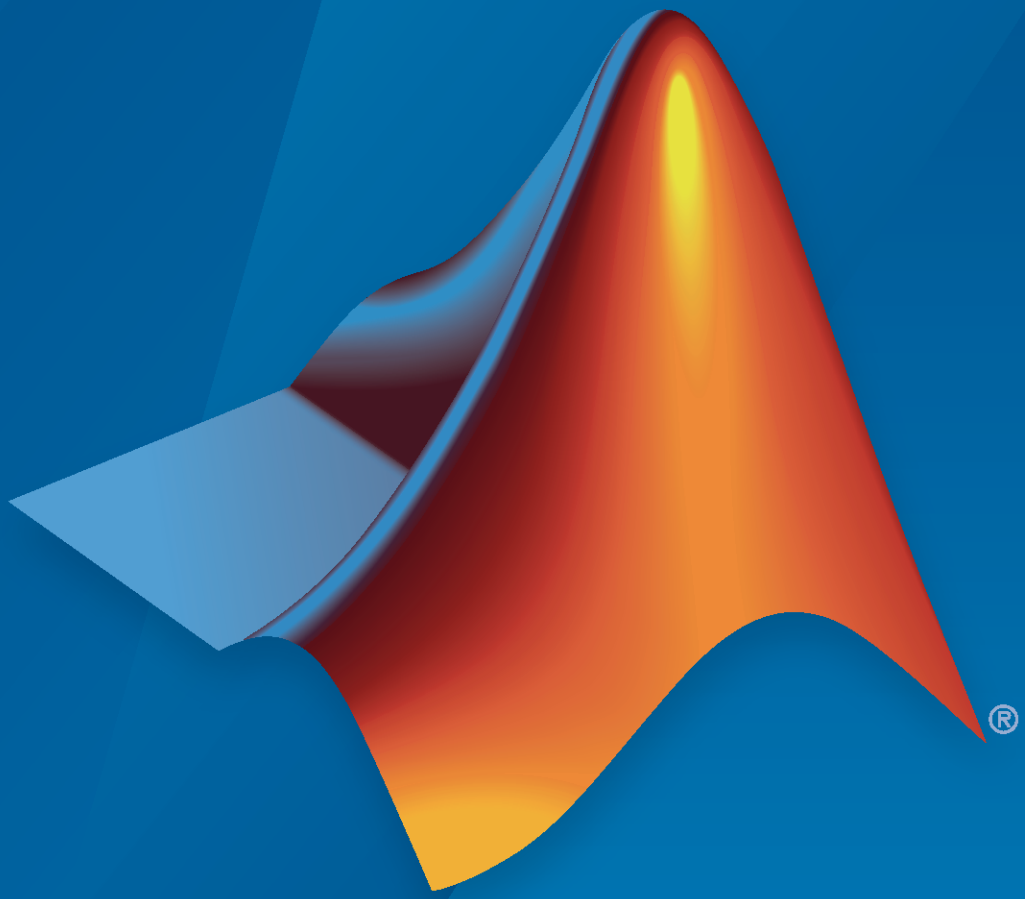
Statistics and Machine Learning Toolbox™

User's Guide



MATLAB®

R2021a

MathWorks®

How to Contact MathWorks

Latest news: `www.mathworks.com`

Sales and services: `www.mathworks.com/sales_and_services`

User community: `www.mathworks.com/matlabcentral`

Technical support: `www.mathworks.com/support/contact_us`

Phone: 508-647-7000

The MathWorks, Inc.
1 Apple Hill Drive
Natick, MA 01760-2098

Statistics and Machine Learning Toolbox™ User's Guide

Revision History

September 1993	First printing	Version 1.0
March 1996	Second printing	Version 2.0
January 1997	Third printing	Version 2.11
November 2000	Fourth printing	Revised for Version 3.0 (Release 12)
May 2001	Fifth printing	Minor revisions
July 2002	Sixth printing	Revised for Version 4.0 (Release 13)
February 2003	Online only	Revised for Version 4.1 (Release 13.0.1)
June 2004	Seventh printing	Revised for Version 5.0 (Release 14)
October 2004	Online only	Revised for Version 5.0.1 (Release 14SP1)
March 2005	Online only	Revised for Version 5.0.2 (Release 14SP2)
September 2005	Online only	Revised for Version 5.1 (Release 14SP3)
March 2006	Online only	Revised for Version 5.2 (Release 2006a)
September 2006	Online only	Revised for Version 5.3 (Release 2006b)
March 2007	Eighth printing	Revised for Version 6.0 (Release 2007a)
September 2007	Ninth printing	Revised for Version 6.1 (Release 2007b)
March 2008	Online only	Revised for Version 6.2 (Release 2008a)
October 2008	Online only	Revised for Version 7.0 (Release 2008b)
March 2009	Online only	Revised for Version 7.1 (Release 2009a)
September 2009	Online only	Revised for Version 7.2 (Release 2009b)
March 2010	Online only	Revised for Version 7.3 (Release 2010a)
September 2010	Online only	Revised for Version 7.4 (Release 2010b)
April 2011	Online only	Revised for Version 7.5 (Release 2011a)
September 2011	Online only	Revised for Version 7.6 (Release 2011b)
March 2012	Online only	Revised for Version 8.0 (Release 2012a)
September 2012	Online only	Revised for Version 8.1 (Release 2012b)
March 2013	Online only	Revised for Version 8.2 (Release 2013a)
September 2013	Online only	Revised for Version 8.3 (Release 2013b)
March 2014	Online only	Revised for Version 9.0 (Release 2014a)
October 2014	Online only	Revised for Version 9.1 (Release 2014b)
March 2015	Online only	Revised for Version 10.0 (Release 2015a)
September 2015	Online only	Revised for Version 10.1 (Release 2015b)
March 2016	Online only	Revised for Version 10.2 (Release 2016a)
September 2016	Online only	Revised for Version 11 (Release 2016b)
March 2017	Online only	Revised for Version 11.1 (Release 2017a)
September 2017	Online only	Revised for Version 11.2 (Release 2017b)
March 2018	Online only	Revised for Version 11.3 (Release 2018a)
September 2018	Online only	Revised for Version 11.4 (Release 2018b)
March 2019	Online only	Revised for Version 11.5 (Release 2019a)
September 2019	Online only	Revised for Version 11.6 (Release 2019b)
March 2020	Online only	Revised for Version 11.7 (Release 2020a)
September 2020	Online only	Revised for Version 12.0 (Release 2020b)
March 2021	Online only	Revised for Version 12.1 (Release 2021a)

Contents

1 Getting Started

2 Organizing Data

3 Descriptive Statistics

4 Statistical Visualization

5 Probability Distributions

6 Gaussian Processes

7 Random Number Generation

8 Hypothesis Tests

9 Analysis of Variance

10 Bayesian Optimization

11 Parametric Regression Analysis

12 Generalized Linear Models

13 Nonlinear Regression

14 Survival Analysis

15 Multivariate Methods

16 Cluster Analysis

19 Decision Trees

20 Discriminant Analysis

21 Naive Bayes

22 Classification and Regression for High-Dimensional Data

23 Classification Learner

24 Regression Learner

25 Support Vector Machines

26 Incremental Learning

27 Markov Models

28 Design of Experiments

29 Statistical Process Control

30 Tall Arrays

31 Parallel Statistics

32 Code Generation

C Bibliography

Getting Started

Statistics and Machine Learning Toolbox Product Description

Analyze and model data using statistics and machine learning

Statistics and Machine Learning Toolbox provides functions and apps to describe, analyze, and model data. You can use descriptive statistics, visualizations, and clustering for exploratory data analysis, fit probability distributions to data, generate random numbers for Monte Carlo simulations, and perform hypothesis tests. Regression and classification algorithms let you draw inferences from data and build predictive models either interactively, using the Classification and Regression Learner apps, or programmatically, using AutoML.

For multidimensional data analysis and feature extraction, the toolbox provides principal component analysis (PCA), regularization, dimensionality reduction, and feature selection methods that let you identify variables with the best predictive power.

The toolbox provides supervised, semi-supervised and unsupervised machine learning algorithms, including support vector machines (SVMs), boosted decision trees, *k*-means, and other clustering methods. You can apply interpretability techniques such as partial dependence plots and LIME, and automatically generate C/C++ code for embedded deployment. Many toolbox algorithms can be used on data sets that are too big to be stored in memory.

Supported Data Types

Statistics and Machine Learning Toolbox supports the following data types for input arguments:

- Numeric scalars, vectors, matrices, or arrays having single- or double-precision entries. These data forms have data type `single` or `double`. Examples include response variables, predictor variables, and numeric values.
- Cell arrays of character vectors; character, string, logical, or categorical arrays; or numeric vectors for categorical variables representing grouping data. These data forms have data types `cell` (specifically `cellstr`), `char`, `string`, `logical`, `categorical`, and `single` or `double`, respectively. An example is an array of class labels in machine learning.
  - You can also use nominal or ordinal arrays for categorical data. However, the `nominal` and `ordinal` data types are not recommended. To work with nominal or ordinal categorical data, use the `categorical` data type instead.
  - You can use signed or unsigned integers, e.g., `int8` or `uint8`. However:
    - Estimation functions might not support signed or unsigned integer data types for nongrouping data.
    - If you recast a `single` or `double` numeric vector containing `NaN` values to a signed or unsigned integer, then the software converts the `NaN` elements to `0`.
- Some functions support tabular arrays for heterogeneous data (for details, see “Tables”). The `table` data type contains variables of any of the data types previously listed. An example is mixed categorical and numerical predictor data for regression analysis.
  - For some functions, you can also use dataset arrays for heterogeneous data. However, the `dataset` data type is not recommended. To work with heterogeneous data, use the `table` data type if the estimation function supports it.
  - Functions that do not support the `table` data type support sample data of type `single` or `double`, e.g., matrices.
- Some functions accept `gpuArray` input arguments so that they execute on the GPU. For the full list of Statistics and Machine Learning Toolbox functions that accept GPU arrays, see Function List (GPU Arrays).
- Some functions accept `tall` array input arguments to work with large data sets. For the full list of Statistics and Machine Learning Toolbox functions that accept tall arrays, see Function List (Tall Arrays).
- Some functions accept sparse matrices, i.e., matrix `A` such that `issparse(A)` returns `1`. For functions that do not accept sparse matrices, recast the data to a full matrix by using `full`.

Statistics and Machine Learning Toolbox does not support the following data types:

- Complex numbers.
- Custom numeric data types, e.g., a variable that is double precision and an object.
- Signed or unsigned numeric integers for nongrouping data, e.g., `unint8` and `int16`.

Note If you specify data of an unsupported type, then the software might return an error or unexpected results.

Organizing Data

Other MATLAB Functions Supporting Nominal and Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the "Categorical Arrays" data type instead.

Notable functions that operate on nominal and ordinal arrays are listed in `nominal` and `ordinal`. In addition to these, many other functions in MATLAB operate on nominal and ordinal arrays in much the same way that they operate on other arrays. A few functions might exhibit special behavior when operating on nominal and ordinal arrays:

- If multiple input arguments are nominal or ordinal arrays, the function often requires that they have the same set of categories, including order if ordinal.
- Relational functions, such as `max` and `gt`, require that the input arrays be `ordinal`.

The following table lists MATLAB functions that operate on nominal and ordinal arrays in addition to other arrays.

`size`	`isequal`	`intersect`	`histogram`	`double`
`length`	`isequaln`	`ismember`	`pietimes`	`single`
`ndims`		`setdiff`		`int8`
`numel`	`eq`	`setxor`	`sort`	`int16`
	`ne`	`unique`	`sortrows`	`int32`
`isrow`	`lt`	`union`	`issorted`	`int64`
`iscolumn`	`le`			`uint8`
	`ge`		`permute`	`uint16`
`cat`	`gt`		`reshape`	`uint32`
`horzcat`			`transpose`	`uint64`
`vertcat`	`min`		`ctranspose`	`char`
	`max`			`cellstr`
	`median`			
	`mode`			

See Also

`nominal` | `ordinal`

Create Nominal and Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

In this section...

Create Nominal Arrays

This example shows how to create nominal arrays using `nominal`.

Load sample data.

The variable `species` is a 150-by-1 cell array of character vectors containing the species name for each observation. The unique species types are setosa, versicolor, and virginica.

```
load fisheriris
unique(species)
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

Create a nominal array.

Convert `species` to a nominal array using the categories occurring in the data.

```
speciesNom = nominal(species);
class(speciesNom)
```

```
ans =
'nominal'
```

Explore category levels.

The nominal array, `speciesNom`, has three levels corresponding to the three unique species. The levels of a nominal array are the set of possible values that its elements can take.

```
getlevels(speciesNom)
```

```
ans = 1x3 nominal
     setosa      versicolor      virginica
```

A nominal array can have more levels than actually appear in the data. For example, a nominal array named `AllSizes` might have levels `small`, `medium`, and `large`, but you might only have observations that are `medium` and `large` in your data. To see which levels of a nominal array are actually present in the data, use `unique`, for instance, `unique(AllSizes)`.

Explore category labels.

Each level has a label. By default, `nominal` labels the category levels with the values occurring in the data. For `speciesNom`, these labels are the species types.

```
getlabels(speciesNom)
```

```
ans = 1x3 cell
    {'setosa'}    {'versicolor'}    {'virginica'}
```

Specify your own category labels.

You can specify your own labels for each category level. You can specify labels when you create the nominal array.

```
speciesNom2 = nominal(species,{'seto','vers','virg'});
getlabels(speciesNom2)
```

```
ans = 1x3 cell
    {'seto'}    {'vers'}    {'virg'}
```

You can also change category labels on an existing nominal array using `setlabels`

Verify new category labels.

Verify that the new labels correspond to the original labels in `speciesNom`.

```
isequal(speciesNom=='setosa',speciesNom2=='seto')
```

```
ans = logical
   1
```

The logical value `1` indicates that the two labels, `'setosa'` and `'seto'`, correspond to the same observations.

Create Ordinal Arrays

This example shows how to create ordinal arrays using `ordinal`.

Load sample data.

```
AllSizes = {'medium','large','small','small','medium',...
            'large','medium','small'};
```

The created variable, `AllSizes`, is a cell array of character vectors containing size measurements on eight objects.

Create an ordinal array.

Create an ordinal array with category levels and labels corresponding to the values in the cell array (the default levels and labels).

```
sizeOrd = ordinal(AllSizes);
getlevels(sizeOrd)
```

```
ans = 1x3 ordinal
     large      medium      small
```

Explore category labels.

By default, `ordinal` uses the original character vectors as category labels. The default order of the categories is ascending alphabetical order.

```
getlabels(sizeOrd)
```

```
ans = 1x3 cell
    {'large'}    {'medium'}    {'small'}
```

Add additional categories.

Suppose that you want to include additional levels for the ordinal array, `xsmall` and `xlarge`, even though they do not occur in the original data. To specify additional levels, use the third input argument to `ordinal`.

```
sizeOrd2 = ordinal(AllSizes,{},...
                   {'xsmall','small','medium','large','xlarge'});
getlevels(sizeOrd2)
```

```
ans = 1x5 ordinal
     xsmall      small      medium      large      xlarge
```

Explore category labels.

To see which levels are actually present in the data, use `unique`.

```
unique(sizeOrd2)
```

```
ans = 1x3 ordinal
     small      medium      large
```

Specify the category order.

Convert `AllSizes` to an ordinal array with categories `small` < `medium` < `large`. Generally, an ordinal array is distinct from a nominal array because there is a natural ordering for levels of an ordinal array. You can use the third input argument to `ordinal` to specify the ascending order of the levels. Here, the order of the levels is smallest to largest.

```
sizeOrd = ordinal(AllSizes,{},{'small','medium','large'});
getlevels(sizeOrd)
```

```
ans = 1x3 ordinal
     small      medium      large
```

The second input argument for `ordinal` is a list of labels for the category levels. When you use braces `{}` for the level labels, `ordinal` uses the labels specified in the third input argument (the labels come from the levels present in the data if only one input argument is used).

Compare elements.

Verify that the first object (with size `medium`) is smaller than the second object (with size `large`).

```
sizeOrd(1) < sizeOrd(2)
```

```
ans = logical
   1
```

The logical value `1` indicates that the inequality holds.

See Also

`getlabels` | `getlevels` | `nominal` | `ordinal`

Related Examples

- “Change Category Labels” on page 2-7
- “Reorder Category Levels” on page 2-9
- “Merge Category Levels” on page 2-16
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Change Category Labels

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Change Category Labels

This example shows how to change the labels for category levels in categorical arrays using `setlabels`. You also have the option to specify labels when creating a categorical array.

Load sample data.

The variable `Cylinders` contains the number of cylinders in 100 sample cars.

```
load carsmall
unique(Cylinders)
```

```
ans = 3×1

     4
     6
     8
```

The sample has 4-, 6-, and 8-cylinder cars.

Create an ordinal array.

Convert `Cylinders` to a nominal array with the default category labels (taken from the values in the data).

```
cyl  = ordinal(Cylinders);
getlabels(cyl)
```

```
ans = 1x3 cell
    {'4'}    {'6'}    {'8'}
```

`ordinal` created labels using the integer values in `Cylinders`, but you should provide labels for numeric data.

Change category labels.

Relabel the categories in `cyl` to `Four`, `Six`, and `Eight`.

```
cyl = setlabels(cyl ,{'Four','Six','Eight'});
getlabels(cyl)
```

```
ans = 1x3 cell
    {'Four'}    {'Six'}    {'Eight'}
```

Alternatively, you can specify category labels when you create a nominal or ordinal array using the second input argument, for example by specifying `ordinal(Cylinders, {'Four','Six','Eight'})`.

See Also

`getlabels` | `nominal` | `ordinal` | `setlabels`

Related Examples

- “Reorder Category Levels” on page 2-9
- “Add and Drop Category Levels” on page 2-18
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Reorder Category Levels

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the "Categorical Arrays" data type instead.

In this section...

Reorder Category Levels in Ordinal Arrays

This example shows how to reorder the category levels in an ordinal array using `reorderlevels`.

Load sample data.

```
AllSizes = {'medium','large','small','small','medium',...
            'large','medium','small'};
```

The created variable, `AllSizes`, is a cell array of character vectors containing size measurements on eight objects.

Create an ordinal array.

Convert `AllSizes` to an ordinal array without specifying the order of the category levels.

```
size = ordinal(AllSizes);
getlevels(size)
```

```
ans = 1x3 ordinal
     large      medium      small
```

By default, the categories are ordered by their labels in ascending alphabetical order, `large` < `medium` < `small`.

Compare elements.

Check whether or not the first object (which has size `medium`) is smaller than the second object (which has size `large`).

```
size(1) < size(2)
```

```
ans = logical
   0
```

The logical value `0` indicates that the medium object is not smaller than the large object.

Reorder category levels.

Reorder the category levels so that `small` < `medium` < `large`.

```
size = reorderlevels(size,{'small','medium','large'});
getlevels(size)
```

```
ans = 1x3 ordinal
     small      medium      large
```

Compare elements.

Verify that the first object is now smaller than the second object.

```
size(1) < size(2)
```

```
ans = logical
   1
```

The logical value 1 indicates that the expected inequality now holds.

Reorder Category Levels in Nominal Arrays

This example shows how to reorder the category levels in nominal arrays using `reorderlevels`. By definition, nominal array categories have no natural ordering. However, you might want to change the order of levels for display or analysis purposes. For example, when fitting a regression model with categorical covariates, `fitlm` uses the first level of a nominal independent variable as the reference category.

Load sample data.

The dataset array, `hospital`, contains variables measured on 100 sample patients. The variable `Weight` contains the weight of each patient. The variable `Sex` is a nominal variable containing the gender, `Male` or `Female`, for each patient.

```
load hospital
getlevels(hospital.Sex)
```

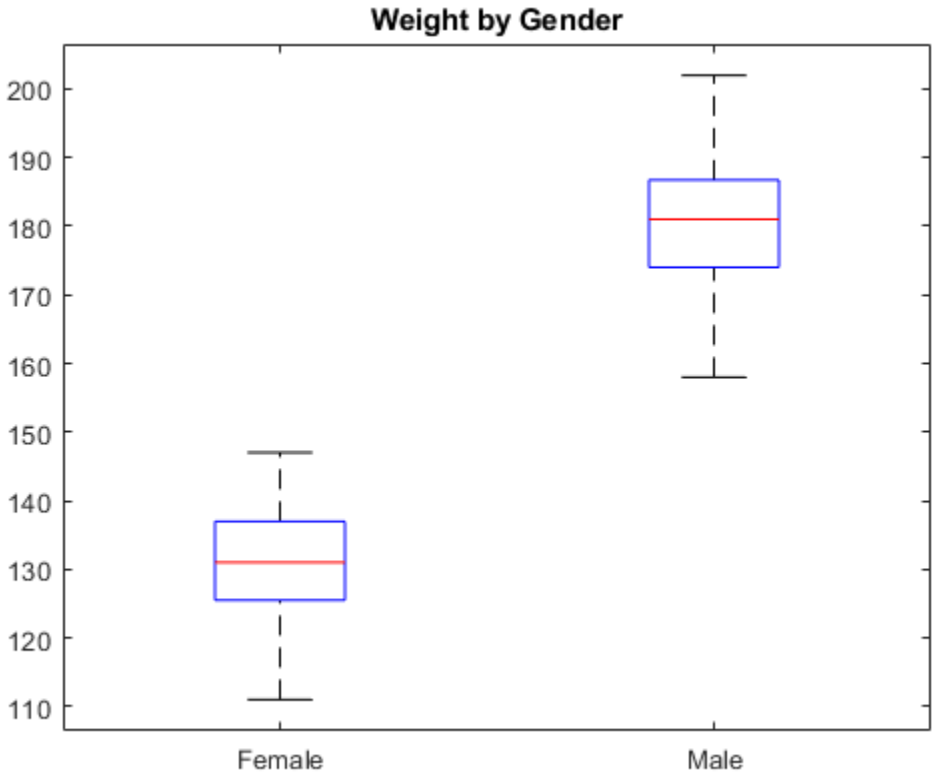
```
ans = 1x2 nominal
     Female      Male
```

By default, the order of the nominal categories is in ascending alphabetical order of the labels.

Plot data grouped by category level.

Draw box plots of weight, grouped by gender.

```
figure
boxplot(hospital.Weight,hospital.Sex)
title('Weight by Gender')
```



The box plots appear in the same alphabetical order returned by `getlevels`.

Change the category order.

Change the order of the category levels.

```
hospital.Sex = reorderlevels(hospital.Sex,{'Male','Female'});
getlevels(hospital.Sex)
```

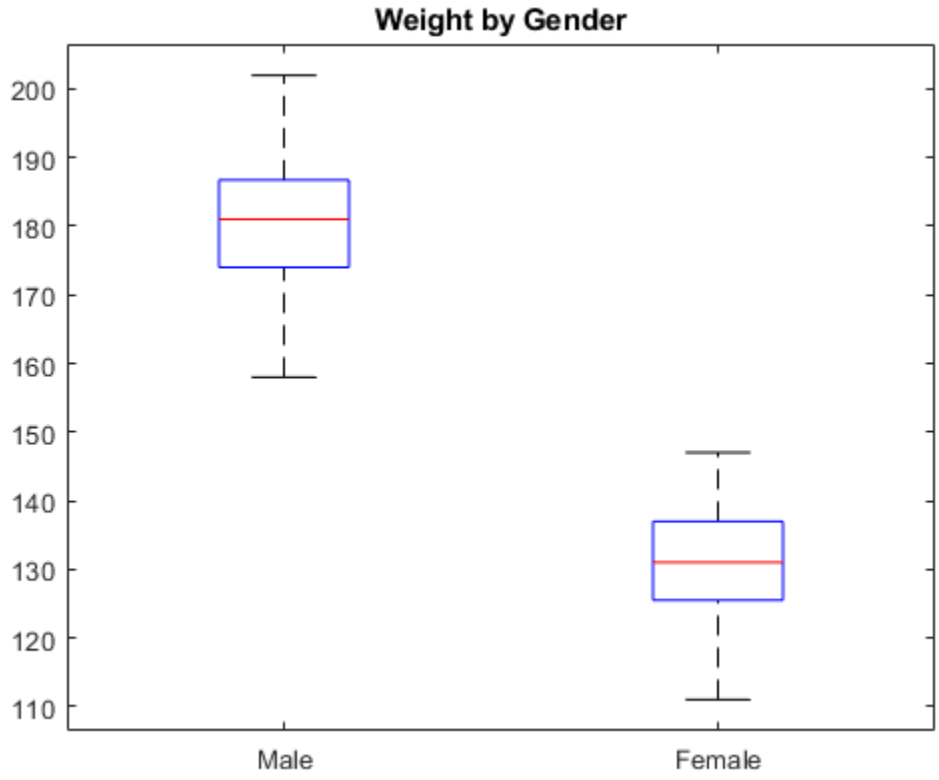
```
ans = 1x2 nominal
     Male      Female
```

The levels are in the newly specified order.

Plot data in new order.

Draw box plots of weight by gender.

```
figure
boxplot(hospital.Weight,hospital.Sex)
title('Weight by Gender')
```



The order of the box plots corresponds to the new level order.

See Also

`fitlm` | `getlevels` | `nominal` | `ordinal` | `reorderlevels`

Related Examples

- “Change Category Labels” on page 2-7
- “Merge Category Levels” on page 2-16
- “Add and Drop Category Levels” on page 2-18
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Categorize Numeric Data

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Categorize Numeric Data

This example shows how to categorize numeric data into a categorical ordinal array using `ordinal`. This is useful for discretizing continuous data.

Load sample data.

The dataset array, `hospital`, contains variables measured on a sample of patients. Compute the minimum, median, and maximum of the variable `Age`.

```
load hospital
quantile(hospital.Age,[0,.5,1])
```

```
ans = 1×3

    25    39    50
```

The patient ages range from 25 to 50.

Convert a numeric array to an ordinal array.

Group patients into the age categories `Under 30`, `30-39`, `Over 40`.

```
hospital.AgeCat = ordinal(hospital.Age,{'Under 30','30-39','Over 40'},...
                          [],[25,30,40,50]);
getlevels(hospital.AgeCat)
```

```
ans = 1x3 ordinal
     Under 30      30-39      Over 40
```

The last input argument to `ordinal` has the endpoints for the categories. The first category begins at age 25, the second at age 30, and so on. The last category contains ages 40 and above, so begins at 40 and ends at 50 (the maximum age in the data set). To specify three categories, you must specify four endpoints (the last endpoint is the upper bound of the last category).

Explore categories.

Display the age and age category for the second patient.

```
dataset({hospital.Age(2),'Age'},...
        {hospital.AgeCat(2),'AgeCategory'})
```

```
ans =
    Age    AgeCategory
    43     Over 40
```

When you discretize a numeric array into categories, the categorical array loses all information about the actual numeric values. In this example, `AgeCat` is not numeric, and you cannot recover the raw data values from it.

Categorize a numeric array into quartiles.

The variable `Weight` has weight measurements for the sample patients. Categorize the patient weights into four categories, by quartile.

```
p = 0:.25:1;
breaks = quantile(hospital.Weight,p);
hospital.WeightQ = ordinal(hospital.Weight,{'Q1','Q2','Q3','Q4'},...
                   [],breaks);
getlevels(hospital.WeightQ)

ans = 1x4 ordinal
     Q1      Q2      Q3      Q4
```

Explore categories.

Display the weight and weight quartile for the second patient.

```
dataset({hospital.Weight(2),'Weight'},...
        {hospital.WeightQ(2),'WeightQuartile'})

ans =
    Weight    WeightQuartile
    163       Q3
```

Summary statistics grouped by category levels.

Compute the mean systolic and diastolic blood pressure for each age and weight category.

```
grpstats(hospital,{'AgeCat','WeightQ'},'mean','DataVars','BloodPressure')

ans =
                   AgeCat      WeightQ    GroupCount    mean_BloodPressure
    Under 30_Q1    Under 30    Q1          6            123.17      79.667
    Under 30_Q2    Under 30    Q2          3            120.33      79.667
    Under 30_Q3    Under 30    Q3          2             127.5        86.5
    Under 30_Q4    Under 30    Q4          4               122          78
    30-39_Q1       30-39       Q1         12            121.75       81.75
    30-39_Q2       30-39       Q2          9            119.56      82.556
    30-39_Q3       30-39       Q3          9               121      83.222
    30-39_Q4       30-39       Q4         11            125.55      87.273
    Over 40_Q1     Over 40     Q1          7            122.14      84.714
    Over 40_Q2     Over 40     Q2         13            123.38      79.385
    Over 40_Q3     Over 40     Q3         14            123.07      84.643
    Over 40_Q4     Over 40     Q4         10             124.6        85.1
```

The variable `BloodPressure` is a matrix with two columns. The first column is systolic blood pressure, and the second column is diastolic blood pressure. The group in the sample with the

highest mean diastolic blood pressure, `87.273`, is aged 30–39 and in the highest weight quartile, `30-39_Q4`.

See Also

`grpstats` | `ordinal`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3
- “Merge Category Levels” on page 2-16
- “Plot Data Grouped by Category” on page 2-21
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Merge Category Levels

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Merge Category Levels

This example shows how to merge categories in a nominal or ordinal array using `mergelevels`. This is useful for collapsing categories with few observations.

Load sample data.

```
load carsmall
```

Create a nominal array.

The variable `Origin` is a character array containing the country of origin for 100 sample cars. Convert `Origin` to a nominal array.

```
Origin = nominal(Origin);
getlevels(Origin)
```

```
ans = 1x6 nominal
     France      Germany      Italy      Japan      Sweden      USA
```

There are six unique countries of origin in the data.

Tabulate category counts.

Explore the elements of the nominal array.

```
tabulate(Origin)
```

```
    Value    Count   Percent
   France        4      4.00%
  Germany        9      9.00%
    Italy        1      1.00%
    Japan       15     15.00%
   Sweden        2      2.00%
      USA       69     69.00%
```

There are relatively few observations in each European country.

Merge categories.

Merge the categories `France`, `Germany`, `Italy`, and `Sweden` into one category called `Europe`.

```
Origin = mergelevels(Origin,{'France','Germany','Italy','Sweden'},...
                     'Europe');
getlevels(Origin)
```

```
ans = 1x3 nominal
     Europe      Japan      USA
```

The variable `Origin` now has only three category levels.

Tabulate category counts.

Explore the elements of the merged categories.

```
tabulate(Origin)
```

```
   Value    Count   Percent
  Europe       16     16.00%
   Japan       15     15.00%
     USA       69     69.00%
```

The category `Europe` has the 16% of observations that were previously distributed across four countries.

See Also

`mergelevels` | `nominal`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3
- “Add and Drop Category Levels” on page 2-18
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Add and Drop Category Levels

This example shows how to add and drop levels from a nominal or ordinal array.

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Load sample data.

```
load('examgrades')
```

The array `grades` contains exam scores from 0 to 100 on five exams for a sample of 120 students.

Create an ordinal array.

Assign letter grades to each student for each test using these categories.

Grade Range	Letter Grade
100	A+
90–99	A
80–89	B
70–79	C
60–69	D

```
letter = ordinal(grades,{'D','C','B','A','A+'},[],...
                 [60,70,80,90,100,100]);
getlevels(letter)
```

```
ans =

     D      C      B      A      A+
```

There are five grade categories, in the specified order `D < C < B < A < A+`.

Check for undefined categories.

Check whether or not there are any exam scores that do not fall into the five letter categories.

```
any(isundefined(letter))
```

```
ans =

     1     0     1     1     0
```

Recall that there are five exam scores for each student. The previous command returns a logical value for each of the five exams, indicating whether there are any scores that are `<undefined>`. There are scores for the first, third, and fourth exams that are `<undefined>`, that is, missing a category level.

Identify elements in undefined categories.

You can find the exam scores that do not have a letter grade using the `isundefined` logical condition.

```
grades(isundefined(letter))

ans =

    55
    59
    58
    59
    54
    57
    56
    59
    59
    50
    59
    52
```

The exam scores that are in the 50s do not have a letter grade.

Add a new category.

Put all scores that are `<undefined>` into a new category labeled `D-`.

```
letter(isundefined(letter)) = 'D-';
getlevels(letter)

Warning: Categorical level 'D-' being added.
> In categorical.subsasgn at 55

ans =

     D      C      B      A      A+      D-
```

The ordinal variable, `letter`, has a new category added to the end.

Reorder category levels.

Reorder the categories so that `D- < D`.

```
letter = reorderlevels(letter,{'D-','D','C','B','A','A+'});
getlevels(letter)

ans =

     D-      D      C      B      A      A+
```

Compare elements.

Now that all exam scores have a letter grade, count how many students received a higher letter grade on the second test than on the first test.

```
sum(letter(:,2) > letter(:,1))

ans =

    32
```

Thirty-two students improved their letter grade between the first two exams.

Explore categories.

Count the number of A+ scores in each of the five exams.

```
sum(letter=='A+')
```

```
ans =

     0     0     0     0     0
```

There are no A+ scores on any of the five exams.

Drop a category.

Drop the category A+ from the ordinal variable, `letter`.

```
letter = droplevels(letter,'A+');
getlevels(letter)
```

```
ans =

     D-      D      C      B      A
```

Category A+ is no longer in the ordinal variable, `letter`.

See Also

`droplevels` | `ordinal` | `reorderlevels`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3
- “Reorder Category Levels” on page 2-9
- “Merge Category Levels” on page 2-16
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Plot Data Grouped by Category

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Plot Data Grouped by Category

This example shows how to plot data grouped by the levels of a categorical variable.

Load sample data.

```
load carsmall
```

The variable `Acceleration` contains acceleration measurements on 100 sample cars. The variable `Origin` is a character array containing the country of origin for each car.

Create a nominal array.

Convert `Origin` to a nominal array.

```
Origin = nominal(Origin);
getlevels(Origin)
```

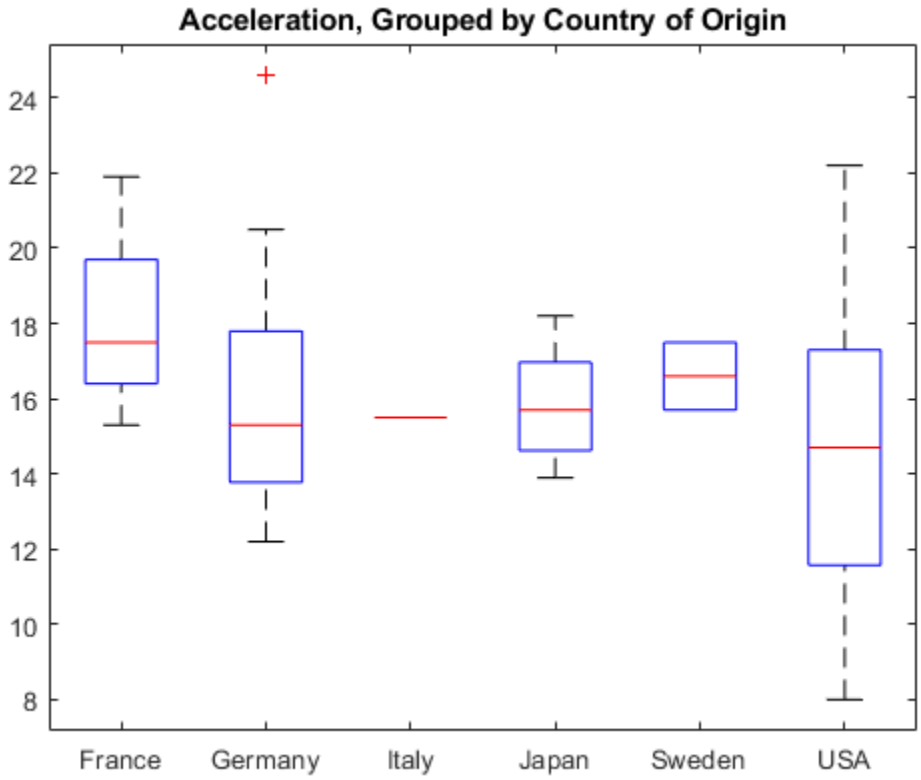
```
ans = 1x6 nominal
     France      Germany      Italy      Japan      Sweden      USA
```

There are six unique countries of origin in the sample. By default, `nominal` orders the countries in ascending alphabetical order.

Plot data grouped by category.

Draw box plots for `Acceleration`, grouped by `Origin`.

```
figure
boxplot(Acceleration,Origin)
title('Acceleration, Grouped by Country of Origin')
```



The box plots appear in the same order as the categorical levels (use `reorderlevels` to change the order of the categories).

Few observations have Italy as the country of origin.

Tabulate category counts.

Tabulate the number of sample cars from each country.

```
tabulate(Origin)
```

```
    Value    Count   Percent
   France        4      4.00%
  Germany        9      9.00%
    Italy        1      1.00%
    Japan       15     15.00%
   Sweden        2      2.00%
      USA       69     69.00%
```

Only one car is made in Italy.

Drop a category.

Delete the Italian car from the sample.

```
Acceleration2 = Acceleration(Origin~='Italy');
Origin2 = Origin(Origin~='Italy');
getlevels(Origin2)
```

```
ans = 1x6 nominal
     France      Germany      Italy      Japan      Sweden      USA
```

Even though the car from Italy is no longer in the sample, the nominal variable, `Origin2`, still has the category `Italy`. Note that this is intentional—the levels of a categorical array do not necessarily coincide with the values.

Drop a category level.

Use `droplevels` to remove the `Italy` category.

```
Origin2 = droplevels(Origin2,'Italy');
tabulate(Origin2)
```

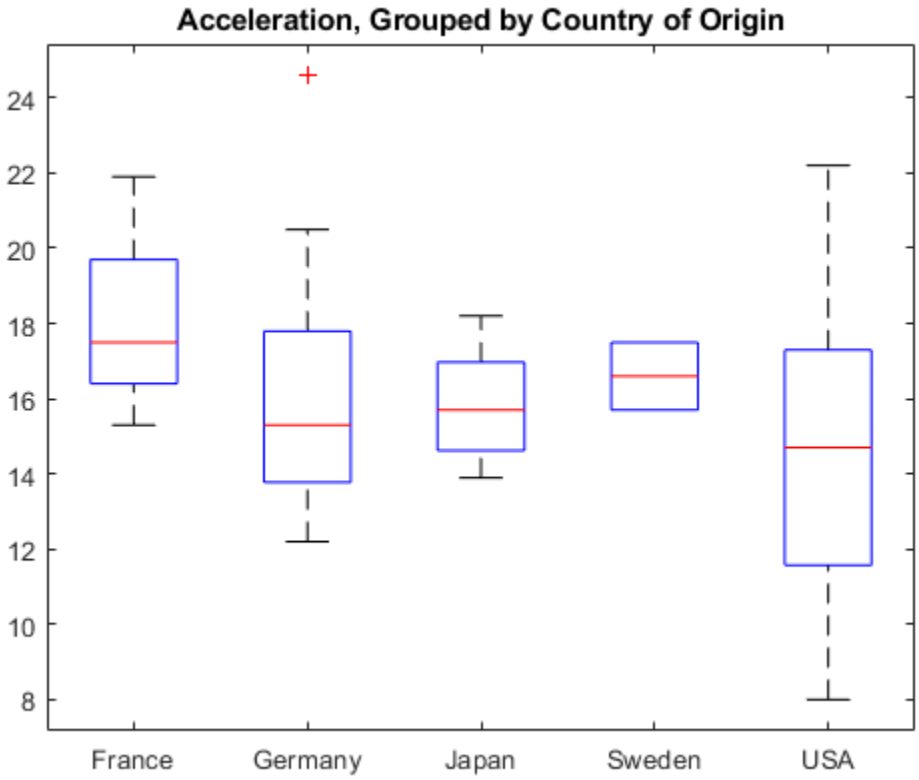
```
    Value    Count   Percent
   France        4      4.04%
  Germany        9      9.09%
    Japan       15     15.15%
   Sweden        2      2.02%
      USA       69     69.70%
```

The `Italy` category is no longer in the nominal array, `Origin2`.

Plot data grouped by category.

Draw box plots of `Acceleration2`, grouped by `Origin2`.

```
figure
boxplot(Acceleration2,Origin2)
title('Acceleration, Grouped by Country of Origin')
```



The plot no longer includes the car from Italy.

See Also

`boxplot` | `droplevels` | `nominal` | `reorderlevels`

Related Examples

- “Test Differences Between Category Means” on page 2-25
- “Summary Statistics Grouped by Category” on page 2-32
- “Linear Regression with Categorical Covariates” on page 2-52

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38
- “Grouping Variables” on page 2-45

Test Differences Between Category Means

This example shows how to test for significant differences between category (group) means using a *t*-test, two-way ANOVA (analysis of variance), and ANOCOVA (analysis of covariance) analysis.

The goal is determining if the expected miles per gallon for a car depends on the decade in which it was manufactured, or the location where it was manufactured.

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Load sample data.

```
load('carsmall')
unique(Model_Year)
```

```
ans =

    70
    76
    82
```

The variable `MPG` has miles per gallon measurements on a sample of 100 cars. The variables `Model_Year` and `Origin` contain the model year and country of origin for each car.

The first factor of interest is the decade of manufacture. There are three manufacturing years in the data.

Create a factor for the decade of manufacture.

Create an ordinal array named `Decade` by merging the observations from years `70` and `76` into a category labeled `1970s`, and putting the observations from `82` into a category labeled `1980s`.

```
Decade = ordinal(Model_Year,{'1970s','1980s'},[],[70 77 82]);
getlevels(Decade)
```

```
ans =

     1970s      1980s
```

Plot data grouped by category.

Draw a box plot of miles per gallon, grouped by the decade of manufacture.

```
figure()
boxplot(MPG,Decade)
title('Miles per Gallon, Grouped by Decade of Manufacture')
```

The box plot suggests that miles per gallon is higher in cars manufactured during the 1980s compared to the 1970s.

Compute summary statistics.

Compute the mean and variance of miles per gallon for each decade.

```
[xbar,s2,grp] = grpstats(MPG,Decade,{'mean','var','gname'})
```

```
xbar =

   19.7857
   31.7097


s2 =

   35.1429
   29.0796


grp =

    '1970s'
    '1980s'
```

This output shows that the mean miles per gallon in the 1980s was `31.71`, compared to `19.79` in the 1970s. The variances in the two groups are similar.

Conduct a two-sample t-test for equal group means.

Conduct a two-sample *t*-test, assuming equal variances, to test for a significant difference between the group means. The hypothesis is

$H_0: \mu_{70} = \mu_{80}$

$H_A: \mu_{70} \neq \mu_{80}$.

```
MPG70 = MPG(Decade=='1970s');
MPG80 = MPG(Decade=='1980s');
[h,p] = ttest2(MPG70,MPG80)
```

```
h =

     1


p =

   3.4809e-15
```

The logical value 1 indicates the null hypothesis is rejected at the default 0.05 significance level. The p-value for the test is very small. There is sufficient evidence that the mean miles per gallon in the 1980s differs from the mean miles per gallon in the 1970s.

Create a factor for the location of manufacture.

The second factor of interest is the location of manufacture. First, convert `Origin` to a nominal array.

```
Location = nominal(Origin);
tabulate(Location)
```

```
tabulate(Location)
    Value    Count   Percent
   France        4      4.00%
  Germany        9      9.00%
    Italy        1      1.00%
    Japan       15     15.00%
   Sweden        2      2.00%
      USA       69     69.00%
```

There are six different countries of manufacture. The European countries have relatively few observations.

Merge categories.

Combine the categories `France`, `Germany`, `Italy`, and `Sweden` into a new category named `Europe`.

```
Location = mergelevels(Location, ...
    {'France','Germany','Italy','Sweden'},'Europe');
tabulate(Location)
```

```
   Value    Count   Percent
   Japan       15     15.00%
     USA       69     69.00%
  Europe       16     16.00%
```

Compute summary statistics.

Compute the mean miles per gallon, grouped by the location of manufacture.

```
[xbar,grp] = grpstats(MPG,Location,{'mean','gname'})
```

```
xbar =

   31.8000
   21.1328
   26.6667


grp =

    'Japan'
    'USA'
    'Europe'
```

This result shows that average miles per gallon is lowest for the sample of cars manufactured in the U.S.

Conduct two-way ANOVA.

Conduct a two-way ANOVA to test for differences in expected miles per gallon between factor levels for `Decade` and `Location`.

The statistical model is

$$MPG_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2; j = 1, 2, 3,$$

where MPG_{ij} is the response, miles per gallon, for cars made in decade i at location j. The treatment effects for the first factor, decade of manufacture, are the α_i terms (constrained to sum to zero). The treatment effects for the second factor, location of manufacture, are the β_j terms (constrained to sum to zero). The ε_{ij} are uncorrelated, normally distributed noise terms.

The hypotheses to test are equality of decade effects,

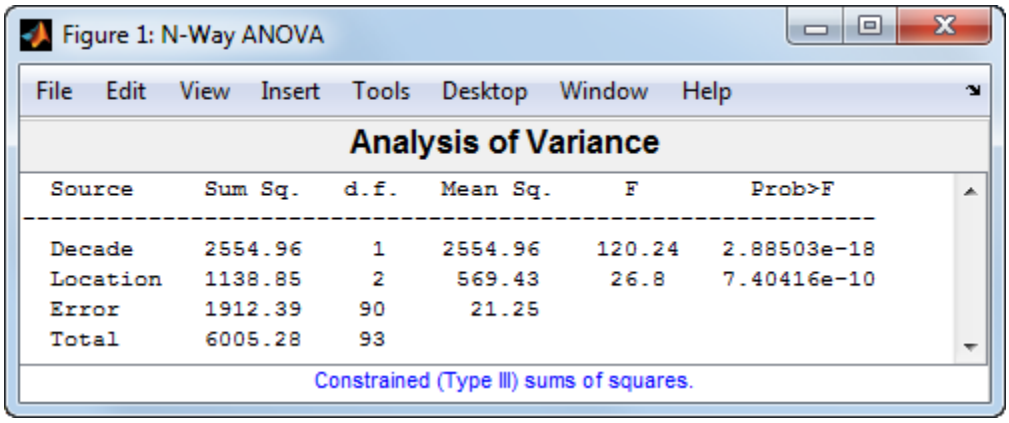
$$H_0: \alpha_1 = \alpha_2 = 0$$
$$H_A: at\ least\ one\ \alpha_i \neq 0,$$

and equality of location effects,

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_A: at\ least\ one\ \beta_j \neq 0.$$

You can conduct a multiple-factor ANOVA using `anovan`.

```
anovan(MPG,{Decade,Location},'varnames',{'Decade','Location'});
```



This output shows the results of the two-way ANOVA. The p-value for testing the equality of decade effects is `2.88503e-18`, so the null hypothesis is rejected at the 0.05 significance level. The p-value for testing the equality of location effects is `7.40416e-10`, so this null hypothesis is also rejected.

Conduct ANOCOVA analysis.

A potential confounder in this analysis is car weight. Cars with greater weight are expected to have lower gas mileage. Include the variable `Weight` as a continuous covariate in the ANOVA; that is, conduct an ANOCOVA analysis.

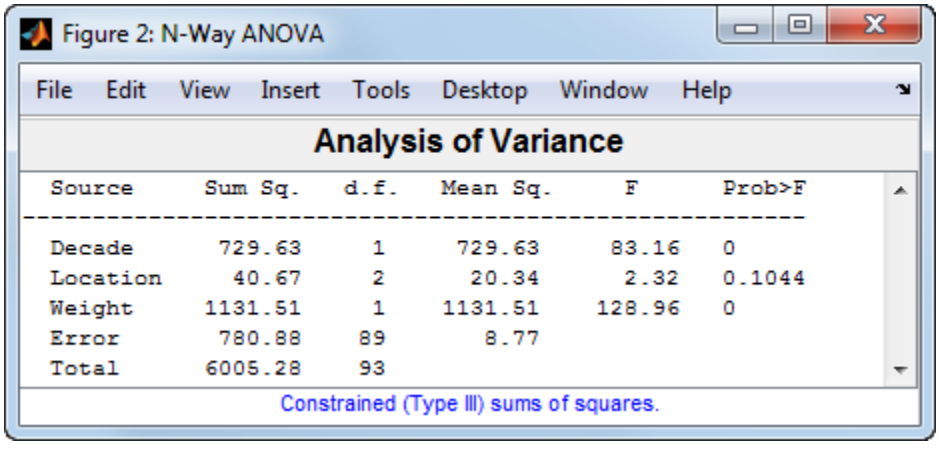
Assuming parallel lines, the statistical model is

$$MPG_{ijk} = \mu + \alpha_i + \beta_j + \gamma Weight_{ijk} + \varepsilon_{ijk}, \quad i = 1, 2; \;\; j = 1, 2, 3; \;\; k = 1, \ldots, 100.$$

The difference between this model and the two-way ANOVA model is the inclusion of the continuous predictor, $Weight_{ijk}$, the weight for the kth car, which was made in the ith decade and in the jth location. The slope parameter is γ.

Add the continuous covariate as a third group in the second `anovan` input argument. Use the name-value pair argument `Continuous` to specify that `Weight` (the third group) is continuous.

```
anovan(MPG,{Decade,Location,Weight},'Continuous',3,...
       'varnames',{'Decade','Location','Weight'});
```



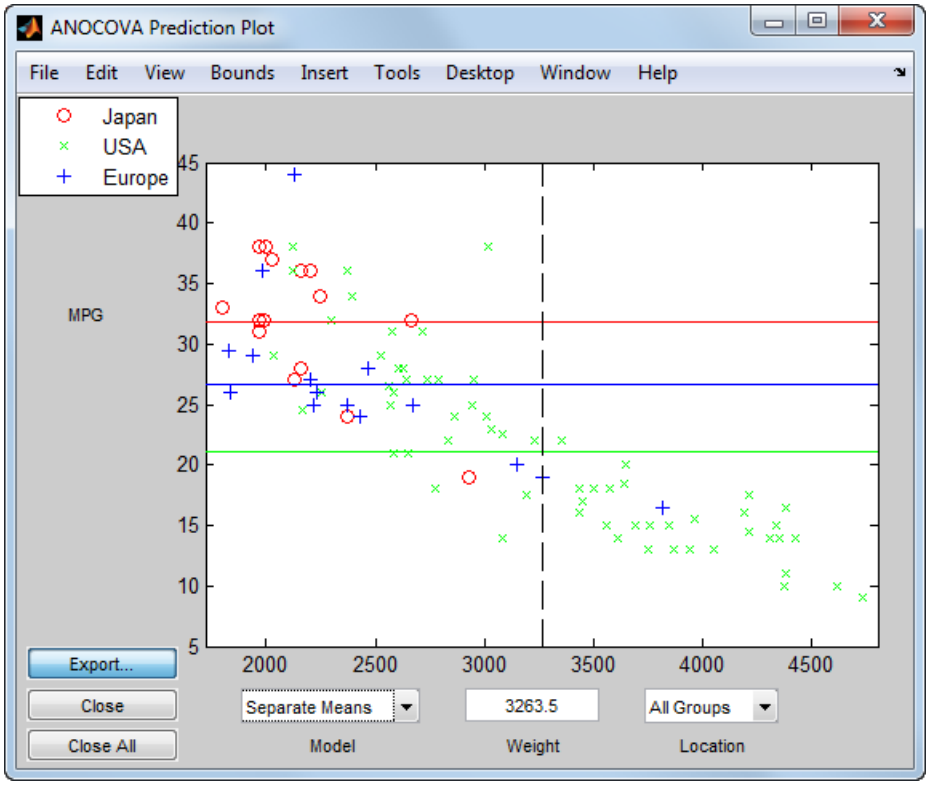
This output shows that when car weight is considered, there is insufficient evidence of a manufacturing location effect (p-value = `0.1044`).

Use interactive tool.

You can use the interactive `aoctool` to explore this result.

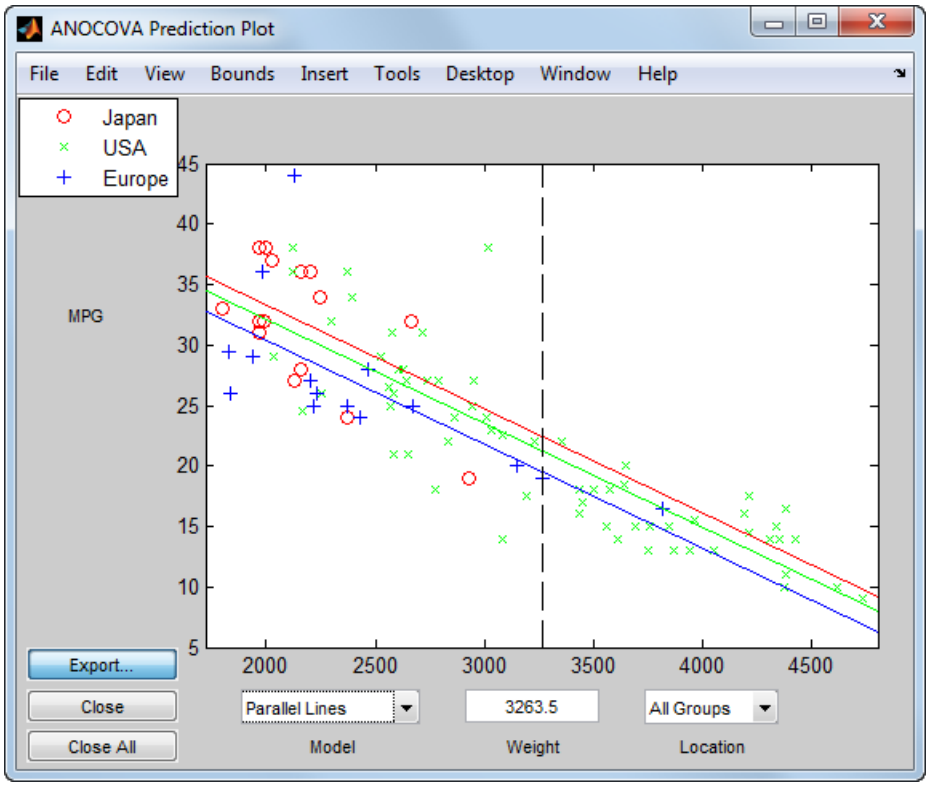
```
aoctool(Weight,MPG,Location);
```

This command opens three dialog boxes. In the ANOCOVA Prediction Plot dialog box, select the **Separate Means** model.



This output shows that when you do not include `Weight` in the model, there are fairly large differences in the expected miles per gallon among the three manufacturing locations. Note that here the model does not adjust for the decade of manufacturing.

Now, select the **Parallel Lines** model.



When you include `Weight` in the model, the difference in expected miles per gallon among the three manufacturing locations is much smaller.

See Also

`anovan` | `aoctool` | `boxplot` | `grpstats` | `nominal` | `ordinal` | `ttest2`

Related Examples

- “Plot Data Grouped by Category” on page 2-21
- “Summary Statistics Grouped by Category” on page 2-32
- “Linear Regression with Categorical Covariates” on page 2-52

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38
- “Grouping Variables” on page 2-45

Summary Statistics Grouped by Category

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Summary Statistics Grouped by Category

This example shows how to compute summary statistics grouped by levels of a categorical variable. You can compute group summary statistics for a numeric array or a dataset array using `grpstats`.

Load sample data.

```
load hospital
```

The dataset array, `hospital`, has 7 variables (columns) and 100 observations (rows).

Compute summary statistics by category.

The variable `Sex` is a nominal array with two levels, `Male` and `Female`. Compute the minimum and maximum weights for each gender.

```
stats = grpstats(hospital,'Sex',{'min','max'},'DataVars','Weight')
```

```
stats =
              Sex       GroupCount    min_Weight    max_Weight
    Female    Female    53            111           147
    Male      Male      47            158           202
```

The dataset array, `stats`, has observations corresponding to the levels of the variable `Sex`. The variable `min_Weight` contains the minimum weight for each group, and the variable `max_Weight` contains the maximum weight for each group.

Compute summary statistics by multiple categories.

The variable `Smoker` is a logical array with value `1` for smokers and value `0` for nonsmokers. Compute the minimum and maximum weights for each gender and smoking combination.

```
stats = grpstats(hospital,{'Sex','Smoker'},{'min','max'},...
                 'DataVars','Weight')
```

```
stats =
                Sex       Smoker    GroupCount    min_Weight    max_Weight
    Female_0    Female    false     40            111           147
    Female_1    Female    true      13            115           146
    Male_0      Male      false     26            158           194
    Male_1      Male      true      21            164           202
```

The dataset array, `stats`, has an observation row for each combination of levels of `Sex` and `Smoker` in the original data.

See Also

`dataset` | `grpstats` | `nominal`

Related Examples

- “Plot Data Grouped by Category” on page 2-21
- “Test Differences Between Category Means” on page 2-25
- “Calculations on Dataset Arrays” on page 2-92

More About

- “Grouping Variables” on page 2-45
- “Nominal and Ordinal Arrays” on page 2-36
- “Dataset Arrays” on page 2-112

Sort Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Sort Ordinal Arrays

This example shows how to determine sorting order for ordinal arrays.

Load sample data.

```
AllSizes = {'medium','large','small','small','medium',...
            'large','medium','small'};
```

The created variable, `AllSizes`, is a cell array of character vectors containing size measurements on eight objects.

Create an ordinal array.

Convert `AllSizes` to an ordinal array with levels `small` < `medium` < `large`.

```
AllSizes = ordinal(AllSizes,{},{'small','medium','large'});
getlevels(AllSizes)
```

```
ans = 1x3 ordinal
     small      medium      large
```

Sort the ordinal array.

When you sort ordinal arrays, the sorted observations are in the same order as the category levels.

```
sizeSort = sort(AllSizes);
sizeSort(:)
```

```
ans = 8x1 ordinal
     small
     small
     small
     medium
     medium
     medium
     large
     large
```

The sorted ordinal array, `sizeSort`, contains the observations ordered from small to large.

See Also

`ordinal`

Related Examples

- “Reorder Category Levels” on page 2-9

- “Add and Drop Category Levels” on page 2-18

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Nominal and Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

In this section...

What Are Nominal and Ordinal Arrays?

Nominal and ordinal arrays are Statistics and Machine Learning Toolbox data types for storing categorical values. Nominal and ordinal arrays store data that have a finite set of discrete levels, which might or might not have a natural order.

- `ordinal` arrays store categorical values with ordered levels. For example, an ordinal variable might have levels {small, medium, large}.
- `nominal` arrays store categorical values with unordered levels. For example, a nominal variable might have levels {red, blue, green}.

In experimental design, these variables are often called factors, with ordered or unordered factor levels.

Nominal and ordinal arrays are convenient and memory efficient containers for storing categorical variables. In addition to storing information about which category each observation belongs to, nominal and ordinal arrays store descriptive metadata including category labels and order.

Nominal and ordinal arrays have associated methods that streamline common tasks such as merging categories, adding or dropping levels, and changing level labels.

Nominal and Ordinal Array Conversion

You can easily convert to and from nominal or ordinal arrays. To create a nominal or ordinal array, use `nominal` or `ordinal`, respectively. You can convert these data types to nominal or ordinal arrays:

- Numeric arrays
- Logical arrays
- Character arrays
- String arrays
- Cell arrays of character vectors

See Also

`nominal` | `ordinal`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3
- “Summary Statistics Grouped by Category” on page 2-32
- “Plot Data Grouped by Category” on page 2-21
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38
- “Grouping Variables” on page 2-45

Advantages of Using Nominal and Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

In this section...

Manipulate Category Levels

When working with categorical variables and their levels, you’ll encounter some typical challenges. This table summarizes the functions you can use with nominal or ordinal arrays to manipulate category levels. For additional functions, type `methods nominal` or `methods ordinal` at the command line, or see the `nominal` and `ordinal` reference pages.

Task	Function
Add new category levels	`addlevels`
Drop category levels	`droplevels`
Combine category levels	`mergelevels`
Reorder category levels	`reorderlevels`
Count the number of observations in each category	`levelcounts`
Change the label or name of category levels	`setlabels`
Create an interaction factor	`times`
Find observations that are not in a defined category	`isundefined`

Analysis Using Nominal and Ordinal Arrays

You can use nominal and ordinal arrays in a variety of statistical analyses. For example, you might want to compute descriptive statistics for data grouped by the category levels, conduct statistical tests on differences between category means, or perform regression analysis using categorical predictors.

Statistics and Machine Learning Toolbox functions that accept a grouping variable as an input argument accept nominal and ordinal arrays. This includes descriptive functions such as:

- `grpstats`
- `gscatter`
- `boxplot`
- `gplotmatrix`

You can also use nominal and ordinal arrays as input arguments to analysis functions and methods based on models, such as:

- `anovan`
- `fitlm`
- `fitglm`

When you use a nominal or ordinal array as a predictor in these functions, the fitting function automatically recognizes the categorical predictor, and constructs appropriate dummy indicator variables for analysis. Alternatively, you can construct your own dummy indicator variables using `dummyvar`.

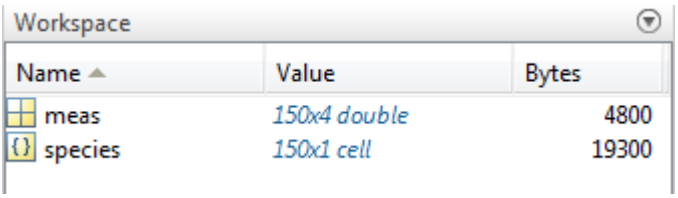
Reduce Memory Requirements

The levels of categorical variables are often defined as text, which can be costly to store and manipulate in a cell array of character vectors or `char` array. Nominal and ordinal arrays separately store category membership and category labels, greatly reducing the amount of memory required to store the variable.

For example, load some sample data:

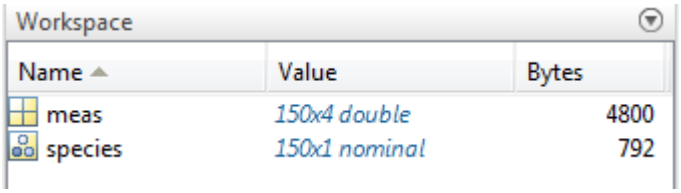
```
load('fisheriris')
```

The variable `species` is a cell array of character vectors requiring 19,300 bytes of memory.



Convert `species` to a nominal array:

```
species = nominal(species);
```



There is a 95% reduction in memory required to store the variable.

See Also

`nominal` | `ordinal`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3
- “Test Differences Between Category Means” on page 2-25
- “Linear Regression with Categorical Covariates” on page 2-52
- “Index and Search Using Nominal and Ordinal Arrays” on page 2-41

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Grouping Variables” on page 2-45
- “Dummy Variables” on page 2-48

Index and Search Using Nominal and Ordinal Arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Index By Category

It is often useful to index and search data by its category, or group. If you store categories as labels inside a cell array of character vectors or `char` array, it can be difficult to index and search the categories. When using nominal or ordinal arrays, you can easily:

- **Index elements from particular categories.** For both nominal and ordinal arrays, you can use the logical operators == and ~= to index the observations that are in, or not in, a particular category. For ordinal arrays, which have an encoded order, you can also use inequalities, >, >=, <, and <=, to find observations in categories above or below a particular category.
- **Search for members of a category.** In addition to the logical operator ==, you can use `ismember` to find observations in a particular group.
- **Find elements that are not in a defined category.** Nominal and ordinal arrays indicate which elements do not belong to a defined category by <undefined>. You can use `isundefined` to find observations missing a category.
- **Delete observations that are in a particular category.** You can use logical operators to include or exclude observations from particular categories. Even if you remove all observations from a category, the category level remains defined unless you remove it using `droplevels`.

Common Indexing and Searching Methods

This example shows several common indexing and searching methods.

Load the sample data.

```
load carsmall;
```

Convert the `char` array, `Origin`, to a nominal array. This variable contains the country of origin, or manufacture, for each sample car.

```
Origin = nominal(Origin);
```

Search for observations in a category. Determine if there are any cars in the sample that were manufactured in Canada.

```
any(Origin=='Canada')
```

```
ans = logical
   0
```

There are no sample cars manufactured in Canada.

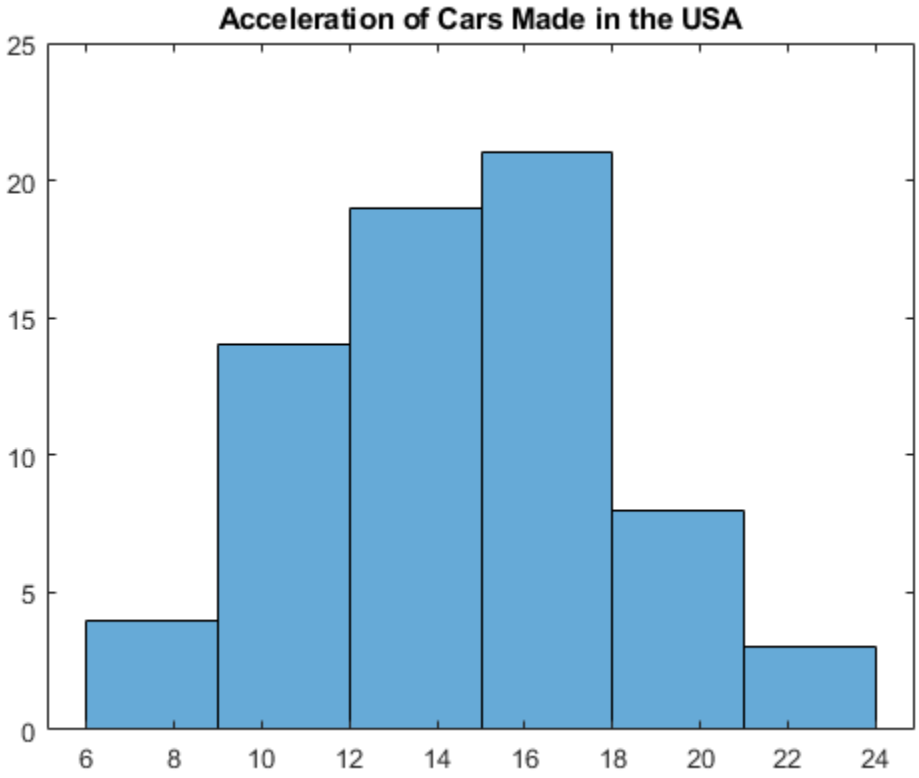
List the countries that are levels of `Origin`.

```
getlevels(Origin)
```

```
ans = 1x6 nominal
     France       Germany       Italy       Japan       Sweden       USA
```

Index elements that are in a particular category. Plot a histogram of the acceleration measurements for cars made in the U.S.

```
figure();
histogram(Acceleration(Origin=='USA'))
title('Acceleration of Cars Made in the USA')
```



Delete observations that are in a particular category. Delete all cars made in Sweden from `Origin`.

```
Origin = Origin(Origin~='Sweden');
any(ismember(Origin,'Sweden'))
```

```
ans = logical
   0
```

The cars made in Sweden are deleted from `Origin`, but `Sweden` is still a level of `Origin`.

```
getlevels(Origin)
```

```
ans = 1x6 nominal
     France       Germany       Italy       Japan       Sweden       USA
```

Remove `Sweden` from the levels of `Origin`.

```
Origin = droplevels(Origin,'Sweden');
getlevels(Origin)
```

```
ans = 1x5 nominal
     France      Germany      Italy      Japan      USA
```

Check for observations not in a defined category. Get the indices for the cars made in France.

```
ix = find(Origin=='France')
```

```
ix = 4×1

    11
    27
    39
    61
```

There are four cars from France. Remove `France` from the levels of `Origin`.

```
Origin = droplevels(Origin,'France');
```

This returns a warning indicating that you are dropping a category level that has elements in it. These observations are no longer in a defined category, indicated by `undefined`.

```
Origin(ix)
```

```
ans = 4x1 nominal
     <undefined>
     <undefined>
     <undefined>
     <undefined>
```

You can use `isundefined` to search for observations with an undefined category.

```
find(isundefined(Origin))
```

```
ans = 4×1

    11
    27
    39
    61
```

These indices correspond to the observations that were in category `France`, before that category was dropped from `Origin`.

See Also

`droplevels` | `nominal` | `ordinal`

Related Examples

- “Create Nominal and Ordinal Arrays” on page 2-3

- “Reorder Category Levels” on page 2-9
- “Merge Category Levels” on page 2-16
- “Add and Drop Category Levels” on page 2-18

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Grouping Variables

In this section...

What Are Grouping Variables?

Grouping variables are utility variables used to group, or categorize, observations. Grouping variables are useful for summarizing or visualizing data by group. A grouping variable can be any of these data types:

- Numeric vector
- Logical vector
- Character array
- String array
- Cell array of character vectors
- Categorical vector

A grouping variable must have the same number of observations (rows) as the table, dataset array, or numeric array you are grouping. Observations that have the same grouping variable value belong to the same group.

For example, the following variables comprise the same groups. Each grouping variable divides five observations into two groups. The first group contains the first and fourth observations. The other three observations are in the second group.

Data Type	Grouping Variable
Numeric vector	`[1 2 2 1 2]`
Logical vector	`[0 1 1 0 1]`
String array	`["Male","Female","Female","Male","Female"]`
Cell array of character vectors	`{'Male','Female','Female','Male','Female'}`
Categorical vector	`Male Female Female Male Female`

Use grouping variables with labels to give each group a meaningful name. A categorical vector is an efficient and flexible choice of grouping variable.

Group Definition

Typically, there are as many groups as unique values in the grouping variable. However, categorical vectors can have levels that are not represented in the data. The groups and the order of the groups depend on the data type of the grouping variable. Suppose `G` is a grouping variable.

- If `G` is a numeric or logical vector, then the groups correspond to the distinct values in `G`, in the sorted order of the unique values.

- If `G` is a character array, string array, or cell array of character vectors, then the groups correspond to the distinct elements in `G`, in the order of their first appearance.
- If `G` is a categorical vector, then the groups correspond to the unique category levels in `G`, in the order returned by `categories`.

Some functions, such as `grpstats`, accept multiple grouping variables specified as a cell array of grouping variables, for example, `{G1,G2,G3}`. In this case, the groups are defined by the unique combinations of values in the grouping variables. The order is decided first by the order of the first grouping variable, then by the order of the second grouping variable, and so on.

Analysis Using Grouping Variables

This table lists common tasks you might want to perform using grouping variables.

Grouping Task	Function Accepting Grouping Variable
Draw side-by-side box plots for data in different groups.	`boxplot`
Draw a scatter plot with markers colored by group.	`gscatter`
Draw a scatter plot matrix with markers colored by group.	`gplotmatrix`
Compute summary statistics by group.	`grpstats`
Test for differences between group means.	`anovan`
Create an index vector from a grouping variable.	`grp2idx`

Missing Group Values

Grouping variables can have missing values provided you include a valid indicator.

Grouping Variable Data Type	Missing Value Indicator
Numeric vector	`NaN`
Logical vector	(Cannot be missing)
Character array	Row of spaces
String array	`<missing>` or `""`
Cell array of character vectors	`''`
Categorical vector	`<undefined>`

See Also

`categorical`

Related Examples

- “Plot Data Grouped by Category” on page 2-21
- “Summary Statistics Grouped by Category” on page 2-32

More About

- “Nominal and Ordinal Arrays” on page 2-36
- “Advantages of Using Nominal and Ordinal Arrays” on page 2-38

Dummy Variables

In this section...

This topic provides an introduction to dummy variables, describes how the software creates them for classification and regression problems, and shows how you can create dummy variables by using the `dummyvar` function.

What Are Dummy Variables?

When you perform classification and regression analysis, you often need to include both continuous (quantitative) and categorical (qualitative) predictor variables. A categorical variable must not be included as a numeric array. Numeric arrays have both order and magnitude. A categorical variable can have order (for example, an ordinal variable), but it does not have magnitude. Using a numeric array implies a known “distance” between the categories. The appropriate way to include categorical predictors is as dummy variables. To define dummy variables, use indicator variables that have the values 0 and 1.

The software chooses one of four schemes to define dummy variables based on the type of analysis, as described in the next sections. For example, suppose you have a categorical variable with three categories: `Cool`, `Cooler`, and `Coolest`.

Full Dummy Variables

Represent the categorical variable with three categories using three dummy variables, one variable for each category.

	X_0	X_1	X_2
Cool	1	0	0
Cooler	0	1	0
Coolest	0	0	1

X_0 is a dummy variable that has the value 1 for `Cool`, and 0 otherwise. X_1 is a dummy variable that has the value 1 for `Cooler`, and 0 otherwise. X_2 is a dummy variable that has the value 1 for `Coolest`, and 0 otherwise.

Dummy Variables with Reference Group

Represent the categorical variable with three categories using two dummy variables with a reference group.

	X_1	X_2
Cool	0	0 ← Reference Group
Cooler	1	0
Coolest	0	1

You can distinguish `Cool`, `Cooler`, and `Coolest` using only X_1 and X_2, without X_0. Observations for `Cool` have 0s for both dummy variables. The category represented by all 0s is the reference group.

Dummy Variables for Ordered Categorical Variable

Assume the mathematical ordering of the categories is `Cool` < `Cooler` < `Coolest`. This coding scheme uses 1 and –1 values, and uses more 1s for higher categories, to indicate the ordering.

	X_1	X_2
Cool	-1	-1
Cooler	1	-1
Coolest	1	1

X_1 is a dummy variable that has the value 1 for `Cooler` and `Coolest`, and –1 for `Cool`. X_2 is a dummy variable that has the value 1 for `Coolest`, and –1 otherwise.

You can indicate that a categorical variable has mathematical ordering by using the `'Ordinal'` name-value pair argument of the `categorical` function.

Dummy Variables Created with Effects Coding

Effects coding uses 1, 0, and –1 to create dummy variables. Instead of using 0 values to represent a reference group, as in “Dummy Variables with Reference Group” on page 2-48, effects coding uses –1 to represent the last category.

	X_1	X_2
Cool	1	0
Cooler	0	1
Coolest	-1	-1

Creating Dummy Variables

Automatic Creation of Dummy Variables

Statistics and Machine Learning Toolbox offers several classification and regression fitting functions that accept categorical predictors. Some fitting functions create dummy variables to handle categorical predictors.

The following is the default behavior of the fitting functions in identifying categorical predictors.

- If the predictor data is in a table, the functions assume that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. The fitting functions that use decision trees assume ordered categorical vectors to be continuous variables.
- If the predictor data is a matrix, the functions assume all predictors are continuous.

To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` or `'CategoricalVars'` name-value pair argument.

The fitting functions handle the identified categorical predictors as follows:

- `fitckernel`, `fitclinear`, `fitcnet`, `fitcsvm`, `fitrgp`, `fitrkernel`, `fitrlinear`, `fitrnet`, and `fitrsvm` use two different schemes to create dummy variables, depending on whether a categorical variable is unordered or ordered.
  - For an unordered categorical variable, the functions use “Full Dummy Variables” on page 2-48.
  - For an ordered categorical variable, the functions use “Dummy Variables for Ordered Categorical Variable” on page 2-49.
- Parametric regression fitting functions such as `fitlm`, `fitglm`, and `fitcox` use “Dummy Variables with Reference Group” on page 2-48. When the functions include the dummy variables, the estimated coefficients of the dummy variables are relative to the reference group. For an example, see “Linear Regression with Categorical Predictor” on page 33-1984.
- `fitlme`, `fitlmematrix` and `fitglme` allow you to specify the scheme for creating dummy variables by using the `'DummyVarCoding'` name-value pair argument. The functions support three schemes: “Full Dummy Variables” on page 2-48 (`'DummyVarCoding','full'`), “Dummy Variables with Reference Group” on page 2-48 (`'DummyVarCoding','reference'`), and “Dummy Variables Created with Effects Coding” on page 2-49 (`'DummyVarCoding','effects'`). Note that these functions do not offer a name-value pair argument for specifying categorical variables.
- `fitrm` uses “Dummy Variables Created with Effects Coding” on page 2-49.
- Other fitting functions that accept categorical predictors use algorithms that can handle categorical predictors without creating dummy variables.

Manual Creation of Dummy Variables

This example shows how to create your own dummy variable design matrix by using the `dummyvar` function. This function accepts grouping variables and returns a matrix containing zeros and ones, whose columns are dummy variables for the grouping variables.

Create a column vector of categorical data specifying gender.

```
gender = categorical({'Male';'Female';'Female';'Male';'Female'});
```

Create dummy variables for `gender`.

```
dv = dummyvar(gender)
```

```
dv = 5×2

     0     1
     1     0
     1     0
     0     1
     1     0
```

`dv` has five rows corresponding to the number of rows in `gender` and two columns for the unique groups, `Female` and `Male`. Column order corresponds to the order of the levels in `gender`. For categorical arrays, the default order is ascending alphabetical. You can check the order by using the `categories` function.

```
categories(gender)
```

```
ans = 2x1 cell
    {'Female'}
```

```
    {'Male'  }
```

To use the dummy variables in a regression model, you must either delete a column (to create a reference group) or fit a regression model with no intercept term. For the gender example, you need only one dummy variable to represent two genders. Notice what happens if you add an intercept term to the complete design matrix `dv`.

```
X = [ones(5,1) dv]
```

```
X = 5×3

     1     0     1
     1     1     0
     1     1     0
     1     0     1
     1     1     0
```

```
rank(X)
```

```
ans = 2
```

The design matrix with an intercept term is not of full rank and is not invertible. Because of this linear dependence, use only c – 1 indicator variables to represent a categorical variable with c categories in a regression model with an intercept term.

See Also

`categorical` | `dummyvar`

Related Examples

- “Linear Regression with Categorical Covariates” on page 2-52
- “Test Differences Between Category Means” on page 2-25

Linear Regression with Categorical Covariates

This example shows how to perform a regression with categorical covariates using categorical arrays and `fitlm`.

Load sample data.

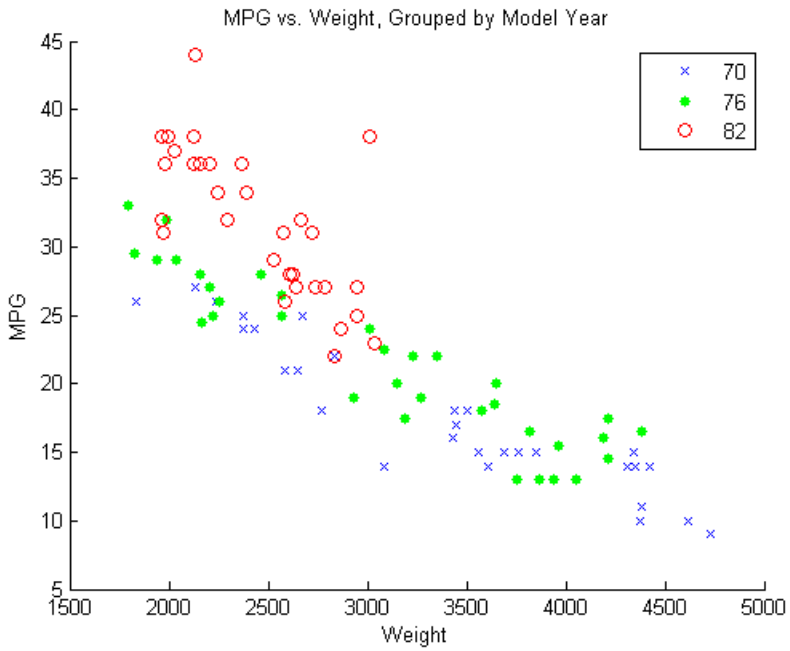
```
load carsmall
```

The variable `MPG` contains measurements on the miles per gallon of 100 sample cars. The model year of each car is in the variable `Model_Year`, and `Weight` contains the weight of each car.

Plot grouped data.

Draw a scatter plot of `MPG` against `Weight`, grouped by model year.

```
figure()
gscatter(Weight,MPG,Model_Year,'bgr','x.o')
title('MPG vs. Weight, Grouped by Model Year')
```



The grouping variable, `Model_Year`, has three unique values, `70`, `76`, and `82`, corresponding to model years 1970, 1976, and 1982.

Create table and categorical array.

Create a table that contains the variables `MPG`, `Weight`, and `Model_Year`. Convert the variable `Model_Year` to a categorical array.

```
cars = table(MPG,Weight,Model_Year);
cars.Model_Year = categorical(cars.Model_Year);
```

Fit a regression model.

Fit a regression model using `fitlm` with `MPG` as the dependent variable, and `Weight` and `Model_Year` as the independent variables. Because `Model_Year` is a categorical covariate with three levels, it should enter the model as two indicator variables.

The scatter plot suggests that the slope of `MPG` against `Weight` might differ for each model year. To assess this, include weight-year interaction terms.

The proposed model is

$$E(MPG) = \beta_0 + \beta_1 Weight + \beta_2 I[1976] + \beta_3 I[1982] + \beta_4 Weight \times I[1976] + \beta_5 Weight \times I[1982],$$

where $I[1976]$ and $I[1982]$ are dummy variables indicating the model years 1976 and 1982, respectively. $I[1976]$ takes the value 1 if model year is 1976 and takes the value 0 if it is not. $I[1982]$ takes the value 1 if model year is 1982 and takes the value 0 if it is not. In this model, 1970 is the reference year.

```
fit = fitlm(cars,'MPG~Weight*Model_Year')
```

```
fit =


Linear regression model:
    MPG ~ 1 + Weight*Model_Year

Estimated Coefficients:
                            Estimate          SE
                           ___________    __________

    (Intercept)                37.399        2.1466
    Weight                  -0.0058437    0.00061765
    Model_Year_76              4.6903        2.8538
    Model_Year_82              21.051         4.157
    Weight:Model_Year_76    -0.00082009    0.00085468
    Weight:Model_Year_82     -0.0050551     0.0015636


                            tStat        pValue
                           ________    __________

    (Intercept)              17.423    2.8607e-30
    Weight                  -9.4612    4.6077e-15
    Model_Year_76            1.6435       0.10384
    Model_Year_82            5.0641    2.2364e-06
    Weight:Model_Year_76    -0.95953       0.33992
    Weight:Model_Year_82     -3.2329     0.0017256


Number of observations: 94, Error degrees of freedom: 88
Root Mean Squared Error: 2.79
R-squared: 0.886,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 137, p-value = 5.79e-40
```

The regression output shows:

- `fitlm` recognizes `Model_Year` as a categorical variable, and constructs the required indicator (dummy) variables. By default, the first level, `70`, is the reference group (use `reordercats` to change the reference group).
- The model specification, `MPG~Weight*Model_Year`, specifies the first-order terms for `Weight` and `Model_Year`, and all interactions.
- The model $R^2 = 0.886$, meaning the variation in miles per gallon is reduced by 88.6% when you consider weight, model year, and their interactions.
- The fitted model is

  $$\widehat{MPG} = 37.4 - 0.006Weight + 4.7I[1976] + 21.1I[1982] - 0.0008Weight \times I[1976] - 0.005Weight \times I[1982].$$

  Thus, the estimated regression equations for the model years are as follows.

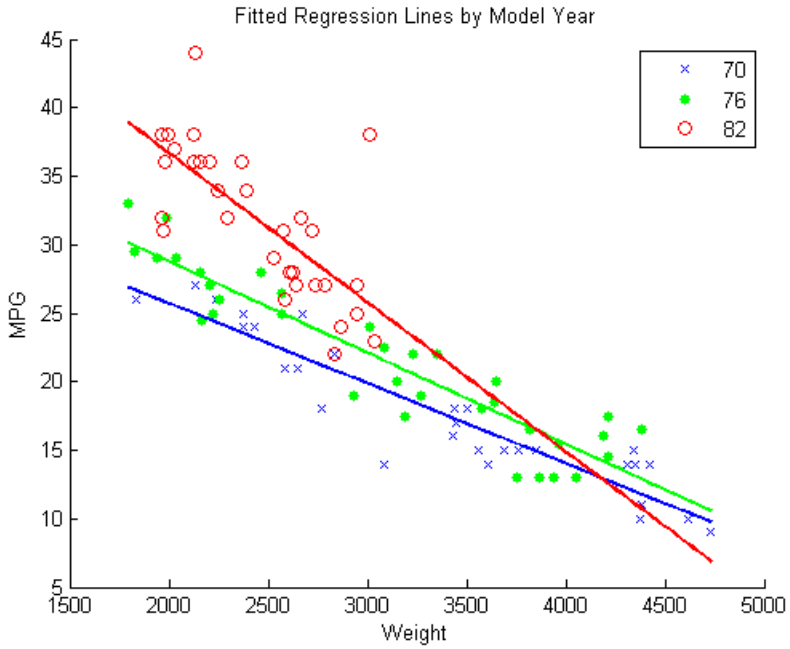
Model Year	Predicted MPG Against Weight
1970	$\widehat{MPG} = 37.4 - 0.006Weight$
1976	$\widehat{MPG} = (37.4 + 4.7) - (0.006 + 0.0008)Weight$
1982	$\widehat{MPG} = (37.4 + 21.1) - (0.006 + 0.005)Weight$

The relationship between `MPG` and `Weight` has an increasingly negative slope as the model year increases.

Plot fitted regression lines.

Plot the data and fitted regression lines.

```
w = linspace(min(Weight),max(Weight));

figure()
gscatter(Weight,MPG,Model_Year,'bgr','x.o')
line(w,feval(fit,w,'70'),'Color','b','LineWidth',2)
line(w,feval(fit,w,'76'),'Color','g','LineWidth',2)
line(w,feval(fit,w,'82'),'Color','r','LineWidth',2)
title('Fitted Regression Lines by Model Year')
```



Test for different slopes.

Test for significant differences between the slopes. This is equivalent to testing the hypothesis

$H_0: \beta_4 = \beta_5 = 0$

$H_A: \beta_i \neq 0$ for at least one i.

```
anova(fit)
```

```
ans =

                       SumSq     DF    MeanSq    F         pValue
    Weight             2050.2     1    2050.2    263.87    3.2055e-28
    Model_Year          807.69    2    403.84    51.976    1.2494e-15
    Weight:Model_Year   81.219    2    40.609    5.2266     0.0071637
    Error               683.74   88    7.7698
```

This output shows that the *p*-value for the test is `0.0072` (from the interaction row, `Weight:Model_Year`), so the null hypothesis is rejected at the 0.05 significance level. The value of the test statistic is `5.2266`. The numerator degrees of freedom for the test is `2`, which is the number of coefficients in the null hypothesis.

There is sufficient evidence that the slopes are not equal for all three model years.

See Also

`anova` | `categorical` | `fitlm` | `reordercats`

Related Examples

- “Test Differences Between Category Means” on page 2-25

- “Linear Regression” on page 11-9
- “Linear Regression Workflow” on page 11-35
- “Interpret Linear Regression Results” on page 11-50

More About

- “Grouping Variables” on page 2-45
- “Dummy Variables” on page 2-48

Create a Dataset Array from Workspace Variables

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB `table` data type instead. See MATLAB `table` documentation for more information.

In this section...

Create a Dataset Array from a Numeric Array

This example shows how to create a dataset array from a numeric array existing in the MATLAB® workspace.

Load sample data.

```
load fisheriris
```

Two variables load into the workspace: `meas`, a 150-by-4 numeric array, and `species`, a 150-by-1 cell array of species labels.

Create a dataset array.

Use `mat2dataset` to convert the numeric array, `meas`, into a dataset array.

```
ds = mat2dataset(meas);
ds(1:10,:)
```

```
ans =
    meas1    meas2    meas3    meas4
    5.1      3.5      1.4      0.2
    4.9        3      1.4      0.2
    4.7      3.2      1.3      0.2
    4.6      3.1      1.5      0.2
      5      3.6      1.4      0.2
    5.4      3.9      1.7      0.4
    4.6      3.4      1.4      0.3
      5      3.4      1.5      0.2
    4.4      2.9      1.4      0.2
    4.9      3.1      1.5      0.1
```

The array, `meas`, has four columns, so the dataset array, `ds`, has four variables. The default variable names are the array name, `meas`, with column numbers appended.

You can specify your own variable or observation names using the name-value pair arguments `VarNames` and `ObsNames`, respectively.

If you use `dataset` to convert a numeric array to a dataset array, by default, the resulting dataset array has one variable that is an array instead of separate variables for each column.

Examine the dataset array.

Return the size of the dataset array, `ds`.

```
size(ds)
```

```
ans = 1×2

   150     4
```

The dataset array, `ds`, is the same size as the numeric array, `meas`. Variable names and observation names do not factor into the size of a dataset array.

Explore dataset array metadata.

Return the metadata properties of the dataset array, `ds`.

```
ds.Properties
```

```
ans = struct with fields:
       Description: ''
    VarDescription: {}
             Units: {}
          DimNames: {'Observations'  'Variables'}
          UserData: []
          ObsNames: {}
          VarNames: {'meas1'  'meas2'  'meas3'  'meas4'}
```

You can also access the properties individually. For example, you can retrieve the variable names using `ds.Properties.VarNames`.

Access data in a dataset array variable.

You can use variable names with dot indexing to access the data in a dataset array. For example, find the minimum value in the first variable, `meas1`.

```
min(ds.meas1)
```

```
ans = 4.3000
```

Change variable names.

The four variables in `ds` are actually measurements of sepal length, sepal width, petal length, and petal width. Modify the variable names to be more descriptive.

```
ds.Properties.VarNames = {'SLength','SWidth','PLength','PWidth'};
```

Add description.

you can add a description for the dataset array.

```
ds.Properties.Description = 'Fisher iris data';
ds.Properties
```

```
ans = struct with fields:
       Description: 'Fisher iris data'
    VarDescription: {}
```

```
           Units: {}
        DimNames: {'Observations'  'Variables'}
        UserData: []
        ObsNames: {}
        VarNames: {'SLength'  'SWidth'  'PLength'  'PWidth'}
```

The dataset array properties are updated with the new variable names and description.

Add a variable to the dataset array.

The variable `species` is a cell array of species labels. Add `species` to the dataset array, `ds`, as a nominal array named `Species`. Display the first five observations in the dataset array.

```
ds.Species = nominal(species);
ds(1:5,:)
```

```
ans =
    SLength    SWidth    PLength    PWidth    Species
    5.1        3.5       1.4        0.2       setosa
    4.9          3       1.4        0.2       setosa
    4.7        3.2       1.3        0.2       setosa
    4.6        3.1       1.5        0.2       setosa
      5        3.6       1.4        0.2       setosa
```

The dataset array, `ds`, now has the fifth variable, `Species`.

Create Dataset Array from Heterogeneous Workspace Variables

This example shows how to create a dataset array from heterogeneous variables existing in the MATLAB® workspace.

Load sample data.

```
load carsmall
```

Create a dataset array.

Create a dataset array from a subset of the workspace variables.

```
ds = dataset(Origin,Acceleration,Cylinders,MPG);
ds.Properties.VarNames(:)
```

```
ans = 4x1 cell
    {'Origin'      }
    {'Acceleration'}
    {'Cylinders'   }
    {'MPG'         }
```

When creating the dataset array, you do not need to enter variable names. `dataset` automatically uses the name of each workspace variable.

Notice that the dataset array, `ds`, contains a collection of variables with heterogeneous data types. `Origin` is a character array, and the other variables are numeric.

Examine a dataset array.

Display the first five observations in the dataset array.

```
ds(1:5,:)
```

```
ans =
    Origin     Acceleration    Cylinders    MPG
    USA          12            8            18
    USA        11.5            8            15
    USA          11            8            18
    USA          12            8            16
    USA        10.5            8            17
```

Apply a function to a dataset array.

Use `datasetfun` to return the data type of each variable in `ds`.

```
varclass = datasetfun(@class,ds,'UniformOutput',false);
varclass(:)
```

```
ans = 4x1 cell
    {'char'  }
    {'double'}
    {'double'}
    {'double'}
```

You can get additional information about the variables using `summary(ds)`.

Modify a dataset array.

`Cylinders` is a numeric variable that has values 4, 6, and 8 for the number of cylinders. Convert `Cylinders` to a nominal array with levels `four`, `six`, and `eight`.

Display the country of origin and number of cylinders for the first 15 cars.

```
ds.Cylinders = nominal(ds.Cylinders,{'four','six','eight'});
ds(1:15,{'Origin','Cylinders'})
```

```
ans =
    Origin     Cylinders
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    USA        eight
    France     four
    USA        eight
    USA        eight
    USA        eight
    USA        eight
```

The variable `Cylinders` has a new data type.

See Also

`dataset` | `datasetfun` | `mat2dataset` | `nominal`

Related Examples

- “Create a Dataset Array from a File” on page 2-62
- “Export Dataset Arrays” on page 2-95
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Create a Dataset Array from a File

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB `table` data type instead. See MATLAB `table` documentation for more information.

In this section...

Create a Dataset Array from a Tab-Delimited Text File

This example shows how to create a dataset array from the contents of a tab-delimited text file.

Create a dataset array using default settings.

Import the text file `hospitalSmall.txt` as a dataset array using the default settings.

```
ds = dataset('File',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.txt'))
```

```
ds =

    name              sex        age    wgt    smoke
    'SMITH'           'm'        38     176    1
    'JOHNSON'         'm'        43     163    0
    'WILLIAMS'        'f'        38     131    0
    'JONES'           'f'        40     133    0
    'BROWN'           'f'        49     119    0
    'DAVIS'           'f'        46     142    0
    'MILLER'          'f'        33     142    1
    'WILSON'          'm'        40     180    0
    'MOORE'           'm'        28     183    0
    'TAYLOR'          'f'        31     132    0
    'ANDERSON'        'f'        45     128    0
    'THOMAS'          'f'        42     137    0
    'JACKSON'         'm'        25     174    0
    'WHITE'           'm'        39     202    1
```

By default, `dataset` uses the first row of the text file for variable names. If the first row does not contain variable names, you can specify the optional name-value pair argument `'ReadVarNames',false` to change the default behavior.

The dataset array contains heterogeneous variables. The variables `id`, `name`, and `sex` are cell arrays of character vectors, and the other variables are numeric.

Summarize the dataset array.

You can see the data type and other descriptive statistics for each variable by using `summary` to summarize the dataset array.

```
summary(ds)
```

```
name: [14x1 cell array of character vectors]

sex: [14x1 cell array of character vectors]

age: [14x1 double]

    min    1st quartile    median    3rd quartile    max
    25     33              39.5      43              49

wgt: [14x1 double]

    min    1st quartile    median    3rd quartile    max
    119    132             142       176             202

smoke: [14x1 double]

    min    1st quartile    median    3rd quartile    max
    0      0               0         0               1
```

Import observation names.

Import the text file again, this time specifying that the first column contains observation names.

```
ds = dataset('File',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.txt'),'ReadO
```

```
ds =

                sex        age    wgt    smoke
    SMITH       'm'        38     176    1
    JOHNSON     'm'        43     163    0
    WILLIAMS    'f'        38     131    0
    JONES       'f'        40     133    0
    BROWN       'f'        49     119    0
    DAVIS       'f'        46     142    0
    MILLER      'f'        33     142    1
    WILSON      'm'        40     180    0
    MOORE       'm'        28     183    0
    TAYLOR      'f'        31     132    0
    ANDERSON    'f'        45     128    0
    THOMAS      'f'        42     137    0
    JACKSON     'm'        25     174    0
    WHITE       'm'        39     202    1
```

The elements of the first column in the text file, last names, are now observation names. Observation names and row names are dataset array properties. You can always add or change the observation names of an existing dataset array by modifying the property `ObsNames`.

Change dataset array properties.

By default, the `DimNames` property of the dataset array has `name` as the descriptor of the observation (row) dimension. `dataset` got this name from the first row of the first column in the text file.

Change the first element of `DimNames` to `LastName`.

```
ds.Properties.DimNames{1} = 'LastName';
ds.Properties
```

```
ans =
```

```
      Description: ''
    VarDescription: {}
             Units: {}
          DimNames: {'LastName'  'Variables'}
          UserData: []
          ObsNames: {14x1 cell}
          VarNames: {'sex'  'age'  'wgt'  'smoke'}
```

Index into dataset array.

You can use observation names to index into a dataset array. For example, return the data for the patient with last name `BROWN`.

```
ds('BROWN',:)
```

```
ans =

              sex         age     wgt     smoke
    BROWN     'f'         49      119     0
```

Note that observation names must be unique.

Create a Dataset Array from a Comma-Separated Text File

This example shows how to create a dataset array from the contents of a comma-separated text file.

Create a dataset array.

Import the file `hospitalSmall.csv` as a dataset array, specifying the comma-delimited format.

```
ds = dataset('File',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.csv'),'Delim
```

```
ds =

    id                name              sex         age     wgt     smoke
    'YPL-320'         'SMITH'           'm'         38      176     1
    'GLI-532'         'JOHNSON'         'm'         43      163     0
    'PNI-258'         'WILLIAMS'        'f'         38      131     0
    'MIJ-579'         'JONES'           'f'         40      133     0
    'XLK-030'         'BROWN'           'f'         49      119     0
    'TFP-518'         'DAVIS'           'f'         46      142     0
    'LPD-746'         'MILLER'          'f'         33      142     1
    'ATA-945'         'WILSON'          'm'         40      180     0
    'VNL-702'         'MOORE'           'm'         28      183     0
    'LQW-768'         'TAYLOR'          'f'         31      132     0
    'QFY-472'         'ANDERSON'        'f'         45      128     0
    'UJG-627'         'THOMAS'          'f'         42      137     0
    'XUE-826'         'JACKSON'         'm'         25      174     0
    'TRW-072'         'WHITE'           'm'         39      202     1
```

By default, `dataset` uses the first row in the text file as variable names.

Add observation names.

Use the unique identifiers in the variable `id` as observation names. Then, delete the variable `id` from the dataset array.

```
ds.Properties.ObsNames = ds.id;
ds.id = []

ds =

              name              sex        age    wgt    smoke
    YPL-320    'SMITH'           'm'        38     176    1
    GLI-532    'JOHNSON'         'm'        43     163    0
    PNI-258    'WILLIAMS'        'f'        38     131    0
    MIJ-579    'JONES'           'f'        40     133    0
    XLK-030    'BROWN'           'f'        49     119    0
    TFP-518    'DAVIS'           'f'        46     142    0
    LPD-746    'MILLER'          'f'        33     142    1
    ATA-945    'WILSON'          'm'        40     180    0
    VNL-702    'MOORE'           'm'        28     183    0
    LQW-768    'TAYLOR'          'f'        31     132    0
    QFY-472    'ANDERSON'        'f'        45     128    0
    UJG-627    'THOMAS'          'f'        42     137    0
    XUE-826    'JACKSON'         'm'        25     174    0
    TRW-072    'WHITE'           'm'        39     202    1
```

Delete observations.

Delete any patients with the last name `BROWN`. You can use `strcmp` to match `'BROWN'` with the elements of the variable containing last names, `name`.

```
toDelete = strcmp(ds.name,'BROWN');
ds(toDelete,:) = []

ds =

              name              sex        age    wgt    smoke
    YPL-320    'SMITH'           'm'        38     176    1
    GLI-532    'JOHNSON'         'm'        43     163    0
    PNI-258    'WILLIAMS'        'f'        38     131    0
    MIJ-579    'JONES'           'f'        40     133    0
    TFP-518    'DAVIS'           'f'        46     142    0
    LPD-746    'MILLER'          'f'        33     142    1
    ATA-945    'WILSON'          'm'        40     180    0
    VNL-702    'MOORE'           'm'        28     183    0
    LQW-768    'TAYLOR'          'f'        31     132    0
    QFY-472    'ANDERSON'        'f'        45     128    0
    UJG-627    'THOMAS'          'f'        42     137    0
    XUE-826    'JACKSON'         'm'        25     174    0
    TRW-072    'WHITE'           'm'        39     202    1
```

One patient having last name `BROWN` is deleted from the dataset array.

Return size of dataset array.

The array now has 13 observations.

```
size(ds)

ans =

    13     5
```

Note that the row and column corresponding to variable and observation names, respectively, are not included in the size of a `dataset` array.

Create a Dataset Array from an Excel File

This example shows how to create a dataset array from the contents of an Excel® spreadsheet file.

Create a dataset array.

Import the data from the first worksheet in the file `hospitalSmall.xlsx`, specifying that the data file is an Excel spreadsheet.

```
ds = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'))
```

```
ds =

    id              name              sex        age    wgt    smoke
    'YPL-320'       'SMITH'           'm'        38     176    1
    'GLI-532'       'JOHNSON'         'm'        43     163    0
    'PNI-258'       'WILLIAMS'        'f'        38     131    0
    'MIJ-579'       'JONES'           'f'        40     133    0
    'XLK-030'       'BROWN'           'f'        49     119    0
    'TFP-518'       'DAVIS'           'f'        46     142    0
    'LPD-746'       'MILLER'          'f'        33     142    1
    'ATA-945'       'WILSON'          'm'        40     180    0
    'VNL-702'       'MOORE'           'm'        28     183    0
    'LQW-768'       'TAYLOR'          'f'        31     132    0
    'QFY-472'       'ANDERSON'        'f'        45     128    0
    'UJG-627'       'THOMAS'          'f'        42     137    0
    'XUE-826'       'JACKSON'         'm'        25     174    0
    'TRW-072'       'WHITE'           'm'        39     202    1
```

By default, `dataset` creates variable names using the contents of the first row in the spreadsheet.

Specify which worksheet to import.

Import the data from the second worksheet into a new dataset array.

```
ds2 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),
```

```
ds2 =

    id              name              sex        age    wgt    smoke
    'TRW-072'       'WHITE'           'm'        39     202    1
    'ELG-976'       'HARRIS'          'f'        36     129    0
    'KOQ-996'       'MARTIN'          'm'        48     181    1
    'YUZ-646'       'THOMPSON'        'm'        32     191    1
    'XBR-291'       'GARCIA'          'f'        27     131    1
    'KPW-846'       'MARTINEZ'        'm'        37     179    0
    'XBA-581'       'ROBINSON'        'm'        50     172    0
    'BKD-785'       'CLARK'           'f'        48     133    0
```

See Also

`dataset` | `summary`

Related Examples

- “Create a Dataset Array from Workspace Variables” on page 2-57
- “Clean Messy and Missing Data” on page 2-97
- “Export Dataset Arrays” on page 2-95
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Add and Delete Observations

This example shows how to add and delete observations in a dataset array. You can also edit dataset arrays using the Variables editor.

Load sample data.

Import the data from the first worksheet in `hospitalSmall.xlsx` into a dataset array.

```
ds = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'));
size(ds)
```

```
ans =

    14     6
```

The dataset array, `ds`, has 14 observations (rows) and 6 variables (columns).

Add observations by concatenation.

The second worksheet in `hospitalSmall.xlsx` has additional patient data. Append the observations in this spreadsheet to the end of `ds`.

```
ds2 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),
dsNew = [ds;ds2];
size(dsNew)
```

```
ans =

    22     6
```

The dataset array `dsNew` has 22 observations. In order to vertically concatenate two dataset arrays, both arrays must have the same number of variables, with the same variable names.

Add observations from a cell array.

If you want to append new observations stored in a cell array, first convert the cell array to a dataset array, and then concatenate the dataset arrays.

```
cellObs = {'id','name','sex','age','wgt','smoke';
               'YQR-965','BAKER','M',36,160,0;
               'LFG-497','WALL' ,'F',28,125,1;
               'KSD-003','REED' ,'M',32,187,0};
dsNew = [dsNew;cell2dataset(cellObs)];
size(dsNew)
```

```
ans =

    25     6
```

Add observations from a structure.

You can also append new observations stored in a structure. Convert the structure to a dataset array, and then concatenate the dataset arrays.

```
structObs(1,1).id = 'GHK-842';
structObs(1,1).name = 'GEORGE';
structObs(1,1).sex = 'M';
```

```
structObs(1,1).age = 45;
structObs(1,1).wgt = 182;
structObs(1,1).smoke = 1;

structObs(2,1).id = 'QRH-308';
structObs(2,1).name = 'BAILEY';
structObs(2,1).sex = 'F';
structObs(2,1).age = 29;
structObs(2,1).wgt = 120;
structObs(2,1).smoke = 0;

dsNew = [dsNew;struct2dataset(structObs)];
size(dsNew)

ans =

    27     6
```

Delete duplicate observations.

Use `unique` to delete any observations in a dataset array that are duplicated.

```
dsNew = unique(dsNew);
size(dsNew)

ans =

    21     6
```

One duplicated observation is deleted.

Delete observations by observation number.

Delete observations 18, 20, and 21 from the dataset array.

```
dsNew([18,20,21],:) = [];
size(dsNew)

ans =

    18     6
```

The dataset array has only 18 observations now.

Delete observations by observation name.

First, specify the variable of identifiers, `id`, as observation names. Then, delete the variable `id` from `dsNew`. You can use the observation name to index observations.

```
dsNew.Properties.ObsNames = dsNew.id;
dsNew.id = [];
dsNew('KOQ-996',:) = [];
size(dsNew)

ans =

    17     5
```

The dataset array now has one less observation and one less variable.

Search for observations to delete.

You can also search for observations in the dataset array. For example, delete observations for any patients with the last name `WILLIAMS`.

```
toDelete = strcmp(dsNew.name,'WILLIAMS');
dsNew(toDelete,:) = [];
size(dsNew)
```

```
ans =

    16     5
```

The dataset array now has one less observation.

See Also

`cell2dataset` | `dataset` | `struct2dataset`

Related Examples

- “Add and Delete Variables” on page 2-71
- “Select Subsets of Observations” on page 2-79
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Add and Delete Variables

This example shows how to add and delete variables in a dataset array. You can also edit dataset arrays using the Variables editor.

Load sample data.

Import the data from the first worksheet in `hospitalSmall.xlsx` into a dataset array.

```
ds = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'));
size(ds)
```

```
ans =

    14     6
```

The dataset array, `ds`, has 14 observations (rows) and 6 variables (columns).

Add variables by concatenating dataset arrays.

The worksheet `Heights` in `hospitalSmall.xlsx` has heights for the patients on the first worksheet. Concatenate the data in this spreadsheet with `ds`.

```
ds2 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),
ds = [ds ds2];
size(ds)
```

```
ans =

    14     7
```

The dataset array now has seven variables. You can only horizontally concatenate dataset arrays with observations in the same position, or with the same observation names.

```
ds.Properties.VarNames{end}
```

```
ans =

hgt
```

The name of the last variable in `ds` is `hgt`, which `dataset` read from the first row of the imported spreadsheet.

Delete variables by variable name.

First, specify the unique identifiers in the variable `id` as observation names. Then, delete the variable `id` from the dataset array.

```
ds.Properties.ObsNames = ds.id;
ds.id = [];
size(ds)
```

```
ans =

    14     6
```

The dataset array now has six variables. List the variable names.

```
ds.Properties.VarNames(:)

ans =

    'name'
    'sex'
    'age'
    'wgt'
    'smoke'
    'hgt'
```

There is no longer a variable called `id`.

Add a new variable by name.

Add a new variable, `bmi`—which contains the body mass index (BMI) for each patient—to the dataset array. BMI is a function of height and weight. Display the last name, gender, and BMI for each patient.

```
ds.bmi = ds.wgt*703./ds.hgt.^2;
ds(:,{'name','sex','bmi'})

ans =

                name              sex        bmi
    YPL-320     'SMITH'           'm'        24.544
    GLI-532     'JOHNSON'         'm'        24.068
    PNI-258     'WILLIAMS'        'f'        23.958
    MIJ-579     'JONES'           'f'        25.127
    XLK-030     'BROWN'           'f'        21.078
    TFP-518     'DAVIS'           'f'        27.729
    LPD-746     'MILLER'          'f'        26.828
    ATA-945     'WILSON'          'm'         24.41
    VNL-702     'MOORE'           'm'        27.822
    LQW-768     'TAYLOR'          'f'        22.655
    QFY-472     'ANDERSON'        'f'        23.409
    UJG-627     'THOMAS'          'f'        25.883
    XUE-826     'JACKSON'         'm'        24.265
    TRW-072     'WHITE'           'm'        29.827
```

The operators `./` and `.^` in the calculation of BMI indicate element-wise division and exponentiation, respectively.

Delete variables by variable number.

Delete the variable `wgt`, the fourth variable in the dataset array.

```
ds(:,4) = [];
ds.Properties.VarNames(:)

ans =

    'name'
    'sex'
    'age'
    'smoke'
    'hgt'
    'bmi'
```

The variable `wgt` is deleted from the dataset array.

See Also

`dataset`

Related Examples

- “Add and Delete Observations” on page 2-68
- “Merge Dataset Arrays” on page 2-85
- “Calculations on Dataset Arrays” on page 2-92
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Access Data in Dataset Array Variables

This example shows how to work with dataset array variables and their data.

Access variables by name.

You can access variable data, or select a subset of variables, by using variable (column) names and dot indexing. Load a sample dataset array. Display the names of the variables in `hospital`.

```
load hospital
hospital.Properties.VarNames(:)
```

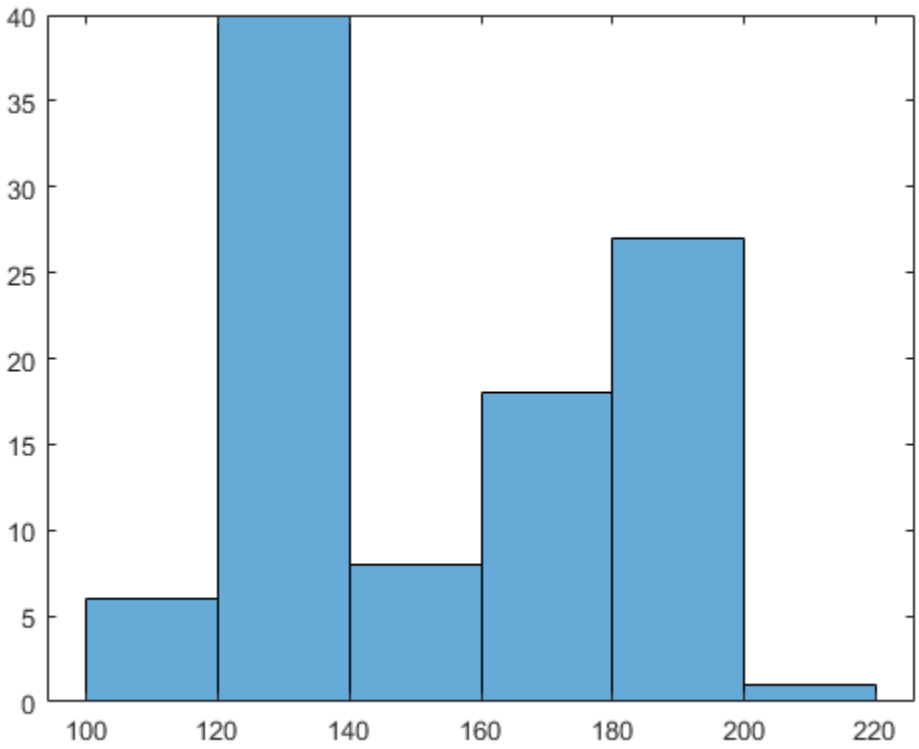
```
ans = 7x1 cell
    {'LastName'     }
    {'Sex'          }
    {'Age'          }
    {'Weight'       }
    {'Smoker'       }
    {'BloodPressure'}
    {'Trials'       }
```

The dataset array has 7 variables (columns) and 100 observations (rows). You can double-click `hospital` in the Workspace window to view the dataset array in the Variables editor.

Plot histogram.

Plot a histogram of the data in the variable `Weight`.

```
figure
histogram(hospital.Weight)
```

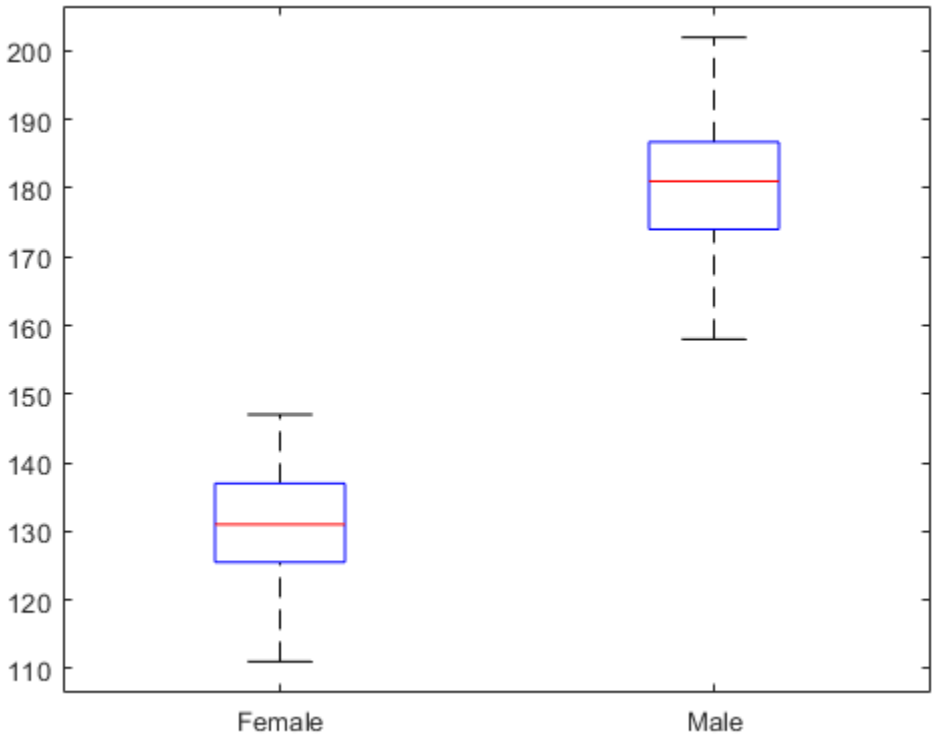


The histogram shows that the weight distribution is bimodal.

Plot data grouped by category.

Draw box plots of `Weight` grouped by the values in `Sex` (Male and Female). That is, use the variable `Sex` as a grouping variable.

```
figure
boxplot(hospital.Weight,hospital.Sex)
```



The box plot suggests that gender accounts for the bimodality in weight.

Select a subset of variables.

Create a new dataset array with only the variables `LastName`, `Sex`, and `Weight`. You can access the variables by name or column number.

```
ds1 = hospital(:,{'LastName','Sex','Weight'});
ds2 = hospital(:,[1,2,4]);
```

The dataset arrays `ds1` and `ds2` are equivalent. Use parentheses `( )` when indexing dataset arrays to preserve the data type; that is, to create a dataset array from a subset of a dataset array. You can also use the Variables editor to create a new dataset array from a subset of variables and observations.

Convert the variable data type.

Convert the data type of the variable `Smoker` from logical to nominal with labels `No` and `Yes`.

```
hospital.Smoker = nominal(hospital.Smoker,{'No','Yes'});
class(hospital.Smoker)
```

```
ans =
'nominal'
```

Explore data.

Display the first 10 elements of `Smoker`.

```
hospital.Smoker(1:10)
```

```
ans = 10x1 nominal
     Yes
     No
     No
     No
     No
     No
     Yes
     No
     No
     No
```

If you want to change the level labels in a nominal array, use `setlabels`.

Add variables.

The variable `BloodPressure` is a 100-by-2 array. The first column corresponds to systolic blood pressure, and the second column to diastolic blood pressure. Separate this array into two new variables, `SysPressure` and `DiaPressure`.

```
hospital.SysPressure = hospital.BloodPressure(:,1);
hospital.DiaPressure = hospital.BloodPressure(:,2);
hospital.Properties.VarNames(:)
```

```
ans = 9x1 cell
    {'LastName'     }
    {'Sex'          }
    {'Age'          }
    {'Weight'       }
    {'Smoker'       }
    {'BloodPressure'}
    {'Trials'       }
    {'SysPressure'  }
    {'DiaPressure'  }
```

The dataset array, `hospital`, has two new variables.

Search for variables by name.

Use `regexp` to find variables in `hospital` with `'Pressure'` in their name. Create a new dataset array containing only these variables.

```
bp = regexp(hospital.Properties.VarNames,'Pressure');
bpIdx = cellfun(@isempty,bp);
bpData = hospital(:,~bpIdx);
bpData.Properties.VarNames(:)
```

```
ans = 3x1 cell
    {'BloodPressure'}
    {'SysPressure'  }
    {'DiaPressure'  }
```

The new dataset array, `bpData`, contains only the blood pressure variables.

Delete variables.

Delete the variable `BloodPressure` from the dataset array, `hospital`.

```
hospital.BloodPressure = [];
hospital.Properties.VarNames(:)
```

```
ans = 8x1 cell
    {'LastName'   }
    {'Sex'        }
    {'Age'        }
    {'Weight'     }
    {'Smoker'     }
    {'Trials'     }
    {'SysPressure'}
    {'DiaPressure'}
```

The variable `BloodPressure` is no longer in the dataset array.

See Also

`dataset`

Related Examples

- “Add and Delete Variables” on page 2-71
- “Calculations on Dataset Arrays” on page 2-92
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112
- “Grouping Variables” on page 2-45

Select Subsets of Observations

This example shows how to select an observation or subset of observations from a dataset array.

Load sample data.

Load the sample dataset array, `hospital`. Dataset arrays can have observation (row) names. This array has observation names corresponding to unique patient identifiers.

```
load hospital
hospital.Properties.ObsNames(1:10)
```

```
ans = 10x1 cell
    {'YPL-320'}
    {'GLI-532'}
    {'PNI-258'}
    {'MIJ-579'}
    {'XLK-030'}
    {'TFP-518'}
    {'LPD-746'}
    {'ATA-945'}
    {'VNL-702'}
    {'LQW-768'}
```

These are the first 10 observation names.

Index an observation by name.

You can use the observation names to index into the dataset array. For example, extract the last name, sex, and age for the patient with identifier `XLK-030`.

```
hospital('XLK-030',{'LastName','Sex','Age'})
```

```
ans =
                LastName         Sex       Age
    XLK-030     {'BROWN'}        Female    49
```

Index a subset of observations by number.

Create a new dataset array containing the first 50 patients.

```
ds50 = hospital(1:50,:);
size(ds50)
```

```
ans = 1×2

    50     7
```

Search observations using a logical condition.

Create a new dataset array containing only male patients. To find the male patients, use a logical condition to search the variable containing gender information.

```
dsMale = hospital(hospital.Sex=='Male',:);
dsMale(1:10,{'LastName','Sex'})
```

```
ans =
                LastName            Sex
    YPL-320    {'SMITH'   }        Male
    GLI-532    {'JOHNSON' }        Male
    ATA-945    {'WILSON'  }        Male
    VNL-702    {'MOORE'   }        Male
    XUE-826    {'JACKSON' }        Male
    TRW-072    {'WHITE'   }        Male
    KOQ-996    {'MARTIN'  }        Male
    YUZ-646    {'THOMPSON'}        Male
    KPW-846    {'MARTINEZ'}        Male
    XBA-581    {'ROBINSON'}        Male
```

Search observations using multiple conditions.

You can use multiple conditions to search the dataset array. For example, create a new dataset array containing only female patients older than 40.

```
dsFemale = hospital(hospital.Sex=='Female' & hospital.Age > 40,:);
dsFemale(1:10,{'LastName','Sex','Age'})
```

```
ans =
                LastName            Sex       Age
    XLK-030    {'BROWN'   }        Female    49
    TFP-518    {'DAVIS'   }        Female    46
    QFY-472    {'ANDERSON'}        Female    45
    UJG-627    {'THOMAS'  }        Female    42
    BKD-785    {'CLARK'   }        Female    48
    VWL-936    {'LEWIS'   }        Female    41
    AAX-056    {'LEE'     }        Female    44
    AFK-336    {'WRIGHT'  }        Female    45
    KKL-155    {'ADAMS'   }        Female    48
    RBA-579    {'SANCHEZ' }        Female    44
```

Select a random subset of observations.

Create a new dataset array containing a random subset of 20 patients from the dataset array `hospital`.

```
rng('default') % For reproducibility
dsRandom = hospital(randsample(length(hospital),20),:);
dsRandom.Properties.ObsNames
```

```
ans = 20x1 cell
    {'DAU-529'}
    {'AGR-528'}
    {'RBO-332'}
    {'QOO-305'}
    {'RVS-253'}
    {'QEQ-082'}
    {'EHE-616'}
    {'HVR-372'}
    {'KOQ-996'}
    {'REV-997'}
    {'PUE-347'}
    {'LQW-768'}
```

```
    {'YLN-495'}
    {'HJQ-495'}
    {'ELG-976'}
    {'XUE-826'}
    {'MEZ-469'}
    {'UDS-151'}
    {'MIJ-579'}
    {'DGC-290'}
```

Delete observations by name.

Delete the data for the patient with observation name `HVR-372`.

```
hospital('HVR-372',:) = [];
size(hospital)
```

```
ans = 1×2

    99     7
```

The dataset array has one less observation.

See Also

`dataset`

Related Examples

- “Add and Delete Observations” on page 2-68
- “Clean Messy and Missing Data” on page 2-97
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Sort Observations in Dataset Arrays” on page 2-82
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Sort Observations in Dataset Arrays

This example shows how to sort observations (rows) in a dataset array using the command line. You can also sort rows using the Variables editor.

Sort observations in ascending order.

Load the sample dataset array, `hospital`. Sort the observations by the values in `Age`, in ascending order.

```
load hospital
dsAgeUp = sortrows(hospital,'Age');
dsAgeUp(1:10,{'LastName','Age'})
```

```
ans =
                LastName              Age
    XUE-826    {'JACKSON'  }        25
    FZR-250    {'HALL'     }        25
    PUE-347    {'YOUNG'    }        25
    LIM-480    {'HILL'     }        25
    SCQ-914    {'JAMES'    }        25
    REV-997    {'ALEXANDER'}        25
    XBR-291    {'GARCIA'   }        27
    VNL-702    {'MOORE'    }        28
    DTT-578    {'WALKER'   }        28
    XAX-646    {'COOPER'   }        28
```

The youngest patients are age 25.

Sort observations in descending order.

Sort the observations by `Age` in descending order.

```
dsAgeDown = sortrows(hospital,'Age','descend');
dsAgeDown(1:10,{'LastName','Age'})
```

```
ans =
                LastName             Age
    XBA-581    {'ROBINSON'}        50
    DAU-529    {'REED'    }        50
    XLK-030    {'BROWN'   }        49
    FLJ-908    {'STEWART' }        49
    GGU-691    {'HUGHES'  }        49
    MEZ-469    {'GRIFFIN' }        49
    KOQ-996    {'MARTIN'  }        48
    BKD-785    {'CLARK'   }        48
    KKL-155    {'ADAMS'   }        48
    NSK-403    {'RAMIREZ' }        48
```

The oldest patients are age 50.

Sort observations by the values of two variables.

Sort the observations in `hospital` by `Age`, and then by `LastName`.

```
dsName = sortrows(hospital,{'Age','LastName'});
dsName(1:10,{'LastName','Age'})

ans =
                LastName             Age
    REV-997    {'ALEXANDER'}        25
    FZR-250    {'HALL'     }        25
    LIM-480    {'HILL'     }        25
    XUE-826    {'JACKSON'  }        25
    SCQ-914    {'JAMES'    }        25
    PUE-347    {'YOUNG'    }        25
    XBR-291    {'GARCIA'   }        27
    XAX-646    {'COOPER'   }        28
    QEQ-082    {'COX'      }        28
    NSU-424    {'JENKINS'  }        28
```

Now the names are sorted alphabetically within increasing age groups.

Sort observations in mixed order.

Sort the observations in `hospital` by `Age` in an increasing order, and then by `Weight` in a decreasing order.

```
dsWeight = sortrows(hospital,{'Age','Weight'},{'ascend','descend'});
dsWeight(1:10,{'LastName','Age','Weight'})

ans =
                LastName             Age    Weight
    FZR-250    {'HALL'     }        25     189
    SCQ-914    {'JAMES'    }        25     186
    XUE-826    {'JACKSON'  }        25     174
    REV-997    {'ALEXANDER'}        25     171
    LIM-480    {'HILL'     }        25     138
    PUE-347    {'YOUNG'    }        25     114
    XBR-291    {'GARCIA'   }        27     131
    NSU-424    {'JENKINS'  }        28     189
    VNL-702    {'MOORE'    }        28     183
    XAX-646    {'COOPER'   }        28     127
```

This shows that the maximum weight among patients that are age 25 is 189 lbs.

Sort observations by observation name.

Sort the observations in `hospital` by the observation names.

```
dsObs = sortrows(hospital,'obsnames');
dsObs(1:10,{'LastName','Age'})

ans =
                LastName           Age
    AAX-056    {'LEE'    }        44
    AFB-271    {'PEREZ'  }        44
    AFK-336    {'WRIGHT' }        45
    AGR-528    {'SIMMONS'}        45
    ATA-945    {'WILSON' }        40
    BEZ-311    {'DIAZ'   }        45
    BKD-785    {'CLARK'  }        48
```

```
DAU-529    {'REED'   }        50
DGC-290    {'BUTLER' }        38
DTT-578    {'WALKER' }        28
```

The observations are sorted by observation name in ascending alphabetical order.

See Also

`dataset` | `sortrows`

Related Examples

- “Select Subsets of Observations” on page 2-79
- “Stack or Unstack Dataset Arrays” on page 2-88
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Merge Dataset Arrays

This example shows how to merge dataset arrays using `join`.

Load sample data.

Import the data from the first worksheet in `hospitalSmall.xlsx` into a dataset array, then keep only a few of the variables.

```
ds1 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'))
ds1 = ds1(:,{'id','name','sex','age'})
```

```
ds1 =

    id              name             sex        age
    'YPL-320'       'SMITH'          'm'        38
    'GLI-532'       'JOHNSON'        'm'        43
    'PNI-258'       'WILLIAMS'       'f'        38
    'MIJ-579'       'JONES'          'f'        40
    'XLK-030'       'BROWN'          'f'        49
    'TFP-518'       'DAVIS'          'f'        46
    'LPD-746'       'MILLER'         'f'        33
    'ATA-945'       'WILSON'         'm'        40
    'VNL-702'       'MOORE'          'm'        28
    'LQW-768'       'TAYLOR'         'f'        31
    'QFY-472'       'ANDERSON'       'f'        45
    'UJG-627'       'THOMAS'         'f'        42
    'XUE-826'       'JACKSON'        'm'        25
    'TRW-072'       'WHITE'          'm'        39
```

The dataset array, `ds1`, has 14 observations (rows) and 4 variables (columns).

Import the data from the worksheet `Heights2` in `hospitalSmall.xlsx`.

```
ds2 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),
```

```
ds2 =

    id              hgt
    'LPD-746'       61
    'PNI-258'       62
    'XUE-826'       71
    'ATA-945'       72
    'XLK-030'       63
```

`ds2` has height measurements for a subset of five individuals from the first dataset array, `ds1`.

Merge only the matching subset of observations.

Use `join` to merge the two dataset arrays, `ds1` and `ds2`, keeping only the subset of observations that are in `ds2`.

```
JoinSmall = join(ds2,ds1)
```

```
JoinSmall =

    id              hgt    name             sex        age
    'LPD-746'       61     'MILLER'         'f'        33
```

```
    'PNI-258'        62     'WILLIAMS'        'f'        38
    'XUE-826'        71     'JACKSON'         'm'        25
    'ATA-945'        72     'WILSON'          'm'        40
    'XLK-030'        63     'BROWN'           'f'        49
```

In `JoinSmall`, the variable `id` only appears once. This is because it is the key variable—the variable that links observations between the two dataset arrays—and has the same variable name in both `ds1` and `ds2`.

Include incomplete observations in the merge.

Merge `ds1` and `ds2` keeping all observations in the larger `ds1`.

```
joinAll = join(ds2,ds1,'type','rightouter','mergekeys',true)
```

```
joinAll =

    id               hgt    name              sex        age
    'ATA-945'         72    'WILSON'          'm'        40
    'GLI-532'        NaN    'JOHNSON'         'm'        43
    'LPD-746'         61    'MILLER'          'f'        33
    'LQW-768'        NaN    'TAYLOR'          'f'        31
    'MIJ-579'        NaN    'JONES'           'f'        40
    'PNI-258'         62    'WILLIAMS'        'f'        38
    'QFY-472'        NaN    'ANDERSON'        'f'        45
    'TFP-518'        NaN    'DAVIS'           'f'        46
    'TRW-072'        NaN    'WHITE'           'm'        39
    'UJG-627'        NaN    'THOMAS'          'f'        42
    'VNL-702'        NaN    'MOORE'           'm'        28
    'XLK-030'         63    'BROWN'           'f'        49
    'XUE-826'         71    'JACKSON'         'm'        25
    'YPL-320'        NaN    'SMITH'           'm'        38
```

Each observation in `ds1` without corresponding height measurements in `ds2` has height value `NaN`. Also, because there is no `id` value in `ds2` for each observation in `ds1`, you need to merge the keys using the option `'MergeKeys',true`. This merges the key variable, `id`.

Merge dataset arrays with different key variable names.

When using `join`, it is not necessary for the key variable to have the same name in the dataset arrays to be merged. Import the data from the worksheet named `Heights3` in `hospitalSmall.xlsx`.

```
ds3 = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),
```

```
ds3 =

    identifier       hgt
    'GLI-532'        69
    'QFY-472'        62
    'MIJ-579'        61
    'VNL-702'        68
    'XLK-030'        63
    'LPD-746'        61
    'TFP-518'        60
    'YPL-320'        71
    'ATA-945'        72
    'LQW-768'        64
    'PNI-258'        62
```

```
    'UJG-627'        61
    'XUE-826'        71
    'TRW-072'        69
```

`ds3` has height measurements for each observation in `ds1`. This dataset array has the same patient identifiers as `ds1`, but they are under the variable name `identifier`, instead of `id` (and in a different order).

Specify key variable.

You can easily change the variable name of the key variable in `ds3` by setting `d3.Properties.VarNames` or using the Variables editor, but it is not required to perform a merge. Instead, you can specify the name of the key variable in each dataset array using `LeftKeys` and `RightKeys`.

```
joinDiff = join(ds3,ds1,'LeftKeys','identifier','RightKeys','id')
```

```
joinDiff =

    identifier       hgt    name              sex        age
    'GLI-532'        69     'JOHNSON'         'm'        43
    'QFY-472'        62     'ANDERSON'        'f'        45
    'MIJ-579'        61     'JONES'           'f'        40
    'VNL-702'        68     'MOORE'           'm'        28
    'XLK-030'        63     'BROWN'           'f'        49
    'LPD-746'        61     'MILLER'          'f'        33
    'TFP-518'        60     'DAVIS'           'f'        46
    'YPL-320'        71     'SMITH'           'm'        38
    'ATA-945'        72     'WILSON'          'm'        40
    'LQW-768'        64     'TAYLOR'          'f'        31
    'PNI-258'        62     'WILLIAMS'        'f'        38
    'UJG-627'        61     'THOMAS'          'f'        42
    'XUE-826'        71     'JACKSON'         'm'        25
    'TRW-072'        69     'WHITE'           'm'        39
```

The merged dataset array, `joinDiff`, has the same key variable order and name as the first dataset array input to `join`, `ds3`.

See Also

`dataset` | `join`

Related Examples

- “Add and Delete Variables” on page 2-71
- “Stack or Unstack Dataset Arrays” on page 2-88
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Stack or Unstack Dataset Arrays

This example shows how to reformat dataset arrays using `stack` and `unstack`.

Load sample data.

Import the data from the comma-separated text file `testScores.csv`.

```
ds = dataset('File','testScores.csv','Delimiter',',')
```

```
ds =

    LastName          Sex             Test1    Test2    Test3    Test4
    'HOWARD'          'male'          90       87       93       92
    'WARD'            'male'          87       85       83       90
    'TORRES'          'male'          86       85       88       86
    'PETERSON'        'female'        75       80       72       77
    'GRAY'            'female'        89       86       87       90
    'RAMIREZ'         'female'        96       92       98       95
    'JAMES'           'male'          78       75       77       77
    'WATSON'          'female'        91       94       92       90
    'BROOKS'          'female'        86       83       85       89
    'KELLY'           'male'          79       76       82       80
```

Each of the 10 students has 4 test scores.

Perform calculations on dataset array.

With the data in this format, you can, for example, calculate the average test score for each student. The test scores are in columns 3 to 6.

```
ds.TestAve = mean(double(ds(:,3:6)),2);
ds(:,{'LastName','Sex','TestAve'})
```

```
ans =

    LastName          Sex             TestAve
    'HOWARD'          'male'           90.5
    'WARD'            'male'          86.25
    'TORRES'          'male'          86.25
    'PETERSON'        'female'           76
    'GRAY'            'female'           88
    'RAMIREZ'         'female'        95.25
    'JAMES'           'male'          76.75
    'WATSON'          'female'        91.75
    'BROOKS'          'female'        85.75
    'KELLY'           'male'          79.25
```

A new variable with average test scores is added to the dataset array, `ds`.

Reformat the dataset array.

Stack the test score variables into a new variable, `Scores`.

```
 dsNew = stack(ds,{'Test1','Test2','Test3','Test4'},...
            'newDataVarName','Scores')
```

```
dsNew =
```

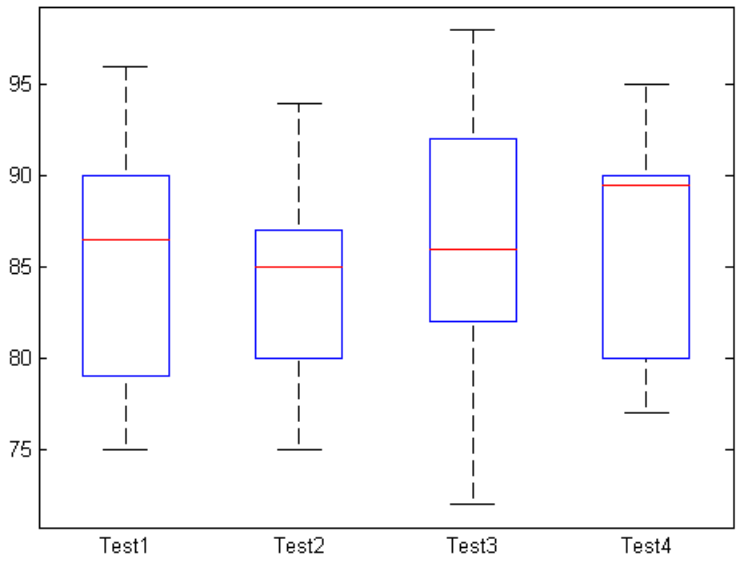
```
    LastName          Sex             TestAve    Scores_Indicator    Scores
    'HOWARD'          'male'           90.5      Test1               90
    'HOWARD'          'male'           90.5      Test2               87
    'HOWARD'          'male'           90.5      Test3               93
    'HOWARD'          'male'           90.5      Test4               92
    'WARD'            'male'          86.25      Test1               87
    'WARD'            'male'          86.25      Test2               85
    'WARD'            'male'          86.25      Test3               83
    'WARD'            'male'          86.25      Test4               90
    'TORRES'          'male'          86.25      Test1               86
    'TORRES'          'male'          86.25      Test2               85
    'TORRES'          'male'          86.25      Test3               88
    'TORRES'          'male'          86.25      Test4               86
    'PETERSON'        'female'           76      Test1               75
    'PETERSON'        'female'           76      Test2               80
    'PETERSON'        'female'           76      Test3               72
    'PETERSON'        'female'           76      Test4               77
    'GRAY'            'female'           88      Test1               89
    'GRAY'            'female'           88      Test2               86
    'GRAY'            'female'           88      Test3               87
    'GRAY'            'female'           88      Test4               90
    'RAMIREZ'         'female'        95.25      Test1               96
    'RAMIREZ'         'female'        95.25      Test2               92
    'RAMIREZ'         'female'        95.25      Test3               98
    'RAMIREZ'         'female'        95.25      Test4               95
    'JAMES'           'male'          76.75      Test1               78
    'JAMES'           'male'          76.75      Test2               75
    'JAMES'           'male'          76.75      Test3               77
    'JAMES'           'male'          76.75      Test4               77
    'WATSON'          'female'        91.75      Test1               91
    'WATSON'          'female'        91.75      Test2               94
    'WATSON'          'female'        91.75      Test3               92
    'WATSON'          'female'        91.75      Test4               90
    'BROOKS'          'female'        85.75      Test1               86
    'BROOKS'          'female'        85.75      Test2               83
    'BROOKS'          'female'        85.75      Test3               85
    'BROOKS'          'female'        85.75      Test4               89
    'KELLY'           'male'          79.25      Test1               79
    'KELLY'           'male'          79.25      Test2               76
    'KELLY'           'male'          79.25      Test3               82
    'KELLY'           'male'          79.25      Test4               80
```

The original test variable names, `Test1`, `Test2`, `Test3`, and `Test4`, appear as levels in the combined test scores indicator variable, `Scores_Indicator`.

Plot data grouped by category.

With the data in this format, you can use `Scores_Indicator` as a grouping variable, and draw box plots of test scores grouped by test.

```
figure()
boxplot(dsNew.Scores,dsNew.Scores_Indicator)
```



Revert the dataset array to the original format.

Reformat `dsNew` back into its original format.

```
dsOrig = unstack(dsNew,'Scores','Scores_Indicator');
dsOrig(:,{'LastName','Test1','Test2','Test3','Test4'})
```

```
ans =

    LastName          Test1    Test2    Test3    Test4
    'HOWARD'          90       87       93       92
    'WARD'            87       85       83       90
    'TORRES'          86       85       88       86
    'PETERSON'        75       80       72       77
    'GRAY'            89       86       87       90
    'RAMIREZ'         96       92       98       95
    'JAMES'           78       75       77       77
    'WATSON'          91       94       92       90
    'BROOKS'          86       83       85       89
    'KELLY'           79       76       82       80
```

The dataset array is back in wide format. `unstack` reassigns the levels of the indicator variable, `Scores_Indicator`, as variable names in the unstacked dataset array.

See Also

`dataset` | `double` | `stack` | `unstack`

Related Examples

- “Access Data in Dataset Array Variables” on page 2-74
- “Calculations on Dataset Arrays” on page 2-92

- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112
- “Grouping Variables” on page 2-45

Calculations on Dataset Arrays

This example shows how to perform calculations on dataset arrays.

Load sample data.

Import the data from the comma-separated text file `testScores.csv`.

```
ds = dataset('File','testScores.csv','Delimiter',',')
```

```
ds =

    LastName          Sex             Test1    Test2    Test3    Test4
    'HOWARD'          'male'          90       87       93       92
    'WARD'            'male'          87       85       83       90
    'TORRES'          'male'          86       85       88       86
    'PETERSON'        'female'        75       80       72       77
    'GRAY'            'female'        89       86       87       90
    'RAMIREZ'         'female'        96       92       98       95
    'JAMES'           'male'          78       75       77       77
    'WATSON'          'female'        91       94       92       90
    'BROOKS'          'female'        86       83       85       89
    'KELLY'           'male'          79       76       82       80
```

There are 4 test scores for each of 10 students, in wide format.

Average dataset array variables.

Compute the average (mean) test score for each student in the dataset array, and store it in a new variable, `TestAvg`. Test scores are in columns 3 to 6.

Use `double` to convert the specified dataset array variables into a numeric array. Then, calculate the mean across the second dimension (across columns) to get the test average for each student.

```
ds.TestAvg = mean(double(ds(:,3:6)),2);
ds(:,{'LastName','TestAvg'})
```

```
ans =

    LastName          TestAvg
    'HOWARD'           90.5
    'WARD'            86.25
    'TORRES'          86.25
    'PETERSON'           76
    'GRAY'               88
    'RAMIREZ'         95.25
    'JAMES'           76.75
    'WATSON'          91.75
    'BROOKS'          85.75
    'KELLY'           79.25
```

Summarize the dataset array using a grouping variable.

Compute the mean and maximum average test scores for each gender.

```
stats = grpstats(ds,'Sex',{'mean','max'},'DataVars','TestAvg')
```

```
stats =
```

```
            Sex             GroupCount    mean_TestAvg    max_TestAvg
    male    'male'          5              83.8            90.5
    female  'female'        5             87.35           95.25
```

This returns a new dataset array containing the specified summary statistics for each level of the grouping variable, `Sex`.

Replace data values.

The denominator for each test score is 100. Convert the test score denominator to 25.

```
scores = double(ds(:,3:6));
newScores = scores*25/100;
ds = replacedata(ds,newScores,3:6)
```

```
ds =

    LastName          Sex             Test1    Test2    Test3    Test4    TestAvg
    'HOWARD'          'male'           22.5    21.75    23.25       23     90.5
    'WARD'            'male'          21.75    21.25    20.75     22.5    86.25
    'TORRES'          'male'           21.5    21.25       22     21.5    86.25
    'PETERSON'        'female'        18.75       20       18    19.25       76
    'GRAY'            'female'        22.25     21.5    21.75     22.5       88
    'RAMIREZ'         'female'           24       23     24.5    23.75    95.25
    'JAMES'           'male'           19.5    18.75    19.25    19.25    76.75
    'WATSON'          'female'        22.75     23.5       23     22.5    91.75
    'BROOKS'          'female'         21.5    20.75    21.25    22.25    85.75
    'KELLY'           'male'          19.75       19     20.5       20    79.25
```

The first two lines of code extract the test data and perform the desired calculation. Then, `replacedata` inserts the new test scores back into the dataset array.

The variable of test score averages, `TestAvg`, is now the final score for each student.

Change variable name.

Change the variable name to `Final`.

```
ds.Properties.VarNames{end} = 'Final';
ds
```

```
ds =

    LastName          Sex             Test1    Test2    Test3    Test4    Final
    'HOWARD'          'male'           22.5    21.75    23.25       23     90.5
    'WARD'            'male'          21.75    21.25    20.75     22.5    86.25
    'TORRES'          'male'           21.5    21.25       22     21.5    86.25
    'PETERSON'        'female'        18.75       20       18    19.25       76
    'GRAY'            'female'        22.25     21.5    21.75     22.5       88
    'RAMIREZ'         'female'           24       23     24.5    23.75    95.25
    'JAMES'           'male'           19.5    18.75    19.25    19.25    76.75
    'WATSON'          'female'        22.75     23.5       23     22.5    91.75
    'BROOKS'          'female'         21.5    20.75    21.25    22.25    85.75
    'KELLY'           'male'          19.75       19     20.5       20    79.25
```

See Also

`dataset` | `double` | `grpstats` | `replacedata`

Related Examples

- “Stack or Unstack Dataset Arrays” on page 2-88
- “Access Data in Dataset Array Variables” on page 2-74
- “Select Subsets of Observations” on page 2-79
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Export Dataset Arrays

This example shows how to export a dataset array from the MATLAB workspace to a text or spreadsheet file.

Load sample data.

```
load('hospital')
```

The dataset array has 100 observations and 7 variables.

Export to a text file.

Export the dataset array, `hospital`, to a text file named `hospital.txt`. By default, `export` writes to a tab-delimited text file with the same name as the dataset array, appended by `.txt`.

```
export(hospital)
```

This creates the file `hospital.txt` in the current working folder, if it does not previously exist. If the file already exists in the current working folder, `export` overwrites the existing file.

By default, variable names are in the first line of the text file. Observation names, if present, are in the first column.

Export without variable names.

Export `hospital` with variable names suppressed to a text file named `NoLabels.txt`.

```
export(hospital,'File','NoLabels.txt','WriteVarNames',false)
```

There are no variable names in the first line of the created text file, `NoLabels.txt`.

Export to a comma-delimited format.

Export `hospital` to a comma-delimited text file, `hospital.csv`.

```
export(hospital,'File','hospital.csv','Delimiter',',')
```

Export to an Excel spreadsheet.

Export `hospital` to an Excel spreadsheet named `hospital.xlsx`.

```
export(hospital,'XLSFile','hospital.xlsx')
```

By default, the first row of `hospital.xlsx` has variable names, and the first column has observation names.

See Also

`dataset` | `export`

Related Examples

- “Create a Dataset Array from Workspace Variables” on page 2-57
- “Create a Dataset Array from a File” on page 2-62

More About

- “Dataset Arrays” on page 2-112

Clean Messy and Missing Data

This example shows how to find, clean, and delete observations with missing data in a dataset array.

Load sample data.

Import the data from the spreadsheet `messy.xlsx`.

```
messyData = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','messy.xlsx'))
```

```
messyData =

    var1          var2          var3         var4          var5
    'afe1'        '3'           'yes'        '3'             3
    'egh3'        '.'           'no'         '7'             7
    'wth4'        '3'           'yes'        '3'             3
    'atn2'        '23'          'no'         '23'           23
    'arg1'        '5'           'yes'        '5'             5
    'jre3'        '34.6'        'yes'        '34.6'       34.6
    'wen9'        '234'         'yes'        '234'         234
    'ple2'        '2'           'no'         '2'             2
    'dbo8'        '5'           'no'         '5'             5
    'oii4'        '5'           'yes'        '5'             5
    'wnk3'        '245'         'yes'        '245'         245
    'abk6'        '563'         ''           '563'         563
    'pnj5'        '463'         'no'         '463'         463
    'wnn3'        '6'           'no'         '6'             6
    'oks9'        '23'          'yes'        '23'           23
    'wba3'        ''            'yes'        'NaN'          14
    'pkn4'        '2'           'no'         '2'             2
    'adw3'        '22'          'no'         '22'           22
    'poj2'        '-99'         'yes'        '-99'         -99
    'bas8'        '23'          'no'         '23'           23
    'gry5'        'NA'          'yes'        'NaN'          21
```

When you import data from a spreadsheet, `dataset` reads any variables with nonnumeric elements as a cell array of character vectors. This is why the variable `var2` is a cell array of character vectors. When importing data from a text file, you have more flexibility to specify which nonnumeric expressions to treat as missing using the option `TreatAsEmpty`.

There are many different missing data indicators in `messy.xlsx`, such as:

- Empty cells
- A period (`.`)
- `NA`
- `NaN`
- `-99`

Find observations with missing values.

Display the subset of observations that have at least one missing value using `ismissing`.

```
ix = ismissing(messyData,'NumericTreatAsMissing',-99,...
                'StringTreatAsMissing',{'NaN','.','NA'});
messyData(any(ix,2),:)
```

```
ans =

    var1          var2         var3         var4         var5
    'egh3'        '.'          'no'         '7'            7
    'abk6'        '563'        ''           '563'        563
    'wba3'        ''           'yes'        'NaN'         14
    'poj2'        '-99'        'yes'        '-99'        -99
    'gry5'        'NA'         'yes'        'NaN'         21
```

By default, `ismissing` recognizes the following missing value indicators:

- `NaN` for numeric arrays
- `''` for character arrays
- `<undefined>` for categorical arrays

Use the `NumericTreatAsMissing` and `StringTreatAsMissing` options to specify other values to treat as missing.

Convert character arrays to double arrays.

You can convert the `char` variables that should be numeric using `str2double`.

```
messyData.var2 = str2double(messyData.var2);
messyData.var4 = str2double(messyData.var4)
```

```
messyData =

    var1          var2    var3         var4    var5
    'afe1'           3    'yes'           3       3
    'egh3'         NaN    'no'            7       7
    'wth4'           3    'yes'           3       3
    'atn2'          23    'no'           23      23
    'arg1'           5    'yes'           5       5
    'jre3'        34.6    'yes'        34.6    34.6
    'wen9'         234    'yes'         234     234
    'ple2'           2    'no'            2       2
    'dbo8'           5    'no'            5       5
    'oii4'           5    'yes'           5       5
    'wnk3'         245    'yes'         245     245
    'abk6'         563    ''            563     563
    'pnj5'         463    'no'          463     463
    'wnn3'           6    'no'            6       6
    'oks9'          23    'yes'          23      23
    'wba3'         NaN    'yes'         NaN      14
    'pkn4'           2    'no'            2       2
    'adw3'          22    'no'           22      22
    'poj2'         -99    'yes'         -99     -99
    'bas8'          23    'no'           23      23
    'gry5'         NaN    'yes'         NaN      21
```

Now, `var2` and `var4` are numeric arrays. During the conversion, `str2double` replaces the nonnumeric elements of the variables `var2` and `var4` with the value `NaN`. However, there are no changes to the numeric missing value indicator, `-99`.

When applying the same function to many dataset array variables, it can sometimes be more convenient to use `datasetfun`. For example, to convert both `var2` and `var4` to numeric arrays simultaneously, you can use:

```
messyData(:,[2,4]) = datasetfun(@str2double,messyData, ...
    'DataVars',[2,4],'DatasetOutput',true);
```

Replace missing value indicators.

Clean the data so that the missing values indicated by the code `-99` have the standard MATLAB numeric missing value indicator, `NaN`.

```
messyData = replaceWithMissing(messyData,'NumericValues',-99)
```

```
messyData =

    var1          var2    var3          var4    var5
    'afe1'           3    'yes'            3       3
    'egh3'         NaN    'no'             7       7
    'wth4'           3    'yes'            3       3
    'atn2'          23    'no'            23      23
    'arg1'           5    'yes'            5       5
    'jre3'        34.6    'yes'         34.6    34.6
    'wen9'         234    'yes'          234     234
    'ple2'           2    'no'             2       2
    'dbo8'           5    'no'             5       5
    'oii4'           5    'yes'            5       5
    'wnk3'         245    'yes'          245     245
    'abk6'         563    ''             563     563
    'pnj5'         463    'no'           463     463
    'wnn3'           6    'no'             6       6
    'oks9'          23    'yes'           23      23
    'wba3'         NaN    'yes'          NaN      14
    'pkn4'           2    'no'             2       2
    'adw3'          22    'no'            22      22
    'poj2'         NaN    'yes'          NaN     NaN
    'bas8'          23    'no'            23      23
    'gry5'         NaN    'yes'          NaN      21
```

Create a dataset array with complete observations.

Create a new dataset array that contains only the complete observations—those without missing data.

```
ix = ismissing(messyData);
completeData = messyData(~any(ix,2),:)
```

```
completeData =

    var1          var2    var3          var4    var5
    'afe1'           3    'yes'            3       3
    'wth4'           3    'yes'            3       3
    'atn2'          23    'no'            23      23
    'arg1'           5    'yes'            5       5
    'jre3'        34.6    'yes'         34.6    34.6
    'wen9'         234    'yes'          234     234
    'ple2'           2    'no'             2       2
    'dbo8'           5    'no'             5       5
    'oii4'           5    'yes'            5       5
    'wnk3'         245    'yes'          245     245
    'pnj5'         463    'no'           463     463
    'wnn3'           6    'no'             6       6
    'oks9'          23    'yes'           23      23
    'pkn4'           2    'no'             2       2
```

```
    'adw3'          22    'no'           22      22
    'bas8'          23    'no'           23      23
```

See Also

`dataset` | `ismissing` | `replaceWithMissing`

Related Examples

- “Select Subsets of Observations” on page 2-79
- “Calculations on Dataset Arrays” on page 2-92
- “Index and Search Dataset Arrays” on page 2-114

More About

- “Dataset Arrays” on page 2-112

Dataset Arrays in the Variables Editor

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB `table` data type instead. See MATLAB `table` documentation for more information.

In this section...

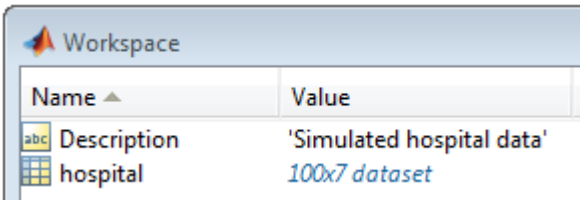
Open Dataset Arrays in the Variables Editor

The MATLAB Variables editor provides a convenient interface for viewing, modifying, and plotting dataset arrays.

First, load the sample data set, `hospital`.

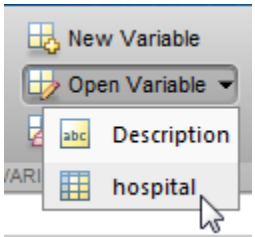
```
load hospital
```

The dataset array, `hospital`, is created in the MATLAB workspace.



The dataset array has 100 observations and 7 variables.

To open `hospital` in the Variables editor, click **Open Variable**, and select `hospital`.



The Variables editor opens, displaying the contents of the dataset array (only the first 10 observations are shown here).

hospital

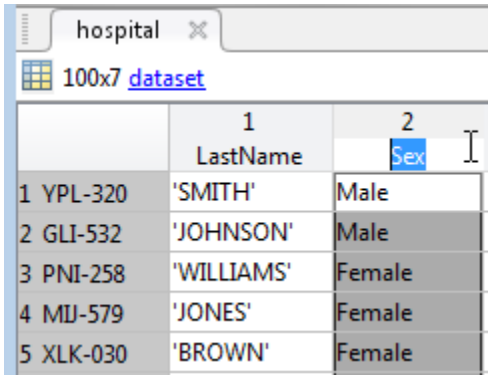
100x7 dataset

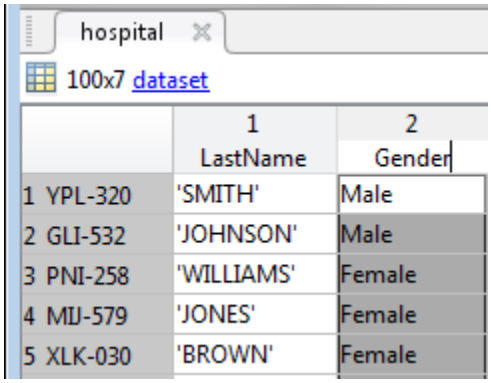
	1 LastName	2 Sex	3 Age	4 Weight	5 Smoker	6 BloodPressure		7 Trials
1	YPL-320	'SMITH'	Male	38	176	1	124	93	18
2	GLI-532	'JOHNSON'	Male	43	163	0	109	77	[11,13,22]
3	PNI-258	'WILLIAMS'	Female	38	131	0	125	83	[]
4	MIJ-579	'JONES'	Female	40	133	0	117	75	[6,12]
5	XLK-030	'BROWN'	Female	49	119	0	122	80	[14,23]
6	TFP-518	'DAVIS'	Female	46	142	0	121	70	19
7	LPD-746	'MILLER'	Female	33	142	1	130	88	13
8	ATA-945	'WILSON'	Male	40	180	0	115	82	[]
9	VNL-702	'MOORE'	Male	28	183	0	115	78	2
10	LQW-768	'TAYLOR'	Female	31	132	0	118	86	11

In the Variables editor, you can see the names of the seven variables along the top row, and the observations names down the first column.

Modify Variable and Observation Names

You can modify variable and observation names by double-clicking a name, and then typing new text.

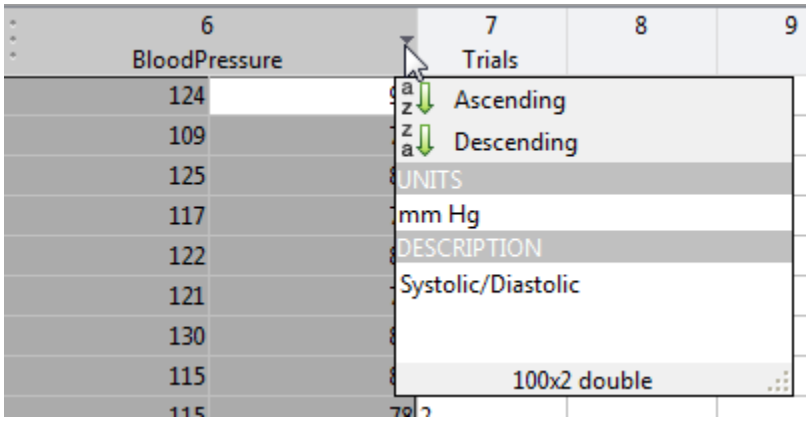




All changes made in the Variables editor are also sent to the command line.

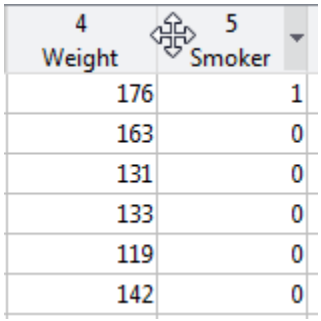
```
Command Window
>> hospital.Properties.VarNames{2} = 'Gender';
>>
```

The sixth variable in the data set, `BloodPressure`, is a numeric array with two columns. The first column shows systolic blood pressure, and the second column shows diastolic blood pressure. Click the arrow that appears on the right side of the variable name cell to see the units and description of the variable. You can type directly in the units and description fields to modify the text. The variable data type and size are shown under the variable description.

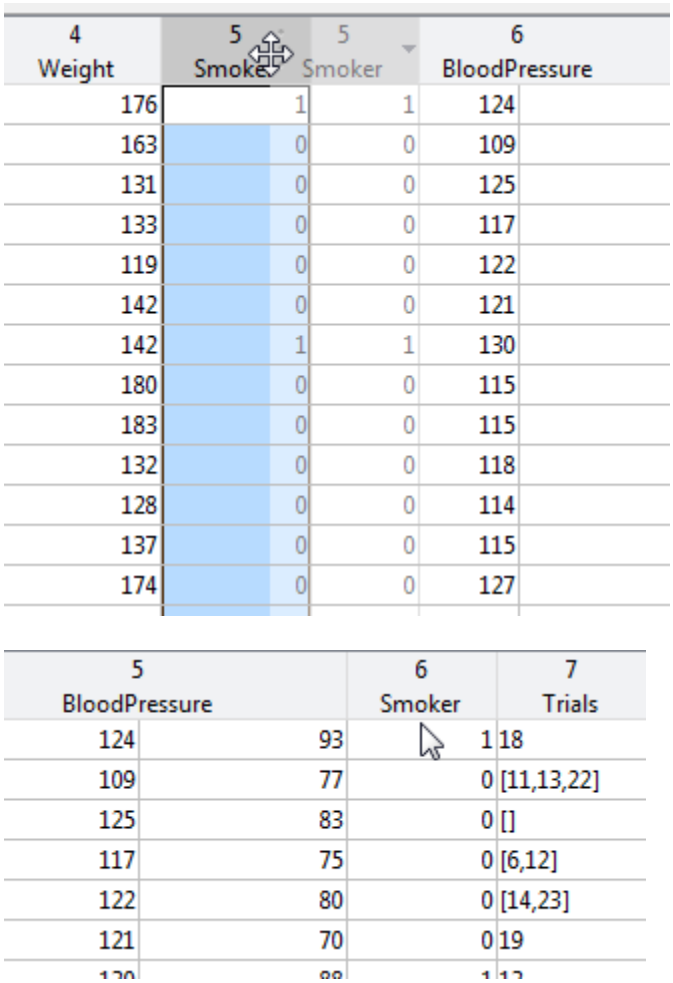


Reorder or Delete Variables

You can reorder variables in a dataset array using the Variables editor. Hover over the left side of a variable name cell until a four-headed arrow appears.



After the arrow appears, click and drag the variable column to a new location.



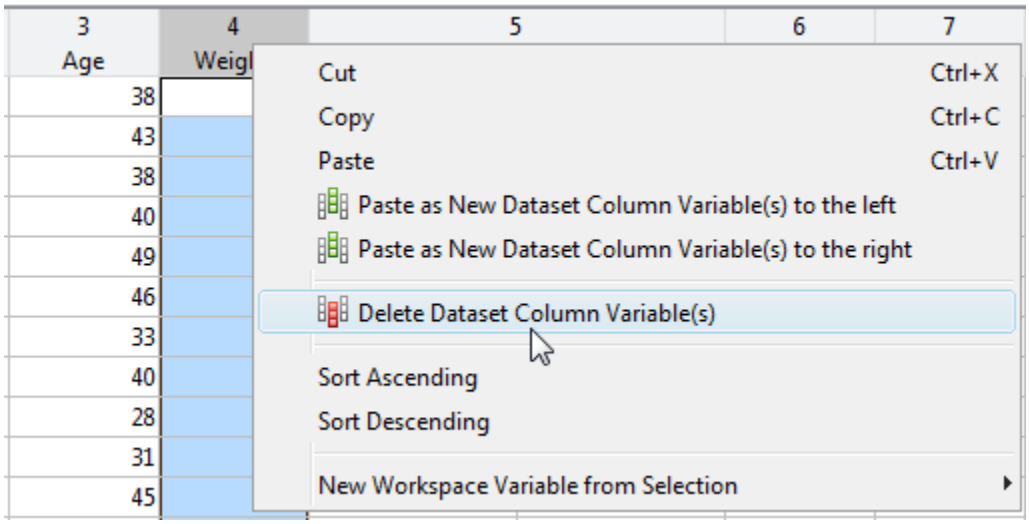
4 Weight	5 Smoke	5 Smoker	6 BloodPressure
176	1	1	124
163	0	0	109
131	0	0	125
133	0	0	117
119	0	0	122
142	0	0	121
142	1	1	130
180	0	0	115
183	0	0	115
132	0	0	118
128	0	0	114
137	0	0	115
174	0	0	127

5 BloodPressure		6 Smoker	7 Trials
124	93	1	18
109	77	0	[11,13,22]
125	83	0	[]
117	75	0	[6,12]
122	80	0	[14,23]
121	70	0	19

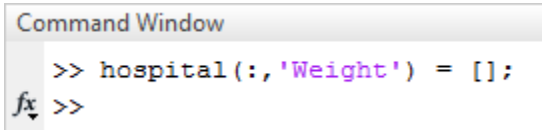
The command for the variable reordering appears in the command line.

```
Command Window
>> hospital = hospital(:,[1:4 6 5 end]);
>>
```

You can delete a variable in the Variables editor by selecting the variable column, right-clicking, and selecting **Delete Column Variable(s)**.

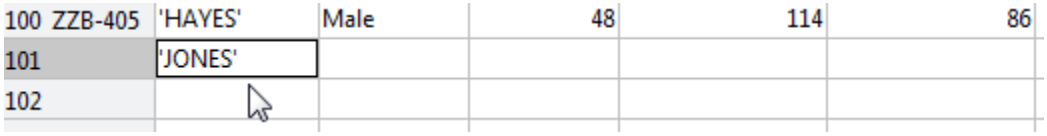


The command for the variable deletion appears in the command line.

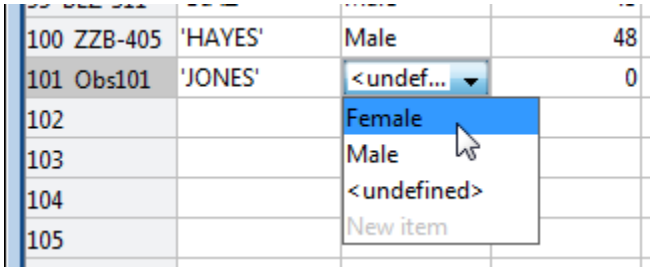


Add New Data

You can enter new data values directly into the Variables editor. For example, you can add a new patient observation to the `hospital` data set. To enter a new last name, add a character vector to the end of the variable `LastName`.



The variable `Gender` is a `nominal` array. The levels of the categorical variable appear in a drop-down list when you double-click a cell in the `Gender` column. You can choose one of the levels previously used, or create a new level by selecting **New Item**.



You can continue to add data for the remaining variables.

To change the observation name, click the observation name and type the new name.

100	ZZB-405	'HAYES'	Male	48
101	Obs101	'JONES'	Female	45
102				

The commands for entering the new data appear at the command line.

```
Command Window
>> hospital.LastName{101} = 'JONES';
Warning: Observations with default values added to dataset
variables.
> In dataset.subsasgn at 584
>> hospital.Sex(101) = 'Female';
>> hospital.Age(101) = 45;
>> hospital.BloodPressure(101,2) = 85;
>> hospital.BloodPressure(101,1) = 120;
>> hospital.Properties.ObsNames{101} = 'QPO-187';
fx >>
```

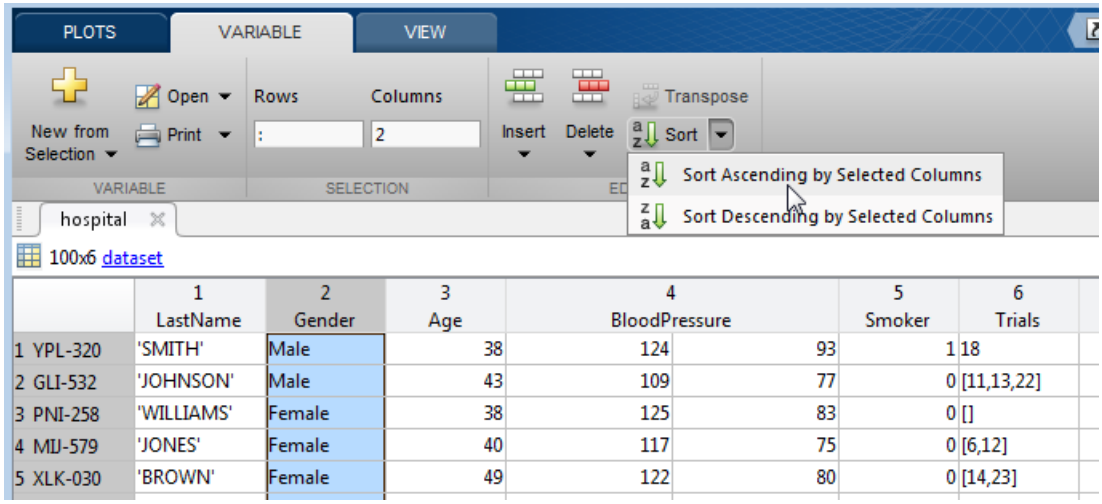
Notice the warning that appears after the first assignment. When you enter the first piece of data in the new observation row—here, the last name—default values are assigned to all other variables. Default assignments are:

- `0` for numeric variables
- `<undefined>` for categorical variables
- `[]` for cell arrays

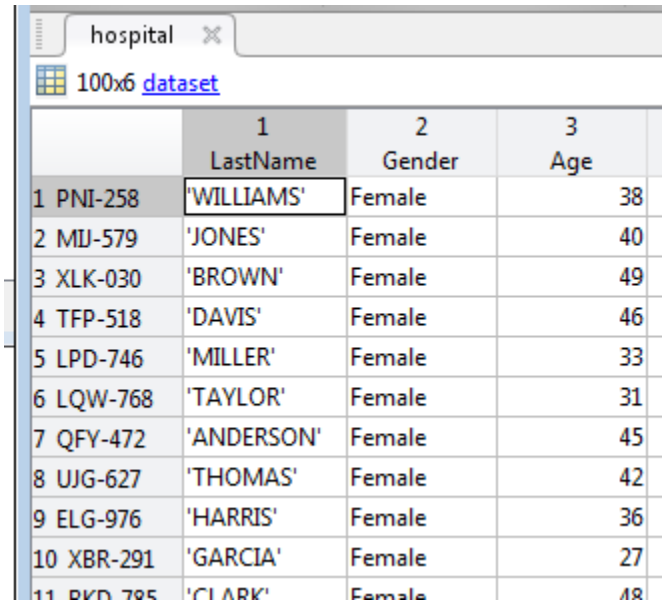
You can also copy and paste data from one dataset array to another using the Variables editor.

Sort Observations

You can use the Variables editor to sort dataset array observations by the values of one or more variables. To sort by gender, for example, select the variable `Gender`. Then click **Sort**, and choose to sort rows by ascending or descending values of the selected variable.



When sorting by variables that are cell arrays of character vectors or of nominal data type, observations are sorted alphabetically. For ordinal variables, rows are sorted by the ordering of the levels. For example, when the observations of `hospital` are sorted by the values in `Gender`, the females are grouped together, followed by the males.



	1 LastName	2 Gender	3 Age
1	PNI-258	'WILLIAMS'	Female	38
2	MIJ-579	'JONES'	Female	40
3	XLK-030	'BROWN'	Female	49
4	TFP-518	'DAVIS'	Female	46
5	LPD-746	'MILLER'	Female	33
6	LQW-768	'TAYLOR'	Female	31
7	QFY-472	'ANDERSON'	Female	45
8	UJG-627	'THOMAS'	Female	42
9	ELG-976	'HARRIS'	Female	36
10	XBR-291	'GARCIA'	Female	27
11	RKD-785	'CLARK'	Female	48

To sort by the values of multiple variables, press **Ctrl** while you select multiple variables.

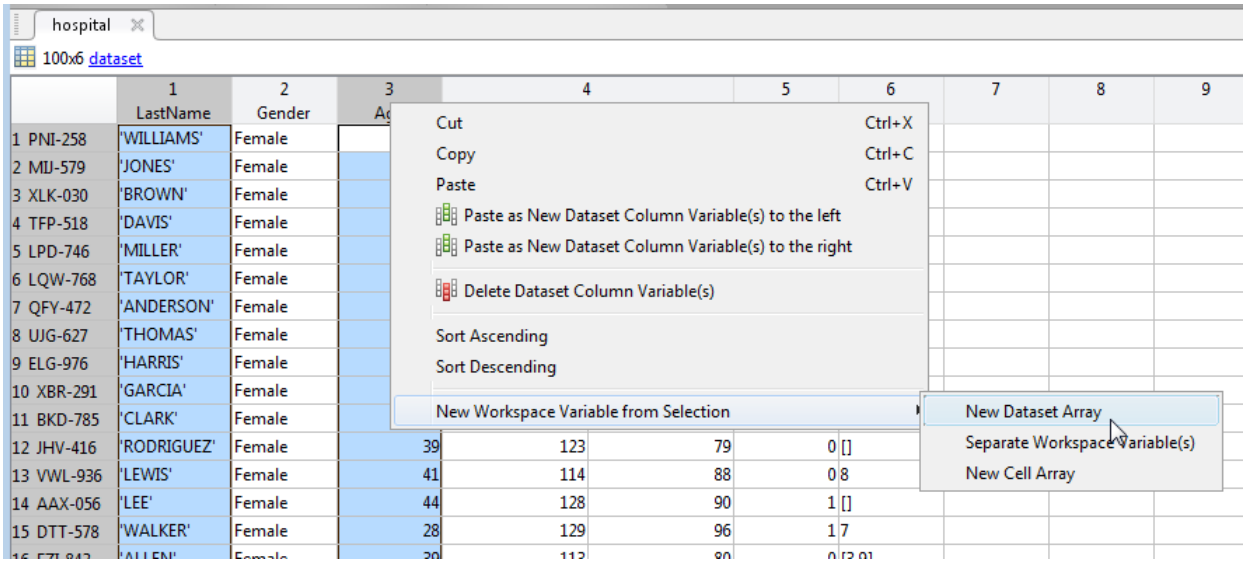
When you use the Variables editor to sort rows, it is the same as calling `sortrows`. You can see this at the command line after executing the sorting.

```
Command Window
>> hospital = sortrows(hospital,'Gender','ascend');
>>
```

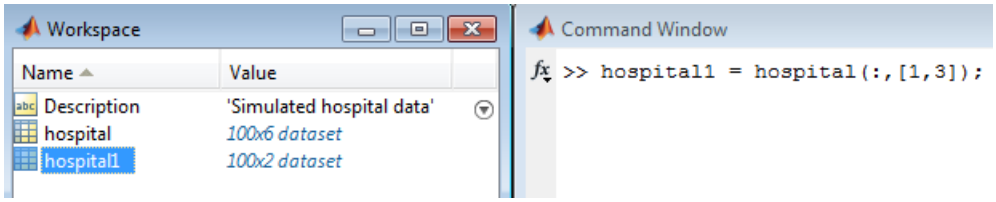
Select a Subset of Data

You can select a subset of data from a dataset array in the Variables editor, and create a new dataset array from the selection. For example, to create a dataset array containing only the variables `LastName` and `Age`:

1. Hold **Ctrl** while you click the variables `LastName` and `Age`.
2. Right-click, and select **New Workspace Variable from Selection** > **New Dataset Array**.

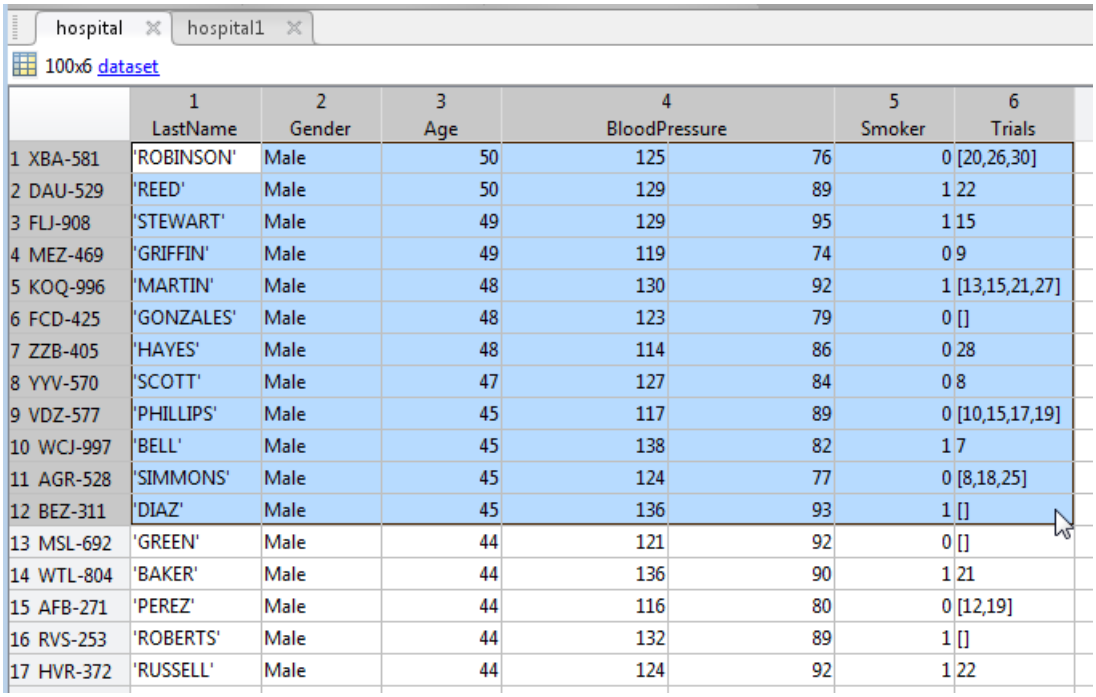


The new dataset array appears in the Workspace window with the name `hospital1`. The Command Window shows the commands that execute the selection.



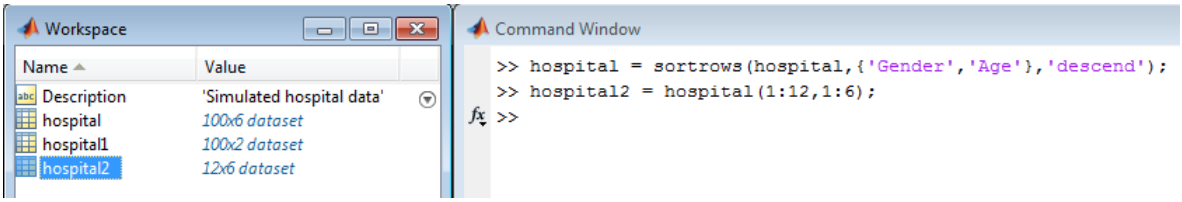
You can use the same steps to select any subset of data. To select observations according to some logical condition, you can use a combination of sorting and selecting. For example, to create a new dataset array containing only males aged 45 and older:

1. Sort the observations of `hospital` by the values in `Gender` and `Age`, descending.
2. Select the male observations with age 45 and older.

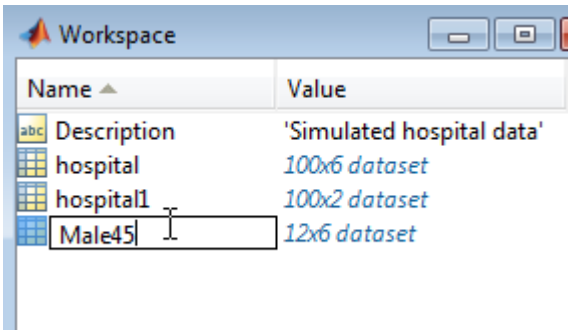


	1 LastName	2 Gender	3 Age	4 BloodPressure		5 Smoker	6 Trials
1	XBA-581	'ROBINSON'	Male	50	125	76	0 [20,26,30]
2	DAU-529	'REED'	Male	50	129	89	1 22
3	FLJ-908	'STEWART'	Male	49	129	95	1 15
4	MEZ-469	'GRIFFIN'	Male	49	119	74	0 9
5	KOQ-996	'MARTIN'	Male	48	130	92	1 [13,15,21,27]
6	FCD-425	'GONZALES'	Male	48	123	79	0 []
7	ZZB-405	'HAYES'	Male	48	114	86	0 28
8	YYV-570	'SCOTT'	Male	47	127	84	0 8
9	VDZ-577	'PHILLIPS'	Male	45	117	89	0 [10,15,17,19]
10	WCJ-997	'BELL'	Male	45	138	82	1 7
11	AGR-528	'SIMMONS'	Male	45	124	77	0 [8,18,25]
12	BEZ-311	'DIAZ'	Male	45	136	93	1 []
13	MSL-692	'GREEN'	Male	44	121	92	0 []
14	WTL-804	'BAKER'	Male	44	136	90	1 21
15	AFB-271	'PEREZ'	Male	44	116	80	0 [12,19]
16	RVS-253	'ROBERTS'	Male	44	132	89	1 []
17	HVR-372	'RUSSELL'	Male	44	124	92	1 22

3 Right-click, and select **New Workspace Variables from Selection > New Dataset Array**. The new dataset array, `hospital2`, is created in the Workspace window.



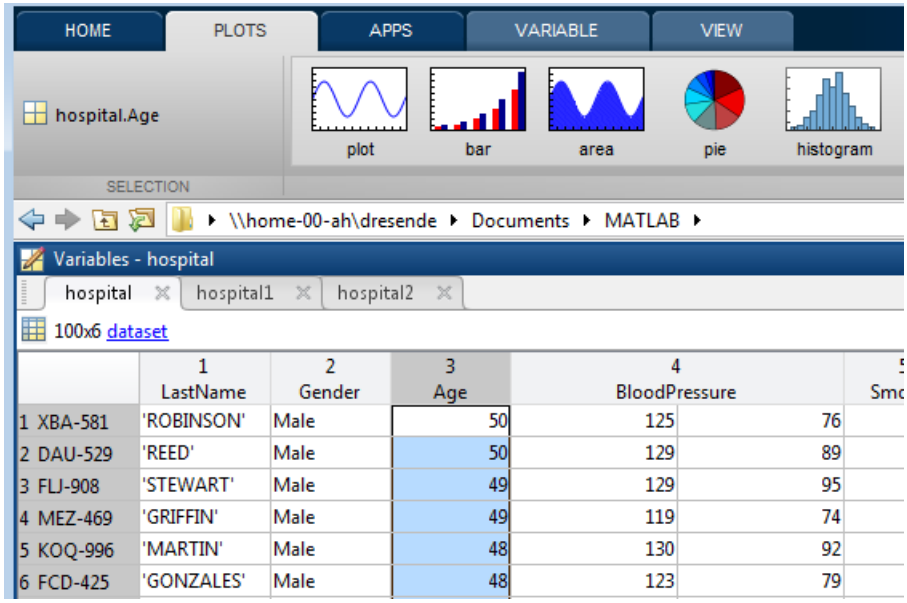
4 You can rename the dataset array in the Workspace window.



Create Plots

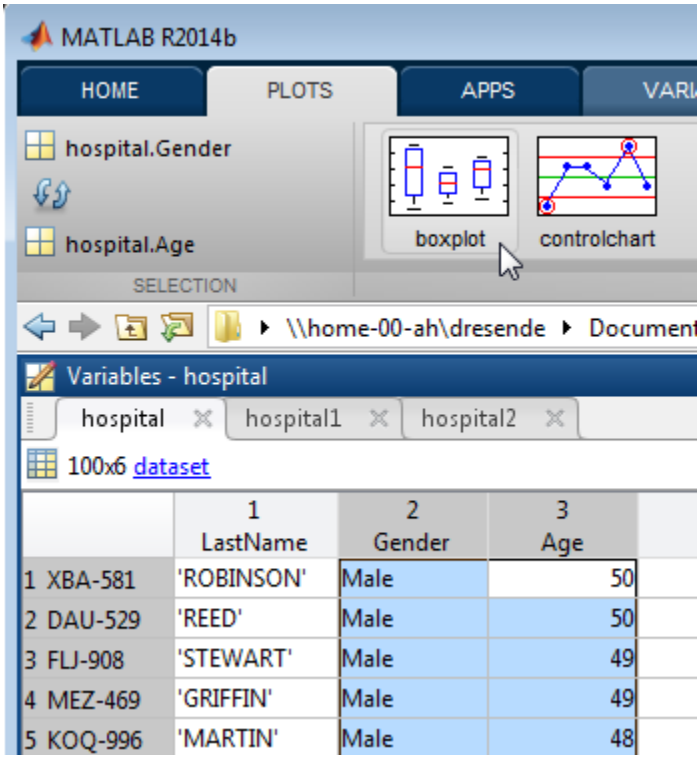
You can plot data from a dataset array using plotting options in the Variables editor. Available plot choices depend on the data types of variables to be plotted.

For example, if you select the variable `Age`, you can see in the **Plots** tab some plotting options that are appropriate for a univariate, numeric variable.



	1 LastName	2 Gender	3 Age	4 BloodPressure	
1	XBA-581	'ROBINSON'	Male	50	125	76
2	DAU-529	'REED'	Male	50	129	89
3	FLJ-908	'STEWART'	Male	49	129	95
4	MEZ-469	'GRIFFIN'	Male	49	119	74
5	KOQ-996	'MARTIN'	Male	48	130	92
6	FCD-425	'GONZALES'	Male	48	123	79

Sometimes, there are plot options for multiple variables, depending on their data types. For example, if you select both `Age` and `Gender`, you can draw box plots of age, grouped by gender.



	1 LastName	2 Gender	3 Age
1	XBA-581	'ROBINSON'	Male	50
2	DAU-529	'REED'	Male	50
3	FLJ-908	'STEWART'	Male	49
4	MEZ-469	'GRIFFIN'	Male	49
5	KOQ-996	'MARTIN'	Male	48

See Also

`dataset` | `sortrows`

Related Examples

- “Add and Delete Observations” on page 2-68
- “Add and Delete Variables” on page 2-71
- “Access Data in Dataset Array Variables” on page 2-74
- “Select Subsets of Observations” on page 2-79
- “Sort Observations in Dataset Arrays” on page 2-82

More About

- “Dataset Arrays” on page 2-112

Dataset Arrays

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB `table` data type instead. See MATLAB `table` documentation for more information.

In this section...

What Are Dataset Arrays?

Statistics and Machine Learning Toolbox has dataset arrays for storing variables with heterogeneous data types. For example, you can combine numeric data, logical data, cell arrays of character vectors, and categorical arrays in one dataset array variable.

Within a dataset array, each variable (column) must be one homogeneous data type, but the different variables can be of heterogeneous data types. A dataset array is usually interpreted as a set of variables measured on many units of observation. That is, each row in a dataset array corresponds to an observation, and each column to a variable. In this sense, a dataset array organizes data like a typical spreadsheet.

Dataset arrays are a unique data type, with a corresponding set of valid operations. Even if a dataset array contains only numeric variables, you cannot operate on the dataset array like a numeric variable. The valid operations for dataset arrays are the methods of the `dataset` class.

Dataset Array Conversion

You can create a dataset array by combining variables that exist in the MATLAB workspace, or directly importing data from a file, such as a text file or spreadsheet. This table summarizes the functions you can use to create dataset arrays.

Data Source	Conversion to Dataset Array
Data from a file	`dataset`
Heterogeneous collection of workspace variables	`dataset`
Numeric array	`mat2dataset`
Cell array	`cell2dataset`
Structure array	`struct2dataset`
Table	`table2dataset`

You can export dataset arrays to text or spreadsheet files using `export`. To convert a dataset array to a cell array or structure array, use `dataset2cell` or `dataset2struct`. To convert a dataset array to a table, use `dataset2table`.

Dataset Array Properties

In addition to storing data in a dataset array, you can store metadata such as:

- Variable and observation names
- Data descriptions
- Units of measurement
- Variable descriptions

This information is stored as dataset array properties. For a dataset array named `ds`, you can view the dataset array metadata by entering `ds.Properties` at the command line. You can access a specific property, such as variable names—property `VarNames`—using `ds.Properties.VarNames`. You can both retrieve and modify property values using this syntax.

Variable and observation names are included in the display of a dataset array. Variable names display across the top row, and observation names, if present, appear in the first column. Note that variable and observation names do not affect the size of a dataset array.

See Also

`cell2dataset` | `dataset` | `dataset2cell` | `dataset2struct` | `dataset2table` | `export` | `mat2dataset` | `struct2dataset` | `table2dataset`

Related Examples

- “Create a Dataset Array from Workspace Variables” on page 2-57
- “Create a Dataset Array from a File” on page 2-62
- “Export Dataset Arrays” on page 2-95
- “Dataset Arrays in the Variables Editor” on page 2-101
- “Index and Search Dataset Arrays” on page 2-114

Index and Search Dataset Arrays

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB `table` data type instead. See MATLAB `table` documentation for more information.

Ways To Index and Search

There are many ways to index into dataset arrays. For example, for a dataset array, `ds`, you can:

- Use `()` to create a new dataset array from a subset of `ds`. For example, `ds1 = ds(1:5,:)` creates a new dataset array, `ds1`, consisting of the first five rows of `ds`. Metadata, including variable and observation names, transfers to the new dataset array.
- Use variable names with dot notation to index individual variables in a dataset array. For example, `ds.Height` indexes the variable named `Height`.
- Use observation names to index individual observations in a dataset array. For example, `ds('Obs1',:)` gives data for the observation named `Obs1`.
- Use observation or variable numbers. For example, `ds(:,[1,3,5])` gives the data in the first, third, and fifth variables (columns) of `ds`.
- Use logical indexing to search for observations in `ds` that satisfy a logical condition. For example, `ds(ds.Gender=='Male',:)` gives the observations in `ds` where the variable named `Gender`, a nominal array, has the value `Male`.
- Use `ismissing` to find missing data in the dataset array.

Examples

Common Indexing and Searching Methods

This example shows several indexing and searching methods for categorical arrays.

Load the sample data.

```
load hospital;
size(hospital)
```

```
ans = 1×2

   100     7
```

The dataset array has 100 observations and 7 variables.

Index a variable by name. Return the minimum age in the dataset array.

```
min(hospital.Age)
```

```
ans = 25
```

Delete the variable `Trials`.

```
hospital.Trials = [];
size(hospital)
```

```
ans = 1×2

   100     6
```

Index an observation by name. Display measurements on the first five variables for the observation named PUE-347.

```
hospital('PUE-347',1:5)
```

```
ans =
                LastName          Sex       Age    Weight    Smoker
    PUE-347     {'YOUNG'}         Female    25     114       false
```

Index variables by number. Create a new dataset array containing the first four variables of hospital.

```
dsNew = hospital(:,1:4);
dsNew.Properties.VarNames(:)
```

```
ans = 4x1 cell
    {'LastName'}
    {'Sex'     }
    {'Age'     }
    {'Weight'  }
```

Index observations by number. Delete the last 10 observations.

```
hospital(end-9:end,:) = [];
size(hospital)
```

```
ans = 1×2

    90     6
```

Search for observations by logical condition. Create a new dataset array containing only females who smoke.

```
dsFS = hospital(hospital.Sex=='Female' & hospital.Smoker==true,:);
dsFS(:,{'LastName','Sex','Smoker'})
```

```
ans =
                LastName             Sex       Smoker
    LPD-746     {'MILLER'   }        Female    true
    XBR-291     {'GARCIA'   }        Female    true
    AAX-056     {'LEE'      }        Female    true
    DTT-578     {'WALKER'   }        Female    true
    AFK-336     {'WRIGHT'   }        Female    true
    RBA-579     {'SANCHEZ'  }        Female    true
    HAK-381     {'MORRIS'   }        Female    true
    NSK-403     {'RAMIREZ'  }        Female    true
    ILS-109     {'WATSON'   }        Female    true
    JDR-456     {'SANDERS'  }        Female    true
    HWZ-321     {'PATTERSON'}        Female    true
    GGU-691     {'HUGHES'   }        Female    true
```

```
    WUS-105    {'FLORES'   }        Female    true
```

See Also

`dataset`

Related Examples

- “Access Data in Dataset Array Variables” on page 2-74
- “Select Subsets of Observations” on page 2-79

More About

- “Dataset Arrays” on page 2-112

Descriptive Statistics

Measures of Central Tendency

Measures of central tendency locate a distribution of data along an appropriate scale.

The following table lists the functions that calculate the measures of central tendency.

Function Name	Description
`geomean`	Geometric mean
`harmmean`	Harmonic mean
`mean`	Arithmetic average
`median`	50th percentile
`mode`	Most frequent value
`trimmean`	Trimmed mean

The average is a simple and popular estimate of location. If the data sample comes from a normal distribution, then the sample mean is also optimal (minimum variance unbiased estimator (MVUE) of μ).

Unfortunately, outliers, data entry errors, or glitches exist in almost all real data. The sample mean is sensitive to these problems. One bad data value can move the average away from the center of the rest of the data by an arbitrarily large distance.

The median and trimmed mean are two measures that are resistant (robust) to outliers. The median is the 50th percentile of the sample, which will only change slightly if you add a large perturbation to any value. The idea behind the trimmed mean is to ignore a small percentage of the highest and lowest values of a sample when determining the center of the sample.

The geometric mean and harmonic mean, like the average, are not robust to outliers. They are useful when the sample is distributed lognormal or heavily skewed.

Measures of Central Tendency

This example shows how to compute and compare measures of location for sample data that contains one outlier.

Generate sample data that contains one outlier.

```
x = [ones(1,6),100]
```

```
x = 1×7

     1     1     1     1     1     1   100
```

Compute the geometric mean, harmonic mean, mean, median, and trimmed mean for the sample data.

```
locate = [geomean(x) harmmean(x) mean(x) median(x)...
          trimmean(x,25)]
```

```
locate = 1×5
```

```
    1.9307    1.1647   15.1429    1.0000    1.0000
```

The mean (`mean`) is far from any data value because of the influence of the outlier. The geometric mean (`geomean`) and the harmonic mean (`harmmean`) are influenced by the outlier, but not as significantly. The median (`median`) and trimmed mean (`trimmean`) ignore the outlier value and describe the location of the rest of the data values.

See Also

Related Examples

- “Exploratory Analysis of Data” on page 3-10

Measures of Dispersion

The purpose of measures of dispersion is to find out how spread out the data values are on the number line. Another term for these statistics is measures of spread.

The table gives the function names and descriptions.

Function Name	Description
`iqr`	Interquartile range
`mad`	Mean absolute deviation
`moment`	Central moment of all orders
`range`	Range
`std`	Standard deviation
`var`	Variance

The range (the difference between the maximum and minimum values) is the simplest measure of spread. But if there is an outlier in the data, it will be the minimum or maximum value. Thus, the range is not robust to outliers.

The standard deviation and the variance are popular measures of spread that are optimal for normally distributed samples. The sample variance is the minimum variance unbiased estimator (MVUE) of the normal parameter σ^2. The standard deviation is the square root of the variance and has the desirable property of being in the same units as the data. That is, if the data is in meters, the standard deviation is in meters as well. The variance is in meters^2, which is more difficult to interpret.

Neither the standard deviation nor the variance is robust to outliers. A data value that is separate from the body of the data can increase the value of the statistics by an arbitrarily large amount.

The mean absolute deviation (MAD) is also sensitive to outliers. But the MAD does not move quite as much as the standard deviation or variance in response to bad data.

The interquartile range (IQR) is the difference between the 75th and 25th percentile of the data. Since only the middle 50% of the data affects this measure, it is robust to outliers.

Compare Measures of Dispersion

This example shows how to compute and compare measures of dispersion for sample data that contains one outlier.

Generate sample data that contains one outlier value.

```
x = [ones(1,6),100]
```

```
x = 1×7

     1     1     1     1     1     1   100
```

Compute the interquartile range, mean absolute deviation, range, and standard deviation of the sample data.

```
stats = [iqr(x),mad(x),range(x),std(x)]
```

```
stats = 1×4

         0   24.2449   99.0000   37.4185
```

The interquartile range (`iqr`) is the difference between the 75th and 25th percentile of the sample data, and is robust to outliers. The range (`range`) is the difference between the maximum and minimum values in the data, and is strongly influenced by the presence of an outlier.

Both the mean absolute deviation (`mad`) and the standard deviation (`std`) are sensitive to outliers. However, the mean absolute deviation is less sensitive than the standard deviation.

See Also

Related Examples

- “Exploratory Analysis of Data” on page 3-10

Quantiles and Percentiles

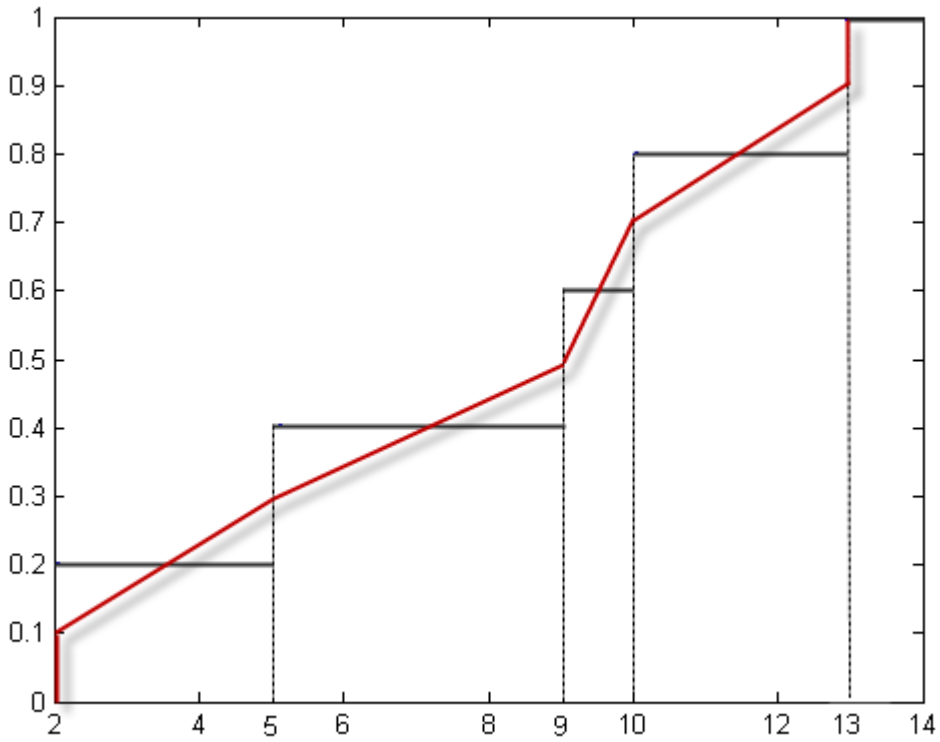
This section explains how the Statistics and Machine Learning Toolbox functions `quantile` and `prctile` compute quantiles and percentiles.

The `prctile` function calculates the percentiles in a similar way as `quantile` calculates quantiles. The following steps in the computation of quantiles are also true for percentiles, given the fact that, for the same data sample, the quantile at the value Q is the same as the percentile at the value P = 100*Q.

1 `quantile` initially assigns the sorted values in X to the (0.5/*n*), (1.5/*n*), ..., ([*n* – 0.5]/*n*) quantiles. For example:

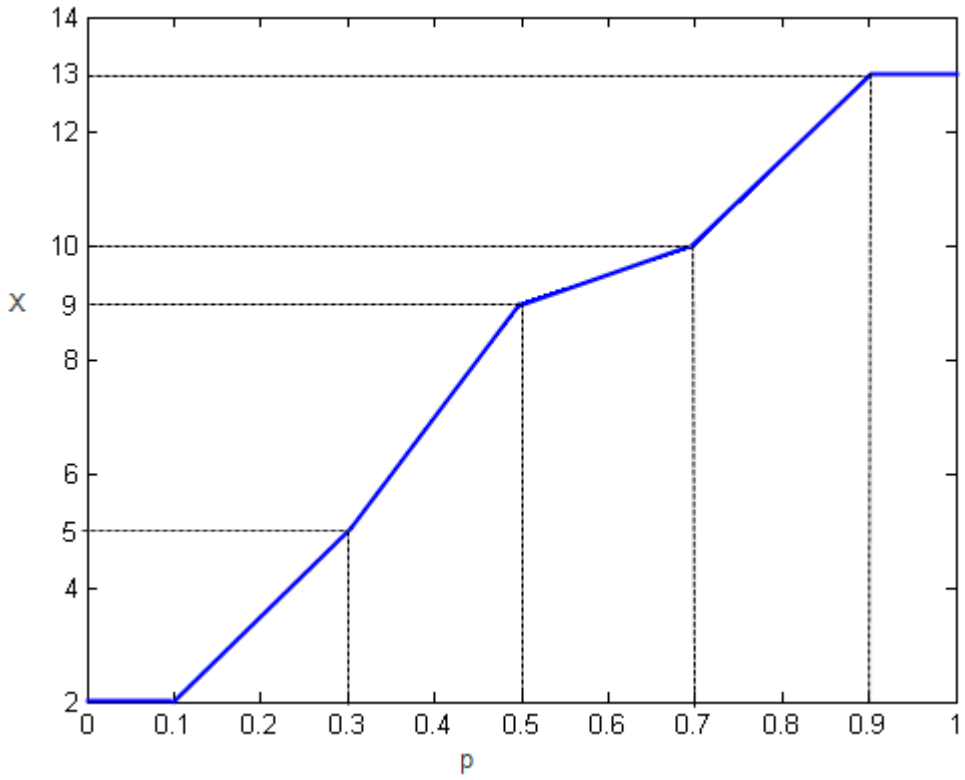
- For a data vector of six elements such as {6, 3, 2, 10, 8, 1}, the sorted elements {1, 2, 3, 6, 8, 10} respectively correspond to the (0.5/6), (1.5/6), (2.5/6), (3.5/6), (4.5/6), and (5.5/6) quantiles.
- For a data vector of five elements such as {2, 10, 5, 9, 13}, the sorted elements {2, 5, 9, 10, 13} respectively correspond to the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles.

The following figure illustrates this approach for data vector X = {2, 10, 5, 9, 13}. The first observation corresponds to the cumulative probability 1/5 = 0.2, the second observation corresponds to the cumulative probability 2/5 = 0.4, and so on. The step function in this figure shows these cumulative probabilities. `quantile` instead places the observations in midpoints, such that the first corresponds to 0.5/5 = 0.1, the second corresponds to 1.5/5 = 0.3, and so on, and then connects these midpoints. The red lines in the following figure connect the midpoints.



Assigning Observations to Quantiles

By switching the axes, as the next figure, you can see the values of the variable X that correspond to the p quantiles.



Quantiles of *X*

2 `quantile` finds any quantiles between the data values using linear interpolation.

Linear interpolation uses linear polynomials to approximate a function f(x) and construct new data points within the range of a known set of data points. Algebraically, given the data points (x_1, y_1) and (x_2, y_2), where y_1 = f(x_1) and y_2 = f(x_2), linear interpolation finds y = f(x) for a given x between x_1 and x_2 as follows:

$$y = f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1).$$

Similarly, if the 1.5/n quantile is $y_{1.5/n}$ and the 2.5/n quantile is $y_{2.5/n}$, then linear interpolation finds the 2.3/n quantile $y_{2.3/n}$ as

$$y_{\frac{2.3}{n}} = y_{\frac{1.5}{n}} + \frac{\left(\frac{2.3}{n} - \frac{1.5}{n}\right)}{\left(\frac{2.5}{n} - \frac{1.5}{n}\right)}\left(y_{\frac{2.5}{n}} - y_{\frac{1.5}{n}}\right).$$

3 `quantile` assigns the first and last values of X to the quantiles for probabilities less than (0.5/n) and greater than ([n–0.5]/n), respectively.

References

[1] Langford, E. “Quartiles in Elementary Statistics”, *Journal of Statistics Education*. Vol. 14, No. 3, 2006.

See Also

`median` | `prctile` | `quantile`

Related Examples

- “Exploratory Analysis of Data” on page 3-10

Exploratory Analysis of Data

This example shows how to explore the distribution of data using descriptive statistics.

Generate sample data.

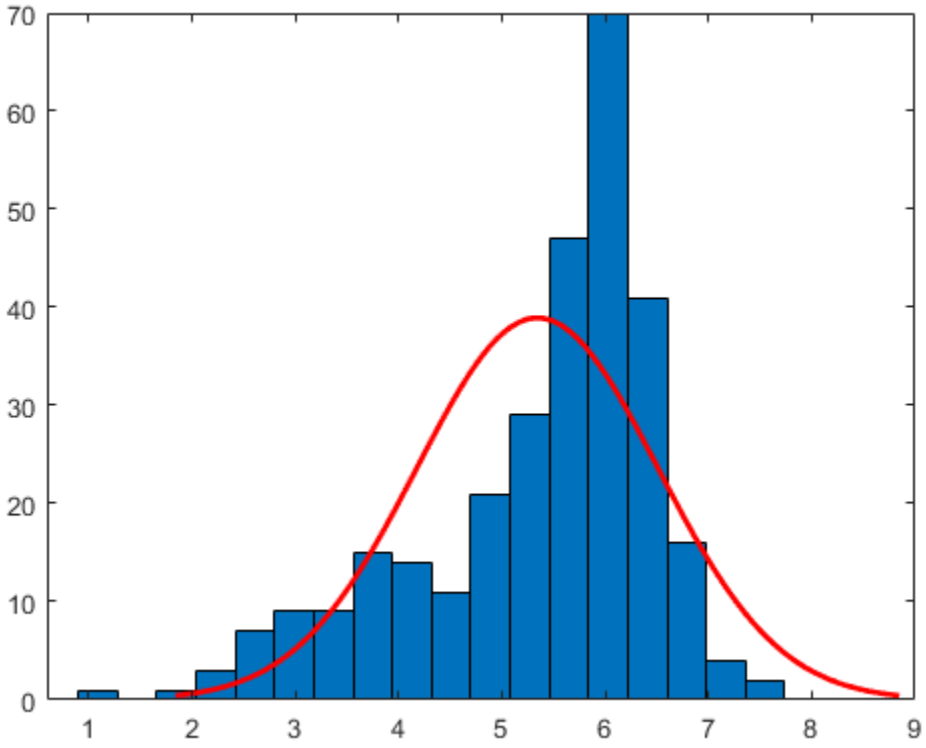
Generate a vector containing randomly-generated sample data.

```
rng default  % For reproducibility
x = [normrnd(4,1,1,100),normrnd(6,0.5,1,200)];
```

Plot a histogram.

Plot a histogram of the sample data with a normal density fit. This provides a visual comparison of the sample data and a normal distribution fitted to the data.

```
histfit(x)
```

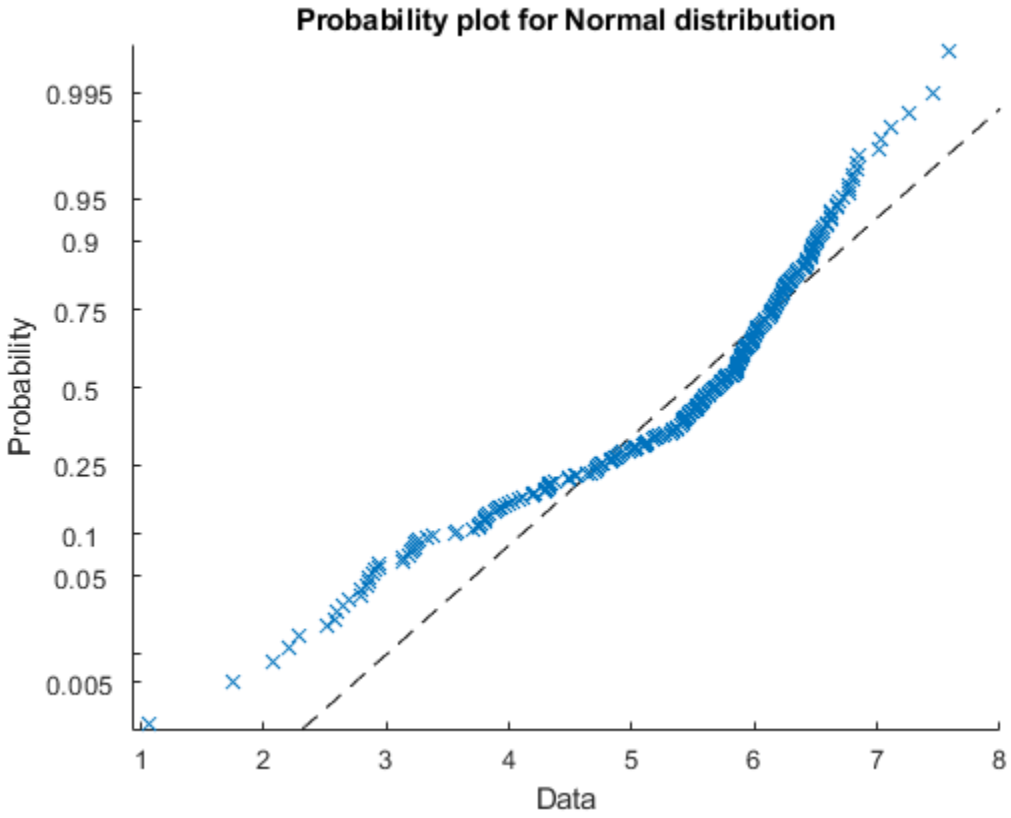


The distribution of the data appears to be left skewed. A normal distribution does not look like a good fit for this sample data.

Obtain a normal probability plot.

Obtain a normal probability plot. This plot provides another way to visually compare the sample data to a normal distribution fitted to the data.

```
probplot('normal',x)
```



The probability plot also shows the deviation of data from normality.

Compute the quantiles.

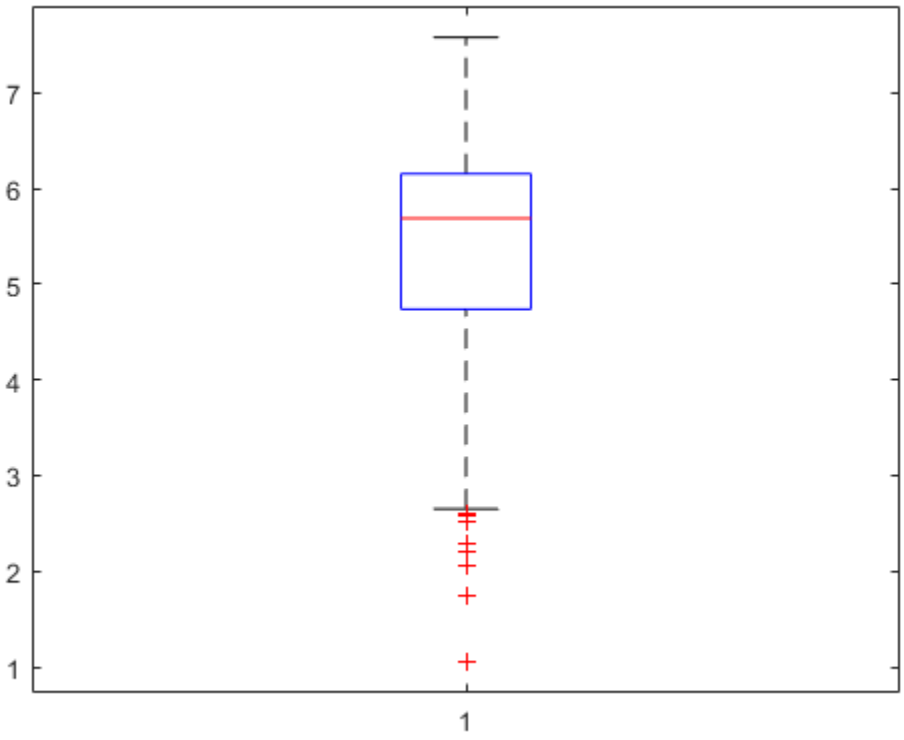
Compute the quantiles of the sample data.

```
p = 0:0.25:1;
y = quantile(x,p);
z = [p;y]
```

```
z = 2×5

         0    0.2500    0.5000    0.7500    1.0000
    1.0557    4.7375    5.6872    6.1526    7.5784
```

Create a box plot to visualize the statistics.

```
boxplot(x)
```



The box plot shows the 0.25, 0.5, and 0.75 quantiles. The long lower tail and plus signs show the lack of symmetry in the sample data values.

Compute descriptive statistics.

Compute the mean and median of the data.

```
y = [mean(x),median(x)]
```

```
y = 1×2

    5.3438    5.6872
```

The mean and median values seem close to each other, but a mean smaller than the median usually indicates that the data is left skewed.

Compute the skewness and kurtosis of the data.

```
y = [skewness(x),kurtosis(x)]
```

```
y = 1×2

   -1.0417    3.5895
```

A negative skewness value means the data is left skewed. The data has a larger peakedness than a normal distribution because the kurtosis value is greater than 3.

Compute z-scores.

Identify possible outliers by computing the z-scores and finding the values that are greater than 3 or less than -3.

```
Z = zscore(x);
find(abs(Z)>3);
```

Based on the z-scores, the 3rd and 35th observations might be outliers.

See Also

`boxplot` | `histfit` | `kurtosis` | `mean` | `median` | `prctile` | `quantile` | `skewness`

More About

- “Compare Grouped Data Using Box Plots” on page 4-4
- “Measures of Central Tendency” on page 3-2
- “Measures of Dispersion” on page 3-4
- “Quantiles and Percentiles” on page 3-6

Resampling Statistics

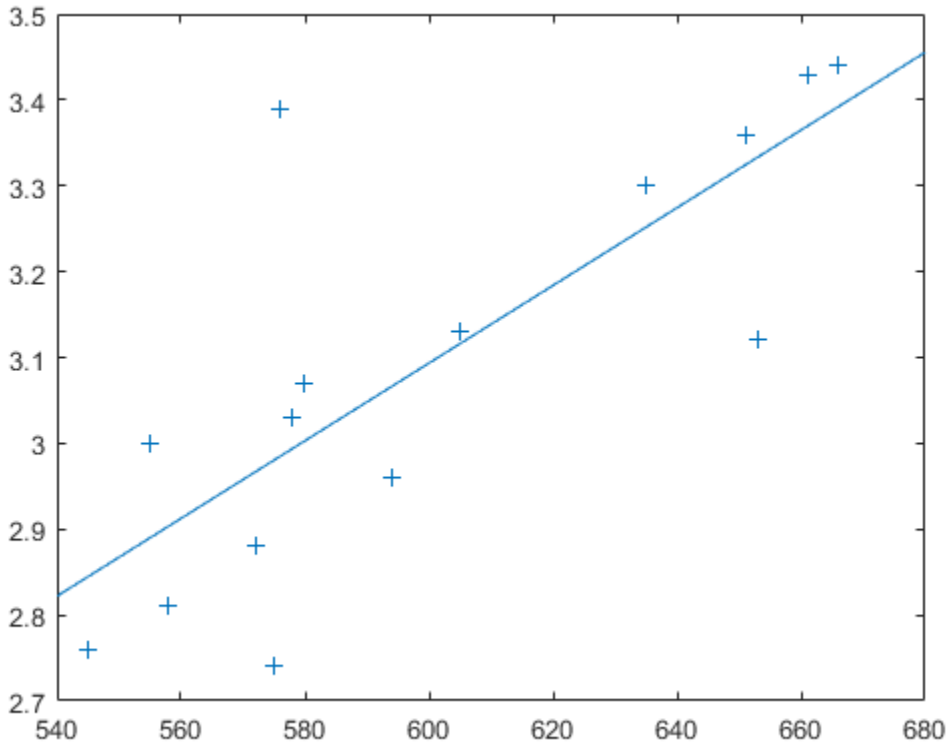
In this section...

Bootstrap Resampling

The bootstrap procedure involves choosing random samples with replacement from a data set and analyzing each sample the same way. Sampling with replacement means that each observation is selected separately at random from the original dataset. So a particular data point from the original data set could appear multiple times in a given bootstrap sample. The number of elements in each bootstrap sample equals the number of elements in the original data set. The range of sample estimates you obtain enables you to establish the uncertainty of the quantity you are estimating.

This example from Efron and Tibshirani compares Law School Admission Test (LSAT) scores and subsequent law school grade point average (GPA) for a sample of 15 law schools.

```
load lawdata
plot(lsat,gpa,'+')
lsline
```



The least-squares fit line indicates that higher LSAT scores go with higher law school GPAs. But how certain is this conclusion? The plot provides some intuition, but nothing quantitative.

You can calculate the correlation coefficient of the variables using the |corr|function.

```
rhohat = corr(lsat,gpa)
```

```
rhohat = 0.7764
```

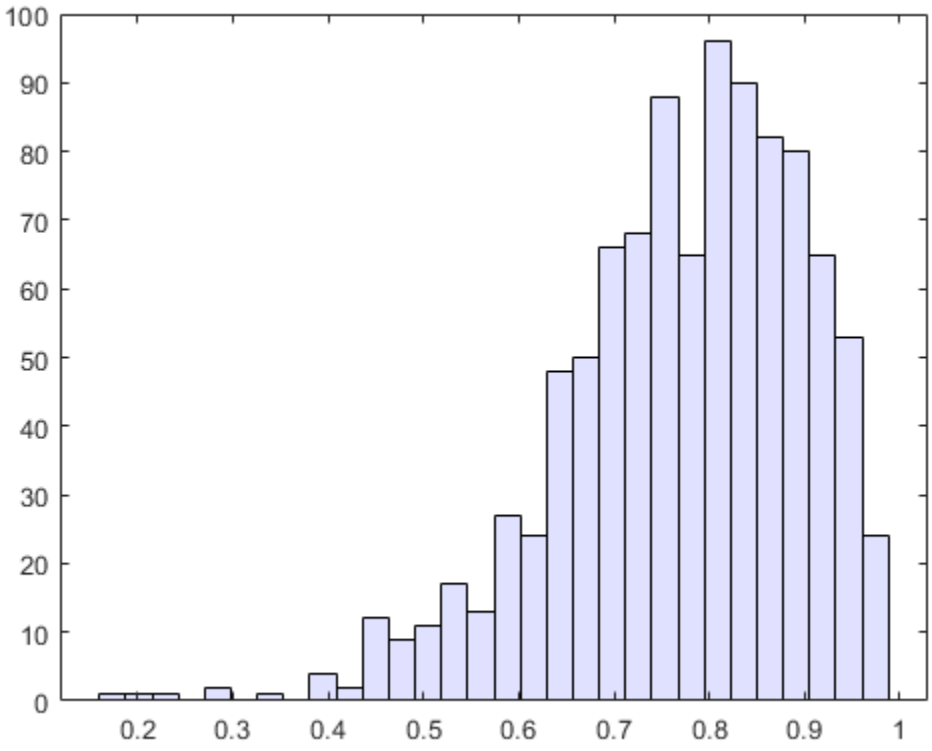
Now you have a number describing the positive connection between LSAT and GPA; though it may seem large, you still do not know if it is statistically significant.

Using the `bootstrp` function you can resample the `lsat` and `gpa` vectors as many times as you like and consider the variation in the resulting correlation coefficients.

```
rng default  % For reproducibility
rhos1000 = bootstrp(1000,'corr',lsat,gpa);
```

This resamples the `lsat` and `gpa` vectors 1000 times and computes the `corr` function on each sample. You can then plot the result in a histogram.

```
histogram(rhos1000,30,'FaceColor',[.8 .8 1])
```



Nearly all the estimates lie on the interval [0.4 1.0].

It is often desirable to construct a confidence interval for a parameter estimate in statistical inferences. Using the `bootci` function, you can use bootstrapping to obtain a confidence interval for the `lsat` and `gpa` data.

```
ci = bootci(5000,@corr,lsat,gpa)
```

```
ci = 2×1

    0.3319
    0.9427
```

Therefore, a 95% confidence interval for the correlation coefficient between LSAT and GPA is [0.33 0.94]. This is strong quantitative evidence that LSAT and subsequent GPA are positively correlated. Moreover, this evidence does not require any strong assumptions about the probability distribution of the correlation coefficient.

Although the `bootci` function computes the Bias Corrected and accelerated (BCa) interval as the default type, it is also able to compute various other types of bootstrap confidence intervals, such as the studentized bootstrap confidence interval.

Jackknife Resampling

Similar to the bootstrap is the jackknife, which uses resampling to estimate the bias of a sample statistic. Sometimes it is also used to estimate standard error of the sample statistic. The jackknife is implemented by the Statistics and Machine Learning Toolbox™ function `jackknife`.

The jackknife resamples systematically, rather than at random as the bootstrap does. For a sample with `n` points, the jackknife computes sample statistics on `n` separate samples of size `n`-1. Each sample is the original data with a single observation omitted.

In the bootstrap example, you measured the uncertainty in estimating the correlation coefficient. You can use the jackknife to estimate the bias, which is the tendency of the sample correlation to over-estimate or under-estimate the true, unknown correlation. First compute the sample correlation on the data.

```
load lawdata
rhohat = corr(lsat,gpa)
```

```
rhohat = 0.7764
```

Next compute the correlations for jackknife samples, and compute their mean.

```
rng default;  % For reproducibility
jackrho = jackknife(@corr,lsat,gpa);
meanrho = mean(jackrho)
```

```
meanrho = 0.7759
```

Now compute an estimate of the bias.

```
n = length(lsat);
biasrho = (n-1) * (meanrho-rhohat)
```

```
biasrho = -0.0065
```

The sample correlation probably underestimates the true correlation by about this amount.

Parallel Computing Support for Resampling Methods

For information on computing resampling statistics in parallel, see Parallel Computing Toolbox™.

Statistical Visualization

Create Scatter Plots Using Grouped Data

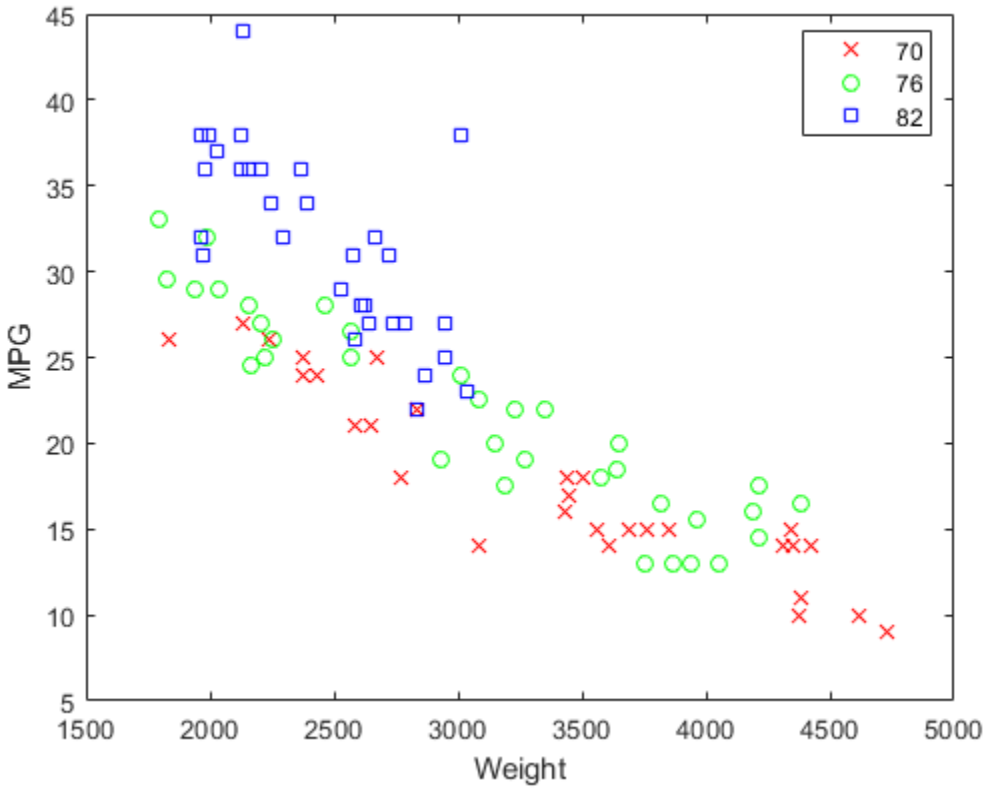
This example shows how to create scatter plots using grouped sample data.

A scatter plot is a simple plot of one variable against another. The MATLAB® functions `plot` and `scatter` produce scatter plots. The MATLAB function `plotmatrix` can produce a matrix of such plots showing the relationship between several pairs of variables.

Statistics and Machine Learning Toolbox™ functions `gscatter` and `gplotmatrix` produce grouped versions of these plots. These functions are useful for determining whether the values of two variables or the relationship between those variables is the same in each group. These functions use different plotting symbols to indicate group membership. You can use `gname` to label points on the plots with a text label or an observation number.

Suppose you want to examine the weight and mileage of cars from three different model years.

```
load carsmall
gscatter(Weight,MPG,Model_Year,'','xos')
```

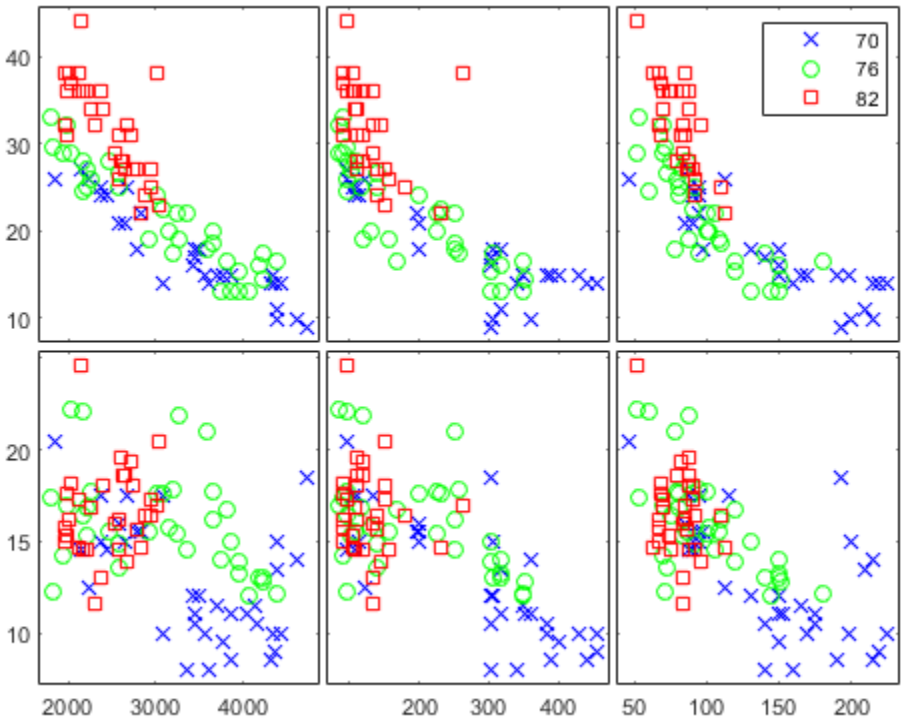


This shows that not only is there a strong relationship between the weight of a car and its mileage, but also that newer cars tend to be lighter and have better gas mileage than older cars.

The default arguments for `gscatter` produce a scatter plot with the different groups shown with the same symbol but different colors. The last two arguments above request that all groups be shown in default colors and with different symbols.

The `carsmall` data set contains other variables that describe different aspects of cars. You can examine several of them in a single display by creating a grouped plot matrix.

```
xvars = [Weight Displacement Horsepower];
yvars = [MPG Acceleration];
gplotmatrix(xvars,yvars,Model_Year,'','xos')
```



The upper right subplot displays `MPG` against `Horsepower`, and shows that over the years the horsepower of the cars has decreased but the gas mileage has improved.

The `gplotmatrix` function can also graph all pairs from a single list of variables, along with histograms for each variable. See “MANOVA” on page 9-49.

See Also

`gname` | `gplotmatrix` | `gscatter`

More About

- “Grouping Variables” on page 2-45

Compare Grouped Data Using Box Plots

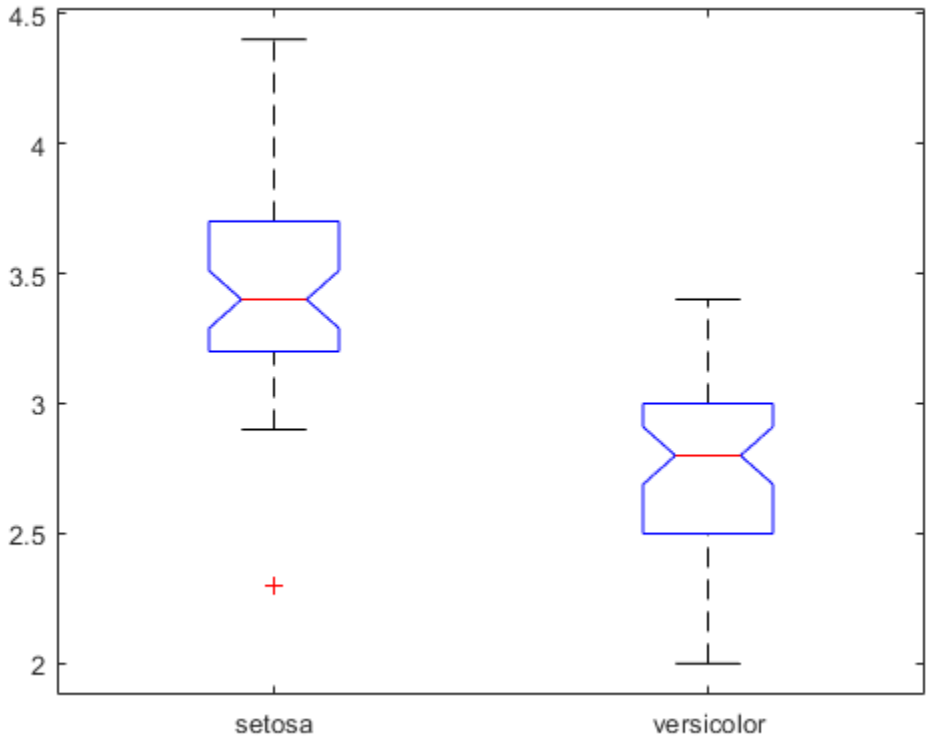
This example shows how to compare two groups of data by creating a notched box plot. Notches display the variability of the median between samples. The width of a notch is computed so that boxes whose notches do not overlap have different medians at the 5% significance level. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions. Comparing box plot medians is like a visual hypothesis test, analogous to the *t* test used for means. For more information on the different features of a box plot, see “Box Plot” on page 33-236.

Load the `fisheriris` data set. The data set contains length and width measurements from the sepals and petals of three species of iris flowers. Store the sepal width data for the setosa irises as `s1`, and the sepal width data for the versicolor irises as `s2`.

```
load fisheriris
s1 = meas(1:50,2);
s2 = meas(51:100,2);
```

Create a notched box plot using the sample data, and label each box with the name of the iris species it represents.

```
boxplot([s1 s2],'Notch','on', ...
        'Labels',{'setosa','versicolor'})
```

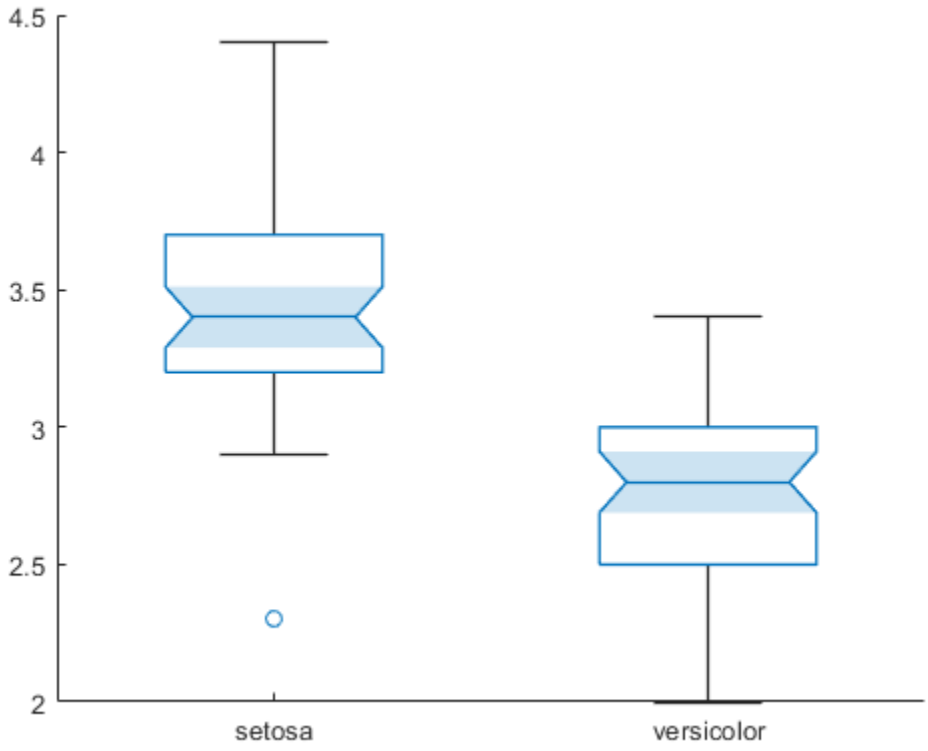


The notches of the two boxes do not overlap, which indicates that the median sepal widths of the setosa and versicolor irises are significantly different at the 5% significance level. Neither the red

median line in the setosa box nor the red median line in the versicolor box appears to be centered inside its box, which indicates that each sample is slightly skewed. Additionally, the setosa data contains one outlier value, while the versicolor data does not contain any outliers.

Instead of using the `boxplot` function, you can use the `boxchart` MATLAB® function to create box plots. Recreate the previous plot by using the `boxchart` function rather than `boxplot`.

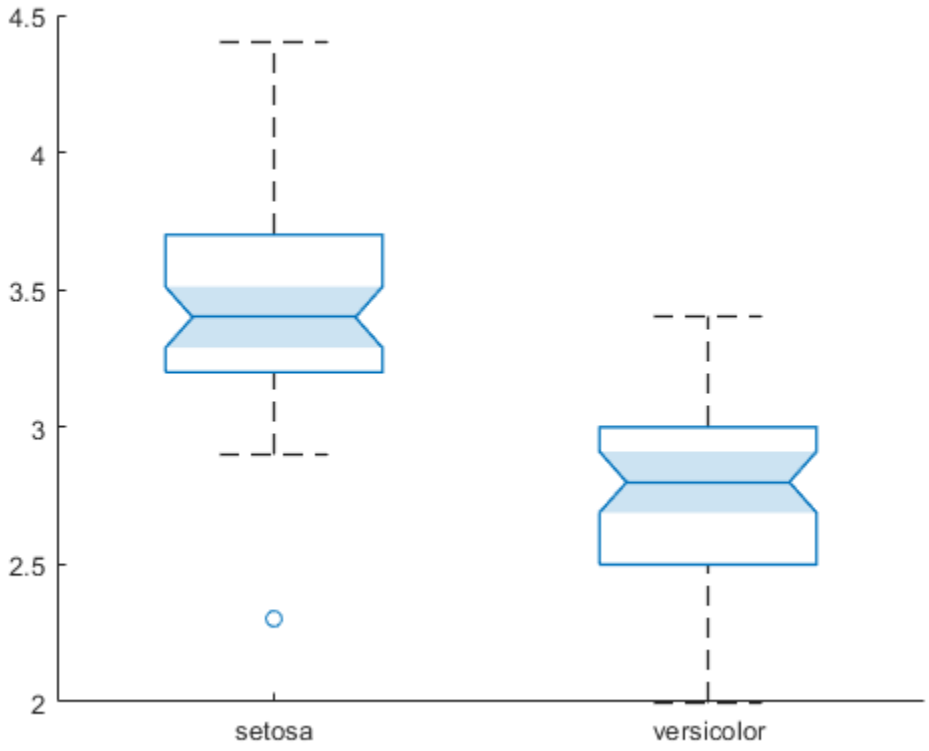
```
speciesName = categorical(species(1:100));
sepalWidth = meas(1:100,2);
b = boxchart(speciesName,sepalWidth,'Notch','on');
```



Each notch created by `boxchart` is a tapered, shaded region around the median line. The shading helps to better identify the notches.

One advantage of using `boxchart` is that the function creates a `BoxChart` object, whose properties you can change easily by using dot notation. For example, you can alter the style of the whiskers by specifying the `WhiskerLineStyle` property of the object `b`.

```
b.WhiskerLineStyle = '--';
```



For more information on the advantages of using `boxchart`, see “Alternative Functionality” on page 33-238.

See Also

`boxchart` | `boxplot` | `iqr` | `median`

More About

- “Exploratory Analysis of Data” on page 3-10
- “Measures of Central Tendency” on page 3-2
- “Measures of Dispersion” on page 3-4
- “Quantiles and Percentiles” on page 3-6
- “Distribution Plots” on page 4-7

Distribution Plots

In this section...

Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution. Use distribution plots in addition to more formal hypothesis tests to determine whether the sample data comes from a specified distribution. To learn about hypothesis tests, see “Hypothesis Testing” on page 8-5.

Statistics and Machine Learning Toolbox offers several distribution plot options:

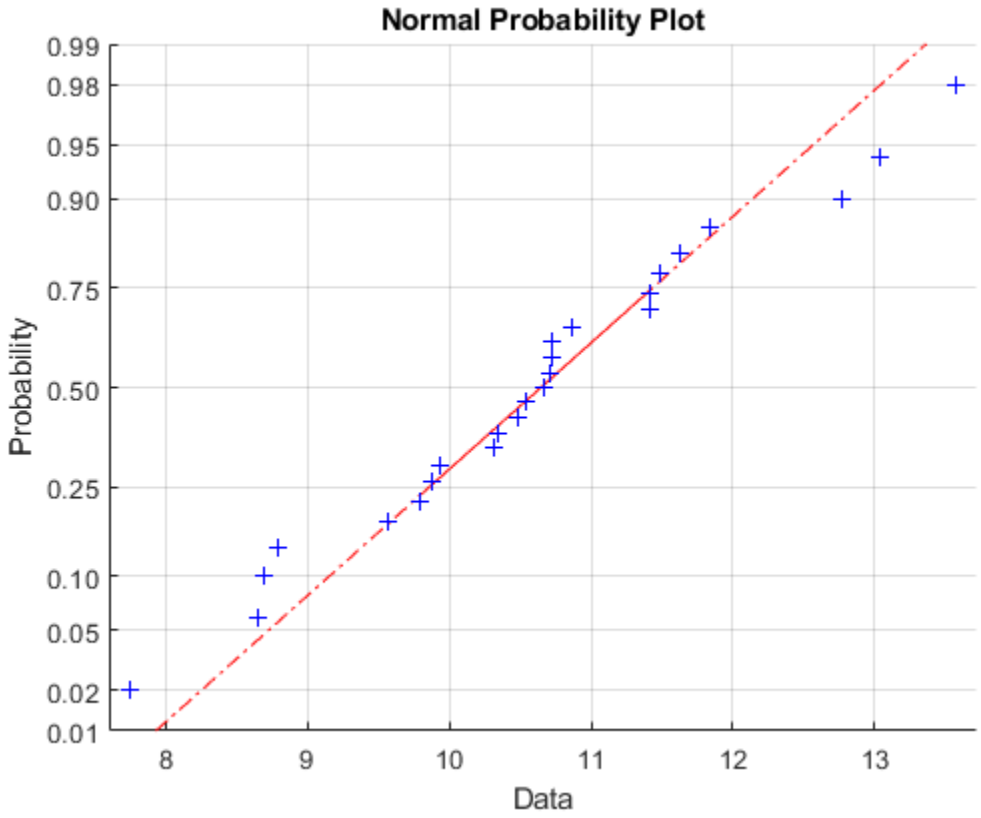
- “Normal Probability Plots” on page 4-7 — Use `normplot` to assess whether sample data comes from a normal distribution. Use `probplot` to create “Probability Plots” on page 4-9 for distributions other than normal, or to explore the distribution of censored data.
- “Quantile-Quantile Plots” on page 4-10 — Use `qqplot` to assess whether two sets of sample data come from the same distribution family. This plot is robust with respect to differences in location and scale.
- “Cumulative Distribution Plots” on page 4-12 — Use `cdfplot` or `ecdf` to display the empirical cumulative distribution function (cdf) of the sample data for visual comparison to the theoretical cdf of a specified distribution.

Normal Probability Plots

Use normal probability plots to assess whether data comes from a normal distribution. Many statistical procedures make the assumption that an underlying distribution is normal. Normal probability plots can provide some assurance to justify this assumption or provide a warning of problems with the assumption. An analysis of normality typically combines normal probability plots with hypothesis tests for normality.

This example generates a data sample of 25 random numbers from a normal distribution with mean 10 and standard deviation 1, and creates a normal probability plot of the data.

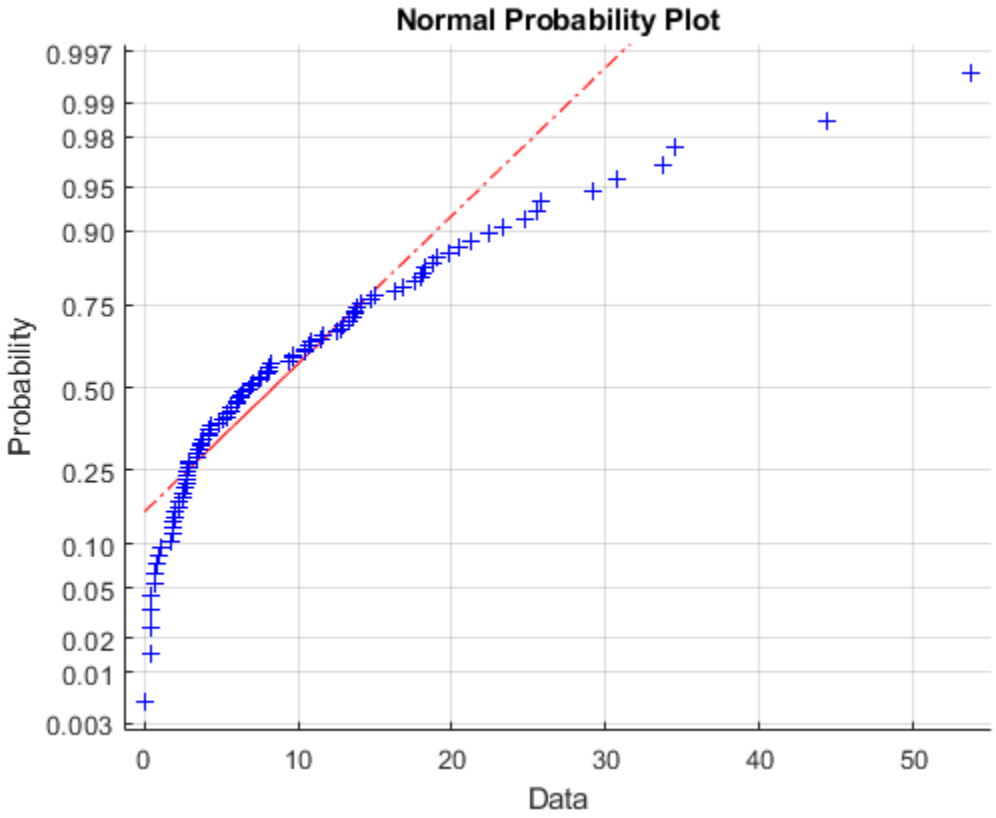
```
rng('default');  % For reproducibility
x = normrnd(10,1,[25,1]);
normplot(x)
```



The plus signs plot the empirical probability versus the data value for each point in the data. A solid line connects the 25th and 75th percentiles in the data, and a dashed line extends it to the ends of the data. The *y*-axis values are probabilities from zero to one, but the scale is not linear. The distance between tick marks on the *y*-axis matches the distance between the quantiles of a normal distribution. The quantiles are close together near the median (50th percentile) and stretch out symmetrically as you move away from the median.

In a normal probability plot, if all the data points fall near the line, an assumption of normality is reasonable. Otherwise, an assumption of normality is not justified. For example, the following generates a data sample of 100 random numbers from an exponential distribution with mean 10, and creates a normal probability plot of the data.

```
x = exprnd(10,100,1);
normplot(x)
```



The plot is strong evidence that the underlying distribution is not normal.

Probability Plots

A probability plot, like the normal probability plot, is just an empirical cdf plot scaled to a particular distribution. The *y*-axis values are probabilities from zero to one, but the scale is not linear. The distance between tick marks is the distance between quantiles of the distribution. In the plot, a line is drawn between the first and third quartiles in the data. If the data falls near the line, it is reasonable to choose the distribution as a model for the data. A distribution analysis typically combines probability plots with hypothesis tests for a particular distribution.

Create Weibull Probability Plot

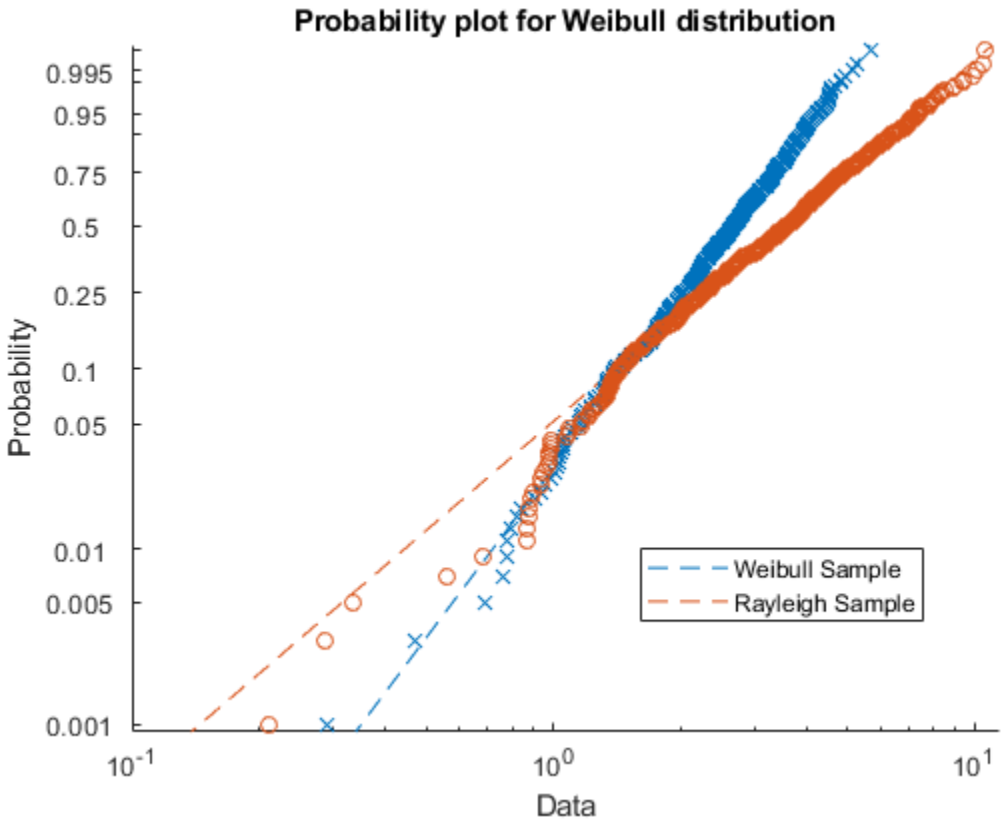
Generate sample data and create a probability plot.

Generate sample data. The sample `x1` contains 500 random numbers from a Weibull distribution with scale parameter `A = 3` and shape parameter `B = 3`. The sample `x2` contains 500 random numbers from a Rayleigh distribution with scale parameter `B = 3`.

```
rng('default');  % For reproducibility
x1 = wblrnd(3,3,[500,1]);
x2 = raylrnd(3,[500,1]);
```

Create a probability plot to assess whether the data in `x1` and `x2` comes from a Weibull distribution.

```
figure
probplot('weibull',[x1 x2])
legend('Weibull Sample','Rayleigh Sample','Location','best')
```



The probability plot shows that the data in `x1` comes from a Weibull distribution, while the data in `x2` does not.

Alternatively, you can use `wblplot` to create a Weibull probability plot.

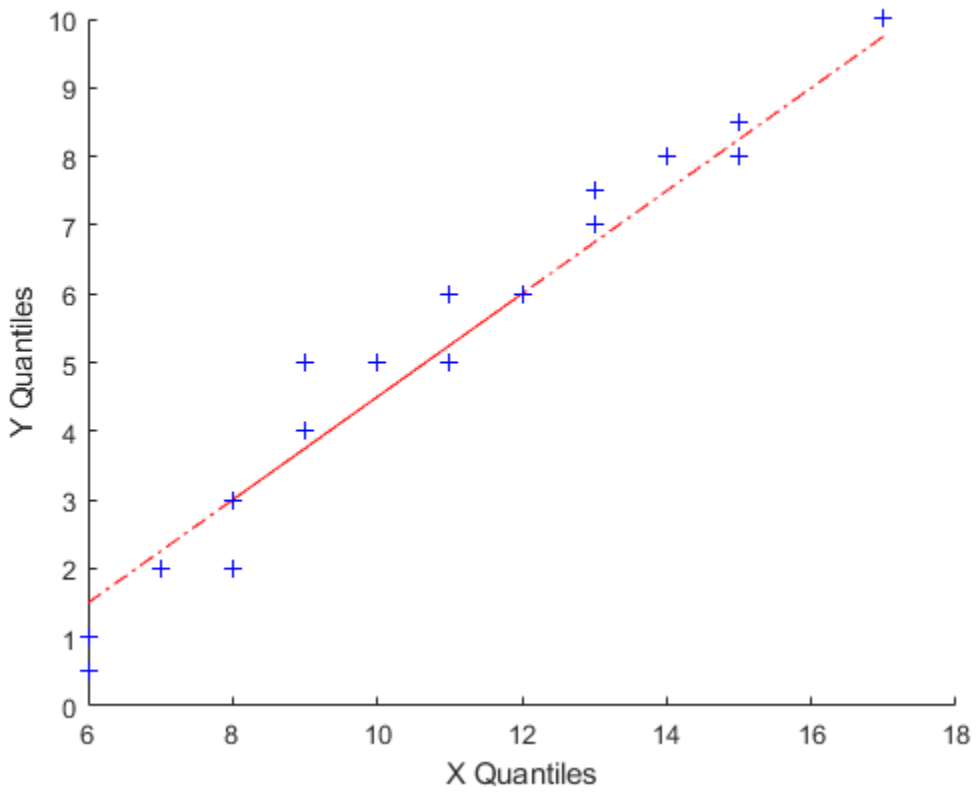
Quantile-Quantile Plots

Use quantile-quantile (q-q) plots to determine whether two samples come from the same distribution family. Q-Q plots are scatter plots of quantiles computed from each sample, with a line drawn between the first and third quartiles. If the data falls near the line, it is reasonable to assume that the two samples come from the same distribution. The method is robust with respect to changes in the location and scale of either distribution.

Create a quantile-quantile plot by using the `qqplot` function.

The following example generates two data samples containing random numbers from Poisson distributions with different parameter values, and creates a quantile-quantile plot. The data in `x` is from a Poisson distribution with mean 10, and the data in `y` is from a Poisson distribution with mean 5.

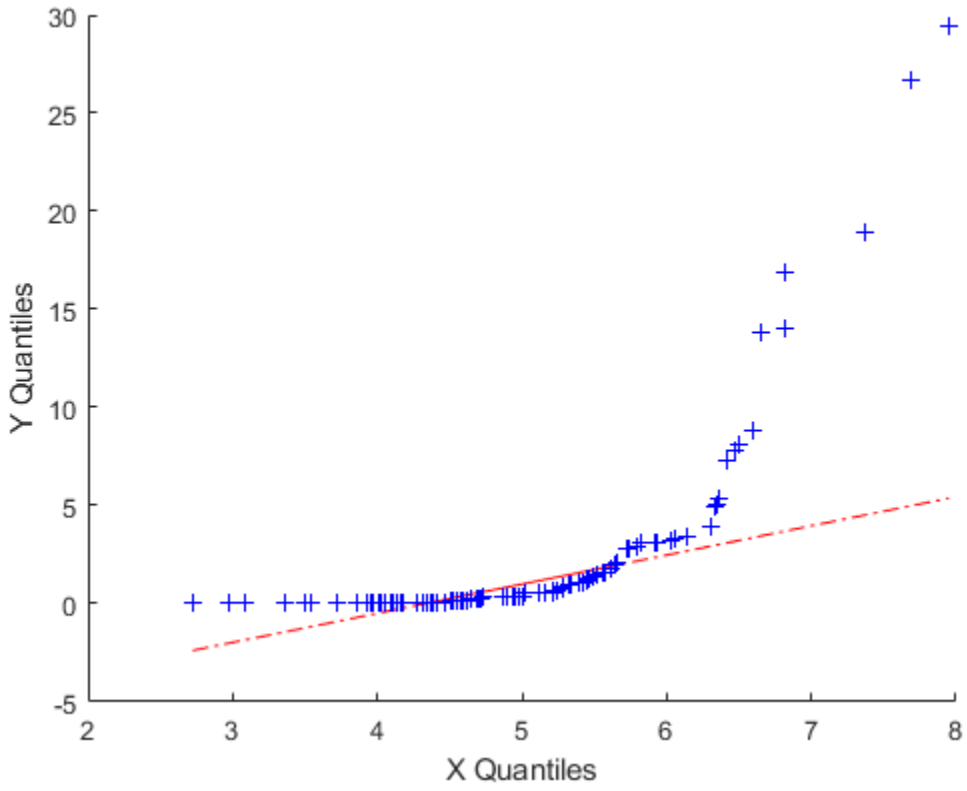
```
x = poissrnd(10,[50,1]);
y = poissrnd(5,[100,1]);
qqplot(x,y)
```



Even though the parameters and sample sizes are different, the approximate linear relationship suggests that the two samples may come from the same distribution family. As with normal probability plots, hypothesis tests can provide additional justification for such an assumption. For statistical procedures that depend on the two samples coming from the same distribution, however, a linear quantile-quantile plot is often sufficient.

The following example shows what happens when the underlying distributions are not the same. Here, `x` contains 100 random numbers generated from a normal distribution with mean 5 and standard deviation 1, while `y` contains 100 random numbers generated from a Weibull distribution with a scale parameter of 2 and a shape parameter of 0.5.

```
x = normrnd(5,1,[100,1]);
y = wblrnd(2,0.5,[100,1]);
qqplot(x,y)
```



The plots indicate that these samples clearly are not from the same distribution family.

Cumulative Distribution Plots

An empirical cumulative distribution function (cdf) plot shows the proportion of data less than or equal to each *x* value, as a function of *x*. The scale on the *y*-axis is linear; in particular, it is not scaled to any particular distribution. Empirical cdf plots are used to compare data cdfs to cdfs for particular distributions.

To create an empirical cdf plot, use the `cdfplot` function or the `ecdf` function.

Compare Empirical cdf to Theoretical cdf

Plot the empirical cdf of a sample data set and compare it to the theoretical cdf of the underlying distribution of the sample data set. In practice, a theoretical cdf can be unknown.

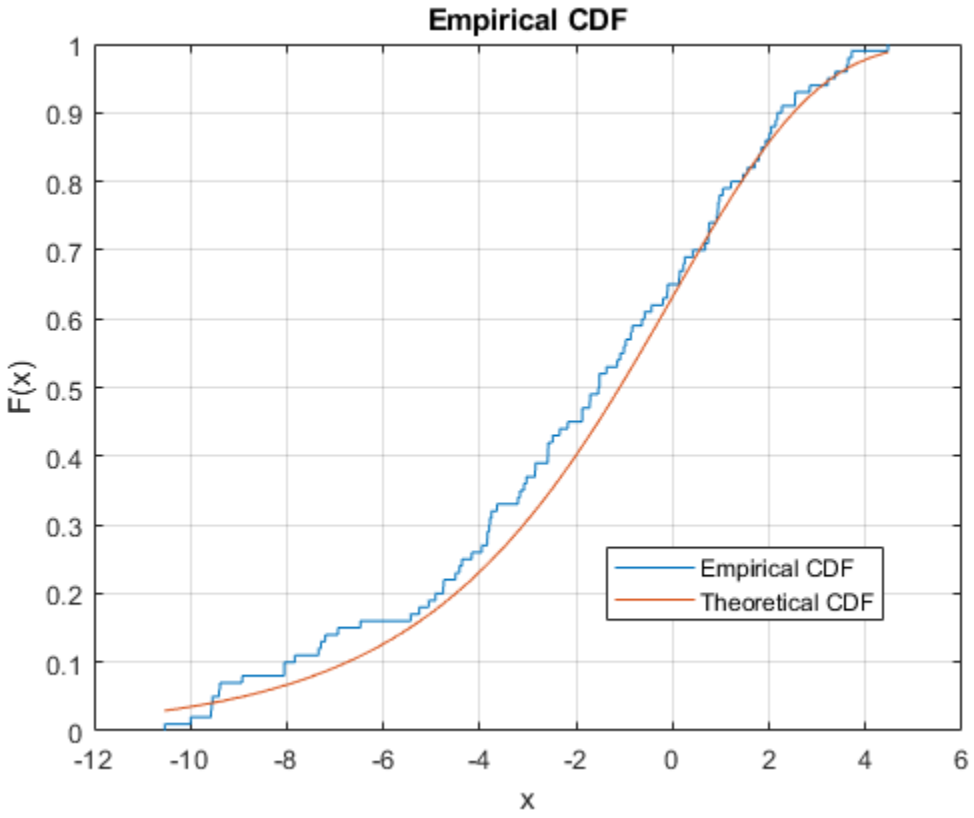
Generate a random sample data set from the extreme value distribution with a location parameter of 0 and a scale parameter of 3.

```
rng('default')  % For reproducibility
y = evrnd(0,3,100,1);
```

Plot the empirical cdf of the sample data set and the theoretical cdf on the same figure.

```
cdfplot(y)
hold on
```

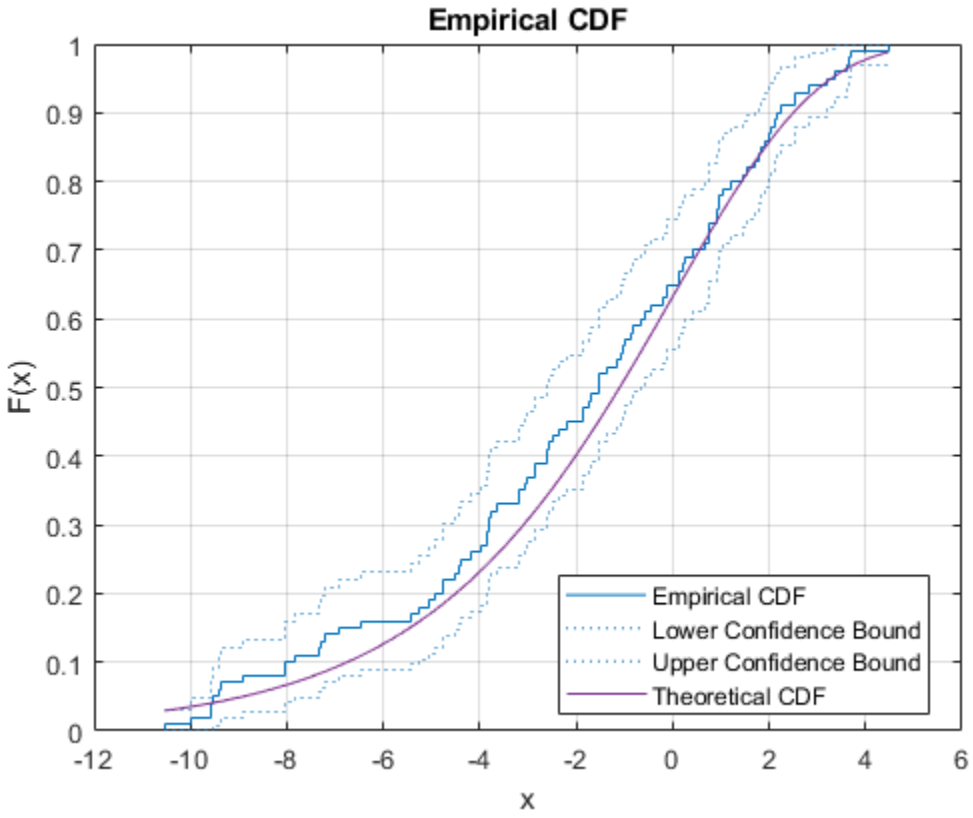
```
x = linspace(min(y),max(y));
plot(x,evcdf(x,0,3))
legend('Empirical CDF','Theoretical CDF','Location','best')
hold off
```



The plot shows the similarity between the empirical cdf and the theoretical cdf.

Alternatively, you can use the `ecdf` function. The `ecdf` function also plots the 95% confidence intervals estimated by using Greenwood's Formula. For details, see “Greenwood’s Formula” on page 33-1279.

```
ecdf(y,'Bounds','on')
hold on
plot(x,evcdf(x,0,3))
grid on
title('Empirical CDF')
legend('Empirical CDF','Lower Confidence Bound','Upper Confidence Bound','Theoretical CDF','Loca
hold off
```



See Also

`cdfplot` | `ecdf` | `normplot` | `probplot` | `qqplot` | `wblplot`

More About

- “Compare Grouped Data Using Box Plots” on page 4-4
- “Hypothesis Testing” on page 8-5

Visualizing Multivariate Data

This example shows how to visualize multivariate data using various statistical plots. Many statistical analyses involve only two variables: a predictor variable and a response variable. Such data are easy to visualize using 2D scatter plots, bivariate histograms, boxplots, etc. It's also possible to visualize trivariate data with 3D scatter plots, or 2D scatter plots with a third variable encoded with, for example color. However, many datasets involve a larger number of variables, making direct visualization more difficult. This example explores some of the ways to visualize high-dimensional data in MATLAB®, using Statistics and Machine Learning Toolbox™.

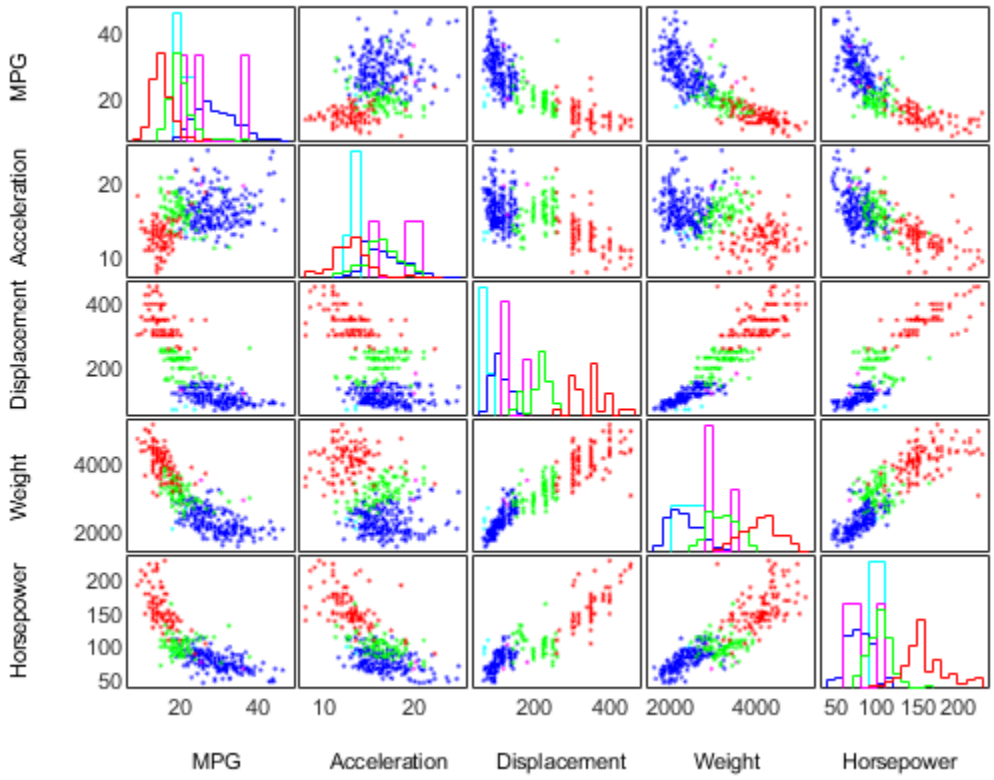
In this example, we'll use the `carbig` dataset, a dataset that contains various measured variables for about 400 automobiles from the 1970's and 1980's. We'll illustrate multivariate visualization using the values for fuel efficiency (in miles per gallon, MPG), acceleration (time from 0-60MPH in sec), engine displacement (in cubic inches), weight, and horsepower. We'll use the number of cylinders to group observations.

```
load carbig
X = [MPG,Acceleration,Displacement,Weight,Horsepower];
varNames = {'MPG'; 'Acceleration'; 'Displacement'; 'Weight'; 'Horsepower'};
```

Scatter Plot Matrices

Viewing slices through lower dimensional subspaces is one way to partially work around the limitation of two or three dimensions. For example, we can use the `gplotmatrix` function to display an array of all the bivariate scatter plots between our five variables, along with a univariate histogram for each variable.

```
figure
gplotmatrix(X,[],Cylinders,['c' 'b' 'm' 'g' 'r'],[],[],false);
text([.08 .24 .43 .66 .83], repmat(-.1,1,5), varNames, 'FontSize',8);
text(repmat(-.12,1,5), [.86 .62 .41 .25 .02], varNames, 'FontSize',8, 'Rotation',90);
```

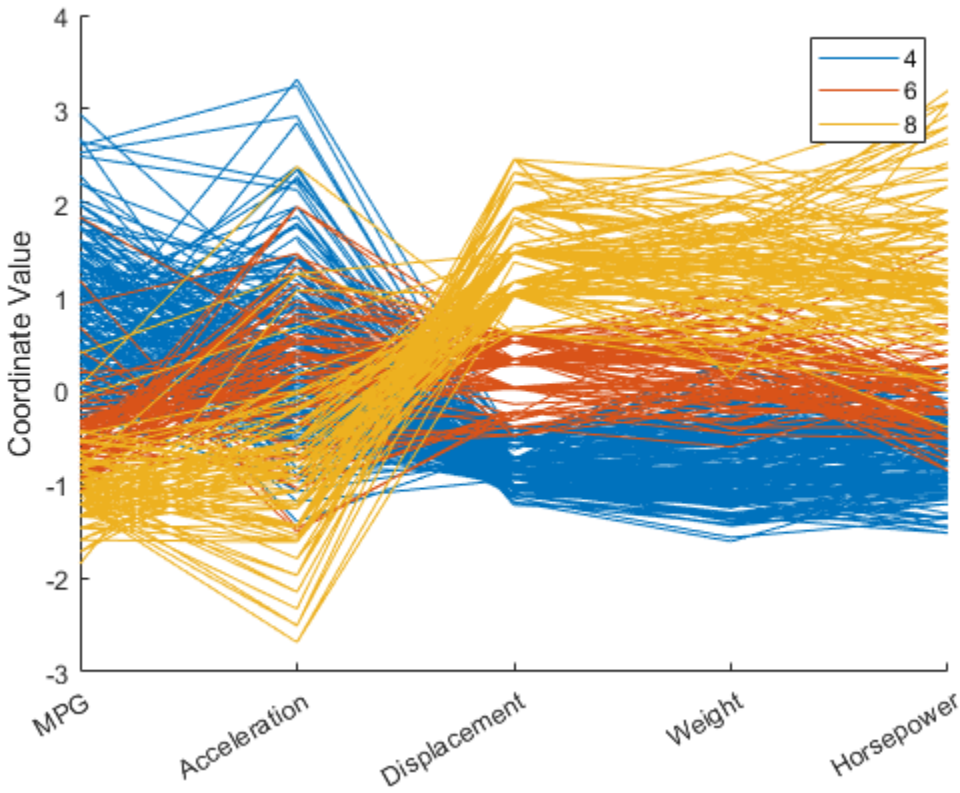


The points in each scatter plot are color-coded by the number of cylinders: blue for 4 cylinders, green for 6, and red for 8. There is also a handful of 5 cylinder cars, and rotary-engined cars are listed as having 3 cylinders. This array of plots makes it easy to pick out patterns in the relationships between pairs of variables. However, there may be important patterns in higher dimensions, and those are not easy to recognize in this plot.

Parallel Coordinates Plots

The scatter plot matrix only displays bivariate relationships. However, there are other alternatives that display all the variables together, allowing you to investigate higher-dimensional relationships among variables. The most straight-forward multivariate plot is the parallel coordinates plot. In this plot, the coordinate axes are all laid out horizontally, instead of using orthogonal axes as in the usual Cartesian graph. Each observation is represented in the plot as a series of connected line segments. For example, we can make a plot of all the cars with 4, 6, or 8 cylinders, and color observations by group.

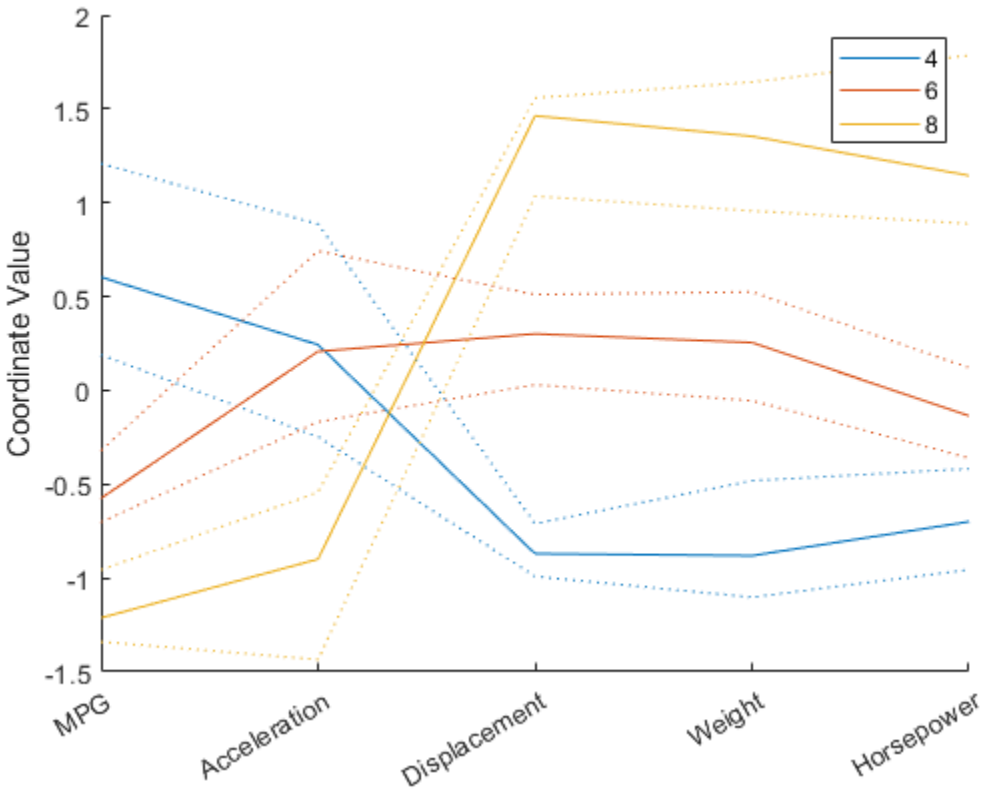
```
Cyl468 = ismember(Cylinders,[4 6 8]);
parallelcoords(X(Cyl468,:), 'group',Cylinders(Cyl468), ...
               'standardize','on', 'labels',varNames)
```



The horizontal direction in this plot represents the coordinate axes, and the vertical direction represents the data. Each observation consists of measurements on five variables, and each measurement is represented as the height at which the corresponding line crosses each coordinate axis. Because the five variables have widely different ranges, this plot was made with standardized values, where each variable has been standardized to have zero mean and unit variance. With the color coding, the graph shows, for example, that 8 cylinder cars typically have low values for MPG and acceleration, and high values for displacement, weight, and horsepower.

Even with color coding by group, a parallel coordinates plot with a large number of observations can be difficult to read. We can also make a parallel coordinates plot where only the median and quartiles (25% and 75% points) for each group are shown. This makes the typical differences and similarities among groups easier to distinguish. On the other hand, it may be the outliers for each group that are most interesting, and this plot does not show them at all.

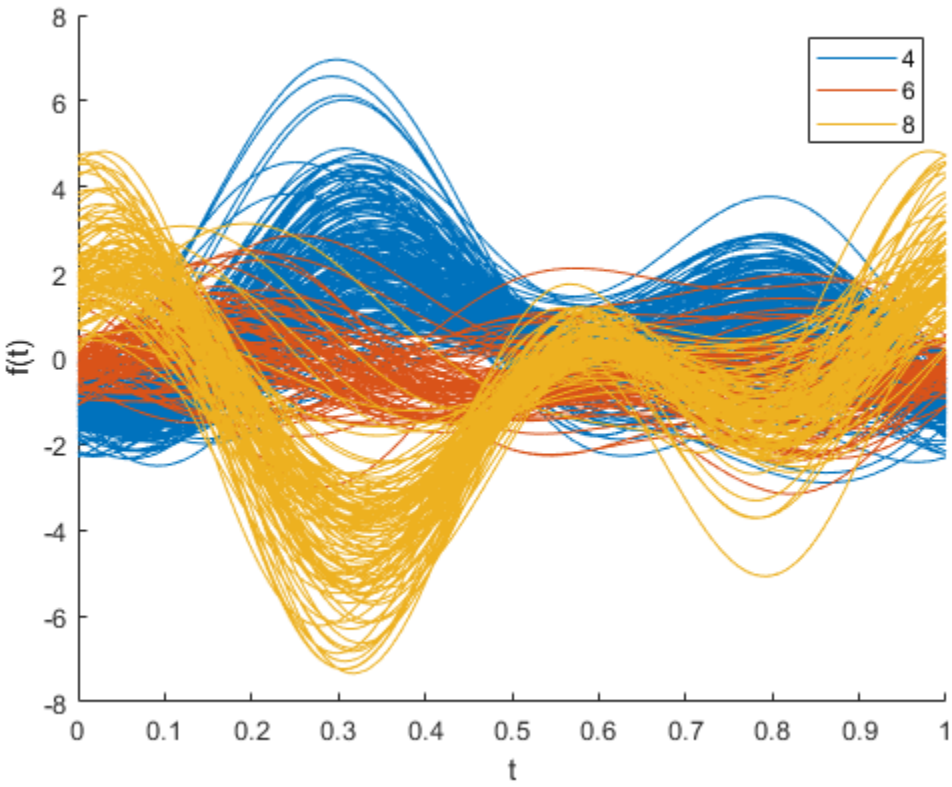
```
parallelcoords(X(Cyl468,:), 'group',Cylinders(Cyl468), ...
               'standardize','on', 'labels',varNames, 'quantile',.25)
```



Andrews Plots

Another similar type of multivariate visualization is the Andrews plot. This plot represents each observation as a smooth function over the interval [0,1].

```
andrewsplot(X(Cyl468,:), 'group',Cylinders(Cyl468), 'standardize','on')
```



Each function is a Fourier series, with coefficients equal to the corresponding observation's values. In this example, the series has five terms: a constant, two sine terms with periods 1 and 1/2, and two similar cosine terms. Effects on the functions' shapes due to the three leading terms are the most apparent in an Andrews plot, so patterns in the first three variables tend to be the ones most easily recognized.

There's a distinct difference between groups at t = 0, indicating that the first variable, MPG, is one of the distinguishing features between 4, 6, and 8 cylinder cars. More interesting is the difference between the three groups at around t = 1/3. Plugging this value into the formula for the Andrews plot functions, we get a set of coefficients that define a linear combination of the variables that distinguishes between groups.

```
t1 = 1/3;
[1/sqrt(2) sin(2*pi*t1) cos(2*pi*t1) sin(4*pi*t1) cos(4*pi*t1)]
```

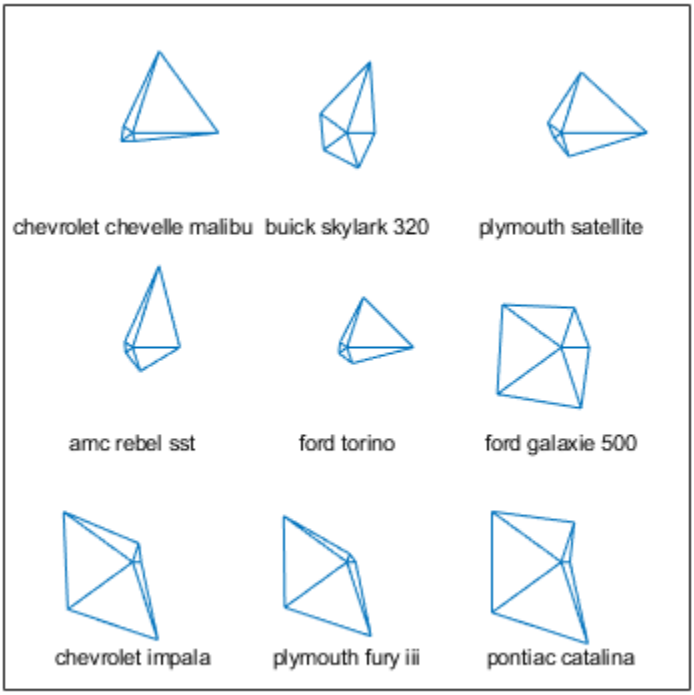
```
ans =

    0.7071    0.8660   -0.5000   -0.8660   -0.5000
```

From these coefficients, we can see that one way to distinguish 4 cylinder cars from 8 cylinder cars is that the former have higher values of MPG and acceleration, and lower values of displacement, horsepower, and particularly weight, while the latter have the opposite. That's the same conclusion we drew from the parallel coordinates plot.

Glyph Plots

Another way to visualize multivariate data is to use "glyphs" to represent the dimensions. The function `glyphplot` supports two types of glyphs: stars, and Chernoff faces. For example, here is a star plot of the first 9 models in the car data. Each spoke in a star represents one variable, and the spoke length is proportional to the value of that variable for that observation.

```
h = glyphplot(X(1:9,:), 'glyph','star', 'varLabels',varNames, 'obslabels',Model(1:9,:));
set(h(:,3),'FontSize',8);
```



In a live MATLAB figure window, this plot would allow interactive exploration of the data values, using data cursors. For example, clicking on the right-hand point of the star for the Ford Torino would show that it has an MPG value of 17.

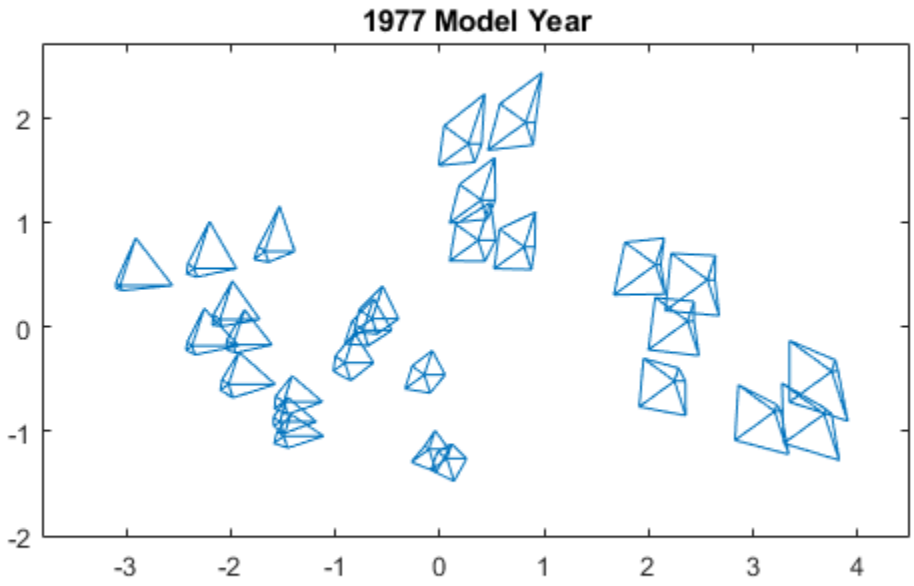
Glyph Plots and Multidimensional Scaling

Plotting stars on a grid, with no particular order, can lead to a figure that is confusing, because adjacent stars can end up quite different-looking. Thus, there may be no smooth pattern for the eye to catch. It's often useful to combine multidimensional scaling (MDS) with a glyph plot. To illustrate, we'll first select all cars from 1977, and use the `zscore` function to standardize each of the five variables to have zero mean and unit variance. Then we'll compute the Euclidean distances among those standardized observations as a measure of dissimilarity. This choice might be too simplistic in a real application, but serves here for purposes of illustration.

```
models77 = find((Model_Year==77));
dissimilarity = pdist(zscore(X(models77,:)));
```

Finally, we use `mdscale` to create a set of locations in two dimensions whose interpoint distances approximate the dissimilarities among the original high-dimensional data, and plot the glyphs using those locations. The distances in this 2D plot may only roughly reproduce the data, but for this type of plot, that's good enough.

```
Y = mdscale(dissimilarity,2);
glyphplot(X(models77,:), 'glyph','star', 'centers',Y, ...
          'varLabels',varNames, 'obslabels',Model(models77,:), 'radius',.5);
title('1977 Model Year');
```

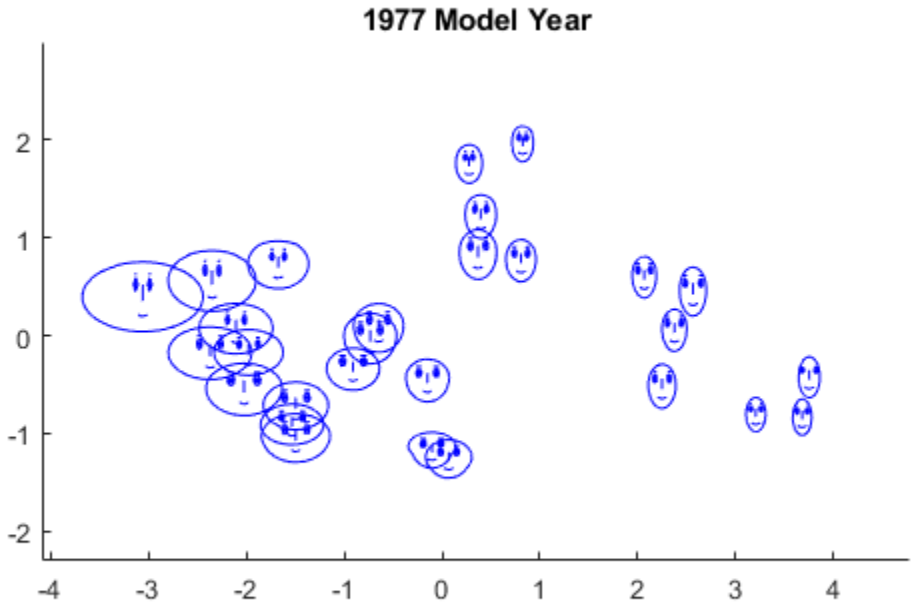


In this plot, we've used MDS as dimension reduction method, to create a 2D plot. Normally that would mean a loss of information, but by plotting the glyphs, we have incorporated all of the high-dimensional information in the data. The purpose of using MDS is to impose some regularity to the variation in the data, so that patterns among the glyphs are easier to see.

Just as with the previous plot, interactive exploration would be possible in a live figure window.

Another type of glyph is the Chernoff face. This glyph encodes the data values for each observation into facial features, such as the size of the face, the shape of the face, position of the eyes, etc.

```
glyphplot(X(models77,:), 'glyph','face', 'centers',Y, ...
          'varLabels',varNames, 'obslabels',Model(models77,:));
title('1977 Model Year');
```



Here, the two most apparent features, face size and relative forehead/jaw size, encode MPG and acceleration, while the forehead and jaw shape encode displacement and weight. Width between eyes encodes horsepower. It's notable that there are few faces with wide foreheads and narrow jaws, or vice-versa, indicating positive linear correlation between the variables displacement and weight. That's also what we saw in the scatter plot matrix.

The correspondence of features to variables determines what relationships are easiest to see, and `glyphplot` allows the choice to be changed easily.

```
close
```

Probability Distributions

Working with Probability Distributions

In this section...

Probability distributions are theoretical distributions based on assumptions about a source population. The distributions assign probability to the event that a random variable has a specific, discrete value, or falls within a specified range of continuous values.

Statistics and Machine Learning Toolbox offers several ways to work with probability distributions.

- Use “Probability Distribution Objects” on page 5-3 to fit a probability distribution object to sample data, or to create a probability distribution object with specified parameter values.
- Use “Probability Distribution Functions” on page 5-6 to work with data input from matrices.
- Use “Probability Distribution Apps and User Interfaces” on page 5-10 to interactively fit, explore, and generate random numbers from probability distributions. Available apps and user interfaces include:
  - The **Distribution Fitter** app
  - The **Probability Distribution Function** user interface
  - The Random Number Generation user interface (`randtool`)

For a list of distributions supported by Statistics and Machine Learning Toolbox, see “Supported Distributions” on page 5-14.

You can define a probability object for a custom distribution and then use the Distribution Fitter app or probability object functions, such as `pdf`, `cdf`, `icdf`, and `random`, to evaluate the distribution, generate random numbers, and so on. For details, see “Define Custom Distributions Using the Distribution Fitter App” on page 5-81. You can also define a custom distribution using a function handle and use the `mle` function to find maximum likelihood estimates. For an example, see “Fit Custom Distribution to Censored Data” on page 33-3970.

Probability Distribution Objects

Probability distribution objects allow you to fit a probability distribution to sample data, or define a distribution by specifying parameter values. You can then perform a variety of analyses on the distribution object.

Create Probability Distribution Objects

Estimate probability distribution parameters from sample data by fitting a probability distribution object to the data using `fitdist`. You can fit a single specified parametric or nonparametric distribution to the sample data. You can also fit multiple distributions of the same type to the sample data based on grouping variables. For most distributions, `fitdist` uses maximum likelihood estimation (MLE) to estimate the distribution parameters from the sample data. For more information and additional syntax options, see `fitdist`.

Alternatively, you can create a probability distribution object with specified parameter values using `makedist`.

Work with Probability Distribution Objects

Once you create a probability distribution object, you can use object functions to:

- Compute confidence intervals for the distribution parameters (`paramci`).
- Compute summary statistics, including mean (`mean`), median (`median`), interquartile range (`iqr`), variance (`var`), and standard deviation (`std`).
- Evaluate the probability density function (`pdf`).
- Evaluate the cumulative distribution function (`cdf`) or the inverse cumulative distribution function (`icdf`).
- Compute the negative loglikelihood (`negloglik`) and profile likelihood function (`proflik`) for the distribution.
- Generate random numbers from the distribution (`random`).
- Truncate the distribution to specified lower and upper limits (`truncate`).

Save a Probability Distribution Object

To save your probability distribution object to a .MAT file:

- In the toolbar, click **Save Workspace**. This option saves all of the variables in your workspace, including any probability distribution objects.
- In the workspace browser, right-click the probability distribution object and select **Save as**. This option saves only the selected probability distribution object, not the other variables in your workspace.

Alternatively, you can save a probability distribution object directly from the command line by using the `save` function. `save` enables you to choose a file name and specify the probability distribution object you want to save. If you do not specify an object (or other variable), MATLAB saves all of the variables in your workspace, including any probability distribution objects, to the specified file name. For more information and additional syntax options, see `save`.

Analyze Distribution Using Probability Distribution Objects

This example shows how to use probability distribution objects to perform a multistep analysis on a fitted distribution.

The analysis illustrates how to:

- Fit a probability distribution to sample data that contains exam grades of 120 students by using `fitdist`.
- Compute the mean of the exam grades by using `mean`.
- Plot a histogram of the exam grade data, overlaid with a plot of the pdf of the fitted distribution, by using `plot` and `pdf`.
- Compute the boundary for the top 10 percent of student grades by using `icdf`.
- Save the fitted probability distribution object by using `save`.

Load the sample data.

```
load examgrades
```

The sample data contains a 120-by-5 matrix of exam grades. The exams are scored on a scale of 0 to 100.

Create a vector containing the first column of exam grade data.

```
x = grades(:,1);
```

Fit a normal distribution to the sample data by using `fitdist` to create a probability distribution object.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

`fitdist` returns a probability distribution object, `pd`, of the type `NormalDistribution`. This object contains the estimated parameter values, `mu` and `sigma`, for the fitted normal distribution. The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

Compute the mean of the students' exam grades using the fitted distribution object, `pd`.

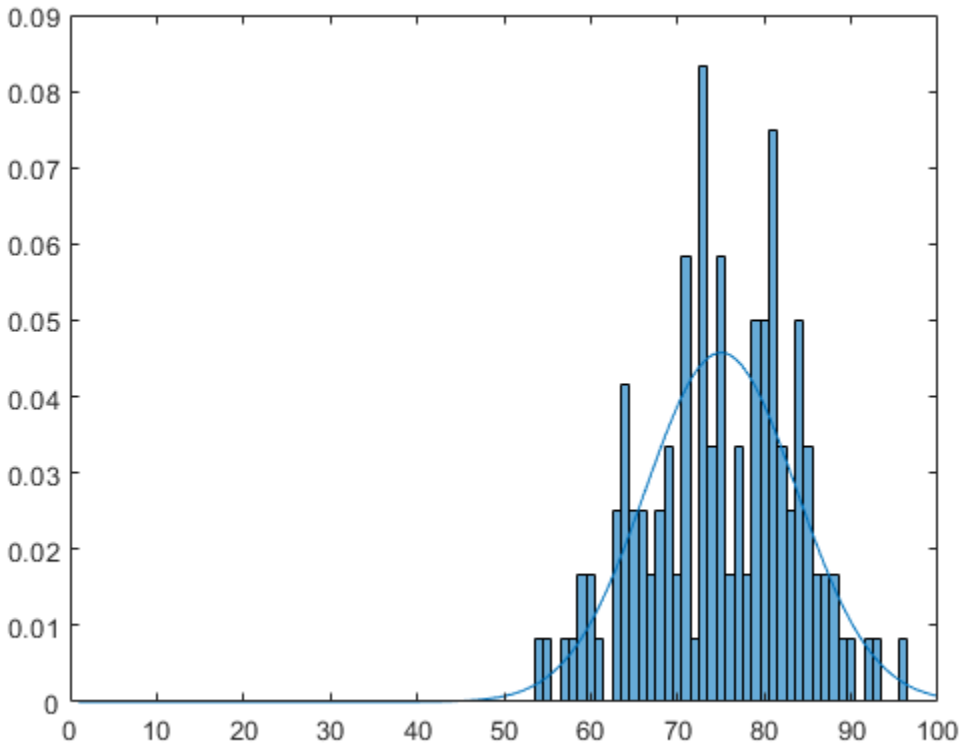
```
m = mean(pd)
```

```
m = 75.0083
```

The mean of the exam grades is equal to the `mu` parameter estimated by `fitdist`.

Plot a histogram of the exam grades. Overlay a plot of the fitted pdf to visually compare the fitted normal distribution with the actual exam grades.

```
x_pdf = [1:0.1:100];
y = pdf(pd,x_pdf);

figure
histogram(x,'Normalization','pdf')
line(x_pdf,y)
```



The pdf of the fitted distribution follows the same shape as the histogram of the exam grades.

Determine the boundary for the upper 10 percent of student exam grades by using the inverse cumulative distribution function (`icdf`). This boundary is equivalent to the value at which the cdf of the probability distribution is equal to 0.9. In other words, 90 percent of the exam grades are less than or equal to the boundary value.

```
A = icdf(pd,0.9)
```

```
A = 86.1837
```

Based on the fitted distribution, 10 percent of students received an exam grade greater than 86.1837. Equivalently, 90 percent of students received an exam grade less than or equal to 86.1837.

Save the fitted probability distribution, `pd`, as a file named `myobject.mat`.

```
save('myobject.mat','pd')
```

Probability Distribution Functions

You can also work with probability distributions using distribution-specific functions. These functions are useful for generating random numbers, computing summary statistics inside a loop or script, and passing a cdf or pdf as a function handle to another function. You can also use these functions to perform computations on arrays of parameter values rather than a single set of parameters. For a list of supported probability distributions, see “Supported Distributions” on page 5-14.

Analyze Distribution Using Distribution-Specific Functions

This example shows how to use distribution-specific functions to perform a multistep analysis on a fitted distribution.

The analysis illustrates how to:

- Fit a probability distribution to sample data that contains exam grades of 120 students by using `normfit`.
- Plot a histogram of the exam grade data, overlaid with a plot of the pdf of the fitted distribution, by using `plot` and `normpdf`.
- Compute the boundary for the top 10 percent of student grades by using `norminv`.
- Save the estimated distribution parameters by using `save`.

You can perform the same analysis using a probability distribution object. See “Analyze Distribution Using Probability Distribution Objects” on page 5-4.

Load the sample data.

```
load examgrades
```

The sample data contains a 120-by-5 matrix of exam grades. The exams are scored on a scale of 0 to 100.

Create a vector containing the first column of exam grade data.

```
x = grades(:,1);
```

Fit a normal distribution to the sample data by using `normfit`.

```
[mu,sigma,muCI,sigmaCI] = normfit(x)
```

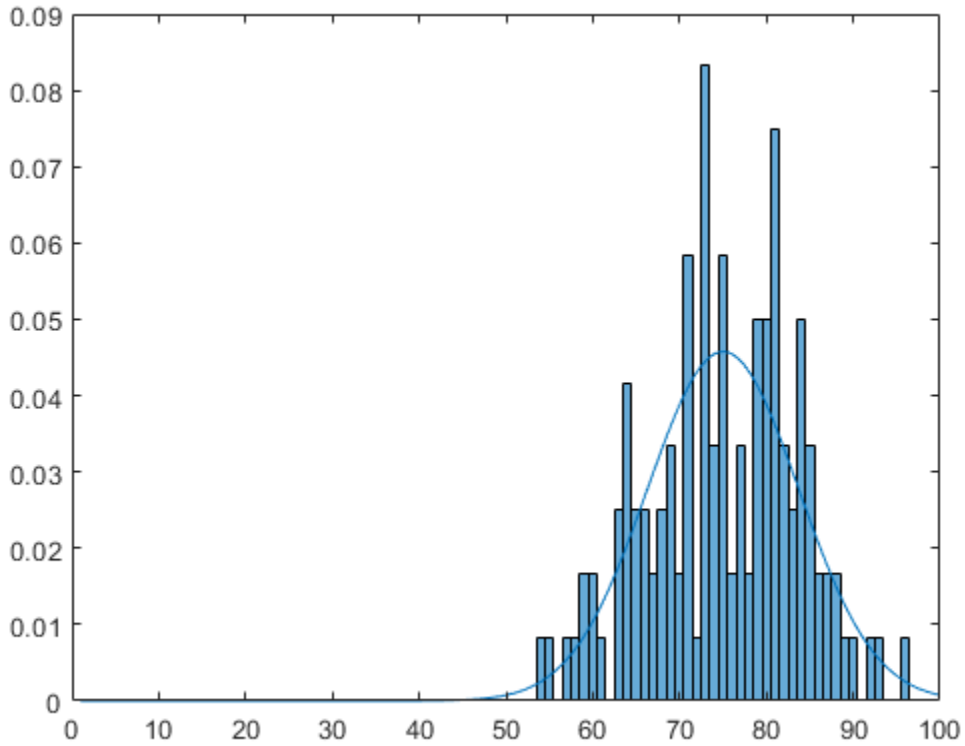
```
mu = 75.0083

sigma = 8.7202

muCI = 2×1

   73.4321
   76.5846


sigmaCI = 2×1

    7.7391
    9.9884
```

The `normfit` function returns the estimates of normal distribution parameters and the 95% confidence intervals for the parameter estimates.

Plot a histogram of the exam grades. Overlay a plot of the fitted pdf to visually compare the fitted normal distribution with the actual exam grades.

```
x_pdf = [1:0.1:100];
y = normpdf(x_pdf,mu,sigma);
```

```
figure
histogram(x,'Normalization','pdf')
line(x_pdf,y)
```



The pdf of the fitted distribution follows the same shape as the histogram of the exam grades.

Determine the boundary for the upper 10 percent of student exam grades by using the normal inverse cumulative distribution function. This boundary is equivalent to the value at which the cdf of the probability distribution is equal to 0.9. In other words, 90 percent of the exam grades are less than or equal to the boundary value.

```
A = norminv(0.9,mu,sigma)
```

```
A = 86.1837
```

Based on the fitted distribution, 10 percent of students received an exam grade greater than 86.1837. Equivalently, 90 percent of students received an exam grade less than or equal to 86.1837.

Save the estimated distribution parameters as a file named `myparameter.mat`.

```
save('myparameter.mat','mu','sigma')
```

Use Probability Distribution Functions as Function Handle

This example shows how to use the probability distribution function `normcdf` as a function handle in the chi-square goodness of fit test (`chi2gof`).

This example tests the null hypothesis that the sample data contained in the input vector, `x`, comes from a normal distribution with parameters μ and σ equal to the mean (`mean`) and standard deviation (`std`) of the sample data, respectively.

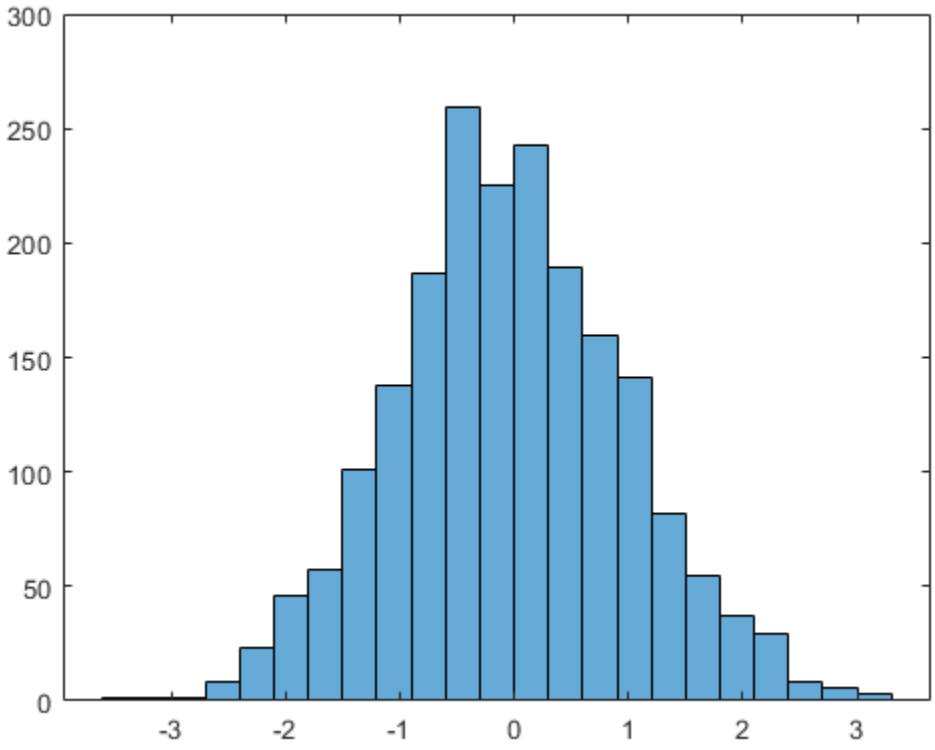
```
rng('default') % For reproducibility
x = normrnd(50,5,100,1);
h = chi2gof(x,'cdf',{@normcdf,mean(x),std(x)})
```

```
h = 0
```

The returned result `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the default 5% significance level.

This next example illustrates how to use probability distribution functions as a function handle in the slice sampler (`slicesample`). The example uses `normpdf` to generate a random sample of 2,000 values from a standard normal distribution, and plots a histogram of the resulting values.

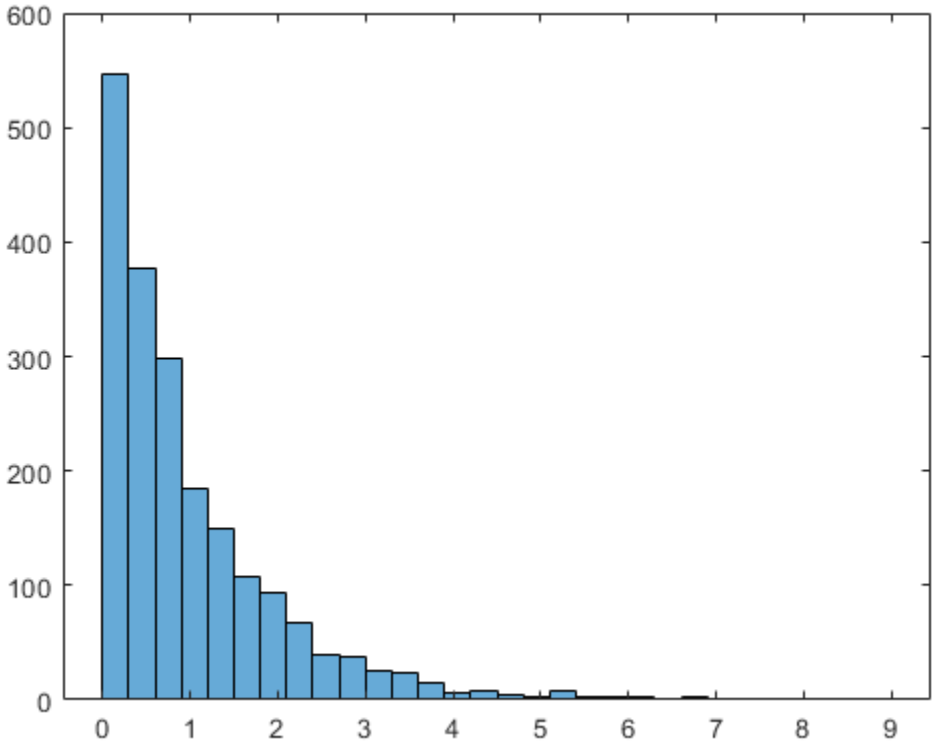
```
rng('default') % For reproducibility
x = slicesample(1,2000,'pdf',@normpdf,'thin',5,'burnin',1000);
histogram(x)
```



The histogram shows that, when using `normpdf`, the resulting random sample has a standard normal distribution.

If you pass the probability distribution function for the exponential distribution pdf (`exppdf`) as a function handle instead of `normpdf`, then `slicesample` generates the 2,000 random samples from an exponential distribution with a default parameter value of μ equal to 1.

```
rng('default') % For reproducibility
x = slicesample(1,2000,'pdf',@exppdf,'thin',5,'burnin',1000);
histogram(x)
```



The histogram shows that the resulting random sample when using `exppdf` has an exponential distribution.

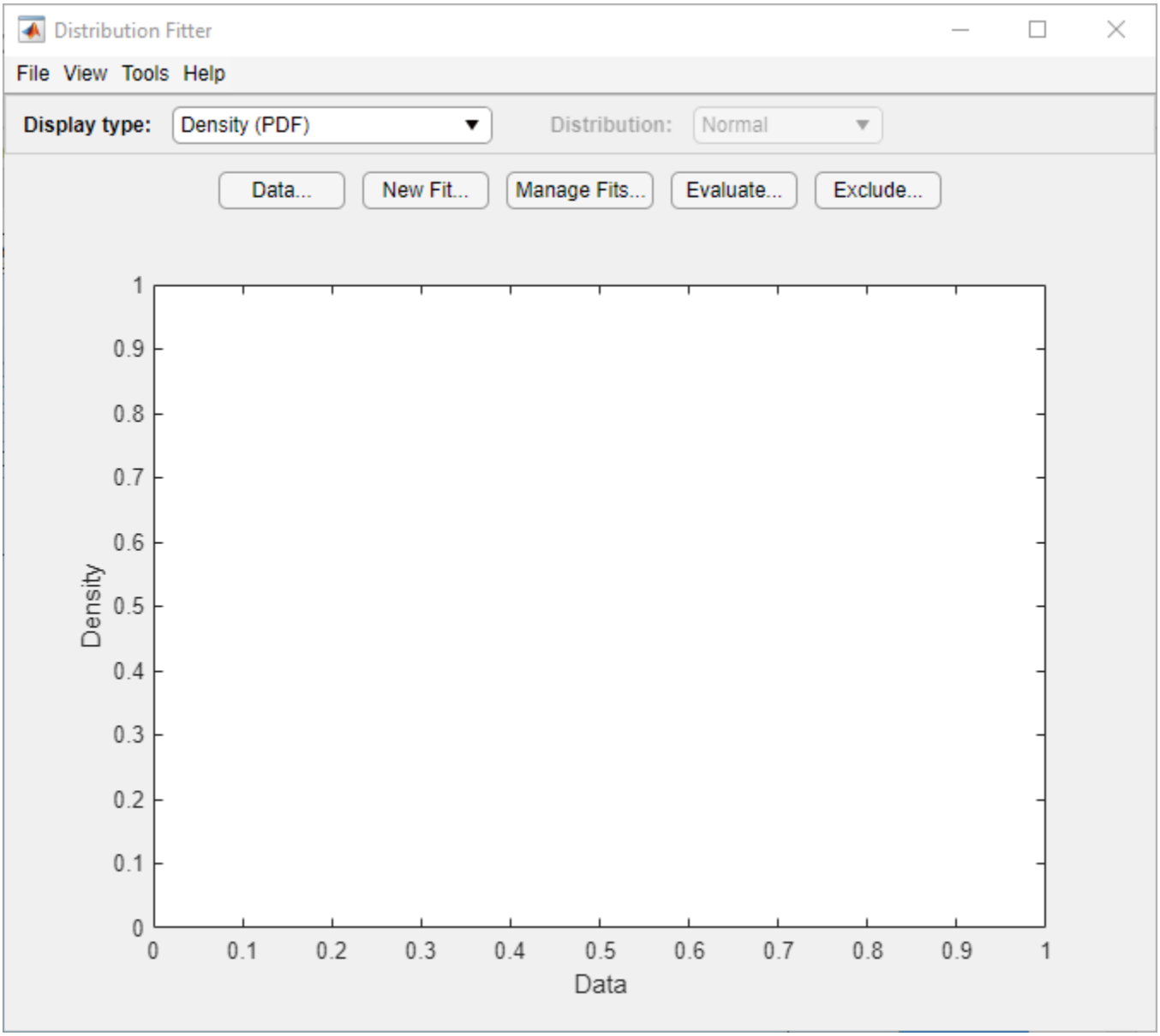
Probability Distribution Apps and User Interfaces

Apps and user interfaces provide an interactive approach to working with parametric and nonparametric probability distributions.

Distribution Fitter App

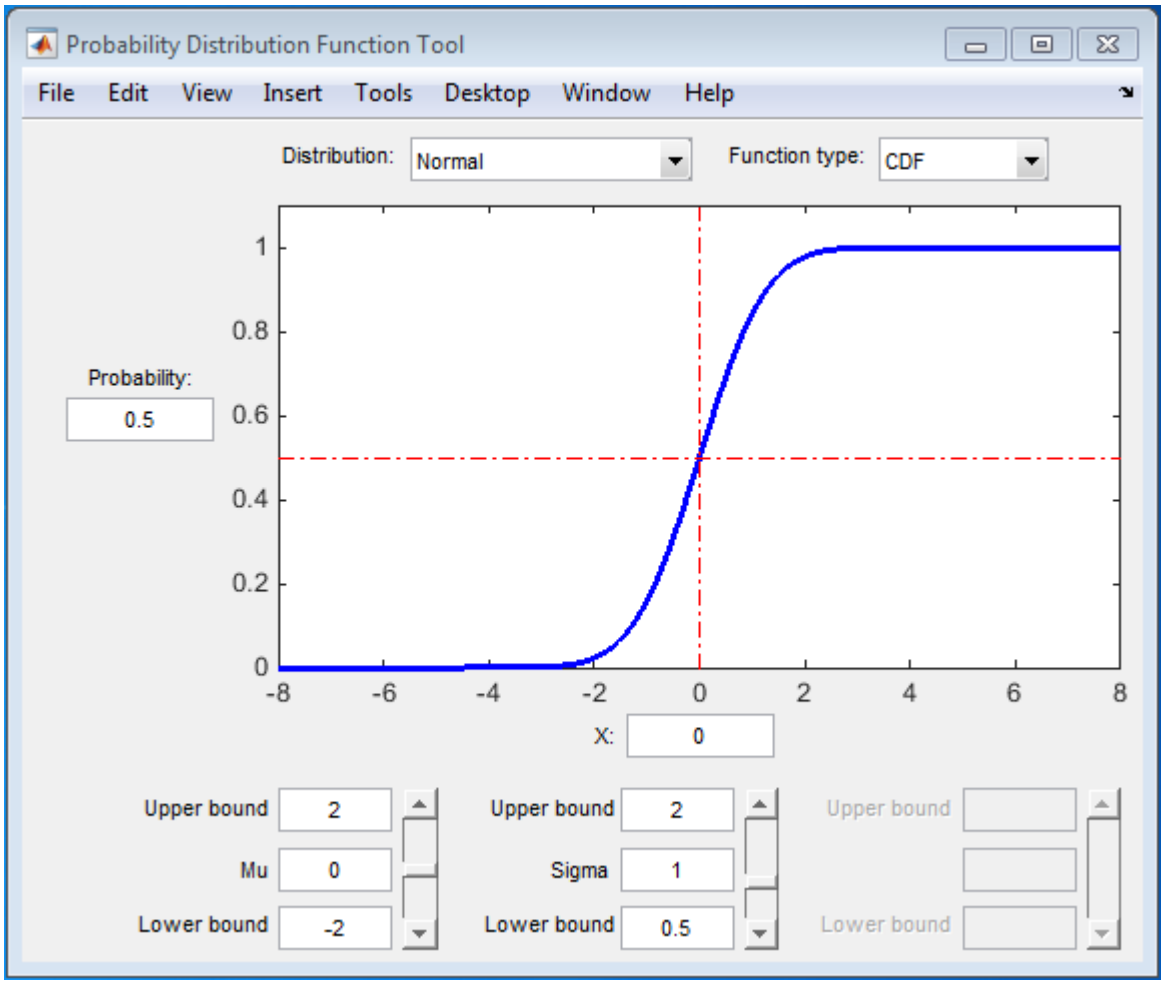
The **Distribution Fitter** app allows you to interactively fit a probability distribution to your data. You can display different types of plots, compute confidence bounds, and evaluate the fit of the data. You can also exclude data from the fit. You can save the data, and export the fit to your workspace as a probability distribution object to perform further analysis.

Load the Distribution Fitter app from the Apps tab, or by entering `distributionFitter` in the command window. For more information, see “Model Data Using the Distribution Fitter App” on page 5-51.



Probability Distribution Function Tool

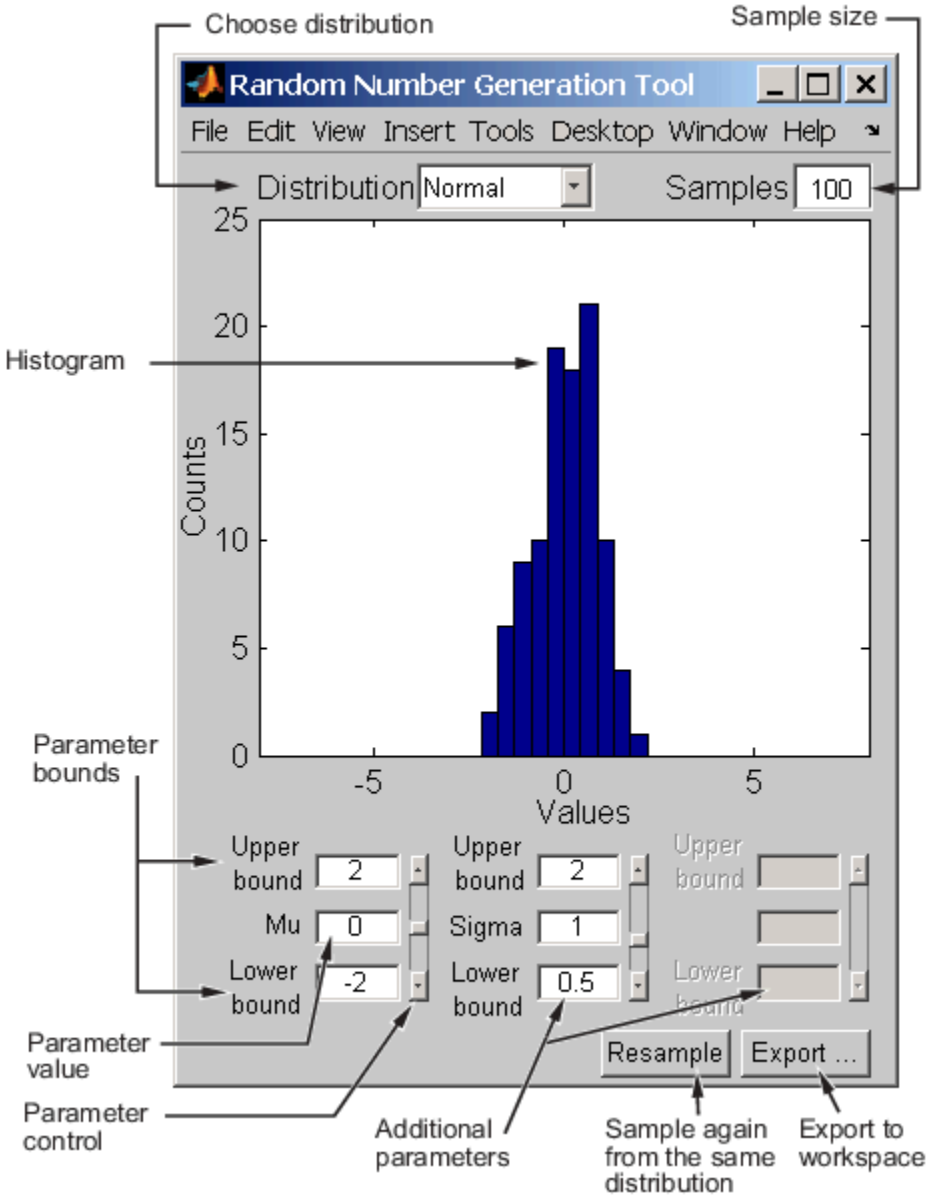
The **Probability Distribution Function** user interface visually explores probability distributions. You can load the Probability Distribution Function user interface by entering `disttool` in the command window.



Random Number Generation Tool

The Random Number Generation user interface generates random data from a specified distribution and exports the results to your workspace. You can use this tool to explore the effects of changing parameters and sample size on the distributions.

The Random Number Generation user interface allows you to set parameter values for the distribution and change their lower and upper bounds; draw another sample from the same distribution, using the same size and parameters; and export the current random sample to your workspace for use in further analysis. A dialog box enables you to provide a name for the sample.



See Also

Distribution Fitter | **Probability Distribution Function** | `fitdist` | `makedist` | `randtool`

More About

- “Multinomial Probability Distribution Objects” on page 5-95
- “Multinomial Probability Distribution Functions” on page 5-98
- “Fit Kernel Distribution Object to Data” on page 5-36
- “Generate Random Numbers Using the Triangular Distribution” on page 5-47
- “Supported Distributions” on page 5-14

Supported Distributions

In this section...

Statistics and Machine Learning Toolbox supports more than 30 probability distributions, including parametric, nonparametric, continuous, and discrete distributions.

The toolbox provides several ways to work with probability distributions.

- Use *probability distribution objects* to fit a probability distribution object to sample data, or to create a probability distribution object with specified parameter values. Once you create a probability distribution object, you can use object functions to:
  - Compute confidence intervals for the distribution parameters (`paramci`).
  - Compute summary statistics, including mean (`mean`), median (`median`), interquartile range (`iqr`), variance (`var`), and standard deviation (`std`).
  - Evaluate the probability density function (`pdf`).
  - Evaluate the cumulative distribution function (`cdf`) or the inverse cumulative distribution function (`icdf`).
  - Compute the negative loglikelihood (`negloglik`) and profile likelihood function (`proflik`) for the distribution.
  - Generate random numbers from the distribution (`random`).
  - Truncate the distribution to specified lower and upper limits (`truncate`).

  Each distribution object page provides information about the object’s properties and the functions you can use to work with the object.
- Use *probability distribution functions* to work with data input from matrices. Some of the supported distributions have distribution-specific functions. These functions use the following abbreviations, as in `normpdf`, `normcdf`, `norminv`, `normstat`, `normfit`, `normlike`, and `normrnd`:
  - `pdf` — Probability density functions
  - `cdf` — Cumulative distribution functions
  - `inv` — Inverse cumulative distribution functions
  - `stat` — Distribution statistics functions
  - `fit` — Distribution Fitter functions
  - `like` — Negative loglikelihood functions
  - `rnd` — Random number generators

  You can also use the following generic functions to work with most of the distributions:

  - `pdf` — Probability density function
  - `cdf` — Cumulative distribution function
  - `icdf` — Inverse cumulative distribution function
  - `random` — Random number generating function
  - `mle` — Distribution fitting function
- Use *probability distribution apps and user interfaces* to interactively fit, explore, and generate random numbers from probability distributions. Available apps and user interfaces include:
  - The **Distribution Fitter** app, to interactively fit a distribution to sample data, and export a probability distribution object to the workspace.
  - The **Probability Distribution Function** user interface, to visually explore the effect on the pdf and cdf of changing the distribution parameter values.
  - The Random Number Generation user interface (`randtool`), to interactively generate random numbers from a probability distribution with specified parameter values and export them to the workspace.

For more information on the different ways to work with probability distributions, see “Working with Probability Distributions” on page 5-3.

Continuous Distributions (Data)

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs

Beta on page B-6	`BetaDistribution`	`betapdf`
		`betacdf`
		`betainv`
		`betastat`
		`betafit`
		`betalike`
		`betarnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Birnbaum-Saunders on page B-18	`BirnbaumSaundersDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Burr Type XII on page B-19	`BurrDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs

Exponential on page B-33	`ExponentialDistribution`	`exppdf`
		`expcdf`
		`expinv`
		`expstat`
		`expfit`
		`explike`
		`exprnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Extreme value on page B-40	`ExtremeValueDistribution`	`evpdf`
		`evcdf`
		`evinv`
		`evstat`
		`evfit`
		`evlike`
		`evrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Gamma on page B-47	`GammaDistribution`	`gampdf`
		`gamcdf`
		`gaminv`
		`gamstat`
		`gamfit`
		`gamlike`
		`gamrnd`
		`randg`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Generalized extreme value on page B-55	`GeneralizedExtremeValueDistribution`	`gevpdf`
		`gevcdf`
		`gevinv`
		`gevstat`
		`gevfit`
		`gevlike`
		`gevrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Generalized Pareto on page B-59	`GeneralizedParetoDistribution`	`gppdf`
		`gpcdf`
		`gpinv`
		`gpstat`
		`gpfit`
		`gplike`
		`gprnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Half-Normal on page B-68	`HalfNormalDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs

Inverse Gaussian on page B-75	`InverseGaussianDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Logistic on page B-85	`LogisticDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Loglogistic on page B-86	`LoglogisticDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Lognormal on page B-88	`LognormalDistribution`	`lognpdf`
		`logncdf`
		`logninv`
		`lognstat`
		`lognfit`
		`lognlike`
		`lognrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Nakagami on page B-108	`NakagamiDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Normal (Gaussian) on page B-119	`NormalDistribution`	`normpdf`
		`normcdf`
		`norminv`
		`normstat`
		`normfit`
		`normlike`
		`normrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Piecewise linear on page B-130	`PiecewiseLinearDistribution`			

Rayleigh on page B-137	`RayleighDistribution`	`raylpdf`
		`raylcdf`
		`raylinv`
		`raylstat`
		`raylfit`
		`raylrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs

Rician on page B-139	`RicianDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Stable on page B-140	`StableDistribution`		`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter

Triangular on page B-158	`TriangularDistribution`			

Uniform (continuous) on page B-163	`UniformDistribution`	`unifpdf`
		`unifcdf`
		`unifinv`
		`unifstat`
		`unifit`
		`unifrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Probability Distribution Function
				`randtool`

Weibull on page B-170	`WeibullDistribution`	`wblpdf`
		`wblcdf`
		`wblinv`
		`wblstat`
		`wblfit`
		`wbllike`
		`wblrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Continuous Distributions (Statistics)

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs

Chi-square on page B-28		`chi2pdf`
		`chi2cdf`
		`chi2inv`
		`chi2stat`
		`chi2rnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
				Probability Distribution Function
				`randtool`

F on page B-45		`fpdf`
		`fcdf`
		`finv`
		`fstat`
		`frnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
				Probability Distribution Function
				`randtool`

Noncentral chi-square on page B-113		`ncx2pdf`
		`ncx2cdf`
		`ncx2inv`
		`ncx2stat`
		`ncx2rnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
				Probability Distribution Function
				`randtool`

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps and UIs
Noncentral *F* on page B-115		ncfpdf ncfcdf ncfinv ncfstat ncfrnd	pdf cdf icdf random	**Probability Distribution Function** randtool
Noncentral *t* on page B-117		nctpdf nctcdf nctinv nctstat nctrnd	pdf cdf icdf random	**Probability Distribution Function** randtool
Student's *t* on page B-149		tpdf tcdf tinv tstat trnd	pdf cdf icdf random	**Probability Distribution Function** randtool
t location-scale on page B-156	tLocationScaleDistribution		pdf cdf icdf random mle	**Distribution Fitter**

Discrete Distributions

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps/UIs
Binomial on page B-10	BinomialDistribution	binopdf binocdf binoinv binostat binofit binornd	pdf cdf icdf random mle	**Distribution Fitter** **Probability Distribution Function** randtool
Bernoulli on page B-2			mle	
Geometric on page B-63		geopdf geocdf geoinv geostat mle geornd	pdf cdf icdf random mle	**Probability Distribution Function** randtool
Hypergeometric on page B-73		hygepdf hygecdf hygeinv hygestat hygernd	pdf cdf icdf random mle	**Probability Distribution Function** randtool

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps/UIs
Multinomial on page B-96	`MultinomialDistribution`	`mnpdf` `mnrnd`		

Negative binomial on page B-109	`NegativeBinomialDistribution`	`nbinpdf`
		`nbincdf`
		`nbininv`
		`nbinstat`
		`nbinfit`
		`nbinrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Poisson on page B-131	`PoissonDistribution`	`poisspdf`
		`poisscdf`
		`poissinv`
		`poisstat`
		`poissfit`
		`poissrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Distribution Fitter
				Probability Distribution Function
				`randtool`

Uniform (discrete) on page B-168		`unidpdf`
		`unidcdf`
		`unidinv`
		`unidstat`
		`unidrnd`
			`pdf`
			`cdf`
			`icdf`
			`random`
			`mle`
				Probability Distribution Function
				`randtool`

Multivariate Distributions

Distribution	Object	Distribution-Specific Functions	Generic Functions	Apps/UI

Copula on page 5-121 (Gaussian copula, *t* copula, Clayton copula, Frank copula, Gumbel copula)		`copulapdf`
		`copulacdf`
		`copulaparam`
		`copulastat`
		`copulafit`
		`copularnd`		

Gaussian Mixture	`gmdistribution`	`fitgmdist`
		`pdf`
		`cdf`
		`random`		

Inverse Wishart on page B-76		`iwishrnd`		

Multivariate normal on page B-98		`mvnpdf`
		`mvncdf`
		`mvnrnd`		

Multivariate *t* on page B-104		`mvtpdf`
		`mvtcdf`
		`mvtrnd`		

Wishart on page B-178		`wishrnd`		

Nonparametric Distributions

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps/UIs
Kernel on page B-78	`KernelDistribution`	`ksdensity`		**Distribution Fitter**
Pareto tails	`paretotails`			

Flexible Distribution Families

Distribution	Distribution Objects	Distribution-Specific Functions	Generic Functions	Apps/UIs
Pearson system on page 7-20		`pearsrnd`		
Johnson system on page 7-20		`johnsrnd`		

See Also

More About

- “Working with Probability Distributions” on page 5-3
- “Nonparametric and Empirical Probability Distributions” on page 5-30

Maximum Likelihood Estimation

The `mle` function computes maximum likelihood estimates (MLEs) for a distribution specified by its name and for a custom distribution specified by its probability density function (pdf), log pdf, or negative log likelihood function.

For some distributions, MLEs can be given in closed form and computed directly. For other distributions, a search for the maximum likelihood must be employed. The search can be controlled with an `options` input argument, created using the `statset` function. For efficient searches, it is important to choose a reasonable distribution model and set appropriate convergence tolerances.

MLEs can be biased, especially for small samples. As sample size increases, however, MLEs become unbiased minimum variance estimators with approximate normal distributions. This is used to compute confidence bounds for the estimates.

For example, consider the following distribution of means from repeated random samples of an exponential distribution:

```
mu = 1; % Population parameter
n = 1e3; % Sample size
ns = 1e4; % Number of samples

rng('default')  % For reproducibility
samples = exprnd(mu,n,ns); % Population samples
means = mean(samples); % Sample means
```

The Central Limit Theorem says that the means will be approximately normally distributed, regardless of the distribution of the data in the samples. The `mle` function can be used to find the normal distribution that best fits the means:

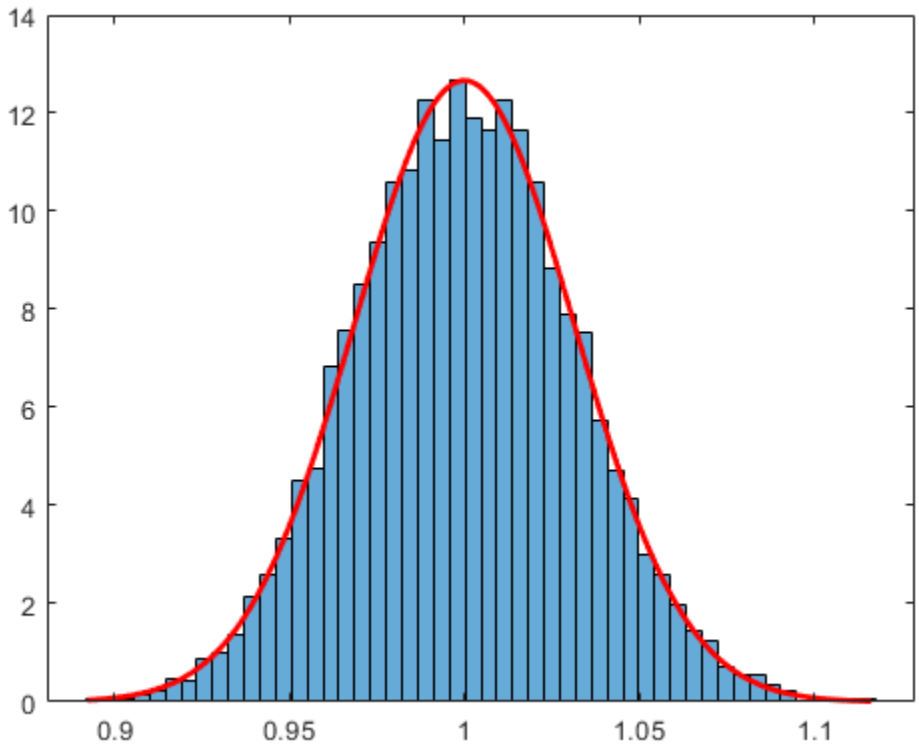
```
[phat,pci] = mle(means)
```

```
phat = 1×2

    1.0000    0.0315


pci = 2×2

    0.9994    0.0311
    1.0006    0.0319
```

`phat(1)` and `phat(2)` are the MLEs for the mean and standard deviation. `pci(:,1)` and `pci(:,1)` are the corresponding 95% confidence intervals.

Visualize the distribution of sample means together with the fitted normal distribution.

```
numbins = 50;
histogram(means,numbins,'Normalization','pdf')
hold on
x = min(means):0.001:max(means);
y = normpdf(x,phat(1),phat(2));
plot(x,y,'r','LineWidth',2)
```



See Also

`histogram` | `mle`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Negative Loglikelihood Functions

Negative loglikelihood functions for supported Statistics and Machine Learning Toolbox distributions all end with `like`, as in `explike`. Each function represents a parametric family of distributions. Input arguments are lists of parameter values specifying a particular member of the distribution family followed by an array of data. Functions return the negative loglikelihood of the parameters, given the data.

To find maximum likelihood estimates (MLEs), you can use a negative loglikelihood function as an objective function of the optimization problem and solve it by using the MATLAB function `fminsearch` or functions in Optimization Toolbox™ and Global Optimization Toolbox. These functions allow you to choose a search algorithm and exercise low-level control over algorithm execution. By contrast, the `mle` function and the distribution fitting functions that end with `fit`, such as `normfit` and `gamfit`, use preset algorithms with options limited to those set by the `statset` function.

You can specify a parametric family of distributions by using the probability density function (pdf) $f(x|\theta)$, where x represents an outcome of a random variable and θ represents the distribution parameters. When you view $f(x|\theta)$ as a function of θ for a fixed x, the function $f(x|\theta)$ is the likelihood of parameters θ for a single outcome x. The likelihood of parameters θ for an independent and identically distributed random sample data set X is:

$$L(\theta) = \prod_{x \in X} f(x|\theta).$$

Given X, MLEs maximize $L(\theta)$ over all possible θ. Numerical algorithms find MLEs that (equivalently) maximize the loglikelihood function, $\log(L(\theta))$. The logarithm transforms the product of potentially small likelihoods into a sum of logs, which is easier to distinguish from 0 in computation. For convenience, Statistics and Machine Learning Toolbox negative loglikelihood functions return the *negative* of this sum because optimization algorithms typically search for minima rather than maxima.

Find MLEs Using Negative Loglikelihood Function

This example shows how to find MLEs by using the `gamlike` and `fminsearch` functions.

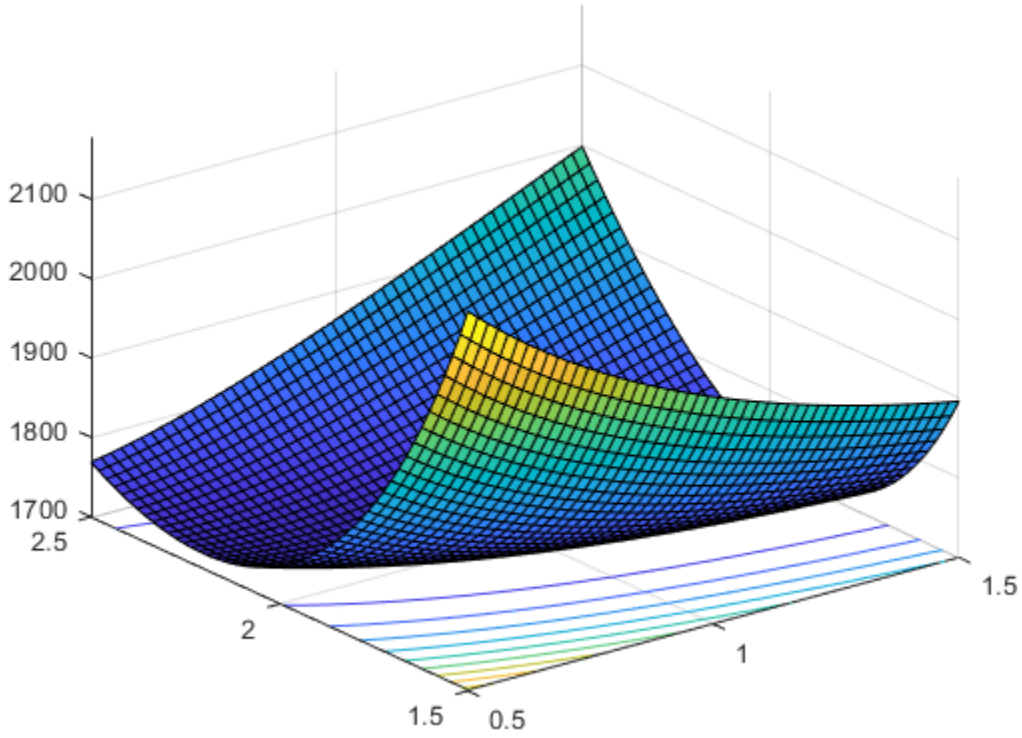
Use the `gamrnd` function to generate a random sample from a specific “Gamma Distribution” on page B-47.

```
rng default;  % For reproducibility
a = [1,2];
X = gamrnd(a(1),a(2),1e3,1);
```

Visualize the likelihood surface in the neighborhood of `a` given `X` by using the `gamlike` function.

```
mesh = 50;
delta = 0.5;
a1 = linspace(a(1)-delta,a(1)+delta,mesh);
a2 = linspace(a(2)-delta,a(2)+delta,mesh);
logL = zeros(mesh); % Preallocate memory
for i = 1:mesh
    for j = 1:mesh
        logL(i,j) = gamlike([a1(i),a2(j)],X);
    end
end
```

```
[A1,A2] = meshgrid(a1,a2);
surfc(A1,A2,logL)
```



Search for the minimum of the likelihood surface by using the `fminsearch` function.

```
LL = @(u)gamlike([u(1),u(2)],X); % Likelihood given X
MLES = fminsearch(LL,[1,2])
```

```
MLES = 1×2

    0.9980    2.0172
```

Compare `MLES` to the estimates returned by the `gamfit` function.

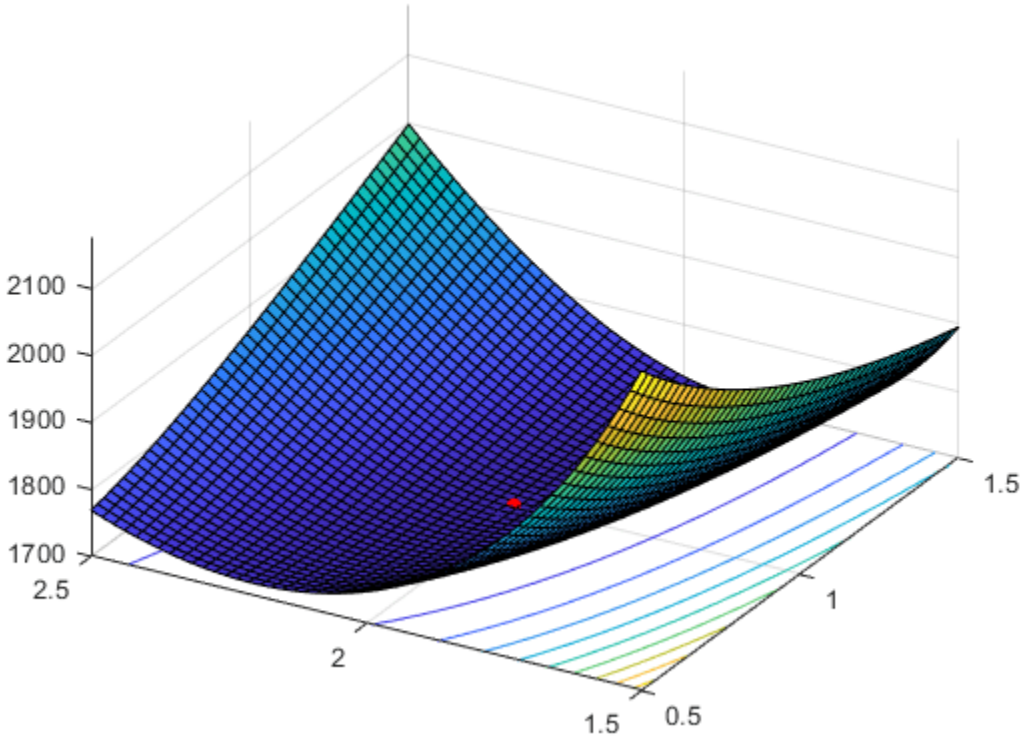
```
ahat = gamfit(X)
```

```
ahat = 1×2

    0.9980    2.0172
```

The difference of each parameter between `MLES` and `ahat` is less than `1e-4`.

Add the MLEs to the surface plot.

```
hold on
plot3(MLES(1),MLES(2),LL(MLES),'ro','MarkerSize',5,'MarkerFaceColor','r')
view([-60 40]) % Rotate to show the minimum
```



See Also

`fminsearch` | `negloglik` | `statset` | `surfc`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Random Number Generation

Statistics and Machine Learning Toolbox supports the generation of random numbers from various distributions. Each random number generator (RNG) represents a parametric family of distributions. RNGs return random numbers from the specified distribution in an array of the specified dimensions.

Other random number generation functions which do not support specific distributions include:

- `cvpartition`
- `datasample`
- `hmmgenerate`
- `lhsdesign`
- `lhsnorm`
- `mhsample`
- `random`
- `randsample`
- `slicesample`

RNGs in Statistics and Machine Learning Toolbox software depend on the default random number stream of MATLAB via the `rand` and `randn` functions. Each RNG uses one of the techniques discussed in “Common Pseudorandom Number Generation Methods” on page 7-2 to generate random numbers from a given distribution.

By controlling the default random number stream and its state, you can control how the RNGs in Statistics and Machine Learning Toolbox software generate random values. For example, to reproduce the same sequence of values from an RNG, you can save and restore the default stream's state, or reset the default stream. For details on managing the default random number stream, see “Managing the Global Stream Using RandStream”.

MATLAB initializes the default random number stream to the same state each time it starts up. Thus, RNGs in Statistics and Machine Learning Toolbox software will generate the same sequence of values for each MATLAB session unless you modify that state at startup. One simple way to do that is to add commands to `startup.m` such as

```
rng shuffle
```

that initialize the default random number stream to a different state for each session.

The following table lists the supported distributions and their respective random number generation functions.

Distribution	Random Number Generation Function
Beta on page B-6	`betarnd`, `random`, `randtool`
Binomial on page B-10	`binornd`, `random`, `randtool`
Birnbaum-Saunders on page B-18	`random`
Burr Type XII on page B-19	`random`, `randtool`
Chi-square on page B-28	`chi2rnd`, `random`, `randtool`
Clayton copula on page 5-121	`copularnd`

Distribution	Random Number Generation Function
Exponential on page B-33	`exprnd`, `random`, `randtool`
Extreme value on page B-40	`evrnd`, `random`, `randtool`
F on page B-45	`frnd`, `random`, `randtool`
Frank copula on page 5-121	`copularnd`
Gamma on page B-47	`gamrnd`, `randg`, `random`, `randtool`
Gaussian copula on page 5-121	`copularnd`
Gaussian Mixture	`random`
Generalized extreme value on page B-55	`gevrnd`, `random`, `randtool`
Generalized Pareto on page B-59	`gprnd`, `random`, `randtool`
Geometric on page B-63	`geornd`, `random`, `randtool`
Gumbel copula on page 5-121	`copularnd`
Half-Normal on page B-68	`random`, `randtool`
Hypergeometric on page B-73	`hygernd`, `random`, `randtool`
Inverse Gaussian on page B-75	`random`
Inverse Wishart on page B-76	`iwishrnd`
Johnson system on page 7-20	`johnsrnd`
Kernel on page B-78	`random`
Logistic on page B-85	`random`
Loglogistic on page B-86	`random`
Lognormal on page B-88	`lognrnd`, `random`, `randtool`
Multinomial on page B-96	`mnrnd`
Multivariate normal on page B-98	`mvnrnd`
Multivariate *t* on page B-104	`mvtrnd`
Nakagami on page B-108	`random`
Negative binomial on page B-109	`nbinrnd`, `random`, `randtool`
Noncentral chi-square on page B-113	`ncx2rnd`, `random`, `randtool`
Noncentral *F* on page B-115	`ncfrnd`, `random`, `randtool`
Noncentral *t* on page B-117	`nctrnd`, `random`, `randtool`
Normal (Gaussian) on page B-119	`normrnd`, `randn`, `random`, `randtool`
Pareto	`random`
Pearson system on page 7-20	`pearsrnd`
Piecewise on page B-130	`random`
Poisson on page B-131	`poissrnd`, `random`, `randtool`
Rayleigh on page B-137	`raylrnd`, `random`, `randtool`
Rician on page B-139	`random`
Stable on page B-140	`random`

Distribution	Random Number Generation Function
Student's *t* on page B-149	`trnd`, `random`, `randtool`
t copula on page 5-121	`copularnd`
t location- scale on page B-156	`random`
Triangular on page B-158	`random`
Uniform (continuous) on page B-163	`unifrnd`, `rand`, `random`
Uniform (discrete) on page B-168	`unidrnd`, `random`, `randtool`
Weibull on page B-170	`wblrnd`, `random`
Wishart on page B-178	`wishrnd`

See Also

More About

- “Generate Random Numbers Using the Triangular Distribution” on page 5-47
- “Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101
- “Generating Pseudorandom Numbers” on page 7-2
- “Generating Quasi-Random Numbers” on page 7-12
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Nonparametric and Empirical Probability Distributions

In this section...

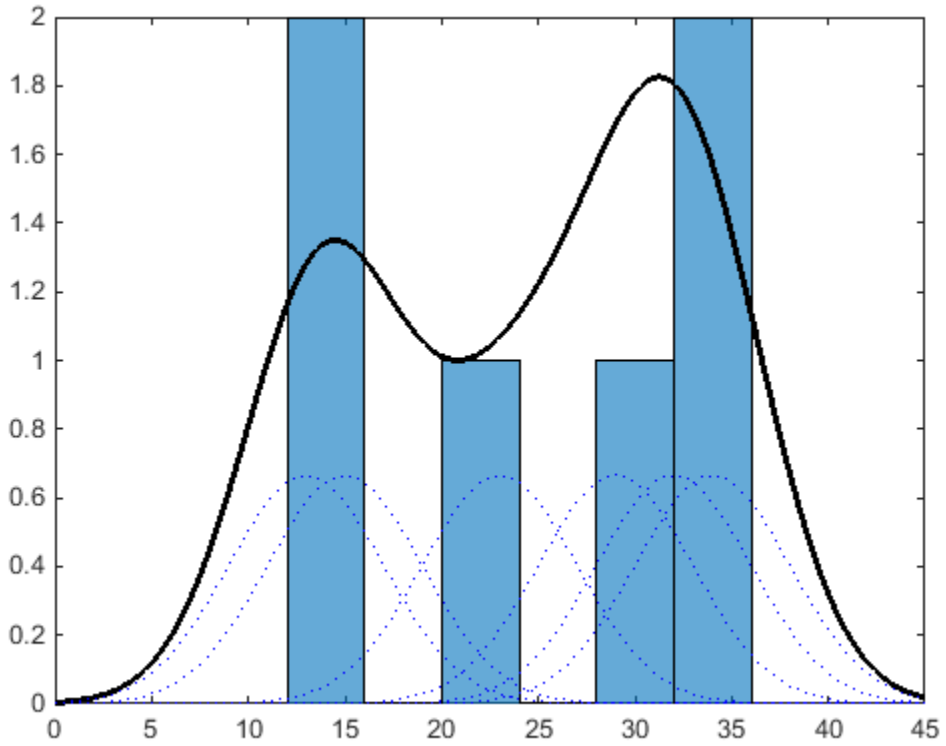
Overview

In some situations, you cannot accurately describe a data sample using a parametric distribution. Instead, the probability density function (pdf) or cumulative distribution function (cdf) must be estimated from the data. Statistics and Machine Learning Toolbox provides several options for estimating the pdf or cdf from sample data.

Kernel Distribution

A kernel distribution on page B-78 produces a nonparametric probability density estimate that adapts itself to the data, rather than selecting a density with a particular parametric form and estimating the parameters. This distribution is defined by a kernel density estimator, a smoothing function that determines the shape of the curve used to generate the pdf, and a bandwidth value that controls the smoothness of the resulting density curve.

Similar to a histogram, the kernel distribution builds a function to represent the probability distribution using the sample data. But unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve. The following plot shows a visual comparison of a histogram and a kernel distribution generated from the same sample data.



A histogram represents the probability distribution by establishing bins and placing each data value in the appropriate bin. Because of this bin count approach, the histogram produces a discrete probability density function. This might be unsuitable for certain applications, such as generating random numbers from a fitted distribution.

Alternatively, the kernel distribution builds the probability density function (pdf) by creating an individual probability density curve for each data value, then summing the smooth curves. This approach creates one smooth, continuous probability density function for the data set.

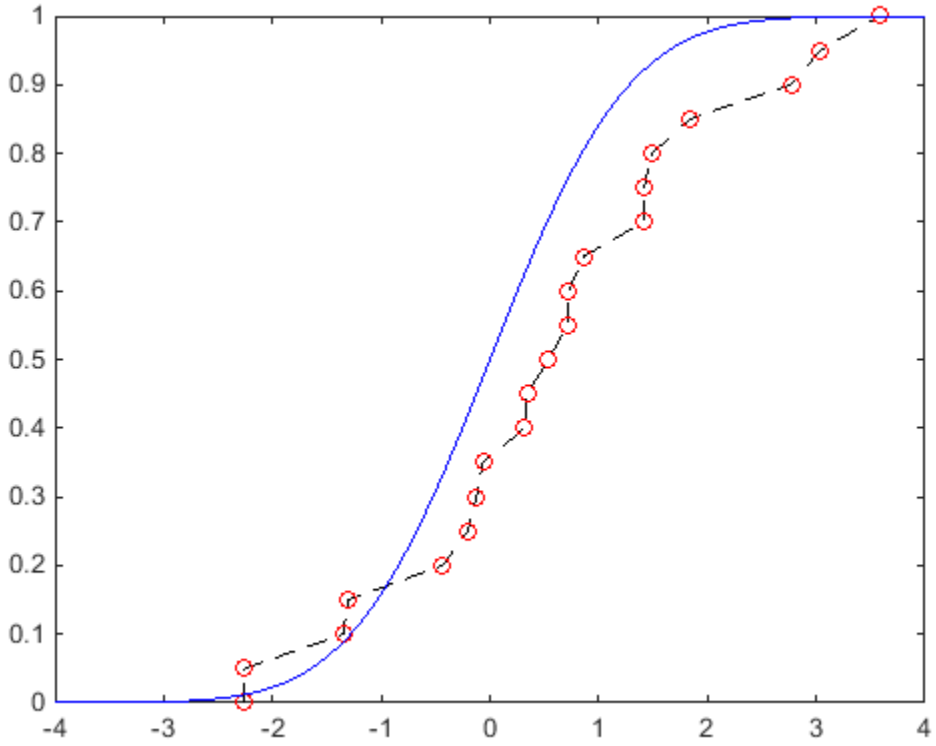
For more general information about kernel distributions, see “Kernel Distribution” on page B-78. For information on how to work with a kernel distribution, see `Using KernelDistribution Objects` and `ksdensity`.

Empirical Cumulative Distribution Function

An empirical cumulative distribution function (`ecdf`) estimates the cdf of a random variable by assigning equal probability to each observation in a sample. Because of this approach, the ecdf is a discrete cumulative distribution function that creates an exact match between the ecdf and the distribution of the sample data.

The following plot shows a visual comparison of the ecdf of 20 random numbers generated from a standard normal distribution, and the theoretical cdf of a standard normal distribution. The circles indicate the value of the ecdf calculated at each sample data point. The dashed line that passes through each circle visually represents the ecdf, although the ecdf is not a continuous function. The

solid line shows the theoretical cdf of the standard normal distribution from which the random numbers in the sample data were drawn.



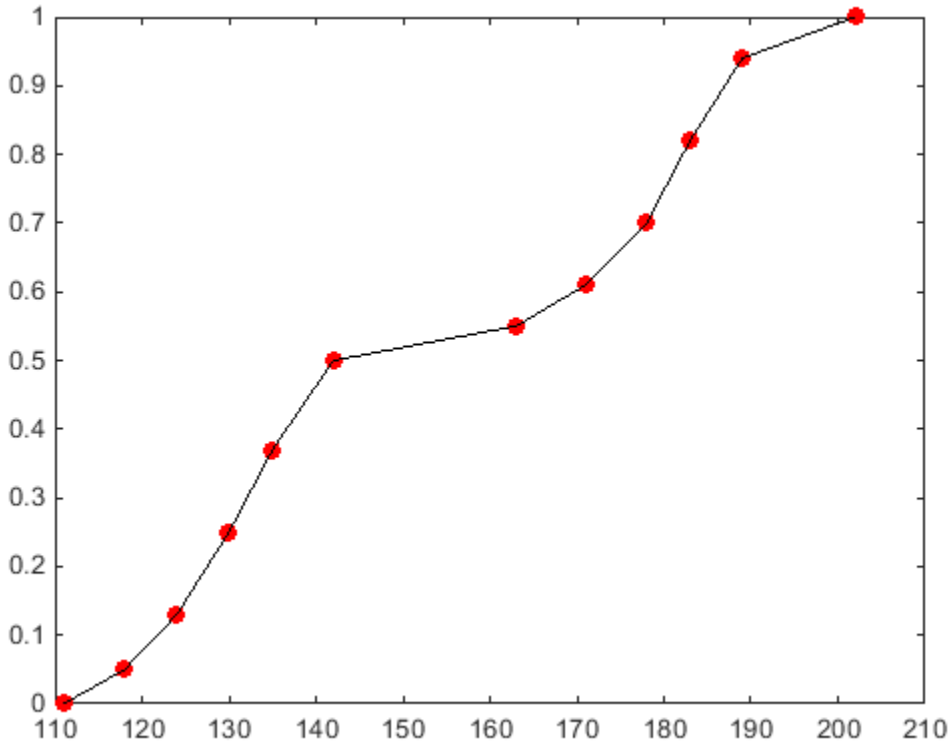
The ecdf is similar in shape to the theoretical cdf, although it is not an exact match. Instead, the ecdf is an exact match to the sample data. The ecdf is a discrete function, and is not smooth, especially in the tails where data might be sparse. You can smooth the distribution with Pareto tails on page 5-33, using the `paretotails` function.

For more information and additional syntax options, see `ecdf`. To construct a continuous function based on cdf values computed from sample data, see “Piecewise Linear Distribution” on page 5-32.

Piecewise Linear Distribution

A piecewise linear distribution on page B-130 estimates an overall cdf for the sample data by computing the cdf value at each individual point, and then linearly connecting these values to form a continuous curve.

The following plot shows the cdf for a piecewise linear distribution based on a sample of hospital patients’ weight measurements. The circles represent each individual data point (weight measurement). The black line that passes through each data point represents the piecewise linear distribution cdf for the sample data.



A piecewise linear distribution linearly connects the cdf values calculated at each sample data point to form a continuous curve. By contrast, an empirical cumulative distribution function on page 5-31 constructed using the `ecdf` function produces a discrete cdf. For example, random numbers generated from the ecdf can only include *x* values contained in the original sample data. Random numbers generated from a piecewise linear distribution can include any *x* value between the lower and upper boundaries of the sample data.

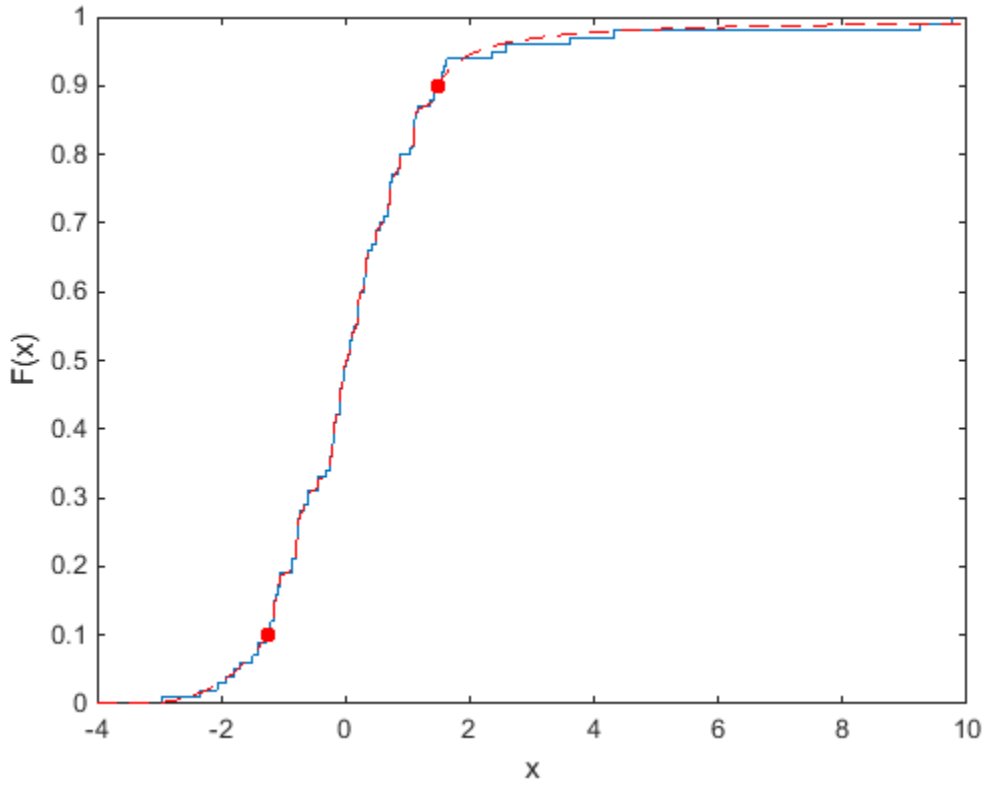
Because the piecewise linear distribution cdf is constructed from the values contained in the sample data, the resulting curve is often not smooth, especially in the tails where data might be sparse. You can smooth the distribution with Pareto tails on page 5-33, using the `paretotails` function.

For information on how to work with a piecewise linear distribution, see Using `PiecewiseLinearDistribution` Objects.

Pareto Tails

Pareto tails use a piecewise approach to improve the fit of a nonparametric cdf by smoothing the tails of the distribution. You can fit a kernel distribution on page 5-30, empirical cdf on page 5-31, or a user-defined estimator to the middle data values, then fit generalized Pareto distribution on page B-59 curves to the tails. This technique is especially useful when the sample data is sparse in the tails.

The following plot shows the empirical cdf (ecdf) of a data sample containing 20 random numbers. The solid line represents the ecdf, and the dashed line represents the empirical cdf with Pareto tails fit to the lower and upper 10 percent of the data. The circles denote the boundaries for the lower and upper 10 percent of the data.

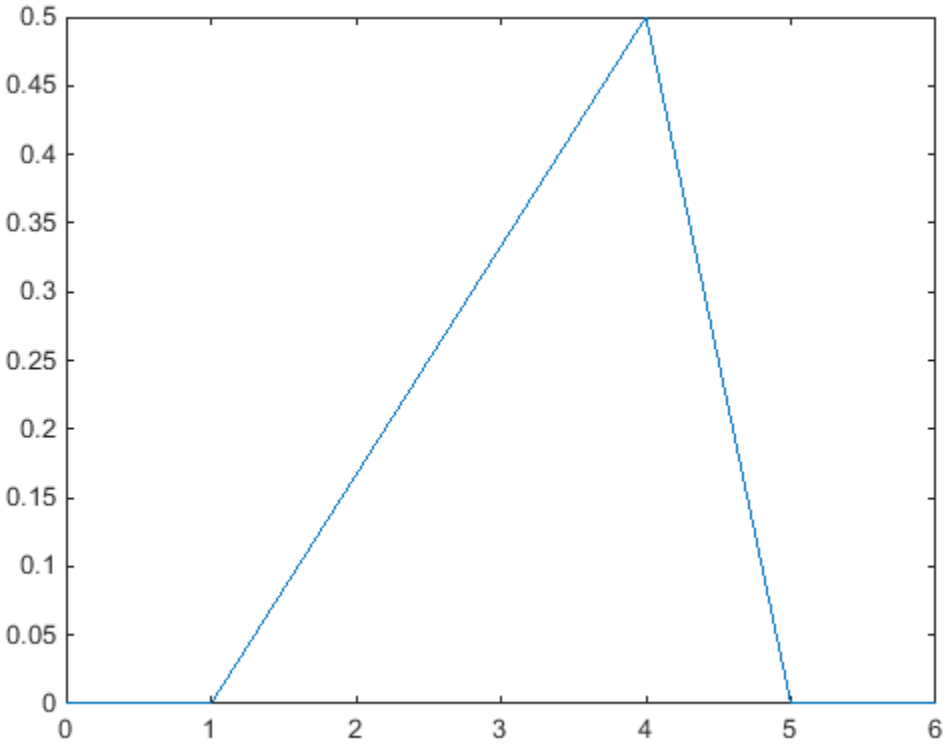


Fitting Pareto tails to the lower and upper 10 percent of the sample data makes the cdf smoother in the tails, where the data is sparse. For more information on working with Pareto tails, see `paretotails`.

Triangular Distribution

A “Triangular Distribution” on page B-158 provides a simplistic representation of the probability distribution when limited sample data is available. This continuous distribution is parameterized by a lower limit, peak location, and upper limit. These points are linearly connected to estimate the pdf of the sample data. You can use the mean, median, or mode of the data as the peak location.

The following plot shows the triangular distribution pdf of a random sample of 10 integers from 0 to 5. The lower limit is the smallest integer in the sample data, and the upper limit is the largest integer. The peak for this plot is at the mode, or most frequently-occurring value, in the sample data.



Business applications such as simulation and project management sometimes use a triangular distribution to create models when limited sample data exists. For more information, see “Triangular Distribution” on page B-158.

See Also

`ecdf` | `ksdensity` | `paretotails`

More About

- “Kernel Distribution” on page B-78
- “Piecewise Linear Distribution” on page B-130
- “Triangular Distribution” on page B-158
- “Generalized Pareto Distribution” on page B-59
- “Fit a Nonparametric Distribution with Pareto Tails” on page 5-43

Fit Kernel Distribution Object to Data

This example shows how to fit a kernel probability distribution object to sample data.

Step 1. Load sample data.

Load the sample data.

```
load carsmall;
```

This data contains miles per gallon (`MPG`) measurements for different makes and models of cars, grouped by country of origin (`Origin`), model year (`Year`), and other vehicle characteristics.

Step 2. Fit a kernel distribution object.

Use `fitdist` to fit a kernel probability distribution object to the miles per gallon (`MPG`) data for all makes of cars.

```
pd = fitdist(MPG,'Kernel')
```

```
pd =
  KernelDistribution

    Kernel = normal
    Bandwidth = 4.11428
    Support = unbounded
```

This creates a `prob.KernelDistribution` object. By default, `fitdist` uses a normal kernel smoothing function and chooses an optimal bandwidth for estimating normal densities, unless you specify otherwise. You can access information about the fit and perform further calculations using the related object functions.

Step 3. Compute descriptive statistics.

Compute the mean, median, and standard deviation of the fitted kernel distribution.

```
m = mean(pd)
```

```
m = 23.7181
```

```
med = median(pd)
```

```
med = 23.4841
```

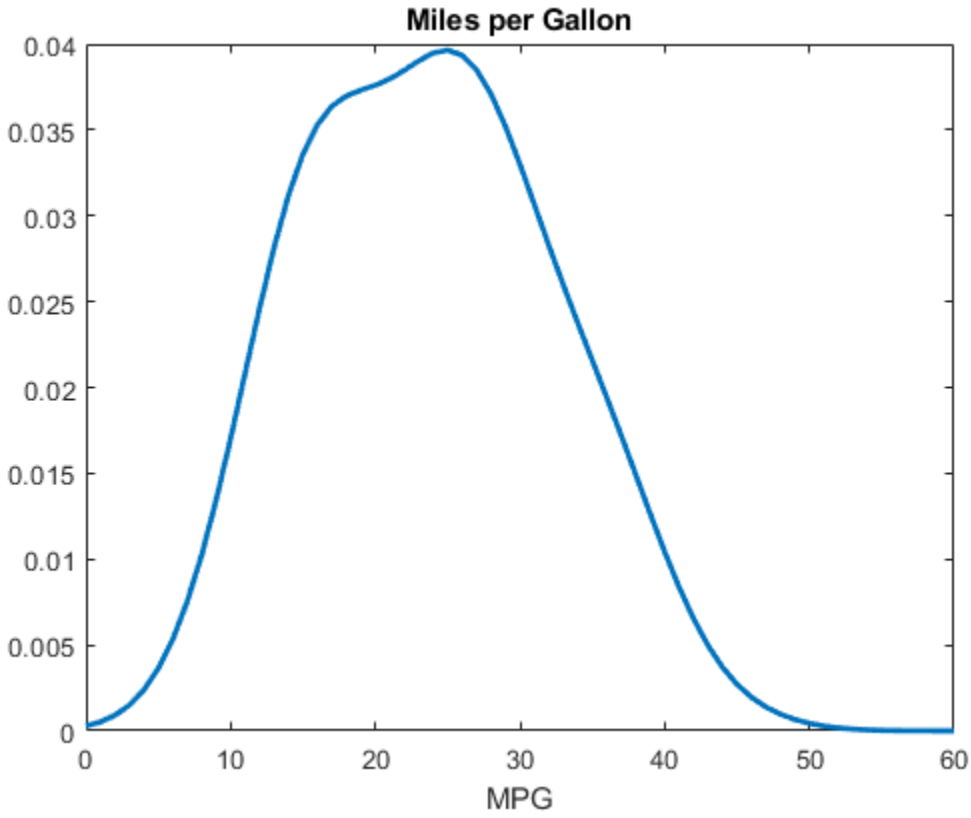
```
s = std(pd)
```

```
s = 8.9896
```

Step 4. Compute and plot the pdf.

Compute and plot the pdf of the fitted kernel distribution.

```
figure
x = 0:1:60;
y = pdf(pd,x);
plot(x,y,'LineWidth',2)
title('Miles per Gallon')
xlabel('MPG')
```

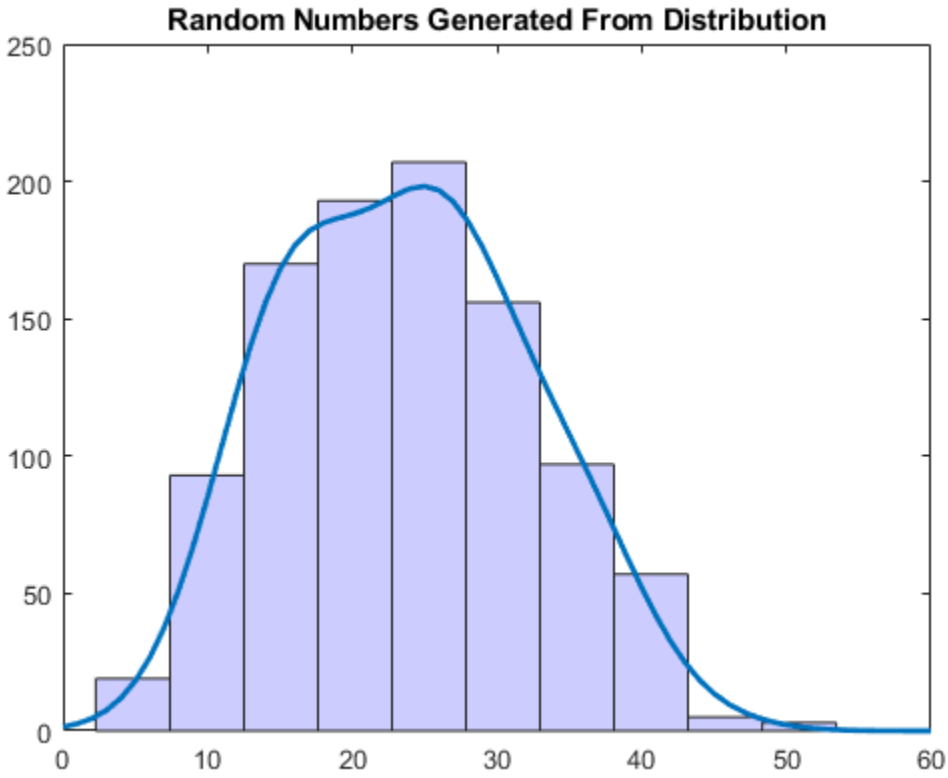


The plot shows the pdf of the kernel distribution fit to the `MPG` data across all makes of cars. The distribution is smooth and fairly symmetrical, although it is slightly skewed with a heavier right tail.

Step 5. Generate random numbers.

Generate a vector of random numbers from the fitted kernel distribution.

```
rng('default')  % For reproducibility
r = random(pd,1000,1);
figure
hist(r);
set(get(gca,'Children'),'FaceColor',[.8 .8 1]);
hold on
y = y*5000;  % Scale pdf to overlay on histogram
plot(x,y,'LineWidth',2)
title('Random Numbers Generated From Distribution')
hold off
```



The histogram has a similar shape to the pdf plot because the random numbers generate from the nonparametric kernel distribution fit to the sample data.

See Also

`KernelDistribution` | `fitdist` | `ksdensity`

More About

- “Kernel Distribution” on page B-78
- “Fit Kernel Distribution Using ksdensity” on page 5-39
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Fit Kernel Distribution Using ksdensity

This example shows how to generate a kernel probability density estimate from sample data using the `ksdensity` function.

Step 1. Load sample data.

Load the sample data.

```
load carsmall;
```

This data contains miles per gallon (`MPG`) measurements for different makes and models of cars, grouped by country of origin (`Origin`), model year (`Year`), and other vehicle characteristics.

Step 2. Generate a kernel probability density estimate.

Use `ksdensity` to generate a kernel probability density estimate for the miles per gallon (`MPG`) data.

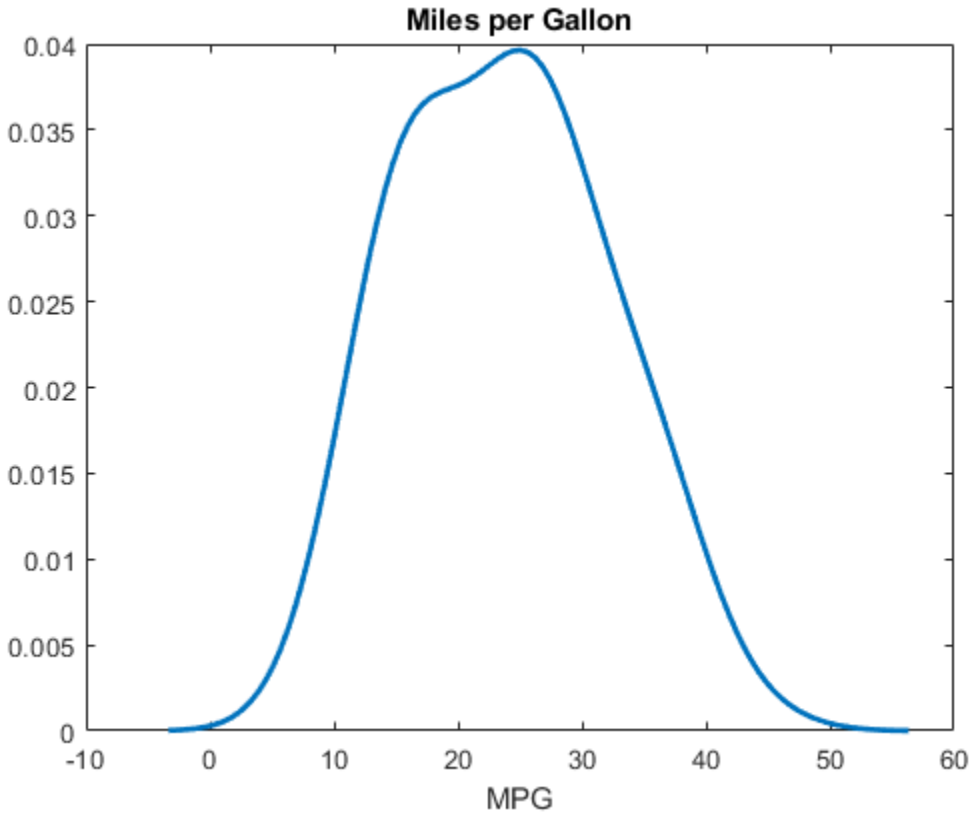
```
[f,xi] = ksdensity(MPG);
```

By default, `ksdensity` uses a normal kernel smoothing function and chooses an optimal bandwidth for estimating normal densities, unless you specify otherwise.

Step 3. Plot the kernel probability density estimate.

Plot the kernel probability density estimate to visualize the `MPG` distribution.

```
plot(xi,f,'LineWidth',2)
title('Miles per Gallon')
xlabel('MPG')
```



The plot shows the pdf of the kernel distribution fit to the `MPG` data across all makes of cars. The distribution is smooth and fairly symmetrical, although it is slightly skewed with a heavier right tail.

See Also

`KernelDistribution` | `fitdist` | `ksdensity`

More About

- “Kernel Distribution” on page B-78
- “Fit Kernel Distribution Object to Data” on page 5-36

Fit Distributions to Grouped Data Using ksdensity

This example shows how to fit kernel distributions to grouped sample data using the `ksdensity` function.

Step 1. Load sample data.

Load the sample data.

```
load carsmall
```

The data contains miles per gallon (`MPG`) measurements for different makes and models of cars, grouped by country of origin (`Origin`), model year (`Model_Year`), and other vehicle characteristics.

Step 2. Group sample data by origin.

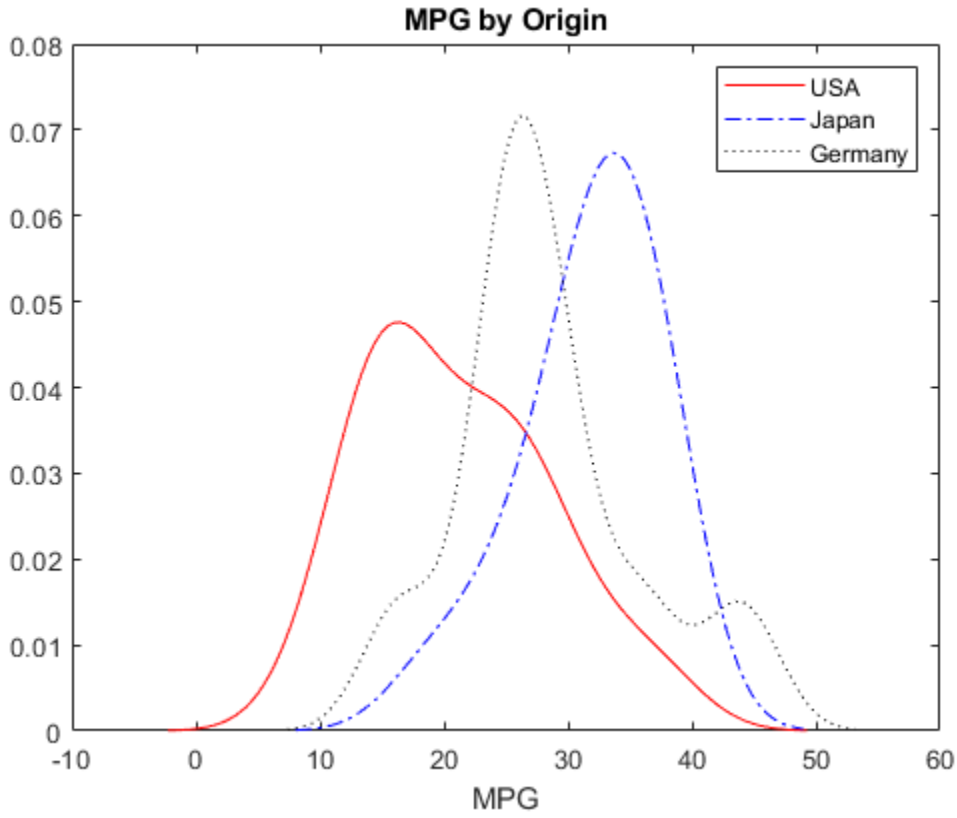
Group the `MPG` data by origin (`Origin`) for cars made in the USA, Japan, and Germany.

```
Origin = categorical(cellstr(Origin));
MPG_USA = MPG(Origin=='USA');
MPG_Japan = MPG(Origin=='Japan');
MPG_Germany = MPG(Origin=='Germany');
```

Step 3. Compute and plot the pdf.

Compute and plot the pdf for each group.

```
[fi,xi] = ksdensity(MPG_USA);
plot(xi,fi,'r-')
hold on

[fj,xj] = ksdensity(MPG_Japan);
plot(xj,fj,'b-.')

[fk,xk] = ksdensity(MPG_Germany);
plot(xk,fk,'k:')

legend('USA','Japan','Germany')
title('MPG by Origin')
xlabel('MPG')
hold off
```



The plot shows how miles per gallon (`MPG`) performance differs by country of origin (`Origin`). Using this data, the USA has the widest distribution, and its peak is at the lowest `MPG` value of the three origins. Japan has the most regular distribution with a slightly heavier left tail, and its peak is at the highest `MPG` value of the three origins. The peak for Germany is between the USA and Japan, and the second bump near 44 miles per gallon suggests that there might be multiple modes in the data.

See Also

`KernelDistribution` | `fitdist` | `ksdensity`

More About

- “Kernel Distribution” on page B-78
- “Grouping Variables” on page 2-45
- “Fit Kernel Distribution Using ksdensity” on page 5-39
- “Fit Probability Distribution Objects to Grouped Data” on page 5-92

Fit a Nonparametric Distribution with Pareto Tails

This example shows how to fit a nonparametric probability distribution to sample data using Pareto tails to smooth the distribution in the tails.

Step 1. Generate sample data.

Generate sample data that contains more outliers than expected from a standard normal distribution.

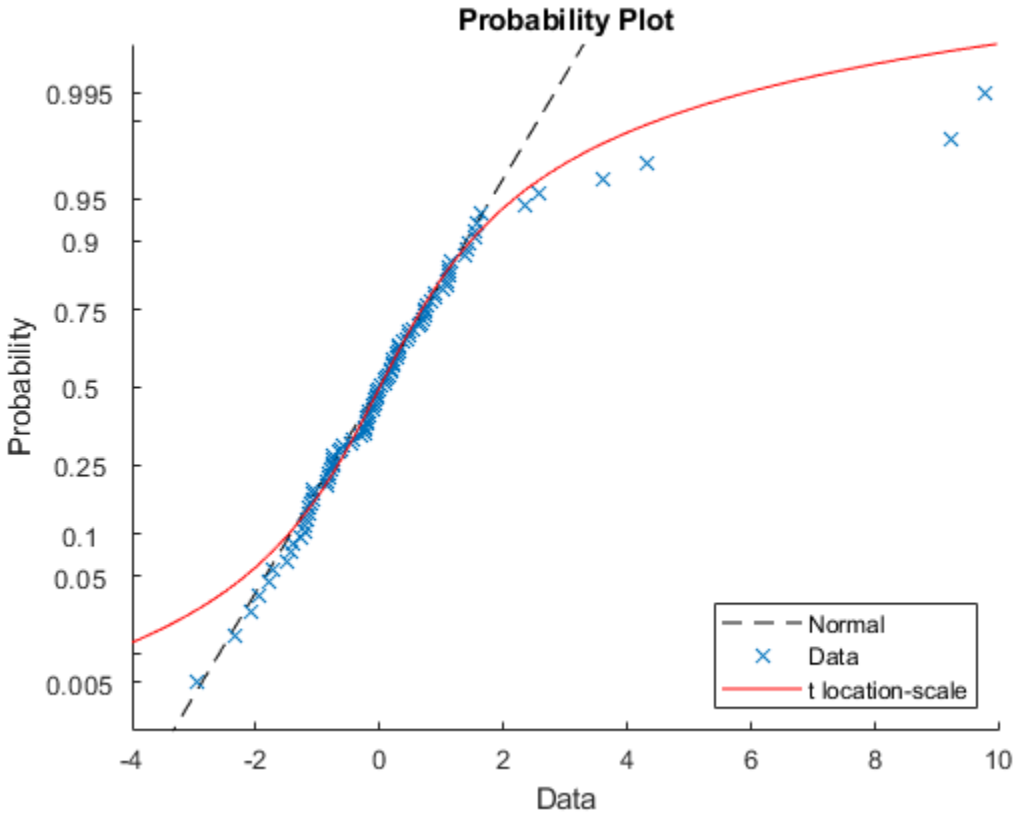
```
rng('default')  % For reproducibility
left_tail = -exprnd(1,10,1);
right_tail = exprnd(5,10,1);
center = randn(80,1);
data = [left_tail;center;right_tail];
```

The data contains 80% values from a standard normal distribution, 10% from an exponential distribution with a mean of 5, and 10% from an exponential distribution with mean of -1. Compared to a standard normal distribution, the exponential values are more likely to be outliers, especially in the upper tail.

Step 2. Fit probability distributions to the data.

Fit a normal distribution and a *t* location-scale distribution to the data, and plot for a visual comparison.

```
probplot(data);
p = fitdist(data,'tlocationscale');
h = probplot(gca,p);
set(h,'color','r','linestyle','-');
title('Probability Plot')
legend('Normal','Data','t location-scale','Location','SE')
```

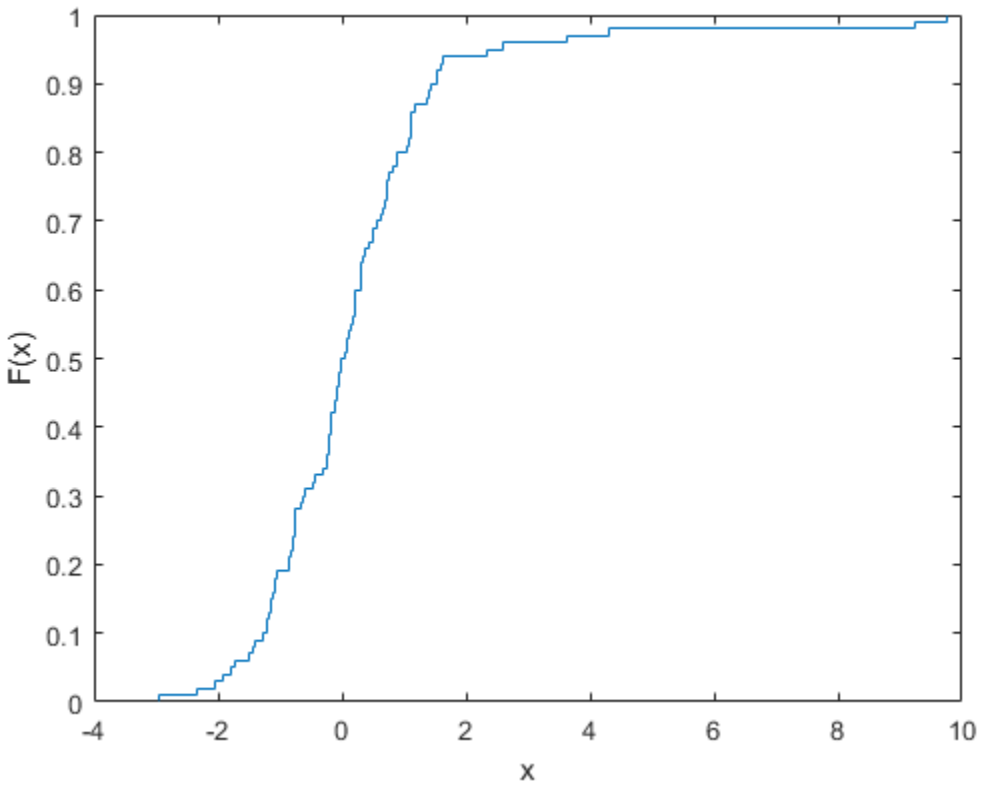


Both distributions appear to fit reasonably well in the center, but neither the normal distribution nor the *t* location-scale distribution fit the tails very well.

Step 3. Generate an empirical distribution.

To obtain a better fit, use `ecdf` to generate an empirical cdf based on the sample data.

```
figure
ecdf(data)
```



The empirical distribution provides a perfect fit, but the outliers make the tails very discrete. Random samples generated from this distribution using the inversion method might include, for example, values near 4.33 and 9.25, but no values in between.

Step 4. Fit a distribution using Pareto tails.

Use `paretotails` to generate an empirical cdf for the middle 80% of the data and fit generalized Pareto distributions to the lower and upper 10%.

```
pfit = paretotails(data,0.1,0.9)
```

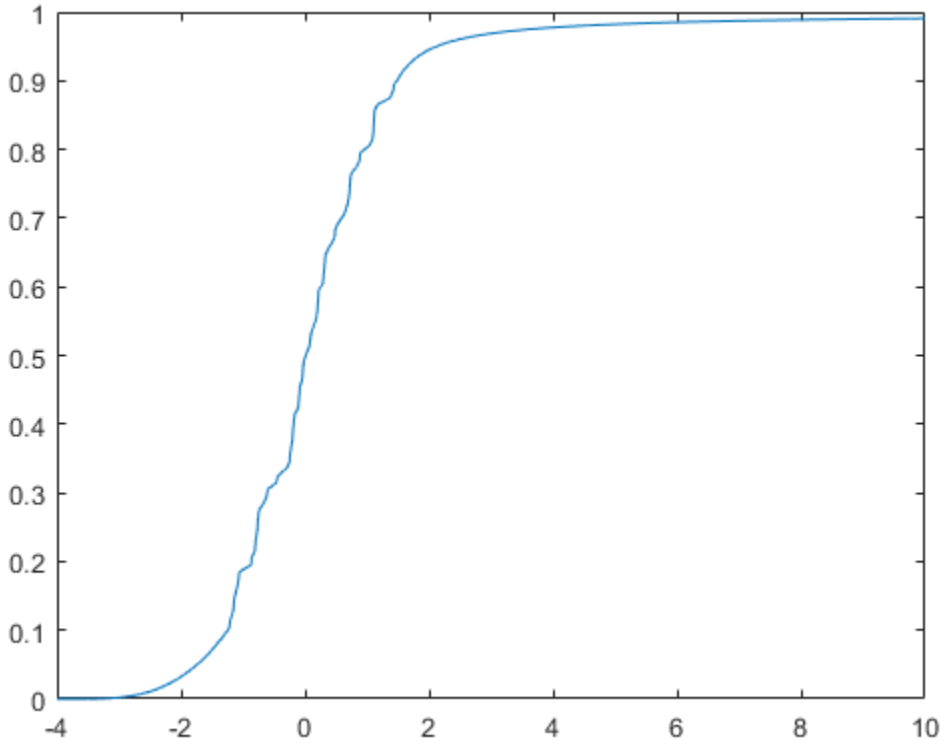
```
pfit =
Piecewise distribution with 3 segments
      -Inf < x < -1.24623    (0 < p < 0.1): lower tail, GPD(-0.334156,0.798745)
   -1.24623 < x < 1.48551  (0.1 < p < 0.9): interpolated empirical cdf
       1.48551 < x < Inf    (0.9 < p < 1): upper tail, GPD(1.23681,0.581868)
```

To obtain a better fit, `paretotails` fits a distribution by piecing together an ecdf or kernel distribution in the center of the sample, and smooth generalized Pareto distributions (GPDs) in the tails. Use `paretotails` to create `paretotails` probability distribution object. You can access information about the fit and perform further calculations on the object using the object functions of the `paretotails` object. For example, you can evaluate the cdf or generate random numbers from the distribution.

Step 5. Compute and plot the cdf.

Compute and plot the cdf of the fitted `paretotails` distribution.

```
x = -4:0.01:10;
plot(x,cdf(pfit,x))
```



The `paretotails` cdf closely fits the data but is smoother in the tails than the ecdf generated in Step 3.

See Also

`ecdf` | `fitdist` | `paretotails`

More About

- “Nonparametric and Empirical Probability Distributions” on page 5-30

Generate Random Numbers Using the Triangular Distribution

This example shows how to create a triangular probability distribution object based on sample data, and generate random numbers for use in a simulation.

Step 1. Input sample data.

Input the data vector `time`, which contains the observed length of time (in seconds) that 10 different cars stopped at a highway tollbooth.

```
time = [6 14 8 7 16 8 23 6 7 15];
```

The data shows that, while most cars stopped for 6 to 16 seconds, one outlier stopped for 23 seconds.

Step 2. Estimate distribution parameters.

Estimate the triangular distribution parameters from the sample data.

```
lower = min(time);
peak = median(time);
upper = max(time);
```

A triangular distribution provides a simplistic representation of the probability distribution when sample data is limited. Estimate the lower and upper boundaries of the distribution by finding the minimum and maximum values of the sample data. For the peak parameter, the median might provide a better estimate of the mode than the mean, since the data includes an outlier.

Step 3. Create a probability distribution object.

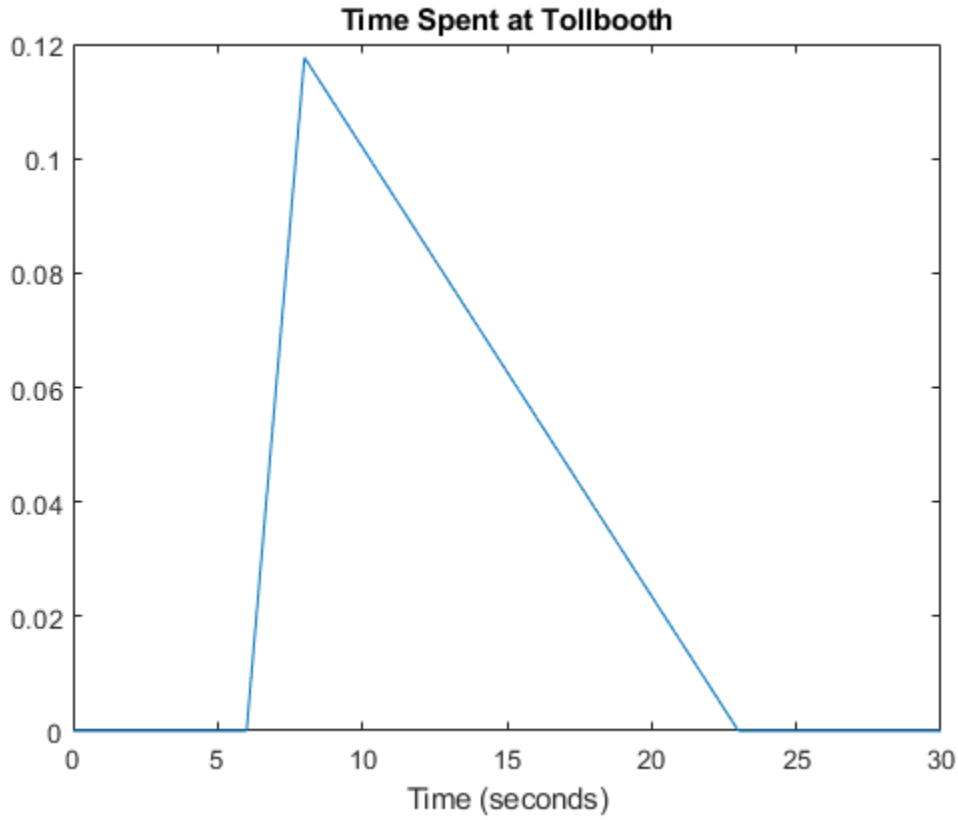
Create a triangular probability distribution object using the estimated parameter values.

```
pd = makedist('Triangular','a',lower,'b',peak,'c',upper)
```

```
pd =
  TriangularDistribution

A = 6, B = 8, C = 23
```

Compute and plot the pdf of the triangular distribution.

```
x = 0:.1:230;
y = pdf(pd,x);
plot(x,y)
title('Time Spent at Tollbooth')
xlabel('Time (seconds)')
xlim([0 30])
```



The plot shows that this triangular distribution is skewed to the right. However, since the estimated peak value is the sample median, the distribution should be symmetrical about the peak. Because of its skew, this model might, for example, generate random numbers that seem unusually high when compared to the initial sample data.

Step 4. Generate random numbers.

Generate random numbers from this distribution to simulate future traffic flow through the tollbooth.

```
rng('default');  % For reproducibility
r = random(pd,10,1)
```

```
r = 10×1

   16.1265
   18.0987
    8.0796
   18.3001
   13.3176
    7.8211
    9.4360
   12.2508
   19.7082
   20.0078
```

The returned values in `r` are the time in seconds that the next 10 simulated cars spend at the tollbooth. These values seem high compared to the values in the original data vector `time` because

the outlier skewed the distribution to the right. Using the second-highest value as the upper limit parameter might mitigate the effects of the outlier and generate a set of random numbers more similar to the initial sample data.

Step 5. Revise estimated parameters.

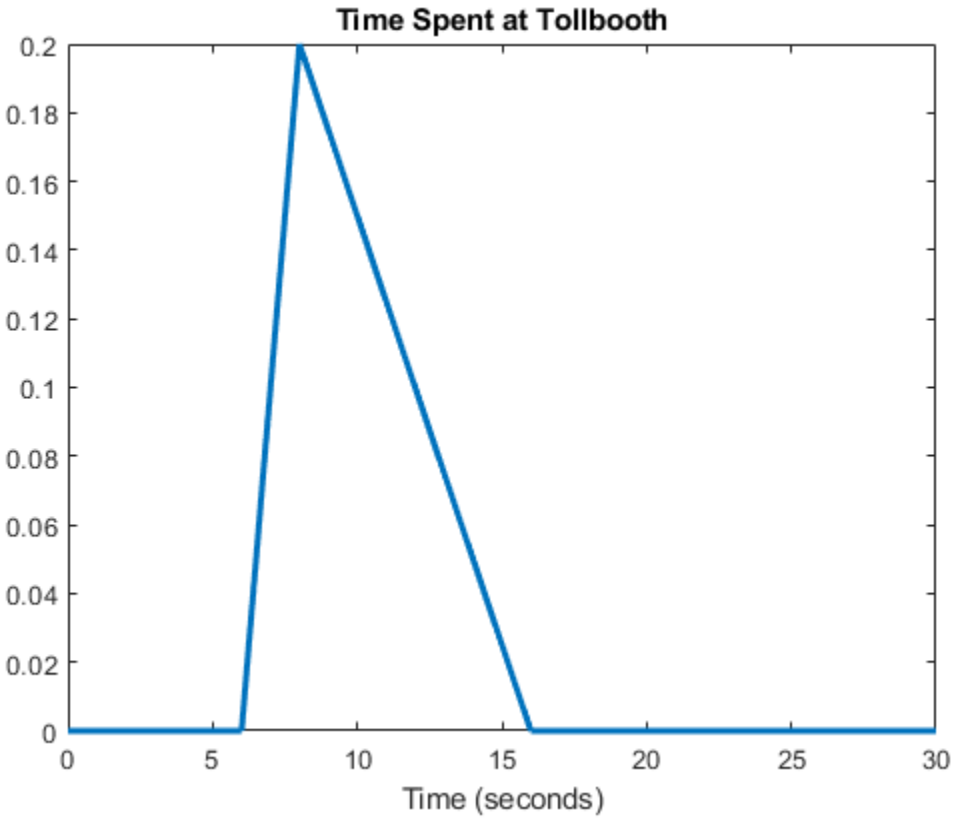
Estimate the upper boundary of the distribution using the second largest value in the sample data.

```
sort_time = sort(time,'descend');
secondLargest = sort_time(2);
```

Step 6. Create a new distribution object and plot the pdf.

Create a new triangular probability distribution object using the revised estimated parameters, and plot its pdf.

```
figure
pd2 = makedist('Triangular','a',lower,'b',peak,'c',secondLargest);
y2 = pdf(pd2,x);
plot(x,y2,'LineWidth',2)
title('Time Spent at Tollbooth')
xlabel('Time (seconds)')
xlim([0 30])
```



The plot shows that this triangular distribution is still slightly skewed to the right. However, it is much more symmetrical about the peak than the distribution that used the maximum sample data value to estimate the upper limit.

Step 7. Generate new random numbers.

Generate new random numbers from the revised distribution.

```
rng('default'); % For reproducibility
r2 = random(pd2,10,1)
```

```
r2 = 10×1

   12.1501
   13.2547
    7.5937
   13.3675
   10.5768
    7.3967
    8.4026
    9.9792
   14.1562
   14.3240
```

These new values more closely resemble those in the original data vector `time`. They are also closer to the sample median than the random numbers generated by the distribution that used the outlier to estimate its upper limit. This example does not remove the outlier from the sample data when computing the median. Other options for parameter estimation include removing outliers from the sample data altogether, or using the mean or mode of the sample data as the peak value.

See Also

`makedist` | `pdf` | `random`

More About

- “Triangular Distribution” on page B-158
- “Random Number Generation” on page 5-27
- “Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101

Model Data Using the Distribution Fitter App

In this section...

The Distribution Fitter app provides a visual, interactive approach to fitting univariate distributions to data.

Explore Probability Distributions Interactively

You can use the Distribution Fitter app to interactively fit probability distributions to data imported from the MATLAB workspace. You can choose from 22 built-in probability distributions, or create your own custom distribution. The app displays the fitted distribution over plots of the empirical distributions, including pdf, cdf, probability plots, and survivor functions. You can export the fit data, including fitted parameter values, to the workspace for further analysis.

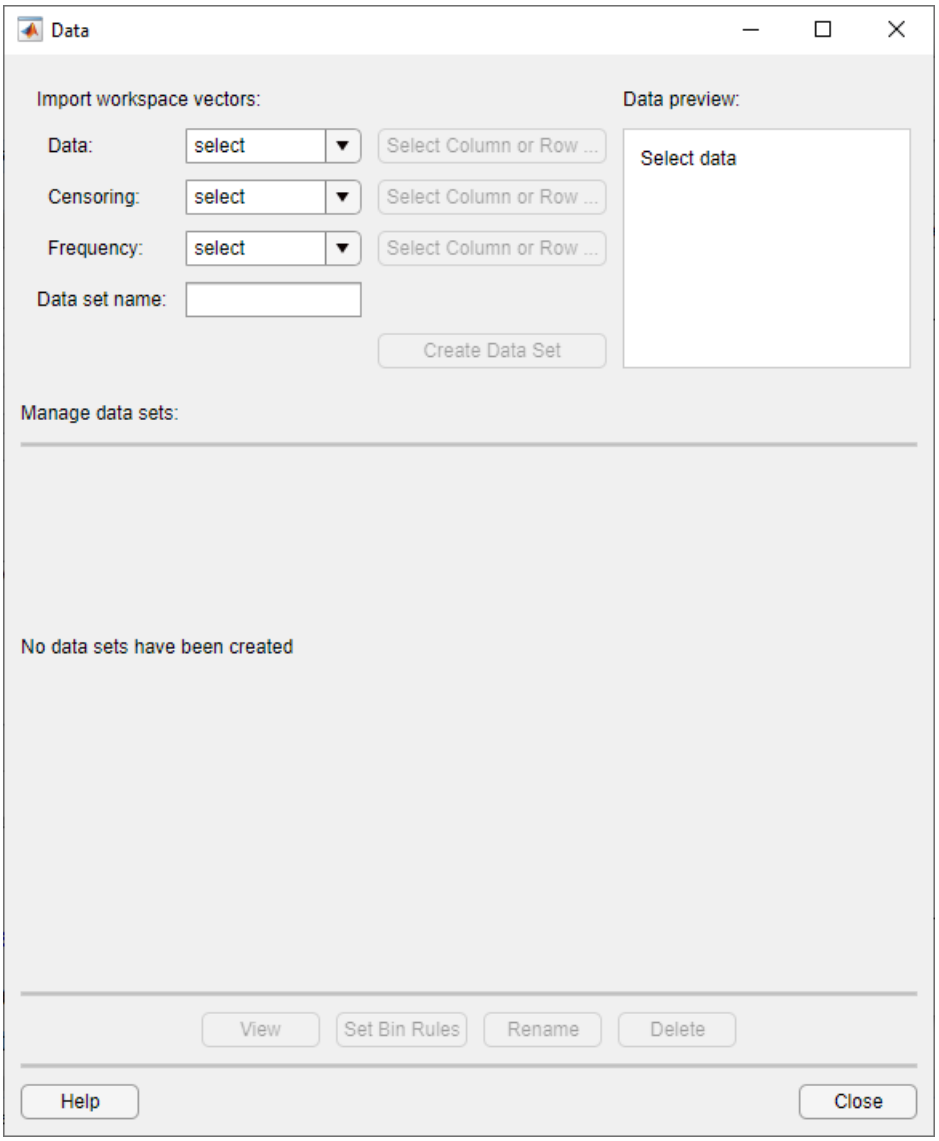
Distribution Fitter App Workflow

To fit a probability distribution to your sample data:

1. On the MATLAB Toolstrip, click the Apps tab. In the Math, Statistics and Optimization group, open the Distribution Fitter app. Alternatively, at the command prompt, enter `distributionFitter`.
2. Import your sample data, or create a data vector directly in the app. You can also manage your data sets and choose which one to fit. See “Create and Manage Data Sets” on page 5-52.
3. Create a new fit for your data. See “Create a New Fit” on page 5-55.
4. Display the results of the fit. You can choose to display the density (pdf), cumulative probability (cdf), quantile (inverse cdf), probability plot (choose one of several distributions), survivor function, and cumulative hazard. See “Display Results” on page 5-59.
5. You can create additional fits, and manage multiple fits from within the app. See “Manage Fits” on page 5-60.
6. Evaluate probability functions for the fit. You can choose to evaluate the density (pdf), cumulative probability (cdf), quantile (inverse cdf), survivor function, and cumulative hazard. See “Evaluate Fits” on page 5-62.
7. Improve the fit by excluding certain data. You can specify bounds for the data to exclude, or you can exclude data graphically using a plot of the values in the sample data. See “Exclude Data” on page 5-64.
8. Save your current Distribution Fitter app session so you can open it later. See “Save and Load Sessions” on page 5-68.

Create and Manage Data Sets

To open the Data dialog box, click the **Data** button in the Distribution Fitter app.



Import Data

Create a data set by importing a vector from the MATLAB workspace using the **Import workspace vectors** pane.

- **Data** — In the **Data** field, the drop-down list contains the names of all matrices and vectors, other than 1-by-1 matrices (scalars) in the MATLAB workspace. Select the array containing the data that you want to fit. The actual data you import must be a vector. If you select a matrix in the **Data** field, the first column of the matrix is imported by default. To select a different column or row of the matrix, click **Select Column or Row**. The matrix displays in the Variables editor. You can select a row or column by highlighting it.

  Alternatively, you can enter any valid MATLAB expression in the **Data** field.

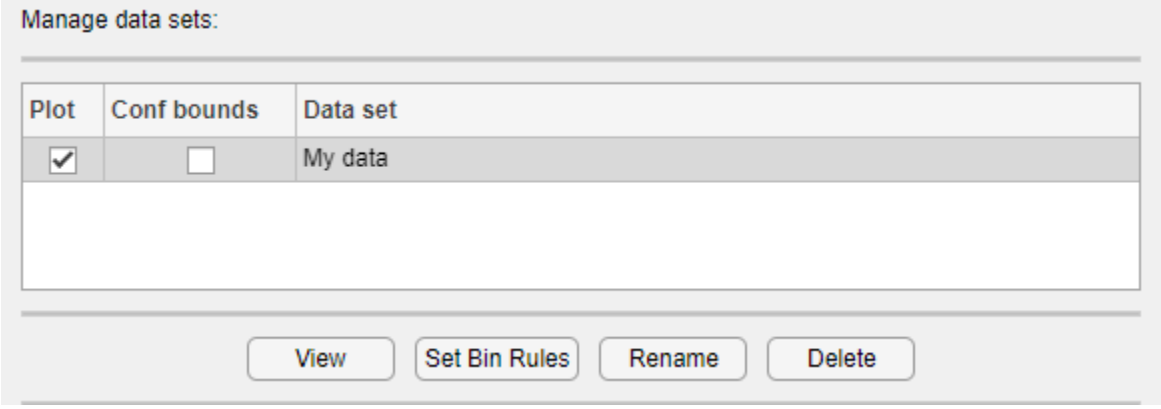
When you select a vector in the **Data** field, a histogram of the data appears in the **Data preview** pane.

- **Censoring** — If some of the points in the data set are censored, enter a Boolean vector of the same size as the data vector, specifying the censored entries of the data. A `1` in the censoring vector specifies that the corresponding entry of the data vector is censored. A `0` specifies that the entry is not censored. If you enter a matrix, you can select a column or row by clicking **Select Column or Row**. If you do not have censored data, leave the **Censoring** field blank.
- **Frequency** — Enter a vector of positive integers of the same size as the data vector to specify the frequency of the corresponding entries of the data vector. For example, a value of 7 in the 15th entry of frequency vector specifies that there are 7 data points corresponding to the value in the 15th entry of the data vector. If all entries of the data vector have frequency 1, leave the **Frequency** field blank.
- **Data set name** — Enter a name for the data set that you import from the workspace, such as `My data`.

After you have entered the information in the preceding fields, click **Create Data Set** to create the data set `My data`.

Manage Data Sets

View and manage the data sets that you create using the **Manage data sets** pane. When you create a data set, its name appears in the **Data sets** list. The following figure shows the **Manage data sets** pane after creating the data set `My data`.



For each data set in the **Data sets** list, you can:

- Select the **Plot** check box to display a plot of the data in the main Distribution Fitter app window. When you create a new data set, **Plot** is selected by default. Clearing the **Plot** check box removes the data from the plot in the main window. You can specify the type of plot displayed in the **Display type** field in the main window.
- If **Plot** is selected, you can also select **Bounds** to display confidence interval bounds for the plot in the main window. These bounds are pointwise confidence bounds around the empirical estimates of these functions. The bounds are displayed only when you set **Display Type** in the main window to one of the following:
  - `Cumulative probability (CDF)`
  - `Survivor function`

- `Cumulative hazard`

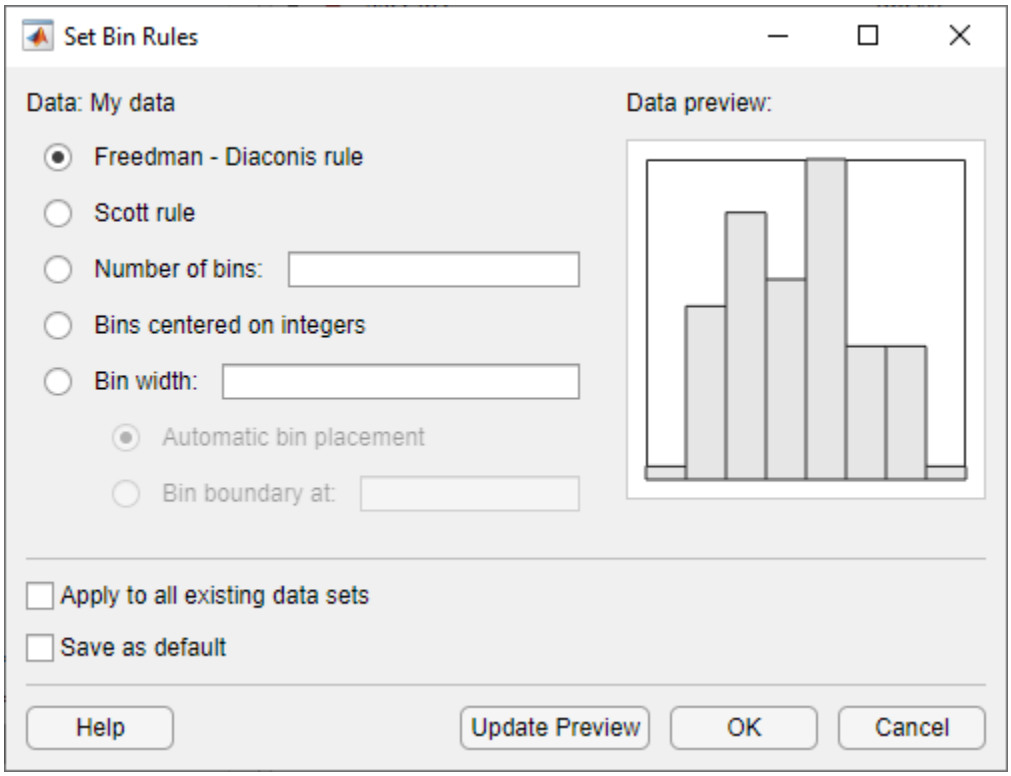
The Distribution Fitter app cannot display confidence bounds on density (`PDF`), quantile (`inverse CDF`), or probability plots. Clearing the **Bounds** check box removes the confidence bounds from the plot in the main window.

When you select a data set from the list, you can access the following buttons:

- **View** — Display the data in a table in a new window.
- **Set Bin Rules** — Defines the histogram bins used in a density (PDF) plot.
- **Rename** — Rename the data set.
- **Delete** — Delete the data set.

Set Bin Rules

To set bin rules for the histogram of a data set, click **Set Bin Rules** to open the **Set Bin Width Rules** dialog box.



You can select from the following rules:

- **Freedman-Diaconis rule** — Algorithm that chooses bin widths and locations automatically, based on the sample size and the spread of the data. This rule, which is the default, is suitable for many kinds of data.
- **Scott rule** — Algorithm intended for data that are approximately normal. The algorithm chooses bin widths and locations automatically.
- **Number of bins** — Enter the number of bins. All bins have equal widths.

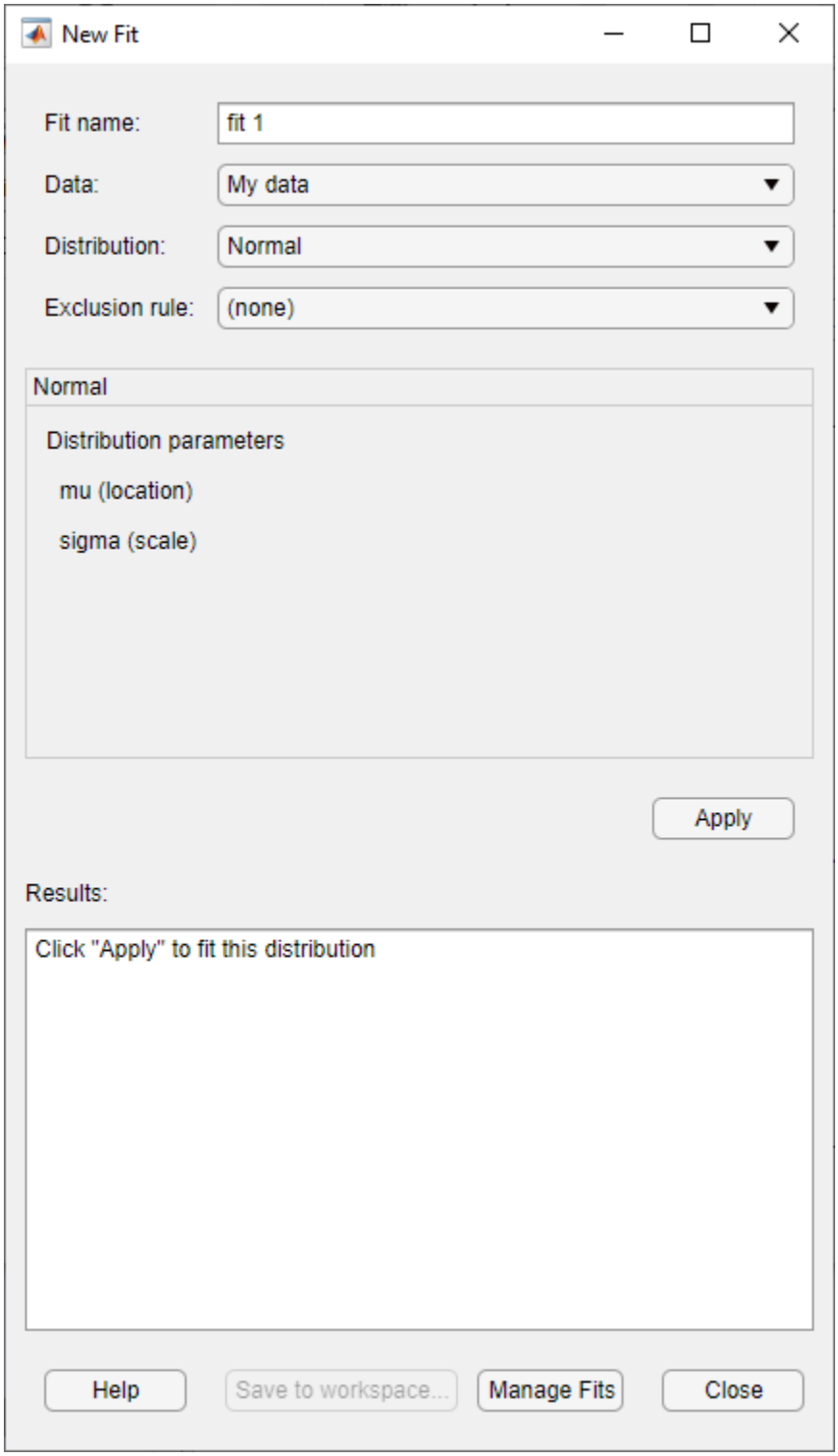
- **Bins centered on integers** — Specifies bins centered on integers.
- **Bin width** — Enter the width of each bin. If you select this option, you can also select:
  - **Automatic bin placement** — Place the edges of the bins at integer multiples of the **Bin width**.
  - **Bin boundary at** — Enter a scalar to specify the boundaries of the bins. The boundary of each bin is equal to this scalar plus an integer multiple of the **Bin width**.

You can also:

- **Apply to all existing data sets** — Apply the rule to all data sets. Otherwise, the rule is applied only to the data set currently selected in the Data dialog box.
- **Save as default** — Apply the current rule to any new data sets that you create. You can set default bin width rules by selecting **Set Default Bin Rules** from the **Tools** menu in the main window.

Create a New Fit

Click the **New Fit** button at the top of the main window to open the New Fit dialog box. If you created the data set `My data`, it appears in the **Data** field.



Field Name	Description
Fit Name	Enter a name for the fit.

Field Name	Description
Data	Select the data set to which you want to fit a distribution from the drop-down list.
Distribution	Select the type of distribution to fit from the **Distribution** drop-down list. Only the distributions that apply to the values of the selected data set appear in the **Distribution** field. For example, when the data include values that are zero or negative, positive distributions are not displayed . You can specify either a parametric or a nonparametric distribution. When you select a parametric distribution from the drop-down list, a description of its parameters appears. Distribution Fitter estimates these parameters to fit the distribution to the data set. If you select the binomial distribution or the generalized extreme value distribution, you must specify a fixed value for one of the parameters. The pane contains a text field into which you can specify that parameter. When you select `Nonparametric fit`, options for the fit appear in the pane, as described in “Further Options for Nonparametric Fits” on page 5-58.
Exclusion rule	Specify a rule to exclude some data. Create an exclusion rule by clicking **Exclude** in the Distribution Fitter app. For more information, see “Exclude Data” on page 5-64.

Apply the New Fit

Click **Apply** to fit the distribution. For a parametric fit, the **Results** pane displays the values of the estimated parameters. For a nonparametric fit, the **Results** pane displays information about the fit.

When you click **Apply**, the Distribution Fitter app displays a plot of the distribution and the corresponding data.

Note When you click **Apply**, the title of the dialog box changes to Edit Fit. You can now make changes to the fit you just created and click **Apply** again to save them. After closing the Edit Fit dialog box, you can reopen it from the Fit Manager dialog box at any time to edit the fit.

After applying the fit, you can save the information to the workspace using probability distribution objects by clicking **Save to workspace**.

Available Distributions

All of the distributions available in the Distribution Fitter app are supported elsewhere in Statistics and Machine Learning Toolbox software. You can use the `fitdist` function to fit any of the distributions supported by the app. Many distributions also have dedicated fitting functions. These functions compute the majority of the fits in the Distribution Fitter app, and are referenced in the following list. Other fits are computed using functions internal to the Distribution Fitter app.

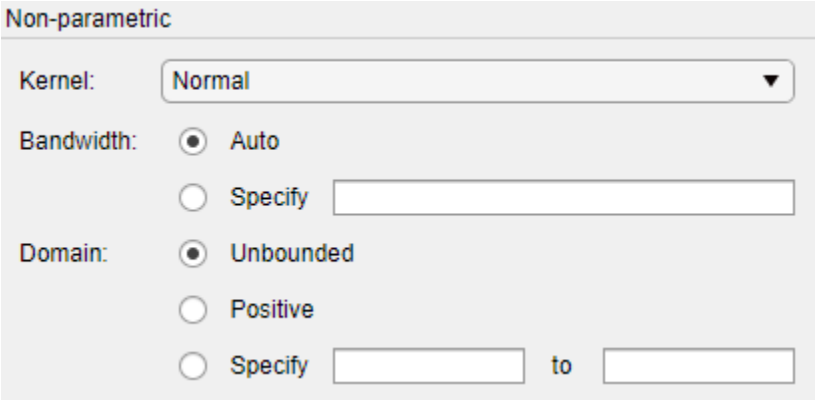
Not all of the distributions listed are available for all data sets. The Distribution Fitter app determines the extent of the data (nonnegative, unit interval, etc.) and displays appropriate distributions in the **Distribution** drop-down list. Distribution data ranges are given parenthetically in the following list.

- Beta on page B-6 (unit interval values) distribution, fit using the function `betafit`.

- Binomial on page B-10 (nonnegative integer values) distribution, fit using the function `binopdf`.
- Birnbaum-Saunders on page B-18 (positive values) distribution.
- Burr Type XII on page B-19 (positive values) distribution.
- Exponential on page B-33 (nonnegative values) distribution, fit using the function `expfit`.
- Extreme value on page B-40 (all values) distribution, fit using the function `evfit`.
- Gamma on page B-47 (positive values) distribution, fit using the function `gamfit`.
- Generalized extreme value on page B-55 (all values) distribution, fit using the function `gevfit`.
- Generalized Pareto on page B-59 (all values) distribution, fit using the function `gpfit`.
- Inverse Gaussian on page B-75 (positive values) distribution.
- Logistic on page B-85 (all values) distribution.
- Loglogistic on page B-86 (positive values) distribution.
- Lognormal on page B-88 (positive values) distribution, fit using the function `lognfit`.
- Nakagami on page B-108 (positive values) distribution.
- Negative binomial on page B-109 (nonnegative integer values) distribution, fit using the function `nbinpdf`.
- Nonparametric on page B-78 (all values) distribution, fit using the function `ksdensity`.
- Normal on page B-119 (all values) distribution, fit using the function `normfit`.
- Poisson on page B-131 (nonnegative integer values) distribution, fit using the function `poisspdf`.
- Rayleigh on page B-137 (positive values) distribution using the function `raylfit`.
- Rician on page B-139 (positive values) distribution.
- *t* location-scale on page B-156 (all values) distribution.
- Weibull on page B-170 (positive values) distribution using the function `wblfit`.

Further Options for Nonparametric Fits

When you select `Non-parametric` in the **Distribution** field, a set of options appears in the **Non-parametric** pane, as shown in the following figure.



The options for nonparametric distributions are:

- **Kernel** — Type of kernel function to use.

  - `Normal`
  - `Box`
  - `Triangle`
  - `Epanechnikov`
- **Bandwidth** — The bandwidth of the kernel smoothing window. Select **Auto** for a default value that is optimal for estimating normal densities. After you click **Apply**, this value appears in the **Fit results** pane. Select **Specify** and enter a smaller value to reveal features such as multiple modes or a larger value to make the fit smoother.
- **Domain** — The allowed *x*-values for the density.
  - **Unbounded** — The density extends over the whole real line.
  - **Positive** — The density is restricted to positive values.
  - **Specify** — Enter lower and upper bounds for the domain of the density.

  When you select **Positive** or **Specify**, the nonparametric fit has zero probability outside the specified domain.

Display Results

The Distribution Fitter app window displays plots of:

- The data sets for which you select **Plot** in the Data dialog box.
- The fits for which you select **Plot** in the Fit Manager dialog box.
- Confidence bounds for:
  - The data sets for which you select **Bounds** in the Data dialog box.
  - The fits for which you select **Bounds** in the Fit Manager dialog box.

The following fields are available.

Display Type

Specify the type of plot to display using the **Display Type** field in the main app window. Each type corresponds to a probability function, for example, a probability density function. You can choose from the following display types:

- `Density (PDF)` — Display a probability density function (PDF) plot for the fitted distribution. The main window displays data sets using a probability histogram, in which the height of each rectangle is the fraction of data points that lie in the bin divided by the width of the bin. This makes the sum of the areas of the rectangles equal to 1.
- `Cumulative probability (CDF)` — Display a cumulative probability plot of the data. The main window displays data sets using a cumulative probability step function. The height of each step is the cumulative sum of the heights of the rectangles in the probability histogram.
- `Quantile (inverse CDF)` — Display a quantile (inverse CDF) plot.
- `Probability plot` — Display a probability plot of the data. Specify the type of distribution used to construct the probability plot in the **Distribution** field. This field is only available when you select `Probability plot`. The choices for the distribution are:
  - `Exponential`

- `Extreme value`
- `Logistic`
- `Log-Logistic`
- `Lognormal`
- `Normal`
- `Rayleigh`
- `Weibull`

You can also create a probability plot against a parametric fit that you create in the **New Fit** pane. When you create these fits, they are added at the bottom of the **Distribution** drop-down list.

- `Survivor function` — Display survivor function plot of the data.
- `Cumulative hazard` — Display cumulative hazard plot of the data.

Note If the plotted data includes `0` or negative values, some distributions are unavailable.

Confidence Bounds

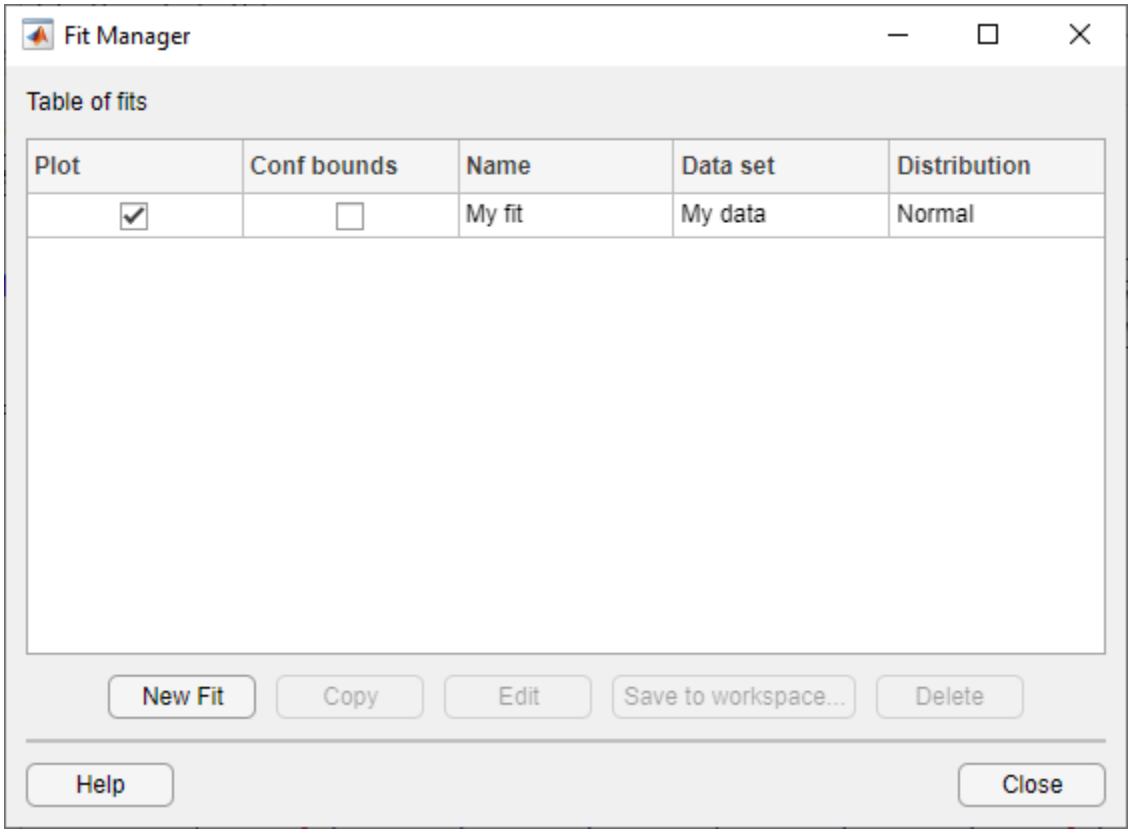
You can display confidence bounds for data sets and fits when you set **Display Type** to `Cumulative probability (CDF)`, `Survivor function`, `Cumulative hazard`, or, for fits only, `Quantile (inverse CDF)`.

- To display bounds for a data set, select **Bounds** next to the data set in the **Data sets** pane of the Data dialog box.
- To display bounds for a fit, select **Bounds** next to the fit in the Fit Manager dialog box. Confidence bounds are not available for all fit types.

To set the confidence level for the bounds, select `Confidence Level` from the **View** menu in the main window and choose from the options.

Manage Fits

Click the **Manage Fits** button to open the **Fit Manager** dialog box.



The **Table of fits** displays a list of the fits that you create, with the following options:

- **Plot** — Displays a plot of the fit in the main window of the Distribution Fitter app. When you create a new fit, **Plot** is selected by default. Clearing the **Plot** check box removes the fit from the plot in the main window.
- **Bounds** — If you select **Plot**, you can also select **Bounds** to display confidence bounds in the plot. The bounds are displayed when you set **Display Type** in the main window to one of the following:

  - `Cumulative probability (CDF)`
  - `Quantile (inverse CDF)`
  - `Survivor function`
  - `Cumulative hazard`

  The Distribution Fitter app cannot display confidence bounds on density (`PDF`) or probability plots. Bounds are not supported for nonparametric fits and some parametric fits.

  Clearing the **Bounds** check box removes the confidence intervals from the plot in the main window.

  When you select a fit in the **Table of fits**, the following buttons are enabled below the table:

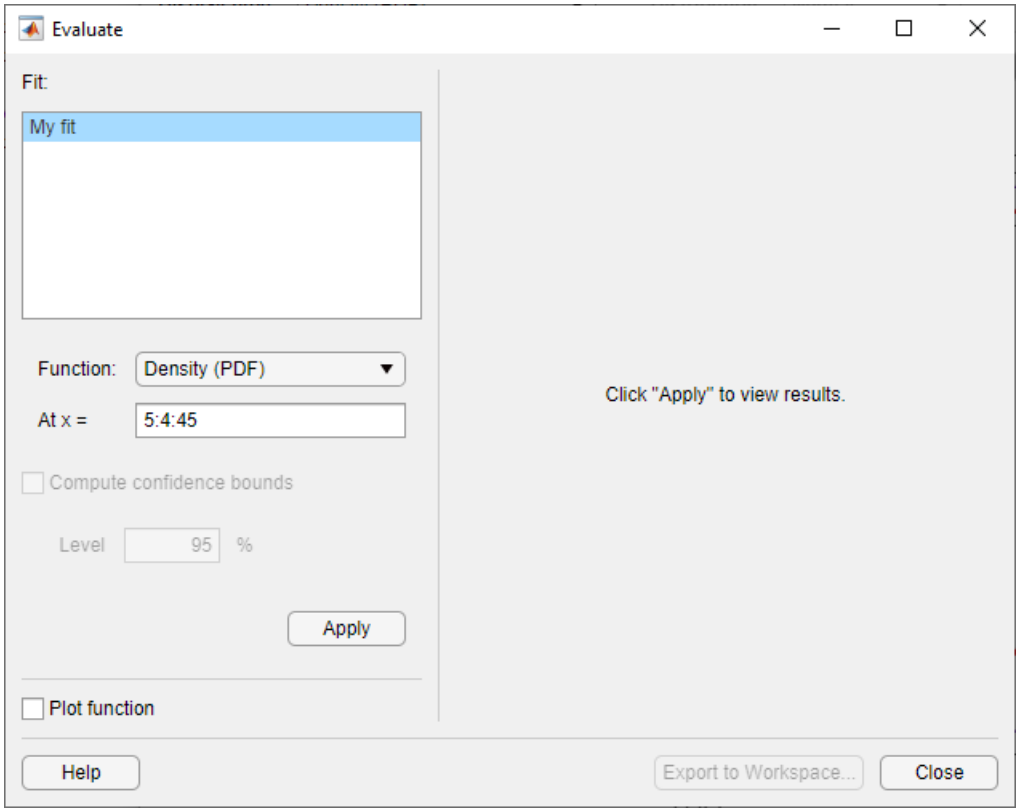
  - **New Fit** — Open a New Fit window.
  - **Copy** — Create a copy of the selected fit.
  - **Edit** — Open an Edit Fit dialog box, to edit the fit.

Note You can edit only the currently selected fit in the Edit Fit dialog box. To edit a different fit, select it in the **Table of fits** and click **Edit** to open another Edit Fit dialog box.

- **Save to workspace** — Save the selected fit as a distribution object.
- **Delete** — Delete the selected fit.

Evaluate Fits

Use the **Evaluate** dialog box to evaluate your fitted distribution at any data points you choose. To open the dialog box, click the **Evaluate** button.



In the **Evaluate** dialog box, choose from the following items:

- **Fit** pane — Display the names of existing fits. Select one or more fits that you want to evaluate. Using your platform specific functionality, you can select multiple fits.
- **Function** — Select the type of probability function that you want to evaluate for the fit. The available functions are:
  - `Density (PDF)` — Computes a probability density function.
  - `Cumulative probability (CDF)` — Computes a cumulative probability function.
  - `Quantile (inverse CDF)` — Computes a quantile (inverse CDF) function.
  - `Survivor function` — Computes a survivor function.
  - `Cumulative hazard` — Computes a cumulative hazard function.
  - `Hazard rate` — Computes the hazard rate.

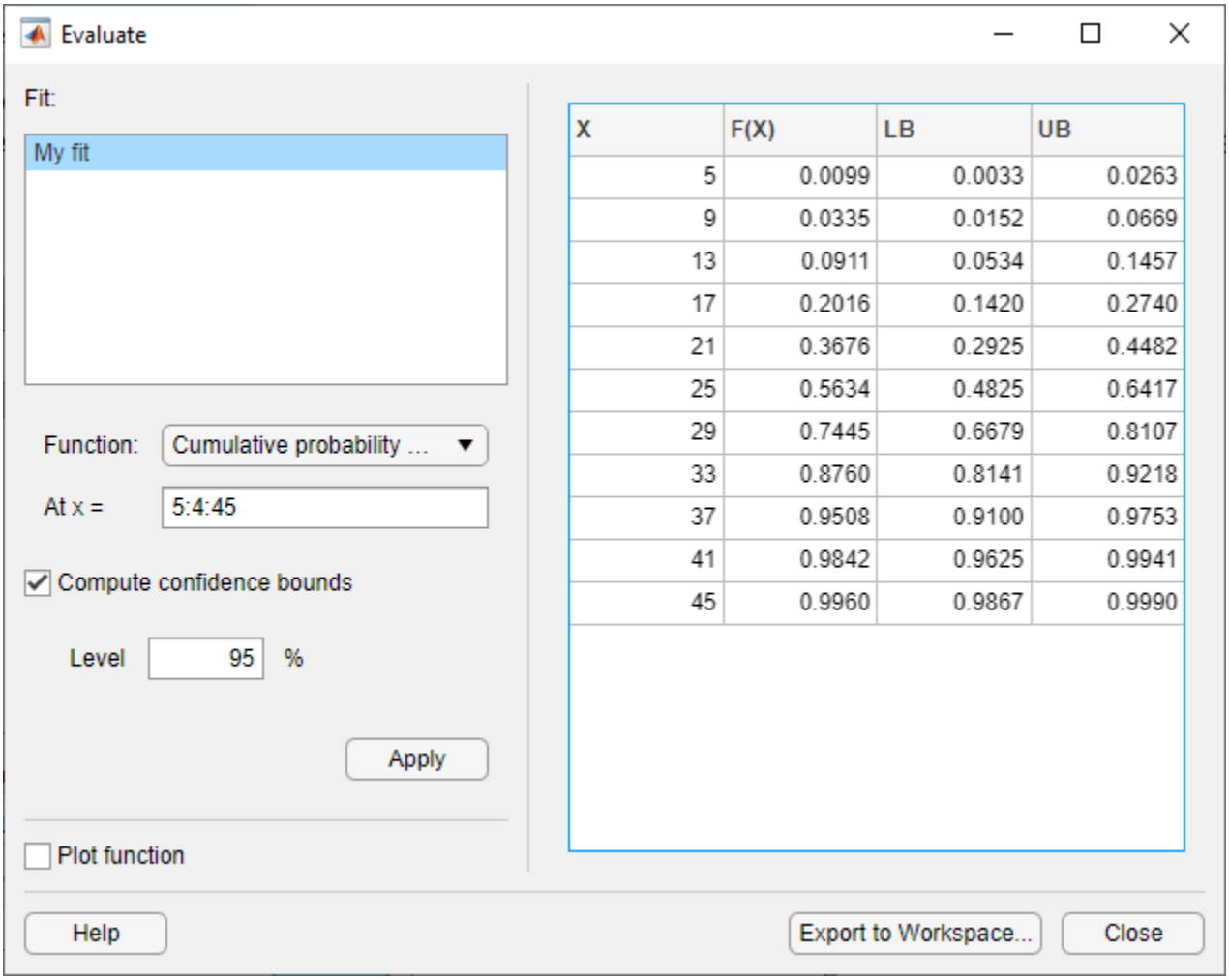
- **At x =** — Enter a vector of points or the name of a workspace variable containing a vector of points at which you want to evaluate the distribution function. If you change **Function** to `Quantile (inverse CDF)`, the field name changes to **At p =**, and you enter a vector of probability values.
- **Compute confidence bounds** — Select this box to compute confidence bounds for the selected fits. The check box is enabled only if you set **Function** to one of the following:

  - `Cumulative probability (CDF)`
  - `Quantile (inverse CDF)`
  - `Survivor function`
  - `Cumulative hazard`

  The Distribution Fitter app cannot compute confidence bounds for nonparametric fits and for some parametric fits. In these cases, it returns `NaN` for the bounds.
- **Level** — Set the level for the confidence bounds.
- **Plot function** — Select this box to display a plot of the distribution function, evaluated at the points you enter in the **At x =** field, in a new window.

Note The settings for **Compute confidence bounds**, **Level**, and **Plot function** do not affect the plots that are displayed in the main window of the Distribution Fitter app. The settings apply only to plots you create by clicking **Plot function** in the Evaluate window.

To apply these evaluation settings to the selected fit, click **Apply**. The following figure shows the results of evaluating the cumulative density function for the fit **My fit**, at the points in the vector `5:4:45`.



X	F(X)	LB	UB
5	0.0099	0.0033	0.0263
9	0.0335	0.0152	0.0669
13	0.0911	0.0534	0.1457
17	0.2016	0.1420	0.2740
21	0.3676	0.2925	0.4482
25	0.5634	0.4825	0.6417
29	0.7445	0.6679	0.8107
33	0.8760	0.8141	0.9218
37	0.9508	0.9100	0.9753
41	0.9842	0.9625	0.9941
45	0.9960	0.9867	0.9990

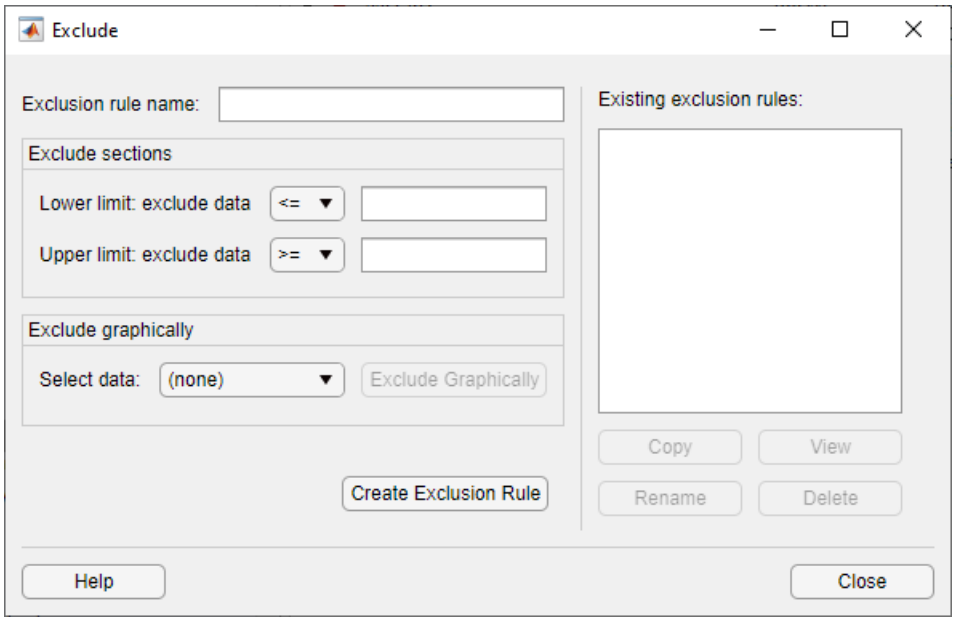
The columns of the table to the right of the **Fit** pane display the following values:

- X — The entries of the vector that you enter in **At x =** field.
- F(X)— The corresponding values of the CDF at the entries of X.
- LB — The lower bounds for the confidence interval, if you select **Compute confidence bounds**.
- UB — The upper bounds for the confidence interval, if you select **Compute confidence bounds**.

To save the data displayed in the table to a matrix in the MATLAB workspace, click **Export to Workspace**.

Exclude Data

To exclude values from fit, open the **Exclude** window by clicking the **Exclude** button. In the **Exclude** window, you can create rules for excluding specified data values. When you create a new fit in the **New Fit** window, you can use these rules to exclude data from the fit.



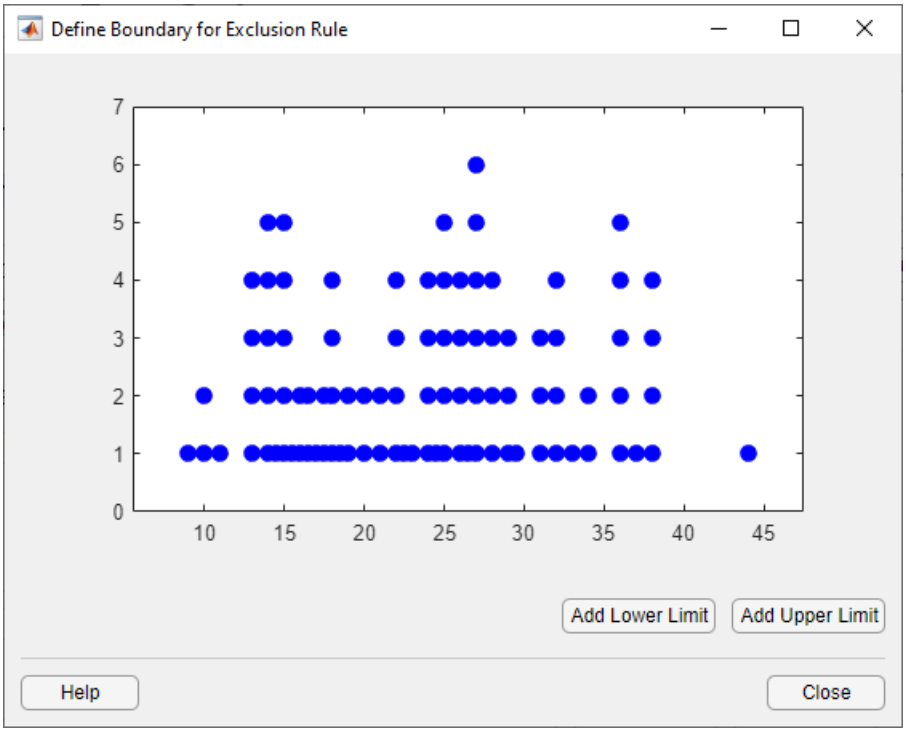
To create an exclusion rule:

1. **Exclusion Rule Name**— Enter a name for the exclusion rule.
2. **Exclude Sections**— Specify bounds for the excluded data:

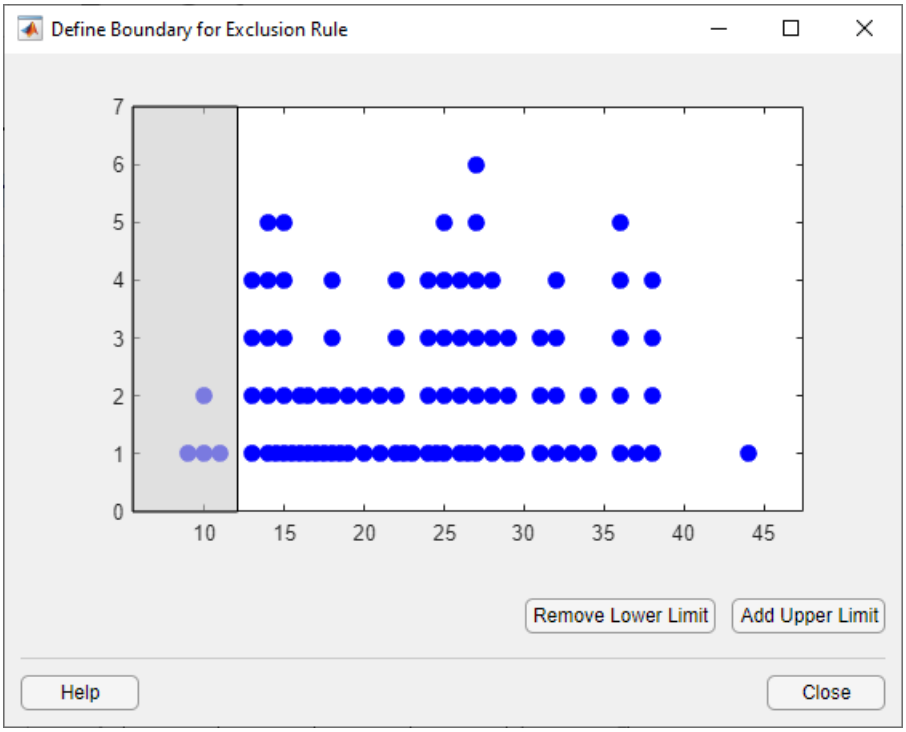
   - In the **Lower limit: exclude data** drop-down list, select <= or < and enter a scalar value in the field to the right. Depending on which operator you select, the app excludes from the fit any data values that are less than or equal to the scalar value, or less than the scalar value, respectively.
   - In the **Upper limit: exclude data** drop-down list, select >= or > and enter a scalar value in the field to the right. Depending on which operator you select, the app excludes from the fit any data values that are greater than or equal to the scalar value, or greater than the scalar value, respectively.

   OR

   Click the **Exclude Graphically** button to define the exclusion rule by displaying a plot of the values in a data set and selecting the bounds for the excluded data. For example, if you created the data set `My data` as described in Create and Manage Data Sets, select it from the drop-down list next to **Exclude graphically**, and then click the **Exclude graphically** button. The app displays the values in `My data` in a new window.



To set a lower limit for the boundary of the excluded region, click **Add Lower Limit**. The app displays a vertical line on the left side of the plot window. Move the line to the point you where you want the lower limit, as shown in the following figure.



Move the vertical line to change the value displayed in the **Lower limit: exclude data** field in the **Exclude** window.

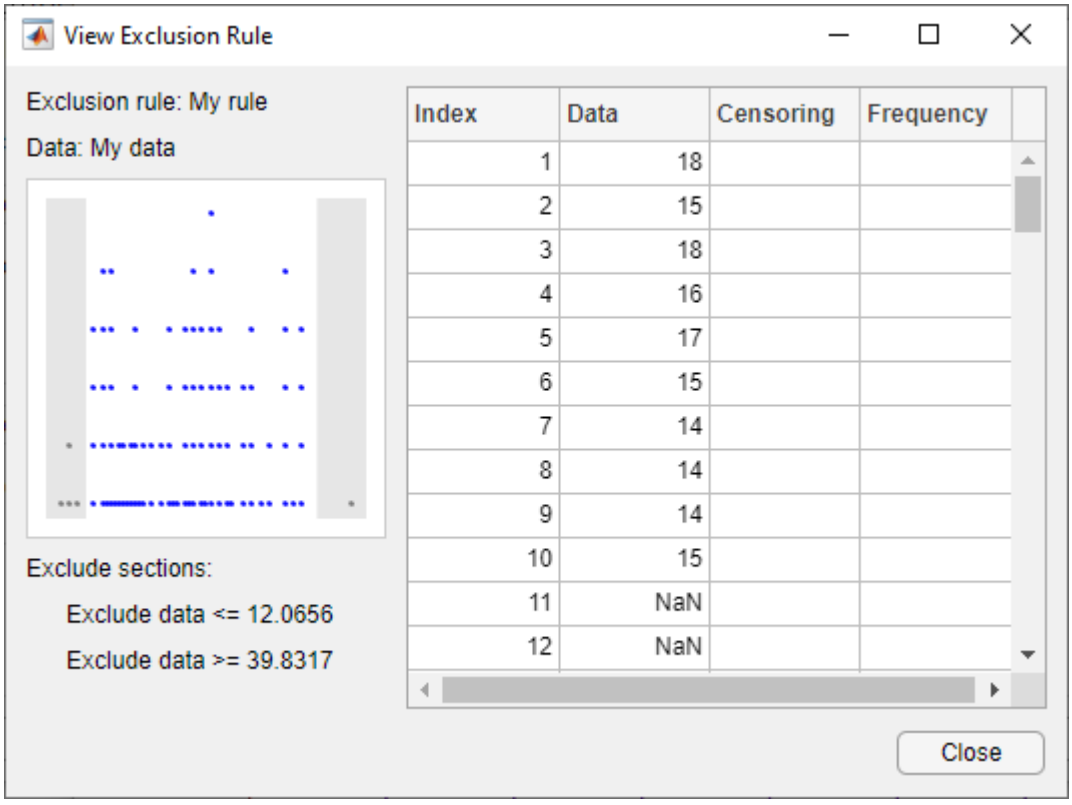
The value displayed corresponds to the *x*-coordinate of the vertical line.

Similarly, you can set the upper limit for the boundary of the excluded region by clicking **Add Upper Limit**, and then moving the vertical line that appears at the right side of the plot window. After setting the lower and upper limits, click **Close** and return to the Exclude window.

3 **Create Exclusion Rule**—Once you have set the lower and upper limits for the boundary of the excluded data, click **Create Exclusion Rule** to create the new rule. The name of the new rule appears in the **Existing exclusion rules** pane.

Selecting an exclusion rule in the **Existing exclusion rules** pane enables the following buttons:

- **Copy** — Creates a copy of the rule, which you can then modify. To save the modified rule under a different name, click **Create Exclusion Rule**.
- **View** — Opens a new window in which you can see the data points excluded by the rule. The following figure shows a typical example.



The shaded areas in the plot graphically display which data points are excluded. The table to the right lists all data points. The shaded rows indicate excluded points.

- **Rename** — Rename the rule.
- **Delete** — Delete the rule.

After you define an exclusion rule, you can use it when you fit a distribution to your data. The rule does not exclude points from the display of the data set.

Save and Load Sessions

Save your work in the current session, and then load it in a subsequent session, so that you can continue working where you left off.

Save a Session

To save the current session, from the **File** menu in the main window, select `Save Session`. A dialog box opens and prompts you to enter a file name, for example`my_session.dfit`. Click **Save** to save the following items created in the current session:

- Data sets
- Fits
- Exclusion rules
- Plot settings
- Bin width rules

Load a Session

To load a previously saved session, from the **File** menu in the main window, select `Load Session`. Enter the name of a previously saved session. Click **Open** to restore the information from the saved session to the current session.

Generate a File to Fit and Plot Distributions

Use the `Generate Code` option in the **File** to create a file that:

- Fits the distributions in the current session to any data vector in the MATLAB workspace.
- Plots the data and the fits.

After you end the current session, you can use the file to create plots in a standard MATLAB figure window, without reopening the Distribution Fitter app.

As an example, if you created the fit described in “Create a New Fit” on page 5-55, do the following steps:

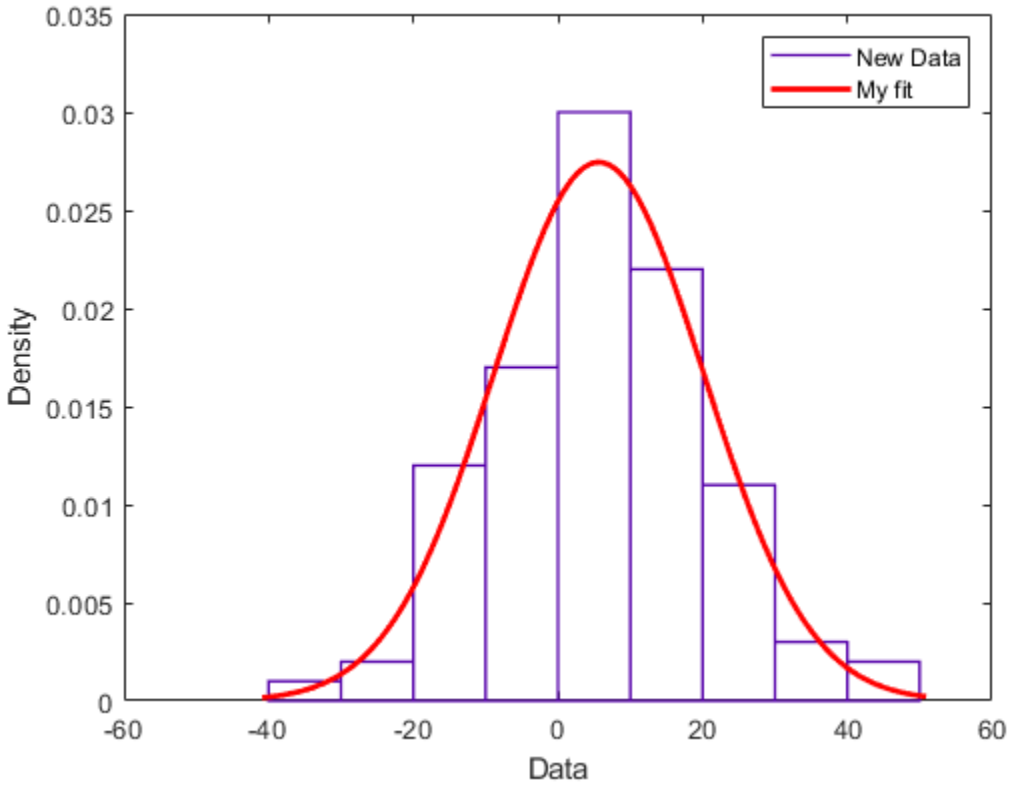
1. From the **File** menu, select `Generate Code`.
2. In the MATLAB Editor window, choose **File > Save as**. Save the file as `normal_fit.m` in a folder on the MATLAB path.

You can then apply the function `normal_fit` to any vector of data in the MATLAB workspace. For example, the following commands:

```
new_data = normrnd(4.1, 12.5, 100, 1);
newfit = normal_fit(new_data)
legend('New Data', 'My fit')
```

generate `newfit`, a fitted normal distribution of the data. The commands also generate a plot of the data and the fit.

```
newfit =

  NormalDistribution

  Normal distribution
       mu = 5.63857   [2.7555, 8.52163]
    sigma =   14.53   [12.7574, 16.8791]
```



Note By default, the file labels the data in the legend using the same name as the data set in the Distribution Fitter app. You can change the label using the `legend` command, as illustrated by the preceding example.

See Also

Distribution Fitter

More About

- “Fit a Distribution Using the Distribution Fitter App” on page 5-71
- “Define Custom Distributions Using the Distribution Fitter App” on page 5-81

Fit a Distribution Using the Distribution Fitter App

In this section...

This example shows how you can use the Distribution Fitter app to interactively fit a probability distribution to data.

Step 1: Load Sample Data

Load the sample data.

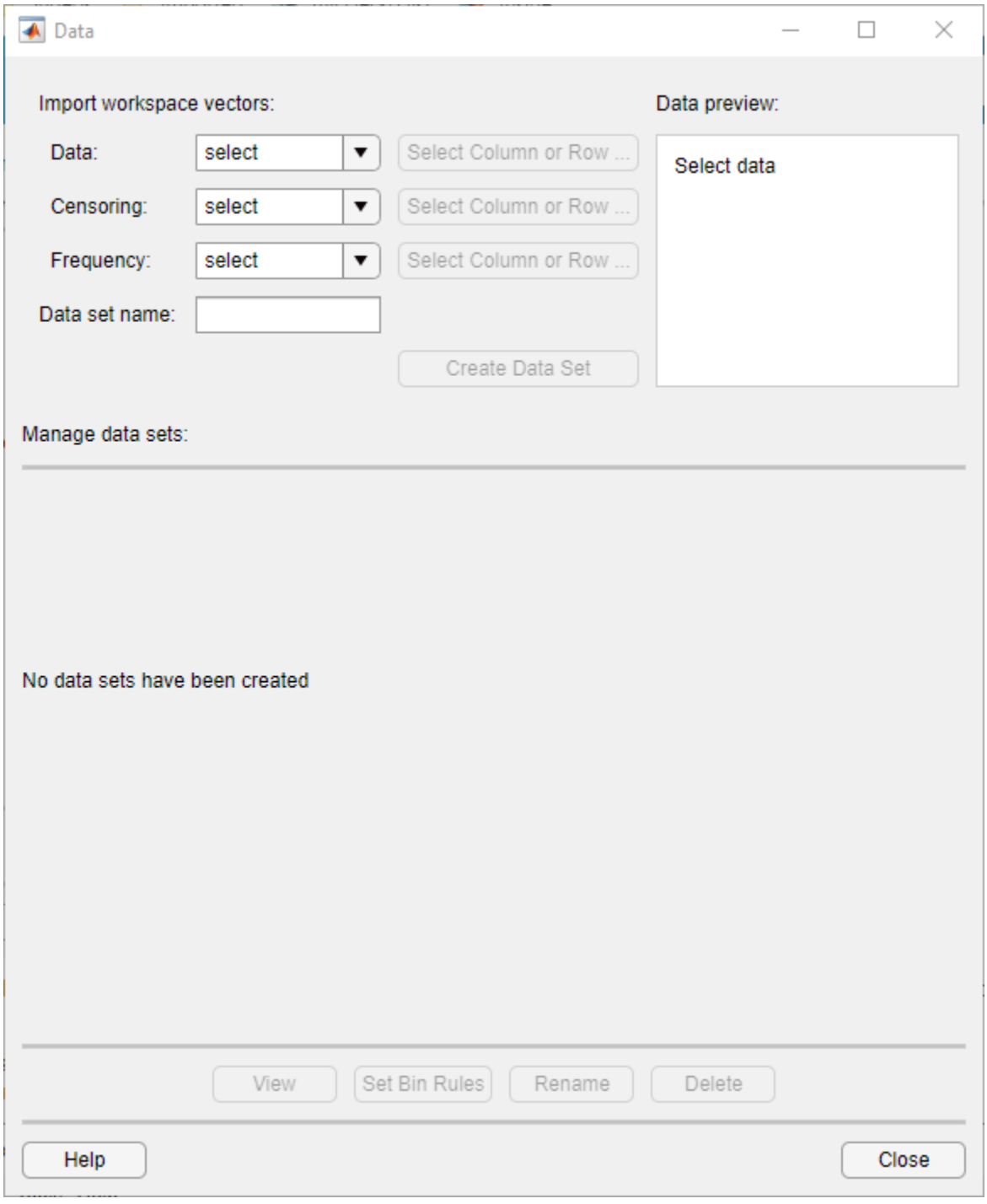
```
load carsmall
```

Step 2: Import Data

Open the Distribution Fitter tool.

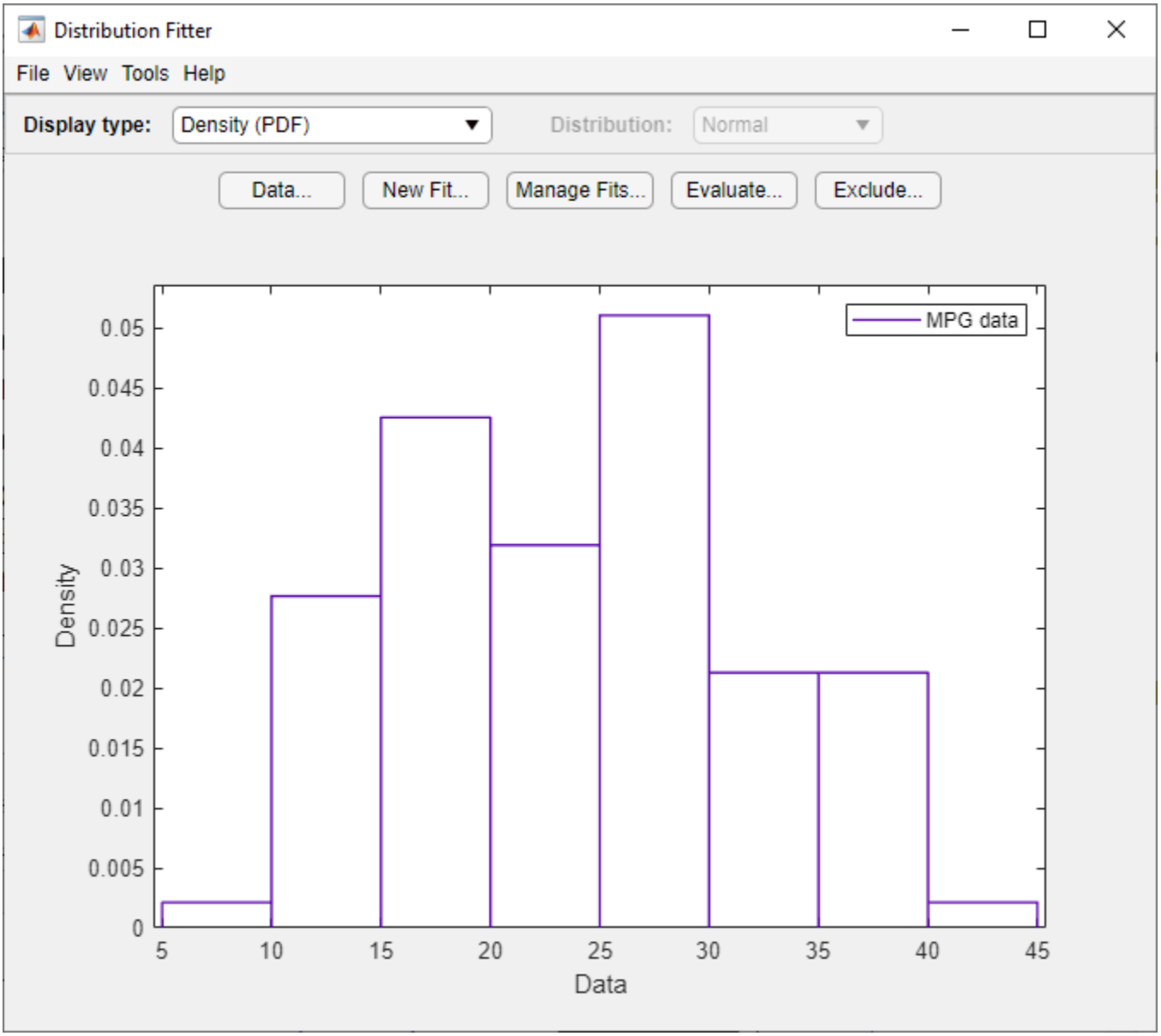
```
distributionFitter
```

To import the vector `MPG` into the Distribution Fitter app, click the **Data** button. The **Data** dialog box opens.



The **Data** field displays all numeric arrays in the MATLAB workspace. From the drop-down list, select `MPG`. A histogram of the selected data appears in the **Data preview** pane.

In the **Data set name** field, type a name for the data set, such as `MPG data`, and click **Create Data Set**. The main window of the Distribution Fitter app now displays a larger version of the histogram in the **Data preview** pane.

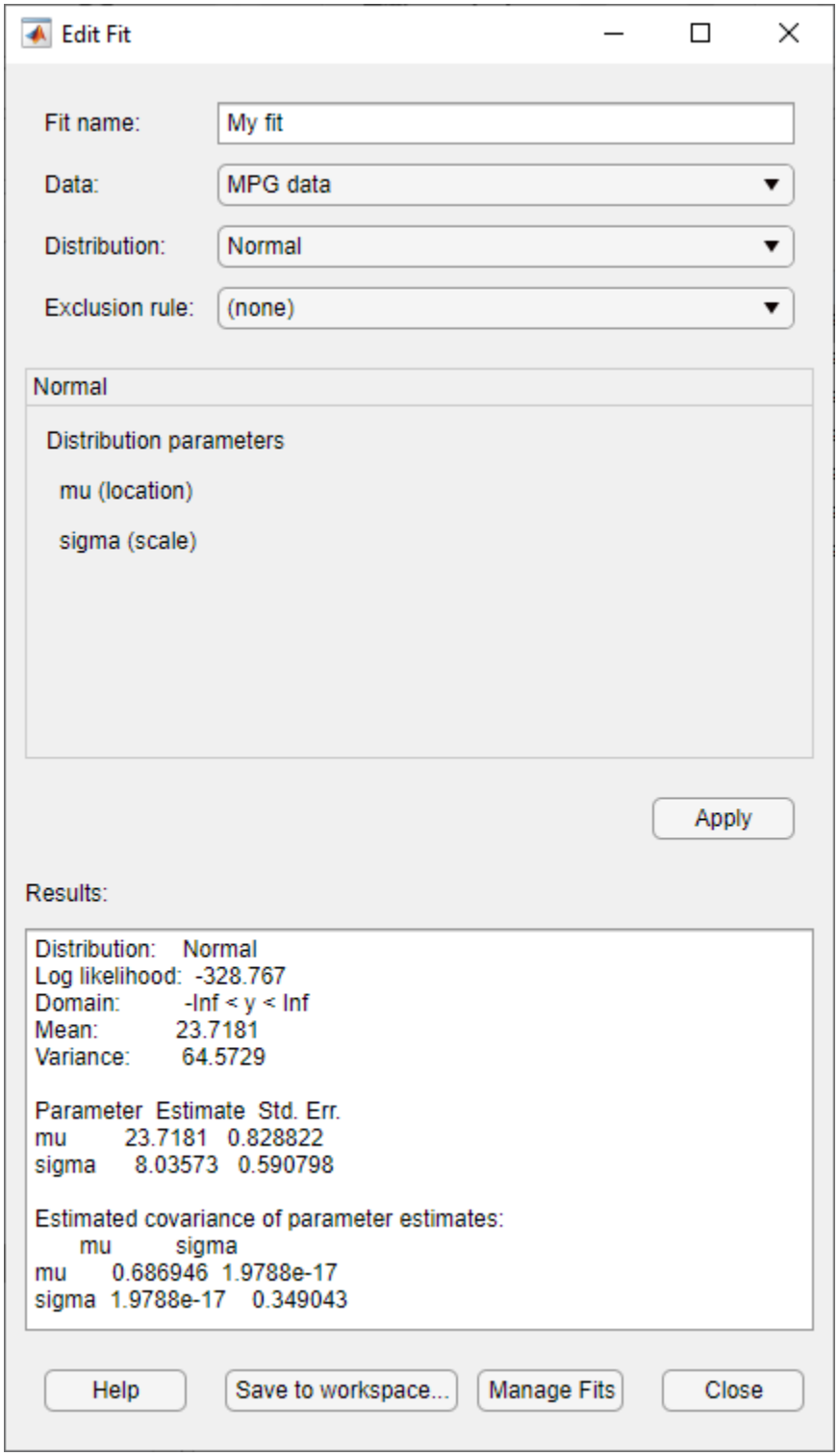


Step 3: Create a New Fit

To fit a distribution to the data, in the main window of the Distribution Fitter app, click **New Fit**.

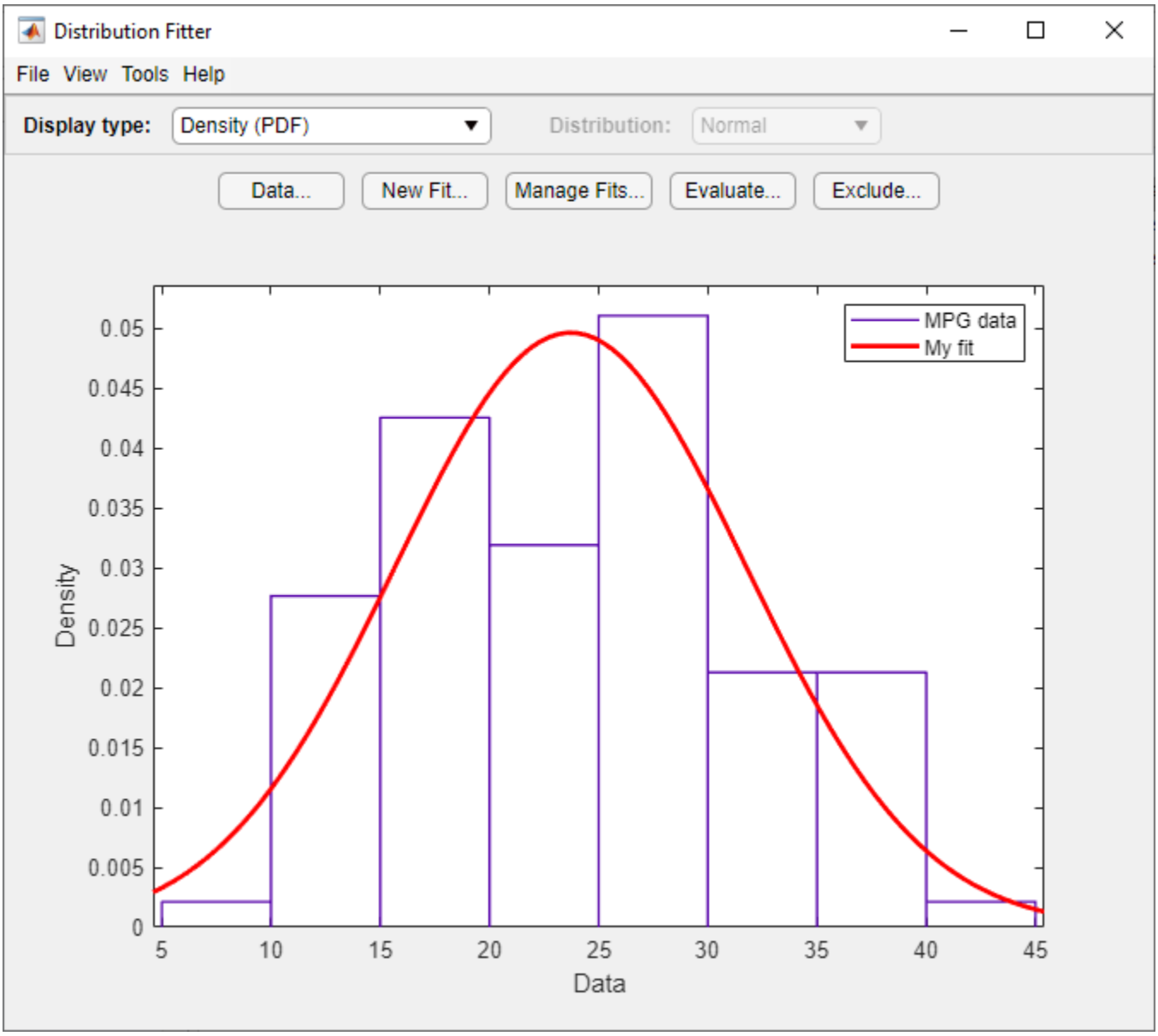
To fit a normal distribution to `MPG data`:

1. In the **Fit name** field, enter a name for the fit, such as `My fit`.
2. From the drop-down list in the **Data** field, select `MPG data`.
3. Confirm that `Normal` is selected from the drop-down menu in the **Distribution** field.
4. Click **Apply**.

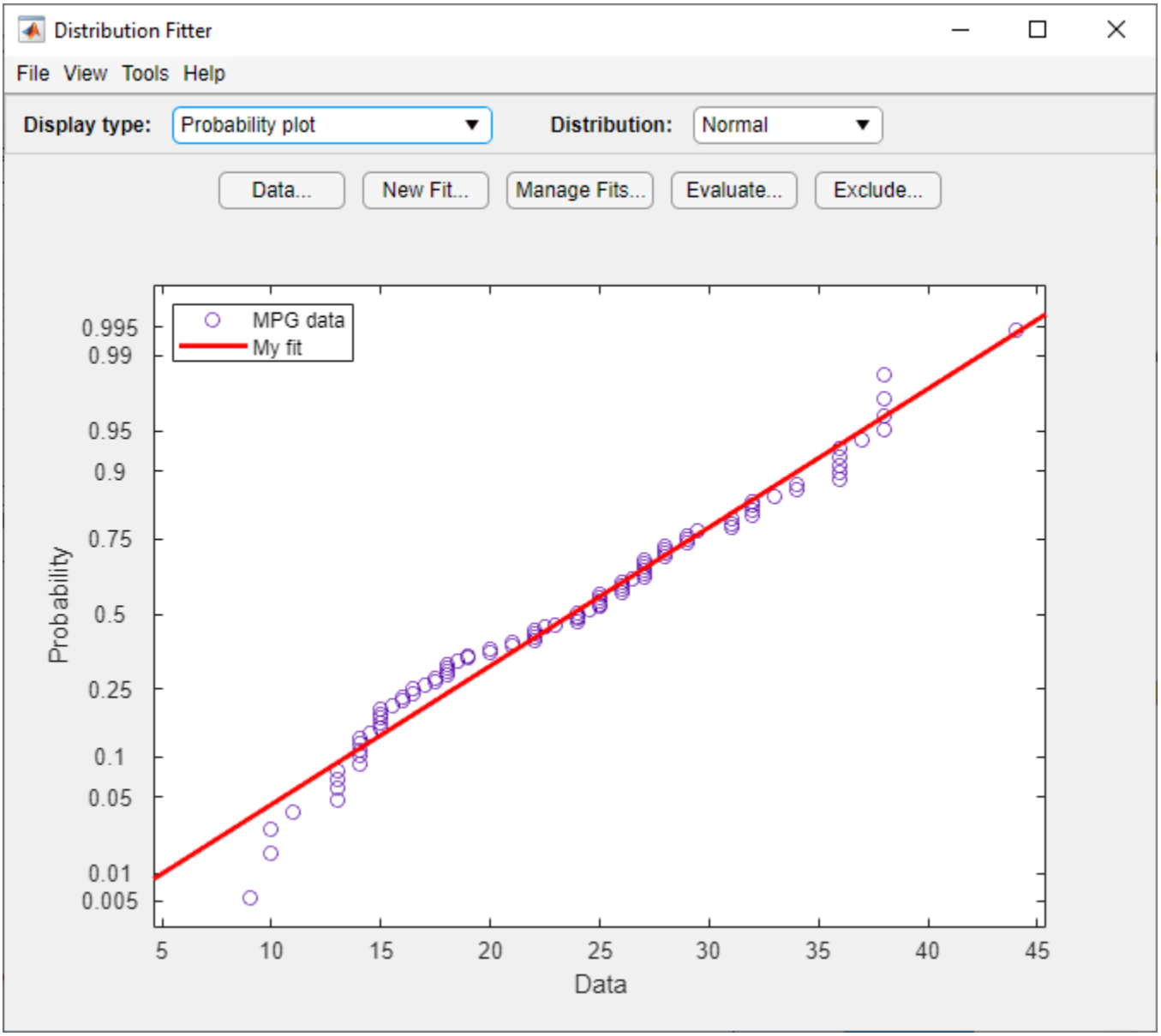


The **Results** pane displays the mean and standard deviation of the normal distribution that best fits `MPG data`.

The Distribution Fitter app main window displays a plot of the normal distribution with this mean and standard deviation.



Based on the plot, a normal distribution does not appear to provide a good fit for the `MPG` data. To obtain a better evaluation, select `Probability plot` from the **Display type** drop-down list. Confirm that the **Distribution** drop-down list is set to `Normal`. The main window displays the following figure.



The normal probability plot shows that the data deviates from normal, especially in the tails.

Step 4: Create and Manage Additional Fits

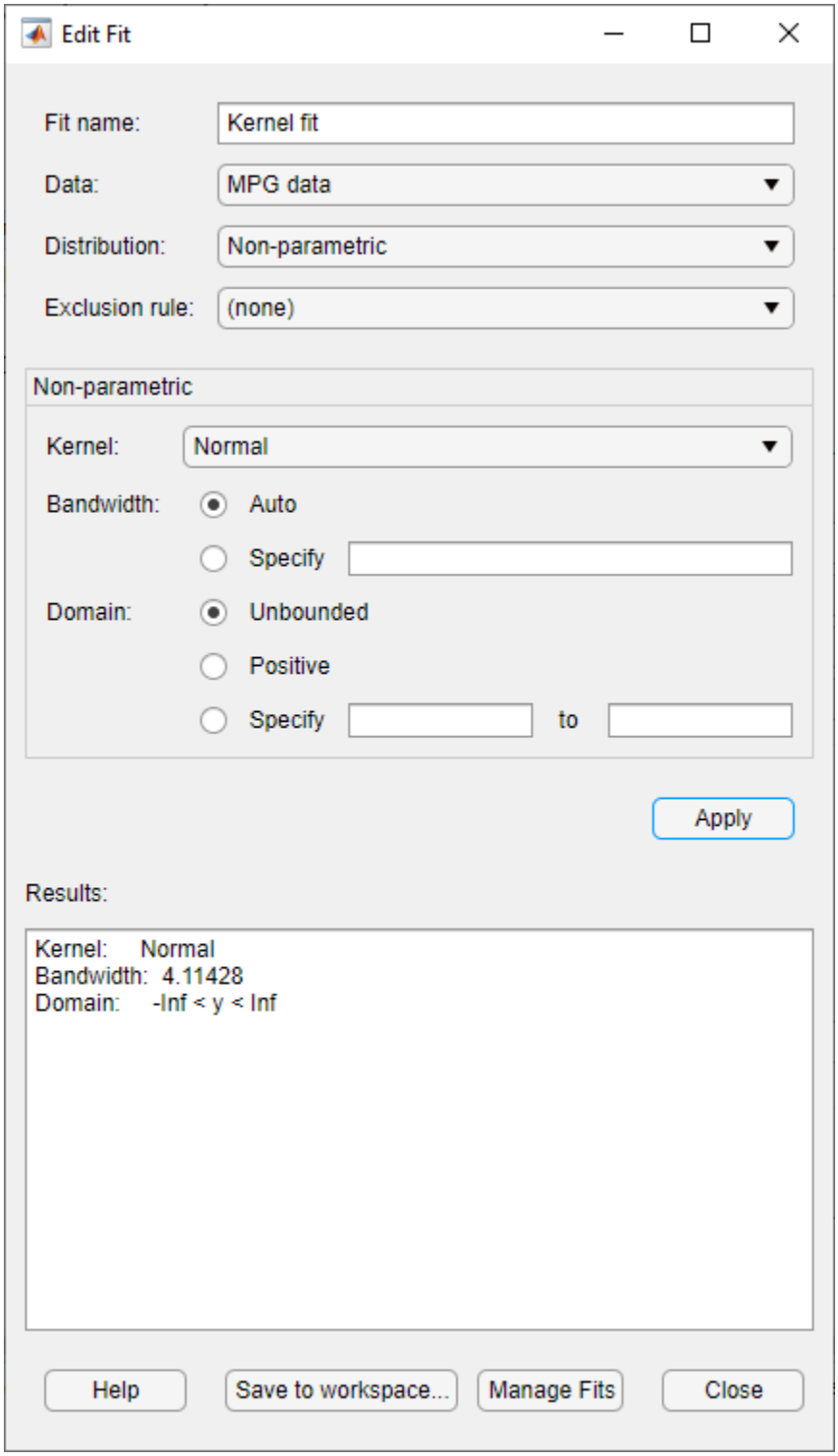
The `MPG` data pdf indicates that the data has two peaks. Try fitting a nonparametric kernel distribution to obtain a better fit for this data.

1. Click **Manage Fits**. In the dialog box, click **New Fit**.
2. In the **Fit name** field, enter a name for the fit, such as `Kernel fit`.
3. From the drop-down list in the **Data** field, select `MPG data`.
4. From the drop-down list in the **Distribution** field, select **Non-parametric**. This enables several options in the **Non-parametric** pane, including **Kernel**, **Bandwidth**, and **Domain**. For now,

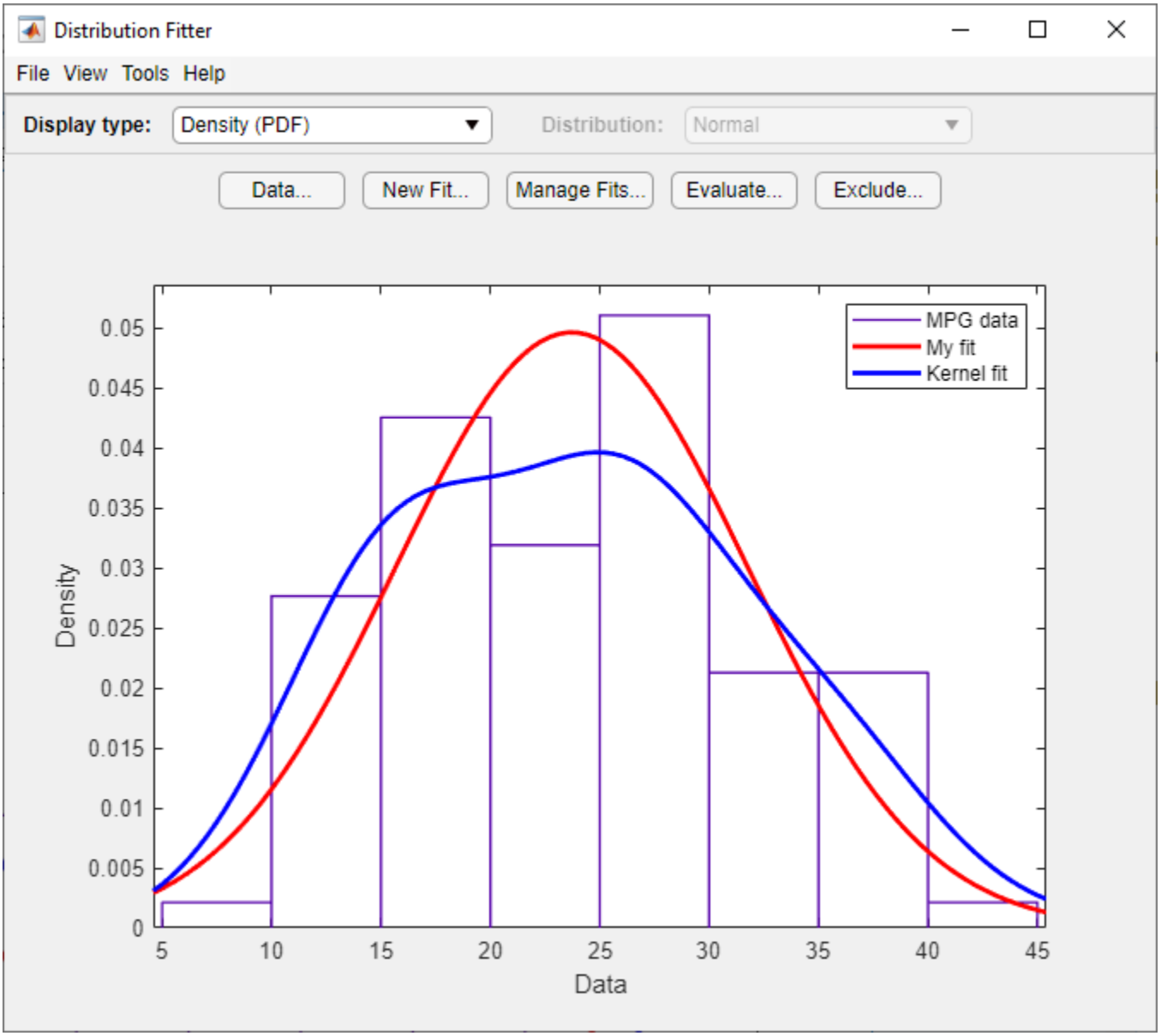
accept the default value to apply a normal kernel shape and automatically determine the kernel bandwidth (using **Auto**). For more information about nonparametric kernel distributions, see “Kernel Distribution” on page B-78.

5 Click **Apply**.

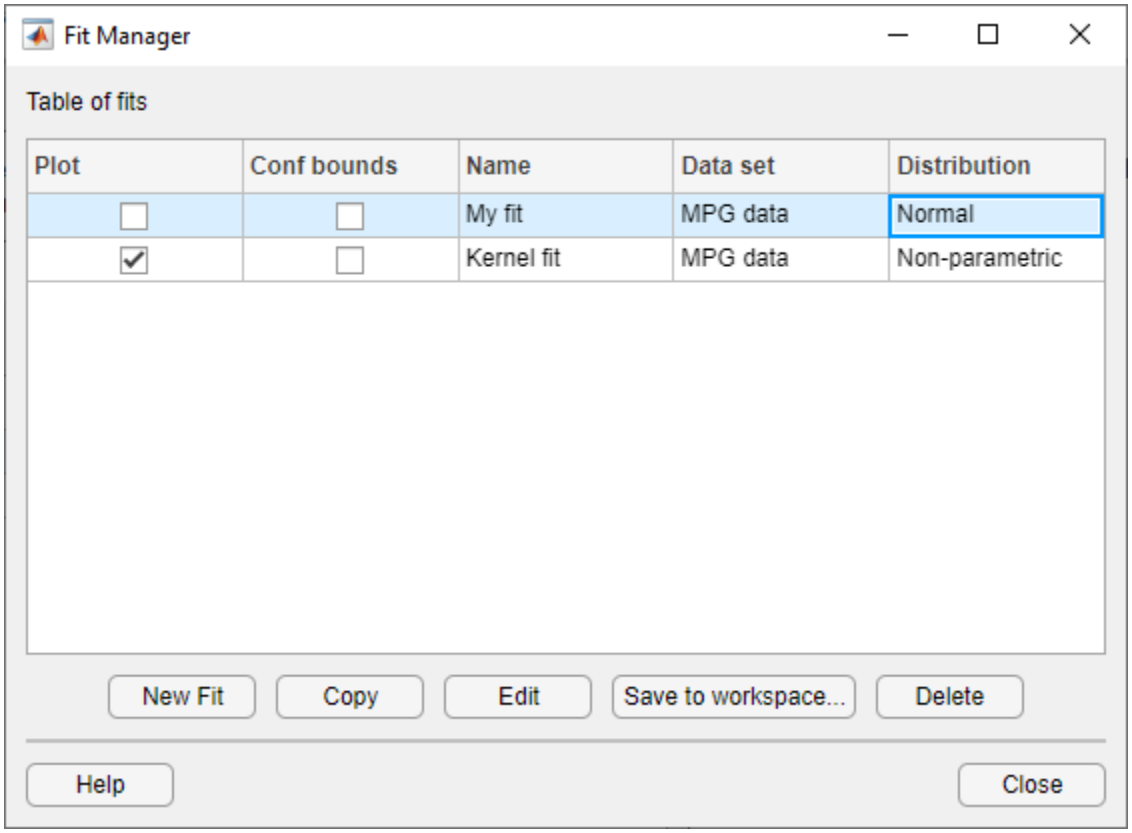
The **Results** pane displays the kernel type, bandwidth, and domain of the nonparametric distribution fit to `MPG data`.



The main window displays plots of the original `MPG data` with the normal distribution and nonparametric kernel distribution overlaid. To visually compare these two fits, select `Density (PDF)` from the **Display type** drop-down list.



To include only the nonparametric kernel fit line (`Kernel fit`) on the plot, click Manage Fits. In the **Table of fits** pane, locate the row for the normal distribution fit (`My fit`) and clear the box in the **Plot** column.



See Also

Distribution Fitter

More About

- “Model Data Using the Distribution Fitter App” on page 5-51
- “Define Custom Distributions Using the Distribution Fitter App” on page 5-81

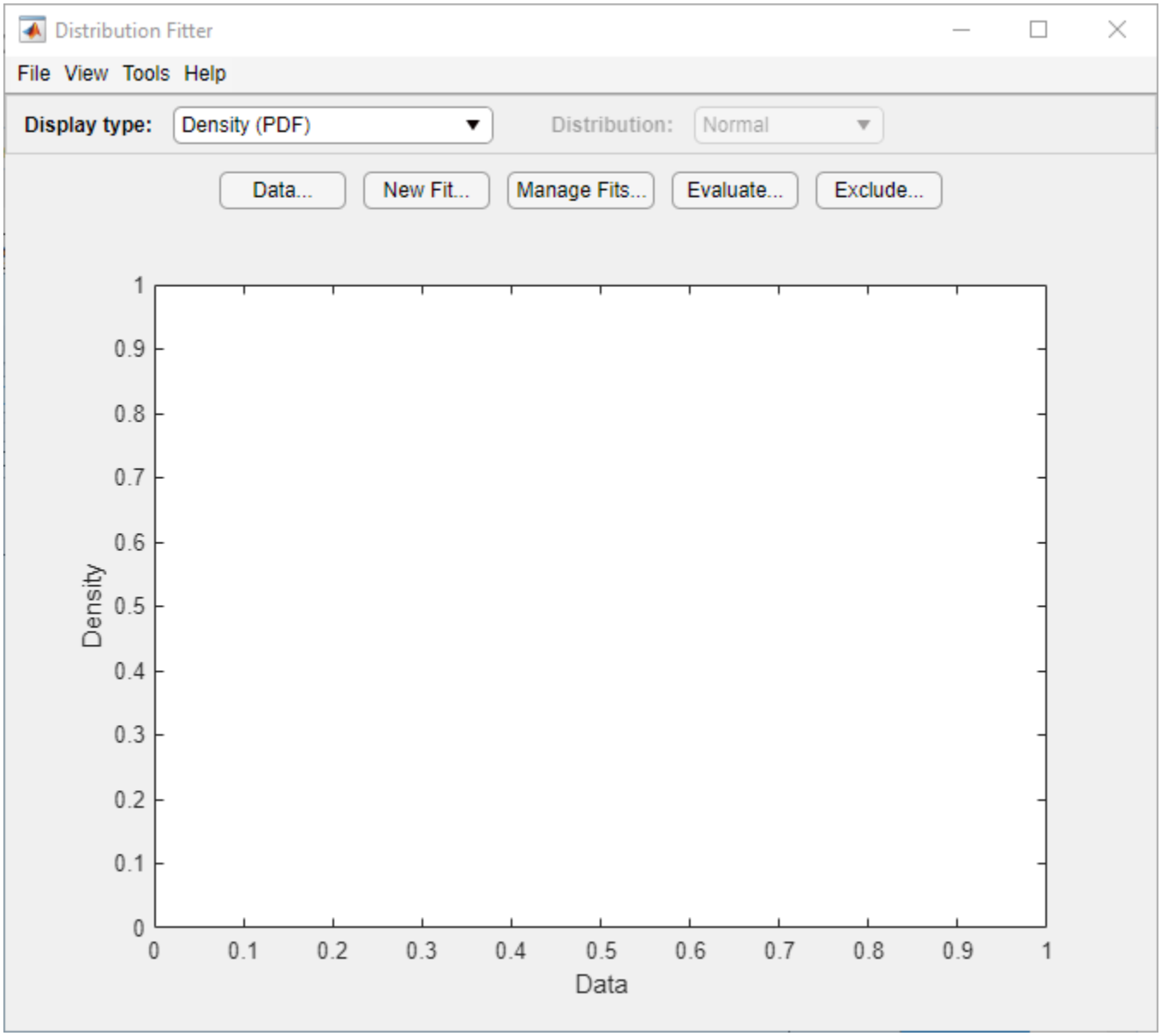
Define Custom Distributions Using the Distribution Fitter App

You can define a probability object for a custom distribution and use the Distribution Fitter app or `fitdist` to fit distributions not supported by Statistics and Machine Learning Toolbox. You can also use a custom probability object as an input argument of probability object functions, such as `pdf`, `cdf`, `icdf`, and `random`, to evaluate the distribution, generate random numbers, and so on.

Open the Distribution Fitter App

- MATLAB Toolstrip: On the **Apps** tab, under **Math, Statistics and Optimization**, click the app icon.
- MATLAB command prompt: Enter `distributionFitter`.

  ```
  distributionFitter
  ```



Define Custom Distribution

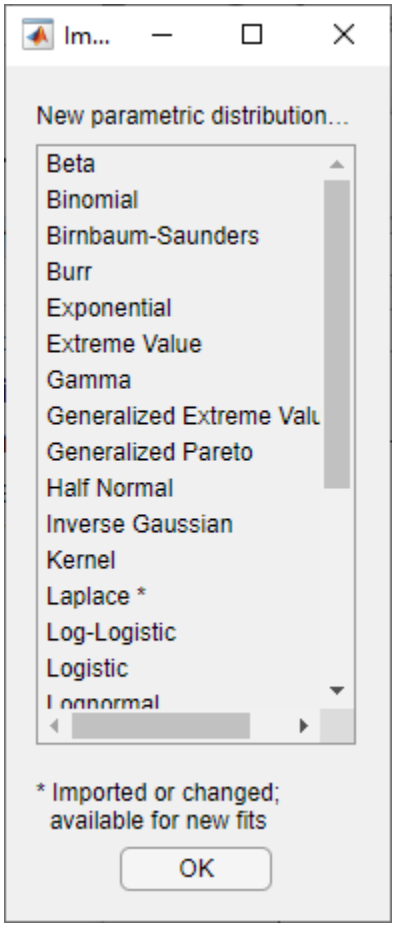
To define a custom distribution using the app, select **File > Define Custom Distributions**. A file template opens in the MATLAB Editor. You then edit this file so that it creates a probability object for the distribution you want.

The template includes sample code that defines a probability object for the Laplace distribution. Follow the instructions in the template to define your own custom distribution.

To save your custom probability object, create a directory named `+prob` on your path. Save the file in this directory using a name that matches your distribution name. For example, save the template as `LaplaceDistribution.m`, and then import the custom distribution as described in the next section.

Import Custom Distribution

To import a custom distribution using the app, select **File > Import Custom Distributions**. The Imported Distributions dialog box opens, in which you can select the file that defines the distribution. For example, if you create the file `LaplaceDistribution.m`, as described in the preceding section, the list in the dialog box includes `Laplace` followed by an asterisk, indicating the file is new or modified and available for fitting.



Alternatively, you can use the `makedist` function to reset the list of distributions so that you do not need to select **File > Import Custom Distributions** in the app.

```
makedist -reset
```

This command resets the list of distributions by searching the path for files contained in a package named `prob` and implementing classes derived from `ProbabilityDistribution`. If you open the app after resetting the list, the distribution list in the app includes the custom distribution that you defined.

Once you import a custom distribution using the Distribution Fitter app or reset the list by using `makedist`, you can use the custom distribution in the app and in the Command Window. The **Distribution** field of the New Fit dialog box, available from the Distribution Fitter app, contains the new custom distribution. In the Command Window, you can create the custom probability distribution object by using `makedist` and fit a data set to the custom distribution by using `fitdist`. Then, you

can use probability object functions, such as `pdf`, `cdf`, `icdf`, and `random`, to evaluate the distribution, generate random numbers, and so on.

See Also

Distribution Fitter | `makedist`

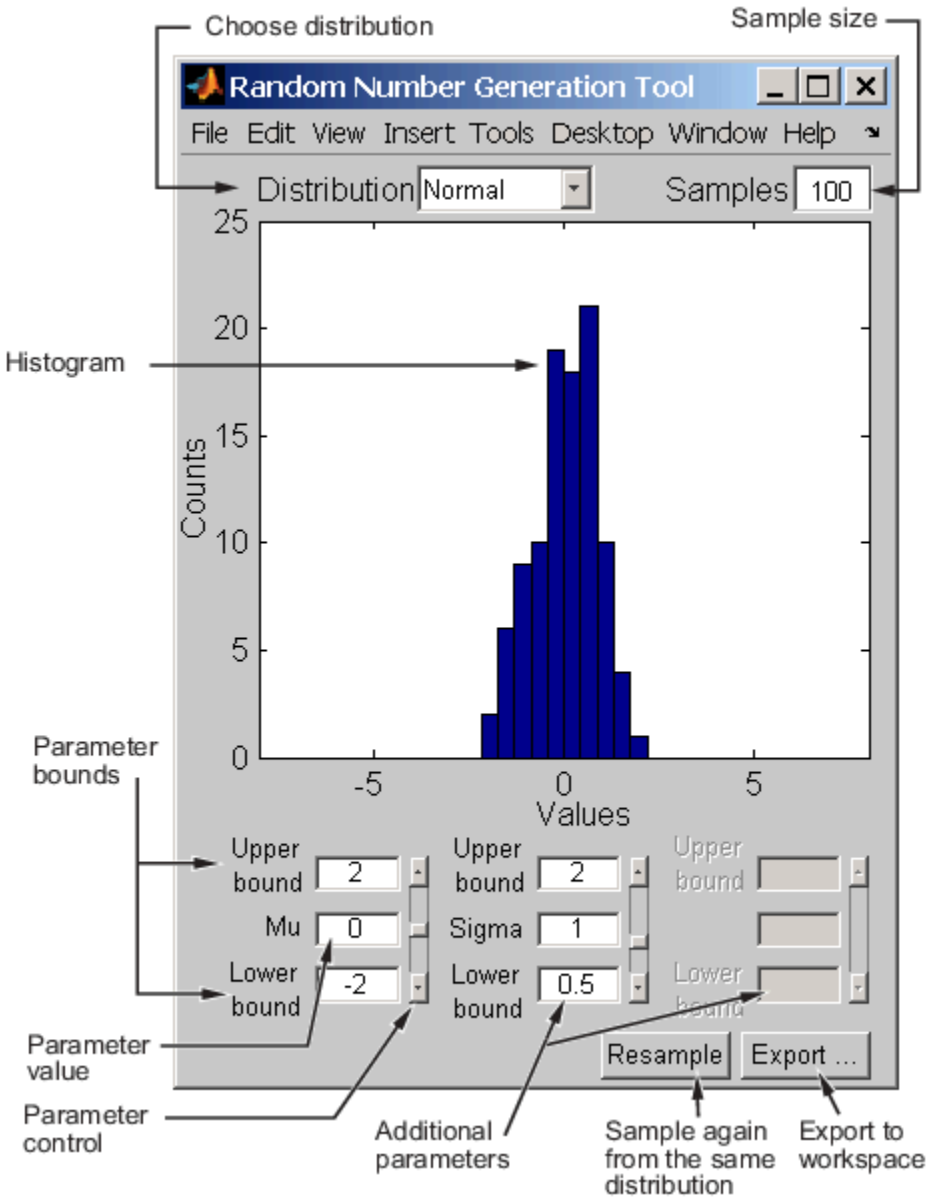
More About

- “Model Data Using the Distribution Fitter App” on page 5-51
- “Fit a Distribution Using the Distribution Fitter App” on page 5-71

Explore the Random Number Generation UI

The Random Number Generation user interface (UI) generates random samples from specified probability distributions, and displays the samples as histograms. Use the interface to explore the effects of changing parameters and sample size on the distributions.

Run the user interface by typing `randtool` at the command line.



Start by selecting a distribution, then enter the desired sample size.

You can also

- Use the controls at the bottom of the window to set parameter values for the distribution and to change their upper and lower bounds.

- Draw another sample from the same distribution, with the same size and parameters.
- Export the current sample to your workspace. A dialog box enables you to provide a name for the sample.

Compare Multiple Distribution Fits

This example shows how to fit multiple probability distribution objects to the same set of sample data, and obtain a visual comparison of how well each distribution fits the data.

Step 1. Load sample data.

Load the sample data.

```
load carsmall
```

This data contains miles per gallon (`MPG`) measurements for different makes and models of cars, grouped by country of origin (`Origin`), model year (`Model_Year`), and other vehicle characteristics.

Step 2. Create a categorical array.

Transform `Origin` into a categorical array and remove the Italian car from the sample data. Since there is only one Italian car, `fitdist` cannot fit a distribution to that group using other than a kernel distribution.

```
Origin = categorical(cellstr(Origin));
MPG2 = MPG(Origin~='Italy');
Origin2 = Origin(Origin~='Italy');
Origin2 = removecats(Origin2,'Italy');
```

Step 3. Fit multiple distributions by group.

Use `fitdist` to fit Weibull, normal, logistic, and kernel distributions to each country of origin group in the `MPG` data.

```
[WeiByOrig,Country] = fitdist(MPG2,'weibull','by',Origin2);
[NormByOrig,Country] = fitdist(MPG2,'normal','by',Origin2);
[LogByOrig,Country] = fitdist(MPG2,'logistic','by',Origin2);
[KerByOrig,Country] = fitdist(MPG2,'kernel','by',Origin2);
```

```
WeiByOrig
```

```
WeiByOrig=1×5 cell array
  Columns 1 through 2

    {1x1 prob.WeibullDistribution}    {1x1 prob.WeibullDistribution}

  Columns 3 through 4

    {1x1 prob.WeibullDistribution}    {1x1 prob.WeibullDistribution}

  Column 5

    {1x1 prob.WeibullDistribution}
```

```
Country
```

```
Country = 5x1 cell
    {'France' }
    {'Germany'}
    {'Japan'  }
    {'Sweden' }
```

```
    {'USA'    }
```

Each country group now has four distribution objects associated with it. For example, the cell array `WeiByOrig` contains five Weibull distribution objects, one for each country represented in the sample data. Likewise, the cell array Norm`ByOrig` contains five normal distribution objects, and so on. Each object contains properties that hold information about the data, distribution, and parameters. The array `Country` lists the country of origin for each group in the same order as the distribution objects are stored in the cell arrays.

Step 4. Compute the pdf for each distribution.

Extract the four probability distribution objects for USA and compute the pdf for each distribution. As shown in Step 3, USA is in position 5 in each cell array.

```
WeiUSA = WeiByOrig{5};
NormUSA = NormByOrig{5};
LogUSA = LogByOrig{5};
KerUSA = KerByOrig{5};

x = 0:1:50;
pdf_Wei = pdf(WeiUSA,x);
pdf_Norm = pdf(NormUSA,x);
pdf_Log = pdf(LogUSA,x);
pdf_Ker = pdf(KerUSA,x);
```

Step 5. Plot pdf the for each distribution.

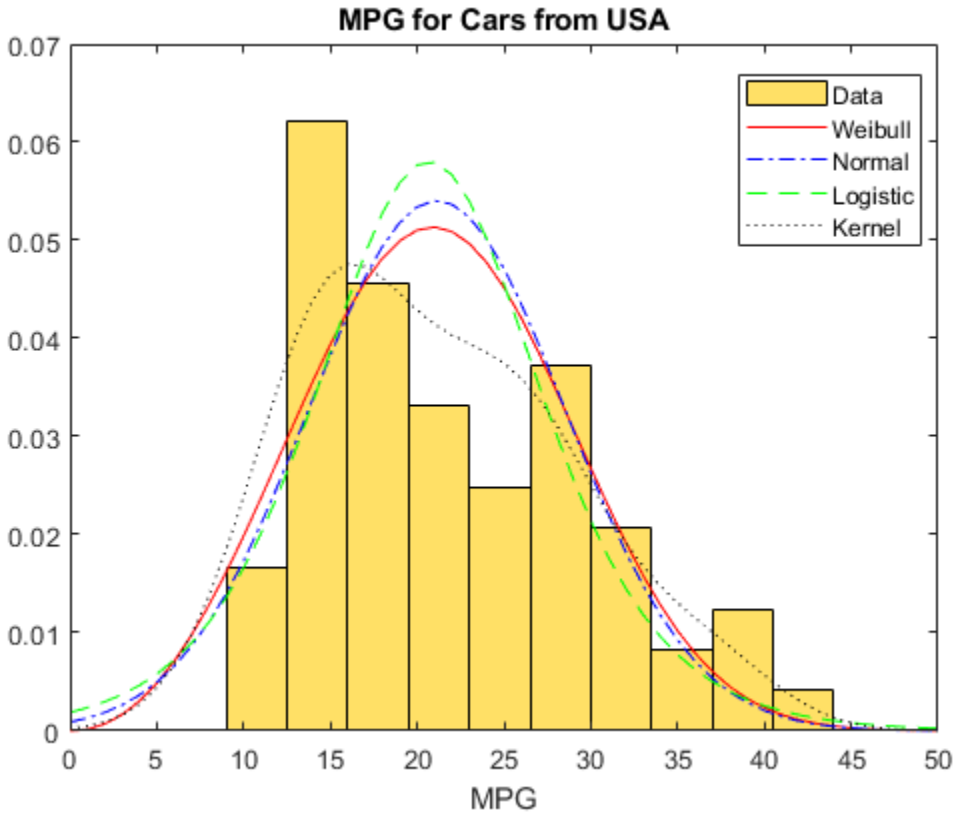
Plot the pdf for each distribution fit to the USA data, superimposed on a histogram of the sample data. Normalize the histogram for easier display.

Create a histogram of the USA sample data.

```
data = MPG(Origin2=='USA');
figure
histogram(data,10,'Normalization','pdf','FaceColor',[1,0.8,0]);
```

Plot the pdf of each fitted distribution.

```
line(x,pdf_Wei,'LineStyle','-','Color','r')
line(x,pdf_Norm,'LineStyle','-.','Color','b')
line(x,pdf_Log,'LineStyle','--','Color','g')
line(x,pdf_Ker,'LineStyle',':','Color','k')
legend('Data','Weibull','Normal','Logistic','Kernel','Location','Best')
title('MPG for Cars from USA')
xlabel('MPG')
```



Superimposing the pdf plots over a histogram of the sample data provides a visual comparison of how well each type of distribution fits the data. Only the nonparametric kernel distribution `KerUSA` comes close to revealing the two modes in the original data.

Step 6. Further group USA data by year.

To investigate the two modes revealed in Step 5, group the `MPG` data by both country of origin (`Origin`) and model year (`Model_Year`), and use `fitdist` to fit kernel distributions to each group.

```
[KerByYearOrig,Names] = fitdist(MPG,'Kernel','By',{Origin Model_Year});
```

Each unique combination of origin and model year now has a kernel distribution object associated with it.

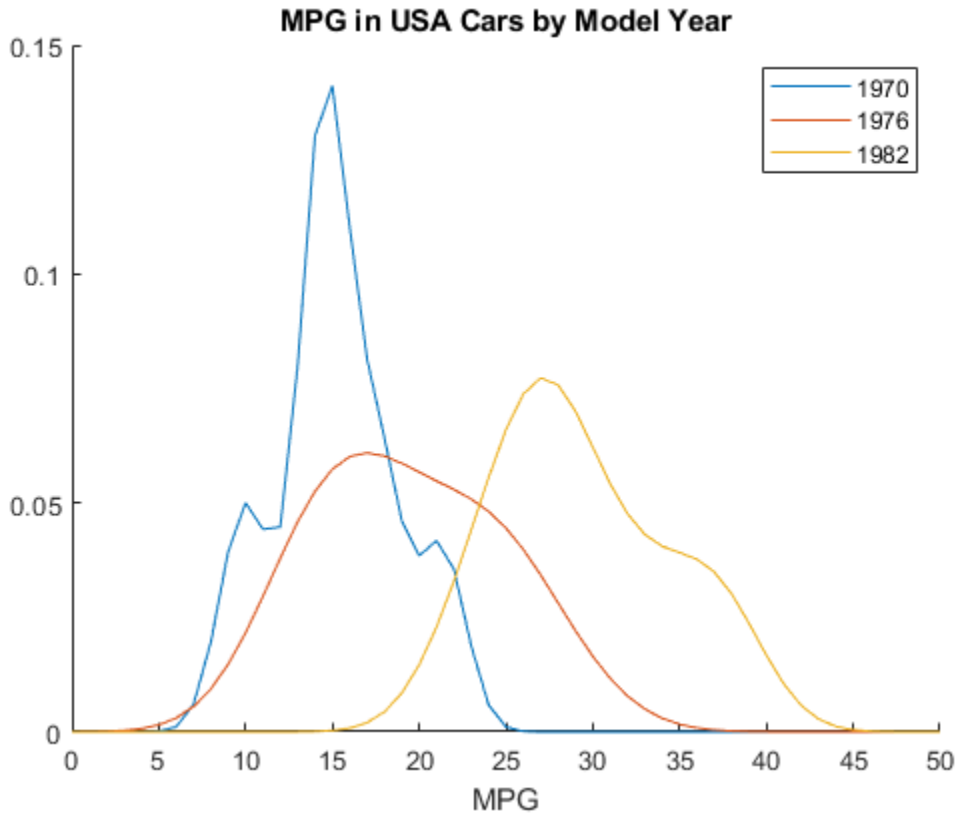
```
Names
```

```
Names = 14x1 cell
    {'France...' }
    {'France...' }
    {'Germany...'}
    {'Germany...'}
    {'Germany...'}
    {'Italy...'  }
    {'Japan...'  }
    {'Japan...'  }
    {'Japan...'  }
    {'Sweden...' }
```

```
    {'Sweden...' }
    {'USA...'    }
    {'USA...'    }
    {'USA...'    }
```

Plot the three probability distributions for each USA model year, which are in positions 12, 13, and 14 in the cell array `KerByYearOrig`.

```
figure
hold on
for i = 12 : 14
    plot(x,pdf(KerByYearOrig{i},x))
end
legend('1970','1976','1982')
title('MPG in USA Cars by Model Year')
xlabel('MPG')
hold off
```



When further grouped by model year, the pdf plots reveal two distinct peaks in the `MPG` data for cars made in the USA — one for the model year 1970 and one for the model year 1982. This explains why the histogram for the combined USA miles per gallon data shows two peaks instead of one.

See Also

`fitdist` | `histogram` | `pdf`

More About

- “Grouping Variables” on page 2-45
- “Fit Probability Distribution Objects to Grouped Data” on page 5-92
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Fit Probability Distribution Objects to Grouped Data

This example shows how to fit probability distribution objects to grouped sample data, and create a plot to visually compare the pdf of each group.

Step 1. Load sample data.

Load the sample data.

```
load carsmall;
```

The data contains miles per gallon (`MPG`) measurements for different makes and models of cars, grouped by country of origin (`Origin`), model year (`Model_Year`), and other vehicle characteristics.

Step 2. Create a categorical array.

Transform `Origin` into a categorical array.

```
Origin = categorical(cellstr(Origin));
```

Step 3. Fit kernel distributions to each group.

Use `fitdist` to fit kernel distributions to each country of origin group in the `MPG` data.

```
[KerByOrig,Country] = fitdist(MPG,'Kernel','by',Origin)
```

```
KerByOrig=1×6 cell array
  Columns 1 through 2

    {1x1 prob.KernelDistribution}    {1x1 prob.KernelDistribution}

  Columns 3 through 4

    {1x1 prob.KernelDistribution}    {1x1 prob.KernelDistribution}

  Columns 5 through 6

    {1x1 prob.KernelDistribution}    {1x1 prob.KernelDistribution}


Country = 6x1 cell
    {'France' }
    {'Germany'}
    {'Italy'  }
    {'Japan'  }
    {'Sweden' }
    {'USA'    }
```

The cell array `KerByOrig` contains six kernel distribution objects, one for each country represented in the sample data. Each object contains properties that hold information about the data, the distribution, and the parameters. The array `Country` lists the country of origin for each group in the same order as the distribution objects are stored in `KerByOrig`.

Step 4. Compute the pdf for each group.

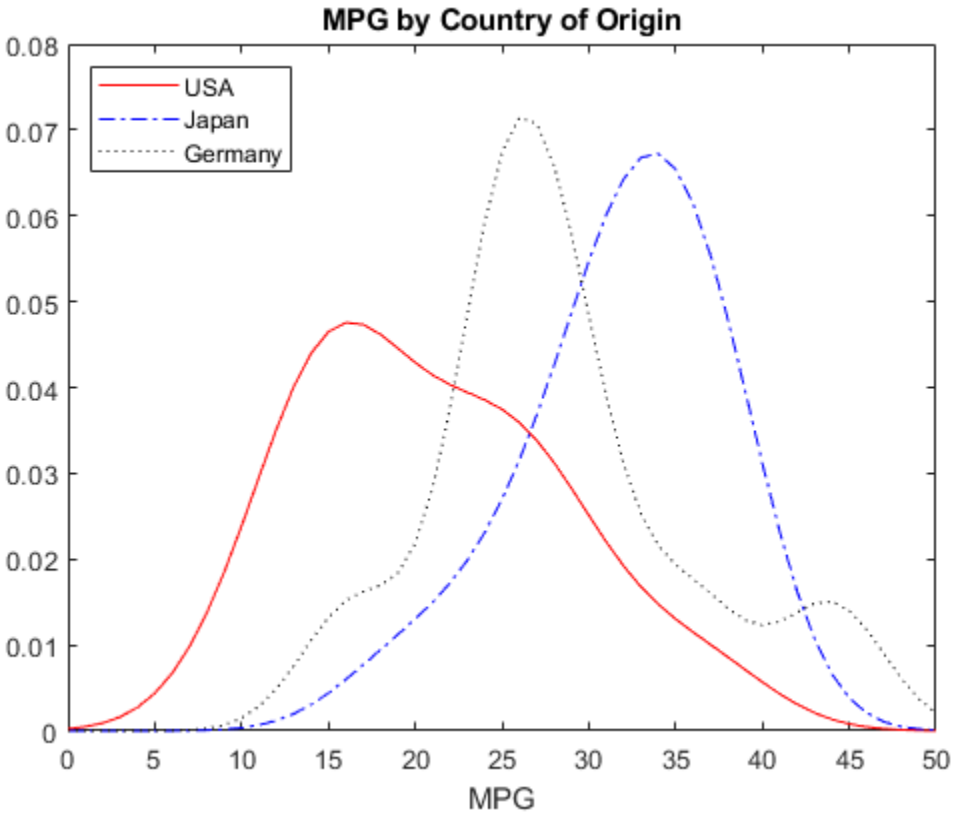
Extract the probability distribution objects for Germany, Japan, and USA. Use the positions of each country in `KerByOrig` shown in Step 3, which indicates that Germany is the second country, Japan is the fourth country, and USA is the sixth country. Compute the pdf for each group.

```
Germany = KerByOrig{2};
Japan = KerByOrig{4};
USA = KerByOrig{6};

x = 0:1:50;

USA_pdf = pdf(USA,x);
Japan_pdf = pdf(Japan,x);
Germany_pdf = pdf(Germany,x);
```

Step 5. Plot the pdf for each group.

Plot the pdf for each group on the same figure.

```
plot(x,USA_pdf,'r-')
hold on
plot(x,Japan_pdf,'b-.')
plot(x,Germany_pdf,'k:')
legend({'USA','Japan','Germany'},'Location','NW')
title('MPG by Country of Origin')
xlabel('MPG')
```



The resulting plot shows how miles per gallon (`MPG`) performance differs by country of origin (`Origin`). Using this data, the USA has the widest distribution, and its peak is at the lowest `MPG` value of the three origins. Japan has the most regular distribution with a slightly heavier left tail, and its peak is at the highest `MPG` value of the three origins. The peak for Germany is between the USA and Japan, and the second bump near 44 miles per gallon suggests that there might be multiple modes in the data.

See Also

`fitdist` | `pdf`

More About

- “Kernel Distribution” on page B-78
- “Grouping Variables” on page 2-45
- “Fit Distributions to Grouped Data Using ksdensity” on page 5-41
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Multinomial Probability Distribution Objects

This example shows how to generate random numbers, compute and plot the pdf, and compute descriptive statistics of a multinomial distribution using probability distribution objects.

Step 1. Define the distribution parameters.

Create a vector `p` containing the probability of each outcome. Outcome 1 has a probability of 1/2, outcome 2 has a probability of 1/3, and outcome 3 has a probability of 1/6. The number of trials `n` in each experiment is 5, and the number of repetitions `reps` of the experiment is 8.

```
p = [1/2 1/3 1/6];
n = 5;
reps = 8;
```

Step 2. Create a multinomial probability distribution object.

Create a multinomial probability distribution object using the specified value `p` for the `Probabilities` parameter.

```
pd = makedist('Multinomial','Probabilities',p)
```

```
pd =
  MultinomialDistribution

  Probabilities:
    0.5000    0.3333    0.1667
```

Step 3. Generate one random number.

Generate one random number from the multinomial distribution, which is the outcome of a single trial.

```
rng('default')  % For reproducibility
r = random(pd)
```

```
r = 2
```

This trial resulted in outcome 2.

Step 4. Generate a matrix of random numbers.

You can also generate a matrix of random numbers from the multinomial distribution, which reports the results of multiple experiments that each contain multiple trials. Generate a matrix that contains the outcomes of an experiment with `n = 5` trials and `reps = 8` repetitions.

```
r = random(pd,reps,n)
```

```
r = 8×5

     3     3     3     2     1
     1     1     2     2     1
     3     3     3     1     2
     2     3     2     2     2
     1     1     1     1     1
     1     2     3     2     3
```

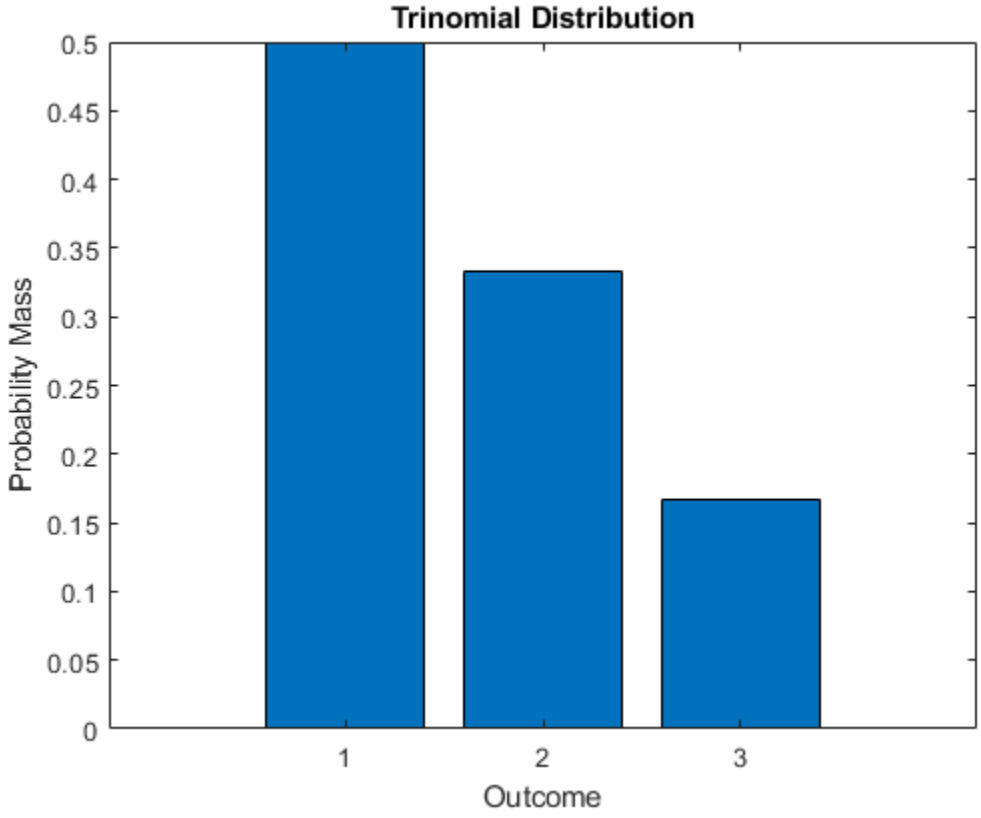
```
     2     1     3     1     1
     3     1     2     1     1
```

Each element in the resulting matrix is the outcome of one trial. The columns correspond to the five trials in each experiment, and the rows correspond to the eight experiments. For example, in the first experiment (corresponding to the first row), one of the five trials resulted in outcome 1, one of the five trials resulted in outcome 2, and three of the five trials resulted in outcome 3.

Step 5. Compute and plot the pdf.

Compute the pdf of the distribution.

```
x = 1:3;
y = pdf(pd,x);
bar(x,y)
xlabel('Outcome')
ylabel('Probability Mass')
title('Trinomial Distribution')
```



The plot shows the probability mass for each k possible outcome. For this distribution, the pdf value for any x other than 1, 2, or 3 is 0.

Step 6. Compute descriptive statistics.

Compute the mean, median, and standard deviation of the distribution.

```
m = mean(pd)
```

```
m = 1.6667
```

```
med = median(pd)
```

```
med = 1
```

```
s = std(pd)
```

```
s = 0.7454
```

See Also

More About

- “Multinomial Probability Distribution Functions” on page 5-98
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Multinomial Probability Distribution Functions

This example shows how to generate random numbers and compute and plot the pdf of a multinomial distribution using probability distribution functions.

Step 1. Define the distribution parameters.

Create a vector `p` containing the probability of each outcome. Outcome 1 has a probability of 1/2, outcome 2 has a probability of 1/3, and outcome 3 has a probability of 1/6. The number of trials in each experiment `n` is 5, and the number of repetitions of the experiment `reps` is 8.

```
p = [1/2 1/3 1/6];
n = 5;
reps = 8;
```

Step 2. Generate one random number.

Generate one random number from the multinomial distribution, which is the outcome of a single trial.

```
rng('default')  % For reproducibility
r = mnrnd(1,p,1)
```

```
r = 1×3

     0     1     0
```

The returned vector `r` contains three elements, which show the counts for each possible outcome. This single trial resulted in outcome 2.

Step 3. Generate a matrix of random numbers.

You can also generate a matrix of random numbers from the multinomial distribution, which reports the results of multiple experiments that each contain multiple trials. Generate a matrix that contains the outcomes of an experiment with `n = 5` trials and `reps = 8` repetitions.

```
r = mnrnd(n,p,reps)
```

```
r = 8×3

     1     1     3
     3     2     0
     1     1     3
     0     4     1
     5     0     0
     1     2     2
     3     1     1
     3     1     1
```

Each row in the resulting matrix contains counts for each of the *k* multinomial bins. For example, in the first experiment (corresponding to the first row), one of the five trials resulted in outcome 1, one of the five trials resulted in outcome 2, and three of the five trials resulted in outcome 3.

Step 4. Compute the pdf.

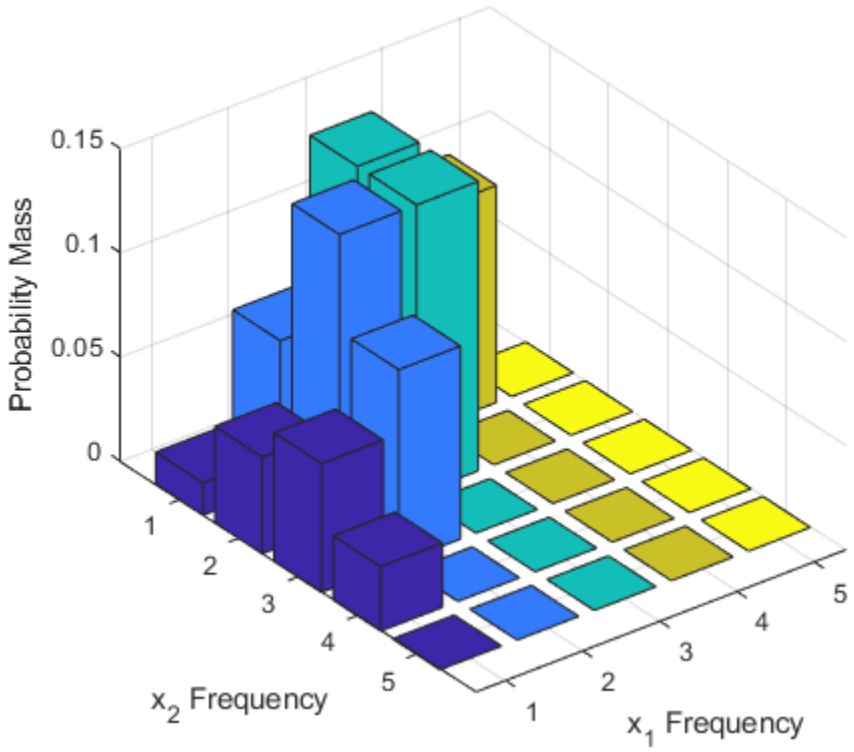
Since multinomial functions work with bin counts, create a multidimensional array of all possible outcome combinations, and compute the pdf using `mnpdf`.

```
count1 = 1:n;
count2 = 1:n;
[x1,x2] = meshgrid(count1,count2);
x3 = n-(x1+x2);
y = mnpdf([x1(:),x2(:),x3(:)],repmat(p,(n)^2,1));
```

Step 5. Plot the pdf.

Create a 3-D bar graph to visualize the pdf for each combination of outcome frequencies.

```
y = reshape(y,n,n);
bar3(y)
set(gca,'XTickLabel',1:n);
set(gca,'YTickLabel',1:n);
xlabel('x_1 Frequency')
ylabel('x_2 Frequency')
zlabel('Probability Mass')
```



The plot shows the probability mass for each possible combination of outcomes. It does not show x_3 , which is determined by the constraint $x_1 + x_2 + x_3 = n$.

See Also

More About

- “Multinomial Probability Distribution Objects” on page 5-95
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Generate Random Numbers Using Uniform Distribution Inversion

This example shows how to generate random numbers using the uniform distribution inversion method. This is useful for distributions when it is possible to compute the inverse cumulative distribution function, but there is no support for sampling from the distribution directly.

Step 1. Generate random numbers from the standard uniform distribution.

Use `rand` to generate 1000 random numbers from the uniform distribution on the interval (0,1).

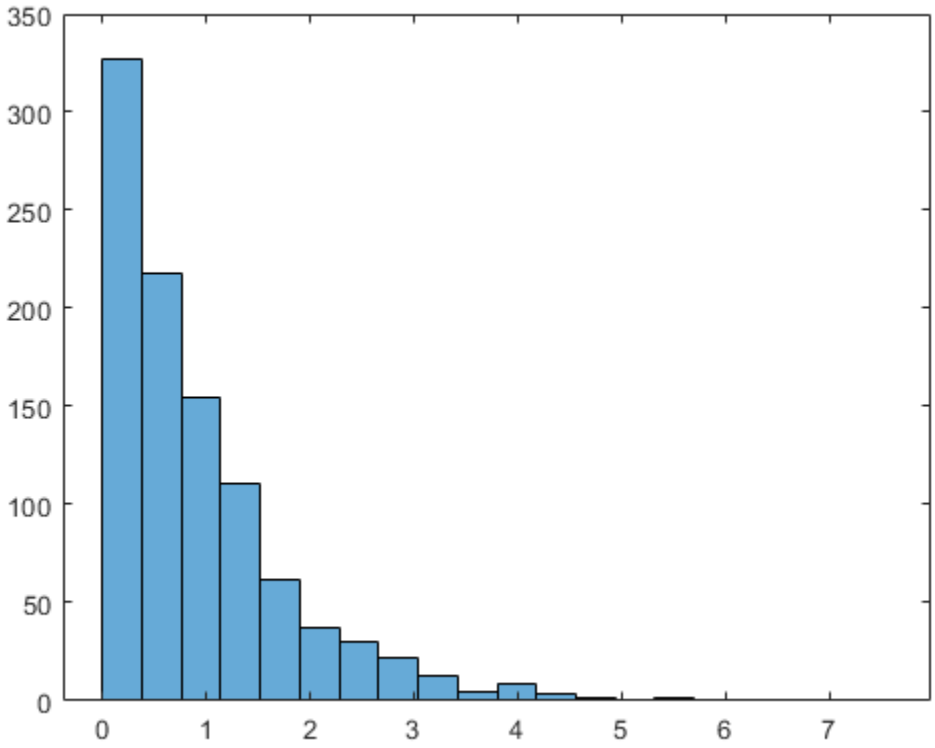
```
rng('default')  % For reproducibility
u = rand(1000,1);
```

The inversion method relies on the principle that continuous cumulative distribution functions (cdfs) range uniformly over the open interval (0,1). If u is a uniform random number on (0,1), then $x = F^{-1}(u)$ generates a random number x from any continuous distribution with the specified cdf `F`.

Step 2. Generate random numbers from the Weibull distribution.

Use the inverse cumulative distribution function to generate the random numbers from a Weibull distribution with parameters `A = 1` and `B = 1` that correspond to the probabilities in `u`. Plot the results.

```
x = wblinv(u,1,1);
histogram(x,20);
```

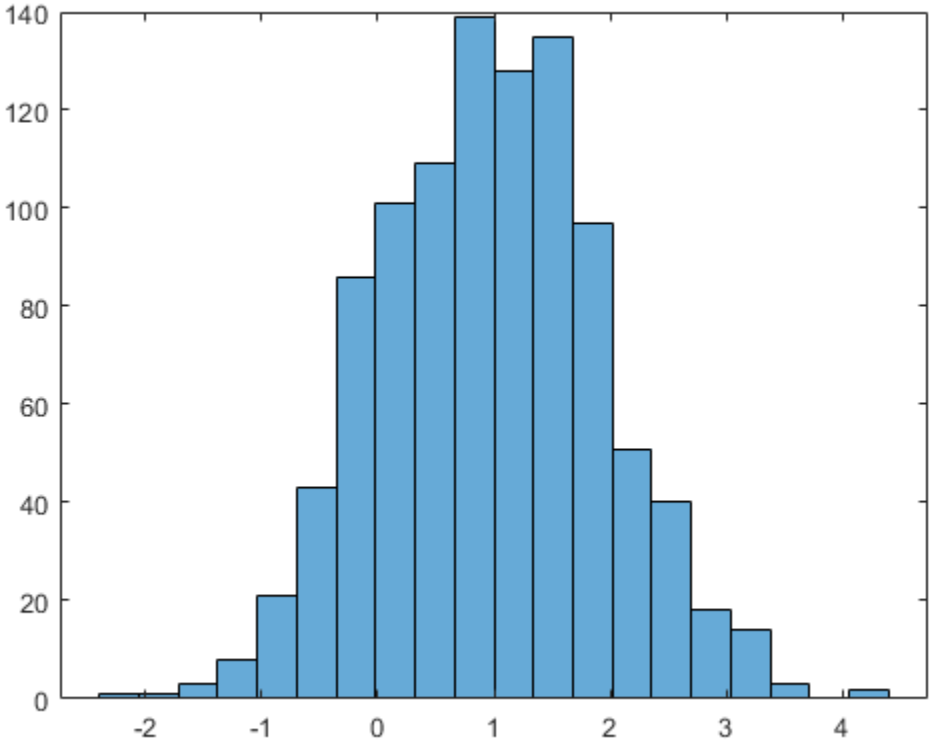


The histogram shows that the random numbers generated using the Weibull inverse cdf function `wblinv` have a Weibull distribution.

Step 3. Generate random numbers from the standard normal distribution.

The same values in `u` can generate random numbers from any distribution, for example the standard normal, by following the same procedure using the inverse cdf of the desired distribution.

```
figure
x_norm = norminv(u,1,1);
histogram(x_norm,20)
```



The histogram shows that, by using the standard normal inverse cdf `norminv`, the random numbers generated from `u` now have a standard normal distribution.

See Also

`hist` | `norminv` | `rand` | `wblinv`

More About

- “Uniform Distribution (Continuous)” on page B-163
- “Weibull Distribution” on page B-170
- “Normal Distribution” on page B-119
- “Random Number Generation” on page 5-27

- “Generate Random Numbers Using the Triangular Distribution” on page 5-47
- “Generating Pseudorandom Numbers” on page 7-2

Represent Cauchy Distribution Using t Location-Scale

This example shows how to use the t location-scale probability distribution object to work with a Cauchy distribution with nonstandard parameter values.

Step 1. Create a probability distribution object.

Create a *t* location-scale probability distribution object with degrees of freedom `nu = 1`. Specify `mu = 3` to set the location parameter equal to 3, and `sigma = 1` to set the scale parameter equal to 1.

```
pd = makedist('tLocationScale','mu',3,'sigma',1,'nu',1)
```

```
pd =
  tLocationScaleDistribution

  t Location-Scale distribution
       mu = 3
    sigma = 1
       nu = 1
```

Step 2. Compute descriptive statistics.

Use object functions to compute descriptive statistics for the Cauchy distribution.

```
med = median(pd)
```

```
med = 3
```

```
r = iqr(pd)
```

```
r = 2
```

```
m = mean(pd)
```

```
m = NaN
```

```
s = std(pd)
```

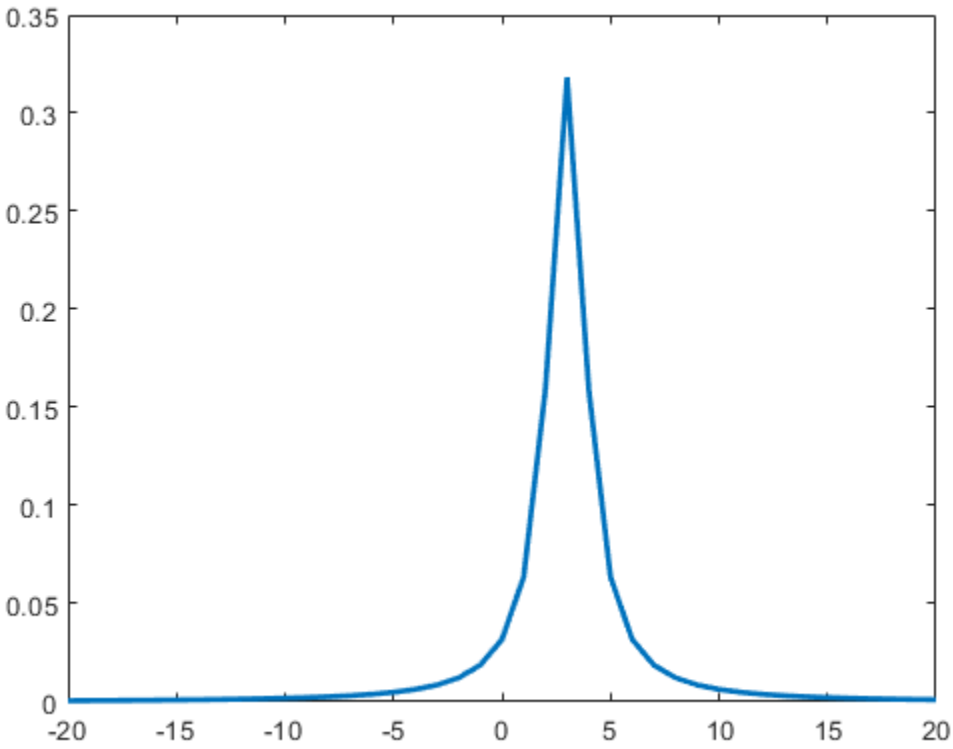
```
s = Inf
```

The median of the Cauchy distribution is equal to its location parameter, and the interquartile range is equal to two times its scale parameter. Its mean and standard deviation are undefined.

Step 3. Compute and plot the pdf.

Compute and plot the pdf of the Cauchy distribution.

```
x = -20:1:20;
y = pdf(pd,x);
plot(x,y,'LineWidth',2)
```



The peak of the pdf is centered at the location parameter `mu = 3`.

Step 4. Generate a vector of Cauchy random numbers.

Generate a column vector containing 10 random numbers from the Cauchy distribution using the `random` function for the *t* location-scale probability distribution object.

```
rng('default');  % For reproducibility
r = random(pd,10,1)
```

```
r = 10×1

    3.2678
    4.6547
    2.0604
    4.7322
    3.1810
    1.6649
    1.8471
    4.2466
    5.4647
    8.8874
```

Step 5. Generate a matrix of Cauchy random numbers.

Generate a 5-by-5 matrix of Cauchy random numbers.

```
r = random(pd,5,5)
```

```
r = 5×5

    2.2867    2.9692   -1.7003    5.5949    1.9806
    2.7421    2.7180    3.2210    2.4233    3.1394
    3.5966    3.9806    1.0182    6.4180    5.1367
    5.4791   15.6472    0.7558    2.8908    5.9031
    1.6863    4.0985    2.9934   13.9506    4.8792
```

See Also

`makedist`

More About

- “t Location-Scale Distribution” on page B-156
- “Generate Cauchy Random Numbers Using Student's t” on page 5-107

Generate Cauchy Random Numbers Using Student's t

This example shows how to use the Student's t distribution to generate random numbers from a standard Cauchy distribution.

Step 1. Generate a vector of random numbers.

Generate a column vector containing 10 random numbers from a standard Cauchy distribution, which has a location parameter `mu = 0` and scale parameter `sigma = 1`. Use `trnd` with degrees of freedom `V = 1`.

```
rng('default');  % For reproducibility
r = trnd(1,10,1)
```

```
r = 10×1

    0.2678
    1.6547
   -0.9396
    1.7322
    0.1810
   -1.3351
   -1.1529
    1.2466
    2.4647
    5.8874
```

Step 2. Generate a matrix of random numbers.

Generate a 5-by-5 matrix of random numbers from a standard Cauchy distribution.

```
r = trnd(1,5,5)
```

```
r = 5×5

   -0.7133   -0.0308   -4.7003    2.5949   -1.0194
   -0.2579   -0.2820    0.2210   -0.5767    0.1394
    0.5966    0.9806   -1.9818    3.4180    2.1367
    2.4791   12.6472   -2.2442   -0.1092    2.9031
   -1.3137    1.0985   -0.0066   10.9506    1.8792
```

See Also

`trnd`

More About

- “Student's t Distribution” on page B-149
- “Represent Cauchy Distribution Using t Location-Scale” on page 5-104

Generate Correlated Data Using Rank Correlation

This example shows how to use a copula and rank correlation to generate correlated data from probability distributions that do not have an inverse cdf function available, such as the Pearson flexible distribution family.

Step 1. Generate Pearson random numbers.

Generate 1000 random numbers from two different Pearson distributions, using the `pearsrnd` function. The first distribution has the parameter values mu equal to 0, sigma equal to 1, skew equal to 1, and kurtosis equal to 4. The second distribution has the parameter values mu equal to 0, sigma equal to 1, skew equal to 0.75, and kurtosis equal to 3.

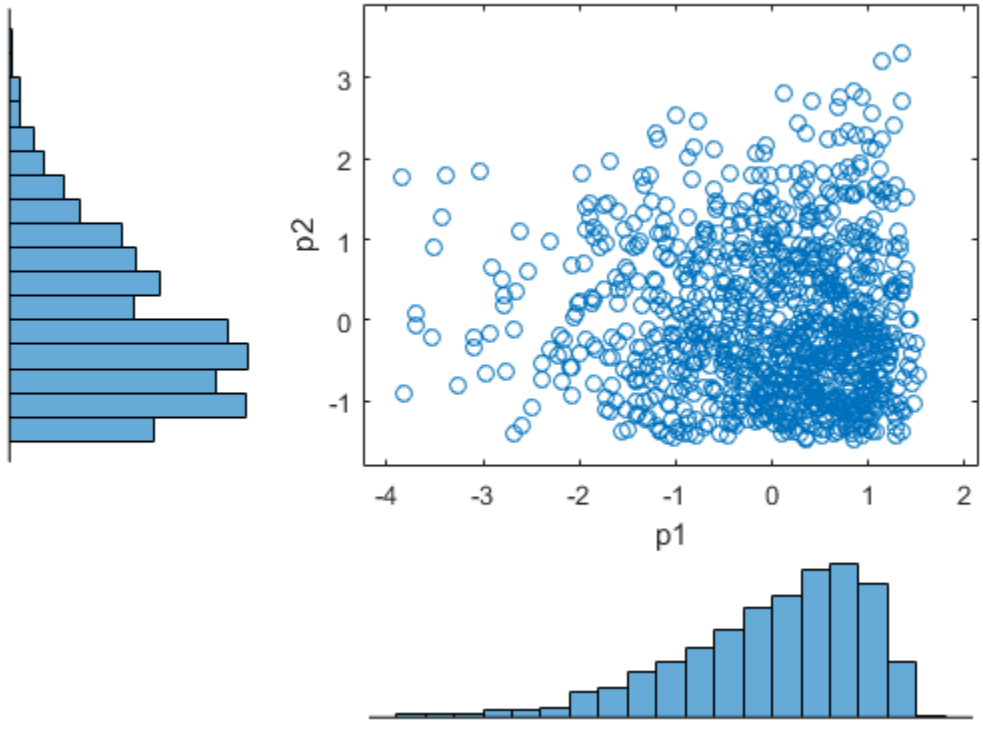
```
rng default  % For reproducibility
p1 = pearsrnd(0,1,-1,4,1000,1);
p2 = pearsrnd(0,1,0.75,3,1000,1);
```

At this stage, `p1` and `p2` are independent samples from their respective Pearson distributions, and are uncorrelated.

Step 2. Plot the Pearson random numbers.

Create a `scatterhist` plot to visualize the Pearson random numbers.

```
figure
scatterhist(p1,p2)
```

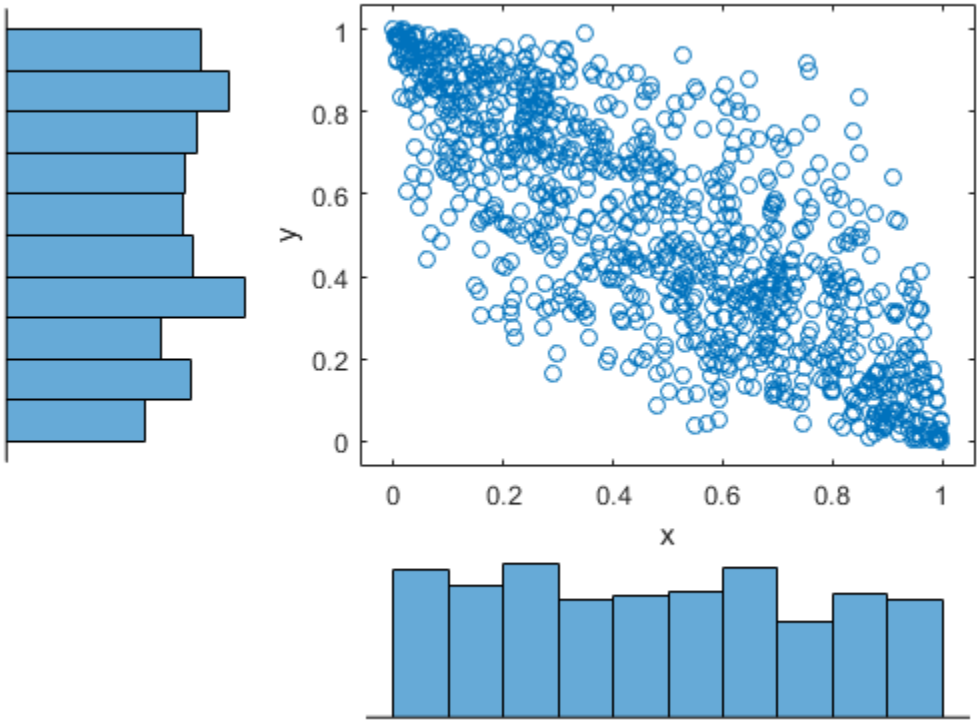


The histograms show the marginal distributions for `p1` and `p2`. The scatterplot shows the joint distribution for `p1` and `p2`. The lack of pattern to the scatterplot shows that `p1` and `p2` are independent.

Step 3. Generate random numbers using a Gaussian copula.

Use `copularnd` to generate 1000 correlated random numbers with a correlation coefficient equal to –0.8, using a Gaussian copula. Create a `scatterhist` plot to visualize the random numbers generated from the copula.

```
u = copularnd('Gaussian',-0.8,1000);
figure
scatterhist(u(:,1),u(:,2))
```



The histograms show that the data in each column of the copula have a marginal uniform distribution. The scatterplot shows that the data in the two columns are negatively correlated.

Step 4. Sort the copula random numbers.

Using Spearman's rank correlation, transform the two independent Pearson samples into correlated data.

Use the `sort` function to sort the copula random numbers from smallest to largest, and to return a vector of indices describing the rearranged order of the numbers.

```
[s1,i1] = sort(u(:,1));
[s2,i2] = sort(u(:,2));
```

`s1` and `s2` contain the numbers from the first and second columns of the copula, `u`, sorted in order from smallest to largest. `i1` and `i2` are index vectors that describe the rearranged order of the elements into `s1` and `s2`. For example, if the first value in the sorted vector `s1` is the third value in the original unsorted vector, then the first value in the index vector `i1` is 3.

Step 5. Transform the Pearson samples using Spearman's rank correlation.

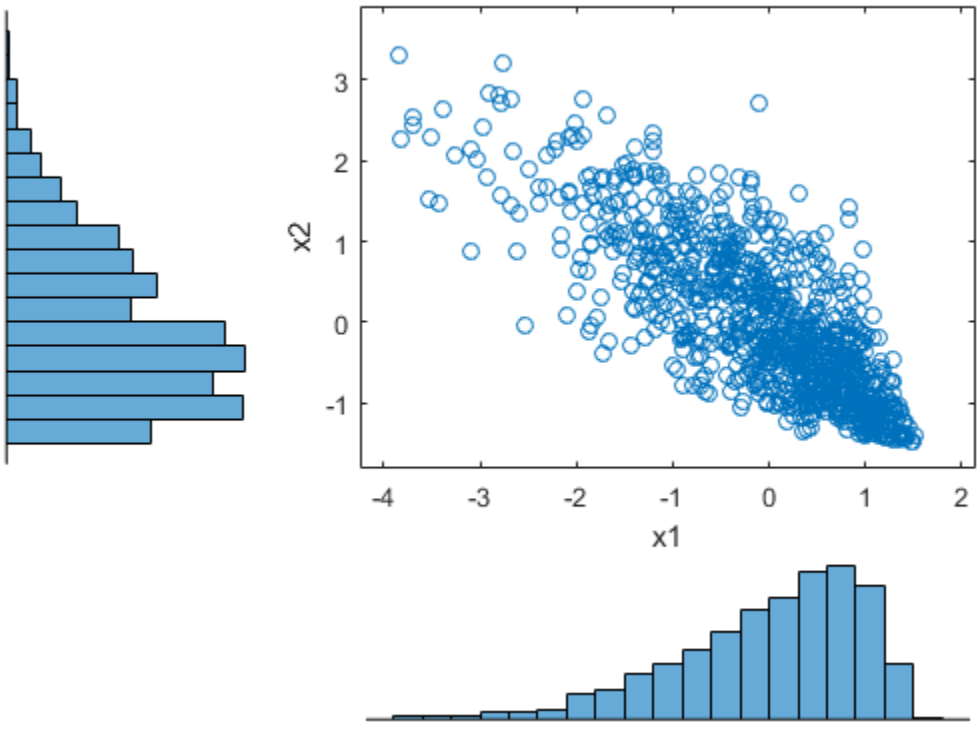
Create two vectors of zeros, `x1` and `x2`, that are the same size as the sorted copula vectors, `s1` and `s2`. Sort the values in `p1` and `p2` from smallest to largest. Place the values into `x1` and `x2`, in the same order as the indices `i1` and `i2` generated by sorting the copula random numbers.

```
x1 = zeros(size(s1));
x2 = zeros(size(s2));

x1(i1) = sort(p1);
x2(i2) = sort(p2);
```

Step 6. Plot the correlated Pearson random numbers.

Create a `scatterhist` plot to visualize the correlated Pearson data.

```
figure
scatterhist(x1,x2)
```



The histograms show the marginal Pearson distributions for each column of data. The scatterplot shows the joint distribution of `p1` and `p2`, and indicates that the data are now negatively correlated.

Step 7. Confirm Spearman rank correlation coefficient values.

Confirm that the Spearman rank correlation coefficient is the same for the copula random numbers and the correlated Pearson random numbers.

```
copula_corr = corr(u,'Type','spearman')
```

```
copula_corr = 2×2

    1.0000   -0.7858
   -0.7858    1.0000
```

```
pearson_corr = corr([x1,x2],'Type','spearman')
```

```
pearson_corr = 2×2

    1.0000   -0.7858
   -0.7858    1.0000
```

The Spearman rank correlation is the same for the copula and the Pearson random numbers.

See Also

`copularnd` | `corr` | `sort`

More About

- “Copulas: Generate Correlated Samples” on page 5-121

Create Gaussian Mixture Model

This example shows how to create a known, or fully specified, Gaussian mixture model (GMM) object using `gmdistribution` and by specifying component means, covariances, and mixture proportions. To create a GMM object by fitting data to a GMM, see “Fit Gaussian Mixture Model to Data” on page 5-115.

Specify the component means, covariances, and mixing proportions for a two-component mixture of bivariate Gaussian distributions.

```
mu = [1 2;-3 -5];                    % Means
sigma = cat(3,[2 0;0 .5],[1 0;0 1]); % Covariances
p = ones(1,2)/2;                     % Mixing proportions
```

The rows of `mu` correspond to the component mean vectors, and the pages of `sigma`, `sigma(:,:,J)`, correspond to the component covariance matrices.

Create a GMM object using `gmdistribution`.

```
gm = gmdistribution(mu,sigma,p);
```

Display the properties of the GMM.

```
properties(gm)
```

```
Properties for class gmdistribution:

    NumVariables
    DistributionName
    NumComponents
    ComponentProportion
    SharedCovariance
    NumIterations
    RegularizationValue
    NegativeLogLikelihood
    CovarianceType
    mu
    Sigma
    AIC
    BIC
    Converged
    ProbabilityTolerance
```

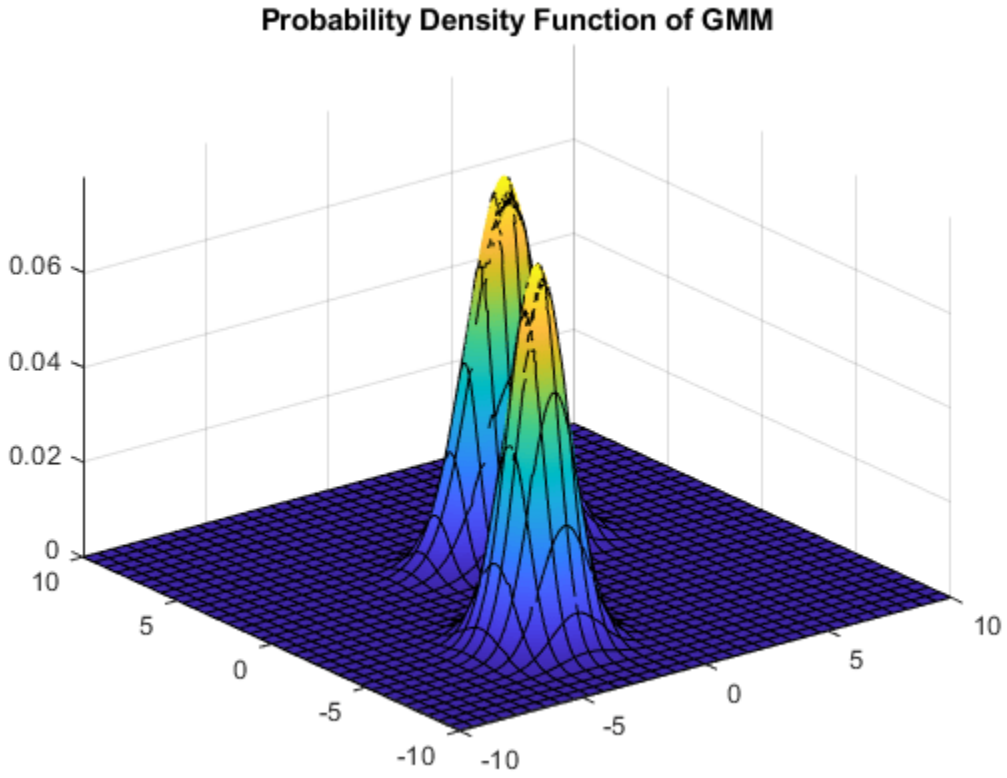
For a description of the properties, see `gmdistribution`. To access the value of a property, use dot notation. For example, access the number of variables of each GMM component.

```
dimension = gm.NumVariables
```

```
dimension = 2
```

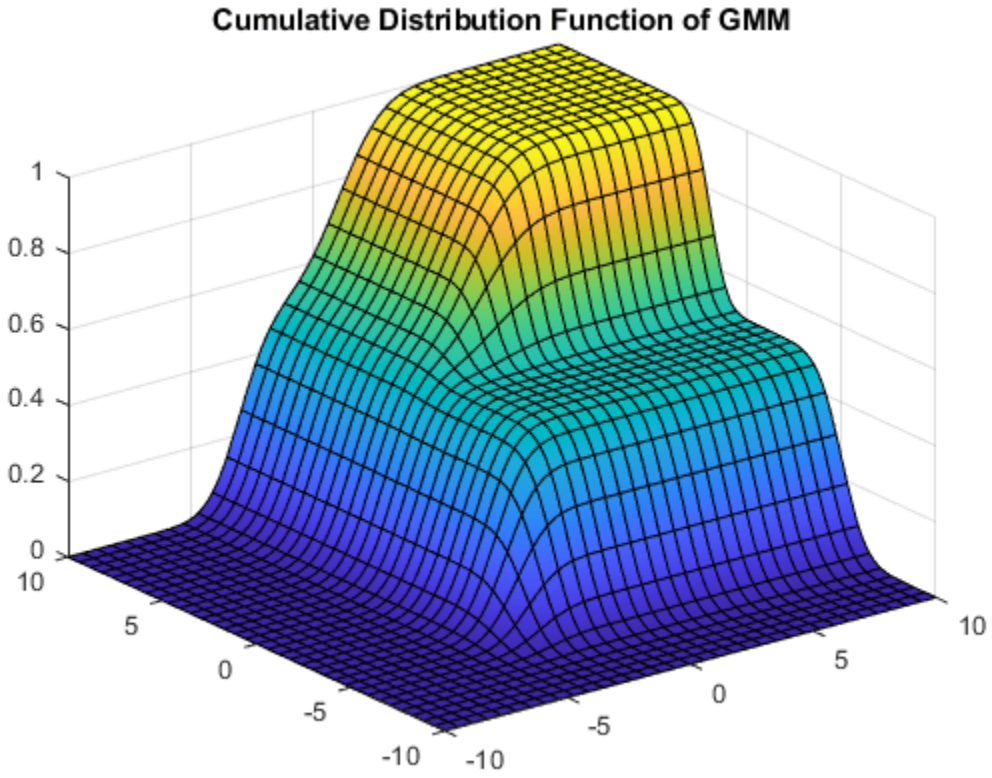
Visualize the probability density function (pdf) of the GMM using `pdf` and the MATLAB® function `fsurf`.

```
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fsurf(gmPDF,[-10 10])
title('Probability Density Function of GMM');
```



Visualize the cumulative distribution function (cdf) of the GMM using `cdf` and `fsurf`.

```
gmCDF = @(x,y) arrayfun(@(x0,y0) cdf(gm,[x0 y0]),x,y);
fsurf(gmCDF,[-10 10])
title('Cumulative Distribution Function of GMM');
```



See Also

`fitgmdist` | `gmdistribution`

More About

- “Fit Gaussian Mixture Model to Data” on page 5-115
- “Simulate Data from Gaussian Mixture Model” on page 5-119
- “Cluster Using Gaussian Mixture Model” on page 16-39

Fit Gaussian Mixture Model to Data

This example shows how to simulate data from a multivariate normal distribution, and then fit a Gaussian mixture model (GMM) to the data using `fitgmdist`. To create a known, or fully specified, GMM object, see “Create Gaussian Mixture Model” on page 5-112.

`fitgmdist` requires a matrix of data and the number of components in the GMM. To create a useful GMM, you must choose `k` carefully. Too few components fails to model the data accurately (i.e., underfitting to the data). Too many components leads to an over-fit model with singular covariance matrices.

Simulate data from a mixture of two bivariate Gaussian distributions using `mvnrnd`.

```
mu1 = [1 2];
sigma1 = [2 0; 0 .5];
mu2 = [-3 -5];
sigma2 = [1 0; 0 1];
rng(1); % For reproducibility
X = [mvnrnd(mu1,sigma1,1000);
     mvnrnd(mu2,sigma2,1000)];
```

Plot the simulated data.

```
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
title('Simulated Data')
```

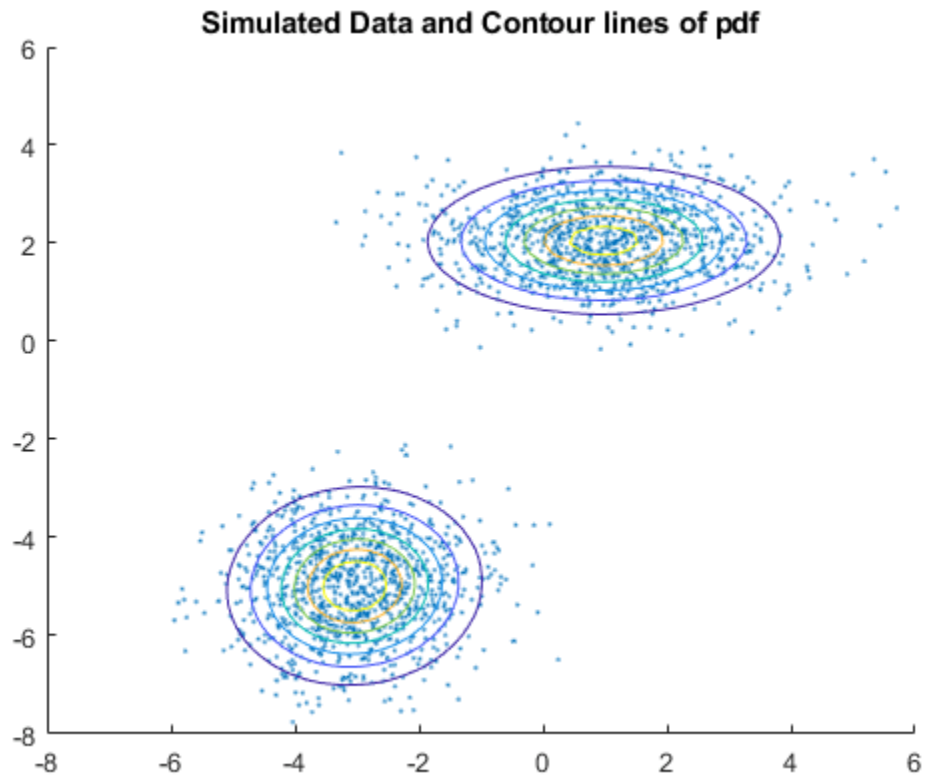
Fit a two-component GMM. Use the `'Options'` name-value pair argument to display the final output of the fitting algorithm.

```
options = statset('Display','final');
gm = fitgmdist(X,2,'Options',options)
```

```
5 iterations, log-likelihood = -7105.71

gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:   -3.0377   -4.9859

Component 2:
Mixing proportion: 0.500000
Mean:    0.9812    2.0563
```

Plot the pdf of the fitted GMM.

```
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
hold on
h = fcontour(gmPDF,[-8 6]);
title('Simulated Data and Contour lines of pdf');
```



Display the estimates for means, covariances, and mixture proportions

```
ComponentMeans = gm.mu
```

```
ComponentMeans = 2×2

   -3.0377   -4.9859
    0.9812    2.0563
```

```
ComponentCovariances = gm.Sigma
```

```
ComponentCovariances =
ComponentCovariances(:,:,1) =

    1.0132    0.0482
    0.0482    0.9796


ComponentCovariances(:,:,2) =

    1.9919    0.0127
    0.0127    0.5533
```

```
MixtureProportions = gm.ComponentProportion
```

```
MixtureProportions = 1×2

    0.5000    0.5000
```

Fit four models to the data, each with an increasing number of components, and compare the Akaike Information Criterion (AIC) values.

```
AIC = zeros(1,4);
gm = cell(1,4);
for k = 1:4
    gm{k} = fitgmdist(X,k);
    AIC(k)= gm{k}.AIC;
end
```

Display the number of components that minimizes the AIC value.

```
[minAIC,numComponents] = min(AIC);
numComponents
```

```
numComponents = 2
```

The two-component model has the smallest AIC value.

Display the two-component GMM.

```
gm2 = gm{numComponents}
```

```
gm2 =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:   -3.0377   -4.9859
```

```
Component 2:
Mixing proportion: 0.500000
Mean:    0.9812    2.0563
```

Both the AIC and Bayesian information criteria (BIC) are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). You can use them to determine an appropriate number of components for a model when the number of components is unspecified.

See Also

`fitgmdist` | `gmdistribution` | `mvnrnd` | `random`

More About

- “Create Gaussian Mixture Model” on page 5-112
- “Simulate Data from Gaussian Mixture Model” on page 5-119
- “Cluster Using Gaussian Mixture Model” on page 16-39

Simulate Data from Gaussian Mixture Model

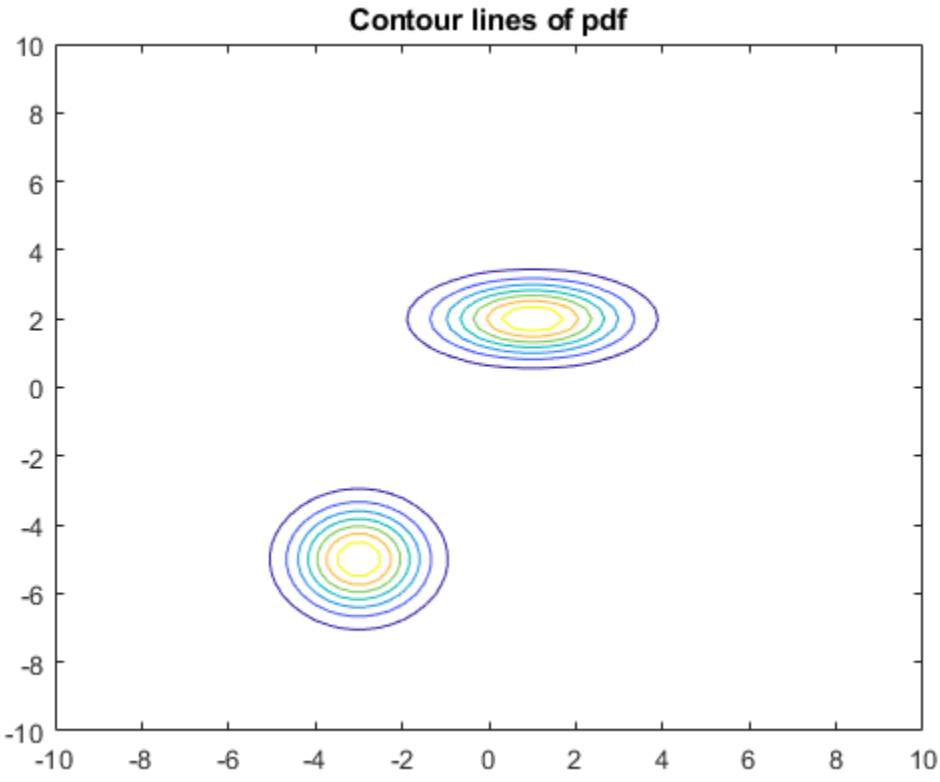
This example shows how to simulate data from a Gaussian mixture model (GMM) using a fully specified `gmdistribution` object and the `random` function.

Create a known, two-component GMM object.

```
mu = [1 2;-3 -5];
sigma = cat(3,[2 0;0 .5],[1 0;0 1]);
p = ones(1,2)/2;
gm = gmdistribution(mu,sigma,p);
```

Plot the contour of the pdf of the GMM.

```
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fcontour(gmPDF,[-10 10]);
title('Contour lines of pdf');
```

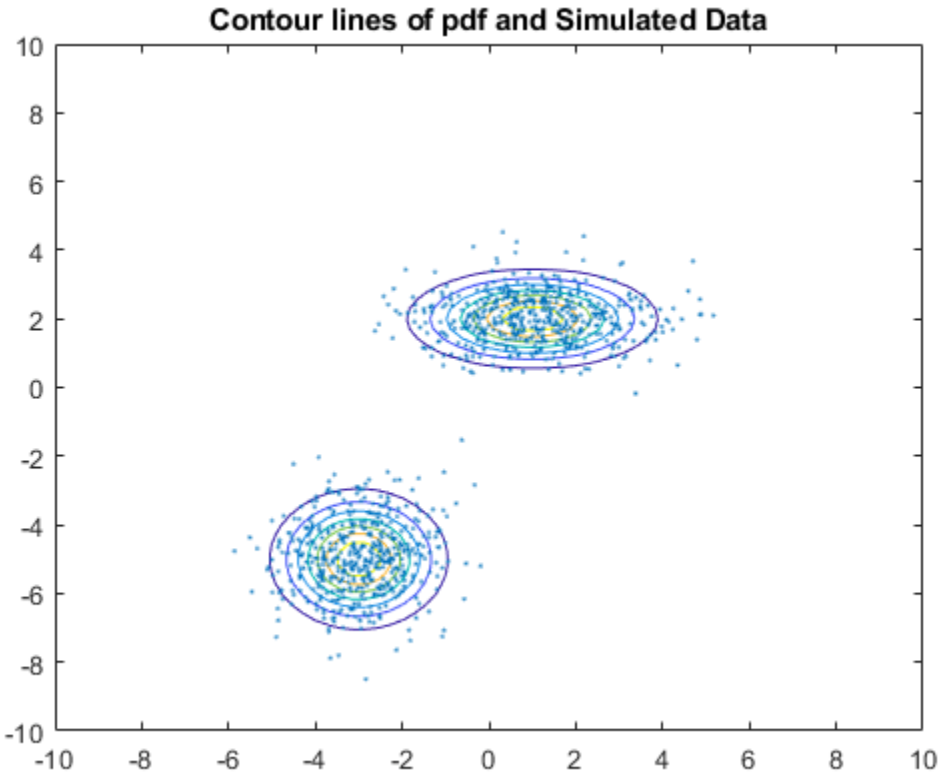


Generate 1000 random variates from the GMM.

```
rng('default') % For reproducibility
X = random(gm,1000);
```

Plot the variates with the pdf contours.

```
hold on
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
title('Contour lines of pdf and Simulated Data')
```



See Also

`fitgmdist` | `gmdistribution` | `mvnrnd` | `random`

More About

- “Create Gaussian Mixture Model” on page 5-112
- “Fit Gaussian Mixture Model to Data” on page 5-115
- “Cluster Using Gaussian Mixture Model” on page 16-39

Copulas: Generate Correlated Samples

In this section...

Copulas are functions that describe dependencies among variables, and provide a way to create distributions that model correlated multivariate data. Using a copula, you can construct a multivariate distribution by specifying marginal univariate distributions, and then choose a copula to provide a correlation structure between variables. Bivariate distributions, as well as distributions in higher dimensions, are possible.

Determining Dependence Between Simulation Inputs

One of the design decisions for a Monte Carlo simulation is a choice of probability distributions for the random inputs. Selecting a distribution for each individual variable is often straightforward, but deciding what dependencies should exist between the inputs may not be. Ideally, input data to a simulation should reflect what you know about dependence among the real quantities you are modeling. However, there may be little or no information on which to base any dependence in the simulation. In such cases, it is useful to experiment with different possibilities in order to determine the model's sensitivity.

It can be difficult to generate random inputs with dependence when they have distributions that are not from a standard multivariate distribution. Further, some of the standard multivariate distributions can model only limited types of dependence. It is always possible to make the inputs independent, and while that is a simple choice, it is not always sensible and can lead to the wrong conclusions.

For example, a Monte-Carlo simulation of financial risk could have two random inputs that represent different sources of insurance losses. You could model these inputs as lognormal random variables. A reasonable question to ask is how dependence between these two inputs affects the results of the simulation. Indeed, you might know from real data that the same random conditions affect both sources; ignoring that in the simulation could lead to the wrong conclusions.

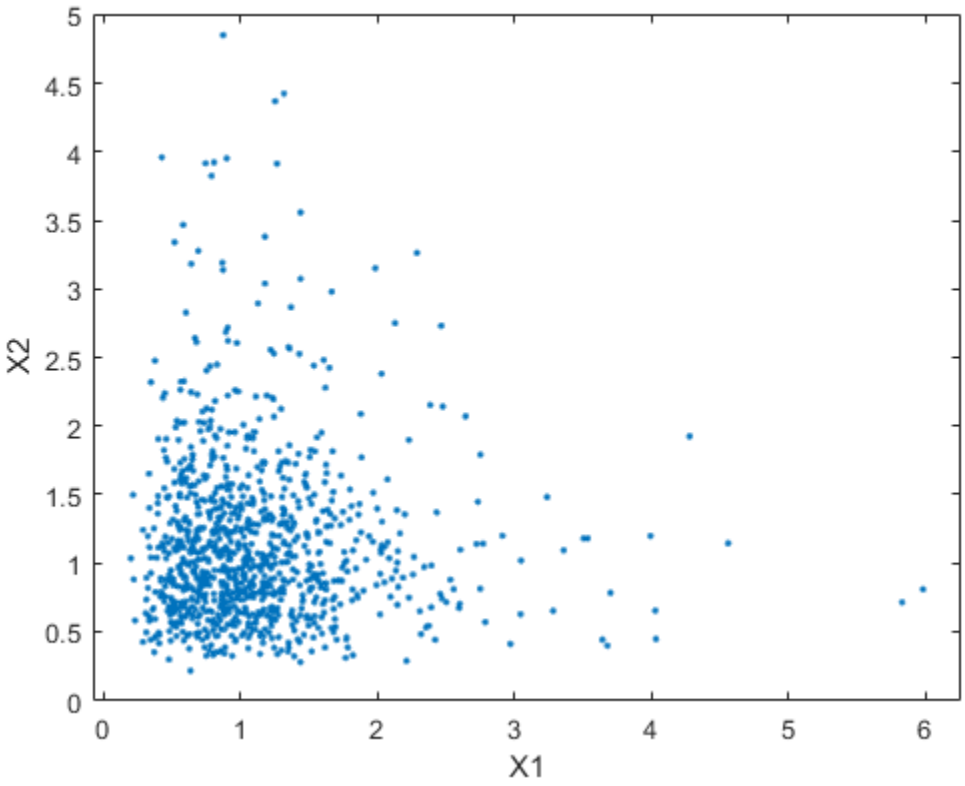
Generate and Exponentiate Normal Random Variables

The `lognrnd` function simulates independent lognormal random variables. In the following example, the `mvnrnd` function generates `n` pairs of independent normal random variables, and then exponentiates them. Notice that the covariance matrix used here is diagonal.

```
n = 1000;

sigma = .5;
SigmaInd = sigma.^2 .* [1 0; 0 1]
```

```
SigmaInd = 2×2

    0.2500         0
         0    0.2500
```

```
rng('default');  % For reproducibility
ZInd = mvnrnd([0 0],SigmaInd,n);
XInd = exp(ZInd);

plot(XInd(:,1),XInd(:,2),'.')
axis([0 5 0 5])
axis equal
xlabel('X1')
ylabel('X2')
```



Dependent bivariate lognormal random variables are also easy to generate using a covariance matrix with nonzero off-diagonal terms.

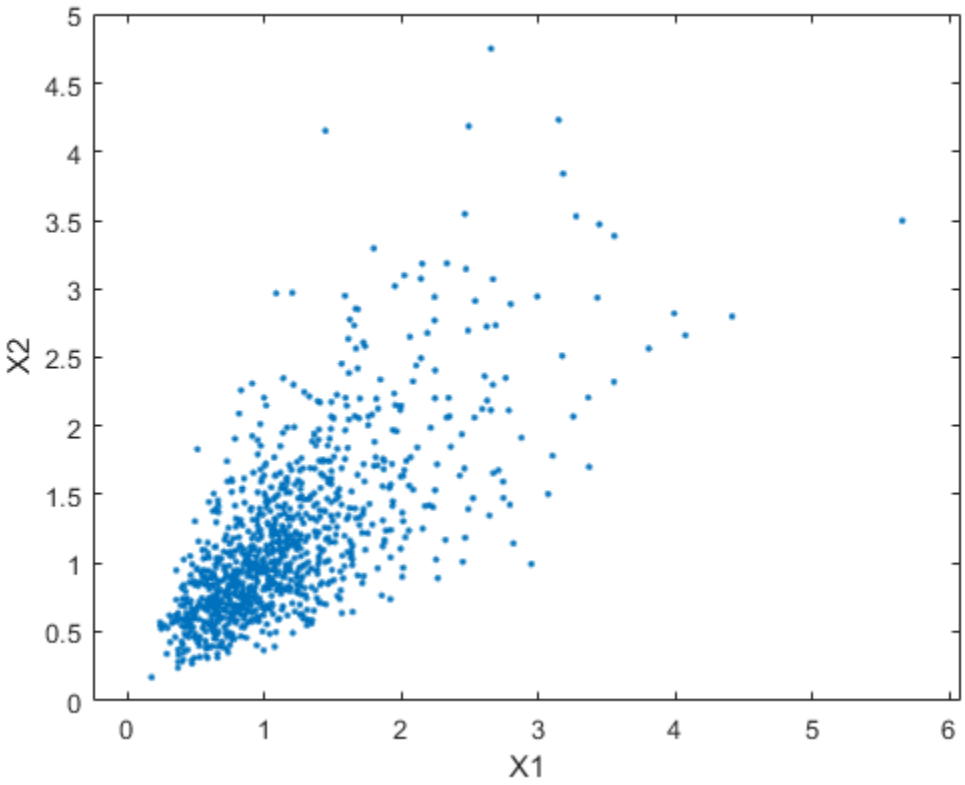
```
rho = .7;

SigmaDep = sigma.^2 .* [1 rho; rho 1]

SigmaDep = 2×2

    0.2500    0.1750
```

```
    0.1750    0.2500
```

```
ZDep = mvnrnd([0 0],SigmaDep,n);
XDep = exp(ZDep);
```

A second scatter plot demonstrates the difference between these two bivariate distributions.

```
plot(XDep(:,1),XDep(:,2),'.')
axis([0 5 0 5])
axis equal
xlabel('X1')
ylabel('X2')
```



It is clear that there is a tendency in the second data set for large values of X1 to be associated with large values of X2, and similarly for small values. The correlation parameter ρ of the underlying bivariate normal determines this dependence. The conclusions drawn from the simulation could well depend on whether you generate X1 and X2 with dependence. The bivariate lognormal distribution is a simple solution in this case; it easily generalizes to higher dimensions in cases where the marginal distributions are different lognormals.

Other multivariate distributions also exist. For example, the multivariate *t* and the Dirichlet distributions simulate dependent *t* and beta random variables, respectively. But the list of simple multivariate distributions is not long, and they only apply in cases where the marginals are all in the same family (or even the exact same distributions). This can be a serious limitation in many situations.

Constructing Dependent Bivariate Distributions

Although the construction discussed in the previous section creates a bivariate lognormal that is simple, it serves to illustrate a method that is more generally applicable.

1. Generate pairs of values from a bivariate normal distribution. There is statistical dependence between these two variables, and each has a normal marginal distribution.
2. Apply a transformation (the exponential function) separately to each variable, changing the marginal distributions into lognormals. The transformed variables still have a statistical dependence.

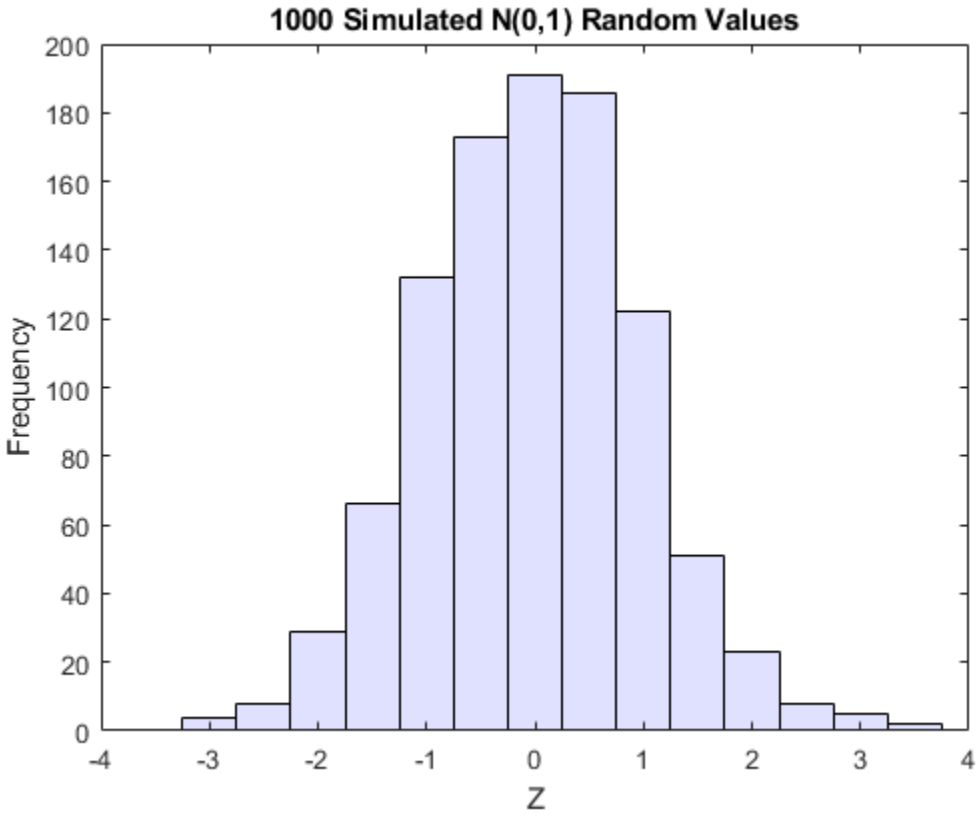
If a suitable transformation can be found, this method can be generalized to create dependent bivariate random vectors with other marginal distributions. In fact, a general method of constructing such a transformation does exist, although it is not as simple as exponentiation alone.

By definition, applying the normal cumulative distribution function (cdf), denoted here by Φ, to a standard normal random variable results in a random variable that is uniform on the interval [0,1]. To see this, if Z has a standard normal distribution, then the cdf of $U = \Phi(Z)$ is

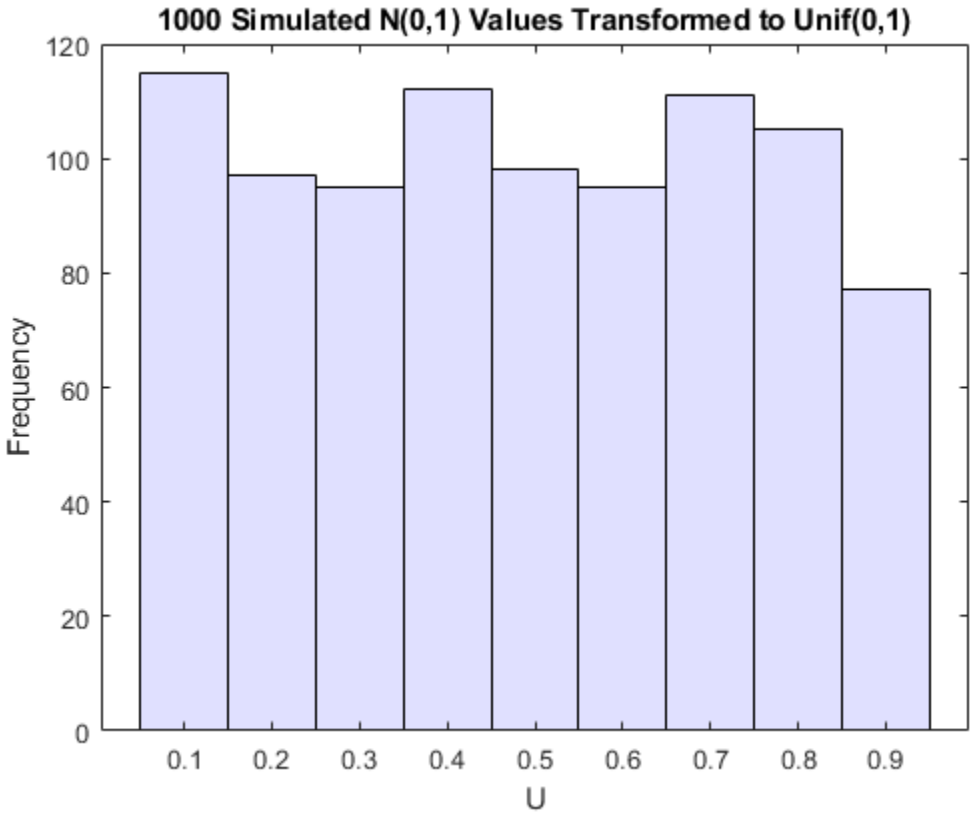
$$\Pr\{U \le u\} = \Pr\{\Phi(Z) \le u\} = \Pr\left\{Z \le \Phi^{-1}(u)\right\} = u$$

and that is the cdf of a Unif(0,1) random variable. Histograms of some simulated normal and transformed values demonstrate that fact:

```
n = 1000;
rng default % for reproducibility
z = normrnd(0,1,n,1); % generate standard normal data

histogram(z,-3.75:.5:3.75,'FaceColor',[.8 .8 1]) % plot the histogram of data
xlim([-4 4])
title('1000 Simulated N(0,1) Random Values')
xlabel('Z')
ylabel('Frequency')
```

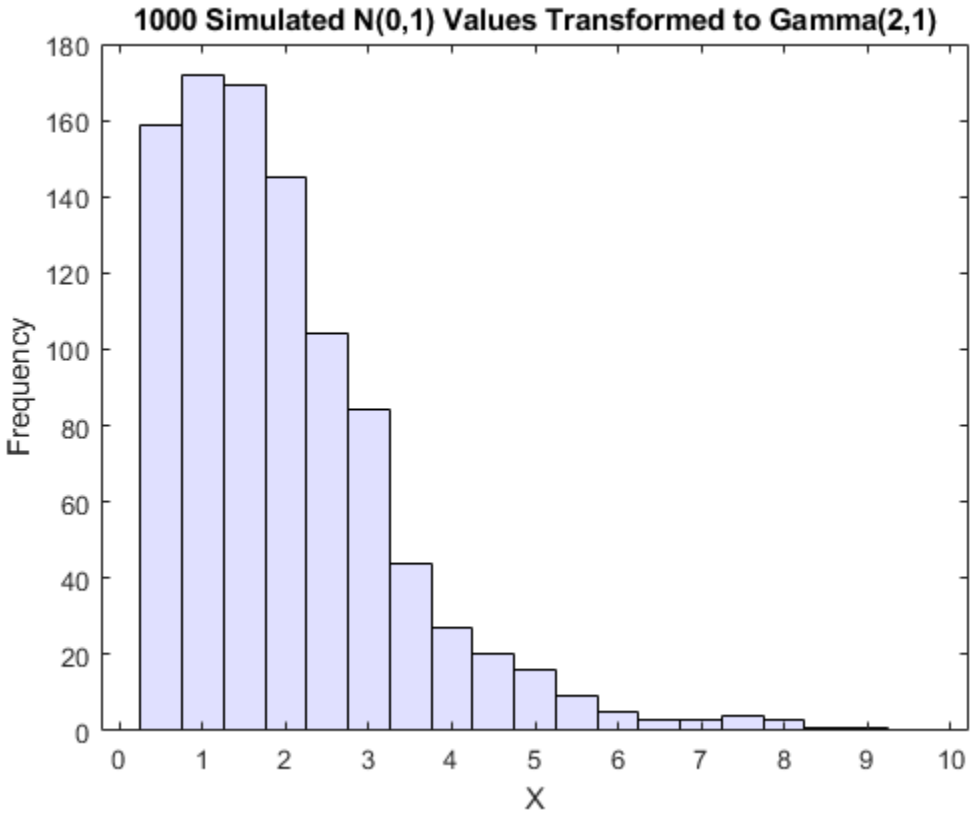



```
u = normcdf(z);  % compute the cdf values of the sample data

figure
histogram(u,.05:.1:.95,'FaceColor',[.8 .8 1]) % plot the histogram of the cdf values
title('1000 Simulated N(0,1) Values Transformed to Unif(0,1)')
xlabel('U')
ylabel('Frequency')
```



Borrowing from the theory of univariate random number generation, applying the inverse cdf of any distribution, *F*, to a Unif(0,1) random variable results in a random variable whose distribution is exactly *F* (see “Inversion Methods” on page 7-3). The proof is essentially the opposite of the preceding proof for the forward case. Another histogram illustrates the transformation to a gamma distribution:

```
x = gaminv(u,2,1); % transform to gamma values

figure
histogram(x,.25:.5:9.75,'FaceColor',[.8 .8 1]) % plot the histogram of gamma values
title('1000 Simulated N(0,1) Values Transformed to Gamma(2,1)')
xlabel('X')
ylabel('Frequency')
```



You can apply this two-step transformation to each variable of a standard bivariate normal, creating dependent random variables with arbitrary marginal distributions. Because the transformation works on each component separately, the two resulting random variables need not even have the same marginal distributions. The transformation is defined as:

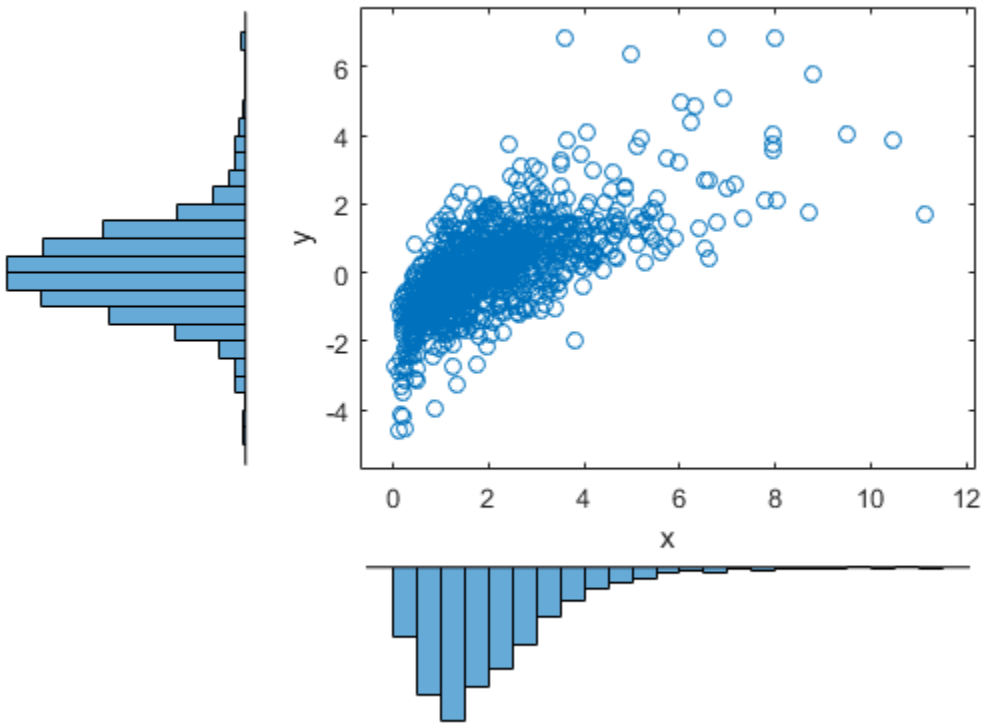
$$Z = [Z_1, Z_2] \sim N\left([0,0], \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

$$U = [\Phi(Z_1), \Phi(Z_2)]$$

$$X = [G_1(U_1), G_2(U_2)]$$

where G_1 and G_2 are inverse cdfs of two possibly different distributions. For example, the following generates random vectors from a bivariate distribution with t_5 and Gamma(2,1) marginals:

```
n = 1000; rho = .7;
Z = mvnrnd([0 0],[1 rho; rho 1],n);
U = normcdf(Z);
X = [gaminv(U(:,1),2,1) tinv(U(:,2),5)];

% draw the scatter plot of data with histograms
figure
scatterhist(X(:,1),X(:,2),'Direction','out')
```



This plot has histograms alongside a scatter plot to show both the marginal distributions, and the dependence.

Using Rank Correlation Coefficients

The correlation parameter, ρ, of the underlying bivariate normal determines the dependence between X1 and X2 in this construction. However, the linear correlation of X1 and X2 is not ρ. For example, in the original lognormal case, a closed form for that correlation is:

$$\mathrm{cor}(X1, X2) = \frac{e^{\rho\sigma^2} - 1}{e^{\sigma^2} - 1}$$

which is strictly less than ρ, unless ρ is exactly 1. In more general cases such as the Gamma/t construction, the linear correlation between X1 and X2 is difficult or impossible to express in terms of ρ, but simulations show that the same effect happens.

That is because the linear correlation coefficient expresses the linear dependence between random variables, and when nonlinear transformations are applied to those random variables, linear correlation is not preserved. Instead, a rank correlation coefficient, such as Kendall's τ or Spearman's ρ, is more appropriate.

Roughly speaking, these rank correlations measure the degree to which large or small values of one random variable associate with large or small values of another. However, unlike the linear

correlation coefficient, they measure the association only in terms of ranks. As a consequence, the rank correlation is preserved under any monotonic transformation. In particular, the transformation method just described preserves the rank correlation. Therefore, knowing the rank correlation of the bivariate normal Z exactly determines the rank correlation of the final transformed random variables, X. While the linear correlation coefficient, ρ, is still needed to parameterize the underlying bivariate normal, Kendall's τ or Spearman's ρ are more useful in describing the dependence between random variables, because they are invariant to the choice of marginal distribution.

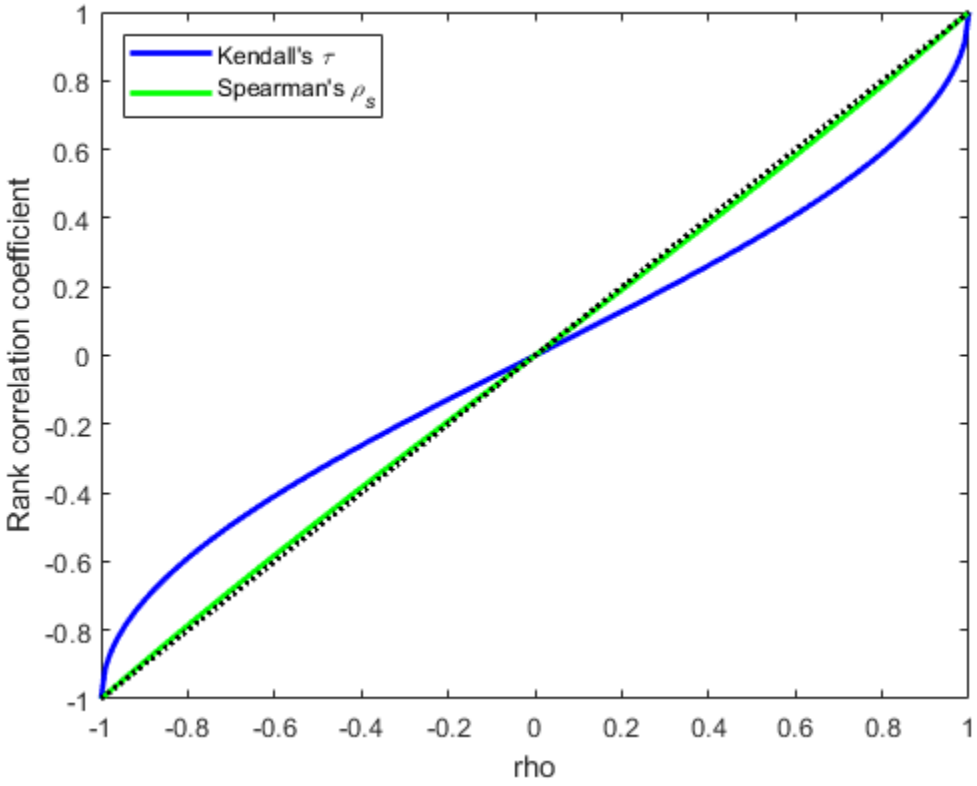
For the bivariate normal, there is a simple one-to-one mapping between Kendall's τ or Spearman's ρ, and the linear correlation coefficient ρ:

$$\tau = \frac{2}{\pi}\arcsin(\rho) \text{ or } \rho = \sin\left(\tau\frac{\pi}{2}\right)$$

$$\rho_S = \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right) \text{ or } \rho = 2\sin\left(\rho_S\frac{\pi}{6}\right)$$

The following plot shows the relationship.

```
rho = -1:.01:1;
tau = 2.*asin(rho)./pi;
rho_s = 6.*asin(rho./2)./pi;

plot(rho,tau,'b-','LineWidth',2)
hold on
plot(rho,rho_s,'g-','LineWidth',2)
plot([-1 1],[-1 1],'k:','LineWidth',2)
axis([-1 1 -1 1])
xlabel('rho')
ylabel('Rank correlation coefficient')
legend('Kendall''s {\it\tau}', ...
       'Spearman''s {\it\rho_s}', ...
       'location','NW')
```



Thus, it is easy to create the desired rank correlation between `X1` and `X2`, regardless of their marginal distributions, by choosing the correct ρ parameter value for the linear correlation between `Z1` and `Z2`.

For the multivariate normal distribution, Spearman's rank correlation is almost identical to the linear correlation. However, this is not true once you transform to the final random variables.

Using Bivariate Copulas

The first step of the construction described in the previous section defines what is known as a bivariate Gaussian copula. A copula is a multivariate probability distribution, where each random variable has a uniform marginal distribution on the unit interval `[0,1]`. These variables may be completely independent, deterministically related (e.g., `U2 = U1`), or anything in between. Because of the possibility for dependence among variables, you can use a copula to construct a new multivariate distribution for dependent variables. By transforming each of the variables in the copula separately using the inversion method, possibly using different cdfs, the resulting distribution can have arbitrary marginal distributions. Such multivariate distributions are often useful in simulations, when you know that the different random inputs are not independent of each other.

Statistics and Machine Learning Toolbox functions compute:

- Probability density functions (`copulapdf`) and the cumulative distribution functions (`copulacdf`) for Gaussian copulas

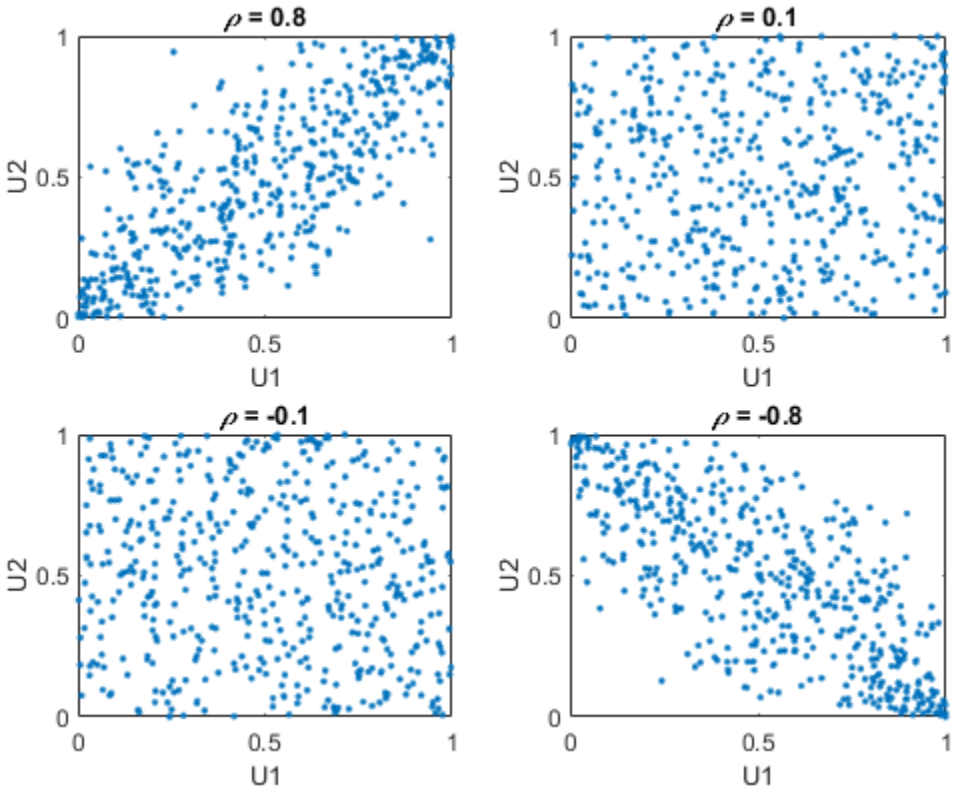
- Rank correlations from linear correlations (`copulastat`) and vice versa (`copulaparam`)
- Random vectors (`copularnd`)
- Parameters for copulas fit to data (`copulafit`)

For example, use the `copularnd` function to create scatter plots of random values from a bivariate Gaussian copula for various levels of ρ, to illustrate the range of different dependence structures. The family of bivariate Gaussian copulas is parameterized by the linear correlation matrix:

$$P = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

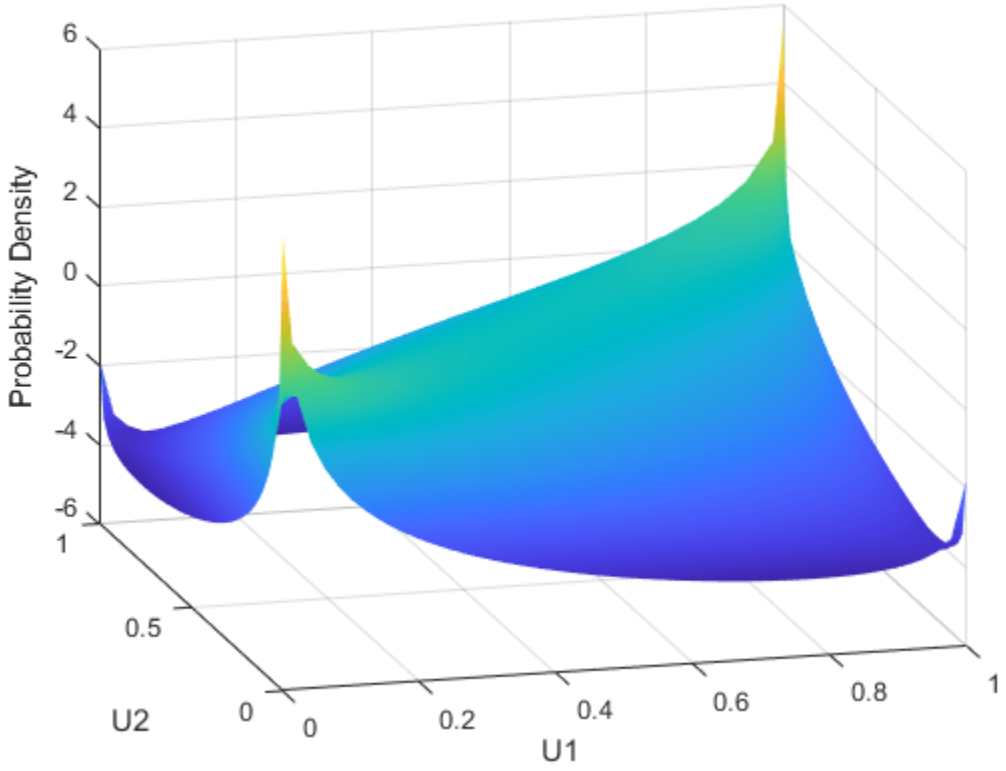
`U1` and `U2` approach linear dependence as ρ approaches ± 1, and approach complete independence as ρ approaches zero:

```
n = 500;

rng('default') % for reproducibility
U = copularnd('Gaussian',[1 .8; .8 1],n);
subplot(2,2,1)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = 0.8')
xlabel('U1')
ylabel('U2')

U = copularnd('Gaussian',[1 .1; .1 1],n);
subplot(2,2,2)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = 0.1')
xlabel('U1')
ylabel('U2')

U = copularnd('Gaussian',[1 -.1; -.1 1],n);
subplot(2,2,3)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = -0.1')
xlabel('U1')
ylabel('U2')

U = copularnd('Gaussian',[1 -.8; -.8 1],n);
subplot(2,2,4)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = -0.8')
xlabel('U1')
ylabel('U2')
```

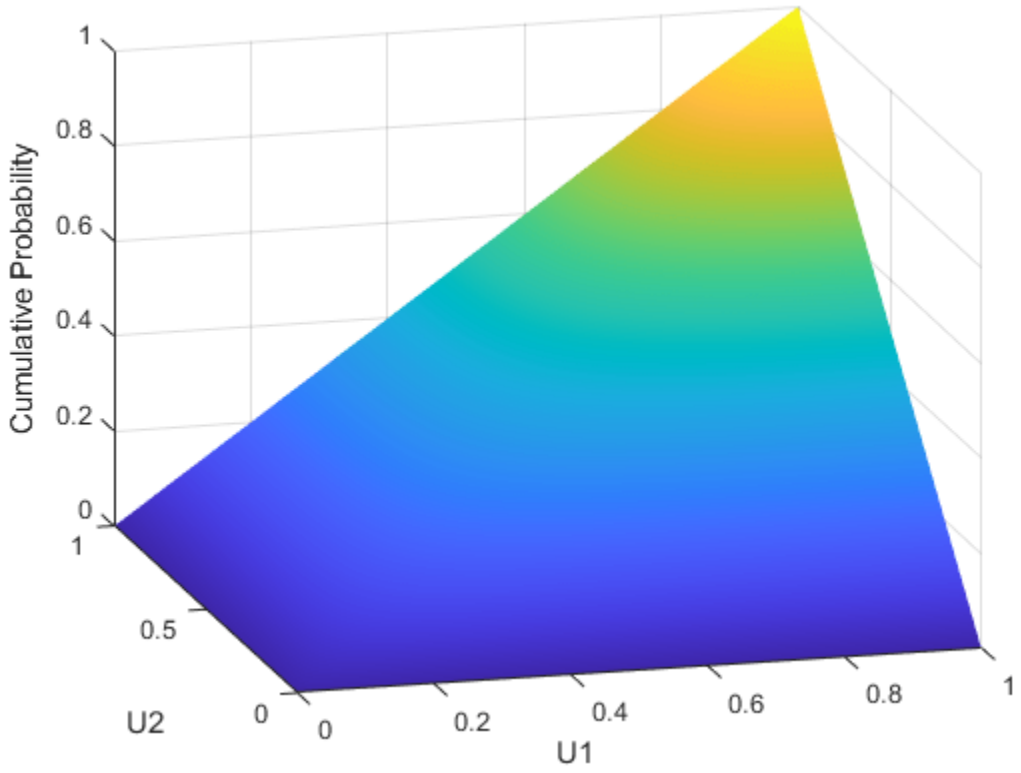


The dependence between `U1` and `U2` is completely separate from the marginal distributions of `X1 = G(U1)` and `X2 = G(U2)`. `X1` and `X2` can be given any marginal distributions, and still have the same rank correlation. This is one of the main appeals of copulas—they allow this separate specification of dependence and marginal distribution. You can also compute the pdf (`copulapdf`) and the cdf (`copulacdf`) for a copula. For example, these plots show the pdf and cdf for $\rho = .8$:

```
u1 = linspace(1e-3,1-1e-3,50);
u2 = linspace(1e-3,1-1e-3,50);
[U1,U2] = meshgrid(u1,u2);
Rho = [1 .8; .8 1];
f = copulapdf('t',[U1(:) U2(:)],Rho,5);
f = reshape(f,size(U1));

figure
surf(u1,u2,log(f),'FaceColor','interp','EdgeColor','none')
view([-15,20])
xlabel('U1')
ylabel('U2')
zlabel('Probability Density')
```




```
u1 = linspace(1e-3,1-1e-3,50);
u2 = linspace(1e-3,1-1e-3,50);
[U1,U2] = meshgrid(u1,u2);
F = copulacdf('t',[U1(:) U2(:)],Rho,5);
F = reshape(F,size(U1));

figure()
surf(u1,u2,F,'FaceColor','interp','EdgeColor','none')
view([-15,20])
xlabel('U1')
ylabel('U2')
zlabel('Cumulative Probability')
```



A different family of copulas can be constructed by starting from a bivariate t distribution and transforming using the corresponding t cdf. The bivariate t distribution is parameterized with P, the linear correlation matrix, and ν, the degrees of freedom. Thus, for example, you can speak of a t_1 or a t_5 copula, based on the multivariate t with one and five degrees of freedom, respectively.

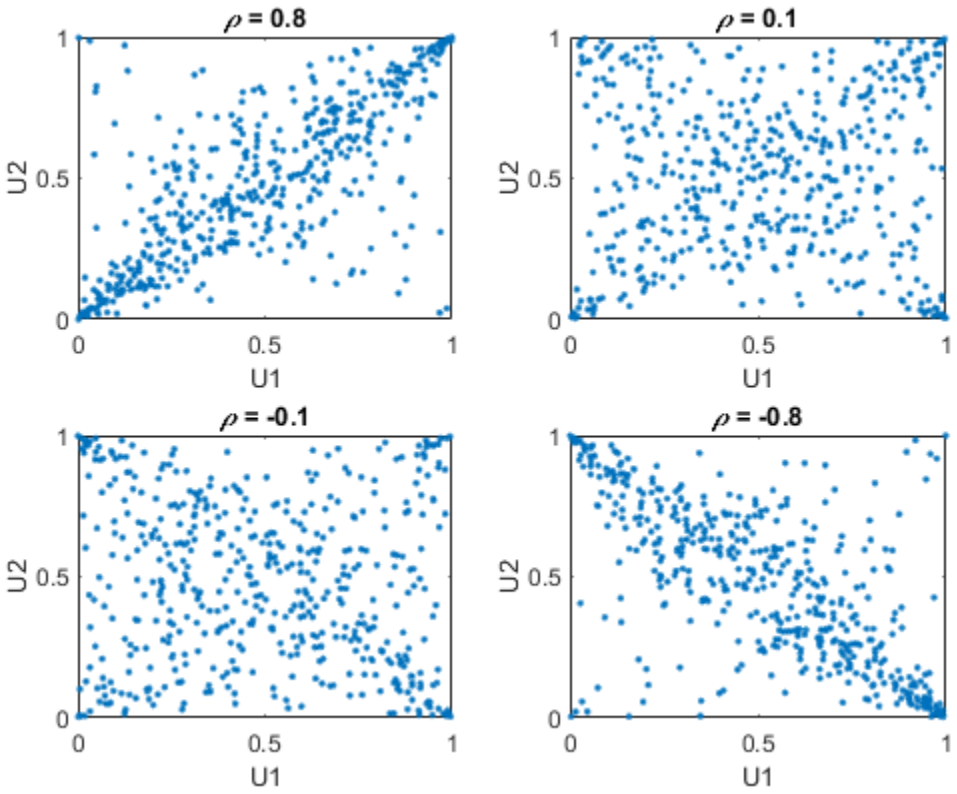
Just as for Gaussian copulas, Statistics and Machine Learning Toolbox functions for t copulas compute:

- Probability density functions (`copulapdf`) and the cumulative distribution functions (`copulacdf`) for t copulas
- Rank correlations from linear correlations (`copulastat`) and vice versa (`copulaparam`)
- Random vectors (`copularnd`)
- Parameters for copulas fit to data (`copulafit`)

For example, use the `copularnd` function to create scatter plots of random values from a bivariate t_1 copula for various levels of ρ, to illustrate the range of different dependence structures:

```
n = 500;
nu = 1;

rng('default') % for reproducibility
U = copularnd('t',[1 .8; .8 1],nu,n);
subplot(2,2,1)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = 0.8')
```

```
xlabel('U1')
ylabel('U2')

U = copularnd('t',[1 .1; .1 1],nu,n);
subplot(2,2,2)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = 0.1')
xlabel('U1')
ylabel('U2')

U = copularnd('t',[1 -.1; -.1 1],nu,n);
subplot(2,2,3)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = -0.1')
xlabel('U1')
ylabel('U2')

U = copularnd('t',[1 -.8; -.8 1],nu, n);
subplot(2,2,4)
plot(U(:,1),U(:,2),'.')
title('{\it\rho} = -0.8')
xlabel('U1')
ylabel('U2')
```

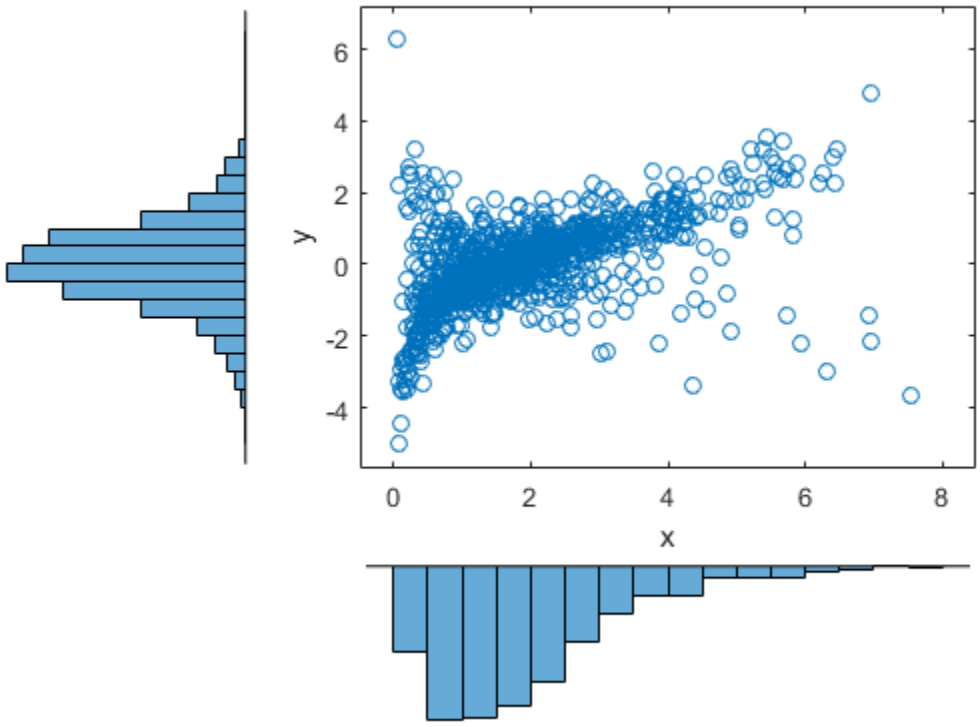


A t copula has uniform marginal distributions for `U1` and `U2`, just as a Gaussian copula does. The rank correlation τ or ρ_s between components in a t copula is also the same function of ρ as for a Gaussian. However, as these plots demonstrate, a t_1 copula differs quite a bit from a Gaussian copula, even

when their components have the same rank correlation. The difference is in their dependence structure. Not surprisingly, as the degrees of freedom parameter ν is made larger, a t_ν copula approaches the corresponding Gaussian copula.

As with a Gaussian copula, any marginal distributions can be imposed over a t copula. For example, using a t copula with 1 degree of freedom, you can again generate random vectors from a bivariate distribution with Gamma(2,1) and t_5 marginals using `copularnd`:

```
n = 1000;
rho = .7;
nu = 1;

rng('default') % for reproducibility
U = copularnd('t',[1 rho; rho 1],nu,n);
X = [gaminv(U(:,1),2,1) tinv(U(:,2),5)];

figure
scatterhist(X(:,1),X(:,2),'Direction','out')
```



Compared to the bivariate Gamma/t distribution constructed earlier, which was based on a Gaussian copula, the distribution constructed here, based on a t_1 copula, has the same marginal distributions and the same rank correlation between variables but a very different dependence structure. This illustrates the fact that multivariate distributions are not uniquely defined by their marginal distributions, or by their correlations. The choice of a particular copula in an application may be based on actual observed data, or different copulas may be used as a way of determining the sensitivity of simulation results to the input distribution.

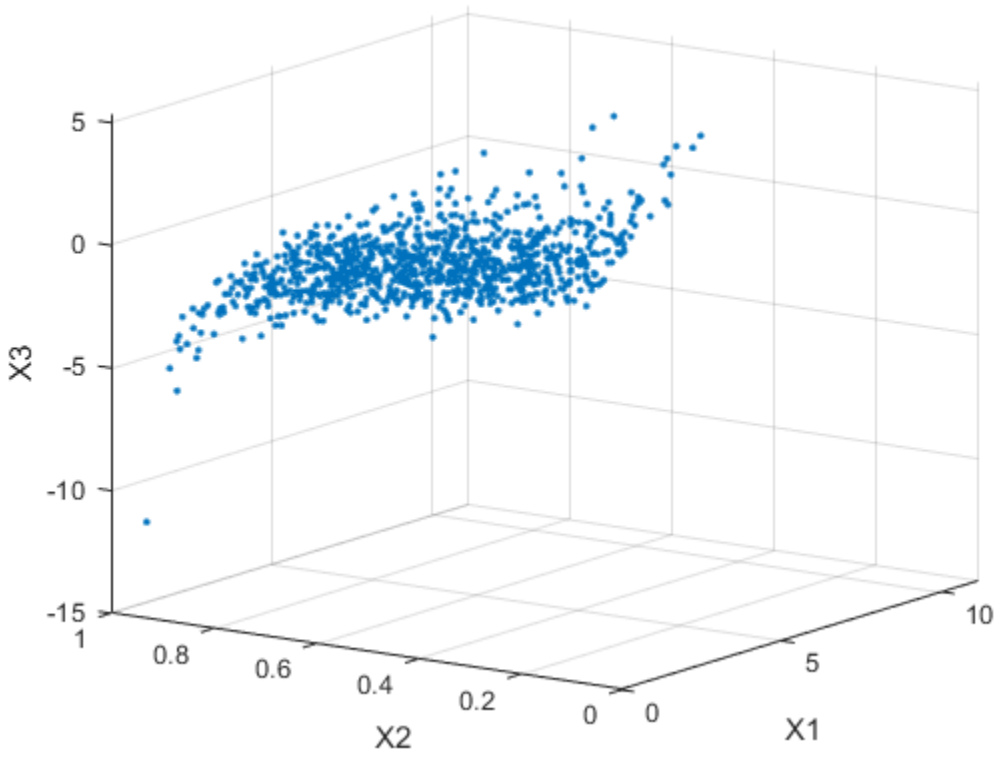
Higher Dimension Copulas

The Gaussian and t copulas are known as elliptical copulas. It is easy to generalize elliptical copulas to a higher number of dimensions. For example, simulate data from a trivariate distribution with Gamma(2,1), Beta(2,2), and t_5 marginals using a Gaussian copula and `copularnd`, as follows:

```
n = 1000;
Rho = [1 .4 .2; .4 1 -.8; .2 -.8 1];
rng('default') % for reproducibility
U = copularnd('Gaussian',Rho,n);
X = [gaminv(U(:,1),2,1) betainv(U(:,2),2,2) tinv(U(:,3),5)];
```

Plot the data.

```
subplot(1,1,1)
plot3(X(:,1),X(:,2),X(:,3),'.')
grid on
view([-55, 15])
xlabel('X1')
ylabel('X2')
zlabel('X3')
```



Notice that the relationship between the linear correlation parameter ρ and, for example, Kendall's τ, holds for each entry in the correlation matrix P used here. You can verify that the sample rank correlations of the data are approximately equal to the theoretical values:

```
tauTheoretical = 2.*asin(Rho)./pi
```

```
tauTheoretical = 3×3

    1.0000    0.2620    0.1282
    0.2620    1.0000   -0.5903
    0.1282   -0.5903    1.0000
```

```
tauSample = corr(X,'type','Kendall')
```

```
tauSample = 3×3

    1.0000    0.2581    0.1414
    0.2581    1.0000   -0.5790
    0.1414   -0.5790    1.0000
```

Archimedean Copulas

Statistics and Machine Learning Toolbox functions are available for three bivariate Archimedean copula families:

- Clayton copulas
- Frank copulas
- Gumbel copulas

These are one-parameter families that are defined directly in terms of their cdfs, rather than being defined constructively using a standard multivariate distribution.

To compare these three Archimedean copulas to the Gaussian and *t* bivariate copulas, first use the `copulastat` function to find the rank correlation for a Gaussian or *t* copula with linear correlation parameter of 0.8, and then use the `copulaparam` function to find the Clayton copula parameter that corresponds to that rank correlation:

```
tau = copulastat('Gaussian',.8 ,'type','kendall')
```

```
tau = 0.5903
```

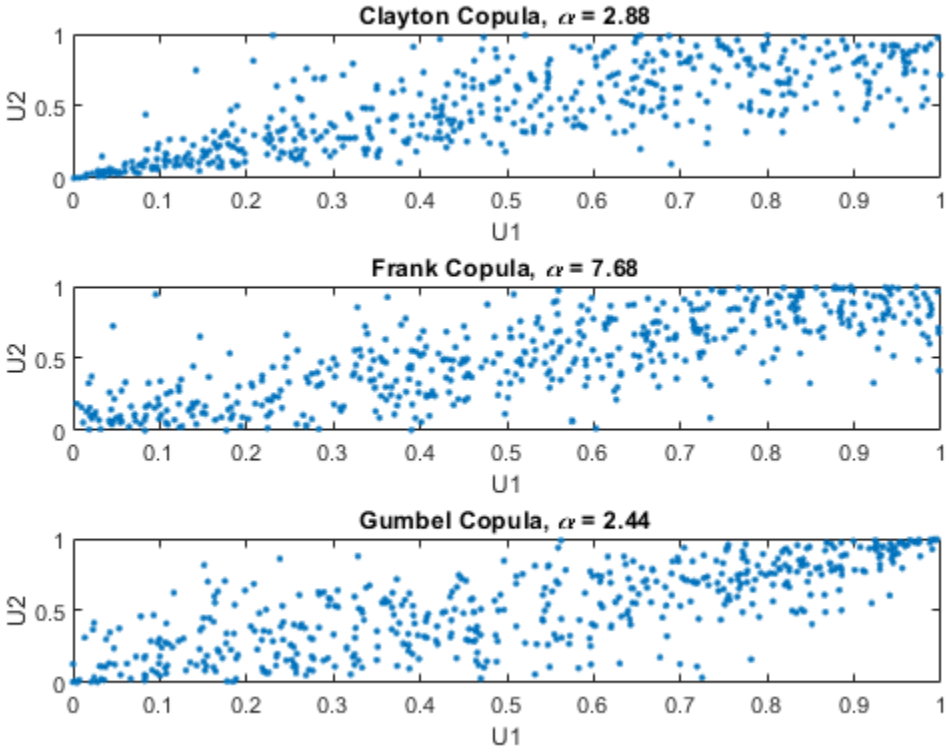
```
alpha = copulaparam('Clayton',tau,'type','kendall')
```

```
alpha = 2.8820
```

Finally, plot a random sample from the Clayton copula with `copularnd`. Repeat the same procedure for the Frank and Gumbel copulas:

```
n = 500;

U = copularnd('Clayton',alpha,n);
subplot(3,1,1)
plot(U(:,1),U(:,2),'.');
title(['Clayton Copula, {\it\alpha} = ',sprintf('%0.2f',alpha)])
xlabel('U1')
ylabel('U2')

alpha = copulaparam('Frank',tau,'type','kendall');
```

```
U = copularnd('Frank',alpha,n);
subplot(3,1,2)
plot(U(:,1),U(:,2),'.')
title(['Frank Copula, {\it\alpha} = ',sprintf('%0.2f',alpha)])
xlabel('U1')
ylabel('U2')

alpha = copulaparam('Gumbel',tau,'type','kendall');
U = copularnd('Gumbel',alpha,n);
subplot(3,1,3)
plot(U(:,1),U(:,2),'.')
title(['Gumbel Copula, {\it\alpha} = ',sprintf('%0.2f',alpha)])
xlabel('U1')
ylabel('U2')
```



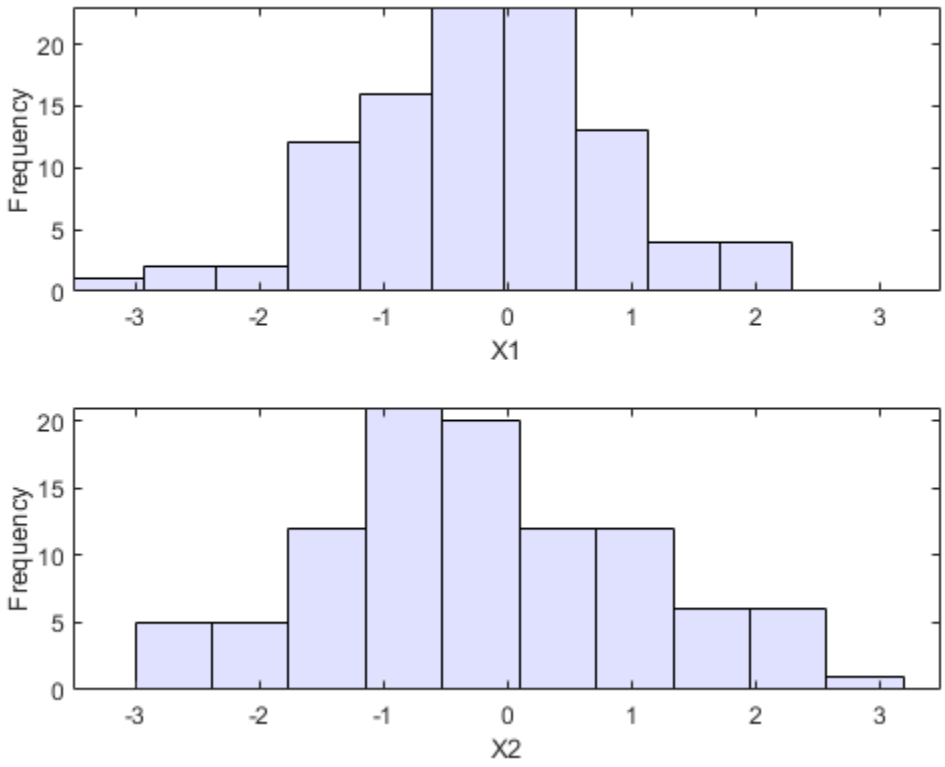
Simulating Dependent Multivariate Data Using Copulas

To simulate dependent multivariate data using a copula, you must specify each of the following:

- The copula family (and any shape parameters)
- The rank correlations among variables
- Marginal distributions for each variable

Suppose you have return data for two stocks and want to run a Monte Carlo simulation with inputs that follow the same distributions as the data:

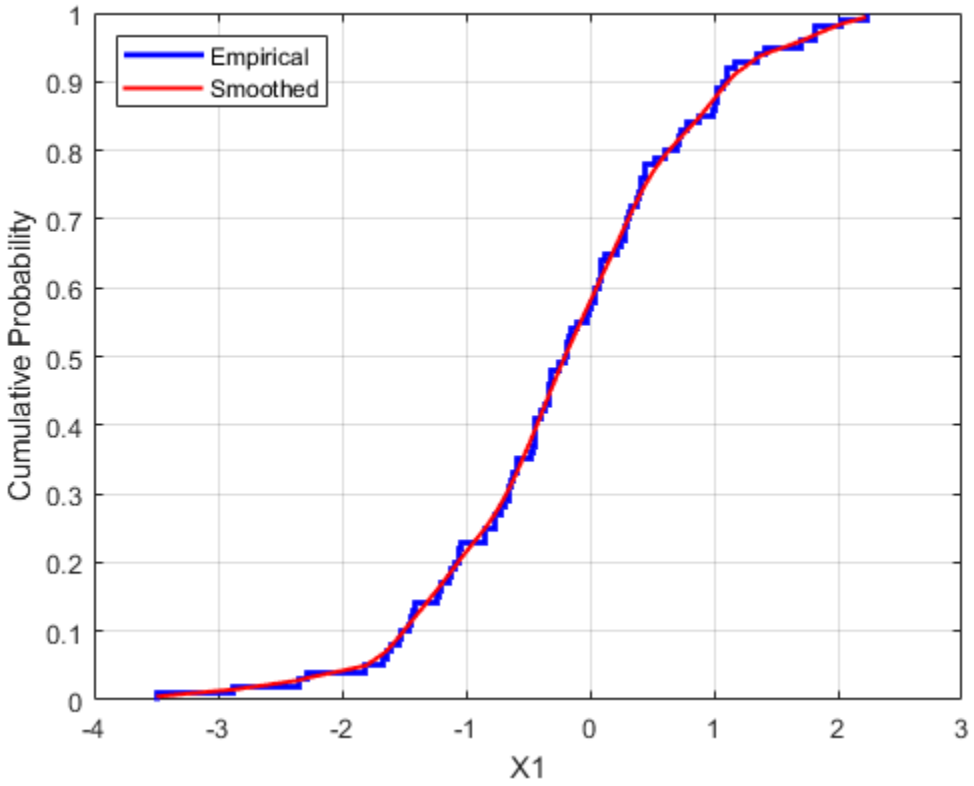
```
load stockreturns
nobs = size(stocks,1);

subplot(2,1,1)
histogram(stocks(:,1),10,'FaceColor',[.8 .8 1])
xlim([-3.5 3.5])
xlabel('X1')
ylabel('Frequency')

subplot(2,1,2)
histogram(stocks(:,2),10,'FaceColor',[.8 .8 1])
xlim([-3.5 3.5])
xlabel('X2')
ylabel('Frequency')
```



You could fit a parametric model separately to each dataset, and use those estimates as the marginal distributions. However, a parametric model may not be sufficiently flexible. Instead, you can use a nonparametric model to transform to the marginal distributions. All that is needed is a way to compute the inverse cdf for the nonparametric model.

The simplest nonparametric model is the empirical cdf, as computed by the `ecdf` function. For a discrete marginal distribution, this is appropriate. However, for a continuous distribution, use a model that is smoother than the step function computed by `ecdf`. One way to do that is to estimate the empirical cdf and interpolate between the midpoints of the steps with a piecewise linear function. Another way is to use kernel smoothing with `ksdensity`. For example, compare the empirical cdf to a kernel smoothed cdf estimate for the first variable:

```
[Fi,xi] = ecdf(stocks(:,1));

figure()
stairs(xi,Fi,'b','LineWidth',2)
hold on

Fi_sm = ksdensity(stocks(:,1),xi,'function','cdf','width',.15);

plot(xi,Fi_sm,'r-','LineWidth',1.5)
xlabel('X1')
ylabel('Cumulative Probability')
legend('Empirical','Smoothed','Location','NW')
grid on
```



For the simulation, experiment with different copulas and correlations. Here, you will use a bivariate *t* copula with a fairly small degrees of freedom parameter. For the correlation parameter, you can compute the rank correlation of the data.

```
nu = 5;
tau = corr(stocks(:,1),stocks(:,2),'type','kendall')
```

```
tau = 0.5180
```

Find the corresponding linear correlation parameter for the *t* copula using `copulaparam`.

```
rho = copulaparam('t', tau, nu, 'type','kendall')
```

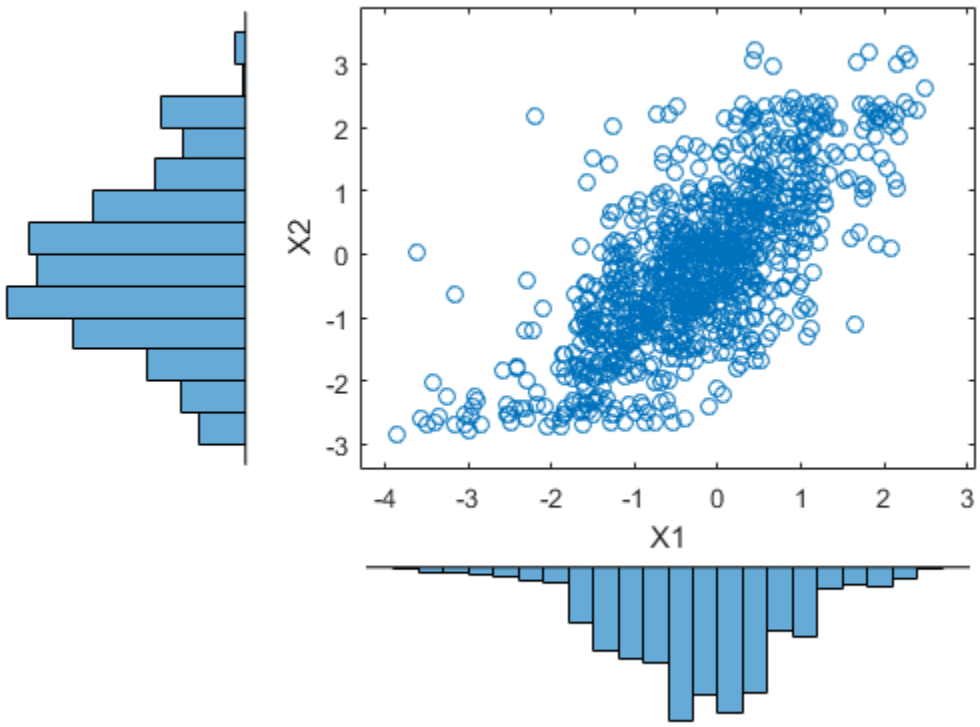
```
rho = 0.7268
```

Next, use `copularnd` to generate random values from the *t* copula and transform using the nonparametric inverse cdfs. The `ksdensity` function allows you to make a kernel estimate of distribution and evaluate the inverse cdf at the copula points all in one step:

```
n = 1000;
U = copularnd('t',[1 rho; rho 1],nu,n);
X1 = ksdensity(stocks(:,1),U(:,1),...
               'function','icdf','width',.15);
X2 = ksdensity(stocks(:,2),U(:,2),...
               'function','icdf','width',.15);
```

Alternatively, when you have a large amount of data or need to simulate more than one set of values, it may be more efficient to compute the inverse cdf over a grid of values in the interval `(0,1)` and use interpolation to evaluate it at the copula points:

```
p = linspace(0.00001,0.99999,1000);
G1 = ksdensity(stocks(:,1),p,'function','icdf','width',0.15);
X1 = interp1(p,G1,U(:,1),'spline');
G2 = ksdensity(stocks(:,2),p,'function','icdf','width',0.15);
X2 = interp1(p,G2,U(:,2),'spline');

scatterhist(X1,X2,'Direction','out')
```



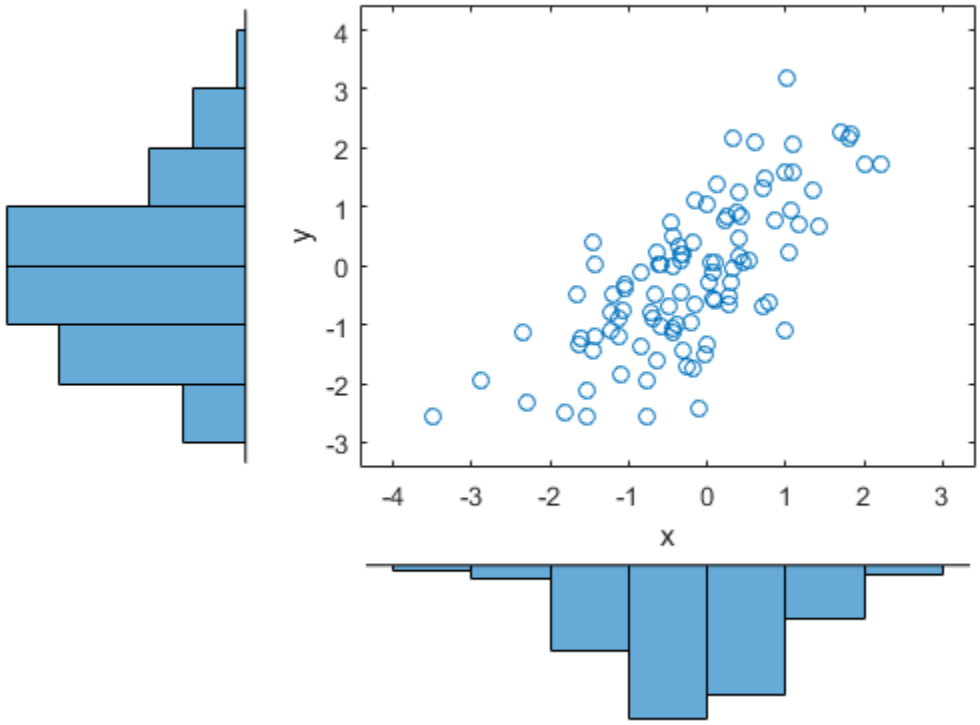
The marginal histograms of the simulated data are a smoothed version of the histograms for the original data. The amount of smoothing is controlled by the bandwidth input to `ksdensity`.

Fitting Copulas to Data

This example shows how to use `copulafit` to calibrate copulas with data. To generate data `Xsim` with a distribution "just like" (in terms of marginal distributions and correlations) the distribution of data in the matrix `X`, you need to fit marginal distributions to the columns of `X`, use appropriate cdf functions to transform `X` to `U`, so that `U` has values between 0 and 1, use `copulafit` to fit a copula to `U`, generate new data `Usim` from the copula, and use appropriate inverse cdf functions to transform `Usim` to `Xsim` .

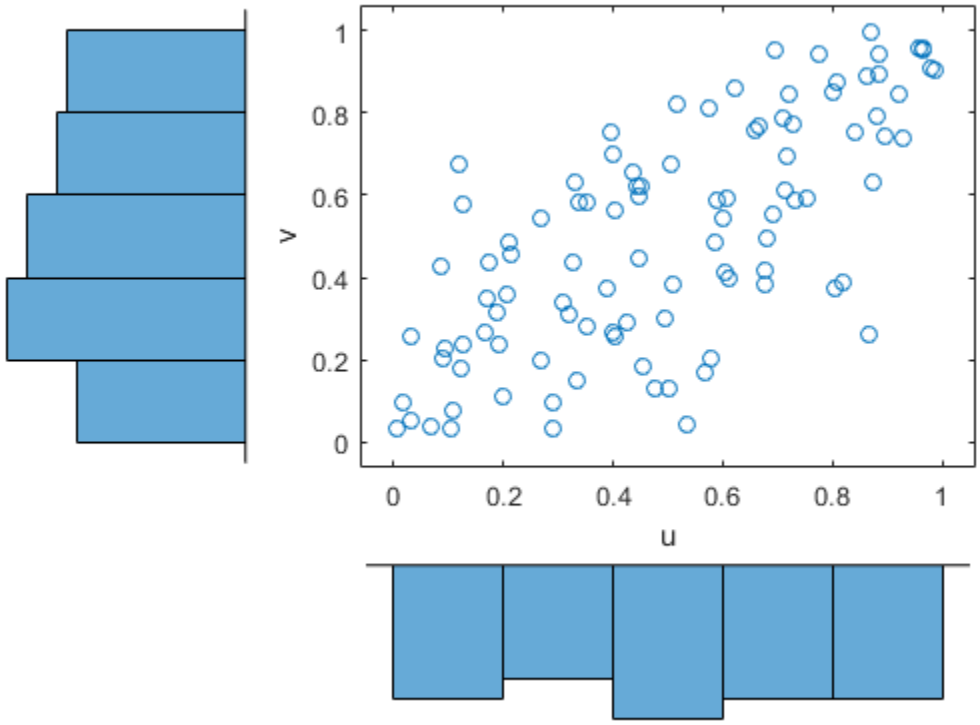
Load and plot the simulated stock return data.

```
load stockreturns
x = stocks(:,1);
y = stocks(:,2);

scatterhist(x,y,'Direction','out')
```



Transform the data to the copula scale (unit square) using a kernel estimator of the cumulative distribution function.

```
u = ksdensity(x,x,'function','cdf');
v = ksdensity(y,y,'function','cdf');

scatterhist(u,v,'Direction','out')
xlabel('u')
ylabel('v')
```



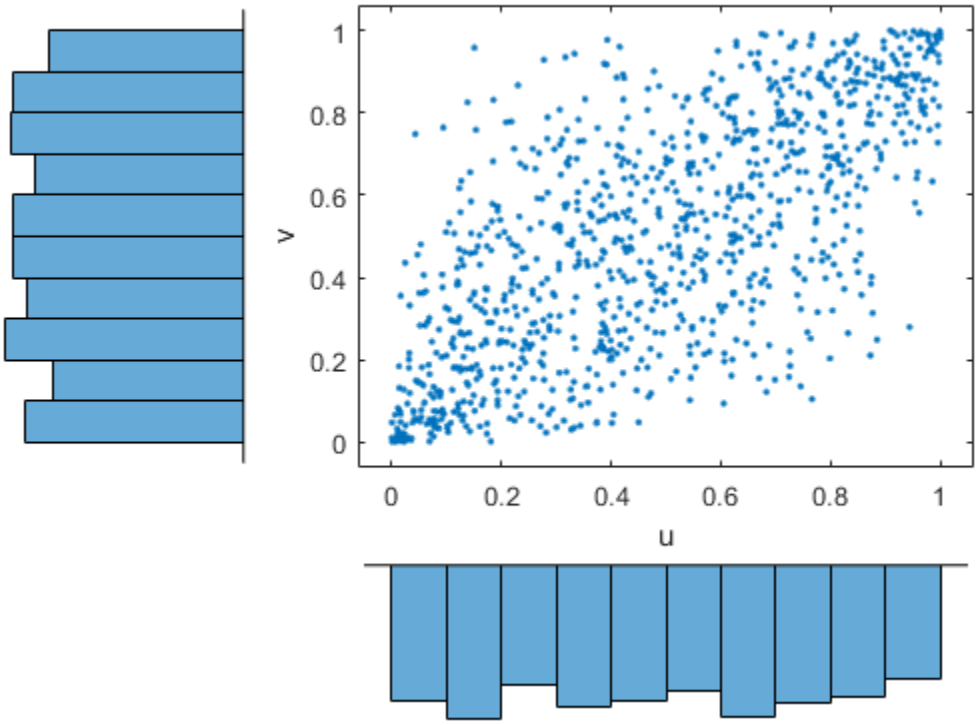
Fit a *t* copula.

```
[Rho,nu] = copulafit('t',[u v],'Method','ApproximateML')
```

```
Rho = 2×2

    1.0000    0.7220
    0.7220    1.0000


nu = 3.4516e+06
```

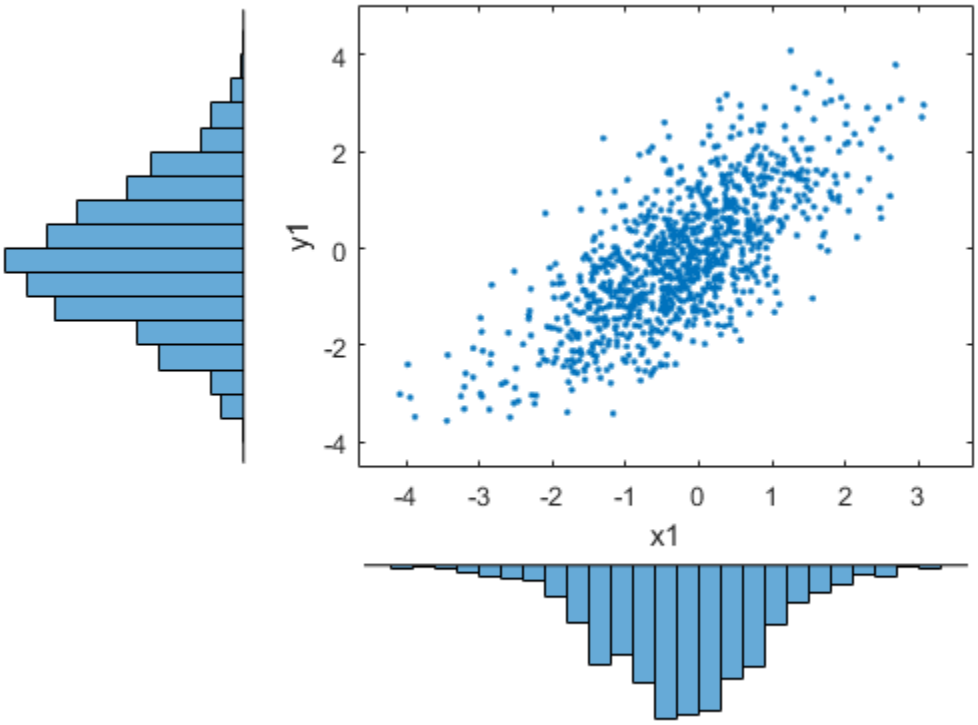
Generate a random sample from the *t* copula.

```
r = copularnd('t',Rho,nu,1000);
u1 = r(:,1);
v1 = r(:,2);

scatterhist(u1,v1,'Direction','out')
xlabel('u')
ylabel('v')
set(get(gca,'children'),'marker','.')
```



Transform the random sample back to the original scale of the data.

```
x1 = ksdensity(x,u1,'function','icdf');
y1 = ksdensity(y,v1,'function','icdf');

scatterhist(x1,y1,'Direction','out')
set(get(gca,'children'),'marker','.')
```



As the example illustrates, copulas integrate naturally with other distribution fitting functions.

See Also

`copulacdf` | `copulafit` | `copulaparam` | `copulapdf` | `copularnd` | `copulastat`

Related Examples

- “Generate Correlated Data Using Rank Correlation” on page 5-108
- “Simulating Dependent Random Variables Using Copulas” on page 5-147

Simulating Dependent Random Variables Using Copulas

This example shows how to use copulas to generate data from multivariate distributions when there are complicated relationships among the variables, or when the individual variables are from different distributions.

MATLAB® is an ideal tool for running simulations that incorporate random inputs or noise. Statistics and Machine Learning Toolbox™ provides functions to create sequences of random data according to many common univariate distributions. The Toolbox also includes a few functions to generate random data from multivariate distributions, such as the multivariate normal and multivariate t. However, there is no built-in way to generate multivariate distributions for all marginal distributions, or in cases where the individual variables are from different distributions.

Recently, copulas have become popular in simulation models. Copulas are functions that describe dependencies among variables, and provide a way to create distributions to model correlated multivariate data. Using a copula, a data analyst can construct a multivariate distribution by specifying marginal univariate distributions, and choosing a particular copula to provide a correlation structure between variables. Bivariate distributions, as well as distributions in higher dimensions, are possible. In this example, we discuss how to use copulas to generate dependent multivariate random data in MATLAB, using Statistics and Machine Learning Toolbox.

Dependence Between Simulation Inputs

One of the design decisions for a Monte-Carlo simulation is a choice of probability distributions for the random inputs. Selecting a distribution for each individual variable is often straightforward, but deciding what dependencies should exist between the inputs may not be. Ideally, input data to a simulation should reflect what is known about dependence among the real quantities being modelled. However, there may be little or no information on which to base any dependence in the simulation, and in such cases, it is a good idea to experiment with different possibilities, in order to determine the model's sensitivity.

However, it can be difficult to actually generate random inputs with dependence when they have distributions that are not from a standard multivariate distribution. Further, some of the standard multivariate distributions can model only very limited types of dependence. It's always possible to make the inputs independent, and while that is a simple choice, it's not always sensible and can lead to the wrong conclusions.

For example, a Monte-Carlo simulation of financial risk might have random inputs that represent different sources of insurance losses. These inputs might be modeled as lognormal random variables. A reasonable question to ask is how dependence between these two inputs affects the results of the simulation. Indeed, it might be known from real data that the same random conditions affect both sources, and ignoring that in the simulation could lead to the wrong conclusions.

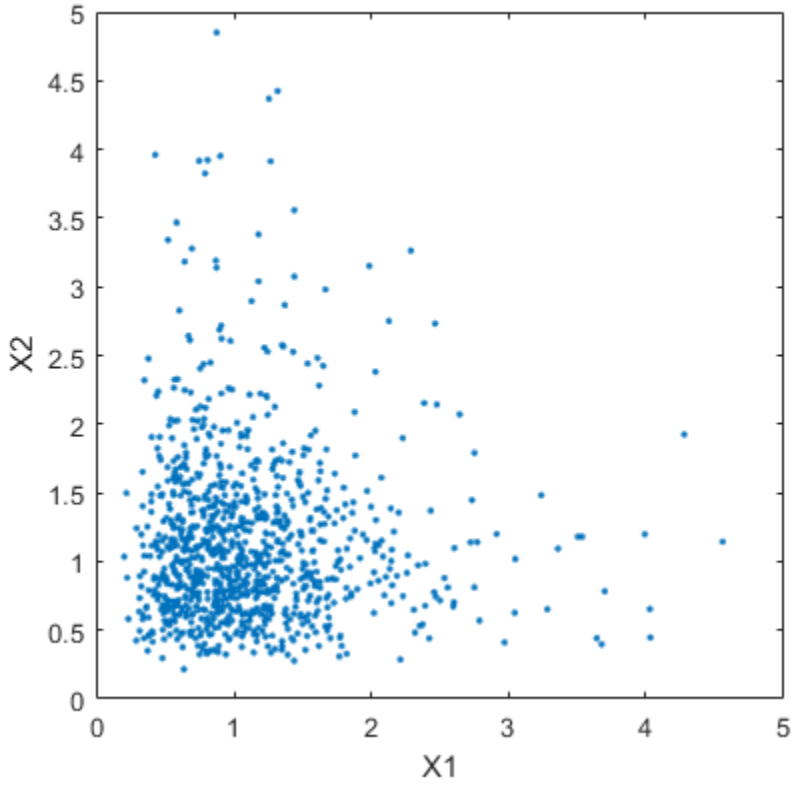
Simulation of independent lognormal random variables is trivial. The simplest way would be to use the `lognrnd` function. Here, we'll use the `mvnrnd` function to generate n pairs of independent normal random variables, and then exponentiate them. Notice that the covariance matrix used here is diagonal, i.e., independence between the columns of Z.

```
n = 1000;
sigma = .5;
SigmaInd = sigma.^2 .* [1 0; 0 1]
```

```
SigmaInd =
```

```
    0.2500         0
         0    0.2500
```

```
ZInd = mvnrnd([0 0], SigmaInd, n);
XInd = exp(ZInd);
plot(XInd(:,1),XInd(:,2),'.');
axis equal;
axis([0 5 0 5]);
xlabel('X1');
ylabel('X2');
```



Dependent bivariate lognormal r.v.'s are also easy to generate, using a covariance matrix with non-zero off-diagonal terms.

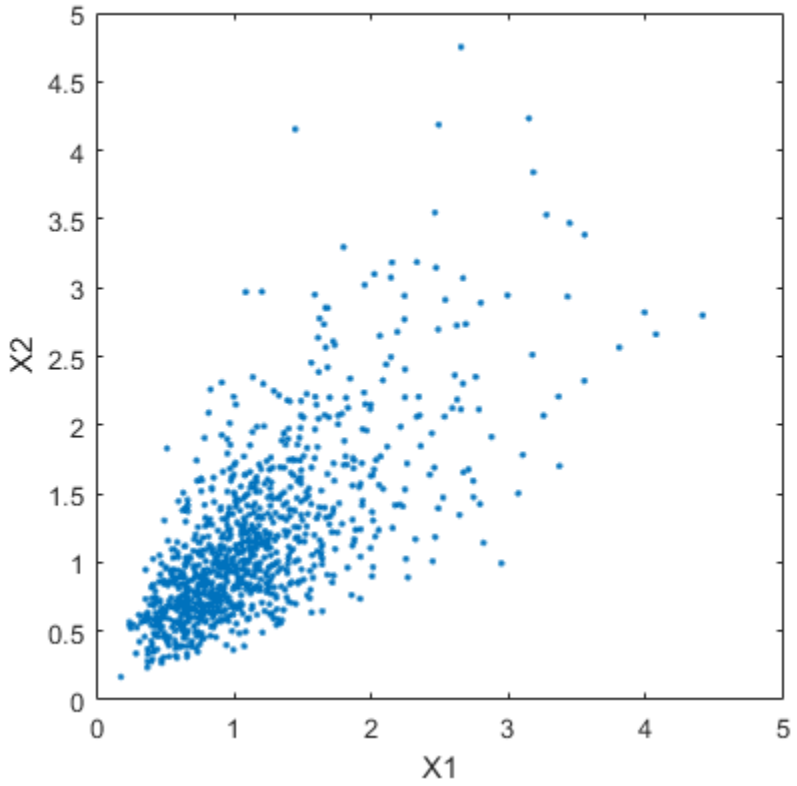
```
rho = .7;
SigmaDep = sigma.^2 .* [1 rho; rho 1]
```

```
SigmaDep =

    0.2500    0.1750
    0.1750    0.2500
```

```
ZDep = mvnrnd([0 0], SigmaDep, n);
XDep = exp(ZDep);
```

A second scatter plot illustrates the difference between these two bivariate distributions.

```
plot(XDep(:,1),XDep(:,2),'.');
axis equal;
axis([0 5 0 5]);
xlabel('X1');
ylabel('X2');
```



It's clear that there is more of a tendency in the second dataset for large values of X1 to be associated with large values of X2, and similarly for small values. This dependence is determined by the correlation parameter, rho, of the underlying bivariate normal. The conclusions drawn from the simulation could well depend on whether or not X1 and X2 were generated with dependence or not.

The bivariate lognormal distribution is a simple solution in this case, and of course easily generalizes to higher dimensions and cases where the marginal distributions are *different* lognormals. Other multivariate distributions also exist, for example, the multivariate t and the Dirichlet distributions are used to simulate dependent t and beta random variables, respectively. But the list of simple multivariate distributions is not long, and they only apply in cases where the marginals are all in the same family (or even the exact same distributions). This can be a real limitation in many situations.

A More General Method for Constructing Dependent Bivariate Distributions

Although the above construction that creates a bivariate lognormal is simple, it serves to illustrate a method which is more generally applicable. First, we generate pairs of values from a bivariate normal distribution. There is statistical dependence between these two variables, and each has a normal marginal distribution. Next, a transformation (the exponential function) is applied separately to each

variable, changing the marginal distributions into lognormals. The transformed variables still have a statistical dependence.

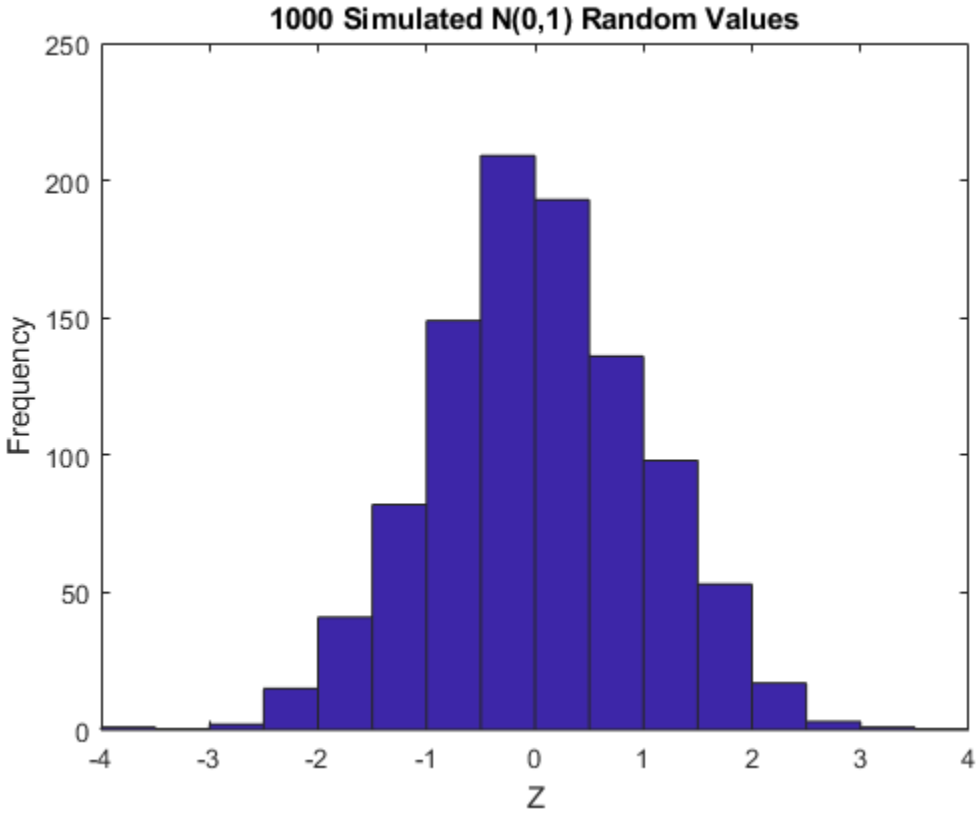
If a suitable transformation could be found, this method could be generalized to create dependent bivariate random vectors with other marginal distributions. In fact, a general method of constructing such a transformation does exist, although not as simple as just exponentiation.

By definition, applying the normal CDF (denoted here by PHI) to a standard normal random variable results in a r.v. that is uniform on the interval [0, 1]. To see this, if Z has a standard normal distribution, then the CDF of U = PHI(Z) is

```
  Pr{U <= u0} = Pr{PHI(Z) <= u0} = Pr{Z <= PHI^(-1)(u0)} = u0,
```

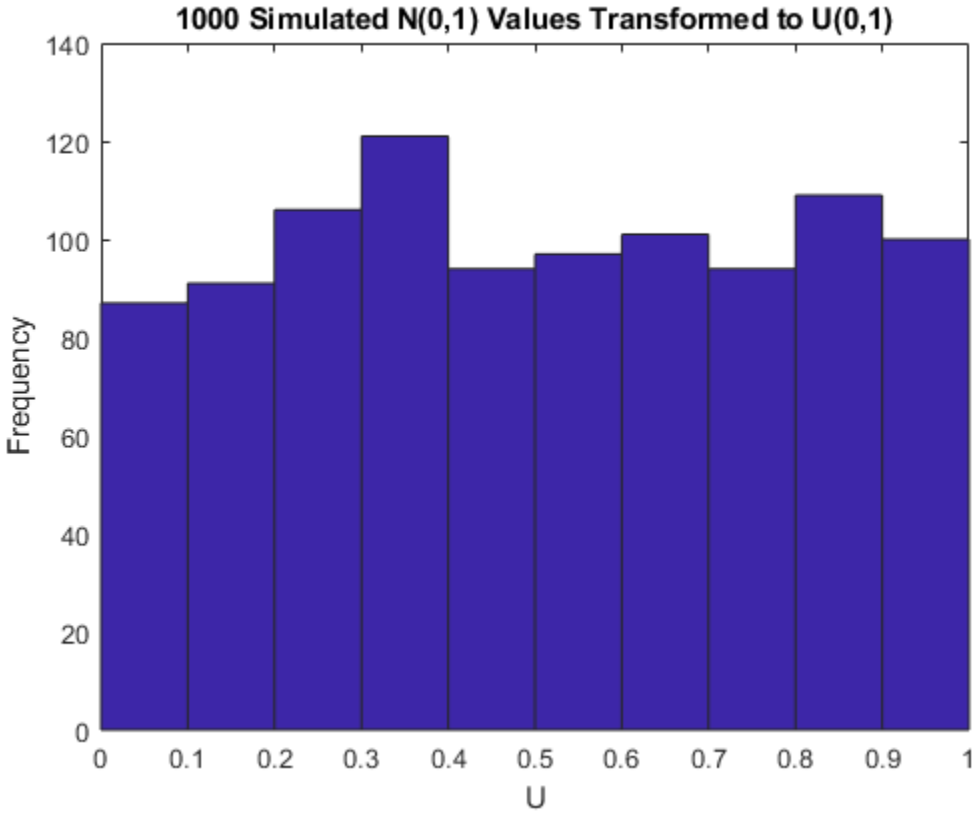
and that is the CDF of a U(0,1) r.v. Histograms of some simulated normal and transformed values demonstrate that fact.

```
n = 1000;
z = normrnd(0,1,n,1);
hist(z,-3.75:.5:3.75);
xlim([-4 4]);
title('1000 Simulated N(0,1) Random Values');
xlabel('Z');
ylabel('Frequency');
```



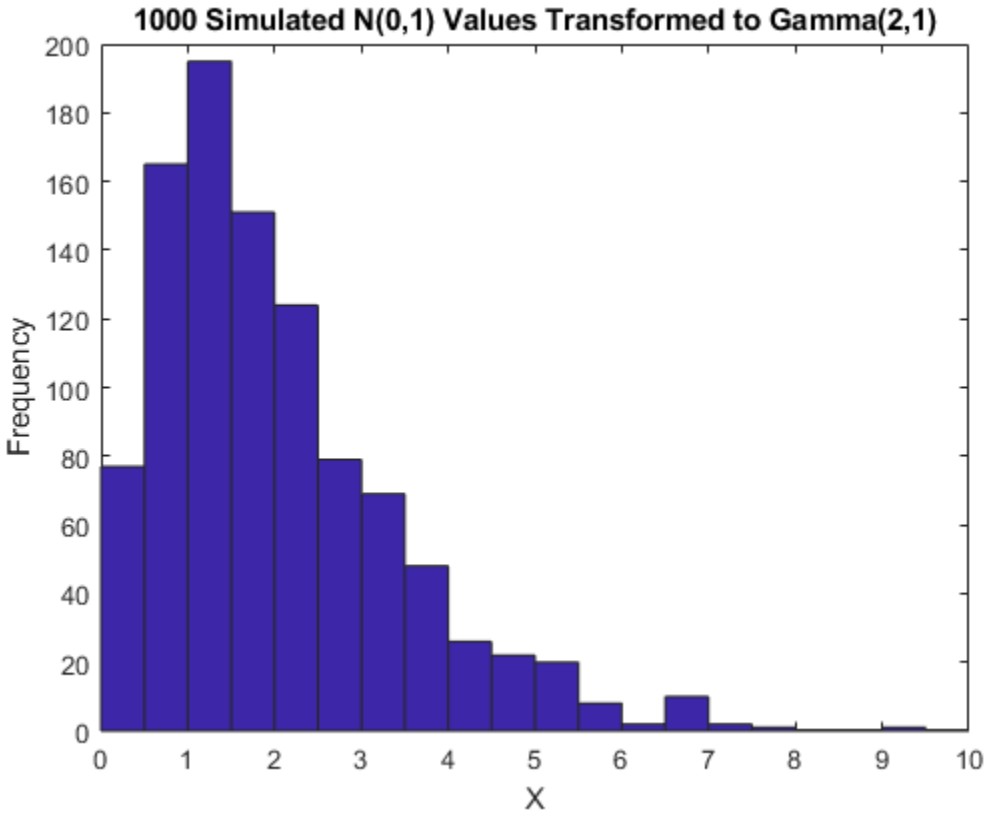

```
u = normcdf(z);
hist(u,.05:.1:.95);
title('1000 Simulated N(0,1) Values Transformed to U(0,1)');
```

```
xlabel('U');
ylabel('Frequency');
```



Now, borrowing from the theory of univariate random number generation, applying the inverse CDF of any distribution F to a U(0,1) random variable results in a r.v. whose distribution is exactly F. This is known as the Inversion Method. The proof is essentially the opposite of the above proof for the forward case. Another histogram illustrates the transformation to a gamma distribution.

```
x = gaminv(u,2,1);
hist(x,.25:.5:9.75);
title('1000 Simulated N(0,1) Values Transformed to Gamma(2,1)');
xlabel('X');
ylabel('Frequency');
```



This two-step transformation can be applied to each variable of a standard bivariate normal, creating dependent r.v.'s with arbitrary marginal distributions. Because the transformation works on each component separately, the two resulting r.v.'s need not even have the same marginal distributions. The transformation is defined as

```
Z = [Z1 Z2] ~ N([0 0],[1 rho; rho 1])
U = [PHI(Z1) PHI(Z2)]
X = [G1(U1) G2(U2)]
```

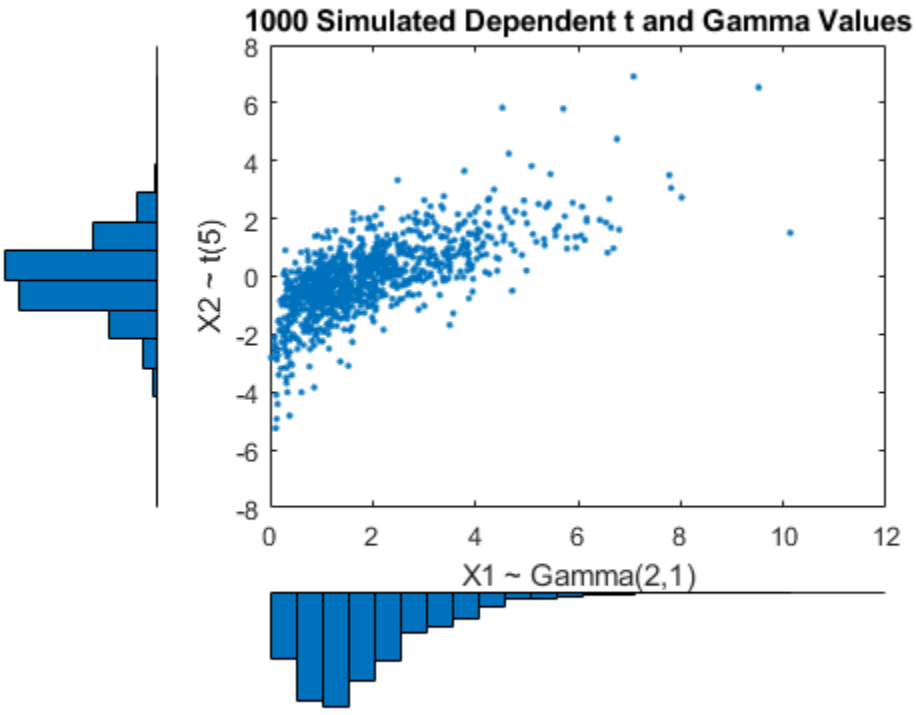
where G1 and G2 are inverse CDFs of two possibly different distributions. For example, we can generate random vectors from a bivariate distribution with t(5) and Gamma(2,1) marginals.

```
n = 1000;
rho = .7;
Z = mvnrnd([0 0], [1 rho; rho 1], n);
U = normcdf(Z);
X = [gaminv(U(:,1),2,1) tinv(U(:,2),5)];
```

This plot has histograms alongside a scatter plot to show both the marginal distributions, and the dependence.

```
[n1,ctr1] = hist(X(:,1),20);
[n2,ctr2] = hist(X(:,2),20);
subplot(2,2,2);
plot(X(:,1),X(:,2),'.');
axis([0 12 -8 8]);
h1 = gca;
```

```
title('1000 Simulated Dependent t and Gamma Values');
xlabel('X1 ~ Gamma(2,1)');
ylabel('X2 ~ t(5)');
subplot(2,2,4);
bar(ctr1,-n1,1);
axis([0 12 -max(n1)*1.1 0]);
axis('off');
h2 = gca;
subplot(2,2,1);
barh(ctr2,-n2,1);
axis([-max(n2)*1.1 0 -8 8]);
axis('off');
h3 = gca;
h1.Position = [0.35 0.35 0.55 0.55];
h2.Position = [.35 .1 .55 .15];
h3.Position = [.1 .35 .15 .55];
colormap([.8 .8 1]);
```



Rank Correlation Coefficients

Dependence between X1 and X2 in this construction is determined by the correlation parameter, rho, of the underlying bivariate normal. However, it is *not* true that the linear correlation of X1 and X2 is rho. For example, in the original lognormal case, there is a closed form for that correlation:

```
cor(X1,X2) = (exp(rho.*sigma.^2) - 1) ./ (exp(sigma.^2) - 1)
```

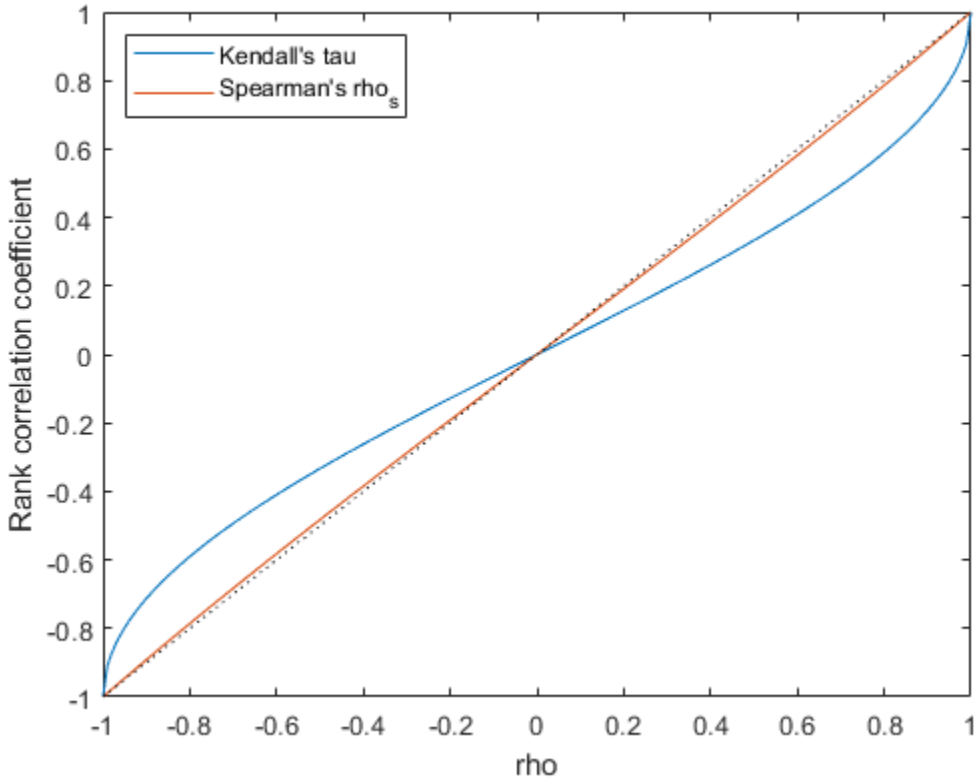
which is strictly less than rho unless rho is exactly one. In more general cases, though, such as the Gamma/t construction above, the linear correlation between X1 and X2 is difficult or impossible to express in terms of rho, but simulations can be used to show that the same effect happens.

That's because the linear correlation coefficient expresses the *linear* dependence between r.v.'s, and when nonlinear transformations are applied to those r.v.'s, linear correlation is not preserved. Instead, a rank correlation coefficient, such as Kendall's tau or Spearman's rho, is more appropriate.

Roughly speaking, these rank correlations measure the degree to which large or small values of one r.v. associate with large or small values of another. However, unlike the linear correlation coefficient, they measure the association only in terms of ranks. As a consequence, the rank correlation is preserved under any monotonic transformation. In particular, the transformation method just described preserves the rank correlation. Therefore, knowing the rank correlation of the bivariate normal Z exactly determines the rank correlation of the final transformed r.v.'s X. While rho is still needed to parameterize the underlying bivariate normal, Kendall's tau or Spearman's rho are more useful in describing the dependence between r.v.'s, because they are invariant to the choice of marginal distribution.

It turns out that for the bivariate normal, there is a simple 1-1 mapping between Kendall's tau or Spearman's rho, and the linear correlation coefficient rho:

```
  tau   = (2/pi)*arcsin(rho)     or   rho = sin(tau*pi/2)
  rho_s = (6/pi)*arcsin(rho/2)   or   rho = 2*sin(rho_s*pi/6)

subplot(1,1,1);
rho = -1:.01:1;
tau = 2.*asin(rho)./pi;
rho_s = 6.*asin(rho./2)./pi;
plot(rho,tau,'-', rho,rho_s,'-', [-1 1],[-1 1],'k:');
axis([-1 1 -1 1]);
xlabel('rho');
ylabel('Rank correlation coefficient');
legend('Kendall''s tau', 'Spearman''s rho_s', 'location','northwest');
```



Thus, it's easy to create the desired rank correlation between X1 and X2, regardless of their marginal distributions, by choosing the correct rho parameter value for the linear correlation between Z1 and Z2.

Notice that for the multivariate normal distribution, Spearman's rank correlation is almost identical to the linear correlation. However, this is not true once we transform to the final random variables.

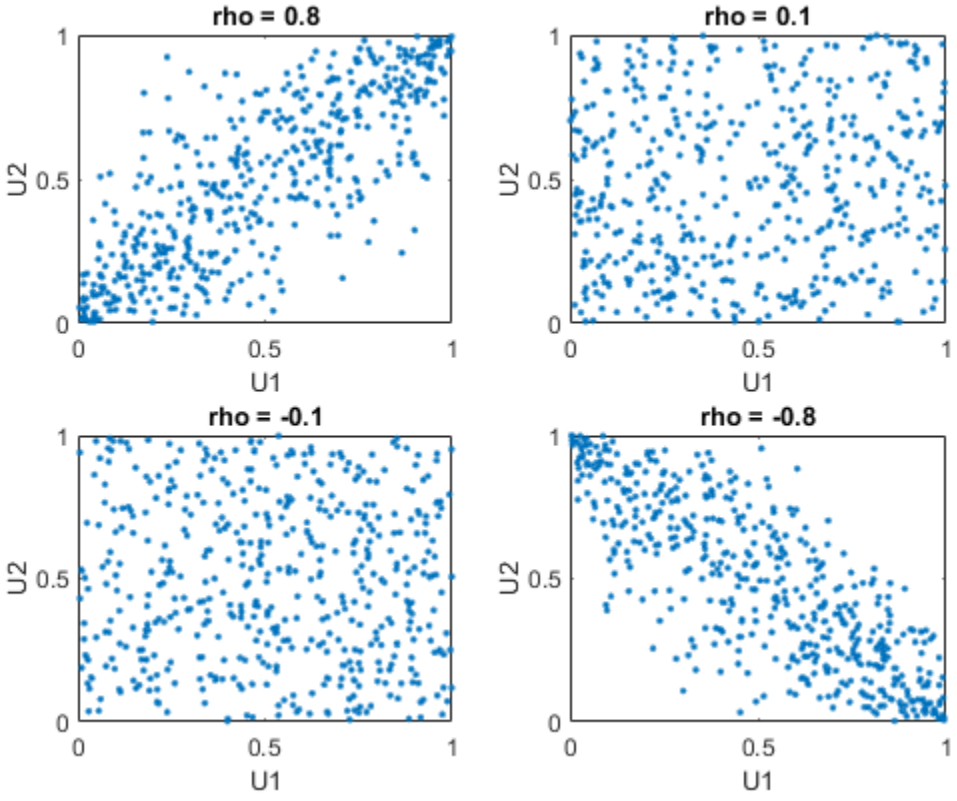
Copulas

The first step of the construction described above defines what is known as a copula, specifically, a Gaussian copula. A bivariate copula is simply a probability distribution on two random variables, each of whose marginal distributions is uniform. These two variables may be completely independent, deterministically related (e.g., U2 = U1), or anything in between. The family of bivariate Gaussian copulas is parameterized by Rho = [1 rho; rho 1], the linear correlation matrix. U1 and U2 approach linear dependence as rho approaches +/- 1, and approach complete independence as rho approaches zero.

Scatter plots of some simulated random values for various levels of rho illustrate the range of different possibilities for Gaussian copulas:

```
n = 500;
Z = mvnrnd([0 0], [1 .8; .8 1], n);
U = normcdf(Z,0,1);
subplot(2,2,1);
plot(U(:,1),U(:,2),'.');
title('rho = 0.8');
```

```
xlabel('U1');
ylabel('U2');
Z = mvnrnd([0 0], [1 .1; .1 1], n);
U = normcdf(Z,0,1);
subplot(2,2,2);
plot(U(:,1),U(:,2),'.');
title('rho = 0.1');
xlabel('U1');
ylabel('U2');
Z = mvnrnd([0 0], [1 -.1; -.1 1], n);
U = normcdf(Z,0,1);
subplot(2,2,3);
plot(U(:,1),U(:,2),'.');
title('rho = -0.1');
xlabel('U1');
ylabel('U2');
Z = mvnrnd([0 0], [1 -.8; -.8 1], n);
U = normcdf(Z,0,1);
subplot(2,2,4);
plot(U(:,1),U(:,2),'.');
title('rho = -0.8');
xlabel('U1');
ylabel('U2');
```



The dependence between U1 and U2 is completely separate from the marginal distributions of X1 = G(U1) and X2 = G(U2). X1 and X2 can be given *any* marginal distributions, and still have the same

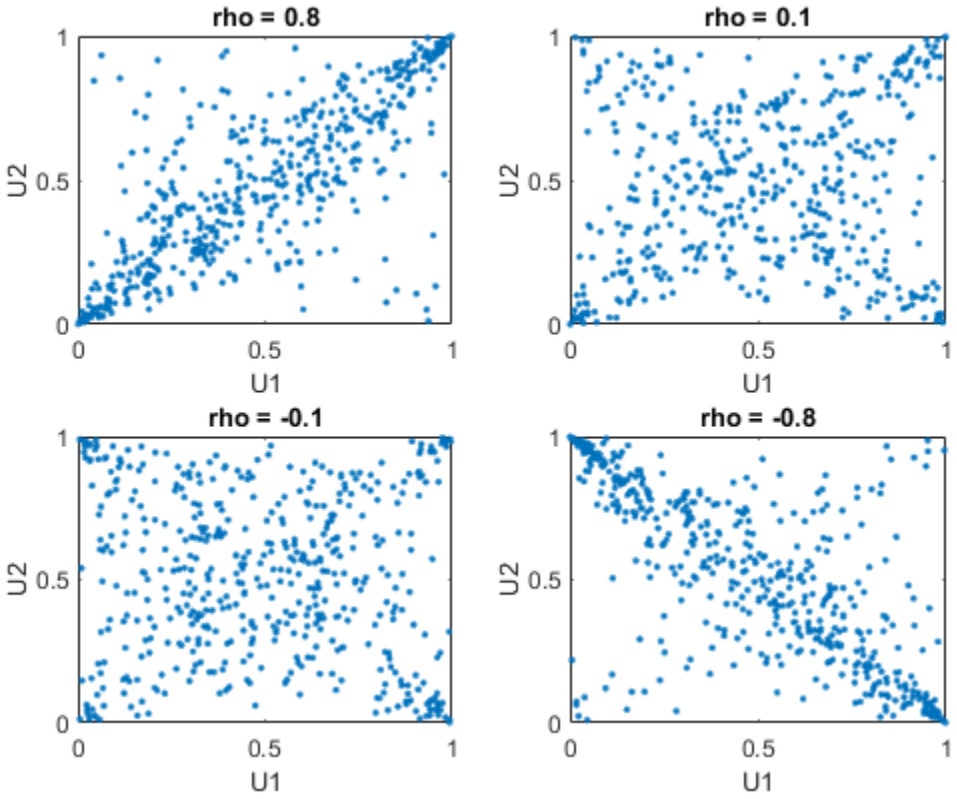
rank correlation. This is one of the main appeals of copulas -- they allow this separate specification of dependence and marginal distribution.

t Copulas

A different family of copulas can be constructed by starting from a bivariate t distribution, and transforming using the corresponding t CDF. The bivariate t distribution is parameterized with Rho, the linear correlation matrix, and nu, the degrees of freedom. Thus, for example, we can speak of a t(1) or a t(5) copula, based on the multivariate t with one and five degrees of freedom, respectively.

Scatter plots of some simulated random values for various levels of rho illustrate the range of different possibilities for t(1) copulas:

```
n = 500;
nu = 1;
T = mvtrnd([1 .8; .8 1], nu, n);
U = tcdf(T,nu);
subplot(2,2,1);
plot(U(:,1),U(:,2),'.');
title('rho = 0.8');
xlabel('U1');
ylabel('U2');
T = mvtrnd([1 .1; .1 1], nu, n);
U = tcdf(T,nu);
subplot(2,2,2);
plot(U(:,1),U(:,2),'.');
title('rho = 0.1');
xlabel('U1');
ylabel('U2');
T = mvtrnd([1 -.1; -.1 1], nu, n);
U = tcdf(T,nu);
subplot(2,2,3);
plot(U(:,1),U(:,2),'.');
title('rho = -0.1');
xlabel('U1');
ylabel('U2');
T = mvtrnd([1 -.8; -.8 1], nu, n);
U = tcdf(T,nu);
subplot(2,2,4);
plot(U(:,1),U(:,2),'.');
title('rho = -0.8');
xlabel('U1');
ylabel('U2');
```

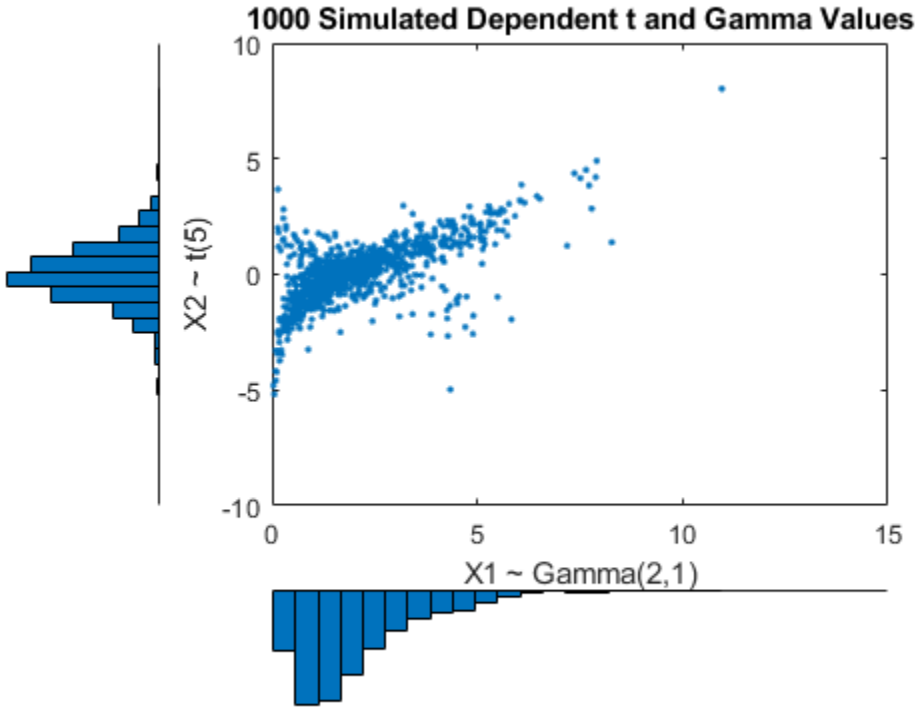


A t copula has uniform marginal distributions for U1 and U2, just as a Gaussian copula does. The rank correlation tau or rho_s between components in a t copula is also the same function of rho as for a Gaussian. However, as these plots demonstrate, a t(1) copula differs quite a bit from a Gaussian copula, even when their components have the same rank correlation. The difference is in their dependence structure. Not surprisingly, as the degrees of freedom parameter nu is made larger, a t(nu) copula approaches the corresponding Gaussian copula.

As with a Gaussian copula, any marginal distributions can be imposed over a t copula. For example, using a t copula with 1 degree of freedom, we can again generate random vectors from a bivariate distribution with Gam(2,1) and t(5) marginals:

```
subplot(1,1,1);
n = 1000;
rho = .7;
nu = 1;
T = mvtrnd([1 rho; rho 1], nu, n);
U = tcdf(T,nu);
X = [gaminv(U(:,1),2,1) tinv(U(:,2),5)];

[n1,ctr1] = hist(X(:,1),20);
[n2,ctr2] = hist(X(:,2),20);
subplot(2,2,2);
plot(X(:,1),X(:,2),'.');
axis([0 15 -10 10]);
h1 = gca;
title('1000 Simulated Dependent t and Gamma Values');
xlabel('X1 ~ Gamma(2,1)');
```

```
ylabel('X2 ~ t(5)');
subplot(2,2,4);
bar(ctr1,-n1,1);
axis([0 15 -max(n1)*1.1 0]);
axis('off');
h2 = gca;
subplot(2,2,1);
barh(ctr2,-n2,1);
axis([-max(n2)*1.1 0 -10 10]);
axis('off');
h3 = gca;
h1.Position = [0.35 0.35 0.55 0.55];
h2.Position = [.35 .1 .55 .15];
h3.Position = [.1 .35 .15 .55];
colormap([.8 .8 1]);
```

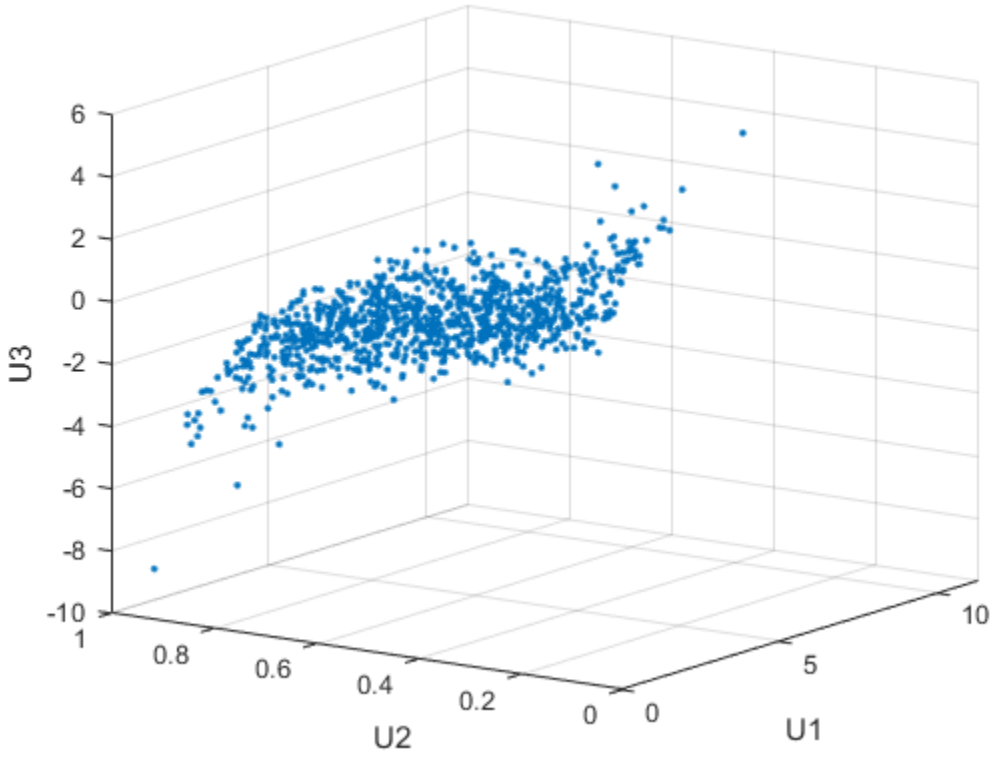


Compared to the bivariate Gamma/t distribution constructed earlier, which was based on a Gaussian copula, the distribution constructed here, based on a t(1) copula, has the same marginal distributions and the same rank correlation between variables, but a very different dependence structure. This illustrates the fact that multivariate distributions are not uniquely defined by their marginal distributions, or by their correlations. The choice of a particular copula in an application may be based on actual observed data, or different copulas may be used as a way of determining the sensitivity of simulation results to the input distribution.

Higher-Order Copulas

The Gaussian and t copulas are known as elliptical copulas. It's easy to generalize elliptical copulas to a higher number of dimensions. For example, we can simulate data from a trivariate distribution with Gamma(2,1), Beta(2,2), and t(5) marginals using a Gaussian copula as follows.

```
subplot(1,1,1);
n = 1000;
Rho = [1 .4 .2; .4 1 -.8; .2 -.8 1];
Z = mvnrnd([0 0 0], Rho, n);
U = normcdf(Z,0,1);
X = [gaminv(U(:,1),2,1) betainv(U(:,2),2,2) tinv(U(:,3),5)];
plot3(X(:,1),X(:,2),X(:,3),'.');
grid on;
view([-55, 15]);
xlabel('U1');
ylabel('U2');
zlabel('U3');
```



Notice that the relationship between the linear correlation parameter rho and, for example, Kendall's tau, holds for each entry in the correlation matrix Rho used here. We can verify that the sample rank correlations of the data are approximately equal to the theoretical values.

```
tauTheoretical = 2.*asin(Rho)./pi
```

```
tauTheoretical =
```

```
    1.0000    0.2620    0.1282
    0.2620    1.0000   -0.5903
    0.1282   -0.5903    1.0000
```

```
tauSample = corr(X, 'type','Kendall')
```

```
tauSample =

    1.0000    0.2655    0.1060
    0.2655    1.0000   -0.6076
    0.1060   -0.6076    1.0000
```

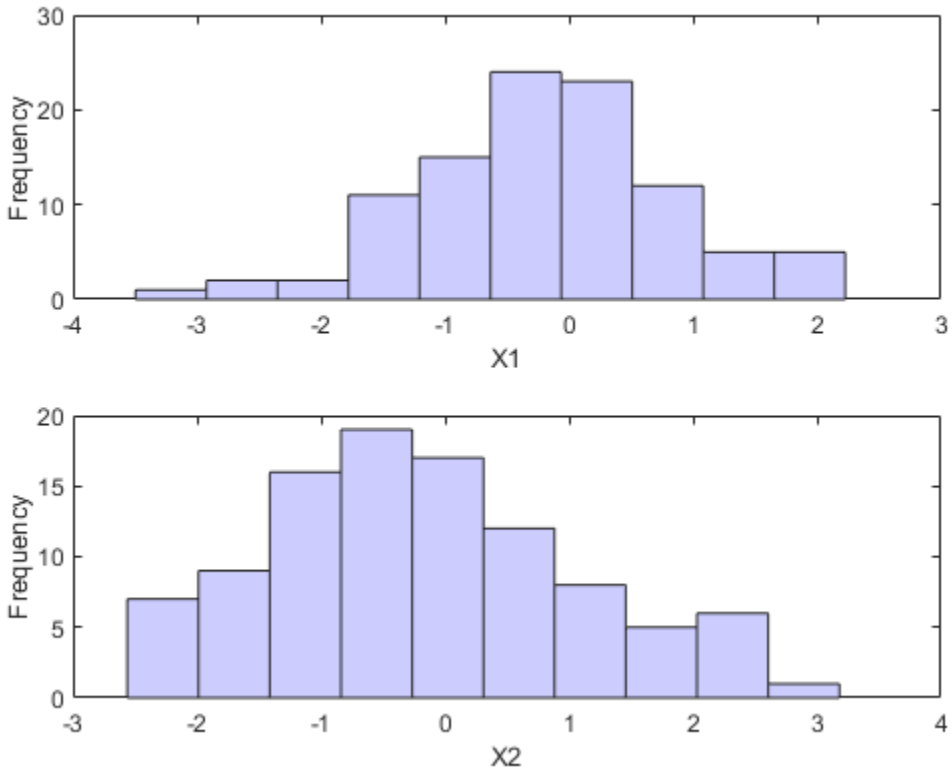
Copulas and Empirical Marginal Distributions

To simulate dependent multivariate data using a copula, we have seen that we need to specify

```
  1) the copula family (and any shape parameters),
  2) the rank correlations among variables, and
  3) the marginal distributions for each variable
```

Suppose we have two sets of stock return data, and we would like to run a Monte Carlo simulation with inputs that follow the same distributions as our data.

```
load stockreturns
nobs = size(stocks,1);
subplot(2,1,1);
hist(stocks(:,1),10);
xlabel('X1');
ylabel('Frequency');
subplot(2,1,2);
hist(stocks(:,2),10);
xlabel('X2');
ylabel('Frequency');
```

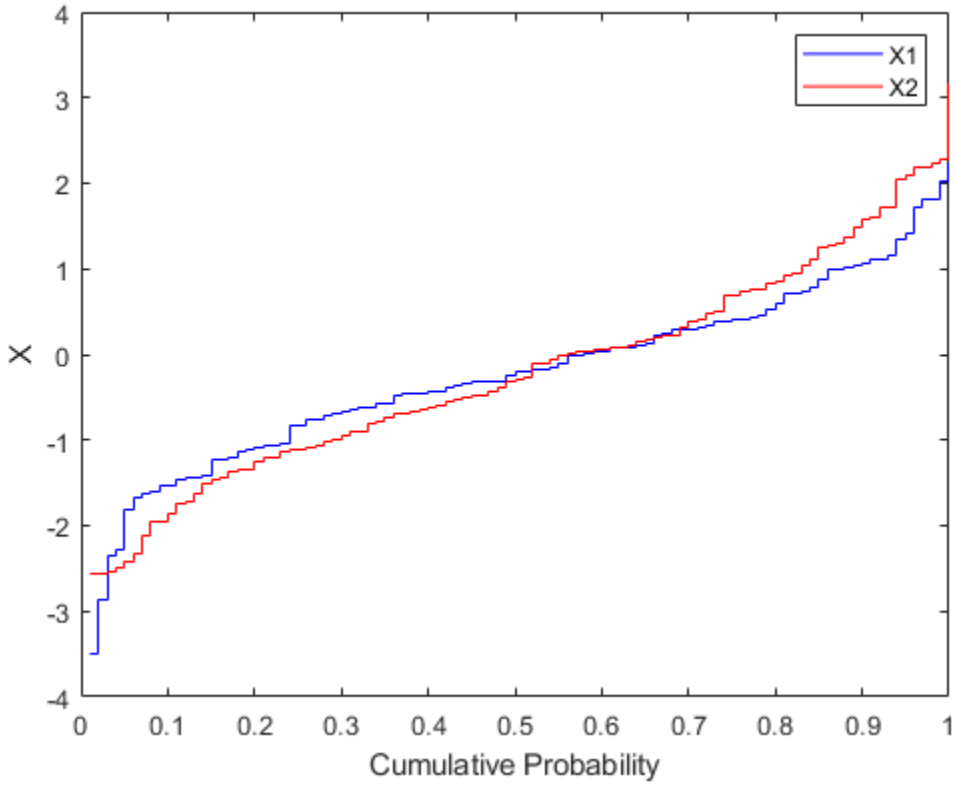


(These two data vectors have the same length, but that is not crucial.)

We could fit a parametric model separately to each dataset, and use those estimates as our marginal distributions. However, a parametric model may not be sufficiently flexible. Instead, we can use an empirical model for the marginal distributions. We only need a way to compute the inverse CDF.

The empirical inverse CDF for these datasets is just a stair function, with steps at the values 1/nobs, 2/nobs, ... 1. The step heights are simply the sorted data.

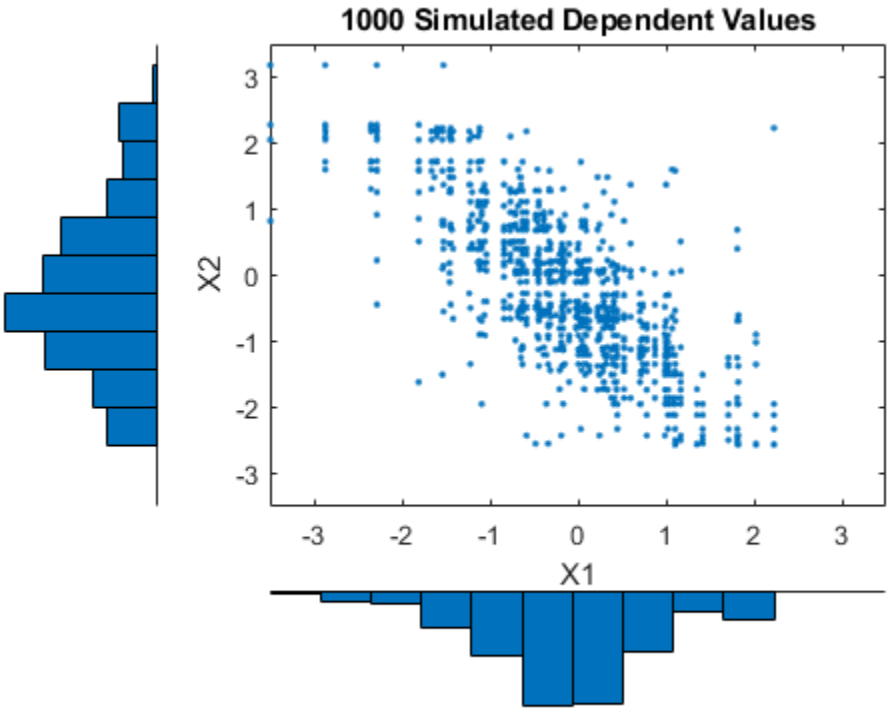
```
invCDF1 = sort(stocks(:,1));
n1 = length(stocks(:,1));
invCDF2 = sort(stocks(:,2));
n2 = length(stocks(:,2));
subplot(1,1,1);
stairs((1:nobs)/nobs, invCDF1,'b');
hold on;
stairs((1:nobs)/nobs, invCDF2,'r');
hold off;
legend('X1','X2');
xlabel('Cumulative Probability');
ylabel('X');
```



For the simulation, we might want to experiment with different copulas and correlations. Here, we'll use a bivariate t(5) copula with a fairly large negative correlation parameter.

```
n = 1000;
rho = -.8;
nu = 5;
T = mvtrnd([1 rho; rho 1], nu, n);
U = tcdf(T,nu);
X = [invCDF1(ceil(n1*U(:,1))) invCDF2(ceil(n2*U(:,2)))];

[n1,ctr1] = hist(X(:,1),10);
[n2,ctr2] = hist(X(:,2),10);
subplot(2,2,2);
plot(X(:,1),X(:,2),'.');
axis([-3.5 3.5 -3.5 3.5]);
h1 = gca;
title('1000 Simulated Dependent Values');
xlabel('X1');
ylabel('X2');
subplot(2,2,4);
bar(ctr1,-n1,1);
axis([-3.5 3.5 -max(n1)*1.1 0]);
axis('off');
h2 = gca;
subplot(2,2,1);
barh(ctr2,-n2,1);
axis([-max(n2)*1.1 0 -3.5 3.5]);
```

```
axis('off');
h3 = gca;
h1.Position = [0.35 0.35 0.55 0.55];
h2.Position = [.35 .1 .55 .15];
h3.Position = [.1 .35 .15 .55];
colormap([.8 .8 1]);
```



The marginal histograms of the simulated data closely match those of the original data, and would become identical as we simulate more pairs of values. Notice that the values are drawn from the original data, and because there are only 100 observations in each dataset, the simulated data are somewhat "discrete". One way to overcome this would be to add a small amount of random variation, possibly normally distributed, to the final simulated values. This is equivalent to using a smoothed version of the empirical inverse CDF.

Fitting Custom Univariate Distributions

This example shows how to use the Statistics and Machine Learning Toolbox™ function `mle` to fit custom distributions to univariate data.

Using `mle`, you can compute maximum likelihood parameter estimates, and estimate their precision, for many kinds of distributions beyond those for which the Toolbox provides specific fitting functions.

To do this, you need to define the distribution using one or more M functions. In the simplest cases, you can write code to compute the probability density function (PDF) for the distribution that you want to fit, and `mle` will do most of the remaining work for you. This example covers those cases. In problems with censored data, you must also write code to compute the cumulative distribution function (CDF) or the survival function (SF). In some other problems, it may be advantageous to define the log-likelihood function (LLF) itself. The second part of this example, “Fitting Custom Univariate Distributions, Part 2” on page 5-176, covers both of those latter cases.

Fitting Custom Distributions: A Zero-Truncated Poisson Example

Count data are often modelled using a Poisson distribution, and you can use the Statistics and Machine Learning Toolbox function `poissfit` to fit a Poisson model. However, in some situations, counts that are zero do not get recorded in the data, and so fitting a Poisson distribution is not straight-forward because of those "missing zeros". This example will show how to fit a Poisson distribution to zero-truncated data, using the function `mle`.

For this example, we'll use simulated data from a zero-truncated Poisson distribution. First, we generate some random Poisson data.

```
rng(18,'twister');
n = 75;
lambda = 1.75;
x = poissrnd(lambda,n,1);
```

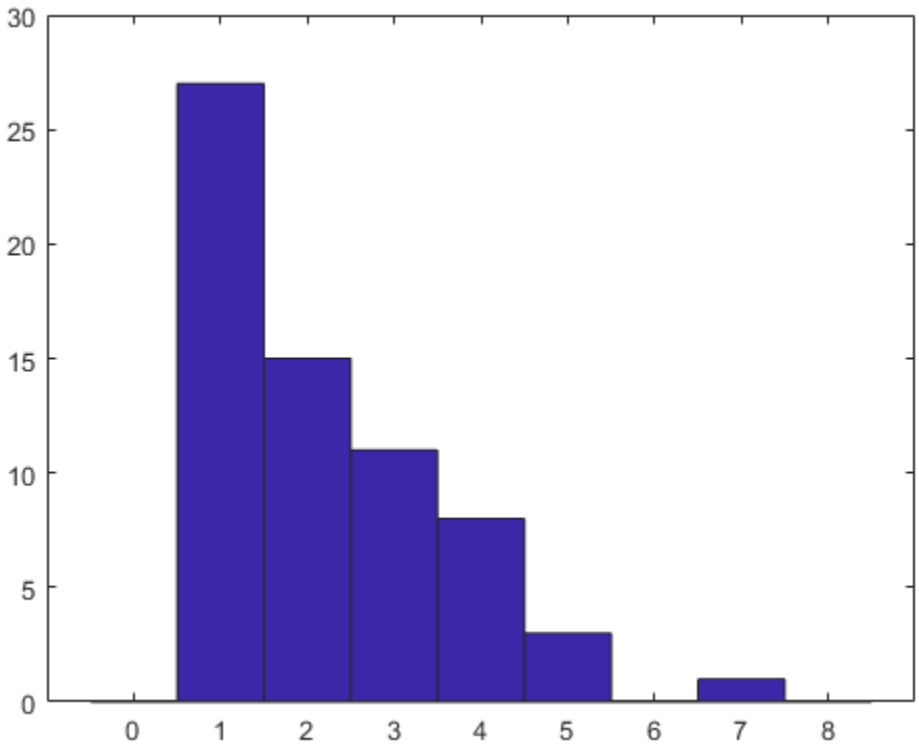
Next, we remove all the zeros from the data to simulate the truncation.

```
x = x(x > 0);
length(x)
```

```
ans = 65
```

Here's a histogram of these simulated data. Notice that the data look reasonably like a Poisson distribution, except that there are no zeros. We will fit them with a distribution that is identical to a Poisson on the positive integers, but that has no probability at zero. In this way, we can estimate the Poisson parameter lambda while accounting for the "missing zeros".

```
hist(x,[0:1:max(x)+1]);
```



The first step is to define the zero-truncated Poisson distribution by its probability function (PF). We will create a function to compute the probability for each point in x, given a value for the Poisson distribution's mean parameter lambda. The PF for a zero-truncated Poisson is just the usual Poisson PF, renormalized so that it sums to one. With zero truncation, the renormalization is just 1-Pr{0}. The easiest way to create a function for the PF is to use an anonymous function.

```
pf_truncpoiss = @(x,lambda) poisspdf(x,lambda) ./ (1-poisscdf(0,lambda));
```

For simplicity, we have assumed that all the x values given to this function will be positive integers, with no checks. Error checking, or a more complicated distribution, would probably take more than a single line of code, suggesting that the function should be defined in a separate file.

The next step is to provide a reasonable rough first guess for the parameter lambda. In this case, we'll just use the sample mean.

```
start = mean(x)
```

```
start = 2.2154
```

We provide `mle` with the data, and with the anonymous function, using the 'pdf' parameter. (The Poisson is discrete, so this is really a probability function, not a PDF.) Because the mean parameter of the Poisson distribution must be positive, we also specify a lower bound for lambda. `mle` returns the maximum likelihood estimate of lambda, and, optionally, approximate 95% confidence intervals for the parameters.

```
[lambdaHat,lambdaCI] = mle(x, 'pdf',pf_truncpoiss, 'start',start, 'lower',0)
```

```
lambdaHat = 1.8760

lambdaCI = 2×1

    1.4990
    2.2530
```

Notice that the parameter estimate is smaller than the sample mean. That's just as it should be, because the maximum likelihood estimate accounts for the missing zeros not present in the data.

We can also compute a standard error estimate for lambda, using the large-sample variance approximation returned by `mlecov`.

```
avar = mlecov(lambdaHat, x, 'pdf',pf_truncpoiss);
stderr = sqrt(avar)

stderr = 0.1923
```

Supplying Additional Values to the Distribution Function: A Truncated Normal Example

It sometimes also happens that continuous data are truncated. For example, observations larger than some fixed value might not be recorded because of limitations in the way the data are collected. This example will show how to fit a normal distribution to truncated data, using the function `mle`.

For this example, we simulate data from a truncated normal distribution. First, we generate some random normal data.

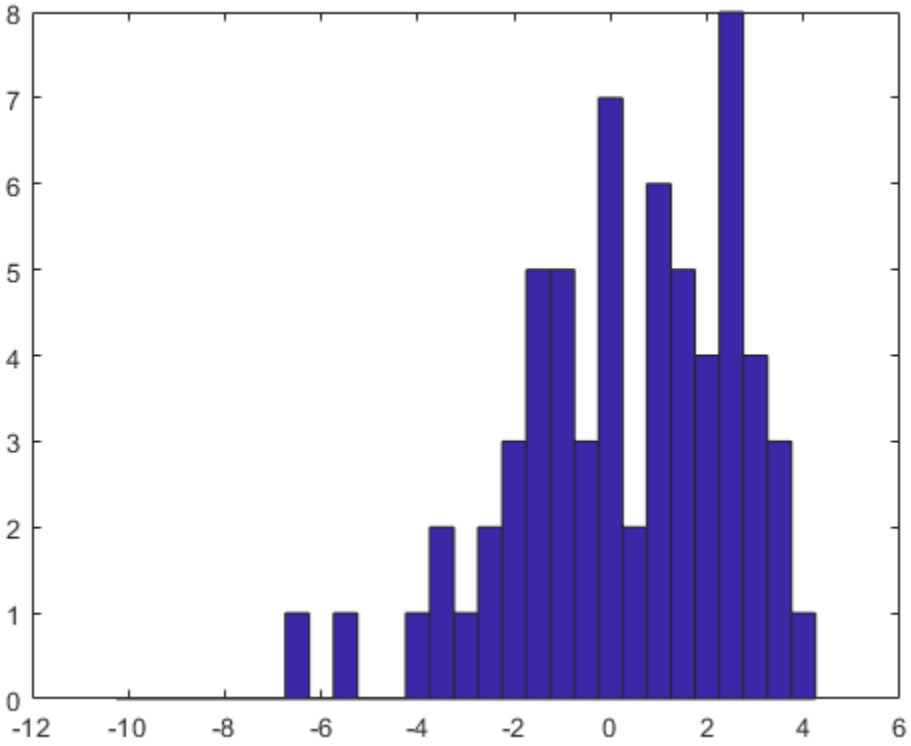
```
n = 75;
mu = 1;
sigma = 3;
x = normrnd(mu,sigma,n,1);
```

Next, we remove any observations that fall beyond the truncation point, xTrunc. Throughout this example, we'll assume that xTrunc is known, and does not need to be estimated.

```
xTrunc = 4;
x = x(x < xTrunc);
length(x)

ans = 64
```

Here's a histogram of these simulated data. We will fit them with a distribution that is identical to a normal for x < xTrunc, but that has zero probability above xTrunc. In this way, we can estimate the normal parameters mu and sigma while accounting for the "missing tail".

```
hist(x,(-10:.5:4));
```



As in the previous example, we will define the truncated normal distribution by its PDF, and create a function to compute the probability density for each point in x, given values for the parameters mu and sigma. With the truncation point fixed and known, the PDF for a truncated normal is just the usual normal PDF, truncated, and then renormalized so that it integrates to one. The renormalization is just the CDF evaluated at xTrunc. For simplicity, we'll assume that all x values are less than xTrunc, without checking. We'll use an anonymous function to define the PDF.

```
pdf_truncnorm = @(x,mu,sigma) normpdf(x,mu,sigma) ./ normcdf(xTrunc,mu,sigma);
```

The truncation point, xTrunc, is not being estimated, and so it is not among the distribution parameters in the PDF function's input argument list. xTrunc is also not part of the data vector input argument. With an anonymous function, we can simply refer to the variable xTrunc that has already been defined in the workspace, and there is no need to worry about passing it in as an additional argument.

We also need to provide a rough starting guess for the parameter estimates. In this case, because the truncation is not too extreme, the sample mean and standard deviation will probably work well.

```
start = [mean(x),std(x)]
```

```
start = 1×2

    0.2735    2.2660
```

We provide `mle` with the data, and with the anonymous function, using the 'pdf' parameter. Because sigma must be positive, we also specify lower parameter bounds. `mle` returns the maximum likelihood

estimates of mu and sigma as a single vector, as well as a matrix of approximate 95% confidence intervals for the two parameters.

```
[paramEsts,paramCIs] = mle(x, 'pdf',pdf_truncnorm, 'start',start, 'lower',[-Inf 0])
```

```
paramEsts = 1×2

    0.9911    2.7800


paramCIs = 2×2

   -0.1605    1.9680
    2.1427    3.5921
```

Notice that the estimates of mu and sigma are quite a bit larger than the sample mean and standard deviation. This is because the model fit has accounted for the "missing" upper tail of the distribution.

We can compute an approximate covariance matrix for the parameter estimates using `mlecov`. The approximation is usually reasonable in large samples, and the standard errors of the estimates can be approximated by the square roots of the diagonal elements.

```
acov = mlecov(paramEsts, x, 'pdf',pdf_truncnorm)
```

```
acov = 2×2

    0.3452    0.1660
    0.1660    0.1717


stderr = sqrt(diag(acov))
```

```
stderr = 2×1

    0.5876
    0.4143
```

Fitting a More Complicated Distribution: A Mixture of Two Normals

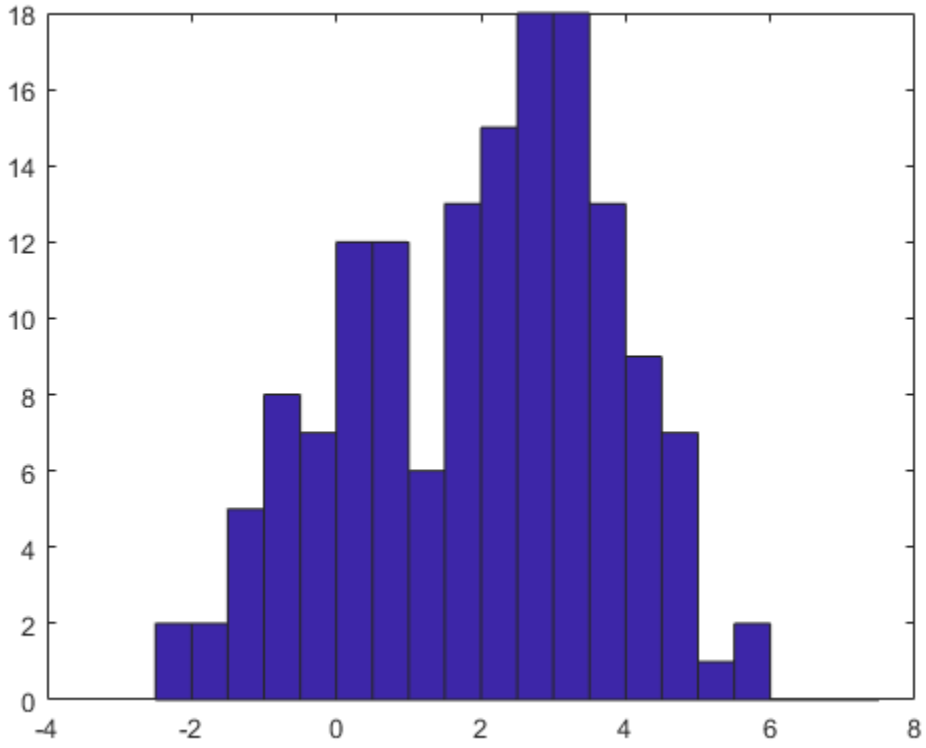
Some datasets exhibit bimodality, or even multimodality, and fitting a standard distribution to such data is usually not appropriate. However, a mixture of simple unimodal distributions can often model such data very well. In fact, it may even be possible to give an interpretation to the source of each component in the mixture, based on application-specific knowledge.

In this example, we will fit a mixture of two normal distributions to some simulated data. This mixture might be described with the following constructive definition for generating a random value:

```
First, flip a biased coin.  If it lands heads, pick a value at random
from a normal distribution with mean mu_1 and standard deviation
sigma_1. If the coin lands tails, pick a value at random from a normal
distribution with mean mu_2 and standard deviation sigma_2.
```

For this example, we'll generate data from a mixture of Student's t distributions rather than using the same model as we are fitting. This is the sort of thing you might do in a Monte-Carlo simulation to test how robust a fitting method is to departures from the assumptions of the model being fit. Here, however, we'll fit just one simulated data set.

```
x = [trnd(20,1,50) trnd(4,1,100)+3];
hist(x,-2.25:.5:7.25);
```



As in the previous examples, we'll define the model to fit by creating a function that computes the probability density. The PDF for a mixture of two normals is just a weighted sum of the PDFs of the two normal components, weighted by the mixture probability. This PDF is simple enough to create using an anonymous function. The function takes six inputs: a vector of data at which to evaluate the PDF, and the distribution's five parameters. Each component has parameters for its mean and standard deviation; the mixture probability makes a total of five.

```
pdf_normmixture = @(x,p,mu1,mu2,sigma1,sigma2) ...
                        p*normpdf(x,mu1,sigma1) + (1-p)*normpdf(x,mu2,sigma2);
```

We'll also need an initial guess for the parameters. The more parameters a model has, the more a reasonable starting point matters. For this example, we'll start with an equal mixture (p = 0.5) of normals, centered at the two quartiles of the data, with equal standard deviations. The starting value for standard deviation comes from the formula for the variance of a mixture in terms of the mean and variance of each component.

```
pStart = .5;
muStart = quantile(x,[.25 .75])
```

```
muStart = 1×2

    0.5970    3.2456
```

```
sigmaStart = sqrt(var(x) - .25*diff(muStart).^2)
```

```
sigmaStart = 1.2153

start = [pStart muStart sigmaStart sigmaStart];
```

Finally, we need to specify bounds of zero and one for the mixing probability, and lower bounds of zero for the standard deviations. The remaining elements of the bounds vectors are set to +Inf or -Inf, to indicate no restrictions.

```
lb = [0 -Inf -Inf 0 0];
ub = [1 Inf Inf Inf Inf];
paramEsts = mle(x, 'pdf',pdf_normmixture, 'start',start, ...
                           'lower',lb, 'upper',ub)

Warning: Maximum likelihood estimation did not converge.  Iteration limit exceeded.

paramEsts = 1×5

    0.3523    0.0257    3.0489    1.0546    1.0629
```

The default for custom distributions is 200 iterations.

```
statset('mlecustom')

ans = struct with fields:
          Display: 'off'
      MaxFunEvals: 400
          MaxIter: 200
           TolBnd: 1.0000e-06
           TolFun: 1.0000e-06
        TolTypeFun: []
             TolX: 1.0000e-06
         TolTypeX: []
          GradObj: 'off'
         Jacobian: []
        DerivStep: 6.0555e-06
      FunValCheck: 'on'
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

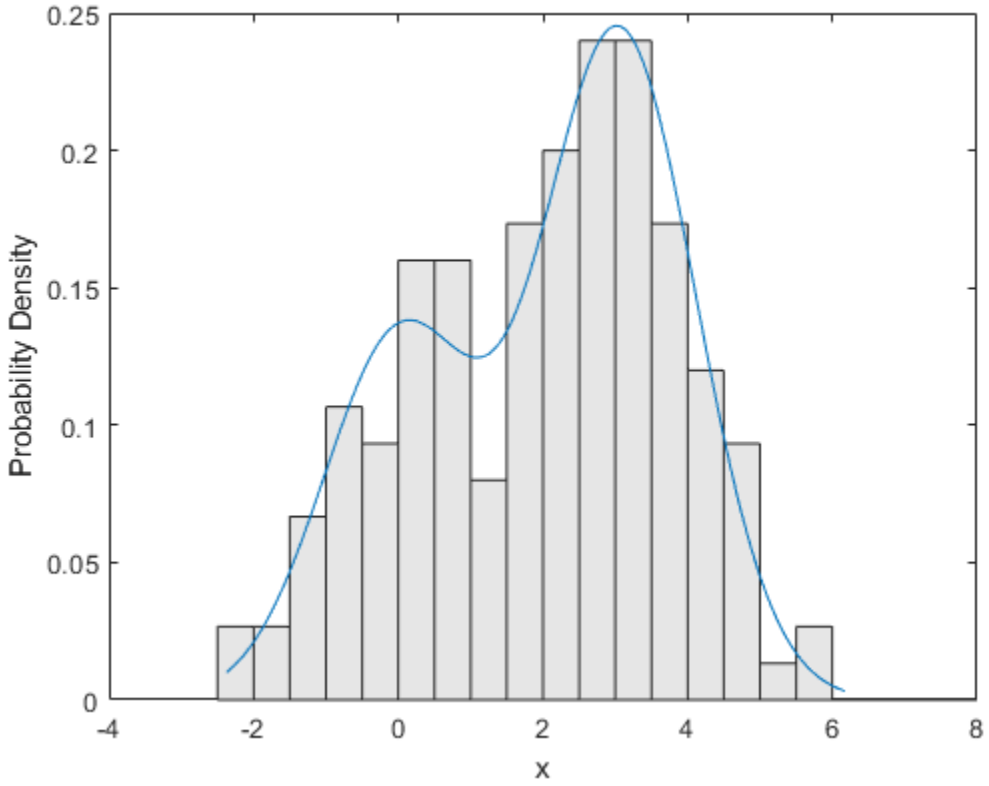
Override that default, using an options structure created with the `statset` function. Also increase the function evaluation limit.

```
options = statset('MaxIter',300, 'MaxFunEvals',600);
paramEsts = mle(x, 'pdf',pdf_normmixture, 'start',start, ...
                           'lower',lb, 'upper',ub, 'options',options)

paramEsts = 1×5

    0.3523    0.0257    3.0489    1.0546    1.0629
```

It appears that the final iterations to convergence mattered only in the last few digits of the result. Nonetheless, it is always a good idea to make sure that convergence has been reached.

Finally, we can plot the fitted density against a probability histogram of the raw data, to check the fit visually.

```
bins = -2.5:.5:7.5;
h = bar(bins,histc(x,bins)/(length(x)*.5),'histc');
h.FaceColor = [.9 .9 .9];
xgrid = linspace(1.1*min(x),1.1*max(x),200);
pdfgrid = pdf_normmixture(xgrid,paramEsts(1),paramEsts(2),paramEsts(3),paramEsts(4),paramEsts(5))
hold on
plot(xgrid,pdfgrid,'-')
hold off
xlabel('x')
ylabel('Probability Density')
```



Using Nested Functions: A Normal Example with Unequal Precisions

It sometimes happens when data are collected that each observation was made with a different precision or reliability. For example, if several experimenters each make a number of independent measurements of the same quantity, but each reports only the average of their measurements, the reliability of each reported data point will depend on the number of raw observations that went into it. If the original raw data are not available, an estimate of their distribution must account for the fact that the data that are available, the averages, each have different variances. This model actually has an explicit solution for maximum likelihood parameter estimates. However, for the purposes of illustration, we will use `mle` to estimate the parameters.

Assume that we have 10 data points, where each one is actually the average of anywhere from 1 to 8 observations. Those original observations are not available, but we know how many there were for each of our data points. We need to estimate the mean and standard deviation of the raw data.

```
x = [0.25 -1.24 1.38 1.39 -1.43 2.79 3.52 0.92 1.44 1.26];
m = [   8     2    1    3     8    4    2    5    2    4];
```

The variance of each data point is inversely proportional to the number of observations that went into it, so we will use 1/m to weight the variance of each data point in a maximum likelihood fit.

```
w = 1 ./ m
```

```
w = 1×10

    0.1250    0.5000    1.0000    0.3333    0.1250    0.2500    0.5000    0.2000    0.5000    0.2
```

In the model we're fitting here, we could define the distribution by its PDF, but using a log PDF is somewhat more natural, because the normal PDF is of the form

```
  c .* exp(-0.5 .* z.^2),
```

and `mle` would have to take the log of the PDF anyway, to compute the log-likelihood. So instead, we will create a function that computes the log PDF directly.

The log PDF function has to compute the log of the probability density for each point in x, given values for mu and sigma. It will also need to account for the different variance weights. Unlike the previous examples, this distribution function is a little more complicated than a one-liner, and is most clearly written as a separate function in its own file. Because the log PDF function needs the observation counts as additional data, the most straight-forward way to accomplish this fit is to use nested functions.

We've created a separate file for a function called `wgtnormfit.m`. This function contains data initialization, a nested function for the log PDF in the weighted normal model, and a call to the `mle` function to actually fit the model. Because sigma must be positive, we must specify lower parameter bounds. The call to `mle` returns the maximum likelihood estimates of mu and sigma in a single vector.

```
type wgtnormfit.m
```

```
function paramEsts = wgtnormfit
%WGTNORMFIT Fitting example for a weighted normal distribution.

%   Copyright 1984-2012 The MathWorks, Inc.


x = [0.25 -1.24 1.38 1.39 -1.43 2.79 3.52 0.92 1.44 1.26]';
m = [   8     2    1    3     8    4    2    5    2    4]';

    function logy = logpdf_wn(x,mu,sigma)
        v = sigma.^2 ./ m;
        logy = -(x-mu).^2 ./ (2.*v) - .5.*log(2.*pi.*v);
    end

paramEsts = mle(x, 'logpdf',@logpdf_wn, 'start',[mean(x),std(x)], 'lower',[-Inf,0]);

end
```

In `wgtnormfit.m`, we pass `mle` a handle to the nested function `logpdf_wn`, using the 'logpdf' parameter. That nested function refers to the observation counts, m, in the computation of the weighted log PDF. Because the vector m is defined in its parent function, `logpdf_wn` has access to it, and there is no need to worry about passing m in as an explicit input argument.

We need to provide a rough first guess for the parameter estimates. In this case, the unweighted sample mean and standard deviation should be ok, and that's what `wgtnormfit.m` uses.

```
start = [mean(x),std(x)]
```

```
start = 1×2

    1.0280    1.5490
```

To fit the model, we run the fitting function.

```
paramEsts = wgtnormfit
```

```
paramEsts = 1×2

    0.6244    2.8823
```

Notice that the estimate of mu is less than two-thirds that of the sample mean. That's just as it should be: the estimate is be influenced most by the most reliable data points, i.e., the ones that were based on the largest number of raw observations. In this dataset, those points tend to pull the estimate down from the unweighted sample mean.

Using a Parameter Transformation: The Normal Example (continued)

In maximum likelihood estimation, confidence intervals for the parameters are usually computed using a large-sample normal approximation for the distribution of the estimators. This is often a reasonable assumption, but with small sample sizes, it is sometimes advantageous to improve that normal approximation by transforming one or more parameters. In this example, we have a location parameter and a scale parameter. Scale parameters are often transformed to their log, and we will do that here with sigma. First, we'll create a new log PDF function, and then recompute the estimates using that parameterization.

The new log PDF function is created as a nested function within the function `wgtnormfit2.m`. As in the first fit, this file contains data initialization, a nested function for the log PDF in the weighted normal model, and a call to the `mle` function to actually fit the model. Because sigma can be any positive value, log(sigma) is unbounded, and we no longer need to specify lower or upper bounds. Also, the call to `mle` in this case returns both the parameter estimates and confidence intervals.

```
type wgtnormfit2.m
```

```
function [paramEsts,paramCIs] = wgtnormfit2
%WGTNORMFIT2 Fitting example for a weighted normal distribution (log(sigma) parameterization).

%   Copyright 1984-2012 The MathWorks, Inc.


x = [0.25 -1.24 1.38 1.39 -1.43 2.79 3.52 0.92 1.44 1.26]';
m = [   8     2    1    3     8    4    2    5    2    4]';

    function logy = logpdf_wn2(x,mu,logsigma)
```

```
        v = exp(logsigma).^2 ./ m;
        logy = -(x-mu).^2 ./ (2.*v) - .5.*log(2.*pi.*v);
    end

[paramEsts,paramCIs] = mle(x, 'logpdf',@logpdf_wn2, 'start',[mean(x),log(std(x))]);

end
```

Notice that `wgtnormfit2.m` uses the same starting point, transformed to the new parameterization, i.e., the log of the sample standard deviation.

```
start = [mean(x),log(std(x))]
```

```
start = 1×2

    1.0280    0.4376
```

```
[paramEsts,paramCIs] = wgtnormfit2
```

```
paramEsts = 1×2

    0.6244    1.0586


paramCIs = 2×2

   -0.2802    0.6203
    1.5290    1.4969
```

Since the parameterization uses log(sigma), we have to transform back to the original scale to get an estimate and confidence interval for sigma. Notice that the estimates for both mu and sigma are the same as in the first fit, because maximum likelihood estimates are invariant to parameterization.

```
muHat = paramEsts(1)
```

```
muHat = 0.6244
```

```
sigmaHat = exp(paramEsts(2))
```

```
sigmaHat = 2.8823
```

```
muCI = paramCIs(:,1)
```

```
muCI = 2×1

   -0.2802
    1.5290
```

```
sigmaCI = exp(paramCIs(:,2))
```

```
sigmaCI = 2×1

    1.8596
    4.4677
```

Fitting Custom Univariate Distributions, Part 2

This example shows how to use some more advanced techniques with the Statistics and Machine Learning Toolbox™ function `mle` to fit custom distributions to univariate data. The techniques include fitting models to censored data, and illustration of some of the numerical details of fitting with custom distributions.

See “Fitting Custom Univariate Distributions” on page 5-165 for additional examples of fitting custom distributions to univariate data.

Fitting Custom Distributions with Censored Data

The extreme value distribution is often used to model failure times of mechanical parts, and experiments in such cases are sometimes only run for a fixed length of time. If not all of the experimental units have failed within that time, then the data are right-censored, that is, the value of some failure times are not known exactly, but only known to be larger than a certain value.

The Statistics and Machine Learning Toolbox includes the function `evfit`, which fits an extreme value distribution to data, including data with censoring. However, for the purposes of this example, we will ignore `evfit`, and show how to use `mle` and custom distributions to fit a model to censored data, using the extreme value distribution.

Because the values for the censored data are not known exactly, maximum likelihood estimation becomes more difficult. In particular, both the PDF and the CDF are needed to compute the log-likelihood. Therefore, you must provide `mle` with functions for both of those in order to fit censored data. The Statistics and Machine Learning Toolbox includes the functions `evpdf` and `evcdf`, so for this example, the work of writing the code has already been done.

We'll fit the model to simulated data. The first step is to generate some uncensored extreme value data.

```
rng(0,'twister');
n = 50;
mu = 5;
sigma = 2.5;
x = evrnd(mu,sigma,n,1);
```

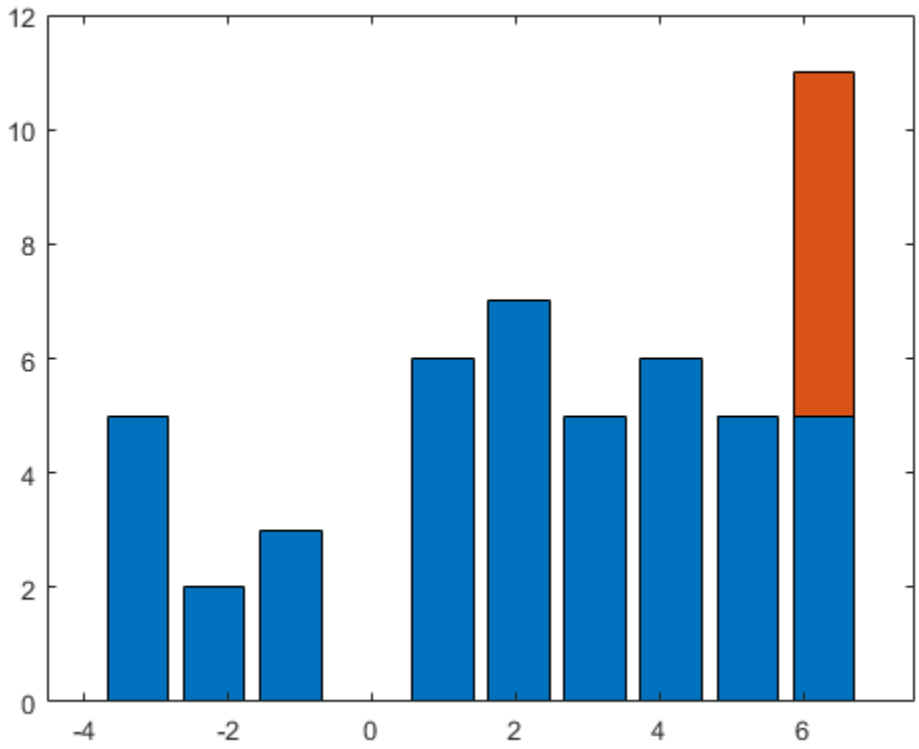
Next, we censor any values that are larger than a predetermined cutoff, by replacing those values with the cutoff value. This is known as Type II censoring, and with the cutoff at 7, about 12% of the original data end up being censored.

```
c = (x > 7);
x(c) = 7;
sum(c)/length(c)
```

```
ans = 0.1200
```

We can plot a histogram of these data, including a stacked bar to represent the censored observations.

```
[uncensCnts,binCtrs] = hist(x(~c));
censCnts = hist(x(c),binCtrs);
bar(binCtrs,[uncensCnts' censCnts'],'stacked');
```



Although there is censoring, the fraction of censored observations is relatively small, and so the method of moments can provide a reasonable starting point for the parameter estimates. We compute the values of mu and sigma that correspond to the observed mean and standard deviation of the uncensored data.

```
sigma0 = sqrt(6)*std(x(~c))./pi
```

```
sigma0 = 2.3495
```

```
mu0 = mean(x(~c))-psi(1).*sigma0
```

```
mu0 = 3.5629
```

In addition to passing the data, x, and handles to the PDF and CDF functions into `mle`, we also use the 'censoring' parameter to pass in the censoring vector, c. Because the scale parameter, sigma, must be positive, we specify lower parameter bounds. `mle` returns the maximum likelihood estimates of the two extreme value distribution parameters, mu and sigma, as well as approximate 95% confidence intervals.

```
[paramEsts,paramCIs] = mle(x, 'censoring',c, 'pdf',@evpdf, 'cdf',@evcdf, ...
                           'start',[mu0 sigma0], 'lower',[-Inf,0])
```

```
paramEsts = 1×2

    4.5530    3.0215


paramCIs = 2×2
```

```
    3.6455    2.2937
    5.4605    3.7494
```

Some Numerical Issues in Fitting Custom Distributions

Fitting a custom distribution requires an initial guess for the parameters, and it's often difficult to know how good or bad a starting point is a priori. In the previous example, if we had picked a starting point that was farther away from the maximum likelihood estimates, some of the observations could have been very far out in the tails of the extreme value distribution corresponding to the starting point. One of two things might then have happened.

First, one of the PDF values might have become so small that it underflowed to zero in double precision arithmetic. Second, one of the CDF values might have become so close to 1 that it rounded up in double precision. (It's also possible that a CDF value might have become so small as to underflow, but that turns out not to be a problem.)

Either of these conditions causes problems when `mle` computes the log-likelihood, because they lead to log-likelihood values of -Inf, and the optimization algorithm in `mle` cannot normally be expected to step out of such regions.

Knowing what the maximum likelihood estimates are, let's see what happens with a different starting point.

```
start = [1 1];
try
    [paramEsts,paramCIs] = mle(x, 'censoring',c, 'pdf',@evpdf, 'cdf',@evcdf, ...
                               'start',start, 'lower',[-Inf,0])
catch ME
    disp(ME.message)
end
```

```
The CDF function returned values greater than or equal to 1.
```

In this case, the second problem has occurred: Some of the CDF values at the initial parameter guess are computed as exactly 1, and so the log-likelihood is infinite. We could try setting `mle`'s 'FunValCheck' control parameter to 'off', which would disable checking for non-finite likelihood values, and then hope for the best. But the right way to solve this numerical problem is at its root, and in this case it's not hard to do.

Notice that the extreme value CDF is of the form

```
  p = 1 - exp( -exp((x-mu)./sigma) )
```

The contribution of the censored observations to the log-likelihood is the log of their survival function (SF) values, i.e., log(1-CDF). For the extreme value distribution, the log of the SF is just -exp((x-mu)./sigma). If we could compute the log-likelihood using the log SF directly, (instead of, in effect, computing log(1 - (1-exp(logSF)))) we would avoid the rounding issues with the CDF. That's because observations whose CDF values are not distinguishable from 1 in double precision have log SF values that are still easily representable as non-zero values. For example, a CDF value of (1 - 1e-20) rounds to 1 in double precision, because double precision `eps` is about 2e-16.

```
SFval = 1e-20;
CDFval = 1 - SFval
```

```
CDFval = 1
```

However, the log of the corresponding SF value, i.e. log(1-CDF), is easily represented.

```
log(SFval)
```

```
ans = -46.0517
```

A similar observation can be made about using the log PDF rather than the PDF itself -- the contribution of uncensored observations to the log-likelihood is the log of their PDF values. Using the log PDF directly (instead of, in effect, computing log(exp(logPDF))) avoids underflow problems where the PDF is not distinguishable from zero in double precision, but the log PDF is still easily representable as a finite negative number. For example, a PDF value of 1e-400 underflows in double precision, because double precision `realmin` is about 2e-308.

```
logPDFval = -921;
PDFval = exp(logPDFval)
```

```
PDFval = 0
```

`mle` provides a syntax for specifying a custom distribution using the log PDF and the log SF (rather than the PDF and CDF), via the 'logpdf' and 'logsf' parameters. Unlike the PDF and CDF functions, there are no existing functions, so we'll create anonymous functions that compute these values:

```
evlogpdf = @(x,mu,sigma) ((x - mu) ./ sigma - exp((x - mu) ./ sigma)) - log(sigma);
evlogsf = @(x,mu,sigma) -exp((x-mu)./sigma);
```

Using the same starting point, the alternate logPDF/logSF specification of the extreme value distribution makes the problem solvable:

```
start = [1 1];
[paramEsts,paramCIs] = mle(x, 'censoring',c, 'logpdf',evlogpdf, 'logsf',evlogsf, ...
                           'start',start, 'lower',[-Inf,0])
```

```
paramEsts = 1×2

    4.5530    3.0215


paramCIs = 2×2

    3.6455    2.2937
    5.4605    3.7494
```

However, this strategy cannot always mitigate a poor starting point, and a careful choice of starting point is always recommended.

Supplying a Gradient

By default, `mle` uses the function `fminsearch` to find parameter values that maximize the log-likelihood for the data. `fminsearch` uses an optimization algorithm that is derivative-free, and is often a good choice.

However, for some problems, choosing an optimization algorithm that uses the derivatives of the log-likelihood function can make the difference between converging to the maximum likelihood estimates or not, especially when the starting point is far away from the final answer. Providing the derivatives can also sometimes speed up the convergence.

If your MATLAB® installation includes the Optimization Toolbox™, `mle` allows you to use the function `fmincon`, which includes optimization algorithms that can use derivative information. To take best advantage of the algorithms in `fmincon`, you can specify a custom distribution using a log-likelihood function, written to return not only the log-likelihood itself, but its gradient as well. The gradient of the log-likelihood function is simply the vector of its partial derivatives with respect to its parameters.

This strategy requires extra preparation, to write code that computes both the log-likelihood and its gradient. For this example, we've the created code to do that for the extreme value distribution as a separate file `evnegloglike.m`.

```
type evnegloglike.m
```

```
function [nll,ngrad] = evnegloglike(params,x,cens,freq)
%EVNEGLOGLIKE Negative log-likelihood for the extreme value distribution.

%   Copyright 1984-2004 The MathWorks, Inc.

if numel(params)~=2
    error(message('stats:probdists:WrongParameterLength',2));
end
mu = params(1);
sigma = params(2);
nunc = sum(1-cens);
z = (x - mu) ./ sigma;
expz = exp(z);
nll = sum(expz) - sum(z(~cens)) + nunc.*log(sigma);
if nargout > 1
    ngrad = [-sum(expz)./sigma + nunc./sigma, ...
             -sum(z.*expz)./sigma + sum(z(~cens))./sigma + nunc./sigma];
end
```

Notice that the function `evnegloglike` returns the *negative* of both the log-likelihood values and of the gradient values, because MLE *minimizes* that negative log-likelihood.

To compute the maximum likelihood estimates using a gradient-based optimization algorithm, we use the `'nloglf'` parameter, specifying that we are providing a handle to a function that computes the negative log-likelihood, and the `'optimfun'` parameter, specifying `fmincon` as the optimization function. `mle` will automatically detect that `evnegloglike` can return both the negative log-likelihood and its gradient.

```
start = [1 1];
[paramEsts,paramCIs] = mle(x, 'censoring',c, 'nloglf',@evnegloglike, ...
                           'start',start, 'lower',[-Inf,0], 'optimfun','fmincon')
```

```
paramEsts = 1×2

    4.5530    3.0215


paramCIs = 2×2

    3.6455    2.2937
    5.4605    3.7493
```

Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses

This example shows how to estimate the cumulative distribution function (CDF) from data in a nonparametric or semiparametric way. It also illustrates the inversion method for generating random numbers from the estimated CDF. The Statistics and Machine Learning Toolbox™ includes more than two dozen random number generator functions for parametric univariate probability distributions. These functions allow you to generate random inputs for a wide variety of simulations, however, there are situations where it is necessary to generate random values to simulate data that are not described by a simple parametric family.

The toolbox also includes the functions `pearsrnd` and `johnsrnd`, for generating random values without having to specify a parametric distribution from which to draw--those functions allow you to specify a distribution in terms of its moments or quantiles, respectively.

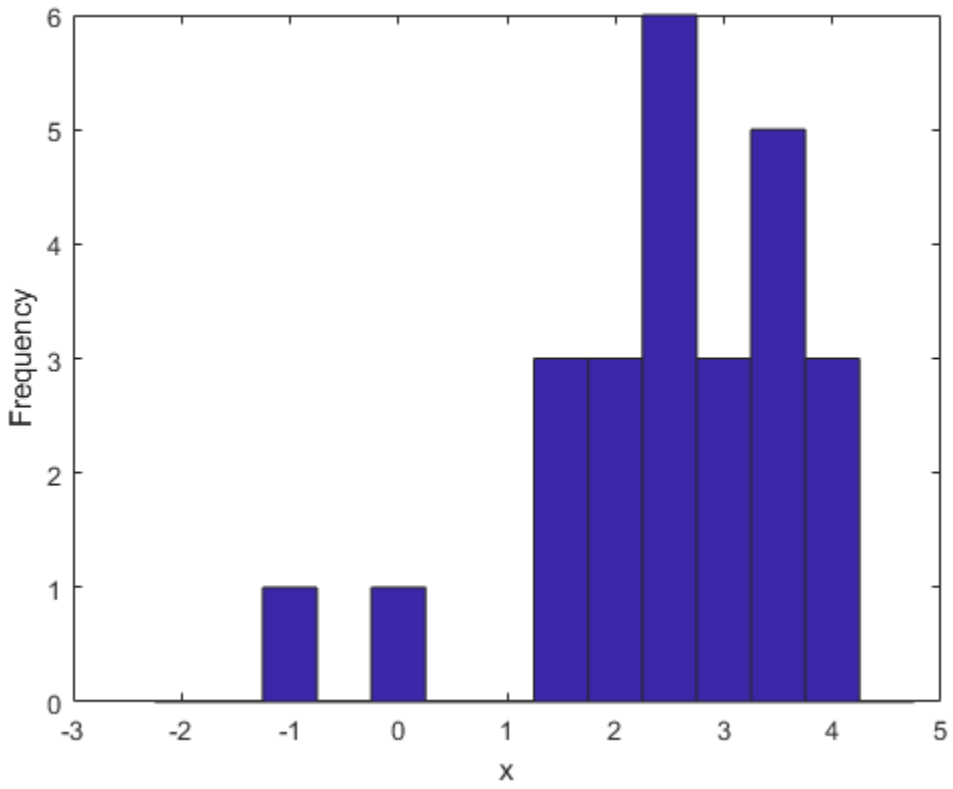
However, there are still situations where even more flexibility is needed, to generate random values that "imitate" data that you have collected even more closely. In this case, you might use a nonparametric estimate of the CDF of those data, and use the inversion method to generate random values. The inversion method involves generating uniform random values on the unit interval, and transforming them to a desired distribution using the inverse CDF for that distribution.

From the opposite perspective, it is sometimes desirable to use a nonparametric estimate of the CDF to transform observed data onto the unit interval, giving them an approximate uniform distribution.

The `ecdf` function computes one type of nonparametric CDF estimate, the empirical CDF, which is a stairstep function. This example illustrates some smoother alternatives, which may be more suitable for simulating or transforming data from a continuous distribution.

For the purpose of illustration, here are some simple simulated data. There are only 25 observations, a small number chosen to make the plots in the example easier to read. The data are also sorted to simplify plotting.

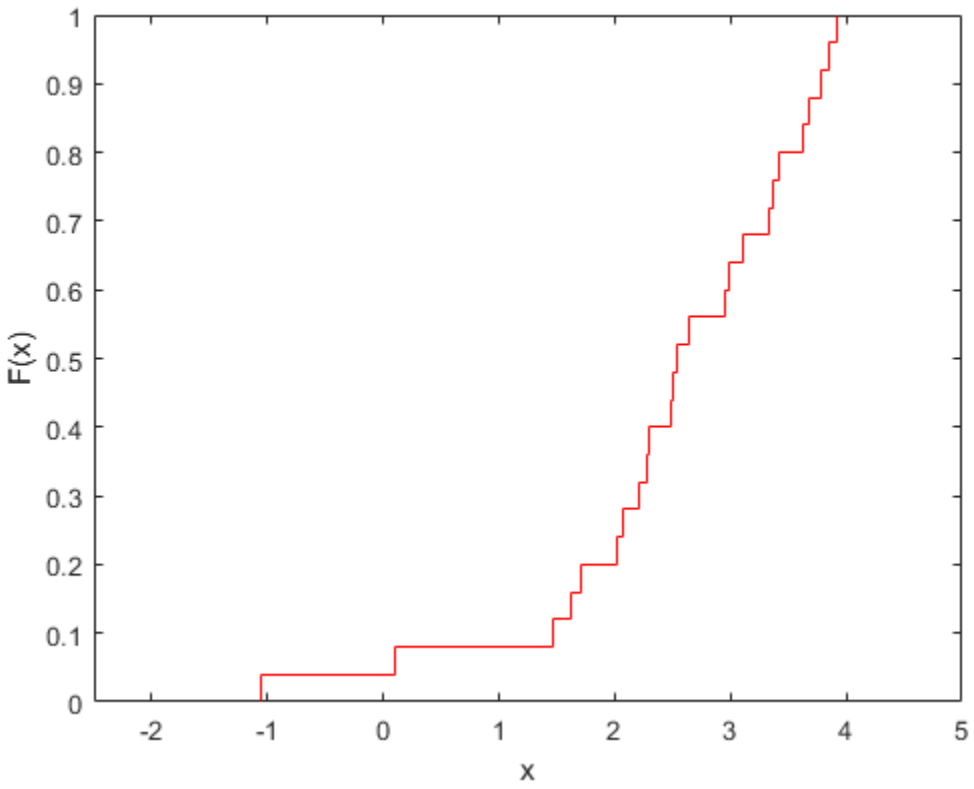
```
rng(19,'twister');
n = 25;
x = evrnd(3,1,n,1); x = sort(x);
hist(x,-2:.5:4.5);
xlabel('x'); ylabel('Frequency');
```



A Piecewise Linear Nonparametric CDF Estimate

The `ecdf` function provides a simple way to compute and plot a "stairstep" empirical CDF for data. In the simplest cases, this estimate makes discrete jumps of 1/n at each data point.

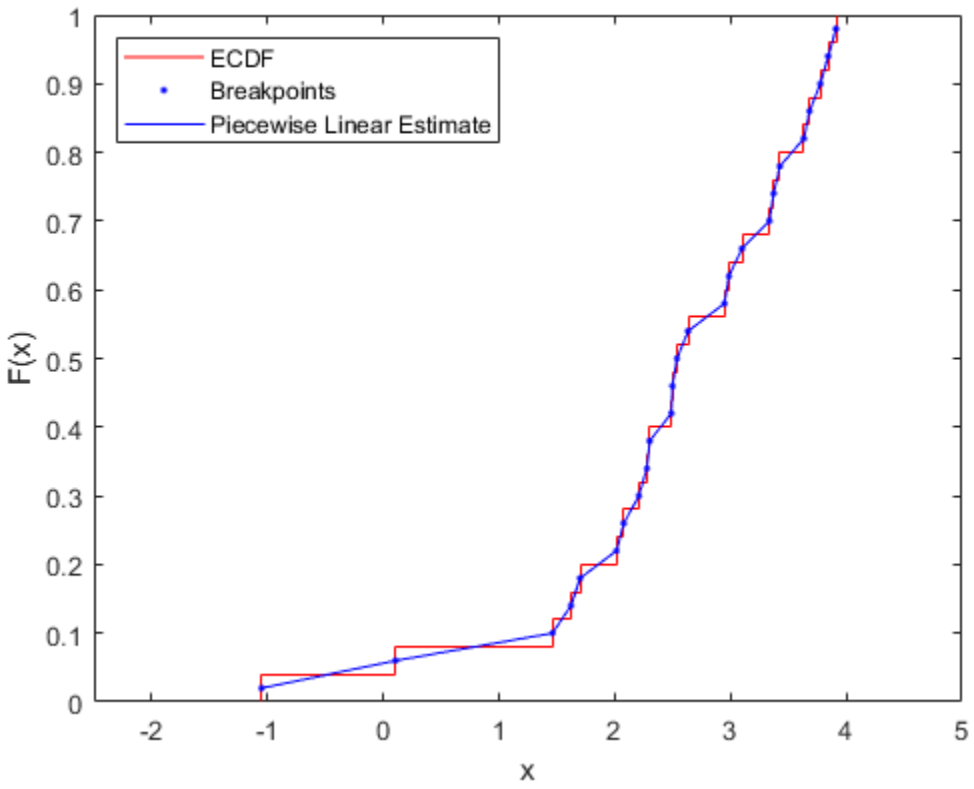
```
[Fi,xi] = ecdf(x);
stairs(xi,Fi,'r');
xlim([-2.5 5]); xlabel('x'); ylabel('F(x)');
```



This estimate is useful for many purposes, including investigating the goodness of fit of a parametric model to data. Its discreteness, however, may make it unsuitable for use in empirically transforming continuous data to or from the unit interval.

It is simple to modify the empirical CDF to address that problems. Instead of taking discrete jumps of 1/n at each data point, define a function that is piecewise linear, with breakpoints at the midpoints of those jumps. The height at each of the data points is then [1/2n, 3/2n, ..., (n-1/2)/n], instead of [(1/n), (2/n), ..., 1]. Use the output of `ecdf` to compute those breakpoints, and then "connect the dots" to define the piecewise linear function.

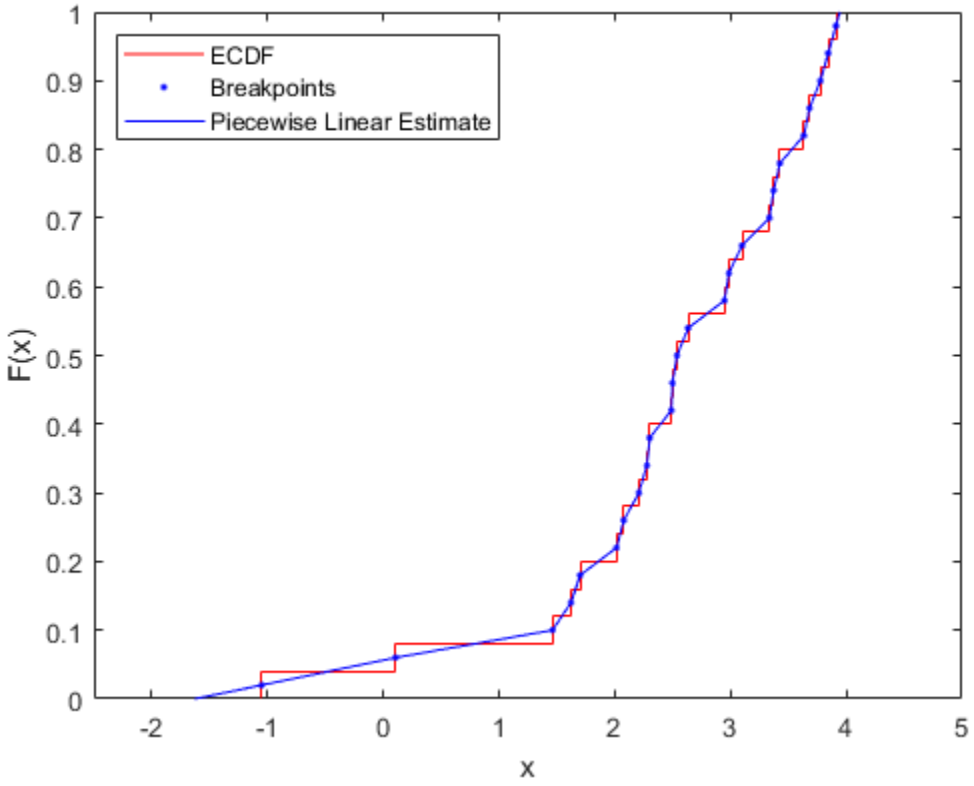
```
xj = xi(2:end);
Fj = (Fi(1:end-1)+Fi(2:end))/2;
hold on
plot(xj,Fj,'b.', xj,Fj,'b-');
hold off
legend({'ECDF' 'Breakpoints' 'Piecewise Linear Estimate'},'location','NW');
```



Because `ecdf` deals appropriately with repeated values and censoring, this calculation works even in cases with more complicated data than in this example.

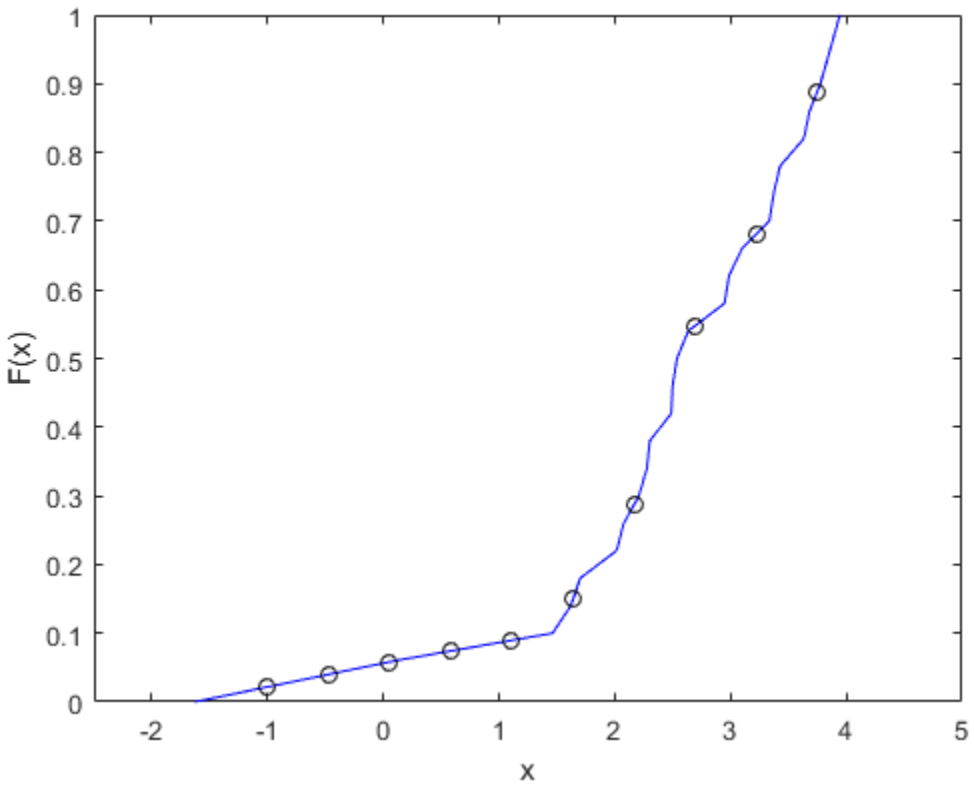
Since the smallest data point corresponds to a height of 1/2n, and the largest to 1-1/2n, the first and last linear segments must be extended beyond the data, to make the function reach 0 and 1.

```
xj = [xj(1)-Fj(1)*(xj(2)-xj(1))/((Fj(2)-Fj(1)));
      xj;
      xj(n)+(1-Fj(n))*((xj(n)-xj(n-1))/(Fj(n)-Fj(n-1)))];
Fj = [0; Fj; 1];
hold on
plot(xj,Fj,'b-','HandleVisibility','off');
hold off
```



This piecewise linear function provides a nonparametric estimate of the CDF that is continuous and symmetric. Evaluating it at points other than the original data is just a matter of linear interpolation, and it can be convenient to define an anonymous function to do that.

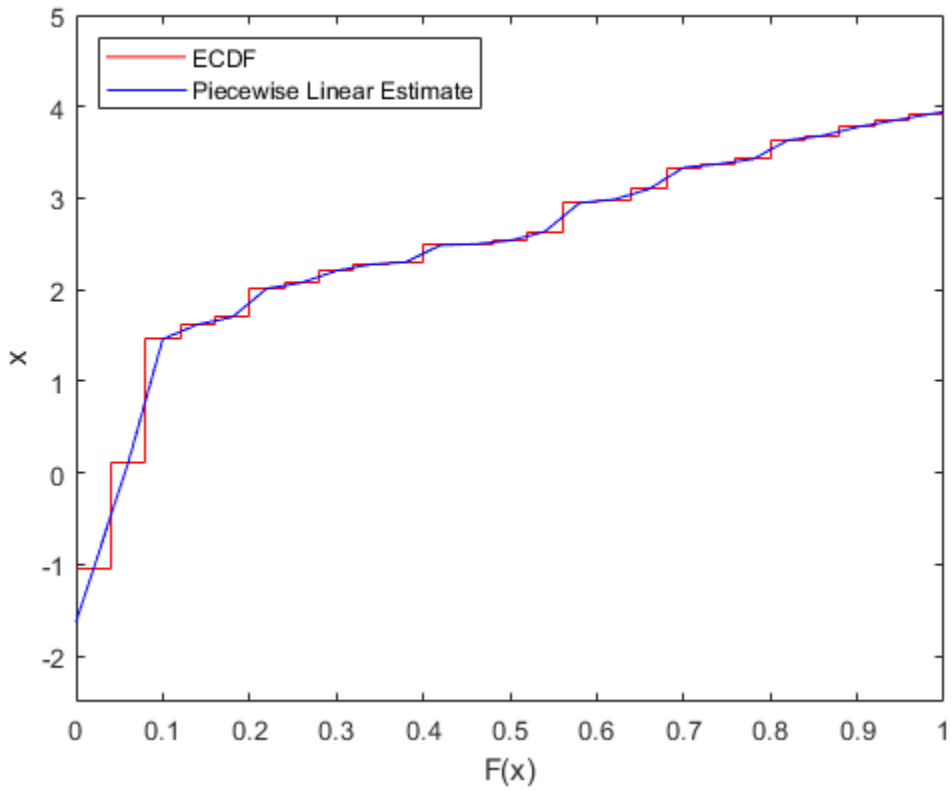
```
F = @(y) interp1(xj,Fj,y,'linear','extrap');
y = linspace(-1,3.75,10);
plot(xj,Fj,'b-',y,F(y),'ko');
xlim([-2.5 5]); xlabel('x'); ylabel('F(x)');
```



A Piecewise Linear Nonparametric Inverse CDF Estimate

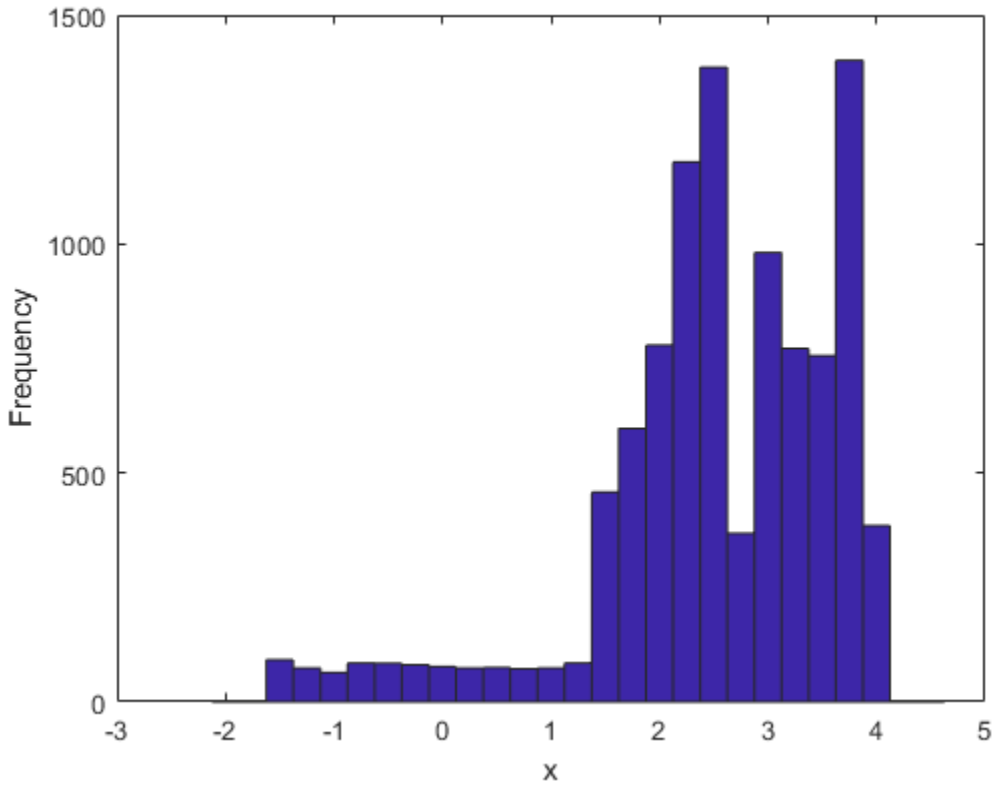
You can use the same calculations to compute a nonparametric estimate of the inverse CDF. In fact, the inverse CDF estimate is just the CDF estimate with the axes swapped.

```
stairs(Fi,[xi(2:end); xi(end)],'r');
hold on
plot(Fj,xj,'b-');
hold off
ylim([-2.5 5]); ylabel('x'); xlabel('F(x)');
legend({'ECDF' 'Piecewise Linear Estimate'},'location','NW');
```



Evaluating this nonparametric inverse CDF at points other than the original breakpoints is again just a matter of linear interpolation. For example, generate uniform random values and use the CDF estimate to transform them back to the scale of your original observed data. This is the inversion method.

```
Finv = @(u) interp1(Fj,xj,u,'linear','extrap');
u = rand(10000,1);
hist(Finv(u),-2:.25:4.5);
xlabel('x'); ylabel('Frequency');
```



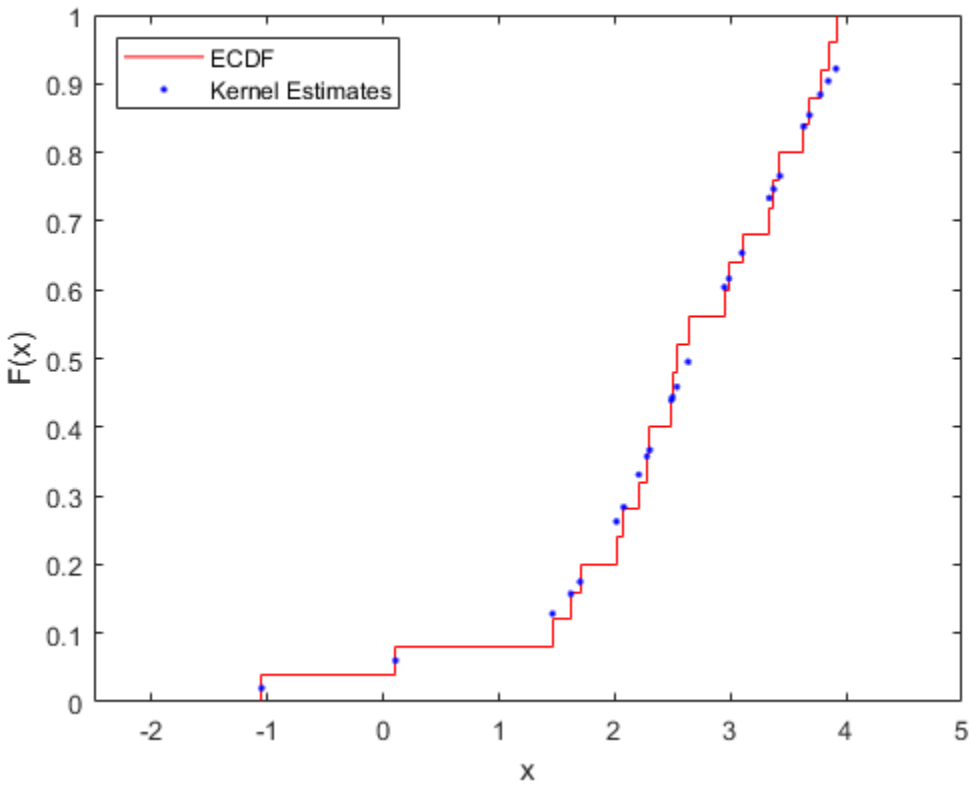
Notice that this histogram of simulated data is more spread out than the histogram of the original data. This is due, in part, to the much larger sample size--the original data consist of only 25 values. But it is also because the piecewise linear CDF estimate, in effect, "spreads out" each of the original observations over an interval, and more so in regions where the individual observations are well-separated.

For example, the two individual observations to the left of zero correspond to a wide, flat region of low density in the simulated data. In contrast, in regions where the data are closely spaced, towards the right tail, for example, the piecewise linear CDF estimate "spreads out" the observations to a lesser extent. In that sense, the method performs a simple version of what is known as variable bandwidth smoothing. However, despite the smoothing, the simulated data retain most of the idiosyncrasies of the original data, i.e., the regions of high and low density.

Kernel Estimators for the CDF and Inverse CDF

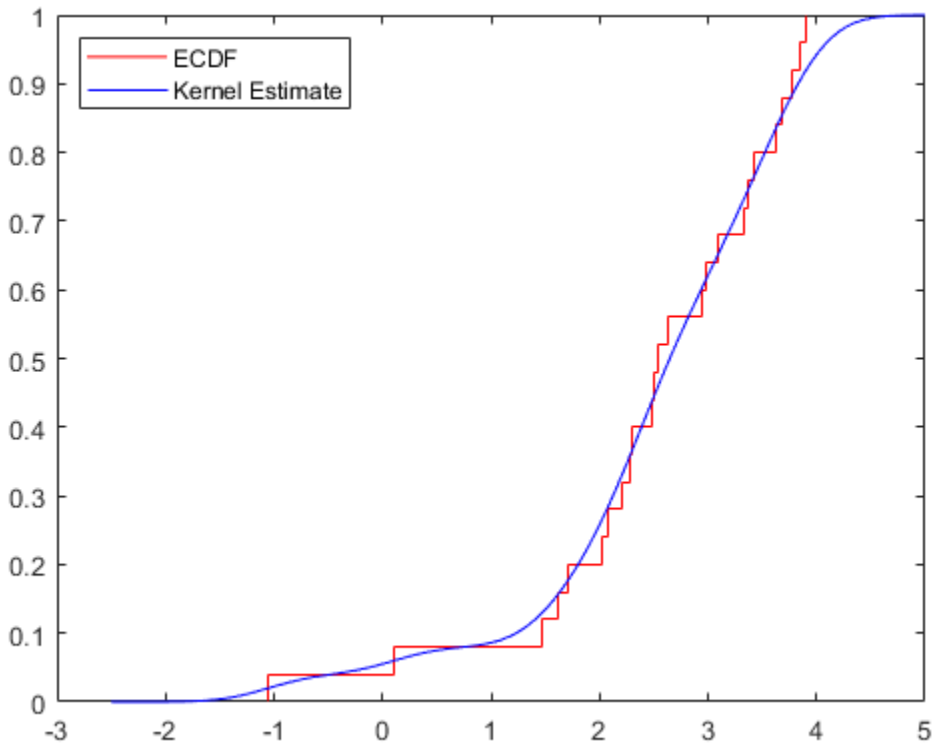
Instead of estimating the CDF using a piecewise linear function, you can perform kernel estimation using the `ksdensity` function to make a smooth nonparametric estimate. Though it is often used to make a nonparametric *density* estimate, `ksdensity` can also estimate other functions. For example, to transform your original data to the unit interval, use it to estimate the CDF.

```
F = ksdensity(x, x, 'function','cdf', 'width',.35);
stairs(xi,Fi,'r');
hold on
plot(x,F,'b.');
hold off
xlim([-2.5 5]); xlabel('x'); ylabel('F(x)');
legend({'ECDF' 'Kernel Estimates'},'location','NW');
```



`ksdensity` also provides a convenient way to evaluate the kernel CDF estimate at points other than the original data. For example, plot the estimate as a smooth curve.

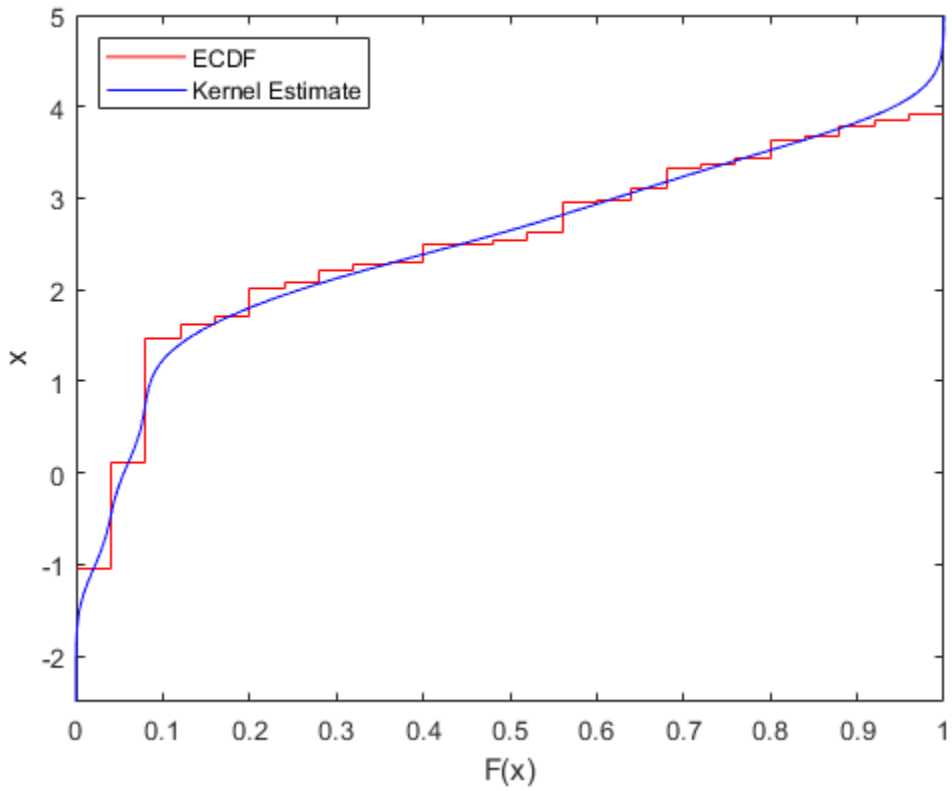
```
y = linspace(-2.5,5,1000);
Fy = ksdensity(x, y, 'function','cdf', 'width',.35);
stairs(xi,Fi,'r');
hold on
plot(y,Fy,'b-');
hold off
legend({'ECDF' 'Kernel Estimate'},'location','NW');
```



`ksdensity` uses a bandwidth parameter to control the amount of smoothing in the estimates it computes, and it is possible to let `ksdensity` choose a default value. The examples here use a fairly small bandwidth to limit the amount of smoothing. Even so, the kernel estimate does not follow the ECDF as closely as the piecewise linear estimate does.

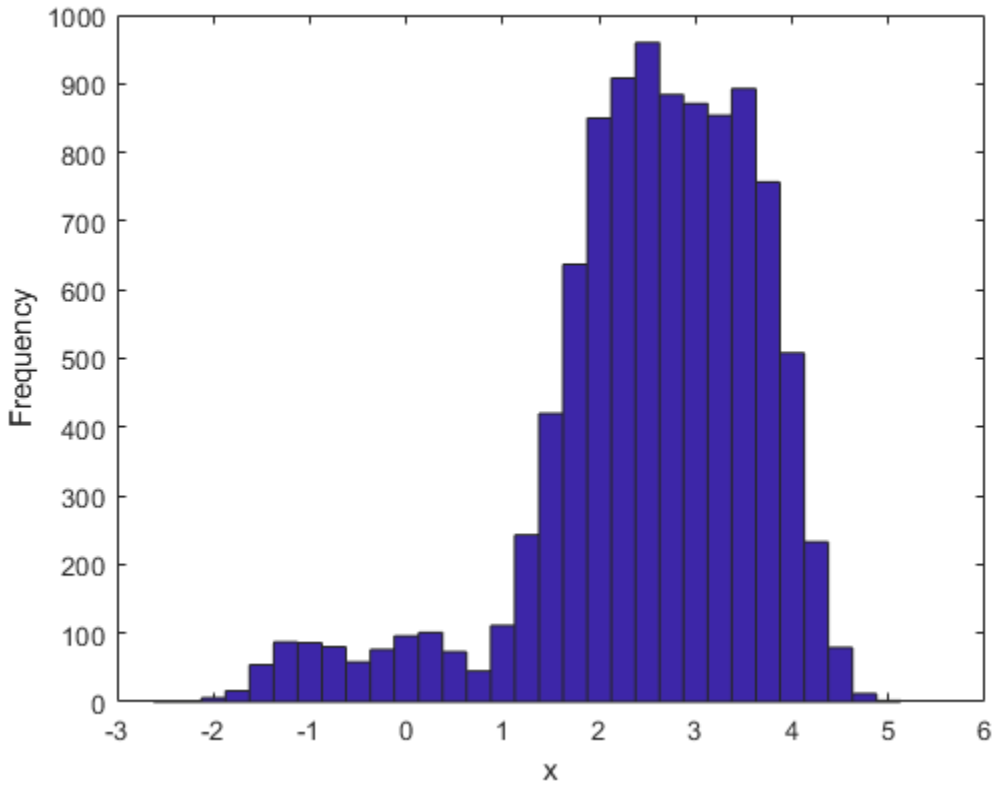
One way to estimate the inverse CDF using kernel estimation is to compute the kernel CDF estimate on a grid of points spanning the range of the original data, and then use the same procedure as for the piecewise linear estimate. For example, to plot the inverse CDF kernel estimate as a smooth curve, simply swap the axes.

```
stairs(Fi,[xi(2:end); xi(end)],'r');
hold on
plot(Fy,y,'b-');
hold off
ylim([-2.5 5]); ylabel('x'); xlabel('F(x)');
legend({'ECDF' 'Kernel Estimate'},'location','NW');
```



To transform uniform random values back to the scale of the original data, interpolate using the grid of CDF estimates.

```
Finv = @(u) interp1(Fy,y,u,'linear','extrap');
hist(Finv(u),-2.5:.25:5);
xlabel('x'); ylabel('Frequency');
```



Notice that the simulated data using the kernel CDF estimate has not completely "smoothed over" the two individual observations to the left of zero present in the original data. The kernel estimate uses a fixed bandwidth. With the particular bandwidth value used in this example, those two observations contribute to two localized areas of density, rather than a wide flat region as was the case with the piecewise linear estimate. In contrast, the kernel estimate has smoothed the data *more* in the right tail than the piecewise linear estimate.

Another way to generate simulated data using kernel estimation is to use `ksdensity` to compute an estimate of the inverse CDF directly, again using the `'function'` parameter. For example, transform those same uniform values.

```
r = ksdensity(x, u, 'function','icdf', 'width',.35);
```

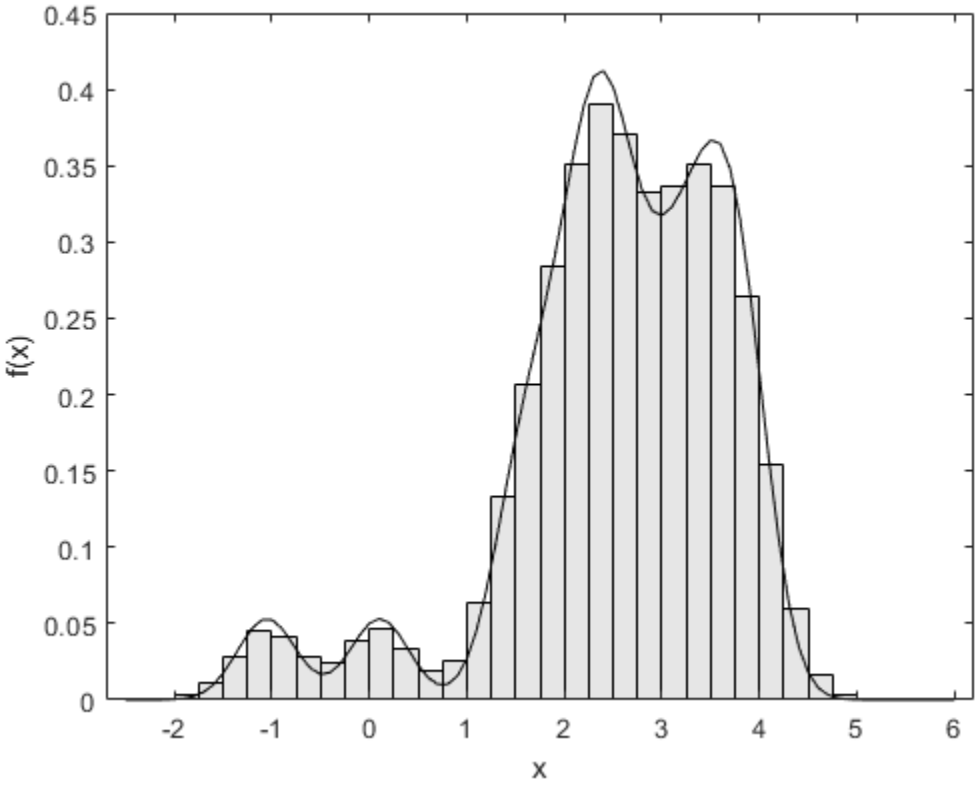
However, using the latter method can be time-consuming for large amounts of data. A simpler, but equivalent, method is to resample with replacement from the original data and add some appropriate random noise.

```
r = datasample(x,100000,'replace',true) + normrnd(0,.35,100000,1);
```

If you generate enough random values, a histogram of the result follows the kernel density estimate of the original data very closely.

```
binwidth = .25;
edges = -2.5:binwidth:6;
ctrs = edges(1:end-1) + binwidth./2;
counts = histc(r,edges); counts = counts(1:end-1);
bar(ctrs,counts./(sum(counts).*binwidth),1,'FaceColor',[.9 .9 .9]);
```

```
hold on
xgrid = edges(1):.1:edges(end);
fgrid = ksdensity(x, xgrid, 'function','pdf', 'width',.3);
plot(xgrid,fgrid,'k-');
hold off
xlabel('x'); ylabel('f(x)');
```



A Semiparametric CDF Estimate

A nonparametric CDF estimate requires a good deal of data to achieve reasonable precision. In addition, data only affect the estimate "locally." That is, in regions where there is a high density of data, the estimate is based on more observations than in regions where there is a low density of data. In particular, nonparametric estimates do not perform well in the tails of a distribution, where data are sparse by definition.

Fitting a semiparametric model to your data using the `paretotails` function allows the best of both the nonparametric and parametric worlds. In the "center" of the distribution, the model uses the piecewise linear nonparametric estimate for the CDF. In each tail, it uses a generalized Pareto distribution. The generalized Pareto is often used as a model for the tail(s) of a dataset, and while it is flexible enough to fit a wide variety of distribution tails, it is sufficiently constrained so that it requires few data to provide a plausible and smooth fit to tail data.

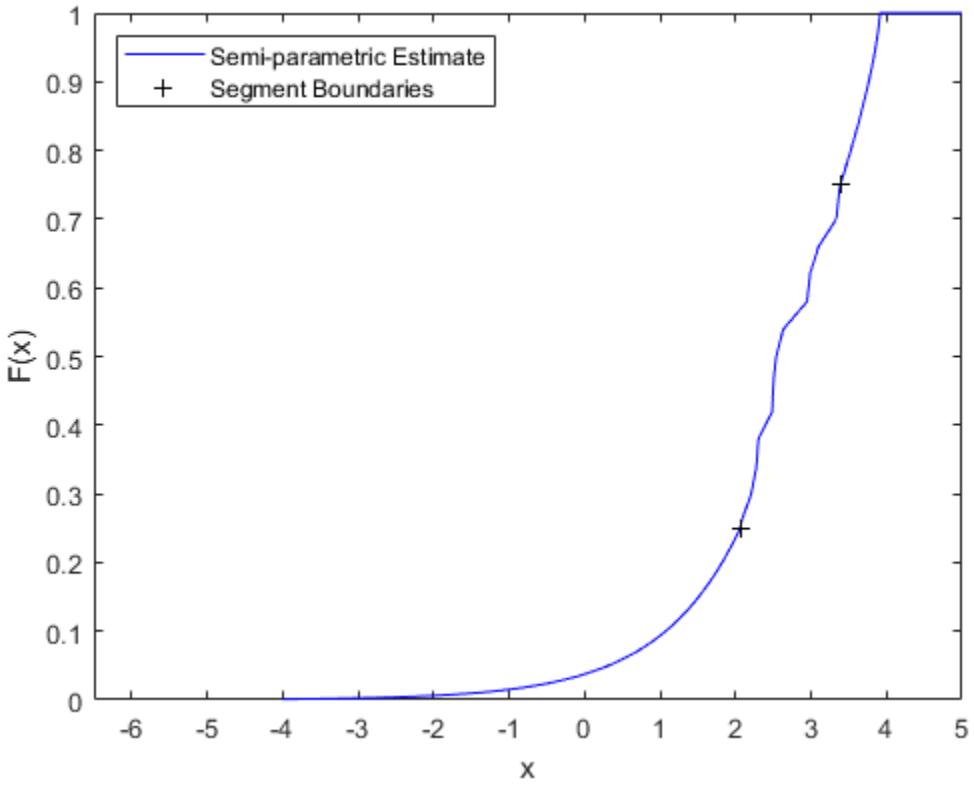
For example, you might define the "center" of the data as the middle 50%, and specify that the transitions between the nonparametric estimate and the Pareto fits take place at the .25 and .75 quantiles of your data. To evaluate the CDF of the semiparametric model fit, use the fit's `cdf` method.

```
semipFit = paretotails(x,.25,.75);
```

```
Warning: Problem fitting generalized Pareto distribution to upper tail. Maximum
likelihood has converged to a boundary point of the parameter space.
```

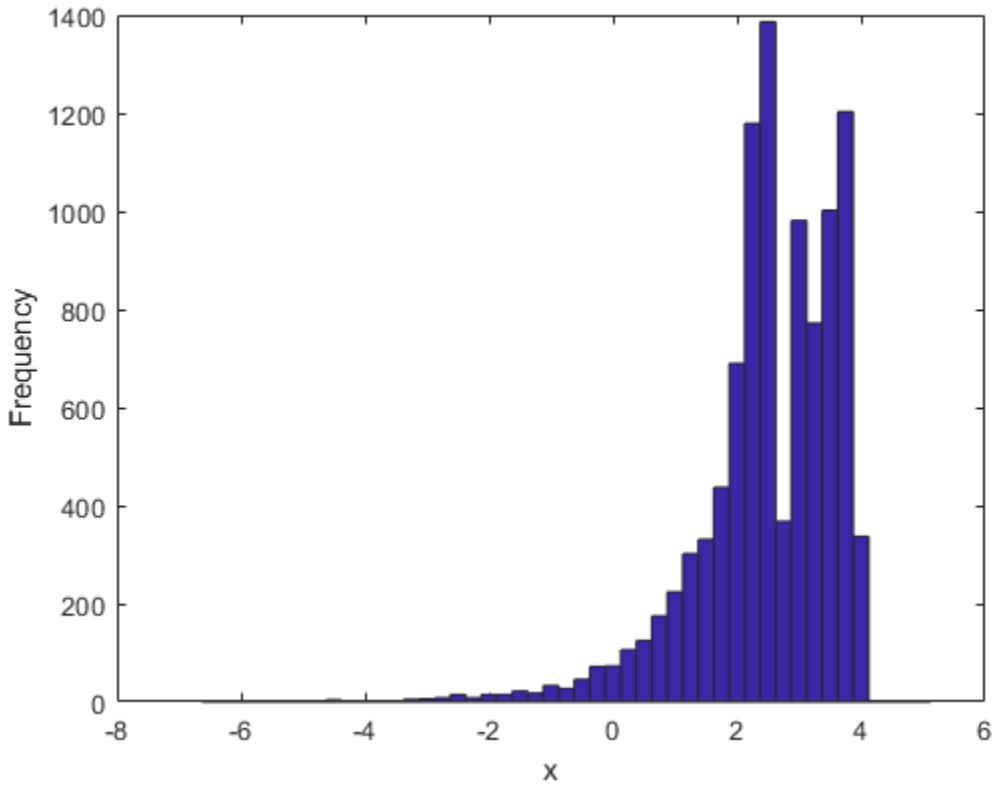
The warning is due to using so few data -- 6 points in each tail in this case -- and indicates that the fitted generalized Pareto distribution in the upper tail extends exactly to the smallest observation, and no further. You can see that in the histogram shown below. In a real application, you would usually have more data, and the warning would typically not occur.

```
[p,q] = boundary(semipFit);
y = linspace(-4,6,1000);
Fy = cdf(semipFit,y);
plot(y,Fy,'b-', q,p,'k+');
xlim([-6.5 5]); xlabel('x'); ylabel('F(x)');
legend({'Semi-parametric Estimate' 'Segment Boundaries'},'location','NW');
```



To transform uniform random values back to the scale of your original data, use the fit's `icdf` method.

```
r = icdf(semipFit,u);
hist(r,-6.5:.25:5);
xlabel('x'); ylabel('Frequency');
```



This semiparametric estimate has smoothed the tails of the data more than the center, because of the parametric model used in the tails. In that sense, the estimate is more similar to the piecewise linear estimate than to the kernel estimate. However, it is also possible to use `paretotails` to create a semiparametric fit that uses kernel estimation in the center of the data.

Conclusions

This example illustrates three methods for computing a nonparametric or semiparametric CDF or inverse CDF estimate from data. The three methods impose different amounts and types of smoothing on the data. Which method you choose depends on how each method captures or fails to capture what you consider the important features of your data.

Modelling Tail Data with the Generalized Pareto Distribution

This example shows how to fit tail data to the Generalized Pareto distribution by maximum likelihood estimation.

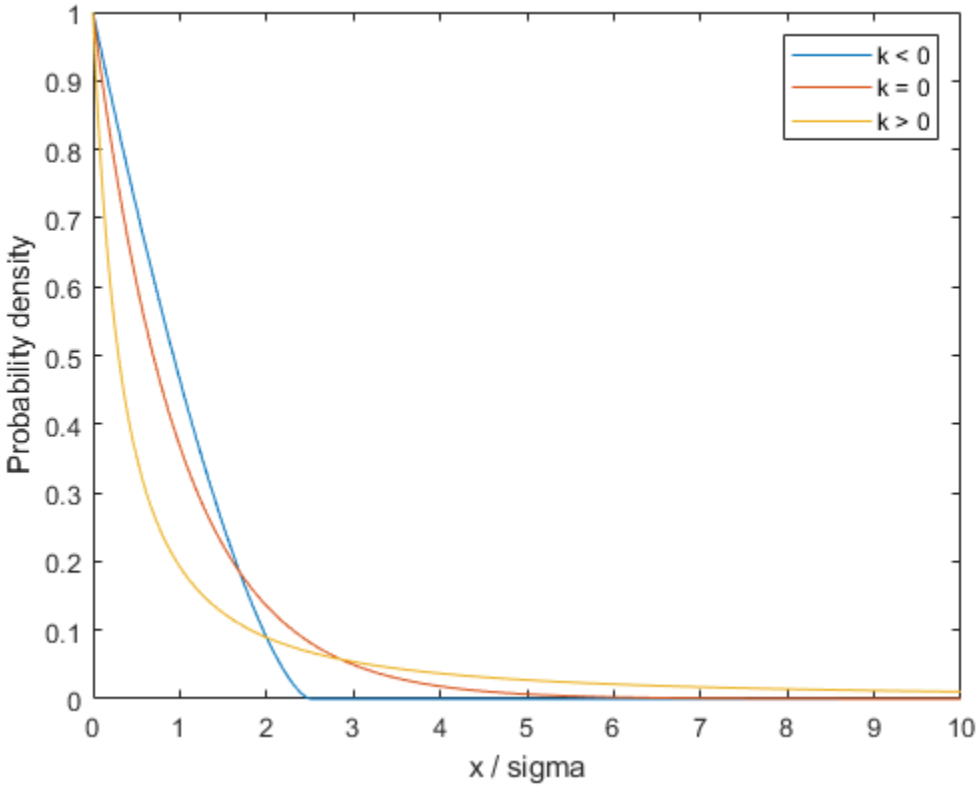
Fitting a parametric distribution to data sometimes results in a model that agrees well with the data in high density regions, but poorly in areas of low density. For unimodal distributions, such as the normal or Student's t, these low density regions are known as the "tails" of the distribution. One reason why a model might fit poorly in the tails is that by definition, there are fewer data in the tails on which to base a choice of model, and so models are often chosen based on their ability to fit data near the mode. Another reason might be that the distribution of real data is often more complicated than the usual parametric models.

However, in many applications, fitting the data in the tail is the main concern. The Generalized Pareto distribution (GP) was developed as a distribution that can model tails of a wide variety of distributions, based on theoretical arguments. One approach to distribution fitting that involves the GP is to use a non-parametric fit (the empirical cumulative distribution function, for example) in regions where there are many observations, and to fit the GP to the tail(s) of the data.

The Generalized Pareto Distribution

The Generalized Pareto (GP) is a right-skewed distribution, parameterized with a shape parameter, k, and a scale parameter, sigma. k is also known as the "tail index" parameter, and can be positive, zero, or negative.

```
x = linspace(0,10,1000);
plot(x,gppdf(x,-.4,1),'-', x,gppdf(x,0,1),'-', x,gppdf(x,2,1),'-');
xlabel('x / sigma');
ylabel('Probability density');
legend({'k < 0' 'k = 0' 'k > 0'});
```



Notice that for k < 0, the GP has zero probability above an upper limit of -(1/k). For k >= 0, the GP has no upper limit. Also, the GP is often used in conjunction with a third, threshold parameter that shifts the lower limit away from zero. We will not need that generality here.

The GP distribution is a generalization of both the exponential distribution (k = 0) and the Pareto distribution (k > 0). The GP includes those two distributions in a larger family so that a continuous range of shapes is possible.

Simulating Exceedance Data

The GP distribution can be defined constructively in terms of exceedances. Starting with a probability distribution whose right tail drops off to zero, such as the normal, we can sample random values independently from that distribution. If we fix a threshold value, throw out all the values that are below the threshold, and subtract the threshold off of the values that are not thrown out, the result is known as exceedances. The distribution of the exceedances is approximately a GP. Similarly, we can set a threshold in the left tail of a distribution, and ignore all values above that threshold. The threshold must be far enough out in the tail of the original distribution for the approximation to be reasonable.

The original distribution determines the shape parameter, k, of the resulting GP distribution. Distributions whose tails fall off as a polynomial, such as Student's t, lead to a positive shape parameter. Distributions whose tails decrease exponentially, such as the normal, correspond to a zero shape parameter. Distributions with finite tails, such as the beta, correspond to a negative shape parameter.

Real-world applications for the GP distribution include modelling extremes of stock market returns, and modelling extreme floods. For this example, we'll use simulated data, generated from a Student's t distribution with 5 degrees of freedom. We'll take the largest 5% of 2000 observations from the t distribution, and then subtract off the 95% quantile to get exceedances.

```
rng(3,'twister');
x = trnd(5,2000,1);
q = quantile(x,.95);
y = x(x>q) - q;
n = numel(y)
```

```
n =

   100
```

Fitting the Distribution Using Maximum Likelihood

The GP distribution is defined for 0 < sigma, and -Inf < k < Inf. However, interpretation of the results of maximum likelihood estimation is problematic when k < -1/2. Fortunately, those cases correspond to fitting tails from distributions like the beta or triangular, and so will not present a problem here.

```
paramEsts = gpfit(y);
kHat      = paramEsts(1)   % Tail index parameter
sigmaHat  = paramEsts(2)   % Scale parameter
```

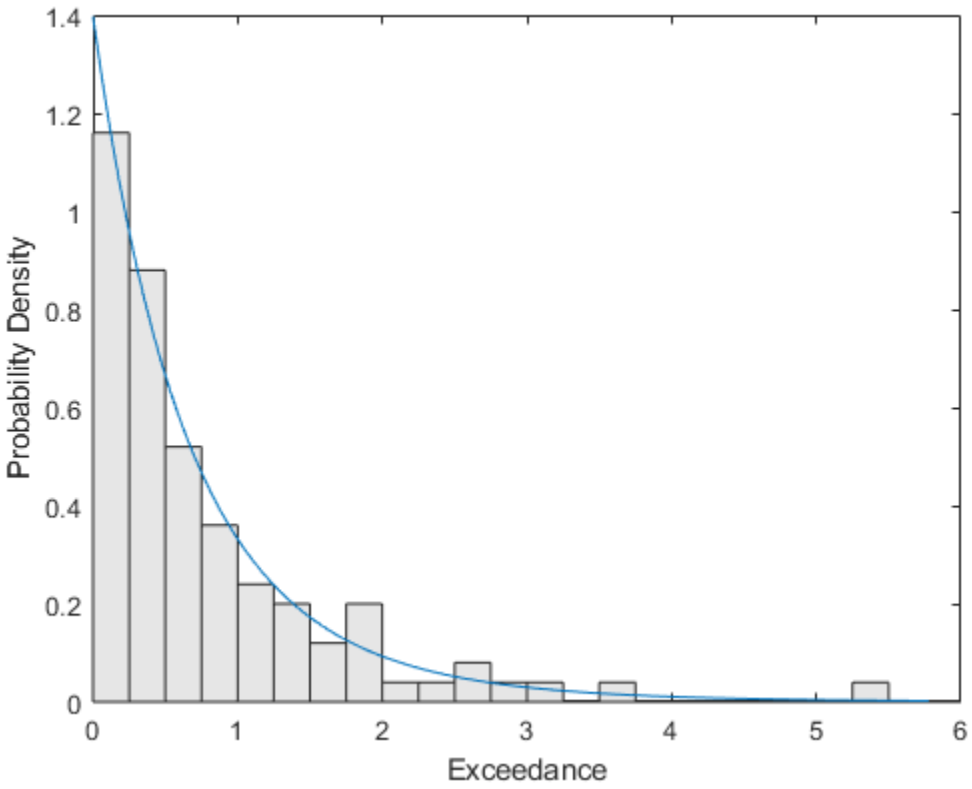
```
kHat =

    0.0987


sigmaHat =

    0.7156
```

As might be expected, since the simulated data were generated using a t distribution, the estimate of k is positive.

Checking the Fit Visually

To visually assess how good the fit is, we'll plot a scaled histogram of the tail data, overlaid with the density function of the GP that we've estimated. The histogram is scaled so that the bar heights times their width sum to 1.

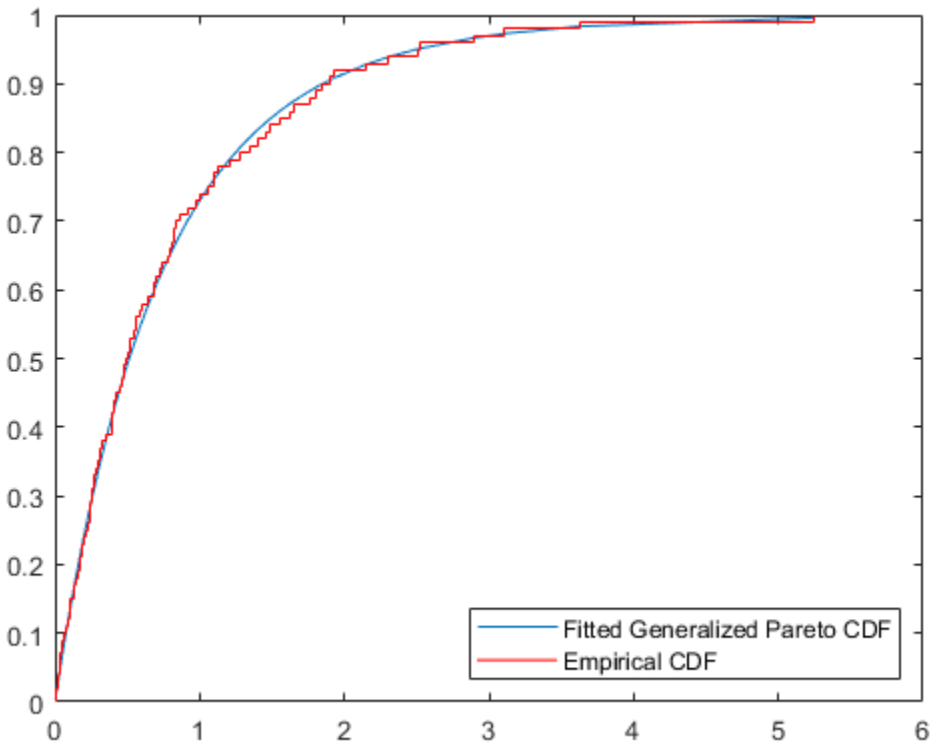
```
bins = 0:.25:7;
h = bar(bins,histc(y,bins)/(length(y)*.25),'histc');
h.FaceColor = [.9 .9 .9];
ygrid = linspace(0,1.1*max(y),100);
line(ygrid,gppdf(ygrid,kHat,sigmaHat));
xlim([0,6]);
xlabel('Exceedance');
ylabel('Probability Density');
```



We've used a fairly small bin width, so there is a good deal of noise in the histogram. Even so, the fitted density follows the shape of the data, and so the GP model seems to be a good choice.

We can also compare the empirical CDF to the fitted CDF.

```
[F,yi] = ecdf(y);
plot(yi,gpcdf(yi,kHat,sigmaHat),'-');
hold on;
stairs(yi,F,'r');
hold off;
legend('Fitted Generalized Pareto CDF','Empirical CDF','location','southeast');
```



Computing Standard Errors for the Parameter Estimates

To quantify the precision of the estimates, we'll use standard errors computed from the asymptotic covariance matrix of the maximum likelihood estimators. The function `gplike` computes, as its second output, a numerical approximation to that covariance matrix. Alternatively, we could have called `gpfit` with two output arguments, and it would have returned confidence intervals for the parameters.

```
[nll,acov] = gplike(paramEsts, y);
stdErr = sqrt(diag(acov))
```

```
stdErr =

    0.1158
    0.1093
```

These standard errors indicate that the relative precision of the estimate for k is quite a bit lower than that for sigma -- its standard error is on the order of the estimate itself. Shape parameters are often difficult to estimate. It's important to keep in mind that computation of these standard errors assumed that the GP model is correct, and that we have enough data for the asymptotic approximation to the covariance matrix to hold.

Checking the Asymptotic Normality Assumption

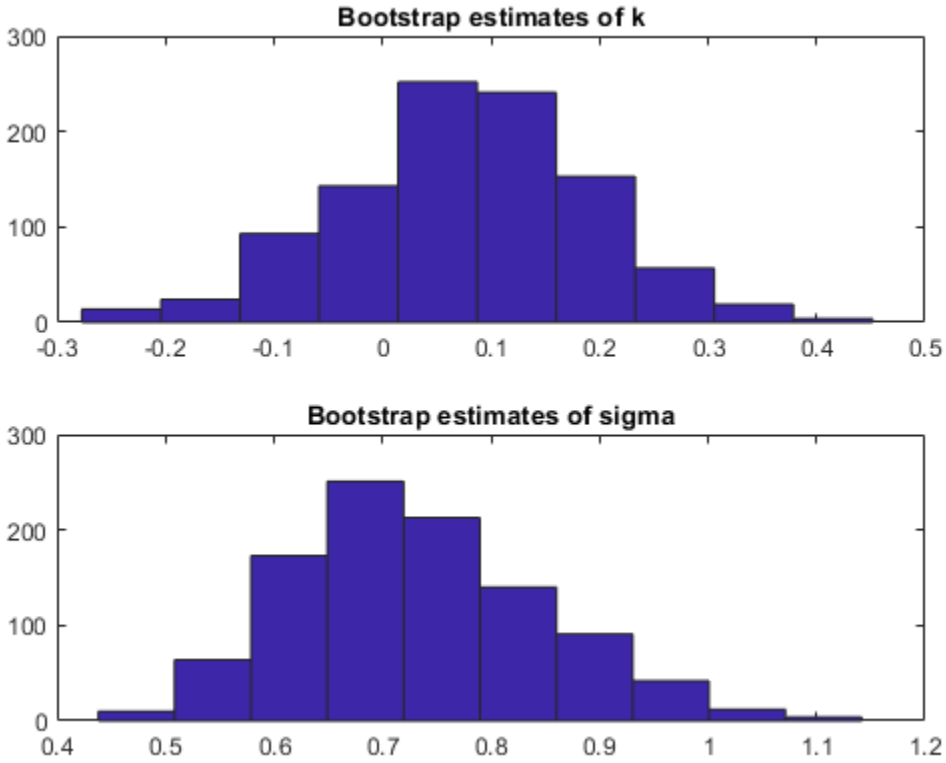
Interpretation of the standard errors usually involves assuming that, if the same fit could be repeated many times on data that came from the same source, the maximum likelihood estimates of the parameters would approximately follow a normal distribution. For example, confidence intervals are often based this assumption.

However, that normal approximation may or may not be a good one. To assess how good it is in this example, we can use a bootstrap simulation. We will generate 1000 replicate datasets by resampling from the data, fit a GP distribution to each one, and save all the replicate estimates.

```
replEsts = bootstrp(1000,@gpfit,y);
```

As a rough check on the sampling distribution of the parameter estimators, we can look at histograms of the bootstrap replicates.

```
subplot(2,1,1);
hist(replEsts(:,1));
title('Bootstrap estimates of k');
subplot(2,1,2);
hist(replEsts(:,2));
title('Bootstrap estimates of sigma');
```

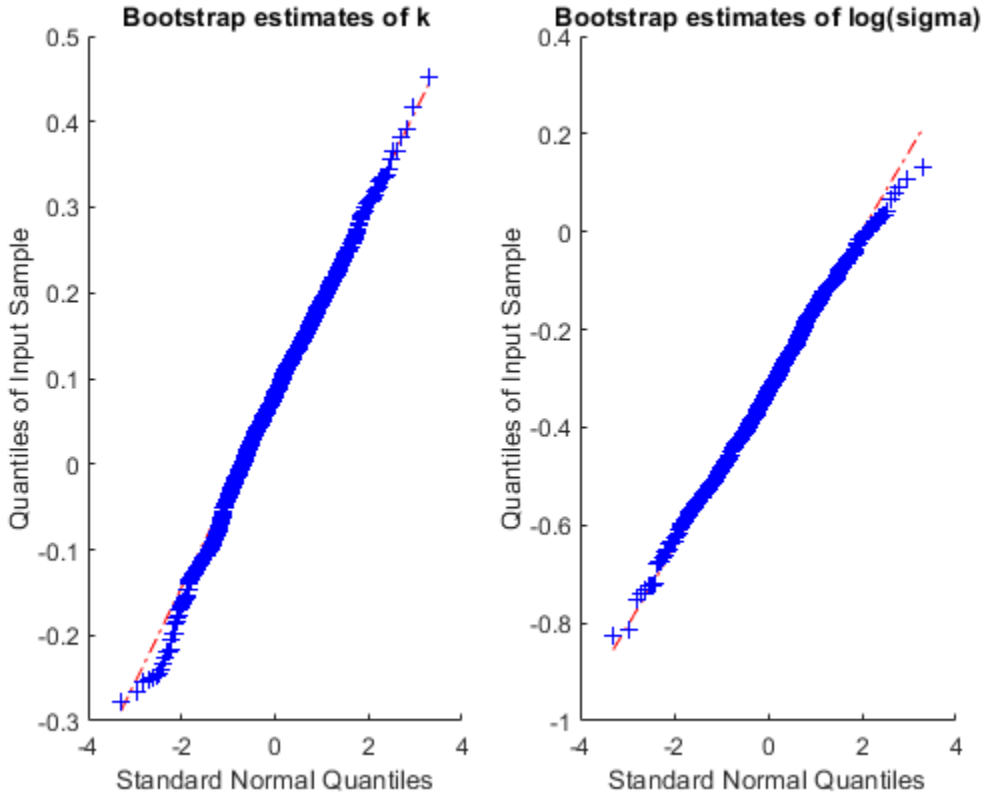


Using a Parameter Transformation

The histogram of the bootstrap estimates for k appears to be only a little asymmetric, while that for the estimates of sigma definitely appears skewed to the right. A common remedy for that skewness is

to estimate the parameter and its standard error on the log scale, where a normal approximation may be more reasonable. A Q-Q plot is a better way to assess normality than a histogram, because non-normality shows up as points that do not approximately follow a straight line. Let's check that to see if the log transform for sigma is appropriate.

```
subplot(1,2,1);
qqplot(replEsts(:,1));
title('Bootstrap estimates of k');
subplot(1,2,2);
qqplot(log(replEsts(:,2)));
title('Bootstrap estimates of log(sigma)');
```



The bootstrap estimates for k and log(sigma) appear acceptably close to normality. A Q-Q plot for the estimates of sigma, on the unlogged scale, would confirm the skewness that we've already seen in the histogram. Thus, it would be more reasonable to construct a confidence interval for sigma by first computing one for log(sigma) under the assumption of normality, and then exponentiating to transform that interval back to the original scale for sigma.

In fact, that's exactly what the function `gpfit` does behind the scenes.

```
[paramEsts,paramCI] = gpfit(y);

kHat
kCI  = paramCI(:,1)
```

```
kHat =
```

```
    0.0987


kCI =

   -0.1283
    0.3258


sigmaHat
sigmaCI  = paramCI(:,2)


sigmaHat =

    0.7156


sigmaCI =

    0.5305
    0.9654
```

Notice that while the 95% confidence interval for k is symmetric about the maximum likelihood estimate, the confidence interval for sigma is not. That's because it was created by transforming a symmetric CI for log(sigma).

Modelling Data with the Generalized Extreme Value Distribution

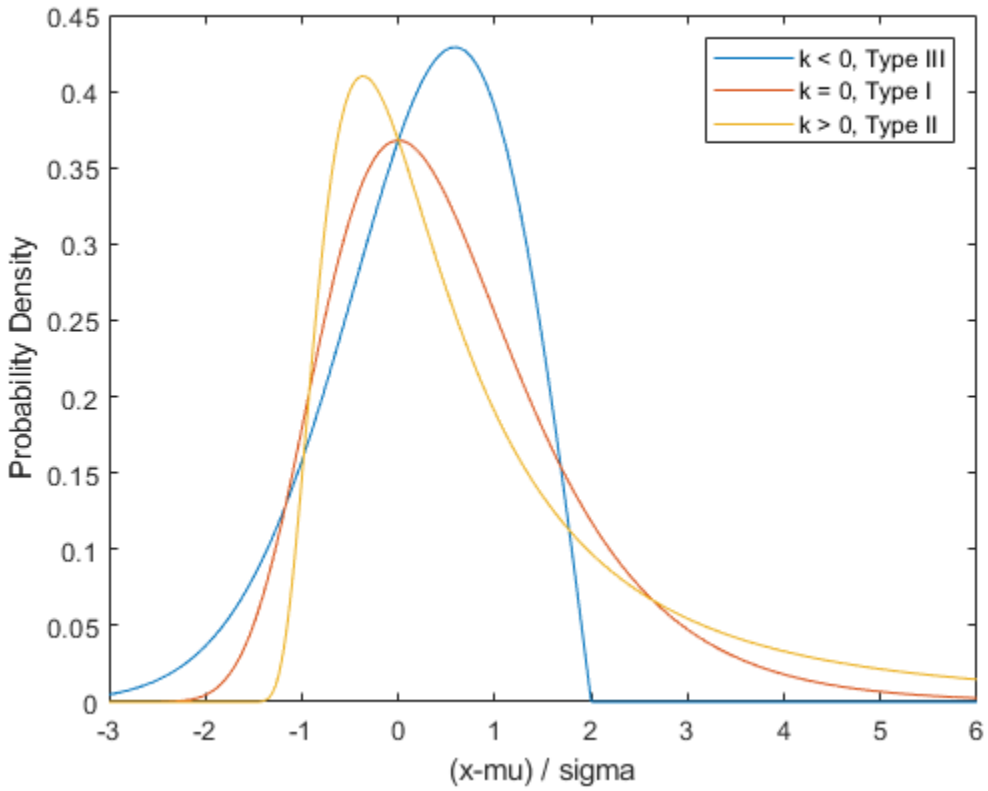
This example shows how to fit the generalized extreme value distribution using maximum likelihood estimation. The extreme value distribution is used to model the largest or smallest value from a group or block of data.

Three types of extreme value distributions are common, each as the limiting case for different types of underlying distributions. For example, the type I extreme value is the limit distribution of the maximum (or minimum) of a block of normally distributed data, as the block size becomes large. In this example, we will illustrate how to fit such data using a single distribution that includes all three types of extreme value distributions as special case, and investigate likelihood-based confidence intervals for quantiles of the fitted distribution.

The Generalized Extreme Value Distribution

The Generalized Extreme Value (GEV) distribution unites the type I, type II, and type III extreme value distributions into a single family, to allow a continuous range of possible shapes. It is parameterized with location and scale parameters, mu and sigma, and a shape parameter, k. When k < 0, the GEV is equivalent to the type III extreme value. When k > 0, the GEV is equivalent to the type II. In the limit as k approaches 0, the GEV becomes the type I.

```
x = linspace(-3,6,1000);
plot(x,gevpdf(x,-.5,1,0),'-', x,gevpdf(x,0,1,0),'-', x,gevpdf(x,.5,1,0),'-');
xlabel('(x-mu) / sigma');
ylabel('Probability Density');
legend({'k < 0, Type III' 'k = 0, Type I' 'k > 0, Type II'});
```



Notice that for k < 0 or k > 0, the density has zero probability above or below, respectively, the upper or lower bound -(1/k). In the limit as k approaches 0, the GEV is unbounded. This can be summarized as the constraint that 1+k*(y-mu)/sigma must be positive.

Simulating Block Maximum Data

The GEV can be defined constructively as the limiting distribution of block maxima (or minima). That is, if you generate a large number of independent random values from a single probability distribution, and take their maximum value, the distribution of that maximum is approximately a GEV.

The original distribution determines the shape parameter, k, of the resulting GEV distribution. Distributions whose tails fall off as a polynomial, such as Student's t, lead to a positive shape parameter. Distributions whose tails decrease exponentially, such as the normal, correspond to a zero shape parameter. Distributions with finite tails, such as the beta, correspond to a negative shape parameter.

Real applications for the GEV might include modelling the largest return for a stock during each month. Here, we will simulate data by taking the maximum of 25 values from a Student's t distribution with two degrees of freedom. The simulated data will include 75 random block maximum values.

```
rng(0,'twister');
y = max(trnd(2,25,75),[],1);
```

Fitting the Distribution by Maximum Likelihood

The function `gevfit` returns both maximum likelihood parameter estimates, and (by default) 95% confidence intervals.

```
[paramEsts,paramCIs] = gevfit(y);

kMLE = paramEsts(1)        % Shape parameter
sigmaMLE = paramEsts(2)    % Scale parameter
muMLE = paramEsts(3)       % Location parameter
```

```
kMLE =

    0.4901


sigmaMLE =

    1.4856


muMLE =

    2.9710
```

```
kCI = paramCIs(:,1)
sigmaCI = paramCIs(:,2)
muCI = paramCIs(:,3)
```

```
kCI =

    0.2020
    0.7782


sigmaCI =

    1.1431
    1.9307


muCI =

    2.5599
    3.3821
```

Notice that the 95% confidence interval for k does not include the value zero. The type I extreme value distribution is apparently not a good model for these data. That makes sense, because the underlying distribution for the simulation had much heavier tails than a normal, and the type II extreme value distribution is theoretically the correct one as the block size becomes large.

As an alternative to confidence intervals, we can also compute an approximation to the asymptotic covariance matrix of the parameter estimates, and from that extract the parameter standard errors.

```
[nll,acov] = gevlike(paramEsts,y);
paramSEs = sqrt(diag(acov))
```

```
paramSEs =

    0.1470
    0.1986
    0.2097
```

Checking the Fit Visually

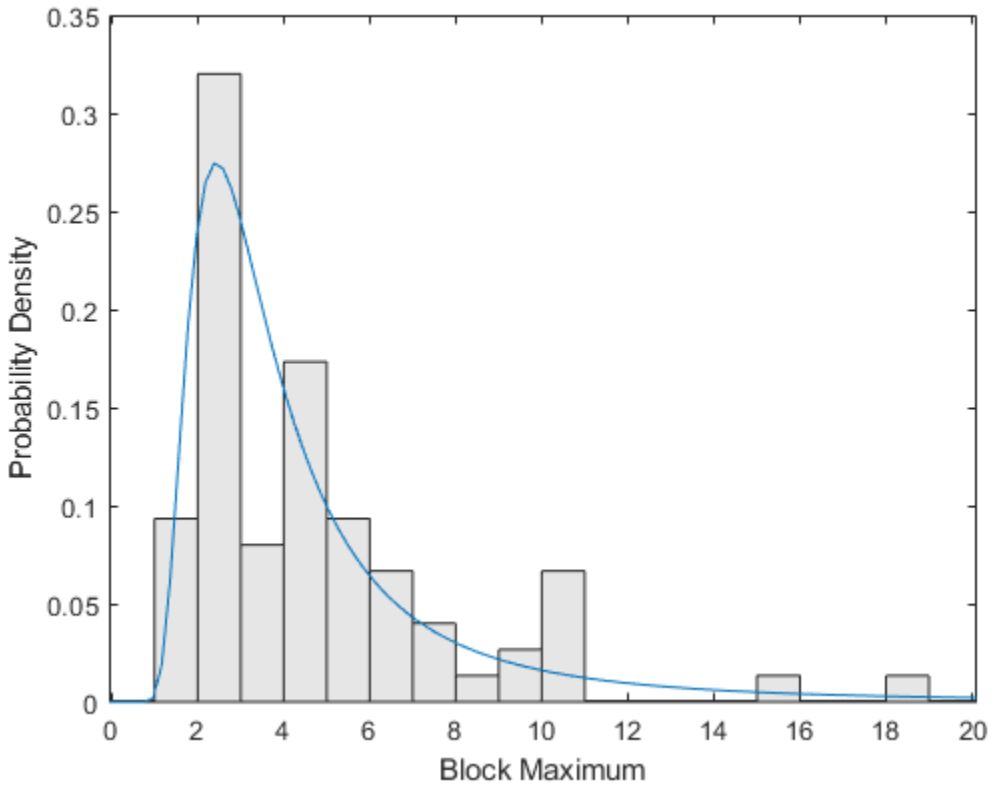
To visually assess how good the fit is, we'll look at plots of the fitted probability density function (PDF) and cumulative distribution function (CDF).

The support of the GEV depends on the parameter values. In this case, the estimate for k is positive, so the fitted distribution has zero probability below a lower bound.

```
lowerBnd = muMLE-sigmaMLE./kMLE;
```

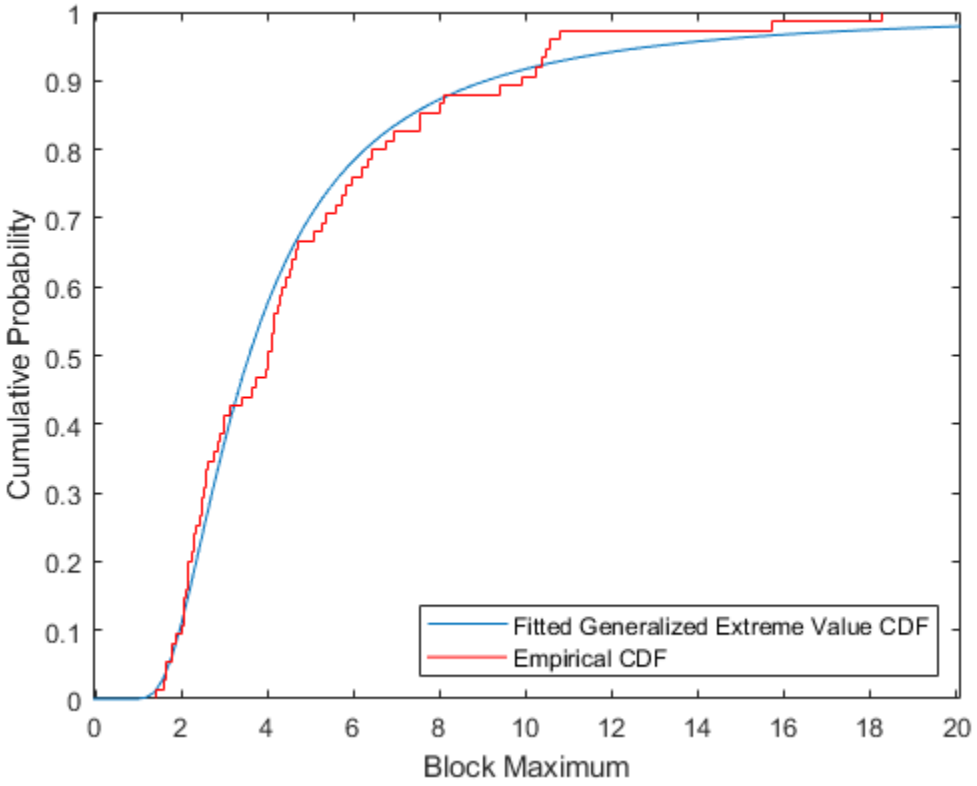
First, we'll plot a scaled histogram of the data, overlaid with the PDF for the fitted GEV model. This histogram is scaled so that the bar heights times their width sum to 1, to make it comparable to the PDF.

```
ymax = 1.1*max(y);
bins = floor(lowerBnd):ceil(ymax);
h = bar(bins,histc(y,bins)/length(y),'histc');
h.FaceColor = [.9 .9 .9];
ygrid = linspace(lowerBnd,ymax,100);
line(ygrid,gevpdf(ygrid,kMLE,sigmaMLE,muMLE));
xlabel('Block Maximum');
ylabel('Probability Density');
xlim([lowerBnd ymax]);
```



We can also compare the fit to the data in terms of cumulative probability, by overlaying the empirical CDF and the fitted CDF.

```
[F,yi] = ecdf(y);
plot(ygrid,gevcdf(ygrid,kMLE,sigmaMLE,muMLE),'-');
hold on;
stairs(yi,F,'r');
hold off;
xlabel('Block Maximum');
ylabel('Cumulative Probability');
legend('Fitted Generalized Extreme Value CDF','Empirical CDF','location','southeast');
xlim([lowerBnd ymax]);
```



Estimating Quantiles of the Model

While the parameter estimates may be important by themselves, a quantile of the fitted GEV model is often the quantity of interest in analyzing block maxima data.

For example, the return level Rm is defined as the block maximum value expected to be exceeded only once in m blocks. That is just the (1-1/m)'th quantile. We can plug the maximum likelihood parameter estimates into the inverse CDF to estimate Rm for m=10.

```
R10MLE = gevinv(1-1./10,kMLE,sigmaMLE,muMLE)
```

```
R10MLE =

    9.0724
```

We could compute confidence limits for R10 using asymptotic approximations, but those may not be valid. Instead, we will use a likelihood-based method to compute confidence limits. This method often produces more accurate results than one based on the estimated covariance matrix of the parameter estimates.

Given any set of values for the parameters mu, sigma, and k, we can compute a log-likelihood -- for example, the MLEs are the parameter values that maximize the GEV log-likelihood. As the parameter values move away from the MLEs, their log-likelihood typically becomes significantly less than the maximum. If we look at the set of parameter values that produce a log-likelihood larger than a specified critical value, this is a complicated region in the parameter space. However, for a suitable

critical value, it is a confidence region for the model parameters. The region contains parameter values that are "compatible with the data". The critical value that determines the region is based on a chi-square approximation, and we'll use 95% as our confidence level. (Note that we will actually work with the negative of the log-likelihood.)

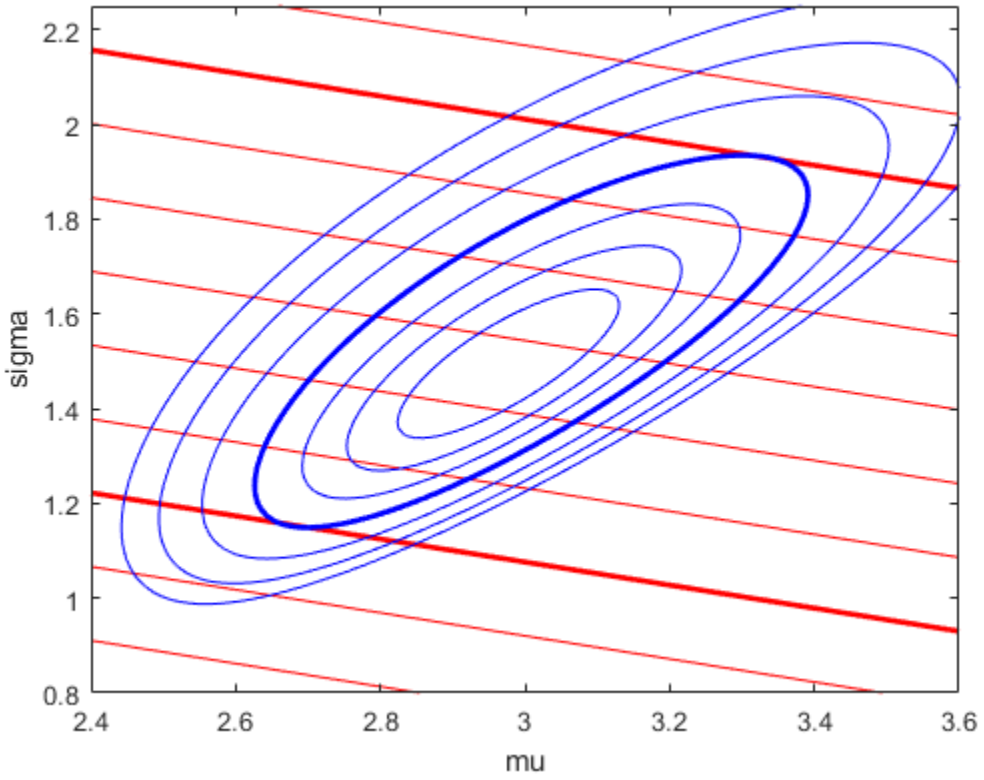
```
nllCritVal = gevlike([kMLE,sigmaMLE,muMLE],y) + .5*chi2inv(.95,1)
```

```
nllCritVal =

  170.3044
```

For any set of parameter values mu, sigma, and k, we can compute R10. Therefore, we can find the smallest R10 value achieved within the critical region of the parameter space where the negative log-likelihood is larger than the critical value. That smallest value is the lower likelihood-based confidence limit for R10.

This is difficult to visualize in all three parameter dimensions, but as a thought experiment, we can fix the shape parameter, k, we can see how the procedure would work over the two remaining parameters, sigma and mu.

```
sigmaGrid = linspace(.8, 2.25, 110);
muGrid = linspace(2.4, 3.6);
nllGrid = zeros(length(sigmaGrid),length(muGrid));
R10Grid = zeros(length(sigmaGrid),length(muGrid));
for i = 1:size(nllGrid,1)
    for j = 1:size(nllGrid,2)
        nllGrid(i,j) = gevlike([kMLE,sigmaGrid(i),muGrid(j)],y);
        R10Grid(i,j) = gevinv(1-1./10,kMLE,sigmaGrid(i),muGrid(j));
    end
end
nllGrid(nllGrid>gevlike([kMLE,sigmaMLE,muMLE],y)+6) = NaN;
contour(muGrid,sigmaGrid,R10Grid,6.14:.64:12.14,'LineColor','r');
hold on
contour(muGrid,sigmaGrid,R10Grid,[7.42 11.26],'LineWidth',2,'LineColor','r');
contour(muGrid,sigmaGrid,nllGrid,[168.7 169.1 169.6 170.3:1:173.3],'LineColor','b');
contour(muGrid,sigmaGrid,nllGrid,[nllCritVal nllCritVal],'LineWidth',2,'LineColor','b');
hold off
axis([2.4 3.6 .8 2.25]);
xlabel('mu');
ylabel('sigma');
```



The blue contours represent the log-likelihood surface, and the bold blue contour is the boundary of the critical region. The red contours represent the surface for R10 -- larger values are to the top right, lower to the bottom left. The contours are straight lines because for fixed k, Rm is a linear function of sigma and mu. The bold red contours are the lowest and highest values of R10 that fall within the critical region. In the full three dimensional parameter space, the log-likelihood contours would be ellipsoidal, and the R10 contours would be surfaces.

Finding the lower confidence limit for R10 is an optimization problem with nonlinear inequality constraints, and so we will use the function `fmincon` from the Optimization Toolbox™. We need to find the smallest R10 value, and therefore the objective to be minimized is R10 itself, equal to the inverse CDF evaluated for p=1-1/m. We'll create a wrapper function that computes Rm specifically for m=10.

```
CIobjfun = @(params) gevinv(1-1./10,params(1),params(2),params(3));
```

To perform the constrained optimization, we'll also need a function that defines the constraint, that is, that the negative log-likelihood be less than the critical value. The constraint function should return positive values when the constraint is violated. We'll create an anonymous function, using the simulated data and the critical log-likelihood value. It also returns an empty value because we're not using any equality constraints here.

```
CIconfun = @(params) deal(gevlike(params,y) - nllCritVal, []);
```

Finally, we call `fmincon`, using the active-set algorithm to perform the constrained optimization.

```
opts = optimset('Algorithm','active-set', 'Display','notify', 'MaxFunEvals',500, ...
                'RelLineSrchBnd',.1, 'RelLineSrchBndDuration',Inf);
```

```
[params,R10Lower,flag,output] = ...
    fmincon(CIobjfun,paramEsts,[],[],[],[],[],[],CIconfun,opts);
```

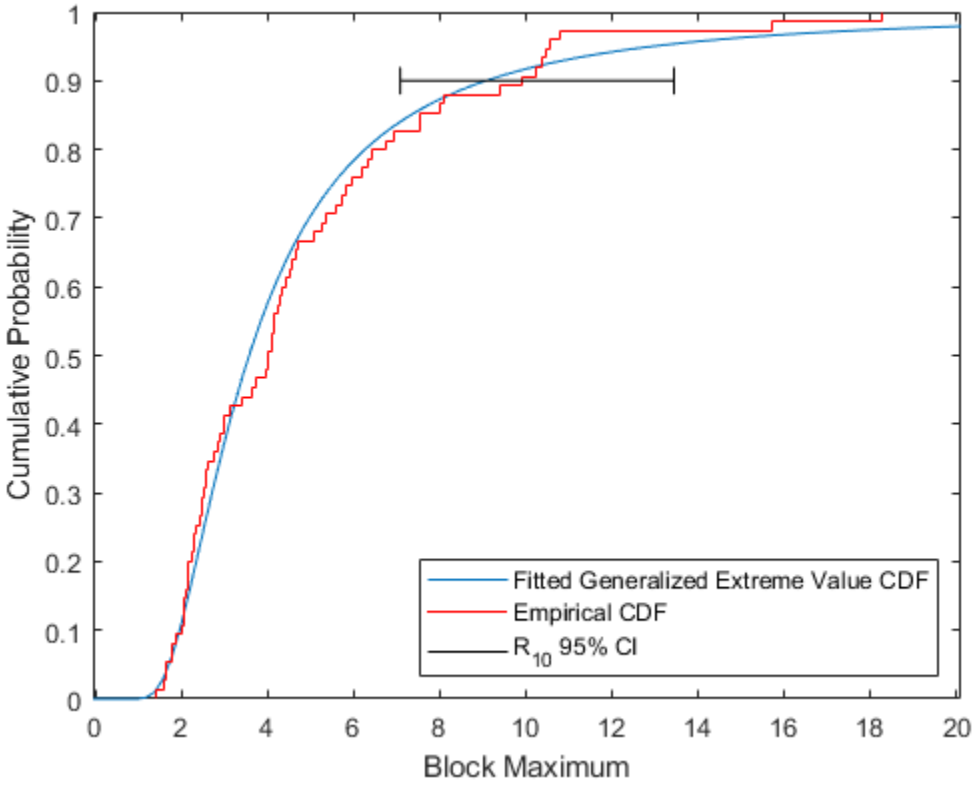
```
Feasible point with lower objective function value found.
```

To find the upper likelihood confidence limit for R10, we simply reverse the sign on the objective function to find the *largest* R10 value in the critical region, and call `fmincon` a second time.

```
CIobjfun = @(params) -gevinv(1-1./10,params(1),params(2),params(3));
[params,R10Upper,flag,output] = ...
    fmincon(CIobjfun,paramEsts,[],[],[],[],[],[],CIconfun,opts);
R10Upper = -R10Upper;

R10CI = [R10Lower, R10Upper]
```

```
R10CI =

    7.0841   13.4452
```

```
plot(ygrid,gevcdf(ygrid,kMLE,sigmaMLE,muMLE),'-');
hold on;
stairs(yi,F,'r');
plot(R10CI([1 1 1 1 2 2 2 2]), [.88 .92 NaN .9 .9 NaN .88 .92],'k-')
hold off;
xlabel('Block Maximum');
ylabel('Cumulative Probability');
legend('Fitted Generalized Extreme Value CDF','Empirical CDF', ...
       'R_{10} 95% CI','location','southeast');
xlim([lowerBnd ymax]);
```



Likelihood Profile for a Quantile

Sometimes just an interval does not give enough information about the quantity being estimated, and a profile likelihood is needed instead. To find the log-likelihood profile for R10, we will fix a possible value for R10, and then maximize the GEV log-likelihood, with the parameters constrained so that they are consistent with that current value of R10. This is a nonlinear equality constraint. If we do that over a range of R10 values, we get a likelihood profile.

As with the likelihood-based confidence interval, we can think about what this procedure would be if we fixed k and worked over the two remaining parameters, sigma and mu. Each red contour line in the contour plot shown earlier represents a fixed value of R10; the profile likelihood optimization consists of stepping along a single R10 contour line to find the highest log-likelihood (blue) contour.

For this example, we'll compute a profile likelihood for R10 over the values that were included in the likelihood confidence interval.

```
R10grid = linspace(R10CI(1)-.05*diff(R10CI), R10CI(2)+.05*diff(R10CI), 51);
```

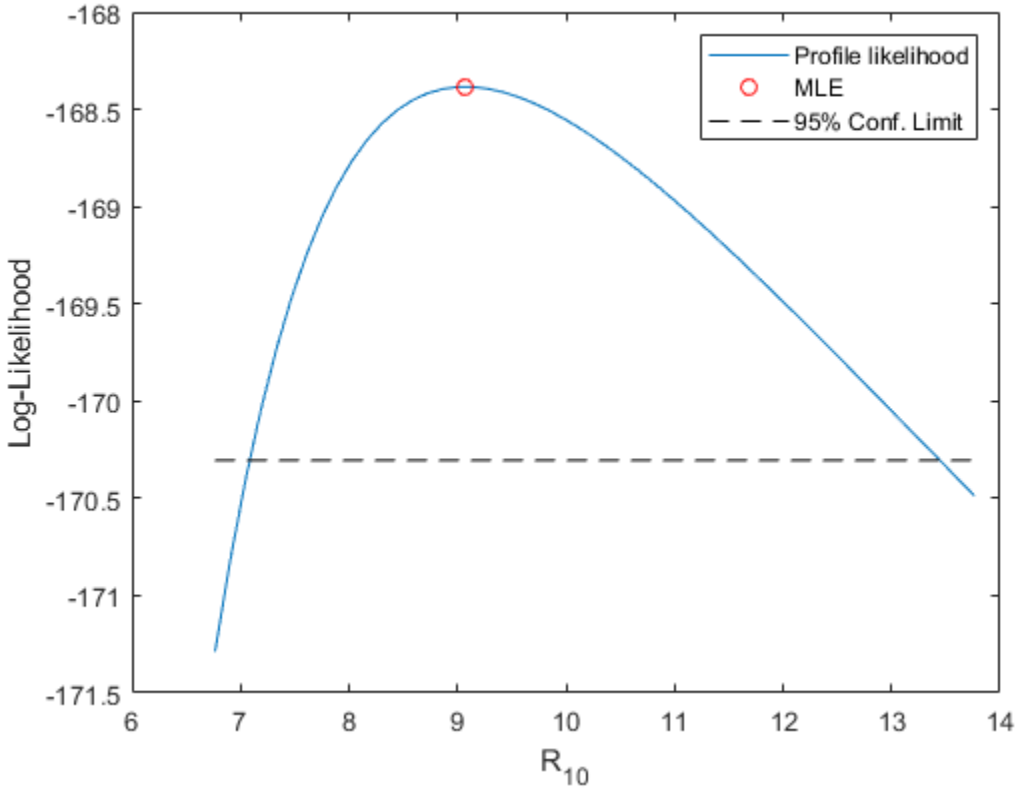
The objective function for the profile likelihood optimization is simply the log-likelihood, using the simulated data.

```
PLobjfun = @(params) gevlike(params,y);
```

To use `fmincon`, we'll need a function that returns non-zero values when the constraint is violated, that is, when the parameters are not consistent with the current value of R10. For each value of R10, we'll create an anonymous function for the particular value of R10 under consideration. It also returns an empty value because we're not using any inequality constraints here.

Finally, we'll call `fmincon` at each value of R10, to find the corresponding constrained maximum of the log-likelihood. We'll start near the maximum likelihood estimate of R10, and work out in both directions.

```
Lprof = nan(size(R10grid));
params = paramEsts;
[dum,peak] = min(abs(R10grid-R10MLE));
for i = peak:1:length(R10grid)
    PLconfun = ...
        @(params) deal([], gevinv(1-1./10,params(1),params(2),params(3)) - R10grid(i));
    [params,Lprof(i),flag,output] = ...
        fmincon(PLobjfun,params,[],[],[],[],[],[],PLconfun,opts);
end
params = paramEsts;
for i = peak-1:-1:1
    PLconfun = ...
        @(params) deal([], gevinv(1-1./10,params(1),params(2),params(3)) - R10grid(i));
    [params,Lprof(i),flag,output] = ...
        fmincon(PLobjfun,params,[],[],[],[],[],[],PLconfun,opts);
end

plot(R10grid,-Lprof,'-', R10MLE,-gevlike(paramEsts,y),'ro', ...
     [R10grid(1), R10grid(end)],[-nllCritVal,-nllCritVal],'k--');
xlabel('R_{10}');
ylabel('Log-Likelihood');
legend('Profile likelihood','MLE','95% Conf. Limit');
```



Curve Fitting and Distribution Fitting

This example shows how to perform curve fitting and distribution fitting, and discusses when each method is appropriate.

Choose Between Curve Fitting and Distribution Fitting

Curve fitting and distribution fitting are different types of data analysis.

- Use curve fitting when you want to model a response variable as a function of a predictor variable.
- Use distribution fitting when you want to model the probability distribution of a single variable.

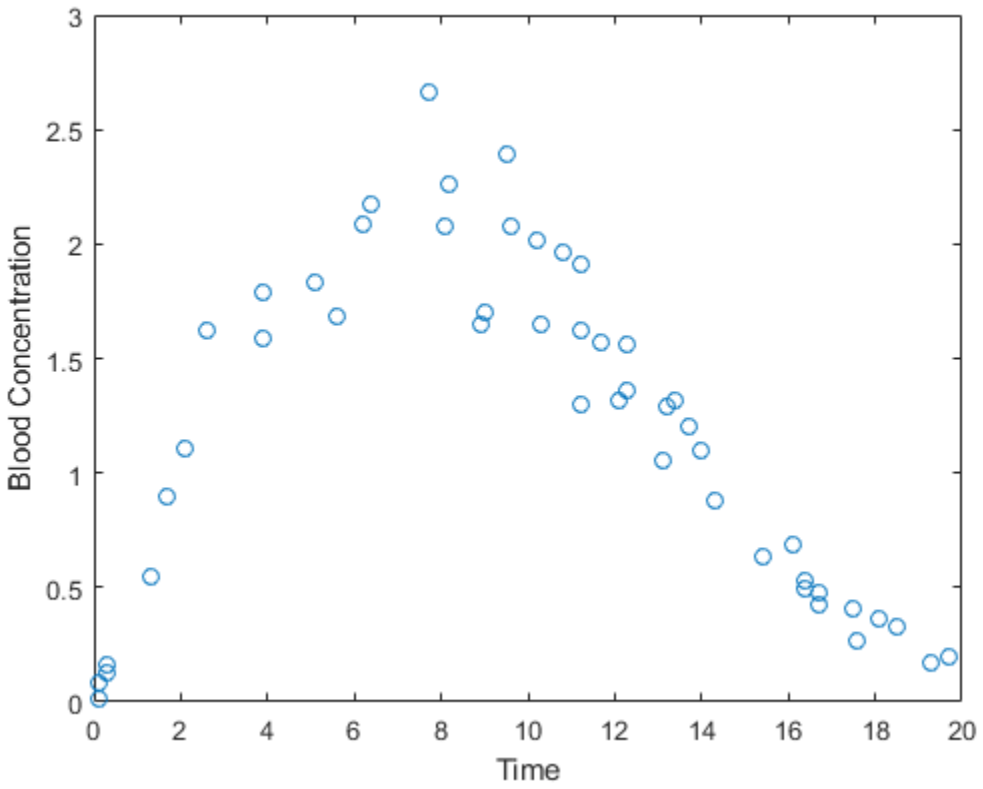
Curve Fitting

In the following experimental data, the predictor variable is `time`, the time after the ingestion of a drug. The response variable is `conc`, the concentration of the drug in the bloodstream. Assume that only the response data `conc` is affected by experimental error.

```
time = [ 0.1   0.1   0.3   0.3   1.3   1.7   2.1   2.6   3.9   3.9 ...
         5.1   5.6   6.2   6.4   7.7   8.1   8.2   8.9   9.0   9.5 ...
         9.6  10.2  10.3  10.8  11.2  11.2  11.2  11.7  12.1  12.3 ...
        12.3  13.1  13.2  13.4  13.7  14.0  14.3  15.4  16.1  16.1 ...
        16.4  16.4  16.7  16.7  17.5  17.6  18.1  18.5  19.3  19.7]';
conc = [0.01  0.08  0.13  0.16  0.55  0.90  1.11  1.62  1.79  1.59 ...
        1.83  1.68  2.09  2.17  2.66  2.08  2.26  1.65  1.70  2.39 ...
        2.08  2.02  1.65  1.96  1.91  1.30  1.62  1.57  1.32  1.56 ...
        1.36  1.05  1.29  1.32  1.20  1.10  0.88  0.63  0.69  0.69 ...
        0.49  0.53  0.42  0.48  0.41  0.27  0.36  0.33  0.17  0.20]';
```

Suppose you want to model blood concentration as a function of time. Plot `conc` against `time`.

```
plot(time,conc,'o');
xlabel('Time');
ylabel('Blood Concentration');
```



Assume that `conc` follows a two-parameter Weibull curve as a function of `time`. A Weibull curve has the form and parameters

$$y = c(x/a)^{(b-1)}e^{-(x/a)^b},$$

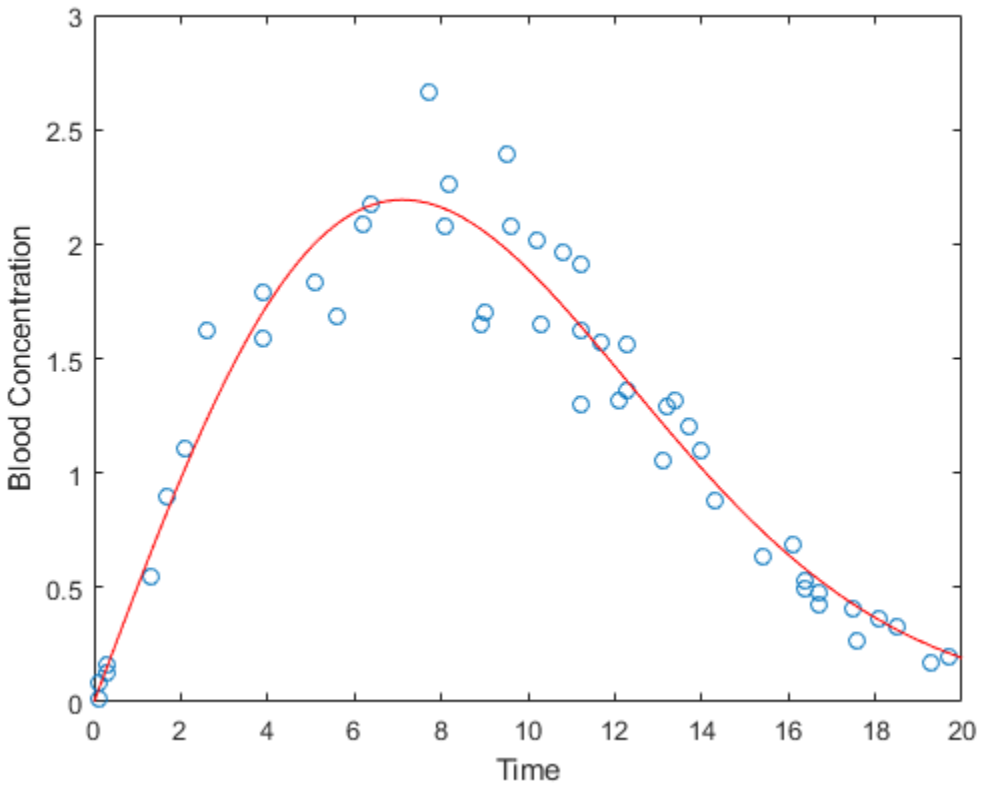
where a is a horizontal scaling, b is a shape parameter, and c is a vertical scaling.

Fit the Weibull model using nonlinear least squares.

```
modelFun =  @(p,x) p(3) .* (x./p(1)).^(p(2)-1) .* exp(-(x./p(1)).^p(2));
startingVals = [10 2 5];
nlModel = fitnlm(time,conc,modelFun,startingVals);
```

Plot the Weibull curve onto the data.

```
xgrid = linspace(0,20,100)';
line(xgrid,predict(nlModel,xgrid),'Color','r');
```



The fitted Weibull model is problematic. `fitnlm` assumes the experimental errors are additive and come from a symmetric distribution with constant variance. However, the scatter plot shows that the error variance is proportional to the height of the curve. Furthermore, the additive, symmetric errors imply that a negative blood concentration measurement is possible.

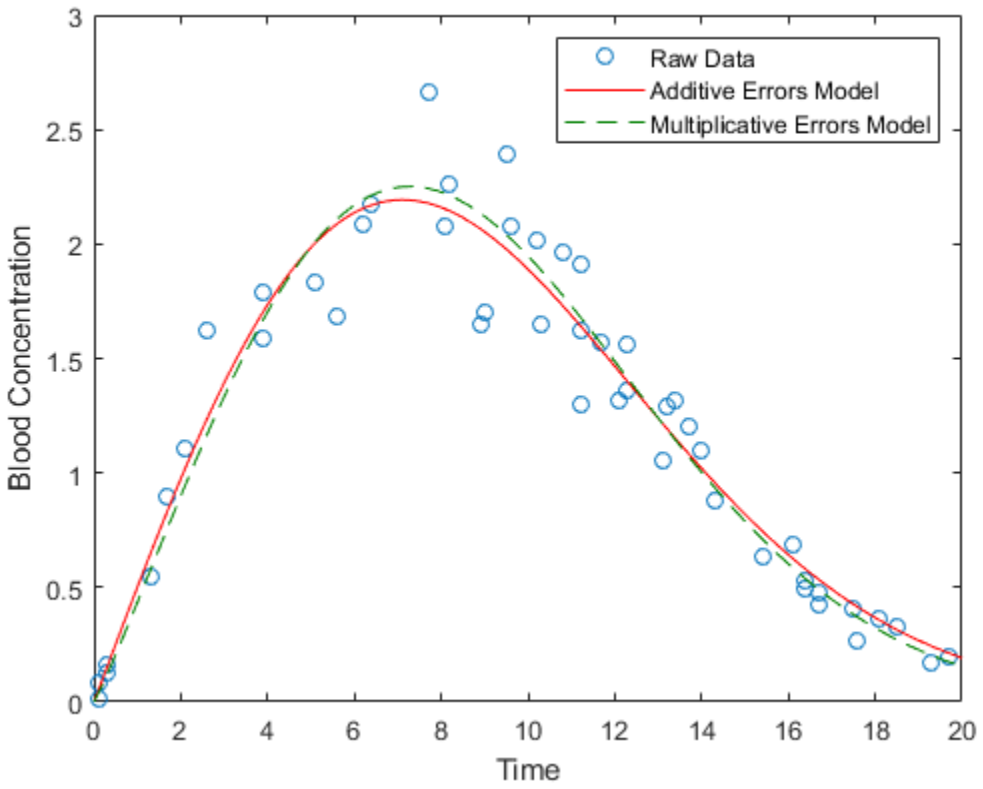
A more realistic assumption is that multiplicative errors are symmetric on the log scale. Under that assumption, fit a Weibull curve to the data by taking the log of both sides. Use nonlinear least squares to fit the curve:

$$\log(y) = \log(c) + (b-1)\log(x/a) - (x/a)^b .$$

```
nlModel2 = fitnlm(time,log(conc),@(p,x) log(modelFun(p,x)),startingVals);
```

Add the new curve to the existing plot.

```
line(xgrid,exp(predict(nlModel2,xgrid)),'Color',[0 .5 0],'LineStyle','--');
legend({'Raw Data','Additive Errors Model','Multiplicative Errors Model'});
```



The model object `nlModel2` contains estimates of precision. A best practice is to check the model's goodness of fit. For example, make residual plots on the log scale to check the assumption of constant variance for the multiplicative errors.

In this example, using the multiplicative errors model has little effect on the model predictions. For an example where the type of model has more of an impact, see “Pitfalls in Fitting Nonlinear Models by Transforming to Linearity” on page 13-53.

Functions for Curve Fitting

- Statistics and Machine Learning Toolbox™ includes these functions for fitting models: `fitnlm` for nonlinear least-squares models, `fitglm` for generalized linear models, `fitrgp` for Gaussian process regression models, and `fitrsvm` for support vector machine regression models.
- Curve Fitting Toolbox™ provides command line and graphical tools that simplify tasks in curve fitting. For example, the toolbox provides automatic choice of starting coefficient values for various models, as well as robust and nonparametric fitting methods.
- Optimization Toolbox™ has functions for performing complicated types of curve fitting analyses, such as analyzing models with constraints on the coefficients.
- The MATLAB® function `polyfit` fits polynomial models, and the MATLAB function `fminsearch` is useful in other kinds of curve fitting.

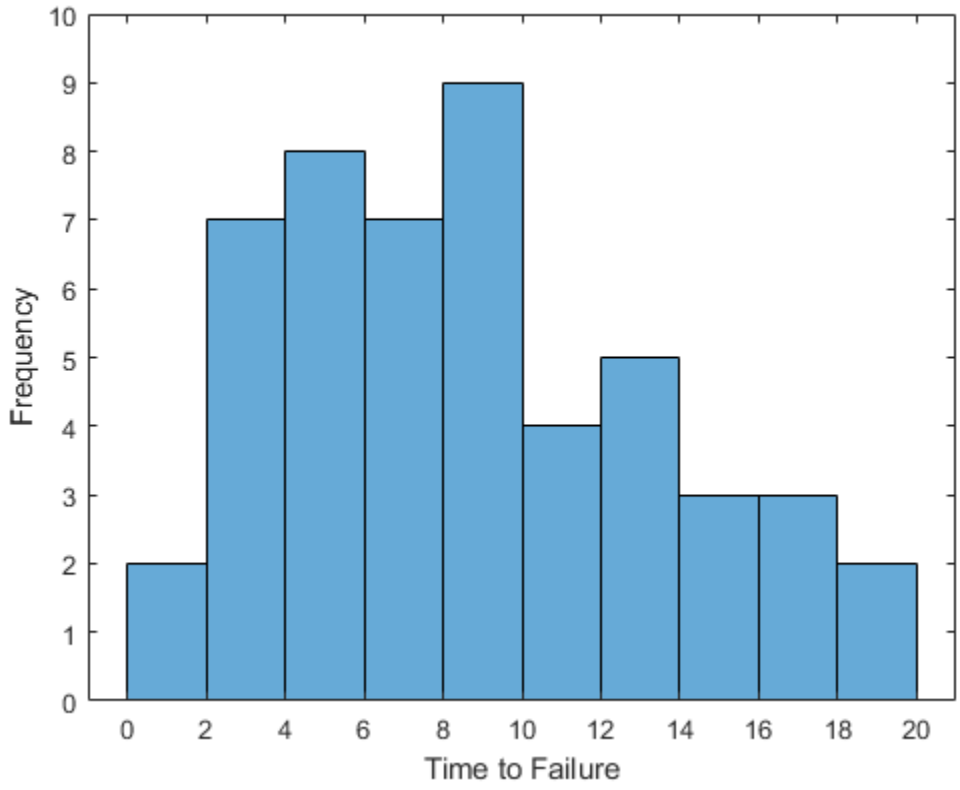
Distribution Fitting

Suppose you want to model the distribution of electrical component lifetimes. The variable `life` measures the time to failure for 50 identical electrical components.

```
life = [ 6.2 16.1 16.3 19.0 12.2  8.1  8.8  5.9  7.3  8.2 ...
        16.1 12.8  9.8 11.3  5.1 10.8  6.7  1.2  8.3  2.3 ...
         4.3  2.9 14.8  4.6  3.1 13.6 14.5  5.2  5.7  6.5 ...
         5.3  6.4  3.5 11.4  9.3 12.4 18.3 15.9  4.0 10.4 ...
         8.7  3.0 12.1  3.9  6.5  3.4  8.5  0.9  9.9  7.9]';
```

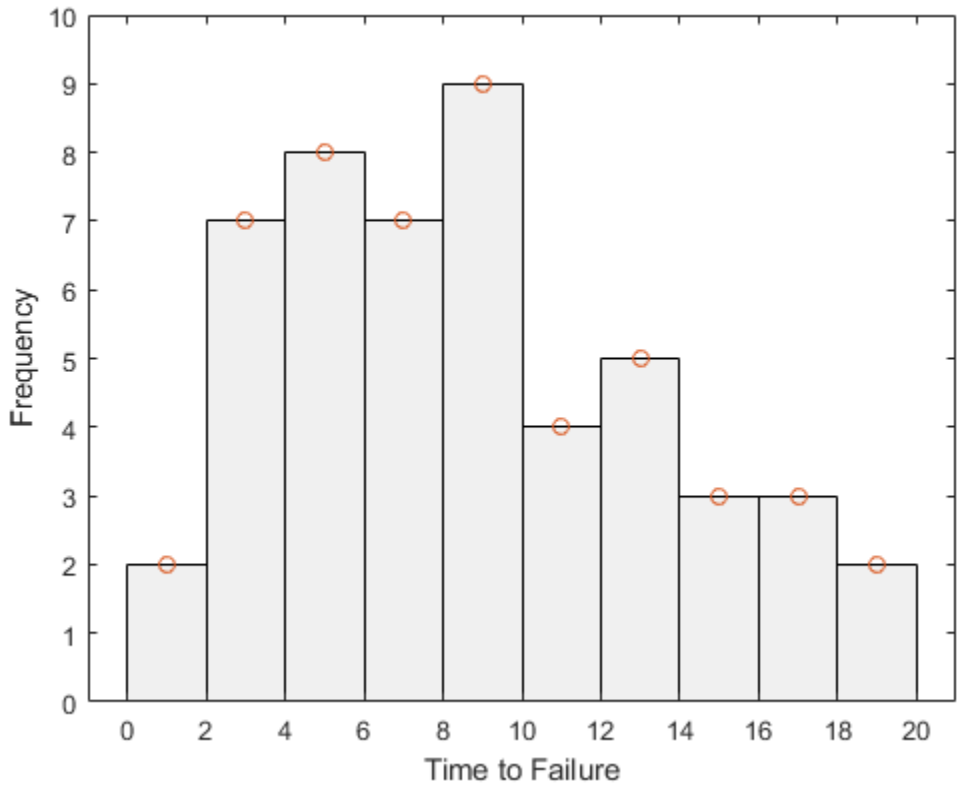
Visualize the data with a histogram.

```
binWidth = 2;
lastVal = ceil(max(life));
binEdges = 0:binWidth:lastVal+1;
h = histogram(life,binEdges);
xlabel('Time to Failure');
ylabel('Frequency');
ylim([0 10]);
```



Because lifetime data often follows a Weibull distribution, one approach might be to use the Weibull curve from the previous curve fitting example to fit the histogram. To try this approach, convert the histogram to a set of points (x,y), where x is a bin center and y is a bin height, and then fit a curve to those points.

```
counts = histcounts(life,binEdges);
binCtrs = binEdges(1:end-1) + binWidth/2;
h.FaceColor = [.9 .9 .9];
hold on
plot(binCtrs,counts,'o');
hold off
```



Fitting a curve to a histogram, however, is problematic and usually not recommended.

1. The process violates basic assumptions of least-squares fitting. The bin counts are nonnegative, implying that measurement errors cannot be symmetric. Also, the bin counts have different variability in the tails than in the center of the distribution. Finally, the bin counts have a fixed sum, implying that they are not independent measurements.
2. If you fit a Weibull curve to the bar heights, you have to constrain the curve because the histogram is a scaled version of an empirical probability density function (pdf).
3. For continuous data, fitting a curve to a histogram rather than data discards information.
4. The bar heights in the histogram are dependent on the choice of bin edges and bin widths.

For many parametric distributions, maximum likelihood is a better way to estimate parameters because it avoids these problems. The Weibull pdf has almost the same form as the Weibull curve:

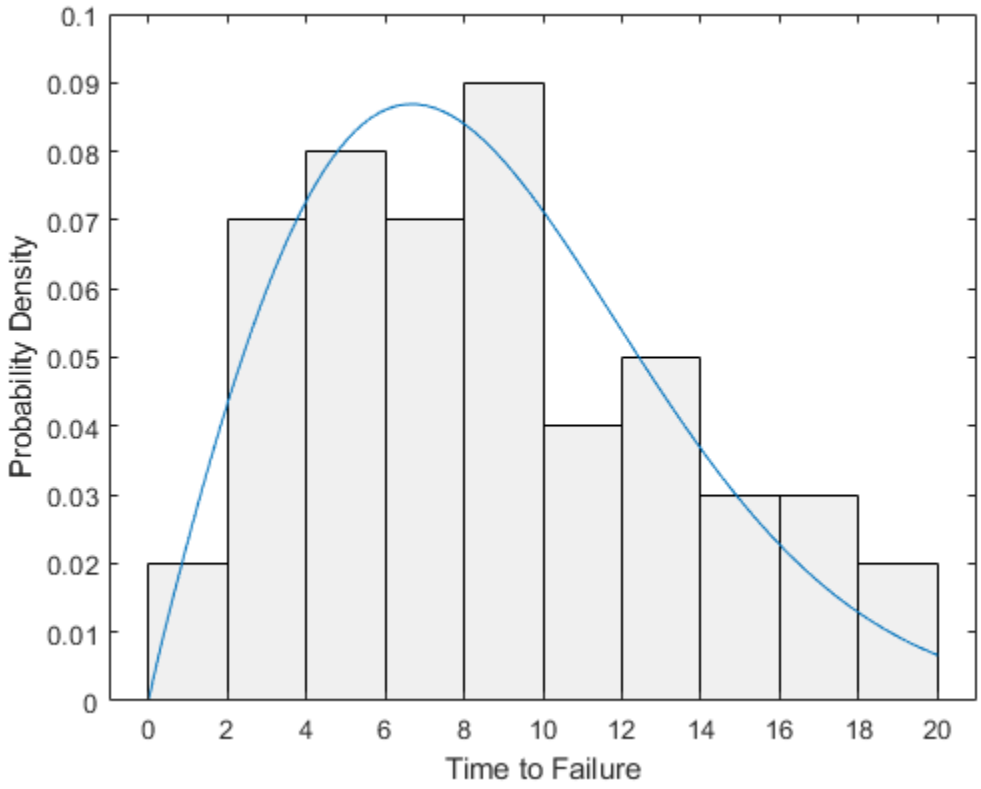
$$y = (b/a)(x/a)^{(b-1)}e^{-(x/a)^b}.$$

However, b/a replaces the scale parameter c because the function must integrate to 1. To fit a Weibull distribution to the data using maximum likelihood, use `fitdist` and specify `'Weibull'` as the distribution name. Unlike least squares, maximum likelihood finds a Weibull pdf that best matches the scaled histogram without minimizing the sum of the squared differences between the pdf and the bar heights.

```
pd = fitdist(life,'Weibull');
```

Plot a scaled histogram of the data and superimpose the fitted pdf.

```
h = histogram(life,binEdges,'Normalization','pdf','FaceColor',[.9 .9 .9]);
xlabel('Time to Failure');
ylabel('Probability Density');
ylim([0 0.1]);
xgrid = linspace(0,20,100)';
pdfEst = pdf(pd,xgrid);
line(xgrid,pdfEst)
```



A best practice is to check the model's goodness of fit.

Although fitting a curve to a histogram is usually not recommended, the process is appropriate in some cases. For an example, see “Fitting Custom Univariate Distributions” on page 5-165.

Functions for Distribution Fitting

- Statistics and Machine Learning Toolbox™ includes the function `fitdist` for fitting probability distribution objects to data. It also includes dedicated fitting functions (such as `wblfit`) for fitting parametric distributions using maximum likelihood, the function `mle` for fitting custom distributions without dedicated fitting functions, and the function `ksdensity` for fitting nonparametric distribution models to data.
- Statistics and Machine Learning Toolbox additionally provides the Distribution Fitter app, which simplifies many tasks in distribution fitting, such as generating visualizations and diagnostic plots.
- Functions in Optimization Toolbox™ enable you to fit complicated distributions, including those with constraints on the parameters.

- The MATLAB® function `fminsearch` provides maximum likelihood distribution fitting.

See Also

Distribution Fitter | `fitdist` | `fitglm` | `fitnlm` | `fitrgp` | `fitrsvm` | `fminsearch` | `ksdensity` | `mle` | `polyfit`

More About

- “Supported Distributions” on page 5-14

Fitting a Univariate Distribution Using Cumulative Probabilities

This example shows how to fit univariate distributions using least squares estimates of the cumulative distribution functions. This is a generally-applicable method that can be useful in cases when maximum likelihood fails, for instance some models that include a threshold parameter.

The most common method for fitting a univariate distribution to data is maximum likelihood. But maximum likelihood does not work in all cases, and other estimation methods, such as the Method of Moments, are sometimes needed. When applicable, maximum likelihood is probably the better choice of methods, because it is often more efficient. But the method described here provides another tool that can be used when needed.

Fitting an Exponential Distribution Using Least Squares

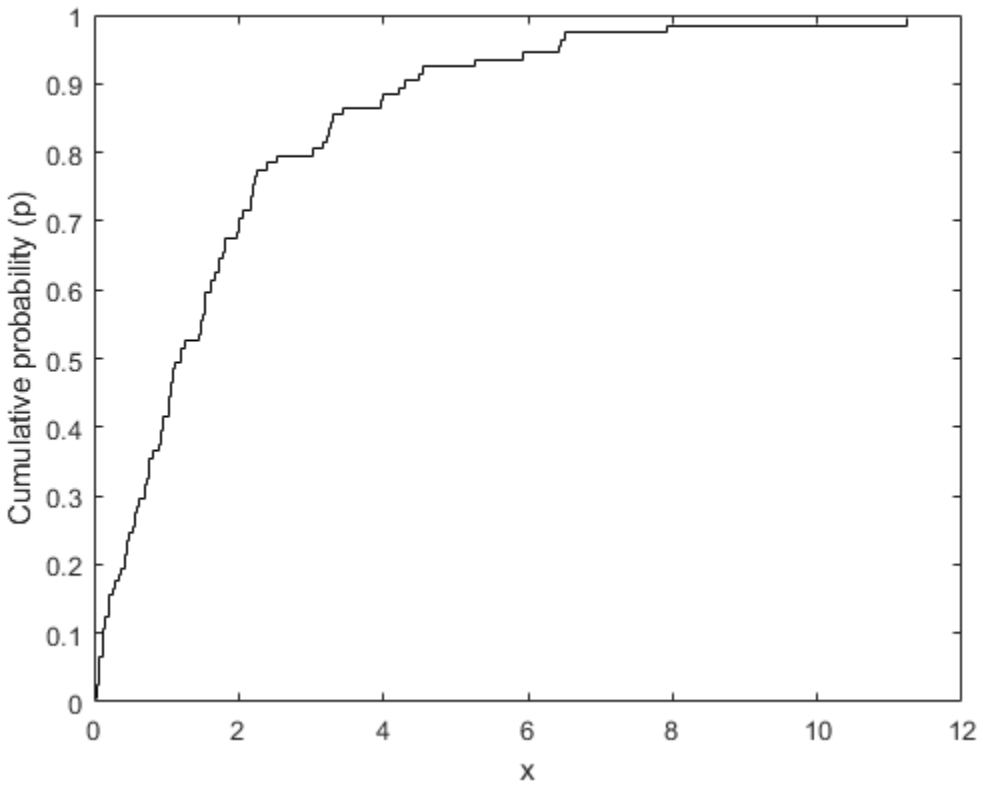
The term "least squares" is most commonly used in the context of fitting a regression line or surface to model a response variable as a function of one or more predictor variables. The method described here is a very different application of least squares: univariate distribution fitting, with only a single variable.

To begin, first simulate some sample data. We'll use an exponential distribution to generate the data. For the purposes of this example, as in practice, we'll assume that the data are not known to have come from a particular model.

```
rng(37,'twister');
n = 100;
x = exprnd(2,n,1);
```

Next, compute the empirical cumulative distribution function (ECDF) of the data. This is simply a step function with a jump in cumulative probability, p, of 1/n at each data point, x.

```
x = sort(x);
p = ((1:n)-0.5)' ./ n;
stairs(x,p,'k-');
xlabel('x');
ylabel('Cumulative probability (p)');
```



We'll fit an exponential distribution to these data. One way to do that is to find the exponential distribution whose cumulative distribution function (CDF) best approximates (in a sense to be explained below) the ECDF of the data. The exponential CDF is p = Pr{X <= x} = 1 - exp(-x/mu). Transforming that to -log(1-p)*mu = x gives a linear relationship between -log(1-p) and x. If the data do come from an exponential, we ought to see, at least approximately, a linear relationship if we plug the computed x and p values from the ECDF into that equation. If we use least squares to fit a straight line through the origin to x vs. -log(1-p), then that fitted line represents the exponential distribution that is "closest" to the data. The slope of the line is an estimate of the parameter mu.

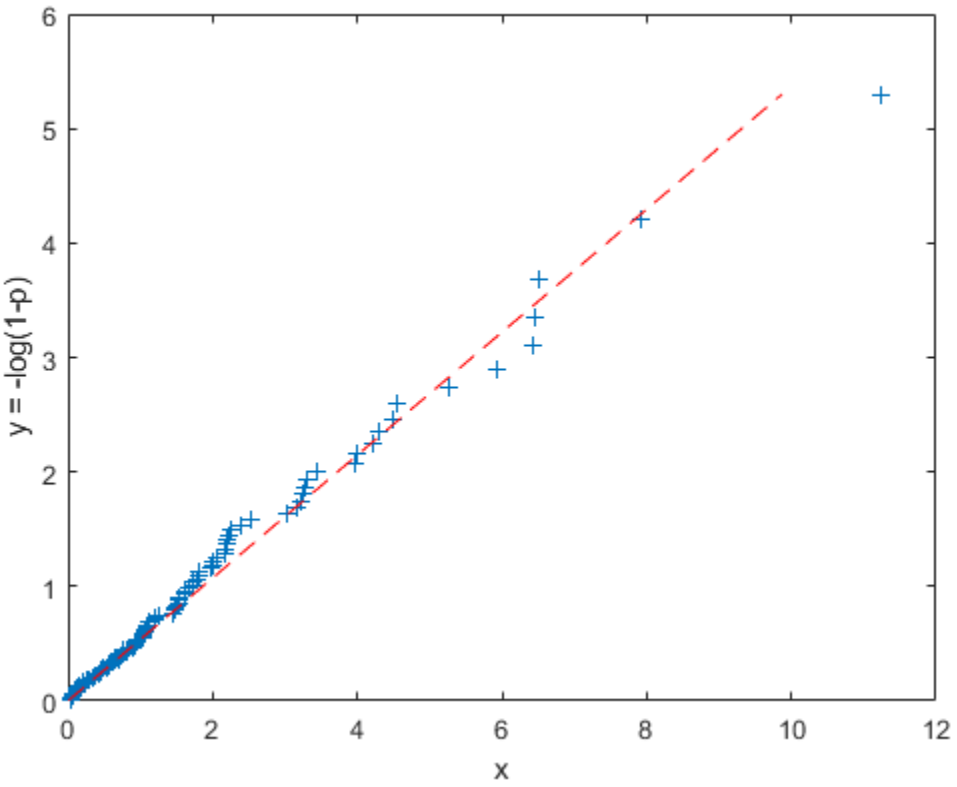
Equivalently, we can think of y = -log(1-p) as an "idealized sample" from a standard (mean 1) exponential distribution. These idealized values are exactly equally spaced on the probability scale. A Q-Q plot of x and y ought to be approximately linear if the data come from an exponential distribution, and we'll fit the least squares line through the origin to x vs. y.

```
y = -log(1 - p);
muHat = y \ x
```

```
muHat =

    1.8627
```

Plot the data and the fitted line.

```
plot(x,y,'+', y*muHat,y,'r--');
xlabel('x');
ylabel('y = -log(1-p)');
```



Notice that the linear fit we've made minimizes the sum of squared errors in the horizontal, or "x", direction. That's because the values for y = -log(1-p) are deterministic, and it's the x values that are random. It's also possible to regress y vs. x, or to use other types of linear fits, for example, weighted regression, orthogonal regression, or even robust regression. We will not explore those possibilities here.

For comparison, fit the data by maximum likelihood.

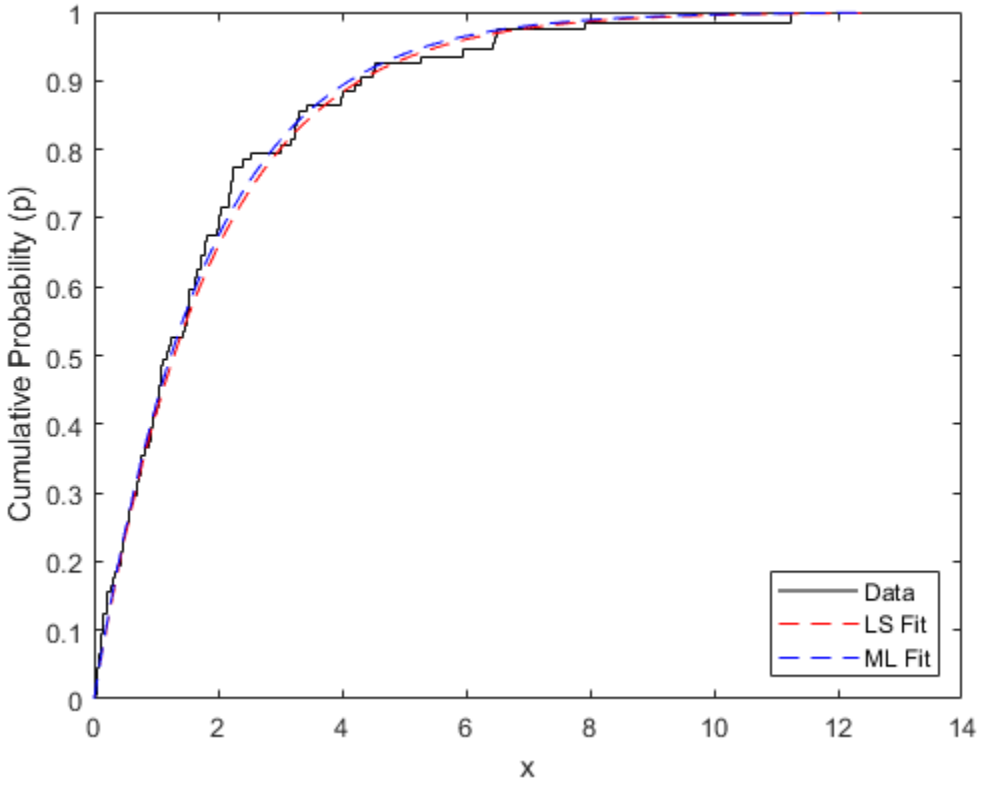
```
muMLE = expfit(x)
```

```
muMLE =

    1.7894
```

Now plot the two estimated distributions on the untransformed cumulative probability scale.

```
stairs(x,p,'k-');
hold on
xgrid = linspace(0,1.1*max(x),100)';
plot(xgrid,expcdf(xgrid,muHat),'r--', xgrid,expcdf(xgrid,muMLE),'b--');
hold off
```

```
xlabel('x'); ylabel('Cumulative Probability (p)');
legend({'Data','LS Fit','ML Fit'},'location','southeast');
```



The two methods give very similar fitted distributions, although the LS fit has been influenced more by observations in the tail of the distribution.

Fitting a Weibull Distribution

For a slightly more complex example, simulate some sample data from a Weibull distribution, and compute the ECDF of x.

```
n = 100;
x = wblrnd(2,1,n,1);
x = sort(x);
p = ((1:n)-0.5)' ./ n;
```

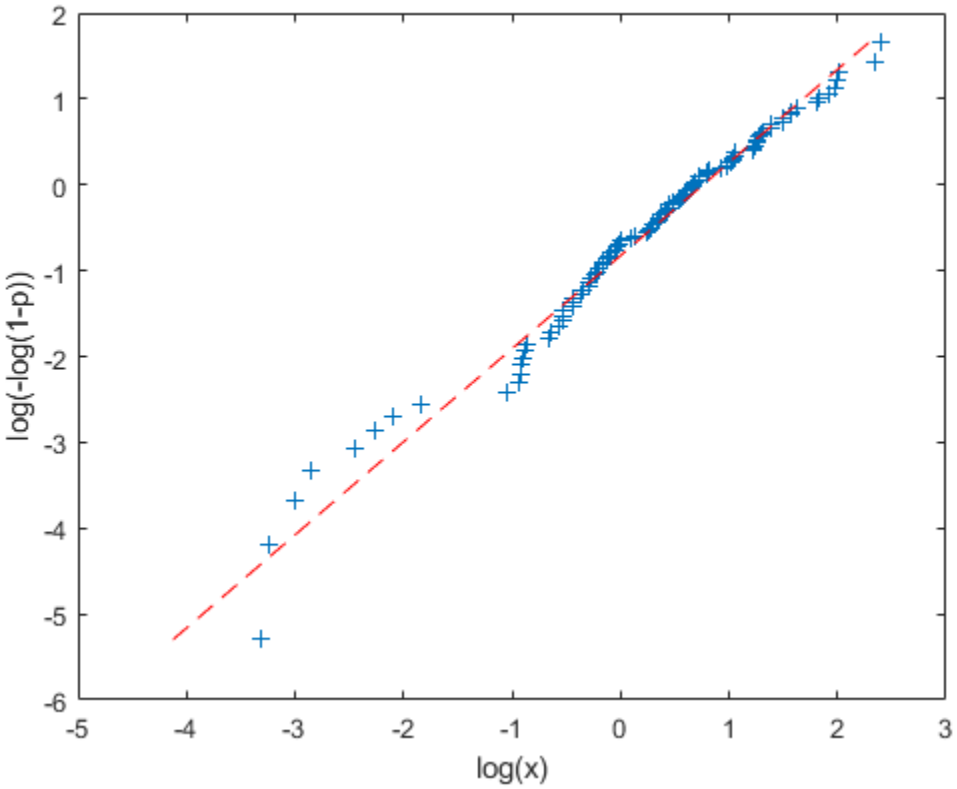
To fit a Weibull distribution to these data, notice that the CDF for the Weibull is p = Pr{X <= x} = 1 - exp(-(x/a)^b). Transforming that to log(a) + log(-log(1-p))*(1/b) = log(x) again gives a linear relationship, this time between log(-log(1-p)) and log(x). We can use least squares to fit a straight line on the transformed scale using p and x from the ECDF, and the slope and intercept of that line lead to estimates of a and b.

```
logx = log(x);
logy = log(-log(1 - p));
poly = polyfit(logy,logx,1);
paramHat = [exp(poly(2)) 1/poly(1)]
```

```
paramHat =

    2.1420    1.0843
```

Plot the data and the fitted line on the transformed scale.

```
plot(logx,logy,'+', log(paramHat(1)) + logy/paramHat(2),logy,'r--');
xlabel('log(x)');
ylabel('log(-log(1-p))');
```

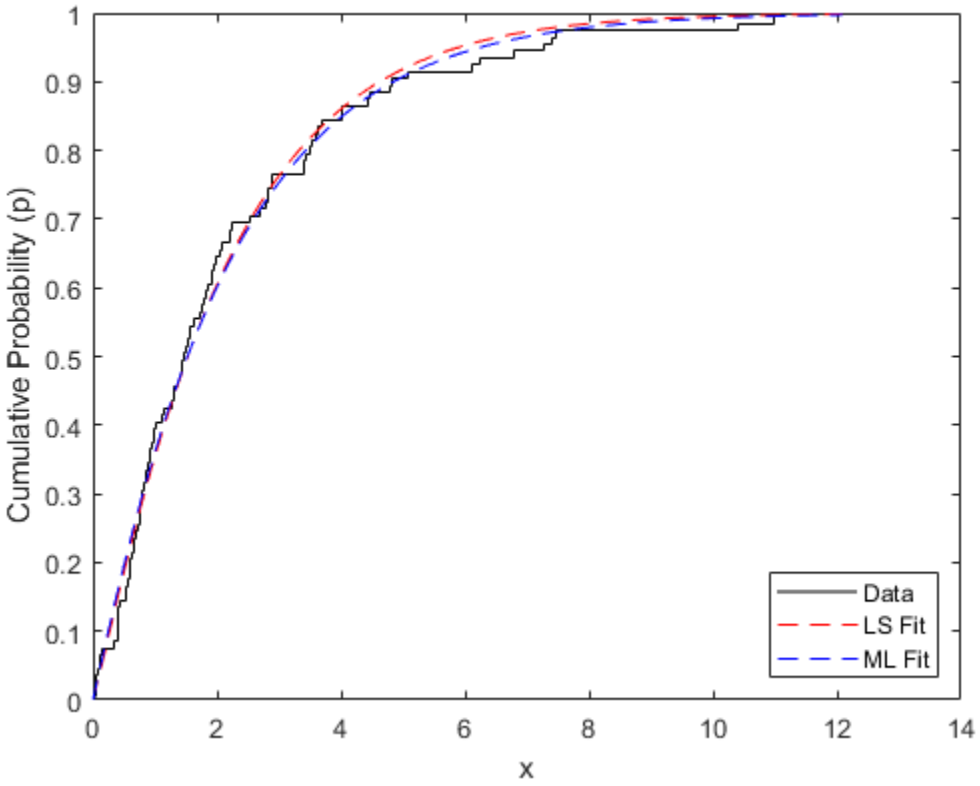


For comparison, fit the data by maximum likelihood, and plot the two estimated distributions on the untransformed scale.

```
paramMLE = wblfit(x)
stairs(x,p,'k');
hold on
xgrid = linspace(0,1.1*max(x),100)';
plot(xgrid,wblcdf(xgrid,paramHat(1),paramHat(2)),'r--', ...
     xgrid,wblcdf(xgrid,paramMLE(1),paramMLE(2)),'b--');
hold off
xlabel('x'); ylabel('Cumulative Probability (p)');
legend({'Data','LS Fit','ML Fit'},'location','southeast');
```

```
paramMLE =
```

```
    2.1685    1.0372
```

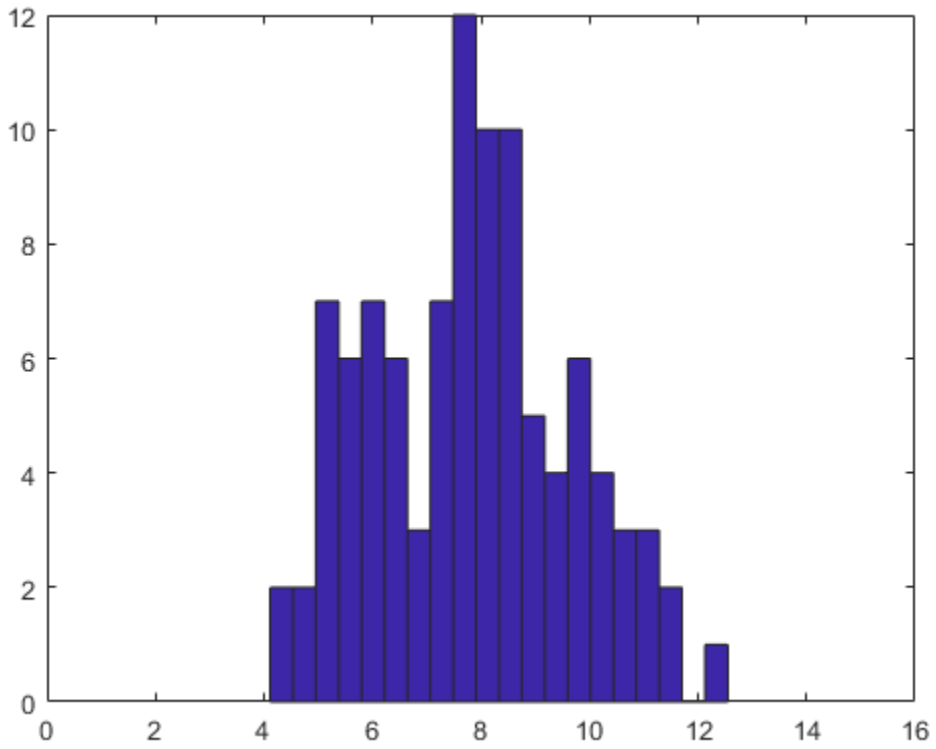


A Threshold Parameter Example

It's sometimes necessary to fit positive distributions like the Weibull or lognormal with a threshold parameter. For example, a Weibull random variable takes values over (0,Inf), and a threshold parameter, c, shifts that range to (c,Inf). If the threshold parameter is known, then there is no difficulty. But if the threshold parameter is not known, it must instead be estimated. These models are difficult to fit with maximum likelihood -- the likelihood can have multiple modes, or even become infinite for parameter values that are not reasonable for the data, and so maximum likelihood is often not a good method. But with a small addition to the least squares procedure, we can get stable estimates.

To illustrate, we'll simulate some data from a three-parameter Weibull distribution, with a threshold value. As above, we'll assume for the purposes of the example that the data are not known to have come from a particular model, and that the threshold is not known.

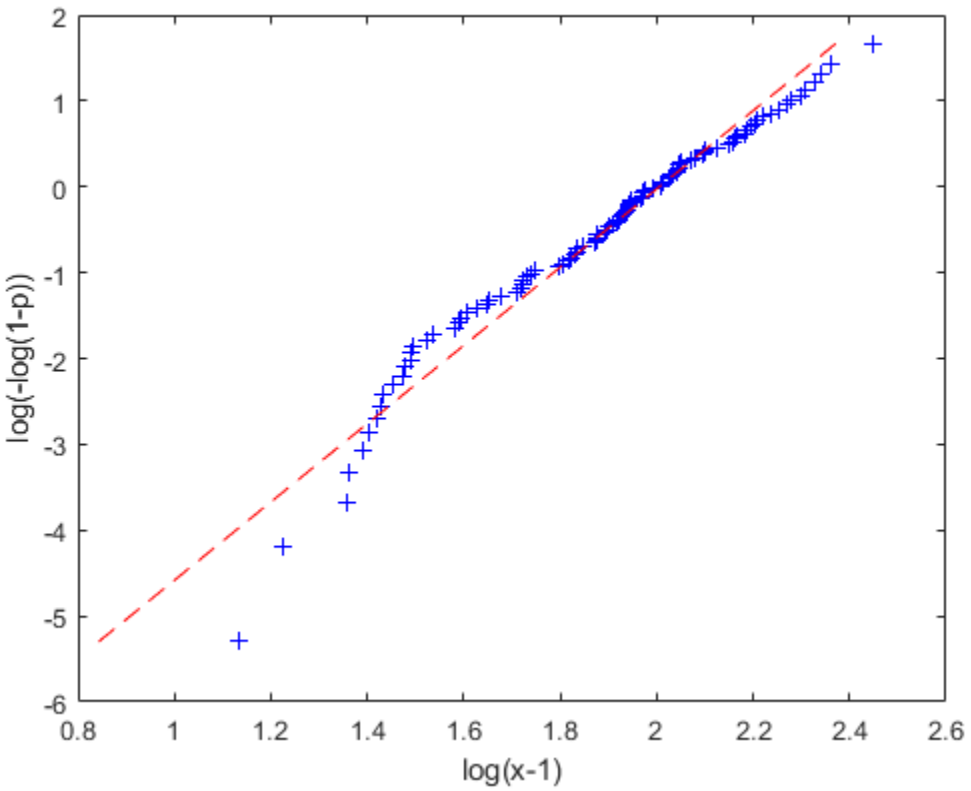
```
n = 100;
x = wblrnd(4,2,n,1) + 4;
hist(x,20); xlim([0 16]);
```



How can we fit a three-parameter Weibull distribution to these data? If we knew what the threshold value was, 1 for example, we could subtract that value from the data and then use the least squares procedure to estimate the Weibull shape and scale parameters.

```
x = sort(x);
p = ((1:n)-0.5)' ./ n;
logy = log(-log(1-p));
logxm1 = log(x-1);
poly1 = polyfit(log(-log(1-p)),log(x-1),1);
paramHat1 = [exp(poly1(2)) 1/poly1(1)]
plot(logxm1,logy,'b+', log(paramHat1(1)) + logy/paramHat1(2),logy,'r--');
xlabel('log(x-1)');
ylabel('log(-log(1-p))');
```

```
paramHat1 =

    7.4305    4.5574
```



That's not a very good fit -- log(x-1) and log(-log(1-p)) do not have a linear relationship. Of course, that's because we don't know the correct threshold value. If we try subtracting different threshold values, we get different plots and different parameter estimates.

```
logxm2 = log(x-2);
poly2 = polyfit(log(-log(1-p)),log(x-2),1);
paramHat2 = [exp(poly2(2)) 1/poly2(1)]
```

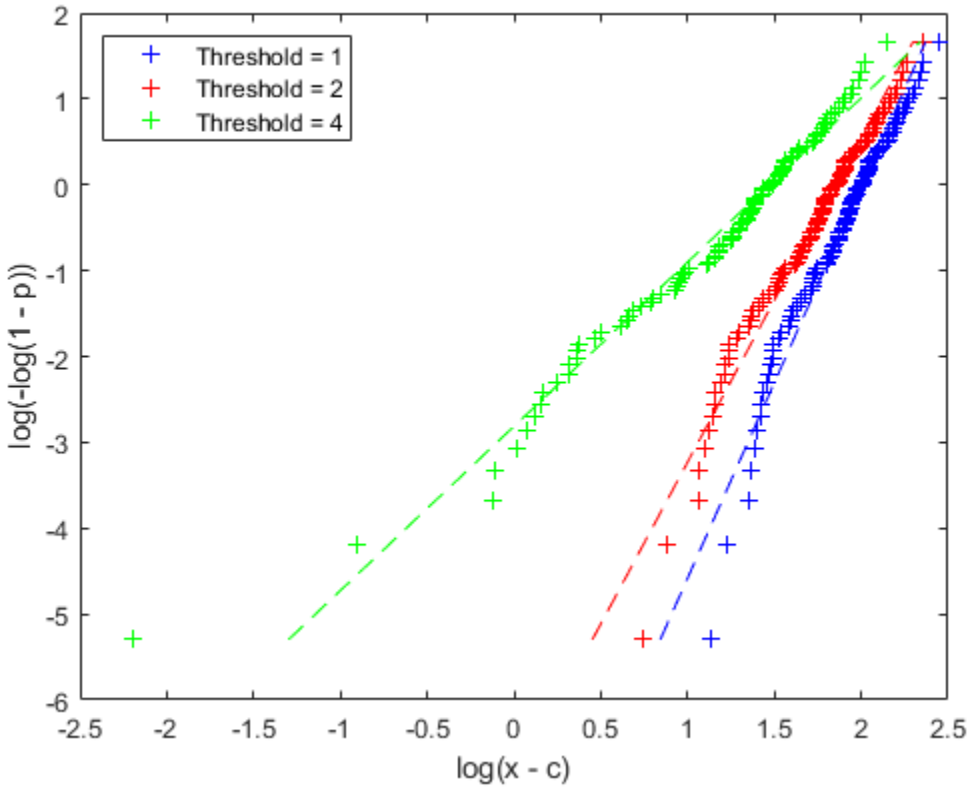
```
paramHat2 =

    6.4046    3.7690
```

```
logxm4 = log(x-4);
poly4 = polyfit(log(-log(1-p)),log(x-4),1);
paramHat4 = [exp(poly4(2)) 1/poly4(1)]
```

```
paramHat4 =

    4.3530    1.9130
```

```
plot(logxm1,logy,'b+', logxm2,logy,'r+', logxm4,logy,'g+', ...
     log(paramHat1(1)) + logy/paramHat1(2),logy,'b--', ...
     log(paramHat2(1)) + logy/paramHat2(2),logy,'r--', ...
     log(paramHat4(1)) + logy/paramHat4(2),logy,'g--');
```

```
xlabel('log(x - c)');
ylabel('log(-log(1 - p))');
legend({'Threshold = 1' 'Threshold = 2' 'Threshold = 4'}, 'location','northwest');
```

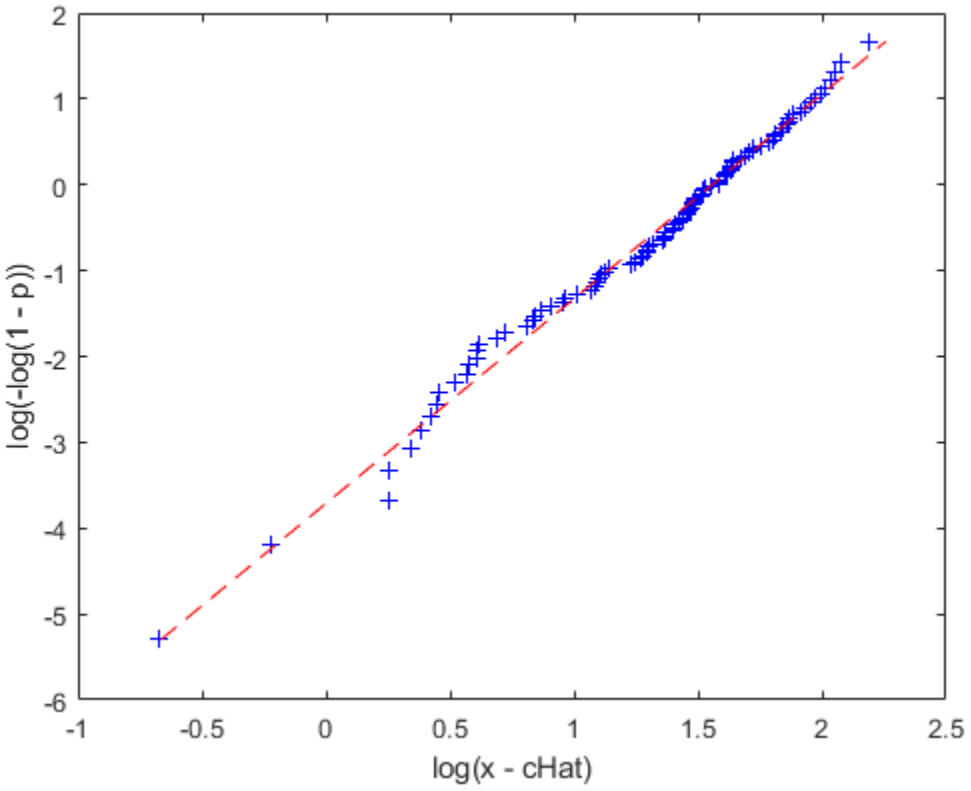


The relationship between log(x-4) and log(-log(1-p)) appears approximately linear. Since we'd expect to see an approximately linear plot if we subtracted the true threshold parameter, this is evidence that 4 might be a reasonable value for the threshold. On the other hand, the plots for 2 and 3 differ more systematically from linear, which is evidence that those values are not consistent with the data.

This argument can be formalized. For each provisional value of the threshold parameter, the corresponding provisional Weibull fit can be characterized as the parameter values that maximize the R^2 value of a linear regression on the transformed variables log(x-c) and log(-log(1-p)). To estimate the threshold parameter, we can carry that one step further, and maximize the R^2 value over all possible threshold values.

```
r2 = @(x,y) 1 - norm(y - polyval(polyfit(x,y,1),x)).^2 / norm(y - mean(y)).^2;
threshObj = @(c) -r2(log(-log(1-p)),log(x-c));
cHat = fminbnd(threshObj,.75*min(x), .9999*min(x));
poly = polyfit(log(-log(1-p)),log(x-cHat),1);
paramHat = [exp(poly(2)) 1/poly(1) cHat]
logx = log(x-cHat);
logy = log(-log(1-p));
plot(logx,logy,'b+', log(paramHat(1)) + logy/paramHat(2),logy,'r--');
xlabel('log(x - cHat)');
ylabel('log(-log(1 - p))');
```

```
paramHat =

    4.7448    2.3839    3.6029
```



Non-Location-Scale Families

The exponential distribution is a scale family, and on the log scale, the Weibull distribution is a location-scale family, so this least squares method was straightforward in those two cases. The general procedure to fit a location-scale distribution is

- Compute the ECDF of the observed data.
- Transform the distribution's CDF to get a linear relationship between some function of the data and some function of the cumulative probability. These two functions do not involve the distribution parameters, but the slope and intercept of the line do.
- Plug the values of x and p from the ECDF into that transformed CDF, and fit a straight line using least squares.
- Solve for the distribution parameters in terms of the slope and intercept of the line.

We also saw that fitting a distribution that is a location-scale family with an additional a threshold parameter is only slightly more difficult.

But other distributions that are not location-scale families, like the gamma, are a bit trickier. There's no transformation of the CDF that will give a relationship that is linear. However, we can use a similar

idea, only this time working on the untransformed cumulative probability scale. A P-P plot is the appropriate way to visualize that fitting procedure.

If the empirical probabilities from the ECDF are plotted against fitted probabilities from a parametric model, a tight scatter along the 1:1 line from zero to one indicates that the parameter values define a distribution that explains the observed data well, because the fitted CDF approximates the empirical CDF well. The idea is to find parameter values that make the probability plot as close to the 1:1 line as possible. That may not even be possible, if the distribution is not a good model for the data. If the P-P plot shows a systematic departure from the 1:1 line, then the model may be questionable. However, it's important to remember that since the points in these plots are not independent, interpretation is not exactly the same as a regression residual plot.

For example, we'll simulate some data and fit a gamma distribution.

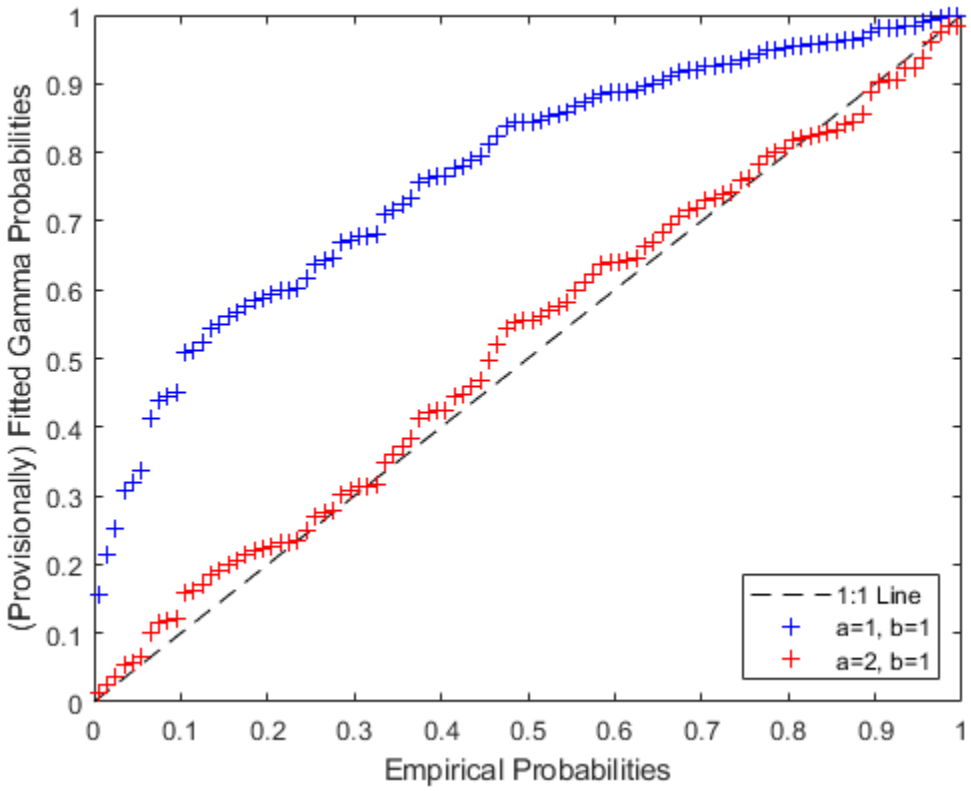
```
n = 100;
x = gamrnd(2,1,n,1);
```

Compute the ECDF of x.

```
x = sort(x);
pEmp = ((1:n)-0.5)' ./ n;
```

We can make a probability plot using any initial guess for the gamma distribution's parameters, a=1 and b=1, say. That guess is not very good -- the probabilities from the parametric CDF are not close to the probabilities from the ECDF. If we tried a different a and b, we'd get a different scatter on the P-P plot, with a different discrepancy from the 1:1 line. Since we know the true a and b in this example, we'll try those values.

```
a0 = 1; b0 = 1;
p0Fit = gamcdf(x,a0,b0);
a1 = 2; b1 = 1;
p1Fit = gamcdf(x,a1,b1);
plot([0 1],[0 1],'k--', pEmp,p0Fit,'b+', pEmp,p1Fit,'r+');
xlabel('Empirical Probabilities');
ylabel('(Provisionally) Fitted Gamma Probabilities');
legend({'1:1 Line','a=1, b=1', 'a=2, b=1'}, 'location','southeast');
```



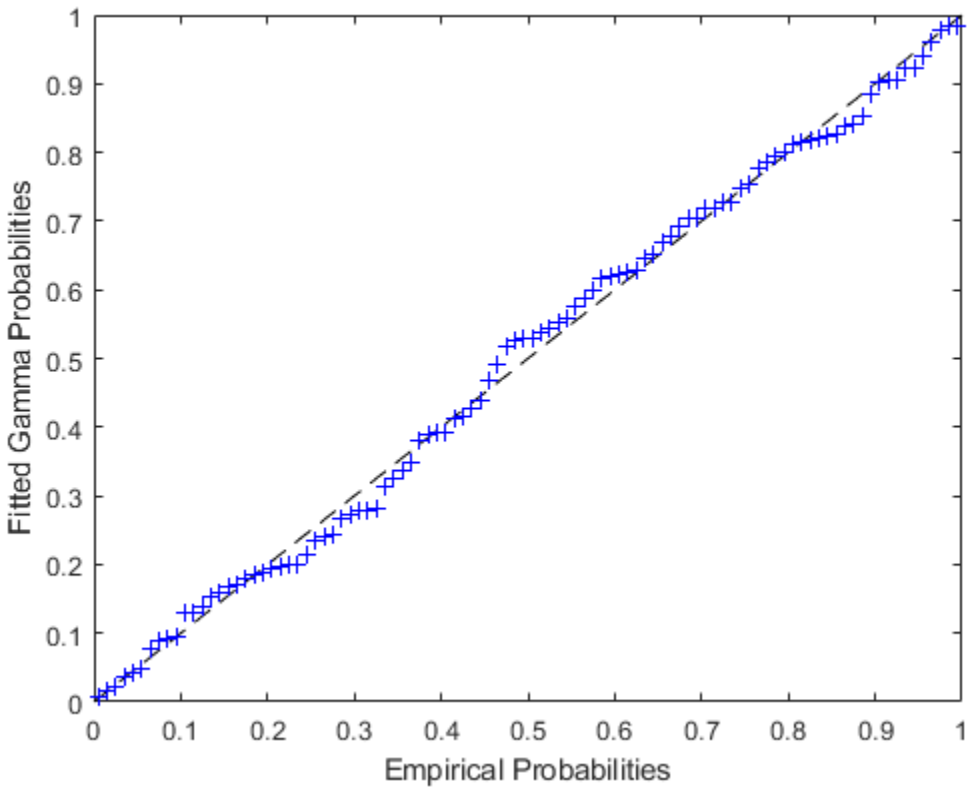
The second set of values for a and b make for a much better plot, and thus are more compatible with the data, if you are measuring "compatible" by how straight you can make the P-P plot.

To make the scatter match the 1:1 line as closely possible, we can find the values of a and b that minimize a weighted sum of the squared distances to the 1:1 line. The weights are defined in terms of the empirical probabilities, and are lowest in the center of the plot and highest at the extremes. These weights compensate for the variance of the fitted probabilities, which is highest near the median and lowest in the tails. This weighted least squares procedure defines the estimator for a and b.

```
wgt = 1 ./ sqrt(pEmp.*(1-pEmp));
gammaObj = @(params) sum(wgt.*(gamcdf(x,exp(params(1)),exp(params(2)))-pEmp).^2);
paramHat = fminsearch(gammaObj,[log(a1),log(b1)]);
paramHat = exp(paramHat)
```

```
paramHat =

    2.2759    0.9059
```

```
pFit = gamcdf(x,paramHat(1),paramHat(2));
plot([0 1],[0 1],'k--', pEmp,pFit,'b+');
xlabel('Empirical Probabilities');
ylabel('Fitted Gamma Probabilities');
```



Notice that in the location-scale cases considered earlier, we could fit the distribution with a single straight line fit. Here, as with the threshold parameter example, we had to iteratively find the best-fit parameter values.

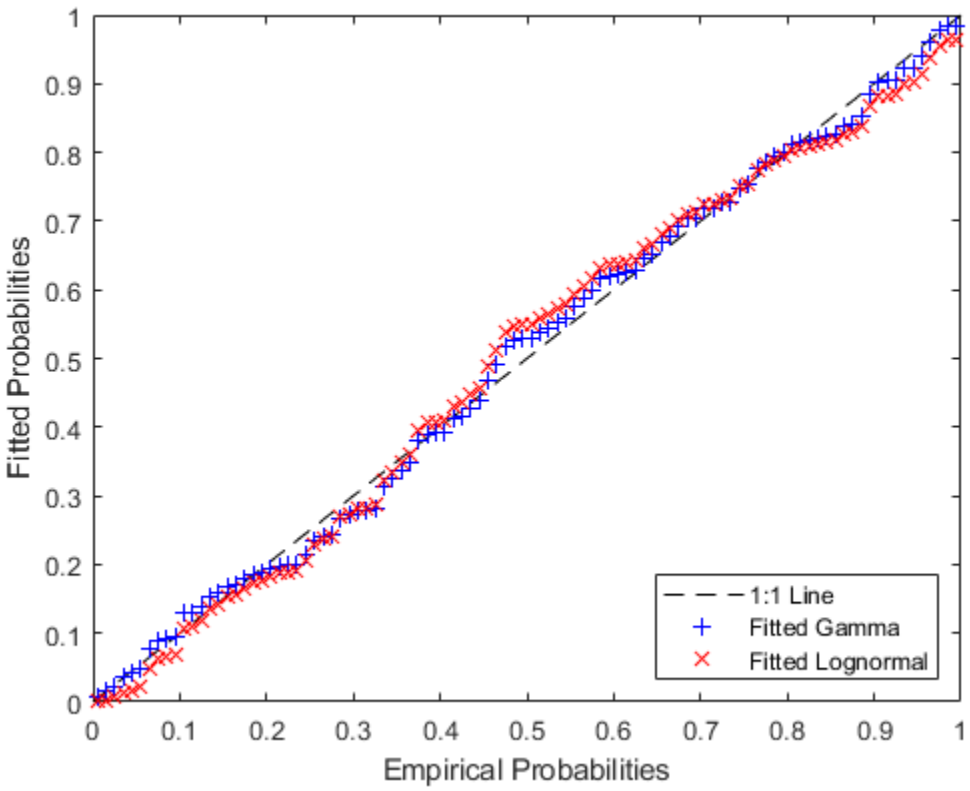
Model Misspecification

The P-P plot can also be useful for comparing fits from different distribution families. What happens if we try to fit a lognormal distribution to these data?

```
wgt = 1 ./ sqrt(pEmp.*(1-pEmp));
LNobj = @(params) sum(wgt.*(logncdf(x,params(1),exp(params(2)))-pEmp).^2);
mu0 = mean(log(x)); sigma0 = std(log(x));
paramHatLN = fminsearch(LNobj,[mu0,log(sigma0)]);
paramHatLN(2) = exp(paramHatLN(2))
```

```
paramHatLN =

    0.5331    0.7038
```

```
pFitLN = logncdf(x,paramHatLN(1),paramHatLN(2));
hold on
plot(pEmp,pFitLN,'rx');
hold off
ylabel('Fitted Probabilities');
legend({'1:1 Line', 'Fitted Gamma', 'Fitted Lognormal'},'location','southeast');
```



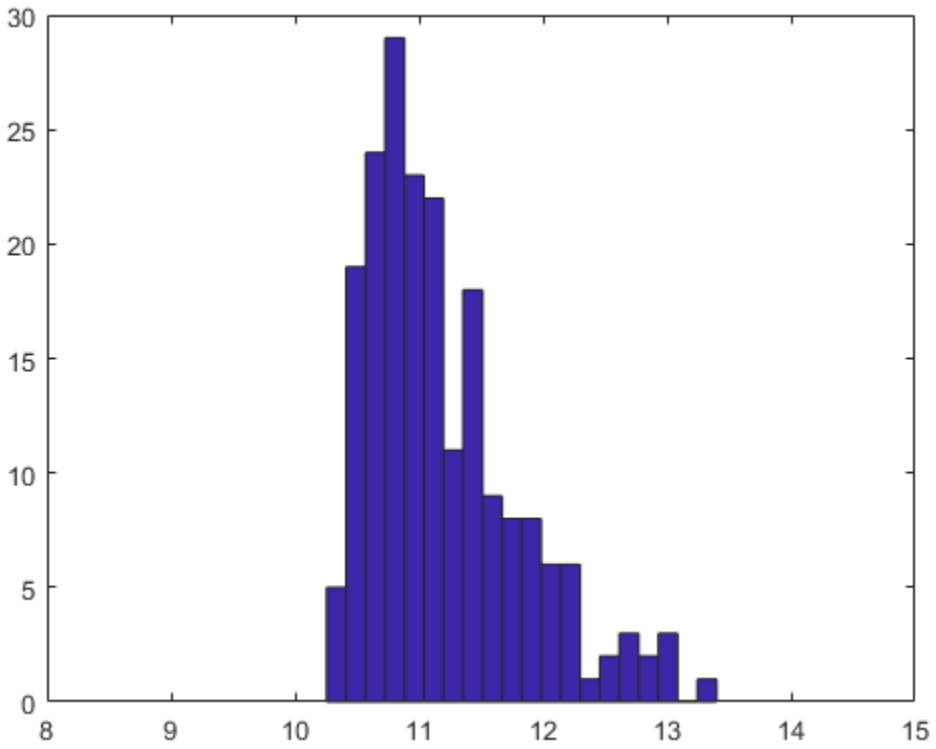
Notice how the lognormal fit differs systematically from the gamma fit in the tails. It grows more slowly in the left tail, and dies more slowly in the right tail. The gamma seems to be a slightly better fit to the data.

A Lognormal Threshold Parameter Example

The lognormal distribution is simple to fit by maximum likelihood, because once the log transformation is applied to the data, maximum likelihood is identical to fitting a normal. But it is sometimes necessary to estimate a threshold parameter in a lognormal model. The likelihood for such a model is unbounded, and so maximum likelihood does not work. However, the least squares method provides a way to make estimates. Since the two-parameter lognormal distribution can be log-transformed to a location-scale family, we could follow the same steps as in the earlier example that showed fitting a Weibull distribution with threshold parameter. Here, however, we'll do the estimation on the cumulative probability scale, as in the previous example showing a fit with the gamma distribution.

To illustrate, we'll simulate some data from a three-parameter lognormal distribution, with a threshold.

```
n = 200;
x = lognrnd(0,.5,n,1) + 10;
hist(x,20); xlim([8 15]);
```

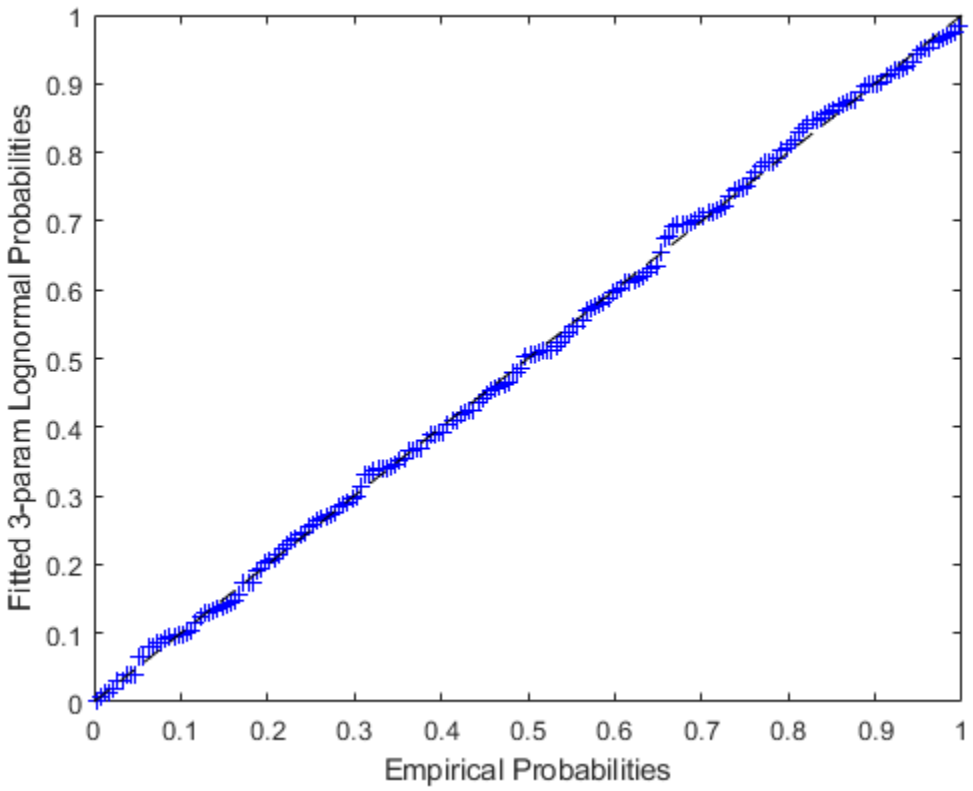


Compute the ECDF of x, and find the parameters for the best-fit three-parameter lognormal distribution.

```
x = sort(x);
pEmp = ((1:n)-0.5)' ./ n;
wgt = 1 ./ sqrt(pEmp.*(1-pEmp));
LN3obj = @(params) sum(wgt.*(logncdf(x-params(3),params(1),exp(params(2)))-pEmp).^2);
c0 = .99*min(x);
mu0 = mean(log(x-c0)); sigma0 = std(log(x-c0));
paramHat = fminsearch(LN3obj,[mu0,log(sigma0),c0]);
paramHat(2) = exp(paramHat(2))
```

```
paramHat =

   -0.0698    0.5930   10.1045
```

```
pFit = logncdf(x-paramHat(3),paramHat(1),paramHat(2));
plot(pEmp,pFit,'b+', [0 1],[0 1],'k--');
xlabel('Empirical Probabilities');
ylabel('Fitted 3-param Lognormal Probabilities');
```



Measures of Precision

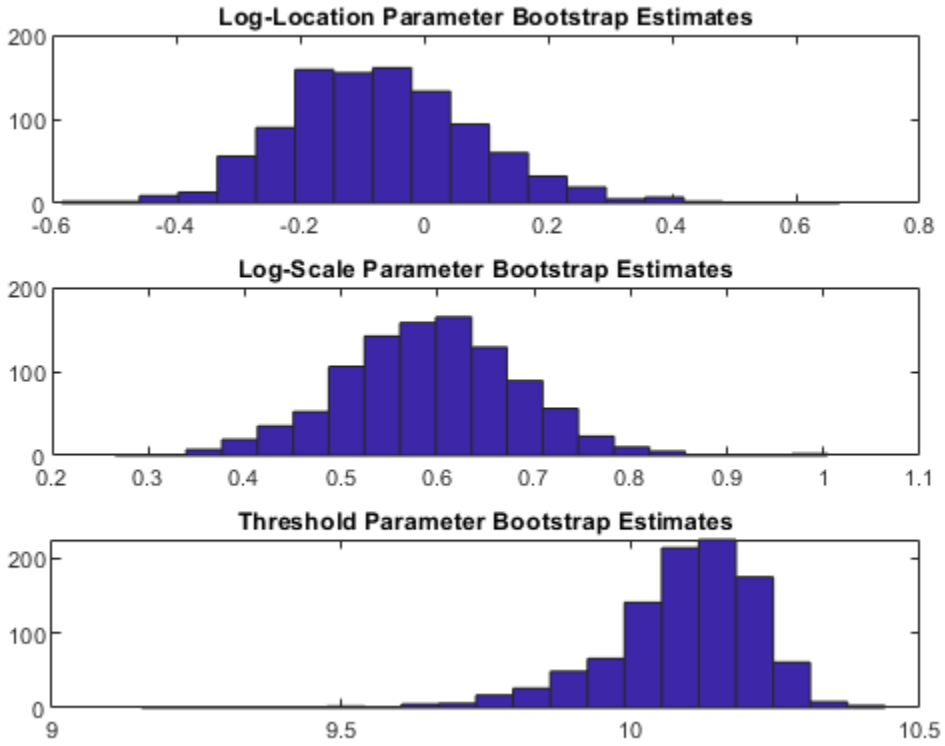
Parameter estimates are only part of the story -- a model fit also needs some measure of how precise the estimates are, typically standard errors. With maximum likelihood, the usual method is to use the information matrix and a large-sample asymptotic argument to approximate the covariance matrix of the estimator over repeated sampling. No such theory exists for these least squares estimators.

However, Monte-Carlo simulation provides another way to estimate standard errors. If we use the fitted model to generate a large number of datasets, we can approximate the standard error of the estimators with the Monte-Carlo standard deviation. For simplicity, we've defined a fitting function in a separate file, `logn3fit.m`.

```
estsSim = zeros(1000,3);
for i = 1:size(estsSim,1)
    xSim = lognrnd(paramHat(1),paramHat(2),n,1) + paramHat(3);
    estsSim(i,:) = logn3fit(xSim);
end
std(estsSim)


ans =

    0.1542    0.0908    0.1303
```

It might also be useful to look at the distribution of the estimates, to check if the assumption of approximate normality is reasonable for this sample size, or to check for bias.

```
subplot(3,1,1), hist(estsSim(:,1),20);
title('Log-Location Parameter Bootstrap Estimates');
subplot(3,1,2), hist(estsSim(:,2),20);
title('Log-Scale Parameter Bootstrap Estimates');
subplot(3,1,3), hist(estsSim(:,3),20);
title('Threshold Parameter Bootstrap Estimates');
```



Clearly, the estimator for the threshold parameter is skewed. This is to be expected, since it is bounded above by the minimum data value. The other two histograms indicate that approximate normality might be a questionable assumption for the log-location parameter (the first histogram) as well. The standard errors computed above must be interpreted with that in mind, and the usual construction for confidence intervals might not be appropriate for the log-location and threshold parameters.

The means of the simulated estimates are close to the parameter values used to generate simulated data, indicating that the procedure is approximately unbiased at this sample size, at least for parameter values near the estimates.

```
[paramHat; mean(estsSim)]
```

```
ans =

   -0.0698    0.5930   10.1045
   -0.0690    0.5926   10.0905
```

Finally, we could also have used the function `bootstrp` to compute bootstrap standard error estimates. These do not make any parametric assumptions about the data.

```
estsBoot = bootstrp(1000,@logn3fit,x);
std(estsBoot)
```

```
ans =

    0.1490    0.0785    0.1180
```

The bootstrap standard errors are not far off from the Monte-Carlo calculations. That's not surprising, since the fitted model is the same one from which the example data were generated.

Summary

The fitting method described here is an alternative to maximum likelihood that can be used to fit univariate distributions when maximum likelihood fails to provide useful parameter estimates. One important application is in fitting distributions involving a threshold parameter, such as the three-parameter lognormal. Standard errors are more difficult to compute than for maximum likelihood estimates, because analytic approximations do not exist, but simulation provides a feasible alternative.

The P-P plots used here to illustrate the fitting method are useful in their own right, as a visual indication of lack of fit when fitting a univariate distribution.

Gaussian Processes

Gaussian Process Regression Models

Gaussian process regression (GPR) models are nonparametric kernel-based probabilistic models. You can train a GPR model using the `fitrgp` function.

Consider the training set $\{(x_i, y_i); i = 1, 2, ..., n\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, drawn from an unknown distribution. A GPR model addresses the question of predicting the value of a response variable y_{new}, given the new input vector x_{new}, and the training data. A linear regression model is of the form

$$y = x^T\beta + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. The error variance σ^2 and the coefficients β are estimated from the data. A GPR model explains the response by introducing latent variables, $f(x_i)$, $i = 1, 2, ..., n$, from a Gaussian process (GP), and explicit basis functions, h. The covariance function of the latent variables captures the smoothness of the response and basis functions project the inputs x into a p-dimensional feature space.

A GP is a set of random variables, such that any finite number of them have a joint Gaussian distribution. If $\left\{f(x), x \in \mathbb{R}^d\right\}$ is a GP, then given n observations $x_1, x_2, ..., x_n$, the joint distribution of the random variables $f(x_1), f(x_2), ..., f(x_n)$ is Gaussian. A GP is defined by its mean function $m(x)$ and covariance function, $k(x, x')$. That is, if $\left\{f(x), x \in \mathbb{R}^d\right\}$ is a Gaussian process, then $E(f(x)) = m(x)$ and $Cov[f(x), f(x')] = E[\{f(x) - m(x)\}\{f(x') - m(x')\}] = k(x, x')$.

Now consider the following model.

$$h(x)^T\beta + f(x),$$

where $f(x) \sim GP(0, k(x, x'))$, that is $f(x)$ are from a zero mean GP with covariance function, $k(x, x')$. $h(x)$ are a set of basis functions that transform the original feature vector x in R^d into a new feature vector $h(x)$ in R^p. β is a p-by-1 vector of basis function coefficients. This model represents a GPR model. An instance of response y can be modeled as

$$P(y_i|f(x_i), x_i) \sim N\left(y_i \middle| h(x_i)^T\beta + f(x_i), \sigma^2\right)$$

Hence, a GPR model is a probabilistic model. There is a latent variable $f(x_i)$ introduced for each observation x_i, which makes the GPR model nonparametric. In vector form, this model is equivalent to

$$P(y \mid f, X) \sim N(y \mid H\beta + f, \sigma^2 I),$$

where

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad H = \begin{pmatrix} h\left(x_1^T\right) \\ h\left(x_2^T\right) \\ \vdots \\ h\left(x_n^T\right) \end{pmatrix}, \quad f = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix}.$$

The joint distribution of latent variables $f(x_1)$, $f(x_2)$, ..., $f(x_n)$ in the GPR model is as follows:

$$P(f \mid X) \sim N(f \mid 0, K(X, X)),$$

close to a linear regression model, where $K(X,X)$ looks as follows:

$$K(X,X) = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n) \end{pmatrix}.$$

The covariance function $k(x,x')$ is usually parameterized by a set of kernel parameters or hyperparameters, θ. Often $k(x,x')$ is written as $k(x,x'|\theta)$ to explicitly indicate the dependence on θ.

`fitrgp` estimates the basis function coefficients, β, the noise variance, σ^2, and the hyperparameters, θ, of the kernel function from the data while training the GPR model. You can specify the basis function, the kernel (covariance) function, and the initial values for the parameters.

Because a GPR model is probabilistic, it is possible to compute the prediction intervals using the trained model (see `predict` and `resubPredict`).

You can also compute the regression error using the trained GPR model (see `loss` and `resubLoss`).

Compare Prediction Intervals of GPR Models

This example fits GPR models to a noise-free data set and a noisy data set. The example compares the predicted responses and prediction intervals of the two fitted GPR models.

Generate two observation data sets from the function $g(x) = x \cdot \sin(x)$.

```
rng('default') % For reproducibility
x_observed = linspace(0,10,21)';
y_observed1 = x_observed.*sin(x_observed);
y_observed2 = y_observed1 + 0.5*randn(size(x_observed));
```

The values in `y_observed1` are noise free, and the values in `y_observed2` include some random noise.

Fit GPR models to the observed data sets.

```
gprMdl1 = fitrgp(x_observed,y_observed1);
gprMdl2 = fitrgp(x_observed,y_observed2);
```

Compute the predicted responses and 95% prediction intervals using the fitted models.

```
x = linspace(0,10)';
[ypred1,~,yint1] = predict(gprMdl1,x);
[ypred2,~,yint2] = predict(gprMdl2,x);
```

Resize a figure to display two plots in one figure.

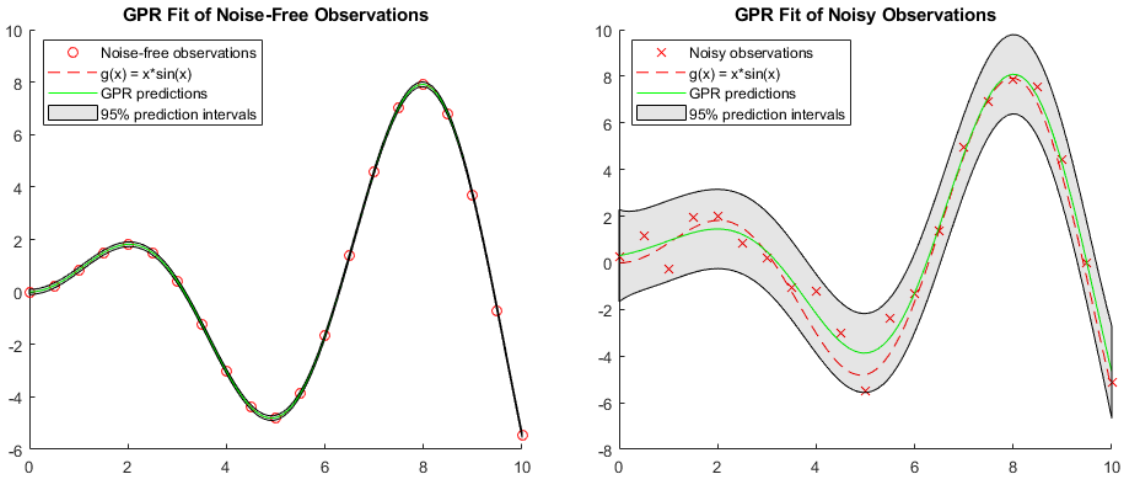
```
fig = figure;
fig.Position(3) = fig.Position(3)*2;
```

Create a 1-by-2 tiled chart layout.

```
tiledlayout(1,2,'TileSpacing','compact')
```

For each tile, draw a scatter plot of observed data points and a function plot of $x \cdot \sin(x)$. Then add a plot of GP predicted responses and a patch of prediction intervals.

```
nexttile
hold on
scatter(x_observed,y_observed1,'r') % Observed data points
fplot(@(x) x.*sin(x),[0,10],'--r')  % Function plot of x*sin(x)
plot(x,ypred1,'g')                  % GPR predictions
patch([x;flipud(x)],[yint1(:,1);flipud(yint1(:,2))],'k','FaceAlpha',0.1); % Prediction intervals
hold off
title('GPR Fit of Noise-Free Observations')
legend({'Noise-free observations','g(x) = x*sin(x)','GPR predictions','95% prediction intervals'}

nexttile
hold on
scatter(x_observed,y_observed2,'xr') % Observed data points
fplot(@(x) x.*sin(x),[0,10],'--r')   % Function plot of x*sin(x)
plot(x,ypred2,'g')                   % GPR predictions
patch([x;flipud(x)],[yint2(:,1);flipud(yint2(:,2))],'k','FaceAlpha',0.1); % Prediction intervals
hold off
title('GPR Fit of Noisy Observations')
legend({'Noisy observations','g(x) = x*sin(x)','GPR predictions','95% prediction intervals'},'Lo
```



When the observations are noise free, the predicted responses of the GPR fit cross the observations. The standard deviation of the predicted response is almost zero. Therefore, the prediction intervals are very narrow. When observations include noise, the predicted responses do not cross the observations, and the prediction intervals become wide.

References

[1] Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press. Cambridge, Massachusetts, 2006.

See Also

`RegressionGP` | `fitrgp` | `predict`

More About

- “Exact GPR Method” on page 6-10
- “Subset of Data Approximation for GPR Models” on page 6-14
- “Subset of Regressors Approximation for GPR Models” on page 6-15
- “Fully Independent Conditional Approximation for GPR Models” on page 6-19
- “Block Coordinate Descent Approximation for GPR Models” on page 6-22

Kernel (Covariance) Function Options

In supervised learning, it is expected that the points with similar predictor values x_i, naturally have close response (target) values y_i. In Gaussian processes, the covariance function expresses this similarity [1]. It specifies the covariance between the two latent variables $f(x_i)$ and $f(x_j)$, where both x_i and x_j are d-by-1 vectors. In other words, it determines how the response at one point x_i is affected by responses at other points x_j, $i \neq j$, $i = 1, 2, \ldots, n$. The covariance function $k(x_i, x_j)$ can be defined by various kernel functions. It can be parameterized in terms of the kernel parameters in vector θ. Hence, it is possible to express the covariance function as $k(x_i, x_j|\theta)$.

For many standard kernel functions, the kernel parameters are based on the signal standard deviation σ_f and the characteristic length scale σ_l. The characteristic length scales briefly define how far apart the input values x_i can be for the response values to become uncorrelated. Both σ_l and σ_f need to be greater than 0, and this can be enforced by the unconstrained parametrization vector θ, such that

$$\theta_1 = \log\sigma_l, \quad \theta_2 = \log\sigma_f .$$

The built-in kernel (covariance) functions *with same length scale for each predictor* are:

- **Squared Exponential Kernel**

  This is one of the most commonly used covariance functions and is the default option for `fitrgp`. The squared exponential kernel function is defined as

  $$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left[-\frac{1}{2}\frac{(x_i - x_j)^T(x_i - x_j)}{\sigma_l^2}\right].$$

  where σ_l is the characteristic length scale, and σ_f is the signal standard deviation.

- **Exponential Kernel**

  You can specify the exponential kernel function using the `'KernelFunction','exponential'` name-value pair argument. This covariance function is defined by

  $$k(x_i, x_j|\theta) = \sigma_f^2 \exp\left(-\frac{r}{\sigma_l}\right),$$

  where σ_l is the characteristic length scale and

  $$r = \sqrt{(x_i - x_j)^T(x_i - x_j)}$$

  is the Euclidean distance between x_i and x_j.

- **Matern 3/2**

  You can specify the Matern 3/2 kernel function using the `'KernelFunction','matern32'` name-value pair argument. This covariance function is defined by

  $$k(x_i, x_j|\theta) = \sigma_f^2\left(1 + \frac{\sqrt{3}r}{\sigma_l}\right)\exp\left(-\frac{\sqrt{3}r}{\sigma_l}\right),$$

  where

$$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$$

is the Euclidean distance between x_i and x_j.

- **Matern 5/2**

  You can specify the Matern 5/2 kernel function using the `'KernelFunction','matern52'` name-value pair argument. The Matern 5/2 covariance function is defined as

  $$k(x_i, x_j) = \sigma_f^2 \left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right) \exp\left(-\frac{\sqrt{5}r}{\sigma_l}\right),$$

  where

  $$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$$

  is the Euclidean distance between x_i and x_j.

- **Rational Quadratic Kernel**

  You can specify the rational quadratic kernel function using the `'KernelFunction','rationalquadratic'` name-value pair argument. This covariance function is defined by

  $$k(x_i, x_j \mid \theta) = \sigma_f^2 \left(1 + \frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha},$$

  where σ_l is the characteristic length scale, α is a positive-valued scale-mixture parameter, and

  $$r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$$

  is the Euclidean distance between x_i and x_j.

It is possible to use a separate length scale σ_m for each predictor m, $m = 1, 2, \ldots, d$. The built-in kernel (covariance) functions with a separate length scale for each predictor implement automatic relevance determination (ARD) [2]. The unconstrained parametrization θ in this case is

$$\theta_m = \log \sigma_m, \quad \text{for } m = 1, 2, \ldots, d$$
$$\theta_{d+1} = \log \sigma_f .$$

The built-in kernel (covariance) functions *with separate length scale for each predictor* are:

- **ARD Squared Exponential Kernel**

  You can specify this kernel function using the `'KernelFunction','ardsquaredexponential'` name-value pair argument. This covariance function is the squared exponential kernel function, with a separate length scale for each predictor. It is defined as

  $$k(x_i, x_j \mid \theta) = \sigma_f^2 \exp\left[-\frac{1}{2} \sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2}\right].$$

- **ARD Exponential Kernel**

You can specify this kernel function using the `'KernelFunction','ardexponential'` name-value pair argument. This covariance function is the exponential kernel function, with a separate length scale for each predictor. It is defined as

$$k(x_i, x_j \mid \theta) = \sigma_f^2 \exp(-r),$$

where

$$r = \sqrt{\sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2}}.$$

- **ARD Matern 3/2**

  You can specify this kernel function using the `'KernelFunction','ardmatern32'` name-value pair argument. This covariance function is the Matern 3/2 kernel function, with a different length scale for each predictor. It is defined as

  $$k(x_i, x_j \mid \theta) = \sigma_f^2 (1 + \sqrt{3}\, r) \exp(-\sqrt{3}\, r),$$

  where

  $$r = \sqrt{\sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2}}.$$

- **ARD Matern 5/2**

  You can specify this kernel function using the `'KernelFunction','ardmatern52'` name-value pair argument. This covariance function is the Matern 5/2 kernel function, with a different length scale for each predictor. It is defined as

  $$k(x_i, x_j \mid \theta) = \sigma_f^2 \left(1 + \sqrt{5}\, r + \frac{5}{3} r^2\right) \exp(-\sqrt{5}\, r),$$

  where

  $$r = \sqrt{\sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2}}.$$

- **ARD Rational Quadratic Kernel**

  You can specify this kernel function using the `'KernelFunction','ardrationalquadratic'` name-value pair argument. This covariance function is the rational quadratic kernel function, with a separate length scale for each predictor. It is defined as

  $$k(x_i, x_j \mid \theta) = \sigma_f^2 \left(1 + \frac{1}{2\alpha} \sum_{m=1}^{d} \frac{(x_{im} - x_{jm})^2}{\sigma_m^2}\right)^{-\alpha}.$$

You can specify the kernel function using the `KernelFunction` name-value pair argument in a call to `fitrgp`. You can either specify one of the built-in kernel parameter options, or specify a custom function. When providing the initial kernel parameter values for a built-in kernel function, input the initial values for signal standard deviation and the characteristic length scale(s) as a numeric vector. When providing the initial kernel parameter values for a custom kernel function, input the initial values the unconstrained parametrization vector θ. `fitrgp` uses analytical derivatives to estimate

parameters when using a built-in kernel function, whereas when using a custom kernel function it uses numerical derivatives.

References

[1] Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press. Cambridge, Massachusetts, 2006.

[2] Neal, R. M. *Bayesian Learning for Neural Networks.* Springer, New York. Lecture Notes in Statistics, 118, 1996.

See Also

`RegressionGP` | `fitrgp`

Related Examples

- “Use Separate Length Scales for Predictors” on page 33-2141
- “Fit GPR Model Using Custom Kernel Function” on page 33-2150
- “Impact of Specifying Initial Kernel Parameter Values” on page 33-2139

Exact GPR Method

In this section...

An instance of response y from a Gaussian process regression (GPR) model on page 6-2 can be modeled as

$$P(y_i|f(x_i),x_i) \sim N\left(y_i\middle|h(x_i)^T\beta + f(x_i),\sigma^2\right)$$

Hence, making predictions for new data from a GPR model requires:

- Knowledge of the coefficient vector, β, of fixed basis functions
- Ability to evaluate the covariance function $k(x,x'|\theta)$ for arbitrary x and x', given the kernel parameters or hyperparameters, θ.
- Knowledge of the noise variance σ^2 that appears in the density $P(y_i|f(x_i),x_i)$

That is, one needs first to estimate β, θ, and σ^2 from the data (X, y).

Parameter Estimation

One approach for estimating the parameters β, θ, and σ^2 of a GPR model is by maximizing the likelihood $P(y\,|\,X)$ as a function of β, θ, and σ^2[1]. That is, if $\widehat{\beta}$, $\widehat{\theta}$, and $\widehat{\sigma}^2$ are the estimates of β, θ, and σ^2, respectively, then:

$$\widehat{\beta},\widehat{\theta},\widehat{\sigma}^2 = \underset{\beta,\theta,\sigma^2}{\arg\max}\log P\left(y\middle|X,\beta,\theta,\sigma^2\right).$$

Because

$$P(y|X) = P\left(y\middle|X,\beta,\theta,\sigma^2\right) = \mathrm{N}\left(y\middle|H\beta,K(X,X|\theta)+\sigma^2 I_n\right),$$

the marginal log likelihood function is as follows:

$$\log P\left(y\middle|X,\beta,\theta,\sigma^2\right) = -\frac{1}{2}(y-H\beta)^T\left[K(X,X|\theta)+\sigma^2 I_n\right]^{-1}(y-H\beta) - \frac{n}{2}\log 2\pi - \frac{1}{2}\log\left|K(X,X|\theta)+\sigma^2 I_n\right|.$$

where H is the vector of explicit basis functions, and $K(X,X|\theta)$ is the covariance function matrix (for more information, see “Gaussian Process Regression Models” on page 6-2).

To estimate the parameters, the software first computes $\widehat{\beta}\left(\theta,\sigma^2\right)$, which maximizes the log likelihood function with respect to β for given θ and σ^2. It then uses this estimate to compute the β-profiled likelihood:

$$\log\left\{P\left(y\middle|X,\widehat{\beta}\left(\theta,\sigma^2\right),\theta,\sigma^2\right)\right\}.$$

The estimate of β for given θ, and σ^2 is

$$\widehat{\beta}\left(\theta,\sigma^2\right)=\left[\,H^T\left[K(X,X\,|\,\theta)+\sigma^2I_n\right]^{-1}H\right]^{-1}H^T\left[K(X,X\,|\,\theta)+\sigma^2I_n\right]^{-1}y\,.$$

Then, the β-profiled log likelihood is given by

$$\begin{aligned}\log P\left(y\middle|X,\widehat{\beta}\,(\theta,\sigma^2),\theta,\sigma^2\right)=&-\frac{1}{2}\left(y-H\widehat{\beta}\left(\theta,\sigma^2\right)\right)^T\left[K(X,X\,|\,\theta)+\sigma^2I_n\right]^{-1}\left(y-H\widehat{\beta}\left(\theta,\sigma^2\right)\right)\\&-\frac{n}{2}\log 2\pi-\frac{1}{2}\log\left|K(X,X\,|\,\theta)+\sigma^2I_n\right|\end{aligned}$$

The software then maximizes the β-profiled log-likelihood over θ, and σ^2 to find their estimates.

Prediction

Making probabilistic predictions from a GPR model with known parameters requires the density $P(y_{new}\,|\,y,X,x_{new})$. Using the definition of conditional probabilities, one can write:

$$P(y_{new}\,|\,y,X,x_{new})=\frac{P(y_{new},y\,|\,X,x_{new})}{P(y\,|\,X,x_{new})}\,.$$

To find the joint density in the numerator, it is necessary to introduce the latent variables f_{new} and f corresponding to y_{new}, and y, respectively. Then, it is possible to use the joint distribution for y_{new}, y, f_{new}, and f to compute $P(y_{new},y|X,x_{new})$:

$$\begin{aligned}P(y_{new},y|X,x_{new})&=\int\int P(y_{new},y,f_{new},f|X,x_{new})dfdf_{new}\\&=\int\int P(y_{new},y|f_{new},f,X,x_{new})P(f_{new},f|X,x_{new})dfdf_{new}\,.\end{aligned}$$

Gaussian process models assume that each response y_i only depends on the corresponding latent variable f_i and the feature vector x_i. Writing $P(y_{new},y|f_{new},f,X,x_{new})$ as a product of conditional densities and based on this assumption produces:

$$P(y_{new},y|f_{new},f,X,x_{new})=P(y_{new}|f_{new},x_{new})\prod_{i=1}^{n}P(y_i|f(x_i),x_i)\,.$$

After integrating with respect to y_{new}, the result only depends on f and X:

$$P(y|f,X)=\prod_{i=1}^{n}P(y_i|f_i,x_i)=\prod_{i=1}^{n}\mathrm{N}\left(y_i|h(x_i)^T\beta+f_i,\sigma^2\right).$$

Hence,

$$P(y_{new},\ y|f_{new},\ f,\ X,\ x_{new})=P(y_{new}|f_{new},\ x_{new})P(y|f,X)\,.$$

Again using the definition of conditional probabilities,

$$P(f_{new},f|X,x_{new})=P(f_{new}|f,X,x_{new})*P(f|X,x_{new}),$$

it is possible to write $P(y_{new}, y|X, x_{new})$ as follows:

$$P(y_{new}, y|X, x_{new}) = \int\int P(y_{new}|f_{new},\ x_{new})P(y|f, X)P(f_{new}|f, X, x_{new})P(f|X, x_{new})dfdf_{new}\,.$$

Using the facts that

$$P(f|X, x_{new}) = P(f|X)$$

and

$$P(y|f, X)P(f|X) = P(y, f|X) = P(f|y, X)P(y|X),$$

one can rewrite $P(y_{new}, y|X, x_{new})$ as follows:

$$P(y_{new}, y|X, x_{new}) = P(y|X)\int\int P(y_{new}|f_{new},\ x_{new})P(f|y, X)P(f_{new}|f, X, x_{new})dfdf_{new}\,.$$

It is also possible to show that

$$P(y|X, x_{new}) = P(y|X)\,.$$

Hence, the required density $P(y_{new}\,|\,y, X, x_{new})$ is:

$$\begin{aligned} P(y_{new}|y, X, x_{new}) &= \frac{P(y_{new}, y\,|\,X, x_{new})}{P(y\,|\,X, x_{new})} = \frac{P(y_{new}, y\,|\,X, x_{new})}{P(y\,|\,X)} \\ &= \int\int \underbrace{P(y_{new}|f_{new},\ x_{new})}_{(1)}\underbrace{P(f|y, X)}_{(2)}\underbrace{P(f_{new}|f, X, x_{new})}_{(3)}dfdf_{new}\,. \end{aligned}$$

It can be shown that

(1) $P(y_{new}|f_{new}, x_{new}) = \mathrm{N}\left(y_{new}\middle|h(x_{new})^T\beta + f_{new}, \sigma^2_{new}\right)$

(2) $P(f|y, X) = \mathrm{N}\left(f\middle|\frac{1}{\sigma^2}\left(\frac{I_n}{\sigma^2} + K(X, X)^{-1}\right)^{-1}(y - H\beta), \left(\frac{I_n}{\sigma^2} + K(X, X)^{-1}\right)^{-1}\right)$

(3) $P(f_{new}|f, X, x_{new}) = \mathrm{N}\left(f_{new}\middle|K\left(x^T_{new}, X\right)K(X, X)^{-1}f, \Delta\right)$,

where $\Delta = k(x_{new}, x_{new}) - K\left(x^T_{new}, X\right)K(X, X)^{-1}K\left(X, x^T_{new}\right)$.

After the integration and required algebra, the density of the new response y_{new} at a new point x_{new}, given y, X is found as

$$P(y_{new}|y, X, x_{new}) = \mathrm{N}\left(y_{new}\middle|h(x_{new})^T\beta + \mu, \sigma^2_{new} + \Sigma\right),$$

where

$$\mu = K\left(x^T_{new}, X\right)\underbrace{\left(K(X, X) + \sigma^2 I_n\right)^{-1}}_{\alpha}(y - H\beta)$$

and

$$\Sigma = k(x_{new}, x_{new}) - K\left(x^T_{new}, X\right)\left(K(X, X) + \sigma^2 I_n\right)^{-1}K\left(X, x^T_{new}\right).$$

The expected value of prediction y_{new} at a new point x_{new} given y, X, and parameters β, θ, and σ^2 is

$$\begin{aligned} E\left(y_{new}|y,\ X, x_{new}, \beta, \theta, \sigma^2\right) &= h(x_{new})^T\beta + K\left(x_{new}^T, X\middle|\theta\right)\alpha \\ &= h(x_{new})^T\beta + \sum_{i=1}^{n} \alpha_i k(x_{new}, x_i|\theta), \end{aligned}$$

where

$$\alpha = \left(K(X,X|\theta) + \sigma^2 I_n\right)^{-1}(y - H\beta)\,.$$

Computational Complexity of Exact Parameter Estimation and Prediction

Training a GPR model with the exact method (when `FitMethod` is `'Exact'`) requires the inversion of an n-by-n kernel matrix $K(X,X)$. The memory requirement for this step scales as $O(n^2)$ since $K(X,X)$ must be stored in memory. One evaluation of $\log P(y|X)$ scales as $O(n^3)$. Therefore, the computational complexity is $O(kn^3)$, where k is the number of function evaluations needed for maximization and n is the number of observations.

Making predictions on new data involves the computation of $\widehat{\alpha}$. If prediction intervals are desired, this step could also involve the computation and storage of the Cholesky factor of $\left(K(X,X) + \sigma^2 I_n\right)$ for later use. The computational complexity of this step using the direct computation of $\widehat{\alpha}$ is $O(n^3)$ and the memory requirement is $O(n^2)$.

Hence, for large n, estimation of parameters or computing predictions can be very expensive. The approximation methods usually involve rearranging the computation so as to avoid the inversion of an n-by-n matrix. For the available approximation methods, please see the related links at the bottom of the page.

References

[1] Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press. Cambridge, Massachusetts, 2006.

See Also

`fitrgp` | `predict`

More About

- “Gaussian Process Regression Models” on page 6-2
- “Subset of Data Approximation for GPR Models” on page 6-14
- “Subset of Regressors Approximation for GPR Models” on page 6-15
- “Fully Independent Conditional Approximation for GPR Models” on page 6-19
- “Block Coordinate Descent Approximation for GPR Models” on page 6-22

Subset of Data Approximation for GPR Models

Training a GPR model with the exact method (when `FitMethod` is `'Exact'`) requires the inversion of an n-by-n matrix. Therefore, the computational complexity is $O(kn^3)$, where k is the number of function evaluations required for estimating β, θ, and σ^2, and n is the number of observations. For large n, estimation of parameters or computing predictions can be very expensive.

One simple way to solve the computational complexity problem with large data sets is to select $m < n$ observations out of n and then apply exact GPR model on page 6-10 to these m points to estimate β, θ, and σ^2 while ignoring the other $(n – m)$ points. This smaller subset is known as the *active set*. And this approximation method is called the Subset of Data (SD) method.

The computational complexity when using SD method is $O(km^3)$, where k is the number of function evaluations and m is the active set size. The storage requirements are $O(m^2)$ since only a part of the full kernel matrix $K(X, X|\theta)$ needs to be stored in memory.

You can specify the SD method for parameter estimation by using the `'FitMethod','sd'` name-value pair argument in the call to `fitrgp`. To specify the SD method for prediction, use the `'PredictMethod','sd'` name-value pair argument.

For estimating parameters using the exact GPR model, see parameter estimation using the exact GPR method on page 6-10. For making predictions using the exact GPR model, see prediction using the exact GPR method on page 6-11.

See Also

`fitrgp` | `predict`

More About

- “Gaussian Process Regression Models” on page 6-2
- “Exact GPR Method” on page 6-10
- “Subset of Regressors Approximation for GPR Models” on page 6-15
- “Fully Independent Conditional Approximation for GPR Models” on page 6-19
- “Block Coordinate Descent Approximation for GPR Models” on page 6-22

Subset of Regressors Approximation for GPR Models

In this section...

The subset of regressors (SR) approximation method consists of replacing the kernel function $k(x, x_r|\theta)$ in the exact GPR method on page 6-10 by its approximation $\widehat{k}_{SR}(x, x_r|\theta, \mathrm{A})$, given the active set $\mathrm{A} \subset \mathrm{N} = \{1, 2, ..., n\}$. You can specify the SR method for parameter estimation by using the `'FitMethod','sr'` name-value pair argument in the call to `fitrgp`. For prediction using SR, you can use the `'PredictMethod','sr'` name-value pair argument in the call to `fitrgp`.

Approximating the Kernel Function

For the exact GPR model on page 6-11, the expected prediction in GPR depends on the set of N functions $\mathrm{S_N} = \{k(x, x_i|\theta), i = 1, 2, ..., n\}$, where $\mathrm{N} = \{1, 2, ..., n\}$ is the set of indices of all observations, and n is the total number of observations. The idea is to approximate the span of these functions by a smaller set of functions, $\mathrm{S_A}$, where $\mathrm{A} \subset \mathrm{N} = \{1, 2, ..., n\}$ is the subset of indices of points selected to be in the active set. Consider $\mathrm{S_A} = \left\{k\left(x, x_j\middle|\theta\right), j \in \mathrm{A}\right\}$. The aim is to approximate the elements of $\mathrm{S_N}$ as linear combinations of the elements of $\mathrm{S_A}$.

Suppose the approximation to $k(x, x_r|\theta)$ using the functions in $\mathrm{S_A}$ is as follows:

$$\widehat{k}\,(x, x_r|\theta) = \sum_{j \in \mathrm{A}} c_{jr} k\left(x, x_j\middle|\theta\right),$$

where $c_{jr} \in \mathbb{R}$ are the coefficients of the linear combination for approximating $k(x, x_r|\theta)$. Suppose C is the matrix that contains all the coefficients c_{jr}. Then, C, is a $|\mathrm{A}| \times n$ matrix such that $C(j, r) = c_{jr}$. The software finds the best approximation to the elements of $\mathrm{S_N}$ using the active set $\mathrm{A} \subset \mathrm{N} = \{1, 2, ..., n\}$ by minimizing the error function

$$E(\mathrm{A}, C) = \sum_{r=1}^{n} \left\| k(x, x_r|\theta) - \widehat{k}\,(x, x_r|\theta) \right\|^2_{\mathscr{H}},$$

where $\mathscr{H}$ is the Reproducing Kernel Hilbert Spaces (RKHS) associated with the kernel function k [1], [2].

The coefficient matrix that minimizes $E(\mathrm{A}, C)$ is

$$\widehat{C}_{\mathrm{A}} = K(X_{\mathrm{A}}, X_{\mathrm{A}}|\theta)^{-1} K(X_{\mathrm{A}}, X|\theta),$$

and an approximation to the kernel function using the elements in the active set $\mathrm{A} \subset \mathrm{N} = \{1, 2, ..., n\}$ is

$$\widehat{k}\,(x, x_r|\theta) = \sum_{j \in \mathrm{A}} c_{jr} k\left(x, x_j\middle|\theta\right) = K\left(x^T, X_{\mathrm{A}}\middle|\theta\right) C(:, r)\,.$$

The SR approximation to the kernel function using the active set $A \subset N = \{1, 2, ..., n\}$ is defined as:

$$\widehat{k}_{SR}(x, x_r|\theta, A) = K\left(x^T, X_A\middle|\theta\right)\widehat{C}_A(:, r) = K\left(x^T, X_A\middle|\theta\right)K(X_A, X_A|\theta)^{-1}K\left(X_A, x_r^T\middle|\theta\right)$$

and the SR approximation to $K(X, X|\theta)$ is:

$$\widehat{K}_{SR}(X, X|\theta, A) = \ K(X, X_A|\theta)\, K(X_A, X_A|\theta)^{-1}\, K(X_A, X|\theta)\,.$$

Parameter Estimation

Replacing $K(X, X|\theta)$ by $\widehat{K}_{SR}(X, X|\theta, A)$ in the marginal log likelihood function produces its SR approximation:

$$\log P_{SR}\left(y\middle|X, \beta, \theta, \sigma^2, A\right) = -\frac{1}{2}(y - H\beta)^T\left[\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right]^{-1}(y - H\beta)$$
$$-\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left|\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right|$$

As in the exact method on page 6-10, the software estimates the parameters by first computing $\widehat{\beta}\left(\theta, \sigma^2\right)$, the optimal estimate of β, given θ and σ^2. Then it estimates θ, and σ^2 using the β-profiled marginal log likelihood. The SR estimate to β for given θ, and σ^2 is:

$$\widehat{\beta}_{SR}\left(\theta, \sigma^2, A\right) = \left[H^T\left[\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right]^{-1}H\right]^{-1}_{*} H^T\left[\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right]^{-1}y,$$

where

$$\left[\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right]^{-1} = \frac{I_N}{\sigma^2} - \frac{K(X, X_A|\theta)}{\sigma^2}A_A^{-1}\frac{K(X_A, X|\theta)}{\sigma^2},$$

$$A_A = K(X_A, X_A|\theta) + \frac{K(X_A, X|\theta)K(X, X_A|\theta)}{\sigma^2},$$

$$* = \frac{H^T H}{\sigma^2} - \frac{H^T K(X, X_A|\theta)}{\sigma^2}A_A^{-1}\frac{K(X_A, X|\theta)H}{\sigma^2},$$

$$** = \frac{H^T y}{\sigma^2} - \frac{H^T K(X, X_A|\theta)}{\sigma^2}A_A^{-1}\frac{K(X_A, X|\theta)y}{\sigma^2}.$$

And the SR approximation to the β-profiled marginal log likelihood is:

$$\log P_{SR}\left(y\middle|X, \widehat{\beta}_{SR}\left(\theta, \sigma^2, A\right), \theta, \sigma^2, A\right) =$$
$$-\frac{1}{2}\left(y - H\widehat{\beta}_{SR}\left(\theta, \sigma^2, A\right)\right)^T\left[\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right]^{-1}\left(y - H\widehat{\beta}_{SR}\left(\theta, \sigma^2, A\right)\right)$$
$$-\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left|\widehat{K}_{SR}(X, X|\theta, A) + \sigma^2 I_n\right|.$$

Prediction

The SR approximation to the distribution of y_{new} given y, X, x_{new} is

$$P(y_{new}|y, X, x_{new}) = \mathrm{N}\left(y_{new}\middle|h(x_{new})^T\beta + \mu_{SR}, \sigma^2_{new} + \Sigma_{SR}\right),$$

where μ_{SR} and Σ_{SR} are the SR approximations to μ and Σ shown in prediction using the exact GPR method on page 6-11.

μ_{SR} and Σ_{SR} are obtained by replacing $k(x, x_r|\theta)$ by its SR approximation $\widehat{k}_{SR}(x, x_r|\theta, \mathrm{A})$ in μ and Σ, respectively.

That is,

$$\mu_{SR} = \underbrace{\widehat{K}_{SR}\left(x^T_{new}, X\middle|\theta, \mathrm{A}\right)}_{(1)}\underbrace{\left(\widehat{K}_{SR}(X, X|\theta, \mathrm{A}) + \sigma^2 I_N\right)^{-1}}_{(2)}(y - H\beta).$$

Since

$$(1) = K\left(x^T_{new}, X_\mathrm{A}\middle|\theta\right) K(X_\mathrm{A}, X_\mathrm{A}|\theta)^{-1}K(X_\mathrm{A}, X|\theta),$$

$$(2) = \frac{I_N}{\sigma^2} - \frac{K(X, X_\mathrm{A}|\theta)}{\sigma^2}\left[K(X_\mathrm{A}, X_\mathrm{A}|\theta) + \frac{K(X_\mathrm{A}, X|\theta)\,K(X, X_\mathrm{A}|\theta)}{\sigma^2}\right]^{-1}\frac{K(X_\mathrm{A}, X|\theta)}{\sigma^2},$$

and from the fact that $I_N - B(A + B)^{-1} = A(A + B)^{-1}$, μ_{SR} can be written as

$$\mu_{SR} = K\left(x^T_{new}, X_\mathrm{A}\middle|\theta\right)\left[K(X_\mathrm{A}, X_\mathrm{A}|\theta) + \frac{K(X_\mathrm{A}, X|\theta)K(X, X_\mathrm{A}|\theta)}{\sigma^2}\right]^{-1}\frac{K(X_\mathrm{A}, X|\theta)}{\sigma^2}(y - H\beta).$$

Similarly, Σ_{SR} is derived as follows:

$$\Sigma_{SR} = \underbrace{\widehat{k}_{SR}(x_{new}, x_{new}|\theta, \mathrm{A})}_{*} - \underbrace{\widehat{K}_{SR}\left(x^T_{new}, X\middle|\theta, \mathrm{A}\right)}_{**}\underbrace{\left(\widehat{K}_{SR}(X, X|\theta, \mathrm{A}) + \sigma^2 I_N\right)^{-1}}_{***}\underbrace{\widehat{K}_{SR}\left(X, x^T_{new}\middle|\theta, \mathrm{A}\right)}_{****}.$$

Because

$$* = K\left(x^T_{new}, X_\mathrm{A}\middle|\theta\right)K(X_\mathrm{A}, X_\mathrm{A}|\theta)^{-1}K\left(X_\mathrm{A}, x^T_{new}\middle|\theta\right),$$

$$** = K\left(x^T_{new}, X_\mathrm{A}\middle|\theta\right)K(X_\mathrm{A}, X_\mathrm{A}|\theta)^{-1}K(X_\mathrm{A}, X|\theta),$$

$*** = (2)$ in the equation of μ_{SR},

$$**** = K(X, X_\mathrm{A}|\theta)K(X_\mathrm{A}, X_\mathrm{A}|\theta)^{-1}K\left(X_\mathrm{A}, x^T_{new}\middle|\theta\right),$$

Σ_{SR} is found as follows:

$$\sum_{SR} = K\left(x^T_{new}, X_\mathrm{A}\middle|\theta\right)\left[K(X_\mathrm{A}, X_\mathrm{A}|\theta) + \frac{K(X_\mathrm{A}, X|\theta)\,K(X, X_\mathrm{A}|\theta))}{\sigma^2}\right]^{-1}K\left(X_\mathrm{A}, x^T_{new}\middle|\theta\right).$$

Predictive Variance Problem

One of the disadvantages of the SR method is that it can give unreasonably small predictive variances when making predictions in a region far away from the chosen active set $\mathrm{A} \subset \mathrm{N} = \{1, 2, ..., n\}$.

Consider making a prediction at a new point x_{new} that is far away from the training set X. In other words, assume that $K\left(x_{new}^T, X\middle|\theta\right) \approx 0$.

For exact GPR, the posterior distribution of f_{new} given y, X and x_{new} would be Normal with mean $\mu = 0$ and variance $\Sigma = k(x_{new}, x_{new}|\theta)$. This value is correct in the sense that, if x_{new} is far from X, then the data (X, y) does not supply any new information about f_{new} and so the posterior distribution of f_{new} given y, X, and x_{new} should reduce to the prior distribution f_{new} given x_{new}, which is a Normal distribution with mean 0 and variance $k(x_{new}, x_{new}|\theta)$.

For the SR approximation, if x_{new} is far away from X (and hence also far away from X_A), then $\mu_{SR} = 0$ and $\Sigma_{SR} = 0$. Thus in this extreme case, μ_{SR} agrees with μ from exact GPR, but Σ_{SR} is unreasonably small compared to Σ from exact GPR.

The fully independent conditional approximation method on page 6-19 can help avoid this problem.

References

[1] Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press. Cambridge, Massachusetts, 2006.

[2] Smola, A. J. and B. Schökopf. "Sparse greedy matrix approximation for machine learning." In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

See Also

`fitrgp` | `predict`

More About

- “Gaussian Process Regression Models” on page 6-2
- “Exact GPR Method” on page 6-10
- “Subset of Data Approximation for GPR Models” on page 6-14
- “Fully Independent Conditional Approximation for GPR Models” on page 6-19
- “Block Coordinate Descent Approximation for GPR Models” on page 6-22

Fully Independent Conditional Approximation for GPR Models

In this section...

The fully independent conditional (FIC) approximation[1] is a way of systematically approximating the true GPR kernel function in a way that avoids the predictive variance problem of the SR approximation on page 6-17 while still maintaining a valid Gaussian process. You can specify the FIC method for parameter estimation by using the `'FitMethod','fic'` name-value pair argument in the call to `fitrgp`. For prediction using FIC, you can use the `'PredictMethod','fic'` name-value pair argument in the call to `fitrgp`.

Approximating the Kernel Function

The FIC approximation to $k\left(x_i, x_j|\theta\right)$ for active set $\mathrm{A} \subset \mathrm{N} = \{1, 2, \ldots, n\}$ is given by:

$$\widehat{k}_{FIC}\left(x_i, x_j|\theta, \mathrm{A}\right) = \widehat{k}_{SR}\left(x_i, x_j|\theta, \mathrm{A}\right) + \delta_{ij}\left(k\left(x_i, x_j|\theta\right) - \widehat{k}_{SR}\left(x_i, x_j|\theta, \mathrm{A}\right)\right),$$

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

That is, the FIC approximation is equal to the SR approximation if $i \neq j$. For $i = j$, the software uses the exact kernel value rather than an approximation. Define an n-by-n diagonal matrix $\Omega(X|\theta, \mathrm{A})$ as follows:

$$\begin{aligned} \left[\Omega(X|\theta, \mathrm{A})\right]_{ij} &= \delta_{ij}\left(k\left(x_i, x_j|\theta\right) - \widehat{k}_{SR}\left(x_i, x_j|\theta, \mathrm{A}\right)\right) \\ &= \begin{cases} k\left(x_i, x_j|\theta\right) - \widehat{k}_{SR}\left(x_i, x_j|\theta, \mathrm{A}\right) & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \end{aligned}$$

The FIC approximation to $K(X, X|\theta)$ is then given by:

$$\begin{aligned} \widehat{K}_{FIC}(X, X|\theta, \mathrm{A}) &= \widehat{K}_{SR}(X, X|\theta, \mathrm{A}) + \Omega(X|\theta, \mathrm{A}) \\ &= K(X, X_{\mathrm{A}}|\theta)K(X_{\mathrm{A}}, X_{\mathrm{A}}|\theta)^{-1}K(X_{\mathrm{A}}, X|\theta) + \Omega(X|\theta, \mathrm{A}). \end{aligned}$$

Parameter Estimation

Replacing $K(X, X|\theta)$ by $\widehat{K}_{FIC}(X, X|\theta, \mathrm{A})$ in the marginal log likelihood function produces its FIC approximation:

$$\begin{aligned} \log P_{FIC}\left(y\middle|X, \beta, \theta, \sigma^2, \mathrm{A}\right) = &-\frac{1}{2}(y - H\beta)^T\left[\widehat{K}_{FIC}(X, X|\theta, \mathrm{A}) + \sigma^2 I_n\right]^{-1}(y - H\beta) \\ &-\frac{N}{2}\log 2\pi - \frac{1}{2}\log\left|\widehat{K}_{FIC}(X, X|\theta, \mathrm{A}) + \sigma^2 I_n\right|. \end{aligned}$$

As in the exact method on page 6-10, the software estimates the parameters by first computing $\widehat{\beta}\left(\theta,\sigma^2\right)$, the optimal estimate of β, given θ and σ^2. Then it estimates θ, and σ^2 using the β-profiled marginal log likelihood. The FIC estimate to β for given θ, and σ^2 is

$$\widehat{\beta}_{FIC}\left(\theta,\sigma^2,\mathrm{A}\right)=\left[H^T\left(\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right)^{-1}H\right]^{-1}H^T\left(\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right)^{-1}y,$$

$$*=H^T\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}H-H^T\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}K(X,X_\mathrm{A}|\theta)B_\mathrm{A}^{-1}K(X_\mathrm{A},X|\theta)\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}H,$$

$$**=H^T\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}y-H^T\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}K(X,X_\mathrm{A}|\theta)B_\mathrm{A}^{-1}K(X_\mathrm{A},X|\theta)\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}y,$$

$$B_\mathrm{A}=K(X_\mathrm{A},X_\mathrm{A}|\theta)+K(X_\mathrm{A},X|\theta)\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}K(X,X_\mathrm{A}|\theta),$$

$$\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)=\Omega(X|\theta,\mathrm{A})+\sigma^2 I_n\,.$$

Using $\widehat{\beta}_{FIC}\left(\theta,\sigma^2,\mathrm{A}\right)$, the β-profiled marginal log likelihood for FIC approximation is:

$$\log P_{FIC}\left(y\middle|X,\widehat{\beta}_{FIC}\left(\theta,\sigma^2,\mathrm{A}\right),\theta,\sigma^2,\mathrm{A}\right)=$$
$$-\frac{1}{2}\left(y-H\widehat{\beta}_{FIC}\left(\theta,\sigma^2,\mathrm{A}\right)\right)^T\left(\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right)^{-1}\left(y-H\widehat{\beta}_{FIC}\left(\theta,\sigma^2,\mathrm{A}\right)\right)$$
$$-\frac{N}{2}\log 2\pi-\frac{1}{2}\log\left|\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right|,$$

where

$$\left(\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right)^{-1}$$
$$=\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}-\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}K(X,X_\mathrm{A}|\theta)B_\mathrm{A}^{-1}K(X_\mathrm{A},X|\theta)\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1},$$
$$\log\left|\widehat{K}_{FIC}(X,X|\theta,\mathrm{A})+\sigma^2 I_N\right|=\log\left|\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)\right|+\log|B_\mathrm{A}|-\log|K(X_\mathrm{A},X_\mathrm{A}|\theta)|\,.$$

Prediction

The FIC approximation to the distribution of y_{new} given y, X, x_{new} is

$$P(y_{new}\,|\,y,X,x_{new})=\mathrm{N}\left(y_{new}\middle|h(x_{new})^T\beta+\mu_{FIC},\sigma_{new}^2+\Sigma_{FIC}\right),$$

where μ_{FIC} and Σ_{FIC} are the FIC approximations to μ and Σ given in prediction using exact GPR method on page 6-11. As in the SR case, μ_{FIC} and Σ_{FIC} are obtained by replacing all occurrences of the true kernel with its FIC approximation. The final forms of μ_{FIC} and Σ_{FIC} are as follows:

$$\mu_{FIC}=K\left(x_{new}^T,X_\mathrm{A}\middle|\theta\right)B_\mathrm{A}^{-1}K(X_\mathrm{A},X|\theta)\,\Lambda\left(\theta,\sigma^2,\mathrm{A}\right)^{-1}(y-H\beta),$$

$$\Sigma_{FIC}=k(x_{new},x_{new}|\theta)-K\left(x_{new}^T,X_\mathrm{A}\middle|\theta\right)K(X_\mathrm{A},X_\mathrm{A}|\theta)^{-1}K\left(X_\mathrm{A},x_{new}^T\middle|\theta\right)$$
$$+K\left(x_{new}^T,X_\mathrm{A}\middle|\theta\right)B_\mathrm{A}^{-1}K\left(X_\mathrm{A},x_{new}^T\middle|\theta\right),$$

where

$$B_{\mathrm{A}} = K(X_{\mathrm{A}}, X_{\mathrm{A}}|\theta) + K(X_{\mathrm{A}}, X|\theta)\Lambda\left(\theta, \sigma^2, \mathrm{A}\right)^{-1}K(X, X_{\mathrm{A}}|\theta),$$

$$\Lambda\left(\theta, \sigma^2, \mathrm{A}\right) = \Omega(X|\theta, \mathrm{A}) + \sigma^2 I_n\,.$$

References

[1] Candela, J. Q. "A Unifying View of Sparse Approximate Gaussian Process Regression." *Journal of Machine Learning Research*. Vol 6, pp. 1939–1959, 2005.

See Also

`fitrgp` | `predict`

More About

- “Gaussian Process Regression Models” on page 6-2
- “Exact GPR Method” on page 6-10
- “Subset of Data Approximation for GPR Models” on page 6-14
- “Subset of Regressors Approximation for GPR Models” on page 6-15
- “Block Coordinate Descent Approximation for GPR Models” on page 6-22

Block Coordinate Descent Approximation for GPR Models

For a large number of observations, using the exact method for parameter estimation and making predictions on new data can be expensive (see “Exact GPR Method” on page 6-10). One of the approximation methods that help deal with this issue for prediction is the Block Coordinate Descent (BCD) method. You can make predictions using the BCD method by first specifying the predict method using the `'PredictMethod','bcd'` name-value pair argument in the call to `fitrgp`, and then using the `predict` function.

The idea of the BCD method is to compute

$$\widehat{\alpha} = \left(K(X,X) + \sigma^2 I_N\right)^{-1}(y - H\beta)$$

in a different way than the exact method. BCD estimates $\widehat{\alpha}$ by solving the following optimization problem:

$$\widehat{\alpha} = \arg\min_{\alpha} f(\alpha)$$

where

$$f(\alpha) = \frac{1}{2}\alpha^T\left[K(X,X) + \sigma^2 I_N\right]\alpha - \alpha^T(y - H\beta).$$

`fitrgp` uses block coordinate descent (BCD) to solve the above optimization problem. For users familiar with support vector machines (SVMs), sequential minimal optimization (SMO) and ISDA (iterative single data algorithm) used to fit SVMs are special cases of BCD. `fitrgp` performs two steps for each BCD update - block selection and block update. In the block selection phase, `fitrgp` uses a combination of random and greedy strategies for selecting the set of α coefficients to optimize next. In the block update phase, the selected α coefficients are optimized while keeping the other α coefficients fixed to their previous values. These two steps are repeated until convergence. BCD does not store the $n*n$ matrix $K(X,X)$ in memory but it just computes chunks of the $K(X,X)$ matrix as needed. As a result BCD is more memory efficient compared to `'PredictMethod','exact'`.

Practical experience indicates that it is not necessary to solve the optimization problem for computing α to very high precision. For example, it is reasonable to stop the iterations when $||\nabla f(\alpha)||$ drops below $\eta||\nabla f(\alpha_0)||$, where α_0 is the initial value of α and η is a small constant (say $\eta = 10^{-3}$). The results from `'PredictMethod','exact'` and `'PredictMethod','bcd'` are equivalent except BCD computes α in a different way. Hence, `'PredictMethod','bcd'` can be thought of as a memory efficient way of doing `'PredictMethod','exact'`. BCD is often faster than `'PredictMethod','exact'` for large n. However, you cannot compute the prediction intervals using the `'PredictMethod','bcd'` option. This is because computing prediction intervals requires solving a problem like the one solved for computing α for every test point, which again is expensive.

Fit GPR Models Using BCD Approximation

This example shows fitting a Gaussian Process Regression (GPR) model to data with a large number of observations, using the Block Coordinate Descent (BCD) Approximation.

Small n - `PredictMethod 'exact' and 'bcd'` Produce the Same Results

Generate a small sample dataset.

```
rng(0,'twister');
n = 1000;
X = linspace(0,1,n)';
X = [X,X.^2];
y = 1 + X*[1;2] + sin(20*X*[1;-2])./(X(:,1)+1) + 0.2*randn(n,1);
```

Create a GPR model using `'FitMethod','exact'` and `'PredictMethod','exact'`.

```
gpr = fitrgp(X,y,'KernelFunction','squaredexponential',...
    'FitMethod','exact','PredictMethod','exact');
```

Create another GPR model using `'FitMethod','exact'` and `'PredictMethod','bcd'`.

```
gprbcd = fitrgp(X,y,'KernelFunction','squaredexponential',...
    'FitMethod','exact','PredictMethod','bcd','BlockSize',200);
```

`'PredictMethod','exact'` and `'PredictMethod','bcd'` should be equivalent. They compute the same $\hat{\alpha}$ but just in different ways. You can also see the iterations by using the `'Verbose',1` name-value pair argument in the call to `fitrgp` .

Compute the fitted values using the two methods and compare them:

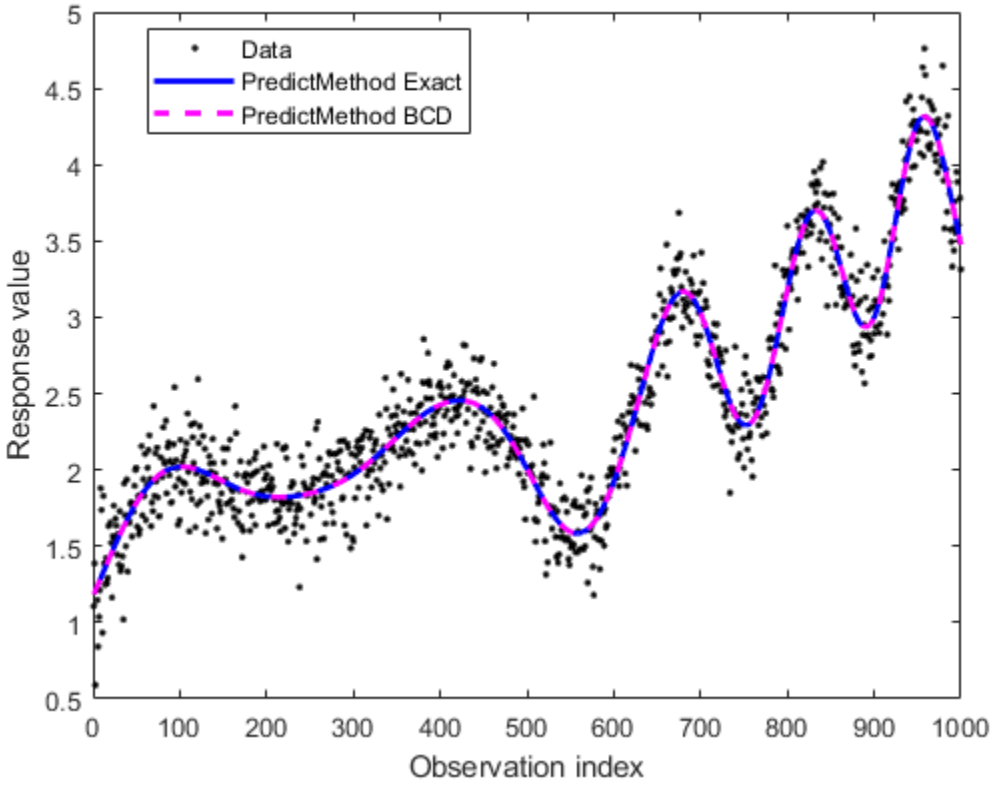
```
ypred    = resubPredict(gpr);
ypredbcd = resubPredict(gprbcd);

max(abs(ypred-ypredbcd))
```

```
ans =

   5.8853e-04
```

The fitted values are close to each other.

Now, plot the data along with fitted values for `'PredictMethod','exact'` and `'PredictMethod','bcd'`.

```
figure;
plot(y,'k.');
hold on;
plot(ypred,'b-','LineWidth',2);
plot(ypredbcd,'m--','LineWidth',2);
legend('Data','PredictMethod Exact','PredictMethod BCD','Location','Best');
xlabel('Observation index');
ylabel('Response value');
```



It can be seen that `'PredictMethod','exact'` and `'PredictMethod','bcd'` produce nearly identical fits.

Large n - Use `'FitMethod','sd'` and `'PredictMethod','bcd'`

Generate a bigger dataset similar to the previous one.

```
rng(0,'twister');
n = 50000;
X = linspace(0,1,n)';
X = [X,X.^2];
y = 1 + X*[1;2] + sin(20*X*[1;-2])./(X(:,1)+1) + 0.2*randn(n,1);
```

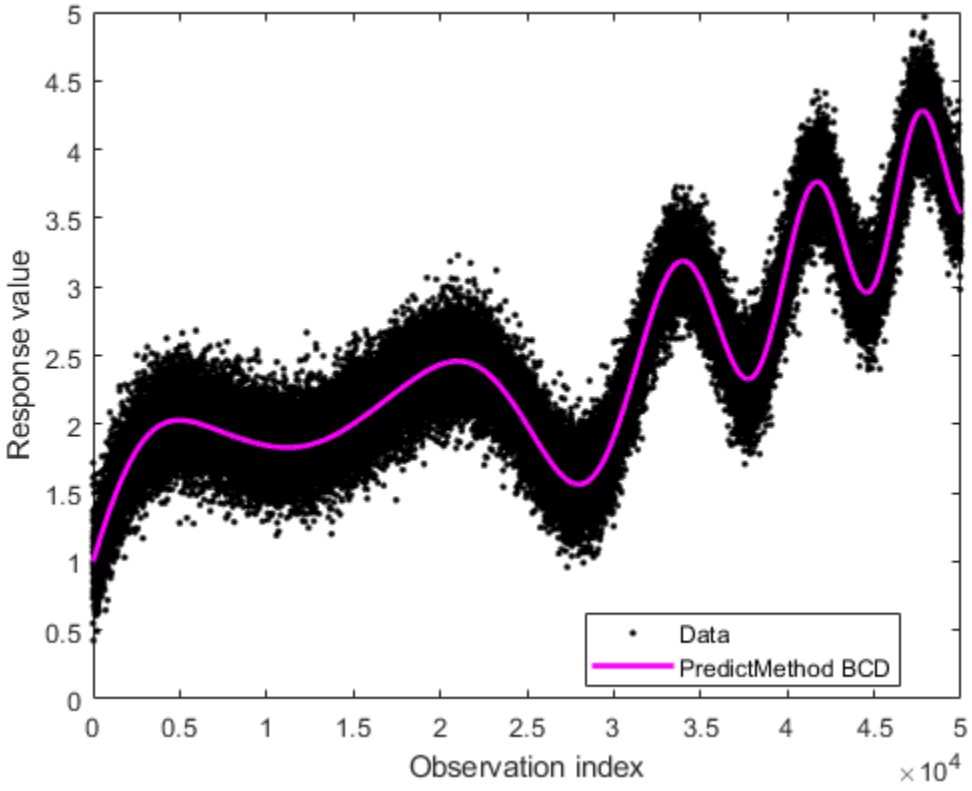
For $n = 50000$, the matrix $K(X, X)$ would be 50000-by-50000. Storing $K(X, X)$ in memory would require around 20 GB of RAM. To fit a GPR model to this dataset, use `'FitMethod','sd'` with a random subset of $m = 2000$ points. The `'ActiveSetSize'` name-value pair argument in the call to `fitrgp` specifies the active set size m. For computing α use `'PredictMethod','bcd'` with a `'BlockSize'` of 5000. The `'BlockSize'` name-value pair argument in `fitrgp` specifies the number of elements of the α vector that the software optimizes at each iteration. A `'BlockSize'` of 5000 assumes that the computer you use can store a 5000-by-5000 matrix in memory.

```
gprbcd = fitrgp(X,y,'KernelFunction','squaredexponential',...,
     'FitMethod','sd','ActiveSetSize',2000,'PredictMethod','bcd','BlockSize',5000);
```

Plot the data and the fitted values.

```
figure;
plot(y,'k.');
```

```
hold on;
plot(resubPredict(gprbcd),'m-','LineWidth',2);
legend('Data','PredictMethod BCD','Location','Best');
xlabel('Observation index');
ylabel('Response value');
```



The plot is similar to the one for smaller n .

References

[1] Grippo, L. and M. Sciandrone. "On the convergence of the block nonlinear gauss-seifel method under convex constraints." *Operations Research Letters.* Vol. 26, pp. 127–136, 2000.

[2] Bo, L. and C. Sminchisescu. "Greed block coordinate descent for large scale Gaussian process regression." In *Proceedings of the Twenty Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008)*: http://arxiv.org/abs/1206.3238, 2012.

See Also

`fitrgp` | `predict`

More About

- “Gaussian Process Regression Models” on page 6-2
- “Exact GPR Method” on page 6-10

Random Number Generation

Generating Pseudorandom Numbers

Pseudorandom numbers are generated by deterministic algorithms. They are "random" in the sense that, on average, they pass statistical tests regarding their distribution and correlation. They differ from true random numbers in that they are generated by an algorithm, rather than a truly random process.

Random number generators (RNGs) like those in MATLAB are algorithms for generating pseudorandom numbers with a specified distribution.

For more information on the GUI for generating random numbers from supported distributions, see “Explore the Random Number Generation UI” on page 5-85.

Common Pseudorandom Number Generation Methods

Methods for generating pseudorandom numbers usually start with uniform random numbers, like the MATLAB `rand` function produces. The methods described in this section detail how to produce random numbers from other distributions.

Direct Methods

Direct methods directly use the definition of the distribution.

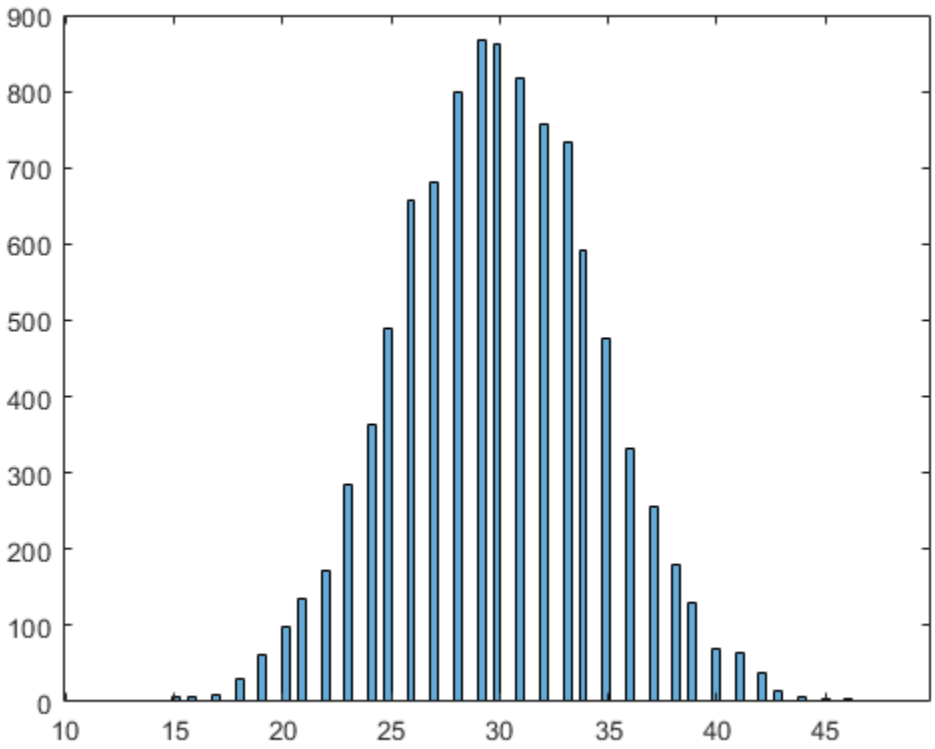
For example, consider binomial random numbers. A binomial random number is the number of heads in N tosses of a coin with probability p of a heads on any single toss. If you generate N uniform random numbers on the interval `(0,1)` and count the number less than p, then the count is a binomial random number with parameters N and p.

This function is a simple implementation of a binomial RNG using the direct approach:

```
function X = directbinornd(N,p,m,n)
    X = zeros(m,n); % Preallocate memory
    for i = 1:m*n
        u = rand(N,1);
        X(i) = sum(u < p);
    end
end
```

For example:

```
rng('default') % For reproducibility
X = directbinornd(100,0.3,1e4,1);
histogram(X,101)
```



The `binornd` function uses a modified direct method, based on the definition of a binomial random variable as the sum of Bernoulli random variables.

You can easily convert the previous method to a random number generator for the Poisson distribution with parameter λ. The "Poisson Distribution" on page B-131 is the limiting case of the binomial distribution as N approaches infinity, p approaches zero, and Np is held fixed at λ. To generate Poisson random numbers, create a version of the previous generator that inputs λ rather than N and p, and internally sets N to some large number and p to λ/N.

The `poissrnd` function actually uses two direct methods:

- A waiting time method for small values of λ
- A method due to Ahrens and Dieter for larger values of λ

Inversion Methods

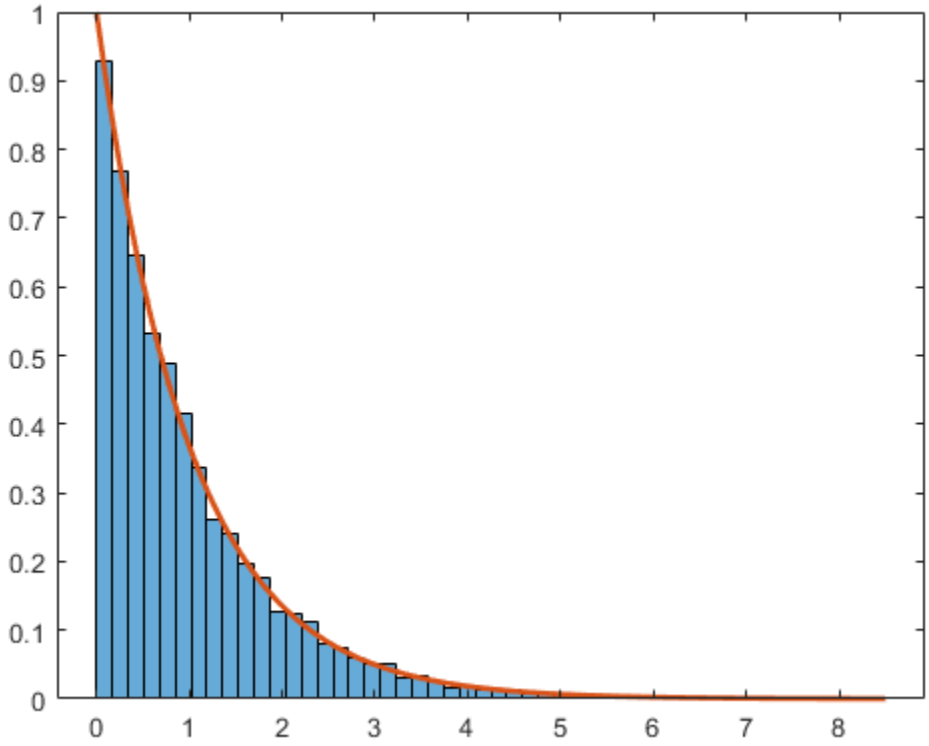
Inversion methods are based on the observation that continuous cumulative distribution functions (cdfs) range uniformly over the interval `(0,1)`. If u is a uniform random number on `(0,1)`, then using $X = F^{-1}(U)$ generates a random number X from a continuous distribution with specified cdf F.

For example, the following code generates random numbers from a specific "Exponential Distribution" on page B-33 using the inverse cdf and the MATLAB® uniform random number generator `rand`:

```
rng('default') % For reproducibility
mu = 1;
X = expinv(rand(1e4,1),mu);
```

Compare the distribution of the generated random numbers to the pdf of the specified exponential.

```
numbins = 50;
h = histogram(X,numbins,'Normalization','pdf');
hold on
x = linspace(h.BinEdges(1),h.BinEdges(end));
y = exppdf(x,mu);
plot(x,y,'LineWidth',2)
hold off
```



Inversion methods also work for discrete distributions. To generate a random number X from a discrete distribution with probability mass vector $P(X = x_i) = p_i$ where $x_0 < x_1 < x_2 < \ldots$, generate a uniform random number u on `(0,1)` and then set $X = x_i$ if $F(x_{i-1}) < u < F(x_i)$.

For example, the following function implements an inversion method for a discrete distribution with probability mass vector p:

```
function X = discreteinvrnd(p,m,n)
    X = zeros(m,n); % Preallocate memory
    for i = 1:m*n
        u = rand;
        I = find(u < cumsum(p));
        X(i) = min(I);
```

```
    end
end
```

Use the function to generate random numbers from any discrete distribution.

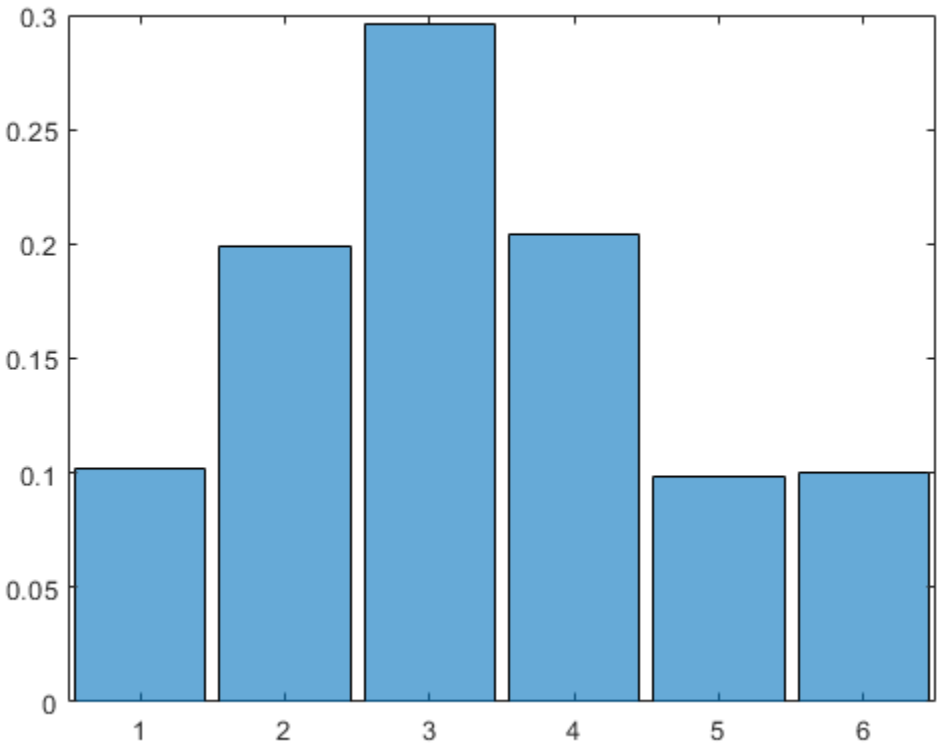
```
p = [0.1 0.2 0.3 0.2 0.1 0.1]; % Probability mass function (pmf) values
X = discreteinvrnd(p,1e4,1);
```

Alternatively, you can use the `discretize` function to generate discrete random numbers.

```
X = discretize(rand(1e4,1),[0 cusmsum(p)]);
```

Plot the histogram of the generated random numbers, and confirm then the distribution follows the specified pmf values.

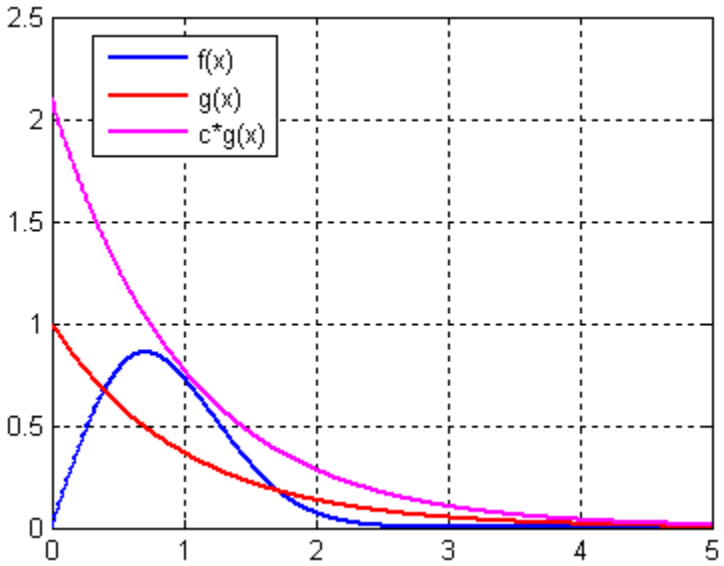
```
histogram(categorical(X),'Normalization','probability')
```



Acceptance-Rejection Methods

The functional form of some distributions makes it difficult or time-consuming to generate random numbers using direct or inversion methods. Acceptance-rejection methods provide an alternative in these cases.

Acceptance-rejection methods begin with uniform random numbers, but require an additional random number generator. If your goal is to generate a random number from a continuous distribution with pdf f, acceptance-rejection methods first generate a random number from a continuous distribution with pdf g satisfying $f(x) \le cg(x)$ for some c and all x.



A continuous acceptance-rejection RNG proceeds as follows:

1. Chooses a density g.
2. Finds a constant c such that $f(x)/g(x) \le c$ for all x.
3. Generates a uniform random number u.
4. Generates a random number v from g.
5. If $cu \le f(v)/g(v)$, accepts and returns v. Otherwise, rejects v and goes to step 3.

For efficiency, a "cheap" method is necessary for generating random numbers from g, and the scalar c should be small. The expected number of iterations to produce a single random number is c.

The following function implements an acceptance-rejection method for generating random numbers from pdf f given f, g, the RNG `grnd` for g, and the constant c:

```
function X = accrejrnd(f,g,grnd,c,m,n)
    X = zeros(m,n); % Preallocate memory
    for i = 1:m*n
        accept = false;
        while accept == false
            u = rand();
            v = grnd();
            if c*u <= f(v)/g(v)
               X(i) = v;
               accept = true;
            end
        end
    end
end
```

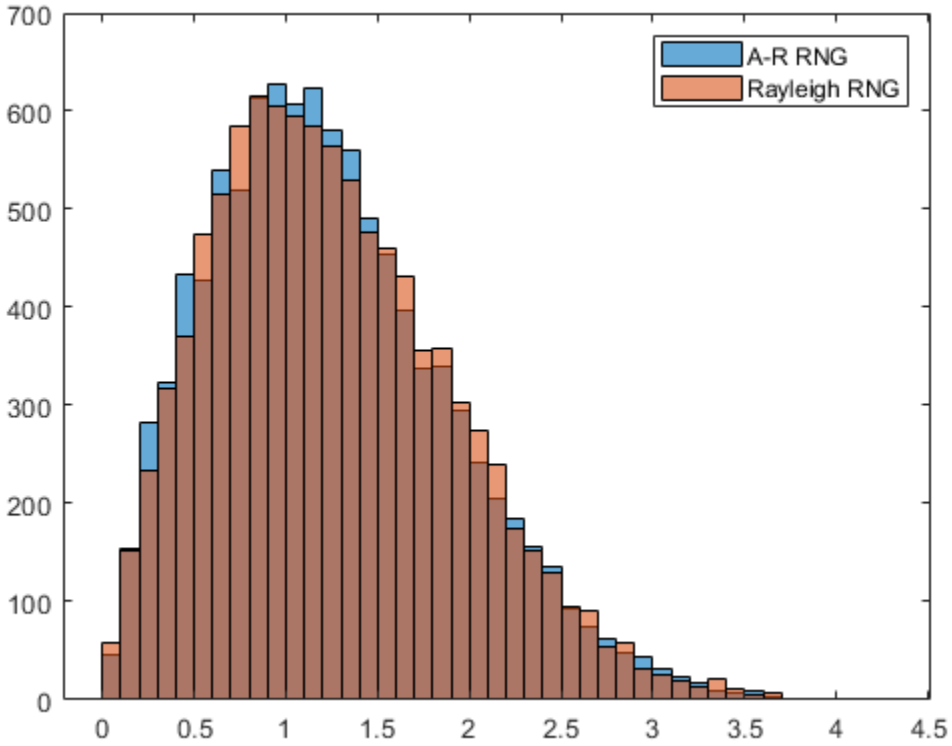
For example, the function $f(x) = xe^{-x^2/2}$ satisfies the conditions for a pdf on $[0, \infty)$ (nonnegative and integrates to 1). The exponential pdf with mean 1, $f(x) = e^{-x}$, dominates g for c greater than about 2.2. Thus, you can use `rand` and `exprnd` to generate random numbers from f:

```
f = @(x)x.*exp(-(x.^2)/2);
g = @(x)exp(-x);
```

```
grnd = @()exprnd(1);
rng('default') % For reproducibility
X = accrejrnd(f,g,grnd,2.2,1e4,1);
```

The pdf f is actually a “Rayleigh Distribution” on page B-137 with shape parameter 1. This example compares the distribution of random numbers generated by the acceptance-rejection method with those generated by `raylrnd`:

```
Y = raylrnd(1,1e4,1);
histogram(X)
hold on
histogram(Y)
legend('A-R RNG','Rayleigh RNG')
```



The `raylrnd` function uses a transformation method, expressing a Rayleigh random variable in terms of a chi-square random variable, which you compute using `randn`.

Acceptance-rejection methods also work for discrete distributions. In this case, the goal is to generate random numbers from a distribution with probability mass $P_p(X = i) = p_i$, assuming that you have a method for generating random numbers from a distribution with probability mass $P_q(X = i) = q_i$. The RNG proceeds as follows:

1. Chooses a density P_q.
2. Finds a constant c such that $p_i/q_i \le c$ for all i.
3. Generates a uniform random number u.

4 Generates a random number v from P_q.

5 If $cu \le p_v/q_v$, accepts and returns v. Otherwise, rejects v and goes to step 3.

See Also

More About

- “Random Number Generation” on page 5-27
- “Generating Quasi-Random Numbers” on page 7-12
- “Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101

Representing Sampling Distributions Using Markov Chain Samplers

In this section...

For more complex probability distributions, you might need more advanced methods for generating samples than the methods described in “Common Pseudorandom Number Generation Methods” on page 7-2. Such distributions arise, for example, in Bayesian data analysis and in the large combinatorial problems of Markov chain Monte Carlo (MCMC) simulations. An alternative is to construct a Markov chain with a stationary distribution equal to the target sampling distribution, using the states of the chain to generate random numbers after an initial burn-in period in which the state distribution converges to the target.

Using the Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm draws samples from a distribution that is only known up to a constant. Random numbers are generated from a distribution with a probability density function that is equal to or proportional to a proposal function.

To generate random numbers:

1. Assume an initial value $x(t)$.
2. Draw a sample, $y(t)$, from a proposal distribution $q(y|x(t))$.
3. Accept $y(t)$ as the next sample $x(t + 1)$ with probability $r(x(t),y(t))$, and keep $x(t)$ as the next sample $x(t + 1)$ with probability $1 – r(x(t),y(t))$, where:

   $$r(x, y) = min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$$

4. Increment $t \rightarrow t + 1$, and repeat steps 2 and 3 until you get the desired number of samples.

Generate random numbers using the Metropolis-Hastings method with the `mhsample` function. To produce quality samples efficiently with the Metropolis-Hastings algorithm, it is crucial to select a good proposal distribution. If it is difficult to find an efficient proposal distribution, use slice sampling (`slicesample`) or Hamiltonian Monte Carlo (`hmcSampler`) instead.

Using Slice Sampling

In instances where it is difficult to find an efficient Metropolis-Hastings proposal distribution, the slice sampling algorithm does not require an explicit specification. The slice sampling algorithm draws samples from the region under the density function using a sequence of vertical and horizontal steps. First, it selects a height at random from 0 to the density function $f(x)$. Then, it selects a new x value at random by sampling from the horizontal “slice” of the density above the selected height. A similar slice sampling algorithm is used for a multivariate distribution.

If a function $f(x)$ proportional to the density function is given, then do the following to generate random numbers:

1. Assume an initial value $x(t)$ within the domain of $f(x)$.
2. Draw a real value y uniformly from $(0, f(x(t)))$, thereby defining a horizontal "slice" as $S = \{x: y < f(x)\}$.
3. Find an interval $I = (L, R)$ around $x(t)$ that contains all, or much of the "slice" S.
4. Draw the new point $x(t + 1)$ within this interval.
5. Increment $t \rightarrow t + 1$ and repeat steps 2 through 4 until you get the desired number of samples.

Slice sampling can generate random numbers from a distribution with an arbitrary form of the density function, provided that an efficient numerical procedure is available to find the interval $I = (L,R)$, which is the "slice" of the density.

Generate random numbers using the slice sampling method with the `slicesample` function.

Using Hamiltonian Monte Carlo

Metropolis-Hastings and slice sampling can produce MCMC chains that mix slowly and take a long time to converge to the stationary distribution, especially in medium-dimensional and high-dimensional problems. Use the gradient-based Hamiltonian Monte Carlo (HMC) sampler to speed up sampling in these situations.

To use HMC sampling, you must specify log $f(x)$ (up to an additive constant) and its gradient. You can use a numerical gradient, but this leads to slower sampling. All sampling variables must be unconstrained, meaning that log $f(x)$ and its gradient are well-defined for all real x. To sample constrained variables, transform these variables into unconstrained ones before using the HMC sampler.

The HMC sampling algorithm introduces a random "momentum vector" z and defines a joint density of z and the "position vector" x as $P(x,z) = f(x)g(z)$. The goal is to sample from this joint distribution and then to ignore the values of z — the marginal distribution of x has the desired density $f(x)$.

The HMC algorithm assigns a Gaussian density with covariance matrix M (the "mass matrix") to z:

$$g(z) \propto \exp\left(-\frac{1}{2}z^T M^{-1} z\right)$$

Then, it defines an "energy function" as

$$E(x, z) = -\log f(x) + \frac{1}{2}z^T M^{-1} z = U(x) + K(z)$$

with $U(x) = -\log f(x)$ the "potential energy" and $K(z) = z^T M^{-1} z/2$ the "kinetic energy". The joint density is given by $P(x,z) \propto \exp\{-E(x,z)\}$.

To generate random samples, the HMC algorithm:

1. Assumes an initial value x of the position vector.
2. Generates a sample of the momentum vector: $z \sim g(z)$.
3. Evolves the state (x, z) for some amount of fictitious time τ to a new state (x', z') using the "equations of motion":

$$\frac{dz}{d\tau} = -\frac{\partial U}{\partial x}$$

$$\frac{dx}{d\tau} = \frac{\partial K}{\partial z}$$

If the equations of motion could be solved exactly, the energy (and hence the density) would remain constant: $E(x,z) = E(x',z')$. In practice, the equations of motions must be solved numerically (usually using so-called leapfrog integration) and the energy is not conserved.

4. Accepts x' as the next sample with probability $p_{acc} = \min(1, \exp\{E(x,z) - E(x',z')\})$, and keeps x as the next sample with probability $1 - p_{acc}$.
5. Repeats steps 2 through 4 until it has generated the desired number of samples.

To use HMC sampling, create a sampler using the `hmcSampler` function. After creating a sampler, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics. For an example of this workflow, see “Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26.

See Also

Functions

`hmcSampler` | `mhsample` | `slicesample`

Generating Quasi-Random Numbers

In this section...

Quasi-Random Sequences

Quasi-random number generators (QRNGs) produce highly uniform samples of the unit hypercube. QRNGs minimize the discrepancy between the distribution of generated points and a distribution with equal proportions of points in each sub-cube of a uniform partition of the hypercube. As a result, QRNGs systematically fill the “holes” in any initial segment of the generated quasi-random sequence.

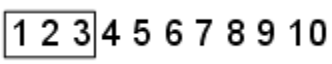
Unlike the pseudorandom sequences described in “Common Pseudorandom Number Generation Methods” on page 7-2, quasi-random sequences fail many statistical tests for randomness. Approximating true randomness, however, is not their goal. Quasi-random sequences seek to fill space uniformly, and to do so in such a way that initial segments approximate this behavior up to a specified density.

QRNG applications include:

- **Quasi-Monte Carlo (QMC) integration.** Monte Carlo techniques are often used to evaluate difficult, multi-dimensional integrals without a closed-form solution. QMC uses quasi-random sequences to improve the convergence properties of these techniques.
- **Space-filling experimental designs.** In many experimental settings, taking measurements at every factor setting is expensive or infeasible. Quasi-random sequences provide efficient, uniform sampling of the design space.
- **Global optimization.** Optimization algorithms typically find a local optimum in the neighborhood of an initial value. By using a quasi-random sequence of initial values, searches for global optima uniformly sample the basins of attraction of all local minima.

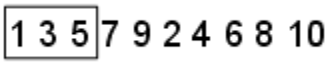
Example: Using Scramble, Leap, and Skip

Imagine a simple 1-D sequence that produces the integers from 1 to 10. This is the basic sequence and the first three points are `[1,2,3]`:

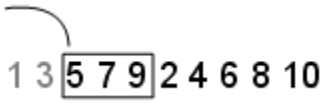


Now look at how `Scramble`, `Skip`, and `Leap` work together:

- `Scramble` — Scrambling shuffles the points in one of several different ways. In this example, assume a scramble turns the sequence into `1,3,5,7,9,2,4,6,8,10`. The first three points are now `[1,3,5]`:



- `Skip` — A `Skip` value specifies the number of initial points to ignore. In this example, set the `Skip` value to 2. The sequence is now `5,7,9,2,4,6,8,10` and the first three points are `[5,7,9]`:



- `Leap` — A `Leap` value specifies the number of points to ignore for each one you take. Continuing the example with the `Skip` set to 2, if you set the `Leap` to 1, the sequence uses every other point. In this example, the sequence is now `5,9,4,8` and the first three points are `[5,9,4]`:

Quasi-Random Point Sets

Statistics and Machine Learning Toolbox functions support these quasi-random sequences:

- **Halton sequences.** Produced by the `haltonset` function. These sequences use different prime bases to form successively finer uniform partitions of the unit interval in each dimension.
- **Sobol sequences.** Produced by the `sobolset` function. These sequences use a base of 2 to form successively finer uniform partitions of the unit interval, and then reorder the coordinates in each dimension.
- **Latin hypercube sequences.** Produced by the `lhsdesign` function. Though not quasi-random in the sense of minimizing discrepancy, these sequences nevertheless produce sparse uniform samples useful in experimental designs.

Quasi-random sequences are functions from the positive integers to the unit hypercube. To be useful in application, an initial point set of a sequence must be generated. Point sets are matrices of size *n*-by-*d*, where *n* is the number of points and *d* is the dimension of the hypercube being sampled. The functions `haltonset` and `sobolset` construct point sets with properties of a specified quasi-random sequence. Initial segments of the point sets are generated by the `net` method of the `haltonset` and `sobolset` classes, but points can be generated and accessed more generally using parenthesis indexing.

Because of the way in which quasi-random sequences are generated, they may contain undesirable correlations, especially in their initial segments, and especially in higher dimensions. To address this issue, quasi-random point sets often skip, leap over, or scramble values in a sequence. The `haltonset` and `sobolset` functions allow you to specify both a `Skip` and a `Leap` property of a quasi-random sequence, and the `scramble` method of the `haltonset` and `sobolset` classes allows you apply a variety of scrambling techniques. Scrambling reduces correlations while also improving uniformity.

Generate a Quasi-Random Point Set

This example shows how to use `haltonset` to construct a 2-D Halton quasi-random point set.

Create a `haltonset` object `p`, that skips the first 1000 values of the sequence and then retains every 101st point.

```
rng default  % For reproducibility
p = haltonset(2,'Skip',1e3,'Leap',1e2)

p =
Halton point set in 2 dimensions (89180190640991 points)

Properties:
               Skip : 1000
               Leap : 100
    ScrambleMethod : none
```

The object `p` encapsulates properties of the specified quasi-random sequence. The point set is finite, with a length determined by the `Skip` and `Leap` properties and by limits on the size of point set indices.

Use `scramble` to apply reverse-radix scrambling.

```
p = scramble(p,'RR2')

p =
Halton point set in 2 dimensions (89180190640991 points)

Properties:
               Skip : 1000
               Leap : 100
    ScrambleMethod : RR2
```

Use `net` to generate the first 500 points.

```
X0 = net(p,500);
```

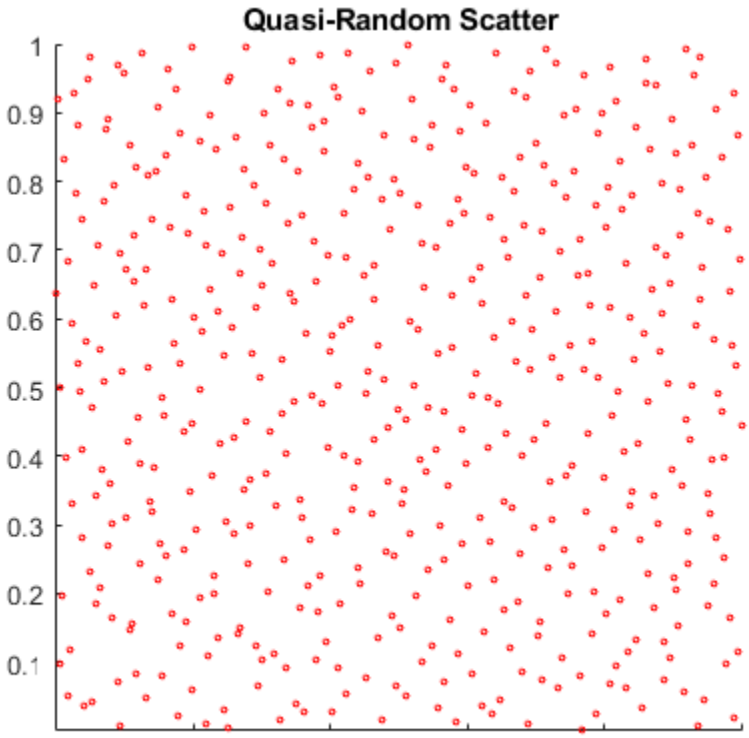
This is equivalent to

```
X0 = p(1:500,:);
```

Values of the point set `X0` are not generated and stored in memory until you access `p` using `net` or parenthesis indexing.

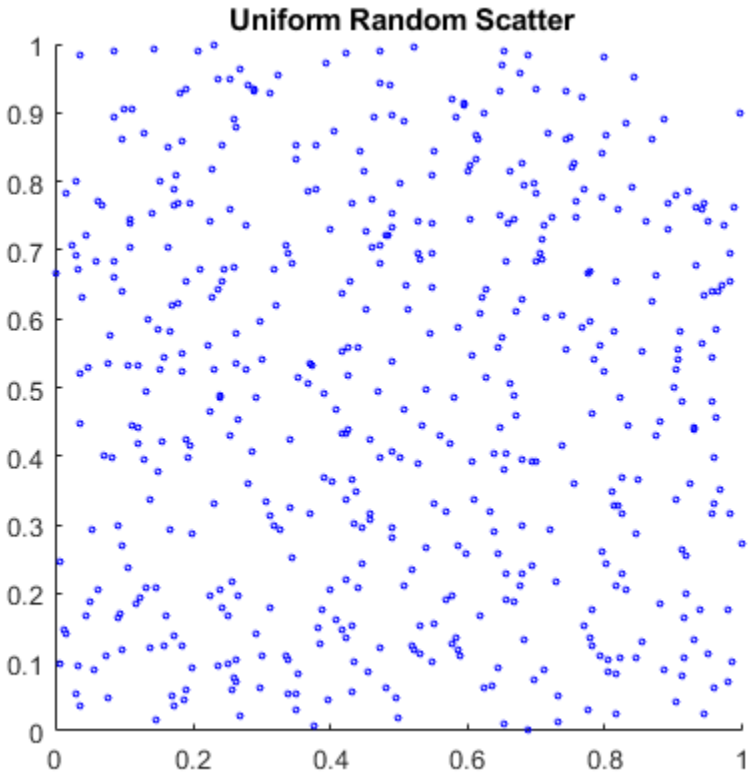
To appreciate the nature of quasi-random numbers, create a scatter plot of the two dimensions in `X0`.

```
scatter(X0(:,1),X0(:,2),5,'r')
axis square
title('{\bf Quasi-Random Scatter}')
```



Compare this to a scatter of uniform pseudorandom numbers generated by the `rand` function.

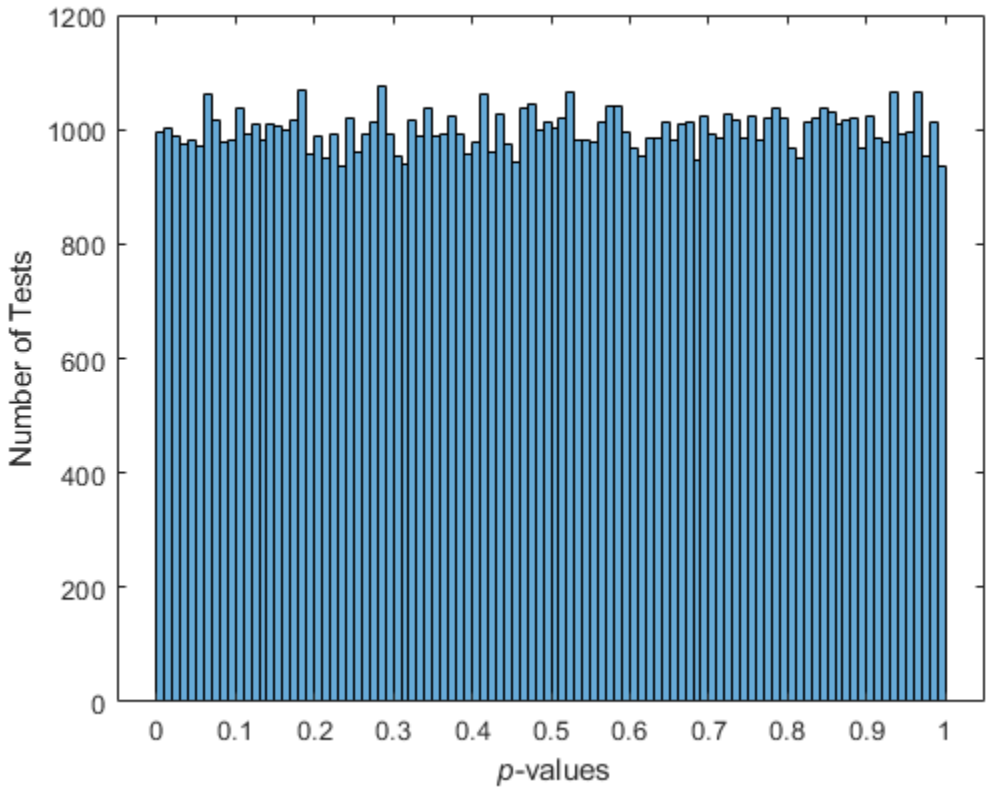
```
X = rand(500,2);
scatter(X(:,1),X(:,2),5,'b')
axis square
title('{\bf Uniform Random Scatter}')
```



The quasi-random scatter appears more uniform, avoiding the clumping in the pseudorandom scatter.

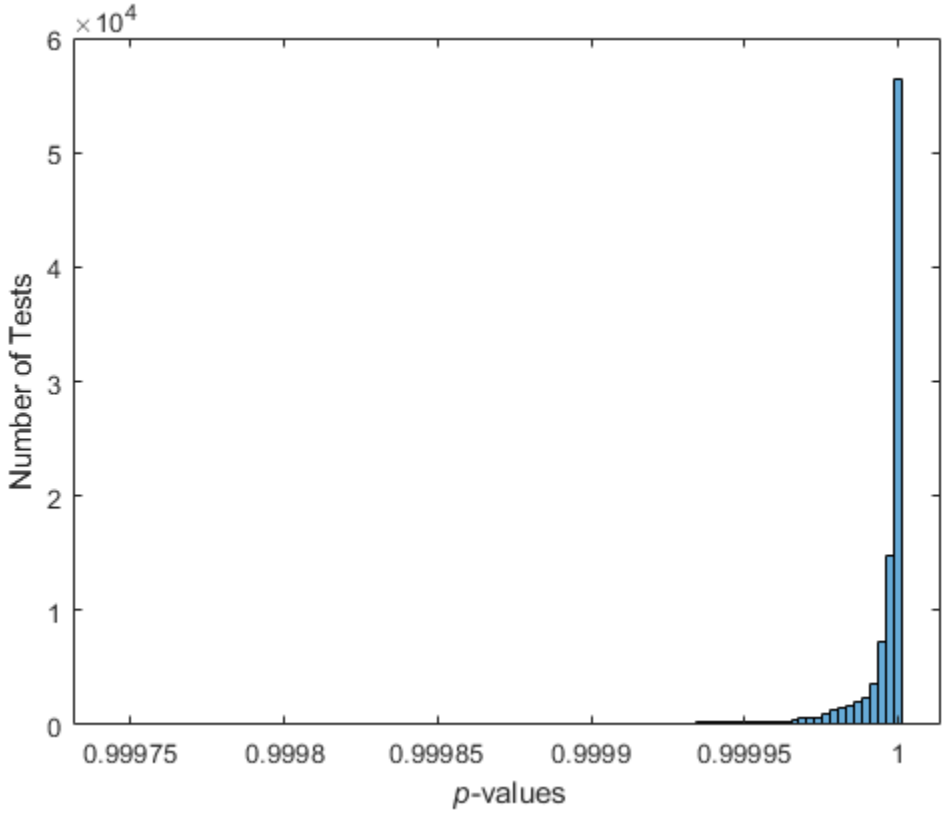
In a statistical sense, quasi-random numbers are too uniform to pass traditional tests of randomness. For example, a Kolmogorov-Smirnov test, performed by `kstest`, is used to assess whether or not a point set has a uniform random distribution. When performed repeatedly on uniform pseudorandom samples, such as those generated by `rand`, the test produces a uniform distribution of *p*-values.

```
nTests = 1e5;
sampSize = 50;
PVALS = zeros(nTests,1);
for test = 1:nTests
    x = rand(sampSize,1);
    [h,pval] = kstest(x,[x,x]);
    PVALS(test) = pval;
end

histogram(PVALS,100)
h = findobj(gca,'Type','patch');
xlabel('{\it p}-values')
ylabel('Number of Tests')
```



The results are quite different when the test is performed repeatedly on uniform quasi-random samples.

```
p = haltonset(1,'Skip',1e3,'Leap',1e2);
p = scramble(p,'RR2');

nTests = 1e5;
sampSize = 50;
PVALS = zeros(nTests,1);
for test = 1:nTests
    x = p(test:test+(sampSize-1),:);
    [h,pval] = kstest(x,[x,x]);
    PVALS(test) = pval;
end

histogram(PVALS,100)
xlabel('{\it p}-values')
ylabel('Number of Tests')
```



Small *p*-values call into question the null hypothesis that the data are uniformly distributed. If the hypothesis is true, about 5% of the *p*-values are expected to fall below 0.05. The results are remarkably consistent in their failure to challenge the hypothesis.

Quasi-Random Streams

Quasi-random streams, produced by the `qrandstream` function, are used to generate sequential quasi-random outputs, rather than point sets of a specific size. Streams are used like pseudoRNGS, such as `rand`, when client applications require a source of quasi-random numbers of indefinite size that can be accessed intermittently. Properties of a quasi-random stream, such as its type (Halton or Sobol), dimension, skip, leap, and scramble, are set when the stream is constructed.

In implementation, quasi-random streams are essentially very large quasi-random point sets, though they are accessed differently. The state of a quasi-random stream is the scalar index of the next point to be taken from the stream. Use the `qrand` method of the `qrandstream` on page 33-5078 class to generate points from the stream, starting from the current state. Use the `reset` method to reset the state to `1`. Unlike point sets, streams do not support parenthesis indexing.

Generate a Quasi-Random Stream

This example shows how to generate samples from a quasi-random point set.

Use `haltonset` to create a quasi-random point set `p`, then repeatedly increment the index into the point set `test` to generate different samples.

```
p = haltonset(1,'Skip',1e3,'Leap',1e2);
p = scramble(p,'RR2');

nTests = 1e5;
sampSize = 50;
PVALS = zeros(nTests,1);
for test = 1:nTests
    x = p(test:test+(sampSize-1),:);
    [h,pval] = kstest(x,[x,x]);
    PVALS(test) = pval;
end
```

The same results are obtained by using `qrandstream` to construct a quasi-random stream `q` based on the point set `p` and letting the stream take care of increments to the index.

```
p = haltonset(1,'Skip',1e3,'Leap',1e2);
p = scramble(p,'RR2');
q = qrandstream(p);

nTests = 1e5;
sampSize = 50;
PVALS = zeros(nTests,1);
for test = 1:nTests
    X = qrand(q,sampSize);
    [h,pval] = kstest(X,[X,X]);
    PVALS(test) = pval;
end
```

See Also

More About

- “Random Number Generation” on page 5-27
- “Generating Pseudorandom Numbers” on page 7-2

Generating Data Using Flexible Families of Distributions

This example shows how to generate data using the Pearson and Johnson systems of distributions.

Pearson and Johnson Systems

As described in “Working with Probability Distributions” on page 5-3, choosing an appropriate parametric family of distributions to model your data can be based on *a priori* or *a posteriori* knowledge of the data-producing process, but the choice is often difficult. The *Pearson and Johnson systems* can make such a choice unnecessary. Each system is a flexible parametric family of distributions that includes a wide range of distribution shapes, and it is often possible to find a distribution within one of these two systems that provides a good match to your data.

Data Input

The following parameters define each member of the Pearson and Johnson systems.

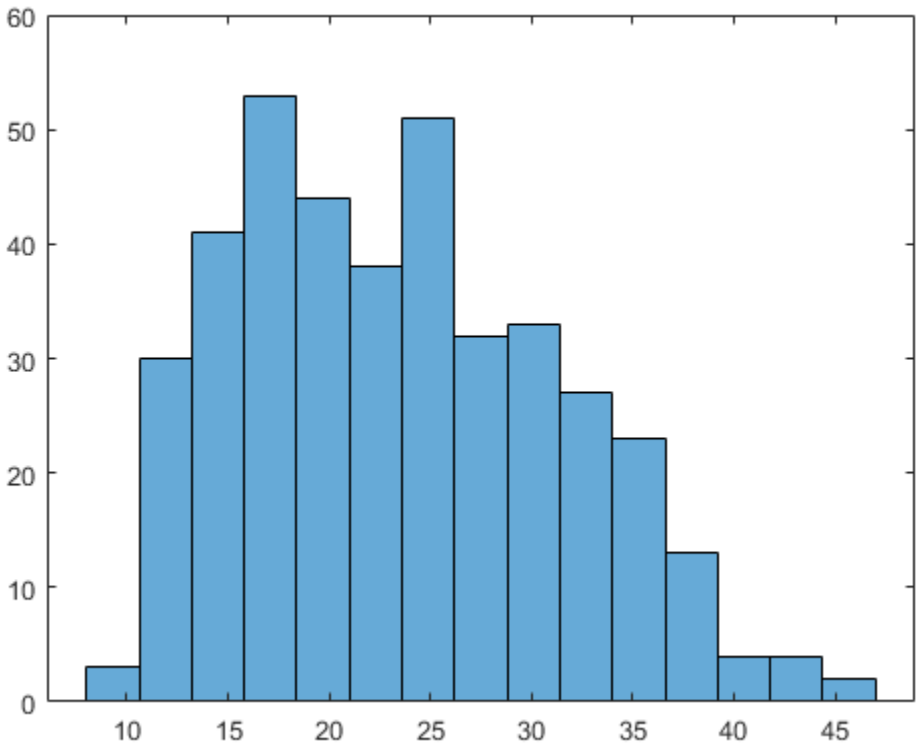
- Mean — Estimated by `mean`
- Standard deviation — Estimated by `std`
- Skewness — Estimated by `skewness`
- Kurtosis — Estimated by `kurtosis`

These statistics can also be computed with the `moment` function. The Johnson system, while based on these four parameters, is more naturally described using quantiles, estimated by the `quantile` function.

The `pearsrnd` and `johnsrnd` functions take input arguments defining a distribution (parameters or quantiles, respectively) and return the type and the coefficients of the distribution in the corresponding system. Both functions also generate random numbers from the specified distribution.

As an example, load the data in `carbig.mat`, which includes a variable `MPG` containing measurements of the gas mileage for each car.

```
load carbig
MPG = MPG(~isnan(MPG));
histogram(MPG,15)
```



The following two sections model the distribution with members of the Pearson and Johnson systems, respectively.

Generating Data Using the Pearson System

The statistician Karl Pearson devised a system, or family, of distributions that includes a unique distribution corresponding to every valid combination of mean, standard deviation, skewness, and kurtosis. If you compute sample values for each of these moments from data, it is easy to find the distribution in the Pearson system that matches these four moments and to generate a random sample.

The Pearson system embeds seven basic types of distribution together in a single parametric framework. It includes common distributions such as the normal and *t* distributions, simple transformations of standard distributions such as a shifted and scaled beta distribution and the inverse gamma distribution, and one distribution—the Type IV—that is not a simple transformation of any standard distribution.

For a given set of moments, there are distributions that are not in the system that also have those same first four moments, and the distribution in the Pearson system may not be a good match to your data, particularly if the data are multimodal. But the system does cover a wide range of distribution shapes, including both symmetric and skewed distributions.

To generate a sample from the Pearson distribution that closely matches the `MPG` data, simply compute the four sample moments and treat those as distribution parameters.

```
moments = {mean(MPG),std(MPG),skewness(MPG),kurtosis(MPG)};
rng('default')  % For reproducibility
[r,type] = pearsrnd(moments{:},10000,1);
```

The optional second output from `pearsrnd` indicates which type of distribution within the Pearson system matches the combination of moments.

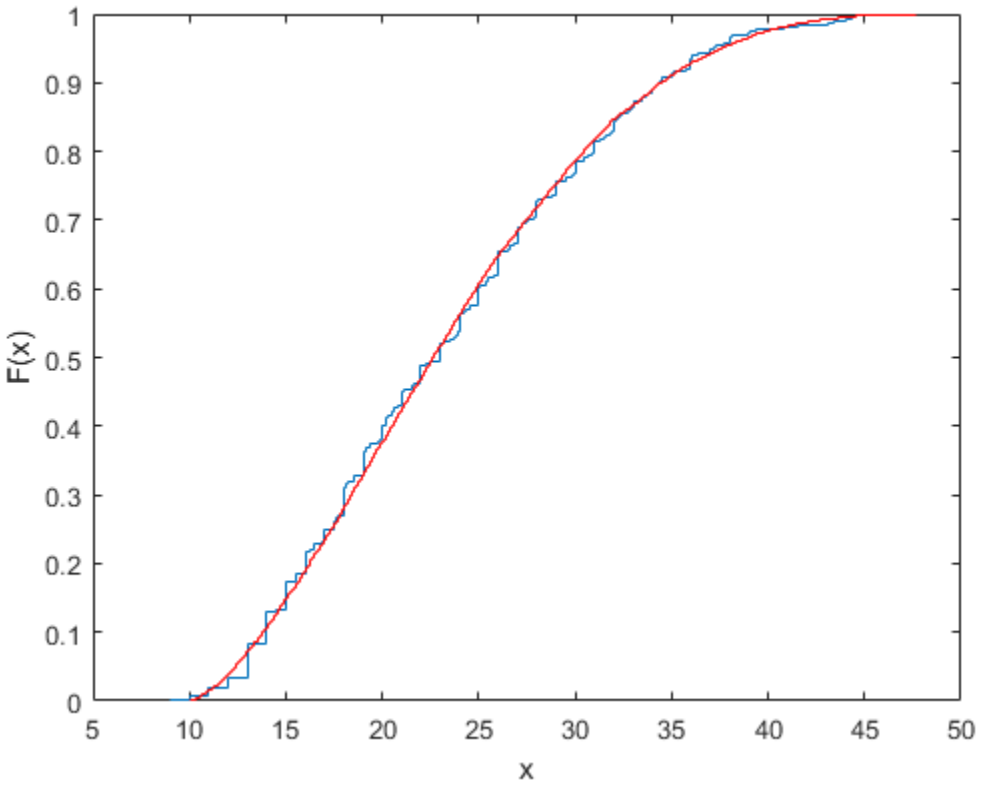
```
type
```

```
type = 1
```

In this case, `pearsrnd` has determined that the data are best described with a Type I Pearson distribution, which is a shifted, scaled beta distribution.

Verify that the sample resembles the original data by overlaying the empirical cumulative distribution functions.

```
ecdf(MPG);
[Fi,xi] = ecdf(r);
hold on;
stairs(xi,Fi,'r');
hold off
```



Generating Data Using the Johnson System

Statistician Norman Johnson devised a different system of distributions that also includes a unique distribution for every valid combination of mean, standard deviation, skewness, and kurtosis.

However, since it is more natural to describe distributions in the Johnson system using quantiles, working with this system is different than working with the Pearson system.

The Johnson system is based on three possible transformations of a normal random variable, plus the identity transformation. The three nontrivial cases are known as `SL` , `SU` , and `SB` , corresponding to exponential, logistic, and hyperbolic sine transformations. All three can be written as

$$X = \gamma + \delta \cdot \Gamma\left(\frac{Z - \xi}{\lambda}\right)$$

where Z is a standard normal random variable, Γ is the transformation, and γ, δ, ξ, and λ are scale and location parameters. The fourth case, `SN` , is the identity transformation.

To generate a sample from the Johnson distribution that matches the `MPG` data, first define the four quantiles to which the four evenly spaced standard normal quantiles of -1.5, -0.5, 0.5, and 1.5 should be transformed. That is, you compute the sample quantiles of the data for the cumulative probabilities of 0.067, 0.309, 0.691, and 0.933.

```
probs = normcdf([-1.5 -0.5 0.5 1.5])
```

```
probs = 1×4

    0.0668    0.3085    0.6915    0.9332
```

```
quantiles = quantile(MPG,probs)
```

```
quantiles = 1×4

   13.0000   18.0000   27.2000   36.0000
```

Then treat those quantiles as distribution parameters.

```
[r1,type] = johnsrnd(quantiles,10000,1);
```

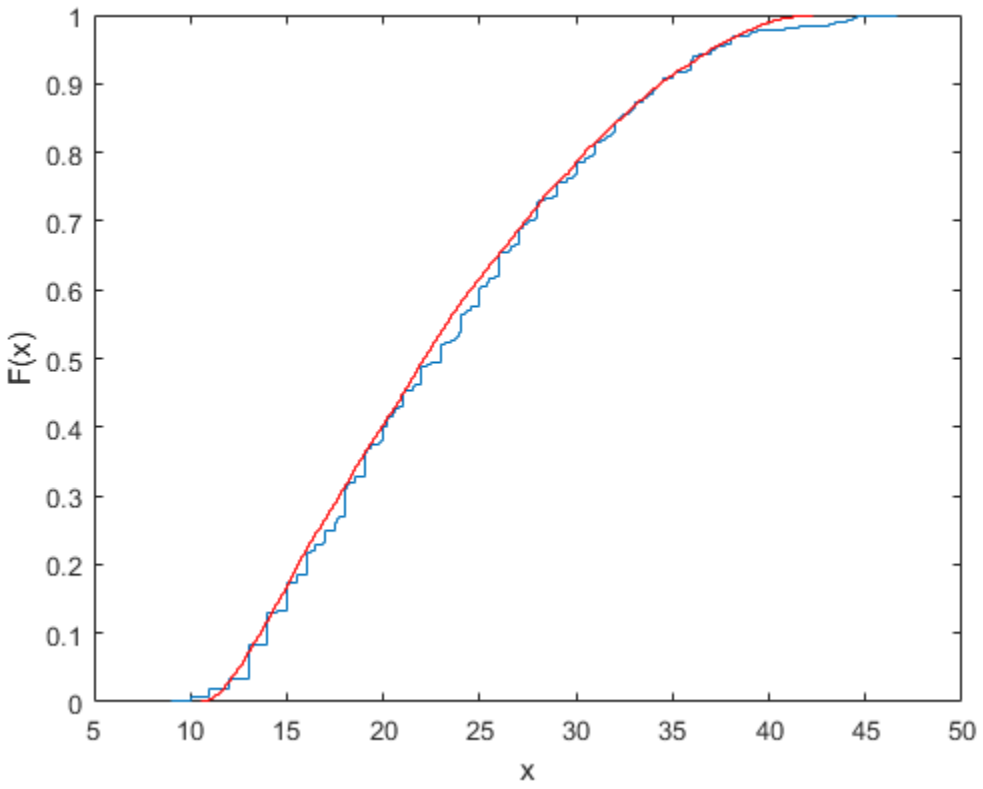
The optional second output from `johnsrnd` indicates which type of distribution within the Johnson system matches the quantiles.

```
type
```

```
type =
'SB'
```

You can verify that the sample resembles the original data by overlaying the empirical cumulative distribution functions.

```
ecdf(MPG);
[Fi,xi] = ecdf(r1);
hold on;
stairs(xi,Fi,'r');
hold off
```

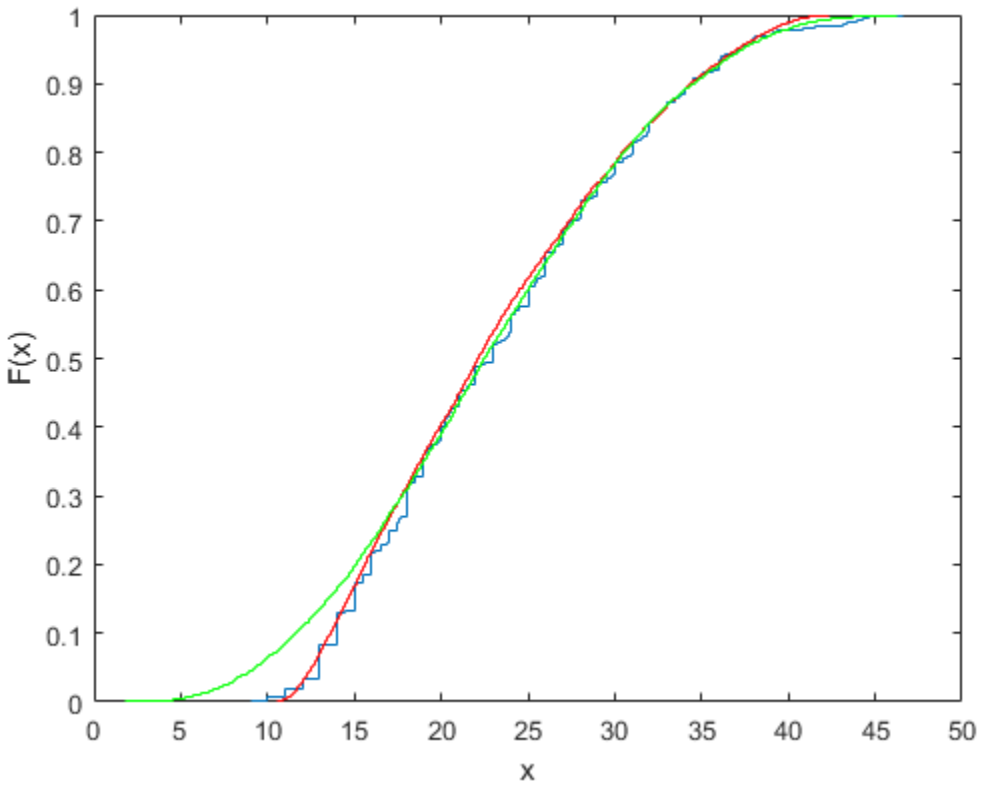


In some applications, it may be important to match the quantiles better in some regions of the data than in others. To do that, specify four evenly spaced standard normal quantiles at which you want to match the data, instead of the default -1.5, -0.5, 0.5, and 1.5. For example, you might care more about matching the data in the right tail than in the left, and so you specify standard normal quantiles that emphasizes the right tail.

```
qnorm = [-.5 .25 1 1.75];
probs = normcdf(qnorm);
qemp = quantile(MPG,probs);
r2 = johnsrnd([qnorm; qemp],10000,1);
```

However, while the new sample matches the original data better in the right tail, it matches much worse in the left tail.

```
[Fj,xj] = ecdf(r2);
hold on;
stairs(xj,Fj,'g');
hold off
```



See Also

`ecdf` | `johnsrnd` | `pearsrnd`

Bayesian Linear Regression Using Hamiltonian Monte Carlo

This example shows how to perform Bayesian inference on a linear regression model using a Hamiltonian Monte Carlo (HMC) sampler.

In Bayesian parameter inference, the goal is to analyze statistical models with the incorporation of prior knowledge of model parameters. The posterior distribution of the free parameters θ combines the likelihood function $P(y|\theta)$ with the prior distribution $P(\theta)$, using Bayes' theorem:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}.$$

Usually, the best way to summarize the posterior distribution is to obtain samples from that distribution using Monte Carlo methods. Using these samples, you can estimate marginal posterior distributions and derived statistics such as the posterior mean, median, and standard deviation. HMC is a gradient-based Markov Chain Monte Carlo sampler that can be more efficient than standard samplers, especially for medium-dimensional and high-dimensional problems.

Linear Regression Model

Analyze a linear regression model with the intercept α, the linear coefficients β (a column vector), and the noise variance σ^2 of the data distribution as free parameters. Assume that each data point has an independent Gaussian distribution:

$$y_i|\theta \sim \mathcal{N}(\mu_i(\theta), \sigma^2).$$

Model the mean μ_i of the Gaussian distribution as a function of the predictors x_i and model parameters as

$$\mu_i = \alpha + x_i^T\beta.$$

In a Bayesian analysis, you also must assign prior distributions to all free parameters. Assign independent Gaussian priors on the intercept and linear coefficients:

$\alpha \sim \mathcal{N}(\alpha_0, \sigma_\alpha^2)$,

$\beta_i \sim \mathcal{N}(\beta_0, \sigma_\beta^2)$.

To use HMC, all sampling variables must be unconstrained, meaning that the posterior density and its gradient must be well-defined for all real parameter values. If you have a parameter that is constrained to an interval, then you must transform this parameter into an unbounded one. To conserve probability, you must multiply the prior distribution by the corresponding Jacobian factor. Also, take this factor into account when calculating the gradient of the posterior.

The noise variance is a (squared) scale parameter that can only be positive. It then can be easier and more natural to consider its logarithm as the free parameter, which is unbounded. Assign a normal prior to the logarithm of the noise variance:

$\log\sigma^2 \sim \mathcal{N}(\kappa, \omega^2)$.

Write the logarithm of the posterior density of the free parameters $\theta = (\alpha; \beta; \log\sigma^2)$ as

$$\log P(\theta|y) = \text{const.} + \log P(y|\theta) + \log P(\theta).$$

Ignore the constant term and call the sum of the last two terms $g(\theta)$. To use HMC, create a function handle that evaluates $g(\theta)$ and its gradient $\partial g/\partial\theta$ for any value of θ. The functions used to calculate $g(\theta)$ are located at the end of the script.

Create Data Set

Define true parameter values for the intercept, the linear coefficients `Beta`, and the noise standard deviation. Knowing the true parameter values makes it possible to compare with the output of the HMC sampler. Only the first predictor affects the response.

```
NumPredictors = 2;

trueIntercept = 2;
trueBeta = [3;0];
trueNoiseSigma = 1;
```

Use these parameter values to create a normally distributed sample data set at random values of the two predictors.

```
NumData = 100;
rng('default') %For reproducibility
X = rand(NumData,NumPredictors);
mu = X*trueBeta + trueIntercept;
y = normrnd(mu,trueNoiseSigma);
```

Define Posterior Probability Density

Choose the means and standard deviations of the Gaussian priors.

```
InterceptPriorMean = 0;
InterceptPriorSigma = 10;
BetaPriorMean = 0;
BetaPriorSigma = 10;
LogNoiseVarianceMean = 0;
LogNoiseVarianceSigma = 2;
```

Save a function `logPosterior` on the MATLAB® path that returns the logarithm of the product of the prior and likelihood, and the gradient of this logarithm. The `logPosterior` function is defined at the end of this example. Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
logpdf = @(Parameters)logPosterior(Parameters,X,y, ...
    InterceptPriorMean,InterceptPriorSigma, ...
    BetaPriorMean,BetaPriorSigma, ...
    LogNoiseVarianceMean,LogNoiseVarianceSigma);
```

Create HMC Sampler

Define the initial point to start sampling from, and then call the `hmcSampler` function to create the Hamiltonian sampler as a `HamiltonianSampler` object. Display the sampler properties.

```
Intercept = randn;
Beta = randn(NumPredictors,1);
LogNoiseVariance = randn;
```

```
startpoint = [Intercept;Beta;LogNoiseVariance];
smp = hmcSampler(logpdf,startpoint,'NumSteps',50);

smp
```

```
smp =

  HamiltonianSampler with properties:

                StepSize: 0.1000
                NumSteps: 50
              MassVector: [4x1 double]
            JitterMethod: 'jitter-both'
      StepSizeTuningMethod: 'dual-averaging'
    MassVectorTuningMethod: 'iterative-sampling'
                   LogPDF: [function_handle]
             VariableNames: {4x1 cell}
                StartPoint: [4x1 double]
```

Estimate MAP Point

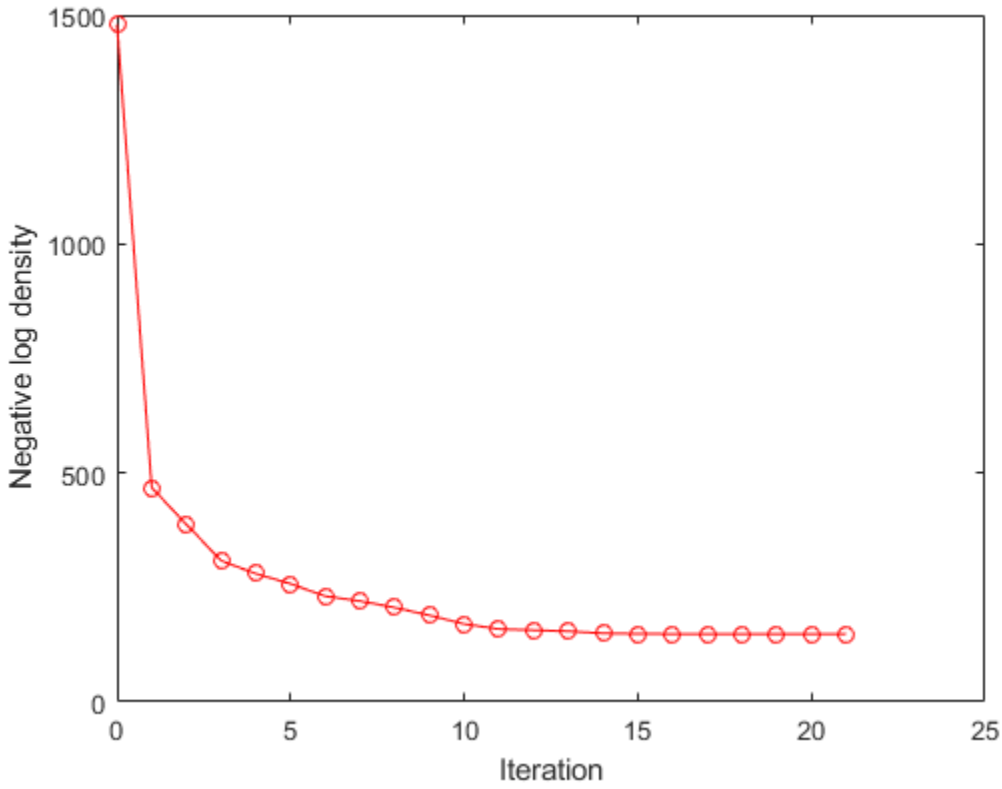
Estimate the MAP (maximum-a-posteriori) point of the posterior density. You can start sampling from any point, but it is often more efficient to estimate the MAP point, and then use it as a starting point for tuning the sampler and drawing samples. Estimate and display the MAP point. You can show more information during optimization by setting the `'VerbosityLevel'` value to 1.

```
[MAPpars,fitInfo] = estimateMAP(smp,'VerbosityLevel',0);
MAPIntercept = MAPpars(1)
MAPBeta = MAPpars(2:end-1)
MAPLogNoiseVariance = MAPpars(end)
```

```
MAPIntercept =

    2.3857


MAPBeta =

    2.5495
   -0.4508


MAPLogNoiseVariance =

   -0.1007
```

To check that the optimization has converged to a local optimum, plot the `fitInfo.Objective` field. This field contains the values of the negative log density at each iteration of the function optimization. The final values are all similar, so the optimization has converged.

```
plot(fitInfo.Iteration,fitInfo.Objective,'ro-');
xlabel('Iteration');
ylabel('Negative log density');
```



Tune Sampler

It is important to select good values for the sampler parameters to get efficient sampling. The best way to find good values is to automatically tune the `MassVector`, `StepSize`, and `NumSteps` parameters using the `tuneSampler` method. Use the method to:

1. Tune the `MassVector` of the sampler.

2. Tune `StepSize` and `NumSteps` for a fixed simulation length to achieve a certain acceptance ratio. The default target acceptance ratio of 0.65 is good in most cases.

Start tuning at the estimated MAP point for more efficient tuning.

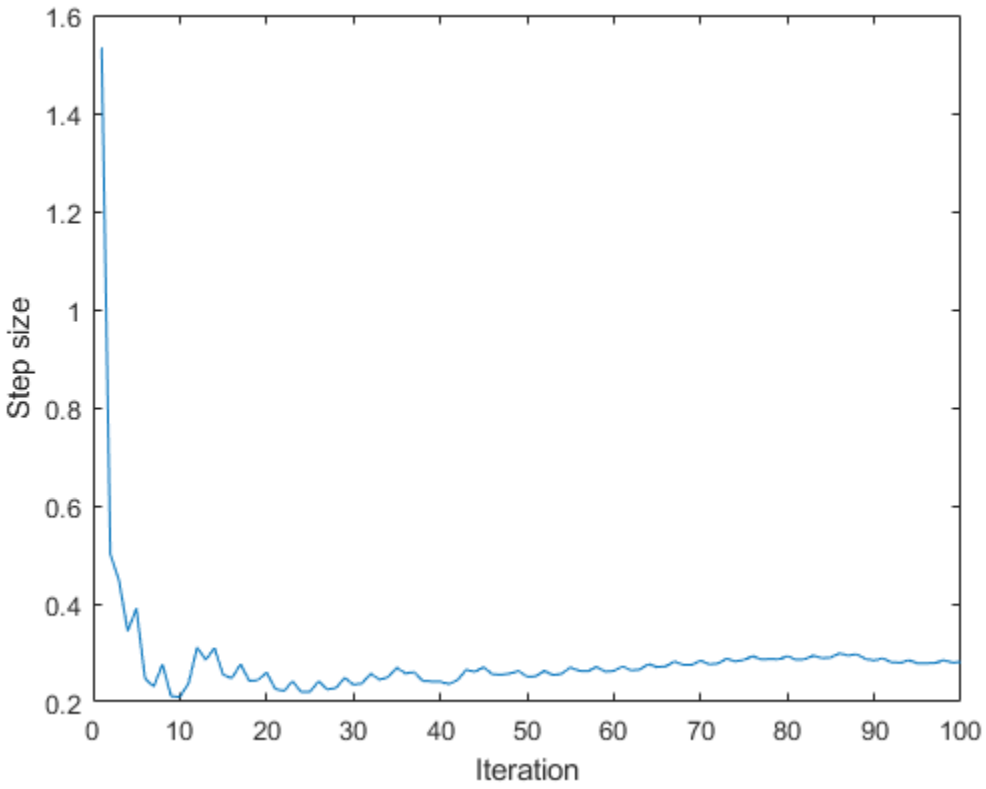
```
[smp,tuneinfo] = tuneSampler(smp,'Start',MAPpars);
```

Plot the evolution of the step size during tuning to ensure that the step size tuning has converged. Display the achieved acceptance ratio.

```
figure;
plot(tuneinfo.StepSizeTuningInfo.StepSizeProfile);
xlabel('Iteration');
ylabel('Step size');

accratio = tuneinfo.StepSizeTuningInfo.AcceptanceRatio
```

```
accratio =
```

```
    0.6400
```



Draw Samples

Draw samples from the posterior density, using a few independent chains. Choose different initial points for the chains, randomly distributed around the estimated MAP point. Specify the number of burn-in samples to discard from the beginning of the Markov chain and the number of samples to generate after the burn-in.

Set the `'VerbosityLevel'` value to print details during sampling for the first chain.

```
NumChains = 4;
chains = cell(NumChains,1);
Burnin = 500;
NumSamples = 1000;
for c = 1:NumChains
    if (c == 1)
        level = 1;
    else
        level = 0;
    end
    chains{c} = drawSamples(smp,'Start',MAPpars + randn(size(MAPpars)), ...
        'Burnin',Burnin,'NumSamples',NumSamples, ...
        'VerbosityLevel',level,'NumPrint',300);
end
```

```
|=====================================================================================|
|   ITER   |    LOG PDF    |  STEP SIZE  |  NUM STEPS  |  ACC RATIO  |  DIVERGENT  |
|=====================================================================================|
|      300 | -1.484164e+02 |   2.532e-01 |          12 |   9.500e-01 |           0 |
|      600 | -1.492436e+02 |   2.128e-02 |           4 |   9.450e-01 |           0 |
|      900 | -1.509753e+02 |   2.171e-01 |           5 |   9.444e-01 |           0 |
|     1200 | -1.493455e+02 |   1.128e-01 |          16 |   9.358e-01 |           0 |
|     1500 | -1.489602e+02 |   2.532e-01 |          12 |   9.373e-01 |           0 |
```

Examine Convergence Diagnostics

Use the `diagnostics` method to compute standard MCMC diagnostics. For each sampling parameter, the method uses all the chains to compute these statistics:

- Posterior mean estimate (`Mean`)
- Estimate of the Monte Carlo standard error (`MCSE`), which is the standard deviation of the posterior mean estimate
- Estimate of the posterior standard deviation (`SD`)
- Estimates of the 5th and 95th quantiles of the marginal posterior distribution (`Q5` and `Q95`)
- Effective sample size for the posterior mean estimate (`ESS`)
- Gelman-Rubin convergence statistic (`RHat`). As a rule of thumb, values of `RHat` less than `1.1` are interpreted as a sign that the chain has converged to the desired distribution. If `RHat` for any variable is larger than `1.1`, then try drawing more samples using the `drawSamples` method.

Display the diagnostics table and the true values of the sampling parameters defined in the beginning of the example. Since the prior distribution is noninformative for this data set, the true values are between (or near) the 5th and 95th quantiles.

```
diags = diagnostics(smp,chains)
truePars = [trueIntercept;trueBeta;log(trueNoiseSigma^2)]
```

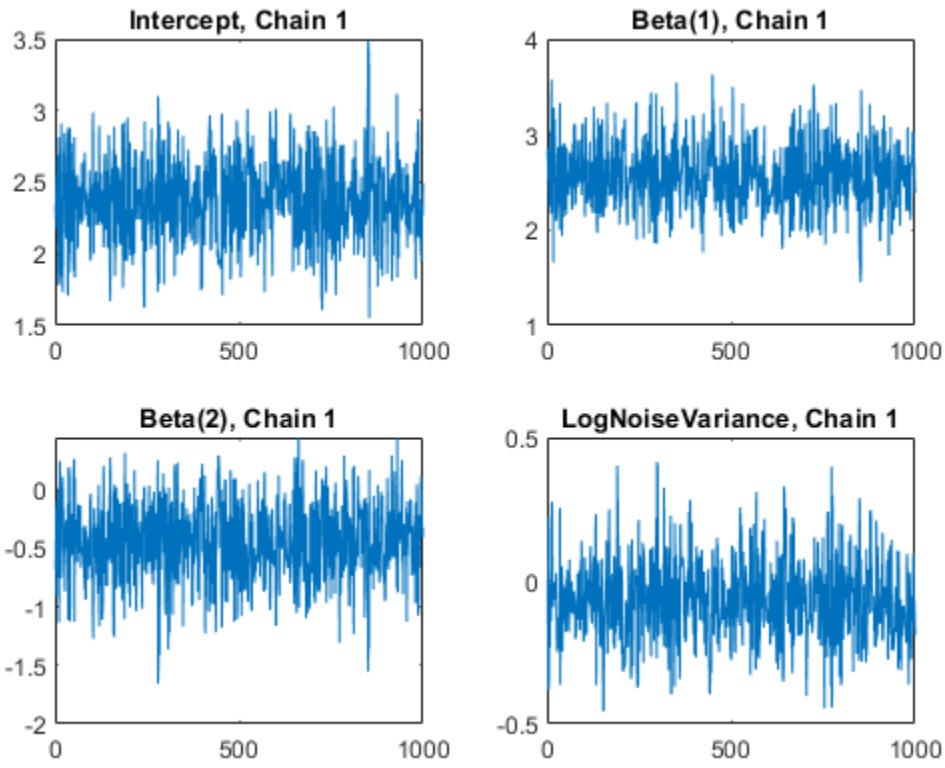
```
diags =

  4x8 table

    Name       Mean         MCSE         SD          Q5          Q95        ESS       RHat
    ______    _________    _________    _______    ________    _______    ______    ______

    {'x1'}      2.3805    0.0053354    0.28028      1.8966     2.8335    2759.7    1.0005
    {'x2'}      2.5544    0.0062264    0.33478      1.9989     3.1026    2890.9    1.0004
    {'x3'}    -0.44516    0.0064468     0.3415     -1.0247    0.10198      2806         1
    {'x4'}    -0.062826    0.0028183    0.14112    -0.28345    0.17997    2507.5    1.0004


truePars =

     2
     3
     0
     0
```

Visualize Samples

Investigate issues such as convergence and mixing to determine whether the drawn samples represent a reasonable set of random realizations from the target distribution. To examine the output, plot the trace plots of the samples using the first chain.

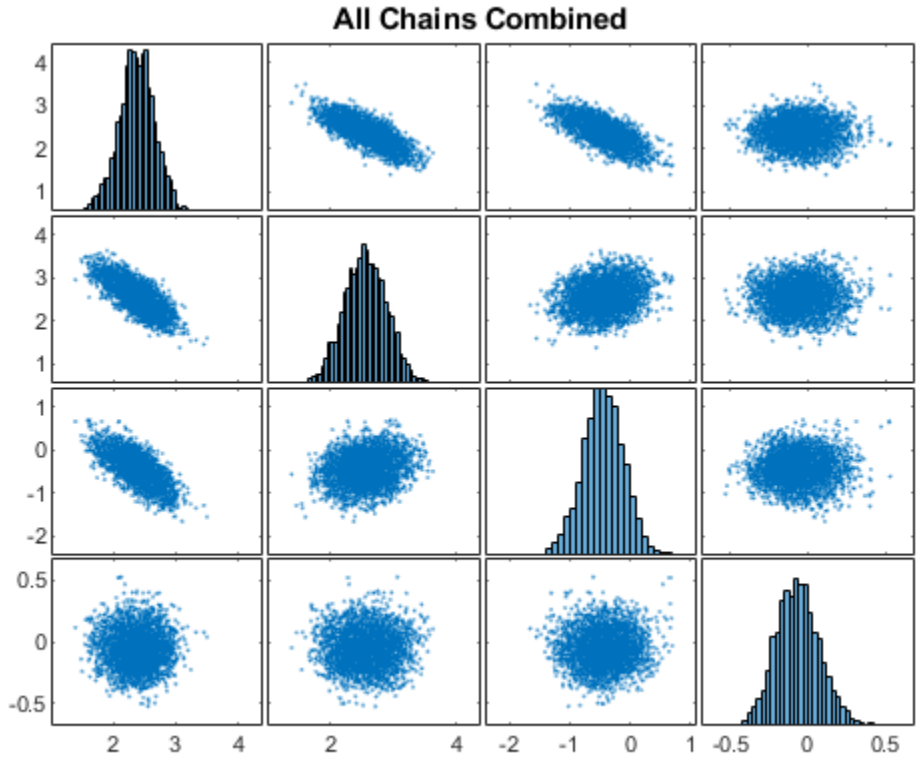
The `drawSamples` method discards burn-in samples from the beginning of the Markov chain to reduce the effect of the sampling starting point. Furthermore, the trace plots look like high-frequency noise, without any visible long-range correlation between the samples. This behavior indicates that the chain is mixed well.

```
figure;
subplot(2,2,1);
plot(chains{1}(:,1));
title(sprintf('Intercept, Chain 1'));
for p = 2:1+NumPredictors
    subplot(2,2,p);
    plot(chains{1}(:,p));
    title(sprintf('Beta(%d), Chain 1',p-1));
end
subplot(2,2,4);
plot(chains{1}(:,end));
title(sprintf('LogNoiseVariance, Chain 1'));
```



Combine the chains into one matrix and create scatter plots and histograms to visualize the 1-D and 2-D marginal posterior distributions.

```
concatenatedSamples = vertcat(chains{:});
figure;
plotmatrix(concatenatedSamples);
title('All Chains Combined');
```



Functions for Computing Posterior Distribution

The `logPosterior` function returns the logarithm of the product of a normal likelihood and a normal prior for the linear model. The input argument `Parameter` has the format `[Intercept;Beta;LogNoiseVariance]`. `X` and `Y` contain the values of the predictors and response, respectively.

The `normalPrior` function returns the logarithm of the multivariate normal probability density with means `Mu` and standard deviations `Sigma`, specified as scalars or columns vectors the same length as `P`. The second output argument is the corresponding gradient.

```
function [logpdf, gradlogpdf] = logPosterior(Parameters,X,Y, ...
    InterceptPriorMean,InterceptPriorSigma, ...
    BetaPriorMean,BetaPriorSigma, ...
    LogNoiseVarianceMean,LogNoiseVarianceSigma)


% Unpack the parameter vector
Intercept        = Parameters(1);
Beta             = Parameters(2:end-1);
LogNoiseVariance = Parameters(end);
% Compute the log likelihood and its gradient
```

```
Sigma                   = sqrt(exp(LogNoiseVariance));
Mu                      = X*Beta + Intercept;
Z                       = (Y - Mu)/Sigma;
loglik                  = sum(-log(Sigma) - .5*log(2*pi) - .5*Z.^2);
gradIntercept1          = sum(Z/Sigma);
gradBeta1               = X'*Z/Sigma;
gradLogNoiseVariance1    = sum(-.5 + .5*(Z.^2));
% Compute log priors and gradients
[LPIntercept, gradIntercept2]           = normalPrior(Intercept,InterceptPriorMean,InterceptPrior
[LPBeta, gradBeta2]                     = normalPrior(Beta,BetaPriorMean,BetaPriorSigma);
[LPLogNoiseVar, gradLogNoiseVariance2]  = normalPrior(LogNoiseVariance,LogNoiseVarianceMean,LogNo
logprior                                = LPIntercept + LPBeta + LPLogNoiseVar;
% Return the log posterior and its gradient
logpdf               = loglik + logprior;
gradIntercept        = gradIntercept1 + gradIntercept2;
gradBeta             = gradBeta1 + gradBeta2;
gradLogNoiseVariance = gradLogNoiseVariance1 + gradLogNoiseVariance2;
gradlogpdf           = [gradIntercept;gradBeta;gradLogNoiseVariance];
end

function [logpdf,gradlogpdf] = normalPrior(P,Mu,Sigma)
Z          = (P - Mu)./Sigma;
logpdf     = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
gradlogpdf = -Z./Sigma;
end
```

See Also

Functions
hmcSampler

Classes
HamiltonianSampler

Bayesian Analysis for a Logistic Regression Model

This example shows how to make Bayesian inferences for a logistic regression model using `slicesample`.

Statistical inferences are usually based on maximum likelihood estimation (MLE). MLE chooses the parameters that maximize the likelihood of the data, and is intuitively appealing. In MLE, parameters are assumed to be unknown but fixed, and are estimated with some confidence. In Bayesian statistics, the uncertainty about the unknown parameters is quantified using probability so that the unknown parameters are regarded as random variables.

Bayesian Inference

Bayesian inference is the process of analyzing statistical models with the incorporation of prior knowledge about the model or model parameters. The root of such inference is Bayes' theorem:

$$P(\text{parameters}|\text{data}) = \frac{P(\text{data}|\text{parameters}) \times P(\text{parameters})}{P(\text{data})} \propto \text{likelihood} \times \text{prior}$$

For example, suppose we have normal observations

$$X|\theta \sim N(\theta, \sigma^2)$$

where sigma is known and the prior distribution for theta is

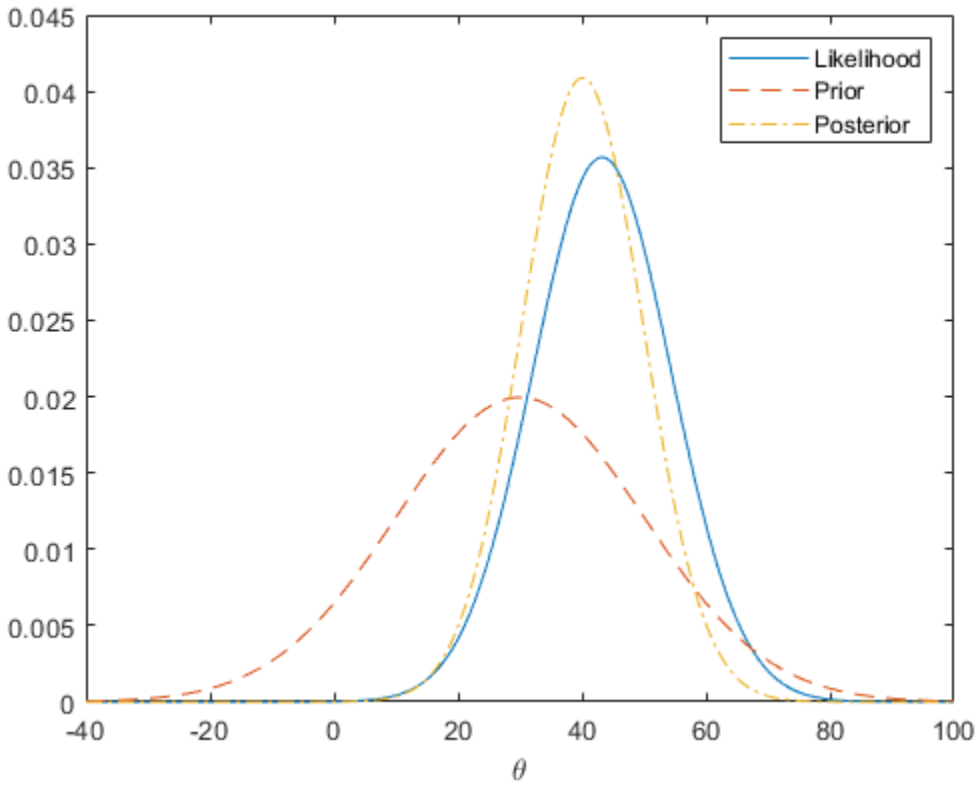
$$\theta \sim N(\mu, \tau^2)$$

In this formula mu and tau, sometimes known as hyperparameters, are also known. If we observe n samples of X, we can obtain the posterior distribution for theta as

$$\theta|X \sim N\left(\frac{\tau^2}{\sigma^2/n + \tau^2}\bar{X} + \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu, \frac{(\sigma^2/n) \times \tau^2}{\sigma^2/n + \tau^2}\right)$$

The following graph shows the prior, likelihood, and posterior for theta.

```
rng(0,'twister');

n = 20;
sigma = 50;
x = normrnd(10,sigma,n,1);
mu = 30;
tau = 20;
theta = linspace(-40, 100, 500);
y1 = normpdf(mean(x),theta,sigma/sqrt(n));
y2 = normpdf(theta,mu,tau);
postMean = tau^2*mean(x)/(tau^2+sigma^2/n) + sigma^2*mu/n/(tau^2+sigma^2/n);
postSD = sqrt(tau^2*sigma^2/n/(tau^2+sigma^2/n));
y3 = normpdf(theta, postMean,postSD);
plot(theta,y1,'-', theta,y2,'--', theta,y3,'-.')
legend('Likelihood','Prior','Posterior')
xlabel('\theta')
```



Car Experiment Data

In some simple problems such as the previous normal mean inference example, it is easy to figure out the posterior distribution in a closed form. But in general problems that involve non-conjugate priors, the posterior distributions are difficult or impossible to compute analytically. We will consider logistic regression as an example. This example involves an experiment to help model the proportion of cars of various weights that fail a mileage test. The data include observations of weight, number of cars tested, and number failed. We will work with a transformed version of the weights to reduce the correlation in our estimates of the regression parameters.

```
% A set of car weights
weight = [2100 2300 2500 2700 2900 3100 3300 3500 3700 3900 4100 4300]';
weight = (weight-2800)/1000;     % recenter and rescale
% The number of cars tested at each weight
total = [48 42 31 34 31 21 23 23 21 16 17 21]';
% The number of cars that have poor mpg performances at each weight
poor = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

Logistic Regression Model

Logistic regression, a special case of a generalized linear model, is appropriate for these data since the response variable is binomial. The logistic regression model can be written as:

$$P(\text{failure}) = \frac{e^{Xb}}{1 + e^{Xb}}$$

where X is the design matrix and b is the vector containing the model parameters. In MATLAB®, we can write this equation as:

```
logitp = @(b,x) exp(b(1)+b(2).*x)./(1+exp(b(1)+b(2).*x));
```

If you have some prior knowledge or some non-informative priors are available, you could specify the prior probability distributions for the model parameters. For instance, in this example, we use normal priors for the intercept `b1` and slope `b2`, i.e.

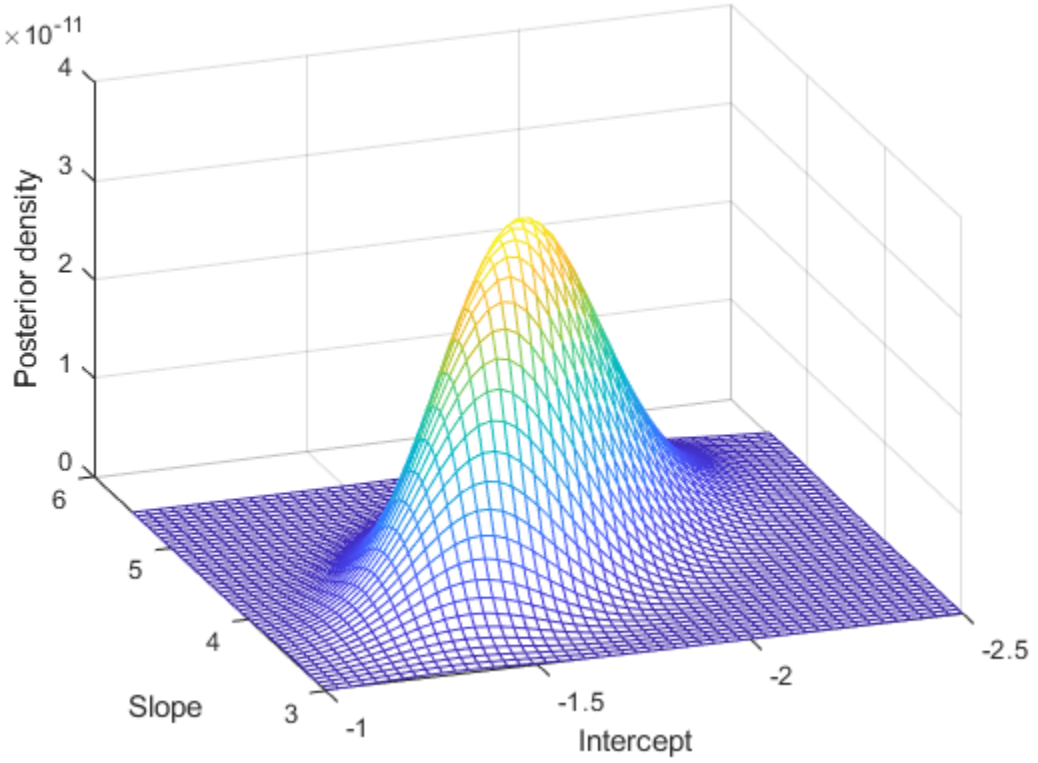
```
prior1 = @(b1) normpdf(b1,0,20);    % prior for intercept
prior2 = @(b2) normpdf(b2,0,20);    % prior for slope
```

By Bayes' theorem, the joint posterior distribution of the model parameters is proportional to the product of the likelihood and priors.

```
post = @(b) prod(binopdf(poor,total,logitp(b,weight))) ...  % likelihood
            * prior1(b(1)) * prior2(b(2));                  % priors
```

Note that the normalizing constant for the posterior in this model is analytically intractable. However, even without knowing the normalizing constant, you can visualize the posterior distribution, if you know the approximate range of the model parameters.

```
b1 = linspace(-2.5, -1, 50);
b2 = linspace(3, 5.5, 50);
simpost = zeros(50,50);
for i = 1:length(b1)
    for j = 1:length(b2)
        simpost(i,j) = post([b1(i), b2(j)]);
    end;
end;
mesh(b2,b1,simpost)
xlabel('Slope')
ylabel('Intercept')
zlabel('Posterior density')
view(-110,30)
```



This posterior is elongated along a diagonal in the parameter space, indicating that, after we look at the data, we believe that the parameters are correlated. This is interesting, since before we collected any data we assumed they were independent. The correlation comes from combining our prior distribution with the likelihood function.

Slice Sampling

Monte Carlo methods are often used in Bayesian data analysis to summarize the posterior distribution. The idea is that, even if you cannot compute the posterior distribution analytically, you can generate a random sample from the distribution and use these random values to estimate the posterior distribution or derived statistics such as the posterior mean, median, standard deviation, etc. Slice sampling is an algorithm designed to sample from a distribution with an arbitrary density function, known only up to a constant of proportionality -- exactly what is needed for sampling from a complicated posterior distribution whose normalization constant is unknown. The algorithm does not generate independent samples, but rather a Markovian sequence whose stationary distribution is the target distribution. Thus, the slice sampler is a Markov Chain Monte Carlo (MCMC) algorithm. However, it differs from other well-known MCMC algorithms because only the scaled posterior need be specified -- no proposal or marginal distributions are needed.

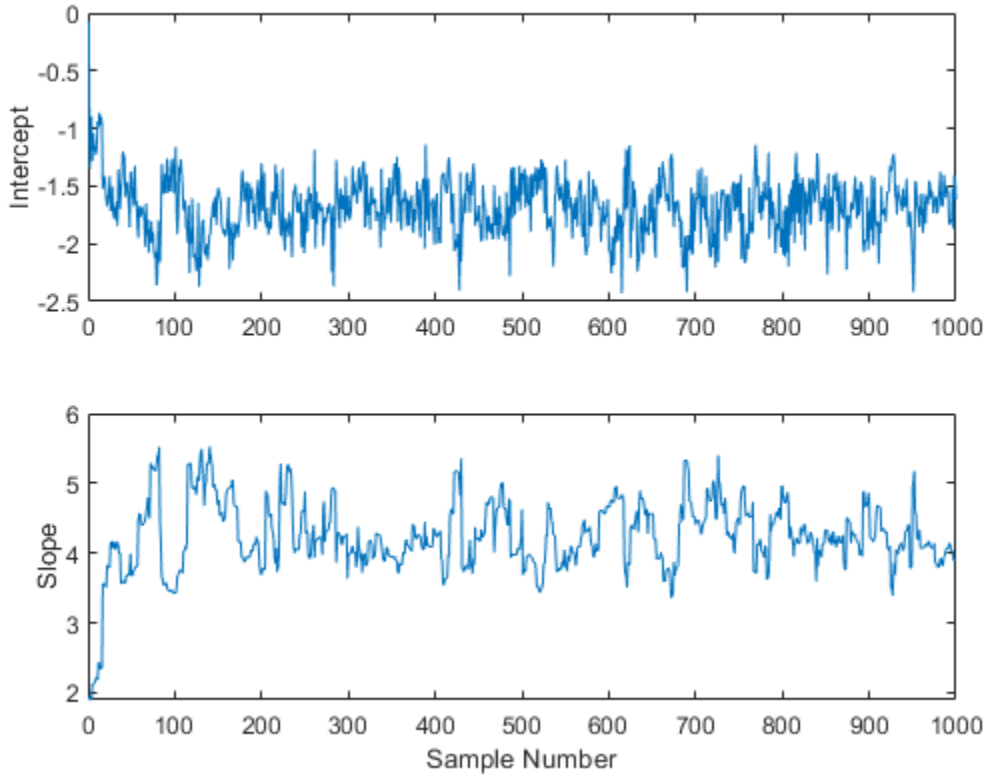
This example shows how to use the slice sampler as part of a Bayesian analysis of the mileage test logistic regression model, including generating a random sample from the posterior distribution for the model parameters, analyzing the output of the sampler, and making inferences about the model parameters. The first step is to generate a random sample.

```
initial = [1 1];
nsamples = 1000;
trace = slicesample(initial,nsamples,'pdf',post,'width',[20 2]);
```

Analysis of Sampler Output

After obtaining a random sample from the slice sampler, it is important to investigate issues such as convergence and mixing, to determine whether the sample can reasonably be treated as a set of random realizations from the target posterior distribution. Looking at marginal trace plots is the simplest way to examine the output.

```
subplot(2,1,1)
plot(trace(:,1))
ylabel('Intercept');
subplot(2,1,2)
plot(trace(:,2))
ylabel('Slope');
xlabel('Sample Number');
```

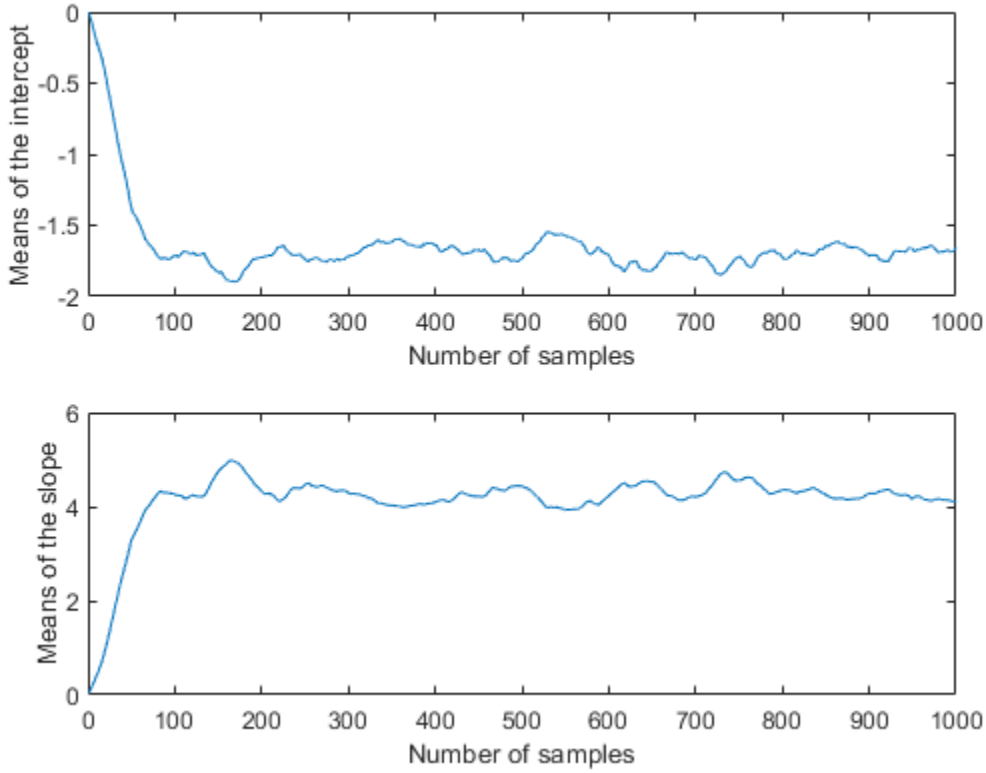


It is apparent from these plots is that the effects of the parameter starting values take a while to disappear (perhaps fifty or so samples) before the process begins to look stationary.

It is also be helpful in checking for convergence to use a moving window to compute statistics such as the sample mean, median, or standard deviation for the sample. This produces a smoother plot than the raw sample traces, and can make it easier to identify and understand any non-stationarity.

```
movavg = filter( (1/50)*ones(50,1), 1, trace);
subplot(2,1,1)
```

```
plot(movavg(:,1))
xlabel('Number of samples')
ylabel('Means of the intercept');
subplot(2,1,2)
plot(movavg(:,2))
xlabel('Number of samples')
ylabel('Means of the slope');
```

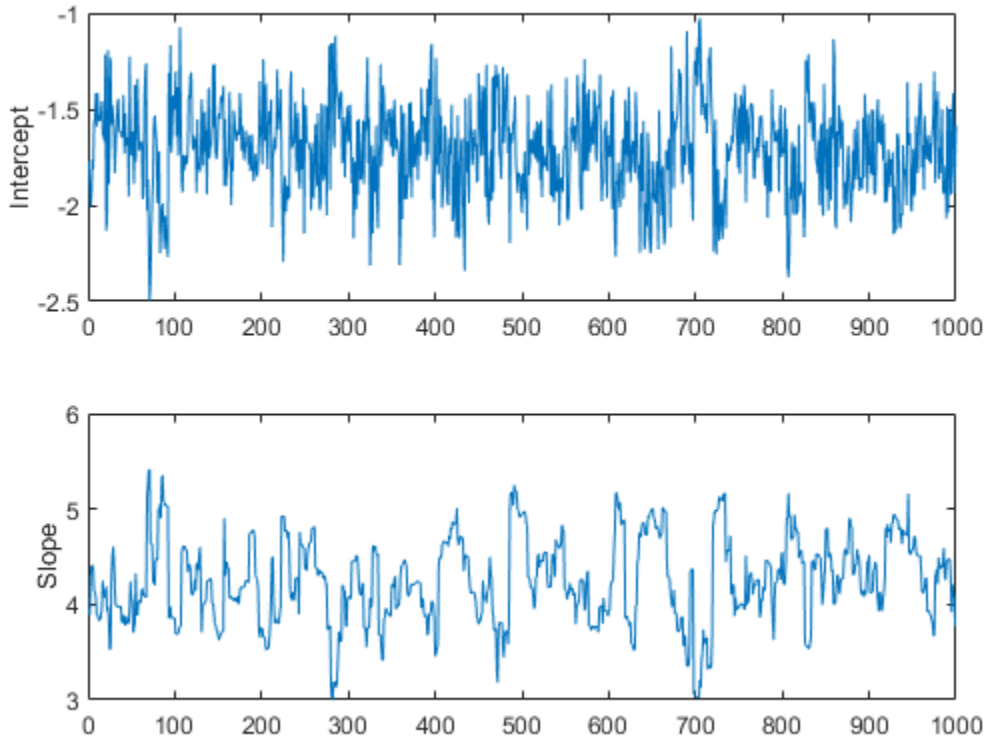


Because these are moving averages over a window of 50 iterations, the first 50 values are not comparable to the rest of the plot. However, the remainder of each plot seems to confirm that the parameter posterior means have converged to stationarity after 100 or so iterations. It is also apparent that the two parameters are correlated with each other, in agreement with the earlier plot of the posterior density.

Since the settling-in period represents samples that cannot reasonably be treated as random realizations from the target distribution, it's probably advisable not to use the first 50 or so values at the beginning of the slice sampler's output. You could just delete those rows of the output, however, it's also possible to specify a "burn-in" period. This is convenient when a suitable burn-in length is already known, perhaps from previous runs.

```
trace = slicesample(initial,nsamples,'pdf',post, ...
                                     'width',[20 2],'burnin',50);
subplot(2,1,1)
plot(trace(:,1))
ylabel('Intercept');
subplot(2,1,2)
```

```
plot(trace(:,2))
ylabel('Slope');
```



These trace plots do not seem to show any non-stationarity, indicating that the burn-in period has done its job.

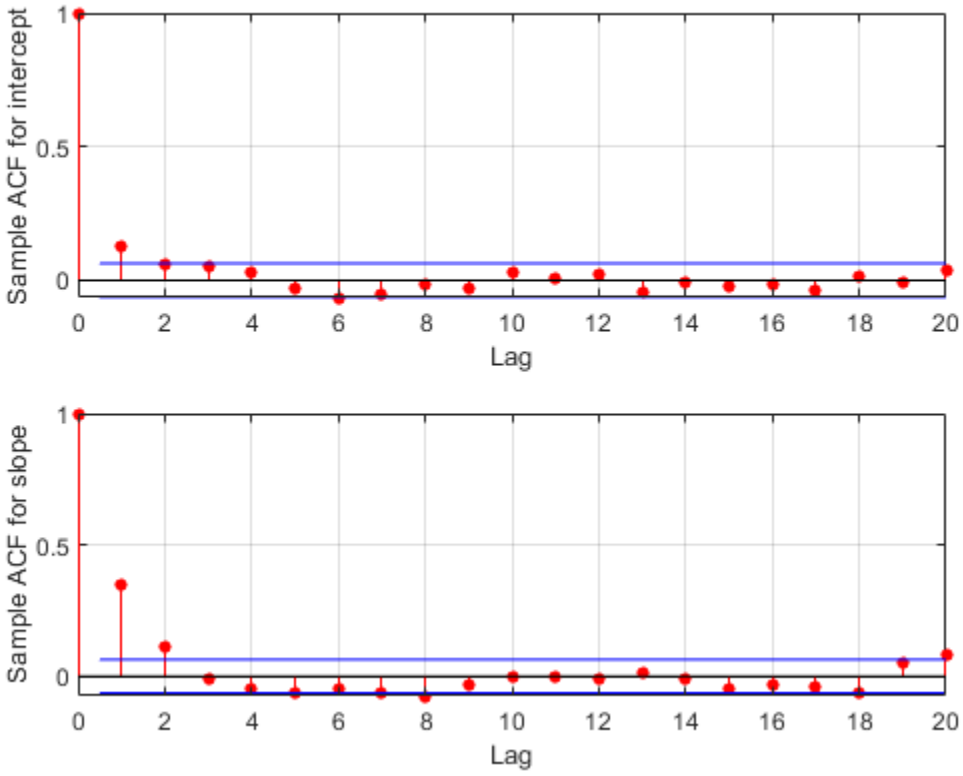
However, there is a second aspect of the trace plots that should also be explored. While the trace for the intercept looks like high frequency noise, the trace for the slope appears to have a lower frequency component, indicating there autocorrelation between values at adjacent iterations. We could still compute the mean from this autocorrelated sample, but it is often convenient to reduce the storage requirements by removing redundancy in the sample. If this eliminated the autocorrelation, it would also allow us to treat this as a sample of independent values. For example, you can thin out the sample by keeping only every 10th value.

```
trace = slicesample(initial,nsamples,'pdf',post,'width',[20 2], ...
                                                 'burnin',50,'thin',10);
```

To check the effect of this thinning, it is useful to estimate the sample autocorrelation functions from the traces and use them to check if the samples mix rapidly.

```
F    =  fft(detrend(trace,'constant'));
F    =  F .* conj(F);
ACF  =  ifft(F);
ACF  =  ACF(1:21,:);                          % Retain lags up to 20.
ACF  =  real([ACF(1:21,1) ./ ACF(1,1) ...
             ACF(1:21,2) ./ ACF(1,2)]);       % Normalize.
bounds  =  sqrt(1/nsamples) * [2 ; -2];       % 95% CI for iid normal
```

```
labs = {'Sample ACF for intercept','Sample ACF for slope' };
for i = 1:2
    subplot(2,1,i)
    lineHandles  =  stem(0:20, ACF(:,i) , 'filled' , 'r-o');
    lineHandles.MarkerSize = 4;
    grid('on')
    xlabel('Lag')
    ylabel(labs{i})
    hold on
    plot([0.5 0.5 ; 20 20] , [bounds([1 1]) bounds([2 2])] , '-b');
    plot([0 20] , [0 0] , '-k');
    hold off
    a  =  axis;
    axis([a(1:3) 1]);
end
```



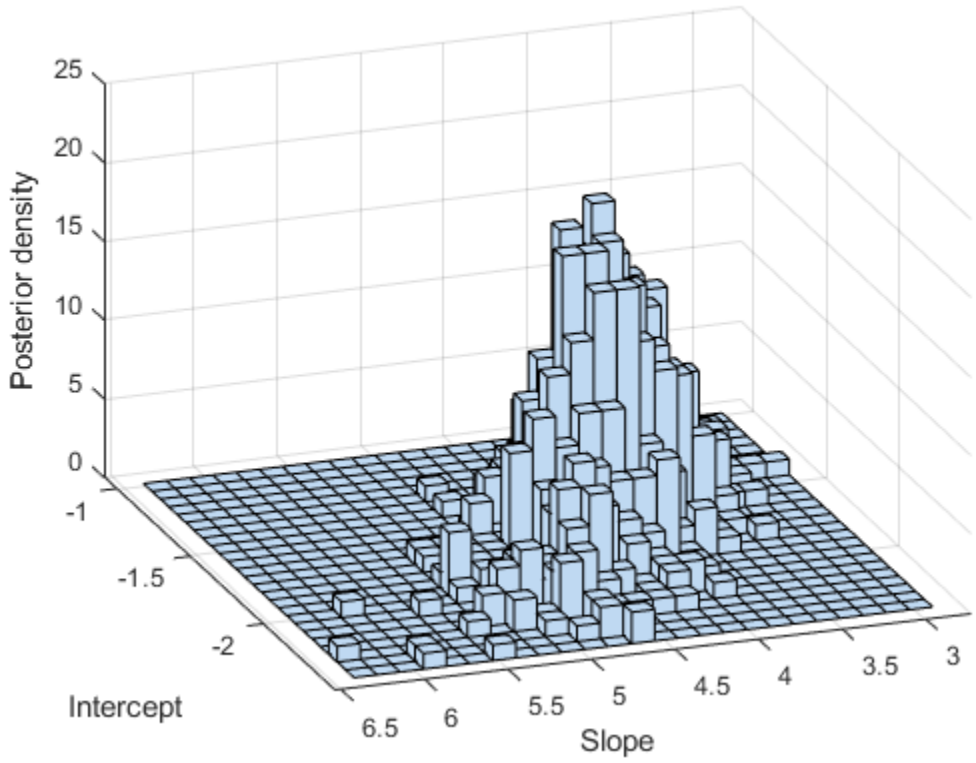
The autocorrelation values at the first lag are significant for the intercept parameter, and even more so for the slope parameter. We could repeat the sampling using a larger thinning parameter in order to reduce the correlation further. For the purposes of this example, however, we'll continue to use the current sample.

Inference for the Model Parameters

As expected, a histogram of the sample mimics the plot of the posterior density.

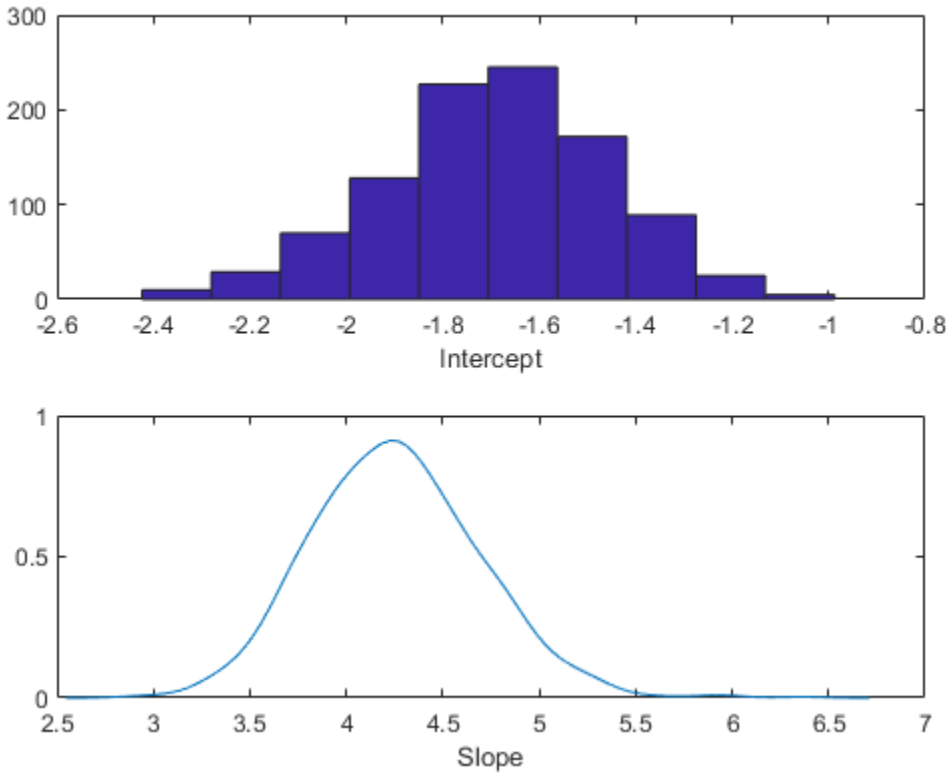
```
subplot(1,1,1)
hist3(trace,[25,25]);
```

```
xlabel('Intercept')
ylabel('Slope')
zlabel('Posterior density')
view(-110,30)
```



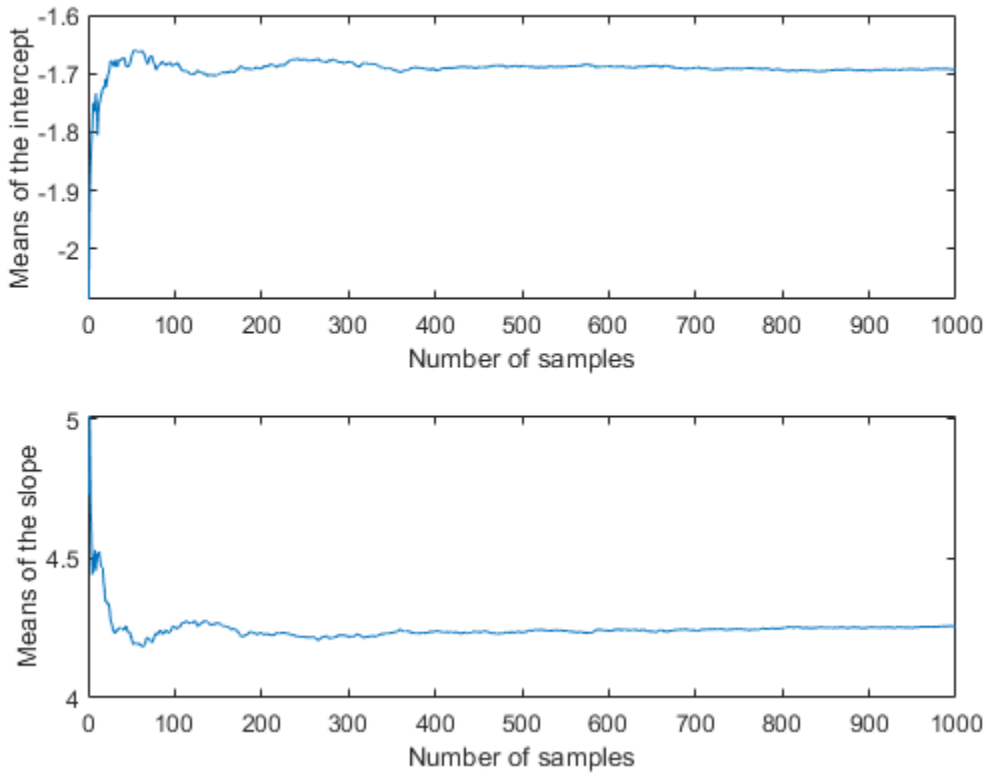
You can use a histogram or a kernel smoothing density estimate to summarize the marginal distribution properties of the posterior samples.

```
subplot(2,1,1)
hist(trace(:,1))
xlabel('Intercept');
subplot(2,1,2)
ksdensity(trace(:,2))
xlabel('Slope');
```



You could also compute descriptive statistics such as the posterior mean or percentiles from the random samples. To determine if the sample size is large enough to achieve a desired precision, it is helpful to monitor the desired statistic of the traces as a function of the number of samples.

```
csum = cumsum(trace);
subplot(2,1,1)
plot(csum(:,1)'./(1:nsamples))
xlabel('Number of samples')
ylabel('Means of the intercept');
subplot(2,1,2)
plot(csum(:,2)'./(1:nsamples))
xlabel('Number of samples')
ylabel('Means of the slope');
```



In this case, it appears that the sample size of 1000 is more than sufficient to give good precision for the posterior mean estimate.

```
bHat = mean(trace)
```

```
bHat =

   -1.6931    4.2569
```

Summary

The Statistics and Machine Learning Toolbox™ offers a variety of functions that allow you to specify likelihoods and priors easily. They can be combined to derive a posterior distribution. The `slicesample` function enables you to carry out Bayesian analysis in MATLAB using Markov Chain Monte Carlo simulation. It can be used even in problems with posterior distributions that are difficult to sample from using standard random number generators.

Hypothesis Tests

Hypothesis Test Terminology

All hypothesis tests share the same basic terminology and structure.

- A *null hypothesis* is an assertion about a population that you would like to test. It is "null" in the sense that it often represents a status quo belief, such as the absence of a characteristic or the lack of an effect. It may be formalized by asserting that a population parameter, or a combination of population parameters, has a certain value. In the example given in the "Hypothesis Testing" on page 8-5, the null hypothesis would be that the average price of gas across the state was $1.15. This is written H_0: $\mu = 1.15$.
- An *alternative hypothesis* is a contrasting assertion about the population that can be tested against the null hypothesis. In the example given in the "Hypothesis Testing" on page 8-5, possible alternative hypotheses are:

  H_1: $\mu \neq 1.15$ — State average was different from $1.15 (two-tailed test)

  H_1: $\mu > 1.15$ — State average was greater than $1.15 (right-tail test)

  H_1: $\mu < 1.15$ — State average was less than $1.15 (left-tail test)
- To conduct a hypothesis test, a random sample from the population is collected and a relevant *test statistic* is computed to summarize the sample. This statistic varies with the type of test, but its distribution under the null hypothesis must be known (or assumed).
- The *p value* of a test is the probability, under the null hypothesis, of obtaining a value of the test statistic as extreme or more extreme than the value computed from the sample.
- The *significance level* of a test is a threshold of probability α agreed to before the test is conducted. A typical value of α is 0.05. If the p value of a test is less than α, the test rejects the null hypothesis. If the p value is greater than α, there is insufficient evidence to reject the null hypothesis. Note that lack of evidence for rejecting the null hypothesis is not evidence for accepting the null hypothesis. Also note that substantive "significance" of an alternative cannot be inferred from the statistical significance of a test.
- The significance level α can be interpreted as the probability of rejecting the null hypothesis when it is actually true—a *type I error*. The distribution of the test statistic under the null hypothesis determines the probability α of a type I error. Even if the null hypothesis is not rejected, it may still be false—a *type II error*. The distribution of the test statistic under the alternative hypothesis determines the probability β of a type II error. Type II errors are often due to small sample sizes. The *power* of a test, $1 - \beta$, is the probability of correctly rejecting a false null hypothesis.
- Results of hypothesis tests are often communicated with a *confidence interval*. A confidence interval is an estimated range of values with a specified probability of containing the true population value of a parameter. Upper and lower bounds for confidence intervals are computed from the sample estimate of the parameter and the known (or assumed) sampling distribution of the estimator. A typical assumption is that estimates will be normally distributed with repeated sampling (as dictated by the Central Limit Theorem). Wider confidence intervals correspond to poor estimates (smaller samples); narrow intervals correspond to better estimates (larger samples). If the null hypothesis asserts the value of a population parameter, the test rejects the null hypothesis when the hypothesized value lies outside the computed confidence interval for the parameter.

See Also

More About

- “Hypothesis Testing” on page 8-5
- “Hypothesis Test Assumptions” on page 8-4
- “Available Hypothesis Tests” on page 8-10

Hypothesis Test Assumptions

Different hypothesis tests make different assumptions about the distribution of the random variable being sampled in the data. These assumptions must be considered when choosing a test and when interpreting the results.

For example, the *z*-test (`ztest`) and the *t*-test (`ttest`) both assume that the data are independently sampled from a normal distribution. Statistics and Machine Learning Toolbox functions are available for testing this assumption, such as `chi2gof`, `jbtest`, `lillietest`, and `normplot`.

Both the *z*-test and the *t*-test are relatively robust with respect to departures from this assumption, so long as the sample size n is large enough. Both tests compute a sample mean $\bar{x}$, which, by the Central Limit Theorem, has an approximately normal sampling distribution with mean equal to the population mean μ, regardless of the population distribution being sampled.

The difference between the *z*-test and the *t*-test is in the assumption of the standard deviation σ of the underlying normal distribution. A *z*-test assumes that σ is known; a *t*-test does not. As a result, a *t*-test must compute an estimate s of the standard deviation from the sample.

Test statistics for the *z*-test and the *t*-test are, respectively,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Under the null hypothesis that the population is distributed with mean μ, the *z*-statistic has a standard normal distribution, $N(0,1)$. Under the same null hypothesis, the *t*-statistic has Student's *t* distribution with $n - 1$ degrees of freedom. For small sample sizes, Student's *t* distribution is flatter and wider than $N(0,1)$, compensating for the decreased confidence in the estimate s. As sample size increases, however, Student's *t* distribution approaches the standard normal distribution, and the two tests become essentially equivalent.

Knowing the distribution of the test statistic under the null hypothesis allows for accurate calculation of *p*-values. Interpreting *p*-values in the context of the test assumptions allows for critical analysis of test results.

Assumptions underlying Statistics and Machine Learning Toolbox hypothesis tests are given in the reference pages for implementing functions.

See Also

More About

- “Hypothesis Testing” on page 8-5
- “Hypothesis Test Terminology” on page 8-2
- “Available Hypothesis Tests” on page 8-10

Hypothesis Testing

Hypothesis testing is a common method of drawing inferences about a population based on statistical evidence from a sample.

As an example, suppose someone says that at a certain time in the state of Massachusetts the average price of a gallon of regular unleaded gas was $1.15. How could you determine the truth of the statement? You could try to find prices at every gas station in the state at the time. That approach would be definitive, but it could be time-consuming, costly, or even impossible.

A simpler approach would be to find prices at a small number of randomly selected gas stations around the state, and then compute the sample average.

Sample averages differ from one another due to chance variability in the selection process. Suppose your sample average comes out to be $1.18. Is the $0.03 difference an artifact of random sampling or significant evidence that the average price of a gallon of gas was in fact greater than $1.15? Hypothesis testing is a statistical method for making such decisions.

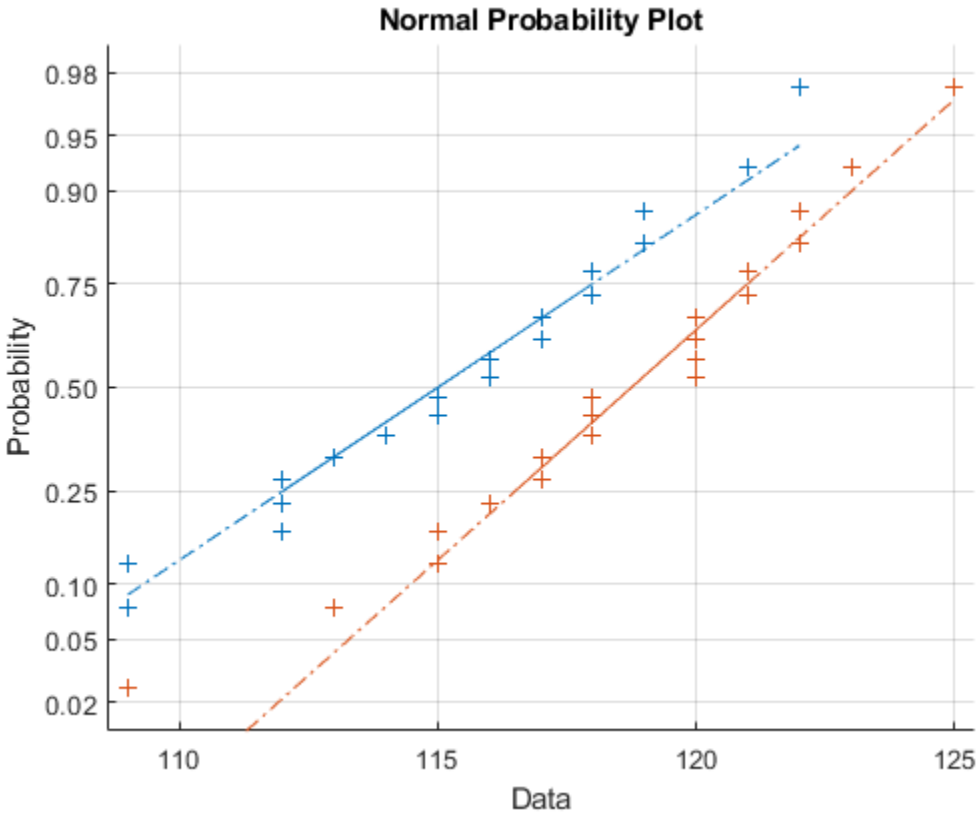
This example shows how to use hypothesis testing to analyze gas prices measured across the state of Massachusetts during two separate months.

This example uses the gas price data in the file `gas.mat`. The file contains two random samples of prices for a gallon of gas around the state of Massachusetts in 1993. The first sample, `price1`, contains 20 random observations around the state on a single day in January. The second sample, `price2`, contains 20 random observations around the state one month later.

```
load gas
prices = [price1 price2];
```

As a first step, you might want to test the assumption that the samples come from normal distributions. A normal probability plot gives a quick idea.

```
normplot(prices)
```



Both scatters approximately follow straight lines through the first and third quartiles of the samples, indicating approximate normal distributions. The February sample (the right-hand line) shows a slight departure from normality in the lower tail. A shift in the mean from January to February is evident. A hypothesis test is used to quantify the test of normality. Since each sample is relatively small, a Lilliefors test is recommended.

```
lillietest(price1)
```

```
ans = 0
```

```
lillietest(price2)
```

```
ans = 0
```

The default significance level of `lillietest` is 5%. The logical 0 returned by each test indicates a failure to reject the null hypothesis that the samples are normally distributed. This failure may reflect normality in the population or it may reflect a lack of strong evidence against the null hypothesis due to the small sample size.

Now compute the sample means.

```
sample_means = mean(prices)
```

```
sample_means = 1×2

  115.1500  118.5000
```

You might want to test the null hypothesis that the mean price across the state on the day of the January sample was $1.15. If you know that the standard deviation in prices across the state has historically, and consistently, been $0.04, then a *z*-test is appropriate.

```
[h,pvalue,ci] = ztest(price1/100,1.15,0.04)
```

```
h = 0

pvalue = 0.8668

ci = 2×1

    1.1340
    1.1690
```

The logical output `h = 0` indicates a failure to reject the null hypothesis at the default significance level of 5%. This is a consequence of the high probability under the null hypothesis, indicated by the *p* value, of observing a value as extreme or more extreme of the *z*-statistic computed from the sample. The 95% confidence interval on the mean [1.1340 1.1690] includes the hypothesized population mean of $1.15.

Does the later sample offer stronger evidence for rejecting a null hypothesis of a state-wide average price of $1.15 in February? The shift shown in the probability plot and the difference in the computed sample means suggest this. The shift might indicate a significant fluctuation in the market, raising questions about the validity of using the historical standard deviation. If a known standard deviation cannot be assumed, a *t*-test is more appropriate.

```
[h,pvalue,ci] = ttest(price2/100,1.15)
```

```
h = 1

pvalue = 4.9517e-04

ci = 2×1

    1.1675
    1.2025
```

The logical output `h = 1` indicates a rejection of the null hypothesis at the default significance level of 5%. In this case, the 95% confidence interval on the mean does not include the hypothesized population mean of $1.15.

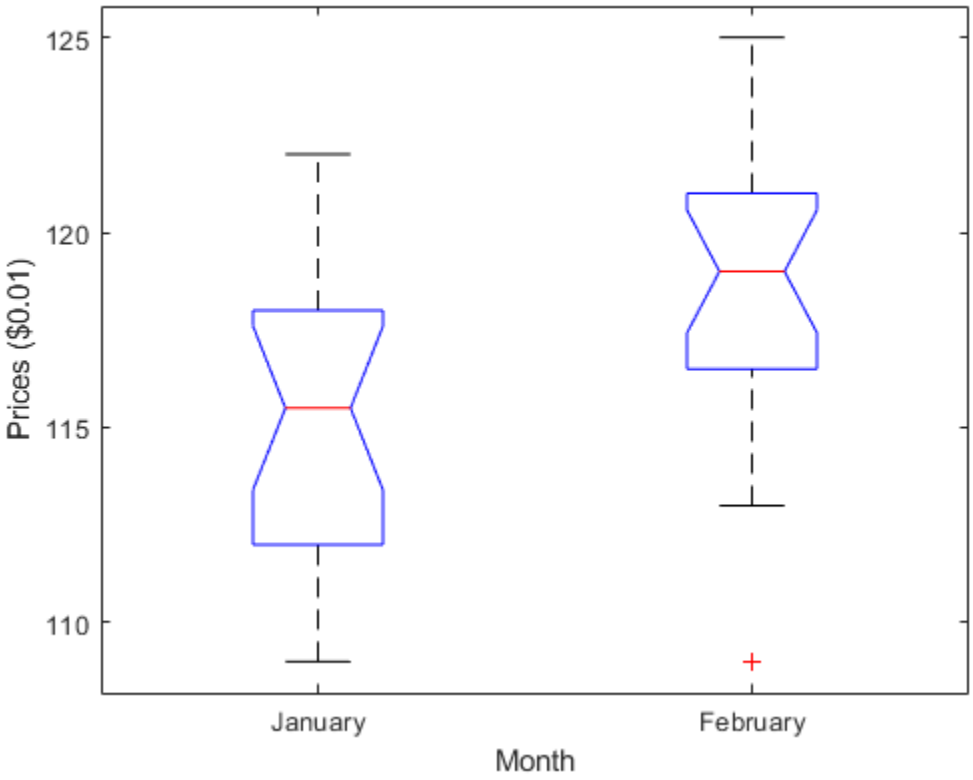
You might want to investigate the shift in prices a little more closely. The function `ttest2` tests if two independent samples come from normal distributions with equal but unknown standard deviations and the same mean, against the alternative that the means are unequal.

```
[h,sig,ci] = ttest2(price1,price2)
```

```
h = 1

sig = 0.0083

ci = 2×1

   -5.7845
   -0.9155
```

The null hypothesis is rejected at the default 5% significance level, and the confidence interval on the difference of means does not include the hypothesized value of 0. A notched box plot is another way to visualize the shift.

```
boxplot(prices,1)
h = gca;
h.XTick = [1 2];
h.XTickLabel = {'January','February'};
xlabel('Month')
ylabel('Prices ($0.01)')
```



The plot displays the distribution of the samples around their medians. The heights of the notches in each box are computed so that the side-by-side boxes have nonoverlapping notches when their medians are different at a default 5% significance level. The computation is based on an assumption of normality in the data, but the comparison is reasonably robust for other distributions. The side-by-side plots provide a kind of visual hypothesis test, comparing medians rather than means. The plot above appears to barely reject the null hypothesis of equal medians.

The nonparametric Wilcoxon rank sum test, implemented by the function `ranksum`, can be used to quantify the test of equal medians. It tests if two independent samples come from identical continuous (not necessarily normal) distributions with equal medians, against the alternative that they do not have equal medians.

```
[p,h] = ranksum(price1,price2)
```

```
p = 0.0095
```

```
h = logical
   1
```

The test rejects the null hypothesis of equal medians at the default 5% significance level.

See Also

`boxplot` | `lillietest` | `normplot` | `ttest` | `ttest2` | `ztest`

More About

- “Hypothesis Test Terminology” on page 8-2
- “Hypothesis Test Assumptions” on page 8-4
- “Available Hypothesis Tests” on page 8-10
- “Distribution Plots” on page 4-7

Available Hypothesis Tests

Function	Description
`ansaribradley`	Ansari-Bradley test. Tests if two independent samples come from the same distribution, against the alternative that they come from distributions that have the same median and shape but different variances.
`barttest`	Bartlett's test. Tests if the variances of the data values along each principal component are equal, against the alternative that the variances are not all equal.
`chi2gof`	Chi-square goodness-of-fit test. Tests if a sample comes from a specified distribution, against the alternative that it does not come from that distribution.
`dwtest`	Durbin-Watson test. Tests if the residuals from a linear regression are uncorrelated, against the alternative that there is autocorrelation among them.
`friedman`	Friedman's test. Tests if the column effects in a two-way layout are all the same, against the alternative that the column effects are not all the same.
`jbtest`	Jarque-Bera test. Tests if a sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution.
`kruskalwallis`	Kruskal-Wallis test. Tests if multiple samples are all drawn from the same populations (or equivalently, from different populations with the same distribution), against the alternative that they are not all drawn from the same population.
`kstest`	One-sample Kolmogorov-Smirnov test. Tests if a sample comes from a continuous distribution with specified parameters, against the alternative that it does not come from that distribution.
`kstest2`	Two-sample Kolmogorov-Smirnov test. Tests if two samples come from the same continuous distribution, against the alternative that they do not come from the same distribution.
`lillietest`	Lilliefors test. Tests if a sample comes from a distribution in the normal family, against the alternative that it does not come from a normal distribution.
`linhyptest`	Linear hypothesis test. Tests if `H*b = c` for parameter estimates `b` with estimated covariance `H` and specified `c`, against the alternative that `H*b ≠ c`.
`ranksum`	Wilcoxon rank sum test. Tests if two independent samples come from identical continuous distributions with equal medians, against the alternative that they do not have equal medians.
`runstest`	Runs test. Tests if a sequence of values comes in random order, against the alternative that the ordering is not random.
`signrank`	One-sample or paired-sample Wilcoxon signed rank test. Tests if a sample comes from a continuous distribution symmetric about a specified median, against the alternative that it does not have that median.

Function	Description
`signtest`	One-sample or paired-sample sign test. Tests if a sample comes from an arbitrary continuous distribution with a specified median, against the alternative that it does not have that median.
`ttest`	One-sample or paired-sample *t*-test. Tests if a sample comes from a normal distribution with unknown variance and a specified mean, against the alternative that it does not have that mean.
`ttest2`	Two-sample *t*-test. Tests if two independent samples come from normal distributions with unknown but equal (or, optionally, unequal) variances and the same mean, against the alternative that the means are unequal.
`vartest`	One-sample chi-square variance test. Tests if a sample comes from a normal distribution with specified variance, against the alternative that it comes from a normal distribution with a different variance.
`vartest2`	Two-sample *F*-test for equal variances. Tests if two independent samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.
`vartestn`	Bartlett multiple-sample test for equal variances. Tests if multiple samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.
`ztest`	One-sample *z*-test. Tests if a sample comes from a normal distribution with known variance and specified mean, against the alternative that it does not have that mean.

Note In addition to the previous functions, Statistics and Machine Learning Toolbox functions are available for analysis of variance (ANOVA), which perform hypothesis tests in the context of linear modeling.

See Also

More About

- “Hypothesis Testing” on page 8-5
- “Hypothesis Test Terminology” on page 8-2
- “Hypothesis Test Assumptions” on page 8-4

Selecting a Sample Size

This example shows how to determine the number of samples or observations needed to carry out a statistical test. It illustrates sample size calculations for a simple problem, then shows how to use the `sampsizepwr` function to compute power and sample size for two more realistic problems. Finally, it illustrates the use of Statistics and Machine Learning Toolbox™ functions to compute the required sample size for a test that the `sampsizepwr` function does not support.

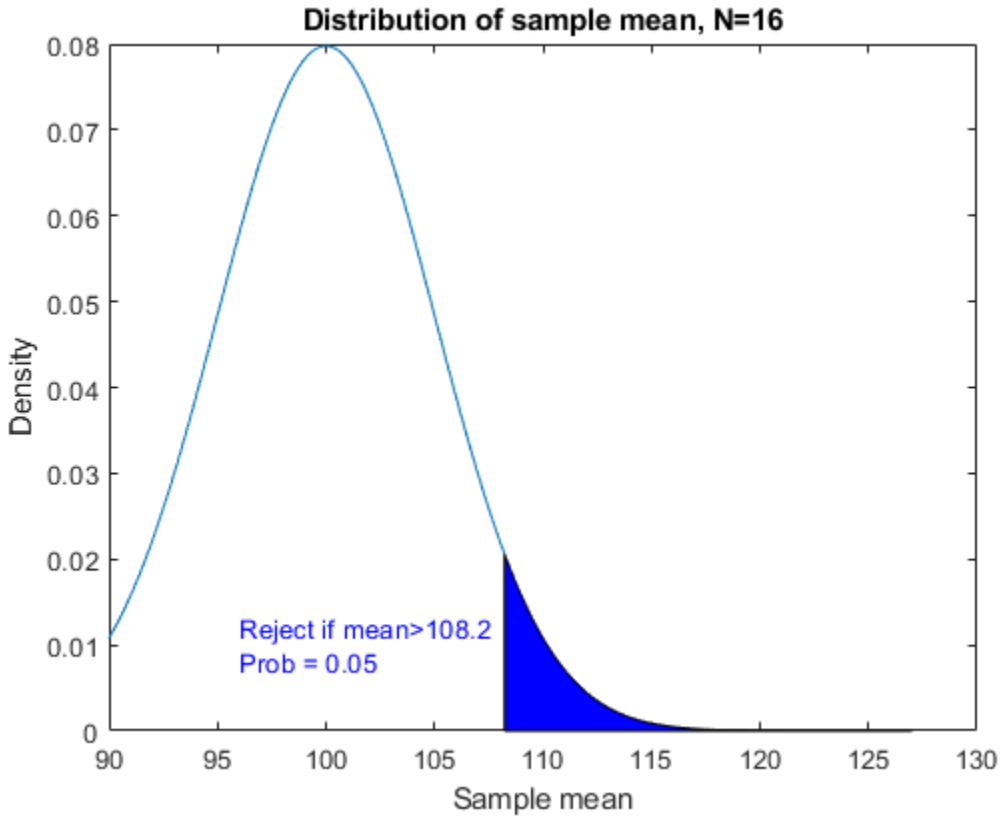
Testing a Normal Mean with Known Standard Deviation, One-Sided

Just to introduce some concepts, let's consider an unrealistically simple example where we want to test a mean and we know the standard deviation. Our data are continuous, and can be modeled with the normal distribution. We need to determine a sample size N so that we can distinguish between a mean of 100 and a mean of 110. We know the standard deviation is 20.

When we carry out a statistical test, we generally test a **null hypothesis** against an **alternative hypothesis**. We find a test statistic T, and look at its distribution under the null hypothesis. If we observe an unusual value, say one that has less than 5% chance of happening if the null hypothesis is true, then we reject the null hypothesis in favor of the alternative. (The 5% probability is known as the **significance level** of the test.) If the value is not unusual, we do not reject the null hypothesis.

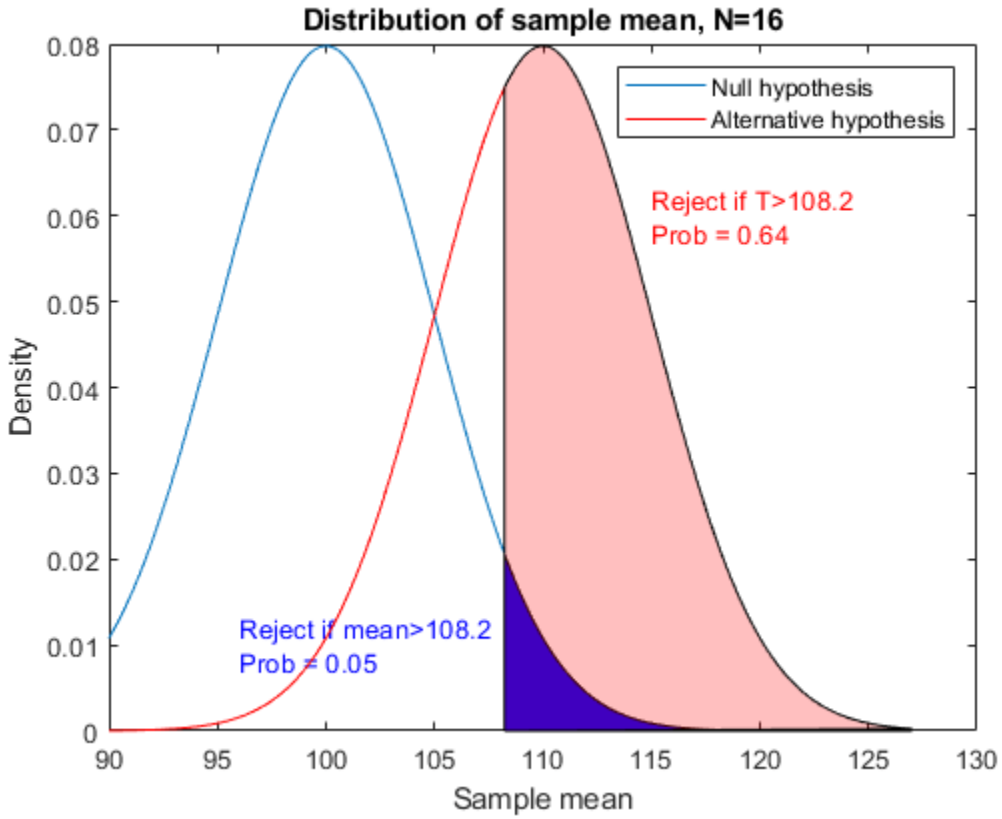
In this case the test statistic T is the sample mean. Under the null hypothesis it has a mean of 100 and a standard deviation of 20/sqrt(N). First let's look at a fixed sample size, say N=16. We will reject the null hypothesis if T is in the shaded region, which is the upper tail of its distribution. That makes this a one-sided test, since we will not reject if T is in the lower tail. The cutoff for this shaded region is 1.6 standard deviations above the mean.

```
rng(0,'twister');
mu0 = 100;
sig = 20;
N = 16;
alpha = 0.05;
conf = 1-alpha;
cutoff = norminv(conf, mu0, sig/sqrt(N));
x = [linspace(90,cutoff), linspace(cutoff,127)];
y = normpdf(x,mu0,sig/sqrt(N));
h1 = plot(x,y);
xhi = [cutoff, x(x>=cutoff)];
yhi = [0, y(x>=cutoff)];
patch(xhi,yhi,'b');
title('Distribution of sample mean, N=16');
xlabel('Sample mean');
ylabel('Density');
text(96,.01,sprintf('Reject if mean>%.4g\nProb = 0.05',cutoff),'Color','b');
```



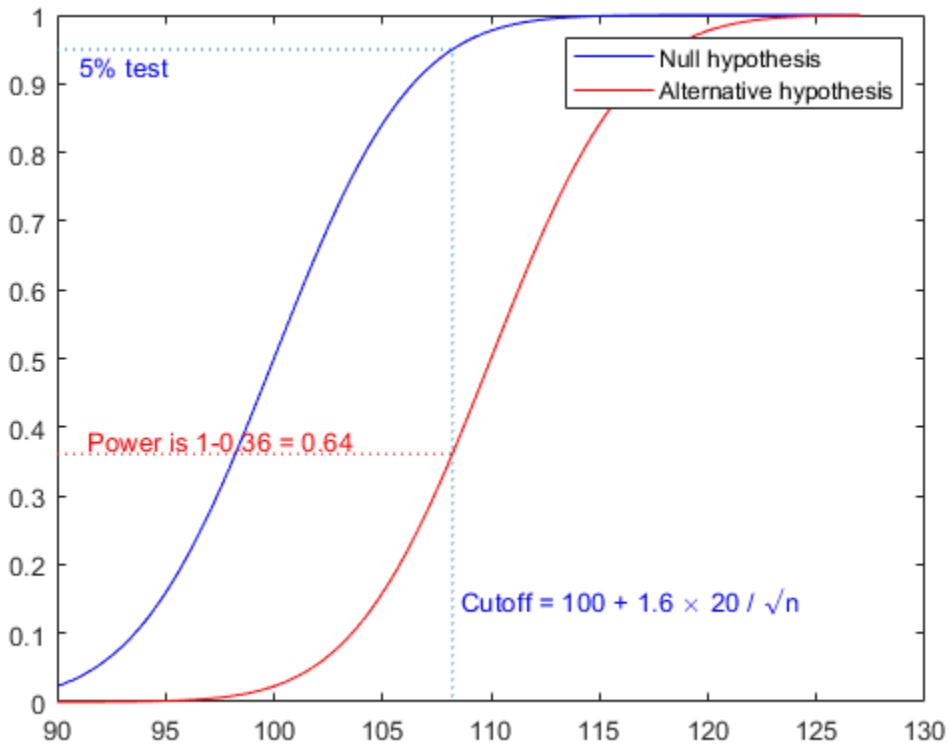
This is how T behaves under the null hypothesis, but what about under an alternative? Our alternative distribution has a mean of 110, as represented by the red curve.

```
mu1 = 110;
y2 = normpdf(x,mu1,sig/sqrt(N));
h2 = line(x,y2,'Color','r');
yhi = [0, y2(x>=cutoff)];
patch(xhi,yhi,'r','FaceAlpha',0.25);
P = 1 - normcdf(cutoff,mu1,sig/sqrt(N));
text(115,.06,sprintf('Reject if T>%.4g\nProb = %.2g',cutoff,P),'Color',[1 0 0]);
legend([h1 h2],'Null hypothesis','Alternative hypothesis');
```



There is a larger chance of rejecting the null hypothesis if the alternative is true. This is just what we want. It's easier to visualize if we look at the cumulative distribution function (cdf) instead of the density (pdf). We can read probabilities directly from this graph, instead of having to compute areas.

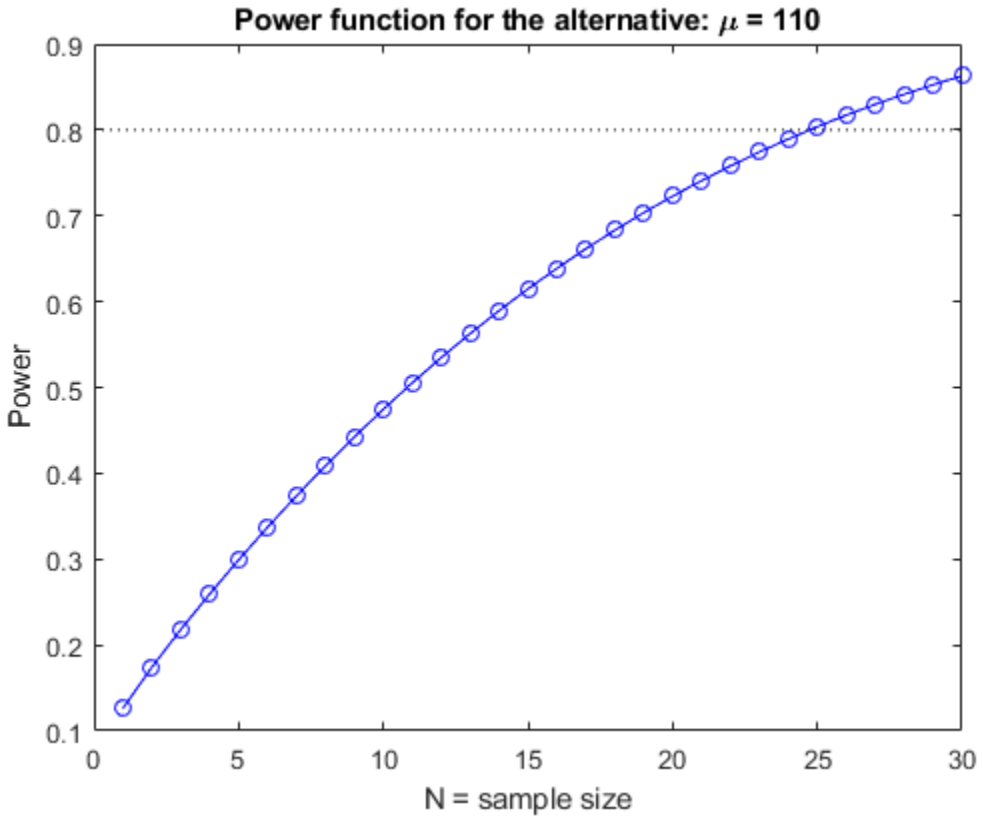
```
ynull = normcdf(x,mu0,sig/sqrt(N));
yalt = normcdf(x,mu1,sig/sqrt(N));
h12 = plot(x,ynull,'b-',x,yalt,'r-');
zval = norminv(conf);
cutoff = mu0 + zval * sig / sqrt(N);
line([90,cutoff,cutoff],[conf, conf, 0],'LineStyle',':');
msg = sprintf(' Cutoff = 100 + %.2g \\times 20 / \\surd{n}',zval);
text(cutoff,.15,msg,'Color','b');
text(min(x),conf,sprintf('   %g%% test',100*alpha),'Color','b',...
                        'verticalalignment','top')
palt = normcdf(cutoff,mu1,sig/sqrt(N));
line([90,cutoff],[palt,palt],'Color','r','LineStyle',':');
text(91,palt+.02,sprintf(' Power is 1-%.2g = %.2g',palt,1-palt),'Color',[1 0 0]);
legend(h12,'Null hypothesis','Alternative hypothesis');
```



This graph shows the probability of getting a significant statistic (rejecting the null hypothesis) for two different mu values when N=16.

The **power function** is defined as the probability of rejecting the null hypothesis when the alternative is true. It depends on the value of the alternative and on the sample size. We'll graph the power (which is one minus the cdf) as a function of N, fixing the alternative at 110. We'll select N to achieve a power of 80%. The graph shows that we need about N=25.

```
DesiredPower = 0.80;
Nvec = 1:30;
cutoff = mu0 + norminv(conf)*sig./sqrt(Nvec);
power = 1 - normcdf(cutoff, mu1, sig./sqrt(Nvec));
plot(Nvec,power,'bo-',[0 30],[DesiredPower DesiredPower],'k:');
xlabel('N = sample size')
ylabel('Power')
title('Power function for the alternative: \mu = 110')
```



In this overly simple example there is a formula to compute the required value directly to get a power of 80%:

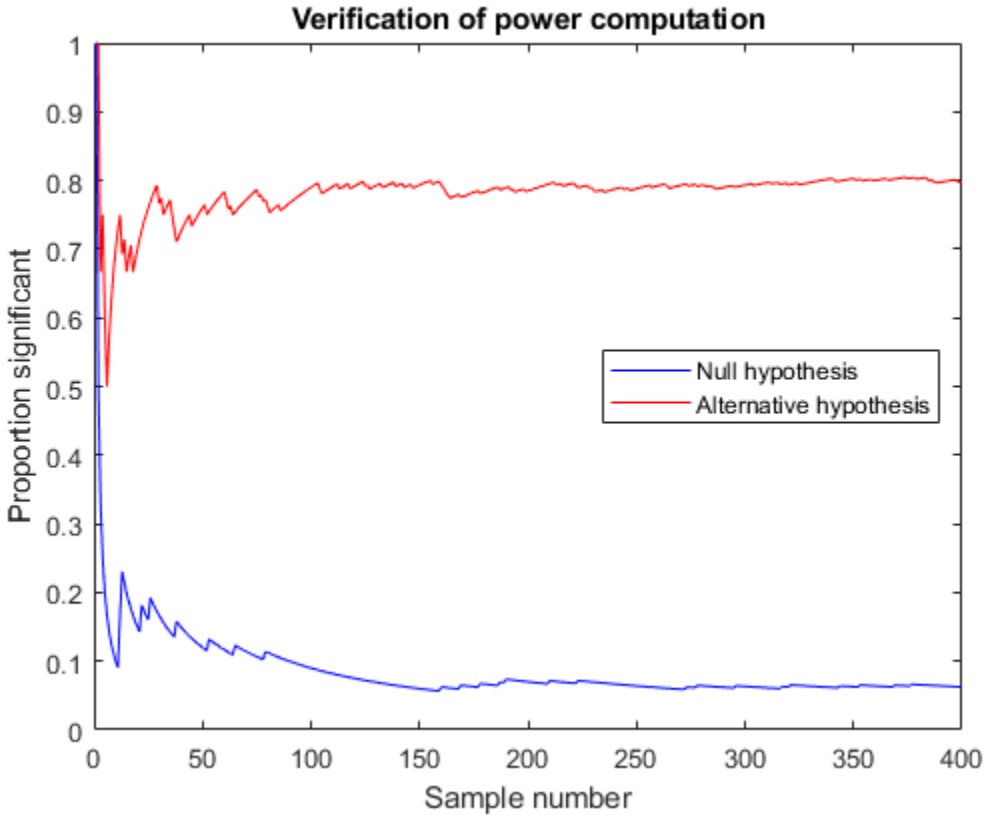
```
mudiff = (mu1 - mu0) / sig;
N80 = ceil(((norminv(1-DesiredPower)-norminv(conf)) / mudiff)^2)
```

```
N80 =

    25
```

To verify that this works, let's do a Monte Carlo simulation and generate 400 samples of size 25 both under the null hypothesis with mean 100, and under the alternative hypothesis with mean 110. If we test each sample to see if its mean is 100, we should expect about 5% of the first group and 80% of the second group to be significant.

```
nsamples = 400;
samplenum = 1:nsamples;
N = 25;
h0 = zeros(1,nsamples);
h1 = h0;
for j = 1:nsamples
    Z0 = normrnd(mu0,sig,N,1);
    h0(j) = ztest(Z0,mu0,sig,alpha,'right');
    Z1 = normrnd(mu1,sig,N,1);
    h1(j) = ztest(Z1,mu0,sig,alpha,'right');
```

```
end
p0 = cumsum(h0) ./ samplenum;
p1 = cumsum(h1) ./ samplenum;
plot(samplenum,p0,'b-',samplenum,p1,'r-')
xlabel('Sample number')
ylabel('Proportion significant')
title('Verification of power computation')
legend('Null hypothesis','Alternative hypothesis','Location','East')
```



Testing a Normal Mean with Unknown Standard Deviation, Two-Sided

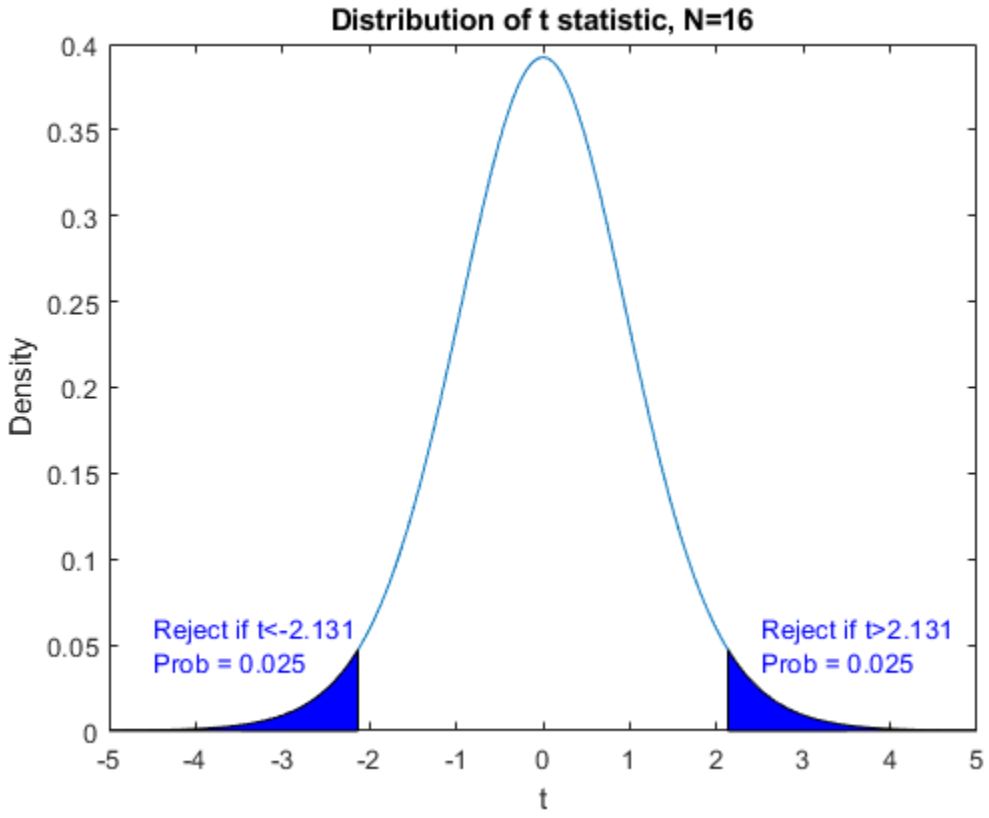
Now let's suppose we don't know the standard deviation, and we want to perform a two-sided test, that is, one that rejects the null hypothesis whether the sample mean is too high or too low.

The test statistic is a t statistic, which is the difference between the sample mean and the mean being tested, divided by the standard error of the mean. Under the null hypothesis, this has Student's t distribution with N-1 degrees of freedom. Under the alternative hypothesis, the distribution is a noncentral t distribution with a noncentrality parameter equal to the normalized difference between the true mean and the mean being tested.

For this two-sided test we have to allocate the 5% chance of an error under the null hypothesis equally to both tails, and reject if the test statistic is too extreme in either direction. We also have to consider both tails under any alternative.

```
N = 16;
df = N-1;
alpha = 0.05;
```

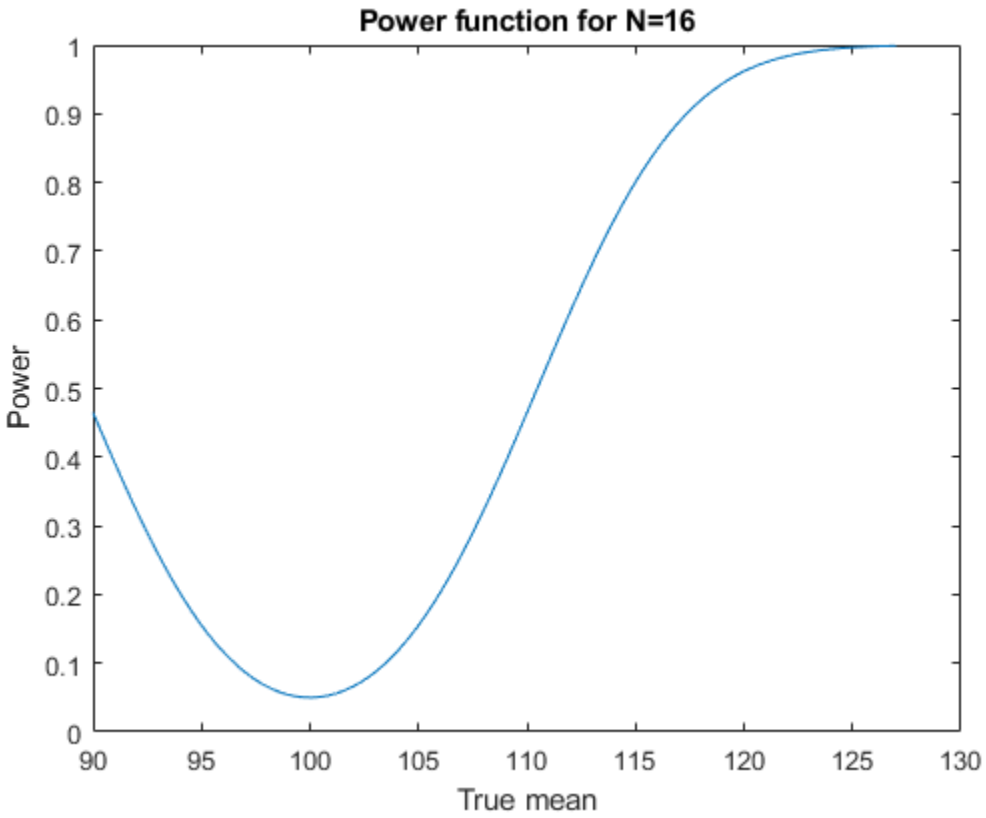
```
conf = 1-alpha;
cutoff1 = tinv(alpha/2,df);
cutoff2 = tinv(1-alpha/2,df);
x = [linspace(-5,cutoff1), linspace(cutoff1,cutoff2),linspace(cutoff2,5)];
y = tpdf(x,df);
h1 = plot(x,y);
xlo = [x(x<=cutoff1),cutoff1];
ylo = [y(x<=cutoff1),0];
xhi = [cutoff2,x(x>=cutoff2)];
yhi = [0, y(x>=cutoff2)];
patch(xlo,ylo,'b');
patch(xhi,yhi,'b');
title('Distribution of t statistic, N=16');
xlabel('t');
ylabel('Density');
text(2.5,.05,sprintf('Reject if t>%.4g\nProb = 0.025',cutoff2),'Color','b');
text(-4.5,.05,sprintf('Reject if t<%.4g\nProb = 0.025',cutoff1),'Color','b');
```



Instead of examining the power function just under the null hypothesis and a single alternative value of mu, we can look at it as a function of mu. The power increases as mu moves away from the null hypothesis value in either direction. We can use the `sampsizepwr` function to compute the power. For a power calculation we need to specify a value for the standard deviation, which we suspect will be roughly 20. Here's a picture of the power function for a sample size N=16.

```
N = 16;
x = linspace(90,127);
power = sampsizepwr('t',[100 20],x,[],N);
```

```
plot(x,power);
xlabel('True mean')
ylabel('Power')
title('Power function for N=16')
```



We want a power of 80% when the mean is 110. According to this graph, our power is less than 50% with a sample size of N=16. What sample size will give the power we want?

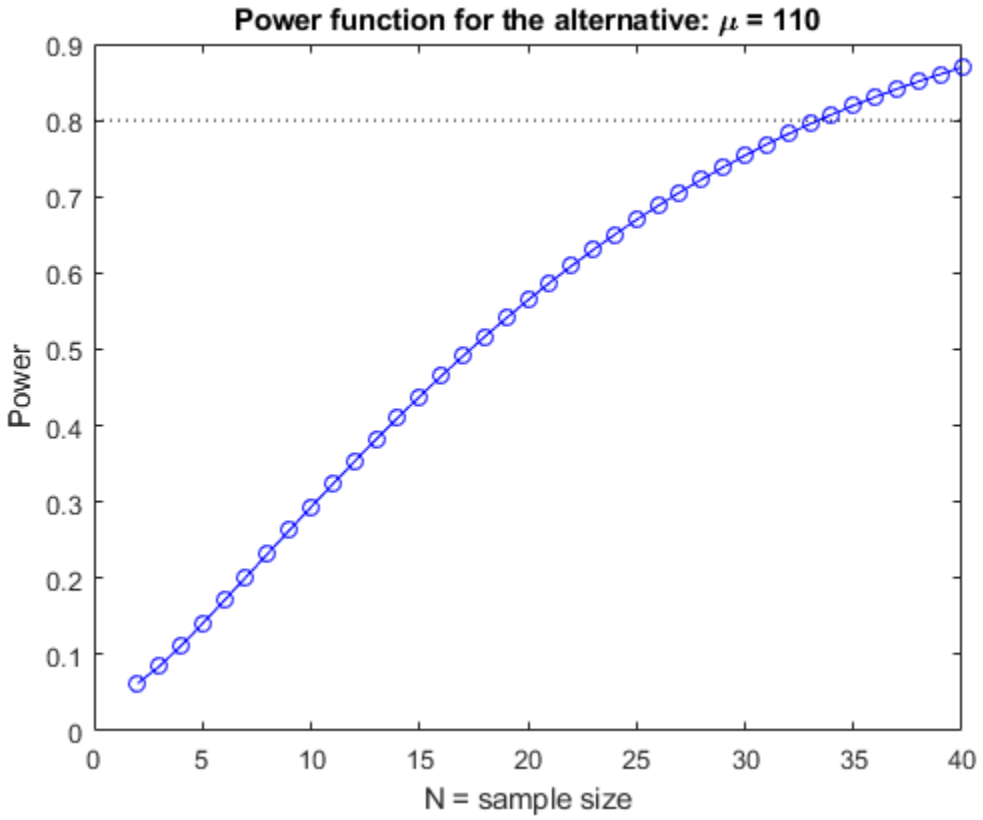
```
N = sampsizepwr('t',[100 20],110,0.8)
```

```
N =

    34
```

We need a sample size of about 34. Compared with the previous example, we need to take nine additional observations to allow a two-sided test and to make up for not knowing the true standard deviation.

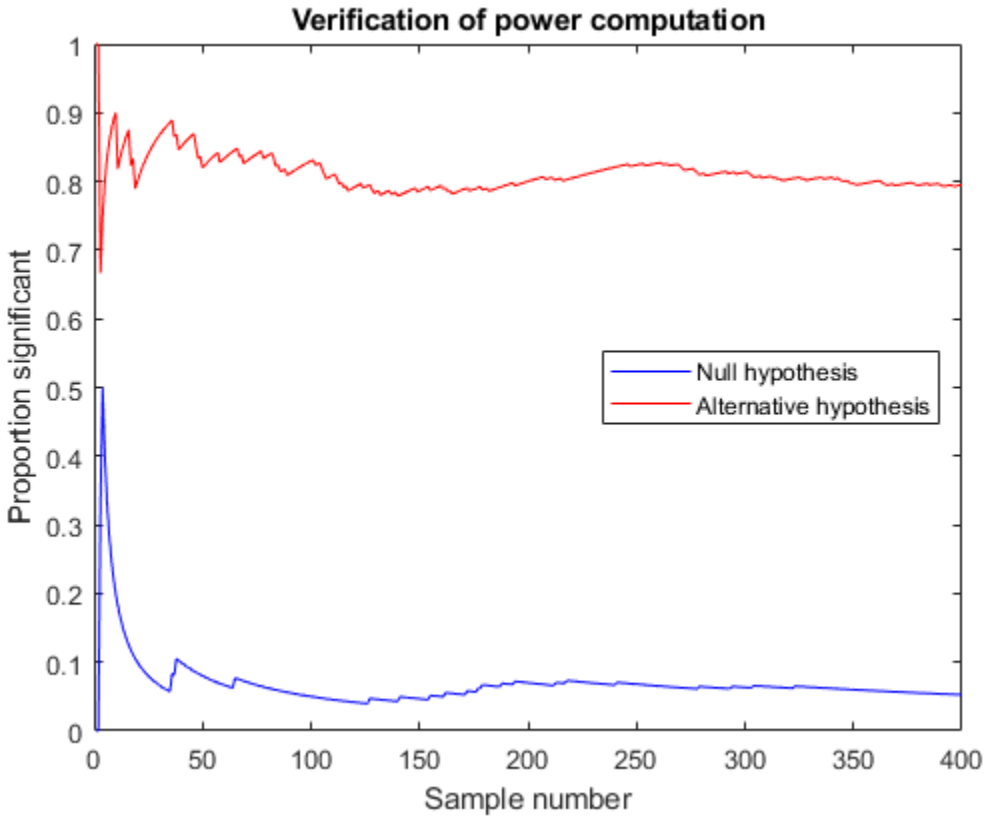
We can make a plot of the power function for various values of N.

```
Nvec = 2:40;
power = sampsizepwr('t',[100 20],110,[],Nvec);
plot(Nvec,power,'bo-',[0 40],[DesiredPower DesiredPower],'k:');
xlabel('N = sample size')
ylabel('Power')
title('Power function for the alternative: \mu = 110')
```



And we can do a simulation similar to the earlier one to verify that we get the power we need.

```
nsamples = 400;
samplenum = 1:nsamples;
N = 34;
h0 = zeros(1,nsamples);
h1 = h0;
for j = 1:nsamples
    Z0 = normrnd(mu0,sig,N,1);
    h0(j) = ttest(Z0,mu0,alpha);
    Z1 = normrnd(mu1,sig,N,1);
    h1(j) = ttest(Z1,mu0,alpha);
end
p0 = cumsum(h0) ./ samplenum;
p1 = cumsum(h1) ./ samplenum;
plot(samplenum,p0,'b-',samplenum,p1,'r-')
xlabel('Sample number')
ylabel('Proportion significant')
title('Verification of power computation')
legend('Null hypothesis','Alternative hypothesis','Location','East')
```



Suppose we could afford to take a sample size of 50. Presumably our power for detecting the alternative value mu=110 would be larger than 80%. If we maintain the power at 80%, what alternative could we detect?

```
mu1 = sampsizepwr('t',[100 20],[],.8,50)
```

```
mu1 =

  108.0837
```

Testing a Proportion

Now let's turn to the problem of determining the sample size needed to distinguish between two proportions. Suppose that we are sampling a population in which about 30% favor some candidate, and we want to sample enough people so we can distinguish this value from 33%.

The idea here is the same as before. Here we can use the sample count as our test statistic. This count has a binomial distribution. For any sample size N we can compute the cutoff for rejecting the null hypothesis P=0.30. For N=100, for instance, we would reject the null hypothesis if the sample count is larger than a cutoff value computed as follows:

```
N = 100;
alpha = 0.05;
p0 = 0.30;
p1 = 0.33;
cutoff = binoinv(1-alpha, N, p0)
```

```
cutoff =

    38
```

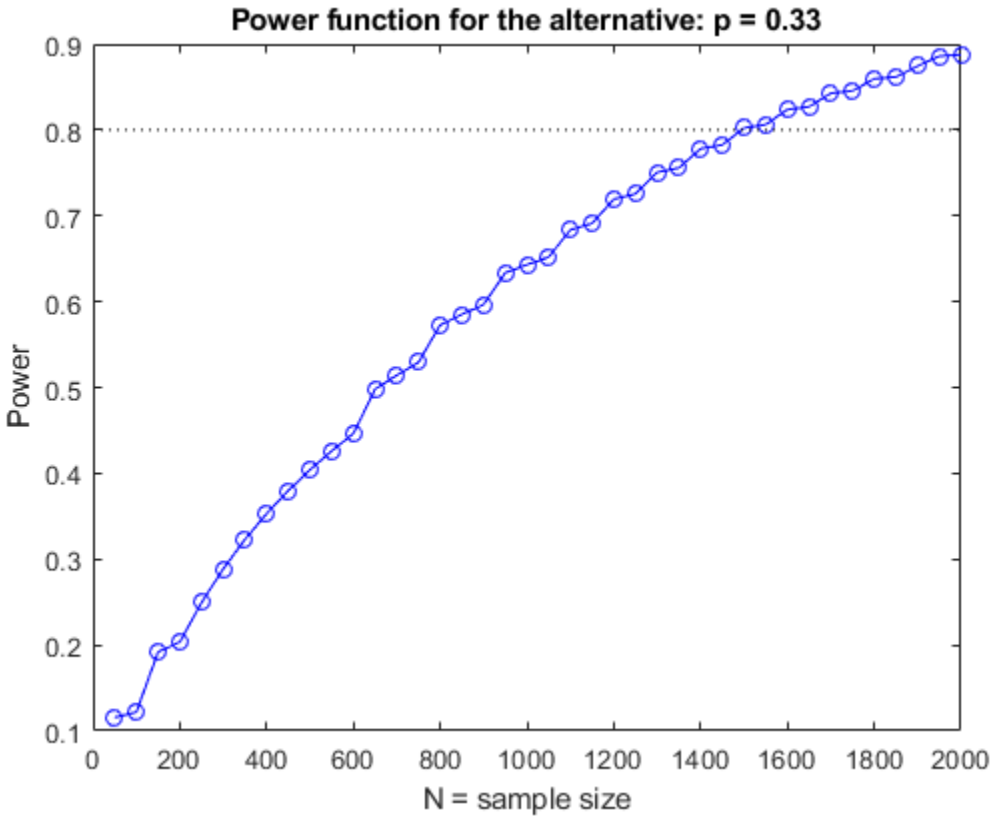
A complication with the binomial distribution comes because it is discrete. The probability of exceeding the cutoff value is not exactly 5%:

```
1 - binocdf(cutoff, N, p0)
```

```
ans =

    0.0340
```

Once again, let's compute the power against the alternative P=0.33 for a range of sample sizes. We'll use a one-sided (right-tailed) test because we're interested only in alternative values greater than 30%.

```
Nvec = 50:50:2000;
power = sampsizepwr('p',p0,p1,[],Nvec,'tail','right');
plot(Nvec,power,'bo-',[0 2000],[DesiredPower DesiredPower],'k:');
xlabel('N = sample size')
ylabel('Power')
title('Power function for the alternative: p = 0.33')
```



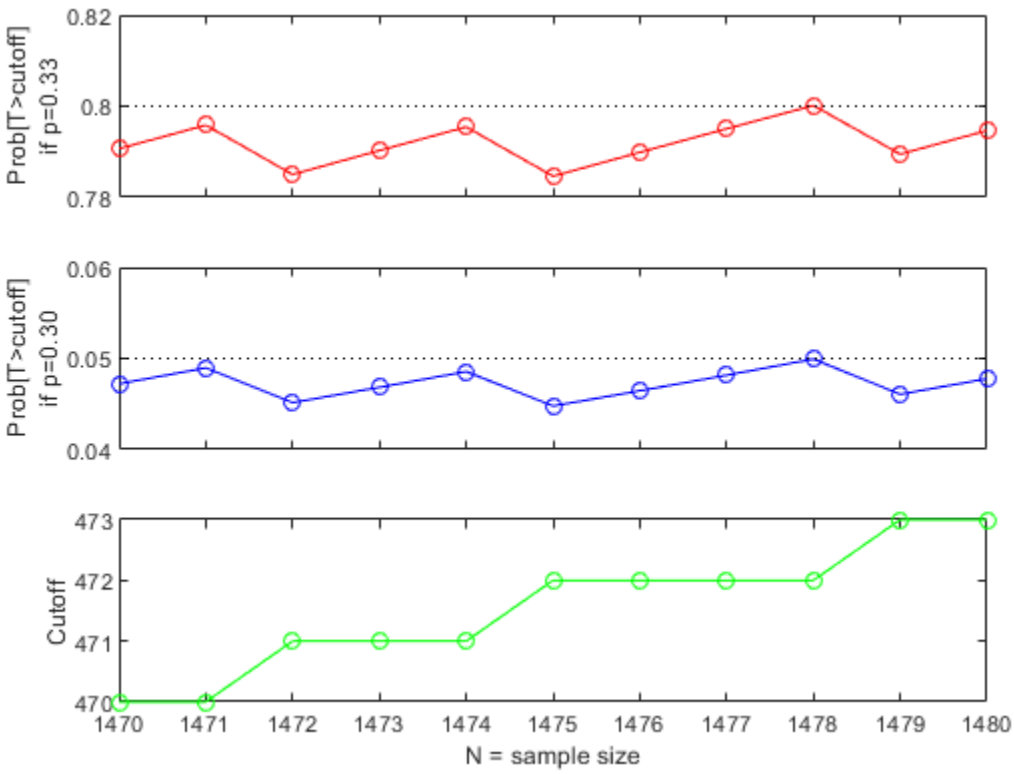
We can use the `sampsizepwr` function to request the sample size required for a power of 80%.

```
approxN = sampsizepwr('p',p0,p1,0.80,[],'tail','right')
```

```
Warning: Values N>200 are approximate.  Plotting the power as a function
of N may reveal lower N values that have the required power.

approxN =

        1500
```

A warning message informs us that the answer is just approximate. If we look at the power function for different sample sizes, we can see that the function is generally increasing, but irregular because the binomial distribution is discrete. Let's look at the probability of rejecting the null hypothesis for both p=0.30 and p=0.33 in the range of samples sizes from 1470 to 1480.

```
subplot(3,1,1);
Nvec = 1470:1480;
power = sampsizepwr('p',p0,p1,[],Nvec,'tail','right');
plot(Nvec,power,'ro-',[min(Nvec),max(Nvec)],[DesiredPower DesiredPower],'k:');
ylabel(sprintf('Prob[T>cutoff]\nif p=0.33'))
h_gca = gca;
h_gca.XTickLabel = '';
ylim([.78, .82]);

subplot(3,1,2);
alf = sampsizepwr('p',p0,p0,[],Nvec,'tail','right');
plot(Nvec,alf,'bo-',[min(Nvec),max(Nvec)],[alpha alpha],'k:');
ylabel(sprintf('Prob[T>cutoff]\nif p=0.30'))
h_gca = gca;
h_gca.XTickLabel = '';
ylim([.04, .06]);

subplot(3,1,3);
cutoff = binoinv(1-alpha, Nvec, p0);
plot(Nvec,cutoff,'go-');
xlabel('N = sample size')
ylabel('Cutoff')
```



This plot reveals that the power function curve (top plot) is not only irregular, but also decreases at some sample sizes. These are the sample sizes for which it is necessary to increase the cutoff value (bottom plot) in order to keep the significance level (middle plot) no larger than 5%. We can find a smaller sample size within this range that gives the desired power of 80%:

```
min(Nvec(power>=0.80))
```

```
ans =

        1478
```

Testing a Correlation

In the examples we've considered so far, we were able to figure out the cutoff for a test statistic to achieve a certain significance level, and to calculate the probability of exceeding that cutoff under an alternative hypothesis. For our final example, let's consider a problem where that is not so easy.

Imagine we can take samples from two variables X and Y, and we want to know what sample size we would need to test whether they are uncorrelated versus the alternative that their correlation is as high as 0.4. Although it is possible to work out the distribution of the sample correlation by transforming it to a t distribution, let's use a method that we can use even in problems where we can't work out the distribution of the test statistic.

For a given sample size, we can use Monte Carlo simulation to determine an approximate cutoff value for a test of the correlation. Let's do a large simulation run so we can get this value accurately. We'll start with a sample size of 25.

```
nsamples = 10000;
N = 25;
alpha = 0.05;
conf = 1-alpha;
r = zeros(1,nsamples);
for j = 1:nsamples
    xy = normrnd(0,1,N,2);
    r(j) = corr(xy(:,1),xy(:,2));
end
cutoff = quantile(r,conf)
```

```
cutoff =

    0.3372
```

Then we can generate samples under the alternative hypothesis, and estimate the power of the test.

```
nsamples = 1000;
mu = [0; 0];
sig = [1 0.4; 0.4 1];
r = zeros(1,nsamples);
for j = 1:nsamples
    xy = mvnrnd(mu,sig,N);
    r(j) = corr(xy(:,1),xy(:,2));
end
[power,powerci] = binofit(sum(r>cutoff),nsamples)
```

```
power =

    0.6470


powerci =

    0.6165    0.6767
```

We estimate the power to be 65%, and we have 95% confidence that the true value is between 62% and 68%. To get a power of 80%, we need a larger sample size. We might try increasing N to 50, estimating the cutoff value for this sample size, and repeating the power simulation.

```
nsamples = 10000;
N = 50;
alpha = 0.05;
conf = 1-alpha;
r = zeros(1,nsamples);
for j = 1:nsamples
    xy = normrnd(0,1,N,2);
    r(j) = corr(xy(:,1),xy(:,2));
end
cutoff = quantile(r,conf)
```

```
nsamples = 1000;
mu = [0; 0];
sig = [1 0.4; 0.4 1];
r = zeros(1,nsamples);
for j = 1:nsamples
    xy = mvnrnd(mu,sig,N);
    r(j) = corr(xy(:,1),xy(:,2));
end
[power,powerci] = binofit(sum(r>cutoff),nsamples)


cutoff =

    0.2315


power =

    0.8990


powerci =

    0.8786    0.9170
```

This sample size gives a power better than our target of 80%. We could continue experimenting this way, trying to find a sample size less than 50 that would meet our requirements.

Conclusion

The probability functions in the Statistics and Machine Learning Toolbox can be used to determine the sample size required to achieve a desired level of power in a hypothesis test. In some problems the sample size can be compute directly; in others it is necessary to search over a range of sample sizes until the right value is found. Random number generators can help verify that the desired power is met, and can also be used to study the power of a specific test under alternative conditions.

Analysis of Variance

Introduction to Analysis of Variance

Analysis of variance (ANOVA) is a procedure for assigning sample variance to different sources and deciding whether the variation arises within or among different population groups. Samples are described in terms of variation around group means and variation of group means around an overall mean. If variations within groups are small relative to variations between groups, a difference in group means may be inferred. Hypothesis tests are used to quantify decisions.

One-Way ANOVA

In this section...

Introduction to One-Way ANOVA

You can use the function `anova1` to perform one-way analysis of variance (ANOVA). The purpose of one-way ANOVA is to determine whether data from several groups (levels) of a factor have a common mean. That is, one-way ANOVA enables you to find out whether different groups of an independent variable have different effects on the response variable *y*. Suppose, a hospital wants to determine if the two new proposed scheduling methods reduce patient wait times more than the old way of scheduling appointments. In this case, the independent variable is the scheduling method, and the response variable is the waiting time of the patients.

One-way ANOVA is a simple special case of the linear model on page 11-6. The one-way ANOVA form of the model is

$$y_{ij} = \alpha_j + \varepsilon_{ij}$$

with the following assumptions:

- y_{ij} is an observation, in which i represents the observation number, and j represents a different group (level) of the variable y. All y_{ij} are independent.
- α_j represents the population mean for the jth group (level or treatment).
- ε_{ij} is the random error, independent and normally distributed, with zero mean and constant variance, i.e., $\varepsilon_{ij} \sim N(0,\sigma^2)$.

This model is also called the *means model*. The model assumes that the columns of y are the constant α_j plus the error component ε_{ij}. ANOVA helps determine if the constants are all the same.

ANOVA tests the hypothesis that all group means are equal ($H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$) against the alternative hypothesis that at least one group is different from the others ($H_1: \alpha_i \neq \alpha_j$ for at least one i and j). `anova1(y)` tests the equality of column means for the data in matrix `y`, where each column is a different group and has the same number of observations (i.e., a balanced design). `anova1(y,group)` tests the equality of group means, specified in `group`, for the data in vector or matrix `y`. In this case, each group or column can have a different number of observations (i.e., an unbalanced design).

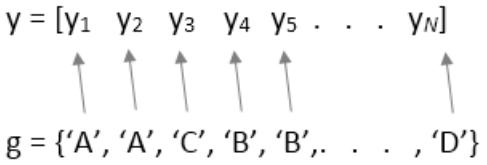
ANOVA is based on the assumption that all sample populations are normally distributed. It is known to be robust to modest violations of this assumption. You can check the normality assumption visually by using a normality plot (`normplot`). Alternatively, you can use one of the Statistics and Machine Learning Toolbox functions that checks for normality: the Anderson-Darling test (`adtest`), the chi-squared goodness of fit test (`chi2gof`), the Jarque-Bera test (`jbtest`), or the Lilliefors test (`lillietest`).

Prepare Data for One-Way ANOVA

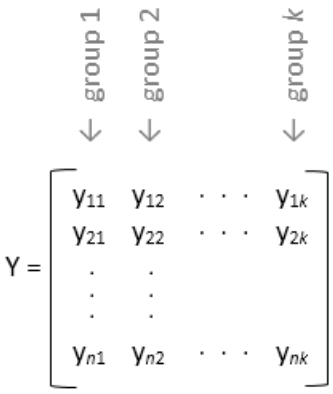
You can provide sample data as a vector or a matrix.

- If the sample data is in a vector, `y`, then you must provide grouping information using the `group` input variable: `anova1(y,group)`.

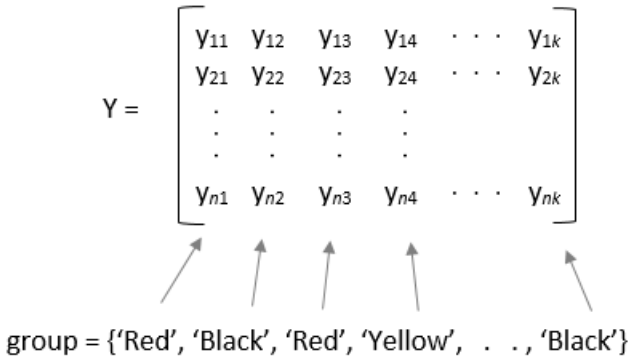
  `group` must be a numeric vector, logical vector, categorical vector, character array, string array, or cell array of character vectors, with one name for each element of `y`. The `anova1` function treats the `y` values corresponding to the same value of `group` as part of the same group. For example,

  Use this design when groups have different numbers of elements (unbalanced ANOVA).
- If the sample data is in a matrix, `y`, providing the group information is optional.
  - If you do not specify the input variable `group`, then `anova1` treats each column of `y` as a separate group, and evaluates whether the population means of the columns are equal. For example,

    Use this form of design when each group has the same number of elements (balanced ANOVA).
  - If you specify the input variable `group`, then each element in `group` represents a group name for the corresponding column in `y`. The `anova1` function treats the columns with the same group name as part of the same group. For example,



`anova1` ignores any `NaN` values in `y`. Also, if `group` contains empty or `NaN` values, `anova1` ignores the corresponding observations in `y`. The `anova1` function performs balanced ANOVA if each group has the same number of observations after the function disregards empty or `NaN` values. Otherwise, `anova1` performs unbalanced ANOVA.

Perform One-Way ANOVA

This example shows how to perform one-way ANOVA to determine whether data from several groups have a common mean.

Load and display the sample data.

```
load hogg
hogg
```

```
hogg = 6×5

    24    14    11     7    19
    15     7     9     7    24
    21    12     7     4    19
    27    17    13     7    15
    33    14    12    12    10
    23    16    18    18    20
```

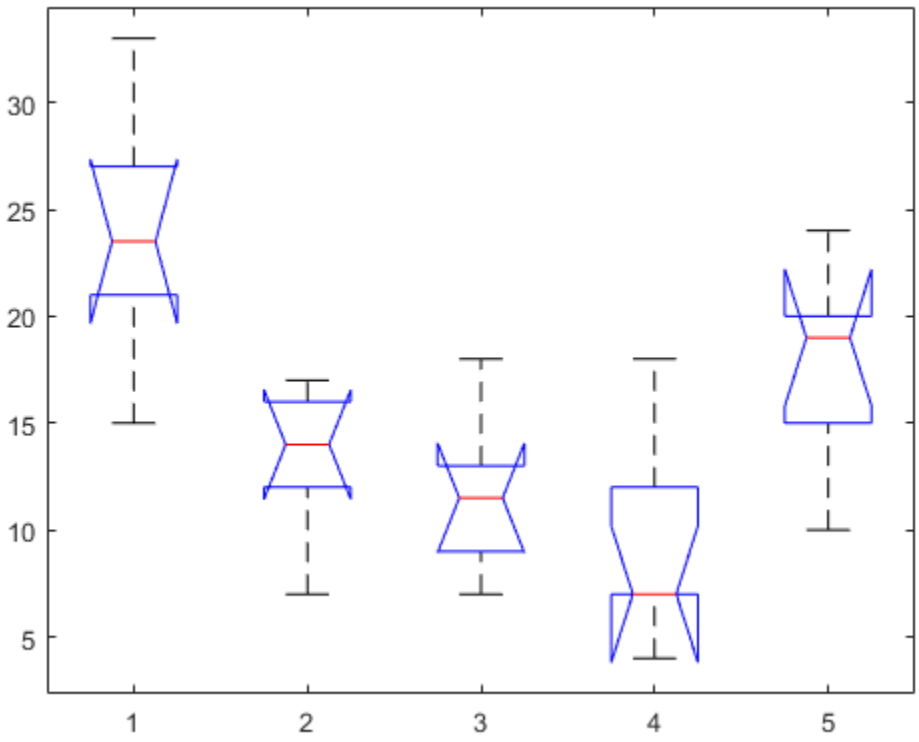
The data comes from a Hogg and Ledolter (1987) study on bacteria counts in shipments of milk. The columns of the matrix `hogg` represent different shipments. The rows are bacteria counts from cartons of milk chosen randomly from each shipment.

Test if some shipments have higher counts than others. By default, `anova1` returns two figures. One is the standard ANOVA table, and the other one is the box plots of data by group.

```
[p,tbl,stats] = anova1(hogg);
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	803	4	200.75	9.01	0.0001
Error	557.17	25	22.287		
Total	1360.17	29			




```
p
```

```
p = 1.1971e-04
```

The small *p*-value of about 0.0001 indicates that the bacteria counts from the different shipments are not the same.

You can get some graphical assurance that the means are different by looking at the box plots. The notches, however, compare the medians, not the means. For more information on this display, see `boxplot`.

View the standard ANOVA table. `anova1` saves the standard ANOVA table as a cell array in the output argument `tbl`.

```
tbl
```

```
tbl=4×6 cell array
  Columns 1 through 5
```

```
    {'Source' }    {'SS'        }    {'df'}    {'MS'      }    {'F'       }
    {'Columns'}    {[  803.0000]}    {[ 4]}    {[200.7500]}    {[  9.0076]}
    {'Error'  }    {[  557.1667]}    {[25]}    {[ 22.2867]}    {0x0 double}
    {'Total'  }    {[1.3602e+03]}    {[29]}    {0x0 double}    {0x0 double}

  Column 6

    {'Prob>F'    }
    {[1.1971e-04]}
    {0x0 double  }
    {0x0 double  }
```

Save the *F*-statistic value in the variable `Fstat`.

```
Fstat = tbl{2,5}
```

```
Fstat = 9.0076
```

View the statistics necessary to make a multiple pairwise comparison of group means. `anova1` saves these statistics in the structure `stats`.

```
stats
```

```
stats = struct with fields:
    gnames: [5x1 char]
         n: [6 6 6 6 6]
    source: 'anova1'
     means: [23.8333 13.3333 11.6667 9.1667 17.8333]
        df: 25
         s: 4.7209
```

ANOVA rejects the null hypothesis that all group means are equal, so you can use the multiple comparisons to determine which group means are different from others. To conduct multiple comparison tests, use the function `multcompare`, which accepts `stats` as an input argument. In this example, `anova1` rejects the null hypothesis that the mean bacteria counts from all four shipments are equal to each other, i.e., $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Perform a multiple comparison test to determine which shipments are different than the others in terms of mean bacteria counts.

```
multcompare(stats)
```


```
ans = 10×6

    1.0000    2.0000    2.4953   10.5000   18.5047    0.0059
    1.0000    3.0000    4.1619   12.1667   20.1714    0.0013
    1.0000    4.0000    6.6619   14.6667   22.6714    0.0001
    1.0000    5.0000   -2.0047    6.0000   14.0047    0.2119
    2.0000    3.0000   -6.3381    1.6667    9.6714    0.9719
    2.0000    4.0000   -3.8381    4.1667   12.1714    0.5544
    2.0000    5.0000  -12.5047   -4.5000    3.5047    0.4806
    3.0000    4.0000   -5.5047    2.5000   10.5047    0.8876
    3.0000    5.0000  -14.1714   -6.1667    1.8381    0.1905
    4.0000    5.0000  -16.6714   -8.6667   -0.6619    0.0292
```

The first two columns show which group means are compared with each other. For example, the first row compares the means for groups 1 and 2. The last column shows the *p*-values for the tests. The *p*-values 0.0059, 0.0013, and 0.0001 indicate that the mean bacteria counts in the milk from the first shipment is different from the ones from the second, third, and fourth shipments. The *p*-value of 0.0292 indicates that the mean bacteria counts in the milk from the fourth shipment is different from the ones from the fifth. The procedure fails to reject the hypotheses that the other group means are different from each other.

The figure also illustrates the same result. The blue bar shows the comparison interval for the first group mean, which does not overlap with the comparison intervals for the second, third, and fourth group means, shown in red. The comparison interval for the mean of fifth group, shown in gray,

overlaps with the comparison interval for the first group mean. Hence, the group means for the first and fifth groups are not significantly different from each other.

Mathematical Details

ANOVA tests for the difference in the group means by partitioning the total variation in the data into two components:

- Variation of group means from the overall mean, i.e., $\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$ (variation between groups), where $\bar{y}_{\cdot j}$ is the sample mean of group j, and $\bar{y}_{\cdot\cdot}$ is the overall sample mean.
- Variation of observations in each group from their group mean estimates, $y_{ij} - \bar{y}_{\cdot j}$ (variation within group).

In other words, ANOVA partitions the total sum of squares (SST) into sum of squares due to between-groups effect (SSR) and sum of squared errors(SSE).

$$\underbrace{\sum_i \sum_j \left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2}_{SST} = \underbrace{\sum_j n_j \left(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}\right)^2}_{SSR} + \underbrace{\sum_i \sum_j \left(y_{ij} - \bar{y}_{\cdot j}\right)^2}_{SSE},$$

where n_j is the sample size for the jth group, $j = 1, 2, \ldots, k$.

Then ANOVA compares the variation between groups to the variation within groups. If the ratio of between-group variation to within-group variation is significantly high, then you can conclude that the group means are significantly different from each other. You can measure this using a test statistic that has an F-distribution with (k – 1, N – k) degrees of freedom:

$$F = \frac{SSR/k-1}{SSE/N-k} = \frac{MSR}{MSE} \sim F_{k-1,\,N-k},$$

where MSR is the mean squared treatment, MSE is the mean squared error, k is the number of groups, and N is the total number of observations. If the p-value for the F-statistic is smaller than the significance level, then the test rejects the null hypothesis that all group means are equal and concludes that at least one of the group means is different from the others. The most common significance levels are 0.05 and 0.01.

ANOVA Table

The ANOVA table captures the variability in the model by source, the F-statistic for testing the significance of this variability, and the p-value for deciding on the significance of this variability. The p-value returned by `anova1` depends on assumptions about the random disturbances ε_{ij} in the model equation. For the p-value to be correct, these disturbances need to be independent, normally distributed, and have constant variance. The standard ANOVA table has this form:

Source	SS	df	MS	F	p-value
Group (Between)	SSR	$k-1$	MSR = SSR/($k-1$)	MSR/MSE	$P(F_{k-1,N-k}) > F$
Error (Within)	SSE	$N-k$	MSE = SSE/($N-k$)		
Total	SST	$N-1$			

`anova1` returns the standard ANOVA table as a cell array with six columns.

Column	Definition
`Source`	Source of the variability.
`SS`	Sum of squares due to each source.
`df`	Degrees of freedom associated with each source. Suppose N is the total number of observations and k is the number of groups. Then, $N – k$ is the within-groups degrees of freedom (`Error`), $k – 1$ is the between-groups degrees of freedom (`Columns`), and $N – 1$ is the total degrees of freedom: $N – 1 = (N – k) + (k – 1)$.
`MS`	Mean squares for each source, which is the ratio `SS/df`.
`F`	F-statistic, which is the ratio of the mean squares.
`Prob>F`	p-value, which is the probability that the F-statistic can take a value larger than the computed test-statistic value. `anova1` derives this probability from the cdf of the F-distribution.

The rows of the ANOVA table show the variability in the data, divided by the source.

Row (Source)	Definition
`Groups` or `Columns`	Variability due to the differences among the group means (variability *between* groups)
`Error`	Variability due to the differences between the data in each group and the group mean (variability *within* groups)
`Total`	Total variability

References

[1] Wu, C. F. J., and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization*, 2000.

[2] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 4th ed. *Applied Linear Statistical Models*. Irwin Press, 1996.

See Also

`anova1` | `kruskalwallis` | `multcompare`

More About

- “Two-Way ANOVA” on page 9-11
- “N-Way ANOVA” on page 9-25
- “Multiple Comparisons” on page 9-18
- “Nonparametric Methods” on page 9-47

Two-Way ANOVA

In this section...

Introduction to Two-Way ANOVA

You can use the function `anova2` to perform a balanced two-way analysis of variance (ANOVA). To perform two-way ANOVA for an unbalanced design, use `anovan`. For an example, see “Two-Way ANOVA for Unbalanced Design” on page 33-82.

As in one-way ANOVA, the data for a two-way ANOVA study can be experimental or observational. The difference between one-way and two-way ANOVA is that in two-way ANOVA, the effects of two factors on a response variable are of interest. These two factors can be independent, and have no interaction effect, or the impact of one factor on the response variable can depend on the group (level) of the other factor. If the two factors have no interactions, the model is called an *additive* model.

Suppose an automobile company has two factories, and each factory makes the same three car models. The gas mileage in the cars can vary from factory to factory and from model to model. These two factors, factory and model, explain the differences in mileage, that is, the response. One measure of interest is the difference in mileage due to the production methods between factories. Another measure of interest is the difference in the mileage of the models (irrespective of the factory) due to different design specifications. The effects of these measures of interest are *additive*. In addition, suppose only one model has different gas mileage between factories, while the mileage of the other two models is the same between factories. This is called an *interaction* effect. To measure an interaction effect, there must be multiple observations for some combination of factory and car model. These multiple observations are called *replications*.

Two-way ANOVA is a special case of the linear model on page 11-6. The two-way ANOVA form of the model is

$$y_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijr}$$

where,

- y_{ijr} is an observation of the response variable.
  - i represents group i of row factor A, $i = 1, 2, \ldots, I$.
  - j represents group j of column factor B, $j = 1, 2, \ldots, J$.
  - r represents the replication number, $r = 1, 2, \ldots, R$.

  There are a total of $N = I*J*R$ observations.
- μ is the overall mean.
- α_i are the deviations of groups defined by row factor A from the overall mean μ. The values of α_i sum to 0.

$$\sum_{i=1}^{I}\alpha_i = 0.$$

- β_j are the deviations of groups defined by column factor B from the overall mean μ. The values of β_j sum to 0.

$$\sum_{j=1}^{J}\beta_j = 0.$$

- $\alpha\beta_{ij}$ are the interactions. The values in each row and in each column of $\alpha\beta_{ij}$ sum to 0.

$$\sum_{i=1}^{I}(\alpha\beta)_{ij} = \sum_{j=1}^{J}(\alpha\beta)_{ij} = 0.$$

- ε_{ijr} are the random disturbances. They are assumed to be independent, normally distributed, and have constant variance.

In the mileage example:

- y_{ijr} are the gas mileage observations, μ is the overall mean gas mileage.
- α_i are the deviations of each car's gas mileage from the mean gas mileage μ due to the car's *model*.
- β_j are the deviations of each car's gas mileage from the mean gas mileage μ due to the car's *factory*.

`anova2` requires that data be balanced, so each combination of model and factory must have the same number of cars.

Two-way ANOVA tests hypotheses about the effects of factors A and B, and their interaction on the response variable y. The hypotheses about the equality of the mean response for groups of row factor A are

$$H_0: \alpha_1 = \alpha_2 \cdots = \alpha_I$$
$$H_1: \text{ at least one } \alpha_i \text{ is different, } i = 1,\ 2,\ \ldots,\ I.$$

The hypotheses about the equality of the mean response for groups of column factor B are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_J$$
$$H_1: \text{ at least one } \beta_j \text{ is different, } j = 1,\ 2,\ \ldots,\ J.$$

The hypotheses about the interaction of the column and row factors are

$$H_0: (\alpha\beta)_{ij} = 0$$
$$H_1: \text{at least one } (\alpha\beta)_{ij} \neq 0$$

Prepare Data for Balanced Two-Way ANOVA

To perform balanced two-way ANOVA using `anova2`, you must arrange data in a specific matrix form. The columns of the matrix must correspond to groups of the column factor, B. The rows must correspond to the groups of the row factor, A, with the same number of replications for each combination of the groups of factors A and B.

Suppose that row factor A has three groups, and column factor B has two groups (levels). Also suppose that each combination of factors A and B has two measurements or observations (`reps = 2`). Then, each group of factor A has six observations and each group of factor B four observations.

$$\begin{array}{cc} B=1 & B=2 \\ \left[\begin{array}{cc} y_{111} & y_{121} \\ y_{112} & y_{122} \\ y_{211} & y_{221} \\ y_{212} & y_{222} \\ y_{311} & y_{321} \\ y_{312} & y_{322} \end{array}\right] & \begin{array}{l} \} A=1 \\ \} A=2 \\ \} A=3 \end{array} \end{array}$$

The subscripts indicate row, column, and replication, respectively. For example, y_{221} corresponds to the measurement for the second group of factor A, the second group of factor B, and the first replication for this combination.

Perform Two-Way ANOVA

This example shows how to perform two-way ANOVA to determine the effect of car model and factory on the mileage rating of cars.

Load and display the sample data.

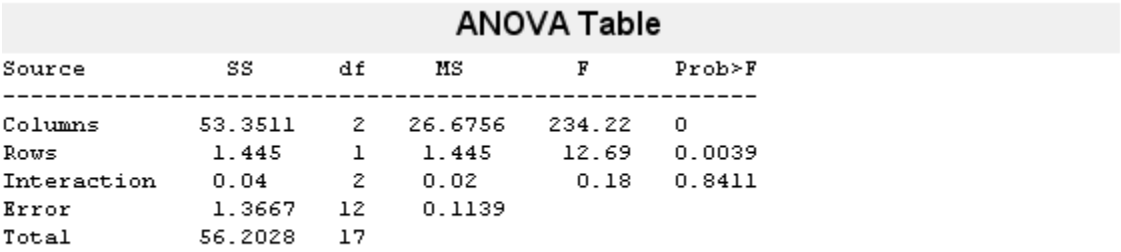
```
load mileage
mileage
```

```
mileage = 6×3

   33.3000   34.5000   37.4000
   33.4000   34.8000   36.8000
   32.9000   33.8000   37.6000
   32.6000   33.4000   36.6000
   32.5000   33.7000   37.0000
   33.0000   33.9000   36.7000
```

There are three car models (columns) and two factories (rows). The data has six mileage rows because each factory provided three cars of each model for the study (i.e., the replication number is three). The data from the first factory is in the first three rows, and the data from the second factory is in the last three rows.

Perform two-way ANOVA. Return the structure of statistics, `stats`, to use in multiple comparisons.

```
nmbcars = 3; % Number of cars from each model, i.e., number of replications
[~,~,stats] = anova2(mileage,nmbcars);
```



You can use the *F*-statistics to do hypotheses tests to find out if the mileage is the same across models, factories, and model - factory pairs. Before performing these tests, you must adjust for the additive effects. `anova2` returns the *p*-value from these tests.

The *p*-value for the model effect (`Columns`) is zero to four decimal places. This result is a strong indication that the mileage varies from one model to another.

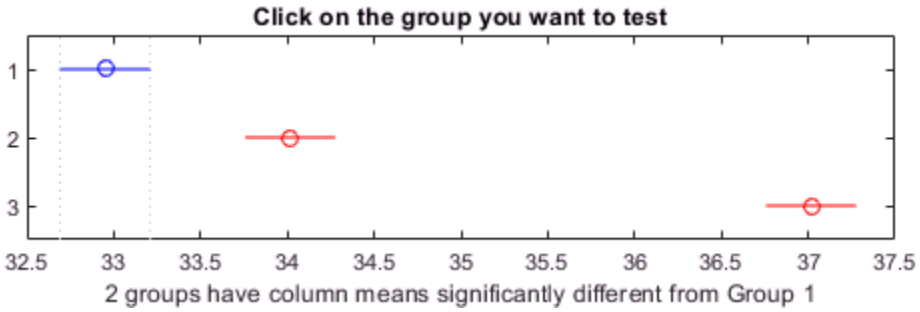
The *p*-value for the factory effect (`Rows`) is 0.0039, which is also highly significant. This value indicates that one factory is out-performing the other in the gas mileage of the cars it produces. The observed *p*-value indicates that an *F*-statistic as extreme as the observed *F* occurs by chance about four out of 1000 times, if the gas mileage were truly equal from factory to factory.

The factories and models appear to have no interaction. The *p*-value, 0.8411, means that the observed result is likely (84 out of 100 times), given that there is no interaction.

Perform “Multiple Comparisons” on page 9-18 to find out which pair of the three car models is significantly different.

```
c = multcompare(stats)
```

```
Note: Your model includes an interaction term.  A test of main effects can be
difficult to interpret when the model includes interactions.
```




```
c = 3×6

    1.0000    2.0000   -1.5865   -1.0667   -0.5469    0.0004
    1.0000    3.0000   -4.5865   -4.0667   -3.5469    0.0000
    2.0000    3.0000   -3.5198   -3.0000   -2.4802    0.0000
```

In the matrix `c`, the first two columns show the pairs of car models that are compared. The last column shows the *p*-values for the test. All *p*-values are small (0.0004, 0, and 0), which indicates that the mean mileage of all car models are significantly different from each other.

In the figure, the blue bar is the comparison interval for the mean mileage of the first car model. The red bars are the comparison intervals for the mean mileage of the second and third car models. None of the second and third comparison intervals overlap with the first comparison interval, indicating that the mean mileage of the first car model is different from the mean mileage of the second and the third car models. If you click on one of the other bars, you can test for the other car models. None of the comparison intervals overlap, indicating that the mean mileage of each car model is significantly different from the other two.

Mathematical Details

The two-factor ANOVA partitions the total variation into the following components:

- Variation of row factor group means from the overall mean, $\bar{y}_{i..} - \bar{y}_{...}$
- Variation of column factor group means from the overall mean, $\bar{y}_{.j.} - \bar{y}_{...}$
- Variation of overall mean plus the replication mean from the column factor group mean plus row factor group mean, $\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$
- Variation of observations from the replication means, $y_{ijk} - \bar{y}_{ij.}$

ANOVA partitions the total sum of squares (SST) into the sum of squares due to row factor A (SS_A), the sum of squares due to column factor B (SS_B), the sum of squares due to interaction between A and B (SS_{AB}), and the sum of squares error (SSE).

$$\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{r=1}^{R}\left(y_{ijk}-\bar{y}_{...}\right)^2}_{SST} = \underbrace{kR\sum_{i=1}^{m}\left(\bar{y}_{i..}-\bar{y}_{...}\right)^2}_{SS_B} + \underbrace{mR\sum_{j=1}^{k}\left(\bar{y}_{.j.}-\bar{y}_{...}\right)^2}_{SS_A}$$
$$+\underbrace{R\sum_{i=1}^{m}\sum_{j=1}^{k}\left(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}\right)^2}_{SS_{AB}} + \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{r=1}^{R}\left(y_{ijk}-\bar{y}_{ij.}\right)^2}_{SSE}$$

ANOVA takes the variation due to the factor or interaction and compares it to the variation due to error. If the ratio of the two variations is high, then the effect of the factor or the interaction effect is statistically significant. You can measure the statistical significance using a test statistic that has an F-distribution.

For the null hypothesis that the mean response for groups of the row factor A are equal, the test statistic is

$$F = \frac{SS_B/(m-1)}{SSE/mk(R-1)} \sim F_{m-1,\,mk(R-1)}.$$

For the null hypothesis that the mean response for groups of the column factor B are equal, the test statistic is

$$F = \frac{SS_A/(k-1)}{SSE/mk(R-1)} \sim F_{k-1,\,mk(R-1)}.$$

For the null hypothesis that the interaction of the column and row factors are equal to zero, the test statistic is

$$F = \frac{SS_{AB}/(m-1)(k-1)}{SSE/mk(R-1)} \sim F_{(m-1)(k-1),\,mk(R-1)}.$$

If the p-value for the F-statistic is smaller than the significance level, then ANOVA rejects the null hypothesis. The most common significance levels are 0.01 and 0.05.

ANOVA Table

The ANOVA table captures the variability in the model by the source, the F-statistic for testing the significance of this variability, and the p-value for deciding on the significance of this variability. The

p-value returned by `anova2` depends on assumptions about the random disturbances, ε_{ij}, in the model equation. For the *p*-value to be correct, these disturbances need to be independent, normally distributed, and have constant variance. The standard ANOVA table has this form:

Source	SS	df	MS	F	*p*-value
Columns	SS_A	$k - 1$	MS_A	MS_A/MSE	$P(F_{k-1,mk(R-1)}) > F$
Rows	SS_B	$m - 1$	MS_B	MS_B/MSE	$P(F_{m-1,mk(R-1)}) > F$
Interaction	SS_{AB}	$(m-1)(k-1)$	MS_{AB}	MS_{AB}/MSE	$P(F_{(m-1)(k-1),mk(R-1)}) > F$
Error	SSE	$mk(R-1)$	MSE		
Total	SST	$mkR - 1$			

`anova2` returns the standard ANOVA table as a cell array with six columns.

Column	Definition
`Source`	The source of the variability.
`SS`	The sum of squares due to each source.
`df`	The degrees of freedom associated with each source. Suppose *J* is the number of groups in the column factor, *I* is the number of groups in the row factor, and *R* is the number of replications. Then, the total number of observations is *IJR* and the total degrees of freedom is *IJR* – 1. *I* – 1 is the degrees of freedom for the row factor,*J* – 1 is the degrees of freedom for the column factor, (*I* – 1)(*J* – 1) is the interaction degrees of freedom, and *IJ*(*R* – 1) is the error degrees of freedom.
`MS`	The mean squares for each source, which is the ratio `SS/df`.
`F`	*F*-statistic, which is the ratio of the mean squares.
`Prob>F`	The *p*-value, which is the probability that the *F*-statistic can take a value larger than the computed test-statistic value. `anova2` derives this probability from the cdf of the *F*-distribution.

The rows of the ANOVA table show the variability in the data that is divided by the source.

Row (Source)	Definition
`Columns`	Variability due to the column factor
`Rows`	Variability due to the row factor
`Interaction`	Variability due to the interaction of the row and column factors
`Error`	Variability due to the differences between the data in each group and the group mean (variability *within* groups)

Row (Source)	Definition
`Total`	Total variability

References

[1] Wu, C. F. J., and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization*, 2000.

[2] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 4th ed. *Applied Linear Statistical Models*. Irwin Press, 1996.

See Also

`anova1` | `anova2` | `anovan` | `multcompare`

Related Examples

- “Two-Way ANOVA for Unbalanced Design” on page 33-82

More About

- “One-Way ANOVA” on page 9-3
- “N-Way ANOVA” on page 9-25
- “Multiple Comparisons” on page 9-18
- “Nonparametric Methods” on page 9-47

Multiple Comparisons

In this section...

Introduction

Analysis of variance (ANOVA) techniques test whether a set of group means (treatment effects) are equal or not. Rejection of the null hypothesis leads to the conclusion that not all group means are the same. This result, however, does not provide further information on which group means are different.

Performing a series of *t*-tests to determine which pairs of means are significantly different is not recommended. When you perform multiple *t*-tests, the probability that the means appear significant, and significant difference results might be due to large number of tests. These *t*-tests use the data from the same sample, hence they are not independent. This fact makes it more difficult to quantify the level of significance for multiple tests.

Suppose that in a single *t*-test, the probability that the null hypothesis (H_0) is rejected when it is actually true is a small value, say 0.05. Suppose also that you conduct six independent *t*-tests. If the significance level for each test is 0.05, then the probability that the tests correctly fail to reject H_0, when H_0 is true for each case, is $(0.95)^6 = 0.735$. And the probability that one of the tests incorrectly rejects the null hypothesis is 1 – 0.735 = 0.265, which is much higher than 0.05.

To compensate for multiple tests, you can use multiple comparison procedures. The Statistics and Machine Learning Toolbox function `multcompare` performs multiple pairwise comparison of the group means, or treatment effects. The options are Tukey’s honestly significant difference criterion (default option), the Bonferroni method, Scheffe’s procedure, Fisher’s least significant differences (lsd) method, and Dunn & Sidák’s approach to *t*-test.

To perform multiple comparisons of group means, provide the structure `stats` as an input for `multcompare`. You can obtain `stats` from one of the following functions :

- `anova1` — One-way ANOVA on page 9-3
- `anova2` — Two-way ANOVA on page 9-11
- `anovan` — *N*-way ANOVA on page 9-25
- `aoctool` — Interactive ANCOVA on page 9-39
- `kruskalwallis` — Nonparametric method on page 9-47 for one-way layout
- `friedman` — Nonparametric method on page 9-47 for two-way layout

For multiple comparison procedure options for repeated measures, see `multcompare` (`RepeatedMeasuresModel`).

Multiple Comparisons Using One-Way ANOVA

Load the sample data.

```
load carsmall
```

`MPG` represents the miles per gallon for each car, and `Cylinders` represents the number of cylinders in each car, either 4, 6, or 8 cylinders.

Test if the mean miles per gallon (mpg) is different across cars that have different numbers of cylinders. Also compute the statistics needed for multiple comparison tests.

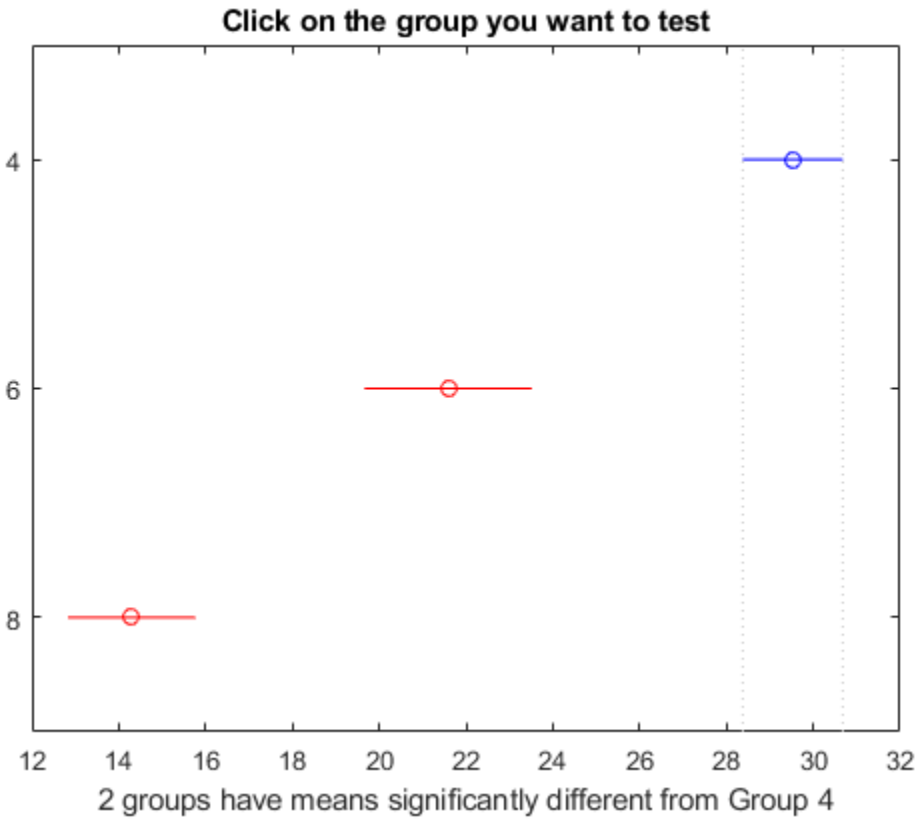
```
[p,~,stats] = anova1(MPG,Cylinders,'off');
p
```

```
p = 4.4902e-24
```

The small *p*-value of about 0 is a strong indication that mean miles per gallon is significantly different across cars with different numbers of cylinders.

Perform a multiple comparison test, using the Bonferroni method, to determine which numbers of cylinders make a difference in the performance of the cars.

```
[results,means] = multcompare(stats,'CType','bonferroni')
```




```
results = 3×6

    1.0000    2.0000    4.8605    7.9418   11.0230    0.0000
    1.0000    3.0000   12.6127   15.2337   17.8548    0.0000
    2.0000    3.0000    3.8940    7.2919   10.6899    0.0000
```

```
means = 3×2

   29.5300    0.6363
   21.5882    1.0913
   14.2963    0.8660
```

In the `results` matrix, 1, 2, and 3 correspond to cars with 4, 6, and 8 cylinders, respectively. The first two columns show which groups are compared. For example, the first row compares the cars with 4 and 6 cylinders. The fourth column shows the mean mpg difference for the compared groups. The third and fifth columns show the lower and upper limits for a 95% confidence interval for the difference in the group means. The last column shows the *p*-values for the tests. All *p*-values are zero, which indicates that the mean mpg for all groups differ across all groups.

In the figure the blue bar represents the group of cars with 4 cylinders. The red bars represent the other groups. None of the red comparison intervals for the mean mpg of cars overlap, which means that the mean mpg is significantly different for cars having 4, 6, or 8 cylinders.

The first column of the `means` matrix has the mean mpg estimates for each group of cars. The second column contains the standard errors of the estimates.

Multiple Comparisons for Three-Way ANOVA

Load the sample data.

```
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';
g1 = [1 2 1 2 1 2 1 2];
g2 = {'hi';'hi';'lo';'lo';'hi';'hi';'lo';'lo'};
g3 = {'may';'may';'may';'may';'june';'june';'june';'june'};
```

`y` is the response vector and `g1`, `g2`, and `g3` are the grouping variables (factors). Each factor has two levels, and every observation in `y` is identified by a combination of factor levels. For example, observation `y(1)` is associated with level 1 of factor `g1`, level `'hi'` of factor `g2`, and level `'may'` of factor `g3`. Similarly, observation `y(6)` is associated with level 2 of factor `g1`, level `'hi'` of factor `g2`, and level `'june'` of factor `g3`.

Test if the response is the same for all factor levels. Also compute the statistics required for multiple comparison tests.

```
[~,~,stats] = anovan(y,{g1 g2 g3},'model','interaction',...
    'varnames',{'g1','g2','g3'});
```

Analysis of Variance

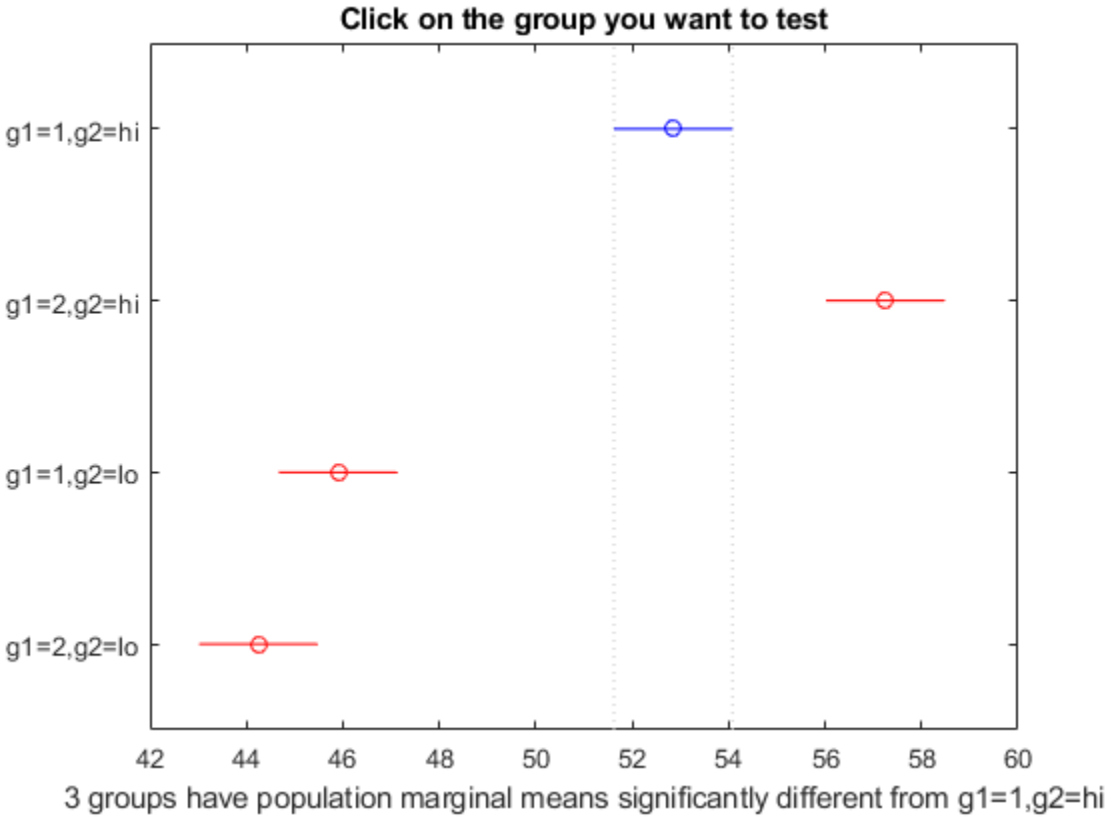
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
g1	3.781	1	3.781	336.11	0.0347
g2	199.001	1	199.001	17689	0.0048
g3	0.061	1	0.061	5.44	0.2578
g1*g2	18.301	1	18.301	1626.78	0.0158
g1*g3	0.211	1	0.211	18.78	0.1444
g2*g3	0.011	1	0.011	1	0.5
Error	0.011	1	0.011		
Total	221.379	7			

Constrained (Type III) sums of squares.

The *p*-value of 0.2578 indicates that the mean responses for levels `'may'` and `'june'` of factor `g3` are not significantly different. The *p*-value of 0.0347 indicates that the mean responses for levels `1` and `2` of factor `g1` are significantly different. Similarly, the *p*-value of 0.0048 indicates that the mean responses for levels `'hi'` and `'lo'` of factor `g2` are significantly different.

Perform multiple comparison tests to find out which groups of the factors `g1` and `g2` are significantly different.

```
results = multcompare(stats,'Dimension',[1 2])
```



```
results = 6×6

    1.0000    2.0000   -6.8604   -4.4000   -1.9396    0.0272
    1.0000    3.0000    4.4896    6.9500    9.4104    0.0170
    1.0000    4.0000    6.1396    8.6000   11.0604    0.0136
    2.0000    3.0000    8.8896   11.3500   13.8104    0.0101
    2.0000    4.0000   10.5396   13.0000   15.4604    0.0087
    3.0000    4.0000   -0.8104    1.6500    4.1104    0.0737
```

`multcompare` compares the combinations of groups (levels) of the two grouping variables, `g1` and `g2`. In the `results` matrix, the number 1 corresponds to the combination of level `1` of `g1` and level `hi` of `g2`, the number 2 corresponds to the combination of level `2` of `g1` and level `hi` of `g2`. Similarly, the number 3 corresponds to the combination of level `1` of `g1` and level `lo` of `g2`, and the number 4 corresponds to the combination of level `2` of `g1` and level `lo` of `g2`. The last column of the matrix contains the *p*-values.

For example, the first row of the matrix shows that the combination of level `1` of `g1` and level `hi` of `g2` has the same mean response values as the combination of level `2` of `g1` and level `hi` of `g2`. The *p*-value corresponding to this test is 0.0280, which indicates that the mean responses are significantly different. You can also see this result in the figure. The blue bar shows the comparison interval for the mean response for the combination of level `1` of `g1` and level `hi` of `g2`. The red bars are the comparison intervals for the mean response for other group combinations. None of the red bars overlap with the blue bar, which means the mean response for the combination of level `1` of `g1` and level `hi` of `g2` is significantly different from the mean response for other group combinations.

You can test the other groups by clicking on the corresponding comparison interval for the group. The bar you click on turns to blue. The bars for the groups that are significantly different are red. The bars for the groups that are not significantly different are gray. For example, if you click on the comparison interval for the combination of level `1` of `g1` and level `lo` of `g2`, the comparison interval for the combination of level `2` of `g1` and level `lo` of `g2` overlaps, and is therefore gray. Conversely, the other comparison intervals are red, indicating significant difference.

Multiple Comparison Procedures

To specify the multiple comparison procedure you want `multcompare` to conduct use the `'CType'` name-value pair argument. `multcompare` provides the following procedures:

- “Tukey’s Honestly Significant Difference Procedure” on page 9-22
- “Bonferroni Method” on page 9-23
- “Dunn & Sidák’s Approach” on page 9-23
- “Least Significant Difference” on page 9-23
- “Scheffe’s Procedure” on page 9-24

Tukey’s Honestly Significant Difference Procedure

You can specify Tukey’s honestly significant difference procedure using the `'CType','Tukey-Kramer'` or `'CType','hsd'` name-value pair argument. The test is based on studentized range distribution. Reject $H_0{:}\alpha_i = \alpha_j$ if

$$|t| = \frac{\left|\bar{y}_i - \bar{y}_j\right|}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > \frac{1}{\sqrt{2}} q_{\alpha, k, N-k,}$$

where $q_{\alpha, k, N-k}$ is the upper 100*(1 – α)th percentile of the studentized range distribution with parameter k and N – k degrees of freedom. k is the number of groups (treatments or marginal means) and N is the total number of observations.

Tukey's honestly significant difference procedure is optimal for balanced one-way ANOVA and similar procedures with equal sample sizes. It has been proven to be conservative for one-way ANOVA with different sample sizes. According to the unproven Tukey-Kramer conjecture, it is also accurate for problems where the quantities being compared are correlated, as in analysis of covariance with unbalanced covariate values.

Bonferroni Method

You can specify the Bonferroni method using the `'CType','bonferroni'` name-value pair. This method uses critical values from Student's t-distribution after an adjustment to compensate for multiple comparisons. The test rejects H_0:$\alpha_i = \alpha_j$ at the $\alpha/2\binom{k}{2}$ significance level, where k is the number of groups if

$$|t| = \frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > t_{\alpha/2\binom{k}{2}, N-k,}$$

where N is the total number of observations and k is the number of groups (marginal means). This procedure is conservative, but usually less so than the Scheffé procedure.

Dunn & Sidák's Approach

You can specify Dunn & Sidák's approach using the `'CType','dunn-sidak'` name-value pair argument. It uses critical values from the t-distribution, after an adjustment for multiple comparisons that was proposed by Dunn and proved accurate by Sidák. This test rejects H_0:$\alpha_i = \alpha_j$ if

$$|t| = \frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > t_{1-\eta/2, \nu,}$$

where

$$\eta = 1 - (1-\alpha)^{1/\binom{k}{2}}$$

and k is the number of groups. This procedure is similar to, but less conservative than, the Bonferroni procedure.

Least Significant Difference

You can specify the least significance difference procedure using the `'CType','lsd'` name-value pair argument. This test uses the test statistic

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

It rejects H_0:$\alpha_i = \alpha_j$ if

$$|\bar{y}_i - \bar{y}_j| > t_{\alpha/2, N-k} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$
$$\text{LSD}$$

Fisher suggests a protection against multiple comparisons by performing LSD only when the null hypothesis H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_k$ is rejected by ANOVA *F*-test. Even in this case, LSD might not reject any of the individual hypotheses. It is also possible that ANOVA does not reject H_0, even when there are differences between some group means. This behavior occurs because the equality of the remaining group means can cause the *F*-test statistic to be nonsignificant. Without any condition, LSD does not provide any protection against the multiple comparison problem.

Scheffe's Procedure

You can specify Scheffe's procedure using the `'CType','scheffe'` name-value pair argument. The critical values are derived from the *F* distribution. The test rejects H_0:$\alpha_i = \alpha_j$ if

$$\frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > \sqrt{(k-1)F_{k-1, N-k, \alpha}}$$

This procedure provides a simultaneous confidence level for comparisons of all linear combinations of the means. It is conservative for comparisons of simple differences of pairs.

References

[1] Milliken G. A. and D. E. Johnson. *Analysis of Messy Data. Volume I: Designed Experiments*. Boca Raton, FL: Chapman & Hall/CRC Press, 1992.

[2] Neter J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman. 4th ed. *Applied Linear Statistical Models*.Irwin Press, 1996.

[3] Hochberg, Y., and A. C. Tamhane. *Multiple Comparison Procedures*. Hoboken, NJ: John Wiley & Sons, 1987.

See Also

`anova1` | `anova2` | `anovan` | `aoctool` | `friedman` | `kruskalwallis` | `multcompare`

Related Examples

- "Perform One-Way ANOVA" on page 9-5
- "Perform Two-Way ANOVA" on page 9-13

N-Way ANOVA

In this section...

Introduction to N-Way ANOVA

You can use the function `anovan` to perform *N*-way ANOVA. Use *N*-way ANOVA to determine if the means in a set of data differ with respect to groups (levels) of multiple factors. By default, `anovan` treats all grouping variables as fixed effects. For an example of ANOVA with random effects, see “ANOVA with Random Effects” on page 9-33. For repeated measures, see `fitrm` and `ranova`.

N-way ANOVA is a generalization of two-way ANOVA. For three factors, for example, the model can be written as

$$y_{ijkr} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkr},$$

where

- y_{ijkr} is an observation of the response variable. i represents group i of factor A, $i = 1, 2, \ldots, I$, j represents group j of factor B, $j = 1, 2, \ldots, J$, k represents group k of factor C, and r represents the replication number, $r = 1, 2, \ldots, R$. For constant R, there are a total of $N = I*J*K*R$ observations, but the number of observations does not have to be the same for each combination of groups of factors.
- μ is the overall mean.
- α_i are the deviations of groups of factor A from the overall mean μ due to factor A. The values of α_i sum to 0.

  $$\sum_{i=1}^{I} \alpha_i = 0.$$
- β_j are the deviations of groups in factor B from the overall mean μ due to factor B. The values of β_j sum to 0.

  $$\sum_{j=1}^{J} \beta_j = 0.$$
- γ_k are the deviations of groups in factor C from the overall mean μ due to factor C. The values of γ_k sum to 0.

  $$\sum_{k=1}^{K} \gamma_k = 0.$$
- $(\alpha\beta)_{ij}$ is the interaction term between factors A and B. $(\alpha\beta)_{ij}$ sum to 0 over either index.

  $$\sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0.$$
- $(\alpha\gamma)_{ik}$ is the interaction term between factors A and C. The values of $(\alpha\gamma)_{ik}$ sum to 0 over either index.

  $$\sum_{i=1}^{I} (\alpha\gamma)_{ik} = \sum_{k=1}^{K} (\alpha\gamma)_{ik} = 0.$$

- $(\beta\gamma)_{jk}$ is the interaction term between factors B and C. The values of $(\beta\gamma)_{jk}$ sum to 0 over either index.

  $$\sum_{j=1}^{J}(\beta\gamma)_{jk} = \sum_{k=1}^{K}(\beta\gamma)_{jk} = 0.$$

- $(\alpha\beta\gamma)_{ijk}$ is the three-way interaction term between factors A, B, and C. The values of $(\alpha\beta\gamma)_{ijk}$ sum to 0 over any index.

  $$\sum_{i=1}^{I}(\alpha\beta\gamma)_{ijk} = \sum_{j=1}^{J}(\alpha\beta\gamma)_{ijk} = \sum_{k=1}^{K}(\alpha\beta\gamma)_{ijk} = 0.$$

- ε_{ijkr} are the random disturbances. They are assumed to be independent, normally distributed, and have constant variance.

Three-way ANOVA tests hypotheses about the effects of factors A, B, C, and their interactions on the response variable y. The hypotheses about the equality of the mean responses for groups of factor A are

$$H_0: \alpha_1 = \alpha_2 \cdots = \alpha_I$$
$$H_1: \text{ at least one } \alpha_i \text{ is different, } i = 1, 2, \ldots, I.$$

The hypotheses about the equality of the mean response for groups of factor B are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_J$$
$$H_1: \text{ at least one } \beta_j \text{ is different, } j = 1, 2, \ldots, J.$$

The hypotheses about the equality of the mean response for groups of factor C are

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_K$$
$$H_1: \text{ at least one } \gamma_k \text{ is different, } k = 1, 2, \ldots, K.$$

The hypotheses about the interaction of the factors are

$$H_0: (\alpha\beta)_{ij} = 0$$
$$H_1: \text{at least one } (\alpha\beta)_{ij} \neq 0$$

$$H_0: (\alpha\gamma)_{ik} = 0$$
$$H_1: \text{at least one } (\alpha\gamma)_{ik} \neq 0$$

$$H_0: (\beta\gamma)_{jk} = 0$$
$$H_1: \text{at least one } (\beta\gamma)_{jk} \neq 0$$

$$H_0: (\alpha\beta\gamma)_{ijk} = 0$$
$$H_1: \text{at least one } (\alpha\beta\gamma)_{ijk} \neq 0$$

In this notation parameters with two subscripts, such as $(\alpha\beta)_{ij}$, represent the interaction effect of two factors. The parameter $(\alpha\beta\gamma)_{ijk}$ represents the three-way interaction. An ANOVA model can have the full set of parameters or any subset, but conventionally it does not include complex interaction terms unless it also includes all simpler terms for those factors. For example, one would generally not include the three-way interaction without also including all two-way interactions.

Prepare Data for N-Way ANOVA

Unlike `anova1` and `anova2`, `anovan` does not expect data in a tabular form. Instead, it expects a vector of response measurements and a separate vector (or text array) containing the values corresponding to each factor. This input data format is more convenient than matrices when there are more than two factors or when the number of measurements per factor combination is not constant.

$$
\begin{array}{rcccccccccc}
y & = & [& y_1, & y_2, & y_3, & y_4, & y_5, & \cdots, & y_N &]' \\
 & & & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & & \uparrow & \\
g1 & = & \{ & 'A', & 'A', & 'C', & 'B', & 'B', & \cdots, & 'D' & \} \\
g2 & = & [& 1 & 2 & 1 & 3 & 1 & \cdots, & 2 &] \\
g3 & = & \{ & \text{'hi'}, & \text{'mid'}, & \text{'low'}, & \text{'mid'}, & \text{'hi'}, & \cdots, & \text{'low'} & \}
\end{array}
$$

Perform N-Way ANOVA

This example shows how to perform N-way ANOVA on car data with mileage and other information on 406 cars made between 1970 and 1982.

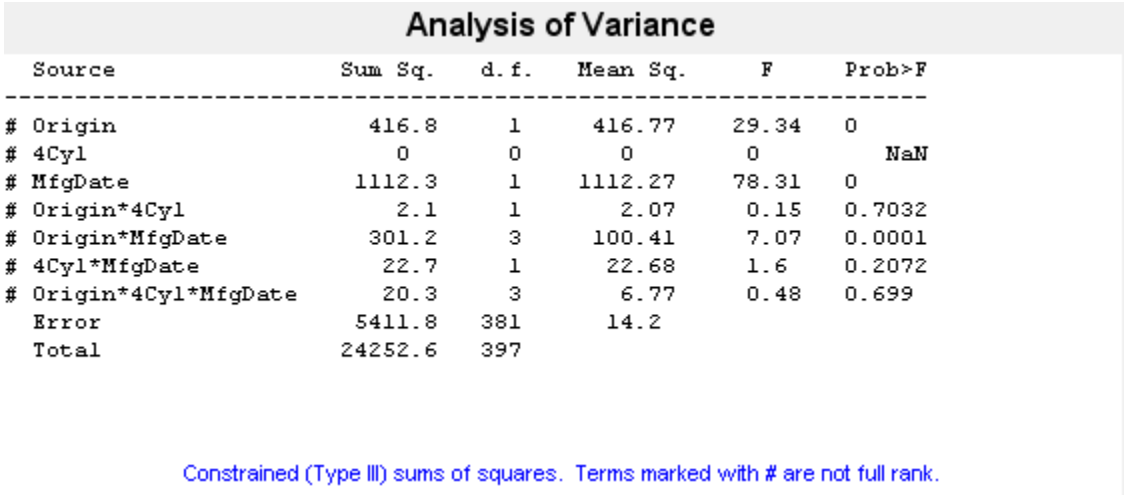
Load the sample data.

```
load carbig
```

The example focusses on four variables. `MPG` is the number of miles per gallon for each of 406 cars (though some have missing values coded as `NaN`). The other three variables are factors: `cyl4` (four-cylinder car or not), `org` (car originated in Europe, Japan, or the USA), and `when` (car was built early in the period, in the middle of the period, or late in the period).

Fit the full model, requesting up to three-way interactions and Type 3 sums-of-squares.

```
varnames = {'Origin';'4Cyl';'MfgDate'};
anovan(MPG,{org cyl4 when},3,3,varnames)
```




```
ans = 7×1

    0.0000
       NaN
```

```
    0.0000
    0.7032
    0.0001
    0.2072
    0.6990
```

Note that many terms are marked by a # symbol as not having full rank, and one of them has zero degrees of freedom and is missing a *p*-value. This can happen when there are missing factor combinations and the model has higher-order terms. In this case, the cross-tabulation below shows that there are no cars made in Europe during the early part of the period with other than four cylinders, as indicated by the 0 in `tbl(2,1,1)`.

```
[tbl,chi2,p,factorvals] = crosstab(org,when,cyl4)
```

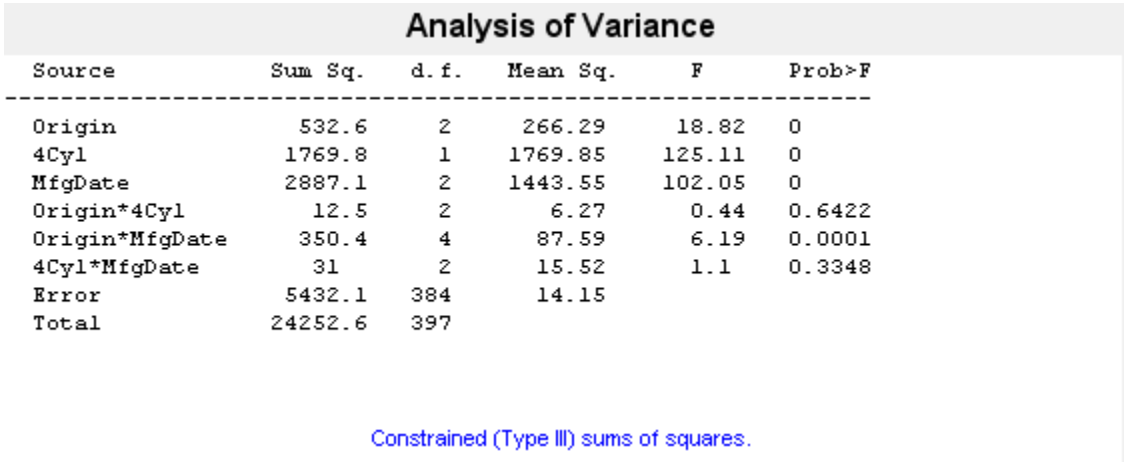
```
tbl =
tbl(:,:,1) =

    82    75    25
     0     4     3
     3     3     4


tbl(:,:,2) =

    12    22    38
    23    26    17
    12    25    32


chi2 = 207.7689

p = 8.0973e-38

factorvals=3×3 cell array
    {'USA'   }    {'Early'}    {'Other'   }
    {'Europe'}    {'Mid'  }    {'Four'    }
    {'Japan' }    {'Late' }    {0x0 double}
```

Consequently it is impossible to estimate the three-way interaction effects, and including the three-way interaction term in the model makes the fit singular.

Using even the limited information available in the ANOVA table, you can see that the three-way interaction has a *p*-value of 0.699, so it is not significant.

Examine only two-way interactions.

```
[p,tbl2,stats,terms] = anovan(MPG,{org cyl4 when},2,3,varnames);
```




```
terms
```

```
terms = 6×3

     1     0     0
     0     1     0
     0     0     1
     1     1     0
     1     0     1
     0     1     1
```

Now all terms are estimable. The *p*-values for interaction term 4 (`Origin*4Cyl`) and interaction term 6 (`4Cyl*MfgDate`) are much larger than a typical cutoff value of 0.05, indicating these terms are not significant. You could choose to omit these terms and pool their effects into the error term. The output `terms` variable returns a matrix of codes, each of which is a bit pattern representing a term.

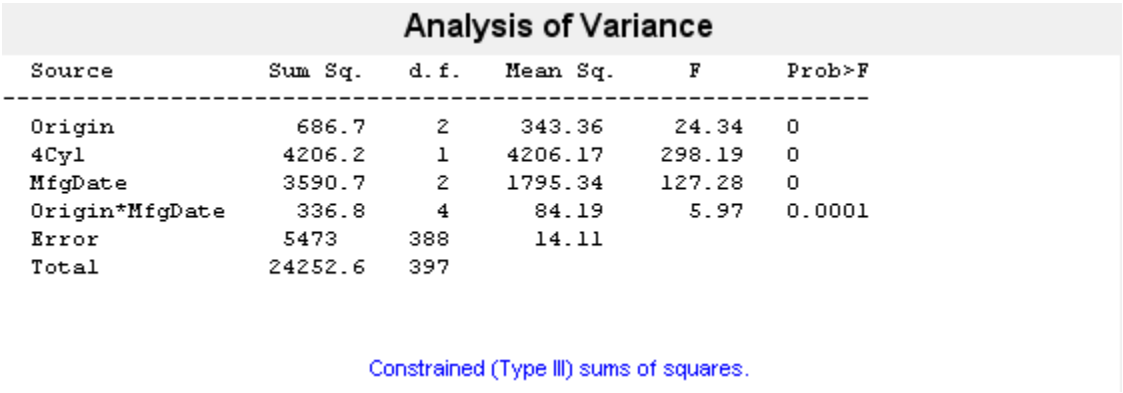
Omit terms from the model by deleting their entries from `terms`.

```
terms([4 6],:) = []
```

```
terms = 4×3

     1     0     0
     0     1     0
     0     0     1
     1     0     1
```

Run `anovan` again, this time supplying the resulting vector as the model argument. Also return the statistics required for multiple comparisons of factors.

```
[~,~,stats] = anovan(MPG,{org cyl4 when},terms,3,varnames)
```

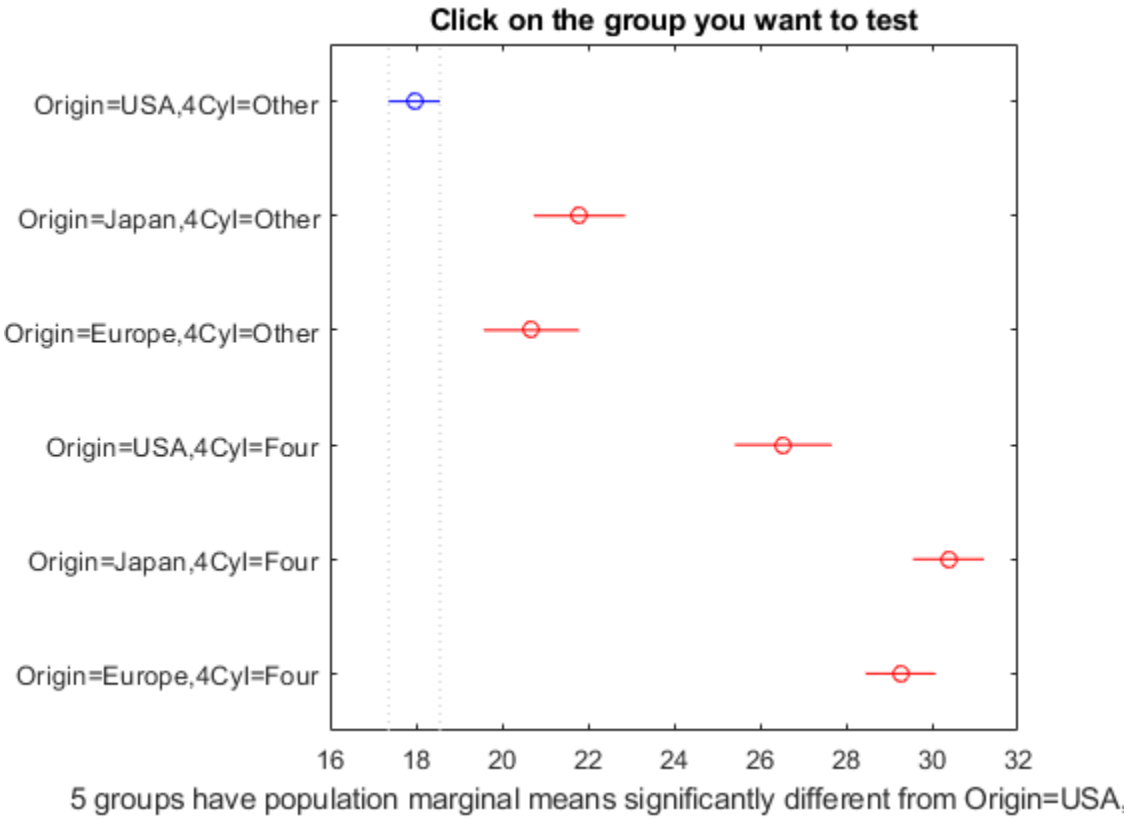



```
stats = struct with fields:
          source: 'anovan'
           resid: [1x406 double]
          coeffs: [18x1 double]
             Rtr: [10x10 double]
        rowbasis: [10x18 double]
             dfe: 388
             mse: 14.1056
     nullproject: [18x10 double]
           terms: [4x3 double]
         nlevels: [3x1 double]
      continuous: [0 0 0]
           vmeans: [3x1 double]
         termcols: [5x1 double]
      coeffnames: {18x1 cell}
            vars: [18x3 double]
        varnames: {3x1 cell}
        grpnames: {3x1 cell}
         vnested: []
             ems: []
           denom: []
         dfdenom: []
         msdenom: []
          varest: []
           varci: []
         txtdenom: []
           txtems: []
          rtnames: []
```

Now you have a more parsimonious model indicating that the mileage of these cars seems to be related to all three factors, and that the effect of the manufacturing date depends on where the car was made.

Perform multiple comparisons for Origin and Cylinder.

```
results = multcompare(stats,'Dimension',[1,2])
```




```
results = 15×6

    1.0000    2.0000   -5.4891   -3.8412   -2.1932    0.0000
    1.0000    3.0000   -4.4146   -2.7251   -1.0356    0.0001
    1.0000    4.0000   -9.9992   -8.5828   -7.1664         0
    1.0000    5.0000  -14.0237  -12.4240  -10.8242         0
    1.0000    6.0000  -12.8980  -11.3080   -9.7180         0
    2.0000    3.0000   -0.7171    1.1160    2.9492    0.5085
    2.0000    4.0000   -7.3655   -4.7417   -2.1179    0.0000
    2.0000    5.0000   -9.9992   -8.5828   -7.1664         0
    2.0000    6.0000   -9.7464   -7.4668   -5.1872    0.0000
    3.0000    4.0000   -8.5396   -5.8577   -3.1757    0.0000
      ⋮
```

See Also

`anova1` | `anovan` | `kruskalwallis` | `multcompare`

Related Examples

- “ANOVA with Random Effects” on page 9-33

More About

- “One-Way ANOVA” on page 9-3

ANOVA with Random Effects

This example shows how to use `anovan` to fit models where a factor's levels represent a random selection from a larger (infinite) set of possible levels.

In an ordinary ANOVA model, each grouping variable represents a fixed factor. The levels of that factor are a fixed set of values. The goal is to determine whether different factor levels lead to different response values.

Set Up the Model

Load the sample data.

```
load mileage
```

The `anova2` function works only with balanced data, and it infers the values of the grouping variables from the row and column numbers of the input matrix. The `anovan` function, on the other hand, requires you to explicitly create vectors of grouping variable values. Create these vectors in the following way.

Create an array indicating the factory for each value in mileage. This array is 1 for the first column, 2 for the second, and 3 for the third.

```
factory  = repmat(1:3,6,1);
```

Create an array indicating the car model for each mileage value. This array is 1 for the first three rows of mileage, and 2 for the remaining three rows.

```
carmod = [ones(3,3); 2*ones(3,3)];
```

Turn these matrices into vectors and display them.

```
mileage = mileage(:);
factory = factory(:);
carmod = carmod(:);
[mileage factory carmod]
```

```
ans = 18×3

   33.3000    1.0000    1.0000
   33.4000    1.0000    1.0000
   32.9000    1.0000    1.0000
   32.6000    1.0000    2.0000
   32.5000    1.0000    2.0000
   33.0000    1.0000    2.0000
   34.5000    2.0000    1.0000
   34.8000    2.0000    1.0000
   33.8000    2.0000    1.0000
   33.4000    2.0000    2.0000
      ⋮
```

Fit a Random Effects Model

Suppose you are studying a few factories but you want information about what would happen if you build these same car models in a different factory, either one that you already have or another that

you might construct. To get this information, fit the analysis of variance model, specifying a model that includes an interaction term and that the factory factor is random.

```
[pvals,tbl,stats] = anovan(mileage, {factory carmod}, ...
'model',2, 'random',1,'varnames',{'Factory' 'Car Model'});
```

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Factory	53.3511	2	26.6756	1333.78	0.0007
Car Model	1.445	1	1.445	72.25	0.0136
Factory*Car Model	0.04	2	0.02	0.18	0.8411
Error	1.3667	12	0.1139		
Total	56.2028	17			

Constrained (Type III) sums of squares.

In the fixed effects version of this fit, which you get by omitting the inputs `'random',1` in the preceding code, the effect of car model is significant, with a *p*-value of 0.0039. But in this example, which takes into account the random variation of the effect of the variable `'Car Model'` from one factory to another, the effect is still significant, but with a higher *p*-value of 0.0136.

F-Statistics for Models with Random Effects

The *F*-statistic in a model having random effects is defined differently than in a model having all fixed effects. In the fixed effects model, you compute the *F*-statistic for any term by taking the ratio of the mean square for that term with the mean square for error. In a random effects model, however, some *F*-statistics use a different mean square in the denominator.

In the example described in **Set Up the Model**, the effect of the variable `'Factory'` could vary across car models. In this case, the interaction mean square takes the place of the error mean square in the *F*-statistic.

Find the *F*-statistic.

```
F = 26.6756 / 0.02
```

```
F = 1.3338e+03
```

The degrees of freedom for the statistic are the degrees of freedom for the numerator (2) and denominator (2) mean squares.

Find the *p*-value.

```
pval = 1 - fcdf(F,2,2)
```

```
pval = 7.4919e-04
```

With random effects, the expected value of each mean square depends not only on the variance of the error term, but also on the variances contributed by the random effects. You can see these dependencies by writing the expected values as linear combinations of contributions from the various model terms.

Find the coefficients of these linear combinations.

```
stats.ems

ans = 4×4

    6.0000    0.0000    3.0000    1.0000
    0.0000    9.0000    3.0000    1.0000
    0.0000    0.0000    3.0000    1.0000
         0         0         0    1.0000
```

This returns the `ems` field of the `stats` structure.

Display text representations of the linear combinations.

```
stats.txtems

ans = 4x1 cell
    {'6*V(Factory)+3*V(Factory*Car Model)+V(Error)'  }
    {'9*Q(Car Model)+3*V(Factory*Car Model)+V(Error)'}
    {'3*V(Factory*Car Model)+V(Error)'               }
    {'V(Error)'                                      }
```

The expected value for the mean square due to car model (second term) includes contributions from a quadratic function of the car model effects, plus three times the variance of the interaction term's effect, plus the variance of the error term. Notice that if the car model effects were all zero, the expression would reduce to the expected mean square for the third term (the interaction term). That is why the *F*-statistic for the car model effect uses the interaction mean square in the denominator.

In some cases there is no single term whose expected value matches the one required for the denominator of the *F*-statistic. In that case, the denominator is a linear combination of mean squares. The stats structure contains fields giving the definitions of the denominators for each *F*-statistic. The `txtdenom` field, `stats.txtdenom`, contains a text representation, and the `denom` field contains a matrix that defines a linear combination of the variances of terms in the model. For balanced models like this one, the `denom` matrix, `stats.denom`, contains zeros and ones, because the denominator is just a single term's mean square.

Display the `txtdenom` field.

```
stats.txtdenom

ans = 3x1 cell
    {'MS(Factory*Car Model)'}
    {'MS(Factory*Car Model)'}
    {'MS(Error)'            }
```

Display the `denom` field.

```
stats.denom

ans = 3×3

    0.0000    1.0000         0
    0.0000    1.0000   -0.0000
   -0.0000    0.0000    1.0000
```

Variance Components

For the model described in **Set Up the Model**, consider the mileage for a particular car of a particular model made at a random factory. The variance of that car is the sum of components, or contributions, one from each of the random terms.

Display the names of the random terms.

```
stats.rtnames
```

```
ans = 3x1 cell
    {'Factory'          }
    {'Factory*Car Model'}
    {'Error'            }
```

You do not know the variances, but you can estimate them from the data. Recall that the `ems` field of the `stats` structure expresses the expected value of each term's mean square as a linear combination of unknown variances for random terms, and unknown quadratic forms for fixed terms. If you take the expected mean square expressions for the random terms, and equate those expected values to the computed mean squares, you get a system of equations that you can solve for the unknown variances. These solutions are the variance component estimates.

Display the variance component estimate for each term.

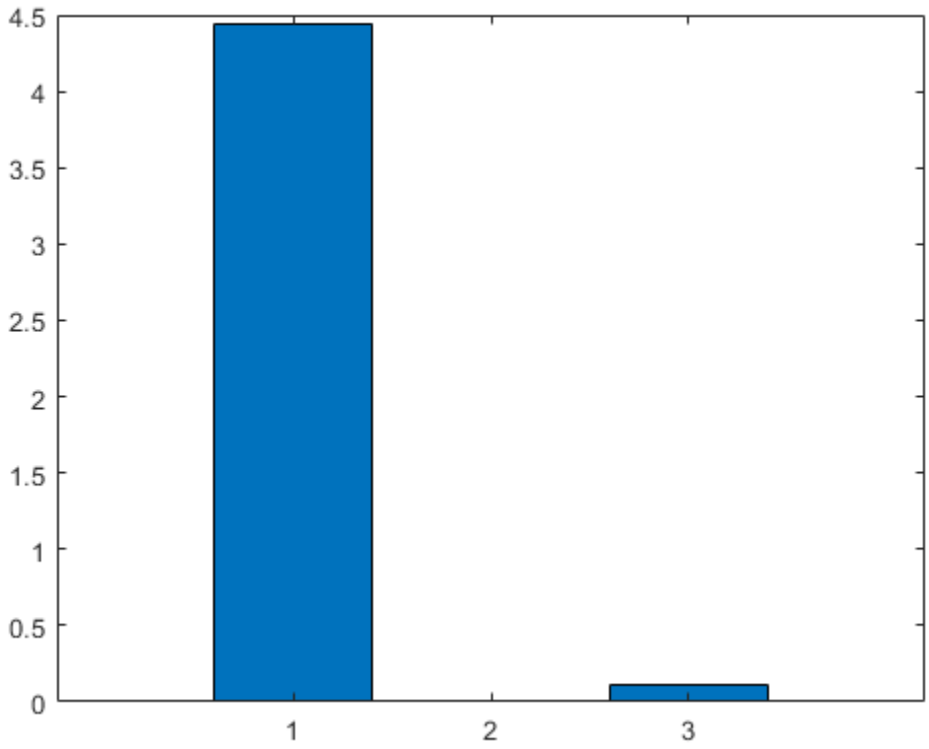
```
stats.varest
```

```
ans = 3×1

    4.4426
   -0.0313
    0.1139
```

Under some conditions, the variability attributed to a term is unusually low, and that term's variance component estimate is negative. In those cases it is common to set the estimate to zero, which you might do, for example, to create a bar graph of the components.

Create a bar graph of the components.

```
bar(max(0,stats.varest))
```




```
gca.xtick = 1:3;
gca.xticklabel = stats.rtnames;
```

You can also compute confidence bounds for the variance estimate. The `anovan` function does this by computing confidence bounds for the variance expected mean squares, and finding lower and upper limits on each variance component containing all of these bounds. This procedure leads to a set of bounds that is conservative for balanced data. (That is, 95% confidence bounds will have a probability of at least 95% of containing the true variances if the number of observations for each combination of grouping variables is the same.) For unbalanced data, these are approximations that are not guaranteed to be conservative.

Display the variance estimates and the confidence limits for the variance estimates of each component.

```
[{'Term' 'Estimate' 'Lower' 'Upper'};
 stats.rtnames, num2cell([stats.varest stats.varci])]
```

```
ans=4×4 cell array
    {'Term'             }    {'Estimate'}    {'Lower' }    {'Upper'   }
    {'Factory'          }    {[  4.4426]}    {[1.0736]}    {[175.6038]}
    {'Factory*Car Model'}    {[ -0.0313]}    {[   NaN]}    {[     NaN]}
    {'Error'            }    {[  0.1139]}    {[0.0586]}    {[  0.3103]}
```

Other ANOVA Models

The `anovan` function also has arguments that enable you to specify two other types of model terms:

- `'nested'` argument specifies a matrix that indicates which factors are nested within other factors. A nested factor is one that takes different values within each level its nested factor.

  Suppose an automobile company has three factories, and each factory makes two car models. The gas mileage in the cars can vary from factory to factory and from model to model. These two factors, factory and model, explain the differences in mileage, that is, the response. One measure of interest is the difference in mileage due to the production methods between factories. Another measure of interest is the difference in the mileage of the models (irrespective of the factory) due to different design specifications. Suppose also that, each factory produces distinct car models for a total of six car models. Then, the car model is nested in factory.

Factory	Car Model
1	1
1	2
2	3
2	4
3	5
3	6

  It is also common with nested models to number the nested factor the same way in each nested factor.
- `'continuous'` argument specifies that some factors are to be treated as continuous variables. The remaining factors are categorical variables. Although the `anovan` function can fit models with multiple continuous and categorical predictors, the simplest model that combines one predictor of each type is known as an *analysis of covariance* model. “Analysis of Covariance” on page 9-39 describes a specialized tool for fitting this model.

See Also

`anova1` | `anova2` | `anovan` | `kruskalwallis` | `multcompare`

Related Examples

- “ANOVA with Random Effects” on page 9-33

More About

- “One-Way ANOVA” on page 9-3
- “Two-Way ANOVA” on page 9-11
- “Multiple Comparisons” on page 9-18
- “N-Way ANOVA” on page 9-25
- “Nonparametric Methods” on page 9-47

Analysis of Covariance

In this section...

Introduction to Analysis of Covariance

Analysis of covariance is a technique for analyzing grouped data having a response (y, the variable to be predicted) and a predictor (x, the variable used to do the prediction). Using analysis of covariance, you can model y as a linear function of x, with the coefficients of the line possibly varying from group to group.

Analysis of Covariance Tool

The `aoctool` function opens an interactive graphical environment for fitting and prediction with analysis of covariance (ANOCOVA) models. It fits the following models for the ith group:

Same mean	$y = \alpha + \varepsilon$
Separate means	$y = (\alpha + \alpha_i) + \varepsilon$
Same line	$y = \alpha + \beta x + \varepsilon$
Parallel lines	$y = (\alpha + \alpha_i) + \beta x + \varepsilon$
Separate lines	$y = (\alpha + \alpha_i) + (\beta + \beta_i)x + \varepsilon$

For example, in the parallel lines model the intercept varies from one group to the next, but the slope is the same for each group. In the same mean model, there is a common intercept and no slope. In order to make the group coefficients well determined, the tool imposes the constraints

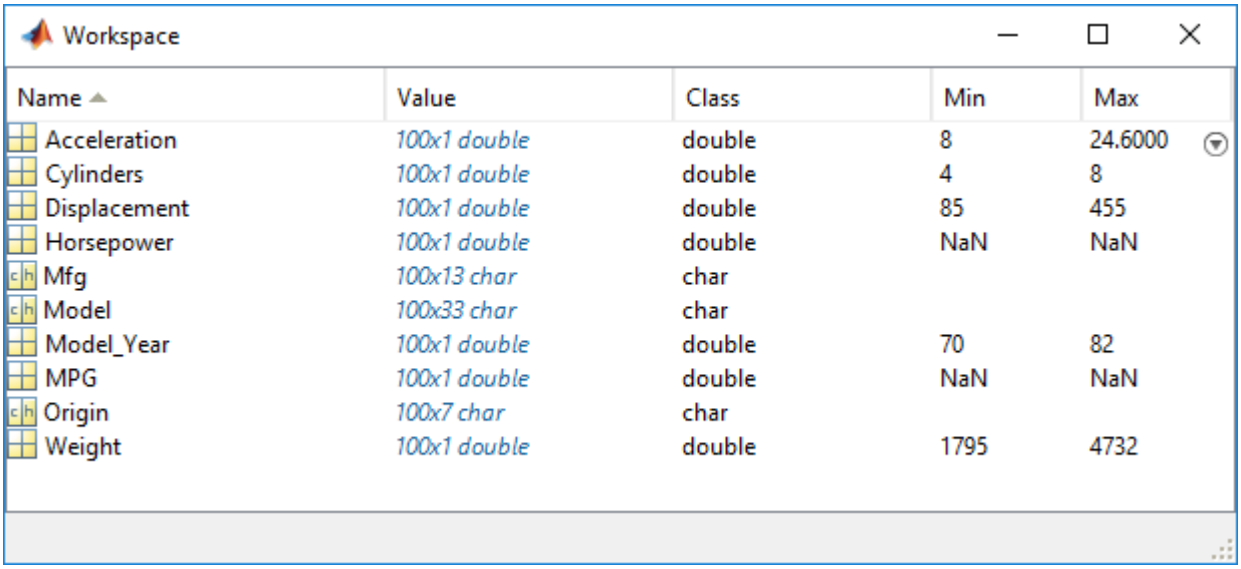
$$\sum \alpha_j = \sum \beta_j = 0$$

The following steps describe the use of `aoctool`.

1. **Load the data.** The Statistics and Machine Learning Toolbox data set `carsmall.mat` contains information on cars from the years 1970, 1976, and 1982. This example studies the relationship between the weight of a car and its mileage, and whether this relationship has changed over the years. To start the demonstration, load the data set.

   ```
   load carsmall
   ```

   The Workspace Browser shows the variables in the data set.



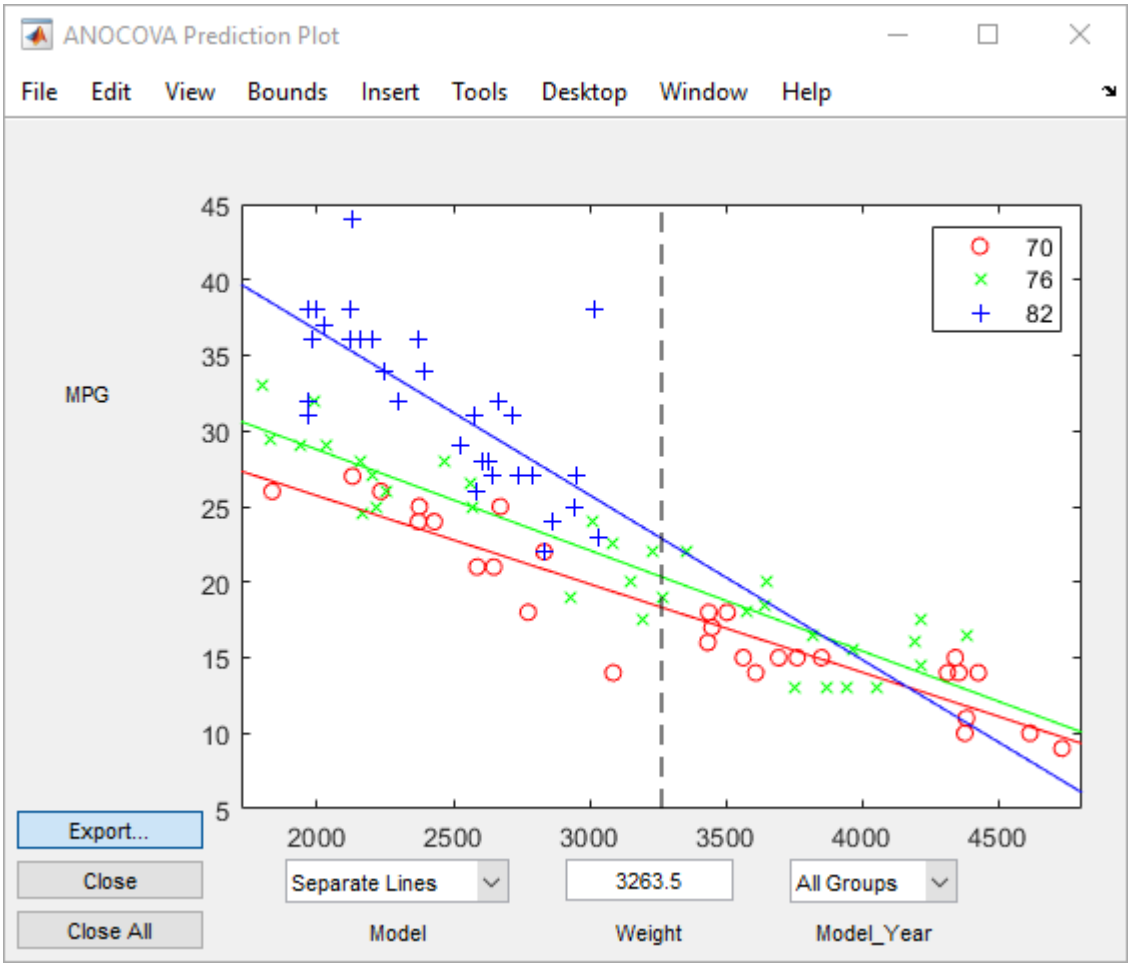
You can also use `aoctool` with your own data.

2 **Start the tool.** The following command calls `aoctool` to fit a separate line to the column vectors `Weight` and `MPG` for each of the three model group defined in `Model_Year`. The initial fit models the *y* variable, `MPG`, as a linear function of the *x* variable, `Weight`.

```
[h,atab,ctab,stats] = aoctool(Weight,MPG,Model_Year);
```

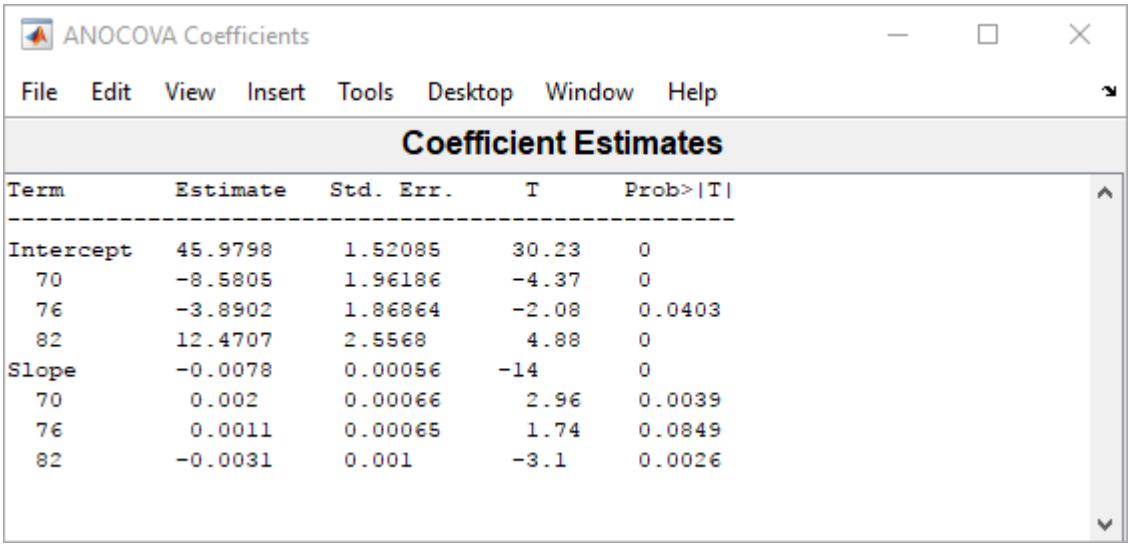
See the `aoctool` function reference page for detailed information about calling `aoctool`.

3 **Examine the output.** The graphical output consists of a main window with a plot, a table of coefficient estimates, and an analysis of variance table. In the plot, each `Model_Year` group has a separate line. The data points for each group are coded with the same color and symbol, and the fit for each group has the same color as the data points.



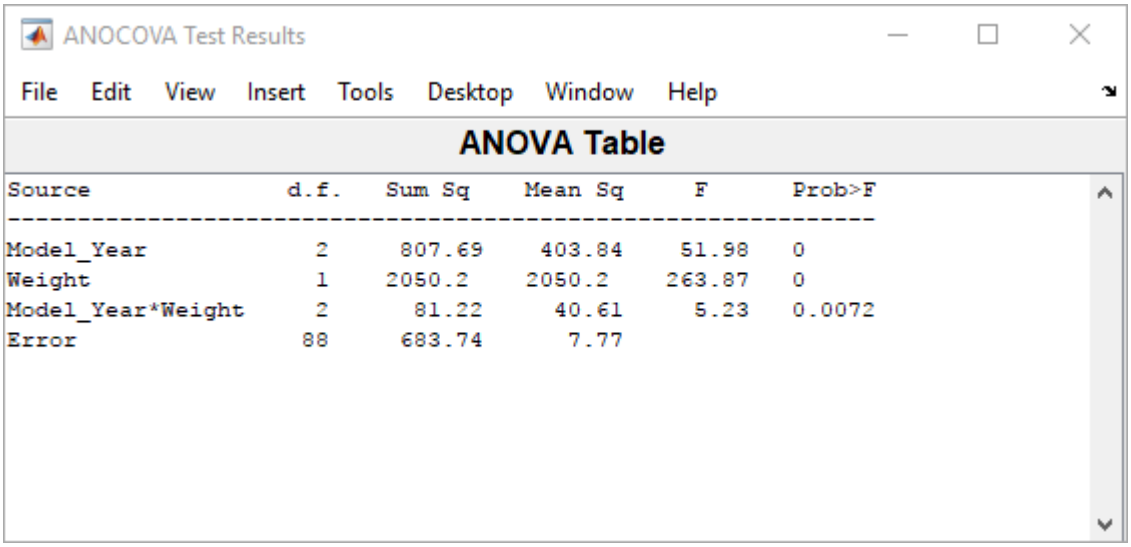
The coefficients of the three lines appear in the figure titled ANOCOVA Coefficients. You can see that the slopes are roughly –0.0078, with a small deviation for each group:

- Model year 1970: $y = (45.9798 - 8.5805) + (-0.0078 + 0.002)x + \varepsilon$
- Model year 1976: $y = (45.9798 - 3.8902) + (-0.0078 + 0.0011)x + \varepsilon$
- Model year 1982: $y = (45.9798 + 12.4707) + (-0.0078 - 0.0031)x + \varepsilon$



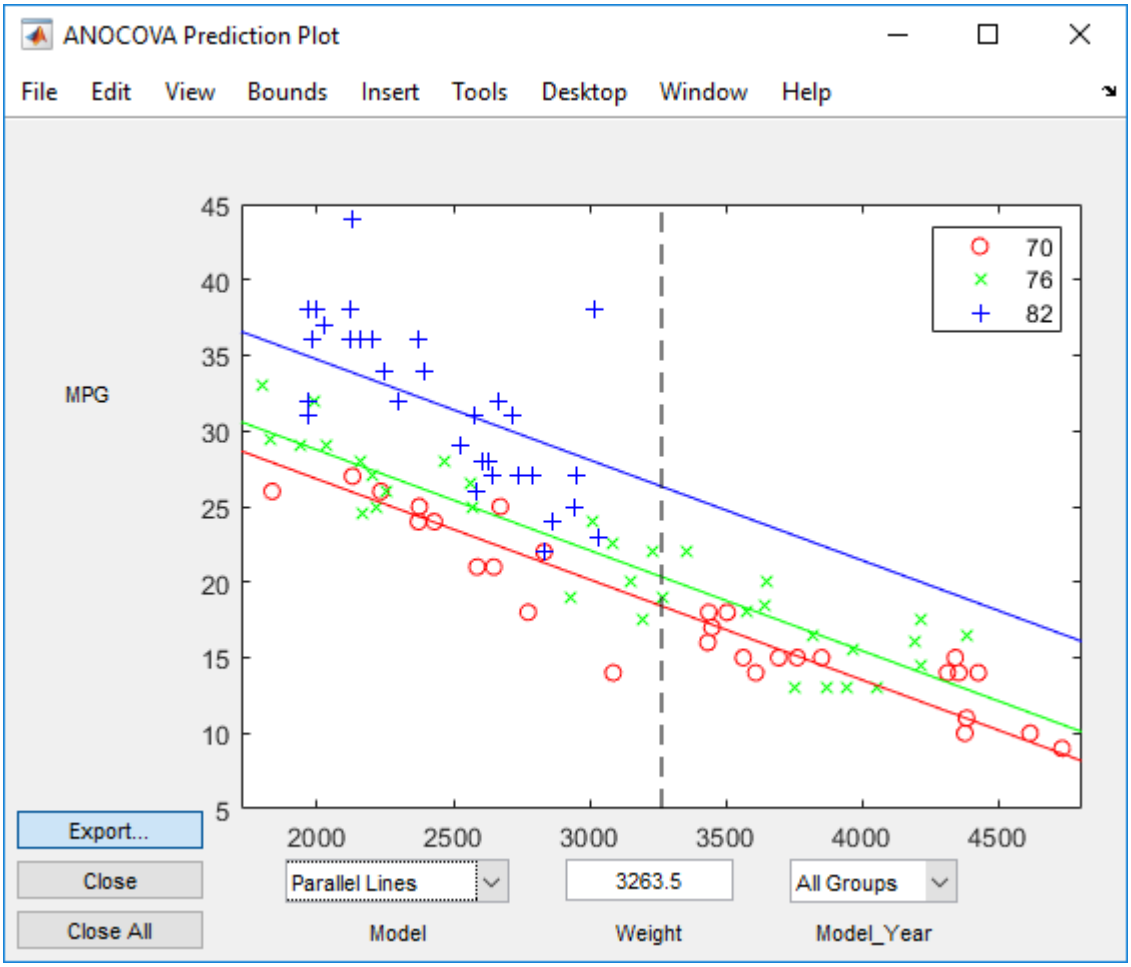
Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	45.9798	1.52085	30.23	0
70	-8.5805	1.96186	-4.37	0
76	-3.8902	1.86864	-2.08	0.0403
82	12.4707	2.5568	4.88	0
Slope	-0.0078	0.00056	-14	0
70	0.002	0.00066	2.96	0.0039
76	0.0011	0.00065	1.74	0.0849
82	-0.0031	0.001	-3.1	0.0026

Because the three fitted lines have slopes that are roughly similar, you may wonder if they really are the same. The `Model_Year*Weight` interaction expresses the difference in slopes, and the ANOVA table shows a test for the significance of this term. With an F statistic of 5.23 and a p value of 0.0072, the slopes are significantly different.



Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
Model_Year	2	807.69	403.84	51.98	0
Weight	1	2050.2	2050.2	263.87	0
Model_Year*Weight	2	81.22	40.61	5.23	0.0072
Error	88	683.74	7.77		

4 **Constrain the slopes to be the same.** To examine the fits when the slopes are constrained to be the same, return to the ANOCOVA Prediction Plot window and use the **Model** pop-up menu to select a `Parallel Lines` model. The window updates to show the following graph.

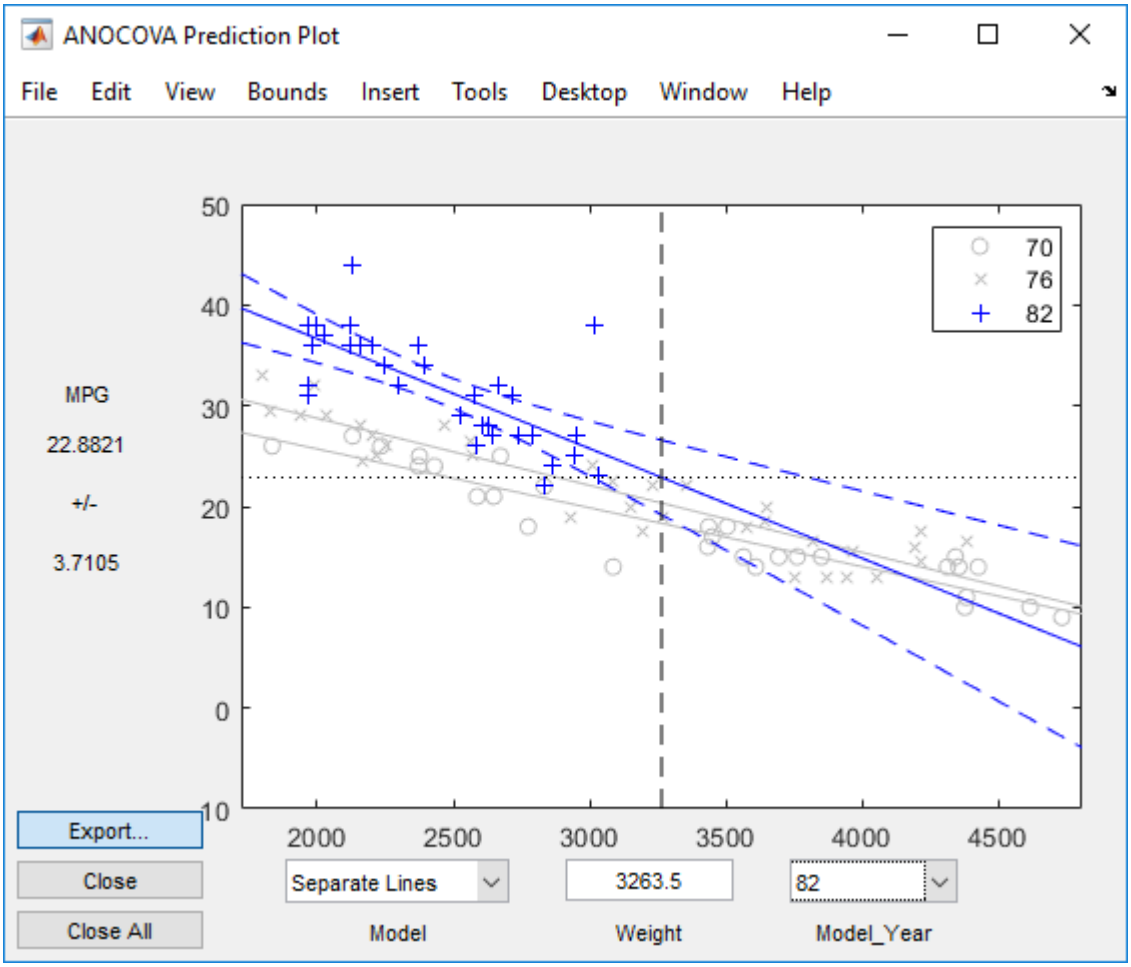


Though this fit looks reasonable, it is significantly worse than the `Separate Lines` model. Use the **Model** pop-up menu again to return to the original model.

Confidence Bounds

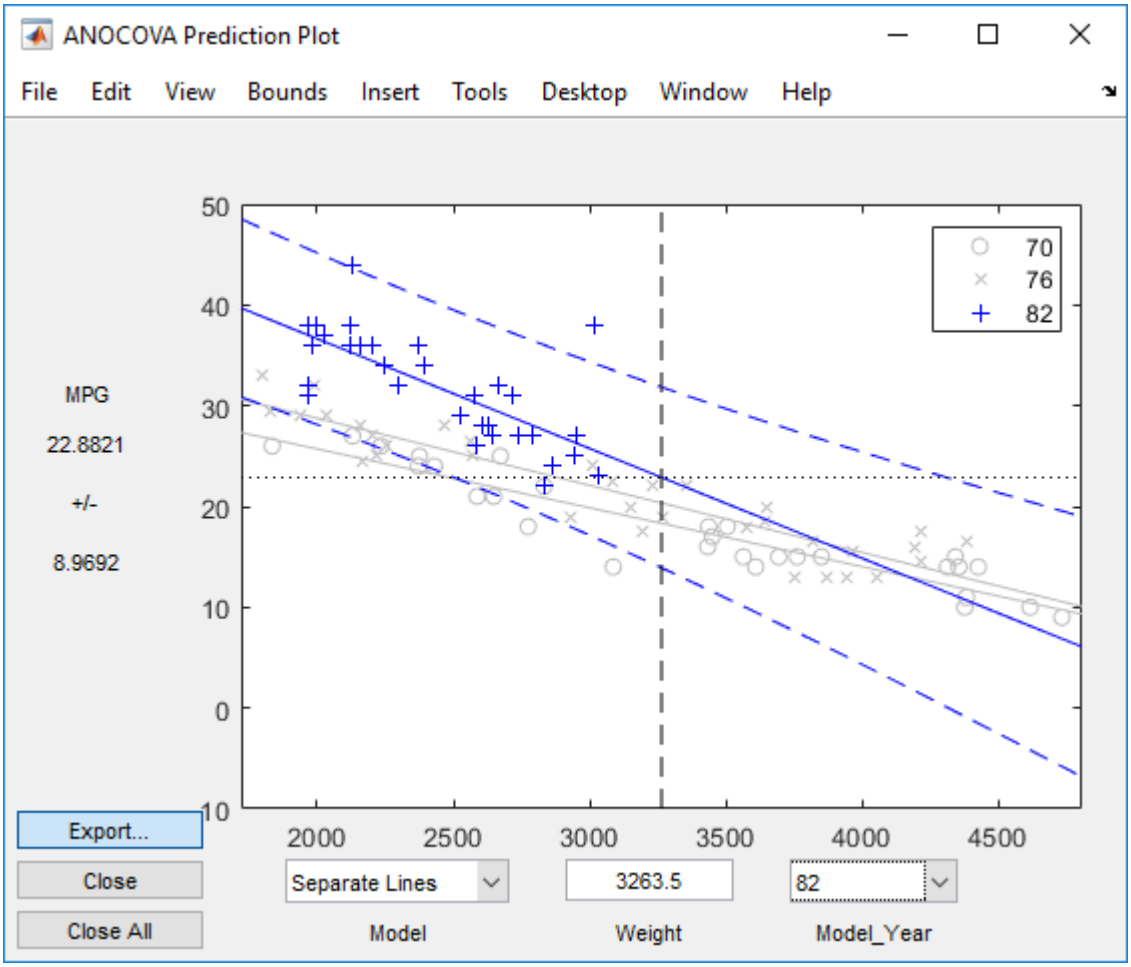
The example in “Analysis of Covariance Tool” on page 9-39 provides estimates of the relationship between `MPG` and `Weight` for each `Model_Year`, but how accurate are these estimates? To find out, you can superimpose confidence bounds on the fits by examining them one group at a time.

1. In the **Model_Year** menu at the lower right of the figure, change the setting from `All Groups` to 82. The data and fits for the other groups are dimmed, and confidence bounds appear around the 82 fit.



The dashed lines form an envelope around the fitted line for model year 82. Under the assumption that the true relationship is linear, these bounds provide a 95% confidence region for the true line. Note that the fits for the other model years are well outside these confidence bounds for `Weight` values between `2000` and `3000`.

2 Sometimes it is more valuable to be able to predict the response value for a new observation, not just estimate the average response value. Use the `aoctool` function **Bounds** menu to change the definition of the confidence bounds from `Line` to `Observation`. The resulting wider intervals reflect the uncertainty in the parameter estimates as well as the randomness of a new observation.



Like the `polytool` function, the `aoctool` function has cross hairs that you can use to manipulate the `Weight` and watch the estimate and confidence bounds along the *y*-axis update. These values appear only when a single group is selected, not when `All Groups` is selected.

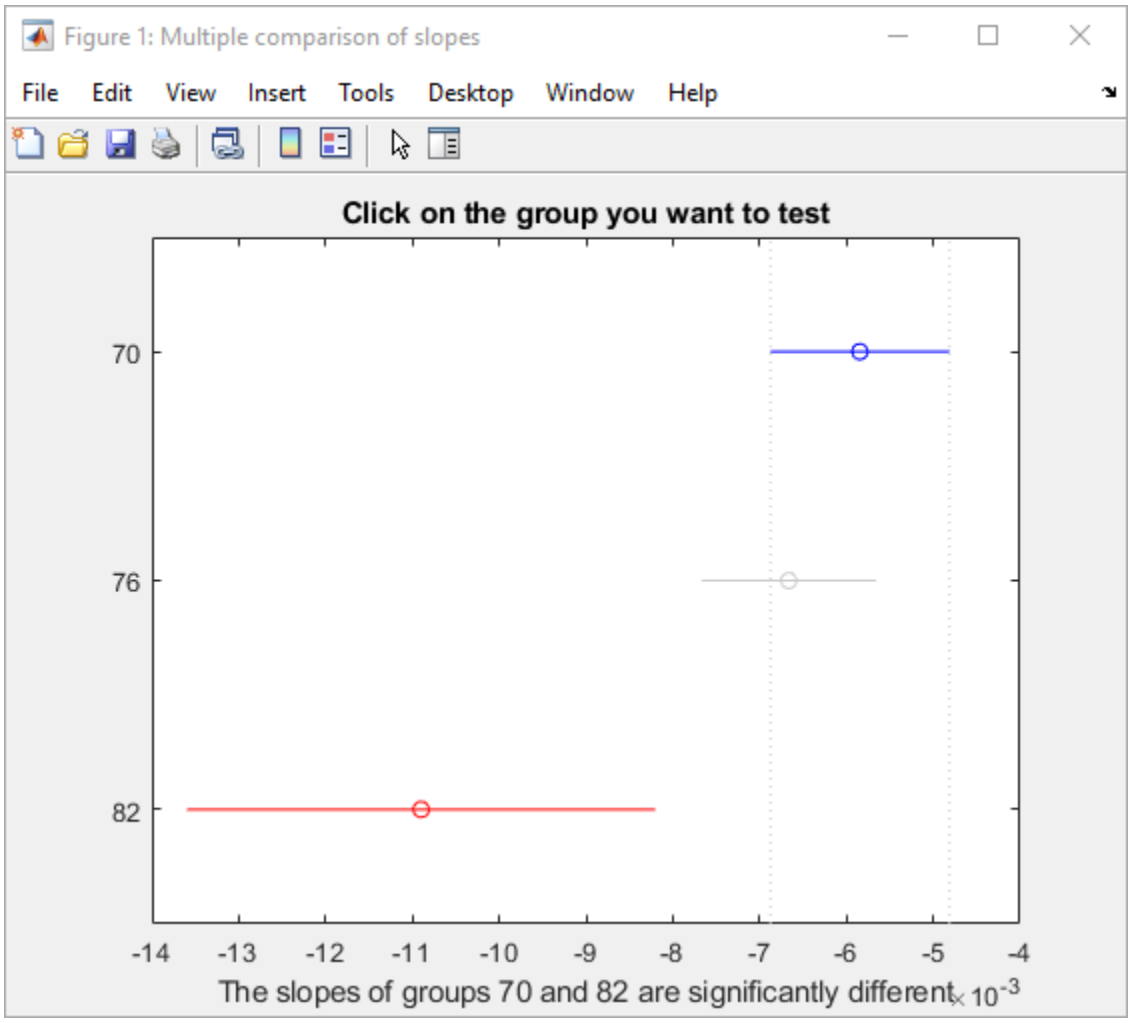
Multiple Comparisons

You can perform a multiple comparison test by using the `stats` output structure from `aoctool` as input to the `multcompare` function. The `multcompare` function can test either slopes, intercepts, or population marginal means (the predicted MPG of the mean weight for each group). The example in “Analysis of Covariance Tool” on page 9-39 shows that the slopes are not all the same, but could it be that two are the same and only the other one is different? You can test that hypothesis.

```
multcompare(stats,0.05,'on','','s')

ans =
    1.0000    2.0000   -0.0012    0.0008    0.0029
    1.0000    3.0000    0.0013    0.0051    0.0088
    2.0000    3.0000    0.0005    0.0042    0.0079
```

This matrix shows that the estimated difference between the intercepts of groups 1 and 2 (1970 and 1976) is 0.0008, and a confidence interval for the difference is [–0.0012, 0.0029]. There is no

significant difference between the two. There are significant differences, however, between the intercept for 1982 and each of the other two. The graph shows the same information.



Note that the `stats` structure was created in the initial call to the `aoctool` function, so it is based on the initial model fit (typically a separate-lines model). If you change the model interactively and want to base your multiple comparisons on the new model, you need to run `aoctool` again to get another `stats` structure, this time specifying your new model as the initial model.

Nonparametric Methods

In this section...

Introduction to Nonparametric Methods

Statistics and Machine Learning Toolbox functions include nonparametric versions of one-way and two-way analysis of variance. Unlike classical tests, nonparametric tests make only mild assumptions about the data, and are appropriate when the distribution of the data is non-normal. On the other hand, they are less powerful than classical methods for normally distributed data.

Both of the nonparametric functions described here will return a `stats` structure that can be used as an input to the `multcompare` function for multiple comparisons.

Kruskal-Wallis Test

The example “Perform One-Way ANOVA” on page 9-5 uses one-way analysis of variance to determine if the bacteria counts of milk varied from shipment to shipment. The one-way analysis rests on the assumption that the measurements are independent, and that each has a normal distribution with a common variance and with a mean that was constant in each column. You can conclude that the column means were not all the same. The following example repeats that analysis using a nonparametric procedure.

The Kruskal-Wallis test is a nonparametric version of one-way analysis of variance. The assumption behind this test is that the measurements come from a continuous distribution, but not necessarily a normal distribution. The test is based on an analysis of variance using the ranks of the data values, not the data values themselves. Output includes a table similar to an ANOVA table, and a box plot.

You can run this test as follows:

```
load hogg

p = kruskalwallis(hogg)
p =
    0.0020
```

The low *p* value means the Kruskal-Wallis test results agree with the one-way analysis of variance results.

Friedman's Test

“Perform Two-Way ANOVA” on page 9-13 uses two-way analysis of variance to study the effect of car model and factory on car mileage. The example tests whether either of these factors has a significant effect on mileage, and whether there is an interaction between these factors. The conclusion of the example is there is no interaction, but that each individual factor has a significant effect. The next example examines whether a nonparametric analysis leads to the same conclusion.

Friedman's test is a nonparametric test for data having a two-way layout (data grouped by two categorical factors). Unlike two-way analysis of variance, Friedman's test does not treat the two factors symmetrically and it does not test for an interaction between them. Instead, it is a test for whether the columns are different after adjusting for possible row differences. The test is based on an analysis of variance using the ranks of the data across categories of the row factor. Output includes a table similar to an ANOVA table.

You can run Friedman's test as follows.

```
load mileage
p = friedman(mileage,3)
p =
  7.4659e-004
```

Recall the classical analysis of variance gave a p value to test column effects, row effects, and interaction effects. This p value is for column effects. Using either this p value or the p value from ANOVA ($p < 0.0001$), you conclude that there are significant column effects.

In order to test for row effects, you need to rearrange the data to swap the roles of the rows in columns. For a data matrix `x` with no replications, you could simply transpose the data and type

```
p = friedman(x')
```

With replicated data it is slightly more complicated. A simple way is to transform the matrix into a three-dimensional array with the first dimension representing the replicates, swapping the other two dimensions, and restoring the two-dimensional shape.

```
x = reshape(mileage, [3 2 3]);
x = permute(x,[1 3 2]);
x = reshape(x,[9 2])
x =
   33.3000   32.6000
   33.4000   32.5000
   32.9000   33.0000
   34.5000   33.4000
   34.8000   33.7000
   33.8000   33.9000
   37.4000   36.6000
   36.8000   37.0000
   37.6000   36.7000

friedman(x,3)
ans =
    0.0082
```

Again, the conclusion is similar to that of the classical analysis of variance. Both this p value and the one from ANOVA ($p = 0.0039$) lead you to conclude that there are significant row effects.

You cannot use Friedman's test to test for interactions between the row and column factors.

MANOVA

In this section...

Introduction to MANOVA

The analysis of variance technique in “Perform One-Way ANOVA” on page 9-5 takes a set of grouped data and determine whether the mean of a variable differs significantly among groups. Often there are multiple response variables, and you are interested in determining whether the entire set of means is different from one group to the next. There is a multivariate version of analysis of variance that can address the problem.

ANOVA with Multiple Responses

The `carsmall` data set has measurements on a variety of car models from the years 1970, 1976, and 1982. Suppose you are interested in whether the characteristics of the cars have changed over time.

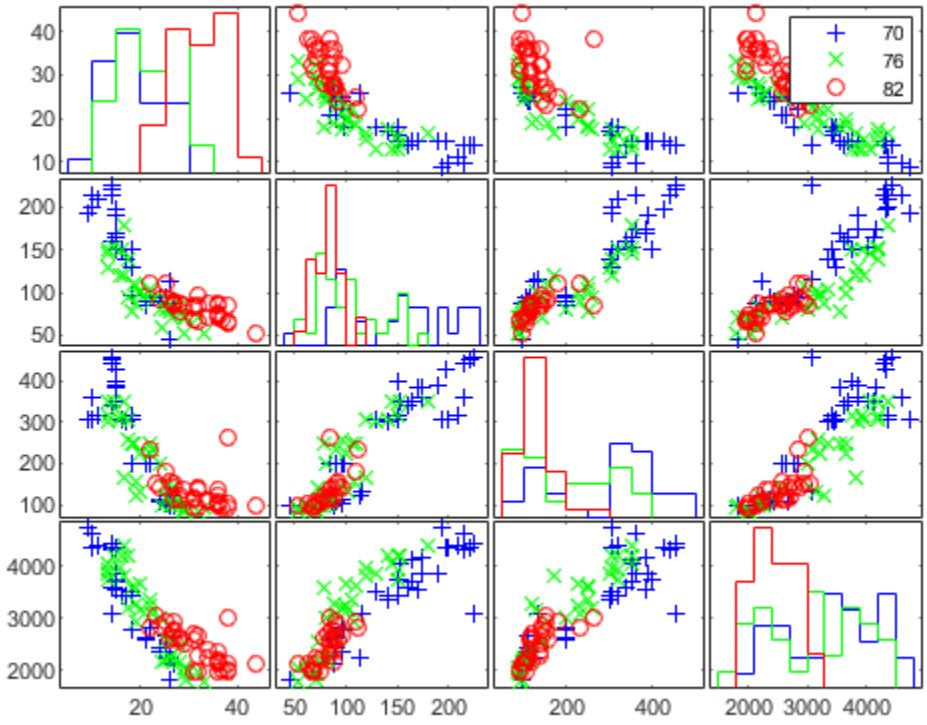
```
load carsmall
whos
```

```
  Name                Size            Bytes  Class     Attributes

  Acceleration      100x1               800  double
  Cylinders         100x1               800  double
  Displacement      100x1               800  double
  Horsepower        100x1               800  double
  MPG               100x1               800  double
  Mfg               100x13             2600  char
  Model             100x33             6600  char
  Model_Year        100x1               800  double
  Origin            100x7              1400  char
  Weight            100x1               800  double
```

Four of these variables (`Acceleration`, `Displacement`, `Horsepower`, and `MPG`) are continuous measurements on individual car models. The variable `Model_Year` indicates the year in which the car was made. You can create a grouped plot matrix of these variables using the `gplotmatrix` function.

Create a grouped plot matrix of these variables using the `gplotmatrix` function.

```
x = [MPG Horsepower Displacement Weight];
gplotmatrix(x,[],Model_Year,[],'+xo')
```



(When the second argument of `gplotmatrix` is empty, the function graphs the columns of the `x` argument against each other, and places histograms along the diagonals. The empty fourth argument produces a graph with the default colors. The fifth argument controls the symbols used to distinguish between groups.)

It appears the cars do differ from year to year. The upper right plot, for example, is a graph of `MPG` versus `Weight`. The 1982 cars appear to have higher mileage than the older cars, and they appear to weigh less on average. But as a group, are the three years significantly different from one another? The `manova1` function can answer that question.

```
[d,p,stats] = manova1(x,Model_Year)
```

```
d = 2

p = 2×1
10^-6 ×

    0.0000
    0.1141


stats = struct with fields:
           W: [4x4 double]
           B: [4x4 double]
           T: [4x4 double]
         dfW: 90
         dfB: 2
```

```
       dfT: 92
    lambda: [2x1 double]
     chisq: [2x1 double]
   chisqdf: [2x1 double]
  eigenval: [4x1 double]
  eigenvec: [4x4 double]
     canon: [100x4 double]
     mdist: [1x100 double]
    gmdist: [3x3 double]
    gnames: {3x1 cell}
```

The `manova1` function produces three outputs:

- The first output `d` is an estimate of the dimension of the group means. If the means were all the same, the dimension would be 0, indicating that the means are at the same point. If the means differed but fell along a line, the dimension would be 1. In the example the dimension is 2, indicating that the group means fall in a plane but not along a line. This is the largest possible dimension for the means of three groups.
- The second output `p` is a vector of *p*-values for a sequence of tests. The first *p*-value tests whether the dimension is 0, the next whether the dimension is 1, and so on. In this case both *p*-values are small. That's why the estimated dimension is 2.
- The third output `stats` is a structure containing several fields, described in the following section.

Fields of the `stats` Structure

The `W`, `B`, and `T` fields are matrix analogs to the within, between, and total sums of squares in ordinary one-way analysis of variance. The next three fields are the degrees of freedom for these matrices. Fields `lambda`, `chisq`, and `chisqdf` are the ingredients of the test for the dimensionality of the group means. (The *p*-values for these tests are the first output argument of `manova1`.)

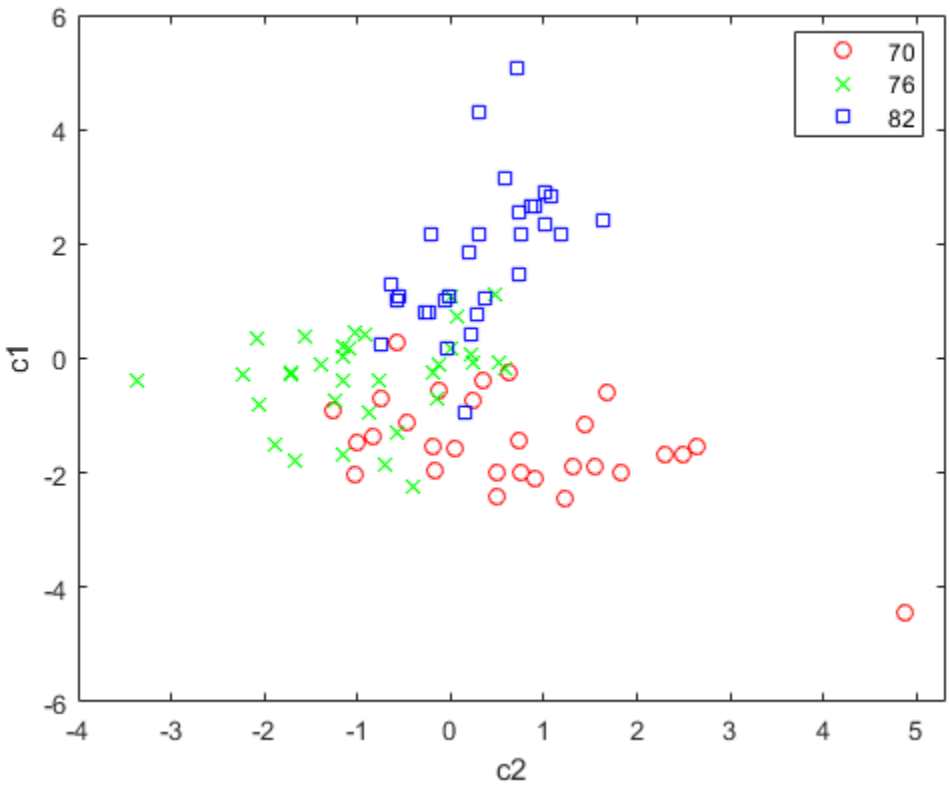
The next three fields are used to do a canonical analysis. Recall that in “Principal Component Analysis (PCA)” on page 15-68 you look for the combination of the original variables that has the largest possible variation. In multivariate analysis of variance, you instead look for the linear combination of the original variables that has the largest separation between groups. It is the single variable that would give the most significant result in a univariate one-way analysis of variance. Having found that combination, you next look for the combination with the second highest separation, and so on.

The `eigenvec` field is a matrix that defines the coefficients of the linear combinations of the original variables. The `eigenval` field is a vector measuring the ratio of the between-group variance to the within-group variance for the corresponding linear combination. The `canon` field is a matrix of the canonical variable values. Each column is a linear combination of the mean-centered original variables, using coefficients from the `eigenvec` matrix.

```
c1 = stats.canon(:,1);
c2 = stats.canon(:,2);
```

Plot the grouped scatter plot of the first two canonical variables.

```
figure
gscatter(c2,c1,Model_Year,[],'oxs')
```



A grouped scatter plot of the first two canonical variables shows more separation between groups then a grouped scatter plot of any pair of original variables. In this example, it shows three clouds of points, overlapping but with distinct centers. One point in the bottom right sits apart from the others. You can mark this point on the plot using the `gname` function.

Roughly speaking, the first canonical variable, `c1`, separates the 1982 cars (which have high values of `c1`) from the older cars. The second canonical variable, `c2`, reveals some separation between the 1970 and 1976 cars.

The final two fields of the `stats` structure are Mahalanobis distances. The `mdist` field measures the distance from each point to its group mean. Points with large values may be outliers. In this data set, the largest outlier is the one in the scatter plot, the Buick Estate station wagon. (Note that you could have supplied the model name to the `gname` function above if you wanted to label the point with its model name rather than its row number.)

Find the largest distance from the group mean.

```
max(stats.mdist)
```

```
ans = 31.5273
```

Find the point that has the largest distance from the group mean.

```
find(stats.mdist == ans)
```

```
ans = 20
```

Find the car model that corresponds to the largest distance from the group mean.

```
Model(20,:)
```

```
ans =
'buick estate wagon (sw)          '
```

The `gmdist` field measures the distances between each pair of group means. Examine the group means using `grpstats`.

```
grpstats(x, Model_Year)
```

```
ans = 3×4
10^3 ×

    0.0177    0.1489    0.2869    3.4413
    0.0216    0.1011    0.1978    3.0787
    0.0317    0.0815    0.1289    2.4535
```

Find the distances between the each pair of group means.

```
stats.gmdist
```

```
ans = 3×3

         0    3.8277   11.1106
    3.8277         0    6.1374
   11.1106    6.1374         0
```

As might be expected, the multivariate distance between the extreme years 1970 and 1982 (11.1) is larger than the difference between more closely spaced years (3.8 and 6.1). This is consistent with the scatter plots, where the points seem to follow a progression as the year changes from 1970 through 1976 to 1982. If you had more groups, you might find it instructive to use the `manovacluster` function to draw a diagram that presents clusters of the groups, formed using the distances between their means.

See Also

`gname` | `gplotmatrix` | `manova1` | `manovacluster`

Model Specification for Repeated Measures Models

Model specification for a repeated measures model is a character vector or string scalar representing a formula in the form

```
'y1-yk ~ terms',
```

where the responses and terms are in Wilkinson notation.

For example, if you have five repeated measures $y1$, $y2$, $y3$, $y4$, and $y5$, and you include the terms $X1$, $X2$, $X3$, $X4$, and $X3$:$X4$ in your linear model, then you can specify `modelspec` as follows:

```
'y1-y5 ~ X1 + X2 + X3*X4'.
```

Wilkinson Notation

Wilkinson notation describes the factors present in models. It does not describe the multipliers (coefficients) of those factors.

Use these rules to specify the responses in `modelspec`.

Wilkinson Notation	Description
`Y1,Y2,Y3`	Specific list of variables
`Y1-Y5`	All table variables from `Y1` through `Y5`

The following rules are for specifying terms in `modelspec`.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	X, X^2, ..., X^k
`X1 + X2`	X1, X2
`X1*X2`	X1, X2, X1*X2
`X1:X2`	X1*X2 only
`-X2`	Do not include X2
`X1*X2 + X3`	X1, X2, X3, X1*X2
`X1 + X2 + X3 + X1:X2`	X1, X2, X3, X1*X2
`X1*X2*X3 - X1:X2:X3`	X1, X2, X3, X1*X2, X1*X3, X2*X3
`X1*(X2 + X3)`	X1, X2, X3, X1*X2, X1*X3

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`.

See Also

`fitrm`

Compound Symmetry Assumption and Epsilon Corrections

The regular *p*-value calculations in the repeated measures anova (`ranova`) are accurate if the theoretical distribution of the response variables has compound symmetry. This means that all response variables have the same variance, and each pair of response variables share a common correlation. That is,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Under the compound symmetry assumption, the *F*-statistics in the repeated measures anova table have an *F*-distribution with degrees of freedom (v_1, v_2). Here, v_1 is the rank of the contrast being tested, and v_2 is the degrees of freedom for error. If the compound symmetry assumption is not true, the *F*-statistic has an approximate *F*-distribution with degrees of freedom (εv_1, εv_2), where ε is the correction factor. Then, the *p*-value must be computed using the adjusted values. The three different correction factor computations are as follows:

- **Greenhouse-Geisser approximation**

  $$\varepsilon_{GG} = \frac{\left(\sum_{i=1}^{p} \lambda_i\right)^2}{d \sum_{i=1}^{p} \lambda_i^2},$$

  where λ_i i = 1, 2, .., p are the eigenvalues of the covariance matrix. p is the number of variables, and d is equal to p-1.

- **Huynh-Feldt approximation**

  $$\varepsilon_{HF} = \min\left(1, \frac{nd\varepsilon_{GG} - 2}{d(n - rx) - d^2\varepsilon_{GG}}\right),$$

  where n is the number of rows in the design matrix and r is the rank of the design matrix.

- **Lower bound on the true *p*-value**

  $$\varepsilon_{LB} = \frac{1}{d}.$$

References

[1] Huynh, H., and L. S. Feldt. “Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs.” *Journal of Educational Statistics*. Vol. 1, 1976, pp. 69–82.

[2] Greenhouse, S. W., and S. Geisser. “An Extension of Box’s Result on the Use of *F*-Distribution in Multivariate Analysis.” *Annals of Mathematical Statistics*. Vol. 29, 1958, pp. 885–891.

See Also

`epsilon` | `mauchly` | `ranova`

More About

- “Mauchly’s Test of Sphericity” on page 9-57

Mauchly’s Test of Sphericity

The regular *p*-value calculations in the repeated measures anova (`ranova`) are accurate if the theoretical distribution of the response variables have compound symmetry. This means that all response variables have the same variance, and each pair of response variables share a common correlation. That is,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

If the compound symmetry assumption is false, then the degrees of freedom for the repeated measures anova test must be adjusted by a factor ε, and the *p*-value must be computed using the adjusted values.

Compound symmetry implies sphericity.

For a repeated measures model with responses $y1$, $y2$, ..., sphericity means that all pair-wise differences $y1 - y2$, $y1 - y3$, ... have the same theoretical variance. Mauchly’s test is the most accepted test for sphericity.

Mauchly’s W statistic is

$$W = \frac{|T|}{(trace(T)/p)^d},$$

where

$$T = M'\widehat{\sum}M.$$

M is a p-by-d orthogonal contrast matrix, Σ is the covariance matrix, p is the number of variables, and $d = p - 1$.

A chi-square test statistic assesses the significance of W. If n is the number of rows in the design matrix, and r is the rank of the design matrix, then the chi-square statistic is

$$C = -(n - r)\log(W)D,$$

where

$$D = 1 - \frac{2d^2 + d + 2}{6d(n - r)}.$$

The C test statistic has a chi-square distribution with $(p(p - 1)/2) - 1$ degrees of freedom. A small *p*-value for the Mauchly’s test indicates that the sphericity assumption does not hold.

The `rmanova` method computes the *p*-values for the repeated measures anova based on the results of the Mauchly’s test and each epsilon value.

References

[1] Mauchly, J. W. “Significance Test for Sphericity of a Normal *n*-Variate Distribution. *The Annals of Mathematical Statistics*. Vol. 11, 1940, pp. 204–209.

See Also

`epsilon` | `mauchly` | `ranova`

More About

- “Compound Symmetry Assumption and Epsilon Corrections” on page 9-55

Multivariate Analysis of Variance for Repeated Measures

Multivariate analysis of variance analysis is a test of the form `A*B*C = D`, where `B` is the *p*-by-*r* matrix of coefficients. *p* is the number of terms, such as the constant, linear predictors, dummy variables for categorical predictors, and products and powers, *r* is the number of repeated measures, and *n* is the number of subjects. `A` is an *a*-by-*p* matrix, with rank $a \le p$, defining hypotheses based on the between-subjects model. `C` is an *r*-by-*c* matrix, with rank $c \le r \le n - p$, defining hypotheses based on the within-subjects model, and `D` is an *a*-by-*c* matrix, containing the hypothesized value.

`manova` tests if the model terms are significant in their effect on the response by measuring how they contribute to the overall covariance. It includes all terms in the between-subjects model. `manova` always takes `D` as zero. The multivariate response for each observation (subject) is the vector of repeated measures.

`manova` uses four different methods to measure these contributions: Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, Roy's maximum root statistic. Define

$$T = A\widehat{B}C - D,$$
$$Z = A(X'X)^{-1}A'.$$

Then, the hypotheses sum of squares and products matrix is

$$Q_h = T'Z^{-1}T,$$

and the residuals sum of squares and products matrix is

$$Q_e = C'(R'R)C,$$

where

$$R = Y - X\widehat{B}.$$

The matrix Q_h is analogous to the numerator of a univariate *F*-test, and Q_e is analogous to the error sum of squares. Hence, the four statistics `manova` uses are:

- **Wilks' lambda**

  $$\Lambda = \frac{|Q_e|}{|Q_h + Q_e|} = \prod \frac{1}{1 + \lambda_i},$$

  where λ_i are the solutions of the characteristic equation $|Q_h - \lambda Q_e| = 0$.
- **Pillai's trace**

  $$V = trace\left(Q_h(Q_h + Q_e)^{-1}\right) = \sum \theta_i,$$

  where θ_i values are the solutions of the characteristic equation $Q_h - \theta(Q_h + Q_e) = 0$.
- **Hotelling-Lawley trace**

  $$U = trace\left(Q_h Q_e^{-1}\right) = \sum \lambda_i.$$
- **Roy's maximum root statistic**

  $$\Theta = \max\left(eig\left(Q_h Q_e^{-1}\right)\right).$$

References

[1] Charles, S. D. *Statistical Methods for the Analysis of Repeated Measurements*. Springer Texts in Statistics. Springer-Verlag, New York, Inc., 2002.

See Also

`coeftest` | `manova`

Bayesian Optimization

Bayesian Optimization Algorithm

In this section...

Algorithm Outline

The Bayesian optimization algorithm attempts to minimize a scalar objective function $f(x)$ for x in a bounded domain. The function can be deterministic or stochastic, meaning it can return different results when evaluated at the same point x. The components of x can be continuous reals, integers, or categorical, meaning a discrete set of names.

Note Throughout this discussion, D represents the number of components of x.

The key elements in the minimization are:

- A Gaussian process model of $f(x)$.
- A Bayesian update procedure for modifying the Gaussian process model at each new evaluation of $f(x)$.
- An acquisition function $a(x)$ (based on the Gaussian process model of f) that you maximize to determine the next point x for evaluation. For details, see “Acquisition Function Types” on page 10-3 and “Acquisition Function Maximization” on page 10-5.

Algorithm outline:

- Evaluate $y_i = f(x_i)$ for `NumSeedPoints` points x_i, taken at random within the variable bounds. `NumSeedPoints` is a `bayesopt` setting. If there are evaluation errors, take more random points until there are `NumSeedPoints` successful evaluations. The probability distribution of each component is either uniform or log-scaled, depending on the `Transform` value in `optimizableVariable`.

Then repeat the following steps:

1. Update the Gaussian process model of $f(x)$ to obtain a posterior distribution over functions $Q(f|x_i, y_i \text{ for } i = 1,\ldots,t)$. (Internally, `bayesopt` uses `fitrgp` to fit a Gaussian process model to the data.)
2. Find the new point x that maximizes the acquisition function $a(x)$.

The algorithm stops after reaching any of the following:

- A fixed number of iterations (default 30).
- A fixed time (default is no time limit).
- A stopping criterion that you supply in “Bayesian Optimization Output Functions” on page 10-19 or “Bayesian Optimization Plot Functions” on page 10-11.

For the algorithmic differences in parallel, see “Parallel Bayesian Algorithm” on page 10-7.

Gaussian Process Regression for Fitting the Model

The underlying probabilistic model for the objective function f is a Gaussian process prior with added Gaussian noise in the observations. So the prior distribution on $f(x)$ is a Gaussian process with mean $\mu(x;\theta)$ and covariance kernel function $k(x,x';\theta)$. Here, θ is a vector of kernel parameters. For the particular kernel function `bayesopt` uses, see “Kernel Function” on page 10-3.

In a bit more detail, denote a set of points $X = x_i$ with associated objective function values $F = f_i$. The prior’s joint distribution of the function values F is multivariate normal, with mean $\mu(X)$ and covariance matrix $K(X,X)$, where $K_{ij} = k(x_i,x_j)$.

Without loss of generality, the prior mean is given as `0`.

Also, the observations are assumed to have added Gaussian noise with variance σ^2. So the prior distribution has covariance $K(X,X;\theta) + \sigma^2 I$.

Fitting a Gaussian process regression model to observations consists of finding values for the noise variance σ^2 and kernel parameters θ. This fitting is a computationally intensive process performed by `fitrgp`.

For details on fitting a Gaussian process to observations, see “Gaussian Process Regression”.

Kernel Function

The kernel function $k(x,x';\theta)$ can significantly affect the quality of a Gaussian process regression. `bayesopt` uses the ARD Matérn 5/2 kernel defined in “Kernel (Covariance) Function Options” on page 6-6.

See Snoek, Larochelle, and Adams [3].

Acquisition Function Types

Six choices of acquisition functions are available for `bayesopt`. There are three basic types, with `expected-improvement` also modified by `per-second` or `plus`:

- `'expected-improvement-per-second-plus'` (default)
- `'expected-improvement'`
- `'expected-improvement-plus'`
- `'expected-improvement-per-second'`
- `'lower-confidence-bound'`
- `'probability-of-improvement'`

The acquisition functions evaluate the “goodness” of a point x based on the posterior distribution function Q. When there are coupled constraints, including the Error constraint (see “Objective Function Errors” on page 10-36), all acquisition functions modify their estimate of “goodness” following a suggestion of Gelbart, Snoek, and Adams [2]. Multiply the “goodness” by an estimate of the probability that the constraints are satisfied, to arrive at the acquisition function.

- “Expected Improvement” on page 10-4
- “Probability of Improvement” on page 10-4
- “Lower Confidence Bound” on page 10-4

Expected Improvement

The `'expected-improvement'` family of acquisition functions evaluates the expected amount of improvement in the objective function, ignoring values that cause an increase in the objective. In other words, define

- x_{best} as the location of the lowest posterior mean.
- $\mu_Q(x_{\text{best}})$ as the lowest value of the posterior mean.

Then the expected improvement

$$EI(x,Q) = E_Q\left[\max\left(0, \mu_Q(x_{\text{best}}) - f(x)\right)\right].$$

Probability of Improvement

The `'probability-of-improvement'` acquisition function makes a similar, but simpler, calculation as `'expected-improvement'`. In both cases, `bayesopt` first calculates x_{best} and $\mu_Q(x_{\text{best}})$. Then for `'probability-of-improvement'`, `bayesopt` calculates the probability PI that a new point x leads to a better objective function value, modified by a “margin” parameter m:

$$PI(x,Q) = P_Q\left(f(x) < \mu_Q(x_{\text{best}}) - m\right).$$

`bayesopt` takes m as the estimated noise standard deviation. `bayesopt` evaluates this probability as

$$PI = \Phi\left(\nu_Q(x)\right),$$

where

$$\nu_Q(x) = \frac{\mu_Q(x_{\text{best}}) - m - \mu_Q(x)}{\sigma_Q(x)}.$$

Here $\Phi(\cdot)$ is the unit normal CDF, and σ_Q is the posterior standard deviation of the Gaussian process at x.

Lower Confidence Bound

The `'lower-confidence-bound'` acquisition function looks at the curve G two standard deviations below the posterior mean at each point:

$$G(x) = \mu_Q(x) - 2\sigma_Q(x).$$

$G(x)$ is the $2\sigma_Q$ lower confidence envelope of the objective function model. `bayesopt` then maximizes the negative of G:

$$LCB = 2\sigma_Q(x) - \mu_Q(x).$$

Per Second

Sometimes, the time to evaluate the objective function can depend on the region. For example, many Support Vector Machine calculations vary in timing a good deal over certain ranges of points. If so, `bayesopt` can obtain better improvement per second by using time-weighting in its acquisition function. The cost-weighted acquisition functions have the phrase `per-second` in their names.

These acquisition functions work as follows. During the objective function evaluations, `bayesopt` maintains another Bayesian model of objective function evaluation time as a function of position x. The expected improvement per second that the acquisition function uses is

$$EIpS(x) = \frac{EI_Q(x)}{\mu_S(x)},$$

where $\mu_S(x)$ is the posterior mean of the timing Gaussian process model.

Plus

To escape a local objective function minimum, the acquisition functions with `plus` in their names modify their behavior when they estimate that they are overexploiting an area. To understand overexploiting, let $\sigma_F(x)$ be the standard deviation of the posterior objective function at x. Let σ be the posterior standard deviation of the additive noise, so that

$$\sigma_Q^2(x) = \sigma_F^2(x) + \sigma^2.$$

Define t_σ to be the value of the `ExplorationRatio` option, a positive number. The `bayesopt` `plus` acquisition functions, after each iteration, evaluate whether the next point x satisfies

$$\sigma_F(x) < t_\sigma\sigma.$$

If so, the algorithm declares that x is overexploiting. Then the acquisition function modifies its “Kernel Function” on page 10-3 by multiplying θ by the number of iterations, as suggested by Bull [1]. This modification raises the variance σ_Q for points in between observations. It then generates a new point based on the new fitted kernel function. If the new point x is again overexploiting, the acquisition function multiplies θ by an additional factor of 10 and tries again. It continues in this way up to five times, trying to generate a point x that is not overexploiting. The algorithm accepts the new x as the next point.

`ExplorationRatio` therefore controls a tradeoff between exploring new points for a better global solution, versus concentrating near points that have already been examined.

Acquisition Function Maximization

Internally, `bayesopt` maximizes an acquisition function using the following general steps:

1. For algorithms starting with `'expected-improvement'` and for `'probability-of-improvement'`, `bayesopt` estimates the smallest feasible mean of the posterior distribution $\mu_Q(x_{\text{best}})$ by sampling several thousand points within the variable bounds, taking several of the best (low mean value) feasible points, and improving them using local search, to find the ostensible best feasible point. Feasible means that the point satisfies constraints (see “Constraints in Bayesian Optimization” on page 10-38).
2. For all algorithms, `bayesopt` samples several thousand points within the variable bounds, takes several of the best (high acquisition function) feasible points, and improves them using local search, to find the ostensible best feasible point. The acquisition function value depends on the modeled posterior distribution, not a sample of the objective function, and so it can be calculated quickly.

References

[1] Bull, A. D. *Convergence rates of efficient global optimization algorithms*. https://arxiv.org/abs/1101.3501v3, 2011.

[2] Gelbart, M., J. Snoek, R. P. Adams. *Bayesian Optimization with Unknown Constraints*. https://arxiv.org/abs/1403.5607, 2014.

[3] Snoek, J., H. Larochelle, R. P. Adams. *Practical Bayesian Optimization of Machine Learning Algorithms*. https://arxiv.org/abs/1206.2944, 2012.

See Also

`BayesianOptimization` | `bayesopt`

Related Examples

- “Bayesian Optimization Workflow” on page 10-25
- “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45
- “Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 10-55
- “Optimize a Boosted Regression Ensemble” on page 10-63
- “Bayesian Optimization Output Function” on page 10-20
- “Bayesian Optimization with Coupled Constraints” on page 10-41
- “Bayesian Optimization with Tall Arrays” on page 30-9

Parallel Bayesian Optimization

In this section...

Optimize in Parallel

Running Bayesian optimization in parallel can save time. Running in parallel requires Parallel Computing Toolbox. `bayesopt` performs parallel objective function evaluations concurrently on parallel workers.

To optimize in parallel:

- `bayesopt` — Set the `UseParallel` name-value pair to `true`. For example,

  ```
  results = bayesopt(fun,vars,'UseParallel',true);
  ```

- Fit functions — Set the `UseParallel` field of the `HyperparameterOptimizationOptions` structure to `true`. For example,

  ```
  Mdl = fitcsvm(X,Y,'OptimizeHyperparameters','auto',...
      'HyperparameterOptimizationOptions',struct('UseParallel',true))
  ```

Parallel Bayesian Algorithm

The parallel Bayesian optimization algorithm is similar to the serial algorithm, which is described in “Bayesian Optimization Algorithm” on page 10-2. The differences are:

- `bayesopt` assigns points to evaluate to the parallel workers, generally one point at a time. `bayesopt` calculates on the client to determine which point to assign.
- After `bayesopt` evaluates the initial random points, it chooses points to evaluate by fitting a Gaussian process (GP) model. To fit a GP model while some workers are still evaluating points, `bayesopt` imputes a value to each point that is still on a worker. The imputed value is the mean of the GP model value at the points it is evaluating, or some other value as specified by the `bayesopt 'ParallelMethod'` name-value pair. For parallel optimization of fit functions, `bayesopt` uses the default `ParallelMethod` imputed value.
- After `bayesopt` assigns a point to evaluate, and before it computes a new point to assign, it checks whether too many workers are idle. The threshold for active workers is determined by the `MinWorkerUtilization` name-value pair. If too many workers are idle, then `bayesopt` assigns random points, chosen uniformly within bounds, to all idle workers. This step causes the workers to be active more quickly, but the workers have random points rather than fitted points. If the number of idle workers does not exceed the threshold, then `bayesopt` chooses a point to evaluate as usual, by fitting a GP model and maximizing the acquisition function.

Note Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results.

Settings for Best Parallel Performance

Fit functions have no special settings for better parallel performance. In contrast, several `bayesopt` settings can help to speed an optimization.

Solver Options

Setting the `GPActiveSetSize` option to a smaller value than the default (`300`) can speed the solution. The cost is potential inaccuracy in the points that `bayesopt` chooses to evaluate, because the GP model of the objective function can be less accurate than with a larger value. Setting the option to a larger value can result in a more accurate GP model, but requires more time to create the model.

Setting the `ParallelMethod` option to `'max-observed'` can lead `bayesopt` to search more widely for a global optimum. This choice can lead to a better solution in less time. However, the default value of `'clipped-model-prediction'` is often best.

Setting the `MinWorkerUtilization` option to a large value can result in higher parallel utilization. However, this setting causes more completely random points to be evaluated, which can lead to less accurate solutions. A large value, in this context, depends on how many workers you have. The default is `floor(0.8*N)`, where `N` is the number of parallel workers. Setting the option to a lower value can give lower parallel utilization, but with the benefit of higher quality points.

Placing the Objective Function on Workers

You can place an objective function on the parallel workers in one of three ways. Some have better performance, but require a more complex setup.

1. Automatic If you give a function handle as the objective function, `bayesopt` sends the handle to all the parallel workers at the beginning of its run. For example,

```
load ionosphere
splits = optimizableVariable('splits',[1,100],'Type','integer');
minleaf = optimizableVariable('minleaf',[1,100],'Type','integer');
fun = @(params)kfoldLoss(fitctree(X,Y,'Kfold',5,...
    'MaxNumSplits',params.splits,'MinLeaf',params.minleaf));

results = bayesopt(fun,[splits,minleaf],'UseParallel',true);
```

This method is effective if the handle is small, or if you run the optimization only once. However, if you plan to run the optimization several times, you can save time by using one of the other two techniques.

2. Parallel constant If you plan to run an optimization several times, save time by transferring the objective function to the workers only once. This technique is especially effective when the function handle incorporates a large amount of data. Transfer the objective once by setting the function handle to a `parallel.pool.Constant` construct, as in this example.

```
load ionosphere
splits = optimizableVariable('splits',[1,100],'Type','integer');
minleaf = optimizableVariable('minleaf',[1,100],'Type','integer');
fun = @(params)kfoldLoss(fitctree(X,Y,'Kfold',5,...
    'MaxNumSplits',params.splits,'MinLeaf',params.minleaf));

C = copyFunctionHandleToWorkers(fun);
```

```
results1 = bayesopt(C,[splits,minleaf],'UseParallel',true);
results2 = bayesopt(C,[splits,minleaf],'UseParallel',true,...
    'MaxObjectiveEvaluations',50);
results3 = bayesopt(C,[splits,minleaf],'UseParallel',true,...
    'AcquisitionFunction','expected-improvement');
```

In this example, `copyFunctionHandleToWorkers` sends the function handle to the workers only once.

3. Create objective function on workers If you have a great deal of data to send to the workers, you can avoid loading the data in the client by using `spmd` to load the data on the workers. Use a `Composite` with `parallel.pool.Constant` to access the distributed objective functions.

```
% makeFun is at the end of this script
spmd
    fun = makeFun();
end

% ObjectiveFunction is now a Composite. Get a parallel.pool.Constant
% that refers to it, without copying it to the client:
C = parallel.pool.Constant(fun);

% You could also use the line
% C = parallel.pool.Constant(@MakeFun);
% In this case, you do not use spmd

% Call bayesopt, passing the Constant
splits = optimizableVariable('splits', [1 100]);
minleaf = optimizableVariable('minleaf', [1 100]);
bo = bayesopt(C,[splits minleaf],'UseParallel',true);

function f = makeFun()
load('ionosphere','X','Y');
f = @fun;
    function L = fun(Params)
        L = kfoldLoss(fitctree(X,Y, ...
            'KFold', 5,...
            'MaxNumSplits',Params.splits, ...
            'MinLeaf', Params.minleaf));
    end
end
```

In this example, the function handle exists only on the workers. The handle never appears on the client.

Differences in Parallel Bayesian Optimization Output

When `bayesopt` runs in parallel, the Bayesian optimization output includes these differences.

- **Iterative Display** — Iterative display includes a column showing the number of active workers. This is the number after `bayesopt` assigns a job to the next worker.
- **Plot Functions**
  - Objective Function Model plot (`@plotObjectiveModel`) shows the pending points (those points executing on parallel workers). The height of the points depends on the `ParallelMethod` name-value pair.

- Elapsed Time plot (`@plotElapsedTime`) shows the total elapsed time with the label **Real time** and the total objective function evaluation time, summed over all workers, with the label **Objective evaluation time (all workers)**. Objective evaluation time includes the time to start a worker on a job.

See Also

`parallel.pool.Constant` | `spmd`

More About

- “Bayesian Optimization Algorithm” on page 10-2
- “Feature Extraction Workflow” on page 15-135

Bayesian Optimization Plot Functions

In this section...

Built-In Plot Functions

There are two sets of built-in plot functions.

Model Plots — Apply When D ≤ 2	Description
`@plotAcquisitionFunction`	Plot the acquisition function surface.
`@plotConstraintModels`	Plot each constraint model surface. Negative values indicate feasible points. Also plot a *P*(feasible) surface. Also plot the error model, if it exists, which ranges from –1 to 1. Negative values mean that the model probably does not error, positive values mean that it probably does error. The model is: Plotted error = 2*Probability(error) – 1.
`@plotObjectiveEvaluationTimeModel`	Plot the objective function evaluation time model surface.
`@plotObjectiveModel`	Plot the `fun` model surface, the estimated location of the minimum, and the location of the next proposed point to evaluate. For one-dimensional problems, plot envelopes one credible interval above and below the mean function, and envelopes one noise standard deviation above and below the mean.

Trace Plots — Apply to All D	Description
`@plotObjective`	Plot each observed function value versus the number of function evaluations.
`@plotObjectiveEvaluationTime`	Plot each observed function evaluation run time versus the number of function evaluations.
`@plotMinObjective`	Plot the minimum observed and estimated function values versus the number of function evaluations.
`@plotElapsedTime`	Plot three curves: the total elapsed time of the optimization, the total function evaluation time, and the total modeling and point selection time, all versus the number of function evaluations.

Note When there are coupled constraints, iterative display and plot functions can give counterintuitive results such as:

- A *minimum objective* plot can increase.
- The optimization can declare a problem infeasible even when it showed an earlier feasible point.

The reason for this behavior is that the decision about whether a point is feasible can change as the optimization progresses. `bayesopt` determines feasibility with respect to its constraint model, and this model changes as `bayesopt` evaluates points. So a "minimum objective" plot can increase when the minimal point is later deemed infeasible, and the iterative display can show a feasible point that is later deemed infeasible.

Custom Plot Function Syntax

A custom plot function has the same syntax as a custom output function (see "Bayesian Optimization Output Functions" on page 10-19):

```
stop = plotfun(results,state)
```

`bayesopt` passes the `results` and `state` variables to your function. Your function returns `stop`, which you set to `true` to halt the iterations, or to `false` to continue the iterations.

`results` is an object of class `BayesianOptimization` that contains the available information on the computations.

`state` has these possible values:

- `'initial'` — `bayesopt` is about to start iterating. Use this state to set up a plot or to perform other initializations.
- `'iteration'` — `bayesopt` just finished an iteration. Generally, you perform most of the plotting or other calculations in this state.
- `'done'` — `bayesopt` just finished its final iteration. Clean up plots or otherwise prepare for the plot function to shut down.

Create a Custom Plot Function

This example shows how to create a custom plot function for `bayesopt`. It further shows how to use information in the `UserData` property of a `BayesianOptimization` object.

Problem Statement

The problem is to find parameters of a Support Vector Machine (SVM) classification to minimize the cross-validated loss. The specific model is the same as in "Optimize a Cross-Validated SVM Classifier Using bayesopt" on page 10-45. Therefore, the objective function is essentially the same, except it also computes `UserData`, in this case the number of support vectors in an SVM model fitted to the current parameters.

Create a custom plot function that plots the number of support vectors in the SVM model as the optimization progresses. To give the plot function access to the number of support vectors, create a third output, `UserData`, to return the number of support vectors.

Objective Function

Create an objective function that computes the cross-validation loss for a fixed cross-validation partition, and that returns the number of support vectors in the resulting model.

```
function [f,viol,nsupp] = mysvmminfn(x,cdata,grp,c)
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'KernelScale',x.sigma,'BoxConstraint',x.box);
f = kfoldLoss(crossval(SVMModel,'CVPartition',c));
viol = [];
nsupp = sum(SVMModel.IsSupportVector);
end
```

Custom Plot Function

Create a custom plot function that uses the information computed in `UserData`. Have the function plot both the current number of constraints and the number of constraints for the model with the best objective function found.

```
function stop = svmsuppvec(results,state)
persistent hs nbest besthist nsupptrace
stop = false;
switch state
    case 'initial'
        hs = figure;
        besthist = [];
        nbest = 0;
        nsupptrace = [];
    case 'iteration'
        figure(hs)
        nsupp = results.UserDataTrace{end};   % get nsupp from UserDataTrace property.
        nsupptrace(end+1) = nsupp; % accumulate nsupp values in a vector.
        if (results.ObjectiveTrace(end) == min(results.ObjectiveTrace)) || (length(results.Objec
            nbest = nsupp;
        end
        besthist = [besthist,nbest];
        plot(1:length(nsupptrace),nsupptrace,'b',1:length(besthist),besthist,'r--')
        xlabel 'Iteration number'
        ylabel 'Number of support vectors'
        title 'Number of support vectors at each iteration'
        legend('Current iteration','Best objective','Location','best')
        drawnow
end
```

Set Up the Model

Generate ten base points for each class.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
```

Generate 100 data points of each class.

```
redpts = zeros(100,2);grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
```

Put the data into one matrix, and make a vector `grp` that labels the class of each point.

```
cdata = [grnpts;redpts];
grp = ones(200,1);
% Green label 1, red label -1
grp(101:200) = -1;
```

Check the basic classification of all the data using the default SVM parameters.

```
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf','ClassNames',[-1 1]);
```

Set up a partition to fix the cross validation. Without this step, the cross validation is random, so the objective function is not deterministic.

```
c = cvpartition(200,'KFold',10);
```

Check the cross-validation accuracy of the original fitted model.

```
loss = kfoldLoss(fitcsvm(cdata,grp,'CVPartition',c,...
    'KernelFunction','rbf','BoxConstraint',SVMModel.BoxConstraints(1),...
    'KernelScale',SVMModel.KernelParameters.Scale))
```

```
loss =

    0.1350
```

Prepare Variables for Optimization

The objective function takes an input `z = [rbf_sigma,boxconstraint]` and returns the cross-validation loss value of `z`. Take the components of `z` as positive, log-transformed variables between `1e-5` and `1e5`. Choose a wide range because you do not know which values are likely to be good.

```
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
```

Set Plot Function and Call the Optimizer

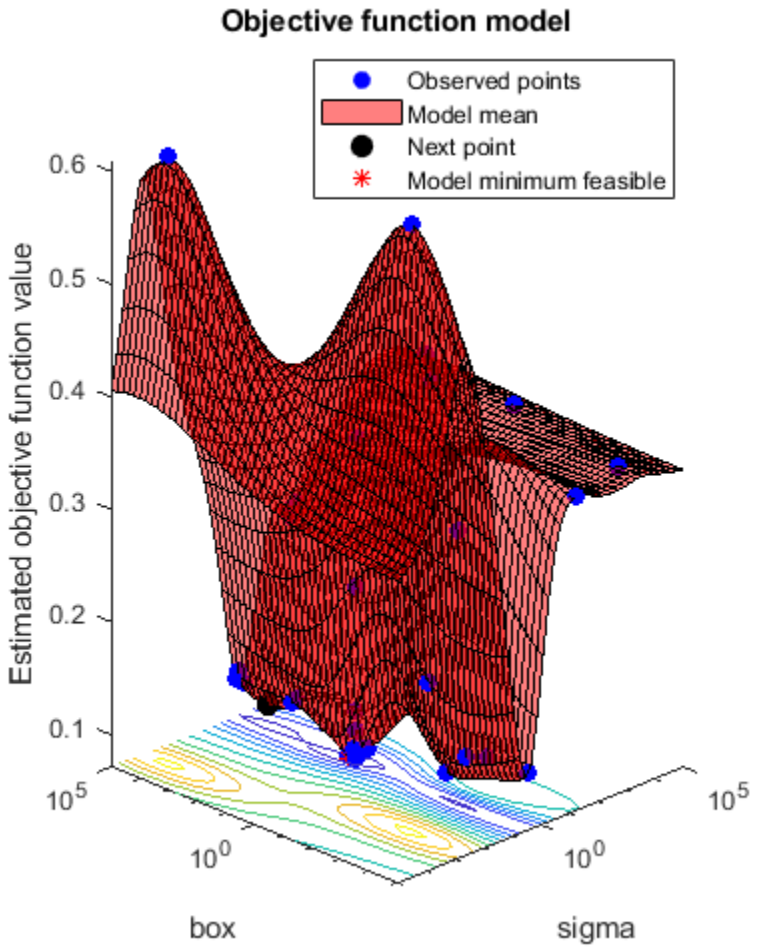
Search for the best parameters `[sigma,box]` using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time, so it can give varying results.

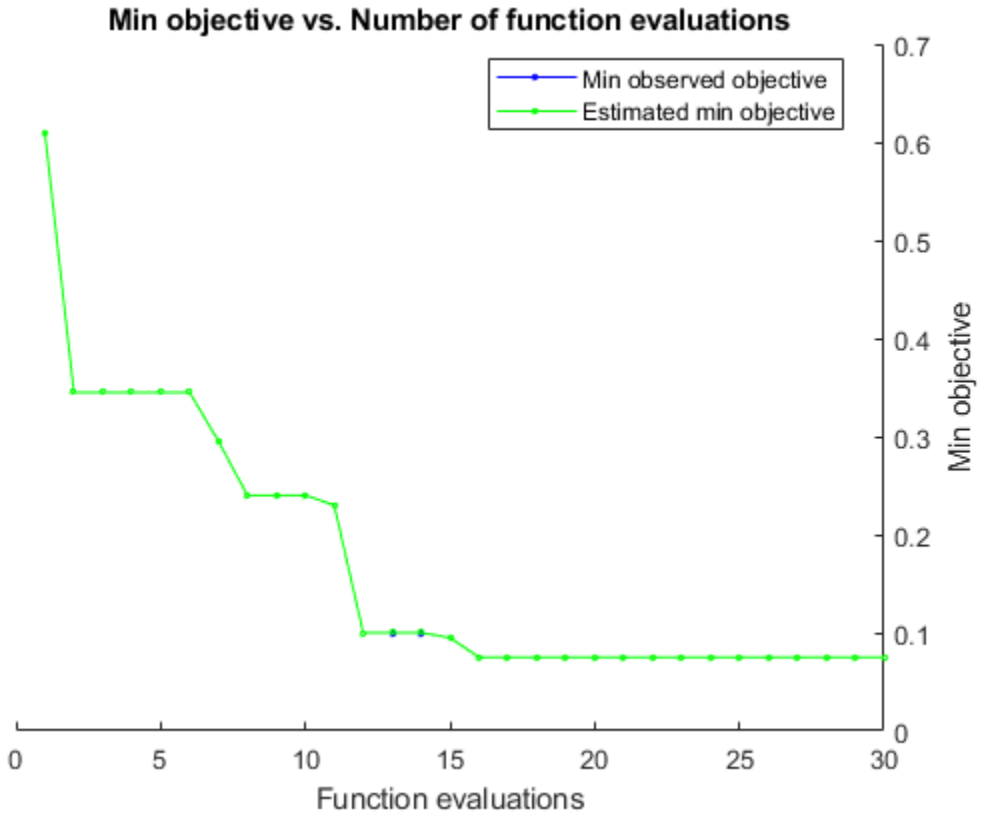
Plot the number of support vectors as a function of the iteration number, and plot the number of support vectors for the best parameters found.

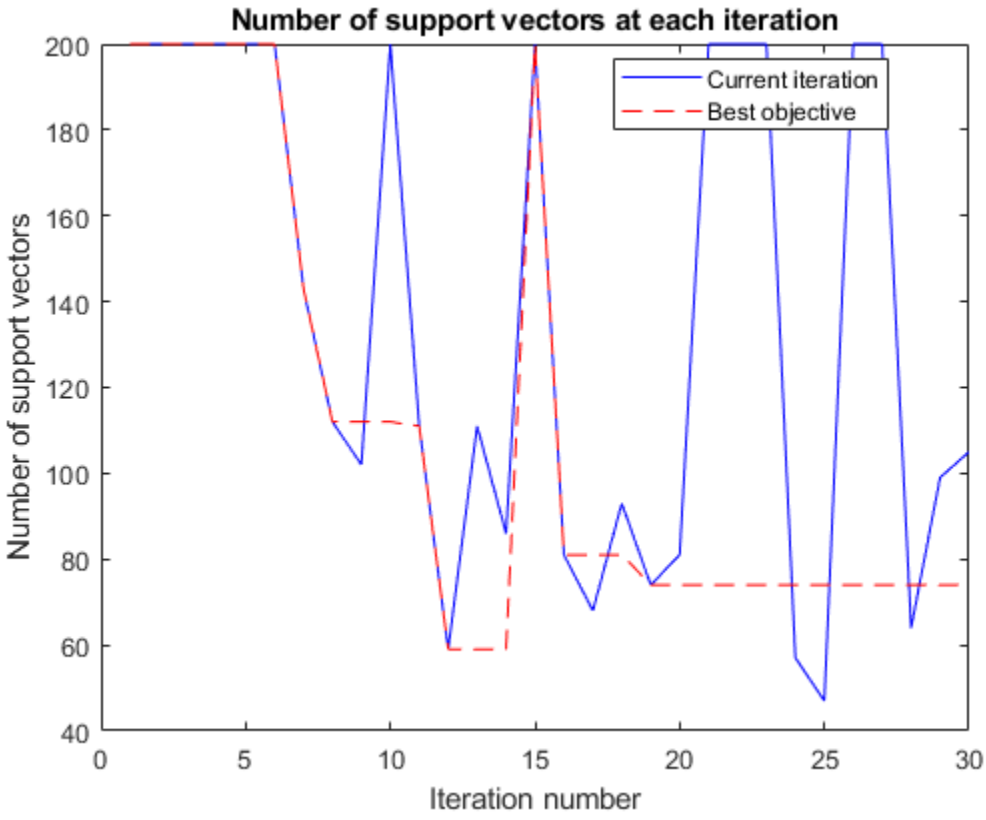
```
obj = @(x)mysvmminfn(x,cdata,grp,c);
results = bayesopt(obj,[sigma,box],...
    'IsObjectiveDeterministic',true,'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus',...
    'PlotFcn',{@svmsuppvec,@plotObjectiveModel,@plotMinObjective})
```

```
results =

  BayesianOptimization with properties:

                      ObjectiveFcn: @(x)mysvmminfn(x,cdata,grp,c)
```

```
          VariableDescriptions: [1x2 optimizableVariable]
                       Options: [1x1 struct]
                  MinObjective: 0.0750
               XAtMinObjective: [1x2 table]
         MinEstimatedObjective: 0.0750
      XAtMinEstimatedObjective: [1x2 table]
       NumObjectiveEvaluations: 30
              TotalElapsedTime: 43.6662
                     NextPoint: [1x2 table]
                        XTrace: [30x2 table]
                ObjectiveTrace: [30x1 double]
              ConstraintsTrace: []
                 UserDataTrace: {30x1 cell}
  ObjectiveEvaluationTimeTrace: [30x1 double]
            IterationTimeTrace: [30x1 double]
                    ErrorTrace: [30x1 double]
              FeasibilityTrace: [30x1 logical]
   FeasibilityProbabilityTrace: [30x1 double]
          IndexOfMinimumTrace: [30x1 double]
        ObjectiveMinimumTrace: [30x1 double]
EstimatedObjectiveMinimumTrace: [30x1 double]
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.6
0.5
0.4
0.3
0.2
0.1
10^5
10^0
box
10^5
10^0
sigma


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25
30



See Also

Related Examples

- “Bayesian Optimization Output Functions” on page 10-19
- “Constraints in Bayesian Optimization” on page 10-38

Bayesian Optimization Output Functions

In this section...

What Is a Bayesian Optimization Output Function?

An output function is a function that is called at the end of every iteration of `bayesopt`. An output function can halt iterations. It can also create plots, save information to your workspace or to a file, or perform any other calculation you like.

Other than halting the iterations, output functions cannot change the course of a Bayesian optimization. They simply monitor the progress of the optimization.

Built-In Output Functions

These built-in output functions save your optimization results to a file or to the workspace.

- `@assignInBase` — Saves your results after each iteration to a variable named `'BayesoptResults'` in your workspace. To choose a different name, pass the `SaveVariableName` name-value pair.
- `@saveToFile` — Saves your results after each iteration to a file named `'BayesoptResults.mat'` in your current folder. To choose a different name or folder, pass the `SaveFileName` name-value pair.

For example, to save the results after each iteration to a workspace variable named `'BayesIterations'`,

```
results = bayesopt(fun,vars,'OutputFcn',@assignInBase, ...
    'SaveVariableName','BayesIterations')
```

Custom Output Functions

Write a custom output function with signature

```
stop = outputfun(results,state)
```

`bayesopt` passes the `results` and `state` variables to your function. Your function returns `stop`, which you set to `true` to halt the iterations, or to `false` to allow the iterations to continue.

`results` is an object of class `BayesianOptimization`. `results` contains the available information on the computations so far.

`state` has possible values:

- `'initial'` — `bayesopt` is about to start iterating.
- `'iteration'` — `bayesopt` just finished an iteration.

- `'done'` — `bayesopt` just finished its final iteration.

For an example, see “Bayesian Optimization Output Function” on page 10-20.

Bayesian Optimization Output Function

This example shows how to use a custom output function with Bayesian optimization. The output function halts the optimization when the objective function, which is the cross-validation error rate, drops below 13%. The output function also plots the time for each iteration.

```
function stop = outputfun(results,state)
persistent h
stop = false;
switch state
    case 'initial'
        h = figure;
    case 'iteration'
        if results.MinObjective < 0.13
            stop = true;
        end
        figure(h)
        tms = results.IterationTimeTrace;
        plot(1:numel(tms),tms')
        xlabel('Iteration Number')
        ylabel('Time for Iteration')
        title('Time for Each Iteration')
        drawnow
end
```

The objective function is the cross validation loss of the KNN classification of the `ionosphere` data. Load the data and, for reproducibility, set the default random stream.

```
load ionosphere
rng default
```

Optimize over neighborhood size from 1 through 30, and for three distance metrics.

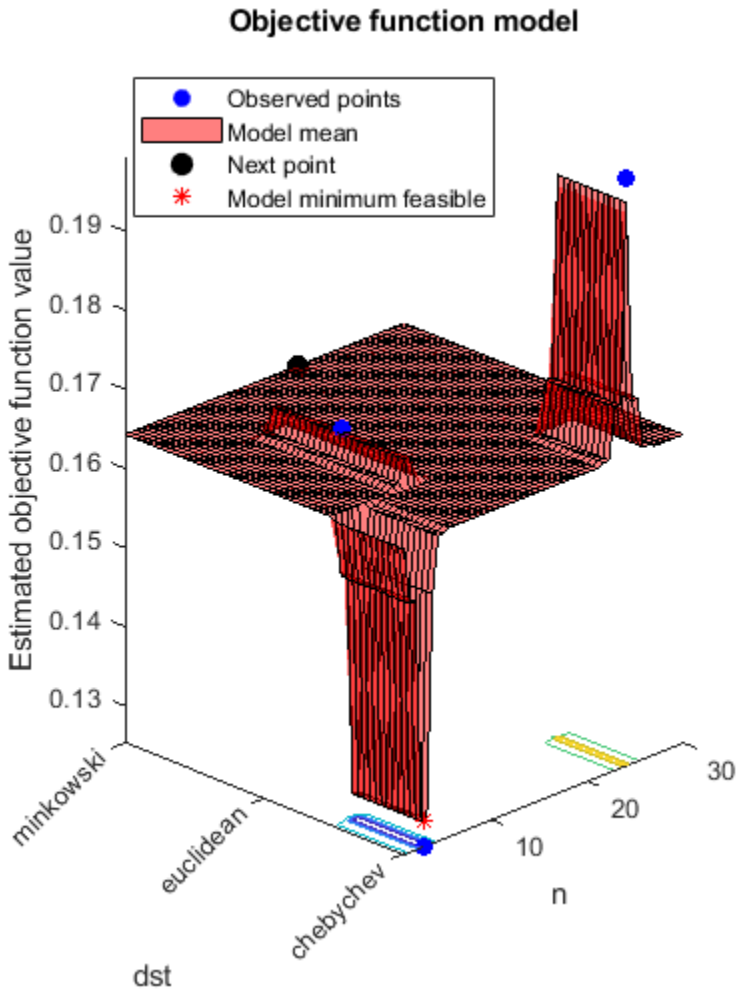
```
num = optimizableVariable('n',[1,30],'Type','integer');
dst = optimizableVariable('dst',{'chebychev','euclidean','minkowski'},'Type','categorical');
vars = [num,dst];
```

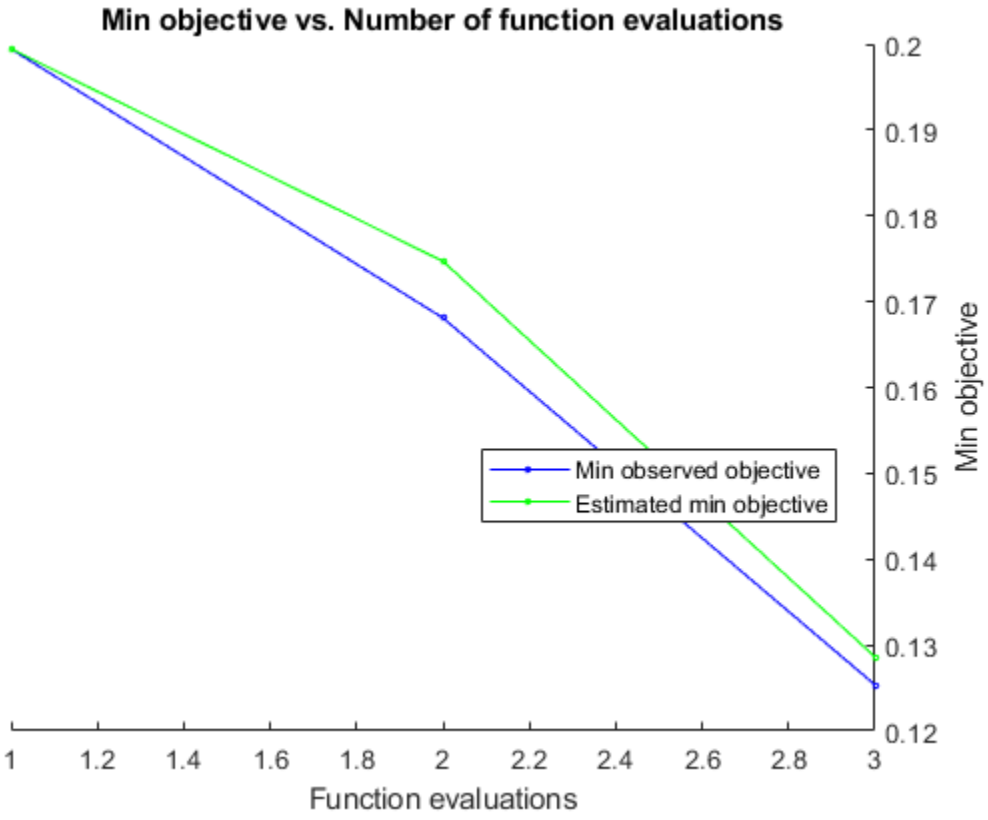
Set the cross-validation partition and objective function. For reproducibility, set the `AcquisitionFunctionName` to `'expected-improvement-plus'`. Run the optimization.

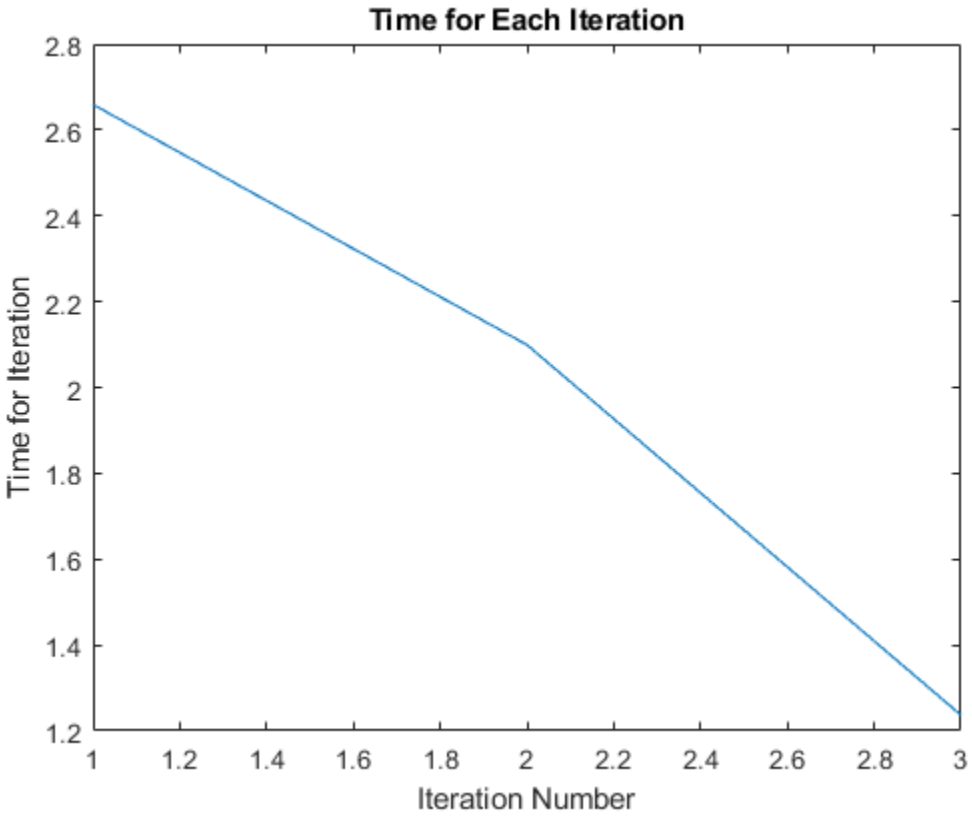
```
c = cvpartition(351,'Kfold',5);
fun = @(x)kfoldLoss(fitcknn(X,Y,'CVPartition',c,'NumNeighbors',x.n,...
    'Distance',char(x.dst),'NSMethod','exhaustive'));
results = bayesopt(fun,vars,'OutputFcn',@outputfun,...
    'AcquisitionFunctionName','expected-improvement-plus');
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |           n |
|      | result |             | runtime     | (observed)  | (estim.)    |             |
|=====================================================================================================
|    1 | Best   |     0.19943 |     0.39391 |     0.19943 |     0.19943 |          24 |    cheby
```

```
|    2 | Best   |     0.16809 |     0.18704 |     0.16809 |      0.1747 |           9 |    eucl
|    3 | Best   |     0.12536 |     0.18892 |     0.12536 |     0.12861 |           3 |    cheby

__________________________________________________________
Optimization completed.
Total function evaluations: 3
Total elapsed time: 6.0781 seconds
Total objective function evaluation time: 0.76987

Best observed feasible point:
    n       dst
    _    _________

    3    chebychev

Observed objective function value = 0.12536
Estimated objective function value = 0.12861
Function evaluation time = 0.18892

Best estimated feasible point (according to models):
    n       dst
    _    _________

    3    chebychev

Estimated objective function value = 0.12861
Estimated function evaluation time = 0.2404
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.19
0.18
0.17
0.16
0.15
0.14
0.13
minkowski
euclidean
chebychev
dst
10
20
30
n


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Function evaluations
Min objective
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.2



See Also

Related Examples

- “Bayesian Optimization Plot Functions” on page 10-11

Bayesian Optimization Workflow

In this section...

What Is Bayesian Optimization?

Optimization, in its most general form, is the process of locating a point that minimizes a real-valued function called the objective function. Bayesian optimization is the name of one such process. Bayesian optimization internally maintains a Gaussian process model of the objective function, and uses objective function evaluations to train the model. One innovation in Bayesian optimization is the use of an acquisition function, which the algorithm uses to determine the next point to evaluate. The acquisition function can balance sampling at points that have low modeled objective functions, and exploring areas that have not yet been modeled well. For details, see “Bayesian Optimization Algorithm” on page 10-2.

Bayesian optimization is part of Statistics and Machine Learning Toolbox because it is well-suited to optimizing hyperparameters of classification and regression algorithms. A hyperparameter is an internal parameter of a classifier or regression function, such as the box constraint of a support vector machine, or the learning rate of a robust classification ensemble. These parameters can strongly affect the performance of a classifier or regressor, and yet it is typically difficult or time-consuming to optimize them. See “Bayesian Optimization Characteristics” on page 10-27.

Typically, optimizing the hyperparameters means that you try to minimize the cross-validation loss of a classifier or regression.

Ways to Perform Bayesian Optimization

You can perform a Bayesian optimization in several ways:

- `fitcauto` and `fitrauto` — Pass predictor and response data to the `fitcauto` or `fitrauto` function to optimize across a selection of model types and hyperparameter values. Unlike other approaches, using `fitcauto` or `fitrauto` does not require you to specify a single model before the optimization; model selection is part of the optimization process. The optimization minimizes cross-validation loss, which is modeled using a multi-`TreeBagger` model in `fitcauto` and a multi-`RegressionGP` model in `fitrauto`, rather than a single Gaussian process regression model as used in other approaches. See “Bayesian Optimization” on page 33-1571 for `fitcauto` and “Bayesian Optimization” on page 33-2113 for `fitrauto`.
- Classification Learner and Regression Learner apps — Choose **Optimizable** models in the machine learning apps and automatically tune their hyperparameter values by using Bayesian optimization. The optimization minimizes the model loss based on the selected validation options. This approach has fewer tuning options than using a fit function, but allows you to perform

Bayesian optimization directly in the apps. See “Hyperparameter Optimization in Classification Learner App” on page 23-54 and “Hyperparameter Optimization in Regression Learner App” on page 24-30.

- Fit function — Include the `OptimizeHyperparameters` name-value pair in many fitting functions to apply Bayesian optimization automatically. The optimization minimizes cross-validation loss. This approach gives you fewer tuning options than using `bayesopt`, but enables you to perform Bayesian optimization more easily. See “Bayesian Optimization Using a Fit Function” on page 10-26.
- `bayesopt` — Exert the most control over your optimization by calling `bayesopt` directly. This approach requires you to write an objective function, which does not have to represent cross-validation loss. See “Bayesian Optimization Using bayesopt” on page 10-26.

Bayesian Optimization Using a Fit Function

To minimize the error in a cross-validated response via Bayesian optimization, follow these steps.

1. Choose your classification or regression solver among `fitcdiscr`, `fitcecoc`, `fitcensemble`, `fitckernel`, `fitcknn`, `fitclinear`, `fitcnb`, `fitcsvm`, `fitctree`, `fitrensemble`, `fitrgp`, `fitrkernel`, `fitrlinear`, `fitrsvm`, or `fitrtree`.
2. Decide on the hyperparameters to optimize, and pass them in the `OptimizeHyperparameters` name-value pair. For each fit function, you can choose from a set of hyperparameters. See Eligible Hyperparameters for Fit Functions, or use the `hyperparameters` function, or consult the fit function reference page.

   You can pass a cell array of parameter names. You can also set `'auto'` as the `OptimizeHyperparameters` value, which chooses a typical set of hyperparameters to optimize, or `'all'` to optimize all available parameters.
3. For ensemble fit functions `fitcecoc`, `fitcensemble`, and `fitrensemble`, also include parameters of the weak learners in the `OptimizeHyperparameters` cell array.
4. Optionally, create an options structure for the `HyperparameterOptimizationOptions` name-value pair. See “Hyperparameter Optimization Options for Fit Functions” on page 10-29.
5. Call the fit function with the appropriate name-value pairs.

For examples, see “Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 10-55 and “Optimize a Boosted Regression Ensemble” on page 10-63. Also, every fit function reference page contains a Bayesian optimization example.

Bayesian Optimization Using bayesopt

To perform a Bayesian optimization using `bayesopt`, follow these steps.

1. Prepare your variables. See “Variables for a Bayesian Optimization” on page 10-33.
2. Create your objective function. See “Bayesian Optimization Objective Functions” on page 10-36. If necessary, create constraints, too. See “Constraints in Bayesian Optimization” on page 10-38. To include extra parameters in an objective function, see “Parameterizing Functions”.
3. Decide on options, meaning the `bayseopt` `Name,Value` on page 33-133 pairs. You are not required to pass any options to `bayesopt` but you typically do, especially when trying to improve a solution.
4. Call `bayesopt`.

5 Examine the solution. You can decide to resume the optimization by using `resume`, or restart the optimization, usually with modified options.

For an example, see “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45.

Bayesian Optimization Characteristics

Bayesian optimization algorithms are best suited to these problem types.

Characteristic	Details
Low dimension	Bayesian optimization works best in a low number of dimensions, typically 10 or fewer. While Bayesian optimization can solve some problems with a few dozen variables, it is not recommended for dimensions higher than about 50.
Expensive objective	Bayesian optimization is designed for objective functions that are slow to evaluate. It has considerable overhead, typically several seconds for each iteration.
Low accuracy	Bayesian optimization does not necessarily give very accurate results. If you have a deterministic objective function, you can sometimes improve the accuracy by starting a standard optimization algorithm from the `bayesopt` solution.
Global solution	Bayesian optimization is a global technique. Unlike many other algorithms, to search for a global solution you do not have to start the algorithm from various initial points.
Hyperparameters	Bayesian optimization is well-suited to optimizing hyperparameters of another function. A hyperparameter is a parameter that controls the behavior of a function. For example, the `fitcsvm` function fits an SVM model to data. It has hyperparameters `BoxConstraint` and `KernelScale` for its `'rbf'` `KernelFunction`. For an example of Bayesian optimization applied to hyperparameters, see “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45.

Parameters Available for Fit Functions

Eligible Hyperparameters for Fit Functions

Function Name	Eligible Parameters

`fitcdiscr`	`Delta`
	`Gamma`
	`DiscrimType`

`fitcecoc`	`Coding` eligible `fitcdiscr` parameters for `'Learners','discriminant'` eligible `fitckernel` parameters for `'Learners','kernel'` eligible `fitcknn` parameters for `'Learners','knn'` eligible `fitclinear` parameters for `'Learners','linear'` eligible `fitcsvm` parameters for `'Learners','svm'` eligible `fitctree` parameters for `'Learners','tree'`

`fitcensemble`	`Method`
	`NumLearningCycles`
	`LearnRate`
	eligible `fitcdiscr` parameters for `'Learners','discriminant'`
	eligible `fitcknn` parameters for `'Learners','knn'`
	eligible `fitctree` parameters for `'Learners','tree'`

`fitckernel`	`'Learner'`
	`'KernelScale'`
	`'Lambda'`
	`'NumExpansionDimensions'`

`fitcknn`	`NumNeighbors`
	`Distance`
	`DistanceWeight`
	`Exponent`
	`Standardize`

`fitclinear`	`Lambda`
	`Learner`
	`Regularization`

`fitcnb`	`DistributionNames`
	`Width`
	`Kernel`

`fitcsvm`	`BoxConstraint`
	`KernelScale`
	`KernelFunction`
	`PolynomialOrder`
	`Standardize`

`fitctree`	`MinLeafSize`
	`MaxNumSplits`
	`SplitCriterion`
	`NumVariablesToSample`

Function Name	Eligible Parameters

`fitrensemble`	`Method`
	`NumLearningCycles`
	`LearnRate`
	eligible `fitrtree` parameters for `'Learners','tree'`:
	`MinLeafSize`
	`MaxNumSplits`
	`NumVariablesToSample`

`fitrgp`	`Sigma`
	`BasisFunction`
	`KernelFunction`
	`KernelScale`
	`Standardize`

`fitrkernel`	`Learner`
	`KernelScale`
	`Lambda`
	`NumExpansionDimensions`
	`Epsilon`

`fitrlinear`	`Lambda`
	`Learner`
	`Regularization`

`fitrsvm`	`BoxConstraint`
	`KernelScale`
	`Epsilon`
	`KernelFunction`
	`PolynomialOrder`
	`Standardize`

`fitrtree`	`MinLeafSize`
	`MaxNumSplits`
	`NumVariablesToSample`

Hyperparameter Optimization Options for Fit Functions

When optimizing using a fit function, you have these options available in the `HyperparameterOptimizationOptions` name-value pair. Give the value as a structure. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

See Also

`BayesianOptimization` | `bayesopt`

More About

- “Bayesian Optimization Algorithm” on page 10-2
- “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45
- “Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 10-55
- “Optimize a Boosted Regression Ensemble” on page 10-63
- “Bayesian Optimization Output Function” on page 10-20
- “Bayesian Optimization with Coupled Constraints” on page 10-41
- “Bayesian Optimization with Tall Arrays” on page 30-9

Variables for a Bayesian Optimization

In this section...

Syntax for Creating Optimization Variables

For each variable in your objective function, create a variable description object using `optimizableVariable`. Each variable has a unique name and a range of values. The minimal syntax for variable creation is

```
variable = optimizableVariable(Name,Range)
```

This function creates a real variable that ranges from the lower bound `Range(1)` to the upper bound `Range(2)`.

You can specify three types of variables in the `Type` name-value pair:

- `'real'` — Continuous real values between finite bounds. Give `Range` as the two-element vector `[lower upper]`, which represent the lower and upper bounds.
- `'integer'` — Integer values between finite bounds, similar to `'real'`.
- `'categorical'` — Cell array of names of possible values, such as `{'red','green','blue'}`, that you specify in the `Range` argument.

For `'real'` or `'integer'` variables, you can specify that `bayesopt` searches in a log-scaled space by setting the `Transform` name-value pair to `'log'`. For this transformation, ensure that the lower bound in the `Range` is strictly positive.

Include variables for `bayesopt` as a vector in the second argument.

```
results = bayesopt(fun,[xvar,ivar,rvar])
```

To exclude a variable from an optimization, set `Optimize` to `false`, either in the name-value pair of `optimizableVariable`, or by dot notation:

```
xvar.Optimize = false;
```

Tip

- There are two names associated with an `optimizableVariable`:
  - The MATLAB workspace variable name
  - The name of the variable in the optimization

  For example,

  ```
  xvar = optimizableVariable('spacevar',[1,100]);
  ```

  `xvar` is the MATLAB workspace variable, and `'spacevar'` is the variable in the optimization.

  Use these names as follows:

- Use `xvar` as an element in the vector of variables you pass to `bayesopt`. For example,

  ```
  results = bayesopt(fun,[xvar,tvar])
  ```

- Use `'spacevar'` as the name of the variable in the optimization. For example, in an objective function,

  ```
  function objective = mysvmfun(x,cdata,grp)
  SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
      'BoxConstraint',x.spacevar,...
      'KernelScale',x.tvar);
  objective = kfoldLoss(crossval(SVMModel));
  ```

Variables for Optimization Examples

Real variable from 0 to 1:

```
var1 = optimizableVariable('xvar',[0 1])
```

```
var1 =
  optimizableVariable with properties:

         Name: 'xvar'
        Range: [0 1]
         Type: 'real'
    Transform: 'none'
     Optimize: 1
```

Integer variable from 1 to 1000 on a log scale:

```
var2 = optimizableVariable('ivar',[1 1000],'Type','integer','Transform','log')
```

```
var2 =
  optimizableVariable with properties:

         Name: 'ivar'
        Range: [1 1000]
         Type: 'integer'
    Transform: 'log'
     Optimize: 1
```

Categorical variable of rainbow colors:

```
var3 = optimizableVariable('rvar',{'r' 'o' 'y' 'g' 'b' 'i' 'v'},'Type','categorical')
```

```
var3 =
  optimizableVariable with properties:

         Name: 'rvar'
        Range: {'r'  'o'  'y'  'g'  'b'  'i'  'v'}
         Type: 'categorical'
    Transform: 'none'
     Optimize: 1
```

See Also

Related Examples

- “Bayesian Optimization Workflow” on page 10-25
- “Bayesian Optimization Objective Functions” on page 10-36

Bayesian Optimization Objective Functions

In this section...

Objective Function Syntax

`bayesopt` attempts to minimize an objective function. If, instead, you want to maximize a function, set the objective function to the negative of the function you want to maximize. See “Maximizing Functions”. To include extra parameters in an objective function, see “Parameterizing Functions”.

`bayesopt` passes a table of variables to the objective function. The variables have the names and types that you declare; see “Variables for a Bayesian Optimization” on page 10-33.

The objective function has the following signature:

```
[objective,coupledconstraints,userdata] = fun(x)
```

1. `objective` — The objective function value at `x`, a real scalar.
2. `coupledconstraints` — Value of coupled constraints, if any (optional output), a vector of real values. A negative value indicates that a constraint is satisfied, a positive value indicates that it is not satisfied. For details, see “Coupled Constraints” on page 10-40.
3. `userdata` — Optional data that your function can return for further uses, such as plotting or logging (optional output). For an example, see “Bayesian Optimization Plot Functions” on page 10-11.

Objective Function Example

This objective function returns the loss in a cross-validated fit of an SVM model with parameters `box` and `sigma`. The objective also returns a coupled constraint function that is positive (infeasible) when the number of support vectors exceeds 100 (100 is feasible, 101 is not).

```
function [objective,constraint] = mysvmfun(x,cdata,grp)
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'BoxConstraint',x.box,...
    'KernelScale',x.sigma);
objective = kfoldLoss(crossval(SVMModel));
constraint = sum(SVMModel.SupportVectors) - 100.5;
```

To use the objective function, assuming that `cdata` and `grp` exist in the workspace, create an anonymous function that incorporates the data, as described in “Parameterizing Functions”.

```
fun = @(x)mysvmfun(x,cdata,grp);
results = bayesopt(fun,vars) % Assumes vars exists
```

Objective Function Errors

`bayesopt` deems your objective function to return an error when the objective function returns anything other than a finite real scalar. For example, if your objective function returns a complex

value, `NaN`, `Inf`, or matrix with more than one entry, then `bayesopt` deems that your objective function errors. If `bayesopt` encounters an error, it continues to optimize, and automatically updates a Bayesian model of points that lead to errors. This Bayesian model is the Error model. `bayesopt` incorporates the Error model as a coupled constraint. See “Coupled Constraints” on page 10-40.

When errors exist, you can plot the Error model by setting the `bayesopt` `PlotFcn` name-value pair `@plotConstraintModels`. Or you can retrospectively call `plot` on the results of a Bayesian optimization, and include `@plotConstraintModels`.

See Also

Related Examples

- “Bayesian Optimization Workflow” on page 10-25
- “Variables for a Bayesian Optimization” on page 10-33
- “Constraints in Bayesian Optimization” on page 10-38

Constraints in Bayesian Optimization

In this section...

Bounds

`bayesopt` requires finite bounds on all variables. (`categorical` variables are, by nature, bounded in their possible values.) Pass the lower and upper bounds for real and integer-valued variables in `optimizableVariable`.

`bayesopt` uses these bounds to sample points, either uniformly or log-scaled. You set the scaling for sampling in `optimizableVariable`.

For example, to constrain a variable `X1` to values between `1e-6` and `1e3`, scaled logarithmically,

```
xvar = optimizableVariable('X1',[1e-6,1e3],'Transform','log')
```

`bayesopt` includes the endpoints in its range. Therefore, you cannot use 0 as a lower bound for a log-transformed variable.

Tip To use a zero lower bound in a log-transformed variable, set the lower bound to `1`, then inside the objective function use `x-1`.

Deterministic Constraints — XConstraintFcn

Sometimes your problem is valid or well-defined only for points in a certain region, called the feasible region. A deterministic constraint is a deterministic function that returns `true` when a point is feasible, and `false` when a point is infeasible. So deterministic constraints are not stochastic, and they are not functions of a group of points, but of individual points.

Tip It is more efficient to use `optimizableVariable` bounds, instead of deterministic constraints, to confine the optimization to a rectangular region.

Write a deterministic constraint function using the signature

```
tf = xconstraint(X)
```

- `X` is a width-D table of arbitrary height.
- `tf` is a logical column vector, where `tf(i) = true` exactly when `X(i,:)` is feasible.

Pass the deterministic constraint function in the `bayesopt` `XConstraintFcn` name-value pair. For example,

```
results = bayesopt(fun,vars,'XConstraintFcn',@xconstraint)
```

`bayesopt` evaluates deterministic constraints on thousands of points, and so runs faster when your constraint function is vectorized. See “Vectorization”.

For example, suppose that the variables named `'x1'` and `'x2'` are feasible when the norm of the vector `[x1 x2]` is less than `6`, and when `x1 <= x2`. The following constraint function evaluates these constraints.

```
function tf = xconstraint(X)
tf1 = sqrt(X.x1.^2 + X.x2.^2) < 6;
tf2 = X.x1 <= X.x2;
tf = tf1 & tf2;
```

Conditional Constraints — ConditionalVariableFcn

Conditional constraints are functions that enforce one of the following two conditions:

- When some variables have certain values, other variables are set to given values.
- When some variables have certain values, other variables have `NaN` or, for categorical variables, `<undefined>` values.

Specify a conditional constraint by setting the `bayesopt` `ConditionalVariableFcn` name-value pair to a function handle, say `@condvariablefcn`. The `@condvariablefcn` function must have the signature

```
Xnew = condvariablefcn(X)
```

- `X` is a width-`D` table of arbitrary height.
- `Xnew` is a table the same type and size as `X`.

`condvariablefcn` sets `Xnew` to be equal to `X`, except it also sets the relevant variables in each row of `Xnew` to the correct values for the constraint.

Note If you have both conditional constraints and deterministic constraints, `bayesopt` applies the conditional constraints first. Therefore, if your conditional constraint function can set variables to `NaN` or `<undefined>`, ensure that your deterministic constraint function can process these values correctly.

Conditional constraints ensure that variable values are sensible. Therefore, `bayesopt` applies conditional constraints first so that all passed values are sensible.

Conditional Constraint That Sets a Variable Value

Suppose that you are optimizing a classification using `fitcdiscr`, and you optimize over both the `'DiscrimType'` and `'Gamma'` name-value pair arguments. When `'DiscrimType'` is one of the quadratic types, `'Gamma'` must be `0` or the solver errors. In that case, use this conditional constraint function:

```
function XTable = fitcdiscrCVF(XTable)
% Gamma must be 0 if discrim type is a quadratic
XTable.Gamma(ismember(XTable.DiscrimType, {'quadratic',...
```

```
        'diagQuadratic','pseudoQuadratic'})) = 0;
end
```

Conditional Constraint That Sets a Variable to NaN

Suppose that you are optimizing a classification using `fitcsvm`, and you optimize over both the `'KernelFunction'` and `'PolynomialOrder'` name-value pair arguments. When `'KernelFunction'` is not `'polynomial'`, the `'PolynomialOrder'` setting does not apply. The following function enforces this conditional constraint.

```
function Xnew = condvariablefcn(X)
Xnew = X;
Xnew.PolynomialOrder(Xnew.KernelFunction ~= 'polynomial') = NaN;
```

You can save a line of code as follows:

```
function X = condvariablefcn(X)
X.PolynomialOrder(Xnew.KernelFunction ~= 'polynomial') = NaN;
```

In addition, define an objective function that does not pass the `'PolynomialOrder'` name-value pair argument to `fitcsvm` when the value of `'PolynomialOrder'` is `NaN`.

```
fun = @(X)mysvmfun(X,predictors,responce,c)

function objective = mysvmfun(X,predictors,response,c)
    args = {predictors,response, ...
        'CVPartition',c, ...
        'KernelFunction',X.KernelFunction};
    if ~isnan(X.PolynomialOrder)
        args = [args,{'PolynomialOrder',X.PolynomialOrder}];
    end
    objective = kfoldLoss(fitcsvm(args{:}));
end
```

Coupled Constraints

Coupled constraints are constraints that you can evaluate only by calling the objective function. These constraints can be stochastic or deterministic. Return these constraint values from your objective function in the second argument. See “Bayesian Optimization Objective Functions” on page 10-36.

The objective function returns a numeric vector for the coupled constraints, one entry for each coupled constraint. For each entry, a negative value indicates that the constraint is satisfied (also called feasible). A positive value indicates that the constraint is not satisfied (infeasible).

`bayesopt` automatically creates a coupled constraint, called the Error constraint, for every run. This constraint enables `bayesopt` to model points that cause errors in objective function evaluation. For details, see “Objective Function Errors” on page 10-36 and `predictError`.

If you have coupled constraints in addition to the Error constraint:

- Include the `NumCoupledConstraints` name-value pair in your `bayesopt` call (required). Do not include the Error constraint in this number.
- If any of your coupled constraints are stochastic, include the `AreCoupledConstraintsDeterministic` name-value pair and pass `false` for any stochastic constraint.

Observe the coupled constraint values in each iteration by setting the `bayesopt` `Verbose` name-value pair to `1` or `2`.

Note When there are coupled constraints, iterative display and plot functions can give counterintuitive results such as:

- A *minimum objective* plot can increase.
- The optimization can declare a problem infeasible even when it showed an earlier feasible point.

The reason for this behavior is that the decision about whether a point is feasible can change as the optimization progresses. `bayesopt` determines feasibility with respect to its constraint model, and this model changes as `bayesopt` evaluates points. So a "minimum objective" plot can increase when the minimal point is later deemed infeasible, and the iterative display can show a feasible point that is later deemed infeasible.

For an example, see "Bayesian Optimization with Coupled Constraints" on page 10-41.

Bayesian Optimization with Coupled Constraints

A coupled constraint is one that can be evaluated only by evaluating the objective function. In this case, the objective function is the cross-validated loss of an SVM model. The coupled constraint is that the number of support vectors is no more than 100. The model details are in "Optimize a Cross-Validated SVM Classifier Using bayesopt" on page 10-45.

Create the data for classification.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
redpts = zeros(100,2);
grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
cdata = [grnpts;redpts];
grp = ones(200,1);
grp(101:200) = -1;
c = cvpartition(200,'KFold',10);
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
```

The objective function is the cross-validation loss of the SVM model for partition `c`. The coupled constraint is the number of support vectors minus 100.5. This ensures that 100 support vectors give a negative constraint value, but 101 support vectors give a positive value. The model has 200 data points, so the coupled constraint values range from -99.5 (there is always at least one support vector) to 99.5. Positive values mean the constraint is not satisfied.

```
function [objective,constraint] = mysvmfun(x,cdata,grp,c)
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'BoxConstraint',x.box,...
    'KernelScale',x.sigma);
```

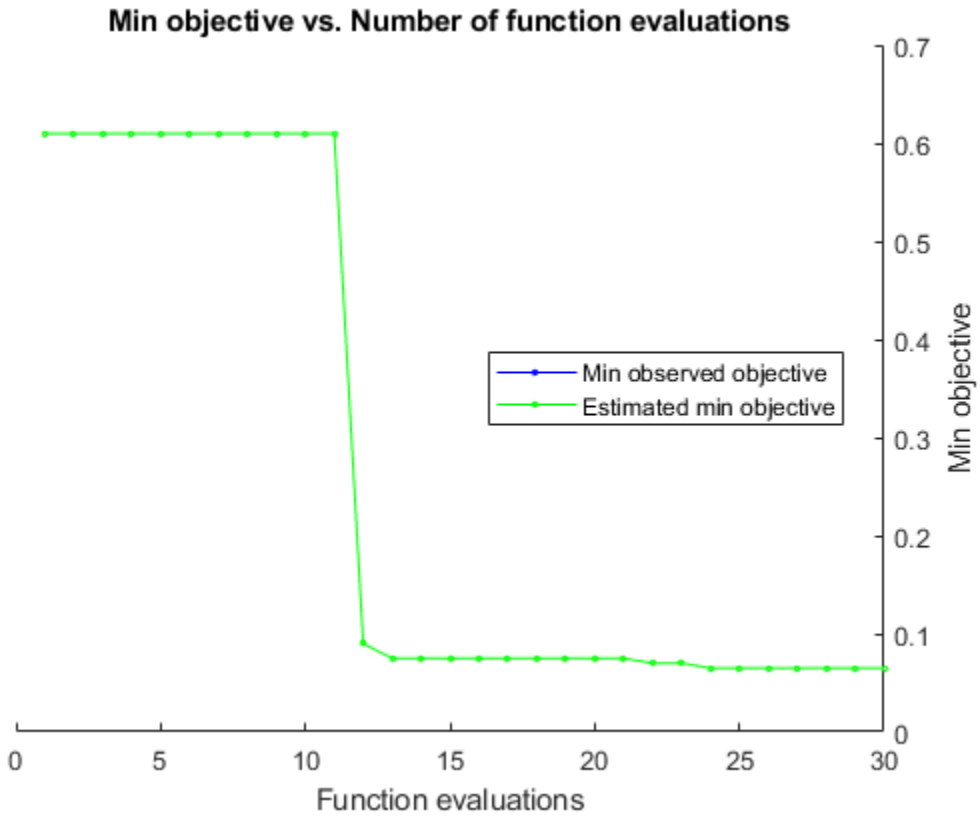
```
cvModel = crossval(SVMModel,'CVPartition',c);
objective = kfoldLoss(cvModel);
constraint = sum(SVMModel.IsSupportVector)-100.5;
```

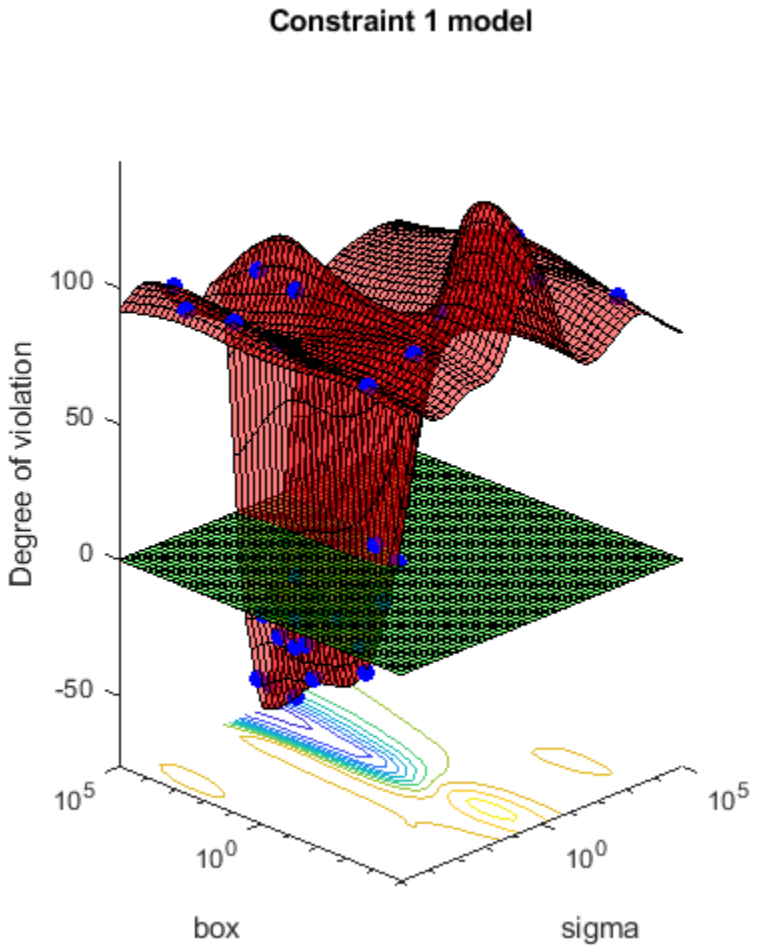
Pass the partition `c` and fitting data `cdata` and `grp` to the objective function `fun` by creating `fun` as an anonymous function that incorporates this data. See “Parameterizing Functions”.

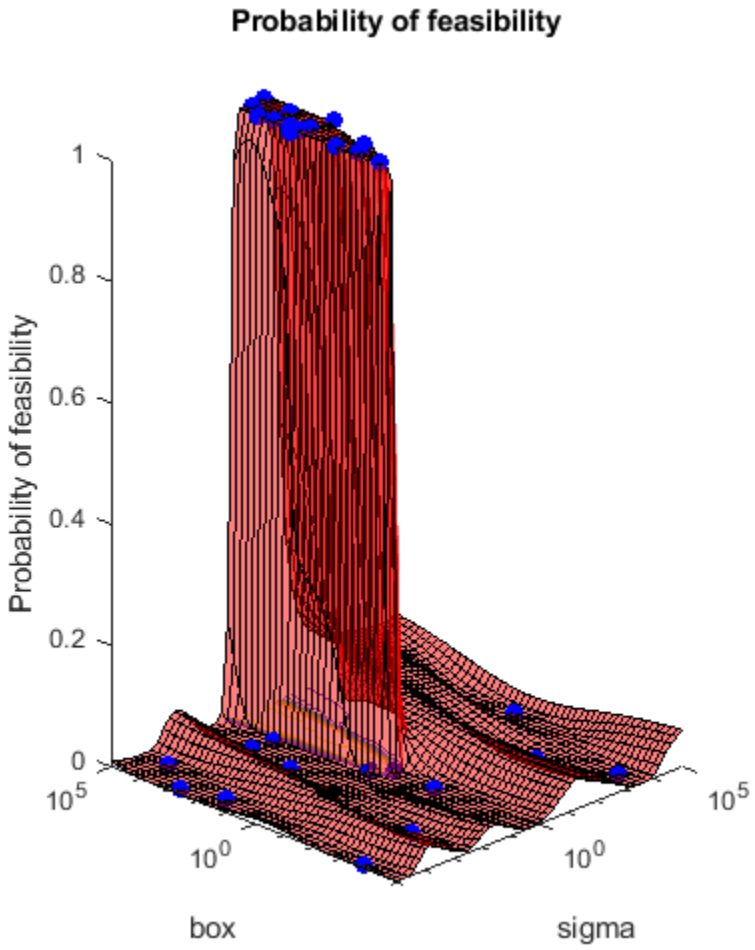
```
fun = @(x)mysvmfun(x,cdata,grp,c);
```

Set the `NumCoupledConstraints` to `1` so the optimizer knows that there is a coupled constraint. Set options to plot the constraint model.

```
results = bayesopt(fun,[sigma,box],'IsObjectiveDeterministic',true,...
    'NumCoupledConstraints',1,'PlotFcn',...
    {@plotMinObjective,@plotConstraintModels},...
    'AcquisitionFunctionName','expected-improvement-plus','Verbose',0);
```




Constraint 1 model
Degree of violation
100
50
0
-50
10^5
10^0
box
10^5
10^0
sigma



Most points lead to an infeasible number of support vectors.

See Also

`bayesopt` | `optimizableVariable`

Related Examples

- “Bayesian Optimization Workflow” on page 10-25
- “Variables for a Bayesian Optimization” on page 10-33
- “Bayesian Optimization Objective Functions” on page 10-36

Optimize a Cross-Validated SVM Classifier Using bayesopt

This example shows how to optimize an SVM classification using the `bayesopt` function. The classification works on locations of points from a Gaussian mixture model. In *The Elements of Statistical Learning*, Hastie, Tibshirani, and Friedman (2009), page 17 describes the model. The model begins with generating 10 base points for a "green" class, distributed as 2-D independent normals with mean (1,0) and unit variance. It also generates 10 base points for a "red" class, distributed as 2-D independent normals with mean (0,1) and unit variance. For each class (green and red), generate 100 random points as follows:

1. Choose a base point *m* of the appropriate color uniformly at random.
2. Generate an independent random point with 2-D normal distribution with mean *m* and variance I/5, where I is the 2-by-2 identity matrix. In this example, use a variance I/50 to show the advantage of optimization more clearly.

After generating 100 green and 100 red points, classify them using `fitcsvm`. Then use `bayesopt` to optimize the parameters of the resulting SVM model with respect to cross validation.

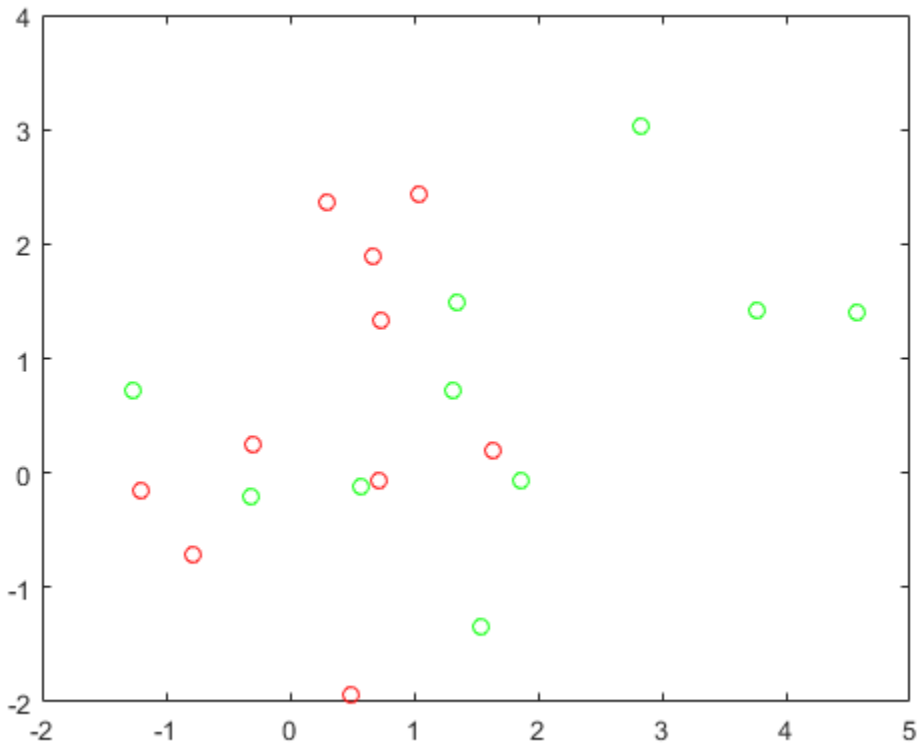
Generate the Points and Classifier

Generate the 10 base points for each class.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
```

View the base points.

```
plot(grnpop(:,1),grnpop(:,2),'go')
hold on
plot(redpop(:,1),redpop(:,2),'ro')
hold off
```



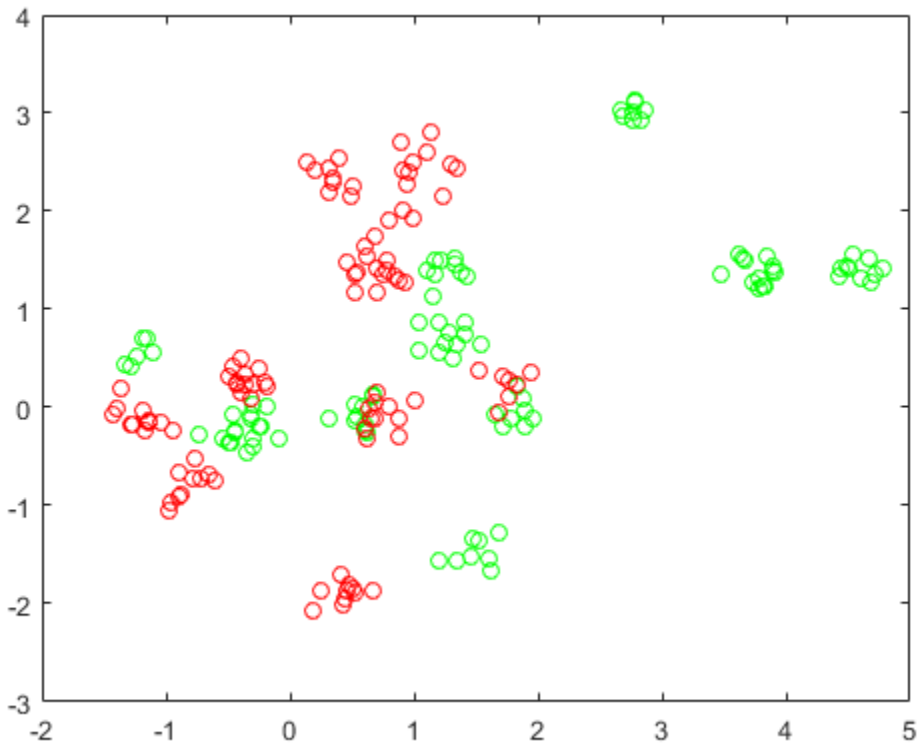
Since some red base points are close to green base points, it can be difficult to classify the data points based on location alone.

Generate the 100 data points of each class.

```
redpts = zeros(100,2);grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
```

View the data points.

```
figure
plot(grnpts(:,1),grnpts(:,2),'go')
hold on
plot(redpts(:,1),redpts(:,2),'ro')
hold off
```



Prepare Data For Classification

Put the data into one matrix, and make a vector `grp` that labels the class of each point.

```
cdata = [grnpts;redpts];
grp = ones(200,1);
% Green label 1, red label -1
grp(101:200) = -1;
```

Prepare Cross-Validation

Set up a partition for cross-validation. This step fixes the train and test sets that the optimization uses at each step.

```
c = cvpartition(200,'KFold',10);
```

Prepare Variables for Bayesian Optimization

Set up a function that takes an input `z = [rbf_sigma,boxconstraint]` and returns the cross-validation loss value of `z`. Take the components of `z` as positive, log-transformed variables between `1e-5` and `1e5`. Choose a wide range, because you don't know which values are likely to be good.

```
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
```

Objective Function

This function handle computes the cross-validation loss at parameters `[sigma,box]`. For details, see `kfoldLoss`.

`bayesopt` passes the variable `z` to the objective function as a one-row table.

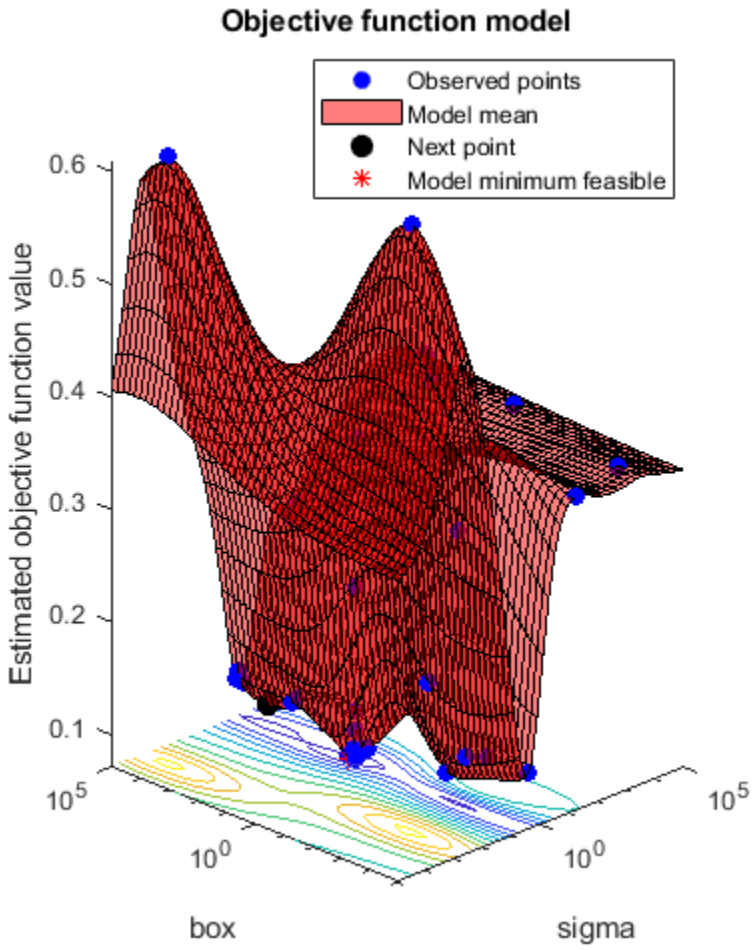
```
minfn = @(z)kfoldLoss(fitcsvm(cdata,grp,'CVPartition',c,...
    'KernelFunction','rbf','BoxConstraint',z.box,...
    'KernelScale',z.sigma));
```

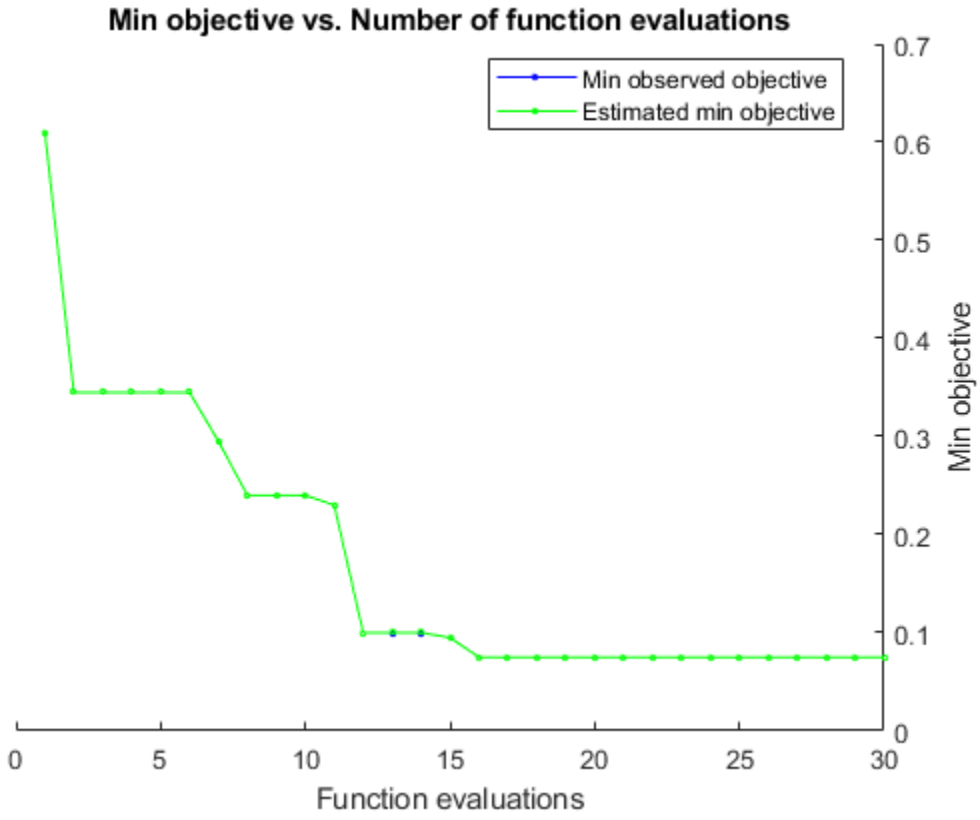
Optimize Classifier

Search for the best parameters `[sigma,box]` using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time, and so can give varying results.

```
results = bayesopt(minfn,[sigma,box],'IsObjectiveDeterministic',true,...
    'AcquisitionFunctionName','expected-improvement-plus')
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        sigma |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |        0.61 |     0.26893 |        0.61 |        0.61 |   0.00013375 |
|    2 | Best   |       0.345 |     0.22004 |       0.345 |       0.345 |        24526 |
|    3 | Accept |        0.61 |     0.28187 |       0.345 |       0.345 |    0.0026459 |   0.0008
|    4 | Accept |       0.345 |     0.22345 |       0.345 |       0.345 |       3506.3 |   6.742
|    5 | Accept |       0.345 |     0.23648 |       0.345 |       0.345 |       9135.2 |       5
|    6 | Accept |       0.345 |     0.21083 |       0.345 |       0.345 |        99701 |
|    7 | Best   |       0.295 |     0.23303 |       0.295 |       0.295 |       455.88 |       9
|    8 | Best   |        0.24 |      2.2293 |        0.24 |        0.24 |        31.56 |
|    9 | Accept |        0.24 |      2.5065 |        0.24 |        0.24 |       10.451 |
|   10 | Accept |        0.35 |      0.1752 |        0.24 |        0.24 |       17.331 |   1.026
|   11 | Best   |        0.23 |      1.3391 |        0.23 |        0.23 |       16.005 |
|   12 | Best   |         0.1 |     0.22906 |         0.1 |         0.1 |      0.36562 |
|   13 | Accept |       0.115 |     0.17834 |         0.1 |         0.1 |       0.1793 |
|   14 | Accept |       0.105 |      0.2158 |         0.1 |         0.1 |       0.2267 |
|   15 | Best   |       0.095 |     0.23586 |       0.095 |       0.095 |       0.28999 |    0.00
|   16 | Best   |       0.075 |     0.22547 |       0.075 |       0.075 |       0.30554 |       8
|   17 | Accept |       0.085 |     0.21066 |       0.075 |       0.075 |       0.41122 |       4
|   18 | Accept |       0.085 |     0.21961 |       0.075 |       0.075 |       0.25565 |       7
|   19 | Accept |       0.075 |     0.22714 |       0.075 |       0.075 |       0.32869 |      1
|   20 | Accept |       0.085 |     0.21682 |       0.075 |       0.075 |       0.32442 |       5
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        sigma |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|   21 | Accept |         0.3 |     0.23247 |       0.075 |       0.075 |       1.3592 |    0.00
|   22 | Accept |        0.12 |     0.21034 |       0.075 |       0.075 |      0.17515 |   0.000
|   23 | Accept |       0.175 |     0.22285 |       0.075 |       0.075 |       0.1252 |     0.0
|   24 | Accept |       0.105 |     0.21393 |       0.075 |       0.075 |       1.1664 |
|   25 | Accept |         0.1 |     0.23242 |       0.075 |       0.075 |      0.57465 |      2
|   26 | Accept |        0.12 |      0.2238 |       0.075 |       0.075 |      0.42922 |   1.160
|   27 | Accept |        0.12 |     0.23202 |       0.075 |       0.075 |      0.42956 |   0.000
|   28 | Accept |       0.095 |     0.20072 |       0.075 |       0.075 |       0.4806 |      1
|   29 | Accept |       0.105 |     0.23571 |       0.075 |       0.075 |      0.19755 |      9
|   30 | Accept |       0.205 |      0.2403 |       0.075 |       0.075 |       3.5051 |      9
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.6
0.5
0.4
0.3
0.2
0.1
10^5
10^0
box
10^5
10^0
sigma




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 54.4117 seconds
Total objective function evaluation time: 12.1281

Best observed feasible point:
     sigma      box
    _______    ______

    0.30554    8.9017

Observed objective function value = 0.075
Estimated objective function value = 0.075
Function evaluation time = 0.22547

Best estimated feasible point (according to models):
     sigma      box
    _______    ______

    0.32869    18.076

Estimated objective function value = 0.075
Estimated function evaluation time = 0.21877

results =
  BayesianOptimization with properties:
```

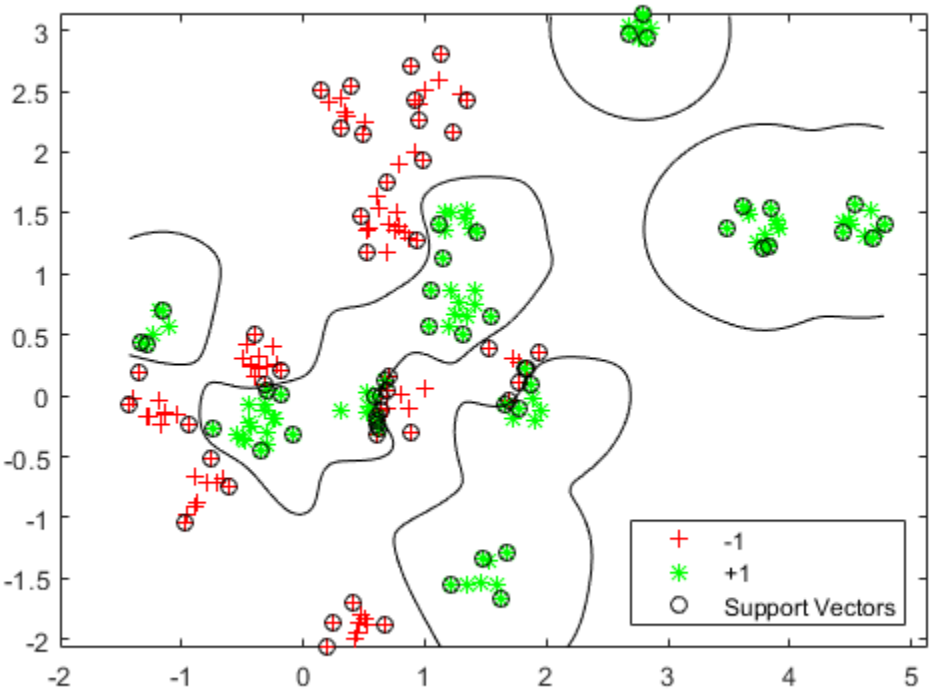
```
                    ObjectiveFcn: [function_handle]
            VariableDescriptions: [1x2 optimizableVariable]
                         Options: [1x1 struct]
                    MinObjective: 0.0750
                 XAtMinObjective: [1x2 table]
           MinEstimatedObjective: 0.0750
        XAtMinEstimatedObjective: [1x2 table]
         NumObjectiveEvaluations: 30
                TotalElapsedTime: 54.4117
                       NextPoint: [1x2 table]
                          XTrace: [30x2 table]
                  ObjectiveTrace: [30x1 double]
                ConstraintsTrace: []
                   UserDataTrace: {30x1 cell}
     ObjectiveEvaluationTimeTrace: [30x1 double]
              IterationTimeTrace: [30x1 double]
                      ErrorTrace: [30x1 double]
                FeasibilityTrace: [30x1 logical]
      FeasibilityProbabilityTrace: [30x1 double]
             IndexOfMinimumTrace: [30x1 double]
           ObjectiveMinimumTrace: [30x1 double]
EstimatedObjectiveMinimumTrace: [30x1 double]
```

Use the results to train a new, optimized SVM classifier.

```
z(1) = results.XAtMinObjective.sigma;
z(2) = results.XAtMinObjective.box;
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'KernelScale',z(1),'BoxConstraint',z(2));
```

Plot the classification boundaries. To visualize the support vector classifier, predict scores over a grid.

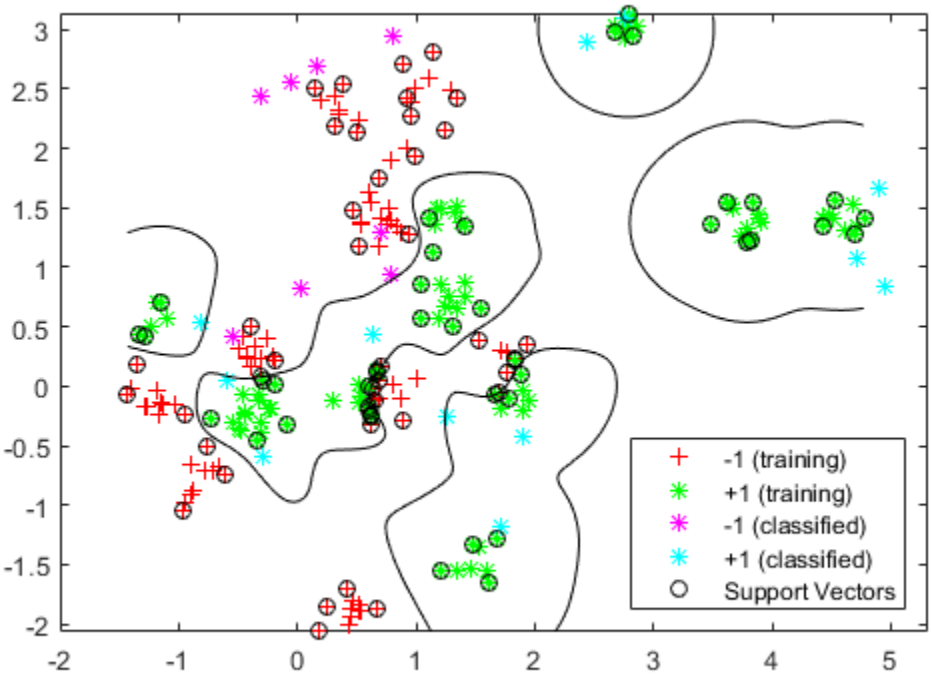
```
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(cdata(:,1)):d:max(cdata(:,1)),...
    min(cdata(:,2)):d:max(cdata(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];
[~,scores] = predict(SVMModel,xGrid);

h = nan(3,1); % Preallocation
figure;
h(1:2) = gscatter(cdata(:,1),cdata(:,2),grp,'rg','+*');
hold on
h(3) = plot(cdata(SVMModel.IsSupportVector,1),...
    cdata(SVMModel.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');
legend(h,{'-1','+1','Support Vectors'},'Location','Southeast');
axis equal
hold off
```



Evaluate Accuracy on New Data

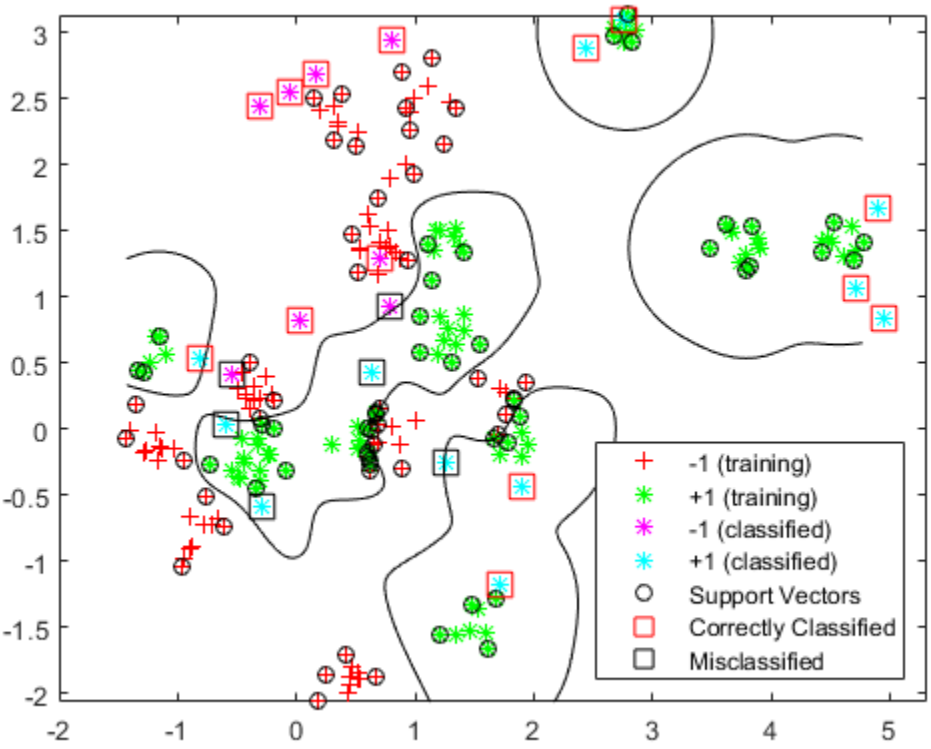
Generate and classify some new data points.

```
grnobj = gmdistribution(grnpop,.2*eye(2));
redobj = gmdistribution(redpop,.2*eye(2));

newData = random(grnobj,10);
newData = [newData;random(redobj,10)];
grpData = ones(20,1);
grpData(11:20) = -1; % red = -1

v = predict(SVMModel,newData);

g = nan(7,1);
figure;
h(1:2) = gscatter(cdata(:,1),cdata(:,2),grp,'rg','+*');
hold on
h(3:4) = gscatter(newData(:,1),newData(:,2),v,'mc','**');
h(5) = plot(cdata(SVMModel.IsSupportVector,1),...
    cdata(SVMModel.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');
legend(h(1:5),{'-1 (training)','+1 (training)','-1 (classified)',...
    '+1 (classified)','Support Vectors'},'Location','Southeast');
axis equal
hold off
```



See which new data points are correctly classified. Circle the correctly classified points in red, and the incorrectly classified points in black.

```
mydiff = (v == grpData); % Classified correctly
figure;
h(1:2) = gscatter(cdata(:,1),cdata(:,2),grp,'rg','+*');
hold on
h(3:4) = gscatter(newData(:,1),newData(:,2),v,'mc','**');
h(5) = plot(cdata(SVMModel.IsSupportVector,1),...
    cdata(SVMModel.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');

for ii = mydiff % Plot red squares around correct pts
    h(6) = plot(newData(ii,1),newData(ii,2),'rs','MarkerSize',12);
end

for ii = not(mydiff) % Plot black squares around incorrect pts
    h(7) = plot(newData(ii,1),newData(ii,2),'ks','MarkerSize',12);
end
legend(h,{'-1 (training)','+1 (training)','-1 (classified)',...
    '+1 (classified)','Support Vectors','Correctly Classified',...
    'Misclassified'},'Location','Southeast');
hold off
```



See Also

`bayesopt` | `optimizableVariable`

Related Examples

- “Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 10-55

Optimize an SVM Classifier Fit Using Bayesian Optimization

This example shows how to optimize an SVM classification using the `fitcsvm` function and `OptimizeHyperparameters` name-value pair. The classification works on locations of points from a Gaussian mixture model. In *The Elements of Statistical Learning*, Hastie, Tibshirani, and Friedman (2009), page 17 describes the model. The model begins with generating 10 base points for a "green" class, distributed as 2-D independent normals with mean (1,0) and unit variance. It also generates 10 base points for a "red" class, distributed as 2-D independent normals with mean (0,1) and unit variance. For each class (green and red), generate 100 random points as follows:

1. Choose a base point *m* of the appropriate color uniformly at random.
2. Generate an independent random point with 2-D normal distribution with mean *m* and variance I/5, where I is the 2-by-2 identity matrix. In this example, use a variance I/50 to show the advantage of optimization more clearly.

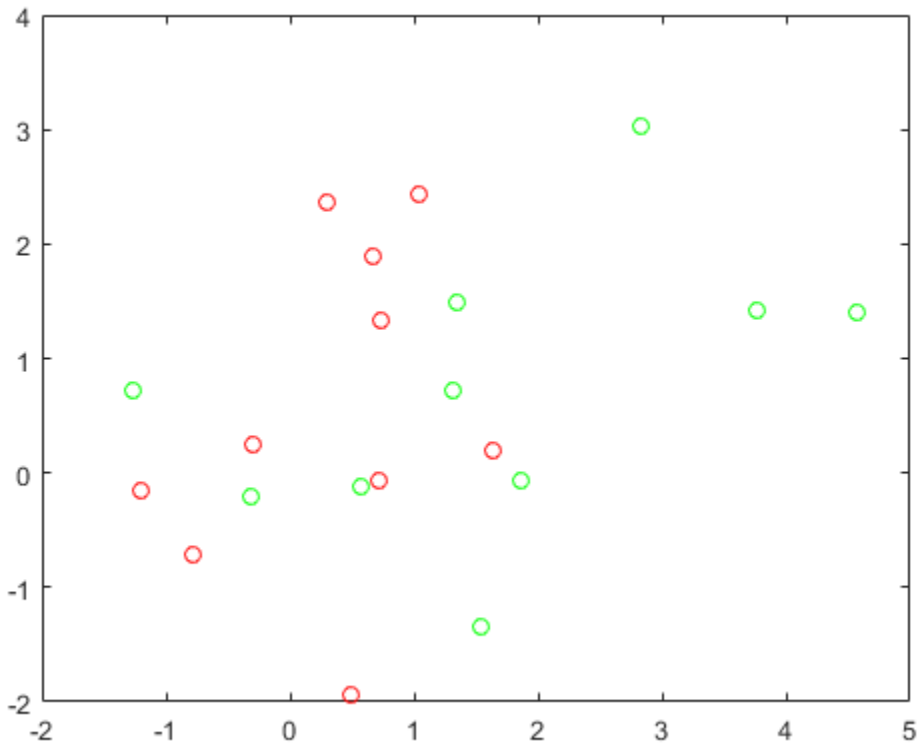
Generate the Points and Classifier

Generate the 10 base points for each class.

```
rng default % For reproducibility
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
```

View the base points.

```
plot(grnpop(:,1),grnpop(:,2),'go')
hold on
plot(redpop(:,1),redpop(:,2),'ro')
hold off
```



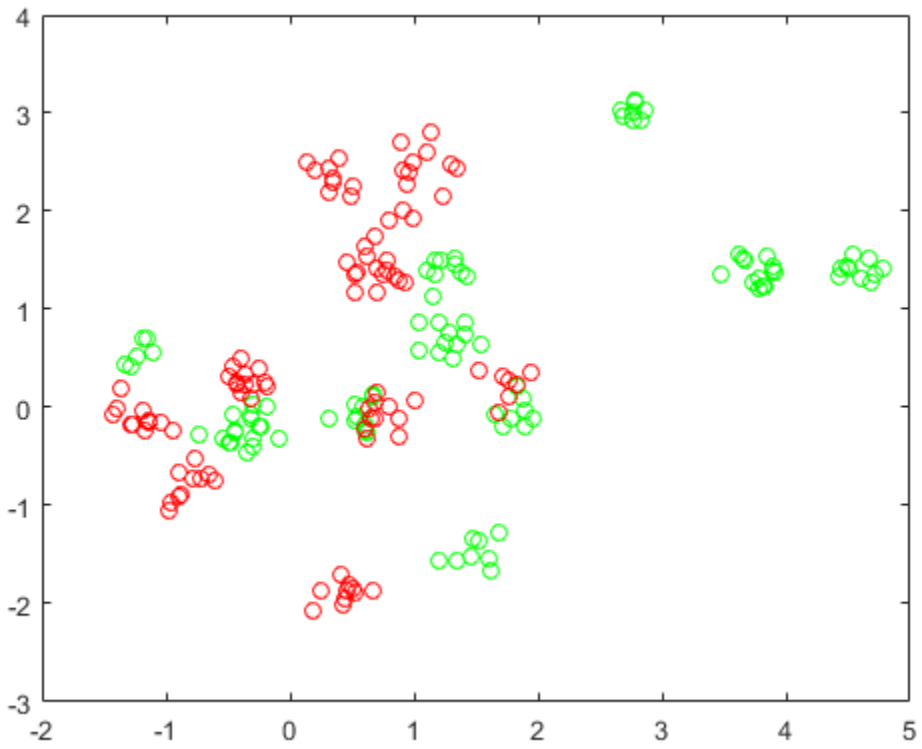
Since some red base points are close to green base points, it can be difficult to classify the data points based on location alone.

Generate the 100 data points of each class.

```
redpts = zeros(100,2);grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
```

View the data points.

```
figure
plot(grnpts(:,1),grnpts(:,2),'go')
hold on
plot(redpts(:,1),redpts(:,2),'ro')
hold off
```



Prepare Data For Classification

Put the data into one matrix, and make a vector `grp` that labels the class of each point.

```
cdata = [grnpts;redpts];
grp = ones(200,1);
% Green label 1, red label -1
grp(101:200) = -1;
```

Prepare Cross-Validation

Set up a partition for cross-validation. This step fixes the train and test sets that the optimization uses at each step.

```
c = cvpartition(200,'KFold',10);
```

Optimize the Fit

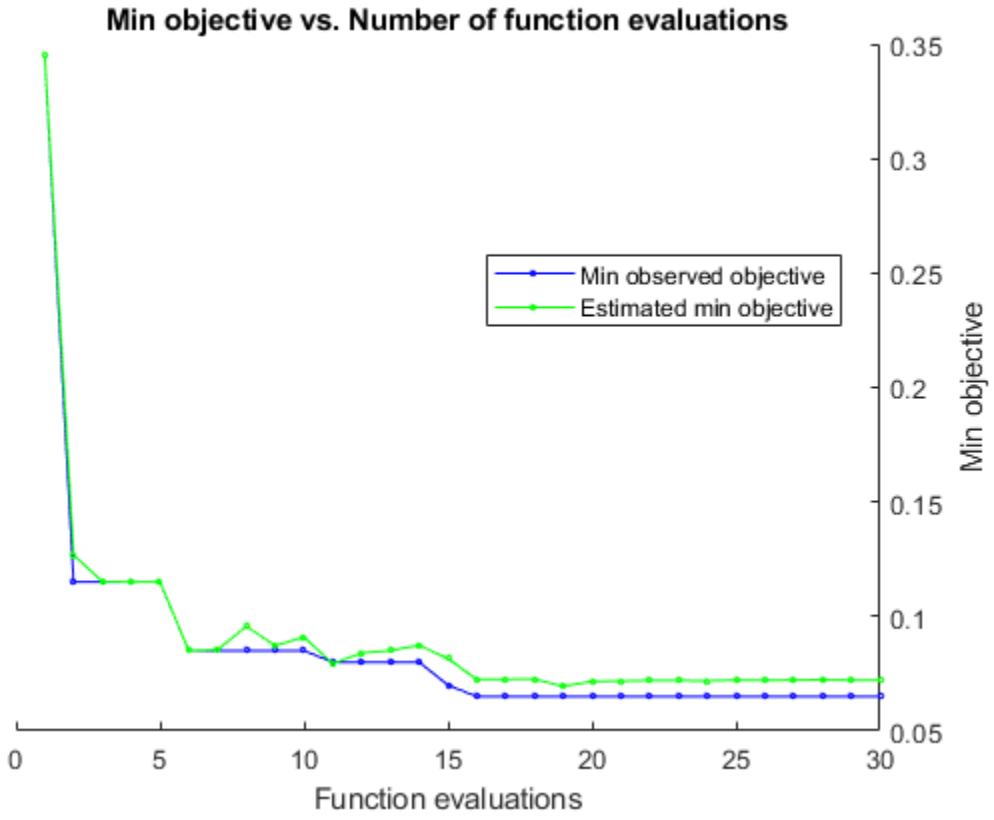
To find a good fit, meaning one with a low cross-validation loss, set options to use Bayesian optimization. Use the same cross-validation partition `c` in all optimizations.

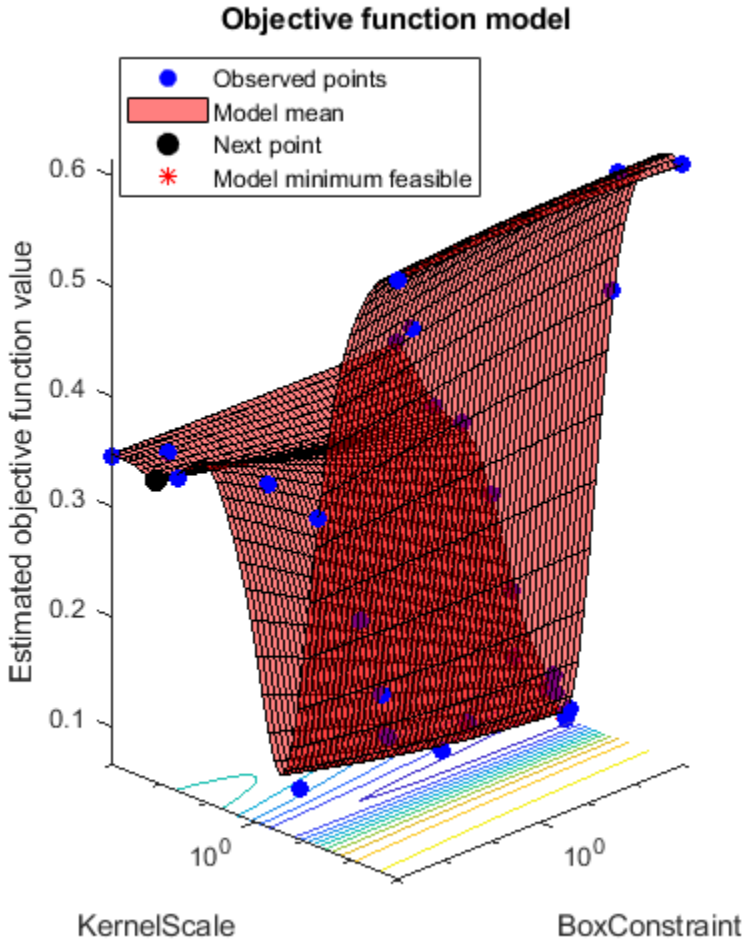
For reproducibility, use the `'expected-improvement-plus'` acquisition function.

```
opts = struct('Optimizer','bayesopt','ShowPlots',true,'CVPartition',c,...
    'AcquisitionFunctionName','expected-improvement-plus');
svmmod = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions',opts)
```

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	BoxConstraint	KernelS
1	Best	0.345	0.32188	0.345	0.345	0.00474	30
2	Best	0.115	0.21093	0.115	0.12678	430.31	1
3	Accept	0.52	0.32736	0.115	0.1152	0.028415	0.01
4	Accept	0.61	0.30316	0.115	0.11504	133.94	0.003
5	Accept	0.34	0.23884	0.115	0.11504	0.010993	5
6	Best	0.085	0.25666	0.085	0.085039	885.63	0.6
7	Accept	0.105	0.2523	0.085	0.085428	0.3057	0.5
8	Accept	0.21	0.24205	0.085	0.09566	0.16044	0.9
9	Accept	0.085	0.33802	0.085	0.08725	972.19	0.4
10	Accept	0.1	0.28635	0.085	0.090952	990.29	0
11	Best	0.08	0.24828	0.08	0.079362	2.5195	0
12	Accept	0.09	0.29983	0.08	0.08402	14.338	0.4
13	Accept	0.1	0.24832	0.08	0.08508	0.0022577	0.2
14	Accept	0.11	0.2402	0.08	0.087378	0.2115	0.3
15	Best	0.07	0.2857	0.07	0.081507	910.2	0.2
16	Best	0.065	0.29082	0.065	0.072457	953.22	0.2
17	Accept	0.075	0.26319	0.065	0.072554	998.74	0.2
18	Accept	0.295	0.26722	0.065	0.072647	996.18	44
19	Accept	0.07	0.29055	0.065	0.06946	985.37	0.2
20	Accept	0.165	0.24204	0.065	0.071622	0.065103	0.1

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	BoxConstraint	KernelS
21	Accept	0.345	0.21508	0.065	0.071764	971.7	99
22	Accept	0.61	0.24321	0.065	0.071967	0.0010168	0.001
23	Accept	0.345	0.25085	0.065	0.071959	0.0010674	99
24	Accept	0.35	0.19215	0.065	0.071863	0.0010003	40
25	Accept	0.24	0.31973	0.065	0.072124	996.55	10
26	Accept	0.61	0.27511	0.065	0.072068	958.64	0.001
27	Accept	0.47	0.25386	0.065	0.07218	993.69	0.02
28	Accept	0.3	0.23806	0.065	0.072291	993.15	17
29	Accept	0.16	0.52113	0.065	0.072104	992.81	3
30	Accept	0.365	0.20023	0.065	0.072112	0.0010017	0.04


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25
30




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 68.9913 seconds
Total objective function evaluation time: 8.1631

Best observed feasible point:
    BoxConstraint    KernelScale
    _____________    ___________

       953.22          0.26253

Observed objective function value = 0.065
Estimated objective function value = 0.073726
Function evaluation time = 0.29082

Best estimated feasible point (according to models):
    BoxConstraint    KernelScale
    _____________    ___________
```

```
       985.37          0.27389

Estimated objective function value = 0.072112
Estimated function evaluation time = 0.29025

svmmod =
  ClassificationSVM
                           ResponseName: 'Y'
                  CategoricalPredictors: []
                             ClassNames: [-1 1]
                         ScoreTransform: 'none'
                        NumObservations: 200
      HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                                  Alpha: [77x1 double]
                                   Bias: -0.2352
                        KernelParameters: [1x1 struct]
                          BoxConstraints: [200x1 double]
                         ConvergenceInfo: [1x1 struct]
                         IsSupportVector: [200x1 logical]
                                  Solver: 'SMO'


  Properties, Methods
```

Find the loss of the optimized model.

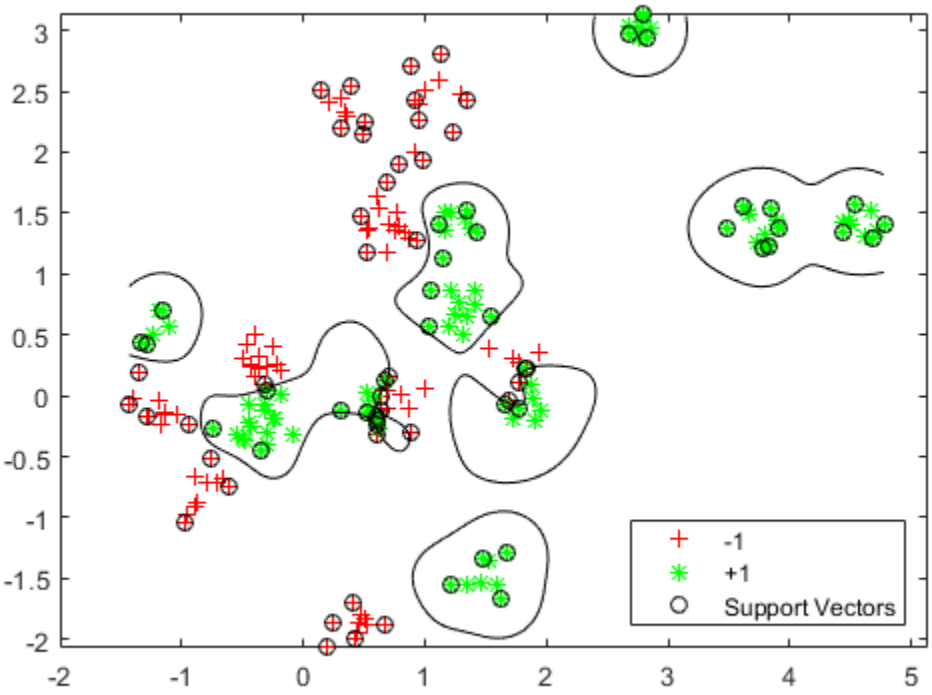
```
lossnew = kfoldLoss(fitcsvm(cdata,grp,'CVPartition',c,'KernelFunction','rbf',...
    'BoxConstraint',svmmod.HyperparameterOptimizationResults.XAtMinObjective.BoxConstraint,...
    'KernelScale',svmmod.HyperparameterOptimizationResults.XAtMinObjective.KernelScale))
```

```
lossnew = 0.0650
```

This loss is the same as the loss reported in the optimization output under "Observed objective function value".

Visualize the optimized classifier.

```
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(cdata(:,1)):d:max(cdata(:,1)),...
    min(cdata(:,2)):d:max(cdata(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];
[~,scores] = predict(svmmod,xGrid);
figure;
h = nan(3,1); % Preallocation
h(1:2) = gscatter(cdata(:,1),cdata(:,2),grp,'rg','+*');
hold on
h(3) = plot(cdata(svmmod.IsSupportVector,1),...
    cdata(svmmod.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');
legend(h,{'-1','+1','Support Vectors'},'Location','Southeast');
axis equal
hold off
```



See Also

`bayesopt` | `fitcsvm`

Related Examples

- “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45

Optimize a Boosted Regression Ensemble

This example shows how to optimize hyperparameters of a boosted regression ensemble. The optimization minimizes the cross-validation loss of the model.

The problem is to model the efficiency in miles per gallon of an automobile, based on its acceleration, engine displacement, horsepower, and weight. Load the `carsmall` data, which contains these and other predictors.

```
load carsmall
X = [Acceleration Displacement Horsepower Weight];
Y = MPG;
```

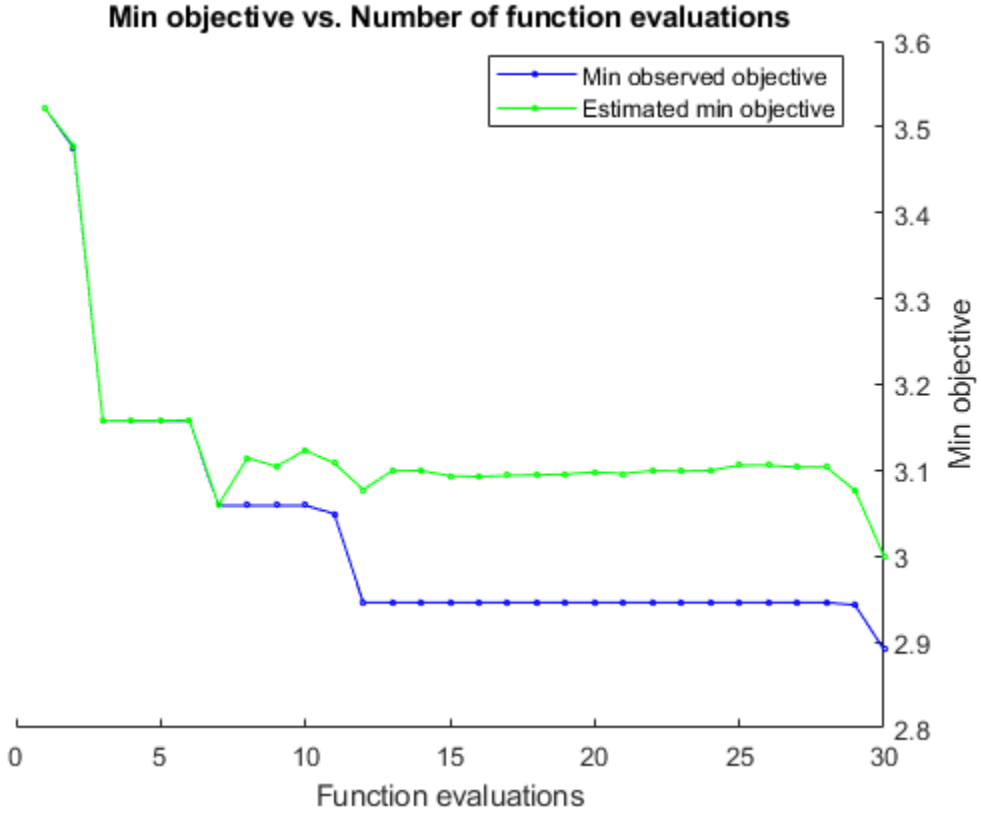
Fit a regression ensemble to the data using the `LSBoost` algorithm, and using surrogate splits. Optimize the resulting model by varying the number of learning cycles, the maximum number of surrogate splits, and the learn rate. Furthermore, allow the optimization to repartition the cross-validation between every iteration.

For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng('default')
Mdl = fitrensemble(X,Y, ...
    'Method','LSBoost', ...
    'Learner',templateTree('Surrogate','on'), ...
    'OptimizeHyperparameters',{'NumLearningCycles','MaxNumSplits','LearnRate'}, ...
    'HyperparameterOptimizationOptions',struct('Repartition',true, ...
    'AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    | NumLearningC-|    Learn |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     | ycles        |          |
|=====================================================================================================|
|    1 | Best   |       3.5219 |       23.163 |       3.5219 |       3.5219 |          383 |      0.5 |
|    2 | Best   |       3.4752 |        1.489 |       3.4752 |       3.4777 |           16 |      0.6 |
|    3 | Best   |       3.1575 |       2.6021 |       3.1575 |       3.1575 |           33 |        0 |
|    4 | Accept |       6.3076 |      0.99623 |       3.1575 |       3.1579 |           13 |    0.005 |
|    5 | Accept |       3.4449 |       17.308 |       3.1575 |       3.1579 |          277 |      0.4 |
|    6 | Accept |       3.9806 |      0.78612 |       3.1575 |       3.1584 |           10 |      0.1 |
|    7 | Best   |        3.059 |      0.78378 |        3.059 |         3.06 |           10 |      0.3 |
|    8 | Accept |       3.1707 |      0.85668 |        3.059 |       3.1144 |           10 |      0.2 |
|    9 | Accept |       3.0937 |      0.89769 |        3.059 |       3.1046 |           10 |      0.3 |
|   10 | Accept |        3.196 |      0.91429 |        3.059 |       3.1233 |           10 |      0.3 |
|   11 | Best   |       3.0495 |      0.82826 |       3.0495 |       3.1083 |           10 |      0.2 |
|   12 | Best   |        2.946 |      0.80825 |        2.946 |       3.0774 |           10 |      0.2 |
|   13 | Accept |       3.2026 |      0.81735 |        2.946 |       3.0995 |           10 |      0.2 |
|   14 | Accept |        5.595 |       26.739 |        2.946 |       3.0996 |          440 |    0.001 |
|   15 | Accept |       3.1976 |       30.757 |        2.946 |       3.0935 |          496 |     0.02 |
|   16 | Accept |       3.9809 |       2.4105 |        2.946 |       3.0927 |           34 |     0.04 |
|   17 | Accept |       3.0512 |       24.393 |        2.946 |       3.0939 |          428 |     0.01 |
|   18 | Accept |       3.4832 |       10.251 |        2.946 |       3.0946 |          205 |      0.9 |
|   19 | Accept |       3.3389 |       4.5493 |        2.946 |       3.0956 |           95 |     0.02 |
|   20 | Accept |       3.2818 |        26.98 |        2.946 |       3.0979 |          494 |     0.02 |
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    | NumLearningC-|    Learn |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     | ycles        |          |
|=====================================================================================================|
```

```
|   21 | Accept |      3.4367 |      23.075 |       2.946 |      3.0962 |          480 |      0.2
|   22 | Accept |      6.2247 |     0.80539 |       2.946 |      3.0995 |           10 |     0.01
|   23 | Accept |      3.2847 |      9.8245 |       2.946 |      3.0991 |          181 |     0.05
|   24 | Accept |       3.142 |      11.168 |       2.946 |      3.0997 |          222 |     0.02
|   25 | Accept |      3.2174 |      1.0394 |       2.946 |       3.106 |           18 |      0.3
|   26 | Accept |       3.064 |      5.0358 |       2.946 |      3.1057 |          108 |      0.1
|   27 | Accept |      3.4532 |      5.2689 |       2.946 |      3.1038 |           93 |      0.2
|   28 | Accept |      3.1992 |      12.674 |       2.946 |      3.1038 |          252 |     0.02
|   29 | Best   |      2.9432 |     0.68054 |      2.9432 |      3.0766 |           10 |      0.3
|   30 | Best   |       2.891 |     0.72016 |       2.891 |           3 |           10 |      0.3
```




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 331.8243 seconds
Total objective function evaluation time: 248.6229

Best observed feasible point:
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           10              0.38339          2

Observed objective function value = 2.891
Estimated objective function value = 2.9674
Function evaluation time = 0.72016
```

```
Best estimated feasible point (according to models):
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           10             0.30126          3

Estimated objective function value = 3
Estimated function evaluation time = 0.80564

Mdl =
  RegressionEnsemble
                          ResponseName: 'Y'
                 CategoricalPredictors: []
                     ResponseTransform: 'none'
                       NumObservations: 94
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                             NumTrained: 10
                                 Method: 'LSBoost'
                          LearnerNames: {'Tree'}
                  ReasonForTermination: 'Terminated normally after completing the requested numbe
                               FitInfo: [10x1 double]
                    FitInfoDescription: {2x1 cell}
                        Regularization: []


  Properties, Methods
```

Compare the loss to that of a boosted, unoptimized model, and to that of the default ensemble.

```
loss = kfoldLoss(crossval(Mdl,'kfold',10))
```

```
loss = 20.6082
```

```
Mdl2 = fitrensemble(X,Y, ...
    'Method','LSBoost', ...
    'Learner',templateTree('Surrogate','on'));
loss2 = kfoldLoss(crossval(Mdl2,'kfold',10))
```

```
loss2 = 36.4539
```

```
Mdl3 = fitrensemble(X,Y);
loss3 = kfoldLoss(crossval(Mdl3,'kfold',10))
```

```
loss3 = 36.6756
```

For a different way of optimizing this ensemble, see “Optimize Regression Ensemble Using Cross-Validation” on page 33-2317.

Parametric Regression Analysis

Choose a Regression Function

Regression is the process of fitting models to data. The models must have numerical responses. For models with categorical responses, see “Parametric Classification” on page 17-2 or “Supervised Learning Workflow and Algorithms” on page 18-3. The regression process depends on the model. If a model is parametric, regression estimates the parameters from the data. If a model is linear in the parameters, estimation is based on methods from linear algebra that minimize the norm of a residual vector. If a model is nonlinear in the parameters, estimation is based on search methods from optimization that minimize the norm of a residual vector.

This table describes which function to use depending on the type of regression problem.

Model Components	Result of Regression	Function to Use
Continuous or categorical predictors, continuous response, linear model	Fitted model coefficients	`fitlm`. See “Linear Regression” on page 11-9.
Continuous or categorical predictors, continuous response, linear model of unknown complexity	Fitted model and fitted coefficients	`stepwiselm`. See “Stepwise Regression” on page 11-99.
Continuous or categorical predictors, response possibly with restrictions such as nonnegative or integer-valued, generalized linear model	Fitted generalized linear model coefficients	`fitglm` or `stepwiseglm`. See “Generalized Linear Models” on page 12-9.
Continuous predictors with a continuous nonlinear response, parametrized nonlinear model	Fitted nonlinear model coefficients	`fitnlm`. See “Nonlinear Regression” on page 13-2.
Continuous predictors, continuous response, linear model	Set of models from ridge, lasso, or elastic net regression	`lasso` or `ridge`. See “Lasso and Elastic Net” on page 11-112 or “Ridge Regression” on page 11-109.
Correlated continuous predictors, continuous response, linear model	Fitted model and fitted coefficients	`plsregress`. See “Partial Least Squares” on page 11-126.
Continuous or categorical predictors, continuous response, unknown model	Nonparametric model	`fitrtree` or `fitrensemble`.
Categorical predictors only	ANOVA	`anova`, `anova1`, `anova2`, `anovan`.
Continuous predictors, multivariable response, linear model	Fitted multivariate regression model coefficients	`mvregress`
Continuous predictors, continuous response, mixed-effects model	Fitted mixed-effects model coefficients	`nlmefit` or `nlmefitsa`. See “Mixed-Effects Models” on page 13-17.

Update Legacy Code with New Fitting Methods

There are several Statistics and Machine Learning Toolbox functions for performing regression. The following sections describe how to replace calls to older functions to new versions:

regress into fitlm

Previous Syntax:

```
[b,bint,r,rint,stats] = regress(y,X)
```

where X contains a column of ones.

Current Syntax:

```
mdl = fitlm(X,y)
```

where you do not add a column of ones to X.

Equivalent values of the previous outputs:

- `b` — `mdl.Coefficients.Estimate`
- `bint` — `coefCI(mdl)`
- `r` — `mdl.Residuals.Raw`
- `rint` — There is no exact equivalent. Try examining `mdl.Residuals.Studentized` to find outliers.
- `stats` — `mdl` contains various properties that replace components of `stats`.

regstats into fitlm

Previous Syntax:

```
stats = regstats(y,X,model,whichstats)
```

Current Syntax:

```
mdl = fitlm(X,y,model)
```

Obtain statistics from the properties and methods of the `LinearModel` object (`mdl`). For example, see the `mdl.Diagnostics` and `mdl.Residuals` properties.

robustfit into fitlm

Previous Syntax:

```
[b,stats] = robustfit(X,y,wfun,tune,const)
```

Current Syntax:

```
mdl = fitlm(X,y,'robust','on') % bisquare
```

Or to use the *`wfun`* weight and the *`tune`* tuning parameter:

```
opt.RobustWgtFun = 'wfun';
opt.Tune = tune; % optional
mdl = fitlm(X,y,'robust',opt)
```

Obtain statistics from the properties and methods of the `LinearModel` object (`mdl`). For example, see the `mdl.Diagnostics` and `mdl.Residuals` properties.

stepwisefit into stepwiselm

Previous Syntax:

```
[b,se,pval,inmodel,stats,nextstep,history] = stepwisefit(X,y,Name,Value)
```

Current Syntax:

```
mdl = stepwiselm(ds,modelspec,Name,Value)
```

or

```
mdl = stepwiselm(X,y,modelspec,Name,Value)
```

Obtain statistics from the properties and methods of the `LinearModel` object (`mdl`). For example, see the `mdl.Diagnostics` and `mdl.Residuals` properties.

glmfit into fitglm

Previous Syntax:

```
[b,dev,stats] = glmfit(X,y,distr,param1,val1,...)
```

Current Syntax:

```
mdl = fitglm(X,y,distr,...)
```

Obtain statistics from the properties and methods of the `GeneralizedLinearModel` object (`mdl`). For example, the deviance is `mdl.Deviance`, and to compare `mdl` against a constant model, use `devianceTest(mdl)`.

nlinfit into fitnlm

Previous Syntax:

```
[beta,r,J,COVB,mse] = nlinfit(X,y,fun,beta0,options)
```

Current Syntax:

```
mdl = fitnlm(X,y,fun,beta0,'Options',options)
```

Equivalent values of the previous outputs:

- `beta` — `mdl.Coefficients.Estimate`
- `r` — `mdl.Residuals.Raw`
- `covb` — `mdl.CoefficientCovariance`
- `mse` — `mdl.mse`

`mdl` does not provide the Jacobian (`J`) output. The primary purpose of `J` was to pass it into `nlparci` or `nlpredci` to obtain confidence intervals for the estimated coefficients (parameters) or predictions. Obtain those confidence intervals as:

```
parci = coefCI(mdl)
[pred,predci] = predict(mdl)
```

See Also

What Is a Linear Regression Model?

A linear regression model describes the relationship between a *dependent variable*, y, and one or more *independent variables*, X. The dependent variable is also called the *response variable*. Independent variables are also called *explanatory* or *predictor variables*. Continuous predictor variables are also called *covariates*, and categorical predictor variables are also called *factors*. The matrix X of observations on predictor variables is usually called the *design matrix*.

A multiple linear regression model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \cdots, n,$$

where

- y_i is the ith response.
- β_k is the kth coefficient, where β_0 is the constant term in the model. Sometimes, design matrices might include information about the constant term. However, `fitlm` or `stepwiselm` by default includes a constant term in the model, so you must not enter a column of 1s into your design matrix X.
- X_{ij} is the ith observation on the jth predictor variable, $j = 1, \ldots, p$.
- ε_i is the ith noise term, that is, random error.

If a model includes only one predictor variable ($p = 1$), then the model is called a simple linear regression model.

In general, a linear regression model can be a model of the form

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k f_k\left(X_{i1}, X_{i2}, \cdots, X_{ip}\right) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $f(.)$ is a scalar-valued function of the independent variables, X_{ij}s. The functions, $f(X)$, might be in any form including nonlinear functions or polynomials. The linearity, in the linear regression models, refers to the linearity of the coefficients β_k. That is, the response variable, y, is a linear function of the coefficients, β_k.

Some examples of linear models are:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^3 + \beta_4 X_{2i}^2 + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \beta_4 \log X_{3i} + \varepsilon_i$$

The following, however, are not linear models since they are not linear in the unknown coefficients, β_k.

$$\log y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 X_{1i} + \frac{1}{\beta_2 X_{2i}} + e^{\beta_3 X_{1i} X_{2i}} + \varepsilon_i$$

The usual assumptions for linear regression models are:

- The noise terms, ε_i, are uncorrelated.

- The noise terms, ε_i, have independent and identical normal distributions with mean zero and constant variance, σ^2. Thus,

$$\begin{aligned} E(y_i) &= E\left(\sum_{k=0}^{K} \beta_k f_k(X_{i1}, X_{i2}, \cdots, X_{ip}) + \varepsilon_i\right) \\ &= \sum_{k=0}^{K} \beta_k f_k(X_{i1}, X_{i2}, \cdots, X_{ip}) + E(\varepsilon_i) \\ &= \sum_{k=0}^{K} \beta_k f_k(X_{i1}, X_{i2}, \cdots, X_{ip}) \end{aligned}$$

  and

$$V(y_i) = V\left(\sum_{k=0}^{K} \beta_k f_k(X_{i1}, X_{i2}, \cdots, X_{ip}) + \varepsilon_i\right) = V(\varepsilon_i) = \sigma^2$$

  So the variance of y_i is the same for all levels of X_{ij}.
- The responses y_i are uncorrelated.

The fitted linear function is

$$\widehat{y}_i = \sum_{k=0}^{K} b_k f_k(X_{i1}, X_{i2}, \cdots, X_{ip}), \quad i = 1, \cdots, n,$$

where $\widehat{y}_i$ is the estimated response and b_ks are the fitted coefficients. The coefficients are estimated so as to minimize the mean squared difference between the prediction vector $\widehat{y}$ and the true response vector y, that is $\widehat{y} - y$. This method is called the *method of least squares*. Under the assumptions on the noise terms, these coefficients also maximize the likelihood of the prediction vector.

In a linear regression model of the form $y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$, the coefficient β_k expresses the impact of a one-unit change in predictor variable, X_j, on the mean of the response E(y), provided that all other variables are held constant. The sign of the coefficient gives the direction of the effect. For example, if the linear model is E(y) = 1.8 – 2.35X_1 + X_2, then –2.35 indicates a 2.35 unit decrease in the mean response with a one-unit increase in X_1, given X_2 is held constant. If the model is E(y) = 1.1 + 1.5X_1^2 + X_2, the coefficient of X_1^2 indicates a 1.5 unit increase in the mean of Y with a one-unit increase in X_1^2 given all else held constant. However, in the case of E(y) = 1.1 + 2.1X_1 + 1.5X_1^2, it is difficult to interpret the coefficients similarly, since it is not possible to hold X_1 constant when X_1^2 changes or vice versa.

References

[1] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. IRWIN, The McGraw-Hill Companies, Inc., 1996.

[2] Seber, G. A. F. *Linear Regression Analysis*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, Inc., 1977.

See Also

`LinearModel` | `fitlm` | `stepwiselm`

Related Examples

- “Linear Regression” on page 11-9
- “Linear Regression Workflow” on page 11-35
- “Interpret Linear Regression Results” on page 11-50
- “Stepwise Regression” on page 11-99
- “Reduce Outlier Effects Using Robust Regression” on page 11-104

Linear Regression

In this section...

Prepare Data

To begin fitting a regression, put your data into a form that fitting functions expect. All regression techniques begin with input data in an array `X` and response data in a separate vector `y`, or input data in a table or dataset array `tbl` and response data as a column in `tbl`. Each row of the input data represents one observation. Each column represents one predictor (variable).

For a table or dataset array `tbl`, indicate the response variable with the `'ResponseVar'` name-value pair:

```
mdl = fitlm(tbl,'ResponseVar','BloodPressure');
```

The response variable is the last column by default.

You can use numeric categorical predictors. A categorical predictor is one that takes values from a fixed set of possibilities.

- For a numeric array `X`, indicate the categorical predictors using the `'Categorical'` name-value pair. For example, to indicate that predictors `2` and `3` out of six are categorical:

  ```
  mdl = fitlm(X,y,'Categorical',[2,3]);
  % or equivalently
  mdl = fitlm(X,y,'Categorical',logical([0 1 1 0 0 0]));
  ```
- For a table or dataset array `tbl`, fitting functions assume that these data types are categorical:

  - Logical vector
  - Categorical vector
  - Character array
  - String array

  If you want to indicate that a numeric predictor is categorical, use the `'Categorical'` name-value pair.

Represent missing numeric data as `NaN`. To represent missing data for other data types, see “Missing Group Values” on page 2-46.

Dataset Array for Input and Response Data

To create a dataset array from an Excel spreadsheet:

```
ds = dataset('XLSFile','hospital.xls', ...
    'ReadObsNames',true);
```

To create a dataset array from workspace variables:

```
load carsmall
ds = dataset(MPG,Weight);
ds.Year = categorical(Model_Year);
```

Table for Input and Response Data

To create a table from an Excel spreadsheet:

```
tbl = readtable('hospital.xls', ...
    'ReadRowNames',true);
```

To create a table from workspace variables:

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
```

Numeric Matrix for Input Data, Numeric Vector for Response

For example, to create numeric arrays from workspace variables:

```
load carsmall
X = [Weight Horsepower Cylinders Model_Year];
y = MPG;
```

To create numeric arrays from an Excel spreadsheet:

```
[X, Xnames] = xlsread('hospital.xls');
y = X(:,4); % response y is systolic pressure
X(:,4) = []; % remove y from the X matrix
```

Notice that the nonnumeric entries, such as `sex`, do not appear in `X`.

Choose a Fitting Method

There are three ways to fit a model to data:

- “Least-Squares Fit” on page 11-10
- “Robust Fit” on page 11-10
- “Stepwise Fit” on page 11-11

Least-Squares Fit

Use `fitlm` to construct a least-squares fit of a model to the data. This method is best when you are reasonably certain of the model’s form, and mainly need to find its parameters. This method is also useful when you want to explore a few models. The method requires you to examine the data manually to discard outliers, though there are techniques to help (see “Examine Quality and Adjust Fitted Model” on page 11-14).

Robust Fit

Use `fitlm` with the `RobustOpts` name-value pair to create a model that is little affected by outliers. Robust fitting saves you the trouble of manually discarding outliers. However, `step` does not work

with robust fitting. This means that when you use robust fitting, you cannot search stepwise for a good model.

Stepwise Fit

Use `stepwiselm` to find a model, and fit parameters to the model. `stepwiselm` starts from one model, such as a constant, and adds or subtracts terms one at a time, choosing an optimal term each time in a greedy fashion, until it cannot improve further. Use stepwise fitting to find a good model, which is one that has only relevant terms.

The result depends on the starting model. Usually, starting with a constant model leads to a small model. Starting with more terms can lead to a more complex model, but one that has lower mean squared error. See “Compare large and small stepwise models” on page 11-99.

You cannot use robust options along with stepwise fitting. So after a stepwise fit, examine your model for outliers (see “Examine Quality and Adjust Fitted Model” on page 11-14).

Choose a Model or Range of Models

There are several ways of specifying a model for linear regression. Use whichever you find most convenient.

- “Brief Name” on page 11-11
- “Terms Matrix” on page 11-12
- “Formula” on page 11-12

For `fitlm`, the model specification you give is the model that is fit. If you do not give a model specification, the default is `'linear'`.

For `stepwiselm`, the model specification you give is the starting model, which the stepwise procedure tries to improve. If you do not give a model specification, the default starting model is `'constant'`, and the default upper bounding model is `'interactions'`. Change the upper bounding model using the `Upper` name-value pair.

Note There are other ways of selecting models, such as using `lasso`, `lassoglm`, `sequentialfs`, or `plsregress`.

Brief Name

Name	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear terms for each predictor.
`'interactions'`	Model contains an intercept, linear terms, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept, linear terms, and squared terms.
`'quadratic'`	Model contains an intercept, linear terms, interactions, and squared terms.

Name	Model Type
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, etc. Use numerals `0` through `9`. For example, `'poly2111'` has a constant plus all linear and product terms, and also contains terms with predictor 1 squared.

For example, to specify an interaction model using `fitlm` with matrix predictors:

```
mdl = fitlm(X,y,'interactions');
```

To specify a model using `stepwiselm` and a table or dataset array `tbl` of predictors, suppose you want to start from a constant and have a linear model upper bound. Assume the response variable in `tbl` is in the third column.

```
mdl2 = stepwiselm(tbl,'constant', ...
    'Upper','linear','ResponseVar',3);
```

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for a model specification is a character vector or string scalar of the form

`'`*`y ~ terms`*`'`,

- *`y`* is the response name.
- *`terms`* contains
  - Variable names
  - `+` to include the next variable
  - `-` to exclude the next variable
  - `:` to define an interaction, a product of terms

- * to define an interaction and all lower-order terms
- ^ to raise the predictor to a power, exactly as in * repeated, so ^ includes lower order terms as well
- ( ) to group terms

Tip Formulas include a constant (intercept) term by default. To exclude a constant term from the model, include `-1` in the formula.

Examples:

`'y ~ x1 + x2 + x3'` is a three-variable linear model with intercept.
`'y ~ x1 + x2 + x3 - 1'` is a three-variable linear model without intercept.
`'y ~ x1 + x2 + x3 + x2^2'` is a three-variable model with intercept and a `x2^2` term.
`'y ~ x1 + x2^2 + x3'` is the same as the previous example, since `x2^2` includes a `x2` term.
`'y ~ x1 + x2 + x3 + x1:x2'` includes an `x1*x2` term.
`'y ~ x1*x2 + x3'` is the same as the previous example, since `x1*x2 = x1 + x2 + x1:x2`.
`'y ~ x1*x2*x3 - x1:x2:x3'` has all interactions among `x1`, `x2`, and `x3`, except the three-way interaction.
`'y ~ x1*(x2 + x3 + x4)'` has all linear terms, plus products of `x1` with each of the other variables.

For example, to specify an interaction model using `fitlm` with matrix predictors:

```
mdl = fitlm(X,y,'y ~ x1*x2*x3 - x1:x2:x3');
```

To specify a model using `stepwiselm` and a table or dataset array `tbl` of predictors, suppose you want to start from a constant and have a linear model upper bound. Assume the response variable in `tbl` is named `'y'`, and the predictor variables are named `'x1'`, `'x2'`, and `'x3'`.

```
mdl2 = stepwiselm(tbl,'y ~ 1','Upper','y ~ x1 + x2 + x3');
```

Fit Model to Data

The most common optional arguments for fitting:

- For robust regression in `fitlm`, set the `'RobustOpts'` name-value pair to `'on'`.
- Specify an appropriate upper bound model in `stepwiselm`, such as set `'Upper'` to `'linear'`.
- Indicate which variables are categorical using the `'CategoricalVars'` name-value pair. Provide a vector with column numbers, such as `[1 6]` to specify that predictors `1` and `6` are categorical. Alternatively, give a logical vector the same length as the data columns, with a `1` entry indicating that variable is categorical. If there are seven predictors, and predictors `1` and `6` are categorical, specify `logical([1,0,0,0,0,1,0])`.
- For a table or dataset array, specify the response variable using the `'ResponseVar'` name-value pair. The default is the last column in the array.

For example,

```
mdl = fitlm(X,y,'linear', ...
    'RobustOpts','on','CategoricalVars',3);
mdl2 = stepwiselm(tbl,'constant', ...
    'ResponseVar','MPG','Upper','quadratic');
```

Examine Quality and Adjust Fitted Model

After fitting a model, examine the result and make adjustments.

Model Display

A linear regression model shows several diagnostics when you enter its name or enter `disp(mdl)`. This display gives some of the basic information to check whether the fitted model represents the data adequately.

For example, fit a linear model to data constructed with two out of five predictors not present and with no intercept term:

```
X = randn(100,5);
y = X*[1;0;3;0;-1] + randn(100,1);
mdl = fitlm(X,y)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)     0.038164    0.099458     0.38372       0.70205
    x1               0.92794    0.087307      10.628    8.5494e-18
    x2              -0.075593     0.10044    -0.75264       0.45355
    x3                 2.8965    0.099879          29    1.1117e-48
    x4              0.045311     0.10832     0.41831       0.67667
    x5              -0.99708     0.11799     -8.4504     3.593e-13


Number of observations: 100, Error degrees of freedom: 94
Root Mean Squared Error: 0.972
R-squared: 0.93,  Adjusted R-Squared: 0.926
F-statistic vs. constant model: 248, p-value = 1.5e-52
```

Notice that:

- The display contains the estimated values of each coefficient in the `Estimate` column. These values are reasonably near the true values `[0;1;0;3;0;-1]`.
- There is a standard error column for the coefficient estimates.
- The reported `pValue` (which are derived from the *t* statistics (`tStat`) under the assumption of normal errors) for predictors 1, 3, and 5 are extremely small. These are the three predictors that were used to create the response data `y`.
- The `pValue` for `(Intercept)`, `x2` and `x4` are much larger than 0.01. These three predictors were not used to create the response data `y`.
- The display contains R^2, adjusted R^2, and F statistics.

ANOVA

To examine the quality of the fitted model, consult an ANOVA table. For example, use `anova` on a linear model with five predictors:

```
tbl = anova(mdl)
```

```
tbl=6×5 table
              SumSq     DF    MeanSq        F         pValue
             _______    __    _______    _______    __________

    x1        106.62     1     106.62     112.96    8.5494e-18
    x2       0.53464     1    0.53464    0.56646       0.45355
    x3        793.74     1     793.74     840.98    1.1117e-48
    x4       0.16515     1    0.16515    0.17498       0.67667
    x5        67.398     1     67.398      71.41     3.593e-13
    Error     88.719    94    0.94382
```

This table gives somewhat different results than the model display. The table clearly shows that the effects of `x2` and `x4` are not significant. Depending on your goals, consider removing `x2` and `x4` from the model.

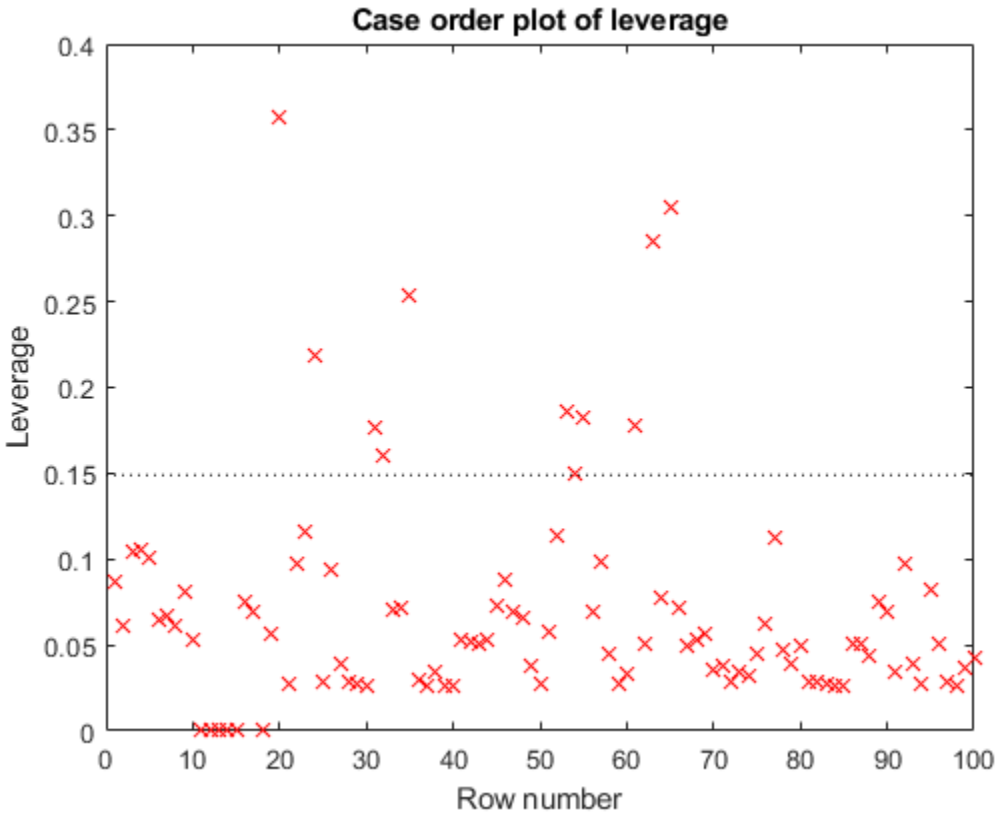
Diagnostic Plots

Diagnostic plots help you identify outliers, and see other problems in your model or fit. For example, load the `carsmall` data, and make a model of `MPG` as a function of `Cylinders` (categorical) and `Weight`:

```
load carsmall
tbl = table(Weight,MPG,Cylinders);
tbl.Cylinders = categorical(tbl.Cylinders);
mdl = fitlm(tbl,'MPG ~ Cylinders*Weight + Weight^2');
```

Make a leverage plot of the data and model.

```
plotDiagnostics(mdl)
```



There are a few points with high leverage. But this plot does not reveal whether the high-leverage points are outliers.

Look for points with large Cook's distance.

```
plotDiagnostics(mdl,'cookd')
```

There is one point with large Cook’s distance. Identify it and remove it from the model. You can use the Data Cursor to click the outlier and identify it, or identify it programmatically:

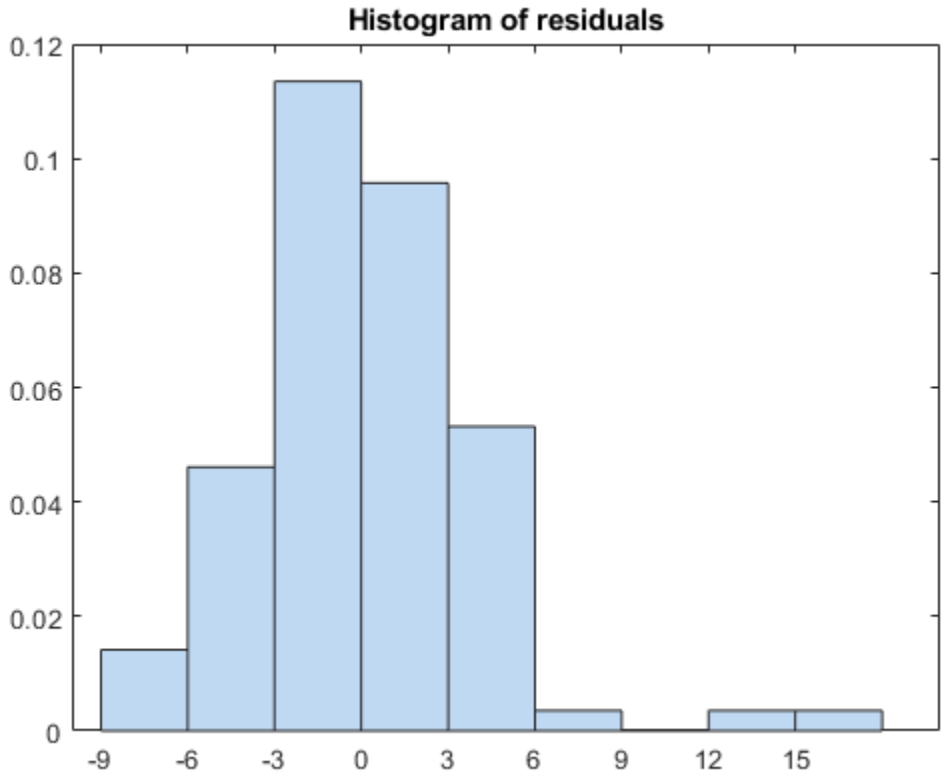
```
[~,larg] = max(mdl.Diagnostics.CooksDistance);
mdl2 = fitlm(tbl,'MPG ~ Cylinders*Weight + Weight^2','Exclude',larg);
```

Residuals — Model Quality for Training Data

There are several residual plots to help you discover errors, outliers, or correlations in the model or data. The simplest residual plots are the default histogram plot, which shows the range of the residuals and their frequencies, and the probability plot, which shows how the distribution of the residuals compares to a normal distribution with matched variance.

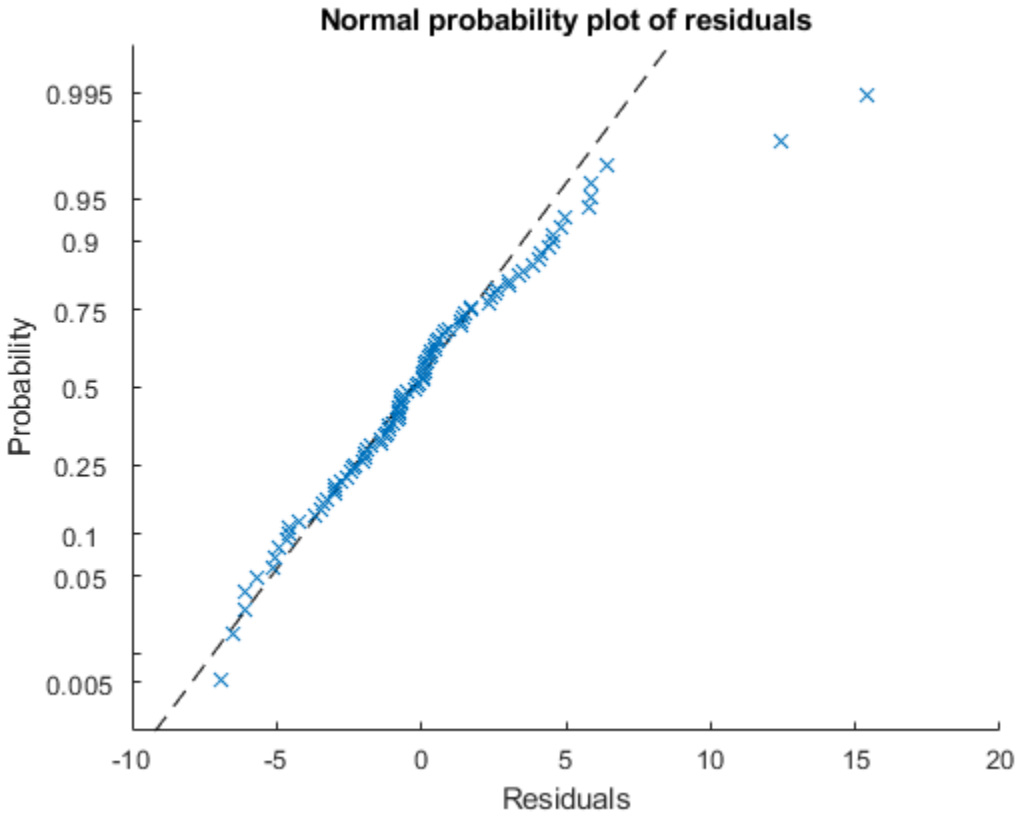
Examine the residuals:

```
plotResiduals(mdl)
```



The observations above 12 are potential outliers.

```
plotResiduals(mdl,'probability')
```



The two potential outliers appear on this plot as well. Otherwise, the probability plot seems reasonably straight, meaning a reasonable fit to normally distributed residuals.

You can identify the two outliers and remove them from the data:

```
outl = find(mdl.Residuals.Raw > 12)
```

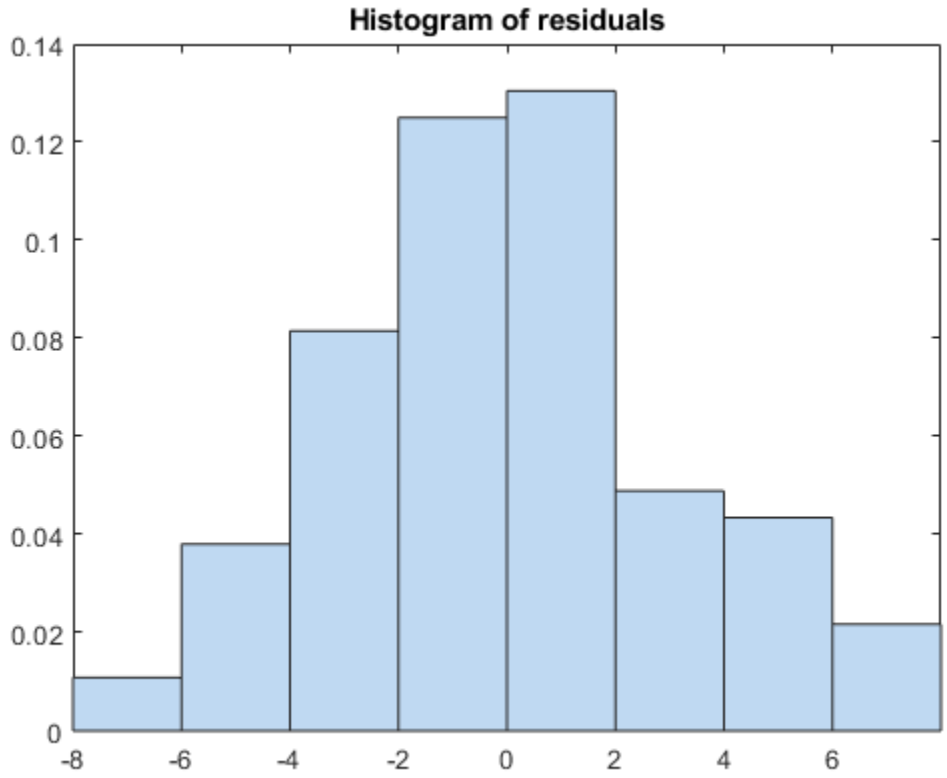
```
outl = 2×1

    90
    97
```

To remove the outliers, use the `Exclude` name-value pair:

```
mdl3 = fitlm(tbl,'MPG ~ Cylinders*Weight + Weight^2','Exclude',outl);
```

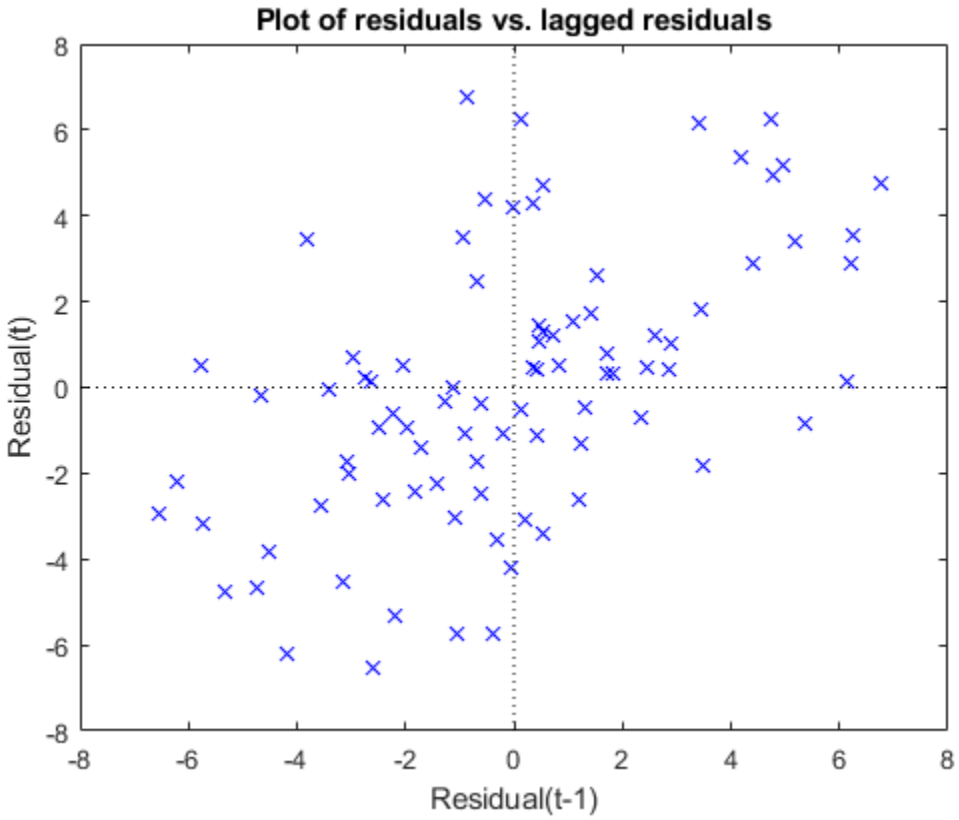
Examine a residuals plot of mdl2:

```
plotResiduals(mdl3)
```



The new residuals plot looks fairly symmetric, without obvious problems. However, there might be some serial correlation among the residuals. Create a new plot to see if such an effect exists.

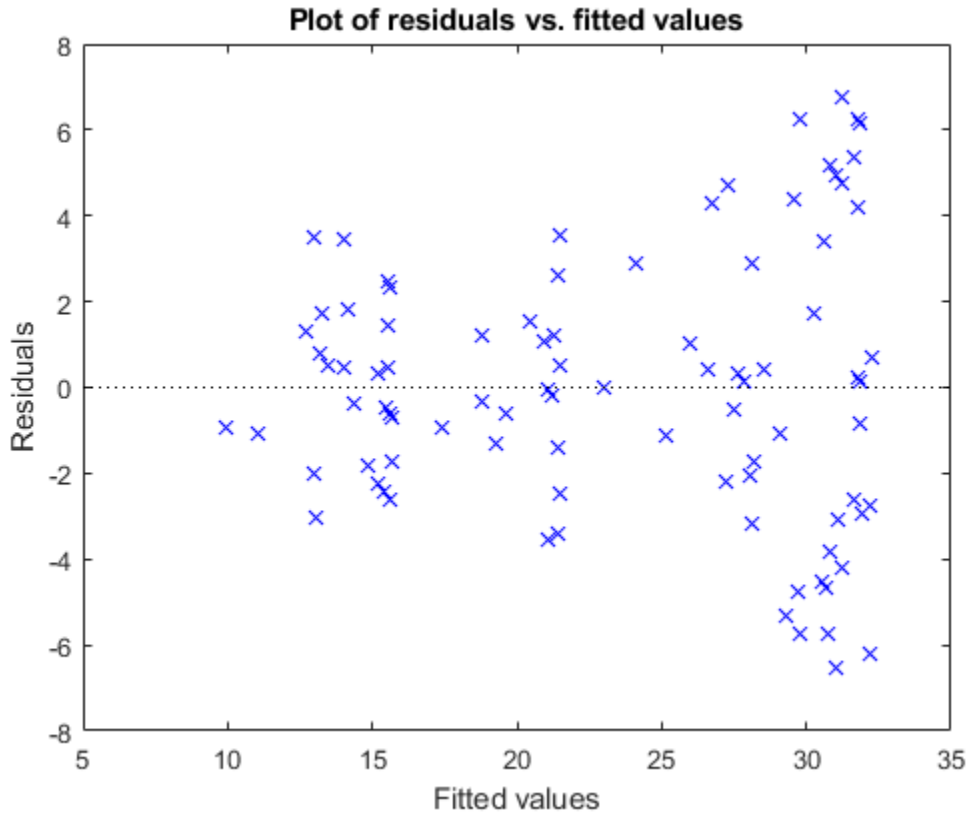
```
plotResiduals(mdl3,'lagged')
```



The scatter plot shows many more crosses in the upper-right and lower-left quadrants than in the other two quadrants, indicating positive serial correlation among the residuals.

Another potential issue is when residuals are large for large observations. See if the current model has this issue.

```
plotResiduals(mdl3,'fitted')
```



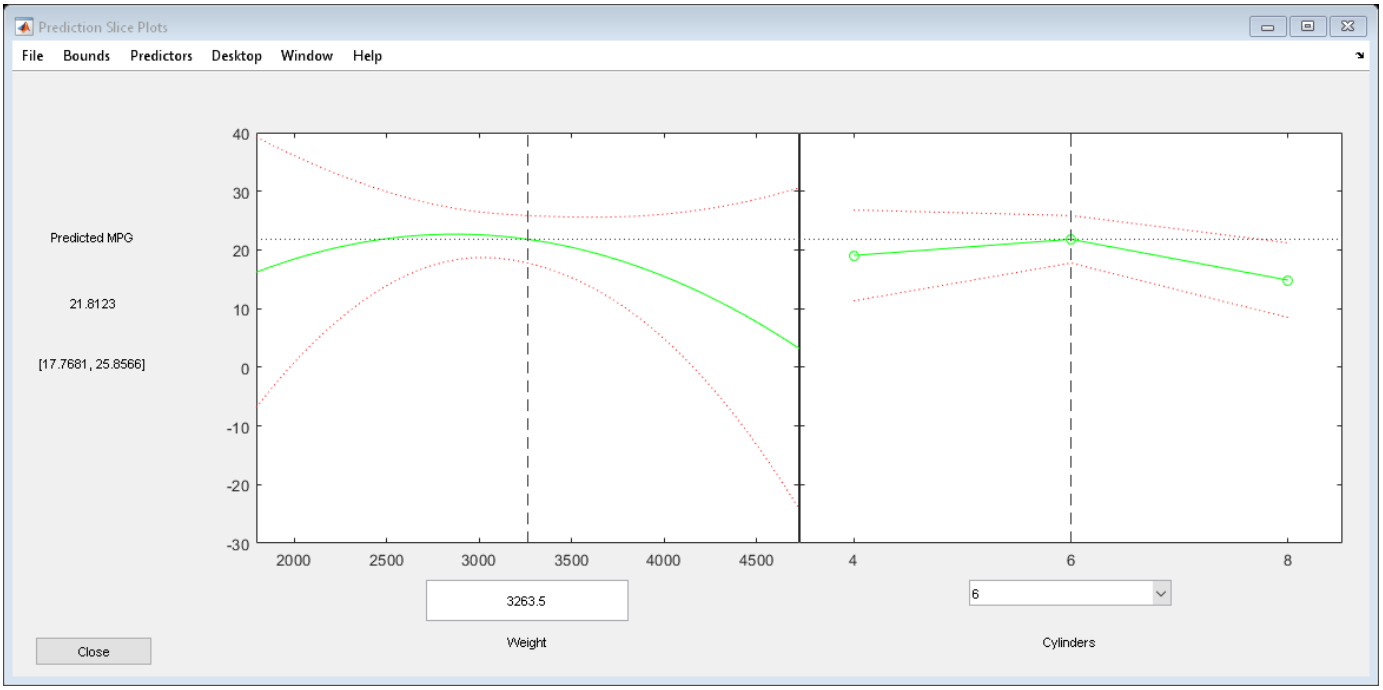
There is some tendency for larger fitted values to have larger residuals. Perhaps the model errors are proportional to the measured values.

Plots to Understand Predictor Effects

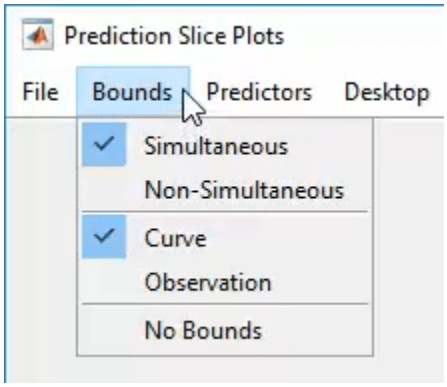
This example shows how to understand the effect each predictor has on a regression model using a variety of available plots.

Examine a slice plot of the responses. This displays the effect of each predictor separately.

```
plotSlice(mdl)
```

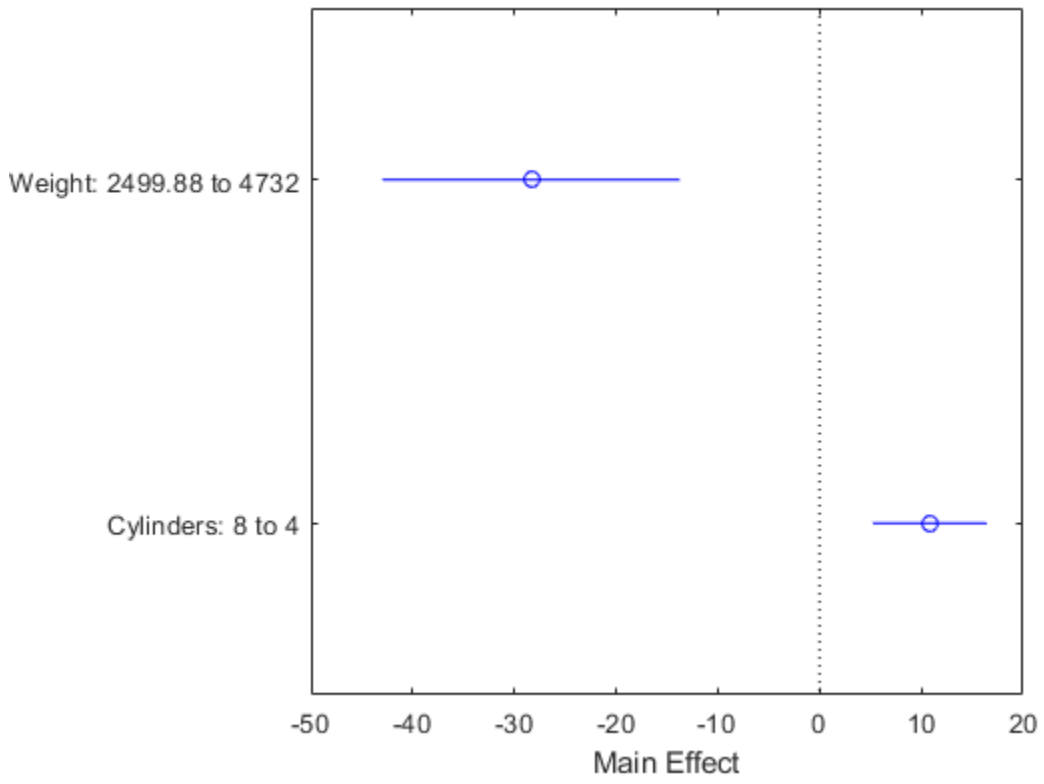


You can drag the individual predictor values, which are represented by dashed blue vertical lines. You can also choose between simultaneous and non-simultaneous confidence bounds, which are represented by dashed red curves.



Use an effects plot to show another view of the effect of predictors on the response.

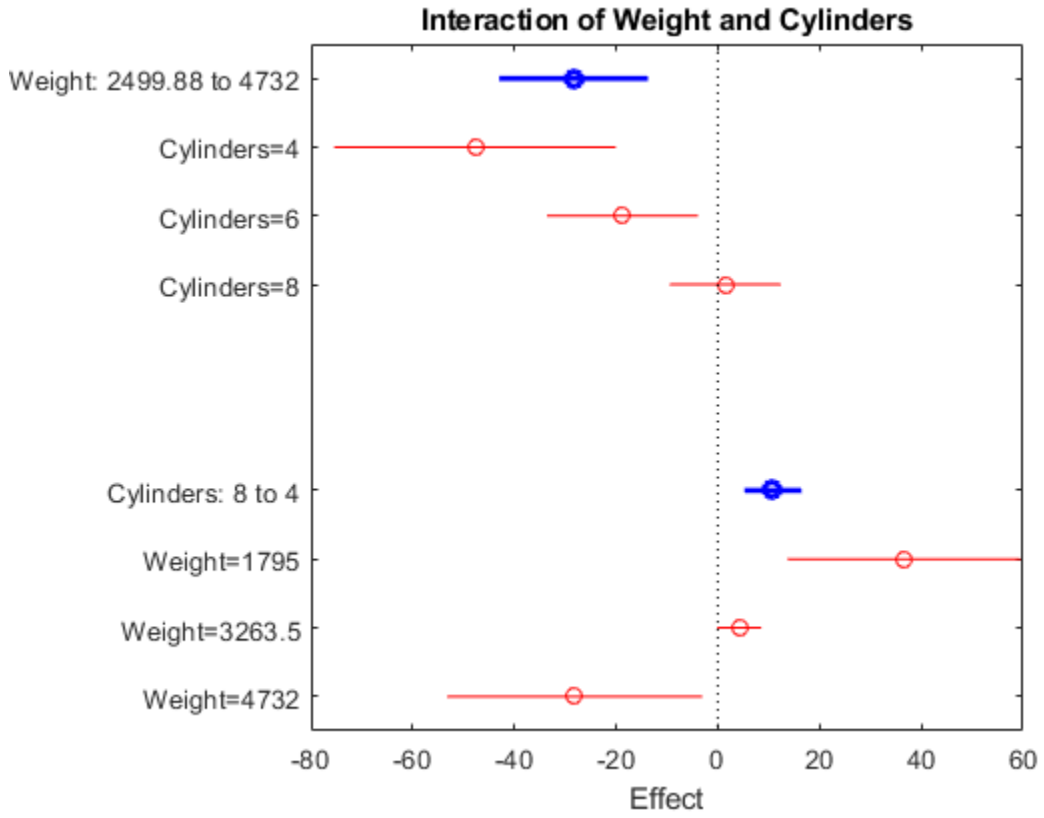
```
plotEffects(mdl)
```



This plot shows that changing `Weight` from about 2500 to 4732 lowers `MPG` by about 30 (the location of the upper blue circle). It also shows that changing the number of cylinders from 8 to 4 raises `MPG` by about 10 (the lower blue circle). The horizontal blue lines represent confidence intervals for these predictions. The predictions come from averaging over one predictor as the other is changed. In cases such as this, where the two predictors are correlated, be careful when interpreting the results.

Instead of viewing the effect of averaging over a predictor as the other is changed, examine the joint interaction in an interaction plot.

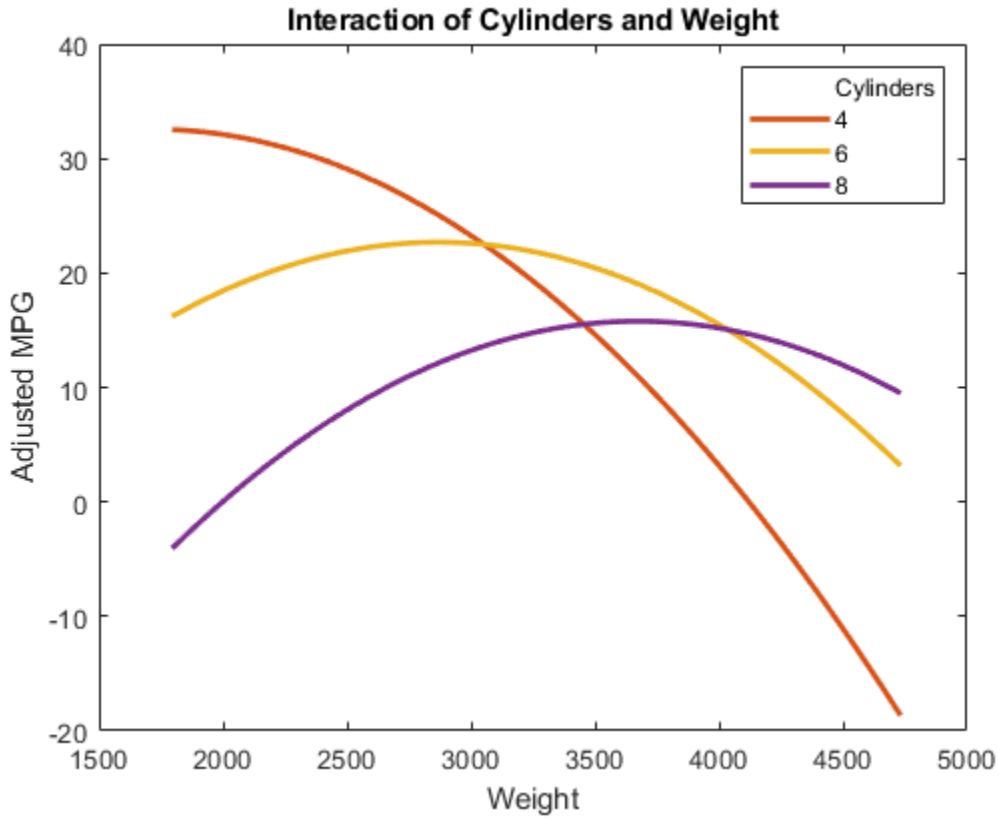
```
plotInteraction(mdl,'Weight','Cylinders')
```



The interaction plot shows the effect of changing one predictor with the other held fixed. In this case, the plot is much more informative. It shows, for example, that lowering the number of cylinders in a relatively light car (`Weight` = 1795) leads to an increase in mileage, but lowering the number of cylinders in a relatively heavy car (`Weight` = 4732) leads to a decrease in mileage.

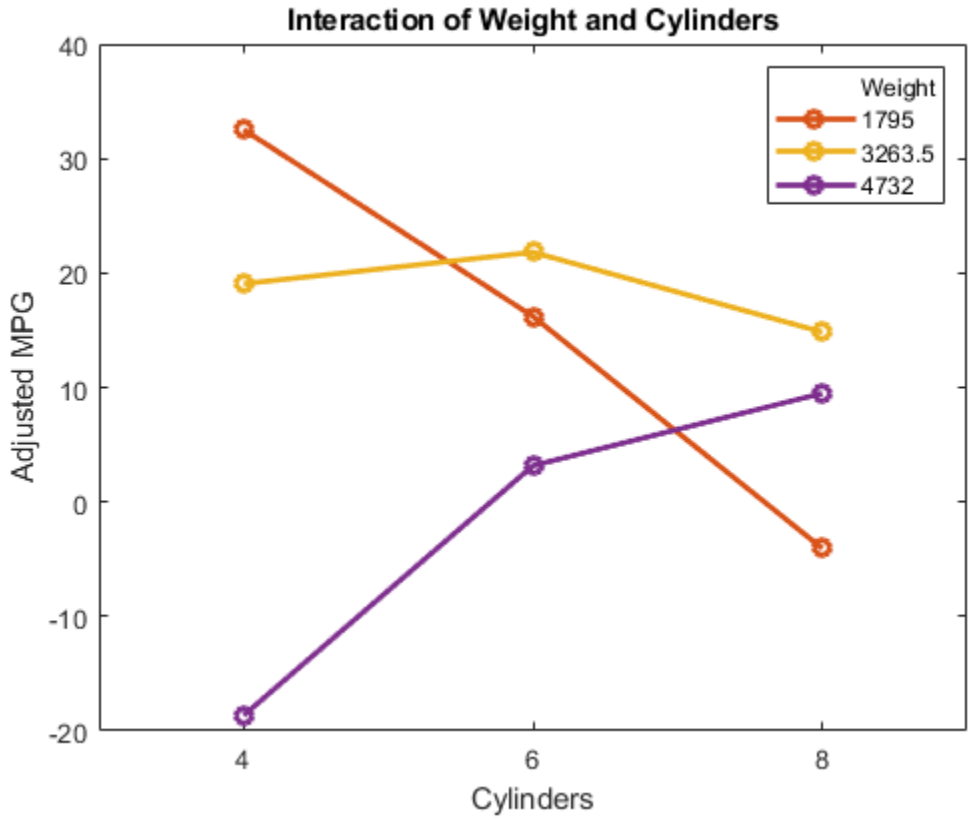
For an even more detailed look at the interactions, look at an interaction plot with predictions. This plot holds one predictor fixed while varying the other, and plots the effect as a curve. Look at the interactions for various fixed numbers of cylinders.

```
plotInteraction(mdl,'Cylinders','Weight','predictions')
```



Now look at the interactions with various fixed levels of weight.

```
plotInteraction(mdl,'Weight','Cylinders','predictions')
```

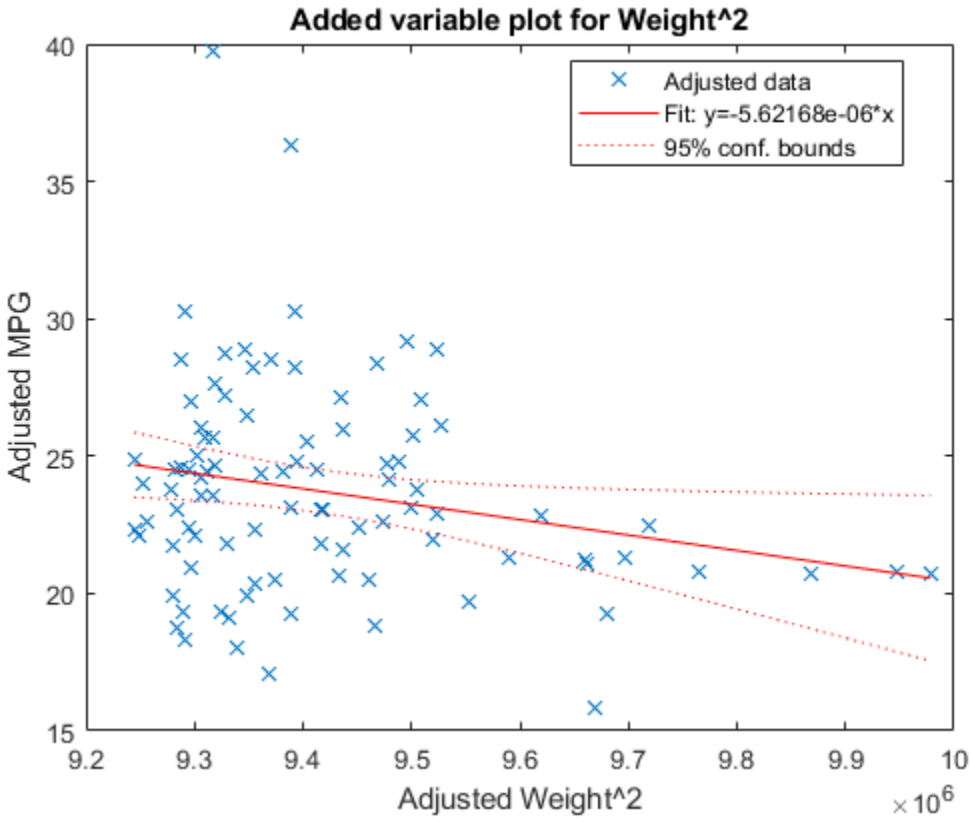


Plots to Understand Terms Effects

This example shows how to understand the effect of each term in a regression model using a variety of available plots.

Create an added variable plot with `Weight^2` as the added variable.

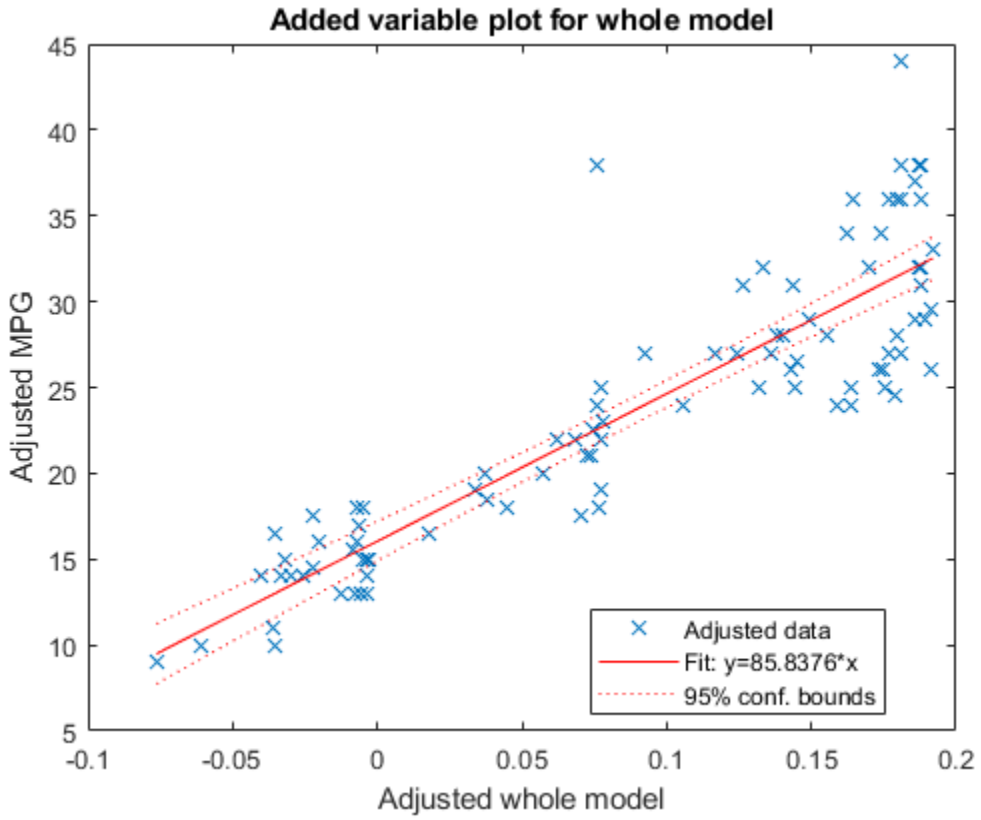
```
plotAdded(mdl,'Weight^2')
```



This plot shows the results of fitting both `Weight^2` and `MPG` to the terms other than `Weight^2`. The reason to use `plotAdded` is to understand what additional improvement in the model you get by adding `Weight^2`. The coefficient of a line fit to these points is the coefficient of `Weight^2` in the full model. The `Weight^2` predictor is just over the edge of significance (`pValue` < 0.05) as you can see in the coefficients table display. You can see that in the plot as well. The confidence bounds look like they could not contain a horizontal line (constant `y`), so a zero-slope model is not consistent with the data.

Create an added variable plot for the model as a whole.

```
plotAdded(mdl)
```



The model as a whole is very significant, so the bounds don't come close to containing a horizontal line. The slope of the line is the slope of a fit to the predictors projected onto their best-fitting direction, or in other words, the norm of the coefficient vector.

Change Models

There are two ways to change a model:

- `step` — Add or subtract terms one at a time, where `step` chooses the most important term to add or remove.
- `addTerms` and `removeTerms` — Add or remove specified terms. Give the terms in any of the forms described in “Choose a Model or Range of Models” on page 11-11.

If you created a model using `stepwiselm`, then `step` can have an effect only if you give different upper or lower models. `step` does not work when you fit a model using `RobustOpts`.

For example, start with a linear model of mileage from the `carbig` data:

```
load carbig
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
mdl = fitlm(tbl,'linear','ResponseVar','MPG')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Acceleration + Displacement + Horsepower + Weight

Estimated Coefficients:
```

```
                     Estimate         SE         tStat        pValue
                    ___________    __________    ________    __________

    (Intercept)         45.251         2.456      18.424    7.0721e-55
    Acceleration     -0.023148        0.1256     -0.1843       0.85388
    Displacement    -0.0060009     0.0067093    -0.89441       0.37166
    Horsepower       -0.043608      0.016573     -2.6312      0.008849
    Weight          -0.0052805    0.00081085     -6.5123    2.3025e-10


Number of observations: 392, Error degrees of freedom: 387
Root Mean Squared Error: 4.25
R-squared: 0.707,  Adjusted R-Squared: 0.704
F-statistic vs. constant model: 233, p-value = 9.63e-102
```

Try to improve the model using step for up to 10 steps:

```
mdl1 = step(mdl,'NSteps',10)
```

```
1. Adding Displacement:Horsepower, FStat = 87.4802, pValue = 7.05273e-19

mdl1 =
Linear regression model:
    MPG ~ 1 + Acceleration + Weight + Displacement*Horsepower

Estimated Coefficients:
                                Estimate         SE         tStat       pValue
                               ___________    __________    _______    __________

    (Intercept)                   61.285         2.8052     21.847    1.8593e-69
    Acceleration                 -0.34401        0.11862      -2.9     0.0039445
    Displacement                 -0.081198      0.010071    -8.0623    9.5014e-15
    Horsepower                    -0.24313      0.026068    -9.3265    8.6556e-19
    Weight                      -0.0014367    0.00084041    -1.7095      0.088166
    Displacement:Horsepower    0.00054236    5.7987e-05     9.3531    7.0527e-19


Number of observations: 392, Error degrees of freedom: 386
Root Mean Squared Error: 3.84
R-squared: 0.761,  Adjusted R-Squared: 0.758
F-statistic vs. constant model: 246, p-value = 1.32e-117
```

`step` stopped after just one change.

To try to simplify the model, remove the `Acceleration` and `Weight` terms from `mdl1`:

```
mdl2 = removeTerms(mdl1,'Acceleration + Weight')
```

```
mdl2 =
Linear regression model:
    MPG ~ 1 + Displacement*Horsepower

Estimated Coefficients:
                                Estimate         SE         tStat       pValue
                               ___________    __________    _______    __________

    (Intercept)                   53.051         1.526      34.765    3.0201e-121
    Displacement                 -0.098046    0.0066817    -14.674     4.3203e-39
    Horsepower                    -0.23434     0.019593     -11.96     2.8024e-28
```

```
    Displacement:Horsepower    0.00058278    5.193e-05     11.222     1.6816e-25


Number of observations: 392, Error degrees of freedom: 388
Root Mean Squared Error: 3.94
R-squared: 0.747,  Adjusted R-Squared: 0.745
F-statistic vs. constant model: 381, p-value = 3e-115
```

`mdl2` uses just `Displacement` and `Horsepower`, and has nearly as good a fit to the data as `mdl1` in the `Adjusted R-Squared` metric.

Predict or Simulate Responses to New Data

A `LinearModel` object offers three functions to predict or simulate the response to new data: `predict`, `feval`, and `random`.

`predict`

Use the `predict` function to predict and obtain confidence intervals on the predictions.

Load the `carbig` data and create a default linear model of the response `MPG` to the `Acceleration`, `Displacement`, `Horsepower`, and `Weight` predictors.

```
load carbig
X = [Acceleration,Displacement,Horsepower,Weight];
mdl = fitlm(X,MPG);
```

Create a three-row array of predictors from the minimal, mean, and maximal values. `X` contains some `NaN` values, so specify the `'omitnan'` option for the `mean` function. The `min` and `max` functions omit `NaN` values in the calculation by default.

```
Xnew = [min(X);mean(X,'omitnan');max(X)];
```

Find the predicted model responses and confidence intervals on the predictions.

```
[NewMPG, NewMPGCI] = predict(mdl,Xnew)
```

```
NewMPG = 3×1

   34.1345
   23.4078
    4.7751


NewMPGCI = 3×2

   31.6115   36.6575
   22.9859   23.8298
    0.6134    8.9367
```

The confidence bound on the mean response is narrower than those for the minimum or maximum responses.

`feval`

Use the `feval` function to predict responses. When you create a model from a table or dataset array, `feval` is often more convenient than `predict` for predicting responses. When you have new

predictor data, you can pass it to `feval` without creating a table or matrix. However, `feval` does not provide confidence bounds.

Load the `carbig` data set and create a default linear model of the response `MPG` to the predictors `Acceleration`, `Displacement`, `Horsepower`, and `Weight`.

```
load carbig
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
mdl = fitlm(tbl,'linear','ResponseVar','MPG');
```

Predict the model response for the mean values of the predictors.

```
NewMPG = feval(mdl,mean(Acceleration,'omitnan'),mean(Displacement,'omitnan'),mean(Horsepower,'om
```

```
NewMPG = 23.4078
```

`random`

Use the `random` function to simulate responses. The `random` function simulates new random response values, equal to the mean prediction plus a random disturbance with the same variance as the training data.

Load the `carbig` data and create a default linear model of the response `MPG` to the `Acceleration`, `Displacement`, `Horsepower`, and `Weight` predictors.

```
load carbig
X = [Acceleration,Displacement,Horsepower,Weight];
mdl = fitlm(X,MPG);
```

Create a three-row array of predictors from the minimal, mean, and maximal values.

```
Xnew = [min(X);mean(X,'omitnan');max(X)];
```

Generate new predicted model responses including some randomness.

```
rng('default') % for reproducibility
NewMPG = random(mdl,Xnew)
```

```
NewMPG = 3×1

   36.4178
   31.1958
   -4.8176
```

Because a negative value of `MPG` does not seem sensible, try predicting two more times.

```
NewMPG = random(mdl,Xnew)
```

```
NewMPG = 3×1

   37.7959
   24.7615
   -0.7783
```

```
NewMPG = random(mdl,Xnew)
```

```
NewMPG = 3×1
```

```
   32.2931
   24.8628
   19.9715
```

Clearly, the predictions for the third (maximal) row of `Xnew` are not reliable.

Share Fitted Models

Suppose you have a linear regression model, such as `mdl` from the following commands.

```
load carbig
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
mdl = fitlm(tbl,'linear','ResponseVar','MPG');
```

To share the model with other people, you can:

- Provide the model display.

```
mdl
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Acceleration + Displacement + Horsepower + Weight

Estimated Coefficients:
                    Estimate         SE         tStat        pValue
                   __________    __________    ________    __________

    (Intercept)         45.251         2.456      18.424    7.0721e-55
    Acceleration     -0.023148        0.1256     -0.1843       0.85388
    Displacement    -0.0060009     0.0067093    -0.89441       0.37166
    Horsepower       -0.043608      0.016573     -2.6312      0.008849
    Weight          -0.0052805    0.00081085     -6.5123    2.3025e-10


Number of observations: 392, Error degrees of freedom: 387
Root Mean Squared Error: 4.25
R-squared: 0.707,  Adjusted R-Squared: 0.704
F-statistic vs. constant model: 233, p-value = 9.63e-102
```

- Provide the model definition and coefficients.

```
mdl.Formula
```

```
ans =
MPG ~ 1 + Acceleration + Displacement + Horsepower + Weight
```

```
mdl.CoefficientNames
```

```
ans = 1x5 cell
  Columns 1 through 4

    {'(Intercept)'}    {'Acceleration'}    {'Displacement'}    {'Horsepower'}

  Column 5

    {'Weight'}
```

```
mdl.Coefficients.Estimate

ans = 5×1

   45.2511
   -0.0231
   -0.0060
   -0.0436
   -0.0053
```

See Also

`LinearModel` | `anova` | `fitlm` | `lasso` | `plotResiduals` | `predict` | `sequentialfs` | `stepwiselm`

More About

- “What Is a Linear Regression Model?” on page 11-6
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “Interpret Linear Regression Results” on page 11-50
- “Linear Regression with Interaction Effects” on page 11-44
- “Linear Regression with Categorical Covariates” on page 2-52
- “Reduce Outlier Effects Using Robust Regression” on page 11-104
- “Stepwise Regression” on page 11-99

Linear Regression Workflow

This example shows how to fit a linear regression model. A typical workflow involves the following: import data, fit a regression, test its quality, modify it to improve the quality, and share it.

Step 1. Import the data into a table.

`hospital.xls` is an Excel® spreadsheet containing patient names, sex, age, weight, blood pressure, and dates of treatment in an experimental protocol. First read the data into a table.

```
patients = readtable('hospital.xls','ReadRowNames',true);
```

Examine the five rows of data.

```
patients(1:5,:)
```

```
ans=5×11 table
                 name          sex      age    wgt    smoke    sys    dia    trial1    trial2
              _____________    _____    ___    ___    _____    ___    ___    ______    ______

    YPL-320    {'SMITH'   }    {'m'}    38     176      1      124    93       18       -99
    GLI-532    {'JOHNSON' }    {'m'}    43     163      0      109    77       11        13
    PNI-258    {'WILLIAMS'}    {'f'}    38     131      0      125    83      -99       -99
    MIJ-579    {'JONES'   }    {'f'}    40     133      0      117    75        6        12
    XLK-030    {'BROWN'   }    {'f'}    49     119      0      122    80       14        23
```

The `sex` and `smoke` fields seem to have two choices each. So change these fields to categorical.

```
patients.smoke = categorical(patients.smoke,0:1,{'No','Yes'});
patients.sex = categorical(patients.sex);
```

Step 2. Create a fitted model.

Your goal is to model the systolic pressure as a function of a patient's age, weight, sex, and smoking status. Create a linear formula for `'sys'` as a function of `'age'`, `'wgt'`, `'sex'`, and `'smoke'` .

```
modelspec = 'sys ~ age + wgt + sex + smoke';
mdl = fitlm(patients,modelspec)
```

```
mdl =
Linear regression model:
    sys ~ 1 + sex + age + wgt + smoke

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      118.28      7.6291      15.504    9.1557e-28
    sex_m           0.88162      2.9473     0.29913       0.76549
    age             0.08602     0.06731       1.278       0.20438
    wgt            -0.016685    0.055714    -0.29947       0.76524
    smoke_Yes          9.884      1.0406       9.498    1.9546e-15


Number of observations: 100, Error degrees of freedom: 95
Root Mean Squared Error: 4.81
```

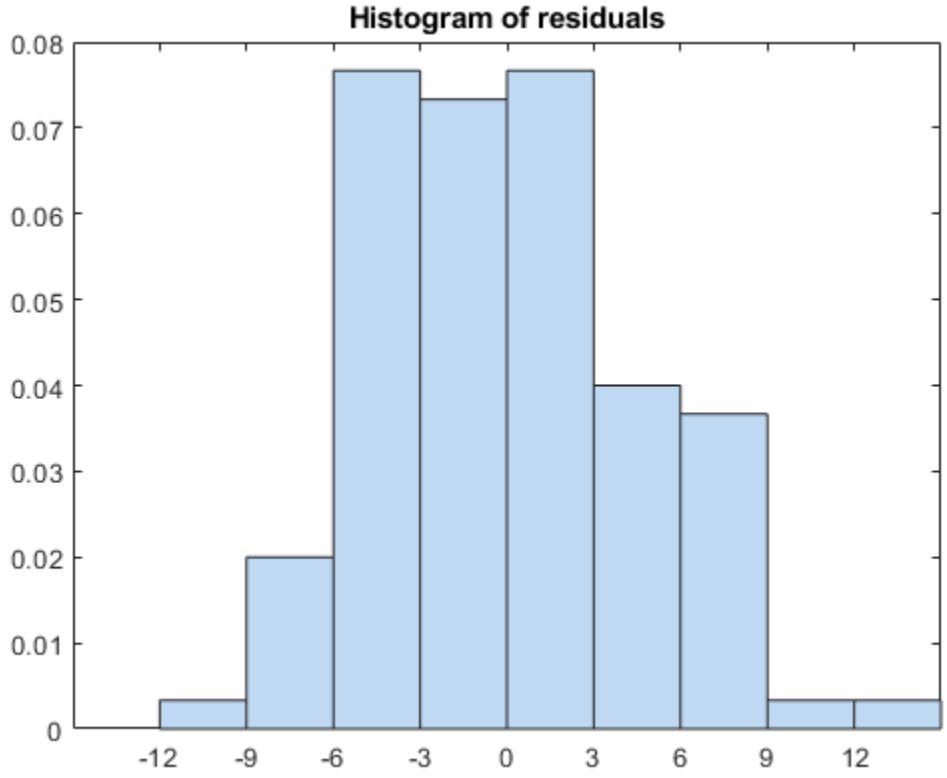
```
R-squared: 0.508,  Adjusted R-Squared: 0.487
F-statistic vs. constant model: 24.5, p-value = 5.99e-14
```

The sex, age, and weight predictors have rather high *p*-values, indicating that some of these predictors might be unnecessary.

Step 3. Locate and remove outliers.

See if there are outliers in the data that should be excluded from the fit. Plot the residuals.

```
plotResiduals(mdl)
```



There is one possible outlier, with a value greater than 12. This is probably not truly an outlier. For demonstration, here is how to find and remove it.

Find the outlier.

```
outlier = mdl.Residuals.Raw > 12;
find(outlier)
```

```
ans = 84
```

Remove the outlier.

```
mdl = fitlm(patients,modelspec,...
    'Exclude',84);

mdl.ObservationInfo(84,:)
```

```
ans=1×4 table
               Weights    Excluded    Missing    Subset
               _______    ________    _______    ______

    WXM-486       1        true        false     false
```

Observation 84 is no longer in the model.

Step 4. Simplify the model.

Try to obtain a simpler model, one with fewer predictors but the same predictive accuracy. `step` looks for a better model by adding or removing one term at a time. Allow `step` take up to 10 steps.

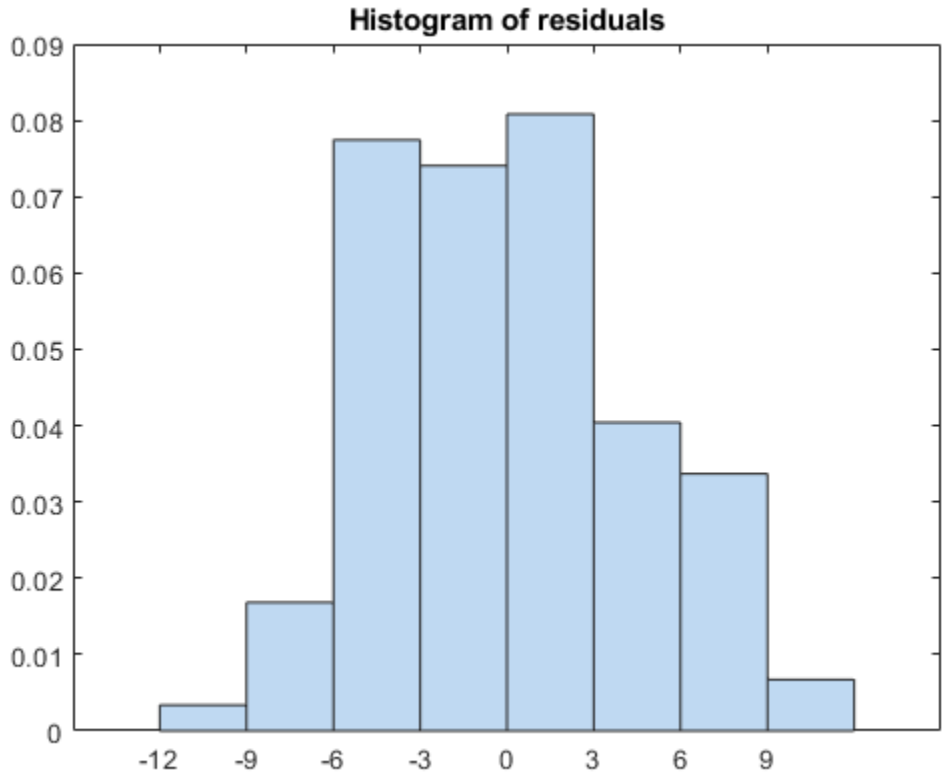
```
mdl1 = step(mdl,'NSteps',10)
```

```
1. Removing wgt, FStat = 4.6001e-05, pValue = 0.9946
2. Removing sex, FStat = 0.063241, pValue = 0.80199

mdl1 =
Linear regression model:
    sys ~ 1 + age + smoke

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     115.11       2.5364    45.383    1.1407e-66
    age            0.10782     0.064844    1.6628       0.09962
    smoke_Yes       10.054      0.97696    10.291    3.5276e-17


Number of observations: 99, Error degrees of freedom: 96
Root Mean Squared Error: 4.61
R-squared: 0.536,  Adjusted R-Squared: 0.526
F-statistic vs. constant model: 55.4, p-value = 1.02e-16
```

`step` took two steps. This means it could not improve the model further by adding or subtracting a single term.

Plot the effectiveness of the simpler model on the training data.

```
plotResiduals(mdl1)
```



The residuals look about as small as those of the original model.

Step 5. Predict responses to new data.

Suppose you have four new people, aged 25, 30, 40, and 65, and the first and third smoke. Predict their systolic pressure using `mdl1`.

```
ages = [25;30;40;65];
smoker = {'Yes';'No';'Yes';'No'};
systolicnew = feval(mdl1,ages,smoker)
```

```
systolicnew = 4×1

  127.8561
  118.3412
  129.4734
  122.1149
```

To make predictions, you need only the variables that `mdl1` uses.

Step 6. Share the model.

You might want others to be able to use your model for prediction. Access the terms in the linear model.

```
coefnames = mdl1.CoefficientNames
```

```
coefnames = 1x3 cell
    {'(Intercept)'}    {'age'}    {'smoke_Yes'}
```

View the model formula.

```
mdl1.Formula
```

```
ans =
sys ~ 1 + age + smoke
```

Access the coefficients of the terms.

```
coefvals = mdl1.Coefficients(:,1).Estimate
```

```
coefvals = 3×1

  115.1066
    0.1078
   10.0540
```

The model is `sys = 115.1066 + 0.1078*age + 10.0540*smoke`, where `smoke` is `1` for a smoker, and `0` otherwise.

See Also

`LinearModel` | `feval` | `fitlm` | `plotResiduals` | `step`

More About

- “What Is a Linear Regression Model?” on page 11-6
- “Linear Regression” on page 11-9
- “Interpret Linear Regression Results” on page 11-50
- “Linear Regression with Interaction Effects” on page 11-44

Regression Using Dataset Arrays

This example shows how to perform linear and stepwise regression analyses using dataset arrays.

Load sample data.

```
load imports-85
```

Store predictor and response variables in dataset array.

```
ds = dataset(X(:,7),X(:,8),X(:,9),X(:,15),'Varnames',...
{'curb_weight','engine_size','bore','price'});
```

Fit linear regression model.

Fit a linear regression model that explains the price of a car in terms of its curb weight, engine size, and bore.

```
fitlm(ds,'price~curb_weight+engine_size+bore')
```

```
ans =
Linear regression model:
    price ~ 1 + curb_weight + engine_size + bore

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    _______    __________

    (Intercept)        64.095        3.703     17.309    2.0481e-41
    curb_weight    -0.0086681    0.0011025    -7.8623      2.42e-13
    engine_size     -0.015806     0.013255    -1.1925       0.23452
    bore              -2.6998       1.3489    -2.0015      0.046711


Number of observations: 201, Error degrees of freedom: 197
Root Mean Squared Error: 3.95
R-squared: 0.674,  Adjusted R-Squared: 0.669
F-statistic vs. constant model: 136, p-value = 1.14e-47
```

The command `fitlm(ds)` also returns the same result because `fitlm`, by default, assumes the predictor variable is in the last column of the dataset array `ds`.

Recreate dataset array and repeat analysis.

This time, put the response variable in the first column of the dataset array.

```
 ds = dataset(X(:,15),X(:,7),X(:,8),X(:,9),'Varnames',...
{'price','curb_weight','engine_size','bore'});
```

When the response variable is in the first column of `ds`, define its location. For example, `fitlm`, by default, assumes that `bore` is the response variable. You can define the response variable in the model using either:

```
fitlm(ds,'ResponseVar','price');
```

or

```
fitlm(ds,'ResponseVar',logical([1 0 0 0]));
```

Perform stepwise regression.

```
stepwiselm(ds,'quadratic','lower','price~1',...
'ResponseVar','price')

1. Removing bore^2, FStat = 0.01282, pValue = 0.90997
2. Removing engine_size^2, FStat = 0.078043, pValue = 0.78027
3. Removing curb_weight:bore, FStat = 0.70558, pValue = 0.40195

ans =
Linear regression model:
    price ~ 1 + curb_weight*engine_size + engine_size*bore + curb_weight^2

Estimated Coefficients:
                              Estimate          SE         tStat       pValue
                             __________    __________    _______    __________

    (Intercept)                 131.13        14.273     9.1873    6.2319e-17
    curb_weight               -0.043315     0.0085114    -5.0891    8.4682e-07
    engine_size                -0.17102       0.13844    -1.2354       0.21819
    bore                        -12.244         4.999    -2.4493      0.015202
    curb_weight:engine_size   -6.3411e-05    2.6577e-05     -2.386      0.017996
    engine_size:bore            0.092554      0.037263     2.4838      0.013847
    curb_weight^2              8.0836e-06    1.9983e-06     4.0451    7.5432e-05


Number of observations: 201, Error degrees of freedom: 194
Root Mean Squared Error: 3.59
R-squared: 0.735,  Adjusted R-Squared: 0.726
F-statistic vs. constant model: 89.5, p-value = 3.58e-53
```

The initial model is a quadratic formula, and the lowest model considered is the constant. Here, `stepwiselm` performs a backward elimination technique to determine the terms in the model. The final model is `price ~ 1 + curb_weight*engine_size + engine_size*bore + curb_weight^2`, which corresponds to

$$P = \beta_0 + \beta_C C + \beta_E E + \beta_B B + \beta_{CE} CE + \beta_{EB} EB + \beta_{C^2} C^2 + \epsilon$$

where P is price, C is curb weight, E is engine size, B is bore, β_i is the coefficient for the corresponding term in the model, and ϵ is the error term. The final model includes all three main effects, the interaction effects for curb weight and engine size and engine size and bore, and the second-order term for curb weight.

See Also

`LinearModel` | `fitlm` | `stepwiselm`

Related Examples

- “Linear Regression” on page 11-9
- “Stepwise Regression” on page 11-99
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “Interpret Linear Regression Results” on page 11-50

Linear Regression Using Tables

This example shows how to perform linear and stepwise regression analyses using tables.

Load sample data.

```
load imports-85
```

Store predictor and response variables in a table.

```
tbl = table(X(:,7),X(:,8),X(:,9),X(:,15),'VariableNames',...
{'curb_weight','engine_size','bore','price'});
```

Fit linear regression model.

Fit a linear regression model that explains the price of a car in terms of its curb weight, engine size, and bore.

```
fitlm(tbl,'price~curb_weight+engine_size+bore')
```

```
ans =
Linear regression model:
    price ~ 1 + curb_weight + engine_size + bore

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    _______    __________

    (Intercept)        64.095        3.703     17.309    2.0481e-41
    curb_weight    -0.0086681    0.0011025    -7.8623      2.42e-13
    engine_size     -0.015806     0.013255    -1.1925       0.23452
    bore              -2.6998       1.3489    -2.0015      0.046711


Number of observations: 201, Error degrees of freedom: 197
Root Mean Squared Error: 3.95
R-squared: 0.674,  Adjusted R-Squared: 0.669
F-statistic vs. constant model: 136, p-value = 1.14e-47
```

The command `fitlm(tbl)` also returns the same result because `fitlm`, by default, assumes the response variable is in the last column of the table `tbl`.

Recreate table and repeat analysis.

This time, put the response variable in the first column of the table.

```
 tbl = table(X(:,15),X(:,7),X(:,8),X(:,9),'VariableNames',...
{'price','curb_weight','engine_size','bore'});
```

When the response variable is in the first column of `tbl`, define its location. For example, `fitlm`, by default, assumes that `bore` is the response variable. You can define the response variable in the model using either:

```
fitlm(tbl,'ResponseVar','price');
```

or

```
fitlm(tbl,'ResponseVar',logical([1 0 0 0]));
```

Perform stepwise regression.

```
stepwiselm(tbl,'quadratic','lower','price~1',...
'ResponseVar','price')

1. Removing bore^2, FStat = 0.01282, pValue = 0.90997
2. Removing engine_size^2, FStat = 0.078043, pValue = 0.78027
3. Removing curb_weight:bore, FStat = 0.70558, pValue = 0.40195

ans =
Linear regression model:
    price ~ 1 + curb_weight*engine_size + engine_size*bore + curb_weight^2

Estimated Coefficients:
                              Estimate          SE         tStat       pValue
                             __________    __________    _______    __________

    (Intercept)                 131.13        14.273     9.1873    6.2319e-17
    curb_weight               -0.043315     0.0085114    -5.0891    8.4682e-07
    engine_size                -0.17102       0.13844    -1.2354       0.21819
    bore                        -12.244         4.999    -2.4493      0.015202
    curb_weight:engine_size   -6.3411e-05    2.6577e-05     -2.386      0.017996
    engine_size:bore            0.092554      0.037263     2.4838      0.013847
    curb_weight^2              8.0836e-06    1.9983e-06     4.0451    7.5432e-05


Number of observations: 201, Error degrees of freedom: 194
Root Mean Squared Error: 3.59
R-squared: 0.735,  Adjusted R-Squared: 0.726
F-statistic vs. constant model: 89.5, p-value = 3.58e-53
```

The initial model is a quadratic formula, and the lowest model considered is the constant. Here, `stepwiselm` performs a backward elimination technique to determine the terms in the model. The final model is `price ~ 1 + curb_weight*engine_size + engine_size*bore + curb_weight^2`, which corresponds to

$$P = \beta_0 + \beta_C C + \beta_E E + \beta_B B + \beta_{CE} CE + \beta_{EB} EB + \beta_{C^2} C^2 + \epsilon$$

where P is price, C is curb weight, E is engine size, B is bore, β_i is the coefficient for the corresponding term in the model, and ϵ is the error term. The final model includes all three main effects, the interaction effects for curb weight and engine size and engine size and bore, and the second-order term for curb weight.

See Also

`LinearModel` | `fitlm` | `stepwiselm`

Related Examples

- “Linear Regression” on page 11-9
- “Stepwise Regression” on page 11-99
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “Interpret Linear Regression Results” on page 11-50

Linear Regression with Interaction Effects

Construct and analyze a linear regression model with interaction effects and interpret the results.

Load sample data.

```
load hospital
```

To retain only the first column of blood pressure, store data in a table.

```
tbl = table(hospital.Sex,hospital.Age,hospital.Weight,hospital.Smoker,hospital.BloodPressure(:,1
    'VariableNames',{'Sex','Age','Weight','Smoker','BloodPressure'});
```

Perform stepwise linear regression.

For the initial model, use the full model with all terms and their pairwise interactions.

```
mdl = stepwiselm(tbl,'interactions')
```

```
1. Removing Sex:Smoker, FStat = 0.050738, pValue = 0.8223
2. Removing Weight:Smoker, FStat = 0.07758, pValue = 0.78124
3. Removing Age:Weight, FStat = 1.9717, pValue = 0.16367
4. Removing Sex:Age, FStat = 0.32389, pValue = 0.57067
5. Removing Age:Smoker, FStat = 2.4939, pValue = 0.11768

mdl =
Linear regression model:
    BloodPressure ~ 1 + Age + Smoker + Sex*Weight

Estimated Coefficients:
                       Estimate       SE        tStat       pValue
                       ________    _________    _______    __________

    (Intercept)         133.17       10.337     12.883      1.76e-22
    Sex_Male           -35.269       17.524    -2.0126      0.047015
    Age                0.11584     0.067664      1.712      0.090198
    Weight             -0.1393     0.080211    -1.7367      0.085722
    Smoker_1            9.8307       1.0229     9.6102    1.2391e-15
    Sex_Male:Weight     0.2341      0.11192     2.0917      0.039162


Number of observations: 100, Error degrees of freedom: 94
Root Mean Squared Error: 4.72
R-squared: 0.53,  Adjusted R-Squared: 0.505
F-statistic vs. constant model: 21.2, p-value = 4e-14
```

The final model in formula form is `BloodPressure ~ 1 + Age + Smoker + Sex*Weight`. This model includes all four main effects (Age, Smoker, Sex, Weight) and the two-way interaction between `Sex` and `Weight`. This model corresponds to

$$BP = \beta_0 + \beta_A X_A + \beta_{Sm} I_{Sm} + \beta_S I_S + \beta_W X_W + \beta_{SW} X_W I_S + \epsilon,$$

where

- BP is the blood pressure
- β_i are the coefficients

- I_{Sm} is the indicator variable for smoking; $I_{Sm} = 1$ indicates a smoking patient whereas $I_{Sm} = 0$ indicates a nonsmoking patient
- I_S is the indicator variable for sex; $I_S = 1$ indicates a male patient whereas $I_S = 0$ indicates a female patient
- X_A is the `Age` variable
- X_W is the `Weight` variable
- ϵ is the error term

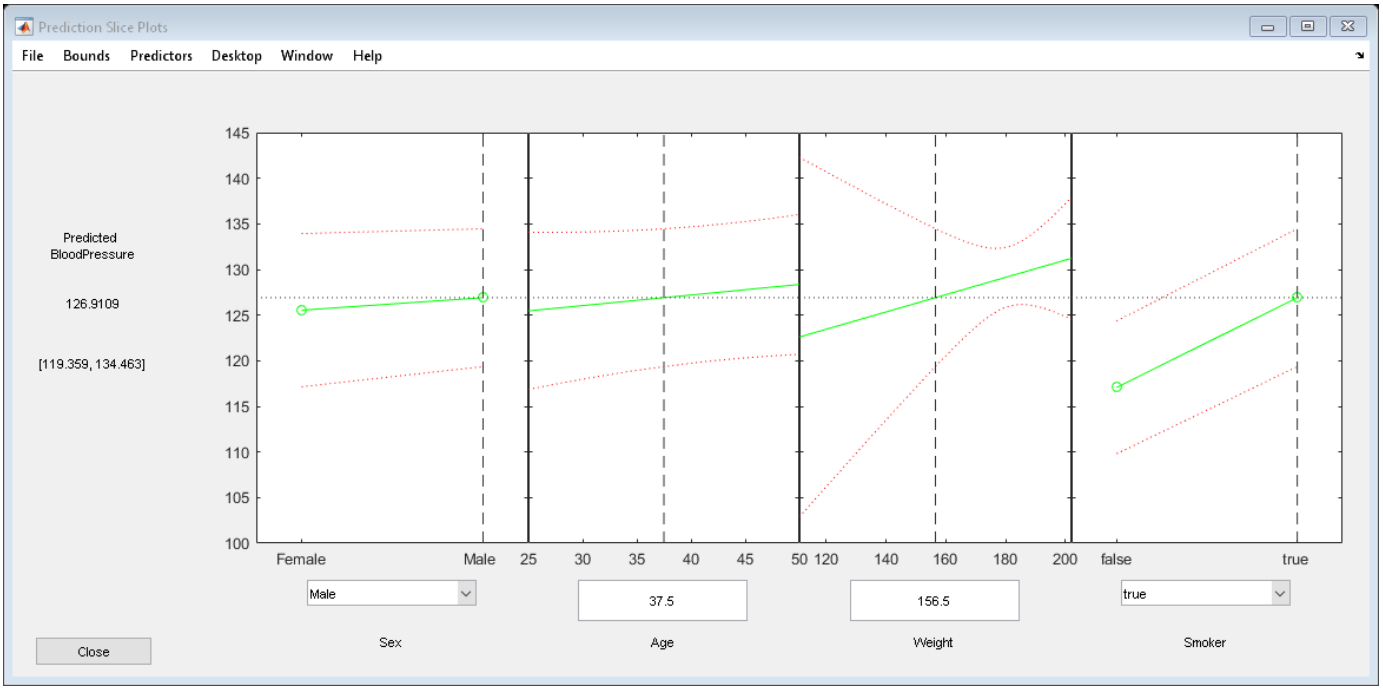
The following table shows the fitted linear model for each gender and smoking combination.

I_{Sm}	I_S	Linear Model
1(Smoker)	1(Male)	$BP = (\beta_0 + \beta_{Sm} + \beta_S) + \beta_A X_A + (\beta_W + \beta_{SW}) X_W$ $\widehat{BP} = 107.5617 + 0.11584 X_A + 0.11826 X_W$
1(Smoker)	0(Female)	$BP = (\beta_0 + \beta_{Sm}) + \beta_A X_A + \beta_W X_W$ $\widehat{BP} = 143.0007 + 0.11584 X_A - 0.1393 X_W$
0(Nonsmoker)	1(Male)	$BP = (\beta_0 + \beta_S) + \beta_A X_A + (\beta_W + \beta_{SW}) X_W$ $\widehat{BP} = 97.901 + 0.11584 X_A + 0.11826 X_W$
0(Nonsmoker)	0(Female)	$BP = \beta_0 + \beta_A X_A + \beta_W X_W$ $\widehat{BP} = 133.17 + 0.11584 X_A - 0.1393 X_W$

As seen from these models, β_{Sm} and β_S show how much the intercept of the response function changes when the indicator variable takes the value 1 compared to when it takes the value 0. β_{SW}, however, shows the effect of the Weight variable on the response variable when the indicator variable for sex takes the value 1 compared to when it takes the value 0. You can explore the main and interaction effects in the final model using the methods of the `LinearModel` class as follows.

Plot prediction slice plots.

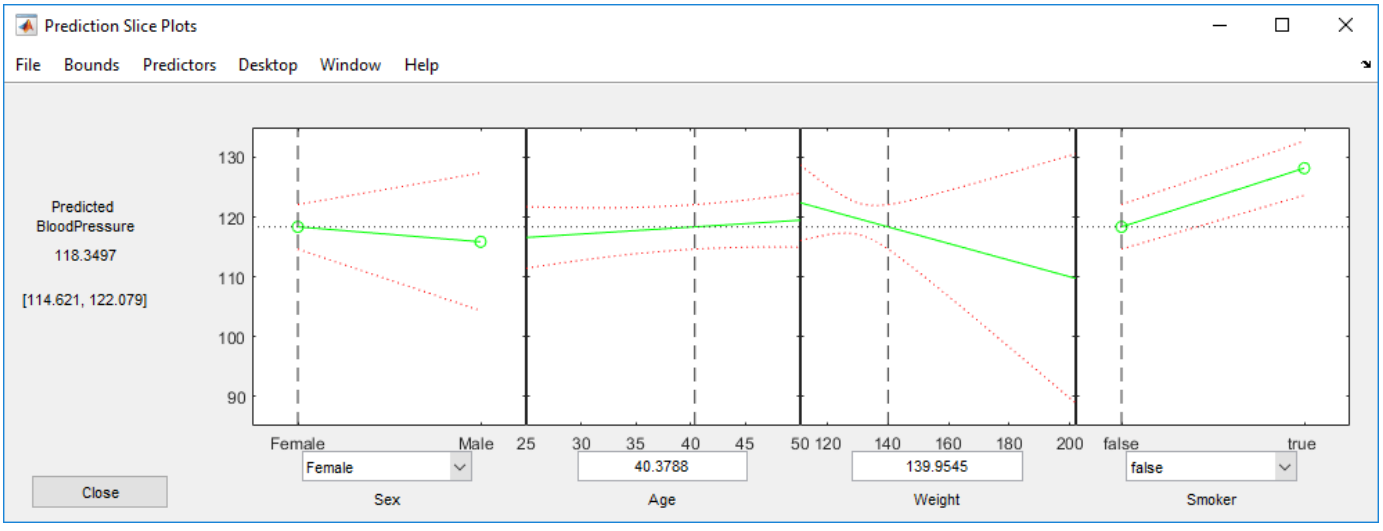
```
figure
plotSlice(mdl)
```



This plot shows the main effects for all predictor variables. The green line in each panel shows the change in the response variable as a function of the predictor variable when all other predictor variables are held constant. For example, for a smoking male patient aged 37.5, the expected blood pressure increases as the weight of the patient increases, given all else the same.

The dashed red curves in each panel show the 95% confidence bounds for the predicted response values.

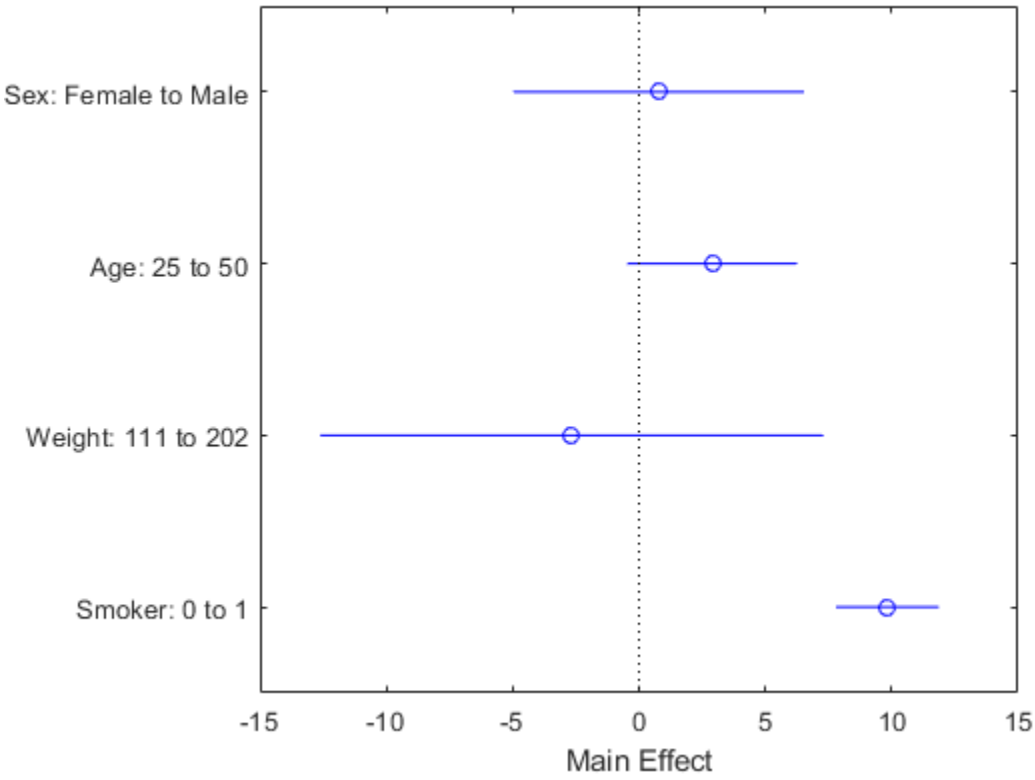
The horizontal dashed line in each panel shows the predicted response for the specific value of the predictor variable corresponding to the vertical dashed line. You can drag these lines to get the predicted response values at other predictor values, as shown next.



For example, the predicted value of the response variable is 118.3497 when a patient is female, nonsmoking, age 40.3788, and weighs 139.9545 pounds. The values in the square brackets, [114.621, 122.079], show the lower and upper limits of a 95% confidence interval for the estimated response. Note that, for a nonsmoking female patient, the expected blood pressure decreases as the weight increases, given all else is held constant.

Plot main effects.

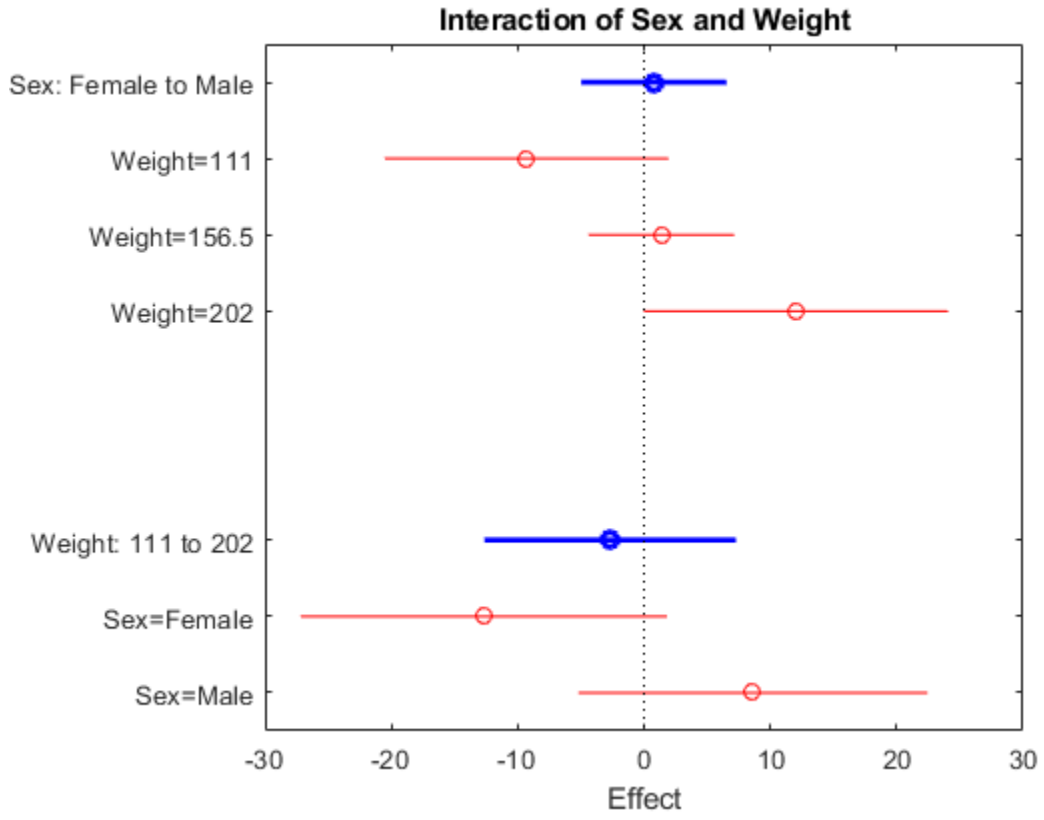
```
plotEffects(mdl)
```



This plot displays the main effects. The circles show the magnitude of the effect and the blue lines show the upper and lower confidence limits for the main effect. For example, being a smoker increases the expected blood pressure by 10 units, compared to being a nonsmoker, given all else is held constant. Expected blood pressure increases about two units for males compared to females, again, given other predictors held constant. An increase in age from 25 to 50 causes an expected increase of 4 units, whereas a change in weight from 111 to 202 causes about a 4-unit decrease in the expected blood pressure, given all else held constant.

Plot interaction effects.

```
figure
plotInteraction(mdl,'Sex','Weight')
```



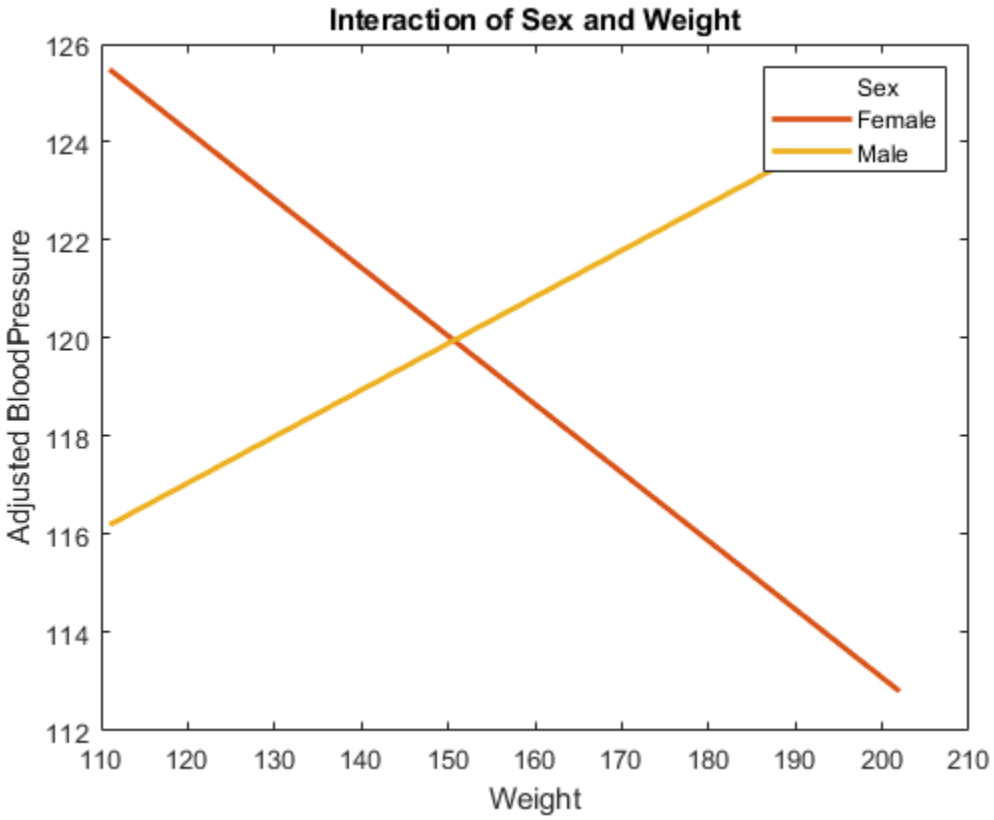
This plot displays the impact of a change in one factor given the other factor is fixed at a value.

Be cautious while interpreting the interaction effects. When there is not enough data on all factor combinations or the data is highly correlated, it might be difficult to determine the interaction effect of changing one factor while keeping the other fixed. In such cases, the estimated interaction effect is an extrapolation from the data.

The blue circles show the main effect of a specific term, as in the main effects plot. The red circles show the impact of a change in one term for fixed values of the other term. For example, in the bottom half of this plot, the red circles show the impact of a weight change in female and male patients, separately. You can see that an increase in a female's weight from 111 to 202 pounds causes about a 14-unit decrease in the expected blood pressure, while an increase of the same amount in the weight of a male patient causes about a 5-unit increase in the expected blood pressure, again given other predictors are held constant.

Plot prediction effects.

```
figure
plotInteraction(mdl,'Sex','Weight','predictions')
```



This plot shows the effect of changing one variable as the other predictor variable is held constant. In this example, the last figure shows the response variable, blood pressure, as a function of weight, when the variable sex is fixed at males and females. The lines for males and females are crossing which indicates a strong interaction between weight and sex. You can see that the expected blood pressure increases as the weight of a male patient increases, but decreases as the weight of a female patient increases.

See Also

`LinearModel` | `fitlm` | `plotEffects` | `plotInteraction` | `plotSlice` | `stepwiselm`

Related Examples

- “Linear Regression” on page 11-9
- “Stepwise Regression” on page 11-99
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “Interpret Linear Regression Results” on page 11-50

Interpret Linear Regression Results

This example shows how to display and interpret linear regression output statistics.

Fit Linear Regression Model

Load the `carsmall` data set, a matrix input data set.

```
load carsmall
X = [Weight,Horsepower,Acceleration];
```

Fit a linear regression model by using `fitlm`.

```
lm = fitlm(X,MPG)
```

```
lm =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                  __________    _________    _________    __________

    (Intercept)        47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
    x3              -0.011583      0.19333    -0.059913       0.95236


Number of observations: 93, Error degrees of freedom: 89
Root Mean Squared Error: 4.09
R-squared: 0.752,  Adjusted R-Squared: 0.744
F-statistic vs. constant model: 90, p-value = 7.38e-27
```

The model display includes the model formula, estimated coefficients, and model summary statistics.

The model formula in the display, `y ~ 1 + x1 + x2 + x3`, corresponds to $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

The model display shows the estimated coefficient information, which is stored in the `Coefficients` property. Display the `Coefficients` property.

```
lm.Coefficients
```

```
ans=4×4 table
                   Estimate        SE          tStat        pValue
                  __________    _________    _________    __________

    (Intercept)        47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
    x3              -0.011583      0.19333    -0.059913       0.95236
```

The `Coefficient` property includes these columns:

- `Estimate` — Coefficient estimates for each corresponding term in the model. For example, the estimate for the constant term (`intercept`) is 47.977.

- `SE` — Standard error of the coefficients.
- `tStat` — *t*-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. Note that `tStat = Estimate/SE`. For example, the *t*-statistic for the intercept is 47.977/3.8785 = 12.37.
- `pValue` — *p*-value for the *t*-statistic of the hypothesis test that the corresponding coefficient is equal to zero or not. For example, the *p*-value of the *t*-statistic for `x2` is greater than 0.05, so this term is not significant at the 5% significance level given the other terms in the model.

The summary statistics of the model are:

- `Number of observations` — Number of rows without any `NaN` values. For example, `Number of observations` is 93 because the `MPG` data vector has six `NaN` values and the `Horsepower` data vector has one `NaN` value for a different observation, where the number of rows in `X` and `MPG` is 100.
- `Error degrees of freedom` — $n - p$, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, the model has four predictors, so the `Error degrees of freedom` is 93 – 4 = 89.
- `Root mean squared error` — Square root of the mean squared error, which estimates the standard deviation of the error distribution.
- `R-squared` and `Adjusted R-squared` — Coefficient of determination and adjusted coefficient of determination, respectively. For example, the `R-squared` value suggests that the model explains approximately 75% of the variability in the response variable `MPG`.
- `F-statistic vs. constant model` — Test statistic for the *F*-test on the regression model, which tests whether the model fits significantly better than a degenerate model consisting of only a constant term.
- `p-value` — *p*-value for the *F*-test on the model. For example, the model is significant with a *p*-value of 7.3816e-27.

ANOVA

Perform analysis of variance (ANOVA) for the model.

```
anova(lm,'summary')
```

```
ans=3×5 table
                SumSq     DF    MeanSq      F         pValue
                ______    __    ______    ______    __________

    Total       6004.8    92    65.269
    Model         4516     3    1505.3    89.987    7.3816e-27
    Residual    1488.8    89    16.728
```

This `anova` display shows the following.

- `SumSq` — Sum of squares for the regression model, `Model`, the error term, `Residual`, and the total, `Total`.
- `DF` — Degrees of freedom for each term. Degrees of freedom is $n - 1$ for the total, $p - 1$ for the model, and $n - p$ for the error term, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, `MPG` data vector has six `NaN` values and one of the data vectors, `Horsepower`, has one `NaN` value for a different observation, so the

total degrees of freedom is 93 – 1 = 92. There are four coefficients in the model, so the model `DF` is 4 – 1 = 3, and the `DF` for error term is 93 – 4 = 89.

- `MeanSq` — Mean squared error for each term. Note that `MeanSq = SumSq/DF`. For example, the mean squared error for the error term is 1488.8/89 = 16.728. The square root of this value is the `root mean squared error` in the linear regression display, or 4.09.
- `F` — *F*-statistic value, which is the same as `F-statistic vs. constant model` in the linear regression display. In this example, it is 89.987, and in the linear regression display this *F*-statistic value is rounded up to 90.
- `pValue` — *p*-value for the *F*-test on the model. In this example, it is 7.3816e-27.

If there are higher-order terms in the regression model, `anova` partitions the model `SumSq` into the part explained by the higher-order terms and the rest. The corresponding *F*-statistics are for testing the significance of the linear terms and higher-order terms as separate groups.

If the data includes replicates, or multiple measurements at the same predictor values, then the `anova` partitions the error `SumSq` into the part for the replicates and the rest. The corresponding *F*-statistic is for testing the lack-of-fit by comparing the model residuals with the model-free variance estimate computed on the replicates.

Decompose ANOVA table for model terms.

```
anova(lm)
```

```
ans=4×5 table
              SumSq      DF     MeanSq          F          pValue
            ________    __    ________    _________    __________

    x1         563.18     1      563.18       33.667    9.8742e-08
    x2         52.187     1      52.187       3.1197       0.08078
    x3       0.060046     1    0.060046    0.0035895       0.95236
    Error      1488.8    89      16.728
```

This `anova` display shows the following:

- First column — Terms included in the model.
- `SumSq` — Sum of squared error for each term except for the constant.
- `DF` — Degrees of freedom. In this example, `DF` is 1 for each term in the model and $n - p$ for the error term, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, the `DF` for the error term in this model is 93 – 4 = 89. If any of the variables in the model is a categorical variable, the `DF` for that variable is the number of indicator variables created for its categories (number of categories – 1).
- `MeanSq` — Mean squared error for each term. Note that `MeanSq = SumSq/DF`. For example, the mean squared error for the error term is 1488.8/89 = 16.728.
- `F` — *F*-values for each coefficient. The *F*-value is the ratio of the mean squared of each term and mean squared error, that is, `F = MeanSq(xi)/MeanSq(Error)`. Each *F*-statistic has an *F* distribution, with the numerator degrees of freedom, `DF` value for the corresponding term, and the denominator degrees of freedom, $n - p$. n is the number of observations, and p is the number of coefficients in the model. In this example, each *F*-statistic has an $F_{(1,89)}$ distribution.
- `pValue` — *p*-value for each hypothesis test on the coefficient of the corresponding term in the linear model. For example, the *p*-value for the *F*-statistic coefficient of `x2` is 0.08078, and is not significant at the 5% significance level given the other terms in the model.

Coefficient Confidence Intervals

Display coefficient confidence intervals.

```
coefCI(lm)
```

```
ans = 4×2

   40.2702   55.6833
   -0.0088   -0.0043
   -0.0913    0.0054
   -0.3957    0.3726
```

The values in each row are the lower and upper confidence limits, respectively, for the default 95% confidence intervals for the coefficients. For example, the first row shows the lower and upper limits, 40.2702 and 55.6833, for the intercept, β_0. Likewise, the second row shows the limits for β_1 and so on. Confidence intervals provide a measure of precision for linear regression coefficient estimates. A $100(1-\alpha)\,\%$ confidence interval gives the range the corresponding regression coefficient will be in with $100(1-\alpha)\,\%$ confidence.

You can also change the confidence level. Find the 99% confidence intervals for the coefficients.

```
coefCI(lm,0.01)
```

```
ans = 4×2

   37.7677   58.1858
   -0.0095   -0.0036
   -0.1069    0.0211
   -0.5205    0.4973
```

Hypothesis Test on Coefficients

Test the null hypothesis that all predictor variable coefficients are equal to zero versus the alternate hypothesis that at least one of them is different from zero.

```
[p,F,d] = coefTest(lm)
```

```
p = 7.3816e-27

F = 89.9874

d = 3
```

Here, `coefTest` performs an *F*-test for the hypothesis that all regression coefficients (except for the intercept) are zero versus at least one differs from zero, which essentially is the hypothesis on the model. It returns *p*, the *p*-value, `F`, the *F*-statistic, and `d`, the numerator degrees of freedom. The *F*-statistic and *p*-value are the same as the ones in the linear regression display and `anova` for the model. The degrees of freedom is 4 – 1 = 3 because there are four predictors (including the intercept) in the model.

Now, perform a hypothesis test on the coefficients of the first and second predictor variables.

```
H = [0 1 0 0; 0 0 1 0];
[p,F,d] = coefTest(lm,H)
```

```
p = 5.1702e-23

F = 96.4873

d = 2
```

The numerator degrees of freedom is the number of coefficients tested, which is 2 in this example. The results indicate that at least one of β_2 and β_3 differs from zero.

See Also

`LinearModel` | `anova` | `fitlm` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Linear Regression” on page 11-9
- “Stepwise Regression” on page 11-99
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “What Is a Linear Regression Model?” on page 11-6

More About

- “Coefficient Standard Errors and Confidence Intervals” on page 11-58
- “Coefficient of Determination (R-Squared)” on page 11-61
- “F-statistic and t-statistic” on page 11-72
- “Summary of Output and Diagnostic Statistics” on page 11-89

Cook’s Distance

Purpose

Cook’s distance is the scaled change in fitted values, which is useful for identifying outliers in the X values (observations for predictor variables). Cook’s distance shows the influence of each observation on the fitted response values. An observation with Cook’s distance larger than three times the mean Cook’s distance might be an outlier.

Definition

Each element in the Cook's distance D is the normalized change in the fitted response values due to the deletion of an observation. The Cook’s distance of observation i is

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{p\,MSE},$$

where

- $\widehat{y}_j$ is the jth fitted response value.
- $\widehat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook’s distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p\,MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right),$$

where r_i is the ith residual, and h_{ii} is the ith leverage value.

How To

After fitting the model `mdl`, for example, you can use `fitlm` or `stepwiselm` to:

- Display the Cook’s distance values by indexing into the property using dot notation.

  ```
  mdl.Diagnostics.CooksDistance
  ```

  `CooksDistance` is an n-by-1 column vector in the `Diagnostics` table of the `LinearModel` object.
- Plot the Cook’s distance values.

  ```
  plotDiagnostics(mdl,'cookd')
  ```

  For details, see the `plotDiagnostics` function of the `LinearModel` object.

Determine Outliers Using Cook's Distance

This example shows how to use Cook's Distance to determine the outliers in the data.

Load the sample data and define the independent and response variables.

```
load hospital
X = double(hospital(:,2:5));
y = hospital.BloodPressure(:,1);
```

Fit the linear regression model.

```
mdl = fitlm(X,y);
```

Plot the Cook's distance values.

```
plotDiagnostics(mdl,'cookd')
```

The dashed line in the figure corresponds to the recommended threshold value, `3*mean(mdl.Diagnostics.CooksDistance)`. The plot has some observations with Cook's distance values greater than the threshold value, which for this example is 3*(0.0108) = 0.0324. In particular, there are two Cook's distance values that are relatively higher than the others, which exceed the threshold value. You might want to find and omit these from your data and rebuild your model.

Find the observations with Cook's distance values that exceed the threshold value.

```
find((mdl.Diagnostics.CooksDistance)>3*mean(mdl.Diagnostics.CooksDistance))
```

```
ans = 10×1

     2
    13
    28
    44
    58
    70
    71
    84
    93
    95
```

Find the observations with Cook's distance values that are relatively larger than the other observations with Cook's distances exceeding the threshold value.

```
find((mdl.Diagnostics.CooksDistance)>5*mean(mdl.Diagnostics.CooksDistance))
```

```
ans = 2×1

     2
    84
```

References

[1] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. 4th ed. Chicago: Irwin, 1996.

See Also

`LinearModel` | `fitlm` | `plotDiagnostics` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Summary of Output and Diagnostic Statistics” on page 11-89

Coefficient Standard Errors and Confidence Intervals

In this section...

Coefficient Covariance and Standard Errors

Purpose

Estimated coefficient variances and covariances capture the precision of regression coefficient estimates. The coefficient variances and their square root, the standard errors, are useful in testing hypotheses for coefficients.

Definition

The estimated covariance matrix is

$$\sum = MSE(X'X)^{-1},$$

where MSE is the mean squared error, and X is the matrix of observations on the predictor variables. `CoefficientCovariance`, a property of the fitted model, is a p-by-p covariance matrix of regression coefficient estimates. p is the number of coefficients in the regression model. The diagonal elements are the variances of the individual coefficients.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can display the coefficient covariances using

```
mdl.CoefficientCovariance
```

Compute Coefficient Covariance and Standard Errors

This example shows how to compute the covariance matrix and standard errors of the coefficients.

Load the sample data and define the predictor and response variables.

```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

Display the coefficient covariance matrix.

```
CM = mdl.CoefficientCovariance
```

```
CM = 5×5

   27.5113   11.0027   -0.1542   -0.2444    0.2702
   11.0027    8.6864    0.0021   -0.1547   -0.0838
```

```
   -0.1542    0.0021    0.0045   -0.0001   -0.0029
   -0.2444   -0.1547   -0.0001    0.0031   -0.0026
    0.2702   -0.0838   -0.0029   -0.0026    1.0829
```

Compute the coefficient standard errors.

```
SE = diag(sqrt(CM))
```

```
SE = 5×1

    5.2451
    2.9473
    0.0673
    0.0557
    1.0406
```

Coefficient Confidence Intervals

Purpose

The coefficient confidence intervals provide a measure of precision for linear regression coefficient estimates. A 100(1–α)% confidence interval gives the range that the corresponding regression coefficient will be in with 100(1–α)% confidence.

Definition

The software finds confidence intervals using the Wald method. The 100*(1 – α)% confidence intervals for regression coefficients are

$$b_i \pm t_{(1-\alpha/2,\, n-p)} SE(b_i),$$

where b_i is the coefficient estimate, $SE(b_i)$ is the standard error of the coefficient estimate, and $t_{(1-\alpha/2,n-p)}$ is the 100(1 – α/2) percentile of t-distribution with n – p degrees of freedom. n is the number of observations and p is the number of regression coefficients.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can obtain the default 95% confidence intervals for coefficients using

```
coefCI(mdl)
```

You can also change the confidence level using

```
coefCI(mdl,alpha)
```

For details, see the `coefCI` function of `LinearModel` object.

Compute Coefficient Confidence Intervals

This example shows how to compute coefficient confidence intervals.

Load the sample data and fit a linear regression model.

```
load hald
mdl = fitlm(ingredients,heat);
```

Display the 95% coefficient confidence intervals.

```
coefCI(mdl)
```

```
ans = 5×2

  -99.1786  223.9893
   -0.1663    3.2685
   -1.1589    2.1792
   -1.6385    1.8423
   -1.7791    1.4910
```

The values in each row are the lower and upper confidence limits, respectively, for the default 95% confidence intervals for the coefficients. For example, the first row shows the lower and upper limits, -99.1786 and 223.9893, for the intercept, β_0 . Likewise, the second row shows the limits for β_1 and so on.

Display the 90% confidence intervals for the coefficients ($\alpha = 0.1$).

```
coefCI(mdl,0.1)
```

```
ans = 5×2

  -67.8949  192.7057
    0.1662    2.9360
   -0.8358    1.8561
   -1.3015    1.5053
   -1.4626    1.1745
```

The confidence interval limits become narrower as the confidence level decreases.

See Also

`LinearModel` | `anova` | `coefCI` | `coefTest` | `fitlm` | `plotDiagnostics` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “F-statistic and t-statistic” on page 11-72
- “Summary of Output and Diagnostic Statistics” on page 11-89

Coefficient of Determination (R-Squared)

Purpose

Coefficient of determination (R-squared) indicates the proportionate amount of variation in the response variable y explained by the independent variables X in the linear regression model. The larger the R-squared is, the more variability is explained by the linear regression model.

Definition

R-squared is the proportion of the total sum of squares explained by the model. `Rsquared`, a property of the fitted model, is a structure with two fields:

- `Ordinary` — Ordinary (unadjusted) R-squared

  $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

- `Adjusted` — R-squared adjusted for the number of coefficients

  $$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}.$$

  SSE is the sum of squared error, SSR is the sum of squared regression, SST is the sum of squared total, n is the number of observations, and p is the number of regression coefficients. Note that p includes the intercept, so for example, p is 2 for a linear fit. Because R-squared increases with added predictor variables in the regression model, the adjusted R-squared adjusts for the number of predictor variables in the model. This makes it more useful for comparing models with a different number of predictors.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can obtain either R-squared value as a scalar by indexing into the property using dot notation, for example,

```
mdl.Rsquared.Ordinary
mdl.Rsquared.Adjusted
```

You can also obtain the SSE, SSR, and SST using the properties with the same name.

```
mdl.SSE
mdl.SSR
mdl.SST
```

Display Coefficient of Determination

This example shows how to display R-squared (coefficient of determination) and adjusted R-squared. Load the sample data and define the response and independent variables.

```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y)

mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)        117.4      5.2451      22.383    1.1667e-39
    x1               0.88162      2.9473     0.29913       0.76549
    x2               0.08602     0.06731       1.278       0.20438
    x3             -0.016685    0.055714    -0.29947       0.76524
    x4                 9.884      1.0406       9.498    1.9546e-15


Number of observations: 100, Error degrees of freedom: 95
Root Mean Squared Error: 4.81
R-squared: 0.508,  Adjusted R-Squared: 0.487
F-statistic vs. constant model: 24.5, p-value = 5.99e-14
```

The R-squared and adjusted R-squared values are 0.508 and 0.487, respectively. Model explains about 50% of the variability in the response variable.

Access the R-squared and adjusted R-squared values using the property of the fitted `LinearModel` object.

```
mdl.Rsquared.Ordinary

ans = 0.5078

mdl.Rsquared.Adjusted

ans = 0.4871
```

The adjusted R-squared value is smaller than the ordinary R-squared value.

See Also

`LinearModel` | `anova` | `fitlm` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Summary of Output and Diagnostic Statistics” on page 11-89

Delete-1 Statistics

In this section...

Delete-1 Change in Covariance (CovRatio)

Purpose

Delete-1 change in covariance (`CovRatio`) identifies the observations that are influential in the regression fit. An influential observation is one where its exclusion from the model might significantly alter the regression function. Values of `CovRatio` larger than 1 + 3*p/n or smaller than 1 – 3*p/n indicate influential points, where p is the number of regression coefficients, and n is the number of observations.

Definition

The `CovRatio` statistic is the ratio of the determinant of the coefficient covariance matrix with observation i deleted to the determinant of the covariance matrix for the full model:

$$\text{CovRatio} = \frac{\det\left\{MSE(i)[X'(i)X(i)]^{-1}\right\}}{\det\left[MSE(X'X)^{-1}\right]}.$$

`CovRatio` is an n-by-1 vector in the `Diagnostics` table of the fitted `LinearModel` object. Each element is the ratio of the generalized variance of the estimated coefficients when the corresponding element is deleted to the generalized variance of the coefficients using all the data.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Display the `CovRatio` by indexing into the property using dot notation

  ```
  mdl.Diagnostics.CovRatio
  ```

- Plot the delete-1 change in covariance using

  ```
  plotDiagnostics(mdl,'CovRatio')
  ```

  For details, see the `plotDiagnostics` method of the `LinearModel` class.

Determine Influential Observations Using CovRatio

This example shows how to use the `CovRatio` statistics to determine the influential points in data. Load the sample data and define the response and predictor variables.

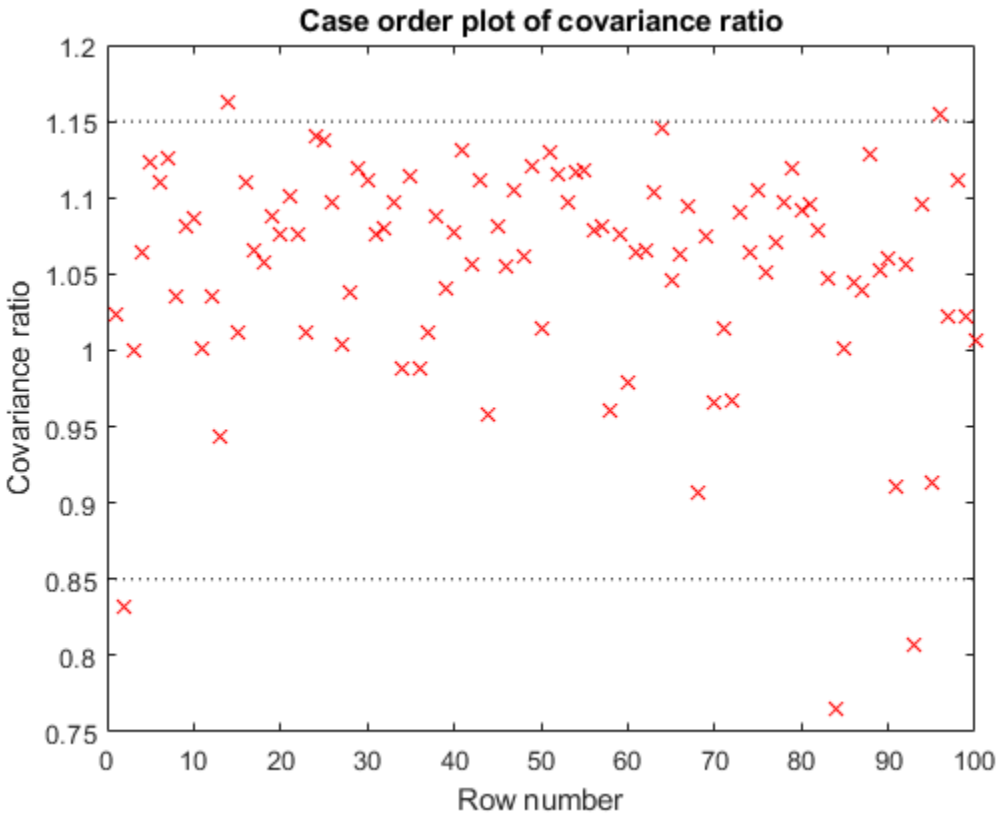
```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

Plot the `CovRatio` statistics.

```
plotDiagnostics(mdl,'CovRatio')
```



For this example, the threshold limits are 1 + 3*5/100 = 1.15 and 1 - 3*5/100 = 0.85. There are a few points beyond the limits, which might be influential points.

Find the observations that are beyond the limits.

```
find((mdl.Diagnostics.CovRatio)>1.15|(mdl.Diagnostics.CovRatio)<0.85)
```

```
ans = 5×1

     2
    14
    84
    93
    96
```

Delete-1 Scaled Difference in Coefficient Estimates (Dfbetas)

Purpose

The sign of a delete-1 scaled difference in coefficient estimate (`Dfbetas`) for coefficient j and observation i indicates whether that observation causes an increase or decrease in the estimate of the regression coefficient. The absolute value of a `Dfbetas` indicates the magnitude of the difference relative to the estimated standard deviation of the regression coefficient. A `Dfbetas` value larger than 3/sqrt(n) in absolute value indicates that the observation has a large influence on the corresponding coefficient.

Definition

`Dfbetas` for coefficient j and observation i is the ratio of the difference in the estimate of coefficient j using all observations and the one obtained by removing observation i, and the standard error of the coefficient estimate obtained by removing observation i. The `Dfbetas` for coefficient j and observation i is

$$Dfbetas_{ij} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)}}(1 - h_{ii})},$$

where b_j is the estimate for coefficient j, $b_{j(i)}$ is the estimate for coefficient j by removing observation i, $MSE_{(i)}$ is the mean squared error of the regression fit by removing observation i, and h_{ii} is the leverage value for observation i. `Dfbetas` is an n-by-p matrix in the `Diagnostics` table of the fitted `LinearModel` object. Each cell of `Dfbetas` corresponds to the `Dfbetas` value for the corresponding coefficient obtained by removing the corresponding observation.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can obtain the `Dfbetas` values as an n-by-p matrix by indexing into the property using dot notation,

```
mdl.Diagnostics.Dfbetas
```

Determine Observations Influential on Coefficients Using Dfbetas

This example shows how to determine the observations that have large influence on coefficients using `Dfbetas`. Load the sample data and define the response and independent variables.

```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

Find the `Dfbetas` values that are high in absolute value.

```
[row,col] = find(abs(mdl.Diagnostics.Dfbetas)>3/sqrt(100));
disp([row col])
```

```
     2     1
    28     1
    84     1
    93     1
```

```
     2     2
    13     3
    84     3
     2     4
    84     4
```

Delete-1 Scaled Change in Fitted Values (Dffits)

Purpose

The delete-1 scaled change in fitted values (`Dffits`) show the influence of each observation on the fitted response values. `Dffits` values with an absolute value larger than 2*sqrt(*p*/*n*) might be influential.

Definition

`Dffits` for observation *i* is

$$\text{Dffits}_i = sr_i\sqrt{\frac{h_{ii}}{1-h_{ii}}},$$

where sr_i is the studentized residual, and h_{ii} is the leverage value of the fitted `LinearModel` object. `Dffits` is an *n*-by-1 column vector in the `Diagnostics` table of the fitted `LinearModel` object. Each element in `Dffits` is the change in the fitted value caused by deleting the corresponding observation and scaling by the standard error.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Display the `Dffits` values by indexing into the property using dot notation

  ```
  mdl.Diagnostics.Dffits
  ```

- Plot the delete-1 scaled change in fitted values using

  ```
  plotDiagnostics(mdl,'Dffits')
  ```

  For details, see the `plotDiagnostics` method of the `LinearModel` class for details.

Determine Observations Influential on Fitted Response Using Dffits

This example shows how to determine the observations that are influential on the fitted response values using `Dffits` values. Load the sample data and define the response and independent variables.

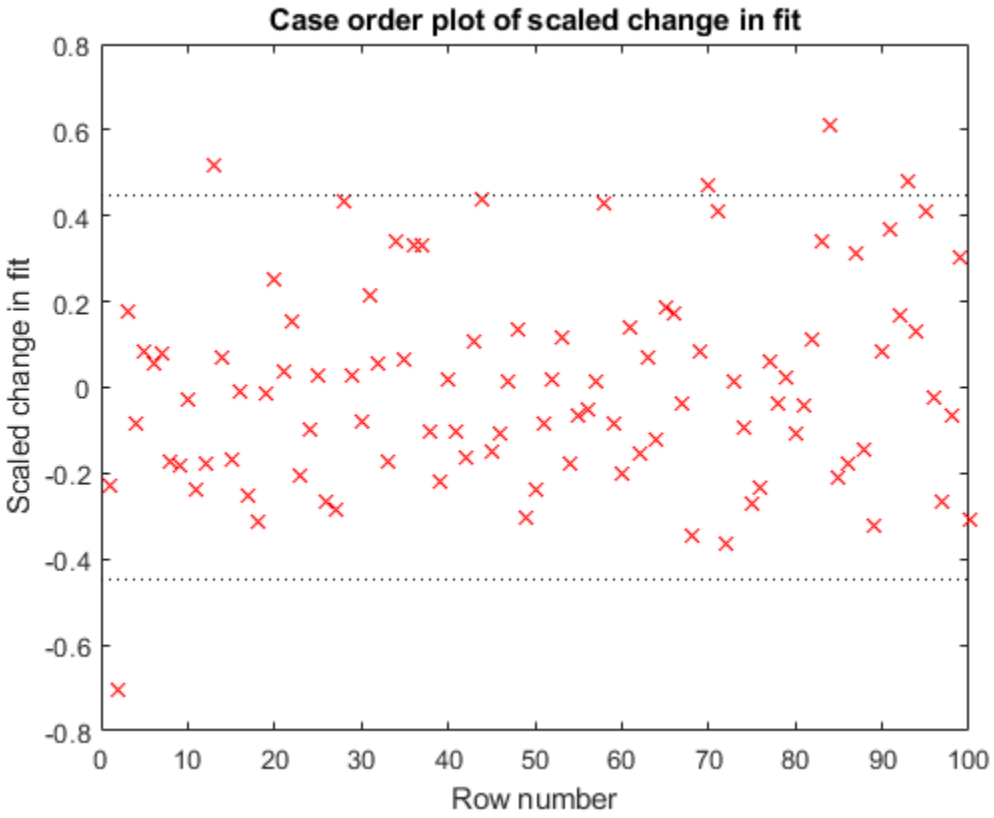
```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

Plot the `Dffits` values.

```
plotDiagnostics(mdl,'Dffits')
```



The influential threshold limit for the absolute value of `Dffits` in this example is 2*sqrt(5/100) = 0.45. Again, there are some observations with `Dffits` values beyond the recommended limits.

Find the `Dffits` values that are large in absolute value.

```
find(abs(mdl.Diagnostics.Dffits)>2*sqrt(4/100))
```

```
ans = 10×1

     2
    13
    28
    44
    58
    70
    71
    84
    93
    95
```

Delete-1 Variance (S2_i)

Purpose

The delete-1 variance (`S2_i`) shows how the mean squared error changes when an observation is removed from the data set. You can compare the `S2_i` values with the value of the mean squared error.

Definition

`S2_i` is a set of residual variance estimates obtained by deleting each observation in turn. The `S2_i` value for observation i is

$$S2\_i = MSE_{(i)} = \frac{\sum_{j \neq i}^{n} \left[y_j - \hat{y}_{j(i)} \right]^2}{n - p - 1},$$

where y_j is the jth observed response value. `S2_i` is an n-by-1 vector in the `Diagnostics` table of the fitted `LinearModel` object. Each element in `S2_i` is the mean squared error of the regression obtained by deleting that observation.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Display the `S2_i` vector by indexing into the property using dot notation

  ```
  mdl.Diagnostics.S2_i
  ```

- Plot the delete-1 variance values using

  ```
  plotDiagnostics(mdl,'S2_i')
  ```

  For details, see the `plotDiagnostics` method of the `LinearModel` class.

Compute and Examine Delete-1 Variance Values

This example shows how to compute and plot S2_i values to examine the change in the mean squared error when an observation is removed from the data. Load the sample data and define the response and independent variables.

```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

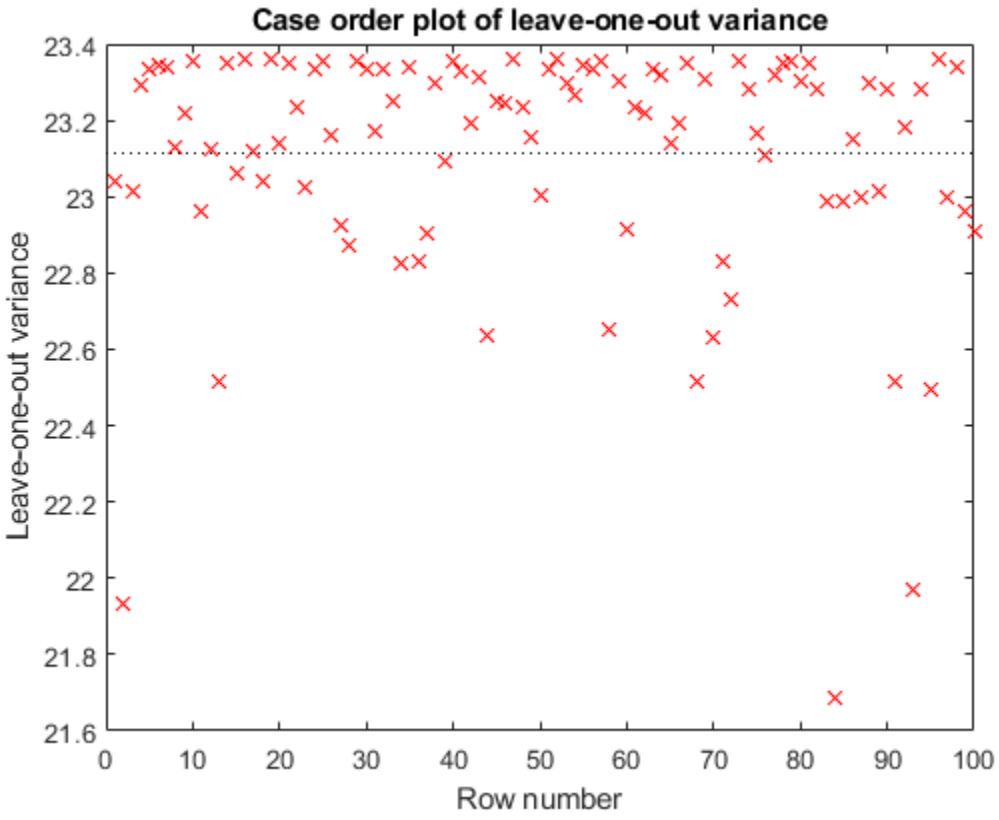
Display the MSE value for the model.

```
mdl.MSE
```

```
ans = 23.1140
```

Plot the S2_i values.

```
plotDiagnostics(mdl,'S2_i')
```



This plot makes it easy to compare the S2_i values to the MSE value of 23.114, indicated by the horizontal dashed lines. You can see how deleting one observation changes the error variance.

See Also

`LinearModel` | `fitlm` | `plotDiagnostics` | `plotResiduals` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Summary of Output and Diagnostic Statistics” on page 11-89

Durbin-Watson Test

Purpose

The Durbin-Watson test assesses whether or not there is autocorrelation among the residuals of time series data.

Definition

The Durbin-Watson test statistic, `DW`, is

$$DW = \frac{\sum_{i=1}^{n-1} (r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2},$$

where r_i is the ith raw residual, and n is the number of observations.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can perform the Durbin-Watson test using

```
dwtest(mdl)
```

For details, see the `dwtest` method of the `LinearModel` class.

Test for Autocorrelation Among Residuals

This example shows how to test for autocorrelation among the residuals of a linear regression model.

Load the sample data and fit a linear regression model.

```
load hald
mdl = fitlm(ingredients,heat);
```

Perform a two-sided Durbin-Watson test to determine if there is any autocorrelation among the residuals of the linear model, `mdl`.

```
[p,DW] = dwtest(mdl,'exact','both')
```

```
p = 0.8421
```

```
DW = 2.0526
```

The value of the Durbin-Watson test statistic is 2.0526. The p-value of 0.8421 suggests that the residuals are not autocorrelated.

See Also

`LinearModel` | `dwtest` | `fitlm` | `plotResiduals` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Summary of Output and Diagnostic Statistics” on page 11-89

F-statistic and t-statistic

In this section...

F-statistic

Purpose

In linear regression, the F-statistic is the test statistic for the analysis of variance (ANOVA) approach to test the significance of the model or the components in the model.

Definition

The F-statistic in the linear model output display is the test statistic for testing the statistical significance of the model. The F-statistic values in the `anova` display are for assessing the significance of the terms or components in the model.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Find the `F-statistic vs. constant model` in the output display or by using

  ```
  disp(mdl)
  ```
- Display the ANOVA for the model using

  ```
  anova(mdl,'summary')
  ```
- Obtain the F-statistic values for the components, except for the constant term using

  ```
  anova(mdl)
  ```

  For details, see the `anova` method of the `LinearModel` class.

Assess Fit of Model Using F-statistic

This example shows how to assess the fit of the model and the significance of the regression coefficients using the F-statistic.

Load the sample data.

```
load hospital
tbl = table(hospital.Age,hospital.Weight,hospital.Smoker,hospital.BloodPressure(:,1), ...
      'VariableNames',{'Age','Weight','Smoker','BloodPressure'});
tbl.Smoker = categorical(tbl.Smoker);
```

Fit a linear regression model.

```
mdl = fitlm(tbl,'BloodPressure ~ Age*Weight + Smoker + Weight^2')
```

```
mdl =
Linear regression model:
    BloodPressure ~ 1 + Smoker + Age*Weight + Weight^2

Estimated Coefficients:
                    Estimate        SE         tStat        pValue
                   __________    _________    ________    __________

    (Intercept)        168.02       27.694       6.067    2.7149e-08
    Age              0.079569      0.39861     0.19962       0.84221
    Weight            -0.69041       0.3435     -2.0099      0.047305
    Smoker_true         9.8027       1.0256      9.5584    1.5969e-15
    Age:Weight     0.00021796    0.0025258    0.086294       0.93142
    Weight^2        0.0021877    0.0011037      1.9822      0.050375


Number of observations: 100, Error degrees of freedom: 94
Root Mean Squared Error: 4.73
R-squared: 0.528,  Adjusted R-Squared: 0.503
F-statistic vs. constant model: 21, p-value = 4.81e-14
```

The F-statistic of the linear fit versus the constant model is 21, with a *p*-value of 4.81e-14. The model is significant at the 5% significance level. The R-squared value of 0.528 means the model explains about 53% of the variability in the response. There might be other predictor (explanatory) variables that are not included in the current model.

Display the ANOVA table for the fitted model.

```
anova(mdl,'summary')
```

```
ans=5×5 table
                    SumSq     DF    MeanSq      F         pValue
                    ______    __    ______    ______    __________

    Total           4461.2    99    45.062
    Model           2354.5     5     470.9    21.012    4.8099e-14
    . Linear        2263.3     3    754.42    33.663    7.2417e-15
    . Nonlinear     91.248     2    45.624    2.0358        0.1363
    Residual        2106.6    94    22.411
```

This display separates the variability in the model into linear and nonlinear terms. Since there are two non-linear terms (`Weight^2` and the interaction between `Weight` and `Age`), the nonlinear degrees of freedom in the `DF` column is 2. There are three linear terms in the model (one `Smoker` indicator variable, `Weight`, and `Age`). The corresponding F-statistics in the `F` column are for testing the significance of the linear and nonlinear terms as separate groups.

When there are replicated observations, the residual term is also separated into two parts; first is the error due to the lack of fit, and second is the pure error independent from the model, obtained from the replicated observations. In that case, the F-statistic is for testing the lack of fit, that is, whether the fit is adequate or not. But, in this example, there are no replicated observations.

Display the ANOVA table for the model terms.

```
anova(mdl)
```

```
ans=6×5 table
                    SumSq      DF     MeanSq         F          pValue
```

```
                ________    __    ________    _________    __________

Age               62.991     1      62.991       2.8107      0.096959
Weight          0.064104     1    0.064104    0.0028604       0.95746
Smoker            2047.5     1      2047.5       91.363    1.5969e-15
Age:Weight       0.16689     1     0.16689    0.0074466       0.93142
Weight^2          88.057     1      88.057       3.9292      0.050375
Error             2106.6    94      22.411
```

This display decomposes the ANOVA table into the model terms. The corresponding F-statistics in the `F` column assess the statistical significance of each term. For example, the F-test for `Smoker` tests whether the coefficient of the indicator variable for `Smoker` is different from zero. That is, the F-test determines whether being a smoker has a significant effect on `BloodPressure`. The degrees of freedom for each model term is the numerator degrees of freedom for the corresponding F-test. All the terms have one degree of freedom. In the case of a categorical variable, the degrees of freedom is the number of indicator variables. `Smoker` has only one indicator variable, so it also has one degree of freedom.

t-statistic

Purpose

In linear regression, the *t*-statistic is useful for making inferences about the regression coefficients. The hypothesis test on coefficient *i* tests the null hypothesis that it is equal to zero – meaning the corresponding term is not significant – versus the alternate hypothesis that the coefficient is different from zero.

Definition

For a hypotheses test on coefficient *i*, with

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$,

the *t*-statistic is:

$$t = \frac{b_i}{SE(b_i)},$$

where $SE(b_i)$ is the standard error of the estimated coefficient b_i.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Find the coefficient estimates, the standard errors of the estimates (`SE`), and the *t*-statistic values of hypothesis tests for the corresponding coefficients (`tStat`) in the output display.
- Call for the display using

  ```
  display(mdl)
  ```

Assess Significance of Regression Coefficients Using t-statistic

This example shows how to test for the significance of the regression coefficients using t-statistic.

Load the sample data and fit the linear regression model.

```
load hald
mdl = fitlm(ingredients,heat)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    ________

    (Intercept)      62.405     70.071      0.8906     0.39913
    x1               1.5511    0.74477      2.0827    0.070822
    x2              0.51017    0.72379     0.70486      0.5009
    x3              0.10191    0.75471     0.13503     0.89592
    x4             -0.14406    0.70905    -0.20317     0.84407


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 2.45
R-squared: 0.982,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 111, p-value = 4.76e-07
```

You can see that for each coefficient, `tStat = Estimate/SE`. The *p*-values for the hypotheses tests are in the `pValue` column. Each *t*-statistic tests for the significance of each term given other terms in the model. According to these results, none of the coefficients seem significant at the 5% significance level, although the R-squared value for the model is really high at 0.97. This often indicates possible multicollinearity among the predictor variables.

Use stepwise regression to decide which variables to include in the model.

```
load hald
mdl = stepwiselm(ingredients,heat)
```

```
1. Adding x4, FStat = 22.7985, pValue = 0.000576232
2. Adding x1, FStat = 108.2239, pValue = 1.105281e-06

mdl =
Linear regression model:
    y ~ 1 + x1 + x4

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)      103.1        2.124      48.54    3.3243e-13
    x1                1.44      0.13842     10.403    1.1053e-06
    x4              -0.61395    0.048645    -12.621    1.8149e-07


Number of observations: 13, Error degrees of freedom: 10
Root Mean Squared Error: 2.73
```

```
R-squared: 0.972,  Adjusted R-Squared: 0.967
F-statistic vs. constant model: 177, p-value = 1.58e-08
```

In this example, `stepwiselm` starts with the constant model (default) and uses forward selection to incrementally add `x4` and `x1`. Each predictor variable in the final model is significant given the other one is in the model. The algorithm stops when adding none of the other predictor variables significantly improves in the model. For details on stepwise regression, see `stepwiselm`.

See Also

`LinearModel` | `anova` | `coefCI` | `coefTest` | `fitlm` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Coefficient Standard Errors and Confidence Intervals” on page 11-58
- “Summary of Output and Diagnostic Statistics” on page 11-89

Hat Matrix and Leverage

In this section...

Hat Matrix

Purpose

The hat matrix provides a measure of leverage. It is useful for investigating whether one or more observations are outlying with regard to their X values, and therefore might be excessively influencing the regression results.

Definition

The hat matrix is also known as the *projection matrix* because it projects the vector of observations, y, onto the vector of predictions, $\widehat{y}$, thus putting the "hat" on y. The hat matrix H is defined in terms of the data matrix X:

$$H = X(X^TX)^{-1}X^T$$

and determines the fitted or predicted values since

$$\widehat{y} = Hy = Xb\,.$$

The diagonal elements of H, h_{ii}, are called leverages and satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where p is the number of coefficients, and n is the number of observations (rows of X) in the regression model. `HatMatrix` is an n-by-n matrix in the `Diagnostics` table.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Display the `HatMatrix` by indexing into the property using dot notation

  ```
  mdl.Diagnostics.HatMatrix
  ```

  When n is large, `HatMatrix` might be computationally expensive. In those cases, you can obtain the diagonal values directly, using

  ```
  mdl.Diagnostics.Leverage
  ```

Leverage

Purpose

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs. In general, the farther a point is from the center of the input space, the more leverage it has. Because the sum of the leverage values is p, an observation i can be considered as an outlier if its leverage substantially exceeds the mean leverage value, p/n, for example, a value larger than $2*p/n$.

Definition

The leverage of observation i is the value of the ith diagonal term, h_{ii}, of the hat matrix, H, where

$H = X(X^TX)^{-1}X^T$.

The diagonal terms satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where p is the number of coefficients in the regression model, and n is the number of observations. The minimum value of h_{ii} is $1/n$ for a model with a constant term. If the fitted model goes through the origin, then the minimum leverage value is 0 for an observation at $x = 0$.

It is possible to express the fitted values, $\widehat{y}$, by the observed values, y, since

$$\widehat{y} = Hy = Xb\,.$$

Hence, h_{ii} expresses how much the observation y_i has impact on $\widehat{y}_i$. A large value of h_{ii} indicates that the ith case is distant from the center of all X values for all n cases and has more leverage. `Leverage` is an n-by-1 column vector in the `Diagnostics` table.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Display the `Leverage` vector by indexing into the property using dot notation

  ```
  mdl.Diagnostics.Leverage
  ```

- Plot the leverage for the values fitted by your model using

  ```
  plotDiagnostics(mdl)
  ```

  See the `plotDiagnostics` method of the `LinearModel` class for details.

Determine High Leverage Observations

This example shows how to compute `Leverage` values and assess high leverage observations. Load the sample data and define the response and independent variables.

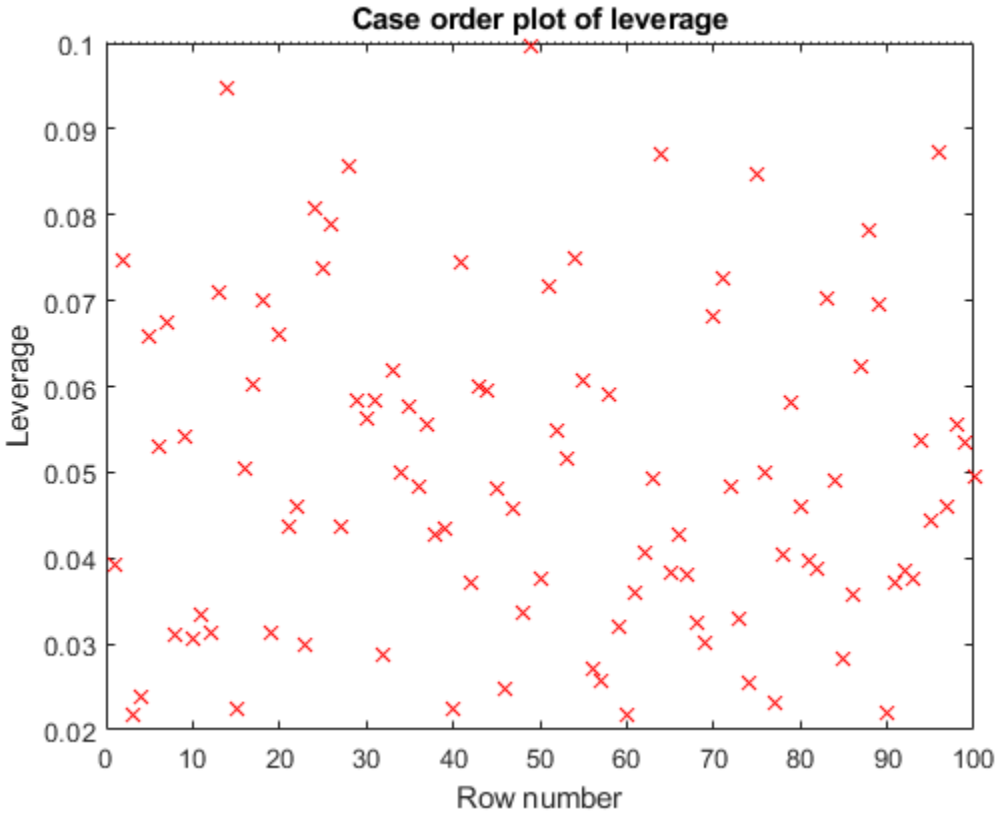
```
load hospital
y = hospital.BloodPressure(:,1);
X = double(hospital(:,2:5));
```

Fit a linear regression model.

```
mdl = fitlm(X,y);
```

Plot the leverage values.

```
plotDiagnostics(mdl)
```



For this example, the recommended threshold value is 2*5/100 = 0.1. There is no indication of high leverage observations.

See Also

`LinearModel` | `fitlm` | `plotDiagnostics` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50
- “Summary of Output and Diagnostic Statistics” on page 11-89

Residuals

Purpose

Residuals are useful for detecting outlying y values and checking the linear regression assumptions with respect to the error term in the regression model. High-leverage observations have smaller residuals because they often shift the regression line or surface closer to them. You can also use residuals to detect some forms of heteroscedasticity and autocorrelation.

Definition

The `Residuals` matrix is an n-by-4 table containing four types of residuals, with one row for each observation.

Raw Residuals

Observed minus fitted values, that is,

$$r_i = y_i - \widehat{y}_i.$$

Pearson Residuals

Raw residuals divided by the root mean squared error, that is,

$$pr_i = \frac{r_i}{\sqrt{MSE}},$$

where r_i is the raw residual and MSE is the mean squared error.

Standardized Residuals

Standardized residuals are raw residuals divided by their estimated standard deviation. The standardized residual for observation i is

$$st_i = \frac{r_i}{\sqrt{MSE(1-h_{ii})}},$$

where MSE is the mean squared error and h_{ii} is the leverage value for observation i.

Studentized Residuals

Studentized residuals are the raw residuals divided by an independent estimate of the residual standard deviation. The residual for observation i is divided by an estimate of the error standard deviation based on all observations except for observation i.

$$sr_i = \frac{r_i}{\sqrt{MSE_{(i)}(1-h_{ii})}},$$

where $MSE_{(i)}$ is the mean squared error of the regression fit calculated by removing observation i, and h_{ii} is the leverage value for observation i. The studentized residual sr_i has a t-distribution with $n - p - 1$ degrees of freedom.

How To

After obtaining a fitted model, say, `mdl`, using `fitlm` or `stepwiselm`, you can:

- Find the `Residuals` table under `mdl` object.
- Obtain any of these columns as a vector by indexing into the property using dot notation, for example,

  ```
  mdl.Residuals.Raw
  ```

- Plot any of the residuals for the values fitted by your model using

  ```
  plotResiduals(mdl)
  ```

  For details, see the `plotResiduals` method of the `LinearModel` class.

Assess Model Assumptions Using Residuals

This example shows how to assess the model assumptions by examining the residuals of a fitted linear regression model.

Load the sample data and store the independent and response variables in a table.

```
 load imports-85
 tbl = table(X(:,7),X(:,8),X(:,9),X(:,15),'VariableNames',...
{'curb_weight','engine_size','bore','price'});
```

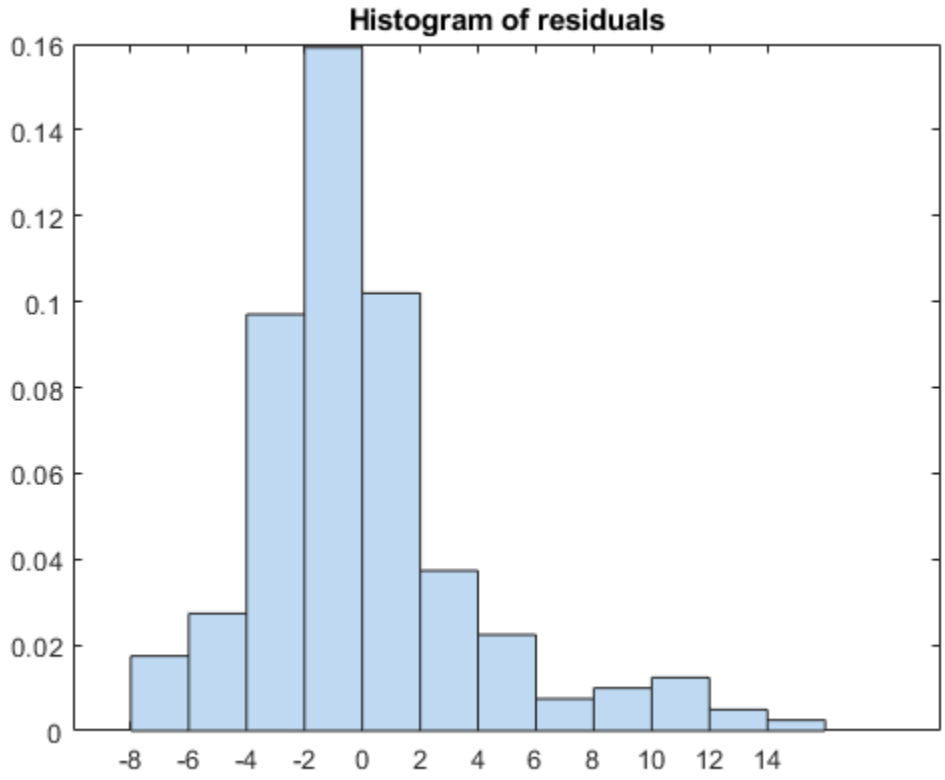
Fit a linear regression model.

```
mdl = fitlm(tbl)
```

```
mdl =
Linear regression model:
    price ~ 1 + curb_weight + engine_size + bore

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    _______    __________

    (Intercept)        64.095        3.703     17.309    2.0481e-41
    curb_weight    -0.0086681    0.0011025    -7.8623      2.42e-13
    engine_size     -0.015806     0.013255    -1.1925       0.23452
    bore              -2.6998       1.3489    -2.0015      0.046711


Number of observations: 201, Error degrees of freedom: 197
Root Mean Squared Error: 3.95
R-squared: 0.674,  Adjusted R-Squared: 0.669
F-statistic vs. constant model: 136, p-value = 1.14e-47
```

Plot the histogram of raw residuals.

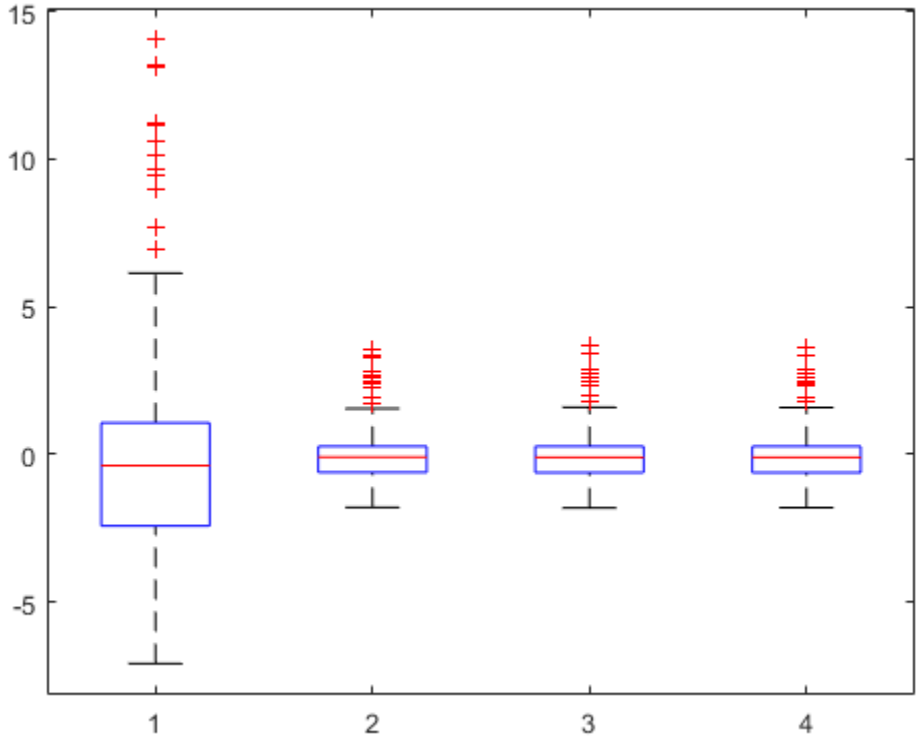
```
plotResiduals(mdl)
```



The histogram shows that the residuals are slightly right skewed.

Plot the box plot of all four types of residuals.

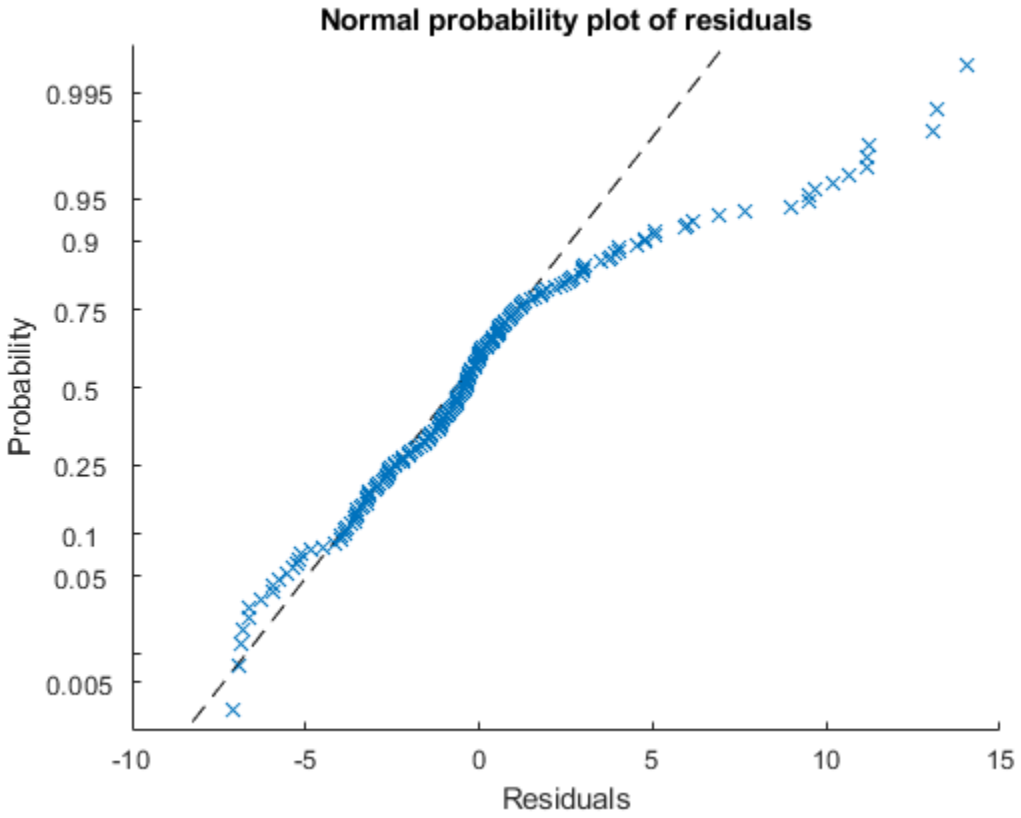
```
Res = table2array(mdl.Residuals);
boxplot(Res)
```



You can see the right-skewed structure of the residuals in the box plot as well.

Plot the normal probability plot of the raw residuals.

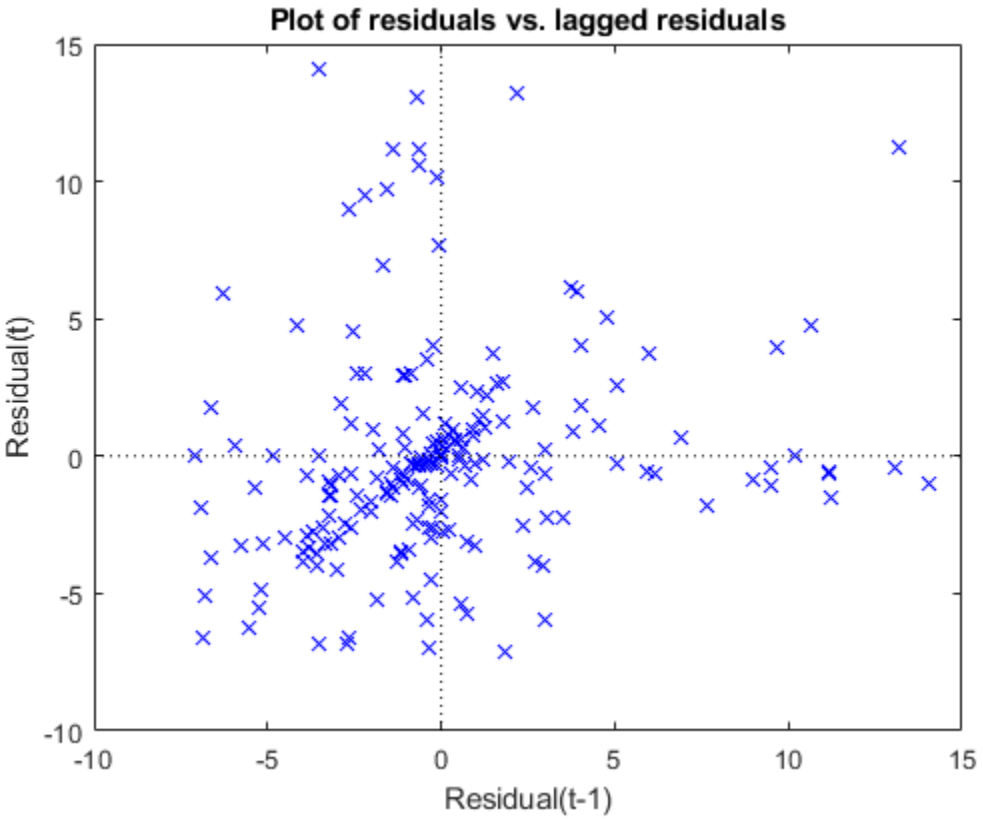
```
plotResiduals(mdl,'probability')
```



This normal probability plot also shows the deviation from normality and the skewness on the right tail of the distribution of residuals.

Plot the residuals versus lagged residuals.

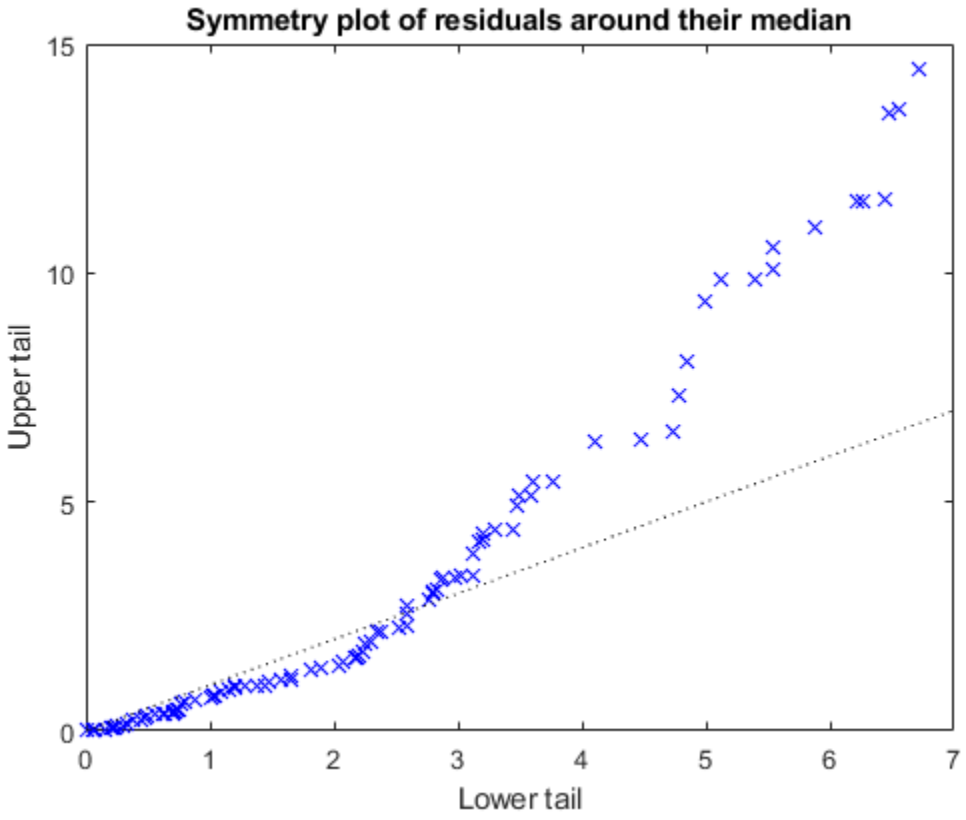
```
plotResiduals(mdl,'lagged')
```



This graph shows a trend, which indicates a possible correlation among the residuals. You can further check this using `dwtest(mdl)`. Serial correlation among residuals usually means that the model can be improved.

Plot the symmetry plot of residuals.

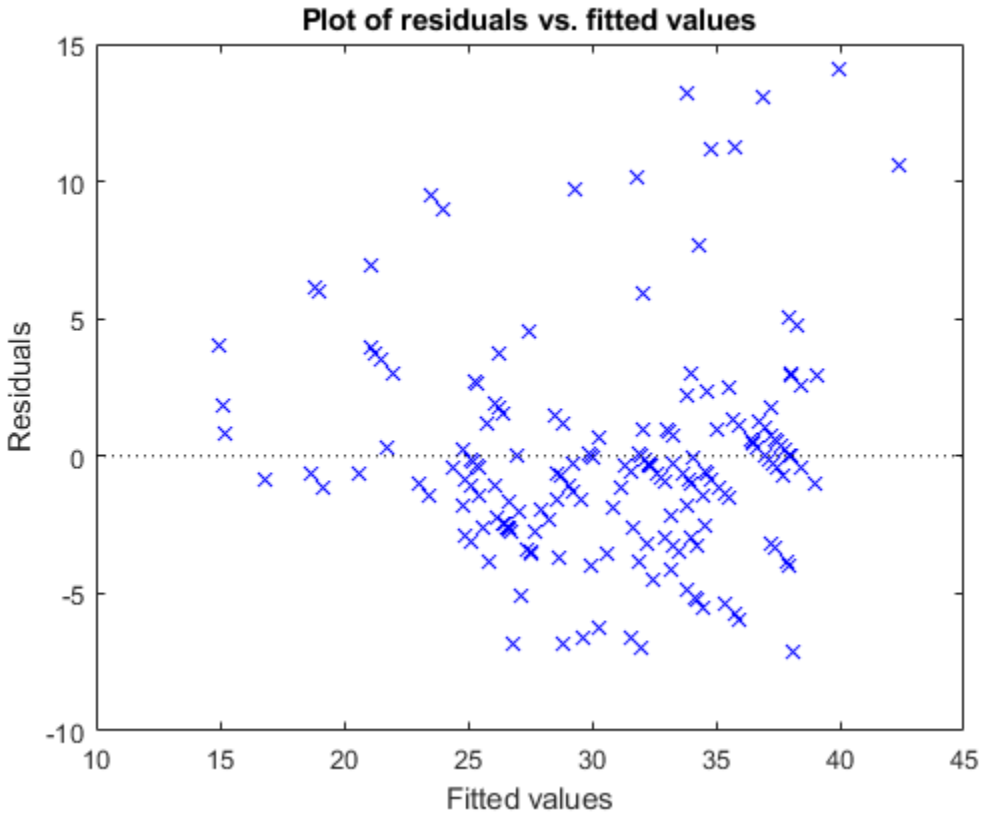
```
plotResiduals(mdl,'symmetry')
```



This plot also suggests that the residuals are not distributed equally around their median, as would be expected for normal distribution.

Plot the residuals versus the fitted values.

```
plotResiduals(mdl,'fitted')
```



The increase in the variance as the fitted values increase suggests possible heteroscedasticity.

References

[1] Atkinson, A. T. *Plots, Transformations, and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis.* New York: Oxford Statistical Science Series, Oxford University Press, 1987.

[2] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*. IRWIN, The McGraw-Hill Companies, Inc., 1996.

[3] Belsley, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, Inc., 1980.

See Also

`LinearModel` | `dwtest` | `fitlm` | `plotDiagnostics` | `plotResiduals` | `stepwiselm`

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50

Summary of Output and Diagnostic Statistics

Diagnostic Statistics	`LinearModel` Properties and Functions	Field Names of `regstats` output
"Cook's Distance" on page 11-55	• See the `CooksDistance` field of the `Diagnostics` property. • Use the `plotDiagnostics` function with the `'cookd'` plot type.	`cookd`
"Coefficient Covariance and Standard Errors" on page 11-58	See the `CoefficientCovariance` property.	`covb`
"Coefficient Confidence Intervals" on page 11-59	Use the `coefCI` function.	N/A
"Coefficient of Determination (R-Squared)" on page 11-61	See the `Ordinary` and `Adjusted` fields of the `Rsquared` property.	`rsquare`, `adjrsquare`
"Delete-1 Change in Covariance (CovRatio)" on page 11-63	• See the `CovRatio` field of the `Diagnostics` property. • Use the `plotDiagnostics` function with the `'covratio'` plot type.	`covratio`
"Delete-1 Scaled Difference in Coefficient Estimates (Dfbetas)" on page 11-65	See the `Dfbetas` field of the `Diagnostics` property.	`dfbetas`
"Delete-1 Scaled Change in Fitted Values (Dffits)" on page 11-66	• See the `Dffits` field of the `Diagnostics` property. • Use the `plotDiagnostics` function with the `'dffits'` plot type.	`dffits`
"Delete-1 Variance (S2_i)" on page 11-68	• See the `S2_i` field of the `Diagnostics` property. • Use the `plotDiagnostics` with the `'s2_i'` plot type.	`s2_i`
"Durbin-Watson Test" on page 11-70	Use the `dwtest` function.	`dwstat`
"F-statistic" on page 11-72	• Find the value of `F-statistic vs. constant model` in the fitted model display at command line. • Use the `anova` function.	`fstat`
"Hat Matrix" on page 11-77	See the `HatMatrix` field of the `Diagnostics` property.	`hatmat`

Diagnostic Statistics	LinearModel Properties and Functions	Field Names of regstats output
“Leverage” on page 11-78	• See the Leverage field of the Diagnostics property. • Use the plotDiagnostics function with the 'leverage' plot type.	leverage
“Residuals” on page 11-80	• See the Raw, Pearson, Standardized, and Studentized fields of the Residuals property. • Use the plotResiduals function.	r, studres, standres
“t-statistic” on page 11-74	See the tStat column of the Coefficients property.	tstat

See Also

LinearModel | fitlm | stepwiselm

Related Examples

- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Interpret Linear Regression Results” on page 11-50

Wilkinson Notation

In this section...

Overview

Wilkinson notation provides a way to describe regression and repeated measures models without specifying coefficient values. This specialized notation identifies the response variable and which predictor variables to include or exclude from the model. You can also include squared and higher-order terms, interaction terms, and grouping variables in the model formula.

Specifying a model using Wilkinson notation provides several advantages:

- You can include or exclude individual predictors and interaction terms from the model. For example, using the `'Interactions'` name-value pair available in each model fitting functions includes interaction terms for all pairs of variables. Using Wilkinson notation instead allows you to include only the interaction terms of interest.
- You can change the model formula without changing the design matrix, if your input data uses the `table` data type. For example, if you fit an initial model using all the available predictor variables, but decide to remove a variable that is not statistically significant, then you can re-write the model formula to include only the variables of interest. You do not need to make any changes to the input data itself.

Statistics and Machine Learning Toolbox offers several model fitting functions that use Wilkinson notation, including:

- Linear models (using `fitlm` and `stepwiselm`)
- Generalized linear models (using `fitglm`)
- Linear mixed-effects models (using `fitlme` and `fitlmematrix`)
- Generalized linear mixed-effects models (using `fitglme`)
- Repeated measures models (using `fitrm`)
- Cox proportional hazards model (using `fitcox`)

Formula Specification

A formula for model specification is a character vector or string scalar of the form `y ~ terms`, where `y` is the name of the response variable, and `terms` defines the model using the predictor variable names and the following operators.

Predictor Variables

Predictor Terms in Model	Wilkinson Notation
intercept	`1`
no intercept	`-1`
x_1	`x1`
x_1, x_2	`x1 + x2`
x_1, x_2, x_1x_2	`x1*x2` or `x1 + x2 + x1:x2`
x_1x_2	`x1:x2`
x_1, x_1^2	`x1^2`
x_1^2	`x1^2 - x1`

Wilkinson notation includes an intercept term in the model by default, even if you do not add 1 to the model formula. To exclude the intercept from the model, use -1 in the formula.

The * operator (for interactions) and the ^ operator (for power and exponents) automatically include all lower-order terms. For example, if you specify `x^3`, the model will automatically include x^3, x^2, and x. If you want to exclude certain variables from the model, use the − operator to remove the unwanted terms.

Random-Effects and Mixed-Effects Models

For random-effects and mixed-effects models, the formula specification includes the names of the predictor variables and the grouping variables. For example, if the predictor variable x_1 is a random effect grouped by the variable g, then represent this in Wilkinson notation as follows:

```
(x1 | g)
```

Repeated Measures Models

For repeated measures models, the formula specification includes all of the repeated measures as responses, and the factors as predictor variables. Specify the response variables for repeated measures models as described in the following table.

Response Terms in Model	Wilkinson Notation
y_1	`y1`
y_1, y_2, y_3	`y1,y2,y3`
y_1, y_2, y_3, y_4, y_5	`y1-y5`

For example, if you have three repeated measures as responses and the factors x_1, x_2, and x_3 as the predictor variables, then you can define the repeated measures model using Wilkinson notation as follows:

```
y1,y2,y3 ~ x1 + x2 + x3
```

or

```
y1-y3 ~ x1 + x2 + x3
```

Variable Names

If the input data (response and predictor variables) is stored in a table or dataset array, you can specify the formula using the variable names. For example, load the `carsmall` sample data. Create a table containing `Weight`, `Acceleration`, and `MPG`. Name each variable using the `'VariableNames'` name-value pair argument of the fitting function `fitlm`. Then fit the following model to the data:

$$MPG = \beta_0 + \beta_1 Weight + \beta_2 Acceleration$$

```
load carsmall
tbl = table(Weight,Acceleration,MPG, ...
    'VariableNames',{'Weight','Acceleration','MPG'});
mdl = fitlm(tbl,'MPG ~ Weight + Acceleration')
```

```
mdl =


Linear regression model:
    MPG ~ 1 + Weight + Acceleration

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________

    (Intercept)         45.155        3.4659     13.028    1.6266e-22
    Weight          -0.0082475    0.00059836    -13.783    5.3165e-24
    Acceleration       0.19694       0.14743     1.3359       0.18493


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 4.12
R-squared: 0.743,  Adjusted R-Squared: 0.738
F-statistic vs. constant model: 132, p-value = 1.38e-27
```

The model object display uses the variable names provided in the input table.

If the input data is stored as a matrix, you can specify the formula using default variable names such as `y`, `x1`, and `x2`. For example, load the `carsmall` sample data. Create a matrix containing the predictor variables `Weight` and `Acceleration`. Then fit the following model to the data:

$$MPG = \beta_0 + \beta_1 Weight + \beta_2 Acceleration$$

```
load carsmall
X = [Weight,Acceleration];
y = MPG;
mdl = fitlm(X,y,'y ~ x1 + x2')
```

```
mdl =


Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________
```

```
    (Intercept)        45.155        3.4659     13.028    1.6266e-22
    x1             -0.0082475    0.00059836    -13.783    5.3165e-24
    x2                0.19694       0.14743     1.3359       0.18493


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 4.12
R-squared: 0.743,  Adjusted R-Squared: 0.738
F-statistic vs. constant model: 132, p-value = 1.38e-27
```

The term `x1` in the model specification formula corresponds to the first column of the predictor variable matrix `X`. The term `x2` corresponds to the second column of the input matrix. The term `y` corresponds to the response variable.

Linear Model Examples

Use `fitlm` and `stepwiselm` to fit linear models.

Intercept and Two Predictors

For a linear regression model with an intercept and two fixed-effects predictors, such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1 + x2'
```

No Intercept and Two Predictors

For a linear regression model with no intercept and two fixed-effects predictors, such as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i\,,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ -1 + x1 + x2'
```

Intercept, Two Predictors, and an Interaction Term

For a linear regression model with an intercept, two fixed-effects predictors, and an interaction term, such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i\,,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1*x2'
```

or

```
'y ~ x1 + x2 + x1:x2'
```

Intercept, Three Predictors, and All Interaction Effects

For a linear regression model with an intercept, three fixed-effects predictors, and interaction effects between all three predictors plus all lower-order terms, such as

$$y_i = \beta_0 + \beta_1 xi_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_1 x_{i2} + \beta_5 x_1 x_{i3} + \beta_6 x_2 x_{i3} + \beta_7 x_{i1} x_{i2} x_{i3} + \varepsilon_i ,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1*x2*x3'
```

Intercept, Three Predictors, and Selected Interaction Effects

For a linear regression model with an intercept, three fixed-effects predictors, and interaction effects between two of the predictors, such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_1 x_{i2} + \varepsilon_i ,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1*x2 + x3'
```

or

```
'y ~ x1 + x2 + x3 + x1:x2'
```

Intercept, Three Predictors, and Lower-Order Interaction Effects Only

For a linear regression model with an intercept, three fixed-effects predictors, and pairwise interaction effects between all three predictors, but excluding an interaction effect between all three predictors simultaneously, such as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_1 x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \varepsilon_i ,$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1*x2*x3 - x1:x2:x3'
```

Linear Mixed-Effects Model Examples

Use `fitlme` and `fitlmematrix` to fit linear mixed-effects models.

Random Effect Intercept, No Predictors

For a linear mixed-effects model that contains a random intercept but no predictor terms, such as

$$y_{im} = \beta_{0m} ,$$

where

$$\beta_{0m} = \beta_{00} + b_{0m} ,\ b_{0m} \sim N\left(0, \sigma_0^2\right)$$

and g is the grouping variable with m levels, specify the model formula using Wilkinson notation as follows:

```
'y ~ (1 | g)'
```

Random Intercept and Fixed Slope for One Predictor

For a linear mixed-effects model that contains a fixed intercept, random intercept, and fixed slope for the continuous predictor variable, such as

$$y_{im} = \beta_{0m} + \beta_1 x_{im},$$

where

$$\beta_{0m} = \beta_{00} + b_{0m},\ b_{0m} \sim N\left(0, \sigma_0^2\right)$$

and g is the grouping variable with m levels, specify the model formula using Wilkinson notation as follows:

```
'y ~ x1 + (1 | g)'
```

Random Intercept and Random Slope for One Predictor

For a linear mixed-effects model that contains a fixed intercept, plus a random intercept and a random slope that have a possible correlation between them, such as

$$y_{im} = \beta_{0m} + \beta_{1m} x_{im},$$

where

$$\beta_{0m} = \beta_{00} + b_{0m}$$

$$\beta_{1m} = \beta_{10} + b_{1m}$$

$$\begin{bmatrix} b_{0m} \\ b_{1m} \end{bmatrix} \sim N\left\{0, \sigma^2 D(\theta)\right\}$$

and D is a 2-by-2 symmetric and positive semidefinite covariance matrix, parameterized by a variance component vector θ, specify the model formula using Wilkinson notation as follows:

```
'y ~ x1 + (x1 | g)'
```

The pattern of the random effects covariance matrix is determined by the model fitting function. To specify the covariance matrix pattern, use the name-value pairs available through `fitlme` when fitting the model. For example, you can specify the assumption that the random intercept and random slope are independent of one another using the `'CovariancePattern'` name-value pair argument in `fitlme`.

Generalized Linear Model Examples

Use `fitglm` and `stepwiseglm` to fit generalized linear models.

In a generalized linear model, the y response variable has a distribution other than normal, but you can represent the model as an equation that is linear in the regression coefficients. Specifying a generalized linear model requires three parts:

- Distribution of the response variable
- Link function
- Linear predictor

The distribution of the response variable and the link function are specified using name-value pair arguments in the fit function `fitglm` or `stepwiseglm`.

The linear predictor portion of the equation, which appears on the right side of the ~ symbol in the model specification formula, uses Wilkinson notation in the same way as for the linear model examples.

A generalized linear model models the link function, rather than the actual response, as y. This is reflected in the output display for the model object.

Intercept and Two Predictors

For a generalized linear regression model with an intercept and two predictors, such as

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

specify the model formula using Wilkinson notation as follows:

```
'y ~ x1 + x2'
```

Generalized Linear Mixed-Effects Model Examples

Use `fitglme` to fit generalized linear mixed-effects models.

In a generalized linear mixed-effects model, the y response variable has a distribution other than normal, but you can represent the model as an equation that is linear in the regression coefficients. Specifying a generalized linear model requires three parts:

- Distribution of the response variable
- Link function
- Linear predictor

The distribution of the response variable and the link function are specified using name-value pair arguments in the fit function `fitglme`.

The linear predictor portion of the equation, which appears on the right side of the ~ symbol in the model specification formula, uses Wilkinson notation in the same way as for the linear mixed-effects model examples.

A generalized linear model models the link function as y, not the response itself. This is reflected in the output display for the model object.

The pattern of the random effects covariance matrix is determined by the model fitting function. To specify the covariance matrix pattern, use the name-value pairs available through `fitglme` when fitting the model. For example, you can specify the assumption that the random intercept and random slope are independent of one another using the `'CovariancePattern'` name-value pair argument in `fitglme`.

Random Intercept and Fixed Slope for One Predictor

For a generalized linear mixed-effects model that contains a fixed intercept, random intercept, and fixed slope for the continuous predictor variable, where the response can be modeled using a Poisson distribution, such as

$$\log(y_{im}) = \beta_0 + \beta_1 x_{im} + b_i ,$$

where

$$b_i \sim N\left(0, \sigma_b^2\right)$$

and g is the grouping variable with m levels, specify the model formula using Wilkinson notation as follows:

```
'y ~ x1 + (1 | g)'
```

Repeated Measures Model Examples

Use `fitrm` to fit repeated measures models.

One Predictor

For a repeated measures model with five response measurements and one predictor variable, specify the model formula using Wilkinson notation as follows:

```
'y1-y5 ~ x1'
```

Three Predictors and an Interaction Term

For a repeated measures model with five response measurements and three predictor variables, plus an interaction between two of the predictor variables, specify the model formula using Wilkinson notation as follows:

```
'y1-y5 ~ x1*x2 + x3'
```

References

[1] Wilkinson, G. N., and C. E. Rogers. "Symbolic description of factorial models for analysis of variance." *J. Royal Statistics Society* 22, pp. 392–399, 1973.

Stepwise Regression

In this section...

Stepwise Regression to Select Appropriate Models

`stepwiselm` creates a linear model and automatically adds to or trims the model. To create a small model, start from a constant model. To create a large model, start with a model containing many terms. A large model usually has lower error as measured by the fit to the original data, but might not have any advantage in predicting new data.

`stepwiselm` can use all the name-value options from `fitlm`, with additional options relating to the starting and bounding models. In particular:

- For a small model, start with the default lower bounding model: `'constant'` (a model that has no predictor terms).
- The default upper bounding model has linear terms and interaction terms (products of pairs of predictors). For an upper bounding model that also includes squared terms, set the `Upper` name-value pair to `'quadratic'`.

Compare large and small stepwise models

This example shows how to compare models that `stepwiselm` returns starting from a constant model and starting from a full interaction model.

Load the `carbig` data and create a table from some of the data.

```
load carbig
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Create a mileage model stepwise starting from the constant model.

```
mdl1 = stepwiselm(tbl,'constant','ResponseVar','MPG')
```

```
1. Adding Weight, FStat = 888.8507, pValue = 2.9728e-103
2. Adding Horsepower, FStat = 3.8217, pValue = 0.00049608
3. Adding Horsepower:Weight, FStat = 64.8709, pValue = 9.93362e-15

mdl1 =
Linear regression model:
    MPG ~ 1 + Horsepower*Weight

Estimated Coefficients:
                        Estimate         SE         tStat       pValue
                       __________    __________    _______    __________

    (Intercept)           63.558        2.3429      27.127    1.2343e-91
    Horsepower           -0.25084      0.027279    -9.1952    2.3226e-18
    Weight               -0.010772    0.00077381    -13.921    5.1372e-36
    Horsepower:Weight    5.3554e-05    6.6491e-06     8.0542    9.9336e-15
```

```
Number of observations: 392, Error degrees of freedom: 388
Root Mean Squared Error: 3.93
R-squared: 0.748,  Adjusted R-Squared: 0.746
F-statistic vs. constant model: 385, p-value = 7.26e-116
```

Create a mileage model stepwise starting from the full interaction model.

```
mdl2 = stepwiselm(tbl,'interactions','ResponseVar','MPG')
```

```
1. Removing Acceleration:Displacement, FStat = 0.024186, pValue = 0.8765
2. Removing Displacement:Weight, FStat = 0.33103, pValue = 0.56539
3. Removing Acceleration:Horsepower, FStat = 1.7334, pValue = 0.18876
4. Removing Acceleration:Weight, FStat = 0.93269, pValue = 0.33477
5. Removing Horsepower:Weight, FStat = 0.64486, pValue = 0.42245

mdl2 =
Linear regression model:
    MPG ~ 1 + Acceleration + Weight + Displacement*Horsepower

Estimated Coefficients:
                              Estimate         SE         tStat       pValue
                             __________    __________    _______    __________

    (Intercept)                 61.285        2.8052     21.847    1.8593e-69
    Acceleration               -0.34401       0.11862      -2.9     0.0039445
    Displacement               -0.081198      0.010071    -8.0623    9.5014e-15
    Horsepower                 -0.24313       0.026068    -9.3265    8.6556e-19
    Weight                     -0.0014367     0.00084041  -1.7095      0.088166
    Displacement:Horsepower    0.00054236     5.7987e-05   9.3531    7.0527e-19


Number of observations: 392, Error degrees of freedom: 386
Root Mean Squared Error: 3.84
R-squared: 0.761,  Adjusted R-Squared: 0.758
F-statistic vs. constant model: 246, p-value = 1.32e-117
```

Notice that:

- `mdl1` has four coefficients (the `Estimate` column), and `mdl2` has six coefficients.
- The adjusted R-squared of `mdl1` is `0.746`, which is slightly less (worse) than that of `mdl2`, `0.758`.

Create a mileage model stepwise with a full quadratic model as the upper bound, starting from the full quadratic model:

```
mdl3 = stepwiselm(tbl,'quadratic','ResponseVar','MPG','Upper','quadratic');
```

```
1. Removing Acceleration:Horsepower, FStat = 0.075209, pValue = 0.78405
2. Removing Acceleration:Weight, FStat = 0.072756, pValue = 0.78751
3. Removing Horsepower:Weight, FStat = 0.12569, pValue = 0.72314
4. Removing Weight^2, FStat = 1.194, pValue = 0.27521
5. Removing Displacement:Weight, FStat = 1.2839, pValue = 0.25789
6. Removing Displacement^2, FStat = 2.069, pValue = 0.15114
7. Removing Horsepower^2, FStat = 0.74063, pValue = 0.39
```

Compare the three model complexities by examining their formulas.

```
mdl1.Formula
```

```
ans =
MPG ~ 1 + Horsepower*Weight
```

```
mdl2.Formula
```

```
ans =
MPG ~ 1 + Acceleration + Weight + Displacement*Horsepower
```

```
mdl3.Formula
```

```
ans =
MPG ~ 1 + Weight + Acceleration*Displacement + Displacement*Horsepower
        + Acceleration^2
```

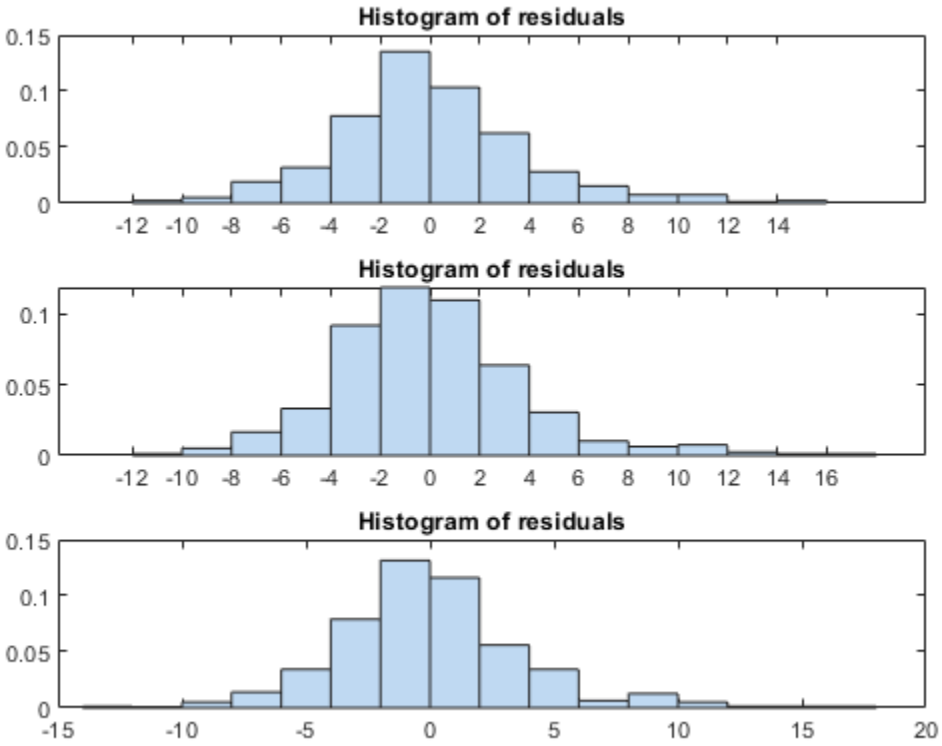
The adjusted R^2 values improve slightly as the models become more complex:

```
RSquared = [mdl1.Rsquared.Adjusted, ...
    mdl2.Rsquared.Adjusted, mdl3.Rsquared.Adjusted]
```

```
RSquared = 1×3

    0.7465    0.7580    0.7599
```

Compare residual plots of the three models.

```
subplot(3,1,1)
plotResiduals(mdl1)
subplot(3,1,2)
plotResiduals(mdl2)
subplot(3,1,3)
plotResiduals(mdl3)
```



The models have similar residuals. It is not clear which fits the data better.

Interestingly, the more complex models have larger maximum deviations of the residuals:

```
Rrange1 = [min(mdl1.Residuals.Raw),max(mdl1.Residuals.Raw)];
Rrange2 = [min(mdl2.Residuals.Raw),max(mdl2.Residuals.Raw)];
Rrange3 = [min(mdl3.Residuals.Raw),max(mdl3.Residuals.Raw)];
Rranges = [Rrange1;Rrange2;Rrange3]
```

```
Rranges = 3×2

  -10.7725   14.7314
  -11.4407   16.7562
  -12.2723   16.7927
```

See Also

`LinearModel` | `fitlm` | `plotResiduals` | `stepwiselm`

More About

- “Linear Regression” on page 11-9
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161

- “Interpret Linear Regression Results” on page 11-50
- “Introduction to Feature Selection” on page 15-49
- “Sequential Feature Selection” on page 15-61

Reduce Outlier Effects Using Robust Regression

In this section...

You can reduce outlier effects in linear regression models by using robust linear regression. This topic defines robust regression, shows how to use it to fit a linear model, and compares the results to a standard fit. You can use `fitlm` with the `'RobustOpts'` name-value pair argument to fit a robust regression model. Or you can use `robustfit` to simply compute the robust regression coefficient parameters.

Why Use Robust Regression?

Robust linear regression is less sensitive to outliers than standard linear regression. Standard linear regression uses ordinary least-squares fitting to compute the model parameters that relate the response data to the predictor data with one or more coefficients. (See “Estimation of Multivariate Regression Models” on page 15-5 for more details.) As a result, outliers have a large influence on the fit, because squaring the residuals magnifies the effects of these extreme data points. Models that use standard linear regression, described in “What Is a Linear Regression Model?” on page 11-6, are based on certain assumptions, such as a normal distribution of errors in the observed responses. If the distribution of errors is asymmetric or prone to outliers, model assumptions are invalidated, and parameter estimates, confidence intervals, and other computed statistics become unreliable.

Robust regression uses a method called iteratively reweighted least squares to assign a weight to each data point. This method is less sensitive to large changes in small parts of the data. As a result, robust linear regression is less sensitive to outliers than standard linear regression.

Iteratively Reweighted Least Squares

In weighted least squares, the fitting process includes the weight as an additional scale factor, which improves the fit. The weights determine how much each response value influences the final parameter estimates. A low-quality data point (for example, an outlier) should have less influence on the fit. To compute the weights w_i, you can use predefined weight functions, such as Tukey's bisquare function (see the name-value pair argument `'RobustOpts'` in `fitlm` for more options).

The *iteratively reweighted least-squares* algorithm automatically and iteratively calculates the weights. At initialization, the algorithm assigns equal weight to each data point, and estimates the model coefficients using ordinary least squares. At each iteration, the algorithm computes the weights w_i, giving lower weight to points farther from model predictions in the previous iteration. The algorithm then computes model coefficients b using weighted least squares. Iteration stops when the values of the coefficient estimates converge within a specified tolerance. This algorithm simultaneously seeks to find the curve that fits the bulk of the data using the least-squares approach, and to minimize the effects of outliers.

For more details, see “Steps for Iteratively Reweighted Least Squares” on page 11-107.

Compare Results of Standard and Robust Least-Squares Fit

This example shows how to use robust regression with the `fitlm` function, and compares the results of a robust fit to a standard least-squares fit.

Load the `moore` data. The predictor data is in the first five columns, and the response data is in the sixth.

```
load moore
X = moore(:,1:5);
y = moore(:,6);
```

Fit the least-squares linear model to the data.

```
mdl = fitlm(X,y)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                    Estimate          SE          tStat       pValue
                   ___________    __________    _________    ________

    (Intercept)        -2.1561       0.91349      -2.3603      0.0333
    x1             -9.0116e-06    0.00051835    -0.017385     0.98637
    x2                0.0013159     0.0012635       1.0415     0.31531
    x3                0.0001278    7.6902e-05       1.6618     0.11876
    x4                0.0078989         0.014      0.56421     0.58154
    x5               0.00014165    7.3749e-05       1.9208    0.075365


Number of observations: 20, Error degrees of freedom: 14
Root Mean Squared Error: 0.262
R-squared: 0.811,  Adjusted R-Squared: 0.743
F-statistic vs. constant model: 12, p-value = 0.000118
```

Fit the robust linear model to the data by using the `'RobustOps'` name-value pair argument.

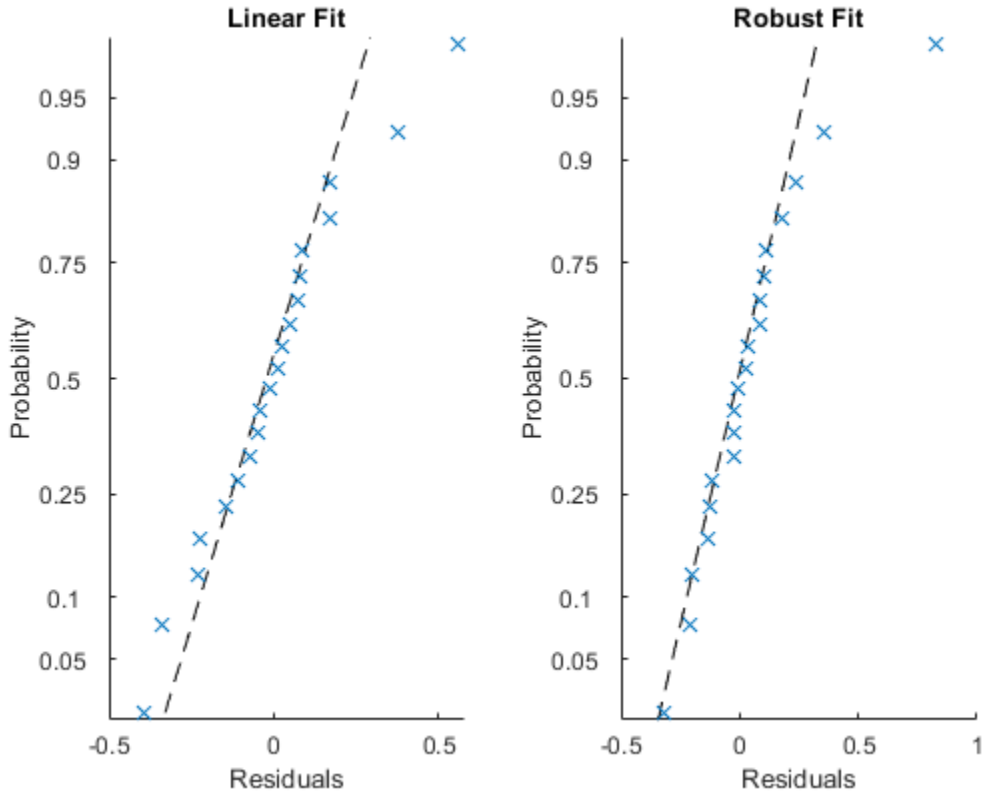
```
mdlr = fitlm(X,y,'RobustOpts','on')
```

```
mdlr =
Linear regression model (robust fit):
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                    Estimate          SE          tStat       pValue
                   ___________    __________    _________    ________

    (Intercept)        -1.7516       0.86953      -2.0144     0.063595
    x1              1.7006e-05    0.00049341     0.034467      0.97299
    x2              0.00088843     0.0012027       0.7387      0.47229
    x3              0.00015729    7.3202e-05       2.1487     0.049639
    x4               0.0060468      0.013326      0.45375      0.65696
    x5              6.8807e-05    7.0201e-05      0.98015      0.34365


Number of observations: 20, Error degrees of freedom: 14
```

```
Root Mean Squared Error: 0.249
R-squared: 0.775,  Adjusted R-Squared: 0.694
F-statistic vs. constant model: 9.64, p-value = 0.000376
```

Visually examine the residuals of the two models.

```
tiledlayout(1,2)
nexttile
plotResiduals(mdl,'probability')
title('Linear Fit')
nexttile
plotResiduals(mdlr,'probability')
title('Robust Fit')
```



The residuals from the robust fit (right half of the plot) are closer to the straight line, except for the one obvious outlier.

Find the index of the outlier.

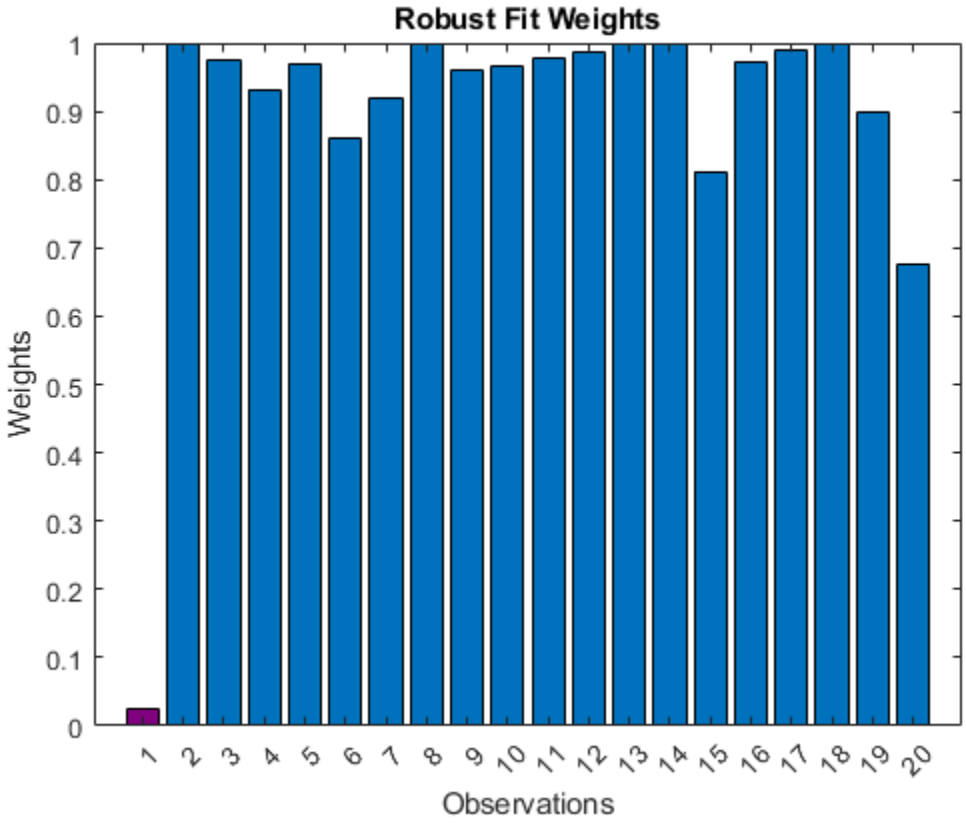
```
outlier = find(isoutlier(mdlr.Residuals.Raw))
```

```
outlier = 1
```

Plot the weights of the observations in the robust fit.

```
figure
b = bar(mdlr.Robust.Weights);
b.FaceColor = 'flat';
```

```
b.CData(outlier,:) = [.5 0 .5];
xticks(1:length(mdlr.Residuals.Raw))
xlabel('Observations')
ylabel('Weights')
title('Robust Fit Weights')
```



The weight of the outlier in the robust fit (purple bar) is much less than the weights of the other observations.

Steps for Iteratively Reweighted Least Squares

The iteratively reweighted least-squares algorithm follows this procedure:

1. Start with an initial estimate of the weights and fit the model by weighted least squares.
2. Compute the adjusted residuals. The adjusted residuals are given by

   $$r_{\text{adj}} = \frac{r_i}{\sqrt{1 - h_i}}$$

   where r_i are the ordinary least-squares residuals, and h_i are the least-squares fit leverage values. Leverages adjust the residuals by reducing the weight of high-leverage data points, which have a large effect on the least-squares fit (see “Hat Matrix and Leverage” on page 11-77).
3. Standardize the residuals. The standardized adjusted residuals are given by

$$u = \frac{r_{\text{adj}}}{Ks} = \frac{r_i}{Ks\sqrt{1-h_i}}$$

where K is a tuning constant, and s is an estimate of the standard deviation of the error term given by `s = MAD/0.6745`.

`MAD` is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution. If the predictor data matrix `X` has p columns, the software excludes the smallest p absolute deviations when computing the median.

4 Compute the robust weights w_i as a function of u. For example, the bisquare weights are given by

$$w_i = \begin{cases} (1-u_i 2)^2 & , |u_i| < 1 \\ 0 & , |u_i| \geq 1 \end{cases}$$

5 Estimate the robust regression coefficients b. The weights modify the expression for the parameter estimates b as follows

$$b = \widehat{\beta} = (X^T WT)^{-1} X^T W y$$

where W is the diagonal weight matrix, X is the predictor data matrix, and y is the response vector.

6 Estimate the weighted least-squares error

$$e = \sum_1^n w_i (y_i - \widehat{y}_i)^2 = \sum_1^n w_i r_i^2$$

where w_i are the weights, y_i are the observed responses, $\hat{y}_i$ are the fitted responses, and r_i are the residuals.

7 Iteration stops if the fit converges or the maximum number of iterations is reached. Otherwise, perform the next iteration of the least-squares fitting by returning to the second step.

See Also

`LinearModel` | `fitlm` | `plotResiduals` | `robustfit`

More About

- “Linear Regression” on page 11-9
- “Linear Regression Workflow” on page 11-35
- “Train Linear Regression Model” on page 11-161
- “Interpret Linear Regression Results” on page 11-50

Ridge Regression

In this section...

Introduction to Ridge Regression

Coefficient estimates for the models described in “Linear Regression” on page 11-9 rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the matrix $(X^TX)^{-1}$ becomes close to singular. As a result, the least-squares estimate

$$\widehat{\beta} = (X^TX)^{-1}X^Ty$$

becomes highly sensitive to random errors in the observed response y, producing a large variance. This situation of multicollinearity can arise, for example, when data are collected without an experimental design.

Ridge regression addresses the problem by estimating regression coefficients using

$$\widehat{\beta} = (X^TX + kI)^{-1}X^Ty$$

where k is the ridge parameter and I is the identity matrix. Small positive values of k improve the conditioning of the problem and reduce the variance of the estimates. While biased, the reduced variance of ridge estimates often result in a smaller mean square error when compared to least-squares estimates.

The Statistics and Machine Learning Toolbox function `ridge` carries out ridge regression.

Ridge Regression

This example shows how to perform ridge regression.

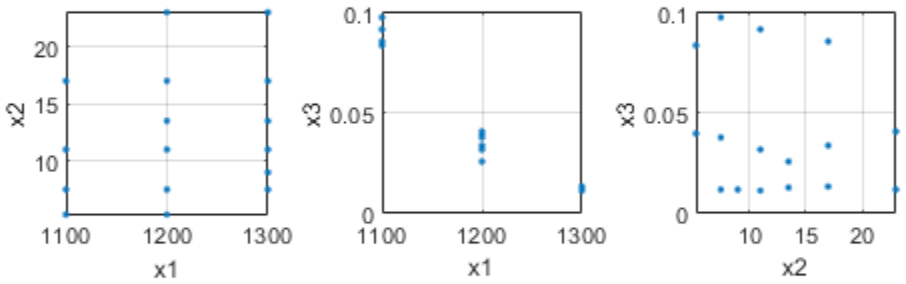
Load the data in `acetylene.mat`, with observations of the predictor variables `x1`, `x2`, `x3`, and the response variable `y`.

```
load acetylene
```

Plot the predictor variables against each other.

```
subplot(1,3,1)
plot(x1,x2,'.')
xlabel('x1')
ylabel('x2')
grid on
axis square

subplot(1,3,2)
plot(x1,x3,'.')
xlabel('x1')
ylabel('x3')
```

```
grid on
axis square

subplot(1,3,3)
plot(x2,x3,'.')
xlabel('x2')
ylabel('x3')
grid on
axis square
```



Note the correlation between `x1` and the other two predictor variables.

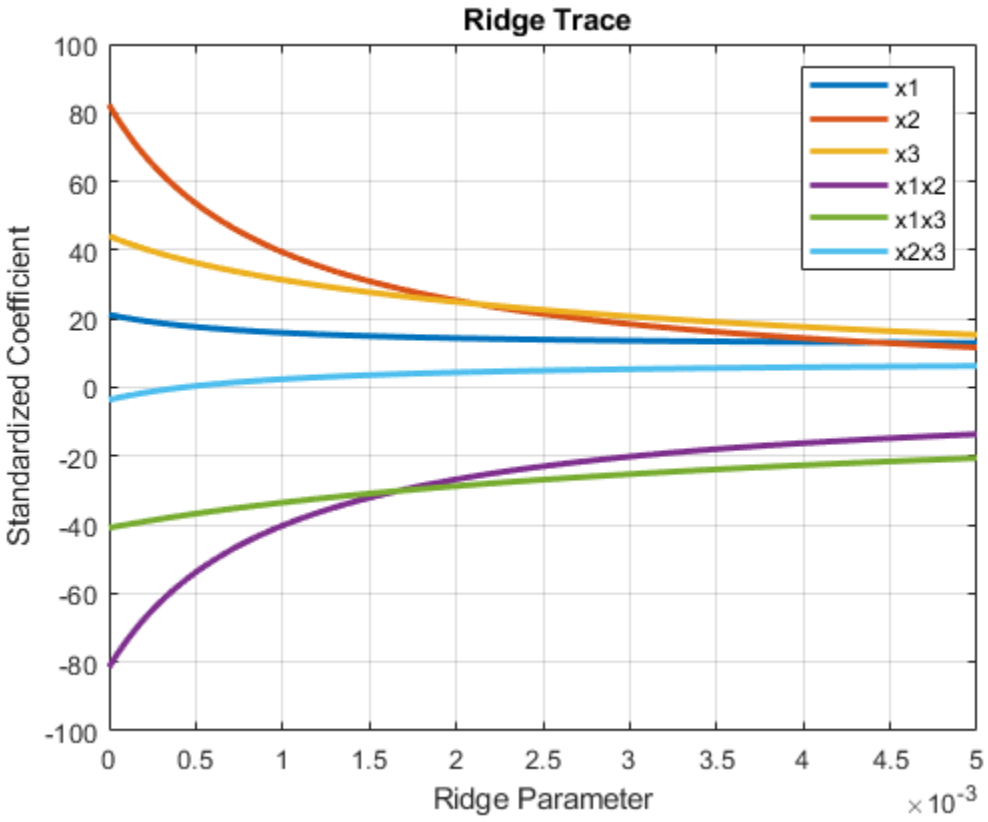
Use `ridge` and `x2fx` to compute coefficient estimates for a multilinear model with interaction terms, for a range of ridge parameters.

```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
k = 0:1e-5:5e-3;
betahat = ridge(y,D,k);
```

Plot the ridge trace.

```
figure
plot(k,betahat,'LineWidth',2)
ylim([-100 100])
grid on
xlabel('Ridge Parameter')
```

```
ylabel('Standardized Coefficient')
title('{\bf Ridge Trace}')
legend('x1','x2','x3','x1x2','x1x3','x2x3')
```



The estimates stabilize to the right of the plot. Note that the coefficient of the `x2x3` interaction term changes sign at a value of the ridge parameter $\approx 5 \times 10^{-4}$.

See Also

`fitrlinear` | `lasso` | `lassoPlot` | `lassoglm` | `ridge`

More About

- “Lasso Regularization” on page 11-120
- “Lasso and Elastic Net with Cross Validation” on page 11-123
- “Wide Data via Lasso and Parallel Computing” on page 11-115
- “Lasso and Elastic Net” on page 11-112

Lasso and Elastic Net

In this section...

What Are Lasso and Elastic Net?

Lasso is a regularization technique. Use `lasso` to:

- Reduce the number of predictors in a regression model.
- Identify important predictors.
- Select among redundant predictors.
- Produce shrinkage estimates with potentially lower predictive errors than ordinary least squares.

Elastic net is a related technique. Use elastic net when you have several highly correlated variables. `lasso` provides elastic net regularization when you set the `Alpha` name-value pair to a number strictly between `0` and `1`.

See “Lasso and Elastic Net Details” on page 11-112.

For lasso regularization of regression ensembles, see `regularize`.

Lasso and Elastic Net Details

Overview of Lasso and Elastic Net

Lasso is a regularization technique for performing linear regression. Lasso includes a penalty term that constrains the size of the estimated coefficients. Therefore, it resembles ridge regression on page 11-109. Lasso is a shrinkage estimator: it generates coefficient estimates that are biased to be small. Nevertheless, a lasso estimator can have smaller mean squared error than an ordinary least-squares estimator when you apply it to new data.

Unlike ridge regression, as the penalty term increases, lasso sets more coefficients to zero. This means that the lasso estimator is a smaller model, with fewer predictors. As such, lasso is an alternative to stepwise regression on page 11-99 and other model selection and dimensionality reduction techniques.

Elastic net is a related technique. Elastic net is a hybrid of ridge regression and lasso regularization. Like lasso, elastic net can generate reduced models by generating zero-valued coefficients. Empirical studies have suggested that the elastic net technique can outperform lasso on data with highly correlated predictors.

Definition of Lasso

The lasso technique solves this regularization problem. For a given value of λ, a nonnegative parameter, `lasso` solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right).$$

- N is the number of observations.
- y_i is the response at observation i.
- x_i is data, a vector of p values at observation i.
- λ is a positive regularization parameter corresponding to one value of `Lambda`.
- The parameters β_0 and β are scalar and p-vector respectively.

As λ increases, the number of nonzero components of β decreases.

The lasso problem involves the L^1 norm of β, as contrasted with the elastic net algorithm.

Definition of Elastic Net

The elastic net technique solves this regularization problem. For an α strictly between 0 and 1, and a nonnegative λ, elastic net solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda P_\alpha(\beta)\right),$$

where

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha\left|\beta_j\right|\right).$$

Elastic net is the same as lasso when $\alpha = 1$. As α shrinks toward 0, elastic net approaches `ridge` regression. For other values of α, the penalty term $P_\alpha(\beta)$ interpolates between the L^1 norm of β and the squared L^2 norm of β.

References

[1] Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society*, Series B, Vol 58, No. 1, pp. 267–288, 1996.

[2] Zou, H. and T. Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society, Series B,* Vol. 67, No. 2, pp. 301–320, 2005.

[3] Friedman, J., R. Tibshirani, and T. Hastie. "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, Vol 33, No. 1, 2010. `https://www.jstatsoft.org/v33/i01`

[4] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, 2nd edition. Springer, New York, 2008.

See Also

`fitrlinear` | `lasso` | `lassoPlot` | `lassoglm` | `ridge`

More About

- “Lasso Regularization” on page 11-120
- “Lasso and Elastic Net with Cross Validation” on page 11-123

- “Wide Data via Lasso and Parallel Computing” on page 11-115
- “Ridge Regression” on page 11-109
- “Introduction to Feature Selection” on page 15-49

Wide Data via Lasso and Parallel Computing

This example shows how to use `lasso` along with cross validation to identify important predictors.

Load the sample data and display the description.

```
load spectra
Description
```

```
Description =

  11×72 char array

    '== Spectral and octane data of gasoline ==                              '
    '                                                                        '
    'NIR spectra and octane numbers of 60 gasoline samples                   '
    '                                                                        '
    'NIR:     NIR spectra, measured in 2 nm intervals from 900 nm to 1700 nm '
    'octane:  octane numbers                                                 '
    'spectra: a dataset array containing variables for NIR and octane        '
    '                                                                        '
    'Reference:                                                              '
    'Kalivas, John H., "Two Data Sets of Near Infrared Spectra," Chemometrics'
    'and Intelligent Laboratory Systems, v.37 (1997) pp.255-259              '
```

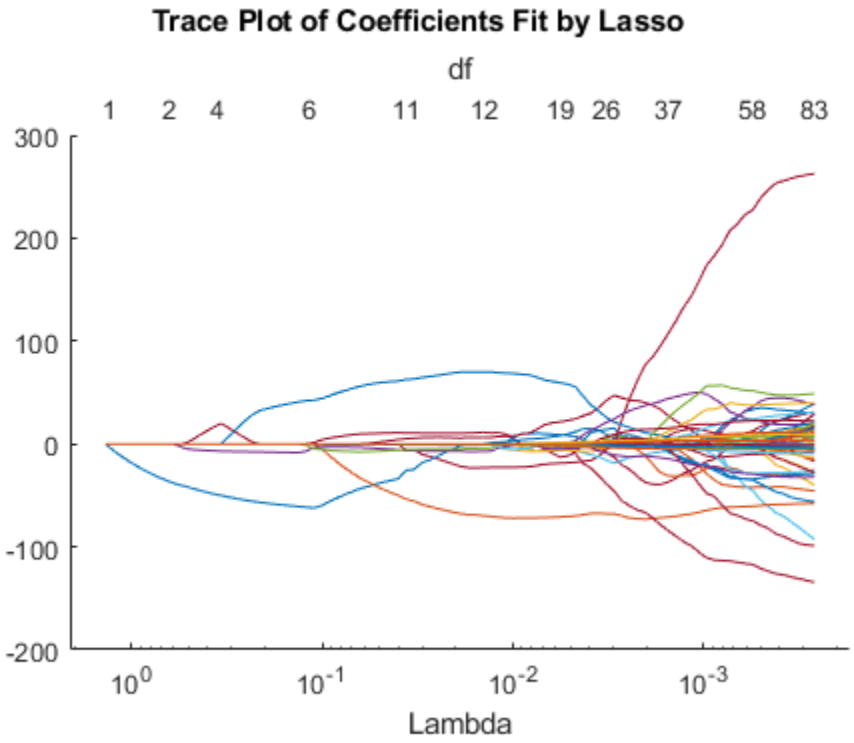
Lasso and elastic net are especially well suited for wide data, that is, data with more predictors than observations with lasso and elastic net. There are redundant predictors in this type of data. You can use `lasso` along with cross validation to identify important predictors.

Compute the default `lasso` fit.

```
[b fitinfo] = lasso(NIR,octane);
```

Plot the number of predictors in the fitted lasso regularization as a function of `Lambda` , using a logarithmic *x* -axis.

```
lassoPlot(b,fitinfo,'PlotType','Lambda','XScale','log');
```



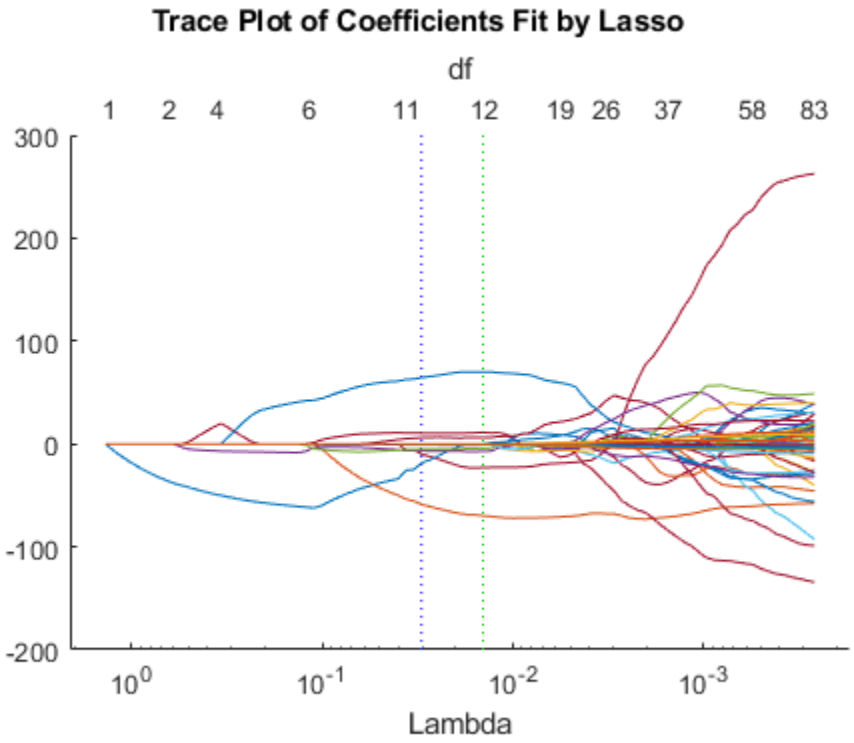
It is difficult to tell which value of `Lambda` is appropriate. To determine a good value, try fitting with cross validation.

```
tic
[b fitinfo] = lasso(NIR,octane,'CV',10);
toc
```

```
Elapsed time is 7.353767 seconds.
```

Plot the result.

```
lassoPlot(b,fitinfo,'PlotType','Lambda','XScale','log');
```



Display the suggested value of `Lambda` .

```
fitinfo.Lambda1SE
```

```
ans =

    0.0302
```

Display the `Lambda` with minimal MSE.

```
fitinfo.LambdaMinMSE
```

```
ans =

    0.0144
```

Examine the quality of the fit for the suggested value of `Lambda` .

```
lambdaindex = fitinfo.Index1SE;
mse = fitinfo.MSE(lambdaindex)
df = fitinfo.DF(lambdaindex)
```

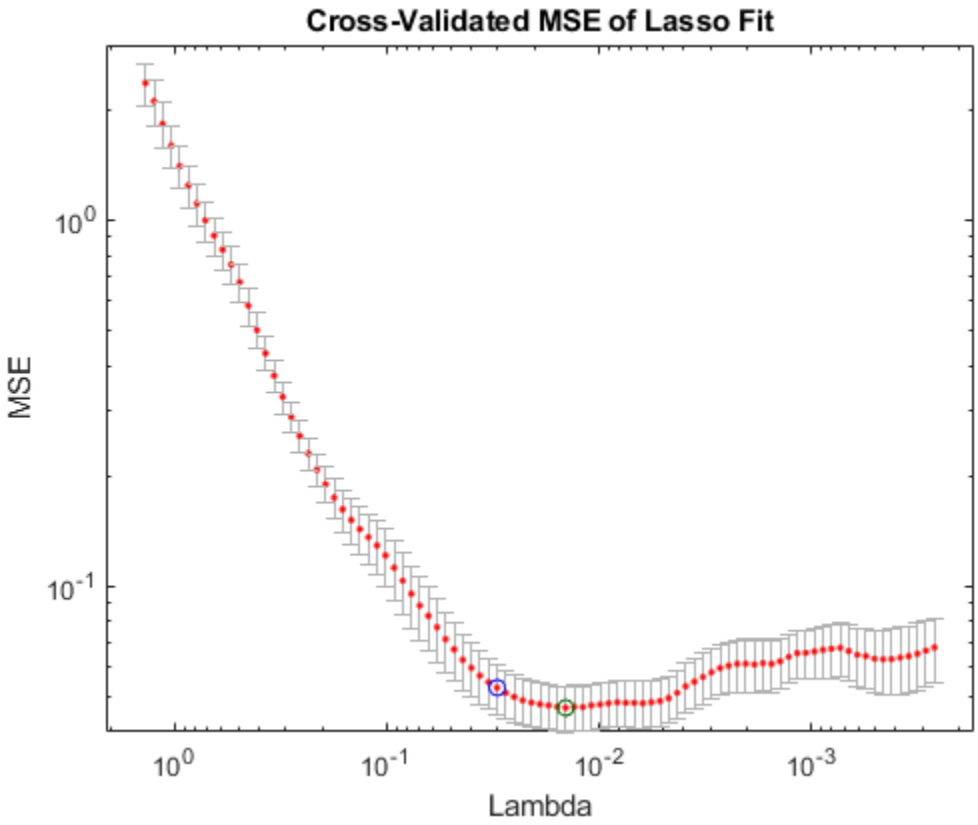
```
mse =
```

```
    0.0528

df =

    11
```

The fit uses just 11 of the 401 predictors and achieves a small cross-validated MSE.

Examine the plot of cross-validated MSE.

```
lassoPlot(b,fitinfo,'PlotType','CV');
% Use a log scale for MSE to see small MSE values better
set(gca,'YScale','log');
```



As `Lambda` increases (toward the left), MSE increases rapidly. The coefficients are reduced too much and they do not adequately fit the responses. As `Lambda` decreases, the models are larger (have more nonzero coefficients). The increasing MSE suggests that the models are overfitted.

The default set of `Lambda` values does not include values small enough to include all predictors. In this case, there does not appear to be a reason to look at smaller values. However, if you want smaller values than the default, use the `LambdaRatio` parameter, or supply a sequence of `Lambda` values using the `Lambda` parameter. For details, see the `lasso` reference page.

Cross validation can be slow. If you have a Parallel Computing Toolbox license, speed the computation of cross-validated lasso estimate using parallel computing. Start a parallel pool.

```
mypool = parpool()
```

```
Starting parallel pool (parpool) using the 'local' profile ...
connected to 6 workers.

mypool =

 Pool with properties:

            Connected: true
           NumWorkers: 6
              Cluster: local
        AttachedFiles: {}
    AutoAddClientPath: true
          IdleTimeout: 30 minutes (30 minutes remaining)
          SpmdEnabled: true
```

Set the parallel computing option and compute the lasso estimate.

```
opts = statset('UseParallel',true);
tic;
[b fitinfo] = lasso(NIR,octane,'CV',10,'Options',opts);
toc
```

```
Elapsed time is 3.799009 seconds.
```

Computing in parallel using two workers is faster on this problem.

Stop parallel pool.

```
delete(mypool)
```

```
Parallel pool using the 'local' profile is shutting down.
```

See Also

`fitrlinear` | `lasso` | `lassoPlot` | `lassoglm` | `ridge`

More About

- “Lasso Regularization” on page 11-120
- “Lasso and Elastic Net with Cross Validation” on page 11-123
- “Lasso and Elastic Net” on page 11-112
- “Ridge Regression” on page 11-109

Lasso Regularization

This example shows how `lasso` identifies and discards unnecessary predictors.

Generate 200 samples of five-dimensional artificial data X from exponential distributions with various means.

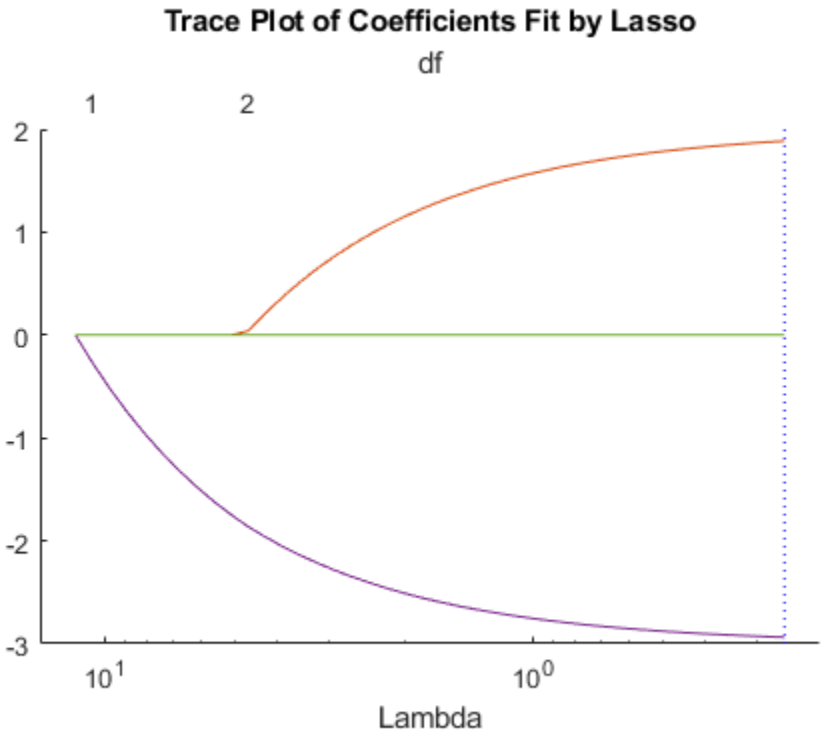
```
rng(3,'twister') % For reproducibility
X = zeros(200,5);
for ii = 1:5
    X(:,ii) = exprnd(ii,200,1);
end
```

Generate response data `Y = X*r + eps` , where `r` has just two nonzero components, and the noise `eps` is normal with standard deviation 0.1.

```
r = [0;2;0;-3;0];
Y = X*r + randn(200,1)*.1;
```

Fit a cross-validated sequence of models with `lasso` , and plot the result.

```
[b,fitinfo] = lasso(X,Y,'CV',10);
lassoPlot(b,fitinfo,'PlotType','Lambda','XScale','log');
```



The plot shows the nonzero coefficients in the regression for various values of the `Lambda` regularization parameter. Larger values of `Lambda` appear on the left side of the graph, meaning more regularization, resulting in fewer nonzero regression coefficients.

The dashed vertical lines represent the `Lambda` value with minimal mean squared error (on the right), and the `Lambda` value with minimal mean squared error plus one standard deviation. This latter value is a recommended setting for `Lambda` . These lines appear only when you perform cross validation. Cross validate by setting the `'CV'` name-value pair argument. This example uses 10-fold cross validation.

The upper part of the plot shows the degrees of freedom (df), meaning the number of nonzero coefficients in the regression, as a function of Lambda. On the left, the large value of Lambda causes all but one coefficient to be 0. On the right all five coefficients are nonzero, though the plot shows only two clearly. The other three coefficients are so small that you cannot visually distinguish them from 0.

For small values of Lambda (toward the right in the plot), the coefficient values are close to the least-squares estimate.

Find the `Lambda` value of the minimal cross-validated mean squared error plus one standard deviation. Examine the MSE and coefficients of the fit at that `Lambda` .

```
lam = fitinfo.Index1SE;
fitinfo.MSE(lam)
```

```
ans = 0.1398
```

```
b(:,lam)
```

```
ans = 5×1

         0
    1.8855
         0
   -2.9367
         0
```

`lasso` did a good job finding the coefficient vector `r` .

For comparison, find the least-squares estimate of `r` .

```
rhat = X\Y
```

```
rhat = 5×1

   -0.0038
    1.9952
    0.0014
   -2.9993
    0.0031
```

The estimate `b(:,lam)` has slightly more mean squared error than the mean squared error of `rhat` .

```
res = X*rhat - Y;     % Calculate residuals
MSEmin = res'*res/200 % b(:,lam) value is 0.1398
```

```
MSEmin = 0.0088
```

But `b(:,lam)` has only two nonzero components, and therefore can provide better predictive estimates on new data.

See Also

`fitrlinear` | `lasso` | `lassoPlot` | `lassoglm` | `ridge`

More About

- “Lasso and Elastic Net with Cross Validation” on page 11-123
- “Wide Data via Lasso and Parallel Computing” on page 11-115
- “Lasso and Elastic Net” on page 11-112
- “Ridge Regression” on page 11-109
- “Introduction to Feature Selection” on page 15-49

Lasso and Elastic Net with Cross Validation

This example shows how to predict the mileage (MPG) of a car based on its weight, displacement, horsepower, and acceleration, using the lasso and elastic net methods.

Load the `carbig` data set.

```
load carbig
```

Extract the continuous (noncategorical) predictors (lasso does not handle categorical predictors).

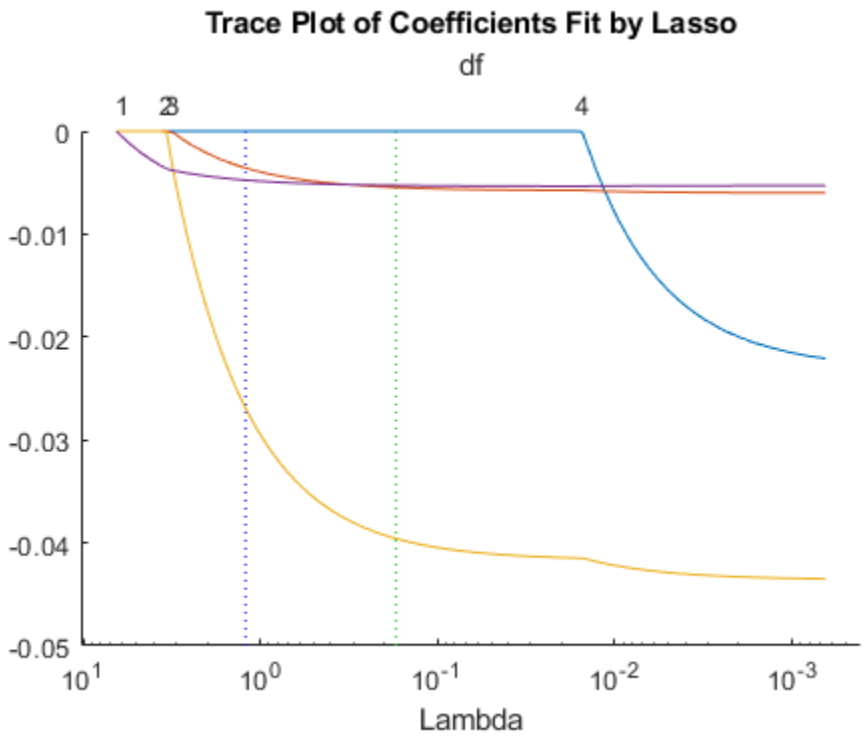
```
X = [Acceleration Displacement Horsepower Weight];
```

Perform a lasso fit with 10-fold cross validation.

```
[b,fitinfo] = lasso(X,MPG,'CV',10);
```

Plot the result.

```
lassoPlot(b,fitinfo,'PlotType','Lambda','XScale','log');
```



Calculate the correlation of the predictors. Eliminate NaNs first.

```
nonan = ~any(isnan([X MPG]),2);
Xnonan = X(nonan,:);
MPGnonan = MPG(nonan,:);
corr(Xnonan)
```

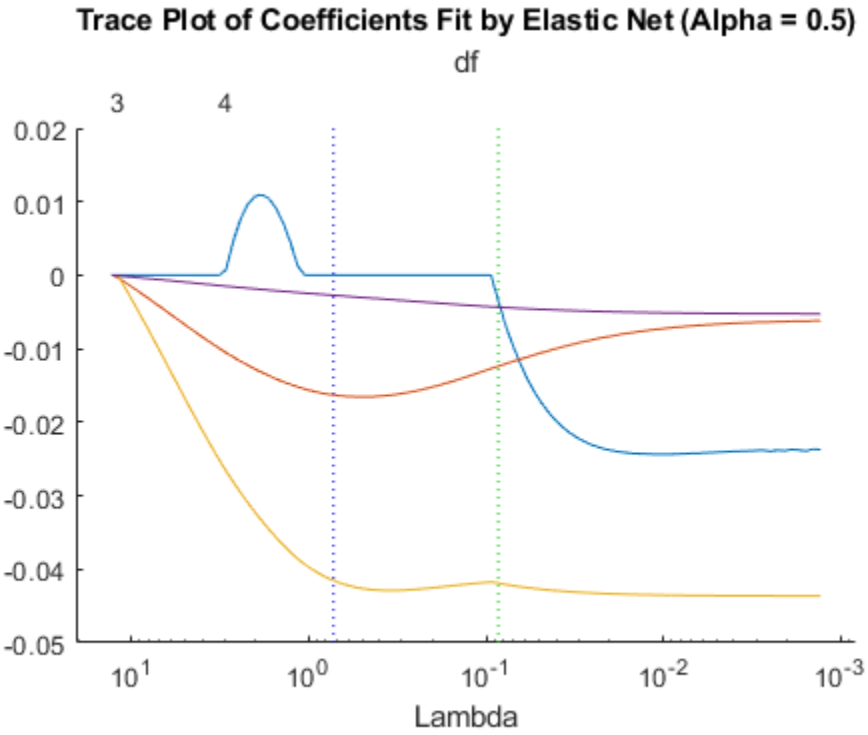
```
ans = 4×4

    1.0000   -0.5438   -0.6892   -0.4168
   -0.5438    1.0000    0.8973    0.9330
   -0.6892    0.8973    1.0000    0.8645
   -0.4168    0.9330    0.8645    1.0000
```

Because some predictors are highly correlated, perform elastic net fitting. Use `Alpha` = 0.5.

```
[ba,fitinfoa] = lasso(X,MPG,'CV',10,'Alpha',.5);
```

Plot the result. Name each predictor so you can tell which curve is which.

```
pnames = {'Acceleration','Displacement','Horsepower','Weight'};
lassoPlot(ba,fitinfoa,'PlotType','Lambda','XScale','log',...
    'PredictorNames',pnames);
```



When you activate the data cursor and click the plot, you see the name of the predictor, the coefficient, the value of `Lambda`, and the index of that point, meaning the column in `b` associated with that fit.

Here, the elastic net and lasso results are not very similar. Also, the elastic net plot reflects a notable qualitative property of the elastic net technique. The elastic net retains three nonzero coefficients as `Lambda` increases (toward the left of the plot), and these three coefficients reach 0 at about the same `Lambda` value. In contrast, the lasso plot shows two of the three coefficients becoming 0 at the same value of `Lambda`, while another coefficient remains nonzero for higher values of `Lambda`.

This behavior exemplifies a general pattern. In general, elastic net tends to retain or drop groups of highly correlated predictors as `Lambda` increases. In contrast, lasso tends to drop smaller groups, or even individual predictors.

See Also

`fitrlinear` | `lasso` | `lassoPlot` | `lassoglm` | `ridge`

More About

- “Lasso Regularization” on page 11-120
- “Wide Data via Lasso and Parallel Computing” on page 11-115
- “Lasso and Elastic Net” on page 11-112
- “Ridge Regression” on page 11-109

Partial Least Squares

In this section...

Introduction to Partial Least Squares

Partial least-squares (PLS) regression is a technique used with data that contain correlated predictor variables. This technique constructs new predictor variables, known as components, as linear combinations of the original predictor variables. PLS constructs these components while considering the observed response values, leading to a parsimonious model with reliable predictive power.

The technique is something of a cross between multiple linear regression on page 11-9 and principal component analysis on page 15-68:

- Multiple linear regression finds a combination of the predictors that best fit a response.
- Principal component analysis finds combinations of the predictors with large variance, reducing correlations. The technique makes no use of response values.
- PLS finds combinations of the predictors that have a large covariance with the response values.

PLS therefore combines information about the variances of both the predictors and the responses, while also considering the correlations among them.

PLS shares characteristics with other regression and feature transformation techniques. It is similar to ridge regression on page 11-109 in that it is used in situations with correlated predictors. It is similar to stepwise regression on page 11-99 (or more general feature selection on page 15-49 techniques) in that it can be used to select a smaller set of model terms. PLS differs from these methods, however, by transforming the original predictor space into the new component space.

The function `plsregress` carries out PLS regression.

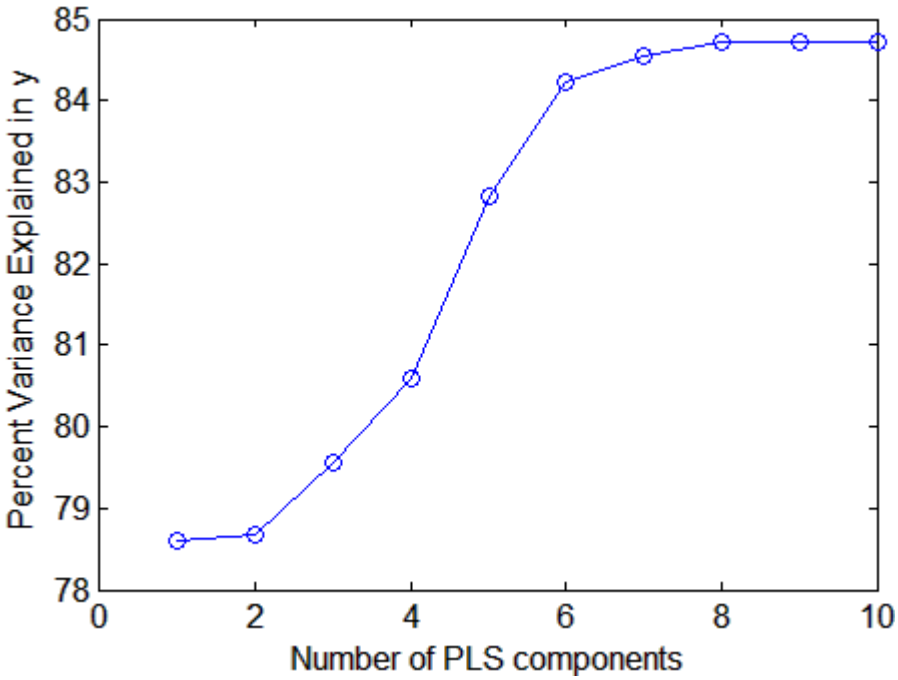
Partial Least Squares

For example, consider the data on biochemical oxygen demand in `moore.mat`, padded with noisy versions of the predictors to introduce correlations:

```
load moore
y = moore(:,6);              % Response
X0 = moore(:,1:5);           % Original predictors
X1 = X0+10*randn(size(X0));  % Correlated predictors
X = [X0,X1];
```

Use `plsregress` to perform PLS regression with the same number of components as predictors, then plot the percentage variance explained in the response as a function of the number of components:

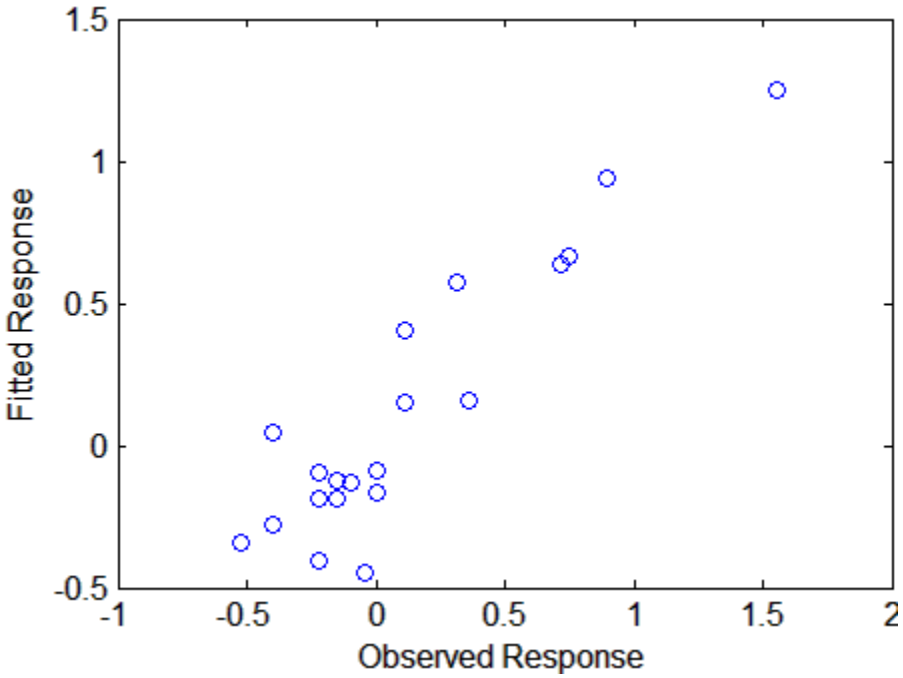
```
[XL,yl,XS,YS,beta,PCTVAR] = plsregress(X,y,10);

plot(1:10,cumsum(100*PCTVAR(2,:)),'-bo');
xlabel('Number of PLS components');
ylabel('Percent Variance Explained in y');
```



Choosing the number of components in a PLS model is a critical step. The plot gives a rough indication, showing nearly 80% of the variance in y explained by the first component, with as many as five additional components making significant contributions.

The following computes the six-component model:

```
[XL,yl,XS,YS,beta,PCTVAR,MSE,stats] = plsregress(X,y,6);
yfit = [ones(size(X,1),1) X]*beta;

plot(y,yfit,'o')
```

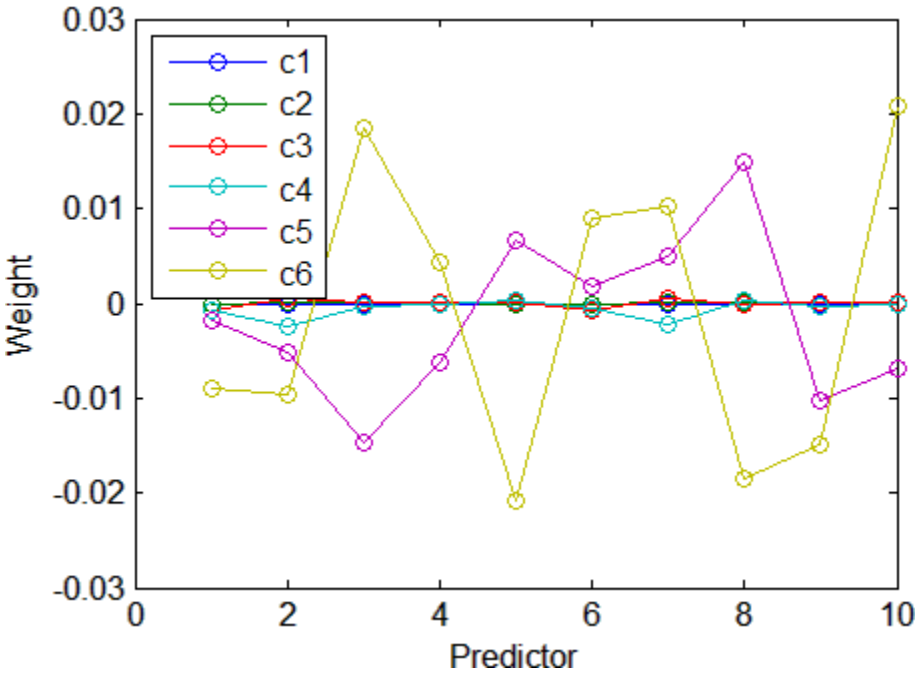


The scatter shows a reasonable correlation between fitted and observed responses, and this is confirmed by the R^2 statistic:

```
TSS = sum((y-mean(y)).^2);
RSS = sum((y-yfit).^2);
Rsquared = 1 - RSS/TSS
Rsquared =
    0.8421
```

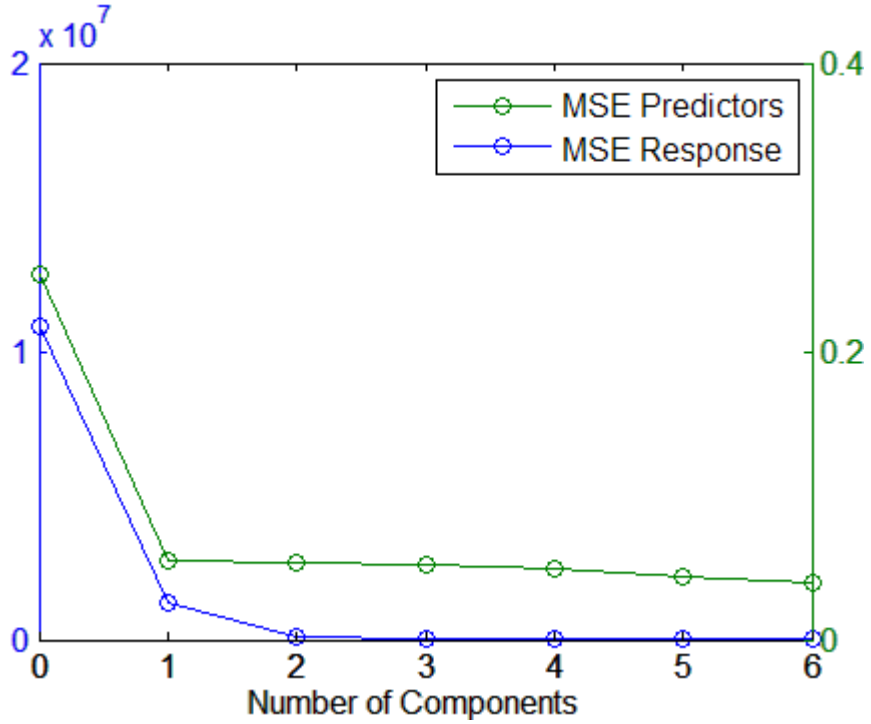
A plot of the weights of the ten predictors in each of the six components shows that two of the components (the last two computed) explain the majority of the variance in X:

```
plot(1:10,stats.W,'o-');
legend({'c1','c2','c3','c4','c5','c6'},'Location','NW')
xlabel('Predictor');
ylabel('Weight');
```



A plot of the mean-squared errors suggests that as few as two components may provide an adequate model:

```
[axes,h1,h2] = plotyy(0:6,MSE(1,:),0:6,MSE(2,:));
set(h1,'Marker','o')
set(h2,'Marker','o')
legend('MSE Predictors','MSE Response')
xlabel('Number of Components')
```



The calculation of mean-squared errors by `plsregress` is controlled by optional parameter name/value pairs specifying cross-validation type and the number of Monte Carlo repetitions.

Linear Mixed-Effects Models

Linear mixed-effects models are extensions of linear regression models for data that are collected and summarized in groups. These models describe the relationship between a response variable and independent variables, with coefficients that can vary with respect to one or more grouping variables. A mixed-effects model consists of two parts, fixed effects and random effects. Fixed-effects terms are usually the conventional linear regression part, and the random effects are associated with individual experimental units drawn at random from a population. The random effects have prior distributions whereas fixed effects do not. Mixed-effects models can represent the covariance structure related to the grouping of data by associating the common random effects to observations that have the same level of a grouping variable. The standard form of a linear mixed-effects model is

$$y = \underbrace{X\beta}_{fixed} + \underbrace{Zb}_{random} + \underbrace{\varepsilon}_{error},$$

where

- y is the n-by-1 response vector, and n is the number of observations.
- X is an n-by-p fixed-effects design matrix.
- β is a p-by-1 fixed-effects vector.
- Z is an n-by-q random-effects design matrix.
- b is a q-by-1 random-effects vector.
- ε is the n-by-1 observation error vector.

The assumptions for the linear mixed-effects model are:

- Random-effects vector, b, and the error vector, ε, have the following prior distributions:

  $$b \sim N\left(0, \sigma^2 D(\theta)\right),$$
  $$\varepsilon \sim N\left(0, \sigma^2 I\right),$$

  where D is a symmetric and positive semidefinite matrix, parameterized by a variance component vector θ, I is an n-by-n identity matrix, and σ^2 is the error variance.
- Random-effects vector, b, and the error vector, ε, are independent from each other.

Mixed-effects models are also called *multilevel models* or *hierarchical models* depending on the context. Mixed-effects models is a more general term than the latter two. Mixed-effects models might include factors that are not necessarily multilevel or hierarchical, for example crossed factors. That is why mixed-effects is the terminology preferred here. Sometimes mixed-effects models are expressed as multilevel regression models (first level and grouping level models) that are fit simultaneously. For example, a varying or random intercept model, with one continuous predictor variable x and one grouping variable with M levels, can be expressed as

$$y_{im} = \beta_{0m} + \beta_1 x_{im} + \varepsilon_{im}, \quad i = 1, 2, .., n, \quad m = 1, 2, ..., M, \quad \varepsilon_{im} \sim N\left(0, \sigma^2\right),$$
$$\beta_{0m} = \beta_{00} + b_{0m}, \quad b_{0m} \sim N\left(0, \sigma_0^2\right),$$

where y_{im} corresponds to data for observation i and group m, n is the total number of observations, and b_{0m} and ε_{im} are independent of each other. After substituting the group-level parameters in the first-level model, the model for the response vector becomes

$$y_{im} = \underbrace{\beta_{00} + \beta_1 x_{im}}_{fixed\ effects} + \underbrace{b_{0m}}_{random\ effects} + \varepsilon_{im}.$$

A random intercept and slope model with one continuous predictor variable x, where both the intercept and slope vary independently by a grouping variable with M levels is

$$y_{im} = \beta_{0m} + \beta_{1m} x_{im} + \varepsilon_{im}, \quad i = 1, 2, \ldots, n, \quad m = 1, 2, \ldots, M, \quad \varepsilon_{im} \sim N\left(0, \sigma^2\right),$$
$$\beta_{0m} = \beta_{00} + b_{0m}, \quad b_{0m} \sim N\left(0, \sigma_0^2\right),$$
$$\beta_{1m} = \beta_{10} + b_{1m}, \quad b_{1m} \sim N\left(0, \sigma_1^2\right),$$

or

$$b_m = \begin{pmatrix} b_{0m} \\ b_{1m} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right).$$

You might also have correlated random effects. In general, for a model with a random intercept and slope, the distribution of the random effects is

$$b_m = \begin{pmatrix} b_{0m} \\ b_{1m} \end{pmatrix} \sim N\left(0, \sigma^2 D(\theta)\right),$$

where D is a 2-by-2 symmetric and positive semidefinite matrix, parameterized by a variance component vector θ.

After substituting the group-level parameters in the first-level model, the model for the response vector is

$$y_{im} = \underbrace{\beta_{00} + \beta_{10} x_{im}}_{fixed\ effects} + \underbrace{b_{0m} + b_{1m} x_{im}}_{random\ effects} + \varepsilon_{im}, \quad i = 1, 2, \ldots, n, \quad m = 1, 2, \ldots, M.$$

If you express the group-level variable, x_{im}, in the random-effects term by z_{im}, this model is

$$y_{im} = \underbrace{\beta_{00} + \beta_{10} x_{im}}_{fixed\ effects} + \underbrace{b_{0m} + b_{1m} z_{im}}_{random\ effects} + \varepsilon_{im}, \quad i = 1, 2, \ldots, n, \quad m = 1, 2, \ldots, M.$$

In this case, the same terms appear in both the fixed-effects design matrix and random-effects design matrix. Each z_{im} and x_{im} correspond to the level m of the grouping variable.

It is also possible to explain more of the group-level variations by adding more group-level predictor variables. A random-intercept and random-slope model with one continuous predictor variable x, where both the intercept and slope vary independently by a grouping variable with M levels, and one group-level predictor variable v_m is

$$y_{im} = \beta_{0im} + \beta_{1im} x_{im} + \varepsilon_{im}, \quad i = 1, 2, \ldots, n, \quad m = 1, 2, \ldots, M, \quad \varepsilon_{im} \sim N\left(0, \sigma^2\right),$$
$$\beta_{0im} = \beta_{00} + \beta_{01} v_{im} + b_{0m}, \quad b_{0m} \sim N\left(0, \sigma_0^2\right),$$
$$\beta_{1im} = \beta_{10} + \beta_{11} v_{im} + b_{1m}, \quad b_{1m} \sim N\left(0, \sigma_1^2\right).$$

This model results in main effects of the group-level predictor and an interaction term between the first-level and group-level predictor variables in the model for the response variable as

$$y_{im} = \beta_{00} + \beta_{01}v_{im} + b_{0m} + (\beta_{10} + \beta_{11}v_{im} + b_{1m})x_{im} + \varepsilon_{im}, \quad i = 1, 2, ..., n, \quad m = 1, 2, ..., M,$$
$$= \underbrace{\beta_{00} + \beta_{10}x_{im} + \beta_{01}v_{im} + \beta_{11}v_{im}x_{im}}_{fixed\ effects} + \underbrace{b_{0m} + b_{1m}x_{im}}_{random\ effects} + \varepsilon_{im}.$$

The term $\beta_{11}v_m x_{im}$ is often called a cross-level interaction in many textbooks on multilevel models. The model for the response variable y can be expressed as

$$y_{im} = \begin{bmatrix}1 & x_{1im} & v_{im} & v_{im}x_{1im}\end{bmatrix}\begin{bmatrix}\beta_{00}\\ \beta_{10}\\ \beta_{01}\\ \beta_{11}\end{bmatrix} + \begin{bmatrix}1 & x_{1im}\end{bmatrix}\begin{bmatrix}b_{0m}\\ b_{1m}\end{bmatrix} + \varepsilon_{im}, \quad i = 1, 2, ..., n, \quad m = 1, 2, ..., M,$$

which corresponds to the standard form given earlier,

$$y = X\beta + Zb + \varepsilon.$$

In general, if there are R grouping variables, and $m(r,i)$ shows the level of grouping variable r, for observation i, then the model for the response variable for observation i is

$$y_i = x_i^T\beta + \sum_{r=1}^{R} z_{ir}b^{(r)}_{m(r,i)} + \varepsilon_i, \quad i = 1, 2, ..., n,$$

where β is a p-by-1 fixed-effects vector, $b^{(r)}_{m(r,i)}$ is a $q(r)$-by-1 random-effects vector for the rth grouping variable and level $m(r,i)$, and ε_i is a 1-by-1 error term for observation i.

References

[1] Pinherio, J. C., and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing Series, Springer, 2004.

[2] Hariharan, S. and J. H. Rogers. “Estimation Procedures for Hierarchical Linear Models.” *Multilevel Modeling of Educational Data* (A. A. Connell and D. B. McCoach, eds.). Charlotte, NC: Information Age Publishing, Inc., 2008.

[3] Hox, J. *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Inc., 2002

[4] Snidjers, T. and R. Bosker. *Multilevel Analysis*. Thousand Oaks, CA: Sage Publications, 1999.

[5] Gelman, A. and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press, 2007.

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

More About

- “Prepare Data for Linear Mixed-Effects Models” on page 11-134

Prepare Data for Linear Mixed-Effects Models

In this section...

Tables and Dataset Arrays

To fit a linear-mixed effects model, you must store your data in a table or dataset array. In your table or dataset array, you must have a column for each variable including the response variable. More specifically, the table or dataset array, say `tbl`, must contain the following:

- A response variable y
- Predictive variables X_j which can be continuous or grouping variables
- Grouping variables g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, a character array, a string array, or a cell array of character vectors, $r = 1, 2, ..., R$.

You must organize your data so that each row represents an observation. And each row should contain the value of variables and the levels of grouping variables corresponding to that observation. For example, if you have data from an experiment with four treatment options, on five different types of individuals chosen randomly from a population of individuals (blocks), the table or dataset array must look like this.

Block	Treatment	Response
1	1	y11
1	2	y12
1	3	y13
1	4	y14
...	...	...
5	1	y51
5	2	y52
5	3	y53
5	4	y54

Now, consider a split-plot experiment, where the effect of four different types of fertilizers on the yield of tomato plants is studied. The soil where the tomato plants are planted is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five types of tomato plants, (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. Then, the tomato plants in the plots are divided into subplots, where each subplot is treated by one of the four fertilizers. The data from this experiment looks like:

Soil	Tomato	Fertilizer	Yield
'Sandy'	'Plum'	1	104

Soil	Tomato	Fertilizer	Yield
'Sandy'	'Plum'	2	136
'Sandy'	'Plum'	3	158
'Sandy'	'Plum'	4	174
'Sandy'	'Cherry'	1	57
'Sandy'	'Cherry'	2	86
...	...	...	...
'Sandy'	'Vine'	3	99
'Sandy'	'Vine'	4	117
'Silty'	'Plum'	1	120
'Silty'	'Plum'	2	115
...	...	...	...
'Loamy'	'Vine'	3	111
'Loamy'	'Vine'	4	105

You must specify the model you want to fit using the `formula` input argument to `fitlme`.

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` contains the fixed-effects terms and `random1, ..., randomR` contain the random-effects terms. For example, for the previous fertilizer experiment, consider the following mixed-effects model

$$y_{imjk} = \beta_0 + \sum_{m=2}^{4} \beta_{1m} I[F]_{im} + \sum_{j=2}^{5} \beta_{2j} I[T]_{ij} + b_{0k} S_k + b_{0jk} (S*T)_{jk} + \varepsilon_{imjk},$$

where $i = 1, 2, ..., 60$, the index m corresponds to the fertilizer types, j corresponds to the tomato types, and $k = 1, 2, 3$ corresponds to the blocks (soil). S_k represents the kth soil type, and $I[F]_{im}$ is the dummy variable representing level m of the fertilizer. Similarly, $I[T]_{ij}$ is the dummy variable representing the level j of the tomato type.

You can fit this model using the formula `'Yield ~ 1 + Fertilizer + Tomato + (1|Soil)+(1|Soil:Tomato)'`.

For detailed information on how to specify your model using formula, see “Relationship Between Formula and Design Matrices” on page 11-138.

Design Matrices

If you cannot easily describe your model using a formula, you can create design matrices to define the fixed and random effects, and fit the model using `fitlmematrix(X,y,Z,G)`. You must create your design matrices as follows.

Fixed-effects and random-effects design matrices `X` and `Z`:

- Enter a column of 1s for the intercept using `ones(n,1)`, where n is the total number of observations.

- If `X1` is a continuous variable, then enter `X1` as it is in a separate column.
- If `X1` is a categorical variable with *m* levels, then there must be *m* – 1 dummy variables for *m* – 1 levels of `X1` in `X`.

  For example, consider an experiment where you want to study the impact of quality of raw materials from four different providers on the productivity of a production line. If you fit a linear mixed-effects model with intercept and provider as the fixed-effects terms, intercept is the random-effects term, and you use reference contrasts coding, then you must construct your fixed- and random-effects design matrices as follows.

  ```
  D = dummyvar(provider); % Create dummy variables
  X = [ones(n,1) D(:,2) D(:,3) D(:,4)];
  Z = [ones(n,1)];
  ```

  Because reference contrast coding uses the first provider as the reference, and the model has an intercept, you must use the dummy variables for only the last three providers.
- If there is an interaction term of predictor variables `X1` and `X2`, then you must enter a column that you form by elementwise product of the vectors `X1` and `X2`.

  For example, if you want to fit a model, where there is an intercept, a continuous treatment factor, a continuous time factor, and their interaction as the fixed-effects in a longitudinal study, and time is the random-effects term, then your fixed- and random-effects design matrices should look like

  ```
  X = [ones(n,1),treatment,time,treatment.*time];
  y = response;
  Z = [time];
  ```

Grouping variables `G`:

There is one column for each grouping variable and a column of elementwise product of the grouping variables in case of a nesting.

For example, if you want to group plots (`plot`) within blocks (`block`), then you must add a column of elementwise product of `plot` by `block`. More specifically, if you want to fit a model where there is intercept and a continuous treatment factor as the fixed-effects in a split-block experiment, and the intercept and treatment are grouped by the plots nested within blocks, then the design matrices should look like this.

```
X = [ones(n,1),treatment];
y = response;
Z = [ones(n,1),treatment];
G = [block.*plot];
```

Suppose in the earlier quality of raw materials example, the raw materials arrive in bulks, and the bulks are nested within providers. If you want to fit a linear mixed-effects model, where intercept is grouped by the bulks within providers, then your design matrices should look like this.

```
D = dummyvar(provider);
X = [ones(n,1) D(:,2) D(:,3) D(:,4)];
y = response;
Z = ones(n,1);
G = [provider.*bulks];
```

In the earlier longitudinal study example, if you want to add random effects for intercept and time grouped by subjects that participated in the study, then your design matrices should look like

```
X = [ones(n,1),treatment,time, treatment.*time];
y = response;
Z = [ones(n,1),time];
G = subject;
```

Relation of Matrix Form to Tables and Dataset Arrays

`fitlme(tbl,formula)` and `fitlmematrix(X,y,Z,G)` are equivalent in functionality, such that

- `y` is the *n*-by-1 response vector.
- `X` is an *n*-by-*p* fixed-effects design matrix. `fitlme` constructs this from the expression `fixed` in `formula`.
- `Z` is an *R*-by-1 cell array with `Z{r}` being an *n*-by-*q*(*r*) random-effects design matrix constructed from the *r*th expression in `random` in `formula`, *r* = 1, 2, ..., *R*.
- `G` is an *R*-by-1 cell array with `G{r}` being an *n*-by-1 grouping variable, g_r, in `formula` with *M*(*r*) levels or groups.

For example, if `tbl` is a table or dataset array containing the response variable `y`, the continuous variables `X1` and `X2`, and the grouping variable `g`, then to fit a linear mixed-effects model that corresponds to the formula expression `'y ~ X1+ X2+ (X1*X2|g)'` using `fitlmematrix(X,y,Z,G)` the input arguments must correspond to the following:

```
y = tbl.y
X = [ones(n,1), tbl.X1, tbl.X2]
Z = [ones(n,1), tbl.X1, tbl.X2, tbl.X1.*tbl.X2]
G = tbl.g
```

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

More About

- “Linear Mixed-Effects Models” on page 11-131
- “Relationship Between Formula and Design Matrices” on page 11-138

Relationship Between Formula and Design Matrices

In this section...

Formula

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For the linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms.

Suppose a table `tbl` contains the following:

- A response variable, `y`
- Predictor variables, X_j, which can be continuous or grouping variables
- Grouping variables, g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form, `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random`$_1$ is a specification of the random-effects design matrix Z_1 corresponding to grouping variable g_1, and similarly `random`$_R$ is a specification of the random-effects design matrix Z_R corresponding to grouping variable g_R. You can express the `fixed` and `random` terms using Wilkinson notation.

Wilkinson notation describes the factors present in models. The notation relates to factors present in models, not to the multipliers (coefficients) of those factors.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	`X`, X^2, ..., X^k
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (elementwise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only
`- X2`	Do not include `X2`
`X1*X2 + X3`	`X1`, `X2`, `X3`, `X1*X2`
`X1 + X2 + X3 + X1:X2`	`X1`, `X2`, `X3`, `X1*X2`
`X1*X2*X3 - X1:X2:X3`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`, `X2*X3`
`X1*(X2 + X3)`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`. Here are some examples for linear mixed-effects model specification.

Examples:

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, `X1` and `X2`. This is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept and fixed effects for `X1` and `X2`. The implicit intercept term is suppressed by including `-1`.
`'y ~ 1 + (1 | g1)'`	Fixed effects for the intercept plus random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

Design Matrices for Fixed and Random Effects

`fitlme` converts the expressions in the `fixed` and `random` parts (not grouping variables) of a formula into design matrices as follows:

- Each term in a formula adds one or more columns to the corresponding design matrix.
- A term containing a single continuous variable adds one column to the design matrix.
- A fixed term containing a categorical variable `X` with k levels adds (k – 1) dummy variables to the design matrix.

  For example, if the variable `Supplier` represents three different suppliers a manufacturer receives parts from, i.e. a categorical variable with three levels, and out of six batches of parts, the first two batches come from supplier 1 (level 1), the second two batches come from supplier 2 (level 2), and the last two batches come from supplier 3 (level 3), such as

```
Supplier =

     1
     1
     2
     2
     3
     3
```

Then, adding `Supplier` to the formula as a fixed-effects or random-effects term adds the following two dummy variables to the corresponding design matrix, using the `'reference'` contrast:

```
 0     0
 0     0
 1     0
 1     0
 0     1
 0     1
```

For more details on dummy variables, see “Dummy Variables” on page 2-48. For other contrast options, see the `'DummyVarCoding'` name-value pair argument of `fitlme`.

- If `X1` and `X2` are continuous variables, the product term `X1:X2` adds one column obtained by elementwise multiplication of `X1` and `X2` to the design matrix.
- If `X1` is continuous and `X2` is categorical with *k* levels, the product term `X1:X2` multiplies elementwise `X1` with the (*k* – 1) dummy variables representing `X2`, and adds these (*k* – 1) columns to the design matrix.

  For example, if `Drug` is the amount of a drug given to patients, a continuous treatment, and `Time` is three distinct points in time when the health measures are taken, a categorical variable with three levels, and out of nine observations, the first three are observed at time point 1, the second three are observed at time point 2, and the last three are observed at time point 3 so that

```
[Drug Time] =

    0.1000    1.0000
    0.2000    1.0000
    0.5000    2.0000
    0.6000    2.0000
    0.3000    3.0000
    0.8000    3.0000
```

  Then, the product term `Drug:Time` adds the following two variables to the design matrix:

```
     0         0
     0         0
0.5000         0
0.6000         0
     0    0.3000
     0    0.8000
```

- If `X1` and `X2` are categorical variables with *k* and *m* levels respectively, the product term `X1:X2` adds (*k* – 1)*(*m* – 1) dummy variables to the design matrix formed by taking the elementwise product of each dummy variable representing `X1` with each dummy variable representing `X2`.

  For example, in an experiment to determine the impact of the type of corn and the popping method on the yield, suppose there are three types of `Corn` and two types of `Method` as follows:

```
    1    oil
    1    oil
    1    air
    1    air
    2    oil
    2    oil
    2    air
    2    air
```

```
3    oil
3    oil
3    air
3    air
```

Then, the interaction term `Corn:Method` adds the following to the design matrix:

```
0     0
0     0
0     0
0     0
1     0
1     0
0     0
0     0
0     1
0     1
0     0
0     0
```

- The term `X1*X2` adds the necessary number of columns for `X1`, `X2`, and `X1:X2` to the design matrix.
- The term `X1^2` adds the necessary number of columns for `X1` and `X1:X1` to the design matrix.
- The symbol `1` (one) in the formula stands for a column of all 1s. By default a column of 1s is included in the design matrix. To exclude a column of ones from the design matrix, you must explicitly specify `−1` as a term in the expression.

Grouping Variables

`fitlme` handles the grouping variables in the `(.|group)` part of a formula as follows:

- If a grouping variable has k levels, then k dummy variables represent this grouping.

  For example, suppose `District` is a categorical grouping variable with three levels, showing the three types of districts, and out of six schools, the first two are in district 1, the second two are in district 2, and the last two are in district 3, so that

```
District =

     1
     1
     2
     2
     3
     3
```

  Then, the dummy variables that represent this grouping are:

```
  1     0     0
  1     0     0
  0     1     0
  0     1     0
  0     0     1
  0     0     1
```

- If `X1` is a continuous random-effects variable and `X2` is a grouping variable with k levels, then the random term `(X1 − 1|X2)` multiplies elementwise `X1` with the k dummy variables representing `X2` and adds these k columns to the random-effects design matrix.

For example, suppose `Score` is a continuous variable showing the scores of students from a math exam in a school, and `Class` is a categorical variable with three levels, showing the three different classes in a school. Also, suppose out of nine observations first three correspond to the scores of students in the first class, the second three correspond to scores of students in the second class, and the last three correspond to the scores of students in the third class, such as

```
[Score Class] =

    78.0000    1.0000
    68.0000    1.0000
    81.0000    2.0000
    53.0000    2.0000
    85.0000    3.0000
    72.0000    3.0000
```

Then, the random term `(Score – 1|Class)` adds the following three columns to the random-effects design matrix:

```
 78.0000          0          0
 68.0000          0          0
       0    81.0000          0
       0    53.0000          0
       0          0    85.0000
       0          0    72.0000
```

- If `X1` is a continuous predictor variable and `X2` and `X3` are grouping variables with k and m levels respectively, the term `(X1|X2:X3)` represents this grouping of `X1` with $k*m$ dummy variables formed by taking the elementwise product of each dummy variable representing `X2` with each dummy variable representing `X3`.

  For example, suppose `Treatment` is a continuous predictor variable, and there are three levels of `Block` and two levels of `Plot` nested within the block as follows:

```
   0.1000    1    a
   0.2000    1    b
   0.5000    2    a
   0.6000    2    b
   0.3000    3    a
   0.8000    3    b
```

  Then, the random term `(Treatment – 1|Block:Plot)` adds the following to the random-effects design matrix:

```
 0.1000         0         0         0         0         0
      0    0.2000         0         0         0         0
      0         0    0.5000         0         0         0
      0         0         0    0.6000         0         0
      0         0         0         0    0.3000         0
      0         0         0         0         0    0.8000
```

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

More About

- “Prepare Data for Linear Mixed-Effects Models” on page 11-134

Estimating Parameters in Linear Mixed-Effects Models

In this section...

A linear mixed-effects model is of the form

$$y = \underbrace{X\beta}_{fixed} + \underbrace{Zb}_{random} + \underbrace{\varepsilon}_{error},$$

where

- y is the n-by-1 response vector, and n is the number of observations.
- X is an n-by-p fixed-effects design matrix.
- β is a p-by-1 fixed-effects vector.
- Z is an n-by-q random-effects design matrix.
- b is a q-by-1 random-effects vector.
- ε is the n-by-1 observation error vector.

The random-effects vector, b, and the error vector, ε, are assumed to have the following independent prior distributions:

$$b \sim N\left(0, \sigma^2 D(\theta)\right),$$
$$\varepsilon \sim N\left(0, \sigma^2 I\right),$$

where D is a symmetric and positive semidefinite matrix, parameterized by a variance component vector θ, I is an n-by-n identity matrix, and σ^2 is the error variance.

In this model, the parameters to estimate are the fixed-effects coefficients β, and the variance components θ and σ^2. The two most commonly used approaches to parameter estimation in linear mixed-effects models are maximum likelihood and restricted maximum likelihood methods.

Maximum Likelihood (ML)

The maximum likelihood estimation includes both regression coefficients and the variance components, that is, both fixed-effects and random-effects terms in the likelihood function.

For a linear mixed-effects model defined above, the conditional response of the response variable y given β, b, θ, and σ^2 is

$$y \left| b, \beta, \theta, \sigma^2 \sim N\left(X\beta + Zb, \sigma^2 I_n\right)\right. .$$

The likelihood of y given β, θ, and σ^2 is

$$P\left(y \left| \beta, \theta, \sigma^2\right.\right) = \int P\left(y \left| b, \beta, \theta, \sigma^2\right.\right) P\left(b \left| \theta, \sigma^2\right.\right) db,$$

where

$$P\left(b\left|\theta,\sigma^2\right.\right)=\frac{1}{\left(2\pi\sigma^2\right)^{q/2}}\frac{1}{\left|D(\theta)\right|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}b^TD^{-1}b\right\}\quad\text{and}$$

$$P\left(y\left|b,\beta,\theta,\sigma^2\right.\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta-Zb)^T(y-X\beta-Zb)\right\}.$$

Suppose $\Lambda(\theta)$ is the lower triangular Cholesky factor of $D(\theta)$ and $\Delta(\theta)$ is the inverse of $\Lambda(\theta)$. Then,

$$D(\theta)^{-1}=\Delta(\theta)^T\Delta(\theta).$$

Define

$$r^2(\beta,b,\theta)=b^T\Delta(\theta)^T\Delta(\theta)b+(y-X\beta-Zb)^T(y-X\beta-Zb),$$

and suppose b^* is the value of b that satisfies

$$\left.\frac{\partial r^2(\beta,b,\theta)}{\partial b}\right|_{b^*}=0$$

for given β and θ. Then, the likelihood function is

$$P\left(y\left|\beta,\theta,\sigma^2\right.\right)=\left(2\pi\sigma^2\right)^{-n/2}\left|D(\theta)\right|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}r^2(\beta,b^*(\beta),\theta)\right\}\frac{1}{\left|\Delta^T\Delta+Z^TZ\right|^{1/2}}.$$

$P(y|\beta,\theta,\sigma^2)$ is first maximized with respect to β and σ^2 for a given θ. Thus the optimized solutions $\widehat{\beta}(\theta)$ and $\widehat{\sigma}^2(\theta)$are obtained as functions of θ. Substituting these solutions into the likelihood function produces $P\left(y\left|\widehat{\beta}(\theta),\theta,\widehat{\sigma}^2(\theta)\right.\right)$. This expression is called a profiled likelihood where β and σ^2 have been profiled out. $P\left(y\left|\widehat{\beta}(\theta),\theta,\widehat{\sigma}^2(\theta)\right.\right)$ is a function of θ, and the algorithm then optimizes it with respect to θ. Once it finds the optimal estimate of θ, the estimates of β and σ^2 are given by $\widehat{\beta}(\theta)$ and $\widehat{\sigma}^2(\theta)$.

The ML method treats β as fixed but unknown quantities when the variance components are estimated, but does not take into account the degrees of freedom lost by estimating the fixed effects. This causes ML estimates to be biased with smaller variances. However, one advantage of ML over REML is that it is possible to compare two models in terms of their fixed- and random-effects terms. On the other hand, if you use REML to estimate the parameters, you can only compare two models, that are nested in their random-effects terms, with the same fixed-effects design.

Restricted Maximum Likelihood (REML)

Restricted maximum likelihood estimation includes only the variance components, that is, the parameters that parameterize the random-effects terms in the linear mixed-effects model. β is estimated in a second step. Assuming a uniform improper prior distribution for β and integrating the likelihood $P(y|\beta,\theta,\sigma^2)$ with respect to β results in the restricted likelihood $P(y|\theta,\sigma^2)$. That is,

$$P\left(y\left|\theta,\sigma^2\right.\right)=\int P\left(y\left|\beta,\theta,\sigma^2\right.\right)P(\beta)d\beta=\int P\left(y\left|\beta,\theta,\sigma^2\right.\right)d\beta.$$

The algorithm first profiles out $\widehat{\sigma}_R^2$ and maximizes remaining objective function with respect to θ to find $\widehat{\theta}_R$. The restricted likelihood is then maximized with respect to σ^2 to find $\widehat{\sigma}_R^2$. Then, it estimates β by finding its expected value with respect to the posterior distribution

$$P\left(\beta \middle| y, \widehat{\theta}_R, \widehat{\sigma}_R^2\right).$$

REML accounts for the degrees of freedom lost by estimating the fixed effects, and makes a less biased estimation of random effects variances. The estimates of θ and σ^2 are invariant to the value of β and less sensitive to outliers in the data compared to ML estimates. However, if you use REML to estimate the parameters, you can only compare two models that have the identical fixed-effects design matrices and are nested in their random-effects terms.

References

[1] Pinherio, J. C., and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing Series, Springer, 2004.

[2] Hariharan, S. and J. H. Rogers. “Estimation Procedures for Hierarchical Linear Models.” *Multilevel Modeling of Educational Data* (A. A. Connell and D. B. McCoach, eds.). Charlotte, NC: Information Age Publishing, Inc., 2008.

[3] Raudenbush, S. W. and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2002.

[4] Hox, J. *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Inc, 2002.

[5] Snidjers, T. and R. Bosker. *Multilevel Analysis*. Thousand Oaks, CA: Sage Publications, 1999.

[6] McCulloch, C.E., R. S. Shayle, and J. M. Neuhaus. *Generalized, Linear, and Mixed Models*. Wiley, 2008.

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

More About

- “Linear Mixed-Effects Models” on page 11-131

Linear Mixed-Effects Model Workflow

This example shows how to fit and analyze a linear mixed-effects model (LME).

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the CDC).

Reorganize and plot the data.

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the response variable `FluRate`, the nominal variable `Region` that shows which region each estimate is from, the nationwide estimate `WtdILI`, and the grouping variable `Date`.

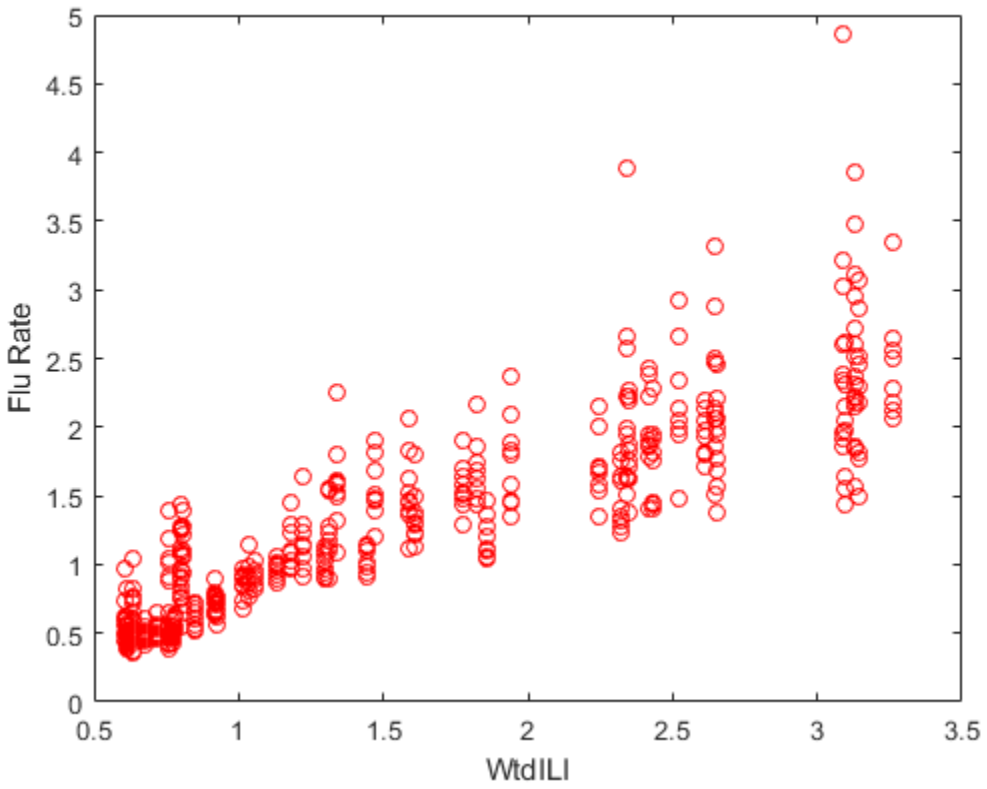
```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate',...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Define `flu2` as a `table`.

```
flu2 = dataset2table(flu2);
```

Plot flu rates versus the nationwide estimate.

```
plot(flu2.WtdILI,flu2.FluRate,'ro')
xlabel('WtdILI')
ylabel('Flu Rate')
```



You can see that the flu rates in regions have a direct relationship with the nationwide estimate.

Fit an LME model and interpret the results.

Fit a linear mixed-effects model with the nationwide estimate as the predictor variable and a random intercept that varies by `Date`.

```
lme = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date)')
```

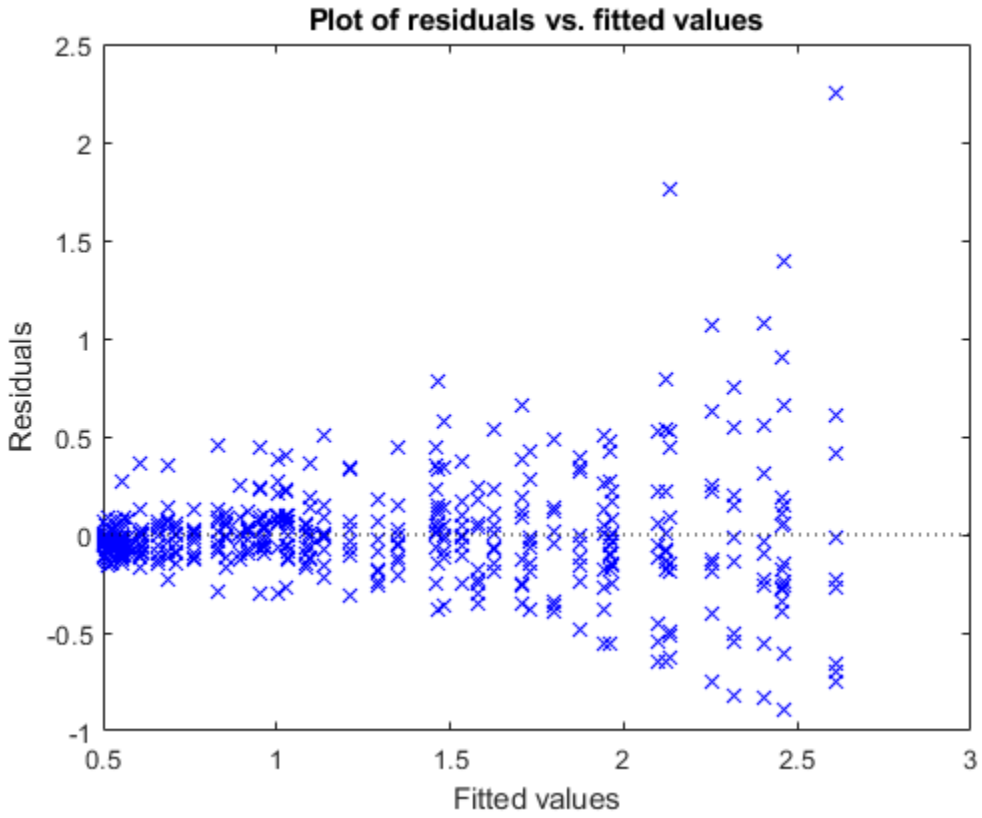
```
lme =


Linear mixed-effects model fit by ML

Model information:
    Number of observations             468
    Fixed effects coefficients           2
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    FluRate ~ 1 + WtdILI + (1 | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    286.24    302.83    -139.12          278.24
```

```
Fixed effects coefficients (95% CIs):
    Name                  Estimate    SE          tStat     DF     pValue
    {'(Intercept)'}       0.16385     0.057525    2.8484    466     0.0045885
    {'WtdILI'     }        0.7236     0.032219    22.459    466    3.0502e-76


    Lower       Upper
    0.050813    0.27689
     0.66028    0.78691

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       0.17146


    Lower      Upper
    0.13227    0.22226

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.30201     0.28217    0.32324
```

The small p-values of 0.0045885 and 3.0502e-76 indicate that both the intercept and nationwide estimate are significant. Also, the confidence limits for the standard deviation of the random-effects term, σ_b, do not include 0 (0.13227, 0.22226), which indicates that the random-effects term is significant.

Plot the raw residuals versus the fitted values.

```
figure();
plotResiduals(lme,'fitted')
```



The variance of residuals increases with increasing fitted response values, which is known as heteroscedasticity.

Find the two observations on the top right that appear like outliers.

```
find(residuals(lme) > 1.5)
```

```
ans =

    98
   107
```

Refit the model by removing these observations.

```
lme = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date)','Exclude',[98,107]);
```

Improve the model.

Determine if including an independent random term for the nationwide estimate grouped by `Date` improves the model.

```
altlme = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date) + (WtdILI-1|Date)',...
'Exclude',[98,107])
```

```
altlme =
```

```
Linear mixed-effects model fit by ML

Model information:
    Number of observations             466
    Fixed effects coefficients           2
    Random effects coefficients        104
    Covariance parameters                3

Formula:
    FluRate ~ 1 + WtdILI + (1 | Date) + (WtdILI | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    179.39    200.11    -84.694          169.39

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat     DF     pValue
    {'(Intercept)'}        0.17837     0.054585    3.2676    464     0.001165
    {'WtdILI'     }        0.70836     0.030594    23.153    464    2.123e-79


    Lower      Upper
     0.0711    0.28563
    0.64824    0.76849

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.16631


    Lower      Upper
    0.12977    0.21313

Group: Date (52 Levels)
    Name1             Name2             Type           Estimate      Lower
    {'WtdILI'}        {'WtdILI'}        {'std'}        4.6939e-08    NaN


    Upper
    NaN

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.26691     0.24934    0.28572
```

The estimated standard deviation of `WtdILI` term is nearly 0 and its confidence interval cannot be computed. This is an indication that the model is overparameterized and the `(WtdILI-1|Date)` term is not significant. You can formally test this using the `compare` method as follows: `compare(lme,altlme,'CheckNesting',true)`.

Add a random effects-term for intercept grouped by Region to the initial model `lme`.

```
lme2 = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date) + (1|Region)',...
'Exclude',[98,107]);
```

Compare the models `lme` and `lme2`.

```
compare(lme,lme2,'CheckNesting',true)
```

```
ans =


    THEORETICAL LIKELIHOOD RATIO TEST

    Model    DF    AIC       BIC       LogLik     LRStat    deltaDF    pValue
    lme      4     177.39    193.97    -84.694
    lme2     5     62.265    82.986    -26.133    117.12    1          0
```

The p-value of 0 indicates that `lme2` is a better fit than `lme`.

Now, check if adding a potentially correlated random-effects term for the intercept and national average improves the model `lme2`.

```
lme3 = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date) + (1 + WtdILI|Region)',...
'Exclude',[98,107])
```

```
lme3 =


Linear mixed-effects model fit by ML

Model information:
    Number of observations             466
    Fixed effects coefficients           2
    Random effects coefficients         70
    Covariance parameters                5

Formula:
    FluRate ~ 1 + WtdILI + (1 | Date) + (1 + WtdILI | Region)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    13.338    42.348    0.33076          -0.66153

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat     DF     pValue
    {'(Intercept)'}         0.1795     0.054953    3.2665    464     0.0011697
    {'WtdILI'     }        0.70719     0.04252     16.632    464    4.6451e-49


    Lower       Upper
    0.071514    0.28749
     0.62363    0.79074

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.17634
```

```
    Lower      Upper
    0.14093    0.22064

Group: Region (9 Levels)
    Name1                  Name2                  Type            Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std' }        0.0077038
    {'WtdILI'     }        {'(Intercept)'}        {'corr'}        -0.059604
    {'WtdILI'     }        {'WtdILI'     }        {'std' }         0.088069


    Lower         Upper
    3.2081e-16    1.8499e+11
      -0.99996       0.99995
      0.051693       0.15004

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.20976     0.19568    0.22486
```

The estimate for the standard deviation of the random-effects term for intercept grouped by Region is 0.0077037, its confidence interval is very large and includes zero. This indicates that the random-effects for intercept grouped by Region is insignificant. The correlation between the random-effects for intercept and `WtdILI` is -0.059604. Its confidence interval is also very large and includes zero. This is an indication that the correlation is not significant.

Refit the model by eliminating the intercept from the `(1 + WtdILI | Region)` random-effects term.

```
lme3 = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date) + (WtdILI - 1|Region)',...
'Exclude',[98,107])
```

```
lme3 =


Linear mixed-effects model fit by ML

Model information:
    Number of observations             466
    Fixed effects coefficients           2
    Random effects coefficients         61
    Covariance parameters                3

Formula:
    FluRate ~ 1 + WtdILI + (1 | Date) + (WtdILI | Region)

Model fit statistics:
    AIC       BIC      LogLikelihood    Deviance
    9.3395    30.06    0.33023          -0.66046

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat     DF     pValue
    {'(Intercept)'}         0.1795     0.054892    3.2702    464     0.0011549
    {'WtdILI'     }        0.70718     0.042486    16.645    464    4.0496e-49


    Lower       Upper
```

```
    0.071637    0.28737
     0.62369    0.79067

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       0.17633


    Lower      Upper
    0.14092    0.22062

Group: Region (9 Levels)
    Name1             Name2             Type          Estimate    Lower
    {'WtdILI'}        {'WtdILI'}        {'std'}       0.087925    0.054474


    Upper
    0.14192

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.20979     0.19585    0.22473
```

All terms in the new model `lme3` are significant.

Compare `lme2` and `lme3`.

```
compare(lme2,lme3,'CheckNesting',true,'NSim',100)
```

```
ans =


    SIMULATED LIKELIHOOD RATIO TEST: NSIM = 100, ALPHA = 0.05

    Model    DF    AIC       BIC       LogLik     LRStat    pValue
    lme2     5     62.265    82.986    -26.133
    lme3     5     9.3395     30.06    0.33023    52.926    0.009901


    Lower         Upper

    0.00025064    0.053932
```

The P-value of 0.009901 indicates that `lme3` is a better fit than `lme2`.

Add a quadratic fixed-effects term to the model `lme3`.

```
lme4 = fitlme(flu2,'FluRate ~ 1 + WtdILI^2 + (1|Date) + (WtdILI - 1|Region)',...
'Exclude',[98,107])
```

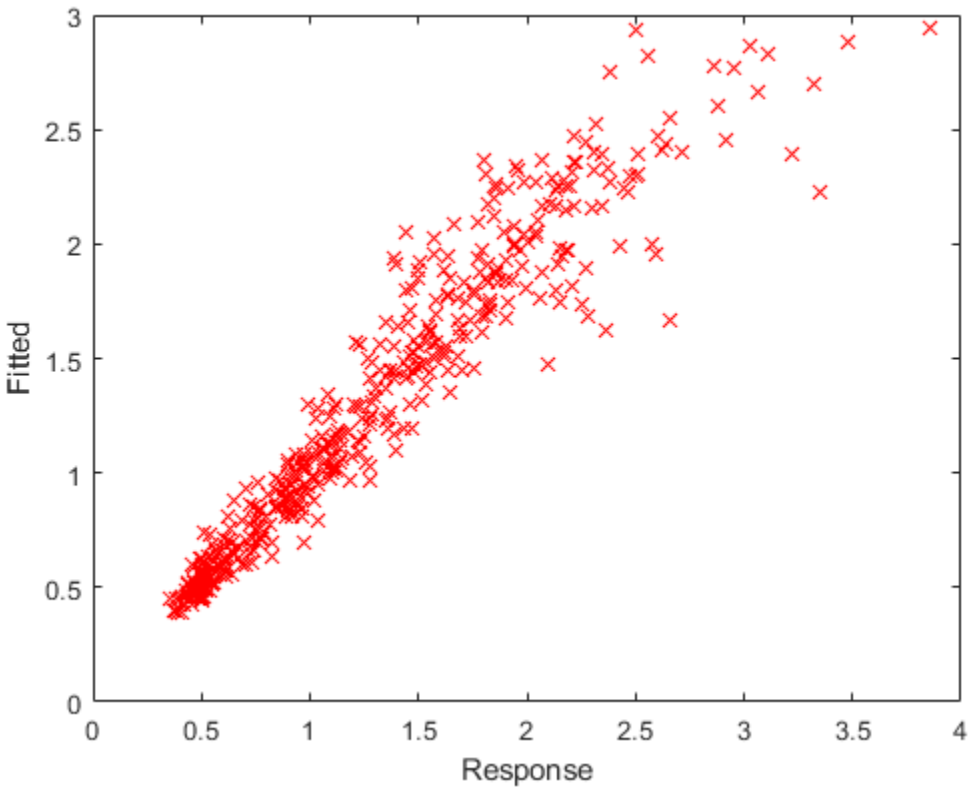
```
lme4 =


Linear mixed-effects model fit by ML
```

```
Model information:
    Number of observations             466
    Fixed effects coefficients           3
    Random effects coefficients         61
    Covariance parameters                3

Formula:
    FluRate ~ 1 + WtdILI + WtdILI^2 + (1 | Date) + (WtdILI | Region)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    6.7234    31.588    2.6383           -5.2766

Fixed effects coefficients (95% CIs):
    Name                   Estimate     SE         tStat       DF     pValue
    {'(Intercept)'}        -0.063406    0.12236    -0.51821    463       0.60456
    {'WtdILI'     }           1.0594    0.16554      6.3996    463    3.8232e-10
    {'WtdILI^2'   }        -0.096919     0.0441     -2.1977    463      0.028463


    Lower       Upper
    -0.30385      0.17704
     0.73406       1.3847
    -0.18358    -0.010259

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.16732


    Lower      Upper
    0.13326    0.21009

Group: Region (9 Levels)
    Name1             Name2             Type           Estimate    Lower
    {'WtdILI'}        {'WtdILI'}        {'std'}        0.087865    0.054443


    Upper
    0.1418

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.20979     0.19585    0.22473
```

The p-value of 0.028463 indicates that the coefficient of the quadratic term `WtdILI^2` is significant.

Plot the fitted response versus the observed response and residuals.

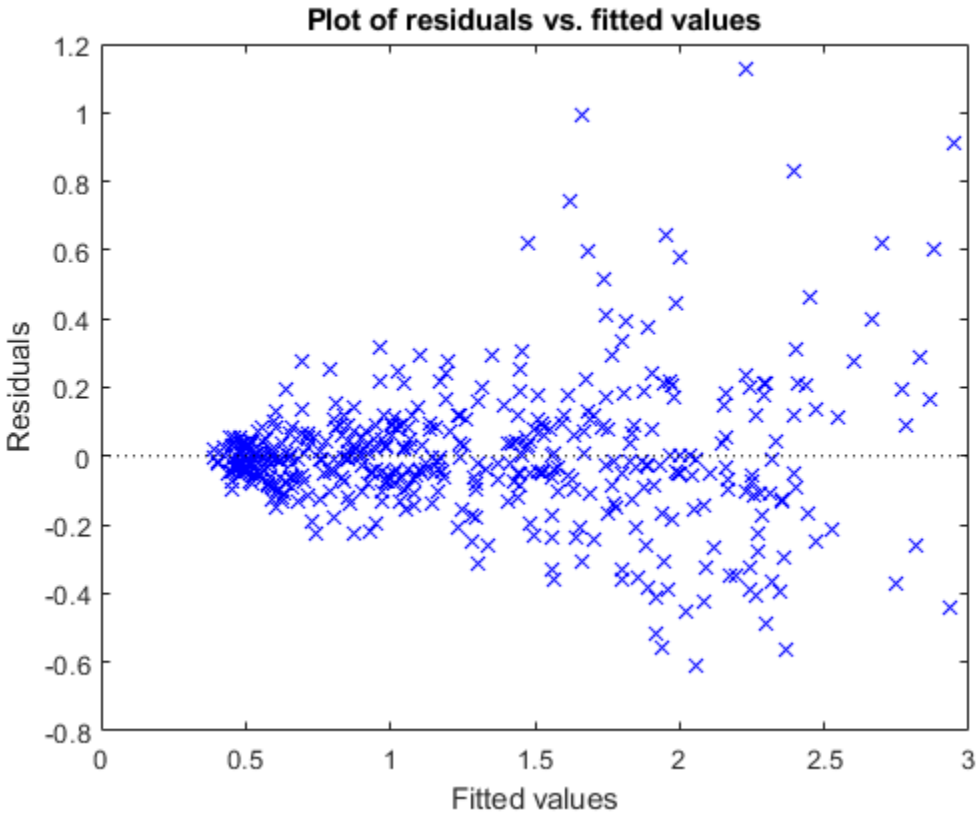
```
F = fitted(lme4);
R = response(lme4);
figure();
plot(R,F,'rx')
xlabel('Response')
ylabel('Fitted')
```



The fitted versus observed response values form almost 45-degree angle indicating a good fit.

Plot the residuals versus the fitted values.

```
figure();
plotResiduals(lme4,'fitted')
```



Although it has improved, you can still see some heteroscedasticity in the model. This might be due to another predictor that does not exist in the data set, hence not in the model.

Find the fitted flu rate value for region ENCentral, date 11/6/2005.

```
F(flu2.Region == 'ENCentral' & flu2.Date == '11/6/2005')
```

```
ans =

    1.4860
```

Randomly generate response values.

Randomly generate response values for a national estimate of 1.625, region MidAtl, and date 4/23/2006. First, define the new table. Because Date and Region are nominal in the original table, you must define them similarly in the new table.

```
tblnew.Date = nominal('4/23/2006');
tblnew.WtdILI = 1.625;
tblnew.Region = nominal('MidAtl');
tblnew = struct2table(tblnew);
```

Now, generate the response value.

```
random(lme4,tblnew)
```

```
ans =

    1.2679
```

Fit Mixed-Effects Spline Regression

This example shows how to fit a mixed-effects linear spline model.

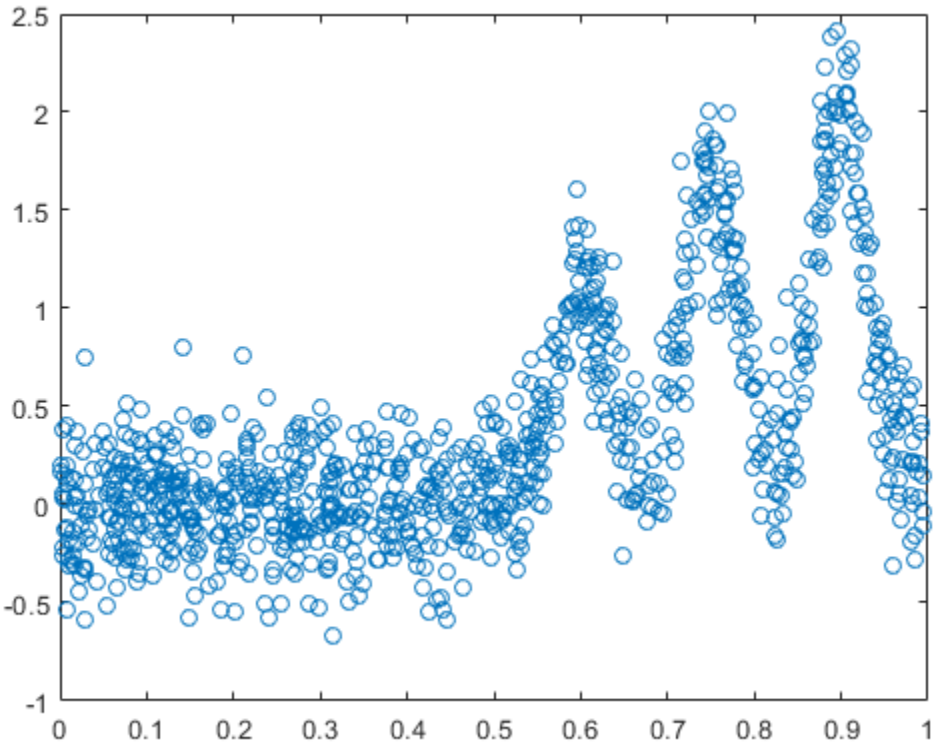
Load the sample data.

```
load('mespline.mat');
```

This is simulated data.

Plot y versus sorted x.

```
[x_sorted,I] = sort(x,'ascend');
plot(x_sorted,y(I),'o')
```



Fit the following mixed-effects linear spline regression model

$$y_i = \beta_1 + \beta_2 x_i + \sum_{j=1}^{K} b_j (x_i - k_j)_+ + \epsilon_i$$

where k_j is the j th knot, and K is the total number of knots. Assume that $b_j \sim N(0, \sigma_b^2)$ and $\epsilon \sim N(0, \sigma^2)$.

Define the knots.

```
k = linspace(0.05,0.95,100);
```

Define the design matrices.

```
X = [ones(1000,1),x];
Z = zeros(length(x),length(k));
for j = 1:length(k)
      Z(:,j) = max(X(:,2) - k(j),0);
end
```

Fit the model with an isotropic covariance structure for the random effects.

```
lme = fitlmematrix(X,y,Z,[],'CovariancePattern','Isotropic');
```

Fit a fixed-effects only model.

```
X = [X Z];
lme_fixed = fitlmematrix(X,y,[],[]);
```

Compare `lme_fixed` and `lme` via a simulated likelihood ratio test.

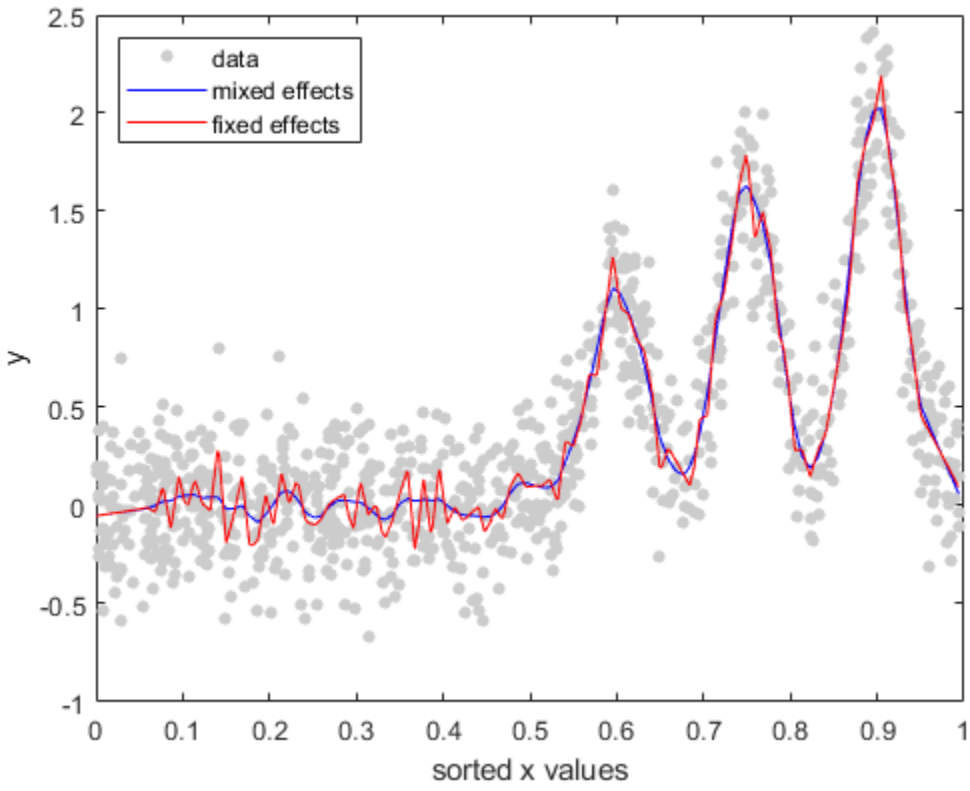
```
compare(lme,lme_fixed,'NSim',500,'CheckNesting',true)
```

```
ans =
    Simulated Likelihood Ratio Test: Nsim = 500, Alpha = 0.05

    Model        DF     AIC       BIC       LogLik     LRStat    pValue
    lme            4    170.62    190.25    -81.309
    lme_fixed    103    113.38    618.88     46.309    255.24    0.68064


    Lower      Upper

    0.63784    0.72129
```

The *p*-value indicates that the fixed-effects only model is not a better fit than the mixed-effects spline regression model.

Plot the fitted values from both models on top of the original response data.

```
R = response(lme);
figure();
plot(x_sorted,R(I),'o', 'MarkerFaceColor',[0.8,0.8,0.8],...
    'MarkerEdgeColor',[0.8,0.8,0.8],'MarkerSize',4);
hold on
F = fitted(lme);
F_fixed = fitted(lme_fixed);
plot(x_sorted,F(I),'b');
plot(x_sorted,F_fixed(I),'r');
legend('data','mixed effects','fixed effects','Location','NorthWest')
xlabel('sorted x values');
ylabel('y');
hold off
```



You can also see from the figure that the mixed-effects model provides a better fit to data than the fixed-effects only model.

Train Linear Regression Model

Statistics and Machine Learning Toolbox™ provides several features for training a linear regression model.

- For greater accuracy on low-dimensional through medium-dimensional data sets, use `fitlm`. After fitting the model, you can use the object functions to improve, evaluate, and visualize the fitted model. To regularize a regression, use `lasso` or `ridge`.
- For reduced computation time on high-dimensional data sets, use `fitrlinear`. This function offers useful options for cross-validation, regularization, and hyperparameter optimization.

This example shows the typical workflow for linear regression analysis using `fitlm`. The workflow includes preparing a data set, fitting a linear regression model, evaluating and improving the fitted model, and predicting response values for new predictor data. The example also describes how to fit and evaluate a linear regression model for tall arrays.

Prepare Data

Load the sample data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 10 variables with information on the sales of properties in New York City in 2015. This example uses some of these variables to analyze the sale prices.

Instead of loading the sample data set `NYCHousing2015`, you can download the data from the NYC Open Data website and import the data as follows.

```
folder = 'Annualized_Rolling_Sales_Update';
ds = spreadsheetDatastore(folder,"TextType","string","NumHeaderLines",4);
ds.Files = ds.Files(contains(ds.Files,"2015"));
ds.SelectedVariableNames = ["BOROUGH","NEIGHBORHOOD","BUILDINGCLASSCATEGORY","RESIDENTIALUNITS",
    "COMMERCIALUNITS","LANDSQUAREFEET","GROSSSQUAREFEET","YEARBUILT","SALEPRICE","SALEDATE"];
NYCHousing2015 = readall(ds);
```

Preprocess the data set to choose the predictor variables of interest. First, change the variable names to lowercase for readability.

```
NYCHousing2015.Properties.VariableNames = lower(NYCHousing2015.Properties.VariableNames);
```

Next, convert the `saledate` variable, specified as a `datetime` array, into two numeric columns `MM` (month) and `DD` (day), and remove the `saledate` variable. Ignore the year values because all samples are for the year 2015.

```
[~,NYCHousing2015.MM,NYCHousing2015.DD] = ymd(NYCHousing2015.saledate);
NYCHousing2015.saledate = [];
```

The numeric values in the `borough` variable indicate the names of the boroughs. Change the variable to a categorical variable using the names.

```
NYCHousing2015.borough = categorical(NYCHousing2015.borough,1:5, ...
    ["Manhattan","Bronx","Brooklyn","Queens","Staten Island"]);
```

The `neighborhood` variable has 254 categories. Remove this variable for simplicity.

```
NYCHousing2015.neighborhood = [];
```

Convert the `buildingclasscategory` variable to a categorical variable, and explore the variable by using the `wordcloud` function.

```
NYCHousing2015.buildingclasscategory = categorical(NYCHousing2015.buildingclasscategory);
wordcloud(NYCHousing2015.buildingclasscategory);
```

Assume that you are interested only in one-, two-, and three-family dwellings. Find the sample indices for these dwellings and delete the other samples. Then, change the data type of the `buildingclasscategory` variable to `double`.

```
idx = ismember(string(NYCHousing2015.buildingclasscategory), ...
    ["01  ONE FAMILY DWELLINGS","02  TWO FAMILY DWELLINGS","03  THREE FAMILY DWELLINGS"]);
NYCHousing2015 = NYCHousing2015(idx,:);
NYCHousing2015.buildingclasscategory = renamecats(NYCHousing2015.buildingclasscategory, ...
    ["01  ONE FAMILY DWELLINGS","02  TWO FAMILY DWELLINGS","03  THREE FAMILY DWELLINGS"], ...
    ["1","2","3"]);
NYCHousing2015.buildingclasscategory = double(NYCHousing2015.buildingclasscategory);
```

The `buildingclasscategory` variable now indicates the number of families in one dwelling.

Explore the response variable `saleprice` using the `summary` function.

```
s = summary(NYCHousing2015);
s.saleprice
```

```
ans = struct with fields:
           Size: [37881 1]
           Type: 'double'
```

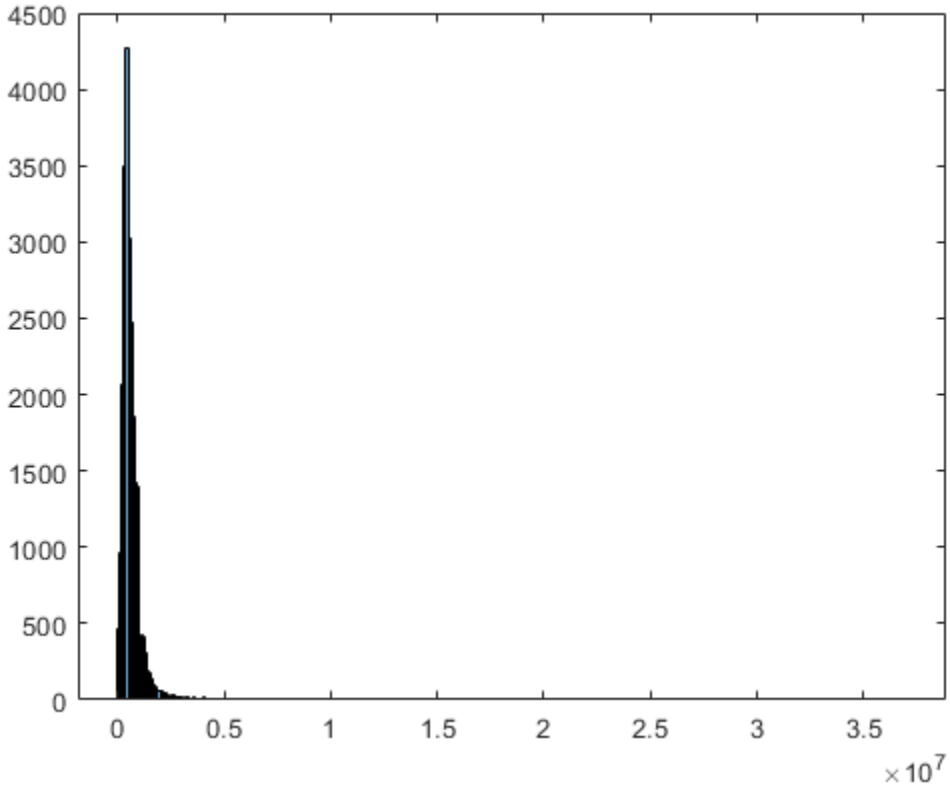
```
    Description: ''
          Units: ''
     Continuity: []
            Min: 0
         Median: 352000
            Max: 37000000
     NumMissing: 0
```

Assume that a `saleprice` less than or equal to $1000 indicates ownership transfer without a cash consideration. Remove the samples that have this `saleprice`.

```
idx0 = NYCHousing2015.saleprice <= 1000;
NYCHousing2015(idx0,:) = [];
```

Create a histogram of the `saleprice` variable.

```
histogram(NYCHousing2015.saleprice)
```

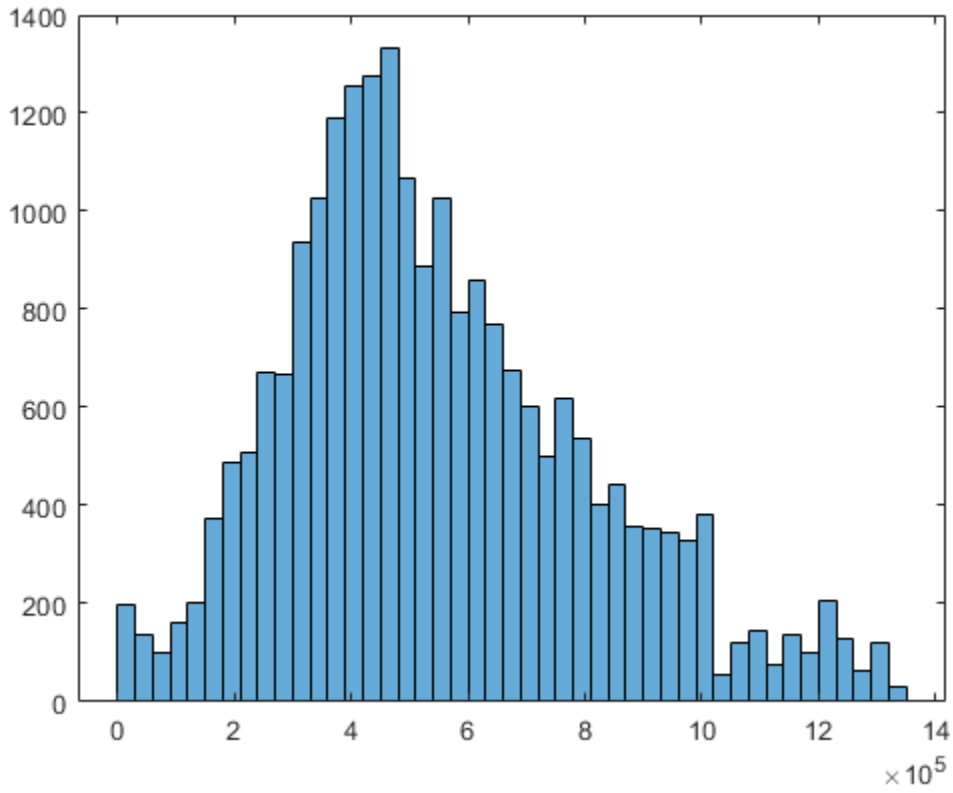


The maximum value of `saleprice` is 3.7×10^7, but most values are smaller than 0.5×10^7. You can identify the outliers of `saleprice` by using the `isoutlier` function.

```
idx = isoutlier(NYCHousing2015.saleprice);
```

Remove the identified outliers and create the histogram again.

```
NYCHousing2015(idx,:) = [];
histogram(NYCHousing2015.saleprice)
```



Partition the data set into a training set and test set by using `cvpartition`.

```
rng('default') % For reproducibility
c = cvpartition(height(NYCHousing2015),"holdout",0.3);
trainData = NYCHousing2015(training(c),:);
testData = NYCHousing2015(test(c),:);
```

Train Model

Fit a linear regression model by using the `fitlm` function.

```
mdl = fitlm(trainData,"PredictorVars",["borough","grosssquarefeet", ...
    "landsquarefeet","buildingclasscategory","yearbuilt","MM","DD"], ...
    "ResponseVar","saleprice")
```

```
mdl =
Linear regression model:
    saleprice ~ 1 + borough + buildingclasscategory + landsquarefeet + grosssquarefeet + yearbui

Estimated Coefficients:
                            Estimate          SE         tStat        pValue
                           __________    __________    ________    __________

    (Intercept)             2.0345e+05    1.0308e+05      1.9736       0.048441
    borough_Bronx          -3.0165e+05         56676     -5.3224     1.0378e-07
    borough_Brooklyn            -41160         56490    -0.72862        0.46624
    borough_Queens              -91136         56537      -1.612        0.10699
    borough_Staten Island  -2.2199e+05         56726     -3.9134     9.1385e-05
```

```
    buildingclasscategory         3165.7        3510.3     0.90185        0.36715
    landsquarefeet                13.149       0.84534      15.555      3.714e-54
    grosssquarefeet               112.34        2.9494       38.09    8.0393e-304
    yearbuilt                     100.07        45.464       2.201        0.02775
    MM                            3850.5        543.79      7.0808     1.4936e-12
    DD                           -367.19        207.56     -1.7691       0.076896


Number of observations: 15848, Error degrees of freedom: 15837
Root Mean Squared Error: 2.32e+05
R-squared: 0.235,  Adjusted R-Squared: 0.235
F-statistic vs. constant model: 487, p-value = 0
```

`mdl` is a `LinearModel` object. The model display includes the model formula, estimated coefficients, and summary statistics.

`borough` is a categorical variable that has five categories: `Manhattan`, `Bronx`, `Brooklyn`, `Queens`, and `Staten Island`. The fitted model `mdl` has four indicator variables. The `fitlm` function uses the first category `Manhattan` as a reference level, so the model does not include the indicator variable for the reference level. `fitlm` fixes the coefficient of the indicator variable for the reference level as zero. The coefficient values of the four indicator variables are relative to `Manhattan`. For more details on how the function treats a categorical predictor, see “Algorithms” on page 33-1997 of `fitlm`.

To learn how to interpret the values in the model display, see “Interpret Linear Regression Results” on page 11-50.

You can use the properties of a `LinearModel` object to investigate a fitted linear regression model. The object properties include information about coefficient estimates, summary statistics, fitting method, and input data. For example, you can find the R-squared and adjusted R-squared values in the `Rsquared` property. You can access the property values through the Workspace browser or using dot notation.

```
mdl.Rsquared
```

```
ans = struct with fields:
    Ordinary: 0.2352
    Adjusted: 0.2348
```

The model display also shows these values. The R-squared value indicates that the model explains approximately 24% of the variability in the response variable. See “Properties” on page 33-3508 of a `LinearModel` object for details about other properties.

Evaluate Model

The model display shows the *p*-value of each coefficient. The *p*-values indicate which variables are significant to the model. For the categorical predictor `borough`, the model uses four indicator variables and displays four *p*-values. To examine the categorical variable as a group of indicator variables, use the object function `anova`. This function returns analysis of variance (ANOVA) statistics of the model.

```
anova(mdl)
```

```
ans=8×5 table
                              SumSq        DF        MeanSq         F         pValue
```

```
borough                 1.123e+14        4    2.8076e+13     520.96             0
buildingclasscategory   4.3833e+10       1    4.3833e+10    0.81334        0.36715
landsquarefeet          1.3039e+13       1    1.3039e+13     241.95      3.714e-54
grosssquarefeet         7.8189e+13       1    7.8189e+13     1450.8    8.0393e-304
yearbuilt               2.6108e+11       1    2.6108e+11     4.8444        0.02775
MM                      2.7021e+12       1    2.7021e+12     50.138     1.4936e-12
DD                      1.6867e+11       1    1.6867e+11     3.1297       0.076896
Error                    8.535e+14   15837    5.3893e+10
```

The *p*-values for the indicator variables `borough_Brooklyn` and `borough_Queens` are large, but the *p*-value of the `borough` variable as a group of four indicator variables is almost zero, which indicates that the `borough` variable is statistically significant.

The *p*-values of `buildingclasscategory` and `DD` are larger than 0.05, which indicates that these variables are not significant at the 5% significance level. Therefore, you can consider removing these variables.

You can also use `coeffCI`, `coeefTest`, and `dwTest` to further evaluate the fitted model.

- `coefCI` returns confidence intervals of the coefficient estimates.
- `coefTest` performs a linear hypothesis test on the model coefficients.
- `dwtest` performs the Durbin-Watson test. (This test is used for time series data, so `dwtest` is not appropriate for the housing data in this example.)

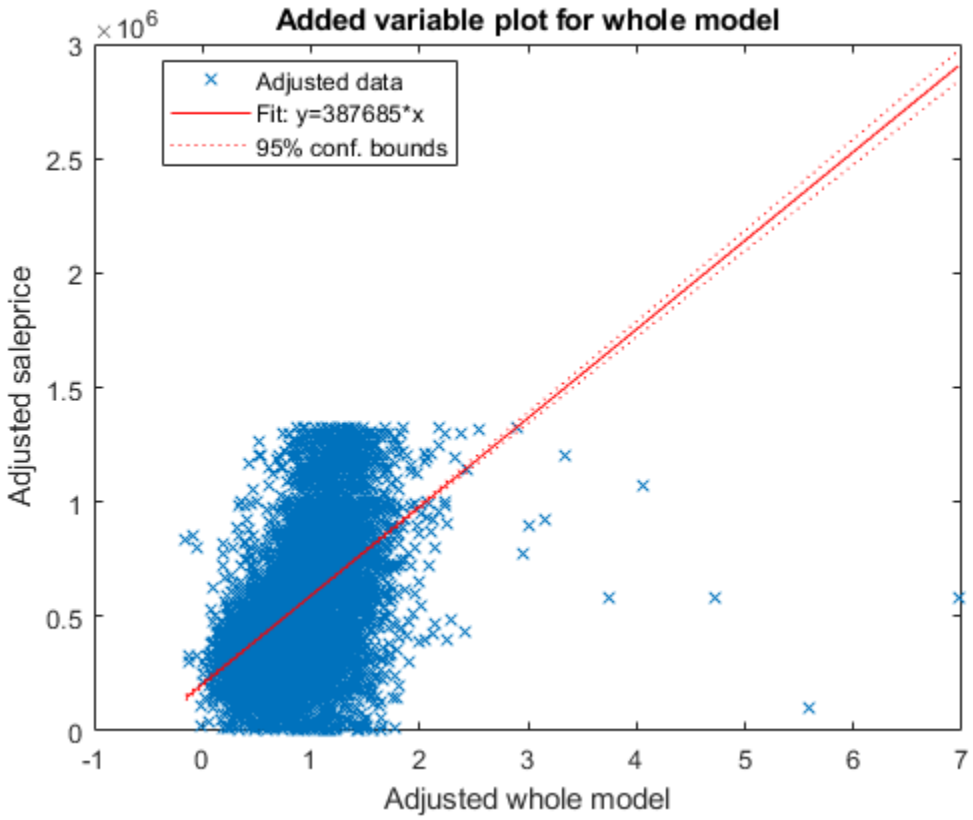
Visualize Model and Summary Statistics

A `LinearModel` object provides multiple plotting functions.

- When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
- When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
- After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to examine the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

In addition, `plot` creates an added variable plot for the whole model, except the intercept term, if `mdl` includes multiple predictor variables.

```
plot(mdl)
```



This plot is equivalent to `plotAdded(mdl)`. The fitted line represents how the model, as a group of variables, can explain the response variable. The slope of the fitted line is not close to zero, and the confidence bound does not include a horizontal line, indicating that the model fits better than a degenerate model consisting of only a constant term. The test statistic value shown in the model display (`F-statistic vs. constant model`) also indicates that the model fits better than the degenerate model.

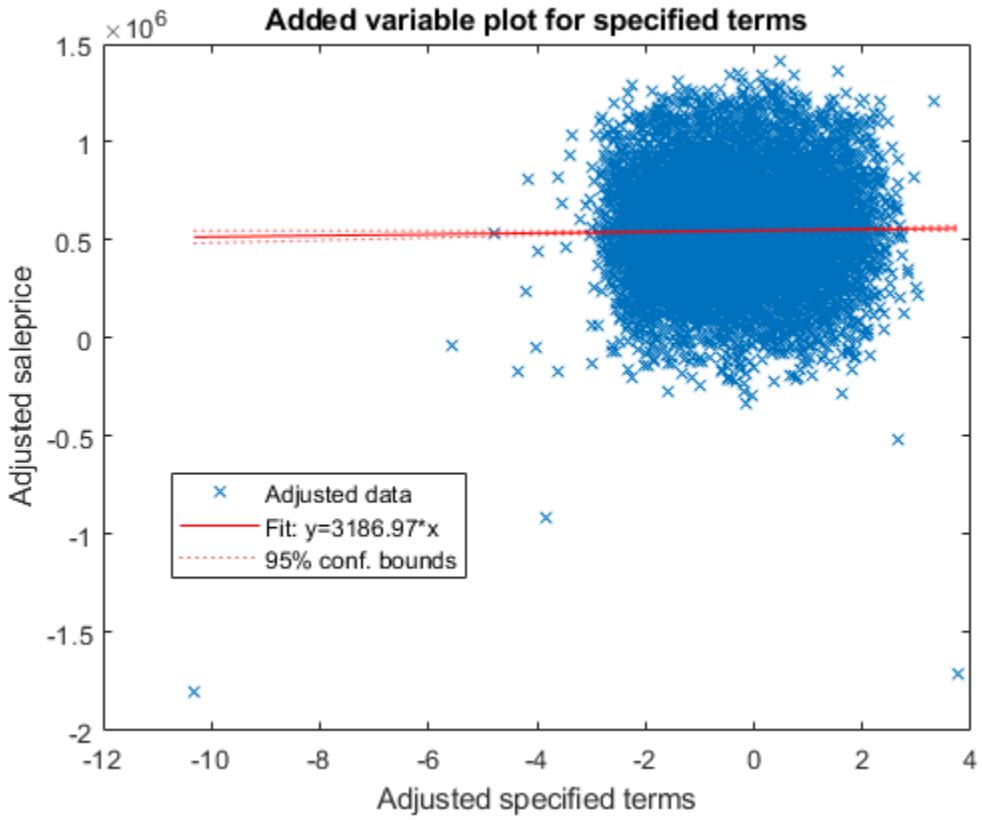
Create an added variable plot for the insignificant variables `buildingclasscategory` and `DD`. The *p*-values of these variables are larger than 0.05. First, find the indices of these coefficients in `mdl.CoefficientNames`.

```
mdl.CoefficientNames
```

```
ans = 1×11 cell
    {'(Intercept)'}    {'borough_Bronx'}    {'borough_Brooklyn'}    {'borough_Queens'}    {'borou
```

`buildingclasscategory` and `DD` are the 6th and 11th coefficients, respectively. Create an added plot for these two variables.

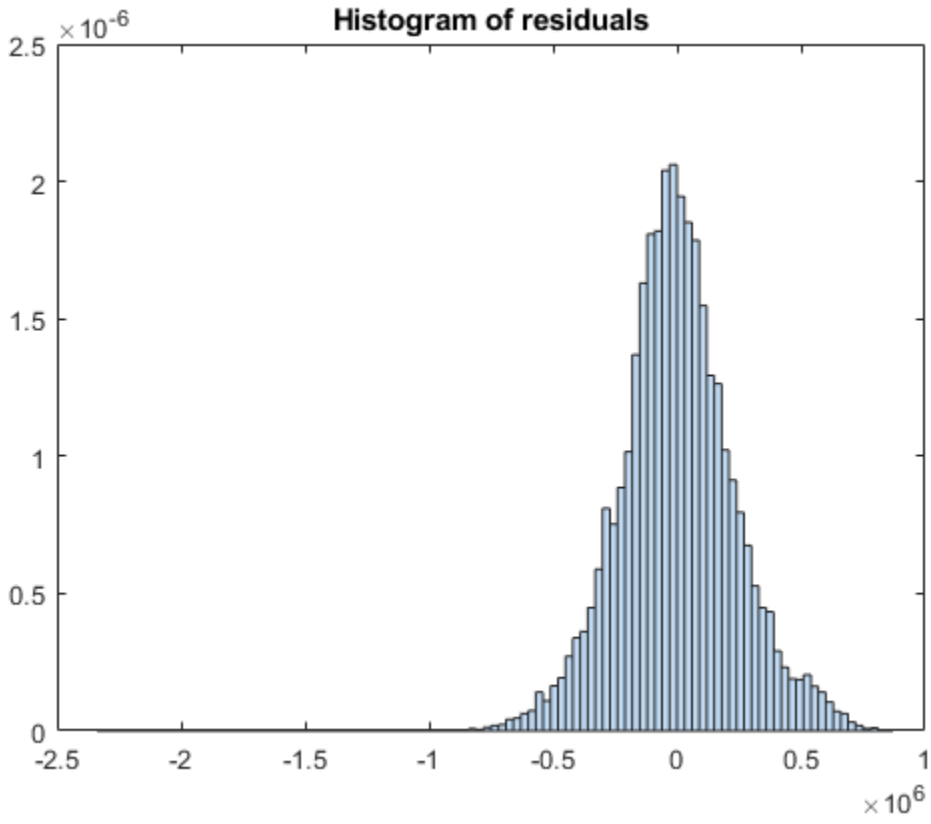
```
plotAdded(mdl,[6,11])
```



The slope of the fitted line is close to zero, indicating that the information from the two variables does not explain the part of the response values not explained by the other predictors. For more details about an added variable plot, see “Added Variable Plot” on page 33-4579.

Create a histogram of the model residuals. `plotResiduals` plots a histogram of the raw residuals using probability density function scaling.

```
plotResiduals(mdl)
```



The histogram shows that a few residuals are smaller than -1×10^6. Identify these outliers.

```
find(mdl.Residuals.Raw < -1*10^6)
```

```
ans = 4×1

        1327
        4136
        4997
       13894
```

Alternatively, you can find the outliers by using `isoutlier`. Specify the `'grubbs'` option to apply Grubb's test. This option is suitable for a normally distributed data set.

```
find(isoutlier(mdl.Residuals.Raw,'grubbs'))
```

```
ans = 3×1

        1327
        4136
        4997
```

The `isoutlier` function does not identify residual `13894` as an outlier. This residual is close to -1×10^6. Display the residual value.

```
mdl.Residuals.Raw(13894)
```

```
ans = -1.0720e+06
```

You can exclude outliers when fitting a linear regression model by using the "'Exclude'" on page 33-0 name-value pair argument. In this case, the example adjusts the fitted model and checks whether the improved model can also explain the outliers.

Adjust Model

Remove the `DD` and `buildingclasscategory` variables using `removeTerms`.

```
newMdl1 = removeTerms(mdl,"DD + buildingclasscategory")
```

```
newMdl1 =
Linear regression model:
    saleprice ~ 1 + borough + landsquarefeet + grosssquarefeet + yearbuilt + MM

Estimated Coefficients:
                            Estimate          SE         tStat        pValue
                          ____________    ____________    ________    __________

    (Intercept)              2.0529e+05    1.0274e+05      1.9981      0.045726
    borough_Bronx           -3.0038e+05         56675        -5.3    1.1739e-07
    borough_Brooklyn             -39704         56488    -0.70286       0.48215
    borough_Queens               -90231         56537      -1.596       0.11052
    borough_Staten Island   -2.2149e+05         56720     -3.9049    9.4652e-05
    landsquarefeet                13.04       0.83912       15.54    4.6278e-54
    grosssquarefeet              113.85        2.5078      45.396             0
    yearbuilt                    96.649        45.395      2.1291      0.033265
    MM                           3875.6        543.49       7.131    1.0396e-12


Number of observations: 15848, Error degrees of freedom: 15839
Root Mean Squared Error: 2.32e+05
R-squared: 0.235,  Adjusted R-Squared: 0.235
F-statistic vs. constant model: 608, p-value = 0
```

Because the two variables are not significant in explaining the response variable, the R-squared and adjusted R-squared values of `newMdl1` are close to the values of `mdl`.

Improve the model by adding or removing variables using `step`. The default upper bound of the model is a model containing an intercept term, the linear term for each predictor, and all products of pairs of distinct predictors (no squared terms), and the default lower bound is a model containing an intercept term. Specify the maximum number of steps to take as 30. The function stops when no single step improves the model.

```
newMdl2 = step(newMdl1,'NSteps',30)
```

```
1. Adding borough:grosssquarefeet, FStat = 58.7413, pValue = 2.63078e-49
2. Adding borough:yearbuilt, FStat = 31.5067, pValue = 3.50645e-26
3. Adding borough:landsquarefeet, FStat = 29.5473, pValue = 1.60885e-24
4. Adding grosssquarefeet:yearbuilt, FStat = 69.312, pValue = 9.08599e-17
5. Adding landsquarefeet:grosssquarefeet, FStat = 33.2929, pValue = 8.07535e-09
6. Adding landsquarefeet:yearbuilt, FStat = 45.2756, pValue = 1.7704e-11
7. Adding yearbuilt:MM, FStat = 18.0785, pValue = 2.13196e-05
8. Adding residentialunits, FStat = 16.0491, pValue = 6.20026e-05
9. Adding residentialunits:landsquarefeet, FStat = 160.2601, pValue = 1.49309e-36
10. Adding residentialunits:grosssquarefeet, FStat = 27.351, pValue = 1.71835e-07
11. Adding commercialunits, FStat = 14.1503, pValue = 0.000169381
```

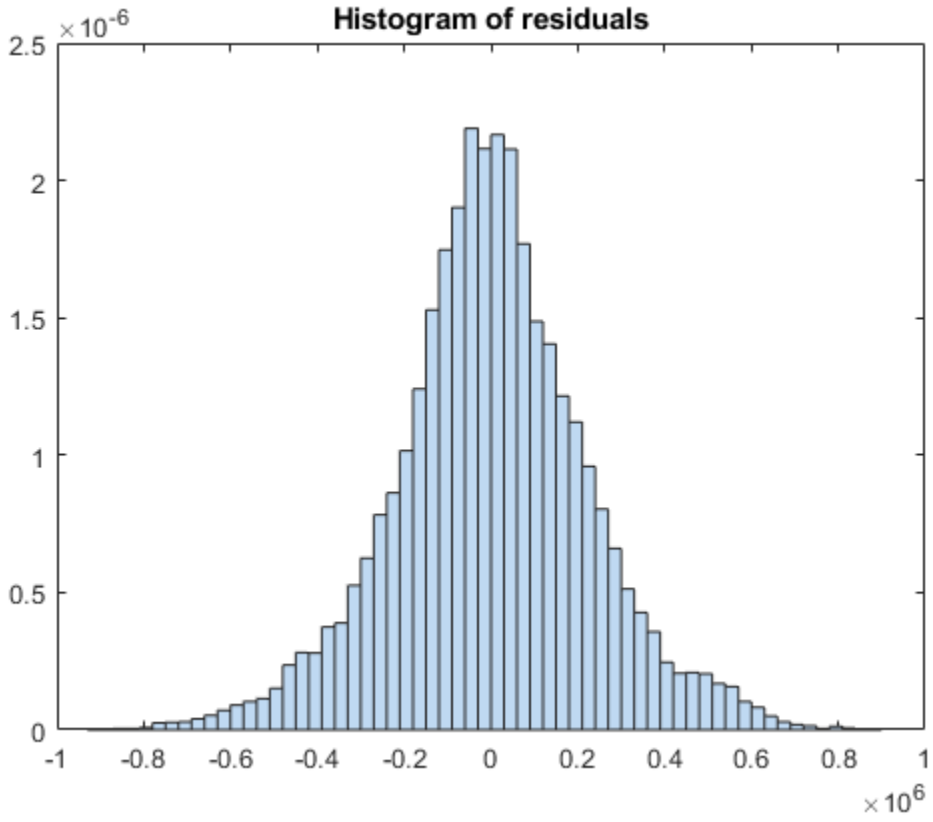
```
12. Adding commercialunits:grosssquarefeet, FStat = 25.6942, pValue = 4.04549e-07
13. Adding borough:commercialunits, FStat = 6.1327, pValue = 6.3015e-05
14. Adding buildingclasscategory, FStat = 11.1412, pValue = 0.00084624
15. Adding buildingclasscategory:landsquarefeet, FStat = 66.9205, pValue = 3.04003e-16
16. Adding buildingclasscategory:yearbuilt, FStat = 15.0776, pValue = 0.0001036
17. Adding buildingclasscategory:grosssquarefeet, FStat = 18.3304, pValue = 1.86812e-05
18. Adding residentialunits:yearbuilt, FStat = 15.0732, pValue = 0.00010384
19. Adding buildingclasscategory:residentialunits, FStat = 13.5644, pValue = 0.00023129
20. Adding borough:buildingclasscategory, FStat = 2.8214, pValue = 0.023567
21. Adding landsquarefeet:MM, FStat = 4.9185, pValue = 0.026585
22. Removing grosssquarefeet:yearbuilt, FStat = 1.6052, pValue = 0.20519

newMdl2 =
Linear regression model:
    saleprice ~ 1 + borough*buildingclasscategory + borough*commercialunits + borough*landsquaref

Estimated Coefficients:
                                                   Estimate          SE         tStat        pVa
                                                  ___________    ___________    ________    _____

    (Intercept)                                    2.2152e+07     1.318e+07      1.6808      0.0
    borough_Bronx                                 -2.3263e+07    1.3176e+07     -1.7656      0.0
    borough_Brooklyn                              -1.8935e+07    1.3174e+07     -1.4373        0
    borough_Queens                                -2.1757e+07    1.3173e+07     -1.6516      0.0
    borough_Staten Island                         -2.3471e+07    1.3177e+07     -1.7813      0.0
    buildingclasscategory                         -7.2403e+05    1.9374e+05      -3.737    0.000
    residentialunits                               6.1912e+05    1.2399e+05      4.9932     6.00
    commercialunits                                4.2016e+05    1.2815e+05      3.2786     0.00
    landsquarefeet                                    -390.54        96.349     -4.0535    5.070
    grosssquarefeet                                    189.33        83.723      2.2614      0.0
    yearbuilt                                          -11556        6958.7     -1.6606      0.0
    MM                                                  95189         31787      2.9946     0.00
    borough_Bronx:buildingclasscategory           -1.1972e+05    1.0481e+05     -1.1422        0
    borough_Brooklyn:buildingclasscategory        -1.4154e+05    1.0448e+05     -1.3548        0
    borough_Queens:buildingclasscategory          -1.1597e+05    1.0454e+05     -1.1093        0
    borough_Staten Island:buildingclasscategory   -1.1851e+05    1.0513e+05     -1.1273        0
    borough_Bronx:commercialunits                 -2.7488e+05    1.3267e+05     -2.0719      0.0
    borough_Brooklyn:commercialunits              -3.8228e+05    1.2835e+05     -2.9784     0.00
    borough_Queens:commercialunits                -3.9818e+05    1.2884e+05     -3.0906     0.00
    borough_Staten Island:commercialunits         -4.9381e+05     1.353e+05     -3.6496    0.000
    borough_Bronx:landsquarefeet                       121.81        77.442       1.573        0
    borough_Brooklyn:landsquarefeet                    113.09        77.413      1.4609        0
    borough_Queens:landsquarefeet                      99.894        77.374      1.2911        0
    borough_Staten Island:landsquarefeet               84.508        77.376      1.0922        0
    borough_Bronx:grosssquarefeet                     -55.417        83.412    -0.66437        0
    borough_Brooklyn:grosssquarefeet                   6.4033        83.031    0.077119        0
    borough_Queens:grosssquarefeet                      38.28        83.144     0.46041        0
    borough_Staten Island:grosssquarefeet               12.539        83.459     0.15024        0
    borough_Bronx:yearbuilt                             12121        6956.8      1.7422      0.0
    borough_Brooklyn:yearbuilt                         9986.5        6955.8      1.4357        0
    borough_Queens:yearbuilt                            11382        6955.3      1.6364        0
    borough_Staten Island:yearbuilt                     12237        6957.1      1.7589      0.0
    buildingclasscategory:residentialunits              21392          5465      3.9143    9.104
    buildingclasscategory:landsquarefeet              -13.099        2.0014      -6.545    6.134
    buildingclasscategory:grosssquarefeet             -30.087        5.2786     -5.6998    1.220
    buildingclasscategory:yearbuilt                    462.31        85.912      5.3813    7.502
    residentialunits:landsquarefeet                   -1.0826       0.13896     -7.7911    7.055
    residentialunits:grosssquarefeet                  -5.1192        1.7923     -2.8563     0.00
```

```
    residentialunits:yearbuilt                  -326.69        63.556     -5.1403    2.776
    commercialunits:grosssquarefeet             -29.839        5.0231     -5.9403    2.904
    landsquarefeet:grosssquarefeet           -0.0055199     0.0010364     -5.3262    1.016
    landsquarefeet:yearbuilt                     0.1766      0.030902      5.7151    1.116
    landsquarefeet:MM                            0.6595       0.30229      2.1817      0.0
    yearbuilt:MM                                 -47.944        16.392     -2.9248     0.00

Number of observations: 15848, Error degrees of freedom: 15804
Root Mean Squared Error: 2.25e+05
R-squared: 0.285,  Adjusted R-Squared: 0.283
F-statistic vs. constant model: 146, p-value = 0
```

The R-squared and adjusted R-squared values of `newMdl2` are larger than the values of `newMdl1`.

Create a histogram of the model residuals by using `plotResiduals`.

```
plotResiduals(newMdl2)
```



The residual histogram of `newMdl2` is symmetric, without outliers.

You can also use `addTerms` to add specific terms. Alternatively, you can use `stepwiselm` to specify terms in a starting model and continue improving the model by using stepwise regression.

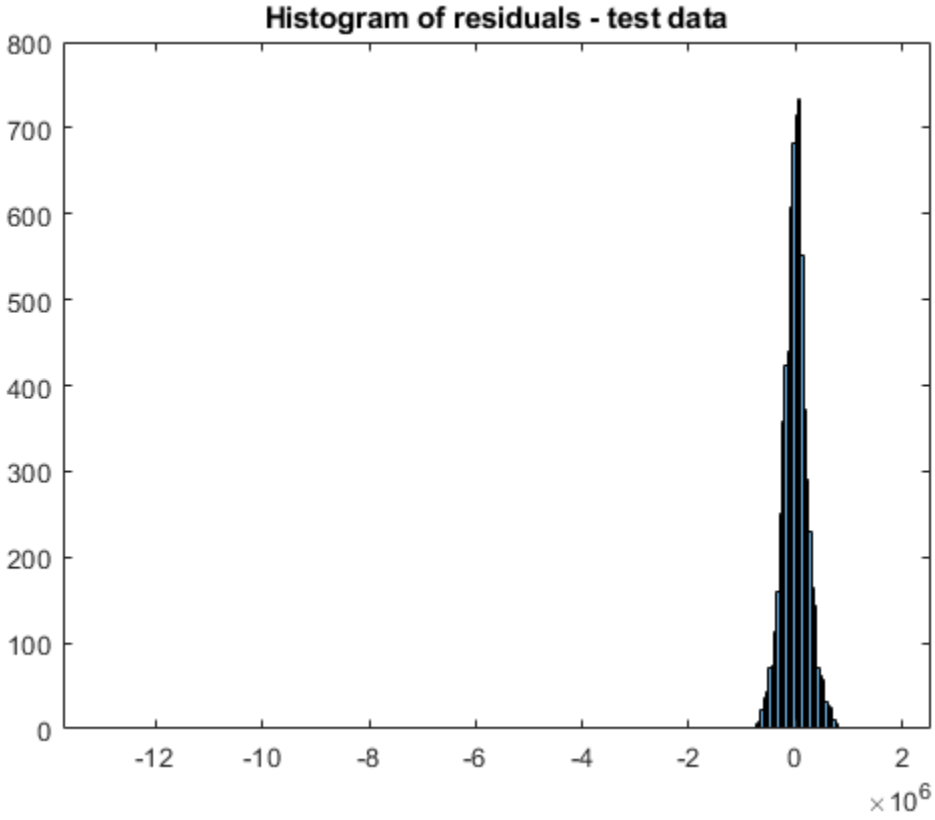
Predict Responses to New Data

Predict responses to the test data set `testData` by using the fitted model `newMdl2` and the object function `predict` to

```
ypred = predict(newMdl2,testData);
```

Plot the residual histogram of the test data set.

```
errs = ypred - testData.saleprice;
histogram(errs)
title("Histogram of residuals - test data")
```



The residual values have a few outliers.

```
errs(isoutlier(errs,'grubbs'))
```

```
ans = 6×1
10^7 ×

    0.1788
   -0.4688
   -1.2981
    0.1019
    0.1122
    0.1331
```

Analyze Using Tall Arrays

The `fitlm` function supports tall arrays for out-of-memory data, with some limitations. For tall data, `fitlm` returns a `CompactLinearModel` object that contains most of the same properties as a `LinearModel` object. The main difference is that the compact object is sensitive to memory

requirements. The compact object does not have properties that include the data, or that include an array of the same size as the data. Therefore, some `LinearModel` object functions that require data do not work with a compact model. See “Object Functions” on page 33-823 for the list of supported object functions. Also, see “Tall Arrays” on page 33-1998 for the usage notes and limitations of `fitlm` for tall arrays.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Assume that all the data in the datastore `ds` does not fit in memory. You can use `tall` instead of `readall` to read `ds`.

```
NYCHousing2015 = tall(ds);
```

For this example, convert the in-memory table `NYCHousing2015` to a tall table by using the `tall` function.

```
NYCHousing2015_t = tall(NYCHousing2015);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Partition the data set into a training set and test set. When you use `cvpartition` with tall arrays, the function partitions the data set based on the variable supplied as the first input argument. For classification problems, you typically use the response variable (a grouping variable) and create a random stratified partition to get even distribution between training and test sets for all groups. For regression problems, this stratification is not adequate, and you can use the `'Stratify'` name-value pair argument to turn off the option.

In this example, specify the predictor variable `NYCHousing2015_t.borough` as the first input argument to make the distribution of boroughs roughly the same across the training and tests sets. For reproducibility, set the seed of the random number generator using `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
tallrng('default') % For reproducibility
c = cvpartition(NYCHousing2015_t.borough,"holdout",0.3);
trainData_t = NYCHousing2015_t(training(c),:);
testData_t = NYCHousing2015_t(test(c),:);
```

Because `fitlm` returns a compact model object for tall arrays, you cannot improve the model using the `step` function. Instead, you can explore the model parameters by using the object functions and then adjust the model as needed. You can also gather a subset of the data into the workspace, use `stepwiselm` to iteratively develop the model in memory, and then scale up to use tall arrays. For details, see Model Development of “Statistics and Machine Learning with Big Data Using Tall Arrays” on page 30-24.

In this example, fit a linear regression model using the model formula of `newMdl2`.

```
mdl_t = fitlm(trainData_t,newMdl2.Formula)
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 7.4 sec
Evaluation completed in 9.2 sec
```

```
mdl_t =
Compact linear regression model:
    saleprice ~ 1 + borough*buildingclasscategory + borough*commercialunits + borough*landsquaref

Estimated Coefficients:
                                                Estimate          SE         tStat        pVa
                                               ___________    ___________    ________    _____

    (Intercept)                                -1.3301e+06    5.1815e+05      -2.567      0.0
    borough_Brooklyn                            4.2583e+06    4.1808e+05      10.185    2.739
    borough_Manhattan                           2.2758e+07    1.3448e+07      1.6923      0.0
    borough_Queens                              1.1395e+06    4.1868e+05      2.7216     0.00
    borough_Staten Island                      -1.1196e+05    4.6677e+05     -0.23986        0
    buildingclasscategory                       -8.08e+05    1.6219e+05     -4.9817    6.370
    residentialunits                            6.0588e+05    1.2669e+05      4.7822    1.749
    commercialunits                                  80197         53311      1.5043        0
    landsquarefeet                                -279.94        53.913     -5.1925    2.100
    grosssquarefeet                                170.02        13.996      12.147    8.383
    yearbuilt                                      683.49        268.34      2.5471      0.0
    MM                                             86488         32725      2.6428     0.00
    borough_Brooklyn:buildingclasscategory        -9852.4         12048     -0.81773        0
    borough_Manhattan:buildingclasscategory     1.3318e+05    1.3592e+05     0.97988        0
    borough_Queens:buildingclasscategory             15621         11671      1.3385        0
    borough_Staten Island:buildingclasscategory      15132         14893       1.016        0
    borough_Brooklyn:commercialunits               -22060         43012     -0.51289        0
    borough_Manhattan:commercialunits           4.8349e+05    2.1757e+05      2.2222      0.0
    borough_Queens:commercialunits                 -42023         44736     -0.93936        0
    borough_Staten Island:commercialunits      -1.3382e+05         56976     -2.3487      0.0
    borough_Brooklyn:landsquarefeet                 9.8263        5.2513      1.8712      0.0
    borough_Manhattan:landsquarefeet               -78.962        78.445     -1.0066        0
    borough_Queens:landsquarefeet                  -3.0855        3.9087     -0.78939        0
    borough_Staten Island:landsquarefeet           -17.325        3.5831     -4.8351    1.343
    borough_Brooklyn:grosssquarefeet                37.689        10.573      3.5646    0.000
    borough_Manhattan:grosssquarefeet               16.107        82.074     0.19625        0
    borough_Queens:grosssquarefeet                  70.381         10.69      6.5837    4.734
    borough_Staten Island:grosssquarefeet           36.396         12.08      3.0129     0.00
    borough_Brooklyn:yearbuilt                     -2110.1        216.32     -9.7546    2.038
    borough_Manhattan:yearbuilt                     -11884        7023.9      -1.692      0.0
    borough_Queens:yearbuilt                       -566.44        216.89     -2.6116     0.00
    borough_Staten Island:yearbuilt                 53.714        239.89     0.22391        0
    buildingclasscategory:residentialunits           24088          5574      4.3215    1.559
    buildingclasscategory:landsquarefeet            5.7964        5.8438      0.9919        0
    buildingclasscategory:grosssquarefeet          -47.079        5.2884     -8.9023    6.055
    buildingclasscategory:yearbuilt                 430.97        83.593      5.1555      2.5
    residentialunits:landsquarefeet                -21.756        5.6485     -3.8517    0.000
    residentialunits:grosssquarefeet                 4.584        1.4586      3.1427     0.00
    residentialunits:yearbuilt                     -310.09        65.429     -4.7393    2.163
    commercialunits:grosssquarefeet                -27.839        11.463     -2.4286      0.0
    landsquarefeet:grosssquarefeet              -0.0068613    0.00094607     -7.2524    4.283
    landsquarefeet:yearbuilt                       0.17489      0.028195      6.2028    5.686
    landsquarefeet:MM                              0.70295        0.2848      2.4682      0.0
    yearbuilt:MM                                   -43.405        16.871     -2.5728      0.0


Number of observations: 15849, Error degrees of freedom: 15805
Root Mean Squared Error: 2.26e+05
R-squared: 0.277,  Adjusted R-Squared: 0.275
F-statistic vs. constant model: 141, p-value = 0
```

`mdl_t` is a `CompactLinearModel` object. `mdl_t` is not exactly the same as `newMdl2` because the partitioned training data set obtained from the tall table is not the same as the one from the in-memory data set.

You cannot use the `plotResiduals` function to create a histogram of the model residuals because `mdl_t` is a compact object. Instead, compute the residuals directly from the compact object and create the histogram using `histogram`.

```
mdl_t_Residual = trainData_t.saleprice - predict(mdl_t,trainData_t);
histogram(mdl_t_Residual)
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 2.5 sec
- Pass 2 of 2: Completed in 0.63 sec
Evaluation completed in 3.8 sec
```

```
title("Histogram of residuals - train data")
```

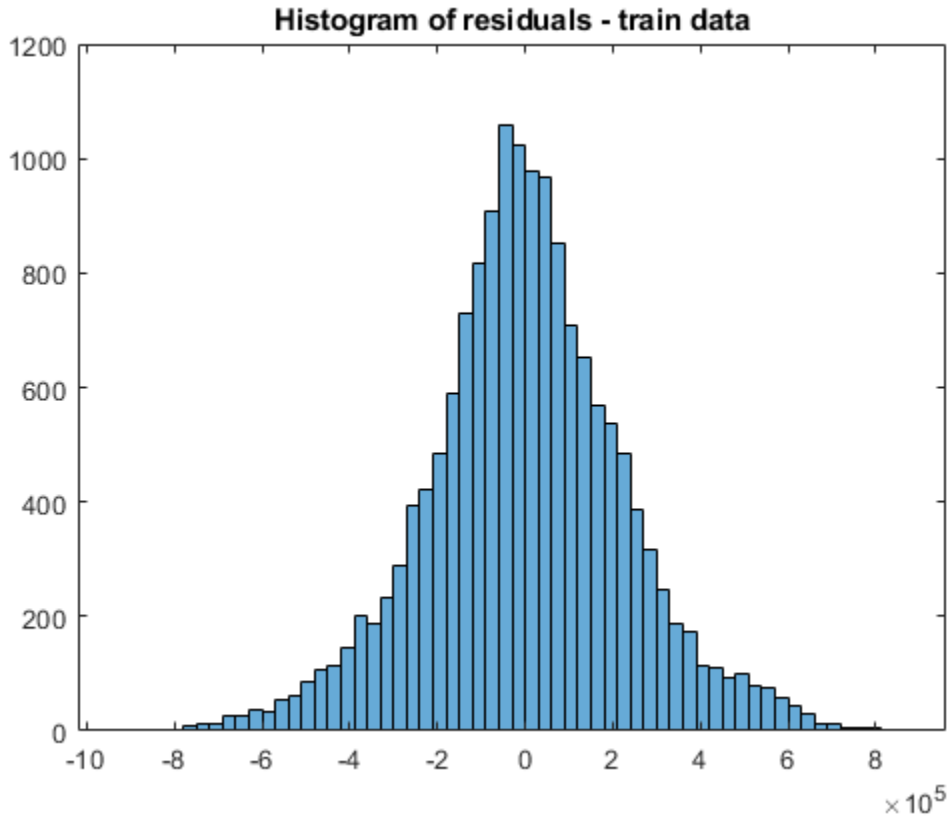
Predict responses to the test data set `testData_t` by using `predict`.

```
ypred_t = predict(mdl_t,testData_t);
```

Plot the residual histogram of the test data set.

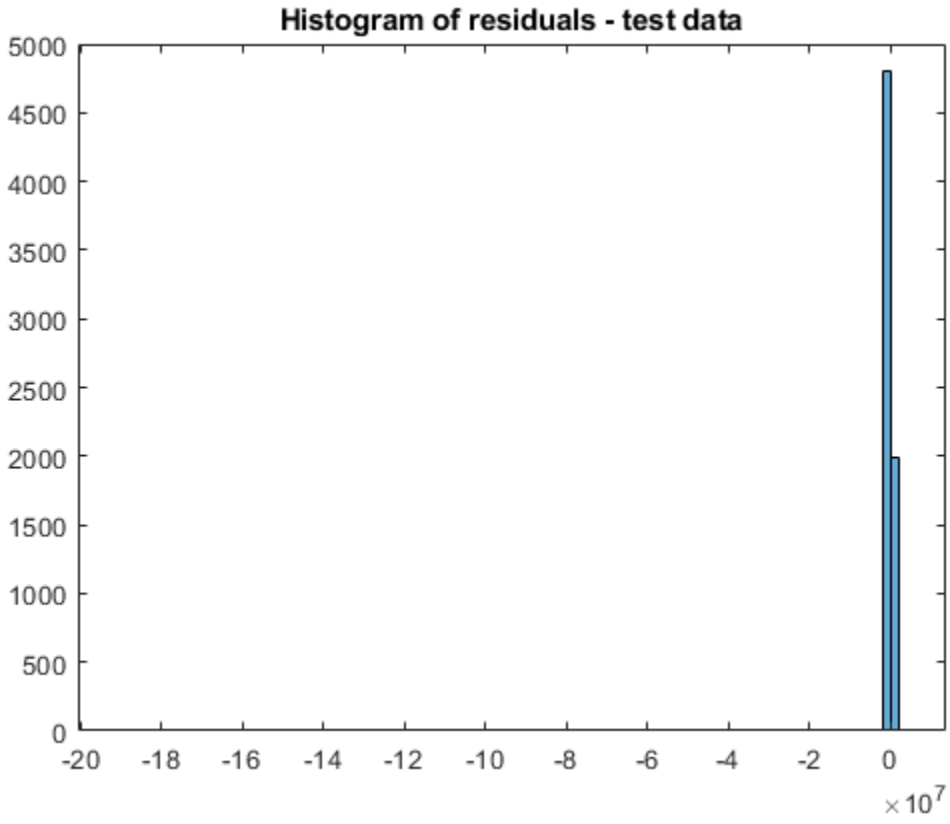
```
errs_t = ypred_t - testData_t.saleprice;
histogram(errs_t)
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: 0% complete
Evaluation 0% complete

- Pass 1 of 2: 6% complete
Evaluation 3% complete
```




```
- Pass 1 of 2: Completed in 0.79 sec
- Pass 2 of 2: Completed in 0.55 sec
Evaluation completed in 2 sec
```

```
title("Histogram of residuals - test data")
```



You can further assess the fitted model using the `CompactLinearModel` object functions. For an example, see Assess and Adjust Model of “Statistics and Machine Learning with Big Data Using Tall Arrays” on page 30-24.

See Also

`CompactLinearModel` | `LinearModel` | `fitlm` | `isoutlier`

More About

- “Linear Regression” on page 11-9
- “Linear Regression Workflow” on page 11-35
- “Interpret Linear Regression Results” on page 11-50
- “Linear Regression with Interaction Effects” on page 11-44
- “Linear Regression with Categorical Covariates” on page 2-52
- “Examine Quality and Adjust Fitted Model” on page 11-14
- “Summary of Output and Diagnostic Statistics” on page 11-89
- “Statistics and Machine Learning with Big Data Using Tall Arrays” on page 30-24

Analyze Time Series Data

This example shows how to visualize and analyze time series data using a `timeseries` object and the `regress` function.

Air Passenger Data

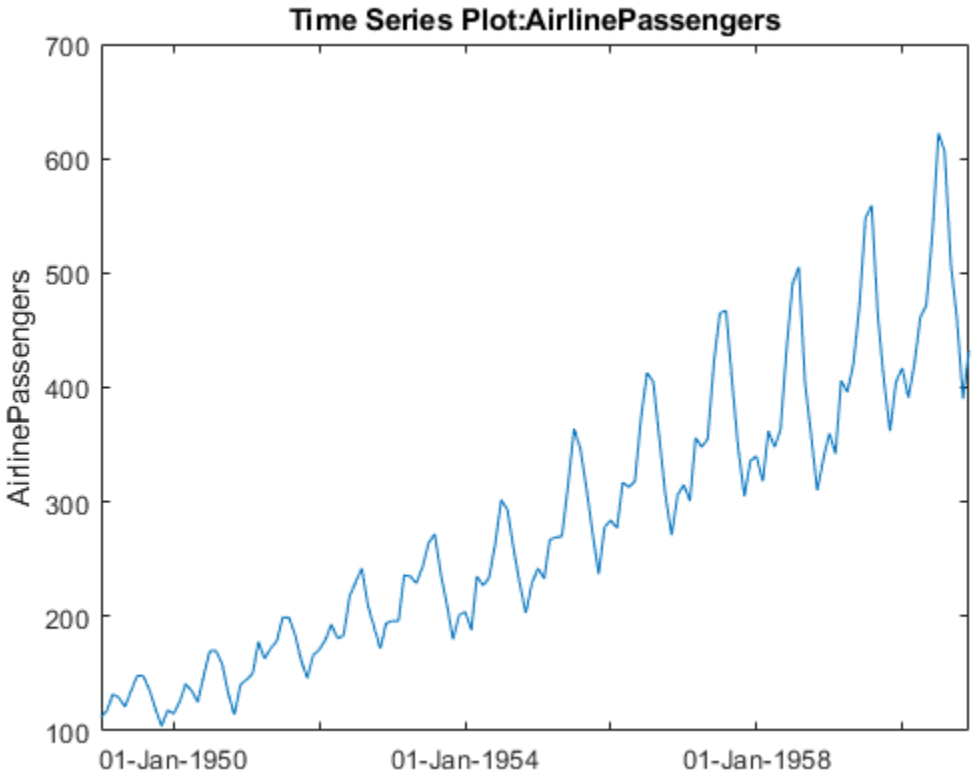
First we create an array of monthly counts of airline passengers, measured in thousands, for the period January 1949 through December 1960.

```
%   1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960
y = [112  115  145  171  196  204  242  284  315  340  360  417    % Jan
     118  126  150  180  196  188  233  277  301  318  342  391    % Feb
     132  141  178  193  236  235  267  317  356  362  406  419    % Mar
     129  135  163  181  235  227  269  313  348  348  396  461    % Apr
     121  125  172  183  229  234  270  318  355  363  420  472    % May
     135  149  178  218  243  264  315  374  422  435  472  535    % Jun
     148  170  199  230  264  302  364  413  465  491  548  622    % Jul
     148  170  199  242  272  293  347  405  467  505  559  606    % Aug
     136  158  184  209  237  259  312  355  404  404  463  508    % Sep
     119  133  162  191  211  229  274  306  347  359  407  461    % Oct
     104  114  146  172  180  203  237  271  305  310  362  390    % Nov
     118  140  166  194  201  229  278  306  336  337  405  432 ]; % Dec
% Source:
% Hyndman, R.J., Time Series Data Library,
% http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/.
% Copied in October, 2005.
```

Create Time Series Object

When we create a time series object, we can keep the time information along with the data values. We have monthly data, so we create an array of dates and use it along with the Y data to create the time series object.

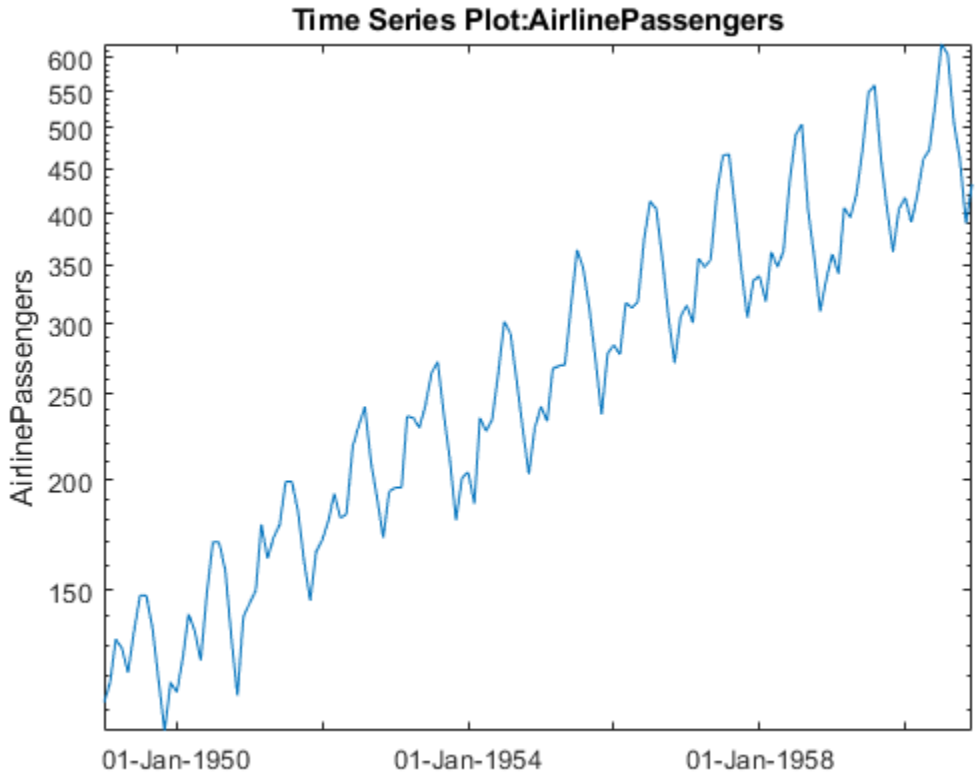
```
yr = repmat((1949:1960),12,1);
mo = repmat((1:12)',1,12);
time = datestr(datenum(yr(:),mo(:),1));
ts = timeseries(y(:),time,'name','AirlinePassengers');
ts.TimeInfo.Format = 'dd-mmm-yyyy';
tscol = tscollection(ts);
plot(ts)
```



Examine Trend and Seasonality

This series seems to have a strong seasonal component, with a trend that may be linear or quadratic. Furthermore, the magnitude of the seasonal variation increases as the general level increases. Perhaps a log transformation would make the seasonal variation be more constant. First we'll change the axis scale.

```
h_gca = gca;
h_gca.YScale = 'log';
```

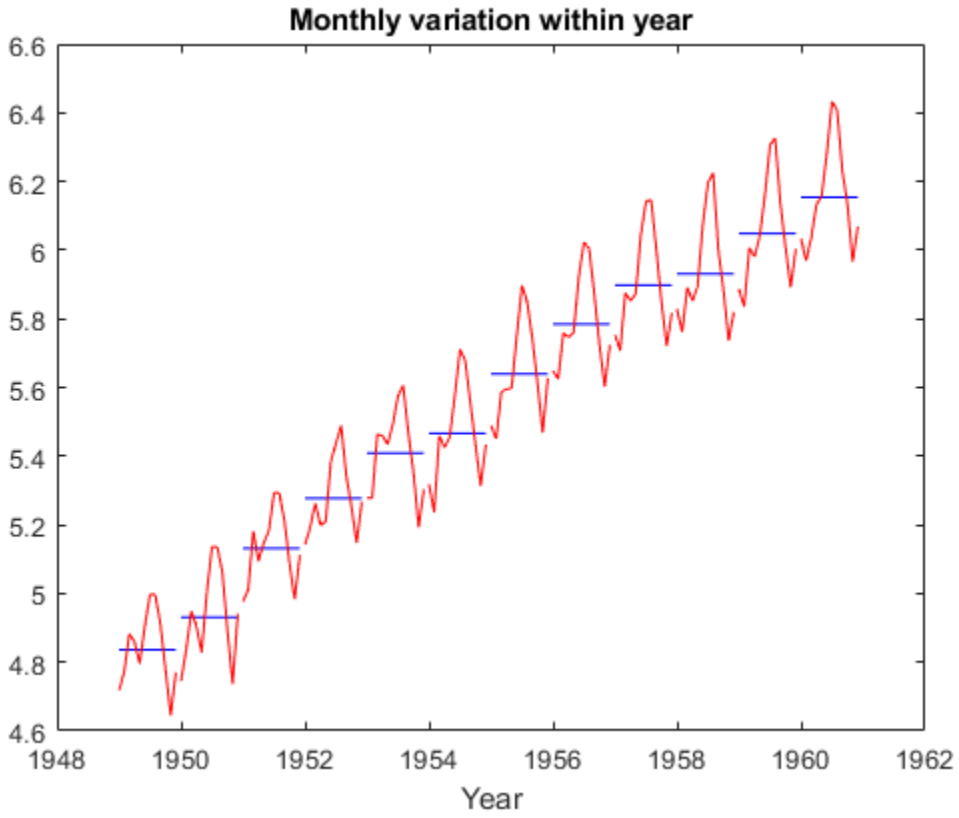


It appears that it would be easier to model the seasonal component on the log scale. We'll create a new time series with a log transformation.

```
tscol = addts(tscol,log(ts.data),'logAirlinePassengers');
logts = tscol.logAirlinePassengers;
```

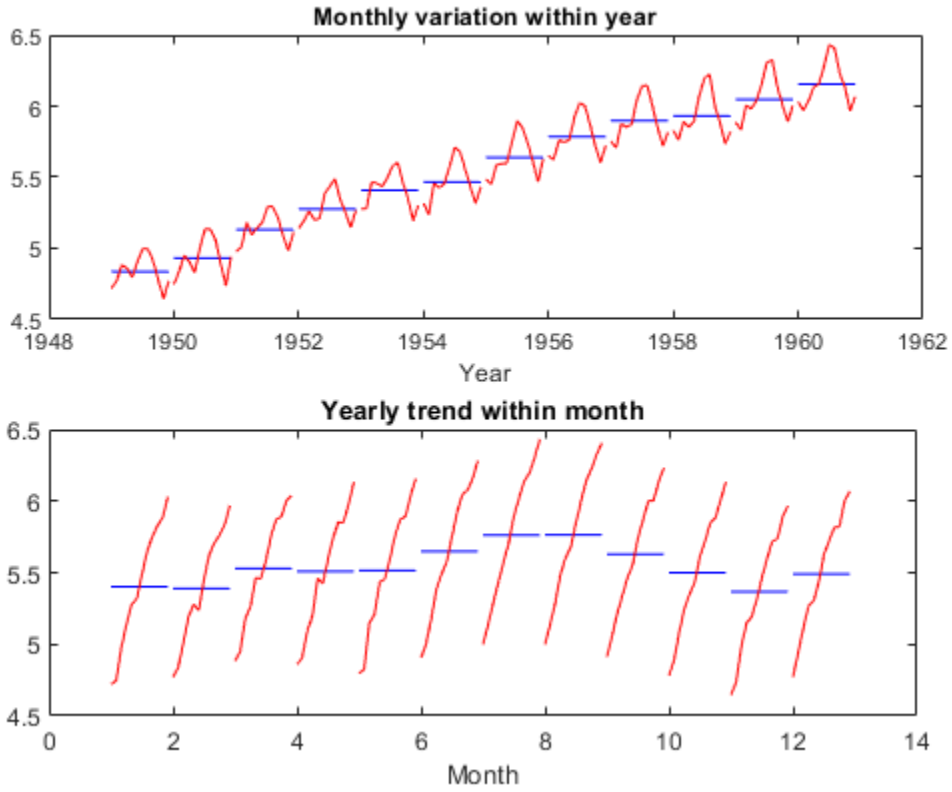
Now let's plot the yearly averages, with monthly deviations superimposed. We want to see if the month-to-month variation within years appears constant. For these manipulations treating the data as a matrix in a month-by-year format, it's more convenient to operate on the original data matrix.

```
t = reshape(datenum(time),12,12);
logy = log(y);
ymean = repmat(mean(logy),12,1);
ydiff = logy - ymean;
x = yr + (mo-1)/12;
plot(x,ymean,'b-',x,ymean+ydiff,'r-')
title('Monthly variation within year')
xlabel('Year')
```



Now let's reverse the years and months, and try to see if the year-to-year trend is constant for each month.

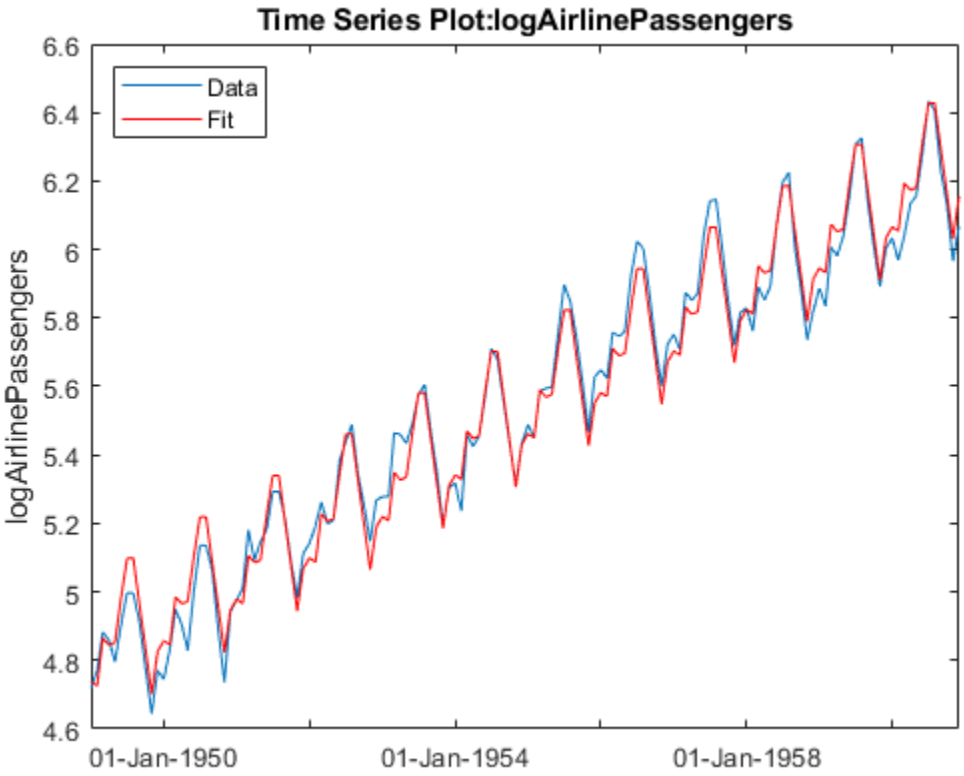
```
h_gca = gca;
h_gca.Position = [0.13 0.58 0.78 0.34];
subplot(2,1,2);
t = reshape(datenum(time),12,12);
mmean = repmat(mean(logy,2),1,12);
mdiff = logy - mmean;
x = mo + (yr-min(yr(:)))/12;
plot(x',mmean','b-',x',(mmean+mdiff)','r-')
title('Yearly trend within month')
xlabel('Month')
```



Model Trend and Seasonality

Let's attempt to model this series as a linear trend plus a seasonal component.

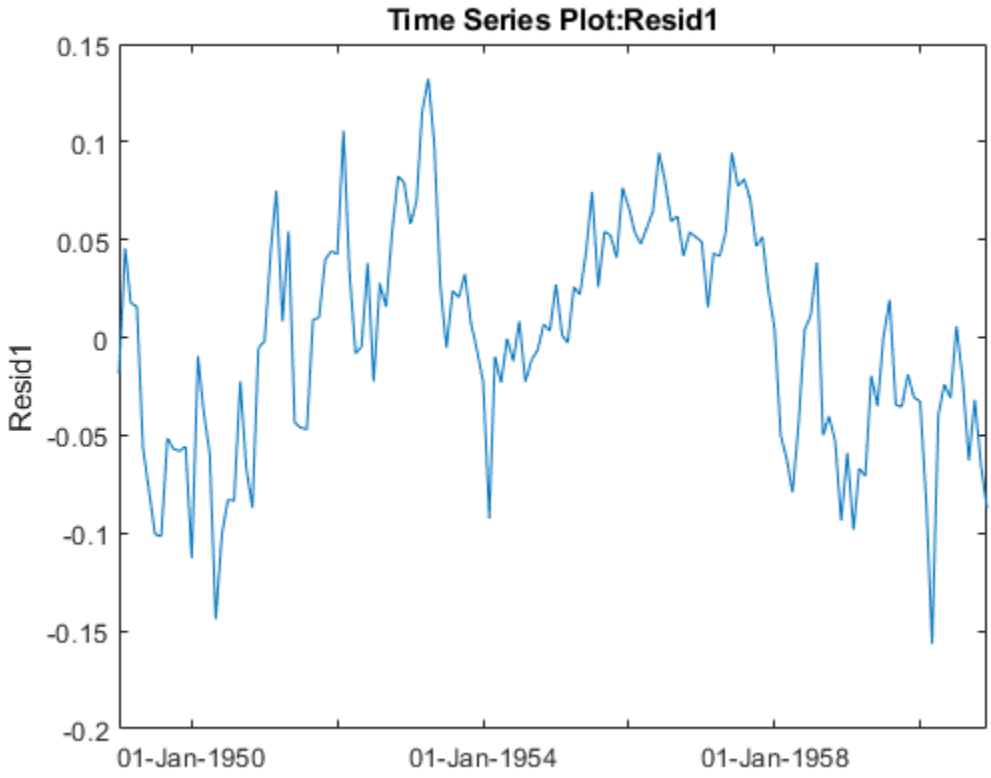
```
subplot(1,1,1);
X = [dummyvar(mo(:)) logts.time];
[b,bint,resid] = regress(logts.data,X);
tscol = addts(tscol,X*b,'Fit1')

Time Series Collection Object: unnamed

Time vector characteristics

      Start date            01-Jan-1949
      End date              01-Dec-1960

Member Time Series Objects:

      AirlinePassengers
      logAirlinePassengers
      Fit1

plot(logts)
hold on
plot(tscol.Fit1,'Color','r')
hold off
legend('Data','Fit','location','NW')
```



Based on this graph, the fit appears to be good. The differences between the actual data and the fitted values may well be small enough for our purposes.

But let's try to investigate this some more. We would like the residuals to look independent. If there is autocorrelation (correlation between adjacent residuals), then there may be an opportunity to model that and make our fit better. Let's create a time series from the residuals and plot it.

```
tscol = addts(tscol,resid,'Resid1');
plot(tscol.Resid1)
```



The residuals do not look independent. In fact, the correlation between adjacent residuals looks quite strong. We can test this formally using a Durbin-Watson test.

```
[p,dw] = dwtest(tscol.Resid1.data,X)
```

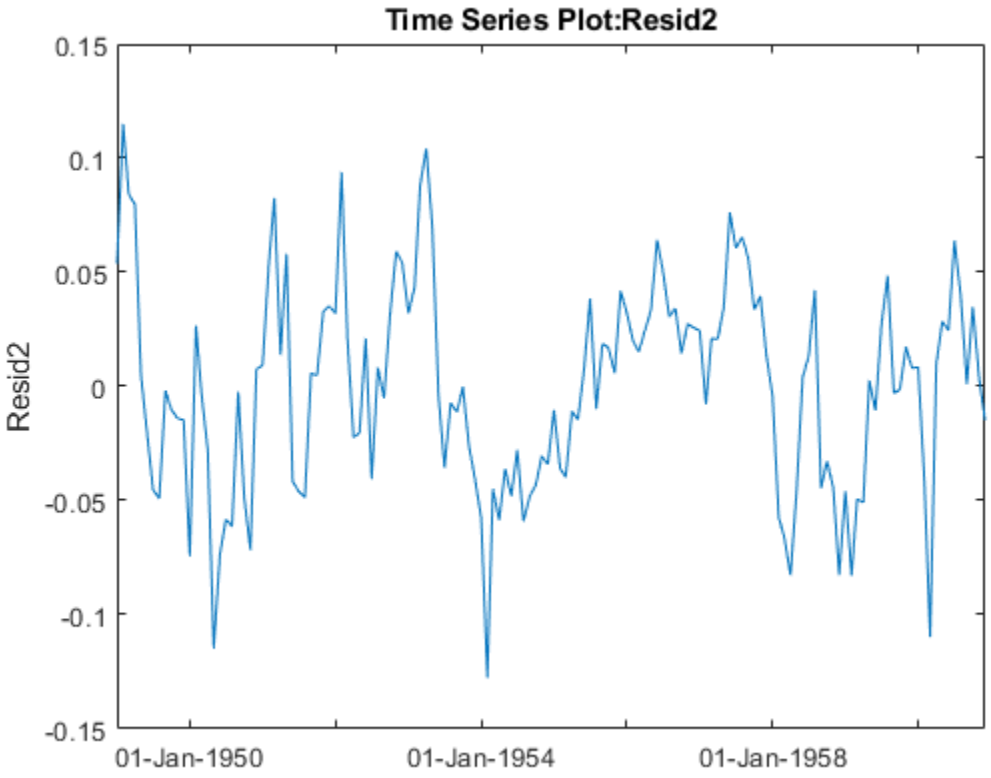
```
p = 7.7787e-30
```

```
dw = 0.4256
```

A low p-value for the Durbin-Watson statistic is an indication that the residuals are correlated across time. A typical cutoff for hypothesis tests is to decide that $p<0.05$ is significant. Here the very small p-value gives strong evidence that the residuals are correlated.

We can attempt to change the model to remove the autocorrelation. The general shape of the curve is high in the middle and low at the ends. This suggests that we should allow for a quadratic trend term. However, it also appears that autocorrelation will remain after we add this term. Let's try it.

```
X = [dummyvar(mo(:)) logts.time logts.time.^2];
[b2,bint,resid2] = regress(logts.data,X);
tscol = addts(tscol,resid2,'Resid2');
plot(tscol.Resid2)
```




```
[p,dw] = dwtest(tscol.Resid2.data,X)
```

```
p = 8.7866e-20
```

```
dw = 0.6487
```

Adding the squared term did remove the pronounced curvature in the original residual plot, but both the plot and the new Durbin-Watson test show that there is still significant correlation in the residuals.

Autocorrelation like this could be the result of other causes that are not captured in our X variable. Perhaps we could collect other data that would help us improve our model and reduce the correlation. In the absence of other data, we might simply add another parameter to the model to represent the autocorrelation. Let's do that, removing the squared term, and using an autoregressive model for the error.

In an autoregressive process, we have two stages:

```
Y(t) = X(t,:)*b + r(t)       % regression model for original data
r(t) = rho * r(t-1) + u(t)   % autoregressive model for residuals
```

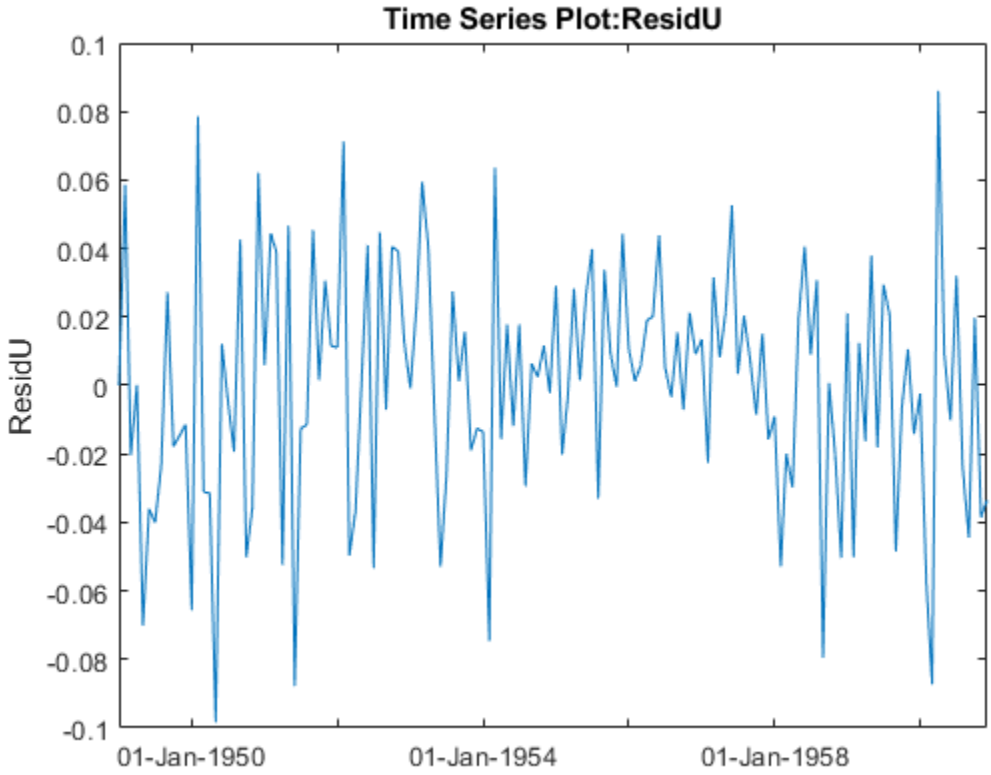
Unlike in the usual regression model when we would like the residual series `r(t)` to be a set of independent values, this model allows the residuals to follow an autoregressive model with its own error term `u(t)` that consists of independent values.

To create this model, we want to write an anonymous function `f` to compute fitted values `Yfit`, so that `Y-Yfit` gives the u values:

```
   Yfit(t) = rho*Y(t-1) + (X(t,:) - rho*X(t-1,:))*b
```

In this anonymous function we combine `[rho; b]` into a single parameter vector `c`. The resulting residuals look much closer to an uncorrelated series.

```
r = corr(resid(1:end-1),resid(2:end));  % initial guess for rho
X = [dummyvar(mo(:)) logts.time];
Y = logts.data;
f = @(c,x) [Y(1); c(1)*Y(1:end-1) + (x(2:end,:)- c(1)*x(1:end-1,:))*c(2:end)];
c = nlinfit(X,Y,f,[r; b]);

u = Y - f(c,X);
tscol = addts(tscol,u,'ResidU');
plot(tscol.ResidU);
```



Summary

This example provides an illustration of how to use the MATLAB® time series object along with features from the Statistics and Machine Learning Toolbox. It is simple to use the `ts.data` notation to extract the data and supply it as input to any function. The `controlchart` function also accepts time series objects directly.

More elaborate analyses are possible by using features specifically designed for time series, such as those in Econometrics Toolbox™ and System Identification Toolbox™.

Partial Least Squares Regression and Principal Components Regression

This example shows how to apply Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR), and discusses the effectiveness of the two methods. PLSR and PCR are both methods to model a response variable when there are a large number of predictor variables, and those predictors are highly correlated or even collinear. Both methods construct new predictor variables, known as components, as linear combinations of the original predictor variables, but they construct those components in different ways. PCR creates components to explain the observed variability in the predictor variables, without considering the response variable at all. On the other hand, PLSR does take the response variable into account, and therefore often leads to models that are able to fit the response variable with fewer components. Whether or not that ultimately translates into a more parsimonious model, in terms of its practical use, depends on the context.

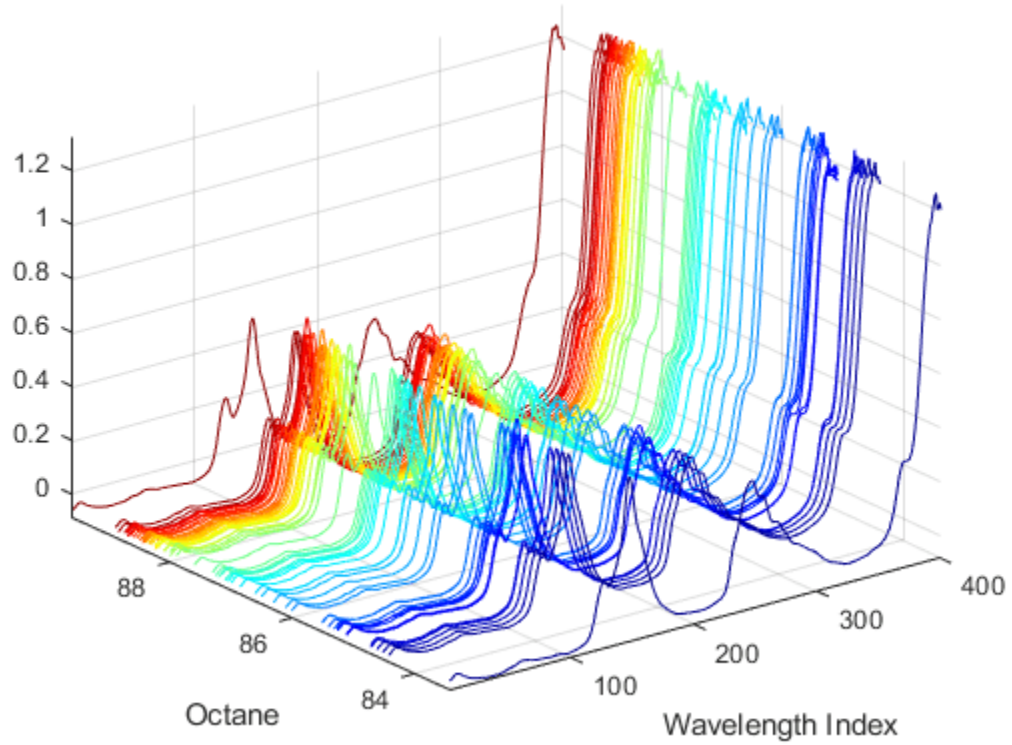
Loading the Data

Load a data set comprising spectral intensities of 60 samples of gasoline at 401 wavelengths, and their octane ratings. These data are described in Kalivas, John H., "Two Data Sets of Near Infrared Spectra," Chemometrics and Intelligent Laboratory Systems, v.37 (1997) pp.255-259.

```
load spectra
whos NIR octane
```

```
  Name         Size              Bytes  Class     Attributes

  NIR         60x401            192480  double
  octane      60x1                 480  double
```

```
[dummy,h] = sort(octane);
oldorder = get(gcf,'DefaultAxesColorOrder');
set(gcf,'DefaultAxesColorOrder',jet(60));
plot3(repmat(1:401,60,1)',repmat(octane(h),1,401)',NIR(h,:)');
set(gcf,'DefaultAxesColorOrder',oldorder);
xlabel('Wavelength Index'); ylabel('Octane'); axis('tight');
grid on
```



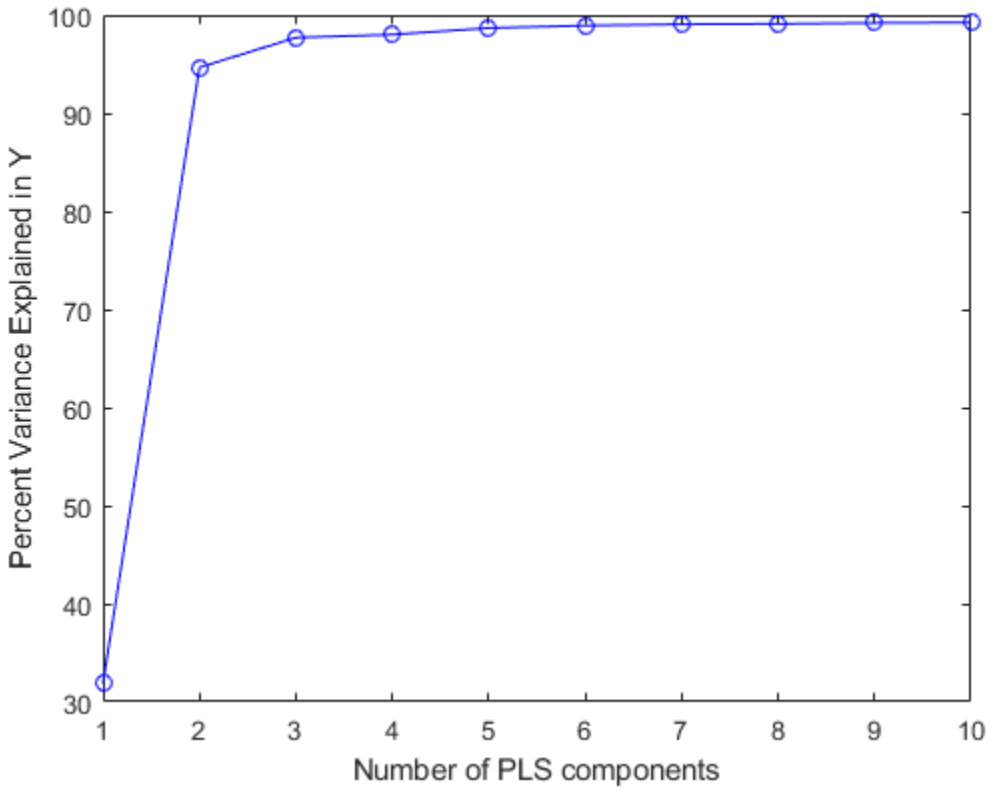
Fitting the Data with Two Components

Use the `plsregress` function to fit a PLSR model with ten PLS components and one response.

```
X = NIR;
y = octane;
[n,p] = size(X);
[Xloadings,Yloadings,Xscores,Yscores,betaPLS10,PLSPctVar] = plsregress(...
    X,y,10);
```

Ten components may be more than will be needed to adequately fit the data, but diagnostics from this fit can be used to make a choice of a simpler model with fewer components. For example, one quick way to choose the number of components is to plot the percent of variance explained in the response variable as a function of the number of components.

```
plot(1:10,cumsum(100*PLSPctVar(2,:)),'-bo');
xlabel('Number of PLS components');
ylabel('Percent Variance Explained in Y');
```



In practice, more care would probably be advisable in choosing the number of components. Cross-validation, for instance, is a widely-used method that will be illustrated later in this example. For now, the above plot suggests that PLSR with two components explains most of the variance in the observed `y`. Compute the fitted response values for the two-component model.

```
[Xloadings,Yloadings,Xscores,Yscores,betaPLS] = plsregress(X,y,2);
yfitPLS = [ones(n,1) X]*betaPLS;
```

Next, fit a PCR model with two principal components. The first step is to perform Principal Components Analysis on `X`, using the `pca` function, and retaining two principal components. PCR is then just a linear regression of the response variable on those two components. It often makes sense to normalize each variable first by its standard deviation when the variables have very different amounts of variability, however, that is not done here.

```
[PCALoadings,PCAScores,PCAVar] = pca(X,'Economy',false);
betaPCR = regress(y-mean(y), PCAScores(:,1:2));
```

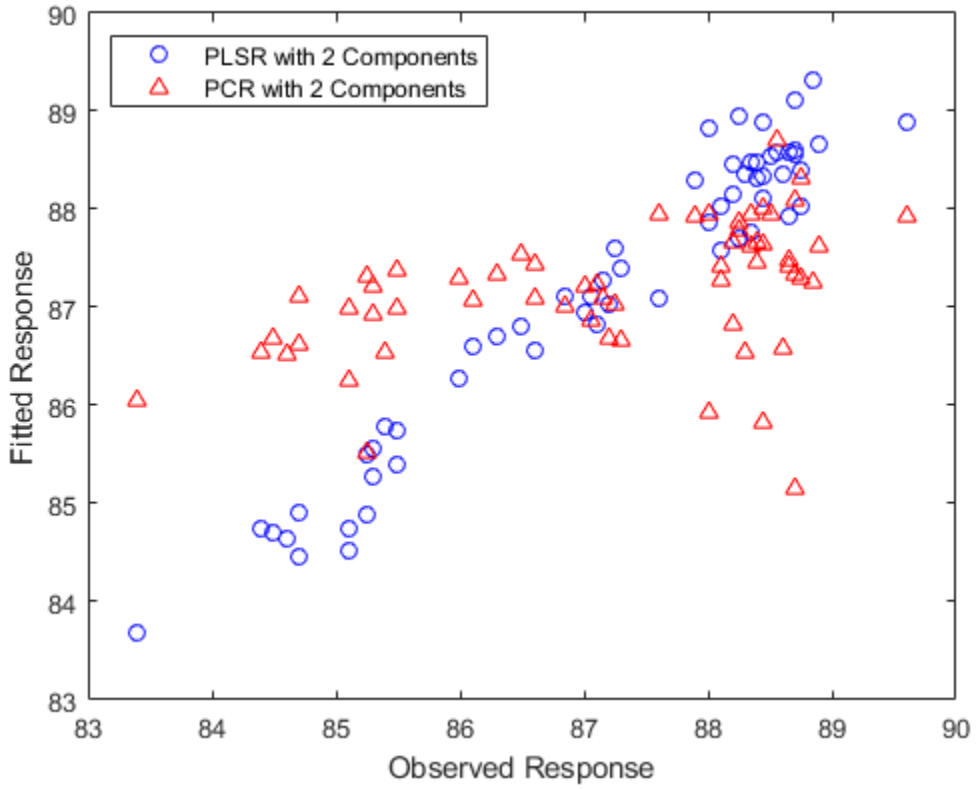
To make the PCR results easier to interpret in terms of the original spectral data, transform to regression coefficients for the original, uncentered variables.

```
betaPCR = PCALoadings(:,1:2)*betaPCR;
betaPCR = [mean(y) - mean(X)*betaPCR; betaPCR];
yfitPCR = [ones(n,1) X]*betaPCR;
```

Plot fitted vs. observed response for the PLSR and PCR fits.

```
plot(y,yfitPLS,'bo',y,yfitPCR,'r^');
xlabel('Observed Response');
```

```
ylabel('Fitted Response');
legend({'PLSR with 2 Components' 'PCR with 2 Components'},  ...
    'location','NW');
```



In a sense, the comparison in the plot above is not a fair one -- the number of components (two) was chosen by looking at how well a two-component PLSR model predicted the response, and there's no reason why the PCR model should be restricted to that same number of components. With the same number of components, however, PLSR does a much better job at fitting `y`. In fact, looking at the horizontal scatter of fitted values in the plot above, PCR with two components is hardly better than using a constant model. The r-squared values from the two regressions confirm that.

```
TSS = sum((y-mean(y)).^2);
RSS_PLS = sum((y-yfitPLS).^2);
rsquaredPLS = 1 - RSS_PLS/TSS
```

```
rsquaredPLS =

    0.9466
```

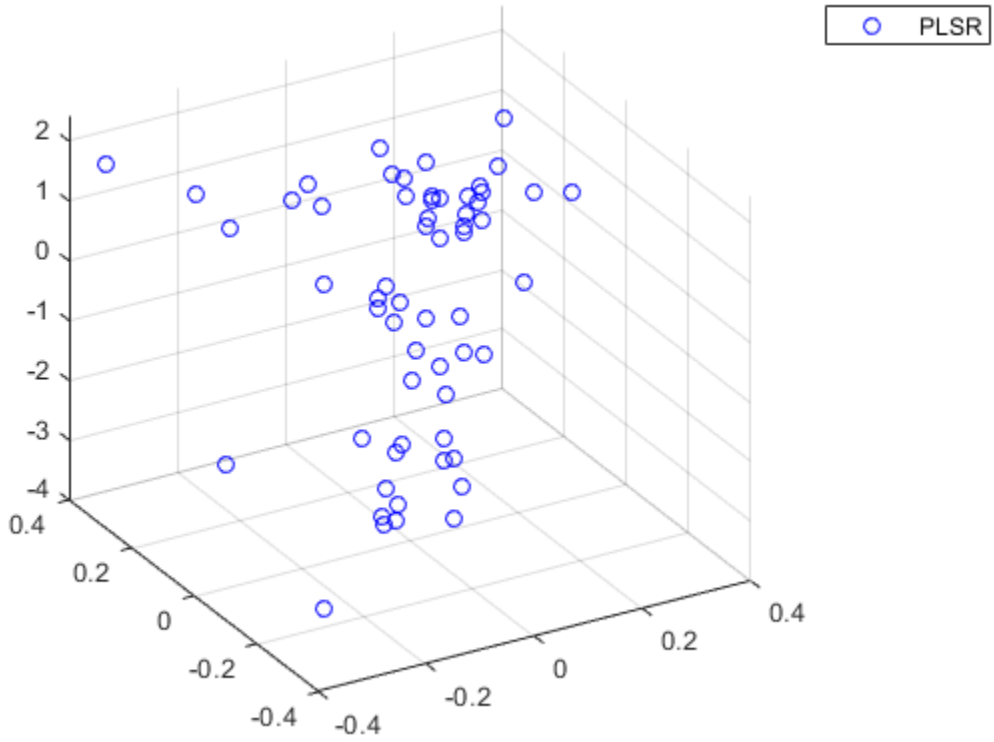
```
RSS_PCR = sum((y-yfitPCR).^2);
rsquaredPCR = 1 - RSS_PCR/TSS
```

```
rsquaredPCR =
```

```
    0.1962
```

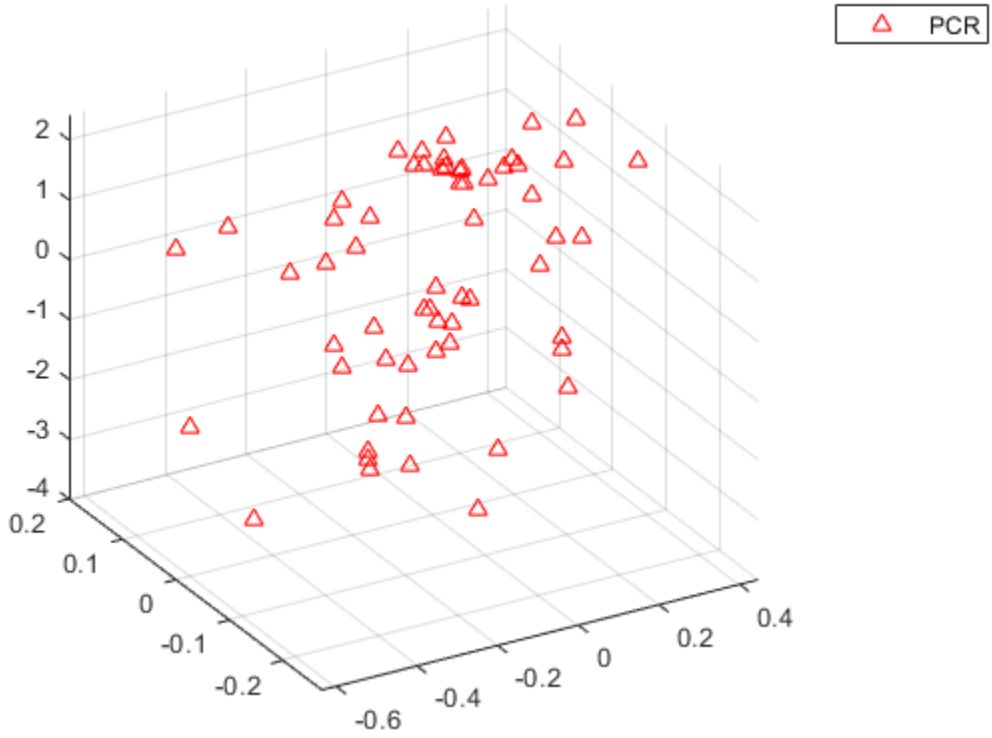
Another way to compare the predictive power of the two models is to plot the response variable against the two predictors in both cases.

```
plot3(Xscores(:,1),Xscores(:,2),y-mean(y),'bo');
legend('PLSR');
grid on; view(-30,30);
```



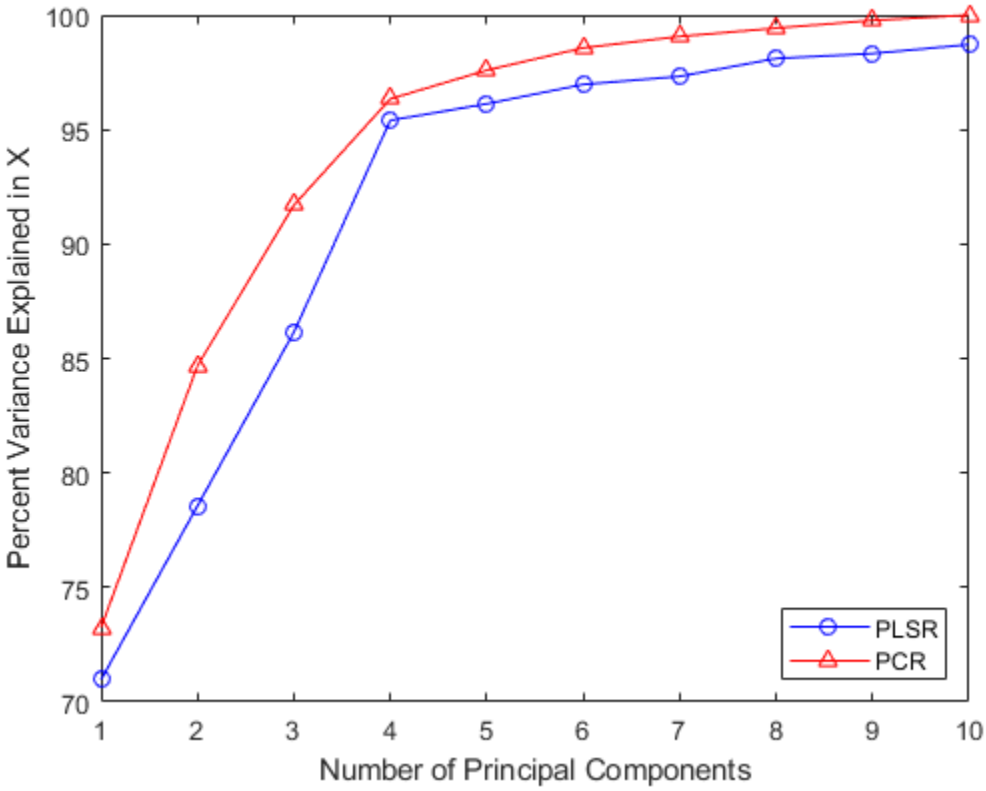
It's a little hard to see without being able to interactively rotate the figure, but the PLSR plot above shows points closely scattered about a plane. On the other hand, the PCR plot below shows a cloud of points with little indication of a linear relationship.

```
plot3(PCAScores(:,1),PCAScores(:,2),y-mean(y),'r^');
legend('PCR');
grid on; view(-30,30);
```



Notice that while the two PLS components are much better predictors of the observed y, the following figure shows that they explain somewhat less variance in the observed X than the first two principal components used in the PCR.

```
plot(1:10,100*cumsum(PLSPctVar(1,:)),'b-o',1:10,  ...
    100*cumsum(PCAVar(1:10))/sum(PCAVar(1:10)),'r-^');
xlabel('Number of Principal Components');
ylabel('Percent Variance Explained in X');
legend({'PLSR' 'PCR'},'location','SE');
```

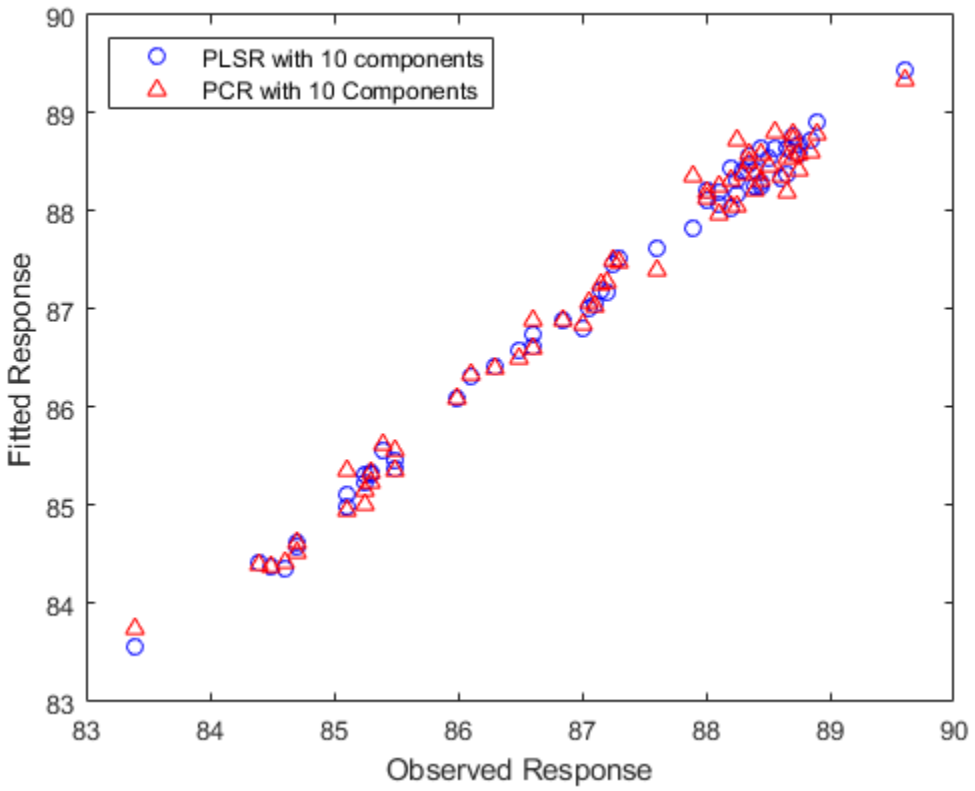


The fact that the PCR curve is uniformly higher suggests why PCR with two components does such a poor job, relative to PLSR, in fitting y. PCR constructs components to best explain X, and as a result, those first two components ignore the information in the data that is important in fitting the observed y.

Fitting with More Components

As more components are added in PCR, it will necessarily do a better job of fitting the original data y, simply because at some point most of the important predictive information in X will be present in the principal components. For example, the following figure shows that the difference in residuals for the two methods is much less dramatic when using ten components than it was for two components.

```
yfitPLS10 = [ones(n,1) X]*betaPLS10;
betaPCR10 = regress(y-mean(y), PCAScores(:,1:10));
betaPCR10 = PCALoadings(:,1:10)*betaPCR10;
betaPCR10 = [mean(y) - mean(X)*betaPCR10; betaPCR10];
yfitPCR10 = [ones(n,1) X]*betaPCR10;
plot(y,yfitPLS10,'bo',y,yfitPCR10,'r^');
xlabel('Observed Response');
ylabel('Fitted Response');
legend({'PLSR with 10 components' 'PCR with 10 Components'},  ...
    'location','NW');
```



Both models fit `y` fairly accurately, although PLSR still makes a slightly more accurate fit. However, ten components is still an arbitrarily-chosen number for either model.

Choosing the Number of Components with Cross-Validation

It's often useful to choose the number of components to minimize the expected error when predicting the response from future observations on the predictor variables. Simply using a large number of components will do a good job in fitting the current observed data, but is a strategy that leads to overfitting. Fitting the current data too well results in a model that does not generalize well to other data, and gives an overly-optimistic estimate of the expected error.

Cross-validation is a more statistically sound method for choosing the number of components in either PLSR or PCR. It avoids overfitting data by not reusing the same data to both fit a model and to estimate prediction error. Thus, the estimate of prediction error is not optimistically biased downwards.

`plsregress` has an option to estimate the mean squared prediction error (MSEP) by cross-validation, in this case using 10-fold C-V.

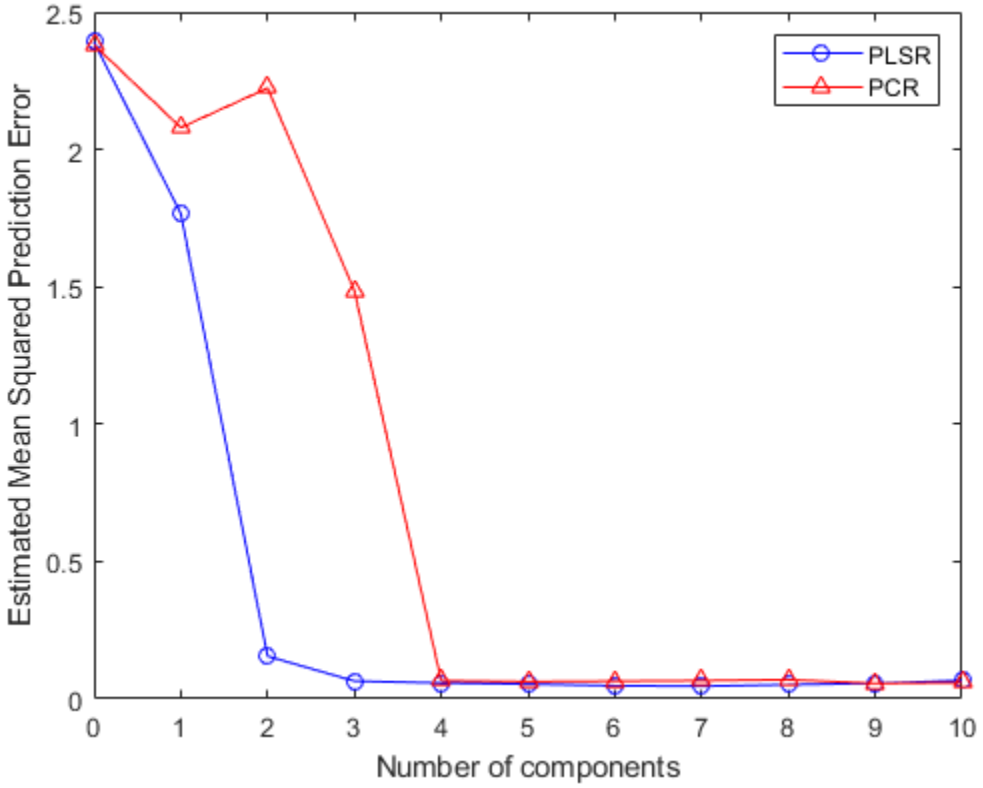
```
[Xl,Yl,Xs,Ys,beta,pctVar,PLSmsep] = plsregress(X,y,10,'CV',10);
```

For PCR, `crossval` combined with a simple function to compute the sum of squared errors for PCR, can estimate the MSEP, again using 10-fold cross-validation.

```
PCRmsep = sum(crossval(@pcrsse,X,y,'KFold',10),1) / n;
```

The MSEP curve for PLSR indicates that two or three components does about as good a job as possible. On the other hand, PCR needs four components to get the same prediction accuracy.

```
plot(0:10,PLSmsep(2,:),'b-o',0:10,PCRmsep,'r-^');
xlabel('Number of components');
ylabel('Estimated Mean Squared Prediction Error');
legend({'PLSR' 'PCR'},'location','NE');
```



In fact, the second component in PCR *increases* the prediction error of the model, suggesting that the combination of predictor variables contained in that component is not strongly correlated with `y`. Again, that's because PCR constructs components to explain variation in `X`, not `y`.

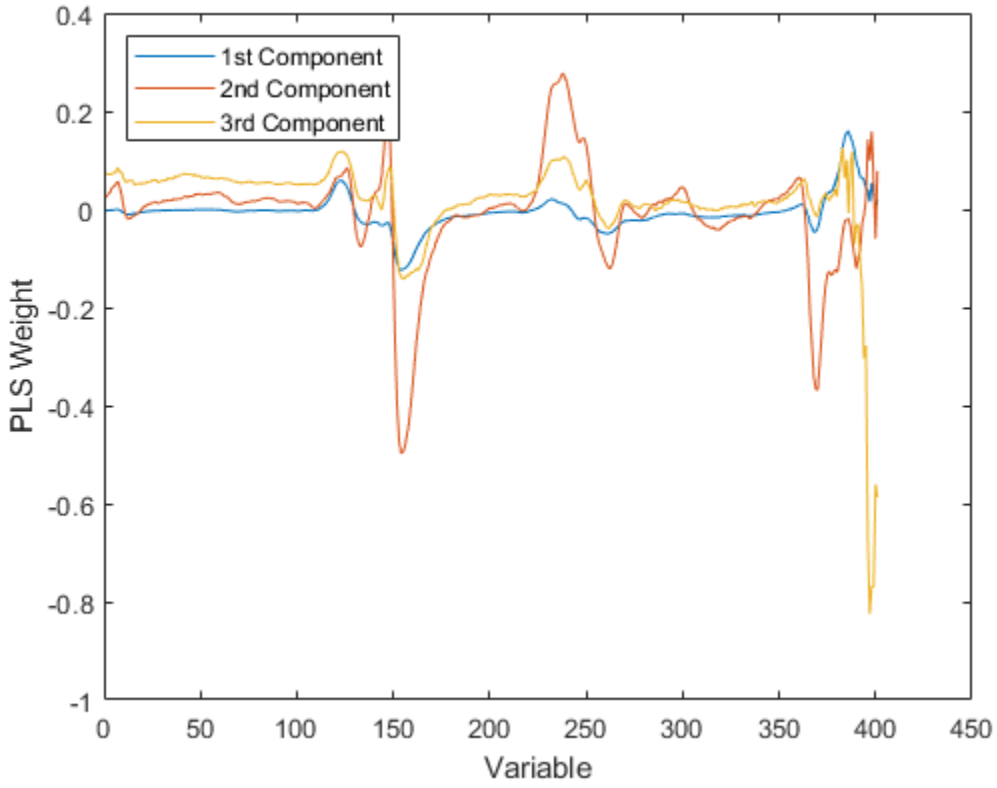
Model Parsimony

So if PCR requires four components to get the same prediction accuracy as PLSR with three components, is the PLSR model more parsimonious? That depends on what aspect of the model you consider.

The PLS weights are the linear combinations of the original variables that define the PLS components, i.e., they describe how strongly each component in the PLSR depends on the original variables, and in what direction.

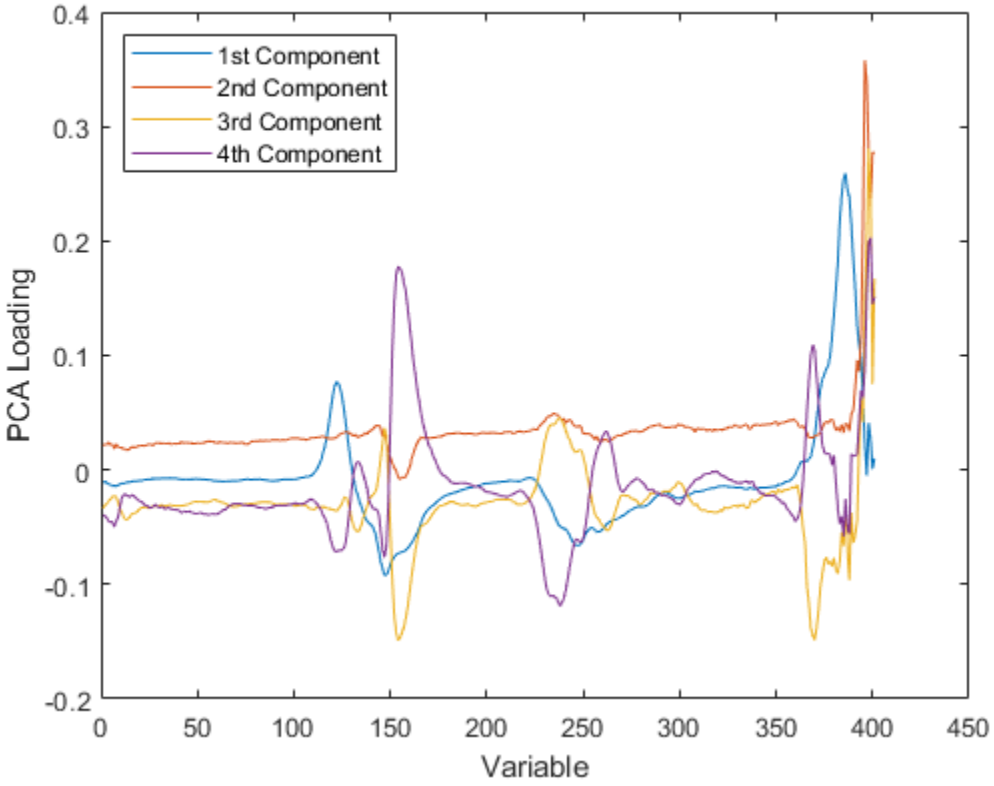
```
[Xl,Yl,Xs,Ys,beta,pctVar,mse,stats] = plsregress(X,y,3);
plot(1:401,stats.W,'-');
xlabel('Variable');
ylabel('PLS Weight');
```

```
legend({'1st Component' '2nd Component' '3rd Component'},  ...
    'location','NW');
```



Similarly, the PCA loadings describe how strongly each component in the PCR depends on the original variables.

```
plot(1:401,PCALoadings(:,1:4),'-');
xlabel('Variable');
ylabel('PCA Loading');
legend({'1st Component' '2nd Component' '3rd Component'  ...
    '4th Component'},'location','NW');
```



For either PLSR or PCR, it may be that each component can be given a physically meaningful interpretation by inspecting which variables it weights most heavily. For instance, with these spectral data it may be possible to interpret intensity peaks in terms of compounds present in the gasoline, and then to observe that weights for a particular component pick out a small number of those compounds. From that perspective, fewer components are simpler to interpret, and because PLSR often requires fewer components to predict the response adequately, it leads to more parsimonious models.

On the other hand, both PLSR and PCR result in one regression coefficient for each of the original predictor variables, plus an intercept. In that sense, neither is more parsimonious, because regardless of how many components are used, both models depend on all predictors. More concretely, for these data, both models need 401 spectral intensity values in order to make a prediction.

However, the ultimate goal may to reduce the original set of variables to a smaller subset still able to predict the response accurately. For example, it may be possible to use the PLS weights or the PCA loadings to select only those variables that contribute most to each component. As shown earlier, some components from a PCR model fit may serve primarily to describe the variation in the predictor variables, and may include large weights for variables that are not strongly correlated with the response. Thus, PCR can lead to retaining variables that are unnecessary for prediction.

For the data used in this example, the difference in the number of components needed by PLSR and PCR for accurate prediction is not great, and the PLS weights and PCA loadings seem to pick out the same variables. That may not be true for other data.

Generalized Linear Models

Multinomial Models for Nominal Responses

The outcome of a response variable might be one of a restricted set of possible values. If there are only two possible outcomes, such as a yes or no answer to a question, these responses are called binary responses. If there are multiple outcomes, then they are called polytomous responses. Some examples include the degree of a disease (mild, medium, severe), preferred districts to live in a city, and so on. When the response variable is *nominal*, there is no natural order among the response variable categories. Nominal response models explain and predict the probability that an observation is in each category of a categorical response variable.

A nominal response model is one of several natural extensions of the binary logit model and is also called a *multinomial logit* model. The multinomial logit model explains the relative risk of being in one category versus being in the reference category, k, using a linear combination of predictor variables. Consequently, the probability of each outcome is expressed as a nonlinear function of p predictor variables. The `'interactions','on'` name-value pair argument in `mnrfit` corresponds to this multinomial model with separate intercept and slopes among categories. `mnrfit` uses the default logit link function for multinomial models. You cannot specify a different link function for multinomial responses.

The multinomial logit model is

$$\ln\left(\frac{\pi_1}{\pi_k}\right) = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p,$$

$$\ln\left(\frac{\pi_2}{\pi_k}\right) = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2p}X_p,$$

$$\vdots$$

$$\ln\left(\frac{\pi_{k-1}}{\pi_k}\right) = \alpha_{(k-1)} + \beta_{(k-1)1}X_1 + \beta_{(k-1)2}X_2 + \cdots + \beta_{(k-1)p}X_p,$$

where $\pi_j = \mathrm{P}(y = j)$ is the probability of an outcome being in category j, k is the number of response categories, and p is the number of predictor variables. Theoretically, any category can be the reference category, but `mnrfit` chooses the last one, k, as the reference category. Thus, `mnrfit` assumes the coefficients of the kth category are zero. The total of j – 1 equations are solved simultaneously to estimate the coefficients. `mnrfit` uses the iteratively weighted least squares algorithm to find the maximum likelihood estimates.

The coefficients in the model express the effects of the predictor variables on the relative risk or the log odds of being in category j versus the reference category, here k. For example, the coefficient β_{23} indicates that the probability of the response variable being in category 2 compared to the probability of being in category k increases $\exp(\beta_{23})$ times for each unit increase in X_3, given all else is held constant. Or it indicates that the relative log odds of the response variable being category 2 versus in category k increases β_{23} times with a one-unit increase in X_3, given all else equal.

Based on the nominal response model, and the assumption that the coefficients for the last category are zero, the probability of being in each category is

$$\pi_j = P(y = j) = \frac{e^{\alpha_j + \sum_{l=1}^{p} \beta_{jl}x_l}}{1 + \sum_{j=1}^{k-1} e^{\alpha_j + \sum_{l=1}^{p} \beta_{jl}x_l}}, \quad j = 1, \cdots, k-1.$$

The probability of the kth category becomes

$$\pi_k = P(y = k) = \frac{1}{1 + \sum_{j=1}^{k-1} e^{\alpha_j + \sum_{l=1}^{p} \beta_{jl} x_l}},$$

which is simply equal to $1 - \pi_1 - \pi_2 - \ldots - \pi_{k-1}$.

After estimating the model coefficients using `mnrfit`, you can estimate the category probabilities or the number in each category using `mnrval` (the default name-value pair is `'type','category'`). This function accepts the coefficient estimates and the model statistics `mnrfit` returns and estimates the categorical probabilities or the number in each category and their confidence bounds. You can also specify the cumulative or conditional probabilities or numbers to estimate using the `'type'` name-value pair argument in `mnrval`.

References

[1] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, 1997.

[3] Dobson, A. J., and A. G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC. Taylor & Francis Group, 2008.

See Also

`fitglm` | `glmfit` | `glmval` | `mnrfit` | `mnrval`

More About

- “Multinomial Models for Ordinal Responses” on page 12-4
- “Hierarchical Multinomial Models” on page 12-7

Multinomial Models for Ordinal Responses

The outcome of a response variable might be one of a restricted set of possible values. If there are only two possible outcomes, such as male and female for gender, these responses are called binary responses. If there are multiple outcomes, then they are called polytomous responses. Some examples of polytomous responses include levels of a disease (mild, medium, severe), preferred districts to live in a city, the species for a certain flower type, and so on. Sometimes there might be a natural order among the response categories. These responses are called *ordinal responses*.

The ordering might be inherent in the category choices, such as an individual being not satisfied, satisfied, or very satisfied with an online customer service. The ordering might also be introduced by categorization of a latent (continuous) variable, such as in the case of an individual being in the low risk, medium risk, or high risk group for developing a certain disease, based on a quantitative medical measure such as blood pressure.

You can specify a multinomial regression model that uses the natural ordering among the response categories. This ordinal model describes the relationship between the cumulative probabilities of the categories and predictor variables.

Different link functions can describe this relationship with logit and probit being the most used.

- **Logit:** The default link function `mnrfit` uses for ordinal categories is the *logit* link function. This models the *log cumulative odds*. The `'link','logit'` name-value pair specifies this in `mnrfit`. Log cumulative odds is the logarithm of the ratio of the probability that a response belongs to a category with a value less than or equal to category j, $P(y \le c_j)$, and the probability that a response belongs to a category with a value greater than category j, $P(y > c_j)$.

  Ordinal models are usually based on the assumption that the effects of predictor variables are the same for all categories on the logarithmic scale. That is, the model has different intercepts but common slopes (coefficients) among categories. This model is called *parallel regression* or the *proportional odds* model. It is the default for ordinal responses, and the `'interactions','off'` name-value pair specifies this model in `mnrfit`.

  The proportional odds model is

$$\ln\left(\frac{P(y \le c_1)}{P(y > c_1)}\right) = \ln\left(\frac{\pi_1}{\pi_2 + \cdots + \pi_k}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

$$\ln\left(\frac{P(y \le c_2)}{P(y > c_2)}\right) = \ln\left(\frac{\pi_1 + \pi_2}{\pi_3 + \cdots + \pi_k}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

$$\vdots$$

$$\ln\left(\frac{P(y \le c_{k-1})}{P(y > c_{k-1})}\right) = \ln\left(\frac{\pi_1 + \pi_2 + \cdots + \pi_{k-1}}{\pi_k}\right) = \alpha_{k-1} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

  where π_j, $j = 1, 2, \ldots, k$, are the category probabilities.

  For example, for a response variable with three categories, there are $3 - 1 = 2$ equations as follows:

$$\ln\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) = \alpha_1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

$$\ln\left(\frac{\pi_1 + \pi_2}{\pi_3}\right) = \alpha_2 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Under the proportional odds assumption, the partial effect of a predictor variable X is invariant to the choice of the response variable category, j. For example, if there are three categories, then the coefficients express the impact of a predictor variable on the relative risk or log odds of the response value being in category 1 versus categories 2 or 3, or in category 1 or 2 versus category 3.

Thus, a unit change in variable X_2 would mean a change in the cumulative odds of the response value being in category 1 versus categories 2 or 3, or category 1 or 2 versus category 3 by a factor of $\exp(\beta_2)$, given all else equal.

You can alternatively fit a model with different intercept and slopes among the categories by using the `'interactions','on'` name-value pair argument. However, using this option for ordinal models when the equal slopes model is true causes a loss of efficiency (you lose the advantage of estimating fewer parameters).

- **Probit:** The `'link','probit'` name-value pair argument uses the *probit* link function which is based on a normally distributed latent variable assumption. For ordinal response variables this is also called an *ordered probit* model. Consider the regression model that describes the relationship of a latent variable y^* of an ordinal process and a vector of predictor variables, X,

  $$y^* = \beta X + \varepsilon,$$

  where the error term ε has a standard normal distribution. Suppose there is the following relationship between the latent variable y^* and the observed variable y:

  $$\begin{aligned} y &= c_1 \quad if \quad \alpha_0 < y^* \le \alpha_1, \\ y &= c_2 \quad if \quad \alpha_1 < y^* \le \alpha_2, \\ &\vdots \qquad\qquad \vdots \\ y &= c_k \quad if \quad \alpha_{k-1} < y^* \le \alpha_k, \end{aligned}$$

  where $\alpha_0 = -\infty$ and $\alpha_k = \infty$. Then, the cumulative probability of y being in category j or one of earlier categories, $P(y \le c_j)$, is equal to

  $$P\left(y \le c_j\right) = P\left(y^* < \alpha_j\right) = P\left(\beta X + \varepsilon < \alpha_j\right) = P\left(\varepsilon < \alpha_j - \beta X\right) = \Phi\left(\alpha_j - \beta X\right),$$

  where Φ is standard normal cumulative distribution function. Thus,

  $$\Phi^{-1}\left(P\left(y \le c_j\right)\right) = \alpha_j - \beta X,$$

  where α_j corresponds to the cut points of the latent variable and the intercept in the regression model. This only holds under the assumptions of a normal latent variable and parallel regression. More generally, for a response variable with k categories and multiple predictors, the ordered probit model is

  $$\begin{aligned} \Phi^{-1}(P(y \le c_1)) &= \alpha_1 + \beta_1 X_1 + \cdots + \beta_p X_p, \\ \Phi^{-1}(P(y \le c_2)) &= \alpha_2 + \beta_1 X_1 + \cdots + \beta_p X_p, \\ &\vdots \qquad\qquad\qquad \vdots \\ \Phi^{-1}(P(y \le c_{k-1})) &= \alpha_{k-1} + \beta_1 X_1 + \cdots + \beta_p X_p, \end{aligned}$$

  where $P(y \le c_j) = \pi_1 + \pi_2 + \ldots + \pi_j$.

  The coefficients indicate the impact of a unit change in the predictor variable on the likelihood of a state. A positive coefficient, β_1, for example, indicates an increase in the underlying latent

variable with an increase in the corresponding predictor variable, X_1. Hence, it causes a decrease in $P(y \le c_1)$ and an increase in $P(y \le c_k)$.

After estimating the model coefficients using `mnrfit`, you can estimate the cumulative probabilities or the cumulative number in each category using `mnrval` with the `'type','cumulative'` name-value pair option. `mnrval` accepts the coefficient estimates and the model statistics `mnrfit` returns, and estimates the categorical probabilities or the number in each category and their confidence intervals. You can specify which category or conditional probabilities or numbers to estimate by changing the value of the `'type'` name-value pair argument.

References

[1] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, 1997.

[3] Dobson, A. J., and A. G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC. Taylor & Francis Group, 2008.

See Also

`fitglm` | `glmfit` | `glmval` | `mnrfit` | `mnrval`

More About

- “Multinomial Models for Nominal Responses” on page 12-2
- “Hierarchical Multinomial Models” on page 12-7

Hierarchical Multinomial Models

The outcome of a response variable might sometimes be one of a restricted set of possible values. If there are only two possible outcomes, such as male and female for gender, these responses are called binary responses. If there are multiple outcomes, then they are called polytomous responses. These responses are usually qualitative rather than quantitative, such as preferred districts to live in a city, the severity level of a disease, the species for a certain flower type, and so on. Polytomous responses might also have categories which are not independent of each other. Instead the response happens in a sequential manner, or one category is nested in the previous one. These types of responses are called *hierarchical*, *or sequential*, or *nested multinomial responses*.

For example, if the response is the number of cigarettes a person smokes in a given day, the first level is whether the person is a smoker or not. Given that he or she is a smoker, the number of cigarettes he or she smokes can be from one to five or more than five a day. Given that it is more than 5, this person might be smoking from 6 to 10 or more than 10 cigarettes a day, and so on. The risk group at each level changes accordingly. At level one, the risk group is all of the individuals of interest (smoker or not), say m. If out of m individuals, y_1 of them are not smokers, then at level two, the risk group is the number of all smoking individuals, $m - y_1$. If y_2 of these $m - y_1$ individuals smoke from one to five cigarettes a day, then at level three, the risk group is $m - y_1 - y_2$. So, at each level, the number of people in that category becomes a conditional binomial observation.

The hierarchical multinomial regression models are extensions of binary regression models based on conditional binary observations. The default is a model with different intercept and slopes (coefficients) among categories, in which case `mnrfit` fits a sequence of conditional binomial models. The `'interactions','on'` name-value pair specifies this in `mnrfit`. The default link function is logit and the `'link','logit'` name-value pair specifies this model in `mnrfit`.

Suppose the probability that an individual is in category j given that he or she is not in the previous categories is π_j, and the cumulative probability that a response belongs to a category j or a previous category is $P(y \le c_j)$. Then the hierarchical model with a logit link function and different slopes assumption is

$$\ln\left(\frac{\pi_1}{1-P(y\le c_1)}\right) = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p,$$

$$\ln\left(\frac{\pi_2}{1-P(y\le c_2)}\right) = \ln\left(\frac{\pi_2}{1-(\pi_1+\pi_2)}\right) = \alpha_2 + \beta_{21}X_2 + \beta_{22}X_2 + \cdots + \beta_{2p}X_p,$$

$$\vdots$$

$$\ln\left(\frac{\pi_{k-1}}{1-P(y\le c_{k-1})}\right) = \ln\left(\frac{\pi_{k-1}}{1-(\pi_1+\cdots+\pi_{k-1})}\right) = \alpha_{k-1} + \beta_{(k-1)1}X_1 + \beta_{(k-1)2}X_2 + \cdots + \beta_{(k-1)p}X_p\,.$$

For example, for a response variable with four sequential categories, there are $4 - 1 = 3$ equations as follows:

$$\ln\left(\frac{\pi_1}{\pi_2+\pi_3+\pi_4}\right) = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}X_p,$$

$$\ln\left(\frac{\pi_2}{\pi_3+\pi_4}\right) = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2p}X_p,$$

$$\ln\left(\frac{\pi_3}{\pi_4}\right) = \alpha_3 + \beta_{31}X_1 + \beta_{32}X_2 + \cdots + \beta_{3p}X_p\,.$$

The coefficients β_{ij} are interpreted within each level. For example, for the previous smoking example, β_{12} shows the impact of X_2 on the log odds of a person being a smoker versus a nonsmoker, provided that everything else is held constant. Alternatively, β_{22} shows the impact of X_2 on the log odds of a person smoking one to five cigarettes versus more than five cigarettes a day, given that he or she is a smoker, provided that everything else is held constant. Similarly, β_{23}, shows the effect of X_2 on the log odds of a person smoking 6 to 10 cigarettes versus more than 10 cigarettes a day, given that he or she smokes more than 5 cigarettes a day, provided that everything else is held constant.

You can specify other link functions for hierarchical models. The `'link','probit'` name-value pair argument uses the probit link function. With the separate slopes assumption, the model becomes

$$\begin{aligned}
\Phi^{-1}(\pi_1) &= \alpha_1 + \beta_{11}X_1 + \cdots + \beta_{1p}X_p, \\
\Phi^{-1}(\pi_2) &= \alpha_2 + \beta_{21}X_1 + \cdots + \beta_{2p}X_p, \\
&\vdots \qquad\qquad\qquad \vdots \\
\Phi^{-1}(\pi_k) &= \alpha_k + \beta_{k1}X_1 + \cdots + \beta_{kp}X_p,
\end{aligned}$$

where π_j is the conditional probability of being in category j, given that it is not in categories previous to category j. And $\Phi^{-1}(.)$ is the inverse of the standard normal cumulative distribution function.

After estimating the model coefficients using `mnrfit`, you can estimate the cumulative probabilities or the cumulative number in each category using `mnrval` with the `'type','conditional'` name-value pair argument. The function `mnrval` accepts the coefficient estimates and the model statistics `mnrfit` returns, and estimates the categorical probabilities or the number in each category and their confidence bounds. You can specify which category or cumulative probabilities or numbers to estimate by changing the value of the `'type'` name-value pair argument in `mnrval`.

References

[1] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] Liao, T. F. *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models* Series: Quantitative Applications in the Social Sciences. Sage Publications, 1994.

See Also

`fitglm` | `glmfit` | `glmval` | `mnrfit` | `mnrval`

More About

- “Multinomial Models for Nominal Responses” on page 12-2
- “Multinomial Models for Ordinal Responses” on page 12-4

Generalized Linear Models

In this section...

What Are Generalized Linear Models?

Linear regression models describe a linear relationship between a response and one or more predictive terms. Many times, however, a nonlinear relationship exists. “Nonlinear Regression” on page 13-2 describes general nonlinear models. A special class of nonlinear models, called *generalized linear models*, uses linear methods.

Recall that linear models have these characteristics:

- At each set of values for the predictors, the response has a normal distribution with mean μ.
- A coefficient vector b defines a linear combination Xb of the predictors X.
- The model is $\mu = Xb$.

In generalized linear models, these characteristics are generalized as follows:

- At each set of values for the predictors, the response has a distribution that can be normal on page B-119, binomial on page B-10, Poisson on page B-131, gamma on page B-47, or inverse Gaussian on page B-75, with parameters including a mean μ.
- A coefficient vector b defines a linear combination Xb of the predictors X.
- A link function f defines the model as $f(\mu) = Xb$.

Prepare Data

To begin fitting a regression, put your data into a form that fitting functions expect. All regression techniques begin with input data in an array `X` and response data in a separate vector `y`, or input data in a table or dataset array `tbl` and response data as a column in `tbl`. Each row of the input data represents one observation. Each column represents one predictor (variable).

For a table or dataset array `tbl`, indicate the response variable with the `'ResponseVar'` name-value pair:

```
mdl = fitglm(tbl,'ResponseVar','BloodPressure');
```

The response variable is the last column by default.

You can use numeric categorical predictors. A categorical predictor is one that takes values from a fixed set of possibilities.

- For a numeric array `X`, indicate the categorical predictors using the `'Categorical'` name-value pair. For example, to indicate that predictors 2 and 3 out of six are categorical:

  ```
  mdl = fitglm(X,y,'Categorical',[2,3]);
  % or equivalently
  mdl = fitglm(X,y,'Categorical',logical([0 1 1 0 0 0]));
  ```

- For a table or dataset array `tbl`, fitting functions assume that these data types are categorical:

  - Logical vector
  - Categorical vector
  - Character array
  - String array

  If you want to indicate that a numeric predictor is categorical, use the `'Categorical'` name-value pair.

Represent missing numeric data as `NaN`. To represent missing data for other data types, see “Missing Group Values” on page 2-46.

- For a `'binomial'` model with data matrix `X`, the response `y` can be:

  - Binary column vector — Each entry represents success (`1`) or failure (`0`).
  - Two-column matrix of integers — The first column is the number of successes in each observation, the second column is the number of trials in that observation.

- For a `'binomial'` model with table or dataset `tbl`:

  - Use the `ResponseVar` name-value pair to specify the column of `tbl` that gives the number of successes in each observation.
  - Use the `BinomialSize` name-value pair to specify the column of `tbl` that gives the number of trials in each observation.

Dataset Array for Input and Response Data

For example, to create a dataset array from an Excel spreadsheet:

```
ds = dataset('XLSFile','hospital.xls',...
     'ReadObsNames',true);
```

To create a dataset array from workspace variables:

```
load carsmall
ds = dataset(MPG,Weight);
ds.Year = ordinal(Model_Year);
```

Table for Input and Response Data

To create a table from workspace variables:

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = ordinal(Model_Year);
```

Numeric Matrix for Input Data, Numeric Vector for Response

For example, to create numeric arrays from workspace variables:

```
load carsmall
X = [Weight Horsepower Cylinders Model_Year];
y = MPG;
```

To create numeric arrays from an Excel spreadsheet:

```
[X, Xnames] = xlsread('hospital.xls');
y = X(:,4); % response y is systolic pressure
X(:,4) = []; % remove y from the X matrix
```

Notice that the nonnumeric entries, such as `sex`, do not appear in `X`.

Choose Generalized Linear Model and Link Function

Often, your data suggests the distribution type of the generalized linear model.

Response Data Type	Suggested Model Distribution Type
Any real number	`'normal'`
Any positive number	`'gamma'` or `'inverse gaussian'`
Any nonnegative integer	`'poisson'`
Integer from 0 to `n`, where `n` is a fixed positive value	`'binomial'`

Set the model distribution type with the `Distribution` name-value pair. After selecting your model type, choose a link function to map between the mean μ and the linear predictor Xb.

Value	Description
`'comploglog'`	$\log(-\log((1 - \mu))) = Xb$
`'identity'`, default for the distribution `'normal'`	$\mu = Xb$
`'log'`, default for the distribution `'poisson'`	$\log(\mu) = Xb$
`'logit'`, default for the distribution `'binomial'`	$\log(\mu/(1 - \mu)) = Xb$
`'loglog'`	$\log(-\log(\mu)) = Xb$
`'probit'`	$\Phi^{-1}(\mu) = Xb$, where Φ is the normal (Gaussian) cumulative distribution function
`'reciprocal'`, default for the distribution `'gamma'`	$\mu^{-1} = Xb$
`p` (a number), default for the distribution `'inverse gaussian'` (with $p = -2$)	$\mu^p = Xb$

Value	Description
A cell array of the form `{FL FD FI}`, containing three function handles created using `@`, which define the link (`FL`), the derivative of the link (`FD`), and the inverse link (`FI`). Or, a structure of function handles with the field `Link` containing `FL`, the field `Derivative` containing `FD`, and the field `Inverse` containing `FI`.	User-specified link function (see “Custom Link Function” on page 12-12)

The nondefault link functions are mainly useful for binomial models. These nondefault link functions are `'comploglog'`, `'loglog'`, and `'probit'`.

Custom Link Function

The link function defines the relationship $f(\mu) = Xb$ between the mean response μ and the linear combination $Xb = X*b$ of the predictors. You can choose one of the built-in link functions or define your own by specifying the link function `FL`, its derivative `FD`, and its inverse `FI`:

- The link function `FL` calculates $f(\mu)$.
- The derivative of the link function `FD` calculates $df(\mu)/d\mu$.
- The inverse function `FI` calculates $g(Xb) = \mu$.

You can specify a custom link function in either of two equivalent ways. Each way contains function handles that accept a single array of values representing μ or Xb, and returns an array the same size. The function handles are either in a cell array or a structure:

- Cell array of the form `{FL FD FI}`, containing three function handles, created using `@`, that define the link (`FL`), the derivative of the link (`FD`), and the inverse link (`FI`).
- Structure `s` with three fields, each containing a function handle created using `@`:
  - `s.Link` — Link function
  - `s.Derivative` — Derivative of the link function
  - `s.Inverse` — Inverse of the link function

For example, to fit a model using the `'probit'` link function:

```
x = [2100 2300 2500 2700 2900 ...
     3100 3300 3500 3700 3900 4100 4300]';
n = [48 42 31 34 31 21 23 23 21 16 17 21]';
y = [1 2 0 3 8 8 14 17 19 15 17 21]';
g = fitglm(x,[y n],...
    'linear','distr','binomial','link','probit')

g =

Generalized Linear regression model:
    probit(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate     SE            tStat     pValue
    (Intercept)      -7.3628       0.66815    -11.02    3.0701e-28
```

```
    x1             0.0023039    0.00021352     10.79    3.8274e-27

12 observations, 10 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 241, p-value = 2.25e-54
```

You can perform the same fit using a custom link function that performs identically to the `'probit'` link function:

```
s = {@norminv,@(x)1./normpdf(norminv(x)),@normcdf};
g = fitglm(x,[y n],...
    'linear','distr','binomial','link',s)

g =

Generalized Linear regression model:
    link(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate     SE            tStat     pValue
    (Intercept)      -7.3628       0.66815    -11.02    3.0701e-28
    x1             0.0023039    0.00021352     10.79    3.8274e-27

12 observations, 10 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 241, p-value = 2.25e-54
```

The two models are the same.

Equivalently, you can write `s` as a structure instead of a cell array of function handles:

```
s.Link = @norminv;
s.Derivative = @(x) 1./normpdf(norminv(x));
s.Inverse = @normcdf;
g = fitglm(x,[y n],...
    'linear','distr','binomial','link',s)

g =

Generalized Linear regression model:
    link(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate     SE            tStat     pValue
    (Intercept)      -7.3628       0.66815    -11.02    3.0701e-28
    x1             0.0023039    0.00021352     10.79    3.8274e-27

12 observations, 10 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 241, p-value = 2.25e-54
```

Choose Fitting Method and Model

There are two ways to create a fitted model.

- Use `fitglm` when you have a good idea of your generalized linear model, or when you want to adjust your model later to include or exclude certain terms.
- Use `stepwiseglm` when you want to fit your model using stepwise regression. `stepwiseglm` starts from one model, such as a constant, and adds or subtracts terms one at a time, choosing an optimal term each time in a greedy fashion, until it cannot improve further. Use stepwise fitting to find a good model, one that has only relevant terms.

  The result depends on the starting model. Usually, starting with a constant model leads to a small model. Starting with more terms can lead to a more complex model, but one that has lower mean squared error.

In either case, provide a model to the fitting function (which is the starting model for `stepwiseglm`).

Specify a model using one of these methods.

- "Brief Model Name" on page 12-14
- "Terms Matrix" on page 12-14
- "Formula" on page 12-15

Brief Model Name

Name	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear terms for each predictor.
`'interactions'`	Model contains an intercept, linear terms, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept, linear terms, and squared terms.
`'quadratic'`	Model contains an intercept, linear terms, interactions, and squared terms.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, etc. Use numerals `0` through `9`. For example, `'poly2111'` has a constant plus all linear and product terms, and also contains terms with predictor 1 squared.

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for a model specification is a character vector or string scalar of the form

`'y ~ terms'`,

- *`y`* is the response name.
- *`terms`* contains
  - Variable names
  - `+` to include the next variable
  - `-` to exclude the next variable
  - `:` to define an interaction, a product of terms
  - `*` to define an interaction and all lower-order terms
  - `^` to raise the predictor to a power, exactly as in `*` repeated, so `^` includes lower order terms as well
  - `()` to group terms

Tip Formulas include a constant (intercept) term by default. To exclude a constant term from the model, include `-1` in the formula.

Examples:

`'y ~ x1 + x2 + x3'` is a three-variable linear model with intercept.
`'y ~ x1 + x2 + x3 - 1'` is a three-variable linear model without intercept.
`'y ~ x1 + x2 + x3 + x2^2'` is a three-variable model with intercept and a `x2^2` term.
`'y ~ x1 + x2^2 + x3'` is the same as the previous example, since `x2^2` includes a `x2` term.
`'y ~ x1 + x2 + x3 + x1:x2'` includes an `x1*x2` term.
`'y ~ x1*x2 + x3'` is the same as the previous example, since `x1*x2 = x1 + x2 + x1:x2`.
`'y ~ x1*x2*x3 - x1:x2:x3'` has all interactions among `x1`, `x2`, and `x3`, except the three-way interaction.
`'y ~ x1*(x2 + x3 + x4)'` has all linear terms, plus products of `x1` with each of the other variables.

Fit Model to Data

Create a fitted model using `fitglm` or `stepwiseglm`. Choose between them as in “Choose Fitting Method and Model” on page 12-13. For generalized linear models other than those with a normal distribution, give a `Distribution` name-value pair as in “Choose Generalized Linear Model and Link Function” on page 12-11. For example,

```
mdl = fitglm(X,y,'linear','Distribution','poisson')
% or
mdl = fitglm(X,y,'quadratic',...
         'Distribution','binomial')
```

Examine Quality and Adjust the Fitted Model

After fitting a model, examine the result.

Model Display

A linear regression model shows several diagnostics when you enter its name or enter `disp(mdl)`. This display gives some of the basic information to check whether the fitted model represents the data adequately.

For example, fit a Poisson model to data constructed with two out of five predictors not affecting the response, and with no intercept term:

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[.4;.2;.3]);
y = poissrnd(mu);
mdl = fitglm(X,y,...
    'linear','Distribution','poisson')

mdl =


Generalized Linear regression model:
    log(y) ~ 1 + x1 + x2 + x3 + x4 + x5
    Distribution = Poisson

Estimated Coefficients:
                   Estimate     SE          tStat       pValue
    (Intercept)     0.039829     0.10793     0.36901       0.71212
    x1               0.38551    0.076116      5.0647    4.0895e-07
    x2             -0.034905    0.086685    -0.40266        0.6872
    x3              -0.17826    0.093552     -1.9054      0.056722
    x4               0.21929     0.09357      2.3436      0.019097
    x5               0.28918      0.1094      2.6432     0.0082126


100 observations, 94 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 44.9, p-value = 1.55e-08
```

Notice that:

- The display contains the estimated values of each coefficient in the `Estimate` column. These values are reasonably near the true values `[0;.4;0;0;.2;.3]`, except possibly the coefficient of `x3` is not terribly near `0`.
- There is a standard error column for the coefficient estimates.
- The reported `pValue` (which are derived from the *t* statistics under the assumption of normal errors) for predictors 1, 4, and 5 are small. These are the three predictors that were used to create the response data `y`.

- The `pValue` for `(Intercept)`, `x2` and `x3` are larger than 0.01. These three predictors were not used to create the response data `y`. The `pValue` for `x3` is just over `.05`, so might be regarded as possibly significant.
- The display contains the Chi-square statistic.

Diagnostic Plots

Diagnostic plots help you identify outliers, and see other problems in your model or fit. To illustrate these plots, consider binomial regression with a logistic link function.

The logistic model is useful for proportion data. It defines the relationship between the proportion p and the weight w by:

$\log[p/(1 - p)] = b_1 + b_2 w$

This example fits a binomial model to data. The data are derived from `carbig.mat`, which contains measurements of large cars of various weights. Each weight in `w` has a corresponding number of cars in `total` and a corresponding number of poor-mileage cars in `poor`.

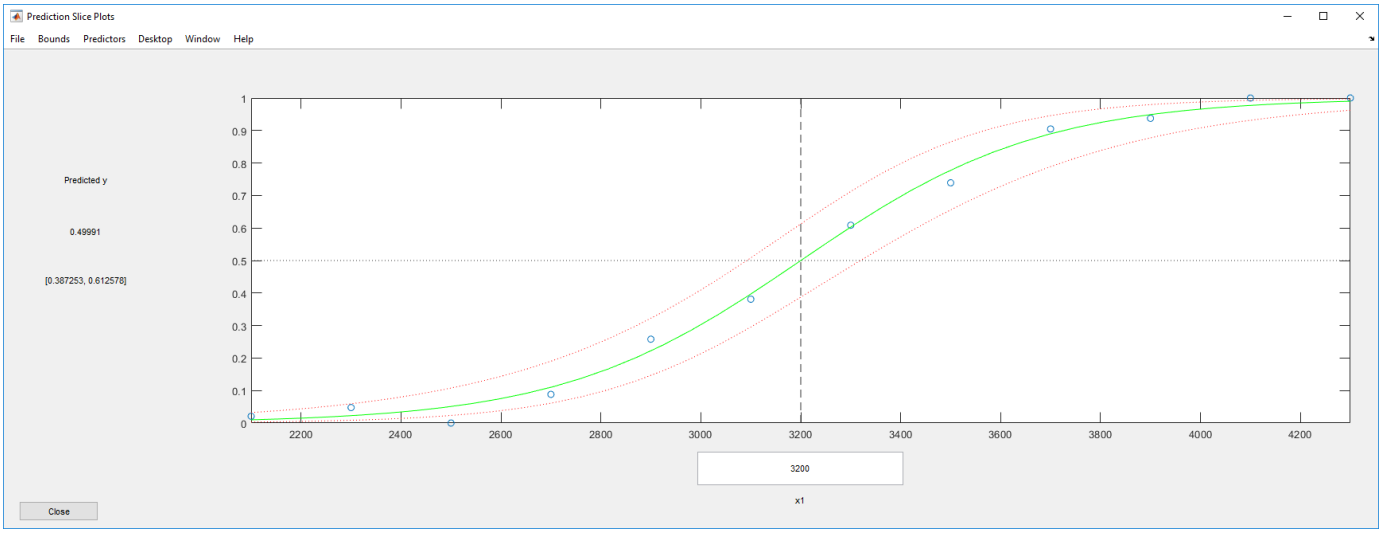
It is reasonable to assume that the values of `poor` follow binomial on page B-10 distributions, with the number of trials given by `total` and the percentage of successes depending on `w`. This distribution can be accounted for in the context of a logistic model by using a generalized linear model with link function $\log(\mu/(1 - \mu)) = Xb$. This link function is called `'logit'`.

```
w = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
total = [48 42 31 34 31 21 23 23 21 16 17 21]';
poor = [1 2 0 3 8 8 14 17 19 15 17 21]';
mdl = fitglm(w,[poor total],...
    'linear','Distribution','binomial','link','logit')
```

```
mdl =

Generalized Linear regression model:
    logit(y) ~ 1 + x1
    Distribution = Binomial

Estimated Coefficients:
                   Estimate     SE            tStat      pValue
    (Intercept)       -13.38         1.394    -9.5986    8.1019e-22
    x1             0.0041812    0.00044258     9.4474    3.4739e-21

12 observations, 10 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 242, p-value = 1.3e-54
```

See how well the model fits the data.

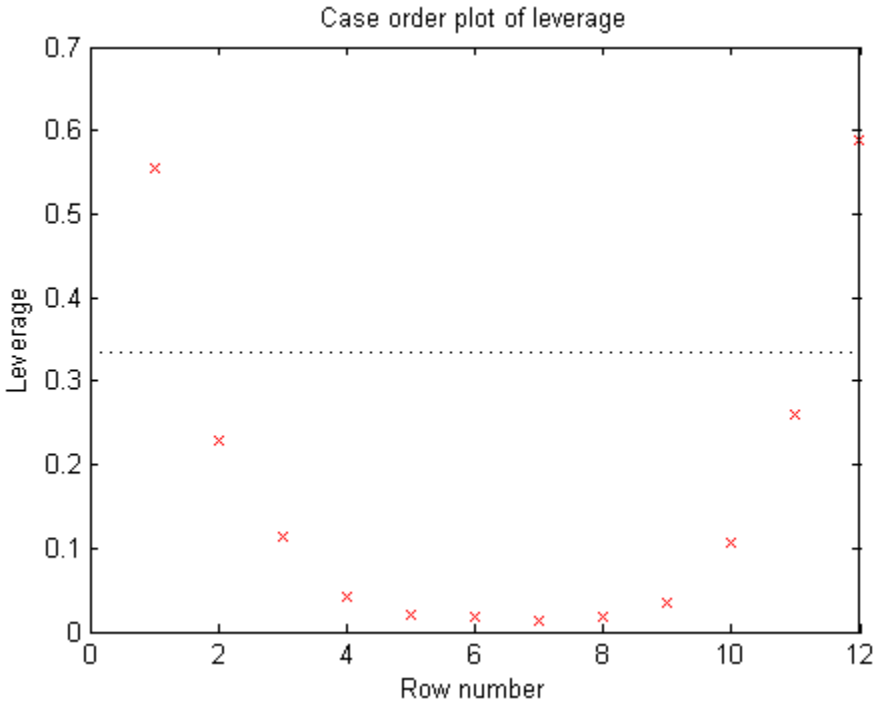
```
plotSlice(mdl)
```



The fit looks reasonably good, with fairly wide confidence bounds.

To examine further details, create a leverage plot.

```
plotDiagnostics(mdl)
```



This is typical of a regression with points ordered by the predictor variable. The leverage of each point on the fit is higher for points with relatively extreme predictor values (in either direction) and low for points with average predictor values. In examples with multiple predictors and with points not ordered by predictor value, this plot can help you identify which observations have high leverage because they are outliers as measured by their predictor values.

Residuals — Model Quality for Training Data

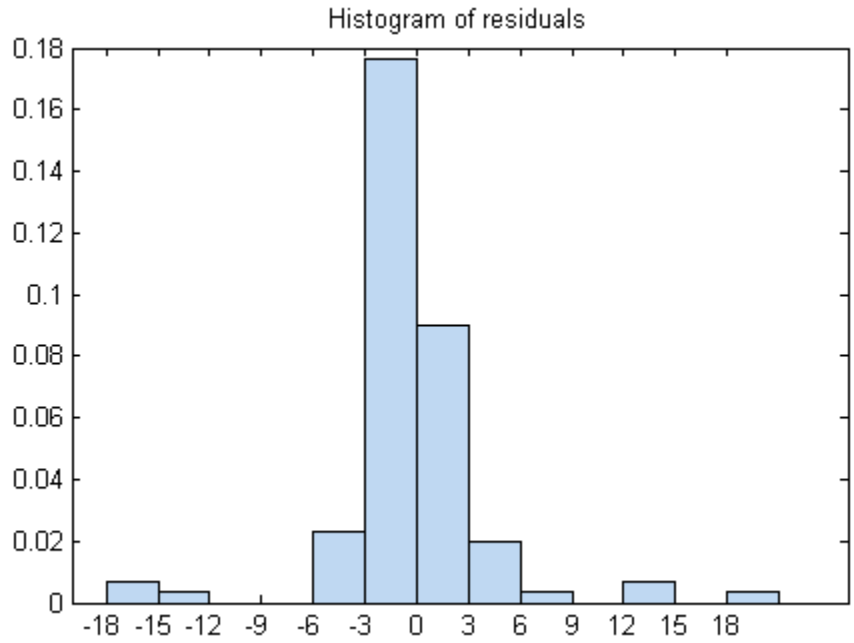
There are several residual plots to help you discover errors, outliers, or correlations in the model or data. The simplest residual plots are the default histogram plot, which shows the range of the residuals and their frequencies, and the probability plot, which shows how the distribution of the residuals compares to a normal distribution with matched variance.

This example shows residual plots for a fitted Poisson model. The data construction has two out of five predictors not affecting the response, and no intercept term:

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
mdl = fitglm(X,y,...
    'linear','Distribution','poisson');
```

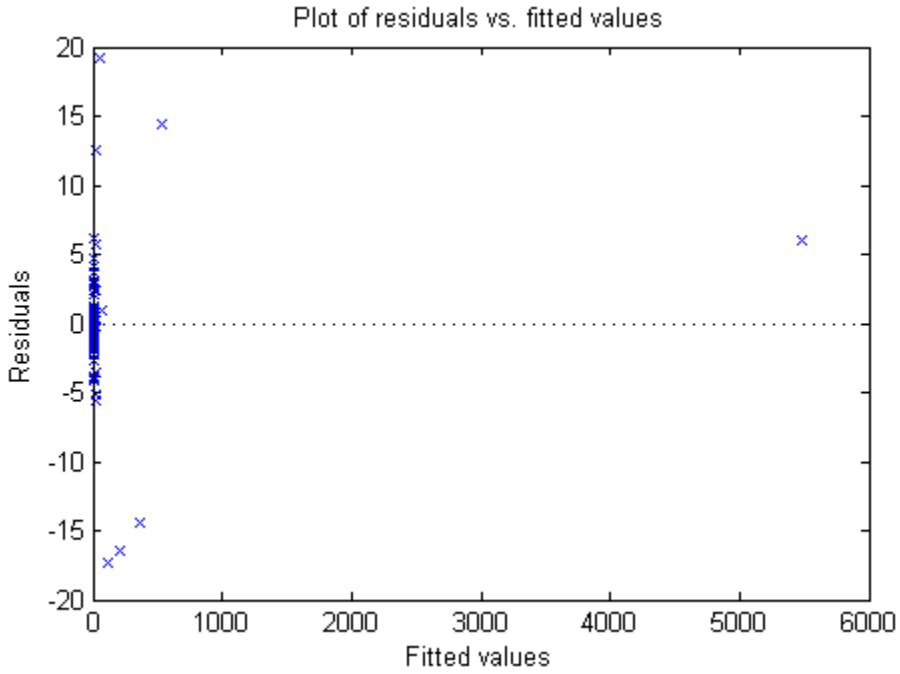
Examine the residuals:

```
plotResiduals(mdl)
```



While most residuals cluster near 0, there are several near ±18. So examine a different residuals plot.

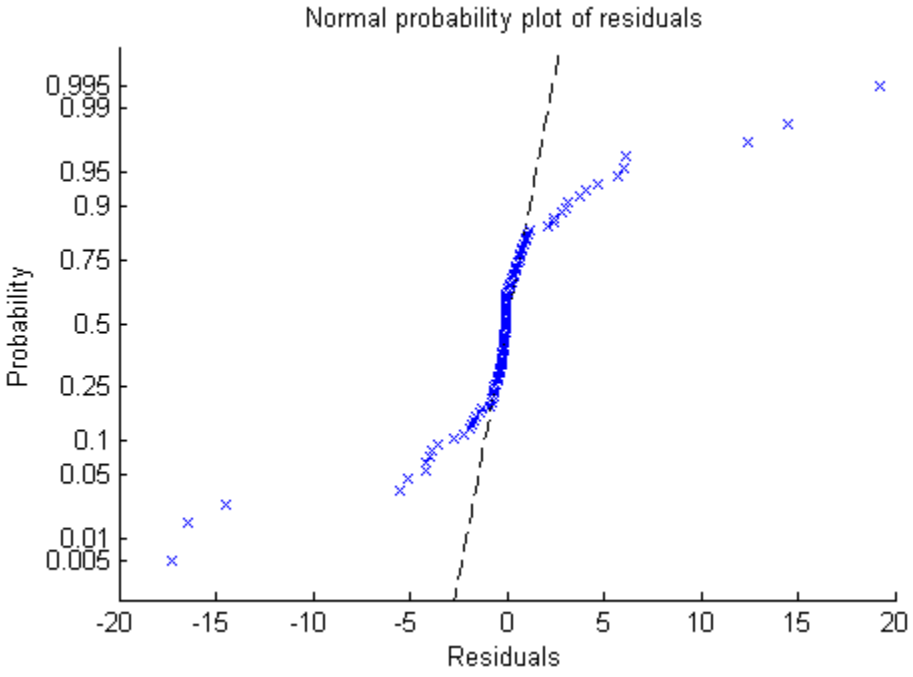
```
plotResiduals(mdl,'fitted')
```



The large residuals don’t seem to have much to do with the sizes of the fitted values.

Perhaps a probability plot is more informative.

```
plotResiduals(mdl,'probability')
```



Now it is clear. The residuals do not follow a normal distribution. Instead, they have fatter tails, much as an underlying Poisson distribution.

Plots to Understand Predictor Effects and How to Modify a Model

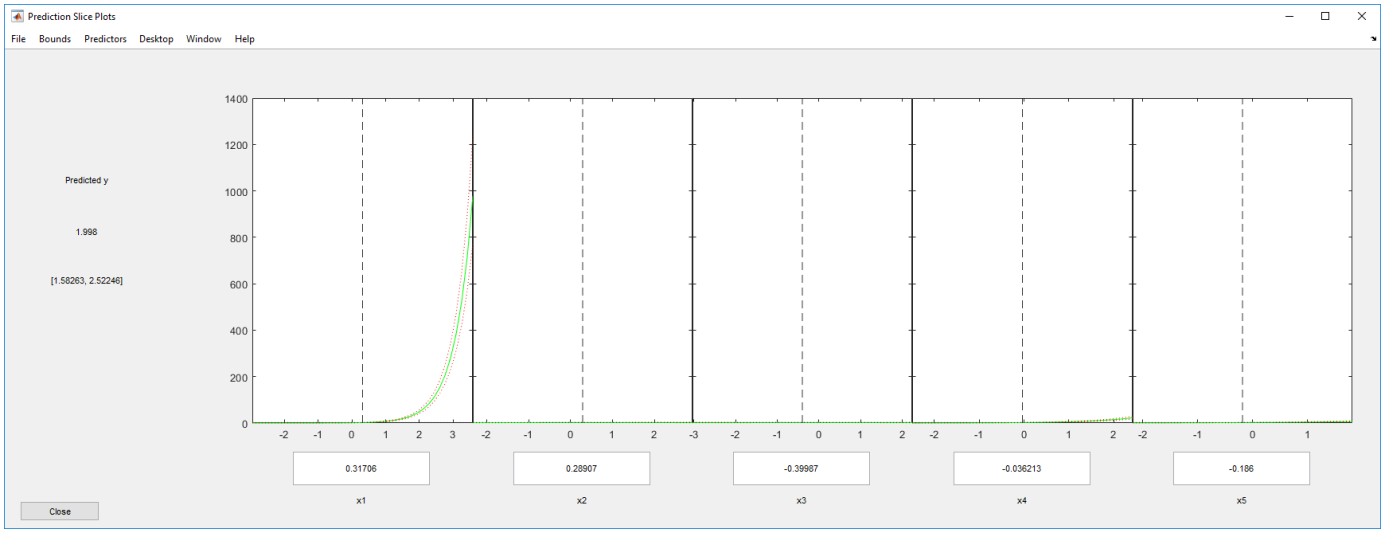
This example shows how to understand the effect each predictor has on a regression model, and how to modify the model to remove unnecessary terms.

1 Create a model from some predictors in artificial data. The data do not use the second and third columns in X. So you expect the model not to show much dependence on those predictors.

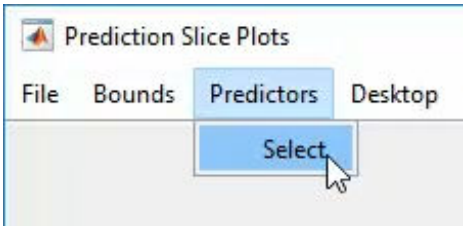
```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
mdl = fitglm(X,y,...
    'linear','Distribution','poisson');
```

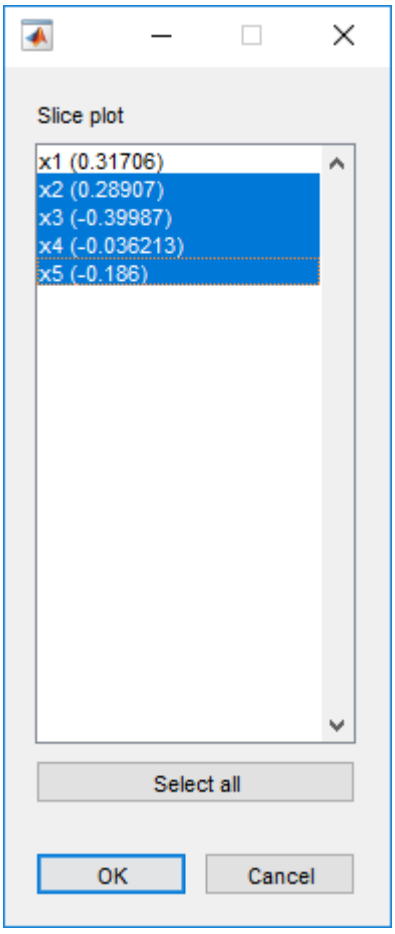
2 Examine a slice plot of the responses. This displays the effect of each predictor separately.

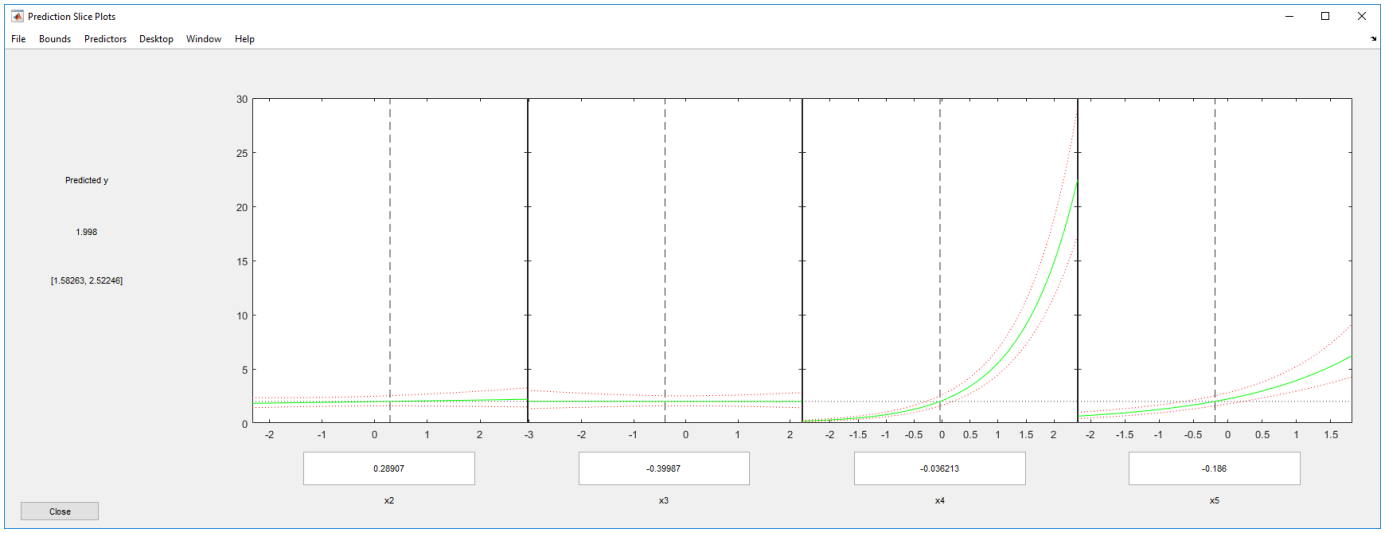
```
plotSlice(mdl)
```



The scale of the first predictor is overwhelming the plot. Disable it using the **Predictors** menu.







Now it is clear that predictors 2 and 3 have little to no effect.

You can drag the individual predictor values, which are represented by dashed blue vertical lines. You can also choose between simultaneous and non-simultaneous confidence bounds, which are

represented by dashed red curves. Dragging the predictor lines confirms that predictors 2 and 3 have little to no effect.

3 Remove the unnecessary predictors using either `removeTerms` or `step`. Using `step` can be safer, in case there is an unexpected importance to a term that becomes apparent after removing another term. However, sometimes `removeTerms` can be effective when `step` does not proceed. In this case, the two give identical results.

```
mdl1 = removeTerms(mdl,'x2 + x3')
```

```
mdl1 =


Generalized Linear regression model:
    log(y) ~ 1 + x1 + x4 + x5
    Distribution = Poisson

Estimated Coefficients:
                   Estimate    SE          tStat     pValue
    (Intercept)    0.17604     0.062215    2.8295       0.004662
    x1              1.9122     0.024638    77.614              0
    x4             0.98521     0.026393    37.328    5.6696e-305
    x5             0.61321     0.038435    15.955     2.6473e-57


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 4.97e+04, p-value = 0
```

```
mdl1 = step(mdl,'NSteps',5,'Upper','linear')
```

```
1. Removing x3, Deviance = 93.856, Chi2Stat = 0.00075551, PValue = 0.97807
2. Removing x2, Deviance = 96.333, Chi2Stat = 2.4769, PValue = 0.11553

mdl1 =


Generalized Linear regression model:
    log(y) ~ 1 + x1 + x4 + x5
    Distribution = Poisson

Estimated Coefficients:
                   Estimate    SE          tStat     pValue
    (Intercept)    0.17604     0.062215    2.8295       0.004662
    x1              1.9122     0.024638    77.614              0
    x4             0.98521     0.026393    37.328    5.6696e-305
    x5             0.61321     0.038435    15.955     2.6473e-57


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 4.97e+04, p-value = 0
```

Predict or Simulate Responses to New Data

There are three ways to use a linear model to predict the response to new data:

- “predict” on page 12-24
- “feval” on page 12-24
- “random” on page 12-25

predict

The `predict` method gives a prediction of the mean responses and, if requested, confidence bounds.

This example shows how to predict and obtain confidence intervals on the predictions using the `predict` method.

1. Create a model from some predictors in artificial data. The data do not use the second and third columns in X. So you expect the model not to show much dependence on these predictors. Construct the model stepwise to include the relevant predictors automatically.

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
mdl = stepwiseglm(X,y,...
    'constant','upper','linear','Distribution','poisson');
```

```
1. Adding x1, Deviance = 2515.02869, Chi2Stat = 47242.9622, PValue = 0
2. Adding x4, Deviance = 328.39679, Chi2Stat = 2186.6319, PValue = 0
3. Adding x5, Deviance = 96.3326, Chi2Stat = 232.0642, PValue = 2.114384e-52
```

2. Generate some new data, and evaluate the predictions from the data.

```
Xnew = randn(3,5) + repmat([1 2 3 4 5],[3,1]); % new data
[ynew,ynewci] = predict(mdl,Xnew)
```

```
ynew =

   1.0e+04 *

    0.1130
    1.7375
    3.7471


ynewci =

   1.0e+04 *

    0.0821    0.1555
    1.2167    2.4811
    2.8419    4.9407
```

feval

When you construct a model from a table or dataset array, `feval` is often more convenient for predicting mean responses than `predict`. However, `feval` does not provide confidence bounds.

This example shows how to predict mean responses using the `feval` method.

1. Create a model from some predictors in artificial data. The data do not use the second and third columns in X. So you expect the model not to show much dependence on these predictors. Construct the model stepwise to include the relevant predictors automatically.

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
```

```
X = array2table(X); % create data table
y = array2table(y);
tbl = [X y];

mdl = stepwiseglm(tbl,...
    'constant','upper','linear','Distribution','poisson');
```

```
1. Adding x1, Deviance = 2515.02869, Chi2Stat = 47242.9622, PValue = 0
2. Adding x4, Deviance = 328.39679, Chi2Stat = 2186.6319, PValue = 0
3. Adding x5, Deviance = 96.3326, Chi2Stat = 232.0642, PValue = 2.114384e-52
```

2 Generate some new data, and evaluate the predictions from the data.

```
Xnew = randn(3,5) + repmat([1 2 3 4 5],[3,1]); % new data
ynew = feval(mdl,Xnew(:,1),Xnew(:,4),Xnew(:,5)) % only need predictors 1,4,5
```

```
ynew =

   1.0e+04 *

    0.1130
    1.7375
    3.7471
```

Equivalently,

```
ynew = feval(mdl,Xnew(:,[1 4 5])) % only need predictors 1,4,5
```

```
ynew =

   1.0e+04 *

    0.1130
    1.7375
    3.7471
```

random

The `random` method generates new random response values for specified predictor values. The distribution of the response values is the distribution used in the model. `random` calculates the mean of the distribution from the predictors, estimated coefficients, and link function. For distributions such as normal, the model also provides an estimate of the variance of the response. For the binomial and Poisson distributions, the variance of the response is determined by the mean; `random` does not use a separate “dispersion” estimate.

This example shows how to simulate responses using the `random` method.

1 Create a model from some predictors in artificial data. The data do not use the second and third columns in `X`. So you expect the model not to show much dependence on these predictors. Construct the model stepwise to include the relevant predictors automatically.

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
mdl = stepwiseglm(X,y,...
    'constant','upper','linear','Distribution','poisson');
```

```
1. Adding x1, Deviance = 2515.02869, Chi2Stat = 47242.9622, PValue = 0
2. Adding x4, Deviance = 328.39679, Chi2Stat = 2186.6319, PValue = 0
3. Adding x5, Deviance = 96.3326, Chi2Stat = 232.0642, PValue = 2.114384e-52
```

2 Generate some new data, and evaluate the predictions from the data.

```
Xnew = randn(3,5) + repmat([1 2 3 4 5],[3,1]); % new data
ysim = random(mdl,Xnew)

ysim =

         1111
        17121
        37457
```

The predictions from `random` are Poisson samples, so are integers.

3 Evaluate the `random` method again, the result changes.

```
ysim = random(mdl,Xnew)

ysim =

         1175
        17320
        37126
```

Share Fitted Models

The model display contains enough information to enable someone else to recreate the model in a theoretical sense. For example,

```
rng('default') % for reproducibility
X = randn(100,5);
mu = exp(X(:,[1 4 5])*[2;1;.5]);
y = poissrnd(mu);
mdl = stepwiseglm(X,y,...
    'constant','upper','linear','Distribution','poisson')

1. Adding x1, Deviance = 2515.02869, Chi2Stat = 47242.9622, PValue = 0
2. Adding x4, Deviance = 328.39679, Chi2Stat = 2186.6319, PValue = 0
3. Adding x5, Deviance = 96.3326, Chi2Stat = 232.0642, PValue = 2.114384e-52

mdl =


Generalized Linear regression model:
    log(y) ~ 1 + x1 + x4 + x5
    Distribution = Poisson

Estimated Coefficients:
                   Estimate    SE          tStat     pValue
    (Intercept)    0.17604     0.062215    2.8295       0.004662
    x1              1.9122     0.024638    77.614             0
    x4             0.98521     0.026393    37.328    5.6696e-305
    x5             0.61321     0.038435    15.955     2.6473e-57


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 4.97e+04, p-value = 0
```

You can access the model description programmatically, too. For example,

```
mdl.Coefficients.Estimate
```

```
ans =

    0.1760
    1.9122
    0.9852
    0.6132
```

```
mdl.Formula
```

```
ans =

log(y) ~ 1 + x1 + x4 + x5
```

References

[1] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

[2] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[3] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[4] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, Fourth Edition. Irwin, Chicago, 1996.

Generalized Linear Model Workflow

This example shows how to fit a generalized linear model and analyze the results. A typical workflow involves these steps: import data, fit a generalized linear model, test its quality, modify the model to improve its quality, and make predictions based on the model. In this example, you use the Fisher iris data to compute the probability that a flower is in one of two classes.

Load Data

Load the Fisher iris data.

```
load fisheriris
```

Extract rows 51 to 150, which have the classification versicolor or virginica.

```
X = meas(51:end,:);
```

Create logical response variables that are `true` for versicolor and `false` for virginica.

```
y = strcmp('versicolor',species(51:end));
```

Fit Generalized Linear Model

Fit a binomial generalized linear model to the data.

```
mdl = fitglm(X,y,'linear','Distribution','binomial')
```

```
mdl =
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    ______    _______    ________

    (Intercept)     42.638     25.708     1.6586    0.097204
    x1              2.4652     2.3943     1.0296     0.30319
    x2              6.6809     4.4796     1.4914     0.13585
    x3             -9.4294     4.7372    -1.9905    0.046537
    x4             -18.286     9.7426    -1.8769    0.060529


100 observations, 95 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 127, p-value = 1.95e-26
```

According to the model display, some *p*-values in the `pValue` column are not small, which implies that you can simplify the model.

Examine and Improve Model

Determine if 95% confidence intervals for the coefficients include 0. If so, you can remove the model terms with those intervals.

```
confint = coefCI(mdl)
```

```
confint = 5×2
```

```
   -8.3984   93.6740
   -2.2881    7.2185
   -2.2122   15.5739
  -18.8339   -0.0248
  -37.6277    1.0554
```

Only the fourth predictor x3 has a coefficient whose confidence interval does not include 0.

The coefficients of x1 and x2 have large *p*-values and their 95% confidence intervals include 0. Test whether both coefficients can be zero. Specify a hypothesis matrix to select the coefficients of x1 and x2.

```
M = [0 1 0 0 0
     0 0 1 0 0];
p = coefTest(mdl,M)
```

```
p = 0.1442
```

The *p*-value is approximately 0.14, which is not small. Remove x1 and x2 from the model.

```
mdl1 = removeTerms(mdl,'x1 + x2')
```

```
mdl1 =
Generalized linear regression model:
    logit(y) ~ 1 + x3 + x4
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    ______    _______    __________

    (Intercept)     45.272     13.612      3.326    0.00088103
    x3             -5.7545     2.3059    -2.4956      0.012576
    x4             -10.447     3.7557    -2.7816     0.0054092


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 118, p-value = 2.3e-26
```

Alternatively, you can identify important predictors using stepwiseglm.

```
mdl2 = stepwiseglm(X,y,'constant','Distribution','binomial','Upper','linear')
```

```
1. Adding x4, Deviance = 33.4208, Chi2Stat = 105.2086, PValue = 1.099298e-24
2. Adding x3, Deviance = 20.5635, Chi2Stat = 12.8573, PValue = 0.000336166
3. Adding x2, Deviance = 13.2658, Chi2Stat = 7.29767, PValue = 0.00690441

mdl2 =
Generalized linear regression model:
    logit(y) ~ 1 + x2 + x3 + x4
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    ______    _______    ________

    (Intercept)     50.527     23.995     2.1057    0.035227
```

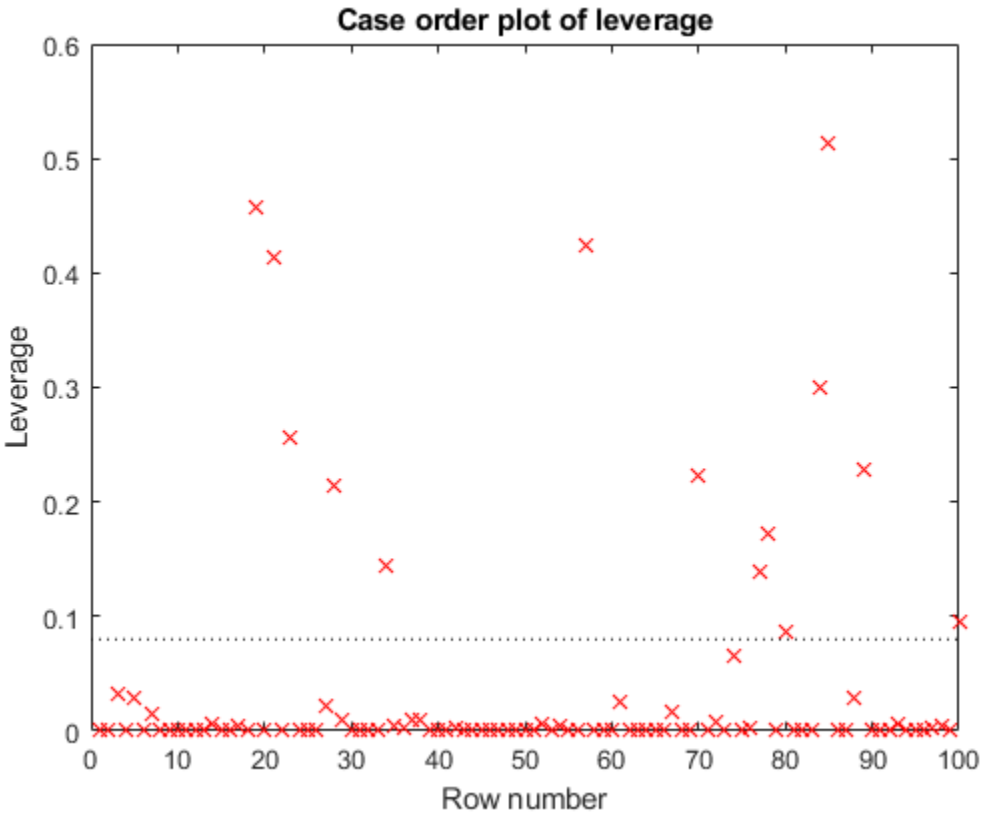
```
    x2              8.3761     4.7612     1.7592    0.078536
    x3             -7.8745     3.8407    -2.0503    0.040334
    x4              -21.43     10.707    -2.0014     0.04535


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 125, p-value = 5.4e-27
```

The *p*-value (`pValue`) for `x2` in the coefficient table is greater than 0.05, but `stepwiseglm` includes `x2` in the model because the *p*-value (`PValue`) for adding `x2` is smaller than 0.05. The `stepwiseglm` function computes `PValue` using the fits with and without `x2`, whereas the function computes `pValue` based on an approximate standard error computed only from the final model. Therefore, `PValue` is more reliable than `pValue`.

Identify Outliers

Examine a leverage plot to look for influential outliers.

```
plotDiagnostics(mdl2,'leverage')
```



An observation can be considered an outlier if its leverage substantially exceeds `p/n`, where `p` is the number of coefficients and `n` is the number of observations. The dotted reference line is a recommended threshold, computed by `2*p/n`, which corresponds to 0.08 in this plot. Some observations have leverage values larger than `10*p/n` (that is, 0.40). Identify these observation points.

```
idxOutliers = find(mdl2.Diagnostics.Leverage > 10*mdl2.NumCoefficients/mdl2.NumObservations)
```

```
idxOutliers = 4×1

    19
    21
    57
    85
```

See if the model coefficients change when you fit a model excluding these points.

```
oldCoeffs = mdl2.Coefficients.Estimate;
mdl3 = fitglm(X,y,'linear','Distribution','binomial', ...
    'PredictorVars',2:4,'Exclude',idxOutliers);
newCoeffs = mdl3.Coefficients.Estimate;
disp([oldCoeffs newCoeffs])
```

```
   50.5268   44.0085
    8.3761    5.6361
   -7.8745   -6.1145
  -21.4296  -18.1236
```

The model coefficients in `mdl3` are different from those in mdl2. This result implies that the responses at the high-leverage points are not consistent with the predicted values from the reduced model.

Predict Probability of Being Versicolor

Use `mdl3` to predict the probability that a flower with average measurements is versicolor. Generate confidence intervals for the prediction.

```
[newf,newc] = predict(mdl3,mean(X))
```

```
newf = 0.4558
```

```
newc = 1×2

    0.1234    0.8329
```

The model gives almost a 46% probability that the average flower is versicolor, with a wide confidence interval.

See Also

`GeneralizedLinearModel` | `coefCI` | `fitglm` | `plotDiagnostics` | `predict` | `removeTerms` | `stepwiseglm`

More About

- “Generalized Linear Models” on page 12-9

Lasso Regularization of Generalized Linear Models

In this section...

What is Generalized Linear Model Lasso Regularization?

Lasso is a regularization technique. Use `lassoglm` to:

- Reduce the number of predictors in a generalized linear model.
- Identify important predictors.
- Select among redundant predictors.
- Produce shrinkage estimates with potentially lower predictive errors than ordinary least squares.

Elastic net is a related technique. Use it when you have several highly correlated variables. `lassoglm` provides elastic net regularization when you set the `Alpha` name-value pair to a number strictly between `0` and `1`.

For details about lasso and elastic net computations and algorithms, see “Generalized Linear Model Lasso and Elastic Net” on page 12-32. For a discussion of generalized linear models, see “What Are Generalized Linear Models?” on page 12-9.

Generalized Linear Model Lasso and Elastic Net

Overview of Lasso and Elastic Net

Lasso is a regularization technique for estimating generalized linear models. Lasso includes a penalty term that constrains the size of the estimated coefficients. Therefore, it resembles “Ridge Regression” on page 11-109. Lasso is a shrinkage estimator: it generates coefficient estimates that are biased to be small. Nevertheless, a lasso estimator can have smaller error than an ordinary maximum likelihood estimator when you apply it to new data.

Unlike ridge regression, as the penalty term increases, the lasso technique sets more coefficients to zero. This means that the lasso estimator is a smaller model, with fewer predictors. As such, lasso is an alternative to stepwise regression on page 11-99 and other model selection and dimensionality reduction techniques.

Elastic net is a related technique. Elastic net is akin to a hybrid of ridge regression and lasso regularization. Like lasso, elastic net can generate reduced models by generating zero-valued coefficients. Empirical studies suggest that the elastic net technique can outperform lasso on data with highly correlated predictors.

Definition of Lasso for Generalized Linear Models

For a nonnegative value of λ, `lassoglm` solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\text{Deviance}(\beta_0,\beta)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right).$$

- The function Deviance in this equation is the deviance of the model fit to the responses using the intercept β_0 and the predictor coefficients β. The formula for Deviance depends on the `distr` parameter you supply to `lassoglm`. Minimizing the λ-penalized deviance is equivalent to maximizing the λ-penalized loglikelihood.
- N is the number of observations.
- λ is a nonnegative regularization parameter corresponding to one value of `Lambda`.
- The parameters β_0 and β are a scalar and a vector of length p, respectively.

As λ increases, the number of nonzero components of β decreases.

The lasso problem involves the L^1 norm of β, as contrasted with the elastic net algorithm.

Definition of Elastic Net for Generalized Linear Models

For α strictly between 0 and 1, and nonnegative λ, elastic net solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\text{Deviance}(\beta_0,\beta)+\lambda P_\alpha(\beta)\right),$$

where

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha\left|\beta_j\right|\right).$$

Elastic net is the same as lasso when $\alpha = 1$. For other values of α, the penalty term $P_\alpha(\beta)$ interpolates between the L^1 norm of β and the squared L^2 norm of β. As α shrinks toward 0, elastic net approaches `ridge` regression.

References

[1] Tibshirani, R. *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, Series B, Vol. 58, No. 1, pp. 267–288, 1996.

[2] Zou, H. and T. Hastie. *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society, Series B, Vol. 67, No. 2, pp. 301–320, 2005.

[3] Friedman, J., R. Tibshirani, and T. Hastie. *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, Vol. 33, No. 1, 2010. `https://www.jstatsoft.org/v33/i01`

[4] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning,* 2nd edition. Springer, New York, 2008.

[5] McCullagh, P., and J. A. Nelder. *Generalized Linear Models,* 2nd edition. Chapman & Hall/CRC Press, 1989.

Regularize Poisson Regression

This example shows how to identify and remove redundant predictors from a generalized linear model.

Create data with 20 predictors, and Poisson responses using just three of the predictors, plus a constant.

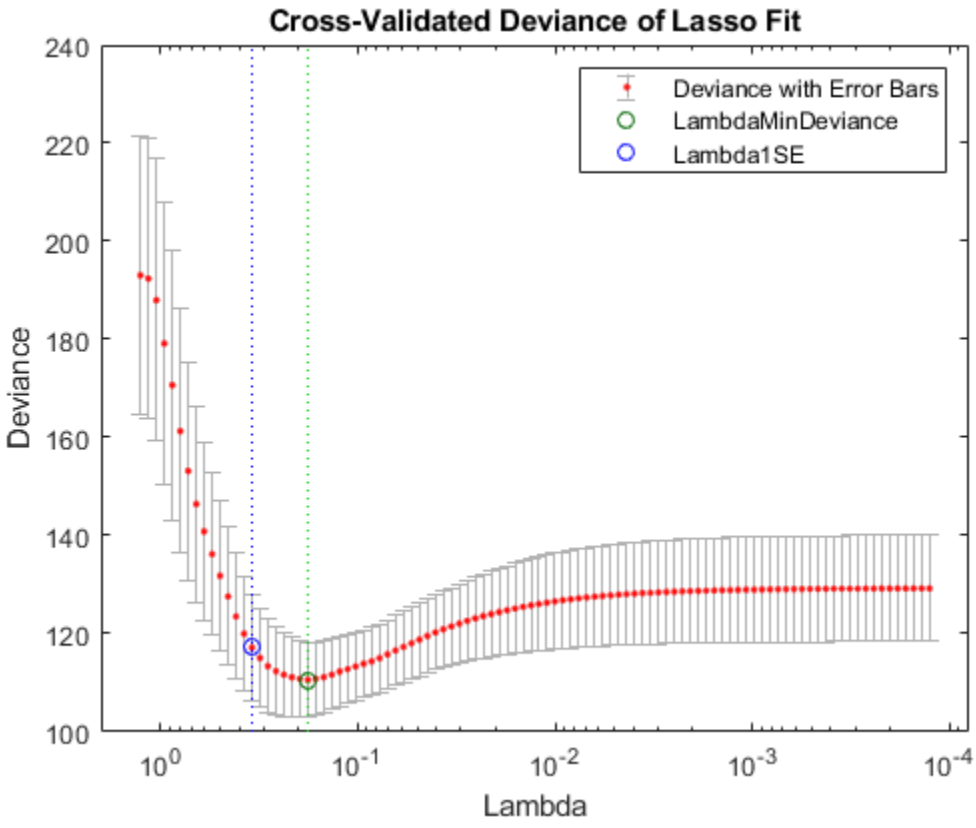
```
rng('default') % for reproducibility
X = randn(100,20);
mu = exp(X(:,[5 10 15])*[.4;.2;.3] + 1);
y = poissrnd(mu);
```

Construct a cross-validated lasso regularization of a Poisson regression model of the data.

```
[B,FitInfo] = lassoglm(X,y,'poisson','CV',10);
```

Examine the cross-validation plot to see the effect of the `Lambda` regularization parameter.

```
lassoPlot(B,FitInfo,'plottype','CV');
legend('show') % show legend
```



The green circle and dashed line locate the `Lambda` with minimal cross-validation error. The blue circle and dashed line locate the point with minimal cross-validation error plus one standard deviation.

Find the nonzero model coefficients corresponding to the two identified points.

```
minpts = find(B(:,FitInfo.IndexMinDeviance))
```

```
minpts = 7×1

     3
     5
     6
    10
    11
    15
    16
```

```
min1pts = find(B(:,FitInfo.Index1SE))
```

```
min1pts = 3×1

     5
    10
    15
```

The coefficients from the minimal plus one standard error point are exactly those coefficients used to create the data.

Find the values of the model coefficients at the minimal plus one standard error point.

```
B(min1pts,FitInfo.Index1SE)
```

```
ans = 3×1

    0.2903
    0.0789
    0.2081
```

The values of the coefficients are, as expected, smaller than the original `[0.4,0.2,0.3]`. Lasso works by "shrinkage," which biases predictor coefficients toward zero.

The constant term is in the `FitInfo.Intercept` vector.

```
FitInfo.Intercept(FitInfo.Index1SE)
```

```
ans = 1.0879
```

The constant term is near 1, which is the value used to generate the data.

Regularize Logistic Regression

This example shows how to regularize binomial regression. The default (canonical) link function for binomial regression is the logistic function.

Step 1. Prepare the data.

Load the `ionosphere` data. The response `Y` is a cell array of `'g'` or `'b'` characters. Convert the cells to logical values, with `true` representing `'g'`. Remove the first two columns of `X` because they have some awkward statistical properties, which are beyond the scope of this discussion.

```
load ionosphere
Ybool = strcmp(Y,'g');
X = X(:,3:end);
```

Step 2. Create a cross-validated fit.

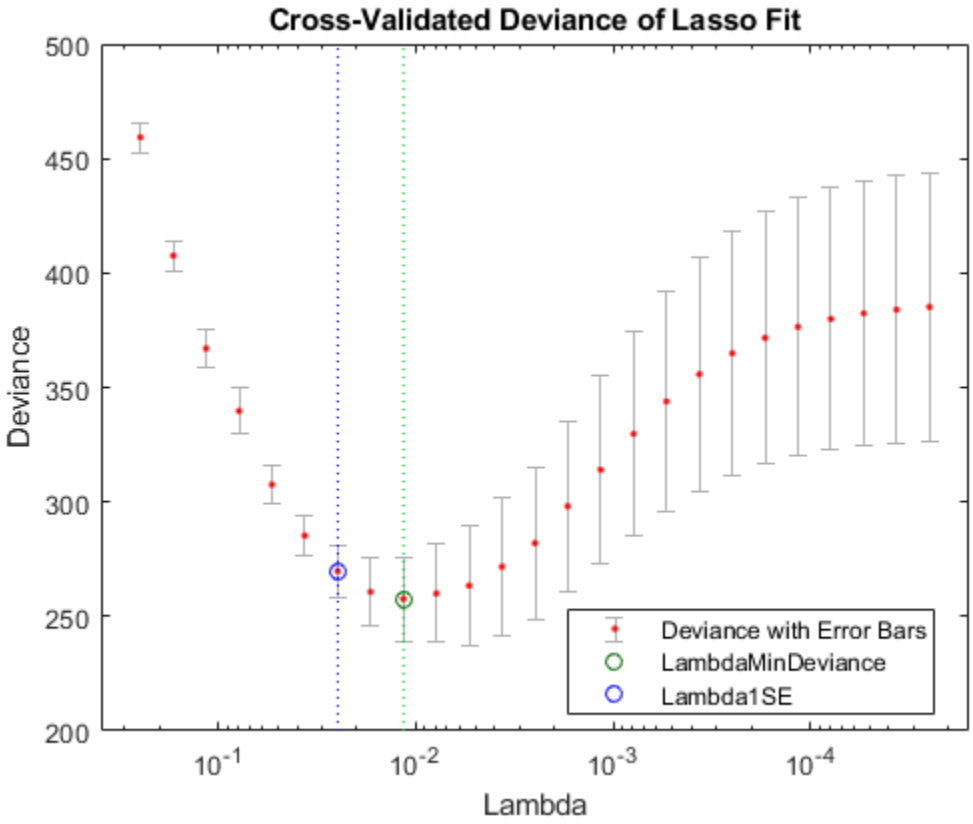
Construct a regularized binomial regression using 25 `Lambda` values and 10-fold cross validation. This process can take a few minutes.

```
rng('default') % for reproducibility
[B,FitInfo] = lassoglm(X,Ybool,'binomial',...
    'NumLambda',25,'CV',10);
```

Step 3. Examine plots to find appropriate regularization.

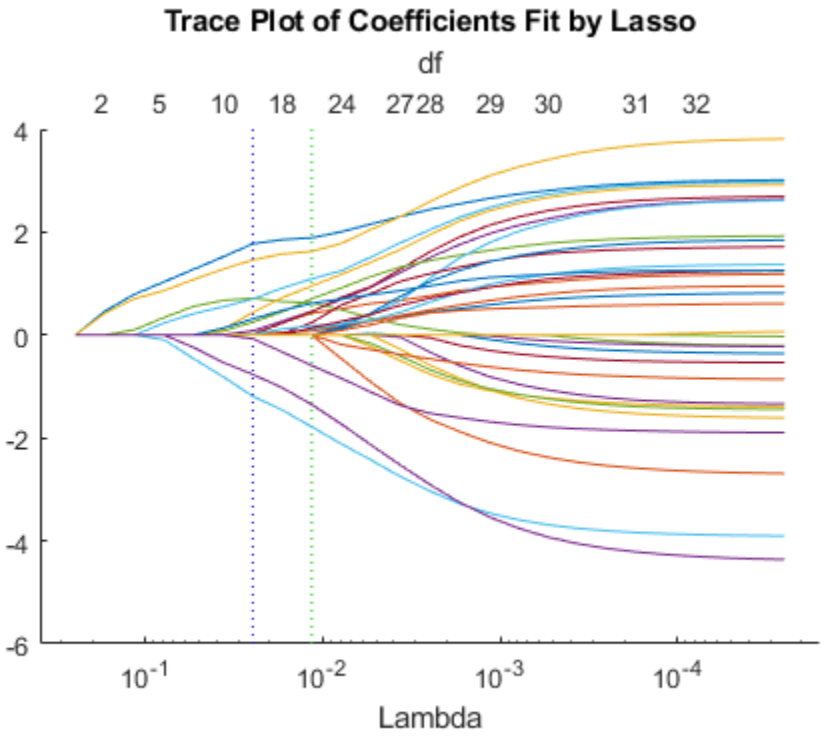
`lassoPlot` can give both a standard trace plot and a cross-validated deviance plot. Examine both plots.

```
lassoPlot(B,FitInfo,'PlotType','CV');
legend('show','Location','best') % show legend
```



The plot identifies the minimum-deviance point with a green circle and dashed line as a function of the regularization parameter `Lambda`. The blue circled point has minimum deviance plus no more than one standard deviation.

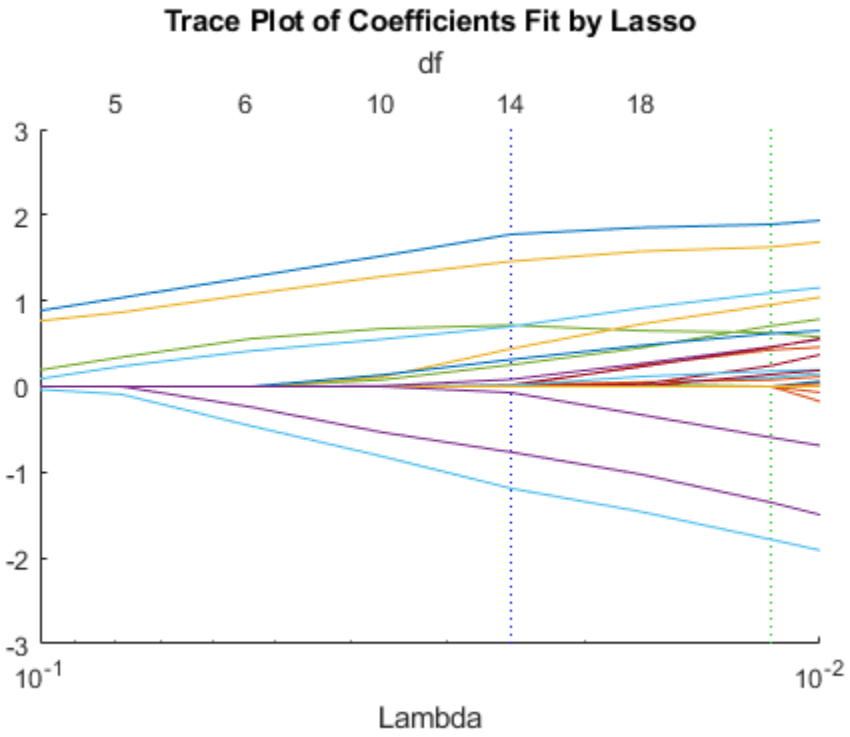
```
lassoPlot(B,FitInfo,'PlotType','Lambda','XScale','log');
```



The trace plot shows nonzero model coefficients as a function of the regularization parameter `Lambda`. Because there are 32 predictors and a linear model, there are 32 curves. As `Lambda` increases to the left, `lassoglm` sets various coefficients to zero, removing them from the model.

The trace plot is somewhat compressed. Zoom in to see more detail.

```
xlim([.01 .1])
ylim([-3 3])
```



As `Lambda` increases toward the left side of the plot, fewer nonzero coefficients remain.

Find the number of nonzero model coefficients at the `Lambda` value with minimum deviance plus one standard deviation point. The regularized model coefficients are in column `FitInfo.Index1SE` of the `B` matrix.

```
indx = FitInfo.Index1SE;
B0 = B(:,indx);
nonzeros = sum(B0 ~= 0)
```

```
nonzeros =

    14
```

When you set `Lambda` to `FitInfo.Index1SE`, `lassoglm` removes over half of the 32 original predictors.

Step 4. Create a regularized model.

The constant term is in the `FitInfo.Index1SE` entry of the `FitInfo.Intercept` vector. Call that value `cnst`.

The model is logit(mu) = log(mu/(1 - mu)) = `X*B0 + cnst` . Therefore, for predictions, mu = `exp(X*B0 + cnst)/(1+exp(x*B0 + cnst))`.

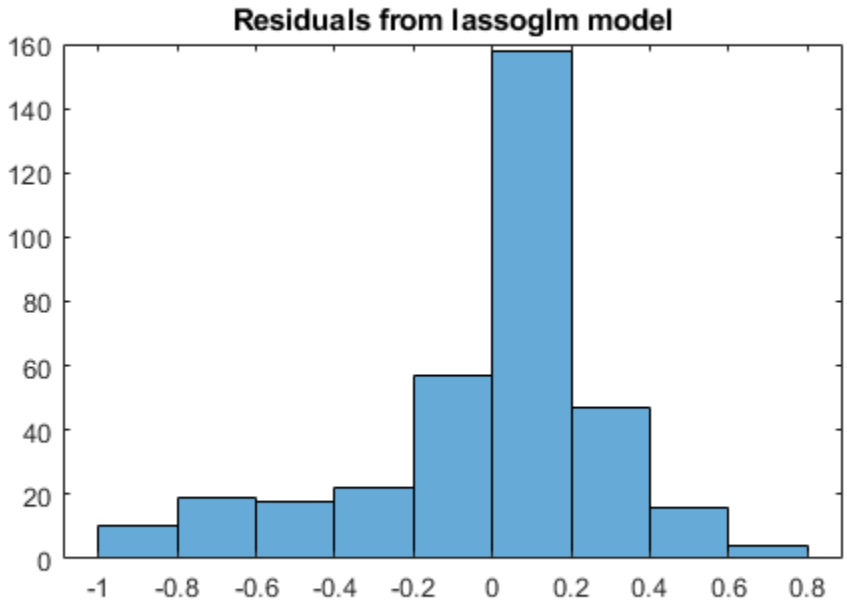
The `glmval` function evaluates model predictions. It assumes that the first model coefficient relates to the constant term. Therefore, create a coefficient vector with the constant term first.

```
cnst = FitInfo.Intercept(indx);
B1 = [cnst;B0];
```

Step 5. Examine residuals.

Plot the training data against the model predictions for the regularized `lassoglm` model.

```
preds = glmval(B1,X,'logit');
histogram(Ybool - preds) % plot residuals
title('Residuals from lassoglm model')
```



Step 6. Alternative: Use identified predictors in a least-squares generalized linear model.

Instead of using the biased predictions from the model, you can make an unbiased model using just the identified predictors.

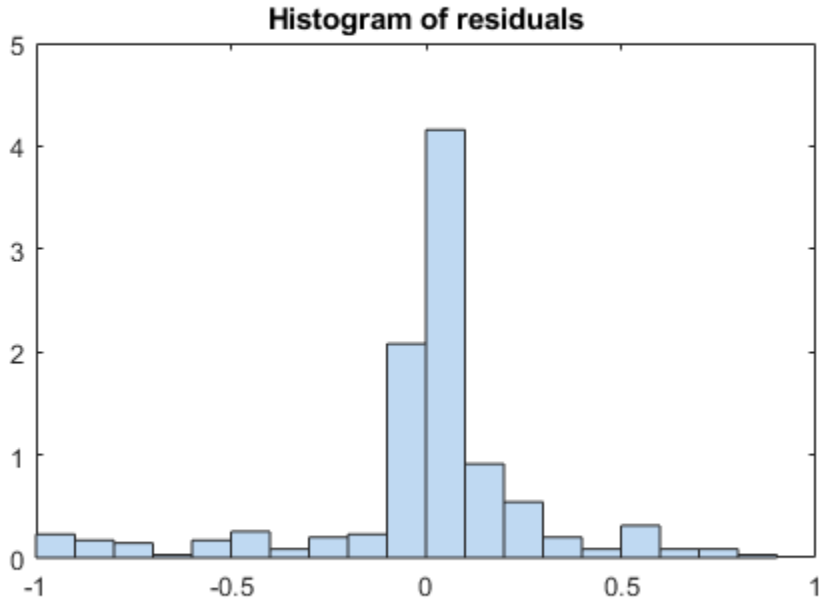
```
predictors = find(B0); % indices of nonzero predictors
mdl = fitglm(X,Ybool,'linear',...
    'Distribution','binomial','PredictorVars',predictors)
```

```
mdl =


Generalized linear regression model:
```

```
    y ~ [Linear formula with 15 terms in 14 predictors]
    Distribution = Binomial

Estimated Coefficients:
                   Estimate       SE        tStat        pValue
                   ________    _______    ________    __________

    (Intercept)      -2.9367    0.50926     -5.7666    8.0893e-09
    x1                 2.492    0.60795       4.099    4.1502e-05
    x3                2.5501    0.63304      4.0284     5.616e-05
    x4               0.48816    0.50336      0.9698       0.33215
    x5                0.6158    0.62192     0.99015        0.3221
    x6                 2.294     0.5421      4.2317    2.3198e-05
    x7               0.77842    0.57765      1.3476        0.1778
    x12               1.7808    0.54316      3.2786     0.0010432
    x16            -0.070993    0.50515    -0.14054       0.88823
    x20              -2.7767    0.55131     -5.0365    4.7402e-07
    x24               2.0212    0.57639      3.5067    0.00045372
    x25              -2.3796    0.58274     -4.0835    4.4363e-05
    x27              0.79564    0.55904      1.4232       0.15467
    x29               1.2689    0.55468      2.2876      0.022162
    x32              -1.5681    0.54336     -2.8859     0.0039035


351 observations, 336 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 262, p-value = 1e-47
```

Plot the residuals of the model.

```
plotResiduals(mdl)
```



As expected, residuals from the least-squares model are slightly smaller than those of the regularized model. However, this does not mean that `mdl` is a better predictor for new data.

Regularize Wide Data in Parallel

This example shows how to regularize a model with many more predictors than observations. *Wide data* is data with more predictors than observations. Typically, with wide data you want to identify important predictors. Use `lassoglm` as an exploratory or screening tool to select a smaller set of variables to prioritize your modeling and research. Use parallel computing to speed up cross validation.

Load the `ovariancancer` data. This data has 216 observations and 4000 predictors in the `obs` workspace variable. The responses are binary, either `'Cancer'` or `'Normal'`, in the `grp` workspace variable. Convert the responses to binary for use in `lassoglm`.

```
load ovariancancer
y = strcmp(grp,'Cancer');
```

Set options to use parallel computing. Prepare to compute in parallel using `parpool`.

```
opt = statset('UseParallel',true);
parpool()
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

ans =

 ProcessPool with properties:

            Connected: true
           NumWorkers: 6
              Cluster: local
        AttachedFiles: {}
    AutoAddClientPath: true
          IdleTimeout: 30 minutes (30 minutes remaining)
          SpmdEnabled: true
```

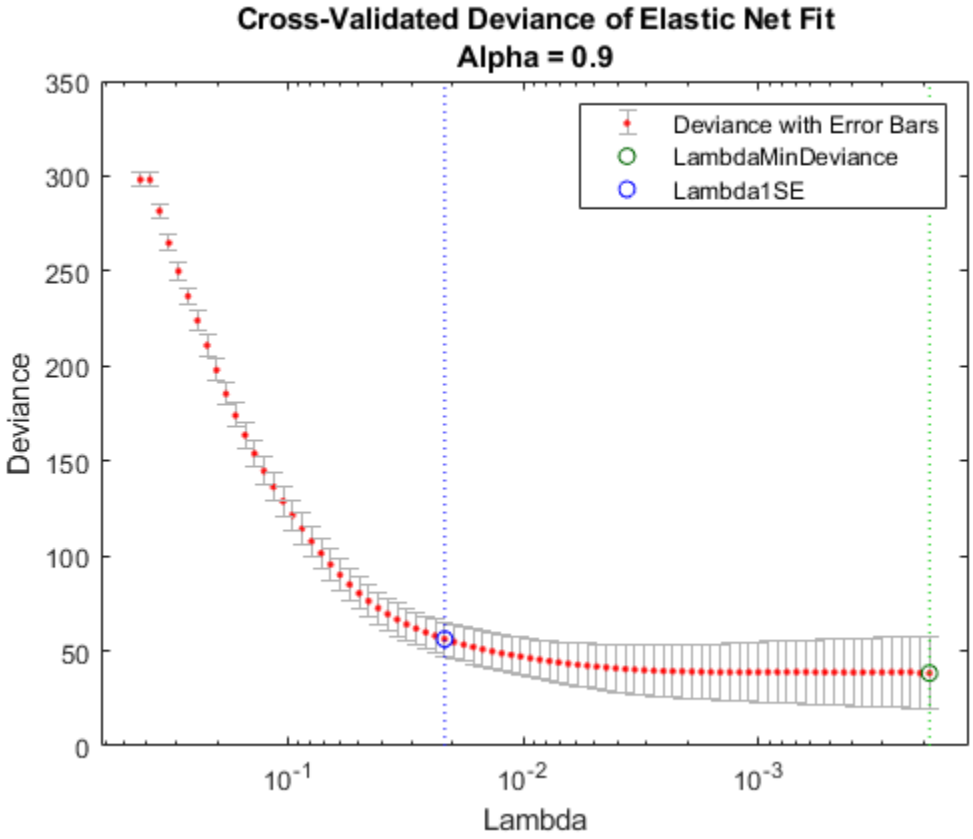
Fit a cross-validated set of regularized models. Use the `Alpha` parameter to favor retaining groups of highly correlated predictors, as opposed to eliminating all but one member of the group. Commonly, you use a relatively large value of `Alpha`.

```
rng('default') % For reproducibility
tic
[B,S] = lassoglm(obs,y,'binomial','NumLambda',100, ...
  'Alpha',0.9,'LambdaRatio',1e-4,'CV',10,'Options',opt);
toc
```

```
Elapsed time is 90.892114 seconds.
```

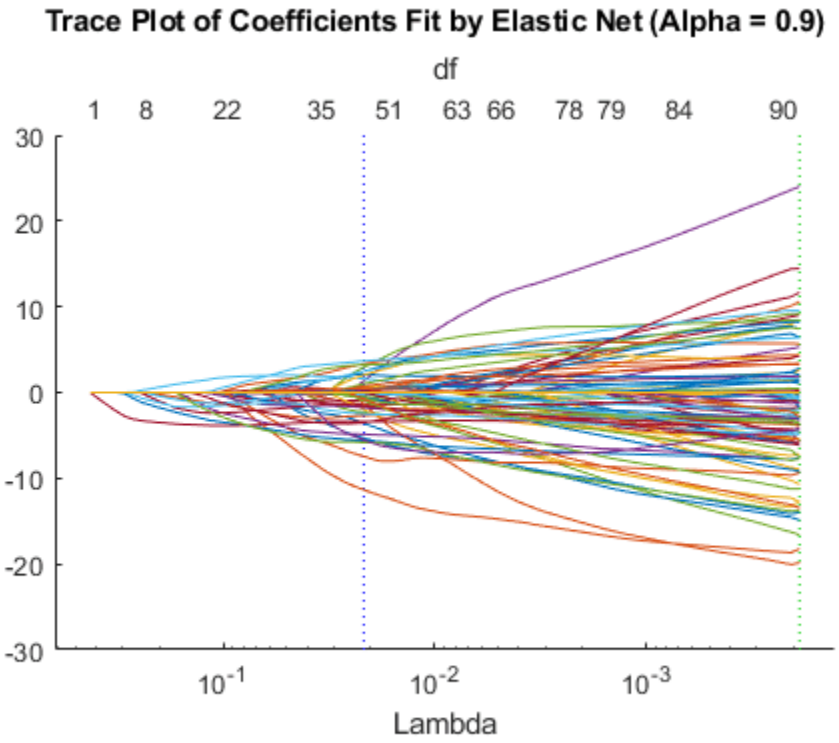
Examine cross-validation plot.

```
lassoPlot(B,S,'PlotType','CV');
legend('show') % Show legend
```



Examine trace plot.

```
lassoPlot(B,S,'PlotType','Lambda','XScale','log')
```



The right (green) vertical dashed line represents the `Lambda` providing the smallest cross-validated deviance. The left (blue) dashed line has the minimal deviance plus no more than one standard deviation. This blue line has many fewer predictors:

```
[S.DF(S.Index1SE) S.DF(S.IndexMinDeviance)]
```

```
ans = 1×2

    50    89
```

You asked `lassoglm` to fit using 100 different `Lambda` values. How many did it use?

```
size(B)
```

```
ans = 1×2

        4000          84
```

`lassoglm` stopped after 84 values because the deviance was too small for small `Lambda` values. To avoid overfitting, `lassoglm` halts when the deviance of the fitted model is too small compared to the deviance in the binary responses, ignoring the predictor variables.

You can force `lassoglm` to include more terms by using the `'Lambda'` name-value pair argument. For example, define a set of `Lambda` values that additionally includes three values smaller than the values in `S.Lambda`.

```
minLambda = min(S.Lambda);
explicitLambda = [minLambda*[.1 .01 .001] S.Lambda];
```

Specify `'Lambda',explicitLambda` when you call the `lassoglm` function. `lassoglm` halts when the deviance of the fitted model is too small, even though you explicitly provide a set of `Lambda` values.

To save time, you can use:

- Fewer `Lambda`, meaning fewer fits
- Fewer cross-validation folds
- A larger value for `LambdaRatio`

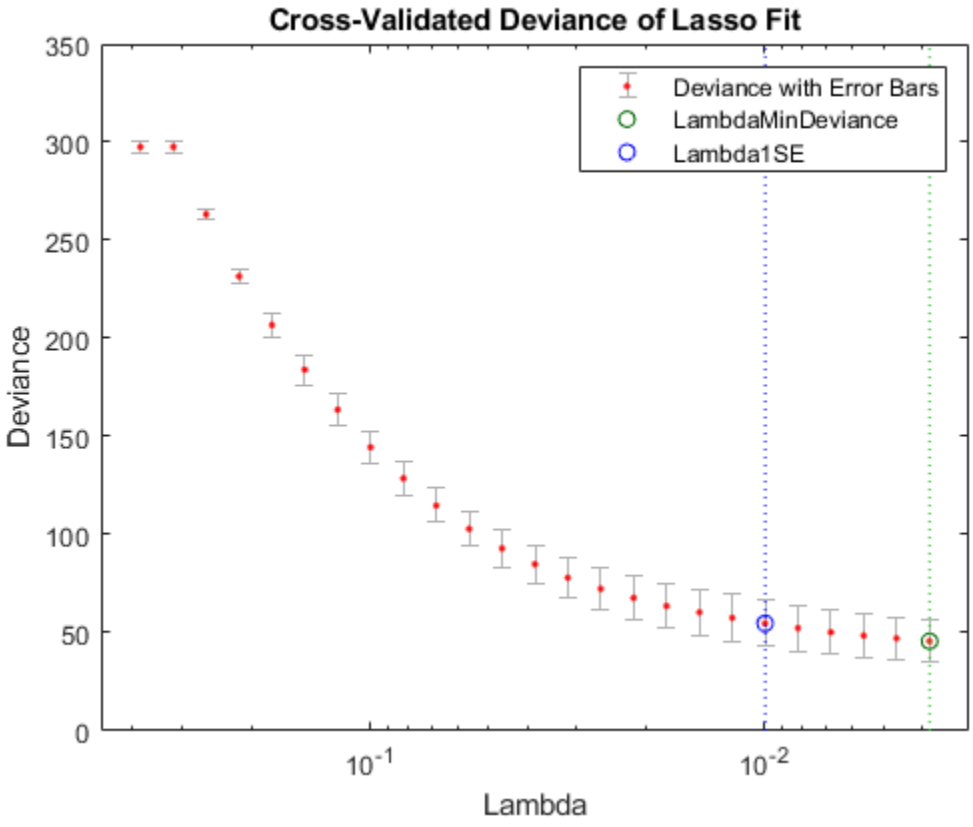
Use serial computation and all three of these time-saving methods:

```
tic
[Bquick,Squick] = lassoglm(obs,y,'binomial','NumLambda',25,...
    'LambdaRatio',1e-2,'CV',5);
toc
```

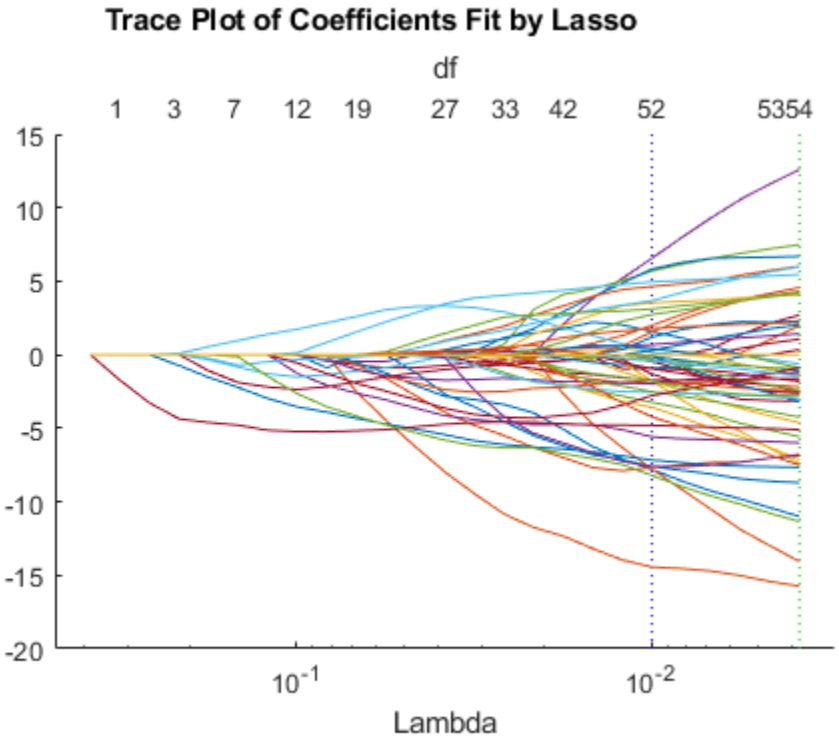
```
Elapsed time is 16.517331 seconds.
```

Graphically compare the new results to the first results.

```
lassoPlot(Bquick,Squick,'PlotType','CV');
legend('show') % Show legend
```



```
lassoPlot(Bquick,Squick,'PlotType','Lambda','XScale','log')
```



The number of nonzero coefficients in the lowest plus one standard deviation model is around 50, similar to the first computation.

Generalized Linear Mixed-Effects Models

In this section...

What Are Generalized Linear Mixed-Effects Models?

Generalized linear mixed-effects (GLME) models describe the relationship between a response variable and independent variables using coefficients that can vary with respect to one or more grouping variables, for data with a response variable distribution other than normal. You can think of GLME models as extensions of generalized linear models on page 12-9 (GLM) for data that are collected and summarized in groups. Alternatively, you can think of GLME models as a generalization of linear mixed-effects models on page 11-131 (LME) for data where the response variable is not normally distributed.

A mixed-effects model consists of fixed-effects and random-effects terms. Fixed-effects terms are usually the conventional linear regression part of the model. Random-effects terms are associated with individual experimental units drawn at random from a population, and account for variations between groups that might affect the response. The random effects have prior distributions, whereas the fixed effects do not.

GLME Model Equations

The standard form of a generalized linear mixed-effects model is

$$y_i | b \sim Distr\left(\mu_i, \frac{\sigma^2}{w_i}\right)$$

$$g(\mu) = X\beta + Zb + \delta\,,$$

where

- y is an n-by-1 response vector, and y_i is its ith element.
- b is the random-effects vector.
- $Distr$ is a specified conditional distribution of y given b.
- μ is the conditional mean of y given b, and μ_i is its ith element.
- σ^2 is the dispersion parameter.
- w is the effective observation weight vector, and w_i is the weight for observation i.

  - For a binomial distribution, the effective observation weight is equal to the prior weight specified using the `'Weights'` name-value pair argument in `fitglme`, multiplied by the binomial size specified using the `'BinomialSize'` name-value pair argument.
  - For all other distributions, the effective observation weight is equal to the prior weight specified using the `'Weights'` name-value pair argument in `fitglme`.
- $g(\mu)$ is a link function that defines the relationship between the mean response μ and the linear combination of the predictors.
- X is an n-by-p fixed-effects design matrix.
- β is a p-by-1 fixed-effects vector.
- Z is an n-by-q random-effects design matrix.
- b is a q-by-1 random-effects vector.
- δ is a model offset vector.

The model for the mean response μ is

$$\mu = g^{-1}(\eta),$$

where g^{-1} is inverse of the link function $g(\mu)$, and $\widehat{\eta}_{ME}$ is the linear predictor of the fixed and random effects of the generalized linear mixed-effects model

$$\eta = X\beta + Zb + \delta .$$

A GLME model is parameterized by β, θ, and σ^2.

The assumptions for generalized linear mixed-effects models are:

- The random effects vector b has the prior distribution:

  $$b \left| \sigma^2, \theta \sim N\left(0, \sigma^2 D(\theta)\right), \right.$$

  where σ^2 is the dispersion parameter, and D is a symmetric and positive semidefinite matrix parameterized by an unconstrained parameter vector θ.
- The observations y_i are conditionally independent given b.

Prepare Data for Model Fitting

To fit a GLME model to your data, use `fitglme`. Format your input data using the `table` data type. Each row of the table represents one observation, and each column represents one predictor variable. For more information on creating and using `table`, see “Create and Work with Tables”.

Input data can include continuous and grouping variables. `fitglme` assumes that predictors using the following data types are categorical:

- Logical
- Categorical
- Character vector or character array
- String array
- Cell array of character vectors

If the input data table contains any `NaN` values, then `fitglme` excludes that entire row of data from the fit. To exclude additional rows of data, you can use the `'Exclude'` name-value pair argument of `fitglme` when fitting the model.

Choose a Distribution Type for the Model

GLME models are used when the response data does not follow a normal distribution. Therefore, when fitting a model using `fitglme`, you must specify the response distribution type using the `'Distribution'` name-value pair argument. Often, the type of response data suggests the appropriate distribution type for the model.

Type of Response Data	Suggested Response Distribution Type
Any real number	`'Normal'`
Any positive number	`'Gamma'` or `'InverseGaussian'`
Any nonnegative integer	`'Poisson'`
Integer from 0 to *n*, where *n* is a fixed positive value	`'Binomial'`

Choose a Link Function for the Model

GLME models use a link function, *g*, to map the relationship between the mean response and the linear combination of the predictors. By default, `fitglme` uses a predefined, commonly accepted link function based on the specified distribution of the response data, as shown in the following table. However, you can specify a different link function from the list of predefined functions, or define your own, using the `'Link'` name-value pair argument of `fitglme`.

Value	Description
`'comploglog'`	`g(mu) = log(-log(1-mu))`
`'identity'`	`g(mu) = mu` Canonical link for the normal distribution.
`'log'`	`g(mu) = log(mu)` Canonical link for the Poisson distribution.
`'logit'`	`g(mu) = log(mu/(1-mu))` Canonical link for the binomial distribution.
`'loglog'`	`g(mu) = log(-log(mu))`
`'probit'`	`g(mu) = norminv(mu)`
`'reciprocal'`	`g(mu) = mu.^(-1)`
Scalar value `P`	`g(mu) = mu.^P`

Value	Description
Structure `S`	A structure containing four fields whose values are function handles: • `S.Link` — Link function • `S.Derivative` — Derivative • `S.SecondDerivative` — Second derivative • `S.Inverse` — Inverse of link If `'FitMethod'` is `'MPL'` or `'REMPL'`, or if `S` represents a canonical link for the specified distribution, you can omit the specification of `S.SecondDerivative`.

When fitting a model to data, `fitglme` uses the canonical link function by default.

Distribution	Default Link Function
`'Normal'`	`'identity'`
`'Binomial'`	`'logit'`
`'Poisson'`	`'log'`
`'Gamma'`	`-1`
`'InverseGaussian'`	`-2`

The link functions `'comploglog'`, `'loglog'`, and `'probit'` are mainly useful for binomial models.

Specify the Model Formula

Model specification for `fitglme` uses Wilkinson notation, which is a character vector or string scalar of the form `'y ~ terms'`, where `y` is the response variable name, and `terms` is written in the following notation.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	$X, X^2, ..., X^k$
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (element-wise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only
`- X2`	Do not include `X2`
`X1*X2 + X3`	`X1`, `X2`, `X3`, `X1*X2`
`X1 + X2 + X3 + X1:X2`	`X1`, `X2`, `X3`, `X1*X2`
`X1*X2*X3 - X1:X2:X3`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`, `X2*X3`
`X1*(X2 + X3)`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`

Formulas include a constant (intercept) term by default. To exclude a constant term from the model, include –1 in the formula.

For generalized linear mixed-effects models, the formula specification is of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms, respectively.

Suppose the input data table contains the following:

- A response variable, `y`
- Predictor variables, `X1`, `X2`, ..., `XJ`, where *J* is the total number of predictor variables (including continuous and grouping variables).
- Grouping variables, `g1`, `g2`, ..., `gR`, where *R* is the number of grouping variables.

The grouping variables in `XJ` and `gR` can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form `'y ~ fixed + (random1|g1) + ... + (randomR|gR)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random1` is a specification of the random-effects design matrix `Z1` corresponding to grouping variable `g1`, and similarly `randomR` is a specification of the random-effects design matrix `ZR` corresponding to grouping variable `gR`. You can express the `fixed` and `random` terms using Wilkinson notation as follows.

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, `X1`, and `X2`. This formula is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept, with fixed effects for `X1` and `X2`. The implicit intercept term is suppressed by including `-1`.
`'y ~ 1 + (1 | g1)'`	A fixed effect for the intercept, plus a random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This formula is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random-effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

For example, the sample data `mfr` contains simulated data from a manufacturing company that operates 50 factories across the world. Each factory runs a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches), and recorded data on processing time (`time_dev`), temperature (`temp_dev`), number of defects (`defects`), and a categorical variable indicating the raw materials supplier (`supplier`) for each batch.

To determine whether the new process (represented by the predictor variable `newprocess`) significantly reduces the number of defects, fit a GLME model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution.

The number of defects can be modeled using a Poisson distribution

$$defects_{ij} \sim Poisson(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 newprocess_{ij} + \beta_2 time\_dev_{ij} + \beta_3 temp\_dev_{ij} + \beta_4 supplier\_C_{ij} + \beta_5 supplier\_B_{ij} + b_i,$$

where

- $defects_{ij}$ is the number of defects observed in the batch produced by factory i (where i = 1, 2, ..., 20) during batch j (where j = 1, 2, ..., 5).
- μ_{ij} is the mean number of defects corresponding to factory i during batch j.
- $supplier\_C_{ij}$ and $supplier\_B_{ij}$ are dummy variables that indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0,\sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

Using Wilkinson notation, specify this model as:

```
'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)'
```

To account for the Poisson distribution of the response variable, when fitting the model using `fitglme`, specify the `'Distribution'` name-value pair argument as `'Poisson'`. By default, `fitglme` uses a log link function for response variables with a Poisson distribution.

Display the Model

The output of the fitting function `fitglme` provides information about generalized linear mixed-effects model.

Using the `mfr` manufacturing experiment data, fit a model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Specify the response distribution as Poisson, the link function as log, and the fit method as Laplace.

```
load mfr

glme = fitglme(mfr,...
        'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
        'Distribution','Poisson','Link','log','FitMethod','Laplace',...
        'DummyVarCoding','effects')
```

```
glme =


Generalized linear mixed-effects model fit by ML

Model information:
```

```
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35

Fixed effects coefficients (95% CIs):
    Name                Estimate     SE          tStat       DF    pValue
    '(Intercept)'          1.4689     0.15988      9.1875    94    9.8194e-15
    'newprocess'         -0.36766     0.17755     -2.0708    94      0.041122
    'time_dev'           -0.094521    0.82849    -0.11409    94       0.90941
    'temp_dev'            -0.28317     0.9617    -0.29444    94       0.76907
    'supplier_C'         -0.071868    0.078024     -0.9211    94       0.35936
    'supplier_B'          0.071072     0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
     -0.72019    -0.015134
      -1.7395       1.5505
      -2.1926       1.6263
     -0.22679     0.083051
    -0.082588      0.22473

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1               Name2               Type         Estimate
    '(Intercept)'       '(Intercept)'       'std'        0.31381

Group: Error
    Name                    Estimate
    'sqrt(Dispersion)'      1
```

The `Model information` table displays the total number of observations in the sample data (100), the number of fixed- and random-effects coefficients (6 and 20, respectively), and the number of covariance parameters (1). It also indicates that the response variable has a `Poisson` distribution, the link function is `Log`, and the fit method is `Laplace`.

`Formula` indicates the model specification using Wilkinson's notation.

The `Model fit statistics` table displays statistics used to assess the goodness of fit of the model. This includes the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`) values, log likelihood (`LogLikelihood`), and deviance (`Deviance`) values.

The `Fixed effects coefficients` table indicates that `fitglme` returned 95% confidence intervals. It contains one row for each fixed-effects predictor, and each column contains statistics corresponding to that predictor. Column 1 (`Name`) contains the name of each fixed-effects coefficient, column 2 (`Estimate`) contains its estimated value, and column 3 (`SE`) contains the standard error of the coefficient. Column 4 (`tStat`) contains the *t*-statistic for a hypothesis test that the coefficient is equal to 0. Column 5 (`DF`) and column 6 (`pValue`) contain the degrees of freedom and *p*-value that correspond to the *t*-statistic, respectively. The last two columns (`Lower` and `Upper`) display the lower and upper limits, respectively, of the 95% confidence interval for each fixed-effects coefficient.

`Random effects covariance parameters` displays a table for each grouping variable (here, only `factory`), including its total number of levels (20), and the type and estimate of the covariance parameter. Here, `std` indicates that `fitglme` returns the standard deviation of the random effect associated with the factory predictor, which has an estimated value of 0.31381. It also displays a table containing the error parameter type (here, the square root of the dispersion parameter), and its estimated value of 1.

The standard display generated by `fitglme` does not provide confidence intervals for the random-effects parameters. To compute and display these values, use `covarianceParameters`.

Work with the Model

After you create a GLME model using `fitglme`, you can use additional functions to work with the model.

Inspect and Test Coefficients and Confidence Intervals

To extract estimates of the fixed- and random-effects coefficients, covariance parameters, design matrices, and related statistics:

- `fixedEffects` extracts estimated fixed-effects coefficients and related statistics from a fitted model. Related statistics include the standard error; the *t*-statistic, degrees of freedom, and *p*-value for a hypothesis test of whether each parameter is equal to 0; and the confidence intervals.
- `randomEffects` extracts estimated random-effects coefficients and related statistics from a fitted GLME model. Related statistics include the estimated empirical Bayes predictor (EBP) of each random effect, the square root of the conditional mean squared error of prediction (CMSEP) given the covariance parameters and the response; the *t*-statistic, estimated degrees of freedom, and *p*-value for a hypothesis test of whether each random effect is equal to 0; and the confidence intervals.
- `covarianceParameters` extracts estimated covariance parameters and related statistics from a fitted GLME model. Related statistics include estimate of the covariance parameter, and the confidence intervals.
- `designMatrix` extracts the fixed- and random-effects design matrices, or a specified subset thereof, from the fitted GLME model.

To conduct customized hypothesis tests for the significance of fixed- and random-effects coefficients, and to compute custom confidence intervals:

- `anova` performs a marginal *F*-test (hypothesis test) on fixed-effects terms, to determine if all coefficients representing the fixed-effects terms are equal to 0. You can use `anova` to test the combined significance of the coefficients of categorical predictors.
- `coefCI` computes confidence intervals for fixed- and random-effects parameters from a fitted GLME model. By default, `fitglme` computes 95% confidence intervals. Use `coefCI` to compute the boundaries at a different confidence level.
- `coefTest` performs custom hypothesis tests on fixed-effects or random-effects vectors of a fitted generalized linear mixed-effects model. For example, you can specify contrast matrices.

Generate New Response Values and Refit Model

To generate new response values, including fitted, predicted, and random responses, based on the fitted GLME model:

- `fitted` computes fitted response values using the original predictor values, and the estimated coefficient and parameter values from the fitted model.
- `predict` computes the predicted conditional or marginal mean of the response using either the original predictor values or new predictor values, and the estimated coefficient and parameter values from the fitted model.
- `random` generates random responses from a fitted model.

- `refit` creates a new fitted GLME model, based on the original model and a new response vector.

Inspect and Visualize Residuals

To extract and visualize residuals from the fitted GLME model:

- `residuals` extracts the raw or Pearson residuals from the fitted model. You can also specify whether to compute the conditional or marginal residuals.
- `plotResiduals` creates plots using the raw or Pearson residuals from the fitted model, including:
  - A histogram of the residuals
  - A scatterplot of the residuals versus fitted values
  - A scatterplot of residuals versus lagged residuals

See Also

`GeneralizedLinearMixedModel` | `fitglme`

Related Examples

- “Fit a Generalized Linear Mixed-Effects Model” on page 12-57

Fit a Generalized Linear Mixed-Effects Model

This example shows how to fit a generalized linear mixed-effects model (GLME) to sample data.

Load the sample data.

Load the sample data.

```
load mfr
```

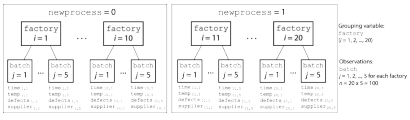
A manufacturing company operates 50 factories across the world, and each runs a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. However, the company wants to test the new process in select factories to ensure that it is effective before rolling it out to all 50 locations.

To test whether the new process significantly reduces the number of defects in each batch, the company selected 20 of its factories at random to participate in an experiment. Ten factories implemented the new process, while the other ten used the old process.

In each of the 20 factories (i = 1, 2, ..., 20), the company ran five batches (j = 1, 2, ..., 5) and recorded the following data in the table `mfr`:

- Flag to indicate use of the new process:
  - If the batch used the new process, then `newprocess = 1`
  - If the batch used the old process, then `newprocess = 0`
- Processing time for the batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Supplier of the chemical used in the batch (`supplier`)
  - `supplier` is a categorical variable with levels `A`, `B`, and `C`, where each level represents one of the three suppliers
- Number of defects in the batch (defects)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours and 20 degrees Celsius. The response variable `defects` has a Poisson distribution. This is simulated data.



The company wants to determine whether the new process significantly reduces the number of defects in each batch, while accounting for quality differences that might exist due to factory-specific variations in time, temperature, and supplier. The number of defects per batch can be modeled using a Poison distribution:

$$defects_{ij} \sim Poisson(\mu_{ij})$$

Use a generalized linear mixed-effects model to model the number of defects per batch:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 newprocess_{ij} + \beta_2 time\_dev_{ij} + \beta_3 temp\_dev_{ij} + \beta_4 supplier\_C_{ij} + \beta_5 supplier\_B_{ij} + b_i,$$

where

- $defects_{ij}$ is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where i = 1, 2, ..., 20) during batch j (where j = 1, 2, ..., 5).
- $newprocess_{ij}$, $time\_dev_{ij}$, and $temp\_dev_{ij}$ are the measurements for each variable that correspond to factory i during batch j. For example, $newprocess_{ij}$ indicates whether the batch produced by factory i during batch j used the new process.
- $supplier\_C_{ij}$ and $supplier\_B_{ij}$ are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0,\sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

Fit a GLME model and interpret the results.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

```
glme = fitglme(mfr,...
'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
'Distribution','Poisson','Link','log','FitMethod','Laplace',...
'DummyVarCoding','effects')
```

```
glme =


Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35

Fixed effects coefficients (95% CIs):
    Name                 Estimate     SE          tStat       DF    pValue
    '(Intercept)'           1.4689     0.15988      9.1875    94    9.8194e-15
    'newprocess'          -0.36766     0.17755     -2.0708    94      0.041122
    'time_dev'            -0.094521    0.82849    -0.11409    94       0.90941
    'temp_dev'            -0.28317     0.9617     -0.29444    94       0.76907
    'supplier_C'          -0.071868    0.078024    -0.9211    94       0.35936
    'supplier_B'           0.071072    0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
     -0.72019    -0.015134
      -1.7395       1.5505
      -2.1926       1.6263
     -0.22679     0.083051
    -0.082588      0.22473
```

```
Random effects covariance parameters:
Group: factory (20 Levels)
    Name1               Name2               Type         Estimate
    '(Intercept)'       '(Intercept)'       'std'        0.31381

Group: Error
    Name                      Estimate
    'sqrt(Dispersion)'        1
```

The `Model information` table displays the total number of observations in the sample data (100), the number of fixed- and random-effects coefficients (6 and 20, respectively), and the number of covariance parameters (1). It also indicates that the response variable has a `Poisson` distribution, the link function is `Log`, and the fit method is `Laplace`.

`Formula` indicates the model specification using Wilkinson's notation.

The `Model fit statistics` table displays statistics used to assess the goodness of fit of the model. This includes the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`) values, log likelihood (`LogLikelihood`), and deviance (`Deviance`) values.

The `Fixed effects coefficients` table indicates that `fitglme` returned 95% confidence intervals. It contains one row for each fixed-effects predictor, and each column contains statistics corresponding to that predictor. Column 1 (`Name`) contains the name of each fixed-effects coefficient, column 2 (`Estimate`) contains its estimated value, and column 3 (`SE`) contains the standard error of the coefficient. Column 4 (`tStat`) contains the *t*-statistic for a hypothesis test that the coefficient is equal to 0. Column 5 (`DF`) and column 6 (`pValue`) contain the degrees of freedom and *p*-value that correspond to the *t*-statistic, respectively. The last two columns (`Lower` and `Upper`) display the lower and upper limits, respectively, of the 95% confidence interval for each fixed-effects coefficient.

`Random effects covariance parameters` displays a table for each grouping variable (here, only `factory`), including its total number of levels (20), and the type and estimate of the covariance parameter. Here, `std` indicates that `fitglme` returns the standard deviation of the random effect associated with the factory predictor, which has an estimated value of 0.31381. It also displays a table containing the error parameter type (here, the square root of the dispersion parameter), and its estimated value of 1.

The standard display generated by `fitglme` does not provide confidence intervals for the random-effects parameters. To compute and display these values, use `covarianceParameters`.

Check significance of random effect.

To determine whether the random-effects intercept grouped by `factory` is statistically significant, compute the confidence intervals for the estimated covariance parameter.

```
[psi,dispersion,stats] = covarianceParameters(glme);
```

`covarianceParameters` returns the estimated covariance parameter in `psi`, the estimated dispersion parameter `dispersion`, and a cell array of related statistics `stats`. The first cell of `stats` contains statistics for `factory`, while the second cell contains statistics for the dispersion parameter.

Display the first cell of `stats` to see the confidence intervals for the estimated covariance parameter for `factory`.

```
stats{1}
```

```
ans =
```

```
    Covariance Type: Isotropic

    Group      Name1               Name2               Type
    factory    '(Intercept)'       '(Intercept)'       'std'


    Estimate    Lower      Upper
    0.31381     0.19253    0.51148
```

The columns `Lower` and `Upper` display the default 95% confidence interval for the estimated covariance parameter for `factory`. Because the interval [0.19253,0.51148] does not contain 0, the random-effects intercept is significant at the 5% significance level. Therefore, the random effect due to factory-specific variation must be considered before drawing any conclusions about the effectiveness of the new manufacturing process.

Compare two models.

Compare the mixed-effects model that includes a random-effects intercept grouped by `factory` with a model that does not include the random effect, to determine which model is a better fit for the data. Fit the first model, `FEglme`, using only the fixed-effects predictors `newprocess`, `time_dev`, `temp_dev`, and `supplier`. Fit the second model, `glme`, using these same fixed-effects predictors, but also including a random-effects intercept grouped by `factory`.

```
FEglme = fitglme(mfr,...
'defects ~ 1 + newprocess + time_dev + temp_dev + supplier',...
'Distribution','Poisson','Link','log','FitMethod','Laplace');

glme = fitglme(mfr,...
'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
'Distribution','Poisson','Link','log','FitMethod','Laplace');
```

Compare the two models using a likelihood ratio test. Specify `'CheckNesting'` as `true`, so `compare` returns a warning if the nesting requirements are not satisfied.

```
results = compare(FEglme,glme,'CheckNesting',true)

results =


    Theoretical Likelihood Ratio Test

    Model     DF    AIC       BIC       LogLik     LRStat    deltaDF
    FEglme    6     431.02    446.65    -209.51
    glme      7     416.35    434.58    -201.17    16.672    1


    pValue

    4.4435e-05
```

`compare` returns the degrees of freedom (`DF`), the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`), and log likelihood values for each model. `glme` has smaller AIC, BIC, and log likelihood values than `FEglme`, which indicates that `glme` (the model containing the random-effects term for intercept grouped by factory) is the better-fitting model for this data. Additionally, the small *p*-value indicates that `compare` rejects the null hypothesis that the response vector was

generated by the fixed-effects-only model `FEglme`, in favor of the alternative that the response vector was generated by the mixed-effects model `glme`.

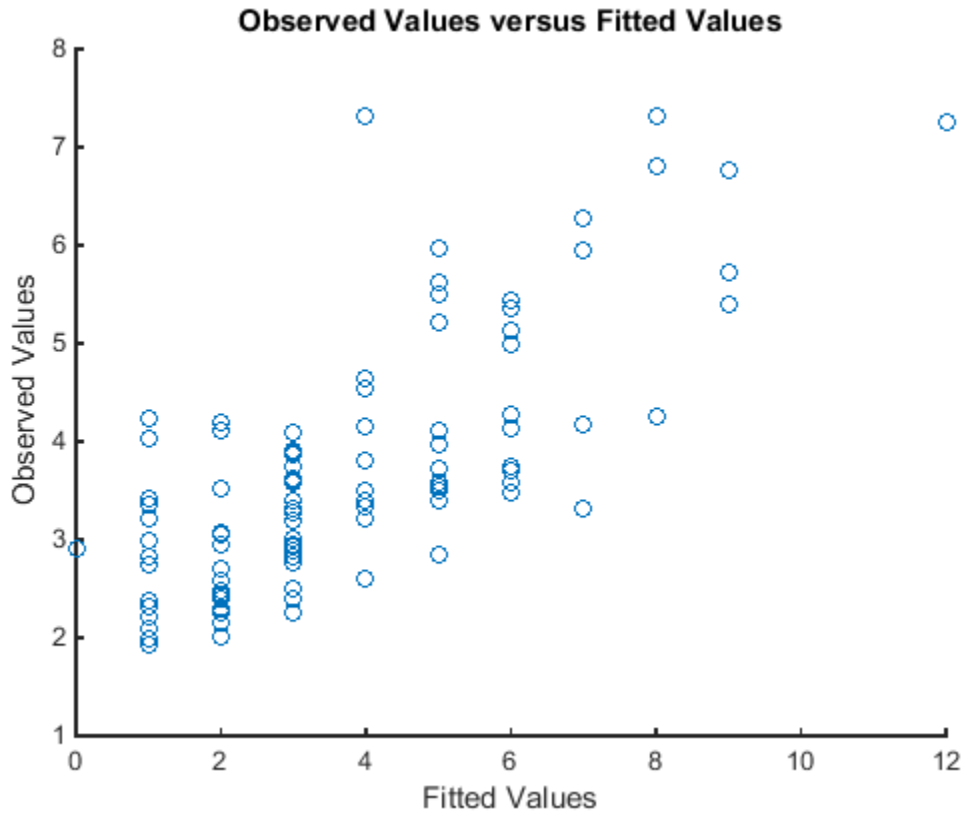
Plot the results.

Generate the fitted conditional mean values for the model.

```
mufit = fitted(glme);
```

Plot the observed response values versus the fitted response values.

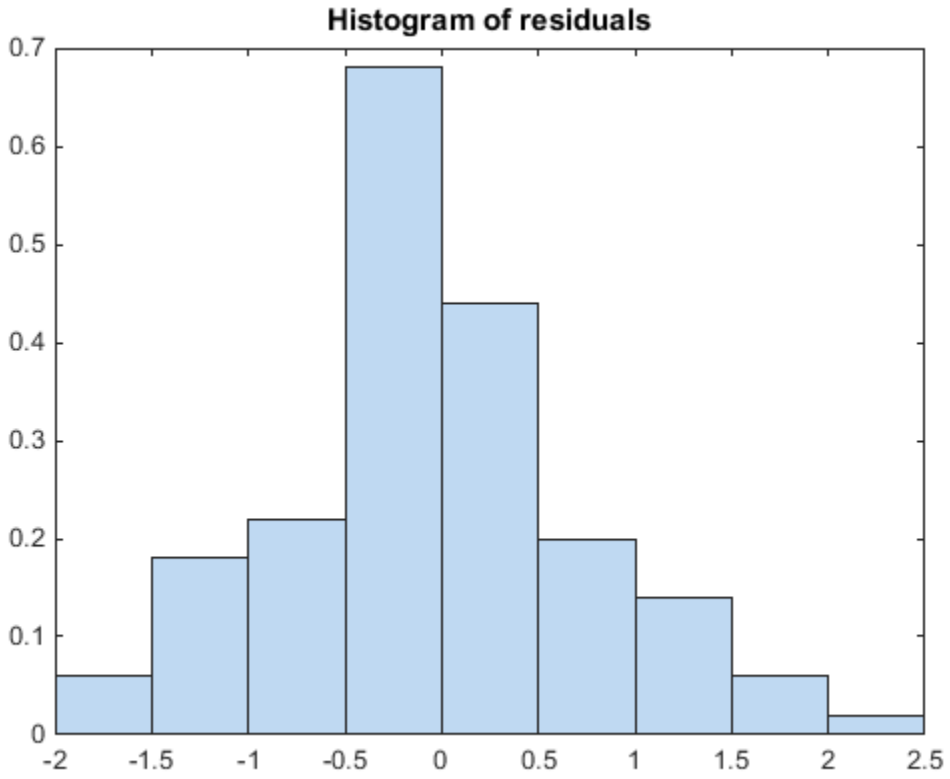
```
figure
scatter(mfr.defects,mufit)
title('Observed Values versus Fitted Values')
xlabel('Fitted Values')
ylabel('Observed Values')
```



Create diagnostic plots using conditional Pearson residuals to test model assumptions. Since raw residuals for generalized linear mixed-effects models do not have a constant variance across observations, use the conditional Pearson residuals instead.

Plot a histogram to visually confirm that the mean of the Pearson residuals is equal to 0. If the model is correct, we expect the Pearson residuals to be centered at 0.

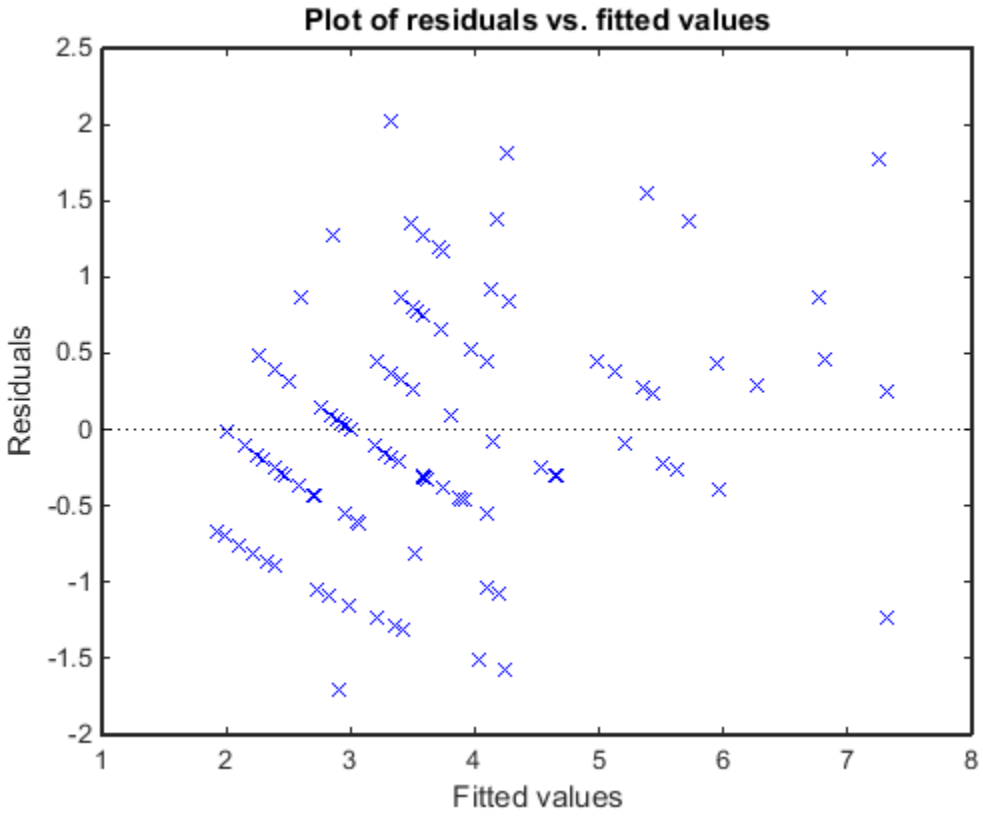
```
plotResiduals(glme,'histogram','ResidualType','Pearson')
```



The histogram shows that the Pearson residuals are centered at 0.

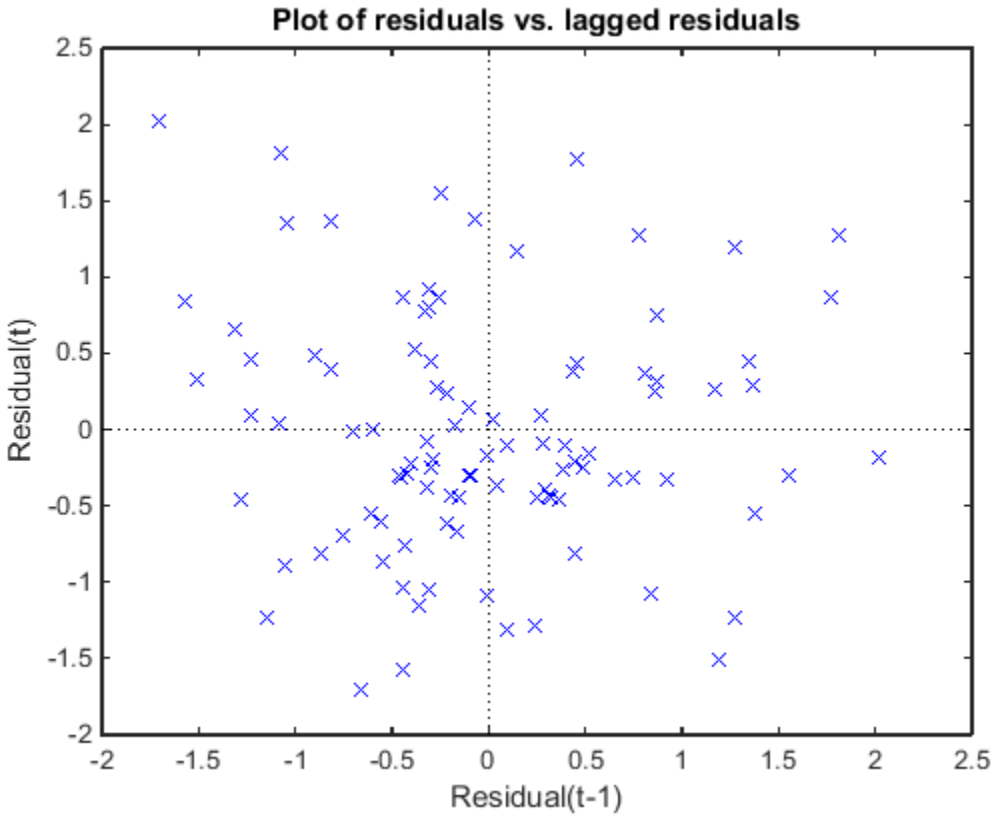
Plot the Pearson residuals versus the fitted values, to check for signs of nonconstant variance among the residuals (heteroscedasticity). We expect the conditional Pearson residuals to have a constant variance. Therefore, a plot of conditional Pearson residuals versus conditional fitted values should not reveal any systematic dependence on the conditional fitted values.

```
plotResiduals(glme,'fitted','ResidualType','Pearson')
```



The plot does not show a systematic dependence on the fitted values, so there are no signs of nonconstant variance among the residuals.

Plot the Pearson residuals versus lagged residuals, to check for correlation among the residuals. The conditional independence assumption in GLME implies that the conditional Pearson residuals are approximately uncorrelated.



There is no pattern to the plot, so there are no signs of correlation among the residuals.

See Also

`GeneralizedLinearMixedModel` | `fitglme`

More About

- “Generalized Linear Models” on page 12-9

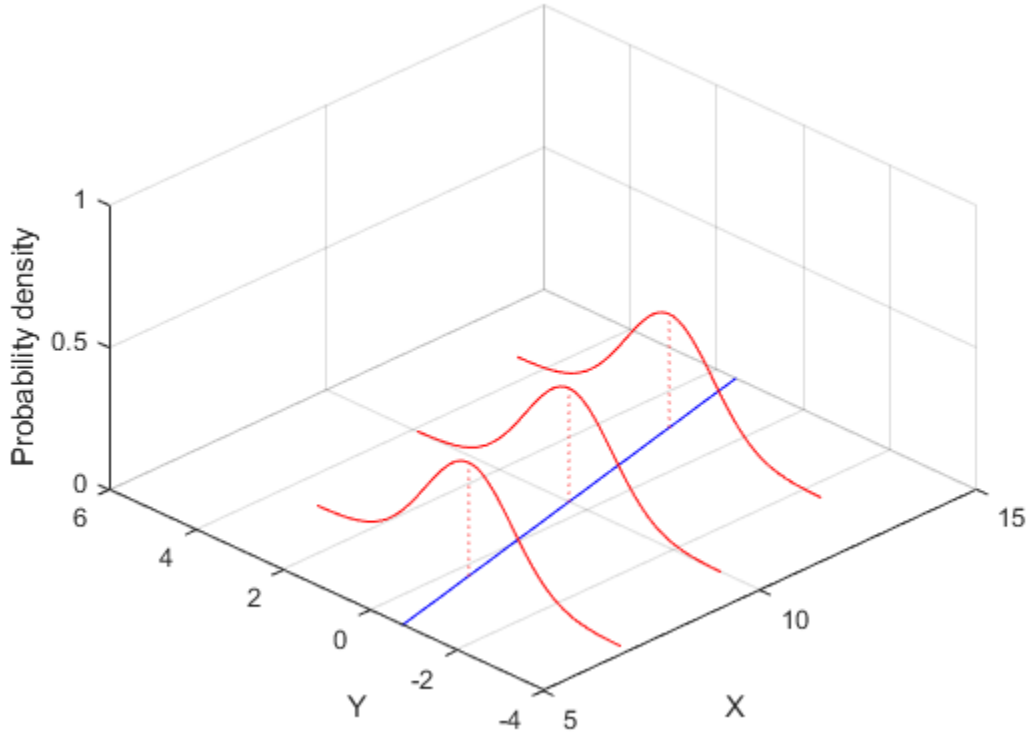
Fitting Data with Generalized Linear Models

This example shows how to fit and evaluate generalized linear models using `glmfit` and `glmval`. Ordinary linear regression can be used to fit a straight line, or any function that is linear in its parameters, to data with normally distributed errors. This is the most commonly used regression model; however, it is not always a realistic one. Generalized linear models extend the linear model in two ways. First, assumption of linearity in the parameters is relaxed, by introducing the link function. Second, error distributions other than the normal can be modeled

Generalized Linear Models

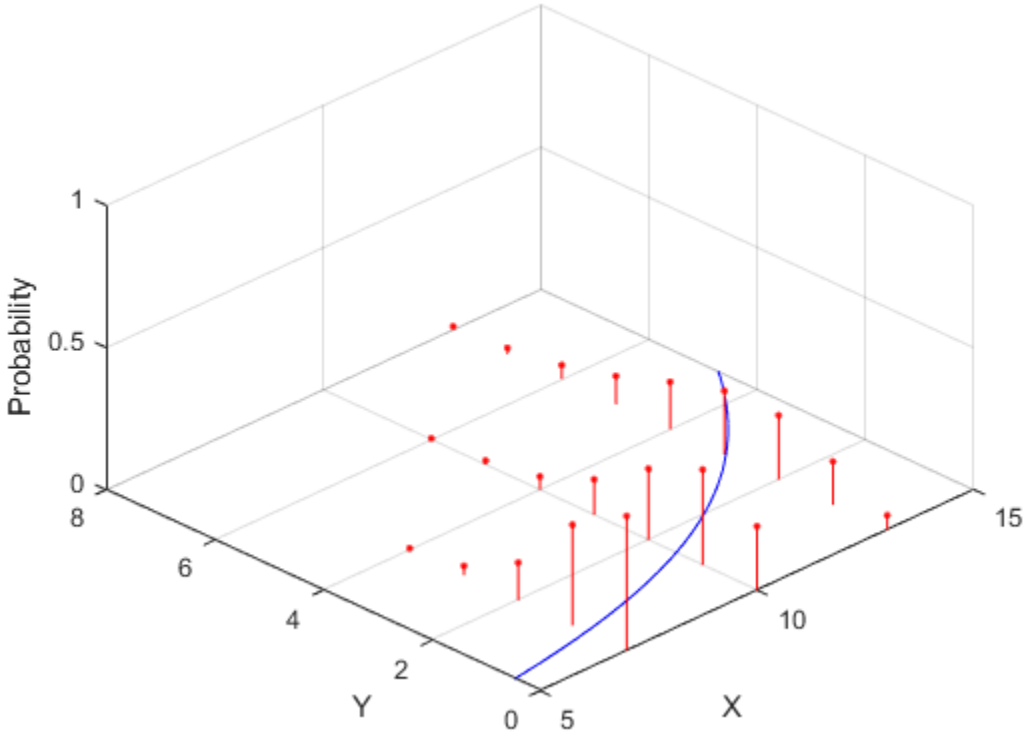
A regression model defines the distribution of a response variable (often generically denoted as y) in terms of one or more predictor variables (often denoted x1, x2, etc.). The most commonly used regression model, the ordinary linear regression, models y as a normal random variable, whose mean is linear function of the predictors, b0 + b1*x1 + ... , and whose variance is constant. In the simplest case of a single predictor x, the model can be represented as a straight line with Gaussian distributions about each point.

```
mu = @(x) -1.9+.23*x;
x = 5:.1:15;
yhat = mu(x);
dy = -3.5:.1:3.5; sz = size(dy); k = (length(dy)+1)/2;
x1 =  7*ones(sz); y1 = mu(x1)+dy; z1 = normpdf(y1,mu(x1),1);
x2 = 10*ones(sz); y2 = mu(x2)+dy; z2 = normpdf(y2,mu(x2),1);
x3 = 13*ones(sz); y3 = mu(x3)+dy; z3 = normpdf(y3,mu(x3),1);
plot3(x,yhat,zeros(size(x)),'b-', ...
      x1,y1,z1,'r-', x1([k k]),y1([k k]),[0 z1(k)],'r:', ...
      x2,y2,z2,'r-', x2([k k]),y2([k k]),[0 z2(k)],'r:', ...
      x3,y3,z3,'r-', x3([k k]),y3([k k]),[0 z3(k)],'r:');
zlim([0 1]);
xlabel('X'); ylabel('Y'); zlabel('Probability density');
grid on; view([-45 45]);
```



In a generalized linear model, the mean of the response is modeled as a monotonic nonlinear transformation of a linear function of the predictors, g(b0 + b1*x1 + ...) . The inverse of the transformation g is known as the "link" function. Examples include the logit (sigmoid) link and the log link. Also, y may have a non-normal distribution, such as the binomial or Poisson. For example, a Poisson regression with log link and a single predictor x can be represented as an exponential curve with Poisson distributions about each point.

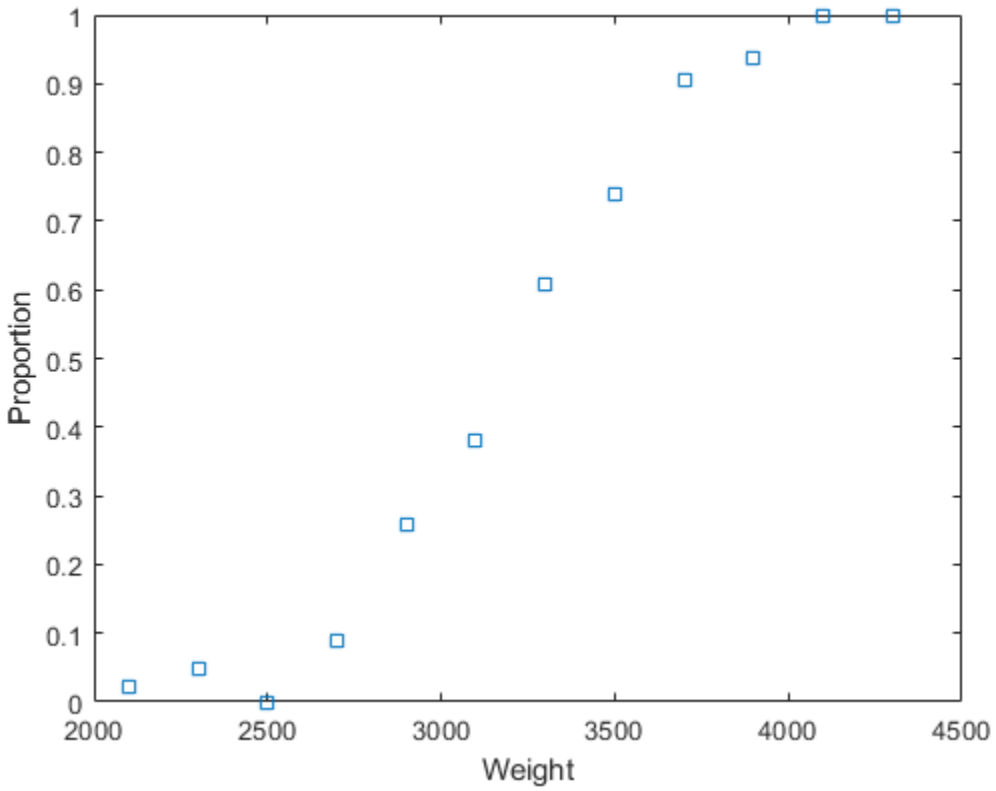
```
mu = @(x) exp(-1.9+.23*x);
x = 5:.1:15;
yhat = mu(x);
x1 =  7*ones(1,5);  y1 = 0:4; z1 = poisspdf(y1,mu(x1));
x2 = 10*ones(1,7); y2 = 0:6; z2 = poisspdf(y2,mu(x2));
x3 = 13*ones(1,9); y3 = 0:8; z3 = poisspdf(y3,mu(x3));
plot3(x,yhat,zeros(size(x)),'b-', ...
      [x1; x1],[y1; y1],[z1; zeros(size(y1))],'r-', x1,y1,z1,'r.', ...
      [x2; x2],[y2; y2],[z2; zeros(size(y2))],'r-', x2,y2,z2,'r.', ...
      [x3; x3],[y3; y3],[z3; zeros(size(y3))],'r-', x3,y3,z3,'r.');
zlim([0 1]);
xlabel('X'); ylabel('Y'); zlabel('Probability');
grid on; view([-45 45]);
```



Fitting a Logistic Regression

This example involves an experiment to help model the proportion of cars of various weights that fail a mileage test. The data include observations of weight, number of cars tested, and number failed.

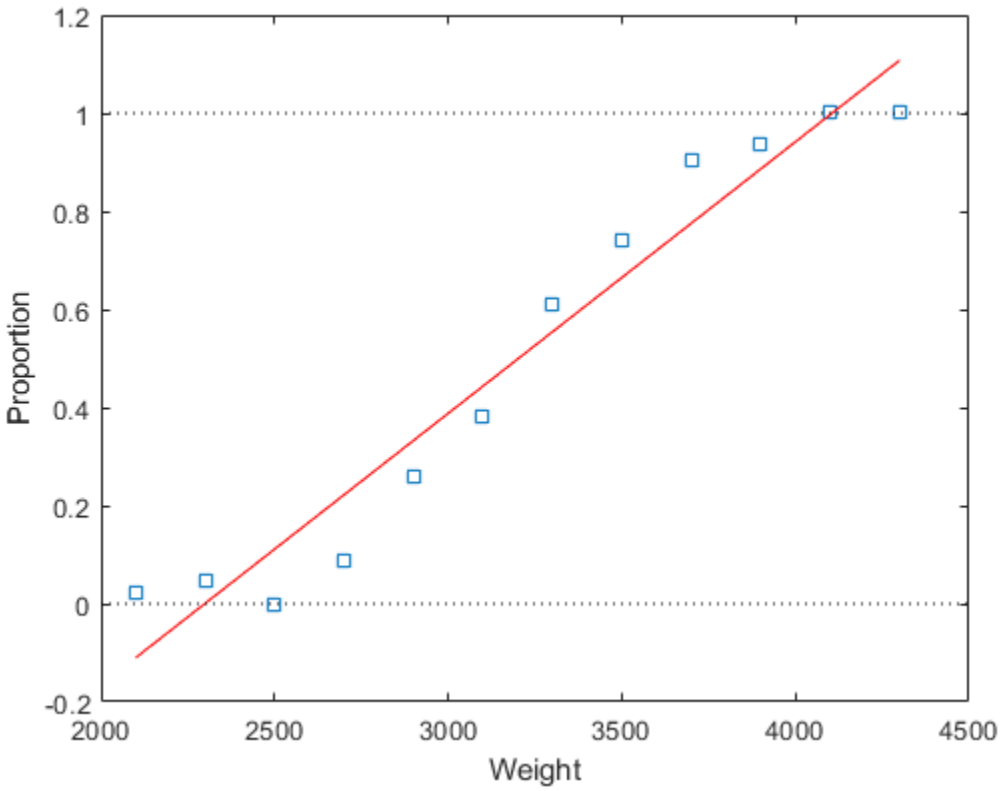
```
% A set of car weights
weight = [2100 2300 2500 2700 2900 3100 3300 3500 3700 3900 4100 4300]';
% The number of cars tested at each weight
tested = [48 42 31 34 31 21 23 23 21 16 17 21]';
% The number of cars failing the test at each weight
failed = [1 2 0 3 8 8 14 17 19 15 17 21]';
% The proportion of cars failing for each weight
proportion = failed ./ tested;

plot(weight,proportion,'s')
xlabel('Weight'); ylabel('Proportion');
```



This graph is a plot of the proportion of cars failing, as a function of weight. It's reasonable to assume that the failure counts came from a binomial distribution, with a probability parameter P that increases with weight. But how exactly should P depend on weight?

We can try fitting a straight line to these data.

```
linearCoef = polyfit(weight,proportion,1);
linearFit = polyval(linearCoef,weight);
plot(weight,proportion,'s', weight,linearFit,'r-', [2000 4500],[0 0],'k:', [2000 4500],[1 1],'k:
xlabel('Weight'); ylabel('Proportion');
```



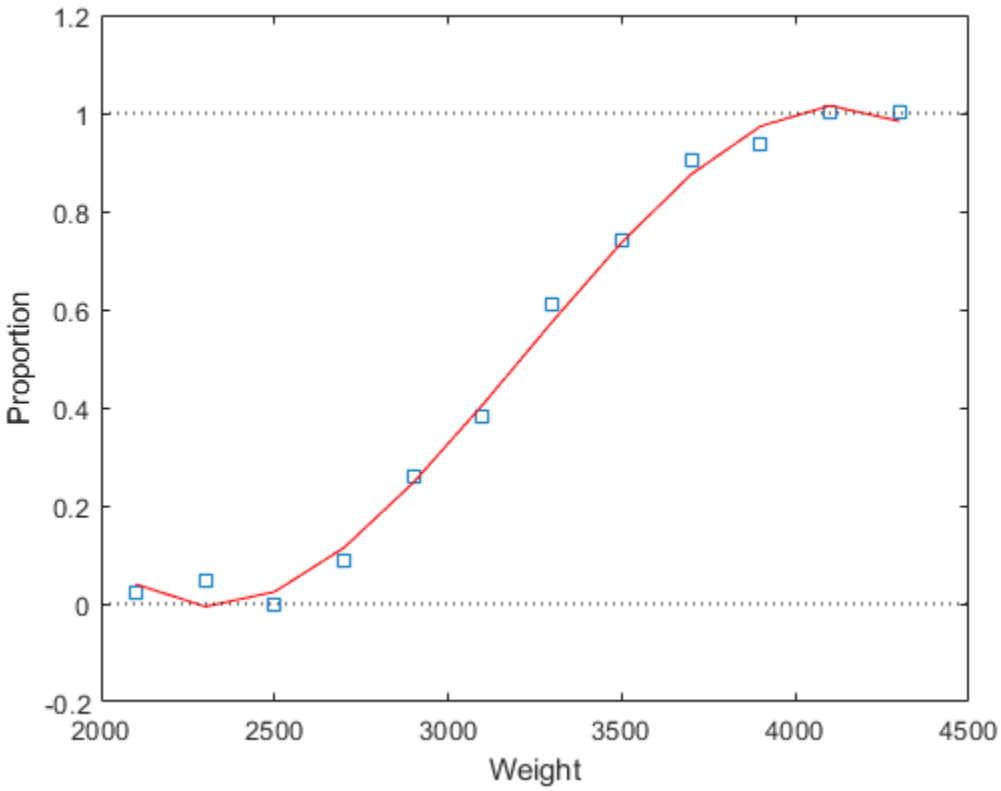
There are two problems with this linear fit:

1) The line predicts proportions less than 0 and greater than 1.

2) The proportions are not normally distributed, since they are necessarily bounded. This violates one of the assumptions required for fitting a simple linear regression model.

Using a higher-order polynomial may appear to help.

```
[cubicCoef,stats,ctr] = polyfit(weight,proportion,3);
cubicFit = polyval(cubicCoef,weight,[],ctr);
plot(weight,proportion,'s', weight,cubicFit,'r-', [2000 4500],[0 0],'k:', [2000 4500],[1 1],'k:'
xlabel('Weight'); ylabel('Proportion');
```

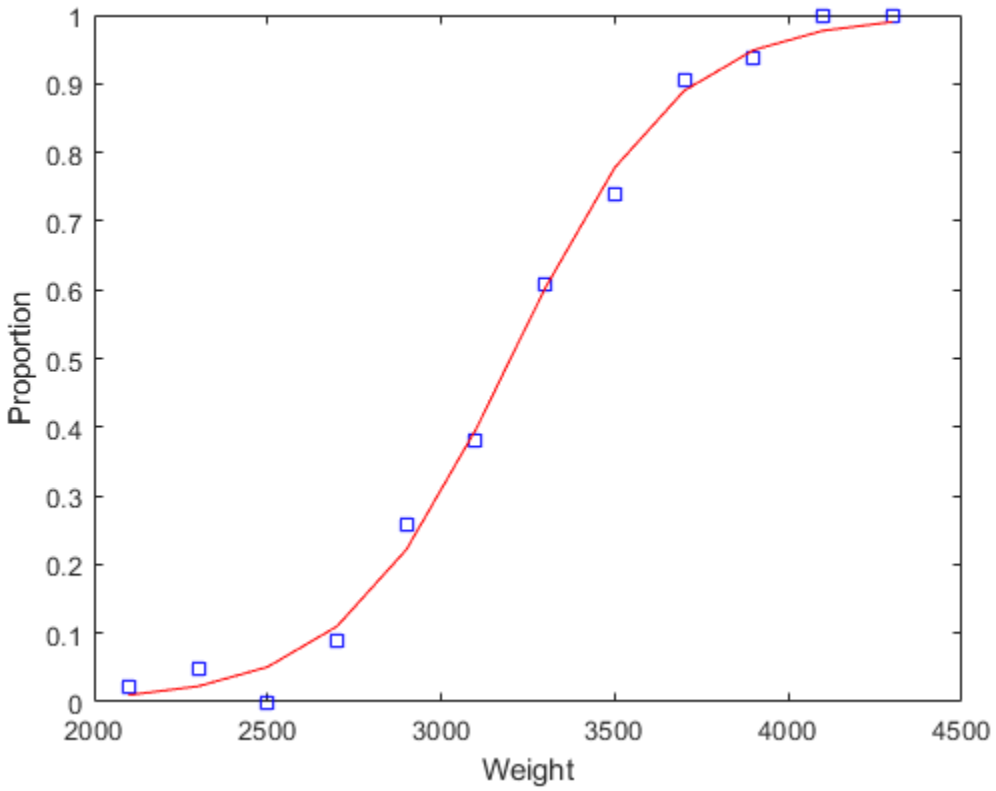


However, this fit still has similar problems. The graph shows that the fitted proportion starts to decrease as weight goes above 4000; in fact it will become negative for larger weight values. And of course, the assumption of a normal distribution is still violated.

Instead, a better approach is to use `glmfit` to fit a logistic regression model. Logistic regression is a special case of a generalized linear model, and is more appropriate than a linear regression for these data, for two reasons. First, it uses a fitting method that is appropriate for the binomial distribution. Second, the logistic link limits the predicted proportions to the range [0,1].

For logistic regression, we specify the predictor matrix, and a matrix with one column containing the failure counts, and one column containing the number tested. We also specify the binomial distribution and the logit link.

```
[logitCoef,dev] = glmfit(weight,[failed tested],'binomial','logit');
logitFit = glmval(logitCoef,weight,'logit');
plot(weight,proportion,'bs', weight,logitFit,'r-');
xlabel('Weight'); ylabel('Proportion');
```



As this plot indicates, the fitted proportions asymptote to zero and one as weight becomes small or large.

Model Diagnostics

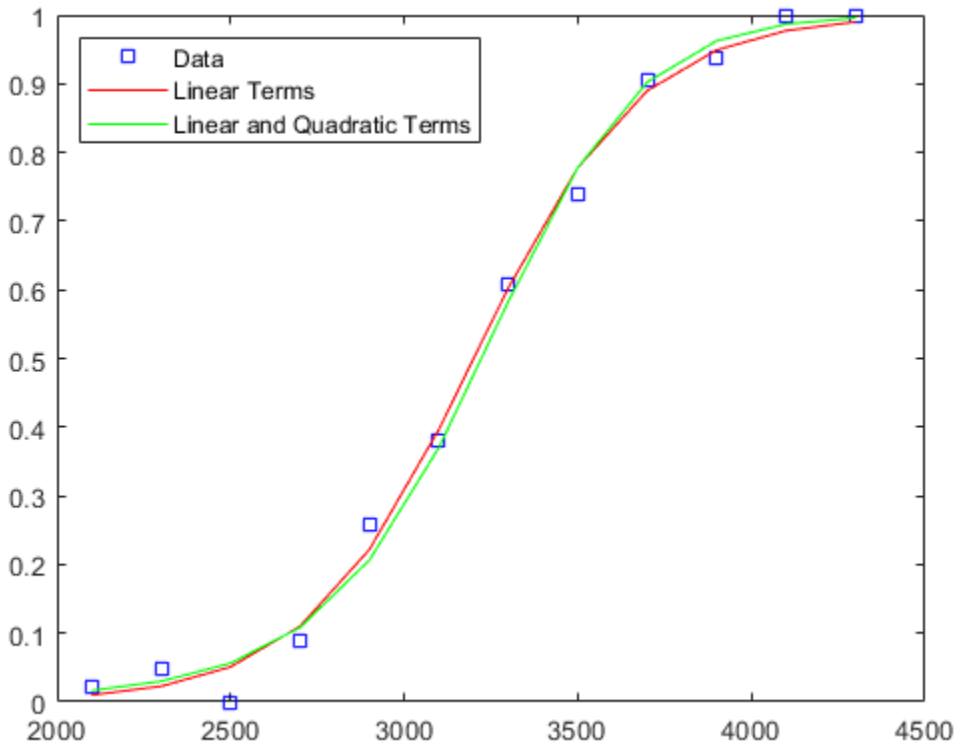
The `glmfit` function provides a number of outputs for examining the fit and testing the model. For example, we can compare the deviance values for two models to determine if a squared term would improve the fit significantly.

```
[logitCoef2,dev2] = glmfit([weight weight.^2],[failed tested],'binomial','logit');
pval = 1 - chi2cdf(dev-dev2,1)
```

```
pval =

    0.4019
```

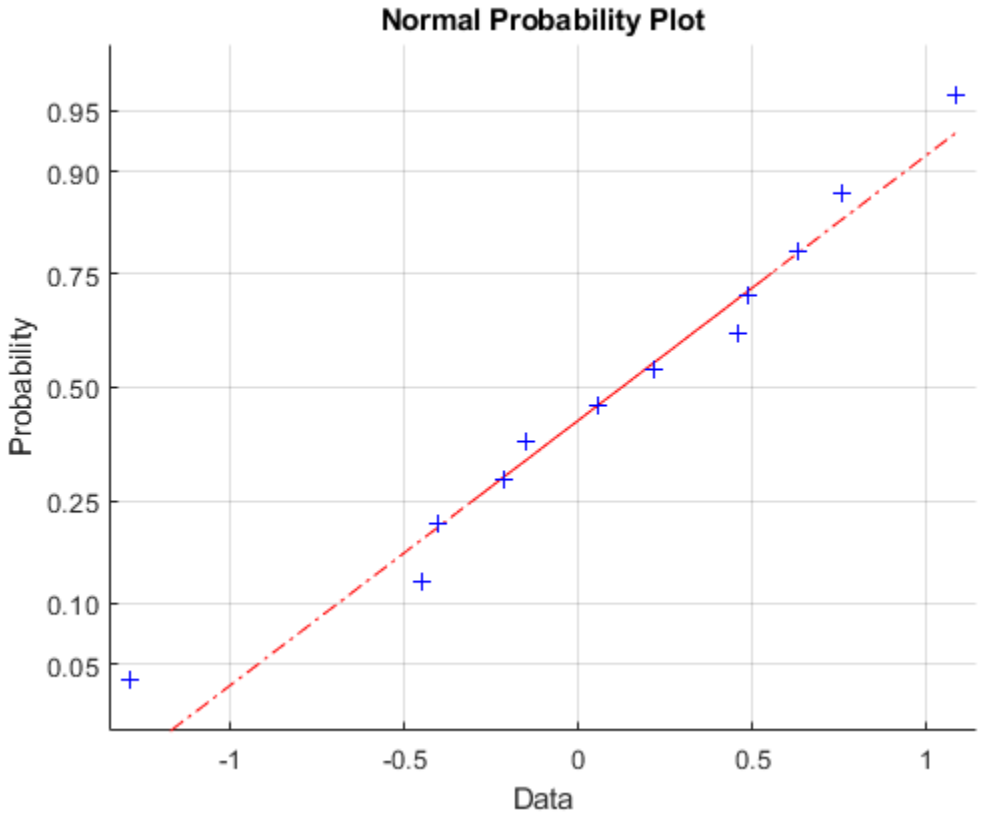
The large p-value indicates that, for these data, a quadratic term does not improve the fit significantly. A plot of the two fits shows there is little difference in the fits.

```
logitFit2 = glmval(logitCoef2,[weight weight.^2],'logit');
plot(weight,proportion,'bs', weight,logitFit,'r-', weight,logitFit2,'g-');
legend('Data','Linear Terms','Linear and Quadratic Terms','Location','northwest');
```



To check the goodness of fit, we can also look at a probability plot of the Pearson residuals. These are normalized so that when the model is a reasonable fit to the data, they have roughly a standard normal distribution. (Without this standardization, the residuals would have different variances.)

```
[logitCoef,dev,stats] = glmfit(weight,[failed tested],'binomial','logit');
normplot(stats.residp);
```

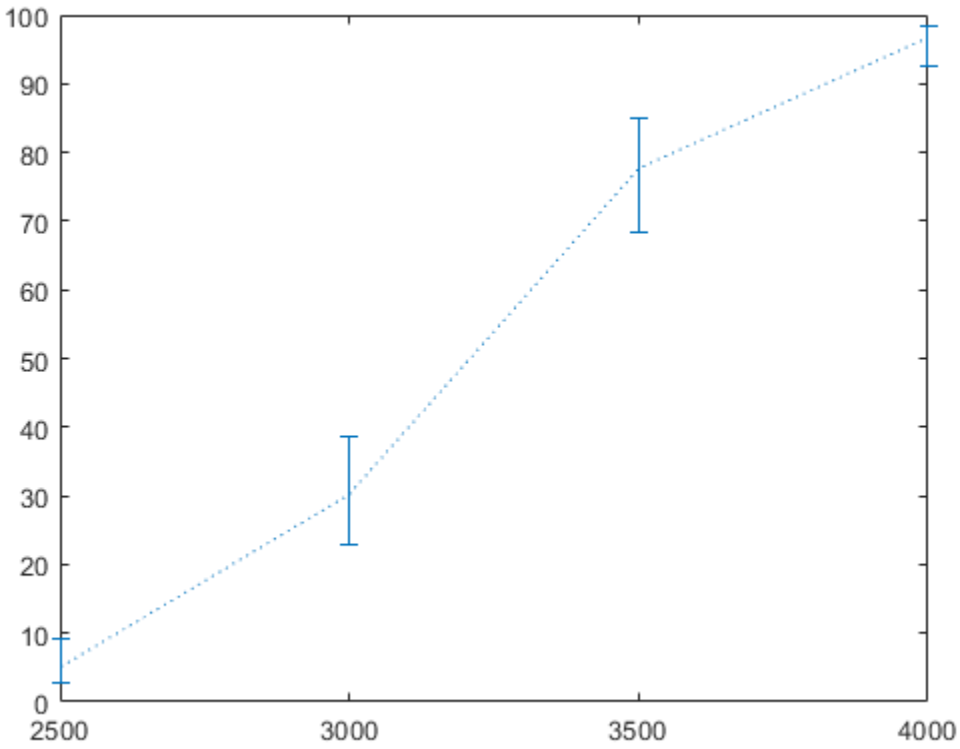


The residual plot shows a nice agreement with the normal distribution.

Evaluating the Model Predictions

Once we are satisfied with the model, we can use it to make predictions, including computing confidence bounds. Here we predict the expected number of cars, out of 100 tested, that would fail the mileage test at each of four weights.

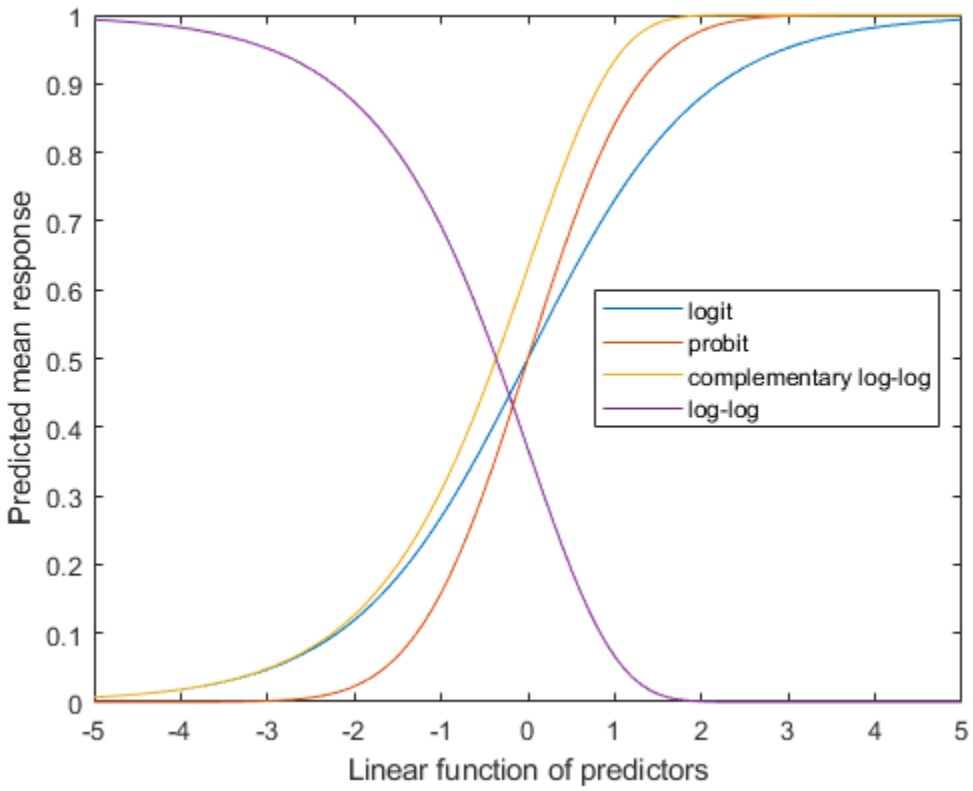
```
weightPred = 2500:500:4000;
[failedPred,dlo,dhi] = glmval(logitCoef,weightPred,'logit',stats,.95,100);
errorbar(weightPred,failedPred,dlo,dhi,':');
```



Link Functions for Binomial Models

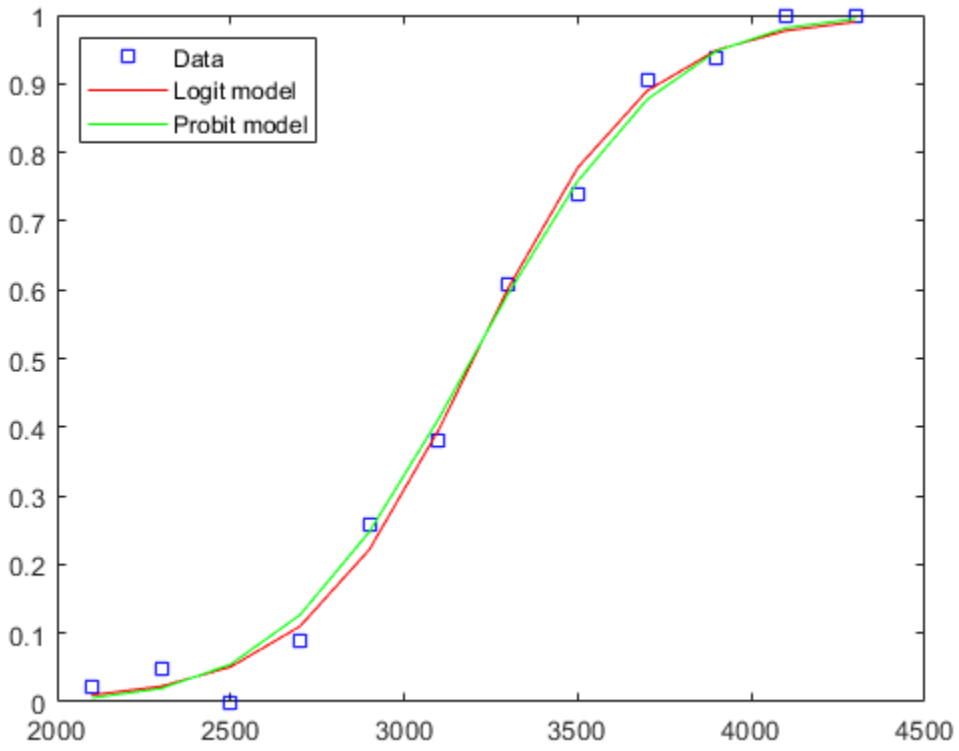
For each of the five distributions that `glmfit` supports, there is a canonical (default) link function. For the binomial distribution, the canonical link is the logit. However, there are also three other links that are sensible for binomial models. All four maintain the mean response in the interval [0, 1].

```
eta = -5:.1:5;
plot(eta,1 ./ (1 + exp(-eta)),'-', eta,normcdf(eta), '-', ...
     eta,1 - exp(-exp(eta)),'-', eta,exp(-exp(eta)),'-');
xlabel('Linear function of predictors'); ylabel('Predicted mean response');
legend('logit','probit','complementary log-log','log-log','location','east');
```



For example, we can compare a fit with the probit link to one with the logit link.

```
probitCoef = glmfit(weight,[failed tested],'binomial','probit');
probitFit = glmval(probitCoef,weight,'probit');
plot(weight,proportion,'bs', weight,logitFit,'r-', weight,probitFit,'g-');
legend('Data','Logit model','Probit model','Location','northwest');
```



It's often difficult for the data to distinguish between these four link functions, and a choice is often made on theoretical grounds.

Train Generalized Additive Model for Binary Classification

This example shows how to train a generalized additive model (GAM) with optimal parameters and how to assess the predictive performance of the trained model. The example first finds the optimal parameter values for a univariate GAM (parameters for linear terms) and then finds the values for a bivariate GAM (parameters for interaction terms). Also, the example explains how to interpret the trained model by examining local effects of terms on a specific prediction and by computing the partial dependence of the predictions on predictors.

Load Sample Data

Load the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

```
load census1994
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. To reduce the running time for this example, subsample 500 training observations and 500 test observations by using the `datasample` function.

```
rng('default')
NumSamples = 5e2;
adultdata = datasample(adultdata,NumSamples,'Replace',false);
adulttest = datasample(adulttest,NumSamples,'Replace',false);
```

Find Optimal Parameters for Univariate GAM

Optimize the parameters for a univariate GAM with respect to cross-validation by using the `bayesopt` function.

Prepare `optimizableVariable` objects for the name-value arguments of a univariate GAM: `MaxNumSplitsPerPredictor`, `NumTreesPerPredictor`, and `InitialLearnRateForPredictors`.

```
maxNumSplitsPerPredictor = optimizableVariable('maxNumSplitsPerPredictor',[1,10],'Type','integer
numTreesPerPredictor = optimizableVariable('numTreesPerPredictor',[1,500],'Type','integer');
initialLearnRateForPredictors = optimizableVariable('initialLearnRateForPredictors',[1e-3,1],'Ty
```

Create an objective function that takes an input `z = [maxNumSplitsPerPredictor,numTreesPerPredictor,initialLearnRateForPredictors]` and returns the cross-validated loss value at the parameters in `z`.

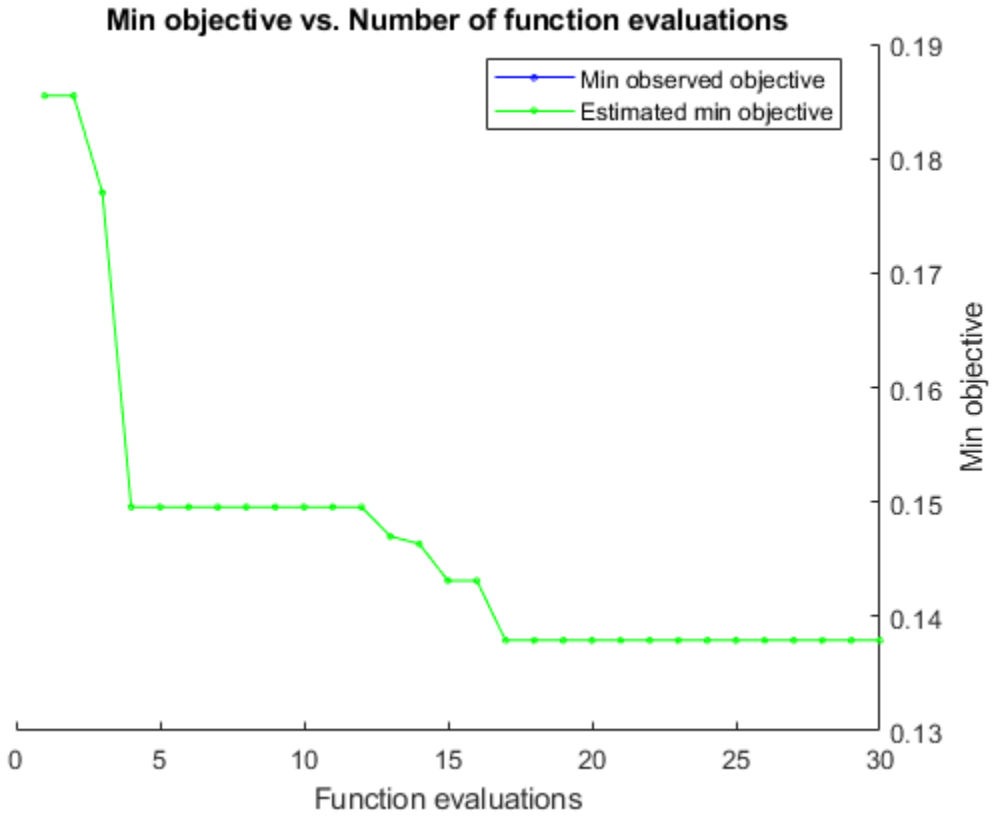
```
minfun1 = @(z)kfoldLoss(fitcgam(adultdata,'salary','Weights','fnlwgt', ...
    'CrossVal','on', ...
    'InitialLearnRateForPredictors',z.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',z.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',z.numTreesPerPredictor));
```

If you specify the cross-validation option `'CrossVal','on'`, then the `fitcgam` function returns a cross-validated model object `ClassificationPartitionedGAM`. The `kfoldLoss` function returns the classification loss obtained by the cross-validated model. Therefore, the function handle `minfun` computes the cross-validation loss at the parameters in `z`.

Search for the best parameters using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time and, therefore, can give varying results.

```
results1 = bayesopt(minfun1, ...
    [initialLearnRateForPredictors,maxNumSplitsPerPredictor,numTreesPerPredictor], ...
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForPredi | PerPredi
|=====================================================================================================
|    1 | Best   |     0.18549 |      5.6957 |     0.18549 |     0.18549 |      0.73503 |
|    2 | Accept |     0.19145 |      20.383 |     0.18549 |     0.18549 |      0.72917 |
|    3 | Best   |     0.17703 |      13.412 |     0.17703 |     0.17703 |     0.079299 |
|    4 | Best   |     0.14955 |       0.402 |     0.14955 |     0.14955 |      0.24236 |
|    5 | Accept |     0.15999 |      12.363 |     0.14955 |     0.14955 |      0.25509 |
|    6 | Accept |     0.15158 |      1.5035 |     0.14955 |     0.14955 |      0.23051 |
|    7 | Accept |     0.16181 |     0.18204 |     0.14955 |     0.14955 |      0.34396 |
|    8 | Accept |     0.15079 |     0.38418 |     0.14955 |     0.14955 |      0.26669 |
|    9 | Accept |     0.16102 |     0.55525 |     0.14955 |     0.14955 |      0.26065 |
|   10 | Accept |     0.19259 |      8.6487 |     0.14955 |     0.14955 |      0.24894 |
|   11 | Accept |     0.18628 |     0.20681 |     0.14955 |     0.14955 |      0.13389 |
|   12 | Accept |     0.15653 |     0.24643 |     0.14955 |     0.14955 |      0.24172 |
|   13 | Best   |     0.14699 |     0.82743 |     0.14699 |     0.14699 |      0.26745 |
|   14 | Best   |     0.14634 |     0.47528 |     0.14634 |     0.14634 |      0.25025 |
|   15 | Best   |     0.14312 |     0.34493 |     0.14312 |     0.14312 |      0.30452 |
|   16 | Accept |     0.14334 |     0.51583 |     0.14312 |     0.14312 |      0.33507 |
|   17 | Best   |     0.13791 |     0.32248 |     0.13791 |     0.13791 |      0.33179 |
|   18 | Accept |     0.14875 |      0.3551 |     0.13791 |     0.13791 |      0.36806 |
|   19 | Accept |      0.1651 |      1.3731 |     0.13791 |     0.13791 |      0.32691 |
|   20 | Accept |     0.15895 |     0.37324 |     0.13791 |     0.13791 |      0.32985 |
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForPredi | PerPredi
|=====================================================================================================
|   21 | Accept |     0.13946 |     0.26793 |     0.13791 |     0.13791 |      0.36721 |
|   22 | Accept |     0.16719 |      1.1276 |     0.13791 |     0.13791 |      0.25385 |
|   23 | Accept |     0.17017 |        1.35 |     0.13791 |     0.13791 |      0.23809 |
|   24 | Accept |     0.15519 |     0.46246 |     0.13791 |     0.13791 |      0.34831 |
|   25 | Accept |     0.15312 |     0.26445 |     0.13791 |     0.13791 |      0.33416 |
|   26 | Accept |     0.15852 |     0.31045 |     0.13791 |     0.13791 |       0.6142 |
|   27 | Accept |     0.16691 |     0.50559 |     0.13791 |     0.13791 |      0.31446 |
|   28 | Accept |     0.14384 |     0.35136 |     0.13791 |     0.13791 |      0.40215 |
|   29 | Accept |     0.14773 |     0.33296 |     0.13791 |     0.13791 |      0.34255 |
|   30 | Accept |     0.17604 |     0.85847 |     0.13791 |     0.13791 |      0.36565 |
```




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 97.6656 seconds
Total objective function evaluation time: 74.4022

Best observed feasible point:
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

              0.33179                           9                          4

Observed objective function value = 0.13791
Estimated objective function value = 0.13791
Function evaluation time = 0.32248

Best estimated feasible point (according to models):
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

              0.33179                           9                          4

Estimated objective function value = 0.13791
Estimated function evaluation time = 0.33084
```

Obtain the best point from `results1`.

```
zbest1 = bestPoint(results1)

zbest1=1×3 table
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

                0.33179                          9                          4
```

Train Univariate GAM with Optimal Parameters

Train an optimized GAM using the `zbest1` values. A recommended practice is to specify the class names.

```
Mdl1 = fitcgam(adultdata,'salary','Weights','fnlwgt', ...
    'ClassNames',categorical({'<=50K','>50K'}), ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor)

Mdl1 =
  ClassificationGAM
           PredictorNames: {'age'  'workClass'  'education'  'education_num'  'marital_status'
             ResponseName: 'salary'
    CategoricalPredictors: [2 3 5 6 7 8 9 13]
               ClassNames: [<=50K    >50K]
           ScoreTransform: 'logit'
                Intercept: -1.7383
          NumObservations: 500


  Properties, Methods
```

`Mdl1` is a `ClassificationGAM` model object. The model display shows a partial list of the model properties. To view the full list of the model properties, double-click the variable name `Mdl1` in the Workspace. The Variables editor opens for `Mdl1`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the `ReasonForTermination` property.

```
Mdl1.ReasonForTermination

ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: ''
```

The `PredictorTrees` field of the property value indicates that `Mdl1` includes the specified number of trees. `NumTreesPerPredictor` of `fitcgam` specifies the maximum number of trees per predictor, and the function can stop before training the requested number of trees. You can use the `ReasonForTermination` property to determine whether the trained model contains the specified number of trees.

If you specify to include interaction terms so that `fitcgam` trains trees for them, then the `InteractionTrees` field contains a nonempty value.

Find Optimal Parameters for Bivariate GAM

Find the parameters for interaction terms of a bivariate GAM by using the `bayesopt` function.

Prepare `optimizableVariable` objects for the name-value arguments for the interaction terms: `InitialLearnRateForInteractions`, `MaxNumSplitsPerInteraction`, `NumTreesPerInteraction`, and `InitialLearnRateForInteractions`.

```
initialLearnRateForInteractions = optimizableVariable('initialLearnRateForInteractions',[1e-3,1]
maxNumSplitsPerInteraction = optimizableVariable('maxNumSplitsPerInteraction',[1,10],'Type','inte
numTreesPerInteraction = optimizableVariable('numTreesPerInteraction',[1,500],'Type','integer');
numInteractions = optimizableVariable('numInteractions',[1,28],'Type','integer');
```

Create an objective function for the optimization. Use the optimal parameter values in `zbest1` so that the software finds optimal parameter values for interaction terms based on the `zbest1` values.

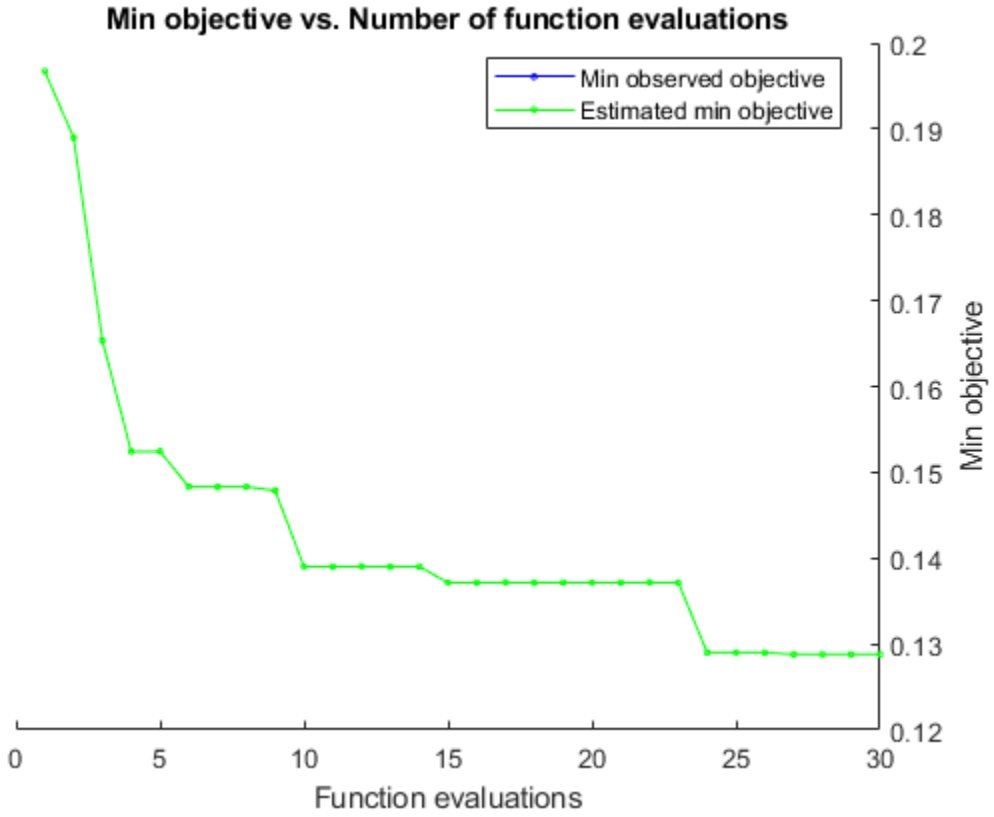
```
minfun2 = @(z)kfoldLoss(fitcgam(adultdata,'salary','Weights','fnlwgt', ...
    'CrossVal','on', ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor, ...
    'InitialLearnRateForInteractions',z.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',z.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',z.numTreesPerInteraction, ...
    'Interactions',z.numInteractions));
```

Search for the best parameters using `bayesopt`. The optimization process trains multiple models and displays warning messages if the models include no interaction terms. Disable all warnings before calling `bayesopt` and restore the warning state after running `bayesopt`. You can leave the warning state unchanged to view the warning messages.

```
orig_state = warning('query');
warning('off')
results2 = bayesopt(minfun2, ...
    [initialLearnRateForInteractions,maxNumSplitsPerInteraction,numTreesPerInteraction,numInterac
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForInter | PerInter
|=====================================================================================================
|    1 | Best   |     0.19671 |      10.999 |     0.19671 |     0.19671 |      0.96444 |
|    2 | Best   |       0.189 |       30.57 |       0.189 |       0.189 |      0.98548 |
|    3 | Best   |     0.16538 |      18.643 |     0.16538 |     0.16538 |      0.28678 |
|    4 | Best   |     0.15243 |      0.4285 |     0.15243 |     0.15243 |      0.28044 |
|    5 | Accept |     0.16065 |     0.69005 |     0.15243 |     0.15243 |      0.20151 |
|    6 | Best   |     0.14831 |     0.36629 |     0.14831 |     0.14831 |     0.032423 |
|    7 | Accept |     0.14887 |     0.36443 |     0.14831 |     0.14831 |     0.021093 |
|    8 | Accept |     0.15039 |     0.42139 |     0.14831 |     0.14831 |     0.012128 |
|    9 | Best   |     0.14787 |     0.42482 |     0.14787 |     0.14787 |      0.10119 |
|   10 | Best   |     0.13902 |     0.38822 |     0.13902 |     0.13902 |       0.1233 |
|   11 | Accept |     0.14721 |     0.39532 |     0.13902 |     0.13902 |     0.065618 |
|   12 | Accept |     0.14586 |     0.39205 |     0.13902 |     0.13902 |      0.18711 |
|   13 | Accept |     0.15073 |       0.383 |     0.13902 |     0.13902 |      0.15072 |
|   14 | Accept |     0.14966 |     0.42744 |     0.13902 |     0.13902 |      0.17155 |
|   15 | Best   |     0.13716 |     0.37599 |     0.13716 |     0.13716 |      0.12601 |
|   16 | Accept |     0.15094 |     0.38197 |     0.13716 |     0.13716 |      0.13962 |
```

```
|   17 | Accept |     0.13972 |      4.5994 |     0.13716 |     0.13716 |    0.0028545 |
|   18 | Accept |     0.14788 |       31.639 |     0.13716 |     0.13716 |    0.0024433 |
|   19 | Accept |     0.14565 |       1.276 |     0.13716 |     0.13716 |     0.013118 |
|   20 | Accept |     0.16502 |       28.315 |     0.13716 |     0.13716 |    0.0063353 |
|======|========|=============|=============|=============|=============|==============|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForInter | PerInter
|======|========|=============|=============|=============|=============|==============|
|   21 | Accept |     0.15693 |      4.9653 |     0.13716 |     0.13716 |     0.016486 |
|   22 | Accept |     0.16312 |       29.942 |     0.13716 |     0.13716 |     0.019904 |
|   23 | Accept |     0.15719 |      4.7423 |     0.13716 |     0.13716 |     0.020155 |
|   24 | Best   |       0.129 |      6.4419 |       0.129 |       0.129 |     0.090858 |
|   25 | Accept |     0.15118 |      6.6757 |       0.129 |       0.129 |      0.15943 |
|   26 | Accept |     0.15343 |      2.2035 |       0.129 |       0.129 |     0.070349 |
|   27 | Best   |     0.12879 |      6.8017 |     0.12879 |     0.12879 |     0.091985 |
|   28 | Accept |     0.19093 |      5.9262 |     0.12879 |     0.12879 |     0.067405 |
|   29 | Accept |     0.16767 |      6.3779 |     0.12879 |     0.12879 |      0.31419 |
|   30 | Accept |     0.17636 |       11.026 |     0.12879 |     0.12879 |     0.054697 |
```




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 239.1035 seconds
Total objective function evaluation time: 216.5833

Best observed feasible point:
```

```
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.091985                             5                        387

Observed objective function value = 0.12879
Estimated objective function value = 0.12879
Function evaluation time = 6.8017

Best estimated feasible point (according to models):
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.091985                             5                        387

Estimated objective function value = 0.12879
Estimated function evaluation time = 6.7245
```

```
warning(orig_state)
```

Obtain the best point from `results2`.

```
zbest2 = bestPoint(results2)
```

```
zbest2=1×4 table
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.091985                             5                        387
```

Train Bivariate GAM with Optimal Parameters

Train an optimized GAM using the `zbest1` and `zbest2` values.

```
Mdl = fitcgam(adultdata,'salary','Weights','fnlwgt', ...
    'ClassNames',categorical({'<=50K','>50K'}), ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor, ...
    'InitialLearnRateForInteractions',zbest2.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',zbest2.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',zbest2.numTreesPerInteraction, ...
    'Interactions',zbest2.numInteractions)
```

```
Mdl =
  ClassificationGAM
           PredictorNames: {'age'  'workClass'  'education'  'education_num'  'marital_status'
             ResponseName: 'salary'
    CategoricalPredictors: [2 3 5 6 7 8 9 13]
               ClassNames: [<=50K    >50K]
          ScoreTransform: 'logit'
               Intercept: -1.7755
             Interactions: [4×2 double]
         NumObservations: 500


  Properties, Methods
```

Alternatively, you can add interaction terms to the univariate GAM by using the `addInteractions` function.

```
Mdl2 = addInteractions(Mdl1,zbest2.numInteractions, ...
    'InitialLearnRateForInteractions',zbest2.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',zbest2.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',zbest2.numTreesPerInteraction);
```

The second input argument specifies the maximum number of interaction terms, and the `NumTreesPerInteraction` name-value argument specifies the maximum number of trees per interaction term. The `addInteractions` function can include fewer interaction terms and stop before training the requested number of trees. You can use the `Interactions` and `ReasonForTermination` properties to check the actual number of interaction terms and number of trees in the trained model.

Display the interaction terms in `Mdl`.

```
Mdl.Interactions
```

```
ans = 4×2

     7    10
     4     7
     7     9
     5    10
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitcgam` adds them to the model.

Display the interaction terms in `Mdl` using the predictor names.

```
Mdl.PredictorNames(Mdl.Interactions)
```

```
ans = 4×2 cell
    {'relationship'  }    {'capital_gain'}
    {'education_num' }    {'relationship'}
    {'relationship'  }    {'sex'         }
    {'marital_status'}    {'capital_gain'}
```

Display the reason for termination to determine whether the model contains the specified number of trees for each linear term and each interaction term.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Terminated after training the requested number of trees.'
```

Assess Predictive Performance on New Observations

Assess the performance of the trained model by using the test sample `adulttest` and the object functions `predict`, `loss`, `edge`, and `margin`. You can use a full or compact model with these functions.

- `predict` — Classify observations
- `loss` — Compute classification loss (misclassification rate in decimal, by default)
- `margin` — Compute classification margins
- `edge` — Compute classification edge (average of classification margins)

If you want to assess the performance of the training data set, use the resubstitution object functions: `resubPredict`, `resubLoss`, `resubMargin`, and `resubEdge`. To use these functions, you must use the full model that contains the training data.

Create a compact model to reduce the size of the trained model.

```
CMdl = compact(Mdl);
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class                                                 Attri

  CMdl      1x1             3272176  classreg.learning.classif.CompactClassificationGAM
  Mdl       1x1             3389515  ClassificationGAM
```

Predict labels and scores for the test data set (`adulttest`), and compute model statistics (loss, margin, and edge) using the test data set.

```
[labels,scores] = predict(CMdl,adulttest);
L = loss(CMdl,adulttest,'Weights',adulttest.fnlwgt);
M = margin(CMdl,adulttest);
E = edge(CMdl,adulttest,'Weights',adulttest.fnlwgt);
```

Predict labels and scores and compute the statistics without including interaction terms in the trained model.

```
[labels_nointeraction,scores_nointeraction] = predict(CMdl,adulttest,'IncludeInteractions',false)
L_nointeractions = loss(CMdl,adulttest,'Weights',adulttest.fnlwgt,'IncludeInteractions',false);
M_nointeractions = margin(CMdl,adulttest,'IncludeInteractions',false);
E_nointeractions = edge(CMdl,adulttest,'Weights',adulttest.fnlwgt,'IncludeInteractions',false);
```

Compare the results obtained by including both linear and interaction terms to the results obtained by including only linear terms.

Create a table containing the observed labels, predicted labels, and scores. Display the first eight rows of the table.

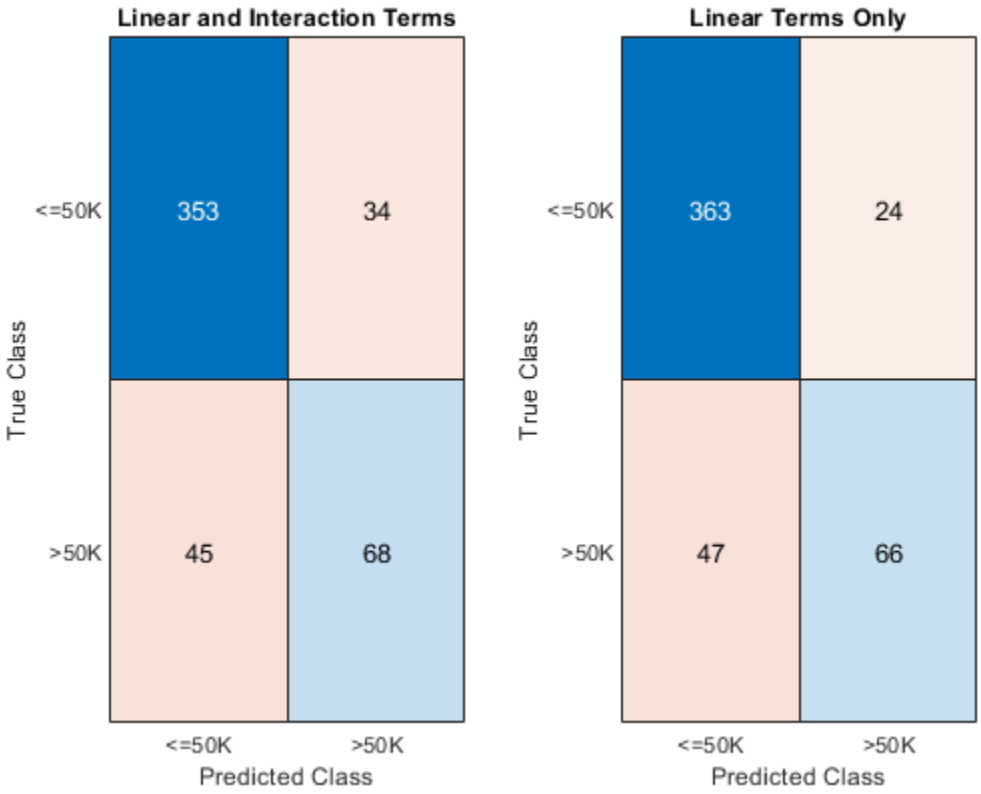
```
t = table(adulttest.salary,labels,scores,labels_nointeraction,scores_nointeraction, ...
    'VariableNames',{'True Labels','Predicted Labels','Scores' ...
    'Predicted Labels without interactions','Scores without interactions'});
head(t)
```

```
ans=8×5 table
    True Labels    Predicted Labels          Scores          Predicted Labels without interac
    ___________    ________________    ____________________    ____________________________

       <=50K           <=50K          0.97921      0.020787                 <=50K
       <=50K           <=50K                1     8.258e-17                 <=50K
       <=50K           <=50K                1    1.8297e-19                 <=50K
       <=50K           <=50K          0.87422       0.12578                 <=50K
       <=50K           <=50K                1    3.5643e-07                 <=50K
       <=50K           <=50K          0.60371       0.39629                 <=50K
       <=50K           >50K           0.49917       0.50083                 >50K
```

```
    >50K            >50K            0.3109        0.6891                    <=50K
```

Create a confusion chart from the true labels `adulttest.salary` and the predicted labels.

```
tiledlayout(1,2);
nexttile
confusionchart(adulttest.salary,labels)
title('Linear and Interaction Terms')
nexttile
confusionchart(adulttest.salary,labels_nointeraction)
title('Linear Terms Only')
```



Display the computed loss and edge values.

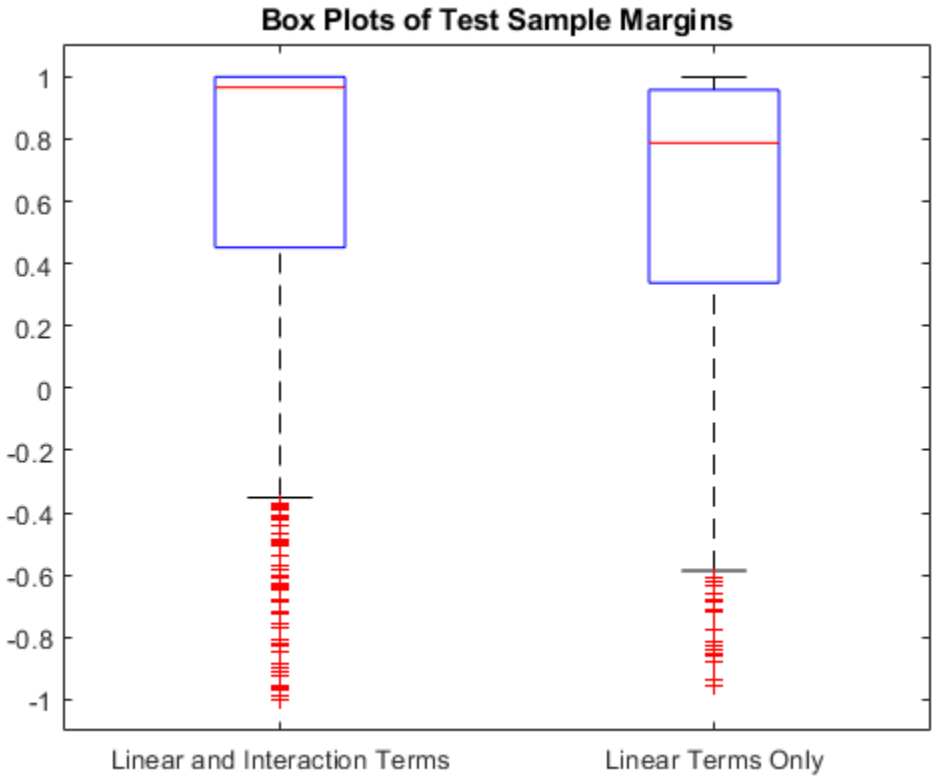
```
table([L; E], [L_nointeractions; E_nointeractions], ...
    'VariableNames',{'Linear and Interaction Terms','Only Linear Terms'}, ...
    'RowNames',{'Loss','Edge'})
```

```
ans=2×2 table
           Linear and Interaction Terms    Only Linear Terms
           ____________________________    _________________

    Loss              0.14868                  0.13852
    Edge              0.63926                  0.58405
```

The model achieves a smaller loss when only linear terms are included, but achieves a higher edge value when both linear and interaction terms are included.

Display the distributions of the margins using box plots.

```
figure
boxplot([M M_nointeractions],'Labels',{'Linear and Interaction Terms','Linear Terms Only'})
title('Box Plots of Test Sample Margins')
```



Interpret Prediction

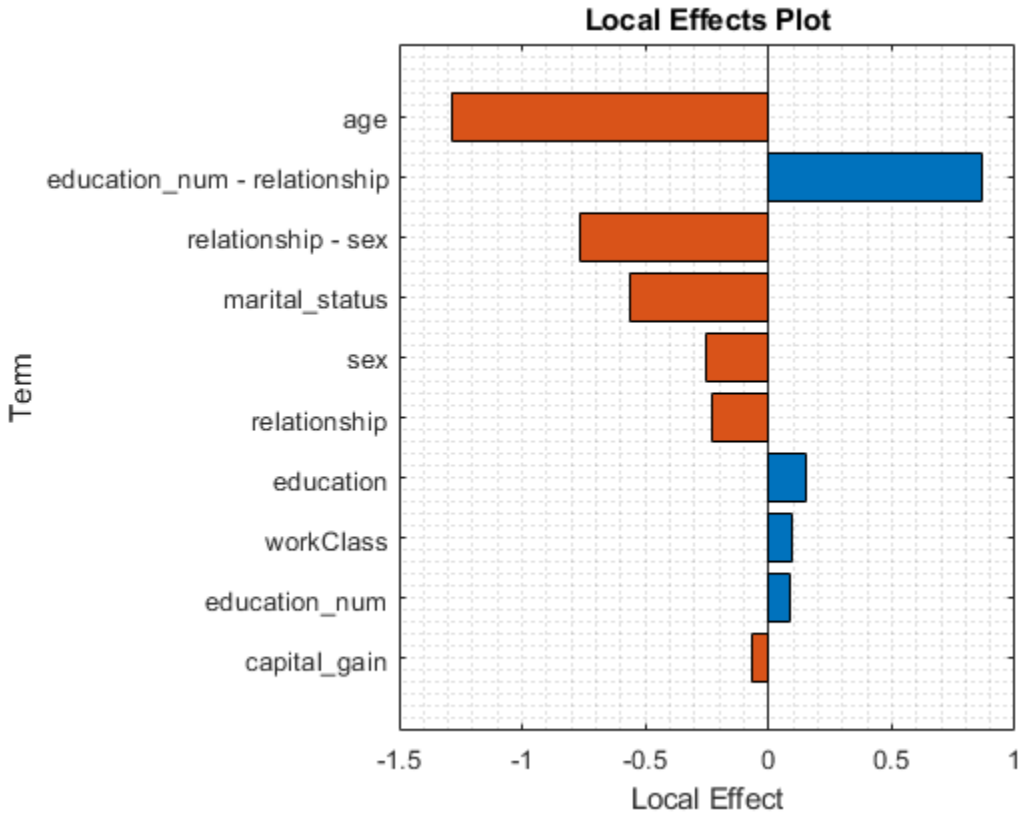
Interpret the prediction for the first test observation by using the `plotLocalEffects` function. Also, create partial dependence plots for some important terms in the model by using the `plotPartialDependence` function.

Classify the first observation of the test data, and plot the local effects of the terms in `CMdl` on the prediction. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
label = predict(CMdl,adulttest(1,:))
```

```
label = categorical
     <=50K
```

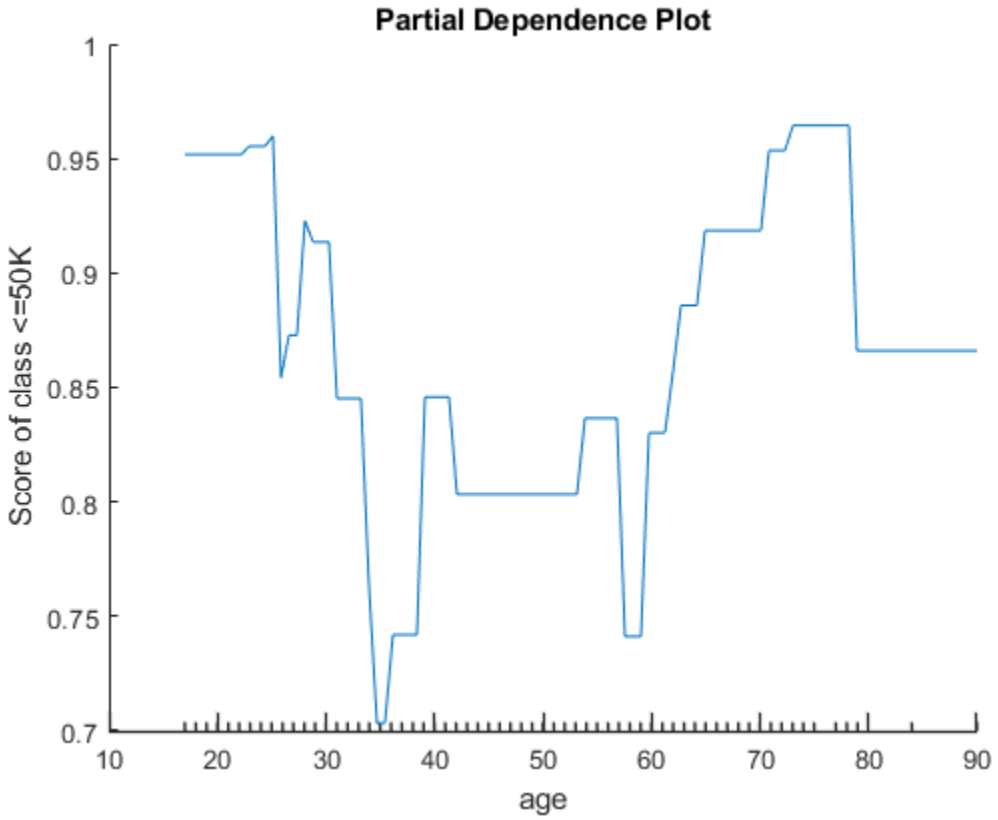
```
f1 = figure;
plotLocalEffects(CMdl,adulttest(1,:))
f1.CurrentAxes.TickLabelInterpreter = 'none';
```



The `predict` function classifies the first observation `adulttest(1,:)` as `'<=50K'`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the 10 most important terms on the prediction. Each local effect value shows the contribution of each term to the classification score for `'<=50K'`, which is the logit of the posterior probability that the classification is `'<=50K'` for the observation.

Create a partial dependence plot for the term `age`. Specify both the training and test data sets to compute the partial dependence values using both sets.

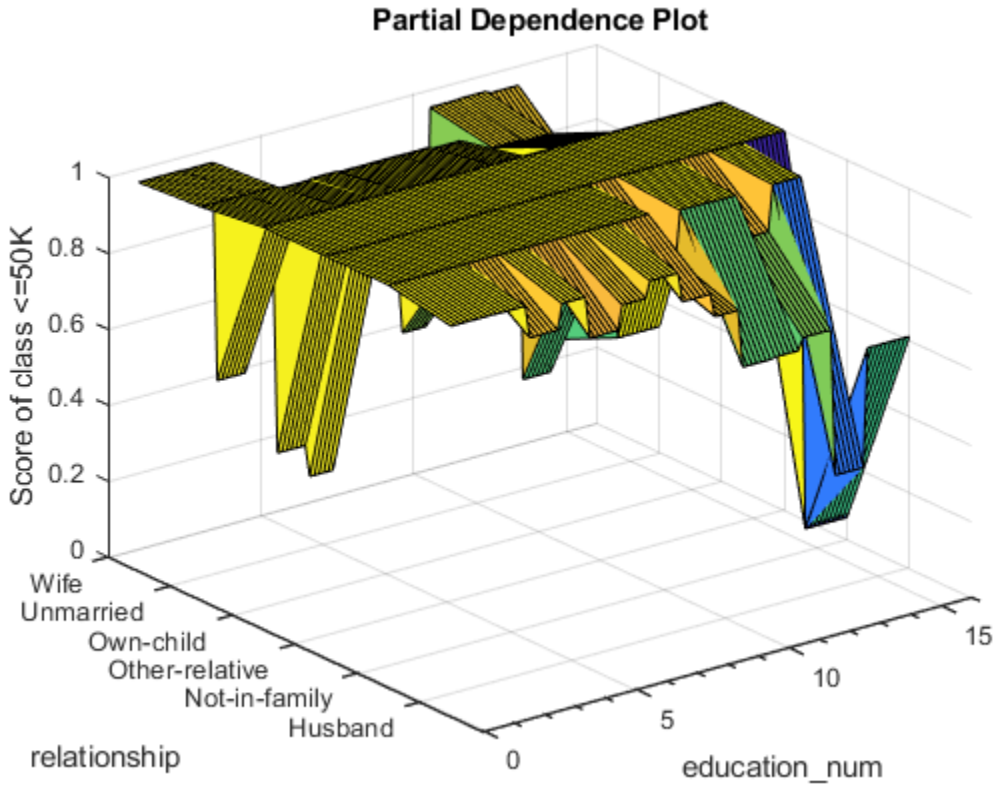
```
figure
plotPartialDependence(CMdl,'age',label,[adultdata; adulttest])
```



The plotted line represents the averaged partial relationships between the predictor `age` and the score of the class `<=50K` in the trained model. The *x*-axis minor ticks represent the unique values in the predictor `age`.

Create partial dependence plots for the terms `education_num` and `relationship`.

```
f2 = figure;
plotPartialDependence(CMdl,["education_num","relationship"],label,[adultdata; adulttest])
f2.CurrentAxes.TickLabelInterpreter = 'none';
```



The plot shows the partial dependence on `education_num`, which has a different trend depending on the `relationship` value.

See Also

`ClassificationGAM` | `CompactClassificationGAM` | `bayesopt` | `fitcgam` | `optimizableVariable` | `plotLocalEffects` | `plotPartialDependence`

Related Examples

- “Train Generalized Additive Model for Regression” on page 12-91

Train Generalized Additive Model for Regression

This example shows how to train a generalized additive model (GAM) with optimal parameters and how to assess the predictive performance of the trained model. The example first finds the optimal parameter values for a univariate GAM (parameters for linear terms) and then finds the values for a bivariate GAM (parameters for interaction terms). Also, The example explains how to interpret the trained model by examining local effects of terms on a specific prediction and by computing the partial dependence of the predictions on predictors.

Load Sample Data

Load the sample data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 10 variables with information on the sales of properties in New York City in 2015. This example uses these variables to analyze the sale prices (`SALEPRICE`).

Preprocess the data set. Assume that a `SALEPRICE` less than or equal to $1000 indicates ownership transfer without a cash consideration. Remove the samples that have this `SALEPRICE`. Also, remove the outliers identified by the `isoutlier` function. Then, convert the `datetime` array (`SALEDATE`) to the month numbers and move the response variable (`SALEPRICE`) to the last column. Change zeros in `LANDSQUAREFEET`, `GROSSSQUAREFEET`, and `YEARBUILT` to `NaN`s.

```
idx1 = NYCHousing2015.SALEPRICE <= 1000;
idx2 = isoutlier(NYCHousing2015.SALEPRICE);
NYCHousing2015(idx1|idx2,:) = [];
NYCHousing2015.SALEDATE = month(NYCHousing2015.SALEDATE);
NYCHousing2015 = movevars(NYCHousing2015,'SALEPRICE','After','SALEDATE');
NYCHousing2015.LANDSQUAREFEET(NYCHousing2015.LANDSQUAREFEET == 0) = NaN;
NYCHousing2015.GROSSSQUAREFEET(NYCHousing2015.GROSSSQUAREFEET == 0) = NaN;
NYCHousing2015.YEARBUILT(NYCHousing2015.YEARBUILT == 0) = NaN;
```

Display the first three rows of the table.

```
head(NYCHousing2015,3)
```

```
ans=3×10 table
    BOROUGH    NEIGHBORHOOD       BUILDINGCLASSCATEGORY        RESIDENTIALUNITS    COMMERCIALUNI
    _______    ____________    ___________________________    ________________    _____________

       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   1
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   2
```

Randomly select 1000 samples by using the `datasample` function, and partition observations into a training set and a test set by using the `cvpartition` function. Specify a 10% holdout sample for testing.

```
rng('default') % For reproducibility
NumSamples = 1e3;
NYCHousing2015 = datasample(NYCHousing2015,NumSamples,'Replace',false);
cv = cvpartition(size(NYCHousing2015,1),'HoldOut',0.10);
```

Extract the training and test indices, and create tables for training and test data sets.

```
tbl_training = NYCHousing2015(training(cv),:);
tbl_test = NYCHousing2015(test(cv),:);
```

Find Optimal Parameters for Univariate GAM

Optimize the parameters for a univariate GAM with respect to cross-validation by using the `bayesopt` function.

Prepare `optimizableVariable` objects for the name-value arguments of a univariate GAM: `MaxNumSplitsPerPredictor`, `NumTreesPerPredictor`, and `InitialLearnRateForPredictors`.

```
maxNumSplitsPerPredictor = optimizableVariable('maxNumSplitsPerPredictor',[1,10],'Type','integer
numTreesPerPredictor = optimizableVariable('numTreesPerPredictor',[1,500],'Type','integer');
initialLearnRateForPredictors = optimizableVariable('initialLearnRateForPredictors',[1e-3,1],'Ty
```

Create an objective function that takes an input `z = [maxNumSplitsPerPredictor,numTreesPerPredictor,initialLearnRateForPredictors]` and returns the cross-validated loss value at the parameters in `z`.

```
minfun1 = @(z)kfoldLoss(fitrgam(tbl_training,'SALEPRICE', ...
    'CrossVal','on', ...
    'InitialLearnRateForPredictors',z.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',z.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',z.numTreesPerPredictor));
```

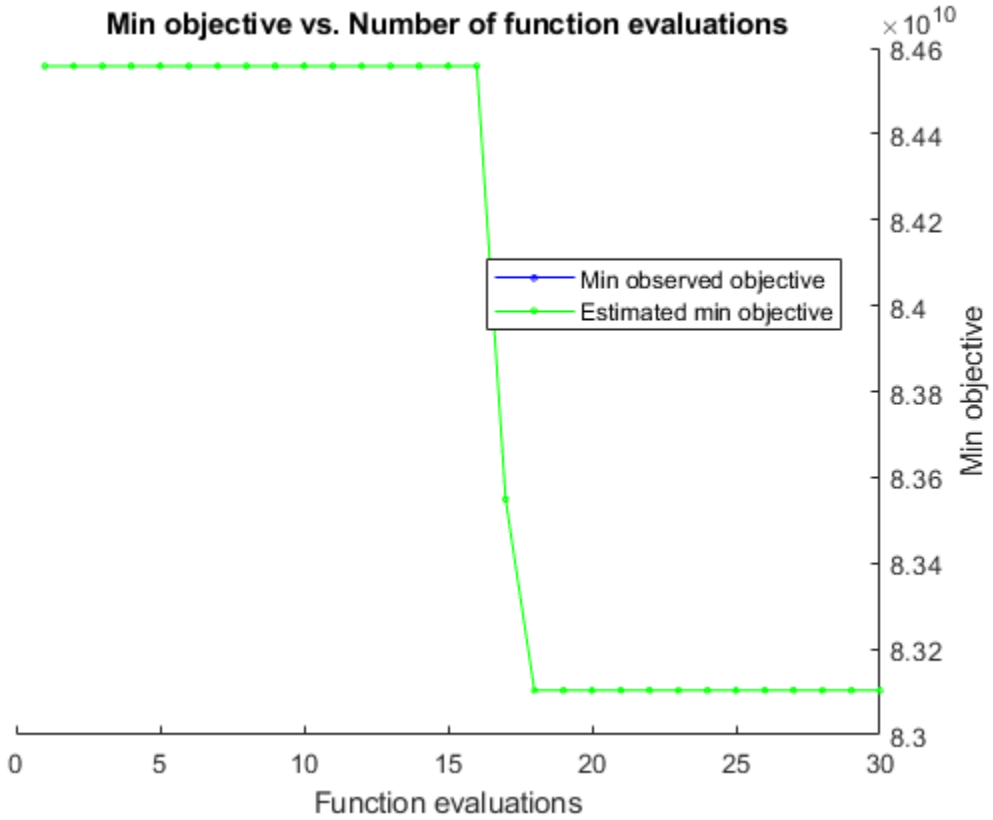
If you specify the cross-validation option `'CrossVal','on'`, then the `fitrgam` function returns a cross-validated model object `RegressionPartitionedGAM`. The `kfoldLoss` function returns the regression loss (mean squared error) obtained by the cross-validated model. Therefore, the function handle `minfun1` computes the cross-validation loss at the parameters in `z`.

Search for the best parameters using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time and, therefore, can give varying results.

```
results1 = bayesopt(minfun1, ...
    [initialLearnRateForPredictors,maxNumSplitsPerPredictor,numTreesPerPredictor], ...
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForPredi | PerPredi
|=====================================================================================================
|    1 | Best   |  8.4558e+10 |      1.5106 |  8.4558e+10 |  8.4558e+10 |      0.36695 |
|    2 | Accept |  8.6891e+10 |       12.01 |  8.4558e+10 |  8.4558e+10 |     0.008213 |
|    3 | Accept |  9.6521e+10 |      1.9121 |  8.4558e+10 |  8.4558e+10 |      0.22984 |
|    4 | Accept |  1.3402e+11 |      14.388 |  8.4558e+10 |  8.4558e+10 |      0.99932 |
|    5 | Accept |  8.7852e+10 |      13.595 |  8.4558e+10 |  8.4558e+10 |      0.16575 |
|    6 | Accept |  9.3041e+10 |      11.002 |  8.4558e+10 |  8.4558e+10 |      0.49477 |
|    7 | Accept |  1.0558e+11 |      7.7647 |  8.4558e+10 |  8.4558e+10 |      0.24562 |
|    8 | Accept |  8.8841e+10 |      1.5763 |  8.4558e+10 |  8.4558e+10 |      0.39298 |
|    9 | Accept |  9.9227e+10 |      14.377 |  8.4558e+10 |  8.4558e+10 |     0.091879 |
|   10 | Accept |  9.8611e+10 |     0.14914 |  8.4558e+10 |  8.4558e+10 |      0.22487 |
|   11 | Accept |  1.2998e+11 |      23.962 |  8.4558e+10 |  8.4558e+10 |      0.25341 |
|   12 | Accept |  8.8968e+10 |      5.0028 |  8.4558e+10 |  8.4558e+10 |      0.33109 |
|   13 | Accept |  1.2018e+11 |      1.8004 |  8.4558e+10 |  8.4558e+10 |    0.0030413 |
```

```
|   14 | Accept |  8.7503e+10 |     0.79283 |  8.4558e+10 |  8.4558e+10 |      0.33877 |
|   15 | Accept |  9.3798e+10 |      2.9578 |  8.4558e+10 |  8.4558e+10 |      0.32926 |
|   16 | Accept |  9.5165e+10 |      8.0635 |  8.4558e+10 |  8.4558e+10 |      0.33878 |
|   17 | Best   |  8.3549e+10 |     0.24446 |  8.3549e+10 |  8.3549e+10 |       0.3552 |
|   18 | Best   |  8.3104e+10 |      1.4534 |  8.3104e+10 |  8.3104e+10 |       0.2526 |
|   19 | Accept |  8.6938e+10 |      3.3234 |  8.3104e+10 |  8.3104e+10 |      0.18293 |
|   20 | Accept |  8.7531e+10 |      2.8096 |  8.3104e+10 |  8.3104e+10 |       0.2781 |
|======|========|=============|=============|=============|=============|==============|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForPredi | PerPredi
|======|========|=============|=============|=============|=============|==============|
|   21 | Accept |  9.1613e+10 |      13.347 |  8.3104e+10 |  8.3104e+10 |      0.31722 |
|   22 | Accept |   8.678e+10 |      10.358 |  8.3104e+10 |  8.3104e+10 |      0.11269 |
|   23 | Accept |  8.3614e+10 |     0.47001 |  8.3104e+10 |  8.3104e+10 |      0.22278 |
|   24 | Accept |  1.3203e+11 |       1.069 |  8.3104e+10 |  8.3104e+10 |    0.0021552 |
|   25 | Accept |    8.66e+10 |       7.233 |  8.3104e+10 |  8.3104e+10 |      0.11469 |
|   26 | Accept |  8.4535e+10 |      8.7657 |  8.3104e+10 |  8.3104e+10 |    0.0090628 |
|   27 | Accept |  1.0315e+11 |      12.297 |  8.3104e+10 |  8.3104e+10 |    0.0014094 |
|   28 | Accept |  9.6736e+10 |      5.8323 |  8.3104e+10 |  8.3104e+10 |    0.0040429 |
|   29 | Accept |  8.3651e+10 |      8.4999 |  8.3104e+10 |  8.3104e+10 |      0.09375 |
|   30 | Accept |  8.7977e+10 |      13.521 |  8.3104e+10 |  8.3104e+10 |     0.016448 |
```




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 245.1541 seconds
```

```
Total objective function evaluation time: 210.0881

Best observed feasible point:
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

                0.2526                           1                        49

Observed objective function value = 83103839919.908
Estimated objective function value = 83103840296.3186
Function evaluation time = 1.4534

Best estimated feasible point (according to models):
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

                0.2526                           1                        49

Estimated objective function value = 83103840296.3186
Estimated function evaluation time = 1.803
```

Obtain the best point from `results1`.

```
zbest1 = bestPoint(results1)
```

```
zbest1=1×3 table
    initialLearnRateForPredictors    maxNumSplitsPerPredictor    numTreesPerPredictor
    _____________________________    ________________________    ____________________

                0.2526                           1                        49
```

Train Univariate GAM with Optimal Parameters

Train an optimized GAM using the `zbest1` values.

```
Mdl1 = fitrgam(tbl_training,'SALEPRICE', ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor)
```

```
Mdl1 =
  RegressionGAM
             PredictorNames: {'BOROUGH'  'NEIGHBORHOOD'  'BUILDINGCLASSCATEGORY'  'RESIDENTIALUNITS
               ResponseName: 'SALEPRICE'
    CategoricalPredictors: [2 3]
        ResponseTransform: 'none'
                Intercept: 4.9806e+05
          NumObservations: 900


  Properties, Methods
```

`Mdl1` is a `RegressionGAM` model object. The model display shows a partial list of the model properties. To view the full list of properties, double-click the variable name `Mdl1` in the Workspace. The Variables editor opens for `Mdl1`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the `ReasonForTermination` property.

```
Mdl1.ReasonForTermination

ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: ''
```

The `PredictorTrees` field of the property value indicates that `Mdl1` includes the specified number of trees. `NumTreesPerPredictor` of `fitrgam` specifies the maximum number of trees per predictor, and the function can stop before training the requested number of trees. You can use the `ReasonForTermination` property to determine whether the trained model contains the specified number of trees.

If you specify to include interaction terms so that `fitrgam` trains trees for them. the `InteractionTrees` field contains a nonempty value.

Find Optimal Parameters for Bivariate GAM

Find the parameters for interaction terms of a bivariate GAM by using the `bayesopt` function.

Prepare `optimizableVariable` for the name-value arguments for the interaction terms: `InitialLearnRateForInteractions`, `MaxNumSplitsPerInteraction`, `NumTreesPerInteraction`, and `InitialLearnRateForInteractions`.

```
initialLearnRateForInteractions = optimizableVariable('initialLearnRateForInteractions',[1e-3,1]
maxNumSplitsPerInteraction = optimizableVariable('maxNumSplitsPerInteraction',[1,10],'Type','inte
numTreesPerInteraction = optimizableVariable('numTreesPerInteraction',[1,500],'Type','integer');
numInteractions = optimizableVariable('numInteractions',[1,28],'Type','integer');
```

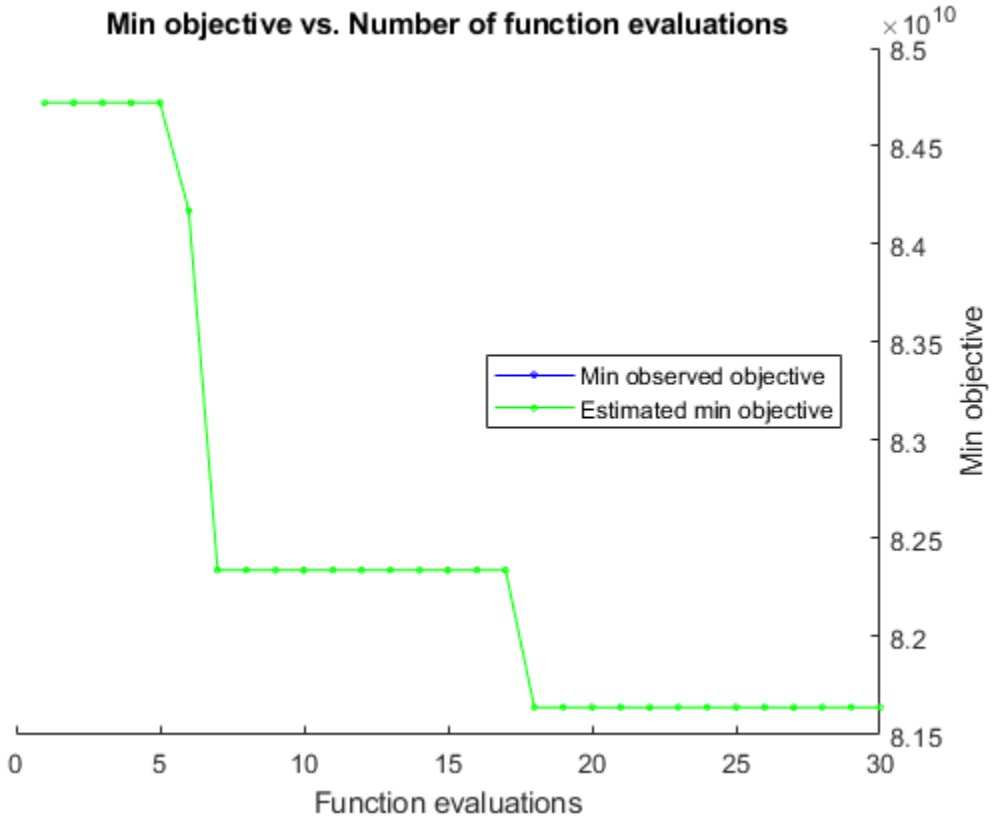
Create an objective function for the optimization. Use the optimal parameter values in `zbest1` so that the software finds optimal parameter values for interaction terms based on the `zbest1` values.

```
minfun2 = @(z)kfoldLoss(fitrgam(tbl_training,'SALEPRICE', ...
    'CrossVal','on', ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor, ...
    'InitialLearnRateForInteractions',z.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',z.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',z.numTreesPerInteraction, ...
    'Interactions',z.numInteractions));
```

Search for the best parameters using `bayesopt`. The optimization process trains multiple models and displays warning messages if the models include no interaction terms. Disable all warnings before calling `bayesopt` and restore the warning state after running `bayesopt`. You can leave the warning state unchanged to view the warning messages.

```
orig_state = warning('query');
warning('off')
results2 = bayesopt(minfun2, ...
    [initialLearnRateForInteractions,maxNumSplitsPerInteraction,numTreesPerInteraction,numInterac
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');

|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | initialLearn-| maxNumSp
|      | result |             | runtime     | (observed)  | (estim.)    | RateForInter | PerInter
```

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	initialLearn-RateForInter	maxNumSp PerInte
1	Best	8.4721e+10	1.6996	8.4721e+10	8.4721e+10	0.41774	
2	Accept	9.1765e+10	8.3313	8.4721e+10	8.4721e+10	0.9565	
3	Accept	9.2116e+10	2.8341	8.4721e+10	8.4721e+10	0.33578	
4	Accept	1.784e+11	76.237	8.4721e+10	8.4721e+10	0.91186	
5	Accept	8.4906e+10	1.8275	8.4721e+10	8.4721e+10	0.296	
6	Best	8.4172e+10	1.73	8.4172e+10	8.4172e+10	0.68133	
7	Best	8.234e+10	1.7164	8.234e+10	8.234e+10	0.13943	
8	Accept	8.3488e+10	1.6382	8.234e+10	8.234e+10	0.46764	
9	Accept	8.7977e+10	1.5655	8.234e+10	8.234e+10	0.8385	
10	Accept	8.4431e+10	1.5744	8.234e+10	8.234e+10	0.95535	
11	Accept	8.5784e+10	1.7478	8.234e+10	8.234e+10	0.023058	
12	Accept	8.6068e+10	1.7304	8.234e+10	8.234e+10	0.77118	
13	Accept	8.7004e+10	1.5903	8.234e+10	8.234e+10	0.016991	
14	Accept	8.3325e+10	1.5895	8.234e+10	8.234e+10	0.9468	
15	Accept	8.4097e+10	1.6357	8.234e+10	8.234e+10	0.97988	
16	Accept	8.3106e+10	1.6081	8.234e+10	8.234e+10	0.024052	
17	Accept	8.469e+10	1.6235	8.234e+10	8.234e+10	0.047902	
18	Best	8.1641e+10	1.5833	8.1641e+10	8.1641e+10	0.99848	
19	Accept	8.5957e+10	1.6305	8.1641e+10	8.1641e+10	0.99826	
20	Accept	8.2486e+10	1.6515	8.1641e+10	8.1641e+10	0.36059	
21	Accept	8.6534e+10	1.647	8.1641e+10	8.1641e+10	0.0089186	
22	Accept	8.5425e+10	1.5316	8.1641e+10	8.1641e+10	0.99842	
23	Accept	8.515e+10	1.5728	8.1641e+10	8.1641e+10	0.99934	
24	Accept	8.593e+10	1.6086	8.1641e+10	8.1641e+10	0.0099052	
25	Accept	8.7394e+10	1.577	8.1641e+10	8.1641e+10	0.96502	
26	Accept	8.618e+10	1.5714	8.1641e+10	8.1641e+10	0.097871	
27	Accept	8.5704e+10	1.665	8.1641e+10	8.1641e+10	0.056356	
28	Accept	9.5451e+10	2.8821	8.1641e+10	8.1641e+10	0.91844	
29	Accept	8.4013e+10	1.5633	8.1641e+10	8.1641e+10	0.68016	
30	Accept	8.3928e+10	1.7715	8.1641e+10	8.1641e+10	0.07259	




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 155.1459 seconds
Total objective function evaluation time: 132.9347

Best observed feasible point:
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.99848                            6                          1

Observed objective function value = 81640836929.8637
Estimated objective function value = 81640841484.6238
Function evaluation time = 1.5833

Best estimated feasible point (according to models):
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.99848                            6                          1

Estimated objective function value = 81640841484.6238
Estimated function evaluation time = 1.5784

warning(orig_state)
```

Obtain the best point from `results2`.

```
zbest2 = bestPoint(results2)
```

```
zbest2=1×4 table
    initialLearnRateForInteractions    maxNumSplitsPerInteraction    numTreesPerInteraction    nu
    _______________________________    __________________________    ______________________    __

                0.99848                             6                          1
```

Train Bivariate GAM with Optimal Parameters

Train an optimized GAM using the `zbest1` and `zbest2` values.

```
Mdl = fitrgam(tbl_training,'SALEPRICE', ...
    'InitialLearnRateForPredictors',zbest1.initialLearnRateForPredictors, ...
    'MaxNumSplitsPerPredictor',zbest1.maxNumSplitsPerPredictor, ...
    'NumTreesPerPredictor',zbest1.numTreesPerPredictor, ...
    'InitialLearnRateForInteractions',zbest2.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',zbest2.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',zbest2.numTreesPerInteraction, ...
    'Interactions',zbest2.numInteractions)
```

```
Mdl =
  RegressionGAM
             PredictorNames: {'BOROUGH'  'NEIGHBORHOOD'  'BUILDINGCLASSCATEGORY'  'RESIDENTIALUNITS
               ResponseName: 'SALEPRICE'
    CategoricalPredictors: [2 3]
        ResponseTransform: 'none'
                Intercept: 4.9741e+05
             Interactions: [3×2 double]
          NumObservations: 900


  Properties, Methods
```

Alternatively, you can add interaction terms to the univariate GAM by using the `addInteractions` function.

```
Mdl2 = addInteractions(Mdl1,zbest2.numInteractions, ...
    'InitialLearnRateForInteractions',zbest2.initialLearnRateForInteractions, ...
    'MaxNumSplitsPerInteraction',zbest2.maxNumSplitsPerInteraction, ...
    'NumTreesPerInteraction',zbest2.numTreesPerInteraction);
```

The second input argument specifies the maximum number of interaction terms, and the `NumTreesPerInteraction` name-value argument specifies the maximum number of trees per interaction term. The `addInteractions` function can include fewer interaction terms and stop before training the requested number of trees. You can use the `Interactions` and `ReasonForTermination` properties to check the actual number of interaction terms and number of trees in the trained model.

Display the interaction terms in `Mdl`.

```
Mdl.Interactions
```

```
ans = 3×2
```

```
     3     6
     4     6
     6     8
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitrgam` adds them to the model.

Display the interaction terms in `Mdl` using the predictor names.

```
Mdl.PredictorNames(Mdl.Interactions)
```

```
ans = 3×2 cell
    {'BUILDINGCLASSCATEGORY'}    {'LANDSQUAREFEET'}
    {'RESIDENTIALUNITS'     }    {'LANDSQUAREFEET'}
    {'LANDSQUAREFEET'       }    {'YEARBUILT'     }
```

Display the reason for termination to determine whether the model contains the specified number of trees for each linear term and each interaction term.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Terminated after training the requested number of trees.'
```

Assess Predictive Performance on New Observations

Assess the performance of the trained model by using the test sample `tbl_test` and the object functions `predict` and `loss`. You can use a full or compact model with these functions.

- `predict` — Predict responses
- `loss` — Compute regression loss (mean squared error, by default)

If you want to assess the performance of the training data set, use the resubstitution object functions: `resubPredict` and `resubLoss`. To use these functions, you must use the full model that contains the training data.

Create a compact model to reduce the size of the trained model.

```
CMdl = compact(Mdl);
whos('Mdl','CMdl')
```

```
  Name      Size             Bytes  Class                                          Attributes

  CMdl      1x1             370211  classreg.learning.regr.CompactRegressionGAM
  Mdl       1x1             528102  RegressionGAM
```

Predict responses and compute mean squared errors for the test data set `tbl_test`.

```
yFit = predict(CMdl,tbl_test);
L = loss(CMdl,tbl_test)
```

```
L = 1.2855e+11
```

Find predicted responses and errors without including interaction terms in the trained model.

```
yFit_nointeraction = predict(CMdl,tbl_test,'IncludeInteractions',false);
L_nointeractions = loss(CMdl,tbl_test,'IncludeInteractions',false)
```

```
L_nointeractions = 1.3031e+11
```

The model achieves a smaller error for the test data set when both linear and interaction terms are included.

Compare the results obtained by including both linear to interaction terms and the results obtained by including only linear terms. Create a table containing the observed responses and predicted responses. Display the first eight rows of the table.

```
t = table(tbl_test.SALEPRICE,yFit,yFit_nointeraction, ...
    'VariableNames',{'Observed Value','Predicted Response','Predicted Response Without Interactio
head(t)
```

```
ans=8×3 table
    Observed Value    Predicted Response    Predicted Response Without Interactions
    ______________    __________________    ______________________________________

         3.6e+05          4.9812e+05                    5.2712e+05
         1.8e+05          2.7349e+05                    2.7415e+05
         1.9e+05          3.3682e+05                    3.3748e+05
        4.26e+05            6.15e+05                    5.6542e+05
        3.91e+05          3.1262e+05                    3.1328e+05
         2.3e+05          1.0606e+05                    1.0672e+05
      4.7333e+05          1.0773e+06                    1.1399e+06
           2e+05          2.9506e+05                     3.305e+05
```

Interpret Prediction

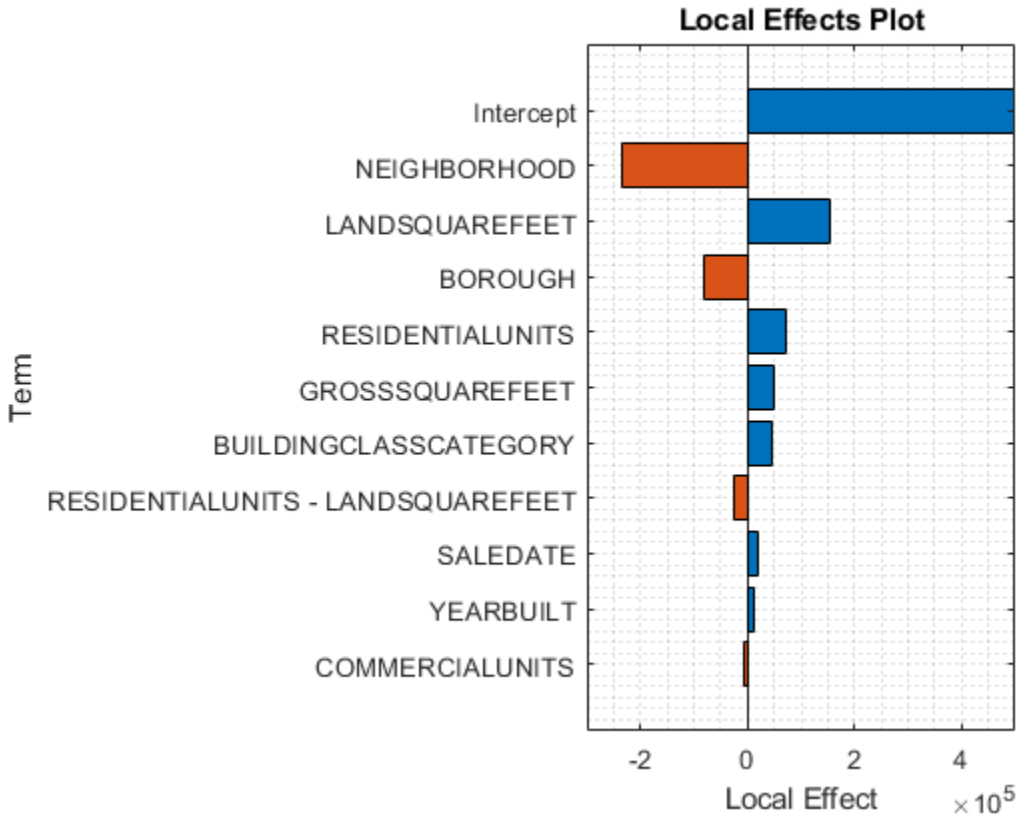
Interpret the prediction for the first test observation by using the `plotLocalEffects` function. Also, create partial dependence plots for some important terms in the model by using the `plotPartialDependence` function.

Predict a response value for the first observation of the test data, and plot the local effects of the terms in `CMdl` on the prediction. Specify `'IncludeIntercept',true` to include the intercept term in the plot.

```
yFit = predict(CMdl,tbl_test(1,:))
```

```
yFit = 4.9812e+05
```

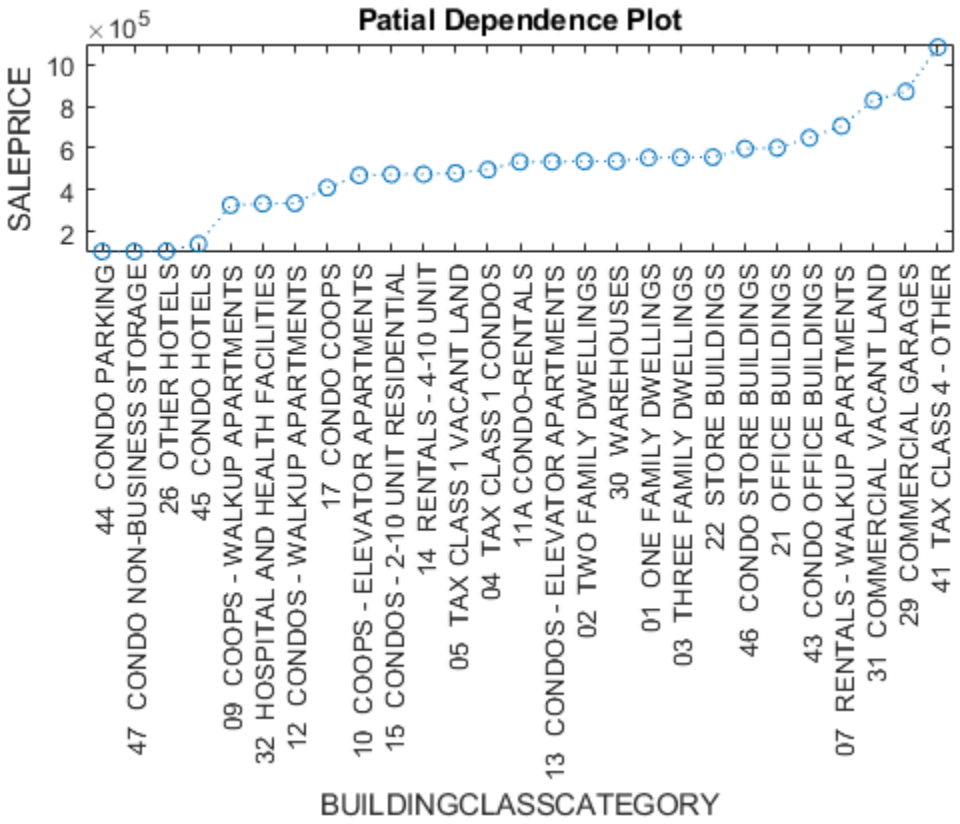
```
plotLocalEffects(CMdl,tbl_test(1,:),'IncludeIntercept',true)
```



The `predict` function returns the sale price for the first observation `tbl_test(1,:)`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the terms in `CMdl` on the prediction. Each local effect value shows the contribution of each term to the predicted sale price for `tbl_test(1,:)`.

Compute the partial dependence values for `BUILDINGCLASSCATEGORY` and plot the sorted values. Specify both the training and test data sets to compute the partial dependence values using both sets.

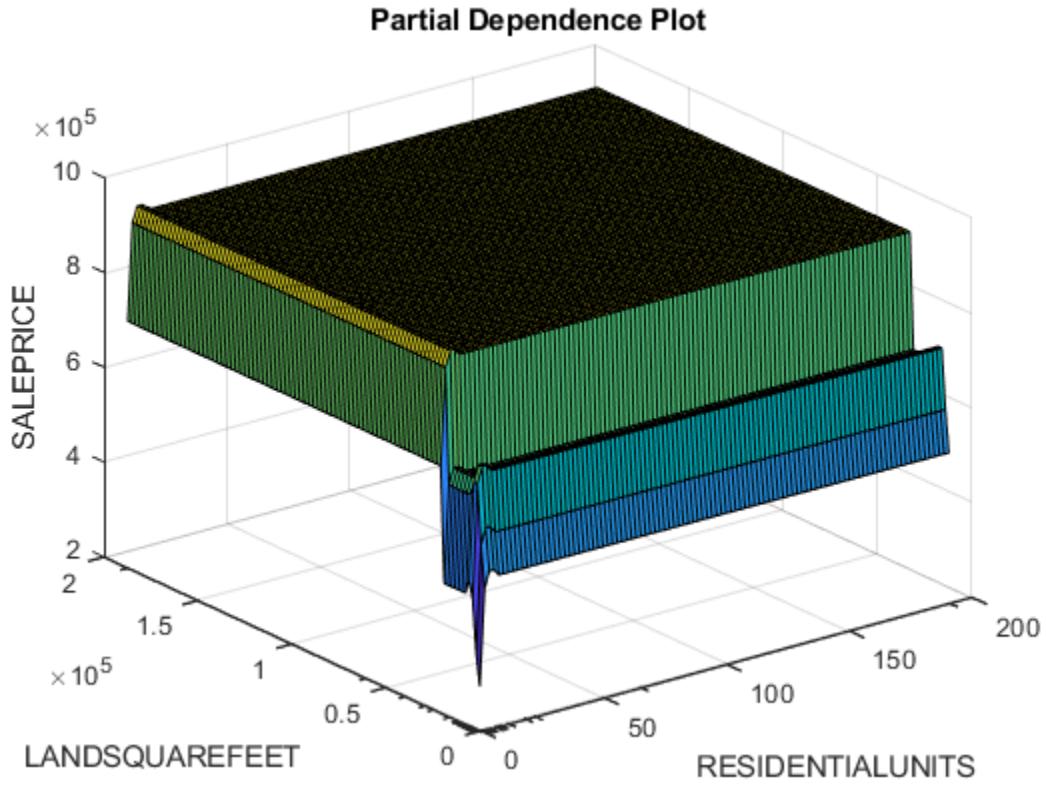
```
[pd,x,y] = partialDependence(CMdl,'BUILDINGCLASSCATEGORY',[tbl_training; tbl_test]);
[pd_sorted,I] = sort(pd);
x_sorted = x(I);
x_sorted = reordercats(x_sorted,I);
figure
plot(x_sorted,pd_sorted,'o:')
xlabel('BUILDINGCLASSCATEGORY')
ylabel('SALEPRICE')
title('Patial Dependence Plot')
```



The plotted line represents the averaged partial relationships between the predictor `BUILDINGCLASSCATEGORY` and the response `SALEPRICE` in the trained model.

Create a partial dependence plot for the terms `RESIDENTIALUNITS` and `LANDSQUAREFEET`.

```
figure
plotPartialDependence(CMdl,["RESIDENTIALUNITS","LANDSQUAREFEET"],[tbl_training; tbl_test])
```



The minor ticks in the *x*-axis (`RESIDENTIALUNITS`) and *y*-axis (`LANDSQUAREFEET`) represent the unique values of the predictors in the specified data. The predictor values include a few outliers, and most of the `RESIDENTIALUNITS` and `LANDSQUAREFEET` values are less than 10 and 50,000, respectively. The plot shows that the `SALEPRICE` values do not vary significantly when the `RESIDENTIALUNITS` and `LANDSQUAREFEET` values are greater than 10 and 50,000.

See Also

`CompactRegressionGAM` | `RegressionGAM` | `bayesopt` | `fitrgam` | `optimizableVariable` | `plotLocalEffects` | `plotPartialDependence`

Related Examples

- “Train Generalized Additive Model for Binary Classification” on page 12-77

Nonlinear Regression

Nonlinear Regression

In this section...

What Are Parametric Nonlinear Regression Models?

Parametric nonlinear models represent the relationship between a continuous response variable and one or more continuous predictor variables in the form

$y = f(X,\beta) + \varepsilon,$

where

- y is an n-by-1 vector of observations of the response variable.
- f is any function of X and β that evaluates each row of X along with the vector β to compute the prediction for the corresponding row of y.
- X is an n-by-p matrix of predictors, with one row for each observation, and one column for each predictor.
- β is a p-by-1 vector of unknown parameters to be estimated.
- ε is an n-by-1 vector of independent, identically distributed random disturbances.

In contrast, nonparametric models do not attempt to characterize the relationship between predictors and response with model parameters. Descriptions are often graphical, as in the case of “Decision Trees” on page 19-2.

`fitnlm` attempts to find values of the parameters β that minimize the mean squared differences between the observed responses y and the predictions of the model $f(X,\beta)$. To do so, it needs a starting value `beta0` before iteratively modifying the vector β to a vector with minimal mean squared error.

Prepare Data

To begin fitting a regression, put your data into a form that fitting functions expect. All regression techniques begin with input data in an array `X` and response data in a separate vector `y`, or input data in a table or dataset array `tbl` and response data as a column in `tbl`. Each row of the input data represents one observation. Each column represents one predictor (variable).

For a table or dataset array `tbl`, indicate the response variable with the `'ResponseVar'` name-value pair:

```
mdl = fitlm(tbl,'ResponseVar','BloodPressure');
```

The response variable is the last column by default.

You cannot use categorical predictors for nonlinear regression. A categorical predictor is one that takes values from a fixed set of possibilities.

Represent missing data as `NaN` for both input data and response data.

Dataset Array for Input and Response Data

For example, to create a dataset array from an Excel spreadsheet:

```
ds = dataset('XLSFile','hospital.xls',...
     'ReadObsNames',true);
```

To create a dataset array from workspace variables:

```
load carsmall
ds = dataset(Weight,Model_Year,MPG);
```

Table for Input and Response Data

To create a table from an Excel spreadsheet:

```
tbl = readtable('hospital.xls',...
     'ReadRowNames',true);
```

To create a table from workspace variables:

```
load carsmall
tbl = table(Weight,Model_Year,MPG);
```

Numeric Matrix for Input Data and Numeric Vector for Response

For example, to create numeric arrays from workspace variables:

```
load carsmall
X = [Weight Horsepower Cylinders Model_Year];
y = MPG;
```

To create numeric arrays from an Excel spreadsheet:

```
[X, Xnames] = xlsread('hospital.xls');
y = X(:,4); % response y is systolic pressure
X(:,4) = []; % remove y from the X matrix
```

Notice that the nonnumeric entries, such as `sex`, do not appear in `X`.

Represent the Nonlinear Model

There are several ways to represent a nonlinear model. Use whichever is most convenient.

The nonlinear model is a required input to `fitnlm`, in the `modelfun` input.

`fitnlm` assumes that the response function $f(X,\beta)$ is smooth in the parameters β. If your function is not smooth, `fitnlm` can fail to provide optimal parameter estimates.

- “Function Handle to Anonymous Function or Function File” on page 13-4

• “Text Representation of Formula” on page 13-4

Function Handle to Anonymous Function or Function File

The function handle @*modelfun*(b,x) accepts a vector b and matrix, table, or dataset array x. The function handle should return a vector f with the same number of rows as x. For example, the function file hougen.m computes

$$\text{hougen}(b,x) = \frac{b(1)x(2) - x(3)/b(5)}{1 + b(2)x(1) + b(3)x(2) + b(4)x(3)}.$$

Examine the function by entering type hougen at the MATLAB command line.

```
function yhat = hougen(beta,x)
%HOUGEN Hougen-Watson model for reaction kinetics.
%   YHAT = HOUGEN(BETA,X) gives the predicted values of the
%   reaction rate, YHAT, as a function of the vector of
%   parameters, BETA, and the matrix of data, X.
%   BETA must have 5 elements and X must have three
%   columns.
%
%   The model form is:
%   y = (b1*x2 - x3/b5)./(1+b2*x1+b3*x2+b4*x3)
%
%   Reference:
%      [1]  Bates, Douglas, and Watts, Donald, "Nonlinear
%      Regression Analysis and Its Applications", Wiley
%      1988 p. 271-272.

%   Copyright 1993-2004 The MathWorks, Inc.
%   B.A. Jones 1-06-95.

b1 = beta(1);
b2 = beta(2);
b3 = beta(3);
b4 = beta(4);
b5 = beta(5);

x1 = x(:,1);
x2 = x(:,2);
x3 = x(:,3);

yhat = (b1*x2 - x3/b5)./(1+b2*x1+b3*x2+b4*x3);
```

You can write an anonymous function that performs the same calculation as hougen.m.

```
modelfun = @(b,x)(b(1)*x(:,2) - x(:,3)/b(5))./...
(1 + b(2)*x(:,1) + b(3)*x(:,2) + b(4)*x(:,3));
```

Text Representation of Formula

For data in a matrix X and response in a vector y:

- Represent the formula using 'x1' as the first predictor (column) in X, 'x2' as the second predictor, etc.
- Represent the vector of parameters to optimize as 'b1', 'b2', etc.

- Write the formula as `'y ~ (mathematical expressions)'`.

For example, to represent the response to the reaction data:

```
modelfun = 'y ~ (b1*x2 - x3/b5)/(1 + b2*x1 + b3*x2 + b4*x3)';
```

For data in a table or dataset array, you can use formulas represented as the variable names from the table or dataset array. Put the response variable name at the left of the formula, followed by a `~`, followed by a character vector representing the response formula.

This example shows how to create a character vector to represent the response to the `reaction` data that is in a dataset array.

1. Load the `reaction` data.

   ```
   load reaction
   ```

2. Put the data into a dataset array, where each variable has a name given in `xn` or `yn`.

   ```
   ds = dataset({reactants,xn(1,:),xn(2,:),xn(3,:)},...
       {rate,yn});
   ```

3. Examine the first row of the dataset array.

   ```
   ds(1,:)

   ans =

       Hydrogen    n_Pentane    Isopentane    ReactionRate
       470         300          10            8.55
   ```

4. Write the `hougen` formula using names in the dataset array.

   ```
   modelfun = ['ReactionRate ~ (b1*n_Pentane - Isopentane/b5) /'...
   ' (1 + Hydrogen*b2 + n_Pentane*b3 + Isopentane*b4)']

   modelfun =
   ReactionRate ~ (b1*n_Pentane - Isopentane/b5) / ...
        (1 + Hydrogen*b2 + n_Pentane*b3 + Isopentane*b4)
   ```

Choose Initial Vector beta0

The initial vector for the fitting iterations, `beta0`, can greatly influence the quality of the resulting fitted model. `beta0` gives the dimensionality of the problem, meaning it needs the correct length. A good choice of `beta0` leads to a quick, reliable model, while a poor choice can lead to a long computation, or to an inadequate model.

It is difficult to give advice on choosing a good `beta0`. If you believe certain components of the vector should be positive or negative, set your `beta0` to have those characteristics. If you know the approximate value of other components, include them in `beta0`. However, if you don’t know good values, try a random vector, such as

```
beta0 = randn(nVars,1);
% or
beta0 = 10*rand(nVars,1);
```

Fit Nonlinear Model to Data

The syntax for fitting a nonlinear regression model using a table or dataset array `tbl` is

```
mdl = fitnlm(tbl,modelfun,beta0)
```

The syntax for fitting a nonlinear regression model using a numeric array `X` and numeric response vector `y` is

```
mdl = fitnlm(X,y,modelfun,beta0)
```

For information on representing the input parameters, see “Prepare Data” on page 13-2, “Represent the Nonlinear Model” on page 13-3, and “Choose Initial Vector beta0” on page 13-5.

`fitnlm` assumes that the response variable in a table or dataset array `tbl` is the last column. To change this, use the `ResponseVar` name-value pair to name the response column.

Examine Quality and Adjust the Fitted Nonlinear Model

There are diagnostic plots to help you examine the quality of a model. `plotDiagnostics(mdl)` gives a variety of plots, including leverage and Cook's distance plots. `plotResiduals(mdl)` gives the difference between the fitted model and the data.

There are also properties of `mdl` that relate to the model quality. `mdl.RMSE` gives the root mean square error between the data and the fitted model. `mdl.Residuals.Raw` gives the raw residuals. `mdl.Diagnostics` contains several fields, such as `Leverage` and `CooksDistance`, that can help you identify particularly interesting observations.

This example shows how to examine a fitted nonlinear model using diagnostic, residual, and slice plots.

Load the sample data.

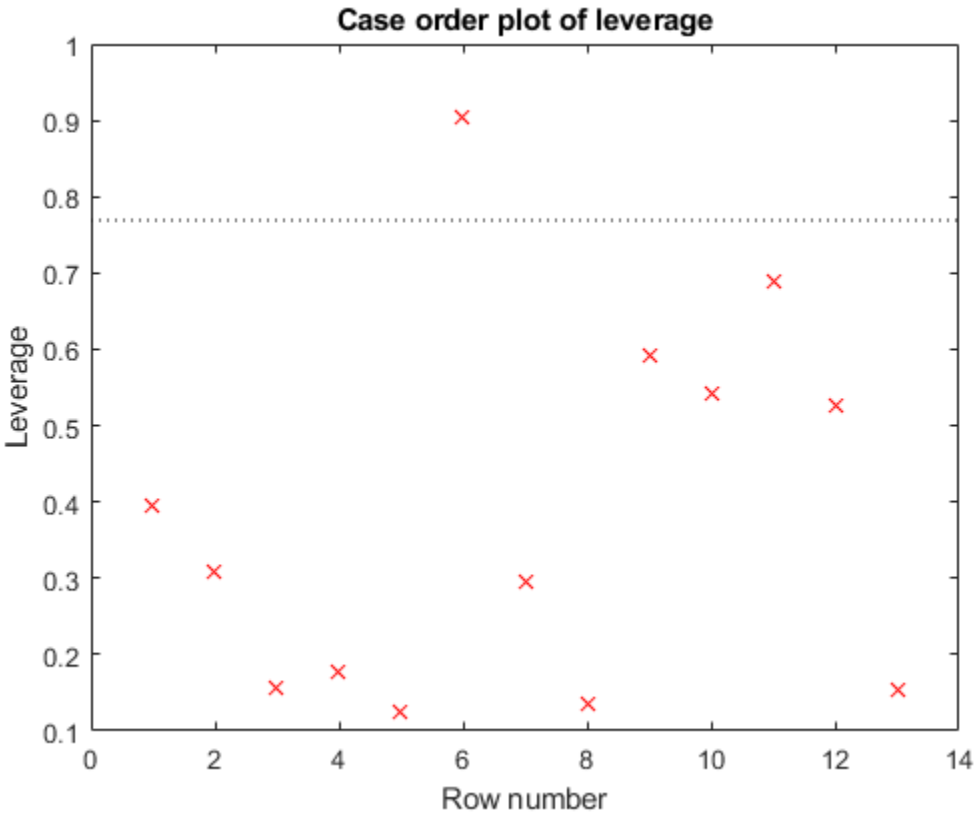
```
load reaction
```

Create a nonlinear model of rate as a function of `reactants` using the `hougen.m` function.

```
beta0 = ones(5,1);
mdl = fitnlm(reactants,...
    rate,@hougen,beta0);
```

Make a leverage plot of the data and model.

```
plotDiagnostics(mdl)
```



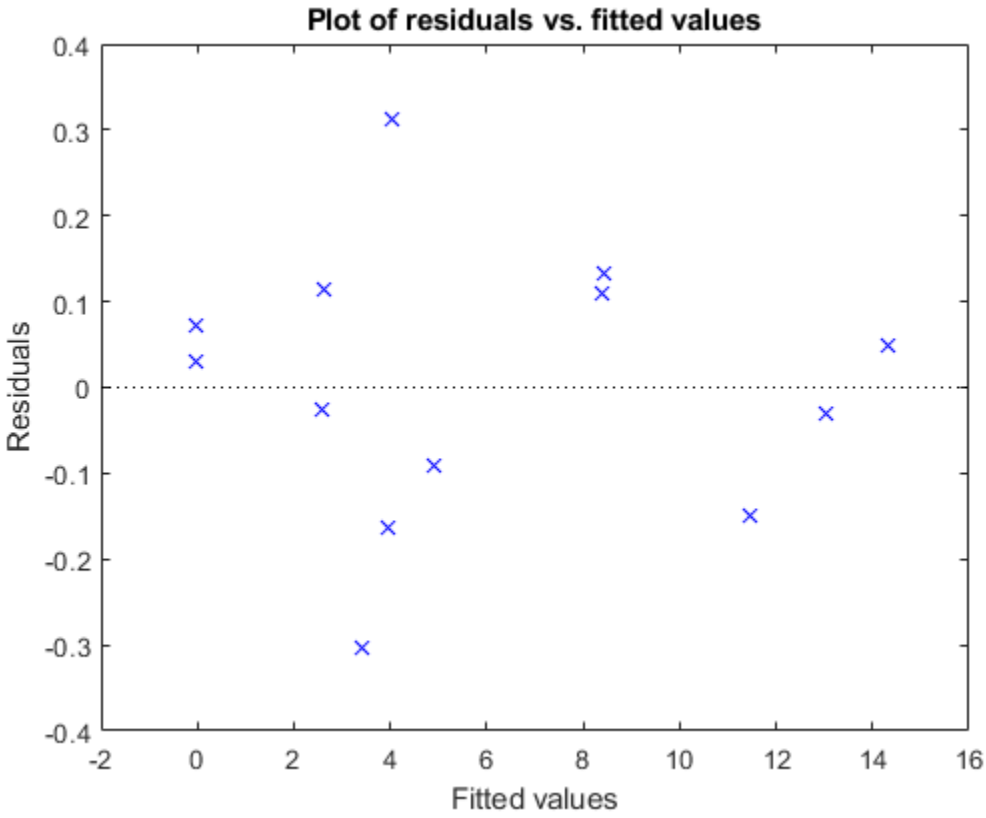
There is one point that has high leverage. Locate the point.

```
[~,maxl] = max(mdl.Diagnostics.Leverage)
```

```
maxl = 6
```

Examine a residuals plot.

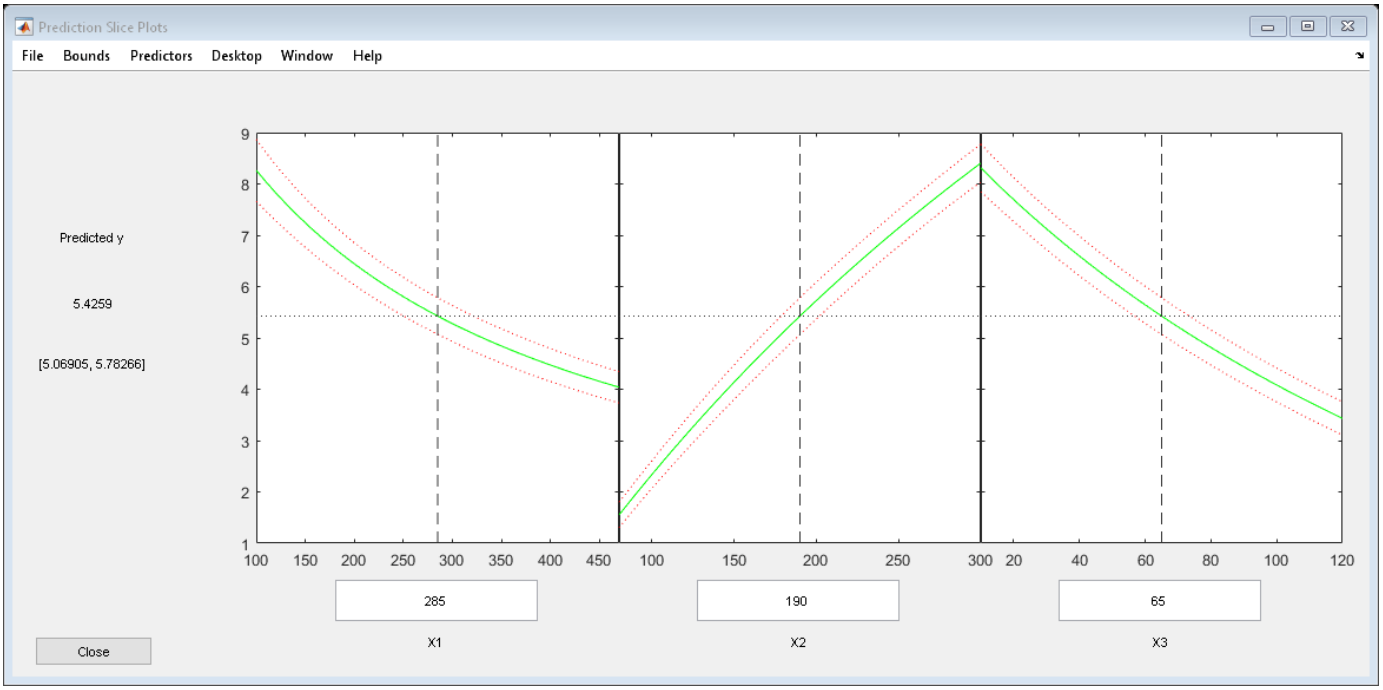
```
plotResiduals(mdl,'fitted')
```



Nothing stands out as an outlier.

Use a slice plot to show the effect of each predictor on the model.

```
plotSlice(mdl)
```



You can drag the vertical dashed blue lines to see the effect of a change in one predictor on the response. For example, drag the X2 line to the right, and notice that the slope of the X3 line changes.

Predict or Simulate Responses Using a Nonlinear Model

This example shows how to use the methods `predict`, `feval`, and `random` to predict and simulate responses to new data.

Randomly generate a sample from a Cauchy distribution.

```
rng('default')
X = rand(100,1);
X = tan(pi*X - pi/2);
```

Generate the response according to the model `y = b1*(pi /2 + atan((x - b2) / b3))` and add noise to the response.

```
modelfun = @(b,x) b(1) * ...
    (pi/2 + atan((x - b(2))/b(3)));
y = modelfun([12 5 10],X) + randn(100,1);
```

Fit a model starting from the arbitrary parameters b = [1,1,1].

```
beta0 = [1 1 1]; % An arbitrary guess
mdl = fitnlm(X,y,modelfun,beta0)
```

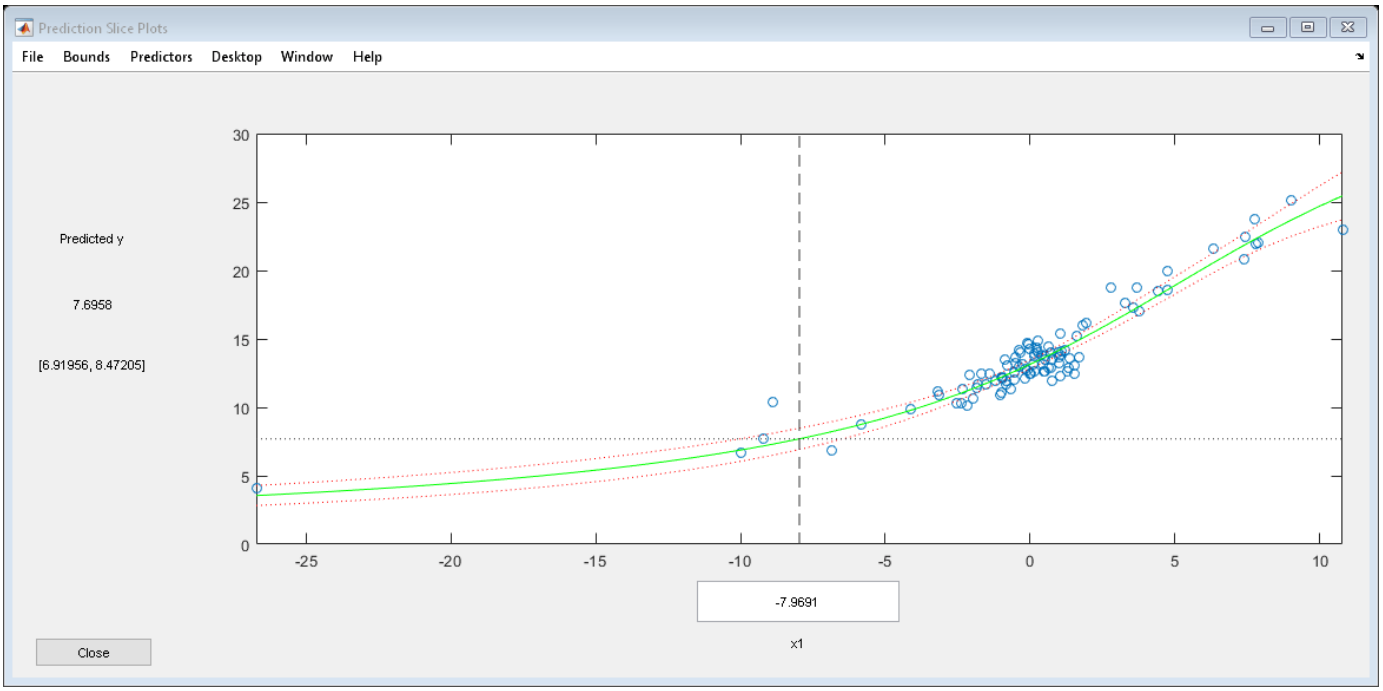
```
mdl =
Nonlinear regression model:
    y ~ b1*(pi/2 + atan((x - b2)/b3))

Estimated Coefficients:
```

```
          Estimate      SE       tStat       pValue
          ________    ________    _______    __________

    b1      12.082     0.80028    15.097    3.3151e-27
    b2      5.0603      1.0825    4.6747    9.5063e-06
    b3        9.64     0.46499    20.732    2.0382e-37


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.02
R-Squared: 0.92,  Adjusted R-Squared 0.918
F-statistic vs. zero model: 6.45e+03, p-value = 1.72e-111
```

The fitted values are within a few percent of the parameters [12,5,10].

Examine the fit.

```
plotSlice(mdl)
```



predict

The `predict` method predicts the mean responses and, if requested, gives confidence bounds. Find the predicted response values and predicted confidence intervals about the response at X values [-15;5;12].

```
Xnew = [-15;5;12];
[ynew,ynewci] = predict(mdl,Xnew)
```

```
ynew = 3×1

    5.4122
   18.9022
   26.5161
```

```
ynewci = 3×2

    4.8233    6.0010
   18.4555   19.3490
   25.0170   28.0151
```

The confidence intervals are reflected in the slice plot.

`feval`

The `feval` method predicts the mean responses. `feval` is often more convenient to use than predict when you construct a model from a dataset array.

Create the nonlinear model from a dataset array.

```
ds = dataset({X,'X'},{y,'y'});
mdl2 = fitnlm(ds,modelfun,beta0);
```

Find the predicted model responses (CDF) at X values [-15;5;12].

```
Xnew = [-15;5;12];
ynew = feval(mdl2,Xnew)
```

```
ynew = 3×1

    5.4122
   18.9022
   26.5161
```

`random`

The `random` method simulates new random response values, equal to the mean prediction plus a random disturbance with the same variance as the training data.

```
Xnew = [-15;5;12];
ysim = random(mdl,Xnew)
```

```
ysim = 3×1

    6.0505
   19.0893
   25.4647
```

Rerun the random method. The results change.

```
ysim = random(mdl,Xnew)
```

```
ysim = 3×1

    6.3813
   19.2157
   26.6541
```

Nonlinear Regression Workflow

This example shows how to do a typical nonlinear regression workflow: import data, fit a nonlinear regression, test its quality, modify it to improve the quality, and make predictions based on the model.

Step 1. Prepare the data.

Load the `reaction` data.

```
load reaction
```

Examine the data in the workspace. `reactants` is a matrix with 13 rows and 3 columns. Each row corresponds to one observation, and each column corresponds to one variable. The variable names are in `xn`:

```
xn
```

```
xn = 3x10 char array
    'Hydrogen  '
    'n-Pentane '
    'Isopentane'
```

Similarly, `rate` is a vector of 13 responses, with the variable name in `yn`:

```
yn
```

```
yn =
'Reaction Rate'
```

The `hougen.m` file contains a nonlinear model of reaction rate as a function of the three predictor variables. For a 5-D vector b and 3-D vector x,

$$hougen(b,x) = \frac{b(1)x(2) - x(3)/b(5)}{1 + b(2)x(1) + b(3)x(2) + b(4)x(3)}$$

As a start point for the solution, take `b` as a vector of ones.

```
beta0 = ones(5,1);
```

Step 2. Fit a nonlinear model to the data.

```
mdl = fitnlm(reactants,...
    rate,@hougen,beta0)
```

```
mdl =
Nonlinear regression model:
    y ~ hougen(b,X)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1      1.2526     0.86702    1.4447    0.18654
    b2    0.062776    0.043562    1.4411    0.18753
    b3    0.040048    0.030885    1.2967    0.23089
    b4     0.11242    0.075158    1.4957    0.17309
    b5      1.1914     0.83671    1.4239     0.1923
```

```
Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 0.193
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 3.91e+03, p-value = 2.54e-13
```

Step 3. Examine the quality of the model.

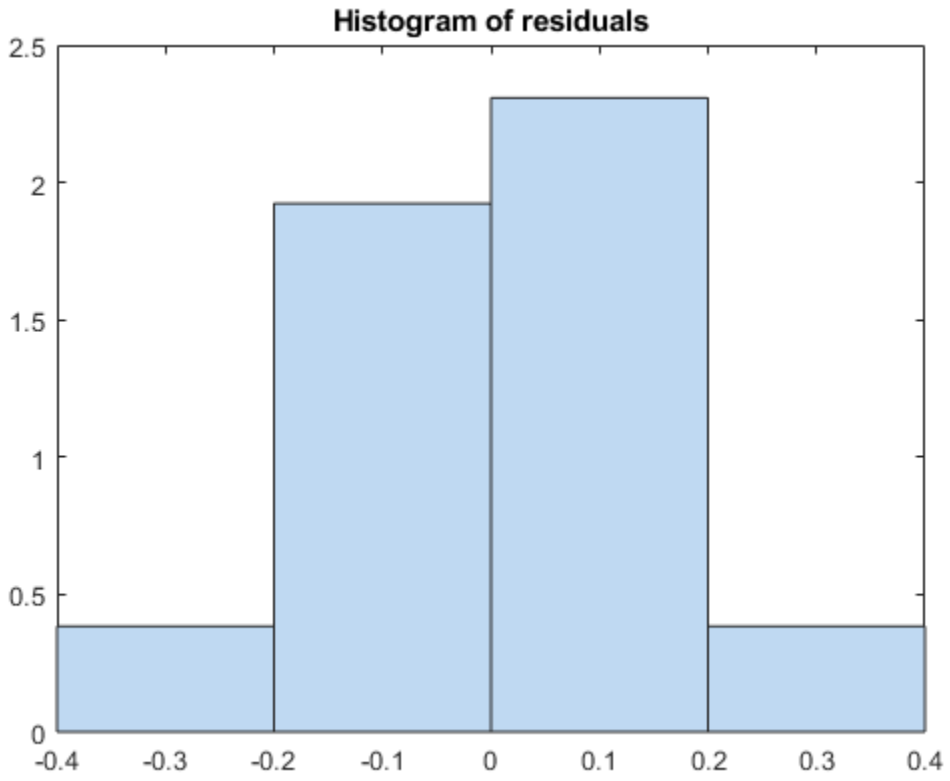
The root mean squared error is fairly low compared to the range of observed values.

```
[mdl.RMSE min(rate) max(rate)]
```

```
ans = 1×3

    0.1933    0.0200   14.3900
```

Examine a residuals plot.

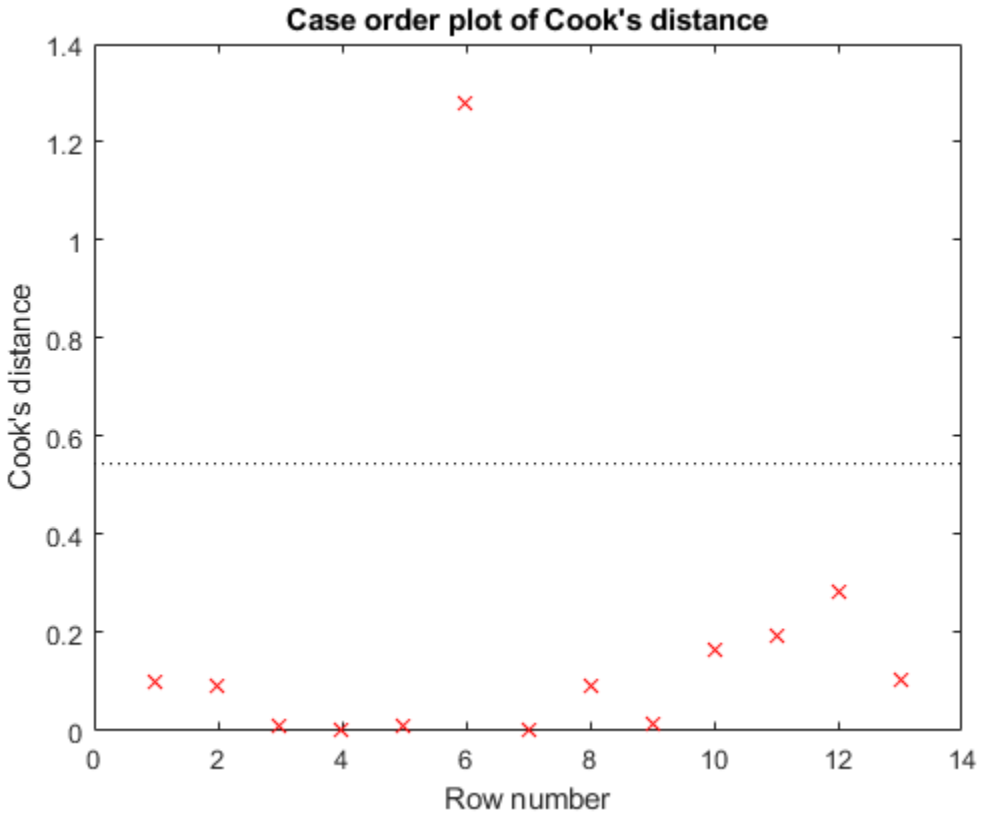
```
plotResiduals(mdl)
```



The model seems adequate for the data.

Examine a diagnostic plot to look for outliers.

```
plotDiagnostics(mdl,'cookd')
```



Observation 6 seems out of line.

Step 4. Remove the outlier.

Remove the outlier from the fit using the `Exclude` name-value pair.

```
mdl1 = fitnlm(reactants,...
    rate,@hougen,ones(5,1),'Exclude',6)
```

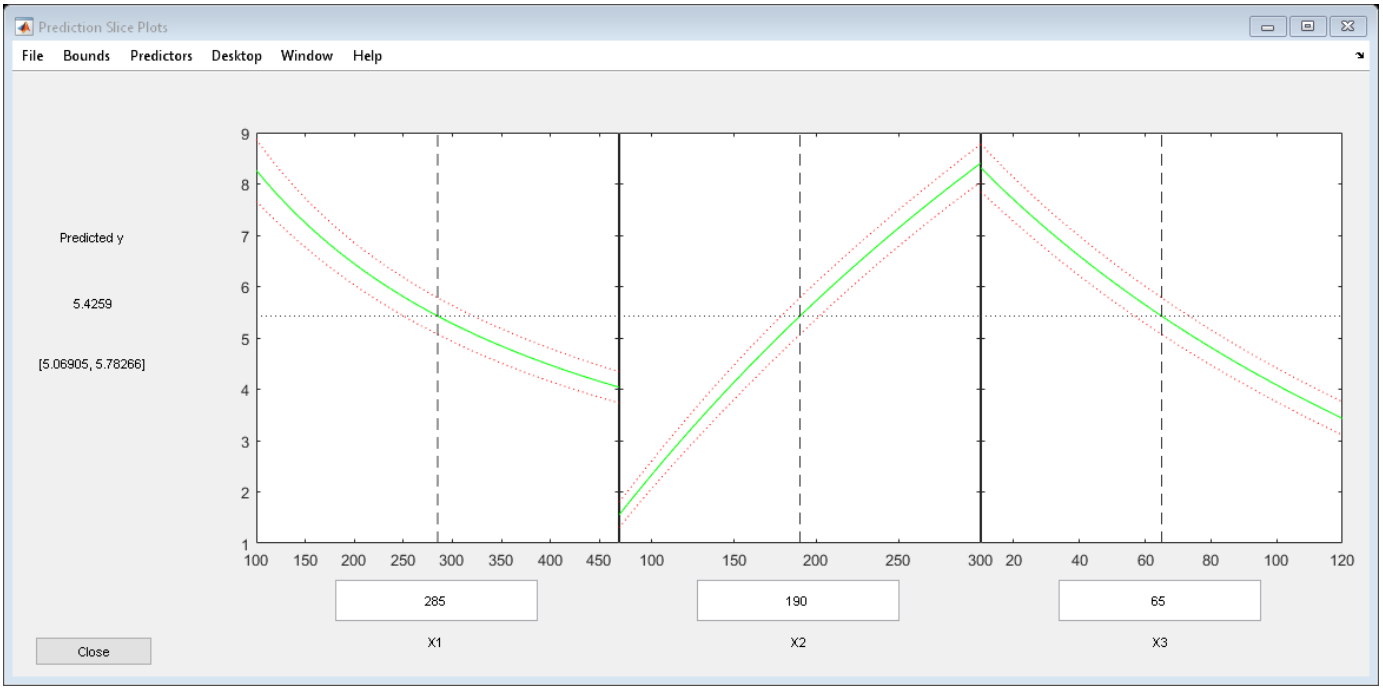
```
mdl1 =
Nonlinear regression model:
    y ~ hougen(b,X)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1       0.619      0.4552    1.3598    0.21605
    b2    0.030377    0.023061    1.3172    0.22924
    b3    0.018927     0.01574    1.2024    0.26828
    b4    0.053411    0.041084       1.3    0.23476
    b5      2.4125      1.7903    1.3475     0.2198


Number of observations: 12, Error degrees of freedom: 7
Root Mean Squared Error: 0.198
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 2.67e+03, p-value = 2.54e-11
```

The model coefficients changed quite a bit from those in `mdl`.

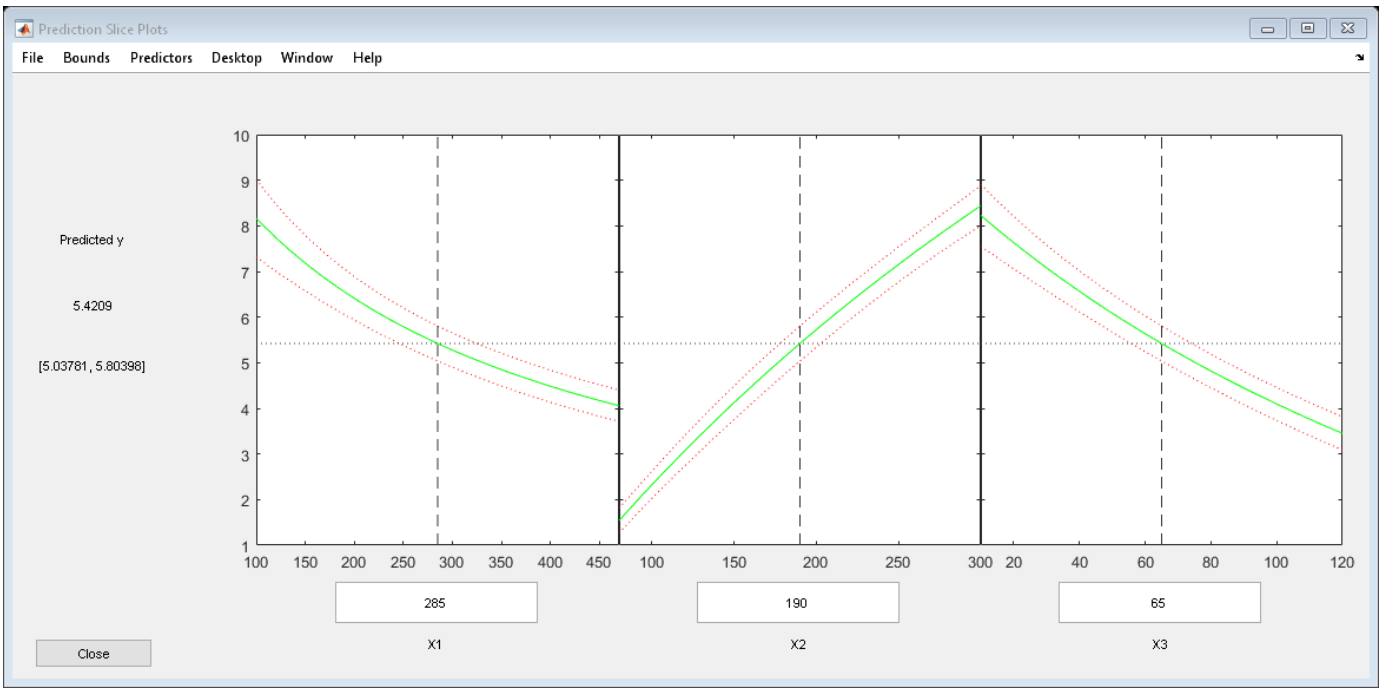
Step 5. Examine slice plots of both models.

To see the effect of each predictor on the response, make a slice plot using `plotSlice(mdl)`.

```
plotSlice(mdl)
```




```
plotSlice(mdl1)
```



The plots look very similar, with slightly wider confidence bounds for `mdl1`. This difference is understandable, since there is one less data point in the fit, representing over 7% fewer observations.

Step 6. Predict for new data.

Create some new data and predict the response from both models.

```
Xnew =  [200,200,200;100,200,100;500,50,5];
[ypred yci] = predict(mdl,Xnew)
```

```
ypred = 3×1

    1.8762
    6.2793
    1.6718


yci = 3×2

    1.6283    2.1242
    5.9789    6.5797
    1.5589    1.7846
```

```
[ypred1 yci1] = predict(mdl1,Xnew)
```

```
ypred1 = 3×1

    1.8984
    6.2555
    1.6594


yci1 = 3×2

    1.6260    2.1708
    5.9323    6.5787
    1.5345    1.7843
```

Even though the model coefficients are dissimilar, the predictions are nearly identical.

Mixed-Effects Models

In this section...

Introduction to Mixed-Effects Models

In statistics, an effect is anything that influences the value of a response variable at a particular setting of the predictor variables. Effects are translated into model parameters. In linear models, effects become coefficients, representing the proportional contributions of model terms. In nonlinear models, effects often have specific physical interpretations, and appear in more general nonlinear combinations.

Fixed effects represent population parameters, assumed to be the same each time data is collected. Estimating fixed effects is the traditional domain of regression modeling. Random effects, by comparison, are sample-dependent random variables. In modeling, random effects act like additional error terms, and their distributions and covariances must be specified.

For example, consider a model of the elimination of a drug from the bloodstream. The model uses time t as a predictor and the concentration of the drug C as the response. The nonlinear model term C_0e^{-rt} combines parameters C_0 and r, representing, respectively, an initial concentration and an elimination rate. If data is collected across multiple individuals, it is reasonable to assume that the elimination rate is a random variable r_i depending on individual i, varying around a population mean $\bar{r}$. The term C_0e^{-rt} becomes

$$C_0e^{-[\bar{r}+(r_i-\bar{r})]t} = C_0e^{-(\beta+b_i)t},$$

where $\beta = \bar{r}$ is a fixed effect and $b_i = r_i - \bar{r}$ is a random effect.

Random effects are useful when data falls into natural groups. In the drug elimination model, the groups are simply the individuals under study. More sophisticated models might group data by an individual's age, weight, diet, etc. Although the groups are not the focus of the study, adding random effects to a model extends the reliability of inferences beyond the specific sample of individuals.

Mixed-effects models account for both fixed and random effects. As with all regression models, their purpose is to describe a response variable as a function of the predictor variables. Mixed-effects models, however, recognize correlations within sample subgroups. In this way, they provide a compromise between ignoring data groups entirely and fitting each group with a separate model.

Mixed-Effects Model Hierarchy

Suppose data for a nonlinear regression model falls into one of m distinct groups $i = 1, \ldots, m$. To account for the groups in a model, write response j in group i as:

$$y_{ij} = f(\varphi, x_{ij}) + \varepsilon_{ij}$$

y_{ij} is the response, x_{ij} is a vector of predictors, φ is a vector of model parameters, and ε_{ij} is the measurement or process error. The index j ranges from 1 to n_i, where n_i is the number of observations in group i. The function f specifies the form of the model. Often, x_{ij} is simply an observation time t_{ij}. The errors are usually assumed to be independent and identically, normally distributed, with constant variance.

Estimates of the parameters in φ describe the population, assuming those estimates are the same for all groups. If, however, the estimates vary by group, the model becomes

$$y_{ij} = f(\varphi_i, x_{ij}) + \varepsilon_{ij}$$

In a mixed-effects model, φ_i may be a combination of a fixed and a random effect:

$$\varphi_i = \beta + b_i$$

The random effects b_i are usually described as multivariate normally distributed, with mean zero and covariance Ψ. Estimating the fixed effects β and the covariance of the random effects Ψ provides a description of the population that does not assume the parameters φ_i are the same across groups. Estimating the random effects b_i also gives a description of specific groups within the data.

Model parameters do not have to be identified with individual effects. In general, design matrices A and B are used to identify parameters with linear combinations of fixed and random effects:

$$\varphi_i = A\beta + Bb_i$$

If the design matrices differ among groups, the model becomes

$$\varphi_i = A_i\beta + B_ib_i$$

If the design matrices also differ among observations, the model becomes

$$\varphi_{ij} = A_{ij}\beta + B_{ij}b_i$$
$$y_{ij} = f(\varphi_{ij}, x_{ij}) + \varepsilon_{ij}$$

Some of the group-specific predictors in x_{ij} may not change with observation j. Calling those v_i, the model becomes

$$y_{ij} = f(\varphi_{ij}, x_{ij}, v_i) + \varepsilon_{ij}$$

Specifying Mixed-Effects Models

Suppose data for a nonlinear regression model falls into one of m distinct groups $i = 1, \ldots, m$. (Specifically, suppose that the groups are not nested.) To specify a general nonlinear mixed-effects model for this data:

1. Define group-specific model parameters φ_i as linear combinations of fixed effects β and random effects b_i.
2. Define response values y_i as a nonlinear function f of the parameters and group-specific predictor variables X_i.

The model is:

$$\varphi_i = A_i\beta + B_i b_i$$
$$y_i = f(\varphi_i, X_i) + \varepsilon_i$$
$$b_i \sim N(0, \Psi)$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

This formulation of the nonlinear mixed-effects model uses the following notation:

φ_i	A vector of group-specific model parameters
β	A vector of fixed effects, modeling population parameters
b_i	A vector of multivariate normally distributed group-specific random effects
A_i	A group-specific design matrix for combining fixed effects
B_i	A group-specific design matrix for combining random effects
X_i	A data matrix of group-specific predictor values
y_i	A data vector of group-specific response values
f	A general, real-valued function of φ_i and X_i
ε_i	A vector of group-specific errors, assumed to be independent, identically, normally distributed, and independent of b_i
Ψ	A covariance matrix for the random effects
σ^2	The error variance, assumed to be constant across observations

For example, consider a model of the elimination of a drug from the bloodstream. The model incorporates two overlapping phases:

- An initial phase p during which drug concentrations reach equilibrium with surrounding tissues
- A second phase q during which the drug is eliminated from the bloodstream

For data on multiple individuals i, the model is

$$y_{ij} = C_{pi}e^{-r_{pi}t_{ij}} + C_{qi}e^{-r_{qi}t_{ij}} + \varepsilon_{ij},$$

where y_{ij} is the observed concentration in individual i at time t_{ij}. The model allows for different sampling times and different numbers of observations for different individuals.

The elimination rates r_{pi} and r_{qi} must be positive to be physically meaningful. Enforce this by introducing the log rates $R_{pi} = \log(r_{pi})$ and $R_{qi} = \log(r_{qi})$ and reparameterizing the model:

$$y_{ij} = C_{pi}e^{-\exp(R_{pi})t_{ij}} + C_{qi}e^{-\exp(R_{qi})t_{ij}} + \varepsilon_{ij}$$

Choosing which parameters to model with random effects is an important consideration when building a mixed-effects model. One technique is to add random effects to all parameters, and use estimates of their variances to determine their significance in the model. An alternative is to fit the model separately to each group, without random effects, and look at the variation of the parameter estimates. If an estimate varies widely across groups, or if confidence intervals for each group have minimal overlap, the parameter is a good candidate for a random effect.

To introduce fixed effects β and random effects b_i for all model parameters, reexpress the model as follows:

$$\begin{aligned} y_{ij} &= [\bar{C}_p + (C_{pi} - \bar{C}_p)]e^{-\exp[\bar{R}_p + (R_{pi} - \bar{R}_p)]t_{ij}} + \\ &\quad [\bar{C}_q + (C_{qi} - \bar{C}_q)]e^{-\exp[\bar{R}_q + (R_{qi} - \bar{R}_q)]t_{ij}} + \varepsilon_{ij} \\ &= (\beta_1 + b_{1i})e^{-\exp(\beta_2 + b_{2i})t_{ij}} + \\ &\quad (\beta_3 + b_{3i})e^{-\exp(\beta_4 + b_{4i})t_{ij}} + \varepsilon_{ij} \end{aligned}$$

In the notation of the general model:

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_4 \end{pmatrix}, b_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{i4} \end{pmatrix}, y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{pmatrix}, X_i = \begin{pmatrix} t_{i1} \\ \vdots \\ t_{in_i} \end{pmatrix},$$

where n_i is the number of observations of individual i. In this case, the design matrices A_i and B_i are, at least initially, 4-by-4 identity matrices. Design matrices may be altered, as necessary, to introduce weighting of individual effects, or time dependency.

Fitting the model and estimating the covariance matrix Ψ often leads to further refinements. A relatively small estimate for the variance of a random effect suggests that it can be removed from the model. Likewise, relatively small estimates for covariances among certain random effects suggests that a full covariance matrix is unnecessary. Since random effects are unobserved, Ψ must be estimated indirectly. Specifying a diagonal or block-diagonal covariance pattern for Ψ can improve convergence and efficiency of the fitting algorithm.

Statistics and Machine Learning Toolbox functions `nlmefit` and `nlmefitsa` fit the general nonlinear mixed-effects model to data, estimating the fixed and random effects. The functions also estimate the covariance matrix Ψ for the random effects. Additional diagnostic outputs allow you to assess tradeoffs between the number of model parameters and the goodness of fit.

Specifying Covariate Models

If the model in “Specifying Mixed-Effects Models” on page 13-18 assumes a group-dependent covariate such as weight (w) the model becomes:

$$\begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & wi \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

Thus, the parameter φ_i for any individual in the `i`th group is:

$$\begin{pmatrix} \varphi_{1_i} \\ \varphi_{2_i} \\ \varphi_{3_i} \end{pmatrix} = \begin{pmatrix} \beta_1 + \beta_4 * w_i \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} b_{1_i} \\ b_{2_i} \\ b_{3_i} \end{pmatrix}$$

To specify a covariate model, use the `'FEGroupDesign'` option.

`'FEGroupDesign'` is a `p-by-q-by-m` array specifying a different `p-by-q` fixed-effects design matrix for each of the `m` groups. Using the previous example, the array resembles the following:

1. Create the array.

```
% Number of parameters in the model (Phi)
num_params = 3;
% Number of covariates
num_cov = 1;
% Assuming number of groups in the data set is 7
num_groups = 7;
% Array of covariate values
covariates = [75; 52; 66; 55; 70; 58; 62 ];
A = repmat(eye(num_params, num_params+num_cov),...
[1,1,num_groups]);
A(1,num_params+1,1:num_groups) = covariates(:,1)
```

2. Create a `struct` with the specified design matrix.

```
options.FEGroupDesign = A;
```

3. Specify the arguments for `nlmefit` (or `nlmefitsa`) as shown in “Mixed-Effects Models Using nlmefit and nlmefitsa” on page 13-33.

Choosing nlmefit or nlmefitsa

Statistics and Machine Learning Toolbox provides two functions, `nlmefit` and `nlmefitsa` for fitting nonlinear mixed-effects models. Each function provides different capabilities, which may help you decide which to use.

Approximation Methods

`nlmefit` provides the following four approximation methods for fitting nonlinear mixed-effects models:

- `'LME'` — Use the likelihood for the linear mixed-effects model at the current conditional estimates of `beta` and `B`. This is the default.
- `'RELME'` — Use the restricted likelihood for the linear mixed-effects model at the current conditional estimates of `beta` and `B`.

- `'FO'` — First-order Laplacian approximation without random effects.
- `'FOCE'` — First-order Laplacian approximation at the conditional estimates of `B`.

`nlmefitsa` provides an additional approximation method, Stochastic Approximation Expectation-Maximization (SAEM) [25] with three steps :

1. Simulation: Generate simulated values of the random effects b from the posterior density $p(b|\Sigma)$ given the current parameter estimates.
2. Stochastic approximation: Update the expected value of the log likelihood function by taking its value from the previous step, and moving part way toward the average value of the log likelihood calculated from the simulated random effects.
3. Maximization step: Choose new parameter estimates to maximize the log likelihood function given the simulated values of the random effects.

Both `nlmefit` and `nlmefitsa` attempt to find parameter estimates to maximize a likelihood function, which is difficult to compute. `nlmefit` deals with the problem by approximating the likelihood function in various ways, and maximizing the approximate function. It uses traditional optimization techniques that depend on things like convergence criteria and iteration limits.

`nlmefitsa`, on the other hand, simulates random values of the parameters in such a way that in the long run they converge to the values that maximize the exact likelihood function. The results are random, and traditional convergence tests don't apply. Therefore `nlmefitsa` provides options to plot the results as the simulation progresses, and to restart the simulation multiple times. You can use these features to judge whether the results have converged to the accuracy you desire.

Parameters Specific to nlmefitsa

The following parameters are specific to `nlmefitsa`. Most control the stochastic algorithm.

- `Cov0` — Initial value for the covariance matrix `PSI`. Must be an r-by-r positive definite matrix. If empty, the default value depends on the values of `BETA0`.
- `ComputeStdErrors` — `true` to compute standard errors for the coefficient estimates and store them in the output `STATS` structure, or `false` (default) to omit this computation.
- `LogLikMethod` — Specifies the method for approximating the log likelihood.
- `NBurnIn` — Number of initial burn-in iterations during which the parameter estimates are not recomputed. Default is 5.
- `NIterations` — Controls how many iterations are performed for each of three phases of the algorithm.
- `NMCMCIterations` — Number of Markov Chain Monte Carlo (MCMC) iterations.

Model and Data Requirements

There are some differences in the capabilities of `nlmefit` and `nlmefitsa`. Therefore some data and models are usable with either function, but some may require you to choose just one of them.

- **Error models** — `nlmefitsa` supports a variety of error models. For example, the standard deviation of the response can be constant, proportional to the function value, or a combination of the two. `nlmefit` fits models under the assumption that the standard deviation of the response is constant. One of the error models, `'exponential'`, specifies that the log of the response has a constant standard deviation. You can fit such models using `nlmefit` by providing the log response as input, and by rewriting the model function to produce the log of the nonlinear function value.

- **Random effects** — Both functions fit data to a nonlinear function with parameters, and the parameters may be simple scalar values or linear functions of covariates. `nlmefit` allows any coefficients of the linear functions to have both fixed and random effects. `nlmefitsa` supports random effects only for the constant (intercept) coefficient of the linear functions, but not for slope coefficients. So in the example in “Specifying Covariate Models” on page 13-20, `nlmefitsa` can treat only the first three beta values as random effects.
- **Model form** — `nlmefit` supports a very general model specification, with few restrictions on the design matrices that relate the fixed coefficients and the random effects to the model parameters. `nlmefitsa` is more restrictive:
  - The fixed effect design must be constant in every group (for every individual), so an observation-dependent design is not supported.
  - The random effect design must be constant for the entire data set, so neither an observation-dependent design nor a group-dependent design is supported.
  - As mentioned under **Random Effects**, the random effect design must not specify random effects for slope coefficients. This implies that the design must consist of zeros and ones.
  - The random effect design must not use the same random effect for multiple coefficients, and cannot use more than one random effect for any single coefficient.
  - The fixed effect design must not use the same coefficient for multiple parameters. This implies that it can have at most one nonzero value in each column.

  If you want to use `nlmefitsa` for data in which the covariate effects are random, include the covariates directly in the nonlinear model expression. Don't include the covariates in the fixed or random effect design matrices.
- **Convergence** — As described in the **Model form**, `nlmefit` and `nlmefitsa` have different approaches to measuring convergence. `nlmefit` uses traditional optimization measures, and `nlmefitsa` provides diagnostics to help you judge the convergence of a random simulation.

In practice, `nlmefitsa` tends to be more robust, and less likely to fail on difficult problems. However, `nlmefit` may converge faster on problems where it converges at all. Some problems may benefit from a combined strategy, for example by running `nlmefitsa` for a while to get reasonable parameter estimates, and using those as a starting point for additional iterations using `nlmefit`.

Using Output Functions with Mixed-Effects Models

The `Outputfcn` field of the `options` structure specifies one or more functions that the solver calls after each iteration. Typically, you might use an output function to plot points at each iteration or to display optimization quantities from the algorithm. To set up an output function:

1. Write the output function as a MATLAB file function or local function.
2. Use `statset` to set the value of `Outputfcn` to be a function handle, that is, the name of the function preceded by the `@` sign. For example, if the output function is `outfun.m`, the command

   ```
   options = statset('OutputFcn', @outfun);
   ```

   specifies `OutputFcn` to be the handle to `outfun`. To specify multiple output functions, use the syntax:

   ```
   options = statset('OutputFcn',{@outfun, @outfun2});
   ```
3. Call the optimization function with `options` as an input argument.

For an example of an output function, see “Sample Output Function” on page 13-27.

Structure of the Output Function

The function definition line of the output function has the following form:

```
stop = outfun(beta,status,state)
```

where

- *beta* is the current fixed effects.
- *status* is a structure containing data from the current iteration. “Fields in status” on page 13-24 describes the structure in detail.
- *state* is the current state of the algorithm. “States of the Algorithm” on page 13-25 lists the possible values.
- *stop* is a flag that is `true` or `false` depending on whether the optimization routine should quit or continue. See “Stop Flag” on page 13-26 for more information.

The solver passes the values of the input arguments to `outfun` at each iteration.

Fields in status

The following table lists the fields of the `status` structure:

Field	Description
`procedure`	• `'ALT'` — alternating algorithm for the optimization of the linear mixed effects or restricted linear mixed effects approximations • `'LAP'` — optimization of the Laplacian approximation for first order or first order conditional estimation
`iteration`	An integer starting from 0.

Field	Description

`inner`	A structure describing the status of the inner iterations within the `ALT` and `LAP` procedures, with the fields:
	• `procedure` — When `procedure` is `'ALT'`:
	  • `'PNLS'` (penalized nonlinear least squares)
	  • `'LME'` (linear mixed-effects estimation)
	  • `'none'`
	  When `procedure` is `'LAP'`,
	  • `'PNLS'` (penalized nonlinear least squares)
	  • `'PLM'` (profiled likelihood maximization)
	  • `'none'`
	• `state` — one of the following:
	  • `'init'`
	  • `'iter'`
	  • `'done'`
	  • `'none'`
	• `iteration` — an integer starting from 0, or `NaN`. For `nlmefitsa` with burn-in iterations, the output function is called after each of those iterations with a negative value for `STATUS.iteration`.

`fval`	The current log likelihood
`Psi`	The current random-effects covariance matrix
`theta`	The current parameterization of `Psi`
`mse`	The current error variance

States of the Algorithm

The following table lists the possible values for `state`:

state	Description
`'init'`	The algorithm is in the initial state before the first iteration.
`'iter'`	The algorithm is at the end of an iteration.
`'done'`	The algorithm is in the final state after the last iteration.

The following code illustrates how the output function might use the value of `state` to decide which tasks to perform at the current iteration:

```
switch state
    case 'iter'
          % Make updates to plot or guis as needed
    case 'init'
          % Setup for plots or guis
    case 'done'
          % Cleanup of plots, guis, or final plot
```

```
otherwise
end
```

Stop Flag

The output argument `stop` is a flag that is `true` or `false`. The flag tells the solver whether it should quit or continue. The following examples show typical ways to use the `stop` flag.

Stopping an Optimization Based on Intermediate Results

The output function can stop the estimation at any iteration based on the values of arguments passed into it. For example, the following code sets stop to true based on the value of the log likelihood stored in the `'fval'`field of the status structure:

```
stop = outfun(beta,status,state)
stop = false;
% Check if loglikelihood is more than 132.
if status.fval > -132
    stop = true;
end
```

Stopping an Iteration Based on GUI Input

If you design a GUI to perform `nlmefit` iterations, you can make the output function stop when a user clicks a **Stop** button on the GUI. For example, the following code implements a dialog to cancel calculations:

```
function retval = stop_outfcn(beta,str,status)
persistent h stop;
if isequal(str.inner.state,'none')
    switch(status)
        case 'init'
            % Initialize dialog
            stop = false;
            h = msgbox('Press STOP to cancel calculations.',...
                'NLMEFIT: Iteration 0 ');
            button = findobj(h,'type','uicontrol');
            set(button,'String','STOP','Callback',@stopper)
            pos = get(h,'Position');
            pos(3) = 1.1 * pos(3);
            set(h,'Position',pos)
            drawnow
        case 'iter'
            % Display iteration number in the dialog title
            set(h,'Name',sprintf('NLMEFIT: Iteration %d',...
                str.iteration))
            drawnow;
        case 'done'
            % Delete dialog
            delete(h);
    end
end
if stop
    % Stop if the dialog button has been pressed
    delete(h)
end
retval = stop;
```

```
    function stopper(varargin)
        % Set flag to stop when button is pressed
        stop = true;
        disp('Calculation stopped.')
    end
end
```

Sample Output Function

`nmlefitoutputfcn` is the sample Statistics and Machine Learning Toolbox output function for `nlmefit` and `nlmefitsa`. It initializes or updates a plot with the fixed-effects (`BETA`) and variance of the random effects (`diag(STATUS.Psi)`). For `nlmefit`, the plot also includes the log-likelihood (`STATUS.fval`).

`nlmefitoutputfcn` is the default output function for `nlmefitsa`. To use it with `nlmefit`, specify a function handle for it in the options structure:

```
opt = statset('OutputFcn', @nlmefitoutputfcn, …)
beta = nlmefit(…, 'Options', opt, …)
```

To prevent `nlmefitsa` from using of this function, specify an empty value for the output function:

```
opt = statset('OutputFcn', [], …)
beta = nlmefitsa(…, 'Options', opt, …)
```

`nlmefitoutputfcn` stops `nlmefit` or `nlmefitsa` if you close the figure that it produces.

Examining Residuals for Model Verification

You can examine the `stats` structure, which is returned by both `nlmefit` and `nlmefitsa`, to determine the quality of your model. The `stats` structure contains fields with conditional weighted residuals (`cwres` field) and individual weighted residuals (`iwres` field). Since the model assumes that residuals are normally distributed, you can examine the residuals to see how well this assumption holds.

This example generates synthetic data using normal distributions. It shows how the fit statistics look:

- Good when testing against the same type of model as generates the data
- Poor when tested against incorrect data models

1 Initialize a 2-D model with 100 individuals:

```
nGroups = 100; % 100 Individuals
nlmefun = @(PHI,t)(PHI(:,1)*5 + PHI(:,2)^2.*t); % Regression fcn
REParamsSelect = [1  2]; % Both Parameters have random effect
errorParam = .03;
beta0 = [ 1.5  5]; % Parameter means
psi = [ 0.35  0; ...  % Covariance Matrix
       0   0.51 ];
time =[0.25;0.5;0.75;1;1.25;2;3;4;5;6];
nParameters = 2;
rng(0,'twister') % for reproducibility
```

2 Generate the data for fitting with a proportional error model:

```
b_i = mvnrnd(zeros(1, numel(REParamsSelect)), psi, nGroups);
individualParameters = zeros(nGroups,nParameters);
individualParameters(:, REParamsSelect) = ...
     bsxfun(@plus,beta0(REParamsSelect), b_i);

groups = repmat(1:nGroups,numel(time),1);
groups = vertcat(groups(:));

y = zeros(numel(time)*nGroups,1);
x = zeros(numel(time)*nGroups,1);
for i = 1:nGroups
    idx = groups == i;
    f = nlmefun(individualParameters(i,:), time);
    % Make a proportional error model for y:
    y(idx) = f + errorParam*f.*randn(numel(f),1);
    x(idx) = time;
end

P = [ 1 0 ; 0 1 ];
```

3 Fit the data using the same regression function and error model as the model generator:

```
[~,~,stats] = nlmefit(x,y,groups, ...
    [],nlmefun,[1 1],'REParamsSelect',REParamsSelect,...
    'ErrorModel','Proportional','CovPattern',P);
```

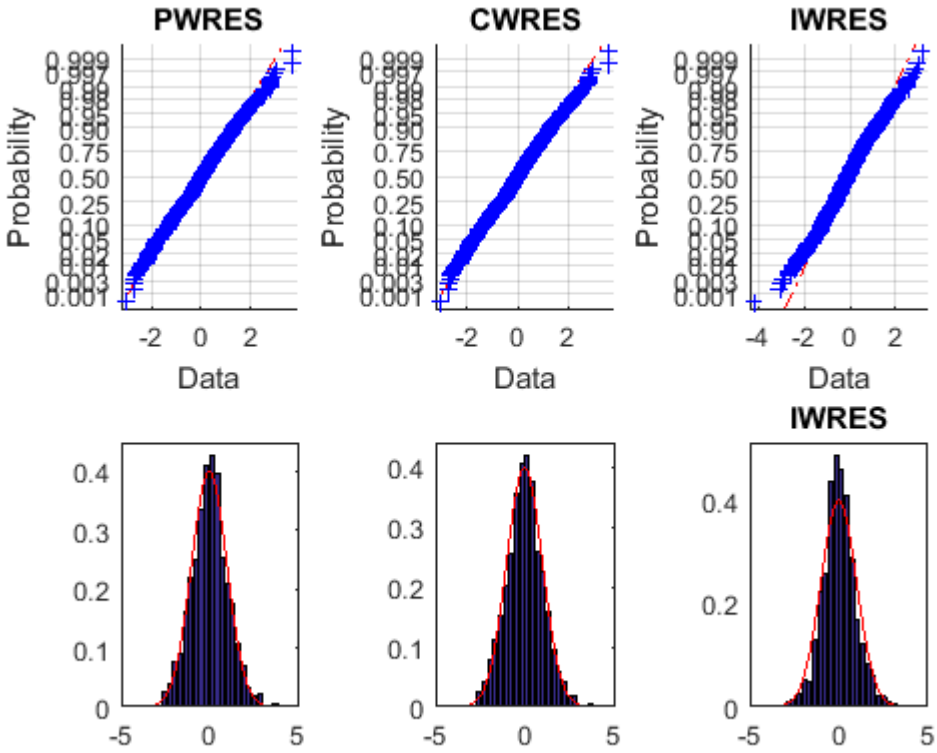
4 Create a plotting routine by copying the following function definition, and creating a file `plotResiduals.m` on your MATLAB path:

```
function plotResiduals(stats)
pwres = stats.pwres;
iwres = stats.iwres;
```

```
cwres = stats.cwres;
figure
subplot(2,3,1);
normplot(pwres); title('PWRES')
subplot(2,3,4);
createhistplot(pwres);

subplot(2,3,2);
normplot(cwres); title('CWRES')
subplot(2,3,5);
createhistplot(cwres);

subplot(2,3,3);
normplot(iwres); title('IWRES')
subplot(2,3,6);
createhistplot(iwres); title('IWRES')

function createhistplot(pwres)
h = histogram(pwres);

% x is the probability/height for each bin
x = h.Values/sum(h.Values*h.BinWidth)

% n is the center of each bin
n = h.BinEdges + (0.5*h.BinWidth)
n(end) = [];

bar(n,x);
ylim([0 max(x)*1.05]);
hold on;
x2 = -4:0.1:4;
f2 = normpdf(x2,0,1);
plot(x2,f2,'r');
end

end
```

5 Plot the residuals using the `plotResiduals` function:

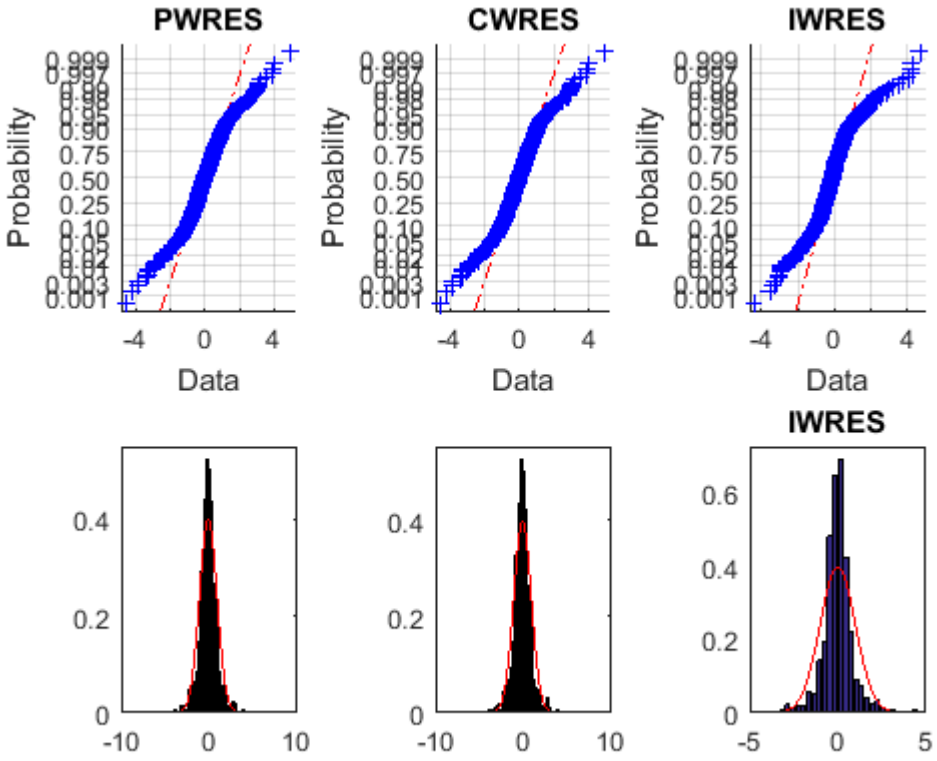
```
plotResiduals(stats);
```



The upper probability plots look straight, meaning the residuals are normally distributed. The bottom histogram plots match the superimposed normal density plot. So you can conclude that the error model matches the data.

6 For comparison, fit the data using a constant error model, instead of the proportional model that created the data:

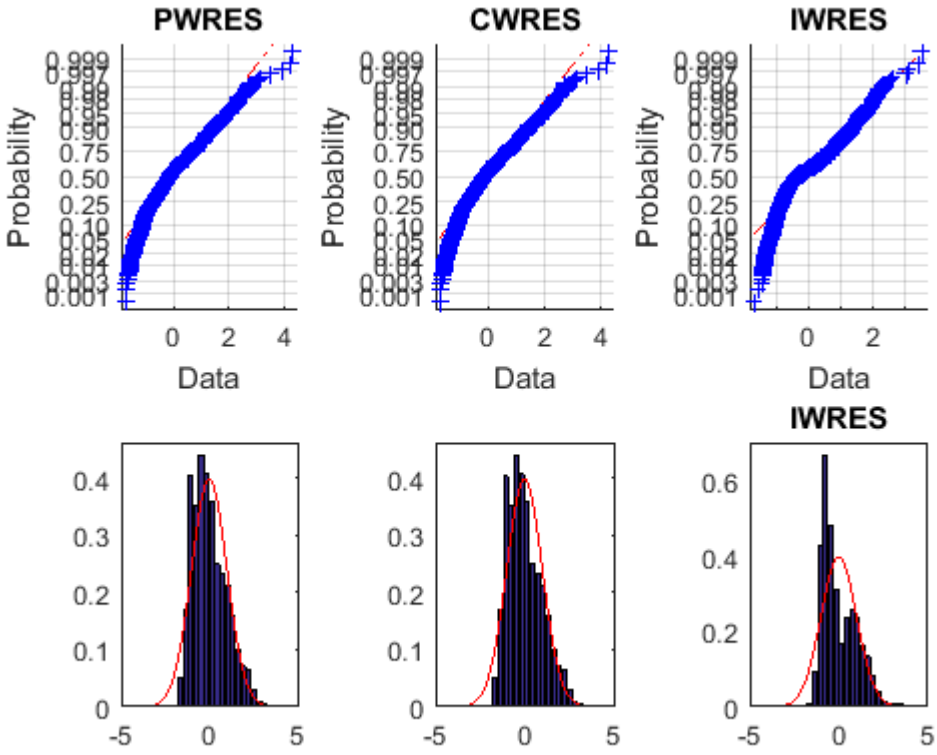
```
[~,~,stats] = nlmefit(x,y,groups, ...
    [],nlmefun,[0 0],'REParamsSelect',REParamsSelect,...
    'ErrorModel','Constant','CovPattern',P);
plotResiduals(stats);
```



The upper probability plots are not straight, indicating the residuals are not normally distributed. The bottom histogram plots are fairly close to the superimposed normal density plots.

7 For another comparison, fit the data to a different structural model than the one that created the data:

```
nlmefun2 = @(PHI,t)(PHI(:,1)*5 + PHI(:,2).*t.^4);
[~,~,stats] = nlmefit(x,y,groups, ...
    [],nlmefun2,[0 0],'REParamsSelect',REParamsSelect,...
    'ErrorModel','constant', 'CovPattern',P);
plotResiduals(stats);
```



The upper probability plots are not straight. Also, the histogram plots are quite skewed compared to the superimposed normal density plots. These residuals are not normally distributed, and do not match the model.

Mixed-Effects Models Using nlmefit and nlmefitsa

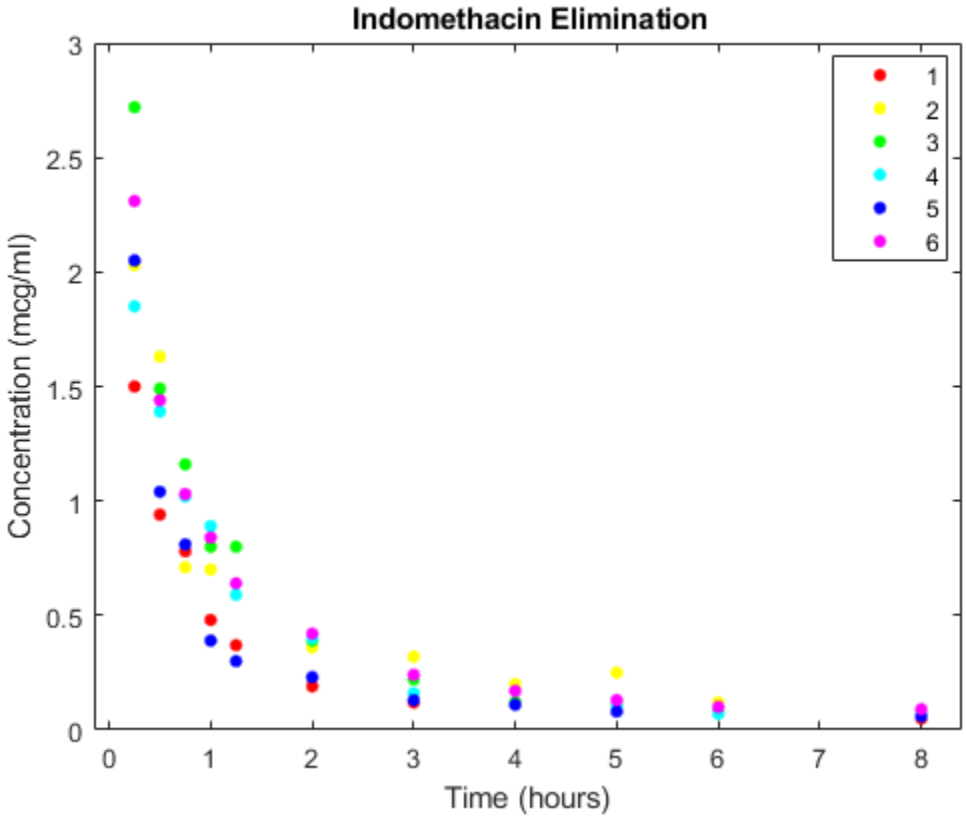
Load the sample data.

```
load indomethacin
```

The data in `indomethacin.mat` records concentrations of the drug indomethacin in the bloodstream of six subjects over eight hours.

Plot the scatter plot of indomethacin in the bloodstream grouped by subject.

```
gscatter(time,concentration,subject)
xlabel('Time (hours)')
ylabel('Concentration (mcg/ml)')
title('{\bf Indomethacin Elimination}')
hold on
```



`Specifying Mixed-Effects Models` page discusses a useful model for this type of data.

Construct the model via an anonymous function.

```
model = @(phi,t)(phi(1)*exp(-exp(phi(2))*t) + ...
                 phi(3)*exp(-exp(phi(4))*t));
```

Use the `nlinfit` function to fit the model to all of the data, ignoring subject-specific effects.

```
phi0 = [1 2 1 1];
[phi,res] = nlinfit(time,concentration,model,phi0);
```

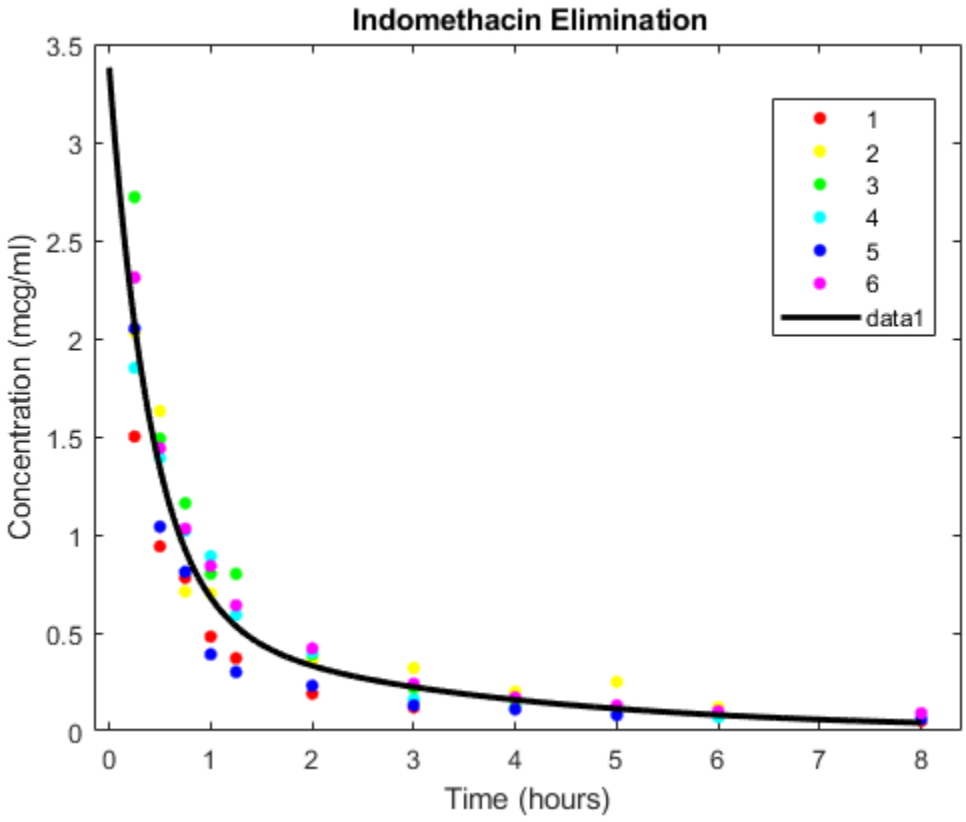
Compute the mean squared error.

```
numObs = length(time);
numParams = 4;
df = numObs-numParams;
mse = (res'*res)/df
```

```
mse =

    0.0304
```

Super impose the model on the scatter plot of data.

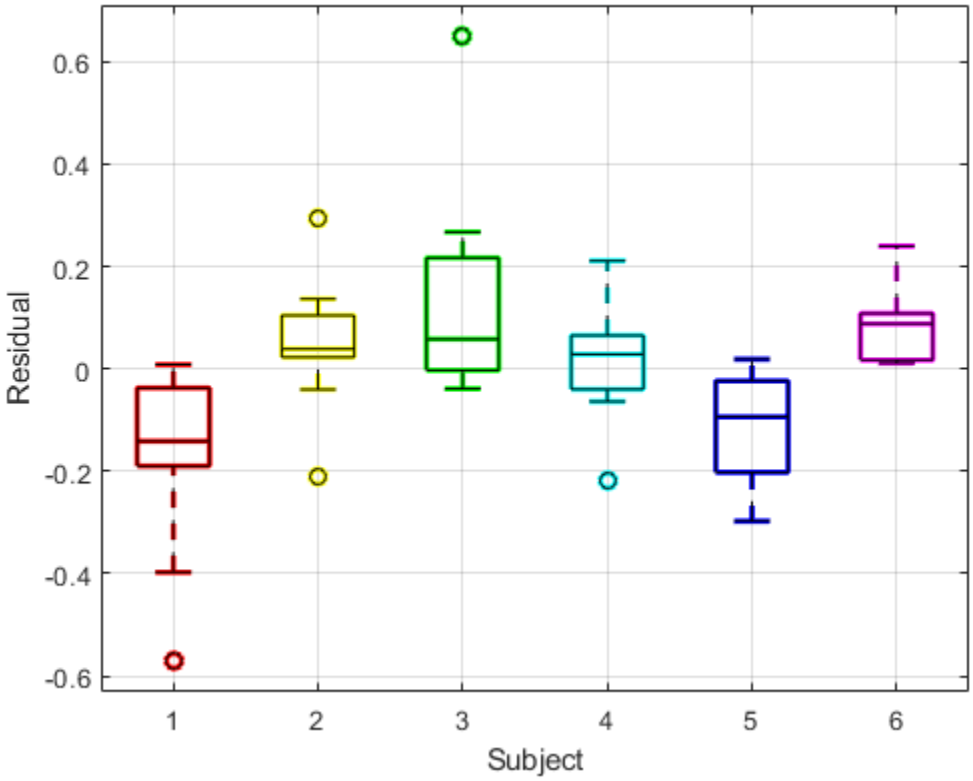
```
tplot = 0:0.01:8;
plot(tplot,model(phi,tplot),'k','LineWidth',2)
hold off
```



Draw the box-plot of residuals by subject.

```
colors = 'rygcbm';
h = boxplot(res,subject,'colors',colors,'symbol','o');
set(h(~isnan(h)),'LineWidth',2)
hold on
boxplot(res,subject,'colors','k','symbol','ko')
grid on
xlabel('Subject')
```

```
ylabel('Residual')
hold off
```



The box plot of residuals by subject shows that the boxes are mostly above or below zero, indicating that the model has failed to account for subject-specific effects.

To account for subject-specific effects, fit the model separately to the data for each subject.

```
phi0 = [1 2 1 1];
PHI = zeros(4,6);
RES = zeros(11,6);
for I = 1:6
    tI = time(subject == I);
    cI = concentration(subject == I);
    [PHI(:,I),RES(:,I)] = nlinfit(tI,cI,model,phi0);
end
PHI
```

```
PHI =

    2.0293    2.8277    5.4683    2.1981    3.5661    3.0023
    0.5794    0.8013    1.7498    0.2423    1.0408    1.0882
    0.1915    0.4989    1.6757    0.2545    0.2915    0.9685
   -1.7878   -1.6354   -0.4122   -1.6026   -1.5069   -0.8731
```

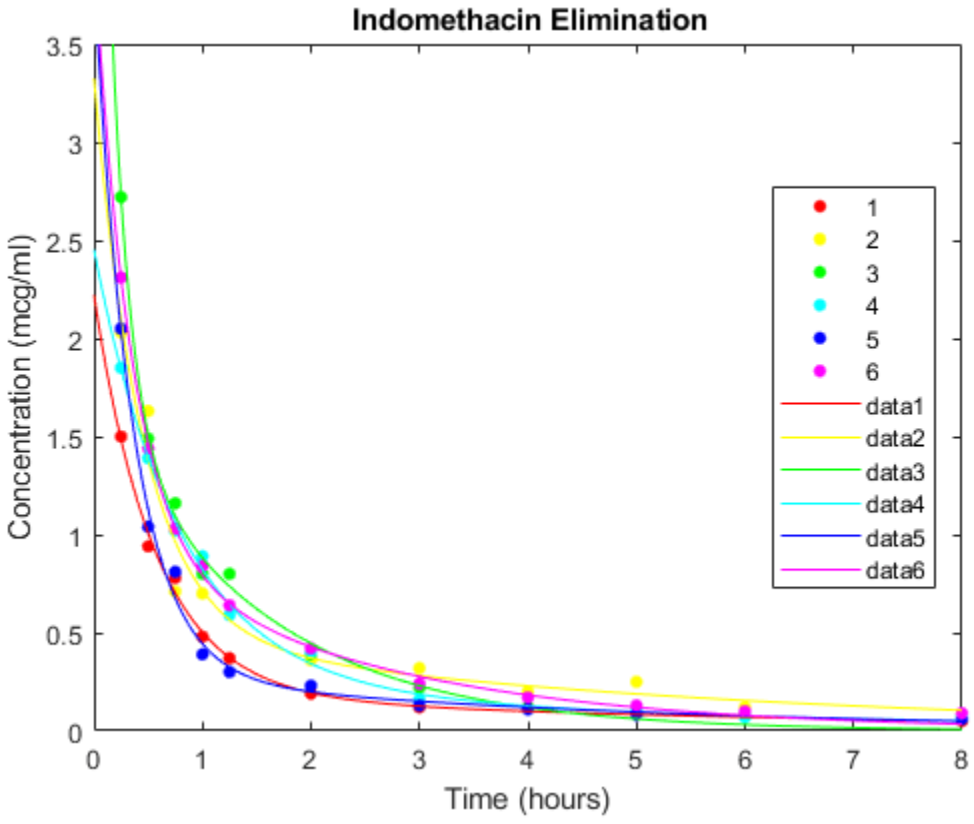
Compute the mean squared error.

```
numParams = 24;
df = numObs-numParams;
mse = (RES(:)'*RES(:))/df
```

```
mse =

    0.0057
```

Plot the scatter plot of the data and superimpose the model for each subject.

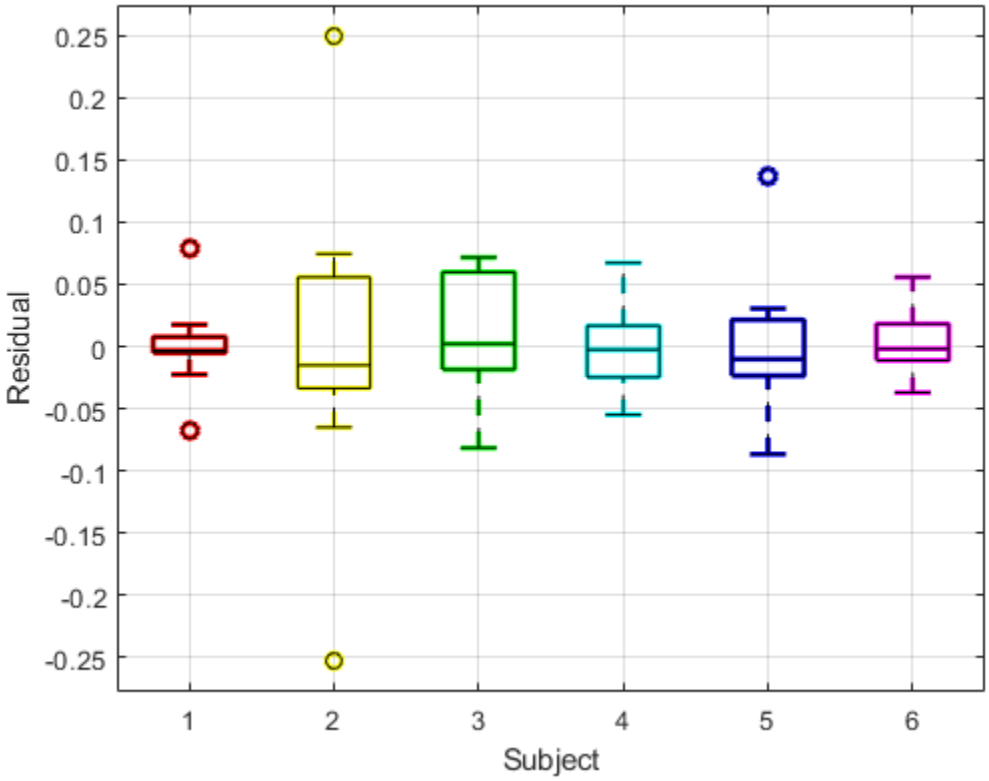
```
gscatter(time,concentration,subject)
xlabel('Time (hours)')
ylabel('Concentration (mcg/ml)')
title('{\bf Indomethacin Elimination}')
hold on
for I = 1:6
    plot(tplot,model(PHI(:,I),tplot),'Color',colors(I))
end
axis([0 8 0 3.5])
hold off
```



`PHI` gives estimates of the four model parameters for each of the six subjects. The estimates vary considerably, but taken as a 24-parameter model of the data, the mean-squared error of 0.0057 is a significant reduction from 0.0304 in the original four-parameter model.

Draw the box plot of residuals by subject.

```
h = boxplot(RES,'colors',colors,'symbol','o');
set(h(~isnan(h)),'LineWidth',2)
hold on
boxplot(RES,'colors','k','symbol','ko')
grid on
xlabel('Subject')
ylabel('Residual')
hold off
```



Now the box plot shows that the larger model accounts for most of the subject-specific effects. The spread of the residuals (the vertical scale of the box plot) is much smaller than in the previous box plot, and the boxes are now mostly centered on zero.

While the 24-parameter model successfully accounts for variations due to the specific subjects in the study, it does not consider the subjects as representatives of a larger population. The sampling distribution from which the subjects are drawn is likely more interesting than the sample itself. The purpose of mixed-effects models is to account for subject-specific variations more broadly, as random effects varying around population means.

Use the `nlmefit` function to fit a mixed-effects model to the data. You can also use `nlmefitsa` in place of `nlmefit` .

The following anonymous function, `nlme_model` , adapts the four-parameter model used by `nlinfit` to the calling syntax of `nlmefit` by allowing separate parameters for each individual. By default, `nlmefit` assigns random effects to all the model parameters. Also by default, `nlmefit` assumes a diagonal covariance matrix (no covariance among the random effects) to avoid overparameterization and related convergence issues.

```
nlme_model = @(PHI,t)(PHI(:,1).*exp(-exp(PHI(:,2)).*t) + ...
                      PHI(:,3).*exp(-exp(PHI(:,4)).*t));
phi0 = [1 2 1 1];
[phi,PSI,stats] = nlmefit(time,concentration,subject, ...
                          [],nlme_model,phi0)


phi =

    2.8277
    0.7729
    0.4606
   -1.3459


PSI =

    0.3264         0         0         0
         0    0.0250         0         0
         0         0    0.0124         0
         0         0         0    0.0000


stats =

  struct with fields:

          dfe: 57
         logl: 54.5882
          mse: 0.0066
         rmse: 0.0787
    errorparam: 0.0815
          aic: -91.1765
          bic: -93.0506
         covb: [4x4 double]
       sebeta: [0.2558 0.1066 0.1092 0.2244]
         ires: [66x1 double]
         pres: [66x1 double]
        iwres: [66x1 double]
        pwres: [66x1 double]
        cwres: [66x1 double]
```

The mean-squared error of 0.0066 is comparable to the 0.0057 of the 24-parameter model without random effects, and significantly better than the 0.0304 of the four-parameter model without random effects.

The estimated covariance matrix `PSI` shows that the variance of the fourth random effect is essentially zero, suggesting that you can remove it to simplify the model. To do this, use the `'REParamsSelect'` name-value pair to specify the indices of the parameters to be modeled with random effects in `nlmefit` .

```
[phi,PSI,stats] = nlmefit(time,concentration,subject, ...
                          [],nlme_model,phi0, ...
                          'REParamsSelect',[1 2 3])


phi =
```

```
    2.8277
    0.7728
    0.4605
   -1.3460


PSI =

    0.3270         0         0
         0    0.0250         0
         0         0    0.0124


stats =

  struct with fields:

          dfe: 58
         logl: 54.5875
          mse: 0.0066
         rmse: 0.0780
    errorparam: 0.0815
          aic: -93.1750
          bic: -94.8410
         covb: [4x4 double]
       sebeta: [0.2560 0.1066 0.1092 0.2244]
         ires: [66x1 double]
         pres: [66x1 double]
        iwres: [66x1 double]
        pwres: [66x1 double]
        cwres: [66x1 double]
```

The log-likelihood `logl` is almost identical to what it was with random effects for all of the parameters, the Akaike information criterion `aic` is reduced from -91.1765 to -93.1750, and the Bayesian information criterion `bic` is reduced from -93.0506 to -94.8410. These measures support the decision to drop the fourth random effect.

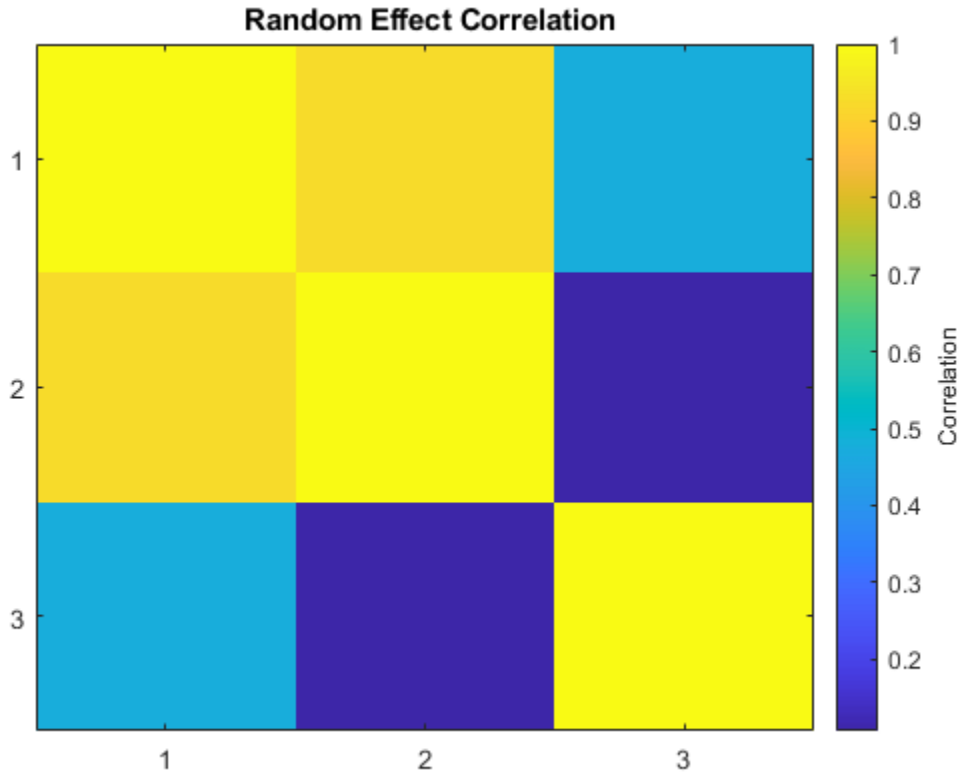
Refitting the simplified model with a full covariance matrix allows for identification of correlations among the random effects. To do this, use the `CovPattern` parameter to specify the pattern of nonzero elements in the covariance matrix.

```
[phi,PSI,stats] = nlmefit(time,concentration,subject, ...
                          [],nlme_model,phi0, ...
                          'REParamsSelect',[1 2 3], ...
                          'CovPattern',ones(3))


phi =

    2.8159
    0.8263
    0.5570
   -1.1478


PSI =
```

```
    0.4737    0.1145    0.0491
    0.1145    0.0323    0.0029
    0.0491    0.0029    0.0227


stats =

  struct with fields:

          dfe: 55
         logl: 58.4419
          mse: 0.0061
         rmse: 0.0783
   errorparam: 0.0781
          aic: -94.8838
          bic: -97.1744
         covb: [4x4 double]
       sebeta: [0.3018 0.1104 0.1170 0.1675]
         ires: [66x1 double]
         pres: [66x1 double]
        iwres: [66x1 double]
        pwres: [66x1 double]
        cwres: [66x1 double]
```

The estimated covariance matrix `PSI` shows that the random effects on the first two parameters have a relatively strong correlation, and both have a relatively weak correlation with the last random effect. This structure in the covariance matrix is more apparent if you convert `PSI` to a correlation matrix using `corrcov` .

```
RHO = corrcov(PSI)
clf;
imagesc(RHO)
set(gca,'XTick',[1 2 3],'YTick',[1 2 3])
title('{\bf Random Effect Correlation}')
h = colorbar;
set(get(h,'YLabel'),'String','Correlation');


RHO =

    1.0000    0.9264    0.4735
    0.9264    1.0000    0.1070
    0.4735    0.1070    1.0000
```



Incorporate this structure into the model by changing the specification of the covariance pattern to block-diagonal.

```
P = [1 1 0;1 1 0;0 0 1] % Covariance pattern
[phi,PSI,stats,b] = nlmefit(time,concentration,subject, ...
                            [],nlme_model,phi0, ...
                            'REParamsSelect',[1 2 3], ...
                            'CovPattern',P)
```

```
P =

     1     1     0
     1     1     0
     0     0     1


phi =

    2.7830
    0.8981
    0.6581
   -1.0000


PSI =

    0.5180    0.1069         0
```

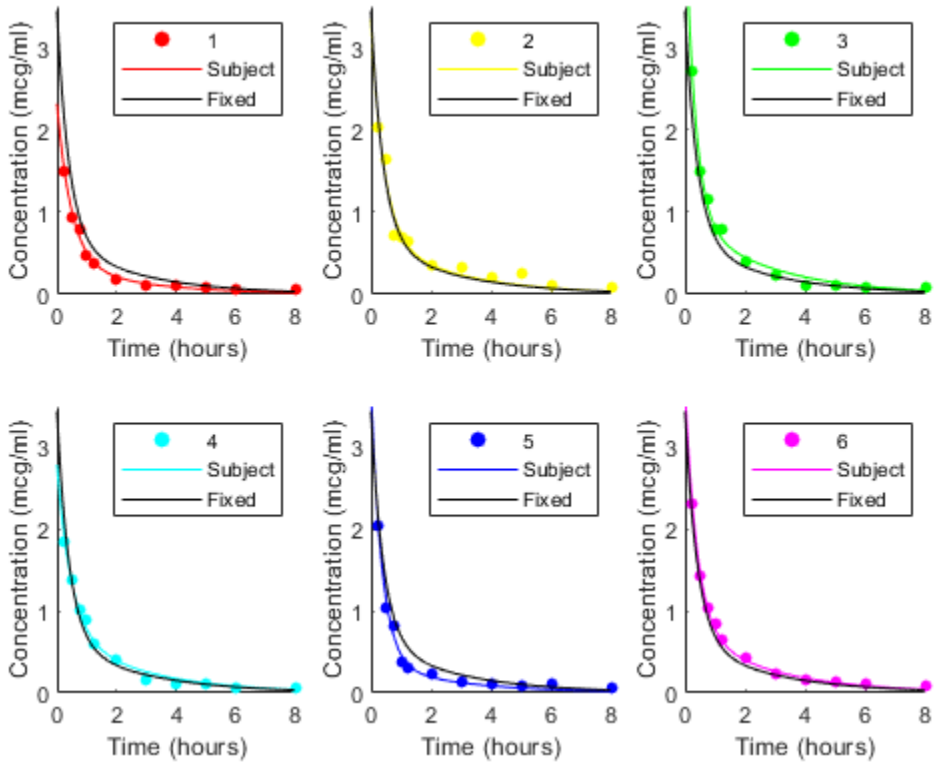
```
    0.1069    0.0221         0
         0         0    0.0454


stats =

  struct with fields:

          dfe: 57
         logl: 58.0804
          mse: 0.0061
         rmse: 0.0768
   errorparam: 0.0782
          aic: -98.1608
          bic: -100.0350
         covb: [4x4 double]
       sebeta: [0.3171 0.1073 0.1384 0.1453]
         ires: [66x1 double]
         pres: [66x1 double]
        iwres: [66x1 double]
        pwres: [66x1 double]
        cwres: [66x1 double]


b =

   -0.8507   -0.1563    1.0427   -0.7559    0.5652    0.1550
   -0.1756   -0.0323    0.2152   -0.1560    0.1167    0.0320
   -0.2756    0.0519    0.2620    0.1064   -0.2835    0.1389
```

The block-diagonal covariance structure reduces `aic` from -94.9462 to -98.1608 and `bic` from -97.2368 to -100.0350 without significantly affecting the log-likelihood. These measures support the covariance structure used in the final model. The output `b` gives predictions of the three random effects for each of the six subjects. These are combined with the estimates of the fixed effects in `phi` to produce the mixed-effects model.

Plot the mixed-effects model for each of the six subjects. For comparison, the model without random effects is also shown.

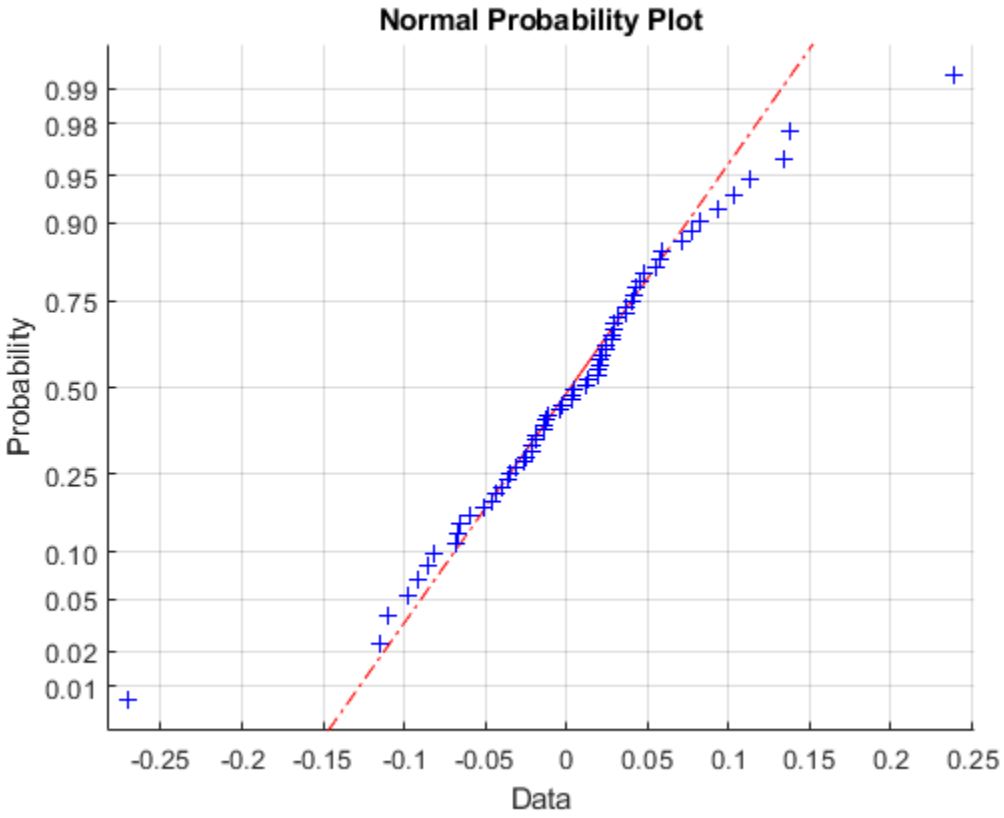
```
PHI = repmat(phi,1,6) + ...                 % Fixed effects
      [b(1,:);b(2,:);b(3,:);zeros(1,6)];    % Random effects
RES = zeros(11,6); % Residuals
colors = 'rygcbm';
for I = 1:6
    fitted_model = @(t)(PHI(1,I)*exp(-exp(PHI(2,I))*t) + ...
                        PHI(3,I)*exp(-exp(PHI(4,I))*t));
    tI = time(subject == I);
    cI = concentration(subject == I);
    RES(:,I) = cI - fitted_model(tI);

    subplot(2,3,I)
    scatter(tI,cI,20,colors(I),'filled')
    hold on
    plot(tplot,fitted_model(tplot),'Color',colors(I))
    plot(tplot,model(phi,tplot),'k')
    axis([0 8 0 3.5])
    xlabel('Time (hours)')
```

```
    ylabel('Concentration (mcg/ml)')
    legend(num2str(I),'Subject','Fixed')
end
```



If obvious outliers in the data (visible in previous box plots) are ignored, a normal probability plot of the residuals shows reasonable agreement with model assumptions on the errors.

```
clf; normplot(RES(:))
```


Normal Probability Plot
Probability
0.99
0.98
0.95
0.90
0.75
0.50
0.25
0.10
0.05
0.02
0.01
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
Data

Weighted Nonlinear Regression

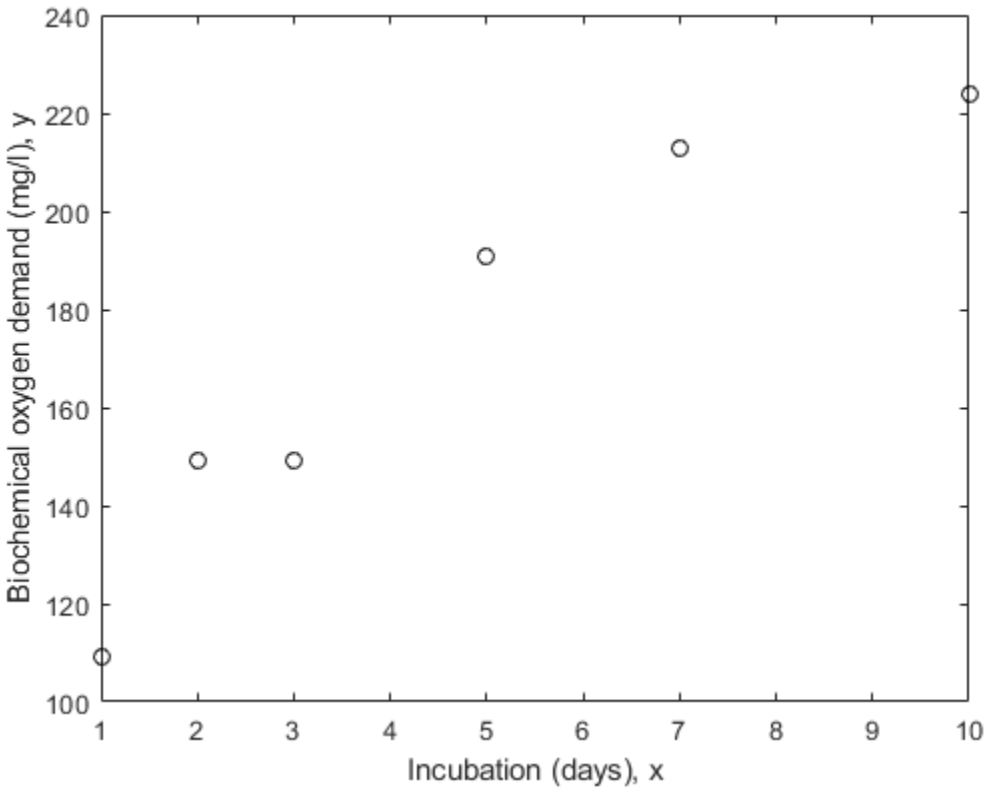
This example shows how to fit a nonlinear regression model for data with nonconstant error variance.

Regular nonlinear least squares algorithms are appropriate when measurement errors all have the same variance. When that assumption is not true, it is appropriate to used a weighted fit. This example shows how to use weights with the `fitnlm` function.

Data and Model for the Fit

We'll use data collected to study water pollution caused by industrial and domestic waste. These data are described in detail in Box, G.P., W.G. Hunter, and J.S. Hunter, Statistics for Experimenters (Wiley, 1978, pp. 483-487). The response variable is biochemical oxygen demand in mg/l, and the predictor variable is incubation time in days.

```
x = [1 2 3 5 7 10]';
y = [109 149 149 191 213 224]';

plot(x,y,'ko');
xlabel('Incubation (days), x'); ylabel('Biochemical oxygen demand (mg/l), y');
```



We'll assume that it is known that the first two observations were made with less precision than the remaining observations. They might, for example, have been made with a different instrument. Another common reason to weight data is that each recorded observation is actually the mean of several measurements taken at the same value of x. In the data here, suppose the first two values represent a single raw measurement, while the remaining four are each the mean of 5 raw

measurements. Then it would be appropriate to weight by the number of measurements that went into each observation.

```
w = [1 1 5 5 5 5]';
```

The model we'll fit to these data is a scaled exponential curve that becomes level as x becomes large.

```
modelFun = @(b,x) b(1).*(1-exp(-b(2).*x));
```

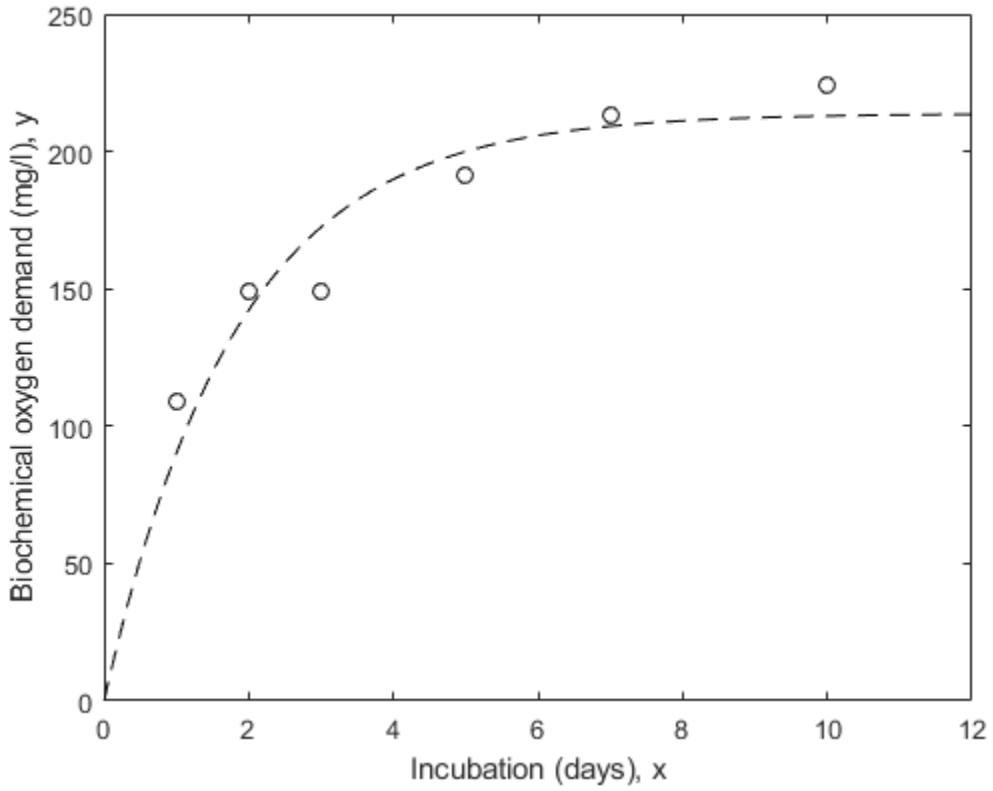
Just based on a rough visual fit, it appears that a curve drawn through the points might level out at a value of around 240 somewhere in the neighborhood of x = 15. So we'll use 240 as the starting value for b1, and since e^(-.5*15) is small compared to 1, we'll use .5 as the starting value for b2.

```
start = [240; .5];
```

Fit the Model without Weights

The danger in ignoring measurement error is that the fit may be overly influenced by imprecise measurements, and may therefore not provide a good fit to measurements that are known precisely. Let's fit the data without weights and compare it to the points.

```
nlm = fitnlm(x,y,modelFun,start);
xx = linspace(0,12)';
line(xx,predict(nlm,xx),'linestyle','--','color','k')
```

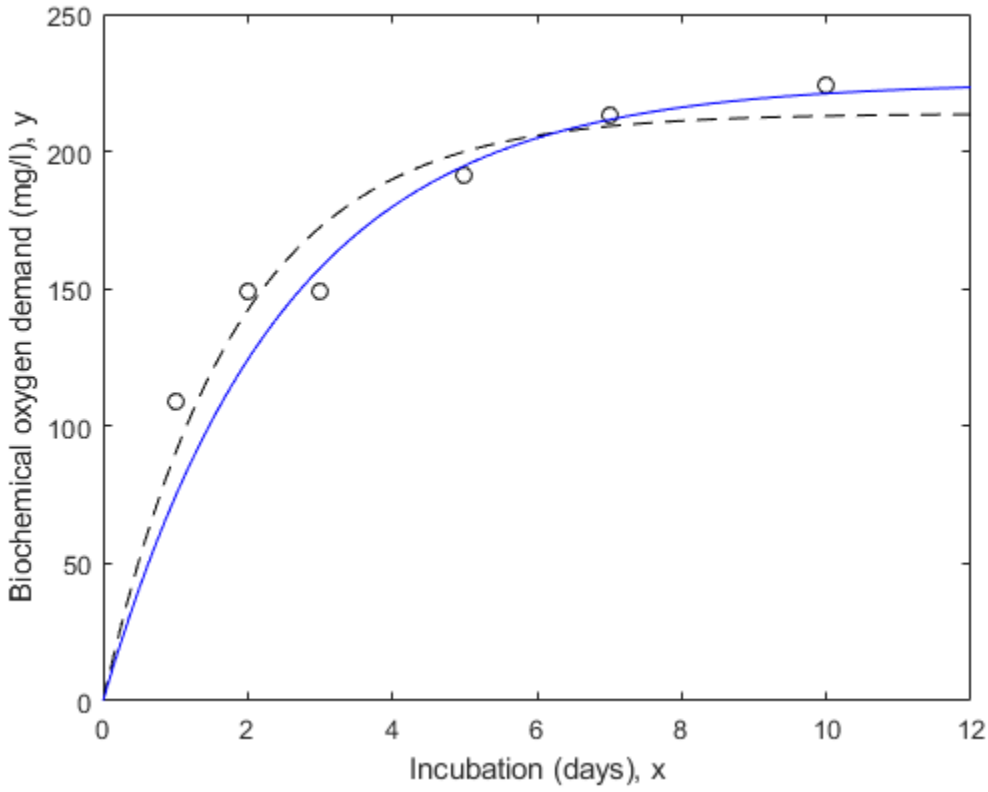


Fit the Model with Weights

Notice that the fitted curve is pulled toward the first two points, but seems to miss the trend of the other points. Let's try repeating the fit using weights.

```
wnlm = fitnlm(x,y,modelFun,start,'Weight',w)
line(xx,predict(wnlm,xx),'color','b')
```

```
wnlm =


Nonlinear regression model:
    y ~ b1*(1 - exp( - b2*x))

Estimated Coefficients:
          Estimate       SE       tStat       pValue
          ________    ________    ______    __________

    b1     225.17         10.7    21.045    3.0134e-05
    b2    0.40078     0.064296    6.2333     0.0033745


Number of observations: 6, Error degrees of freedom: 4
Root Mean Squared Error: 24
R-Squared: 0.908,  Adjusted R-Squared 0.885
F-statistic vs. zero model: 696, p-value = 8.2e-06
```



The estimated population standard deviation in this case describes the average variation for a "standard" observation with a weight, or measurement precision, of 1.

```
wnlm.RMSE
```

```
ans =

   24.0096
```

An important part of any analysis is an estimate of the precision of the model fit. The coefficient display shows standard errors for the parameters, but we can also compute confidence intervals for them.

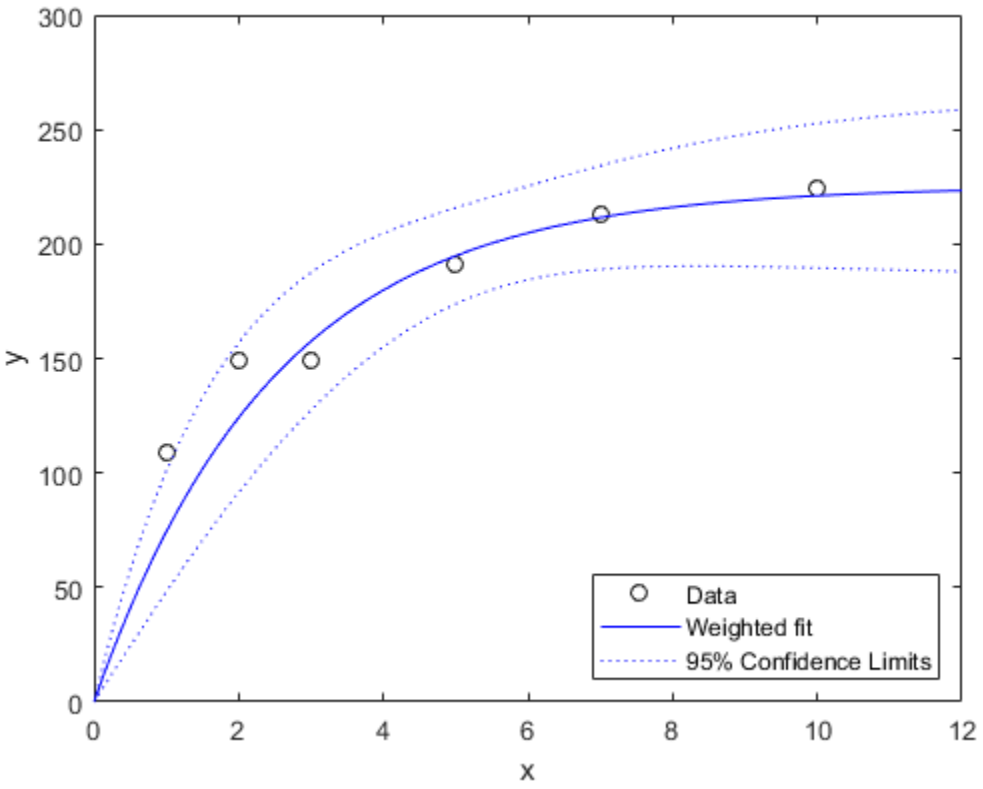
```
coefCI(wnlm)
```

```
ans =

  195.4650  254.8788
    0.2223    0.5793
```

Estimate the Response Curve

Next, we'll compute the fitted response values and confidence intervals for them. By default, those widths are for pointwise confidence bounds for the predicted value, but we will request simultaneous intervals for the entire curve.

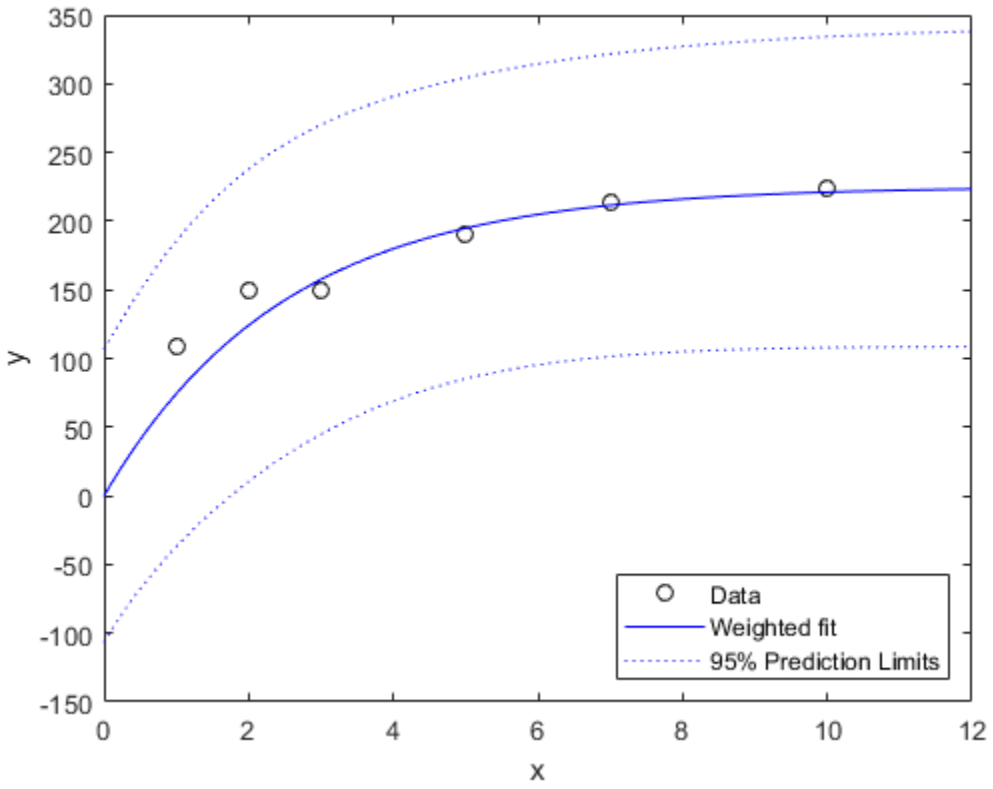
```
[ypred,ypredci] = predict(wnlm,xx,'Simultaneous',true);
plot(x,y,'ko', xx,ypred,'b-', xx,ypredci,'b:');
xlabel('x'); ylabel('y');
legend({'Data', 'Weighted fit', '95% Confidence Limits'},'location','SouthEast');
```



Notice that the two downweighted points are not fit as well by the curve as the remaining points. That's as you would expect for a weighted fit.

It's also possible to estimate prediction intervals for future observations at specified values of x. Those intervals will in effect assume a weight, or measurement precision, of 1.

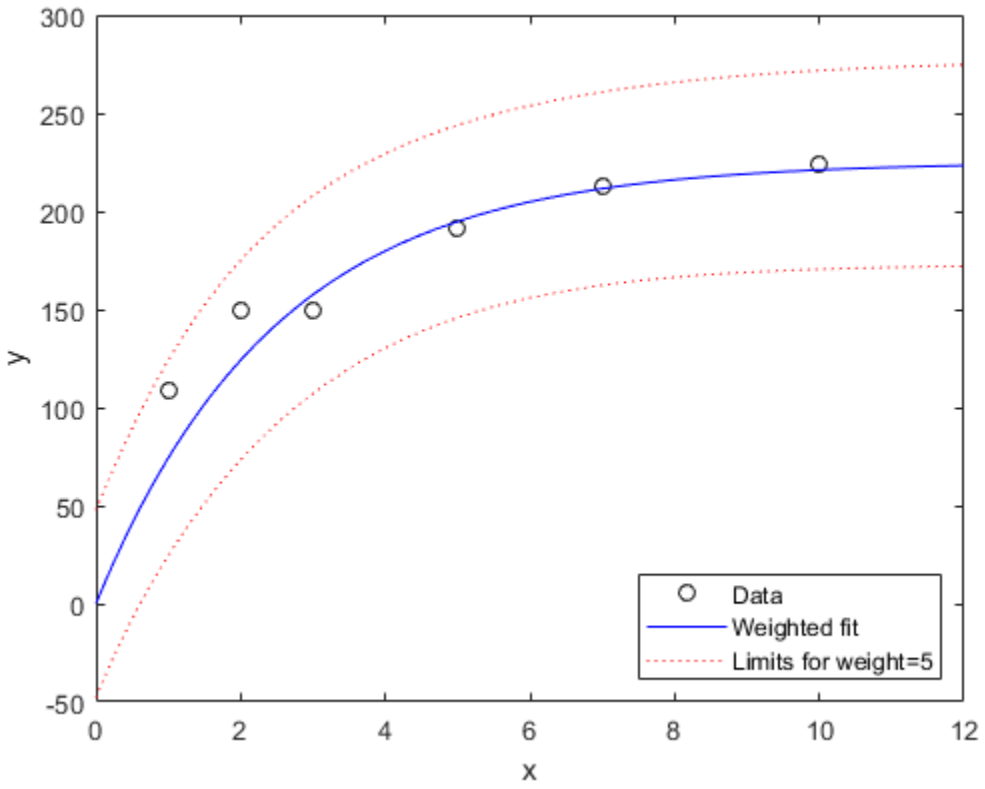
```
[ypred,ypredci] = predict(wnlm,xx,'Simultaneous',true,'Prediction','observation');
plot(x,y,'ko', xx,ypred,'b-', xx,ypredci,'b:');
xlabel('x'); ylabel('y');
legend({'Data', 'Weighted fit', '95% Prediction Limits'},'location','SouthEast');
```



The absolute scale of the weights actually doesn't affect the parameter estimates. Rescaling the weights by any constant would have given us the same estimates. But they do affect the confidence bounds, since the bounds represent an observation with weight 1. Here you can see that the points with higher weight seem too close to the fitted line, compared with the confidence limits.

While the `predict` method doesn't allow us to change the weights, it is possible for us to do some post-processing and investigate how the curve would look for a more precise estimate. Suppose we are interested in a new observation that is based on the average of five measurements, just like the last four points in this plot. We could reduce the width of the intervals by a factor of sqrt(5).

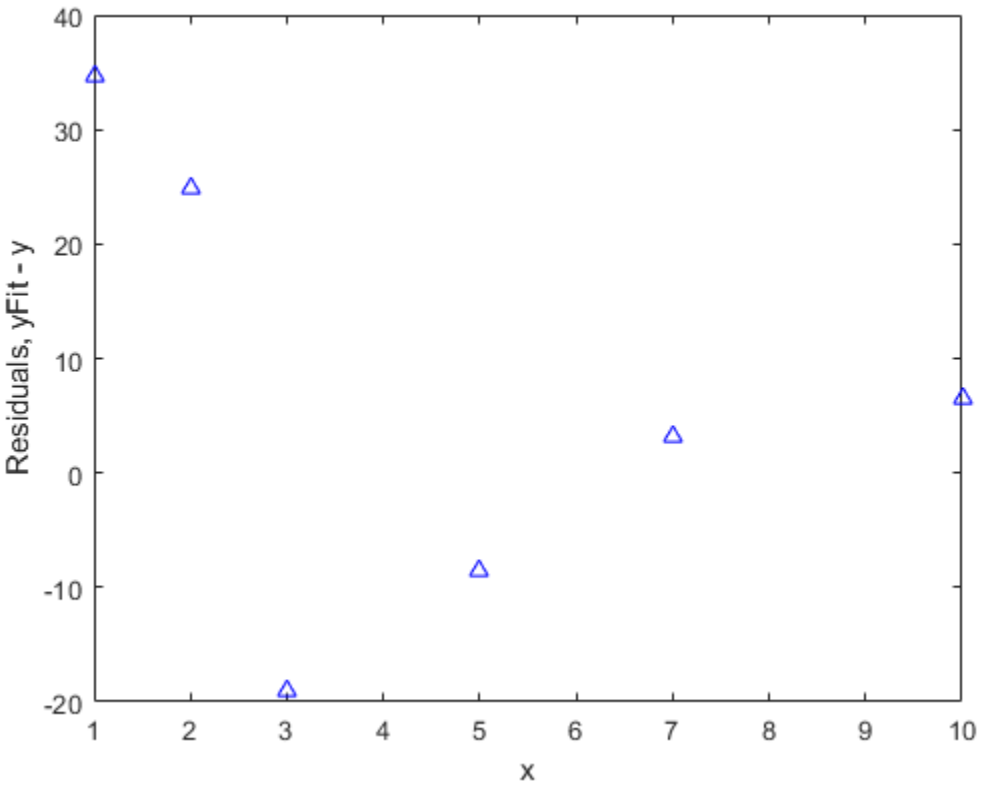
```
halfwidth = ypredci(:,2)-ypred;
newwidth = halfwidth/sqrt(5);
newci = [ypred-newwidth, ypred+newwidth];
plot(x,y,'ko', xx,ypred,'b-', xx,newci,'r:');
xlabel('x'); ylabel('y');
legend({'Data', 'Weighted fit', 'Limits for weight=5'},'location','SouthEast');
```



Residual Analysis

In addition to plotting the data and the fit, we'll plot residuals from a fit against the predictors, to diagnose any problems with the model. The residuals should appear independent and identically distributed but with a variance proportional to the inverse of the weights. We can standardize this variance to make the plot easier to interpret.

```
r = wnlm.Residuals.Raw;
plot(x,r.*sqrt(w),'b^');
xlabel('x'); ylabel('Residuals, yFit - y');
```



There is some evidence of systematic patterns in this residual plot. Notice how the last four residuals have a linear trend, suggesting that the model might not increase fast enough as x increases. Also, the magnitude of the residuals tends to decrease as x increases, suggesting that measurement error may depend on x. These deserve investigation, however, there are so few data points, that it's hard to attach significance to these apparent patterns.

Pitfalls in Fitting Nonlinear Models by Transforming to Linearity

This example shows pitfalls that can occur when fitting a nonlinear model by transforming to linearity. Imagine that we have collected measurements on two variables, x and y, and we want to model y as a function of x. Assume that x is measured exactly, while measurements of y are affected by additive, symmetric, zero-mean errors.

```
x = [5.72 4.22 5.72 3.59 5.04 2.66 5.02 3.11 0.13 2.26 ...
     5.39 2.57 1.20 1.82 3.23 5.46 3.15 1.84 0.21 4.29 ...
     4.61 0.36 3.76 1.59 1.87 3.14 2.45 5.36 3.44 3.41]';
y = [2.66 2.91 0.94 4.28 1.76 4.08 1.11 4.33 8.94 5.25 ...
     0.02 3.88 6.43 4.08 4.90 1.33 3.63 5.49 7.23 0.88 ...
     3.08 8.12 1.22 4.24 6.21 5.48 4.89 2.30 4.13 2.17]';
```

Let's also assume that theory tells us that these data should follow a model of exponential decay, y = p1*exp(p2*x), where p1 is positive and p2 is negative. To fit this model, we could use nonlinear least squares.

```
modelFun = @(p,x) p(1)*exp(p(2)*x);
```

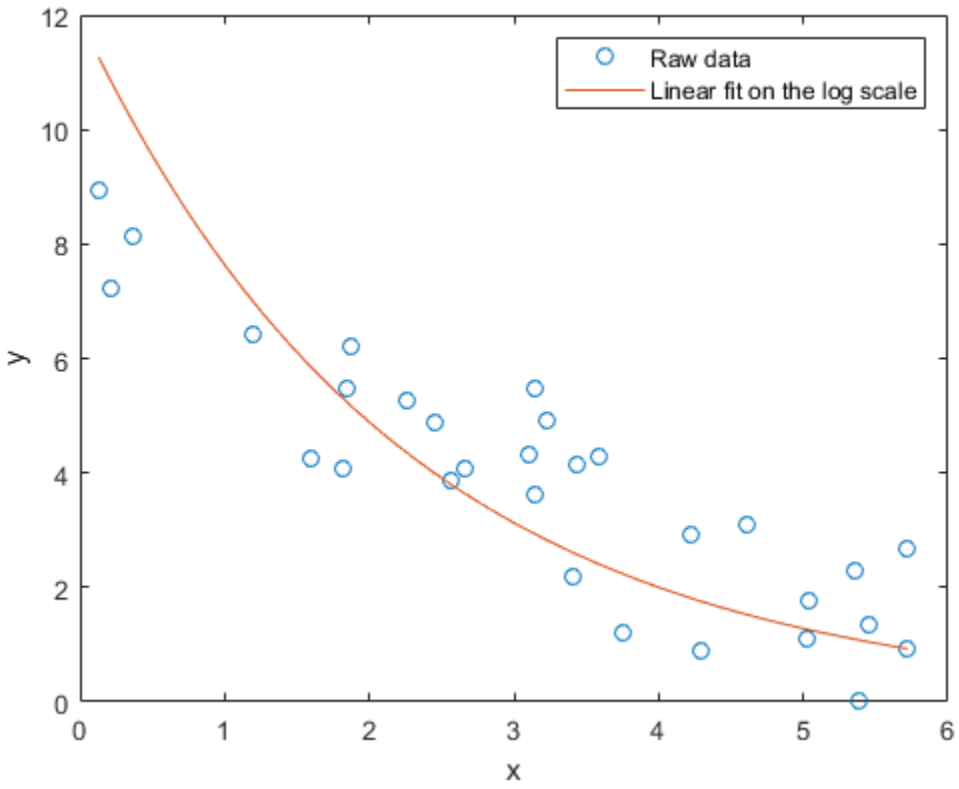
But the nonlinear model can also be transformed to a linear one by taking the log on both sides, to get log(y) = log(p1) + p2*x. That's tempting, because we can fit that linear model by ordinary linear least squares. The coefficients we'd get from a linear least squares would be log(p1) and p2.

```
paramEstsLin = [ones(size(x)), x] \ log(y);
paramEstsLin(1) = exp(paramEstsLin(1))
```

```
paramEstsLin =

   11.9312
   -0.4462
```

How did we do? We can superimpose the fit on the data to find out.

```
xx = linspace(min(x), max(x));
yyLin = modelFun(paramEstsLin, xx);
plot(x,y,'o', xx,yyLin,'-');
xlabel('x'); ylabel('y');
legend({'Raw data','Linear fit on the log scale'},'location','NE');
```

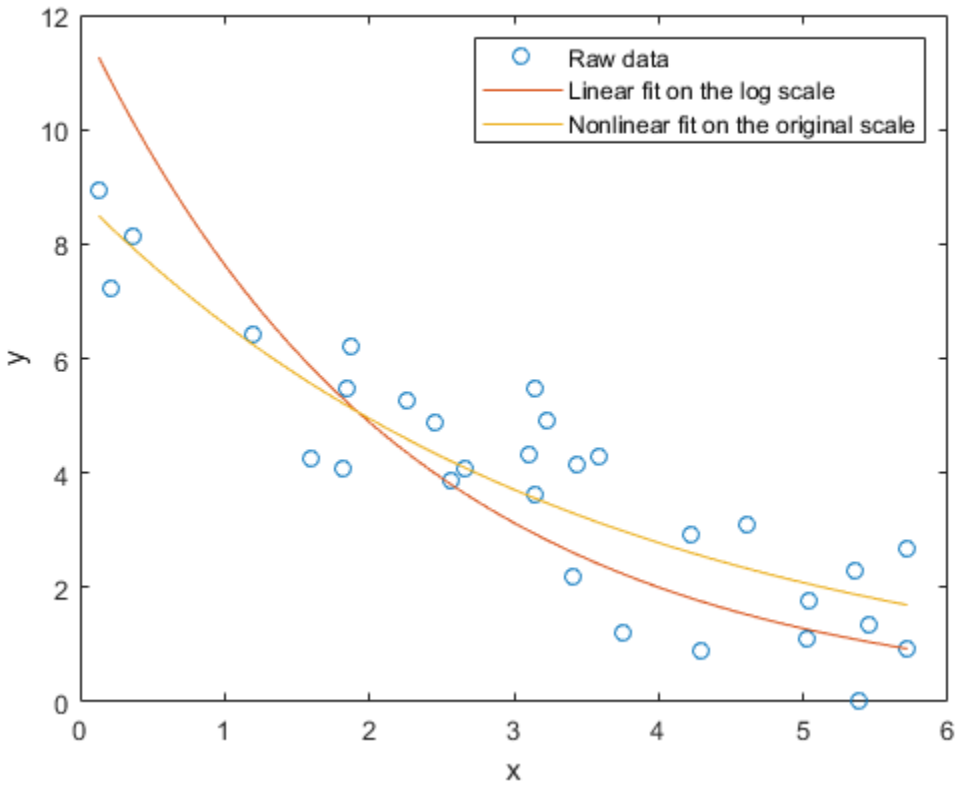


Something seems to have gone wrong, because the fit doesn't really follow the trend that we can see in the raw data. What kind of fit would we get if we used `nlinfit` to do nonlinear least squares instead? We'll use the previous fit as a rough starting point, even though it's not a great fit.

```
paramEsts = nlinfit(x, y, modelFun, paramEstsLin)
```

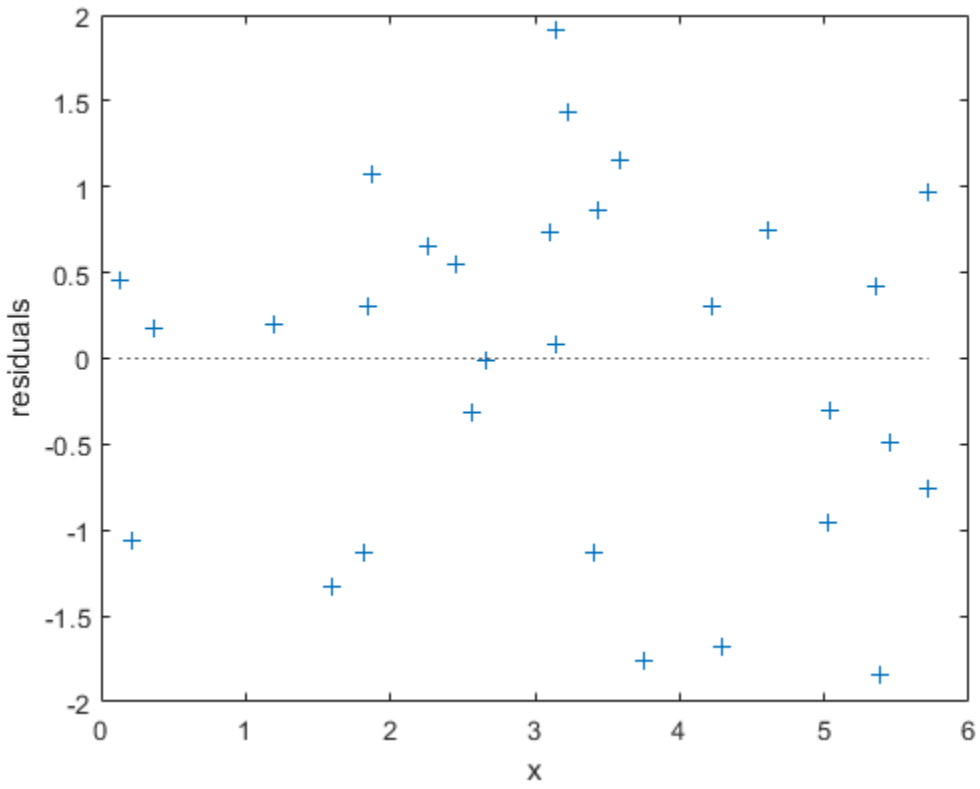
```
paramEsts =

    8.8145
   -0.2885
```

```
yy = modelFun(paramEsts,xx);
plot(x,y,'o', xx,yyLin,'-', xx,yy,'-');
xlabel('x'); ylabel('y');
legend({'Raw data','Linear fit on the log scale',  ...
    'Nonlinear fit on the original scale'},'location','NE');
```



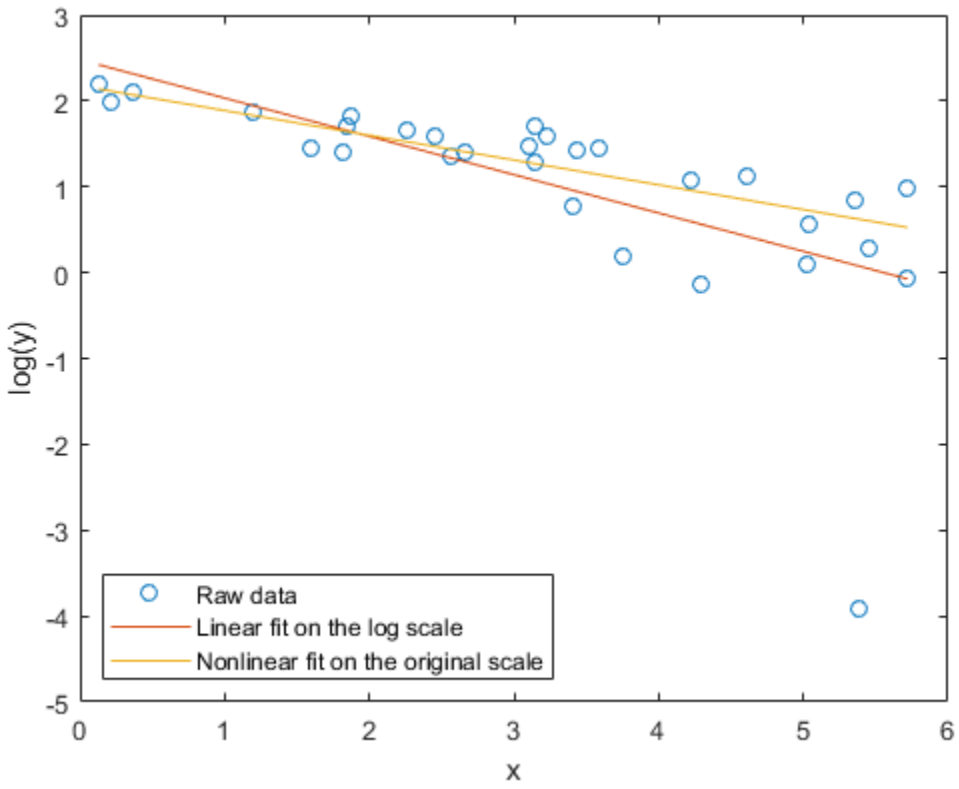
The fit using `nlinfit` more or less passes through the center of the data point scatter. A residual plot shows something approximately like an even scatter about zero.

```
r = y-modelFun(paramEsts,x);
plot(x,r,'+', [min(x) max(x)],[0 0],'k:');
xlabel('x'); ylabel('residuals');
```



So what went wrong with the linear fit? The problem is in log transform. If we plot the data and the two fits on the log scale, we can see that there's an extreme outlier.

```
plot(x,log(y),'o', xx,log(yyLin),'-', xx,log(yy),'-');
xlabel('x'); ylabel('log(y)');
ylim([-5,3]);
legend({'Raw data', 'Linear fit on the log scale',  ...
    'Nonlinear fit on the original scale'},'location','SW');
```



That observation is not an outlier in the original data, so what happened to make it one on the log scale? The log transform is exactly the right thing to straighten out the trend line. But the log is a very nonlinear transform, and so symmetric measurement errors on the original scale have become asymmetric on the log scale. Notice that the outlier had the smallest y value on the original scale -- close to zero. The log transform has "stretched out" that smallest y value more than its neighbors. We made the linear fit on the log scale, and so it is very much affected by that outlier.

Had the measurement at that one point been slightly different, the two fits might have been much more similar. For example,

```
y(11) = 1;
paramEsts = nlinfit(x, y, modelFun, [10;-.3])
```

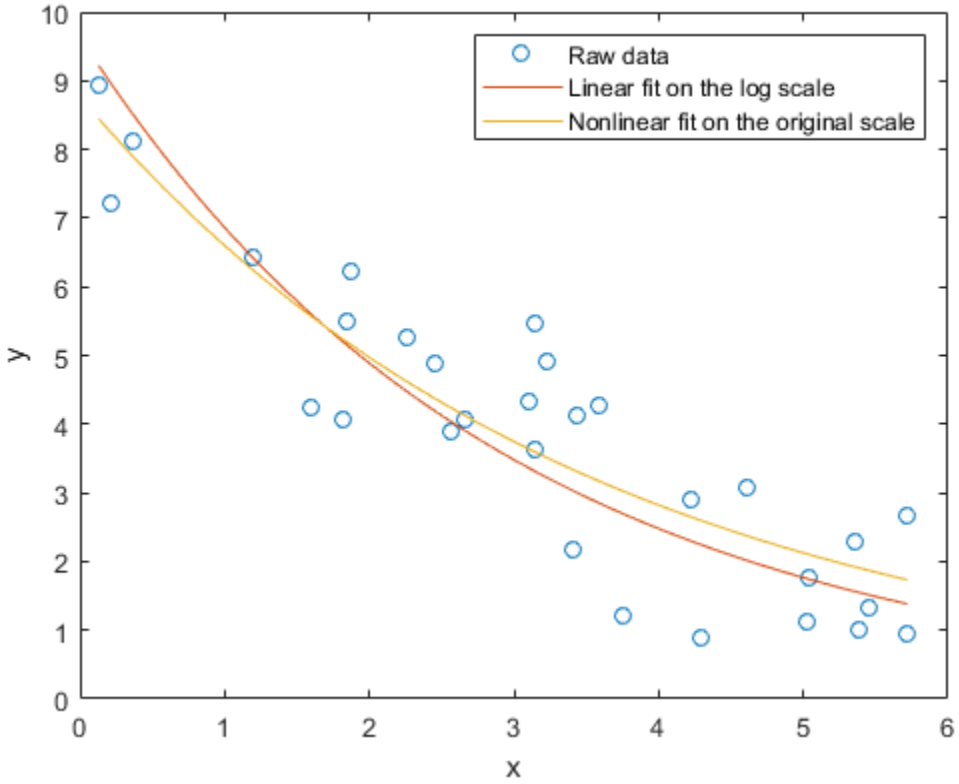
```
paramEsts =

    8.7618
   -0.2833
```

```
paramEstsLin = [ones(size(x)), x] \ log(y);
paramEstsLin(1) = exp(paramEstsLin(1))
```

```
paramEstsLin =

    9.6357
```

```
    -0.3394
```

```
yy = modelFun(paramEsts,xx);
yyLin = modelFun(paramEstsLin, xx);
plot(x,y,'o', xx,yyLin,'-', xx,yy,'-');
xlabel('x'); ylabel('y');
legend({'Raw data', 'Linear fit on the log scale',  ...
    'Nonlinear fit on the original scale'},'location','NE');
```



Still, the two fits are different. Which one is "right"? To answer that, suppose that instead of additive measurement errors, measurements of y were affected by multiplicative errors. These errors would not be symmetric, and least squares on the original scale would not be appropriate. On the other hand, the log transform would make the errors symmetric on the log scale, and the linear least squares fit on that scale is appropriate.

So, which method is "right" depends on what assumptions you are willing to make about your data. In practice, when the noise term is small relative to the trend, the log transform is "locally linear" in the sense that y values near the same x value will not be stretched out too asymmetrically. In that case, the two methods lead to essentially the same fit. But when the noise term is not small, you should consider what assumptions are realistic, and choose an appropriate fitting method.

Nonlinear Logistic Regression

This example shows two ways of fitting a nonlinear logistic regression model. The first method uses maximum likelihood (ML) and the second method uses generalized least squares (GLS) via the function `fitnlm` from Statistics and Machine Learning Toolbox™.

Problem Description

Logistic regression is a special type of regression in which the goal is to model the probability of something as a function of other variables. Consider a set of predictor vectors $x_1, \ldots, x_N$ where N is the number of observations and x_i is a column vector containing the values of the d predictors for the i th observation. The response variable for x_i is Z_i where Z_i represents a Binomial random variable with parameters n, the number of trials, and μ_i, the probability of success for trial i. The normalized response variable is $Y_i = Z_i/n$ - the proportion of successes in n trials for observation i. Assume that responses Y_i are independent for $i = 1, \ldots, N$. For each i:

$$E(Y_i) = \mu_i$$

$$Var(Y_i) = \frac{\mu_i(1-\mu_i)}{n}.$$

Consider modeling μ_i as a function of predictor variables x_i.

In linear logistic regression, you can use the function `fitglm` to model μ_i as a function of x_i as follows:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i^T\beta$$

with β representing a set of coefficients multiplying the predictors in x_i. However, suppose you need a nonlinear function on the right-hand-side:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = f(x_i, \beta)\,.$$

There are functions in Statistics and Machine Learning Toolbox™ for fitting nonlinear regression models, but not for fitting nonlinear logistic regression models. This example shows how you can use toolbox functions to fit those models.

Direct Maximum Likelihood (ML)

The ML approach maximizes the log likelihood of the observed data. The likelihood is easily computed using the Binomial probability (or density) function as computed by the `binopdf` function.

Generalized Least Squares (GLS)

You can estimate a nonlinear logistic regression model using the function `fitnlm`. This might seem surprising at first since `fitnlm` does not accommodate Binomial distribution or any link functions. However, `fitnlm` can use Generalized Least Squares (GLS) for model estimation if you specify the mean and variance of the response. If GLS converges, then it solves the same set of nonlinear equations for estimating β as solved by ML. You can also use GLS for quasi-likelihood estimation of generalized linear models. In other words, we should get the same or equivalent solutions from GLS

and ML. To implement GLS estimation, provide the nonlinear function to fit, and the variance function for the Binomial distribution.

Mean or model function

The model function describes how μ_i changes with β. For `fitnlm`, the model function is:

$$\mu_i = \frac{1}{1 + \exp\{-f(x_i, \beta)\}}$$

Weight function

`fitnlm` accepts observation weights as a function handle using the `'Weights'` name-value pair argument. When using this option, `fitnlm` assumes the following model:

$$E(Y_i) = \mu_i$$

$$Var(Y_i) = \frac{\sigma^2}{w(\mu_i)}$$

where responses Y_i are assumed to be independent, and w is a custom function handle that accepts μ_i and returns an observation weight. In other words, the weights are inversely proportional to the response variance. For the Binomial distribution used in the logistic regression model, create the weight function as follows:

$$w(\mu_i) = \frac{1}{Var(y_i)} = \frac{n}{\mu_i(1-\mu_i)}$$

`fitnlm` models the variance of the response Y_i as $\sigma^2/w(\mu_i)$ where σ^2 is an extra parameter that is present in GLS estimation, but absent in the logistic regression model. However, this typically does not affect the estimation of β, and it provides a "dispersion parameter" to check on the assumption that the Z_i values have a Binomial distribution.

An advantage of using `fitnlm` over direct ML is that you can perform hypothesis tests and compute confidence intervals on the model coefficients.

Generate Example Data

To illustrate the differences between ML and GLS fitting, generate some example data. Assume that x_i is one dimensional and suppose the true function f in the nonlinear logistic regression model is the Michaelis-Menten model parameterized by a 2×1 vector β:

$$f(x_i, \beta) = \frac{\beta_1 x_i}{\beta_2 + x_i}.$$

```
myf = @(beta,x) beta(1)*x./(beta(2) + x);
```

Create a model function that specifies the relationship between μ_i and β.

```
mymodelfun = @(beta,x) 1./(1 + exp(-myf(beta,x)));
```

Create a vector of one dimensional predictors and the true coefficient vector β.

```
rng(300,'twister');
x    = linspace(-1,1,200)';
beta = [10;2];
```

Compute a vector of μ_i values for each predictor.

```
mu = mymodelfun(beta,x);
```

Generate responses z_i from a Binomial distribution with success probabilities μ_i and number of trials n.

```
n = 50;
z = binornd(n,mu);
```

Normalize the responses.

```
y = z./n;
```

ML Approach

The ML approach defines the negative log likelihood as a function of the β vector, and then minimizes it with an optimization function such as `fminsearch`. Specify `beta0` as the starting value for β.

```
mynegloglik = @(beta) -sum(log(binopdf(z,n,mymodelfun(beta,x))));
beta0 = [3;3];
opts = optimset('fminsearch');
opts.MaxFunEvals = Inf;
opts.MaxIter = 10000;
betaHatML = fminsearch(mynegloglik,beta0,opts)
```

```
betaHatML = 2×1

    9.9259
    1.9720
```

The estimated coefficients in `betaHatML` are close to the true values of `[10;2]`.

GLS Approach

The GLS approach creates a weight function for `fitnlm` previously described.

```
wfun = @(xx) n./(xx.*(1-xx));
```

Call `fitnlm` with custom mean and weight functions. Specify `beta0` as the starting value for β.

```
nlm = fitnlm(x,y,mymodelfun,beta0,'Weights',wfun)
```

```
nlm =
Nonlinear regression model:
    y ~ F(beta,x)

Estimated Coefficients:
          Estimate      SE       tStat       pValue
          ________    _______    ______    __________

    b1     9.926      0.83135     11.94     4.193e-25
    b2     1.972      0.16438    11.996    2.8182e-25
```

```
Number of observations: 200, Error degrees of freedom: 198
Root Mean Squared Error: 1.16
R-Squared: 0.995,  Adjusted R-Squared 0.995
F-statistic vs. zero model: 1.88e+04, p-value = 2.04e-226
```

Get an estimate of β from the fitted `NonLinearModel` object `nlm`.

```
betaHatGLS = nlm.Coefficients.Estimate
```

```
betaHatGLS = 2×1

    9.9260
    1.9720
```

As in the ML method, the estimated coefficients in `betaHatGLS` are close to the true values of `[10;2]`. The small p-values for both β_1 and β_2 indicate that both coefficients are significantly different from 0.

Compare ML and GLS Approaches

Compare estimates of β.

```
max(abs(betaHatML - betaHatGLS))
```

```
ans = 1.1460e-05
```

Compare fitted values using ML and GLS

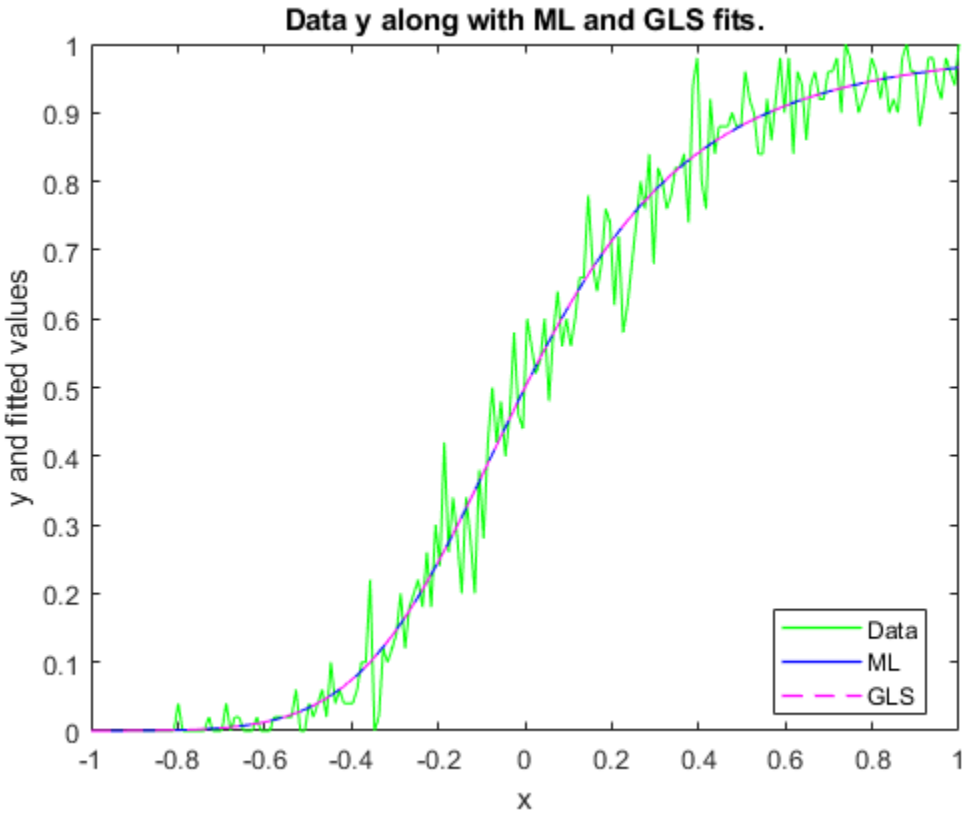
```
yHatML  = mymodelfun(betaHatML ,x);
yHatGLS = mymodelfun(betaHatGLS,x);
max(abs(yHatML - yHatGLS))
```

```
ans = 1.2746e-07
```

ML and GLS approaches produce similar solutions.

Plot fitted values using ML and GLS

```
figure
plot(x,y,'g','LineWidth',1)
hold on
plot(x,yHatML ,'b'  ,'LineWidth',1)
plot(x,yHatGLS,'m--','LineWidth',1)
legend('Data','ML','GLS','Location','Best')
xlabel('x')
ylabel('y and fitted values')
title('Data y along with ML and GLS fits.')
```

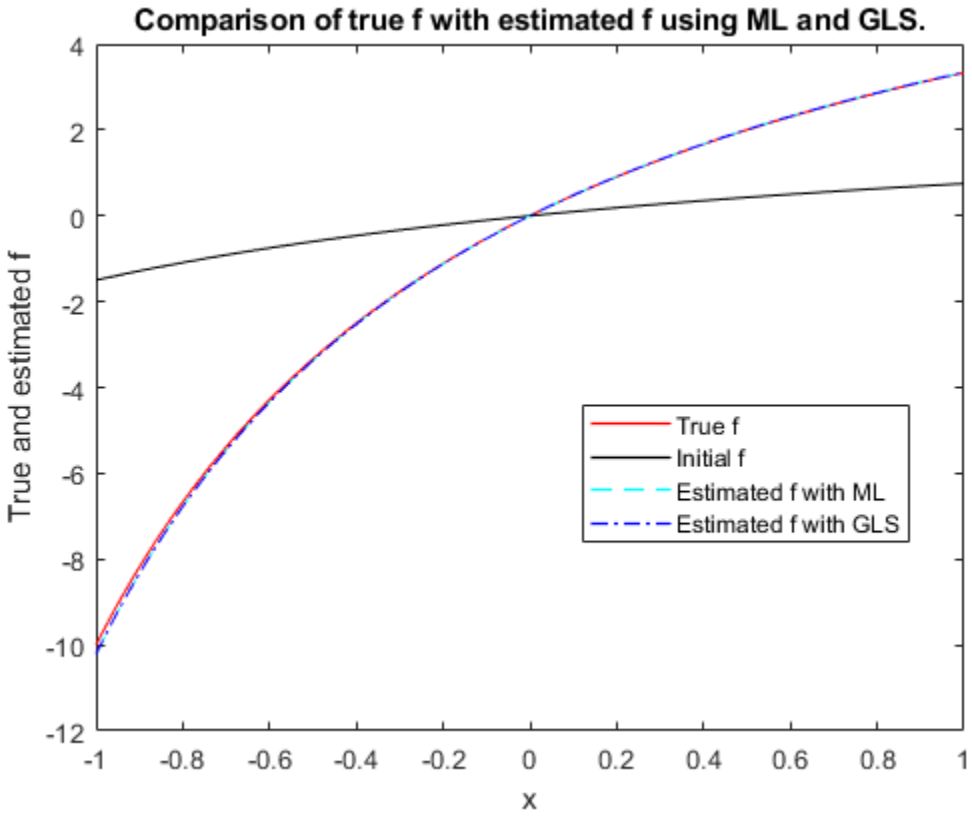


ML and GLS produce similar fitted values.

Plot estimated nonlinear function using ML and GLS

Plot true model for $f(x_i, \beta)$. Add plot for the initial estimate of $f(x_i, \beta)$ using $\beta = \beta_0$ and plots for ML and GLS based estimates of $f(x_i, \beta)$.

```
figure
plot(x,myf(beta,x),'r','LineWidth',1)
hold on
plot(x,myf(beta0,x),'k','LineWidth',1)
plot(x,myf(betaHatML,x),'c--','LineWidth',1)
plot(x,myf(betaHatGLS,x),'b-.','LineWidth',1)
legend('True f','Initial f','Estimated f with ML','Estimated f with GLS','Location','Best')
xlabel('x')
ylabel('True and estimated f')
title('Comparison of true f with estimated f using ML and GLS.')
```



The estimated nonlinear function f using both ML and GLS methods is close to the true nonlinear function f. You can use a similar technique to fit other nonlinear generalized linear models like nonlinear Poisson regression.

Survival Analysis

What Is Survival Analysis?

In this section...

Introduction

Survival analysis is time-to-event analysis, that is, when the outcome of interest is the time until an event occurs. Examples of time-to-events are the time until infection, reoccurrence of a disease, or recovery in health sciences, duration of unemployment in economics, time until the failure of a machine part or lifetime of light bulbs in engineering, and so on. Survival analysis is a part of reliability studies in engineering. In this case, it is usually used to study the lifetime of industrial components. In reliability analyses, survival times are usually called failure times as the variable of interest is how much time a component functions properly before it fails.

Survival analysis consists of parametric, semiparametric, and nonparametric methods. You can use these to estimate the most commonly used measures in survival studies, survivor and hazard functions, compare them for different groups, and assess the relationship of predictor variables to survival time. Some statistical probability distributions describe survival times well. Commonly used distributions are exponential, Weibull, lognormal, Burr, and Birnbaum-Saunders distributions. Statistics and Machine Learning Toolbox functions `ecdf` and `ksdensity` compute the empirical and kernel density estimates of the cdf, cumulative hazard, and survivor functions. `coxphfit` fits the Cox proportional hazards model to the data.

Censoring

One important concept in survival analysis is censoring. The survival times of some individuals might not be fully observed due to different reasons. In life sciences, this might happen when the survival study (e.g., the clinical trial) stops before the full survival times of all individuals can be observed, or a person drops out of a study, or for long-term studies, when the patient is lost to follow up. In the industrial context, not all components might have failed before the end of the reliability study. In such cases, the individual survives beyond the time of the study, and the exact survival time is unknown. This is called right censoring.

During a survival study either the individual is observed to fail at time T, or the observation on that individual ceases at time c. Then the observation is $\min(T,c)$ and an indicator variable I_c shows if the individual is censored or not. The calculations for hazard and survivor functions must be adjusted to account for censoring. Statistics and Machine Learning Toolbox functions such as `ecdf`, `ksdensity`, `coxphfit`, and `mle` account for censoring.

Data

Survival data usually consists of the time until an event of interest occurs and the censoring information for each individual or component. The following table shows the fictitious unemployment time of individuals in a 6-month study. Two individuals are right censored (indicated by a censoring

value of 1). One individual was still unemployed after the 24th week, when the study ended. Contact with the other censored individual was lost at the end of the 21st week.

Unemployment Time (Weeks)	Censoring
14	0
23	0
7	0
21	1
19	0
16	0
24	1
8	0

Survival data might also include the number of failures at a certain time (the number of times a particular survival or failure time was observed). The following table shows the simulated time until a light-emitting diodes drops to 70% of its full light output level, in hours, in an accelerated life test.

Failure Time (hrs)	Frequency
8600	6
15300	19
22000	11
28600	20
35300	17
42000	14
48700	8
55400	2
62100	0
68800	2

Data might also have information on the predictor variables, to use in semi-parametric regression-like methods such as Cox proportional hazards regression.

Time Until Recovery (weeks)	Censoring	Gender	Systolic Blood Pressure	Diastolic Blood Pressure
12	1	Male	124	93
20	0	Female	109	77
7	0	Female	125	83
13	0	Male	117	75
9	1	Male	122	80
15	0	Female	121	70
17	1	Male	130	88

Time Until Recovery (weeks)	Censoring	Gender	Systolic Blood Pressure	Diastolic Blood Pressure
8	0	Female	115	82
14	0	Male	118	86

Survivor Function

The survivor function is the probability of survival as a function of time. It is also called the survival function. It gives the probability that the survival time of an individual exceeds a certain value. Since the cumulative distribution function, $F(t)$, is the probability that the survival time is less than or equal to a given point in time, the survival function for a continuous distribution, $S(t)$, is the complement of the cumulative distribution function:

$S(t) = 1 - F(t)$.

The survivor function is also related to the hazard function on page 14-6. If the data has the hazard function, $h(t)$, then the survivor function is

$$S(t) = \exp\left(-\int_0^t h(u)du\right),$$

which corresponds to

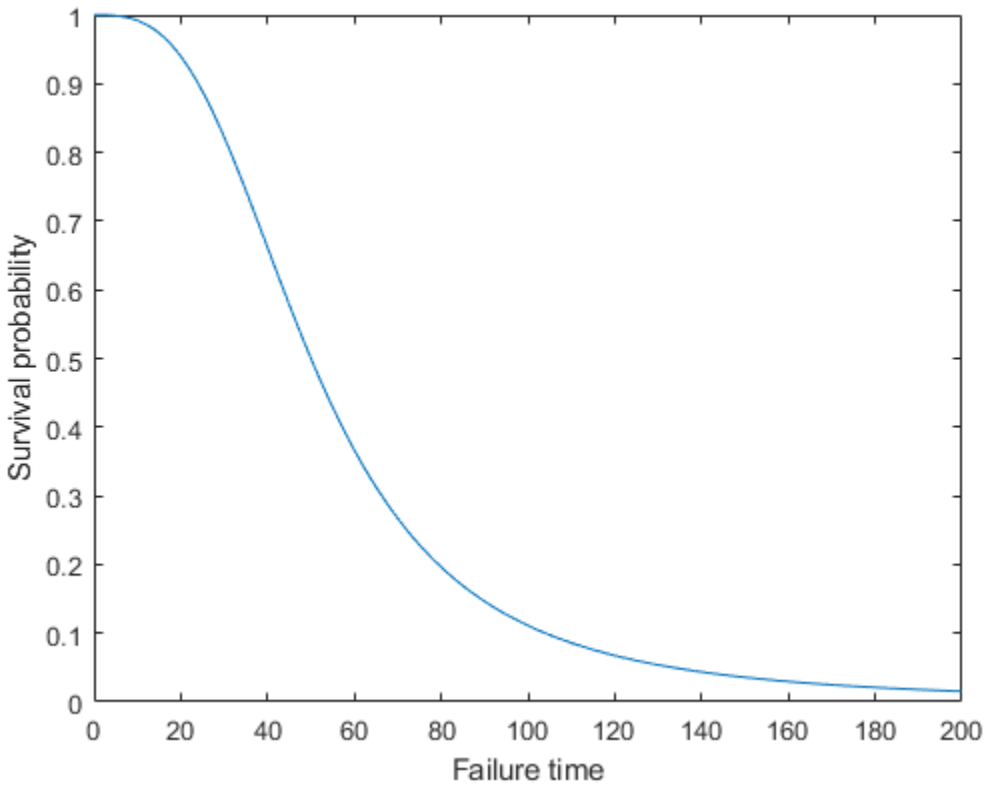
$$S(t) = \exp(-H(t)),$$

where $H(t)$ is the cumulative hazard function.

Burr Distribution Survivor Function

Calculate and plot the survivor function of a Burr distribution with parameters `50`, `3`, and `1`.

```
x = 0:0.1:200;
figure()
plot(x,1-cdf('Burr',x,50,3,1))
xlabel('Failure time');
ylabel('Survival probability');
```



Survivor Function from Data

This example shows how to estimate the survivor function from data.

Load the sample data.

```
load readmissiontimes
```

The column vector `ReadmissionTime` shows the readmission times for 100 patients. The column vector `Censored` has the censorship information for each patient, where 1 indicates censored data, and 0 that indicates the exact readmission times are observed. This data is simulated.

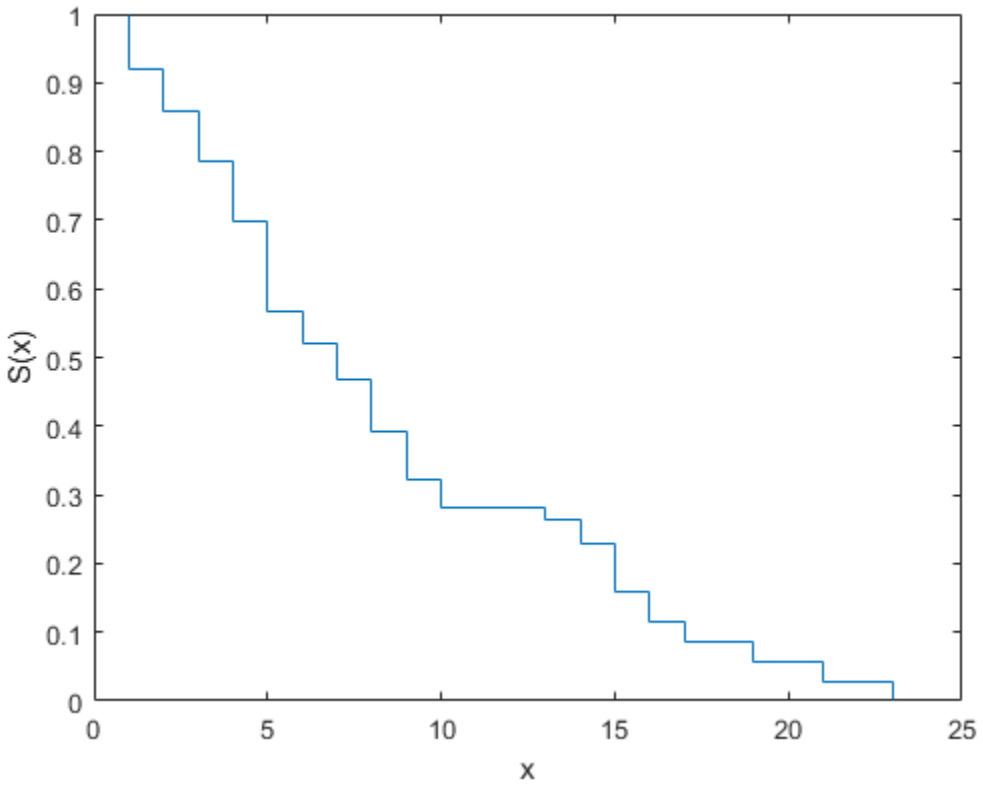
```
[ReadmissionTime Censored]
```

```
ans = 100×2

     5     1
     3     1
    19     0
    17     0
     9     0
    16     0
     4     0
     2     0
     3     0
    15     0
```

The first two readmission times, 5 and 3, are both censored.

Display the empirical survivor function with censoring using `ecdf` with the name-value pair arguments `'function','survivor'` and `'censoring',Censored`.

```
ecdf(ReadmissionTime,'censoring',Censored,'function','survivor')
```



Hazard Function

The hazard function gives the instantaneous failure rate of an individual conditioned on the fact that the individual survived until a given time. That is,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \,|\, T \ge t)}{\Delta t},$$

where Δt is a very small time interval. The hazard rate, therefore, is sometimes called the conditional failure rate. The hazard function always takes a positive value. However, these values do not correspond to probabilities and might be greater than 1.

The hazard function is related to the probability density function, $f(t)$, cumulative distribution function, $F(t)$, and survivor function, $S(t)$, as follows:

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)},$$

which is also equivalent to

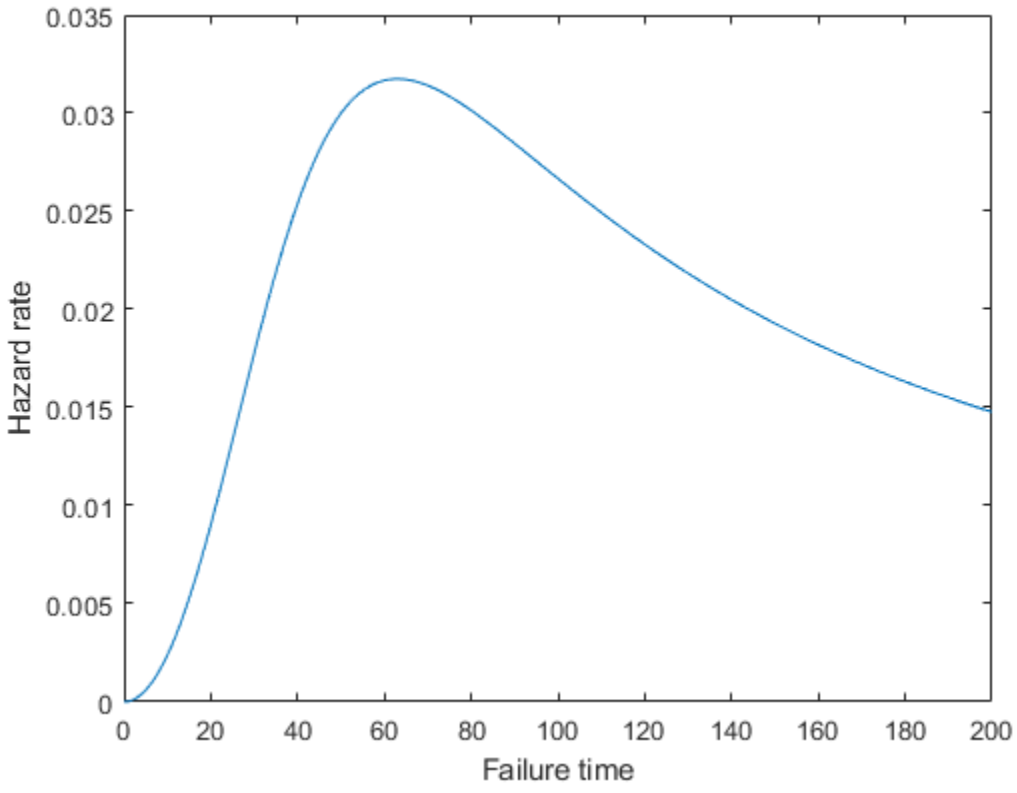
$$h(t) = -\frac{d}{dt}\ln S(t).$$

So, if you know the shape of the survival function, you can also derive the corresponding hazard function.

Burr Distribution Hazard Function

Calculate and plot the hazard function of a Burr distribution with parameters `50`, `3`, and `1`.

```
x = 0:1:200;
Burrhazard = pdf('Burr',x,50,3,1)./(1-cdf('Burr',x,50,3,1));
figure()
plot(x,Burrhazard)
xlabel('Failure time');
ylabel('Hazard rate');
```

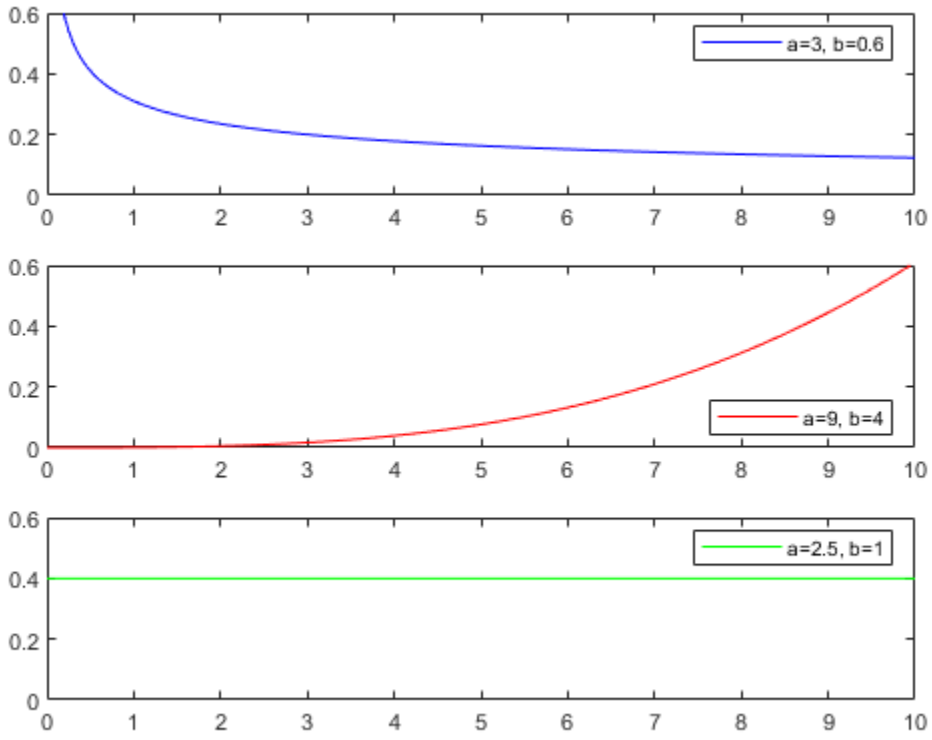


Weibull Hazard Functions

There are different types of hazard functions. The previous figure shows a situation when the hazard rate increases for the early time periods and then gradually decreases. The hazard rate might also be

monotonically decreasing, increasing, or constant over time. The following figure shows examples of different types of hazard functions for data coming from different Weibull distributions.

```
figure
ax1 = subplot(3,1,1);
x1 = 0:0.05:10;
hazard1 = pdf('wbl',x1,3,0.6)./(1-cdf('wbl',x1,3,0.6));
plot(x1,hazard1,'color','b')
set(ax1,'Ylim',[0 0.6]);
legend(ax1,'a=3, b=0.6');

ax2 = subplot(3,1,2);
x2 = 0:0.05:10;
hazard2 = pdf('wbl',x2,9,4)./(1-cdf('wbl',x2,9,4));
plot(x2,hazard2,'color','r')
set(ax2,'Ylim',[0 0.6]);
legend(ax2,'a=9, b=4','location','southeast');

ax3 = subplot(3,1,3);
x3 = 0:0.05:10;
hazard3 = pdf('wbl',x3,2.5,1)./(1-cdf('wbl',x3,2.5,1));
plot(x3,hazard3,'color','g')
set(ax3,'Ylim',[0 0.6]);
legend(ax3,'a=2.5, b=1');
```



In the third case, the Weibull distribution has a shape parameter value of 1, which corresponds to the exponential distribution. The exponential distribution always has a constant hazard rate over time.

References

[1] Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[3] Kleinbaum, D. G., and M. Klein. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2005.

See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model for Censored Data” on page 14-31
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “Kaplan-Meier Method” on page 14-10
- “Cox Proportional Hazards Model” on page 14-26

Kaplan-Meier Method

The Statistics and Machine Learning Toolbox function `ecdf` produces the empirical cumulative hazard, survivor, and cumulative distribution functions by using the Kaplan-Meier nonparametric method. The Kaplan-Meier estimator for the survivor function is also called the *product-limit estimator*.

The Kaplan-Meier method uses survival data summarized in life tables. Life tables order data according to ascending failure times, but you don't have to enter the failure/survival times in an ordered manner to use `ecdf`.

A life table usually consists of:

- Failure times
- Number of items failed at a time/time period
- Number of items censored at a time/time period
- Number of items at risk at the beginning of a time/time period

The number at risk is the total number of survivors at the beginning of each period. The number at risk at the beginning of the first period is all individuals in the lifetime study. At the beginning of each remaining period, the number at risk is reduced by the number of failures plus individuals censored at the end of the previous period.

This life table shows fictitious survival data. At the beginning of the first failure time, there are seven items at risk. At time 4, three fail. So at the beginning of time 7, there are four items at risk. Only one fails at time 7, so the number at risk at the beginning of time 11 is three. Two fail at time 11, so at the beginning of time 12, the number at risk is one. The remaining item fails at time 12.

Failure Time (t)	Number Failed	Number at Risk
4	3	7
7	1	4
11	2	3
12	1	1

You can estimate the hazard, cumulative hazard, survival, and cumulative distribution functions using the life tables as described next.

Cumulative Hazard Rate (Failure Rate)

The hazard rate at each period is the number of failures in the given period divided by the number of surviving individuals at the beginning of the period (number at risk).

Failure Time (t)	Hazard Rate ($h(t)$)	Cumulative Hazard Rate
0	0	0
t_1	d_1/r_1	d_1/r_1
t_2	d_2/r_2	$h(t_1) + d_2/r_2$
...	...	...
t_n	d_n/r_n	$h(t_{n-1}) + d_n/r_n$

Survival Probability

For each period, the survival probability is the product of the complement of hazard rates. The initial survival probability at the beginning of the first time period is 1. If the hazard rate for the each period is $h(t_i)$, then the survivor probability is as shown.

Time (*t*)	Survival Probability (*S*(*t*))
0	1
t_1	$1*(1 - h(t_1))$
t_2	$S(t_1)*(1 - h(t_2))$
...	...
t_n	$S(t_{n-1})*(1 - h(t_n))$

Cumulative Distribution Function

Because the cumulative distribution function (cdf) and the survivor function are complements of each other, you can find the cdf from the life tables using $F(t) = 1 - S(t)$.

You can compute the cumulative hazard rate, survival rate, and cumulative distribution function for the simulated data in the first table on this page as follows.

t	Number Failed (*d*)	Number at Risk (*r*)	Hazard Rate	Survival Probability	Cumulative Distribution Function
4	3	7	3/7	1 – 3/7 = 4/7 = 0.5714	0.4286
7	1	4	1/4	4/7*(1 – 1/4) = 3/7 = .4286	0.5714
11	2	3	2/3	3/7*(1 – 2/3) = 1/7 = 0.1429	0.8571
12	1	1	1/1	1/7*(1 – 1) = 0	1

This rates in this example are based on the discrete failure times, and hence the calculations do not necessarily follow the derivative-based definition in “What Is Survival Analysis?” on page 14-2

Here is how you can enter the data and calculate these measures using `ecdf`. The data does not necessarily have to be in ascending order. Suppose the failure times are stored in an array `y`.

```
y = [4 7 11 12];
freq = [3 1 2 1];
[f,x] = ecdf(y,'frequency',freq)
```

```
f =

         0
    0.4286
    0.5714
    0.8571
    1.0000
```

```
x =

     4
     4
     7
    11
    12
```

When you have censored data, the life table might look like the following:

Time (*t*)	Number failed (*d*)	Censoring	Number at Risk (*r*)	Hazard Rate	Survival Probability	Cumulative Distribution Function
4	2	1	7	2/7	1 – 2/7 = 0.7143	0.2857
7	1	0	4	1/4	0.7143*(1 – 1/4) = 0.5357	0.4643
11	1	1	3	2/3	0.5357*(1 – 1/3) = 0.3571	0.6429
12	1	0	1	1/1	0.3571*(1 – 1) = 0	1.0000

At any given time, the censored items are also considered in the total of number at risk, and the hazard rate formula is based on the number failed and the total number at risk. While updating the number at risk at the beginning of each period, the total number failed and censored in the previous period is reduced from the number at risk at the beginning of that period.

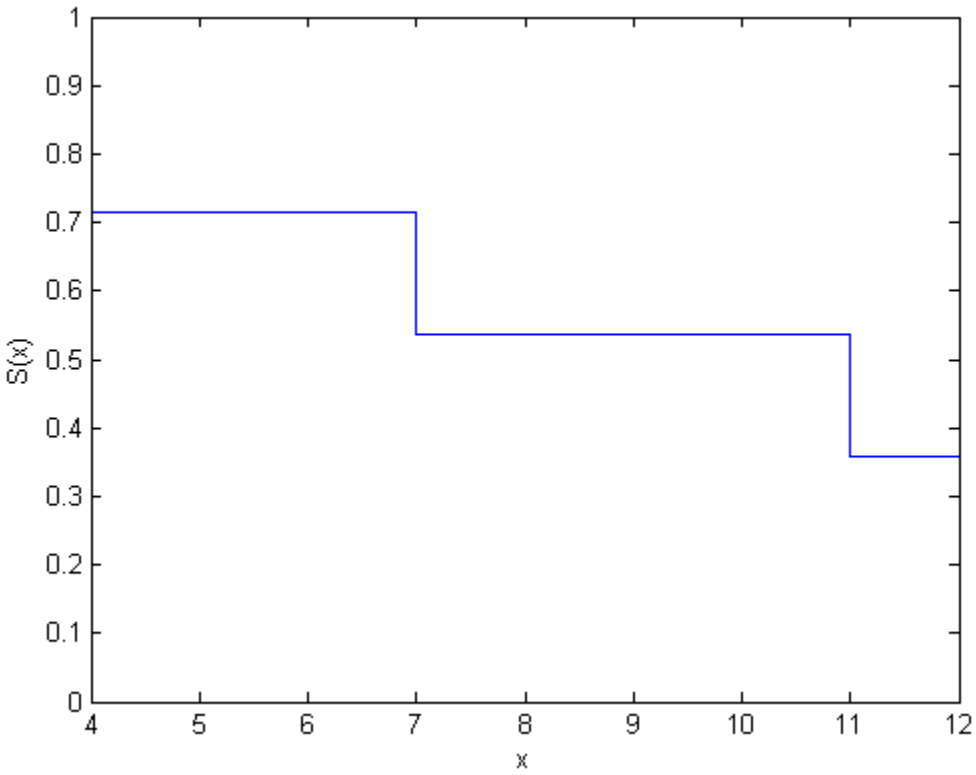
While using `ecdf`, you must also enter the censoring information using an array of binary variables. Enter 1 for censored data, and enter 0 for exact failure time.

```
y = [4 4 4 7 11 11 12];
cens = [0 1 0 0 1 0 0];
[f,x] = ecdf(y,'censoring',cens)
```

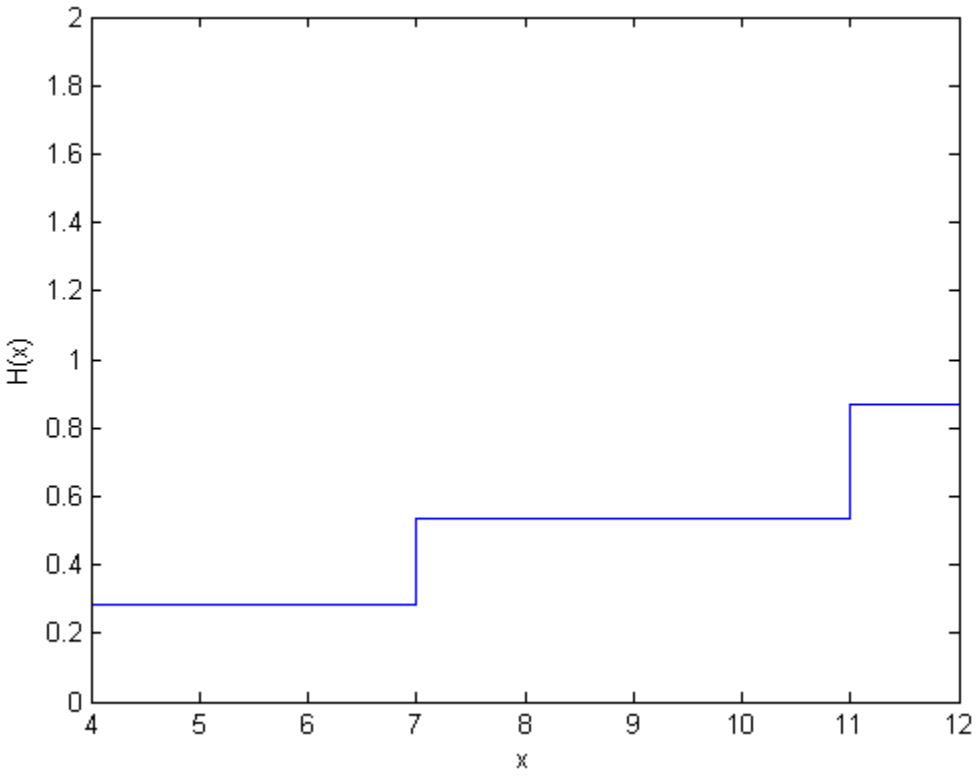
```
f =

         0
    0.2857
    0.4643
    0.6429
    1.0000

x =

     4
     4
     7
    11
    12
```

`ecdf`, by default, produces the cumulative distribution function values. You have to specify the survivor function or the hazard function using optional name-value pair arguments. You can also plot the results as follows.

```
figure()
ecdf(y,'censoring',cens,'function','survivor');
```




```
figure()
ecdf(y,'censoring',cens,'function','cumulative hazard');
```



References

[1] Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[3] Kleinbaum, D. G., and M. Klein. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2005.

See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model for Censored Data” on page 14-31
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “What Is Survival Analysis?” on page 14-2

- “Cox Proportional Hazards Model” on page 14-26

Hazard and Survivor Functions for Different Groups

This example shows how to estimate and plot the cumulative hazard and survivor functions for different groups.

Step 1. Load and organize sample data.

Load the sample data.

```
load('readmissiontimes.mat')
```

The data has readmission times of patients with information on their gender, age, weight, smoking status, and censorship. This is simulated data.

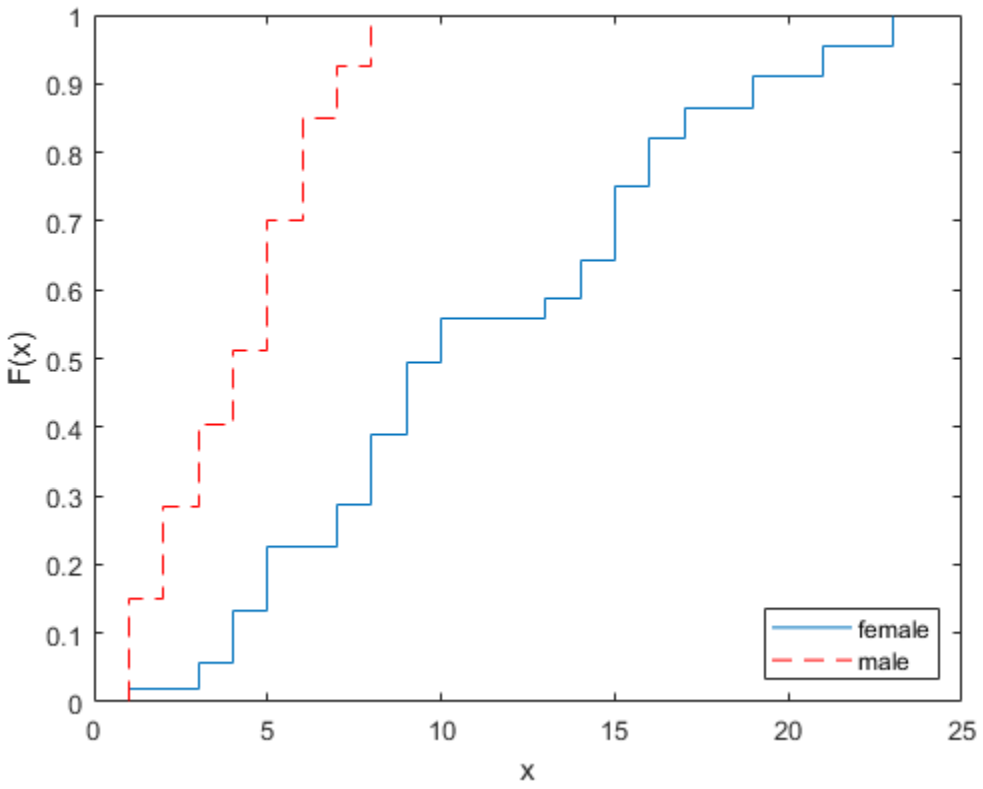
Create a matrix of readmission times and censoring for each gender.

```
female = [ReadmissionTime(Sex==1),Censored(Sex==1)];
male = [ReadmissionTime(Sex==0),Censored(Sex==0)];
```

Step 2. Estimate and plot cumulative distribution function for each gender.

Plot the Kaplan-Meier estimate of the cumulative distribution function for female and male patients.

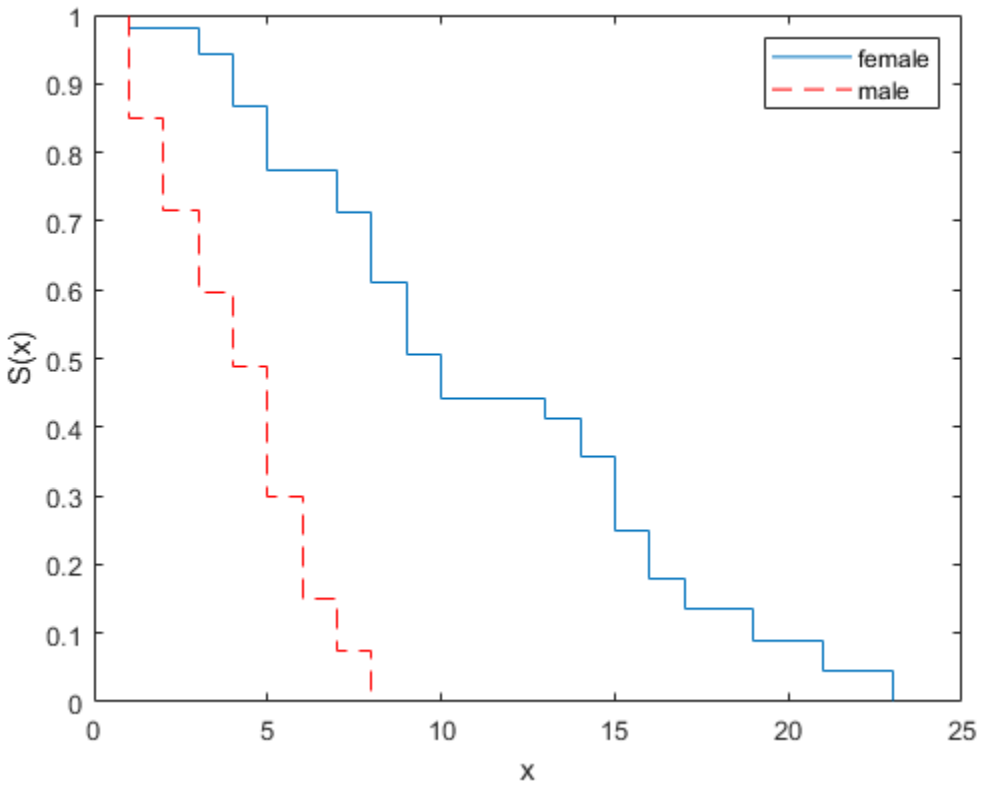
```
figure()
ecdf(gca,female(:,1),'Censoring',female(:,2));
hold on
[f,x] = ecdf(male(:,1),'Censoring',male(:,2));
stairs(x,f,'--r')
hold off
legend('female','male','Location','SouthEast')
```



Step 3. Plot survivor functions.

Compare the survivor functions for female and male patients.

```
figure()
ax1 = gca;
ecdf(ax1,female(:,1),'Censoring',female(:,2),'function','survivor');
hold on
[f,x] = ecdf(male(:,1),'Censoring',male(:,2),'function','survivor');
stairs(x,f,'--r')
legend('female','male')
```



This figure shows that readmission times are shorter for male patients than female patients.

Step 4. Fit Weibull survivor functions.

Fit Weibull distributions to readmission times of female and male patients.

```
pd = fitdist(female(:,1),'wbl','Censoring',female(:,2))
```

```
pd =

  WeibullDistribution

  Weibull distribution
    A = 12.5593   [10.749, 14.6745]
    B = 1.99834   [1.56489, 2.55185]
```

```
pd2 = fitdist(male(:,1),'wbl','Censoring',male(:,2))
```

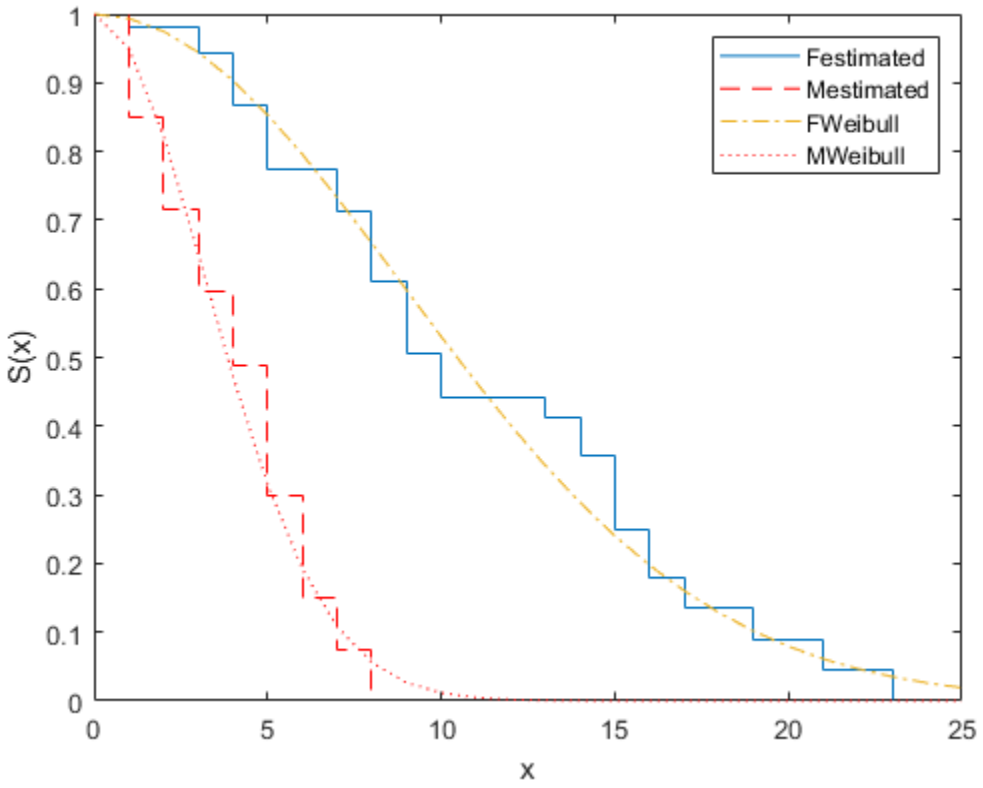
```
pd2 =

  WeibullDistribution

  Weibull distribution
    A = 4.63991   [3.91039, 5.50551]
    B = 1.94422   [1.48496, 2.54552]
```

```
pd2 = fitdist(male(:,1),'wbl','Censoring',male(:,2))
```

```
pd2 =

  WeibullDistribution

  Weibull distribution
    A = 4.63991   [3.91039, 5.50551]
    B = 1.94422   [1.48496, 2.54552]
```

Plot the Weibull survivor functions for female and male patients on estimated survivor functions.

```
plot(0:1:25,1-cdf('wbl',0:1:25,12.5593,1.99834),'-.')
plot(0:1:25,1-cdf('wbl',0:1:25,4.63991,1.94422),':r')
hold off
legend('Festimated','Mestimated','FWeibull','MWeibull')
```



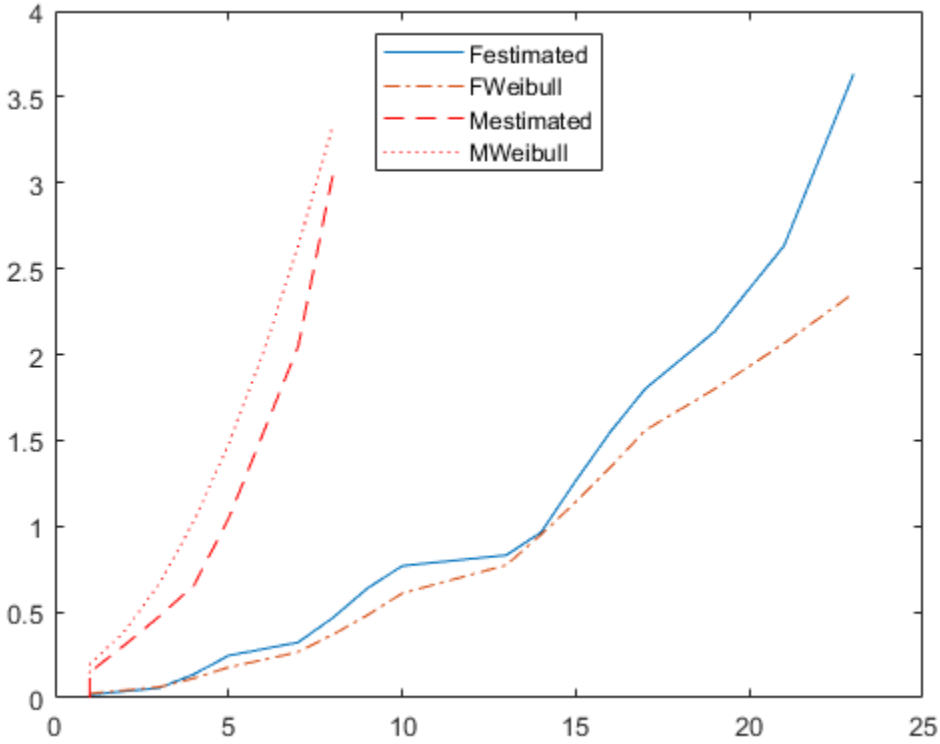
Weibull distribution provides a good fit for the data.

Step 5. Estimate cumulative hazard and fit Weibull cumulative hazard functions.

Estimate the cumulative hazard function for the genders and fit Weibull cumulative hazard functions.

```
figure()
[f,x] = ecdf(female(:,1),'Censoring',female(:,2),...
'function','cumhazard');
```

```
plot(x,f)
hold on
plot(x,cumsum(pdf(pd,x)./(1-cdf(pd,x))),'-.')
[f,x] = ecdf(male(:,1),'Censoring',male(:,2),...
'function','cumhazard');
plot(x,f,'--r')
plot(x,cumsum(pdf(pd2,x)./(1-cdf(pd2,x))),':r')
legend('Festimated','FWeibull','Mestimated','MWeibull',...
'Location','North')
```



See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model for Censored Data” on page 14-31
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “What Is Survival Analysis?” on page 14-2
- “Kaplan-Meier Method” on page 14-10

- “Cox Proportional Hazards Model” on page 14-26

Survivor Functions for Two Groups

This example shows how to find the empirical survivor functions and the parametric survivor functions using the Burr type XII distribution fit to data for two groups.

Step 1. Load and prepare sample data.

Load the sample data.

```
load('lightbulb.mat')
```

The first column of the data has the lifetime (in hours) of two types of light bulbs. The second column has information about the type of light bulb. 0 indicates fluorescent bulbs whereas 1 indicates the incandescent bulb. The third column has censoring information. 1 indicates censored data, and 0 indicates the exact failure time. This is simulated data.

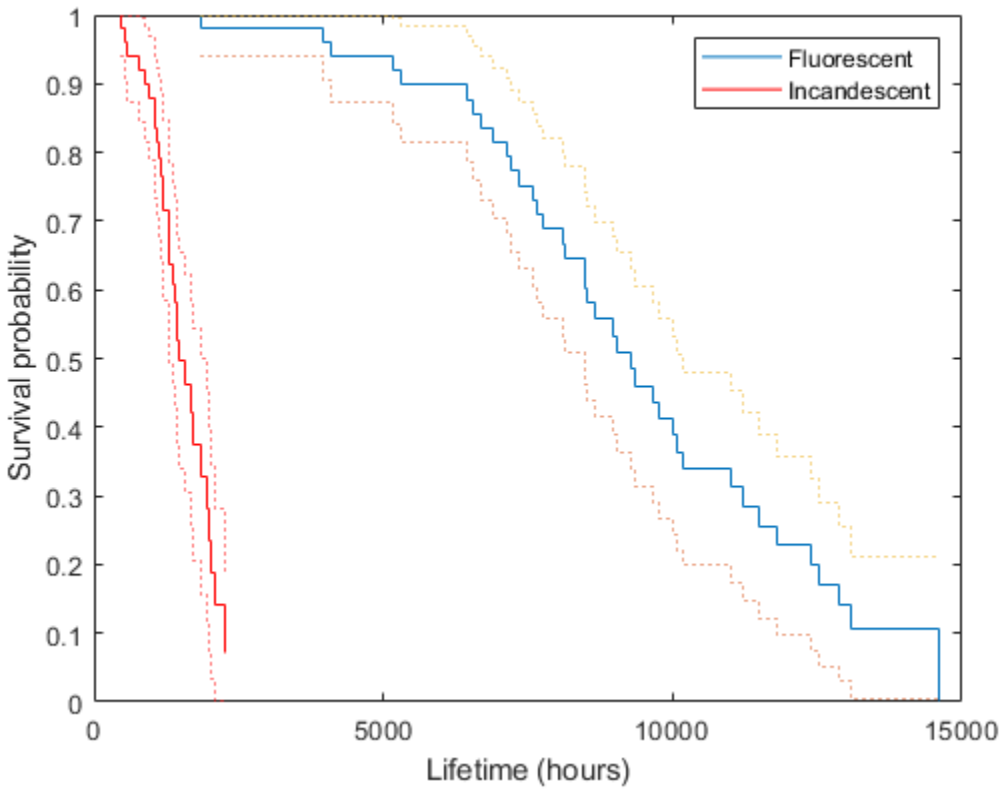
Create a variable for each light bulb type and also include the censorship information.

```
fluo = [lightbulb(lightbulb(:,2)==0,1),...
            lightbulb(lightbulb(:,2)==0,3)];
insc = [lightbulb(lightbulb(:,2)==1,1),...
            lightbulb(lightbulb(:,2)==1,3)];
```

Step 2. Plot estimated survivor functions.

Plot the estimated survivor functions for the two different types of light bulbs.

```
figure()
[f,x,flow,fup] = ecdf(fluo(:,1),'censoring',fluo(:,2),...
                 'function','survivor');
ax1 = stairs(x,f);
hold on
stairs(x,flow,':')
stairs(x,fup,':')
[f,x,flow,fup] = ecdf(insc(:,1),'censoring',insc(:,2),...
                 'function','survivor');
ax2 = stairs(x,f,'color','r');
stairs(x,flow,':r')
stairs(x,fup,':r')
legend([ax1,ax2],{'Fluorescent','Incandescent'})
xlabel('Lifetime (hours)')
ylabel('Survival probability')
```



You can see that the survival probability of incandescent light bulbs is much smaller than that of fluorescent light bulbs.

Step 3. Fit Burr Type XII distribution.

Fit Burr distribution to the lifetime data of fluorescent and incandescent type bulbs.

```
pd = fitdist(fluo(:,1),'burr','Censoring',fluo(:,2))
```

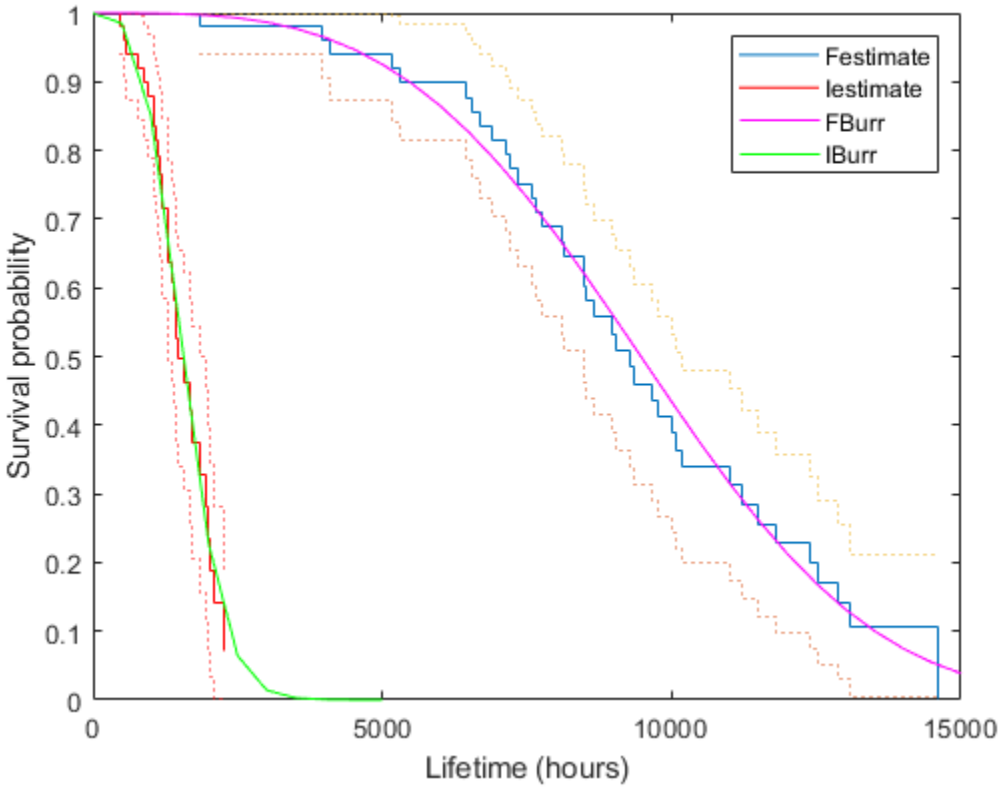
```
pd =
  BurrDistribution

  Burr distribution
    alpha = 29143.4   [0.903922, 9.39617e+08]
        c = 3.44582   [2.13013, 5.57417]
        k = 33.7039   [8.10737e-14, 1.40114e+16]
```

```
pd2 = fitdist(insc(:,1),'burr','Censoring',insc(:,2))
```

```
pd2 =
  BurrDistribution

  Burr distribution
    alpha = 2650.76   [430.773, 16311.4]
        c = 3.41898   [2.16794, 5.39197]
        k =  4.5891   [0.0307809, 684.185]
```

Superimpose Burr type XII survivor functions.

```
ax3 = plot(0:500:15000,1-cdf('burr',0:500:15000,29143.5,...
           3.44582,33.704),'m');
ax4 = plot(0:500:5000,1-cdf('burr',0:500:5000,2650.76,...
           3.41898,4.5891),'g');
legend([ax1;ax2;ax3;ax4],'Festimate','Iestimate','FBurr','IBurr')
```



Burr distribution provides a good fit for the lifetime of light bulbs in this example.

Step 4. Fit a Cox proportional hazards model.

Fit a Cox proportional hazards regression where the type of the bulb is the explanatory variable.

```
[b,logl,H,stats] = coxphfit(lightbulb(:,2),lightbulb(:,1),...
'Censoring',lightbulb(:,3));
stats
```

```
stats = struct with fields:
                 covb: 1.0757
                 beta: 4.7262
                   se: 1.0372
                    z: 4.5568
                    p: 5.1936e-06
               csres: [100x1 double]
              devres: [100x1 double]
             martres: [100x1 double]
              schres: [100x1 double]
```

```
              sschres: [100x1 double]
               scores: [100x1 double]
              sscores: [100x1 double]
    LikelihoodRatioTestP: 0
```

The *p*-value, `p`, indicates that the type of light bulb is statistically significant. The estimate of the hazard ratio is $exp(b) = 112.8646$. This means that the hazard for the incandescent bulbs is 112.86 times the hazard for the fluorescent bulbs.

See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Cox Proportional Hazards Model for Censored Data” on page 14-31
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “What Is Survival Analysis?” on page 14-2
- “Kaplan-Meier Method” on page 14-10
- “Cox Proportional Hazards Model” on page 14-26

Cox Proportional Hazards Model

In this section...

Introduction

Cox proportional hazards regression is a semiparametric method for adjusting survival rate estimates to quantify the effect of predictor variables. The method represents the effects of explanatory variables as a multiplier of a common baseline hazard function, $h_0(t)$. The hazard function is the nonparametric part of the Cox proportional hazards regression function, whereas the impact of the predictor variables is a loglinear regression. For a baseline relative to 0, this model corresponds to

$$h(X_i, t) = h_0(t)\exp\left[\sum_{j=1}^{p} x_{ij}b_j\right],$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is the predictor variable for the ith subject, $h(X_i,t)$ is the hazard rate at time t for X_i, and $h_0(t)$ is the baseline hazard rate function.

Hazard Ratio

The Cox proportional hazards model relates the hazard rate for individuals or items at the value X_i, to the hazard rate for individuals or items at the baseline value. It produces an estimate for the hazard ratio:

$$HR(X_i) = \frac{h(X_i, t)}{h_0(t)} = \exp\left[\sum_{j=1}^{p} x_{ij}b_j\right].$$

The model is based on the assumption that the baseline hazard function depends on time, t, but the predictor variables do not. This assumption is also called the proportional hazards assumption, which states that the hazard ratio does not change over time for any individual.

The hazard ratio represents the relative risk of instant failure for individuals or items having the predictive variable value X_i compared to the ones having the baseline values. For example, if the predictive variable is smoking status, where nonsmoking is the baseline category, the hazard ratio shows the relative instant failure rate of smokers compared to the baseline category, that is, nonsmokers. For a baseline relative to X^* and the predictor variable value X_i, the hazard ratio is

$$HR(X_i) = \frac{h(X_i, t)}{h(X^*, t)} = \exp\left[\sum_{j=1}^{p} \left(x_{ij} - x_j^*\right)b_j\right].$$

For example, if the baseline is the mean values of the predictor variables (`mean(X)`), then the hazard ratio becomes

$$HR(X_i) = \frac{h(X_i, t)}{h(\bar{X}, t)} = \exp\left[\sum_{j=1}^{p} (x_{ij} - \bar{x}_j) b_j\right].$$

Hazard rates are related to survival rates, such that the survival rate at time t for an individual with the explanatory variable value X_i is

$$S_{X_i}(t) = S_0(t)^{HR(X_i)},$$

where $S_0(t)$ is the survivor function with the baseline hazard rate function $h_0(t)$, and $HR(X_i)$ is the hazard ratio of the predictor variable value X_i relative to the baseline value.

Extension of Cox Proportional Hazards Model

When you have variables that do not satisfy the proportional hazards (PH) assumption, you can consider using two extensions of Cox proportional hazards model: the stratified Cox model and the Cox model with time-dependent variables.

If the variables that do not satisfy the PH assumption are categorizable, use the stratified Cox model:

$$h_s(X_i, t) = h_{0s}(t)\exp\left[\sum_{j=1}^{p} x_{ij} b_j\right],$$

where the subscript s indicates the sth stratum. The stratified Cox model has a different baseline hazard rate function for each stratum but shares coefficients. Therefore, it has the same hazard ratio across all strata if the predictor variable values are the same. You can include stratification variables in `coxphfit` by using the name-value pair `'Strata'`.

If the variables that do not satisfy the PH assumption are time-dependent variables, use the Cox model with time-dependent variables:

$$h(X_i, t) = h_0(t)\exp\left[\sum_{j=1}^{p_1} x_{ij} b_j + \sum_{k=1}^{p_2} x_{ik}(t) c_k\right],$$

where x_{ij} is an element of a time-independent predictor and $x_{ik}(t)$ is an element of a time-dependent predictor. For an example of how to include time-dependent variables in `coxphfit`, see “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35.

Partial Likelihood Function

A point estimate of the effect of each explanatory variable, that is, the estimated hazard ratio for the effect of each explanatory variable is $\exp(b)$, given all other variables are held constant, where b is the coefficient estimate for that variable. The coefficient estimates are found by maximizing the partial likelihood function of the model. The partial likelihood function for the proportional hazards regression model is based on the observed order of events. It is the product of partial likelihoods of failures estimated for each failure time. If there are n failures at n distinct failure times, $t_1 < t_2 < \cdots < t_n$, then the partial likelihood is

$$L = \frac{HR(X_1)}{\sum_{j=1}^{n} HR(X_j)} \times \frac{HR(X_2)}{\sum_{j=2}^{n} HR(X_j)} \times \cdots \times \frac{HR(X_n)}{HR(X_n)} = \prod_{i=1}^{n} \frac{HR(X_i)}{\sum_{j=i}^{n} HR(X_j)}.$$

You can rewrite the partial likelihood by using a risk set R_i:

$$L = \prod_{i=1}^{n} \frac{HR(X_i)}{\sum_{j \in R_i} HR(X_j)},$$

where R_i represents the index set of subjects who are under study but do not experience the event until the ith failure time.

You can use a likelihood ratio test to assess the significance of adding a term or terms in a model. Consider the two models where the first model has p predictive variables and the second model has $p + r$ predictive variables. Then, comparing the two models, $-2*(L_1/L_2)$ has a chi-square distribution with r degrees of freedom (the number of terms being tested).

Partial Likelihood Function for Tied Events

When you have tied events, `coxphfit` approximates the partial likelihood of the model by either Breslow's method (default) or Efron's method, instead of computing the exact partial likelihood. Computing the exact partial likelihood requires a large amount of computation, which involves an entire permutation of the risk sets for the tied event times.

The simplest approximation method is Breslow's method. This method uses the same denominator for each tied set.

$$L = \prod_{i=1}^{d} \prod_{j \in D_i} \frac{HR(X_j)}{\sum_{k \in R_i} HR(X_k)},$$

where d is the number of distinct event times, and D_i is the index set of all subjects whose event time is equal to the ith event time.

Efron's method is more accurate than Breslow's method, yet simple. This method adjusts the denominator of the tied events as follows:

$$L = \prod_{i=1}^{d} \prod_{j \in D_i} \frac{HR(X_j)}{\sum_{k \in R_i} HR(X_k) - \frac{j-1}{d_i} \sum_{k \in D_i} HR(X_k)},$$

where d_i is the number of indexes in D_i.

For an example, assume that the first two events are tied, that is, $t_1 = t_2$ and $t_2 < t_3 < \cdots < t_n$. In Breslow's method, the denominators of the first two terms are the same:

$$L = \frac{HR(X_1)}{\sum_{j=1}^{n} HR(X_j)} \times \frac{HR(X_2)}{\sum_{j=1}^{n} HR(X_j)} \times \frac{HR(X_3)}{\sum_{j=3}^{n} HR(X_j)} \times \frac{HR(X_4)}{\sum_{j=4}^{n} HR(X_j)} \times \cdots \times \frac{HR(X_n)}{HR(X_n)}.$$

Efron's method adjusts the denominator of the second term:

$$L = \frac{HR(X_1)}{\sum_{j=1}^{n} HR(X_j)} \times \frac{HR(X_2)}{0.5HR(X_1) + 0.5HR(X_2) + \sum_{j=3}^{n} HR(X_j)} \times \frac{HR(X_3)}{\sum_{j=3}^{n} HR(X_j)} \times \frac{HR(X_4)}{\sum_{j=4}^{n} HR(X_j)} \times \cdots \times \frac{HR(X_n, t_n)}{HR(X_n, t_n)}.$$

You can specify an approximation method by using the name-value pair `'Ties'` in `coxphfit`.

Frequency or Weights of Observations

The Cox proportional hazards model can incorporate with the frequency or weights of observations. Let w_i be the weight of the ith observation. Then, the partial likelihoods of the Cox model with weights become as follows:

- Partial likelihood with weights

$$L = \prod_{i=1}^{n} \frac{HR_w(X_i)}{\sum_{j \in R_i} w_j HR(X_j)},$$

  where

$$HR_w(X_i) = \exp\left[\sum_{j=1}^{p} w_j x_{ij} b_j\right].$$

- Partial likelihood with weights and Breslow's method

$$L = \prod_{i=1}^{d} \prod_{j \in D_i} \frac{HR_w(X_j)}{\left[\sum_{k \in R_i} w_k HR(X_k)\right]^{\frac{1}{d_i}\sum_{j \in D_i} w_j}}$$

- Partial likelihood with weights and Efron's method

$$L = \prod_{i=1}^{d} \prod_{j \in D_i} \frac{HR_w(X_j)}{\left[\sum_{k \in R_i} w_k HR(X_k) - \frac{j-1}{d_i}\sum_{k \in D_i} w_k HR(X_k)\right]^{\frac{1}{d_i}\sum_{j \in D_i} w_j}}$$

You can specify the frequency or weights of observations by using the name-value pair `'Frequency'` in `coxphfit`.

References

[1] Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[3] Kleinbaum, D. G., and M. Klein. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2005.

[4] Klein, J. P., and M. L. Moeschberger. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2003.

See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model for Censored Data” on page 14-31
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “What Is Survival Analysis?” on page 14-2
- “Kaplan-Meier Method” on page 14-10

Cox Proportional Hazards Model for Censored Data

This example shows how to construct a Cox proportional hazards model, and assess the significance of the predictor variables.

Step 1. Load sample data.

Load the sample data.

```
load readmissiontimes
```

The response variable is `ReadmissionTime`, which shows the readmission times for 100 patients. The predictor variables are `Age`, `Sex`, `Weight`, and the smoking status of each patient, `Smoker`. 1 indicates the patient is a smoker, and 0 indicates that the patient does not smoke. The column vector `Censored` has the censorship information for each patient, where 1 indicates censored data, and 0 indicates the exact readmission times are observed. This is simulated data.

Step 2. Fit Cox proportional hazards function.

Fit a Cox proportional hazard function with the variable Sex as the predictor variable, taking the censoring into account.

```
X = Sex;
[b,logl,H,stats] = coxphfit(X,ReadmissionTime,'censoring',Censored);
```

Assess the statistical significance of the term Sex.

```
stats
```

```
stats = struct with fields:
                    covb: 0.1016
                    beta: -1.7642
                      se: 0.3188
                       z: -5.5335
                       p: 3.1392e-08
                   csres: [100x1 double]
                  devres: [100x1 double]
                 martres: [100x1 double]
                  schres: [100x1 double]
                 sschres: [100x1 double]
                  scores: [100x1 double]
                 sscores: [100x1 double]
    LikelihoodRatioTestP: 5.9825e-09
```

The *p*-value, `p`, indicates that the term Sex is statistically significant.

Save the loglikelihood value with a different name. You will use this to assess the significance of the extended models.

```
loglSex = logl
```

```
loglSex = -262.1365
```

Step 3. Add Age and Weight to the model.

Fit a Cox proportional hazards model with the variables Sex, Age, and Weight.

```
X = [Sex Age Weight];
[b,logl,H,stats] = coxphfit(X,ReadmissionTime,'censoring',Censored);
```

Assess the significance of the terms.

```
stats.beta
```

```
ans = 3×1

   -0.5441
    0.0143
    0.0250
```

```
stats.p
```

```
ans = 3×1

    0.4953
    0.3842
    0.0960
```

None of the terms, adjusted for others, is statistically significant.

Assess the significance of the terms using the log likelihood ratio. You can assess the significance of the new model using the likelihood ratio statistic. First find the difference between the log-likelihood statistic of the model without the terms Age and Weight and the log-likelihood of the model with Sex, Age, and Weight.

```
-2*[loglSex - logl]
```

```
ans = 3.6705
```

Now, compute the *p*-value for the likelihood ratio statistic. The likelihood ratio statistic has a Chi-square distribution with a degrees of freedom equal to the number of predictor variables being assessed. In this case, the degrees of freedom is 2.

```
p = 1 - cdf('chi2',3.6705,2)
```

```
p = 0.1596
```

The *p*-value of 0.1596 indicates that the terms Age and Weight are not statistically significant, given the term Sex in the model.

Step 4. Add Smoker to the model.

Fit a Cox proportional hazards model with the variables Sex and Smoker.

```
X = [Sex Smoker];
[b,logl,H,stats] = coxphfit(X,ReadmissionTime,...
'censoring',Censored);
```

Assess the significance of the terms in the model.

```
stats.p
```

```
ans = 2×1
```

```
    0.0000
    0.0148
```

Compare this model to the first model where Sex is the only term.

```
 -2*[loglSex - logl]
```

```
ans = 5.5789
```

Compute the *p*-value for the likelihood ratio statistic. The likelihood ratio statistic has a Chi-square distribution with a degree of freedom of 1.

```
p = 1 - cdf('chi2',5.5789,1)
```

```
p = 0.0182
```

The *p*-value of 0.0182 indicates that Sex and Smoker are statistically significant given the other is in the model. The model with Sex and Smoker is a better fit compared to the model with only Sex.

Request the coefficient estimates.

```
 stats.beta
```

```
ans = 2×1

   -1.7165
    0.6338
```

The default baseline is the mean of X, so the final model for the hazard ratio is

$$HR = \frac{h_X(t)}{h_{\bar{X}}(t)} = \exp\left[\beta_s\left(X_s - \bar{X}_s\right) + \beta_\alpha\left(X_\alpha - \bar{X}_\alpha\right)\right].$$

Fit a Cox proportional hazards model with a baseline of 0.

```
X = [Sex Smoker];
[b,logl,H,stats] = coxphfit(X,ReadmissionTime,...
'censoring',Censored,'baseline',0);
```

The model for the hazard ratio is

$$HR = \frac{h_X(t)}{h_0(t)} = \exp\left[\beta_s X_s + \beta_\alpha X_\alpha\right].$$

Request the coefficient estimates.

```
 stats.beta
```

```
ans = 2×1

   -1.7165
    0.6338
```

The coefficients are not affected, but the hazard rate differs from when the baseline is the mean of X.

See Also

`coxphfit` | `ecdf` | `ksdensity`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35

More About

- “What Is Survival Analysis?” on page 14-2
- “Kaplan-Meier Method” on page 14-10
- “Cox Proportional Hazards Model” on page 14-26

Cox Proportional Hazards Model with Time-Dependent Covariates

This example shows how to convert survival data to counting process form and then construct a Cox proportional hazards model with time-dependent covariates.

Step 1. Compare standard layout and counting process form.

A Cox model with time-dependent covariates requires survival data to be in counting process form and not in standard layout. To see the difference between survival data in standard layout and in counting process form, load the following sample data.

```
load simplesurvivaldata
```

This sample data contains two tables: `relapseS` and `relapseCP`. These two tables represent the same simple survival data in standard layout and in counting process form, respectively.

Display the data in standard layout.

```
relapseS
```

```
relapseS=2×5 table
    ID    Time    Censoring    Age    StopTreatment
    __    ____    _________    ___    _____________

    1      5          0        20          NaN
    2     20          1        30           12
```

This data represents two patients whose treatment status changes over time. Patient 1 was not taking treatment for the interval from week 0 to 5 and relapsed at the end of the interval. Patient 2 was taking treatment for the interval from week 0 to 12, but not for the interval from week 12 to 20. Patient 2 did not relapse and left the study after week 20.

Now display the same data in counting process form.

```
relapseCP
```

```
relapseCP=3×6 table
    ID    tStart    tStop    Censoring    Age    TreatmentStatus
    __    ______    _____    _________    ___    _______________

    1       0         5          0        20            0
    2       0        12          1        30            1
    2      12        20          1        30            0
```

In counting process form, each row represents the risk interval `(tStart,tStop]` instead of a single value of an event time. `Censoring` is 0 if the event is observed at the end of the risk interval, and 1 if it is not. `TreatmentStatus` corresponds to a time-dependent covariate, which represents the same information with `StopTreatment` in standard layout. Note that a Cox model assumes time-dependent covariates to be constant in each risk interval.

Step 2. Load sample data.

Next, load sample data to convert.

```
load survivaldatacp
```

This sample data contains a table `labS`, which is simulated survival data including repeated measurement for each patient in standard layout.

Display the simulated survival data in standard layout.

```
labS
```

```
labS=6×7 table
    ID    Time    Censoring    Sex    Lab_0    Lab_50    Lab_100
    __    ____    _________    ___    _____    ______    _______

    1      46         0         1      0.3       NaN       NaN
    2     138         1         0      0.2      0.23      0.39
    3      94         0         1     0.18      0.22       NaN
    4      50         0         0     0.21       0.2       NaN
    5     106         0         0     0.25      0.21      0.42
    6      98         0         0     0.21      0.22       NaN
```

In standard layout, each row of the table shows information for one patient.

- `ID` indicates the ID of a patient. You do not include `ID` as an input of a Cox model. Include `ID` in a data set to confirm that the data set is correctly converted to counting process form.
- `Time` represents time to event in days, which corresponds to a response variable.
- `Censoring` has the censorship information for each patient, where 1 indicates censored data and 0 indicates that the exact time to event is observed at the end of the observation period.
- `Sex` is a time-independent predictor where 1 indicates female, and 0 indicates male.
- `Lab_0`, `Lab_50`, and `Lab_100` represent three consecutive laboratory results measured at day 0, 50, and 100, which correspond to a time-dependent predictor.

Step 3. Convert survival data to counting process form.

To convert the survival data `labS` to counting process form, execute the code below. This code converts `Time` to a risk interval `(tStart,tStop]` and combines three vectors of the time-dependent predictor, `Lab_0`, `Lab_50`, and `Lab_100`, into one vector, `Lab`.

```
mTime = [0 50 100]; % Measurement time
threeLabs = [labS.Lab_0 labS.Lab_50 labS.Lab_100];
nLabMeasure = sum(sum(~isnan(threeLabs))); % Number of lab measurements
data = zeros(nLabMeasure,6); % One row for each observation
oID = 0; % Observation ID
for i = 1 : size(labS,1)
    idx = find(mTime <= labS.Time(i));
    for j = 1 : length(idx)-1
        oID = oID + 1;
        data(oID,:) = [labS.ID(i) mTime(j:j+1) 1 labS.Sex(i) threeLabs(i,j)];
    end
    oID = oID + 1;
    data(oID,:) = [labS.ID(i) mTime(length(idx)) labS.Time(i) ...
            labS.Censoring(i) labS.Sex(i) threeLabs(i,length(idx))];
end
labCP = table(data(:,1),data(:,2),data(:,3),data(:,4),data(:,5),data(:,6), ...
    'VariableNames', {'ID','tStart','tStop','Censoring','Sex','Lab'});
```

Display the survival data in counting process form.

```
labCP
```

```
labCP=13×6 table
    ID    tStart    tStop    Censoring    Sex    Lab
    __    ______    _____    _________    ___    ____

    1        0        46          0         1      0.3
    2        0        50          1         0      0.2
    2       50       100          1         0     0.23
    2      100       138          1         0     0.39
    3        0        50          1         1     0.18
    3       50        94          0         1     0.22
    4        0        50          1         0     0.21
    4       50        50          0         0      0.2
    5        0        50          1         0     0.25
    5       50       100          1         0     0.21
    5      100       106          0         0     0.42
    6        0        50          1         0     0.21
    6       50        98          0         0     0.22
```

In counting process form, each row of table `labCP` shows information of one observation corresponding to one risk interval. Note that a Cox model assumes `Lab` to be constant in the risk interval `(tStart,tStop]`. The value in `Censoring` is 0 if an event is observed at the end of the risk interval, and 1 if an event is not observed.

For example, patient 3 has two laboratory measurements at day 0 and 50, so there are two rows of data for patient 3 in counting process form. A Cox model assumes the lab results 0.18 and 0.22 to be constant in the interval (0,50] and (50,94], respectively. `Censoring` is 1 in (0,50] and 0 in (50,94] because the exact event time of patient 3 is observed at day 94.

Step 4. Adjust zero-length risk interval.

Find a patient who has a zero-length risk interval.

```
idxInvalid = labCP.ID(find(labCP.tStart == labCP.tStop))
```

```
idxInvalid = 4
```

Review the data for patient 4.

```
labCP(find(labCP.ID==idxInvalid),:)
```

```
ans=2×6 table
    ID    tStart    tStop    Censoring    Sex    Lab
    __    ______    _____    _________    ___    ____

    4        0        50          1         0     0.21
    4       50        50          0         0      0.2
```

The time to event of patient 4 coincides with the measurement day 50. However, (50,50] is an invalid risk interval for a Cox model because the model does not accept a zero length interval. Adjust the risk interval to be valid. You can choose any value less than the time unit as an adjustment amount. The choice of an adjustment amount is arbitrary, and it does not change the result.

```
idxAdjust = find(labCP.ID==idxInvalid);
labCP.tStop(idxAdjust(1)) = labCP.tStop(idxAdjust(1))-0.5;
labCP.tStart(idxAdjust(2)) = labCP.tStart(idxAdjust(2))-0.5;
labCP(idxAdjust,:)
```

```
ans=2×6 table
    ID    tStart    tStop    Censoring    Sex    Lab
    __    ______    _____    _________    ___    ____

    4          0     49.5          1         0     0.21
    4       49.5       50          0         0      0.2
```

Step 5. Construct a Cox proportional hazards model.

Fit a Cox proportional hazards model with the time-independent variable `Sex` and time-dependent variable `Lab`.

```
X = [labCP.Sex labCP.Lab];
T = [labCP.tStart labCP.tStop];
b = coxphfit(X,T,'Censoring',labCP.Censoring,'Baseline',0)
```

```
b = 2×1

    2.0054
   29.7530
```

For details on how to assess a Cox proportional hazards model, see “Cox Proportional Hazards Model for Censored Data” on page 14-31.

See Also

`coxphfit`

Related Examples

- “Hazard and Survivor Functions for Different Groups” on page 14-16
- “Survivor Functions for Two Groups” on page 14-22
- “Cox Proportional Hazards Model for Censored Data” on page 14-31

More About

- “What Is Survival Analysis?” on page 14-2
- “Kaplan-Meier Method” on page 14-10
- “Cox Proportional Hazards Model” on page 14-26

Cox Proportional Hazards Model Object

This example shows how to fit and analyze a Cox proportional hazards model using a `CoxModel` object. Cox proportional hazards models relate to lifetime or failure time data. For background, see “What Is Survival Analysis?” on page 14-2 and “Cox Proportional Hazards Model” on page 14-26.

Generate the data for three lifetime models with the following types of hazard rates. These models correspond to three stratification levels; see “Extension of Cox Proportional Hazards Model” on page 14-27.

- Bathtub, whose failure rate is high at the beginning, decreases to a low level, then climbs toward a constant level
- Logarithmically increasing: log(x)/10
- Constant 1/4

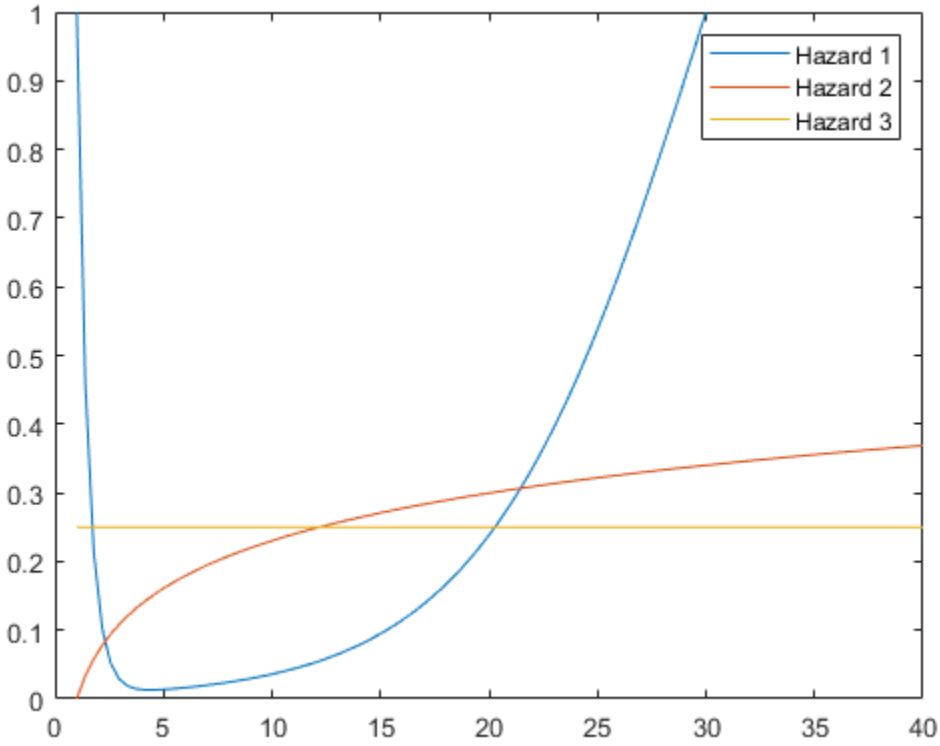
The three models share a predictor with three multipliers for the base hazard rates:

- `1`
- `1/20`
- `1/100`

In total, the data has nine types of population members, one from each stratification level and one from each predictor level. The functions for creating the bathtub and logarithmic hazard rates are in the Helper Functions on page 14-0 section at the end of this example.

Plot the three hazard rates when the predictor value is `1`.

```
t = linspace(1,40);
plot(t,hazard(t),t,log(t)/10,t,1/4*ones(size(t)))
legend('Hazard 1','Hazard 2','Hazard 3','Location','northeast')
ylim([0 1])
```



Create Data for Fitting

Create pseudorandom data for the lifetimes associated with the nine models. Create 9000 samples chosen randomly (about 1000 of each type) by inverting the cumulative distributions. (For details, see Inverse transform sampling).

```
N = 9000;
rng default
mults = [1;1/20;1/100]; % Three predictors
strata = randi(3,N,1); % Three strata
t1 = zeros(N,1);
a0 = randi(3,N,1); % Predictor
a1 = mults(a0);
v1 = rand(N,1);
for i = 1:N
    switch strata(i)
        case 1 % Bathtub
            t1(i) = zeropt(a1(i),v1(i));
        case 2 % Logarithmic
            t1(i) = zeroptold(a1(i),v1(i));
        case 3 % Constant
            t1(i) = 1 + exprnd(4/a1(i));
    end
end
```

Place data into a table.

```
a3 = categorical(a1);
tbldata = table(t1,a3,strata,'VariableNames',["Lifetime" "Predictors" "Strata"]);
```

Fit Cox Model

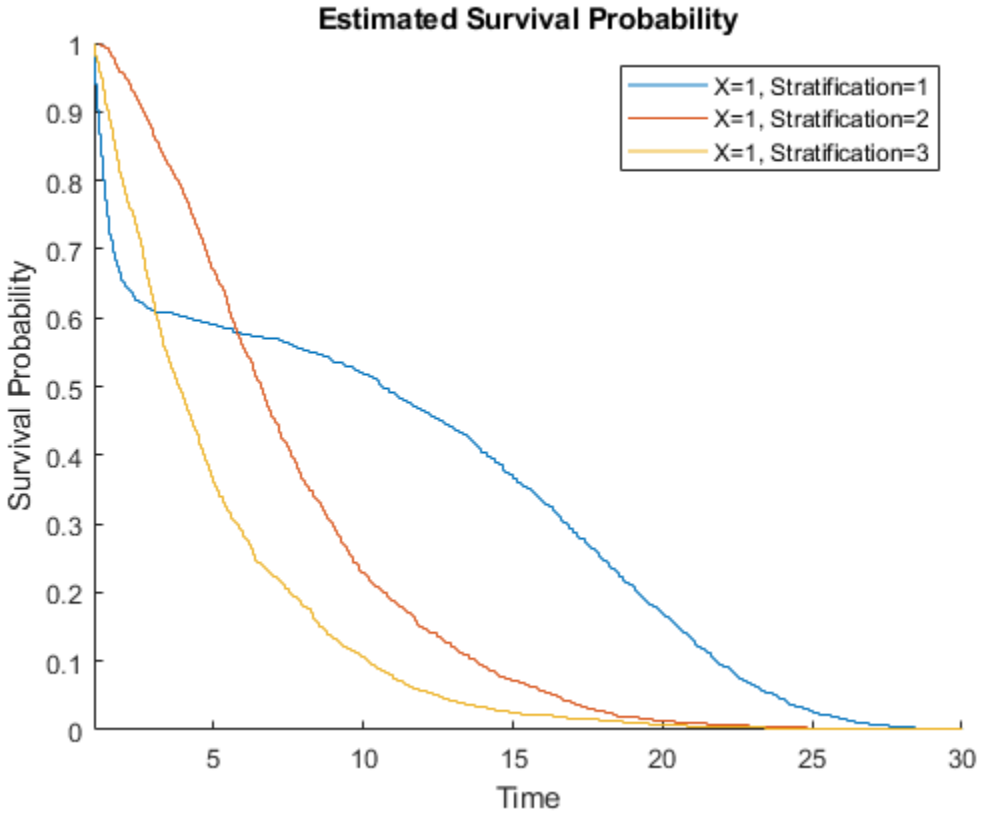
Fit a Cox proportional hazards model to the table data.

```
coxMdl = fitcox(tbldata,'Lifetime ~ Predictors',"Stratification",'Strata');
```

Plot Survival

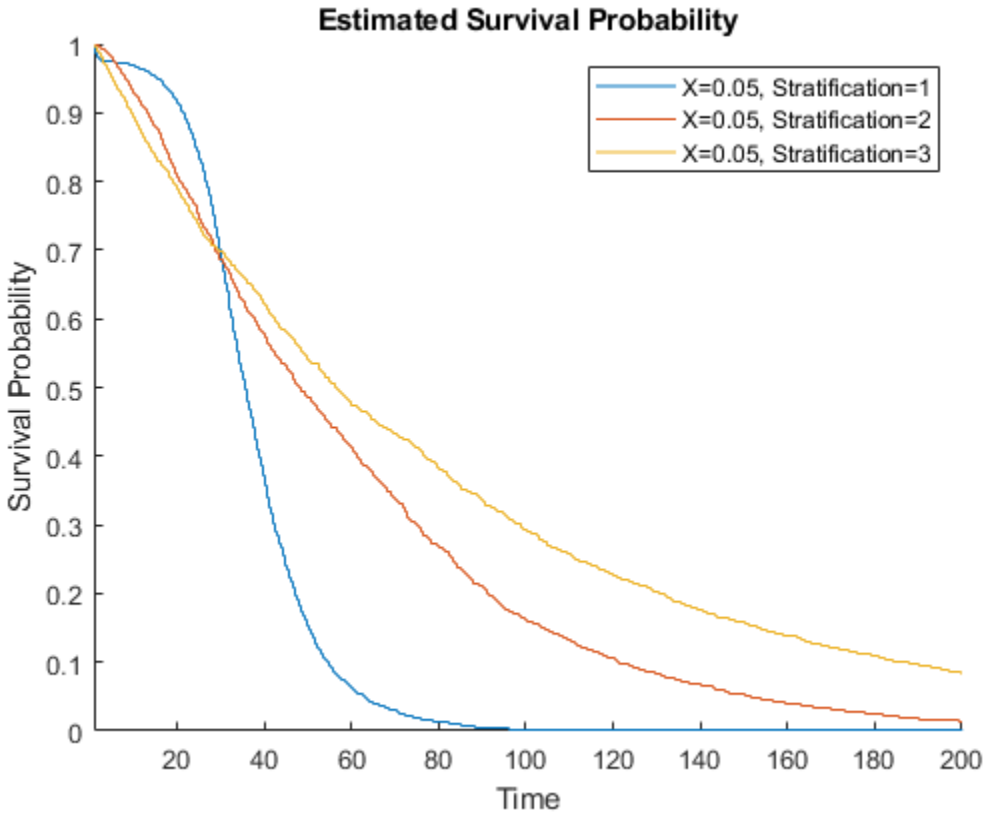
Plot the probability of survival as a function of time for predictor value `1` and the three stratification levels.

```
oo = categorical(1);
plotSurvival(coxMdl,[oo;oo;oo],[1;2;3])
xlim([1,30])
```



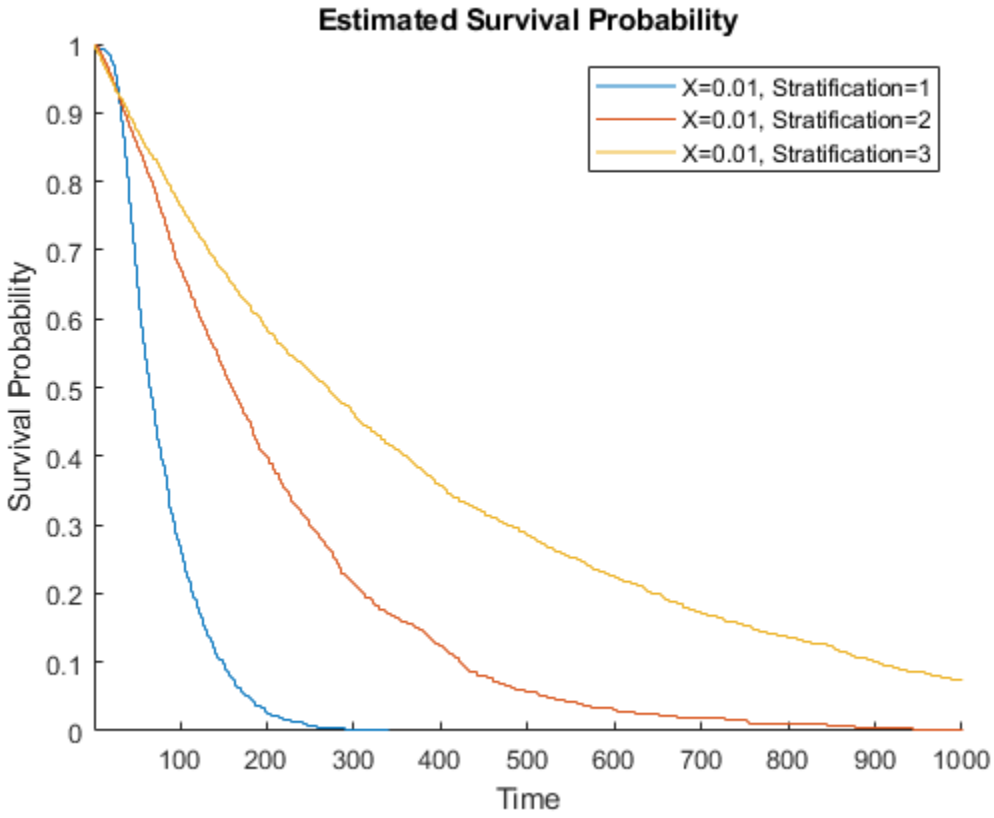
Plot the probability of survival as a function of time for predictor value `1/20` and the three stratification levels.

```
tt = categorical(1/20);
figure
plotSurvival(coxMdl,[tt;tt;tt],[1;2;3])
xlim([1,200])
```



Plot the probability of survival as a function of time for predictor value `1/100` and the three stratification levels.

```
uu = categorical(1/100);
figure
plotSurvival(coxMdl,[uu;uu;uu],[1;2;3])
xlim([1,1000])
```



Even though the hazard rates are proportional for the different predictors, the three survival plots are not proportional.

Analyze Fit

Examine the coefficients of the fitted model.

```
disp(coxMdl.Coefficients)
```

```
                      Beta        SE       zStat     pValue
                    ______    ________    ______    ______

    Predictors_0.05    1.5301    0.031783    48.143      0
    Predictors_1       4.5593    0.052149    87.427      0
```

Notice that the `pValue` entries are `0`, which means that the listed predictor `Beta` values are not zero.

View the confidence intervals for the model coefficients at the 0.01 level of significance.

```
coefci(coxMdl,0.01)
```

```
ans = 2×2

    1.4483    1.6120
    4.4249    4.6936
```

To infer the hazard for the 0.01 level predictor, recall the definition of the Cox model:

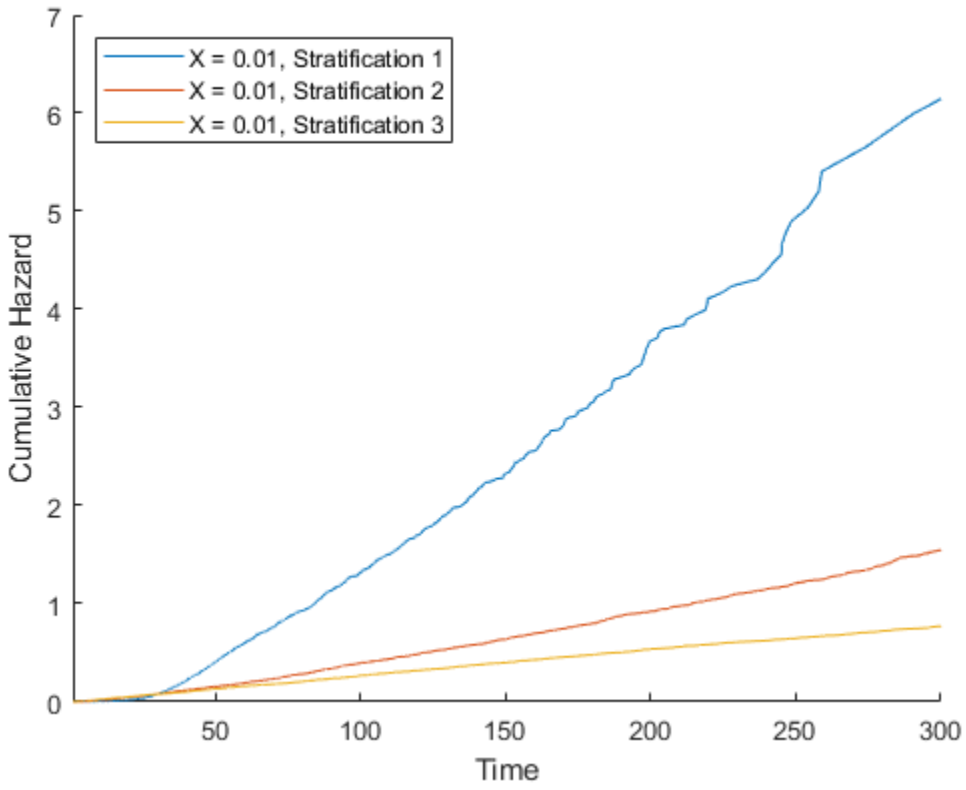
$h(X_i, t) = h_0(t)\exp\left(\sum\ x_{ij}b_j\right)$.

The `fitcox` function uses dummy variables with a reference group to handle categorical data. In this case, the function treats the 0.01 level predictor as the reference group, and encodes the predictor as all 0s when fitting the model. If you enter all 0s into the hazard function, you get

$$h([0,0],t) = h_0(t)\exp(0*b1 + 0*b2) = h_0(t)\exp(0) = h_0(t)\,.$$

$h_0(t)$ is the baseline hazard, which is stored in `coxMdl.Hazard`. Therefore, to get the hazard for the 0.01 level predictor, you can examine `coxMdl.Hazard`. Plot the baseline cumulative hazard for the three stratification levels.

```
figure
hold on
for i = 1:3
    pred3 = find(coxMdl.Hazard(:,3) == i); % Find indices of stratification i
    plot(coxMdl.Hazard(pred3,1),coxMdl.Hazard(pred3,2))
end
xlabel('Time')
ylabel('Cumulative Hazard')
xlim([1 300])
legend('X = 0.01, Stratification 1',...
    'X = 0.01, Stratification 2',...
    'X = 0.01, Stratification 3','Location','northwest')
hold off
```



The cumulative hazard for the other predictor values is exp(Beta) times the baseline cumulative hazard, where Beta is the inferred coefficient.

```
disp(exp(coxMdl.Coefficients.Beta))
```

```
    4.6188
   95.5127
```

These relative hazard values are close to the theoretical values for the data, which was generated with multipliers 1, 1/20, and 1/100. The baseline value corresponds to the 1/100 multiplier, so the theoretical multipliers are 5 and 100.

View the `linhyptest` table.

```
linhyptest(coxMdl)
```

```
ans=2×2 table
         Predictor          pValue
    ____________________    ______

    {'Empty Model'    }       0
    {'Predictors_0.05'}       0
```

Again, the model requires the `1/20` value predictor and the 1 value predictor.

Check whether the data supports the hypothesis that the data is from a Cox proportional hazards model.

```
supports = coxMdl.ProportionalHazardsPValueGlobal
```

```
supports = 0.9730
```

The null hypothesis for this test is that the data is from a Cox proportional hazards model. To reject this hypothesis, `supports` must be less than 0.05 or some other small significance level. The statistic indicates support for the hypothesis that the data is consistent with a Cox model.

Examine Hazard Ratios

Calculate the hazard ratio for the predictor values `1`, `1/20`, and `1/100` compared to a baseline of `0` for the three stratification levels.

```
hazards = hazardratio(coxMdl,...
    categorical([1;1;1;1/20;1/20;1/20;1/100;1/100;1/100]),...
    [1;2;3;1;2;3;1;2;3],'Baseline',0)
```

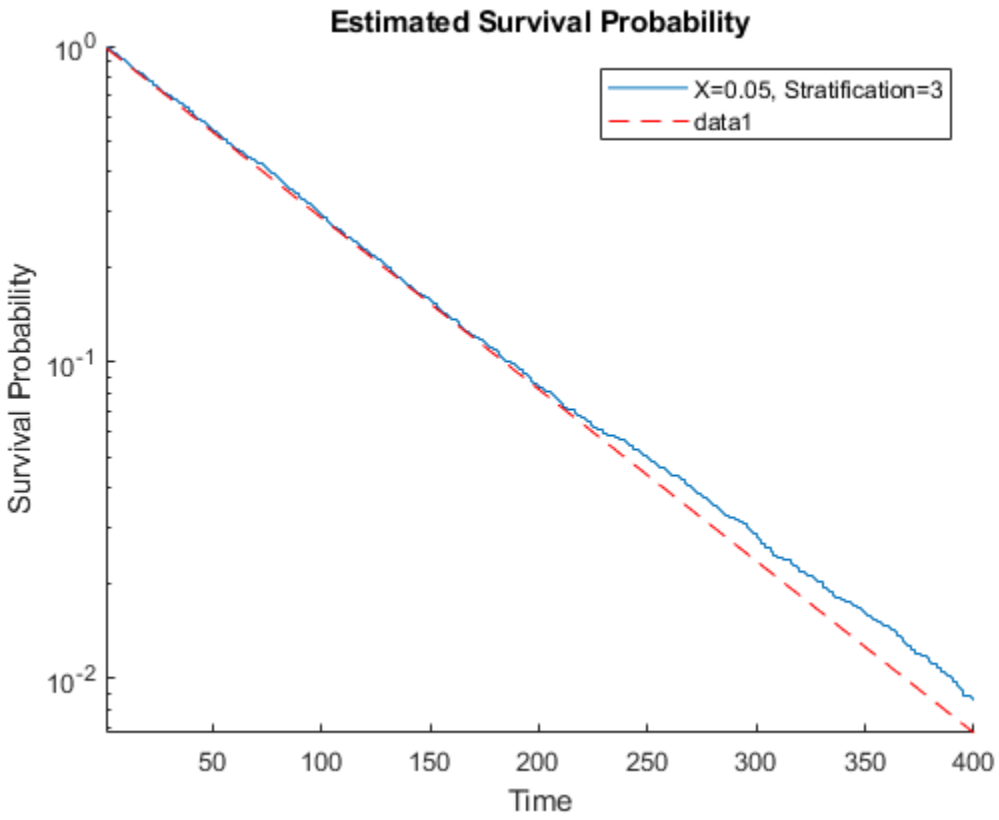
```
hazards = 9×1

   95.5127
   95.5127
   95.5127
    4.6188
    4.6188
    4.6188
    1.0000
    1.0000
    1.0000
```

The hazard ratios are near their theoretical values of 100, 5, and 1, as explained in the previous section. The hazard ratios do not depend on the stratification level.

How Well Does the Constant Hazard Stratification Level Match Theory?

Stratification level 3 has a constant hazard rate of 1/4. Theoretically, a constant hazard rate means an exponential survival function, whose logarithm is a straight line. Plot the survival of level 3 stratification using the predictor value `1/20`, which leads to an exponential rate of 1/80.

```
tt = categorical(1/20);
h = figure;
axes1 = axes('Parent',h);
plotSurvival(coxMdl,axes1,tt,3);
xlim([1 400]);
axes1.YScale = 'log';
hold on
tms = linspace(1,400);
semilogy(tms,exp(-tms/80),'r--')
hold off
```



The data matches the theoretical line for probabilities well above 1/100.

Helper Functions

This function creates the bathtub hazard rate.

```
function h = hazard(x)
h = exp(-2*(x - 1)) + (1 + tanh(x/10 - 3));
end
```

This function creates the integral of the bathtub hazard rate from `1` to `x`.

```
function eh = exphazard(x)
eh = 1/2 - exp(-2*(x-1))/2;
eh2 = (10*log(cosh(x/10 - 3)) - 10*log(cosh(1/10 - 3)) + x - 1);
eh = eh + eh2;
end
```

This function solves for the root of the cumulative hazard rate with multiplier `a` to level `v`.

```
function zz = zeropt(a,v)
zz = fzero(@(x)(1 - exp(-a*exphazard(x)) - v),[1 100*max(1,1/a)]);
end
```

This function creates the integral of the logarithmic hazard rate with multiplier `1/10` from `1` to `x`.

```
function cr = cumrisk(x)
cr = 1/10*(x.*(log(x) - 1) + 1);
end
```

This function solves for the root of the cumulative hazard rate with multiplier `a` to level `v`.

```
function zz = zeroptold(a,v)
zz = fzero(@(x)(1 - exp(-a*cumrisk(x)) - v),[1 50*max(1,1/a)]);
end
```

See Also

`CoxModel` | `fitcox`

Related Examples

- “Analysis of Lifetime Data”

Analyzing Survival or Reliability Data

This example shows how to analyze lifetime data with censoring. In biological or medical applications, this is known as survival analysis, and the times may represent the survival time of an organism or the time until a disease is cured. In engineering applications, this is known as reliability analysis, and the times may represent the time to failure of a piece of equipment.

Our example models the time to failure of a throttle from an automobile fuel injection system.

Special Properties of Lifetime Data

Some features of lifetime data distinguish them other types of data. First, the lifetimes are always positive values, usually representing time. Second, some lifetimes may not be observed exactly, so that they are known only to be larger than some value. Third, the distributions and analysis techniques that are commonly used are fairly specific to lifetime data

Let's simulate the results of testing 100 throttles until failure. We'll generate data that might be observed if most throttles had a fairly long lifetime, but a small percentage tended to fail very early.

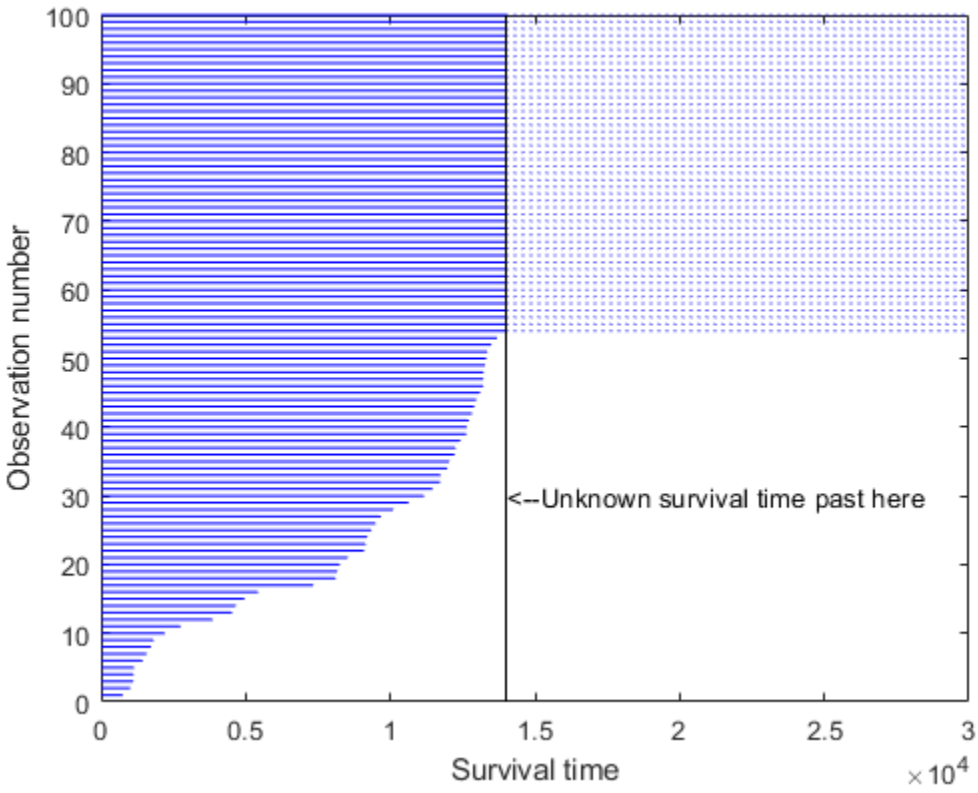
```
rng(2,'twister');
lifetime = [wblrnd(15000,3,90,1); wblrnd(1500,3,10,1)];
```

In this example, assume that we are testing the throttles under stressful conditions, so that each hour of testing is equivalent to 100 hours of actual use in the field. For pragmatic reasons, it's often the case that reliability tests are stopped after a fixed amount of time. For this example, we will use 140 hours, equivalent to a total of 14,000 hours of real service. Some items fail during the test, while others survive the entire 140 hours. In a real test, the times for the latter would be recorded as 14,000, and we mimic this in the simulated data. It is also common practice to sort the failure times.

```
T = 14000;
obstime = sort(min(T, lifetime));
```

We know that any throttles that survive the test will fail eventually, but the test is not long enough to observe their actual time to failure. Their lifetimes are only known to be greater than 14,000 hours. These values are said to be censored. This plot shows that about 40% of our data are censored at 14,000.

```
failed = obstime(obstime<T); nfailed = length(failed);
survived = obstime(obstime==T); nsurvived = length(survived);
censored = (obstime >= T);
plot([zeros(size(obstime)),obstime]', repmat(1:length(obstime),2,1), ...
     'Color','b','LineStyle','-')
line([T;3e4], repmat(nfailed+(1:nsurvived), 2, 1), 'Color','b','LineStyle',':');
line([T;T], [0;nfailed+nsurvived],'Color','k','LineStyle','-')
text(T,30,'<--Unknown survival time past here')
xlabel('Survival time'); ylabel('Observation number')
```



Ways of Looking at Distributions

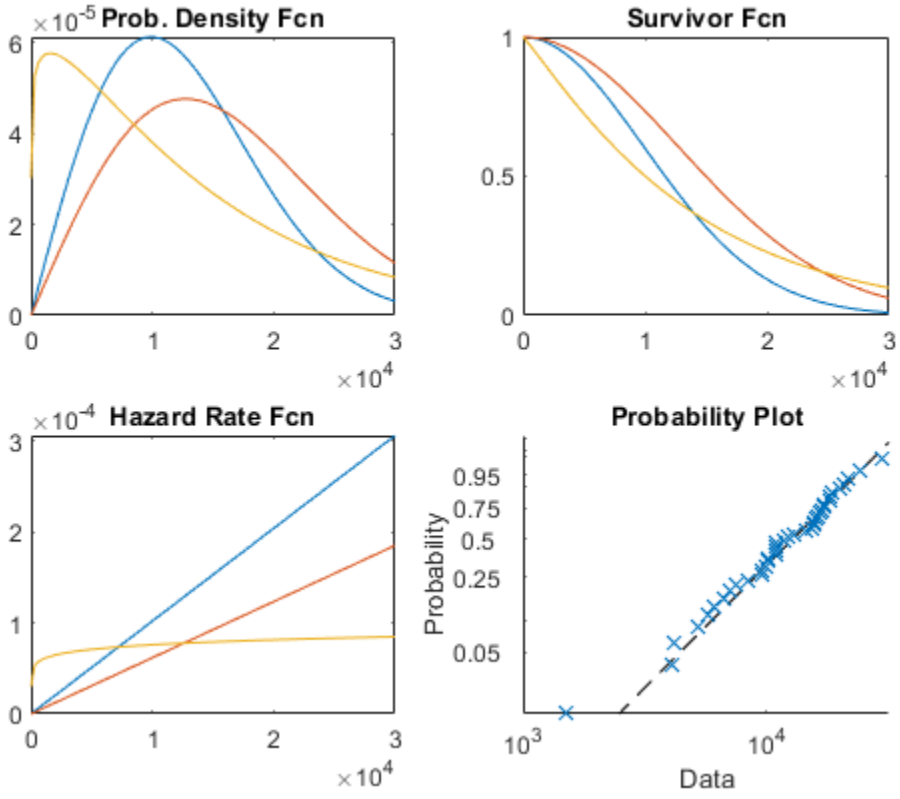
Before we examine the distribution of the data, let's consider different ways of looking at a probability distribution.

- A probability density function (PDF) indicates the relative probability of failure at different times.
- A survivor function gives the probability of survival as a function of time, and is simply one minus the cumulative distribution function (1-CDF).
- The hazard rate gives the instantaneous probability of failure given survival to a given time. It is the PDF divided by the survivor function. In this example the hazard rates turn out to be increasing, meaning the items are more susceptible to failure as time passes (aging).
- A probability plot is a re-scaled CDF, and is used to compare data to a fitted distribution.

Here are examples of those four plot types, using the Weibull distribution to illustrate. The Weibull is a common distribution for modeling lifetime data.

```
x = linspace(1,30000);
subplot(2,2,1);
plot(x,wblpdf(x,14000,2),x,wblpdf(x,18000,2),x,wblpdf(x,14000,1.1))
title('Prob. Density Fcn')
subplot(2,2,2);
plot(x,1-wblcdf(x,14000,2),x,1-wblcdf(x,18000,2),x,1-wblcdf(x,14000,1.1))
title('Survivor Fcn')
subplot(2,2,3);
wblhaz = @(x,a,b) (wblpdf(x,a,b) ./ (1-wblcdf(x,a,b)));
plot(x,wblhaz(x,14000,2),x,wblhaz(x,18000,2),x,wblhaz(x,14000,1.1))
```

```
title('Hazard Rate Fcn')
subplot(2,2,4);
probplot('weibull',wblrnd(14000,2,40,1))
title('Probability Plot')
```



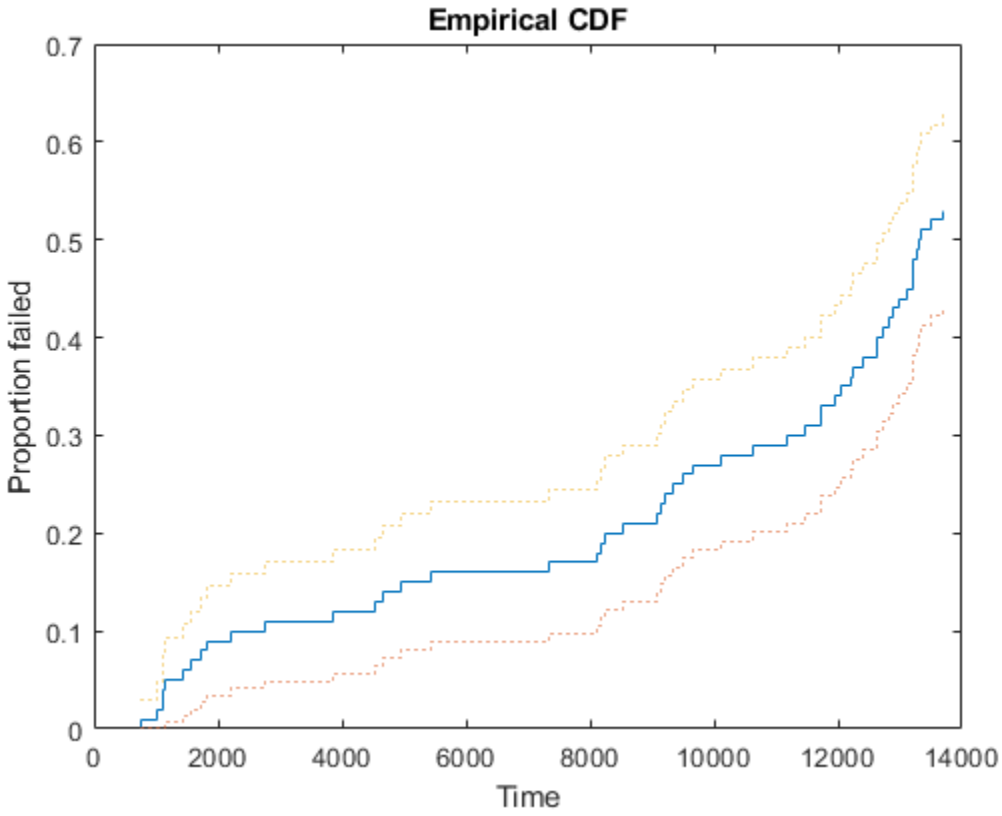
Fitting a Weibull Distribution

The Weibull distribution is a generalization of the exponential distribution. If lifetimes follow an exponential distribution, then they have a constant hazard rate. This means that they do not age, in the sense that the probability of observing a failure in an interval, given survival to the start of that interval, doesn't depend on where the interval starts. A Weibull distribution has a hazard rate that may increase or decrease.

Other distributions used for modeling lifetime data include the lognormal, gamma, and Birnbaum-Saunders distributions.

We will plot the empirical cumulative distribution function of our data, showing the proportion failing up to each possible survival time. The dotted curves give 95% confidence intervals for these probabilities.

```
subplot(1,1,1);
[empF,x,empFlo,empFup] = ecdf(obstime,'censoring',censored);
stairs(x,empF);
hold on;
stairs(x,empFlo,':'); stairs(x,empFup,':');
hold off
xlabel('Time'); ylabel('Proportion failed'); title('Empirical CDF')
```



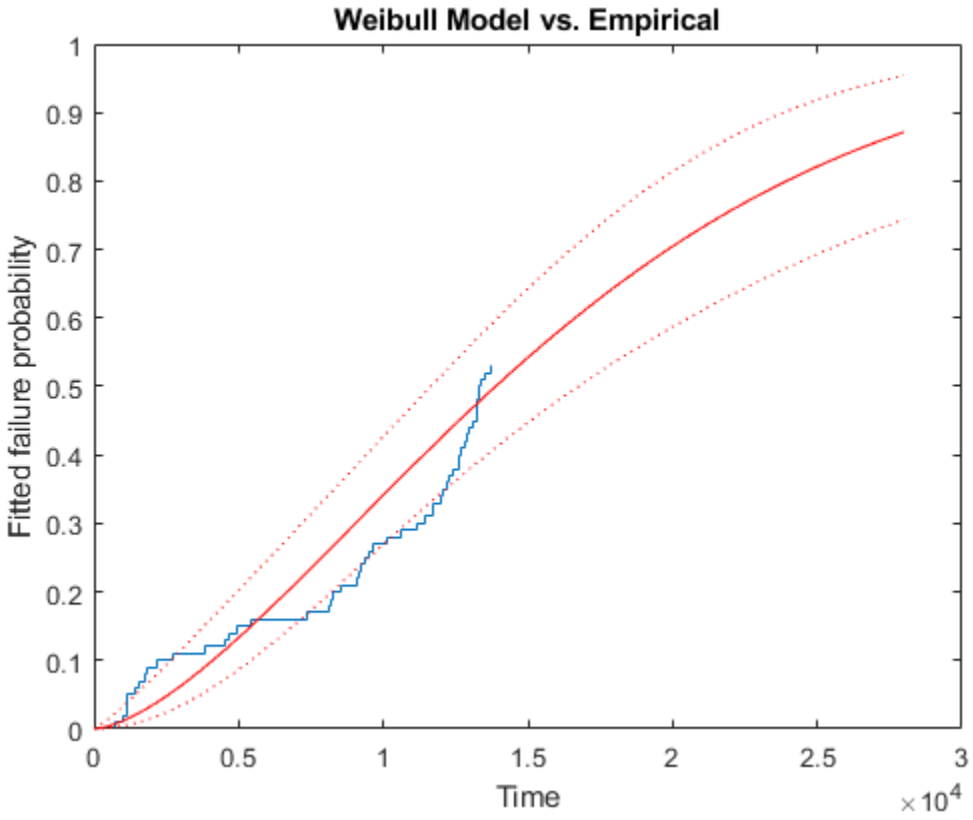
This plot shows, for instance, that the proportion failing by time 4,000 is about 12%, and a 95% confidence bound for the probability of failure by this time is from 6% to 18%. Notice that because our test only ran 14,000 hours, the empirical CDF only allows us to compute failure probabilities out to that limit. Almost half of the data were censored at 14,000, and so the empirical CDF only rises to about 0.53, instead of 1.0.

The Weibull distribution is often a good model for equipment failure. The function `wblfit` fits the Weibull distribution to data, including data with censoring. After computing parameter estimates, we'll evaluate the CDF for the fitted Weibull model, using those estimates. Because the CDF values are based on estimated parameters, we'll compute confidence bounds for them.

```
paramEsts = wblfit(obstime,'censoring',censored);
[nlogl,paramCov] = wbllike(paramEsts,obstime,censored);
xx = linspace(1,2*T,500);
[wblF,wblFlo,wblFup] = wblcdf(xx,paramEsts(1),paramEsts(2),paramCov);
```

We can superimpose plots of the empirical CDF and the fitted CDF, to judge how well the Weibull distribution models the throttle reliability data.

```
stairs(x,empF);
hold on
handles = plot(xx,wblF,'r-',xx,wblFlo,'r:',xx,wblFup,'r:');
hold off
xlabel('Time'); ylabel('Fitted failure probability'); title('Weibull Model vs. Empirical')
```

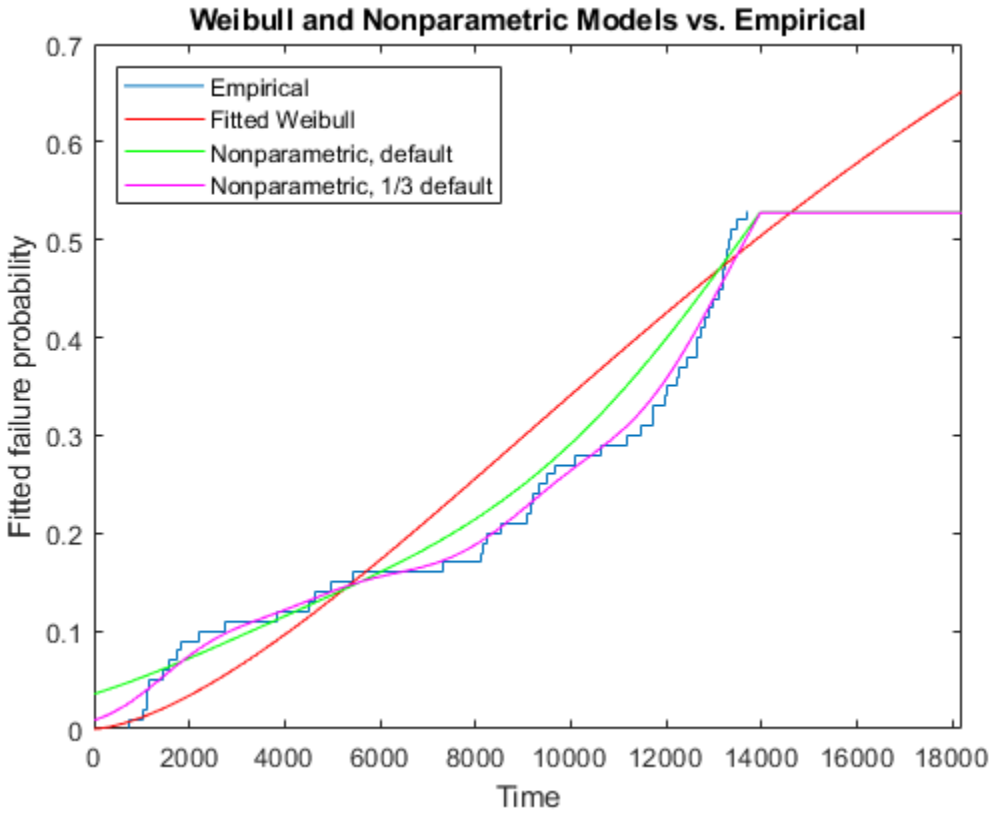


Notice that the Weibull model allows us to project out and compute failure probabilities for times beyond the end of the test. However, it appears the fitted curve does not match our data well. We have too many early failures before time 2,000 compared with what the Weibull model would predict, and as a result, too few for times between about 7,000 and about 13,000. This is not surprising -- recall that we generated data with just this sort of behavior.

Adding a Smooth Nonparametric Estimate

The pre-defined functions provided with the Statistics and Machine Learning Toolbox™ don't include any distributions that have an excess of early failures like this. Instead, we might want to draw a smooth, nonparametric curve through the empirical CDF, using the function `ksdensity`. We'll remove the confidence bands for the Weibull CDF, and add two curves, one with the default smoothing parameter, and one with a smoothing parameter 1/3 the default value. The smaller smoothing parameter makes the curve follow the data more closely.

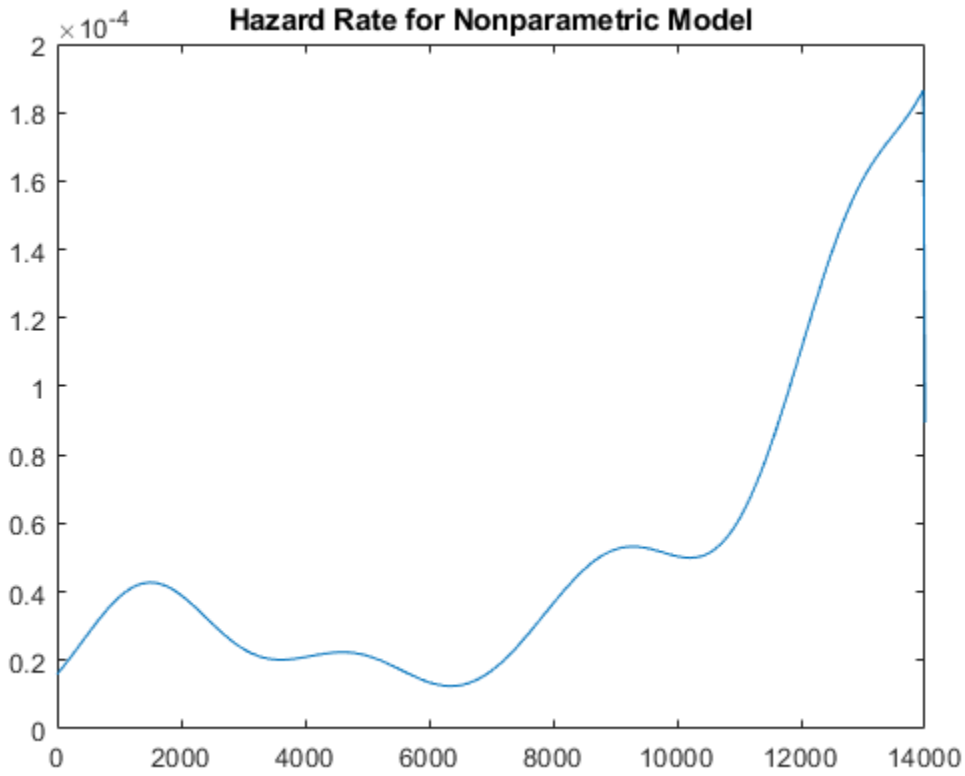
```
delete(handles(2:end))
[npF,ignore,u] = ksdensity(obstime,xx,'cens',censored,'function','cdf');
line(xx,npF,'Color','g');
npF3 = ksdensity(obstime,xx,'cens',censored,'function','cdf','width',u/3);
line(xx,npF3,'Color','m');
xlim([0 1.3*T])
title('Weibull and Nonparametric Models vs. Empirical')
legend('Empirical','Fitted Weibull','Nonparametric, default','Nonparametric, 1/3 default', ...
       'location','northwest');
```



The nonparametric estimate with the smaller smoothing parameter matches the data well. However, just as for the empirical CDF, it is not possible to extrapolate the nonparametric model beyond the end of the test -- the estimated CDF levels off above the last observation.

Let's compute the hazard rate for this nonparametric fit and plot it over the range of the data.

```
hazrate = ksdensity(obstime,xx,'cens',censored,'width',u/3) ./ (1-npF3);
plot(xx,hazrate)
title('Hazard Rate for Nonparametric Model')
xlim([0 T])
```



This curve has a bit of a "bathtub" shape, with a hazard rate that is high near 2,000, drops to lower values, then rises again. This is typical of the hazard rate for a component that is more susceptible to failure early in its life (infant mortality), and again later in its life (aging).

Also notice that the hazard rate cannot be estimated above the largest uncensored observation for the nonparametric model, and the graph drops to zero.

Alternative Models

For the simulated data we've used for this example, we found that a Weibull distribution was not a suitable fit. We were able to fit the data well with a nonparametric fit, but that model was only useful within the range of the data.

One alternative would be to use a different parametric distribution. The Statistics and Machine Learning Toolbox includes functions for other common lifetime distributions such as the lognormal, gamma, and Birnbaum-Saunders, as well as many other distributions that are not commonly used in lifetime models. You can also define and fit custom parametric models to lifetime data, as described in the “Fitting Custom Univariate Distributions, Part 2” on page 5-176 example.

Another alternative would be to use a mixture of two parametric distributions -- one representing early failure and the other representing the rest of the distribution. Fitting mixtures of distributions is described in the “Fitting Custom Univariate Distributions” on page 5-165 example.

Multivariate Methods

Multivariate Linear Regression

In this section...

Introduction to Multivariate Methods

Large, high-dimensional data sets are common in the modern era of computer-based instrumentation and electronic data storage. High-dimensional data present many challenges for statistical visualization, analysis, and modeling.

Data visualization, of course, is impossible beyond a few dimensions. As a result, pattern recognition, data preprocessing, and model selection must rely heavily on numerical methods.

A fundamental challenge in high-dimensional data analysis is the so-called curse of dimensionality. Observations in a high-dimensional space are necessarily sparser and less representative than those in a low-dimensional space. In higher dimensions, data over-represent the edges of a sampling distribution, because regions of higher-dimensional space contain the majority of their volume near the surface. (A d-dimensional spherical shell has a volume, relative to the total volume of the sphere, that approaches 1 as d approaches infinity.) In high dimensions, typical data points at the interior of a distribution are sampled less frequently.

Often, many of the dimensions in a data set—the measured features—are not useful in producing a model. Features may be irrelevant or redundant. Regression and classification algorithms may require large amounts of storage and computation time to process raw data, and even if the algorithms are successful the resulting models may contain an incomprehensible number of terms.

Because of these challenges, multivariate statistical methods often begin with some type of dimension reduction, in which data are approximated by points in a lower-dimensional space. Dimension reduction is the goal of the methods presented in this chapter. Dimension reduction often leads to simpler models and fewer measured variables, with consequent benefits when measurements are expensive and visualization is important.

Multivariate Linear Regression Model

The multivariate linear regression model expresses a d-dimensional continuous response vector as a linear combination of predictor terms plus a vector of error terms with a multivariate normal distribution. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})'$ denote the response vector for observation i, $i = 1,\ldots,n$. In the most general case, given the d-by-K design matrix $\mathbf{X}_i$ and the K-by-1 vector of coefficientsβ, the multivariate linear regression model is

$$\mathbf{y}_i = \mathbf{X}_i\beta + \varepsilon_i,$$

where the d-dimensional vector of error terms follows a multivariate normal distribution,

$$\varepsilon_i \sim MVN_d(\mathbf{0}, \Sigma)\,.$$

The model assumes independence between observations, meaning the error variance-covariance matrix for the n stacked d-dimensional response vectors is

$$\mathbf{I}_n \otimes \Sigma = \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix}.$$

If $\mathbf{y}$ denotes the nd-by-1 vector of stacked d-dimensional responses, and $\mathbf{X}$ denotes the nd-by-K matrix of stacked design matrices, then the distribution of the response vector is

$$\mathbf{y} \sim MVN_{nd}(\mathbf{X}\beta, \mathbf{I}_n \otimes \Sigma).$$

Solving Multivariate Regression Problems

To fit multivariate linear regression models of the form

$$\mathbf{y}_i = \mathbf{X}_i\beta + \varepsilon_i, \quad \varepsilon_i \sim MVN_d(\mathbf{0}, \Sigma)$$

in Statistics and Machine Learning Toolbox, use `mvregress`. This function fits multivariate regression models with a diagonal (heteroscedastic) or unstructured (heteroscedastic and correlated) error variance-covariance matrix, Σ, using least squares or maximum likelihood estimation.

Many variations of multivariate regression might not initially appear to be of the form supported by `mvregress`, such as:

- Multivariate general linear model
- Multivariate analysis of variance (MANOVA)
- Longitudinal analysis
- Panel data analysis
- Seemingly unrelated regression (SUR)
- Vector autoregressive (VAR) model

In many cases, you can frame these problems in the form used by `mvregress` (but `mvregress` does not support parameterized error variance-covariance matrices). For the special case of one-way MANOVA, you can alternatively use `manova1`. Econometrics Toolbox™ has functions for VAR estimation.

Note The multivariate linear regression model is distinct from the multiple linear regression model, which models a *univariate* continuous response as a linear combination of exogenous terms plus an independent and identically distributed error term. To fit a multiple linear regression model, use `fitlm`.

See Also

`fitlm` | `manova1` | `mvregress` | `mvregresslike`

Related Examples

- “Set Up Multivariate Regression Problems” on page 15-11
- “Multivariate General Linear Model” on page 15-20
- “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24
- “Longitudinal Analysis” on page 15-30

More About

- “Estimation of Multivariate Regression Models” on page 15-5

Estimation of Multivariate Regression Models

In this section...

Least Squares Estimation

Ordinary Least Squares

When you fit multivariate linear regression models using `mvregress`, you can use the optional name-value pair `'algorithm','cwls'` to choose least squares estimation. In this case, by default, `mvregress` returns ordinary least squares (OLS) estimates using $\Sigma = \mathbf{I}_d$. Alternatively, if you specify a covariance matrix for weighting, you can return covariance-weighted least squares (CWLS) estimates. If you combine OLS and CWLS, you can get feasible generalized least squares (FGLS) estimates.

The OLS estimate for the coefficient vector is the vector $\mathbf{b}$ that minimizes

$$\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}).$$

Let $\mathbf{y}$ denote the nd-by-1 vector of stacked d-dimensional responses, and $\mathbf{X}$ denote the nd-by-K matrix of stacked design matrices. The K-by-1 vector of OLS regression coefficient estimates is

$$\mathbf{b}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This is the first `mvregress` output.

Given $\Sigma = \mathbf{I}_d$ (the `mvregress` OLS default), the variance-covariance matrix of the OLS estimates is

$$V(\mathbf{b}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}.$$

This is the fourth `mvregress` output. The standard errors of the OLS regression coefficients are the square root of the diagonal of this variance-covariance matrix.

If your data is not scaled such that $\Sigma = \sigma^2\mathbf{I}_d$, then you can multiply the `mvregress` variance-covariance matrix by the mean squared error (MSE), an unbiased estimate of σ^2. To compute the MSE, return the n-by-d matrix of residuals, $\mathbf{E}$ (the third `mvregress` output). Then,

$$\text{MSE} = \frac{\sum_{i=1}^{n} \mathbf{e}_i\mathbf{e}_i'}{n-K},$$

where $\mathbf{e}_i = (\mathbf{y}_i - \mathbf{X}_i\beta)'$ is the ith row of $\mathbf{E}$.

Covariance-Weighted Least Squares

For most multivariate problems, an identity error covariance matrix is insufficient, and leads to inefficient or biased standard error estimates. You can specify a matrix for CWLS estimation using the optional name-value pair argument `covar0`, for example, an invertible d-by-d matrix named $\mathbf{C}_0$.

Usually, $\mathbf{C}_0$ is a diagonal matrix such that the inverse matrix $\mathbf{C}_0^{-1}$ contains weights for each dimension to model heteroscedasticity. However, $\mathbf{C}_0$ can also be a nondiagonal matrix that models correlation.

Given $\mathbf{C}_0$, the CWLS solution is the vector $\mathbf{b}$ that minimizes

$$\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})'\mathbf{C}_0(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}).$$

In this case, the K-by-1 vector of CWLS regression coefficient estimates is

$$\mathbf{b}_{CWLS} = \left(\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{C}_0)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{C}_0)^{-1}\mathbf{y}.$$

This is the first `mvregress` output.

If $\Sigma = \mathbf{C}_0$, this is the generalized least squares (GLS) solution. The corresponding variance-covariance matrix of the CWLS estimates is

$$V(\mathbf{b}_{CWLS}) = \left(\mathbf{X}'(\mathbf{I}_n \otimes \mathbf{C}_0)^{-1}\mathbf{X}\right)^{-1}.$$

This is the fourth `mvregress` output. The standard errors of the CWLS regression coefficients are the square root of the diagonal of this variance-covariance matrix.

If you only know the error covariance matrix up to a proportion, that is, $\Sigma = \sigma^2\mathbf{C}_0$, you can multiply the `mvregress` variance-covariance matrix by the MSE, as described in “Ordinary Least Squares” on page 15-5.

Error Covariance Estimation

Regardless of which least squares method you use, the estimate for the error variance-covariance matrix is

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{12} & \cdots & \widehat{\sigma}_{1d} \\ \widehat{\sigma}_{12} & \widehat{\sigma}_2^2 & \cdots & \widehat{\sigma}_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\sigma}_{1d} & \widehat{\sigma}_{2d} & \cdots & \widehat{\sigma}_d^2 \end{pmatrix} = \frac{\mathbf{E}'\mathbf{E}}{n},$$

where $\mathbf{E}$ is the n-by-d matrix of residuals. The ith row of $\mathbf{E}$ is $\mathbf{e}_i = (\mathbf{y}_i - \mathbf{X}_i\mathbf{b})'$.

The error covariance estimate, $\widehat{\Sigma}$, is the second `mvregress` output, and the matrix of residuals, $\mathbf{E}$, is the third output. If you specify the optional name-value pair `'covtype','diagonal'`, then `mvregress` returns $\widehat{\Sigma}$ with zeros in the off-diagonal entries,

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\sigma}_1^2 & & 0 \\ & \ddots & \\ 0 & & \widehat{\sigma}_d^2 \end{pmatrix}.$$

Feasible Generalized Least Squares

The generalized least squares estimate is the CWLS estimate with a known covariance matrix. That is, given Σ is known, the GLS solution is

$$\mathbf{b}_{GLS} = \left(\mathbf{X}'(\mathbf{I}_n \otimes \Sigma)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{I}_n \otimes \Sigma)^{-1}\mathbf{y},$$

with variance-covariance matrix

$$V(\mathbf{b}_{GLS}) = \left(\mathbf{X}'(\mathbf{I}_n \otimes \Sigma)^{-1}\mathbf{X}\right)^{-1}.$$

In most cases, the error covariance is unknown. The feasible generalized least squares (FGLS) estimate uses $\widehat{\Sigma}$ in place of Σ. You can obtain two-step FGLS estimates as follows:

1. Perform OLS regression, and return an estimate $\widehat{\Sigma}$.
2. Perform CWLS regression, using $\mathbf{C}_0 = \widehat{\Sigma}$.

You can also iterate between these two steps until convergence is reached.

For some data, the OLS estimate $\widehat{\Sigma}$ is positive semidefinite, and has no unique inverse. In this case, you cannot get the FGLS estimate using `mvregress`. As an alternative, you can use `lscov`, which uses a generalized inverse to return weighted least squares solutions for positive semidefinite covariance matrices.

Panel Corrected Standard Errors

An alternative to FGLS is to use OLS coefficient estimates (which are consistent) and make a standard error correction to improve efficiency. One such standard error adjustment—which does not require inversion of the covariance matrix—is panel corrected standard errors (PCSE) [1]. The panel corrected variance-covariance matrix for OLS estimates is

$$V_{pcse}(\mathbf{b}_{OLS}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}_n \otimes \Sigma)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

The PCSE are the square root of the diagonal of this variance-covariance matrix. “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24 illustrates PCSE computation.

Maximum Likelihood Estimation

Maximum Likelihood Estimates

The default estimation algorithm used by `mvregress` is maximum likelihood estimation (MLE). The loglikelihood function for the multivariate linear regression model is

$$\log L(\beta, \Sigma | \mathbf{y}, \mathbf{X}) = \frac{1}{2} nd \log(2\pi) + \frac{1}{2} n \log(\det(\Sigma)) + \frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{X}_i \beta)' \Sigma^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta).$$

The MLEs for β and Σ are the values that maximize the loglikelihood objective function.

`mvregress` finds the MLEs using an iterative two-stage algorithm. At iteration $m + 1$, the estimates are

$$\mathbf{b}_{MLE}^{(m+1)} = \left(\mathbf{X}' \left(\mathbf{I}_n \otimes \Sigma^{(m)} \right)^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \left(\mathbf{I}_n \otimes \Sigma^{(m)} \right)^{-1} \mathbf{y}$$

and

$$\widehat{\Sigma}^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{MLE}^{(m+1)} \right) \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_{MLE}^{(m+1)} \right)'.$$

The algorithm terminates when the changes in the coefficient estimates and loglikelihood objective function are less than a specified tolerance, or when the specified maximum number of iterations is reached. The optional name-value pair arguments for changing these convergence criteria are `tolbeta`, `tolobj`, and `maxiter`, respectively.

Standard Errors

The variance-covariance matrix of the MLEs is an optional `mvregress` output. By default, `mvregress` returns the variance-covariance matrix for only the regression coefficients, but you can also get the variance-covariance matrix of $\widehat{\Sigma}$ using the optional name-value pair `'vartype','full'`. In this case, `mvregress` returns the variance-covariance matrix for all K regression coefficients, and d or $d(d + 1)/2$ covariance terms (depending on whether the error covariance is diagonal or full).

By default, the variance-covariance matrix is the inverse of the observed Fisher information matrix (the `'hessian'` option). You can request the expected Fisher information matrix using the optional name-value pair `'vartype','fisher'`. Provided there is no missing response data, the observed and expected Fisher information matrices are the same. If response data is missing, the observed Fisher information accounts for the added uncertainty due to the missing values, whereas the expected Fisher information matrix does not.

The variance-covariance matrix for the regression coefficient MLEs is

$$V(\mathbf{b}_{MLE}) = \left(\mathbf{X}' \left(\mathbf{I}_n \otimes \widehat{\Sigma} \right)^{-1} \mathbf{X} \right)^{-1},$$

evaluated at the MLE of the error covariance matrix. This is the fourth `mvregress` output. The standard errors of the MLEs are the square root of the diagonal of this variance-covariance matrix.

For $\widehat{\Sigma}$, let θ denote the vector of parameters in the estimated error variance-covariance matrix. For example, if $d = 2$, then:

- If the estimated covariance matrix is diagonal, then $\theta = (\widehat{\sigma}_1^2, \widehat{\sigma}_2^2)$.
- If the estimated covariance matrix is full, then $\theta = (\widehat{\sigma}_1^2, \widehat{\sigma}_{12}, \widehat{\sigma}_2^2)$.

The Fisher information matrix for θ, $I(\theta)$, has elements

$$I(\theta)_{u,v} = \frac{1}{2} tr\left(\widehat{\Sigma}^{-1}\frac{\partial\widehat{\Sigma}}{\partial\theta_u}\widehat{\Sigma}^{-1}\frac{\partial\widehat{\Sigma}}{\partial\theta_v}\right), \quad u, v = 1, \ldots, n_\theta,$$

where n_θ is the length of θ (either d or $d(d + 1)/2$). The resulting variance-covariance matrix is

$$V(\theta) = I(\theta)^{-1}.$$

When you request the full variance-covariance matrix, `mvregress` returns (as the fourth output) the block diagonal matrix

$$\begin{pmatrix} V(\mathbf{b}_{MLE}) & \mathbf{0} \\ \mathbf{0} & V(\theta) \end{pmatrix}.$$

Missing Response Data

Expectation/Conditional Maximization

If any response values are missing, indicated by `NaN`, `mvregress` uses an expectation/conditional maximization (ECM) algorithm for estimation (if enough data is available). In this case, the algorithm is iterative for both least squares and maximum likelihood estimation. During each iteration, `mvregress` imputes missing response values using their conditional expectation.

Consider organizing the data so that the joint distribution of the missing and observed responses, denoted $\tilde{\mathbf{y}}$ and $\mathbf{y}$ respectively, can be written as

$$\begin{pmatrix} \tilde{\mathbf{y}} \\ \mathbf{y} \end{pmatrix} \sim MVN\left\{\begin{pmatrix} \widetilde{\mathbf{X}}\beta \\ \mathbf{X}\beta \end{pmatrix}, \begin{pmatrix} \Sigma_{\tilde{y}} & \Sigma_{\tilde{y}y} \\ \Sigma_{y\tilde{y}} & \Sigma_y \end{pmatrix}\right\}.$$

Using properties of the multivariate normal distribution, the conditional expectation of the missing responses given the observed responses is

$$\mathrm{E}(\tilde{\mathbf{y}}|\mathbf{y}) = \widetilde{\mathbf{X}}\beta + \Sigma_{\tilde{y}y}\Sigma_y^{-1}(\mathbf{y} - \mathbf{X}\beta).$$

Also, the variance-covariance matrix of the conditional distribution is

$$\mathrm{COV}(\tilde{\mathbf{y}}|\mathbf{y}) = \Sigma_{\tilde{y}} - \Sigma_{\tilde{y}y}\Sigma_y^{-1}\Sigma_{y\tilde{y}}.$$

At each iteration of the ECM algorithm, `mvregress` uses the parameter values from the previous iteration to:

- Update the regression coefficients using the combined vector of observed responses and conditional expectations of missing responses.
- Update the variance-covariance matrix, adjusting for missing responses using the variance-covariance matrix of the conditional distribution.

Finally, the residuals that `mvregress` returns for missing responses are the difference between the conditional expectation and the fitted value, both evaluated at the final parameter estimates.

If you prefer to ignore any observations that have missing response values, use the name-value pair `'algorithm','mvn'`. Note that `mvregress` always ignores observations that have missing predictor values.

Observed Information Matrix

By default, `mvregress` uses the observed Fisher information matrix (the `'hessian'` option) to compute the variance-covariance matrix of the regression parameters. This accounts for the additional uncertainty due to missing response values.

The observed information matrix includes contributions from only the observed responses. That is, the observed Fisher information matrix for the parameters in the error variance-covariance matrix has elements

$$I(\theta)_{u,v} = \frac{1}{2}\sum_{i=1}^{n} tr\left(\widehat{\Sigma}_i^{-1}\frac{\partial\widehat{\Sigma}_i}{\partial\theta_u}\widehat{\Sigma}_i^{-1}\frac{\partial\widehat{\Sigma}_i}{\partial\theta_v}\right),\quad u,v = 1,\ldots,n_\theta,$$

where $\widehat{\Sigma}_i$ is the subset of $\widehat{\Sigma}$ corresponding to the observed responses in $\mathbf{y}_i$.

For example, if $d = 3$, but y_{i2} is missing, then

$$\widehat{\Sigma}_i = \begin{pmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{13} \\ \widehat{\sigma}_{13} & \widehat{\sigma}_3^2 \end{pmatrix}.$$

The observed Fisher information for the regression coefficients has similar contributions from the design and covariance matrices.

References

[1] Beck, N. and J. N. Katz. "What to Do (and Not to Do) with Time-Series-Cross-Section Data in Comparative Politics." *American Political Science Review,* Vol. 89, No. 3, pp. 634–647, 1995.

See Also

`mvregress` | `mvregresslike`

Related Examples

- “Set Up Multivariate Regression Problems” on page 15-11
- “Multivariate General Linear Model” on page 15-20
- “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24
- “Longitudinal Analysis” on page 15-30

More About

- “Multivariate Linear Regression” on page 15-2

Set Up Multivariate Regression Problems

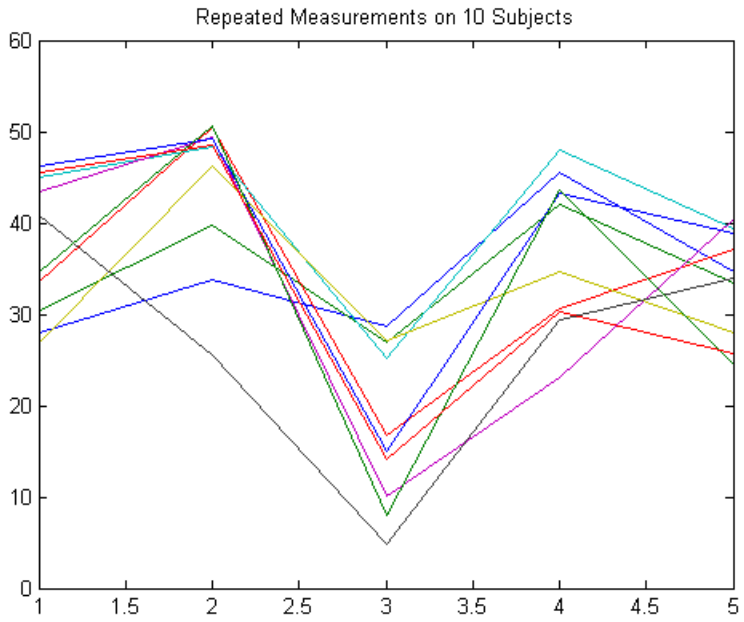
In this section...

Response Matrix

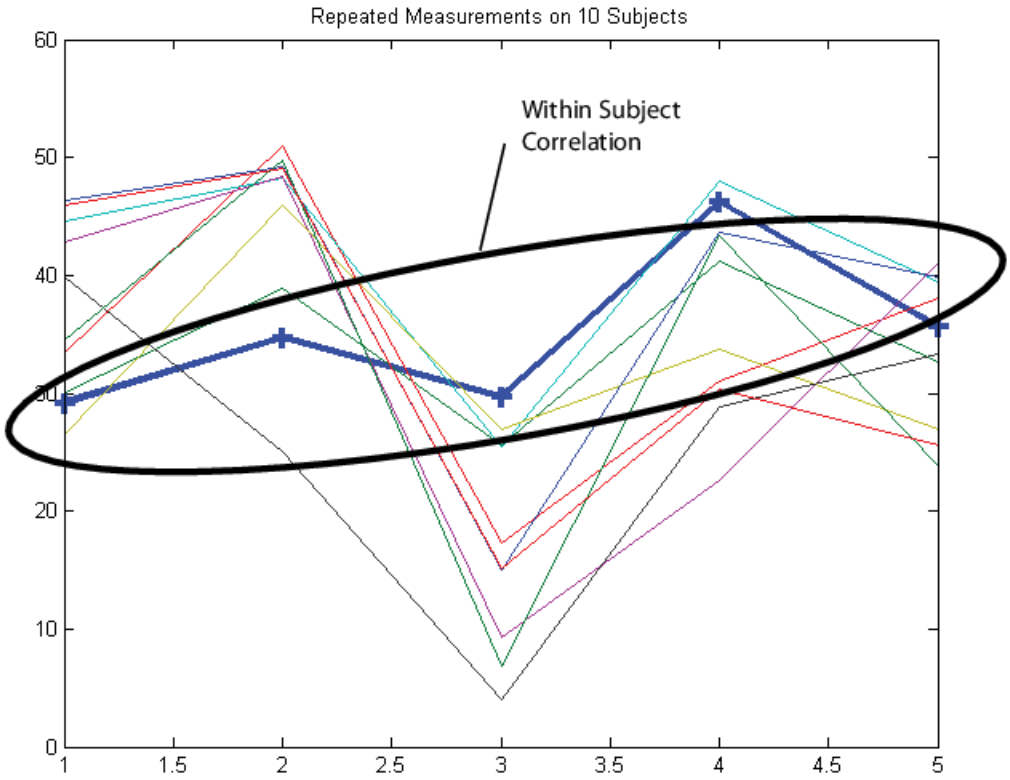
To fit a multivariate linear regression model using `mvregress`, you must set up your response matrix and design matrices in a particular way. Given properly formatted inputs, `mvregress` can handle a variety of multivariate regression problems.

`mvregress` expects the n observations of potentially correlated d-dimensional responses to be in an n-by-d matrix, named `Y`, for example. That is, set up your responses so that the dependency structure is between observations in the same *row*. If you specify `Y` as a vector of length n (either a row or column vector), then `mvregress` assumes that $d = 1$, and treats the elements as n independent observations. It does *not* model the vector as one realization of a correlated series (such as a time series).

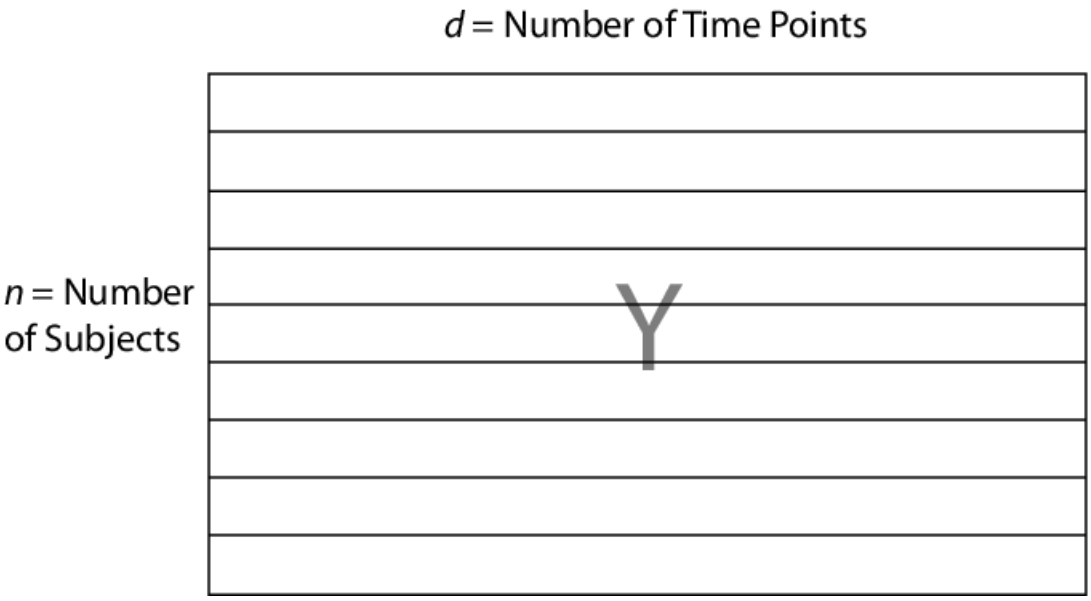
To illustrate how to set up a response matrix, suppose that your multivariate responses are repeated measurements made on subjects at multiple time points, as in the following figure.



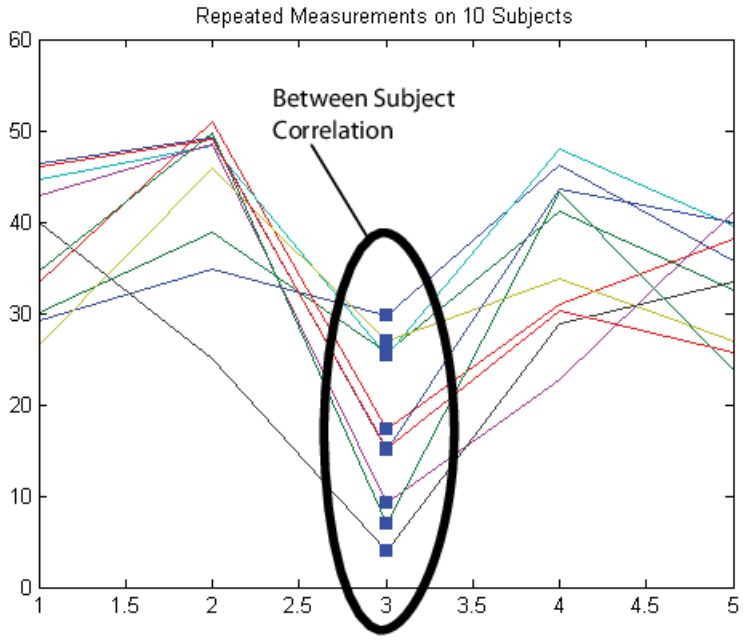
Suppose that observations within a subject are correlated.



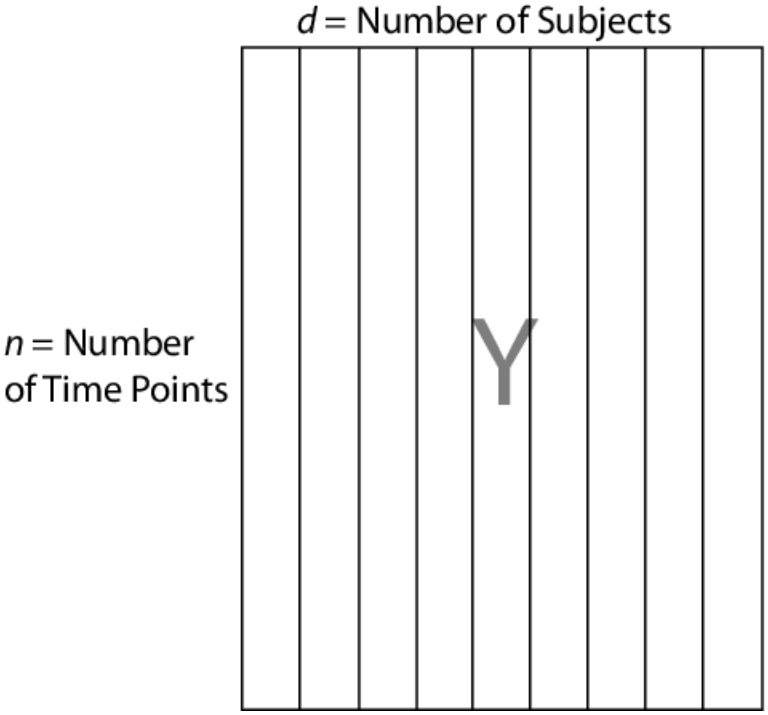
In this case, set up the response matrix Y such that each row corresponds to a subject, and each column corresponds to a time point.



Then again, suppose that observations made on subjects at the same time are correlated (concurrent correlation).



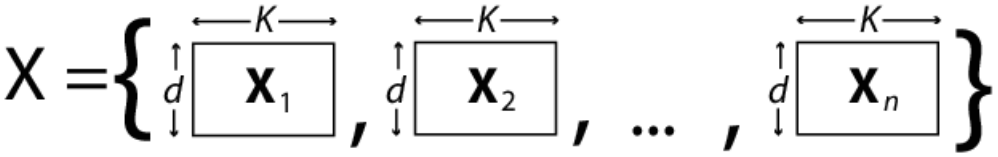
In this case, set up the response matrix Y such that each row corresponds to a time point, and each column corresponds to a subject.



Design Matrices

In the multivariate linear regression model, each d-dimensional response has a corresponding design matrix. Depending on the model, the design matrix might be comprised of exogenous predictor variables, dummy variables, lagged responses, or a combination of these and other covariate terms.

- If $d > 1$ and all d dimensions have the same design matrix, then specify one n-by-p design matrix, where p is the number of predictor variables. To determine an intercept for each dimension, add a column of ones to the design matrix. In this case, `mvregress` applies the design matrix to all d dimensions.
- If $d > 1$ and all d dimensions do not have the same design matrix, then specify the design matrices using a length-n cell array of d-by-K arrays, named `X`, for example. K is the total number of regression coefficients in the model. Note that the rows of the arrays in `X` correspond to the columns of the response matrix, `Y`.



  If all n observations have the same design matrix, you can specify a cell array containing one d-by-K design matrix. In this case, `mvregress` applies the design matrix to all n observations. For example, this situation might arise if the predictors are functions of time, and all observations were measured at the same time points.
- In the special case that $d = 1$, you can specify one n-by-K design matrix (not in a cell array). However, you should consider using `fitlm` to fit regression models to univariate, continuous responses.

The following sections illustrate how to set up the some common multivariate regression problems for estimation using `mvregress`.

Common Multivariate Regression Problems

Multivariate General Linear Model

The multivariate general linear model is of the form

$$\mathbf{Y}_{n\times d} = \mathbf{X}_{n\times (p+1)}\mathbf{B}_{(p+1)\times d} + \mathbf{E}_{n\times d}.$$

In expanded form,

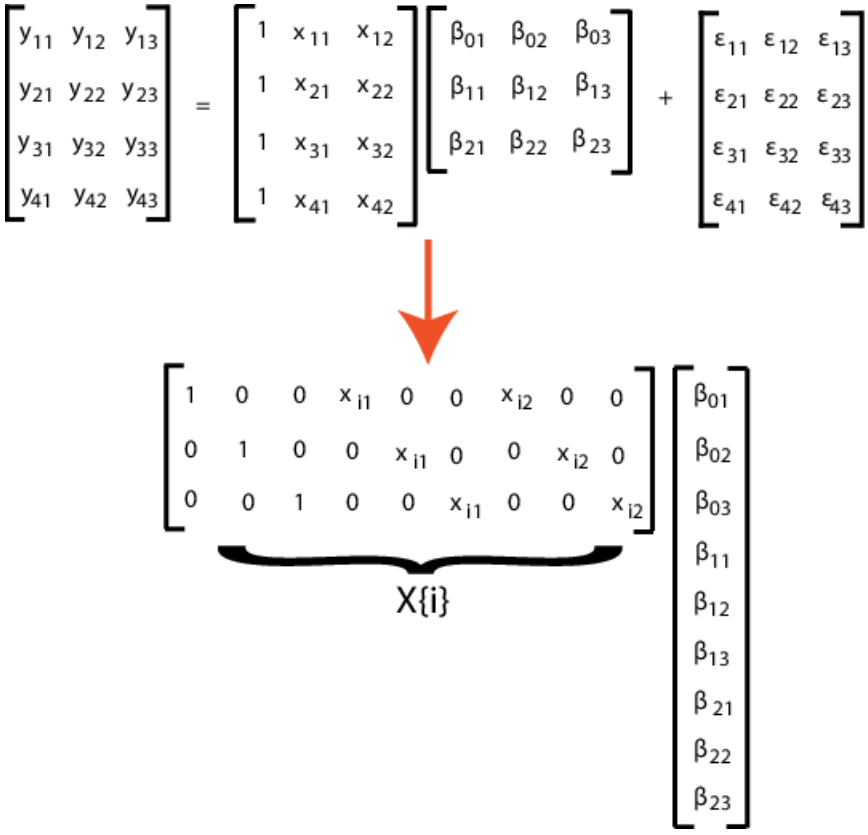
$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0d} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pd} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1d} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nd} \end{bmatrix}.$$

That is, each d-dimensional response has an intercept and p predictor variables, and each dimension has its own set of regression coefficients. In this form, the least squares solution is `B = X\Y`. To estimate this model using `mvregress`, use the n-by-d matrix of responses, as above.

If all d dimensions have the same design matrix, use the n-by-(p+1) design matrix, as above. Adding a column of ones to the p predictor variables computes the intercept for each dimension.

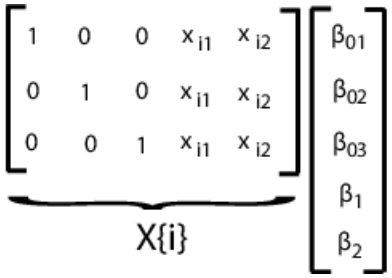
If all d dimensions do not have the same design matrix, reformat the n-by-(p + 1) design matrix into a length-n cell array of d-by-K matrices. Here, $K = (p + 1)d$ for an intercept and slopes for each dimension.

For example, suppose $n = 4$, $d = 3$, and $p = 2$ (two predictor terms in addition to an intercept). This figure shows how to format the ith element in the cell array.



If you prefer, you can reshape the K-by-1 vector of coefficients back into a $(p + 1)$-by-d matrix after estimation.

To put constraints on the model parameters, adjust the design matrix accordingly. For example, suppose that the three dimensions in the previous example have a common slope. That is, $\beta_{11} = \beta_{12} = \beta_{13} = \beta_1$ and $\beta_{21} = \beta_{22} = \beta_{23} = \beta_2$. In this case, each design matrix is 3-by-5, as shown in the following figure.



Longitudinal Analysis

In a longitudinal analysis, you might measure responses on n subjects at d time points, with correlation between observations made on the same subject. For example, suppose that you measure responses y_{ij} at times t_{ij}, $i = 1,...,n$ and $j = 1,...,d$. In addition, suppose that each subject is in one of two groups (such as male or female), specified by the indicator variable G_i. You could model y_{ij} as a function of G_i and t_{ij}, with group-specific intercepts and slopes, as follows:

$$y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 G_i \times t_{ij} + \varepsilon_{ij}, \;\; i = 1, \ldots, n; \;\; j = 1, \ldots, d,$$

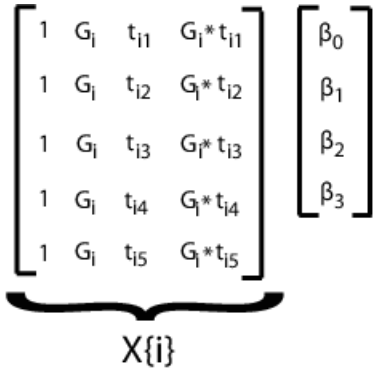
where

$$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})' \sim MVN(\mathbf{0}, \Sigma).$$

Most longitudinal models include time as an explicit predictor.

To fit this model using `mvregress`, arrange the responses in an n-by-d matrix, where n is the number of subjects and d is the number of time points. Specify the design matrices in an n-length cell array of d-by-K matrices, where here $K = 4$ for the four regression coefficients.

For example, suppose $d = 5$ (five observations per subject). The ith design matrix and corresponding parameter vector for the specified model are shown in the following figure.



Panel Analysis

In a panel analysis, you might measure responses and covariates on d subjects (such as individuals or countries) at n time points. For example, suppose you measure responses y_{tj} and covariates x_{tj} on subjects $j = 1,...,d$ at times $t = 1,...,n$. A fixed effects panel model, with subject-specific fixed effects, and concurrent correlation might look like:

$$y_{tj} = \alpha_j + \beta x_{tj} + \varepsilon_{tj},$$

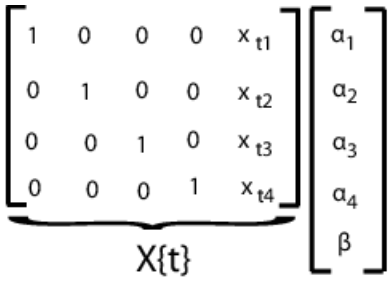
where

$$\varepsilon_t = (\varepsilon_{t1}, \ldots, \varepsilon_{td})' \sim MVN(\mathbf{0}, \Sigma).$$

In contrast to longitudinal models, the panel analysis model typically includes covariates measured at each time point, instead of using time as an explicit predictor.

To fit this model using `mvregress`, arrange the responses in an n-by-d matrix, such that each column corresponds to a subject. Specify the design matrices in an n-length cell array of d-by-K matrices, where here $K = d + 1$ for the d intercepts and a slope term.

For example, suppose $d = 4$ (four subjects). The tth design matrix and corresponding parameter vector are shown in the following figure.



Seemingly Unrelated Regression

In a seemingly unrelated regression (SUR), you model d separate regressions, each with its own intercept and slope, but a common error variance-covariance matrix. For example, suppose you measure responses y_{ij} and covariates x_{ij} for regression models $j = 1,...,d$, with $i = 1,...,n$ observations to fit each regression. The SUR model might look like:

$$y_{ij} = \beta_{0j} + \beta_j x_{ij} + \varepsilon_{ij},$$

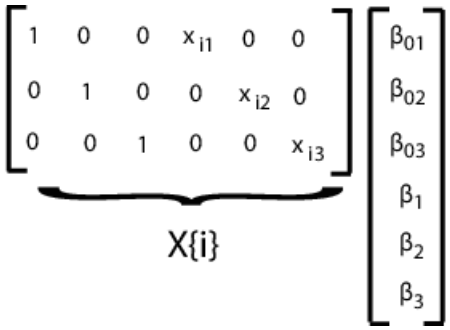
where

$$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})' \sim MVN(\mathbf{0}, \Sigma).$$

This model is very similar to the multivariate general linear model, except that it has different covariates for each dimension.

To fit this model using `mvregress`, arrange the responses in an n-by-d matrix, such that each column has the data for the jth regression model. Specify the design matrices in an n-length cell array of d-by-K matrices, where here $K = 2d$ for d intercepts and d slopes.

For example, suppose $d = 3$ (three regressions). The ith design matrix and corresponding parameter vector are shown in the following figure.



Vector Autoregressive Model

The VAR(p) vector autoregressive model expresses d-dimensional time series responses as a linear function of p lagged d-dimensional responses from previous times. For example, suppose you measure responses y_{tj} for time series $j = 1,...,d$ at times $t = 1,...,n$. The VAR(p) model might look like:

$$\begin{bmatrix} y_{t1} \\ y_{t2} \\ \vdots \\ y_{td} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{bmatrix} + \begin{bmatrix} \varphi_{11}^{(1)} & \varphi_{12}^{(1)} & \cdots & \varphi_{1d}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{d1}^{(1)} & \varphi_{d2}^{(1)} & \cdots & \varphi_{dd}^{(1)} \end{bmatrix} \begin{bmatrix} y_{t-1,1} \\ y_{t-1,2} \\ \vdots \\ y_{t-1,d} \end{bmatrix} + \cdots + \begin{bmatrix} \varphi_{11}^{(p)} & \varphi_{12}^{(p)} & \cdots & \varphi_{1d}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{d1}^{(p)} & \varphi_{d2}^{(p)} & \cdots & \varphi_{dd}^{(p)} \end{bmatrix} \begin{bmatrix} y_{t-p,1} \\ y_{t-p,2} \\ \vdots \\ y_{t-p,d} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \\ \vdots \\ \varepsilon_{td} \end{bmatrix},$$

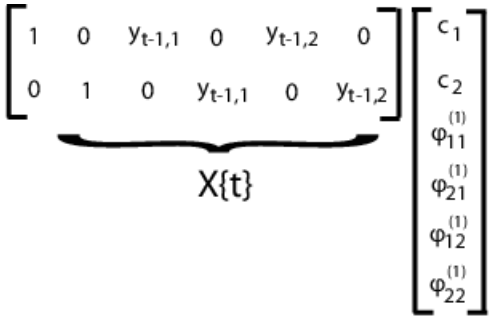
where

$$\varepsilon_t = (\varepsilon_{t1}, \ldots, \varepsilon_{td})' \sim MVN(\mathbf{0}, \Sigma).$$

When estimating vector autoregressive models, you typically need to use the first p observations to initiate the model, or provide some other presample response values.

To fit this model using `mvregress`, arrange the responses in an n-by-d matrix, such that each column corresponds to a time series. Specify the design matrices in an n-length cell array of d-by-K matrices, where here $K = d + pd^2$.

For example, suppose $d = 2$ (two time series) and $p = 1$ (one lag). The tth design matrix and corresponding parameter vector are shown in the following figure.



Alternatively, Econometrics Toolbox has functions for fitting and forecasting VAR(p) models, including the option to specify exogenous predictor variables.

See Also

`mvregress` | `mvregresslike`

Related Examples

- “Multivariate General Linear Model” on page 15-20
- “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24
- “Longitudinal Analysis” on page 15-30

More About

- “Multivariate Linear Regression” on page 15-2
- “Estimation of Multivariate Regression Models” on page 15-5

Multivariate General Linear Model

This example shows how to set up a multivariate general linear model for estimation using `mvregress`.

Load sample data.

This data contains measurements on a sample of 205 auto imports from 1985.

Here, model the bivariate response of city and highway MPG (columns 14 and 15).

For predictors, use wheel base (column 3), curb weight (column 7), and fuel type (column 18). The first two predictors are continuous, and for this example are centered and scaled. Fuel type is a categorical variable with two categories (`11` and `20`), so a dummy indicator variable is needed for the regression.

```
load('imports-85')
Y = X(:,14:15);
[n,d] = size(Y);

X1 = zscore(X(:,3));
X2 = zscore(X(:,7));
X3 = X(:,18)==20;

Xmat = [ones(n,1) X1 X2 X3];
```

The variable `X3` is coded to have value `1` for the fuel type 20, and value `0` otherwise.

For convenience, the three predictors (wheel base, curb weight, and fuel type indicator) are combined into one design matrix, with an added intercept term.

Set up design matrices.

Given these predictors, the multivariate general linear model for the bivariate MPG response is

$$\begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \\ \vdots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} \end{bmatrix},$$

where $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2})' - MVN(0, \Sigma)$. There are $K = 8$ regression coefficients in total.

Create a length $n = 205$ cell array of 2-by-8 (d-by-K) matrices for use with `mvregress`. The ith matrix in the cell array is

$$X(i) = \begin{bmatrix} 1 & 0 & x_{i1} & 0 & x_{i2} & 0 & x_{i3} & 0 \\ 0 & 1 & 0 & x_{i1} & 0 & x_{i2} & 0 & x_{i3} \end{bmatrix}.$$

```
Xcell = cell(1,n);
for i = 1:n
    Xcell{i} = [kron([Xmat(i,:)],eye(d))];
end
```

Given this specification of the design matrices, the corresponding parameter vector is

$$\beta = \begin{bmatrix} \beta_{01} \\ \beta_{02} \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \\ \beta_{31} \\ \beta_{32} \end{bmatrix}.$$

Estimate regression coefficients.

Fit the model using maximum likelihood estimation.

```
[beta,sigma,E,V] = mvregress(Xcell,Y);
beta
```

```
beta = 8×1

   33.5476
   38.5720
    0.9723
    0.3950
   -6.3064
   -6.3584
   -9.2284
   -8.6663
```

These coefficient estimates show:

- The expected city and highway MPG for cars of average wheel base, curb weight, and fuel type 11 are `33.5` and `38.6`, respectively. For fuel type 20, the expected city and highway MPG are `33.5476 - 9.2284 = 24.3192` and `38.5720 - 8.6663 = 29.9057`.
- An increase of one standard deviation in curb weight has almost the same effect on expected city and highway MPG. Given all else is equal, the expected MPG decreases by about `6.3` with each one standard deviation increase in curb weight, for both city and highway MPG.
- For each one standard deviation increase in wheel base, the expected city MPG increases `0.972`, while the expected highway MPG increases by only `0.395`, given all else is equal.

Compute standard errors.

The standard errors for the regression coefficients are the square root of the diagonal of the variance-covariance matrix, `V`.

```
se = sqrt(diag(V))
```

```
se = 8×1

    0.7365
    0.7599
    0.3589
    0.3702
    0.3497
```

```
    0.3608
    0.7790
    0.8037
```

Reshape coefficient matrix.

You can easily reshape the regression coefficients into the original 4-by-2 matrix.

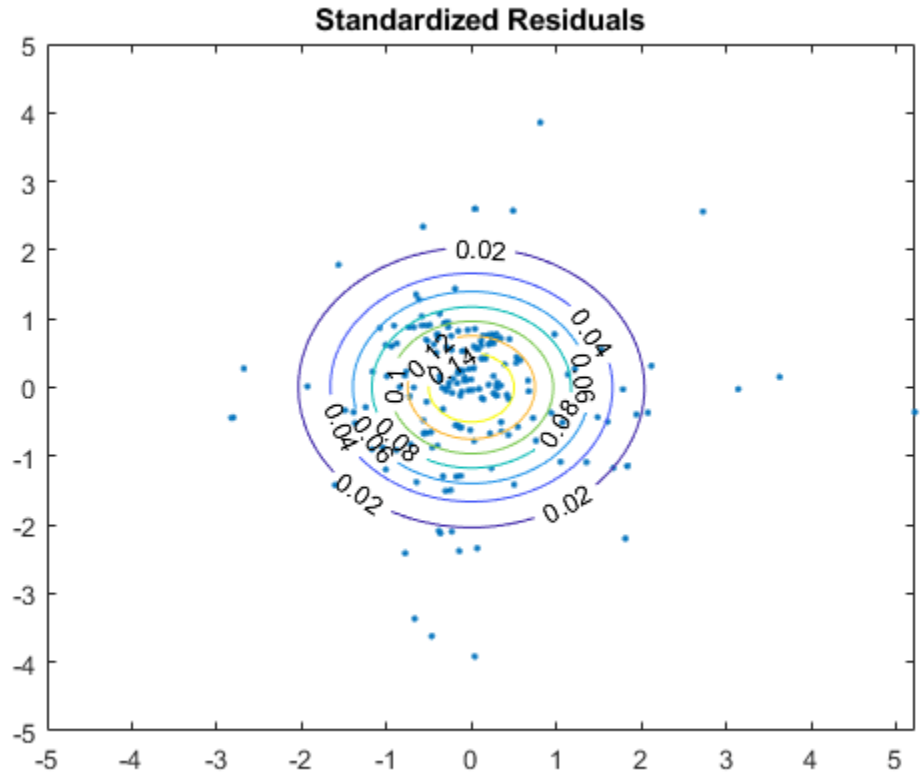
```
B = reshape(beta,2,4)'
```

```
B = 4×2

   33.5476   38.5720
    0.9723    0.3950
   -6.3064   -6.3584
   -9.2284   -8.6663
```

Check model assumptions.

Under the model assumptions, $z = \mathbf{E}\Sigma^{-1/2}$ should be independent, with a bivariate standard normal distribution. In this 2-D case, you can assess the validity of this assumption using a scatter plot.

```
z = E/chol(sigma);
figure()
plot(z(:,1),z(:,2),'.')
title('Standardized Residuals')
hold on

% Overlay standard normal contours
z1 = linspace(-5,5);
z2 = linspace(-5,5);
[zx,zy] = meshgrid(z1,z2);
zgrid = [reshape(zx,100^2,1),reshape(zy,100^2,1)];
zn = reshape(mvnpdf(zgrid),100,100);
[c,h] = contour(zx,zy,zn);
clabel(c,h)
```



Several residuals are larger than expected, but overall, there is little evidence against the multivariate normality assumption.

See Also

`mvregress` | `mvregresslike`

Related Examples

- “Set Up Multivariate Regression Problems” on page 15-11
- “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24
- “Longitudinal Analysis” on page 15-30

More About

- “Multivariate Linear Regression” on page 15-2
- “Estimation of Multivariate Regression Models” on page 15-5

Fixed Effects Panel Model with Concurrent Correlation

This example shows how to perform panel data analysis using `mvregress`. First, a fixed effects model with concurrent correlation is fit by ordinary least squares (OLS) to some panel data. Then, the estimated error covariance matrix is used to get panel corrected standard errors for the regression coefficients.

Load sample data.

Load the sample panel data.

```
load panelData
```

The dataset array, `panelData`, contains yearly observations on eight cities for 6 years. This is simulated data.

Define variables.

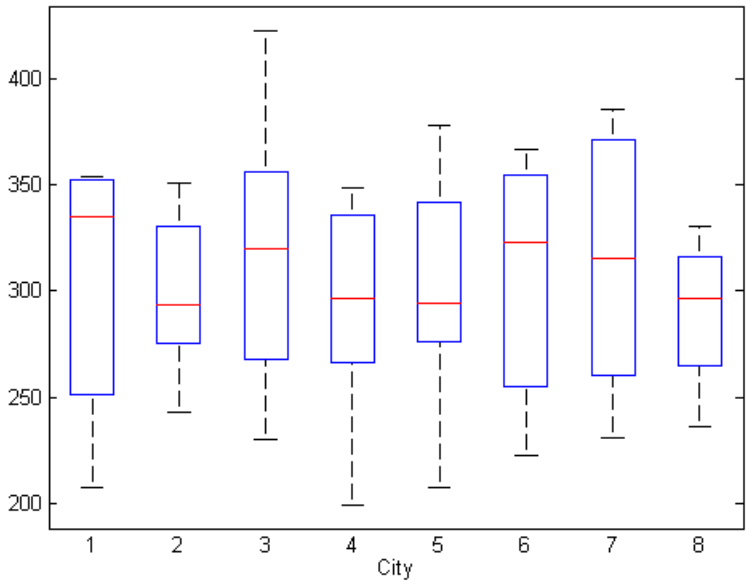
The first variable, `Growth`, measures economic growth (the response variable). The second and third variables are city and year indicators, respectively. The last variable, `Employ`, measures employment (the predictor variable).

```
y = panelData.Growth;
city = panelData.City;
year = panelData.Year;
x = panelData.Employ;
```

Plot data grouped by category.

To look for potential city-specific fixed effects, create a box plot of the response grouped by city.

```
figure()
boxplot(y,city)
xlabel('City')
```

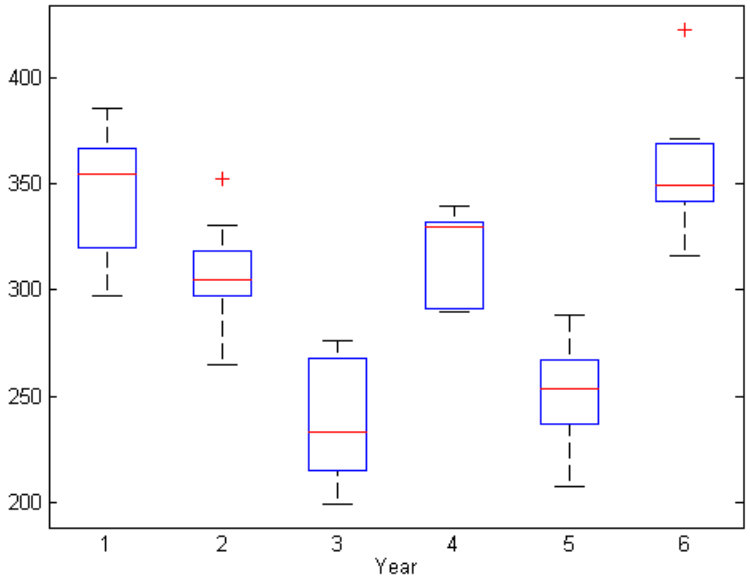


There does not appear to be any systematic differences in the mean response among cities.

Plot data grouped by a different category.

To look for potential year-specific fixed effects, create a box plot of the response grouped by year.

```
figure()
boxplot(y,year)
xlabel('Year')
```



Some evidence of systematic differences in the mean response between years seems to exist.

Format response data.

Let y_{ij} denote the response for city $j = 1,...,d$, in year $i = 1,...,n$. Similarly, x_{ij} is the corresponding value of the predictor variable. In this example, $n = 6$ and $d = 8$.

Consider fitting a year-specific fixed effects model with a constant slope and concurrent correlation among cities in the same year,

$$y_{ij} = \alpha_i + \beta_1 x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \; j = 1, \ldots, d,$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})' \sim MVN(\mathbf{0}, \Sigma)$. The concurrent correlation accounts for any unmeasured, time-static factors that might impact growth similarly for some cities. For example, cities with close spatial proximity might be more likely to have similar economic growth.

To fit this model using `mvregress`, reshape the response data into an n-by-d matrix.

```
n = 6; d = 8;
Y = reshape(y,n,d);
```

Format design matrices.

Create a length-n cell array of d-by-K design matrices. For this model, there are $K = 7$ parameters ($d = 6$ intercept terms and a slope).

Suppose the vector of parameters is arranged as

$$\beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_6 \\ \beta_1 \end{pmatrix}.$$

In this case, the first design matrix for year 1 looks like

$$X\left\{1\right\} = \begin{pmatrix} 1 & 0 & \cdots & 0 & x_{11} \\ 1 & 0 & \cdots & 0 & x_{12} \\ \vdots & \vdots & \cdots & 0 & \vdots \\ 1 & 0 & \cdots & 0 & x_{18} \end{pmatrix},$$

and the second design matrix for year 2 looks like

$$X\left\{2\right\} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & x_{21} \\ 0 & 1 & 0 & \cdots & 0 & x_{22} \\ \vdots & \vdots & 0 & \cdots & 0 & \vdots \\ 0 & 1 & 0 & \cdots & 0 & x_{28} \end{pmatrix}.$$

The design matrices for the remaining 4 years are similar.

```
K = 7; N = n*d;
X = cell(n,1);
for i = 1:n
    x0 = zeros(d,K-1);
    x0(:,i) = 1;
    X{i} = [x0,x(i:n:N)];
end
```

Fit the model.

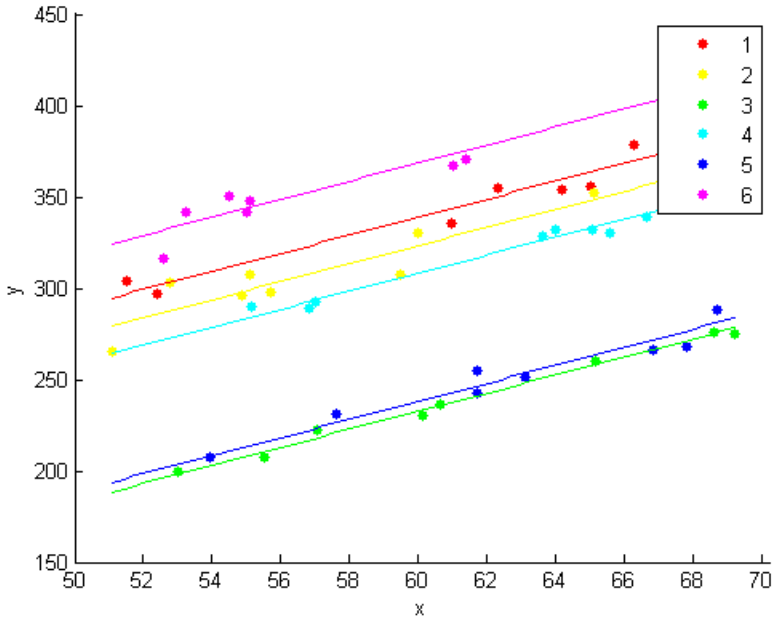
Fit the model using ordinary least squares (OLS).

```
[b,sig,E,V] = mvregress(X,Y,'algorithm','cwls');
b
```

```
b =

   41.6878
   26.1864
  -64.5107
   11.0924
  -59.1872
   71.3313
    4.9525
```

Plot fitted model.

```
xx = linspace(min(x),max(x));
axx = repmat(b(1:K-1),1,length(xx));
bxx = repmat(b(K)*xx,n,1);
yhat =  axx + bxx;

figure()
hPoints = gscatter(x,y,year);
hold on
hLines = plot(xx,yhat);
for i=1:n
   set(hLines(i),'color',get(hPoints(i),'color'));
end
hold off
```

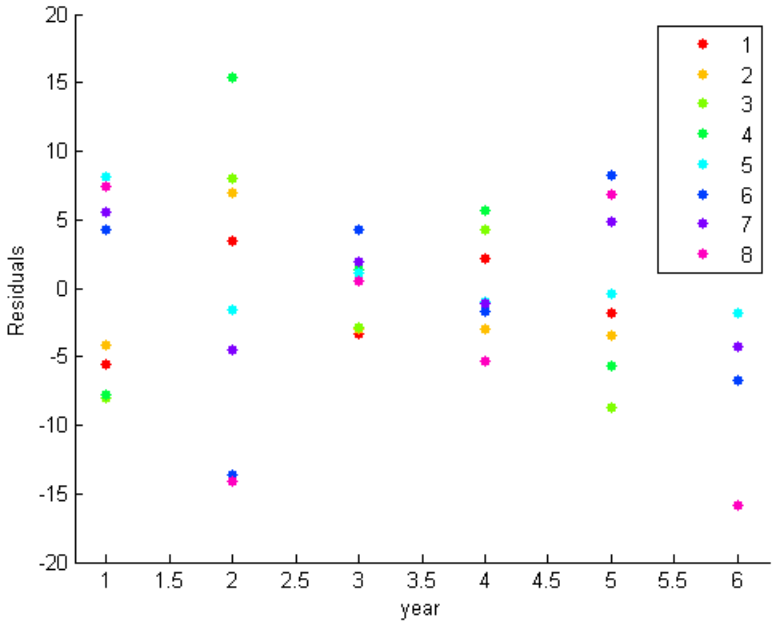


The model with year-specific intercepts and common slope appears to fit the data quite well.

Residual correlation.

Plot the residuals, grouped by year.

```
figure()
gscatter(year,E(:),city)
ylabel('Residuals')
```



The residual plot suggests concurrent correlation is present. For examples, cities 1, 2, 3, and 4 are consistently above or below average as a group in any given year. The same is true for the collection of cities 5, 6, 7, and 8. As seen in the exploratory plots, there are no systematic city-specific effects.

Panel corrected standard errors.

Use the estimated error variance-covariance matrix to compute panel corrected standard errors for the regression coefficients.

```
XX = cell2mat(X);
S = kron(eye(n),sig);
Vpcse = inv(XX'*XX)*XX'*S*XX*inv(XX'*XX);
se = sqrt(diag(Vpcse))
```

```
se =

    9.3750
    8.6698
    9.3406
    9.4286
    9.5729
    8.8207
    0.1527
```

See Also

`mvregress` | `mvregresslike`

Related Examples

- “Set Up Multivariate Regression Problems” on page 15-11
- “Multivariate General Linear Model” on page 15-20

- “Longitudinal Analysis” on page 15-30

More About

- “Multivariate Linear Regression” on page 15-2
- “Estimation of Multivariate Regression Models” on page 15-5

Longitudinal Analysis

This example shows how to perform longitudinal analysis using `mvregress`.

Load sample data.

Load the sample longitudinal data.

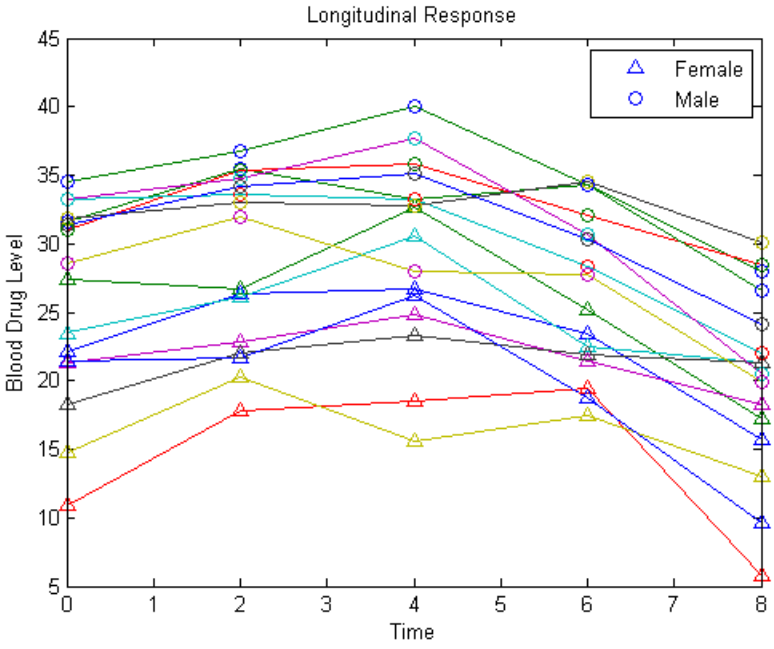
```
load longitudinalData
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (t = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual, and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Plot data.

Plot the data for all 16 subjects.

```
figure()
t = [0,2,4,6,8];
plot(t,Y)
hold on

hf = plot(t,Y(1:8,:),'^');
hm = plot(t,Y(9:16,:),'o');
legend([hf(1),hm(1)],'Female','Male','Location','NorthEast')

title('Longitudinal Response')
ylabel('Blood Drug Level')
xlabel('Time')
hold off
```



Define design matrices.

Let y_{ij} denote the response for individual $i = 1,...,n$ measured at times t_{ij}, $j = 1,...,d$. In this example, $n = 16$ and $d = 5$. Let G_i denote the gender of individual i, where $G_i = 1$ for males and 0 for females.

Consider fitting a quadratic longitudinal model, with a separate slope and intercept for each gender,

$$y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 G_i \times t_{ij} + \beta_5 G_i \times t_{ij}^2 + \varepsilon_{ij},$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})' \sim MVN(\mathbf{0}, \Sigma)$. The error correlation accounts for clustering within an individual.

To fit this model using `mvregress`, the response data should be in an n-by-d matrix. `Y` is already in the proper format.

Next, create a length-n cell array of d-by-K design matrices. For this model, there are $K = 6$ parameters.

For individual i, the 5-by-6 design matrix is

$$X\left\{i\right\} = \begin{pmatrix} 1 & G_i & t_{i1} & t_{i1}^2 & G_i \times t_{i1} & G_i \times t_{i1}^2 \\ 1 & G_i & t_{i2} & t_{i2}^2 & G_i \times t_{i2} & G_i \times t_{i2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & G_i & t_{i5} & t_{i5}^2 & G_i \times t_{i5} & G_i \times t_{i5}^2 \end{pmatrix},$$

corresponding to the parameter vector

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_5 \end{pmatrix}.$$

The matrix `X1` has the design matrix for a female, and `X2` has the design matrix for a male.

Create a cell array of design matrices. The first eight individuals are females, and the second eight are males.

```
X = cell(8,1);
X(1:8) = {X1};
X(9:16) = {X2};
```

Fit the model.

Fit the model using maximum likelihood estimation. Display the estimated coefficients and standard errors.

```
[b,sig,E,V,loglikF] = mvregress(X,Y);
[b sqrt(diag(V))]
```

```
ans =

   18.8619    0.7432
   13.0942    1.0511
```

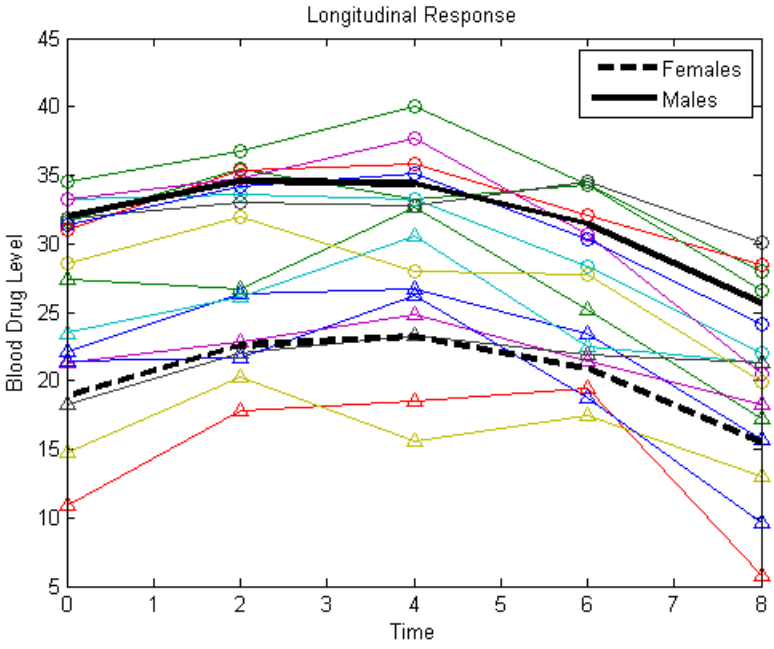
```
    2.5968    0.2845
   -0.3771    0.0398
   -0.5929    0.4023
    0.0290    0.0563
```

The coefficients on the interaction terms (in the last two rows of `b`) do not appear significant. You can use the value of the loglikelihood objective function for this fit, `loglikF`, to compare this model to one without the interaction terms using a likelihood ratio test.

Plot fitted model.

Plot the fitted lines for females and males.

```
Yhatf = X1*b;
Yhatm = X2*b;

figure()
plot(t,Y)
hold on

plot(t,Y(1:8,:),'^',t,Y(9:16,:),'o')

hf = plot(t,Yhatf,'k--','LineWidth',3);
hm = plot(t,Yhatm,'k','LineWidth',3);
legend([hf,hm],'Females','Males','Location','NorthEast')

title('Longitudinal Response')
ylabel('Blood Drug Level')
xlabel('Time')
hold off
```



Define a reduced model.

Fit the model without interaction terms,

$$y_{ij} = \beta_0 + \beta_1 G_i + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \varepsilon_{ij},$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})' \sim MVN(\mathbf{0}, \Sigma)$.

This model has four coefficients, which correspond to the first four columns of the design matrices `X1` and `X2` (for females and males, respectively).

```
X1R = X1(:,1:4);
X2R = X2(:,1:4);

XR = cell(8,1);
XR(1:8) = {X1R};
XR(9:16) = {X2R};
```

Fit the reduced model.

Fit this model using maximum likelihood estimation. Display the estimated coefficients and their standard errors.

```
[bR,sigR,ER,VR,loglikR] = mvregress(XR,Y);
[bR,sqrt(diag(VR))]
```

```
ans =

   19.3765    0.6898
   12.0936    0.8591
    2.2919    0.2139
   -0.3623    0.0283
```

Conduct a likelihood ratio test.

Compare the two models using a likelihood ratio test. The null hypothesis is that the reduced model is sufficient. The alternative is that the reduced model is inadequate (compared to the full model with the interaction terms).

The likelihood ratio test statistic is compared to a chi-squared distribution with two degrees of freedom (for the two coefficients being dropped).

```
LR = 2*(loglikF-loglikR);
pval = 1 - chi2cdf(LR,2)
```

```
pval =

    0.0803
```

The *p*-value `0.0803` indicates that the null hypothesis is not rejected at the 5% significance level. Therefore, there is insufficient evidence that the extra terms improve the fit.

See Also

`mvregress` | `mvregresslike`

Related Examples

- “Set Up Multivariate Regression Problems” on page 15-11
- “Multivariate General Linear Model” on page 15-20

- “Fixed Effects Panel Model with Concurrent Correlation” on page 15-24

More About

- “Multivariate Linear Regression” on page 15-2
- “Estimation of Multivariate Regression Models” on page 15-5

Multidimensional Scaling

One of the most important goals in visualizing data is to get a sense of how near or far points are from each other. Often, you can do this with a scatter plot. However, for some analyses, the data that you have might not be in the form of points at all, but rather in the form of pairwise similarities or dissimilarities between cases, observations, or subjects. There are no points to plot.

Even if your data are in the form of points rather than pairwise distances, a scatter plot of those data might not be useful. For some kinds of data, the relevant way to measure how near two points are might not be their Euclidean distance. While scatter plots of the raw data make it easy to compare Euclidean distances, they are not always useful when comparing other kinds of inter-point distances, city block distance for example, or even more general dissimilarities. Also, with a large number of variables, it is very difficult to visualize distances unless the data can be represented in a small number of dimensions. Some sort of dimension reduction is usually necessary.

Multidimensional scaling (MDS) is a set of methods that address all these problems. MDS allows you to visualize how near points are to each other for many kinds of distance or dissimilarity metrics and can produce a representation of your data in a small number of dimensions. MDS does not require raw data, but only a matrix of pairwise distances or dissimilarities.

See Also

`cmdscale` | `mdscale`

Related Examples

- “Nonclassical and Nonmetric Multidimensional Scaling” on page 15-36
- “Classical Multidimensional Scaling” on page 15-40

Nonclassical and Nonmetric Multidimensional Scaling

In this section...

Perform nonclassical multidimensional scaling using `mdscale`.

Nonclassical Multidimensional Scaling

The function `mdscale` performs nonclassical multidimensional scaling. As with `cmdscale`, you use `mdscale` either to visualize dissimilarity data for which no “locations” exist, or to visualize high-dimensional data by reducing its dimensionality. Both functions take a matrix of dissimilarities as an input and produce a configuration of points. However, `mdscale` offers a choice of different criteria to construct the configuration, and allows missing data and weights.

For example, the cereal data include measurements on 10 variables describing breakfast cereals. You can use `mdscale` to visualize these data in two dimensions. First, load the data. For clarity, this example code selects a subset of 22 of the observations.

```
load cereal.mat
X = [Calories Protein Fat Sodium Fiber ...
    Carbo Sugars Shelf Potass Vitamins];
% Take a subset from a single manufacturer
mfg1 = strcmp('G',cellstr(Mfg));
X = X(mfg1,:);
size(X)
ans =
    22 10
```

Then use `pdist` to transform the 10-dimensional data into dissimilarities. The output from `pdist` is a symmetric dissimilarity matrix, stored as a vector containing only the (23*22/2) elements in its upper triangle.

```
dissimilarities = pdist(zscore(X),'cityblock');
size(dissimilarities)
ans =
     1   231
```

This example code first standardizes the cereal data, and then uses city block distance as a dissimilarity. The choice of transformation to dissimilarities is application-dependent, and the choice here is only for simplicity. In some applications, the original data are already in the form of dissimilarities.

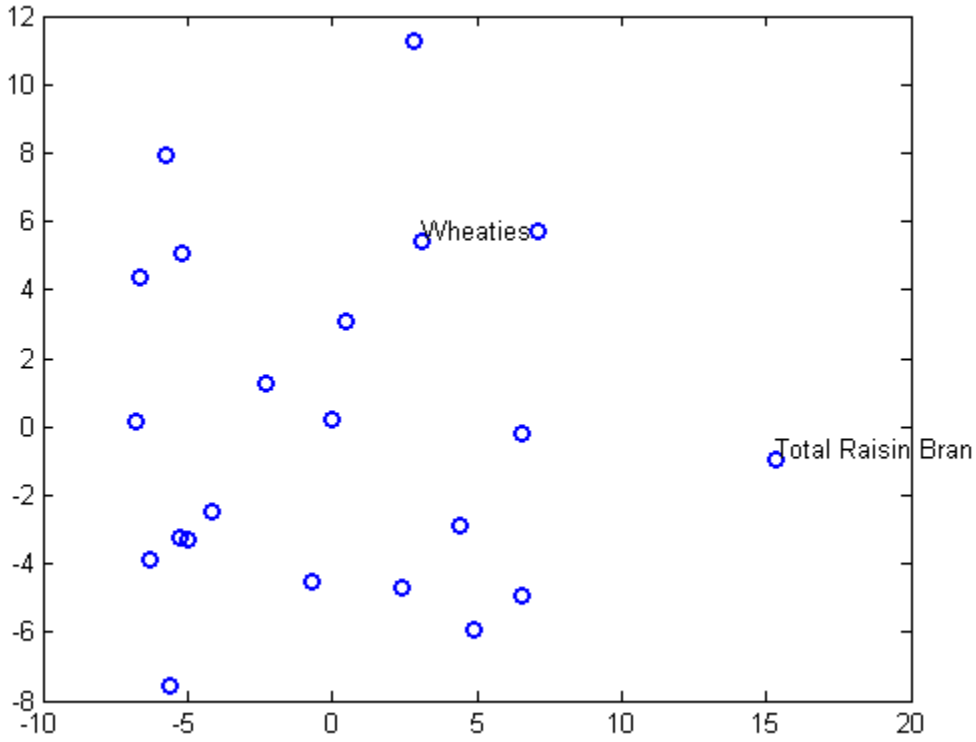
Next, use `mdscale` to perform metric MDS. Unlike `cmdscale`, you must specify the desired number of dimensions, and the method to use to construct the output configuration. For this example, use two dimensions. The metric STRESS criterion is a common method for computing the output; for other choices, see the `mdscale` reference page in the online documentation. The second output from `mdscale` is the value of that criterion evaluated for the output configuration. It measures the how well the inter-point distances of the output configuration approximate the original input dissimilarities:

```
[Y,stress] =...
mdscale(dissimilarities,2,'criterion','metricstress');
```

```
stress
stress =
    0.1856
```

A scatterplot of the output from `mdscale` represents the original 10-dimensional data in two dimensions, and you can use the `gname` function to label selected points:

```
plot(Y(:,1),Y(:,2),'o','LineWidth',2);
gname(Name(mfg1))
```



Nonmetric Multidimensional Scaling

Metric multidimensional scaling creates a configuration of points whose inter-point distances approximate the given dissimilarities. This is sometimes too strict a requirement, and non-metric scaling is designed to relax it a bit. Instead of trying to approximate the dissimilarities themselves, non-metric scaling approximates a nonlinear, but monotonic, transformation of them. Because of the monotonicity, larger or smaller distances on a plot of the output will correspond to larger or smaller dissimilarities, respectively. However, the nonlinearity implies that `mdscale` only attempts to preserve the ordering of dissimilarities. Thus, there may be contractions or expansions of distances at different scales.

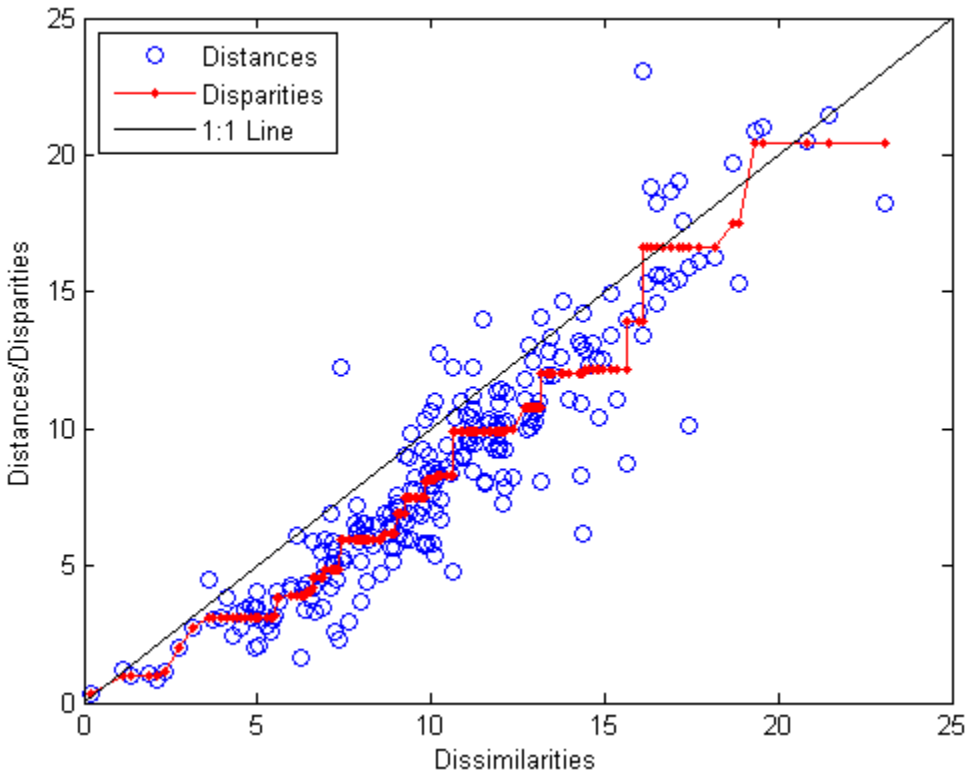
You use `mdscale` to perform nonmetric MDS in much the same way as for metric scaling. The nonmetric STRESS criterion is a common method for computing the output; for more choices, see the `mdscale` reference page in the online documentation. As with metric scaling, the second output from `mdscale` is the value of that criterion evaluated for the output configuration. For nonmetric scaling,

however, it measures the how well the inter-point distances of the output configuration approximate the disparities. The disparities are returned in the third output. They are the transformed values of the original dissimilarities:

```
[Y,stress,disparities] = ...
mdscale(dissimilarities,2,'criterion','stress');
stress
stress =
    0.1562
```

To check the fit of the output configuration to the dissimilarities, and to understand the disparities, it helps to make a Shepard plot:

```
distances = pdist(Y);
[dum,ord] = sortrows([disparities(:) dissimilarities(:)]);
plot(dissimilarities,distances,'bo', ...
     dissimilarities(ord),disparities(ord),'r.-', ...
     [0 25],[0 25],'k-')
xlabel('Dissimilarities')
ylabel('Distances/Disparities')
legend({'Distances' 'Disparities' '1:1 Line'},...
       'Location','NorthWest');
```



This plot shows that `mdscale` has found a configuration of points in two dimensions whose inter-point distances approximates the disparities, which in turn are a nonlinear transformation of the original dissimilarities. The concave shape of the disparities as a function of the dissimilarities indicates that

fit tends to contract small distances relative to the corresponding dissimilarities. This may be perfectly acceptable in practice.

`mdscale` uses an iterative algorithm to find the output configuration, and the results can often depend on the starting point. By default, `mdscale` uses `cmdscale` to construct an initial configuration, and this choice often leads to a globally best solution. However, it is possible for `mdscale` to stop at a configuration that is a local minimum of the criterion. Such cases can be diagnosed and often overcome by running `mdscale` multiple times with different starting points. You can do this using the `'start'` and `'replicates'` name-value pair arguments. The following code runs five replicates of MDS, each starting at a different randomly-chosen initial configuration. The criterion value is printed out for each replication; `mdscale` returns the configuration with the best fit.

```
opts = statset('Display','final');
[Y,stress] =...
mdscale(dissimilarities,2,'criterion','stress',...
'start','random','replicates',5,'Options',opts);

35 iterations, Final stress criterion = 0.156209
31 iterations, Final stress criterion = 0.156209
48 iterations, Final stress criterion = 0.171209
33 iterations, Final stress criterion = 0.175341
32 iterations, Final stress criterion = 0.185881
```

Notice that `mdscale` finds several different local solutions, some of which do not have as low a stress value as the solution found with the `cmdscale` starting point.

Classical Multidimensional Scaling

This example shows how to use `cmdscale` to perform classical (metric) multidimensional scaling, also known as principal coordinates analysis.

`cmdscale` takes as an input a matrix of inter-point distances and creates a configuration of points. Ideally, those points are in two or three dimensions, and the Euclidean distances between them reproduce the original distance matrix. Thus, a scatter plot of the points created by `cmdscale` provides a visual representation of the original distances.

As a very simple example, you can reconstruct a set of points from only their inter-point distances. First, create some four dimensional points with a small component in their fourth coordinate, and reduce them to distances.

```
rng default;  % For reproducibility
X = [normrnd(0,1,10,3),normrnd(0,.1,10,1)];
D = pdist(X,'euclidean');
```

Next, use `cmdscale` to find a configuration with those inter-point distances. `cmdscale` accepts distances as either a square matrix, or, as in this example, in the vector upper-triangular form produced by `pdist`.

```
[Y,eigvals] = cmdscale(D);
```

`cmdscale` produces two outputs. The first output, `Y`, is a matrix containing the reconstructed points. The second output, `eigvals`, is a vector containing the sorted eigenvalues of what is often referred to as the "scalar product matrix," which, in the simplest case, is equal to `Y*Y'`. The relative magnitudes of those eigenvalues indicate the relative contribution of the corresponding columns of `Y` in reproducing the original distance matrix `D` with the reconstructed points.

```
format short g
[eigvals eigvals/max(abs(eigvals))]
```

```
ans = 10×2

        35.41            1
       11.158      0.31511
       1.6894      0.04771
       0.1436    0.0040553
   5.1216e-15   1.4464e-16
   3.4766e-15   9.8181e-17
   1.9662e-15   5.5527e-17
   1.1713e-15   3.3079e-17
  -1.5864e-15    -4.48e-17
  -3.2498e-15  -9.1776e-17
```

If `eigvals` contains only positive and zero (within round-off error) eigenvalues, the columns of `Y` corresponding to the positive eigenvalues provide an exact reconstruction of `D`, in the sense that their inter-point Euclidean distances, computed using `pdist`, for example, are identical (within round-off) to the values in `D`.

```
maxerr4 = max(abs(D - pdist(Y)))   % Exact reconstruction
```

```
maxerr4 =
   2.6645e-15
```

If two or three of the eigenvalues in `eigvals` are much larger than the rest, then the distance matrix based on the corresponding columns of `Y` nearly reproduces the original distance matrix `D`. In this sense, those columns form a lower-dimensional representation that adequately describes the data. However it is not always possible to find a good low-dimensional reconstruction.

```
maxerr3 = max(abs(D - pdist(Y(:,1:3))))  % Good reconstruction in 3D
```

```
maxerr3 =
     0.043142
```

```
maxerr2 = max(abs(D - pdist(Y(:,1:2))))  % Poor reconstruction in 2D
```

```
maxerr2 =
      0.98315
```

The reconstruction in three dimensions reproduces `D` very well, but the reconstruction in two dimensions has errors that are of the same order of magnitude as the largest values in `D`.

```
max(max(D))
```

```
ans =
       5.8974
```

Often, `eigvals` contains some negative eigenvalues, indicating that the distances in `D` cannot be reproduced exactly. That is, there might not be any configuration of points whose inter-point Euclidean distances are given by `D`. If the largest negative eigenvalue is small in magnitude relative to the largest positive eigenvalues, then the configuration returned by `cmdscale` might still reproduce `D` well.

Procrustes Analysis

In this section...

Compare Landmark Data

The `procrustes` function analyzes the distribution of a set of shapes using Procrustes analysis. This analysis method matches landmark data (geometric locations representing significant features in a given shape) to calculate the best shape-preserving Euclidean transformations. These transformations minimize the differences in location between compared landmark data.

Procrustes analysis is also useful in conjunction with multidimensional scaling. In “Construct a Map Using Multidimensional Scaling” on page 33-647 there is an observation that the orientation of the reconstructed points is arbitrary. Two different applications of multidimensional scaling could produce reconstructed points that are very similar in principle, but that look different because they have different orientations. The `procrustes` function transforms one set of points to make them more comparable to the other.

Data Input

The `procrustes` function takes two matrices as input:

- The target shape matrix X has dimension $n \times p$, where n is the number of landmarks in the shape and p is the number of measurements per landmark.
- The comparison shape matrix Y has dimension $n \times q$ with $q \le p$. If there are fewer measurements per landmark for the comparison shape than the target shape ($q < p$), the function adds columns of zeros to Y, yielding an $n \times p$ matrix.

The equation to obtain the transformed shape, Z, is

$$Z = bYT + c \tag{15-1}$$

where:

- b is a scaling factor that stretches ($b > 1$) or shrinks ($b < 1$) the points.
- T is the orthogonal rotation and reflection matrix.
- c is a matrix with constant values in each column, used to shift the points.

The `procrustes` function chooses b, T, and c to minimize the distance between the target shape X and the transformed shape Z as measured by the least squares criterion:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(X_{ij} - Z_{ij})^2$$

Preprocess Data for Accurate Results

Procrustes analysis is appropriate when all `p` measurement dimensions have similar scales. The analysis would be inaccurate, for example, if the columns of Z had different scales:

- The first column is measured in milliliters ranging from 2,000 to 6,000.
- The second column is measured in degrees Celsius ranging from 10 to 25.
- The third column is measured in kilograms ranging from 50 to 230.

In such cases, standardize your variables by:

1. Subtracting the sample mean from each variable.
2. Dividing each resultant variable by its sample standard deviation.

Use the `zscore` function to perform this standardization.

See Also

`procrustes`

Related Examples

- “Compare Handwritten Shapes Using Procrustes Analysis” on page 15-44

Compare Handwritten Shapes Using Procrustes Analysis

This example shows how to use Procrustes analysis to compare two handwritten number threes. Visually and analytically explore the effects of forcing size and reflection changes.

Load and Display the Original Data

Input landmark data for two handwritten number threes.

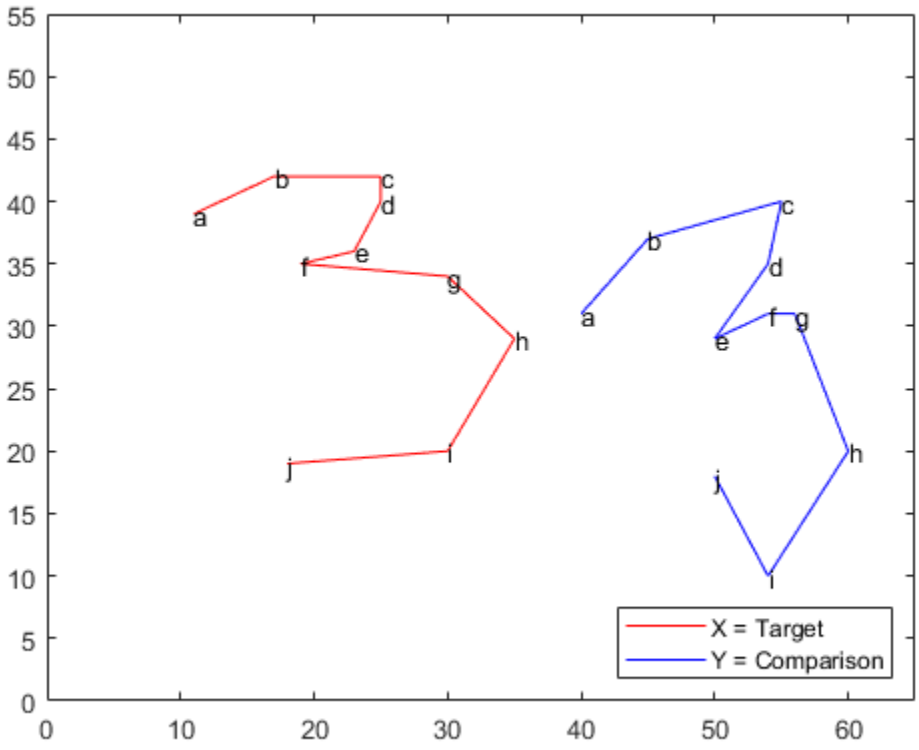
```
A = [11 39;17 42;25 42;25 40;23 36;19 35;30 34;35 29;...
30 20;18 19];
B = [15 31;20 37;30 40;29 35;25 29;29 31;31 31;35 20;...
29 10;25 18];
```

Create X and Y from A and B , moving B to the side to make each shape more visible.

```
X = A;
Y = B + repmat([25 0], 10,1);
```

Plot the shapes, using letters to designate the landmark points. Lines in the figure join the points to indicate the drawing path of each shape.

```
plot(X(:,1), X(:,2),'r-', Y(:,1), Y(:,2),'b-');
text(X(:,1), X(:,2),('abcdefghij')')
text(Y(:,1), Y(:,2),('abcdefghij')')
legend('X = Target','Y = Comparison','location','SE')
xlim([0 65]);
ylim([0 55]);
```



Calculate the Best Transformation

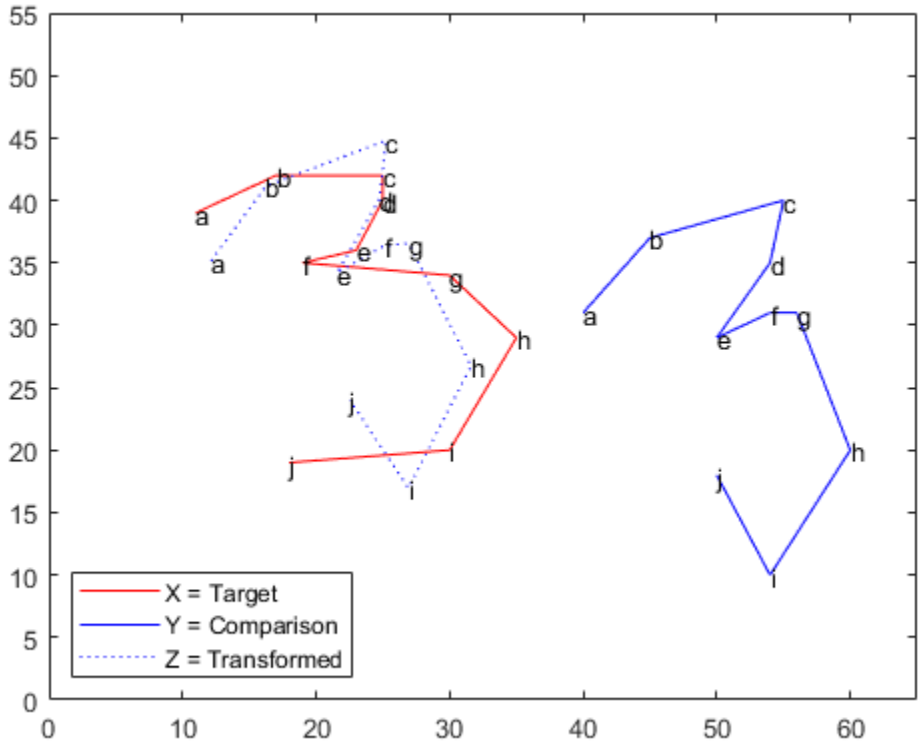
Use Procrustes analysis to find the transformation that minimizes distances between landmark data points.

```
[d,Z,tr] = procrustes(X,Y);
```

The outputs of the function are `d` (a standardized dissimilarity measure), `Z` (a matrix of the transformed landmarks), and `tr` (a structure array of the computed transformation with fields `T` , `b` , and `c` which correspond to the transformation equation).

Visualize the transformed shape, `Z` , using a dashed blue line.

```
plot(X(:,1), X(:,2),'r-', Y(:,1), Y(:,2),'b-',...
Z(:,1),Z(:,2),'b:');
text(X(:,1), X(:,2),('abcdefghij')')
text(Y(:,1), Y(:,2),('abcdefghij')')
text(Z(:,1), Z(:,2),('abcdefghij')')
legend('X = Target','Y = Comparison',...
'Z = Transformed','location','SW')
xlim([0 65]);
ylim([0 55]);
```



Examine the Similarity of the Two Shapes

Use two different numerical values, the dissimilarity measure `d` and the scaling measure `b` , to assess the similarity of the target shape and the transformed shape.

The dissimilarity measure d gives a number between 0 and 1 describing the difference between the target shape and the transformed shape. Values near 0 imply more similar shapes, while values near 1 imply dissimilarity.

```
d
```

```
d = 0.1502
```

The small value of d in this case shows that the two shapes are similar. `procrustes` calculates d by comparing the sum of squared deviations between the set of points with the sum of squared deviations of the original points from their column means.

```
numerator = sum(sum((X-Z).^2))
```

```
numerator = 166.5321
```

```
denominator = sum(sum(bsxfun(@minus,X,mean(X)).^2))
```

```
denominator = 1.1085e+03
```

```
ratio = numerator/denominator
```

```
ratio = 0.1502
```

The resulting measure d is independent of the scale of the size of the shapes and takes into account only the similarity of landmark data.

Examine the size similarity of the shapes.

```
tr.b
```

```
ans = 0.9291
```

The sizes of the target and comparison shapes in the previous figure appear similar. This visual impression is reinforced by the value of b = % 0.93, which implies that the best transformation results in shrinking the comparison shape by a factor .93 (only 7%).

Restrict the Form of the Transformations

Explore the effects of manually adjusting the scaling and reflection coefficients.

Force b to equal 1 (set `'Scaling'` to false) to examine the amount of dissimilarity in size of the target and transformed figures.

```
ds = procrustes(X,Y,'Scaling',false)
```

```
ds = 0.1552
```

In this case, setting `'Scaling'` to false increases the calculated value of d only 0.0049, which further supports the similarity in the size of the two number threes. A larger increase in d would have indicated a greater size discrepancy.

This example requires only a rotation, not a reflection, to align the shapes. You can show this by observing that the determinant of the matrix T is 1 in this analysis.

```
det(tr.T)
```

```
ans = 1.0000
```

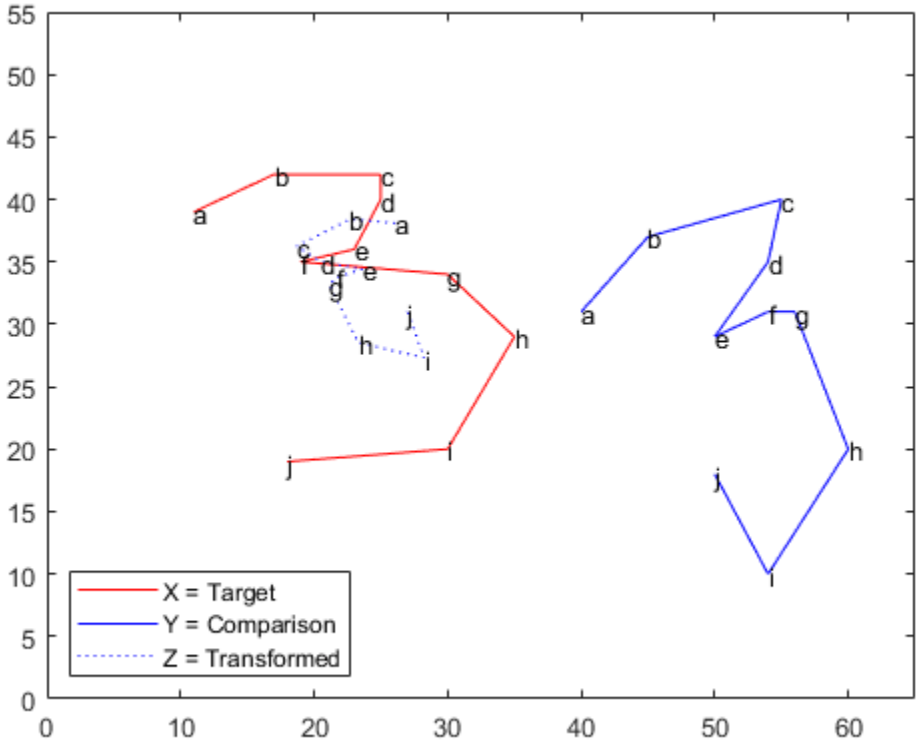
If you need a reflection in the transformation, the determinant of T is -1. You can force a reflection into the transformation as follows.

```
[dr,Zr,trr] = procrustes(X,Y,'Reflection',true);
dr
```

```
dr = 0.8130
```

The d value increases dramatically, indicating that a forced reflection leads to a poor transformation of the landmark points. A plot of the transformed shape shows a similar result.

```
plot(X(:,1), X(:,2),'r-', Y(:,1), Y(:,2),'b-',...
Zr(:,1),Zr(:,2),'b:');
text(X(:,1), X(:,2),('abcdefghij')')
text(Y(:,1), Y(:,2),('abcdefghij')')
text(Zr(:,1), Zr(:,2),('abcdefghij')')
legend('X = Target','Y = Comparison',...
'Z = Transformed','Location','SW')
xlim([0 65]);
ylim([0 55]);
```



The landmark data points are now further away from their target counterparts. The transformed three is now an undesirable mirror image of the target three.

It appears that the shapes might be better matched if you flipped the transformed shape upside down. Flipping the shapes would make the transformation even worse, however, because the

landmark data points would be further away from their target counterparts. From this example, it is clear that manually adjusting the scaling and reflection parameters is generally not optimal.

See Also

`procrustes`

More About

- “Procrustes Analysis” on page 15-42

Introduction to Feature Selection

This topic provides an introduction to feature selection algorithms and describes the feature selection functions available in Statistics and Machine Learning Toolbox.

Feature Selection Algorithms

Feature selection reduces the dimensionality of data by selecting only a subset of measured features (predictor variables) to create a model. Feature selection algorithms search for a subset of predictors that optimally models measured responses, subject to constraints such as required or excluded features and the size of the subset. The main benefits of feature selection are to improve prediction performance, provide faster and more cost-effective predictors, and provide a better understanding of the data generation process [1]. Using too many features can degrade prediction performance even when all features are relevant and contain information about the response variable.

You can categorize feature selection algorithms into three types:

- "Filter Type Feature Selection" on page 15-50 — The filter type feature selection algorithm measures feature importance based on the characteristics of the features, such as feature variance and feature relevance to the response. You select important features as part of a data preprocessing step and then train a model using the selected features. Therefore, filter type feature selection is uncorrelated to the training algorithm.
- "Wrapper Type Feature Selection" on page 15-53 — The wrapper type feature selection algorithm starts training using a subset of features and then adds or removes a feature using a selection criterion. The selection criterion directly measures the change in model performance that results from adding or removing a feature. The algorithm repeats training and improving a model until its stopping criteria are satisfied.
- "Embedded Type Feature Selection" on page 15-54 — The embedded type feature selection algorithm learns feature importance as part of the model learning process. Once you train a model, you obtain the importance of the features in the trained model. This type of algorithm selects features that work well with a particular learning process.

In addition, you can categorize feature selection algorithms according to whether or not an algorithm ranks features sequentially. The minimum redundancy maximum relevance (MRMR) algorithm and stepwise regression are two examples of the sequential feature selection algorithm. For details, see "Sequential Feature Selection" on page 15-61.

You can compare the importance of predictor variables visually by creating partial dependence plots (PDP) and individual conditional expectation (ICE) plots. For details, see `plotPartialDependence`.

For classification problems, after selecting features, you can train two models (for example, a full model and a model trained with a subset of predictors) and compare the accuracies of the models by using the `compareHoldout`, `testcholdout`, or `testckfold` functions.

Feature selection is preferable to feature transformation when the original features and their units are important and the modeling goal is to identify an influential subset. When categorical features are present, and numerical transformations are inappropriate, feature selection becomes the primary means of dimension reduction.

Feature Selection Functions

Statistics and Machine Learning Toolbox offers several functions for feature selection. Choose the appropriate feature selection function based on your problem and the data types of the features.

Filter Type Feature Selection

Function	Supported Problem	Supported Data Type	Description
`fscchi2`	Classification	Categorical and continuous features	Examine whether each predictor variable is independent of a response variable by using individual chi-square tests, and then rank features using the *p*-values of the chi-square test statistics. For examples, see the function reference page `fscchi2`.
`fscmrmr`	Classification	Categorical and continuous features	Rank features sequentially using the "Minimum Redundancy Maximum Relevance (MRMR) Algorithm" on page 33-2482. For examples, see the function reference page `fscmrmr`.
`fscnca*`	Classification	Continuous features	Determine the feature weights by using a diagonal adaptation of neighborhood component analysis (NCA). This algorithm works best for estimating feature importance for distance-based supervised models that use pairwise distances between observations to predict the response. For details, see the function reference page `fscnca` and these topics: • "Neighborhood Component Analysis (NCA) Feature Selection" on page 15-99 • "Tune Regularization Parameter to Detect Features Using NCA for Classification" on page 15-210

Function	Supported Problem	Supported Data Type	Description
`fsrftest`	Regression	Categorical and continuous features	Examine the importance of each predictor individually using an *F*-test, and then rank features using the *p*-values of the *F*-test statistics. Each *F*-test tests the hypothesis that the response values grouped by predictor variable values are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. For examples, see the function reference page `fsrftest`.
`fsrnca*`	Regression	Continuous features	Determine the feature weights by using a diagonal adaptation of neighborhood component analysis (NCA). This algorithm works best for estimating feature importance for distance-based supervised models that use pairwise distances between observations to predict the response. For details, see the function reference page `fsrnca` and these topics: • “Neighborhood Component Analysis (NCA) Feature Selection” on page 15-99 • “Robust Feature Selection Using NCA for Regression” on page 15-85
`fsulaplacian`	Unsupervised learning	Continuous features	Rank features using the “Laplacian Score” on page 33-2538. For examples, see the function reference page `fsulaplacian`.

Function	Supported Problem	Supported Data Type	Description
`relieff`	Classification and regression	Either all categorical or all continuous features	Rank features using the “ReliefF” on page 33-5461 algorithm for classification and the “RReliefF” on page 33-5462 algorithm for regression. This algorithm works best for estimating feature importance for distance-based supervised models that use pairwise distances between observations to predict the response. For examples, see the function reference page `relieff`.
`sequentialfs`	Classification and regression	Either all categorical or all continuous features	Select features sequentially using a custom criterion. Define a function that measures the characteristics of data to select features, and pass the function handle to the `sequentialfs` function. You can specify sequential forward selection or sequential backward selection by using the `'Direction'` name-value pair argument. `sequentialfs` evaluates the criterion using cross-validation.

*You can also consider `fscnca` and `fsrnca` as embedded type feature selection functions because they return a trained model object and you can use the object functions `predict` and `loss`. However, you typically use these object functions to tune the regularization parameter of the algorithm. After selecting features using the `fscnca` or `fsrnca` function as part of a data preprocessing step, you can apply another classification or regression algorithm for your problem.

Wrapper Type Feature Selection

Function	Supported Problem	Supported Data Type	Description
`sequentialfs`	Classification and regression	Either all categorical or all continuous features	Select features sequentially using a custom criterion. Define a function that implements a supervised learning algorithm or a function that measures performance of a learning algorithm, and pass the function handle to the `sequentialfs` function. You can specify sequential forward selection or sequential backward selection by using the `'Direction'` name-value pair argument. `sequentialfs` evaluates the criterion using cross-validation. For examples, see the function reference page `sequentialfs` and these topics: • “Select Subset of Features with Comparative Predictive Power” on page 15-61 • “Selecting Features for Classifying High-dimensional Data” on page 15-171

Embedded Type Feature Selection

Function	Supported Problem	Supported Data Type	Description
`DeltaPredictor` property of a `ClassificationDiscriminant` model object	Linear discriminant analysis classification	Continuous features	Create a linear discriminant analysis classifier by using `fitcdiscr`. A trained classifier, returned as `ClassificationDiscriminant`, stores the coefficient magnitude in the `DeltaPredictor` property. You can use the values in `DeltaPredictor` as measures of the predictor importance. This classifier uses the two regularization parameters "Gamma and Delta" on page 33-1124 to identify and remove redundant predictors. You can obtain appropriate values for these parameters by using the `cvshrink` function or the `'OptimizeHyperparameters'` name-value pair argument. For examples, see these topics: • "Regularize Discriminant Analysis Classifier" on page 20-21 • "Optimize Discriminant Analysis Model" on page 33-1574
`fitcecoc` with `templateLinear`	Linear classification for multiclass learning with high-dimensional data	Continuous features	Train a linear classification model by using `fitcecoc` and linear binary learners defined by `templateLinear`. Specify `'Regularization'` of `templatelinear` as `'lasso'` to use lasso regularization. For an example, see "Find Good Lasso Penalty Using Cross-Validation" on page 33-514. This example determines a good lasso-penalty strength by evaluating models with different strength values using `kfoldLoss`. You can also evaluate models using `kfoldEdge`, `kfoldMargin`, `edge`, `loss`, or `margin`.

Function	Supported Problem	Supported Data Type	Description
`fitclinear`	Linear classification for binary learning with high-dimensional data	Continuous features	Train a linear classification model by using `fitclinear`. Specify `'Regularization'` of `fitclinear` as `'lasso'` to use lasso regularization. For an example, see "Find Good Lasso Penalty Using Cross-Validated AUC" on page 33-3286. This example determines a good lasso-penalty strength by evaluating models with different strength values using the AUC values. Compute the cross-validated posterior class probabilities by using `kfoldPredict`, and compute the AUC values by using `perfcurve`. You can also evaluate models using `kfoldEdge`, `kfoldLoss`, `kfoldMargin`, `edge`, `loss`, `margin`, or `predict`.
`fitrgp`	Regression	Categorical and continuous features	Train a Gaussian process regression (GPR) model by using `fitrgp`. Set the `'KernelFunction'` name-value pair argument to use automatic relevance determination (ARD). Available options are `'ardsquaredexponential'`, `'ardexponential'`, `'ardmatern32'`, `'ardmatern52'`, and `'ardrationalquadratic'`. Find the predictor weights by taking the exponential of the negative learned length scales, stored in the `KernelInformation` property. For examples, see these topics: • "Specify Initial Step Size for LBFGS Optimization" on page 33-2151 • "Compare NCA and ARD Feature Selection" on page 33-2507

Function	Supported Problem	Supported Data Type	Description
`fitrlinear`	Linear regression with high-dimensional data	Continuous features	Train a linear regression model by using `fitrlinear`. Specify `'Regularization'` of `fitrlinear` as `'lasso'` to use lasso regularization. For examples, see these topics: • "Find Good Lasso Penalty Using Regression Loss" on page 33-5815 • "Find Good Lasso Penalty Using Cross-Validation" on page 33-2178
`lasso`	Linear regression	Continuous features	Train a linear regression model with "Lasso" on page 33-3431 regularization by using `lasso`. You can specify the weight of lasso versus ridge optimization by using the `'Alpha'` name-value pair argument. For examples, see the function reference page `lasso` and these topics: • "Lasso Regularization" on page 11-120 • "Lasso and Elastic Net with Cross Validation" on page 11-123 • "Wide Data via Lasso and Parallel Computing" on page 11-115
`lassoglm`	Generalized linear regression	Continuous features	Train a generalized linear regression model with "Lasso" on page 33-3446 regularization by using `lassoglm`. You can specify the weight of lasso versus ridge optimization by using the `'Alpha'` name-value pair argument. For details, see the function reference page `lassoglm` and these topics: • "Lasso Regularization of Generalized Linear Models" on page 12-32 • "Regularize Poisson Regression" on page 12-34 • "Regularize Logistic Regression" on page 12-36 • "Regularize Wide Data in Parallel" on page 12-43

Function	Supported Problem	Supported Data Type	Description
`oobPermutedPredictorImportance**` of `ClassificationBaggedEnsemble`	Classification with an ensemble of bagged decision trees (for example, random forest)	Categorical and continuous features	Train a bagged classification ensemble with tree learners by using `fitcensemble` and specifying `'Method'` as `'bag'`. Then, use `oobPermutedPredictorImportance` to compute “Out-of-Bag, Predictor Importance Estimates by Permutation” on page 33-4351. The function measures how influential the predictor variables in the model are at predicting the response. For examples, see the function reference page and the topic `oobPermutedPredictorImportance`.
`oobPermutedPredictorImportance**` of `RegressionBaggedEnsemble`	Regression with an ensemble of bagged decision trees (for example, random forest)	Categorical and continuous features	Train a bagged regression ensemble with tree learners by using `fitrensemble` and specifying `'Method'` as `'bag'`. Then, use `oobPermutedPredictorImportance` to compute “Out-of-Bag, Predictor Importance Estimates by Permutation” on page 33-4358. The function measures how influential the predictor variables in the model are at predicting the response. For examples, see the function reference page `oobPermutedPredictorImportance` and “Select Predictors for Random Forests” on page 18-60.
`predictorImportance**` of `ClassificationEnsemble`	Classification with an ensemble of decision trees	Categorical and continuous features	Train a classification ensemble with tree learners by using `fitcensemble`. Then, use `predictorImportance` to compute estimates of “Predictor Importance” on page 33-5032 for the ensemble by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. For examples, see the function reference page `predictorImportance`.

Function	Supported Problem	Supported Data Type	Description
`predictorImportance**` of `ClassificationTree`	Classification with a decision tree	Categorical and continuous features	Train a classification tree by using `fitctree`. Then, use `predictorImportance` to compute estimates of "Predictor Importance" on page 33-5039 for the tree by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. For examples, see the function reference page `predictorImportance`.
`predictorImportance**` of `RegressionEnsemble`	Regression with an ensemble of decision trees	Categorical and continuous features	Train a regression ensemble with tree learners by using `fitrensemble`. Then, use `predictorImportance` to compute estimates of "Predictor Importance" on page 33-5044 for the ensemble by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. For examples, see the function reference page `predictorImportance`.
`predictorImportance**` of `RegressionTree`	Regression with a decision tree	Categorical and continuous features	Train a regression tree by using `fitrtree`. Then, use `predictorImportance` to compute estimates of "Predictor Importance" on page 33-5049 for the tree by summing changes in the mean squared error (MSE) due to splits on every predictor and dividing the sum by the number of branch nodes. For examples, see the function reference page `predictorImportance`.

Function	Supported Problem	Supported Data Type	Description
`stepwiseglm`***	Generalized linear regression	Categorical and continuous features	Fit a generalized linear regression model using stepwise regression by using `stepwiseglm`. Alternatively, you can fit a linear regression model by using `fitglm` and then adjust the model by using `step`. Stepwise regression is a systematic method for adding and removing terms from the model based on their statistical significance in explaining the response variable. For details, see the function reference page `stepwiseglm` and these topics: • “Generalized Linear Model Using Stepwise Algorithm” on page 33-6004 • “Generalized Linear Models” on page 12-9 • “Generalized Linear Model Workflow” on page 12-28
`stepwiselm`***	Linear regression	Categorical and continuous features	Fit a linear regression model using stepwise regression by using `stepwiselm`. Alternatively, you can fit a linear regression model by using `fitlm` and then adjust the model by using `step`. Stepwise regression is a systematic method for adding and removing terms from the model based on their statistical significance in explaining the response variable. For details, see the function reference page `stepwiselm` and these topics: • “Stepwise Regression” on page 11-99 • “Linear Regression with Interaction Effects” on page 11-44 • “Assess Significance of Regression Coefficients Using t-statistic” on page 11-75

**For a tree-based algorithm, specify `'PredictorSelection'` as `'interaction-curvature'` to use the interaction test for selecting the best split predictor. The interaction test is useful in identifying important variables in the presence of many irrelevant variables. Also, if the training data includes many predictors, then specify `'NumVariablesToSample'` as `'all'` for training.

Otherwise, the software might not select some predictors, underestimating their importance. For details, see `fitctree`, `fitrtree`, and `templateTree`.

***`stepwiseglm` and `stepwiselm` are not wrapper type functions because you cannot use them as a wrapper for another training function. However, these two functions use the wrapper type algorithm to find important features.

References

[1] Guyon, Isabelle, and A. Elisseeff. "An introduction to variable and feature selection." *Journal of Machine Learning Research*. Vol. 3, 2003, pp. 1157–1182.

See Also

`rankfeatures`

More About

- “Sequential Feature Selection” on page 15-61
- “Interpret Machine Learning Models” on page 18-256

Sequential Feature Selection

This topic introduces to sequential feature selection and provides an example that selects features sequentially using a custom criterion and the `sequentialfs` function.

Introduction to Sequential Feature Selection

A common method of Feature Selection on page 15-49 is sequential feature selection. This method has two components:

- An objective function, called the criterion, which the method seeks to minimize over all feasible feature subsets. Common criteria are mean squared error (for regression models) and misclassification rate (for classification models).
- A sequential search algorithm, which adds or removes features from a candidate subset while evaluating the criterion. Since an exhaustive comparison of the criterion value at all 2^n subsets of an n-feature data set is typically infeasible (depending on the size of n and the cost of objective calls), sequential searches move in only one direction, always growing or always shrinking the candidate set.

The method has two variants:

- Sequential forward selection (SFS), in which features are sequentially added to an empty candidate set until the addition of further features does not decrease the criterion.
- Sequential backward selection (SBS), in which features are sequentially removed from a full candidate set until the removal of further features increase the criterion.

Statistics and Machine Learning Toolbox offers several sequential feature selection functions:

- Stepwise regression is a sequential feature selection technique designed specifically for least-squares fitting. The functions `stepwiselm` and `stepwiseglm` use optimizations that are possible only with least-squares criteria. Unlike other sequential feature selection algorithms, stepwise regression can remove features that have been added or add features that have been removed, based on the criterion specified by the `'Criterion'` name-value pair argument.
- `sequentialfs` performs sequential feature selection using a custom criterion. Input arguments include predictor data, response data, and a function handle to a file implementing the criterion function. You can define a criterion function that measures the characteristics of data or the performance of a learning algorithm. Optional inputs allow you to specify SFS or SBS, required or excluded features, and the size of the feature subset. The function calls `cvpartition` and `crossval` to evaluate the criterion at different candidate sets.
- `fscmrmr` ranks features using the minimum redundancy maximum relevance (MRMR) algorithm for classification problems.

Select Subset of Features with Comparative Predictive Power

This example selects a subset of features using a custom criterion that measures predictive power for a generalized linear regression problem.

Consider a data set with 100 observations of 10 predictors. Generate the random data from a logistic model, with a binomial distribution of responses at each set of values for the predictors. Some coefficients are set to zero so that not all of the predictors affect the response.

```
rng(456) % Set the seed for reproducibility
n = 100;
m = 10;
X = rand(n,m);
b = [1 0 0 2 .5 0 0 0.1 0 1];
Xb = X*b';
p = 1./(1+exp(-Xb));
N = 50;
y = binornd(N,p);
```

Fit a logistic model to the data using `fitglm`.

```
Y = [y N*ones(size(y))];
model0 = fitglm(X,Y,'Distribution','binomial')
```

```
model0 =
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10
    Distribution = Binomial

Estimated Coefficients:
                   Estimate       SE        tStat        pValue
                   ________    _______    ________    __________

    (Intercept)      0.22474    0.30043     0.74806       0.45443
    x1               0.68782    0.17207      3.9973     6.408e-05
    x2                0.2003    0.18087      1.1074       0.26811
    x3             -0.055328    0.18871    -0.29319       0.76937
    x4                2.2576     0.1813      12.452    1.3566e-35
    x5               0.54603    0.16836      3.2432     0.0011821
    x6              0.069701    0.17738     0.39294       0.69437
    x7              -0.22562    0.16957     -1.3306       0.18334
    x8              -0.19712    0.17317     -1.1383       0.25498
    x9              -0.20373    0.16796      -1.213       0.22514
    x10              0.99741    0.17247      5.7832    7.3296e-09


100 observations, 89 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 222, p-value = 4.92e-42
```

Display the deviance of the fit.

```
dev0 = model0.Deviance
```

```
dev0 = 101.5648
```

This model is the full model, with all of the features and an initial constant term. Sequential feature selection searches for a subset of the features in the full model with comparative predictive power.

Before performing feature selection, you must specify a criterion for selecting the features. In this case, the criterion is the deviance of the fit (a generalization of the residual sum of squares). The `critfun` function (shown at the end of this example) calls `fitglm` and returns the deviance of the fit.

If you use the live script file for this example, the `critfun` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB path.

Perform feature selection. `sequentialfs` calls the criterion function via a function handle.

```
maxdev = chi2inv(.95,1);
opt = statset('display','iter',...
              'TolFun',maxdev,...
              'TolTypeFun','abs');

inmodel = sequentialfs(@critfun,X,Y,...
                       'cv','none',...
                       'nullmodel',true,...
                       'options',opt,...
                       'direction','forward');
```

```
Start forward sequential feature selection:
Initial columns included:  none
Columns that can not be included:  none
Step 1, used initial columns, criterion value 323.173
Step 2, added column 4, criterion value 184.794
Step 3, added column 10, criterion value 139.176
Step 4, added column 1, criterion value 119.222
Step 5, added column 5, criterion value 107.281
Final columns included:  1 4 5 10
```

The iterative display shows a decrease in the criterion value as each new feature is added to the model. The final result is a reduced model with only four of the original ten features: columns `1`, `4`, `5`, and `10` of `X`, as indicated in the logical vector `inmodel` returned by `sequentialfs`.

The deviance of the reduced model is higher than the deviance of the full model. However, the addition of any other single feature would not decrease the criterion value by more than the absolute tolerance, `maxdev`, set in the options structure. Adding a feature with no effect reduces the deviance by an amount that has a chi-square distribution with one degree of freedom. Adding a significant feature results in a larger change in the deviance. By setting `maxdev` to `chi2inv(.95,1)`, you instruct `sequentialfs` to continue adding features provided that the change in deviance is more than the change expected by random chance.

Create the reduced model with an initial constant term.

```
model = fitglm(X(:,inmodel),Y,'Distribution','binomial')
```

```
model =
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE         tStat        pValue
                   __________    _______    _________    __________

    (Intercept)    -0.0052025    0.16772    -0.031018       0.97525
    x1                0.73814    0.16316       4.5241    6.0666e-06
    x2                 2.2139    0.17402       12.722    4.4369e-37
    x3                0.54073     0.1568       3.4485    0.00056361
    x4                 1.0694    0.15916       6.7191    1.8288e-11


100 observations, 95 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 216, p-value = 1.44e-45
```

This code creates the function `critfun`.

```
function dev = critfun(X,Y)
model = fitglm(X,Y,'Distribution','binomial');
dev = model.Deviance;
end
```

See Also

`sequentialfs` | `stepwiseglm` | `stepwiselm`

More About

- “Introduction to Feature Selection” on page 15-49

Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a dimension-reduction technique based on a low-rank approximation of the feature space. Besides providing a reduction in the number of features, NMF guarantees that the features are nonnegative, producing additive models that respect, for example, the nonnegativity of physical quantities.

Given a nonnegative m-by-n matrix X and a positive integer $k < \min(m,n)$, NMF finds nonnegative m-by-k and k-by-n matrices W and H, respectively, that minimize the norm of the difference $X – WH$. W and H are thus approximate nonnegative factors of X.

The k columns of W represent transformations of the variables in X; the k rows of H represent the coefficients of the linear combinations of the original n variables in X that produce the transformed variables in W. Since k is generally smaller than the rank of X, the product WH provides a compressed approximation of the data in X. A range of possible values for k is often suggested by the modeling context.

The function `nnmf` carries out nonnegative matrix factorization. `nnmf` uses one of two iterative algorithms that begin with random initial values for W and H. Because the norm of the residual $X – WH$ may have local minima, repeated calls to `nnmf` may yield different factorizations. Sometimes the algorithm converges to a solution of lower rank than k, which may indicate that the result is not optimal.

See Also

`nnmf`

Related Examples

- “Perform Nonnegative Matrix Factorization” on page 15-66

Perform Nonnegative Matrix Factorization

This example shows how to perform nonnegative matrix factorization.

Load the sample data.

```
load moore
X = moore(:,1:5);
rng('default'); % For reproducibility
```

Compute a rank-two approximation of X using a multiplicative update algorithm that begins from five random initial values for W and H.

```
opt = statset('MaxIter',10,'Display','final');
[W0,H0] = nnmf(X,2,'replicates',5,'options',opt,'algorithm','mult');
```

```
    rep       iteration       rms resid      |delta x|
      1          10         358.296      0.00190554
      2          10         78.3556     0.000351747
      3          10         230.962       0.0172839
      4          10         326.347      0.00739552
      5          10         361.547      0.00705539
Final root mean square residual = 78.3556
```

The `'mult'` algorithm is sensitive to initial values, which makes it a good choice when using `'replicates'` to find W and H from multiple random starting values.

Now perform the factorization using alternating least-squares algorithm, which converges faster and more consistently. Run 100 times more iterations, beginning from the initial W0 and H0 identified above.

```
opt = statset('Maxiter',1000,'Display','final');
[W,H] = nnmf(X,2,'w0',W0,'h0',H0,'options',opt,'algorithm','als');
```

```
    rep       iteration       rms resid      |delta x|
      1           2         77.5315     0.000830334
Final root mean square residual = 77.5315
```

The two columns of W are the transformed predictors. The two rows of H give the relative contributions of each of the five predictors in X to the predictors in W. Display H.

```
H
```

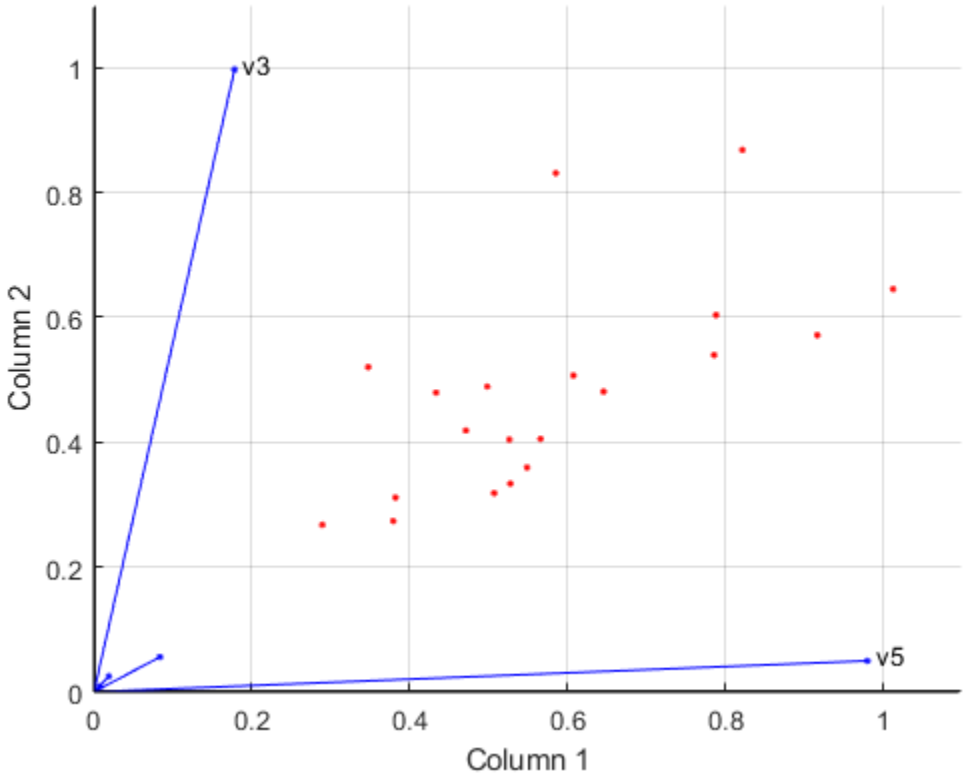
```
H = 2×5

    0.0835    0.0190    0.1782    0.0072    0.9802
    0.0559    0.0250    0.9969    0.0085    0.0497
```

The fifth predictor in X (weight 0.9802) strongly influences the first predictor in W. The third predictor in X (weight 0.9969) strongly influences the second predictor in W.

Visualize the relative contributions of the predictors in X with `biplot`, showing the data and original variables in the column space of W.

```
biplot(H','scores',W,'varlabels',{'','','v3','','v5'});
axis([0 1.1 0 1.1])
```

```
xlabel('Column 1')
ylabel('Column 2')
```



See Also

`nnmf`

More About

- “Nonnegative Matrix Factorization” on page 15-65

Principal Component Analysis (PCA)

One of the difficulties inherent in multivariate statistics is the problem of visualizing data that has many variables. The function `plot` displays a graph of the relationship between two variables. The `plot3` and `surf` commands display different three-dimensional views. But when there are more than three variables, it is more difficult to visualize their relationships.

Fortunately, in data sets with many variables, groups of variables often move together. One reason for this is that more than one variable might be measuring the same driving principle governing the behavior of the system. In many systems there are only a few such driving forces. But an abundance of instrumentation enables you to measure dozens of system variables. When this happens, you can take advantage of this redundancy of information. You can simplify the problem by replacing a group of variables with a single new variable.

Principal component analysis is a quantitatively rigorous method for achieving this simplification. The method generates a new set of variables, called *principal components*. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data.

There are an infinite number of ways to construct an orthogonal basis for several columns of data. What is so special about the principal component basis?

The first principal component is a single axis in space. When you project each observation on that axis, the resulting values form a new variable. And the variance of this variable is the maximum among all possible choices of the first axis.

The second principal component is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis.

The full set of principal components is as large as the original set of variables. But it is commonplace for the sum of the variances of the first few principal components to exceed 80% of the total variance of the original data. By examining plots of these few new variables, researchers often develop a deeper understanding of the driving forces that generated the original data.

You can use the function `pca` to find the principal components. To use `pca`, you need to have the actual measured data you want to analyze. However, if you lack the actual data, but have the sample covariance or correlation matrix for the data, you can still use the function `pcacov` to perform a principal components analysis. See the reference page for `pcacov` for a description of its inputs and outputs.

See Also

`pca` | `pcacov` | `pcares` | `ppca`

Related Examples

- “Analyze Quality of Life in U.S. Cities Using PCA” on page 15-69

Analyze Quality of Life in U.S. Cities Using PCA

This example shows how to perform a weighted principal components analysis and interpret the results.

Load sample data.

Load the sample data. The data includes ratings for 9 different indicators of the quality of life in 329 U.S. cities. These are climate, housing, health, crime, transportation, education, arts, recreation, and economics. For each category, a higher rating is better. For example, a higher rating for crime means a lower crime rate.

Display the `categories` variable.

```
load cities
categories
```

```
categories =
   climate
   housing
   health
   crime
   transportation
   education
   arts
   recreation
   economics
```

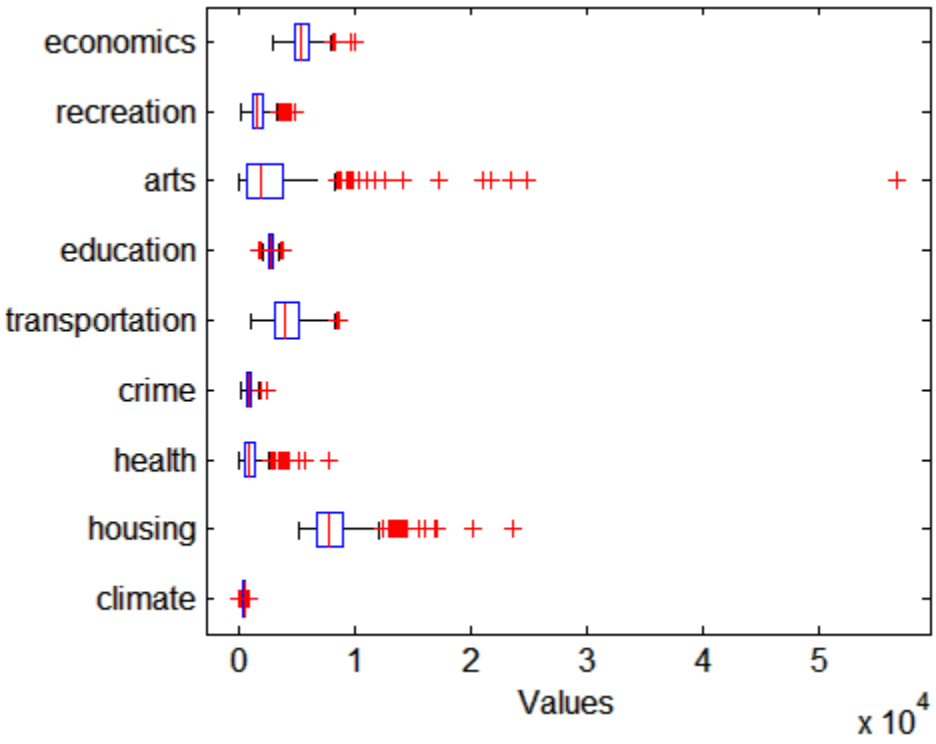
In total, the `cities` data set contains three variables:

- `categories`, a character matrix containing the names of the indices
- `names`, a character matrix containing the 329 city names
- `ratings`, the data matrix with 329 rows and 9 columns

Plot data.

Make a box plot to look at the distribution of the `ratings` data.

```
figure()
boxplot(ratings,'Orientation','horizontal','Labels',categories)
```



There is more variability in the ratings of the arts and housing than in the ratings of crime and climate.

Check pairwise correlation.

Check the pairwise correlation between the variables.

```
C = corr(ratings,ratings);
```

The correlation among some variables is as high as 0.85. Principal components analysis constructs independent new variables which are linear combinations of the original variables.

Compute principal components.

When all variables are in the same unit, it is appropriate to compute principal components for raw data. When the variables are in different units or the difference in the variance of different columns is substantial (as in this case), scaling of the data or use of weights is often preferable.

Perform the principal component analysis by using the inverse variances of the ratings as weights.

```
w = 1./var(ratings);
[wcoeff,score,latent,tsquared,explained] = pca(ratings,...
'VariableWeights',w);
```

Or equivalently:

```
[wcoeff,score,latent,tsquared,explained] = pca(ratings,...
'VariableWeights','variance');
```

The following sections explain the five outputs of `pca`.

Component coefficients.

The first output, `wcoeff`, contains the coefficients of the principal components.

The first three principal component coefficient vectors are:

```
c3 = wcoeff(:,1:3)
```

```
c3 = wcoeff(:,1:3)
c3 =

   1.0e+03 *

    0.0249   -0.0263   -0.0834
    0.8504   -0.5978   -0.4965
    0.4616    0.3004   -0.0073
    0.1005   -0.1269    0.0661
    0.5096    0.2606    0.2124
    0.0883    0.1551    0.0737
    2.1496    0.9043   -0.1229
    0.2649   -0.3106   -0.0411
    0.1469   -0.5111    0.6586
```

These coefficients are weighted, hence the coefficient matrix is not orthonormal.

Transform coefficients.

Transform the coefficients so that they are orthonormal.

```
 coefforth = inv(diag(std(ratings)))*wcoeff;
```

Note that if you use a weights vector, `w`, while conducting the `pca`, then

```
 coefforth = diag(sqrt(w))*wcoeff;
```

Check coefficients are orthonormal.

The transformed coefficients are now orthonormal.

```
I = coefforth'*coefforth;
I(1:3,1:3)
```

```
ans =
    1.0000   -0.0000   -0.0000
   -0.0000    1.0000   -0.0000
   -0.0000   -0.0000    1.0000
```

Component scores.

The second output, `score`, contains the coordinates of the original data in the new coordinate system defined by the principal components. The `score` matrix is the same size as the input data matrix. You can also obtain the component scores using the orthonormal coefficients and the standardized ratings as follows.

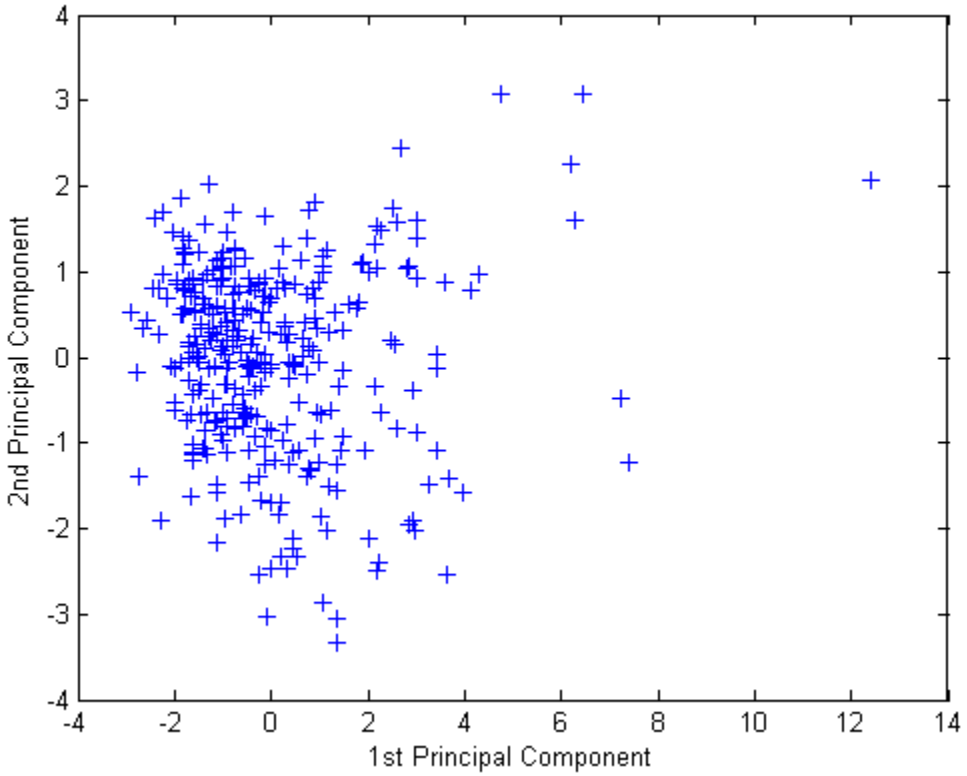
```
cscores = zscore(ratings)*coefforth;
```

`cscores` and `score` are identical matrices.

Plot component scores.

Create a plot of the first two columns of `score`.

```
figure()
plot(score(:,1),score(:,2),'+')
xlabel('1st Principal Component')
ylabel('2nd Principal Component')
```



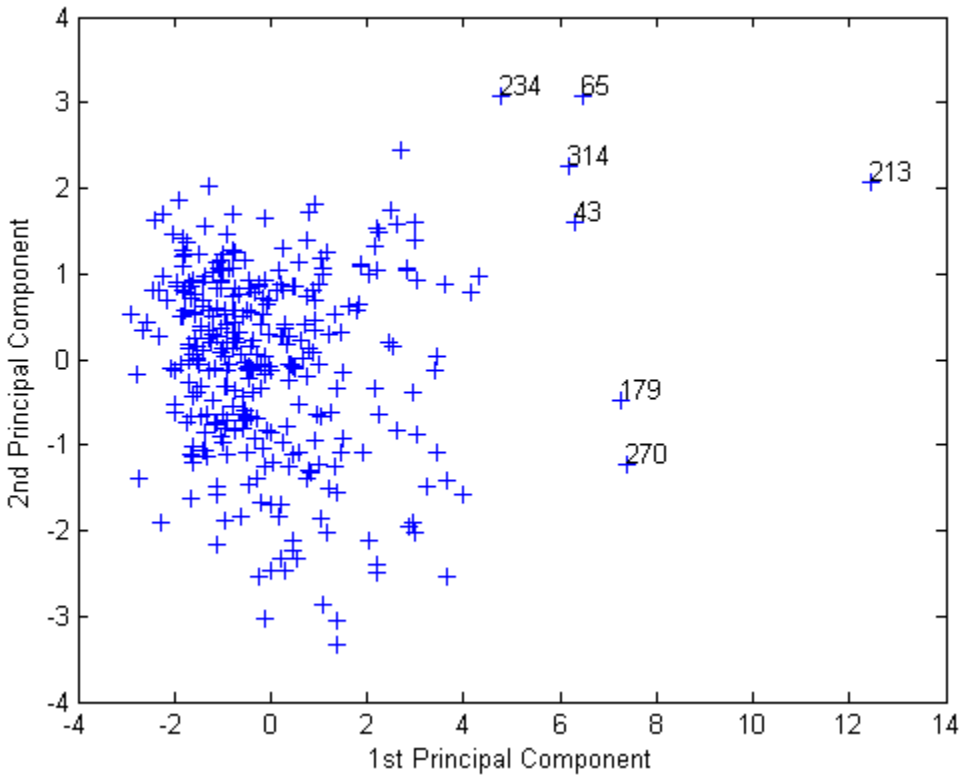
This plot shows the centered and scaled ratings data projected onto the first two principal components. `pca` computes the scores to have mean zero.

Explore plot interactively.

Note the outlying points in the right half of the plot. You can graphically identify these points as follows.

```
gname
```

Move your cursor over the plot and click once near the rightmost seven points. This labels the points by their row numbers as in the following figure.



After labeling points, press **Return**.

Extract observation names.

Create an index variable containing the row numbers of all the cities you chose and get the names of the cities.

```
metro = [43 65 179 213 234 270 314];
names(metro,:)
```

```
ans =
   Boston, MA
   Chicago, IL
   Los Angeles, Long Beach, CA
   New York, NY
   Philadelphia, PA-NJ
   San Francisco, CA
   Washington, DC-MD-VA
```

These labeled cities are some of the biggest population centers in the United States and they appear more extreme than the remainder of the data.

Component variances.

The third output, `latent`, is a vector containing the variance explained by the corresponding principal component. Each column of `score` has a sample variance equal to the corresponding row of `latent`.

```
latent
```

```
latent =

    3.4083
    1.2140
    1.1415
    0.9209
    0.7533
    0.6306
    0.4930
    0.3180
    0.1204
```

Percent variance explained.

The fifth output, `explained`, is a vector containing the percent variance explained by the corresponding principal component.

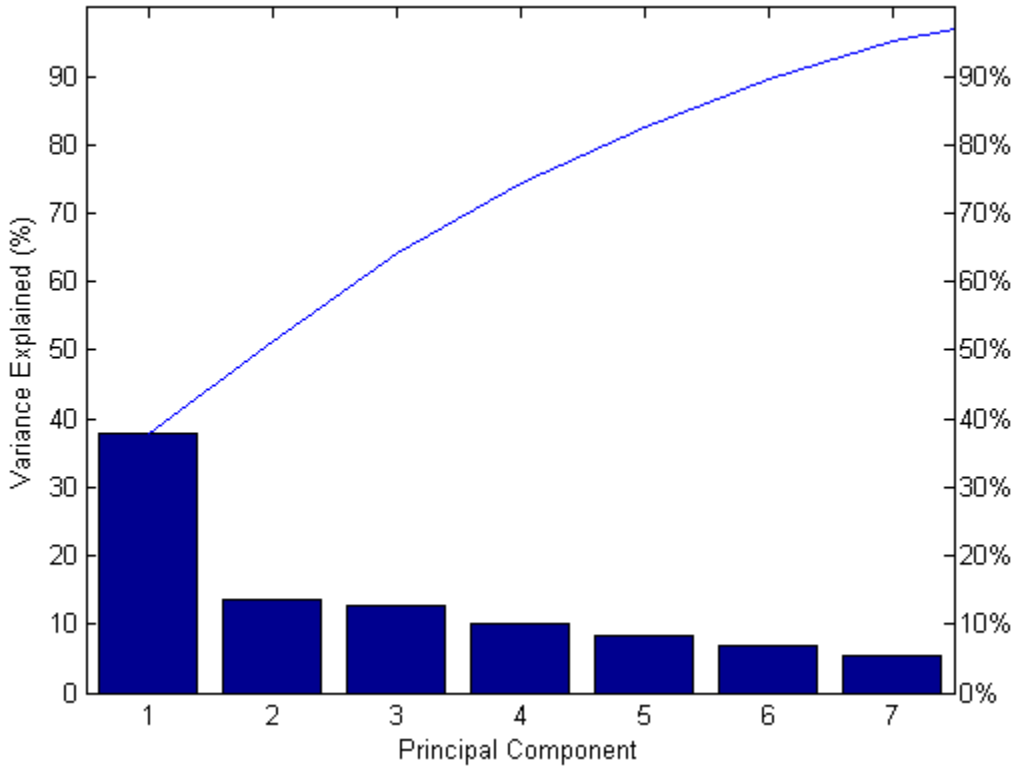
```
explained
```

```
explained =

   37.8699
   13.4886
   12.6831
   10.2324
    8.3698
    7.0062
    5.4783
    3.5338
    1.3378
```

Create scree plot.

Make a scree plot of the percent variability explained by each principal component.

```
figure()
pareto(explained)
xlabel('Principal Component')
ylabel('Variance Explained (%)')
```



This scree plot only shows the first seven (instead of the total nine) components that explain 95% of the total variance. The only clear break in the amount of variance accounted for by each component is between the first and second components. However, the first component by itself explains less than 40% of the variance, so more components might be needed. You can see that the first three principal components explain roughly two-thirds of the total variability in the standardized ratings, so that might be a reasonable way to reduce the dimensions.

Hotelling's T-squared statistic.

The last output from `pca` is `tsquared`, which is Hotelling's T^2, a statistical measure of the multivariate distance of each observation from the center of the data set. This is an analytical way to find the most extreme points in the data.

```
[st2,index] = sort(tsquared,'descend'); % sort in descending order
extreme = index(1);
names(extreme,:)
```

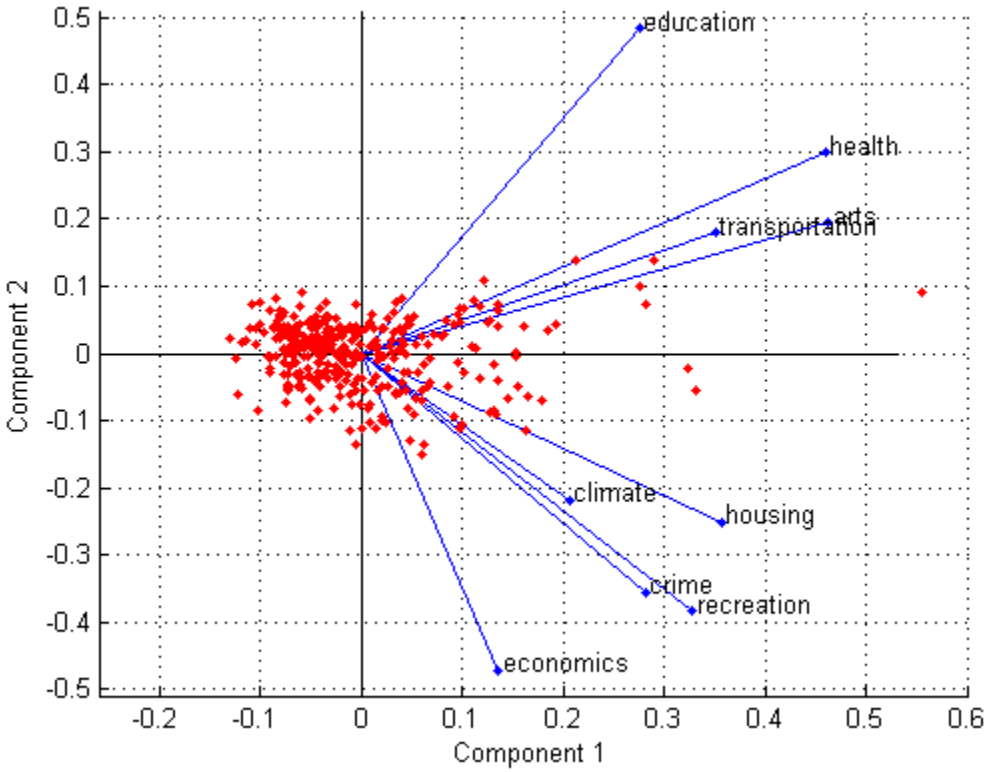
```
ans =

New York, NY
```

The ratings for New York are the furthest from the average U.S. city.

Visualize the results.

Visualize both the orthonormal principal component coefficients for each variable and the principal component scores for each observation in a single plot.

```
biplot(coefforth(:,1:2),'Scores',score(:,1:2),'Varlabels',categories);
axis([-.26 0.6 -.51 .51]);
```



All nine variables are represented in this bi-plot by a vector, and the direction and length of the vector indicate how each variable contributes to the two principal components in the plot. For example, the first principal component, on the horizontal axis, has positive coefficients for all nine variables. That is why the nine vectors are directed into the right half of the plot. The largest coefficients in the first principal component are the third and seventh elements, corresponding to the variables `health` and `arts`.

The second principal component, on the vertical axis, has positive coefficients for the variables `education`, `health`, `arts`, and `transportation`, and negative coefficients for the remaining five variables. This indicates that the second component distinguishes among cities that have high values for the first set of variables and low for the second, and cities that have the opposite.

The variable labels in this figure are somewhat crowded. You can either exclude the `VarLabels` name-value pair argument when making the plot, or select and drag some of the labels to better positions using the Edit Plot tool from the figure window toolbar.

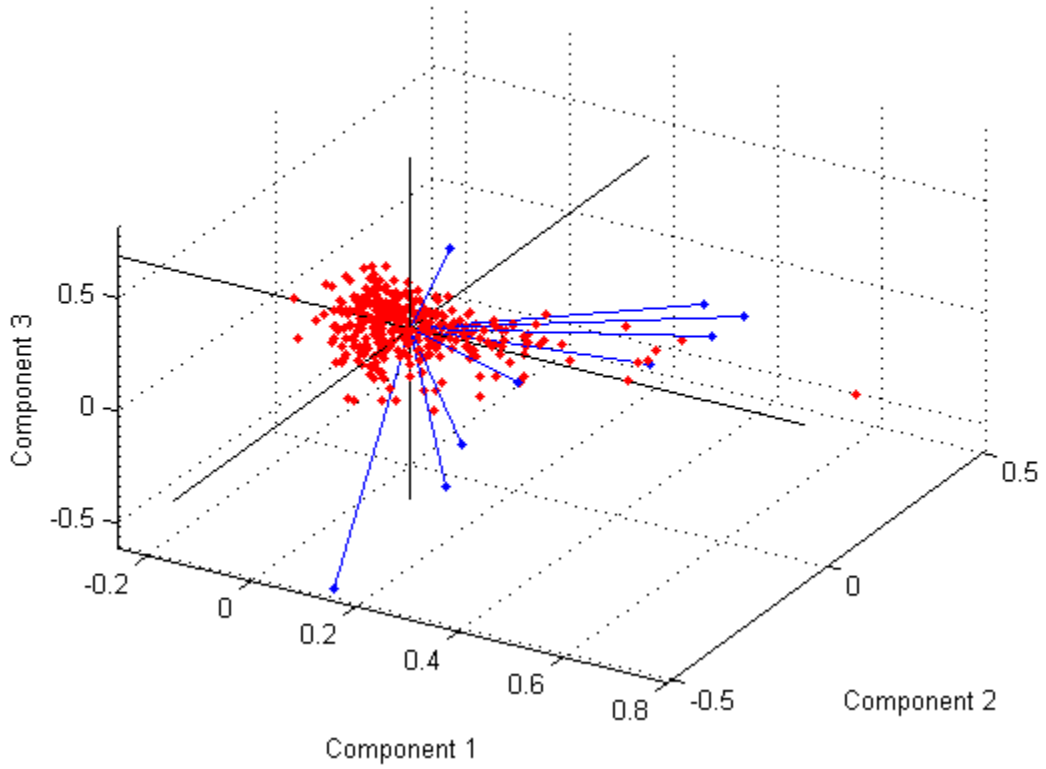
This 2-D bi-plot also includes a point for each of the 329 observations, with coordinates indicating the score of each observation for the two principal components in the plot. For example, points near the left edge of this plot have the lowest scores for the first principal component. The points are scaled with respect to the maximum score value and maximum coefficient length, so only their relative locations can be determined from the plot.

You can identify items in the plot by selecting **Tools>Data Cursor** in the figure window. By clicking a variable (vector), you can read the variable label and coefficients for each principal component. By clicking an observation (point), you can read the observation name and scores for each principal component. You can specify `'Obslabels',names` to show the observation names instead of the observation numbers in the data cursor display.

Create a three-dimensional bi-plot.

You can also make a bi-plot in three dimensions.

```
figure()
biplot(coefforth(:,1:3),'Scores',score(:,1:3),'Obslabels',names);
axis([-.26 0.8 -.51 .51 -.61 .81]);
view([30 40]);
```



This graph is useful if the first two principal coordinates do not explain enough of the variance in your data. You can also rotate the figure to see it from different angles by selecting the**Tools**> `Rotate 3D`.

See Also

`biplot` | `boxplot` | `pca` | `pcacov` | `pcares` | `ppca`

More About

- “Principal Component Analysis (PCA)” on page 15-68

Factor Analysis

Multivariate data often includes a large number of measured variables, and sometimes those variables overlap, in the sense that groups of them might be dependent. For example, in a decathlon, each athlete competes in 10 events, but several of them can be thought of as speed events, while others can be thought of as strength events, etc. Thus, you can think of a competitor's 10 event scores as largely dependent on a smaller set of three or four types of athletic ability.

Factor analysis is a way to fit a model to multivariate data to estimate just this sort of interdependence. In a factor analysis model, the measured variables depend on a smaller number of unobserved (latent) factors. Because each factor might affect several variables in common, they are known as common factors. Each variable is assumed to be dependent on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as specific variance because it is specific to one variable.

Specifically, factor analysis assumes that the covariance matrix of your data is of the form

$$\sum_X = \Lambda\Lambda^T + \Psi$$

where Λ is the matrix of loadings, and the elements of the diagonal matrix Ψ are the specific variances. The function `factoran` fits the Factor Analysis model using maximum likelihood.

See Also

`factoran`

Related Examples

- “Analyze Stock Prices Using Factor Analysis” on page 15-79

Analyze Stock Prices Using Factor Analysis

This example shows how to analyze if companies within the same sector experience similar week-to-week changes in stock price.

Factor Loadings

Load the sample data.

```
load stockreturns
```

Suppose that over the course of 100 weeks, the percent change in stock prices for ten companies has been recorded. Of the ten companies, the first four can be classified as primarily technology, the next three as financial, and the last three as retail. It seems reasonable that the stock prices for companies that are in the same sector might vary together as economic conditions change. Factor analysis can provide quantitative evidence.

First specify a model fit with three common factors. By default, `factoran` computes rotated estimates of the loadings to try and make their interpretation simpler. But in this example, specify an unrotated solution.

```
[Loadings,specificVar,T,stats] = factoran(stocks,3,'rotate','none');
```

The first two `factoran` output arguments are the estimated loadings and the estimated specific variances. Each row of the loadings matrix represents one of the ten stocks, and each column corresponds to a common factor. With unrotated estimates, interpretation of the factors in this fit is difficult because most of the stocks contain fairly large coefficients for two or more factors.

```
Loadings
```

```
Loadings = 10×3

    0.8885    0.2367   -0.2354
    0.7126    0.3862    0.0034
    0.3351    0.2784   -0.0211
    0.3088    0.1113   -0.1905
    0.6277   -0.6643    0.1478
    0.4726   -0.6383    0.0133
    0.1133   -0.5416    0.0322
    0.6403    0.1669    0.4960
    0.2363    0.5293    0.5770
    0.1105    0.1680    0.5524
```

Factor rotation helps to simplify the structure in the `Loadings` matrix, to make it easier to assign meaningful interpretations to the factors.

From the estimated specific variances, you can see that the model indicates that a particular stock price varies quite a lot beyond the variation due to the common factors. Display estimated specific variances.

```
specificVar
```

```
specificVar = 10×1

    0.0991
    0.3431
```

```
    0.8097
    0.8559
    0.1429
    0.3691
    0.6928
    0.3162
    0.3311
    0.6544
```

A specific variance of 1 would indicate that there is no common factor component in that variable, while a specific variance of 0 would indicate that the variable is entirely determined by common factors. These data seem to fall somewhere in between.

Display the *p*-value.

```
stats.p
```

```
ans = 0.8144
```

The *p*-value returned in the `stats` structure fails to reject the null hypothesis of three common factors, suggesting that this model provides a satisfactory explanation of the covariation in these data.

Fit a model with two common factors to determine whether fewer than three factors can provide an acceptable fit.

```
[Loadings2,specificVar2,T2,stats2] = factoran(stocks, 2,'rotate','none');
```

Display the *p*-value.

```
stats2.p
```

```
ans = 3.5610e-06
```

The *p*-value for this second fit is highly significant, and rejects the hypothesis of two factors, indicating that the simpler model is not sufficient to explain the pattern in these data.

Factor Rotation

As the results illustrate, the estimated loadings from an unrotated factor analysis fit can have a complicated structure. The goal of factor rotation is to find a parameterization in which each variable has only a small number of large loadings. That is, each variable is affected by a small number of factors, preferably only one. This can often make it easier to interpret what the factors represent.

If you think of each row of the loadings matrix as coordinates of a point in M-dimensional space, then each factor corresponds to a coordinate axis. Factor rotation is equivalent to rotating those axes and computing new loadings in the rotated coordinate system. There are various ways to do this. Some methods leave the axes orthogonal, while others are oblique methods that change the angles between them. For this example, you can rotate the estimated loadings by using the promax criterion, a common oblique method.

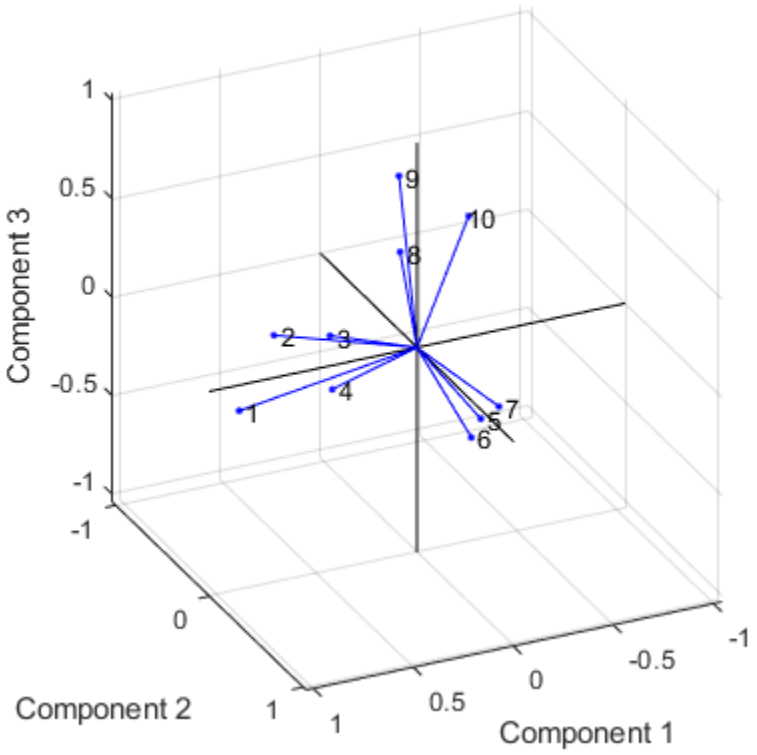
```
[LoadingsPM,specVarPM] = factoran(stocks,3,'rotate','promax');
LoadingsPM
```

```
LoadingsPM = 10×3
```

```
    0.9452    0.1214   -0.0617
    0.7064   -0.0178    0.2058
    0.3885   -0.0994    0.0975
    0.4162   -0.0148   -0.1298
    0.1021    0.9019    0.0768
    0.0873    0.7709   -0.0821
   -0.1616    0.5320   -0.0888
    0.2169    0.2844    0.6635
    0.0016   -0.1881    0.7849
   -0.2289    0.0636    0.6475
```

Promax rotation creates a simpler structure in the loadings, one in which most of the stocks have a large loading on only one factor. To see this structure more clearly, you can use the biplot function to plot each stock using its factor loadings as coordinates.

```
biplot(LoadingsPM,'varlabels',num2str((1:10)'));
axis square
view(155,27);
```

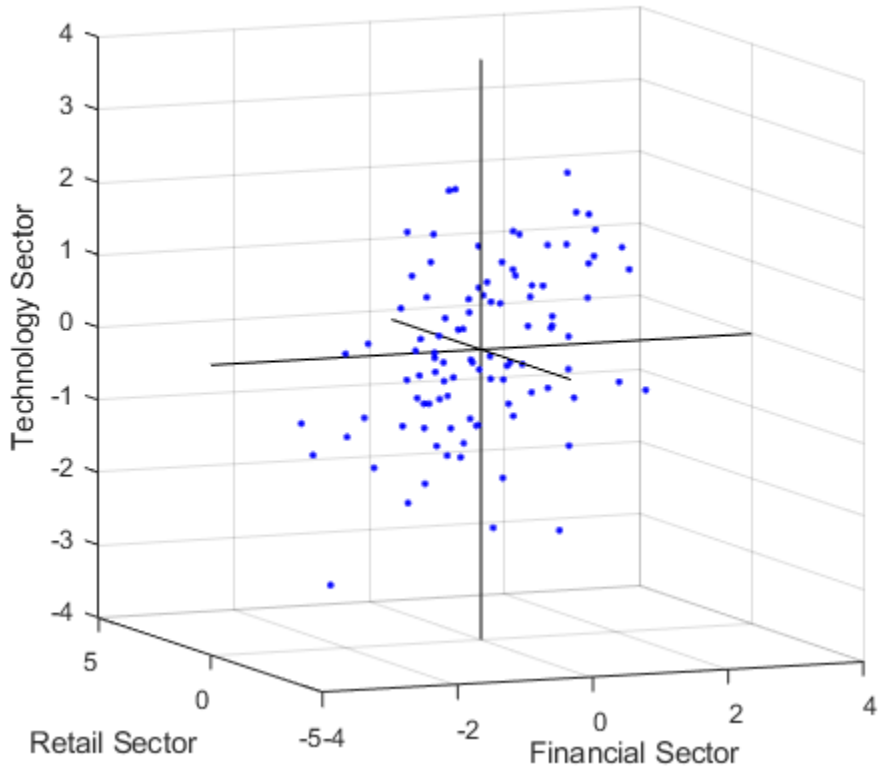


This plot shows that promax has rotated the factor loadings to a simpler structure. Each stock depends primarily on only one factor, and it is possible to describe each factor in terms of the stocks that it affects. Based on which companies are near which axes, you could reasonably conclude that the first factor axis represents the financial sector, the second retail, and the third technology. The original conjecture, that stocks vary primarily within sector, is apparently supported by the data.

Factor Scores

Sometimes, it is useful to be able to classify an observation based on its factor scores. For example, if you accepted the three-factor model and the interpretation of the rotated factors, you might want to categorize each week in terms of how favorable it was for each of the three stock sectors, based on the data from the 10 observed stocks. Because the data in this example are the raw stock price changes, and not just their correlation matrix, you can have `factoran` return estimates of the value of each of the three rotated common factors for each week. You can then plot the estimated scores to see how the different stock sectors were affected during each week.

```
[LoadingsPM,specVarPM,TPM,stats,F] = factoran(stocks, 3,'rotate','promax');

plot3(F(:,1),F(:,2),F(:,3),'b.')
line([-4 4 NaN 0 0 NaN 0 0],[0 0 NaN -4 4 NaN 0 0],[0 0 NaN 0 0 NaN -4 4], 'Color','black')
xlabel('Financial Sector')
ylabel('Retail Sector')
zlabel('Technology Sector')
grid on
axis square
view(-22.5, 8)
```



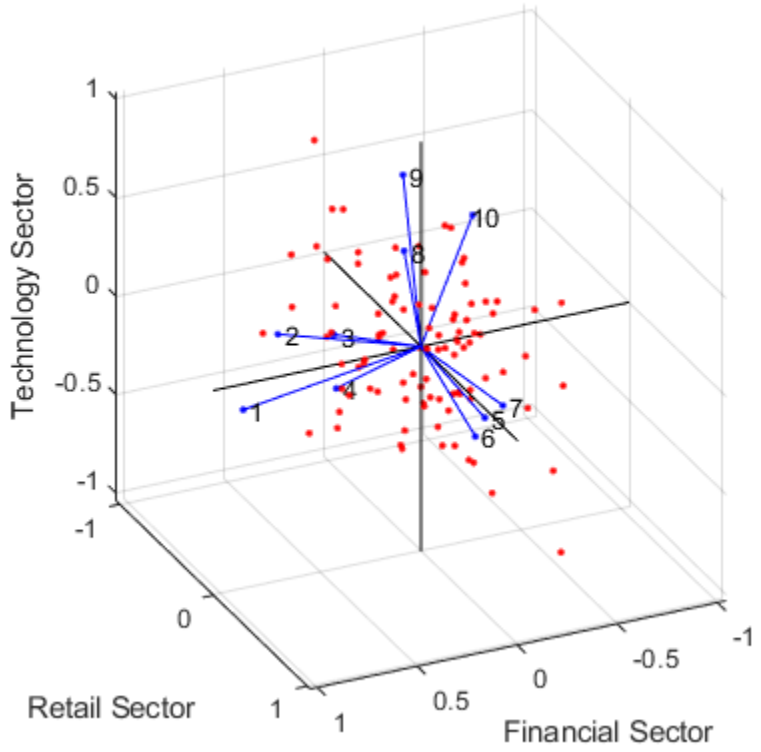
Oblique rotation often creates factors that are correlated. This plot shows some evidence of correlation between the first and third factors, and you can investigate further by computing the estimated factor correlation matrix.

```
inv(TPM'*TPM);
```

Visualize the Results

You can use the biplot function to help visualize both the factor loadings for each variable and the factor scores for each observation in a single plot. For example, the following command plots the results from the factor analysis on the stock data and labels each of the 10 stocks.

```
biplot(LoadingsPM,'scores',F,'varlabels',num2str((1:10)'))
xlabel('Financial Sector')
ylabel('Retail Sector')
zlabel('Technology Sector')
axis square
view(155,27)
```



In this case, the factor analysis includes three factors, and so the biplot is three-dimensional. Each of the 10 stocks is represented in this plot by a vector, and the direction and length of the vector indicates how each stock depends on the underlying factors. For example, you have seen that after promax rotation, the first four stocks have positive loadings on the first factor, and unimportant loadings on the other two factors. That first factor, interpreted as a financial sector effect, is represented in this biplot as one of the horizontal axes. The dependence of those four stocks on that factor corresponds to the four vectors directed approximately along that axis. Similarly, the dependence of stocks 5, 6, and 7 primarily on the second factor, interpreted as a retail sector effect, is represented by vectors directed approximately along that axis.

Each of the 100 observations is represented in this plot by a point, and their locations indicate the score of each observation for the three factors. For example, points near the top of this plot have the

highest scores for the technology sector factor. The points are scaled to fit within the unit square, so only their relative locations can be determined from the plot.

You can use the **Data Cursor** tool from the **Tools** menu in the figure window to identify the items in this plot. By clicking a stock (vector), you can read off that stock's loadings for each factor. By clicking an observation (point), you can read off that observation's scores for each factor.

Robust Feature Selection Using NCA for Regression

Perform feature selection that is robust to outliers using a custom robust loss function in NCA.

Generate data with outliers

Generate sample data for regression where the response depends on three of the predictors, namely predictors 4, 7, and 13.

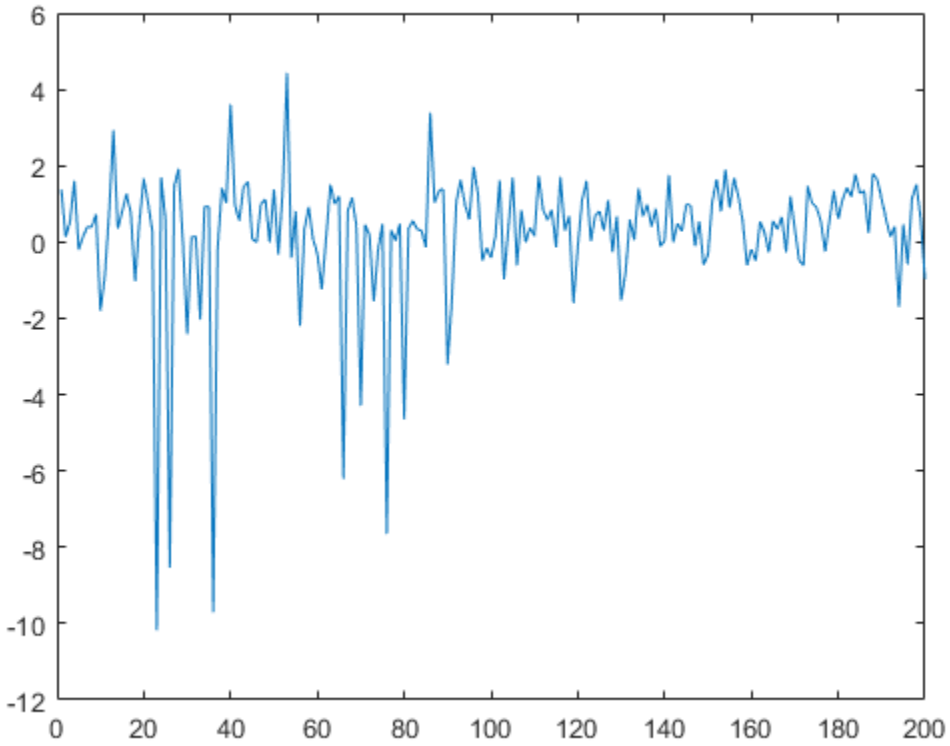
```
rng(123,'twister') % For reproducibility
n = 200;
X = randn(n,20);
y = cos(X(:,7)) + sin(X(:,4).*X(:,13)) + 0.1*randn(n,1);
```

Add outliers to data.

```
numoutliers = 25;
outlieridx = floor(linspace(10,90,numoutliers));
y(outlieridx) = 5*randn(numoutliers,1);
```

Plot the data.

```
figure
plot(y)
```



Use non-robust loss function

The performance of the feature selection algorithm highly depends on the value of the regularization parameter. A good practice is to tune the regularization parameter for the best value to use in feature selection. Tune the regularization parameter using five-fold cross validation. Use the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_j\right)^2$$

First, partition the data into five folds. In each fold, the software uses 4/5th of the data for training and 1/5th of the data for validation (testing).

```
cvp = cvpartition(length(y),'kfold',5);
numtestsets = cvp.NumTestSets;
```

Compute the lambda values to test for and create an array to store the loss values.

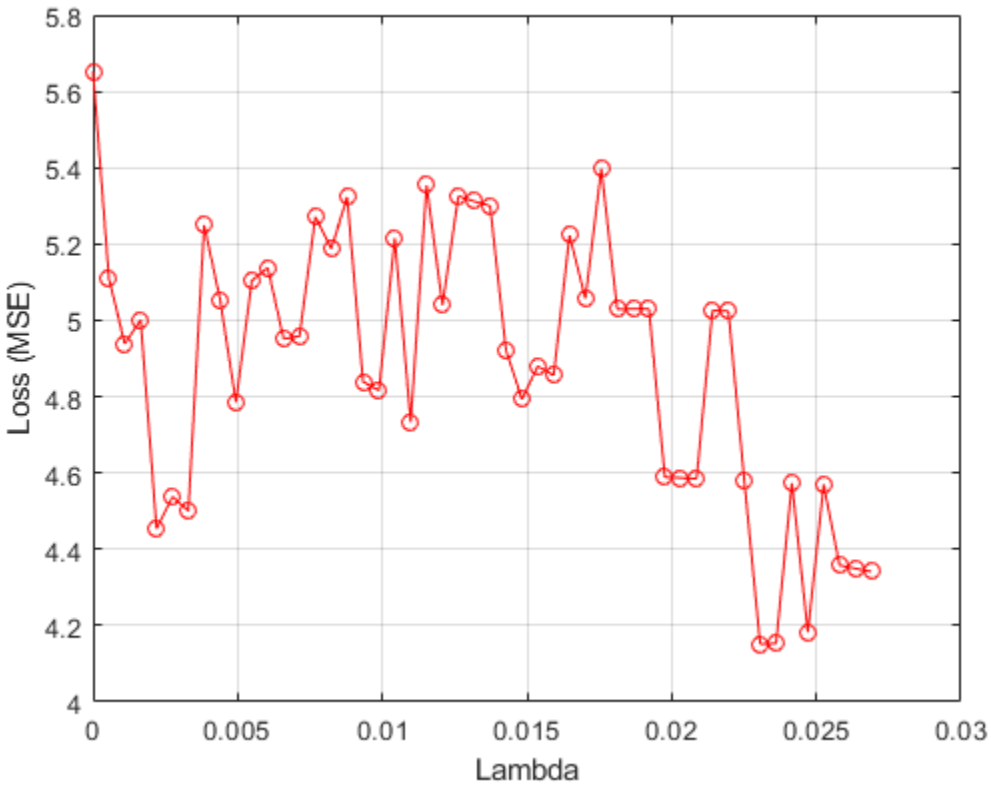
```
lambdavals = linspace(0,3,50)*std(y)/length(y);
lossvals = zeros(length(lambdavals),numtestsets);
```

Perform NCA and compute the loss for each λ value and each fold.

```
for i = 1:length(lambdavals)
    for k = 1:numtestsets
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);
        Xtest = X(cvp.test(k),:);
        ytest = y(cvp.test(k),:);

        nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
            'Solver','lbfgs','Verbose',0,'Lambda',lambdavals(i), ...
            'LossFunction','mse');

        lossvals(i,k) = loss(nca,Xtest,ytest,'LossFunction','mse');
    end
end
```

Plot the mean loss corresponding to each lambda value.

```
figure
meanloss = mean(lossvals,2);
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the λ value that produces the minimum average loss.

```
[~,idx] = min(mean(lossvals,2));
bestlambda = lambdavals(idx)
```

```
bestlambda = 0.0231
```

Perform feature selection using the best λ value and MSE.

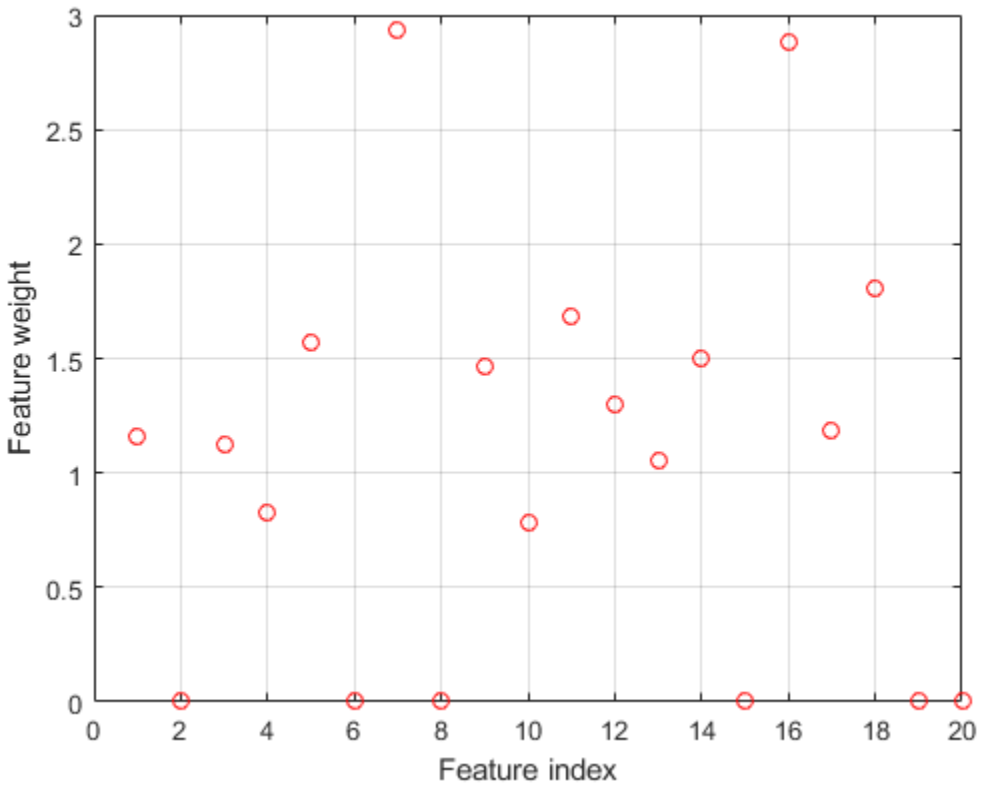
```
nca = fsrnca(X,y,'FitMethod','exact','Solver','lbfgs', ...
    'Verbose',1,'Lambda',bestlambda,'LossFunction','mse');
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 |  6.414642e+00 |   8.430e-01 |   0.000e+00 |        |   7.117e-01 |   0.000e+00 |
|        1 |  6.066100e+00 |   9.952e-01 |   1.264e+00 |    OK  |   3.741e-01 |   1.000e+00 |
|        2 |  5.498221e+00 |   4.267e-01 |   4.250e-01 |    OK  |   4.016e-01 |   1.000e+00 |
|        3 |  5.108548e+00 |   3.933e-01 |   8.564e-01 |    OK  |   3.599e-01 |   1.000e+00 |
|        4 |  4.808456e+00 |   2.505e-01 |   9.352e-01 |    OK  |   8.798e-01 |   1.000e+00 |
|        5 |  4.677382e+00 |   2.085e-01 |   6.014e-01 |    OK  |   1.052e+00 |   1.000e+00 |
|        6 |  4.487789e+00 |   4.726e-01 |   7.374e-01 |    OK  |   5.593e-01 |   1.000e+00 |
|        7 |  4.310099e+00 |   2.484e-01 |   4.253e-01 |    OK  |   3.367e-01 |   1.000e+00 |
|        8 |  4.258539e+00 |   3.629e-01 |   4.521e-01 |    OK  |   4.705e-01 |   5.000e-01 |
|        9 |  4.175345e+00 |   1.972e-01 |   2.608e-01 |    OK  |   4.018e-01 |   1.000e+00 |
|       10 |  4.122340e+00 |   9.169e-02 |   2.947e-01 |    OK  |   3.487e-01 |   1.000e+00 |
```

```
|       11 |  4.095525e+00 |   9.798e-02 |   2.529e-01 |    OK  |   1.188e+00 |   1.000e+00 |   Y
|       12 |  4.059690e+00 |   1.584e-01 |   5.213e-01 |    OK  |   9.930e-01 |   1.000e+00 |   Y
|       13 |  4.029208e+00 |   7.411e-02 |   2.076e-01 |    OK  |   4.886e-01 |   1.000e+00 |   Y
|       14 |  4.016358e+00 |   1.068e-01 |   2.696e-01 |    OK  |   6.919e-01 |   1.000e+00 |   Y
|       15 |  4.004521e+00 |   5.434e-02 |   1.136e-01 |    OK  |   5.647e-01 |   1.000e+00 |   Y
|       16 |  3.986929e+00 |   6.158e-02 |   2.993e-01 |    OK  |   1.353e+00 |   1.000e+00 |   Y
|       17 |  3.976342e+00 |   4.966e-02 |   2.213e-01 |    OK  |   7.668e-01 |   1.000e+00 |   Y
|       18 |  3.966646e+00 |   5.458e-02 |   2.529e-01 |    OK  |   1.988e+00 |   1.000e+00 |   Y
|       19 |  3.959586e+00 |   1.046e-01 |   4.169e-01 |    OK  |   1.858e+00 |   1.000e+00 |   Y

|=========================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=========================================================================================================
|       20 |  3.953759e+00 |   8.248e-02 |   2.892e-01 |    OK  |   1.040e+00 |   1.000e+00 |   Y
|       21 |  3.945475e+00 |   3.119e-02 |   1.698e-01 |    OK  |   1.095e+00 |   1.000e+00 |   Y
|       22 |  3.941567e+00 |   2.350e-02 |   1.293e-01 |    OK  |   1.117e+00 |   1.000e+00 |   Y
|       23 |  3.939468e+00 |   1.296e-02 |   1.805e-01 |    OK  |   2.287e+00 |   1.000e+00 |   Y
|       24 |  3.938662e+00 |   8.591e-03 |   5.955e-02 |    OK  |   1.553e+00 |   1.000e+00 |   Y
|       25 |  3.938239e+00 |   6.421e-03 |   5.334e-02 |    OK  |   1.102e+00 |   1.000e+00 |   Y
|       26 |  3.938013e+00 |   5.449e-03 |   6.773e-02 |    OK  |   2.085e+00 |   1.000e+00 |   Y
|       27 |  3.937896e+00 |   6.226e-03 |   3.368e-02 |    OK  |   7.541e-01 |   1.000e+00 |   Y
|       28 |  3.937820e+00 |   2.497e-03 |   2.397e-02 |    OK  |   7.940e-01 |   1.000e+00 |   Y
|       29 |  3.937791e+00 |   2.004e-03 |   1.339e-02 |    OK  |   1.863e+00 |   1.000e+00 |   Y
|       30 |  3.937784e+00 |   2.448e-03 |   1.265e-02 |    OK  |   9.667e-01 |   1.000e+00 |   Y
|       31 |  3.937778e+00 |   6.973e-04 |   2.906e-03 |    OK  |   4.672e-01 |   1.000e+00 |   Y
|       32 |  3.937778e+00 |   3.038e-04 |   9.502e-04 |    OK  |   1.060e+00 |   1.000e+00 |   Y
|       33 |  3.937777e+00 |   2.327e-04 |   1.069e-03 |    OK  |   1.597e+00 |   1.000e+00 |   Y
|       34 |  3.937777e+00 |   1.959e-04 |   1.537e-03 |    OK  |   4.026e+00 |   1.000e+00 |   Y
|       35 |  3.937777e+00 |   1.162e-04 |   1.464e-03 |    OK  |   3.418e+00 |   1.000e+00 |   Y
|       36 |  3.937777e+00 |   8.353e-05 |   3.660e-04 |    OK  |   7.304e-01 |   5.000e-01 |   Y
|       37 |  3.937777e+00 |   1.412e-05 |   1.412e-04 |    OK  |   7.842e-01 |   1.000e+00 |   Y
|       38 |  3.937777e+00 |   1.277e-05 |   3.808e-05 |    OK  |   1.021e+00 |   1.000e+00 |   Y
|       39 |  3.937777e+00 |   8.614e-06 |   3.698e-05 |    OK  |   2.561e+00 |   1.000e+00 |   Y

|=========================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=========================================================================================================
|       40 |  3.937777e+00 |   3.159e-06 |   5.299e-05 |    OK  |   4.331e+00 |   1.000e+00 |   Y
|       41 |  3.937777e+00 |   2.657e-06 |   1.080e-05 |    OK  |   7.038e-01 |   5.000e-01 |   Y
|       42 |  3.937777e+00 |   7.054e-07 |   7.036e-06 |    OK  |   9.519e-01 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 7.054e-07
             Two norm of the final step     = 7.036e-06, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 7.054e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

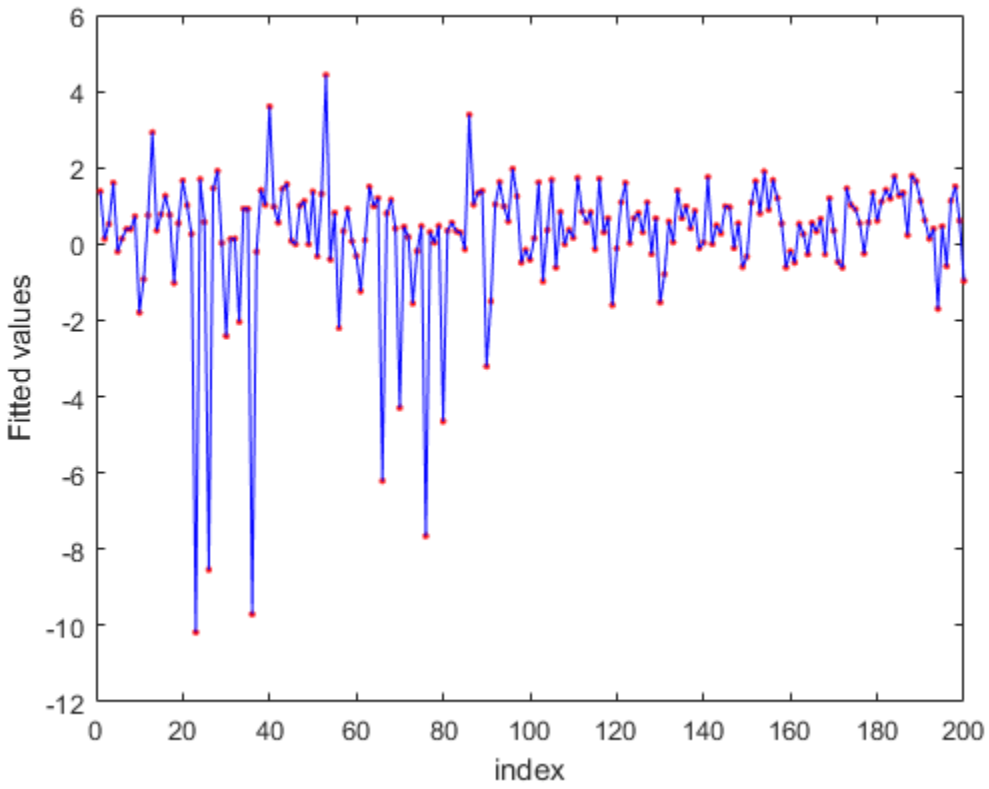
Plot selected features.

```
figure
plot(nca.FeatureWeights,'ro')
grid on
xlabel('Feature index')
ylabel('Feature weight')
```



Predict the response values using the `nca` model and plot the fitted (predicted) response values and the actual response values.

```
figure
fitted = predict(nca,X);
plot(y,'r.')
hold on
plot(fitted,'b-')
xlabel('index')
ylabel('Fitted values')
```



`fsrnca` tries to fit every point in data including the outliers. As a result it assigns nonzero weights to many features besides predictors 4, 7, and 13.

Use built-in robust loss function

Repeat the same process of tuning the regularization parameter, this time using the built-in ϵ-insensitive loss function:

$$l(y_i, y_j) = \max(0, |y_i - y_j| - \epsilon)$$

ϵ-insensitive loss function is more robust to outliers than mean squared error.

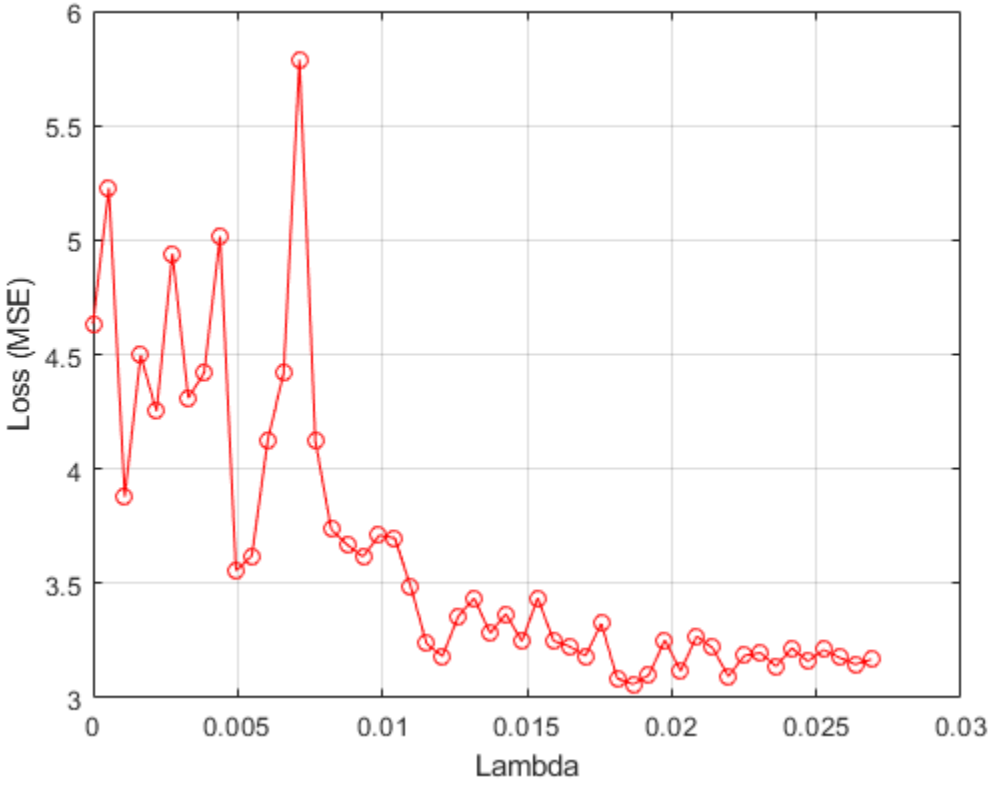
```
lambdavals = linspace(0,3,50)*std(y)/length(y);
cvp = cvpartition(length(y),'kfold',5);
numtestsets = cvp.NumTestSets;
lossvals = zeros(length(lambdavals),numtestsets);

for i = 1:length(lambdavals)
    for k = 1:numtestsets
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);
        Xtest = X(cvp.test(k),:);
        ytest = y(cvp.test(k),:);

        nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
            'Solver','sgd','Verbose',0,'Lambda',lambdavals(i), ...
            'LossFunction','epsiloninsensitive','Epsilon',0.8);
```

```
        lossvals(i,k) = loss(nca,Xtest,ytest,'LossFunction','mse');
    end
end
```

The ϵ value to use depends on the data and the best value can be determined using cross-validation as well. But choosing the ϵ value is out of scope of this example. The choice of ϵ in this example is mainly for illustrating the robustness of the method.

Plot the mean loss corresponding to each lambda value.

```
figure
meanloss = mean(lossvals,2);
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the lambda value that produces the minimum average loss.

```
[~,idx] = min(mean(lossvals,2));
bestlambda = lambdavals(idx)
```

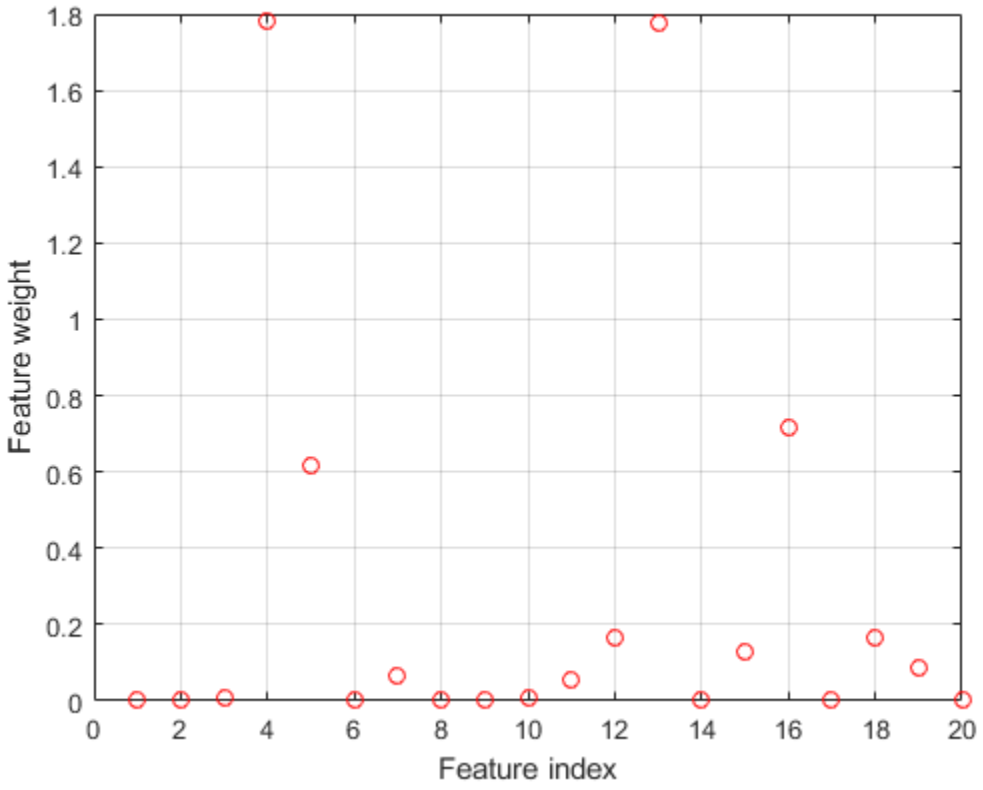
```
bestlambda = 0.0187
```

Fit neighborhood component analysis model using ϵ-insensitive loss function and best lambda value.

```
nca = fsrnca(X,y,'FitMethod','exact','Solver','sgd', ...
    'Lambda',bestlambda,'LossFunction','epsiloninsensitive','Epsilon',0.8);
```

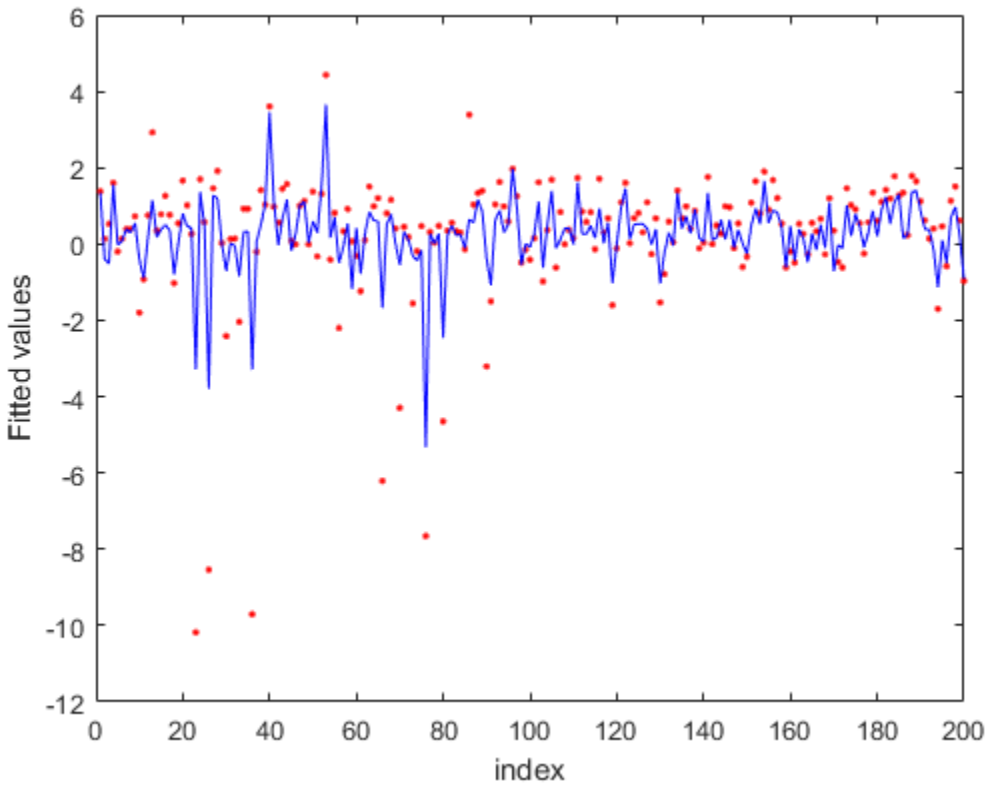
Plot selected features.

```
figure
plot(nca.FeatureWeights,'ro')
grid on
xlabel('Feature index')
ylabel('Feature weight')
```



Plot fitted values.

```
figure
fitted = predict(nca,X);
plot(y,'r.')
hold on
plot(fitted,'b-')
xlabel('index')
ylabel('Fitted values')
```



ϵ-insensitive loss seems more robust to outliers. It identified fewer features than mse as relevant. The fit shows that it is still impacted by some of the outliers.

Use custom robust loss function

Define a custom robust loss function that is robust to outliers to use in feature selection for regression:

$$f(y_i, y_j) = 1 - \exp(-|y_i - y_j|)$$

```
customlossFcn = @(yi,yj) 1 - exp(-abs(yi-yj'));
```

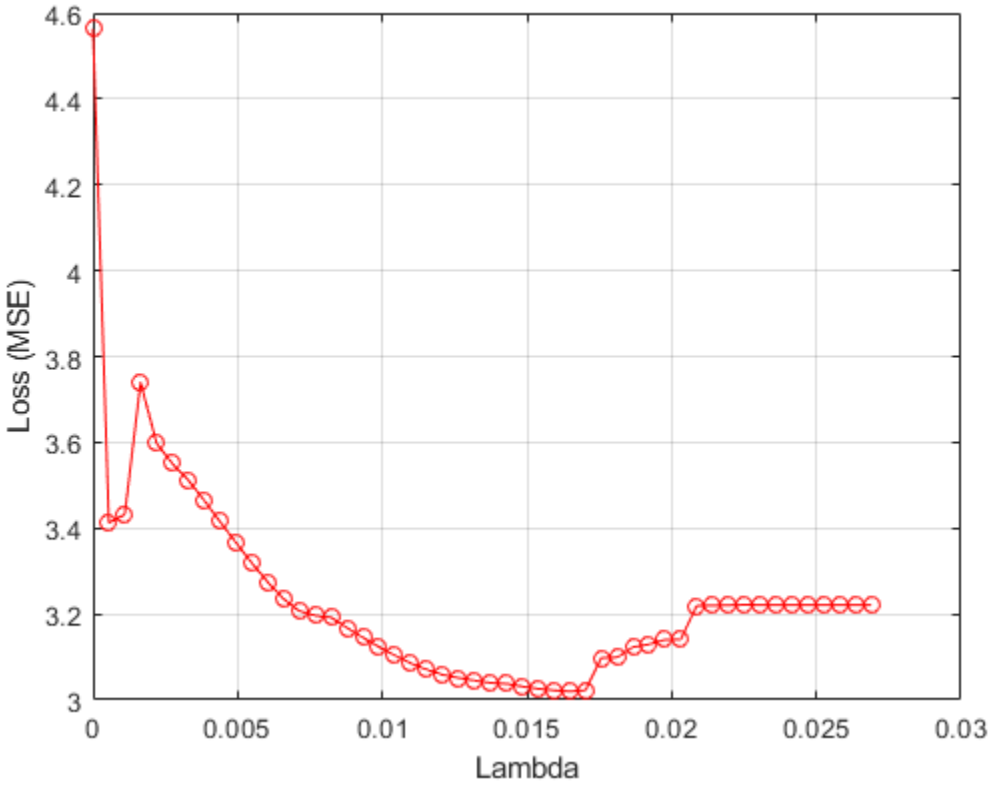
Tune the regularization parameter using the custom-defined robust loss function.

```
lambdavals = linspace(0,3,50)*std(y)/length(y);
cvp = cvpartition(length(y),'kfold',5);
numtestsets = cvp.NumTestSets;
lossvals = zeros(length(lambdavals),numtestsets);

for i = 1:length(lambdavals)
    for k = 1:numtestsets
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);
        Xtest = X(cvp.test(k),:);
        ytest = y(cvp.test(k),:);

        nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
```

```
            'Solver','lbfgs','Verbose',0,'Lambda',lambdavals(i), ...
            'LossFunction',customlossFcn);

        lossvals(i,k) = loss(nca,Xtest,ytest,'LossFunction','mse');
    end
end
```

Plot the mean loss corresponding to each lambda value.

```
figure
meanloss = mean(lossvals,2);
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the λ value that produces the minimum average loss.

```
[~,idx] = min(mean(lossvals,2));
bestlambda = lambdavals(idx)
```

```
bestlambda = 0.0165
```

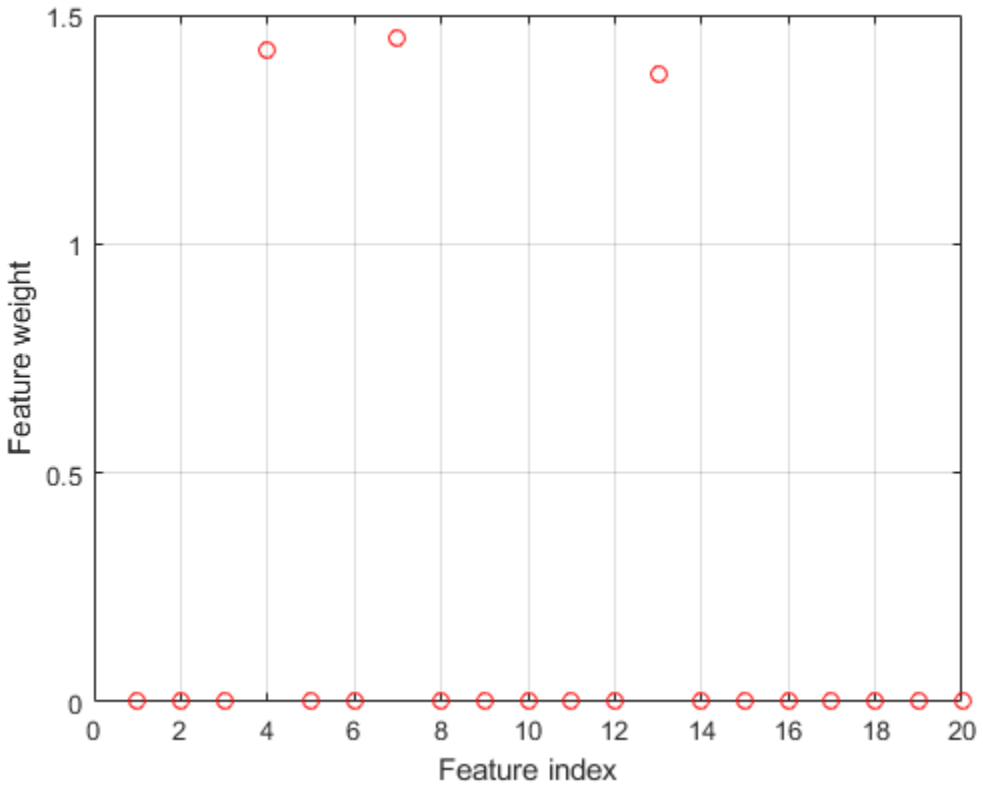
Perform feature selection using the custom robust loss function and best λ value.

```
nca = fsrnca(X,y,'FitMethod','exact','Solver','lbfgs', ...
    'Verbose',1,'Lambda',bestlambda,'LossFunction',customlossFcn);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=============================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=============================================================================================
|        0 |  8.610073e-01 |   4.921e-02 |   0.000e+00 |        |   1.219e+01 |   0.000e+00 |
|        1 |  6.582278e-01 |   2.328e-02 |   1.820e+00 |    OK  |   2.177e+01 |   1.000e+00 |
|        2 |  5.706490e-01 |   2.241e-02 |   2.360e+00 |    OK  |   2.541e+01 |   1.000e+00 |
|        3 |  5.677090e-01 |   2.666e-02 |   7.583e-01 |    OK  |   1.092e+01 |   1.000e+00 |
|        4 |  5.620806e-01 |   5.524e-03 |   3.335e-01 |    OK  |   9.973e+00 |   1.000e+00 |
|        5 |  5.616054e-01 |   1.428e-03 |   1.025e-01 |    OK  |   1.736e+01 |   1.000e+00 |
|        6 |  5.614779e-01 |   4.446e-04 |   8.350e-02 |    OK  |   2.507e+01 |   1.000e+00 |
|        7 |  5.614653e-01 |   4.118e-04 |   2.466e-02 |    OK  |   2.105e+01 |   1.000e+00 |
|        8 |  5.614620e-01 |   1.307e-04 |   1.373e-02 |    OK  |   2.002e+01 |   1.000e+00 |
|        9 |  5.614615e-01 |   9.318e-05 |   4.128e-03 |    OK  |   3.683e+01 |   1.000e+00 |
|       10 |  5.614611e-01 |   4.579e-05 |   8.785e-03 |    OK  |   6.170e+01 |   1.000e+00 |
|       11 |  5.614610e-01 |   1.232e-05 |   1.582e-03 |    OK  |   2.000e+01 |   5.000e-01 |
|       12 |  5.614610e-01 |   3.174e-06 |   4.742e-04 |    OK  |   2.510e+01 |   1.000e+00 |
|       13 |  5.614610e-01 |   7.896e-07 |   1.683e-04 |    OK  |   2.959e+01 |   1.000e+00 |

         Infinity norm of the final gradient = 7.896e-07
             Two norm of the final step     = 1.683e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 7.896e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

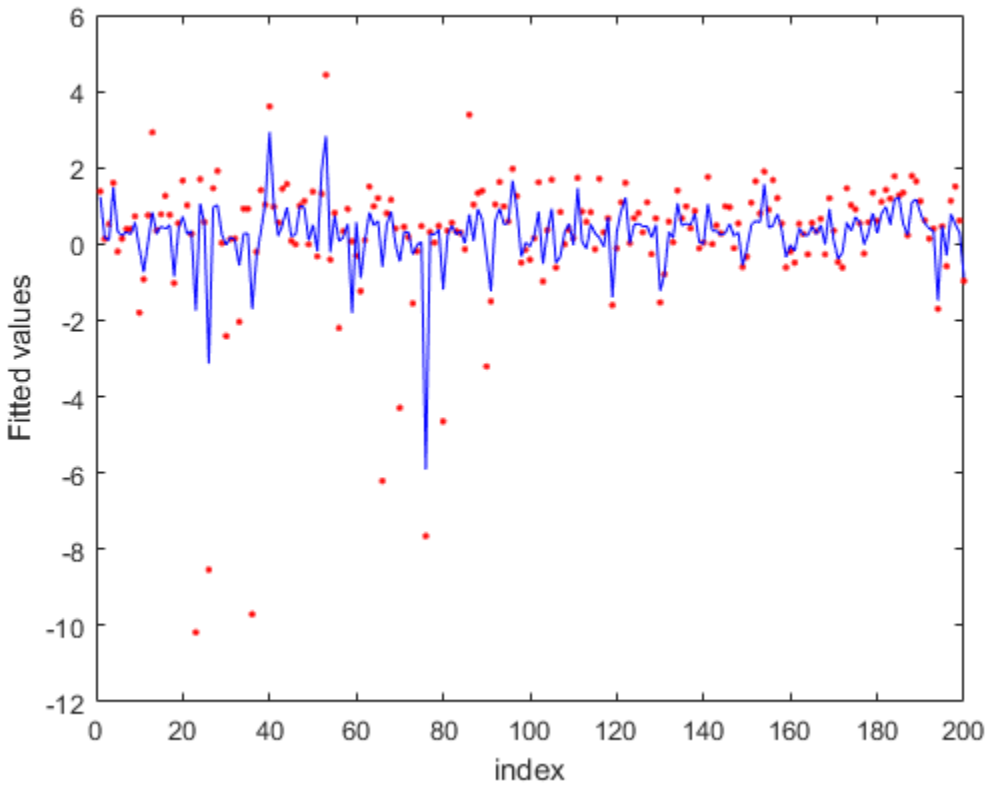
Plot selected features.

```
figure
plot(nca.FeatureWeights,'ro')
grid on
xlabel('Feature index')
ylabel('Feature weight')
```



Plot fitted values.

```
figure
fitted = predict(nca,X);
plot(y,'r.')
hold on
plot(fitted,'b-')
xlabel('index')
ylabel('Fitted values')
```



In this case, the loss is not affected by the outliers and results are based on most of the observation values. `fsrnca` detects the predictors 4, 7, and 13 as relevant features and does not select any other features.

Why does the loss function choice affect the results?

First, compute the loss functions for a series of values for the difference between two observations.

```
deltay = linspace(-10,10,1000)';
```

Compute custom loss function values.

```
customlossvals = customlossFcn(deltay,0);
```

Compute epsilon insensitive loss function and values.

```
epsinsensitive = @(yi,yj,E) max(0,abs(yi-yj')-E);
epsinsenvals = epsinsensitive(deltay,0,0.5);
```

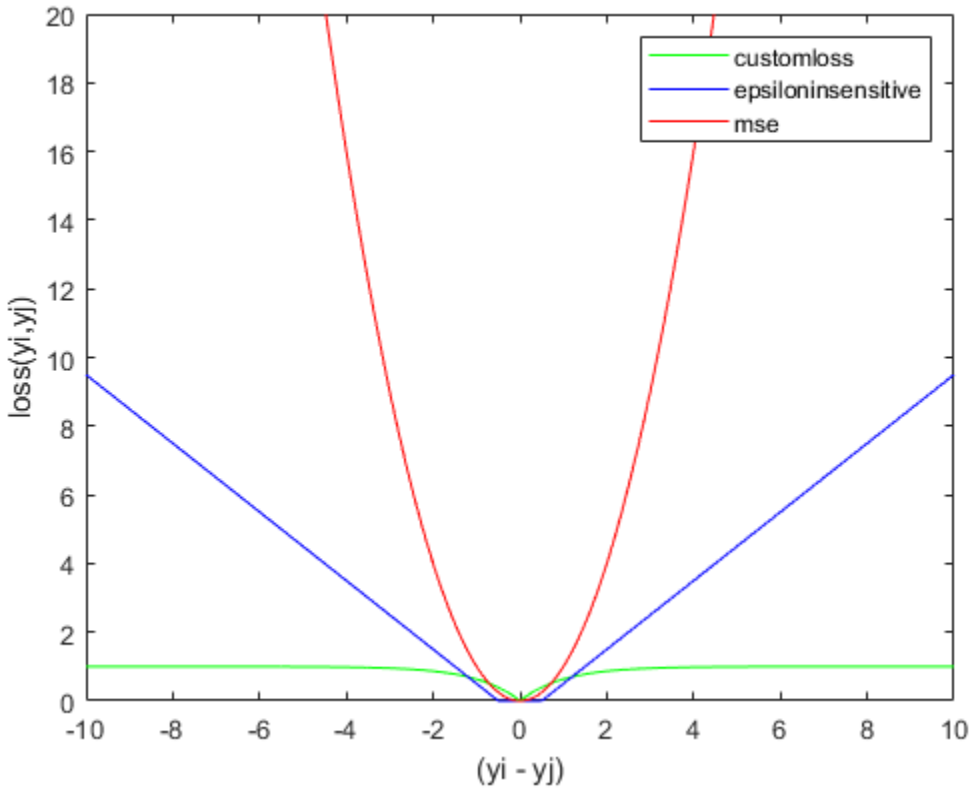
Compute MSE loss function and values.

```
mse = @(yi,yj) (yi-yj').^2;
msevals = mse(deltay,0);
```

Now, plot the loss functions to see their difference and why they affect the results in the way they do.

```
figure
plot(deltay,customlossvals,'g-',deltay,epsinsenvals,'b-',deltay,msevals,'r-')
```

```
xlabel('(yi - yj)')
ylabel('loss(yi,yj)')
legend('customloss','epsiloninsensitive','mse')
ylim([0 20])
```



As the difference between two response values increases, mse increases quadratically, which makes it very sensitive to outliers. As `fsrnca` tries to minimize this loss, it ends up identifying more features as relevant. The epsilon insensitive loss is more resistant to outliers than mse, but eventually it does start to increase linearly as the difference between two observations increase. As the difference between two observations increase, the robust loss function does approach 1 and stays at that value even though the difference between the observations keeps increasing. Out of three, it is the most robust to outliers.

See Also

`FeatureSelectionNCARegression` | `fsrnca` | `loss` | `predict` | `refit`

More About

- “Neighborhood Component Analysis (NCA) Feature Selection” on page 15-99
- “Introduction to Feature Selection” on page 15-49

Neighborhood Component Analysis (NCA) Feature Selection

In this section...

Neighborhood component analysis (NCA) is a non-parametric method for selecting features with the goal of maximizing prediction accuracy of regression and classification algorithms. The Statistics and Machine Learning Toolbox functions `fscnca` and `fsrnca` perform NCA feature selection with regularization to learn feature weights for minimization of an objective function that measures the average leave-one-out classification or regression loss over the training data.

NCA Feature Selection for Classification

Consider a multi-class classification problem with a training set containing n observations:

$$S = \{(x_i, y_i), i = 1, 2, ..., n\},$$

where $x_i \in \mathbb{R}^p$ are the feature vectors, $y_i \in \{1, 2, ..., c\}$ are the class labels, and c is the number of classes. The aim is to learn a classifier $f: \mathbb{R}^p \to \{1, 2, ..., c\}$ that accepts a feature vector and makes a prediction $f(x)$ for the true label y of x.

Consider a randomized classifier that:

- Randomly picks a point, Ref(x), from S as the ‘reference point’ for x
- Labels x using the label of the reference point Ref(x).

This scheme is similar to that of a 1-NN classifier where the reference point is chosen to be the nearest neighbor of the new point x. In NCA, the reference point is chosen randomly and all points in S have some probability of being selected as the reference point. The probability $P\left(\text{Ref}(x) = x_j \middle| S\right)$ that point x_j is picked from S as the reference point for x is higher if x_j is closer to x as measured by the distance function d_w, where

$$d_w(x_i, x_j) = \sum_{r=1}^{p} w_r^2 \left| x_{ir} - x_{jr} \right|,$$

and w_r are the feature weights. Assume that

$$P\left(\text{Ref}(x) = x_j \middle| S\right) \propto k\left(d_w\left(x, x_j\right)\right),$$

where k is some kernel or a similarity function that assumes large values when $d_w\left(x, x_j\right)$ is small. Suppose it is

$$k(z) = \exp\left(-\frac{z}{\sigma}\right),$$

as suggested in [1]. The reference point for x is chosen from S, so sum of $P\left(\text{Ref}(x) = x_j \middle| S\right)$ for all j must be equal to 1. Therefore, it is possible to write

$$P\left(\text{Ref}(x) = x_j \middle| S\right) = \frac{k\left(d_w\left(x, x_j\right)\right)}{\sum_{j=1}^{n} k\left(d_w\left(x, x_j\right)\right)}.$$

Now consider the leave-one-out application of this randomized classifier, that is, predicting the label of x_i using the data in S^{-i}, the training set S excluding the point (x_i, y_i). The probability that point x_j is picked as the reference point for x_i is

$$p_{ij} = P\left(\text{Ref}(x_i) = x_j \middle| S^{-i}\right) = \frac{k\left(d_w\left(x_i, x_j\right)\right)}{\sum_{j=1, j\neq i}^{n} k\left(d_w\left(x_i, x_j\right)\right)}.$$

The average leave-one-out probability of correct classification is the probability p_i that the randomized classifier correctly classifies observation i using S^{-i}.

$$p_i = \sum_{j=1, j\neq i}^{n} P\left(\text{Ref}(x_i) = x_j \middle| S^{-i}\right) I\left(y_i = y_j\right) = \sum_{j=1, j\neq i}^{n} p_{ij} y_{ij},$$

where

$$y_{ij} = I\left(y_i = y_j\right) = \begin{cases} 1 & \text{if } y_i = y_j, \\ 0 & \text{otherwise.} \end{cases}$$

The average leave-one-out probability of correct classification using the randomized classifier can be written as

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} p_i.$$

The right hand side of $F(w)$ depends on the weight vector w. The goal of neighborhood component analysis is to maximize $F(w)$ with respect to w. `fscnca` uses the regularized objective function as introduced in [1].

$$\begin{aligned} F(w) &= \frac{1}{n} \sum_{i=1}^{n} p_i - \lambda \sum_{r=1}^{p} w_r^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \Bigg[\underbrace{\sum_{j=1, j\neq i}^{n} p_{ij} y_{ij}}_{F_i(w)} - \lambda \sum_{r=1}^{p} w_r^2 \Bigg], \\ &= \frac{1}{n} \sum_{i=1}^{n} F_i(w) \end{aligned}$$

where λ is the regularization parameter. The regularization term drives many of the weights in w to 0.

After choosing the kernel parameter σ in p_{ij} as 1, finding the weight vector w can be expressed as the following minimization problem for given λ.

$$\widehat{w} = \underset{w}{\operatorname{argmin}} f(w) = \underset{w}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} f_i(w),$$

where $f(w) = -F(w)$ and $f_i(w) = -F_i(w)$.

Note that

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} = 1,$$

and the argument of the minimum does not change if you add a constant to an objective function. Therefore, you can rewrite the objective function by adding the constant 1.

$$\begin{aligned}
\widehat{w} &= \underset{w}{\operatorname{argmin}} \{1 + f(w)\} \\
&= \underset{w}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} y_{ij} + \lambda \sum_{r=1}^{p} w_r^2 \right\} \\
&= \underset{w}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} (1 - y_{ij}) + \lambda \sum_{r=1}^{p} w_r^2 \right\} \\
&= \underset{w}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} l(y_i, y_j) + \lambda \sum_{r=1}^{p} w_r^2 \right\},
\end{aligned}$$

where the loss function is defined as

$$l(y_i, y_j) = \begin{cases} 1 & \text{if } y_i \neq y_j, \\ 0 & \text{otherwise.} \end{cases}$$

The argument of the minimum is the weight vector that minimizes the classification error. You can specify a custom loss function using the `LossFunction` name-value pair argument in the call to `fscnca`.

NCA Feature Selection for Regression

The `fsrnca` function performs NCA feature selection modified for regression. Given n observations

$$S = \{(x_i, y_i), i = 1, 2, \ldots, n\},$$

the only difference from the classification problem is that the response values $y_i \in \mathbb{R}$ are continuous. In this case, the aim is to predict the response y given the training set S.

Consider a randomized regression model that:

- Randomly picks a point (Ref(x)) from Sas the 'reference point' for x
- Sets the response value at x equal to the response value of the reference point Ref(x).

Again, the probability $P\left(\text{Ref}(x) = x_j \middle| S\right)$ that point x_j is picked from S as the reference point for x is

$$P\left(\text{Ref}(x) = x_j \middle| S\right) = \frac{k\left(d_w\left(x, x_j\right)\right)}{\sum_{j=1}^{n} k\left(d_w\left(x, x_j\right)\right)}.$$

Now consider the leave-one-out application of this randomized regression model, that is, predicting the response for x_i using the data in S^{-i}, the training set S excluding the point (x_i, y_i). The probability that point x_j is picked as the reference point for x_i is

$$p_{ij} = P\left(\text{Ref}(x_i) = x_j \middle| S^{-i}\right) = \frac{k\left(d_w\left(x_i, x_j\right)\right)}{\sum_{j=1, j\neq i}^{n} k\left(d_w\left(x_i, x_j\right)\right)}.$$

Let $\widehat{y}_i$ be the response value the randomized regression model predicts and y_i be the actual response for x_i. And let $l: \mathbb{R}^2 \to \mathbb{R}$ be a loss function that measures the disagreement between $\widehat{y}_i$ and y_i. Then, the average value of $l(y_i, \widehat{y}_i)$ is

$$l_i = E\left(l(y_i, \widehat{y}_i) \middle| S^{-i}\right) = \sum_{j=1, j\neq i}^{n} p_{ij} l\left(y_i, y_j\right).$$

After adding the regularization term, the objective function for minimization is:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} l_i + \lambda \sum_{r=1}^{p} w_r^2.$$

The default loss function $l\left(y_i, y_j\right)$ for NCA for regression is mean absolute deviation, but you can specify other loss functions, including a custom one, using the `LossFunction` name-value pair argument in the call to `fsrnca`.

Impact of Standardization

The regularization term derives the weights of irrelevant predictors to zero. In the objective functions for NCA for classification or regression, there is only one regularization parameter λ for all weights. This fact requires the magnitudes of the weights to be comparable to each other. When the feature vectors x_i in S are in different scales, this might result in weights that are in different scales and not meaningful. To avoid this situation, standardize the predictors to have zero mean and unit standard deviation before applying NCA. You can standardize the predictors using the `'Standardize',true` name-value pair argument in the call to `fscnca` or `fsrnca`.

Choosing the Regularization Parameter Value

It is usually necessary to select a value of the regularization parameter by calculating the accuracy of the randomized NCA classifier or regression model on an independent test set. If you use cross-validation instead of a single test set, select the λ value that minimizes the average loss across the cross-validation folds. For examples, see “Tune Regularization Parameter to Detect Features Using NCA for Classification” on page 15-210 and “Tune Regularization Parameter in NCA for Regression” on page 33-2503.

References

[1] Yang, W., K. Wang, W. Zuo. "Neighborhood Component Feature Selection for High-Dimensional Data." Journal of Computers. Vol. 7, Number 1, January, 2012.

See Also

`FeatureSelectionNCAClassification` | `FeatureSelectionNCARegression` | `fscnca` | `fsrnca`

More About

- “Robust Feature Selection Using NCA for Regression” on page 15-85
- “Tune Regularization Parameter to Detect Features Using NCA for Classification” on page 15-210
- “Introduction to Feature Selection” on page 15-49

t-SNE

In this section...

What Is t-SNE?

t-SNE (`tsne`) is an algorithm for dimensionality reduction that is well-suited to visualizing high-dimensional data. The name stands for *t*-distributed Stochastic Neighbor Embedding. The idea is to embed high-dimensional points in low dimensions in a way that respects similarities between points. Nearby points in the high-dimensional space correspond to nearby embedded low-dimensional points, and distant points in high-dimensional space correspond to distant embedded low-dimensional points. (Generally, it is impossible to match distances exactly between high-dimensional and low-dimensional spaces.)

The `tsne` function creates a set of low-dimensional points from high-dimensional data. Typically, you visualize the low-dimensional points to see natural clusters in the original high-dimensional data.

The algorithm takes the following general steps to embed the data in low dimensions.

1. Calculate the pairwise distances between the high-dimensional points.
2. Create a standard deviation σ_i for each high-dimensional point i so that the perplexity of each point is at a predetermined level. For the definition of perplexity, see “Compute Distances, Gaussian Variances, and Similarities” on page 15-105.
3. Calculate the similarity matrix. This is the joint probability distribution of X, defined by “Equation 15-2”.
4. Create an initial set of low-dimensional points.
5. Iteratively update the low-dimensional points to minimize the Kullback-Leibler divergence between a Gaussian distribution in the high-dimensional space and a t distribution in the low-dimensional space. This optimization procedure is the most time-consuming part of the algorithm.

See van der Maaten and Hinton [1].

t-SNE Algorithm

The basic t-SNE algorithm performs the following steps.

- “Prepare Data” on page 15-105
- “Compute Distances, Gaussian Variances, and Similarities” on page 15-105
- “Initialize the Embedding and Divergence” on page 15-106
- “Gradient Descent of Kullback-Leibler Divergence” on page 15-106

Prepare Data

`tsne` first removes each row of the input data X that contains any `NaN` values. Then, if the `Standardize` name-value pair is `true`, `tsne` centers X by subtracting the mean of each column, and scales X by dividing its columns by their standard deviations.

The original authors van der Maaten and Hinton [1] recommend reducing the original data X to a lower-dimensional version using “Principal Component Analysis (PCA)” on page 15-68. You can set the `tsne` `NumPCAComponents` name-value pair to the number of dimensions you like, perhaps 50. To exercise more control over this step, preprocess the data using the `pca` function.

Compute Distances, Gaussian Variances, and Similarities

After the preprocessing, `tsne` calculates the distance $d(x_i,x_j)$ between each pair of points x_i and x_j in X. You can choose various distance metrics using the `Distance` name-value pair. By default, `tsne` uses the standard Euclidean metric. `tsne` uses the square of the distance metric in its subsequent calculations.

Then for each row i of X, `tsne` calculates a standard deviation σ_i so that the perplexity of row i is equal to the `Perplexity` name-value pair. The perplexity is defined in terms of a model Gaussian distribution as follows. As van der Maaten and Hinton [1] describe, “The similarity of data point x_j to data point x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. For nearby data points, $p_{j|i}$ is relatively high, whereas for widely separated data points, $p_{j|i}$ will be almost infinitesimal (for reasonable values of the variance of the Gaussian, σ_i).”

Define the conditional probability of j given i as

$$p_{j|i} = \frac{\exp\left(-d\left(x_i, x_j\right)^2/\left(2\sigma_i^2\right)\right)}{\sum_{k \neq i} \exp\left(-d\left(x_i, x_k\right)^2/\left(2\sigma_i^2\right)\right)}$$

$$p_{i|i} = 0.$$

Then define the joint probability p_{ij} by symmetrizing the conditional probabilities:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \tag{15-2}$$

where N is the number of rows of X.

The distributions still do not have their standard deviations σ_i defined in terms of the `Perplexity` name-value pair. Let P_i represents the conditional probability distribution over all other data points given data point x_i. The perplexity of the distribution is

$$\text{perplexity}(P_i) = 2^{H(P_i)},$$

where $H(P_i)$ is the Shannon entropy of P_i:

$$H(P_i) = -\sum_j p_{j|i} \log_2\left(p_{j|i}\right).$$

The perplexity measures the effective number of neighbors of point i. `tsne` performs a binary search over the σ_i to achieve a fixed perplexity for each point i.

Initialize the Embedding and Divergence

To embed the points in X into a low-dimensional space, `tsne` performs an optimization. `tsne` attempts to minimize the Kullback-Leibler divergence between the model Gaussian distribution of the points in X and a Student t distribution of points Y in the low-dimensional space.

The minimization procedure begins with an initial set of points Y. `tsne` create the points by default as random Gaussian-distributed points. You can also create these points yourself and include them in the `'InitialY'` name-value pair for `tsne`. `tsne` then calculates the similarities between each pair of points in Y.

The probability model q_{ij} of the distribution of the distances between points y_i and y_j is

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

$$q_{ii} = 0.$$

Using this definition and the model of distances in X given by “Equation 15-2”, the Kullback-Leibler divergence between the joint distribution P and Q is

$$KL(P||Q) = \sum_j \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

For consequences of this definition, see “Helpful Nonlinear Distortion” on page 15-107.

Gradient Descent of Kullback-Leibler Divergence

To minimize the Kullback-Leibler divergence, the `'exact'` algorithm uses a modified gradient descent procedure. The gradient with respect to the points in Y of the divergence is

$$\frac{\partial KL(P||Q)}{\partial y_i} = 4 \sum_{j \neq i} Z(p_{ij} - q_{ij}) q_{ij} (y_i - y_j),$$

where the normalization term

$$Z = \sum_k \sum_{l \neq k} \left(1 + \|y_k - y_l\|^2\right)^{-1}.$$

The modified gradient descent algorithm uses a few tuning parameters to attempt to reach a good local minimum.

- `'Exaggeration'` — During the first 99 gradient descent steps, `tsne` multiplies the probabilities p_{ij} from “Equation 15-2” by the exaggeration value. This step tends to create more space between clusters in the output Y.
- `'LearnRate'` — `tsne` uses adaptive learning to improve the convergence of the gradient descent iterations. The descent algorithm has iterative steps that are a linear combination of the previous step in the descent and the current gradient. `'LearnRate'` is a multiplier of the current gradient for the linear combination. For details, see Jacobs [3].

Barnes-Hut Variation of t-SNE

To speed the t-SNE algorithm and to cut down on its memory usage, `tsne` offers an approximate optimization scheme. The Barnes-Hut algorithm groups nearby points together to lower the complexity and memory usage of the t-SNE optimization step. The Barnes-Hut algorithm is an approximate optimizer, not an exact optimizer. There is a nonnegative tuning parameter `Theta` that effects a tradeoff between speed and accuracy. Larger values of `'Theta'` give faster but less accurate optimization results. The algorithm is relatively insensitive to `'Theta'` values in the range (0.2,0.8).

The Barnes-Hut algorithm groups nearby points in the low-dimensional space, and performs an approximate gradient descent based on these groups. The idea, originally used in astrophysics, is that the gradient is similar for nearby points, so the computations can be simplified.

See van der Maaten [2].

Characteristics of t-SNE

- “Cannot Use Embedding to Classify New Data” on page 15-107
- “Performance Depends on Data Sizes and Algorithm” on page 15-107
- “Helpful Nonlinear Distortion” on page 15-107

Cannot Use Embedding to Classify New Data

Because t-SNE often separates data clusters well, it can seem that t-SNE can classify new data points. However, t-SNE cannot classify new points. The t-SNE embedding is a nonlinear map that is data-dependent. To embed a new point in the low-dimensional space, you cannot use the previous embedding as a map. Instead, run the entire algorithm again.

Performance Depends on Data Sizes and Algorithm

t-SNE can take a good deal of time to process data. If you have N data points in D dimensions that you want to map to Y dimensions, then

- Exact t-SNE takes of order $D*N^2$ operations.
- Barnes-Hut t-SNE takes of order $D*N\log(N)*\exp(\text{dimension}(Y))$ operations.

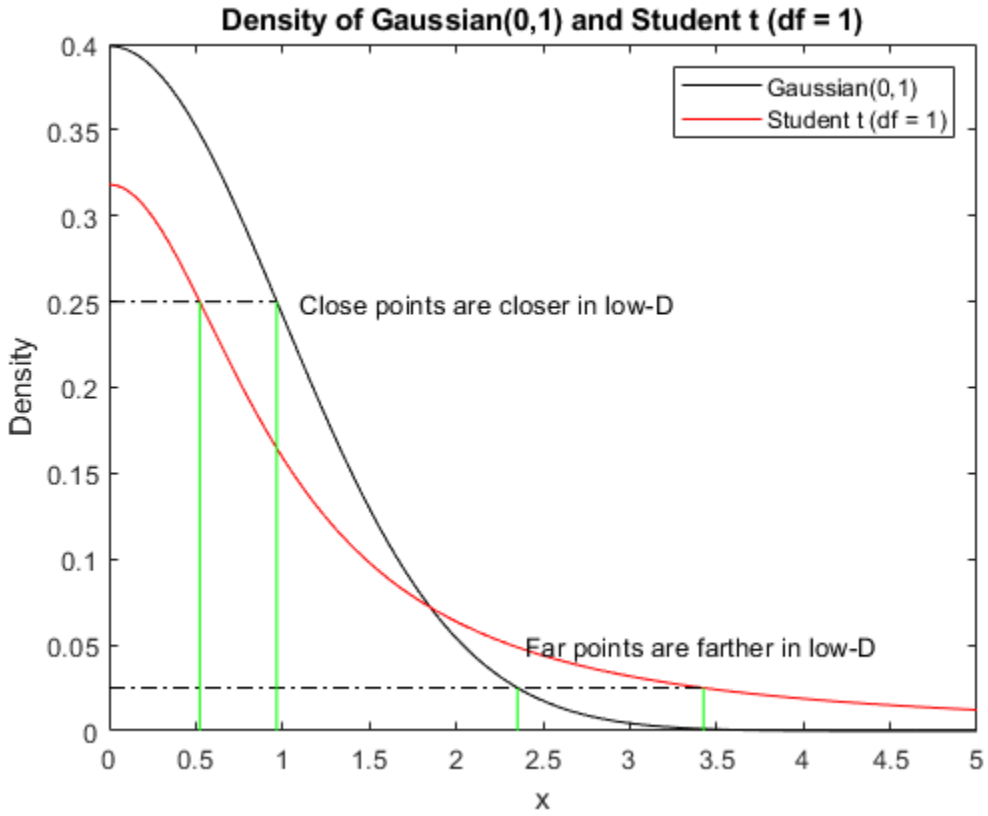
So for large data sets, where N is greater than 1000 or so, and where the embedding dimension Y is 2 or 3, the Barnes-Hut algorithm can be faster than the exact algorithm.

Helpful Nonlinear Distortion

T-SNE maps high-dimensional distances to distorted low-dimensional analogues. Because of the fatter tail of the Student t distribution in the low-dimensional space, `tsne` often moves close points closer together, and moves far points farther apart than in the high-dimensional space, as illustrated in the following figure. The figure shows both Gaussian and Student t distributions at the points where the densities are at 0.25 and 0.025. The Gaussian density relates to high-dimensional distances, and the t density relates to low-dimensional distances. The t density corresponds to close points being closer, and far points being farther, compared to the Gaussian density.

```
t = linspace(0,5);
y1 = normpdf(t,0,1);
y2 = tpdf(t,1);
```

```
plot(t,y1,'k',t,y2,'r')
hold on
x1 = fzero(@(x)normpdf(x,0,1)-0.25,[0,2]);
x2 = fzero(@(x)tpdf(x,1)-0.25,[0,2]);
z1 = fzero(@(x)normpdf(x,0,1)-0.025,[0,5]);
z2 = fzero(@(x)tpdf(x,1)-0.025,[0,5]);
plot([0,x1],[0.25,0.25],'k-.')
plot([0,z2],[0.025,0.025],'k-.')
plot([x1,x1],[0,0.25],'g-',[x2,x2],[0,0.25],'g-')
plot([z1,z1],[0,0.025],'g-',[z2,z2],[0,0.025],'g-')
text(1.1,.25,'Close points are closer in low-D')
text(2.4,.05,'Far points are farther in low-D')
legend('Gaussian(0,1)','Student t (df = 1)')
xlabel('x')
ylabel('Density')
title('Density of Gaussian(0,1) and Student t (df = 1)')
hold off
```



This distortion is helpful when it applies. It does not apply in cases such as when the Gaussian variance is high, which lowers the Gaussian peak and flattens the distribution. In such a case, `tsne` can move close points farther apart than in the original space. To achieve a helpful distortion,

- Set the `'Verbose'` name-value pair to 2.
- Adjust the `'Perplexity'` name-value pair so the reported range of variances is not too far from 1, and the mean variance is near 1.

If you can achieve this range of variances, then the diagram applies, and the `tsne` distortion is helpful.

For effective ways to tune `tsne`, see Wattenberg, Viégas and Johnson [4].

References

[1] van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing Data using *t*-SNE." *J. Machine Learning Research* 9, 2008, pp. 2579–2605.

[2] van der Maaten, Laurens. *Barnes-Hut-SNE*. arXiv:1301.3342 [cs.LG], 2013.

[3] Jacobs, Robert A. "Increased rates of convergence through learning rate adaptation." *Neural Networks* 1.4, 1988, pp. 295–307.

[4] Wattenberg, Martin, Fernanda Viégas, and Ian Johnson. "How to Use t-SNE Effectively." *Distill*, 2016. Available at `How to Use t-SNE Effectively`.

See Also

Related Examples

- “Visualize High-Dimensional Data Using t-SNE” on page 15-113
- “t-SNE Output Function” on page 15-110
- “tsne Settings” on page 15-117

More About

- “Principal Component Analysis (PCA)” on page 15-68
- “Classical Multidimensional Scaling” on page 15-40
- “Factor Analysis” on page 15-78

t-SNE Output Function

In this section...

t-SNE Output Function Description

A `tsne` output function is a function that runs after every `NumPrint` optimization iterations of the t-SNE algorithm. An output function can create plots, or log data to a file or to a workspace variable. The function cannot change the progress of the algorithm, but can halt the iterations.

Set output functions using the `Options` name-value pair argument to the `tsne` function. Set `Options` to a structure created using `statset` or `struct`. Set the `'OutputFcn'` field of the `Options` structure to a function handle or cell array of function handles.

For example, to set an output function named `outfun.m`, use the following commands.

```
opts = statset('OutputFcn',@outfun);
Y = tsne(X,'Options',opts);
```

Write an output function using the following syntax.

```
function stop = outfun(optimValues,state)

stop = false; % do not stop by default
switch state
    case 'init'
        % Set up plots or open files
    case 'iter'
        % Draw plots or update variables
    case 'done'
        % Clean up plots or files
end
```

`tsne` passes the `state` and `optimValues` variables to your function. `state` takes on the values `'init'`, `'iter'`, or `'done'` as shown in the code snippet.

tsne optimValues Structure

optimValues Field	Description
`'iteration'`	Iteration number
`'fval'`	Kullback-Leibler divergence, modified by exaggeration during the first 99 iterations
`'grad'`	Gradient of the Kullback-Leibler divergence, modified by exaggeration during the first 99 iterations
`'Exaggeration'`	Value of the exaggeration parameter in use in the current iteration
`'Y'`	Current embedding

t-SNE Custom Output Function

This example shows how to use an output function in `tsne`.

Custom Output Function

The following code is an output function that performs these tasks:

- Keep a history of the Kullback-Leibler divergence and the norm of its gradient in a workspace variable.
- Plot the solution and the history as the iterations proceed.
- Display a `Stop` button on the plot to stop the iterations early without losing any information.

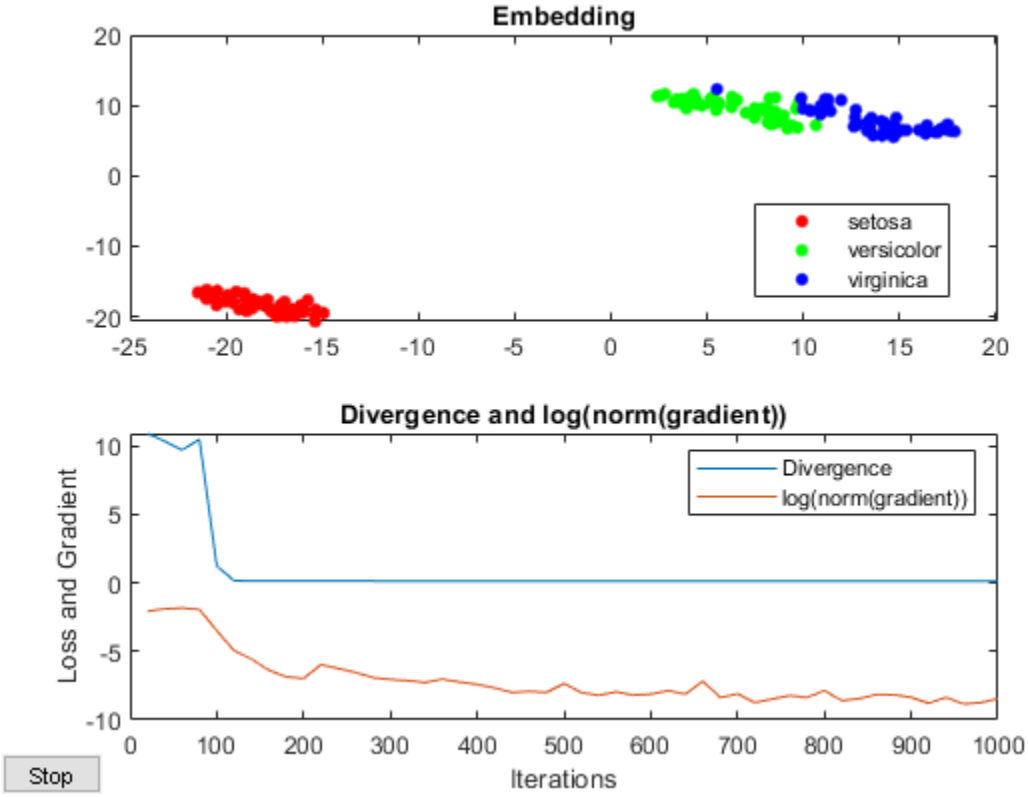
The output function has an extra input variable, `species`, that enables its plots to show the correct classification of the data. For information on including extra parameters such as `species` in a function, see “Parameterizing Functions”.

```
function stop = KLLogging(optimValues,state,species)
persistent h kllog iters stopnow
switch state
    case 'init'
        stopnow = false;
        kllog = [];
        iters = [];
        h = figure;
        c = uicontrol('Style','pushbutton','String','Stop','Position', ...
            [10 10 50 20],'Callback',@stopme);
    case 'iter'
        kllog = [kllog; optimValues.fval,log(norm(optimValues.grad))];
        assignin('base','history',kllog)
        iters = [iters; optimValues.iteration];
        if length(iters) > 1
            figure(h)
            subplot(2,1,2)
            plot(iters,kllog);
            xlabel('Iterations')
            ylabel('Loss and Gradient')
            legend('Divergence','log(norm(gradient))')
            title('Divergence and log(norm(gradient))')
            subplot(2,1,1)
            gscatter(optimValues.Y(:,1),optimValues.Y(:,2),species)
            title('Embedding')
            drawnow
        end
    case 'done'
        % Nothing here
end
stop = stopnow;

function stopme(~,~)
stopnow = true;
end
end
```

Use the Custom Output Function

Plot the Fisher iris data, a 4-D data set, in two dimensions using `tsne`. There is a drop in the Divergence value at iteration 100 because the divergence is scaled by the exaggeration value for earlier iterations. The embedding remains largely unchanged for the last several hundred iterations, so you can save time by clicking the `Stop` button during the iterations.

```
load fisheriris
rng default % for reproducibility
opts = statset('OutputFcn',@(optimValues,state) KLLogging(optimValues,state,species));
Y = tsne(meas,'Options',opts,'Algorithm','exact');
```



See Also

Related Examples

- “Visualize High-Dimensional Data Using t-SNE” on page 15-113
- “tsne Settings” on page 15-117

Visualize High-Dimensional Data Using t-SNE

This example shows how to visualize the MNIST data [1], which consists of images of handwritten digits, using the `tsne` function. The images are 28-by-28 pixels in grayscale. Each image has an associated label from 0 through 9, which is the digit that the image represents. `tsne` reduces the dimension of the data from 784 original dimensions to 50 using PCA, and then to two or three using the t-SNE Barnes-Hut algorithm.

Obtain Data

Begin by obtaining image and label data from

http://yann.lecun.com/exdb/mnist/

Unzip the files. For this example, use the `t10k-images` data.

```
imageFileName = 't10k-images.idx3-ubyte';
labelFileName = 't10k-labels.idx1-ubyte';
```

Process the files to load them in the workspace. The code for this processing function appears at the end of this example.

```
[X,L] = processMNISTdata(imageFileName,labelFileName);
```

```
Read MNIST image data...
Number of images in the dataset:  10000 ...
Each image is of 28 by 28 pixels...
The image data is read to a matrix of dimensions:  10000 by  784...
End of reading image data.

Read MNIST label data...
Number of labels in the dataset:  10000 ...
The label data is read to a matrix of dimensions:  10000 by  1...
End of reading label data.
```

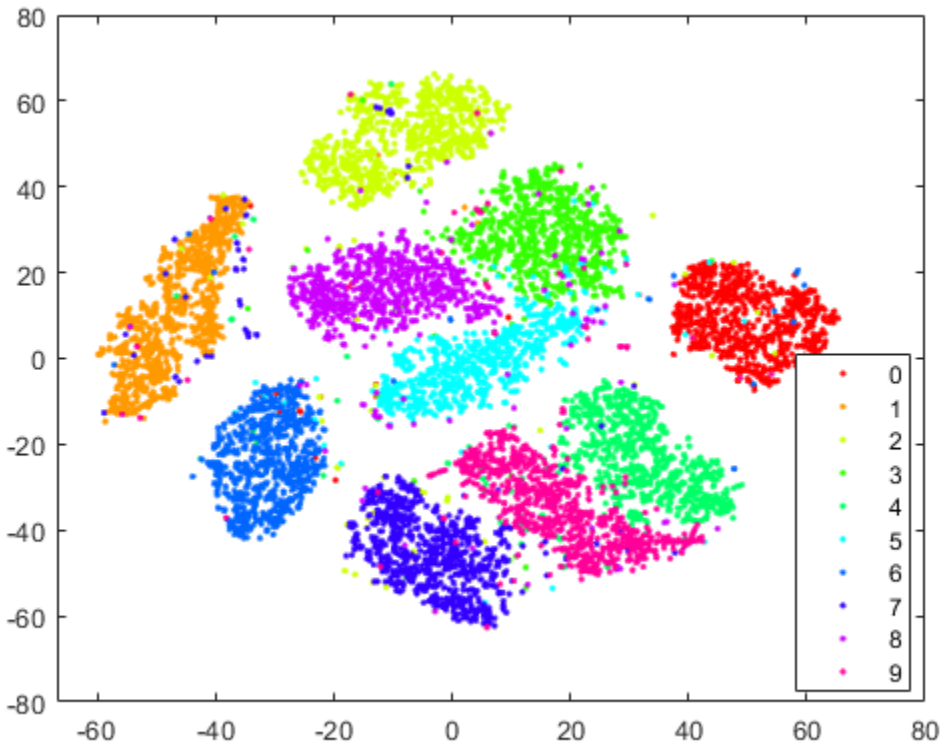
Reduce Dimension of Data to Two

Obtain two-dimensional analogues of the data clusters using t-SNE. Use PCA to reduce the initial dimensionality to 50. Use the Barnes-Hut variant of the t-SNE algorithm to save time on this relatively large data set.

```
rng default % for reproducibility
Y = tsne(X,'Algorithm','barneshut','NumPCAComponents',50);
```

Display the result, colored with the correct labels.

```
figure
gscatter(Y(:,1),Y(:,2),L)
```

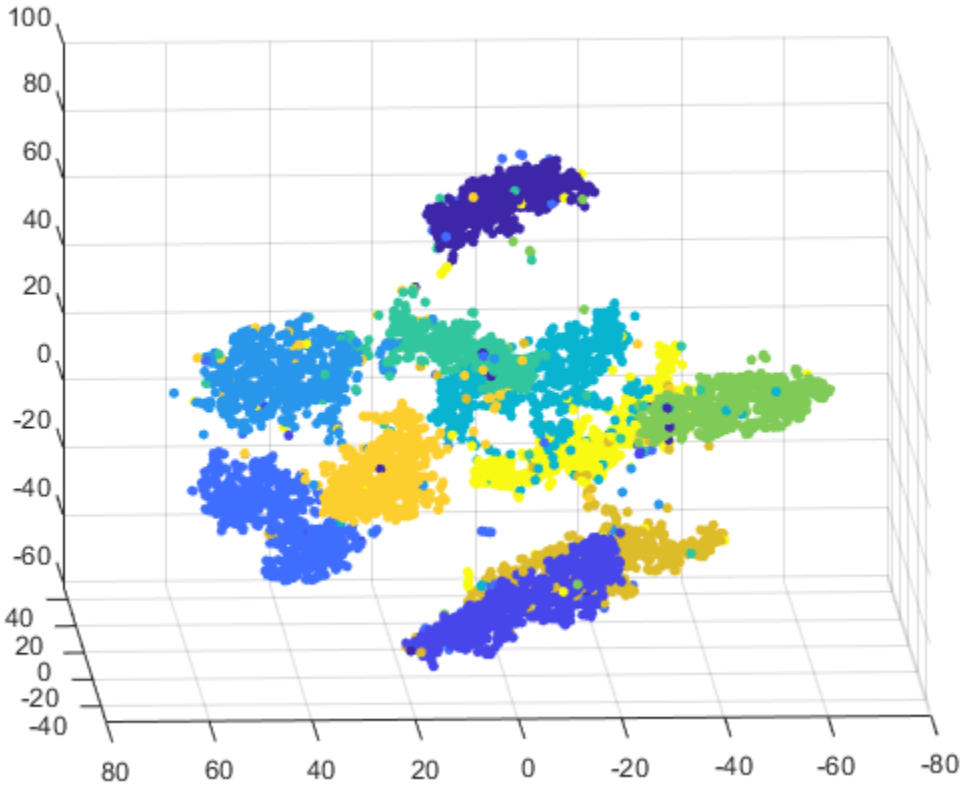


t-SNE creates clusters of points based solely on their relative similarities that correspond closely to the true labels.

Reduce Dimension of Data to Three

t-SNE can also reduce the data to three dimensions. Set the `tsne` `'NumDimensions'` name-value pair to 3.

```
rng default % for fair comparison
Y3 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'NumDimensions',3);
figure
scatter3(Y3(:,1),Y3(:,2),Y3(:,3),15,L,'filled');
view(-93,14)
```



Here is the code of the function that reads the data into the workspace.

```
function [X,L] = processMNISTdata(imageFileName,labelFileName)

[fileID,errmsg] = fopen(imageFileName,'r','b');
if fileID < 0
    error(errmsg);
end
%%
% First read the magic number. This number is 2051 for image data, and
% 2049 for label data
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2051
    fprintf('\nRead MNIST image data...\n')
end
%%
% Then read the number of images, number of rows, and number of columns
numImages = fread(fileID,1,'int32',0,'b');
fprintf('Number of images in the dataset: %6d ...\n',numImages);
numRows = fread(fileID,1,'int32',0,'b');
numCols = fread(fileID,1,'int32',0,'b');
fprintf('Each image is of %2d by %2d pixels...\n',numRows,numCols);
%%
% Read the image data
X = fread(fileID,inf,'unsigned char');
%%
```

```
% Reshape the data to array X
X = reshape(X,numCols,numRows,numImages);
X = permute(X,[2 1 3]);
%%
% Then flatten each image data into a 1 by (numRows*numCols) vector, and
% store all the image data into a numImages by (numRows*numCols) array.
X = reshape(X,numRows*numCols,numImages)';
fprintf(['The image data is read to a matrix of dimensions: %6d by %4d...\n',...
    'End of reading image data.\n'],size(X,1),size(X,2));
%%
% Close the file
fclose(fileID);
%%
% Similarly, read the label data.
[fileID,errmsg] = fopen(labelFileName,'r','b');
if fileID < 0
    error(errmsg);
end
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2049
    fprintf('\nRead MNIST label data...\n')
end
numItems = fread(fileID,1,'int32',0,'b');
fprintf('Number of labels in the dataset: %6d ...\n',numItems);

L = fread(fileID,inf,'unsigned char');
fprintf(['The label data is read to a matrix of dimensions: %6d by %2d...\n',...
    'End of reading label data.\n'],size(L,1),size(L,2));
fclose(fileID);
```

References

[1] Yann LeCun (Courant Institute, NYU) and Corinna Cortes (Google Labs, New York) hold the copyright of MNIST dataset, which is a derivative work from original NIST datasets. MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license, https://creativecommons.org/licenses/by-sa/3.0/

See Also

Related Examples

- “tsne Settings” on page 15-117

More About

- “t-SNE” on page 15-104

tsne Settings

This example shows the effects of various `tsne` settings.

Obtain Data

Begin by obtaining the MNIST [1] image and label data from

http://yann.lecun.com/exdb/mnist/

Unzip the files. For this example, use the `t10k-images` data.

```
imageFileName = 't10k-images.idx3-ubyte';
labelFileName = 't10k-labels.idx1-ubyte';
```

Process the files to load them in the workspace. The code for this processing function appears at the end of this example.

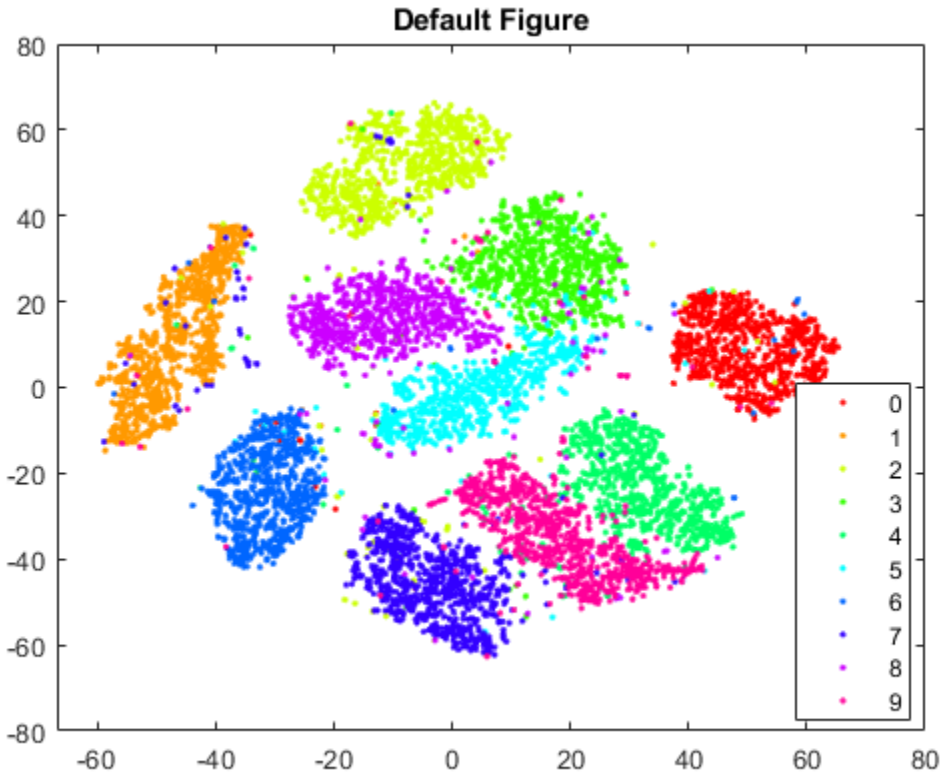
```
[X,L] = processMNISTdata(imageFileName,labelFileName);
```

```
Read MNIST image data...
Number of images in the dataset:  10000 ...
Each image is of 28 by 28 pixels...
The image data is read to a matrix of dimensions:  10000 by  784...
End of reading image data.

Read MNIST label data...
Number of labels in the dataset:  10000 ...
The label data is read to a matrix of dimensions:  10000 by  1...
End of reading label data.
```

Process Data Using t-SNE

Obtain two-dimensional analogs of the data clusters using t-SNE. Use the Barnes-Hut algorithm for better performance on this large data set. Use PCA to reduce the initial dimensions from 784 to 50.

```
rng default % for reproducibility
Y = tsne(X,'Algorithm','barneshut','NumPCAComponents',50);
figure
gscatter(Y(:,1),Y(:,2),L)
title('Default Figure')
```

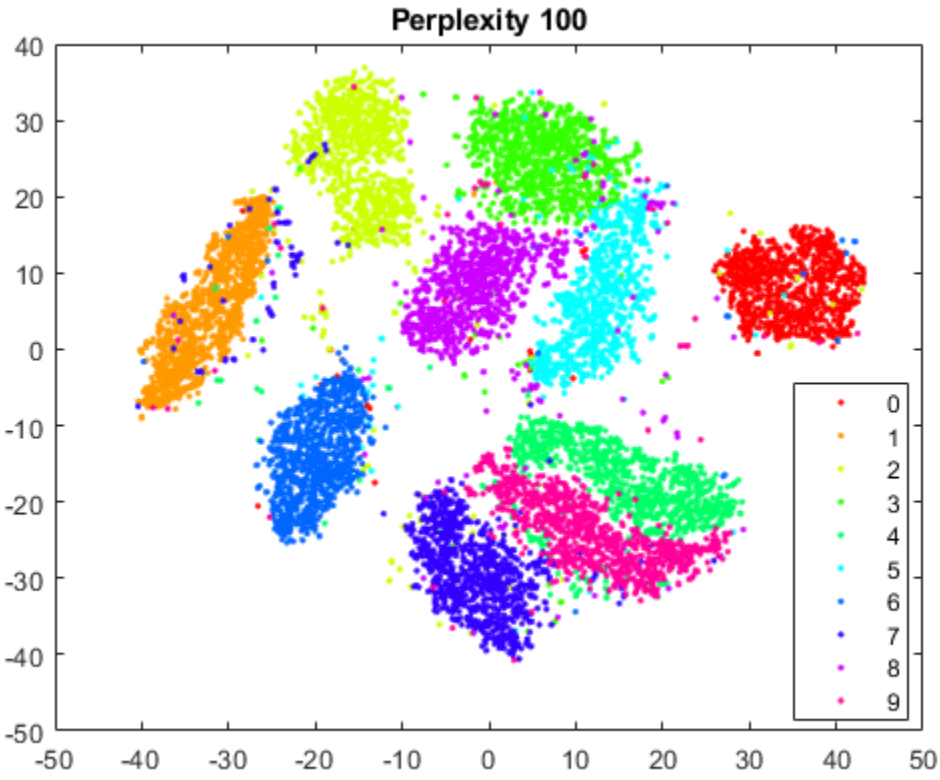


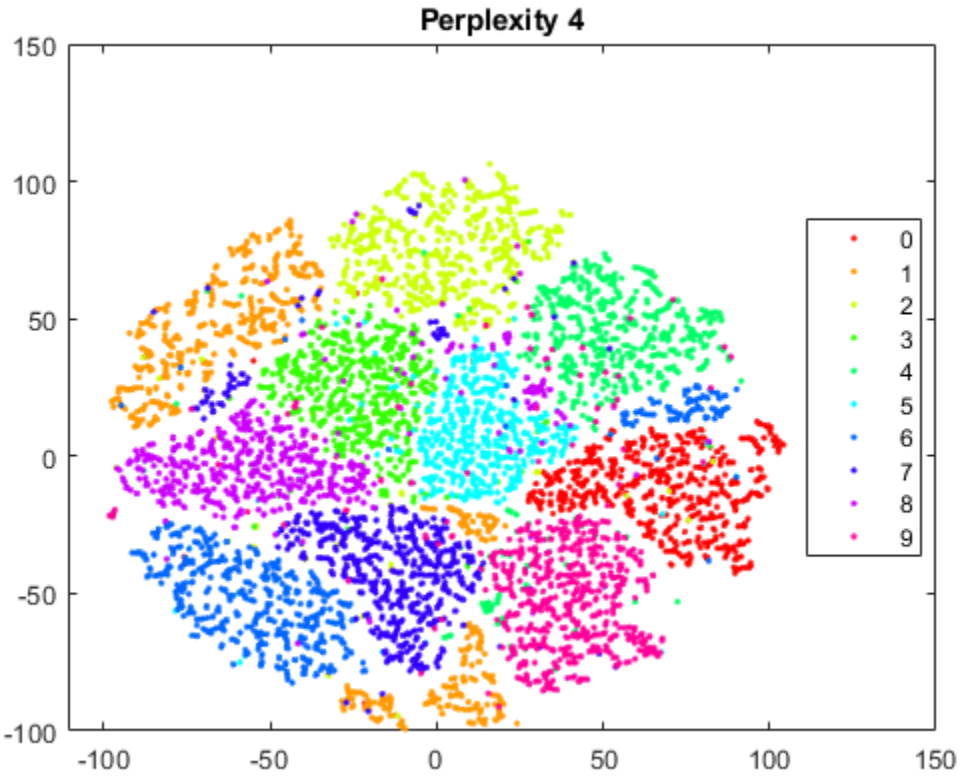
t-SNE creates a figure with well separated clusters and relatively few data points that seem misplaced.

Perplexity

Try altering the perplexity setting to see the effect on the figure.

```
rng default % for fair comparison
Y100 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Perplexity',100);
figure
gscatter(Y100(:,1),Y100(:,2),L)
title('Perplexity 100')

rng default % for fair comparison
Y4 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Perplexity',4);
figure
gscatter(Y4(:,1),Y4(:,2),L)
title('Perplexity 4')
```


Perplexity 100
40
30
20
10
0
-10
-20
-30
-40
-50
-50
-40
-30
-20
-10
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8
9

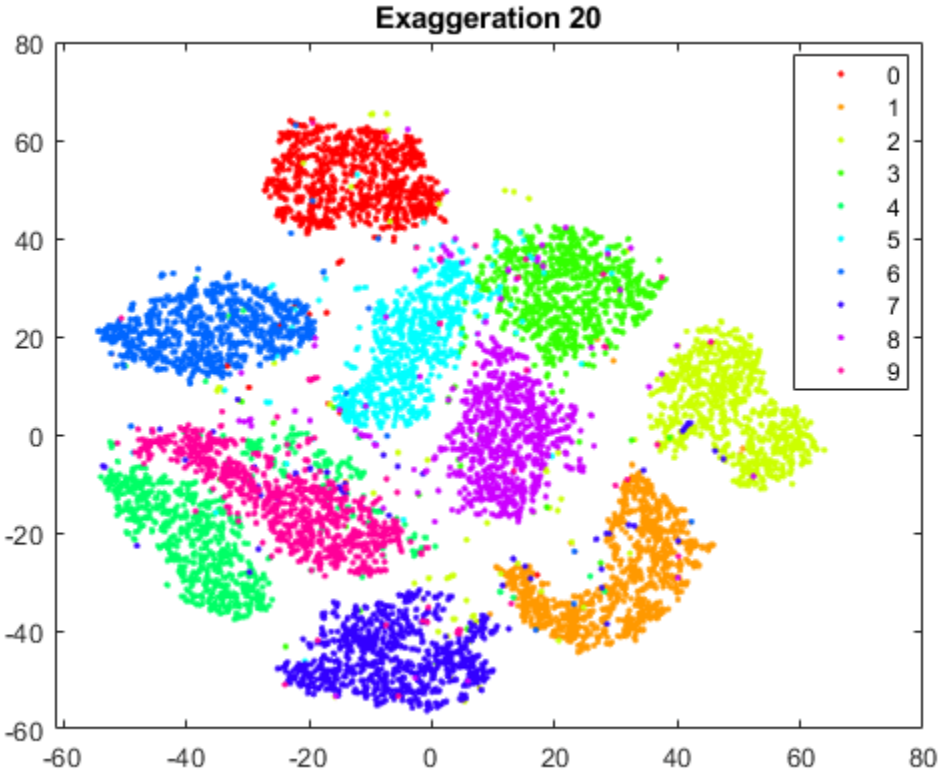


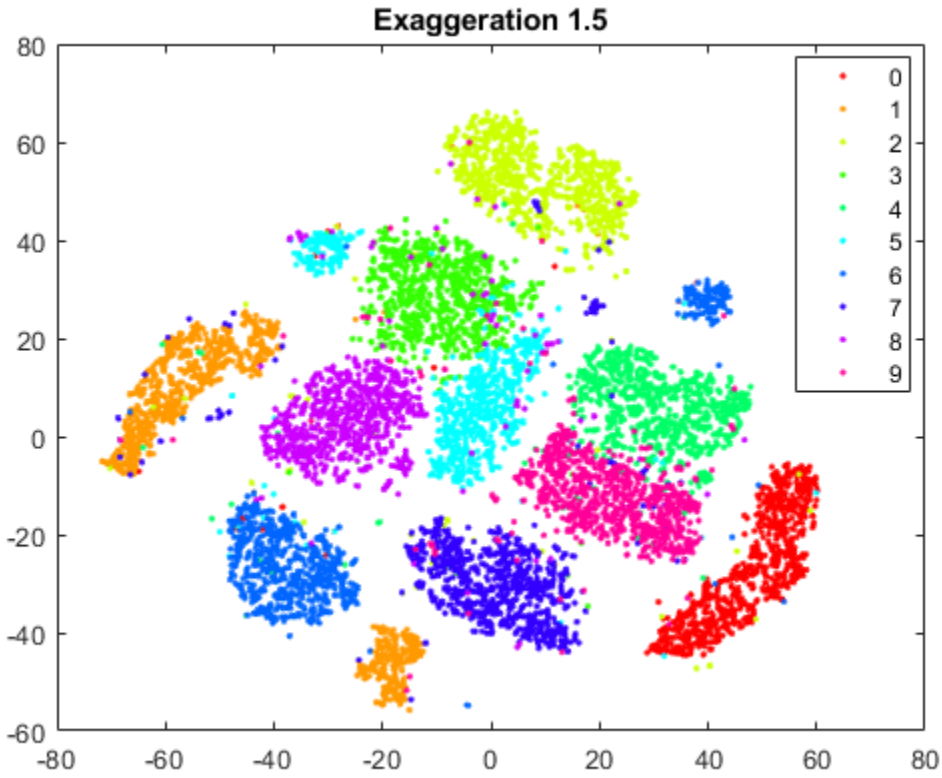
Setting the perplexity to 100 yields a figure that is largely similar to the default figure. The clusters are tighter than with the default setting. However, setting the perplexity to 4 gives a figure without well separated clusters. The clusters are looser than with the default setting.

Exaggeration

Try altering the exaggeration setting to see the effect on the figure.

```
rng default % for fair comparison
YEX0 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Exaggeration',20);
figure
gscatter(YEX0(:,1),YEX0(:,2),L)
title('Exaggeration 20')

rng default % for fair comparison
YEx15 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Exaggeration',1.5);
figure
gscatter(YEx15(:,1),YEx15(:,2),L)
title('Exaggeration 1.5')
```


Exaggeration 20
80
60
40
20
0
-20
-40
-60
-60
-40
-20
0
20
40
60
80
0
1
2
3
4
5
6
7
8
9

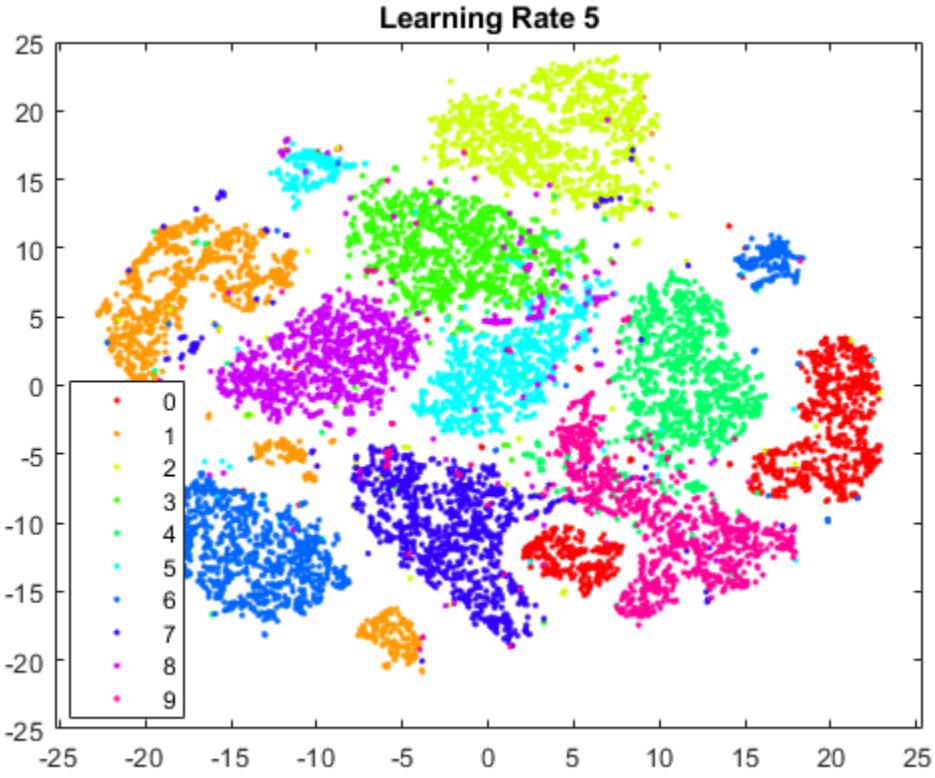


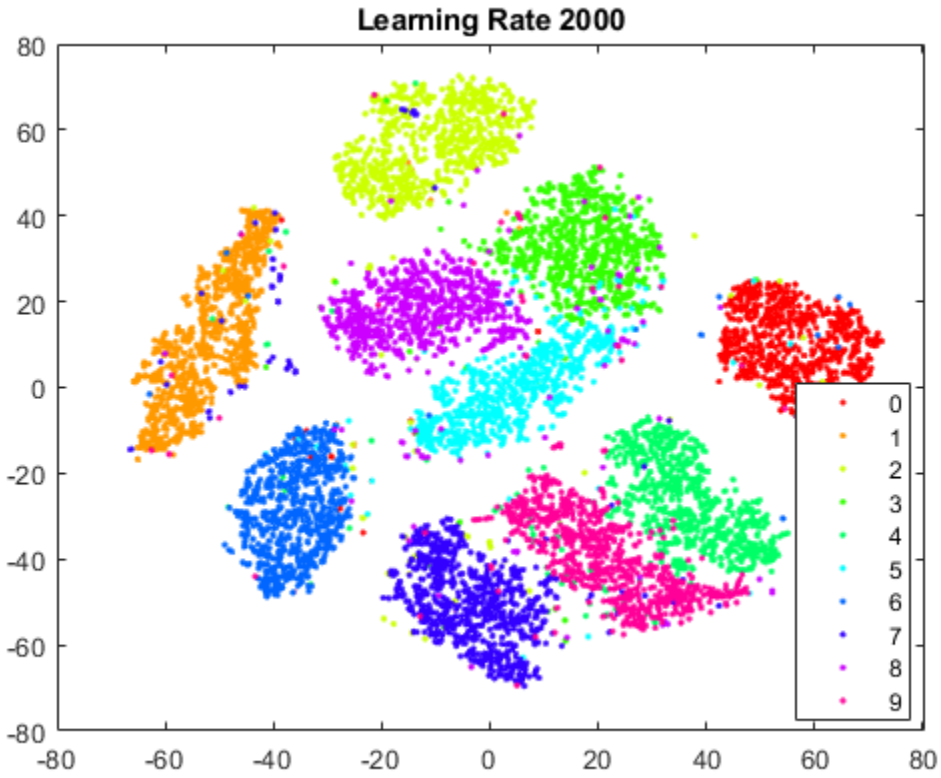
While the exaggeration setting has an effect on the figure, it is not clear whether any nondefault setting gives a better picture than the default setting. The figure with an exaggeration of 20 is similar to the default figure. In general, a larger exaggeration creates more empty space between embedded clusters. An exaggeration of 1.5 causes the groups labeled 1 and 6 to split into two groups each, an undesirable outcome. Exaggerating the values in the joint distribution of X makes the values in the joint distribution of Y smaller. This makes it much easier for the embedded points to move relative to one another. The splitting of clusters 1 and 6 reflects this effect.

Learning Rate

Try altering the learning rate setting to see the effect on the figure.

```
rng default % for fair comparison
YL5 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'LearnRate',5);
figure
gscatter(YL5(:,1),YL5(:,2),L)
title('Learning Rate 5')

rng default % for fair comparison
YL2000 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'LearnRate',2000);
figure
gscatter(YL2000(:,1),YL2000(:,2),L)
title('Learning Rate 2000')
```


Learning Rate 5
25
20
15
10
5
0
-5
-10
-15
-20
-25
-25
-20
-15
-10
-5
0
5
10
15
20
25
0
1
2
3
4
5
6
7
8
9



The figure with a learning rate of 5 has several clusters that split into two or more pieces. This shows that if the learning rate is too small, the minimization process can get stuck in a bad local minimum. A learning rate of 2000 gives a figure similar to the default figure.

Initial Behavior with Various Settings

Large learning rates or large exaggeration values can lead to undesirable initial behavior. To see this, set large values of these parameters and set `NumPrint` and `Verbose` to 1 to show all the iterations. Stop the iterations after 10, as the goal of this experiment is simply to look at the initial behavior.

Begin by setting the exaggeration to 200.

```
rng default % for fair comparison
opts = statset('MaxIter',10);
YEX200 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Exaggeration',200,...
    'NumPrint',1,'Verbose',1,'Options',opts);
```

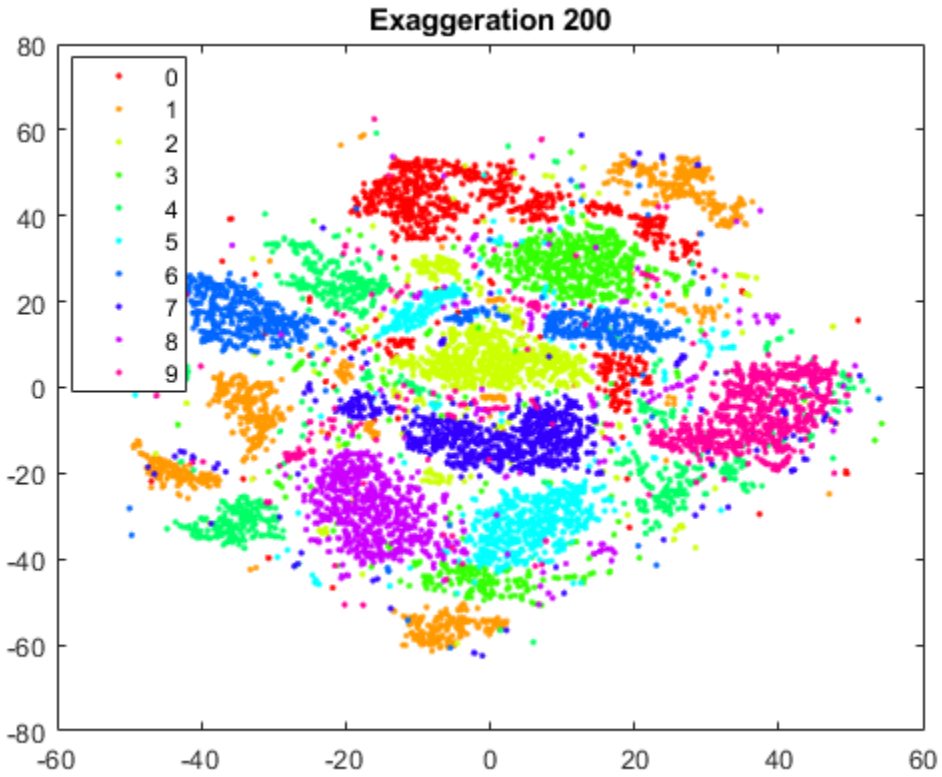
```
|============================================|
|   ITER   | KL DIVERGENCE   | NORM GRAD USING |
|          | FUN VALUE USING | EXAGGERATED DIST|
|          | EXAGGERATED DIST| OF X            |
|          | OF X            |                 |
|============================================|
|        1 |    2.190347e+03 |    6.078667e-05 |
|        2 |    2.190352e+03 |    4.769050e-03 |
|        3 |    2.204061e+03 |    9.423678e-02 |
```

```
|         4 |    2.464585e+03 |    2.113271e-02 |
|         5 |    2.501222e+03 |    2.616407e-02 |
|         6 |    2.529362e+03 |    3.022570e-02 |
|         7 |    2.553233e+03 |    3.108418e-02 |
|         8 |    2.562822e+03 |    3.278873e-02 |
|         9 |    2.538056e+03 |    3.222265e-02 |
|        10 |    2.504932e+03 |    3.671708e-02 |
```

The Kullback-Leibler divergence increases during the first few iterations, and the norm of the gradient increases as well.

To see the final result of the embedding, allow the algorithm to run to completion using the default stopping criteria.

```
rng default % for fair comparison
YEX200 = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'Exaggeration',200);
figure
gscatter(YEX200(:,1),YEX200(:,2),L)
title('Exaggeration 200')
```



This exaggeration value does not give a clean separation into clusters.

Show the initial behavior when the learning rate is 100,000.

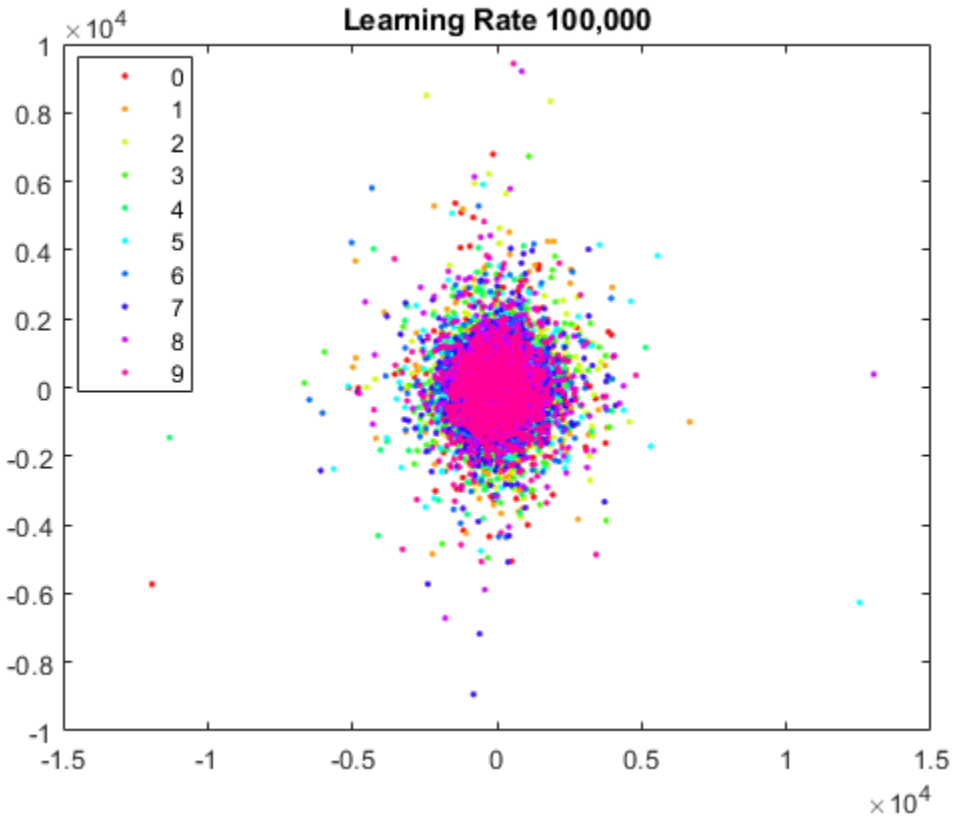
```
rng default % for fair comparison
YL100k = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'LearnRate',1e5,...
    'NumPrint',1,'Verbose',1,'Options',opts);
```

```
|=============================================|
|   ITER   | KL DIVERGENCE   | NORM GRAD USING |
|          | FUN VALUE USING | EXAGGERATED DIST|
|          | EXAGGERATED DIST| OF X            |
|          | OF X            |                 |
|=============================================|
|        1 |    2.815885e+01 |    1.024049e-06 |
|        2 |    2.816002e+01 |    2.902059e-04 |
|        3 |    3.195873e+01 |    7.355889e-04 |
|        4 |    3.348151e+01 |    3.958901e-04 |
|        5 |    3.365935e+01 |    2.876905e-04 |
|        6 |    3.342462e+01 |    3.906245e-04 |
|        7 |    3.303205e+01 |    4.037983e-04 |
|        8 |    3.263320e+01 |    5.665630e-04 |
|        9 |    3.235384e+01 |    4.319099e-04 |
|       10 |    3.211238e+01 |    4.803526e-04 |
```

Again, the Kullback-Leibler divergence increases during the first few iterations, and the norm of the gradient increases as well.

To see the final result of the embedding, allow the algorithm to run to completion using the default stopping criteria.

```
rng default % for fair comparison
YL100k = tsne(X,'Algorithm','barneshut','NumPCAComponents',50,'LearnRate',1e5);
figure
gscatter(YL100k(:,1),YL100k(:,2),L)
title('Learning Rate 100,000')
```



The learning rate is far too large, and gives no useful embedding.

Conclusion

`tsne` with default settings does a good job of embedding the high-dimensional initial data into two-dimensional points that have well defined clusters. The effects of algorithm settings are difficult to predict. Sometimes they can improve the clustering, but for the most part the default settings seem good. While speed is not part of this investigation, settings can affect the speed of the algorithm. In particular, the Barnes-Hut algorithm is notably faster on this data.

Code to Process MNIST Data

Here is the code of the function that reads the data into the workspace.

```
function [X,L] = processMNISTdata(imageFileName,labelFileName)

[fileID,errmsg] = fopen(imageFileName,'r','b');
if fileID < 0
    error(errmsg);
end
%%
% First read the magic number. This number is 2051 for image data, and
% 2049 for label data
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2051
    fprintf('\nRead MNIST image data...\n')
```

```
end
%%
% Then read the number of images, number of rows, and number of columns
numImages = fread(fileID,1,'int32',0,'b');
fprintf('Number of images in the dataset: %6d ...\n',numImages);
numRows = fread(fileID,1,'int32',0,'b');
numCols = fread(fileID,1,'int32',0,'b');
fprintf('Each image is of %2d by %2d pixels...\n',numRows,numCols);
%%
% Read the image data
X = fread(fileID,inf,'unsigned char');
%%
% Reshape the data to array X
X = reshape(X,numCols,numRows,numImages);
X = permute(X,[2 1 3]);
%%
% Then flatten each image data into a 1 by (numRows*numCols) vector, and
% store all the image data into a numImages by (numRows*numCols) array.
X = reshape(X,numRows*numCols,numImages)';
fprintf(['The image data is read to a matrix of dimensions: %6d by %4d...\n',...
    'End of reading image data.\n'],size(X,1),size(X,2));
%%
% Close the file
fclose(fileID);
%%
% Similarly, read the label data.
[fileID,errmsg] = fopen(labelFileName,'r','b');
if fileID < 0
    error(errmsg);
end
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2049
    fprintf('\nRead MNIST label data...\n')
end
numItems = fread(fileID,1,'int32',0,'b');
fprintf('Number of labels in the dataset: %6d ...\n',numItems);

L = fread(fileID,inf,'unsigned char');
fprintf(['The label data is read to a matrix of dimensions: %6d by %2d...\n',...
    'End of reading label data.\n'],size(L,1),size(L,2));
fclose(fileID);
```

References

[1] Yann LeCun (Courant Institute, NYU) and Corinna Cortes (Google Labs, New York) hold the copyright of MNIST dataset, which is a derivative work from original NIST datasets. MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license, https://creativecommons.org/licenses/by-sa/3.0/

See Also

Related Examples

- “Visualize High-Dimensional Data Using t-SNE” on page 15-113

- “t-SNE Output Function” on page 15-110

More About

- “t-SNE” on page 15-104

External Websites

- How to Use t-SNE Effectively

Feature Extraction

In this section...

What Is Feature Extraction?

Feature extraction is a set of methods that map input features to new output features. Many feature extraction methods use unsupervised learning to extract features. Unlike some feature extraction methods such as PCA and NNMF, the methods described in this section can increase dimensionality (and decrease dimensionality). Internally, the methods involve optimizing nonlinear objective functions. For details, see “Sparse Filtering Algorithm” on page 15-130 or “Reconstruction ICA Algorithm” on page 15-132.

One typical use of feature extraction is finding features in images. Using these features can lead to improved classification accuracy. For an example, see “Feature Extraction Workflow” on page 15-135. Another typical use is extracting individual signals from superpositions, which is often termed blind source separation. For an example, see “Extract Mixed Signals” on page 15-164.

There are two feature extraction functions: `rica` and `sparsefilt`. Associated with these functions are the objects that they create: `ReconstructionICA` and `SparseFiltering`.

Sparse Filtering Algorithm

The sparse filtering algorithm begins with a data matrix `X` that has `n` rows and `p` columns. Each row represents one observation and each column represents one measurement. The columns are also called the features or predictors. The algorithm then takes either an initial random `p`-by-`q` weight matrix `W` or uses the weight matrix passed in the `InitialTransformWeights` name-value pair. `q` is the requested number of features that `sparsefilt` computes.

The algorithm attempts to minimize the “Sparse Filtering Objective Function” on page 15-131 by using a standard limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton optimizer. See Nocedal and Wright [2]. This optimizer takes up to `IterationLimit` iterations. It stops iterating earlier when it takes a step whose norm is less than `StepTolerance`, or when it computes that the norm of the gradient at the current point is less than `GradientTolerance` times a scalar τ, where

$$\tau = \max\left(1, \min\left(|f|, \|g_0\|_\infty\right)\right).$$

$|f|$ is the norm of the objective function, and $\|g_0\|_\infty$ is the infinity norm of the initial gradient.

The objective function attempts to simultaneously obtain few nonzero features for each data point, and for each resulting feature to have nearly equal weight. To understand how the objective function attempts to achieve these goals, see Ngiam, Koh, Chen, Bhaskar, and Ng [1].

Frequently, you obtain good features by setting a relatively small value of `IterationLimit`, from as low as 5 to a few hundred. Allowing the optimizer to continue can result in overtraining, where the extracted features do not generalize well to new data.

After constructing a `SparseFiltering` object, use the `transform` method to map input data to the new output features.

Sparse Filtering Objective Function

To compute an objective function, the sparse filtering algorithm uses the following steps. The objective function depends on the `n`-by-`p` data matrix `X` and a weight matrix `W` that the optimizer varies. The weight matrix `W` has dimensions `p`-by-`q`, where `p` is the number of original features and `q` is the number of requested features.

1. Compute the `n`-by-`q` matrix `X*W`. Apply the approximate absolute value function $\phi(u) = \sqrt{u^2 + 10^{-8}}$ to each element of `X*W` to obtain the matrix `F`. ϕ is a smooth nonnegative symmetric function that closely approximates the absolute value function.
2. Normalize the columns of `F` by the approximate L^2 norm. In other words, define the normalized matrix $\tilde{F}(i, j)$ by

$$\|F(j)\| = \sqrt{\sum_{i=1}^{n} (F(i, j))^2 + 10^{-8}}$$

$$\tilde{F}(i, j) = F(i, j) / \|F(j)\| .$$

3. Normalize the rows of $\tilde{F}(i, j)$ by the approximate L^2 norm. In other words, define the normalized matrix $\widehat{F}(i, j)$ by

$$\left\|\tilde{F}(i)\right\| = \sqrt{\sum_{j=1}^{q} \left(\tilde{F}(i, j)\right)^2 + 10^{-8}}$$

$$\widehat{F}(i, j) = \tilde{F}(i, j) / \left\|\tilde{F}(i)\right\| .$$

   The matrix $\widehat{F}$ is the matrix of converted features in `X`. Once `sparsefilt` finds the weights `W` that minimize the objective function h (see below), which the function stores in the output object `Mdl` in the `Mdl.TransformWeights` property, the `transform` function can follow the same transformation steps to convert new data to output features.

4. Compute the objective function h(`W`) as the 1–norm of the matrix $\widehat{F}(i, j)$, meaning the sum of all the elements in the matrix (which are nonnegative by construction):

$$h(W) = \sum_{j=1}^{q} \sum_{i=1}^{n} \widehat{F}(i, j) .$$

5. If you set the `Lambda` name-value pair to a strictly positive value, `sparsefilt` uses the following modified objective function:

$$h(W) = \sum_{j=1}^{q} \sum_{i=1}^{n} \widehat{F}(i, j) + \lambda \sum_{j=1}^{q} w_j^T w_j .$$

   Here, w_j is the jth column of the matrix `W` and λ is the value of `Lambda`. The effect of this term is to shrink the weights `W`. If you plot the columns of `W` as images, with positive `Lambda` these images appear smooth compared to the same images with zero `Lambda`.

Reconstruction ICA Algorithm

The Reconstruction Independent Component Analysis (RICA) algorithm is based on minimizing an objective function. The algorithm maps input data to output features.

The ICA source model is the following. Each observation x is generated by a random vector s according to

$$x = \mu + As\,.$$

- x is a column vector of length `p`.
- μ is a column vector of length `p` representing a constant term.
- s is a column vector of length `q` whose elements are zero mean, unit variance random variables that are statistically independent of each other.
- A is a mixing matrix of size `p`-by-`q`.

You can use this model in `rica` to estimate A from observations of x. See “Extract Mixed Signals” on page 15-164.

The RICA algorithm begins with a data matrix `X` that has `n` rows and `p` columns consisting of the observations x_i:

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}.$$

Each row represents one observation and each column represents one measurement. The columns are also called the features or predictors. The algorithm then takes either an initial random `p`-by-`q` weight matrix `W` or uses the weight matrix passed in the `InitialTransformWeights` name-value pair. `q` is the requested number of features that `rica` computes. The weight matrix `W` is composed of columns w_i of size `p`-by-1:

$$W = \begin{bmatrix} w_1 & w_2 & \dots & w_q \end{bmatrix}.$$

The algorithm attempts to minimize the “Reconstruction ICA Objective Function” on page 15-133 by using a standard limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton optimizer. See Nocedal and Wright [2]. This optimizer takes up to `IterationLimit` iterations. It stops iterating when it takes a step whose norm is less than `StepTolerance`, or when it computes that the norm of the gradient at the current point is less than `GradientTolerance` times a scalar τ, where

$$\tau = \max\left(1, \min\left(|f|, \|g_0\|_\infty\right)\right).$$

$|f|$ is the norm of the objective function, and $\|g_0\|_\infty$ is the infinity norm of the initial gradient.

The objective function attempts to obtain a nearly orthonormal weight matrix that minimizes the sum of elements of $g(\mathtt{XW})$, where g is a function (described below) that is applied elementwise to `XW`. To understand how the objective function attempts to achieve these goals, see Le, Karpenko, Ngiam, and Ng [3].

After constructing a `ReconstructionICA` object, use the `transform` method to map input data to the new output features.

Reconstruction ICA Objective Function

The objective function uses a contrast function, which you specify by using the `ContrastFcn` name-value pair. The contrast function is a smooth convex function that is similar to an absolute value. By default, the contrast function is $g = \frac{1}{2}\log(\cosh(2x))$. For other available contrast functions, see `ContrastFcn`.

For an `n`-by-`p` data matrix `X` and `q` output features, with a regularization parameter λ as the value of the `Lambda` name-value pair, the objective function in terms of the `p`-by-`q` matrix `W` is

$$h = \frac{\lambda}{n}\sum_{i=1}^{n}\left\|WW^{T}x_i - x_i\right\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q}\sigma_j g\left(w_j^{T}x_i\right)$$

The σ_j are known constants that are ±1. When $\sigma_j = +1$, minimizing the objective function h encourages the histogram of $w_j^{T}x_i$ to be sharply peaked at 0 (super Gaussian). When $\sigma_j = -1$, minimizing the objective function h encourages the histogram of $w_j^{T}x_i$ to be flatter near 0 (sub Gaussian). Specify the σ_j values using the `rica` `NonGaussianityIndicator` name-value pair.

The objective function h can have a spurious minimum of zero when λ is zero. Therefore, `rica` minimizes h over W that are normalized to 1. In other words, each column w_j of W is defined in terms of a column vector v_j by

$$w_j = \frac{v_j}{\sqrt{v_j^{T}v_j + 10^{-8}}}.$$

`rica` minimizes over the v_j. The resulting minimal matrix `W` provides the transformation from input data `X` to output features `XW`.

References

[1] Ngiam, Jiquan, Zhenghao Chen, Sonia A. Bhaskar, Pang W. Koh, and Andrew Y. Ng. “Sparse Filtering.” *Advances in Neural Information Processing Systems.* Vol. 24, 2011, pp. 1125–1133. `https://papers.nips.cc/paper/4334-sparse-filtering.pdf`.

[2] Nocedal, J. and S. J. Wright. *Numerical Optimization*, Second Edition. Springer Series in Operations Research, Springer Verlag, 2006.

[3] Le, Quoc V., Alexandre Karpenko, Jiquan Ngiam, and Andrew Y. Ng. “ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning.” *Advances in Neural Information Processing Systems.* Vol. 24, 2011, pp. 1017–1025. `https://papers.nips.cc/paper/4467-ica-with-reconstruction-cost-for-efficient-overcomplete-feature-learning.pdf`.

See Also

`ReconstructionICA` | `SparseFiltering` | `rica` | `sparsefilt`

Related Examples

- “Feature Extraction Workflow” on page 15-135
- “Extract Mixed Signals” on page 15-164

Feature Extraction Workflow

This example shows a complete workflow for feature extraction from image data.

Obtain Data

This example uses the MNIST image data [1], which consists of images of handwritten digits. The images are 28-by-28 pixels in gray scale. Each image has an associated label from 0 through 9, which is the digit that the image represents.

Begin by obtaining image and label data from

http://yann.lecun.com/exdb/mnist/

Unzip the files. For better performance on this long example, use the test data as training data and the training data as test data.

```
imageFileName = 't10k-images.idx3-ubyte';
labelFileName = 't10k-labels.idx1-ubyte';
```

Process the files to load them in the workspace. The code for this processing function appears at the end of this example.

```
[Xtrain,LabelTrain] = processMNISTdata(imageFileName,labelFileName);
```

```
Read MNIST image data...
Number of images in the dataset:  10000 ...
Each image is of 28 by 28 pixels...
The image data is read to a matrix of dimensions:  10000 by  784...
End of reading image data.

Read MNIST label data...
Number of labels in the dataset:  10000 ...
The label data is read to a matrix of dimensions:  10000 by  1...
End of reading label data.
```

View a few of the images.

```
rng('default') % For reproducibility
numrows = size(Xtrain,1);
ims = randi(numrows,4,1);
imgs = Xtrain(ims,:);
for i = 1:4
    pp{i} = reshape(imgs(i,:),28,28);
end
ppf = [pp{1},pp{2};pp{3},pp{4}];
imshow(ppf);
```

Choose New Feature Dimensions

There are several considerations in choosing the number of features to extract:

- More features use more memory and computational time.
- Fewer features can produce a poor classifier.

For this example, choose 100 features.

```
q = 100;
```

Extract Features

There are two feature extraction functions, `sparsefilt` and `rica`. Begin with the `sparsefilt` function. Set the number of iterations to 10 so that the extraction does not take too long.

Typically, you get good results by running the `sparsefilt` algorithm for a few iterations to a few hundred iterations. Running the algorithm for too many iterations can lead to decreased classification accuracy, a type of overfitting problem.

Use `sparsefilt` to obtain the sparse filtering model while using 10 iterations.

```
Mdl = sparsefilt(Xtrain,q,'IterationLimit',10);
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

`sparsefilt` warns that the internal LBFGS optimizer did not converge. The optimizer did not converge because you set the iteration limit to 10. Nevertheless, you can use the result to train a classifier.

Create Classifier

Transform the original data into the new feature representation.

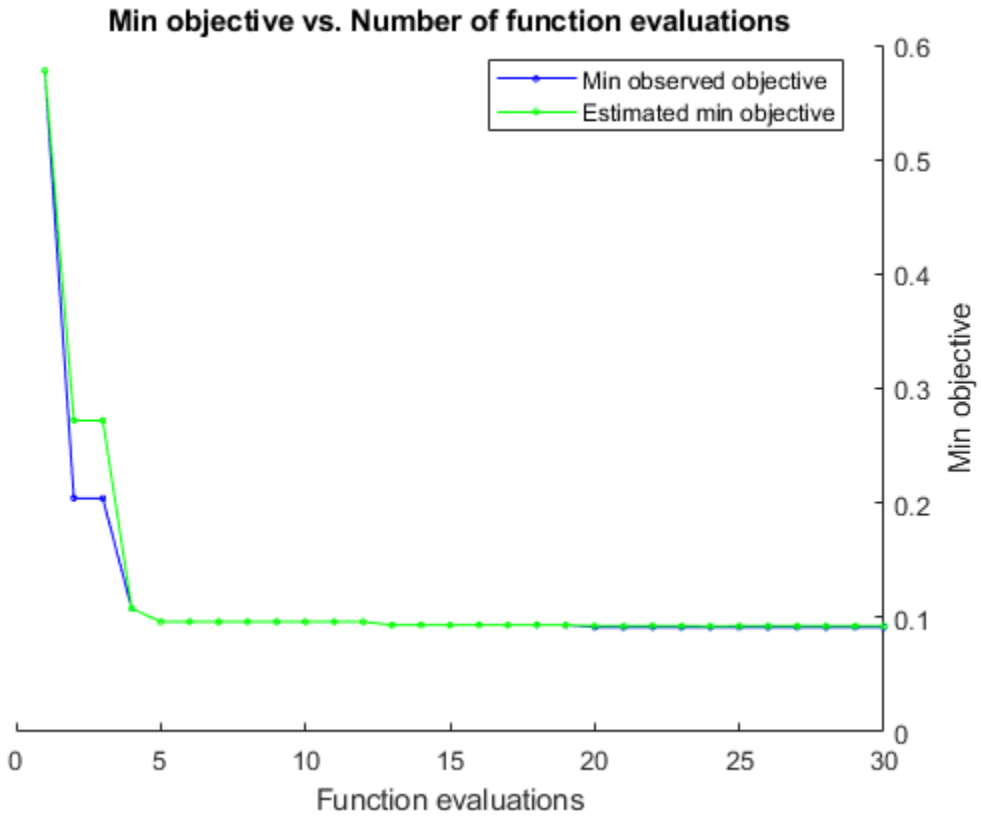
```
NewX = transform(Mdl,Xtrain);
```

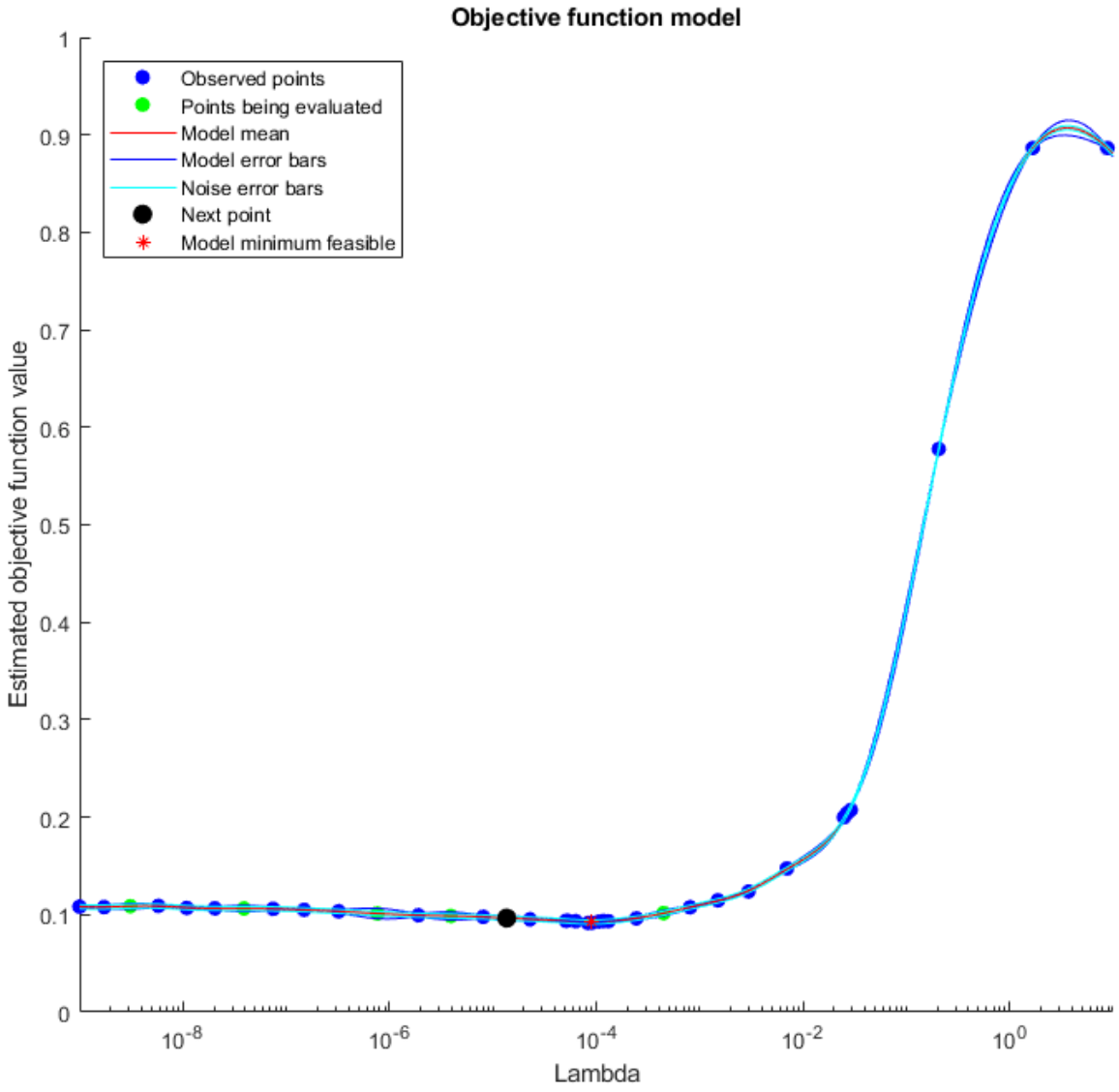
Train a linear classifier based on the transformed data and the correct classification labels in `LabelTrain`. The accuracy of the learned model is sensitive to the `fitcecoc` regularization parameter `Lambda`. Try to find the best value for `Lambda` by using the `OptimizeHyperparameters` name-value pair. Be aware that this optimization takes time. If you have a Parallel Computing Toolbox™ license, use parallel computing for faster execution. If you don't have a parallel license, remove the `UseParallel` calls before running this script.

```
t = templateLinear('Solver','lbfgs');
options = struct('UseParallel',true);
```

```
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);

Copying objective function to workers...
Done copying objective function to workers.
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|    1 |       6 | Best   |      0.5777 |      8.5334 |      0.5777 |      0.5777 |      0.20606 |
|    2 |       5 | Accept |      0.8865 |      8.9062 |      0.2041 |     0.27206 |       8.8234 |
|    3 |       5 | Best   |      0.2041 |      9.7024 |      0.2041 |     0.27206 |     0.026804 |
|    4 |       6 | Best   |      0.1077 |      14.629 |      0.1077 |     0.10773 |   1.7309e-09 |
|    5 |       6 | Best   |      0.0962 |      15.767 |      0.0962 |    0.096203 |    0.0002442 |
|    6 |       6 | Accept |      0.1999 |      6.4363 |      0.0962 |     0.09622 |     0.024862 |
|    7 |       6 | Accept |      0.2074 |      6.4171 |      0.0962 |    0.096222 |     0.029034 |
|    8 |       6 | Accept |      0.1065 |      12.974 |      0.0962 |    0.096222 |    2.037e-08 |
|    9 |       6 | Accept |      0.0977 |      22.976 |      0.0962 |    0.096216 |   8.0495e-06 |
|   10 |       6 | Accept |      0.1237 |      8.5033 |      0.0962 |    0.096199 |    0.0029745 |
|   11 |       6 | Accept |      0.1076 |      10.653 |      0.0962 |    0.096208 |   0.00080903 |
|   12 |       6 | Accept |      0.1034 |      16.761 |      0.0962 |      0.0962 |   3.2145e-07 |
|   13 |       6 | Best   |      0.0933 |      16.715 |      0.0933 |    0.093293 |   6.3327e-05 |
|   14 |       6 | Accept |       0.109 |      12.946 |      0.0933 |     0.09328 |   5.7887e-09 |
|   15 |       6 | Accept |      0.0994 |      18.805 |      0.0933 |    0.093312 |   1.8981e-06 |
|   16 |       6 | Accept |       0.106 |      15.088 |      0.0933 |    0.093306 |   7.4684e-08 |
|   17 |       6 | Accept |      0.0952 |      20.372 |      0.0933 |    0.093285 |   2.2831e-05 |
|   18 |       6 | Accept |      0.0933 |      14.528 |      0.0933 |    0.093459 |   0.00013097 |
|   19 |       6 | Accept |      0.1082 |      12.764 |      0.0933 |    0.093458 |   1.0001e-09 |
|   20 |       6 | Best   |      0.0915 |      16.157 |      0.0915 |    0.092391 |   8.3234e-05 |
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|   21 |       6 | Accept |      0.8865 |      6.6373 |      0.0915 |    0.092387 |       1.6749 |
|   22 |       6 | Accept |      0.0929 |      17.306 |      0.0915 |    0.092457 |   0.00010668 |
|   23 |       6 | Accept |      0.0937 |      19.046 |      0.0915 |    0.092535 |   5.0962e-05 |
|   24 |       6 | Accept |      0.0916 |      17.932 |      0.0915 |    0.092306 |    9.023e-05 |
|   25 |       6 | Accept |      0.0935 |       17.53 |      0.0915 |    0.092431 |   0.00011726 |
|   26 |       6 | Accept |      0.1474 |      8.3795 |      0.0915 |    0.092397 |     0.006997 |
|   27 |       6 | Accept |      0.0939 |      19.188 |      0.0915 |    0.092427 |   5.2557e-05 |
|   28 |       6 | Accept |      0.1147 |      10.686 |      0.0915 |    0.092432 |    0.0015036 |
|   29 |       6 | Accept |      0.1049 |      16.609 |      0.0915 |    0.092434 |   1.4871e-07 |
|   30 |       6 | Accept |      0.1069 |      13.929 |      0.0915 |    0.092435 |   1.0899e-08 |

__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 83.1976 seconds.
Total objective function evaluation time: 416.8767

Best observed feasible point:
      Lambda
    __________

    8.3234e-05
```

```
Observed objective function value = 0.0915
Estimated objective function value = 0.09245
Function evaluation time = 16.1569

Best estimated feasible point (according to models):
     Lambda
    _________

    9.023e-05

Estimated objective function value = 0.092435
Estimated function evaluation time = 17.0972
```





Evaluate Classifier

Check the error of the classifier when applied to test data. First, load the test data.

```
imageFileName = 'train-images.idx3-ubyte';
labelFileName = 'train-labels.idx1-ubyte';
[Xtest,LabelTest] = processMNISTdata(imageFileName,labelFileName);
```

```
Read MNIST image data...
Number of images in the dataset:  60000 ...
Each image is of 28 by 28 pixels...
The image data is read to a matrix of dimensions:  60000 by  784...
End of reading image data.
```

```
Read MNIST label data...
Number of labels in the dataset:  60000 ...
The label data is read to a matrix of dimensions:  60000 by  1...
End of reading label data.
```

Calculate the classification loss when applying the classifier to the test data.

```
TestX = transform(Mdl,Xtest);
Loss = loss(Cmdl,TestX,LabelTest)
```

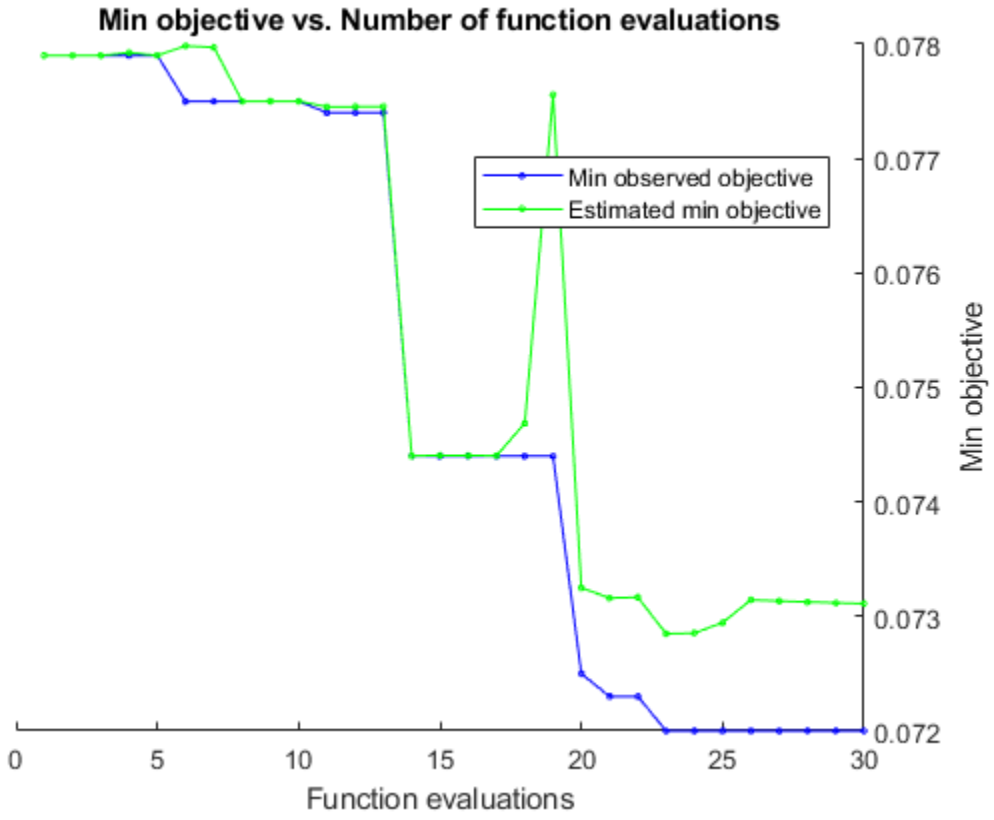
```
Loss =

    0.1009
```

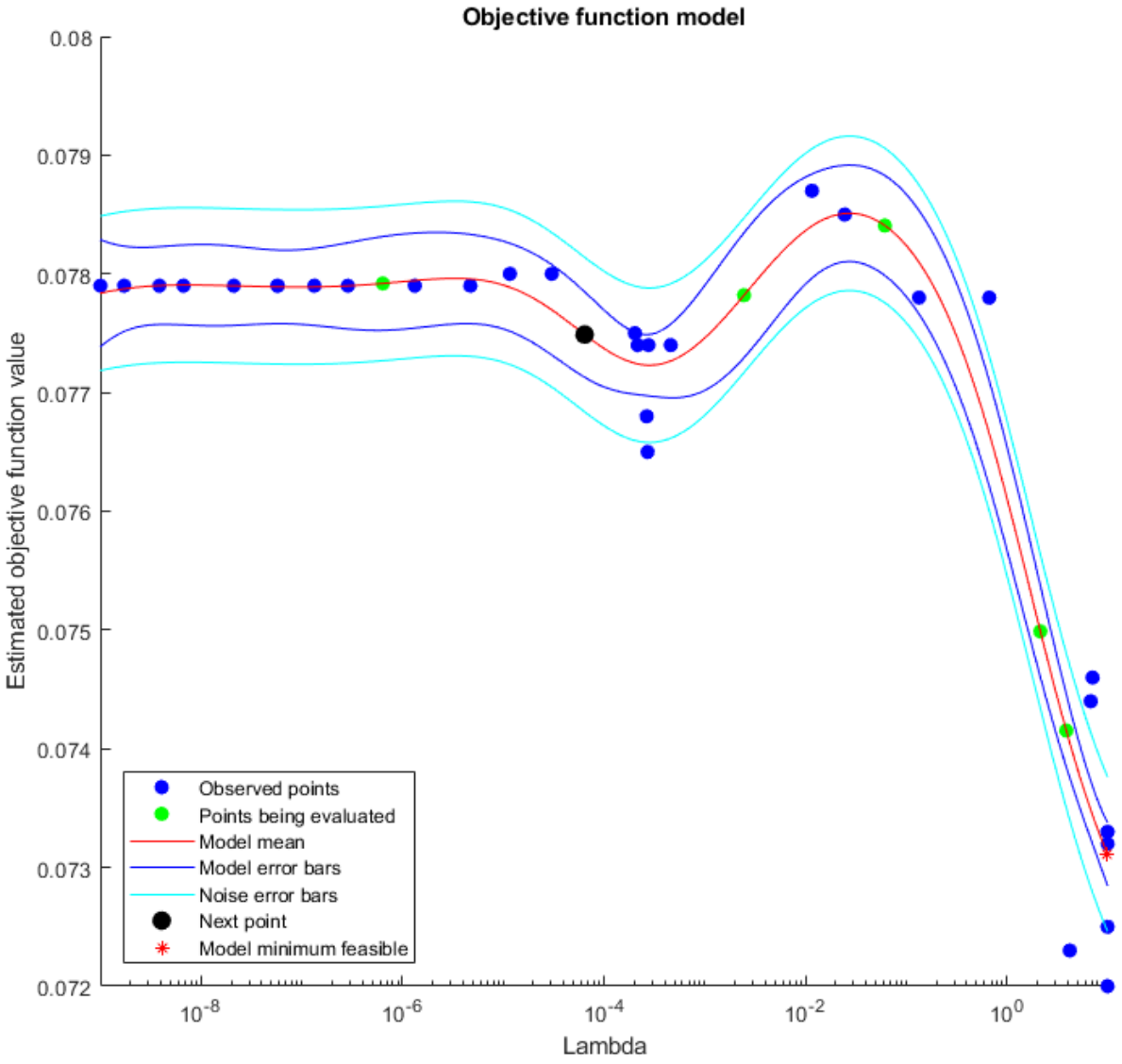
Did this transformation result in a better classifier than one trained on the original data? Create a classifier based on the original training data and evaluate its loss.

```
Omdl = fitcecoc(Xtrain,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);
Losso = loss(Omdl,Xtest,LabelTest)
```

```
Copying objective function to workers...
Done copying objective function to workers.
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|    1 |       5 | Best   |      0.0779 |      46.965 |      0.0779 |      0.0779 |   5.7933e-08 |
|    2 |       5 | Accept |      0.0779 |      47.003 |      0.0779 |      0.0779 |   3.8643e-09 |
|    3 |       5 | Accept |      0.0779 |      47.068 |      0.0779 |      0.0779 |   1.3269e-06 |
|    4 |       6 | Accept |       0.078 |      60.714 |      0.0779 |    0.077925 |   3.0332e-05 |
|    5 |       6 | Accept |      0.0787 |      133.21 |      0.0779 |      0.0779 |     0.011605 |
|    6 |       6 | Best   |      0.0775 |      135.97 |      0.0775 |    0.077983 |   0.00020291 |
|    7 |       6 | Accept |      0.0779 |      44.642 |      0.0775 |    0.077971 |    5.735e-08 |
|    8 |       6 | Accept |      0.0785 |      123.19 |      0.0775 |      0.0775 |     0.024589 |
|    9 |       6 | Accept |      0.0779 |      43.574 |      0.0775 |      0.0775 |   1.0042e-09 |
|   10 |       6 | Accept |      0.0779 |      43.038 |      0.0775 |      0.0775 |   4.7227e-06 |
|   11 |       6 | Best   |      0.0774 |      137.51 |      0.0774 |    0.077451 |   0.00021639 |
|   12 |       6 | Accept |      0.0779 |       44.07 |      0.0774 |    0.077452 |   6.7132e-09 |
|   13 |       6 | Accept |      0.0779 |      44.822 |      0.0774 |    0.077453 |    2.873e-07 |
|   14 |       6 | Best   |      0.0744 |      233.12 |      0.0744 |    0.074402 |        6.805 |
|   15 |       6 | Accept |      0.0778 |      140.49 |      0.0744 |    0.074406 |      0.66889 |
|   16 |       6 | Accept |      0.0774 |      149.32 |      0.0744 |    0.074405 |    0.0002769 |
|   17 |       6 | Accept |      0.0774 |         155 |      0.0744 |    0.074404 |   0.00046083 |
|   18 |       6 | Accept |      0.0765 |      152.63 |      0.0744 |    0.074687 |   0.00027101 |
|   19 |       6 | Accept |      0.0768 |      156.32 |      0.0744 |    0.077558 |   0.00026573 |
|   20 |       6 | Best   |      0.0725 |      255.51 |      0.0725 |    0.073249 |       9.9961 |
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|   21 |       6 | Best   |      0.0723 |       221.5 |      0.0723 |    0.073161 |        4.212 |
|   22 |       6 | Accept |      0.0732 |      259.51 |      0.0723 |    0.073166 |       9.9916 |
|   23 |       6 | Best   |       0.072 |      261.94 |       0.072 |    0.072848 |       9.9883 |
```

```
|   24 |       6 | Accept |      0.0778 |      122.56 |       0.072 |    0.072854 |      0.13413
|   25 |       6 | Accept |      0.0733 |      258.54 |       0.072 |    0.072946 |       9.9904
|   26 |       6 | Accept |      0.0746 |      244.53 |       0.072 |    0.073144 |       7.0911
|   27 |       6 | Accept |      0.0779 |      44.573 |       0.072 |    0.073134 |   2.1183e-08
|   28 |       6 | Accept |       0.078 |      45.478 |       0.072 |    0.073126 |   1.1663e-05
|   29 |       6 | Accept |      0.0779 |      43.954 |       0.072 |    0.073118 |    1.336e-07
|   30 |       6 | Accept |      0.0779 |      44.574 |       0.072 |    0.073112 |   1.7282e-09

__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 690.8688 seconds.
Total objective function evaluation time: 3741.3176

Best observed feasible point:
    Lambda
    _______

    9.9883

Observed objective function value = 0.072
Estimated objective function value = 0.073112
Function evaluation time = 261.9357

Best estimated feasible point (according to models):
    Lambda
    _______

    9.9961

Estimated objective function value = 0.073112
Estimated function evaluation time = 257.9556


Losso =

    0.0865
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.078
0.077
0.076
0.075
0.074
0.073
0.072
0
5
10
15
20
25
30

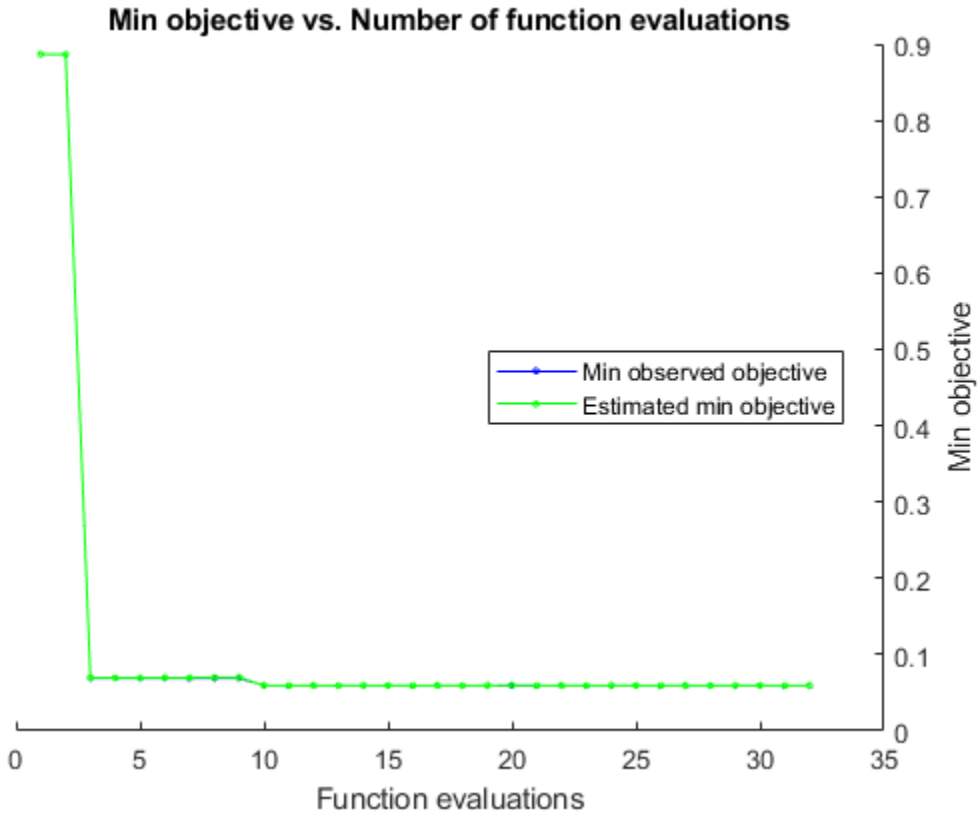


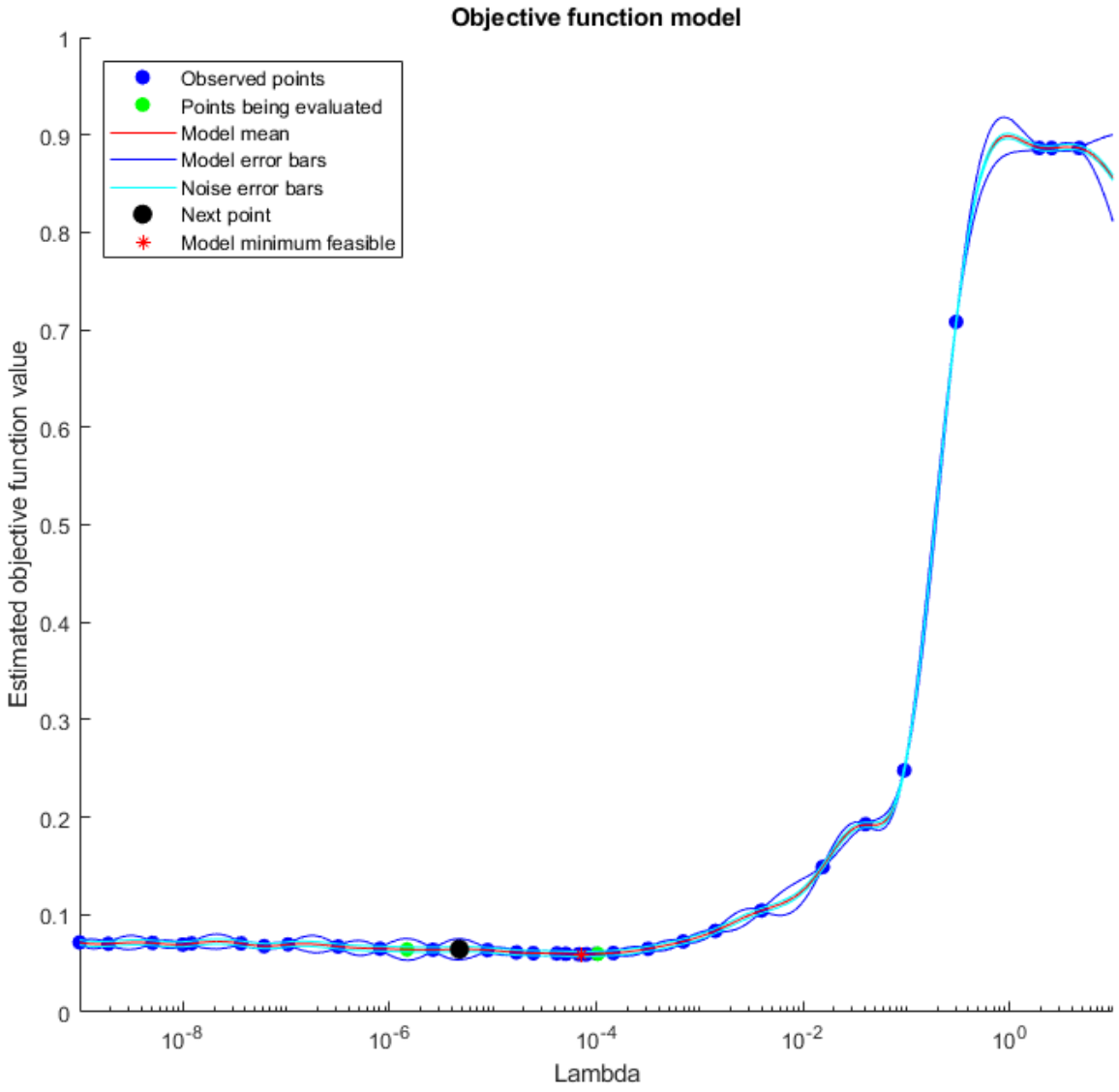
The classifier based on sparse filtering has a somewhat higher loss than the classifier based on the original data. However, the classifier uses only 100 features rather than the 784 features in the original data, and is much faster to create. Try to make a better sparse filtering classifier by increasing `q` from 100 to 200, which is still far less than 784.

```
q = 200;
Mdl2 = sparsefilt(Xtrain,q,'IterationLimit',10);
NewX = transform(Mdl2,Xtrain);
TestX = transform(Mdl2,Xtest);
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);
Loss2 = loss(Cmdl,TestX,LabelTest)
```

```
Warning: Solver LBFGS was not able to converge to a solution.
Copying objective function to workers...
Done copying objective function to workers.
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|    1 |       5 | Best   |      0.8865 |      7.3578 |      0.8865 |      0.8865 |         1.93 |
|    2 |       5 | Accept |      0.8865 |      7.3408 |      0.8865 |      0.8865 |       2.5549 |
|    3 |       6 | Best   |      0.0693 |      9.0077 |      0.0693 |    0.069376 |   9.9515e-09 |
|    4 |       5 | Accept |      0.0705 |      9.1067 |      0.0693 |    0.069374 |   1.2123e-08 |
|    5 |       5 | Accept |      0.1489 |      9.5685 |      0.0693 |    0.069374 |     0.015542 |
|    6 |       6 | Accept |      0.8865 |      7.5032 |      0.0693 |     0.06943 |       4.7067 |
|    7 |       6 | Accept |       0.071 |      8.8044 |      0.0693 |    0.069591 |   5.0861e-09 |
|    8 |       6 | Accept |      0.0715 |      8.9517 |      0.0693 |    0.070048 |    1.001e-09 |
|    9 |       6 | Accept |      0.0833 |      14.393 |      0.0693 |    0.069861 |    0.0014191 |
|   10 |       6 | Best   |      0.0594 |      25.565 |      0.0594 |    0.059458 |    6.767e-05 |
|   11 |       6 | Accept |      0.0651 |      20.074 |      0.0594 |    0.059463 |    8.078e-07 |
|   12 |       6 | Accept |      0.0695 |      14.495 |      0.0594 |    0.059473 |   1.0381e-07 |
|   13 |       6 | Accept |      0.1042 |      12.085 |      0.0594 |    0.059386 |    0.0039745 |
|   14 |       6 | Accept |       0.065 |      20.235 |      0.0594 |    0.059416 |   0.00031759 |
|   15 |       6 | Accept |      0.0705 |      10.929 |      0.0594 |    0.059416 |   3.6503e-08 |
|   16 |       6 | Accept |      0.0637 |      30.593 |      0.0594 |    0.059449 |   8.8718e-06 |
|   17 |       6 | Accept |       0.064 |      25.084 |      0.0594 |    0.059464 |   2.6286e-06 |
|   18 |       6 | Accept |      0.0605 |      31.964 |      0.0594 |    0.059387 |    2.459e-05 |
|   19 |       6 | Accept |      0.0606 |      23.149 |      0.0594 |    0.059312 |    0.0001464 |
|   20 |       6 | Accept |      0.0602 |      32.178 |      0.0594 |    0.059874 |   4.1437e-05 |
|=====================================================================================================|
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |
|      | workers | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|   21 |       6 | Accept |      0.0594 |      27.686 |      0.0594 |    0.059453 |   8.0717e-05 |
|   22 |       6 | Accept |      0.0612 |      33.427 |      0.0594 |    0.059476 |   1.6878e-05 |
|   23 |       6 | Accept |      0.0673 |      17.444 |      0.0594 |    0.059475 |   3.1788e-07 |
|   24 |       6 | Best   |      0.0593 |      26.262 |      0.0593 |     0.05944 |   7.8179e-05 |
|   25 |       6 | Accept |       0.248 |      7.6345 |      0.0593 |    0.059409 |     0.095654 |
|   26 |       6 | Accept |      0.0598 |      28.536 |      0.0593 |    0.059465 |   5.0819e-05 |
|   27 |       6 | Accept |      0.0701 |      9.0545 |      0.0593 |    0.059466 |   1.8937e-09 |
|   28 |       5 | Accept |      0.7081 |      7.1176 |      0.0593 |    0.059372 |      0.30394 |
|   29 |       5 | Accept |      0.0676 |      11.782 |      0.0593 |    0.059372 |   6.1136e-08 |
|   30 |       3 | Accept |        0.06 |      23.556 |      0.0593 |    0.059422 |   0.00010144 |
|   31 |       3 | Accept |      0.0725 |      16.069 |      0.0593 |    0.059422 |   0.00069403 |
|   32 |       3 | Accept |      0.1928 |      8.3732 |      0.0593 |    0.059422 |     0.040402 |

__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 32
Total elapsed time: 97.7946 seconds.
Total objective function evaluation time: 545.3255

Best observed feasible point:
      Lambda
    __________

    7.8179e-05

Observed objective function value = 0.0593
```

```
Estimated objective function value = 0.059422
Function evaluation time = 26.2624

Best estimated feasible point (according to models):
      Lambda
    __________

    7.8179e-05

Estimated objective function value = 0.059422
Estimated function evaluation time = 26.508


Loss2 =

    0.0682
```





This time the classification loss is lower than that of the original data classifier.

Try RICA

Try the other feature extraction function, `rica`. Extract 200 features, create a classifier, and examine its loss on the test data. Use more iterations for the `rica` function, because `rica` can perform better with more iterations than `sparsefilt` uses.

Often prior to feature extraction, you "prewhiten" the input data as a data preprocessing step. The prewhitening step includes two transforms, decorrelation and standardization, which make the predictors have zero mean and identity covariance. `rica` supports only the standardization transform. You use the `Standardize` name-value pair argument to make the predictors have zero

mean and unit variance. Alternatively, you can transform images for contrast normalization individually by applying the `zscore` transformation before calling `sparsefilt` or `rica`.

```
Mdl3 = rica(Xtrain,q,'IterationLimit',400,'Standardize',true);
NewX = transform(Mdl3,Xtrain);
TestX = transform(Mdl3,Xtest);
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);
Loss3 = loss(Cmdl,TestX,LabelTest)
```

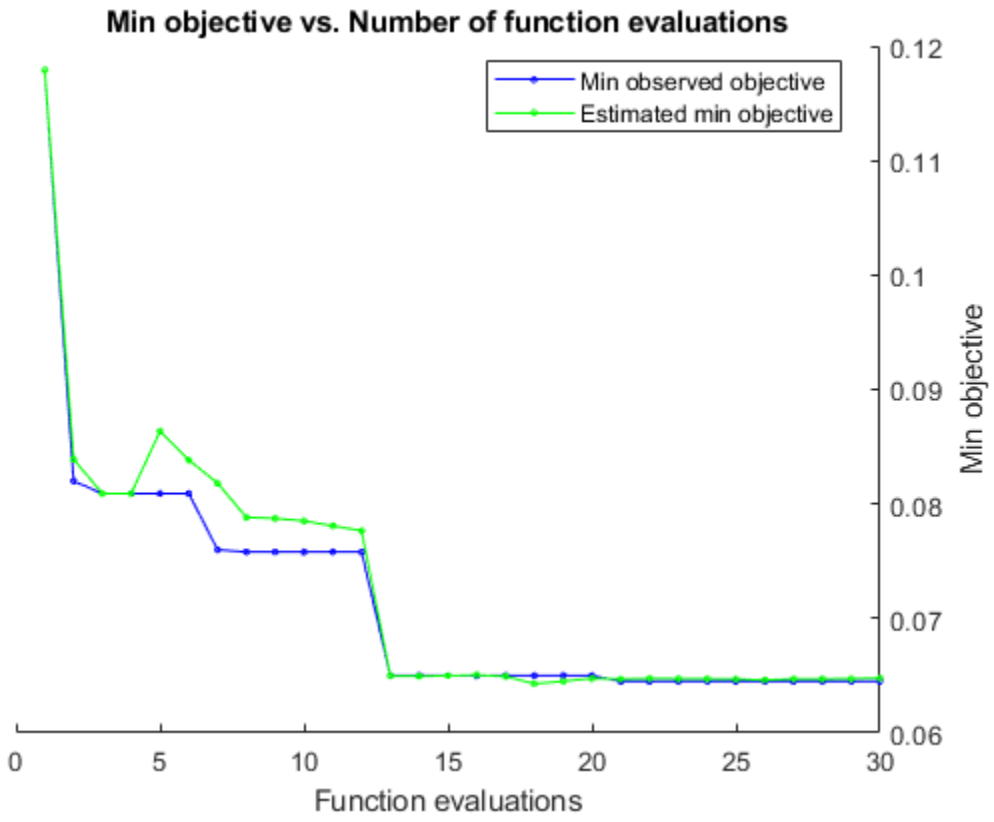
```
Warning: Solver LBFGS was not able to converge to a solution.
Copying objective function to workers...
Done copying objective function to workers.
```

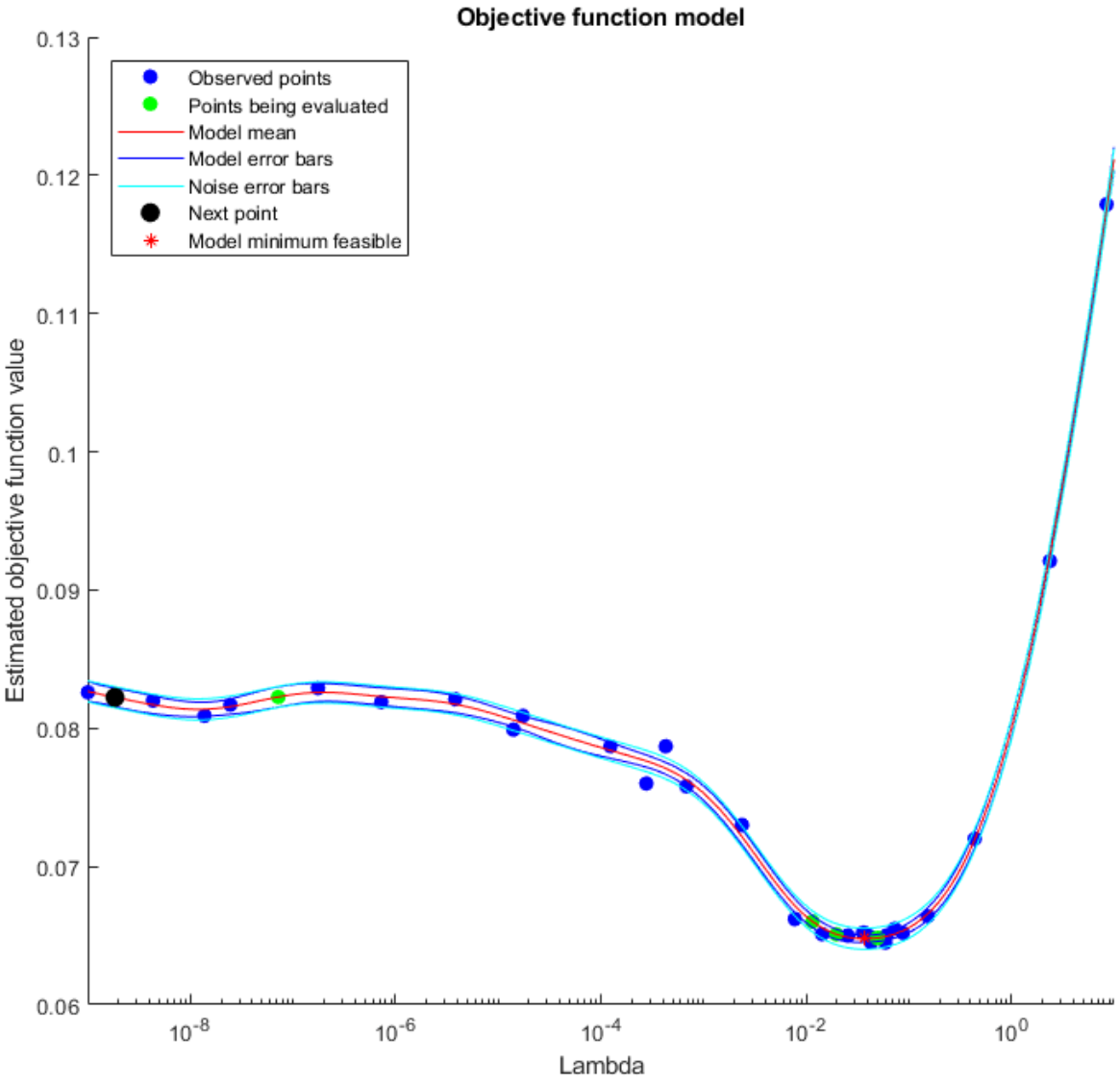
Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
1	6	Best	0.1179	12.012	0.1179	0.1179	8.4727
2	6	Best	0.082	13.384	0.082	0.083897	4.3291e-09
3	6	Best	0.0809	18.917	0.0809	0.080902	1.738e-05
4	6	Accept	0.0821	19.172	0.0809	0.08091	3.8101e-06
5	6	Accept	0.0921	14.445	0.0809	0.086349	2.3753
6	6	Accept	0.0809	13.393	0.0809	0.083836	1.3757e-08
7	6	Best	0.076	28.075	0.076	0.081808	0.00027773
8	6	Best	0.0758	29.686	0.0758	0.078829	0.00068195
9	6	Accept	0.0829	13.373	0.0758	0.078733	1.7543e-07
10	6	Accept	0.0826	14.031	0.0758	0.078512	1.0045e-09
11	6	Accept	0.0817	13.662	0.0758	0.078077	2.4568e-08
12	6	Accept	0.0799	19.311	0.0758	0.077658	1.4061e-05
13	6	Best	0.065	25.148	0.065	0.064974	0.060326
14	6	Accept	0.0787	23.434	0.065	0.064947	0.00012407
15	6	Accept	0.072	19.167	0.065	0.064997	0.43899
16	6	Accept	0.073	28.39	0.065	0.065053	0.0023721
17	6	Accept	0.0787	29.887	0.065	0.064928	0.00042914
18	6	Accept	0.0662	26.374	0.065	0.064295	0.0077638
19	6	Accept	0.0652	24.937	0.065	0.064502	0.087389
20	6	Accept	0.0655	25.416	0.065	0.064762	0.072931

Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
21	6	Best	0.0645	25.529	0.0645	0.064691	0.059245
22	6	Accept	0.065	23.832	0.0645	0.06474	0.025521
23	6	Accept	0.0819	20.343	0.0645	0.064732	7.2593e-07
24	6	Accept	0.0664	23.732	0.0645	0.064718	0.1534
25	6	Accept	0.0651	24.796	0.0645	0.064693	0.038371
26	6	Accept	0.0651	25.449	0.0645	0.064613	0.014318
27	6	Accept	0.0652	25.092	0.0645	0.064713	0.037107
28	6	Accept	0.0645	24.404	0.0645	0.0647	0.042959
29	6	Accept	0.0649	24.704	0.0645	0.064729	0.042776
30	6	Accept	0.0652	24.341	0.0645	0.064786	0.035788

```
__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 124.9755 seconds.
```

```
Total objective function evaluation time: 654.4364

Best observed feasible point:
     Lambda
    ________

    0.059245

Observed objective function value = 0.0645
Estimated objective function value = 0.064932
Function evaluation time = 25.5294

Best estimated feasible point (according to models):
     Lambda
    ________

    0.042776

Estimated objective function value = 0.064786
Estimated function evaluation time = 24.7849


Loss3 =

    0.0749
```





The `rica`-based classifier has somewhat higher test loss compared to the sparse filtering classifier.

Try More Features

The feature extraction functions have few tuning parameters. One parameter that can affect results is the number of requested features. See how well classifiers work when based on 1000 features, rather than the 200 features previously tried, or the 784 features in the original data. Using more features than appear in the original data is called "overcomplete" learning. Conversely, using fewer features is called "undercomplete" learning. Overcomplete learning can lead to increased classification accuracy, while undercomplete learning can save memory and time.

```
q = 1000;
Mdl4 = sparsefilt(Xtrain,q,'IterationLimit',10);
```

```
NewX = transform(Mdl4,Xtrain);
TestX = transform(Mdl4,Xtest);
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);
Loss4 = loss(Cmdl,TestX,LabelTest)
```

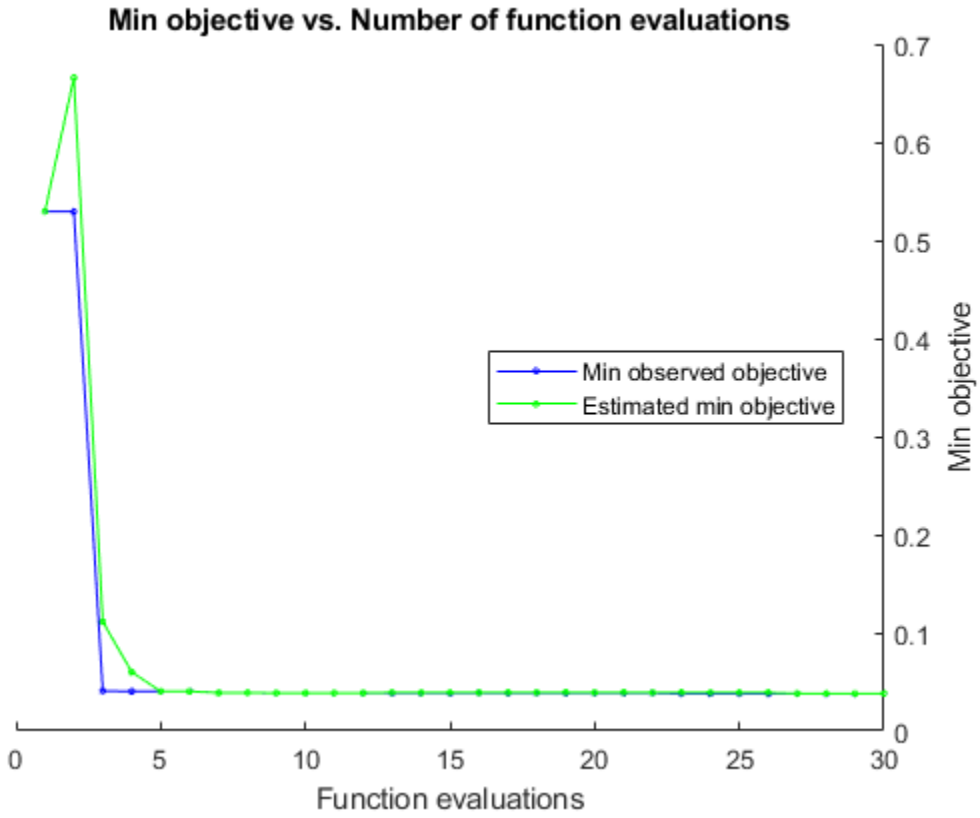
```
Warning: Solver LBFGS was not able to converge to a solution.
Copying objective function to workers...
Done copying objective function to workers.
```

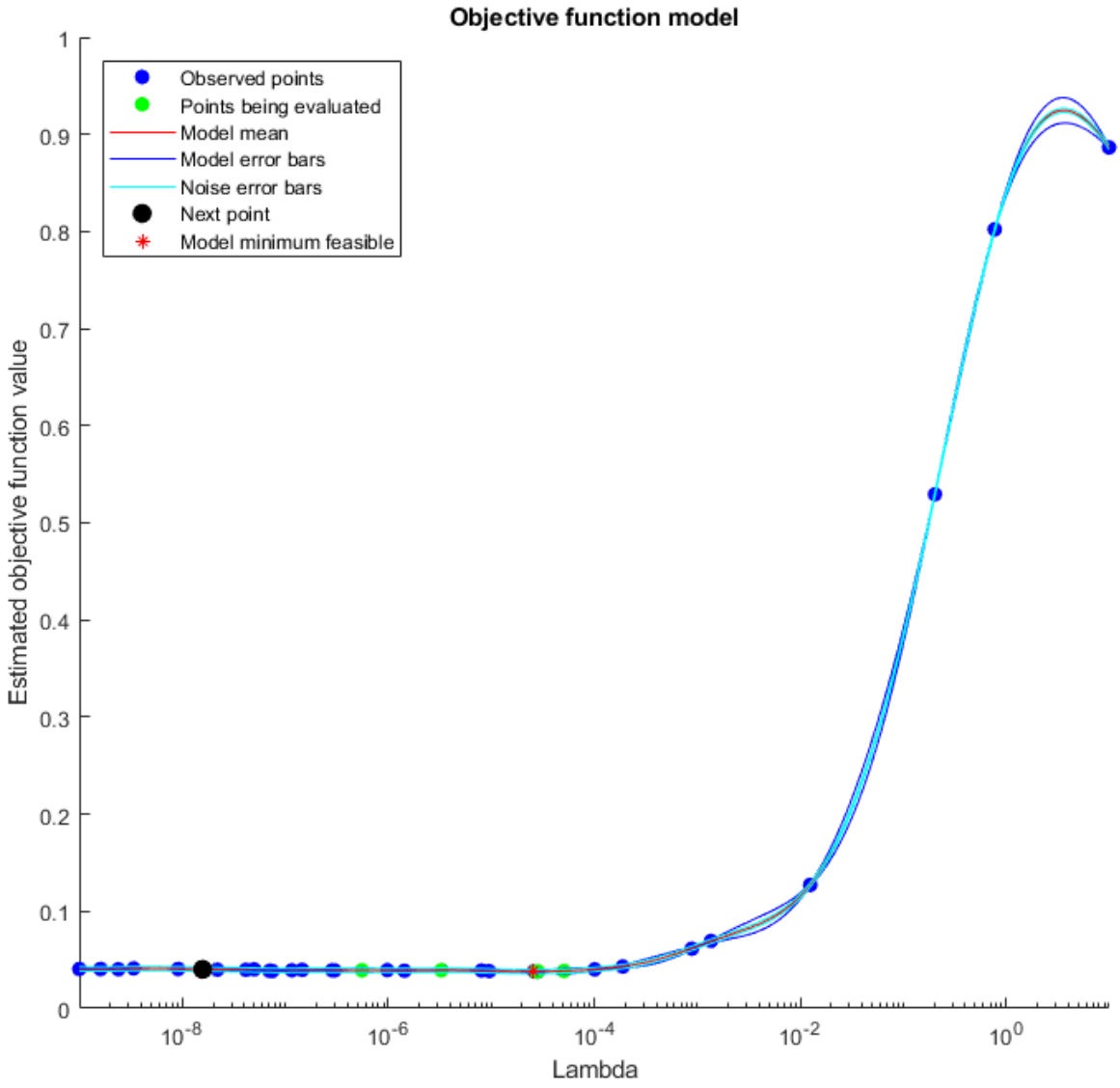
Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
1	6	Best	0.5293	39.885	0.5293	0.5293	0.20333
2	6	Accept	0.8022	43.475	0.5293	0.66575	0.77337
3	6	Best	0.0406	52.594	0.0406	0.11113	9.1082e-09
4	6	Best	0.0403	54.73	0.0403	0.060037	2.3947e-09
5	6	Accept	0.0695	124.96	0.0403	0.040319	0.001361
6	6	Accept	0.0406	53.691	0.0403	0.040207	1.0005e-09
7	6	Best	0.0388	178.69	0.0388	0.038811	1.4358e-06
8	6	Accept	0.0615	138.53	0.0388	0.038817	0.00088731
9	6	Best	0.0385	61.81	0.0385	0.038557	7.4709e-08
10	6	Accept	0.0399	54.198	0.0385	0.038555	2.1909e-08
11	6	Accept	0.0402	234.55	0.0385	0.038639	0.000101
12	6	Accept	0.0431	198.09	0.0385	0.038636	0.00018896
13	6	Accept	0.0393	75.811	0.0385	0.039016	1.1597e-07
14	6	Accept	0.0387	61.281	0.0385	0.038908	7.0518e-08
15	6	Accept	0.0393	125.73	0.0385	0.038931	2.8429e-07
16	6	Accept	0.0397	89.804	0.0385	0.039106	1.4603e-07
17	6	Accept	0.0391	126.88	0.0385	0.039081	3.0065e-07
18	6	Accept	0.0398	56.157	0.0385	0.039123	4.1563e-08
19	6	Accept	0.0406	55.25	0.0385	0.039122	1.0014e-09
20	6	Accept	0.0385	272.92	0.0385	0.039127	9.568e-06

Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
21	6	Accept	0.0412	55.191	0.0385	0.039124	3.3737e-09
22	6	Accept	0.0394	229.72	0.0385	0.039117	3.2757e-06
23	6	Best	0.0379	295.55	0.0379	0.039116	2.8439e-05
24	6	Accept	0.0394	168.74	0.0379	0.039111	9.778e-07
25	6	Accept	0.039	281.91	0.0379	0.039112	8.0694e-06
26	6	Accept	0.8865	54.865	0.0379	0.038932	9.9885
27	6	Accept	0.0381	300.7	0.0379	0.037996	2.6027e-05
28	6	Accept	0.0406	54.611	0.0379	0.037996	1.6057e-09
29	6	Accept	0.1272	76.648	0.0379	0.037997	0.012507
30	6	Accept	0.0403	57.931	0.0379	0.037997	4.9907e-08

```
__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 724.6036 seconds.
Total objective function evaluation time: 3674.8899

Best observed feasible point:
      Lambda
```

```
    __________

    2.8439e-05

Observed objective function value = 0.0379
Estimated objective function value = 0.03801
Function evaluation time = 295.5515

Best estimated feasible point (according to models):
      Lambda
    __________

    2.6027e-05

Estimated objective function value = 0.037997
Estimated function evaluation time = 297.6756


Loss4 =

    0.0440
```





The classifier based on overcomplete sparse filtering with 1000 extracted features has the lowest test loss of any classifier yet tested.

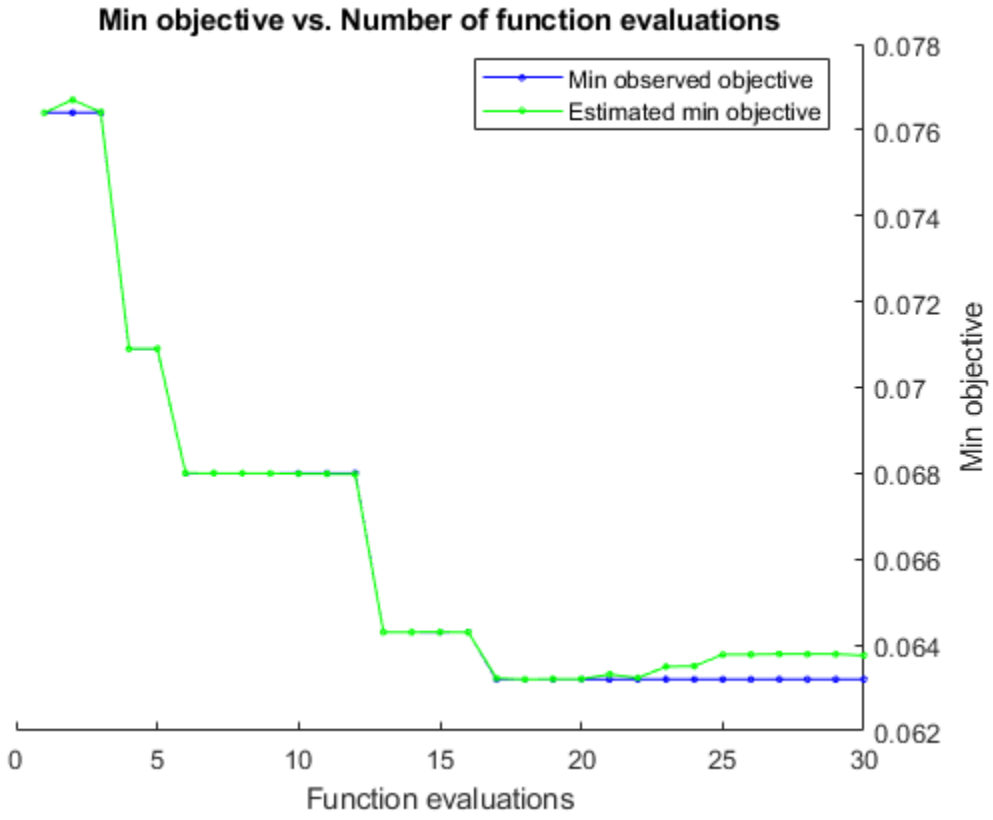
```
Mdl5 = rica(Xtrain,q,'IterationLimit',400,'Standardize',true);
NewX = transform(Mdl5,Xtrain);
TestX = transform(Mdl5,Xtest);
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t, ...
    'OptimizeHyperparameters',{'Lambda'}, ...
    'HyperparameterOptimizationOptions',options);
Loss5 = loss(Cmdl,TestX,LabelTest)
```

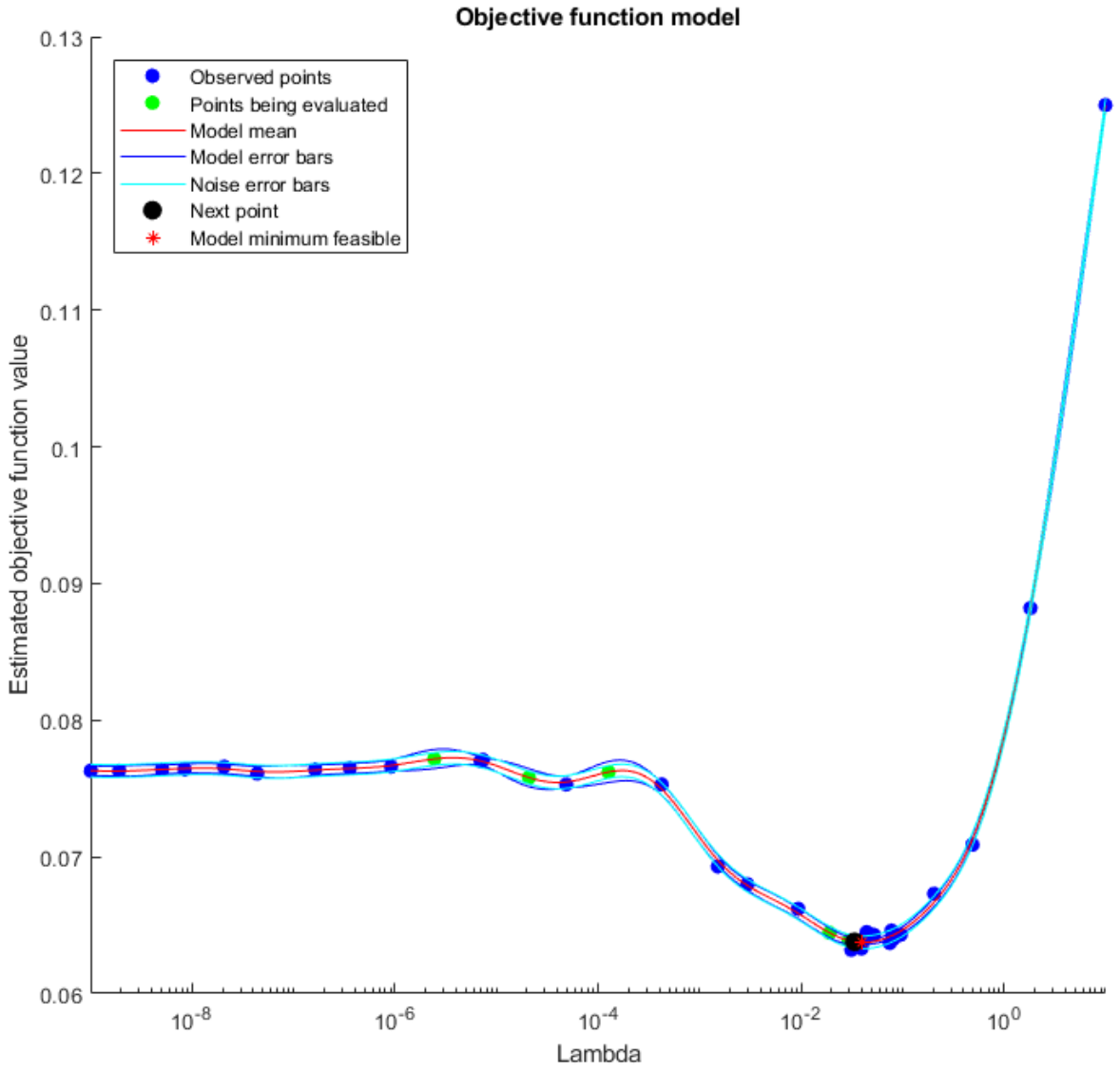
```
Warning: Solver LBFGS was not able to converge to a solution.
Copying objective function to workers...
Done copying objective function to workers.
```

Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
1	6	Best	0.0764	46.206	0.0764	0.0764	8.4258e-09
2	6	Accept	0.077	141.95	0.0764	0.0767	6.9536e-06
3	6	Accept	0.0771	146.87	0.0764	0.076414	7.3378e-06
4	6	Best	0.0709	182.51	0.0709	0.0709	0.48851
5	6	Accept	0.0764	46.923	0.0709	0.070903	5.0695e-09
6	6	Best	0.068	294.89	0.068	0.068004	0.0029652
7	6	Accept	0.125	99.095	0.068	0.068001	9.9814
8	6	Accept	0.0693	321.66	0.068	0.067999	0.0015167
9	6	Accept	0.0882	138.03	0.068	0.068	1.8203
10	6	Accept	0.0753	285.07	0.068	0.067991	0.00042423
11	6	Accept	0.0764	47.704	0.068	0.067984	1.6326e-07
12	6	Accept	0.0763	46.514	0.068	0.06798	1.0048e-09
13	6	Best	0.0643	252.2	0.0643	0.0643	0.095965
14	6	Accept	0.0766	168.37	0.0643	0.0643	9.1336e-07
15	6	Accept	0.0753	153.29	0.0643	0.064301	4.8641e-05
16	6	Accept	0.0662	256.65	0.0643	0.064298	0.0093576
17	6	Best	0.0632	224.2	0.0632	0.063226	0.031314
18	6	Accept	0.0673	219.59	0.0632	0.063201	0.20528
19	6	Accept	0.0637	244.17	0.0632	0.063208	0.075001
20	6	Accept	0.064	234.85	0.0632	0.06321	0.081232
Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	Lambda
21	6	Accept	0.0646	242.2	0.0632	0.063315	0.078081
22	6	Accept	0.0633	217.97	0.0632	0.063233	0.039495
23	6	Accept	0.0643	224.22	0.0632	0.063496	0.052107
24	6	Accept	0.0761	45.102	0.0632	0.063509	4.3946e-08
25	6	Accept	0.0645	221.24	0.0632	0.063778	0.044455
26	6	Accept	0.0763	44.572	0.0632	0.063778	1.9139e-09
27	6	Accept	0.0639	216.9	0.0632	0.063791	0.041759
28	6	Accept	0.0766	45.609	0.0632	0.06379	2.0642e-08
29	6	Accept	0.0765	121.35	0.0632	0.063789	3.5882e-07
30	6	Accept	0.0636	215.47	0.0632	0.063755	0.038062

```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 952.7987 seconds.
Total objective function evaluation time: 5145.3787

Best observed feasible point:
     Lambda
    ________

    0.031314

Observed objective function value = 0.0632
Estimated objective function value = 0.063828
Function evaluation time = 224.2018

Best estimated feasible point (according to models):
     Lambda
```

```
    ________

    0.044455

Estimated objective function value = 0.063755
Estimated function evaluation time = 219.4845


Loss5 =

    0.0748
```





The classifier based on RICA with 1000 extracted features has a similar test loss to the RICA classifier based on 200 extracted features.

Optimize Hyperparameters by Using `bayesopt`

Feature extraction functions have these tuning parameters:

- Iteration limit
- Function, either `rica` or `sparsefilt`
- Parameter `Lambda`
- Number of learned features `q`

The `fitcecoc` regularization parameter also affects the accuracy of the learned classifier. Include that parameter in the list of hyperparameters as well.

To search among the available parameters effectively, try `bayesopt`. Use the following objective function, which includes parameters passed from the workspace.

```
function objective = filterica(x,Xtrain,Xtest,LabelTrain,LabelTest,winit)

initW = winit(1:size(Xtrain,2),1:x.q);
if char(x.solver) == 'r'
    Mdl = rica(Xtrain,x.q,'Lambda',x.lambda,'IterationLimit',x.iterlim, ...
        'InitialTransformWeights',initW,'Standardize',true);
else
    Mdl = sparsefilt(Xtrain,x.q,'Lambda',x.lambda,'IterationLimit',x.iterlim, ...
        'InitialTransformWeights',initW);
end

NewX = transform(Mdl,Xtrain);
TestX = transform(Mdl,Xtest);
t = templateLinear('Lambda',x.lambdareg,'Solver','lbfgs');
Cmdl = fitcecoc(NewX,LabelTrain,'Learners',t);
objective = loss(Cmdl,TestX,LabelTest);
```

To remove sources of variation, fix an initial transform weight matrix.

```
W = randn(1e4,1e3);
```

Create hyperparameters for the objective function.

```
iterlim = optimizableVariable('iterlim',[5,500],'Type','integer');
lambda = optimizableVariable('lambda',[0,10]);
solver = optimizableVariable('solver',{'r','s'},'Type','categorical');
qvar = optimizableVariable('q',[10,1000],'Type','integer');
lambdareg = optimizableVariable('lambdareg',[1e-6,1],'Transform','log');
vars = [iterlim,lambda,solver,qvar,lambdareg];
```

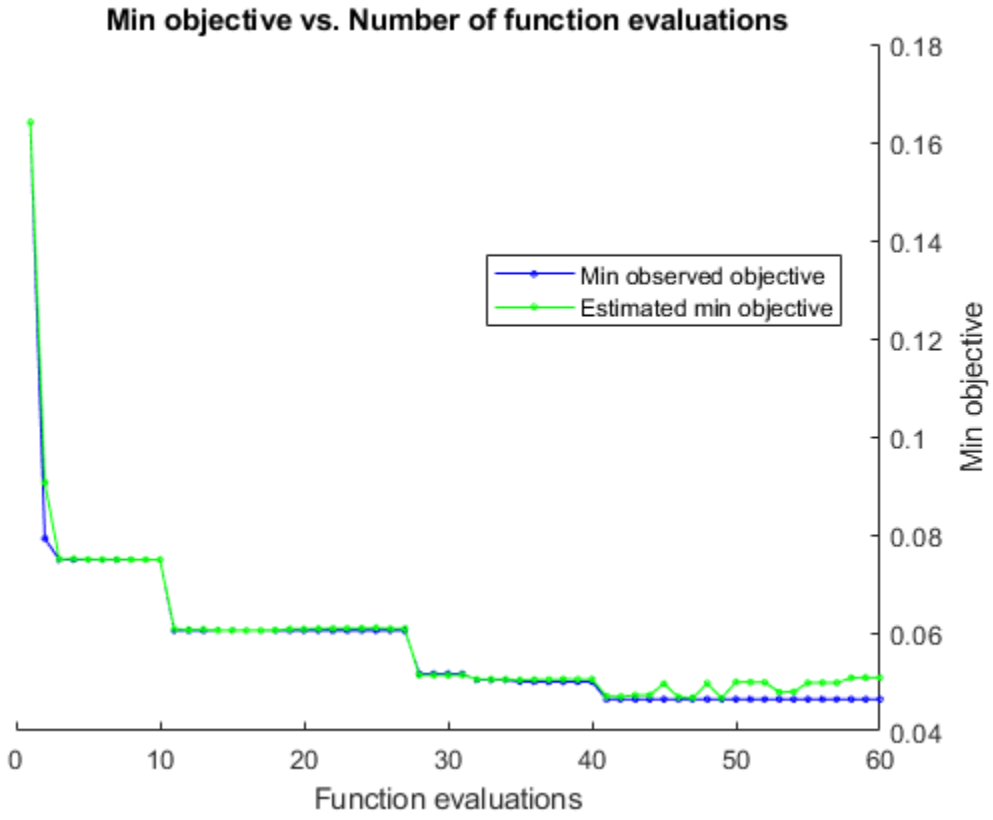
Run the optimization without the warnings that occur when the internal optimizations do not run to completion. Run for 60 iterations instead of the default 30 to give the optimization a better chance of locating a good value.

```
warning('off','stats:classreg:learning:fsutils:Solver:LBFGSUnableToConverge');
results = bayesopt(@(x) filterica(x,Xtrain,Xtest,LabelTrain,LabelTest,W),vars, ...
    'UseParallel',true,'MaxObjectiveEvaluations',60);
warning('on','stats:classreg:learning:fsutils:Solver:LBFGSUnableToConverge');
```

```
Copying objective function to workers...
Done copying objective function to workers.
|=====================================================================================================
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |      iterlim
|      | workers | result |             | runtime     | (observed)  | (estim.)    |
|=====================================================================================================
|    1 |       6 | Best   |     0.16408 |      33.743 |     0.16408 |     0.16408 |          140
|    2 |       6 | Best   |    0.079213 |      51.975 |    0.079213 |     0.09064 |           10
|    3 |       6 | Best   |    0.074897 |      82.031 |    0.074897 |    0.074983 |           32
|    4 |       6 | Accept |     0.07546 |      93.221 |    0.074897 |    0.075073 |          178
|    5 |       6 | Accept |     0.13924 |      30.444 |    0.074897 |    0.074933 |          282
```

```
|    6 |       6 | Accept |    0.083964 |         133 |    0.074897 |    0.074933 |           58
|    7 |       6 | Accept |     0.08128 |      33.609 |    0.074897 |    0.074957 |            8
|    8 |       6 | Accept |    0.090751 |      203.96 |    0.074897 |    0.074913 |          131
|    9 |       6 | Accept |    0.090001 |      172.38 |    0.074897 |    0.074904 |          146
|   10 |       6 | Accept |    0.080191 |       316.8 |    0.074897 |    0.074897 |          164
|   11 |       6 | Best   |    0.060472 |      40.777 |    0.060472 |    0.060731 |            5
|   12 |       6 | Accept |    0.079027 |      45.841 |    0.060472 |    0.060632 |            8
|   13 |       6 | Accept |    0.074823 |      237.43 |    0.060472 |     0.06067 |          109
|   14 |       6 | Accept |     0.84009 |      85.121 |    0.060472 |    0.060468 |          306
|   15 |       6 | Accept |     0.15637 |      200.13 |    0.060472 |    0.060451 |           90
|   16 |       6 | Accept |     0.69006 |      14.273 |    0.060472 |     0.06047 |            6
|   17 |       6 | Accept |    0.093035 |      205.83 |    0.060472 |    0.060469 |          263
|   18 |       6 | Accept |     0.18753 |      6.0238 |    0.060472 |    0.060527 |           36
|   19 |       6 | Accept |       0.119 |      749.98 |    0.060472 |    0.060751 |          482
|   20 |       6 | Accept |    0.076414 |      751.21 |    0.060472 |    0.060754 |          387
|==========================================================================================================
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |      iterlim
|      | workers | result |             | runtime     | (observed)  | (estim.)    |
|==========================================================================================================
|   21 |       6 | Accept |    0.099332 |      7.2298 |    0.060472 |    0.060828 |           20
|   22 |       6 | Accept |    0.090139 |      7.9815 |    0.060472 |    0.060858 |           11
|   23 |       6 | Accept |    0.076696 |      323.64 |    0.060472 |    0.060872 |          120
|   24 |       6 | Accept |    0.098003 |      50.544 |    0.060472 |    0.060876 |          492
|   25 |       6 | Accept |     0.10383 |      56.568 |    0.060472 |     0.06101 |           11
|   26 |       6 | Accept |     0.14405 |      30.426 |    0.060472 |    0.060797 |          477
|   27 |       6 | Accept |     0.09046 |      53.398 |    0.060472 |    0.060815 |           13
|   28 |       6 | Best   |    0.051641 |      99.452 |    0.051641 |    0.051368 |           23
|   29 |       6 | Accept |     0.10016 |      6.4162 |    0.051641 |    0.051365 |            6
|   30 |       6 | Accept |     0.10943 |      40.676 |    0.051641 |    0.051391 |          488
|   31 |       6 | Accept |    0.086761 |      7.8419 |    0.051641 |    0.051393 |           24
|   32 |       6 | Best   |      0.0504 |      96.816 |      0.0504 |    0.050526 |           14
|   33 |       6 | Accept |    0.088789 |      81.158 |      0.0504 |    0.050525 |           14
|   34 |       6 | Accept |    0.083083 |      887.17 |      0.0504 |     0.05052 |          351
|   35 |       6 | Best   |    0.050023 |      99.493 |    0.050023 |    0.050372 |           19
|   36 |       6 | Accept |    0.053338 |      113.36 |    0.050023 |    0.050499 |            7
|   37 |       6 | Accept |    0.089024 |      70.047 |    0.050023 |      0.0505 |           15
|   38 |       6 | Accept |    0.052029 |      95.822 |    0.050023 |    0.050551 |            7
|   39 |       6 | Accept |    0.085992 |      73.422 |    0.050023 |    0.050528 |            5
|   40 |       6 | Accept |    0.091159 |      5.8348 |    0.050023 |     0.05052 |           15
|==========================================================================================================
| Iter | Active  | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |      iterlim
|      | workers | result |             | runtime     | (observed)  | (estim.)    |
|==========================================================================================================
|   41 |       6 | Best   |    0.046444 |      152.93 |    0.046444 |    0.047062 |           30
|   42 |       6 | Accept |    0.052712 |      58.107 |    0.046444 |     0.04698 |           12
|   43 |       6 | Accept |    0.058005 |      91.928 |    0.046444 |    0.047263 |           10
|   44 |       6 | Accept |    0.055413 |      103.25 |    0.046444 |    0.047306 |            7
|   45 |       6 | Accept |    0.052517 |      96.201 |    0.046444 |    0.049604 |           10
|   46 |       6 | Accept |    0.089527 |      76.617 |    0.046444 |    0.046888 |           20
|   47 |       6 | Accept |    0.050062 |      99.709 |    0.046444 |    0.046735 |           12
|   48 |       6 | Accept |     0.21166 |      90.117 |    0.046444 |    0.049716 |          495
|   49 |       6 | Accept |    0.054535 |        79.1 |    0.046444 |    0.046679 |            6
|   50 |       6 | Accept |     0.12385 |      964.74 |    0.046444 |    0.049963 |          474
|   51 |       6 | Accept |    0.052016 |      76.098 |    0.046444 |    0.049914 |           10
|   52 |       6 | Accept |    0.048984 |      95.054 |    0.046444 |    0.049891 |           12
|   53 |       6 | Accept |      0.1948 |      889.11 |    0.046444 |    0.047903 |          466
|   54 |       6 | Accept |     0.10652 |       5.076 |    0.046444 |    0.047961 |           10
|   55 |       6 | Accept |    0.074194 |      319.41 |    0.046444 |     0.04981 |          130
```

```
|   56 |       6 | Accept |      0.1014 |      45.184 |    0.046444 |    0.049828 |          480 |
|   57 |       6 | Accept |     0.33214 |      3.1996 |    0.046444 |    0.049785 |           12 |
|   58 |       6 | Accept |    0.054348 |      96.616 |    0.046444 |    0.050832 |           12 |
|   59 |       6 | Accept |     0.71471 |      3.0555 |    0.046444 |    0.050852 |           10 |
|   60 |       6 | Accept |    0.074353 |      67.118 |    0.046444 |     0.05084 |            8 |

__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 60 reached.
Total function evaluations: 60
Total elapsed time: 1921.1117 seconds.
Total objective function evaluation time: 9107.7006

Best observed feasible point:
    iterlim    lambda    solver     q     lambdareg
    _______    ______    ______    ___    _________

      30       4.0843      s       997    7.279e-06

Observed objective function value = 0.046444
Estimated objective function value = 0.053743
Function evaluation time = 152.932

Best estimated feasible point (according to models):
    iterlim    lambda    solver     q     lambdareg
    _______    ______    ______    ___    _________

      10       1.0798      s       922    1.133e-06

Estimated objective function value = 0.05084
Estimated function evaluation time = 90.9315
```



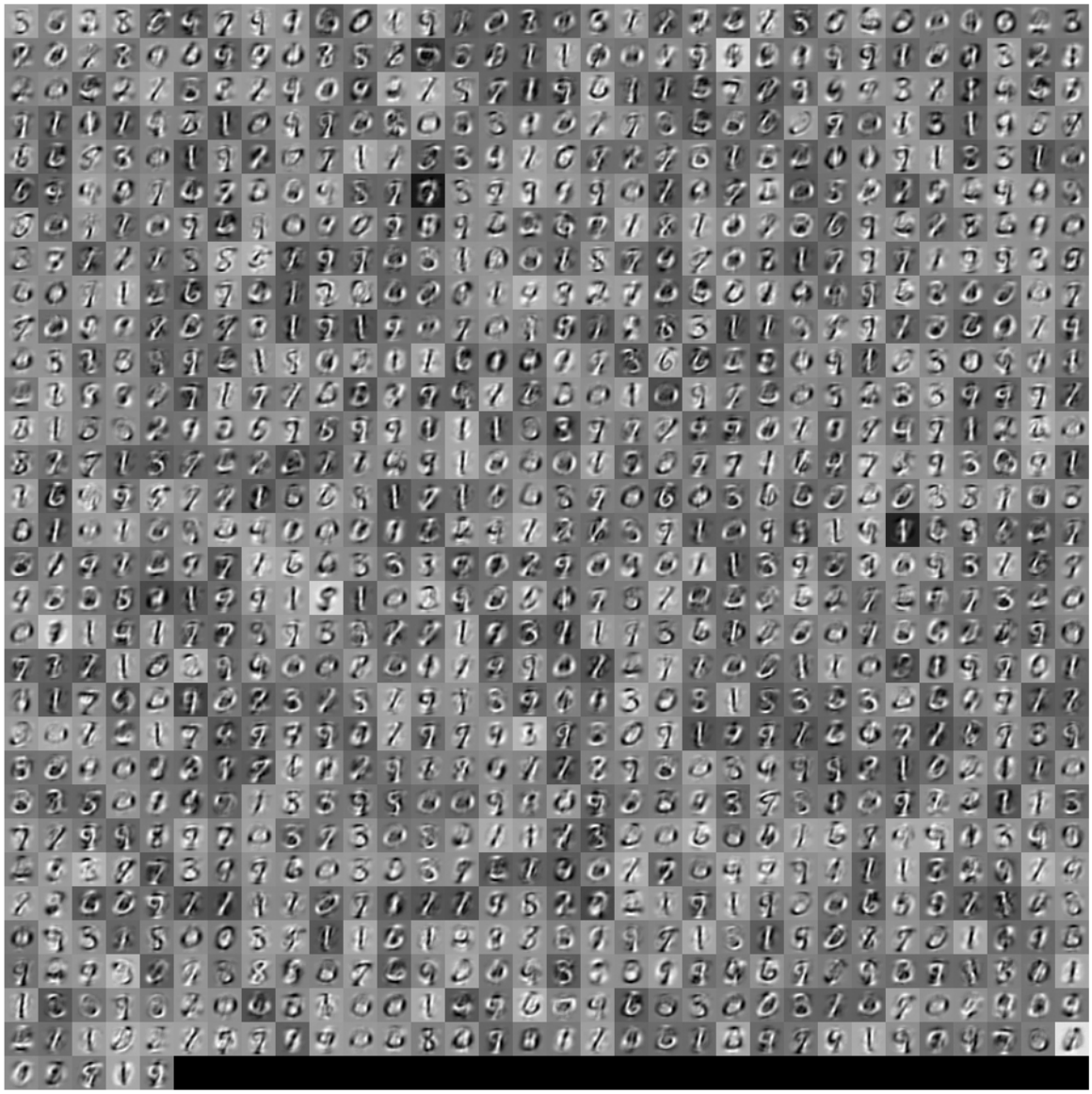
The resulting classifier does not have better (lower) loss than the classifier using `sparsefilt` for 1000 features, trained for 10 iterations.

View the filter coefficients for the best hyperparameters that `bayesopt` found. The resulting images show the shapes of the extracted features. These shapes are recognizable as portions of handwritten digits.

```
Xtbl = results.XAtMinObjective;
Q = Xtbl.q;
initW = W(1:size(Xtrain,2),1:Q);
if char(Xtbl.solver) == 'r'
    Mdl = rica(Xtrain,Q,'Lambda',Xtbl.lambda,'IterationLimit',Xtbl.iterlim, ...
        'InitialTransformWeights',initW,'Standardize',true);
else
    Mdl = sparsefilt(Xtrain,Q,'Lambda',Xtbl.lambda,'IterationLimit',Xtbl.iterlim, ...
        'InitialTransformWeights',initW);
end
Wts = Mdl.TransformWeights;
Wts = reshape(Wts,[28,28,Q]);
[dx,dy,~,~] = size(Wts);
for f = 1:Q
    Wvec = Wts(:,:,f);
    Wvec = Wvec(:);
    Wvec =(Wvec - min(Wvec))/(max(Wvec) - min(Wvec));
    Wts(:,:,f) = reshape(Wvec,dx,dy);
end
m   = ceil(sqrt(Q));
```

```
n   = m;
img = zeros(m*dx,n*dy);
f   = 1;
for i = 1:m
    for j = 1:n
        if (f <= Q)
            img((i-1)*dx+1:i*dx,(j-1)*dy+1:j*dy,:) = Wts(:,:,f);
            f = f+1;
        end
    end
end
imshow(img);
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```



Code for Reading MNIST Data

The code of the function that reads the data into the workspace is:

```
function [X,L] = processMNISTdata(imageFileName,labelFileName)

[fileID,errmsg] = fopen(imageFileName,'r','b');
if fileID < 0
    error(errmsg);
end
%%
% First read the magic number. This number is 2051 for image data, and
```

```
% 2049 for label data
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2051
    fprintf('\nRead MNIST image data...\n')
end
%%
% Then read the number of images, number of rows, and number of columns
numImages = fread(fileID,1,'int32',0,'b');
fprintf('Number of images in the dataset: %6d ...\n',numImages);
numRows = fread(fileID,1,'int32',0,'b');
numCols = fread(fileID,1,'int32',0,'b');
fprintf('Each image is of %2d by %2d pixels...\n',numRows,numCols);
%%
% Read the image data
X = fread(fileID,inf,'unsigned char');
%%
% Reshape the data to array X
X = reshape(X,numCols,numRows,numImages);
X = permute(X,[2 1 3]);
%%
% Then flatten each image data into a 1 by (numRows*numCols) vector, and
% store all the image data into a numImages by (numRows*numCols) array.
X = reshape(X,numRows*numCols,numImages)';
fprintf(['The image data is read to a matrix of dimensions: %6d by %4d...\n',...
    'End of reading image data.\n'],size(X,1),size(X,2));
%%
% Close the file
fclose(fileID);
%%
% Similarly, read the label data.
[fileID,errmsg] = fopen(labelFileName,'r','b');
if fileID < 0
    error(errmsg);
end
magicNum = fread(fileID,1,'int32',0,'b');
if magicNum == 2049
    fprintf('\nRead MNIST label data...\n')
end
numItems = fread(fileID,1,'int32',0,'b');
fprintf('Number of labels in the dataset: %6d ...\n',numItems);

L = fread(fileID,inf,'unsigned char');
fprintf(['The label data is read to a matrix of dimensions: %6d by %2d...\n',...
    'End of reading label data.\n'],size(L,1),size(L,2));
fclose(fileID);
```

References

[1] Yann LeCun (Courant Institute, NYU) and Corinna Cortes (Google Labs, New York) hold the copyright of MNIST dataset, which is a derivative work from original NIST datasets. MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license, https://creativecommons.org/licenses/by-sa/3.0/

See Also

`ReconstructionICA` | `SparseFiltering` | `rica` | `sparsefilt`

Related Examples

- “Extract Mixed Signals” on page 15-164

More About

- “Feature Extraction” on page 15-130

Extract Mixed Signals

This example shows how to use `rica` to disentangle mixed audio signals. You can use `rica` to perform independent component analysis (ICA) when prewhitening is included as a preprocessing step. The ICA model is

$$x = \mu + As.$$

Here, x is a p-by-1 vector of mixed signals, μ is a p-by-1 vector of offset values, A is a p-by-q mixing matrix, and s is a q-by-1 vector of original signals. Suppose first that A is a square matrix. If you know μ and A, you can recover an original signal s from the data x:

$$s = A^{-1}(x - \mu).$$

Using the `rica` function, you can perform this recovery even without knowing the mixing matrix A or the mean μ. Given a set of several observations $x(1)$, $x(2)$, ..., `rica` extracts the original signals $s(1)$, $s(2)$, ...

Load Data

Load a set of six audio files, which ship with MATLAB®. Trim each file to 10,000 samples.

```
files = {'chirp.mat'
         'gong.mat'
         'handel.mat'
         'laughter.mat'
         'splat.mat'
         'train.mat'};

S = zeros(10000,6);
for i = 1:6
    test     = load(files{i});
    y        = test.y(1:10000,1);
    S(:,i)   = y;
end
```

Mix Signals

Mix the signals together by using a random mixing matrix and add a random offset.

```
rng default % For reproducibility
mixdata = S*randn(6) + randn(1,6);
```

To listen to the original sounds, execute this code:

```
for i = 1:6
    disp(i);
    sound(S(:,i));
    pause;
end
```

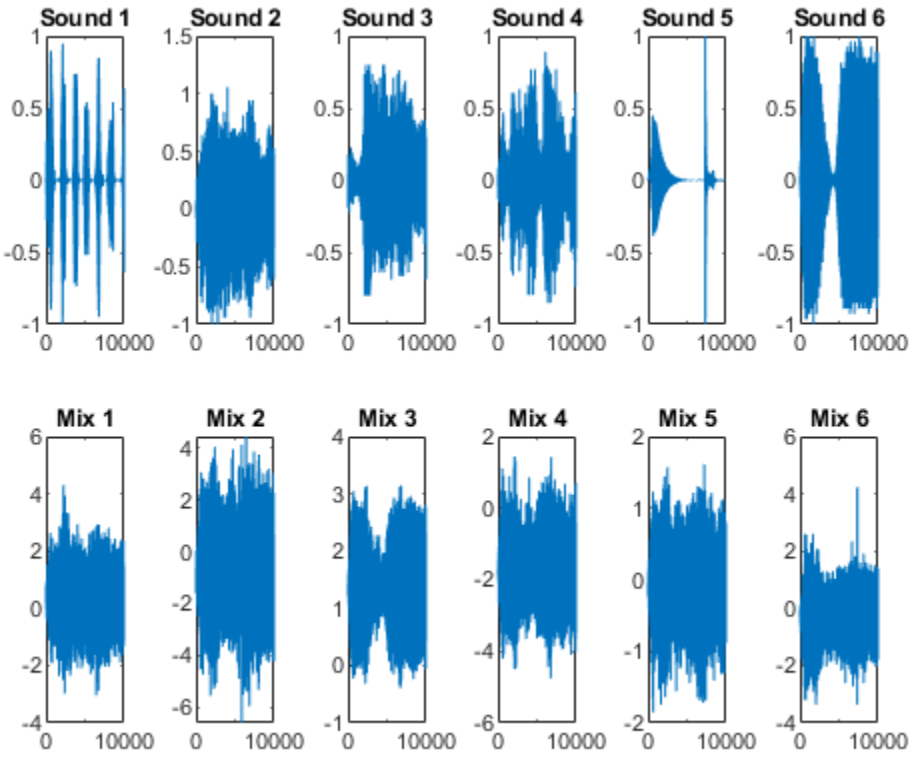
To listen to the mixed sounds, execute this code:

```
for i = 1:6
    disp(i);
```

```
        sound(mixdata(:,i));
        pause;
    end
```

Plot the signals.

```
figure
for i = 1:6
    subplot(2,6,i)
    plot(S(:,i))
    title(['Sound ',num2str(i)])
    subplot(2,6,i+6)
    plot(mixdata(:,i))
    title(['Mix ',num2str(i)])
end
```



The original signals have clear structure. The mixed signals have much less structure.

Prewhiten Mixed Signals

To separate the signals effectively, "prewhiten" the signals by using the `prewhiten` function that appears at the end of this example. This function transforms `mixdata` so that it has zero mean and identity covariance.

The idea is the following. If s is a zero-mean source with statistically independent components, then

$$E(s) = 0$$

$$E(ss^T) = I.$$

Then the mean and covariance of x are

$$E(x) = \mu$$

$$\mathrm{Cov}(x) = AA^T = C.$$

Suppose that you know μ and C. In practice, you would estimate these quantities from the sample mean and covariance of the columns of x. You can solve for s in terms of x by

$$s = A^{-1}(x - \mu) = (A^T A)^{-1} A^T (x - \mu).$$

The latter equation holds even when A is not a square invertible matrix.

Suppose that U is a p-by-q matrix of left eigenvectors of the positive semidefinite matrix C, and Σ is the q-by-q matrix of eigenvalues. Then

$$C = U\Sigma U^T$$

$$U^T U = I.$$

Then

$$AA^T = U\Sigma U^T.$$

There are many mixing matrices A that satisfy this last equation. If W is a q-by-q orthonormal matrix, then

$$W^T W = WW^T = I$$

$$A = U\Sigma^{1/2} W.$$

Substituting into the equation for s,

$$s = W^T \tilde{x}, \text{ where}$$

$$\tilde{x} = \Sigma^{-1/2} U^T (x - \mu).$$

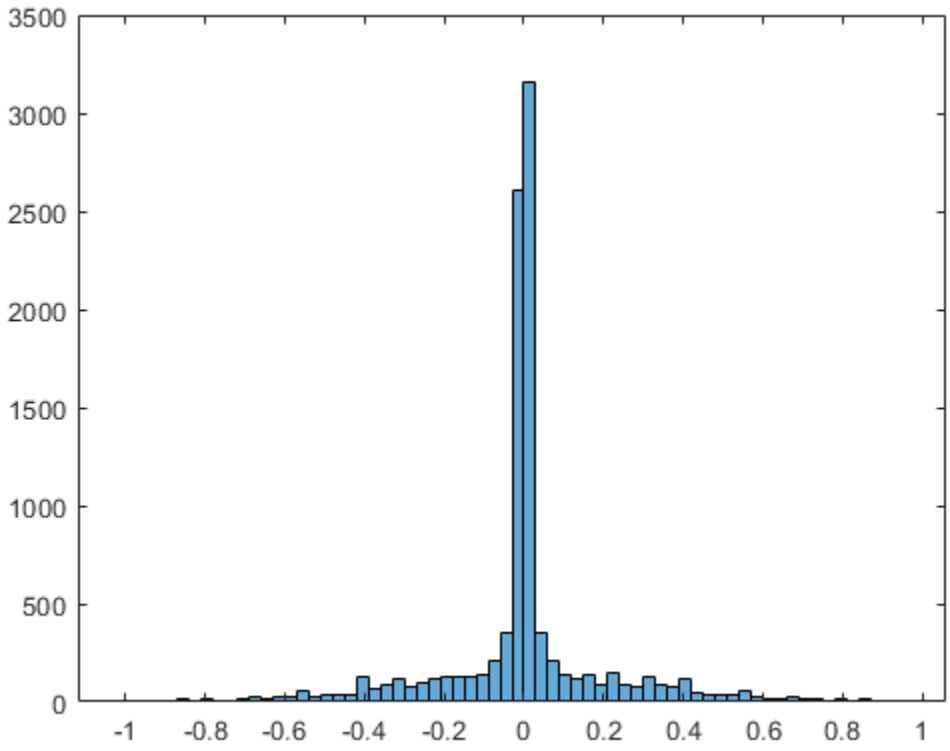
$\tilde{x}$ is the prewhitened data. `rica` computes the unknown matrix W under the assumption that the components of s are as independent as possible.

```
mixdata = prewhiten(mixdata);
```

Separate All Signals

A super-Gaussian source has a sharp peak near zero, such as a histogram of sound 1 shows.

```
figure
histogram(S(:,1))
```



Perform Reconstruction ICA while asking for six features. Indicate that each source is super-Gaussian.

```
q = 6;
Mdl = rica(mixdata,q,'NonGaussianityIndicator',ones(6,1));
```

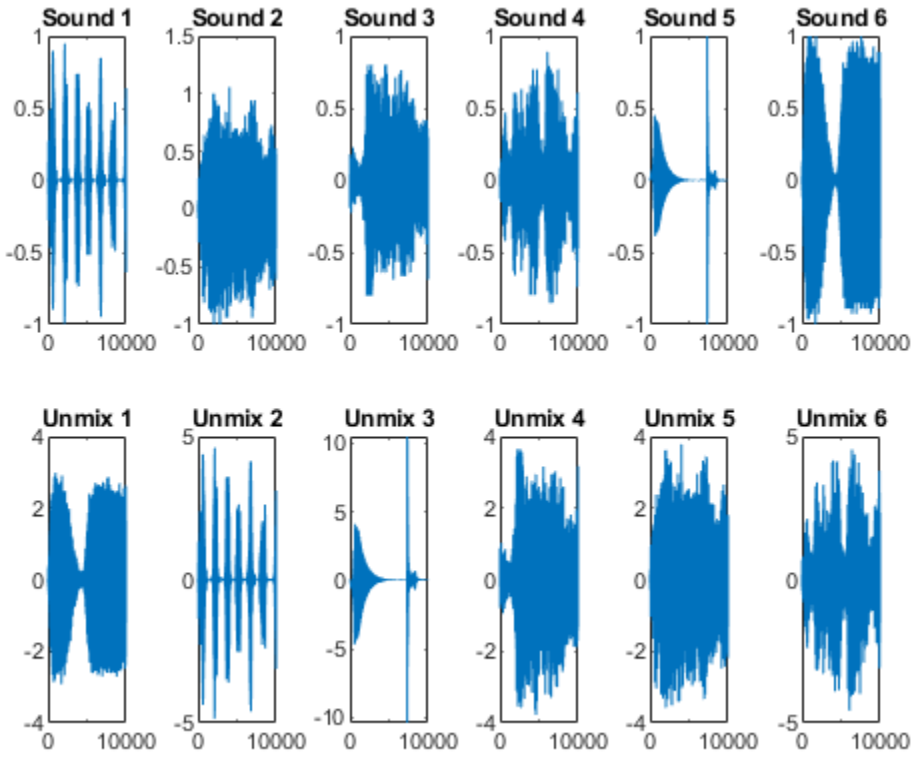
Extract the features. If the unmixing procedure is successful, the features are proportional to the original signals.

```
unmixed = transform(Mdl,mixdata);
```

Compare Unmixed Signals To Original Signals

Plot the original and unmixed signals.

```
figure
for i = 1:6
    subplot(2,6,i)
    plot(S(:,i))
    title(['Sound ',num2str(i)])
    subplot(2,6,i+6)
    plot(unmixed(:,i))
    title(['Unmix ',num2str(i)])
end
```

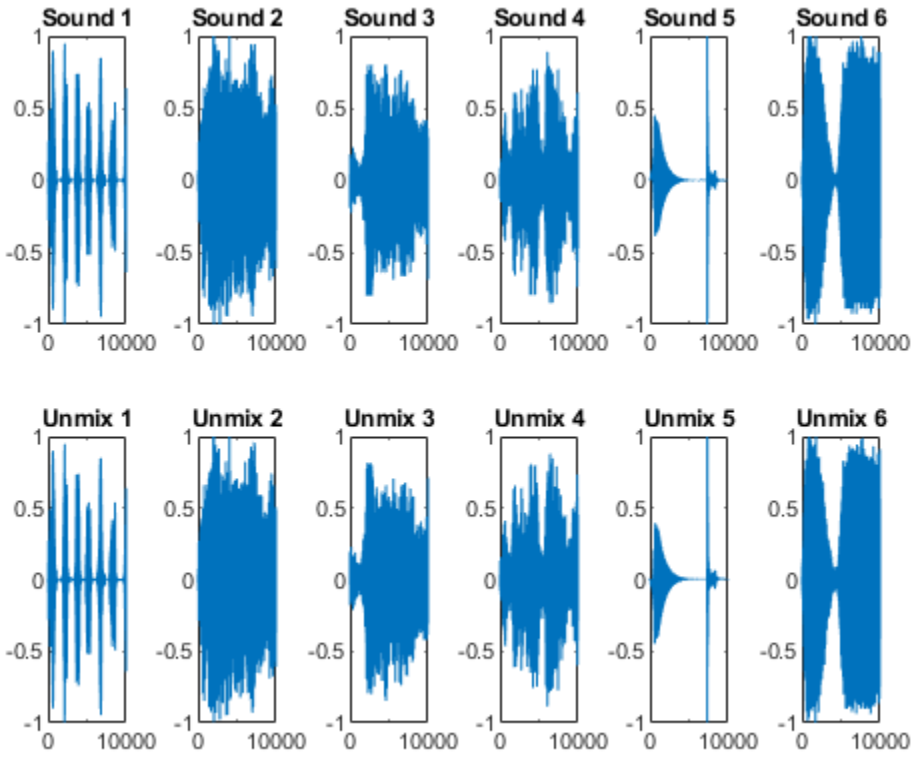


The order of the unmixed signals is different than the original order. Reorder the columns so that the unmixed signals match the corresponding original signals. Scale the unmixed signals to have the same norms as the corresponding original signals. (`rica` cannot identify the scale of the original signals because any scale can lead to the same signal mixture.)

```
unmixed = unmixed(:,[2,5,4,6,3,1]);
for i = 1:6
    unmixed(:,i) = unmixed(:,i)/norm(unmixed(:,i))*norm(S(:,i));
end
```

Plot the original and unmixed signals.

```
figure
for i = 1:6
    subplot(2,6,i)
    plot(S(:,i))
    ylim([-1,1])
    title(['Sound ',num2str(i)])
    subplot(2,6,i+6)
    plot(unmixed(:,i))
    ylim([-1,1])
    title(['Unmix ',num2str(i)])
end
```



The unmixed signals look similar to the original signals. To listen to the unmixed sounds, execute this code.

```
for i = 1:6
    disp(i);
    sound(unmixed(:,i));
    pause;
end
```

Here is the code for the `prewhiten` function.

```
function Z = prewhiten(X)
% X = N-by-P matrix for N observations and P predictors
% Z = N-by-P prewhitened matrix

    % 1. Size of X.
    [N,P] = size(X);
    assert(N >= P);

    % 2. SVD of covariance of X. We could also use svd(X) to proceed but N
    % can be large and so we sacrifice some accuracy for speed.
    [U,Sig] = svd(cov(X));
    Sig     = diag(Sig);
    Sig     = Sig(:)';

    % 3. Figure out which values of Sig are non-zero.
    tol = eps(class(X));
```

```
    idx = (Sig > max(Sig)*tol);
    assert(~all(idx == 0));

    % 4. Get the non-zero elements of Sig and corresponding columns of U.
    Sig = Sig(idx);
    U   = U(:,idx);

    % 5. Compute prewhitened data.
    mu = mean(X,1);
    Z = bsxfun(@minus,X,mu);
    Z = bsxfun(@times,Z*U,1./sqrt(Sig));
end
```

See Also

`ReconstructionICA` | `SparseFiltering` | `rica` | `sparsefilt`

Related Examples

- “Feature Extraction Workflow” on page 15-135

More About

- “Feature Extraction” on page 15-130

Selecting Features for Classifying High-dimensional Data

This example shows how to select features for classifying high-dimensional data. More specifically, it shows how to perform sequential feature selection, which is one of the most popular feature selection algorithms. It also shows how to use holdout and cross-validation to evaluate the performance of the selected features.

Reducing the number of features (dimensionality) is important in statistical learning. For many data sets with a large number of features and a limited number of observations, such as bioinformatics data, usually many features are not useful for producing a desired learning result and the limited observations may lead the learning algorithm to overfit to the noise. Reducing features can also save storage and computation time and increase comprehensibility.

There are two main approaches to reducing features: feature selection and feature transformation. Feature selection algorithms select a subset of features from the original feature set; feature transformation methods transform data from the original high-dimensional feature space to a new space with reduced dimensionality.

Loading the Data

Serum proteomic pattern diagnostics can be used to differentiate observations from patients with and without disease. Profile patterns are generated using surface-enhanced laser desorption and ionization (SELDI) protein mass spectrometry. These features are ion intensity levels at specific mass/charge values.

This example uses the high-resolution ovarian cancer data set that was generated using the WCX2 protein array. After some pre-processing steps, similar to those shown in the Bioinformatics Toolbox™ example Pre-processing Raw Mass Spectrometry Data, the data set has two variables `obs` and `grp`. The `obs` variable consists 216 observations with 4000 features. Each element in `grp` defines the group to which the corresponding row of `obs` belongs.

```
load ovariancancer;
whos
```

```
  Name        Size                Bytes  Class     Attributes

  grp       216x1                 25056  cell
  obs       216x4000            3456000  single
```

Dividing Data Into a Training Set and a Test Set

Some of the functions used in this example call MATLAB® built-in random number generation functions. To duplicate the exact results shown in this example, execute the command below to set the random number generator to a known state. Otherwise, your results may differ.

```
rng(8000,'twister');
```

The performance on the training data (resubstitution performance) is not a good estimate for a model's performance on an independent test set. Resubstitution performance will usually be over-optimistic. To predict the performance of a selected model, you need to assess its performance on another data set that was not used to build the model. Here, we use `cvpartition` to divide data into a training set of size 160 and a test set of size 56. Both the training set and the test set have roughly the same group proportions as in `grp`. We select features using the training data and judge the performance of the selected features on the test data. This is often called holdout validation. Another

simple and widely-used method for evaluating and selecting a model is cross-validation, which will be illustrated later in this example.

```
holdoutCVP = cvpartition(grp,'holdout',56)
```

```
holdoutCVP =
Hold-out cross validation partition
   NumObservations: 216
       NumTestSets: 1
         TrainSize: 160
          TestSize: 56
```

```
dataTrain = obs(holdoutCVP.training,:);
grpTrain = grp(holdoutCVP.training);
```

The Problem of Classifying Data Using All the Features

Without first reducing the number of features, some classification algorithms would fail on the data set used in this example, since the number of features is much larger than the number of observations. In this example, we use Quadratic Discriminant Analysis (QDA) as the classification algorithm. If we apply QDA on the data using all the features, as shown in the following, we will get an error because there are not enough samples in each group to estimate a covariance matrix.

```
try
   yhat = classify(obs(test(holdoutCVP),:), dataTrain, grpTrain,'quadratic');
catch ME
   display(ME.message);
end
```

```
The covariance matrix of each group in TRAINING must be positive definite.
```

Selecting Features Using a Simple Filter Approach

Our goal is to reduce the dimension of the data by finding a small set of important features which can give good classification performance. Feature selection algorithms can be roughly grouped into two categories: filter methods and wrapper methods. Filter methods rely on general characteristics of the data to evaluate and to select the feature subsets without involving the chosen learning algorithm (QDA in this example). Wrapper methods use the performance of the chosen learning algorithm to evaluate each candidate feature subset. Wrapper methods search for features better fit for the chosen learning algorithm, but they can be significantly slower than filter methods if the learning algorithm takes a long time to run. The concepts of "filters" and "wrappers" are described in John G. Kohavi R. (1997) "Wrappers for feature subset selection", Artificial Intelligence, Vol.97, No.1-2, pp.272-324. This example shows one instance of a filter method and one instance of a wrapper method.

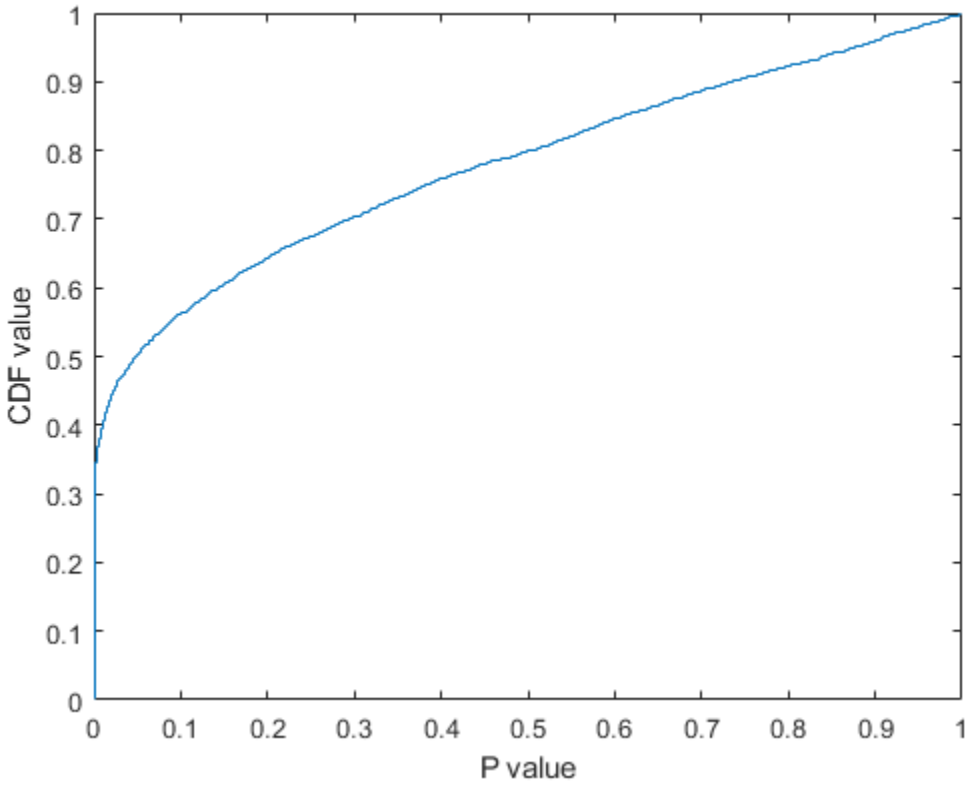
Filters are usually used as a pre-processing step since they are simple and fast. A widely-used filter method for bioinformatics data is to apply a univariate criterion separately on each feature, assuming that there is no interaction between features.

For example, we might apply the *t*-test on each feature and compare *p*-value (or the absolute values of *t*-statistics) for each feature as a measure of how effective it is at separating groups.

```
dataTrainG1 = dataTrain(grp2idx(grpTrain)==1,:);
dataTrainG2 = dataTrain(grp2idx(grpTrain)==2,:);
[h,p,ci,stat] = ttest2(dataTrainG1,dataTrainG2,'Vartype','unequal');
```

In order to get a general idea of how well-separated the two groups are by each feature, we plot the empirical cumulative distribution function (CDF) of the *p*-values:

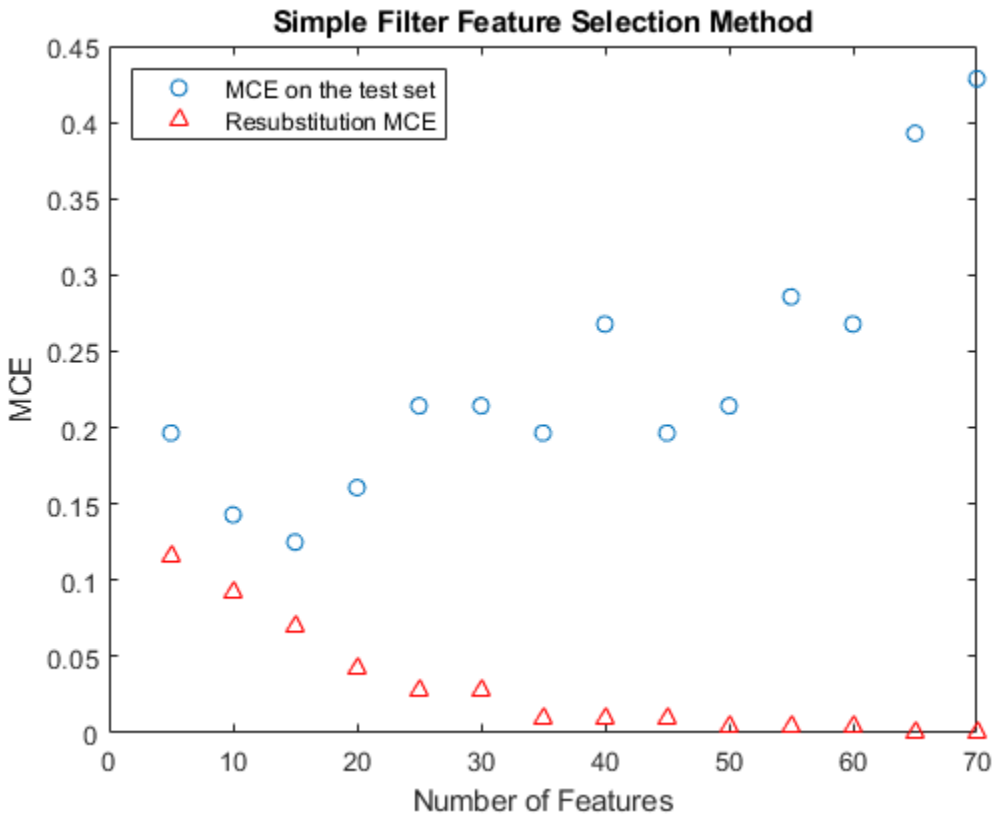
```
ecdf(p);
xlabel('P value');
ylabel('CDF value')
```



There are about 35% of features having *p*-values close to zero and over 50% of features having *p*-values smaller than 0.05, meaning there are more than 2500 features among the original 5000 features that have strong discrimination power. One can sort these features according to their *p*-values (or the absolute values of the *t*-statistic) and select some features from the sorted list. However, it is usually difficult to decide how many features are needed unless one has some domain knowledge or the maximum number of features that can be considered has been dictated in advance based on outside constraints.

One quick way to decide the number of needed features is to plot the MCE (misclassification error, i.e., the number of misclassified observations divided by the number of observations) on the test set as a function of the number of features. Since there are only 160 observations in the training set, the largest number of features for applying QDA is limited, otherwise, there may not be enough samples in each group to estimate a covariance matrix. Actually, for the data used in this example, the holdout partition and the sizes of two groups dictate that the largest allowable number of features for applying QDA is about 70. Now we compute MCE for various numbers of features between 5 and 70 and show the plot of MCE as a function of the number of features. In order to reasonably estimate the performance of the selected model, it is important to use the 160 training samples to fit the QDA model and compute the MCE on the 56 test observations (blue circular marks in the following plot). To illustrate why resubstitution error is not a good error estimate of the test error, we also show the resubstitution MCE using red triangular marks.

```
[~,featureIdxSortbyP] = sort(p,2); % sort the features
testMCE = zeros(1,14);
resubMCE = zeros(1,14);
nfs = 5:5:70;
classf = @(xtrain,ytrain,xtest,ytest) ...
             sum(~strcmp(ytest,classify(xtest,xtrain,ytrain,'quadratic')));
resubCVP = cvpartition(length(grp),'resubstitution')

resubCVP =
Resubstitution (no partition of data)
   NumObservations: 216
       NumTestSets: 1
         TrainSize: 216
          TestSize: 216

for i = 1:14
   fs = featureIdxSortbyP(1:nfs(i));
   testMCE(i) = crossval(classf,obs(:,fs),grp,'partition',holdoutCVP)...
       /holdoutCVP.TestSize;
   resubMCE(i) = crossval(classf,obs(:,fs),grp,'partition',resubCVP)/...
       resubCVP.TestSize;
end
 plot(nfs, testMCE,'o',nfs,resubMCE,'r^');
 xlabel('Number of Features');
 ylabel('MCE');
 legend({'MCE on the test set' 'Resubstitution MCE'},'location','NW');
 title('Simple Filter Feature Selection Method');
```



For convenience, `classf` is defined as an anonymous function. It fits QDA on the given training set and returns the number of misclassified samples for the given test set. If you were developing your own classification algorithm, you might want to put it in a separate file, as follows:

```
%  function err = classf(xtrain,ytrain,xtest,ytest)
%       yfit = classify(xtest,xtrain,ytrain,'quadratic');
%        err = sum(~strcmp(ytest,yfit));
```

The resubstitution MCE is over-optimistic. It consistently decreases when more features are used and drops to zero when more than 60 features are used. However, if the test error increases while the resubstitution error still decreases, then overfitting may have occurred. This simple filter feature selection method gets the smallest MCE on the test set when 15 features are used. The plot shows overfitting begins to occur when 20 or more features are used. The smallest MCE on the test set is 12.5%:

```
testMCE(3)
```

```
ans = 0.1250
```

These are the first 15 features that achieve the minimum MCE:

```
featureIdxSortbyP(1:15)
```

```
ans = 1×15

        2814        2813        2721        2720        2452        2645        2644        2642
```

Applying Sequential Feature Selection

The above feature selection algorithm does not consider interaction between features; besides, features selected from the list based on their individual ranking may also contain redundant information, so that not all the features are needed. For example, the linear correlation coefficient between the first selected feature (column 2814) and the second selected feature (column 2813) is almost 0.95.

```
corr(dataTrain(:,featureIdxSortbyP(1)),dataTrain(:,featureIdxSortbyP(2)))
```

```
ans = single
    0.9447
```

This kind of simple feature selection procedure is usually used as a pre-processing step since it is fast. More advanced feature selection algorithms improve the performance. Sequential feature selection is one of the most widely used techniques. It selects a subset of features by sequentially adding (forward search) or removing (backward search) until certain stopping conditions are satisfied.

In this example, we use forward sequential feature selection in a wrapper fashion to find important features. More specifically, since the typical goal of classification is to minimize the MCE, the feature selection procedure performs a sequential search using the MCE of the learning algorithm QDA on each candidate feature subset as the performance indicator for that subset. The training set is used to select the features and to fit the QDA model, and the test set is used to evaluate the performance of the finally selected feature. During the feature selection procedure, to evaluate and to compare the performance of the each candidate feature subset, we apply stratified 10-fold cross-validation to the training set. We will illustrate later why applying cross-validation to the training set is important.

First we generate a stratified 10-fold partition for the training set:

```
tenfoldCVP = cvpartition(grpTrain,'kfold',10)
```

```
tenfoldCVP =
K-fold cross validation partition
   NumObservations: 160
       NumTestSets: 10
         TrainSize: 144  144  144  144  144  144  144  144  144  144
          TestSize: 16  16  16  16  16  16  16  16  16  16
```

Then we use the filter results from the previous section as a pre-processing step to select features. For instance, we select 150 features here:

```
fs1 = featureIdxSortbyP(1:150);
```

We apply forward sequential feature selection on these 150 features. The function `sequentialfs` provides a simple way (the default option) to decide how many features are needed. It stops when the first local minimum of the cross-validation MCE is found.

```
 fsLocal = sequentialfs(classf,dataTrain(:,fs1),grpTrain,'cv',tenfoldCVP);
```

The selected features are the following:

```
fs1(fsLocal)
```

```
ans = 1×3

        2337         864        3288
```

To evaluate the performance of the selected model with these three features, we compute the MCE on the 56 test samples.

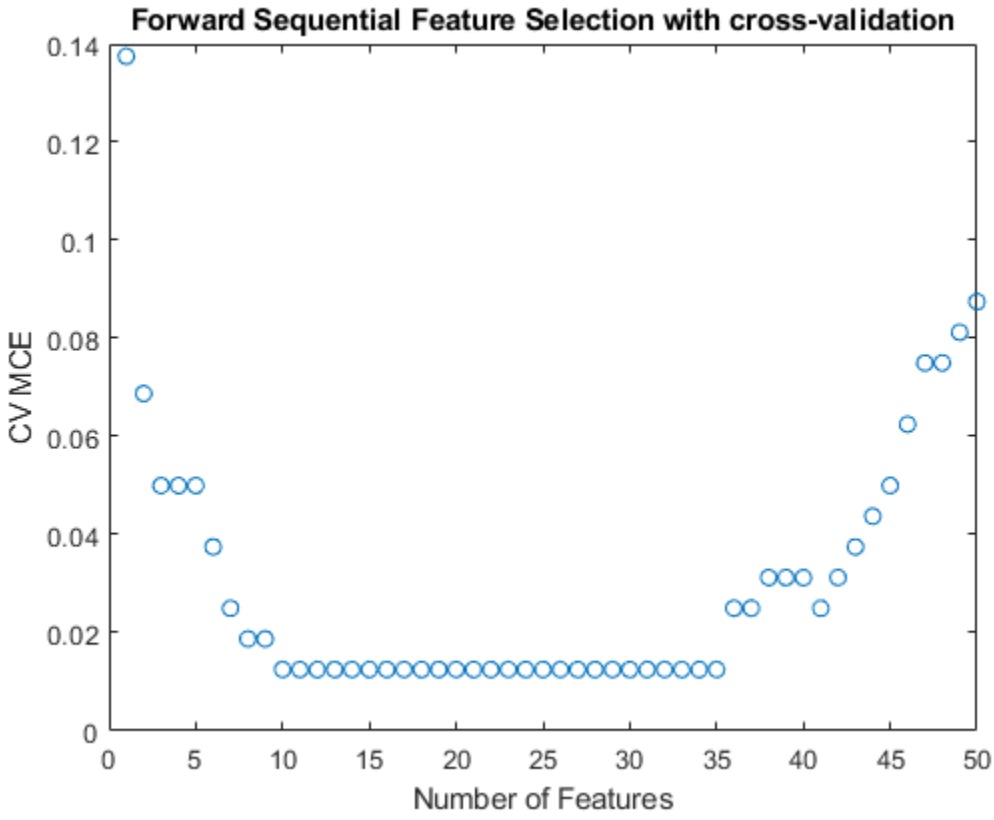
```
testMCELocal = crossval(classf,obs(:,fs1(fsLocal)),grp,'partition',...
    holdoutCVP)/holdoutCVP.TestSize
```

```
testMCELocal = 0.0714
```

With only three features being selected, the MCE is only a little over half of the smallest MCE using the simple filter feature selection method.

The algorithm may have stopped prematurely. Sometimes a smaller MCE is achievable by looking for the minimum of the cross-validation MCE over a reasonable range of number of features. For instance, we draw the plot of the cross-validation MCE as a function of the number of features for up to 50 features.

```
[fsCVfor50,historyCV] = sequentialfs(classf,dataTrain(:,fs1),grpTrain,...
    'cv',tenfoldCVP,'Nf',50);
plot(historyCV.Crit,'o');
xlabel('Number of Features');
ylabel('CV MCE');
title('Forward Sequential Feature Selection with cross-validation');
```



The cross-validation MCE reaches the minimum value when 10 features are used and this curve stays flat over the range from 10 features to 35 features. Also, the curve goes up when more than 35 features are used, which means overfitting occurs there.

It is usually preferable to have fewer features, so here we pick 10 features:

```
fsCVfor10 = fs1(historyCV.In(10,:))
```

```
fsCVfor10 = 1×10

        2814        2721        2720        2452        2650        2731        2337        2658
```

To show these 10 features in the order in which they are selected in the sequential forward procedure, we find the row in which they first become true in the `historyCV` output:

```
[orderlist,ignore] = find( [historyCV.In(1,:); diff(historyCV.In(1:10,:) )]' );
fs1(orderlist)
```

```
ans = 1×10

        2337         864        3288        2721        2814        2658        2452        2731
```

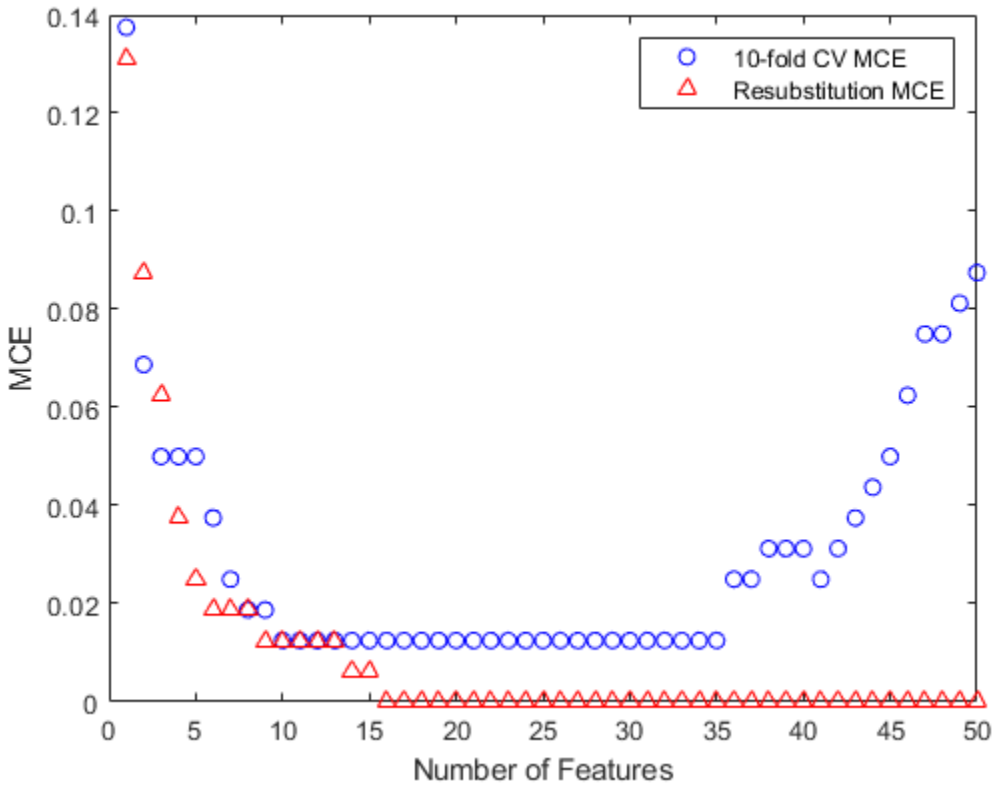
To evaluate these 10 features, we compute their MCE for QDA on the test set. We get the smallest MCE value so far:

```
testMCECVfor10 = crossval(classf,obs(:,fsCVfor10),grp,'partition',...
    holdoutCVP)/holdoutCVP.TestSize
```

```
testMCECVfor10 = 0.0357
```

It is interesting to look at the plot of resubstitution MCE values on the training set (i.e., without performing cross-validation during the feature selection procedure) as a function of the number of features:

```
[fsResubfor50,historyResub] = sequentialfs(classf,dataTrain(:,fs1),...
     grpTrain,'cv','resubstitution','Nf',50);
plot(1:50, historyCV.Crit,'bo',1:50, historyResub.Crit,'r^');
xlabel('Number of Features');
ylabel('MCE');
legend({'10-fold CV MCE' 'Resubstitution MCE'},'location','NE');
```



Again, the resubstitution MCE values are overly optimistic here. Most are smaller than the cross-validation MCE values, and the resubstitution MCE goes to zero when 16 features are used. We can compute the MCE value of these 16 features on the test set to see their real performance:

```
fsResubfor16 = fs1(historyResub.In(16,:));
testMCEResubfor16 = crossval(classf,obs(:,fsResubfor16),grp,'partition',...
    holdoutCVP)/holdoutCVP.TestSize
```

```
testMCEResubfor16 = 0.0714
```

`testMCEResubfor16`, the performance of these 16 features (chosen by resubstitution during the feature selection procedure) on the test set, is about double that for `testMCECVfor10`, the

performance of the 10 features (chosen by 10-fold cross-validation during the feature selection procedure) on the test set. It again indicates that the resubstitution error generally is not a good performance estimate for evaluating and selecting features. We may want to avoid using resubstitution error, not only during the final evaluation step, but also during the feature selection procedure.

See Also

`sequentialfs`

More About

- “Introduction to Feature Selection” on page 15-49
- “Sequential Feature Selection” on page 15-61

Perform Factor Analysis on Exam Grades

This example shows how to perform factor analysis using Statistics and Machine Learning Toolbox™.

Multivariate data often include a large number of measured variables, and sometimes those variables "overlap" in the sense that groups of them may be dependent. For example, in a decathlon, each athlete competes in 10 events, but several of them can be thought of as "speed" events, while others can be thought of as "strength" events, etc. Thus, a competitor's 10 event scores might be thought of as largely dependent on a smaller set of 3 or 4 types of athletic ability.

Factor analysis is a way to fit a model to multivariate data to estimate just this sort of interdependence.

The Factor Analysis Model

In the factor analysis model, the measured variables depend on a smaller number of unobserved (latent) factors. Because each factor may affect several variables in common, they are known as "common factors". Each variable is assumed to depend on a linear combination of the common factors, and the coefficients are known as loadings. Each measured variable also includes a component due to independent random variability, known as "specific variance" because it is specific to one variable.

Specifically, factor analysis assumes that the covariance matrix of your data is of the form

```
SigmaX = Lambda*Lambda' + Psi
```

where Lambda is the matrix of loadings, and the elements of the diagonal matrix Psi are the specific variances. The function `factoran` fits the factor analysis model using maximum likelihood.

Example: Finding Common Factors Affecting Exam Grades

120 students have each taken five exams, the first two covering mathematics, the next two on literature, and a comprehensive fifth exam. It seems reasonable that the five grades for a given student ought to be related. Some students are good at both subjects, some are good at only one, etc. The goal of this analysis is to determine if there is quantitative evidence that the students' grades on the five different exams are largely determined by only two types of ability.

First load the data, then call `factoran` and request a model fit with a single common factor.

```
load examgrades
[Loadings1,specVar1,T,stats] = factoran(grades,1);
```

`factoran`'s first two return arguments are the estimated loadings and the estimated specific variances. From the estimated loadings, you can see that the one common factor in this model puts large positive weight on all five variables, but most weight on the fifth, comprehensive exam.

```
Loadings1
```

```
Loadings1 =

    0.6021
    0.6686
    0.7704
    0.7204
```

```
    0.9153
```

One interpretation of this fit is that a student might be thought of in terms of their "overall ability", for which the comprehensive exam would be the best available measurement. A student's grade on a more subject-specific test would depend on their overall ability, but also on whether or not the student was strong in that area. This would explain the lower loadings for the first four exams.

From the estimated specific variances, you can see that the model indicates that a particular student's grade on a particular test varies quite a lot beyond the variation due to the common factor.

```
specVar1
```

```
specVar1 =

    0.6375
    0.5530
    0.4065
    0.4810
    0.1623
```

A specific variance of 1 would indicate that there is *no* common factor component in that variable, while a specific variance of 0 would indicate that the variable is *entirely* determined by common factors. These exam grades seem to fall somewhere in between, although there is the least amount of specific variation for the comprehensive exam. This is consistent with the interpretation given above of the single common factor in this model.

The p-value returned in the `stats` structure rejects the null hypothesis of a single common factor, so we refit the model.

```
stats.p
```

```
ans =

    0.0332
```

Next, use two common factors to try and better explain the exam scores. With more than one factor, you could rotate the estimated loadings to try and make their interpretation simpler, but for the moment, ask for an unrotated solution.

```
[Loadings2,specVar2,T,stats] = factoran(grades,2,'rotate','none');
```

From the estimated loadings, you can see that the first unrotated factor puts approximately equal weight on all five variables, while the second factor contrasts the first two variables with the second two.

```
Loadings2
```

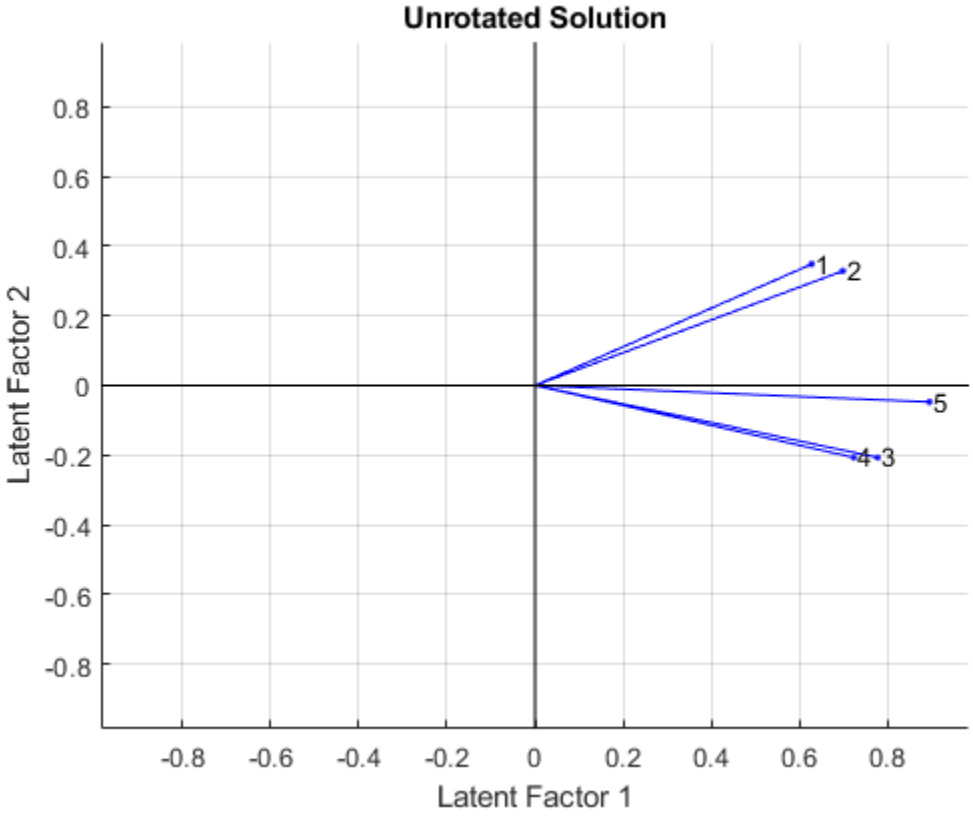
```
Loadings2 =

    0.6289    0.3485
    0.6992    0.3287
    0.7785   -0.2069
```

```
    0.7246   -0.2070
    0.8963   -0.0473
```

You might interpret these factors as "overall ability" and "quantitative vs. qualitative ability", extending the interpretation of the one-factor fit made earlier.

A plot of the variables, where each loading is a coordinate along the corresponding factor's axis, illustrates this interpretation graphically. The first two exams have a positive loading on the second factor, suggesting that they depend on "quantitative" ability, while the second two exams apparently depend on the opposite. The fifth exam has only a small loading on this second factor.

```
biplot(Loadings2, 'varlabels',num2str((1:5)'));
title('Unrotated Solution');
xlabel('Latent Factor 1'); ylabel('Latent Factor 2');
```



From the estimated specific variances, you can see that this two-factor model indicates somewhat less variation beyond that due to the common factors than the one-factor model did. Again, the least amount of specific variance occurs for the fifth exam.

```
specVar2
```

```
specVar2 =

    0.4829
    0.4031
```

```
    0.3512
    0.4321
    0.1944
```

The `stats` structure shows that there is only a single degree of freedom in this two-factor model.

```
stats.dfe
```

```
ans =

     1
```

With only five measured variables, you cannot fit a model with more than two factors.

Factor Analysis from a Covariance/Correlation Matrix

You made the fits above using the raw test scores, but sometimes you might only have a sample covariance matrix that summarizes your data. `factoran` accepts either a covariance or correlation matrix, using the `'Xtype'` parameter, and gives an identical result to that from the raw data.

```
Sigma = cov(grades);
[LoadingsCov,specVarCov] = ...
        factoran(Sigma,2,'Xtype','cov','rotate','none');
LoadingsCov
```

```
LoadingsCov =

    0.6289    0.3485
    0.6992    0.3287
    0.7785   -0.2069
    0.7246   -0.2070
    0.8963   -0.0473
```

Factor Rotation

Sometimes, the estimated loadings from a factor analysis model can give a large weight on several factors for some of the measured variables, making it difficult to interpret what those factors represent. The goal of factor rotation is to find a solution for which each variable has only a small number of large loadings, i.e., is affected by a small number of factors, preferably only one.

If you think of each row of the loadings matrix as coordinates of a point in M-dimensional space, then each factor corresponds to a coordinate axis. Factor rotation is equivalent to rotating those axes, and computing new loadings in the rotated coordinate system. There are various ways to do this. Some methods leave the axes orthogonal, while others are oblique methods that change the angles between them.

Varimax is one common criterion for orthogonal rotation. `factoran` performs varimax rotation by default, so you do not need to ask for it explicitly.

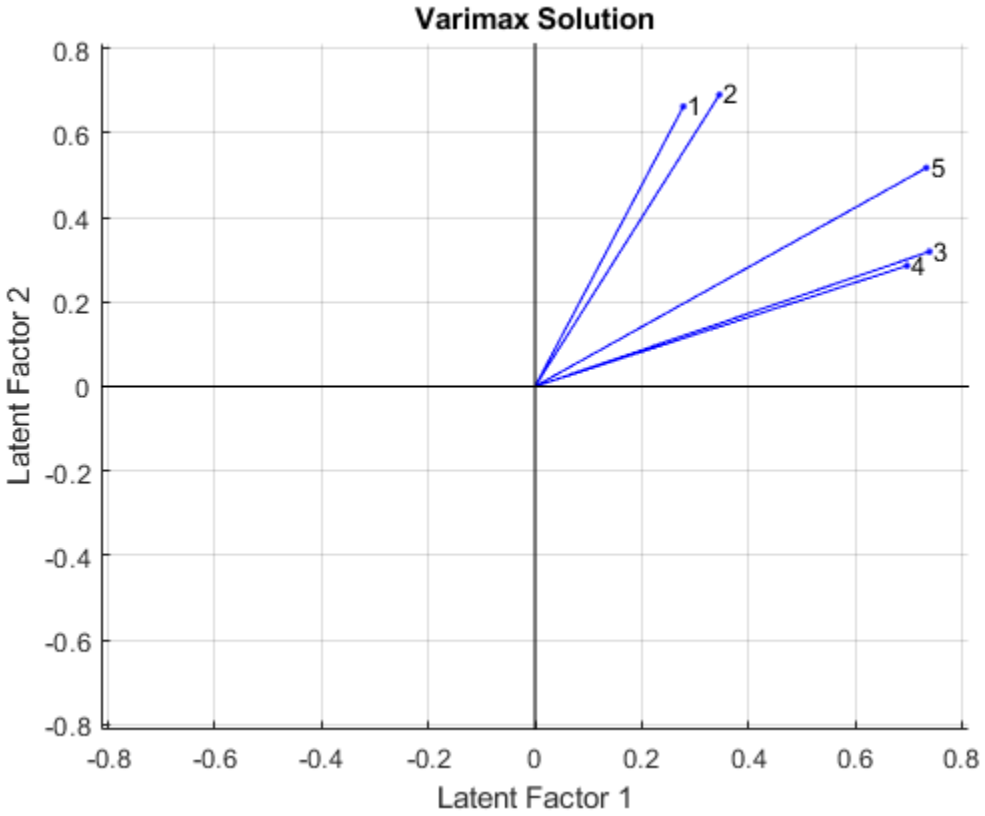
```
[LoadingsVM,specVarVM,rotationVM] = factoran(grades,2);
```

A quick check of the varimax rotation matrix returned by `factoran` confirms that it is orthogonal. Varimax, in effect, rotates the factor axes in the figure above, but keeps them at right angles.

```
rotationVM'*rotationVM


ans =

     1     0
     0     1
```

A biplot of the five variables on the rotated factors shows the effect of varimax rotation.

```
biplot(LoadingsVM, 'varlabels',num2str((1:5)'));
title('Varimax Solution');
xlabel('Latent Factor 1'); ylabel('Latent Factor 2');
```



Varimax has rigidly rotated the axes in an attempt to make all of the loadings close to zero or one. The first two exams are closest to the second factor axis, while the third and fourth are closest to the first axis and the fifth exam is at an intermediate position. These two rotated factors can probably be best interpreted as "quantitative ability" and "qualitative ability". However, because none of the variables are near a factor axis, the biplot shows that orthogonal rotation has not succeeded in providing a simple set of factors.

Because the orthogonal rotation was not entirely satisfactory, you can try using promax, a common oblique rotation criterion.

```
[LoadingsPM,specVarPM,rotationPM] = ...
                factoran(grades,2,'rotate','promax');
```

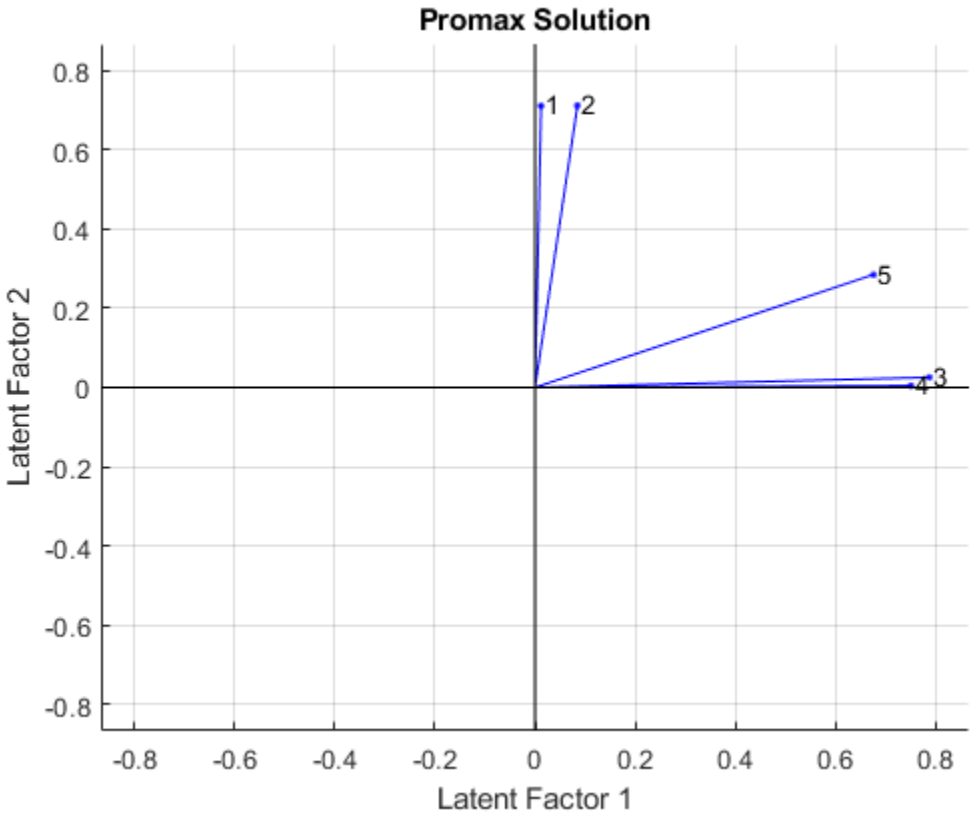
A check on the promax rotation matrix returned by `factoran` shows that it is not orthogonal. Promax, in effect, rotates the factor axes in the first figure separately, allowing them to have an oblique angle between them.

```
rotationPM'*rotationPM
```

```
ans =

    1.9405   -1.3509
   -1.3509    1.9405
```

A biplot of the variables on the new rotated factors shows the effect of promax rotation.

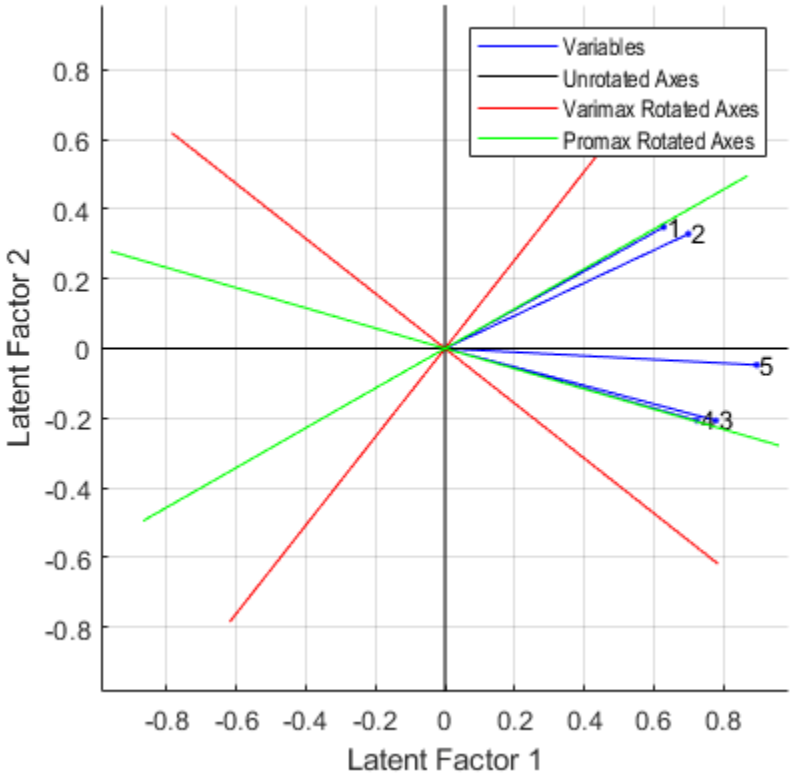
```
biplot(LoadingsPM, 'varlabels',num2str((1:5)'));
title('Promax Solution');
xlabel('Latent Factor 1'); ylabel('Latent Factor 2');
```



Promax has performed a non-rigid rotation of the axes, and has done a much better job than varimax at creating a "simple structure". The first two exams are close to the second factor axis, while the third and fourth are close to the first axis, and the fifth exam is in an intermediate position. This makes an interpretation of these rotated factors as "quantitative ability" and "qualitative ability" more precise.

Instead of plotting the variables on the different sets of rotated axes, it's possible to overlay the rotated axes on an unrotated biplot to get a better idea of how the rotated and unrotated solutions are related.

```
h1 = biplot(Loadings2, 'varlabels',num2str((1:5)'));
xlabel('Latent Factor 1'); ylabel('Latent Factor 2');
hold on
invRotVM = inv(rotationVM);
h2 = line([-invRotVM(1,1) invRotVM(1,1) NaN -invRotVM(2,1) invRotVM(2,1)], ...
          [-invRotVM(1,2) invRotVM(1,2) NaN -invRotVM(2,2) invRotVM(2,2)],'Color',[1 0 0]);
invRotPM = inv(rotationPM);
h3 = line([-invRotPM(1,1) invRotPM(1,1) NaN -invRotPM(2,1) invRotPM(2,1)], ...
          [-invRotPM(1,2) invRotPM(1,2) NaN -invRotPM(2,2) invRotPM(2,2)],'Color',[0 1 0]);
hold off
axis square
lgndHandles = [h1(1) h1(end) h2 h3];
lgndLabels = {'Variables','Unrotated Axes','Varimax Rotated Axes','Promax Rotated Axes'};
legend(lgndHandles, lgndLabels, 'location','northeast', 'fontname','arial narrow');
```



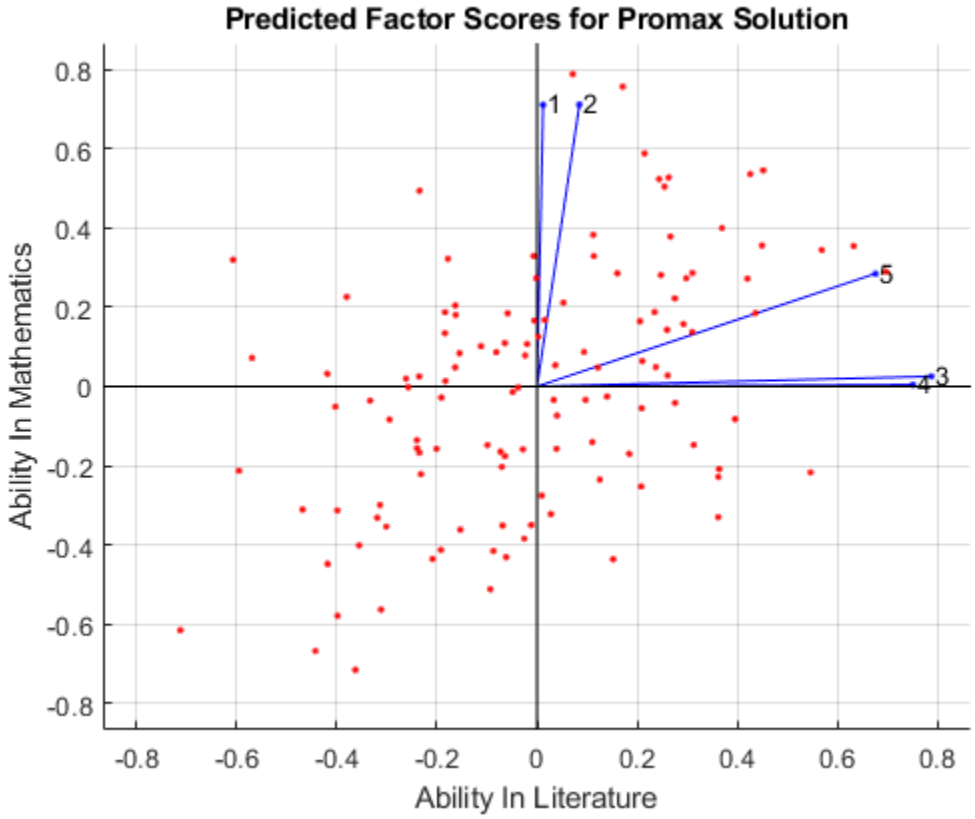
Predicting Factor Scores

Sometimes, it is useful to be able to classify an observation based on its factor scores. For example, if you accepted the two-factor model and the interpretation of the promax rotated factors, you might want to predict how well a student would do on a mathematics exam in the future.

Since the data are the raw exam grades, and not just their covariance matrix, we can have `factoran` return predictions of the value of each of the two rotated common factors for each student.

```
[Loadings,specVar,rotation,stats,preds] = ...
              factoran(grades,2,'rotate','promax','maxit',200);
biplot(Loadings, 'varlabels',num2str((1:5)'), 'Scores',preds);
```

```
title('Predicted Factor Scores for Promax Solution');
xlabel('Ability In Literature'); ylabel('Ability In Mathematics');
```



This plot shows the model fit in terms of both the original variables (vectors) and the predicted scores for each observation (points). The fit suggests that, while some students do well in one subject but not the other (second and fourth quadrants), most students do either well or poorly in both mathematics and literature (first and third quadrants). You can confirm this by looking at the estimated correlation matrix of the two factors.

```
inv(rotation'*rotation)
```

```
ans =

    1.0000    0.6962
    0.6962    1.0000
```

A Comparison of Factor Analysis and Principal Components Analysis

There is a good deal of overlap in terminology and goals between Principal Components Analysis (PCA) and Factor Analysis (FA). Much of the literature on the two methods does not distinguish between them, and some algorithms for fitting the FA model involve PCA. Both are dimension-reduction techniques, in the sense that they can be used to replace a large set of observed variables with a smaller set of new variables. They also often give similar results. However, the two methods are different in their goals and in their underlying models. Roughly speaking, you should use PCA when you simply need to summarize or approximate your data using fewer dimensions (to visualize it,

for example), and you should use FA when you need an explanatory model for the correlations among your data.

Classical Multidimensional Scaling Applied to Nonspatial Distances

This example shows how to perform classical multidimensional scaling using the `cmdscale` function in Statistics and Machine Learning Toolbox™. Classical multidimensional scaling, also known as Principal Coordinates Analysis, takes a matrix of interpoint distances, and creates a configuration of points. Ideally, those points can be constructed in two or three dimensions, and the Euclidean distances between them approximately reproduce the original distance matrix. Thus, a scatter plot of the those points provides a visual representation of the original distances.

This example illustrates applications of multidimensional scaling to dissimilarity measures other than spatial distance, and shows how to construct a configuration of points to visualize those dissimilarities.

This example describes classical multidimensional scaling. The `mdscale` function performs nonclassical MDS, which is sometimes more flexible than the classical method. Nonclassical MDS is described in the “Nonclassical Multidimensional Scaling” on page 15-197 example.

Reconstructing Spatial Locations from Nonspatial Distances

Suppose you have measured the genetic "distance", or dissimilarity, between a number of local subpopulations of a single species of animal. You also know their geographic locations, and would like to know how closely their genetic and spatial distances correspond. If they do, that is evidence that interbreeding between the subpopulations is affected by their geographic locations.

Below are the spatial locations of the subpopulations, and the upper-triangle of the matrix of genetic distances, in the same vector format produced by `pdist`.

```
X = [39.1     18.7;
     40.7     21.2;
     41.5     21.5;
     39.2     21.8;
     38.7     20.6;
     41.7     20.1;
     40.1     22.1;
     39.2     21.6];

D = [4.69 6.79 3.50 3.11 4.46 5.57 3.00 ...
          2.10 2.27 2.65 2.36 1.99 1.74 ...
               3.78 4.53 2.83 2.44 3.79 ...
                    1.98 4.35 2.07 0.53 ...
                         3.80 3.31 1.47 ...
                              4.35 3.82 ...
                                   2.57];
```

Although this vector format for `D` is space-efficient, it's often easier to see the distance relationships if you reformat the distances to a square matrix.

```
squareform(D)
```

```
ans = 8×8

         0    4.6900    6.7900    3.5000    3.1100    4.4600    5.5700    3.0000
    4.6900         0    2.1000    2.2700    2.6500    2.3600    1.9900    1.7400
    6.7900    2.1000         0    3.7800    4.5300    2.8300    2.4400    3.7900
```

```
    3.5000    2.2700    3.7800         0    1.9800    4.3500    2.0700    0.5300
    3.1100    2.6500    4.5300    1.9800         0    3.8000    3.3100    1.4700
    4.4600    2.3600    2.8300    4.3500    3.8000         0    4.3500    3.8200
    5.5700    1.9900    2.4400    2.0700    3.3100    4.3500         0    2.5700
    3.0000    1.7400    3.7900    0.5300    1.4700    3.8200    2.5700         0
```

`cmdscale` recognizes either of the two formats.

```
[Y,eigvals] = cmdscale(D);
```

`cmdscale`'s first output, `Y`, is a matrix of points created to have interpoint distances that reproduce the distances in `D`. With eight species, the points (rows of `Y`) could have as many as eight dimensions (columns of `Y`). Visualization of the genetic distances depends on using points in only two or three dimensions. Fortunately, `cmdscale`'s second output, `eigvals`, is a set of sorted eigenvalues whose relative magnitudes indicate how many dimensions you can safely use. If only the first two or three eigenvalues are large, then only those coordinates of the points in `Y` are needed to accurately reproduce `D`. If more than three eigenvalues are large, then it is not possible to find a good low-dimensional configuration of points, and it will not be easy to visualize the distances.

```
[eigvals eigvals/max(abs(eigvals))]
```

```
ans = 8×2

   29.0371    1.0000
   13.5746    0.4675
    2.0987    0.0723
    0.7418    0.0255
    0.3403    0.0117
    0.0000    0.0000
   -0.4542   -0.0156
   -3.1755   -0.1094
```

Notice that there are only two large positive eigenvalues, so the configuration of points created by `cmdscale` can be plotted in two dimensions. The two negative eigenvalues indicate that the genetic distances are not Euclidean, that is, no configuration of points can reproduce `D` exactly. Fortunately, the negative eigenvalues are small relative to the largest positive ones, and the reduction to the first two columns of `Y` should be fairly accurate. You can check this by looking at the error in the distances between the two-dimensional configuration and the original distances.

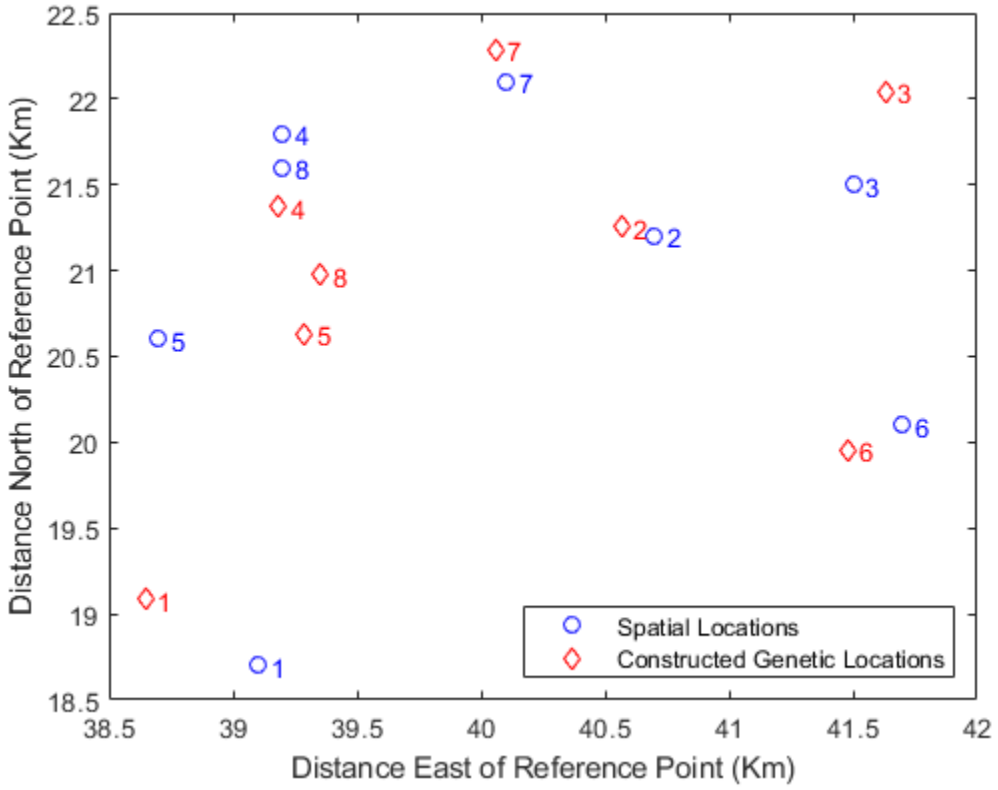
```
maxrelerr = max(abs(D - pdist(Y(:,1:2)))) / max(D)
```

```
maxrelerr = 0.1335
```

Now you can compare the "genetic locations" created by `cmdscale` to the actual geographic locations. Because the configuration returned by `cmdscale` is unique only up to translation, rotation, and reflection, the genetic locations probably won't match the geographic locations. They will also have the wrong scale. But you can use the `procrustes` command to match up the two sets of points best in the least squares sense.

```
[D,Z] = procrustes(X,Y(:,1:2));
plot(X(:,1),X(:,2),'bo',Z(:,1),Z(:,2),'rd');
labels = num2str((1:8)');
text(X(:,1)+.05,X(:,2),labels,'Color','b');
text(Z(:,1)+.05,Z(:,2),labels,'Color','r');
xlabel('Distance East of Reference Point (Km)');
```

```
ylabel('Distance North of Reference Point (Km)');
legend({'Spatial Locations','Constructed Genetic Locations'},'Location','SE');
```



This plot shows the best match of the reconstructed points in the same coordinates as the actual spatial locations. Apparently, the genetic distances do have a close link to the spatial distances between the subpopulations.

Visualizing a Correlation Matrix Using Multidimensional Scaling

Suppose you have computed the following correlation matrix for a set of 10 variables. It's obvious that these variables are all positively correlated, and that there are some very strong pairwise correlations. But with this many variables, it's not easy to get a good feel for the relationships among all 10.

```
Rho = ...
  [1       0.3906  0.3746  0.3318  0.4141  0.4279  0.4216  0.4703  0.4362  0.2066;
   0.3906  1       0.3200  0.3629  0.2211  0.9520  0.9811  0.9052  0.4567  0     ;
   0.3746  0.3200  1       0.8993  0.7999  0.3589  0.3460  0.3333  0.8639  0.6527;
   0.3318  0.3629  0.8993  1       0.7125  0.3959  0.3663  0.3394  0.8719  0.5726;
   0.4141  0.2211  0.7999  0.7125  1       0.2374  0.2079  0.2335  0.7050  0.7469;
   0.4279  0.9520  0.3589  0.3959  0.2374  1       0.9657  0.9363  0.4791  0.0254;
   0.4216  0.9811  0.3460  0.3663  0.2079  0.9657  1       0.9123  0.4554  0.0011;
   0.4703  0.9052  0.3333  0.3394  0.2335  0.9363  0.9123  1       0.4418  0.0099;
   0.4362  0.4567  0.8639  0.8719  0.7050  0.4791  0.4554  0.4418  1       0.5272;
   0.2066  0       0.6527  0.5726  0.7469  0.0254  0.0011  0.0099  0.5272  1     ];
```

Multidimensional scaling is often thought of as a way to (re)construct points using only pairwise distances. But it can also be used with dissimilarity measures that are more general than distance, to spatially visualize things that are not "points in space" in the usual sense. The variables described by Rho are an example, and you can use `cmdscale` to plot a visual representation of their interdependencies.

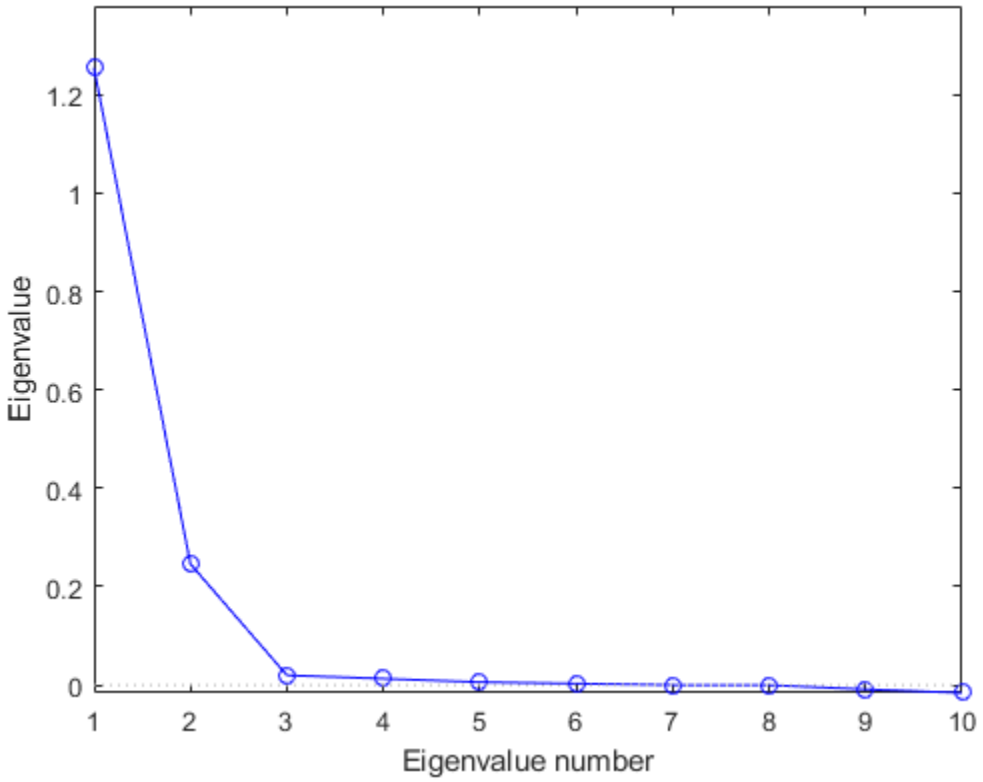
Correlation actually measures similarity, but it is easy to transform it to a measure of dissimilarity. Because all the correlations here are positive, you can simply use

```
D = 1 - Rho;
```

although other choices might also make sense. If `Rho` contained negative correlations, you would have to decide whether, for example, a correlation of -1 indicated more or less of a dissimilarity than a correlation of 0, and choose a transformation accordingly.

It's important to decide whether visualization of the information in the correlation matrix is even possible, that is, whether the number of dimensions can be reduced from ten down to two or three. The eigenvalues returned by `cmdscale` give you a way to decide. In this case, a scree plot of those eigenvalues indicates that two dimensions are enough to represent the variables. (Notice that some of the eigenvalues in the plot below are negative, but small relative to the first two.)

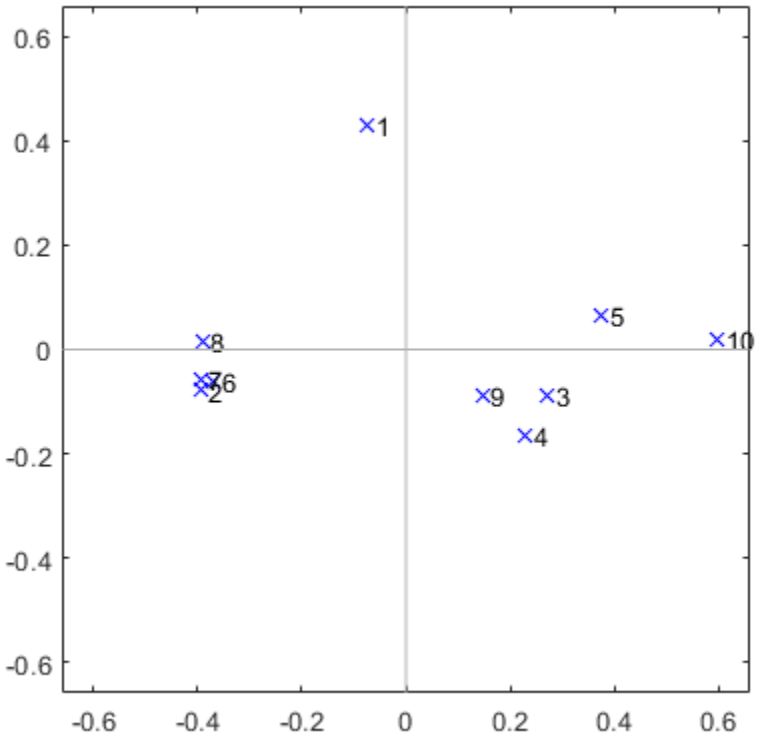
```
[Y,eigvals] = cmdscale(D);
plot(1:length(eigvals),eigvals,'bo-');
line([1,length(eigvals)],[0 0],'LineStyle',':','XLimInclude','off',...
     'Color',[.7 .7 .7])
axis([1,length(eigvals),min(eigvals),max(eigvals)*1.1]);
xlabel('Eigenvalue number');
ylabel('Eigenvalue');
```



In a more independent set of variables, more dimensions might be needed. If more than three variables are needed, the visualization isn't all that useful.

A 2-D plot of the configuration returned by `cmdscale` indicates that there are two subsets of variables that are most closely correlated among themselves, plus a single variable that is more or less on its own. One of the clusters is tight, while the other is relatively loose.

```
labels = {' 1',' 2',' 3',' 4',' 5',' 6',' 7',' 8',' 9',' 10'};
plot(Y(:,1),Y(:,2),'bx');
axis(max(max(abs(Y))) * [-1.1,1.1,-1.1,1.1]); axis('square');
text(Y(:,1),Y(:,2),labels,'HorizontalAlignment','left');
line([-1,1],[0 0],'XLimInclude','off','Color',[.7 .7 .7])
line([0 0],[-1,1],'YLimInclude','off','Color',[.7 .7 .7])
```

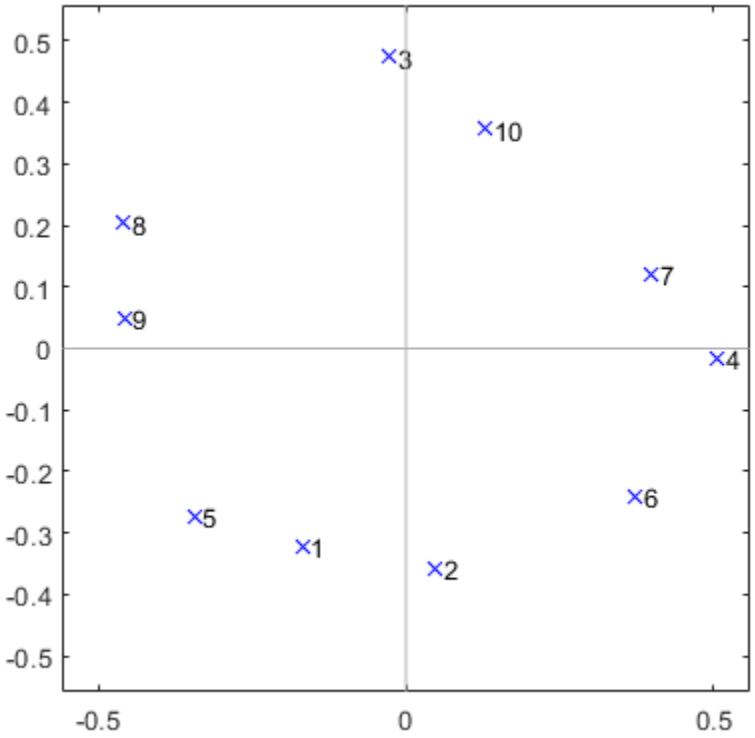


On the other hand, the results from `cmdscale` for the following correlation matrix indicates a much different structure: there are no real groups among the variables. Rather, there is a kind of "circular" dependency, where each variable has a pair of "closest neighbors" but is less well correlated with the remaining variables.

```
Rho = ...
  [1       0.7946  0.1760  0.2560  0.7818  0.4496  0.2732  0.3995  0.5305  0.2827;
   0.7946  1       0.1626  0.4227  0.5674  0.6183  0.4004  0.2283  0.3495  0.2777;
   0.1760  0.1626  1       0.2644  0.1864  0.1859  0.4330  0.4656  0.3947  0.8057;
   0.2560  0.4227  0.2644  1       0.1017  0.7426  0.8340  0       0.0499  0.4853;
   0.7818  0.5674  0.1864  0.1017  1       0.2733  0.1484  0.4890  0.6138  0.2025;
   0.4496  0.6183  0.1859  0.7426  0.2733  1       0.6303  0.0648  0.1035  0.3242;
   0.2732  0.4004  0.4330  0.8340  0.1484  0.6303  1       0.1444  0.1357  0.6291;
   0.3995  0.2283  0.4656  0       0.4890  0.0648  0.1444  1       0.8599  0.3948;
   0.5305  0.3495  0.3947  0.0499  0.6138  0.1035  0.1357  0.8599  1       0.3100;
   0.2827  0.2777  0.8057  0.4853  0.2025  0.3242  0.6291  0.3948  0.3100  1     ];

[Y,eigvals] = cmdscale(1-Rho);
[eigvals eigvals./max(abs(eigvals))]

ans = 10×2

    1.1416    1.0000
    0.7742    0.6782
    0.0335    0.0294
    0.0280    0.0245
    0.0239    0.0210
```

```
    0.0075    0.0066
    0.0046    0.0040
   -0.0000   -0.0000
   -0.0151   -0.0132
   -0.0472   -0.0413
```

```
plot(Y(:,1),Y(:,2),'bx');
axis(max(max(abs(Y))) * [-1.1,1.1,-1.1,1.1]); axis('square');
text(Y(:,1),Y(:,2),labels,'HorizontalAlignment','left');
line([0 0],[-1,1],'XLimInclude','off','Color',[.7 .7 .7])
line([-1,1],[0 0],'YLimInclude','off','Color',[.7 .7 .7])
```



A Comparison of Principal Components Analysis and Classical Multidimensional Scaling

Multidimensional scaling is most often used to visualize data when only their distances or dissimilarities are available. However, when the original data are available, multidimensional scaling can also be used as a dimension reduction method, by reducing the data to a distance matrix, creating a new configuration of points using `cmdscale`, and retaining only the first few dimensions of those points. This application of multidimensional scaling is much like Principal Components Analysis, and in fact, when you call `cmdscale` using the Euclidean distances between the points, the results are identical to PCA, up to a change in sign.

```
n = 10; m = 5;
X = randn(n,m);
D = pdist(X,'Euclidean');
```

```
[Y,eigvals] = cmdscale(D);
[PC,Score,latent] = pca(X);

Y
```

```
Y = 10×5

   -1.4505    1.6602    0.8106    0.5834    0.5952
    2.6140   -1.0513   -1.1962    0.7221   -0.2299
   -2.2399   -1.6699   -0.7881   -0.6659    0.0398
   -0.4956    0.2265    1.2682   -0.5123   -0.5702
    0.1004   -2.3659    1.2672    0.4837   -0.2888
   -2.5996    1.0635   -0.8532    0.1392   -0.1216
   -1.5565    0.4215   -0.0931    0.2863    0.0299
    0.4656   -0.6250   -0.7608   -0.3233    0.2786
    2.3961    2.6933   -0.2020   -0.2572   -0.4374
    2.7660   -0.3529    0.5474   -0.4560    0.7044
```

```
Score
```

```
Score = 10×5

   -1.4505    1.6602   -0.8106   -0.5834   -0.5952
    2.6140   -1.0513    1.1962   -0.7221    0.2299
   -2.2399   -1.6699    0.7881    0.6659   -0.0398
   -0.4956    0.2265   -1.2682    0.5123    0.5702
    0.1004   -2.3659   -1.2672   -0.4837    0.2888
   -2.5996    1.0635    0.8532   -0.1392    0.1216
   -1.5565    0.4215    0.0931   -0.2863   -0.0299
    0.4656   -0.6250    0.7608    0.3233   -0.2786
    2.3961    2.6933    0.2020    0.2572    0.4374
    2.7660   -0.3529   -0.5474    0.4560   -0.7044
```

Even the nonzero eigenvalues are identical up to a scale factor.

```
[eigvals(1:m) (n-1)*latent]
```

```
ans = 5×2

   36.9993   36.9993
   21.3766   21.3766
    7.5792    7.5792
    2.2815    2.2815
    1.5981    1.5981
```

Nonclassical Multidimensional Scaling

This example shows how to visualize dissimilarity data using nonclassical forms of multidimensional scaling (MDS).

Dissimilarity data arises when we have some set of objects, and instead of measuring the characteristics of each object, we can only measure how similar or dissimilar each pair of objects is. For example, instead of knowing the latitude and longitude of a set of cities, we may only know their inter-city distances. However, MDS also works with dissimilarities that are more abstract than physical distance. For example, we may have asked consumers to rate how similar they find several brands of peanut butter.

The typical goal of MDS is to create a configuration of points in one, two, or three dimensions, whose inter-point distances are "close" to the original dissimilarities. The different forms of MDS use different criteria to define "close". These points represent the set of objects, and so a plot of the points can be used as a visual representation of their dissimilarities.

Some applications of "classical" MDS are described in the “Classical Multidimensional Scaling Applied to Nonspatial Distances” on page 15-189 example.

Rothkopf's Morse Code Dataset

To demonstrate MDS, we'll use data collected in an experiment to investigate perception of Morse code (Rothkopf, E.Z., J.Exper.Psych., 53(2):94-101). Subjects in the study listened to two Morse code signals (audible sequences of one or more "dots" and "dashes", representing the 36 alphanumeric characters) played in succession, and were asked whether the signals were the same or different. The subjects did not know Morse code. The dissimilarity between two different characters is the frequency with which those characters were correctly distinguished.

The 36x36 matrix of dissimilarities is stored as a 630-element vector containing the subdiagonal elements of the matrix. You can use the function `squareform` to transform between the vector format and the full matrix form. Here are the first 5 letters and their dissimilarities, reconstructed in matrix form.

```
load morse
morseChars(1:5,:)
```

```
ans = 5x2 cell
    {'A'}    {'.-'  }
    {'B'}    {'-...'}
    {'C'}    {'-.-.'}
    {'D'}    {'-..' }
    {'E'}    {'.'   }
```

```
dissMatrix = squareform(dissimilarities);
dissMatrix(1:5,1:5)
```

```
ans = 5×5

     0   167   169   159   180
   167     0    96    79   163
   169    96     0   141   166
   159    79   141     0   172
   180   163   166   172     0
```

In these data, larger values indicate that more experimental subjects were able to distinguish the two signals, and so the signals were more dissimilar.

Metric Scaling

Metric MDS creates a configuration of points such that their interpoint distances approximate the original dissimilarities. One measure of the goodness of fit of that approximation is known as the "stress", and that's what we'll use initially. To compute the configuration, we provide the `mdscale` function with the dissimilarity data, the number of dimensions in which we want to create the points (two), and the name of the goodness-of-fit criterion we are using.

```
Y1 = mdscale(dissimilarities, 2, 'criterion','metricstress');
size(Y1)
```

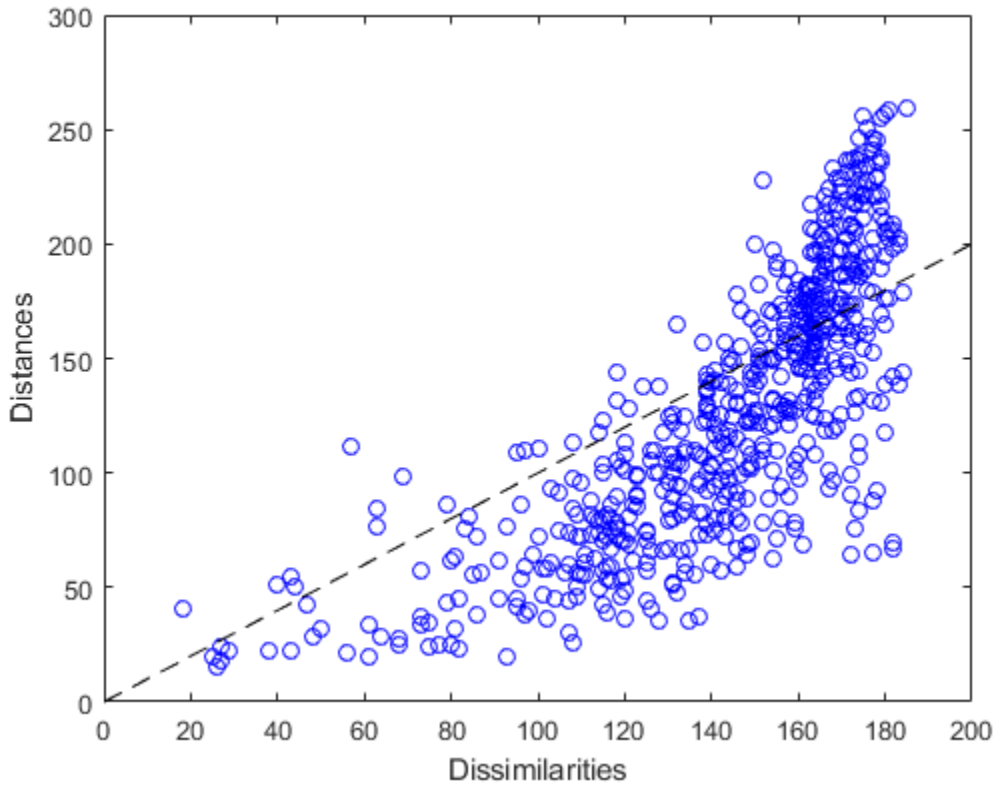
```
ans = 1×2

    36     2
```

`mdscale` returns a set of points in, for this example, two dimensions. We could plot them, but before using this solution (i.e. the configuration) to visualize the data, we'll make some plots to help check whether the interpoint distances from this solution recreate the original dissimilarities.

The Shepard Plot

The Shepard plot is a scatterplot of the interpoint distances (there are n(n-1)/2 of them) vs. the original dissimilarities. This can help determine goodness of fit of the MDS solution. If the fit is poor, then visualization could be misleading, because large (small) distances between points might not correspond to large (small) dissimilarities in the data. In the Shepard plot, a narrow scatter around a 1:1 line indicates a good fit of the distances to the dissimilarities, while a large scatter or a nonlinear pattern indicates a lack of fit.

```
distances1 = pdist(Y1);
plot(dissimilarities,distances1,'bo', [0 200],[0 200],'k--');
xlabel('Dissimilarities')
ylabel('Distances')
```



This plot indicates that this metric solution in two dimensions is probably not appropriate, because it shows both a nonlinear pattern and a large scatter. The former implies that many of the largest dissimilarities would tend to be somewhat exaggerated in the visualization, while moderate and small dissimilarities would tend to be understated. The latter implies that distance in the visualization would generally be a poor reflection of dissimilarity. In particular, a good fraction of the large dissimilarities would be badly understated.

Comparing Metric Criteria

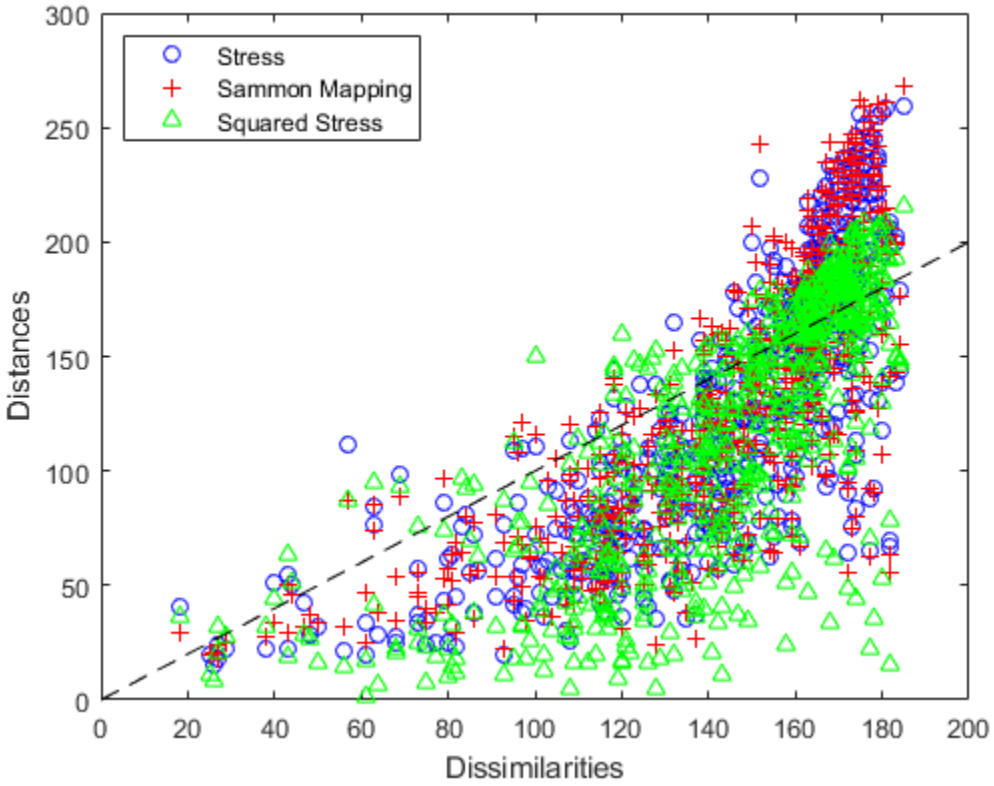
We could try using a third dimension to improve the fidelity of the visualization, because with more degrees of freedom, the fit should improve. We can also try a different criterion. Two other popular metric criteria are known as Sammon Mapping and squared stress ("sstress"). Each leads to a different solution, and one or the other might be more useful in visualizing the original dissimilarities.

```
Y2 = mdscale(dissimilarities,2, 'criterion','sammon');
distances2 = pdist(Y2);
Y3 = mdscale(dissimilarities,2, 'criterion','metricsstress');
distances3 = pdist(Y3);
```

A Shepard plot shows the differences in the three solutions so far.

```
plot(dissimilarities,distances1,'bo', ...
     dissimilarities,distances2,'r+', ...
     dissimilarities,distances3,'g^', ...
     [0 200],[0 200],'k--');
xlabel('Dissimilarities')
```

```
ylabel('Distances')
legend({'Stress', 'Sammon Mapping', 'Squared Stress'}, 'Location','NorthWest');
```



Notice that at the largest dissimilarity values, the scatter for the squared stress criterion tends to be closer to the 1:1 line than for the other two criteria. Thus, for these data, squared stress is somewhat better at preserving the largest dissimilarities, although it badly understates some of those. At smaller dissimilarity values, the scatter for the Sammon Mapping criterion tends to be somewhat closer to the 1:1 line than for the other two criteria. Thus, Sammon Mapping is a little better at preserving small dissimilarities. Stress is somewhere in between. All three criteria show a certain amount of nonlinearity, indicating that metric scaling may not be suitable. However, the choice of criterion depends on the goal of the visualization.

Nonmetric Scaling

Nonmetric scaling is a second form of MDS that has a slightly less ambitious goal than metric scaling. Instead of attempting to create a configuration of points for which the pairwise distances approximate the original dissimilarities, nonmetric MDS attempts only to approximate the *ranks* of the dissimilarities. Another way of saying this is that nonmetric MDS creates a configuration of points whose interpoint distances approximate a *monotonic transformation* of the original dissimilarities.

The practical use of such a construction is that large interpoint distances correspond to large dissimilarities, and small interpoint distances to small dissimilarities. This is often sufficient to convey the relationships among the items or categories being studied.

First, we'll create a configuration of points in 2D. Nonmetric scaling with Kruskal's nonmetric stress criterion is the default for `mdscale`.

```
[Y,stress,disparities] = mdscale(dissimilarities,2);
stress
```

```
stress = 0.1800
```

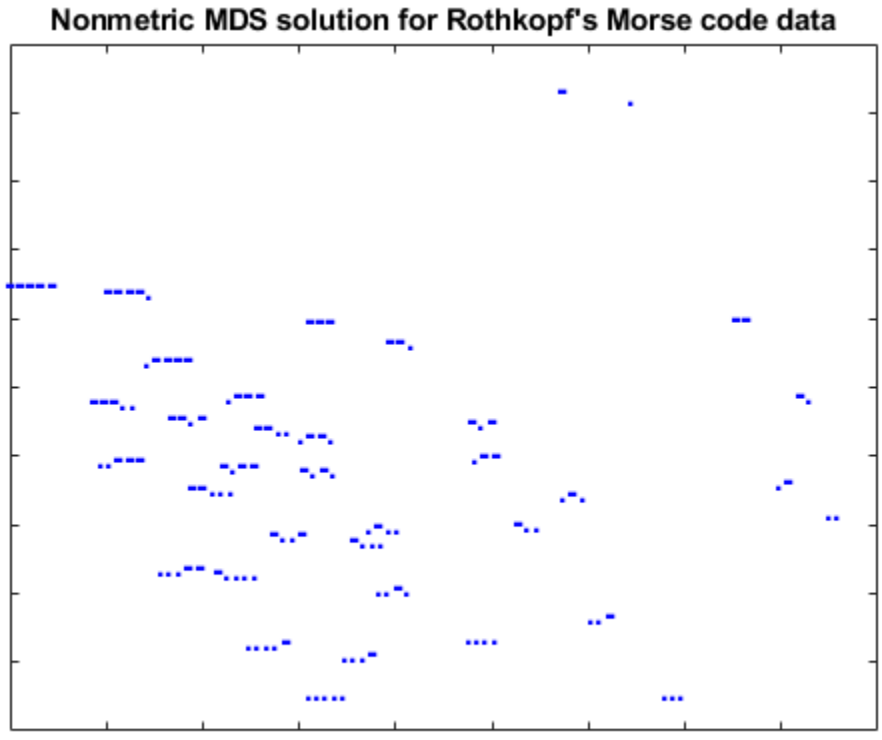
The second output of `mdscale` is the value of the criterion being used, as a measure of how well the solution recreates the dissimilarities. Smaller values indicate a better fit. The stress for this configuration, about 18%, is considered poor to fair for the nonmetric stress criterion. The ranges of acceptable criterion values differ for the different criteria.

The third output of `mdscale` is a vector of what are known as disparities. These are simply the monotonic transformation of the dissimilarities. They will be used in a nonmetric scaling Shepard plot below.

Visualizing the Dissimilarity Data

Although this fit is not as good as we would like, the 2D representation is easiest to visualize. We can plot each signal's dots and dashes to help see why the subjects perceive differences among the characters. The orientation and scale of this configuration is completely arbitrary, so no axis labels or values have been shown.

```
plot(Y(:,1),Y(:,2),'.', 'Marker','none');
text(Y(:,1),Y(:,2),char(morseChars(:,2)), 'Color','b', ...
    'FontSize',12,'FontWeight','bold', 'HorizontalAlignment','center');
h_gca = gca;
h_gca.XTickLabel = [];
h_gca.YTickLabel = [];
title('Nonmetric MDS solution for Rothkopf''s Morse code data');
```

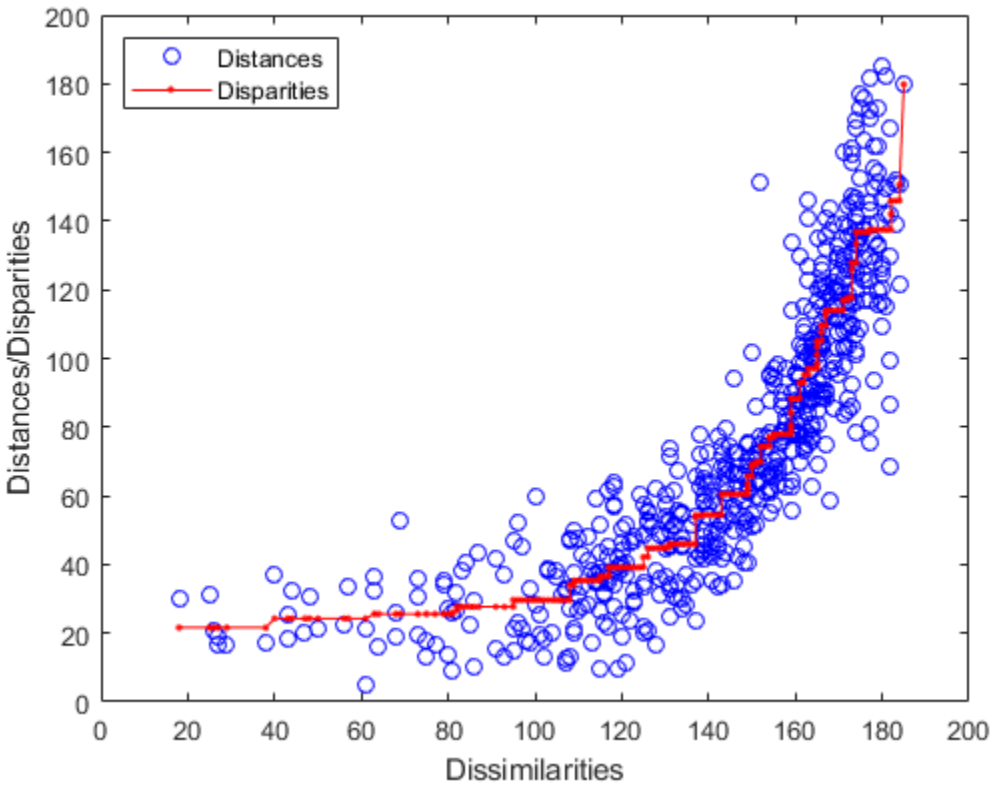


This reconstruction indicates that the characters can be described in terms of two axes: roughly speaking, the northwest/southeast direction discriminates signal length, while the southwest/northeast direction discriminates dots from dashes. The two characters with the shortest signals, 'E' and 'T', are somewhat out of position in that interpretation.

The Nonmetric Shepard Plot

In nonmetric scaling, it is customary to show the disparities as well as the distances in a Shepard plot. This provides a check on how well the distances recreate the disparities, as well as how nonlinear the monotonic transformation from dissimilarities to disparities is.

```
distances = pdist(Y);
[dum,ord] = sortrows([disparities(:) dissimilarities(:)]);
plot(dissimilarities,distances,'bo', ...
     dissimilarities(ord),disparities(ord),'r.-');
xlabel('Dissimilarities')
ylabel('Distances/Disparities')
legend({'Distances' 'Disparities'}, 'Location','NorthWest');
```



This plot shows how the distances in nonmetric scaling approximate the disparities (the scatter of blue circles about the red line), and the disparities reflect the ranks of the dissimilarities (the red line is nonlinear but increasing). Comparing this plot to the Shepard plot from metric scaling shows the difference in the two methods. Nonmetric scaling attempts to recreate not the original dissimilarities, but rather a nonlinear transformation of them (the disparities).

In doing that, nonmetric scaling has made a trade-off: the nonmetric distances recreate the disparities better than the metric distances recreated the dissimilarities -- the scatter in this plot is

smaller that in the metric plot. However, the disparities are quite nonlinear as a function of the dissimilarities. Thus, while we can be more certain that with the nonmetric solution, small distances in the visualization correspond to small dissimilarities in the data, it's important to remember that absolute distances between points in that visualization should not be taken too literally -- only relative distances.

Nonmetric Scaling in 3D

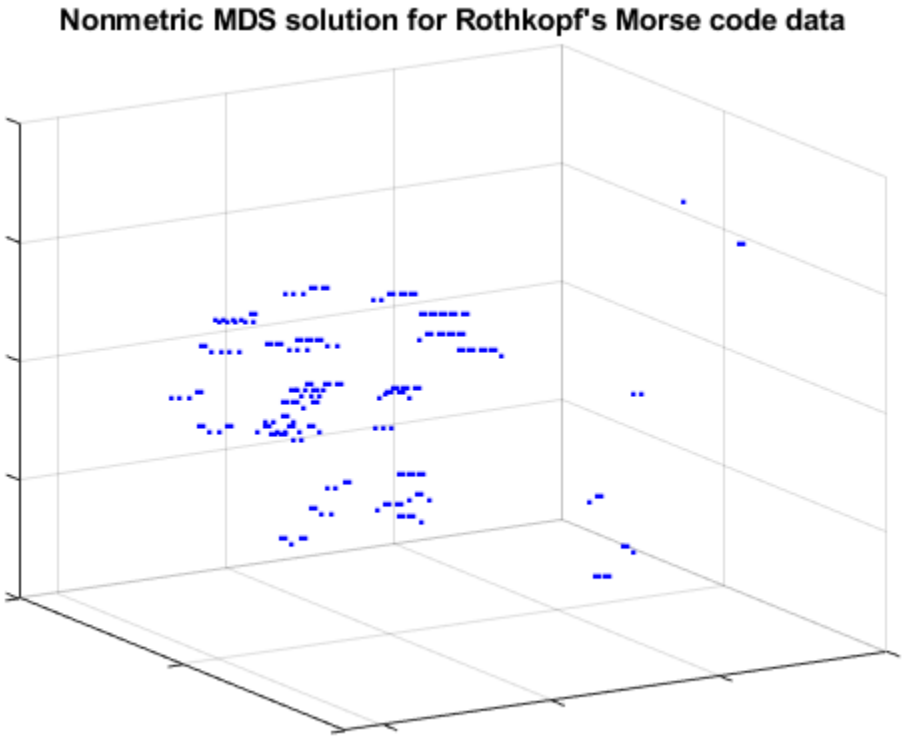
Because the stress in the 2D construction was somewhat high, we can try a 3D configuration.

```
[Y,stress,disparities] = mdscale(dissimilarities,3);
stress
```

```
stress = 0.1189
```

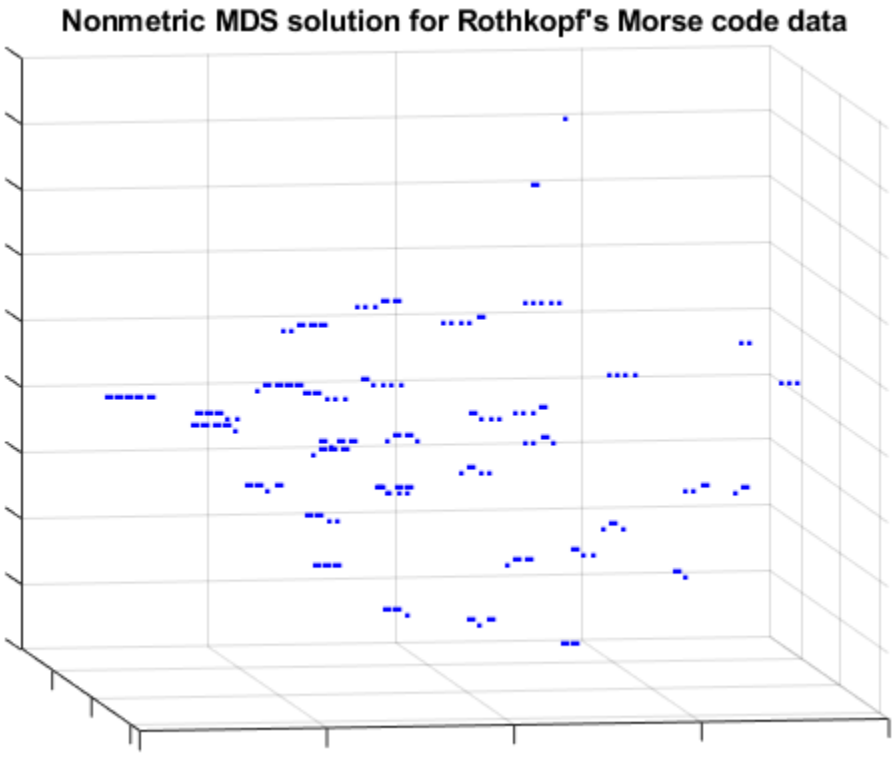
This stress value is quite a bit lower, indicating a better fit. We can plot the configuration in 3 dimensions. A live MATLAB® figure can be rotated interactively; here we will settle for looking from two different angles.

```
plot3(Y(:,1),Y(:,2),Y(:,3),'.', 'Marker','none');
text(Y(:,1),Y(:,2),Y(:,3),char(morseChars(:,2)), 'Color','b', ...
    'FontSize',12,'FontWeight','bold', 'HorizontalAlignment','center');
set(gca,'XTickLabel',[], 'YTickLabel',[], 'ZTickLabel',[]);
title('Nonmetric MDS solution for Rothkopf''s Morse code data');
view(59,18);
grid on
```



From this angle, we can see that the characters with one- and two-symbol signals are well-separated from the characters with longer signals, and from each other, because they are the easiest to distinguish. If we rotate the view to a different perspective, we can see that the longer characters can, as in the 2D configuration, roughly be described in terms of the number of symbols and the number of dots or dashes. (From this second angle, some of the shorter characters spuriously appear to be interspersed with the longer ones.)

```
view(-9,8);
```



This 3D configuration reconstructs the distances more accurately than the 2D configuration, however, the message is essentially the same: the subjects perceive the signals primarily in terms of how many symbols they contain, and how many dots vs. dashes. In practice, the 2D configuration might be perfectly acceptable.

Fitting an Orthogonal Regression Using Principal Components Analysis

This example shows how to use Principal Components Analysis (PCA) to fit a linear regression. PCA minimizes the perpendicular distances from the data to the fitted model. This is the linear case of what is known as Orthogonal Regression or Total Least Squares, and is appropriate when there is no natural distinction between predictor and response variables, or when all variables are measured with error. This is in contrast to the usual regression assumption that predictor variables are measured exactly, and only the response variable has an error component.

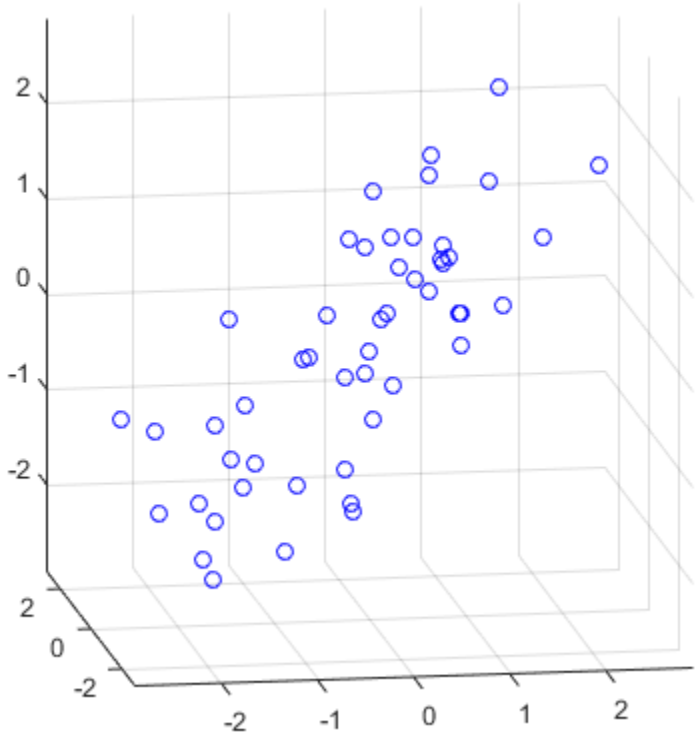
For example, given two data vectors x and y, you can fit a line that minimizes the perpendicular distances from each of the points (x(i), y(i)) to the line. More generally, with p observed variables, you can fit an r-dimensional hyperplane in p-dimensional space (r < p). The choice of r is equivalent to choosing the number of components to retain in PCA. It may be based on prediction error, or it may simply be a pragmatic choice to reduce data to a manageable number of dimensions.

In this example, we fit a plane and a line through some data on three observed variables. It's easy to do the same thing for any number of variables, and for any dimension of model, although visualizing a fit in higher dimensions would obviously not be straightforward.

Fitting a Plane to 3-D Data

First, we generate some trivariate normal data for the example. Two of the variables are fairly strongly correlated.

```
rng(5,'twister');
X = mvnrnd([0 0 0], [1 .2 .7; .2 1 0; .7 0 1],50);
plot3(X(:,1),X(:,2),X(:,3),'bo');
grid on;
maxlim = max(abs(X(:)))*1.1;
axis([-maxlim maxlim -maxlim maxlim -maxlim maxlim]);
axis square
view(-9,12);
```



Next, we fit a plane to the data using PCA. The coefficients for the first two principal components define vectors that form a basis for the plane. The third PC is orthogonal to the first two, and its coefficients define the normal vector of the plane.

```
[coeff,score,roots] = pca(X);
basis = coeff(:,1:2)
```

```
basis = 3×2

    0.6774   -0.0790
    0.2193    0.9707
    0.7022   -0.2269
```

```
normal = coeff(:,3)
```

```
normal = 3×1

    0.7314
   -0.0982
   -0.6749
```

That's all there is to the fit. But let's look closer at the results, and plot the fit along with the data.

Because the first two components explain as much of the variance in the data as is possible with two dimensions, the plane is the best 2-D linear approximation to the data. Equivalently, the third component explains the least amount of variation in the data, and it is the error term in the

regression. The latent roots (or eigenvalues) from the PCA define the amount of explained variance for each component.

```
pctExplained = roots' ./ sum(roots)
```

```
pctExplained = 1×3

    0.6226    0.2976    0.0798
```

The first two coordinates of the principal component scores give the projection of each point onto the plane, in the coordinate system of the plane. To get the coordinates of the fitted points in terms of the original coordinate system, we multiply each PC coefficient vector by the corresponding score, and add back in the mean of the data. The residuals are simply the original data minus the fitted points.

```
[n,p] = size(X);
meanX = mean(X,1);
Xfit = repmat(meanX,n,1) + score(:,1:2)*coeff(:,1:2)';
residuals = X - Xfit;
```

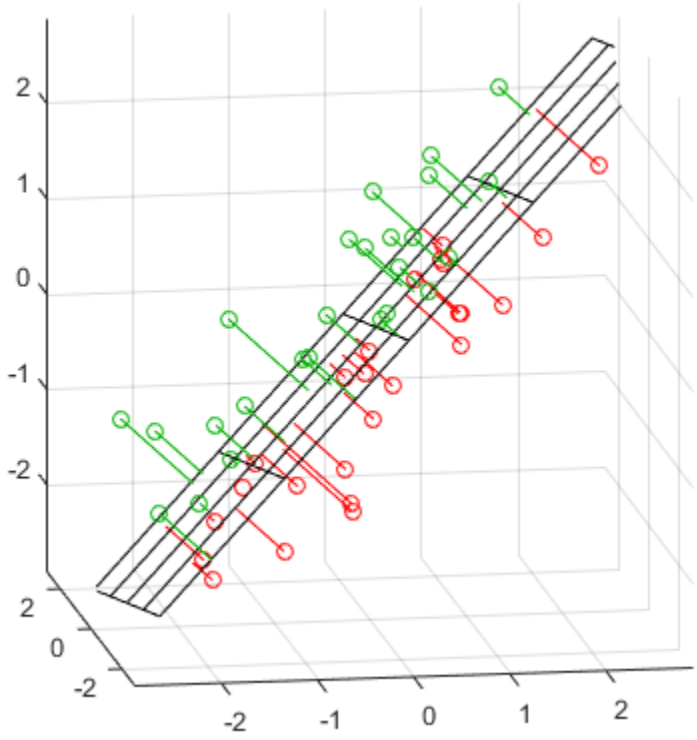
The equation of the fitted plane, satisfied by each of the fitted points in `Xfit`, is `([x1 x2 x3] - meanX)*normal = 0`. The plane passes through the point `meanX`, and its perpendicular distance to the origin is `meanX*normal`. The perpendicular distance from each point in `X` to the plane, i.e., the norm of the residuals, is the dot product of each centered point with the normal to the plane. The fitted plane minimizes the sum of the squared errors.

```
error = abs((X - repmat(meanX,n,1))*normal);
sse = sum(error.^2)
```

```
sse = 15.5142
```

To visualize the fit, we can plot the plane, the original data, and their projection to the plane.

```
[xgrid,ygrid] = meshgrid(linspace(min(X(:,1)),max(X(:,1)),5), ...
                         linspace(min(X(:,2)),max(X(:,2)),5));
zgrid = (1/normal(3)) .* (meanX*normal - (xgrid.*normal(1) + ygrid.*normal(2)));
h = mesh(xgrid,ygrid,zgrid,'EdgeColor',[0 0 0],'FaceAlpha',0);

hold on
above = (X-repmat(meanX,n,1))*normal < 0;
below = ~above;
nabove = sum(above);
X1 = [X(above,1) Xfit(above,1) nan*ones(nabove,1)];
X2 = [X(above,2) Xfit(above,2) nan*ones(nabove,1)];
X3 = [X(above,3) Xfit(above,3) nan*ones(nabove,1)];
plot3(X1',X2',X3','-', X(above,1),X(above,2),X(above,3),'o', 'Color',[0 .7 0]);
nbelow = sum(below);
X1 = [X(below,1) Xfit(below,1) nan*ones(nbelow,1)];
X2 = [X(below,2) Xfit(below,2) nan*ones(nbelow,1)];
X3 = [X(below,3) Xfit(below,3) nan*ones(nbelow,1)];
plot3(X1',X2',X3','-', X(below,1),X(below,2),X(below,3),'o', 'Color',[1 0 0]);

hold off
maxlim = max(abs(X(:)))*1.1;
axis([-maxlim maxlim -maxlim maxlim -maxlim maxlim]);
axis square
view(-9,12);
```



Green points are above the plane, red points are below.

Fitting a Line to 3-D Data

Fitting a straight line to the data is even simpler, and because of the nesting property of PCA, we can use the components that have already been computed. The direction vector that defines the line is given by the coefficients for the first principal component. The second and third PCs are orthogonal to the first, and their coefficients define directions that are perpendicular to the line. The simplest equation to describe the line is `meanX + t*dirVect`, where `t` parameterizes the position along the line.

```
dirVect = coeff(:,1)
```

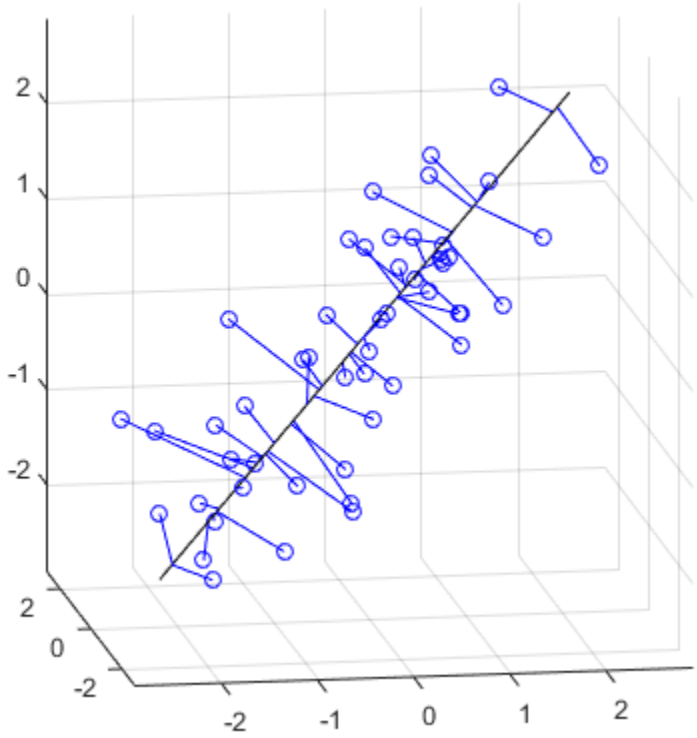
```
dirVect = 3×1

    0.6774
    0.2193
    0.7022
```

The first coordinate of the principal component scores gives the projection of each point onto the line. As with the 2-D fit, the PC coefficient vectors multiplied by the scores the gives the fitted points in the original coordinate system.

```
Xfit1 = repmat(meanX,n,1) + score(:,1)*coeff(:,1)';
```

Plot the line, the original data, and their projection to the line.

```
t = [min(score(:,1))-.2, max(score(:,1))+.2];
endpts = [meanX + t(1)*dirVect'; meanX + t(2)*dirVect'];
plot3(endpts(:,1),endpts(:,2),endpts(:,3),'k-');

X1 = [X(:,1) Xfit1(:,1) nan*ones(n,1)];
X2 = [X(:,2) Xfit1(:,2) nan*ones(n,1)];
X3 = [X(:,3) Xfit1(:,3) nan*ones(n,1)];
hold on
plot3(X1',X2',X3','b-', X(:,1),X(:,2),X(:,3),'bo');
hold off
maxlim = max(abs(X(:)))*1.1;
axis([-maxlim maxlim -maxlim maxlim -maxlim maxlim]);
axis square
view(-9,12);
grid on
```



While it appears that many of the projections in this plot are not perpendicular to the line, that's just because we're plotting 3-D data in two dimensions. In a live MATLAB® figure window, you could interactively rotate the plot to different perspectives to verify that the projections are indeed perpendicular, and to get a better feel for how the line fits the data.

Tune Regularization Parameter to Detect Features Using NCA for Classification

This example shows how to tune the regularization parameter in `fscnca` using cross-validation. Tuning the regularization parameter helps to correctly detect the relevant features in the data.

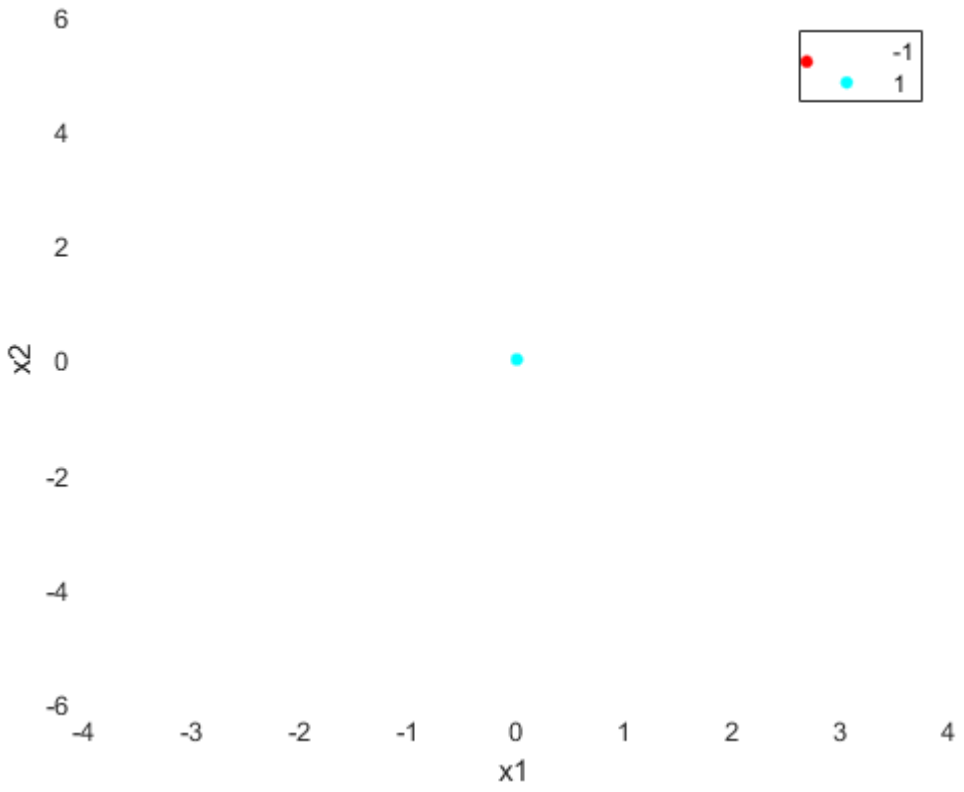
Load the sample data.

```
load('twodimclassdata.mat')
```

This dataset is simulated using the scheme described in [1]. This is a two-class classification problem in two dimensions. Data from the first class are drawn from two bivariate normal distributions $N(\mu_1, \Sigma)$ or $N(\mu_2, \Sigma)$ with equal probability, where $\mu_1 = [-0.75, -1.5]$, $\mu_2 = [0.75, 1.5]$ and $\Sigma = I_2$. Similarly, data from the second class are drawn from two bivariate normal distributions $N(\mu_3, \Sigma)$ or $N(\mu_4, \Sigma)$ with equal probability, where $\mu_3 = [1.5, -1.5]$, $\mu_4 = [-1.5, 1.5]$ and $\Sigma = I_2$. The normal distribution parameters used to create this data set results in tighter clusters in data than the data used in [1].

Create a scatter plot of the data grouped by the class.

```
figure
gscatter(X(:,1),X(:,2),y)
xlabel('x1')
ylabel('x2')
```



Add 100 irrelevant features to X. First generate data from a Normal distribution with a mean of 0 and a variance of 20.

```
n = size(X,1);
rng('default')
XwithBadFeatures = [X,randn(n,100)*sqrt(20)];
```

Normalize the data so that all points are between 0 and 1.

```
XwithBadFeatures = (XwithBadFeatures-min(XwithBadFeatures,[],1))./range(XwithBadFeatures,1);
X = XwithBadFeatures;
```

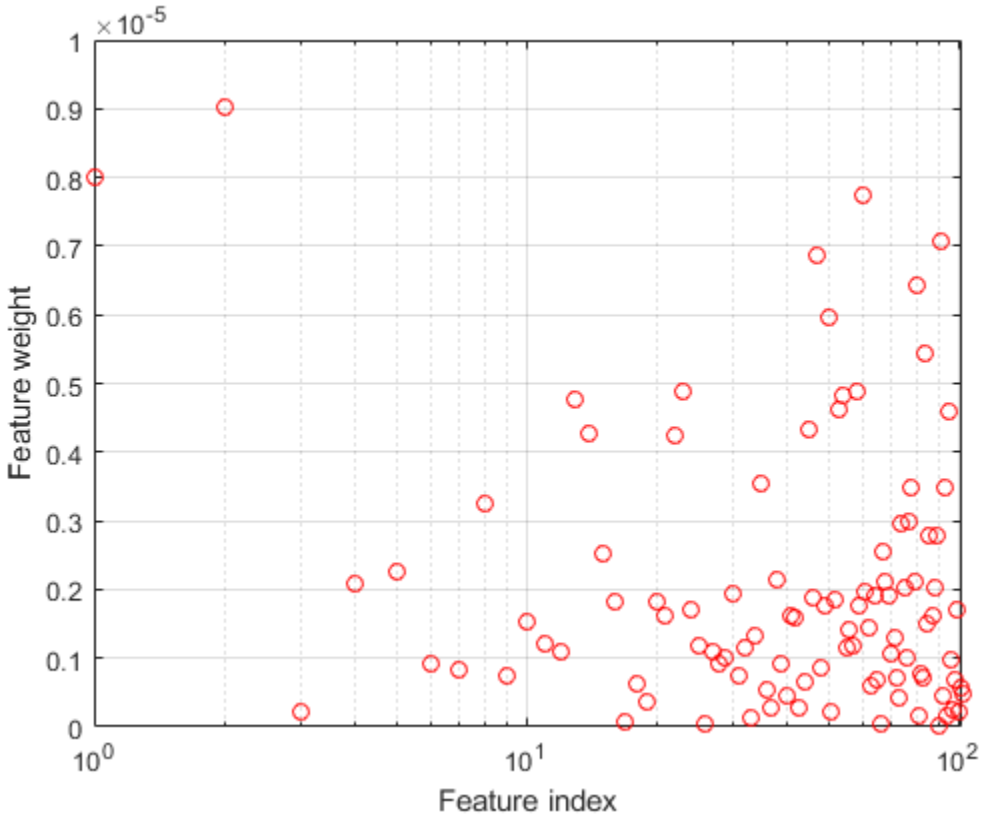
Fit an nca model to the data using the default `Lambda` (regularization parameter, λ) value. Use the LBFGS solver and display the convergence information.

```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
    'Solver','lbfgs');
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 |  9.519258e-03 |   1.494e-02 |   0.000e+00 |        |   4.015e+01 |   0.000e+00 |
|        1 | -3.093574e-01 |   7.186e-03 |   4.018e+00 |    OK  |   8.956e+01 |   1.000e+00 |
|        2 | -4.809455e-01 |   4.444e-03 |   7.123e+00 |    OK  |   9.943e+01 |   1.000e+00 |
|        3 | -4.938877e-01 |   3.544e-03 |   1.464e+00 |    OK  |   9.366e+01 |   1.000e+00 |
|        4 | -4.964759e-01 |   2.901e-03 |   6.084e-01 |    OK  |   1.554e+02 |   1.000e+00 |
|        5 | -4.972077e-01 |   1.323e-03 |   6.129e-01 |    OK  |   1.195e+02 |   5.000e-01 |
|        6 | -4.974743e-01 |   1.569e-04 |   2.155e-01 |    OK  |   1.003e+02 |   1.000e+00 |
|        7 | -4.974868e-01 |   3.844e-05 |   4.161e-02 |    OK  |   9.835e+01 |   1.000e+00 |
|        8 | -4.974874e-01 |   1.417e-05 |   1.073e-02 |    OK  |   1.043e+02 |   1.000e+00 |
|        9 | -4.974874e-01 |   4.893e-06 |   1.781e-03 |    OK  |   1.530e+02 |   1.000e+00 |
|       10 | -4.974874e-01 |   9.404e-08 |   8.947e-04 |    OK  |   1.670e+02 |   1.000e+00 |

         Infinity norm of the final gradient = 9.404e-08
             Two norm of the final step     = 8.947e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 9.404e-08, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the feature weights. The weights of the irrelevant features should be very close to zero.

```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



All weights are very close to zero. This indicates that the value of λ used in training the model is too large. When $\lambda \to \infty$, all features weights approach to zero. Hence, it is important to tune the regularization parameter in most cases to detect the relevant features.

Use five-fold cross-validation to tune λ for feature selection using `fscnca`. Tuning λ means finding the λ value that will produce the minimum classification loss. Here are the steps for tuning λ using cross-validation:

1. First partition the data into five folds. For each fold, `cvpartition` assigns 4/5th of the data as a training set, and 1/5th of the data as a test set.

```
cvp           = cvpartition(y,'kfold',5);
numtestsets   = cvp.NumTestSets;
lambdavalues  = linspace(0,2,20)/length(y);
lossvalues    = zeros(length(lambdavalues),numtestsets);
```

2. Train the neighborhood component analysis (nca) model for each λ value using the training set in each fold.

3. Compute the classification loss for the corresponding test set in the fold using the nca model. Record the loss value.

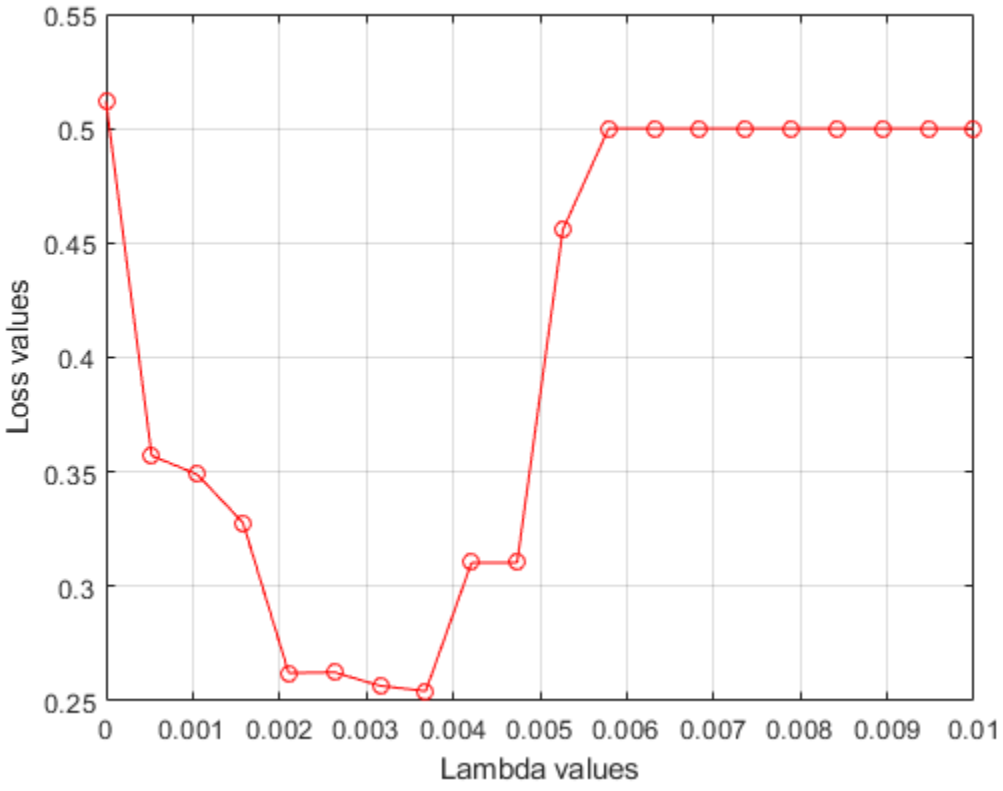
4. Repeat this for all folds and all λ values.

```
for i = 1:length(lambdavalues)
    for k = 1:numtestsets
```

```
        % Extract the training set from the partition object
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);

        % Extract the test set from the partition object
        Xtest = X(cvp.test(k),:);
        ytest = y(cvp.test(k),:);

        % Train an nca model for classification using the training set
        ncaMdl = fscnca(Xtrain,ytrain,'FitMethod','exact', ...
            'Solver','lbfgs','Lambda',lambdavalues(i));

        % Compute the classification loss for the test set using the nca
        % model
        lossvalues(i,k) = loss(ncaMdl,Xtest,ytest, ...
            'LossFunction','quadratic');

    end
end
```

Plot the average loss values of the folds versus the λ values.

```
figure
plot(lambdavalues,mean(lossvalues,2),'ro-')
xlabel('Lambda values')
ylabel('Loss values')
grid on
```



Find the λ value that corresponds to the minimum average loss.

```
[~,idx] = min(mean(lossvalues,2)); % Find the index
bestlambda = lambdavalues(idx) % Find the best lambda value
```

```
bestlambda = 0.0037
```

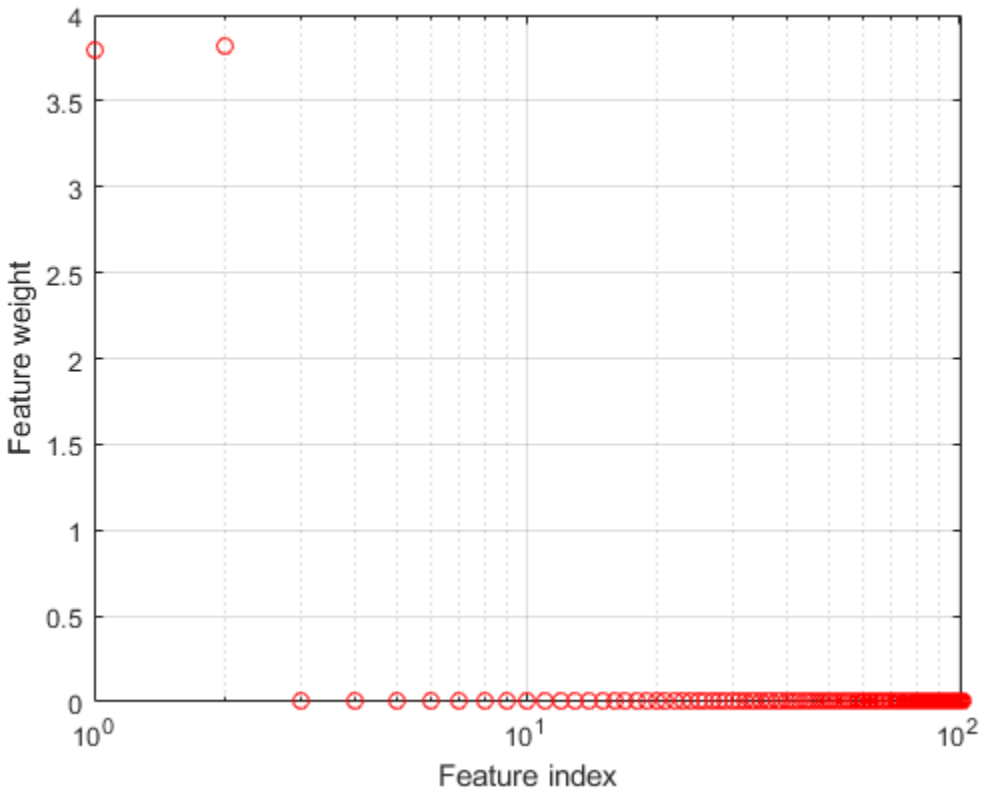
Fit the nca model to all of the data using the best λ value. Use the LBFGS solver and display the convergence information.

```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
     'Solver','lbfgs','Lambda',bestlambda);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=====================================================================================================
|        0 | -1.246913e-01 |   1.231e-02 |   0.000e+00 |        |   4.873e+01 |   0.000e+00 |   Y
|        1 | -3.411330e-01 |   5.717e-03 |   3.618e+00 |    OK  |   1.068e+02 |   1.000e+00 |   Y
|        2 | -5.226111e-01 |   3.763e-02 |   8.252e+00 |    OK  |   7.825e+01 |   1.000e+00 |   Y
|        3 | -5.817731e-01 |   8.496e-03 |   2.340e+00 |    OK  |   5.591e+01 |   5.000e-01 |   Y
|        4 | -6.132632e-01 |   6.863e-03 |   2.526e+00 |    OK  |   8.228e+01 |   1.000e+00 |   Y
|        5 | -6.135264e-01 |   9.373e-03 |   7.341e-01 |    OK  |   3.244e+01 |   1.000e+00 |   Y
|        6 | -6.147894e-01 |   1.182e-03 |   2.933e-01 |    OK  |   2.447e+01 |   1.000e+00 |   Y
|        7 | -6.148714e-01 |   6.392e-04 |   6.688e-02 |    OK  |   3.195e+01 |   1.000e+00 |   Y
|        8 | -6.149524e-01 |   6.521e-04 |   9.934e-02 |    OK  |   1.236e+02 |   1.000e+00 |   Y
|        9 | -6.149972e-01 |   1.154e-04 |   1.191e-01 |    OK  |   1.171e+02 |   1.000e+00 |   Y
|       10 | -6.149990e-01 |   2.922e-05 |   1.983e-02 |    OK  |   7.365e+01 |   1.000e+00 |   Y
|       11 | -6.149993e-01 |   1.556e-05 |   8.354e-03 |    OK  |   1.288e+02 |   1.000e+00 |   Y
|       12 | -6.149994e-01 |   1.147e-05 |   7.256e-03 |    OK  |   2.332e+02 |   1.000e+00 |   Y
|       13 | -6.149995e-01 |   1.040e-05 |   6.781e-03 |    OK  |   2.287e+02 |   1.000e+00 |   Y
|       14 | -6.149996e-01 |   9.015e-06 |   6.265e-03 |    OK  |   9.974e+01 |   1.000e+00 |   Y
|       15 | -6.149996e-01 |   7.763e-06 |   5.206e-03 |    OK  |   2.919e+02 |   1.000e+00 |   Y
|       16 | -6.149997e-01 |   8.374e-06 |   1.679e-02 |    OK  |   6.878e+02 |   1.000e+00 |   Y
|       17 | -6.149997e-01 |   9.387e-06 |   9.542e-03 |    OK  |   1.284e+02 |   5.000e-01 |   Y
|       18 | -6.149997e-01 |   3.250e-06 |   5.114e-03 |    OK  |   1.225e+02 |   1.000e+00 |   Y
|       19 | -6.149997e-01 |   1.574e-06 |   1.275e-03 |    OK  |   1.808e+02 |   1.000e+00 |   Y

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=====================================================================================================
|       20 | -6.149997e-01 |   5.764e-07 |   6.765e-04 |    OK  |   2.905e+02 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 5.764e-07
             Two norm of the final step     = 6.765e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 5.764e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the feature weights.

```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



`fscnca` correctly figures out that the first two features are relevant and the rest are not. Note that the first two features are not individually informative but when taken together result in an accurate classification model.

References

1. Yang, W., K. Wang, W. Zuo. "Neighborhood Component Feature Selection for High-Dimensional Data." *Journal of Computers*. Vol. 7, Number 1, January, 2012.

See Also

`FeatureSelectionNCAClassification` | `fscnca` | `loss` | `predict` | `refit`

More About

- “Neighborhood Component Analysis (NCA) Feature Selection” on page 15-99
- “Introduction to Feature Selection” on page 15-49

Cluster Analysis

Choose Cluster Analysis Method

This topic provides a brief overview of the available clustering methods in Statistics and Machine Learning Toolbox.

Clustering Methods

Cluster analysis, also called segmentation analysis or taxonomy analysis, is a common unsupervised learning method. Unsupervised learning is used to draw inferences from data sets consisting of input data without labeled responses. For example, you can use cluster analysis for exploratory data analysis to find hidden patterns or groupings in unlabeled data.

Cluster analysis creates groups, or *clusters*, of data. Objects that belong to the same cluster are similar to one another and distinct from objects that belong to different clusters. To quantify "similar" and "distinct," you can use a dissimilarity measure (or distance metric on page 18-12) that is specific to the domain of your application and your data set. Also, depending on your application, you might consider scaling (or standardizing) the variables in your data to give them equal importance during clustering.

Statistics and Machine Learning Toolbox provides functionality for these clustering methods:

- “Hierarchical Clustering” on page 16-2
- “k-Means and k-Medoids Clustering” on page 16-2
- “Density-Based Spatial Clustering of Applications with Noise (DBSCAN)” on page 16-3
- “Gaussian Mixture Model” on page 16-3
- “k-Nearest Neighbor Search and Radius Search” on page 16-3
- “Spectral Clustering” on page 16-3

Hierarchical Clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree, or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level combine to form clusters at the next level. This multilevel hierarchy allows you to choose the level, or scale, of clustering that is most appropriate for your application. Hierarchical clustering assigns every point in your data to a cluster.

Use `clusterdata` to perform hierarchical clustering on input data. `clusterdata` incorporates the `pdist`, `linkage`, and `cluster` functions, which you can use separately for more detailed analysis. The `dendrogram` function plots the cluster tree. For more information, see “Introduction to Hierarchical Clustering” on page 16-6.

k-Means and k-Medoids Clustering

k-means clustering and *k*-medoids clustering partition data into *k* mutually exclusive clusters. These clustering methods require that you specify the number of clusters *k*. Both *k*-means and *k*-medoids clustering assign every point in your data to a cluster; however, unlike hierarchical clustering, these methods operate on actual observations (rather than dissimilarity measures), and create a single level of clusters. Therefore, *k*-means or *k*-medoids clustering is often more suitable than hierarchical clustering for large amounts of data.

Use `kmeans` and `kmedoids` to implement *k*-means clustering and *k*-medoids clustering, respectively. For more information, see Introduction to *k*-Means Clustering on page 16-33 and *k*-Medoids Clustering on page 33-3345.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based algorithm that identifies arbitrarily shaped clusters and outliers (noise) in data. During clustering, DBSCAN identifies points that do not belong to any cluster, which makes this method useful for density-based outlier detection. Unlike *k*-means and *k*-medoids clustering, DBSCAN does not require prior knowledge of the number of clusters.

Use `dbscan` to perform clustering on an input data matrix or on pairwise distances between observations. For more information, see “Introduction to DBSCAN” on page 16-19.

Gaussian Mixture Model

A Gaussian mixture model (GMM) forms clusters as a mixture of multivariate normal density components. For a given observation, the GMM assigns posterior probabilities to each component density (or cluster). The posterior probabilities indicate that the observation has some probability of belonging to each cluster. A GMM can perform *hard* clustering by selecting the component that maximizes the posterior probability as the assigned cluster for the observation. You can also use a GMM to perform *soft*, or *fuzzy*, clustering by assigning the observation to multiple clusters based on the scores or posterior probabilities of the observation for the clusters. A GMM can be a more appropriate method than *k*-means clustering when clusters have different sizes and different correlation structures within them.

Use `fitgmdist` to fit a `gmdistribution` object to your data. You can also use `gmdistribution` to create a GMM object by specifying the distribution parameters. When you have a fitted GMM, you can cluster query data by using the `cluster` function. For more information, see “Cluster Using Gaussian Mixture Model” on page 16-39.

k-Nearest Neighbor Search and Radius Search

k-nearest neighbor search finds the *k* closest points in your data to a query point or set of query points. In contrast, radius search finds all points in your data that are within a specified distance from a query point or set of query points. The results of these methods depend on the distance metric on page 18-12 that you specify.

Use the `knnsearch` function to find *k*-nearest neighbors or the `rangesearch` function to find all neighbors within a specified distance of your input data. You can also create a searcher object using a training data set, and pass the object and query data sets to the object functions (`knnsearch` and `rangesearch`). For more information, see “Classification Using Nearest Neighbors” on page 18-12.

Spectral Clustering

Spectral clustering is a graph-based algorithm for finding *k* arbitrarily shaped clusters in data. The technique involves representing the data in a low dimension. In the low dimension, clusters in the data are more widely separated, enabling you to use algorithms such as *k*-means or *k*-medoids clustering. This low dimension is based on eigenvectors of a Laplacian matrix. A Laplacian matrix is one way of representing a similarity graph that models the local neighborhood relationships between data points as an undirected graph.

Use `spectralcluster` to perform spectral clustering on an input data matrix or on a similarity matrix of a similarity graph. `spectralcluster` requires that you specify the number of clusters.

However, the algorithm for spectral clustering also provides a way to estimate the number of clusters in your data. For more information, see “Partition Data Using Spectral Clustering” on page 16-26.

Comparison of Clustering Methods

This table compares the features of available clustering methods in Statistics and Machine Learning Toolbox.

Method	Basis of Algorithm	Input to Algorithm	Requires Specified Number of Clusters	Cluster Shapes Identified	Useful for Outlier Detection
“Hierarchical Clustering” on page 16-6	Distance between objects	Pairwise distances between observations	No	Arbitrarily shaped clusters, depending on the specified `'Linkage'` algorithm	No
“k-Means Clustering” on page 16-33 and k-Medoids Clustering on page 33-3345	Distance between objects and centroids	Actual observations	Yes	Spheroidal clusters with equal diagonal covariance	No
Density-Based Spatial Clustering of Applications with Noise (“DBSCAN” on page 16-19)	Density of regions in the data	Actual observations or pairwise distances between observations	No	Arbitrarily shaped clusters	Yes
“Gaussian Mixture Models”	Mixture of Gaussian distributions	Actual observations	Yes	Spheroidal clusters with different covariance structures	Yes
Nearest Neighbors	Distance between objects	Actual observations	No	Arbitrarily shaped clusters	Yes, depending on the specified number of neighbors
Spectral Clustering (“Partition Data Using Spectral Clustering” on page 16-26)	Graph representing connections between data points	Actual observations or similarity matrix	Yes, but the algorithm also provides a way to estimate the number of clusters	Arbitrarily shaped clusters	No

See Also

More About

- “Hierarchical Clustering” on page 16-6
- “k-Means Clustering” on page 16-33
- “DBSCAN” on page 16-19
- “Cluster Using Gaussian Mixture Model” on page 16-39
- “Partition Data Using Spectral Clustering” on page 16-26

Hierarchical Clustering

In this section...

Introduction to Hierarchical Clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows you to decide the level or scale of clustering that is most appropriate for your application. The function `clusterdata` supports agglomerative clustering and performs all of the necessary steps for you. It incorporates the `pdist`, `linkage`, and `cluster` functions, which you can use separately for more detailed analysis. The `dendrogram` function plots the cluster tree.

Algorithm Description

To perform agglomerative hierarchical cluster analysis on a data set using Statistics and Machine Learning Toolbox functions, follow this procedure:

1. **Find the similarity or dissimilarity between every pair of objects in the data set.** In this step, you calculate the *distance* between objects using the `pdist` function. The `pdist` function supports many different ways to compute this measurement. See “Similarity Measures” on page 16-7 for more information.
2. **Group the objects into a binary, hierarchical cluster tree.** In this step, you link pairs of objects that are in close proximity using the `linkage` function. The `linkage` function uses the distance information generated in step 1 to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed. See “Linkages” on page 16-8 for more information.
3. **Determine where to cut the hierarchical tree into clusters.** In this step, you use the `cluster` function to prune branches off the bottom of the hierarchical tree, and assign all the objects below each cut to a single cluster. This creates a partition of the data. The `cluster` function can create these clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point.

The following sections provide more information about each of these steps.

Note The function `clusterdata` performs all of the necessary steps for you. You do not need to execute the `pdist`, `linkage`, or `cluster` functions separately.

Similarity Measures

You use the `pdist` function to calculate the distance between every pair of objects in a data set. For a data set made up of *m* objects, there are *m**(*m* – 1)/2 pairs in the data set. The result of this computation is commonly known as a distance or dissimilarity matrix.

There are many ways to calculate this distance information. By default, the `pdist` function calculates the Euclidean distance between objects; however, you can specify one of several other options. See `pdist` for more information.

Note You can optionally normalize the values in the data set before calculating the distance information. In a real world data set, variables can be measured against different scales. For example, one variable can measure Intelligence Quotient (IQ) test scores and another variable can measure head circumference. These discrepancies can distort the proximity calculations. Using the `zscore` function, you can convert all the values in the data set to use the same proportional scale. See `zscore` for more information.

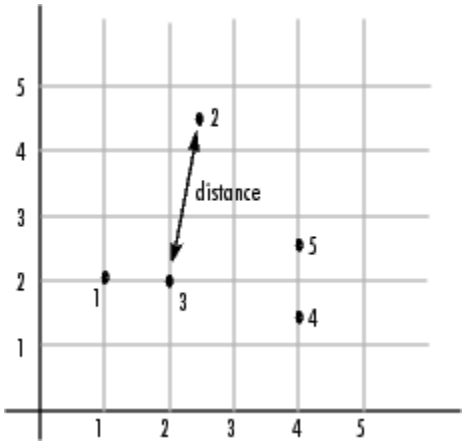
For example, consider a data set, `X`, made up of five objects where each object is a set of *x*,*y* coordinates.

- **Object 1**: 1, 2
- **Object 2**: 2.5, 4.5
- **Object 3**: 2, 2
- **Object 4**: 4, 1.5
- **Object 5**: 4, 2.5

You can define this data set as a matrix

```
rng default;  % For reproducibility
X = [1 2;2.5 4.5;2 2;4 1.5;...
    4 2.5];
```

and pass it to `pdist`. The `pdist` function calculates the distance between object 1 and object 2, object 1 and object 3, and so on until the distances between all the pairs have been calculated. The following figure plots these objects in a graph. The Euclidean distance between object 2 and object 3 is shown to illustrate one interpretation of distance.



Distance Information

The `pdist` function returns this distance information in a vector, `Y`, where each element contains the distance between a pair of objects.

```
Y = pdist(X)
```

```
Y = 1×10

    2.9155    1.0000    3.0414    3.0414    2.5495    3.3541    2.5000    2.0616    2.0616    1.0
```

To make it easier to see the relationship between the distance information generated by `pdist` and the objects in the original data set, you can reformat the distance vector into a matrix using the `squareform` function. In this matrix, element *i*,*j* corresponds to the distance between object *i* and object *j* in the original data set. In the following example, element 1,1 represents the distance between object 1 and itself (which is zero). Element 1,2 represents the distance between object 1 and object 2, and so on.

```
squareform(Y)
```

```
ans = 5×5

         0    2.9155    1.0000    3.0414    3.0414
    2.9155         0    2.5495    3.3541    2.5000
    1.0000    2.5495         0    2.0616    2.0616
    3.0414    3.3541    2.0616         0    1.0000
    3.0414    2.5000    2.0616    1.0000         0
```

Linkages

Once the proximity between objects in the data set has been computed, you can determine how objects in the data set should be grouped into clusters, using the `linkage` function. The `linkage` function takes the distance information generated by `pdist` and links pairs of objects that are close together into binary clusters (clusters made up of two objects). The `linkage` function then links these newly formed clusters to each other and to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree.

For example, given the distance vector `Y` generated by `pdist` from the sample data set of *x*- and *y*-coordinates, the `linkage` function generates a hierarchical cluster tree, returning the linkage information in a matrix, `Z`.

```
Z = linkage(Y)
```

```
Z = 4×3

    4.0000    5.0000    1.0000
    1.0000    3.0000    1.0000
    6.0000    7.0000    2.0616
    2.0000    8.0000    2.5000
```

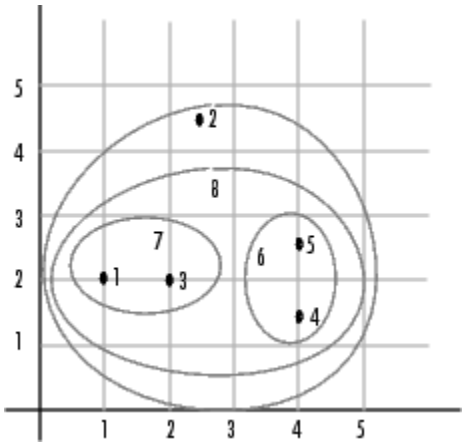
In this output, each row identifies a link between objects or clusters. The first two columns identify the objects that have been linked. The third column contains the distance between these objects. For the sample data set of *x*- and *y*-coordinates, the `linkage` function begins by grouping objects 4

and 5, which have the closest proximity (distance value = 1.0000). The `linkage` function continues by grouping objects 1 and 3, which also have a distance value of 1.0000.

The third row indicates that the `linkage` function grouped objects 6 and 7. If the original sample data set contained only five objects, what are objects 6 and 7? Object 6 is the newly formed binary cluster created by the grouping of objects 4 and 5. When the `linkage` function groups two objects into a new cluster, it must assign the cluster a unique index value, starting with the value *m* + 1, where *m* is the number of objects in the original data set. (Values 1 through *m* are already used by the original data set.) Similarly, object 7 is the cluster formed by grouping objects 1 and 3.

`linkage` uses distances to determine the order in which it clusters objects. The distance vector `Y` contains the distances between the original objects 1 through 5. But linkage must also be able to determine distances involving clusters that it creates, such as objects 6 and 7. By default, `linkage` uses a method known as single linkage. However, there are a number of different methods available. See the `linkage` reference page for more information.

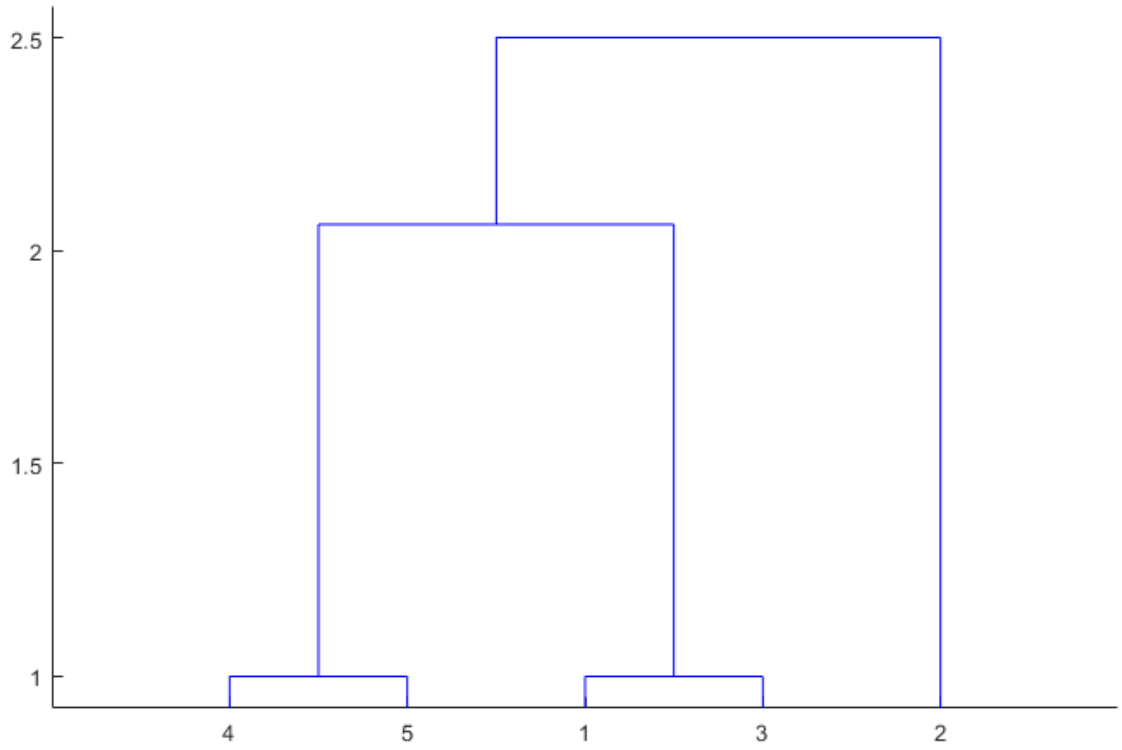
As the final cluster, the `linkage` function grouped object 8, the newly formed cluster made up of objects 6 and 7, with object 2 from the original data set. The following figure graphically illustrates the way `linkage` groups the objects into a hierarchy of clusters.



Dendrograms

The hierarchical, binary cluster tree created by the `linkage` function is most easily understood when viewed graphically. The function `dendrogram` plots the tree as follows.

```
dendrogram(Z)
```



In the figure, the numbers along the horizontal axis represent the indices of the objects in the original data set. The links between objects are represented as upside-down U-shaped lines. The height of the U indicates the distance between the objects. For example, the link representing the cluster containing objects 1 and 3 has a height of 1. The link representing the cluster that groups object 2 together with objects 1, 3, 4, and 5, (which are already clustered as object 8) has a height of 2.5. The height represents the distance `linkage` computes between objects 2 and 8. For more information about creating a dendrogram diagram, see the `dendrogram` reference page.

Verify the Cluster Tree

After linking the objects in a data set into a hierarchical cluster tree, you might want to verify that the distances (that is, heights) in the tree reflect the original distances accurately. In addition, you might want to investigate natural divisions that exist among links between objects. Statistics and Machine Learning Toolbox functions are available for both of these tasks, as described in the following sections.

- “Verify Dissimilarity” on page 16-10
- “Verify Consistency” on page 16-11

Verify Dissimilarity

In a hierarchical cluster tree, any two objects in the original data set are eventually linked together at some level. The height of the link represents the distance between the two clusters that contain those two objects. This height is known as the *cophenetic distance* between the two objects. One way to

measure how well the cluster tree generated by the `linkage` function reflects your data is to compare the cophenetic distances with the original distance data generated by the `pdist` function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The `cophenet` function compares these two sets of values and computes their correlation, returning a value called the *cophenetic correlation coefficient*. The closer the value of the cophenetic correlation coefficient is to 1, the more accurately the clustering solution reflects your data.

You can use the cophenetic correlation coefficient to compare the results of clustering the same data set using different distance calculation methods or clustering algorithms. For example, you can use the `cophenet` function to evaluate the clusters created for the sample data set.

```
c = cophenet(Z,Y)
```

```
c = 0.8615
```

`Z` is the matrix output by the `linkage` function and `Y` is the distance vector output by the `pdist` function.

Execute `pdist` again on the same data set, this time specifying the city block metric. After running the `linkage` function on this new `pdist` output using the average linkage method, call `cophenet` to evaluate the clustering solution.

```
Y = pdist(X,'cityblock');
Z = linkage(Y,'average');
c = cophenet(Z,Y)
```

```
c = 0.9047
```

The cophenetic correlation coefficient shows that using a different distance and linkage method creates a tree that represents the original distances slightly better.

Verify Consistency

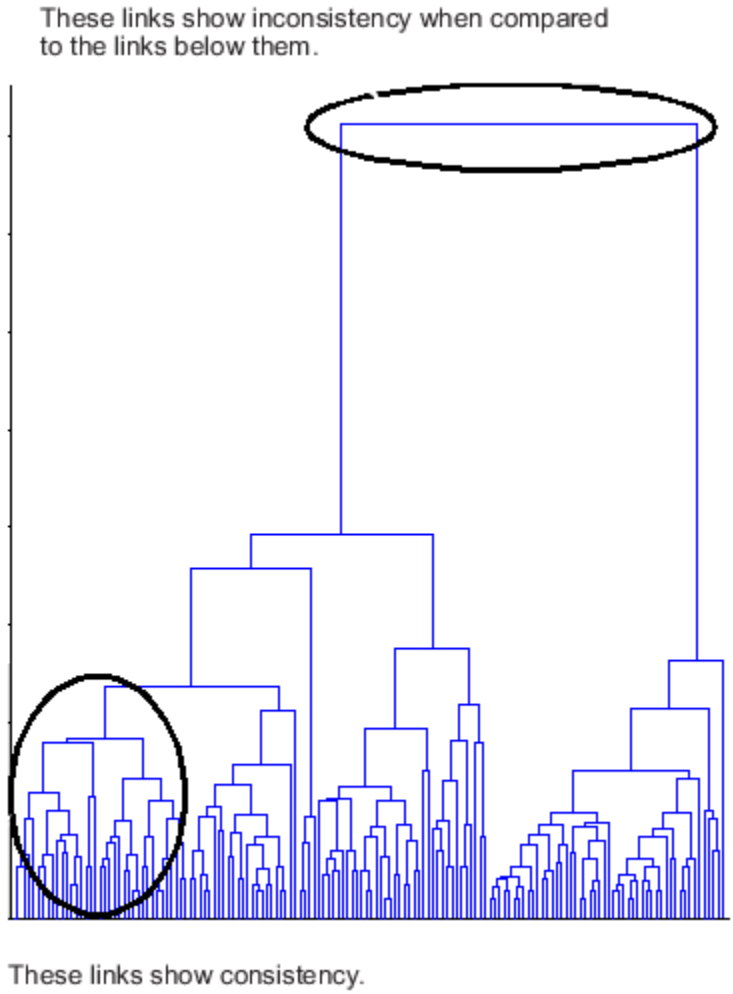
One way to determine the natural cluster divisions in a data set is to compare the height of each link in a cluster tree with the heights of neighboring links below it in the tree.

A link that is approximately the same height as the links below it indicates that there are no distinct divisions between the objects joined at this level of the hierarchy. These links are said to exhibit a high level of consistency, because the distance between the objects being joined is approximately the same as the distances between the objects they contain.

On the other hand, a link whose height differs noticeably from the height of the links below it indicates that the objects joined at this level in the cluster tree are much farther apart from each other than their components were when they were joined. This link is said to be inconsistent with the links below it.

In cluster analysis, inconsistent links can indicate the border of a natural division in a data set. The `cluster` function uses a quantitative measure of inconsistency to determine where to partition your data set into clusters.

The following dendrogram illustrates inconsistent links. Note how the objects in the dendrogram fall into two groups that are connected by links at a much higher level in the tree. These links are inconsistent when compared with the links below them in the hierarchy.



The relative consistency of each link in a hierarchical cluster tree can be quantified and expressed as the *inconsistency coefficient*. This value compares the height of a link in a cluster hierarchy with the average height of links below it. Links that join distinct clusters have a high inconsistency coefficient; links that join indistinct clusters have a low inconsistency coefficient.

To generate a listing of the inconsistency coefficient for each link in the cluster tree, use the `inconsistent` function. By default, the `inconsistent` function compares each link in the cluster hierarchy with adjacent links that are less than two levels below it in the cluster hierarchy. This is called the *depth* of the comparison. You can also specify other depths. The objects at the bottom of the cluster tree, called leaf nodes, that have no further objects below them, have an inconsistency coefficient of zero. Clusters that join two leaves also have a zero inconsistency coefficient.

For example, you can use the `inconsistent` function to calculate the inconsistency values for the links created by the `linkage` function in “Linkages” on page 16-8.

First, recompute the distance and linkage values using the default settings.

```
Y = pdist(X);
Z = linkage(Y);
```

Next, use `inconsistent` to calculate the inconsistency values.

```
I = inconsistent(Z)
```

```
I = 4×4

    1.0000         0    1.0000         0
    1.0000         0    1.0000         0
    1.3539    0.6129    3.0000    1.1547
    2.2808    0.3100    2.0000    0.7071
```

The `inconsistent` function returns data about the links in an (*m*-1)-by-4 matrix, whose columns are described in the following table.

Column	Description
1	Mean of the heights of all the links included in the calculation
2	Standard deviation of all the links included in the calculation
3	Number of links included in the calculation
4	Inconsistency coefficient

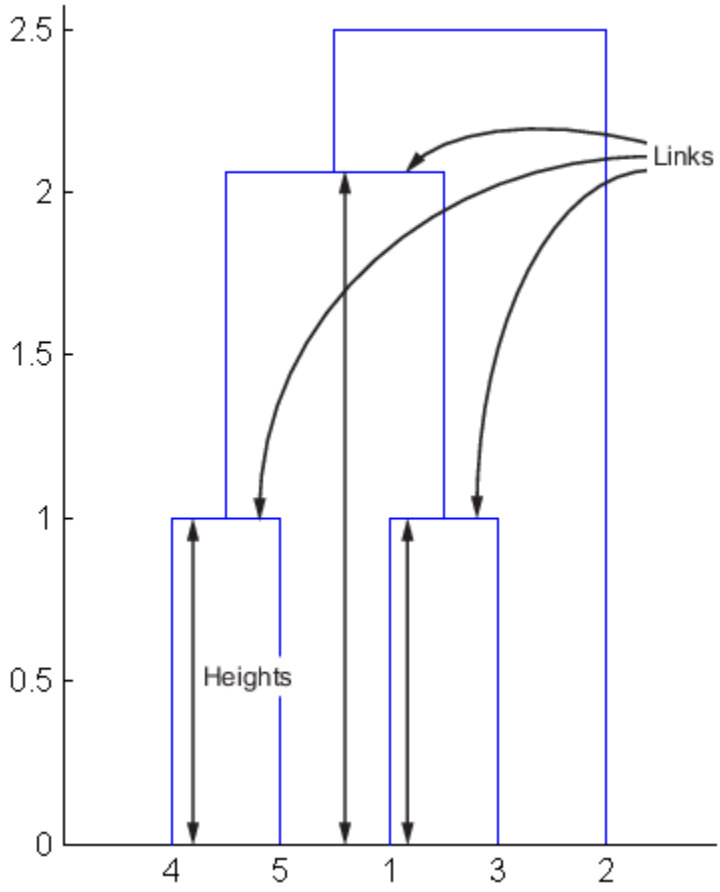
In the sample output, the first row represents the link between objects 4 and 5. This cluster is assigned the index 6 by the `linkage` function. Because both 4 and 5 are leaf nodes, the inconsistency coefficient for the cluster is zero. The second row represents the link between objects 1 and 3, both of which are also leaf nodes. This cluster is assigned the index 7 by the linkage function.

The third row evaluates the link that connects these two clusters, objects 6 and 7. (This new cluster is assigned index 8 in the `linkage` output). Column 3 indicates that three links are considered in the calculation: the link itself and the two links directly below it in the hierarchy. Column 1 represents the mean of the heights of these links. The `inconsistent` function uses the height information output by the `linkage` function to calculate the mean. Column 2 represents the standard deviation between the links. The last column contains the inconsistency value for these links, 1.1547. It is the difference between the current link height and the mean, normalized by the standard deviation.

```
(2.0616 - 1.3539) / .6129
```

```
ans = 1.1547
```

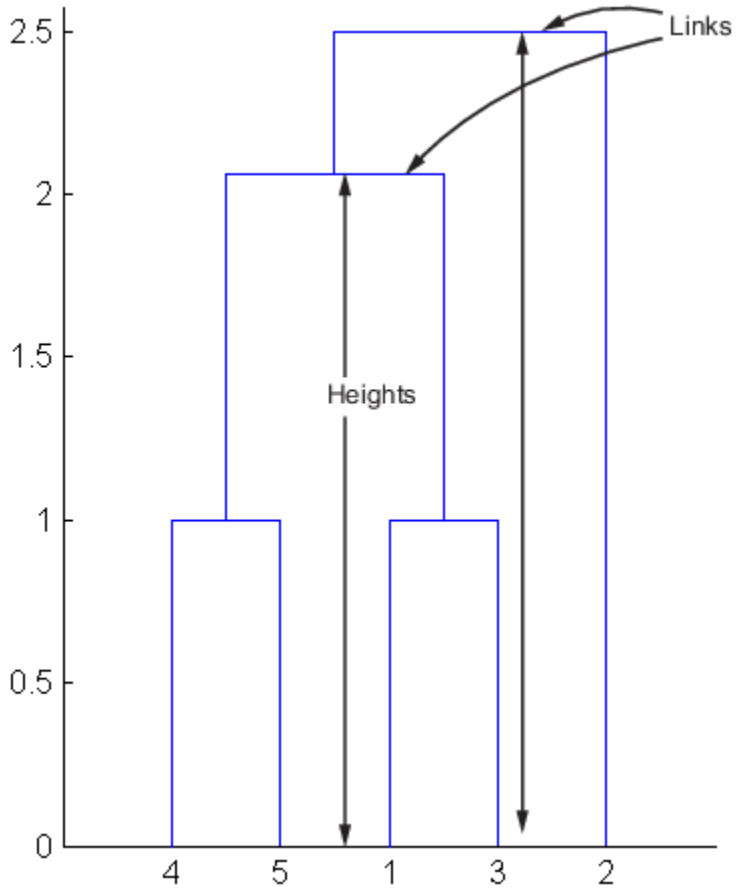
The following figure illustrates the links and heights included in this calculation.



Note In the preceding figure, the lower limit on the *y*-axis is set to `0` to show the heights of the links. To set the lower limit to `0`, select `Axes Properties` from the **Edit** menu, click the **Y Axis** tab, and enter `0` in the field immediately to the right of **Y Limits**.

Row 4 in the output matrix describes the link between object 8 and object 2. Column 3 indicates that two links are included in this calculation: the link itself and the link directly below it in the hierarchy. The inconsistency coefficient for this link is 0.7071.

The following figure illustrates the links and heights included in this calculation.



Create Clusters

After you create the hierarchical tree of binary clusters, you can prune the tree to partition your data into clusters using the `cluster` function. The `cluster` function lets you create clusters in two ways, as discussed in the following sections:

- “Find Natural Divisions in Data” on page 16-15
- “Specify Arbitrary Clusters” on page 16-16

Find Natural Divisions in Data

The hierarchical cluster tree may naturally divide the data into distinct, well-separated clusters. This can be particularly evident in a dendrogram diagram created from data where groups of objects are densely packed in certain areas and not in others. The inconsistency coefficient of the links in the cluster tree can identify these divisions where the similarities between objects change abruptly. (See “Verify the Cluster Tree” on page 16-10 for more information about the inconsistency coefficient.) You can use this value to determine where the `cluster` function creates cluster boundaries.

For example, if you use the `cluster` function to group the sample data set into clusters, specifying an inconsistency coefficient threshold of `1.2` as the value of the `cutoff` argument, the `cluster` function groups all the objects in the sample data set into one cluster. In this case, none of the links in the cluster hierarchy had an inconsistency coefficient greater than `1.2`.

```
T = cluster(Z,'cutoff',1.2)
```

```
T = 5×1

     1
     1
     1
     1
     1
```

The `cluster` function outputs a vector, `T`, that is the same size as the original data set. Each element in this vector contains the number of the cluster into which the corresponding object from the original data set was placed.

If you lower the inconsistency coefficient threshold to `0.8`, the `cluster` function divides the sample data set into three separate clusters.

```
T = cluster(Z,'cutoff',0.8)
```

```
T = 5×1

     3
     2
     3
     1
     1
```

This output indicates that objects 1 and 3 are in one cluster, objects 4 and 5 are in another cluster, and object 2 is in its own cluster.

When clusters are formed in this way, the cutoff value is applied to the inconsistency coefficient. These clusters may, but do not necessarily, correspond to a horizontal slice across the dendrogram at a certain height. If you want clusters corresponding to a horizontal slice of the dendrogram, you can either use the `criterion` option to specify that the cutoff should be based on distance rather than inconsistency, or you can specify the number of clusters directly as described in the following section.

Specify Arbitrary Clusters

Instead of letting the `cluster` function create clusters determined by the natural divisions in the data set, you can specify the number of clusters you want created.

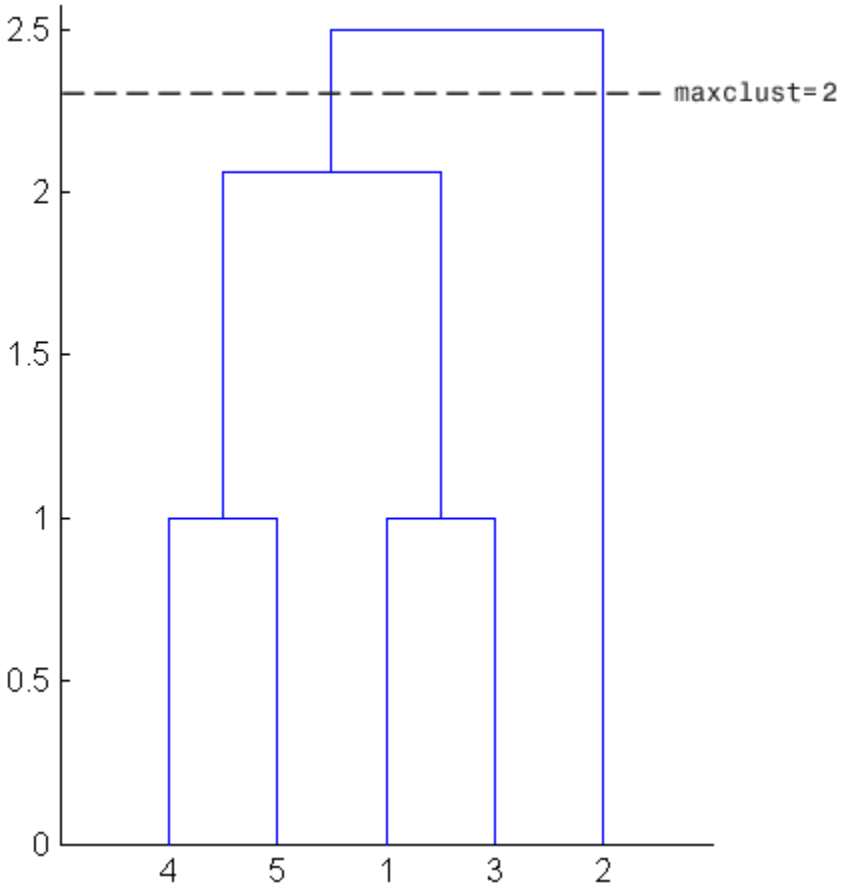
For example, you can specify that you want the `cluster` function to partition the sample data set into two clusters. In this case, the `cluster` function creates one cluster containing objects 1, 3, 4, and 5 and another cluster containing object 2.

```
T = cluster(Z,'maxclust',2)
```

```
T = 5×1

     2
     1
     2
     2
     2
```

To help you visualize how the `cluster` function determines these clusters, the following figure shows the dendrogram of the hierarchical cluster tree. The horizontal dashed line intersects two lines of the

dendrogram, corresponding to setting `'maxclust'` to 2. These two lines partition the objects into two clusters: the objects below the left-hand line, namely 1, 3, 4, and 5, belong to one cluster, while the object below the right-hand line, namely 2, belongs to the other cluster.

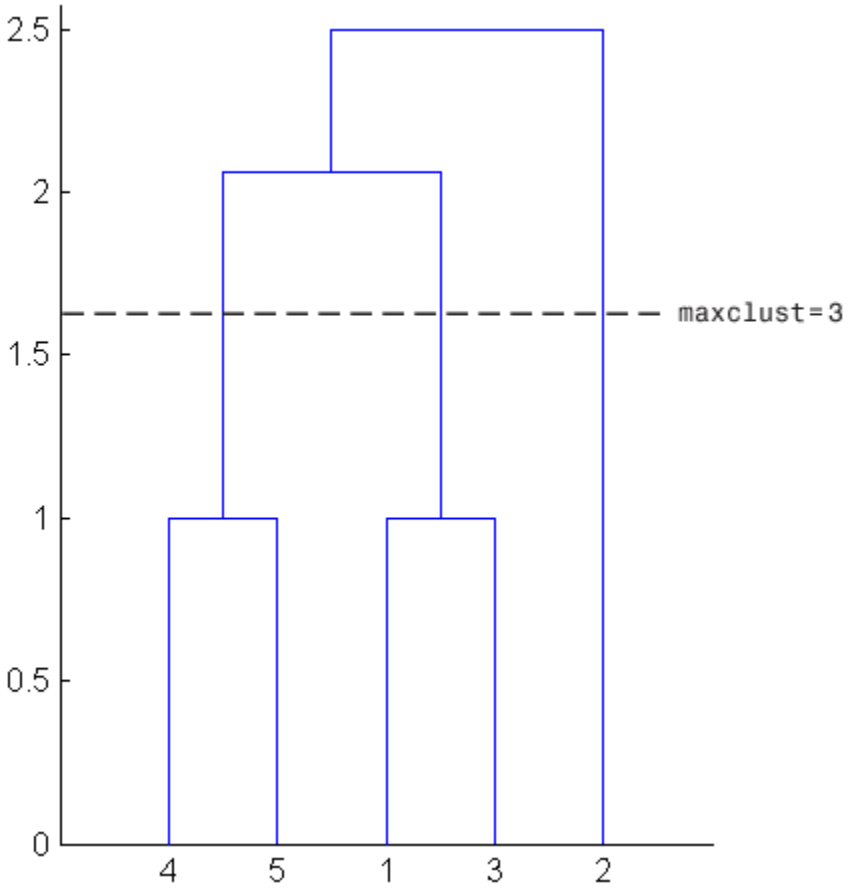


On the other hand, if you set `'maxclust'` to 3, the cluster function groups objects 4 and 5 in one cluster, objects 1 and 3 in a second cluster, and object 2 in a third cluster. The following command illustrates this.

```
T = cluster(Z,'maxclust',3)
```

```
T = 5×1

     2
     3
     2
     1
     1
```

This time, the `cluster` function cuts off the hierarchy at a lower point, corresponding to the horizontal line that intersects three lines of the dendrogram in the following figure.



See Also

More About

- “Choose Cluster Analysis Method” on page 16-2

DBSCAN

In this section...

Introduction to DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies arbitrarily shaped clusters and noise (outliers) in data. The Statistics and Machine Learning Toolbox function `dbscan` performs clustering on an input data matrix or on pairwise distances between observations. `dbscan` returns the cluster indices and a vector indicating the observations that are core points (points inside clusters). Unlike *k*-means clustering, the DBSCAN algorithm does not require prior knowledge of the number of clusters, and clusters are not necessarily spheroidal. DBSCAN is also useful for density-based outlier detection, because it identifies points that do not belong to any cluster.

For a point to be assigned to a cluster, it must satisfy the condition that its epsilon neighborhood (`epsilon`) contains at least a minimum number of neighbors (`minpts`). Or, the point can lie within the epsilon neighborhood of another point that satisfies the `epsilon` and `minpts` conditions. The DBSCAN algorithm identifies three kinds of points:

- Core point — A point in a cluster that has at least `minpts` neighbors in its epsilon neighborhood
- Border point — A point in a cluster that has fewer than `minpts` neighbors in its epsilon neighborhood
- Noise point — An outlier that does not belong to any cluster

DBSCAN works with a wide range of distance metrics on page 33-0 , and you can define a custom distance metric for your particular application. The choice of a distance metric determines the shape of the neighborhood.

Algorithm Description

For specified values of the epsilon neighborhood `epsilon` and the minimum number of neighbors `minpts` required for a core point, the `dbscan` function implements DBSCAN as follows:

1. From the input data set `X`, select the first unlabeled observation x_1 as the current point, and initialize the first cluster label C to 1.
2. Find the set of points within the epsilon neighborhood `epsilon` of the current point. These points are the neighbors.
   a. If the number of neighbors is less than `minpts`, then label the current point as a noise point (or an outlier). Go to step 4.

      Note `dbscan` can reassign noise points to clusters if the noise points later satisfy the constraints set by `epsilon` and `minpts` from some other point in `X`. This process of reassigning points happens for border points of a cluster.

   b. Otherwise, label the current point as a core point belonging to cluster C.

3. Iterate over each neighbor (new current point) and repeat step 2 until no new neighbors are found that can be labeled as belonging to the current cluster *C*.
4. Select the next unlabeled point in X as the current point, and increase the cluster count by 1.
5. Repeat steps 2–4 until all points in X are labeled.

Determine Values for DBSCAN Parameters

This example shows how to select values for the `epsilon` and `minpts` parameters of `dbscan`. The data set is a Lidar scan, stored as a collection of 3-D points, that contains the coordinates of objects surrounding a vehicle.

Load, preprocess, and visualize the data set.

Load the *x*, *y*, *z* coordinates of the objects.

```
load('lidar_subset.mat')
X = lidar_subset;
```

To highlight the environment around the vehicle, set the region of interest to span 20 meters to the left and right of the vehicle, 20 meters in front and back of the vehicle, and the area above the surface of the road.

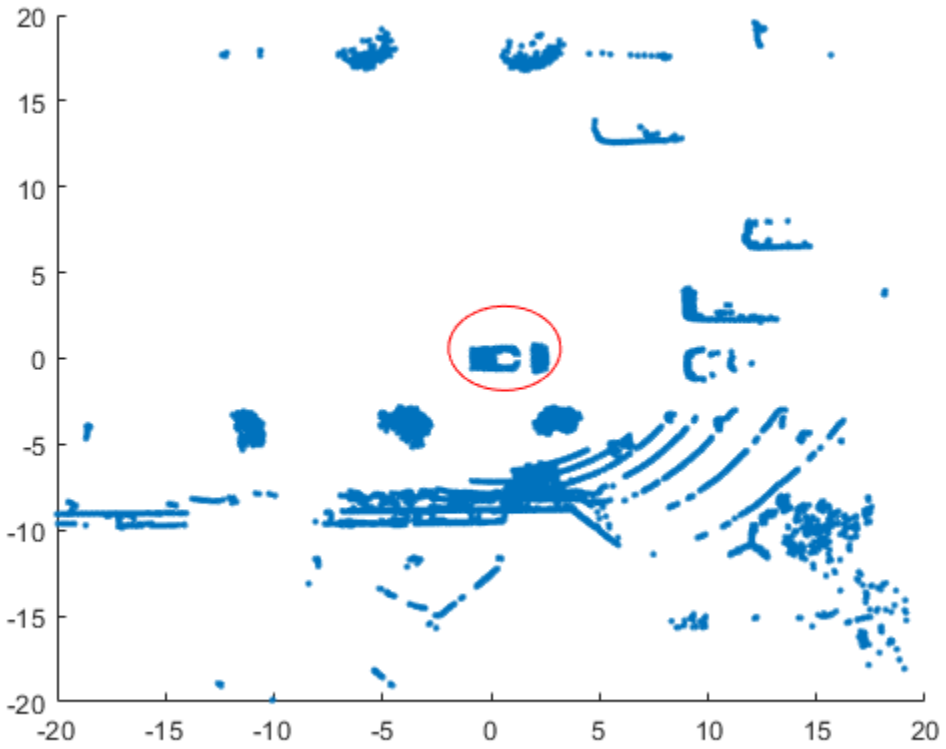
```
xBound  = 20; % in meters
yBound  = 20; % in meters
zLowerBound = 0; % in meters
```

Crop the data to contain only points within the specified region.

```
indices = X(:,1) <= xBound & X(:,1) >= -xBound ...
    & X(:,2) <= yBound & X(:,2) >= -yBound ...
    & X(:,3) > zLowerBound;
X = X(indices,:);
```

Visualize the data as a 2-D scatter plot. Annotate the plot to highlight the vehicle.

```
scatter(X(:,1),X(:,2),'.');
annotation('ellipse',[0.48 0.48 .1 .1],'Color','red')
```



The center of the set of points (circled in red) contains the roof and hood of the vehicle. All other points are obstacles.

Select a value for `minpts`.

To select a value for `minpts`, consider a value greater than or equal to one plus the number of dimensions of the input data [1]. For example, for an `n`-by-`p` matrix `X`, set the value of `'minpts'` greater than or equal to `p+1`.

For the given data set, specify a `minpts` value greater than or equal to 4, specifically the value 50.

```
minpts = 50; % Minimum number of neighbors for a core point
```

Select a value for `epsilon`.

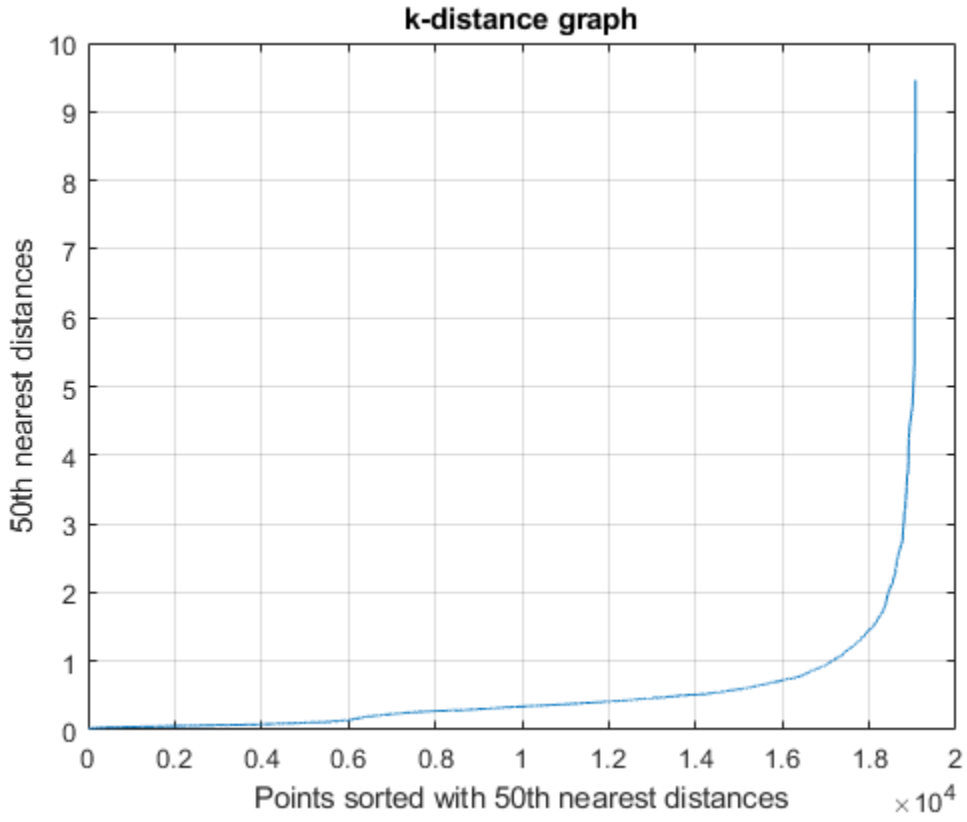
One strategy for estimating a value for `epsilon` is to generate a *k*-distance graph for the input data `X`. For each point in `X`, find the distance to the *k*th nearest point, and plot sorted points against this distance. The graph contains a knee. The distance that corresponds to the knee is generally a good choice for `epsilon`, because it is the region where points start tailing off into outlier (noise) territory [1].

Before plotting the *k*-distance graph, first find the `minpts` smallest pairwise distances for observations in `X`, in ascending order.

```
kD = pdist2(X,X,'euc','Smallest',minpts); % The minpts smallest pairwise distances
```

Plot the *k*-distance graph.

```
plot(sort(kD(end,:)));
title('k-distance graph')
xlabel('Points sorted with 50th nearest distances')
ylabel('50th nearest distances')
grid
```



The knee appears to be around 2; therefore, set the value of `epsilon` to 2.

```
epsilon = 2;
```

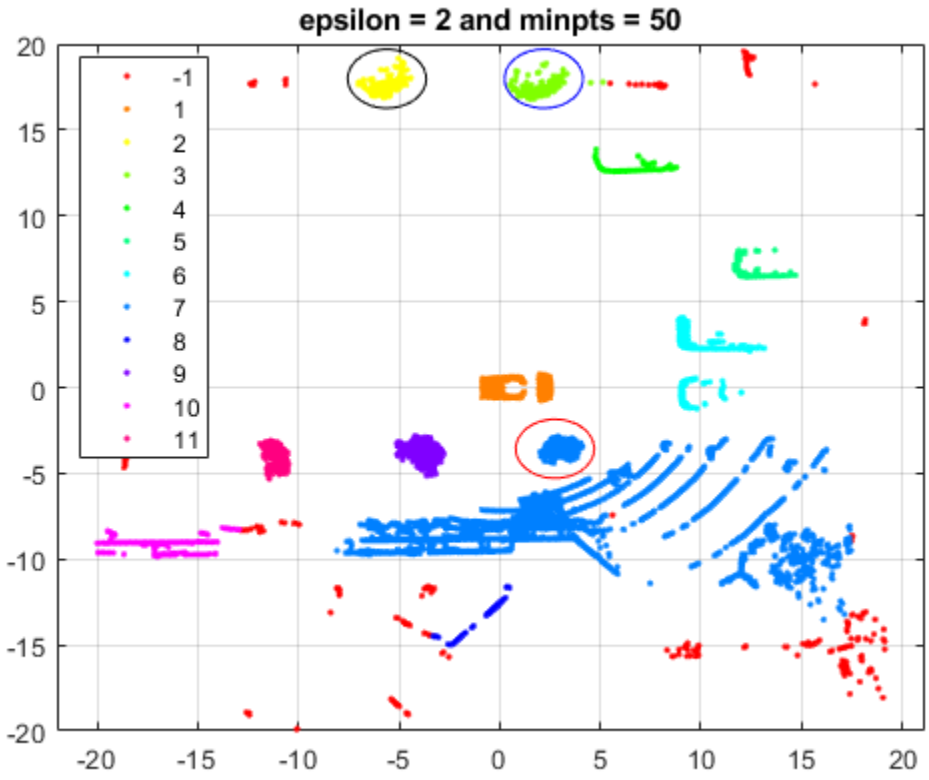
Cluster using dbscan.

Use `dbscan` with the values of `minpts` and `epsilon` that were determined in the previous steps.

```
labels = dbscan(X,epsilon,minpts);
```

Visualize the clustering and annotate the figure to highlight specific clusters.

```
gscatter(X(:,1),X(:,2),labels);
title('epsilon = 2 and minpts = 50')
grid
annotation('ellipse',[0.54 0.41 .07 .07],'Color','red')
annotation('ellipse',[0.53 0.85 .07 .07],'Color','blue')
annotation('ellipse',[0.39 0.85 .07 .07],'Color','black')
```



`dbscan` identifies 11 clusters and a set of noise points. The algorithm also identifies the vehicle at the center of the set of points as a distinct cluster.

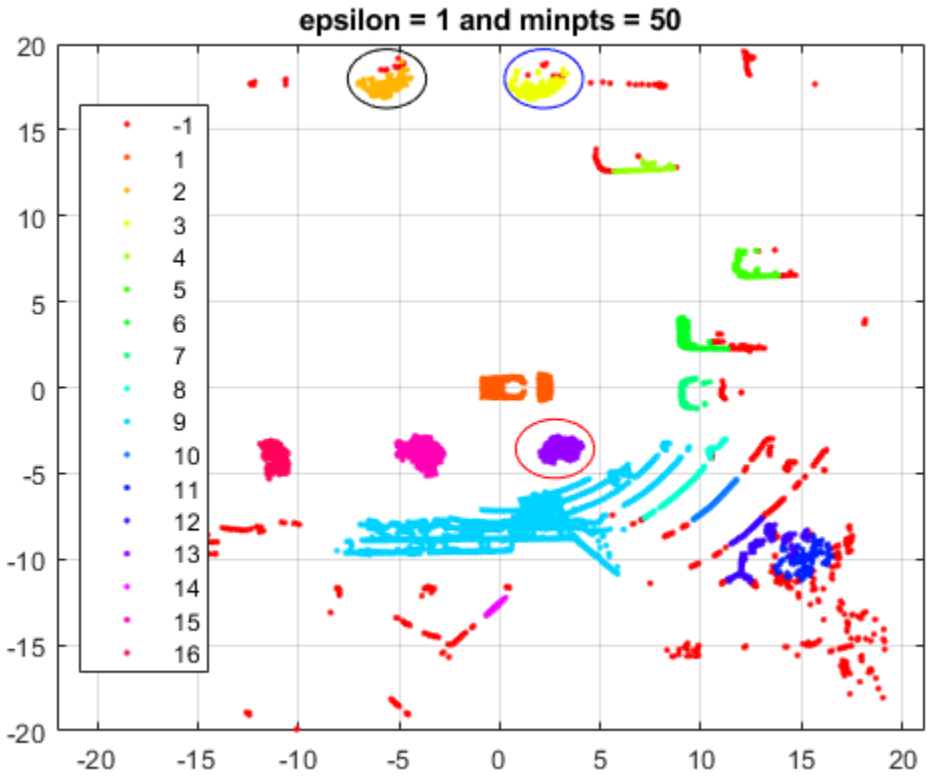
`dbscan` identifies some distinct clusters, such as the cluster circled in black (and centered around (–6,18)) and the cluster circled in blue (and centered around (2.5,18)). The function also assigns the group of points circled in red (and centered around (`3,–4`)) to the same cluster (group 7) as the group of points in the southeast quadrant of the plot. The expectation is that these groups should be in separate clusters.

Use a smaller value for `epsilon` to split up large clusters and further partition the points.

```
epsilon2 = 1;
labels2 = dbscan(X,epsilon2,minpts);
```

Visualize the clustering and annotate the figure to highlight specific clusters.

```
gscatter(X(:,1),X(:,2),labels2);
title('epsilon = 1 and minpts = 50')
grid
annotation('ellipse',[0.54 0.41 .07 .07],'Color','red')
annotation('ellipse',[0.53 0.85 .07 .07],'Color','blue')
annotation('ellipse',[0.39 0.85 .07 .07],'Color','black')
```



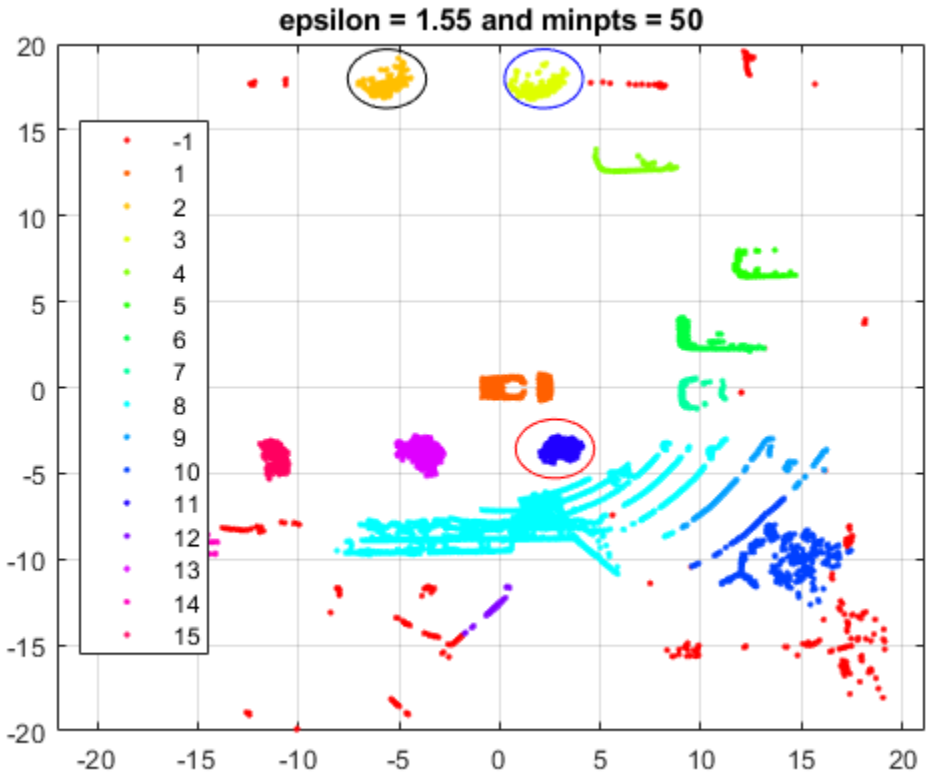
By using a smaller epsilon value, `dbscan` is able to assign the group of points circled in red to a distinct cluster (group 13). However, some clusters that `dbscan` correctly identified before are now split between cluster points and outliers. For example, see cluster group 2 (circled in black) and cluster group 3 (circled in blue). The correct `epsilon` value is somewhere between 1 and 2.

Use an `epsilon` value of `1.55` to cluster the data.

```
epsilon3 = 1.55;
labels3 = dbscan(X,epsilon3,minpts);
```

Visualize the clustering and annotate the figure to highlight specific clusters.

```
gscatter(X(:,1),X(:,2),labels3);
title('epsilon = 1.55 and minpts = 50')
grid
annotation('ellipse',[0.54 0.41 .07 .07],'Color','red')
annotation('ellipse',[0.53 0.85 .07 .07],'Color','blue')
annotation('ellipse',[0.39 0.85 .07 .07],'Color','black')
```



`dbscan` does a better job of identifying the clusters when `epsilon` is set to `1.55`. For example, the function identifies the distinct clusters circled in red, black, and blue (with centers around `(3,–4)`, (–6,18), and (2.5,18), respectively).

References

[1] Ester, M., H.-P. Kriegel, J. Sander, and X. Xiaowei. "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining*, 226-231. Portland, OR: AAAI Press, 1996.

See Also

More About

- "Choose Cluster Analysis Method" on page 16-2

Partition Data Using Spectral Clustering

This topic provides an introduction to spectral clustering and an example that estimates the number of clusters and performs spectral clustering.

Introduction to Spectral Clustering

Spectral clustering is a graph-based algorithm for partitioning data points, or observations, into k clusters. The Statistics and Machine Learning Toolbox function `spectralcluster` performs clustering on an input data matrix or on a similarity matrix on page 33-5958 of a similarity graph derived from the data. `spectralcluster` returns the cluster indices, a matrix containing k eigenvectors of the Laplacian matrix on page 33-5959, and a vector of eigenvalues corresponding to the eigenvectors.

`spectralcluster` requires you to specify the number of clusters k. However, you can verify that your estimate for k is correct by using one of these methods:

- Count the number of zero eigenvalues of the Laplacian matrix. The multiplicity of the zero eigenvalues is an indicator of the number of clusters in your data.
- Find the number of connected components in your similarity matrix by using the MATLAB function `conncomp`.

Algorithm Description

Spectral clustering is a graph-based algorithm for finding k arbitrarily shaped clusters in data. The technique involves representing the data in a low dimension. In the low dimension, clusters in the data are more widely separated, enabling you to use algorithms such as k-means or k-medoids clustering. This low dimension is based on the eigenvectors corresponding to the k smallest eigenvalues of a Laplacian matrix. A Laplacian matrix is one way of representing a similarity graph that models the local neighborhood relationships between data points as an undirected graph. The spectral clustering algorithm derives a similarity matrix of a similarity graph from your data, finds the Laplacian matrix, and uses the Laplacian matrix to find k eigenvectors for splitting the similarity graph into k partitions. You can use spectral clustering when you know the number of clusters, but the algorithm also provides a way to estimate the number of clusters in your data.

By default, the algorithm for `spectralcluster` computes the normalized random-walk Laplacian matrix using the method described by Shi-Malik [1]. `spectralcluster` also supports the unnormalized Laplacian matrix and the normalized symmetric Laplacian matrix which uses the Ng-Jordan-Weiss method [2]. The `spectralcluster` function implements clustering as follows:

1. For each data point in `X`, define a local neighborhood using either the radius search method or nearest neighbor method, as specified by the `'SimilarityGraph'` name-value pair argument (see “Similarity Graph” on page 33-5958). Then, find the pairwise distances $Dist_{i,j}$ for all points i and j in the neighborhood.
2. Convert the distances to similarity measures using the kernel transformation $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$. The matrix S is the similarity matrix on page 33-5958, and σ is the scale factor for the kernel, as specified using the `'KernelScale'` name-value pair argument.
3. Calculate the unnormalized Laplacian matrix on page 33-5959 L , the normalized random-walk Laplacian matrix L_{rw}, or the normalized symmetric Laplacian matrix L_s, depending on the value of the `'LaplacianNormalization'` name-value pair argument.

4. Create a matrix $V \in \mathbb{R}^{n \times k}$ containing columns $v_1, ..., v_k$, where the columns are the k eigenvectors that correspond to the k smallest eigenvalues of the Laplacian matrix. If using L_s, normalize each row of V to have unit length.
5. Treating each row of V as a point, cluster the n points using k-means clustering (default) or k-medoids clustering, as specified by the `'ClusterMethod'` name-value pair argument.
6. Assign the original points in `X` to the same clusters as their corresponding rows in V.

Estimate Number of Clusters and Perform Spectral Clustering

This example demonstrates two approaches for performing spectral clustering.

- The first approach estimates the number of clusters using the eigenvalues of the Laplacian matrix and performs spectral clustering on the data set.
- The second approach estimates the number of clusters using the similarity graph and performs spectral clustering on the similarity matrix.

Generate Sample Data

Randomly generate a sample data set with three well-separated clusters, each containing 20 points.

```
rng('default'); % For reproducibility
n = 20;
X = [randn(n,2)*0.5+3;
     randn(n,2)*0.5;
     randn(n,2)*0.5-3];
```

Perform Spectral Clustering on Data

Estimate the number of clusters in the data by using the eigenvalues of the Laplacian matrix, and perform spectral clustering on the data set.

Compute the five smallest eigenvalues (in magnitude) of the Laplacian matrix by using the `spectralcluster` function. By default, the function uses the normalized random-walk Laplacian matrix.

```
[~,V_temp,D_temp] = spectralcluster(X,5)
```

```
V_temp = 60×5

    0.0000    0.2236   -0.0000   -0.1534   -0.0000
    0.0000    0.2236   -0.0000    0.3093    0.0000
    0.0000    0.2236   -0.0000   -0.2225    0.0000
    0.0000    0.2236   -0.0000   -0.1776   -0.0000
    0.0000    0.2236   -0.0000   -0.1331    0.0000
    0.0000    0.2236   -0.0000   -0.2176   -0.0000
    0.0000    0.2236   -0.0000   -0.1967    0.0000
    0.0000    0.2236   -0.0000    0.0088    0.0000
    0.0000    0.2236   -0.0000    0.2844    0.0000
    0.0000    0.2236   -0.0000    0.3275   -0.0000
      ⋮


D_temp = 5×1

   -0.0000
```

```
   -0.0000
   -0.0000
    0.0876
    0.1653
```

Only the first three eigenvalues are approximately zero. The number of zero eigenvalues is a good indicator of the number of connected components in a similarity graph and, therefore, is a good estimate of the number of clusters in your data. So, `k=3` is a good estimate of the number of clusters in `X`.

Perform spectral clustering on observations by using the `spectralcluster` function. Specify `k=3` clusters.

```
k = 3;
[idx1,V,D] = spectralcluster(X,k)
```

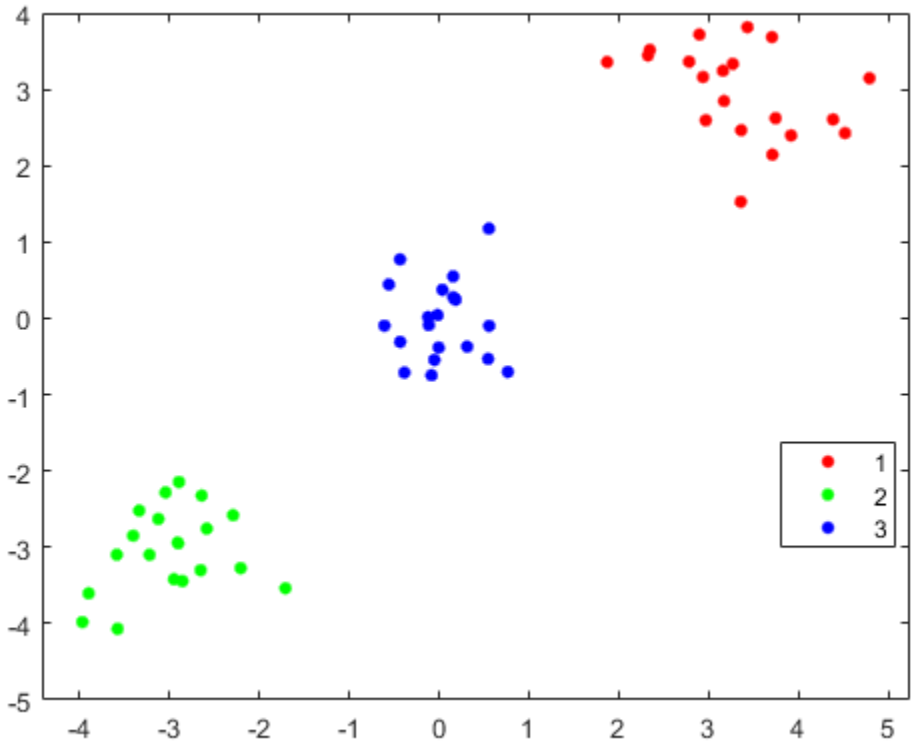
```
idx1 = 60×1

     1
     1
     1
     1
     1
     1
     1
     1
     1
     1
      ⋮


V = 60×3

    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
    0.0000   -0.2236    0.0000
      ⋮


D = 3×1
10^-16 ×

   -0.1031
   -0.1601
   -0.3754
```

Elements of `D` correspond to the three smallest eigenvalues of the Laplacian matrix. The columns of `V` contain the eigenvectors corresponding to the eigenvalues in `D`. For well-separated clusters, the eigenvectors are indicator vectors. The eigenvectors have values of zero (or close to zero) for points

that do not belong to a particular cluster, and nonzero values for points that belong to a particular cluster.

Visualize the result of clustering.

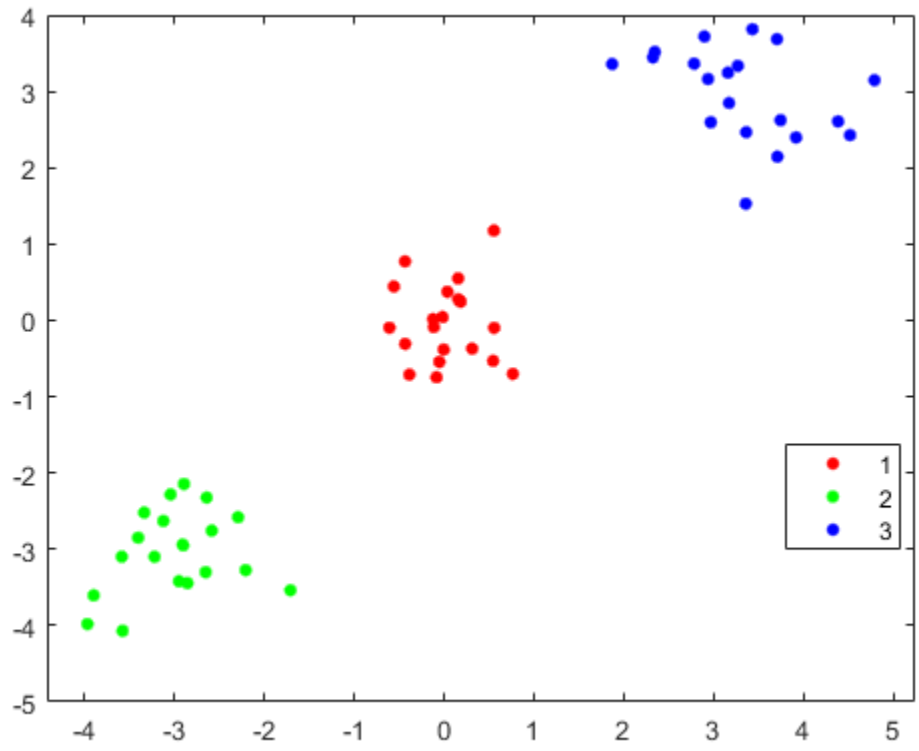
```
gscatter(X(:,1),X(:,2),idx1);
```



The `spectralcluster` function correctly identifies the three clusters in the data set.

Instead of using the `spectralcluster` function again, you can pass `V_temp` to the `kmeans` function to cluster the data points.

```
idx2 = kmeans(V_temp(:,1:3),3);
gscatter(X(:,1),X(:,2),idx2);
```



The order of cluster assignments in `idx1` and `idx2` is different even though the data points are clustered in the same way.

Perform Spectral Clustering on Similarity Matrix

Estimate the number of clusters using the similarity graph and perform spectral clustering on the similarity matrix.

Find the distance between each pair of observations in `X` by using the `pdist` and `squareform` functions with the default Euclidean distance metric.

```
dist_temp = pdist(X);
dist = squareform(dist_temp);
```

Construct the similarity matrix from the pairwise distance and confirm that the similarity matrix is symmetric.

```
S = exp(-dist.^2);
issymmetric(S)
```

```
ans = logical
   1
```

Limit the similarity values to `0.5` so that the similarity graph connects only points whose pairwise distances are smaller than the search radius.

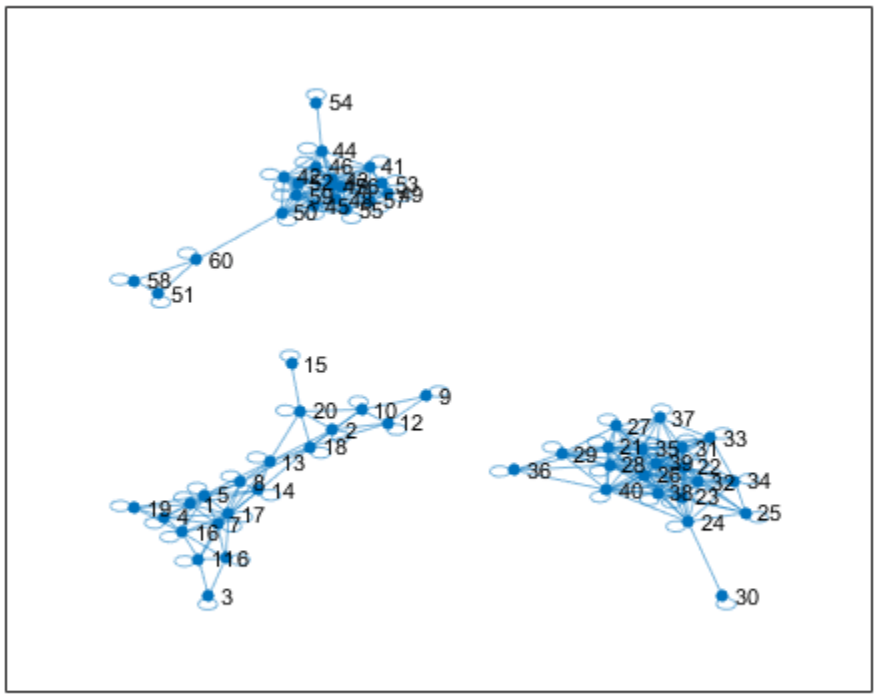
```
S_eps = S;
S_eps(S_eps<0.5) = 0;
```

Create a `graph` object from `S`.

```
G_eps = graph(S_eps);
```

Visualize the similarity graph.

```
plot(G_eps)
```



Identify the number of connected components in graph `G_eps` by using the `unique` and `conncomp` functions.

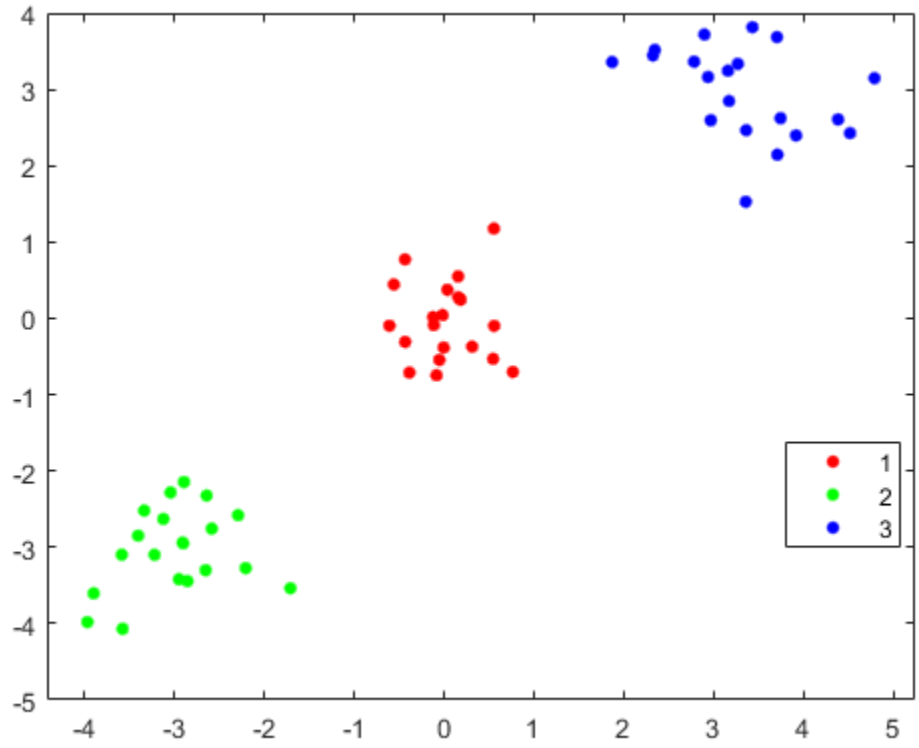
```
unique(conncomp(G_eps))
```

```
ans = 1×3

     1     2     3
```

The similarity graph shows three sets of connected components. The number of connected components in the similarity graph is a good estimate of the number of clusters in your data. Therefore, `k=3` is a good choice for the number of clusters in `X`.

Perform spectral clustering on the similarity matrix derived from the data set `X`.

```
k = 3;
idx3 = spectralcluster(S_eps,k,'Distance','precomputed');
gscatter(X(:,1),X(:,2),idx3);
```



The `spectralcluster` function correctly identifies the three clusters in the data set.

References

[1] Shi, J., and J. Malik. "Normalized cuts and image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 22, 2000, pp. 888–905.

[2] Ng, A.Y., M. Jordan, and Y. Weiss. "On spectral clustering: Analysis and an algorithm." In *Proceedings of the Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.

See Also

`adjacency` | `conncomp` | `pdist` | `pdist2` | `spectralcluster` | `squareform`

More About

- "Choose Cluster Analysis Method" on page 16-2

k-Means Clustering

In this section...

This topic provides an introduction to *k*-means clustering and an example that uses the Statistics and Machine Learning Toolbox function `kmeans` to find the best clustering solution for a data set.

Introduction to k-Means Clustering

k-means clustering is a partitioning method. The function `kmeans` partitions data into *k* mutually exclusive clusters and returns the index of the cluster to which it assigns each observation. `kmeans` treats each observation in your data as an object that has a location in space. The function finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. You can choose a distance metric on page 33-0 to use with `kmeans` based on attributes of your data. Like many clustering methods, *k*-means clustering requires you to specify the number of clusters *k* before clustering.

Unlike hierarchical clustering, *k*-means clustering operates on actual observations rather than the dissimilarity between every pair of observations in the data. Also, *k*-means clustering creates a single level of clusters, rather than a multilevel hierarchy of clusters. Therefore, *k*-means clustering is often more suitable than hierarchical clustering for large amounts of data.

Each cluster in a *k*-means partition consists of member objects and a centroid (or center). In each cluster, `kmeans` minimizes the sum of the distances between the centroid and all member objects of the cluster. `kmeans` computes centroid clusters differently for the supported distance metrics. For details, see `'Distance'`.

You can control the details of the minimization using name-value pair arguments available to `kmeans`; for example, you can specify the initial values of the cluster centroids and the maximum number of iterations for the algorithm. By default, `kmeans` uses the *k*-means++ algorithm on page 33-3329 to initialize cluster centroids, and the squared Euclidean distance metric to determine distances.

When performing *k*-means clustering, follow these best practices:

- Compare *k*-means clustering solutions for different values of *k* to determine an optimal number of clusters for your data.
- Evaluate clustering solutions by examining silhouette plots and silhouette values. You can also use the `evalclusters` function to evaluate clustering solutions based on criteria such as gap values, silhouette values, Davies-Bouldin index values, and Calinski-Harabasz index values.
- Replicate clustering from different randomly selected centroids and return the solution with the lowest total sum of distances among all the replicates.

Compare k-Means Clustering Solutions

This example explores *k*-means clustering on a four-dimensional data set. The example shows how to determine the correct number of clusters for the data set by using silhouette plots and values to analyze the results of different *k*-means clustering solutions. The example also shows how to use the `'Replicates'` name-value pair argument to test a specified number of possible solutions and return the one with the lowest total sum of distances.

Load Data Set

Load the `kmeansdata` data set.

```
rng('default')  % For reproducibility
load('kmeansdata.mat')
size(X)
```

```
ans = 1×2

   560     4
```

The data set is four-dimensional and cannot be visualized easily. However, `kmeans` enables you to investigate whether a group structure exists in the data.

Create Clusters and Examine Separation

Partition the data set into three clusters using *k*-means clustering. Specify the city block distance metric, and use the default *k*-means++ algorithm for cluster center initialization. Use the `'Display'` name-value pair argument to print the final sum of distances for the solution.

```
[idx3,C,sumdist3] = kmeans(X,3,'Distance','cityblock','Display','final');
```

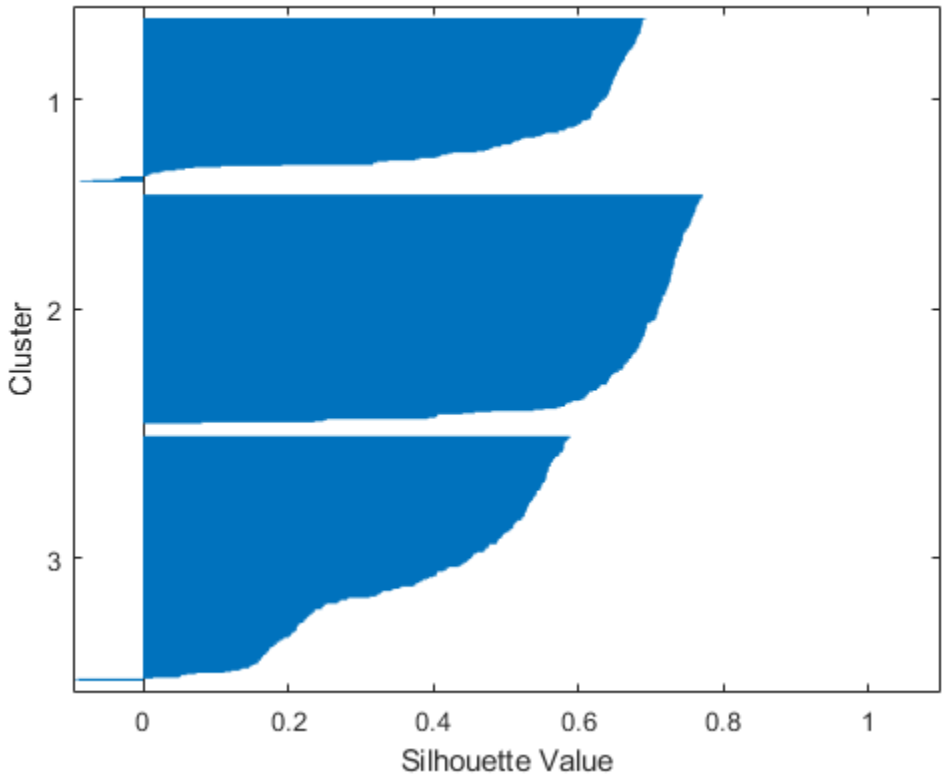
```
Replicate 1, 7 iterations, total sum of distances = 2459.98.
Best total sum of distances = 2459.98
```

`idx3` contains cluster indices that indicate the cluster assignment for each row in `X`. To see if the resulting clusters are well separated, you can create a silhouette plot.

A silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from 1 (indicating points that are very distant from neighboring clusters) through 0 (points that are not distinctly in one cluster or another) to –1 (points that are probably assigned to the wrong cluster). `silhouette` returns these values in its first output.

Create a silhouette plot from `idx3`. Specify `'cityblock'` for the distance metric to indicate that the *k*-means clustering is based on the sum of absolute differences.

```
[silh3,h] = silhouette(X,idx3,'cityblock');
xlabel('Silhouette Value')
ylabel('Cluster')
```



The silhouette plot shows that most points in the second cluster have a large silhouette value (greater than 0.6), indicating that the cluster is somewhat separated from neighboring clusters. However, the third cluster contains many points with low silhouette values, and the first and third clusters contain a few points with negative values, indicating that these two clusters are not well separated.

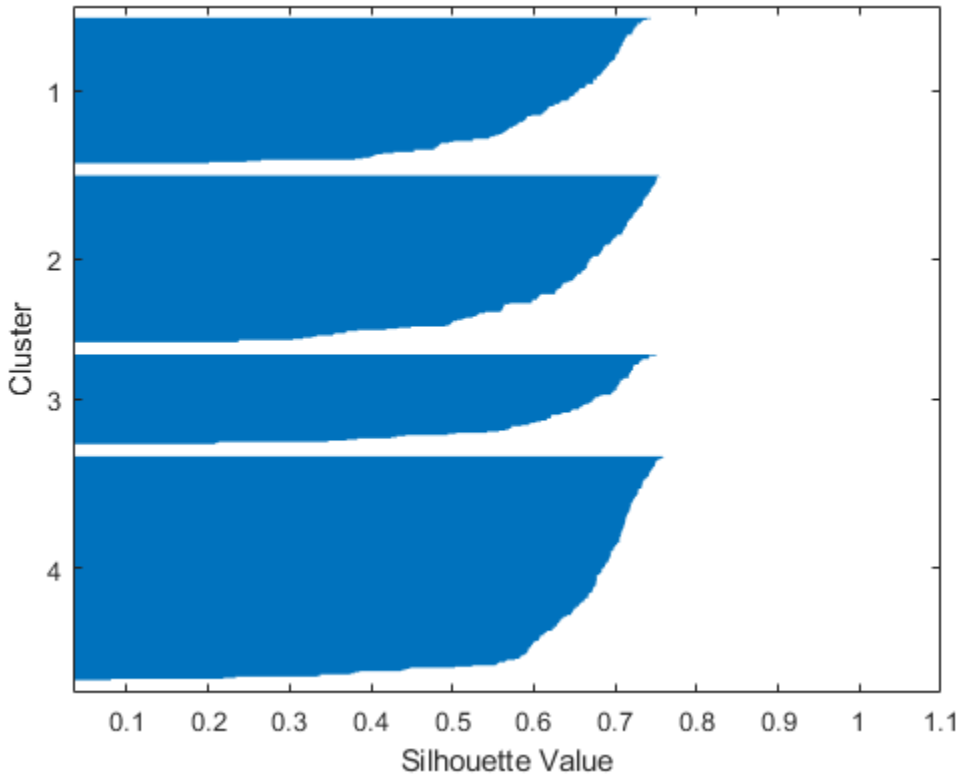
To see if `kmeans` can find a better grouping of the data, increase the number of clusters to four. Print information about each iteration by using the `'Display'` name-value pair argument.

```
idx4 = kmeans(X,4,'Distance','cityblock','Display','iter');
```

```
  iter     phase         num             sum
     1         1         560         1792.72
     2         1           6          1771.1
Best total sum of distances = 1771.1
```

Create a silhouette plot for the four clusters.

```
[silh4,h] = silhouette(X,idx4,'cityblock');
xlabel('Silhouette Value')
ylabel('Cluster')
```



The silhouette plot indicates that these four clusters are better separated than the three clusters in the previous solution. You can take a more quantitative approach to comparing the two solutions by computing the average silhouette values for the two cases.

Compute the average silhouette values.

```
cluster3 = mean(silh3)
```

```
cluster3 = 0.5352
```

```
cluster4 = mean(silh4)
```

```
cluster4 = 0.6400
```

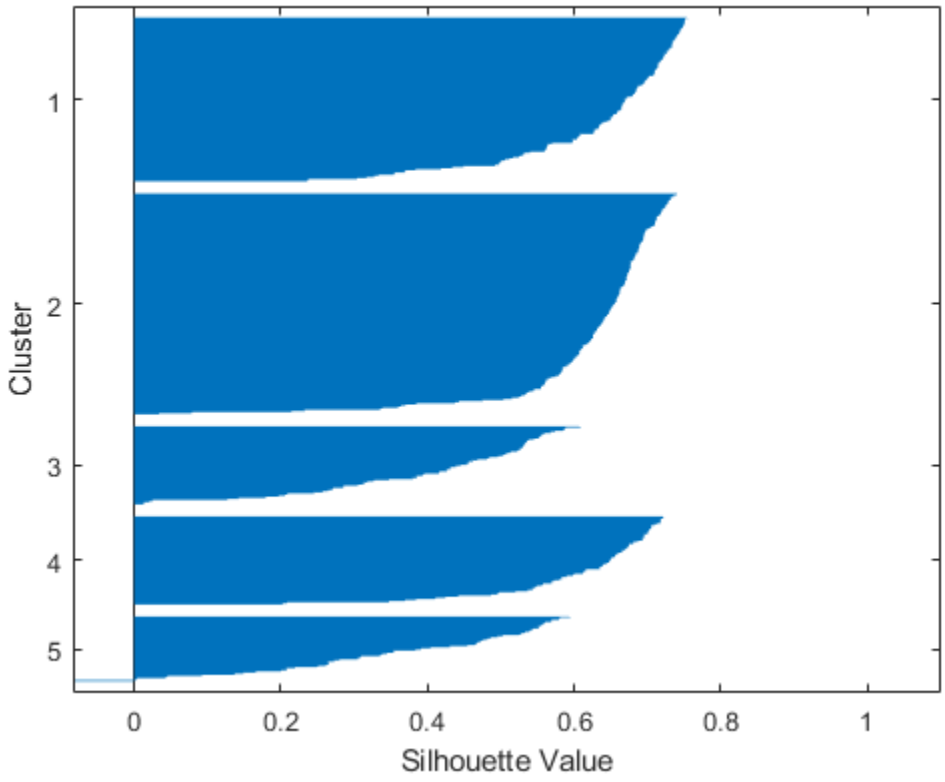
The average silhouette value of the four clusters is higher than the average value of the three clusters. These values support the conclusion represented in the silhouette plots.

Finally, find five clusters in the data. Create a silhouette plot and compute the average silhouette values for the five clusters.

```
idx5 = kmeans(X,5,'Distance','cityblock','Display','final');
```

```
Replicate 1, 7 iterations, total sum of distances = 1647.26.
Best total sum of distances = 1647.26
```

```
[silh5,h] = silhouette(X,idx5,'cityblock');
xlabel('Silhouette Value')
ylabel('Cluster')
```




```
mean(silh5)
```

```
ans = 0.5721
```

The silhouette plot indicates that five is probably not the right number of clusters, because two clusters contain points with mostly low silhouette values, and the fifth cluster contains a few points with negative values. Also, the average silhouette value for the five clusters is lower than the value for the four clusters. Without knowing how many clusters are in the data, it is a good idea to experiment with a range of values for `k`, the number of clusters.

Note that the sum of distances decreases as the number of clusters increases. For example, the sum of distances decreases from `2459.98` to `1771.1` to `1647.26` as the number of clusters increases from 3 to 4 to 5. Therefore, the sum of distances is not useful for determining the optimal number of clusters.

Avoid Local Minima

By default, `kmeans` begins the clustering process using a randomly selected set of initial centroid locations. The `kmeans` algorithm can converge to a solution that is a local (nonglobal) minimum; that is, `kmeans` can partition the data such that moving any single point to a different cluster increases the total sum of distances. However, as with many other types of numerical minimizations, the solution that `kmeans` reaches sometimes depends on the starting points. Therefore, other solutions (local minima) that have lower total sums of distances can exist for the data. You can use the `'Replicates'` name-value pair argument to test different solutions. When you specify more than one replicate, `kmeans` repeats the clustering process starting from different randomly selected

centroids for each replicate, and returns the solution with the lowest total sum of distances among all the replicates.

Find four clusters in the data and replicate the clustering five times. Also, specify the city block distance metric, and use the `'Display'` name-value pair argument to print the final sum of distances for each solution.

```
[idx4,cent4,sumdist] = kmeans(X,4,'Distance','cityblock', ...
                       'Display','final','Replicates',5);
```

```
Replicate 1, 2 iterations, total sum of distances = 1771.1.
Replicate 2, 3 iterations, total sum of distances = 1771.1.
Replicate 3, 3 iterations, total sum of distances = 1771.1.
Replicate 4, 6 iterations, total sum of distances = 2300.23.
Replicate 5, 2 iterations, total sum of distances = 1771.1.
Best total sum of distances = 1771.1
```

In replicate 4, `kmeans` finds a local minimum. Because each replicate begins from a different randomly selected set of initial centroids, `kmeans` sometimes finds more than one local minimum. However, the final solution that `kmeans` returns is the one with the lowest total sum of distances over all replicates.

Find the total of the within-cluster sums of point-to-centroid distances for the final solution returned by `kmeans`.

```
sum(sumdist)
```

```
ans = 1.7711e+03
```

See Also

`kmeans` | `silhouette`

More About

- “Choose Cluster Analysis Method” on page 16-2

Cluster Using Gaussian Mixture Model

This topic provides an introduction to clustering with a Gaussian mixture model (GMM) using the Statistics and Machine Learning Toolbox function `cluster`, and an example that shows the effects of specifying optional parameters when fitting the GMM model using `fitgmdist`.

How Gaussian Mixture Models Cluster Data

Gaussian mixture models (GMMs) are often used for data clustering. You can use GMMs to perform either *hard* clustering or *soft* clustering on query data.

To perform *hard* clustering, the GMM assigns query data points to the multivariate normal components that maximize the component posterior probability, given the data. That is, given a fitted GMM, `cluster` assigns query data to the component yielding the highest posterior probability. Hard clustering assigns a data point to exactly one cluster. For an example showing how to fit a GMM to data, cluster using the fitted model, and estimate component posterior probabilities, see “Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46.

Additionally, you can use a GMM to perform a more flexible clustering on data, referred to as *soft* (or *fuzzy*) clustering. Soft clustering methods assign a score to a data point for each cluster. The value of the score indicates the association strength of the data point to the cluster. As opposed to hard clustering methods, soft clustering methods are flexible because they can assign a data point to more than one cluster. When you perform GMM clustering, the score is the posterior probability. For an example of soft clustering with a GMM, see “Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52.

GMM clustering can accommodate clusters that have different sizes and correlation structures within them. Therefore, in certain applications,, GMM clustering can be more appropriate than methods such as *k*-means clustering. Like many clustering methods, GMM clustering requires you to specify the number of clusters before fitting the model. The number of clusters specifies the number of components in the GMM.

For GMMs, follow these best practices:

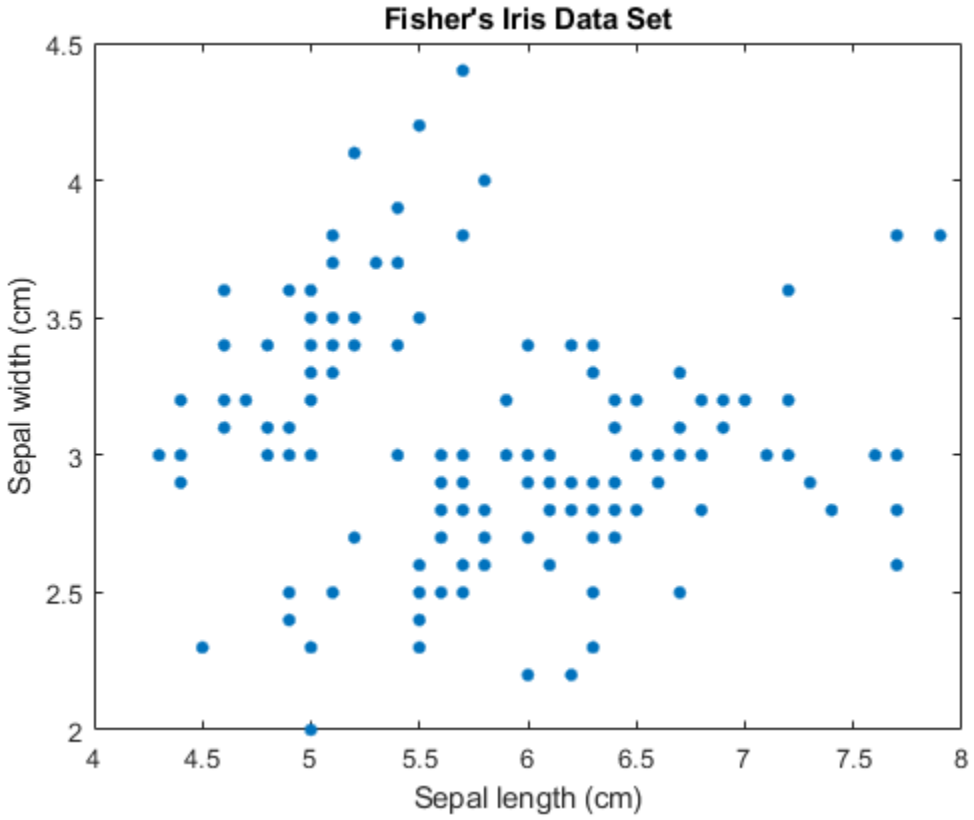
- Consider the component covariance structure. You can specify diagonal or full covariance matrices, and whether all components have the same covariance matrix.
- Specify initial conditions. The Expectation-Maximization (EM) algorithm fits the GMM. As in the *k*-means clustering algorithm, EM is sensitive to initial conditions and might converge to a local optimum. You can specify your own starting values for the parameters, specify initial cluster assignments for data points or let them be selected randomly, or specify use of the *k*-means++ algorithm on page 33-1976.
- Implement regularization. For example, if you have more predictors than data points, then you can regularize for estimation stability.

Fit GMM with Different Covariance Options and Initial Conditions

This example explores the effects of specifying different options for covariance structure and initial conditions when you perform GMM clustering.

Load Fisher's iris data set. Consider clustering the sepal measurements, and visualize the data in 2-D using the sepal measurements.

```
load fisheriris;
X = meas(:,1:2);
[n,p] = size(X);

plot(X(:,1),X(:,2),'.','MarkerSize',15);
title('Fisher''s Iris Data Set');
xlabel('Sepal length (cm)');
ylabel('Sepal width (cm)');
```



The number of components k in a GMM determines the number of subpopulations, or clusters. In this figure, it is difficult to determine if two, three, or perhaps more Gaussian components are appropriate. A GMM increases in complexity as k increases.

Specify Different Covariance Structure Options

Each Gaussian component has a covariance matrix. Geometrically, the covariance structure determines the shape of a confidence ellipsoid drawn over a cluster. You can specify whether the covariance matrices for all components are diagonal or full, and whether all components have the same covariance matrix. Each combination of specifications determines the shape and orientation of the ellipsoids.

Specify three GMM components and 1000 maximum iterations for the EM algorithm. For reproducibility, set the random seed.

```
rng(3);
k = 3; % Number of GMM components
options = statset('MaxIter',1000);
```

Specify covariance structure options.

```
Sigma = {'diagonal','full'}; % Options for covariance matrix type
nSigma = numel(Sigma);

SharedCovariance = {true,false}; % Indicator for identical or nonidentical covariance matrices
SCtext = {'true','false'};
nSC = numel(SharedCovariance);
```

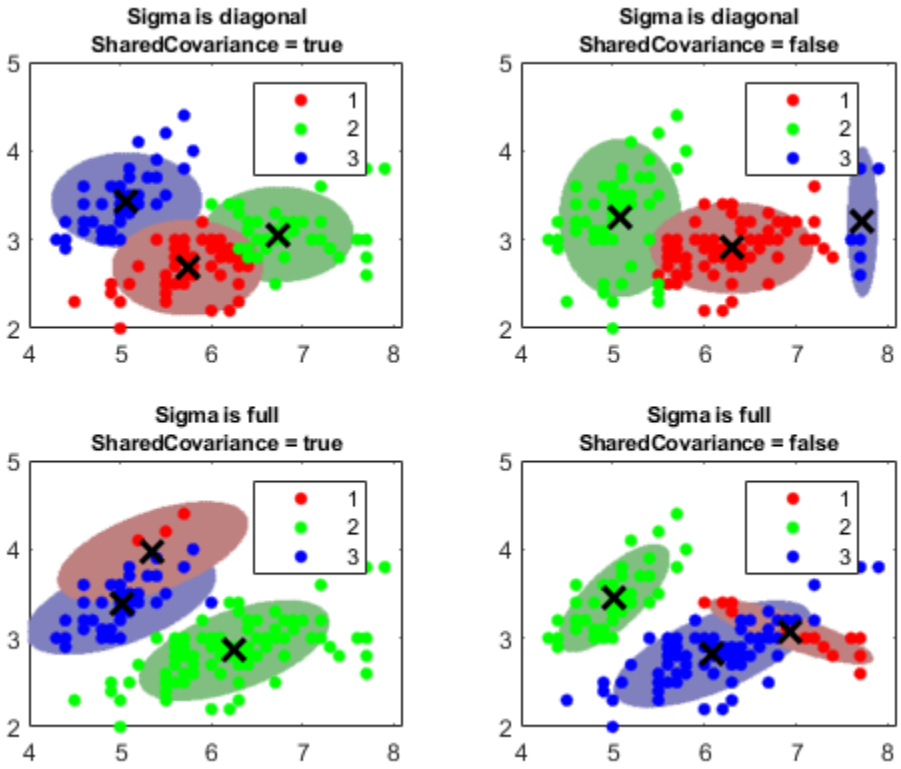
Create a 2-D grid covering the plane composed of extremes of the measurements. You will use this grid later to draw confidence ellipsoids over the clusters.

```
d = 500; % Grid length
x1 = linspace(min(X(:,1))-2, max(X(:,1))+2, d);
x2 = linspace(min(X(:,2))-2, max(X(:,2))+2, d);
[x1grid,x2grid] = meshgrid(x1,x2);
X0 = [x1grid(:) x2grid(:)];
```

Specify the following:

- For all combinations of the covariance structure options, fit a GMM with three components.
- Use the fitted GMM to cluster the 2-D grid.
- Obtain the score that specifies a 99% probability threshold for each confidence region. This specification determines the length of the major and minor axes of the ellipsoids.
- Color each ellipsoid using a similar color as its cluster.

```
threshold = sqrt(chi2inv(0.99,2));
count = 1;
for i = 1:nSigma
    for j = 1:nSC
        gmfit = fitgmdist(X,k,'CovarianceType',Sigma{i}, ...
            'SharedCovariance',SharedCovariance{j},'Options',options); % Fitted GMM
        clusterX = cluster(gmfit,X); % Cluster index
        mahalDist = mahal(gmfit,X0); % Distance from each grid point to each GMM component
        % Draw ellipsoids over each GMM component and show clustering result.
        subplot(2,2,count);
        h1 = gscatter(X(:,1),X(:,2),clusterX);
        hold on
            for m = 1:k
                idx = mahalDist(:,m)<=threshold;
                Color = h1(m).Color*0.75 - 0.5*(h1(m).Color - 1);
                h2 = plot(X0(idx,1),X0(idx,2),'.','Color',Color,'MarkerSize',1);
                uistack(h2,'bottom');
            end
        plot(gmfit.mu(:,1),gmfit.mu(:,2),'kx','LineWidth',2,'MarkerSize',10)
        title(sprintf('Sigma is %s\nSharedCovariance = %s',Sigma{i},SCtext{j}),'FontSize',8)
        legend(h1,{'1','2','3'})
        hold off
        count = count + 1;
    end
end
```



The probability threshold for the confidence region determines the length of the major and minor axes, and the covariance type determines the orientation of the axes. Note the following about options for the covariance matrices:

- **Diagonal covariance matrices** indicate that the predictors are uncorrelated. The major and minor axes of the ellipses are parallel or perpendicular to the x and y axes. This specification increases the total number of parameters by p, the number of predictors, for each component, but is more parsimonious than the full covariance specification.
- **Full covariance matrices** allow for correlated predictors with no restriction to the orientation of the ellipses relative to the x and y axes. Each component increases the total number of parameters by $p(p + 1)/2$, but captures the correlation structure among the predictors. This specification can cause overfitting.
- **Shared covariance matrices** indicate that all components have the same covariance matrix. All ellipses are the same size and have the same orientation. This specification is more parsimonious than the unshared specification because the total number of parameters increases by the number of covariance parameters for one component only.
- **Unshared covariance matrices** indicate that each component has its own covariance matrix. The size and orientation of all ellipses might differ. This specification increases the number of parameters by k times the number of covariance parameters for a component, but can capture covariance differences among components.

The figure also shows that `cluster` does not always preserve cluster order. If you cluster several fitted `gmdistribution` models, `cluster` can assign different cluster labels for similar components.

Specify Different Initial Conditions

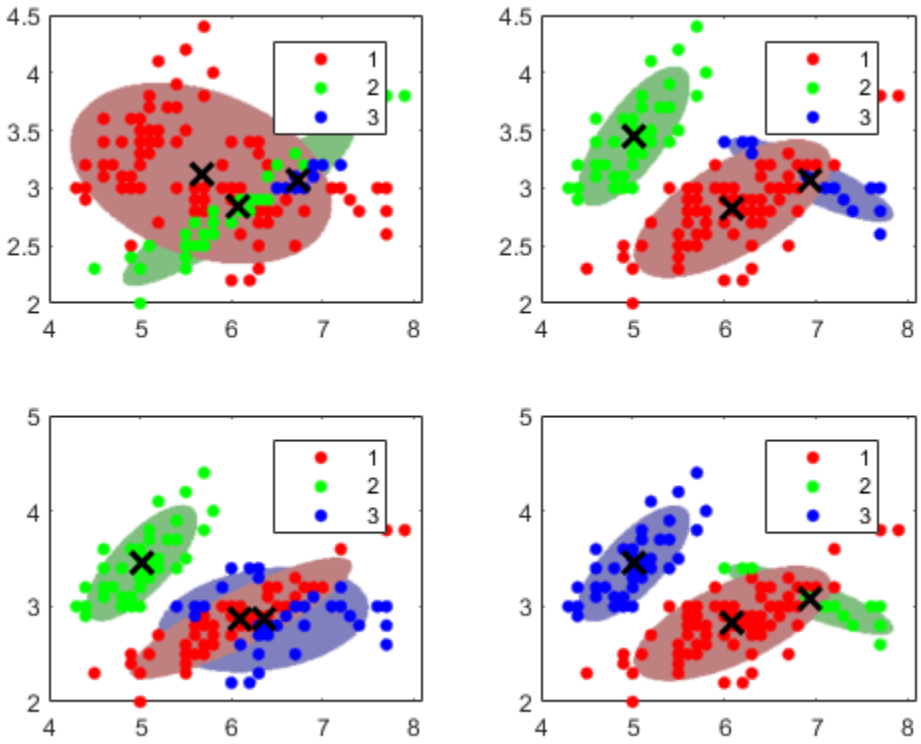
The algorithm that fits a GMM to the data can be sensitive to initial conditions. To illustrate this sensitivity, fit four different GMMs as follows:

1. For the first GMM, assign most data points to the first cluster.
2. For the second GMM, randomly assign data points to clusters.
3. For the third GMM, make another random assignment of data points to clusters.
4. For the fourth GMM, use *k*-means++ to obtain initial cluster centers.

```
initialCond1 = [ones(n-8,1); [2; 2; 2; 2]; [3; 3; 3; 3]]; % For the first GMM
initialCond2 = randsample(1:k,n,true); % For the second GMM
initialCond3 = randsample(1:k,n,true); % For the third GMM
initialCond4 = 'plus'; % For the fourth GMM
cluster0 = {initialCond1; initialCond2; initialCond3; initialCond4};
```

For all instances, use $k = 3$ components, unshared and full covariance matrices, the same initial mixture proportions, and the same initial covariance matrices. For stability when you try different sets of initial values, increase the number of EM algorithm iterations. Also, draw confidence ellipsoids over the clusters.

```
converged = nan(4,1);

for j = 1:4
    gmfit = fitgmdist(X,k,'CovarianceType','full', ...
        'SharedCovariance',false,'Start',cluster0{j}, ...
        'Options',options);
    clusterX = cluster(gmfit,X); % Cluster index
    mahalDist = mahal(gmfit,X0); % Distance from each grid point to each GMM component
    % Draw ellipsoids over each GMM component and show clustering result.
    subplot(2,2,j);
    h1 = gscatter(X(:,1),X(:,2),clusterX); % Distance from each grid point to each GMM component
    hold on;
    nK = numel(unique(clusterX));
    for m = 1:nK
        idx = mahalDist(:,m)<=threshold;
        Color = h1(m).Color*0.75 + -0.5*(h1(m).Color - 1);
        h2 = plot(X0(idx,1),X0(idx,2),'.','Color',Color,'MarkerSize',1);
        uistack(h2,'bottom');
    end
    plot(gmfit.mu(:,1),gmfit.mu(:,2),'kx','LineWidth',2,'MarkerSize',10)
    legend(h1,{'1','2','3'});
    hold off
    converged(j) = gmfit.Converged; % Indicator for convergence
end
```




```
sum(converged)
```

```
ans = 4
```

All algorithms converged. Each starting cluster assignment for the data points leads to a different, fitted cluster assignment. You can specify a positive integer for the name-value pair argument "'Replicates'" on page 33-0 , which runs the algorithm the specified number of times. Subsequently, `fitgmdist` chooses the fit that yields the largest likelihood.

When to Regularize

Sometimes, during an iteration of the EM algorithm, a fitted covariance matrix can become ill conditioned, which means the likelihood is escaping to infinity. This problem can happen if one or more of the following conditions exist:

- You have more predictors than data points.
- You specify fitting with too many components.
- Variables are highly correlated.

To overcome this problem, you can specify a small, positive number using the `'RegularizationValue'` name-value pair argument. `fitgmdist` adds this number to the diagonal elements of all covariance matrices, which ensures that all matrices are positive definite. Regularizing can reduce the maximal likelihood value.

Model Fit Statistics

In most applications, the number of components k and appropriate covariance structure Σ are unknown. One way you can tune a GMM is by comparing information criteria. Two popular information criteria are the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Both the AIC and BIC take the optimized, negative loglikelihood and then penalize it with the number of parameters in the model (the model complexity). However, the BIC penalizes for complexity more severely than the AIC. Therefore, the AIC tends to choose more complex models that might overfit, and the BIC tends to choose simpler models that might underfit. A good practice is to look at both criteria when evaluating a model. Lower AIC or BIC values indicate better fitting models. Also, ensure that your choices for k and the covariance matrix structure are appropriate for your application. `fitgmdist` stores the AIC and BIC of fitted `gmdistribution` model objects in the properties AIC and BIC. You can access these properties by using dot notation. For an example showing how to choose the appropriate parameters, see “Tune Gaussian Mixture Models” on page 16-57.

See Also

`cluster` | `fitgmdist` | `gmdistribution`

More About

- “Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46
- “Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52
- “Tune Gaussian Mixture Models” on page 16-57
- “Choose Cluster Analysis Method” on page 16-2

Cluster Gaussian Mixture Data Using Hard Clustering

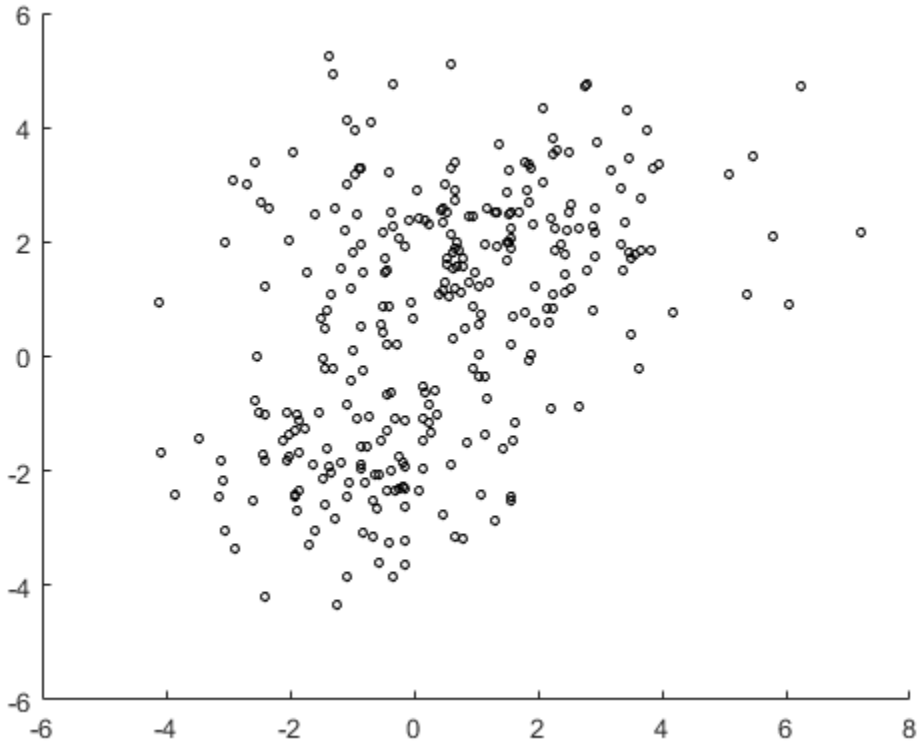
This example shows how to implement hard clustering on simulated data from a mixture of Gaussian distributions.

Gaussian mixture models can be used for clustering data, by realizing that the multivariate normal components of the fitted model can represent clusters.

Simulate Data from a Mixture of Gaussian Distributions

Simulate data from a mixture of two bivariate Gaussian distributions using `mvnrnd`.

```
rng('default')  % For reproducibility
mu1 = [1 2];
sigma1 = [3 .2; .2 2];
mu2 = [-1 -2];
sigma2 = [2 0; 0 1];
X = [mvnrnd(mu1,sigma1,200); mvnrnd(mu2,sigma2,100)];
n = size(X,1);

figure
scatter(X(:,1),X(:,2),10,'ko')
```



Fit a Gaussian Mixture Model to the Simulated Data

Fit a two-component Gaussian mixture model (GMM). Here, you know the correct number of components to use. In practice, with real data, this decision would require comparing models with

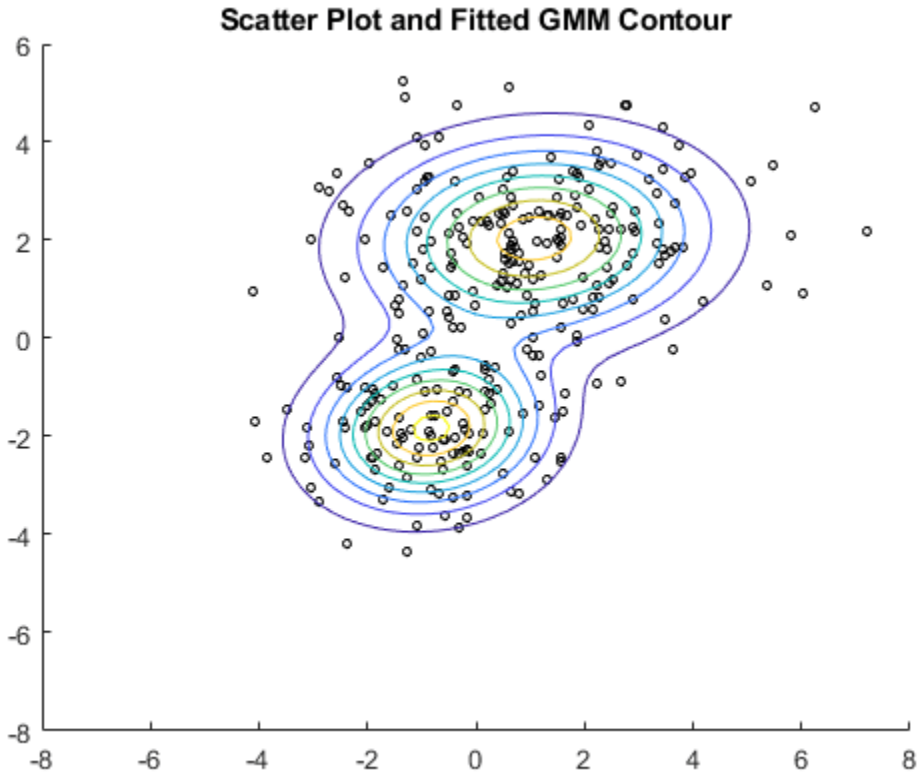
different numbers of components. Also, request to display the final iteration of the expectation-maximization fitting routine.

```
options = statset('Display','final');
gm = fitgmdist(X,2,'Options',options)
```

```
26 iterations, log-likelihood = -1210.59

gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.629514
Mean:    1.0756    2.0421

Component 2:
Mixing proportion: 0.370486
Mean:   -0.8296   -1.8488
```

Plot the estimated probability density contours for the two-component mixture distribution. The two bivariate normal components overlap, but their peaks are distinct. This suggests that the data could reasonably be divided into two clusters.

```
hold on
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0,y0]),x,y);
fcontour(gmPDF,[-8,6])
title('Scatter Plot and Fitted GMM Contour')
hold off
```

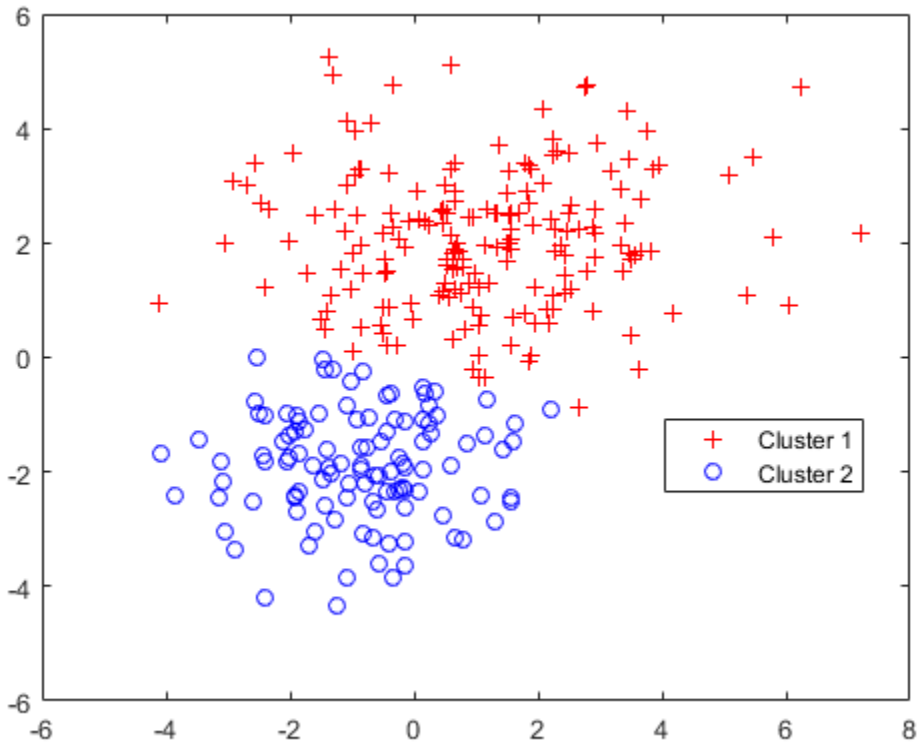


Cluster the Data Using the Fitted GMM

`cluster` implements "hard clustering", a method that assigns each data point to exactly one cluster. For GMM, `cluster` assigns each point to one of the two mixture components in the GMM. The center of each cluster is the corresponding mixture component mean. For details on "soft clustering," see “Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52.

Partition the data into clusters by passing the fitted GMM and the data to `cluster`.

```
idx = cluster(gm,X);
cluster1 = (idx == 1); % |1| for cluster 1 membership
cluster2 = (idx == 2); % |2| for cluster 2 membership

figure
gscatter(X(:,1),X(:,2),idx,'rb','+o')
legend('Cluster 1','Cluster 2','Location','best')
```



Each cluster corresponds to one of the bivariate normal components in the mixture distribution. `cluster` assigns data to clusters based on a cluster membership score. Each cluster membership scores is the estimated posterior probability that the data point came from the corresponding component. `cluster` assigns each point to the mixture component corresponding to the highest posterior probability.

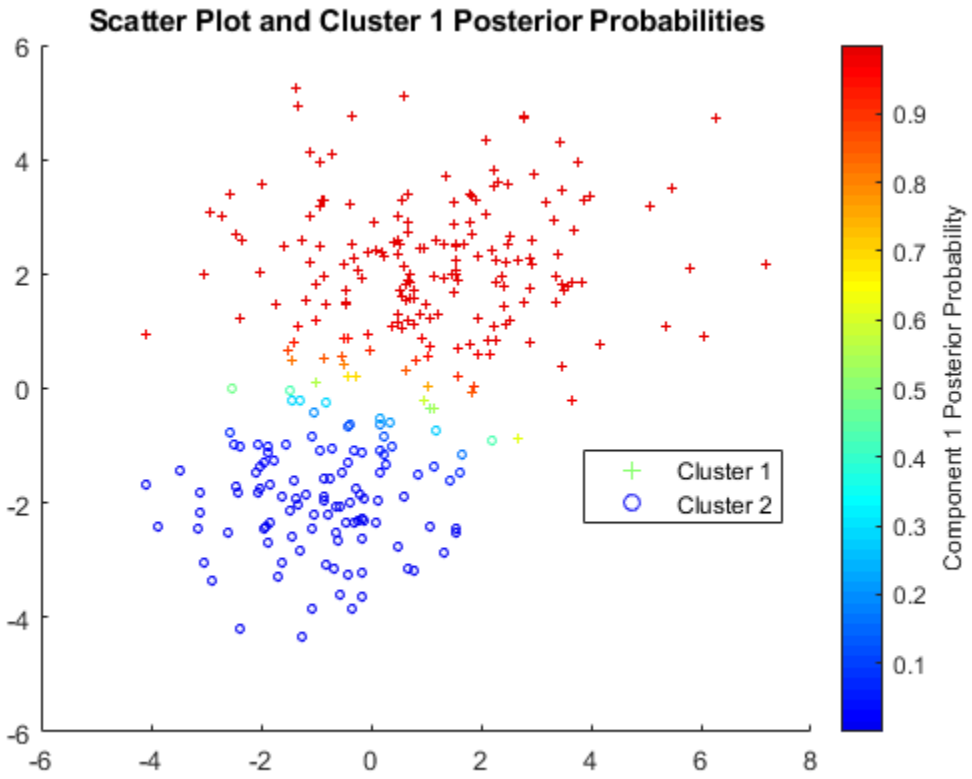
You can estimate cluster membership posterior probabilities by passing the fitted GMM and data to either:

- `posterior`
- `cluster`, and request to return the third output argument

Estimate Cluster Membership Posterior Probabilities

Estimate and plot the posterior probability of the first component for each point.

```
P = posterior(gm,X);

figure
scatter(X(cluster1,1),X(cluster1,2),10,P(cluster1,1),'+')
hold on
scatter(X(cluster2,1),X(cluster2,2),10,P(cluster2,1),'o')
hold off
clrmap = jet(80);
colormap(clrmap(9:72,:))
ylabel(colorbar,'Component 1 Posterior Probability')
legend('Cluster 1','Cluster 2','Location','best')
title('Scatter Plot and Cluster 1 Posterior Probabilities')
```



`P` is an `n`-by-2 matrix of cluster membership posterior probabilities. The first column contains the probabilities for cluster 1 and the second column corresponds to cluster 2.

Assign New Data to Clusters

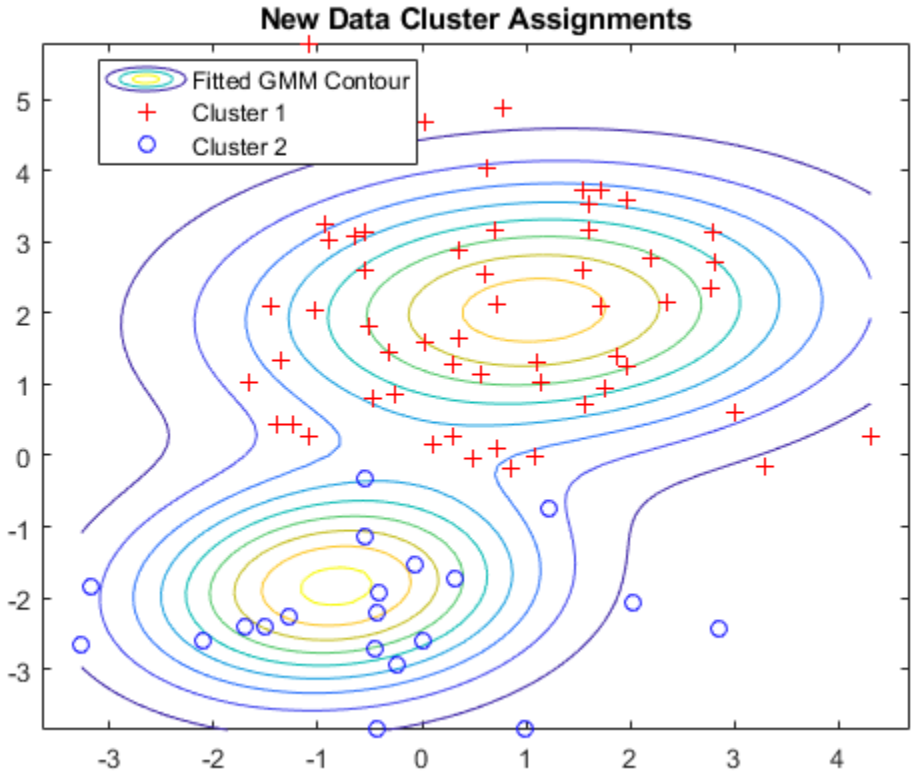
You can also use the `cluster` method to assign new data points to the mixture components found in the original data.

Simulate new data from a mixture of Gaussian distributions. Rather than using `mvnrnd`, you can create a GMM with the true mixture component means and standard deviations using `gmdistribution`, and then pass the GMM to `random` to simulate data.

```
Mu = [mu1; mu2];
Sigma = cat(3,sigma1,sigma2);
p = [0.75 0.25]; % Mixing proportions

gmTrue = gmdistribution(Mu,Sigma,p);
X0 = random(gmTrue,75);
```

Assign clusters to the new data by pass the fitted GMM (`gm`) and the new data to `cluster`. Request cluster membership posterior probabilities.

```
[idx0,~,P0] = cluster(gm,X0);

figure
fcontour(gmPDF,[min(X0(:,1)) max(X0(:,1)) min(X0(:,2)) max(X0(:,2))])
hold on
gscatter(X0(:,1),X0(:,2),idx0,'rb','+o')
legend('Fitted GMM Contour','Cluster 1','Cluster 2','Location','best')
title('New Data Cluster Assignments')
hold off
```



For `cluster` to provide meaningful results when clustering new data, `X0` should come from the same population as `X`, the original data used to create the mixture distribution. In particular, when

computing the posterior probabilities for `X0`, `cluster` and `posterior` use the estimated mixing probabilities.

See Also

`cluster` | `fitgmdist` | `gmdistribution` | `posterior` | `random`

More About

- “Cluster Using Gaussian Mixture Model” on page 16-39
- “Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52
- “Tune Gaussian Mixture Models” on page 16-57
- “Choose Cluster Analysis Method” on page 16-2

Cluster Gaussian Mixture Data Using Soft Clustering

This example shows how to implement soft clustering on simulated data from a mixture of Gaussian distributions.

`cluster` estimates cluster membership posterior probabilities, and then assigns each point to the cluster corresponding to the maximum posterior probability. Soft clustering is an alternative clustering method that allows some data points to belong to multiple clusters. To implement soft clustering:

1. Assign a cluster membership score to each data point that describes how similar each point is to each cluster's archetype. For a mixture of Gaussian distributions, the cluster archetype is corresponding component mean, and the component can be the estimated cluster membership posterior probability.
2. Rank the points by their cluster membership score.
3. Inspect the scores and determine cluster memberships.

For algorithms that use posterior probabilities as scores, a data point is a member of the cluster corresponding to the maximum posterior probability. However, if there are other clusters with corresponding posterior probabilities that are close to the maximum, then the data point can also be a member of those clusters. It is good practice to determine the threshold on scores that yield multiple cluster memberships before clustering.

This example follows from “Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46.

Simulate data from a mixture of two bivariate Gaussian distributions.

```
rng(0,'twister')  % For reproducibility
mu1 = [1 2];
sigma1 = [3 .2; .2 2];
mu2 = [-1 -2];
sigma2 = [2 0; 0 1];
X = [mvnrnd(mu1,sigma1,200); mvnrnd(mu2,sigma2,100)];
```

Fit a two-component Gaussian mixture model (GMM). Because there are two components, suppose that any data point with cluster membership posterior probabilities in the interval [0.4,0.6] can be a member of both clusters.

```
gm = fitgmdist(X,2);
threshold = [0.4 0.6];
```

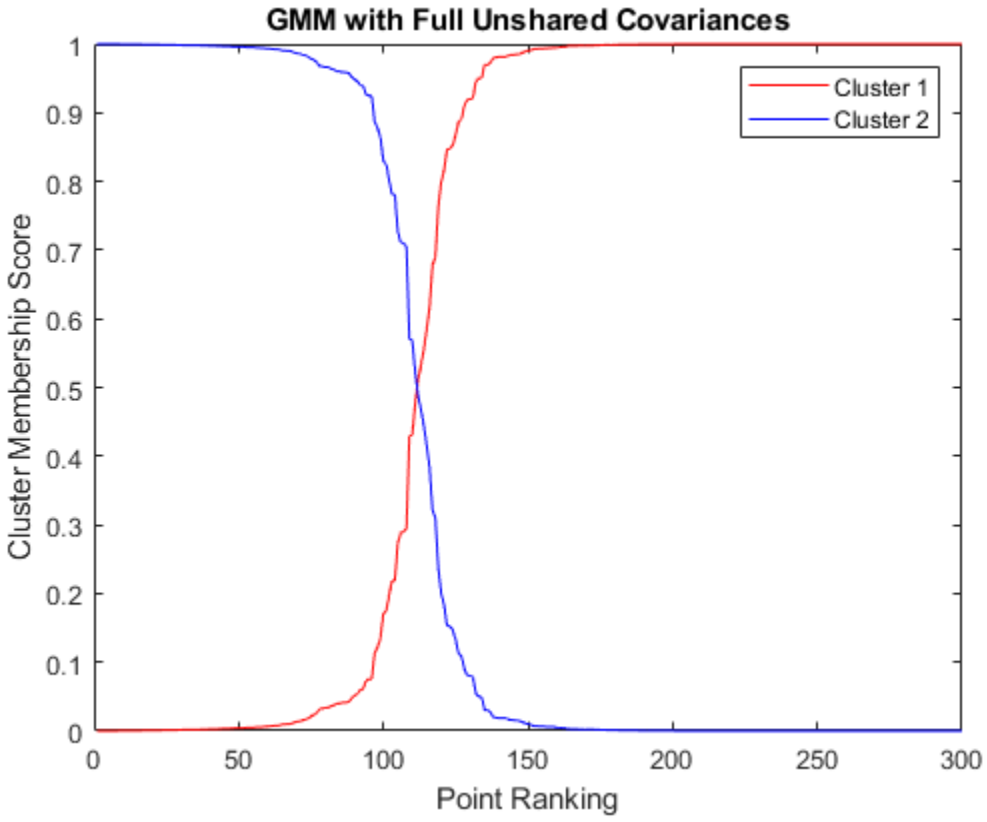
Estimate component-member posterior probabilities for all data points using the fitted GMM `gm`. These represent cluster membership scores.

```
P = posterior(gm,X);
```

For each cluster, rank the membership scores for all data points. For each cluster, plot each data points membership score with respect to its ranking relative to all other data points.

```
n = size(X,1);
[~,order] = sort(P(:,1));

figure
plot(1:n,P(order,1),'r-',1:n,P(order,2),'b-')
legend({'Cluster 1', 'Cluster 2'})
```

```
ylabel('Cluster Membership Score')
xlabel('Point Ranking')
title('GMM with Full Unshared Covariances')
```



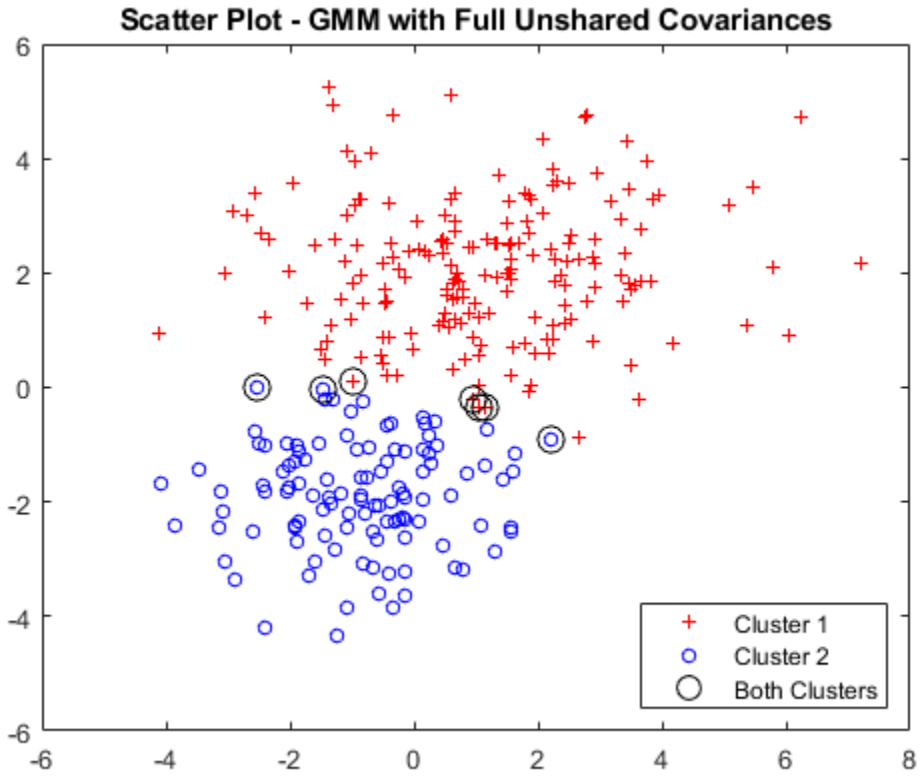
Although a clear separation is hard to see in a scatter plot of the data, plotting the membership scores indicates that the fitted distribution does a good job of separating the data into groups.

Plot the data and assign clusters by maximum posterior probability. Identify points that could be in either cluster.

```
idx = cluster(gm,X);
idxBoth = find(P(:,1)>=threshold(1) & P(:,1)<=threshold(2));
numInBoth = numel(idxBoth)
```

```
numInBoth = 7
```

```
figure
gscatter(X(:,1),X(:,2),idx,'rb','+o',5)
hold on
plot(X(idxBoth,1),X(idxBoth,2),'ko','MarkerSize',10)
legend({'Cluster 1','Cluster 2','Both Clusters'},'Location','SouthEast')
title('Scatter Plot - GMM with Full Unshared Covariances')
hold off
```



Using the score threshold interval, seven data points can be in either cluster.

Soft clustering using a GMM is similar to fuzzy *k*-means clustering, which also assigns each point to each cluster with a membership score. The fuzzy *k*-means algorithm assumes that clusters are roughly spherical in shape, and all of roughly equal size. This is comparable to a Gaussian mixture distribution with a single covariance matrix that is shared across all components, and is a multiple of the identity matrix. In contrast, `gmdistribution` allows you to specify different covariance structures. The default is to estimate a separate, unconstrained covariance matrix for each component. A more restricted option, closer to *k*-means, is to estimate a shared, diagonal covariance matrix.

Fit a GMM to the data, but specify that the components share the same, diagonal covariance matrix. This specification is similar to implementing fuzzy *k*-means clustering, but provides more flexibility by allowing unequal variances for different variables.

```
gmSharedDiag = fitgmdist(X,2,'CovType','Diagonal','SharedCovariance',true');
```

Estimate component-member posterior probabilities for all data points using the fitted GMM `gmSharedDiag`. Estimate soft cluster assignments.

```
[idxSharedDiag,~,PSharedDiag] = cluster(gmSharedDiag,X);
idxBothSharedDiag = find(PSharedDiag(:,1)>=threshold(1) & ...
    PSharedDiag(:,1)<=threshold(2));
numInBoth = numel(idxBothSharedDiag)
```

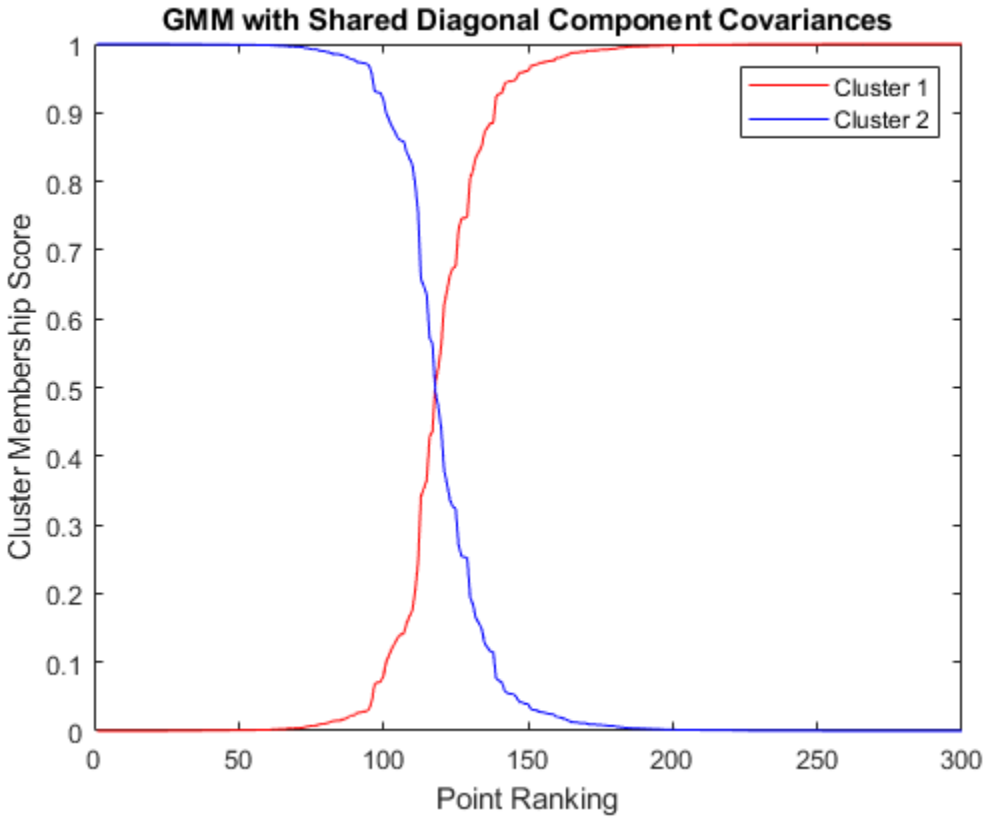
```
numInBoth = 5
```

Assuming shared, diagonal covariances among components, five data points could be in either cluster.

For each cluster:

1. Rank the membership scores for all data points.
2. Plot each data points membership score with respect to its ranking relative to all other data points.

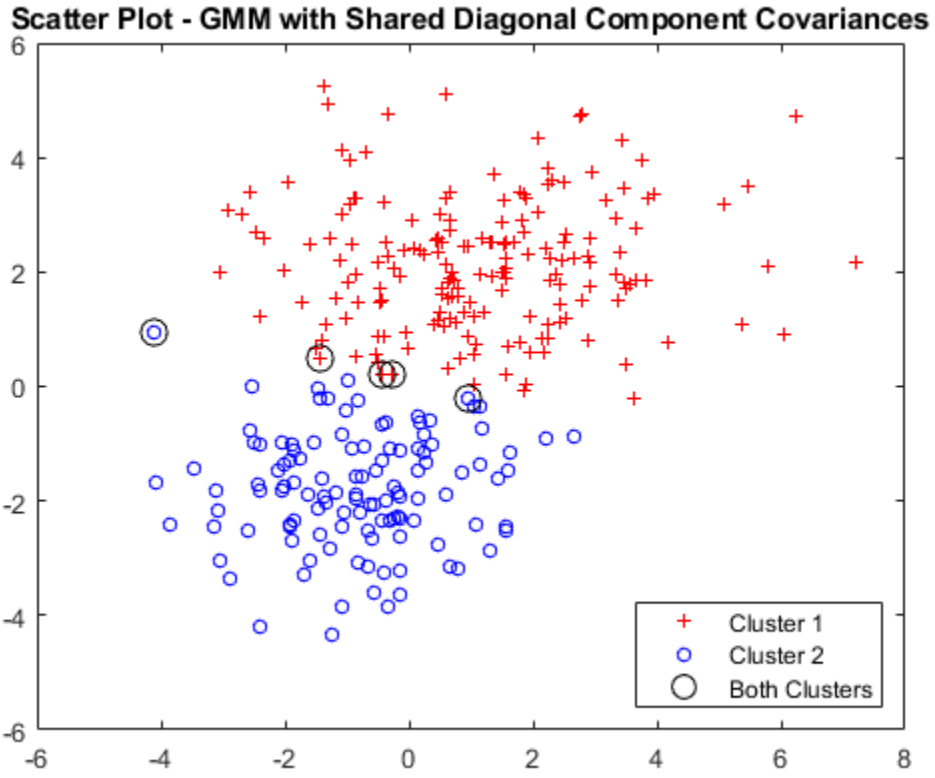
```
[~,orderSharedDiag] = sort(PSharedDiag(:,1));

figure
plot(1:n,PSharedDiag(orderSharedDiag,1),'r-',...
    1:n,PSharedDiag(orderSharedDiag,2),'b-')
legend({'Cluster 1' 'Cluster 2'},'Location','NorthEast')
ylabel('Cluster Membership Score')
xlabel('Point Ranking')
title('GMM with Shared Diagonal Component Covariances')
```



Plot the data and identify the hard, clustering assignments from the GMM analysis assuming the shared, diagonal covariances among components. Also, identify those data points that could be in either cluster.

```
figure
gscatter(X(:,1),X(:,2),idxSharedDiag,'rb','+o',5)
hold on
plot(X(idxBothSharedDiag,1),X(idxBothSharedDiag,2),'ko','MarkerSize',10)
legend({'Cluster 1','Cluster 2','Both Clusters'},'Location','SouthEast')
```

```
title('Scatter Plot - GMM with Shared Diagonal Component Covariances')
hold off
```



See Also

`cluster` | `fitgmdist` | `gmdistribution`

More About

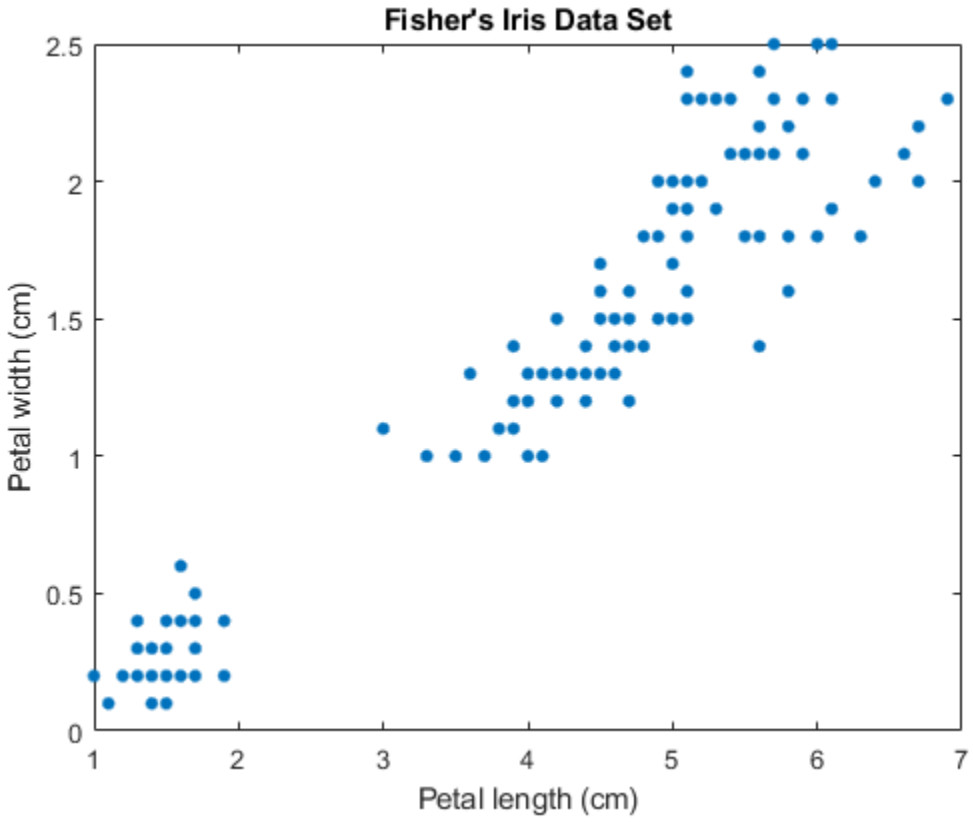
- “Cluster Using Gaussian Mixture Model” on page 16-39
- “Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46
- “Tune Gaussian Mixture Models” on page 16-57
- “Choose Cluster Analysis Method” on page 16-2

Tune Gaussian Mixture Models

This example shows how to determine the best Gaussian mixture model (GMM) fit by adjusting the number of components and the component covariance matrix structure.

Load Fisher's iris data set. Consider the petal measurements as predictors.

```
load fisheriris;
X = meas(:,3:4);
[n,p] = size(X);
rng(1); % For reproducibility

figure;
plot(X(:,1),X(:,2),'.','MarkerSize',15);
title('Fisher''s Iris Data Set');
xlabel('Petal length (cm)');
ylabel('Petal width (cm)');
```



Suppose k is the number of desired components or clusters, and Σ is the covariance structure for all components. Follow these steps to tune a GMM.

1. Choose a (k, Σ) pair, and then fit a GMM using the chosen parameter specification and the entire data set.
2. Estimate the AIC and BIC.
3. Repeat steps 1 and 2 until you exhaust all (k, Σ) pairs of interest.

4 Choose the fitted GMM that balances low AIC with simplicity.

For this example, choose a grid of values for k that include 2 and 3, and some surrounding numbers. Specify all available choices for covariance structure. If k is too high for the data set, then the estimated component covariances can be badly conditioned. Specify to use regularization to avoid badly conditioned covariance matrices. Increase the number of EM algorithm iterations to 10000.

```
k = 1:5;
nK = numel(k);
Sigma = {'diagonal','full'};
nSigma = numel(Sigma);
SharedCovariance = {true,false};
SCtext = {'true','false'};
nSC = numel(SharedCovariance);
RegularizationValue = 0.01;
options = statset('MaxIter',10000);
```

Fit the GMMs using all parameter combination. Compute the AIC and BIC for each fit. Track the terminal convergence status of each fit.

```
% Preallocation
gm = cell(nK,nSigma,nSC);
aic = zeros(nK,nSigma,nSC);
bic = zeros(nK,nSigma,nSC);
converged = false(nK,nSigma,nSC);

% Fit all models
for m = 1:nSC;
    for j = 1:nSigma;
        for i = 1:nK;
            gm{i,j,m} = fitgmdist(X,k(i),...
                'CovarianceType',Sigma{j},...
                'SharedCovariance',SharedCovariance{m},...
                'RegularizationValue',RegularizationValue,...
                'Options',options);
            aic(i,j,m) = gm{i,j,m}.AIC;
            bic(i,j,m) = gm{i,j,m}.BIC;
            converged(i,j,m) = gm{i,j,m}.Converged;
        end
    end
end

allConverge = (sum(converged(:)) == nK*nSigma*nSC)
```

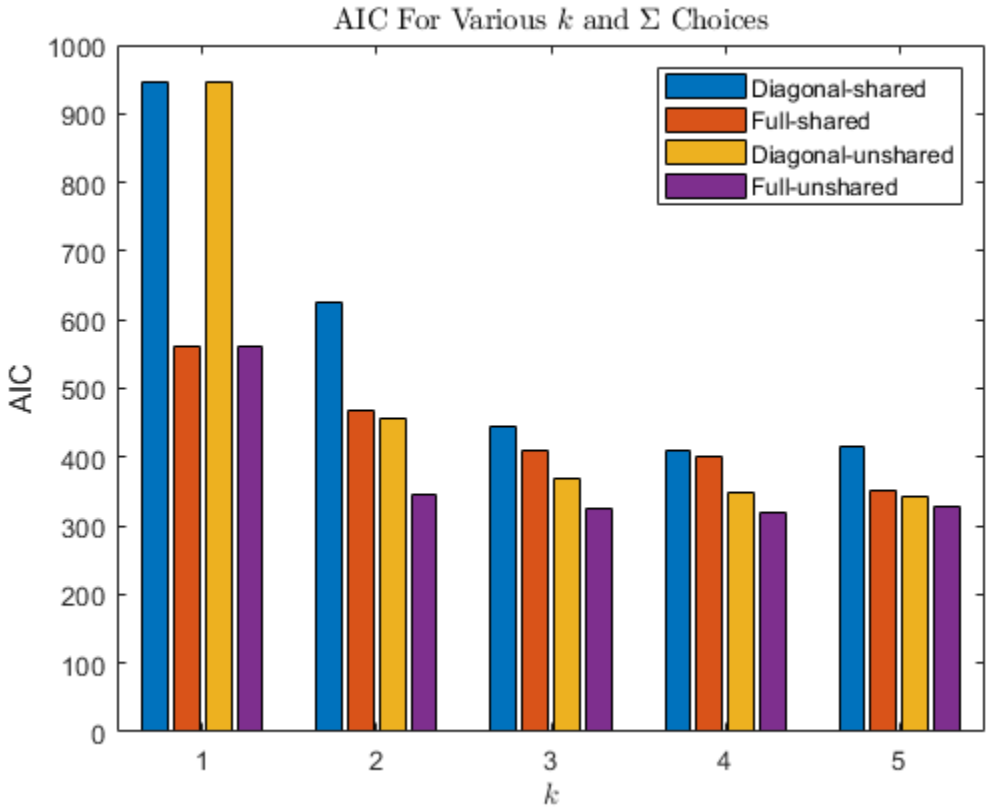
```
allConverge =

  logical

   1
```

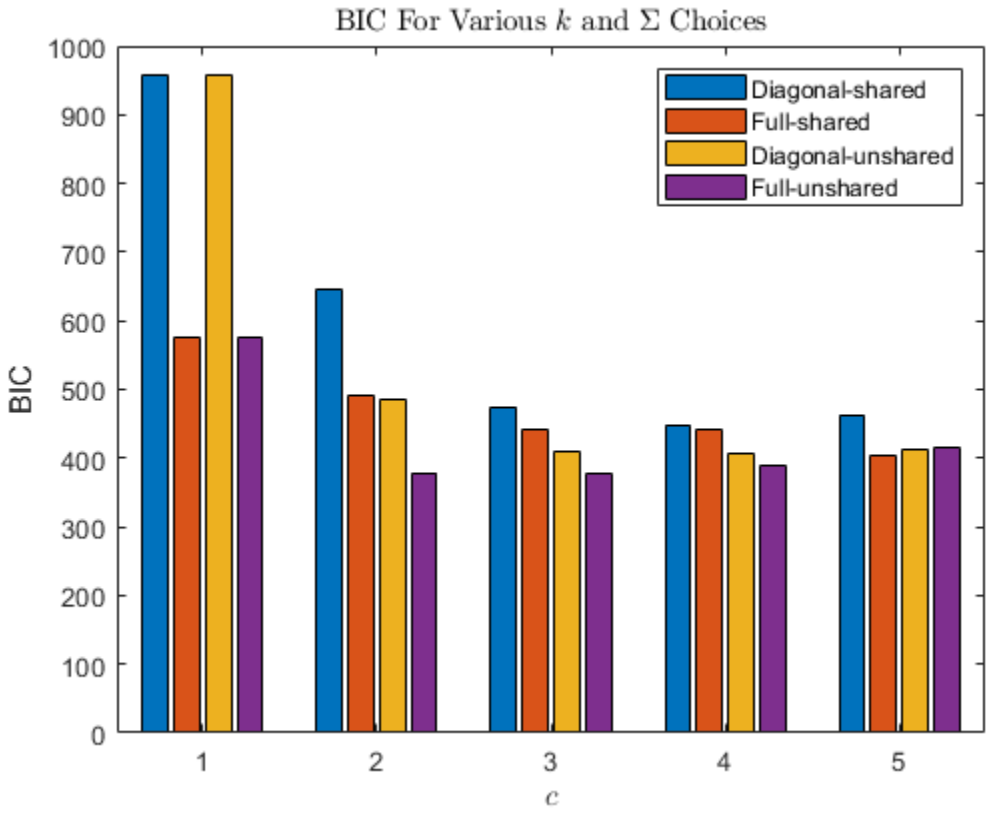
`gm` is a cell array containing all of the fitted `gmdistribution` model objects. All of the fitting instances converged.

Plot separate bar charts to compare the AIC and BIC among all fits. Group the bars by k.

```
figure;
bar(reshape(aic,nK,nSigma*nSC));
```

```
title('AIC For Various $k$ and $\Sigma$ Choices','Interpreter','latex');
xlabel('$k$','Interpreter','Latex');
ylabel('AIC');
legend({'Diagonal-shared','Full-shared','Diagonal-unshared',...
    'Full-unshared'});

figure;
bar(reshape(bic,nK,nSigma*nSC));
title('BIC For Various $k$ and $\Sigma$ Choices','Interpreter','latex');
xlabel('$c$','Interpreter','Latex');
ylabel('BIC');
legend({'Diagonal-shared','Full-shared','Diagonal-unshared',...
    'Full-unshared'});
```



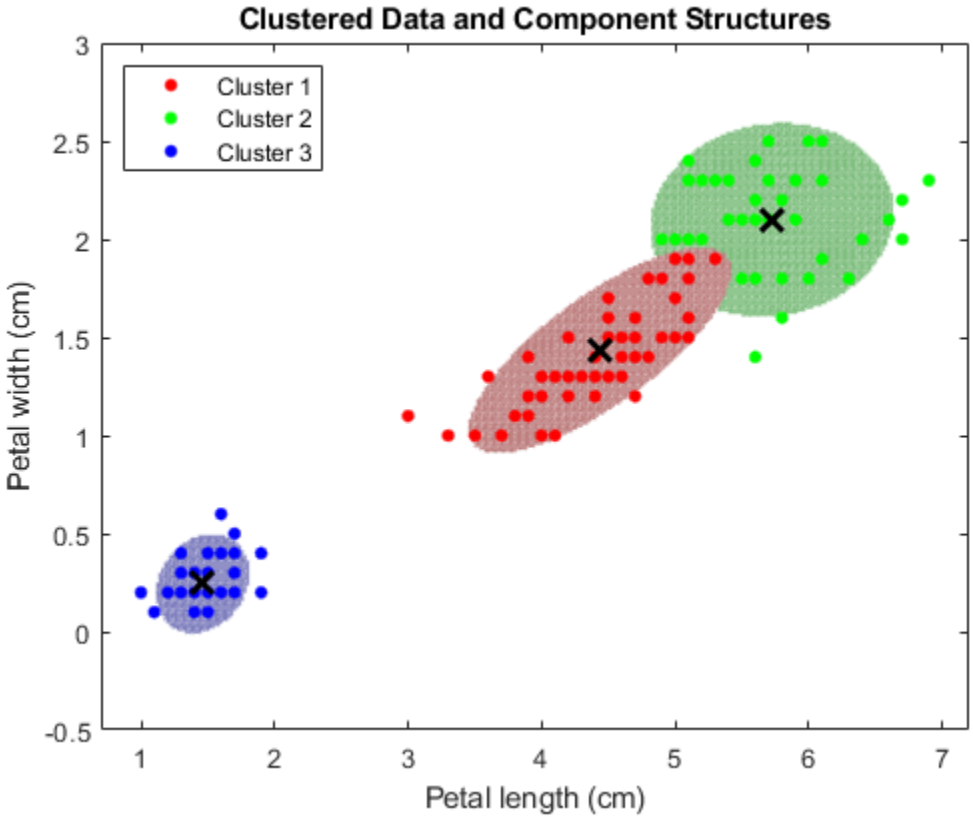


According to the AIC and BIC values, the best model has 3 components and a full, unshared covariance matrix structure.

Cluster the training data using the best fitting model. Plot the clustered data and the component ellipses.

```
gmBest = gm{3,2,2};
clusterX = cluster(gmBest,X);
kGMM = gmBest.NumComponents;
d = 500;
x1 = linspace(min(X(:,1)) - 2,max(X(:,1)) + 2,d);
x2 = linspace(min(X(:,2)) - 2,max(X(:,2)) + 2,d);
[x1grid,x2grid] = meshgrid(x1,x2);
X0 = [x1grid(:) x2grid(:)];
mahalDist = mahal(gmBest,X0);
threshold = sqrt(chi2inv(0.99,2));

figure;
h1 = gscatter(X(:,1),X(:,2),clusterX);
hold on;
for j = 1:kGMM;
    idx = mahalDist(:,j)<=threshold;
    Color = h1(j).Color*0.75 + -0.5*(h1(j).Color - 1);
    h2 = plot(X0(idx,1),X0(idx,2),'.','Color',Color,'MarkerSize',1);
    uistack(h2,'bottom');
end
h3 = plot(gmBest.mu(:,1),gmBest.mu(:,2),'kx','LineWidth',2,'MarkerSize',10);
```

```
title('Clustered Data and Component Structures');
xlabel('Petal length (cm)');
ylabel('Petal width (cm)');
legend(h1,'Cluster 1','Cluster 2','Cluster 3','Location','NorthWest');
hold off
```



This data set includes labels. Determine how well `gmBest` clusters the data by comparing each prediction to the true labels.

```
species = categorical(species);
Y = zeros(n,1);
Y(species == 'versicolor') = 1;
Y(species == 'virginica') = 2;
Y(species == 'setosa') = 3;

miscluster = Y ~= clusterX;
clusterError = sum(miscluster)/n
```

```
clusterError =

    0.0800
```

The best fitting GMM groups 8% of the observations into the wrong cluster.

`cluster` does not always preserve cluster order. That is, if you cluster several fitted `gmdistribution` models, `cluster` might assign different cluster labels for similar components.

See Also

`cluster` | `fitgmdist` | `gmdistribution`

More About

- “Cluster Using Gaussian Mixture Model” on page 16-39
- “Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46
- “Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52
- “Choose Cluster Analysis Method” on page 16-2

Cluster Evaluation

This example shows how to identify clusters in Fisher's iris data.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
y = categorical(species);
```

`X` is a numeric matrix that contains two petal measurements for 150 irises. `Y` is a cell array of character vectors that contains the corresponding iris species.

Evaluate multiple clusters from 1 to 10.

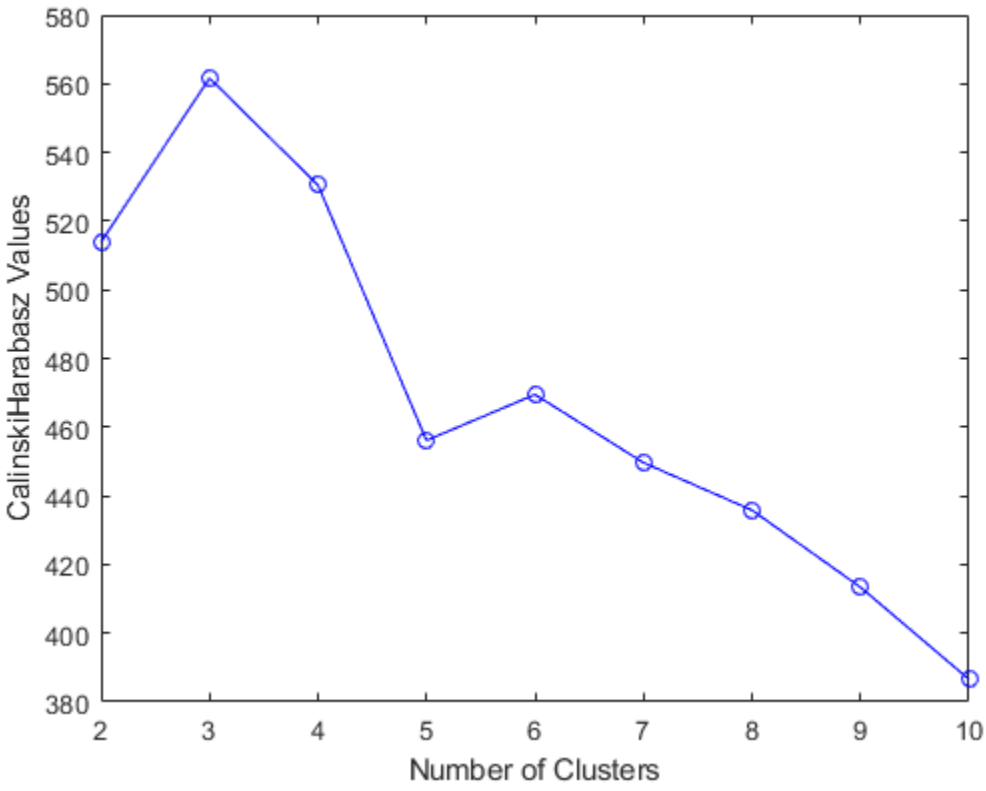
```
eva = evalclusters(X,'kmeans','CalinskiHarabasz','KList',1:10)
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6 7 8 9 10]
    CriterionValues: [1x10 double]
           OptimalK: 1
```

The `OptimalK` value indicates that, based on the Calinski-Harabasz criterion, the optimal number of clusters is three.

Visualize `eva` to see the results for each number of clusters.

```
plot(eva)
```



Most clustering algorithms need prior knowledge of the number of clusters. When this information is not available, use cluster evaluation techniques to determine the number of clusters present in the data based on a specified metric.

Three clusters is consistent with the three species in the data.

```
categories(y)
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

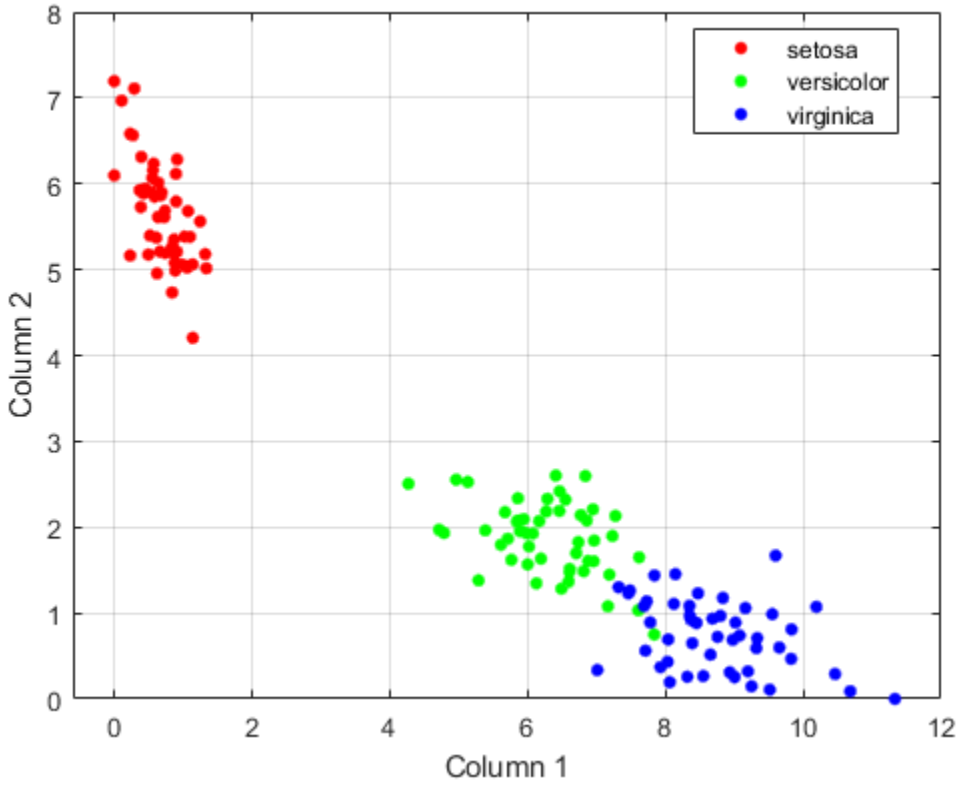
Compute a nonnegative rank-two approximation of the data for visualization purposes.

```
Xred = nnmf(X,2);
```

The original features are reduced to two features. Since none of the features are negative, `nnmf` also guarantees that the features are nonnegative.

Confirm the three clusters visually using a scatter plot.

```
gscatter(Xred(:,1),Xred(:,2),y)
xlabel('Column 1')
ylabel('Column 2')
grid on
```



See Also

`evalclusters` | `nnmf`

Cluster Analysis

This example shows how to examine similarities and dissimilarities of observations or objects using cluster analysis in Statistics and Machine Learning Toolbox™. Data often fall naturally into groups (or clusters) of observations, where the characteristics of objects in the same cluster are similar and the characteristics of objects in different clusters are dissimilar.

K-Means and Hierarchical Clustering

The Statistics and Machine Learning Toolbox includes functions to perform K-means clustering and hierarchical clustering.

K-means clustering is a partitioning method that treats observations in your data as objects having locations and distances from each other. It partitions the objects into K mutually exclusive clusters, such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster is characterized by its centroid, or center point. Of course, the distances used in clustering often do not represent spatial distances.

Hierarchical clustering is a way to investigate grouping in your data, simultaneously over a variety of scales of distance, by creating a cluster tree. The tree is not a single set of clusters, as in K-Means, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows you to decide what scale or level of clustering is most appropriate in your application.

Some of the functions used in this example call MATLAB® built-in random number generation functions. To duplicate the exact results shown in this example, you should execute the command below, to set the random number generator to a known state. If you do not set the state, your results may differ in trivial ways, for example, you may see clusters numbered in a different order. There is also a chance that a suboptimal cluster solution may result (the example includes a discussion of suboptimal solutions, including ways to avoid them).

```
rng(6,'twister')
```

Fisher's Iris Data

In the 1920's, botanists collected measurements on the sepal length, sepal width, petal length, and petal width of 150 iris specimens, 50 from each of three species. The measurements became known as Fisher's iris data set.

Each observation in this data set comes from a known species, and so there is already an obvious way to group the data. For the moment, we will ignore the species information and cluster the data using only the raw measurements. When we are done, we can compare the resulting clusters to the actual species, to see if the three types of iris possess distinct characteristics.

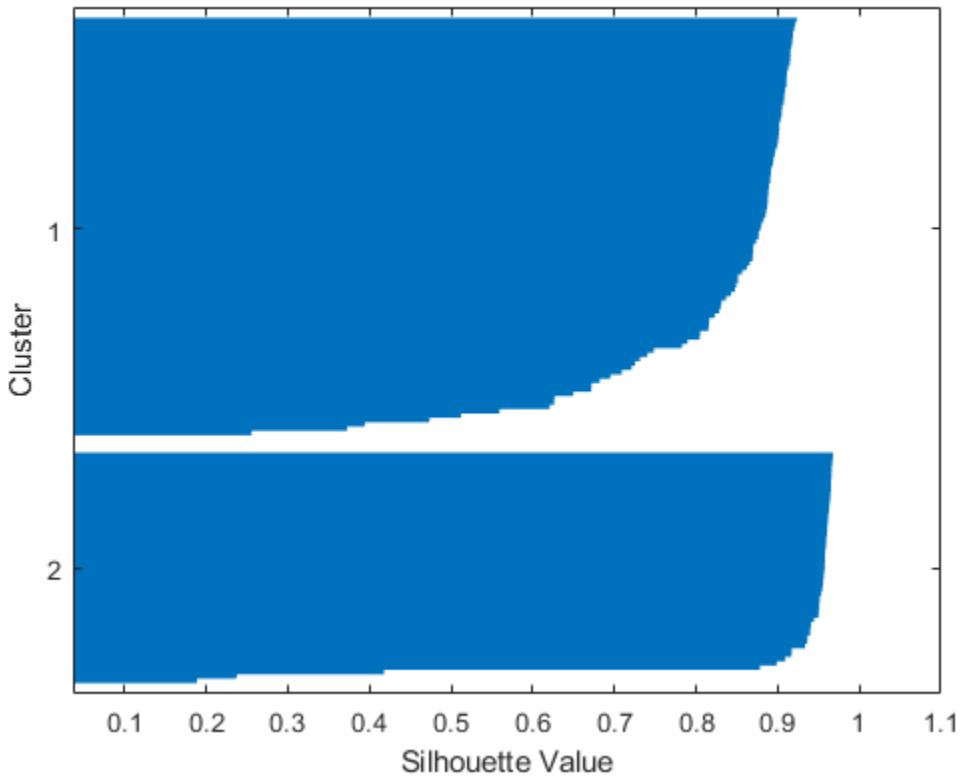
Clustering Fisher's Iris Data Using K-Means Clustering

The function `kmeans` performs K-Means clustering, using an iterative algorithm that assigns objects to clusters so that the sum of distances from each object to its cluster centroid, over all clusters, is a minimum. Used on Fisher's iris data, it will find the natural groupings among iris specimens, based on their sepal and petal measurements. With K-means clustering, you must specify the number of clusters that you want to create.

First, load the data and call `kmeans` with the desired number of clusters set to 2, and using squared Euclidean distance. To get an idea of how well-separated the resulting clusters are, you can make a

silhouette plot. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

```
load fisheriris
[cidx2,cmeans2] = kmeans(meas,2,'dist','sqeuclidean');
[silh2,h] = silhouette(meas,cidx2,'sqeuclidean');
```

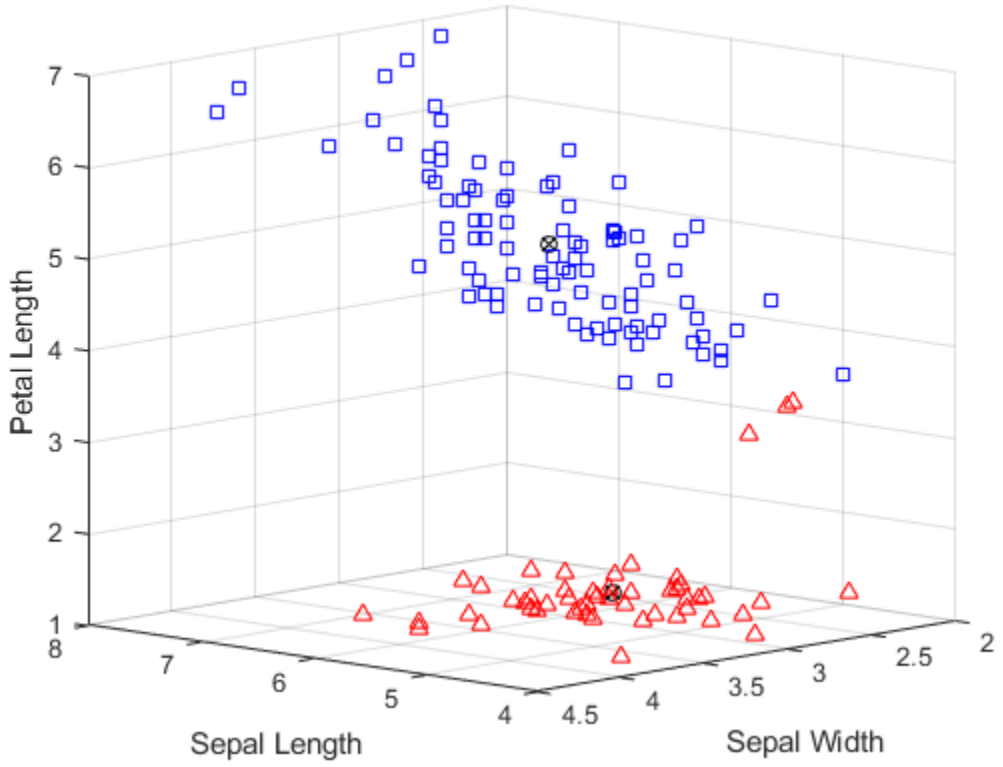


From the silhouette plot, you can see that most points in both clusters have a large silhouette value, greater than 0.8, indicating that those points are well-separated from neighboring clusters. However, each cluster also contains a few points with low silhouette values, indicating that they are nearby to points from other clusters.

It turns out that the fourth measurement in these data, the petal width, is highly correlated with the third measurement, the petal length, and so a 3-D plot of the first three measurements gives a good representation of the data, without resorting to four dimensions. If you plot the data, using different symbols for each cluster created by `kmeans`, you can identify the points with small silhouette values as those points that are close to points from other clusters.

```
ptsymb = {'bs','r^','md','go','c+'};
for i = 1:2
    clust = find(cidx2==i);
    plot3(meas(clust,1),meas(clust,2),meas(clust,3),ptsymb{i});
    hold on
end
plot3(cmeans2(:,1),cmeans2(:,2),cmeans2(:,3),'ko');
plot3(cmeans2(:,1),cmeans2(:,2),cmeans2(:,3),'kx');
hold off
```

```
xlabel('Sepal Length');
ylabel('Sepal Width');
zlabel('Petal Length');
view(-137,10);
grid on
```



The centroids of each cluster are plotted using circled X's. Three of the points from the lower cluster (plotted with triangles) are very close to points from the upper cluster (plotted with squares). Because the upper cluster is so spread out, those three points are closer to the centroid of the lower cluster than to that of the upper cluster, even though the points are separated from the bulk of the points in their own cluster by a gap. Because K-means clustering only considers distances, and not densities, this kind of result can occur.

You can increase the number of clusters to see if `kmeans` can find further grouping structure in the data. This time, use the optional `'Display'` name-value pair argument to print out information about each iteration in the clustering algorithm.

```
[cidx3,cmeans3] = kmeans(meas,3,'Display','iter');
```

```
  iter     phase         num            sum
     1         1         150        146.424
     2         1           5        144.333
     3         1           4        143.924
     4         1           3         143.61
     5         1           1        143.542
     6         1           2        143.414
     7         1           2        143.023
```

```
     8         1           2         142.823
     9         1           1         142.786
    10         1           1         142.754
Best total sum of distances = 142.754
```

At each iteration, the `kmeans` algorithm (see “Algorithms” on page 33-3330) reassigns points among clusters to decrease the sum of point-to-centroid distances, and then recomputes cluster centroids for the new cluster assignments. Notice that the total sum of distances and the number of reassignments decrease at each iteration until the algorithm reaches a minimum. The algorithm used in `kmeans` consists of two phases. In the example here, the second phase of the algorithm did not make any reassignments, indicating that the first phase reached a minimum after only a few iterations.

By default, `kmeans` begins the clustering process using a randomly selected set of initial centroid locations. The `kmeans` algorithm can converge to a solution that is a local minimum; that is, `kmeans` can partition the data such that moving any single point to a different cluster increases the total sum of distances. However, as with many other types of numerical minimizations, the solution that `kmeans` reaches sometimes depends on the starting points. Therefore, other solutions (local minima) that have a lower total sum of distances can exist for the data. You can use the optional `'Replicates'` name-value pair argument to test different solutions. When you specify more than one replicate, `kmeans` repeats the clustering process starting from different randomly selected centroids for each replicate. `kmeans` then returns the solution with the lowest total sum of distances among all the replicates.

```
[cidx3,cmeans3,sumd3] = kmeans(meas,3,'replicates',5,'display','final');
```

```
Replicate 1, 9 iterations, total sum of distances = 78.8557.
Replicate 2, 10 iterations, total sum of distances = 78.8557.
Replicate 3, 8 iterations, total sum of distances = 78.8557.
Replicate 4, 8 iterations, total sum of distances = 78.8557.
Replicate 5, 1 iterations, total sum of distances = 78.8514.
Best total sum of distances = 78.8514
```

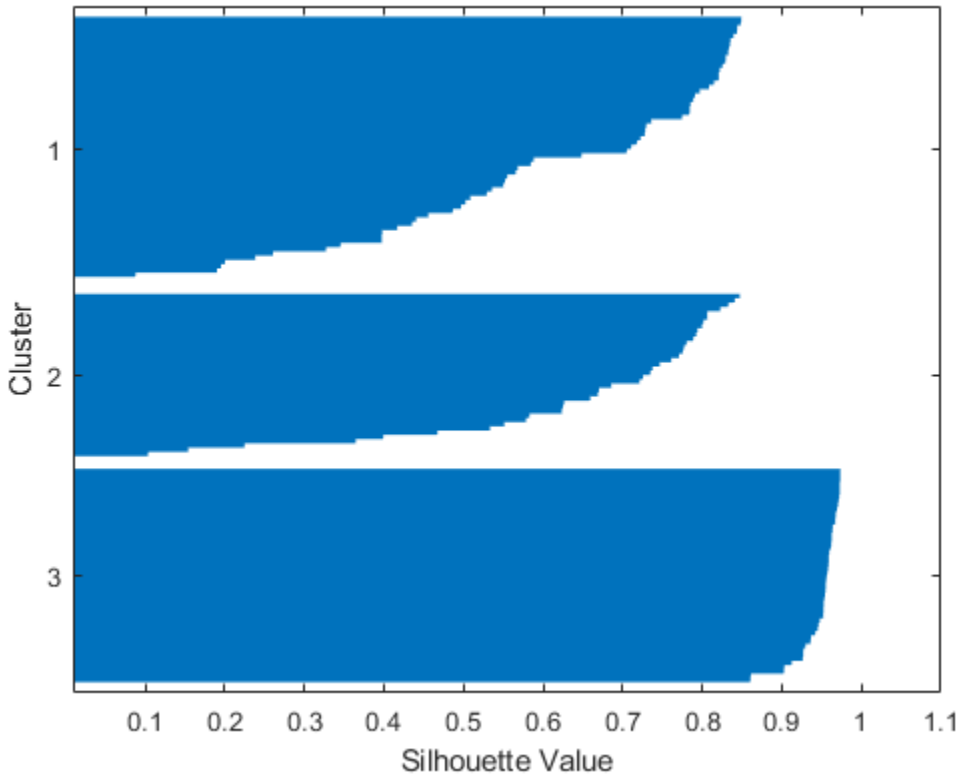
The output shows that, even for this relatively simple problem, non-global minima do exist. Each of these five replicates began from a different set of initial centroids. Depending on where it started from, `kmeans` reached one of two different solutions. However, the final solution that `kmeans` returns is the one with the lowest total sum of distances, over all replicates. The third output argument contains the sum of distances within each cluster for that best solution.

```
sum(sumd3)
```

```
ans =

   78.8514
```

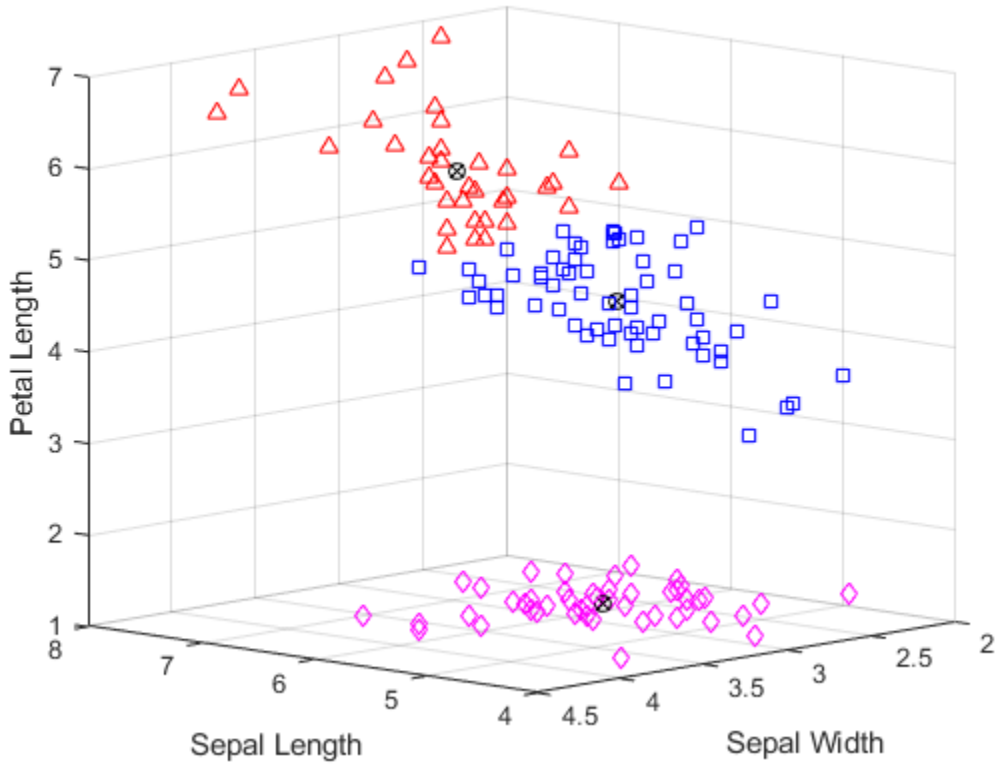
A silhouette plot for this three-cluster solution indicates that there is one cluster that is well-separated, but that the other two clusters are not very distinct.

```
[silh3,h] = silhouette(meas,cidx3,'sqeuclidean');
```



Again, you can plot the raw data to see how `kmeans` has assigned the points to clusters.

```
for i = 1:3
    clust = find(cidx3==i);
    plot3(meas(clust,1),meas(clust,2),meas(clust,3),ptsymb{i});
    hold on
end
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'ko');
plot3(cmeans3(:,1),cmeans3(:,2),cmeans3(:,3),'kx');
hold off
xlabel('Sepal Length');
ylabel('Sepal Width');
zlabel('Petal Length');
view(-137,10);
grid on
```



You can see that `kmeans` has split the upper cluster from the two-cluster solution, and that those two clusters are very close to each other. Depending on what you intend to do with these data after clustering them, this three-cluster solution may be more or less useful than the previous, two-cluster, solution. The first output argument from `silhouette` contains the silhouette values for each point, which you can use to compare the two solutions quantitatively. The average silhouette value was larger for the two-cluster solution, indicating that it is a better answer purely from the point of view of creating distinct clusters.

```
[mean(silh2) mean(silh3)]
```

```
ans =

    0.8504    0.7357
```

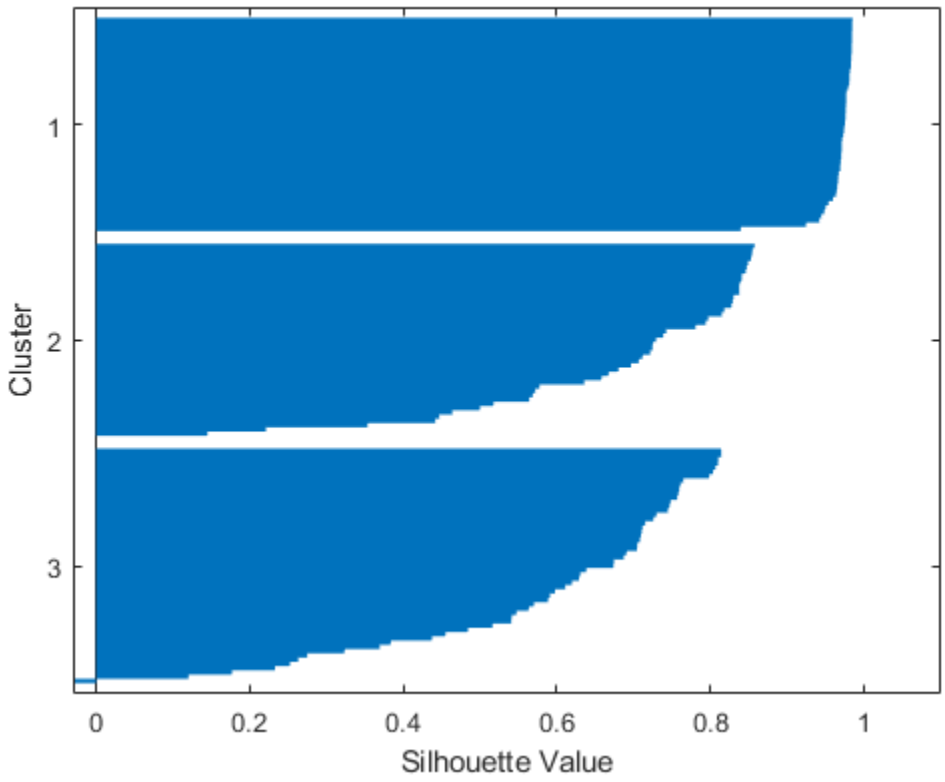
You can also cluster these data using a different distance. The cosine distance might make sense for these data because it would ignore absolute sizes of the measurements, and only consider their relative sizes. Thus, two flowers that were different sizes, but which had similarly shaped petals and sepals, might not be close with respect to squared Euclidean distance, but would be close with respect to cosine distance.

```
[cidxCos,cmeansCos] = kmeans(meas,3,'dist','cos');
```

From the silhouette plot, these clusters appear to be only slightly better separated than those found using squared Euclidean distance.

```
[silhCos,h] = silhouette(meas,cidxCos,'cos');
[mean(silh2) mean(silh3) mean(silhCos)]
```

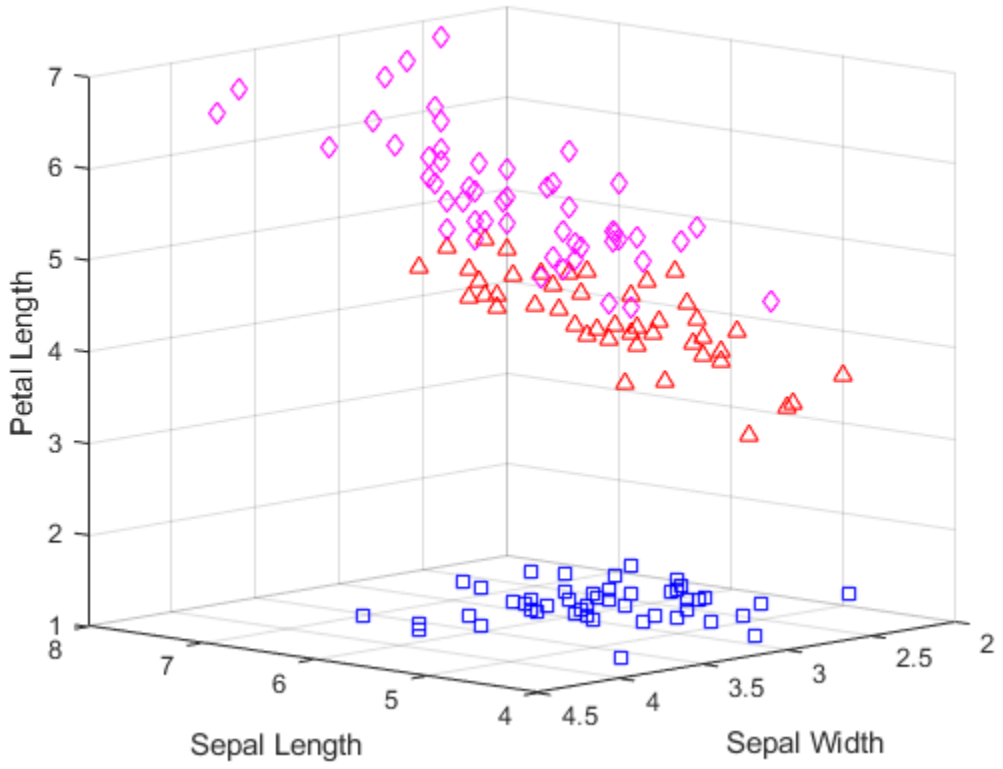
```
ans =

    0.8504    0.7357    0.7491
```



Notice that the order of the clusters is different than in the previous silhouette plot. This is because `kmeans` chooses initial cluster assignments at random.

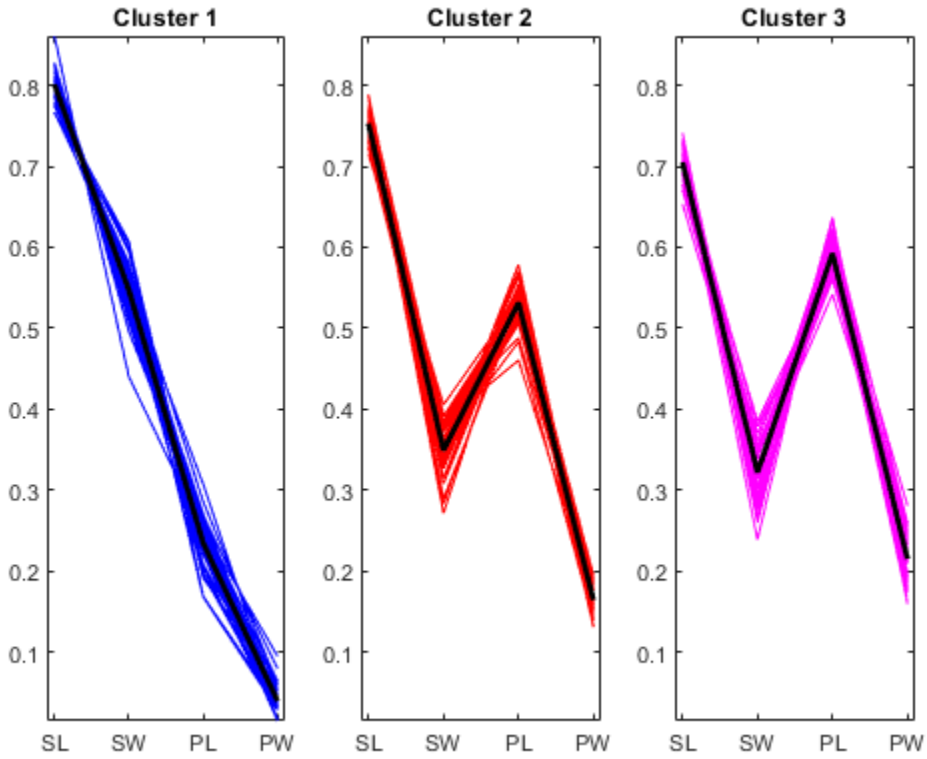
By plotting the raw data, you can see the differences in the cluster shapes created using the two different distances. The two solutions are similar, but the two upper clusters are elongated in the direction of the origin when using cosine distance.

```
for i = 1:3
    clust = find(cidxCos==i);
    plot3(meas(clust,1),meas(clust,2),meas(clust,3),ptsymb{i});
    hold on
end
hold off
xlabel('Sepal Length');
ylabel('Sepal Width');
zlabel('Petal Length');
view(-137,10);
grid on
```



This plot does not include the cluster centroids, because a centroid with respect to the cosine distance corresponds to a half-line from the origin in the space of the raw data. However, you can make a parallel coordinate plot of the normalized data points to visualize the differences between cluster centroids.

```
lnsymb = {'b-','r-','m-'};
names = {'SL','SW','PL','PW'};
meas0 = meas ./ repmat(sqrt(sum(meas.^2,2)),1,4);
ymin = min(min(meas0));
ymax = max(max(meas0));
for i = 1:3
    subplot(1,3,i);
    plot(meas0(cidxCos==i,:)',lnsymb{i});
    hold on;
    plot(cmeansCos(i,:)','k-','LineWidth',2);
    hold off;
    title(sprintf('Cluster %d',i));
    xlim([.9, 4.1]);
    ylim([ymin, ymax]);
    h_gca = gca;
    h_gca.XTick = 1:4;
    h_gca.XTickLabel = names;
end
```

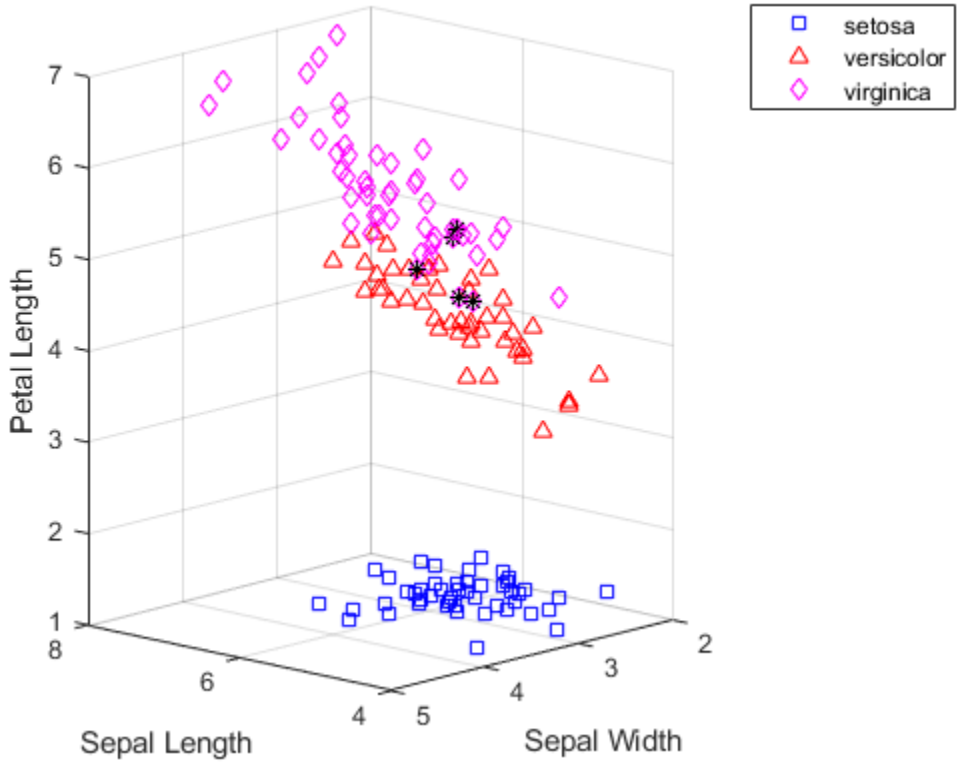


It's clear from this plot that specimens from each of the three clusters have distinctly different relative sizes of petals and sepals on average. The first cluster has petals that are strictly smaller than their sepals. The second two clusters' petals and sepals overlap in size, however, those from the third cluster overlap more than the second. You can also see that the second and third clusters include some specimens which are very similar to each other.

Because we know the species of each observation in the data, you can compare the clusters discovered by `kmeans` to the actual species, to see if the three species have discernibly different physical characteristics. In fact, as the following plot shows, the clusters created using cosine distance differ from the species groups for only five of the flowers. Those five points, plotted with stars, are all near the boundary of the upper two clusters.

```
subplot(1,1,1);
for i = 1:3
    clust = find(cidxCos==i);
    plot3(meas(clust,1),meas(clust,2),meas(clust,3),ptsymb{i});
    hold on
end
xlabel('Sepal Length');
ylabel('Sepal Width');
zlabel('Petal Length');
view(-137,10);
grid on
sidx = grp2idx(species);
miss = find(cidxCos ~= sidx);
plot3(meas(miss,1),meas(miss,2),meas(miss,3),'k*');
```

```
legend({'setosa','versicolor','virginica'});
hold off
```



Clustering Fisher's Iris Data Using Hierarchical Clustering

K-Means clustering produced a single partition of the iris data, but you might also want to investigate different scales of grouping in your data. Hierarchical clustering lets you do just that, by creating a hierarchical tree of clusters.

First, create a cluster tree using distances between observations in the iris data. Begin by using Euclidean distance.

```
eucD = pdist(meas,'euclidean');
clustTreeEuc = linkage(eucD,'average');
```

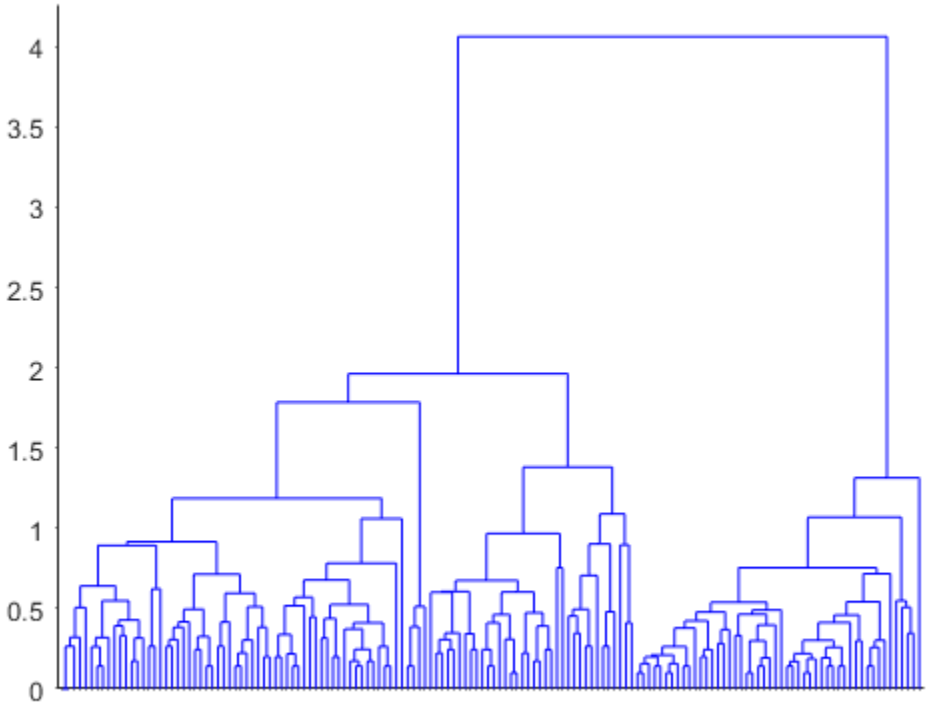
The cophenetic correlation is one way to verify that the cluster tree is consistent with the original distances. Large values indicate that the tree fits the distances well, in the sense that pairwise linkages between observations correlate with their actual pairwise distances. This tree seems to be a fairly good fit to the distances.

```
cophenet(clustTreeEuc,eucD)
```

```
ans =

    0.8770
```

To visualize the hierarchy of clusters, you can plot a dendrogram.

```
[h,nodes] = dendrogram(clustTreeEuc,0);
h_gca = gca;
h_gca.TickDir = 'out';
h_gca.TickLength = [.002 0];
h_gca.XTickLabel = [];
```



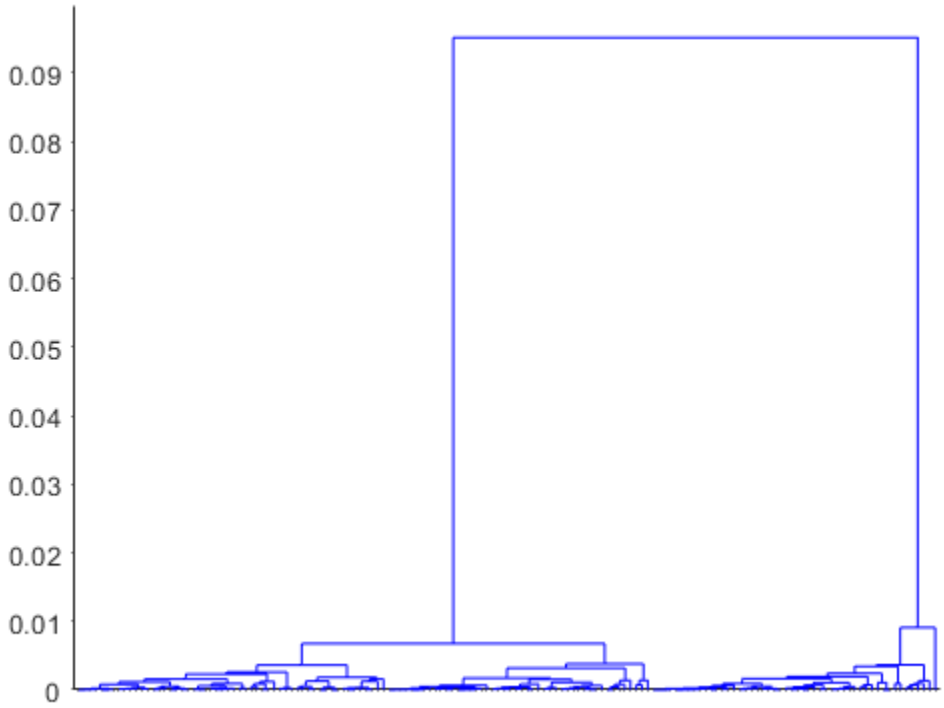
The root node in this tree is much higher than the remaining nodes, confirming what you saw from K-Means clustering: there are two large, distinct groups of observations. Within each of those two groups, you can see that lower levels of groups emerge as you consider smaller and smaller scales in distance. There are many different levels of groups, of different sizes, and at different degrees of distinctness.

Based on the results from K-Means clustering, cosine might also be a good choice of distance measure. The resulting hierarchical tree is quite different, suggesting a very different way to look at group structure in the iris data.

```
cosD = pdist(meas,'cosine');
clustTreeCos = linkage(cosD,'average');
cophenet(clustTreeCos,cosD)
```

```
ans =

    0.9360
```

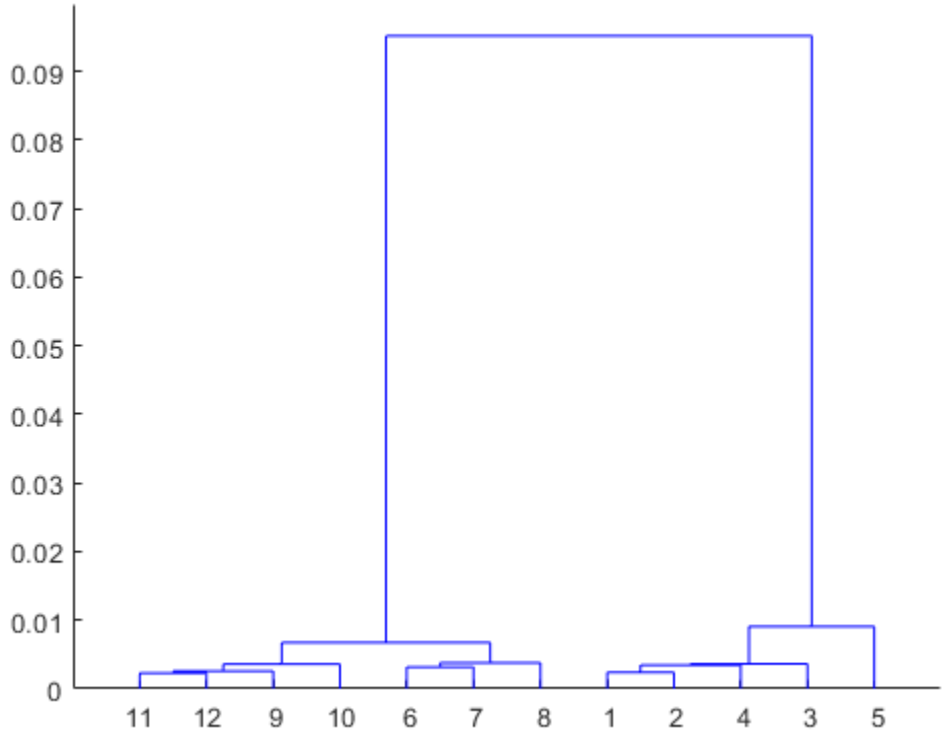
```
[h,nodes] = dendrogram(clustTreeCos,0);
h_gca = gca;
h_gca.TickDir = 'out';
h_gca.TickLength = [.002 0];
h_gca.XTickLabel = [];
```



The highest level of this tree separates iris specimens into two very distinct groups. The dendrogram shows that, with respect to cosine distance, the within-group differences are much smaller relative to the between-group differences than was the case for Euclidean distance. This is exactly what you would expect for these data, since the cosine distance computes a zero pairwise distance for objects that are in the same "direction" from the origin.

With 150 observations, the plot is cluttered, but you can make a simplified dendrogram that does not display the very lowest levels of the tree.

```
[h,nodes] = dendrogram(clustTreeCos,12);
```



The three highest nodes in this tree separate out three equally-sized groups, plus a single specimen (labeled as leaf node 5) that is not near any others.

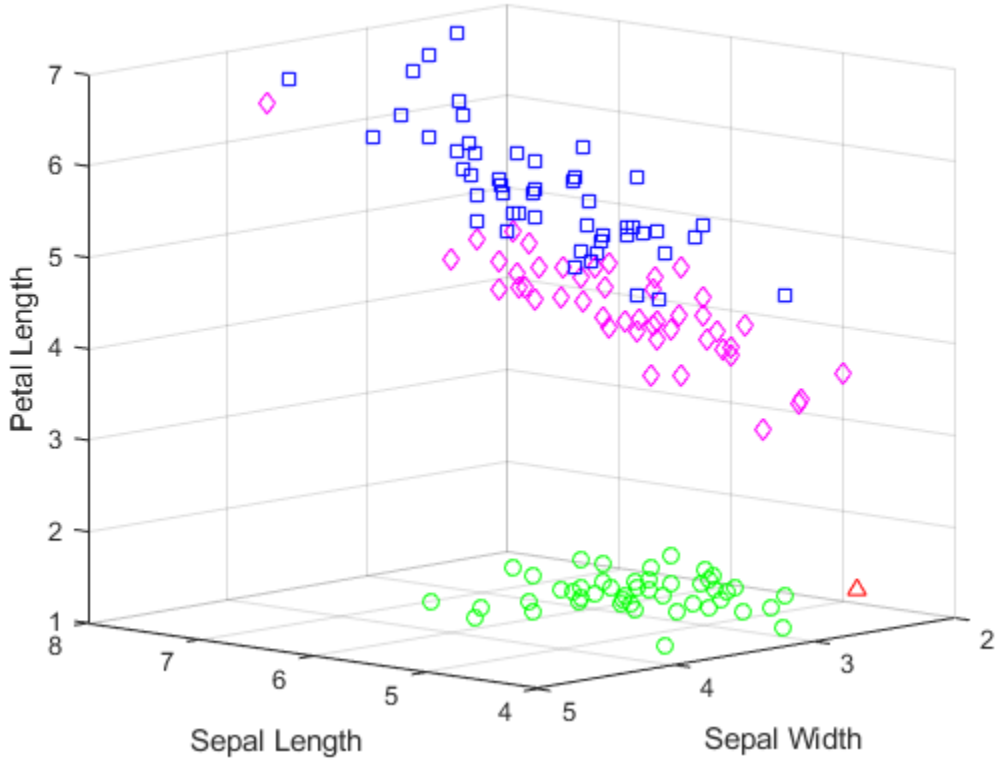
```
[sum(ismember(nodes,[11 12 9 10])) sum(ismember(nodes,[6 7 8])) ...
                  sum(ismember(nodes,[1 2 4 3])) sum(nodes==5)]
```

```
ans =

    54    46    49     1
```

For many purposes, the dendrogram might be a sufficient result. However, you can go one step further, and use the `cluster` function to cut the tree and explicitly partition observations into specific clusters, as with K-Means. Using the hierarchy from the cosine distance to create clusters, specify a linkage height that will cut the tree below the three highest nodes, and create four clusters, then plot the clustered raw data.

```
hidx = cluster(clustTreeCos,'criterion','distance','cutoff',.006);
for i = 1:5
    clust = find(hidx==i);
    plot3(meas(clust,1),meas(clust,2),meas(clust,3),ptsymb{i});
    hold on
end
hold off
xlabel('Sepal Length');
ylabel('Sepal Width');
```

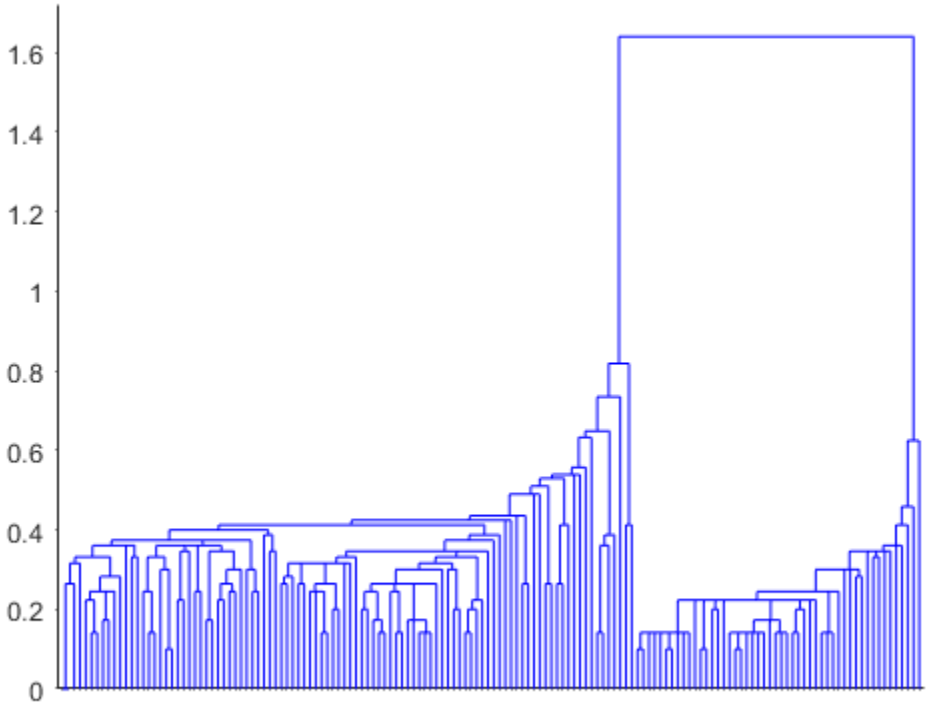
```
zlabel('Petal Length');
view(-137,10);
grid on
```



This plot shows that the results from hierarchical clustering with cosine distance are qualitatively similar to results from K-Means, using three clusters. However, creating a hierarchical cluster tree allows you to visualize, all at once, what would require considerable experimentation with different values for K in K-Means clustering.

Hierarchical clustering also allows you to experiment with different linkages. For example, clustering the iris data with single linkage, which tends to link together objects over larger distances than average distance does, gives a very different interpretation of the structure in the data.

```
clustTreeSng = linkage(eucD,'single');
[h,nodes] = dendrogram(clustTreeSng,0);
h_gca = gca;
h_gca.TickDir = 'out';
h_gca.TickLength = [.002 0];
h_gca.XTickLabel = [];
```


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0

Parametric Classification

Parametric Classification

Models of data with a categorical response are called classifiers. A classifier is built from training data, for which classifications are known. The classifier assigns new test data to one of the categorical levels of the response.

Parametric methods, like “Discriminant Analysis Classification” on page 20-2, fit a parametric model to the training data and interpolate to classify test data.

Nonparametric methods, like classification and regression trees, use other means to determine classifications.

See Also

`fitcdiscr` | `fitcnb`

Related Examples

- “Discriminant Analysis Classification” on page 20-2
- “Naive Bayes Classification” on page 21-2

Performance Curves

In this section...

Introduction to Performance Curves

After a classification algorithm such as `ClassificationNaiveBayes` or `TreeBagger` has trained on data, you may want to examine the performance of the algorithm on a specific test dataset. One common way of doing this would be to compute a gross measure of performance such as quadratic loss or accuracy, averaged over the entire test dataset.

What are ROC Curves?

You may want to inspect the classifier performance more closely, for example, by plotting a Receiver Operating Characteristic (ROC) curve. By definition, a ROC curve [1,2] shows true positive rate versus false positive rate (equivalently, sensitivity versus 1–specificity) for different thresholds of the classifier output. You can use it, for example, to find the threshold that maximizes the classification accuracy or to assess, in more broad terms, how the classifier performs in the regions of high sensitivity and high specificity.

Evaluate Classifier Performance Using perfcurve

`perfcurve` computes measures for a plot of classifier performance. You can use this utility to evaluate classifier performance on test data after you train the classifier. Various measures such as mean squared error, classification error, or exponential loss can summarize the predictive power of a classifier in a single number. However, a performance curve offers more information as it lets you explore the classifier performance across a range of thresholds on its output.

You can use `perfcurve` with any classifier or, more broadly, with any method that returns a numeric score for an instance of input data. By convention adopted here,

- A high score returned by a classifier for any given instance signifies that the instance is likely from the positive class.
- A low score signifies that the instance is likely from the negative classes.

For some classifiers, you can interpret the score as the posterior probability of observing an instance of the positive class at point `X`. An example of such a score is the fraction of positive observations in a leaf of a decision tree. In this case, scores fall into the range from 0 to 1 and scores from positive and negative classes add up to unity. Other methods can return scores ranging between minus and plus infinity, without any obvious mapping from the score to the posterior class probability.

`perfcurve` does not impose any requirements on the input score range. Because of this lack of normalization, you can use `perfcurve` to process scores returned by any classification, regression, or fit method. `perfcurve` does not make any assumptions about the nature of input scores or relationships between the scores for different classes. As an example, consider a problem with three classes, `A`, `B`, and `C`, and assume that the scores returned by some classifier for two instances are as follows:

	A	B	C
instance 1	0.4	0.5	0.1
instance 2	0.4	0.1	0.5

If you want to compute a performance curve for separation of classes `A` and `B`, with `C` ignored, you need to address the ambiguity in selecting `A` over `B`. You could opt to use the score ratio, `s(A)/s(B)`, or score difference, `s(A)-s(B)`; this choice could depend on the nature of these scores and their normalization. `perfcurve` always takes one score per instance. If you only supply scores for class `A`, `perfcurve` does not distinguish between observations 1 and 2. The performance curve in this case may not be optimal.

`perfcurve` is intended for use with classifiers that return scores, not those that return only predicted classes. As a counter-example, consider a decision tree that returns only hard classification labels, 0 or 1, for data with two classes. In this case, the performance curve reduces to a single point because classified instances can be split into positive and negative categories in one way only.

For input, `perfcurve` takes true class labels for some data and scores assigned by a classifier to these data. By default, this utility computes a Receiver Operating Characteristic (ROC) curve and returns values of 1–specificity, or false positive rate, for `X` and sensitivity, or true positive rate, for `Y`. You can choose other criteria for `X` and `Y` by selecting one out of several provided criteria or specifying an arbitrary criterion through an anonymous function. You can display the computed performance curve using `plot(X,Y)`.

`perfcurve` can compute values for various criteria to plot either on the *x*- or the *y*-axis. All such criteria are described by a 2-by-2 confusion matrix, a 2-by-2 cost matrix, and a 2-by-1 vector of scales applied to class counts.

The `confusionchart` matrix, `C`, is defined as

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

where

- *P* stands for "positive".
- *N* stands for "negative".
- *T* stands for "true".
- *F* stands for "false".

For example, the first row of the confusion matrix defines how the classifier identifies instances of the positive class: `C(1,1)` is the count of correctly identified positive instances and `C(1,2)` is the count of positive instances misidentified as negative.

The cost matrix defines the cost of misclassification for each category:

$$\begin{pmatrix} Cost(P|P) & Cost(N|P) \\ Cost(P|N) & Cost(N|N) \end{pmatrix}$$

where `Cost(I|J)` is the cost of assigning an instance of class `J` to class `I`. Usually `Cost(I|J)=0` for `I=J`. For flexibility, `perfcurve` allows you to specify nonzero costs for correct classification as well.

The two scales include prior information about class probabilities. `perfcurve` computes these scales by taking `scale(P)=prior(P)*N` and `scale(N)=prior(N)*P` and normalizing the sum `scale(P)`

+scale(N) to 1. P=TP+FN and N=TN+FP are the total instance counts in the positive and negative class, respectively. The function then applies the scales as multiplicative factors to the counts from the corresponding class: perfcurve multiplies counts from the positive class by scale(P) and counts from the negative class by scale(N). Consider, for example, computation of positive predictive value, PPV = TP/(TP+FP). TP counts come from the positive class and FP counts come from the negative class. Therefore, you need to scale TP by scale(P) and FP by scale(N), and the modified formula for PPV with prior probabilities taken into account is now:

$$PPV = \frac{scale(P)*TP}{scale(P)*TP + scale(N)*FP}$$

If all scores in the data are above a certain threshold, perfcurve classifies all instances as 'positive'. This means that TP is the total number of instances in the positive class and FP is the total number of instances in the negative class. In this case, PPV is simply given by the prior:

$$PPV = \frac{prior(P)}{prior(P) + prior(N)}$$

The perfcurve function returns two vectors, X and Y, of performance measures. Each measure is some function of confusion, cost, and scale values. You can request specific measures by name or provide a function handle to compute a custom measure. The function you provide should take confusion, cost, and scale as its three inputs and return a vector of output values.

The criterion for X must be a monotone function of the positive classification count, or equivalently, threshold for the supplied scores. If perfcurve cannot perform a one-to-one mapping between values of the X criterion and score thresholds, it exits with an error message.

By default, perfcurve computes values of the X and Y criteria for all possible score thresholds. Alternatively, it can compute a reduced number of specific X values supplied as an input argument. In either case, for M requested values, perfcurve computes M+1 values for X and Y. The first value out of these M+1 values is special. perfcurve computes it by setting the TP instance count to zero and setting TN to the total count in the negative class. This value corresponds to the 'reject all' threshold. On a standard ROC curve, this translates into an extra point placed at (0,0).

If there are NaN values among input scores, perfcurve can process them in either of two ways:

- It can discard rows with NaN scores.
- It can add them to false classification counts in the respective class.

That is, for any threshold, instances with NaN scores from the positive class are counted as false negative (FN), and instances with NaN scores from the negative class are counted as false positive (FP). In this case, the first value of X or Y is computed by setting TP to zero and setting TN to the total count minus the NaN count in the negative class. For illustration, consider an example with two rows in the positive and two rows in the negative class, each pair having a NaN score:

Class	Score
Negative	0.2
Negative	NaN
Positive	0.7
Positive	NaN

If you discard rows with NaN scores, then as the score cutoff varies, perfcurve computes performance measures as in the following table. For example, a cutoff of 0.5 corresponds to the middle row where rows 1 and 3 are classified correctly, and rows 2 and 4 are omitted.

TP	FN	FP	TN
0	1	0	1
1	0	0	1
1	0	1	0

If you add rows with NaN scores to the false category in their respective classes, perfcurve computes performance measures as in the following table. For example, a cutoff of 0.5 corresponds to the middle row where now rows 2 and 4 are counted as incorrectly classified. Notice that only the FN and FP columns differ between these two tables.

TP	FN	FP	TN
0	2	1	1
1	1	1	1
1	1	2	0

For data with three or more classes, perfcurve takes one positive class and a list of negative classes for input. The function computes the X and Y values using counts in the positive class to estimate TP and FN, and using counts in all negative classes to estimate TN and FP. perfcurve can optionally compute Y values for each negative class separately and, in addition to Y, return a matrix of size M-by-C, where M is the number of elements in X or Y and C is the number of negative classes. You can use this functionality to monitor components of the negative class contribution. For example, you can plot TP counts on the X-axis and FP counts on the Y-axis. In this case, the returned matrix shows how the FP component is split across negative classes.

You can also use perfcurve to estimate confidence intervals. perfcurve computes confidence bounds using either cross-validation or bootstrap. If you supply cell arrays for labels and scores, perfcurve uses cross-validation and treats elements in the cell arrays as cross-validation folds. If you set input parameter NBoot to a positive integer, perfcurve generates nboot bootstrap replicas to compute pointwise confidence bounds.

perfcurve estimates the confidence bounds using one of two methods:

- Vertical averaging (VA) — estimate confidence bounds on Y and T at fixed values of X. Use the XVals input parameter to use this method for computing confidence bounds.
- Threshold averaging (TA) — estimate confidence bounds for X and Y at fixed thresholds for the positive class score. Use the TVals input parameter to use this method for computing confidence bounds.

To use observation weights instead of observation counts, you can use the 'Weights' parameter in your call to perfcurve. When you use this parameter, to compute X, Y and T or to compute confidence bounds by cross-validation, perfcurve uses your supplied observation weights instead of observation counts. To compute confidence bounds by bootstrap, perfcurve samples N out of N with replacement using your weights as multinomial sampling probabilities.

See Also

confusionchart | perfcurve

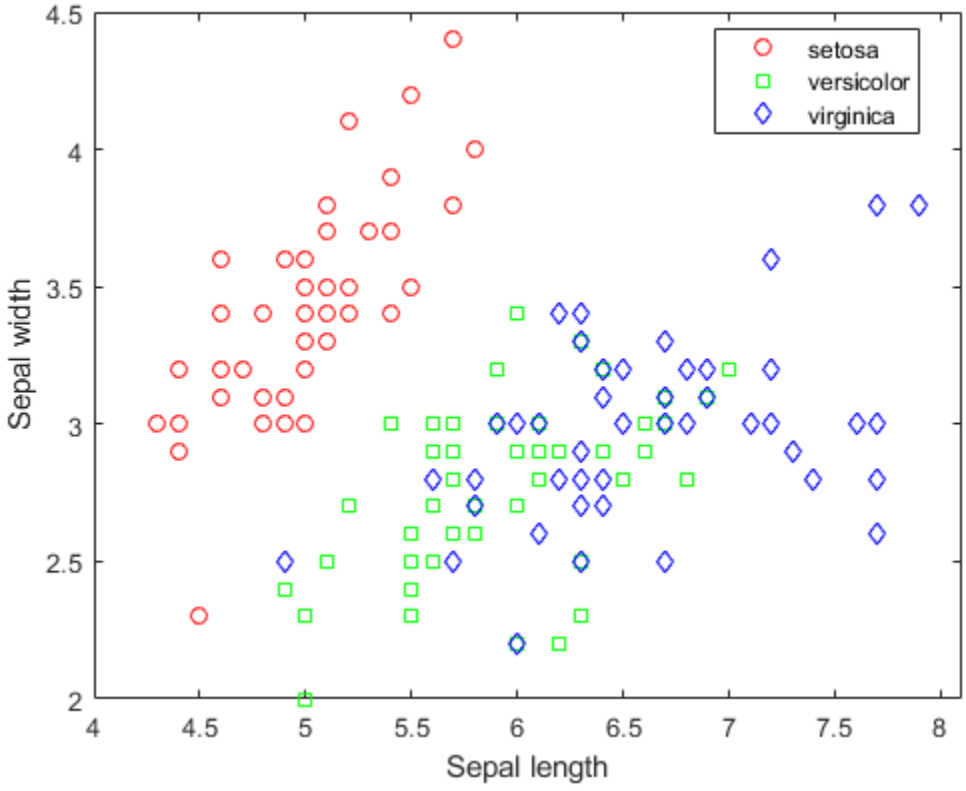
Classification

This example shows how to perform classification using discriminant analysis, naive Bayes classifiers, and decision trees. Suppose you have a data set containing observations with measurements on different variables (called predictors) and their known class labels. If you obtain predictor values for new observations, could you determine to which classes those observations probably belong? This is the problem of classification.

Fisher's Iris Data

Fisher's iris data consists of measurements on the sepal length, sepal width, petal length, and petal width for 150 iris specimens. There are 50 specimens from each of three species. Load the data and see how the sepal measurements differ between species. You can use the two columns containing sepal measurements.

```
load fisheriris
f = figure;
gscatter(meas(:,1), meas(:,2), species,'rgb','osd');
xlabel('Sepal length');
ylabel('Sepal width');
N = size(meas,1);
```



Suppose you measure a sepal and petal from an iris, and you need to determine its species on the basis of those measurements. One approach to solving this problem is known as discriminant analysis.

Linear and Quadratic Discriminant Analysis

The `fitcdiscr` function can perform classification using different types of discriminant analysis. First classify the data using the default linear discriminant analysis (LDA).

```
lda = fitcdiscr(meas(:,1:2),species);
ldaClass = resubPredict(lda);
```

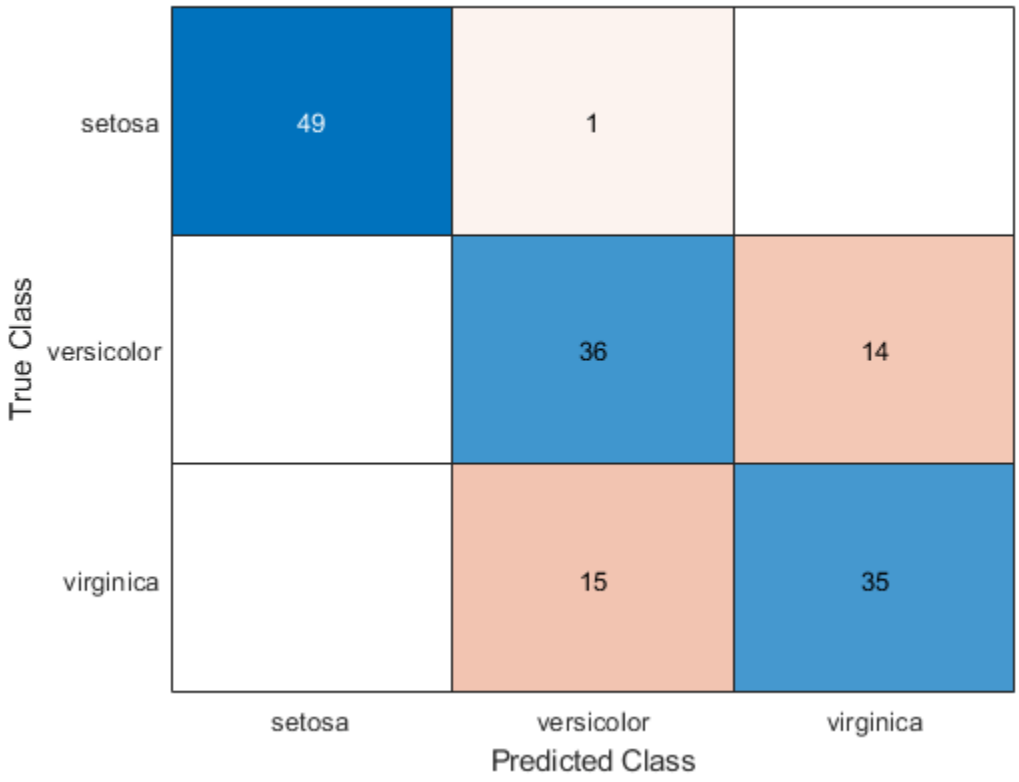
The observations with known class labels are usually called the training data. Now compute the resubstitution error, which is the misclassification error (the proportion of misclassified observations) on the training set.

```
ldaResubErr = resubLoss(lda)
```

```
ldaResubErr =

    0.2000
```

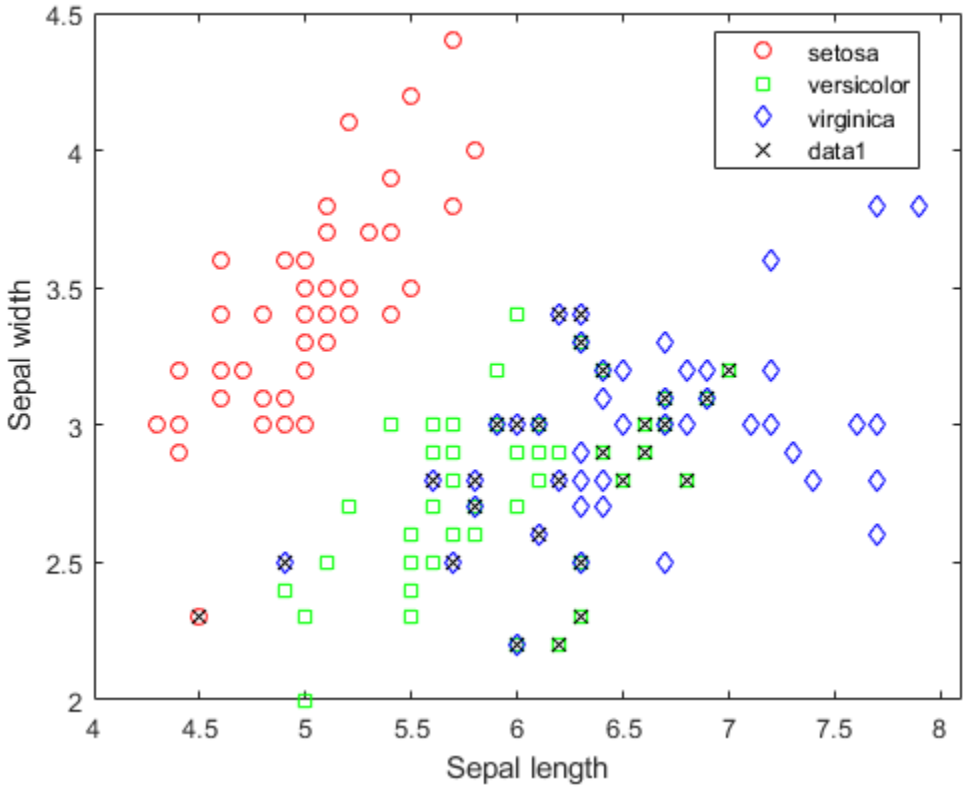
You can also compute the confusion matrix on the training set. A confusion matrix contains information about known class labels and predicted class labels. Generally speaking, the (i,j) element in the confusion matrix is the number of samples whose known class label is class i and whose predicted class is j. The diagonal elements represent correctly classified observations.

```
figure
ldaResubCM = confusionchart(species,ldaClass);
```



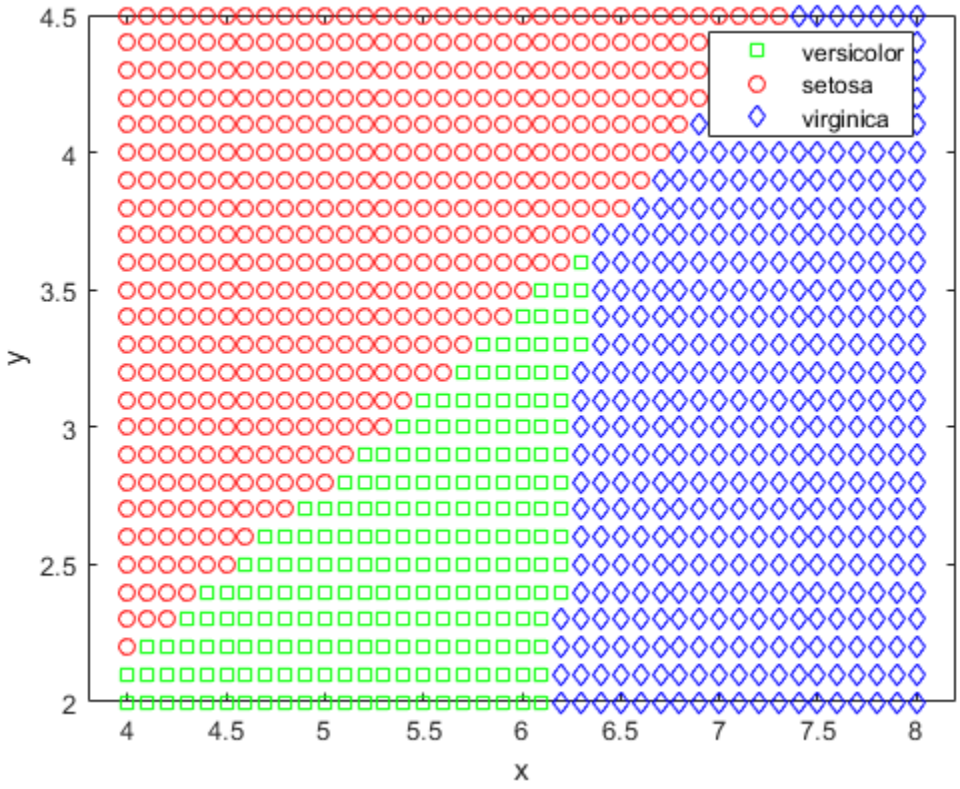
Of the 150 training observations, 20% or 30 observations are misclassified by the linear discriminant function. You can see which ones they are by drawing X through the misclassified points.

```
figure(f)
bad = ~strcmp(ldaClass,species);
hold on;
plot(meas(bad,1), meas(bad,2), 'kx');
hold off;
```



The function has separated the plane into regions divided by lines, and assigned different regions to different species. One way to visualize these regions is to create a grid of (x,y) values and apply the classification function to that grid.

```
[x,y] = meshgrid(4:.1:8,2:.1:4.5);
x = x(:);
y = y(:);
j = classify([x y],meas(:,1:2),species);
gscatter(x,y,j,'grb','sod')
```



For some data sets, the regions for the various classes are not well separated by lines. When that is the case, linear discriminant analysis is not appropriate. Instead, you can try quadratic discriminant analysis (QDA) for our data.

Compute the resubstitution error for quadratic discriminant analysis.

```
qda = fitcdiscr(meas(:,1:2),species,'DiscrimType','quadratic');
qdaResubErr = resubLoss(qda)
```

```
qdaResubErr =

    0.2000
```

You have computed the resubstitution error. Usually people are more interested in the test error (also referred to as generalization error), which is the expected prediction error on an independent set. In fact, the resubstitution error will likely under-estimate the test error.

In this case you don't have another labeled data set, but you can simulate one by doing cross-validation. A stratified 10-fold cross-validation is a popular choice for estimating the test error on classification algorithms. It randomly divides the training set into 10 disjoint subsets. Each subset has roughly equal size and roughly the same class proportions as in the training set. Remove one subset, train the classification model using the other nine subsets, and use the trained model to classify the removed subset. You could repeat this by removing each of the ten subsets one at a time.

Because cross-validation randomly divides data, its outcome depends on the initial random seed. To reproduce the exact results in this example, execute the following command:

```
rng(0,'twister');
```

First use `cvpartition` to generate 10 disjoint stratified subsets.

```
cp = cvpartition(species,'KFold',10)
```

```
cp =

K-fold cross validation partition
   NumObservations: 150
       NumTestSets: 10
         TrainSize: 135  135  135  135  135  135  135  135  135  135
          TestSize: 15  15  15  15  15  15  15  15  15  15
```

The `crossval` and `kfoldLoss` methods can estimate the misclassification error for both LDA and QDA using the given data partition `cp`.

Estimate the true test error for LDA using 10-fold stratified cross-validation.

```
cvlda = crossval(lda,'CVPartition',cp);
ldaCVErr = kfoldLoss(cvlda)
```

```
ldaCVErr =

    0.2000
```

The LDA cross-validation error has the same value as the LDA resubstitution error on this data.

Estimate the true test error for QDA using 10-fold stratified cross-validation.

```
cvqda = crossval(qda,'CVPartition',cp);
qdaCVErr = kfoldLoss(cvqda)
```

```
qdaCVErr =

    0.2200
```

QDA has a slightly larger cross-validation error than LDA. It shows that a simpler model may get comparable, or better performance than a more complicated model.

Naive Bayes Classifiers

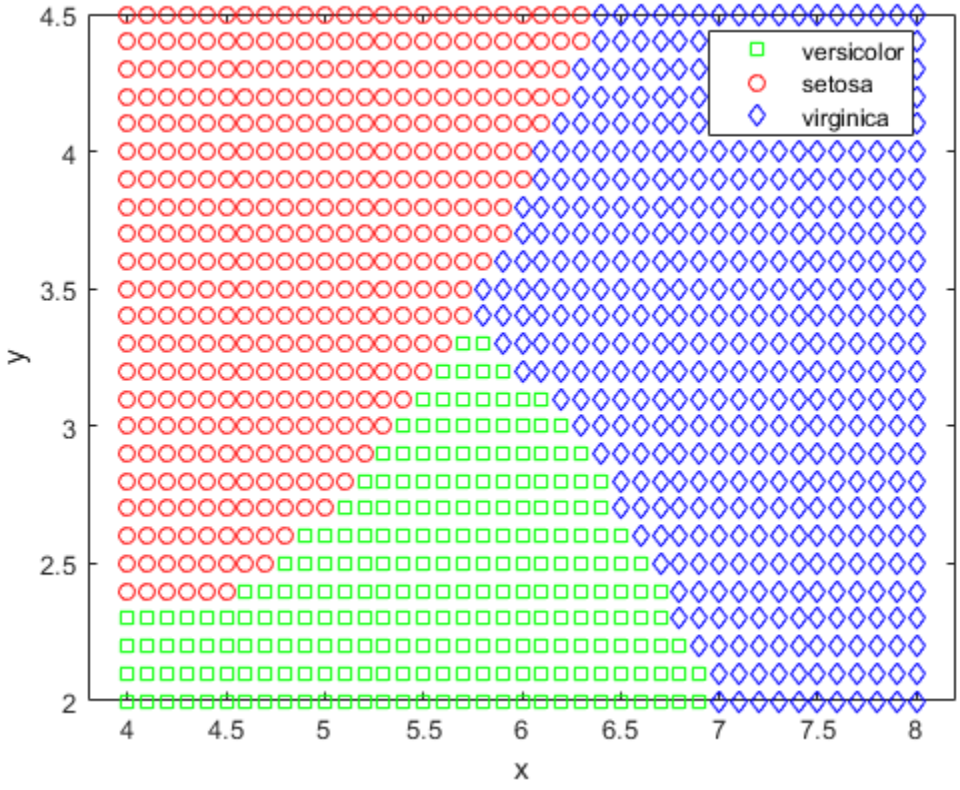
The `fitcdiscr` function has other two other types, 'DiagLinear' and 'DiagQuadratic'. They are similar to 'linear' and 'quadratic', but with diagonal covariance matrix estimates. These diagonal choices are specific examples of a naive Bayes classifier, because they assume the variables are conditionally independent given the class label. Naive Bayes classifiers are among the most popular classifiers. While the assumption of class-conditional independence between variables is not true in general, naive Bayes classifiers have been found to work well in practice on many data sets.

The `fitcnb` function can be used to create a more general type of naive Bayes classifier.

First model each variable in each class using a Gaussian distribution. You can compute the resubstitution error and the cross-validation error.

```
nbGau = fitcnb(meas(:,1:2), species);
nbGauResubErr = resubLoss(nbGau)
nbGauCV = crossval(nbGau, 'CVPartition',cp);
nbGauCVErr = kfoldLoss(nbGauCV)

labels = predict(nbGau, [x y]);
gscatter(x,y,labels,'grb','sod')
```

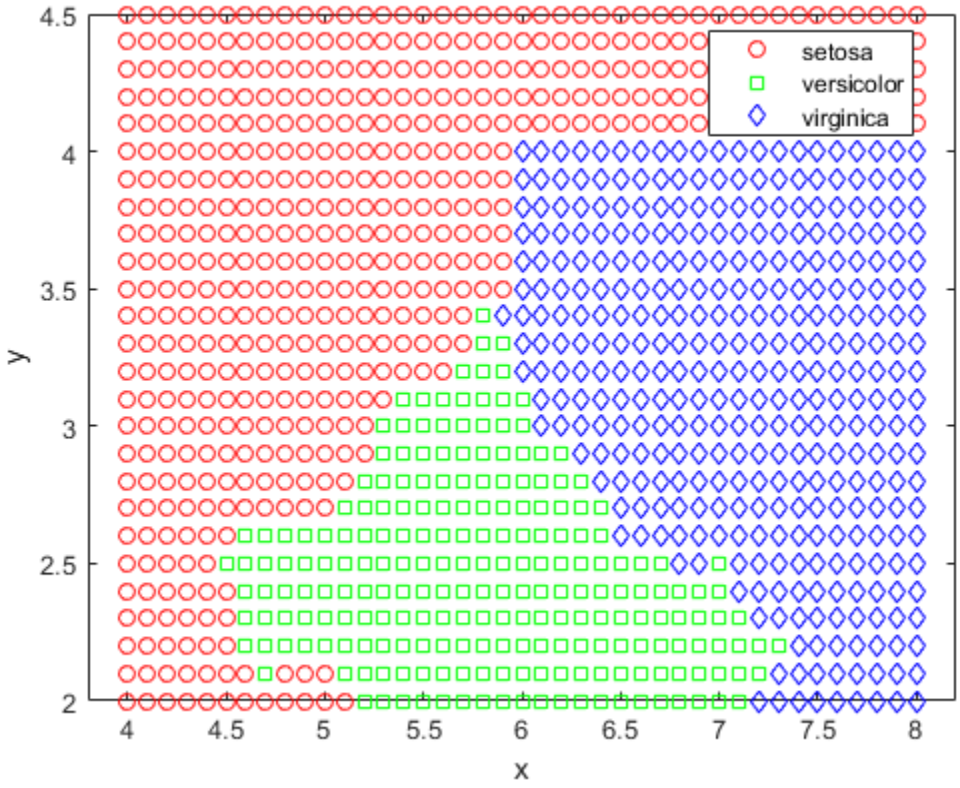
```
nbGauResubErr =

    0.2200


nbGauCVErr =

    0.2200
```



So far you have assumed the variables from each class have a multivariate normal distribution. Often that is a reasonable assumption, but sometimes you may not be willing to make that assumption or you may see clearly that it is not valid. Now try to model each variable in each class using a kernel density estimation, which is a more flexible nonparametric technique. Here we set the kernel to `box`.

```
nbKD = fitcnb(meas(:,1:2), species, 'DistributionNames','kernel', 'Kernel','box');
nbKDResubErr = resubLoss(nbKD)
nbKDCV = crossval(nbKD, 'CVPartition',cp);
nbKDCVErr = kfoldLoss(nbKDCV)

labels = predict(nbKD, [x y]);
gscatter(x,y,labels,'rgb','osd')
```

```
nbKDResubErr =

    0.2067


nbKDCVErr =

    0.2133
```



For this data set, the naive Bayes classifier with kernel density estimation gets smaller resubstitution error and cross-validation error than the naive Bayes classifier with a Gaussian distribution.

Decision Tree

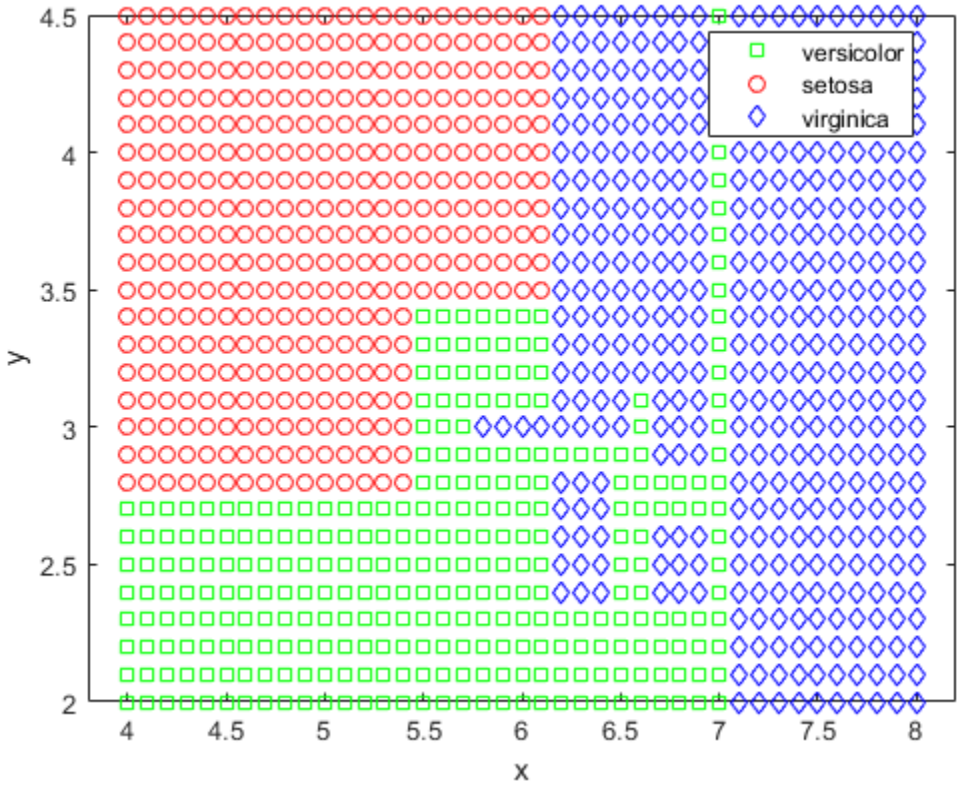
Another classification algorithm is based on a decision tree. A decision tree is a set of simple rules, such as "if the sepal length is less than 5.45, classify the specimen as setosa." Decision trees are also nonparametric because they do not require any assumptions about the distribution of the variables in each class.

The `fitctree` function creates a decision tree. Create a decision tree for the iris data and see how well it classifies the irises into species.

```
t = fitctree(meas(:,1:2), species,'PredictorNames',{'SL' 'SW' });
```

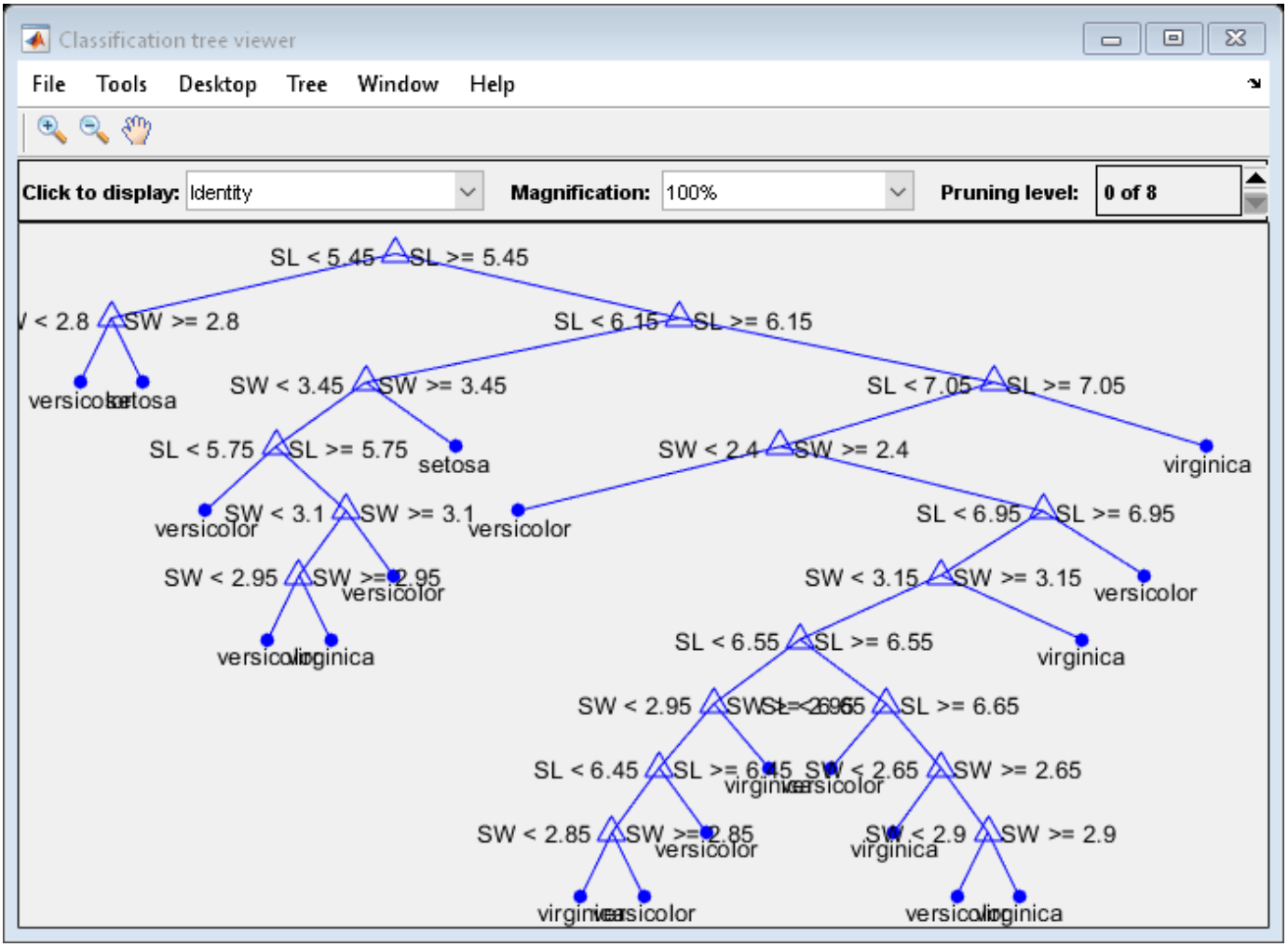
It's interesting to see how the decision tree method divides the plane. Use the same technique as above to visualize the regions assigned to each species.

```
[grpname,node] = predict(t,[x y]);
gscatter(x,y,grpname,'grb','sod')
```



Another way to visualize the decision tree is to draw a diagram of the decision rule and class assignments.

```
view(t,'Mode','graph');
```



This cluttered-looking tree uses a series of rules of the form "SL < 5.45" to classify each specimen into one of 19 terminal nodes. To determine the species assignment for an observation, start at the top node and apply the rule. If the point satisfies the rule you take the left path, and if not you take the right path. Ultimately you reach a terminal node that assigns the observation to one of the three species.

Compute the resubstitution error and the cross-validation error for decision tree.

```
dtResubErr = resubLoss(t)

cvt = crossval(t,'CVPartition',cp);
dtCVErr = kfoldLoss(cvt)
```

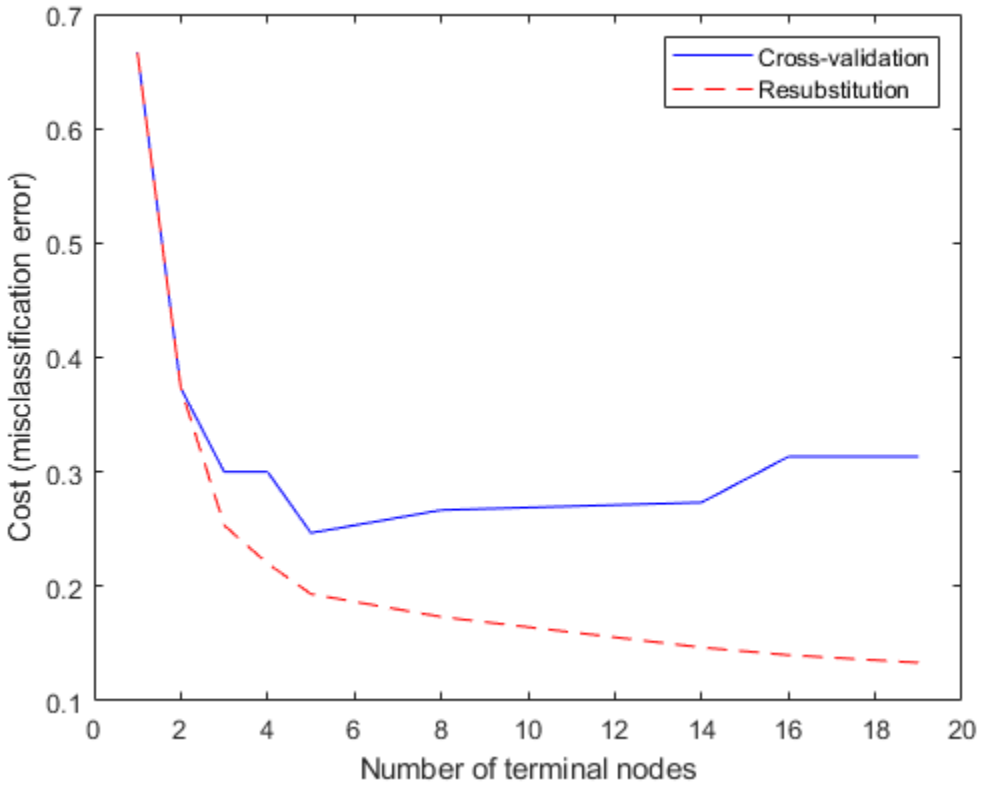
```
dtResubErr =

    0.1333


dtCVErr =
```

```
    0.3000
```

For the decision tree algorithm, the cross-validation error estimate is significantly larger than the resubstitution error. This shows that the generated tree overfits the training set. In other words, this is a tree that classifies the original training set well, but the structure of the tree is sensitive to this particular training set so that its performance on new data is likely to degrade. It is often possible to find a simpler tree that performs better than a more complex tree on new data.

Try pruning the tree. First compute the resubstitution error for various subsets of the original tree. Then compute the cross-validation error for these sub-trees. A graph shows that the resubstitution error is overly optimistic. It always decreases as the tree size grows, but beyond a certain point, increasing the tree size increases the cross-validation error rate.

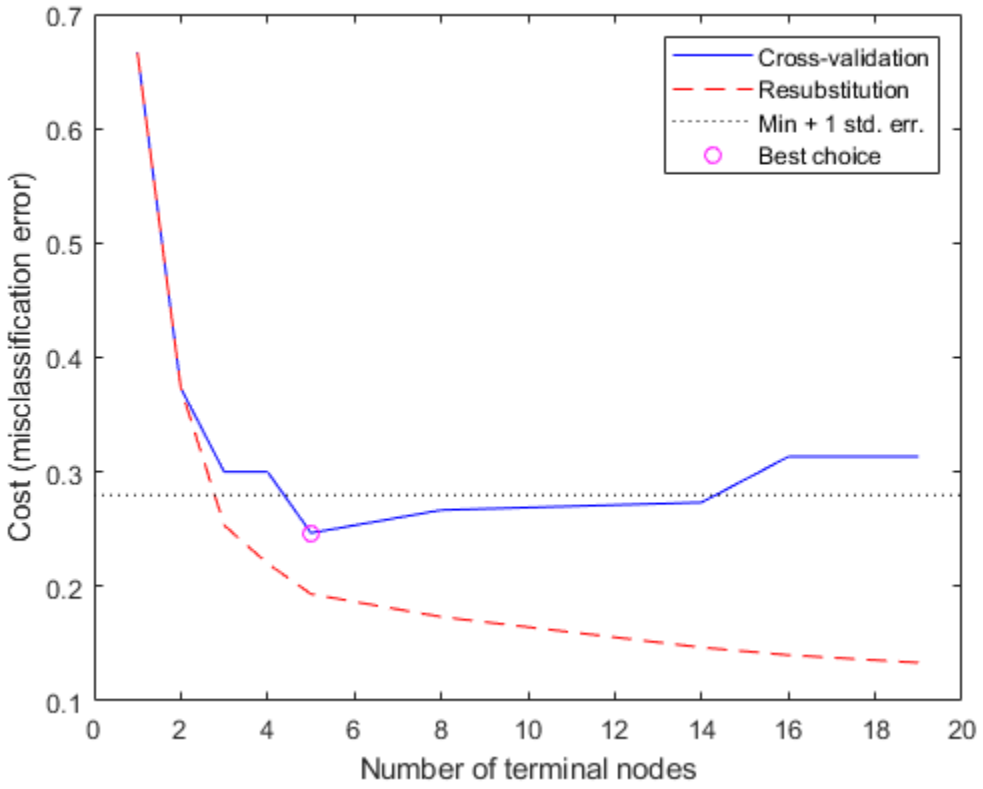
```
resubcost = resubLoss(t,'Subtrees','all');
[cost,secost,ntermnodes,bestlevel] = cvloss(t,'Subtrees','all');
plot(ntermnodes,cost,'b-', ntermnodes,resubcost,'r--')
figure(gcf);
xlabel('Number of terminal nodes');
ylabel('Cost (misclassification error)')
legend('Cross-validation','Resubstitution')
```



Which tree should you choose? A simple rule would be to choose the tree with the smallest cross-validation error. While this may be satisfactory, you might prefer to use a simpler tree if it is roughly as good as a more complex tree. For this example, take the simplest tree that is within one standard error of the minimum. That's the default rule used by the `cvloss` method of `ClassificationTree`.

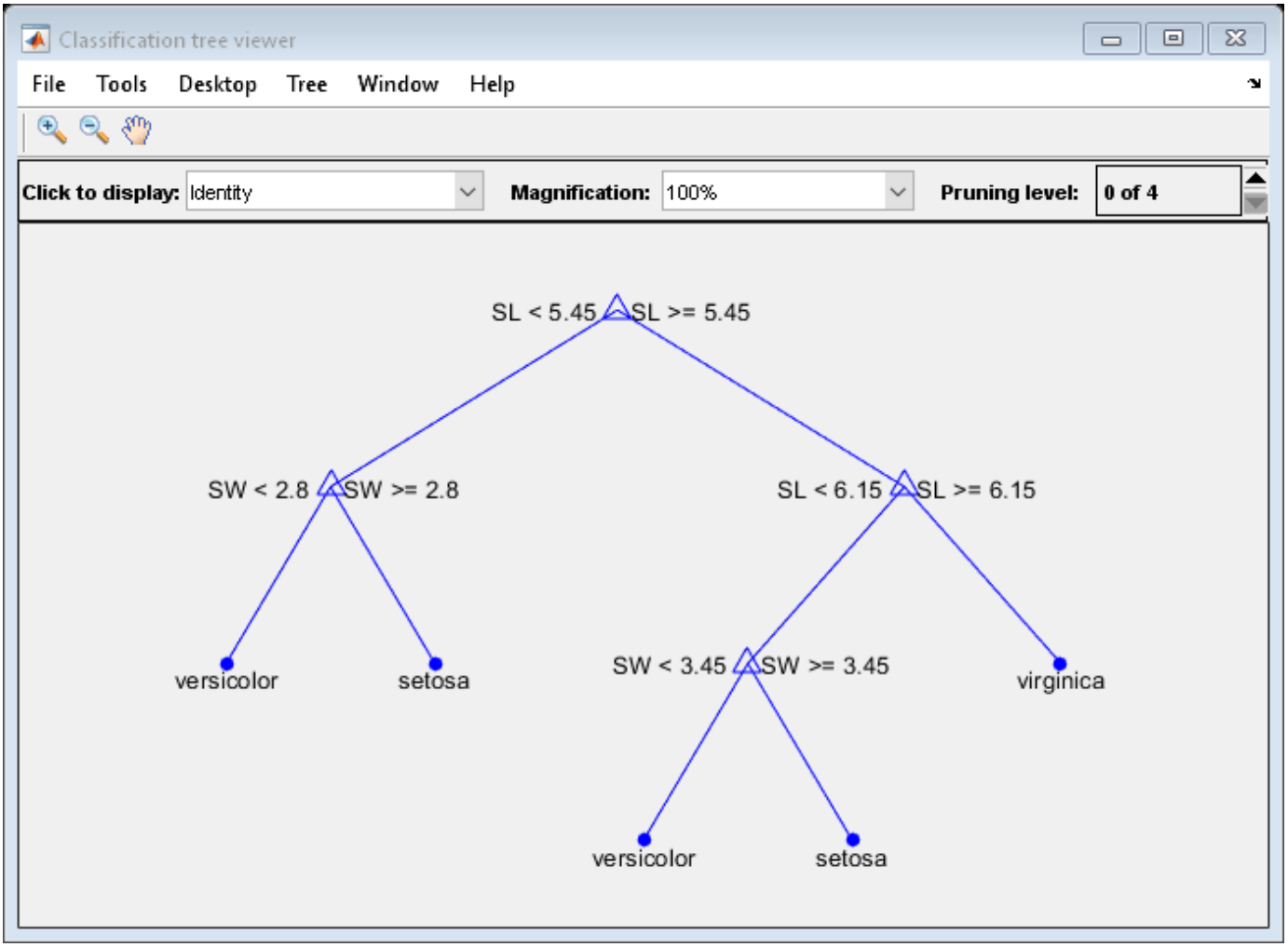
You can show this on the graph by computing a cutoff value that is equal to the minimum cost plus one standard error. The "best" level computed by the `cvloss` method is the smallest tree under this cutoff. (Note that `bestlevel=0` corresponds to the unpruned tree, so you have to add 1 to use it as an index into the vector outputs from `cvloss`.)

```
[mincost,minloc] = min(cost);
cutoff = mincost + secost(minloc);
hold on
plot([0 20], [cutoff cutoff], 'k:')
plot(ntermnodes(bestlevel+1), cost(bestlevel+1), 'mo')
legend('Cross-validation','Resubstitution','Min + 1 std. err.','Best choice')
hold off
```



Finally, you can look at the pruned tree and compute the estimated misclassification error for it.

```
pt = prune(t,'Level',bestlevel);
view(pt,'Mode','graph')
```




```
cost(bestlevel+1)

ans =

    0.2467
```

Conclusions

This example shows how to perform classification in MATLAB® using Statistics and Machine Learning Toolbox™ functions.

This example is not meant to be an ideal analysis of the Fisher iris data, In fact, using the petal measurements instead of, or in addition to, the sepal measurements may lead to better classification. Also, this example is not meant to compare the strengths and weaknesses of different classification algorithms. You may find it instructive to perform the analysis on other data sets and compare different algorithms. There are also Toolbox functions that implement other classification algorithms. For instance, you can use `TreeBagger` to perform bootstrap aggregation for an ensemble of decision trees, as described in the example “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124.

Nonparametric Supervised Learning

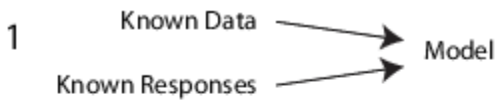
Supervised Learning Workflow and Algorithms

In this section...

What is Supervised Learning?

The aim of supervised, machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. As adaptive algorithms identify patterns in data, a computer "learns" from the observations. When exposed to more observations, the computer improves its predictive performance.

Specifically, a supervised learning algorithm takes a known set of input data and known responses to the data (output), and trains a model to generate reasonable predictions for the response to new data.



For example, suppose you want to predict whether someone will have a heart attack within a year. You have a set of data on previous patients, including age, weight, height, blood pressure, etc. You know whether the previous patients had heart attacks within a year of their measurements. So, the problem is combining all the existing data into a model that can predict whether a new person will have a heart attack within a year.

You can think of the entire set of input data as a heterogeneous matrix. Rows of the matrix are called observations, examples, or instances, and each contain a set of measurements for a subject (patients in the example). Columns of the matrix are called predictors, attributes, or features, and each are variables representing a measurement taken on every subject (age, weight, height, etc. in the example). You can think of the response data as a column vector where each row contains the output of the corresponding observation in the input data (whether the patient had a heart attack). To fit or train a supervised learning model, choose an appropriate algorithm, and then pass the input and response data to it.

Supervised learning splits into two broad categories: classification and regression.

- In classification, the goal is to assign a class (or label) from a finite set of classes to an observation. That is, responses are categorical variables. Applications include spam filters, advertisement recommendation systems, and image and speech recognition. Predicting whether a patient will have a heart attack within a year is a classification problem, and the possible classes

are `true` and `false`. Classification algorithms usually apply to nominal response values. However, some algorithms can accommodate ordinal classes (see `fitcecoc`).

- In regression, the goal is to predict a continuous measurement for an observation. That is, the responses variables are real numbers. Applications include forecasting stock prices, energy consumption, or disease incidence.

Statistics and Machine Learning Toolbox supervised learning functionalities comprise a stream-lined, object framework. You can efficiently train a variety of algorithms, combine models into an ensemble, assess model performances, cross-validate, and predict responses for new data.

Steps in Supervised Learning

While there are many Statistics and Machine Learning Toolbox algorithms for supervised learning, most use the same basic workflow for obtaining a predictor model. (Detailed instruction on the steps for ensemble learning is in “Framework for Ensemble Learning” on page 18-31.) The steps for supervised learning are:

1. “Prepare Data” on page 18-4
2. “Choose an Algorithm” on page 18-5
3. “Fit a Model” on page 18-5
4. “Choose a Validation Method” on page 18-5
5. “Examine Fit and Update Until Satisfied” on page 18-6
6. “Use Fitted Model for Predictions” on page 18-6

Prepare Data

All supervised learning methods start with an input data matrix, usually called `X` here. Each row of `X` represents one observation. Each column of `X` represents one variable, or predictor. Represent missing entries with `NaN` values in `X`. Statistics and Machine Learning Toolbox supervised learning algorithms can handle `NaN` values, either by ignoring them or by ignoring any row with a `NaN` value.

You can use various data types for response data `Y`. Each element in `Y` represents the response to the corresponding row of `X`. Observations with missing `Y` data are ignored.

- For regression, `Y` must be a numeric vector with the same number of elements as the number of rows of `X`.
- For classification, `Y` can be any of these data types. This table also contains the method of including missing entries.

Data Type	Missing Entry
Numeric vector	`NaN`
Categorical vector	`<undefined>`
Character array	Row of spaces
String array	`<missing>` or `""`
Cell array of character vectors	`''`
Logical vector	(Cannot represent)

Choose an Algorithm

There are tradeoffs between several characteristics of algorithms, such as:

- Speed of training
- Memory usage
- Predictive accuracy on new data
- Transparency or interpretability, meaning how easily you can understand the reasons an algorithm makes its predictions

Details of the algorithms appear in “Characteristics of Classification Algorithms” on page 18-7. More detail about ensemble algorithms is in “Choose an Applicable Ensemble Aggregation Method” on page 18-32.

Fit a Model

The fitting function you use depends on the algorithm you choose.

Algorithm	Fitting Function
Classification Trees	`fitctree`
Regression Trees	`fitrtree`
Discriminant Analysis (classification)	`fitcdiscr`
k-Nearest Neighbors (classification)	`fitcknn`
Naive Bayes (classification)	`fitcnb`
Support Vector Machines (SVM) for classification	`fitcsvm`
SVM for regression	`fitrsvm`
Multiclass models for SVM or other classifiers	`fitcecoc`
Classification Ensembles	`fitcensemble`
Regression Ensembles	`fitrensemble`
Classification or Regression Tree Ensembles (e.g., Random Forests [1])	`TreeBagger`

For a comparison of these algorithms, see “Characteristics of Classification Algorithms” on page 18-7.

Choose a Validation Method

The three main methods to examine the accuracy of the resulting fitted model are:

- Examine the resubstitution error. For examples, see:
  - “Classification Tree Resubstitution Error” on page 19-13
  - “Cross Validate a Regression Tree” on page 19-14
  - “Test Ensemble Quality” on page 18-66
  - “Example: Resubstitution Error of a Discriminant Analysis Classifier” on page 20-16
- Examine the cross-validation error. For examples, see:

- “Cross Validate a Regression Tree” on page 19-14
- “Test Ensemble Quality” on page 18-66
- “Estimate Generalization Error of Boosting Ensemble” on page 33-1653
- “Cross Validating a Discriminant Analysis Classifier” on page 20-17

- Examine the out-of-bag error for bagged decision trees. For examples, see:
  - “Test Ensemble Quality” on page 18-66
  - “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
  - “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

Examine Fit and Update Until Satisfied

After validating the model, you might want to change it for better accuracy, better speed, or to use less memory.

- Change fitting parameters to try to get a more accurate model. For examples, see:
  - “Tune RobustBoost” on page 18-101
  - “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84
  - “Improving Discriminant Analysis Models” on page 20-15
- Change fitting parameters to try to get a smaller model. This sometimes gives a model with more accuracy. For examples, see:
  - “Select Appropriate Tree Depth” on page 19-16
  - “Prune a Classification Tree” on page 19-20
  - “Surrogate Splits” on page 18-91
  - “Ensemble Regularization” on page 18-70
  - “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
  - “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
- Try a different algorithm. For applicable choices, see:
  - “Characteristics of Classification Algorithms” on page 18-7
  - “Choose an Applicable Ensemble Aggregation Method” on page 18-32

When satisfied with a model of some types, you can trim it using the appropriate `compact` function (`compact` for classification trees, `compact` for regression trees, `compact` for discriminant analysis, `compact` for naive Bayes, `compact` for SVM, `compact` for ECOC models, `compact` for classification ensembles, and `compact` for regression ensembles). `compact` removes training data and other properties not required for prediction, e.g., pruning information for decision trees, from the model to reduce memory consumption. Because *k*NN classification models require all of the training data to predict labels, you cannot reduce the size of a `ClassificationKNN` model.

Use Fitted Model for Predictions

To predict classification or regression response for most fitted models, use the `predict` method:

```
Ypredicted = predict(obj,Xnew)
```

- `obj` is the fitted model or fitted compact model.
- `Xnew` is the new input data.
- `Ypredicted` is the predicted response, either classification or regression.

Characteristics of Classification Algorithms

This table shows typical characteristics of the various supervised learning algorithms. The characteristics in any particular case can vary from the listed ones. Use the table as a guide for your initial choice of algorithms. Decide on the tradeoff you want in speed, memory usage, flexibility, and interpretability.

Tip Try a decision tree or discriminant first, because these classifiers are fast and easy to interpret. If the models are not accurate enough predicting the response, try other classifiers with higher flexibility.

To control flexibility, see the details for each classifier type. To avoid overfitting, look for a model of lower flexibility that provides sufficient accuracy.

Classifier	Multiclass Support	Categorical Predictor Support on page 18-8	Prediction Speed	Memory Usage	Interpretability
"Decision Trees" on page 19-2 — `fitctree`	Yes	Yes	Fast	Small	Easy
Discriminant analysis on page 20-2 — `fitcdiscr`	Yes	No	Fast	Small for linear, large for quadratic	Easy
SVM on page 18-150 — `fitcsvm`	No. Combine multiple binary SVM classifiers using `fitcecoc`.	Yes	Medium for linear. Slow for others.	Medium for linear. All others: medium for multiclass, large for binary.	Easy for linear SVM. Hard for all other kernel types.
Naive Bayes on page 21-2 — `fitcnb`	Yes	Yes	Medium for simple distributions. Slow for kernel distributions or high-dimensional data	Small for simple distributions. Medium for kernel distributions or high-dimensional data	Easy
Nearest neighbor on page 18-12 — `fitcknn`	Yes	Yes	Slow for cubic. Medium for others.	Medium	Hard

Classifier	Multiclass Support	Categorical Predictor Support on page 18-8	Prediction Speed	Memory Usage	Interpretability
Ensembles on page 18-31 — `fitcensemble` and `fitrensemble`	Yes	Yes	Fast to medium depending on choice of algorithm	Low to high depending on choice of algorithm.	Hard

The results in this table are based on an analysis of many data sets. The data sets in the study have up to 7000 observations, 80 predictors, and 50 classes. This list defines the terms in the table.

Speed:

- Fast — 0.01 second
- Medium — 1 second
- Slow — 100 seconds

Memory

- Small — 1MB
- Medium — 4MB
- Large — 100MB

Note The table provides a general guide. Your results depend on your data and the speed of your machine.

Categorical Predictor Support

This table describes the data-type support of predictors for each classifier.

Classifier	All predictors numeric	All predictors categorical	Some categorical, some numeric
Decision Trees	Yes	Yes	Yes
Discriminant Analysis	Yes	No	No
SVM	Yes	Yes	Yes
Naive Bayes	Yes	Yes	Yes
Nearest Neighbor	Euclidean distance only	Hamming distance only	No
Ensembles	Yes	Yes, except subspace ensembles of discriminant analysis classifiers	Yes, except subspace ensembles

References

[1] Breiman, L. "Random Forests." *Machine Learning* 45, 2001, pp. 5–32.

Visualize Decision Surfaces of Different Classifiers

This example shows how to plot the decision surface of different classification algorithms.

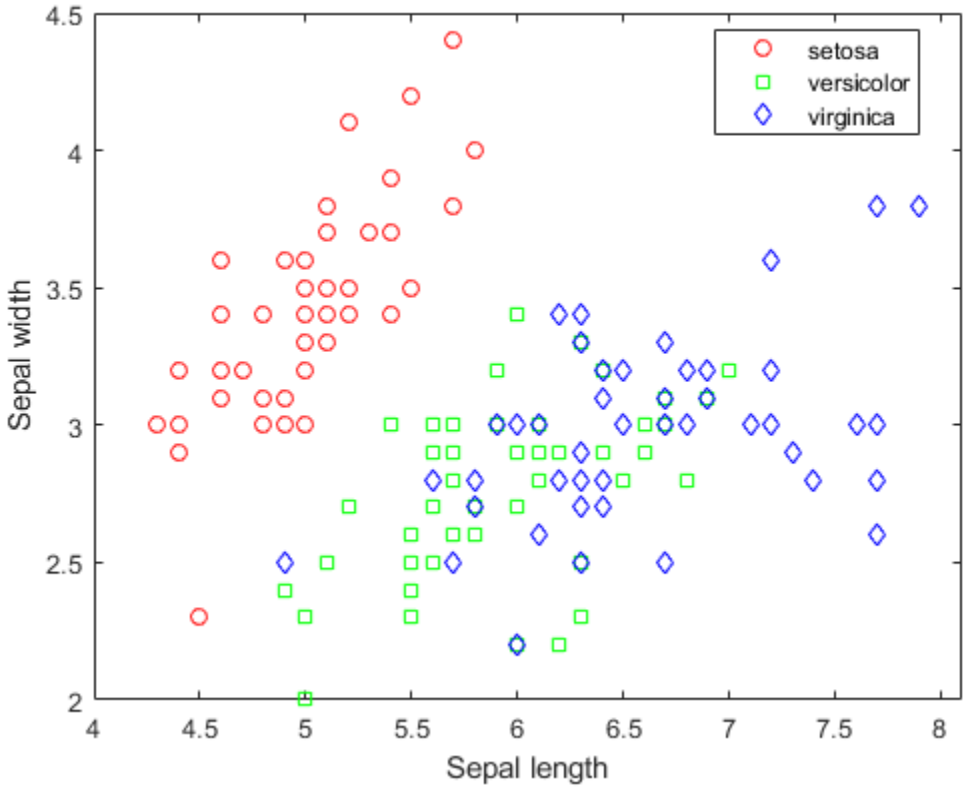
Load Fisher's iris data set.

```
load fisheriris
X = meas(:,1:2);
y = categorical(species);
labels = categories(y);
```

X is a numeric matrix that contains two petal measurements for 150 irises. Y is a cell array of character vectors that contains the corresponding iris species.

Visualize the data using a scatter plot. Group the variables by iris species.

```
gscatter(X(:,1),X(:,2),species,'rgb','osd');
xlabel('Sepal length');
ylabel('Sepal width');
```



Train four different classifiers and store the models in a cell array.

```
classifier_name = {'Naive Bayes','Discriminant Analysis','Classification Tree','Nearest Neighbor
```

Train a naive Bayes model.

```
classifier{1} = fitcnb(X,y);
```

Train a discriminant analysis classifier.

```
classifier{2} = fitcdiscr(X,y);
```

Train a classification decision tree.

```
classifier{3} = fitctree(X,y);
```

Train a *k*-nearest neighbor classifier.

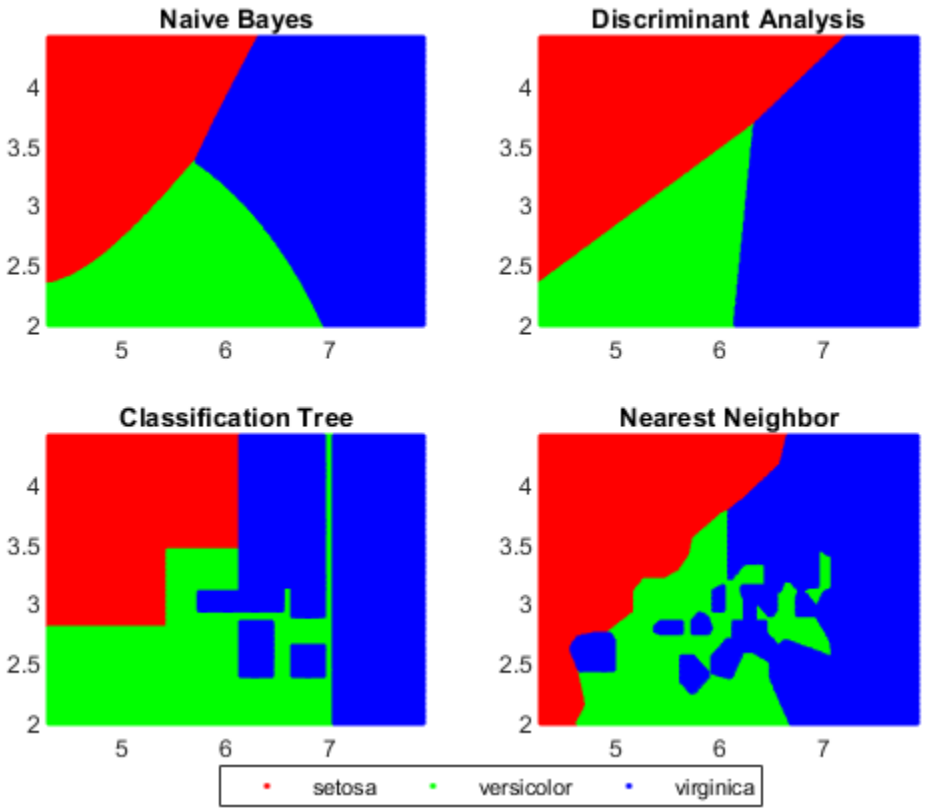
```
classifier{4} = fitcknn(X,y);
```

Create a grid of points spanning the entire space within some bounds of the actual data values.

```
x1range = min(X(:,1)):.01:max(X(:,1));
x2range = min(X(:,2)):.01:max(X(:,2));
[xx1, xx2] = meshgrid(x1range,x2range);
XGrid = [xx1(:) xx2(:)];
```

Predict the iris species of each observation in `XGrid` using all classifiers. Plot the a scatter plot of the results.

```
for i = 1:numel(classifier)
   predictedspecies = predict(classifier{i},XGrid);

   subplot(2,2,i);
   gscatter(xx1(:), xx2(:), predictedspecies,'rgb');

   title(classifier_name{i})
   legend off, axis tight
end

legend(labels,'Location',[0.35,0.01,0.35,0.05],'Orientation','Horizontal')
```



Each classification algorithm generates different decision making rules. A decision surface can help you visualize these rules.

See Also

Functions

`fitcdiscr` | `fitcknn` | `fitcnb` | `fitctree`

Related Examples

- “Plot Posterior Classification Probabilities” on page 21-5

Classification Using Nearest Neighbors

In this section...

Pairwise Distance Metrics

Categorizing query points based on their distance to points in a training data set can be a simple yet effective way of classifying new points. You can use various metrics to determine the distance, described next. Use `pdist2` to find the distance between a set of data and query points.

Distance Metrics

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, x_2, ..., x_{mx}, and an my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, y_2, ...,y_{my}, the various distances between the vector x_s and y_t are defined as follows:

- Euclidean distance

  $$d_{st}^2 = (x_s - y_t)(x_s - y_t)' .$$

  The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.
- Standardized Euclidean distance

  $$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)',$$

  where V is the n-by-n diagonal matrix whose jth diagonal element is $(S(j))^2$, where S is a vector of scaling factors for each dimension.
- Mahalanobis distance

  $$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)',$$

  where C is the covariance matrix.
- City block distance

  $$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right| .$$

  The city block distance is a special case of the Minkowski distance, where $p = 1$.

- Minkowski distance

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left|x_{sj} - y_{tj}\right|^p}.$$

For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.

- Chebychev distance

$$d_{st} = \max_j\left\{\left|x_{sj} - y_{tj}\right|\right\}.$$

The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- Cosine distance

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right).$$

- Correlation distance

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}},$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

and

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}.$$

- Hamming distance

$$d_{st} = \left(\#\left(x_{sj} \neq y_{tj}\right)/n\right).$$

- Jaccard distance

$$d_{st} = \frac{\#\left[\left(x_{sj} \neq y_{tj}\right) \cap \left(\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right)\right]}{\#\left[\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right]}.$$

- Spearman distance

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$, as computed by `tiedrank`.
  - r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, that is, $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.

- $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.
- $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

k-Nearest Neighbor Search and Radius Search

Given a set X of n points and a distance function, k-nearest neighbor (kNN) search lets you find the k closest points in X to a query point or set of points `Y`. The kNN search technique and kNN-based algorithms are widely used as benchmark learning rules. The relative simplicity of the kNN search technique makes it easy to compare the results from other classification techniques to kNN results. The technique has been used in various areas such as:

- bioinformatics
- image processing and data compression
- document retrieval
- computer vision
- multimedia database
- marketing data analysis

You can use kNN search for other machine learning algorithms, such as:

- kNN classification
- local weighted regression
- missing data imputation and interpolation
- density estimation

You can also use kNN search with many distance-based learning functions, such as K-means clustering.

In contrast, for a positive real value `r`, `rangesearch` finds all points in `X` that are within a distance `r` of each point in `Y`. This fixed-radius search is closely related to kNN search, as it supports the same distance metrics and search classes, and uses the same search algorithms.

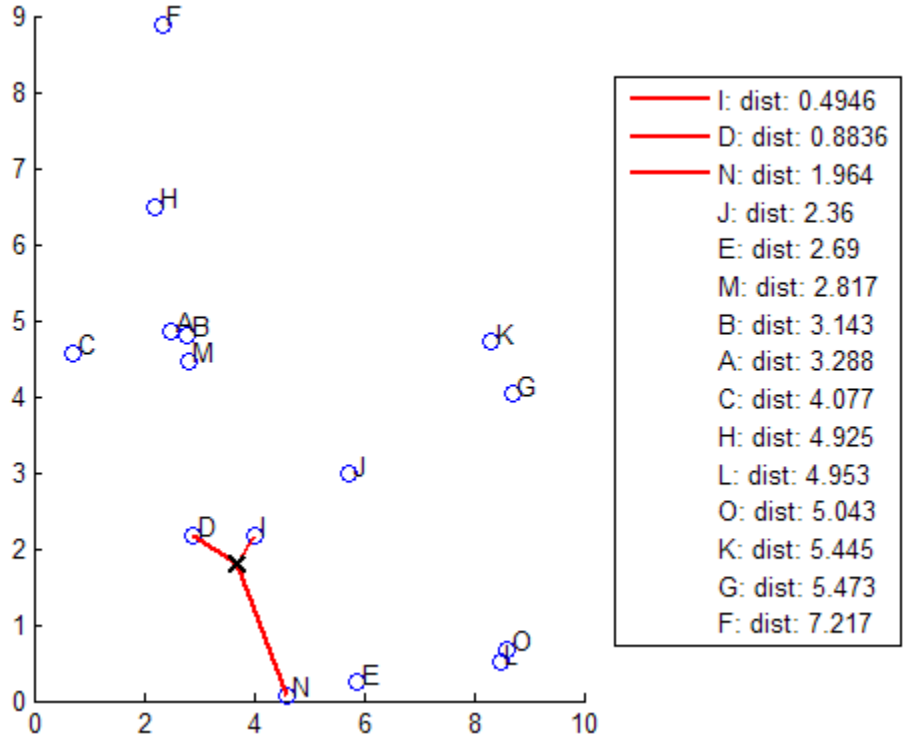
k-Nearest Neighbor Search Using Exhaustive Search

When your input data meets any of the following criteria, `knnsearch` uses the exhaustive search method by default to find the k-nearest neighbors:

- The number of columns of `X` is more than 10.
- `X` is sparse.
- The distance metric is either:
  - `'seuclidean'`
  - `'mahalanobis'`
  - `'cosine'`
  - `'correlation'`
  - `'spearman'`

- `'hamming'`
- `'jaccard'`
- A custom distance function

`knnsearch` also uses the exhaustive search method if your search object is an `ExhaustiveSearcher` model object. The exhaustive search method finds the distance from each query point to every point in `X`, ranks them in ascending order, and returns the k points with the smallest distances. For example, this diagram shows the $k = 3$ nearest neighbors.



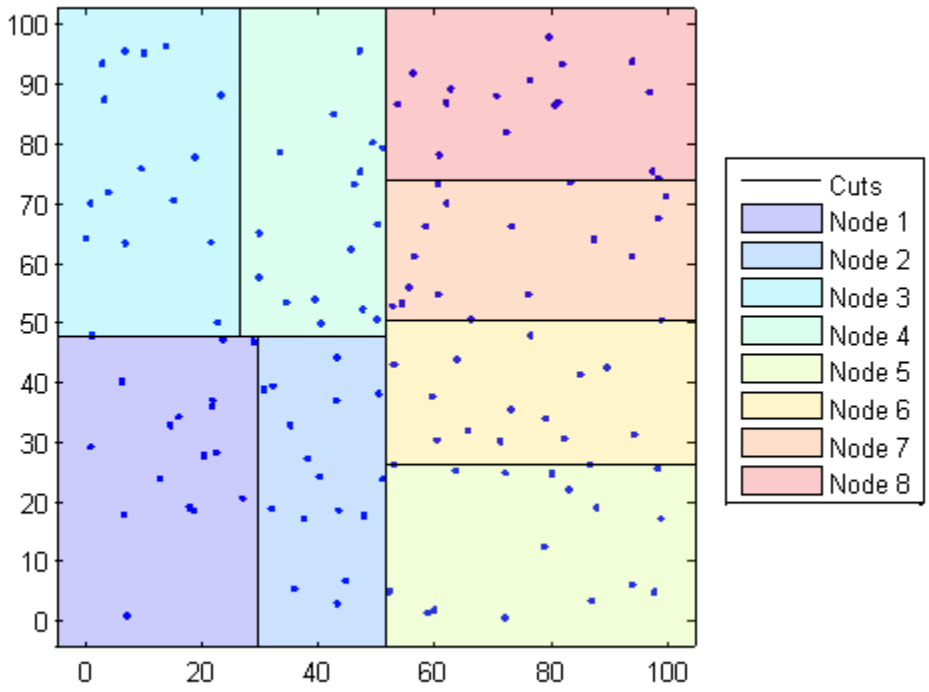
k-Nearest Neighbor Search Using a Kd-Tree

When your input data meets all of the following criteria, `knnsearch` creates a *K*d-tree by default to find the k-nearest neighbors:

- The number of columns of `X` is less than 10.
- `X` is not sparse.
- The distance metric is either:
  - `'euclidean'` (default)
  - `'cityblock'`
  - `'minkowski'`
  - `'chebychev'`

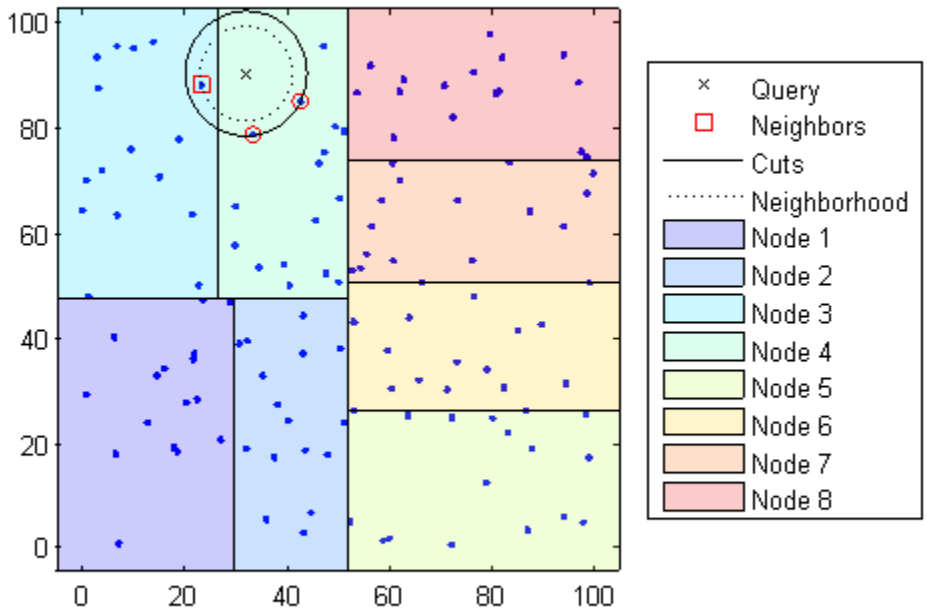
`knnsearch` also uses a *K*d-tree if your search object is a `KDTreeSearcher` model object.

*K*d-trees divide your data into nodes with at most `BucketSize` (default is 50) points per node, based on coordinates (as opposed to categories). The following diagrams illustrate this concept using `patch` objects to color code the different "buckets."



When you want to find the *k*-nearest neighbors to a given query point, `knnsearch` does the following:

1. Determines the node to which the query point belongs. In the following example, the query point (32,90) belongs to Node 4.
2. Finds the closest *k* points within that node and its distance to the query point. In the following example, the points in red circles are equidistant from the query point, and are the closest points to the query point within Node 4.
3. Chooses all other nodes having any area that is within the same distance, in any direction, from the query point to the *k*th closest point. In this example, only Node 3 overlaps the solid black circle centered at the query point with radius equal to the distance to the closest points within Node 4.
4. Searches nodes within that range for any points closer to the query point. In the following example, the point in a red square is slightly closer to the query point than those within Node 4.



Using a *K*d-tree for large data sets with fewer than 10 dimensions (columns) can be much more efficient than using the exhaustive search method, as `knnsearch` needs to calculate only a subset of the distances. To maximize the efficiency of *K*d-trees, use a `KDTreeSearcher` model.

What Are Search Model Objects?

Basically, model objects are a convenient way of storing information. Related models have the same properties with values and types relevant to a specified search method. In addition to storing information within models, you can perform certain actions on models.

You can efficiently perform a *k*-nearest neighbors search on your search model using `knnsearch`. Or, you can search for all neighbors within a specified radius using your search model and `rangesearch`. In addition, there are a generic `knnsearch` and `rangesearch` functions that search without creating or using a model.

To determine which type of model and search method is best for your data, consider the following:

- Does your data have many columns, say more than 10? The `ExhaustiveSearcher` model may perform better.
- Is your data sparse? Use the `ExhaustiveSearcher` model.
- Do you want to use one of these distance metrics to find the nearest neighbors? Use the `ExhaustiveSearcher` model.
  - `'seuclidean'`

  - `'mahalanobis'`
  - `'cosine'`
  - `'correlation'`
  - `'spearman'`
  - `'hamming'`
  - `'jaccard'`
  - A custom distance function
- Is your data set huge (but with fewer than 10 columns)? Use the `KDTreeSearcher` model.
- Are you searching for the nearest neighbors for a large number of query points? Use the `KDTreeSearcher` model.

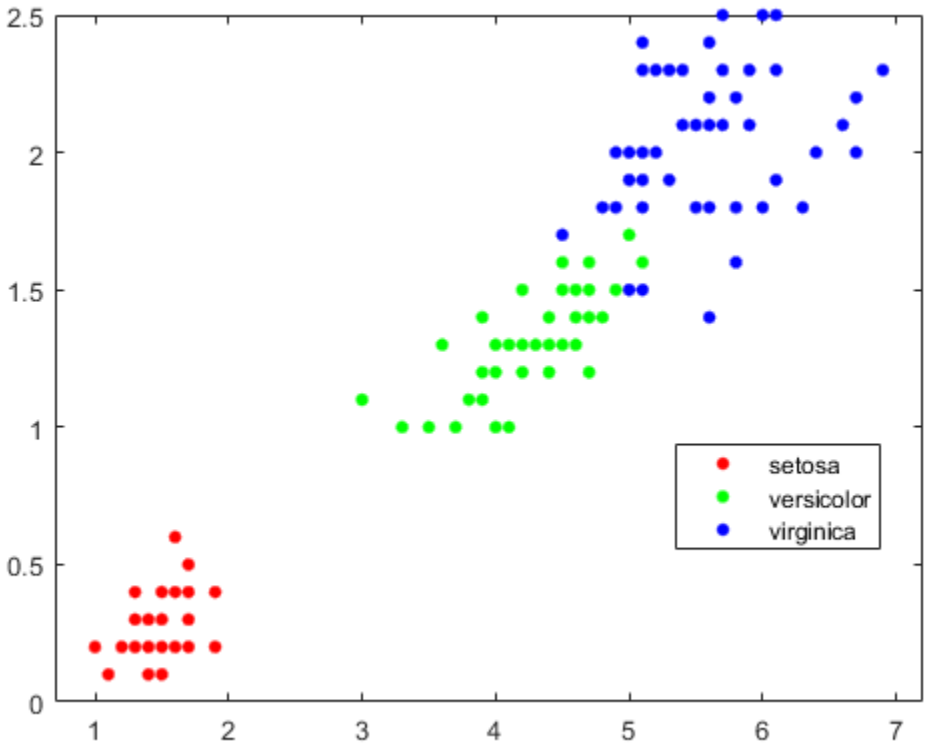
Classify Query Data

This example shows how to classify query data by:

1. Growing a *K*d-tree
2. Conducting a *k* nearest neighbor search using the grown tree.
3. Assigning each query point the class with the highest representation among their respective nearest neighbors.

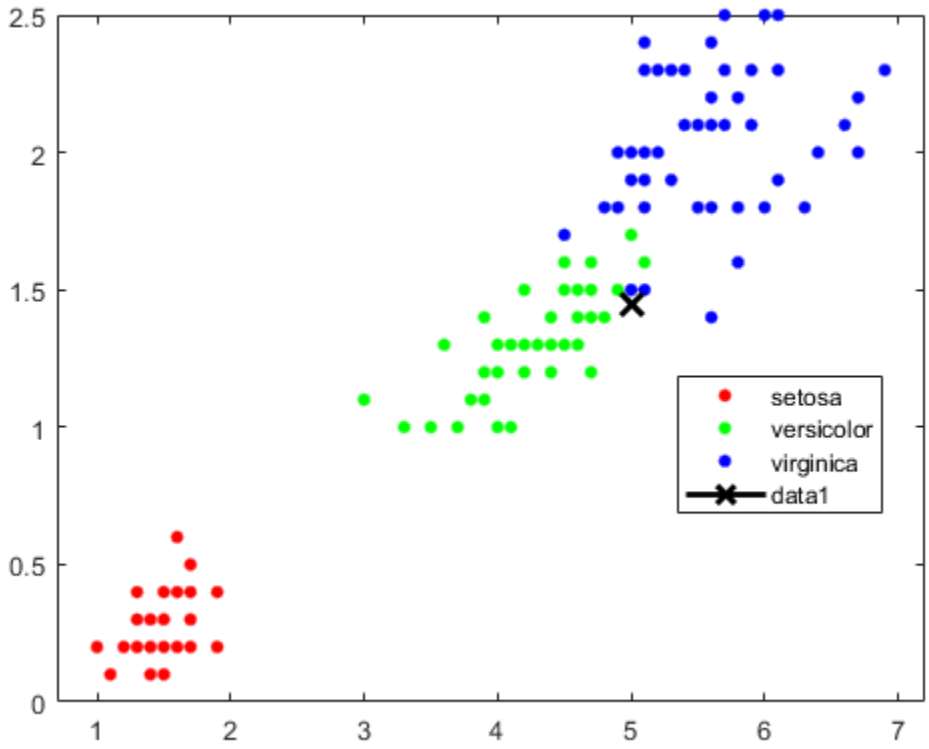
Classify a new point based on the last two columns of the Fisher iris data. Using only the last two columns makes it easier to plot.

```
load fisheriris
x = meas(:,3:4);
gscatter(x(:,1),x(:,2),species)
legend('Location','best')
```



Plot the new point.

```
newpoint = [5 1.45];
line(newpoint(1),newpoint(2),'marker','x','color','k',...
   'markersize',10,'linewidth',2)
```



Prepare a *K*d-tree neighbor searcher model.

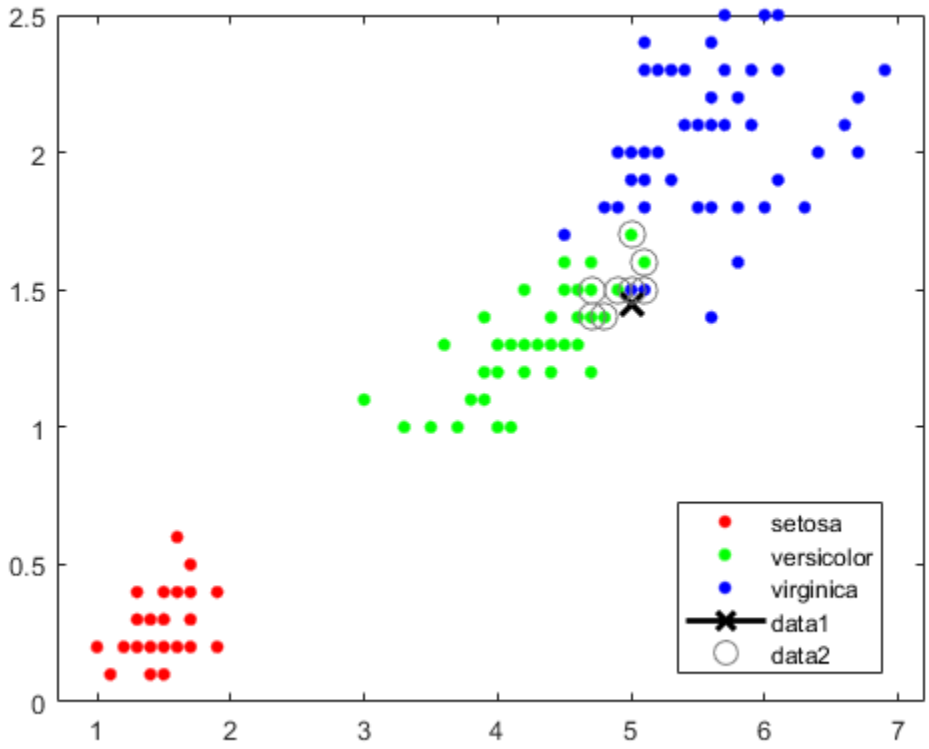
```
Mdl = KDTreeSearcher(x)
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
                X: [150x2 double]
```

`Mdl` is a `KDTreeSearcher` model. By default, the distance metric it uses to search for neighbors is Euclidean distance.

Find the 10 sample points closest to the new point.

```
[n,d] = knnsearch(Mdl,newpoint,'k',10);
line(x(n,1),x(n,2),'color',[.5 .5 .5],'marker','o',...
     'linestyle','none','markersize',10)
```



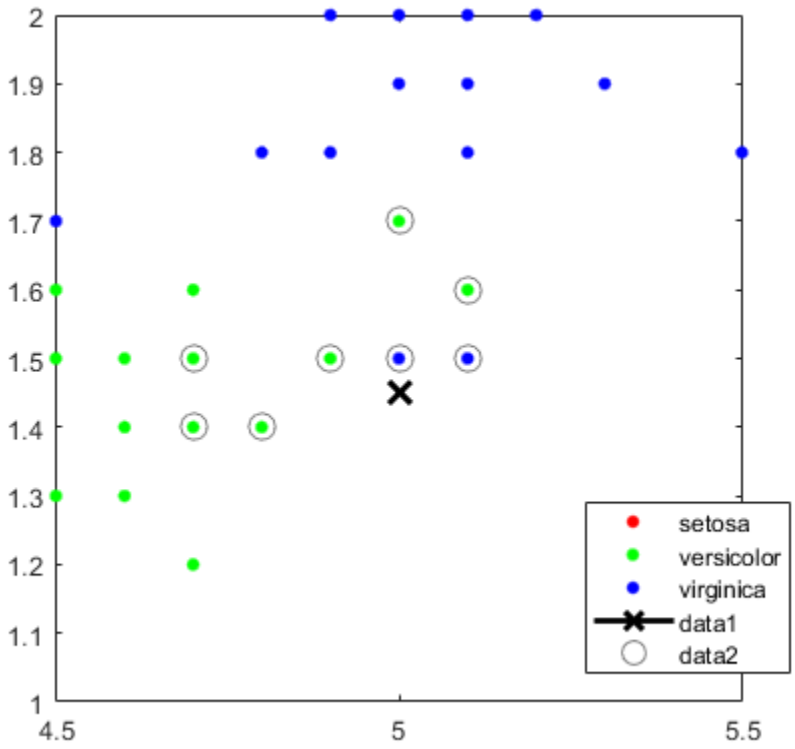
It appears that `knnsearch` has found only the nearest eight neighbors. In fact, this particular dataset contains duplicate values.

```
x(n,:)
```

```
ans = 10×2

    5.0000    1.5000
    4.9000    1.5000
    4.9000    1.5000
    5.1000    1.5000
    5.1000    1.6000
    4.8000    1.4000
    5.0000    1.7000
    4.7000    1.4000
    4.7000    1.4000
    4.7000    1.5000
```

Make the axes equal so the calculated distances correspond to the apparent distances on the plot axis equal and zoom in to see the neighbors better.

```
xlim([4.5 5.5]);
ylim([1 2]);
axis square
```



Find the species of the 10 neighbors.

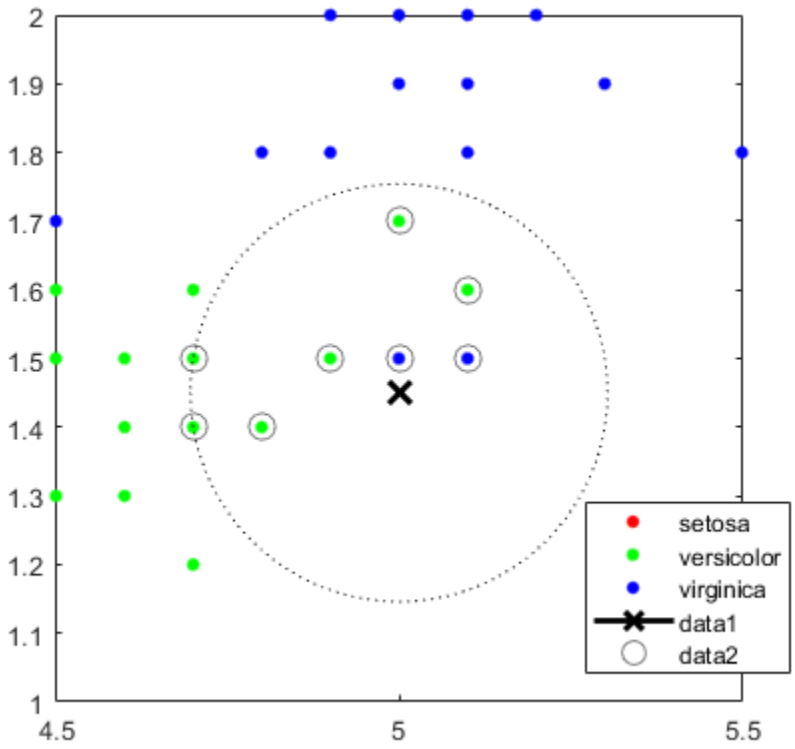
```
tabulate(species(n))
```

```
       Value    Count   Percent
   virginica        2     20.00%
  versicolor        8     80.00%
```

Using a rule based on the majority vote of the 10 nearest neighbors, you can classify this new point as a versicolor.

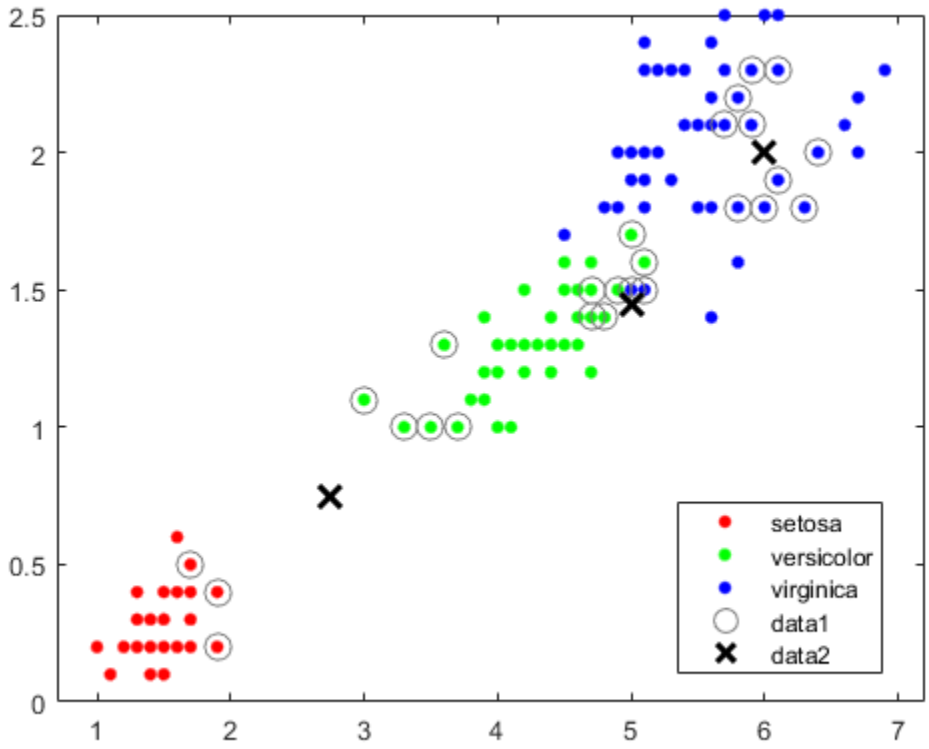
Visually identify the neighbors by drawing a circle around the group of them. Define the center and diameter of a circle, based on the location of the new point.

```
ctr = newpoint - d(end);
diameter = 2*d(end);
% Draw a circle around the 10 nearest neighbors.
h = rectangle('position',[ctr,diameter,diameter],...
    'curvature',[1 1]);
h.LineStyle = ':';
```



Using the same dataset, find the 10 nearest neighbors to three new points.

```
figure
newpoint2 = [5 1.45;6 2;2.75 .75];
gscatter(x(:,1),x(:,2),species)
legend('location','best')
[n2,d2] = knnsearch(Mdl,newpoint2,'k',10);
line(x(n2,1),x(n2,2),'color',[.5 .5 .5],'marker','o',...
   'linestyle','none','markersize',10)
line(newpoint2(:,1),newpoint2(:,2),'marker','x','color','k',...
   'markersize',10,'linewidth',2,'linestyle','none')
```



Find the species of the 10 nearest neighbors for each new point.

```
tabulate(species(n2(1,:)))
```

```
       Value    Count   Percent
   virginica        2     20.00%
  versicolor        8     80.00%
```

```
tabulate(species(n2(2,:)))
```

```
      Value    Count   Percent
  virginica       10    100.00%
```

```
tabulate(species(n2(3,:)))
```

```
       Value    Count   Percent
  versicolor        7     70.00%
      setosa        3     30.00%
```

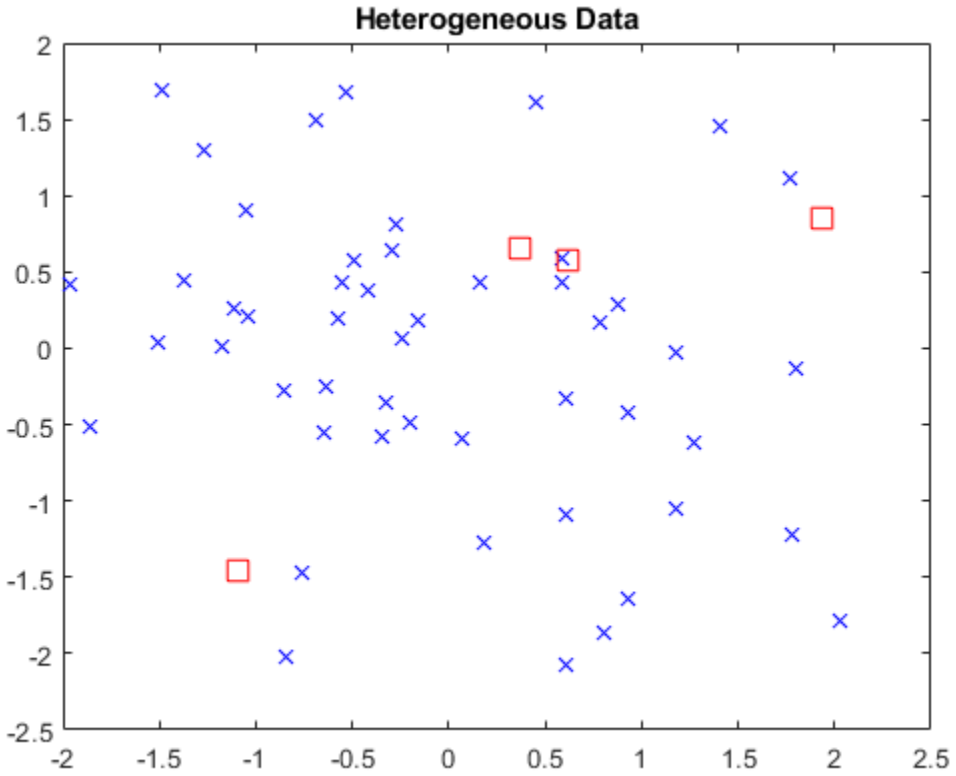
For more examples using `knnsearch` methods and function, see the individual reference pages.

Find Nearest Neighbors Using a Custom Distance Metric

This example shows how to find the indices of the three nearest observations in X to each observation in Y with respect to the chi-square distance. This distance metric is used in correspondence analysis, particularly in ecological applications.

Randomly generate normally distributed data into two matrices. The number of rows can vary, but the number of columns must be equal. This example uses 2-D data for plotting.

```
rng(1) % For reproducibility
X = randn(50,2);
Y = randn(4,2);

h = zeros(3,1);
figure
h(1) = plot(X(:,1),X(:,2),'bx');
hold on
h(2) = plot(Y(:,1),Y(:,2),'rs','MarkerSize',10);
title('Heterogeneous Data')
```



The rows of X and Y correspond to observations, and the columns are, in general, dimensions (for example, predictors).

The chi-square distance between j-dimensional points x and z is

$$\chi(x,z) = \sqrt{\sum_{j=1}^{J} w_j (x_j - z_j)^2},$$

where w_j is the weight associated with dimension j.

Choose weights for each dimension, and specify the chi-square distance function. The distance function must:

- Take as input arguments one row of X, e.g., x, and the matrix Z.
- Compare x to each row of Z.
- Return a vector D of length n_z, where n_z is the number of rows of Z. Each element of D is the distance between the observation corresponding to x and the observations corresponding to each row of Z.

```
w = [0.4; 0.6];
chiSqrDist = @(x,Z)sqrt((bsxfun(@minus,x,Z).^2)*w);
```

This example uses arbitrary weights for illustration.

Find the indices of the three nearest observations in X to each observation in Y.

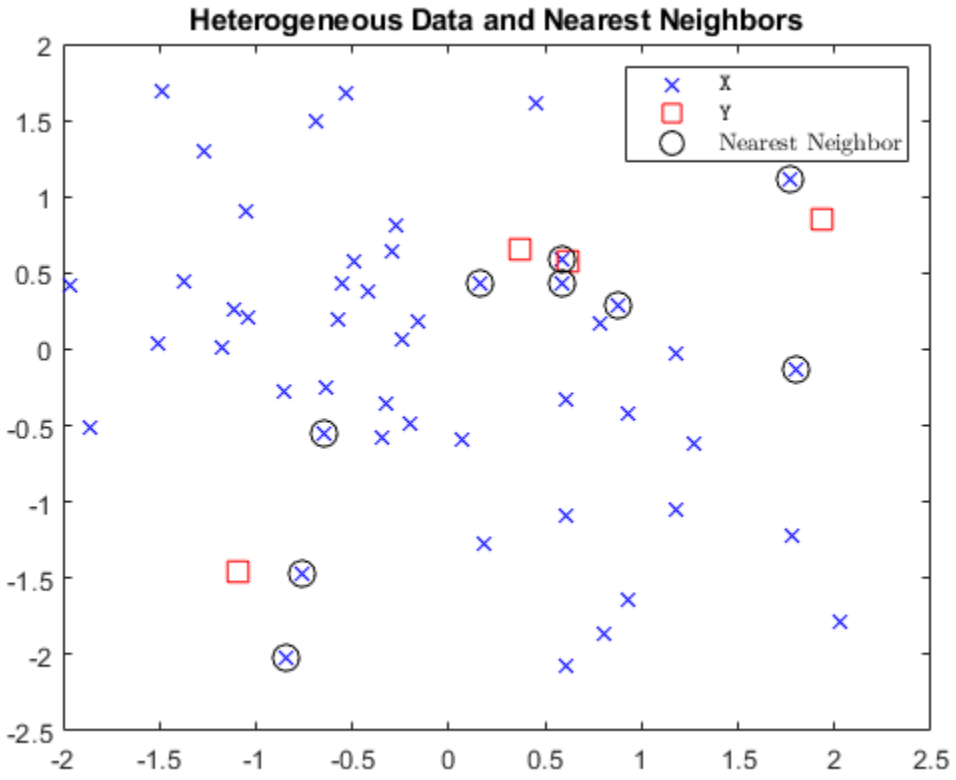
```
k = 3;
[Idx,D] = knnsearch(X,Y,'Distance',chiSqrDist,'k',k);
```

idx and D are 4-by-3 matrices.

- idx(j,1) is the row index of the closest observation in X to observation *j* of Y, and D(j,1) is their distance.
- idx(j,2) is the row index of the next closest observation in X to observation *j* of Y, and D(j,2) is their distance.
- And so on.

Identify the nearest observations in the plot.

```
for j = 1:k
    h(3) = plot(X(Idx(:,j),1),X(Idx(:,j),2),'ko','MarkerSize',10);
end
legend(h,{'\texttt{X}','\texttt{Y}','Nearest Neighbor'},'Interpreter','latex')
title('Heterogeneous Data and Nearest Neighbors')
hold off
```



Several observations of Y share nearest neighbors.

Verify that the chi-square distance metric is equivalent to the Euclidean distance metric, but with an optional scaling parameter.

```
[IdxE,DE] = knnsearch(X,Y,'Distance','seuclidean','k',k, ...
    'Scale',1./(sqrt(w)));
AreDiffIdx = sum(sum(Idx ~= IdxE))
```

```
AreDiffIdx = 0
```

```
AreDiffDist = sum(sum(abs(D - DE) > eps))
```

```
AreDiffDist = 0
```

The indices and distances between the two implementations of three nearest neighbors are practically equivalent.

K-Nearest Neighbor Classification for Supervised Learning

The `ClassificationKNN` classification model lets you:

- “Construct KNN Classifier” on page 18-28
- “Examine Quality of KNN Classifier” on page 18-28
- “Predict Classification Using KNN Classifier” on page 18-29

- “Modify KNN Classifier” on page 18-29

Prepare your data for classification according to the procedure in “Steps in Supervised Learning” on page 18-4. Then, construct the classifier using `fitcknn`.

Construct KNN Classifier

This example shows how to construct a *k*-nearest neighbor classifier for the Fisher iris data.

Load the Fisher iris data.

```
load fisheriris
X = meas;    % Use all data for fitting
Y = species; % Response data
```

Construct the classifier using `fitcknn`.

```
Mdl = fitcknn(X,Y)
```

```
Mdl =
  ClassificationKNN
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
                 Distance: 'euclidean'
             NumNeighbors: 1


  Properties, Methods
```

A default *k*-nearest neighbor classifier uses a single nearest neighbor only. Often, a classifier is more robust with more neighbors than that.

Change the neighborhood size of `Mdl` to 4, meaning that `Mdl` classifies using the four nearest neighbors.

```
Mdl.NumNeighbors = 4;
```

Examine Quality of KNN Classifier

This example shows how to examine the quality of a *k*-nearest neighbor classifier using resubstitution and cross validation.

Construct a KNN classifier for the Fisher iris data as in “Construct KNN Classifier” on page 18-28.

```
load fisheriris
X = meas;
Y = species;
rng(10); % For reproducibility
Mdl = fitcknn(X,Y,'NumNeighbors',4);
```

Examine the resubstitution loss, which, by default, is the fraction of misclassifications from the predictions of `Mdl`. (For nondefault cost, weights, or priors, see `loss`.).

```
rloss = resubLoss(Mdl)
```

```
rloss = 0.0400
```

The classifier predicts incorrectly for 4% of the training data.

Construct a cross-validated classifier from the model.

```
CVMdl = crossval(Mdl);
```

Examine the cross-validation loss, which is the average loss of each cross-validation model when predicting on data that is not used for training.

```
kloss = kfoldLoss(CVMdl)
```

```
kloss = 0.0333
```

The cross-validated classification accuracy resembles the resubstitution accuracy. Therefore, you can expect `Mdl` to misclassify approximately 4% of new data, assuming that the new data has about the same distribution as the training data.

Predict Classification Using KNN Classifier

This example shows how to predict classification for a *k*-nearest neighbor classifier.

Construct a KNN classifier for the Fisher iris data as in “Construct KNN Classifier” on page 18-28.

```
load fisheriris
X = meas;
Y = species;
Mdl = fitcknn(X,Y,'NumNeighbors',4);
```

Predict the classification of an average flower.

```
flwr = mean(X); % an average flower
flwrClass = predict(Mdl,flwr)
```

```
flwrClass = 1x1 cell array
    {'versicolor'}
```

Modify KNN Classifier

This example shows how to modify a *k*-nearest neighbor classifier.

Construct a KNN classifier for the Fisher iris data as in “Construct KNN Classifier” on page 18-28.

```
load fisheriris
X = meas;
Y = species;
Mdl = fitcknn(X,Y,'NumNeighbors',4);
```

Modify the model to use the three nearest neighbors, rather than the default one nearest neighbor.

```
Mdl.NumNeighbors = 3;
```

Compare the resubstitution predictions and cross-validation loss with the new number of neighbors.

```
loss = resubLoss(Mdl)
```

```
loss = 0.0400
```

```
rng(10); % For reproducibility
CVMdl = crossval(Mdl,'KFold',5);
kloss = kfoldLoss(CVMdl)
```

```
kloss = 0.0333
```

In this case, the model with three neighbors has the same cross-validated loss as the model with four neighbors (see “Examine Quality of KNN Classifier” on page 18-28).

Modify the model to use cosine distance instead of the default, and examine the loss. To use cosine distance, you must recreate the model using the exhaustive search method.

```
CMdl = fitcknn(X,Y,'NSMethod','exhaustive','Distance','cosine');
CMdl.NumNeighbors = 3;
closs = resubLoss(CMdl)
```

```
closs = 0.0200
```

The classifier now has lower resubstitution error than before.

Check the quality of a cross-validated version of the new model.

```
CVCMdl = crossval(CMdl);
kcloss = kfoldLoss(CVCMdl)
```

```
kcloss = 0.0200
```

`CVCMdl` has a better cross-validated loss than `CVMdl`. However, in general, improving the resubstitution error does not necessarily produce a model with better test-sample predictions.

See Also

`ClassificationKNN` | `ExhaustiveSearcher` | `KDTreeSearcher` | `fitcknn`

Framework for Ensemble Learning

Using various methods, you can meld results from many weak learners into one high-quality ensemble predictor. These methods closely follow the same syntax, so you can try different methods with minor changes in your commands.

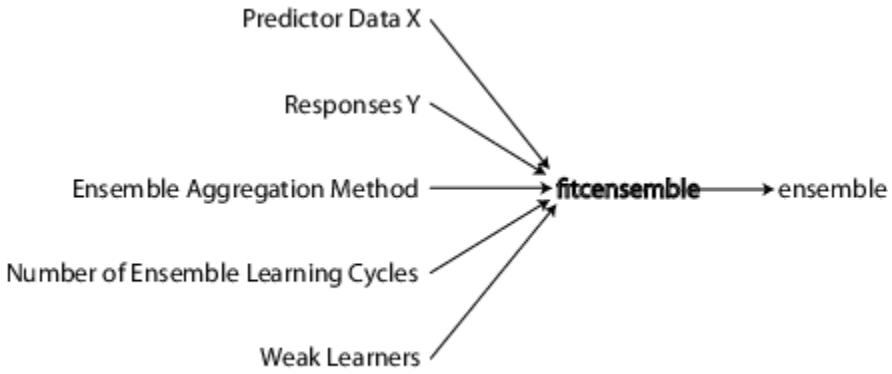
You can create an ensemble for classification by using `fitcensemble` or for regression by using `fitrensemble`.

To train an ensemble for classification using `fitcensemble`, use this syntax.

```
ens = fitcensemble(X,Y,Name,Value)
```

- `X` is the matrix of data. Each row contains one observation, and each column contains one predictor variable.
- `Y` is the vector of responses, with the same number of observations as the rows in `X`.
- `Name,Value` specify additional options using one or more name-value pair arguments. For example, you can specify the ensemble aggregation method with the `'Method'` argument, the number of ensemble learning cycles with the `'NumLearningCycles'` argument, and the type of weak learners with the `'Learners'` argument. For a complete list of name-value pair arguments, see the `fitcensemble` function page.

This figure shows the information you need to create a classification ensemble.



Similarly, you can train an ensemble for regression by using `fitrensemble`, which follows the same syntax as `fitcensemble`. For details on the input arguments and name-value pair arguments, see the `fitrensemble` function page.

For all classification or nonlinear regression problems, follow these steps to create an ensemble:

1. “Prepare the Predictor Data” on page 18-32
2. “Prepare the Response Data” on page 18-32
3. “Choose an Applicable Ensemble Aggregation Method” on page 18-32
4. “Set the Number of Ensemble Members” on page 18-35
5. “Prepare the Weak Learners” on page 18-35
6. “Call fitcensemble or fitrensemble” on page 18-37

Prepare the Predictor Data

All supervised learning methods start with predictor data, usually called X in this documentation. X can be stored in a matrix or a table. Each row of X represents one observation, and each column of X represents one variable or predictor.

Prepare the Response Data

You can use a wide variety of data types for the response data.

- For regression ensembles, Y must be a numeric vector with the same number of elements as the number of rows of X.
- For classification ensembles, Y can be a numeric vector, categorical vector, character array, string array, cell array of character vectors, or logical vector.

  For example, suppose your response data consists of three observations in the following order: `true`, `false`, `true`. You could express Y as:

  - `[1;0;1]` (numeric vector)
  - `categorical({'true','false','true'})` (categorical vector)
  - `[true;false;true]` (logical vector)
  - `['true ';'false';'true ']` (character array, padded with spaces so each row has the same length)
  - `["true","false","true"]` (string array)
  - `{'true','false','true'}` (cell array of character vectors)

  Use whichever data type is most convenient. Because you cannot represent missing values with logical entries, do not use logical entries when you have missing values in Y.

`fitcensemble` and `fitrensemble` ignore missing values in Y when creating an ensemble. This table contains the method of including missing entries.

Data Type	Missing Entry
Numeric vector	`NaN`
Categorical vector	`<undefined>`
Character array	Row of spaces
String array	`<missing>` or `""`
Cell array of character vectors	`''`
Logical vector	(not possible to represent)

Choose an Applicable Ensemble Aggregation Method

To create classification and regression ensembles with `fitcensemble` and `fitrensemble`, respectively, choose appropriate algorithms from this list.

- For classification with two classes:
  - `'AdaBoostM1'`

  - `'LogitBoost'`
  - `'GentleBoost'`
  - `'RobustBoost'` (requires Optimization Toolbox)
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For classification with three or more classes:
  - `'AdaBoostM2'`
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For regression:
  - `'LSBoost'`
  - `'Bag'`

For descriptions of the various algorithms, see “Ensemble Algorithms” on page 18-39.

See “Suggestions for Choosing an Appropriate Ensemble Algorithm” on page 18-34.

This table lists characteristics of the various algorithms. In the table titles:

- **Imbalance** — Good for imbalanced data (one class has many more observations than the other)
- **Stop** — Algorithm self-terminates
- **Sparse** — Requires fewer weak learners than other ensemble algorithms

Algorithm	Regression	Binary Classification	Multiclass Classification	Class Imbalance	Stop	Sparse
`Bag`	×	×	×			
`AdaBoostM1`		×				
`AdaBoostM2`			×			
`LogitBoost`		×				
`GentleBoost`		×				
`RobustBoost`		×				
`LPBoost`		×	×		×	×
`TotalBoost`		×	×		×	×
`RUSBoost`		×	×	×		

Algorithm	Regression	Binary Classification	Multiclass Classification	Class Imbalance	Stop	Sparse
LSBoost	×					
Subspace		×	×			

RobustBoost, LPBoost, and TotalBoost require an Optimization Toolbox license. Try TotalBoost before LPBoost, as TotalBoost can be more robust.

Suggestions for Choosing an Appropriate Ensemble Algorithm

- **Regression** — Your choices are LSBoost or Bag. See "General Characteristics of Ensemble Algorithms" on page 18-34 for the main differences between boosting and bagging.
- **Binary Classification** — Try AdaBoostM1 first, with these modifications:

Data Characteristic	Recommended Algorithm
Many predictors	Subspace
Skewed data (many more observations of one class)	RUSBoost
Label noise (some training data has the wrong class)	RobustBoost
Many observations	Avoid LPBoost and TotalBoost

- **Multiclass Classification** — Try AdaBoostM2 first, with these modifications:

Data Characteristic	Recommended Algorithm
Many predictors	Subspace
Skewed data (many more observations of one class)	RUSBoost
Many observations	Avoid LPBoost and TotalBoost

For details of the algorithms, see "Ensemble Algorithms" on page 18-39.

General Characteristics of Ensemble Algorithms

- Boost algorithms generally use very shallow trees. This construction uses relatively little time or memory. However, for effective predictions, boosted trees might need more ensemble members than bagged trees. Therefore it is not always clear which class of algorithms is superior.
- Bag generally constructs deep trees. This construction is both time consuming and memory-intensive. This also leads to relatively slow predictions.
- Bag can estimate the generalization error without additional cross validation. See oobLoss.
- Except for Subspace, all boosting and bagging algorithms are based on decision tree on page 19-2 learners. Subspace can use either discriminant analysis on page 20-2 or *k*-nearest neighbor on page 18-12 learners.

For details of the characteristics of individual ensemble members, see "Characteristics of Classification Algorithms" on page 18-7.

Set the Number of Ensemble Members

Choosing the size of an ensemble involves balancing speed and accuracy.

- Larger ensembles take longer to train and to generate predictions.
- Some ensemble algorithms can become overtrained (inaccurate) when too large.

To set an appropriate size, consider starting with several dozen to several hundred members in an ensemble, training the ensemble, and then checking the ensemble quality, as in “Test Ensemble Quality” on page 18-66. If it appears that you need more members, add them using the `resume` method (classification) or the `resume` method (regression). Repeat until adding more members does not improve ensemble quality.

Tip For classification, the `LPBoost` and `TotalBoost` algorithms are self-terminating, meaning you do not have to investigate the appropriate ensemble size. Try setting `NumLearningCycles` to `500`. The algorithms usually terminate with fewer members.

Prepare the Weak Learners

Currently the weak learner types are:

- `'Discriminant'` (recommended for `Subspace` ensemble)
- `'KNN'` (only for `Subspace` ensemble)
- `'Tree'` (for any ensemble except `Subspace`)

There are two ways to set the weak learner type in an ensemble.

- To create an ensemble with default weak learner options, specify the value of the `'Learners'` name-value pair argument as the character vector or string scalar of the weak learner name. For example:

  ```
  ens = fitcensemble(X,Y,'Method','Subspace', ...
     'NumLearningCycles',50,'Learners','KNN');
  % or
  ens = fitrensemble(X,Y,'Method','Bag', ...
     'NumLearningCycles',50,'Learners','Tree');
  ```

- To create an ensemble with nondefault weak learner options, create a nondefault weak learner using the appropriate `template` method.

  For example, if you have missing data, and want to use classification trees with surrogate splits for better accuracy:

  ```
  templ = templateTree('Surrogate','all');
  ens = fitcensemble(X,Y,'Method','AdaBoostM2', ...
     'NumLearningCycles',50,'Learners',templ);
  ```

  To grow trees with leaves containing a number of observations that is at least 10% of the sample size:

  ```
  templ = templateTree('MinLeafSize',size(X,1)/10);
  ens = fitcensemble(X,Y,'Method','AdaBoostM2', ...
     'NumLearningCycles',50,'Learners',templ);
  ```

Alternatively, choose the maximal number of splits per tree:

```
templ = templateTree('MaxNumSplits',4);
ens = fitcensemble(X,Y,'Method','AdaBoostM2', ...
   'NumLearningCycles',50,'Learners',templ);
```

You can also use nondefault weak learners in `fitrensemble`.

While you can give `fitcensemble` and `fitrensemble` a cell array of learner templates, the most common usage is to give just one weak learner template.

For examples using a template, see “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84 and “Surrogate Splits” on page 18-91.

Decision trees can handle `NaN` values in `X`. Such values are called “missing”. If you have some missing values in a row of `X`, a decision tree finds optimal splits using nonmissing values only. If an entire row consists of `NaN`, `fitcensemble` and `fitrensemble` ignore that row. If you have data with a large fraction of missing values in `X`, use surrogate decision splits. For examples of surrogate splits, see “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84 and “Surrogate Splits” on page 18-91.

Common Settings for Tree Weak Learners

- The depth of a weak learner tree makes a difference for training time, memory usage, and predictive accuracy. You control the depth these parameters:
  - `MaxNumSplits` — The maximal number of branch node splits is `MaxNumSplits` per tree. Set large values of `MaxNumSplits` to get deep trees. The default for bagging is `size(X,1) - 1`. The default for boosting is `1`.
  - `MinLeafSize` — Each leaf has at least `MinLeafSize` observations. Set small values of `MinLeafSize` to get deep trees. The default for classification is `1` and `5` for regression.
  - `MinParentSize` — Each branch node in the tree has at least `MinParentSize` observations. Set small values of `MinParentSize` to get deep trees. The default for classification is `2` and `10` for regression.

  If you supply both `MinParentSize` and `MinLeafSize`, the learner uses the setting that gives larger leaves (shallower trees):

  ```
  MinParent = max(MinParent,2*MinLeaf)
  ```

  If you additionally supply `MaxNumSplits`, then the software splits a tree until one of the three splitting criteria is satisfied.
- `Surrogate` — Grow decision trees with surrogate splits when `Surrogate` is `'on'`. Use surrogate splits when your data has missing values.

  Note Surrogate splits cause slower training and use more memory.
- `PredictorSelection` — `fitcensemble`, `fitrensemble`, and `TreeBagger` grow trees using the standard CART algorithm [11] by default. If the predictor variables are heterogeneous or there are predictors having many levels and other having few levels, then standard CART tends to select predictors having many levels as split predictors. For split-predictor selection that is robust to the number of levels that the predictors have, consider specifying `'curvature'` or `'interaction-curvature'`. These specifications conduct chi-square tests of association between each predictor

and the response or each pair of predictors and the response, respectively. The predictor that yields the minimal *p*-value is the split predictor for a particular node. For more details, see “Choose Split Predictor Selection Technique” on page 19-14.

Note When boosting decision trees, selecting split predictors using the curvature or interaction tests is not recommended.

Call fitcensemble or fitrensemble

The syntaxes for `fitcensemble` and `fitrensemble` are identical. For `fitrensemble`, the syntax is:

```
ens = fitrensemble(X,Y,Name,Value)
```

- `X` is the matrix of data. Each row contains one observation, and each column contains one predictor variable.
- `Y` is the responses, with the same number of observations as rows in `X`.
- `Name,Value` specify additional options using one or more name-value pair arguments. For example, you can specify the ensemble aggregation method with the `'Method'` argument, the number of ensemble learning cycles with the `'NumLearningCycles'` argument, and the type of weak learners with the `'Learners'` argument. For a complete list of name-value pair arguments, see the `fitrensemble` function page.

The result of `fitrensemble` and `fitcensemble` is an ensemble object, suitable for making predictions on new data. For a basic example of creating a regression ensemble, see “Train Regression Ensemble” on page 18-57. For a basic example of creating a classification ensemble, see “Train Classification Ensemble” on page 18-54.

Where to Set Name-Value Pairs

There are several name-value pairs you can pass to `fitcensemble` or `fitrensemble`, and several that apply to the weak learners (`templateDiscriminant`, `templateKNN`, and `templateTree`). To determine which name-value pair argument is appropriate, the ensemble or the weak learner:

- Use template name-value pairs to control the characteristics of the weak learners.
- Use `fitcensemble` or `fitrensemble` name-value pair arguments to control the ensemble as a whole, either for algorithms or for structure.

For example, for an ensemble of boosted classification trees with each tree deeper than the default, set the `templateTree` name-value pair arguments `MinLeafSize` and `MinParentSize` to smaller values than the defaults. Or, `MaxNumSplits` to a larger value than the defaults. The trees are then leafier (deeper).

To name the predictors in a classification ensemble (part of the structure of the ensemble), use the `PredictorNames` name-value pair in `fitcensemble`.

See Also

`fitcensemble` | `fitrensemble` | `oobLoss` | `resume` | `resume` | `templateDiscriminant` | `templateKNN` | `templateTree`

Related Examples

- “Train Classification Ensemble” on page 18-54
- “Train Regression Ensemble” on page 18-57
- “Ensemble Algorithms” on page 18-39
- “Decision Trees” on page 19-2
- “Choose Split Predictor Selection Technique” on page 19-14

Ensemble Algorithms

In this section...

This topic provides descriptions of ensemble learning algorithms supported by Statistics and Machine Learning Toolbox, including bagging, random space, and various boosting algorithms. You can specify the algorithm by using the `'Method'` name-value pair argument of `fitcensemble`, `fitrensemble`, or `templateEnsemble`. Use `fitcensemble` or `fitrensemble` to create an ensemble of learners for classification or regression, respectively. Use `templateEnsemble` to create an ensemble learner template, and pass the template to `fitcecoc` to specify ensemble binary learners for ECOC multiclass learning.

For bootstrap aggregation (bagging) and random forest, you can use `TreeBagger` as well.

Value of 'Method'	Algorithm	Supported Problems	Examples
'Bag'	"Bootstrap Aggregation (Bagging) and Random Forest" on page 18-42 ([1], [2], [3])	Binary and multiclass classification, regression	• "Select Predictors for Random Forests" on page 18-60 • "Test Ensemble Quality" on page 18-66 • "Surrogate Splits" on page 18-91 • "Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger" on page 18-124 • "Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger" on page 18-113 • "Tune Random Forest Using Quantile Error and Bayesian Optimization" on page 18-145 • "Detect Outliers Using Quantile Regression" on page 18-137 • "Conditional Quantile Estimation Using Kernel Smoothing" on page 18-142
'Subspace'	"Random Subspace" on page 18-45 ([9])	Binary and multiclass classification	"Random Subspace Classification" on page 18-104

Value of 'Method'	Algorithm	Supported Problems	Examples
'AdaBoostM1'	"Adaptive Boosting for Binary Classification" on page 18-46 ([5], [6], [7], [11])	Binary classification	• "Create an Ensemble Template for ECOC Multiclass Learning" on page 33-6101 • "Conduct Cost-Sensitive Comparison of Two Classification Models" on page 33-898
'AdaBoostM2'	"Adaptive Boosting for Multiclass Classification" on page 18-47 ([5])	Multiclass classification	"Predict Class Labels Using Classification Ensemble" on page 33-4830
'GentleBoost'	"Gentle Adaptive Boosting" on page 18-48 ([7])	Binary classification	• "Speed Up Training ECOC Classifiers Using Binning and Parallel Computing" on page 33-6102 • "Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles" on page 18-84
'LogitBoost'	"Adaptive Logistic Regression" on page 18-48 ([7])	Binary classification	• "Train Classification Ensemble" on page 18-54 • "Speed Up Training by Binning Numeric Predictor Values" on page 33-1651
'LPBoost'	"Linear Programming Boosting" on page 18-49 ([13])	Binary and multiclass classification	"LPBoost and TotalBoost for Small Ensembles" on page 18-96

Value of 'Method'	Algorithm	Supported Problems	Examples
'LSBoost'	"Least-Squares Boosting" on page 18-50 ([2], [8])	Regression	• "Ensemble Regularization" on page 18-70 • "Optimize a Boosted Regression Ensemble" on page 10-63 • "Train Regression Ensemble" on page 18-57
'RobustBoost'	"Robust Boosting" on page 18-50 ([4])	Binary classification	"Tune RobustBoost" on page 18-101
'RUSBoost'	"Random Undersampling Boosting" on page 18-51 ([12])	Binary and multiclass classification	"Classification with Imbalanced Data" on page 18-79
'TotalBoost'	"Totally Corrective Boosting" on page 18-51 ([13])	Binary and multiclass classification	"LPBoost and TotalBoost for Small Ensembles" on page 18-96

To learn about how to choose an appropriate algorithm, see "Choose an Applicable Ensemble Aggregation Method" on page 18-32.

Note that usage of some algorithms, such as `LPBoost`, `TotalBoost`, and `RobustBoost`, requires Optimization Toolbox.

Bootstrap Aggregation (Bagging) and Random Forest

Statistics and Machine Learning Toolbox offers three objects for bagging and random forest:

- `ClassificationBaggedEnsemble` created by `fitcensemble` for classification
- `RegressionBaggedEnsemble` created by `fitrensemble` for regression
- `TreeBagger` created by `TreeBagger` for classification and regression

For details about the differences between `TreeBagger` and bagged ensembles (`ClassificationBaggedEnsemble` and `RegressionBaggedEnsemble`), see "Comparison of TreeBagger and Bagged Ensembles" on page 18-44.

Bootstrap aggregation (bagging) is a type of ensemble learning. To bag a weak learner such as a decision tree on a data set, generate many bootstrap replicas of the data set and grow decision trees on the replicas. Obtain each bootstrap replica by randomly selecting `N` out of `N` observations with replacement, where `N` is the data set size. In addition, every tree in the ensemble can randomly select predictors for each decision split, a technique called random forest [2] known to improve the accuracy of bagged trees. By default, the number of predictors to select at random for each split is equal to the square root of the number of predictors for classification, and one third of the number of predictors for regression. After training a model, you can find the predicted response of a trained ensemble for new data by using the `predict` function. `predict` takes an average over predictions from individual trees.

By default, the minimum number of observations per leaf for bagged trees is set to `1` for classification and `5` for regression. Trees grown with the default leaf size are usually very deep. These settings are close to optimal for the predictive power of an ensemble. Often you can grow trees with larger leaves without losing predictive power. Doing so reduces training and prediction time, as well as memory usage for the trained ensemble. You can control the minimum number of observations per leaf by using the `'MinLeafSize'` name-value pair argument of `templateTree` or `TreeBagger`. Note that you use the `templateTree` function to specify the options of tree learners when you create a bagged ensemble by using `fitcensemble` or `fitrensemble`.

Several features of bagged decision trees make them a unique algorithm. Drawing `N` out of `N` observations with replacement omits of observations, on average 37%, for each decision tree. These omitted observations are called “out-of-bag” observations. `TreeBagger` and bagged ensembles (`ClassificationBaggedEnsemble` and `RegressionBaggedEnsemble`) have properties and object functions, whose names start with `oob`, that use out-of-bag observations.

- Use the `oobPredict` function to estimate predictive power and feature importance. For each observation, `oobPredict` estimates the out-of-bag prediction by averaging predictions from all trees in the ensemble for which the observation is out of bag.
- Estimate the average out-of-bag error by using `oobError` (for `TreeBagger`) or `oobLoss` (for bagged ensembles). These functions compare the out-of-bag predicted responses against the observed responses for all observations used for training. The out-of-bag average is an unbiased estimator of the true ensemble error.
- Obtain out-of-bag estimates of feature importance by using the `OOBPermutedPredictorDeltaError` property (for `TreeBagger`) or `oobPermutedPredictorImportance` property (for bagged ensembles). The software randomly permutes out-of-bag data across one variable or column at a time and estimates the increase in the out-of-bag error due to this permutation. The larger the increase, the more important the feature. Therefore, you do not need to supply test data for bagged ensembles because you can obtain reliable estimates of predictive power and feature importance in the process of training.

`TreeBagger` also offers the proximity matrix in the `Proximity` property. Every time two observations land on the same leaf of a tree, their proximity increases by 1. For normalization, sum these proximities over all trees in the ensemble and divide by the number of trees. The resulting matrix is symmetric with diagonal elements equal to 1 and off-diagonal elements ranging from 0 to 1. You can use this matrix to find outlier observations and discover clusters in the data through multidimensional scaling.

For examples using bagging, see:

- “Select Predictors for Random Forests” on page 18-60
- “Test Ensemble Quality” on page 18-66
- “Surrogate Splits” on page 18-91
- “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
- “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
- “Tune Random Forest Using Quantile Error and Bayesian Optimization” on page 18-145
- “Detect Outliers Using Quantile Regression” on page 18-137
- “Conditional Quantile Estimation Using Kernel Smoothing” on page 18-142

Comparison of TreeBagger and Bagged Ensembles

`TreeBagger` and bagged ensembles (`ClassificationBaggedEnsemble` and `RegressionBaggedEnsemble`) share most functionalities, but not all. Additionally, some functionalities have different names.

`TreeBagger` features not in bagged ensembles

Feature	`TreeBagger` Property	`TreeBagger` Method
Computation of proximity matrix	`Proximity`	`fillprox`, `mdsprox` When you estimate the proximity matrix and outliers of a `TreeBagger` model using `fillprox`, MATLAB must fit an *n*-by-*n* matrix in memory, where *n* is the number of observations. Therefore, if *n* is moderate to large, avoid estimating the proximity matrix and outliers.
Computation of outliers	`OutlierMeasure`	N/A
Out-of-bag estimates of predictor importance using classification margins	`OOBPermutedPredictorDeltaMeanMargin` and `OOBPermutedPredictorCountRaiseMargin`	N/A
Merging two ensembles trained separately	N/A	`append`
Quantile regression	N/A	`quantilePredict`, `quantileError`, `oobQuantilePredict`, `oobQuantileError`
Tall array support for creating ensemble	N/A	For details, see “Tall Arrays” on page 33-6274.

Bagged ensemble features not in `TreeBagger`

Feature	Description
Hyperparameter optimization	Use the `'OptimizeHyperparameters'` name-value pair argument.
Binning numeric predictors to speed up training	Use the `'NumBins'` name-value pair argument.
Code generation for `predict`	After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder™. For details, see “Introduction to Code Generation” on page 32-2.

Different names for `TreeBagger` and bagged ensembles

Feature	`TreeBagger`	Bagged Ensembles
Split criterion contributions for each predictor	`DeltaCriterionDecisionSplit` property	First output of `predictorImportance` (classification) or `predictorImportance` (regression)
Predictor associations	`SurrogateAssociation` property	Second output of `predictorImportance` (classification) or `predictorImportance` (regression)
Out-of-bag estimates of predictor importance	`OOBPermutedPredictorDeltaError` property	Output of `oobPermutedPredictorImportance` (classification) or `oobPermutedPredictorImportance` (regression)
Error (misclassification probability or mean-squared error)	`error` and `oobError` methods	`loss` and `oobLoss` methods (classification) or `loss` and `oobLoss` methods (regression)
Training additional trees and adding them to ensemble	`growTrees` method	`resume` method (classification) or `resume` method (regression)
Mean classification margin per tree	`meanMargin` and `oobMeanMargin` methods	`edge` and `oobEdge` methods (classification)

In addition, two important differences exist when you train a model and predict responses:

- If you pass a misclassification cost matrix to `TreeBagger`, it passes the matrix along to the trees. If you pass a misclassification cost matrix to `fitcensemble`, it uses the matrix to adjust the class prior probabilities. `fitcensemble` then passes the adjusted prior probabilities and the default cost matrix to the trees. The default cost matrix is `ones(K)–eye(K)` for `K` classes.
- Unlike the `loss` and `edge` methods in `ClassificationBaggedEnsemble`, the `TreeBagger` `error` and `meanMargin` methods do not normalize input observation weights of the prior probabilities in the respective class.

Random Subspace

Use random subspace ensembles (`Subspace`) to improve the accuracy of discriminant analysis (`ClassificationDiscriminant`) or *k*-nearest neighbor (`ClassificationKNN`) classifiers. `Subspace` ensembles also have the advantage of using less memory than ensembles with all predictors, and can handle missing values (`NaN`s).

The basic random subspace algorithm uses these parameters.

- *m* is the number of dimensions (variables) to sample in each learner. Set *m* using the `NPredToSample` name-value pair.
- *d* is the number of dimensions in the data, which is the number of columns (predictors) in the data matrix `X`.

- n is the number of learners in the ensemble. Set n using the `NLearn` input.

The basic random subspace algorithm performs the following steps:

1. Choose without replacement a random set of m predictors from the d possible values.
2. Train a weak learner using just the m chosen predictors.
3. Repeat steps 1 and 2 until there are n weak learners.
4. Predict by taking an average of the `score` prediction of the weak learners, and classify the category with the highest average `score`.

You can choose to create a weak learner for every possible set of m predictors from the d dimensions. To do so, set n, the number of learners, to `'AllPredictorCombinations'`. In this case, there are `nchoosek(size(X,2),NPredToSample)` weak learners in the ensemble.

`fitcensemble` downweights predictors after choosing them for a learner, so subsequent learners have a lower chance of using a predictor that was previously used. This weighting tends to make predictors more evenly distributed among learners than in uniform weighting.

For examples using `Subspace`, see “Random Subspace Classification” on page 18-104.

Boosting Algorithms

Adaptive Boosting for Binary Classification

Adaptive boosting named `AdaBoostM1` is a very popular boosting algorithm for binary classification. The algorithm trains learners sequentially. For every learner with index t, `AdaBoostM1` computes the weighted classification error

$$\varepsilon_t = \sum_{n=1}^{N} d_n^{(t)} \mathbb{I}(y_n \neq h_t(x_n)),$$

where

- x_n is a vector of predictor values for observation n.
- y_n is the true class label.
- h_t is the prediction of learner (hypothesis) with index t.
- $\mathbb{I}$ is the indicator function.
- $d_n^{(t)}$ is the weight of observation n at step t.

`AdaBoostM1` then increases weights for observations misclassified by learner t and reduces weights for observations correctly classified by learner t. The next learner $t + 1$ is then trained on the data with updated weights $d_n^{(t+1)}$.

After training finishes, `AdaBoostM1` computes prediction for new data using

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x),$$

where

$$\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$$

are weights of the weak hypotheses in the ensemble.

Training by `AdaBoostM1` can be viewed as stagewise minimization of the exponential loss

$$\sum_{n=1}^{N} w_n \exp(-y_n f(x_n)),$$

where

- $y_n \in \{-1,+1\}$ is the true class label.
- w_n are observation weights normalized to add up to 1.
- $f(x_n) \in (-\infty,+\infty)$ is the predicted classification score.

The observation weights w_n are the original observation weights you passed to `fitcensemble`.

The second output from the `predict` method of an `AdaBoostM1` classification ensemble is an N-by-2 matrix of classification scores for the two classes and N observations. The second column in this matrix is always equal to minus the first column. The `predict` method returns two scores to be consistent with multiclass models, though this is redundant because the second column is always the negative of the first.

Most often `AdaBoostM1` is used with decision stumps (default) or shallow trees. If boosted stumps give poor performance, try setting the minimal parent node size to one quarter of the training data.

By default, the learning rate for boosting algorithms is `1`. If you set the learning rate to a lower number, the ensemble learns at a slower rate, but can converge to a better solution. `0.1` is a popular choice for the learning rate. Learning at a rate less than `1` is often called “shrinkage”.

For examples using `AdaBoostM1`, see “Conduct Cost-Sensitive Comparison of Two Classification Models” on page 33-898 and “Create an Ensemble Template for ECOC Multiclass Learning” on page 33-6101.

Adaptive Boosting for Multiclass Classification

Adaptive boosting named `AdaBoostM2` is an extension of `AdaBoostM1` for multiple classes. Instead of weighted classification error, `AdaBoostM2` uses weighted pseudo-loss for N observations and K classes

$$\varepsilon_t = \frac{1}{2}\sum_{n=1}^{N}\sum_{k \neq y_n} d_{n,k}^{(t)}(1 - h_t(x_n, y_n) + h_t(x_n, k)),$$

where

- $h_t(x_n,k)$ is the confidence of prediction by learner at step t into class k ranging from 0 (not at all confident) to 1 (highly confident).
- $d_{n,k}^{(t)}$ are observation weights at step t for class k.
- y_n is the true class label taking one of the K values.
- The second sum is over all classes other than the true class y_n.

Interpreting the pseudo-loss is harder than classification error, but the idea is the same. Pseudo-loss can be used as a measure of the classification accuracy from any learner in an ensemble. Pseudo-loss typically exhibits the same behavior as a weighted classification error for `AdaBoostM1`: the first few learners in a boosted ensemble give low pseudo-loss values. After the first few training steps, the ensemble begins to learn at a slower pace, and the pseudo-loss value approaches 0.5 from below.

For an example using `AdaBoostM2`, see “Predict Class Labels Using Classification Ensemble” on page 33-4830.

Gentle Adaptive Boosting

Gentle adaptive boosting (`GentleBoost`, also known as Gentle AdaBoost) combines features of `AdaBoostM1` and `LogitBoost`. Like `AdaBoostM1`, `GentleBoost` minimizes the exponential loss. But its numeric optimization is set up differently. Like `LogitBoost`, every weak learner fits a regression model to response values $y_n \in \{-1,+1\}$.

`fitcensemble` computes and stores the mean-squared error in the `FitInfo` property of the ensemble object. The mean-squared error is

$$\sum_{n=1}^{N} d_n^{(t)} (\tilde{y}_n - h_t(x_n))^2,$$

where

- $d_n^{(t)}$ are observation weights at step t (the weights add up to 1).
- $h_t(x_n)$ are predictions of the regression model h_t fitted to response values y_n.

As the strength of individual learners weakens, the weighted mean-squared error approaches 1.

For examples using `GentleBoost`, see “Speed Up Training ECOC Classifiers Using Binning and Parallel Computing” on page 33-6102 and “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84.

Adaptive Logistic Regression

Adaptive logistic regression (`LogitBoost`) is another popular algorithm for binary classification. `LogitBoost` works similarly to `AdaBoostM1`, except it minimizes binomial deviance

$$\sum_{n=1}^{N} w_n \log(1 + \exp(-2 y_n f(x_n))),$$

where

- $y_n \in \{-1,+1\}$ is the true class label.
- w_n are observation weights normalized to add up to 1.
- $f(x_n) \in (-\infty,+\infty)$ is the predicted classification score.

Binomial deviance assigns less weight to badly misclassified observations (observations with large negative values of $y_n f(x_n)$). `LogitBoost` can give better average accuracy than `AdaBoostM1` for data with poorly separable classes.

Learner t in a `LogitBoost` ensemble fits a regression model to response values

$$\tilde{y}_n = \frac{y^*_n - p_t(x_n)}{p_t(x_n)(1 - p_t(x_n))},$$

where

- $y^*_n \in \{0,+1\}$ are relabeled classes (0 instead of –1).
- $p_t(x_n)$ is the current ensemble estimate of the probability for observation x_n to be of class 1.

`fitcensemble` computes and stores the mean-squared error in the `FitInfo` property of the ensemble object. The mean-squared error is

$$\sum_{n=1}^{N} d_n^{(t)} (\tilde{y}_n - h_t(x_n))^2,$$

where

- $d_n^{(t)}$ are observation weights at step t (the weights add up to 1).
- $h_t(x_n)$ are predictions of the regression model h_t fitted to response values $\tilde{y}_n$.

Values y_n can range from –∞ to +∞, so the mean-squared error does not have well-defined bounds.

For examples using `LogitBoost`, see “Train Classification Ensemble” on page 18-54 and “Speed Up Training by Binning Numeric Predictor Values” on page 33-1651.

Linear Programming Boosting

Linear programming boosting (`LPBoost`), like `TotalBoost`, performs multiclass classification by attempting to maximize the minimal margin in the training set. This attempt uses optimization algorithms, namely linear programming for `LPBoost`. So you need an Optimization Toolbox license to use `LPBoost` or `TotalBoost`.

The margin of a classification is the difference between the predicted soft classification score for the true class, and the largest score for the false classes. For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node. For more information, see “More About” on page 33-3921 in `margin`.

Why maximize the minimal margin? For one thing, the generalization error (the error on new data) is the probability of obtaining a negative margin. Schapire and Singer [10] establish this inequality on the probability of obtaining a negative margin:

$$P_{\text{test}}(m \le 0) \le P_{\text{train}}(m \le \theta) + O\left(\frac{1}{\sqrt{N}}\sqrt{\frac{V\log^2(N/V)}{\theta^2} + \log(1/\delta)}\right).$$

Here m is the margin, θ is any positive number, V is the Vapnik-Chervonenkis dimension of the classifier space, N is the size of the training set, and δ is a small positive number. The inequality holds with probability $1–\delta$ over many i.i.d. training and test sets. This inequality says: To obtain a low generalization error, minimize the number of observations below margin θ in the training set.

`LPBoost` iteratively maximizes the minimal margin through a sequence of linear programming problems. Equivalently, by duality, `LPBoost` minimizes the maximal edge, where edge is the weighted

mean margin (see “More About” on page 33-1323). At each iteration, there are more constraints in the problem. So, for large problems, the optimization problem becomes increasingly constrained, and slow to solve.

`LPBoost` typically creates ensembles with many learners having weights that are orders of magnitude smaller than those of other learners. Therefore, to better enable you to remove the unimportant ensemble members, the `compact` method reorders the members of an `LPBoost` ensemble from largest weight to smallest. Therefore, you can easily remove the least important members of the ensemble using the `removeLearners` method.

For an example using `LPBoost`, see “LPBoost and TotalBoost for Small Ensembles” on page 18-96.

Least-Squares Boosting

Least-squares boosting (`LSBoost`) fits regression ensembles. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously. The ensemble fits to minimize mean-squared error.

You can use `LSBoost` with shrinkage by passing in the `LearnRate` parameter. By default this parameter is set to `1`, and the ensemble learns at the maximal speed. If you set `LearnRate` to a value from `0` to `1`, the ensemble fits every new learner to $y_n - \eta f(x_n)$, where

- y_n is the observed response.
- $f(x_n)$ is the aggregated prediction from all weak learners grown so far for observation x_n.
- η is the learning rate.

For examples using `LSBoost`, see “Train Regression Ensemble” on page 18-57, “Optimize a Boosted Regression Ensemble” on page 10-63, and “Ensemble Regularization” on page 18-70.

Robust Boosting

Boosting algorithms such as `AdaBoostM1` and `LogitBoost` increase weights for misclassified observations at every boosting step. These weights can become very large. If this happens, the boosting algorithm sometimes concentrates on a few misclassified observations and neglects the majority of training data. Consequently the average classification accuracy suffers. In this situation, you can try using robust boosting (`RobustBoost`). This algorithm does not assign almost the entire data weight to badly misclassified observations. It can produce better average classification accuracy. You need an Optimization Toolbox license to use `RobustBoost`.

Unlike `AdaBoostM1` and `LogitBoost`, `RobustBoost` does not minimize a specific loss function. Instead, it maximizes the number of observations with the classification margin above a certain threshold.

`RobustBoost` trains based on time evolution. The algorithm starts at $t = 0$. At every step, `RobustBoost` solves an optimization problem to find a positive step in time Δt and a corresponding positive change in the average margin for training data Δm. `RobustBoost` stops training and exits if at least one of these three conditions is true:

- Time t reaches 1.
- `RobustBoost` cannot find a solution to the optimization problem with positive updates Δt and Δm.
- `RobustBoost` grows as many learners as you requested.

Results from `RobustBoost` can be usable for any termination condition. Estimate the classification accuracy by cross validation or by using an independent test set.

To get better classification accuracy from `RobustBoost`, you can adjust three parameters in `fitcensemble`: `RobustErrorGoal`, `RobustMaxMargin`, and `RobustMarginSigma`. Start by varying values for `RobustErrorGoal` from 0 to 1. The maximal allowed value for `RobustErrorGoal` depends on the two other parameters. If you pass a value that is too high, `fitcensemble` produces an error message showing the allowed range for `RobustErrorGoal`.

For an example using `RobustBoost`, see “Tune RobustBoost” on page 18-101.

Random Undersampling Boosting

Random undersampling boosting (`RUSBoost`) is especially effective at classifying imbalanced data, meaning some class in the training data has many fewer members than another. RUS stands for Random Under Sampling. The algorithm takes N, the number of members in the class with the fewest members in the training data, as the basic unit for sampling. Classes with more members are under sampled by taking only N observations of every class. In other words, if there are K classes, then, for each weak learner in the ensemble, `RUSBoost` takes a subset of the data with N observations from each of the K classes. The boosting procedure follows the procedure in “Adaptive Boosting for Multiclass Classification” on page 18-47 for reweighting and constructing the ensemble.

When you construct a `RUSBoost` ensemble, there is an optional name-value pair called `RatioToSmallest`. Give a vector of K values, each value representing the multiple of N to sample for the associated class. For example, if the smallest class has $N = 100$ members, then `RatioToSmallest = [2,3,4]` means each weak learner has 200 members in class 1, 300 in class 2, and 400 in class 3. If `RatioToSmallest` leads to a value that is larger than the number of members in a particular class, then `RUSBoost` samples the members with replacement. Otherwise, `RUSBoost` samples the members without replacement.

For an example using `RUSBoost`, see “Classification with Imbalanced Data” on page 18-79.

Totally Corrective Boosting

Totally corrective boosting (`TotalBoost`), like linear programming boost (`LPBoost`), performs multiclass classification by attempting to maximize the minimal margin in the training set. This attempt uses optimization algorithms, namely quadratic programming for `TotalBoost`. So you need an Optimization Toolbox license to use `LPBoost` or `TotalBoost`.

The margin of a classification is the difference between the predicted soft classification score for the true class, and the largest score for the false classes. For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node. For more information, see “More About” on page 33-3921 in `margin`.

Why maximize the minimal margin? For one thing, the generalization error (the error on new data) is the probability of obtaining a negative margin. Schapire and Singer [10] establish this inequality on the probability of obtaining a negative margin:

$$P_{\text{test}}(m \le 0) \le P_{\text{train}}(m \le \theta) + O\left(\frac{1}{\sqrt{N}}\sqrt{\frac{V\log^2(N/V)}{\theta^2} + \log(1/\delta)}\right).$$

Here m is the margin, θ is any positive number, V is the Vapnik-Chervonenkis dimension of the classifier space, N is the size of the training set, and δ is a small positive number. The inequality holds with probability $1–\delta$ over many i.i.d. training and test sets. This inequality says: To obtain a low generalization error, minimize the number of observations below margin θ in the training set.

`TotalBoost` minimizes a proxy of the Kullback-Leibler divergence between the current weight distribution and the initial weight distribution, subject to the constraint that the edge (the weighted margin) is below a certain value. The proxy is a quadratic expansion of the divergence:

$$D(W, W_0) = \sum_{n=1}^{N} \log\frac{W(n)}{W_0(n)} \approx \sum_{n=1}^{N} \left(1 + \frac{W(n)}{W_0(n)}\right)\Delta + \frac{1}{2W(n)}\Delta^2,$$

where Δ is the difference between $W(n)$, the weights at the current and next iteration, and W_0, the initial weight distribution, which is uniform. This optimization formulation keeps weights from becoming zero. At each iteration, there are more constraints in the problem. So, for large problems, the optimization problem becomes increasingly constrained, and slow to solve.

`TotalBoost` typically creates ensembles with many learners having weights that are orders of magnitude smaller than those of other learners. Therefore, to better enable you to remove the unimportant ensemble members, the `compact` method reorders the members of a `TotalBoost` ensemble from largest weight to smallest. Therefore you can easily remove the least important members of the ensemble using the `removeLearners` method.

For an example using `TotalBoost`, see “LPBoost and TotalBoost for Small Ensembles” on page 18-96.

References

[1] Breiman, L. "Bagging Predictors." *Machine Learning* 26, 1996, pp. 123–140.

[2] Breiman, L. "Random Forests." *Machine Learning* 45, 2001, pp. 5–32.

[3] Breiman, L. `https://www.stat.berkeley.edu/~breiman/RandomForests/`

[4] Freund, Y. "A more robust boosting algorithm." arXiv:0905.2138v1, 2009.

[5] Freund, Y. and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *J. of Computer and System Sciences,* Vol. 55, 1997, pp. 119–139.

[6] Friedman, J. "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, Vol. 29, No. 5, 2001, pp. 1189–1232.

[7] Friedman, J., T. Hastie, and R. Tibshirani. "Additive logistic regression: A statistical view of boosting." *Annals of Statistics*, Vol. 28, No. 2, 2000, pp. 337–407.

[8] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. New York: Springer, 2008.

[9] Ho, T. K. "The random subspace method for constructing decision forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 20, No. 8, 1998, pp. 832–844.

[10] Schapire, R., and Y. Singer. "Improved boosting algorithms using confidence-rated predictions." *Machine Learning,* Vol. 37, No. 3, 1999, pp. 297–336.

[11] Schapire, R. E. et al. "Boosting the margin: A new explanation for the effectiveness of voting methods." *Annals of Statistics,* Vol. 26, No. 5, 1998, pp. 1651–1686.

[12] Seiffert, C., T. Khoshgoftaar, J. Hulse, and A. Napolitano. "RUSBoost: Improving classification performance when training data is skewed." *19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[13] Warmuth, M., J. Liao, and G. Ratsch. "Totally corrective boosting algorithms that maximize the margin." *Proc. 23rd Int'l. Conf. on Machine Learning, ACM,* New York, 2006, pp. 1001–1008.

See Also

`ClassificationBaggedEnsemble` | `ClassificationDiscriminant` | `ClassificationEnsemble` | `ClassificationKNN` | `ClassificationPartitionedEnsemble` | `CompactClassificationEnsemble` | `CompactRegressionEnsemble` | `RegressionBaggedEnsemble` | `RegressionEnsemble` | `RegressionPartitionedEnsemble` | `TreeBagger` | `fitcensemble` | `fitrensemble`

Related Examples

- "Framework for Ensemble Learning" on page 18-31
- "Tune RobustBoost" on page 18-101
- "Surrogate Splits" on page 18-91
- "Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles" on page 18-84
- "LPBoost and TotalBoost for Small Ensembles" on page 18-96
- "Random Subspace Classification" on page 18-104

Train Classification Ensemble

This example shows how to create a classification tree ensemble for the `ionosphere` data set, and use it to predict the classification of a radar return with average measurements.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a classification ensemble. For binary classification problems, `fitcensemble` aggregates 100 classification trees using LogitBoost.

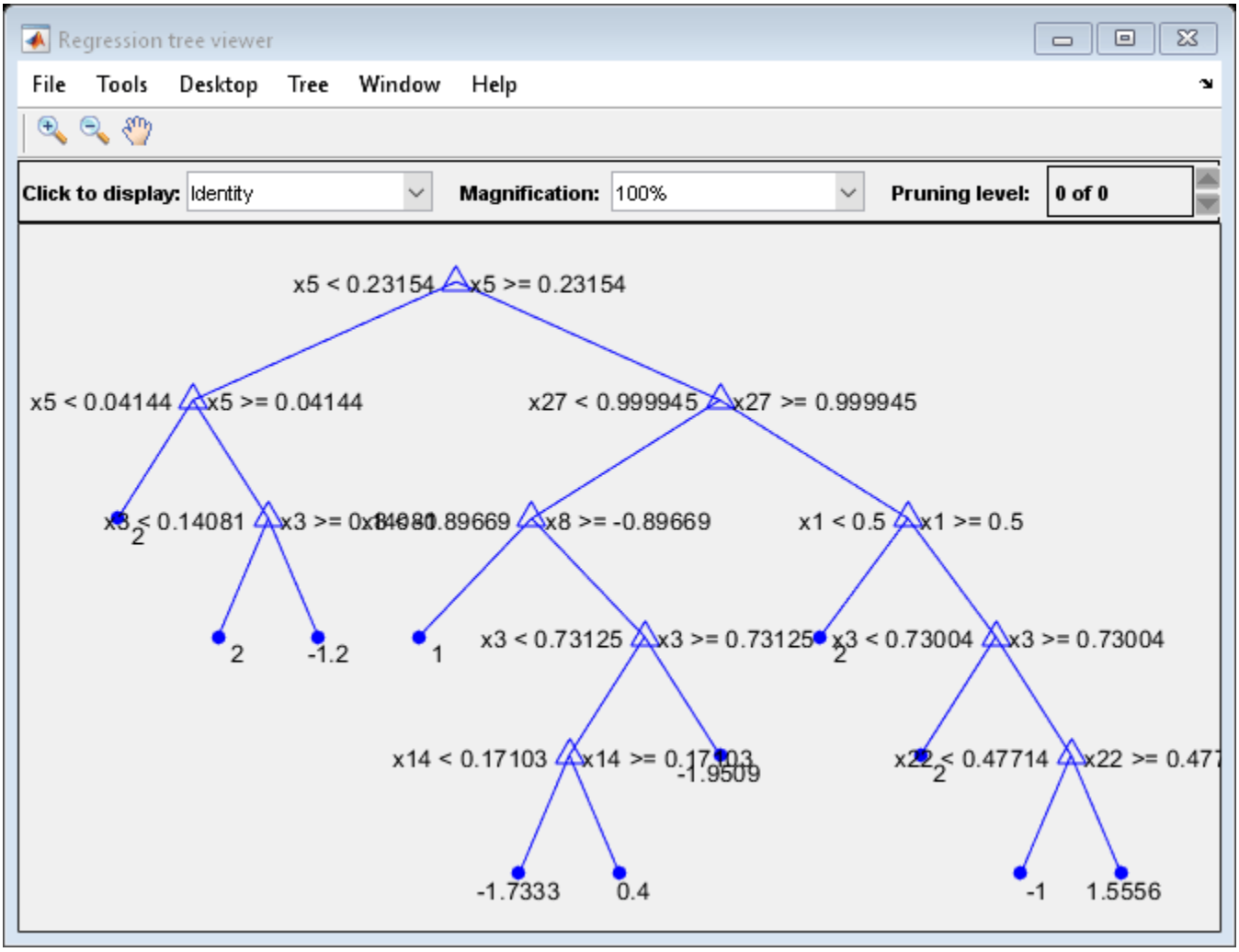
```
Mdl = fitcensemble(X,Y)
```

```
Mdl =
  ClassificationEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
                ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
                NumTrained: 100
                    Method: 'LogitBoost'
              LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}


  Properties, Methods
```

`Mdl` is a `ClassificationEnsemble` model.

Plot a graph of the first trained classification tree in the ensemble.

```
view(Mdl.Trained{1}.CompactRegressionLearner,'Mode','graph');
```



By default, `fitcensemble` grows shallow trees for boosting algorithms. You can alter the tree depth by passing a tree template object to `fitcensemble`. For more details, see `templateTree`.

Predict the quality of a radar return with average predictor measurements.

```
label = predict(Mdl,mean(X))
```

```
label = 1x1 cell array
    {'g'}
```

See Also

`fitcensemble` | `predict`

Related Examples

- “Train Regression Ensemble” on page 18-57
- “Select Predictors for Random Forests” on page 18-60

- “Decision Trees” on page 19-2
- “Ensemble Algorithms” on page 18-39
- “Framework for Ensemble Learning” on page 18-31

Train Regression Ensemble

This example shows how to create a regression ensemble to predict mileage of cars based on their horsepower and weight, trained on the `carsmall` data.

Load the `carsmall` data set.

```
load carsmall
```

Prepare the predictor data.

```
X = [Horsepower Weight];
```

The response data is `MPG`. The only available boosted regression ensemble type is `LSBoost`. For this example, arbitrarily choose an ensemble of 100 trees, and use the default tree options.

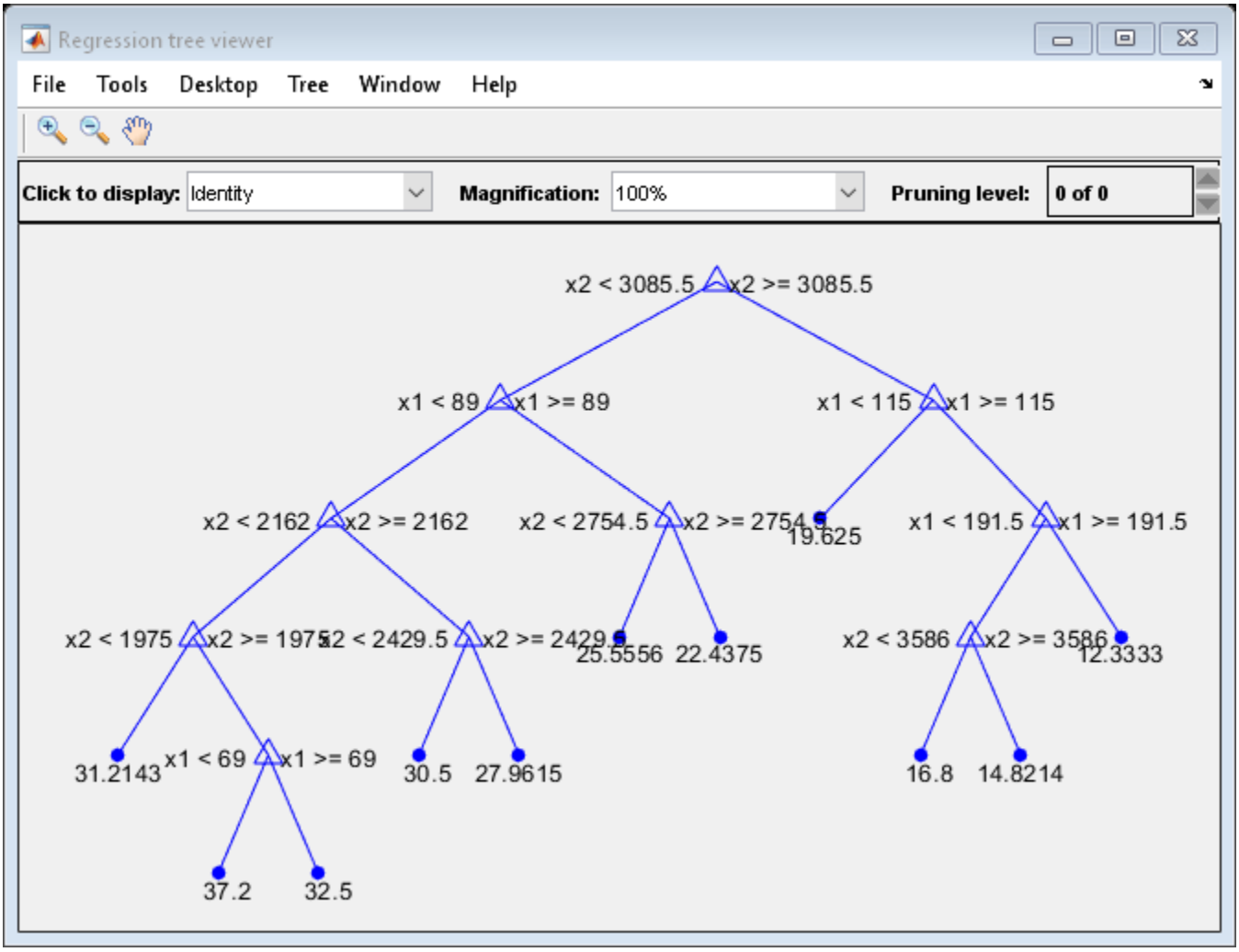
Train an ensemble of regression trees.

```
Mdl = fitrensemble(X,MPG,'Method','LSBoost','NumLearningCycles',100)
```

```
Mdl =
  RegressionEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94
                NumTrained: 100
                    Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

Plot a graph of the first trained regression tree in the ensemble.

```
view(Mdl.Trained{1},'Mode','graph');
```



By default, `fitrensemble` grows shallow trees for LSBoost.

Predict the mileage of a car with 150 horsepower weighing 2750 lbs.

```
mileage = predict(Mdl,[150 2750])
```

```
mileage = 23.6713
```

See Also

`fitrensemble` | `predict`

Related Examples

- “Train Classification Ensemble” on page 18-54
- “Select Predictors for Random Forests” on page 18-60
- “Decision Trees” on page 19-2
- “Ensemble Algorithms” on page 18-39

- “Framework for Ensemble Learning” on page 18-31

Select Predictors for Random Forests

This example shows how to choose the appropriate split predictor selection technique for your data set when growing a random forest of regression trees. This example also shows how to decide which predictors are most important to include in the training data.

Load and Preprocess Data

Load the `carbig` data set. Consider a model that predicts the fuel economy of a car given its number of cylinders, engine displacement, horsepower, weight, acceleration, model year, and country of origin. Consider `Cylinders`, `Model_Year`, and `Origin` as categorical variables.

```
load carbig
Cylinders = categorical(Cylinders);
Model_Year = categorical(Model_Year);
Origin = categorical(cellstr(Origin));
X = table(Cylinders,Displacement,Horsepower,Weight,Acceleration,Model_Year,Origin);
```

Determine Levels in Predictors

The standard CART algorithm tends to split predictors with many unique values (levels), e.g., continuous variables, over those with fewer levels, e.g., categorical variables. If your data is heterogeneous, or your predictor variables vary greatly in their number of levels, then consider using the curvature or interaction tests for split-predictor selection instead of standard CART.

For each predictor, determine the number of levels in the data. One way to do this is define an anonymous function that:

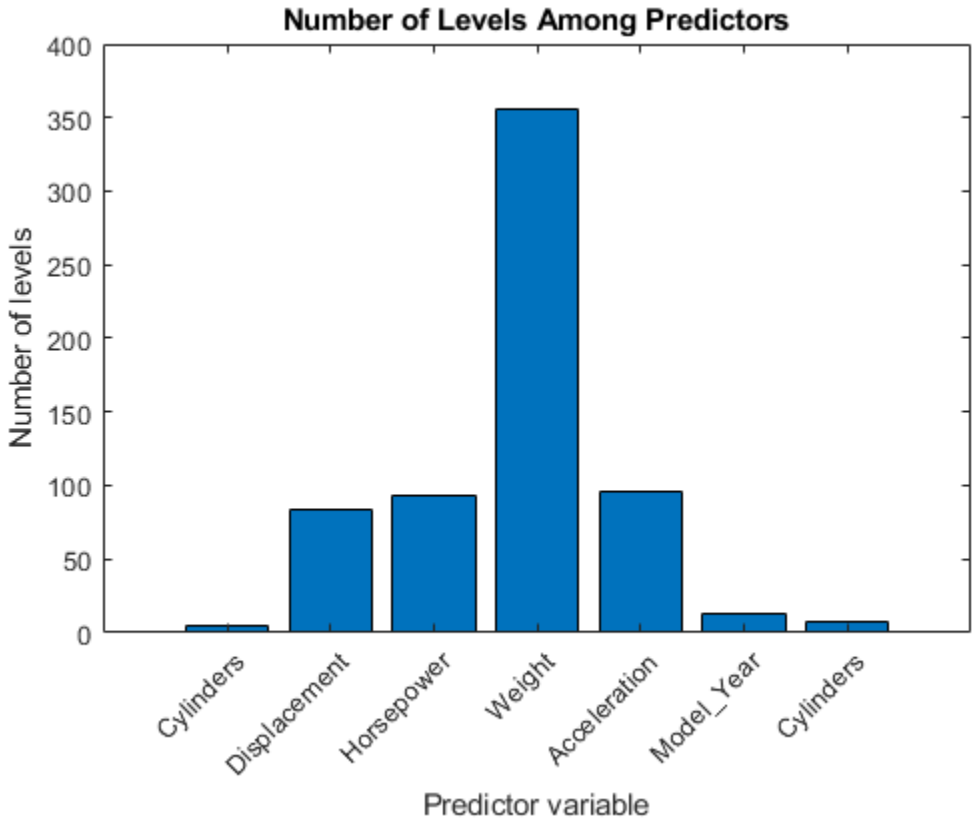
1. Converts all variables to the categorical data type using `categorical`
2. Determines all unique categories while ignoring missing values using `categories`
3. Counts the categories using `numel`

Then, apply the function to each variable using `varfun`.

```
countLevels = @(x)numel(categories(categorical(x)));
numLevels = varfun(countLevels,X,'OutputFormat','uniform');
```

Compare the number of levels among the predictor variables.

```
figure
bar(numLevels)
title('Number of Levels Among Predictors')
xlabel('Predictor variable')
ylabel('Number of levels')
h = gca;
h.XTickLabel = X.Properties.VariableNames(1:end-1);
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



The continuous variables have many more levels than the categorical variables. Because the number of levels among the predictors varies so much, using standard CART to select split predictors at each node of the trees in a random forest can yield inaccurate predictor importance estimates. In this case, use the curvature test or interaction test. Specify the algorithm by using the `'PredictorSelection'` name-value pair argument. For more details, see “Choose Split Predictor Selection Technique” on page 19-14.

Train Bagged Ensemble of Regression Trees

Train a bagged ensemble of 200 regression trees to estimate predictor importance values. Define a tree learner using these name-value pair arguments:

- `'NumVariablesToSample','all'` — Use all predictor variables at each node to ensure that each tree uses all predictor variables.
- `'PredictorSelection','interaction-curvature'` — Specify usage of the interaction test to select split predictors.
- `'Surrogate','on'` — Specify usage of surrogate splits to increase accuracy because the data set includes missing values.

```
t = templateTree('NumVariablesToSample','all',...
    'PredictorSelection','interaction-curvature','Surrogate','on');
rng(1); % For reproducibility
Mdl = fitrensemble(X,MPG,'Method','Bag','NumLearningCycles',200, ...
    'Learners',t);
```

`Mdl` is a `RegressionBaggedEnsemble` model.

Estimate the model R^2 using out-of-bag predictions.

```
yHat = oobPredict(Mdl);
R2 = corr(Mdl.Y,yHat)^2
```

```
R2 = 0.8744
```

`Mdl` explains 87% of the variability around the mean.

Predictor Importance Estimation

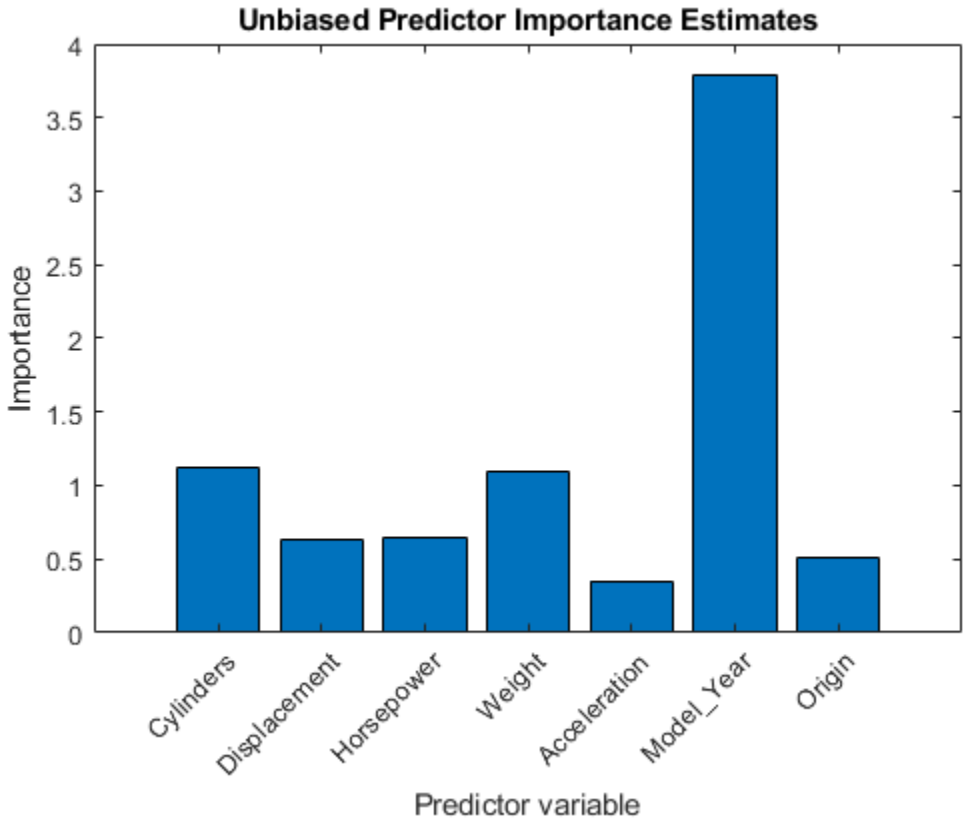
Estimate predictor importance values by permuting out-of-bag observations among the trees.

```
impOOB = oobPermutedPredictorImportance(Mdl);
```

`impOOB` is a 1-by-7 vector of predictor importance estimates corresponding to the predictors in `Mdl.PredictorNames`. The estimates are not biased toward predictors containing many levels.

Compare the predictor importance estimates.

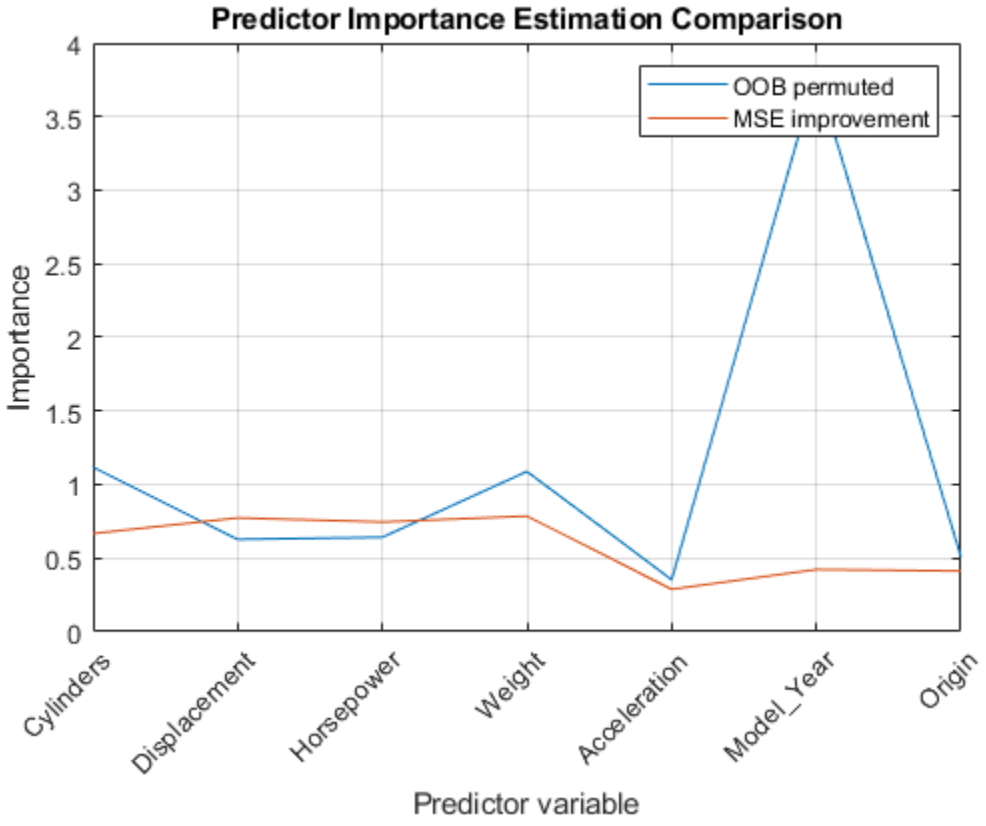
```
figure
bar(impOOB)
title('Unbiased Predictor Importance Estimates')
xlabel('Predictor variable')
ylabel('Importance')
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



Greater importance estimates indicate more important predictors. The bar graph suggests that `Model_Year` is the most important predictor, followed by `Cylinders` and `Weight`. The `Model_Year` and `Cylinders` variables have only 13 and 5 distinct levels, respectively, whereas the `Weight` variable has over 300 levels.

Compare predictor importance estimates by permuting out-of-bag observations and those estimates obtained by summing gains in the mean squared error due to splits on each predictor. Also, obtain predictor association measures estimated by surrogate splits.

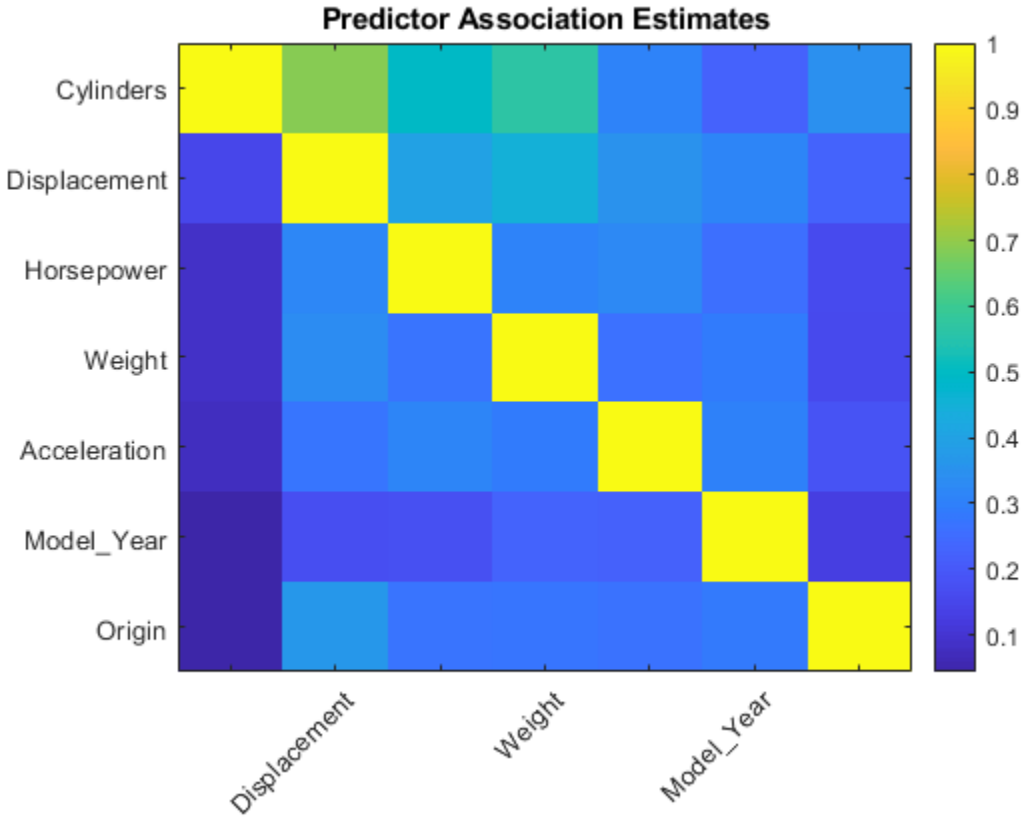
```
[impGain,predAssociation] = predictorImportance(Mdl);

figure
plot(1:numel(Mdl.PredictorNames),[impOOB' impGain'])
title('Predictor Importance Estimation Comparison')
xlabel('Predictor variable')
ylabel('Importance')
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
legend('OOB permuted','MSE improvement')
grid on
```



According to the values of `impGain`, the variables `Displacement`, `Horsepower`, and `Weight` appear to be equally important.

`predAssociation` is a 7-by-7 matrix of predictor association measures. Rows and columns correspond to the predictors in `Mdl.PredictorNames`. The “Predictive Measure of Association” on page 33-5044 is a value that indicates the similarity between decision rules that split observations. The best surrogate decision split yields the maximum predictive measure of association. You can infer the strength of the relationship between pairs of predictors using the elements of `predAssociation`. Larger values indicate more highly correlated pairs of predictors.

```
figure
imagesc(predAssociation)
title('Predictor Association Estimates')
colorbar
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
h.YTickLabel = Mdl.PredictorNames;
```




```
predAssociation(1,2)
```

```
ans = 0.6871
```

The largest association is between `Cylinders` and `Displacement`, but the value is not high enough to indicate a strong relationship between the two predictors.

Grow Random Forest Using Reduced Predictor Set

Because prediction time increases with the number of predictors in random forests, a good practice is to create a model using as few predictors as possible.

Grow a random forest of 200 regression trees using the best two predictors only. The default `'NumVariablesToSample'` value of `templateTree` is one third of the number of predictors for regression, so `fitrensemble` uses the random forest algorithm.

```
t = templateTree('PredictorSelection','interaction-curvature','Surrogate','on', ...
    'Reproducible',true); % For reproducibility of random predictor selections
MdlReduced = fitrensemble(X(:,{'Model_Year' 'Weight'}),MPG,'Method','Bag', ...
    'NumLearningCycles',200,'Learners',t);
```

Compute the R^2 of the reduced model.

```
yHatReduced = oobPredict(MdlReduced);
r2Reduced = corr(Mdl.Y,yHatReduced)^2
```

```
r2Reduced = 0.8653
```

The R^2 for the reduced model is close to the R^2 of the full model. This result suggests that the reduced model is sufficient for prediction.

See Also

`corr` | `fitrensemble` | `oobPermutedPredictorImportance` | `oobPredict` | `predictorImportance` | `templateTree`

Related Examples

- “Improving Classification Trees and Regression Trees” on page 19-13
- “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
- “Surrogate Splits” on page 18-91
- “Introduction to Feature Selection” on page 15-49
- “Interpret Machine Learning Models” on page 18-256

Test Ensemble Quality

You cannot evaluate the predictive quality of an ensemble based on its performance on training data. Ensembles tend to "overtrain," meaning they produce overly optimistic estimates of their predictive power. This means the result of `resubLoss` for classification (`resubLoss` for regression) usually indicates lower error than you get on new data.

To obtain a better idea of the quality of an ensemble, use one of these methods:

- Evaluate the ensemble on an independent test set (useful when you have a lot of training data).
- Evaluate the ensemble by cross validation (useful when you don't have a lot of training data).
- Evaluate the ensemble on out-of-bag data (useful when you create a bagged ensemble with `fitcensemble` or `fitrensemble`).

This example uses a bagged ensemble so it can use all three methods of evaluating ensemble quality.

Generate an artificial dataset with 20 predictors. Each entry is a random number from 0 to 1. The initial classification is $Y = 1$ if $X_1 + X_2 + X_3 + X_4 + X_5 > 2.5$ and $Y = 0$ otherwise.

```
rng(1,'twister') % For reproducibility
X = rand(2000,20);
Y = sum(X(:,1:5),2) > 2.5;
```

In addition, to add noise to the results, randomly switch 10% of the classifications.

```
idx = randsample(2000,200);
Y(idx) = ~Y(idx);
```

Independent Test Set

Create independent training and test sets of data. Use 70% of the data for a training set by calling `cvpartition` using the `holdout` option.

```
cvpart = cvpartition(Y,'holdout',0.3);
Xtrain = X(training(cvpart),:);
Ytrain = Y(training(cvpart),:);
Xtest = X(test(cvpart),:);
Ytest = Y(test(cvpart),:);
```

Create a bagged classification ensemble of 200 trees from the training data.

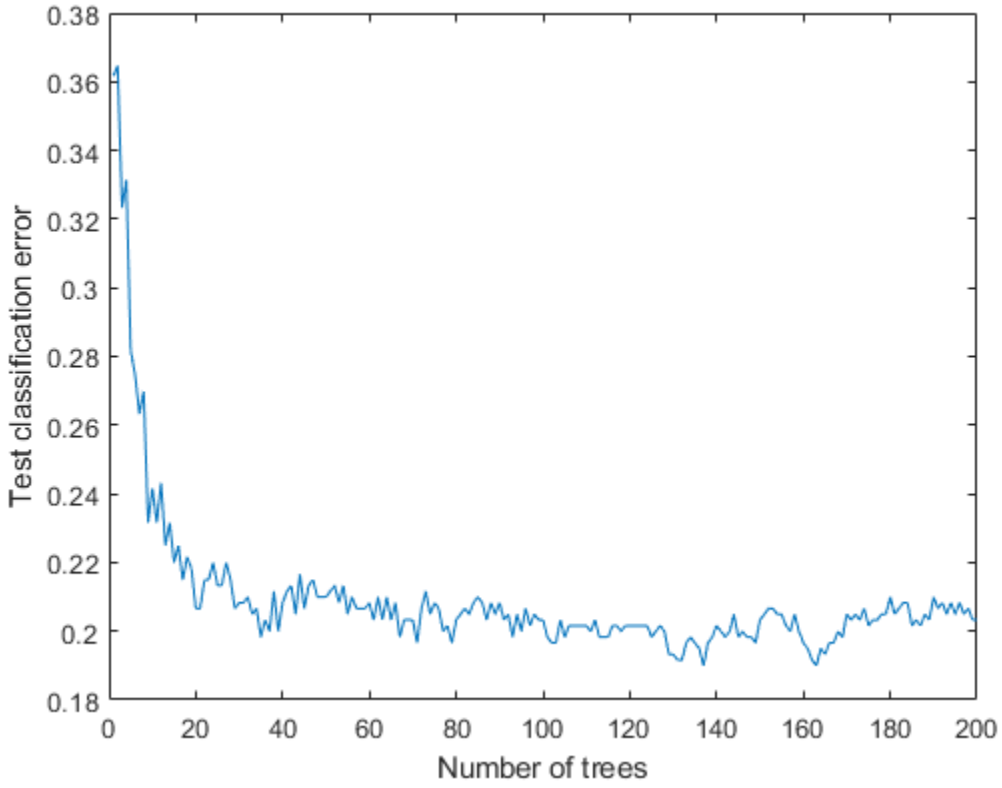
```
t = templateTree('Reproducible',true);  % For reproducibility of random predictor selections
bag = fitcensemble(Xtrain,Ytrain,'Method','Bag','NumLearningCycles',200,'Learners',t)
```

```
bag =
  ClassificationBaggedEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [0 1]
           ScoreTransform: 'none'
          NumObservations: 1400
               NumTrained: 200
                   Method: 'Bag'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: []
```

```
       FitInfoDescription: 'None'
                FResample: 1
                  Replace: 1
          UseObsForLearner: [1400x200 logical]


  Properties, Methods
```

Plot the loss (misclassification) of the test data as a function of the number of trained trees in the ensemble.

```
figure
plot(loss(bag,Xtest,Ytest,'mode','cumulative'))
xlabel('Number of trees')
ylabel('Test classification error')
```



Cross Validation

Generate a five-fold cross-validated bagged ensemble.

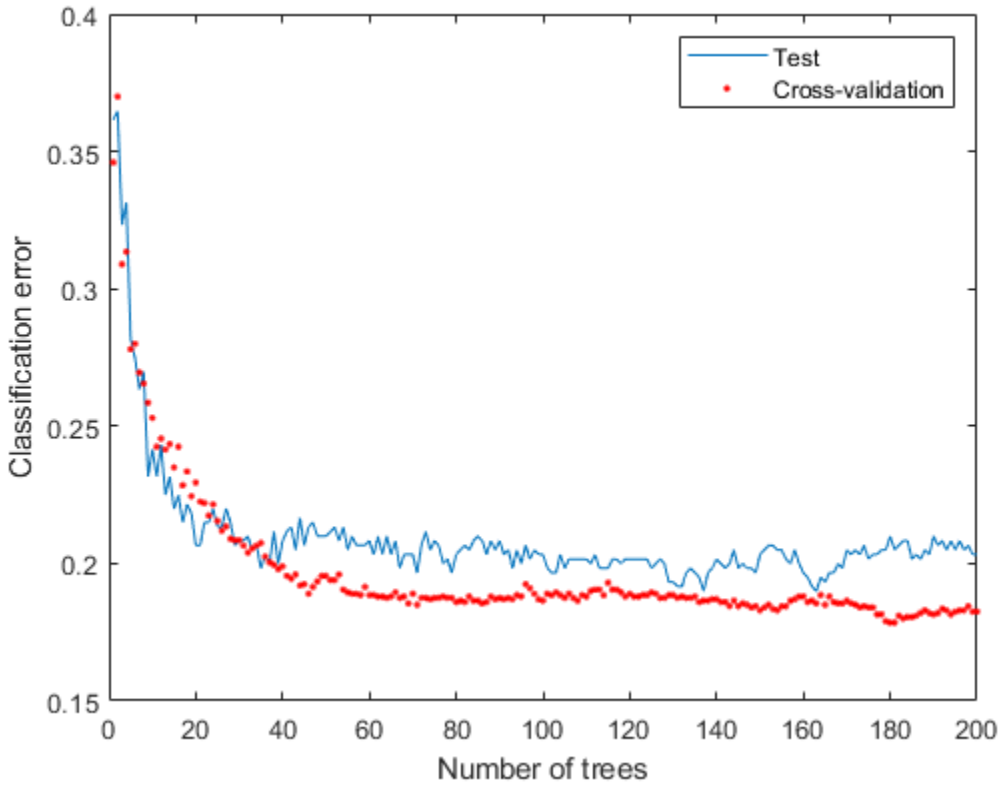
```
cv = fitcensemble(X,Y,'Method','Bag','NumLearningCycles',200,'Kfold',5,'Learners',t)
```

```
cv =
  ClassificationPartitionedEnsemble
    CrossValidatedModel: 'Bag'
         PredictorNames: {1x20 cell}
           ResponseName: 'Y'
```

```
       NumObservations: 2000
                 KFold: 5
             Partition: [1x1 cvpartition]
    NumTrainedPerFold: [200 200 200 200 200]
           ClassNames: [0 1]
       ScoreTransform: 'none'


  Properties, Methods
```

Examine the cross-validation loss as a function of the number of trees in the ensemble.

```
figure
plot(loss(bag,Xtest,Ytest,'mode','cumulative'))
hold on
plot(kfoldLoss(cv,'mode','cumulative'),'r.')
hold off
xlabel('Number of trees')
ylabel('Classification error')
legend('Test','Cross-validation','Location','NE')
```

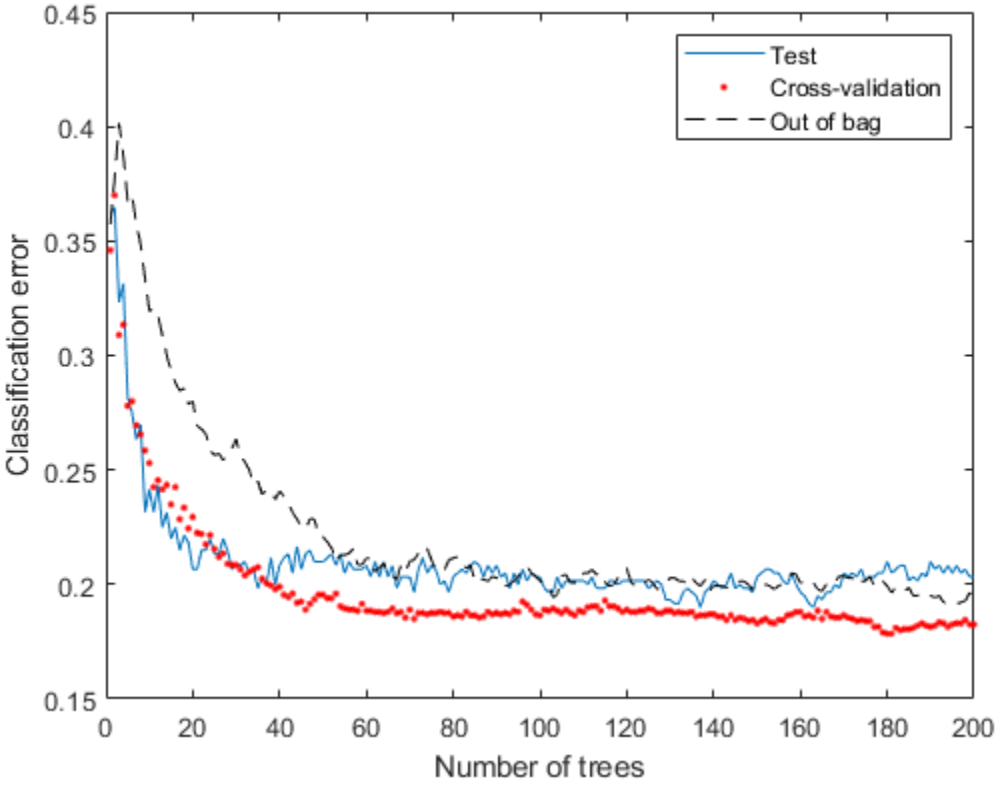


Cross validating gives comparable estimates to those of the independent set.

Out-of-Bag Estimates

Generate the loss curve for out-of-bag estimates, and plot it along with the other curves.

```
figure
plot(loss(bag,Xtest,Ytest,'mode','cumulative'))
hold on
plot(kfoldLoss(cv,'mode','cumulative'),'r.')
plot(oobLoss(bag,'mode','cumulative'),'k--')
hold off
xlabel('Number of trees')
ylabel('Classification error')
legend('Test','Cross-validation','Out of bag','Location','NE')
```



The out-of-bag estimates are again comparable to those of the other methods.

See Also

`cvpartition` | `fitcensemble` | `fitrensemble` | `kfoldLoss` | `loss` | `oobLoss` | `resubLoss` | `resubLoss`

Related Examples

- “Framework for Ensemble Learning” on page 18-31
- “Ensemble Algorithms” on page 18-39
- “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

Ensemble Regularization

Regularization is a process of choosing fewer weak learners for an ensemble in a way that does not diminish predictive performance. Currently you can regularize regression ensembles. (You can also regularize a discriminant analysis classifier in a non-ensemble context; see “Regularize Discriminant Analysis Classifier” on page 20-21.)

The `regularize` method finds an optimal set of learner weights α_t that minimize

$$\sum_{n=1}^{N} w_n g\left(\left(\sum_{t=1}^{T} \alpha_t h_t(x_n)\right), y_n\right) + \lambda \sum_{t=1}^{T} |\alpha_t| .$$

Here

- $\lambda \geq 0$ is a parameter you provide, called the lasso parameter.
- h_t is a weak learner in the ensemble trained on N observations with predictors x_n, responses y_n, and weights w_n.
- $g(f,y) = (f - y)^2$ is the squared error.

The ensemble is regularized on the same (x_n,y_n,w_n) data used for training, so

$$\sum_{n=1}^{N} w_n g\left(\left(\sum_{t=1}^{T} \alpha_t h_t(x_n)\right), y_n\right)$$

is the ensemble resubstitution error. The error is measured by mean squared error (MSE).

If you use $\lambda = 0$, `regularize` finds the weak learner weights by minimizing the resubstitution MSE. Ensembles tend to overtrain. In other words, the resubstitution error is typically smaller than the true generalization error. By making the resubstitution error even smaller, you are likely to make the ensemble accuracy worse instead of improving it. On the other hand, positive values of λ push the magnitude of the α_t coefficients to 0. This often improves the generalization error. Of course, if you choose λ too large, all the optimal coefficients are 0, and the ensemble does not have any accuracy. Usually you can find an optimal range for λ in which the accuracy of the regularized ensemble is better or comparable to that of the full ensemble without regularization.

A nice feature of lasso regularization is its ability to drive the optimized coefficients precisely to 0. If a learner's weight α_t is 0, this learner can be excluded from the regularized ensemble. In the end, you get an ensemble with improved accuracy and fewer learners.

Regularize a Regression Ensemble

This example uses data for predicting the insurance risk of a car based on its many attributes.

Load the `imports-85` data into the MATLAB workspace.

```
load imports-85;
```

Look at a description of the data to find the categorical variables and predictor names.

```
Description
```

```
Description = 9x79 char array
    '1985 Auto Imports Database from the UCI repository                          '
    'http://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names'
    'Variables have been reordered to place variables with numeric values (referred '
    'to as "continuous" on the UCI site) to the left and categorical values to the  '
    'right. Specifically, variables 1:16 are: symboling, normalized-losses,          '
    'wheel-base, length, width, height, curb-weight, engine-size, bore, stroke,      '
    'compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, and price.      '
    'Variables 17:26 are: make, fuel-type, aspiration, num-of-doors, body-style,     '
    'drive-wheels, engine-location, engine-type, num-of-cylinders, and fuel-system. '
```

The objective of this process is to predict the "symboling," the first variable in the data, from the other predictors. "symboling" is an integer from `-3` (good insurance risk) to `3` (poor insurance risk). You could use a classification ensemble to predict this risk instead of a regression ensemble. When you have a choice between regression and classification, you should try regression first.

Prepare the data for ensemble fitting.

```
Y = X(:,1);
X(:,1) = [];
VarNames = {'normalized-losses' 'wheel-base' 'length' 'width' 'height' ...
  'curb-weight' 'engine-size' 'bore' 'stroke' 'compression-ratio' ...
  'horsepower' 'peak-rpm' 'city-mpg' 'highway-mpg' 'price' 'make' ...
  'fuel-type' 'aspiration' 'num-of-doors' 'body-style' 'drive-wheels' ...
  'engine-location' 'engine-type' 'num-of-cylinders' 'fuel-system'};
catidx = 16:25; % indices of categorical predictors
```

Create a regression ensemble from the data using 300 trees.

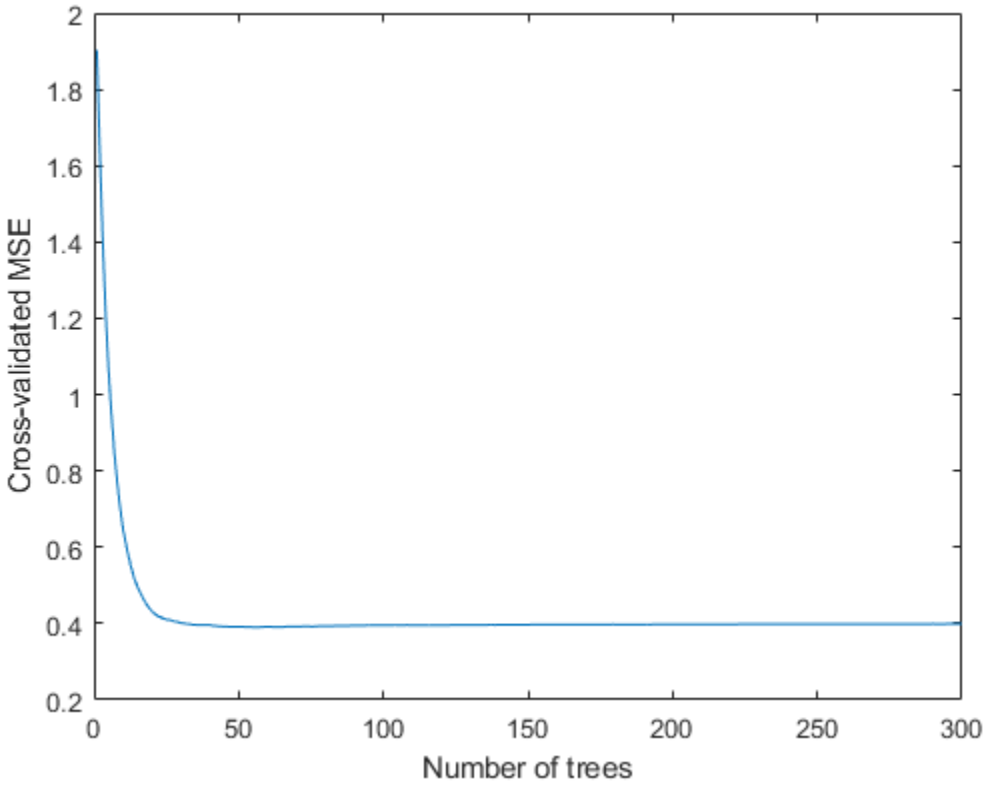
```
ls = fitrensemble(X,Y,'Method','LSBoost','NumLearningCycles',300, ...
    'LearnRate',0.1,'PredictorNames',VarNames, ...
    'ResponseName','Symboling','CategoricalPredictors',catidx)
```

```
ls =
  RegressionEnsemble
            PredictorNames: {1x25 cell}
              ResponseName: 'Symboling'
    CategoricalPredictors: [16 17 18 19 20 21 22 23 24 25]
        ResponseTransform: 'none'
          NumObservations: 205
                NumTrained: 300
                    Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [300x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

The final line, `Regularization`, is empty ([]). To regularize the ensemble, you have to use the `regularize` method.

```
cv = crossval(ls,'KFold',5);
figure;
```

```
plot(kfoldLoss(cv,'Mode','Cumulative'));
xlabel('Number of trees');
ylabel('Cross-validated MSE');
ylim([0.2,2])
```



It appears you might obtain satisfactory performance from a smaller ensemble, perhaps one containing from 50 to 100 trees.

Call the `regularize` method to try to find trees that you can remove from the ensemble. By default, `regularize` examines 10 values of the lasso (`Lambda`) parameter spaced exponentially.

```
ls = regularize(ls)
```

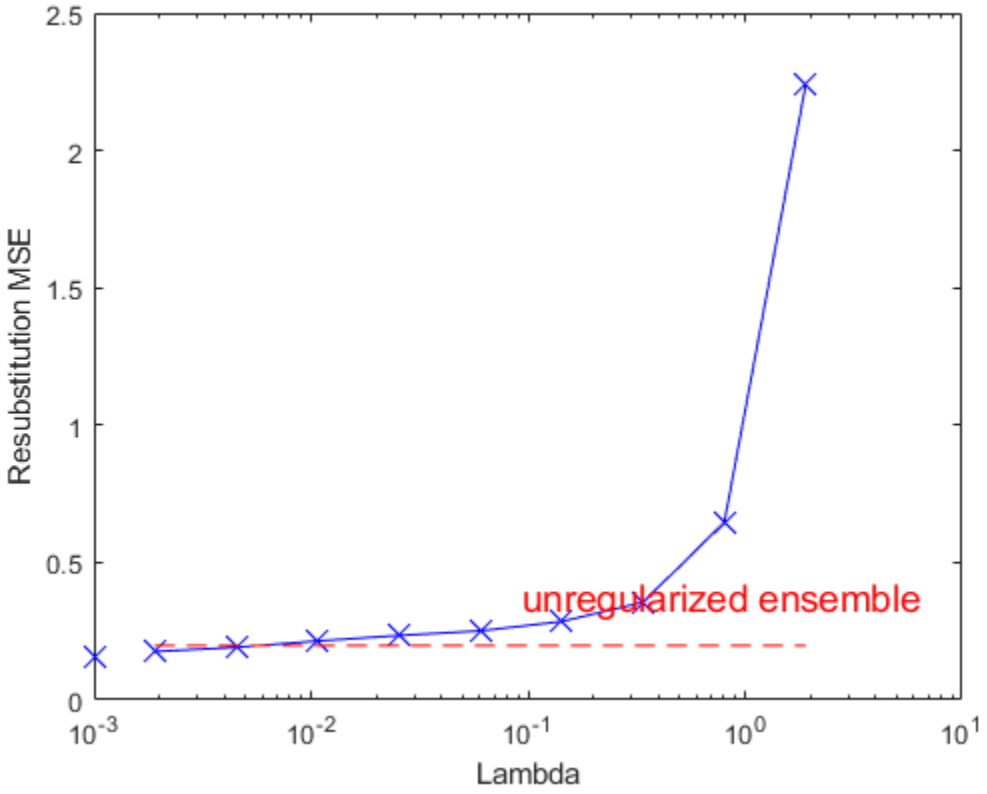
```
ls =
  RegressionEnsemble
            PredictorNames: {1x25 cell}
              ResponseName: 'Symboling'
    CategoricalPredictors: [16 17 18 19 20 21 22 23 24 25]
        ResponseTransform: 'none'
          NumObservations: 205
               NumTrained: 300
                   Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [300x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: [1x1 struct]
```

```
  Properties, Methods
```

The `Regularization` property is no longer empty.

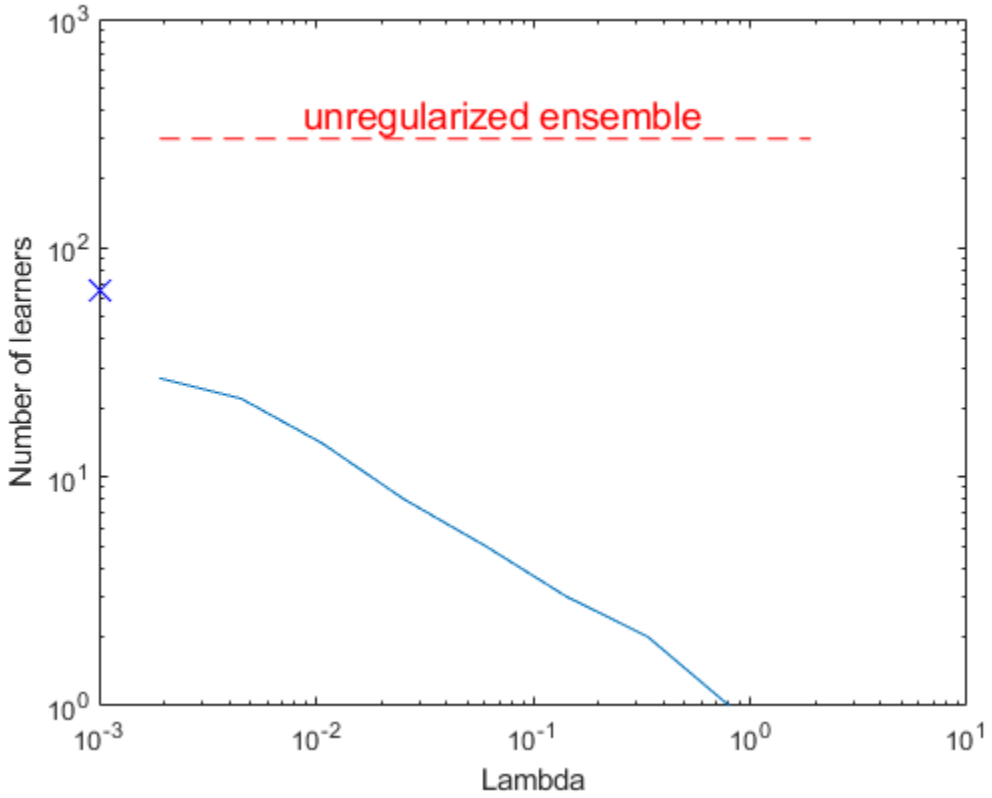
Plot the resubstitution mean-squared error (MSE) and number of learners with nonzero weights against the lasso parameter. Separately plot the value at `Lambda = 0`. Use a logarithmic scale because the values of `Lambda` are exponentially spaced.

```
figure;
semilogx(ls.Regularization.Lambda,ls.Regularization.ResubstitutionMSE, ...
    'bx-','Markersize',10);
line([1e-3 1e-3],[ls.Regularization.ResubstitutionMSE(1) ...
     ls.Regularization.ResubstitutionMSE(1)],...
    'Marker','x','Markersize',10,'Color','b');
r0 = resubLoss(ls);
line([ls.Regularization.Lambda(2) ls.Regularization.Lambda(end)],...
     [r0 r0],'Color','r','LineStyle','--');
xlabel('Lambda');
ylabel('Resubstitution MSE');
annotation('textbox',[0.5 0.22 0.5 0.05],'String','unregularized ensemble', ...
    'Color','r','FontSize',14,'LineStyle','none');
```




```
figure;
loglog(ls.Regularization.Lambda,sum(ls.Regularization.TrainedWeights>0,1));
line([1e-3 1e-3],...
```

```
    [sum(ls.Regularization.TrainedWeights(:,1)>0) ...
    sum(ls.Regularization.TrainedWeights(:,1)>0)],...
    'marker','x','markersize',10,'color','b');
line([ls.Regularization.Lambda(2) ls.Regularization.Lambda(end)],...
    [ls.NTrained ls.NTrained],...
    'color','r','LineStyle','--');
xlabel('Lambda');
ylabel('Number of learners');
annotation('textbox',[0.3 0.8 0.5 0.05],'String','unregularized ensemble',...
    'color','r','FontSize',14,'LineStyle','none');
```



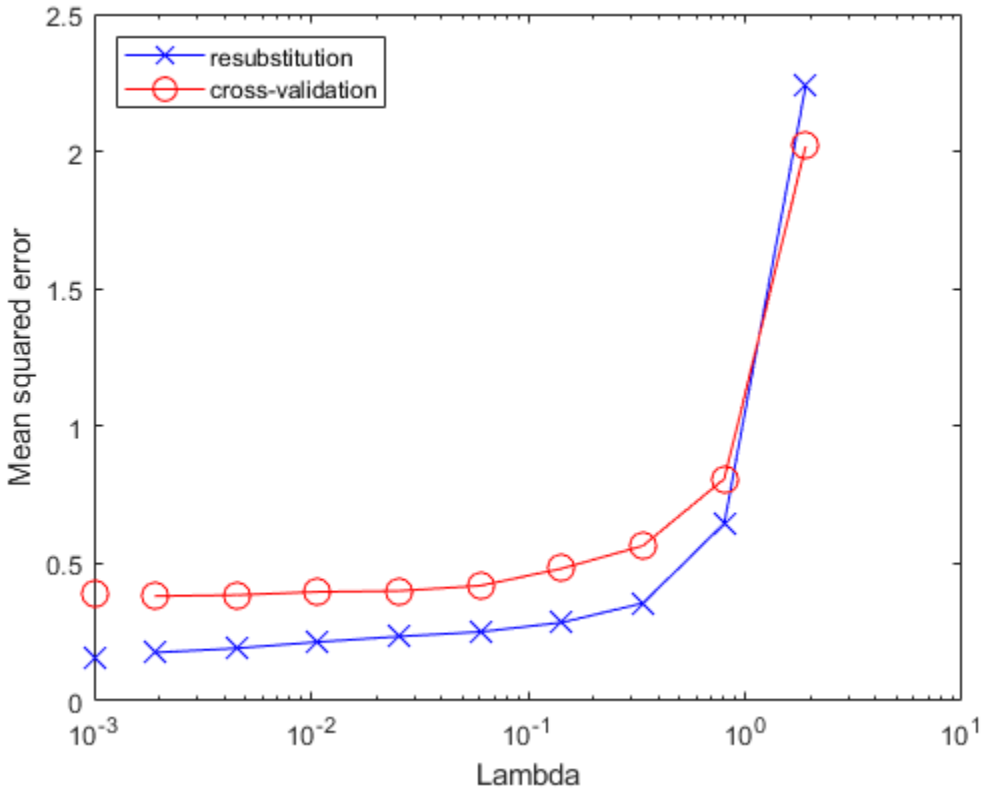
The resubstitution MSE values are likely to be overly optimistic. To obtain more reliable estimates of the error associated with various values of `Lambda`, cross validate the ensemble using `cvshrink`. Plot the resulting cross-validation loss (MSE) and number of learners against `Lambda`.

```
rng(0,'Twister') % for reproducibility
[mse,nlearn] = cvshrink(ls,'Lambda',ls.Regularization.Lambda,'KFold',5);
```

```
Warning: Some folds do not have any trained weak learners.
```

```
figure;
semilogx(ls.Regularization.Lambda,ls.Regularization.ResubstitutionMSE, ...
    'bx-','Markersize',10);
hold on;
semilogx(ls.Regularization.Lambda,mse,'ro-','Markersize',10);
hold off;
xlabel('Lambda');
```

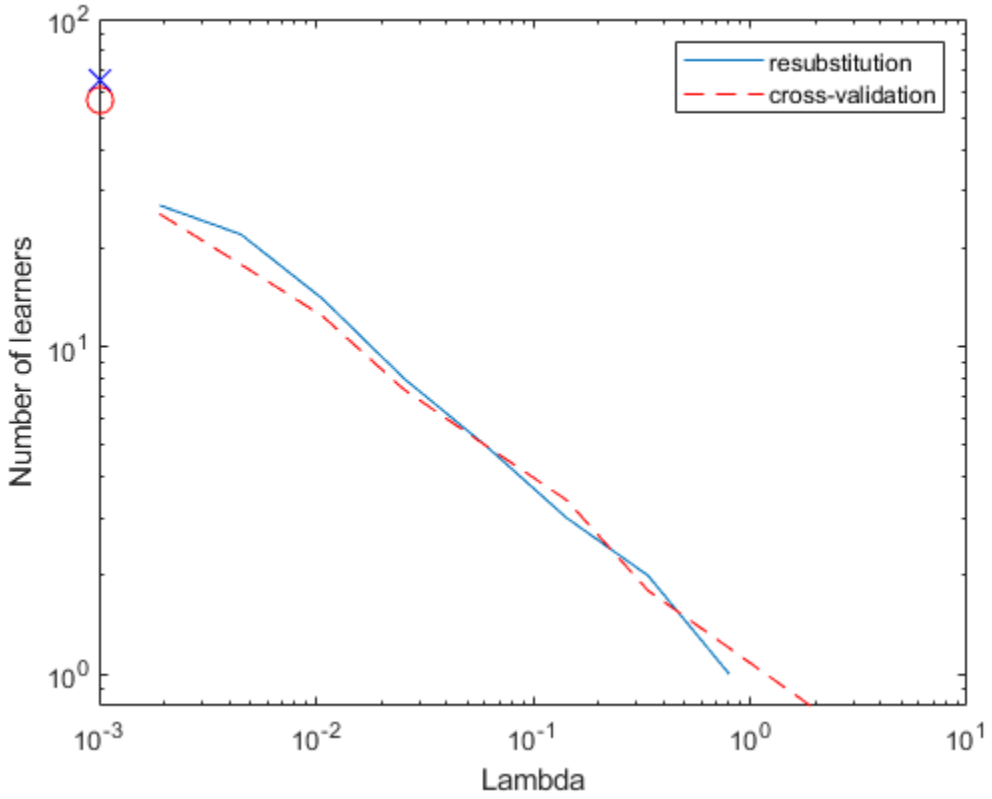
```
ylabel('Mean squared error');
legend('resubstitution','cross-validation','Location','NW');
line([1e-3 1e-3],[ls.Regularization.ResubstitutionMSE(1) ...
     ls.Regularization.ResubstitutionMSE(1)],...
    'Marker','x','Markersize',10,'Color','b','HandleVisibility','off');
line([1e-3 1e-3],[mse(1) mse(1)],'Marker','o',...
    'Markersize',10,'Color','r','LineStyle','--','HandleVisibility','off');
```




```
figure;
loglog(ls.Regularization.Lambda,sum(ls.Regularization.TrainedWeights>0,1));
hold;
```

```
Current plot held
```

```
loglog(ls.Regularization.Lambda,nlearn,'r--');
hold off;
xlabel('Lambda');
ylabel('Number of learners');
legend('resubstitution','cross-validation','Location','NE');
line([1e-3 1e-3],...
    [sum(ls.Regularization.TrainedWeights(:,1)>0) ...
    sum(ls.Regularization.TrainedWeights(:,1)>0)],...
    'Marker','x','Markersize',10,'Color','b','HandleVisibility','off');
line([1e-3 1e-3],[nlearn(1) nlearn(1)],'marker','o',...
    'Markersize',10,'Color','r','LineStyle','--','HandleVisibility','off');
```



Examining the cross-validated error shows that the cross-validation MSE is almost flat for `Lambda` up to a bit over `1e-2`.

Examine `ls.Regularization.Lambda` to find the highest value that gives MSE in the flat region (up to a bit over `1e-2`).

```
jj = 1:length(ls.Regularization.Lambda);
[jj;ls.Regularization.Lambda]
```

```
ans = 2×10

    1.0000    2.0000    3.0000    4.0000    5.0000    6.0000    7.0000    8.0000    9.0000   10.0
         0    0.0019    0.0045    0.0107    0.0254    0.0602    0.1428    0.3387    0.8033    1.9
```

Element `5` of `ls.Regularization.Lambda` has value `0.0254`, the largest in the flat range.

Reduce the ensemble size using the `shrink` method. `shrink` returns a compact ensemble with no training data. The generalization error for the new compact ensemble was already estimated by cross validation in `mse(5)`.

```
cmp = shrink(ls,'weightcolumn',5)
```

```
cmp =
  CompactRegressionEnsemble
            PredictorNames: {1x25 cell}
              ResponseName: 'Symboling'
    CategoricalPredictors: [16 17 18 19 20 21 22 23 24 25]
```

```
        ResponseTransform: 'none'
                NumTrained: 8


  Properties, Methods
```

The number of trees in the new ensemble has notably reduced from the 300 in `ls`.

Compare the sizes of the ensembles.

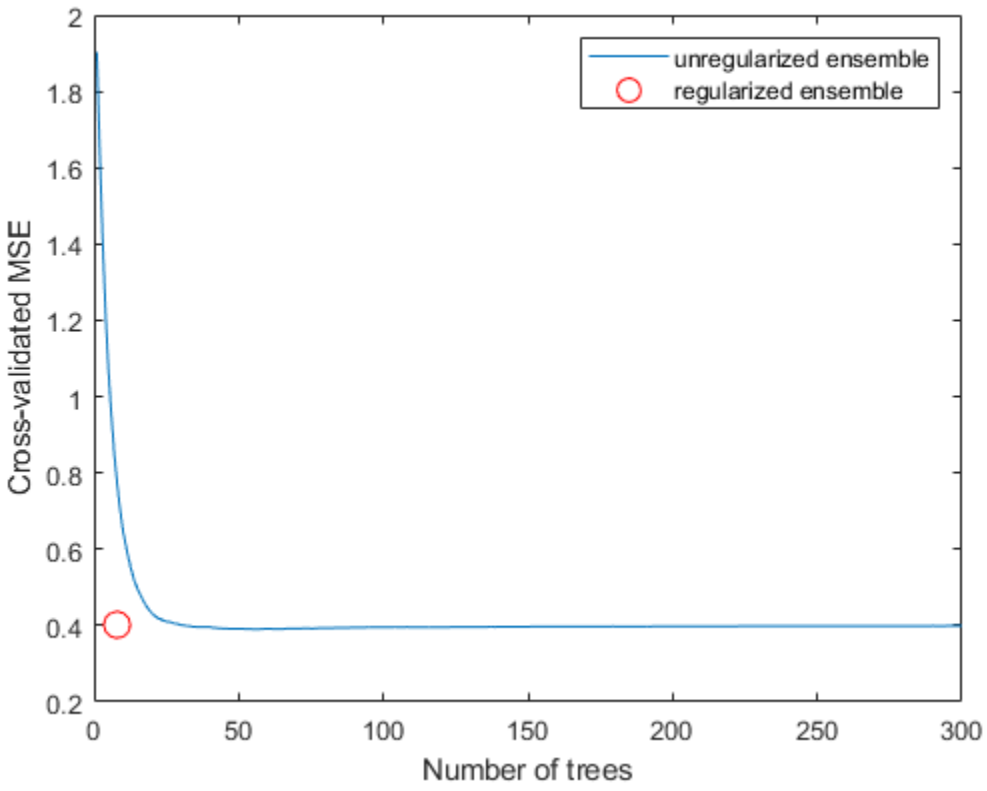
```
sz(1) = whos('cmp'); sz(2) = whos('ls');
[sz(1).bytes sz(2).bytes]
```

```
ans = 1×2

       91209     3226892
```

The size of the reduced ensemble is a fraction of the size of the original. Note that your ensemble sizes can vary depending on your operating system.

Compare the MSE of the reduced ensemble to that of the original ensemble.

```
figure;
plot(kfoldLoss(cv,'mode','cumulative'));
hold on
plot(cmp.NTrained,mse(5),'ro','MarkerSize',10);
xlabel('Number of trees');
ylabel('Cross-validated MSE');
legend('unregularized ensemble','regularized ensemble',...
    'Location','NE');
hold off
```



The reduced ensemble gives low loss while using many fewer trees.

See Also

`crossval` | `cvshrink` | `fitrensemble` | `kfoldLoss` | `regularize` | `resubLoss` | `shrink`

Related Examples

- “Regularize Discriminant Analysis Classifier” on page 20-21
- “Framework for Ensemble Learning” on page 18-31
- “Ensemble Algorithms” on page 18-39

Classification with Imbalanced Data

This example shows how to perform classification when one class has many more observations than another. You use the `RUSBoost` algorithm first, because it is designed to handle this case. Another way to handle imbalanced data is to use the name-value pair arguments `'Prior'` or `'Cost'`. For details, see “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84.

This example uses the "Cover type" data from the UCI machine learning archive, described in https://archive.ics.uci.edu/ml/datasets/Covertype. The data classifies types of forest (ground cover), based on predictors such as elevation, soil type, and distance to water. The data has over 500,000 observations and over 50 predictors, so training and using a classifier is time consuming.

Blackard and Dean [1] describe a neural net classification of this data. They quote a 70.6% classification accuracy. `RUSBoost` obtains over 81% classification accuracy.

Obtain the data

Import the data into your workspace. Extract the last data column into a variable named `Y`.

```
gunzip('https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz')
load covtype.data
Y = covtype(:,end);
covtype(:,end) = [];
```

Examine the response data

```
tabulate(Y)
```

```
  Value    Count   Percent
      1   211840     36.46%
      2   283301     48.76%
      3    35754      6.15%
      4     2747      0.47%
      5     9493      1.63%
      6    17367      2.99%
      7    20510      3.53%
```

There are hundreds of thousands of data points. Those of class 4 are less than 0.5% of the total. This imbalance indicates that `RUSBoost` is an appropriate algorithm.

Partition the data for quality assessment

Use half the data to fit a classifier, and half to examine the quality of the resulting classifier.

```
rng(10,'twister')         % For reproducibility
part = cvpartition(Y,'Holdout',0.5);
istrain = training(part); % Data for fitting
istest = test(part);      % Data for quality assessment
tabulate(Y(istrain))
```

```
  Value    Count   Percent
      1   105919     36.46%
      2   141651     48.76%
      3    17877      6.15%
      4     1374      0.47%
      5     4747      1.63%
```

```
      6     8684      2.99%
      7    10254      3.53%
```

Create the ensemble

Use deep trees for higher ensemble accuracy. To do so, set the trees to have maximal number of decision splits of `N`, where `N` is the number of observations in the training sample. Set `LearnRate` to `0.1` in order to achieve higher accuracy as well. The data is large, and, with deep trees, creating the ensemble is time consuming.

```
N = sum(istrain);         % Number of observations in the training sample
t = templateTree('MaxNumSplits',N);
tic
rusTree = fitcensemble(covtype(istrain,:),Y(istrain),'Method','RUSBoost', ...
    'NumLearningCycles',1000,'Learners',t,'LearnRate',0.1,'nprint',100);
```

```
Training RUSBoost...
Grown weak learners: 100
Grown weak learners: 200
Grown weak learners: 300
Grown weak learners: 400
Grown weak learners: 500
Grown weak learners: 600
Grown weak learners: 700
Grown weak learners: 800
Grown weak learners: 900
Grown weak learners: 1000
```

```
toc
```

```
Elapsed time is 242.836734 seconds.
```

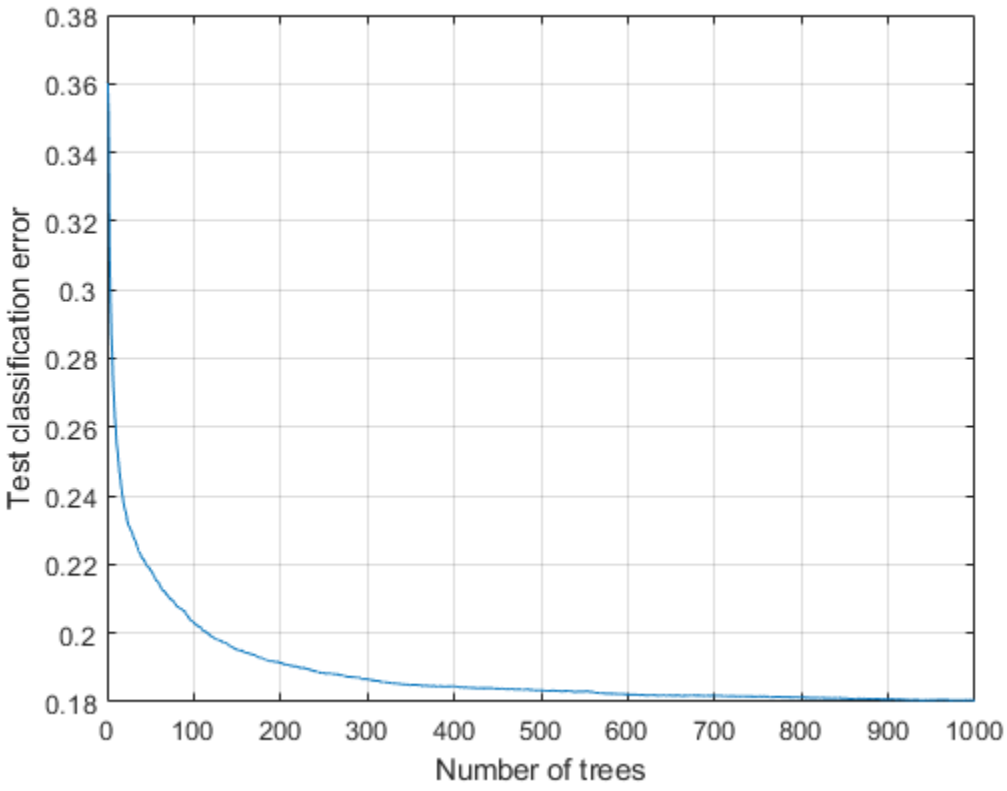
Inspect the classification error

Plot the classification error against the number of members in the ensemble.

```
figure;
tic
plot(loss(rusTree,covtype(istest,:),Y(istest),'mode','cumulative'));
toc
```

```
Elapsed time is 164.470086 seconds.
```

```
grid on;
xlabel('Number of trees');
ylabel('Test classification error');
```



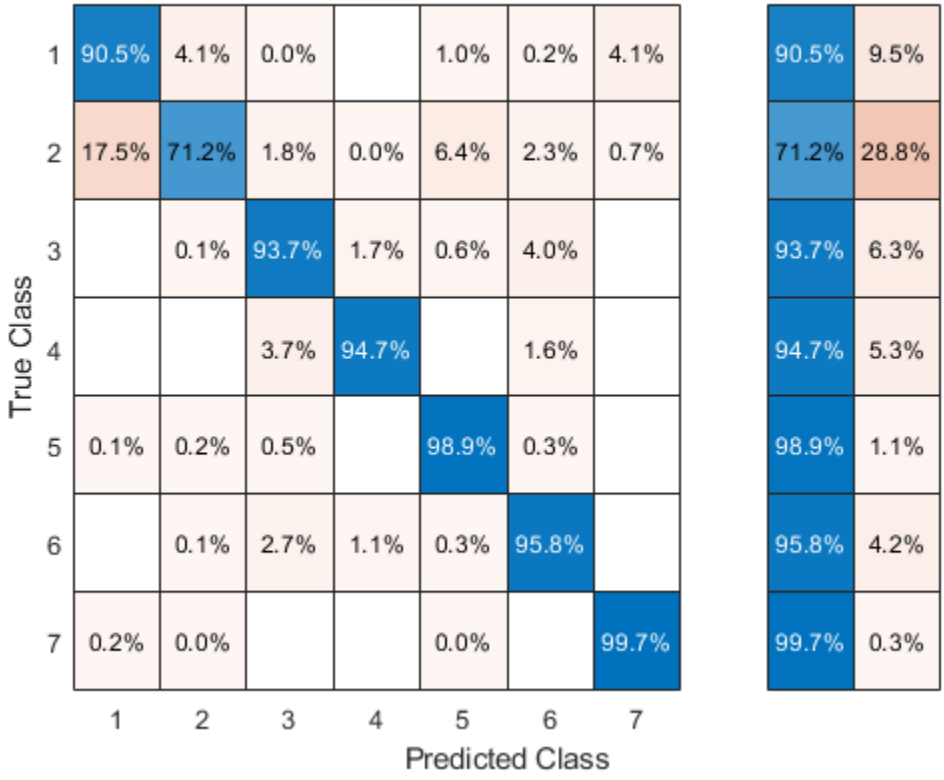
The ensemble achieves a classification error of under 20% using 116 or more trees. For 500 or more trees, the classification error decreases at a slower rate.

Examine the confusion matrix for each class as a percentage of the true class.

```
tic
Yfit = predict(rusTree,covtype(istest,:));
toc
```

```
Elapsed time is 132.353489 seconds.
```

```
confusionchart(Y(istest),Yfit,'Normalization','row-normalized','RowSummary','row-normalized')
```



All classes except class 2 have over 90% classification accuracy. But class 2 makes up close to half the data, so the overall accuracy is not that high.

Compact the ensemble

The ensemble is large. Remove the data using the `compact` method.

```
cmpctRus = compact(rusTree);

sz(1) = whos('rusTree');
sz(2) = whos('cmpctRus');
[sz(1).bytes sz(2).bytes]
```

```
ans = 1×2
10^9 ×

    1.6579    0.9423
```

The compacted ensemble is about half the size of the original.

Remove half the trees from `cmpctRus`. This action is likely to have minimal effect on the predictive performance, based on the observation that 500 out of 1000 trees give nearly optimal accuracy.

```
cmpctRus = removeLearners(cmpctRus,[500:1000]);

sz(3) = whos('cmpctRus');
sz(3).bytes
```

```
ans = 452868660
```

The reduced compact ensemble takes about a quarter of the memory of the full ensemble. Its overall loss rate is under 19%:

```
L = loss(cmpctRus,covtype(istest,:),Y(istest))
```

```
L = 0.1833
```

The predictive accuracy on new data might differ, because the ensemble accuracy might be biased. The bias arises because the same data used for assessing the ensemble was used for reducing the ensemble size. To obtain an unbiased estimate of requisite ensemble size, you should use cross validation. However, that procedure is time consuming.

References

[1] Blackard, J. A. and D. J. Dean. "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables". *Computers and Electronics in Agriculture* Vol. 24, Issue 3, 1999, pp. 131–151.

See Also

`compact` | `confusionchart` | `cvpartition` | `fitcensemble` | `loss` | `predict` | `removeLearners` | `tabulate` | `templateTree` | `test` | `training`

Related Examples

- “Surrogate Splits” on page 18-91
- “Ensemble Algorithms” on page 18-39
- “Test Ensemble Quality” on page 18-66
- “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84
- “LPBoost and TotalBoost for Small Ensembles” on page 18-96
- “Tune RobustBoost” on page 18-101

Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles

In many applications, you might prefer to treat classes in your data asymmetrically. For example, the data might have many more observations of one class than any other. Or misclassifying observations of one class has more severe consequences than misclassifying observations of another class. In such situations, you can either use the RUSBoost algorithm (specify `'Method'` as `'RUSBoost'`) or use the name-value pair argument `'Prior'` or `'Cost'` of `fitcensemble`.

If some classes are underrepresented or overrepresented in your training set, use either the `'Prior'` name-value pair argument or the RUSBoost algorithm. For example, suppose you obtain your training data by simulation. Because simulating class `A` is more expensive than simulating class `B`, you choose to generate fewer observations of class `A` and more observations of class `B`. The expectation, however, is that class `A` and class `B` are mixed in a different proportion in real (nonsimulated) situations. In this case, use `'Prior'` to set prior probabilities for class `A` and `B` approximately to the values you expect to observe in a real situation. The `fitcensemble` function normalizes prior probabilities to make them add up to `1`. Multiplying all prior probabilities by the same positive factor does not affect the result of classification. Another way to handle imbalanced data is to use the `RUSBoost` algorithm (`'Method','RUSBoost'`). You do not need to adjust the prior probabilities when using this algorithm. For details, see “Random Undersampling Boosting” on page 18-51 and “Classification with Imbalanced Data” on page 18-79.

If classes are adequately represented in the training data but you want to treat them asymmetrically, use the `'Cost'` name-value pair argument. Suppose you want to classify benign and malignant tumors in cancer patients. Failure to identify a malignant tumor (false negative) has far more severe consequences than misidentifying benign as malignant (false positive). You should assign high cost to misidentifying malignant as benign and low cost to misidentifying benign as malignant.

You must pass misclassification costs as a square matrix with nonnegative elements. Element `C(i,j)` of this matrix is the cost of classifying an observation into class `j` if the true class is `i`. The diagonal elements `C(i,i)` of the cost matrix must be `0`. For the previous example, you can choose malignant tumor to be class 1 and benign tumor to be class 2. Then you can set the cost matrix to

$$\begin{bmatrix} 0 & c \\ 1 & 0 \end{bmatrix}$$

where $c > 1$ is the cost of misidentifying a malignant tumor as benign. Costs are relative—multiplying all costs by the same positive factor does not affect the result of classification.

If you have only two classes, `fitcensemble` adjusts their prior probabilities using $\tilde{P}_i = C_{ij}P_i$ for class $i = 1,2$ and $j \neq i$. P_i are prior probabilities either passed into `fitcensemble` or computed from class frequencies in the training data, and $\tilde{P}_i$ are adjusted prior probabilities. Then `fitcensemble` uses the default cost matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and these adjusted probabilities for training its weak learners. Manipulating the cost matrix is thus equivalent to manipulating the prior probabilities.

If you have three or more classes, `fitcensemble` also converts input costs into adjusted prior probabilities. This conversion is more complex. First, `fitcensemble` attempts to solve a matrix

equation described in Zhou and Liu [1]. If it fails to find a solution, `fitcensemble` applies the "average cost" adjustment described in Breiman et al. [2]. For more information, see Zadrozny, Langford, and Abe [3].

Train Ensemble With Unequal Classification Costs

This example shows how to train an ensemble of classification trees with unequal classification costs. This example uses data on patients with hepatitis to see if they live or die as a result of the disease. The data set is described at UCI Machine Learning Data Repository.

Read the hepatitis data set from the UCI repository as a character array. Then convert the result to a cell array of character vectors using `textscan`. Specify a cell array of character vectors containing the variable names.

```
options = weboptions('ContentType','text');
hepatitis = textscan(webread(['http://archive.ics.uci.edu/ml/' ...
    'machine-learning-databases/hepatitis/hepatitis.data'],options),...
    '%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f%f','Delimiter',',',...
    'EndOfLine','\n','TreatAsEmpty','?');
size(hepatitis)
```

```
ans = 1×2

     1    20
```

```
VarNames = {'dieOrLive' 'age' 'sex' 'steroid' 'antivirals' 'fatigue' ...
    'malaise' 'anorexia' 'liverBig' 'liverFirm' 'spleen' ...
    'spiders' 'ascites' 'varices' 'bilirubin' 'alkPhosphate' 'sgot' ...
    'albumin' 'protime' 'histology'};
```

`hepatitis` is a 1-by-20 cell array of character vectors. The cells correspond to the response (`liveOrDie`) and 19 heterogeneous predictors.

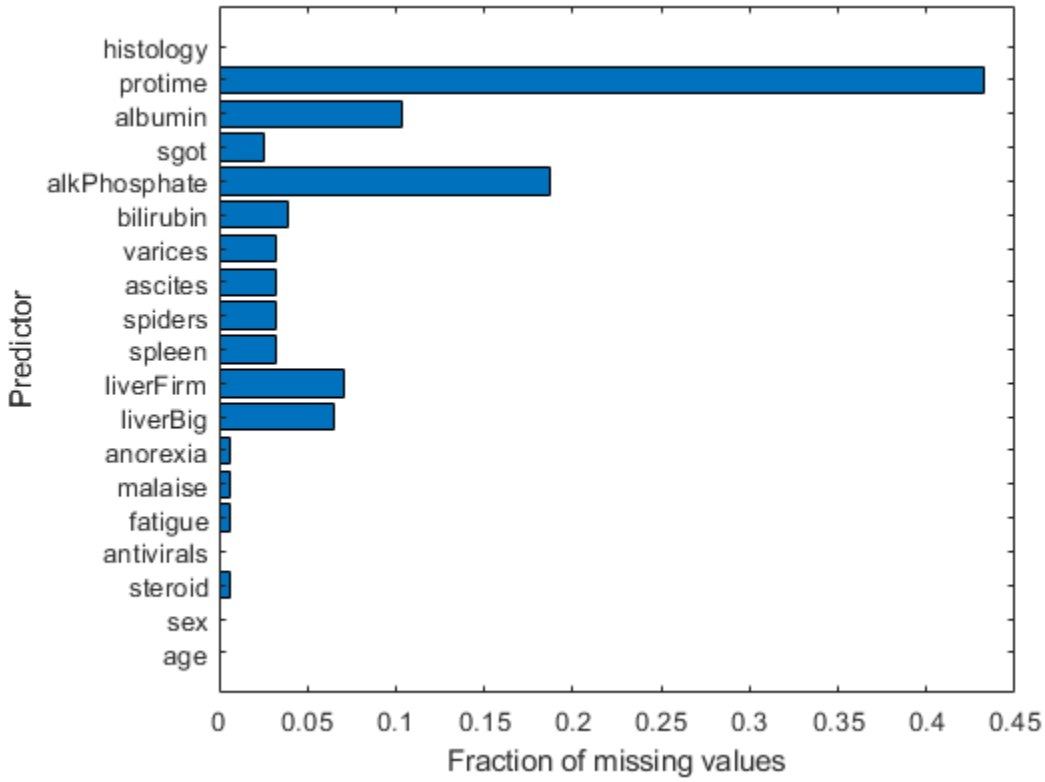
Specify a numeric matrix containing the predictors and a cell vector containing `'Die'` and `'Live'`, which are response categories. The response contains two values: `1` indicates that a patient died, and `2` indicates that a patient lived. Specify a cell array of character vectors for the response using the response categories. The first variable in `hepatitis` contains the response.

```
X = cell2mat(hepatitis(2:end));
ClassNames = {'Die' 'Live'};
Y = ClassNames(hepatitis{:,1});
```

`X` is a numeric matrix containing the 19 predictors. `Y` is a cell array of character vectors containing the response.

Inspect the data for missing values.

```
figure
barh(sum(isnan(X),1)/size(X,1))
h = gca;
h.YTick = 1:numel(VarNames) - 1;
h.YTickLabel = VarNames(2:end);
ylabel('Predictor')
xlabel('Fraction of missing values')
```



Most predictors have missing values, and one has nearly 45% of the missing values. Therefore, use decision trees with surrogate splits for better accuracy. Because the data set is small, training time with surrogate splits should be tolerable.

Create a classification tree template that uses surrogate splits.

```
rng(0,'twister') % For reproducibility
t = templateTree('surrogate','all');
```

Examine the data or the description of the data to see which predictors are categorical.

```
X(1:5,:)
```

```
ans = 5×19

   30.0000    2.0000    1.0000    2.0000    2.0000    2.0000    2.0000    1.0000    2.0000    2.0
   50.0000    1.0000    1.0000    2.0000    1.0000    2.0000    2.0000    1.0000    2.0000    2.0
   78.0000    1.0000    2.0000    2.0000    1.0000    2.0000    2.0000    2.0000    2.0000    2.0
   31.0000    1.0000       NaN    1.0000    2.0000    2.0000    2.0000    2.0000    2.0000    2.0
   34.0000    1.0000    2.0000    2.0000    2.0000    2.0000    2.0000    2.0000    2.0000    2.0
```

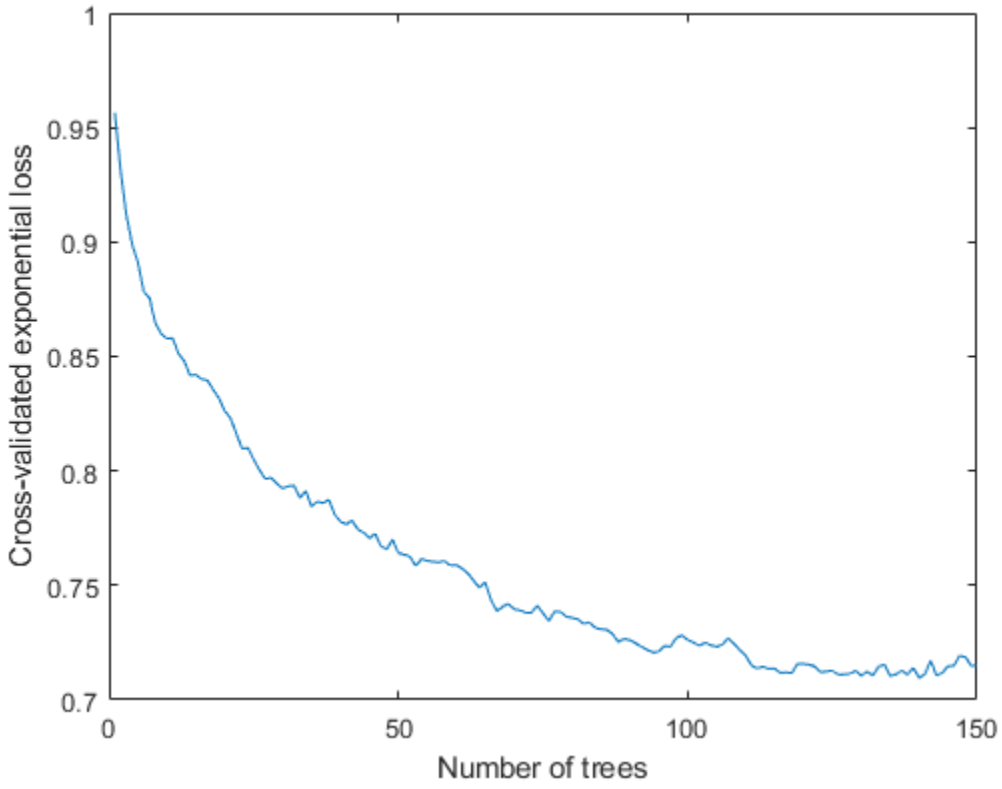
It appears that predictors 2 through 13 are categorical, as well as predictor 19. You can confirm this inference using the data set description at UCI Machine Learning Data Repository.

List the categorical variables.

```
catIdx = [2:13,19];
```

Create a cross-validated ensemble using 150 learners and the `GentleBoost` algorithm.

```
Ensemble = fitcensemble(X,Y,'Method','GentleBoost', ...
    'NumLearningCycles',150,'Learners',t,'PredictorNames',VarNames(2:end), ...
    'LearnRate',0.1,'CategoricalPredictors',catIdx,'KFold',5);
figure
plot(kfoldLoss(Ensemble,'Mode','cumulative','LossFun','exponential'))
xlabel('Number of trees')
ylabel('Cross-validated exponential loss')
```



Inspect the confusion matrix to see which patients the ensemble predicts correctly.

```
[yFit,sFit] = kfoldPredict(Ensemble);
confusionchart(Y,yFit)
```

Of the 123 patient who live, the ensemble predicts correctly that 112 will live. But for the 32 patients who die of hepatitis, the ensemble only predicts correctly that about half will die of hepatitis.

There are two types of error in the predictions of the ensemble:

- Predicting that the patient lives, but the patient dies
- Predicting that the patient dies, but the patient lives

Suppose you believe that the first error is five times worse than the second. Create a new classification cost matrix that reflects this belief.

```
cost.ClassNames = ClassNames;
cost.ClassificationCosts = [0 5; 1 0];
```

Create a new cross-validated ensemble using `cost` as the misclassification cost, and inspect the resulting confusion matrix.

```
EnsembleCost = fitcensemble(X,Y,'Method','GentleBoost', ...
    'NumLearningCycles',150,'Learners',t,'PredictorNames',VarNames(2:end), ...
    'LearnRate',0.1,'CategoricalPredictors',catIdx,'KFold',5,'Cost',cost);
[yFitCost,sFitCost] = kfoldPredict(EnsembleCost);
confusionchart(Y,yFitCost)
```

As expected, the new ensemble does a better job classifying the patients who die. Somewhat surprisingly, the new ensemble also does a better job classifying the patients who live, though the result is not statistically significantly better. The results of the cross validation are random, so this result is simply a statistical fluctuation. The result seems to indicate that the classification of patients who live is not very sensitive to the cost.

References

[1] Zhou, Z.-H. and X.-Y. Liu. “On Multi-Class Cost-Sensitive Learning.” *Computational Intelligence.* Vol. 26, Issue 3, 2010, pp. 232–257 CiteSeerX.

[2] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall, 1984.

[3] Zadrozny, B., J. Langford, and N. Abe. “Cost-Sensitive Learning by Cost-Proportionate Example Weighting.” *Third IEEE International Conference on Data Mining*, 435–442. 2003.

See Also

`confusionchart` | `fitcensemble` | `kfoldLoss` | `kfoldPredict` | `templateTree`

Related Examples

- “Surrogate Splits” on page 18-91

Surrogate Splits

When the value of the optimal split predictor for an observation is missing, if you specify to use surrogate splits, the software sends the observation to the left or right child node using the best surrogate predictor. When you have missing data, trees and ensembles of trees with surrogate splits give better predictions. This example shows how to improve the accuracy of predictions for data with missing values by using decision trees with surrogate splits.

Load Sample Data

Load the `ionosphere` data set.

```
load ionosphere
```

Partition the data set into training and test sets. Hold out 30% of the data for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'Holdout',0.3);
```

Identify the training and testing data.

```
Xtrain = X(training(cv),:);
Ytrain = Y(training(cv));
Xtest = X(test(cv),:);
Ytest = Y(test(cv));
```

Suppose half of the values in the test set are missing. Set half of the values in the test set to `NaN`.

```
Xtest(rand(size(Xtest))>0.5) = NaN;
```

Train Random Forest

Train a random forest of 150 classification trees without surrogate splits.

```
templ = templateTree('Reproducible',true);  % For reproducibility of random predictor selections
Mdl = fitcensemble(Xtrain,Ytrain,'Method','Bag','NumLearningCycles',150,'Learners',templ);
```

Create a decision tree template that uses surrogate splits. A tree using surrogate splits does not discard the entire observation when it includes missing data in some predictors.

```
templS = templateTree('Surrogate','On','Reproducible',true);
```

Train a random forest using the template `templS`.

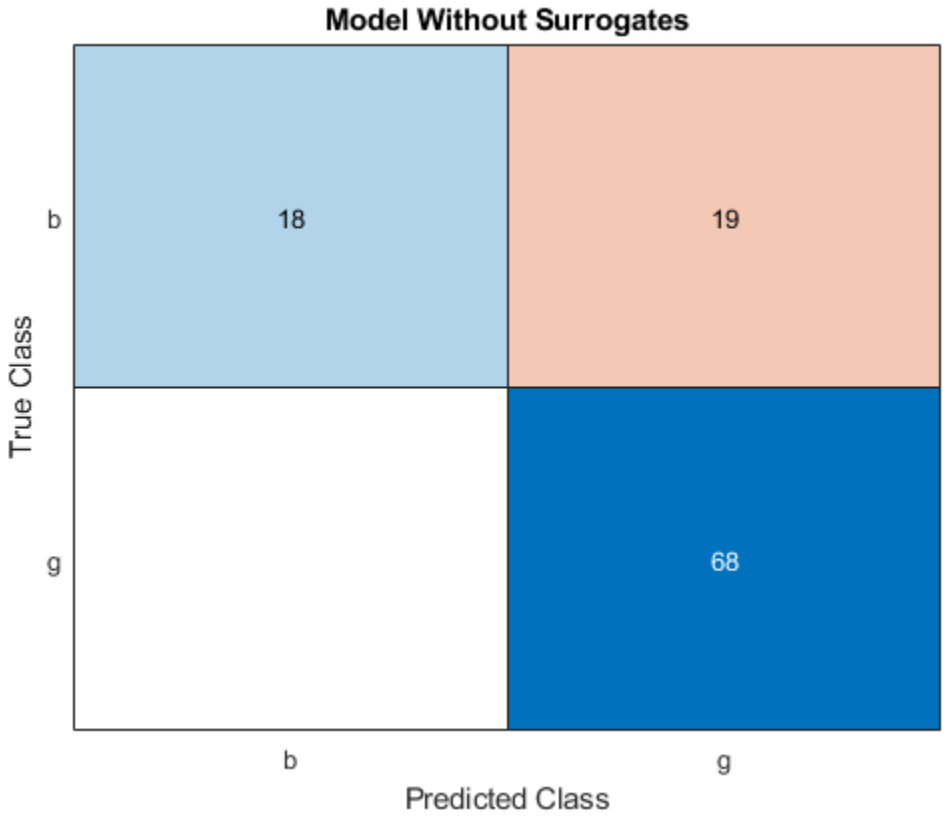
```
Mdls = fitcensemble(Xtrain,Ytrain,'Method','Bag','NumLearningCycles',150,'Learners',templS);
```

Test Accuracy

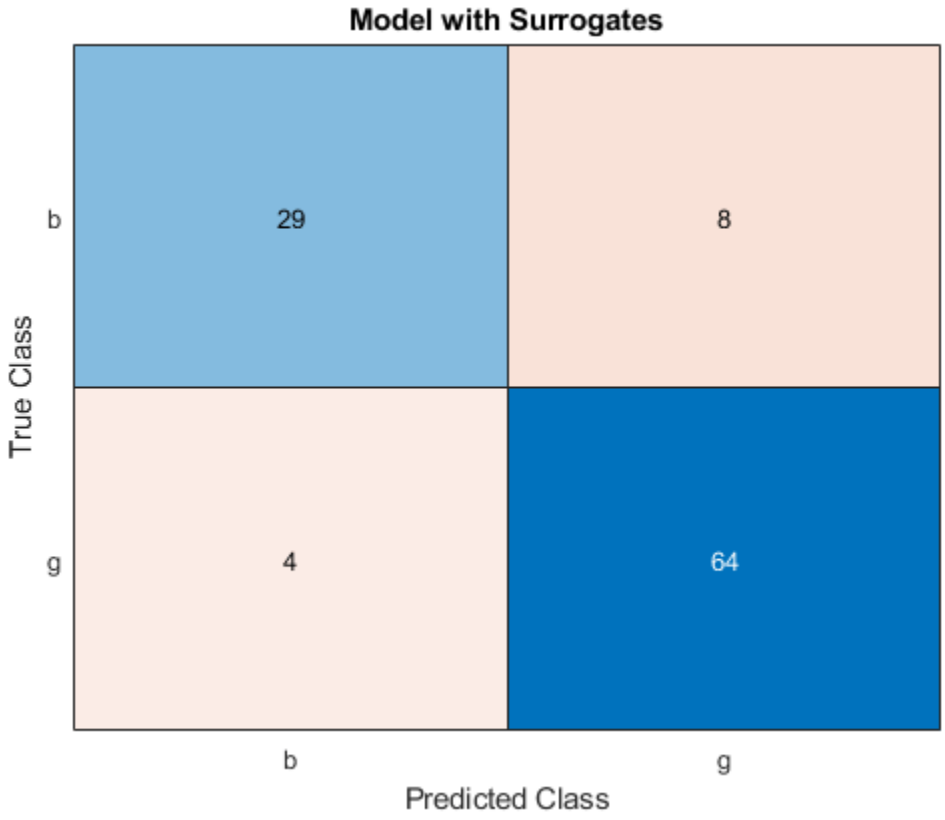
Test the accuracy of predictions with and without surrogate splits.

Predict responses and create confusion matrix charts using both approaches.

```
Ytest_pred = predict(Mdl,Xtest);
figure
cm = confusionchart(Ytest,Ytest_pred);
cm.Title = 'Model Without Surrogates';
```



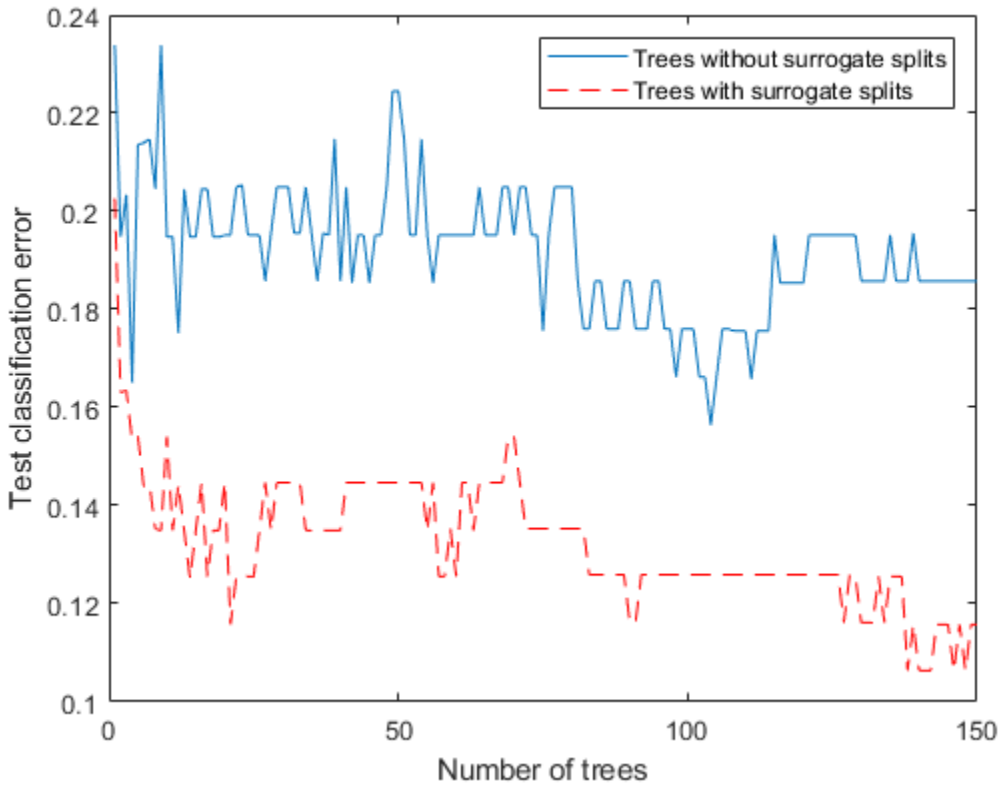

```
Ytest_preds = predict(Mdls,Xtest);
figure
cms = confusionchart(Ytest,Ytest_preds);
cms.Title = 'Model with Surrogates';
```



All off-diagonal elements on the confusion matrix represent misclassified data. A good classifier yields a confusion matrix that looks dominantly diagonal. In this case, the classification error is lower for the model trained with surrogate splits.

Estimate cumulative classification errors. Specify `'Mode','Cumulative'` when estimating classification errors by using the `loss` function. The `loss` function returns a vector in which element `J` indicates the error using the first `J` learners.

```
figure
plot(loss(Mdl,Xtest,Ytest,'Mode','Cumulative'))
hold on
plot(loss(Mdls,Xtest,Ytest,'Mode','Cumulative'),'r--')
legend('Trees without surrogate splits','Trees with surrogate splits')
xlabel('Number of trees')
ylabel('Test classification error')
```



The error value decreases as the number of trees increases, which indicates good performance. The classification error is lower for the model trained with surrogate splits.

Check the statistical significance of the difference in results with by using `compareHoldout`. This function uses the McNemar test.

```
[~,p] = compareHoldout(Mdls,Mdl,Xtest,Xtest,Ytest,'Alternative','greater')
```

```
p = 0.0384
```

The low *p*-value indicates that the ensemble with surrogate splits is better in a statistically significant manner.

Estimate Predictor Importance

Predictor importance estimates can vary depending on whether or not a tree uses surrogate splits. Estimate predictor importance measures by permuting out-of-bag observations. Then, find the five most important predictors.

```
imp = oobPermutedPredictorImportance(Mdl);
[~,ind] = maxk(imp,5)
```

```
ind = 1×5

     5     3    27     8    14
```

```
imps = oobPermutedPredictorImportance(Mdls);
[~,inds] = maxk(imps,5)
```

```
inds = 1×5

     3     5     8    27     7
```

After estimating predictor importance, you can exclude unimportant predictors and train a model again. Eliminating unimportant predictors saves time and memory for predictions, and makes predictions easier to understand.

If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` of the `templateTree` function as `'all'` for the tree learners of the ensemble. Otherwise, the software might not select some predictors, underestimating their importance. For an example, see “Select Predictors for Random Forests” on page 18-60.

See Also

`compareHoldout` | `fitcensemble` | `fitrensemble`

Related Examples

- “Ensemble Algorithms” on page 18-39
- “Test Ensemble Quality” on page 18-66
- “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84
- “Classification with Imbalanced Data” on page 18-79
- “LPBoost and TotalBoost for Small Ensembles” on page 18-96
- “Tune RobustBoost” on page 18-101

LPBoost and TotalBoost for Small Ensembles

This example shows how to obtain the benefits of the `LPBoost` and `TotalBoost` algorithms. These algorithms share two beneficial characteristics:

- They are self-terminating, which means you do not have to figure out how many members to include.
- They produce ensembles with some very small weights, enabling you to safely remove ensemble members.

Load the data

Load the `ionosphere` data set.

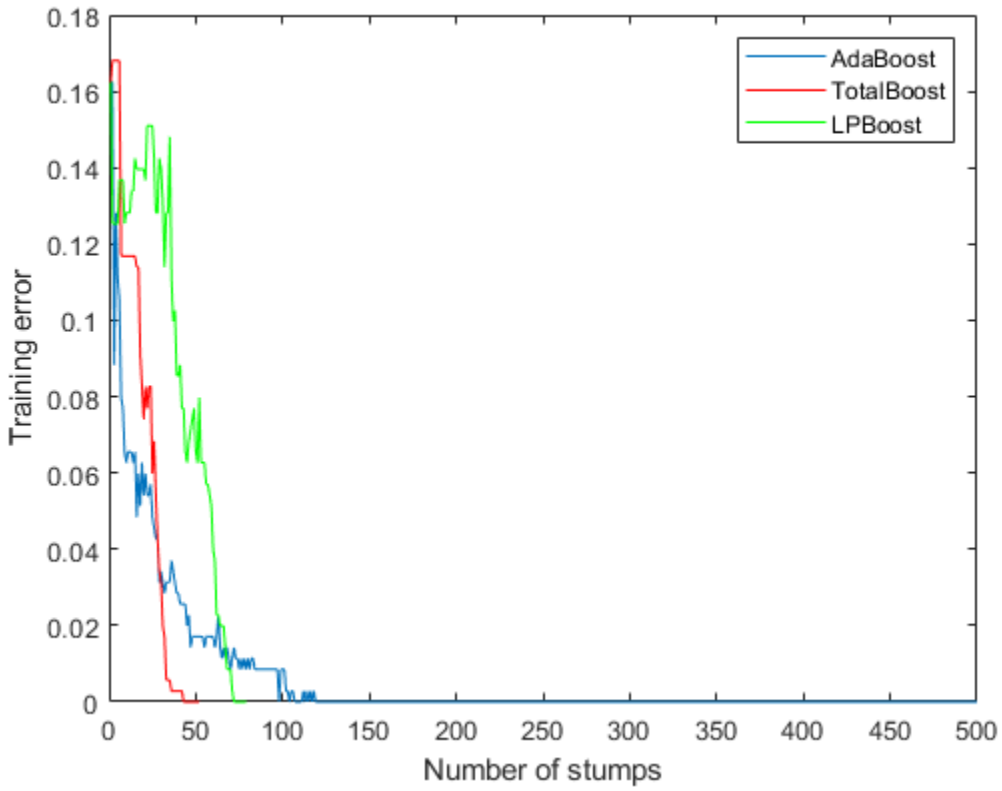
```
load ionosphere
```

Create the classification ensembles

Create ensembles for classifying the `ionosphere` data using the `LPBoost`, `TotalBoost`, and, for comparison, `AdaBoostM1` algorithms. It is hard to know how many members to include in an ensemble. For `LPBoost` and `TotalBoost`, try using `500`. For comparison, also use `500` for `AdaBoostM1`.

The default weak learners for boosting methods are decision trees with the `MaxNumSplits` property set to `10`. These trees tend to fit better than tree stumps (with `1` maximum split) and may overfit more. Therefore, to prevent overfitting, use tree stumps as weak learners for the ensembles.

```
rng('default') % For reproducibility
T = 500;
treeStump = templateTree('MaxNumSplits',1);
adaStump = fitcensemble(X,Y,'Method','AdaBoostM1','NumLearningCycles',T,'Learners',treeStump);
totalStump = fitcensemble(X,Y,'Method','TotalBoost','NumLearningCycles',T,'Learners',treeStump);
lpStump = fitcensemble(X,Y,'Method','LPBoost','NumLearningCycles',T,'Learners',treeStump);

figure
plot(resubLoss(adaStump,'Mode','Cumulative'));
hold on
plot(resubLoss(totalStump,'Mode','Cumulative'),'r');
plot(resubLoss(lpStump,'Mode','Cumulative'),'g');
hold off
xlabel('Number of stumps');
ylabel('Training error');
legend('AdaBoost','TotalBoost','LPBoost','Location','NE');
```



All three algorithms achieve perfect prediction on the training data after a while.

Examine the number of members in all three ensembles.

```
[adaStump.NTrained totalStump.NTrained lpStump.NTrained]
```

```
ans = 1×3

   500    52    79
```

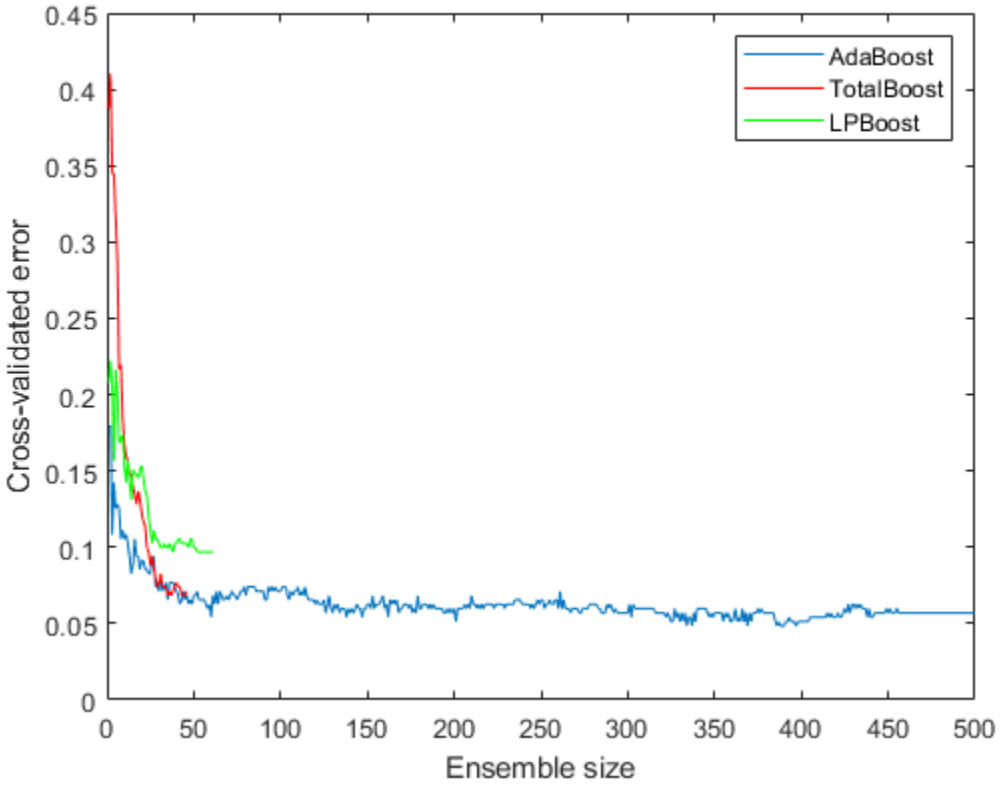
`AdaBoostM1` trained all `500` members. The other two algorithms stopped training early.

Cross validate the ensembles

Cross validate the ensembles to better determine ensemble accuracy.

```
cvlp = crossval(lpStump,'KFold',5);
cvtotal = crossval(totalStump,'KFold',5);
cvada = crossval(adaStump,'KFold',5);

figure
plot(kfoldLoss(cvada,'Mode','Cumulative'));
hold on
plot(kfoldLoss(cvtotal,'Mode','Cumulative'),'r');
plot(kfoldLoss(cvlp,'Mode','Cumulative'),'g');
hold off
xlabel('Ensemble size');
```

```
ylabel('Cross-validated error');
legend('AdaBoost','TotalBoost','LPBoost','Location','NE');
```

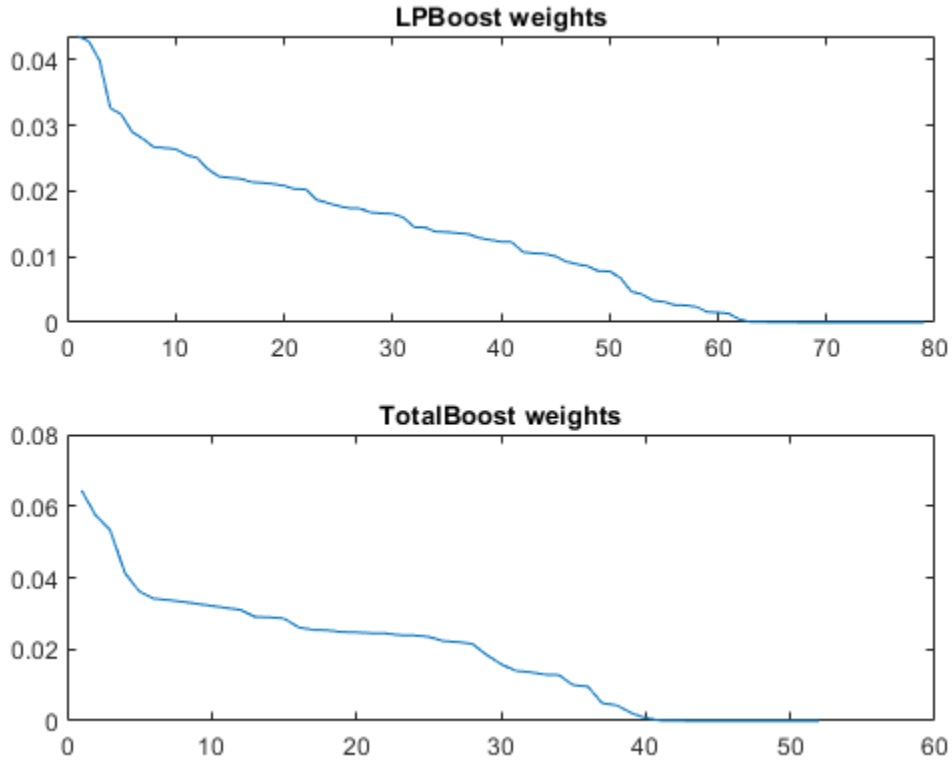


The results show that each boosting algorithm achieves a loss of 10% or lower with 50 ensemble members.

Compact and remove ensemble members

To reduce the ensemble sizes, compact them, and then use `removeLearners`. The question is, how many learners should you remove? The cross-validated loss curves give you one measure. For another, examine the learner weights for `LPBoost` and `TotalBoost` after compacting.

```
cada = compact(adaStump);
clp = compact(lpStump);
ctotal = compact(totalStump);

figure
subplot(2,1,1)
plot(clp.TrainedWeights)
title('LPBoost weights')
subplot(2,1,2)
plot(ctotal.TrainedWeights)
title('TotalBoost weights')
```



Both `LPBoost` and `TotalBoost` show clear points where the ensemble member weights become negligible.

Remove the unimportant ensemble members.

```
cada = removeLearners(cada,150:cada.NTrained);
clp = removeLearners(clp,60:clp.NTrained);
ctotal = removeLearners(ctotal,40:ctotal.NTrained);
```

Check that removing these learners does not affect ensemble accuracy on the training data.

```
[loss(cada,X,Y) loss(clp,X,Y) loss(ctotal,X,Y)]
```

```
ans = 1×3

     0     0     0
```

Check the resulting compact ensemble sizes.

```
s(1) = whos('cada');
s(2) = whos('clp');
s(3) = whos('ctotal');
s.bytes
```

```
ans = 590844

ans = 236030
```

```
ans = 157190
```

The sizes of the compact ensembles are approximately proportional to the number of members in each.

See Also

`compact` | `crossval` | `fitcensemble` | `kfoldLoss` | `loss` | `removeLearners` | `resubLoss`

Related Examples

- “Surrogate Splits” on page 18-91
- “Ensemble Algorithms” on page 18-39
- “Test Ensemble Quality” on page 18-66
- “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84
- “Classification with Imbalanced Data” on page 18-79
- “Tune RobustBoost” on page 18-101

Tune RobustBoost

The `RobustBoost` algorithm can make good classification predictions even when the training data has noise. However, the default `RobustBoost` parameters can produce an ensemble that does not predict well. This example shows one way of tuning the parameters for better predictive accuracy.

Generate data with label noise. This example has twenty uniform random numbers per observation, and classifies the observation as `1` if the sum of the first five numbers exceeds 2.5 (so is larger than average), and `0` otherwise:

```
rng(0,'twister') % for reproducibility
Xtrain = rand(2000,20);
Ytrain = sum(Xtrain(:,1:5),2) > 2.5;
```

To add noise, randomly switch 10% of the classifications:

```
idx = randsample(2000,200);
Ytrain(idx) = ~Ytrain(idx);
```

Create an ensemble with `AdaBoostM1` for comparison purposes:

```
ada = fitcensemble(Xtrain,Ytrain,'Method','AdaBoostM1', ...
    'NumLearningCycles',300,'Learners','Tree','LearnRate',0.1);
```

Create an ensemble with `RobustBoost`. Because the data has 10% incorrect classification, perhaps an error goal of 15% is reasonable.

```
rb1 = fitcensemble(Xtrain,Ytrain,'Method','RobustBoost', ...
    'NumLearningCycles',300,'Learners','Tree','RobustErrorGoal',0.15, ...
    'RobustMaxMargin',1);
```

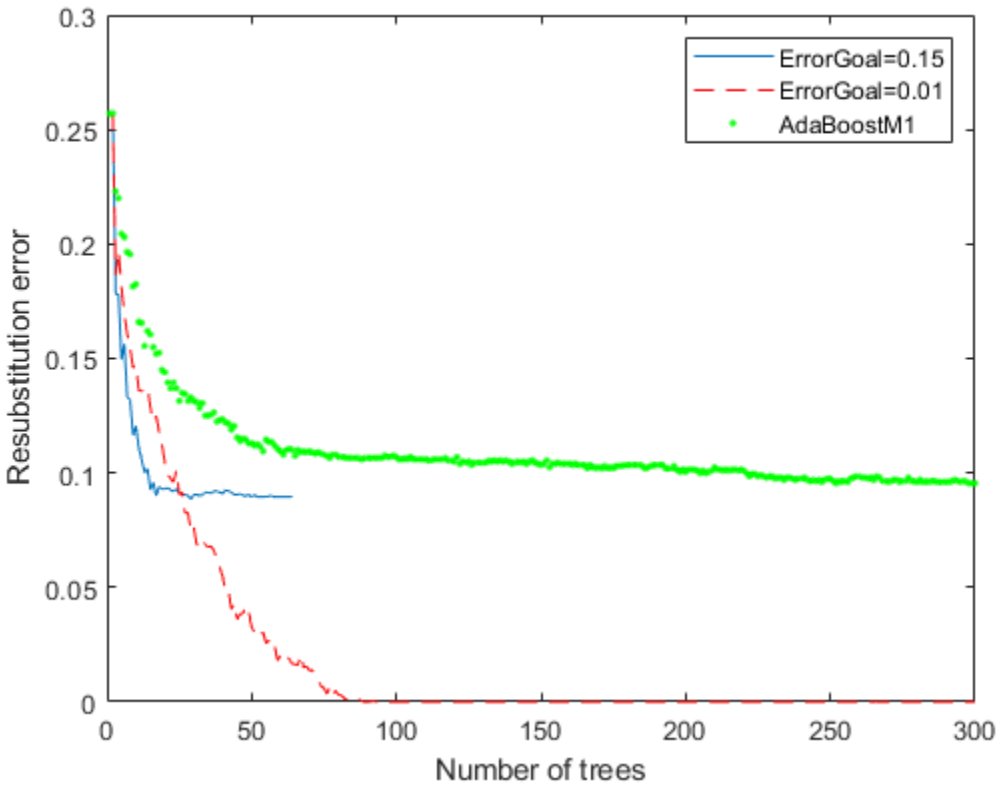
Note that if you set the error goal to a high enough value, then the software returns an error.

Create an ensemble with very optimistic error goal, `0.01`:

```
rb2 = fitcensemble(Xtrain,Ytrain,'Method','RobustBoost', ...
    'NumLearningCycles',300,'Learners','Tree','RobustErrorGoal',0.01);
```

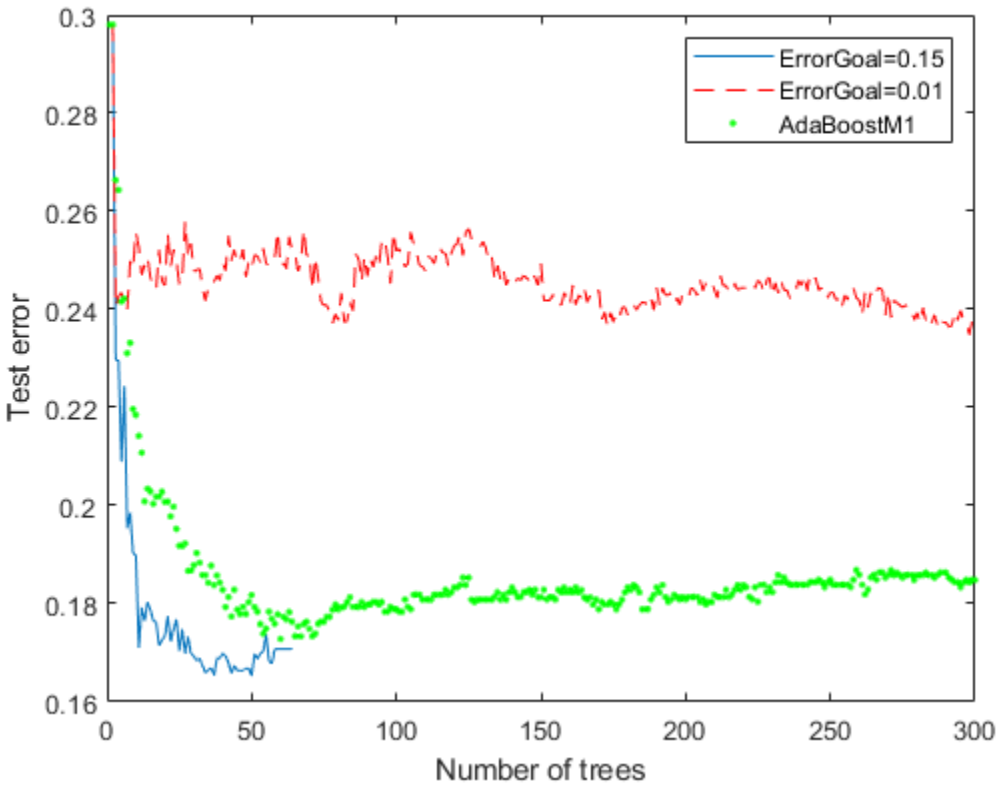
Compare the resubstitution error of the three ensembles:

```
figure
plot(resubLoss(rb1,'Mode','Cumulative'));
hold on
plot(resubLoss(rb2,'Mode','Cumulative'),'r--');
plot(resubLoss(ada,'Mode','Cumulative'),'g.');
hold off;
xlabel('Number of trees');
ylabel('Resubstitution error');
legend('ErrorGoal=0.15','ErrorGoal=0.01',...
    'AdaBoostM1','Location','NE');
```



All the `RobustBoost` curves show lower resubstitution error than the `AdaBoostM1` curve. The error goal of `0.01` curve shows the lowest resubstitution error over most of the range.

```
Xtest = rand(2000,20);
Ytest = sum(Xtest(:,1:5),2) > 2.5;
idx = randsample(2000,200);
Ytest(idx) = ~Ytest(idx);
figure;
plot(loss(rb1,Xtest,Ytest,'Mode','Cumulative'));
hold on
plot(loss(rb2,Xtest,Ytest,'Mode','Cumulative'),'r--');
plot(loss(ada,Xtest,Ytest,'Mode','Cumulative'),'g.');
hold off;
xlabel('Number of trees');
ylabel('Test error');
legend('ErrorGoal=0.15','ErrorGoal=0.01',...
    'AdaBoostM1','Location','NE');
```



The error curve for error goal 0.15 is lowest (best) in the plotted range. `AdaBoostM1` has higher error than the curve for error goal 0.15. The curve for the too-optimistic error goal 0.01 remains substantially higher (worse) than the other algorithms for most of the plotted range.

See Also

`fitcensemble` | `loss` | `resubLoss`

Related Examples

- “Surrogate Splits” on page 18-91
- “Ensemble Algorithms” on page 18-39
- “Test Ensemble Quality” on page 18-66
- “Handle Imbalanced Data or Unequal Misclassification Costs in Classification Ensembles” on page 18-84
- “Classification with Imbalanced Data” on page 18-79
- “LPBoost and TotalBoost for Small Ensembles” on page 18-96

Random Subspace Classification

This example shows how to use a random subspace ensemble to increase the accuracy of classification. It also shows how to use cross validation to determine good parameters for both the weak learner template and the ensemble.

Load the data

Load the `ionosphere` data. This data has 351 binary responses to 34 predictors.

```
load ionosphere;
[N,D] = size(X)
```

```
N = 351
```

```
D = 34
```

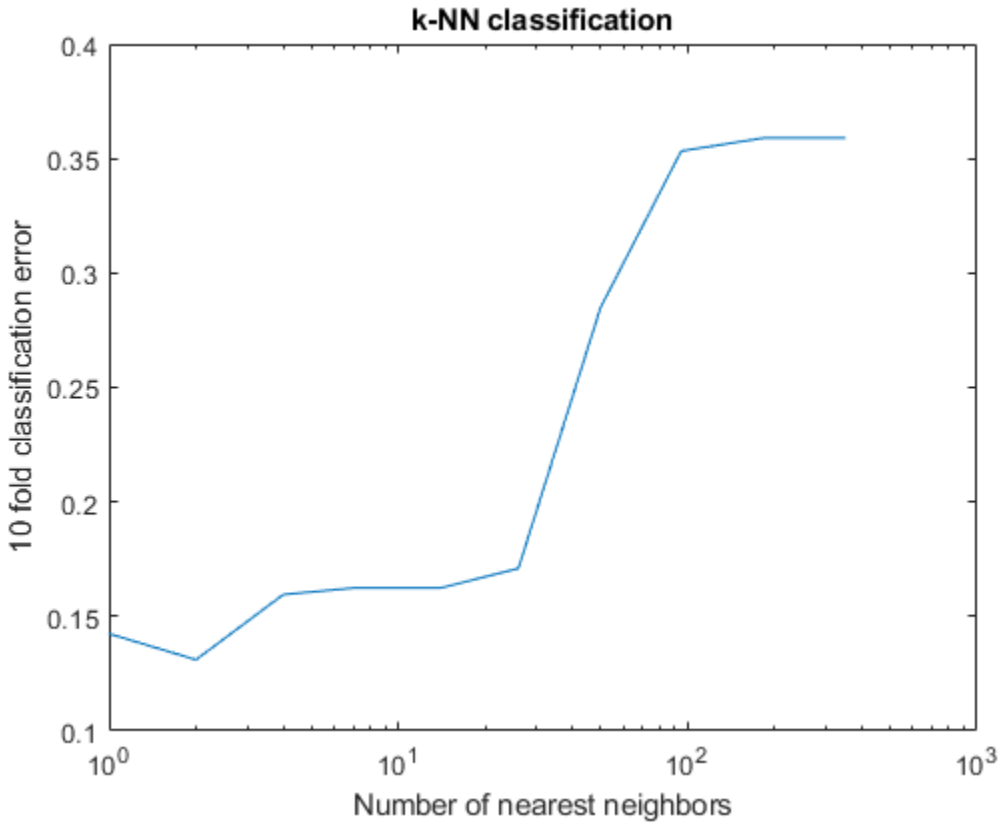
```
resp = unique(Y)
```

```
resp = 2x1 cell
    {'b'}
    {'g'}
```

Choose the number of nearest neighbors

Find a good choice for `k`, the number of nearest neighbors in the classifier, by cross validation. Choose the number of neighbors approximately evenly spaced on a logarithmic scale.

```
rng(8000,'twister') % for reproducibility
K = round(logspace(0,log10(N),10)); % number of neighbors
cvloss = zeros(numel(K),1);
for k=1:numel(K)
    knn = fitcknn(X,Y,...
        'NumNeighbors',K(k),'CrossVal','On');
    cvloss(k) = kfoldLoss(knn);
end
figure; % Plot the accuracy versus k
semilogx(K,cvloss);
xlabel('Number of nearest neighbors');
ylabel('10 fold classification error');
title('k-NN classification');
```



The lowest cross-validation error occurs for `k = 2`.

Create the ensembles

Create ensembles for 2-nearest neighbor classification with various numbers of dimensions, and examine the cross-validated loss of the resulting ensembles.

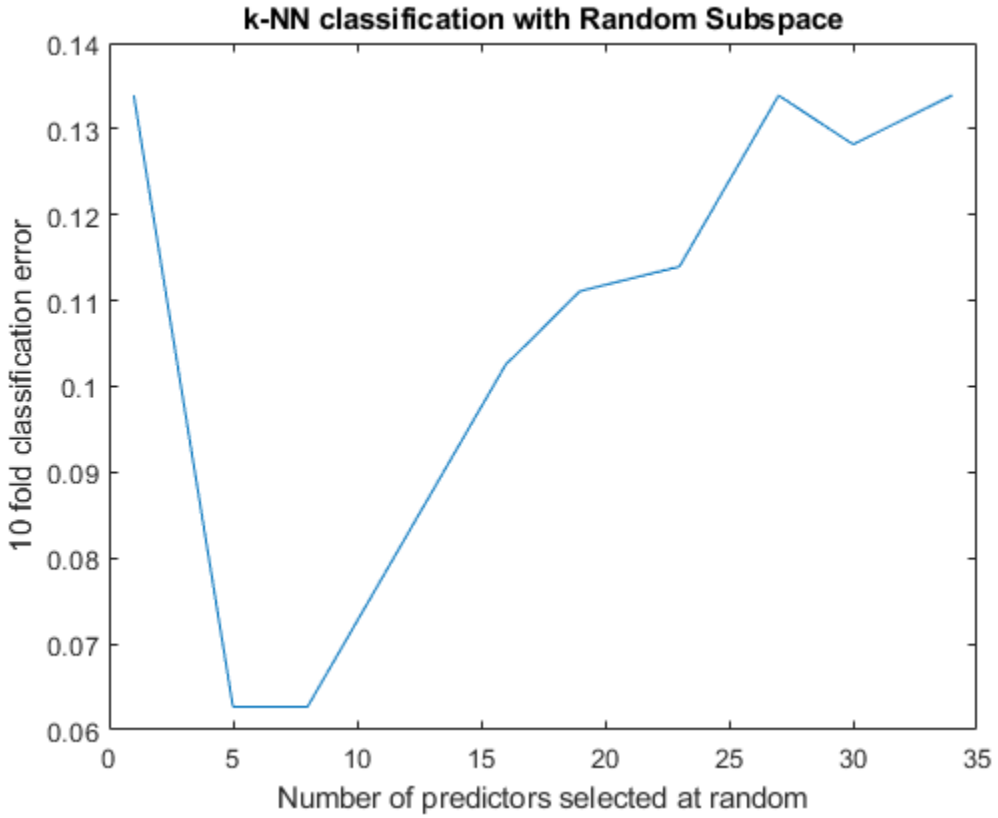
This step takes a long time. To keep track of the progress, print a message as each dimension finishes.

```
NPredToSample = round(linspace(1,D,10)); % linear spacing of dimensions
cvloss = zeros(numel(NPredToSample),1);
learner = templateKNN('NumNeighbors',2);
for npred=1:numel(NPredToSample)
   subspace = fitcensemble(X,Y,'Method','Subspace','Learners',learner, ...
       'NPredToSample',NPredToSample(npred),'CrossVal','On');
   cvloss(npred) = kfoldLoss(subspace);
   fprintf('Random Subspace %i done.\n',npred);
end
```

```
Random Subspace 1 done.
Random Subspace 2 done.
Random Subspace 3 done.
Random Subspace 4 done.
Random Subspace 5 done.
Random Subspace 6 done.
Random Subspace 7 done.
```

```
Random Subspace 8 done.
Random Subspace 9 done.
Random Subspace 10 done.
```

```
figure; % plot the accuracy versus dimension
plot(NPredToSample,cvloss);
xlabel('Number of predictors selected at random');
ylabel('10 fold classification error');
title('k-NN classification with Random Subspace');
```

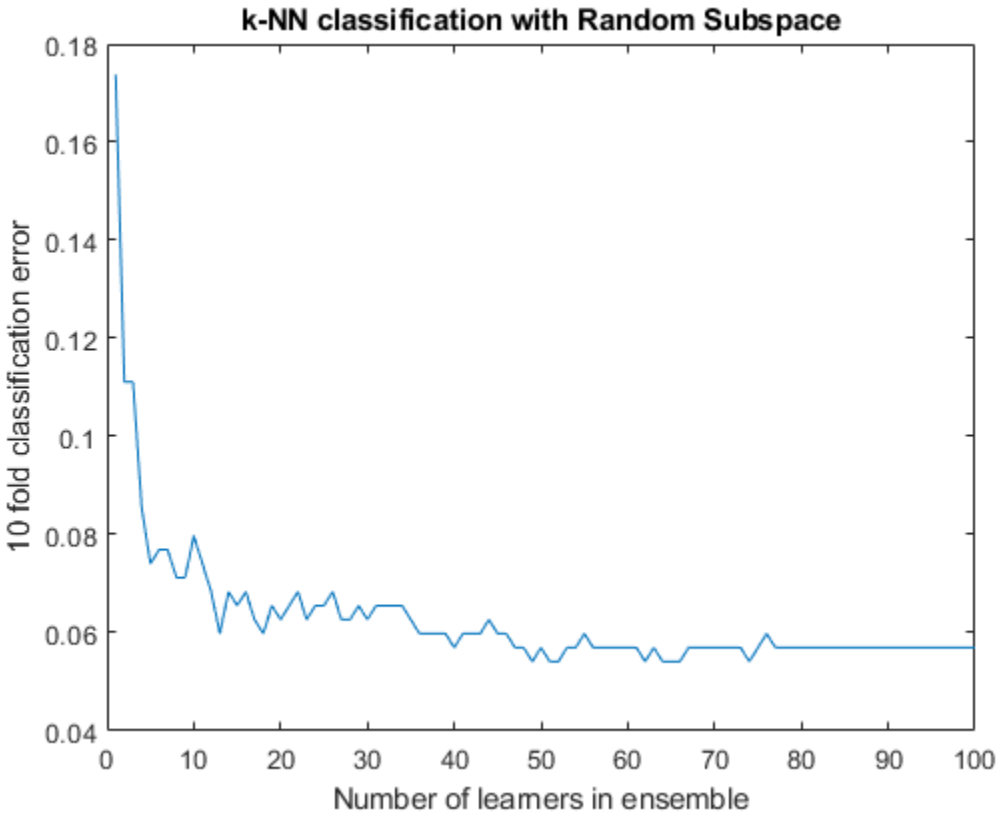


The ensembles that use five and eight predictors per learner have the lowest cross-validated error. The error rate for these ensembles is about 0.06, while the other ensembles have cross-validated error rates that are approximately 0.1 or more.

Find a good ensemble size

Find the smallest number of learners in the ensemble that still give good classification.

```
ens = fitcensemble(X,Y,'Method','Subspace','Learners',learner, ...
       'NPredToSample',5,'CrossVal','on');
figure; % Plot the accuracy versus number in ensemble
plot(kfoldLoss(ens,'Mode','Cumulative'))
xlabel('Number of learners in ensemble');
ylabel('10 fold classification error');
title('k-NN classification with Random Subspace');
```



There seems to be no advantage in an ensemble with more than 50 or so learners. It is possible that 25 learners gives good predictions.

Create a final ensemble

Construct a final ensemble with 50 learners. Compact the ensemble and see if the compacted version saves an appreciable amount of memory.

```
ens = fitcensemble(X,Y,'Method','Subspace','NumLearningCycles',50,...
    'Learners',learner,'NPredToSample',5);
cens = compact(ens);
s1 = whos('ens');
s2 = whos('cens');
[s1.bytes s2.bytes] % si.bytes = size in bytes
```

```
ans = 1×2

     1748467     1518820
```

The compact ensemble is about 10% smaller than the full ensemble. Both give the same predictions.

See Also

`compact` | `fitcensemble` | `fitcknn` | `kfoldLoss` | `templateKNN`

Related Examples

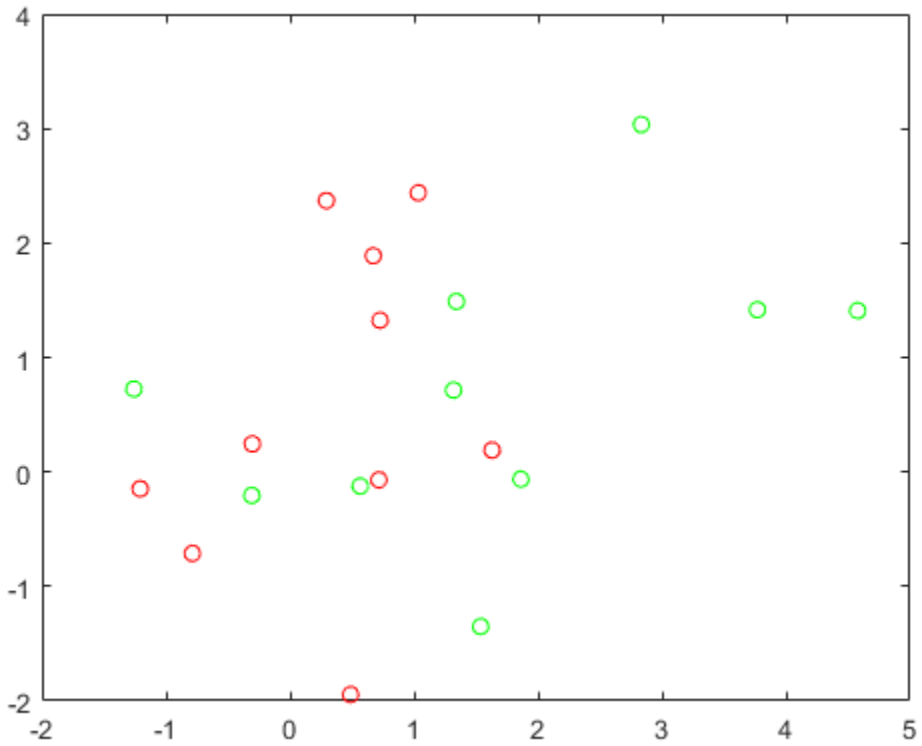
- “Framework for Ensemble Learning” on page 18-31
- “Ensemble Algorithms” on page 18-39
- “Train Classification Ensemble” on page 18-54
- “Test Ensemble Quality” on page 18-66

Train Classification Ensemble in Parallel

This example shows how to train a classification ensemble in parallel. The model has ten red and ten green base locations, and red and green populations that are normally distributed and centered at the base locations. The objective is to classify points based on their locations. These classifications are ambiguous because some base locations are near the locations of the other color.

Create and plot ten base locations of each color.

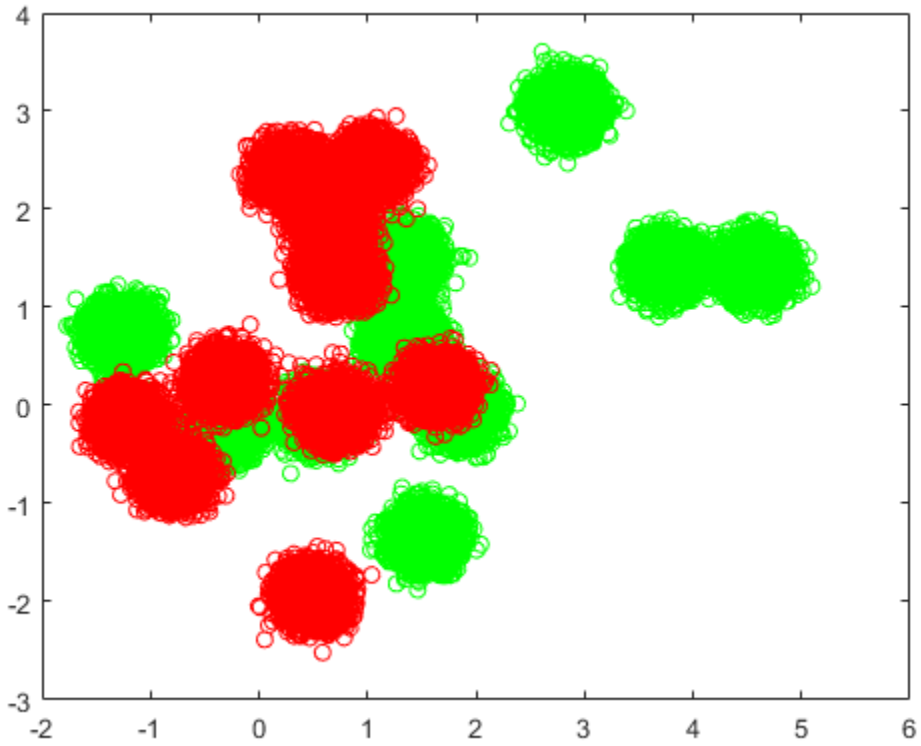
```
rng default % For reproducibility
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
plot(grnpop(:,1),grnpop(:,2),'go')
hold on
plot(redpop(:,1),redpop(:,2),'ro')
hold off
```



Create 40,000 points of each color centered on random base points.

```
N = 40000;
redpts = zeros(N,2);grnpts = redpts;
for i = 1:N
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
figure
plot(grnpts(:,1),grnpts(:,2),'go')
```

```
hold on
plot(redpts(:,1),redpts(:,2),'ro')
hold off
```




```
cdata = [grnpts;redpts];
grp = ones(2*N,1);
% Green label 1, red label -1
grp(N+1:2*N) = -1;
```

Fit a bagged classification ensemble to the data. For comparison with parallel training, fit the ensemble in serial and return the training time.

```
tic
mdl = fitcensemble(cdata,grp,'Method','Bag');
stime = toc
```

```
stime = 9.0782
```

Evaluate the out-of-bag loss for the fitted model.

```
myerr = oobLoss(mdl)
```

```
myerr = 0.0572
```

Create a bagged classification model in parallel, using a reproducible tree template and parallel substreams. You can create a parallel pool on a cluster or a parallel pool of thread workers on your local machine. To choose the appropriate parallel environment, see Choose Between Thread-Based and Process-Based Environments.

```
parpool
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

ans =

 ProcessPool with properties:

            Connected: true
           NumWorkers: 6
              Cluster: local
        AttachedFiles: {}
    AutoAddClientPath: true
          IdleTimeout: 30 minutes (30 minutes remaining)
          SpmdEnabled: true
```

```
s = RandStream('mrg32k3a');
options = statset("UseParallel",true,"UseSubstreams",true,"Streams",s);
t = templateTree("Reproducible",true);
tic
mdl2 = fitcensemble(cdata,grp,'Method','Bag','Learners',t,'Options',options);
ptime = toc
```

```
ptime = 6.2527
```

On this six-core system, the training process in parallel is faster.

```
speedup = stime/ptime
```

```
speedup = 1.4519
```

Evaluate the out-of-bag loss for this model.

```
myerr2 = oobLoss(mdl2)
```

```
myerr2 = 0.0577
```

The error rate is similar to the rate of the first model.

To demonstrate the reproducibility of the model, reset the random number stream and fit the model again.

```
reset(s);
tic
mdl2 = fitcensemble(cdata,grp,'Method','Bag','Learners',t,'Options',options);
toc
```

```
Elapsed time is 3.953355 seconds.
```

Check that the loss is the same as the previous loss.

```
myerr2 = oobLoss(mdl2)
```

```
myerr2 = 0.0577
```

See Also

`fitcensemble` | `fitrensemble`

Related Examples

- “Classification Ensembles”
- “Regression Tree Ensembles”

Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger

Statistics and Machine Learning Toolbox™ offers two objects that support bootstrap aggregation (bagging) of regression trees: `TreeBagger` created by using `TreeBagger` and `RegressionBaggedEnsemble` created by using `fitrensemble`. See “Comparison of TreeBagger and Bagged Ensembles” on page 18-44 for differences between `TreeBagger` and `RegressionBaggedEnsemble`.

This example shows the workflow for classification using the features in `TreeBagger` only.

Use a database of 1985 car imports with 205 observations, 25 predictors, and 1 response, which is insurance risk rating, or "symboling." The first 15 variables are numeric and the last 10 are categorical. The symboling index takes integer values from -3 to 3.

Load the data set and split it into predictor and response arrays.

```
load imports-85
Y = X(:,1);
X = X(:,2:end);
isCategorical = [zeros(15,1);ones(size(X,2)-15,1)]; % Categorical variable flag
```

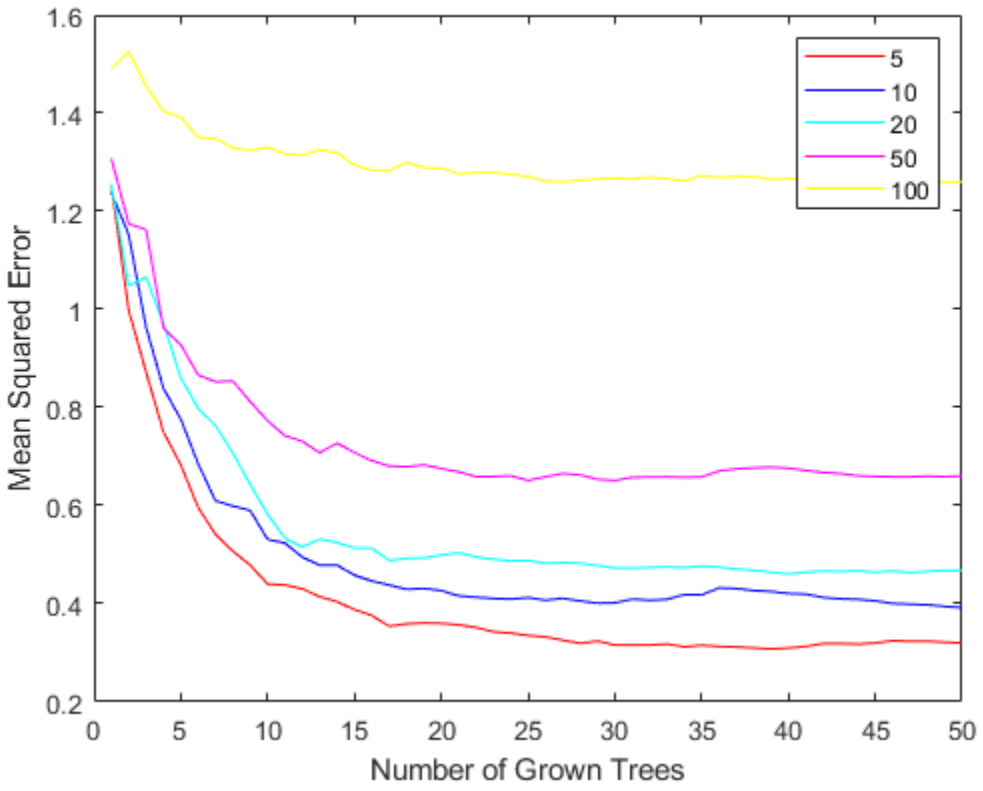
Because bagging uses randomized data drawings, its exact outcome depends on the initial random seed. To reproduce the results in this example, use the random stream settings.

```
rng(1945,'twister')
```

Finding the Optimal Leaf Size

For regression, the general rule is to the set leaf size to 5 and select one third of the input features for decision splits at random. In the following step, verify the optimal leaf size by comparing mean squared errors obtained by regression for various leaf sizes. `oobError` computes MSE versus the number of grown trees. You must set `OOBPred` to `'On'` to obtain out-of-bag predictions later.

```
leaf = [5 10 20 50 100];
col = 'rbcmy';
figure
for i=1:length(leaf)
    b = TreeBagger(50,X,Y,'Method','R','OOBPrediction','On',...
            'CategoricalPredictors',find(isCategorical == 1),...
            'MinLeafSize',leaf(i));
    plot(oobError(b),col(i))
    hold on
end
xlabel('Number of Grown Trees')
ylabel('Mean Squared Error')
legend({'5' '10' '20' '50' '100'},'Location','NorthEast')
hold off
```



The red curve (leaf size 5) yields the lowest MSE values.

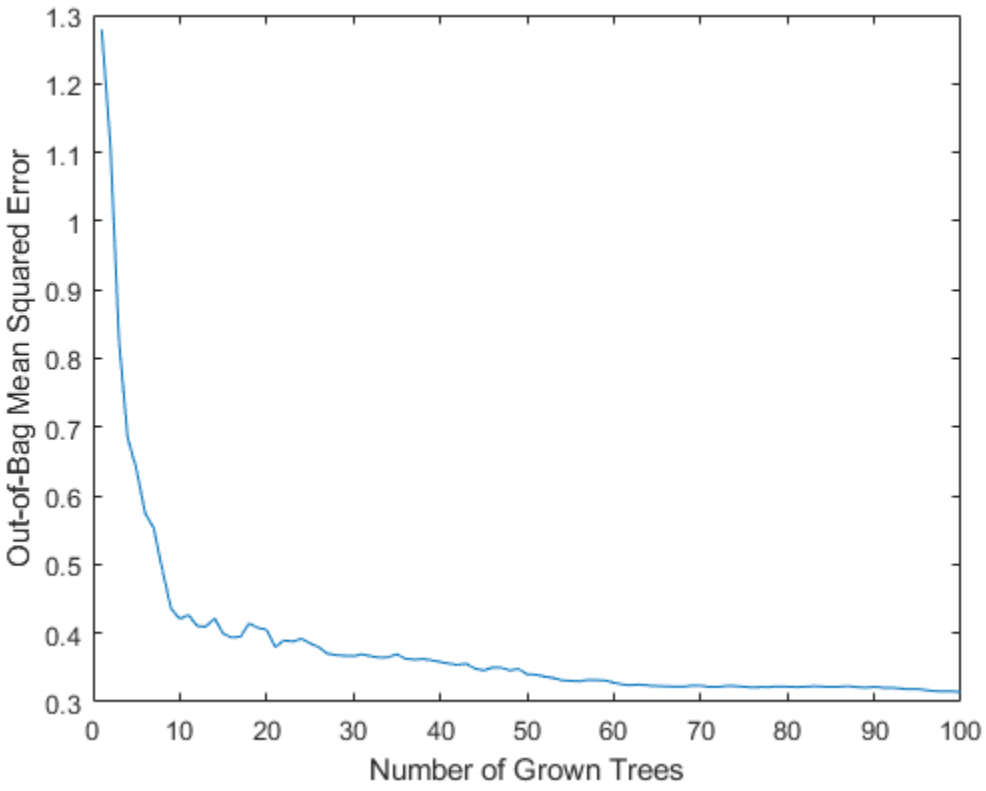
Estimating Feature Importance

In practical applications, you typically grow ensembles with hundreds of trees. For example, the previous code block uses 50 trees for faster processing. Now that you have estimated the optimal leaf size, grow a larger ensemble with 100 trees and use it to estimate feature importance.

```
b = TreeBagger(100,X,Y,'Method','R','OOBPredictorImportance','On',...
    'CategoricalPredictors',find(isCategorical == 1),...
    'MinLeafSize',5);
```

Inspect the error curve again to make sure nothing went wrong during training.

```
figure
plot(oobError(b))
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Mean Squared Error')
```

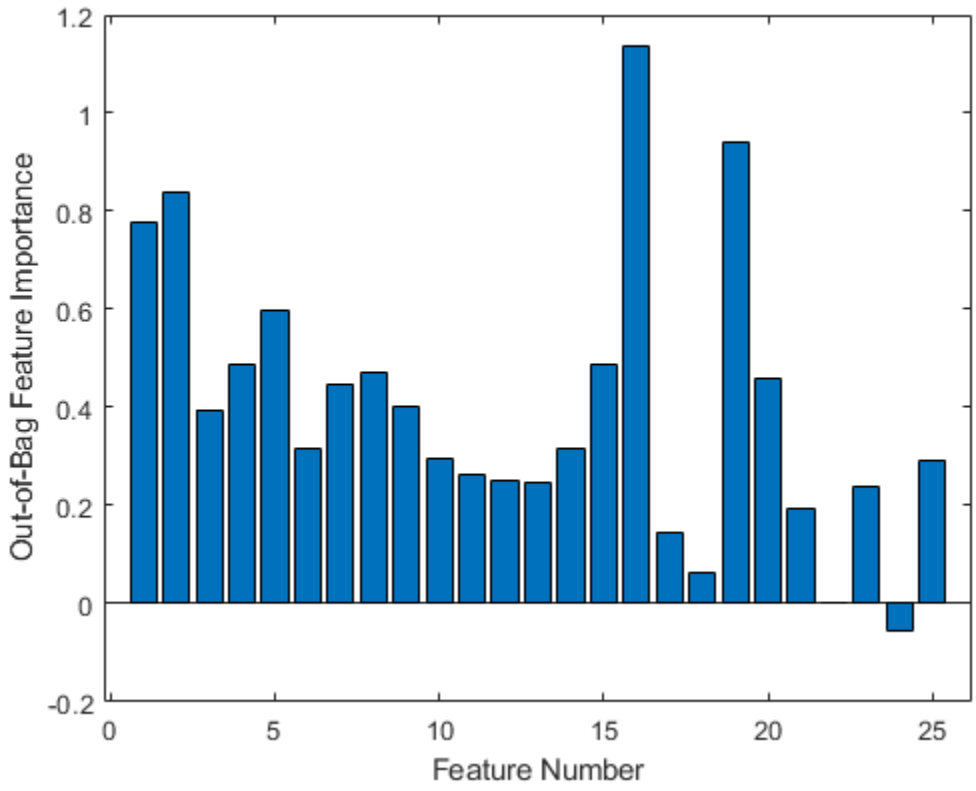


Prediction ability should depend more on important features than unimportant features. You can use this idea to measure feature importance.

For each feature, permute the values of this feature across every observation in the data set and measure how much worse the MSE becomes after the permutation. You can repeat this for each feature.

Plot the increase in MSE due to permuting out-of-bag observations across each input variable. The `OOBPermutedPredictorDeltaError` array stores the increase in MSE averaged over all trees in the ensemble and divided by the standard deviation taken over the trees, for each variable. The larger this value, the more important the variable. Imposing an arbitrary cutoff at 0.7, you can select the four most important features.

```
figure
bar(b.OOBPermutedPredictorDeltaError)
xlabel('Feature Number')
ylabel('Out-of-Bag Feature Importance')
```



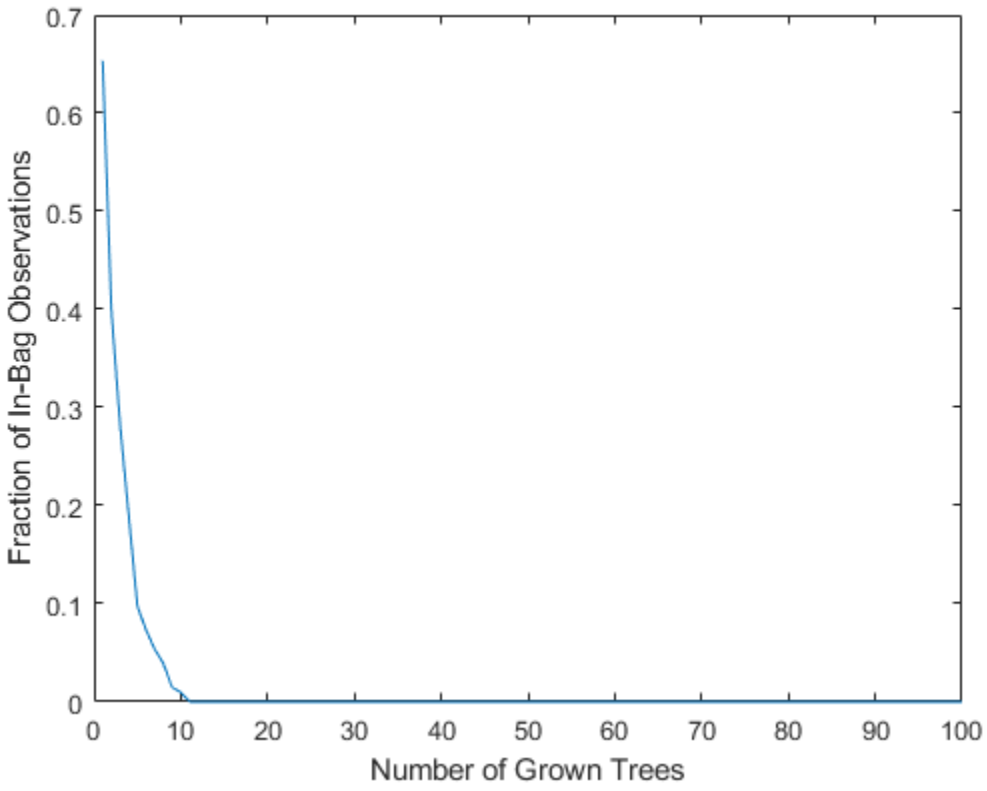

```
idxvar = find(b.OOBPermutedPredictorDeltaError>0.7)
```

```
idxvar = 1×4

     1     2    16    19
```

```
idxCategorical = find(isCategorical(idxvar)==1);
```

The `OOBIndices` property of `TreeBagger` tracks which observations are out of bag for what trees. Using this property, you can monitor the fraction of observations in the training data that are in bag for all trees. The curve starts at approximately 2/3, which is the fraction of unique observations selected by one bootstrap replica, and goes down to 0 at approximately 10 trees.

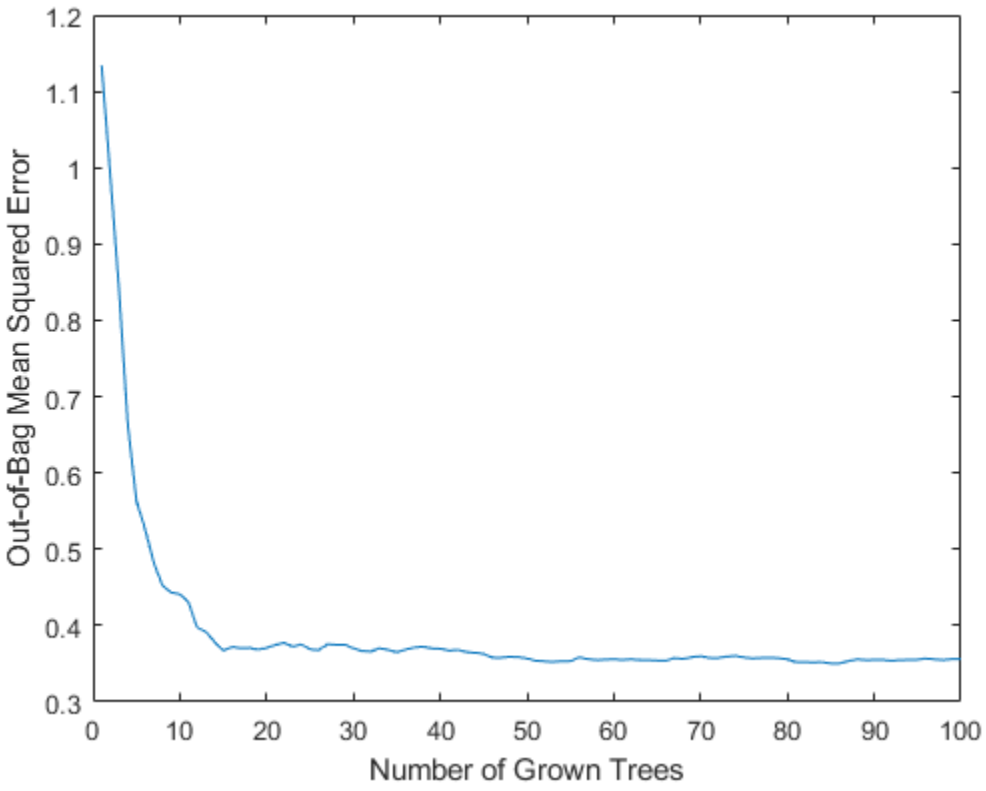
```
finbag = zeros(1,b.NTrees);
for t=1:b.NTrees
    finbag(t) = sum(all(~b.OOBIndices(:,1:t),2));
end
finbag = finbag / size(X,1);
figure
plot(finbag)
xlabel('Number of Grown Trees')
ylabel('Fraction of In-Bag Observations')
```



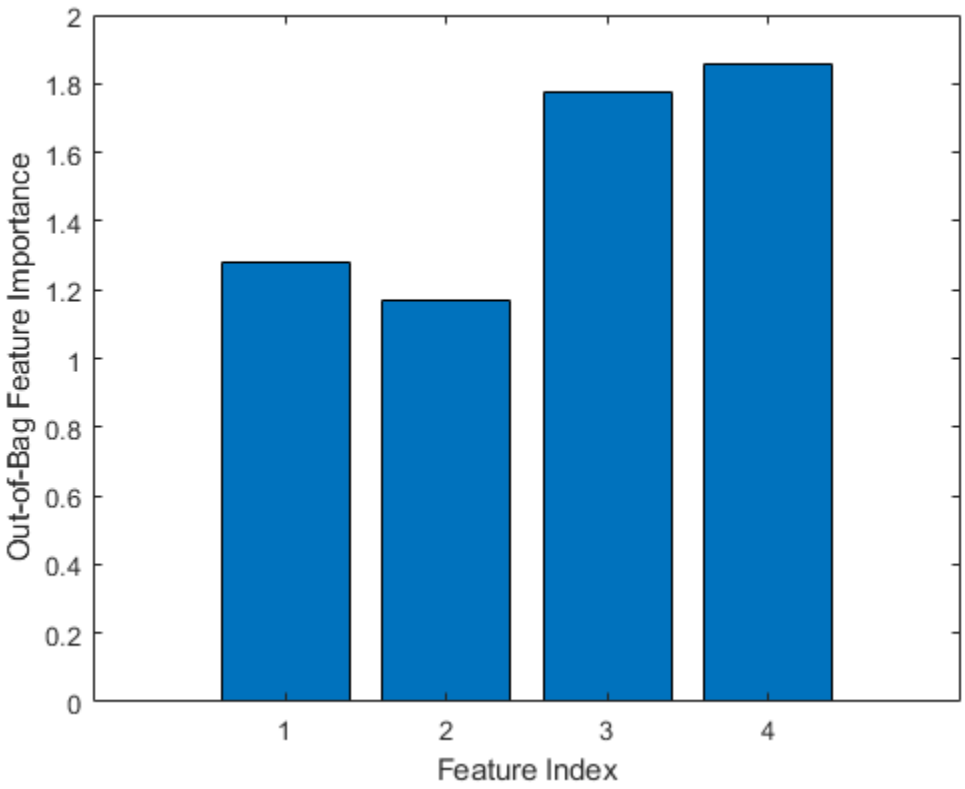
Growing Trees on a Reduced Set of Features

Using just the four most powerful features, determine if it is possible to obtain a similar predictive power. To begin, grow 100 trees on these features only. The first two of the four selected features are numeric and the last two are categorical.

```
b5v = TreeBagger(100,X(:,idxvar),Y,'Method','R',...
    'OOBPredictorImportance','On','CategoricalPredictors',idxCategorical,...
    'MinLeafSize',5);
figure
plot(oobError(b5v))
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Mean Squared Error')
```




```
figure
bar(b5v.OOBPermutedPredictorDeltaError)
xlabel('Feature Index')
ylabel('Out-of-Bag Feature Importance')
```



These four most powerful features give the same MSE as the full set, and the ensemble trained on the reduced set ranks these features similarly to each other. If you remove features 1 and 2 from the reduced set, then the predictive power of the algorithm might not decrease significantly.

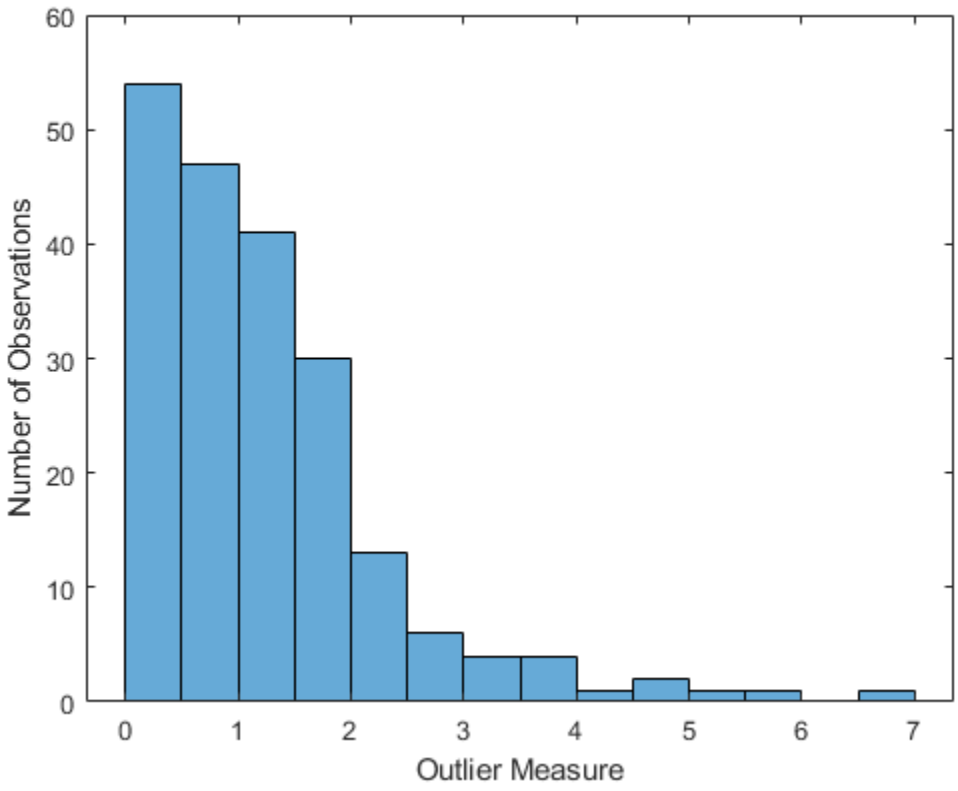
Finding Outliers

To find outliers in the training data, compute the proximity matrix using `fillProximities`.

```
b5v = fillProximities(b5v);
```

The method normalizes this measure by subtracting the mean outlier measure for the entire sample. Then it takes the magnitude of this difference and divides the result by the median absolute deviation for the entire sample.

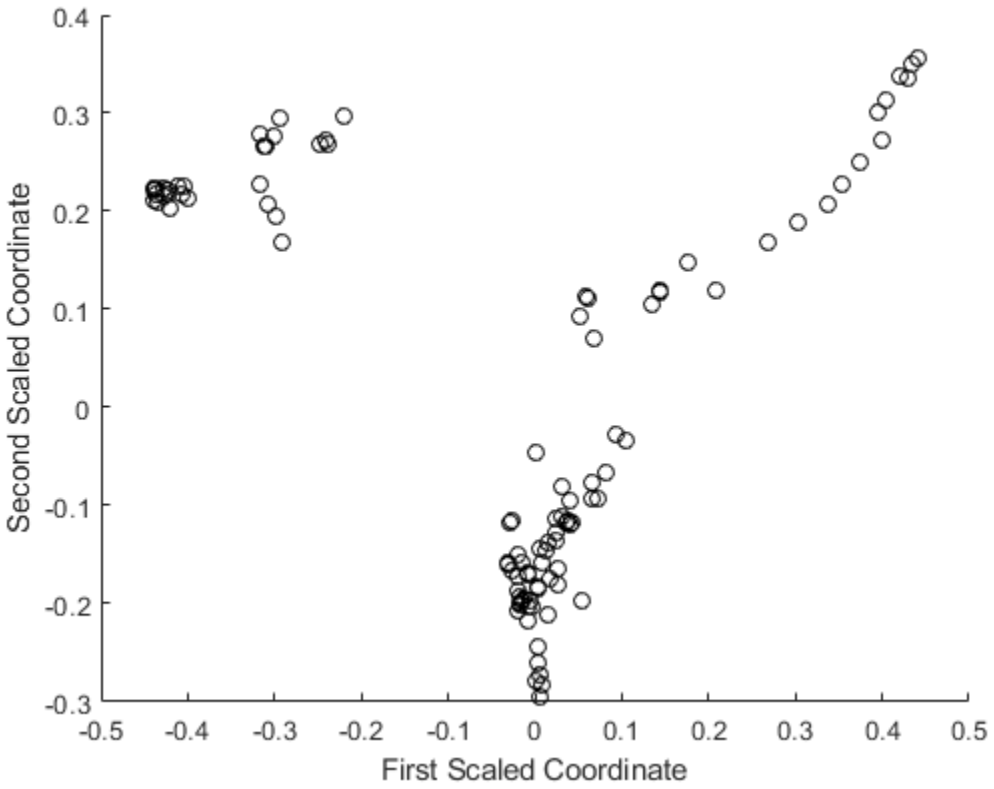
```
figure
histogram(b5v.OutlierMeasure)
xlabel('Outlier Measure')
ylabel('Number of Observations')
```



Discovering Clusters in the Data

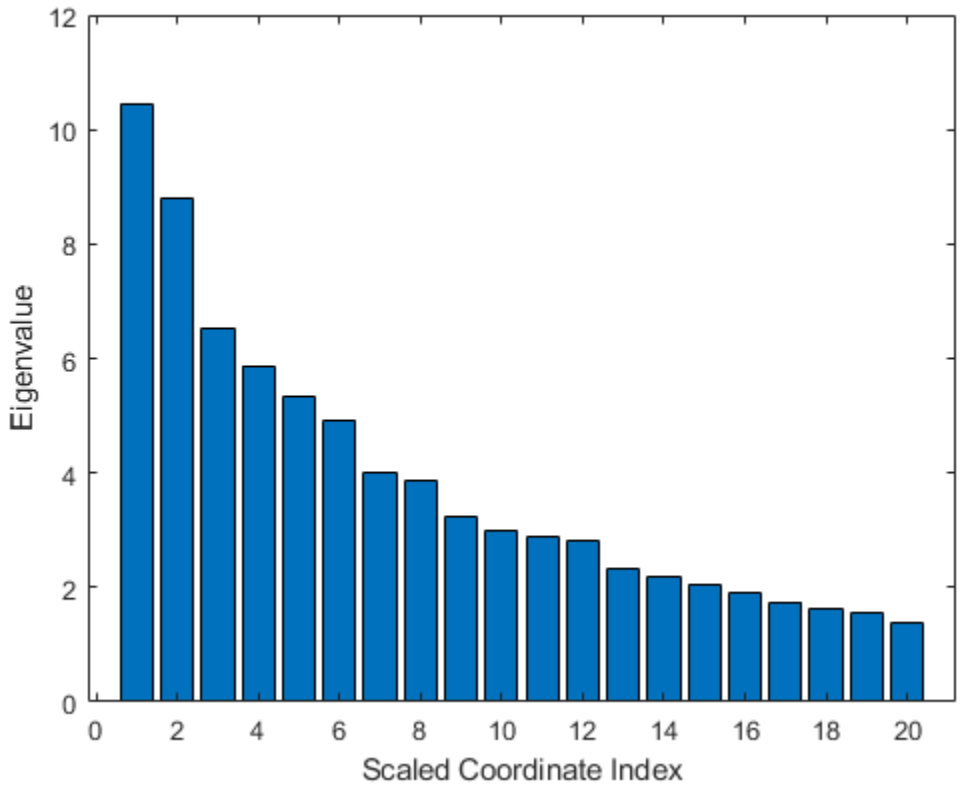
By applying multidimensional scaling to the computed matrix of proximities, you can inspect the structure of the input data and look for possible clusters of observations. The `mdsProx` method returns scaled coordinates and eigenvalues for the computed proximity matrix. If you run it with the `Colors` name-value-pair argument, then this method creates a scatter plot of two scaled coordinates.

```
figure(8)
[~,e] = mdsProx(b5v,'Colors','K');
xlabel('First Scaled Coordinate')
ylabel('Second Scaled Coordinate')
```



Assess the relative importance of the scaled axes by plotting the first 20 eigenvalues.

```
figure
bar(e(1:20))
xlabel('Scaled Coordinate Index')
ylabel('Eigenvalue')
```



Saving the Ensemble Configuration for Future Use

To use the trained ensemble for predicting the response on unseen data, store the ensemble to disk and retrieve it later. If you do not want to compute predictions for out-of-bag data or reuse training data in any other way, there is no need to store the ensemble object itself. Saving the compact version of the ensemble is enough in this case. Extract the compact object from the ensemble.

```
c = compact(b5v)
```

```
c =
  CompactTreeBagger
Ensemble with 100 bagged decision trees:
              Method:           regression
      NumPredictors:                    4

  Properties, Methods
```

You can save the resulting `CompactTreeBagger` model in a `*.mat` file.

See Also

`TreeBagger` | `compact` | `fillprox` | `fitrensemble` | `mdsprox` | `oobError`

Related Examples

- “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

- “Comparison of TreeBagger and Bagged Ensembles” on page 18-44
- “Use Parallel Processing for Regression TreeBagger Workflow” on page 31-6

Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger

Statistics and Machine Learning Toolbox™ offers two objects that support bootstrap aggregation (bagging) of classification trees: `TreeBagger` created by using `TreeBagger` and `ClassificationBaggedEnsemble` created by using `fitcensemble`. See “Comparison of TreeBagger and Bagged Ensembles” on page 18-44 for differences between `TreeBagger` and `ClassificationBaggedEnsemble`.

This example shows the workflow for classification using the features in `TreeBagger` only.

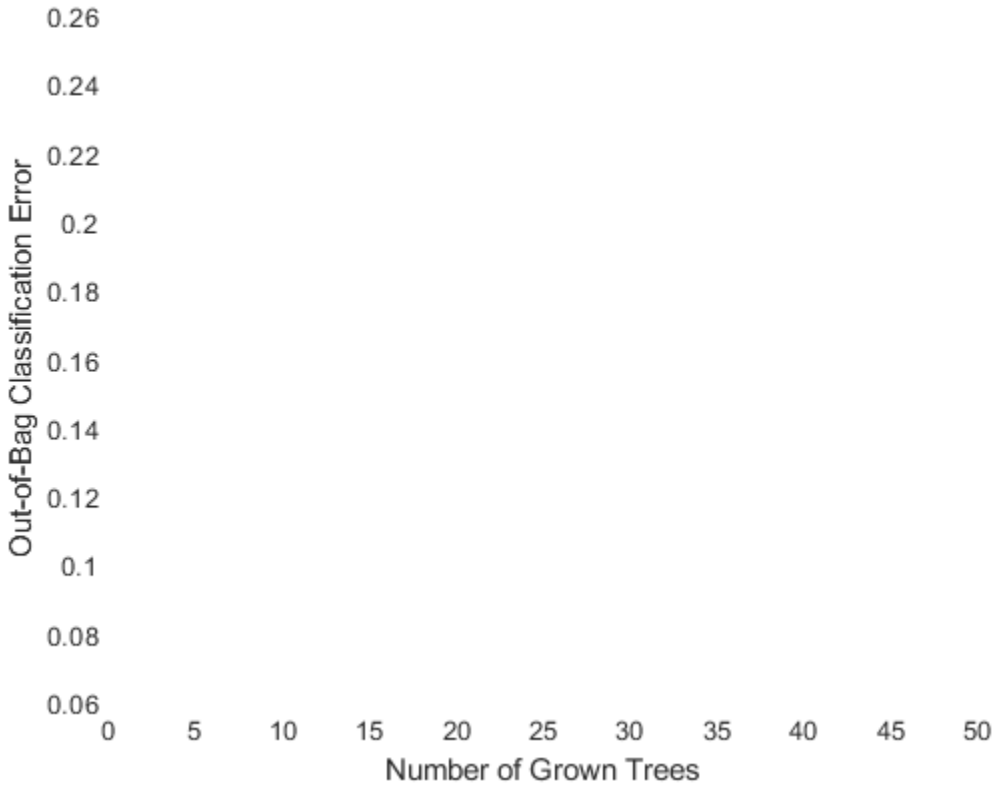
Use ionosphere data with 351 observations and 34 real-valued predictors. The response variable is categorical with two levels:

- `'g'` represents good radar returns.
- `'b'` represents bad radar returns.

The goal is to predict good or bad returns using a set of 34 measurements.

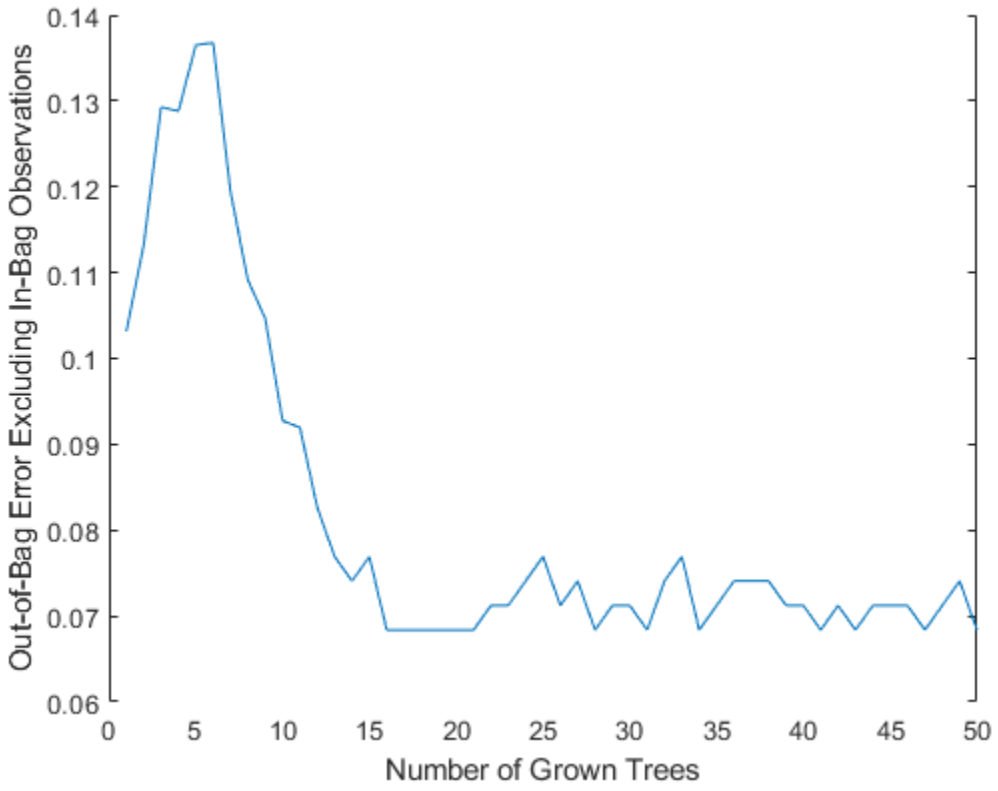
Fix the initial random seed, grow 50 trees, inspect how the ensemble error changes with accumulation of trees, and estimate feature importance. For classification, it is best to set the minimal leaf size to 1 and select the square root of the total number of features for each decision split at random. These settings are defaults for `TreeBagger` used for classification.

```
load ionosphere
rng(1945,'twister')
b = TreeBagger(50,X,Y,'OOBPredictorImportance','On');
figure
plot(oobError(b))
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Classification Error')
```



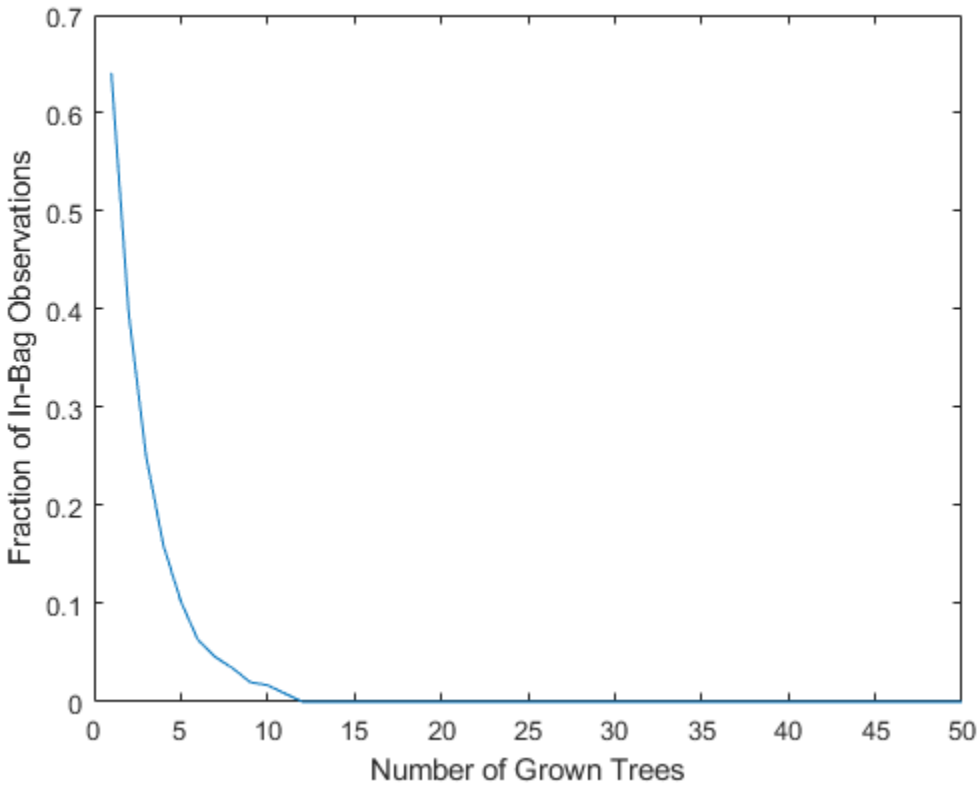
The method trains ensembles with few trees on observations that are in bag for all trees. For such observations, it is impossible to compute the true out-of-bag prediction, and `TreeBagger` returns the most probable class for classification and the sample mean for regression. You can change the default value returned for in-bag observations using the `DefaultYfit` property. If you set the default value to an empty character vector for classification, the method excludes in-bag observations from computation of the out-of-bag error. In this case, the curve is more variable when the number of trees is small, either because some observations are never out of bag (and are therefore excluded) or because their predictions are based on few trees.

```
b.DefaultYfit = '';
figure
plot(oobError(b))
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Error Excluding In-Bag Observations')
```



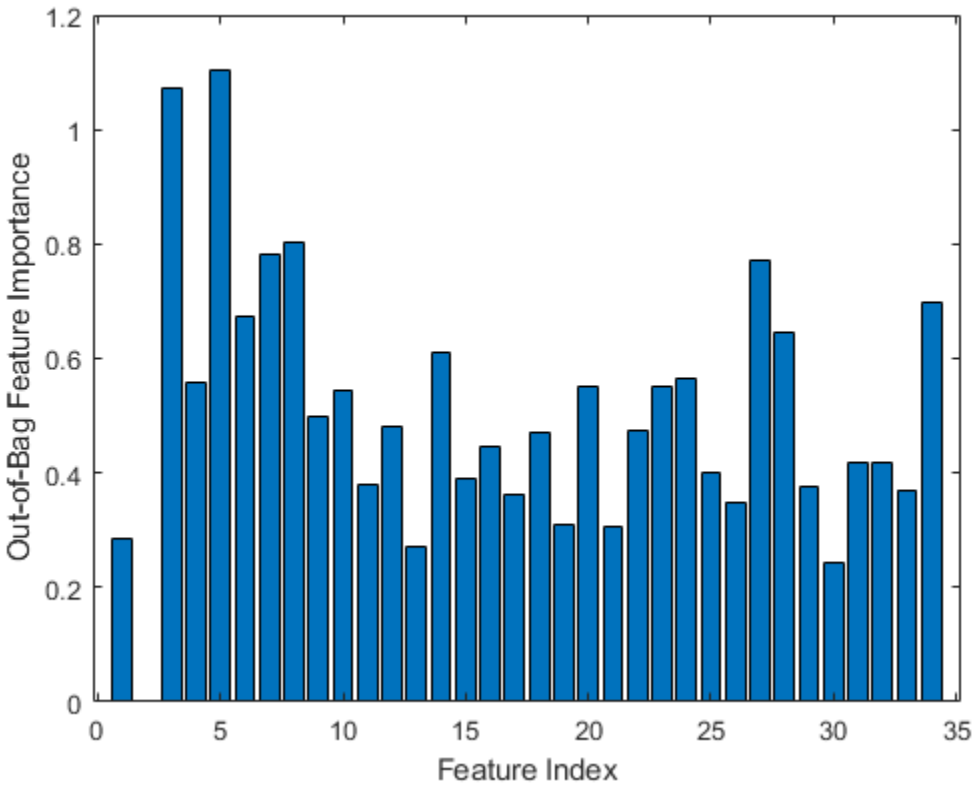
The `OOBIndices` property of `TreeBagger` tracks which observations are out of bag for what trees. Using this property, you can monitor the fraction of observations in the training data that are in bag for all trees. The curve starts at approximately 2/3, which is the fraction of unique observations selected by one bootstrap replica, and goes down to 0 at approximately 10 trees.

```
finbag = zeros(1,b.NumTrees);
for t=1:b.NTrees
    finbag(t) = sum(all(~b.OOBIndices(:,1:t),2));
end
finbag = finbag / size(X,1);
figure
plot(finbag)
xlabel('Number of Grown Trees')
ylabel('Fraction of In-Bag Observations')
```



Estimate feature importance.

```
figure
bar(b.OOBPermutedPredictorDeltaError)
xlabel('Feature Index')
ylabel('Out-of-Bag Feature Importance')
```



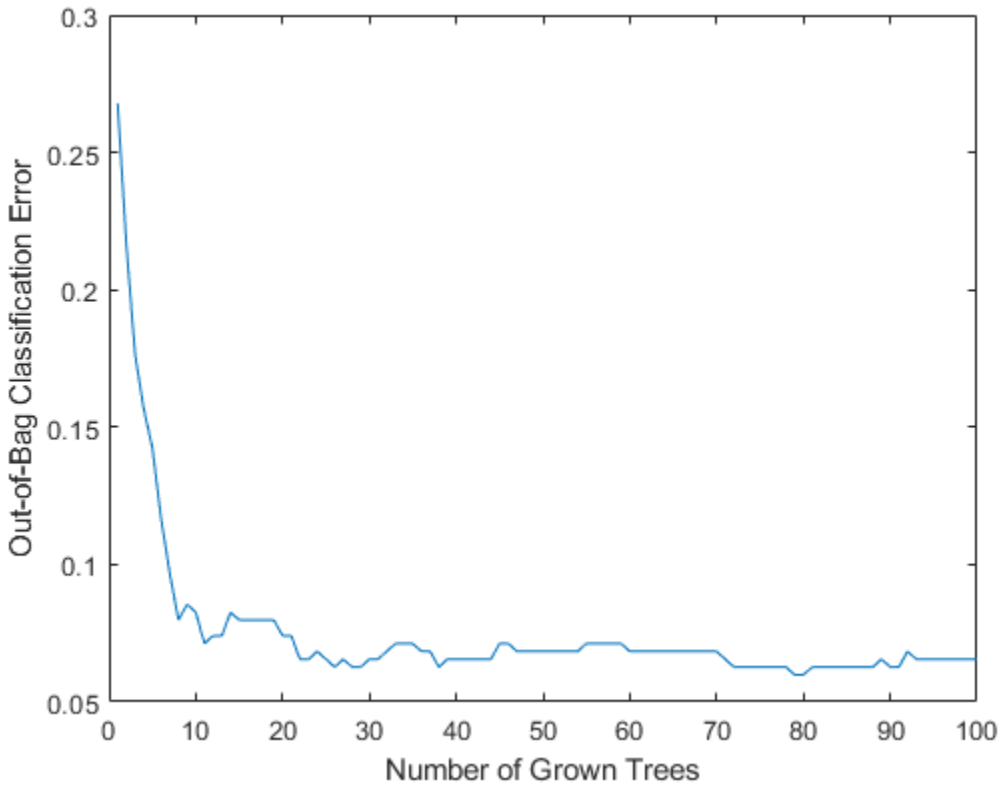
Select the features yielding an importance measure greater than 0.75. This threshold is chosen arbitrarily.

```
idxvar = find(b.OOBPermutedPredictorDeltaError>0.75)
```

```
idxvar = 1×5

     3     5     7     8    27
```

Having selected the most important features, grow a larger ensemble on the reduced feature set. Save time by not permuting out-of-bag observations to obtain new estimates of feature importance for the reduced feature set (set `OOBVarImp` to `'off'`). You would still be interested in obtaining out-of-bag estimates of classification error (set `OOBPred` to `'on'`).

```
b5v = TreeBagger(100,X(:,idxvar),Y,'OOBPredictorImportance','off','OOBPrediction','on');
figure
plot(oobError(b5v))
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Classification Error')
```

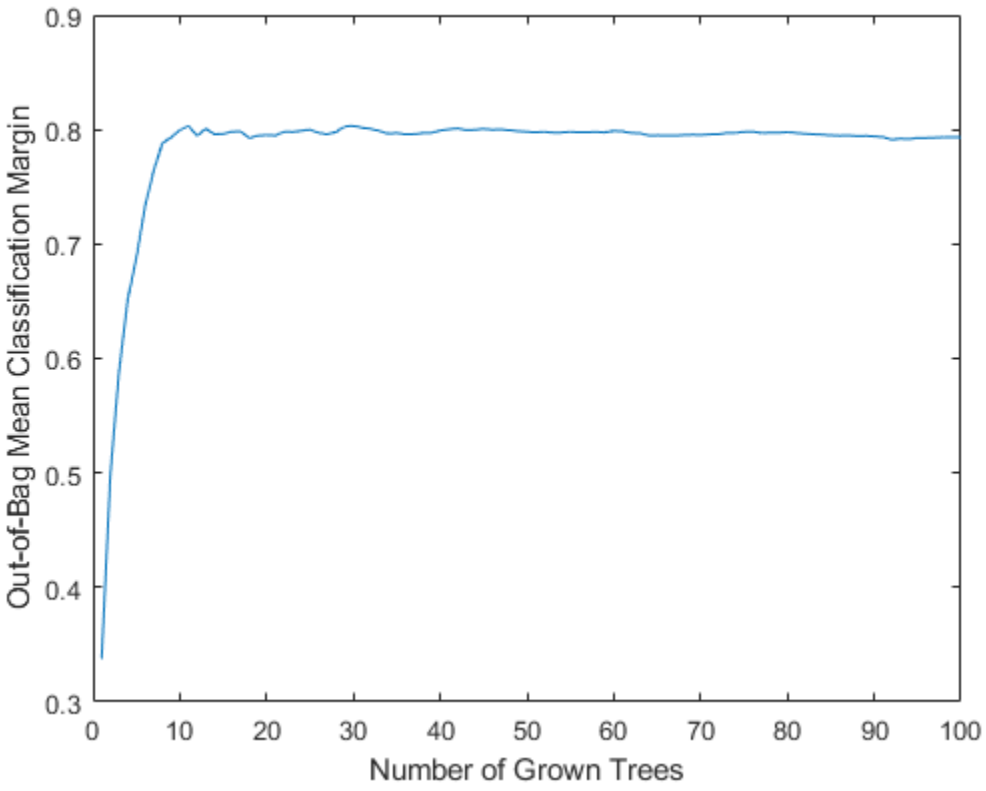


For classification ensembles, in addition to classification error (fraction of misclassified observations), you can also monitor the average classification margin. For each observation, the *margin* is defined as the difference between the score for the true class and the maximal score for other classes predicted by this tree. The cumulative classification margin uses the scores averaged over all trees and the mean cumulative classification margin is the cumulative margin averaged over all observations. The `oobMeanMargin` method with the `'mode'` argument set to `'cumulative'` (default) shows how the mean cumulative margin changes as the ensemble grows: every new element in the returned array represents the cumulative margin obtained by including a new tree in the ensemble. If training is successful, you would expect to see a gradual increase in the mean classification margin.

The method trains ensembles with few trees on observations that are in bag for all trees. For such observations, it is impossible to compute the true out-of-bag prediction, and `TreeBagger` returns the most probable class for classification and the sample mean for regression.

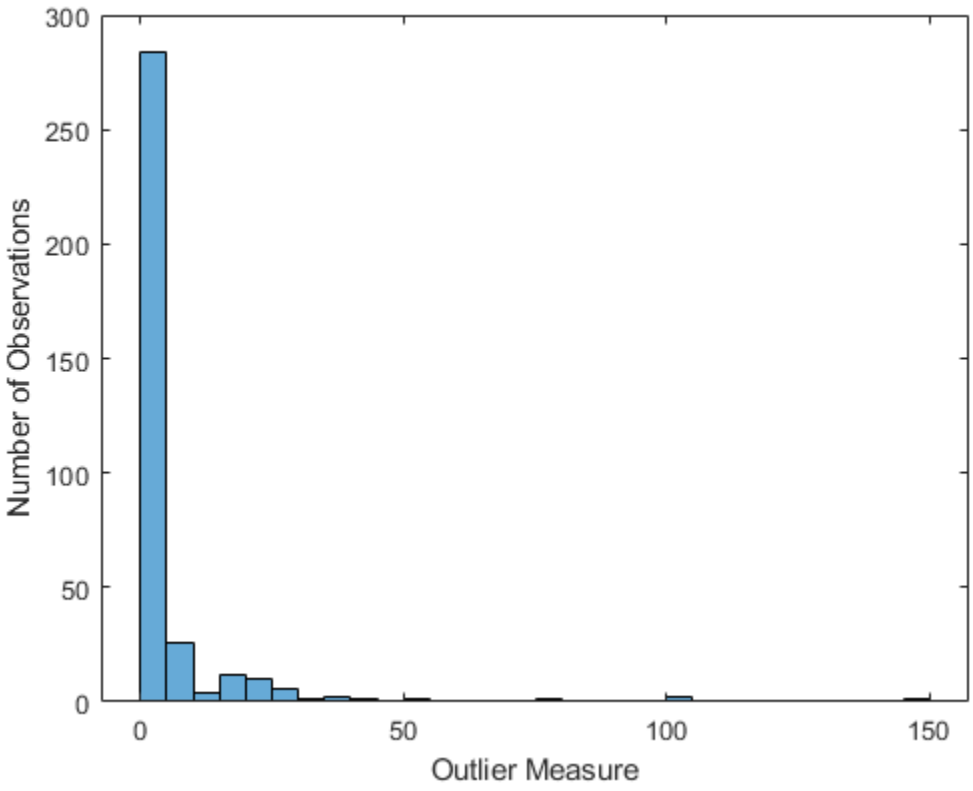
For decision trees, a classification score is the probability of observing an instance of this class in this tree leaf. For example, if the leaf of a grown decision tree has five `'good'` and three `'bad'` training observations in it, the scores returned by this decision tree for any observation fallen on this leaf are 5/8 for the `'good'` class and 3/8 for the `'bad'` class. These probabilities are called `'scores'` for consistency with other classifiers that might not have an obvious interpretation for numeric values of returned predictions.

```
figure
plot(oobMeanMargin(b5v));
xlabel('Number of Grown Trees')
ylabel('Out-of-Bag Mean Classification Margin')
```



Compute the matrix of proximities and examine the distribution of outlier measures. Unlike regression, outlier measures for classification ensembles are computed within each class separately.

```
b5v = fillProximities(b5v);
figure
histogram(b5v.OutlierMeasure)
xlabel('Outlier Measure')
ylabel('Number of Observations')
```



Find the class of the extreme outliers.

```
extremeOutliers = b5v.Y(b5v.OutlierMeasure>40)
```

```
extremeOutliers = 6x1 cell
    {'g'}
    {'g'}
    {'g'}
    {'g'}
    {'g'}
    {'g'}
```

```
percentGood = 100*sum(strcmp(extremeOutliers,'g'))/numel(extremeOutliers)
```

```
percentGood = 100
```

All of the extreme outliers are labeled `'good'`.

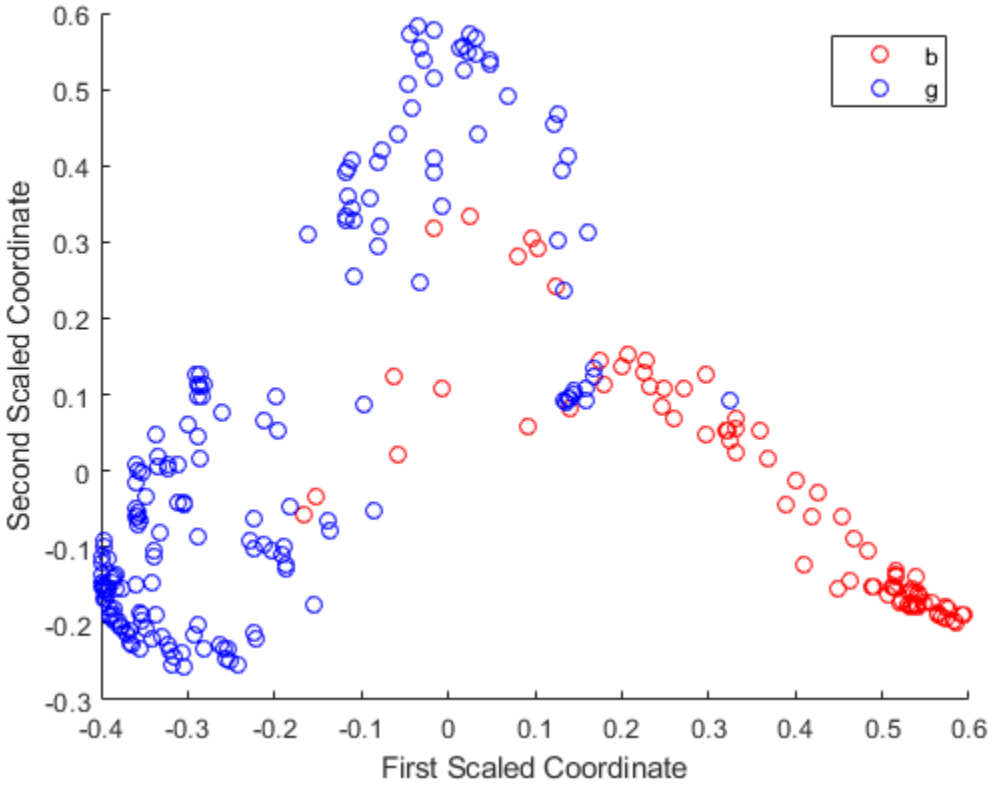
As for regression, you can plot scaled coordinates, displaying the two classes in different colors using the 'Colors' name-value pair argument of `mdsProx`. This argument takes a character vector in which every character represents a color. The software does not rank class names. Therefore, it is best practice to determine the position of the classes in the `ClassNames` property of the ensemble.

```
gPosition = find(strcmp('g',b5v.ClassNames))
```

```
gPosition = 2
```

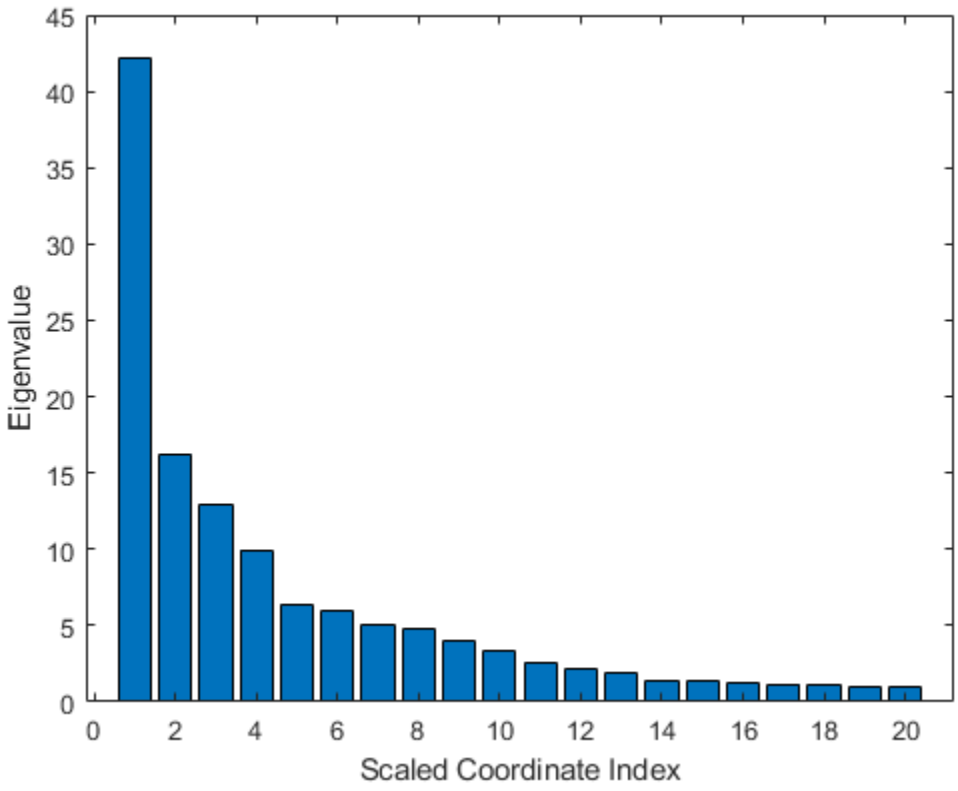
The `'bad'` class is first and the `'good'` class is second. Display scaled coordinates using red for the `'bad'` class and blue for the `'good'` class observations.

```
figure
[s,e] = mdsProx(b5v,'Colors','rb');
xlabel('First Scaled Coordinate')
ylabel('Second Scaled Coordinate')
```



Plot the first 20 eigenvalues obtained by scaling. The first eigenvalue clearly dominates and the first scaled coordinate is most important.

```
figure
bar(e(1:20))
xlabel('Scaled Coordinate Index')
ylabel('Eigenvalue')
```

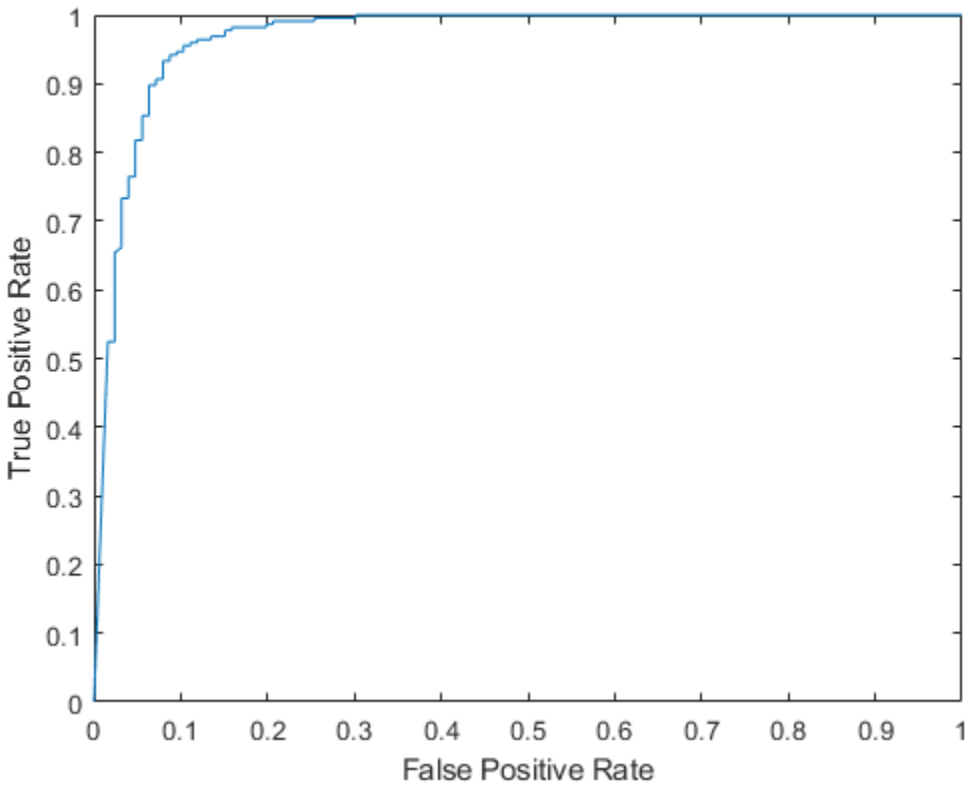


Another way of exploring the performance of a classification ensemble is to plot its Receiver Operating Characteristic (ROC) curve or another performance curve suitable for the current problem. Obtain predictions for out-of-bag observations. For a classification ensemble, the `oobPredict` method returns a cell array of classification labels as the first output argument and a numeric array of scores as the second output argument. The returned array of scores has two columns, one for each class. In this case, the first column is for the `'bad'` class and the second column is for the `'good'` class. One column in the score matrix is redundant because the scores represent class probabilities in tree leaves and by definition add up to 1.

```
[Yfit,Sfit] = oobPredict(b5v);
```

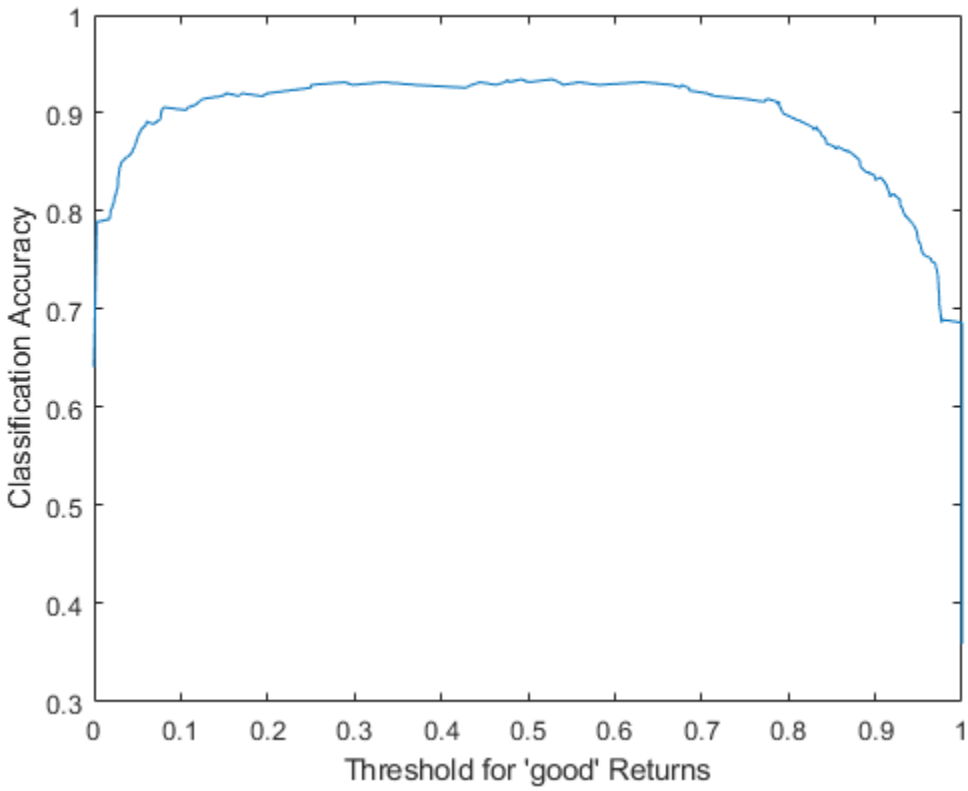
Use `perfcurve` to compute a performance curve. By default, `perfcurve` returns the standard ROC curve, which is the true positive rate versus the false positive rate. `perfcurve` requires true class labels, scores, and the positive class label for input. In this case, choose the `'good'` class as positive.

```
[fpr,tpr] = perfcurve(b5v.Y,Sfit(:,gPosition),'g');
figure
plot(fpr,tpr)
xlabel('False Positive Rate')
ylabel('True Positive Rate')
```



Instead of the standard ROC curve, you might want to plot, for example, ensemble accuracy versus threshold on the score for the `'good'` class. The `ycrit` input argument of `perfcurve` lets you specify the criterion for the y-axis, and the third output argument of `perfcurve` returns an array of thresholds for the positive class score. Accuracy is the fraction of correctly classified observations, or equivalently, 1 minus the classification error.

```
[fpr,accu,thre] = perfcurve(b5v.Y,Sfit(:,gPosition),'g','YCrit','Accu');
figure(20)
plot(thre,accu)
xlabel('Threshold for ''good'' Returns')
ylabel('Classification Accuracy')
```



The curve shows a flat region indicating that any threshold from 0.2 to 0.6 is a reasonable choice. By default, the `perfcurve` assigns classification labels using 0.5 as the boundary between the two classes. You can find exactly what accuracy this corresponds to.

```
accu(abs(thre-0.5)<eps)
```

```
ans = 0.9316
```

The maximal accuracy is a little higher than the default one.

```
[maxaccu,iaccu] = max(accu)
```

```
maxaccu = 0.9345
```

```
iaccu = 99
```

The optimal threshold is therefore.

```
thre(iaccu)
```

```
ans = 0.5278
```

See Also

`TreeBagger` | `compact` | `fitcensemble` | `mdsprox` | `oobError` | `oobMeanMargin` | `oobPredict` | `perfcurve`

Related Examples

- “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
- “Comparison of TreeBagger and Bagged Ensembles” on page 18-44
- “Use Parallel Processing for Regression TreeBagger Workflow” on page 31-6

Detect Outliers Using Quantile Regression

This example shows how to detect outliers using quantile random forest. Quantile random forest can detect outliers with respect to the conditional distribution of Y given X. However, this method cannot detect outliers in the predictor data. For outlier detection in the predictor data using a bag of decision trees, see the `OutlierMeasure` property of a `TreeBagger` model.

An *outlier* is an observation that is located far enough from most of the other observations in a data set and can be considered anomalous. Causes of outlying observations include inherent variability or measurement error. Outliers significant affect estimates and inference, so it is important to detect them and decide whether to remove them or consider a robust analysis.

Statistics and Machine Learning Toolbox™ provides several functions to detect outliers, including:

- `zscore` — Compute z scores of observations.
- `trimmean` — Estimate mean of data, excluding outliers.
- `boxplot` — Draw box plot of data.
- `probplot` — Draw probability plot.
- `robustcov` — Estimate robust covariance of multivariate data.
- `fitcsvm` — Fit a one-class support vector machine (SVM) to determine which observations are located far from the decision boundary.
- `dbscan` — Partition observations into clusters and identify outliers using the density-based spatial clustering of application with noise (DBSCAN) algorithm.

Also, MATLAB® provides the `isoutlier` function, which finds outliers in data.

To demonstrate outlier detection, this example:

1. Generates data from a nonlinear model with heteroscedasticity and simulates a few outliers.
2. Grows a quantile random forest of regression trees.
3. Estimates conditional quartiles (Q_1, Q_2, and Q_3) and the interquartile range (IQR) within the ranges of the predictor variables.
4. Compares the observations to the *fences*, which are the quantities $F_1 = Q_1 - 1.5IQR$ and $F_2 = Q_3 + 1.5IQR$. Any observation that is less than F_1 or greater than F_2 is an outlier.

Generate Data

Generate 500 observations from the model

$$y_t = 10 + 3t + t\sin(2t) + \varepsilon_t .$$

t is uniformly distributed between 0 and 4π, and $\varepsilon_t \sim N(0, t + 0.01)$. Store the data in a table.

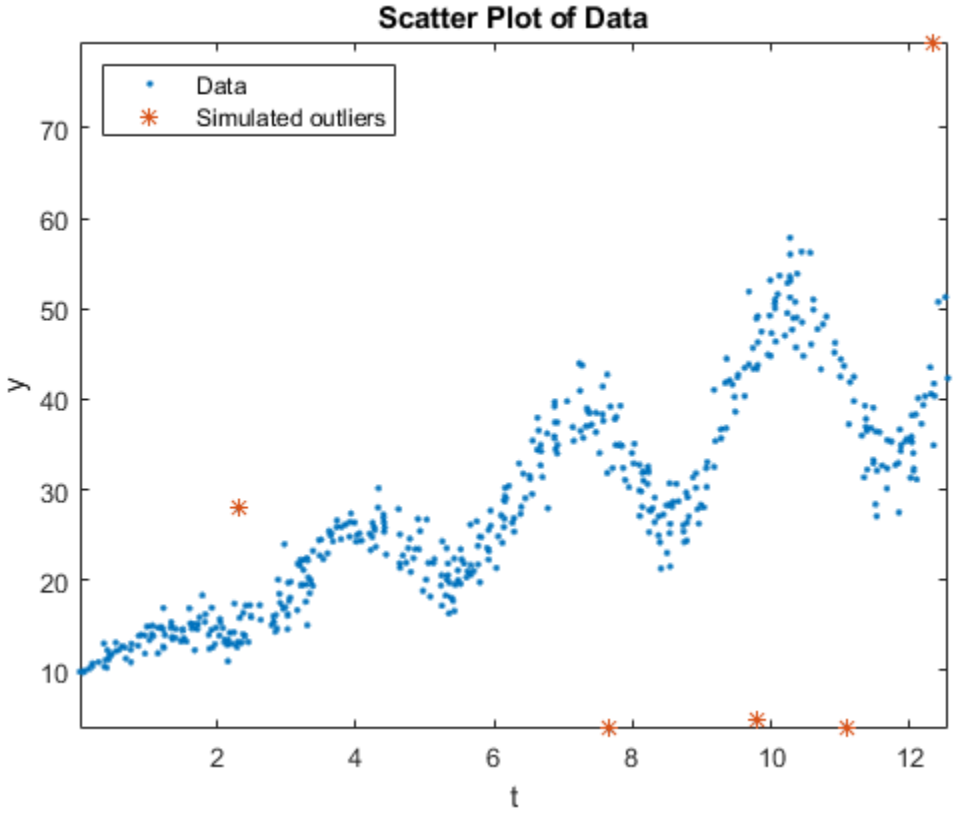
```
n = 500;
rng('default'); % For reproducibility
t = randsample(linspace(0,4*pi,1e6),n,true)';
epsilon = randn(n,1).*sqrt((t+0.01));
y = 10 + 3*t + t.*sin(2*t) + epsilon;

Tbl = table(t,y);
```

Move five observations in a random vertical direction by 90% of the value of the response.

```
numOut = 5;
[~,idx] = datasample(Tbl,numOut);
Tbl.y(idx) = Tbl.y(idx) + randsample([-1 1],numOut,true)'.*(0.9*Tbl.y(idx));
```

Draw a scatter plot of the data and identify the outliers.

```
figure;
plot(Tbl.t,Tbl.y,'.');
hold on
plot(Tbl.t(idx),Tbl.y(idx),'*');
axis tight;
ylabel('y');
xlabel('t');
title('Scatter Plot of Data');
legend('Data','Simulated outliers','Location','NorthWest');
```



Grow Quantile Random Forest

Grow a bag of 200 regression trees using `TreeBagger`.

```
Mdl = TreeBagger(200,Tbl,'y','Method','regression');
```

`Mdl` is a `TreeBagger` ensemble.

Predict Conditional Quartiles and Interquartile Ranges

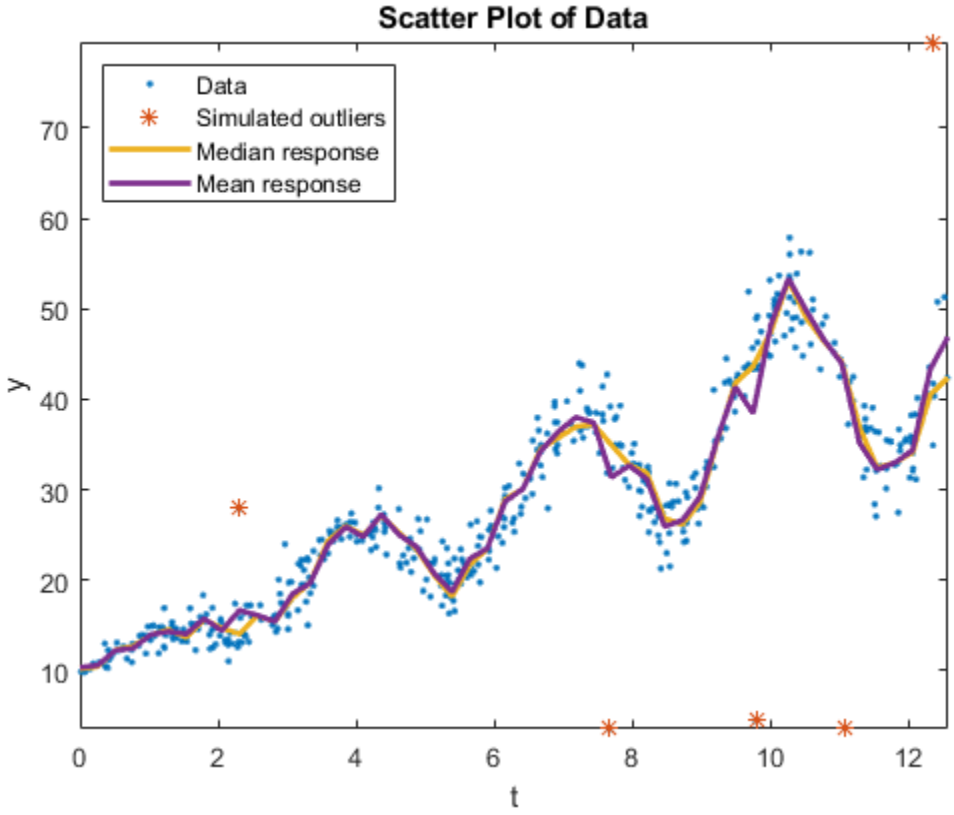
Using quantile regression, estimate the conditional quartiles of 50 equally spaced values within the range of `t`.

```
tau = [0.25 0.5 0.75];
predT = linspace(0,4*pi,50)';
quartiles = quantilePredict(Mdl,predT,'Quantile',tau);
```

`quartiles` is a 500-by-3 matrix of conditional quartiles. Rows correspond to the observations in `t`, and columns correspond to the probabilities in `tau`.

On the scatter plot of the data, plot the conditional mean and median responses.

```
meanY = predict(Mdl,predT);

plot(predT,[quartiles(:,2) meanY],'LineWidth',2);
legend('Data','Simulated outliers','Median response','Mean response',...
    'Location','NorthWest');
hold off;
```



Although the conditional mean and median curves are close, the simulated outliers can affect the mean curve.

Compute the conditional IQR, F_1, and F_2.

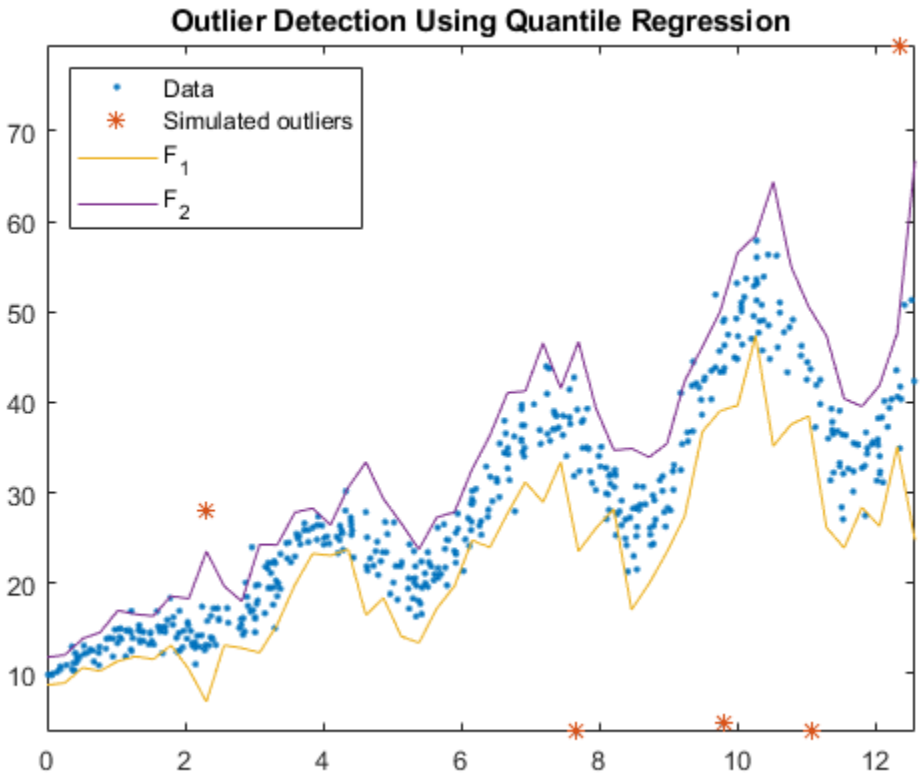
```
iqr = quartiles(:,3) - quartiles(:,1);
k = 1.5;
f1 = quartiles(:,1) - k*iqr;
f2 = quartiles(:,3) + k*iqr;
```

`k = 1.5` means that all observations less than `f1` or greater than `f2` are considered outliers, but this threshold does not disambiguate from extreme outliers. A `k` of 3 identifies extreme outliers.

Compare Observations to Fences

Plot the observations and the fences.

```
figure;
plot(Tbl.t,Tbl.y,'.');
hold on
plot(Tbl.t(idx),Tbl.y(idx),'*');
plot(predT,[f1 f2]);
legend('Data','Simulated outliers','F_1','F_2','Location','NorthWest');
axis tight
title('Outlier Detection Using Quantile Regression')
hold off
```



All simulated outliers fall outside $[F_1, F_2]$, and some observations are outside this interval as well.

See Also

Classes
`TreeBagger`

Functions
`TreeBagger` | `predict` | `quantilePredict`

Related Examples

- “Conditional Quantile Estimation Using Kernel Smoothing” on page 18-142
- “Tune Random Forest Using Quantile Error and Bayesian Optimization” on page 18-145

Conditional Quantile Estimation Using Kernel Smoothing

This example shows how to estimate conditional quantiles of a response given predictor data using quantile random forest and by estimating the conditional distribution function of the response using kernel smoothing.

For quantile-estimation speed, `quantilePredict`, `oobQuantilePredict`, `quantileError`, and `oobQuantileError` use linear interpolation to predict quantiles in the conditional distribution of the response. However, you can obtain response weights, which comprise the distribution function, and then pass them to `ksdensity` to possibly gain accuracy at the cost of computation speed.

Generate 2000 observations from the model

$$y_t = 0.5 + t + \varepsilon_t.$$

t is uniformly distributed between 0 and 1, and $\varepsilon_t \sim N(0, t^2/2 + 0.01)$. Store the data in a table.

```
n = 2000;
rng('default'); % For reproducibility
t = randsample(linspace(0,1,1e2),n,true)';
epsilon = randn(n,1).*sqrt(t.^2/2 + 0.01);
y = 0.5 + t + epsilon;

Tbl = table(t,y);
```

Train an ensemble of bagged regression trees using the entire data set. Specify 200 weak learners and save the out-of-bag indices.

```
rng('default'); % For reproducibility
Mdl = TreeBagger(200,Tbl,'y','Method','regression',...
    'OOBPrediction','on');
```

`Mdl` is a `TreeBagger` ensemble.

Predict out-of-bag, conditional 0.05 and 0.95 quantiles (90% confidence intervals) for all training-sample observations using `oobQuantilePredict`, that is, by interpolation. Request response weights. Record the execution time.

```
tau = [0.05 0.95];
tic
[quantInterp,yw] = oobQuantilePredict(Mdl,'Quantile',tau);
timeInterp = toc;
```

`quantInterp` is a 94-by-2 matrix of predicted quantiles; rows correspond to the observations in `Mdl.X` and columns correspond to the quantile probabilities in `tau`. `yw` is a 94-by-94 sparse matrix of response weights; rows correspond to training-sample observations and columns correspond to the observations in `Mdl.X`. Response weights are independent of `tau`.

Predict out-of-bag, conditional 0.05 and 0.95 quantiles using kernel smoothing and record the execution time.

```
n = numel(Tbl.y);
quantKS = zeros(n,numel(tau)); % Preallocation

tic
for j = 1:n
```

```
    quantKS(j,:) = ksdensity(Tbl.y,tau,'Function','icdf','Weights',yw(:,j));
end
timeKS = toc;
```

`quantKS` is commensurate with `quantInterp`.

Evaluate the ratio of execution times between kernel smoothing estimation and interpolation.

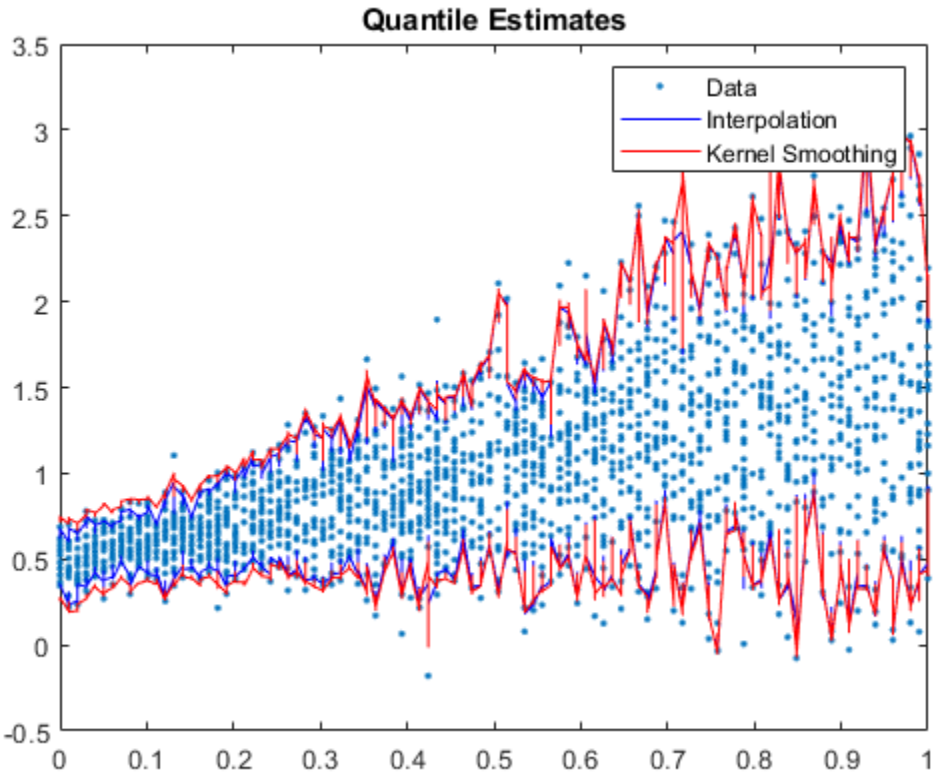
```
timeKS/timeInterp
```

```
ans = 7.7973
```

It takes much more time to execute kernel smoothing than interpolation. This ratio is dependent on the memory of your machine, so your results will vary.

Plot the data with both sets of predicted quantiles.

```
[sT,idx] = sort(t);

figure;
h1 = plot(t,y,'.');
hold on
h2 = plot(sT,quantInterp(idx,:),'b');
h3 = plot(sT,quantKS(idx,:),'r');
legend([h1 h2(1) h3(1)],'Data','Interpolation','Kernel Smoothing');
title('Quantile Estimates')
hold off
```



Both sets of estimated quantiles agree fairly well. However, the quantile intervals from interpolation appear slightly tighter for smaller values of `t` than the ones from kernel smoothing.

See Also

`TreeBagger` | `TreeBagger` | `ksdensity` | `oobQuantilePredict`

Related Examples

- “Detect Outliers Using Quantile Regression” on page 18-137
- “Tune Random Forest Using Quantile Error and Bayesian Optimization” on page 18-145

Tune Random Forest Using Quantile Error and Bayesian Optimization

This example shows how to implement Bayesian optimization to tune the hyperparameters of a random forest of regression trees using quantile error. Tuning a model using quantile error, rather than mean squared error, is appropriate if you plan to use the model to predict conditional quantiles rather than conditional means.

Load and Preprocess Data

Load the `carsmall` data set. Consider a model that predicts the median fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
rng('default'); % For reproducibility
```

Specify Tuning Parameters

Consider tuning:

- The complexity (depth) of the trees in the forest. Deep trees tend to over-fit, but shallow trees tend to underfit. Therefore, specify that the minimum number of observations per leaf be at most 20.
- When growing the trees, the number of predictors to sample at each node. Specify sampling from 1 through all of the predictors.

`bayesopt`, the function that implements Bayesian optimization, requires you to pass these specifications as `optimizableVariable` objects.

```
maxMinLS = 20;
minLS = optimizableVariable('minLS',[1,maxMinLS],'Type','integer');
numPTS = optimizableVariable('numPTS',[1,size(X,2)-1],'Type','integer');
hyperparametersRF = [minLS; numPTS];
```

`hyperparametersRF` is a 2-by-1 array of `OptimizableVariable` objects.

You should also consider tuning the number of trees in the ensemble. `bayesopt` tends to choose random forests containing many trees because ensembles with more learners are more accurate. If available computation resources is a consideration, and you prefer ensembles with as fewer trees, then consider tuning the number of trees separately from the other parameters or penalizing models containing many learners.

Define Objective Function

Define an objective function for the Bayesian optimization algorithm to optimize. The function should:

- Accept the parameters to tune as an input.
- Train a random forest using `TreeBagger`. In the `TreeBagger` call, specify the parameters to tune and specify returning the out-of-bag indices.

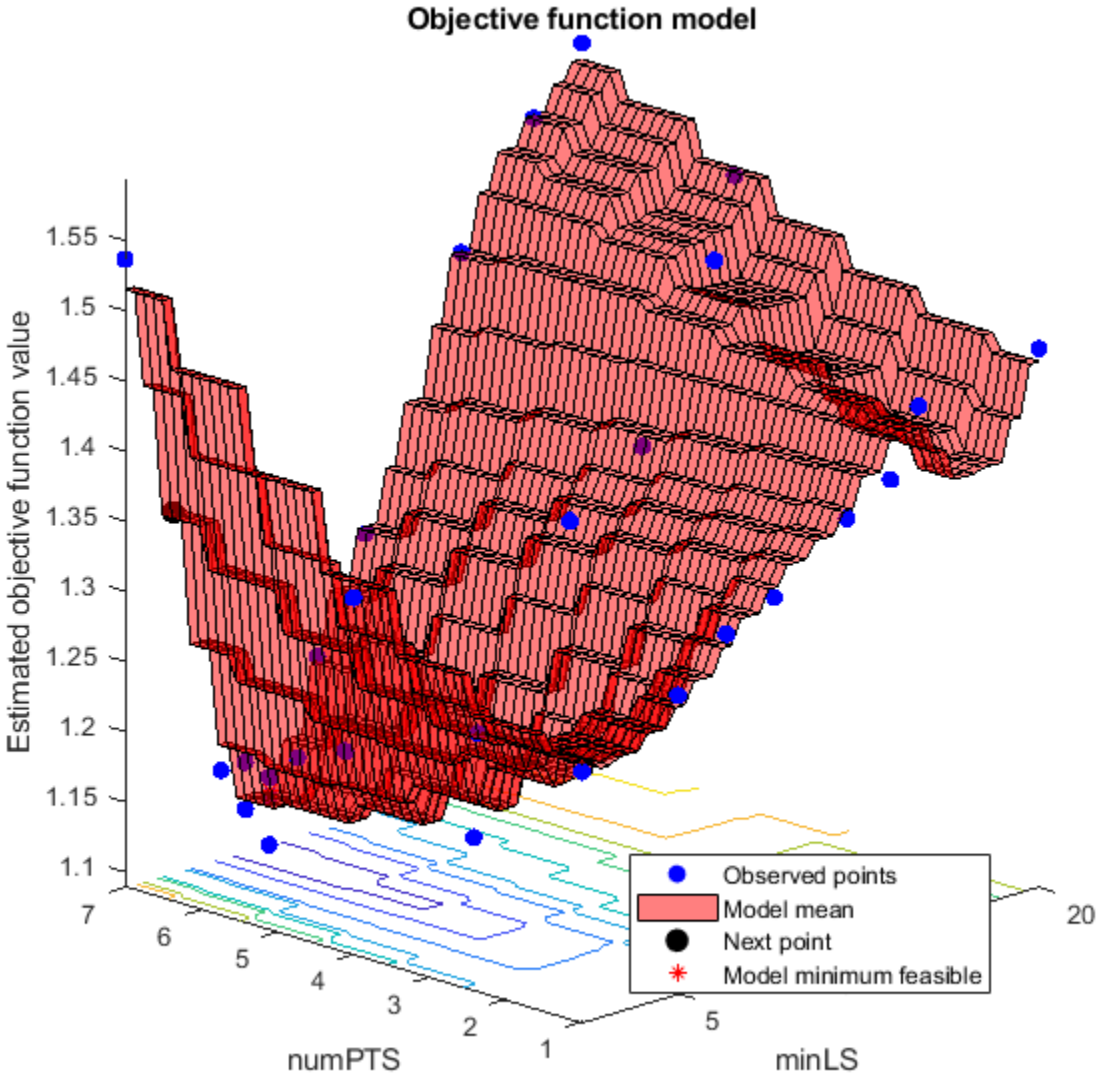
- Estimate the out-of-bag quantile error based on the median.
- Return the out-of-bag quantile error.

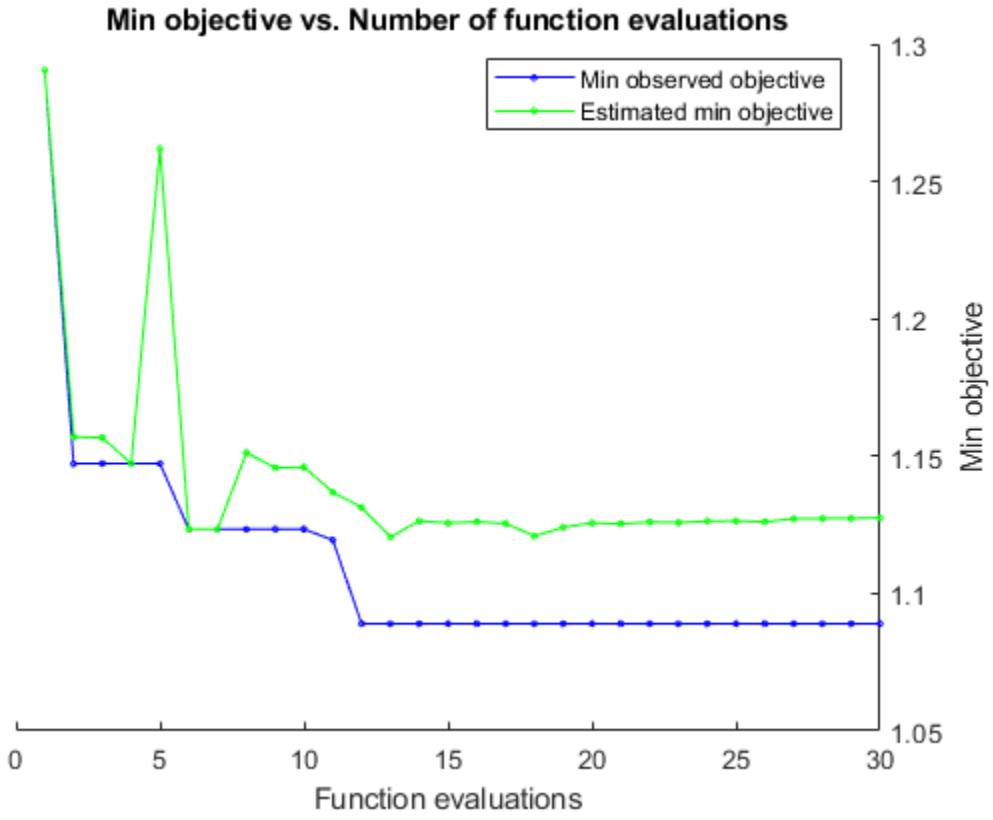
```
function oobErr = oobErrRF(params,X)
%oobErrRF Trains random forest and estimates out-of-bag quantile error
%   oobErr trains a random forest of 300 regression trees using the
%   predictor data in X and the parameter specification in params, and then
%   returns the out-of-bag quantile error based on the median. X is a table
%   and params is an array of OptimizableVariable objects corresponding to
%   the minimum leaf size and number of predictors to sample at each node.
randomForest = TreeBagger(300,X,'MPG','Method','regression',...
    'OOBPrediction','on','MinLeafSize',params.minLS,...
    'NumPredictorstoSample',params.numPTS);
oobErr = oobQuantileError(randomForest);
end
```

Minimize Objective Using Bayesian Optimization

Find the model achieving the minimal, penalized, out-of-bag quantile error with respect to tree complexity and number of predictors to sample at each node using Bayesian optimization. Specify the expected improvement plus function as the acquisition function and suppress printing the optimization information.

```
results = bayesopt(@(params)oobErrRF(params,X),hyperparametersRF,...
    'AcquisitionFunctionName','expected-improvement-plus','Verbose',0);
```


Objective function model
Estimated objective function value
1.55
1.5
1.45
1.4
1.35
1.3
1.25
1.2
1.15
1.1
7
6
5
4
3
2
1
numPTS
5
20
minLS
Observed points
Model mean
Next point
Model minimum feasible



`results` is a `BayesianOptimization` object containing, among other things, the minimum of the objective function and the optimized hyperparameter values.

Display the observed minimum of the objective function and the optimized hyperparameter values.

```
bestOOBErr = results.MinObjective
bestHyperparameters = results.XAtMinObjective
```

```
bestOOBErr =

    1.0890


bestHyperparameters =

  1×2 table

    minLS    numPTS
    _____    ______

      7        7
```

Train Model Using Optimized Hyperparameters

Train a random forest using the entire data set and the optimized hyperparameter values.

```
Mdl = TreeBagger(300,X,'MPG','Method','regression',...
    'MinLeafSize',bestHyperparameters.minLS,...
    'NumPredictorstoSample',bestHyperparameters.numPTS);
```

`Mdl` is `TreeBagger` object optimized for median prediction. You can predict the median fuel economy given predictor data by passing `Mdl` and the new data to `quantilePredict`.

See Also

`TreeBagger` | `TreeBagger` | `bayesopt` | `oobQuantileError` | `optimizableVariable`

Related Examples

- “Detect Outliers Using Quantile Regression” on page 18-137
- “Conditional Quantile Estimation Using Kernel Smoothing” on page 18-142

Support Vector Machines for Binary Classification

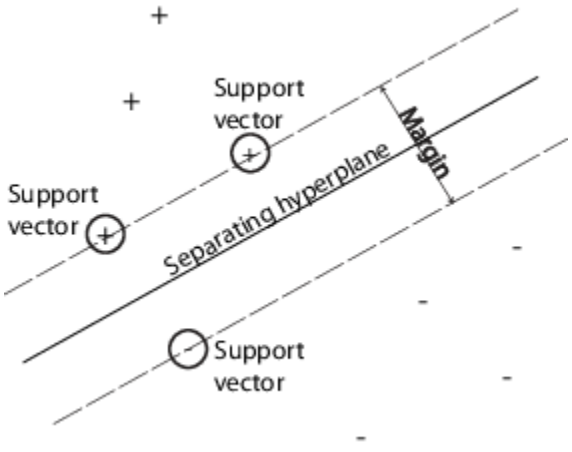
In this section...

Understanding Support Vector Machines

Separable Data

You can use a support vector machine (SVM) when your data has exactly two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.

The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. The following figure illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.



Mathematical Formulation: Primal

This discussion follows Hastie, Tibshirani, and Friedman [1] and Christianini and Shawe-Taylor [2].

The data for training is a set of points (vectors) x_j along with their categories y_j. For some dimension d, the $x_j \in R^d$, and the $y_j = \pm 1$. The equation of a hyperplane is

$$f(x) = x'\beta + b = 0$$

where $\beta \in R^d$ and b is a real number.

The following problem defines the best separating hyperplane (i.e., the decision boundary). Find β and b that minimize $||\beta||$ such that for all data points (x_j, y_j),

$$y_j f(x_j) \geq 1.$$

The support vectors are the x_j on the boundary, those for which $y_j f(x_j) = 1$.

For mathematical convenience, the problem is usually given as the equivalent problem of minimizing $\|\beta\|$. This is a quadratic programming problem. The optimal solution $\left(\widehat{\beta}, \widehat{b}\right)$ enables classification of a vector z as follows:

$$\text{class}(z) = \text{sign}\left(z'\widehat{\beta} + \widehat{b}\right) = \text{sign}\left(\widehat{f}(z)\right).$$

$\widehat{f}(z)$ is the classification score and represents the distance z is from the decision boundary.

Mathematical Formulation: Dual

It is computationally simpler to solve the dual quadratic programming problem. To obtain the dual, take positive Lagrange multipliers α_j multiplied by each constraint, and subtract from the objective function:

$$L_P = \frac{1}{2}\beta'\beta - \sum_j \alpha_j \left(y_j \left(x_j'\beta + b\right) - 1\right),$$

where you look for a stationary point of L_P over β and b. Setting the gradient of L_P to 0, you get

$$\begin{aligned} \beta &= \sum_j \alpha_j y_j x_j \\ 0 &= \sum_j \alpha_j y_j . \end{aligned} \tag{18-1}$$

Substituting into L_P, you get the dual L_D:

$$L_D = \sum_j \alpha_j - \frac{1}{2}\sum_j \sum_k \alpha_j \alpha_k y_j y_k x_j' x_k,$$

which you maximize over $\alpha_j \geq 0$. In general, many α_j are 0 at the maximum. The nonzero α_j in the solution to the dual problem define the hyperplane, as seen in “Equation 18-1”, which gives β as the sum of $\alpha_j y_j x_j$. The data points x_j corresponding to nonzero α_j are the support vectors.

The derivative of L_D with respect to a nonzero α_j is 0 at an optimum. This gives

$$y_j f(x_j) - 1 = 0.$$

In particular, this gives the value of b at the solution, by taking any j with nonzero α_j.

The dual is a standard quadratic programming problem. For example, the Optimization Toolbox `quadprog` solver solves this type of problem.

Nonseparable Data

Your data might not allow for a separating hyperplane. In that case, SVM can use a soft margin, meaning a hyperplane that separates many, but not all data points.

There are two standard formulations of soft margins. Both involve adding slack variables ξ_j and a penalty parameter C.

- The L^1-norm problem is:

  $$\min_{\beta, b, \xi}\left(\frac{1}{2}\beta'\beta + C\sum_j \xi_j\right)$$

  such that

  $$\begin{aligned} y_j f(x_j) &\geq 1 - \xi_j \\ \xi_j &\geq 0. \end{aligned}$$

  The L^1-norm refers to using ξ_j as slack variables instead of their squares. The three solver options `SMO`, `ISDA`, and `L1QP` of `fitcsvm` minimize the L^1-norm problem.
- The L^2-norm problem is:

  $$\min_{\beta, b, \xi}\left(\frac{1}{2}\beta'\beta + C\sum_j \xi_j^2\right)$$

  subject to the same constraints.

In these formulations, you can see that increasing C places more weight on the slack variables ξ_j, meaning the optimization attempts to make a stricter separation between classes. Equivalently, reducing C towards 0 makes misclassification less important.

Mathematical Formulation: Dual

For easier calculations, consider the L^1 dual problem to this soft-margin formulation. Using Lagrange multipliers μ_j, the function to minimize for the L^1-norm problem is:

$$L_P = \frac{1}{2}\beta'\beta + C\sum_j \xi_j - \sum_j \alpha_j\left(y_i f(x_j) - (1 - \xi_j)\right) - \sum_j \mu_j \xi_j,$$

where you look for a stationary point of L_P over β, b, and positive ξ_j. Setting the gradient of L_P to 0, you get

$$\begin{aligned} \beta &= \sum_j \alpha_j y_j x_j \\ \sum_j \alpha_j y_j &= 0 \\ \alpha_j &= C - \mu_j \\ \alpha_j, \mu_j, \xi_j &\geq 0. \end{aligned}$$

These equations lead directly to the dual formulation:

$$\max_{\alpha} \sum_j \alpha_j - \frac{1}{2} \sum_j \sum_k \alpha_j \alpha_k y_j y_k x_j' x_k$$

subject to the constraints

$$\sum_j y_j \alpha_j = 0$$
$$0 \le \alpha_j \le C .$$

The final set of inequalities, $0 \le \alpha_j \le C$, shows why C is sometimes called a box constraint. C keeps the allowable values of the Lagrange multipliers α_j in a "box", a bounded region.

The gradient equation for b gives the solution b in terms of the set of nonzero α_j, which correspond to the support vectors.

You can write and solve the dual of the L^2-norm problem in an analogous manner. For details, see Christianini and Shawe-Taylor [2], Chapter 6.

fitcsvm Implementation

Both dual soft-margin problems are quadratic programming problems. Internally, `fitcsvm` has several different algorithms for solving the problems.

- For one-class or binary classification, if you do not set a fraction of expected outliers in the data (see `OutlierFraction`), then the default solver is Sequential Minimal Optimization (SMO). SMO minimizes the one-norm problem by a series of two-point minimizations. During optimization, SMO respects the linear constraint $\sum_i \alpha_i y_i = 0$, and explicitly includes the bias term in the model. SMO is relatively fast. For more details on SMO, see [3].
- For binary classification, if you set a fraction of expected outliers in the data, then the default solver is the Iterative Single Data Algorithm. Like SMO, ISDA solves the one-norm problem. Unlike SMO, ISDA minimizes by a series on one-point minimizations, does not respect the linear constraint, and does not explicitly include the bias term in the model. For more details on ISDA, see [4].
- For one-class or binary classification, and if you have an Optimization Toolbox license, you can choose to use `quadprog` to solve the one-norm problem. `quadprog` uses a good deal of memory, but solves quadratic programs to a high degree of precision. For more details, see "Quadratic Programming Definition" (Optimization Toolbox).

Nonlinear Transformation with Kernels

Some binary classification problems do not have a simple hyperplane as a useful separating criterion. For those problems, there is a variant of the mathematical approach that retains nearly all the simplicity of an SVM separating hyperplane.

This approach uses these results from the theory of reproducing kernels:

- There is a class of functions $G(x_1,x_2)$ with the following property. There is a linear space S and a function φ mapping x to S such that

  $G(x_1,x_2) = <\varphi(x_1),\varphi(x_2)>$.

  The dot product takes place in the space S.

- This class of functions includes:
  - Polynomials: For some positive integer p,

    $G(x_1,x_2) = (1 + x_1'x_2)^p$.
  - Radial basis function (Gaussian):

    $G(x_1,x_2) = \exp(-\|x_1-x_2)\|^2)$.
  - Multilayer perceptron or sigmoid (neural network): For a positive number p_1 and a negative number p_2,

    $G(x_1,x_2) = \tanh(p_1x_1'x_2 + p_2)$.

    Note

    - Not every set of p_1 and p_2 yields a valid reproducing kernel.
    - `fitcsvm` does not support the sigmoid kernel. Instead, you can define the sigmoid kernel and specify it by using the `'KernelFunction'` name-value pair argument. For details, see "Train SVM Classifier Using Custom Kernel" on page 18-159.

The mathematical approach using kernels relies on the computational method of hyperplanes. All the calculations for hyperplane classification use nothing more than dot products. Therefore, nonlinear kernels can use identical calculations and solution algorithms, and obtain classifiers that are nonlinear. The resulting classifiers are hypersurfaces in some space S, but the space S does not have to be identified or examined.

Using Support Vector Machines

As with any supervised learning model, you first train a support vector machine, and then cross validate the classifier. Use the trained machine to classify (predict) new data. In addition, to obtain satisfactory predictive accuracy, you can use various SVM kernel functions, and you must tune the parameters of the kernel functions.

Training an SVM Classifier

Train, and optionally cross validate, an SVM classifier using `fitcsvm`. The most common syntax is:

```
SVMModel = fitcsvm(X,Y,'KernelFunction','rbf',...
    'Standardize',true,'ClassNames',{'negClass','posClass'});
```

The inputs are:

- `X` — Matrix of predictor data, where each row is one observation, and each column is one predictor.
- `Y` — Array of class labels with each row corresponding to the value of the corresponding row in `X`. `Y` can be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors.

- `KernelFunction` — The default value is `'linear'` for two-class learning, which separates the data by a hyperplane. The value `'gaussian'` (or `'rbf'`) is the default for one-class learning, and specifies to use the Gaussian (or radial basis function) kernel. An important step to successfully train an SVM classifier is to choose an appropriate kernel function.
- `Standardize` — Flag indicating whether the software should standardize the predictors before training the classifier.
- `ClassNames` — Distinguishes between the negative and positive classes, or specifies which classes to include in the data. The negative class is the first element (or row of a character array), e.g., `'negClass'`, and the positive class is the second element (or row of a character array), e.g., `'posClass'`. `ClassNames` must be the same data type as `Y`. It is good practice to specify the class names, especially if you are comparing the performance of different classifiers.

The resulting, trained model (`SVMModel`) contains the optimized parameters from the SVM algorithm, enabling you to classify new data.

For more name-value pairs you can use to control the training, see the `fitcsvm` reference page.

Classifying New Data with an SVM Classifier

Classify new data using `predict`. The syntax for classifying new data using a trained SVM classifier (`SVMModel`) is:

```
[label,score] = predict(SVMModel,newX);
```

The resulting vector, `label`, represents the classification of each row in `X`. `score` is an *n*-by-2 matrix of soft scores. Each row corresponds to a row in `X`, which is a new observation. The first column contains the scores for the observations being classified in the negative class, and the second column contains the scores observations being classified in the positive class.

To estimate posterior probabilities rather than scores, first pass the trained SVM classifier (`SVMModel`) to `fitPosterior`, which fits a score-to-posterior-probability transformation function to the scores. The syntax is:

```
ScoreSVMModel = fitPosterior(SVMModel,X,Y);
```

The property `ScoreTransform` of the classifier `ScoreSVMModel` contains the optimal transformation function. Pass `ScoreSVMModel` to `predict`. Rather than returning the scores, the output argument `score` contains the posterior probabilities of an observation being classified in the negative (column 1 of `score`) or positive (column 2 of `score`) class.

Tuning an SVM Classifier

Use the `'OptimizeHyperparameters'` name-value pair argument of `fitcsvm` to find parameter values that minimize the cross-validation loss. The eligible parameters are `'BoxConstraint'`, `'KernelFunction'`, `'KernelScale'`, `'PolynomialOrder'`, and `'Standardize'`. For an example, see “Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 18-163. Alternatively, you can use the `bayesopt` function, as shown in “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45. The `bayesopt` function allows more flexibility to customize optimization. You can use the `bayesopt` function to optimize any parameters, including parameters that are not eligible to optimize when you use the `fitcsvm` function.

You can also try tuning parameters of your classifier manually according to this scheme:

1. Pass the data to `fitcsvm`, and set the name-value pair argument `'KernelScale','auto'`. Suppose that the trained SVM model is called `SVMModel`. The software uses a heuristic

procedure to select the kernel scale. The heuristic procedure uses subsampling. Therefore, to reproduce results, set a random number seed using `rng` before training the classifier.

2. Cross validate the classifier by passing it to `crossval`. By default, the software conducts 10-fold cross validation.
3. Pass the cross-validated SVM model to `kfoldLoss` to estimate and retain the classification error.
4. Retrain the SVM classifier, but adjust the `'KernelScale'` and `'BoxConstraint'` name-value pair arguments.

   - `BoxConstraint` — One strategy is to try a geometric sequence of the box constraint parameter. For example, take 11 values, from `1e-5` to `1e5` by a factor of 10. Increasing `BoxConstraint` might decrease the number of support vectors, but also might increase training time.
   - `KernelScale` — One strategy is to try a geometric sequence of the RBF sigma parameter scaled at the original kernel scale. Do this by:

     a. Retrieving the original kernel scale, e.g., `ks`, using dot notation: `ks = SVMModel.KernelParameters.Scale`.

     b. Use as new kernel scales factors of the original. For example, multiply `ks` by the 11 values `1e-5` to `1e5`, increasing by a factor of 10.

Choose the model that yields the lowest classification error. You might want to further refine your parameters to obtain better accuracy. Start with your initial parameters and perform another cross-validation step, this time using a factor of 1.2.

Train SVM Classifiers Using a Gaussian Kernel

This example shows how to generate a nonlinear classifier with Gaussian kernel function. First, generate one class of points inside the unit disk in two dimensions, and another class of points in the annulus from radius 1 to radius 2. Then, generates a classifier based on the data with the Gaussian radial basis function kernel. The default linear classifier is obviously unsuitable for this problem, since the model is circularly symmetric. Set the box constraint parameter to `Inf` to make a strict classification, meaning no misclassified training points. Other kernel functions might not work with this strict box constraint, since they might be unable to provide a strict classification. Even though the rbf classifier can separate the classes, the result can be overtrained.

Generate 100 points uniformly distributed in the unit disk. To do so, generate a radius r as the square root of a uniform random variable, generate an angle t uniformly in $(0, 2\pi)$, and put the point at $(r\cos(t), r\sin(t))$.

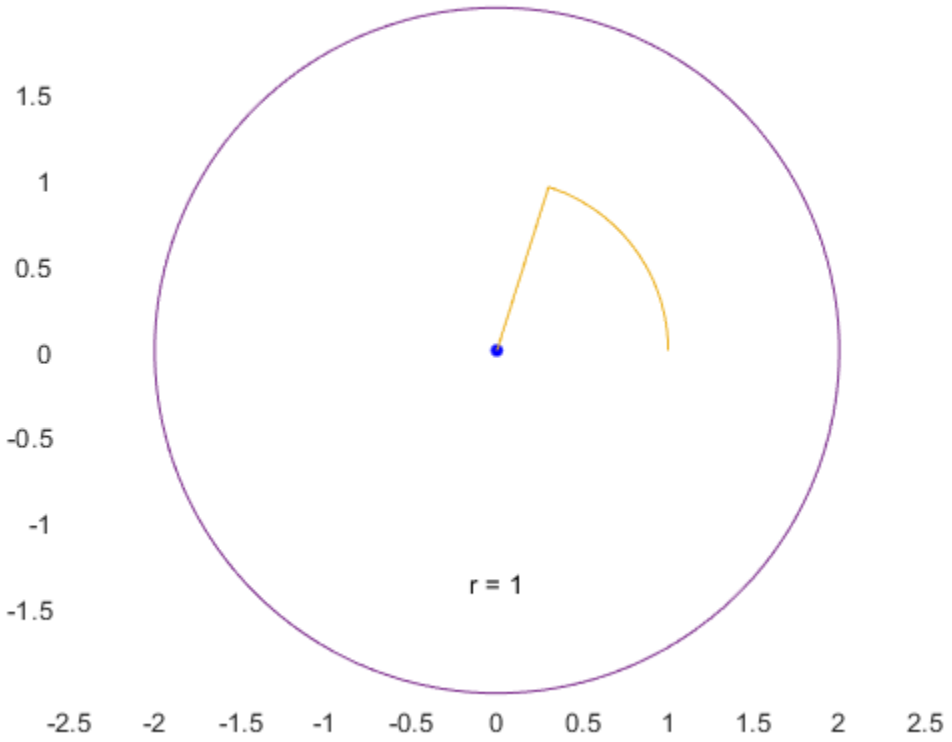
```
rng(1); % For reproducibility
r = sqrt(rand(100,1)); % Radius
t = 2*pi*rand(100,1);  % Angle
data1 = [r.*cos(t), r.*sin(t)]; % Points
```

Generate 100 points uniformly distributed in the annulus. The radius is again proportional to a square root, this time a square root of the uniform distribution from 1 through 4.

```
r2 = sqrt(3*rand(100,1)+1); % Radius
t2 = 2*pi*rand(100,1);      % Angle
data2 = [r2.*cos(t2), r2.*sin(t2)]; % points
```

Plot the points, and plot circles of radii 1 and 2 for comparison.

```
figure;
plot(data1(:,1),data1(:,2),'r.','MarkerSize',15)
hold on
plot(data2(:,1),data2(:,2),'b.','MarkerSize',15)
ezpolar(@(x)1);ezpolar(@(x)2);
axis equal
hold off
```



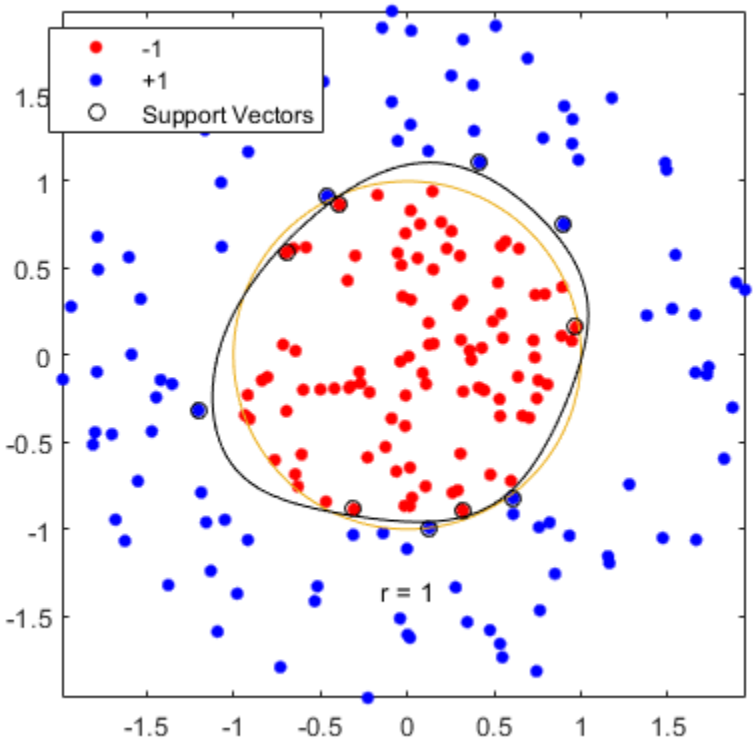
Put the data in one matrix, and make a vector of classifications.

```
data3 = [data1;data2];
theclass = ones(200,1);
theclass(1:100) = -1;
```

Train an SVM classifier with `KernelFunction` set to `'rbf'` and `BoxConstraint` set to `Inf`. Plot the decision boundary and flag the support vectors.

```
%Train the SVM Classifier
cl = fitcsvm(data3,theclass,'KernelFunction','rbf',...
    'BoxConstraint',Inf,'ClassNames',[-1,1]);

% Predict scores over the grid
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(data3(:,1)):d:max(data3(:,1)),...
    min(data3(:,2)):d:max(data3(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];
[~,scores] = predict(cl,xGrid);
```

```
% Plot the data and the decision boundary
figure;
h(1:2) = gscatter(data3(:,1),data3(:,2),theclass,'rb','.');
hold on
ezpolar(@(x)1);
h(3) = plot(data3(cl.IsSupportVector,1),data3(cl.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');
legend(h,{'-1','+1','Support Vectors'});
axis equal
hold off
```

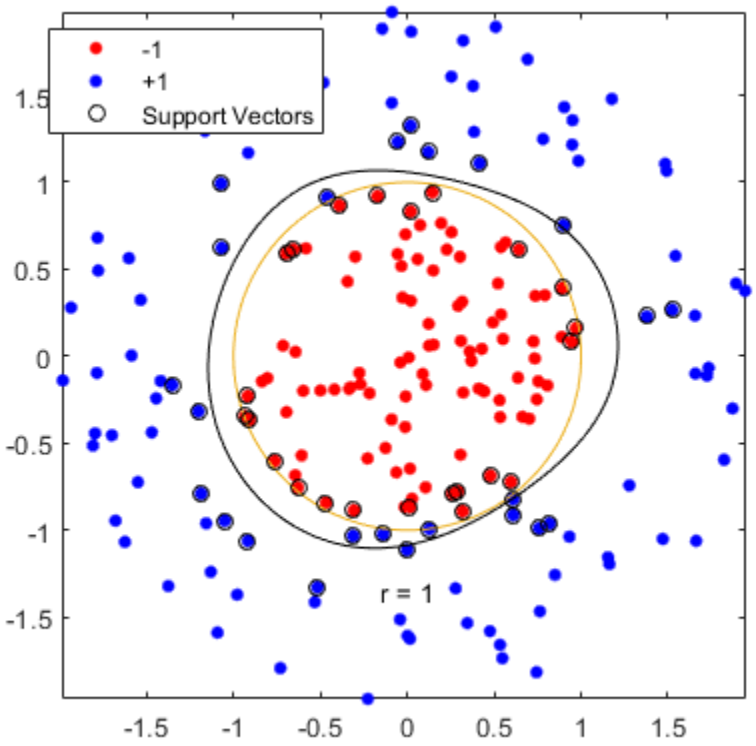


`fitcsvm` generates a classifier that is close to a circle of radius 1. The difference is due to the random training data.

Training with the default parameters makes a more nearly circular classification boundary, but one that misclassifies some training data. Also, the default value of `BoxConstraint` is `1`, and, therefore, there are more support vectors.

```
cl2 = fitcsvm(data3,theclass,'KernelFunction','rbf');
[~,scores2] = predict(cl2,xGrid);

figure;
h(1:2) = gscatter(data3(:,1),data3(:,2),theclass,'rb','.');
hold on
ezpolar(@(x)1);
h(3) = plot(data3(cl2.IsSupportVector,1),data3(cl2.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores2(:,2),size(x1Grid)),[0 0],'k');
```

```
legend(h,{'-1','+1','Support Vectors'});
axis equal
hold off
```



Train SVM Classifier Using Custom Kernel

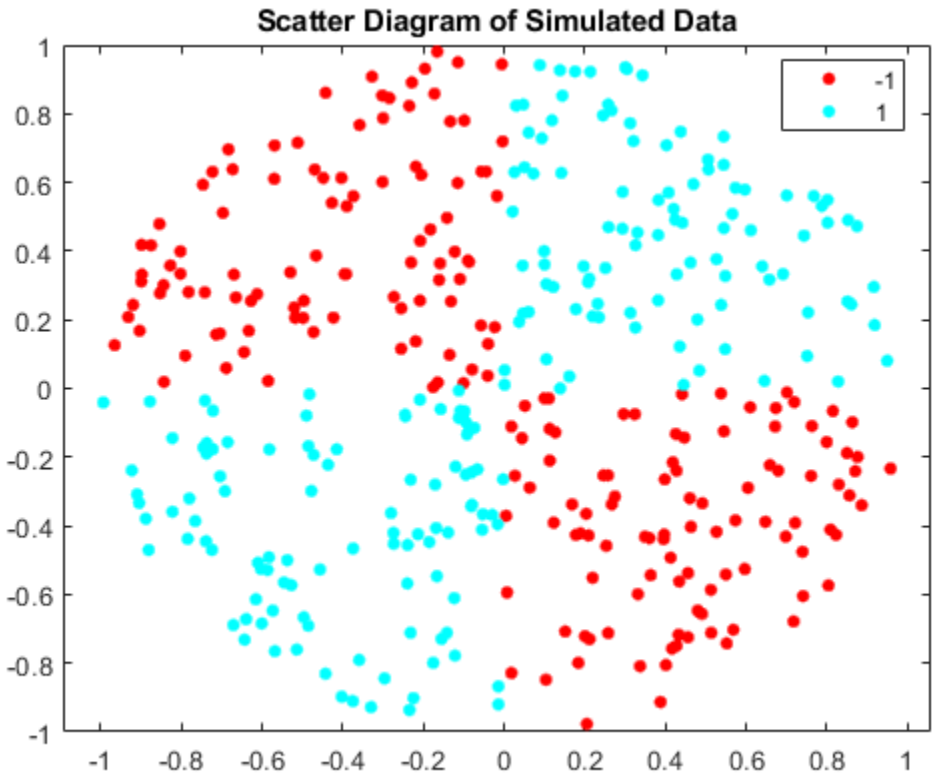
This example shows how to use a custom kernel function, such as the sigmoid kernel, to train SVM classifiers, and adjust custom kernel function parameters.

Generate a random set of points within the unit circle. Label points in the first and third quadrants as belonging to the positive class, and those in the second and fourth quadrants in the negative class.

```
rng(1);  % For reproducibility
n = 100; % Number of points per quadrant

r1 = sqrt(rand(2*n,1));                     % Random radii
t1 = [pi/2*rand(n,1); (pi/2*rand(n,1)+pi)]; % Random angles for Q1 and Q3
X1 = [r1.*cos(t1) r1.*sin(t1)];             % Polar-to-Cartesian conversion

r2 = sqrt(rand(2*n,1));
t2 = [pi/2*rand(n,1)+pi/2; (pi/2*rand(n,1)-pi/2)]; % Random angles for Q2 and Q4
X2 = [r2.*cos(t2) r2.*sin(t2)];

X = [X1; X2];        % Predictors
Y = ones(4*n,1);
Y(2*n + 1:end) = -1; % Labels
```

Plot the data.

```
figure;
gscatter(X(:,1),X(:,2),Y);
title('Scatter Diagram of Simulated Data')
```



Write a function that accepts two matrices in the feature space as inputs, and transforms them into a Gram matrix using the sigmoid kernel.

```
function G = mysigmoid(U,V)
% Sigmoid kernel function with slope gamma and intercept c
gamma = 1;
c = -1;
G = tanh(gamma*U*V' + c);
end
```

Save this code as a file named `mysigmoid` on your MATLAB® path.

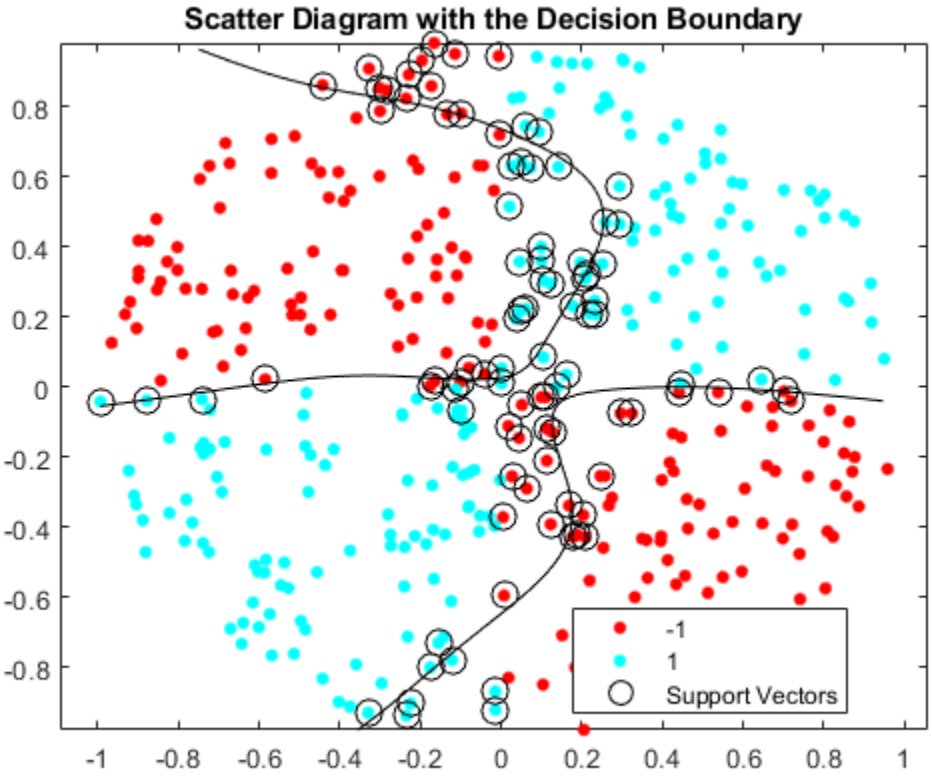
Train an SVM classifier using the sigmoid kernel function. It is good practice to standardize the data.

```
Mdl1 = fitcsvm(X,Y,'KernelFunction','mysigmoid','Standardize',true);
```

`Mdl1` is a `ClassificationSVM` classifier containing the estimated parameters.

Plot the data, and identify the support vectors and the decision boundary.

```
    % Compute the scores over a grid
d = 0.02; % Step size of the grid
[x1Grid,x2Grid] = meshgrid(min(X(:,1)):d:max(X(:,1)),...
    min(X(:,2)):d:max(X(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];        % The grid
[~,scores1] = predict(Mdl1,xGrid); % The scores

figure;
h(1:2) = gscatter(X(:,1),X(:,2),Y);
hold on
h(3) = plot(X(Mdl1.IsSupportVector,1),...
    X(Mdl1.IsSupportVector,2),'ko','MarkerSize',10);
    % Support vectors
contour(x1Grid,x2Grid,reshape(scores1(:,2),size(x1Grid)),[0 0],'k');
    % Decision boundary
title('Scatter Diagram with the Decision Boundary')
legend({'-1','1','Support Vectors'},'Location','Best');
hold off
```



You can adjust the kernel parameters in an attempt to improve the shape of the decision boundary. This might also decrease the within-sample misclassification rate, but, you should first determine the out-of-sample misclassification rate.

Determine the out-of-sample misclassification rate by using 10-fold cross validation.

```
CVMdl1 = crossval(Mdl1);
misclass1 = kfoldLoss(CVMdl1);
misclass1
```

```
misclass1 =

    0.1350
```

The out-of-sample misclassification rate is 13.5%.

Write another sigmoid function, but Set `gamma = 0.5;`.

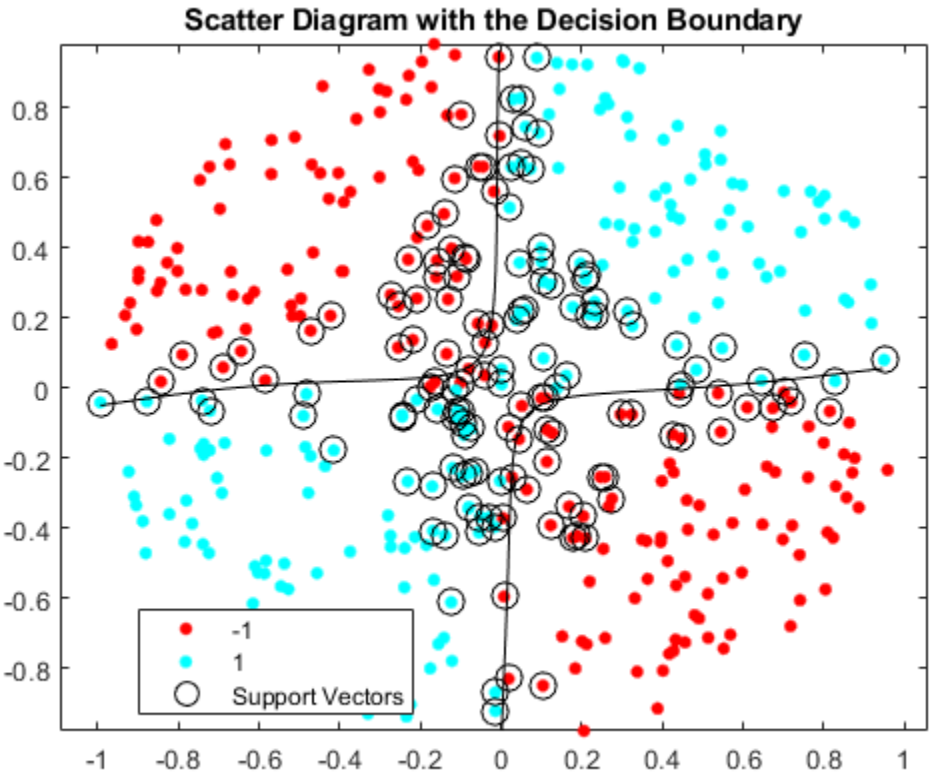
```
function G = mysigmoid2(U,V)
% Sigmoid kernel function with slope gamma and intercept c
gamma = 0.5;
c = -1;
G = tanh(gamma*U*V' + c);
end
```

Save this code as a file named `mysigmoid2` on your MATLAB® path.

Train another SVM classifier using the adjusted sigmoid kernel. Plot the data and the decision region, and determine the out-of-sample misclassification rate.

```
Mdl2 = fitcsvm(X,Y,'KernelFunction','mysigmoid2','Standardize',true);
[~,scores2] = predict(Mdl2,xGrid);

figure;
h(1:2) = gscatter(X(:,1),X(:,2),Y);
hold on
h(3) = plot(X(Mdl2.IsSupportVector,1),...
    X(Mdl2.IsSupportVector,2),'ko','MarkerSize',10);
title('Scatter Diagram with the Decision Boundary')
contour(x1Grid,x2Grid,reshape(scores2(:,2),size(x1Grid)),[0 0],'k');
legend({'-1','1','Support Vectors'},'Location','Best');
hold off

CVMdl2 = crossval(Mdl2);
misclass2 = kfoldLoss(CVMdl2);
misclass2
```

```
misclass2 =

    0.0450
```



After the sigmoid slope adjustment, the new decision boundary seems to provide a better within-sample fit, and the cross-validation rate contracts by more than 66%.

Optimize an SVM Classifier Fit Using Bayesian Optimization

This example shows how to optimize an SVM classification using the `fitcsvm` function and `OptimizeHyperparameters` name-value pair. The classification works on locations of points from a Gaussian mixture model. In *The Elements of Statistical Learning*, Hastie, Tibshirani, and Friedman (2009), page 17 describes the model. The model begins with generating 10 base points for a "green" class, distributed as 2-D independent normals with mean (1,0) and unit variance. It also generates 10 base points for a "red" class, distributed as 2-D independent normals with mean (0,1) and unit variance. For each class (green and red), generate 100 random points as follows:

1. Choose a base point *m* of the appropriate color uniformly at random.
2. Generate an independent random point with 2-D normal distribution with mean *m* and variance I/5, where I is the 2-by-2 identity matrix. In this example, use a variance I/50 to show the advantage of optimization more clearly.

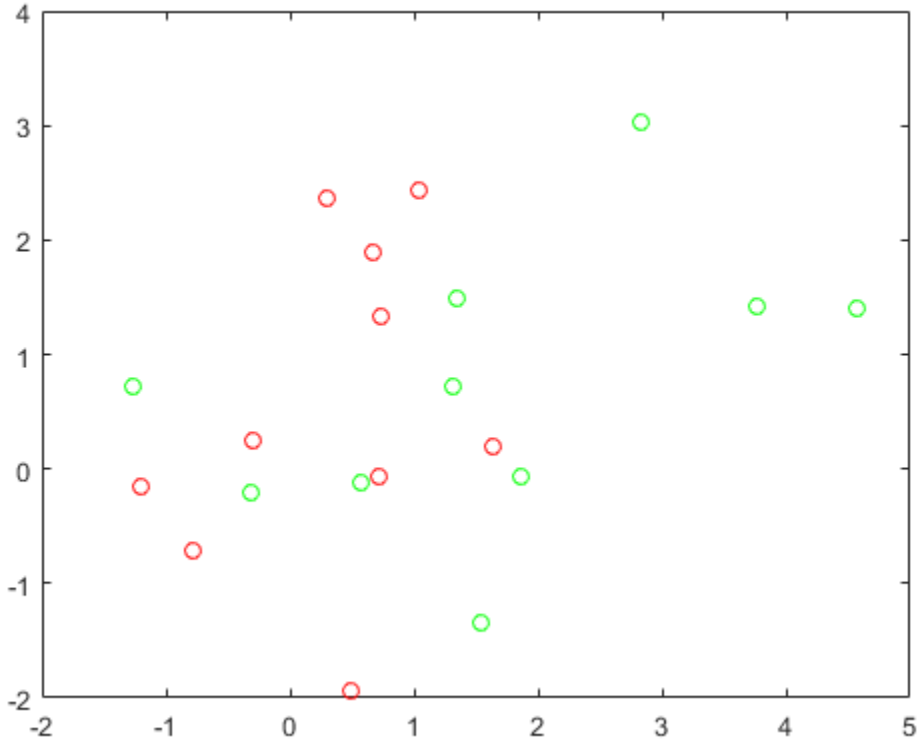
Generate the Points and Classifier

Generate the 10 base points for each class.

```
rng default % For reproducibility
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
```

View the base points.

```
plot(grnpop(:,1),grnpop(:,2),'go')
hold on
plot(redpop(:,1),redpop(:,2),'ro')
hold off
```



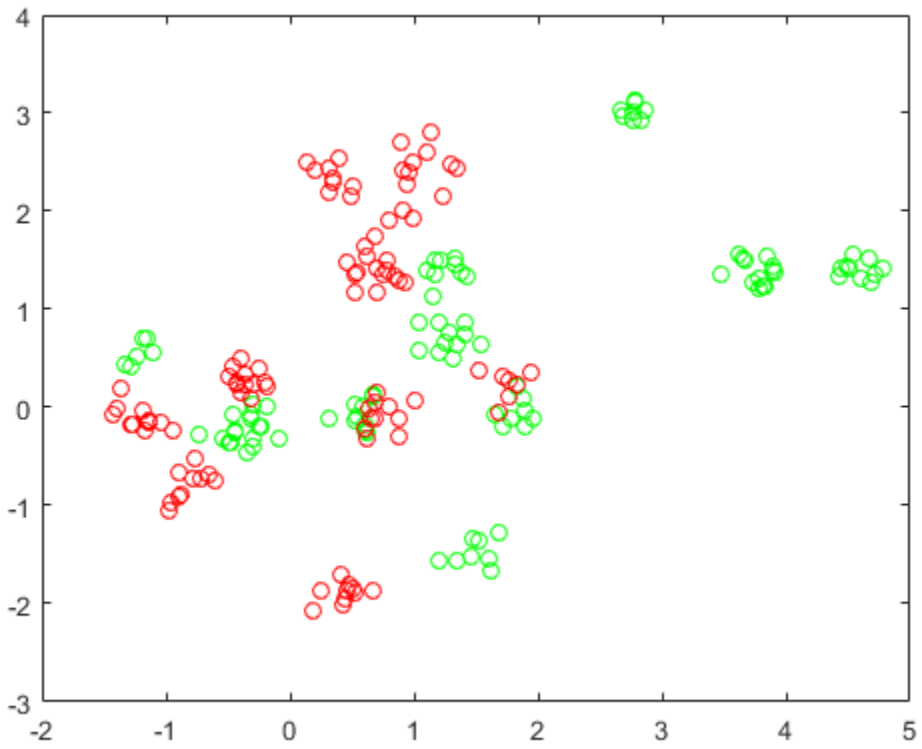
Since some red base points are close to green base points, it can be difficult to classify the data points based on location alone.

Generate the 100 data points of each class.

```
redpts = zeros(100,2);grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
```

View the data points.

```
figure
plot(grnpts(:,1),grnpts(:,2),'go')
hold on
plot(redpts(:,1),redpts(:,2),'ro')
hold off
```



Prepare Data For Classification

Put the data into one matrix, and make a vector `grp` that labels the class of each point.

```
cdata = [grnpts;redpts];
grp = ones(200,1);
% Green label 1, red label -1
grp(101:200) = -1;
```

Prepare Cross-Validation

Set up a partition for cross-validation. This step fixes the train and test sets that the optimization uses at each step.

```
c = cvpartition(200,'KFold',10);
```

Optimize the Fit

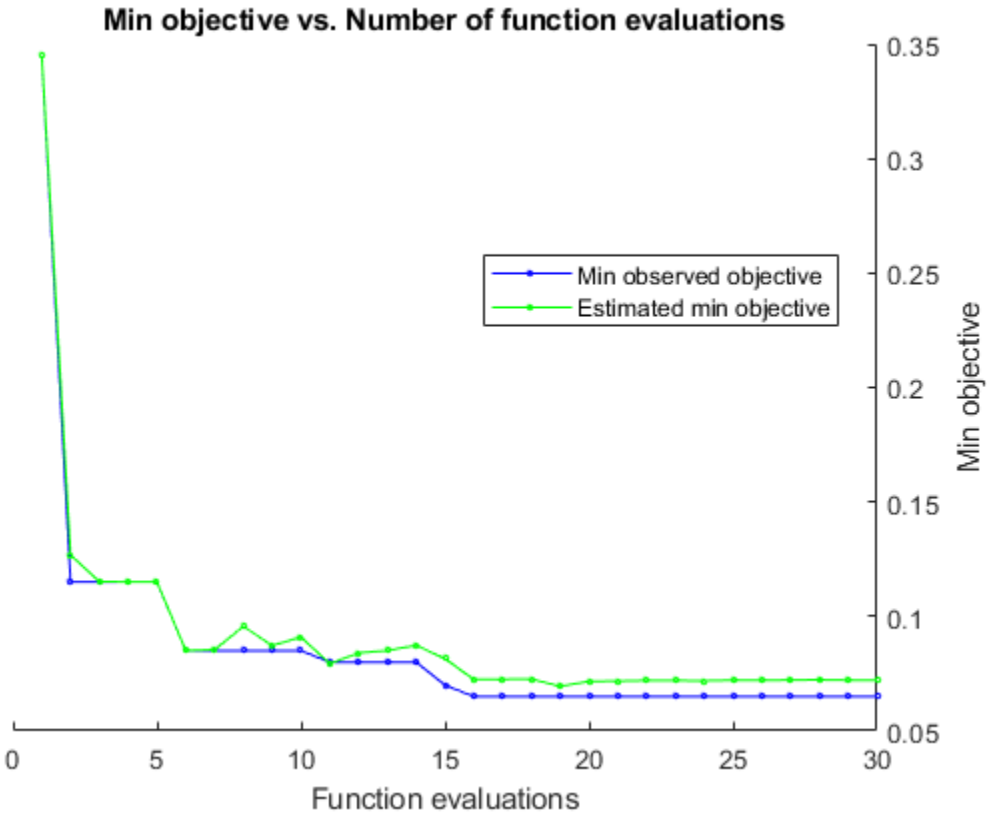
To find a good fit, meaning one with a low cross-validation loss, set options to use Bayesian optimization. Use the same cross-validation partition `c` in all optimizations.

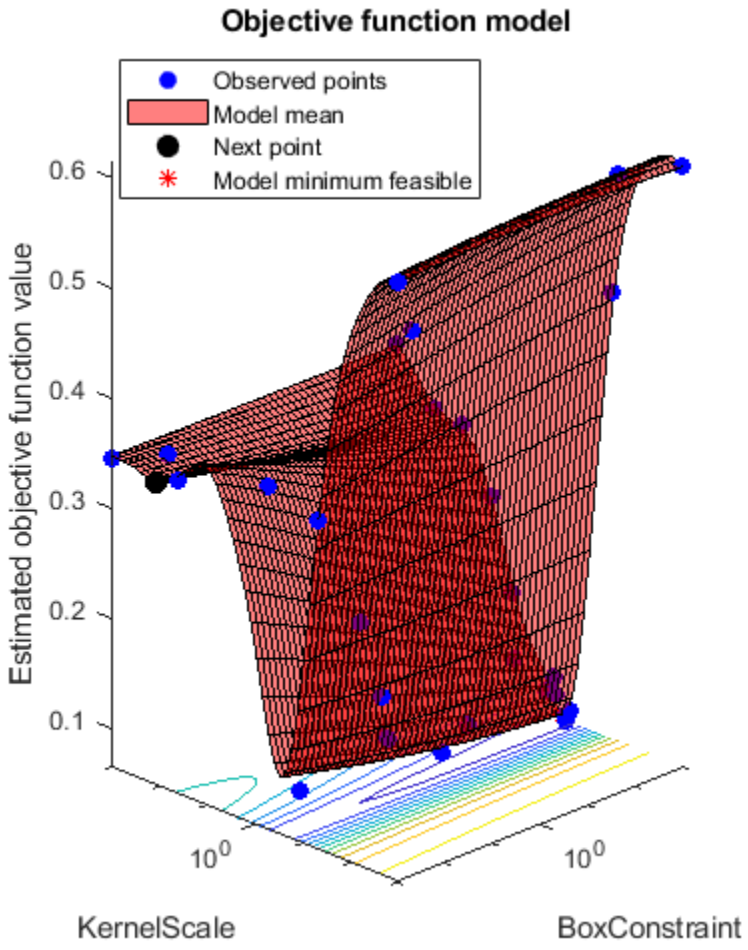
For reproducibility, use the `'expected-improvement-plus'` acquisition function.

```
opts = struct('Optimizer','bayesopt','ShowPlots',true,'CVPartition',c,...
    'AcquisitionFunctionName','expected-improvement-plus');
svmmod = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions',opts)
```

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	BoxConstraint	KernelS
1	Best	0.345	0.32188	0.345	0.345	0.00474	30
2	Best	0.115	0.21093	0.115	0.12678	430.31	1
3	Accept	0.52	0.32736	0.115	0.1152	0.028415	0.01
4	Accept	0.61	0.30316	0.115	0.11504	133.94	0.003
5	Accept	0.34	0.23884	0.115	0.11504	0.010993	5
6	Best	0.085	0.25666	0.085	0.085039	885.63	0.6
7	Accept	0.105	0.2523	0.085	0.085428	0.3057	0.5
8	Accept	0.21	0.24205	0.085	0.09566	0.16044	0.9
9	Accept	0.085	0.33802	0.085	0.08725	972.19	0.4
10	Accept	0.1	0.28635	0.085	0.090952	990.29	0
11	Best	0.08	0.24828	0.08	0.079362	2.5195	0
12	Accept	0.09	0.29983	0.08	0.08402	14.338	0.4
13	Accept	0.1	0.24832	0.08	0.08508	0.0022577	0.2
14	Accept	0.11	0.2402	0.08	0.087378	0.2115	0.3
15	Best	0.07	0.2857	0.07	0.081507	910.2	0.2
16	Best	0.065	0.29082	0.065	0.072457	953.22	0.2
17	Accept	0.075	0.26319	0.065	0.072554	998.74	0.2
18	Accept	0.295	0.26722	0.065	0.072647	996.18	44
19	Accept	0.07	0.29055	0.065	0.06946	985.37	0.2
20	Accept	0.165	0.24204	0.065	0.071622	0.065103	0.1

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	BoxConstraint	KernelS
21	Accept	0.345	0.21508	0.065	0.071764	971.7	99
22	Accept	0.61	0.24321	0.065	0.071967	0.0010168	0.001
23	Accept	0.345	0.25085	0.065	0.071959	0.0010674	99
24	Accept	0.35	0.19215	0.065	0.071863	0.0010003	40
25	Accept	0.24	0.31973	0.065	0.072124	996.55	10
26	Accept	0.61	0.27511	0.065	0.072068	958.64	0.001
27	Accept	0.47	0.25386	0.065	0.07218	993.69	0.02
28	Accept	0.3	0.23806	0.065	0.072291	993.15	17
29	Accept	0.16	0.52113	0.065	0.072104	992.81	3
30	Accept	0.365	0.20023	0.065	0.072112	0.0010017	0.04


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25
30




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 68.9913 seconds
Total objective function evaluation time: 8.1631

Best observed feasible point:
    BoxConstraint    KernelScale
    _____________    ___________

       953.22          0.26253

Observed objective function value = 0.065
Estimated objective function value = 0.073726
Function evaluation time = 0.29082

Best estimated feasible point (according to models):
    BoxConstraint    KernelScale
    _____________    ___________
```

```
       985.37          0.27389

Estimated objective function value = 0.072112
Estimated function evaluation time = 0.29025

svmmod =
  ClassificationSVM
                           ResponseName: 'Y'
                  CategoricalPredictors: []
                             ClassNames: [-1 1]
                         ScoreTransform: 'none'
                        NumObservations: 200
      HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                                  Alpha: [77x1 double]
                                   Bias: -0.2352
                        KernelParameters: [1x1 struct]
                          BoxConstraints: [200x1 double]
                         ConvergenceInfo: [1x1 struct]
                         IsSupportVector: [200x1 logical]
                                  Solver: 'SMO'


  Properties, Methods
```

Find the loss of the optimized model.

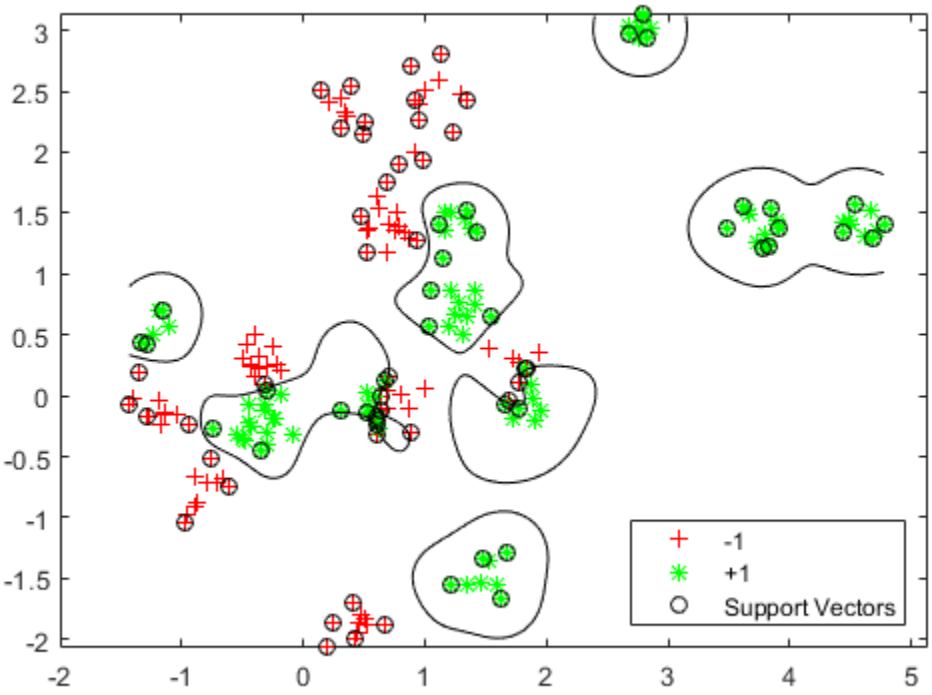
```
lossnew = kfoldLoss(fitcsvm(cdata,grp,'CVPartition',c,'KernelFunction','rbf',...
    'BoxConstraint',svmmod.HyperparameterOptimizationResults.XAtMinObjective.BoxConstraint,...
    'KernelScale',svmmod.HyperparameterOptimizationResults.XAtMinObjective.KernelScale))
```

```
lossnew = 0.0650
```

This loss is the same as the loss reported in the optimization output under "Observed objective function value".

Visualize the optimized classifier.

```
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(cdata(:,1)):d:max(cdata(:,1)),...
    min(cdata(:,2)):d:max(cdata(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];
[~,scores] = predict(svmmod,xGrid);
figure;
h = nan(3,1); % Preallocation
h(1:2) = gscatter(cdata(:,1),cdata(:,2),grp,'rg','+*');
hold on
h(3) = plot(cdata(svmmod.IsSupportVector,1),...
    cdata(svmmod.IsSupportVector,2),'ko');
contour(x1Grid,x2Grid,reshape(scores(:,2),size(x1Grid)),[0 0],'k');
legend(h,{'-1','+1','Support Vectors'},'Location','Southeast');
axis equal
hold off
```



Plot Posterior Probability Regions for SVM Classification Models

This example shows how to predict posterior probabilities of SVM models over a grid of observations, and then plot the posterior probabilities over the grid. Plotting posterior probabilities exposes decision boundaries.

Load Fisher's iris data set. Train the classifier using the petal lengths and widths, and remove the virginica species from the data.

```
load fisheriris
classKeep = ~strcmp(species,'virginica');
X = meas(classKeep,3:4);
y = species(classKeep);
```

Train an SVM classifier using the data. It is good practice to specify the order of the classes.

```
SVMModel = fitcsvm(X,y,'ClassNames',{'setosa','versicolor'});
```

Estimate the optimal score transformation function.

```
rng(1); % For reproducibility
[SVMModel,ScoreParameters] = fitPosterior(SVMModel);
```

```
Warning: Classes are perfectly separated. The optimal score-to-posterior transformation is a step
```

```
ScoreParameters
```

```
ScoreParameters = struct with fields:
                       Type: 'step'
                 LowerBound: -0.8431
                 UpperBound: 0.6897
    PositiveClassProbability: 0.5000
```

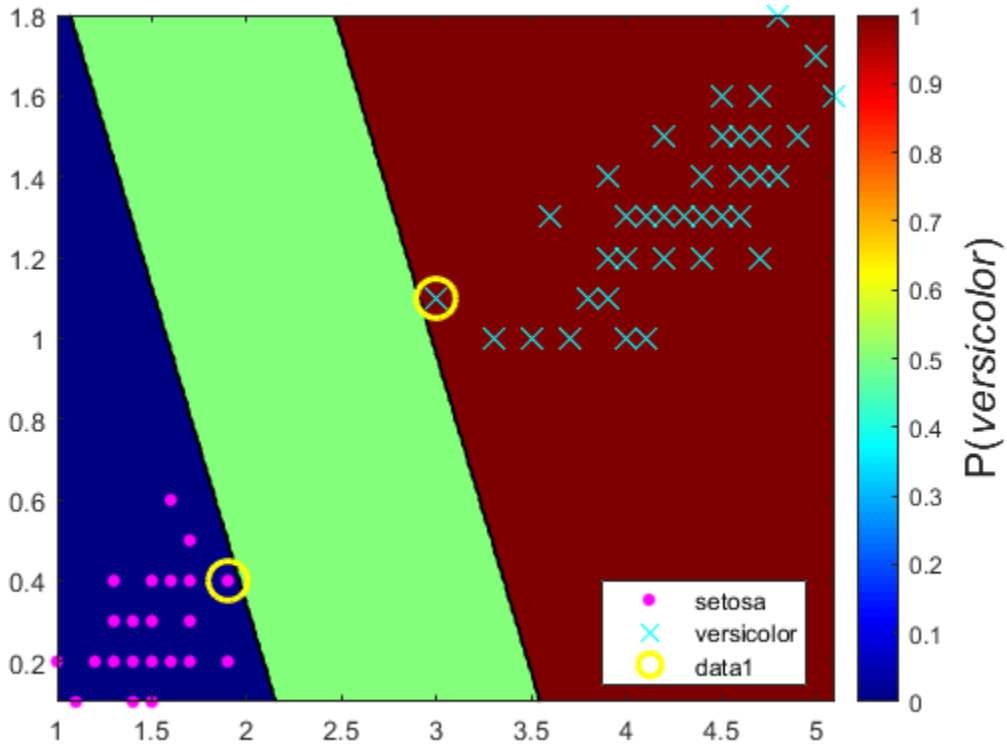
The optimal score transformation function is the step function because the classes are separable. The fields `LowerBound` and `UpperBound` of `ScoreParameters` indicate the lower and upper end points of the interval of scores corresponding to observations within the class-separating hyperplanes (the margin). No training observation falls within the margin. If a new score is in the interval, then the software assigns the corresponding observation a positive class posterior probability, i.e., the value in the `PositiveClassProbability` field of `ScoreParameters`.

Define a grid of values in the observed predictor space. Predict the posterior probabilities for each instance in the grid.

```
xMax = max(X);
xMin = min(X);
d = 0.01;
[x1Grid,x2Grid] = meshgrid(xMin(1):d:xMax(1),xMin(2):d:xMax(2));

[~,PosteriorRegion] = predict(SVMModel,[x1Grid(:),x2Grid(:)]);
```

Plot the positive class posterior probability region and the training data.

```
figure;
contourf(x1Grid,x2Grid,...
        reshape(PosteriorRegion(:,2),size(x1Grid,1),size(x1Grid,2)));
h = colorbar;
h.Label.String = 'P({\it{versicolor}})';
h.YLabel.FontSize = 16;
caxis([0 1]);
colormap jet;

hold on
gscatter(X(:,1),X(:,2),y,'mc','.x',[15,10]);
sv = X(SVMModel.IsSupportVector,:);
plot(sv(:,1),sv(:,2),'yo','MarkerSize',15,'LineWidth',2);
axis tight
hold off
```



In two-class learning, if the classes are separable, then there are three regions: one where observations have positive class posterior probability `0`, one where it is `1`, and the other where it is the positive class prior probability.

Analyze Images Using Linear Support Vector Machines

This example shows how to determine which quadrant of an image a shape occupies by training an error-correcting output codes (ECOC) model comprised of linear SVM binary learners. This example also illustrates the disk-space consumption of ECOC models that store support vectors, their labels, and the estimated α coefficients.

Create the Data Set

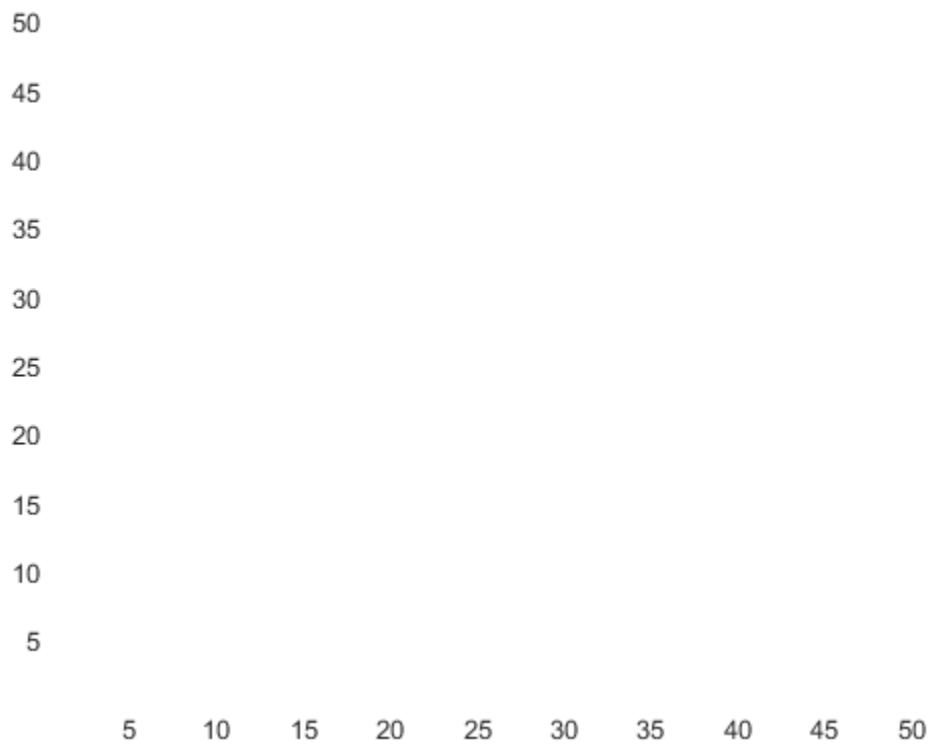
Randomly place a circle with radius five in a 50-by-50 image. Make 5000 images. Create a label for each image indicating the quadrant that the circle occupies. Quadrant 1 is in the upper right, quadrant 2 is in the upper left, quadrant 3 is in the lower left, and quadrant 4 is in the lower right. The predictors are the intensities of each pixel.

```
d = 50;  % Height and width of the images in pixels
n = 5e4; % Sample size

X = zeros(n,d^2); % Predictor matrix preallocation
Y = zeros(n,1);   % Label preallocation
theta = 0:(1/d):(2*pi);
r = 5;            % Circle radius
```

```
rng(1);           % For reproducibility

for j = 1:n
    figmat = zeros(d);                          % Empty image
    c = datasample((r + 1):(d - r - 1),2);   % Random circle center
    x = r*cos(theta) + c(1);                    % Make the circle
    y = r*sin(theta) + c(2);
    idx = sub2ind([d d],round(y),round(x));  % Convert to linear indexing
    figmat(idx) = 1;                            % Draw the circle
    X(j,:) = figmat(:);                 % Store the data
    Y(j) = (c(2) >= floor(d/2)) + 2*(c(2) < floor(d/2)) + ...
        (c(1) < floor(d/2)) + ...
        2*((c(1) >= floor(d/2)) & (c(2) < floor(d/2))); % Determine the quadrant
end
```

Plot an observation.

```
figure
imagesc(figmat)
h = gca;
h.YDir = 'normal';
title(sprintf('Quadrant %d',Y(end)))
```



Train the ECOC Model

Use a 25% holdout sample and specify the training and holdout sample indices.

```
p = 0.25;
CVP = cvpartition(Y,'Holdout',p); % Cross-validation data partition
isIdx = training(CVP);            % Training sample indices
oosIdx = test(CVP);               % Test sample indices
```

Create an SVM template that specifies storing the support vectors of the binary learners. Pass it and the training data to `fitcecoc` to train the model. Determine the training sample classification error.

```
t = templateSVM('SaveSupportVectors',true);
MdlSV = fitcecoc(X(isIdx,:),Y(isIdx),'Learners',t);
isLoss = resubLoss(MdlSV)
```

```
isLoss = 0
```

`MdlSV` is a trained `ClassificationECOC` multiclass model. It stores the training data and the support vectors of each binary learner. For large data sets, such as those in image analysis, the model can consume a lot of memory.

Determine the amount of disk space that the ECOC model consumes.

```
infoMdlSV = whos('MdlSV');
mbMdlSV = infoMdlSV.bytes/1.049e6
```

```
mbMdlSV = 763.6150
```

The model consumes 763.6 MB.

Improve Model Efficiency

You can assess out-of-sample performance. You can also assess whether the model has been overfit with a compacted model that does not contain the support vectors, their related parameters, and the training data.

Discard the support vectors and related parameters from the trained ECOC model. Then, discard the training data from the resulting model by using `compact`.

```
Mdl = discardSupportVectors(MdlSV);
CMdl = compact(Mdl);
info = whos('Mdl','CMdl');
[bytesCMdl,bytesMdl] = info.bytes;
memReduction = 1 - [bytesMdl bytesCMdl]/infoMdlSV.bytes
```

```
memReduction = 1×2

    0.0626    0.9996
```

In this case, discarding the support vectors reduces the memory consumption by about 6%. Compacting and discarding support vectors reduces the size by about 99.96%.

An alternative way to manage support vectors is to reduce their numbers during training by specifying a larger box constraint, such as 100. Though SVM models that use fewer support vectors are more desirable and consume less memory, increasing the value of the box constraint tends to increase the training time.

Remove `MdlSV` and `Mdl` from the workspace.

```
clear Mdl MdlSV
```

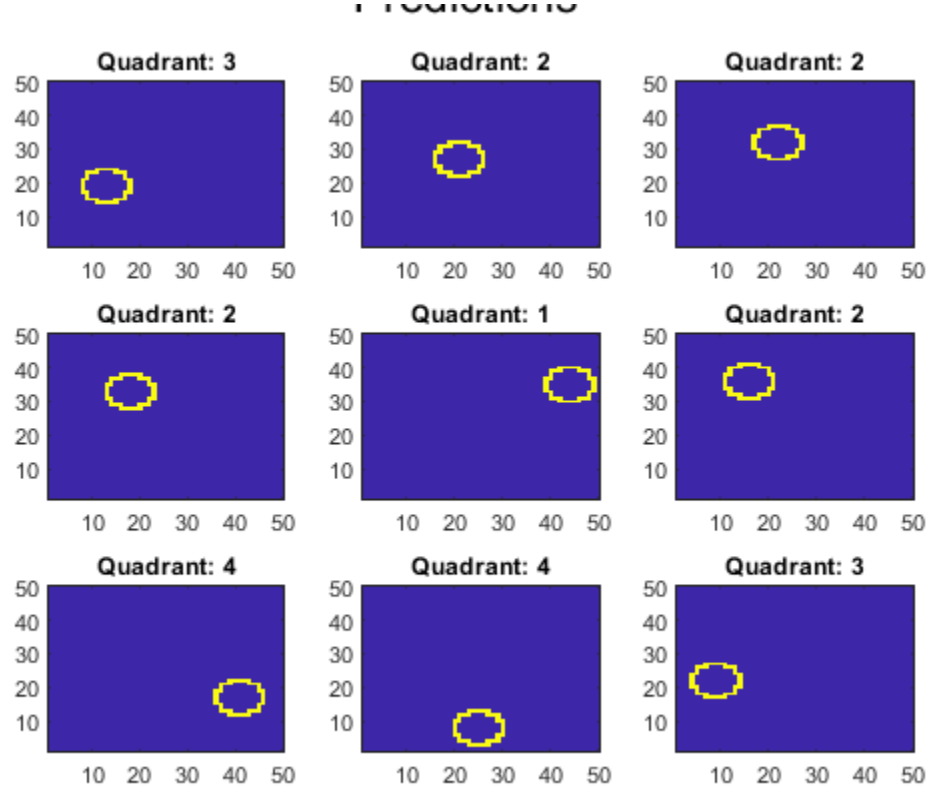
Assess Holdout Sample Performance

Calculate the classification error of the holdout sample. Plot a sample of the holdout sample predictions.

```
oosLoss = loss(CMdl,X(oosIdx,:),Y(oosIdx))
```

```
oosLoss = 0
```

```
yHat = predict(CMdl,X(oosIdx,:));
nVec = 1:size(X,1);
oosIdx = nVec(oosIdx);

figure;
for j = 1:9
    subplot(3,3,j)
    imagesc(reshape(X(oosIdx(j),:),[d d]))
    h = gca;
    h.YDir = 'normal';
    title(sprintf('Quadrant: %d',yHat(j)))
end
text(-1.33*d,4.5*d + 1,'Predictions','FontSize',17)
```



The model does not misclassify any holdout sample observations.

See Also

`bayesopt` | `fitcsvm` | `kfoldLoss`

More About

- “Train Support Vector Machines Using Classification Learner App” on page 23-98
- “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. New York: Springer, 2008.

[2] Christianini, N., and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[3] Fan, R.-E., P.-H. Chen, and C.-J. Lin. “Working set selection using second order information for training support vector machines.” *Journal of Machine Learning Research*, Vol 6, 2005, pp. 1889–1918.

[4] Kecman V., T. -M. Huang, and M. Vogt. “Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance.” In *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

Assess Neural Network Classifier Performance

Create a feedforward neural network classifier with fully connected layers using `fitcnet`. Use validation data for early stopping of the training process to prevent overfitting the model. Then, use the object functions of the classifier to assess the performance of the model on test data.

Load and Preprocess Sample Data

This example uses the 1994 census data stored in `census1994.mat`. The data set consists of demographic information from the US Census Bureau that you can use to predict whether an individual makes over $50,000 per year.

Load the sample data `census1994`, which contains the training data `adultdata` and the test data `adulttest`. Preview the first few rows of the training data set.

```
load census1994
head(adultdata)
```

```
ans=8×15 table
    age       workClass          fnlwgt      education    education_num       marital_status
    ___    _________________    __________    _________    _____________    ______________________

    39     State-gov                 77516    Bachelors         13          Never-married
    50     Self-emp-not-inc          83311    Bachelors         13          Married-civ-spouse
    38     Private                2.1565e+05    HS-grad            9          Divorced
    53     Private                2.3472e+05    11th               7          Married-civ-spouse
    28     Private                3.3841e+05    Bachelors         13          Married-civ-spouse
    37     Private                2.8458e+05    Masters           14          Married-civ-spouse
    49     Private                1.6019e+05    9th                5          Married-spouse-absent
    52     Self-emp-not-inc       2.0964e+05    HS-grad            9          Married-civ-spouse
```

Each row contains the demographic information for one adult. The last column, `salary`, shows whether a person has a salary less than or equal to $50,000 per year or greater than $50,000 per year.

Combine the `education_num` and `education` variables in both the training and test data to create a single ordered categorical variable that shows the highest level of education a person has achieved.

```
edOrder = unique(adultdata.education_num,"stable");
edCats = unique(adultdata.education,"stable");
[~,edIdx] = sort(edOrder);

adultdata.education = categorical(adultdata.education, ...
    edCats(edIdx),"Ordinal",true);
adultdata.education_num = [];

adulttest.education = categorical(adulttest.education, ...
    edCats(edIdx),"Ordinal",true);
adulttest.education_num = [];
```

Partition Training Data

Split the training data further using a stratified holdout partition. Create a separate validation data set to stop the model training process early. Reserve approximately 30% of the observations for the validation data set and use the rest of the observations to train the neural network classifier.

```
rng("default") % For reproducibility of the partition
c = cvpartition(adultdata.salary,"Holdout",0.30);
trainingIndices = training(c);
validationIndices = test(c);
tblTrain = adultdata(trainingIndices,:);
tblValidation = adultdata(validationIndices,:);
```

Train Neural Network

Train a neural network classifier by using the training set. Specify the `salary` column of `tblTrain` as the response and the `fnlwgt` column as the observation weights, and standardize the numeric predictors. Evaluate the model at each iteration by using the validation set. Specify to display the training information at each iteration by using the `Verbose` name-value argument. By default, the training process ends early if the validation cross-entropy loss is greater than or equal to the minimum validation cross-entropy loss computed so far, six times in a row. To change the number of times the validation loss is allowed to be greater than or equal to the minimum, specify the `ValidationPatience` name-value argument.

```
Mdl = fitcnet(tblTrain,"salary","Weights","fnlwgt", ...
    "Standardize",true,"ValidationData",tblValidation, ...
    "Verbose",1);
```

```
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|           1|    0.297812|    0.078920|    0.703981|    0.012127|    0.296816|           0|
|           2|    0.281110|    0.054594|    0.149850|    0.012486|    0.280132|           0|
|           3|    0.252648|    0.062041|    1.004181|    0.011339|    0.247863|           0|
|           4|    0.211868|    0.023567|    0.267214|    0.011319|    0.208988|           0|
|           5|    0.207039|    0.057528|    0.320942|    0.010288|    0.206781|           0|
|           6|    0.196838|    0.022492|    0.089842|    0.011197|    0.195583|           0|
|           7|    0.186133|    0.025551|    0.295975|    0.010426|    0.184900|           0|
|           8|    0.178779|    0.023714|    0.244525|    0.010237|    0.179370|           0|
|           9|    0.174531|    0.027149|    0.306182|    0.012911|    0.178175|           0|
|          10|    0.173217|    0.013365|    0.037475|    0.011084|    0.176371|           0|
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|          11|    0.168160|    0.016506|    0.307350|    0.010016|    0.170415|           0|
|          12|    0.164460|    0.025136|    0.473227|    0.011512|    0.165902|           0|
|          13|    0.162895|    0.014983|    0.473367|    0.010969|    0.164582|           0|
|          14|    0.160791|    0.005187|    0.113760|    0.011720|    0.162947|           0|
|          15|    0.159742|    0.004035|    0.138748|    0.010260|    0.162074|           0|
|          16|    0.159290|    0.005774|    0.108266|    0.010400|    0.161728|           0|
|          17|    0.158593|    0.004977|    0.152142|    0.010603|    0.161272|           0|
|          18|    0.157437|    0.003660|    0.193303|    0.010510|    0.160299|           0|
|          19|    0.156642|    0.007722|    0.430859|    0.010069|    0.160145|           0|
|          20|    0.155954|    0.001908|    0.121039|    0.010041|    0.159066|           0|
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|          21|    0.155824|    0.001645|    0.025159|    0.010557|    0.158992|           0|
|          22|    0.155486|    0.003232|    0.119915|    0.010829|    0.158731|           0|
|          23|    0.155398|    0.006845|    0.083105|    0.031846|    0.158963|           1|
|          24|    0.155261|    0.004374|    0.065660|    0.010762|    0.158816|           2|
```

```
|          25|    0.154955|    0.002505|    0.264106|    0.011437|    0.158687|           0|
|          26|    0.154799|    0.002183|    0.040876|    0.010903|    0.158538|           0|
|          27|    0.154466|    0.002881|    0.219478|    0.012409|    0.158033|           0|
|          28|    0.154250|    0.002724|    0.196190|    0.012062|    0.157980|           0|
|          29|    0.153918|    0.002189|    0.135392|    0.009862|    0.157605|           0|
|          30|    0.153707|    0.001449|    0.111574|    0.010851|    0.157456|           0|
|========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|========================================================================================|
|          31|    0.153214|    0.002050|    0.528628|    0.010212|    0.157379|           0|
|          32|    0.152671|    0.002542|    0.488640|    0.010013|    0.156687|           0|
|          33|    0.152303|    0.004554|    0.223206|    0.010334|    0.156778|           1|
|          34|    0.152093|    0.002856|    0.121284|    0.010188|    0.156639|           0|
|          35|    0.151871|    0.003145|    0.135909|    0.010108|    0.156446|           0|
|          36|    0.151741|    0.001441|    0.225342|    0.010452|    0.156517|           1|
|          37|    0.151626|    0.002500|    0.396782|    0.010487|    0.156429|           0|
|          38|    0.151488|    0.005053|    0.148248|    0.010312|    0.156201|           0|
|          39|    0.151250|    0.002552|    0.110278|    0.009895|    0.155968|           0|
|          40|    0.151013|    0.002506|    0.123906|    0.010837|    0.155812|           0|
|========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|========================================================================================|
|          41|    0.150821|    0.002536|    0.109515|    0.010627|    0.155742|           0|
|          42|    0.150509|    0.001418|    0.223296|    0.010561|    0.155648|           0|
|          43|    0.150340|    0.003437|    0.185351|    0.010131|    0.155435|           0|
|          44|    0.150280|    0.004746|    0.115075|    0.010432|    0.155797|           1|
|          45|    0.150194|    0.002758|    0.082143|    0.010068|    0.155575|           2|
|          46|    0.150061|    0.001122|    0.094288|    0.011405|    0.155334|           0|
|          47|    0.149978|    0.001259|    0.127677|    0.010628|    0.155305|           0|
|          48|    0.149879|    0.001523|    0.107816|    0.011331|    0.155044|           0|
|          49|    0.149749|    0.004572|    0.156869|    0.009953|    0.155043|           0|
|          50|    0.149617|    0.000965|    0.186502|    0.009702|    0.155106|           1|
|========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|========================================================================================|
|          51|    0.149579|    0.001302|    0.062687|    0.010743|    0.155160|           2|
|          52|    0.149519|    0.001407|    0.086000|    0.010335|    0.155216|           3|
|          53|    0.149405|    0.001243|    0.147530|    0.009753|    0.155309|           4|
|          54|    0.149203|    0.002749|    0.186920|    0.010267|    0.155337|           5|
|          55|    0.149040|    0.001217|    0.310011|    0.012444|    0.155460|           6|
|========================================================================================|
```

Use the information inside the `TrainingHistory` property of the object `Mdl` to check the iteration that corresponds to the minimum validation cross-entropy loss. The final returned model `Mdl` is the model trained at this iteration.

```
iteration = Mdl.TrainingHistory.Iteration;
valLosses = Mdl.TrainingHistory.ValidationLoss;
[~,minIdx] = min(valLosses);
iteration(minIdx)
```

```
ans = 49
```

Evaluate Test Set Performance

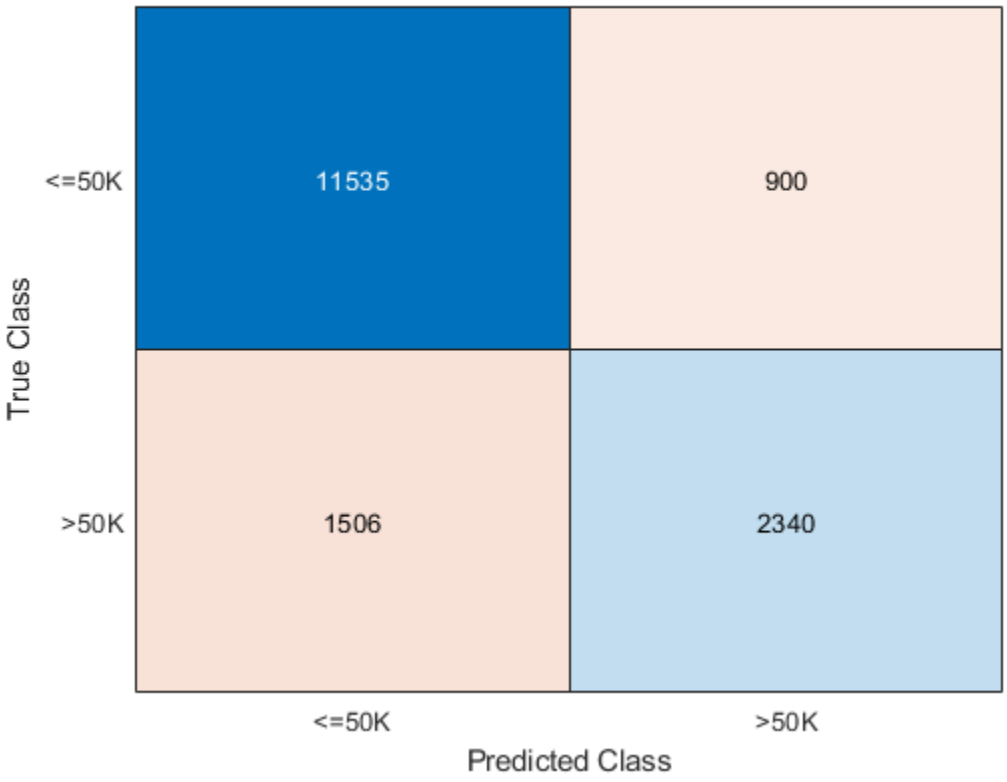
Evaluate the performance of the trained classifier `Mdl` on the test set `adulttest` by using the `predict`, `loss`, `margin`, and `edge` object functions.

Find the predicted labels and classification scores for the observations in the test set.

```
[labels,Scores] = predict(Mdl,adulttest);
```

Create a confusion matrix from the test set results. The diagonal elements indicate the number of correctly classified instances of a given class. The off-diagonal elements are instances of misclassified observations.

```
confusionchart(adulttest.salary,labels)
```



Compute the test set classification accuracy.

```
error = loss(Mdl,adulttest,"salary");
accuracy = (1-error)*100
```

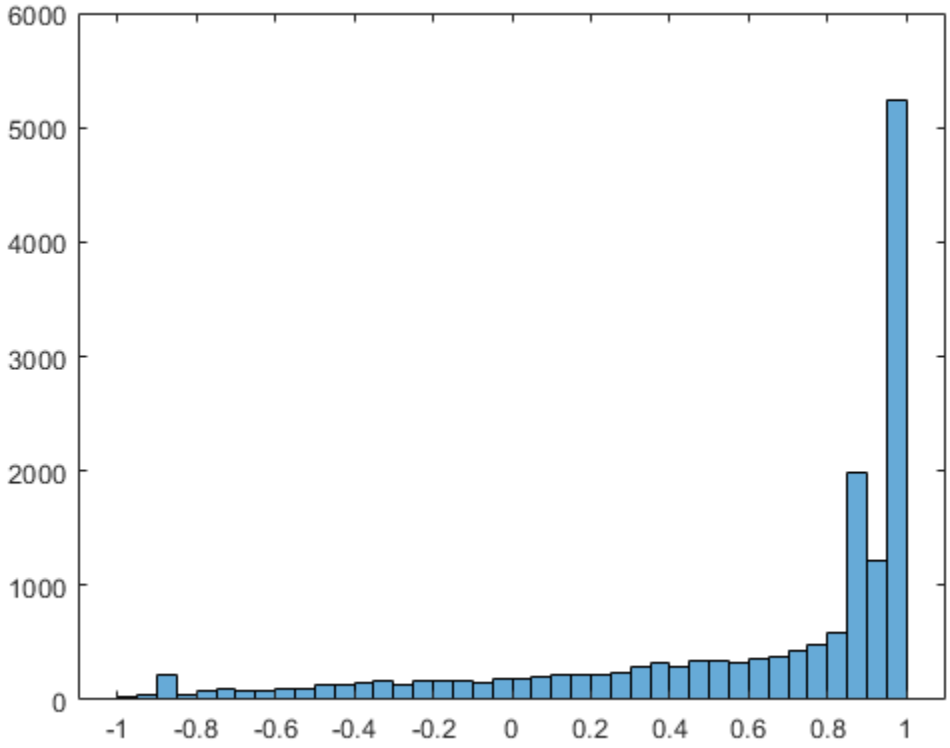
```
accuracy = 85.1306
```

The neural network classifier correctly classifies approximately 85% of the test set observations.

Compute the test set classification margins for the trained neural network. Display a histogram of the margins.

The classification margins are the difference between the classification score for the true class and the classification score for the false class. Because neural network classifiers return scores that are posterior probabilities, classification margins close to 1 indicate confident classifications and negative margin values indicate misclassifications.

```
m = margin(Mdl,adulttest,"salary");
histogram(m)
```



Use the classification edge, or mean of the classification margins, to assess the overall performance of the classifier.

```
meanMargin = edge(Mdl,adulttest,"salary")
```

```
meanMargin = 0.5983
```

Alternatively, compute the weighted classification edge by using observation weights.

```
weightedMeanMargin = edge(Mdl,adulttest,"salary", ...
    "Weight","fnlwgt")
```

```
weightedMeanMargin = 0.6072
```

Visualize the predicted labels and classification scores using scatter plots, in which each point corresponds to an observation. Use the predicted labels to set the color of the points, and use the maximum scores to set the transparency of the points. Points with less transparency are labeled with greater confidence.

First, find the maximum classification score for each test set observation.

```
maxScores = max(Scores,[],2);
```

Create a scatter plot comparing maximum scores across the number of work hours per week and level of education. Because the education variable is categorical, randomly jitter (or space out) the points along the y-dimension.

Change the colormap so that maximum scores corresponding to salaries that are less than or equal to $50,000 per year appear as blue, and maximum scores corresponding to salaries greater than $50,000 per year appear as red.

```
scatter(adulttest.hours_per_week,adulttest.education,[],labels, ...
    "filled","MarkerFaceAlpha","flat","AlphaData",maxScores, ...
    "YJitter","rand");
xlabel("Number of Work Hours Per Week")
ylabel("Education")
```

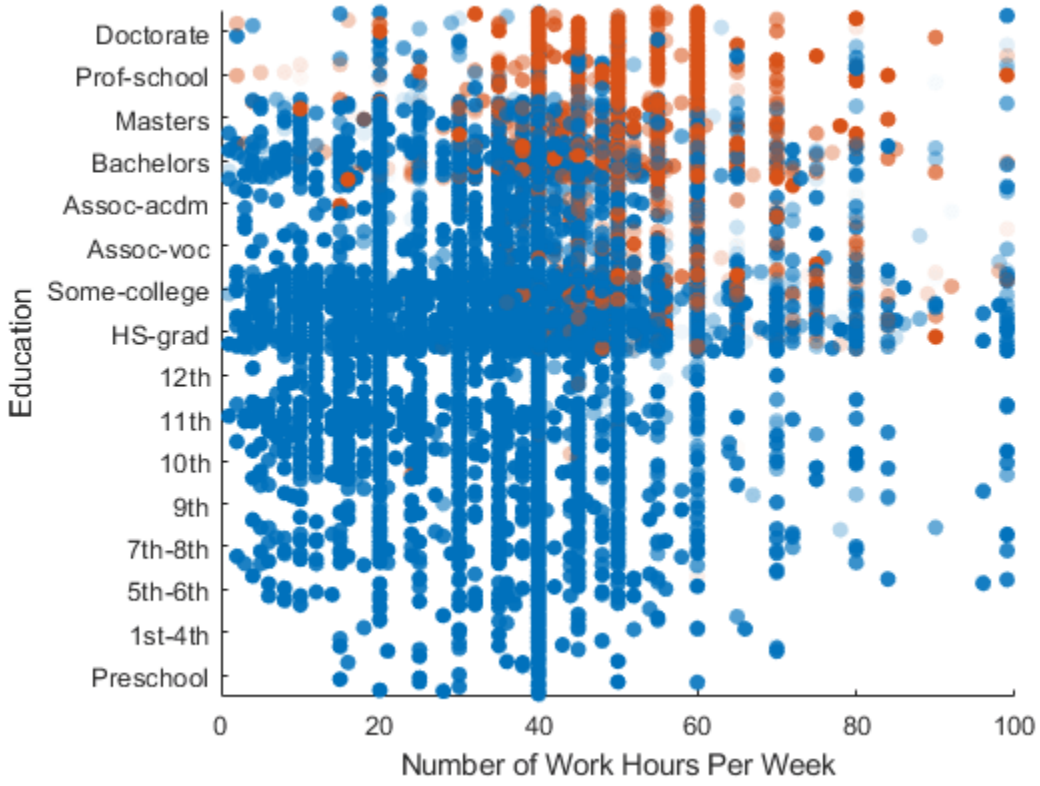
```
Mdl.ClassNames
```

```
ans = 2×1 categorical
     <=50K
     >50K
```

```
colors = lines(2)
```

```
colors = 2×3

         0    0.4470    0.7410
    0.8500    0.3250    0.0980
```

```
colormap(colors);
```



The colors in the scatter plot indicate that, in general, the neural network predicts that people with lower levels of education (12th grade or below) have salaries less than or equal to $50,000 per year. The transparency of some of the points in the lower right of the plot indicates that the model is less confident in this prediction for people who work many hours per week (60 hours or more).

See Also

`ClassificationNeuralNetwork` | `confusionchart` | `edge` | `fitcnet` | `loss` | `margin` | `predict` | `scatter`

Assess Regression Neural Network Performance

Create a feedforward regression neural network model with fully connected layers using `fitrnet`. Use validation data for early stopping of the training process to prevent overfitting the model. Then, use the object functions of the model to assess its performance on test data.

Load Sample Data

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Convert the `Origin` variable to a categorical variable. Then create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `MPG`. Each row contains the measurements for a single car.

```
Origin = categorical(cellstr(Origin));
Tbl = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,Origin,Weight,MPG);
```

Partition Data

Split the data into training, validation, and test sets. First, reserve approximately one third of the observations for the test set. Then, split the remaining data in half to create the training and validation sets.

```
rng("default") % For reproducibility of the data partitions
cvp1 = cvpartition(size(Tbl,1),"Holdout",1/3);
testTbl = Tbl(test(cvp1),:);
remainingTbl = Tbl(training(cvp1),:);

cvp2 = cvpartition(size(remainingTbl,1),"Holdout",1/2);
validationTbl = remainingTbl(test(cvp2),:);
trainTbl = remainingTbl(training(cvp2),:);
```

Train Neural Network

Train a regression neural network model by using the training set. Specify the `MPG` column of `tblTrain` as the response variable, and standardize the numeric predictors. Evaluate the model at each iteration by using the validation set. Specify to display the training information at each iteration by using the `Verbose` name-value argument. By default, the training process ends early if the validation loss is greater than or equal to the minimum validation loss computed so far, six times in a row. To change the number of times the validation loss is allowed to be greater than or equal to the minimum, specify the `ValidationPatience` name-value argument.

```
Mdl = fitrnet(trainTbl,"MPG","Standardize",true, ...
    "ValidationData",validationTbl, ...
    "Verbose",1);
```

```
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|           1|   71.063537|   22.623354|    6.466959|    0.001272|   72.648960|           0|
|           2|   48.608700|   22.384995|    1.022929|    0.001808|   43.435698|           0|
|           3|   30.584887|   13.433471|    0.537190|    0.000903|   29.134447|           0|
|           4|   17.781636|   11.159801|    1.401355|    0.000461|   16.542207|           0|
```

```
|           5|   13.075804|    4.605991|    0.419875|    0.000387|   12.946670|           0|
|           6|   11.697936|    3.197944|    0.226945|    0.000543|   12.025502|           0|
|           7|    9.494801|    2.269831|    0.751711|    0.000452|   12.596499|           1|
|           8|    8.390979|    1.970589|    0.337301|    0.000398|   11.490990|           0|
|           9|    6.853097|    1.029078|    0.866974|    0.000378|    9.449945|           0|
|          10|    6.531678|    0.924820|    0.306913|    0.000429|    9.350721|           0|
|========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|========================================================================================|
|          11|    6.152995|    1.872684|    0.457744|    0.000403|    9.223829|           0|
|          12|    5.924852|    0.718386|    0.447879|    0.000402|    9.656166|           1|
|          13|    5.792836|    0.500170|    0.216351|    0.000387|    9.733226|           2|
|          14|    5.613473|    1.151197|    0.316828|    0.000531|    9.788646|           3|
|          15|    5.415889|    1.513493|    0.327937|    0.000485|    9.607953|           4|
|          16|    5.008195|    1.398069|    1.085660|    0.000430|    9.251971|           5|
|          17|    5.004176|    2.070041|    0.890201|    0.000383|    8.719334|           0|
|          18|    4.738386|    0.483667|    0.338897|    0.000374|    8.523728|           0|
|          19|    4.680213|    0.437918|    0.107667|    0.000371|    8.369271|           0|
|          20|    4.587350|    0.510639|    0.146276|    0.000385|    8.100236|           0|
|========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|========================================================================================|
|          21|    4.479929|    0.565635|    0.228198|    0.000381|    8.062927|           0|
|          22|    4.380618|    0.892717|    0.377776|    0.000554|    7.843234|           0|
|          23|    4.189344|    0.403227|    0.362307|    0.000434|    7.834582|           0|
|          24|    4.182775|    1.150234|    1.908768|    0.000408|    9.436226|           1|
|          25|    3.985939|    0.908479|    0.518217|    0.000570|    8.973756|           2|
|          26|    3.873835|    0.826655|    0.477740|    0.000505|    8.863599|           3|
|          27|    3.830830|    0.331936|    0.220000|    0.000539|    8.574682|           4|
|          28|    3.796605|    0.232756|    0.075643|    0.000492|    8.591758|           5|
|          29|    3.706326|    0.470116|    0.249292|    0.000396|    8.517317|           6|
|========================================================================================|
```

Use the information inside the `TrainingHistory` property of the object `Mdl` to check the iteration that corresponds to the minimum validation mean squared error (MSE). The final returned model `Mdl` is the model trained at this iteration.

```
iteration = Mdl.TrainingHistory.Iteration;
valLosses = Mdl.TrainingHistory.ValidationLoss;
[~,minIdx] = min(valLosses);
iteration(minIdx)
```

```
ans = 23
```

Evaluate Test Set Performance

Evaluate the performance of the trained model `Mdl` on the test set `testTbl` by using the `loss` and `predict` object functions.

Compute the test set mean squared error (MSE). Smaller MSE values indicate better performance.

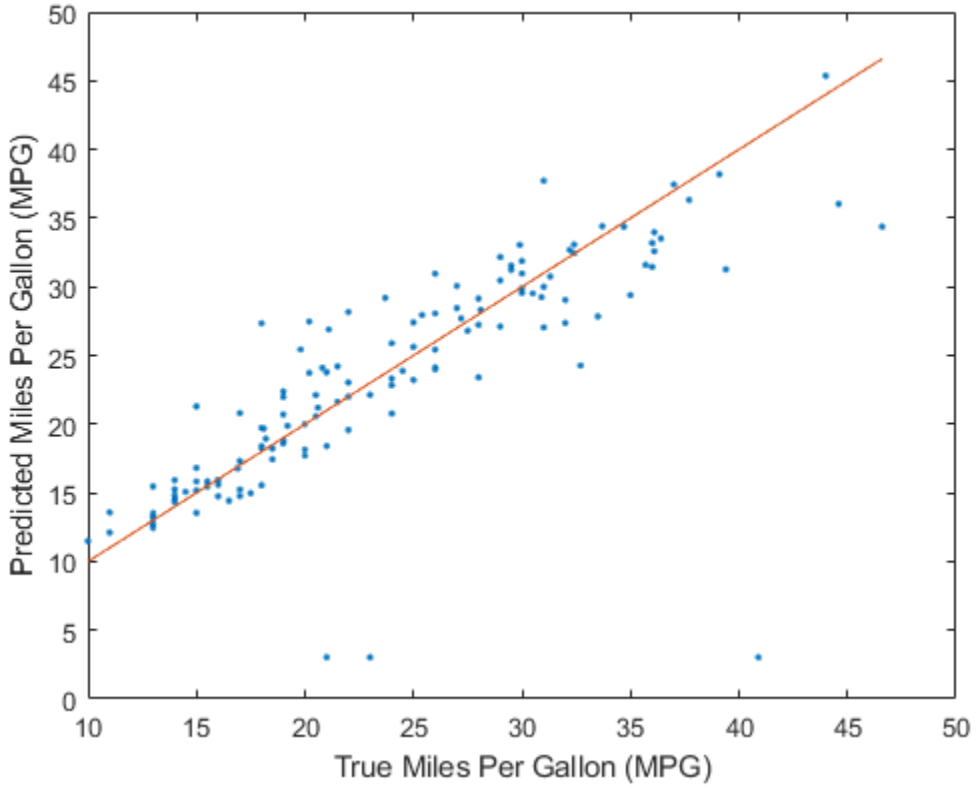
```
mse = loss(Mdl,testTbl,"MPG")
```

```
mse = 25.4145
```

Compare the predicted test set response values to the true response values. Plot the predicted miles per gallon (MPG) along the vertical axis and the true MPG along the horizontal axis. Points on the

reference line indicate correct predictions. A good model produces predictions that are scattered near the line.

```
predictedY = predict(Mdl,testTbl);

plot(testTbl.MPG,predictedY,".")
hold on
plot(testTbl.MPG,testTbl.MPG)
hold off
xlabel("True Miles Per Gallon (MPG)")
ylabel("Predicted Miles Per Gallon (MPG)")
```

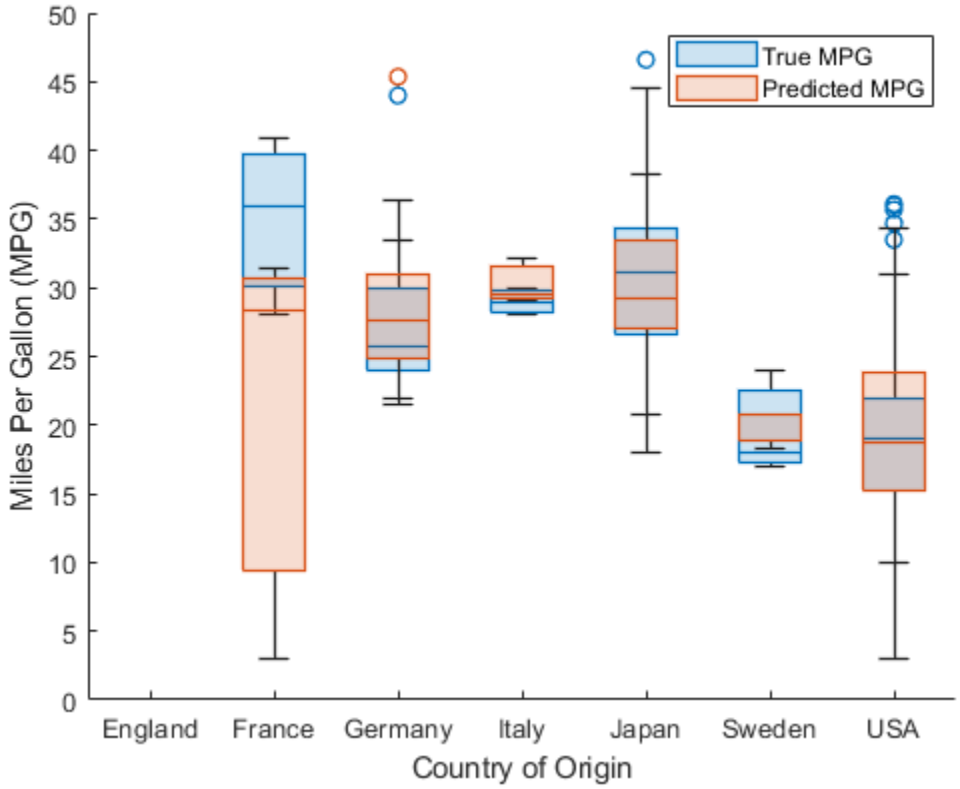


Use box plots to compare the distribution of predicted and true MPG values by country of origin. Create the box plots by using the `boxchart` function. Each box plot displays the median, the lower and upper quartiles, any outliers (computed using the interquartile range), and the minimum and maximum values that are not outliers. In particular, the line inside each box is the sample median, and the circular markers indicate outliers.

For each country of origin, compare the red box plot (showing the distribution of predicted MPG values) to the blue box plot (showing the distribution of true MPG values). Similar distributions for the predicted and true MPG values indicate good predictions.

```
boxchart(testTbl.Origin,testTbl.MPG)
hold on
boxchart(testTbl.Origin,predictedY)
hold off
legend(["True MPG","Predicted MPG"])
```

```
xlabel("Country of Origin")
ylabel("Miles Per Gallon (MPG)")
```



For most countries, the predicted and true MPG values have similar distributions. However, the neural network model tends to underestimate the MPG values for cars made in France. This discrepancy is possibly due to the small number of French cars in the training and test sets.

Compare the range of MPG values for French cars in the training and test sets.

```
trainSummary = grpstats(trainTbl(:,["MPG","Origin"]),"Origin", ...
    ["min","max"])
```

```
trainSummary=6×4 table
               Origin     GroupCount    min_MPG    max_MPG
               _______    __________    _______    _______

    France     France          3          16.2         27
    Germany    Germany        11            20       44.3
    Italy      Italy           1          37.3       37.3
    Japan      Japan          24            20       40.8
    Sweden     Sweden          3            19       21.6
    USA        USA            94             9         39
```

```
testSummary = grpstats(testTbl(:,["MPG","Origin"]),"Origin", ...
    ["min","max"])
```

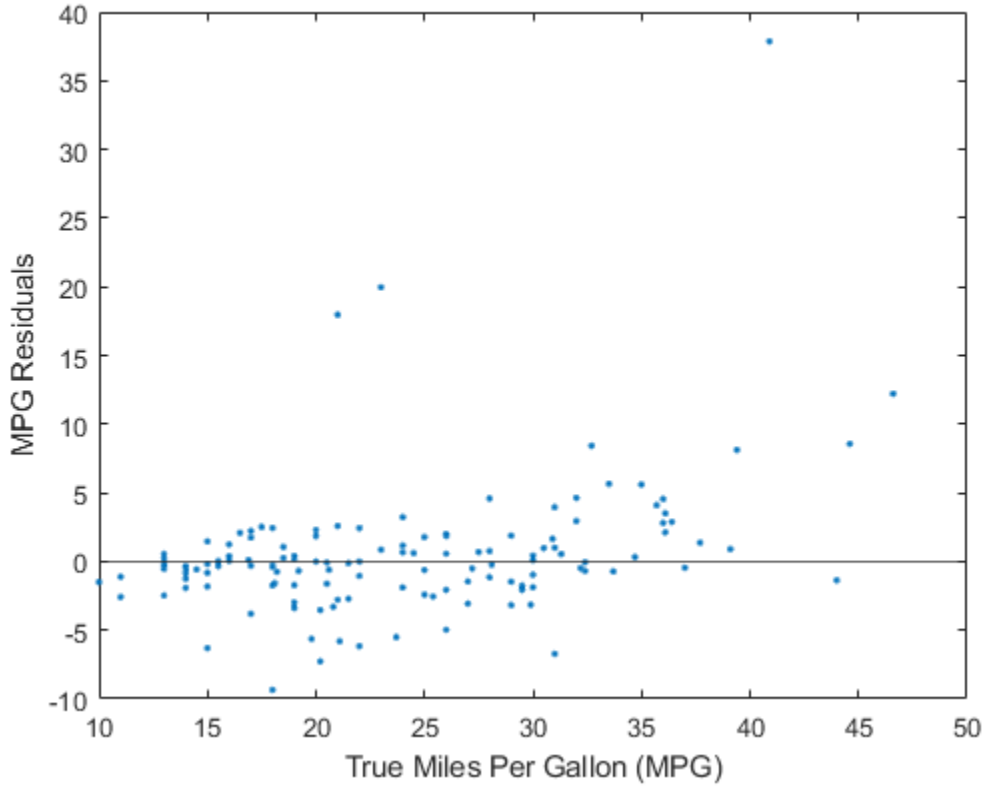
```
testSummary=6×4 table
               Origin     GroupCount    min_MPG    max_MPG
```

France	France	3	28.1	40.9
Germany	Germany	12	21.5	44
Italy	Italy	3	28	30
Japan	Japan	32	18	46.6
Sweden	Sweden	3	17	24
USA	USA	82	10	36.1

In the training set, the MPG values for cars made in France range from 16.2 to 27. However, in the test set, the MPG values for cars made in France range from 28.1 to 40.9.

Plot the test set residuals. A good model usually has residuals scattered roughly symmetrically around 0. Clear patterns in the residuals are a sign that you can improve your model.

```
residuals = testTbl.MPG - predictedY;
plot(testTbl.MPG,residuals,".")
hold on
yline(0)
hold off
xlabel("True Miles Per Gallon (MPG)")
ylabel("MPG Residuals")
```



The plot suggests that some residual values are outliers. Find out more information about the observation with the greatest residual.

```
[outlierResidual,outlierIdx] = max(residuals)
```

```
outlierResidual = 37.8727

outlierIdx = 113

testTbl(outlierIdx,:)

ans=1×7 table
    Acceleration    Displacement    Horsepower    Model_Year    Origin    Weight    MPG
    ____________    ____________    __________    __________    ______    ______    ____

        17.3             85            NaN            80        France     1835     40.9
```

The observation corresponds to a car whose `Horsepower` value is missing and whose country of origin is France, a category with few observations.

See Also

`RegressionNeuralNetwork` | `boxchart` | `fitrnet` | `loss` | `predict`

Automated Feature Engineering for Classification

The `gencfeatures` function enables you to automate the feature engineering process in the context of a machine learning workflow. Before passing tabular training data to a classifier, you can create new features from the predictors in the data by using `gencfeatures`. Use the returned data to train the classifier.

Generate new features based on your machine learning workflow.

- To generate features for an interpretable binary classifier, use the default `TargetLearner` value of `'linear'` in the call to `gencfeatures`. You can then use the returned data to train a binary linear classifier. For an example, see “Interpret Linear Model with Generated Features” on page 18-190.
- To generate features that can lead to better model accuracy, specify `'TargetLearner','bag'` in the call to `gencfeatures`. You can then use the returned data to train a bagged ensemble classifier. For an example, see “Generate New Features to Improve Bagged Ensemble Accuracy” on page 18-193.

To better understand the generated features, use the `describe` function of the `FeatureTransformer` object. To apply the same training set feature transformations to a test or validation set, use the `transform` function of the `FeatureTransformer` object.

Interpret Linear Model with Generated Features

Use automated feature engineering to generate new features. Train a linear classifier using the generated features. Interpret the relationship between the generated features and the trained model.

Load the `patients` data set. Create a table from a subset of the variables.

```
load patients
Tbl = table(Age,Diastolic,Gender,Height,SelfAssessedHealthStatus, ...
    Systolic,Weight,Smoker);
```

Generate 10 new features from the variables in `Tbl`. Specify the `Smoker` variable as the response. By default, `gencfeatures` assumes that the new features will be used to train a binary linear classifier.

```
rng("default") % For reproducibility
[T,NewTbl] = gencfeatures(Tbl,"Smoker",10)
```

```
T =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'linear'
    NumEngineeredFeatures: 10
      NumOriginalFeatures: 0
         TotalNumFeatures: 10


NewTbl=100×11 table
    zsc(Systolic.^2)    eb8(Diastolic)    q8(Systolic)    eb8(Systolic)    q8(Diastolic)    zsc(
    ________________    ______________    ____________    _____________    _____________    ____

         0.15379              8               6               4               8            -1.
         -1.9421              2               1               1               2           -0.2
```

```
    0.30311     4     6     5     5     0.5
   -0.85785     2     2     2     2     0.8
   -0.14125     3     5     4     4       1
   -0.28697     1     4     3     1     0.6
     1.0677     6     8     6     6    -0.4
    -1.1361     4     2     2     5    -0.7
    -1.1361     3     2     2     3    -0.8
   -0.71693     5     3     3     6     0.3
    -1.2734     2     1     1     2     1.2
    -1.1361     1     2     2     1       1
    0.60534     1     6     5     1    -0.9
     1.0677     8     8     6     8    -0.2
    -1.2734     3     1     1     4     0.9
     1.0677     7     8     6     8    -0.9
⋮
```

`T` is a `FeatureTransformer` object that can be used to transform new data, and `newTbl` contains the new features generated from the `Tbl` data.

To better understand the generated features, use the `describe` object function of the `FeatureTransformer` object. For example, inspect the first two generated features.

```
describe(T,1:2)
```

```
                        Type         IsOriginal    InputVariables                      T
                    ___________    ___________    ______________    ____________________________

    zsc(Systolic.^2)    Numeric          false         Systolic       power(  ,2)
                                                                    Standardization with z-sc
    eb8(Diastolic)      Categorical      false         Diastolic      Equal-width binning (numbe
```

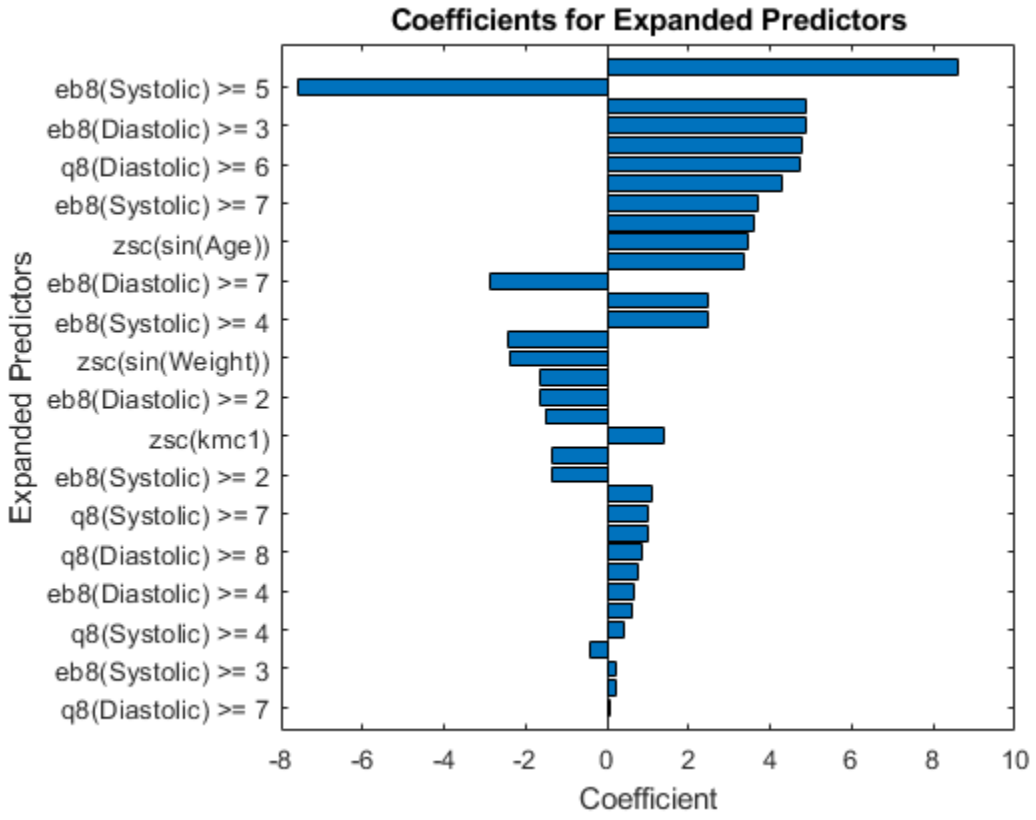
The first feature in `newTbl` is a numeric variable, created by first squaring the values of the `Systolic` variable and then converting the results to z-scores. The second feature in `newTbl` is a categorical variable, created by binning the values of the `Systolic` variable into 50 equiprobable bins.

Use the generated features to fit a linear classifier without any regularization.

```
Mdl = fitclinear(NewTbl,"Smoker","Lambda",0);
```

Plot the coefficients of the predictors used to train `Mdl`. Note that `fitclinear` expands categorical predictors before fitting a model.

```
p = length(Mdl.Beta);
[sortedCoefs,expandedIndex] = sort(Mdl.Beta,"ComparisonMethod","abs");
sortedExpandedPreds = Mdl.ExpandedPredictorNames(expandedIndex);
bar(sortedCoefs,"Horizontal","on")
yticks(1:2:p)
yticklabels(sortedExpandedPreds(1:2:end))
xlabel("Coefficient")
ylabel("Expanded Predictors")
title("Coefficients for Expanded Predictors")
```



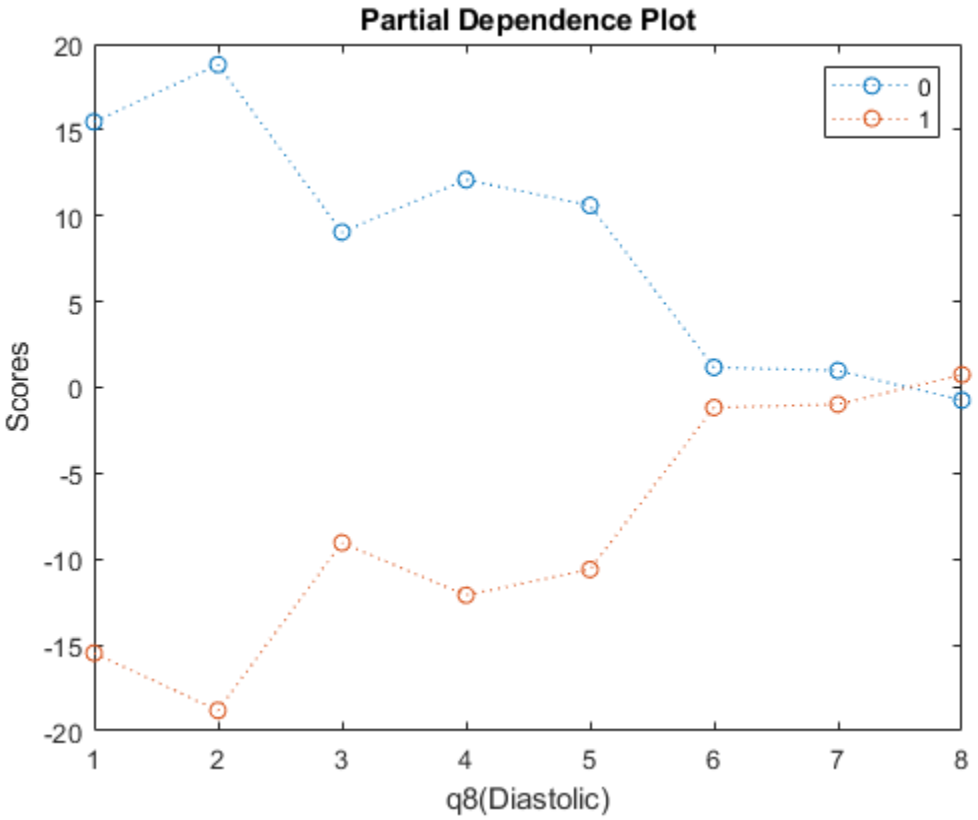
Identify the predictors whose coefficients have larger absolute values.

```
bigCoefs = abs(sortedCoefs) >= 4;
flip(sortedExpandedPreds(bigCoefs))
```

```
ans = 1×7 cell
    {'zsc(Systolic.^2)'}    {'eb8(Systolic) >= 5'}    {'q8(Diastolic) >= 3'}    {'eb8(Diastolic)
```

You can use partial dependence plots to analyze the categorical features whose levels have large coefficients in terms of absolute value. For example, inspect the partial dependence plot for the `q8(Diastolic)` variable, whose levels `q8(Diastolic) >= 3` and `q8(Diastolic) >= 6` have coefficients with large absolute values. These two levels correspond to noticeable changes in the predicted scores.

```
plotPartialDependence(Mdl,"q8(Diastolic)",Mdl.ClassNames,NewTbl);
```



Generate New Features to Improve Bagged Ensemble Accuracy

Use `gencfeatures` to engineer new features before training a bagged ensemble classifier. Before making predictions on new data, apply the same feature transformations to the new data set. Compare the test set performance of the ensemble that uses the engineered features to the test set performance of the ensemble that uses the original features.

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable("CreditRating_Historical.dat");
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____     _____     _______    ________    _____    ________    _______

    62394     0.013     0.104     0.036       0.447      0.142        3       {'BB' }
    48608     0.232     0.335     0.062       1.969      0.281        8       {'A'  }
    42444     0.311     0.367     0.074       1.935      0.366        1       {'A'  }
    48631     0.194     0.263     0.062       1.017      0.228        4       {'BBB'}
    43768     0.121     0.413     0.057       3.647      0.466       12       {'AAA'}
    39255    -0.117    -0.799      0.01       0.179      0.082        4       {'CCC'}
```

```
    62236     0.087     0.158     0.049      0.816      0.324        2        {'BBB'}
    39354     0.005     0.181     0.034      2.597      0.388        7        {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,"ID");
creditrating.Industry = categorical(creditrating.Industry);
```

Convert the `Rating` response variable to an ordinal `categorical` variable.

```
creditrating.Rating = categorical(creditrating.Rating, ...
    ["AAA","AA","A","BBB","BB","B","CCC"],"Ordinal",true);
```

Partition the data into training and test sets. Use approximately 75% of the observations as training data, and 25% of the observations as test data. Partition the data using `cvpartition`.

```
rng("default") % For reproducibility of the partition
c = cvpartition(creditrating.Rating,"Holdout",0.25);
trainingIndices = training(c); % Indices for the training set
testIndices = test(c); % Indices for the test set
creditTrain = creditrating(trainingIndices,:);
creditTest = creditrating(testIndices,:);
```

Use the training data to generate 40 new features to fit a bagged ensemble. By default, the 40 features can include original features if the software considers them to be important variables.

```
[T,newCreditTrain] = gencfeatures(creditTrain,"Rating",40, ...
    "TargetLearner","bag");
T
```

```
T =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'bag'
    NumEngineeredFeatures: 34
      NumOriginalFeatures: 6
         TotalNumFeatures: 40
```

Because `T.NumOriginalFeatures` is `6`, the function keeps all the original predictors.

Create `newCreditTest` by applying the transformations stored in the object `T` to the test data.

```
newCreditTest = transform(T,creditTest);
```

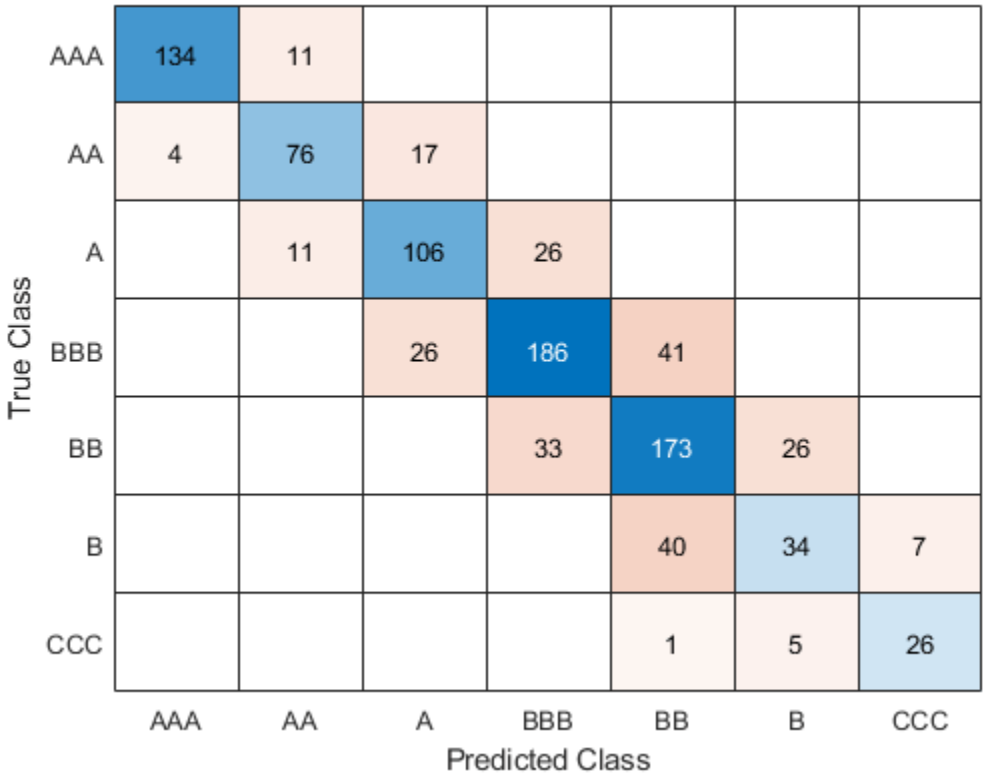
Compare the test set performances of a bagged ensemble trained on the original features and a bagged ensemble trained on the new features.

Train a bagged ensemble using the original training set `creditTrain`. Compute the accuracy of the model on the original test set `creditTest`. Visualize the results using a confusion matrix.

```
originalMdl = fitcensemble(creditTrain,"Rating","Method","Bag");
originalTestAccuracy = 1 - loss(originalMdl,creditTest, ...
    "Rating","LossFun","classiferror")
```

```
originalTestAccuracy = 0.7481
```

```
predictedTestLabels = predict(originalMdl,creditTest);
confusionchart(creditTest.Rating,predictedTestLabels);
```

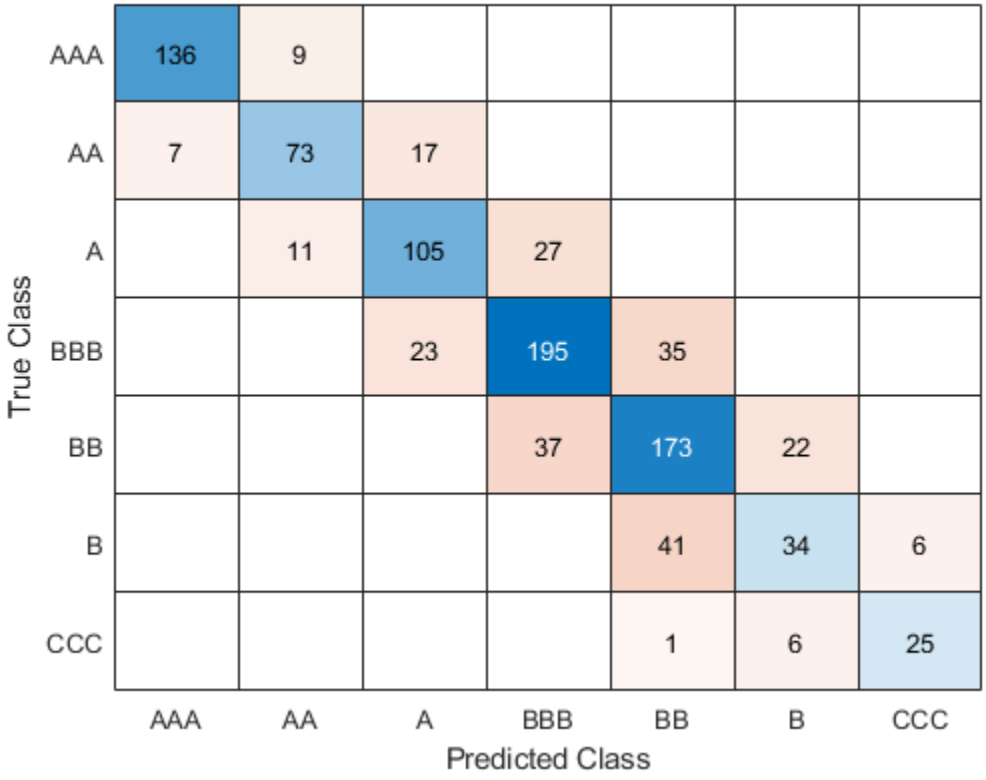


Train a bagged ensemble using the transformed training set `newCreditTrain`. Compute the accuracy of the model on the transformed test set `newCreditTest`. Visualize the results using a confusion matrix.

```
newMdl = fitcensemble(newCreditTrain,"Rating","Method","Bag");
newTestAccuracy = 1 - loss(newMdl,newCreditTest, ...
    "Rating","LossFun","classiferror")
```

```
newTestAccuracy = 0.7543
```

```
newPredictedTestLabels = predict(newMdl,newCreditTest);
confusionchart(newCreditTest.Rating,newPredictedTestLabels)
```



The bagged ensemble trained on the transformed data seems to outperform the bagged ensemble trained on the original data.

See Also

`FeatureTransformer` | `describe` | `fitcensemble` | `fitclinear` | `gencfeatures` | `plotPartialDependence` | `transform`

Moving Towards Automating Model Selection Using Bayesian Optimization

This example shows how to build multiple classification models for a given training data set, optimize their hyperparameters using Bayesian optimization, and select the model that performs the best on a test data set.

Training several models and tuning their hyperparameters can often take days or weeks. Creating a script to develop and compare multiple models automatically can be much faster. You can also use Bayesian optimization to speed up the process. Instead of training each model with different sets of hyperparameters, you select a few different models and tune their default hyperparameters using Bayesian optimization. Bayesian optimization finds an optimal set of hyperparameters for a given model by minimizing the objective function of the model. This optimization algorithm strategically selects new hyperparameters in each iteration and typically arrives at the optimal set of hyperparameters more quickly than a simple grid search. You can use the script in this example to train several classification models using Bayesian optimization for a given training data set and identify the model that performs best on a test data set.

Alternatively, to choose a classification model automatically across a selection of classifier types and hyperparameter values, use `fitcauto`. For an example, see “Automated Classifier Selection with Bayesian Optimization” on page 18-205.

Load Sample Data

This example uses the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

Load the sample data `census1994` and display the variables in the data set.

```
load census1994
whos
```

```
  Name                 Size              Bytes  Class    Attributes

  Description         20x74               2960  char
  adultdata        32561x15            1872567  table
  adulttest        16281x15             944467  table
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. For this example, to reduce the running time, subsample 5000 training and test observations each, from the original tables `adultdata` and `adulttest`, by using the `datasample` function. (You can skip this step if you want to use the complete data sets.)

```
NumSamples = 5000;
s = RandStream('mlfg6331_64'); % For reproducibility
adultdata = datasample(s,adultdata,NumSamples,'Replace',false);
adulttest = datasample(s,adulttest,NumSamples,'Replace',false);
```

Preview the first few rows of the training data set.

```
head(adultdata)
```

```
ans=8×15 table
    age     workClass       fnlwgt       education      education_num      marital_status
```

```
___    ____________    __________    _____________    _____________    ____________________    _
39     Private           4.91e+05    Bachelors             13          Never-married          Ex
25     Private         2.2022e+05    11th                   7          Never-married          Ha
24     Private         2.2761e+05    10th                   6          Divorced               Ha
51     Private         1.7329e+05    HS-grad                9          Divorced               Of
54     Private         2.8029e+05    Some-college          10          Married-civ-spouse     Sa
53     Federal-gov          39643    HS-grad                9          Widowed                Ex
52     Private              81859    HS-grad                9          Married-civ-spouse     Ma
37     Private         1.2429e+05    Some-college          10          Married-civ-spouse     Ad
```

Each row represents the attributes of one adult, such as age, education, and occupation. The last column `salary` shows whether a person has a salary less than or equal to $50,000 per year or greater than $50,000 per year.

Understand Data and Choose Classification Models

Statistics and Machine Learning Toolbox™ provides several options for classification, including classification trees, discriminant analysis, naive Bayes, nearest neighbors, support vector machines (SVMs), and classification ensembles. For the complete list of algorithms, see “Classification”.

Before choosing the algorithms to use for your problem, inspect your data set. The census data has several noteworthy characteristics:

- The data is tabular and contains both numeric and categorical variables.
- The data contains missing values.
- The response variable (`salary`) has two classes (binary classification).

Without making any assumptions or using prior knowledge of algorithms that you expect to work well on your data, you simply train all the algorithms that support tabular data and binary classification. Error-correcting output codes (ECOC) models are used for data with more than two classes. Discriminant analysis and nearest neighbor algorithms do not analyze data that contains both numeric and categorical variables. Therefore, the algorithms appropriate for this example are SVMs, a decision tree, an ensemble of decision trees, and a naive Bayes model.

Build Models and Tune Hyperparameters

To speed up the process, customize the hyperparameter optimization options. Specify `'ShowPlots'` as `false` and `'Verbose'` as 0 to disable plot and message displays, respectively. Also, specify `'UseParallel'` as `true` to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox™. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results.

```
hypopts = struct('ShowPlots',false,'Verbose',0,'UseParallel',true);
```

Start a parallel pool.

```
poolobj = gcp;
```

You can fit the training data set and tune parameters easily by calling each fitting function and setting its `'OptimizeHyperparameters'` name-value pair argument to `'auto'`. Create the classification models.

```
% SVMs: SVM with polynomial kernel & SVM with Gaussian kernel
mdls{1} = fitcsvm(adultdata,'salary','KernelFunction','polynomial','Standardize','on', ...
```

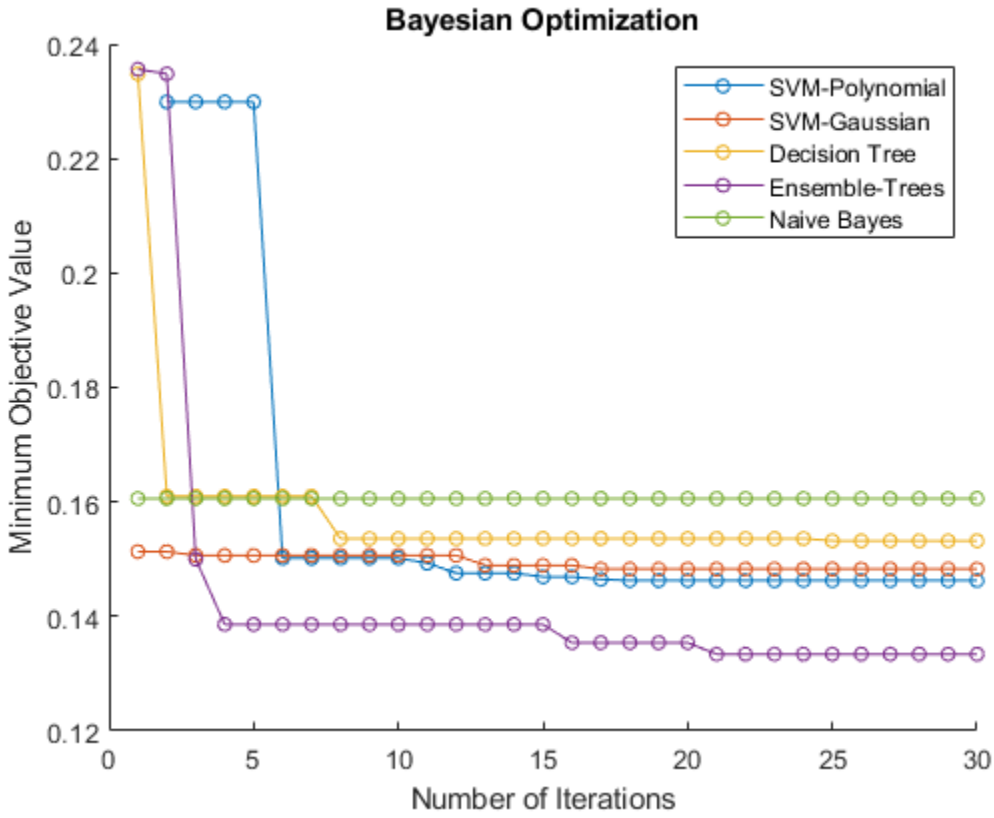
```
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions', hypopts);
mdls{2} = fitcsvm(adultdata,'salary','KernelFunction','gaussian','Standardize','on', ...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions', hypopts);

% Decision tree
mdls{3} = fitctree(adultdata,'salary', ...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions', hypopts);

% Ensemble of Decision trees
mdls{4} = fitcensemble(adultdata,'salary','Learners','tree', ...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions', hypopts);

% Naive Bayes
mdls{5} = fitcnb(adultdata,'salary', ...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions', hypopts);
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the
```

Plot Minimum Objective Curves

Extract the Bayesian optimization results from each model and plot the minimum observed value of the objective function for each model over every iteration of the hyperparameter optimization. The objective function value corresponds to the misclassification rate measured by five-fold cross-validation using the training data set. The plot compares the performance of each model.

```
figure
hold on
N = length(mdls);
for i = 1:N
    mdl = mdls{i};
    results = mdls{i}.HyperparameterOptimizationResults;
    plot(results.ObjectiveMinimumTrace,'Marker','o','MarkerSize',5);
end
names = {'SVM-Polynomial','SVM-Gaussian','Decision Tree','Ensemble-Trees','Naive Bayes'};
legend(names,'Location','northeast')
title('Bayesian Optimization')
xlabel('Number of Iterations')
ylabel('Minimum Objective Value')
```



Using Bayesian optimization to find better hyperparameter sets improves the performance of models over several iterations. In this case, the plot indicates that the ensemble of decision trees has the best prediction accuracy for the data. This model performs well consistently over several iterations and different sets of Bayesian optimization hyperparameters.

Check Performance with Test Set

Check the classifier performance with the test data set by using the confusion matrix and the receiver operating characteristic (ROC) curve.

Find the predicted labels and the score values of the test data set.

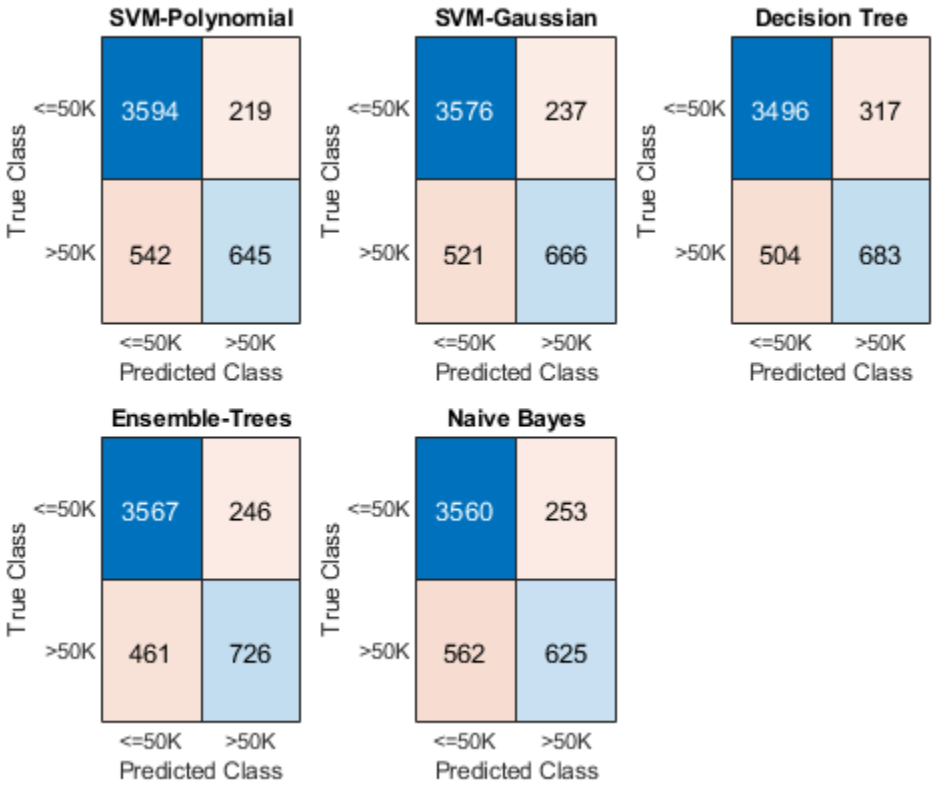
```
label = cell(N,1);
score = cell(N,1);
for i = 1:N
    [label{i},score{i}] = predict(mdls{i},adulttest);
end
```

Confusion Matrix

Obtain the most likely class for each test observation by using the `predict` function of each model. Then compute the confusion matrix with the predicted classes and the known (true) classes of the test data set by using the `confusionchart` function.

```
figure
c = cell(N,1);
for i = 1:N
```

```
    subplot(2,3,i)
    c{i} = confusionchart(adulttest.salary,label{i});
    title(names{i})
end
```



The diagonal elements indicate the number of correctly classified instances of a given class. The off-diagonal elements are instances of misclassified observations.

ROC Curve

Inspect the classifier performance more closely by plotting an ROC curve for each classifier. Use the `perfcurve` function to obtain the X and Y coordinates of the ROC curve and the area under the curve (AUC) value for the computed X and Y.

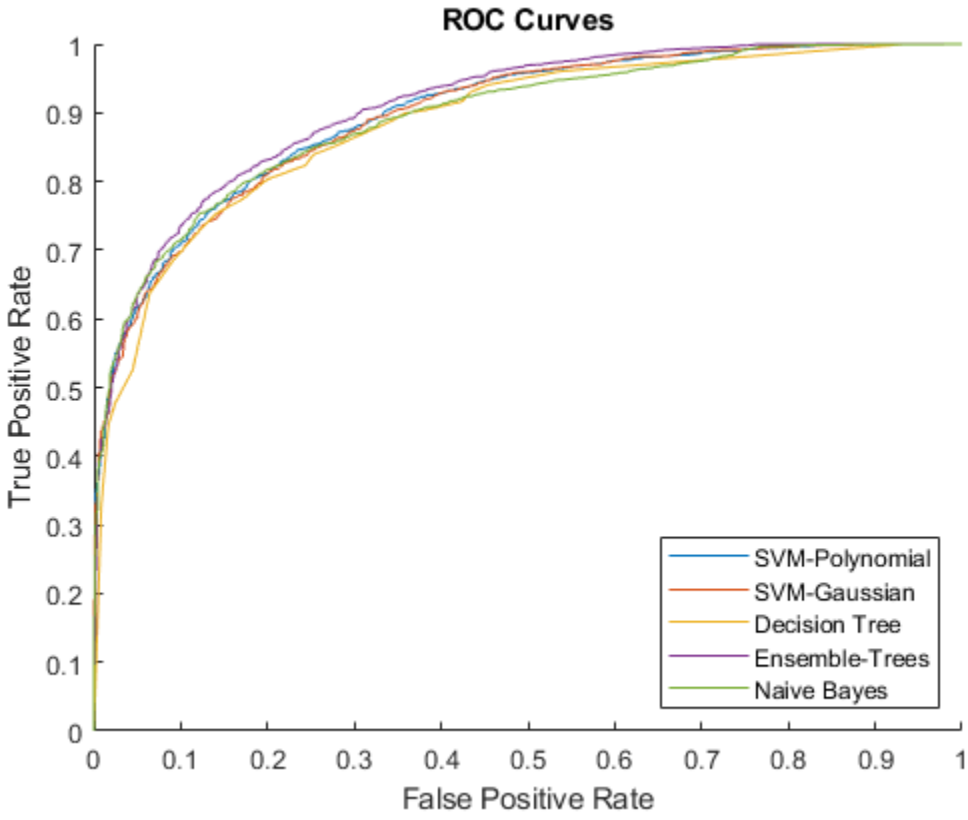
To plot the ROC curves for the score values corresponding to the label `'<=50K'`, check the column order of the score values returned from the `predict` function. The column order is the same as the category order of the response variable in the training data set. Display the category order.

```
c = categories(adultdata.salary)
```

```
c = 2×1 cell
    {'<=50K'}
    {'>50K' }
```

Plot the ROC curves.

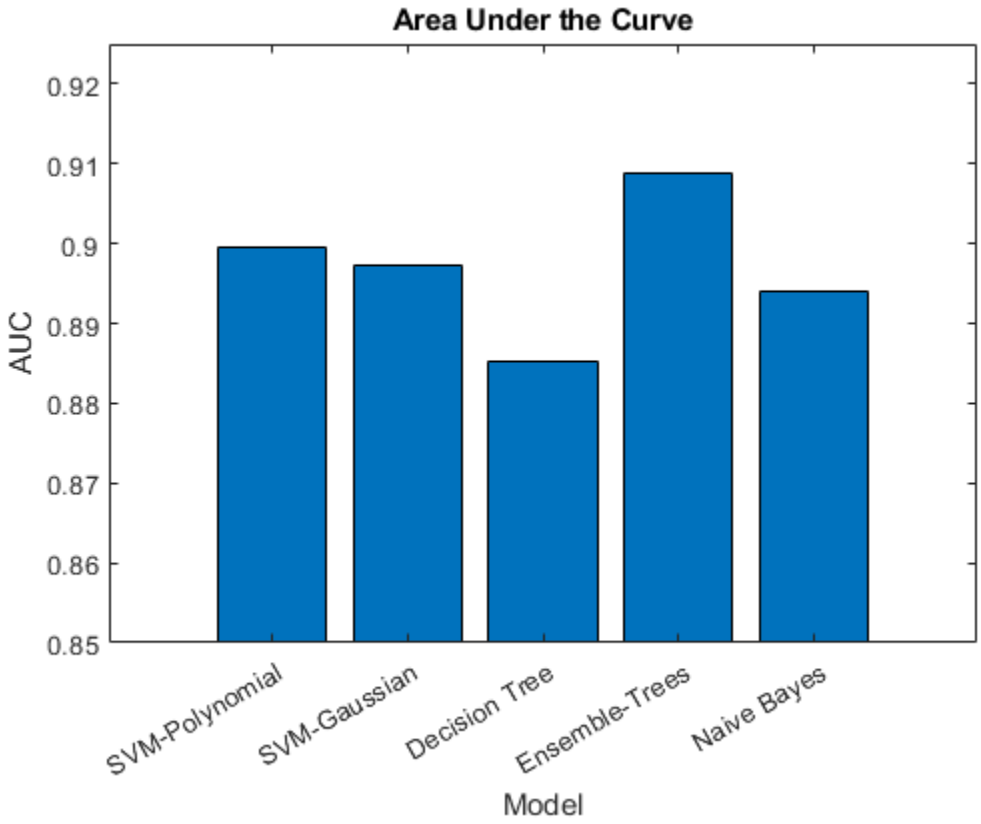
```
figure
hold on
AUC = zeros(1,N);
for i = 1:N
    [X,Y,~,AUC(i)] = perfcurve(adulttest.salary,score{i}(:,1),'<=50K');
    plot(X,Y)
end
title('ROC Curves')
xlabel('False Positive Rate')
ylabel('True Positive Rate')
legend(names,'Location','southeast')
```



An ROC curve shows the true positive rate versus the false positive rate (or, sensitivity versus 1–specificity) for different thresholds of the classifier output.

Now plot the AUC values using a bar graph. For a perfect classifier, whose true positive rate is always 1 regardless of the thresholds, AUC = 1. For a classifier that randomly assigns observations to classes, AUC = 0.5. Larger AUC values indicate better classifier performance.

```
figure
bar(AUC)
title('Area Under the Curve')
xlabel('Model')
ylabel('AUC')
xticklabels(names)
xtickangle(30)
ylim([0.85,0.925])
```

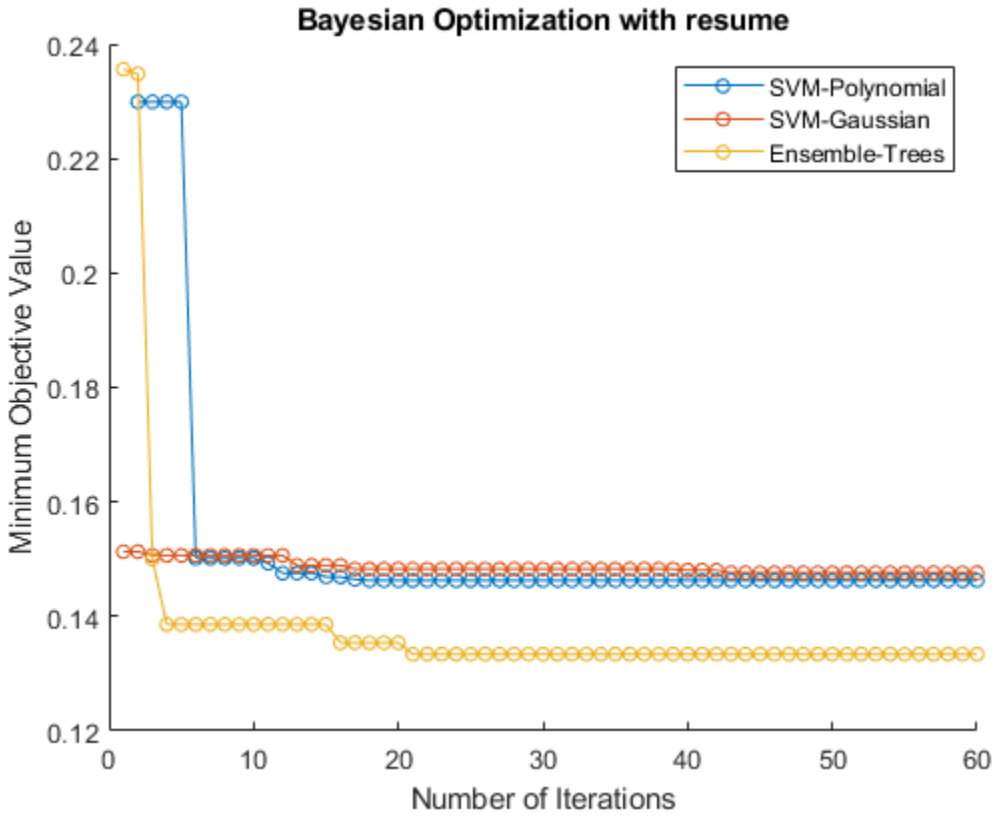


Based on the confusion matrix and the AUC bar graph, the ensemble of decision trees and SVM models achieve better accuracy than the decision tree and naive Bayes models.

Resume Optimization of Most Promising Models

Running Bayesian optimization on all models for further iterations can be computationally expensive. Instead, select a subset of models that have performed well so far and continue the optimization for 30 more iterations by using the `resume` function. Plot the minimum observed values of the objective function for each iteration of Bayesian optimization.

```
figure
hold on
selectedMdls = mdls([1,2,4]);
newresults = cell(1,length(selectedMdls));
for i = 1:length(selectedMdls)
    newresults{i} = resume(selectedMdls{i}.HyperparameterOptimizationResults,'MaxObjectiveEvalua
    plot(newresults{i}.ObjectiveMinimumTrace,'Marker','o','MarkerSize',5)
end
title('Bayesian Optimization with resume')
xlabel('Number of Iterations')
ylabel('Minimum Objective Value')
legend({'SVM-Polynomial','SVM-Gaussian','Ensemble-Trees'},'Location','northeast')
```



The first 30 iterations correspond to the first round of Bayesian optimization. The next 30 iterations correspond to the results of the `resume` function. Resuming optimization is useful because the loss continues to reduce further after the first 30 iterations.

See Also

`BayesianOptimization` | `confusionchart` | `perfcurve` | `resume`

More About

- “Bayesian Optimization Workflow” on page 10-25

Automated Classifier Selection with Bayesian Optimization

This example shows how to use `fitcauto` to automatically try a selection of classification model types with different hyperparameter values, given training predictor and response data. The function uses Bayesian optimization to select models and their hyperparameter values, and computes the cross-validation classification error for each model. After the optimization is complete, `fitcauto` returns the model, trained on the entire data set, that is expected to best classify new data. Check the model performance on test data.

Load Sample Data

This example uses the 1994 census data stored in `census1994.mat`. The data set consists of demographic information from the US Census Bureau that can be used to predict whether an individual makes over $50,000 per year.

Load the sample data `census1994`, which contains the training data `adultdata` and the test data `adulttest`. Preview the first few rows of the training data set.

```
load census1994
head(adultdata)
```

```
ans=8×15 table
    age       workClass          fnlwgt      education    education_num       marital_status
    ___    _________________    __________    _________    _____________    ____________________

    39     State-gov                 77516    Bachelors         13          Never-married
    50     Self-emp-not-inc          83311    Bachelors         13          Married-civ-spouse
    38     Private                2.1565e+05    HS-grad            9          Divorced
    53     Private                2.3472e+05    11th               7          Married-civ-spouse
    28     Private                3.3841e+05    Bachelors         13          Married-civ-spouse
    37     Private                2.8458e+05    Masters           14          Married-civ-spouse
    49     Private                1.6019e+05    9th                5          Married-spouse-absent
    52     Self-emp-not-inc       2.0964e+05    HS-grad            9          Married-civ-spouse
```

Each row contains the demographic information for one adult. The last column `salary` shows whether a person has a salary less than or equal to $50,000 per year or greater than $50,000 per year.

Use Automated Model Selection

Use `fitcauto` to automatically find an appropriate classifier for the data in `adultdata`. Set the observation weights, and specify to run the Bayesian optimization in parallel, which requires Parallel Computing Toolbox™. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results.

Because of the complexity of the optimization, this process can take some time, especially for larger data sets. By default, `fitcauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-1568.

```
options = struct('UseParallel',true);
[mdl,results] = fitcauto(adultdata,'salary','Weights','fnlwgt', ...
    'HyperparameterOptimizationOptions',options);
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the
```

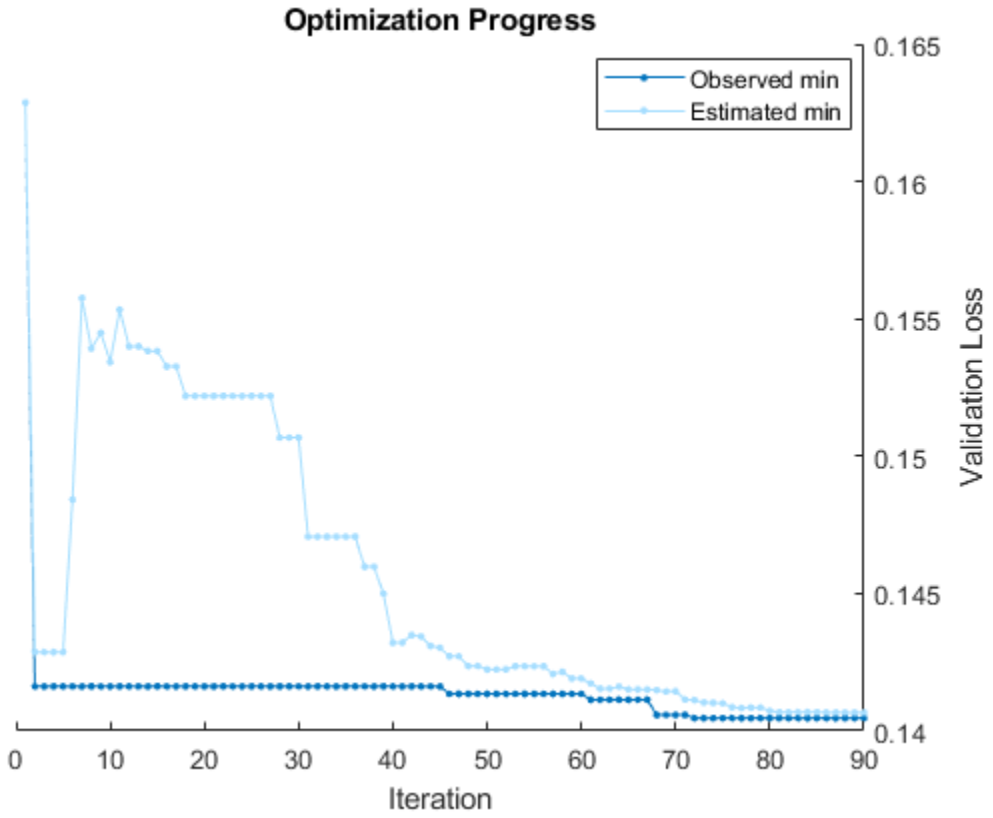
```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
Copying objective function to workers...
Done copying objective function to workers.

Learner types to explore: ensemble, nb, tree
Total iterations (MaxObjectiveEvaluations): 90
Total time (MaxTime): Inf

|=====================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================|
|    1 |       6 | Best   |    0.16287 |            4.3468 |         0.16287 |         0.16287 |
|      |         |        |            |                   |                 |                 |
|    2 |       5 | Accept |    0.14389 |            6.1049 |         0.14162 |         0.14287 |
|    3 |       5 | Best   |    0.14162 |            5.6195 |         0.14162 |         0.14287 |
|    4 |       6 | Accept |    0.15626 |            74.156 |         0.14162 |         0.14287 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    5 |       6 | Accept |    0.15603 |            77.293 |         0.14162 |         0.14287 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    6 |       6 | Accept |    0.16027 |            5.6224 |         0.14162 |         0.14842 |
|    7 |       6 | Accept |    0.17343 |            8.6209 |         0.14162 |         0.15576 |
|    8 |       6 | Accept |    0.15103 |            4.8867 |         0.14162 |         0.15392 |
|    9 |       6 | Accept |    0.17642 |            1.1808 |         0.14162 |         0.15449 |
|   10 |       6 | Accept |    0.15927 |            5.0734 |         0.14162 |         0.15343 |
|=====================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================|
|   11 |       6 | Accept |    0.17009 |            1.6504 |         0.14162 |         0.15533 |
|   12 |       6 | Accept |    0.17869 |            1.0308 |         0.14162 |           0.154 |
|   13 |       6 | Accept |    0.17961 |            116.64 |         0.14162 |           0.154 |
|      |         |        |            |                   |                 |                 |
|   14 |       5 | Accept |    0.15128 |            118.36 |         0.14162 |         0.15383 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   15 |       5 | Accept |    0.15177 |            115.42 |         0.14162 |         0.15383 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   16 |       5 | Accept |    0.15116 |            115.49 |         0.14162 |         0.15326 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   17 |       6 | Accept |    0.14887 |            63.412 |         0.14162 |         0.15326 |
|      |         |        |            |                   |                 |                 |
```

```
|   18 |       6 | Accept |    0.17869 |           0.89318 |         0.14162 |         0.15219 |
|   19 |       6 | Accept |    0.17676 |            59.781 |         0.14162 |         0.15219 |
|   20 |       6 | Accept |    0.15086 |             81.42 |         0.14162 |         0.15219 |
|======|=========|========|============|===================|=================|=================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======|=========|========|============|===================|=================|=================|
|   21 |       6 | Accept |    0.16287 |           0.64656 |         0.14162 |         0.15219 |
|   22 |       6 | Accept |    0.14943 |            75.578 |         0.14162 |         0.15219 |
|   23 |       6 | Accept |    0.16287 |           0.49489 |         0.14162 |         0.15219 |
|   24 |       6 | Accept |    0.14926 |            68.642 |         0.14162 |         0.15219 |
|   25 |       6 | Accept |    0.16287 |            0.5124 |         0.14162 |         0.15219 |
|   26 |       6 | Accept |    0.15609 |            58.267 |         0.14162 |         0.15219 |
|   27 |       6 | Accept |    0.16287 |           0.93385 |         0.14162 |         0.15219 |
|   28 |       6 | Accept |    0.15554 |            4.3668 |         0.14162 |         0.15067 |
|   29 |       6 | Accept |    0.15087 |            127.01 |         0.14162 |         0.15067 |
|   30 |       6 | Accept |    0.15142 |            127.39 |         0.14162 |         0.15067 |
|======|=========|========|============|===================|=================|=================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======|=========|========|============|===================|=================|=================|
|   31 |       6 | Accept |    0.14177 |            2.6306 |         0.14162 |         0.14707 |
|   32 |       6 | Accept |    0.16287 |            1.1225 |         0.14162 |         0.14707 |
|   33 |       6 | Accept |    0.15737 |            56.258 |         0.14162 |         0.14707 |
|   34 |       6 | Accept |    0.15158 |            97.559 |         0.14162 |         0.14707 |
```

```
|   35 |       6 | Accept |     0.1719 |            96.392 |         0.14162 |         0.14707 |
|      |         |        |            |                   |                 |                 |
|   36 |       6 | Accept |    0.16287 |           0.42054 |         0.14162 |         0.14707 |
|      |         |        |            |                   |                 |                 |
|   37 |       6 | Accept |    0.14441 |            3.5932 |         0.14162 |         0.14598 |
|   38 |       6 | Accept |    0.16287 |           0.34693 |         0.14162 |         0.14598 |
|      |         |        |            |                   |                 |                 |
|   39 |       6 | Accept |    0.14432 |            3.4661 |         0.14162 |           0.145 |
|   40 |       6 | Accept |    0.14291 |            2.3121 |         0.14162 |         0.14321 |

|=====================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================|
|   41 |       6 | Accept |    0.15278 |            96.086 |         0.14162 |         0.14321 |
|      |         |        |            |                   |                 |                 |
|   42 |       6 | Accept |    0.15068 |            1.9847 |         0.14162 |         0.14348 |
|   43 |       6 | Accept |    0.14705 |            2.1122 |         0.14162 |         0.14343 |
|   44 |       6 | Accept |    0.14186 |            2.3835 |         0.14162 |         0.14309 |
|   45 |       6 | Accept |    0.16209 |            1.9821 |         0.14162 |         0.14302 |
|   46 |       5 | Accept |    0.15783 |            53.627 |         0.14135 |         0.14271 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   47 |       5 | Best   |    0.14135 |            3.1329 |         0.14135 |         0.14271 |
|   48 |       4 | Accept |    0.15637 |            63.578 |         0.14135 |         0.14236 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   49 |       4 | Accept |     0.1448 |            2.1012 |         0.14135 |         0.14236 |
|   50 |       3 | Accept |     0.1513 |            114.35 |         0.14135 |         0.14224 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |

|=====================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================|
|   51 |       3 | Accept |    0.14271 |            2.2737 |         0.14135 |         0.14224 |
|   52 |       6 | Accept |    0.14349 |            1.9707 |         0.14135 |         0.14224 |
|   53 |       3 | Accept |    0.15337 |            1.6887 |         0.14135 |         0.14235 |
|   54 |       3 | Accept |    0.17869 |             1.049 |         0.14135 |         0.14235 |
|   55 |       3 | Accept |     0.1785 |            0.9639 |         0.14135 |         0.14235 |
|   56 |       3 | Accept |    0.18062 |           0.63917 |         0.14135 |         0.14235 |
|   57 |       6 | Accept |    0.14673 |            3.2067 |         0.14135 |         0.14207 |
|   58 |       6 | Accept |    0.14238 |            2.3081 |         0.14135 |         0.14215 |
```

```
|   59 |       5 | Accept |    0.16352 |            125.94 |         0.14135 |          0.1419 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   60 |       5 | Accept |    0.14162 |             2.849 |         0.14135 |          0.1419 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   61 |       5 | Best   |    0.14113 |            2.6499 |         0.14113 |         0.14173 |

|   62 |       5 | Accept |    0.14178 |            2.9853 |         0.14113 |         0.14153 |

|   63 |       5 | Accept |    0.14157 |            2.8701 |         0.14113 |         0.14153 |

|   64 |       5 | Accept |    0.15886 |            1.7188 |         0.14113 |         0.14161 |

|   65 |       5 | Accept |    0.14529 |            3.6593 |         0.14113 |         0.14151 |

|   66 |       4 | Accept |    0.23856 |            41.472 |         0.14113 |         0.14151 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   67 |       4 | Accept |    0.14702 |            4.0559 |         0.14113 |         0.14151 |

|   68 |       4 | Best   |    0.14058 |            2.8472 |         0.14058 |         0.14148 |

|   69 |       4 | Accept |    0.14168 |            2.1868 |         0.14058 |         0.14143 |

|   70 |       4 | Accept |    0.14072 |            2.9698 |         0.14058 |         0.14144 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   71 |       4 | Accept |    0.14117 |            2.8824 |         0.14058 |         0.14114 |

|   72 |       4 | Best   |    0.14046 |            2.8853 |         0.14046 |         0.14112 |

|   73 |       4 | Accept |    0.14184 |            2.8532 |         0.14046 |         0.14103 |

|   74 |       4 | Accept |    0.14112 |            2.7998 |         0.14046 |         0.14102 |

|   75 |       4 | Accept |    0.14331 |            3.0835 |         0.14046 |           0.141 |

|   76 |       4 | Accept |    0.14089 |            2.9637 |         0.14046 |         0.14086 |

|   77 |       4 | Accept |    0.14046 |            3.0017 |         0.14046 |         0.14083 |

|   78 |       3 | Accept |    0.15093 |            91.952 |         0.14046 |         0.14085 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   79 |       3 | Accept |    0.14046 |            2.9993 |         0.14046 |         0.14085 |

|   80 |       6 | Accept |    0.14046 |            2.7739 |         0.14046 |         0.14073 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   81 |       2 | Accept |    0.18178 |            101.13 |         0.14046 |         0.14068 |
|      |         |        |            |                   |                 |                 |
```

```
|   82 |        2 | Accept |     0.14184 |             3.2218 |          0.14046 |          0.14068 |
|   83 |        2 | Accept |     0.17807 |            0.82685 |          0.14046 |          0.14068 |
|   84 |        2 | Accept |     0.15989 |             1.8729 |          0.14046 |          0.14068 |
|   85 |        2 | Accept |     0.15103 |             3.8835 |          0.14046 |          0.14068 |

|   86 |        6 | Accept |     0.14046 |             2.5909 |          0.14046 |          0.14067 |

|   87 |        6 | Accept |     0.14331 |             3.5433 |          0.14046 |          0.14067 |

|   88 |        6 | Accept |     0.23856 |             47.904 |          0.14046 |          0.14067 |
|      |          |        |             |                    |                  |                  |
|      |          |        |             |                    |                  |                  |

|   89 |        6 | Accept |     0.14914 |             59.665 |          0.14046 |          0.14067 |
|      |          |        |             |                    |                  |                  |

|   90 |        6 | Accept |     0.15604 |             68.731 |          0.14046 |          0.14067 |
|      |          |        |             |                    |                  |                  |
|      |          |        |             |                    |                  |                  |
```




```
Optimization completed.
Total iterations: 90
Total elapsed time: 577.1419 seconds
Total time for training and validation: 2558.1542 seconds

Best observed learner is a tree model with:
    MinLeafSize:             25
Observed validation loss: 0.14046
```

```
Time for training and validation: 2.8853 seconds

Best estimated learner (returned model) is a tree model with:
    MinLeafSize:             25
Estimated validation loss: 0.14067
Estimated time for training and validation: 2.8824 seconds

Documentation for fitcauto display
```

The final model returned by `fitcauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`adultdata`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

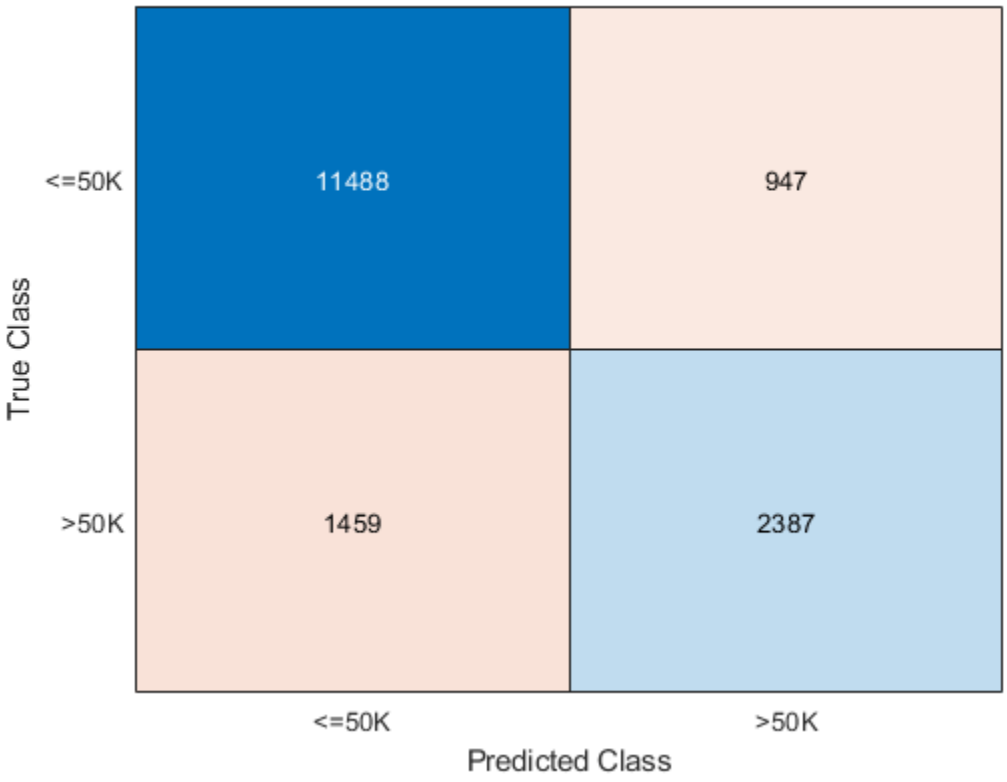
Evaluate the performance of the returned model `mdl` on the test set `adulttest` by using a confusion matrix and a receiver operating characteristic (ROC) curve.

Find the predicted labels and score values for the test set.

```
[labels,scores] = predict(mdl,adulttest);
```

Create a confusion matrix from the test set results. The diagonal elements indicate the number of correctly classified instances of a given class. The off-diagonal elements are instances of misclassified observations.

```
confusionchart(adulttest.salary,labels)
```



Compute the test set classification accuracy. `accuracy` is the percentage of correctly classified test set observations.

```
accuracy = (1-loss(mdl,adulttest,'salary'))*100
```

```
accuracy = 85.1513
```

To plot the ROC curve for the score values corresponding to the label `'<=50K'`, find the column of `scores` that corresponds to that label. The column order of `scores` matches the order of the classes in the trained model.

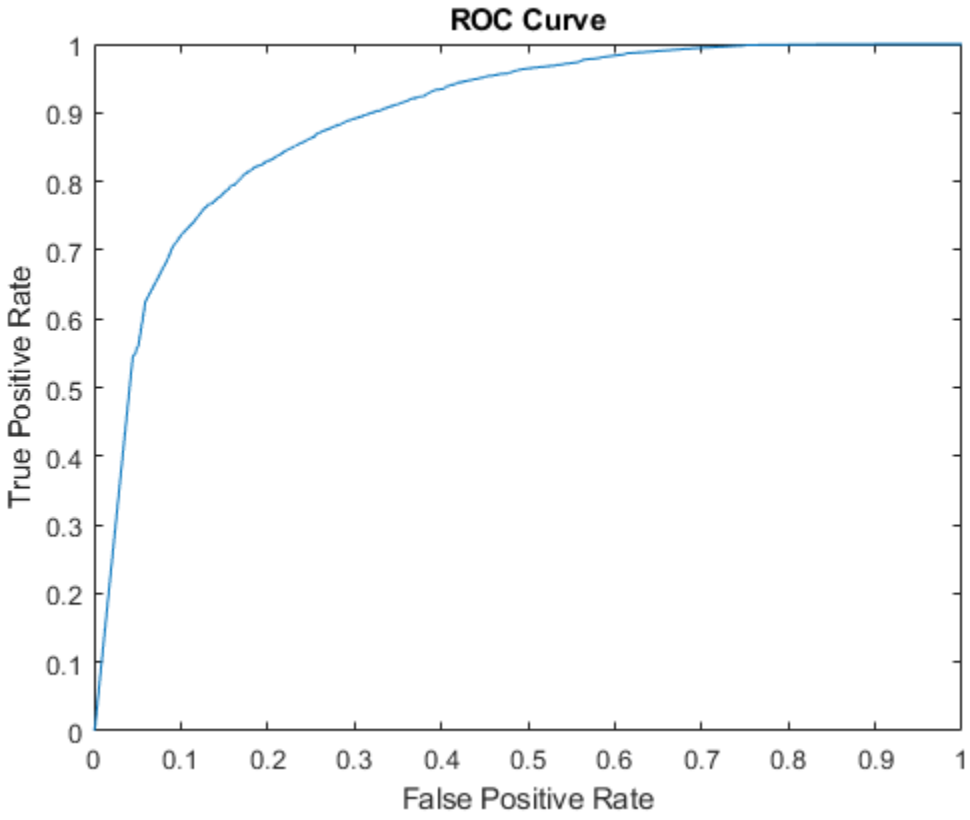
```
mdl.ClassNames
```

```
ans = 2×1 categorical
     <=50K
     >50K
```

Because `'<=50K'` is listed first, the first column of `scores` corresponds to that label.

Plot the ROC curve, and compute the area under the curve (AUC). The ROC curve shows the true positive rate versus the false positive rate for different thresholds of the classifier output. For a perfect classifier, whose true positive rate is always 1 regardless of the threshold, AUC = 1. For a binary classifier that randomly assigns observations to classes, AUC = 0.5. A large AUC value (close to 1) indicates good classifier performance.

```
[X,Y,~,AUC] = perfcurve(adulttest.salary,scores(:,1),'<=50K');
plot(X,Y)
title('ROC Curve')
xlabel('False Positive Rate')
ylabel('True Positive Rate')
```




```
AUC
```

```
AUC = 0.8947
```

Based on the accuracy and AUC values, the classifier performs well on the test data.

See Also

`BayesianOptimization` | `confusionchart` | `fitcauto` | `perfcurve`

More About

- “Bayesian Optimization Workflow” on page 10-25
- “Hyperparameter Optimization in Classification Learner App” on page 23-54

Automated Regression Model Selection with Bayesian Optimization

This example shows how to use the `fitrauto` function to automatically try a selection of regression model types with different hyperparameter values, given training predictor and response data. The function uses Bayesian optimization to select models and their hyperparameter values, and computes the following for each model: log(1 + *valLoss*), where *valLoss* is the cross-validation mean squared error (MSE). After the optimization is complete, `fitrauto` returns the model, trained on the entire data set, that is expected to best predict the responses for new data. Check the model performance on test data.

Prepare Data

Load the sample data set `NYCHousing2015`, which includes 10 variables with information on the sales of properties in New York City in 2015. This example uses some of these variables to analyze the sale prices.

```
load NYCHousing2015
```

Instead of loading the sample data set `NYCHousing2015`, you can download the data from the NYC Open Data website and import the data as follows.

```
folder = 'Annualized_Rolling_Sales_Update';
ds = spreadsheetDatastore(folder,"TextType","string","NumHeaderLines",4);
ds.Files = ds.Files(contains(ds.Files,"2015"));
ds.SelectedVariableNames = ["BOROUGH","NEIGHBORHOOD","BUILDINGCLASSCATEGORY","RESIDENTIALUNITS",
    "COMMERCIALUNITS","LANDSQUAREFEET","GROSSSQUAREFEET","YEARBUILT","SALEPRICE","SALEDATE"];
NYCHousing2015 = readall(ds);
```

Preprocess the data set to choose the predictor variables of interest. Some of the preprocessing steps match those in the example “Train Linear Regression Model” on page 11-161.

First, change the variable names to lowercase for readability.

```
NYCHousing2015.Properties.VariableNames = lower(NYCHousing2015.Properties.VariableNames);
```

Next, remove samples with certain problematic values. For example, retain only those samples where at least one of the area measurements `grosssquarefeet` or `landsquarefeet` is nonzero. Assume that a `saleprice` of $0 indicates an ownership transfer without a cash consideration, and remove the samples with that `saleprice` value. Assume that a `yearbuilt` value of 1500 or less is a typo, and remove the corresponding samples.

```
NYCHousing2015(NYCHousing2015.grosssquarefeet == 0 & NYCHousing2015.landsquarefeet == 0,:) = [];
NYCHousing2015(NYCHousing2015.saleprice == 0,:) = [];
NYCHousing2015(NYCHousing2015.yearbuilt <= 1500,:) = [];
```

Convert the `saledate` variable, specified as a `datetime` array, into two numeric columns `MM` (month) and `DD` (day), and remove the `saledate` variable. Ignore the year values because all samples are for the year 2015.

```
[~,NYCHousing2015.MM,NYCHousing2015.DD] = ymd(NYCHousing2015.saledate);
NYCHousing2015.saledate = [];
```

The numeric values in the `borough` variable indicate the names of the boroughs. Change the variable to a categorical variable using the names.

```
NYCHousing2015.borough = categorical(NYCHousing2015.borough,1:5, ...
    ["Manhattan","Bronx","Brooklyn","Queens","Staten Island"]);
```

The `neighborhood` variable has 254 categories. Remove this variable for simplicity.

```
NYCHousing2015.neighborhood = [];
```

Convert the `buildingclasscategory` variable to a categorical variable, and explore the variable by using the `wordcloud` function.

```
NYCHousing2015.buildingclasscategory = categorical(NYCHousing2015.buildingclasscategory);
wordcloud(NYCHousing2015.buildingclasscategory);
```

Assume that you are interested only in one-, two-, and three-family dwellings. Find the sample indices for these dwellings and delete the other samples. Then, change the `buildingclasscategory` variable to an ordinal categorical variable, with integer-valued category names.

```
idx = ismember(string(NYCHousing2015.buildingclasscategory), ...
    ["01  ONE FAMILY DWELLINGS","02  TWO FAMILY DWELLINGS","03  THREE FAMILY DWELLINGS"]);
NYCHousing2015 = NYCHousing2015(idx,:);
NYCHousing2015.buildingclasscategory = categorical(NYCHousing2015.buildingclasscategory, ...
    ["01  ONE FAMILY DWELLINGS","02  TWO FAMILY DWELLINGS","03  THREE FAMILY DWELLINGS"], ...
    ["1","2","3"],'Ordinal',true);
```

The `buildingclasscategory` variable now indicates the number of families in one dwelling.

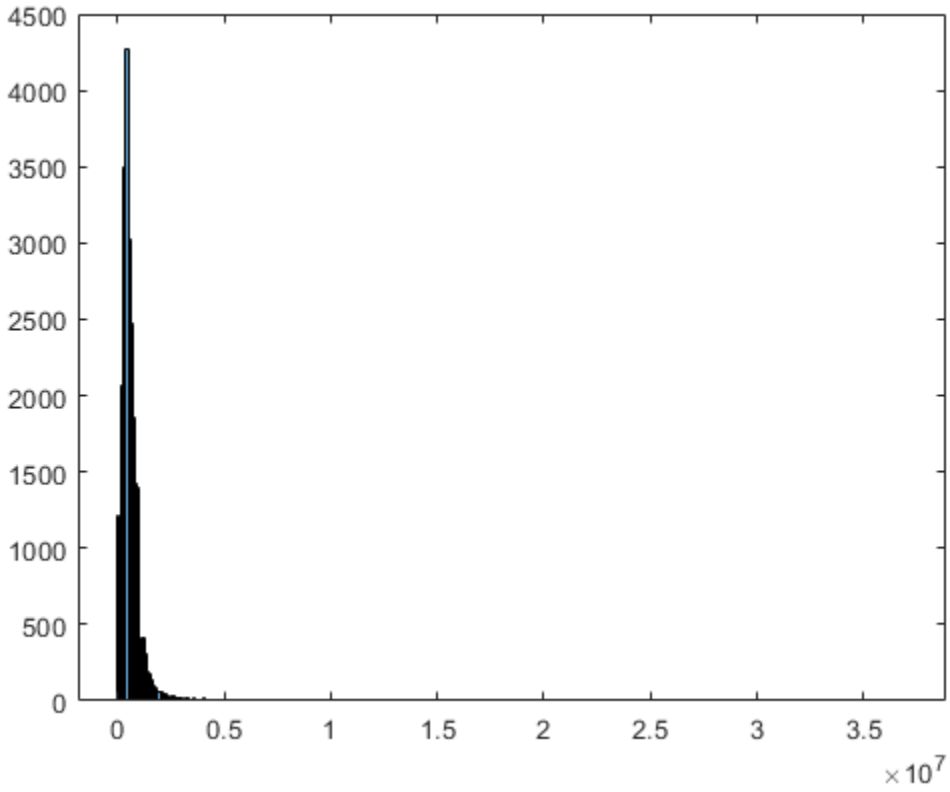
Explore the response variable `saleprice` by using the `summary` function.

```
s = summary(NYCHousing2015);
s.saleprice
```

```
ans = struct with fields:
           Size: [24972 1]
           Type: 'double'
    Description: ''
          Units: ''
     Continuity: []
            Min: 1
         Median: 515000
            Max: 37000000
     NumMissing: 0
```

Create a histogram of the `saleprice` variable.

```
histogram(NYCHousing2015.saleprice)
```



Because the distribution of `saleprice` values is right-skewed, with all values greater than 0, log transform the `saleprice` variable.

```
NYCHousing2015.saleprice = log(NYCHousing2015.saleprice);
```

Similarly, transform the `grosssquarefeet` and `landsquarefeet` variables. Add a value of 1 before taking the logarithm of each variable, in case the variable is equal to 0.

```
NYCHousing2015.grosssquarefeet = log(1 + NYCHousing2015.grosssquarefeet);
NYCHousing2015.landsquarefeet = log(1 + NYCHousing2015.landsquarefeet);
```

Partition Data and Remove Outliers

Partition the data set into a training set and a test set by using `cvpartition`. Use approximately 80% of the observations for the model selection and hyperparameter tuning process, and the other 20% to test the performance of the final model returned by `fitrauto`.

```
rng('default') % For reproducibility of the partition
c = cvpartition(length(NYCHousing2015.saleprice),'Holdout',0.2);
trainData = NYCHousing2015(training(c),:);
testData = NYCHousing2015(test(c),:);
```

Identify and remove the outliers of `saleprice`, `grosssquarefeet`, and `landsquarefeet` from the training data by using the `isoutlier` function.

```
[priceIdx,priceL,priceU] = isoutlier(trainData.saleprice);
trainData(priceIdx,:) = [];

[grossIdx,grossL,grossU] = isoutlier(trainData.grosssquarefeet);
trainData(grossIdx,:) = [];

[landIdx,landL,landU] = isoutlier(trainData.landsquarefeet);
trainData(landIdx,:) = [];
```

Remove the outliers of `saleprice`, `grosssquarefeet`, and `landsquarefeet` from the test data by using the same lower and upper thresholds computed on the training data.

```
testData(testData.saleprice < priceL | testData.saleprice > priceU,:) = [];
testData(testData.grosssquarefeet < grossL | testData.grosssquarefeet > grossU,:) = [];
testData(testData.landsquarefeet < landL | testData.landsquarefeet > landU,:) = [];
```

Use Automated Model Selection

Find an appropriate regression model for the data in `trainData` by using `fitrauto`. Try tree and ensemble learners and run the Bayesian optimization in parallel, which requires Parallel Computing Toolbox™. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. To reduce computational time, use 3-fold cross-validation, rather than 5-fold cross-validation, as part of the optimization process.

Because of the complexity of the optimization, this process can take some time, especially for larger data sets. By default, `fitrauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-2111.

```
options = struct('UseParallel',true,'Kfold',3);
[mdl,results] = fitrauto(trainData,'saleprice', ...
    'Learners',{'tree','ensemble'},'HyperparameterOptimizationOptions',options);
```

```
Copying objective function to workers...
Done copying objective function to workers.

Learner types to explore: ensemble, tree
Total iterations (MaxObjectiveEvaluations): 60
Total time (MaxTime): Inf

|=====================================================================================================
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated m
```

	workers	result		& validation (sec)	log(1 + valLoss)	log(1 + va
1	5	Accept	0.25922	0.067333	0.18985	0
2	5	Best	0.18985	0.14568	0.18985	0
3	2	Accept	0.25126	0.86908	0.1849	0
4	2	Best	0.1849	1.0049	0.1849	0
5	2	Accept	0.25922	0.5705	0.1849	0
6	2	Accept	0.25126	0.87986	0.1849	0
7	6	Accept	0.21227	0.069611	0.1849	0
8	4	Accept	0.18763	0.15728	0.1849	0
9	4	Accept	2.9803	0.47009	0.1849	0
10	4	Accept	0.1914	0.23832	0.1849	0

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + va
11	4	Accept	0.23248	0.057309	0.1849	0
12	4	Accept	0.22354	2.9816	0.1849	0
13	4	Accept	4.7611	3.9969	0.1849	0
14	3	Best	0.17937	1.6426	0.17937	0
15	3	Accept	0.25923	0.31222	0.17937	0
16	5	Best	0.17799	4.7324	0.17799	0
17	5	Accept	0.19076	0.10145	0.17799	0
18	4	Accept	0.22517	0.079648	0.17799	0
19	4	Accept	0.21507	0.32105	0.17799	0
20	4	Accept	0.18797	0.11404	0.17799	0

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
21	3	Accept	0.17862	5.6403	0.17799	0
22	3	Accept	0.19413	0.16212	0.17799	0
23	6	Accept	0.24396	0.74378	0.17799	0
24	5	Accept	0.18986	0.16919	0.17799	0
25	5	Accept	0.19608	0.19077	0.17799	0
26	4	Accept	0.17828	11.436	0.17799	0
27	4	Accept	0.1809	3.7272	0.17799	0
28	4	Accept	0.18171	1.9361	0.17799	0
29	4	Accept	0.17959	8.6553	0.17799	0
30	3	Accept	0.20204	15.893	0.17762	0

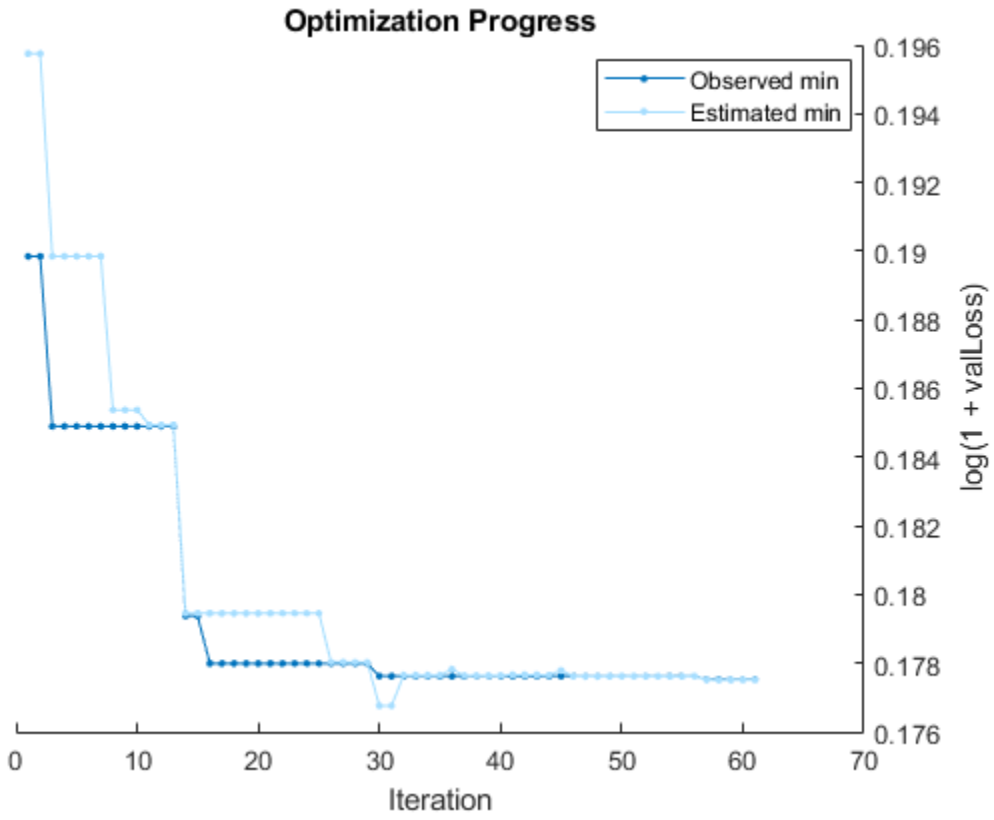
Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
31	3	Best	0.17762	6.6202	0.17762	0
32	5	Accept	0.19444	5.563	0.17762	0
33	5	Accept	0.18056	0.75592	0.17762	0
34	4	Accept	0.18768	0.78702	0.17762	0

35	4	Accept	0.26635	1.1019	0.17762	0
36	4	Accept	0.206	4.2801	0.17762	0
37	3	Accept	0.21503	11.309	0.17762	0
38	3	Accept	0.3789	0.55978	0.17762	0
39	6	Accept	0.23053	0.51383	0.17762	0
40	6	Accept	0.17996	2.7932	0.17762	0

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
41	6	Accept	0.23707	12.289	0.17762	0
42	5	Accept	0.18527	26.878	0.17762	0
43	5	Accept	0.18276	4.678	0.17762	0
44	5	Accept	0.25057	22.826	0.17762	0
45	5	Accept	4.61	0.60326	0.17762	0
46	4	Accept	0.2001	25.14	0.17762	0
47	4	Accept	0.20319	3.6011	0.17762	0

```
|   48 |       4 | Accept |          0.18106 |            26.642 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   49 |       4 | Accept |          0.25922 |            7.6886 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   50 |       4 | Accept |          0.18162 |            4.8133 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|==========================================================================================================
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + val
|==========================================================================================================
|   51 |       5 | Accept |          0.17788 |              5.18 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   52 |       5 | Accept |           0.1858 |            36.145 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   53 |       5 | Accept |          0.17946 |            1.3924 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   54 |       6 | Accept |          0.17837 |            8.0619 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   55 |       6 | Accept |          0.17766 |            25.949 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   56 |       6 | Accept |           1.2951 |            25.877 |          0.17762 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   57 |       6 | Best   |          0.17753 |            28.063 |          0.17753 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   58 |       6 | Accept |          0.17978 |            2.0228 |          0.17753 |           0
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   59 |       6 | Accept |           2.6442 |           0.64641 |          0.17753 |           0
|      |         |        |                  |                   |                  |
```

```
|      |         |        |                 |                 |                 |
|      |         |        |                 |                 |                 |
|   60 |       5 | Accept |         0.97054 |           16.16 |         0.17753 |           0
|      |         |        |                 |                 |                 |
|      |         |        |                 |                 |                 |
|      |         |        |                 |                 |                 |
|======================================================================================================
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min    | Estimated m
|      | workers | result |                 | & validation (sec)| log(1 + valLoss) | log(1 + va
|======================================================================================================
|   61 |       5 | Accept |          0.2084 |          0.6174 |         0.17753 |           0
|      |         |        |                 |                 |                 |
|      |         |        |                 |                 |                 |
|      |         |        |                 |                 |                 |
```




```
_____________________________________________________________
Optimization completed.
Total iterations: 61
Total elapsed time: 150.2458 seconds
Total time for training and validation: 386.9239 seconds

Best observed learner is an ensemble model with:
    Method:             LSBoost
    NumLearningCycles:      486
    LearnRate:         0.038349
    MinLeafSize:             39
Observed log(1 + valLoss): 0.17753
```

```
Time for training and validation: 28.0634 seconds

Best estimated learner (returned model) is an ensemble model with:
    Method:             LSBoost
    NumLearningCycles:      486
    LearnRate:          0.038349
    MinLeafSize:              39
Estimated log(1 + valLoss): 0.17751
Estimated time for training and validation: 28.7325 seconds

Documentation for fitrauto display
```

The final model returned by `fitrauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`trainData`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

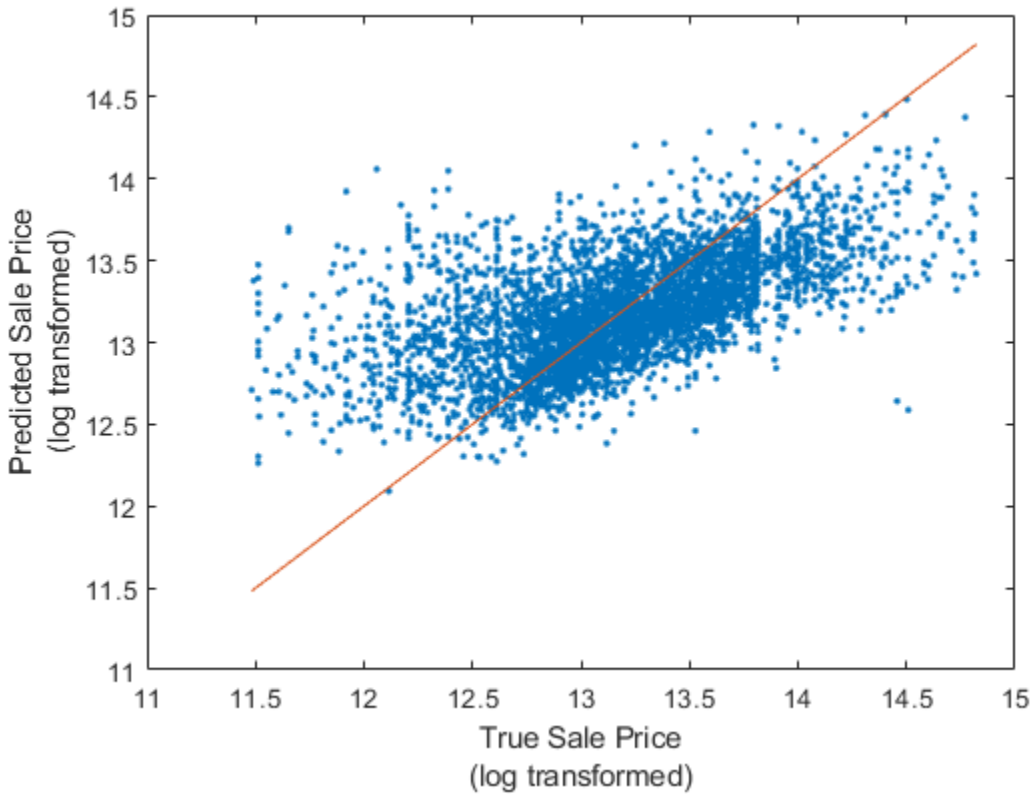
Evaluate the performance of the returned model `mdl` on the test set `testData`. Compute the test set mean squared error (MSE), and take a log transform of the MSE to match the values in the verbose display of `fitrauto`. Smaller MSE (and log-transformed MSE) values indicate better performance.

```
testMSE = loss(mdl,testData,'saleprice');
testError = log(1 + testMSE)
```

```
testError = 0.1791
```

Compare the predicted test set response values to the true response values. Plot the predicted sale price along the vertical axis and the true sale price along the horizontal axis. Points on the reference line indicate correct predictions. A good model produces predictions that are scattered near the line.

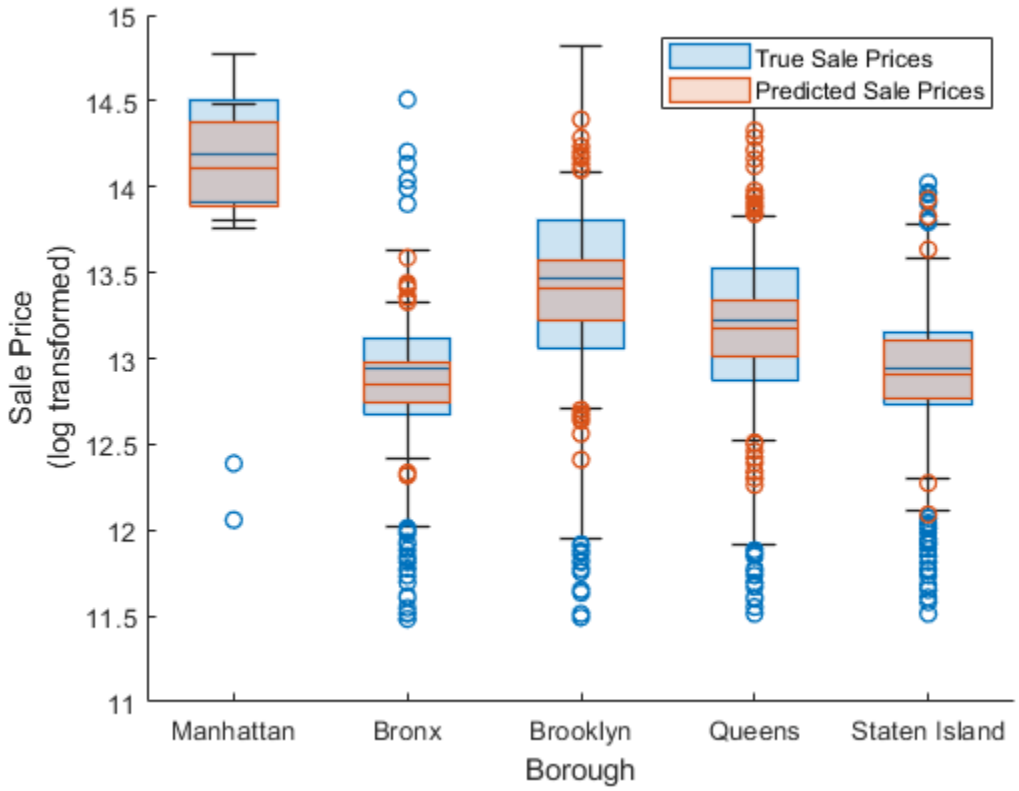
```
testPredictions = predict(mdl,testData);

plot(testData.saleprice,testPredictions,'.')
hold on
plot(testData.saleprice,testData.saleprice) % Reference line
hold off
xlabel(["True Sale Price","(log transformed)"])
ylabel(["Predicted Sale Price","(log transformed)"])
```



Use box plots to compare the distribution of predicted and true sale prices by borough. Create the box plots by using the `boxchart` function. Each box plot displays the median, the lower and upper quartiles, any outliers (computed using the interquartile range), and the minimum and maximum values that are not outliers. In particular, the line inside each box is the sample median, and the circular markers indicate outliers.

For each borough, compare the red box plot (showing the distribution of predicted prices) to the blue box plot (showing the distribution of true prices). Similar distributions for the predicted and true sale prices indicate good predictions.

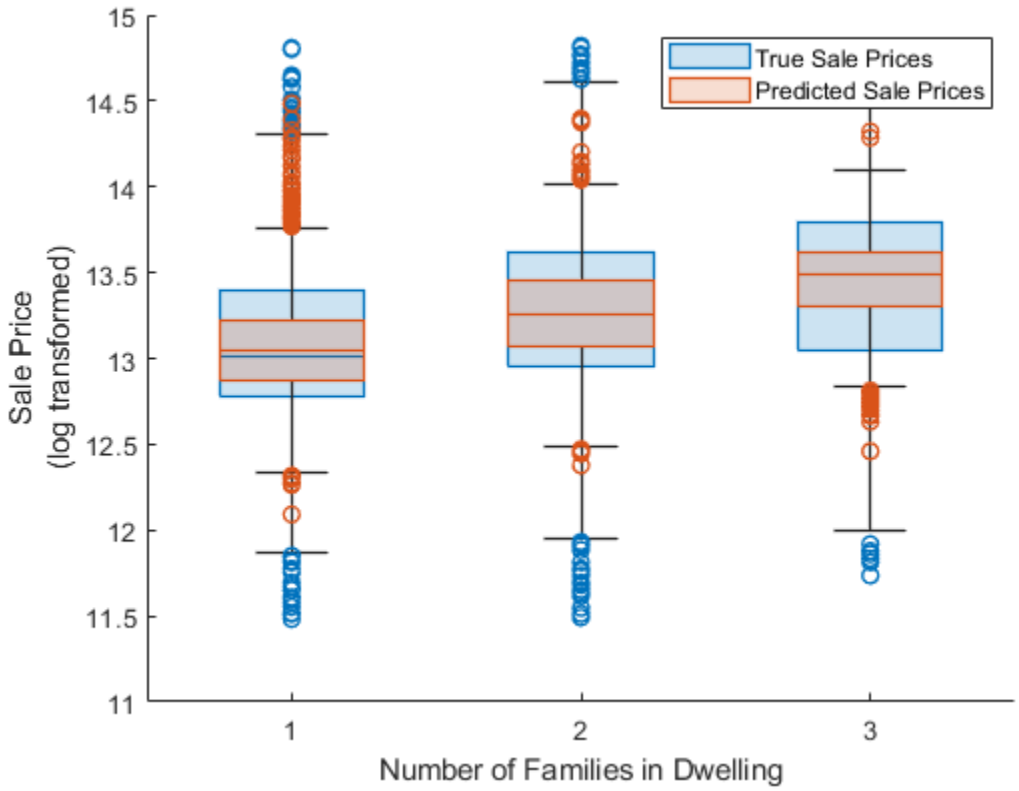
```
boxchart(testData.borough,testData.saleprice)
hold on
boxchart(testData.borough,testPredictions)
hold off
legend(["True Sale Prices","Predicted Sale Prices"])
xlabel("Borough")
ylabel(["Sale Price","(log transformed)"])
```



For all the boroughs, the predicted median sale price closely matches the median true sale price. The predicted sale prices seem to vary less than the true sale prices.

Display box charts that compare the distribution of predicted and true sale prices by the number of families in a dwelling.

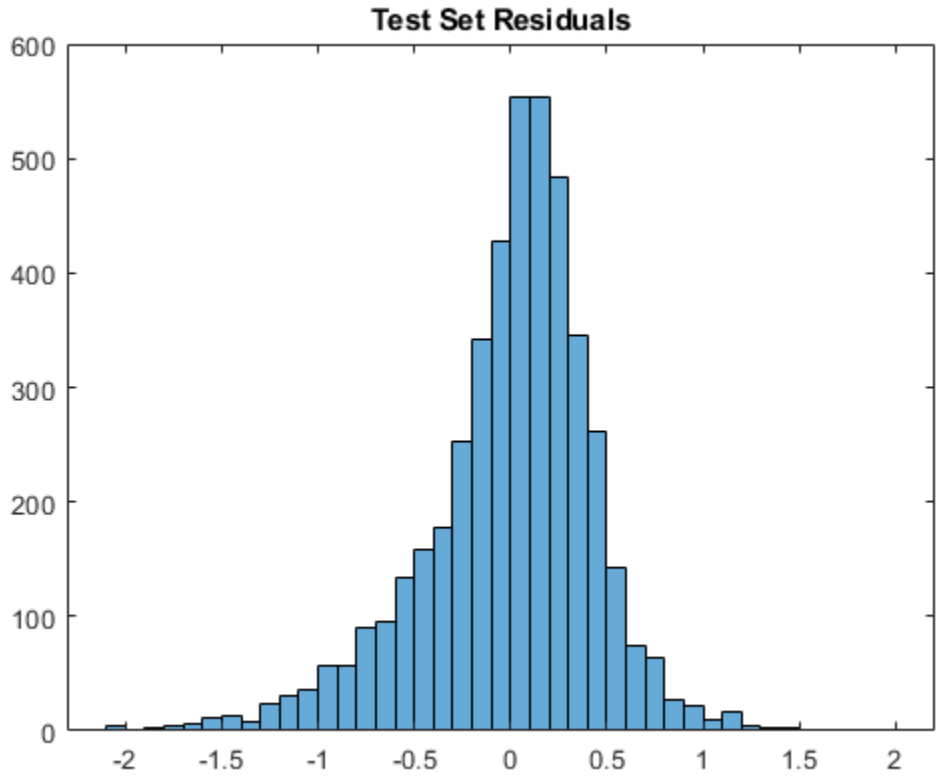
```
boxchart(testData.buildingclasscategory,testData.saleprice)
hold on
boxchart(testData.buildingclasscategory,testPredictions)
hold off
legend(["True Sale Prices","Predicted Sale Prices"])
xlabel("Number of Families in Dwelling")
ylabel(["Sale Price","(log transformed)"])
```



For all dwellings, the predicted median sale price closely matches the median true sale price. The predicted sale prices seem to vary less than the true sale prices.

Plot a histogram of the test set residuals, and check that they are normally distributed. (Recall that the sale prices are log-transformed.)

```
testResiduals = testData.saleprice - testPredictions;
histogram(testResiduals)
title('Test Set Residuals')
```



Although the histogram is slightly left-skewed, it is approximately symmetric about 0.

See Also

`BayesianOptimization` | `boxchart` | `fitrauto` | `histogram`

More About

- “Bayesian Optimization Workflow” on page 10-25
- “Hyperparameter Optimization in Regression Learner App” on page 24-30

Credit Rating by Bagging Decision Trees

This example shows how to build an automated credit rating tool.

One of the fundamental tasks in credit risk management is to assign a credit grade to a borrower. Grades are used to rank customers according to their perceived creditworthiness: better grades mean less risky customers; similar grades mean similar level of risk. Grades come in two categories: credit ratings and credit scores. Credit ratings are a small number of discrete classes, usually labeled with letters, such as 'AAA', 'BB-', etc. Credit scores are numeric grades such as '640' or '720'. Credit grades are one of the key elements in regulatory frameworks, such as Basel II (see Basel Committee on Banking Supervision [3]).

Assigning a credit grade involves analyzing information on the borrower. If the borrower is an individual, information of interest could be the individual's income, outstanding debt (mortgage, credit cards), household size, residential status, etc. For corporate borrowers, one may consider certain financial ratios (e.g., sales divided by total assets), industry, etc. Here, we refer to these pieces of information about a borrower as *features* or *predictors*. Different institutions use different predictors, and they may also have different rating classes or score ranges to rank their customers. For relatively small loans offered to a large market of potential borrowers (e.g., credit cards), it is common to use credit scores, and the process of grading a borrower is usually automated. For larger loans, accessible to small- to medium-sized companies and larger corporations, credit ratings are usually used, and the grading process may involve a combination of automated algorithms and expert analysis.

There are rating agencies that keep track of the creditworthiness of companies. Yet, most banks develop an internal methodology to assign credit grades for their customers. Rating a customer internally can be a necessity if the customer has not been rated by a rating agency, but even if a third-party rating exists, an internal rating offers a complementary assessment of a customer's risk profile.

This example shows how MATLAB® can help with the automated stage of a credit rating process. In particular, we take advantage of one of the statistical learning tools readily available in Statistics and Machine Learning Toolbox™, a classification algorithm known as a *bagged decision tree*.

We assume that historical information is available in the form of a data set where each record contains the features of a borrower and the credit rating that was assigned to it. These may be internal ratings, assigned by a committee that followed policies and procedures already in place. Alternatively, the ratings may come from a rating agency, whose ratings are being used to "jump start" a new internal credit rating system.

The existing historical data is the starting point, and it is used to *train* the bagged decision tree that will automate the credit rating. In the vocabulary of statistical learning, this training process falls in the category of *supervised learning*. The classifier is then used to assign ratings to new customers. In practice, these automated or *predicted* ratings would most likely be regarded as tentative, until a credit committee of experts reviews them. The type of classifier we use here can also facilitate the revision of these ratings, because it provides a measure of certainty for the predicted ratings, a *classification score*.

In practice, one needs to train a classifier first, then use it to assign a credit rating to new customers, and finally one also needs to *profile* or *evaluate the quality* or accuracy of the classifier, a process also known as *validation* or *back-testing*. We discuss some readily available back-testing tools, as well.

Loading the Existing Credit Rating Data

We load the historical data from the comma-delimited text file `CreditRating_Historical.dat`. We choose to work with text files here, but users with access to Database Toolbox™ can certainly load this information directly from a database.

The data set contains financial ratios, industry sector, and credit ratings for a list of corporate customers. This is simulated, not real data. The first column is a customer ID. Then we have five columns of financial ratios. These are the same ratios used in Altman's z-score (see Altman [1]; see also Loeffler and Posch [4] for a related analysis).

- Working capital / Total Assets (WC_TA)
- Retained Earnings / Total Assets (RE_TA)
- Earnings Before Interests and Taxes / Total Assets (EBIT_TA)
- Market Value of Equity / Book Value of Total Debt (MVE_BVTD)
- Sales / Total Assets (S_TA)

Next, we have an industry sector label, an integer value ranging from 1 to 12. The last column has the credit rating assigned to the customer. We load the data into a `table` array.

```
creditDS = readtable('CreditRating_Historical.dat');
```

We copy the features into a matrix `X`, and the corresponding classes, the ratings, into a vector `Y`. This is not a required step, since we could access this information directly from the `dataset` or `table` array, but we do it here to simplify some repeated function calls below.

The features to be stored in the matrix `X` are the five financial ratios, and the industry label. `Industry` is a categorical variable, *nominal* in fact, because there is no ordering in the industry sectors. The response variable, the credit ratings, is also categorical, though this is an *ordinal* variable, because, by definition, ratings imply a *ranking* of creditworthiness. We can use this variable "as is" to train our classifier. Here we choose to copy it into an *ordinal array* because this way the outputs come out in the natural order of the ratings and are easier to read. The ordering of the ratings is established by the cell array we pass as a third argument in the definition of `Y`. The credit ratings can also be mapped into numeric values, which can be useful to try alternative methods to analyze the data (e.g., regression). It is always recommended to try different methods in practice.

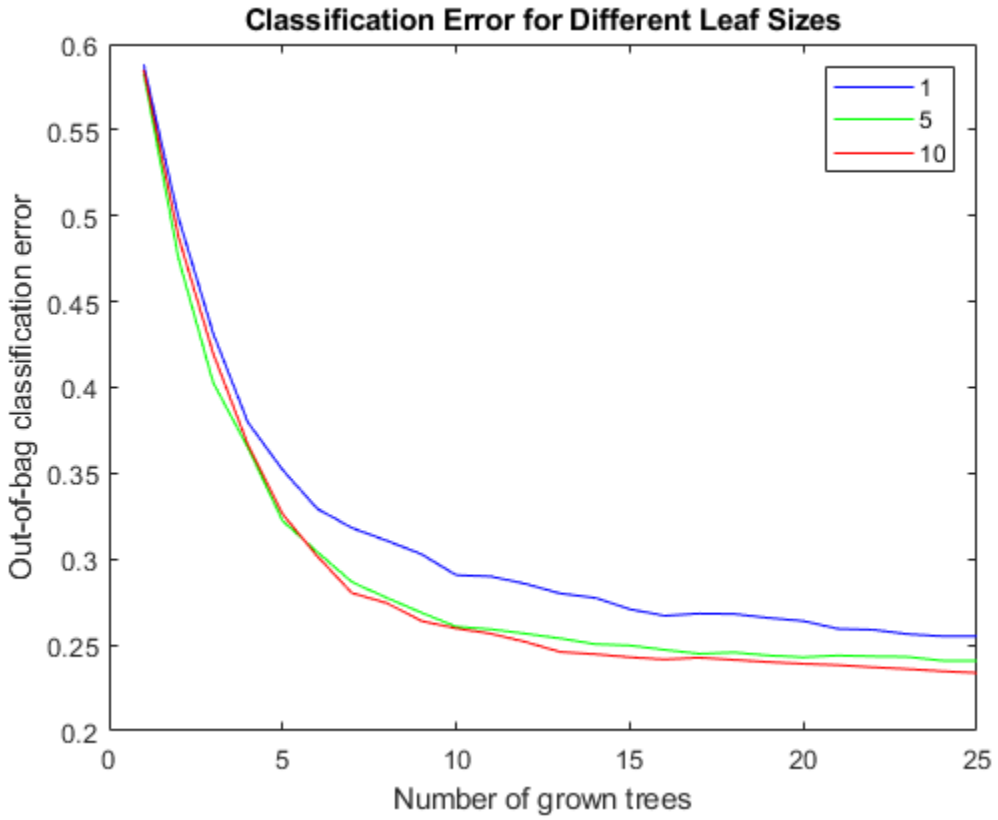
```
X = [creditDS.WC_TA creditDS.RE_TA creditDS.EBIT_TA creditDS.MVE_BVTD...
     creditDS.S_TA creditDS.Industry];
Y = ordinal(creditDS.Rating);
```

We use the predictors `X` and the response `Y` to fit a particular type of classification ensemble called a *bagged decision tree*. "Bagging," in this context, stands for "bootstrap aggregation." The methodology consists in generating a number of sub-samples, or *bootstrap replicas*, from the data set. These sub-samples are randomly generated, sampling with replacement from the list of customers in the data set. For each replica, a decision tree is grown. Each decision tree is a trained classifier on its own, and could be used in isolation to classify new customers. The predictions of two trees grown from two different bootstrap replicas may be different, though. The ensemble *aggregates* the predictions of all the decision trees that are grown for all the bootstrap replicas. If the majority of the trees predict one particular class for a new customer, it is reasonable to consider that prediction to be more robust than the prediction of any single tree alone. Moreover, if a different class is predicted by a smaller set of trees, that information is useful, too. In fact, the proportion of trees that predict different classes is the basis for the *classification scores* that are reported by the ensemble when classifying new data.

Constructing the Tree Bagger

The first step to construct our classification ensemble will be to find a good leaf size for the individual trees; here we try sizes of 1, 5 and 10. (See Statistics and Machine Learning Toolbox documentation for more on `TreeBagger`.) We start with a small number of trees, 25 only, because we mostly want to compare the initial trend in the classification error for different leaf sizes. For reproducibility and fair comparisons, we reinitialize the random number generator, which is used to sample with replacement from the data, each time we build a classifier.

```
leaf = [1 5 10];
nTrees = 25;
rng(9876,'twister');
savedRng = rng; % save the current RNG settings

color = 'bgr';
for ii = 1:length(leaf)
   % Reinitialize the random number generator, so that the
   % random samples are the same for each leaf size
   rng(savedRng)
   % Create a bagged decision tree for each leaf size and plot out-of-bag
   % error 'oobError'
   b = TreeBagger(nTrees,X,Y,'OOBPrediction','on',...
                             'CategoricalPredictors',6,...
                             'MinLeafSize',leaf(ii));
   plot(oobError(b),color(ii))
   hold on
end
xlabel('Number of grown trees')
ylabel('Out-of-bag classification error')
legend({'1', '5', '10'},'Location','NorthEast')
title('Classification Error for Different Leaf Sizes')
hold off
```

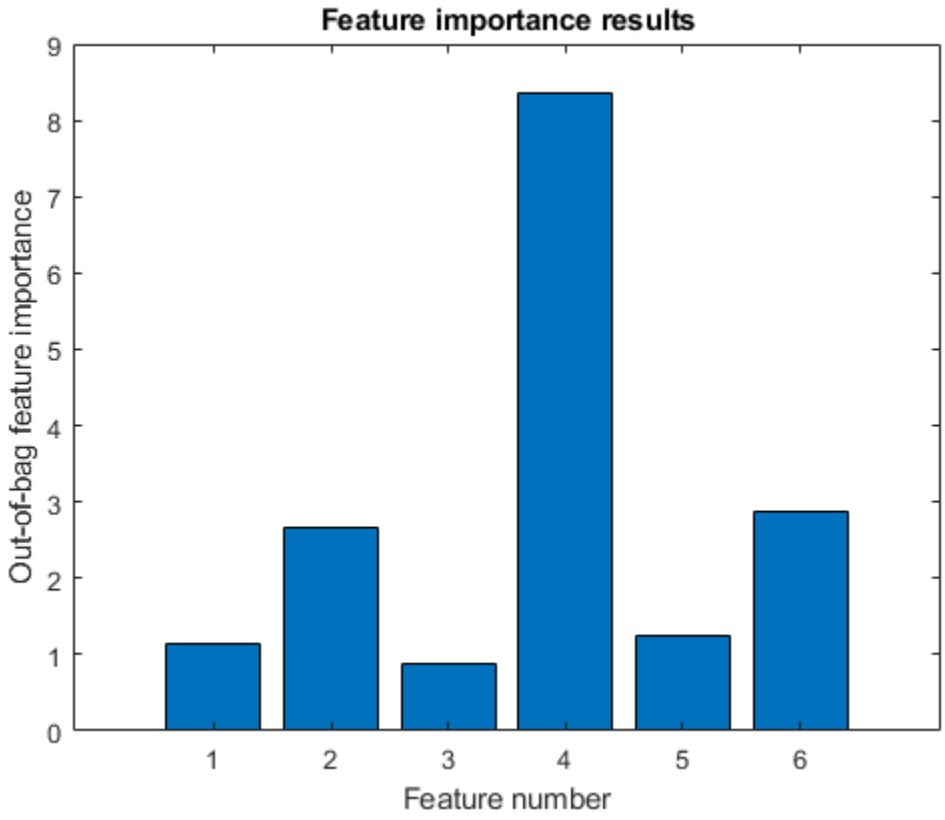


The errors are comparable for the three leaf-size options. We will therefore work with a leaf size of 10, because it results in leaner trees and more efficient computations.

Note that we did not have to split the data into *training* and *test* subsets. This is done internally, it is implicit in the sampling procedure that underlies the method. At each bootstrap iteration, the bootstrap replica is the training set, and any customers left out ("out-of-bag") are used as test points to estimate the out-of-bag classification error reported above.

Next, we want to find out whether all the features are important for the accuracy of our classifier. We do this by turning on the *feature importance* measure (`OOBPredictorImportance`), and plot the results to visually find the most important features. We also try a larger number of trees now, and store the classification error, for further comparisons below.

```
nTrees = 50;
leaf = 10;
rng(savedRng);
b = TreeBagger(nTrees,X,Y,'OOBPredictorImportance','on',...
                          'CategoricalPredictors',6,...
                          'MinLeafSize',leaf);

bar(b.OOBPermutedPredictorDeltaError)
xlabel('Feature number')
ylabel('Out-of-bag feature importance')
title('Feature importance results')

oobErrorFullX = oobError(b);
```



Features 2, 4 and 6 stand out from the rest. Feature 4, market value of equity / book value of total debt (`MVE_BVTD`), is the most important predictor for this data set. This ratio is closely related to the predictors of creditworthiness in structural models, such as Merton's model [5], where the value of the firm's equity is compared to its outstanding debt to determine the default probability.

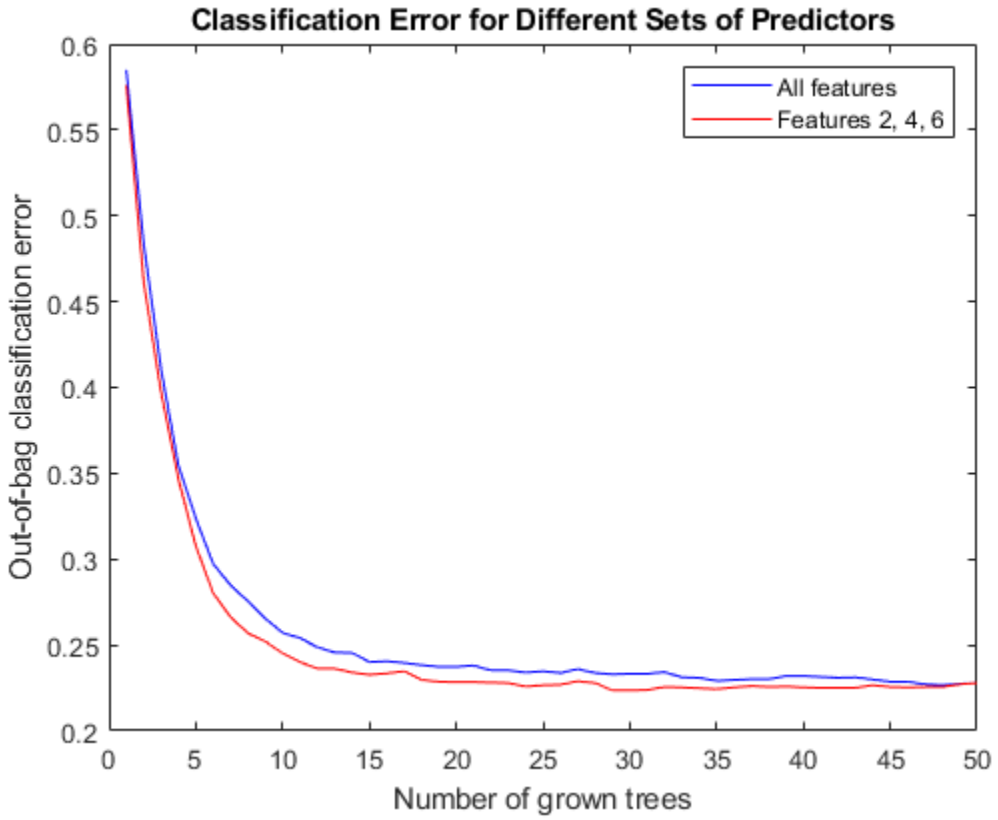
Information on the industry sector, feature 6 (`Industry`), is also relatively more important than other variables to assess the creditworthiness of a firm for this data set.

Although not as important as `MVE_BVTD`, feature 2, retained earnings / total assets (`RE_TA`), stands out from the rest. There is a correlation between retained earnings and the age of a firm (the longer a firm has existed, the more earnings it can accumulate, in general), and in turn the age of a firm is correlated to its creditworthiness (older firms tend to be more likely to survive in tough times).

Let us fit a new classification ensemble using only predictors `RE_TA`, `MVE_BVTD`, and `Industry`. We compare its classification error with the previous classifier, which uses all features.

```
X = [creditDS.RE_TA creditDS.MVE_BVTD creditDS.Industry];

rng(savedRng)
b = TreeBagger(nTrees,X,Y,'OOBPrediction','on',...
                          'CategoricalPredictors',3,...
                          'MinLeafSize',leaf);

oobErrorX246 = oobError(b);

plot(oobErrorFullX,'b')
hold on
```

```
plot(oobErrorX246,'r')
xlabel('Number of grown trees')
ylabel('Out-of-bag classification error')
legend({'All features', 'Features 2, 4, 6'},'Location','NorthEast')
title('Classification Error for Different Sets of Predictors')
hold off
```



The accuracy of the classification does not deteriorate significantly when we remove the features with relatively low importance (1, 3, and 5), so we will use the more parsimonious classification ensemble for our predictions.

In this example, we have started with a set of six features only, and used the feature importance measure of the classifier, and the out-of-bag classification error as criteria to screen out three of the variables. Feature selection can be a time consuming process when the initial set of potential predictors contains dozens of variables. Besides the tools we have used here (variable importance and a "visual" comparison of out-of-bag errors), other variable selection tools in Statistics and Machine Learning Toolbox can be helpful for these types of analyses (see documentation). However, in the end, a successful feature selection process requires a combination of quantitative tools and an analyst's judgement.

For example, the variable importance measure we used here is a ranking mechanism that estimates the relative impact of a feature by measuring how much the predictive accuracy of the classifier deteriorates when this feature's values are randomly permuted. The idea is that when the feature in question adds little to the predictive power of the classifier, using altered (in this case permuted) values should not impact the classification results. Relevant information, on the other hand, cannot be randomly swapped without degrading the predictions. Now, if two highly correlated features are

important, they will both rank high in this analysis. In that case, keeping one of these features should suffice for accurate classifications, but one would not know that from the ranking results alone. One would have to check the correlations separately, or use an expert's judgement. That is to say, tools like variable importance or `sequentialfs` can greatly help for feature selection, but an analyst's judgment is a key piece in this process.

At this point, the classifier could be saved (e.g., `save classifier.mat b`), to be loaded in a future session (`load classifier`) to classify new customers. For efficiency, it is recommended to keep a compact version of the classifier once the training process is finished.

```
b = compact(b);
```

Classifying New Data

Here we use the previously constructed classification ensemble to assign credit ratings to new customers. Because the ratings of existing customers need to be reviewed, too, on a regular basis, especially when their financial information has substantially changed, the data set could also contain a list of existing customers under review. We start by loading the new data.

```
newDS = readtable('CreditRating_NewCompanies.dat');
```

To predict the credit rating for this new data, we call the `predict` method on the classifier. The method returns two arguments, the predicted class and the classification score. We certainly want to get both output arguments, since the classification scores contain information on how certain the predicted ratings seem to be. We could copy variables `RE_TA`, `MVE_BVTD` and `Industry` into a matrix `X`, as before, but since we will make only one call to `predict`, we can skip this step and use `newDS` directly.

```
[predClass,classifScore] = predict(b,[newDS.RE_TA newDS.MVE_BVTD newDS.Industry]);
```

At this point, we can create a report. Here we only display on the screen a small report for the first three customers, for illustration purposes, but MATLAB deployment tools could greatly improve the workflow here. For example, credit analysts could run this classification remotely, using a web browser, and get a report, without even having MATLAB on their desktops.

```
for i = 1:3
   fprintf('Customer %d:\n',newDS.ID(i));
   fprintf('   RE/TA    = %5.2f\n',newDS.RE_TA(i));
   fprintf('   MVE/BVTD = %5.2f\n',newDS.MVE_BVTD(i));
   fprintf('   Industry = %2d\n',newDS.Industry(i));
   fprintf('   Predicted Rating : %s\n',predClass{i});
   fprintf('   Classification score : \n');
   for j = 1:length(b.ClassNames)
      if (classifScore(i,j)>0)
         fprintf('      %s : %5.4f \n',b.ClassNames{j},classifScore(i,j));
      end
   end
end
```

```
Customer 60644:
   RE/TA    =  0.22
   MVE/BVTD =  2.40
   Industry =  6
   Predicted Rating : AA
   Classification score :
      A : 0.2874
      AA : 0.6919
```

```
      AAA : 0.0156
      BBB : 0.0051
Customer 33083:
   RE/TA    =  0.24
   MVE/BVTD =  1.51
   Industry =  4
   Predicted Rating : BBB
   Classification score :
      A : 0.0751
      BB : 0.0017
      BBB : 0.9232
Customer 63830:
   RE/TA    =  0.18
   MVE/BVTD =  1.69
   Industry =  7
   Predicted Rating : A
   Classification score :
      A : 0.6629
      AA : 0.0067
      B : 0.0008
      BB : 0.0005
      BBB : 0.3291
```

Keeping records of the predicted ratings and corresponding scores can be useful for periodic assessments of the quality of the classifier. We store this information here in the `table` array `predDS`.

```
classnames = b.ClassNames;
predDS = [table(newDS.ID,predClass),array2table(classifScore)];
predDS.Properties.VariableNames = {'ID','PredRating',classnames{:}};
```

This information could be saved, for example, to a comma-delimited text file `PredictedRatings.dat` using the command

```
  writetable(predDS,'PredictedRatings.dat');
```

or written directly to a database using Database Toolbox.

Back-Testing: Profiling the Classification Process

Validation or *back-testing* is the process of profiling or assessing the quality of the credit ratings. There are many different measures and tests related to this task (see, for example, Basel Committee on Banking Supervision [2]). Here, we focus on the following two questions:

- How accurate are the predicted ratings, as compared to the actual ratings? Here "predicted ratings" refers to those obtained from the automated classification process, and "actual ratings" to those assigned by a credit committee that puts together the predicted ratings and their classification scores, and other pieces of information, such as news and the state of the economy to determine a final rating.
- How well do the actual ratings rank customers according to their creditworthiness? This is done in an *ex-post* analysis performed, for example, one year later, when it is known which companies defaulted during the year.

The file `CreditRating_ExPost.dat` contains "follow up" data on the same companies considered in the previous section. It contains the actual ratings that the committee assigned to these companies,

as well as a "default flag" that indicates whether the corresponding company defaulted within one year of the rating process (if 1) or not (if 0).

```
exPostDS = readtable('CreditRating_ExPost.dat');
```

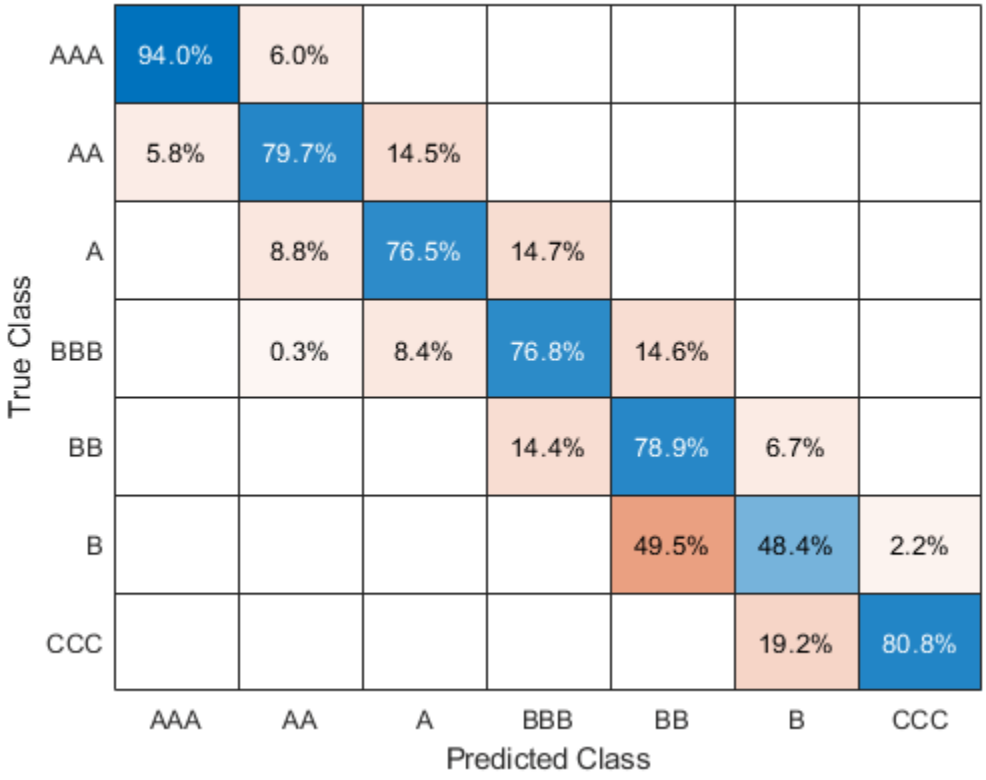
Comparing predicted ratings vs. actual ratings. The rationale to train an automated classifier is to expedite the work of the credit committee. The more accurate the predicted ratings are, the less time the committee has to spend reviewing the predicted ratings. So it is conceivable that the committee wants to have regular checks on how closely the predicted ratings match the final ratings they assign, and to recommend re-training the automated classifier (and maybe include new features, for example) if the mismatch seems concerning.

The first tool we can use to compare predicted vs. actual ratings is a *confusion matrix*, readily available in Statistics and Machine Learning Toolbox:

```
C = confusionchart(exPostDS.Rating,predDS.PredRating);
sortClasses(C,{'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'})
```

The rows correspond to the actual ratings, and the columns to the predicted ratings. The amount in the position `(i,j)` in the confusion matrix indicates how many customers received an actual rating `i` and were predicted as rating `j`. For example, position `(3,2)` tells us how many customers received a rating of 'A' by the credit committee, but were predicted as 'AA' with the automated classifier. One can also present this matrix in percentage. Normalize each value by the number of observations that has the same true rating.

```
C.Normalization = 'row-normalized';
```

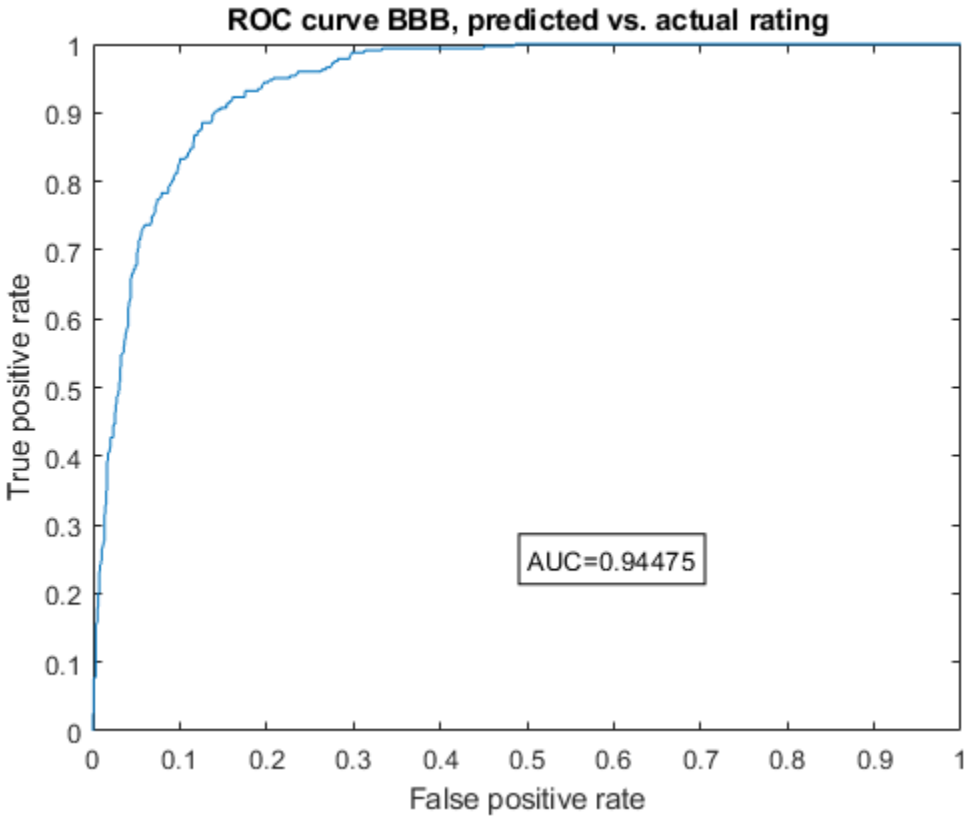


Good agreement between the predicted and the actual ratings would result in values in the main diagonal that dominate the rest of the values in a row, ideally values close to 1. In this case, we actually see an important disagreement for 'B,' since about half of the customers that were rated as 'B' by the credit committee had been predicted as 'BB' by the automated classifier. On the other hand, it is good to see that ratings differ in at most one notch in most cases, with the only exception of 'BBB.'

A confusion matrix could also be used to compare the internal ratings assigned by the institution against third-party ratings; this is often done in practice.

For each specific rating, we can compute yet another measure of agreement between predicted and actual ratings. We can build a *Receiver Operating Characteristic (ROC) curve* using the `perfcurve` function from Statistics and Machine Learning Toolbox, and check the *area under the curve (AUC)*. The `perfcurve` function takes as an argument the actual ratings, which are our benchmark, the standard we are comparing against, and the 'BBB' classification scores determined by the automated process. Let us build a ROC and calculate the AUC for rating 'BBB' in our example.

```
[xVal,yVal,~,auc] = perfcurve(exPostDS.Rating,predDS.BBB,'BBB');
plot(xVal,yVal)
xlabel('False positive rate')
ylabel('True positive rate')
text(0.5,0.25,strcat('AUC=',num2str(auc)),'EdgeColor','k')
title('ROC curve BBB, predicted vs. actual rating')
```



Here is an explanation of how the ROC is built. Recall that for each customer the automated classifier returns a classification score for each of the credit ratings, in particular, for 'BBB,' which can be interpreted as how likely it is that this particular customer should be rated 'BBB.' In order to build the ROC curve, one needs to vary the *classification threshold*. That is, the minimum score to classify a customer as 'BBB.' In other words, if the threshold is `t`, we only classify customers as 'BBB' if their 'BBB' score is greater than or equal to `t`. For example, suppose that company *XYZ* had a 'BBB' score of 0.87. If the actual rating of *XYZ* (the information in `exPostDS.Rating`) is 'BBB,' then *XYZ* would be correctly classified as 'BBB' for any threshold of up to 0.87. This would be a *true positive*, and it would increase what is call the *sensitivity* of the classifier. For any threshold greater than 0.87, this company would not receive a 'BBB' rating, and we would have a *false negative* case. To complete the description, suppose now that *XYZ*'s actual rating is 'BB.' Then it would be correctly rejected as a 'BBB' for thresholds of more than 0.87, becoming a *true negative*, and thus increasing the so called *specificity* of the classifier. However, for thresholds of up to 0.87, it would become a *false positive* (it would be classified as 'BBB,' when it actually is a 'BB'). The ROC curve is constructed by plotting the proportion of true positives (sensitivity), versus false positives (1-specificity), as the threshold varies from 0 to 1.

The AUC, as its name indicates, is the area under the ROC curve. The closer the AUC is to 1, the more accurate the classifier (a perfect classifier would have an AUC of 1). In this example, the AUC seems high enough, but it would be up to the committee to decide which level of AUC for the ratings should trigger a recommendation to improve the automated classifier.

Comparing actual ratings vs. defaults in the following year. A common tool used to assess the ranking of customers implicit in the credit ratings is the *Cumulative Accuracy Profile (CAP)*, and the associated *accuracy ratio* measure. The idea is to measure the relationship between the credit ratings

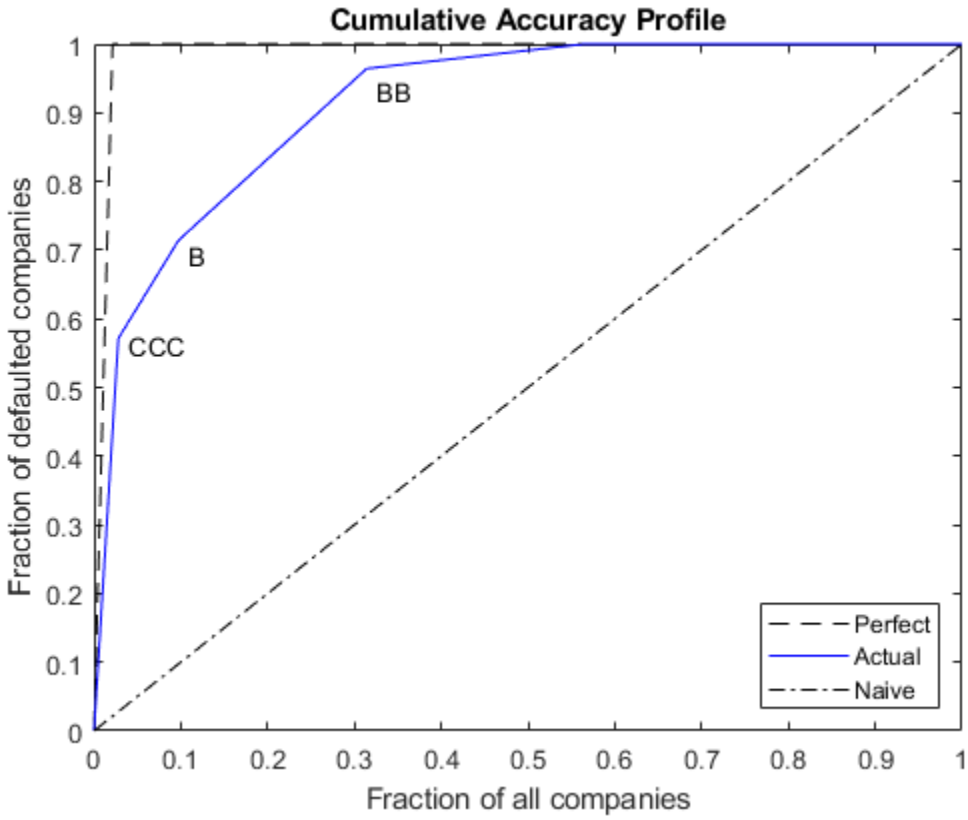
assigned and the number of defaults observed in the following year. One would expect that fewer defaults are observed for better rating classes. If the default rate were the same for all ratings, the rating system would be no different from a naive (and useless) classification system in which customers were randomly assigned a rating, independently of their creditworthiness.

It is not hard to see that the `perfcurve` function can also be used to construct the CAP. The standard we compare against is not a rating, as before, but the default flag that we loaded from the `CreditRating_ExPost.dat` file. The score we use is a "dummy score" that indicates the ranking in creditworthiness implicit in the list of ratings. The dummy score only needs to satisfy that better ratings get lower dummy scores (they are "less likely to have a default flag of 1"), and that any two customers with the same rating get the same dummy score. A default probability could be passed as a score, of course, but we do not have default probabilities here, and in fact *we do not need to have estimates of the default probabilities to construct the CAP*, because we are not validating default probabilities. All we are assessing with this tool is how well the ratings *rank* customers according to their creditworthiness.

Usually, the CAP of the rating system under consideration is plotted together with the CAP of the "perfect rating system." The latter is a hypothetical credit rating system for which the lowest rating includes all the defaulters, and no other customers. The area under this perfect curve is the maximum possible AUC attainable by a rating system. By convention, the AUC is adjusted for CAPs to subtract the area under the *naive system*'s CAP, that is, the CAP of the system that randomly assigns ratings to customers. The naive system's CAP is simply a straight line from the origin to (1,1), with an AUC of 0.5. The *accuracy ratio* for a rating system is then defined as the ratio of the adjusted AUC (AUC of the system in consideration minus AUC of the naive system) to the maximum accuracy (AUC of the perfect system minus AUC of the naive system).

```
ratingsList = {'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'};
Nratings = length(ratingsList);
dummyDelta = 1/(Nratings+1);
dummyRank = linspace(dummyDelta,1-dummyDelta,Nratings)';

D = exPostDS.Def_tplus1;
fracTotDef = sum(D)/length(D);
maxAcc = 0.5 - 0.5 * fracTotDef;

R = double(ordinal(exPostDS.Rating,[],ratingsList));
S = dummyRank(R);
[xVal,yVal,~,auc] = perfcurve(D,S,1);

accRatio = (auc-0.5)/maxAcc;
fprintf('Accuracy ratio for actual ratings: %5.3f\n',accRatio);

xPerfect(1) = 0; xPerfect(2) = fracTotDef; xPerfect(3) = 1;
yPerfect(1) = 0; yPerfect(2) = 1; yPerfect(3) = 1;
xNaive(1) = 0; xNaive(2) = 1;
yNaive(1) = 0; yNaive(2) = 1;

plot(xPerfect,yPerfect,'--k',xVal,yVal,'b',xNaive,yNaive,'-.k')
xlabel('Fraction of all companies')
ylabel('Fraction of defaulted companies')
title('Cumulative Accuracy Profile')
legend({'Perfect','Actual','Naive'},'Location','SouthEast')
text(xVal(2)+0.01,yVal(2)-0.01,'CCC')
text(xVal(3)+0.01,yVal(3)-0.02,'B')
text(xVal(4)+0.01,yVal(4)-0.03,'BB')
```

```
Accuracy ratio for actual ratings: 0.850
```

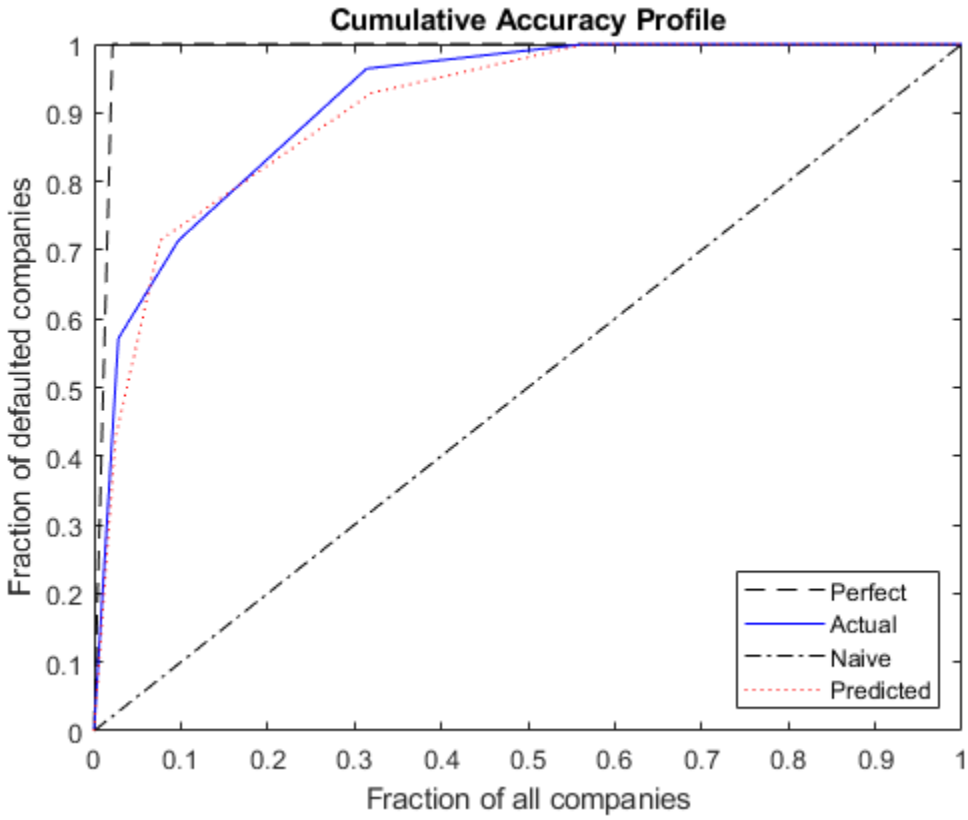


The key to reading the information of the CAP is in the "kinks," labeled in the plot for ratings 'CCC,' 'B,' and 'BB.' For example, the second kink is associated with the second lowest rating, 'B,' and it is located at (0.097, 0.714). This means that 9.7% of the customers were ranked 'B' *or lower*, and they account for 71.4% of the defaults observed.

In general, the accuracy ratio should be treated as a relative, rather than an absolute measure. For example, we can add the CAP of the predicted ratings in the same plot, and compute its accuracy ratio to compare it with the accuracy ratio of the actual ratings.

```
Rpred = double(ordinal(predDS.PredRating,[],ratingsList));
Spred = dummyRank(Rpred);
[xValPred,yValPred,~,aucPred] = perfcurve(D,Spred,1);

accRatioPred = (aucPred-0.5)/maxAcc;
fprintf('Accuracy ratio for predicted ratings: %5.3f\n',accRatioPred);

plot(xPerfect,yPerfect,'--k',xVal,yVal,'b',xNaive,yNaive,'-.k',...
   xValPred,yValPred,':r')
xlabel('Fraction of all companies')
ylabel('Fraction of defaulted companies')
title('Cumulative Accuracy Profile')
legend({'Perfect','Actual','Naive','Predicted'},'Location','SouthEast')
```

```
Accuracy ratio for predicted ratings: 0.830
```



The accuracy ratio of the predicted rating is smaller, and its CAP is mostly below the CAP of the actual rating. This is reasonable, since the actual ratings are assigned by the credit committees that take into consideration the predicted ratings *and* extra information that can be important to fine-tune the ratings.

Final Remarks

MATLAB offers a wide range of machine learning tools, besides bagged decision trees, that can be used in the context of credit rating. In Statistics and Machine Learning Toolbox you can find classification tools such as discriminant analysis and naive Bayes classifiers. MATLAB also offers Deep Learning Toolbox™. Also, Database Toolbox and MATLAB deployment tools may provide you with more flexibility to adapt the workflow presented here to your own preferences and needs.

No probabilities of default have been computed here. For credit ratings, the probabilities of default are usually computed based on credit-rating migration history. See the `transprob` reference page in Financial Toolbox™ for more information.

Bibliography

[1] Altman, E., "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance*, Vol. 23, No. 4, (Sep., 1968), pp. 589-609.

[2] Basel Committee on Banking Supervision, "Studies on the Validation of Internal Rating Systems," Bank for International Settlements (BIS), Working Papers No. 14, revised version, May 2005. Available at: https://www.bis.org/publ/bcbs_wp14.htm.

[3] Basel Committee on Banking Supervision, "International Convergence of Capital Measurement and Capital Standards: A Revised Framework," Bank for International Settlements (BIS), comprehensive version, June 2006. Available at: https://www.bis.org/publ/bcbsca.htm.

[4] Loeffler, G., and P. N. Posch, *Credit Risk Modeling Using Excel and VBA*, West Sussex, England: Wiley Finance, 2007.

[5] Merton, R., "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *Journal of Finance*, Vol. 29, No. 2, (May, 1974), pp. 449-70.

Combine Heterogeneous Models into Stacked Ensemble

This example shows how to build multiple machine learning models for a given training data set, and then combine the models using a technique called *stacking* to improve the accuracy on a test data set compared to the accuracy of the individual models.

Stacking is a technique used to combine several heterogeneous models by training an additional model, often referred to as a *stacked ensemble model*, or *stacked learner*, on the *k*-fold cross-validated predictions (classification scores for classification models and predicted responses for regression models) of the original (base) models. The concept behind stacking is that certain models might correctly classify a test observation while others might fail to do so. The algorithm learns from this diversity of predictions and attempts to combine the models to improve upon the predicted accuracy of the base models.

In this example, you train several heterogeneous classification models on a data set, and then combine the models using stacking.

Load Sample Data

This example uses the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

Load the sample data `census1994` and display the variables in the data set.

```
load census1994
whos
```

```
  Name                 Size              Bytes  Class    Attributes

  Description         20x74               2960  char
  adultdata        32561x15            1872567  table
  adulttest        16281x15             944467  table
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. For this example, to reduce the running time, subsample 5000 training and test observations each, from the original tables `adultdata` and `adulttest`, by using the `datasample` function. (You can skip this step if you want to use the complete data sets.)

```
NumSamples = 5e3;
s = RandStream('mlfg6331_64','seed',0); % For reproducibility
adultdata = datasample(s,adultdata,NumSamples,'Replace',false);
adulttest = datasample(s,adulttest,NumSamples,'Replace',false);
```

Some models, such as support vector machines (SVMs), remove observations containing missing values whereas others, such as decision trees, do not remove such observations. To maintain consistency between the models, remove rows containing missing values before fitting the models.

```
adultdata = rmmissing(adultdata);
adulttest = rmmissing(adulttest);
```

Preview the first few rows of the training data set.

```
head(adultdata)
```

```
ans=8×15 table
    age     workClass       fnlwgt       education      education_num      marital_status
    ___    ___________    __________    _____________    _____________    __________________

    39     Private          4.91e+05    Bachelors             13          Never-married
    25     Private        2.2022e+05    11th                   7          Never-married
    24     Private        2.2761e+05    10th                   6          Divorced
    51     Private        1.7329e+05    HS-grad                9          Divorced
    54     Private        2.8029e+05    Some-college          10          Married-civ-spouse
    53     Federal-gov         39643    HS-grad                9          Widowed
    52     Private             81859    HS-grad                9          Married-civ-spouse
    37     Private        1.2429e+05    Some-college          10          Married-civ-spouse
```

Each row represents the attributes of one adult, such as age, education, and occupation. The last column `salary` shows whether a person has a salary less than or equal to $50,000 per year or greater than $50,000 per year.

Understand Data and Choose Classification Models

Statistics and Machine Learning Toolbox™ provides several options for classification, including classification trees, discriminant analysis, naive Bayes, nearest neighbors, SVMs, and classification ensembles. For the complete list of algorithms, see “Classification”.

Before choosing the algorithms to use for your problem, inspect your data set. The census data has several noteworthy characteristics:

- The data is tabular and contains both numeric and categorical variables.
- The data contains missing values.
- The response variable (`salary`) has two classes (binary classification).

Without making any assumptions or using prior knowledge of algorithms that you expect to work well on your data, you simply train all the algorithms that support tabular data and binary classification. Error-correcting output codes (ECOC) models are used for data with more than two classes. Discriminant analysis and nearest neighbor algorithms do not analyze data that contains both numeric and categorical variables. Therefore, the algorithms appropriate for this example are an SVM, a decision tree, an ensemble of decision trees, and a naive Bayes model.

Build Base Models

Fit two SVM models, one with a Gaussian kernel and one with a polynomial kernel. Also, fit a decision tree, a naive Bayes model, and an ensemble of decision trees.

```
% SVM with Gaussian kernel
rng('default') % For reproducibility
mdls{1} = fitcsvm(adultdata,'salary','KernelFunction','gaussian', ...
    'Standardize',true,'KernelScale','auto');

% SVM with polynomial kernel
rng('default')
mdls{2} = fitcsvm(adultdata,'salary','KernelFunction','polynomial', ...
    'Standardize',true,'KernelScale','auto');

% Decision tree
rng('default')
mdls{3} = fitctree(adultdata,'salary');
```

```
% Naive Bayes
rng('default')
mdls{4} = fitcnb(adultdata,'salary');

% Ensemble of decision trees
rng('default')
mdls{5} = fitcensemble(adultdata,'salary');
```

Combine Models Using Stacking

If you use only the prediction scores of the base models on the training data, the stacked ensemble might be subject to overfitting. To reduce overfitting, use the *k*-fold cross-validated scores instead. To ensure that you train each model using the same *k*-fold data split, create a `cvpartition` object and pass that object to the `crossval` function of each base model. This example is a binary classification problem, so you only need to consider scores for either the positive or negative class.

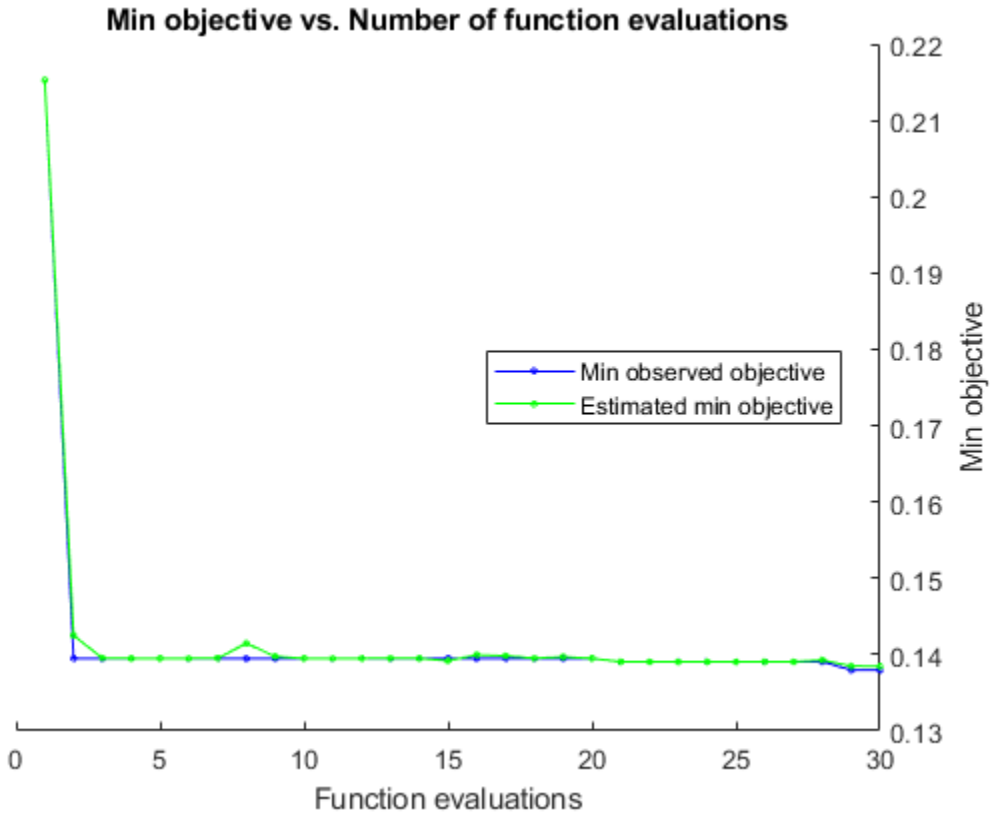
Obtain *k*-fold cross-validation scores.

```
rng('default') % For reproducibility
N = numel(mdls);
Scores = zeros(size(adultdata,1),N);
cv = cvpartition(adultdata.salary,"KFold",5);
for ii = 1:N
    m = crossval(mdls{ii},'cvpartition',cv);
    [~,s] = kfoldPredict(m);
    Scores(:,ii) = s(:,m.ClassNames=='<=50K');
end
```

Create the stacked ensemble by training it on the cross-validated classification scores `Scores` with these options:

- To obtain the best results for the stacked ensemble, optimize its hyperparameters. You can fit the training data set and tune parameters easily by calling the fitting function and setting its `'OptimizeHyperparameters'` name-value pair argument to `'auto'`.
- Specify `'Verbose'` as 0 to disable message displays.
- For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function. Also, for reproducibility of the random forest algorithm, specify the `'Reproducible'` name-value pair argument as `true` for tree learners.

```
rng('default') % For reproducibility
t = templateTree('Reproducible',true);
stckdMdl = fitcensemble(Scores,adultdata.salary, ...
    'OptimizeHyperparameters','auto', ...
    'Learners',t, ...
    'HyperparameterOptimizationOptions',struct('Verbose',0,'AcquisitionFunctionName','expected-i
```



Compare Predictive Accuracy

Check the classifier performance with the test data set by using the confusion matrix and McNemar's hypothesis test.

Predict Labels and Scores on Test Data

Find the predicted labels, scores, and loss values of the test data set for the base models and the stacked ensemble.

First, iterate over the base models to the compute predicted labels, scores, and loss values.

```
label = [];
score = zeros(size(adulttest,1),N);
mdlLoss = zeros(1,numel(mdls));
for i = 1:N
    [lbl,s] = predict(mdls{i},adulttest);
    label = [label,lbl];
    score(:,i) = s(:,m.ClassNames=='<=50K');
    mdlLoss(i) = mdls{i}.loss(adulttest);
end
```

Attach the predictions from the stacked ensemble to `label` and `mdlLoss`.

```
[lbl,s] = predict(stckdMdl,score);
label = [label,lbl];
mdlLoss(end+1) = stckdMdl.loss(score,adulttest.salary);
```

Concatenate the score of the stacked ensemble to the scores of the base models.

```
score = [score,s(:,1)];
```

Display the loss values.

```
names = {'SVM-Gaussian','SVM-Polynomial','Decision Tree','Naive Bayes', ...
    'Ensemble of Decision Trees','Stacked Ensemble'};
array2table(mdlLoss,'VariableNames',names)
```

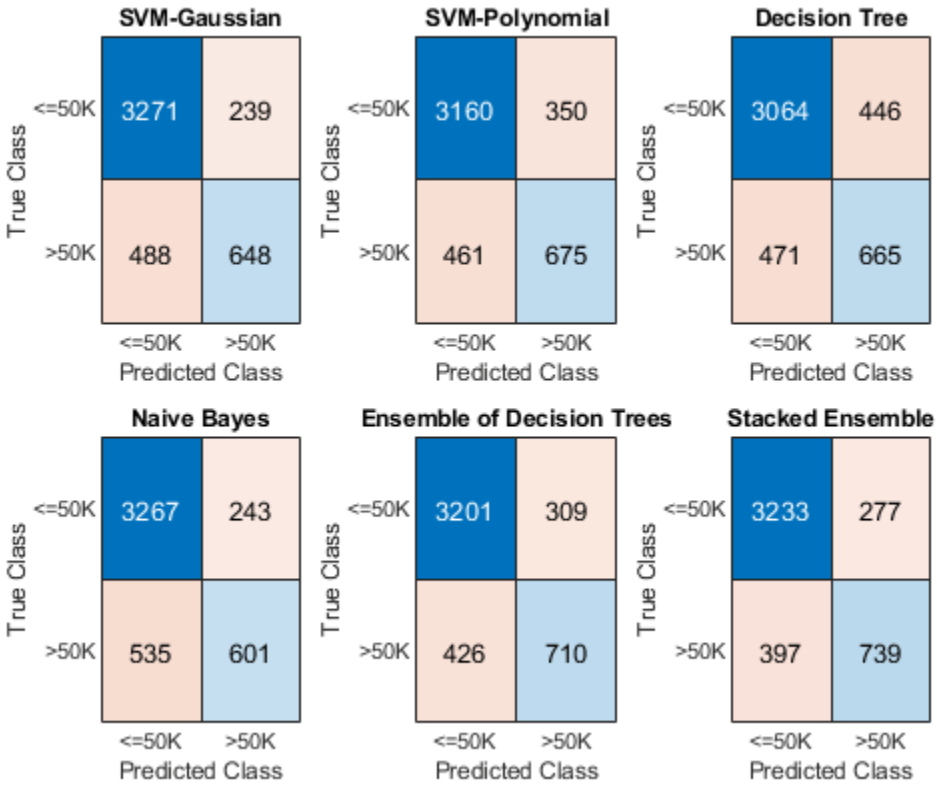
```
ans=1×6 table
    SVM-Gaussian    SVM-Polynomial    Decision Tree    Naive Bayes    Ensemble of Decision Trees
    ____________    ______________    _____________    ___________    __________________________

      0.15668          0.17473           0.1975          0.16764               0.15833
```

The loss value of the stacked ensemble is lower than the loss values of the base models.

Confusion Matrix

Compute the confusion matrix with the predicted classes and the known (true) classes of the test data set by using the `confusionchart` function.

```
figure
c = cell(N+1,1);
for i = 1:numel(c)
    subplot(2,3,i)
    c{i} = confusionchart(adulttest.salary,label(:,i));
    title(names{i})
end
```



The diagonal elements indicate the number of correctly classified instances of a given class. The off-diagonal elements are instances of misclassified observations.

McNemar's Hypothesis Test

To test whether the improvement in prediction is significant, use the `testcholdout` function, which conducts McNemar's hypothesis test. Compare the stacked ensemble to the naive Bayes model.

```
[hNB,pNB] = testcholdout(label(:,6),label(:,4),adulttest.salary)
```

```
hNB = logical
   1


pNB = 9.7646e-07
```

Compare the stacked ensemble to the ensemble of decision trees.

```
[hE,pE] = testcholdout(label(:,6),label(:,5),adulttest.salary)
```

```
hE = logical
   1


pE = 1.9357e-04
```

In both cases, the low *p*-value of the stacked ensemble confirms that its predictions are statistically superior to those of the other models.

Label Data Using Semi-Supervised Learning Techniques

This example shows how to use graph-based and self-training semi-supervised learning techniques to label data.

Semi-supervised learning combines aspects of supervised learning, where all of the training data is labeled, and unsupervised learning, where true labels are unknown. That is, some training observations are labeled, but the vast majority are unlabeled. Semi-supervised learning methods try to leverage the underlying structure of the data to fit labels to the unlabeled data.

Statistics and Machine Learning Toolbox™ provides these semi-supervised learning functions for classification:

- `fitsemigraph` constructs a similarity graph with labeled and unlabeled observations as nodes, and distributes label information from labeled observations to unlabeled observations.
- `fitsemiself` iteratively trains a classifier on the data. First, the function trains a classifier on the labeled data alone, and then uses that classifier to make label predictions for the unlabeled data. `fitsemiself` provides scores for the predictions, and then treats the predictions as true labels for the next training cycle of the classifier if the scores are above a certain threshold. This process repeats until the label predictions converge.

Generate Data

Generate data from two half-moon shapes. Determine which moon new points belong to by using graph-based and self-training semi-supervised techniques.

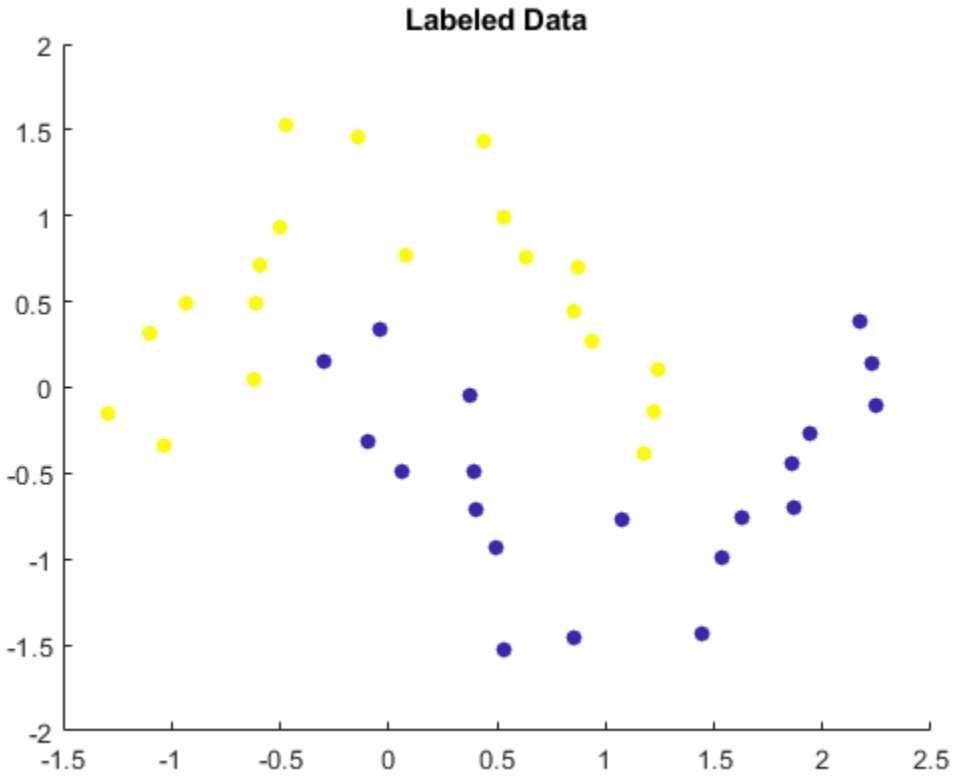
Create the custom function `twomoons` (shown at the end of this example). This function takes an input argument `n` and creates `n` points in each of two interlaced half-moons: a top moon that is concave down and a bottom moon that is concave up.

Generate a set of 40 labeled data points by using the `twomoons` function. Each point in `X` is in one of the two moons, with the corresponding moon label stored in the vector `label`.

```
rng('default') % For reproducibility
[X,label] = twomoons(20);
```

Visualize the points by using a scatter plot. Points in the same moon have the same color.

```
scatter(X(:,1),X(:,2),[],label,'filled')
title('Labeled Data')
```



Generate a set of 400 unlabeled data points by using the `twomoons` function. Each point in `newX` belongs to one of the two moons, but the corresponding moon label is unknown.

```
newX = twomoons(200);
```

Label Data Using Graph-Based Method

Label the unlabeled data in `newX` by using a semi-supervised graph-based method. By default, `fitsemigraph` constructs a similarity graph from the data in `X` and `newX`, and uses a label propagation technique to fit labels to `newX`.

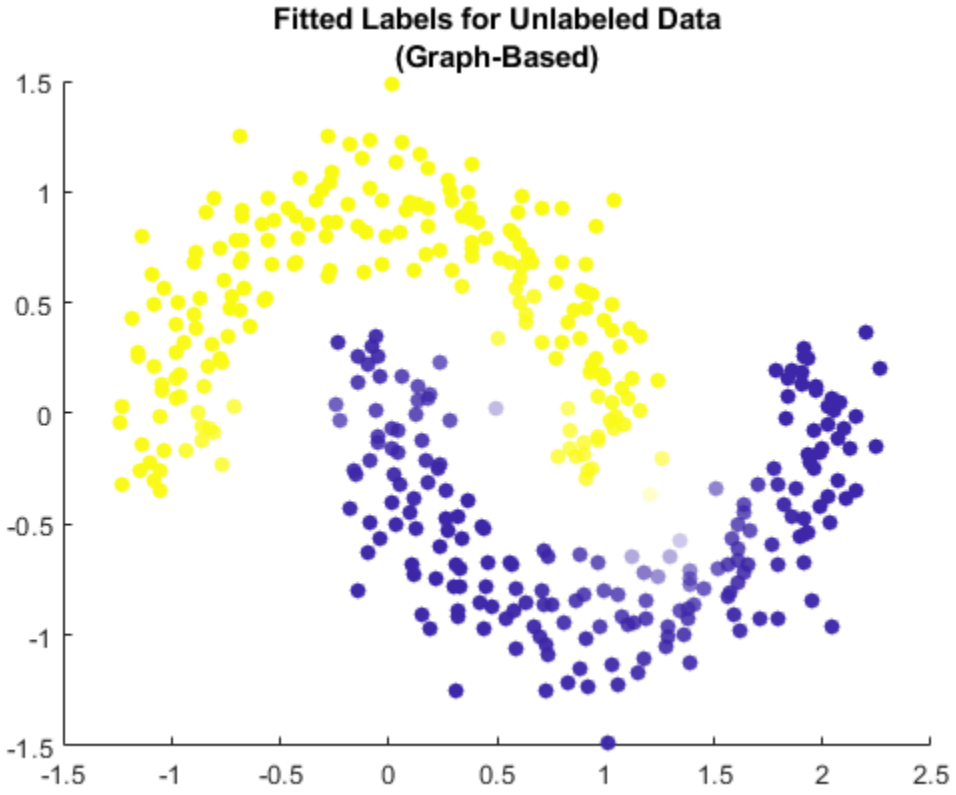
```
graphMdl = fitsemigraph(X,label,newX)
```

```
graphMdl =
  SemiSupervisedGraphModel with properties:

            FittedLabels: [400x1 double]
             LabelScores: [400x2 double]
               ClassNames: [1 2]
            ResponseName: 'Y'
    CategoricalPredictors: []
                  Method: 'labelpropagation'


  Properties, Methods
```

The function returns a `SemiSupervisedGraphModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

Visualize the fitted label results by using a scatter plot. Use the fitted labels to set the color of the points, and use the maximum label scores to set the transparency of the points. Points with less transparency are labeled with greater confidence.

```
maxGraphScores = max(graphMdl.LabelScores,[],2);
rescaledGraphScores = rescale(maxGraphScores,0.05,0.95);
scatter(newX(:,1),newX(:,2),[],graphMdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledGraphScores);
title(["Fitted Labels for Unlabeled Data","(Graph-Based)"])
```



This method seems to label the `newX` points accurately. The two moons are visually distinct, and the points that are labeled with the most uncertainty lie on the boundary between the two shapes.

Label Data Using Self-Training Method

Label the unlabeled data in `newX` by using a semi-supervised self-training method. By default, `fitsemiself` uses a support vector machine (SVM) model with a Gaussian kernel to label the data iteratively.

```
selfSVMMdl = fitsemiself(X,label,newX)
```

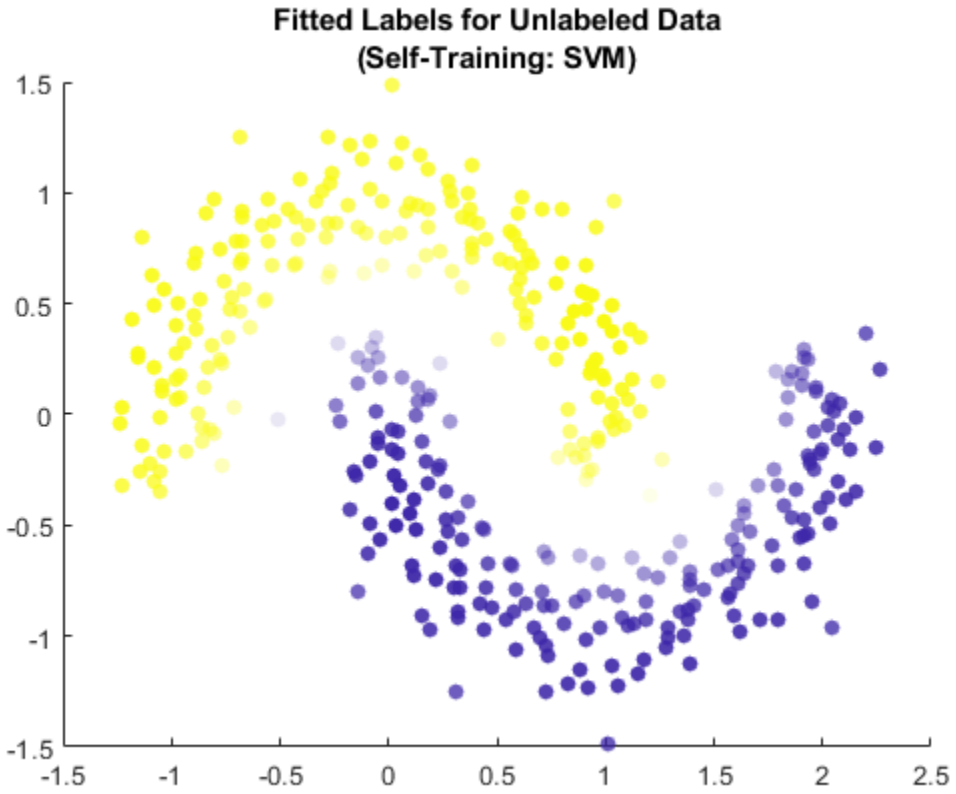
```
selfSVMMdl =
  SemiSupervisedSelfTrainingModel with properties:
```

```
           FittedLabels: [400x1 double]
            LabelScores: [400x2 double]
              ClassNames: [1 2]
           ResponseName: 'Y'
  CategoricalPredictors: []
                Learner: [1x1 classreg.learning.classif.CompactClassificationSVM]


  Properties, Methods
```

The function returns a `SemiSupervisedSelfTrainingModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

Visualize the fitted label results by using a scatter plot. As before, use the fitted labels to set the color of the points, and use the maximum label scores to set the transparency of the points.

```
maxSVMScores = max(selfSVMMdl.LabelScores,[],2);
rescaledSVMScores = rescale(maxSVMScores,0.05,0.95);
scatter(newX(:,1),newX(:,2),[],selfSVMMdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledSVMScores);
title(["Fitted Labels for Unlabeled Data","(Self-Training: SVM)"])
```



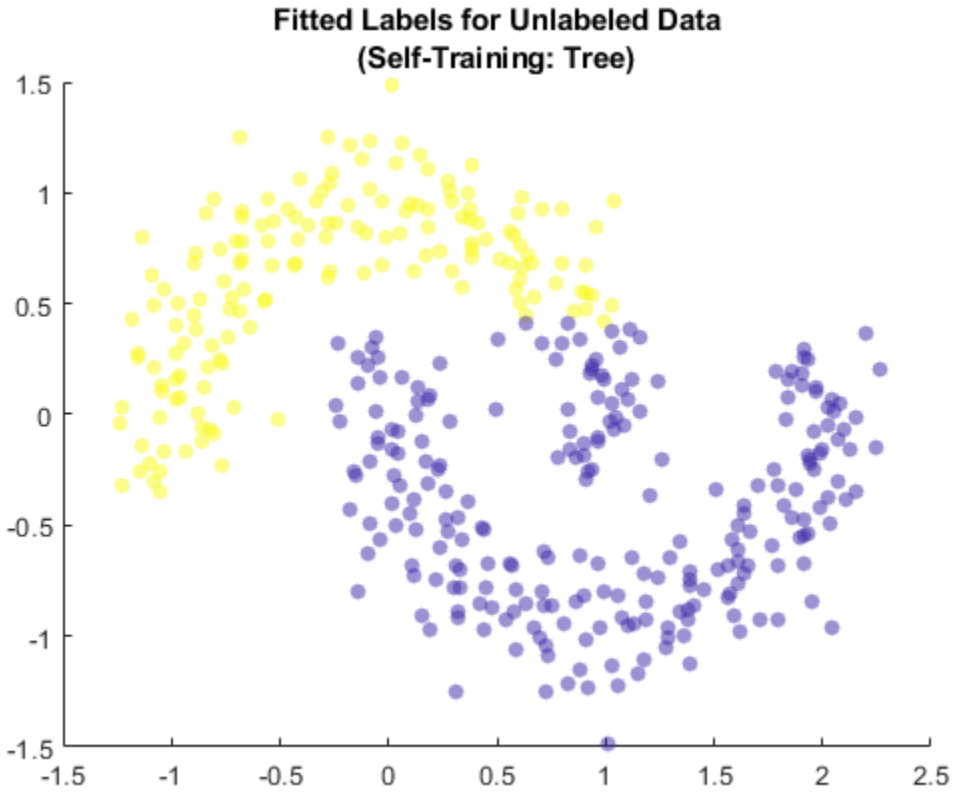
This method, with an SVM learner, also seems to label the newX points accurately. The two moons are visually distinct, and the points that are labeled with the most uncertainty lie on the boundary between the two shapes.

Some learners might not label the unlabeled data as effectively, however. For example, use a tree model instead of the default SVM model to label the data in newX.

```
selfTreeMdl = fitsemiself(X,label,newX,'Learner','tree');
```

Visualize the fitted label results.

```
maxTreeScores = max(selfTreeMdl.LabelScores,[],2);
rescaledTreeScores = rescale(maxTreeScores,0.05,0.95);
scatter(newX(:,1),newX(:,2),[],selfTreeMdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledTreeScores);
title(["Fitted Labels for Unlabeled Data","(Self-Training: Tree)"])
```



This method, with a tree learner, mislabels many of the points in the top moon. When you use a semi-supervised self-training method, make sure to use an underlying learner that is appropriate for the structure of your data.

This code creates the function twomoons.

```
function [X,label] = twomoons(n) % Generate two moons, with n points in each moon.

% Specify the radius and relevant angles for the two moons.
noise = (1/6).*randn(n,1);
```

```
radius = 1 + noise;
angle1 = pi + pi/10;
angle2 = pi/10;

% Create the bottom moon with a center at (1,0).
bottomTheta = linspace(-angle1,angle2,n)';
bottomX1 = radius.*cos(bottomTheta) + 1;
bottomX2 = radius.*sin(bottomTheta);

% Create the top moon with a center at (0,0).
topTheta = linspace(angle1,-angle2,n)';
topX1 = radius.*cos(topTheta);
topX2 = radius.*sin(topTheta);

% Return the moon points and their labels.
X = [bottomX1 bottomX2; topX1 topX2];
label = [ones(n,1); 2*ones(n,1)];
end
```

See Also

`fitsemigraph` | `fitsemiself`

Interpret Machine Learning Models

This topic introduces Statistics and Machine Learning Toolbox features for model interpretation and shows how to interpret a machine learning model (classification and regression).

A machine learning model is often referred to as a "black box" model because it can be difficult to understand how the model makes predictions. Interpretability tools help you overcome this aspect of machine learning algorithms and reveal how predictors contribute (or do not contribute) to predictions. Also, you can validate whether the model uses the correct evidence for its predictions, and find model biases that are not immediately apparent.

Features for Model Interpretation

Use `lime`, `shapley`, and `plotPartialDependence` to explain the contribution of individual predictors to the predictions of a trained classification or regression model.

- `lime` — Local interpretable model-agnostic explanations (LIME [1]) interpret a prediction for a query point by fitting a simple interpretable model for the query point. The simple model acts as an approximation for the trained model and explains model predictions around the query point. The simple model can be either a linear model or a decision tree model. You can use the estimated coefficients of a linear model or the estimated predictor importance of a decision tree model to explain the contribution of individual predictors to the prediction for the query point. For more details, see “LIME” on page 33-3505.
- `shapley` — The Shapley value [2][3] of a predictor for a query point explains the deviation of the prediction (response for regression or class scores for classification) for the query point from the average prediction, due to the predictor. For a query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average. For more details, see “Shapley Values for Machine Learning Model” on page 18-272.
- `plotPartialDependence` and `partialDependence` — A partial dependence plot (PDP [4]) shows the relationships between a predictor (or a pair of predictors) and the prediction (response for regression or class scores for classification) in the trained model. The partial dependence on the selected predictor is defined by the averaged prediction obtained by marginalizing out the effect of the other variables. Therefore, the partial dependence is a function of the selected predictor that shows the average effect of the selected predictor over the data set. You can also create a set of individual conditional expectation (ICE [5]) plots for each observation, showing the effect of the selected predictor on a single observation. For more details, see “More About” on page 33-4659 on the `plotPartialDependence` reference page.

Some machine learning models support embedded type feature selection, where the model learns predictor importance as part of the model learning process. You can use the estimated predictor importance to explain model predictions. For example:

- Train an ensemble (`ClassificationBaggedEnsemble` or `RegressionBaggedEnsemble`) of bagged decision trees (for example, random forest) and use the `predictorImportance` and `oobPermutedPredictorImportance` functions.
- Train a linear model with lasso regularization, which shrinks the coefficients of the least important predictors. Then use the estimated coefficients as measures for predictor importance. For example, use `fitclinear` or `fitrlinear` and specify the `'Regularization'` name-value argument as `'lasso'`.

For a list of machine learning models that support embedded type feature selection, see “Embedded Type Feature Selection” on page 15-54.

Use Statistics and Machine Learning Toolbox features for three levels of model interpretation: local, cohort, and global.

Level	Objective	Use Case	Statistics and Machine Learning Toolbox Feature
Local interpretation	Explain a prediction for a single query point.	• Identify important predictors for an individual prediction. • Examine a counterintuitive prediction.	Use `lime` and `shapley` for a specified query point.
Cohort interpretation	Explain how a trained model makes predictions for a subset of the entire data set.	Validate predictions for a particular group of samples.	• Use `lime` and `shapley` for multiple query points. After creating a `lime` or `shapley` object, you can call the object function `fit` multiple times to interpret predictions for other query points. • Pass a subset of data when you call `lime`, `shapley`, and `plotPartialDependence`. The features interpret the trained model using the specified subset instead of the entire training data set.
Global interpretation	Explain how a trained model makes predictions for the entire data set.	• Demonstrate how a trained model works. • Compare different models.	• Use `plotPartialDependence` to create PDPs and ICE plots for the predictors of interest. • Find important predictors from a trained model that supports “Embedded Type Feature Selection” on page 15-54.

Interpret Classification Model

This example trains an ensemble of bagged decision trees using the random forest algorithm, and interprets the trained model using interpretability features. Use the object functions (`oobPermutedPredictorImportance` and `predictorImportance`) of the trained model to find important predictors in the model. Also, use `lime` and `shapley` to interpret the predictions for specified query points. Then use `plotPartialDependence` to create a plot that shows the relationships between an important predictor and predicted classification scores.

Train Classification Ensemble Model

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    ______

    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

Create a table of predictor variables by removing the columns containing customer IDs and ratings from `tbl`.

```
tblX = removevars(tbl,["ID","Rating"]);
```

Train an ensemble of bagged decision trees by using the `fitcensemble` function and specifying the ensemble aggregation method as random forest (`'Bag'`). For reproducibility of the random forest algorithm, specify the `'Reproducible'` name-value argument as `true` for tree learners. Also, specify the class names to set the order of the classes in the trained model.

```
rng('default') % For reproducibility
t = templateTree('Reproducible',true);
blackbox = fitcensemble(tblX,tbl.Rating, ...
    'Method','Bag','Learners',t, ...
    'CategoricalPredictors','Industry', ...
    'ClassNames',{'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'});
```

`blackbox` is a `ClassificationBaggedEnsemble` model.

Use Model-Specific Interpretability Features

`ClassificationBaggedEnsemble` supports two object functions, `oobPermutedPredictorImportance` and `predictorImportance`, which find important predictors in the trained model.

Estimate out-of-bag predictor importance by using the `oobPermutedPredictorImportance` function. The function randomly permutes out-of-bag data across one predictor at a time, and estimates the increase in the out-of-bag error due to this permutation. The larger the increase, the more important the feature.

```
Imp1 = oobPermutedPredictorImportance(blackbox);
```

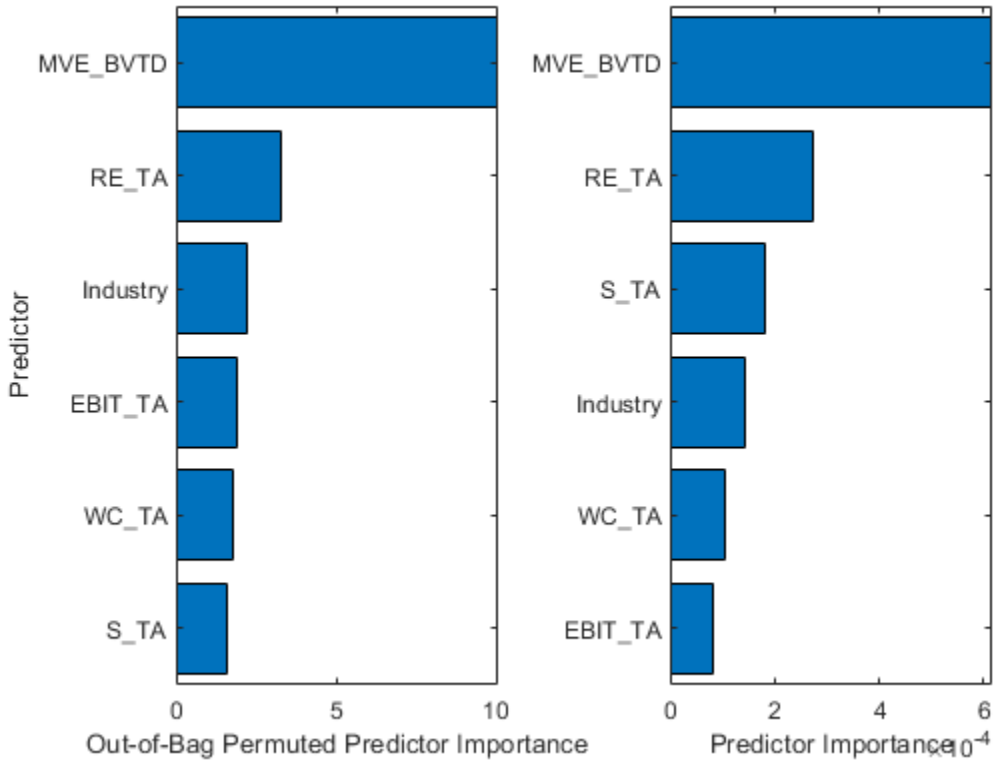
Estimate predictor importance by using the `predictorImportance` function. The function estimates predictor importance by summing changes in the node risk due to splits on each predictor and dividing the sum by the number of branch nodes.

```
Imp2 = predictorImportance(blackbox);
```

Create a table containing the predictor importance estimates, and use the table to create horizontal bar graphs. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
table_Imp = table(Imp1',Imp2', ...
    'VariableNames',{'Out-of-Bag Permuted Predictor Importance','Predictor Importance'}, ...
    'RowNames',blackbox.PredictorNames);
tiledlayout(1,2)
```

```
ax1 = nexttile;
table_Imp1 = sortrows(table_Imp,'Out-of-Bag Permuted Predictor Importance');
barh(categorical(table_Imp1.Row,table_Imp1.Row),table_Imp1.('Out-of-Bag Permuted Predictor Impor
xlabel('Out-of-Bag Permuted Predictor Importance')
ylabel('Predictor')
ax2 = nexttile;
table_Imp2 = sortrows(table_Imp,'Predictor Importance');
barh(categorical(table_Imp2.Row,table_Imp2.Row),table_Imp2.('Predictor Importance'))
xlabel('Predictor Importance')
ax1.TickLabelInterpreter = 'none';
ax2.TickLabelInterpreter = 'none';
```



Both object functions identify `MVE_BVTD` and `RE_TA` as the two most important predictors.

Specify Query Point

Find the observations whose `Rating` is `'AAA'` and choose four query points among them.

```
tblX_AAA = tblX(strcmp(tbl.Rating,'AAA'),:);
queryPoint = datasample(tblX_AAA,4,'Replace',false)
```

```
queryPoint=4×6 table
    WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry
    _____    _____    _______    ________    _____    ________

    0.331    0.531     0.077      7.116      0.522       12
     0.26    0.515     0.065      3.394      0.515        1
    0.121    0.413     0.057      3.647      0.466       12
```

```
    0.617    0.766     0.126      4.442      0.483        9
```

Use LIME with Linear Simple Models

Explain the predictions for the query points using `lime` with linear simple models. `lime` generates a synthetic data set and fits a simple model to the synthetic data set.

Create a `lime` object using `tblX_AAA` so that `lime` generates a synthetic data set using only the observations whose `Rating` is `'AAA'`, not the entire data set.

```
explainer_lime = lime(blackbox,tblX_AAA);
```

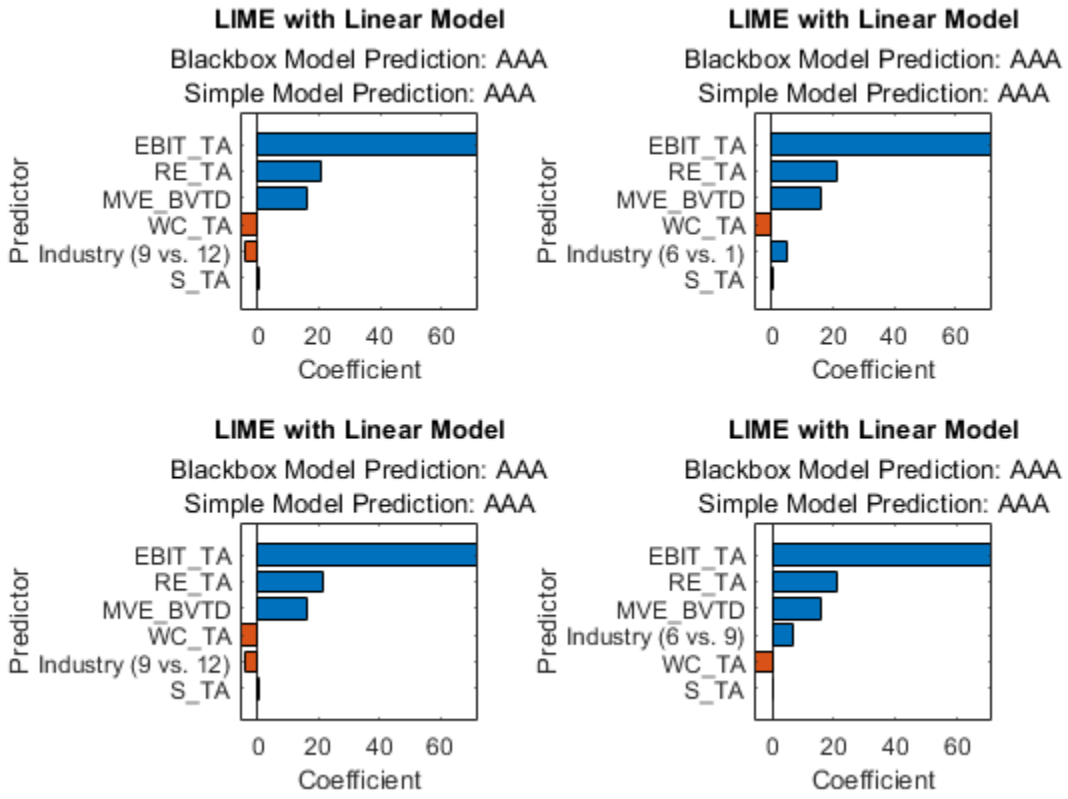
The default value of “'DataLocality'” on page 33-0 for `lime` is `'global'`, which implies that, by default, `lime` generates a global synthetic data set and uses it for any query points. `lime` uses different observation weights so that weight values are more focused on the observations near the query point. Therefore, you can interpret each simple model as an approximation of the trained model for a specific query point.

Fit simple models for the four query points by using the object function `fit`. Specify the third input (the number of important predictors to use in the simple model) as 6 to use all six predictors.

```
explainer_lime1 = fit(explainer_lime,queryPoint(1,:),6);
explainer_lime2 = fit(explainer_lime,queryPoint(2,:),6);
explainer_lime3 = fit(explainer_lime,queryPoint(3,:),6);
explainer_lime4 = fit(explainer_lime,queryPoint(4,:),6);
```

Plot the coefficients of the simple models by using the object function `plot`.

```
tiledlayout(2,2)
ax1 = nexttile; plot(explainer_lime1);
ax2 = nexttile; plot(explainer_lime2);
ax3 = nexttile; plot(explainer_lime3);
ax4 = nexttile; plot(explainer_lime4);
ax1.TickLabelInterpreter = 'none';
ax2.TickLabelInterpreter = 'none';
ax3.TickLabelInterpreter = 'none';
ax4.TickLabelInterpreter = 'none';
```



All simple models identify `EBIT_TA`, `RE_TA`, and `MVE_BVTD` as the three most important predictors. The positive coefficients for the predictors suggest that increasing the predictor values leads to an increase in the predicted scores in the simple models.

For a categorical predictor, the `plot` function displays only the most important dummy variable of the categorical predictor. Therefore, each bar graph displays a different dummy variable.

Compute Shapley Values

The Shapley value of a predictor for a query point explains the deviation of the predicted score for the query point from the average score, due to the predictor. Create a `shapley` object using `tblX_AAA` so that `shapley` computes the expected contribution based on the samples for `'AAA'`.

```
explainer_shapley = shapley(blackbox,tblX_AAA);
```

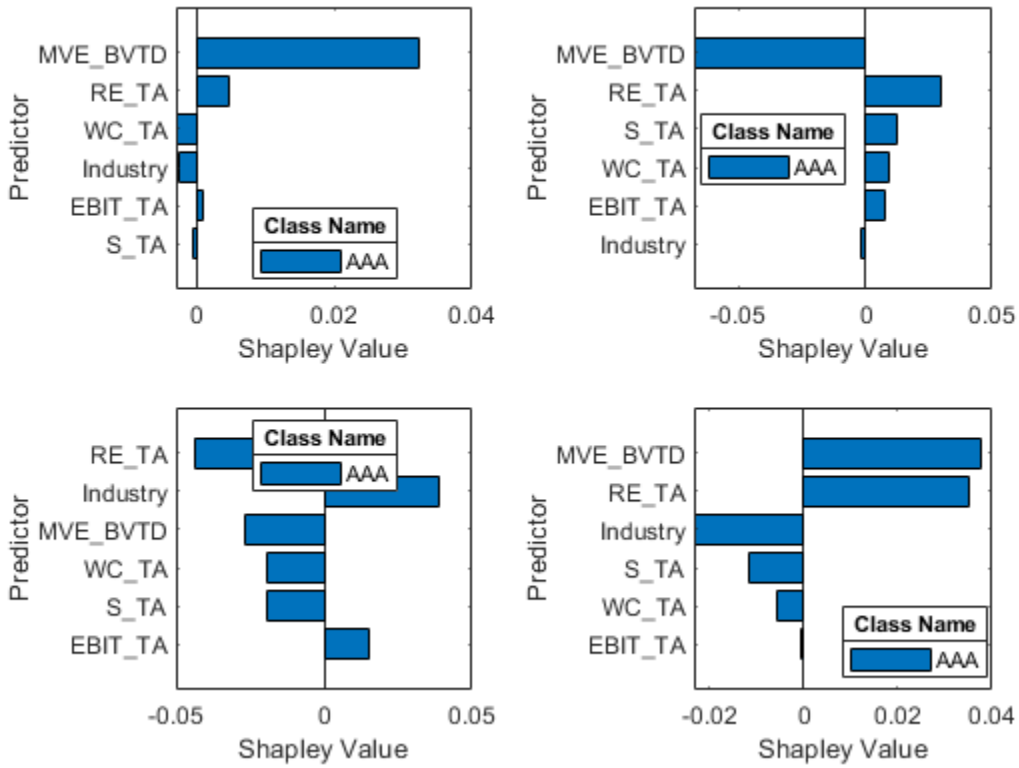
Compute the Shapley values for the query points by using the object function `fit`.

```
explainer_shapley1 = fit(explainer_shapley,queryPoint(1,:));
explainer_shapley2 = fit(explainer_shapley,queryPoint(2,:));
explainer_shapley3 = fit(explainer_shapley,queryPoint(3,:));
explainer_shapley4 = fit(explainer_shapley,queryPoint(4,:));
```

Plot the Shapley values by using the object function `plot`.

```
tiledlayout(2,2)
ax1 = nexttile; plot(explainer_shapley1)
ax2 = nexttile; plot(explainer_shapley2)
```

```
ax3 = nexttile; plot(explainer_shapley3)
ax4 = nexttile; plot(explainer_shapley4)
ax1.TickLabelInterpreter = 'none';
ax2.TickLabelInterpreter = 'none';
ax3.TickLabelInterpreter = 'none';
ax4.TickLabelInterpreter = 'none';
```



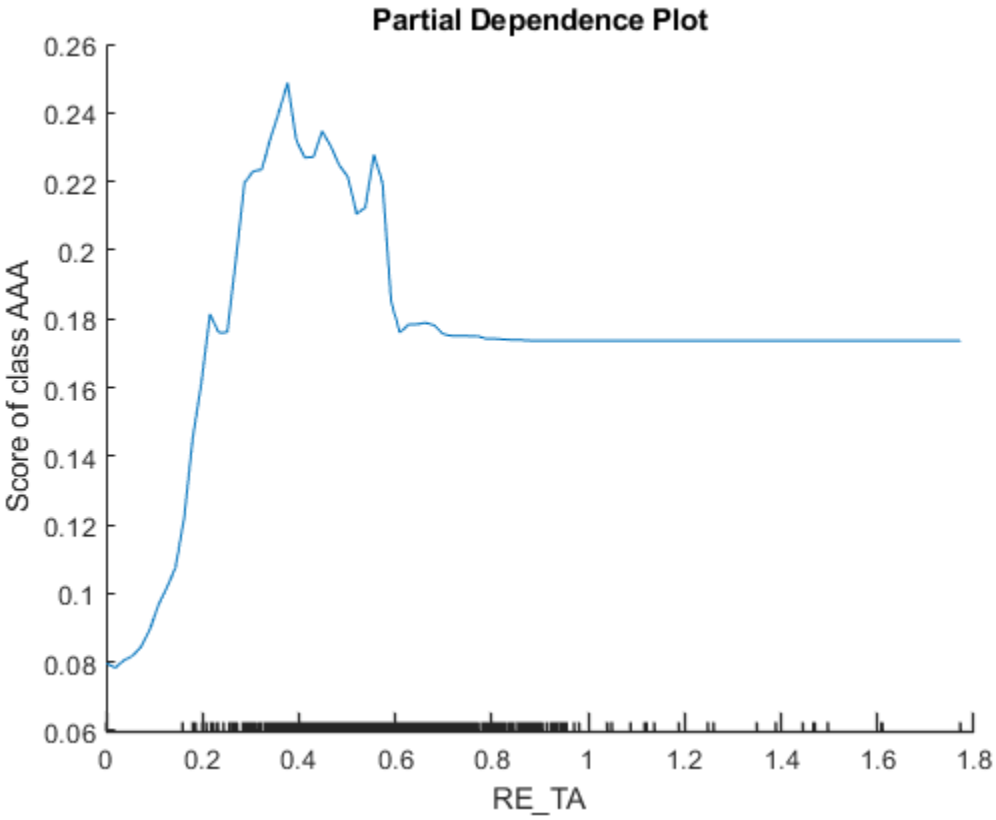
`MVE_BVTD` and `RE_TA` are two of the three most important predictors for all four query points.

The Shapley values of `MVE_BVTD` are positive for the first and fourth query points, and negative for the second and third query points. The `MVE_BVTD` values are about 7 and 4 for the first and fourth query points, respectively, and the value for both the second and third query points is about 3.5. According to the Shapley values for the four query points, a large `MVE_BVTD` value leads to an increase in the predicted score, and a small `MVE_BVTD` value leads to a decrease in the predicted scores compared to the average. The results are consistent with the results from `lime`.

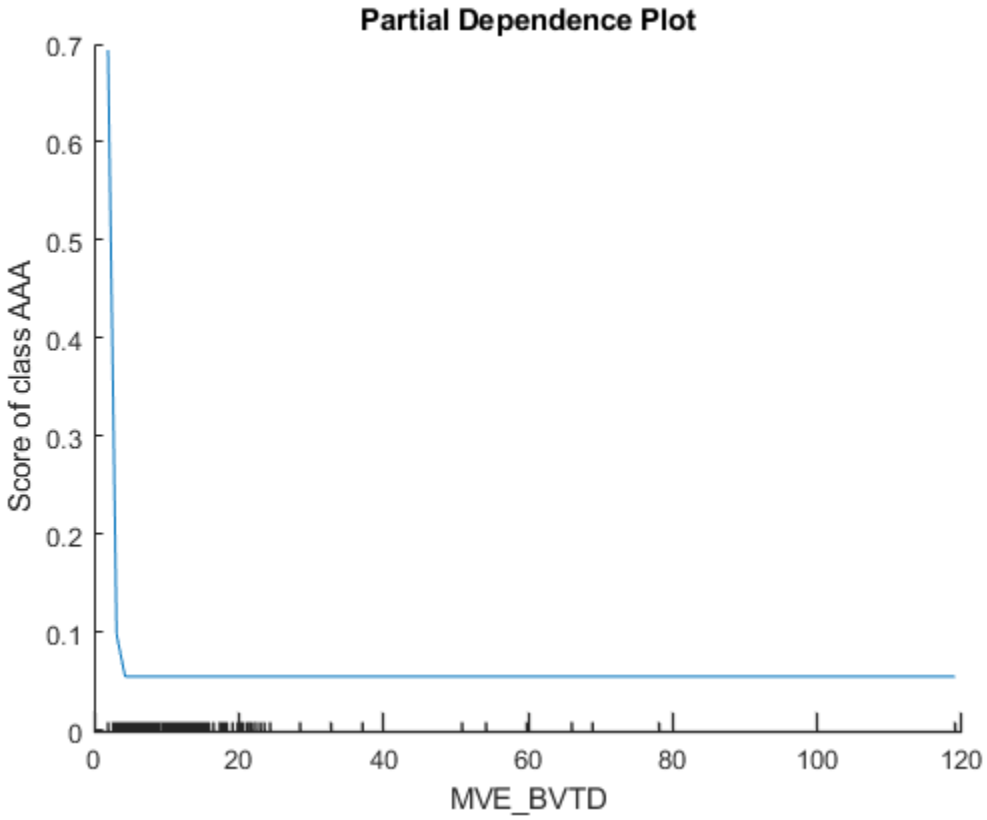
Create Partial Dependence Plot (PDP)

A PDP plot shows the averaged relationships between the predictor and the predicted score in the trained model. Create PDPs for `RE_TA` and `MVE_BVTD`, which the other interpretability tools identify as important predictors. Pass `tblx_AAA` to `plotPartialDependence` so that the function computes the expectation of the predicted scores using only the samples for `'AAA'`.

```
figure
plotPartialDependence(blackbox,'RE_TA','AAA',tblX_AAA)
```




```
plotPartialDependence(blackbox,'MVE_BVTD','AAA',tblX_AAA)
```



The minor ticks in the *x*-axis represent the unique values of the predictor in `tbl_AAA`. The plot for `MVE_BVTD` shows that the predicted score is large when the `MVE_BVTD` value is small. The score value decreases as the `MVE_BVTD` value increases until it reaches about 5, and then the score value stays unchanged as the `MVE_BVTD` value increases. The dependency on `MVE_BVTD` in the subset `tbl_AAA` identified by `plotPartialDependence` is not consistent with the local contributions of `MVE_BVTD` at the four query points identified by `lime` and `shapley`.

Interpret Regression Model

The model interpretation workflow for a regression problem is similar to the workflow for a classification problem, as demonstrated in the example “Interpret Classification Model” on page 18-257.

This example trains a Gaussian process regression (GPR) model and interprets the trained model using interpretability features. Use a kernel parameter of the GPR model to estimate predictor weights. Also, use `lime` and `shapley` to interpret the predictions for specified query points. Then use `plotPartialDependence` to create a plot that shows the relationships between an important predictor and predicted responses.

Train GPR Model

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight);
```

Train a GPR model of the response variable `MPG` by using the `fitrgp` function. Specify `'KernelFunction'` as `'ardsquaredexponential'` to use the squared exponential kernel with a separate length scale per predictor.

```
blackbox = fitrgp(tbl,MPG,'ResponseName','MPG','CategoricalPredictors',[2 5], ...
    'KernelFunction','ardsquaredexponential');
```

`blackbox` is a `RegressionGP` model.

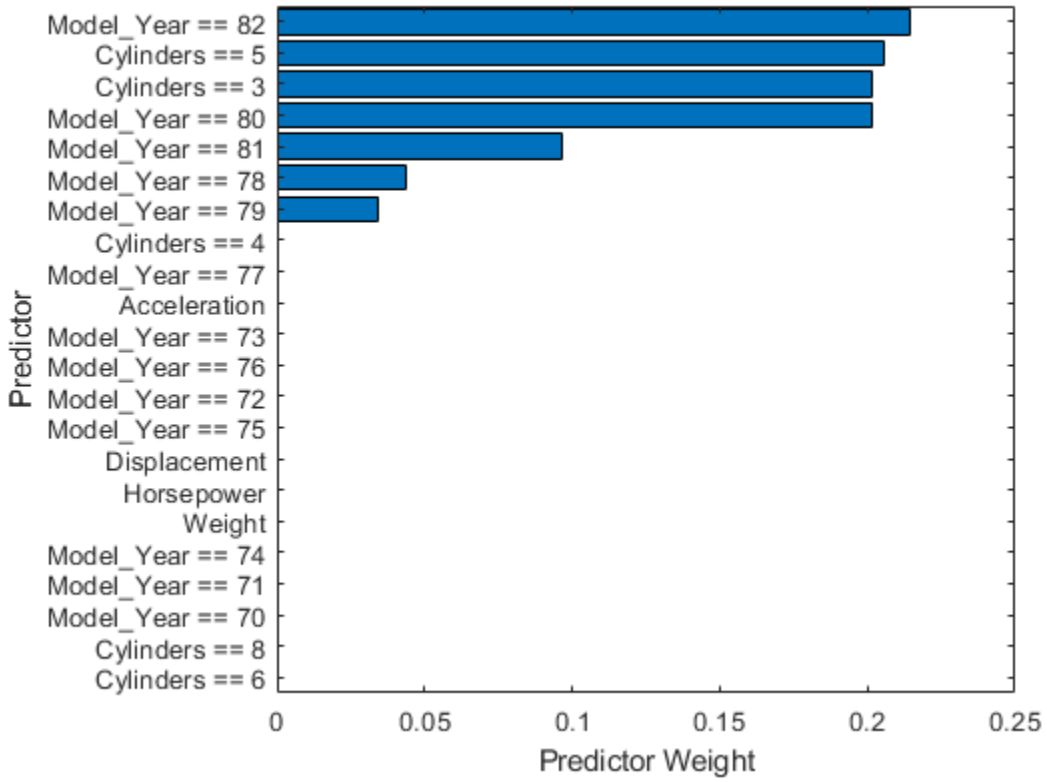
Use Model-Specific Interpretability Features

You can compute predictor weights (predictor importance) from the learned length scales of the kernel function used in the model. The length scales define how far apart a predictor can be for the response values to become uncorrelated. Find the normalized predictor weights by taking the exponential of the negative learned length scales.

```
sigmaL = blackbox.KernelInformation.KernelParameters(1:end-1); % Learned length scales
weights = exp(-sigmaL); % Predictor weights
weights = weights/sum(weights); % Normalized predictor weights
```

Create a table containing the normalized predictor weights, and use the table to create horizontal bar graphs. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
tbl_weight = table(weights,'VariableNames',{'Predictor Weight'}, ...
    'RowNames',blackbox.ExpandedPredictorNames);
tbl_weight = sortrows(tbl_weight,'Predictor Weight');
b = barh(categorical(tbl_weight.Row,tbl_weight.Row),tbl_weight.('Predictor Weight'));
b.Parent.TickLabelInterpreter = 'none';
xlabel('Predictor Weight')
ylabel('Predictor')
```



The predictor weights indicate that multiple dummy variables for the categorical predictors `Model_Year` and `Cylinders` are important.

Specify Query Point

Find the observations whose `MPG` values are smaller than the 0.25 quantile of `MPG`. From the subset, choose four query points that do not include missing values.

```
rng('default') % For reproducibility
idx_subset = find(MPG < quantile(MPG,0.25));
tbl_subset = tbl(idx_subset,:);
queryPoint = datasample(rmmissing(tbl_subset),4,'Replace',false)
```

```
queryPoint=4×6 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight
    ____________    _________    ____________    __________    __________    ______

        13.2            8            318            150           76          3940
        14.9            8            302            130           77          4295
          14            8            360            215           70          4615
        13.7            8            318            145           77          4140
```

Use LIME with Tree Simple Models

Explain the predictions for the query points using `lime` with decision tree simple models. `lime` generates a synthetic data set and fits a simple model to the synthetic data set.

Create a `lime` object using `tbl_subset` so that `lime` generates a synthetic data set using the subset instead of the entire data set. Specify `'SimpleModelType'` as `'tree'` to use a decision tree simple model.

```
explainer_lime = lime(blackbox,tbl_subset,'SimpleModelType','tree');
```

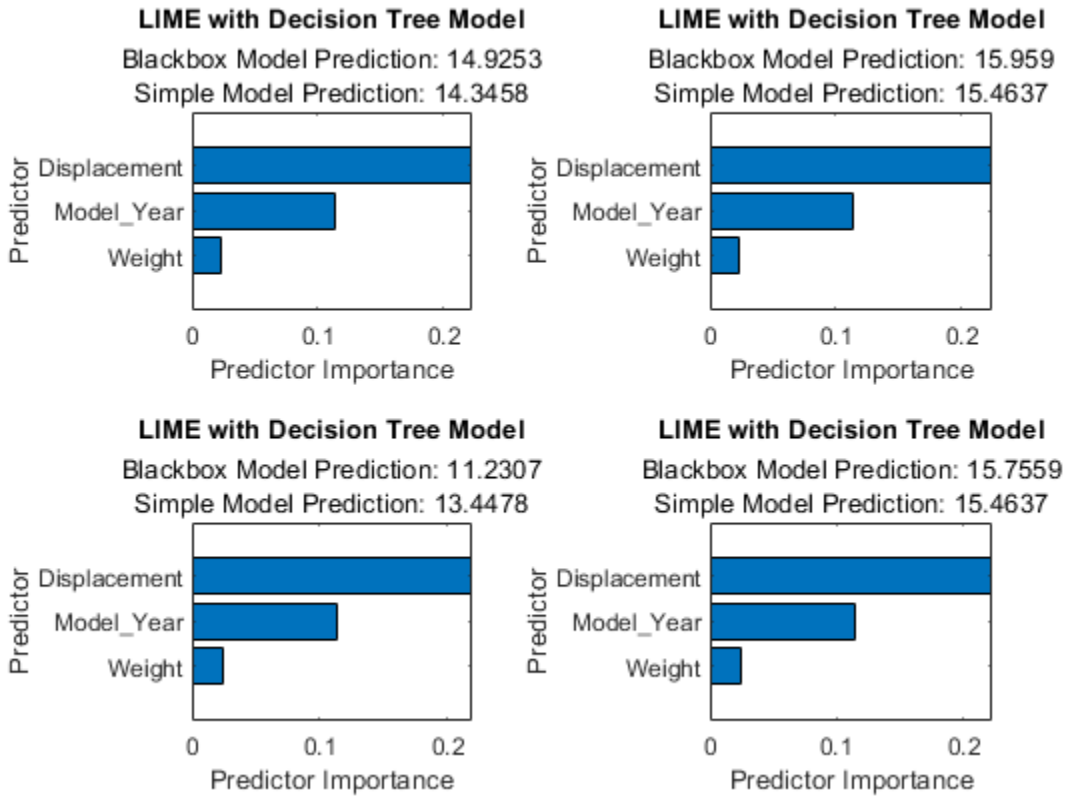
The default value of "'DataLocality'" on page 33-0 for `lime` is `'global'`, which implies that, by default, `lime` generates a global synthetic data set and uses it for any query points. `lime` uses different observation weights so that weight values are more focused on the observations near the query point. Therefore, you can interpret each simple model as an approximation of the trained model for a specific query point.

Fit simple models for the four query points by using the object function `fit`. Specify the third input (the number of important predictors to use in the simple model) as 6. With this setting, the software specifies the maximum number of decision splits (or branch nodes) as 6 so that the fitted decision tree uses at most all predictors.

```
explainer_lime1 = fit(explainer_lime,queryPoint(1,:),6);
explainer_lime2 = fit(explainer_lime,queryPoint(2,:),6);
explainer_lime3 = fit(explainer_lime,queryPoint(3,:),6);
explainer_lime4 = fit(explainer_lime,queryPoint(4,:),6);
```

Plot the predictor importance by using the object function `plot`.

```
tiledlayout(2,2)
ax1 = nexttile; plot(explainer_lime1);
ax2 = nexttile; plot(explainer_lime2);
ax3 = nexttile; plot(explainer_lime3);
ax4 = nexttile; plot(explainer_lime4);
ax1.TickLabelInterpreter = 'none';
ax2.TickLabelInterpreter = 'none';
ax3.TickLabelInterpreter = 'none';
ax4.TickLabelInterpreter = 'none';
```



All simple models identify `Displacement`, `Model_Year`, and `Weight` as important predictors.

Compute Shapley Values

The Shapley value of a predictor for a query point explains the deviation of the predicted response for the query point from the average response, due to the predictor. Create a `shapley` object for the model `blackbox` using `tbl_subset` so that `shapley` computes the expected contribution based on the observations in `tbl_subset`.

```
explainer_shapley = shapley(blackbox,tbl_subset);
```

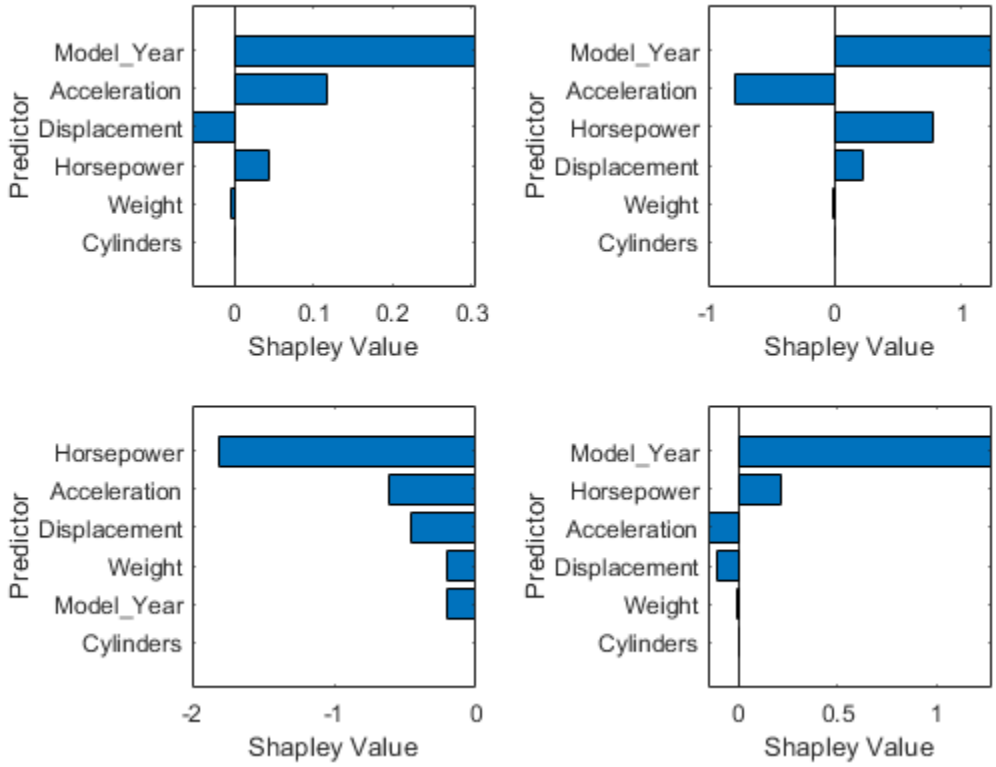
Compute the Shapley values for the query points by using the object function `fit`.

```
explainer_shapley1 = fit(explainer_shapley,queryPoint(1,:));
explainer_shapley2 = fit(explainer_shapley,queryPoint(2,:));
explainer_shapley3 = fit(explainer_shapley,queryPoint(3,:));
explainer_shapley4 = fit(explainer_shapley,queryPoint(4,:));
```

Plot the Shapley values by using the object function `plot`.

```
tiledlayout(2,2)
ax1 = nexttile; plot(explainer_shapley1)
ax2 = nexttile; plot(explainer_shapley2)
ax3 = nexttile; plot(explainer_shapley3)
ax4 = nexttile; plot(explainer_shapley4)
ax1.TickLabelInterpreter = 'none';
ax2.TickLabelInterpreter = 'none';
```

```
ax3.TickLabelInterpreter = 'none';
ax4.TickLabelInterpreter = 'none';
```

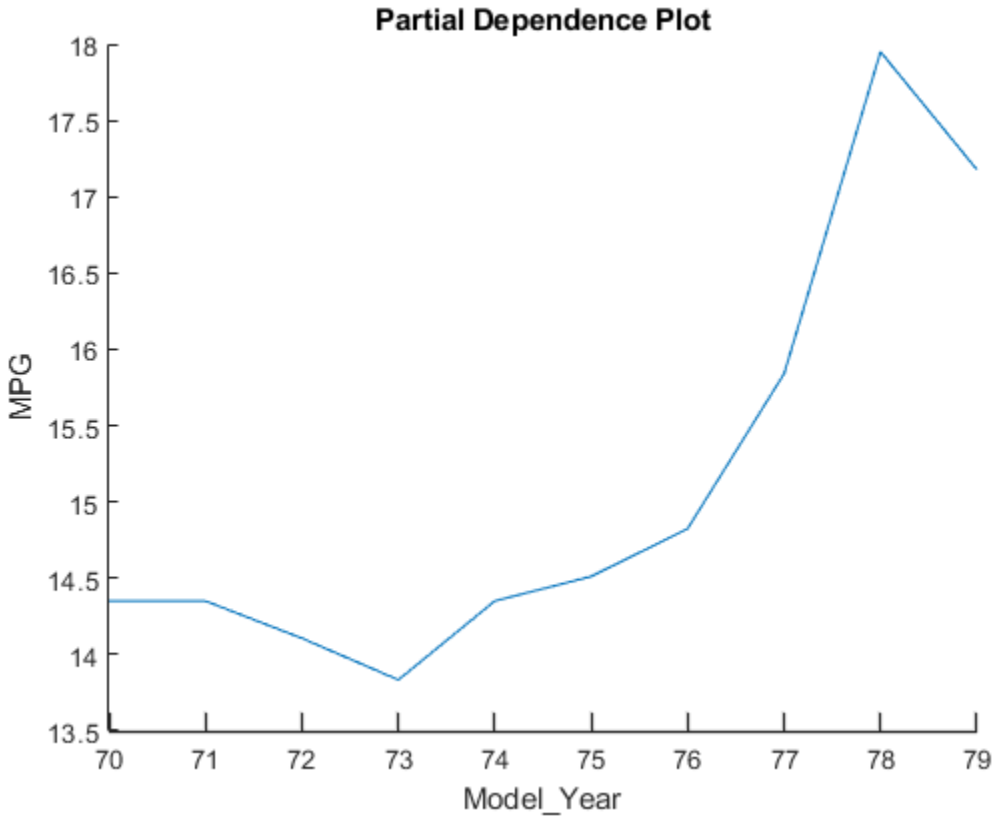


`Model_Year` is the most important predictor for the first, second, and fourth query points, and the Shapley values of `Model_Year` are positive for the three query points. The `Model_Year` value is 76 or 77 for these three points, and the value for the third query point is 70. According to the Shapley values for the four query points, a small `Model_Year` value leads to a decrease in the predicted response, and a large `Model_Year` value leads to an increase in the predicted response compared to the average.

Create Partial Dependence Plot (PDP)

A PDP plot shows the averaged relationships between the predictor and the predicted response in the trained model. Create a PDP for `Model_Year`, which the other interpretability tools identify as an important predictor. Pass `tbl_subset` to `plotPartialDependence` so that the function computes the expectation of the predicted responses using only the samples in `tbl_subset`.

```
figure
plotPartialDependence(blackbox,'Model_Year',tbl_subset)
```



The plot shows the same trend identified by the Shapley values for the four query points. The predicted response (`MPG`) value increases as the `Model_Year` value increases.

References

[1] Ribeiro, Marco Tulio, S. Singh, and C. Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. San Francisco, California: ACM, 2016.

[2] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[3] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

[4] Friedman, Jerome. H. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29, no. 5 (2001): 1189-1232.

[5] Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. “Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation.” *Journal of Computational and Graphical Statistics* 24, no. 1 (January 2, 2015): 44–65.

See Also

`lime` | `plotPartialDependence` | `shapley`

Related Examples

- “Shapley Values for Machine Learning Model” on page 18-272
- “Introduction to Feature Selection” on page 15-49
- “Interpret Deep Network Predictions on Tabular Data Using LIME” (Deep Learning Toolbox)

Shapley Values for Machine Learning Model

You can compute Shapley values for a machine learning model by using a `shapley` object. Use the values to interpret the contributions of individual features in the model to the prediction for a query point. There are two ways to compute Shapley values:

- Create a `shapley` object for a machine learning model with a specified query point by using the `shapley` function. The function computes the Shapley values of all features in the model for the query point.
- Create a `shapley` object for a machine learning model by using the `shapley` function and, then compute the Shapley values for a specified query point by using the `fit` function.

This topic defines Shapley values, describes two available algorithms for computing Shapley values, provides examples for each, and shows how to reduce the cost of computing Shapley values.

What Is a Shapley Value?

In game theory, the Shapley value of a player is the average marginal contribution of the player in a cooperative game. That is, Shapley values are fair allocations, to individual players, of the total gain generated from a cooperative game. In the context of machine learning prediction, the Shapley value of a feature for a query point explains the contribution of the feature to a prediction (response for regression or score of each class for classification) at the specified query point. The Shapley value corresponds to the deviation of the prediction for the query point from the average prediction, due to the feature. For each query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average.

The Shapley value of the ith feature for the query point x is defined by the value function v:

$$\varphi_i(v_x) = \frac{1}{M} \sum_{S \subseteq \mathcal{M} \setminus \{i\}} \frac{v_x(S \cup \{i\}) - v_x(S)}{\frac{(M-1)!}{|S|!(M-|S|-1)!}} \tag{18-2}$$

- M is the number of all features.
- $\mathcal{M}$ is the set of all features.
- $|S|$ is the cardinality of the set S, or the number of elements in the set S.
- $v_x(S)$ is the value function of the features in a set S for the query point x. The value of the function indicates the expected contribution of the features in S to the prediction for the query point x.

Shapley Value Computation Algorithms

`shapley` offers two algorithms: kernelSHAP [1], which uses interventional distributions for the value function, and the extension to kernelSHAP [2], which uses conditional distributions for the value function. You can specify the algorithm to use by setting the `'Method'` name-value argument of the `shapley` function or the `fit` function.

The difference between the two algorithms is the definition of the value function. Both algorithms define the value function such that the sum of the Shapley values of a query point over all features corresponds to the total deviation of the prediction for the query point from the average.

$$\sum_{i=1}^{M} \varphi_i(v_x) = f(x) - E[f(x)] .$$

Therefore, the value function $v_x(S)$ must correspond to the expected contribution of the features in S to the prediction (f) for the query point x. The two algorithms compute the expected contribution by using artificial samples created from the specified data (X). You must provide X through the machine learning model input or a separate data input argument when you create a `shapley` object. In the artificial samples, the values for the features in S come from the query point. For the rest of the features (features in S^c, the complement of S), the kernelSHAP algorithm generates samples using interventional distributions, whereas the extension to the kernelSHAP algorithm generates samples using conditional distributions.

KernelSHAP ('Method','interventional-kernel')

`shapley` uses the kernelSHAP algorithm by default.

The kernelSHAP algorithm defines the value function of the features in S at the query point x as the expected prediction with respect to the interventional distribution D, which is the joint distribution of the features in S^c:

$$v_x(S) = E_D[f(x_S, X_{S^c})],$$

where x_S is the query point value for the features in S, and X_{S^c} are the features in S^c.

To evaluate the value function $v_x(S)$ at the query point x, with the assumption that the features are not highly correlated, `shapley` uses the values in the data X as samples of the interventional distribution D for the features in S^c:

$$v_x(S) = E_D[f(x_S, X_{S^c})] \approx \frac{1}{N}\sum_{j=1}^{N} f\left(x_S, (X_{S^c})_j\right),$$

where N is the number of observations, and $(X_{S^c})_j$ contains the values of the features in S^c for the jth observation.

For example, suppose you have three features in X and four observations: (x_{11},x_{12},x_{13}), (x_{21},x_{22},x_{23}), (x_{31},x_{32},x_{33}), and (x_{41},x_{42},x_{43}). Assume that S includes the first feature, and S^c includes the rest. In this case, the value function of the first feature evaluated at the query point (x_{41},x_{42},x_{43}) is

$$v_x(S) = \frac{1}{4}[f(x_{41}, x_{12}, x_{13}) + f(x_{41}, x_{22}, x_{23}) + f(x_{41}, x_{32}, x_{33}) + f(x_{41}, x_{42}, x_{43})].$$

The kernelSHAP algorithm is computationally less expensive than the extension to the kernelSHAP algorithm, supports ordered categorical predictors, and can handle missing values in X. However, the algorithm requires the feature independence assumption and uses out-of-distribution samples [3]. The artificial samples created with a mix of the query point and the data X can contain unrealistic observations. For example, (x_{41},x_{12},x_{13}) might be a sample that does not occur in the full joint distribution of the three features.

Extension to KernelSHAP ('Method','conditional-kernel')

Specify `'Method','conditional-kernel'` to use the extension to the kernelSHAP algorithm.

the extension to the kernelSHAP algorithm defines the value function of the features in S at the query point x using the conditional distribution of X_{S^c}, given that X_S has the query point values:

$$v_x(S) = E_{X_{S^c}|X_S = x_S}[f(x_S, X_{S^c})].$$

To evaluate the value function $v_x(S)$ at the query point x, `shapley` uses nearest neighbors of the query point, which correspond to 10% of the observations in the data X. This approach uses more realistic samples than the kernelSHAP algorithm and does not require the feature independence assumption. However, this algorithm is computationally more expensive, does not support ordered categorical predictors, and cannot handle `NaN`s in continuous features. Also, the algorithm might assign a nonzero Shapley value to a dummy feature that does not contribute to the prediction, if the dummy feature is correlated with an important feature [3].

Specify Shapley Value Computation Algorithm

This example trains a linear classification model and computes Shapley values using both the kernelSHAP algorithm (`'Method','interventional-kernel'`) and the extension to the kernelSHAP algorithm (`'Method','conditional-kernel'`).

Train Linear Classification Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a linear classification model. Specify the objective function minimization technique (`'Solver'` name-value argument) as the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (`'lbfgs'`) for better accuracy of linear coefficients.

```
Mdl = fitclinear(X,Y,'Solver','lbfgs')
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: {'b'  'g'}
    ScoreTransform: 'none'
              Beta: [34x1 double]
              Bias: -3.7100
            Lambda: 0.0028
           Learner: 'svm'


  Properties, Methods
```

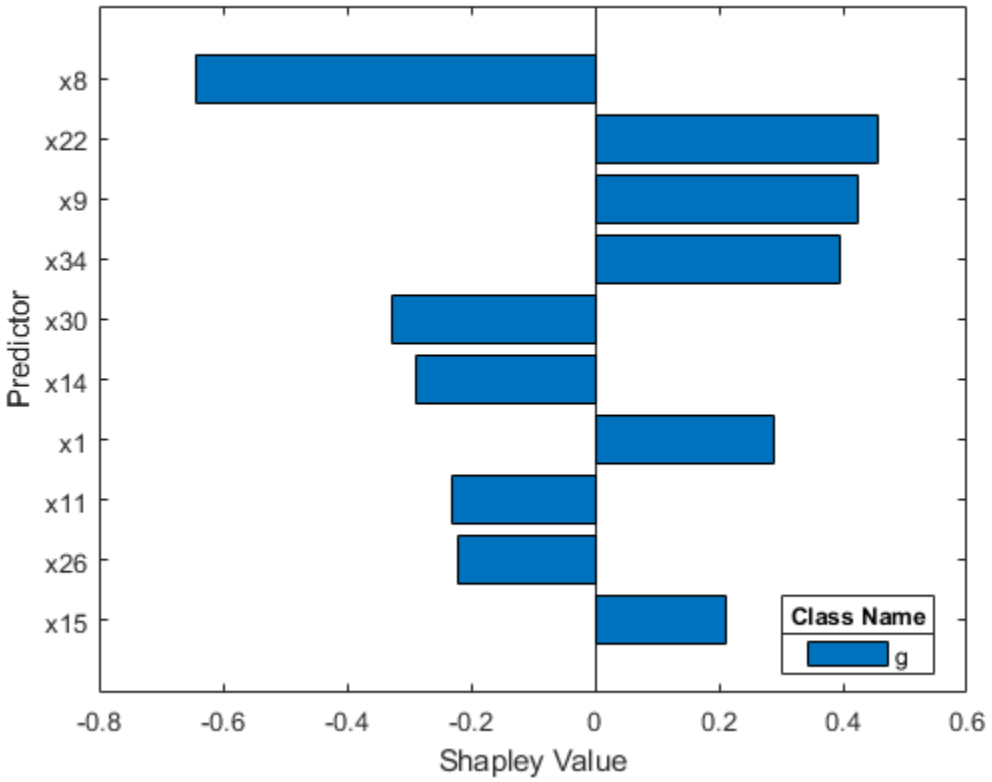
Shapley Values with Interventional Distribution

Compute the Shapley values for the first observation using the kernelSHAP algorithm, which uses the interventional distribution for the value function evaluation. You do not have to specify the `'Method'` value because `'interventional-kernel'` is the default.

```
queryPoint = X(1,:);
explainer1 = shapley(Mdl,X,'QueryPoint',queryPoint);
```

For a classification model, `shapley` computes Shapley values using the predicted class score for each class. Plot the Shapley values for the predicted class by using the `plot` function.

```
plot(explainer1)
```



The horizontal bar graph shows the Shapley values for the 10 most important variables, sorted by their absolute values. Each Shapley value explains the deviation of the score for the query point from the average score of the predicted class, due to the corresponding variable.

For a linear model where you assume features are independent from one another, you can compute the interventional Shapley values for the positive class (or the second class in `Mdl.ClassNames`, `'g'`) from the estimated coefficients (`Mdl.Beta`) [1].

```
linearSHAPValues = (Mdl.Beta'.*(queryPoint-mean(X)))';
```

Create a table containing the Shapley values computed from the kernelSHAP algorithm and the values from the coefficients.

```
t = table(explainer1.ShapleyValues.Predictor,explainer1.ShapleyValues.g,linearSHAPValues, ...
    'VariableNames',{'Predictor','KernelSHAP Value','LinearSHAP Value'})
```

```
t=34×3 table
    Predictor    KernelSHAP Value    LinearSHAP Value
    _________    ________________    ________________

      "x1"             0.28789             0.28789
      "x2"           -2.6619e-15                   0
      "x3"             0.20822             0.20822
      "x4"            -0.01998            -0.01998
      "x5"             0.20872             0.20872
      "x6"           -0.076991           -0.076991
      "x7"             0.19188             0.19188
```

```
"x8"            -0.64386          -0.64386
"x9"             0.42348           0.42348
"x10"           -0.030049         -0.030049
"x11"           -0.23132          -0.23132
"x12"             0.1422            0.1422
"x13"           -0.045973         -0.045973
"x14"           -0.29022          -0.29022
"x15"            0.21051           0.21051
"x16"            0.13382           0.13382
⋮
```

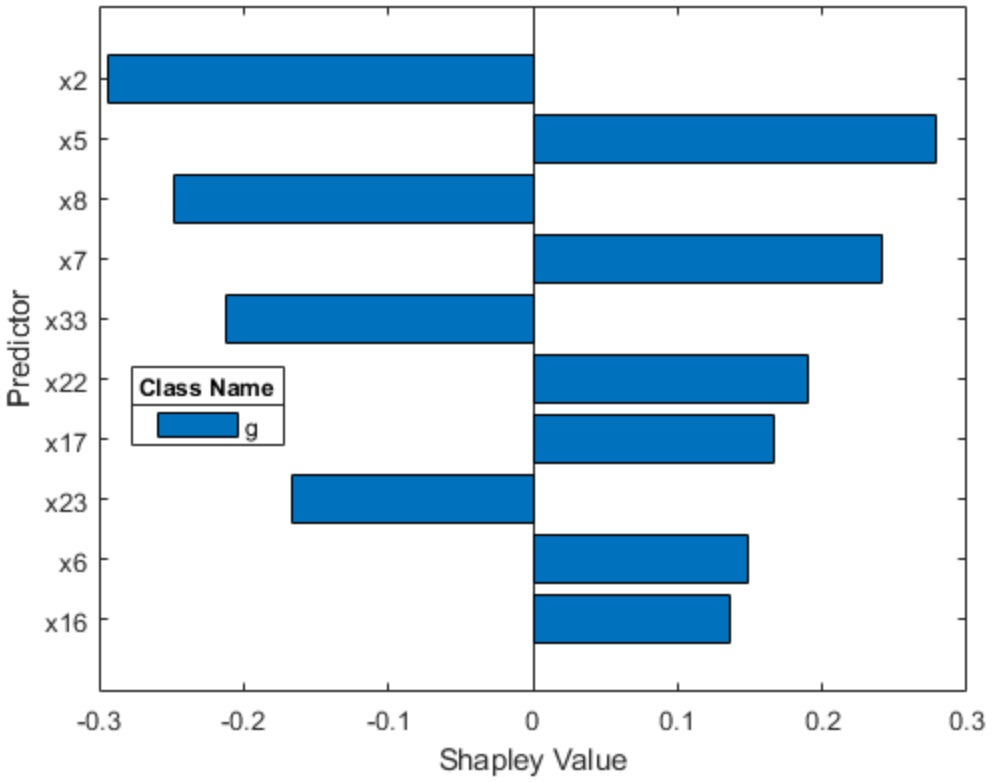
Shapley Values with Conditional Distribution

Compute the Shapley values for the first observation using the extension to the kernelSHAP algorithm, which uses the conditional distribution for the value function evaluation.

```
explainer2 = shapley(Mdl,X,'QueryPoint',queryPoint,'Method','conditional-kernel');
```

Plot the Shapley values.

```
plot(explainer2)
```



The two algorithms identify different sets for the 10 most important variables. Only the two variables `x8` and `x22` are common to both sets.

Complexity of Computing Shapley Values

The computational cost for Shapley values increases if the number of observations or features is large.

Large Number of Observations

Computing the value function (v) can be computationally expensive if you have a large number of observations, for example, more than 1000. For faster computation, use a smaller sample of the observations when you create a `shapley` object, or compute Shapley values in parallel by specifying `'UseParallel'` as `true` when you compute the values using the `shapley` or `fit` function. Computing in parallel requires Parallel Computing Toolbox.

Large Number of Features

Computing the summand in "Equation 18-2" for all available subsets S can be computationally expensive when M (the number of features) is large. The total number of subsets to consider is 2^M. Instead of computing the summand for all subsets, you can specify the maximum number of subsets for Shapley value computation by using the `'MaxNumSubsets'` name-value argument. `shapley` chooses subsets to use based on their weight values. The weight of a subset is proportional to 1/(denominator of the summand), which corresponds to 1 over the binomial coefficient: $1/\binom{M-1}{|S|}$. Therefore, a subset with a high or low value of cardinality has a large weight value. `shapley` includes the subsets with the highest weight first, and then includes the other subsets in descending order based on their weight values.

Reduce Computational Cost

This example shows how to reduce the computational cost for Shapley values when you have a large number of both observations and features.

Train Regression Ensemble

Load the sample data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 55,246 observations of 10 variables with information on the sales of properties in New York City in 2015. This example uses these variables to analyze the sale prices (`SALEPRICE`).

Preprocess the data set. Convert the `datetime` array (`SALEDATE`) to the month numbers.

```
NYCHousing2015.SALEDATE = month(NYCHousing2015.SALEDATE);
```

Train a regression ensemble.

```
Mdl = fitrensemble(NYCHousing2015,'SALEPRICE');
```

Compute Shapley Values with Default Options

Compute the Shapley values of all predictor variables for the first observation. Measure the time required to compute the Shapley values by using `tic` and `toc`.

```
tic
explainer1 = shapley(Mdl,'QueryPoint',NYCHousing2015(1,:));

Warning: Computation can be slow because the predictor data has over 1000 observations. Use a sma

toc

Elapsed time is 492.365307 seconds.
```

As the warning message indicates, the computation can be slow because the predictor data has over 1000 observations.

Specify Options to Reduce Computational Cost

`shapley` provides several options to reduce the computational cost when you have a large number of observations or features.

- Large number of observations — Use a smaller sample of the training data and compute Shapley values in parallel by specifying `'UseParallel'` as `true`.
- Large number of features — Specify the `'MaxNumSubsets'` name-value argument to limit the number of subsets included in the computation.

Compute the Shapley values again using a smaller sample of the training data and the parallel computing option. Also, specify the maximum number of subsets as `2^5`.

```
NumSamples = 5e2;
Tbl = datasample(NYCHousing2015,NumSamples,'Replace',false);
tic
explainer2 = shapley(Mdl,Tbl,'QueryPoint',NYCHousing2015(1,:), ...
    'UseParallel',true,'MaxNumSubsets',2^5);

Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

toc

Elapsed time is 52.183287 seconds.
```

Specifying the additional options reduces the time required to compute the Shapley values.

References

[1] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[2] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

[3] Kumar, I. Elizabeth, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. "Problems with Shapley-Value-Based Explanations as Feature Importance Measures." *arXiv:2002.11097* (2020).

See Also

`fit` | `plot` | `shapley`

Bibliography

[1] Agresti, A. *Categorical Data Analysis*, 2nd Ed. Hoboken, NJ: John Wiley & Sons, Inc., 2002.

[2] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[3] Alpaydin, E. “Combined 5 x 2 CV F Test for Comparing Supervised Classification Learning Algorithms.” *Neural Computation*, Vol. 11, No. 8, 1999, pp. 1885–1992.

[4] Blackard, J. A. and D. J. Dean. "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables". *Computers and Electronics in Agriculture* Vol. 24, Issue 3, 1999, pp. 131–151.

[5] Bottou, L., and Chih-Jen Lin. “Support Vector Machine Solvers.” *Large Scale Kernel Machines* (L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, eds.). Cambridge, MA: MIT Press, 2007.

[6] Bouckaert. R. “Choosing Between Two Learning Algorithms Based on Calibrated Tests.” *International Conference on Machine Learning*, pp. 51–58, 2003.

[7] Bouckaert, R. and E. Frank. “Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms.” *In Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference*, 2004, pp. 3–12.

[8] Breiman, L. "Bagging Predictors." *Machine Learning* 26, 1996, pp. 123–140.

[9] Breiman, L. "Random Forests." *Machine Learning* 45, 2001, pp. 5–32.

[10] Breiman, L. `https://www.stat.berkeley.edu/~breiman/RandomForests/`

[11] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall, 1984.

[12] Christianini, N., and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[13] Dietterich, T. “Approximate statistical tests for comparing supervised classification learning algorithms.” *Neural Computation*, Vol. 10, No. 7, 1998, pp. 1895–1923.

[14] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[15] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[16] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

[17] Fan, R.-E., P.-H. Chen, and C.-J. Lin. “Working set selection using second order information for training support vector machines.” *Journal of Machine Learning Research*, Vol 6, 2005, pp. 1889–1918.

[18] Fagerlan, M.W., S Lydersen, P. Laake. “The McNemar Test for Binary Matched-Pairs Data: Mid-p and Asymptotic Are Better Than Exact Conditional.” *BMC Medical Research Methodology*. Vol. 13, 2013, pp. 1–8.

[19] Freund, Y. "A more robust boosting algorithm." arXiv:0905.2138v1, 2009.

[20] Freund, Y. and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *J. of Computer and System Sciences*, Vol. 55, 1997, pp. 119–139.

[21] Friedman, J. "Greedy function approximation: A gradient boosting machine." *Annals of Statistics,* Vol. 29, No. 5, 2001, pp. 1189–1232.

[22] Friedman, J., T. Hastie, and R. Tibshirani. "Additive logistic regression: A statistical view of boosting." *Annals of Statistics*, Vol. 28, No. 2, 2000, pp. 337–407.

[23] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[24] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. New York: Springer, 2008.

[25] Ho, C. H. and C. J. Lin. “Large-Scale Linear Support Vector Regression.” *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 3323–3348.

[26] Ho, T. K. "The random subspace method for constructing decision forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 20, No. 8, 1998, pp. 832–844.

[27] Hsieh, C. J., K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. “A Dual Coordinate Descent Method for Large-Scale Linear SVM.” *Proceedings of the 25th International Conference on Machine Learning, ICML ’08*, 2001, pp. 408–415.

[28] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. *A Practical Guide to Support Vector Classification*. Available at `https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[29] Hu, Q., X. Che, L. Zhang, and D. Yu. “Feature Evaluation and Selection Based on Neighborhood Soft Margin.” *Neurocomputing*. Vol. 73, 2010, pp. 2114–2124.

[30] Kecman V., T. -M. Huang, and M. Vogt. “Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance.” In *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

[31] Kohavi, R. “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid.” *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[32] Lancaster, H.O. “Significance Tests in Discrete Distributions.” *JASA*, Vol. 56, Number 294, 1961, pp. 223–234.

[33] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[34] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[35] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

[36] McNemar, Q. “Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages.” *Psychometrika*, Vol. 12, Number 2, 1947, pp. 153–157.

[37] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

[38] Mosteller, F. “Some Statistical Problems in Measuring the Subjective Response to Drugs.” *Biometrics*, Vol. 8, Number 3, 1952, pp. 220–226.

[39] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

[40] Schapire, R. E. et al. "Boosting the margin: A new explanation for the effectiveness of voting methods." *Annals of Statistics,* Vol. 26, No. 5, 1998, pp. 1651–1686.

[41] Schapire, R., and Y. Singer. "Improved boosting algorithms using confidence-rated predictions." *Machine Learning,* Vol. 37, No. 3, 1999, pp. 297–336.

[42] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[43] Seiffert, C., T. Khoshgoftaar, J. Hulse, and A. Napolitano. "RUSBoost: Improving classification performance when training data is skewed." *19th International Conference on Pattern Recognition,* 2008, pp. 1–4.

[44] Warmuth, M., J. Liao, and G. Ratsch. "Totally corrective boosting algorithms that maximize the margin." *Proc. 23rd Int’l. Conf. on Machine Learning, ACM,* New York, 2006, pp. 1001–1008.

[45] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[46] Wright, S. J., R. D. Nowak, and M. A. T. Figueiredo. “Sparse Reconstruction by Separable Approximation.” *Trans. Sig. Proc.*, Vol. 57, No 7, 2009, pp. 2479–2493.

[47] Xiao, Lin. “Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization.” *J. Mach. Learn. Res.*, Vol. 11, 2010, pp. 2543–2596.

[48] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

[49] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

[50] Zadrozny, B., J. Langford, and N. Abe. “Cost-Sensitive Learning by Cost-Proportionate Example Weighting.” *Third IEEE International Conference on Data Mining*, 435–442. 2003.

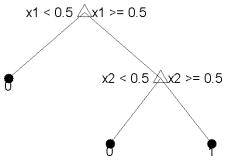
[51] Zhou, Z.-H. and X.-Y. Liu. “On Multi-Class Cost-Sensitive Learning.” *Computational Intelligence.* Vol. 26, Issue 3, 2010, pp. 232–257 CiteSeerX.

Decision Trees

Decision Trees

Decision trees, or classification trees and regression trees, predict responses to data. To predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. Classification trees give responses that are nominal, such as `'true'` or `'false'`. Regression trees give numeric responses.

Statistics and Machine Learning Toolbox trees are binary. Each step in a prediction involves checking the value of one predictor (variable). For example, here is a simple classification tree:



This tree predicts classifications based on two predictors, `x1` and `x2`. To predict, start at the top node, represented by a triangle (Δ). The first decision is whether `x1` is smaller than `0.5`. If so, follow the left branch, and see that the tree classifies the data as type `0`.

If, however, `x1` exceeds `0.5`, then follow the right branch to the lower-right triangle node. Here the tree asks if `x2` is smaller than `0.5`. If so, then follow the left branch to see that the tree classifies the data as type `0`. If not, then follow the right branch to see that the tree classifies the data as type `1`.

To learn how to prepare your data for classification or regression using decision trees, see “Steps in Supervised Learning” on page 18-4.

Train Classification Tree

This example shows how to train a classification tree.

Create a classification tree using the entire `ionosphere` data set.

```
load ionosphere % Contains X and Y variables
Mdl = fitctree(X,Y)
```

```
Mdl =
  ClassificationTree
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351


  Properties, Methods
```

Train Regression Tree

This example shows how to train a regression tree.

Create a regression tree using all observation in the `carsmall` data set. Consider the `Horsepower` and `Weight` vectors as predictor variables, and the `MPG` vector as the response.

```
load carsmall % Contains Horsepower, Weight, MPG
X = [Horsepower Weight];

Mdl = fitrtree(X,MPG)

Mdl =
  RegressionTree
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94


  Properties, Methods
```

References

[1] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall, 1984.

See Also

`ClassificationTree` | `RegressionTree` | `fitctree` | `fitrtree`

Related Examples

- “View Decision Tree” on page 19-4
- “Growing Decision Trees” on page 19-7
- “Prediction Using Classification and Regression Trees” on page 19-9
- “Improving Classification Trees and Regression Trees” on page 19-13

View Decision Tree

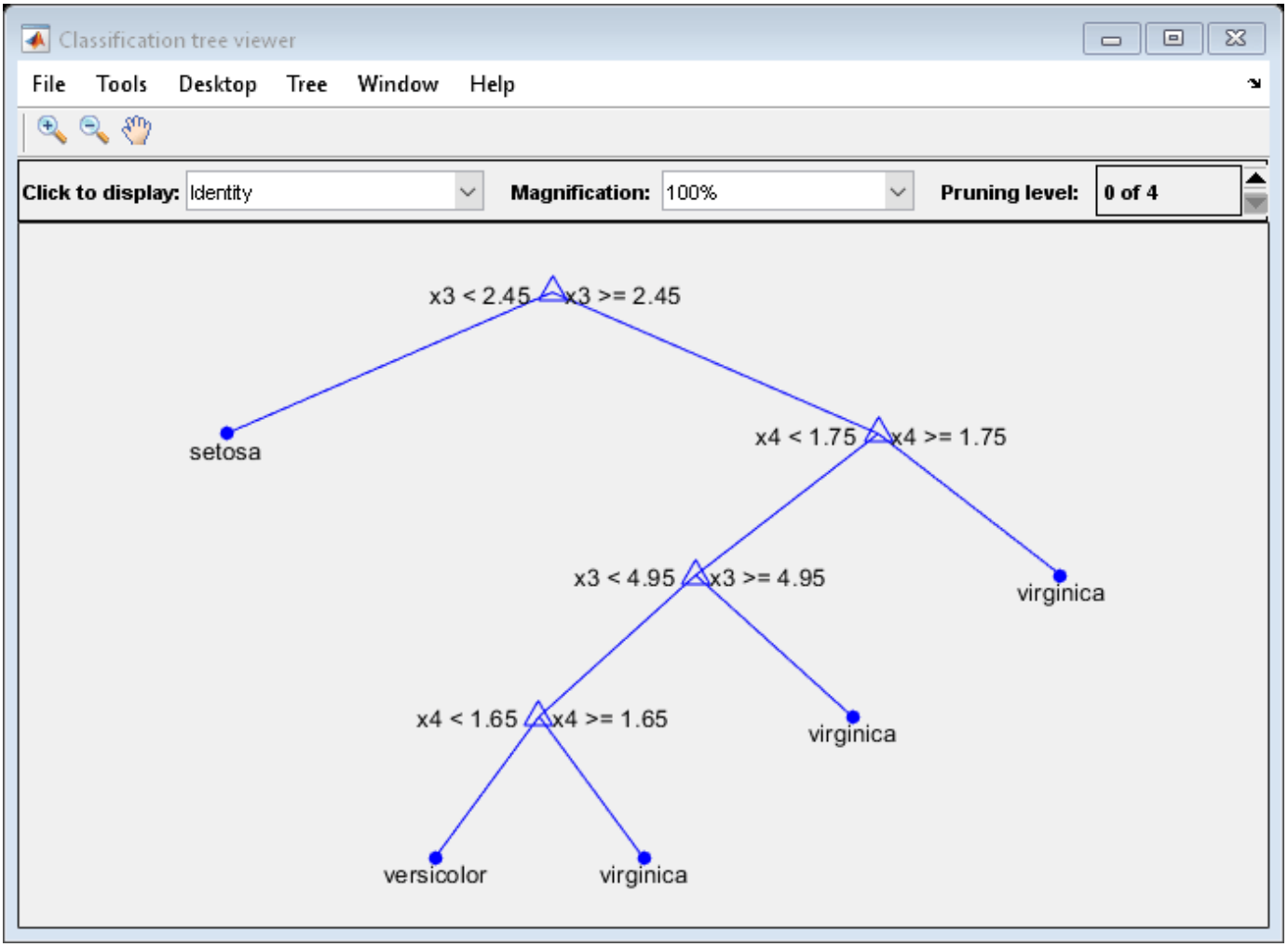
This example shows how to view a classification or regression tree. There are two ways to view a tree: `view(tree)` returns a text description and `view(tree,'mode','graph')` returns a graphic description of the tree.

Create and view a classification tree.

```
load fisheriris % load the sample data
ctree = fitctree(meas,species); % create classification tree
view(ctree) % text description
```

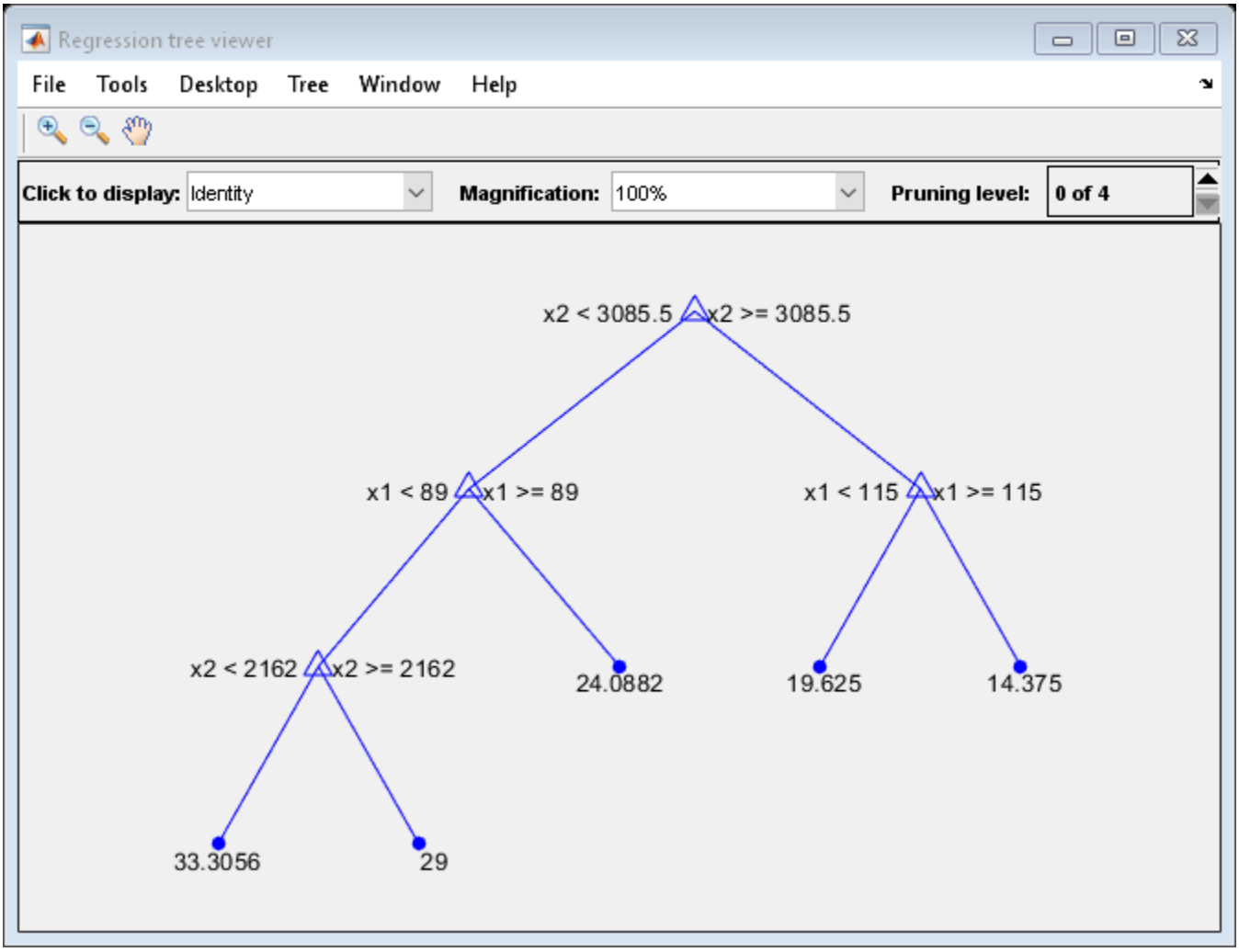
```
Decision tree for classification
1  if x3<2.45 then node 2 elseif x3>=2.45 then node 3 else setosa
2  class = setosa
3  if x4<1.75 then node 4 elseif x4>=1.75 then node 5 else versicolor
4  if x3<4.95 then node 6 elseif x3>=4.95 then node 7 else versicolor
5  class = virginica
6  if x4<1.65 then node 8 elseif x4>=1.65 then node 9 else versicolor
7  class = virginica
8  class = versicolor
9  class = virginica
```

```
view(ctree,'mode','graph') % graphic description
```



Now, create and view a regression tree.

```
load carsmall % load the sample data, contains Horsepower, Weight, MPG
X = [Horsepower Weight];
rtree = fitrtree(X,MPG,'MinParent',30); % create classification tree
view(rtree) % text description

Decision tree for regression
1  if x2<3085.5 then node 2 elseif x2>=3085.5 then node 3 else 23.7181
2  if x1<89 then node 4 elseif x1>=89 then node 5 else 28.7931
3  if x1<115 then node 6 elseif x1>=115 then node 7 else 15.5417
4  if x2<2162 then node 8 elseif x2>=2162 then node 9 else 30.9375
5  fit = 24.0882
6  fit = 19.625
7  fit = 14.375
8  fit = 33.3056
9  fit = 29

view(rtree,'mode','graph') % graphic description
```



See Also

`fitctree` | `fitrtree` | `view (CompactClassificationTree)` | `view (CompactRegressionTree)`

Related Examples

- “Decision Trees” on page 19-2

Growing Decision Trees

By default, `fitctree` and `fitrtree` use the standard CART algorithm [1] to create decision trees. That is, they perform the following steps:

1. Start with all input data, and examine all possible binary splits on every predictor.
2. Select a split with best optimization criterion.
   - A split might lead to a child node having too few observations (less than the `MinLeafSize` parameter). To avoid this, the software chooses a split that yields the best optimization criterion subject to the `MinLeafSize` constraint.
3. Impose the split.
4. Repeat recursively for the two child nodes.

The explanation requires two more items: description of the optimization criterion and stopping rule.

Stopping rule: Stop splitting when any of the following hold:

- The node is pure.
  - For classification, a node is pure if it contains only observations of one class.
  - For regression, a node is pure if the mean squared error (MSE) for the observed response in this node drops below the MSE for the observed response in the entire data multiplied by the tolerance on quadratic error per node (`QuadraticErrorTolerance` parameter).
- There are fewer than `MinParentSize` observations in this node.
- Any split imposed on this node produces children with fewer than `MinLeafSize` observations.
- The algorithm splits `MaxNumSplits` nodes.

Optimization criterion:

- Regression: mean-squared error (MSE). Choose a split to minimize the MSE of predictions compared to the training data.
- Classification: One of three measures, depending on the setting of the `SplitCriterion` name-value pair:
  - `'gdi'` (Gini's diversity index, the default)
  - `'twoing'`
  - `'deviance'`

  For details, see `ClassificationTree` “More About” on page 33-585.

For alternative split predictor selection techniques, see “Choose Split Predictor Selection Technique” on page 19-14.

For a continuous predictor, a tree can split halfway between any two adjacent unique values found for this predictor. For a categorical predictor with L levels, a classification tree needs to consider $2^{L-1}-1$ splits to find the optimal split. Alternatively, you can choose a heuristic algorithm to find a good split, as described in “Splitting Categorical Predictors in Classification Trees” on page 19-25.

For dual-core systems and above, `fitctree` and `fitrtree` parallelize training decision trees using Intel® Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall, 1984.

See Also

`ClassificationTree` | `RegressionTree` | `fitctree` | `fitrtree`

Related Examples

- “Decision Trees” on page 19-2
- “Splitting Categorical Predictors in Classification Trees” on page 19-25

Prediction Using Classification and Regression Trees

This example shows how to predict class labels or responses using trained classification and regression trees.

After creating a tree, you can easily predict responses for new data. Suppose `Xnew` is new data that has the same number of columns as the original data `X`. To predict the classification or regression based on the tree (`Mdl`) and the new data, enter

```
Ynew = predict(Mdl,Xnew)
```

For each row of data in `Xnew`, `predict` runs through the decisions in `Mdl` and gives the resulting prediction in the corresponding element of `Ynew`. For more information on classification tree prediction, see the `predict`. For regression, see `predict`.

For example, find the predicted classification of a point at the mean of the `ionosphere` data.

```
load ionosphere
CMdl = fitctree(X,Y);
Ynew = predict(CMdl,mean(X))
```

```
Ynew = 1x1 cell array
    {'g'}
```

Find the predicted `MPG` of a point at the mean of the `carsmall` data.

```
load carsmall
X = [Horsepower Weight];
RMdl = fitrtree(X,MPG);
Ynew = predict(RMdl,mean(X))
```

```
Ynew = 28.7931
```

See Also

`ClassificationTree` | `RegressionTree` | `fitctree` | `fitrtree` | `predict (CompactClassificationTree)` | `predict (CompactRegressionTree)`

Related Examples

- “Decision Trees” on page 19-2
- “Predict Out-of-Sample Responses of Subtrees” on page 19-10
- “Improving Classification Trees and Regression Trees” on page 19-13

Predict Out-of-Sample Responses of Subtrees

This example hows how to predict out-of-sample responses of regression trees, and then plot the results.

Load the `carsmall` data set. Consider `Weight` as a predictor of the response `MPG`.

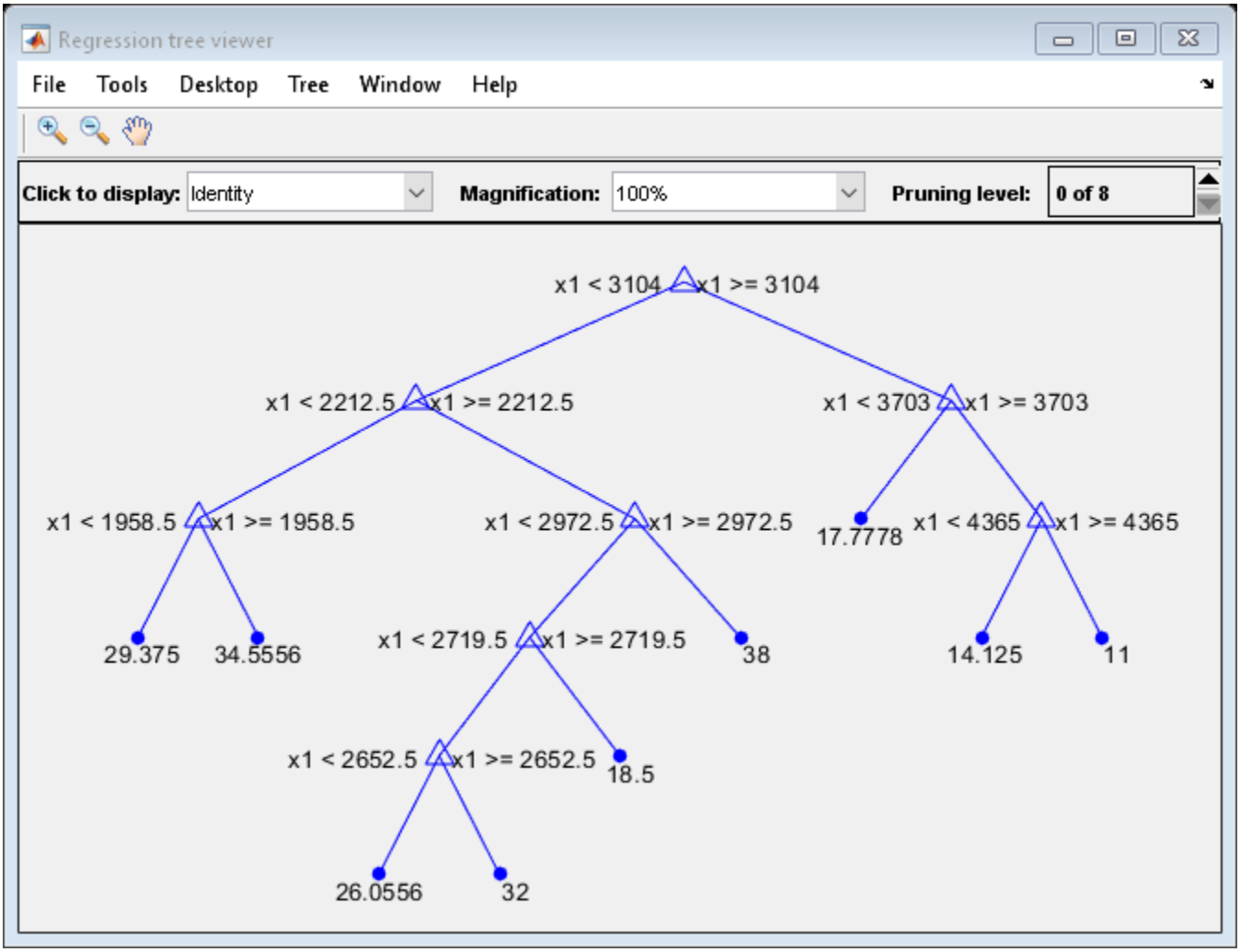
```
load carsmall
idxNaN = isnan(MPG + Weight);
X = Weight(~idxNaN);
Y = MPG(~idxNaN);
n = numel(X);
```

Partition the data into training (50%) and validation (50%) sets.

```
rng(1) % For reproducibility
idxTrn = false(n,1);
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a regression tree using the training observations.

```
Mdl = fitrtree(X(idxTrn),Y(idxTrn));
view(Mdl,'Mode','graph')
```



Compute fitted values of the validation observations for each of several subtrees.

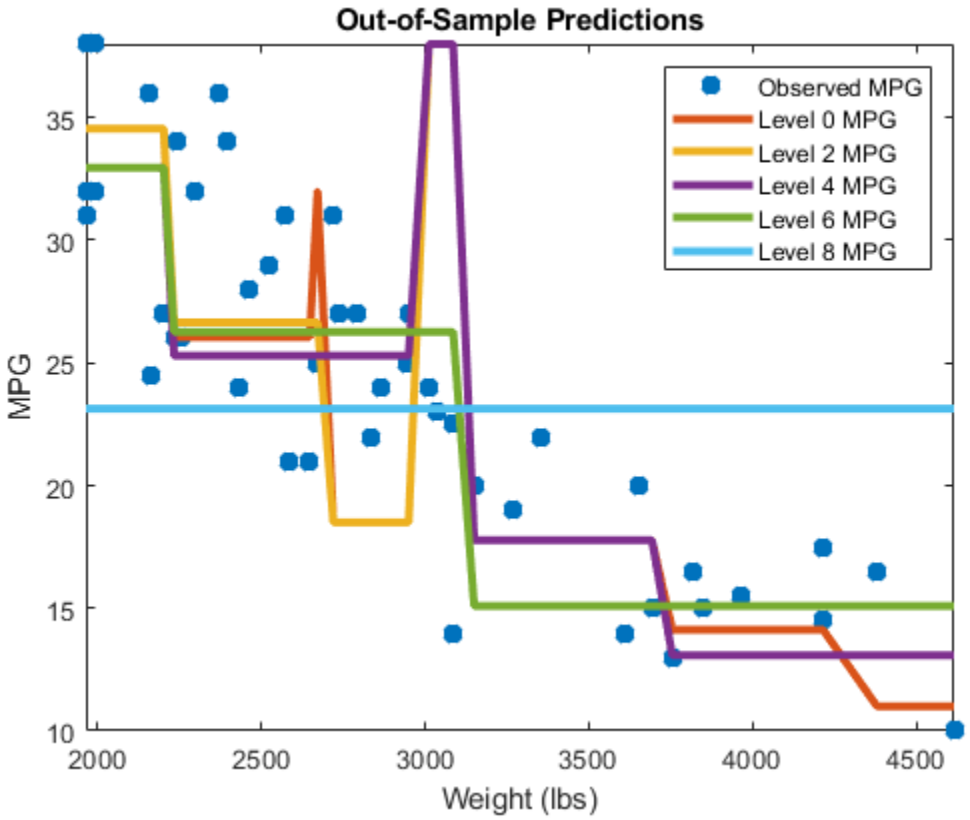
```
m = max(Mdl.PruneList);
pruneLevels = 0:2:m; % Pruning levels to consider
z = numel(pruneLevels);
Yfit = predict(Mdl,X(idxVal),'SubTrees',pruneLevels);
```

`Yfit` is an n-by- z matrix of fitted values in which the rows correspond to observations and the columns correspond to a subtree.

Plot `Yfit` and `Y` against `X`.

```
figure;
sortDat = sortrows([X(idxVal) Y(idxVal) Yfit],1); % Sort all data with respect to X
plot(sortDat(:,1),sortDat(:,2),'*');
hold on;
plot(repmat(sortDat(:,1),1,size(Yfit,2)),sortDat(:,3:end));
lev = cellstr(num2str((pruneLevels)','Level %d MPG'));
legend(['Observed MPG'; lev])
title 'Out-of-Sample Predictions'
xlabel 'Weight (lbs)';
```

```
ylabel 'MPG';
h = findobj(gcf);
axis tight;
set(h(4:end),'LineWidth',3) % Widen all lines
```



The values of `Yfit` for lower pruning levels tend to follow the data more closely than higher levels. Higher pruning levels tend to be flat for large `X` intervals.

See Also

`RegressionTree` | `predict`

Related Examples

- “Decision Trees” on page 19-2
- “Improving Classification Trees and Regression Trees” on page 19-13
- “Prediction Using Classification and Regression Trees” on page 19-9

Improving Classification Trees and Regression Trees

In this section...

You can tune trees by setting name-value pairs in `fitctree` and `fitrtree`. The remainder of this section describes how to determine the quality of a tree, how to decide which name-value pairs to set, and how to control the size of a tree.

Examining Resubstitution Error

Resubstitution error is the difference between the response training data and the predictions the tree makes of the response based on the input training data. If the resubstitution error is high, you cannot expect the predictions of the tree to be good. However, having low resubstitution error does not guarantee good predictions for new data. Resubstitution error is often an overly optimistic estimate of the predictive error on new data.

Classification Tree Resubstitution Error

This example shows how to examine the resubstitution error of a classification tree.

Load Fisher's iris data.

```
load fisheriris
```

Train a default classification tree using the entire data set.

```
Mdl = fitctree(meas,species);
```

Examine the resubstitution error.

```
resuberror = resubLoss(Mdl)
```

```
resuberror = 0.0200
```

The tree classifies nearly all the Fisher iris data correctly.

Cross Validation

To get a better sense of the predictive accuracy of your tree for new data, cross validate the tree. By default, cross validation splits the training data into 10 parts at random. It trains 10 new trees, each one on nine parts of the data. It then examines the predictive accuracy of each new tree on the data not included in training that tree. This method gives a good estimate of the predictive accuracy of the resulting tree, since it tests the new trees on new data.

Cross Validate a Regression Tree

This example shows how to examine the resubstitution and cross-validation accuracy of a regression tree for predicting mileage based on the `carsmall` data.

Load the `carsmall` data set. Consider acceleration, displacement, horsepower, and weight as predictors of MPG.

```
load carsmall
X = [Acceleration Displacement Horsepower Weight];
```

Grow a regression tree using all of the observations.

```
rtree = fitrtree(X,MPG);
```

Compute the in-sample error.

```
resuberror = resubLoss(rtree)
```

```
resuberror = 4.7188
```

The resubstitution loss for a regression tree is the mean-squared error. The resulting value indicates that a typical predictive error for the tree is about the square root of 4.7, or a bit over 2.

Estimate the cross-validation MSE.

```
rng 'default';
cvrtree = crossval(rtree);
cvloss = kfoldLoss(cvrtree)
```

```
cvloss = 23.5706
```

The cross-validated loss is almost 25, meaning a typical predictive error for the tree on new data is about 5. This demonstrates that cross-validated loss is usually higher than simple resubstitution loss.

Choose Split Predictor Selection Technique

The standard CART algorithm tends to select continuous predictors that have many levels. Sometimes, such a selection can be spurious and can also mask more important predictors that have fewer levels, such as categorical predictors. That is, the predictor-selection process at each node is biased. Also, standard CART tends to miss the important interactions between pairs of predictors and the response.

To mitigate selection bias and increase detection of important interactions, you can specify usage of the curvature or interaction tests using the `'PredictorSelection'` name-value pair argument. Using the curvature or interaction test has the added advantage of producing better predictor importance estimates than standard CART.

This table summarizes the supported predictor-selection techniques.

Technique	'PredictorSelection' Value	Description	Training speed	When to specify
Standard CART [1]	Default	Selects the split predictor that maximizes the split-criterion gain over all possible splits of all predictors.	Baseline for comparison	Specify if any of these conditions are true: • All predictors are continuous • Predictor importance is not the analysis goal • For boosting decision trees
Curvature test [2][3]	'curvature'	Selects the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response.	Comparable to standard CART	Specify if any of these conditions are true: • The predictor variables are heterogeneous • Predictor importance is an analysis goal • Enhance tree interpretation
Interaction test [3]	'interaction-curvature'	Chooses the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response (that is, conducts curvature tests), and that minimizes the *p*-value of a chi-square test of independence between each pair of predictors and response.	Slower than standard CART, particularly when data set contains many predictor variables.	Specify if any of these conditions are true: • The predictor variables are heterogeneous • You suspect associations between pairs of predictors and the response • Predictor importance is an analysis goal • Enhance tree interpretation

For more details on predictor selection techniques:

- For classification trees, see `PredictorSelection` and “Node Splitting Rules” on page 33-1928.
- For regression trees, see `PredictorSelection` and “Node Splitting Rules” on page 33-2413.

Control Depth or “Leafiness”

When you grow a decision tree, consider its simplicity and predictive power. A deep tree with many leaves is usually highly accurate on the training data. However, the tree is not guaranteed to show a comparable accuracy on an independent test set. A leafy tree tends to overtrain (or overfit), and its

test accuracy is often far less than its training (resubstitution) accuracy. In contrast, a shallow tree does not attain high training accuracy. But a shallow tree can be more robust — its training accuracy could be close to that of a representative test set. Also, a shallow tree is easy to interpret. If you do not have enough data for training and test, estimate tree accuracy by cross validation.

`fitctree` and `fitrtree` have three name-value pair arguments that control the depth of resulting decision trees:

- `MaxNumSplits` — The maximal number of branch node splits is `MaxNumSplits` per tree. Set a large value for `MaxNumSplits` to get a deep tree. The default is `size(X,1) - 1`.
- `MinLeafSize` — Each leaf has at least `MinLeafSize` observations. Set small values of `MinLeafSize` to get deep trees. The default is `1`.
- `MinParentSize` — Each branch node in the tree has at least `MinParentSize` observations. Set small values of `MinParentSize` to get deep trees. The default is `10`.

If you specify `MinParentSize` and `MinLeafSize`, the learner uses the setting that yields trees with larger leaves (i.e., shallower trees):

```
MinParent = max(MinParentSize,2*MinLeafSize)
```

If you supply `MaxNumSplits`, the software splits a tree until one of the three splitting criteria is satisfied.

For an alternative method of controlling the tree depth, see “Pruning” on page 19-19.

Select Appropriate Tree Depth

This example shows how to control the depth of a decision tree, and how to choose an appropriate depth.

Load the `ionosphere` data.

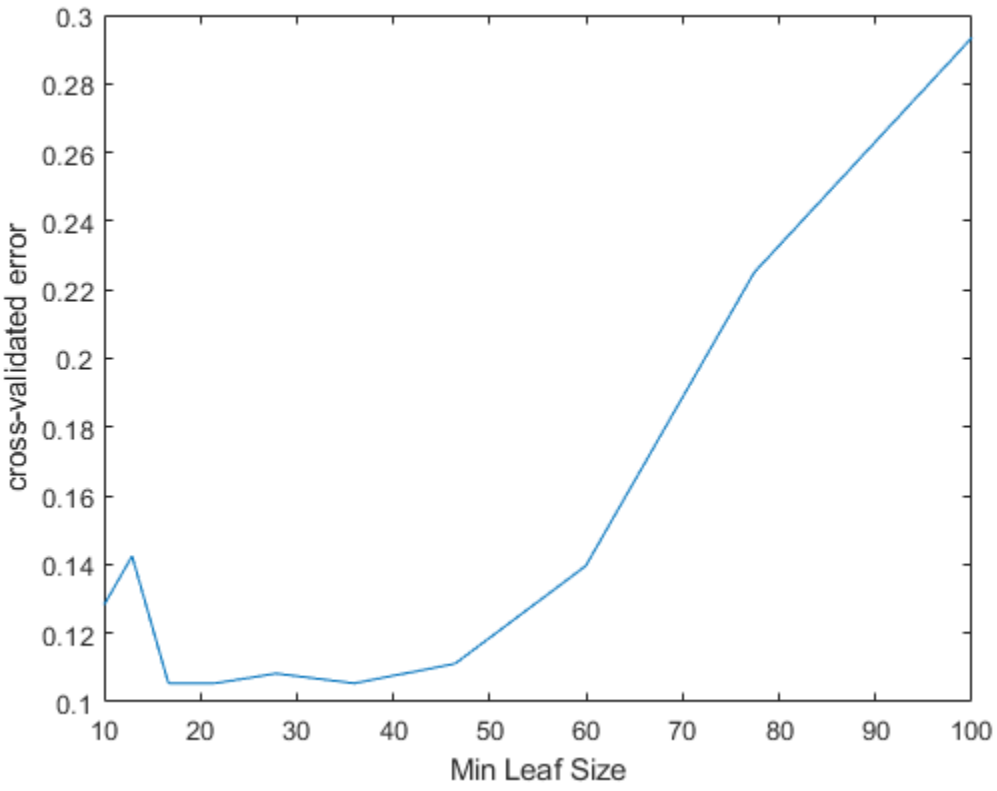
```
load ionosphere
```

Generate an exponentially spaced set of values from `10` through `100` that represent the minimum number of observations per leaf node.

```
leafs = logspace(1,2,10);
```

Create cross-validated classification trees for the `ionosphere` data. Specify to grow each tree using a minimum leaf size in `leafs`.

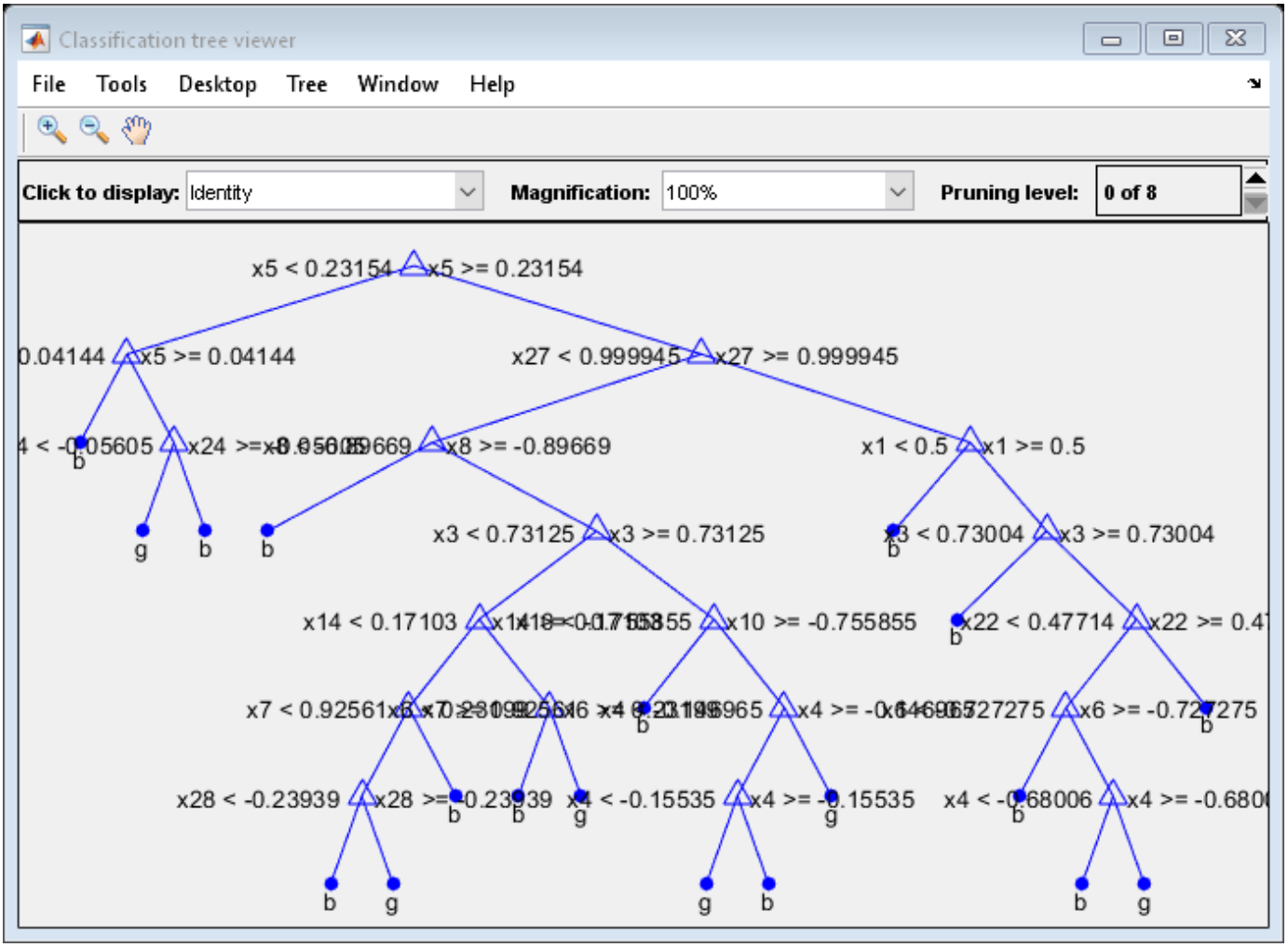
```
rng('default')
N = numel(leafs);
err = zeros(N,1);
for n=1:N
    t = fitctree(X,Y,'CrossVal','On',...
        'MinLeafSize',leafs(n));
    err(n) = kfoldLoss(t);
end
plot(leafs,err);
xlabel('Min Leaf Size');
ylabel('cross-validated error');
```



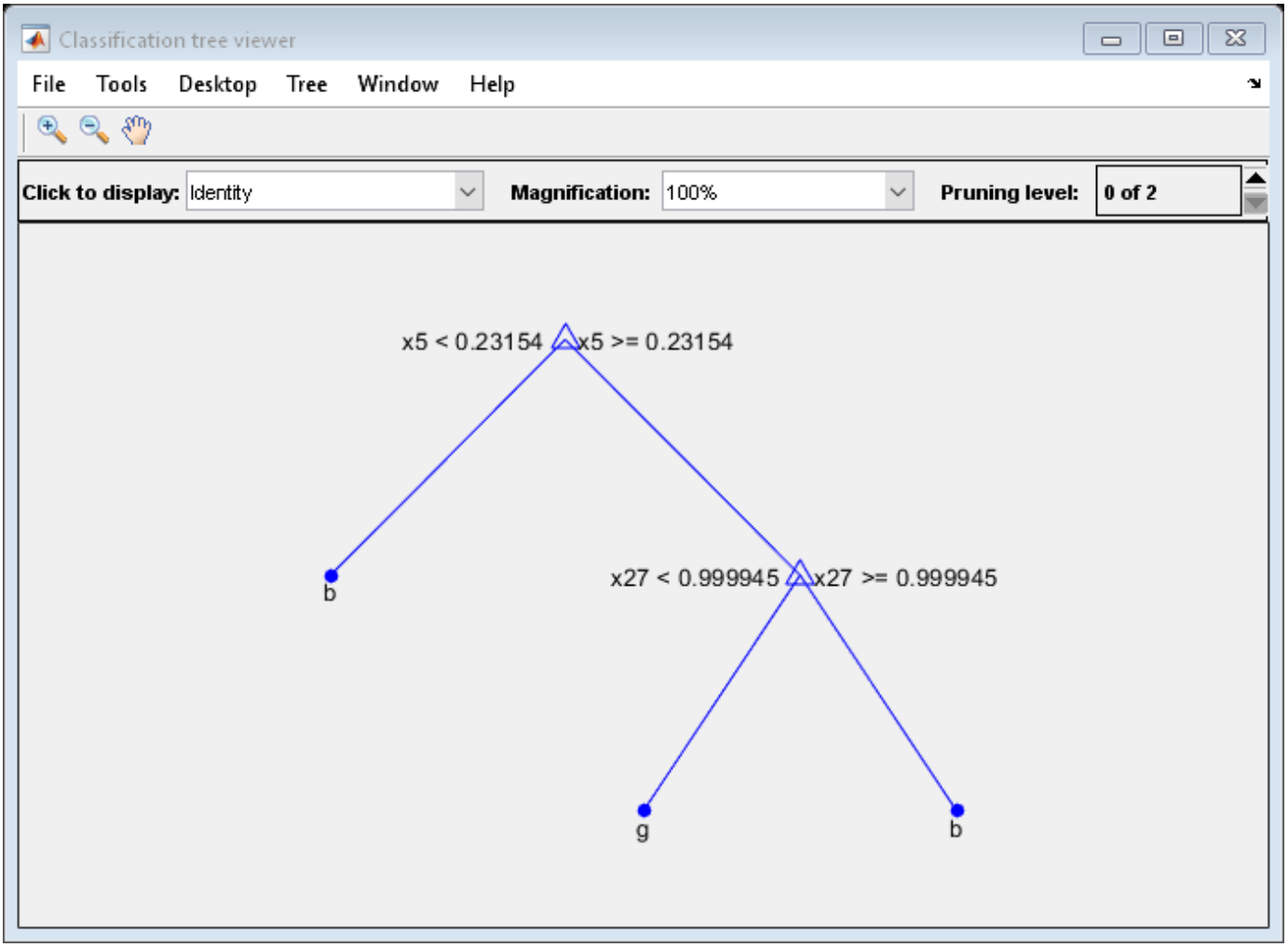
The best leaf size is between about `20` and `50` observations per leaf.

Compare the near-optimal tree with at least `40` observations per leaf with the default tree, which uses `10` observations per parent node and `1` observation per leaf.

```
DefaultTree = fitctree(X,Y);
view(DefaultTree,'Mode','Graph')
```




```
OptimalTree = fitctree(X,Y,'MinLeafSize',40);
view(OptimalTree,'mode','graph')
```




```
resubOpt = resubLoss(OptimalTree);
lossOpt = kfoldLoss(crossval(OptimalTree));
resubDefault = resubLoss(DefaultTree);
lossDefault = kfoldLoss(crossval(DefaultTree));
resubOpt,resubDefault,lossOpt,lossDefault
```

```
resubOpt = 0.0883

resubDefault = 0.0114

lossOpt = 0.1054

lossDefault = 0.1054
```

The near-optimal tree is much smaller and gives a much higher resubstitution error. Yet, it gives similar accuracy for cross-validated data.

Pruning

Pruning optimizes tree depth (leafiness) by merging leaves on the same tree branch. “Control Depth or “Leafiness”” on page 19-15 describes one method for selecting the optimal depth for a tree. Unlike

in that section, you do not need to grow a new tree for every node size. Instead, grow a deep tree, and prune it to the level you choose.

Prune a tree at the command line using the `prune` method (classification) or `prune` method (regression). Alternatively, prune a tree interactively with the tree viewer:

```
view(tree,'mode','graph')
```

To prune a tree, the tree must contain a pruning sequence. By default, both `fitctree` and `fitrtree` calculate a pruning sequence for a tree during construction. If you construct a tree with the `'Prune'` name-value pair set to `'off'`, or if you prune a tree to a smaller level, the tree does not contain the full pruning sequence. Generate the full pruning sequence with the `prune` method (classification) or `prune` method (regression).

Prune a Classification Tree

This example creates a classification tree for the `ionosphere` data, and prunes it to a good level.

Load the `ionosphere` data:

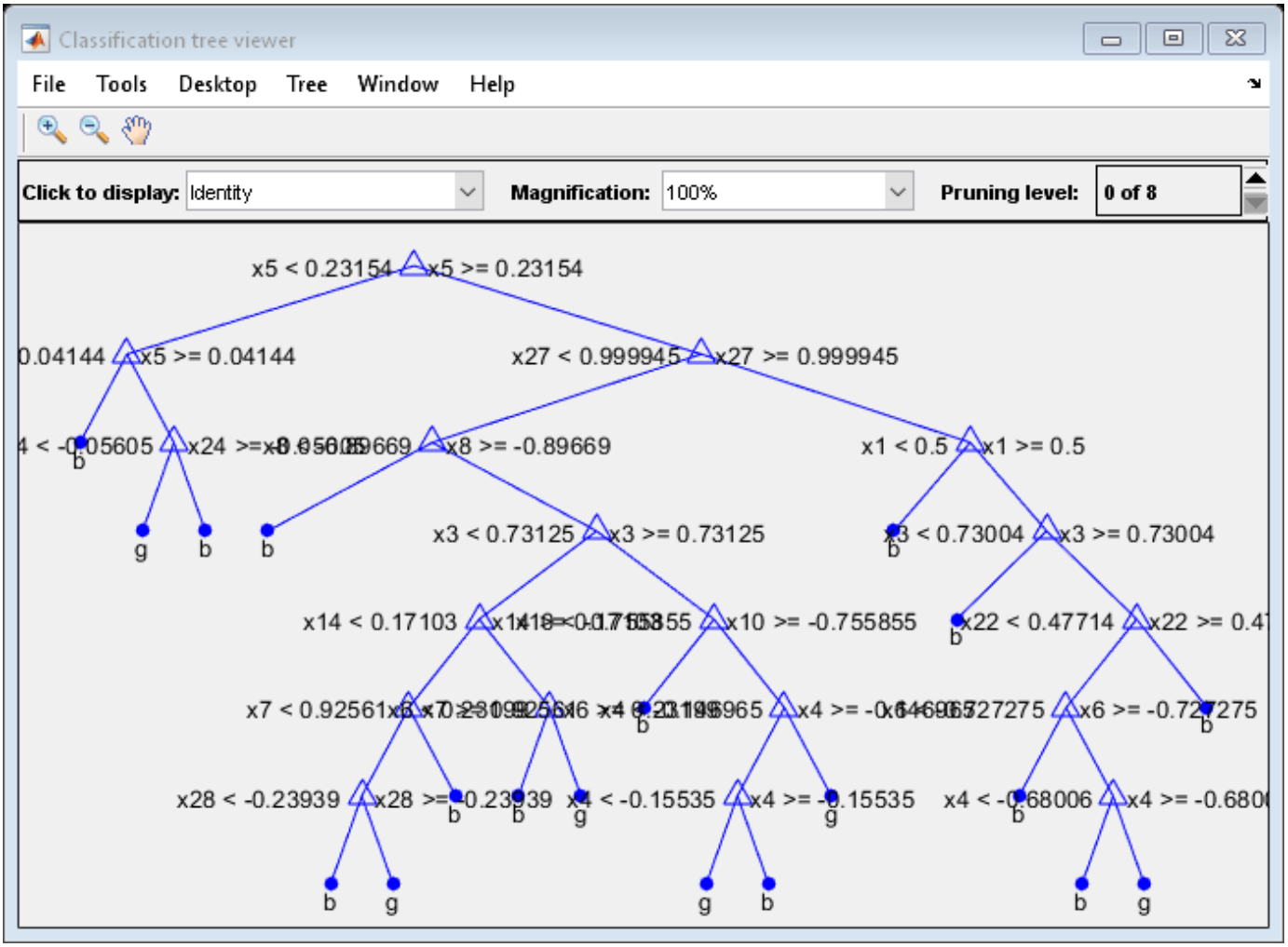
```
load ionosphere
```

Construct a default classification tree for the data:

```
tree = fitctree(X,Y);
```

View the tree in the interactive viewer:

```
view(tree,'Mode','Graph')
```



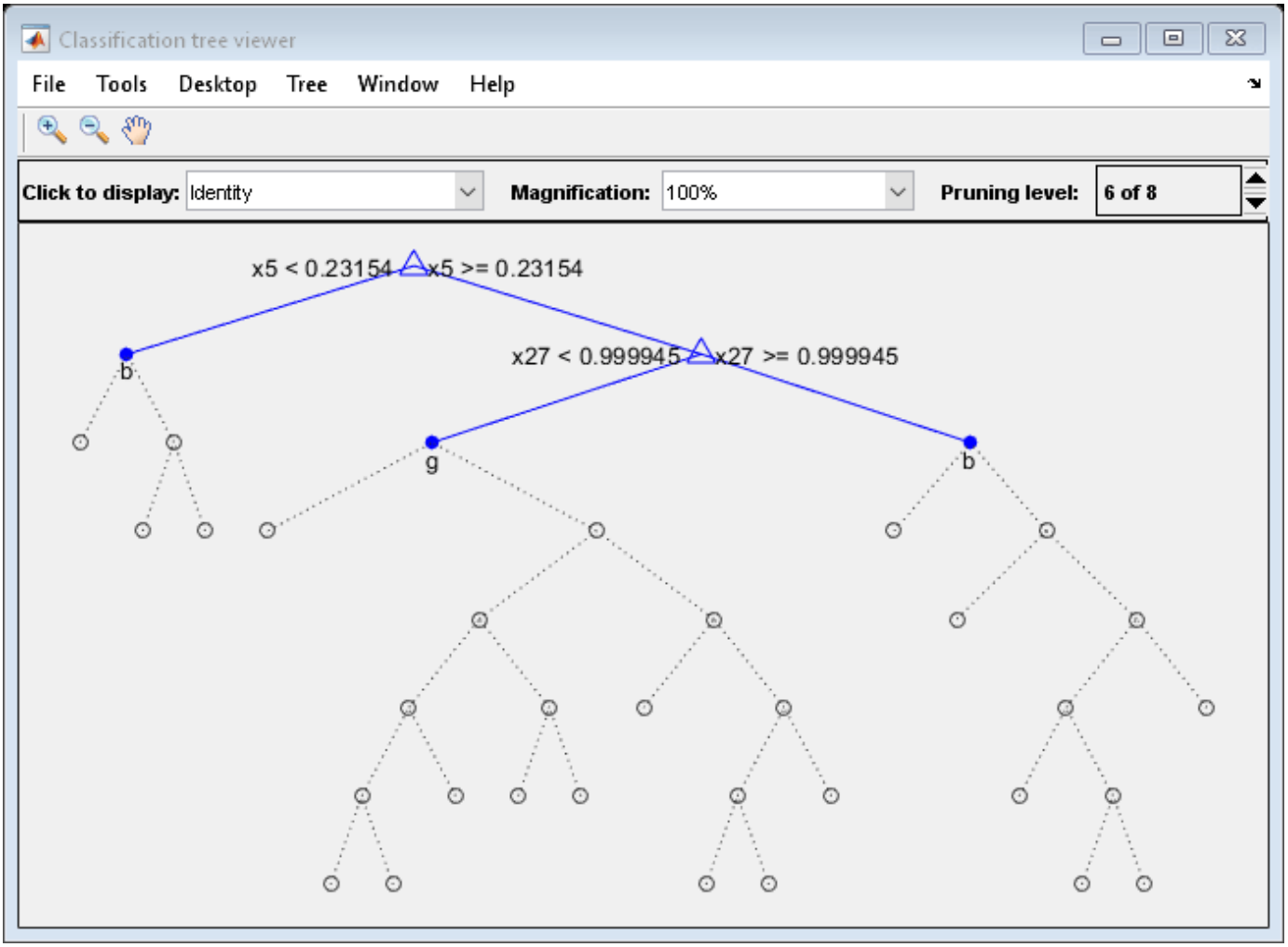
Find the optimal pruning level by minimizing cross-validated loss:

```
[~,~,~,bestlevel] = cvLoss(tree,...
    'SubTrees','All','TreeSize','min')
```

```
bestlevel = 6
```

Prune the tree to level 6:

```
view(tree,'Mode','Graph','Prune',6)
```



Alternatively, use the interactive window to prune the tree.

The pruned tree is the same as the near-optimal tree in the "Select Appropriate Tree Depth" example.

Set `'TreeSize'` to `'SE'` (default) to find the maximal pruning level for which the tree error does not exceed the error from the best level plus one standard deviation:

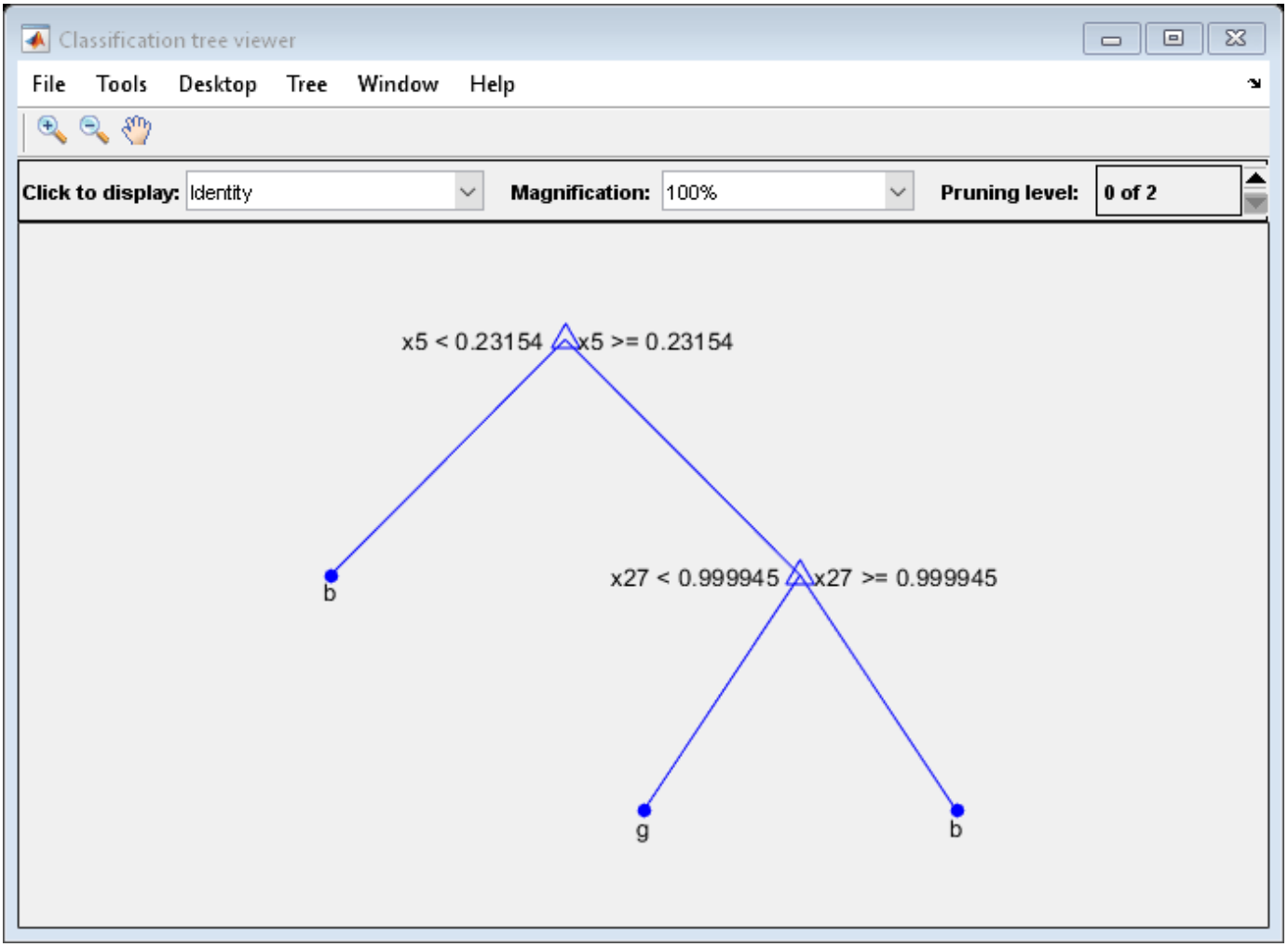
```
[~,~,~,bestlevel] = cvLoss(tree,'SubTrees','All')
```

```
bestlevel = 6
```

In this case the level is the same for either setting of `'TreeSize'`.

Prune the tree to use it for other purposes:

```
tree = prune(tree,'Level',6);
view(tree,'Mode','Graph')
```



References

[1] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Boca Raton, FL: Chapman & Hall, 1984.

[2] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

[3] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

See Also

`ClassificationTree` | `RegressionTree` | `fitctree` | `fitrtree` | `predict (CompactClassificationTree)` | `predict (CompactRegressionTree)` | `prune (ClassificationTree)` | `prune (RegressionTree)`

Related Examples

- “Decision Trees” on page 19-2
- “Prediction Using Classification and Regression Trees” on page 19-9
- “Predict Out-of-Sample Responses of Subtrees” on page 19-10

Splitting Categorical Predictors in Classification Trees

In this section...

Challenges in Splitting Multilevel Predictors

When you grow a classification tree, finding an optimal binary split for a categorical predictor with many levels is more computationally challenging than finding a split for a continuous predictor. For a continuous predictor, a tree can split halfway between any two adjacent, unique values of the predictor. In contrast, to find an exact, optimal binary split for a categorical predictor with L levels, a classification tree must consider $2^{L-1}–1$ splits. To obtain this formula:

1. Count the number of ways to assign L distinct values to the left and right nodes. There are 2^L ways.
2. Divide 2^L by 2, because left and right can be swapped.
3. Discard the case that has an empty node.

For regression problems and binary classification problems, the software uses the exact search algorithm through a computational shortcut[1]. The tree can order the categories by mean response (for regression) or class probability for one of the classes (for classification). Then, the optimal split is one of the L – 1 splits for the ordered list. Therefore, computational challenges arise only when you grow classification trees for data with $K \geq 3$ classes.

Algorithms for Categorical Predictor Split

To reduce computation, the software offers several heuristic algorithms for finding a good split. You can choose an algorithm for splitting categorical predictors by using the `'AlgorithmForCategorical'` name-value pair argument when you grow a classification tree using `fitctree` or when you create a classification learner using `templateTree` for a classification ensemble (`fitcensemble`) or a multiclass ECOC model (`fitcecoc`).

If you do not specify an algorithm, the software selects the optimal algorithm for each split using the known number of classes and levels of a categorical predictor. If the predictor has at most `MaxNumCategories` levels, the software splits categorical predictors using the exact search algorithm. Otherwise, the software chooses a heuristic search algorithm based on the number of classes and levels. The default `MaxNumCategories` level is 10. Depending on your platform, the software cannot perform an exact search on categorical predictors with more than 32 or 64 levels.

The available heuristic algorithms are: pull left by purity, principal component-based partitioning, and one-versus-all by class.

Pull Left by Purity

The pull left by purity algorithm starts with all L categorical levels on the right branch. The algorithm then takes these actions:

1. Inspect the K categories that have the largest class probabilities for each class.

2. Move the category with the maximum value of the split criterion to the left branch.
3. Continue moving categories from right to left, recording the split criterion at each move, until the right child has only one category remaining.

Out of this sequence, the selected split is the one that maximizes the split criterion.

Select this algorithm by specifying `'AlgorithmForCategorical','PullLeft'` in `fitctree` or `templateTree`.

Principal Component-Based Partitioning

The principal component-based partitioning algorithm[2] finds a close-to-optimal binary partition of the L predictor levels by searching for a separating hyperplane. The hyperplane is perpendicular to the first principal component of the weighted covariance matrix of the centered class probability matrix.

The algorithm assigns a score to each of the L categories, computed as the inner product between the found principal component and the vector of class probabilities for that category. Then, the selected split is one of the L – 1 splits that maximizes the split criterion.

Select this algorithm by specifying `'AlgorithmForCategorical','PCA'` in `fitctree` or `templateTree`.

One-Versus-All by Class

The one-versus-all by class algorithm starts with all L categorical levels on the right branch. For each of the K classes, the algorithm orders the categories based on their probability for that class.

For the first class, the algorithm moves each category to the left branch in order, recording the split criterion at each move. Then the algorithm repeats this process for the remaining classes. Out of this sequence, the selected split is the one that maximizes the split criterion.

Select this algorithm by specifying `'AlgorithmForCategorical','OVAbyClass'` in `fitctree` or `templateTree`.

Inspect Data with Multilevel Categorical Predictors

This example shows how to inspect a data set that includes categorical predictors with many levels (categories) and how to train a binary decision tree for classification.

Load Sample Data

Load the `census1994` file. This data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 a year. Specify a cell array of character vectors containing the variable names.

```
load census1994
VarNames = adultdata.Properties.VariableNames;
```

Some variable names in the `adultdata` table contain the _ character. Replace instances of _ with a space.

```
VarNames = strrep(VarNames,'_',' ');
```

Specify the predictor data `tbl` and the response vector `Y`.

```
tbl = adultdata(:,1:end-1);
Y = categorical(adultdata.salary);
```

Inspect Categorical Predictors

Some categorical variables have many levels (categories). Count the number of levels of each categorical predictor.

Find the indexes of categorical predictors that are not numeric in the `tbl` table by using `varfun` and `isnumeric`. The `varfun` function applies the `isnumeric` function to each variable of the table `tbl`.

```
cat = ~varfun(@isnumeric,tbl,'OutputFormat','uniform');
```

Define an anonymous function to count the number of categories in a categorical predictor using `numel` and `categories`.

```
countNumCats = @(var)numel(categories(categorical(var)));
```

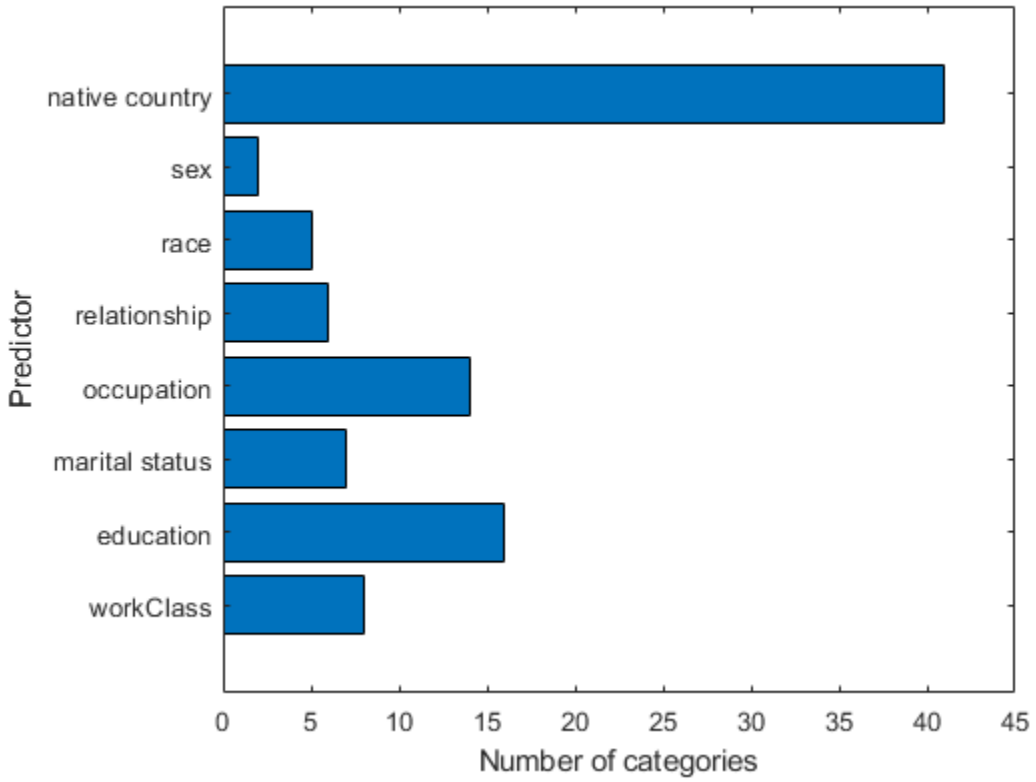
The anonymous function `countNumCats` converts a predictor to a categorical array, then counts the unique, nonempty categories of the predictor.

Use `varfun` and `countNumCats` to count the number of categories for the categorical predictors in `tbl`.

```
numCat = varfun(@(var)countNumCats(var),tbl(:,cat),'OutputFormat','uniform');
```

Plot the number of categories for each categorical predictor.

```
figure
barh(numCat);
h = gca;
h.YTickLabel = VarNames(cat);
ylabel('Predictor')
xlabel('Number of categories')
```



Train Model

For binary classification, the software uses a computational shortcut to find an optimal split for categorical predictors with many categories. For classification with more than two classes, you can choose an exact algorithm or a heuristic algorithm to find a good split by using the `'AlgorithmForCategorical'` name-value pair argument of `fitctree` or `templateTree`. By default, the software selects the optimal subset of algorithms for each split using the known number of classes and levels of a categorical predictor.

Train a classification tree using `tbl` and `Y`. The response vector `Y` has two classes, so the software uses the exact algorithm for categorical predictor splits.

```
Mdl = fitctree(tbl,Y)
```

```
Mdl =
  ClassificationTree
             PredictorNames: {1x14 cell}
               ResponseName: 'Y'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
           ScoreTransform: 'none'
          NumObservations: 32561
```

```
Properties, Methods
```

References

[1] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall, 1984.

[2] Coppersmith, D., S. J. Hong, and J. R. M. Hosking. "Partitioning Nominal Attributes in Decision Trees." *Data Mining and Knowledge Discovery*, Vol. 3, 1999, pp. 197–217.

See Also

`fitcensemble` | `fitctree` | `fitrtree` | `templateTree`

Related Examples

- "Decision Trees" on page 19-2
- "Growing Decision Trees" on page 19-7
- "Ensemble Algorithms" on page 18-39

Discriminant Analysis

Discriminant Analysis Classification

Discriminant analysis is a classification method. It assumes that different classes generate data based on different Gaussian distributions.

- To train (create) a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class (see “Creating Discriminant Analysis Model” on page 20-4).
- To predict the classes of new data, the trained classifier finds the class with the smallest misclassification cost (see “Prediction Using Discriminant Analysis Models” on page 20-6).

Linear discriminant analysis is also known as the Fisher discriminant, named for its inventor, Sir R. A. Fisher [1].

Create Discriminant Analysis Classifiers

This example shows how to train a basic discriminant analysis classifier to classify irises in Fisher's iris data.

Load the data.

```
load fisheriris
```

Create a default (linear) discriminant analysis classifier.

```
MdlLinear = fitcdiscr(meas,species);
```

To visualize the classification boundaries of a 2-D linear classification of the data, see “Create and Visualize Discriminant Analysis Classifier” on page 20-9.

Classify an iris with average measurements.

```
meanmeas = mean(meas);
meanclass = predict(MdlLinear,meanmeas)
```

```
meanclass = 1x1 cell array
    {'versicolor'}
```

Create a quadratic classifier.

```
MdlQuadratic = fitcdiscr(meas,species,'DiscrimType','quadratic');
```

To visualize the classification boundaries of a 2-D quadratic classification of the data, see “Create and Visualize Discriminant Analysis Classifier” on page 20-9.

Classify an iris with average measurements using the quadratic classifier.

```
meanclass2 = predict(MdlQuadratic,meanmeas)
```

```
meanclass2 = 1x1 cell array
    {'versicolor'}
```

References

[1] Fisher, R. A. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics,* Vol. 7, pp. 179–188, 1936. Available at https://digital.library.adelaide.edu.au/dspace/handle/2440/15227.

See Also

Functions
`fitcdiscr`

Objects
`ClassificationDiscriminant`

Related Examples

- “Creating Discriminant Analysis Model” on page 20-4
- “Create and Visualize Discriminant Analysis Classifier” on page 20-9
- “Improving Discriminant Analysis Models” on page 20-15
- “Regularize Discriminant Analysis Classifier” on page 20-21
- “Examine the Gaussian Mixture Assumption” on page 20-27
- “Prediction Using Discriminant Analysis Models” on page 20-6

Creating Discriminant Analysis Model

The model for discriminant analysis is:

- Each class (Y) generates data (X) using a multivariate normal distribution. In other words, the model assumes X has a Gaussian mixture distribution (`gmdistribution`).
  - For linear discriminant analysis, the model has the same covariance matrix for each class; only the means vary.
  - For quadratic discriminant analysis, both means and covariances of each class vary.

Under this modeling assumption, `fitcdiscr` infers the mean and covariance parameters of each class.

- For linear discriminant analysis, it computes the sample mean of each class. Then it computes the sample covariance by first subtracting the sample mean of each class from the observations of that class, and taking the empirical covariance matrix of the result.
- For quadratic discriminant analysis, it computes the sample mean of each class. Then it computes the sample covariances by first subtracting the sample mean of each class from the observations of that class, and taking the empirical covariance matrix of each class.

The `fit` method does not use prior probabilities or costs for fitting.

Weighted Observations

`fitcdiscr` constructs weighted classifiers using the following scheme. Suppose M is an N-by-K class membership matrix:

$M_{nk} = 1$ if observation n is from class k
$M_{nk} = 0$ otherwise.

The estimate of the class mean for unweighted data is

$$\widehat{\mu}_k = \frac{\sum_{n=1}^{N} M_{nk} x_n}{\sum_{n=1}^{N} M_{nk}}.$$

For weighted data with positive weights w_n, the natural generalization is

$$\widehat{\mu}_k = \frac{\sum_{n=1}^{N} M_{nk} w_n x_n}{\sum_{n=1}^{N} M_{nk} w_n}.$$

The unbiased estimate of the pooled-in covariance matrix for unweighted data is

$$\widehat{\Sigma} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} M_{nk} (x_n - \widehat{\mu}_k)(x_n - \widehat{\mu}_k)^T}{N - K}.$$

For quadratic discriminant analysis, `fitcdiscr` uses $K = 1$.

For weighted data, assuming the weights sum to 1, the unbiased estimate of the pooled-in covariance matrix is

$$\widehat{\Sigma} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} M_{nk} w_n (x_n - \widehat{\mu}_k)(x_n - \widehat{\mu}_k)^T}{1 - \sum_{k=1}^{K} \frac{W_k^{(2)}}{W_k}},$$

where

- $W_k = \sum_{n=1}^{N} M_{nk} w_n$ is the sum of the weights for class k.
- $W_k^{(2)} = \sum_{n=1}^{N} M_{nk} w_n^2$ is the sum of squared weights per class.

See Also

Functions
`fitcdiscr`

Objects
`ClassificationDiscriminant` | `gmdistribution`

Related Examples

- “Discriminant Analysis Classification” on page 20-2
- “Examine the Gaussian Mixture Assumption” on page 20-27

Prediction Using Discriminant Analysis Models

`predict` uses three quantities to classify observations: posterior probability on page 20-6, prior probability on page 20-6, and cost on page 20-7.

`predict` classifies so as to minimize the expected classification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{k=1}^{K} \widehat{P}(k|x) C(y|k),$$

where

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(k|x)$ is the posterior probability of class k for observation x.
- $C(y|k)$ is the cost of classifying an observation as y when its true class is k.

The space of `X` values divides into regions where a classification `Y` is a particular value. The regions are separated by straight lines for linear discriminant analysis, and by conic sections (ellipses, hyperbolas, or parabolas) for quadratic discriminant analysis. For a visualization of these regions, see “Create and Visualize Discriminant Analysis Classifier” on page 20-9.

Posterior Probability

The posterior probability that a point x belongs to class k is the product of the prior probability on page 20-6 and the multivariate normal density. The density function of the multivariate normal with 1-by-d mean μ_k and d-by-d covariance Σ_k at a 1-by-d point x is

$$P(x|k) = \frac{1}{\left((2\pi)^d |\Sigma_k|\right)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)\Sigma_k^{-1}(x - \mu_k)^T\right),$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix.

Let $P(k)$ represent the prior probability of class k. Then the posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)},$$

where $P(x)$ is a normalization constant, namely, the sum over k of $P(x|k)P(k)$.

Prior Probability

The prior probability is one of three choices:

- `'uniform'` — The prior probability of class `k` is 1 over the total number of classes.
- `'empirical'` — The prior probability of class `k` is the number of training samples of class `k` divided by the total number of training samples.
- A numeric vector — The prior probability of class `k` is the `j`th element of the `Prior` vector. See `fitcdiscr`.

After creating a classifier `obj`, you can set the prior using dot notation:

```
obj.Prior = v;
```

where `v` is a vector of positive elements representing the frequency with which each element occurs. You do not need to retrain the classifier when you set a new prior.

Cost

There are two costs associated with discriminant analysis classification: the true misclassification cost per class, and the expected misclassification cost per observation.

True Misclassification Cost per Class

`Cost(i,j)` is the cost of classifying an observation into class `j` if its true class is `i`. By default, `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`. In other words, the cost is `0` for correct classification, and `1` for incorrect classification.

You can set any cost matrix you like when creating a classifier. Pass the cost matrix in the `Cost` name-value pair in `fitcdiscr`.

After you create a classifier `obj`, you can set a custom cost using dot notation:

```
obj.Cost = B;
```

`B` is a square matrix of size `K`-by-`K` when there are `K` classes. You do not need to retrain the classifier when you set a new cost.

Expected Misclassification Cost per Observation

Suppose you have `Nobs` observations that you want to classify with a trained discriminant analysis classifier `obj`. Suppose you have `K` classes. You place the observations into a matrix `Xnew` with one observation per row. The command

```
[label,score,cost] = predict(obj,Xnew)
```

returns, among other outputs, a cost matrix of size `Nobs`-by-`K`. Each row of the cost matrix contains the expected (average) cost of classifying the observation into each of the `K` classes. `cost(n,k)` is

$$\sum_{i=1}^{K} \widehat{P}(i|X(n))C(k|i),$$

where

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability on page 20-6 of class i for observation $Xnew(n)$.
- $C(k|i)$ is the cost on page 20-7 of classifying an observation as k when its true class is i.

See Also

Functions
`fitcdiscr` | `predict`

Objects
`ClassificationDiscriminant` | `CompactClassificationDiscriminant`

Related Examples

- “Discriminant Analysis Classification” on page 20-2

Create and Visualize Discriminant Analysis Classifier

This example shows how to perform linear and quadratic classification of Fisher iris data.

Load the sample data.

```
load fisheriris
```

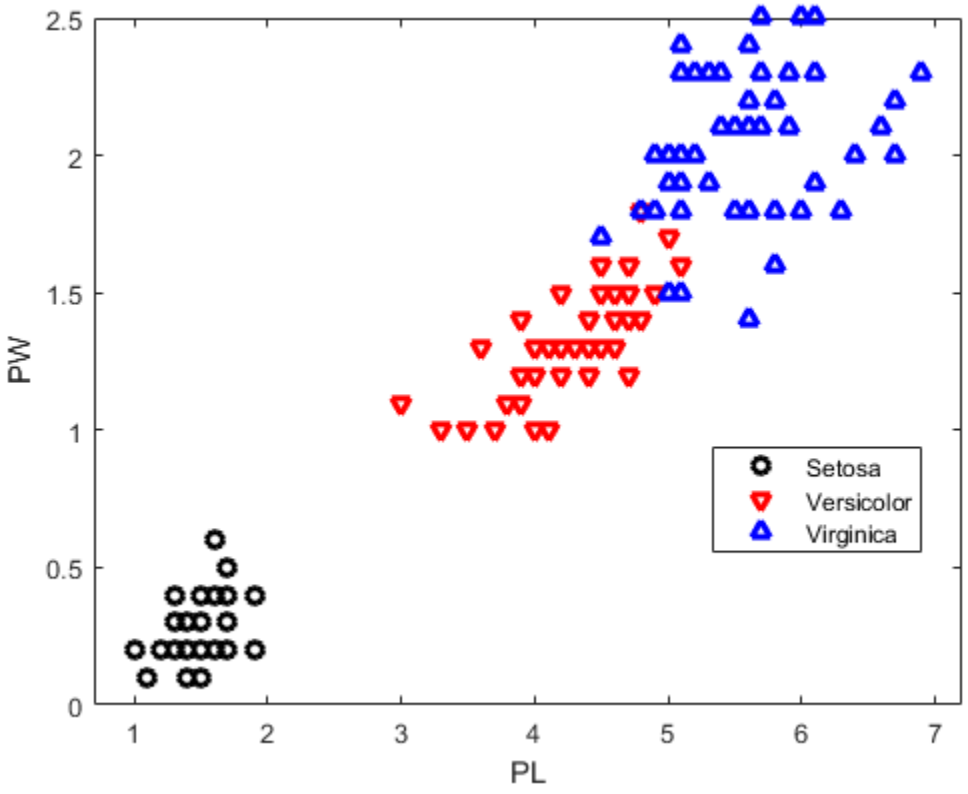
The column vector, `species` , consists of iris flowers of three different species, setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers, the length and width of sepals and petals in centimeters, respectively.

Use petal length (third column in `meas`) and petal width (fourth column in `meas`) measurements. Save these as variables PL and PW, respectively.

```
PL = meas(:,3);
PW = meas(:,4);
```

Plot the data, showing the classification, that is, create a scatter plot of the measurements, grouped by species.

```
h1 = gscatter(PL,PW,species,'krb','ov^',[],'off');
h1(1).LineWidth = 2;
h1(2).LineWidth = 2;
h1(3).LineWidth = 2;
legend('Setosa','Versicolor','Virginica','Location','best')
hold on
```



Create a linear classifier.

```
X = [PL,PW];
MdlLinear = fitcdiscr(X,species);
```

Retrieve the coefficients for the linear boundary between the second and third classes.

```
MdlLinear.ClassNames([2 3])
```

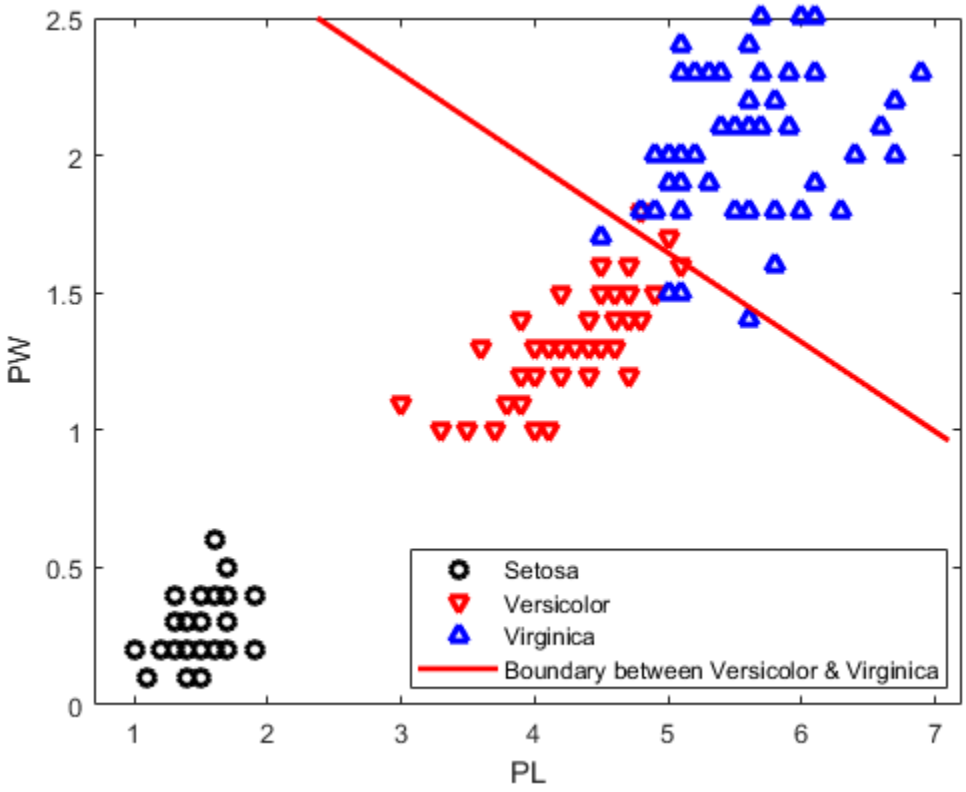
```
ans = 2x1 cell
    {'versicolor'}
    {'virginica' }
```

```
K = MdlLinear.Coeffs(2,3).Const;
L = MdlLinear.Coeffs(2,3).Linear;
```

Plot the curve that separates the second and third classes

$$K + [x_1\ x_2]L = 0\,.$$

```
f = @(x1,x2) K + L(1)*x1 + L(2)*x2;
h2 = fimplicit(f,[.9 7.1 0 2.5]);
h2.Color = 'r';
h2.LineWidth = 2;
h2.DisplayName = 'Boundary between Versicolor & Virginica';
```



Retrieve the coefficients for the linear boundary between the first and second classes.

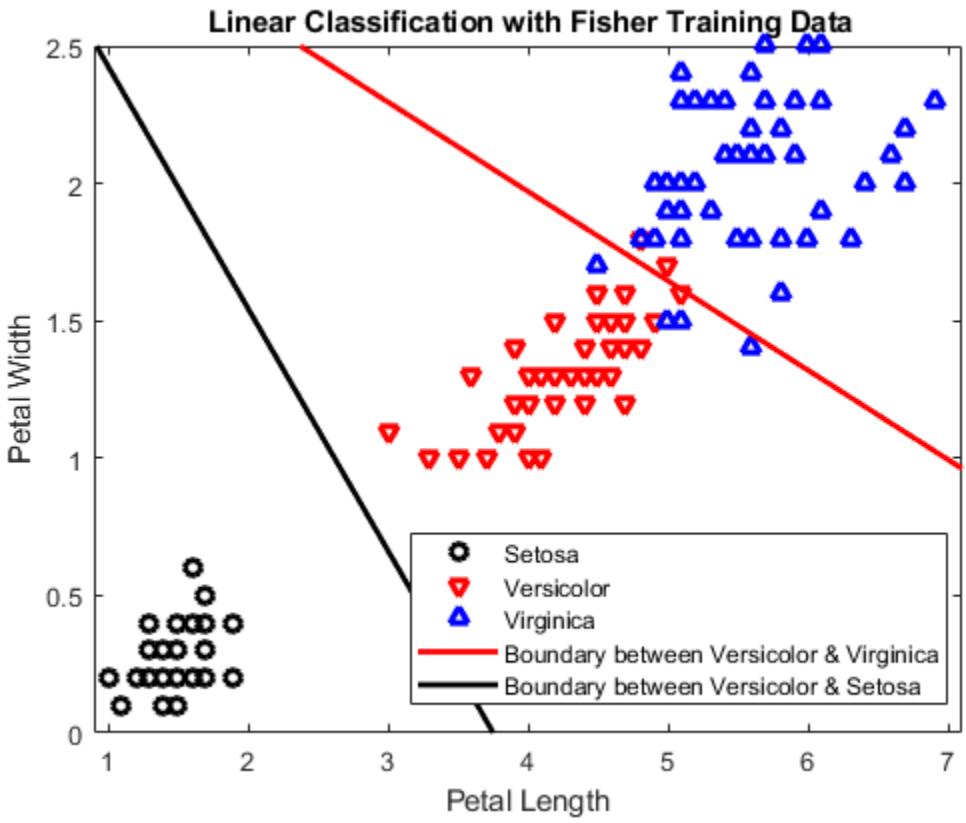
```
MdlLinear.ClassNames([1 2])
```

```
ans = 2x1 cell
    {'setosa'    }
    {'versicolor'}
```

```
K = MdlLinear.Coeffs(1,2).Const;
L = MdlLinear.Coeffs(1,2).Linear;
```

Plot the curve that separates the first and second classes.

```
f = @(x1,x2) K + L(1)*x1 + L(2)*x2;
h3 = fimplicit(f,[.9 7.1 0 2.5]);
h3.Color = 'k';
h3.LineWidth = 2;
h3.DisplayName = 'Boundary between Versicolor & Setosa';
axis([.9 7.1 0 2.5])
xlabel('Petal Length')
ylabel('Petal Width')
title('{\bf Linear Classification with Fisher Training Data}')
```

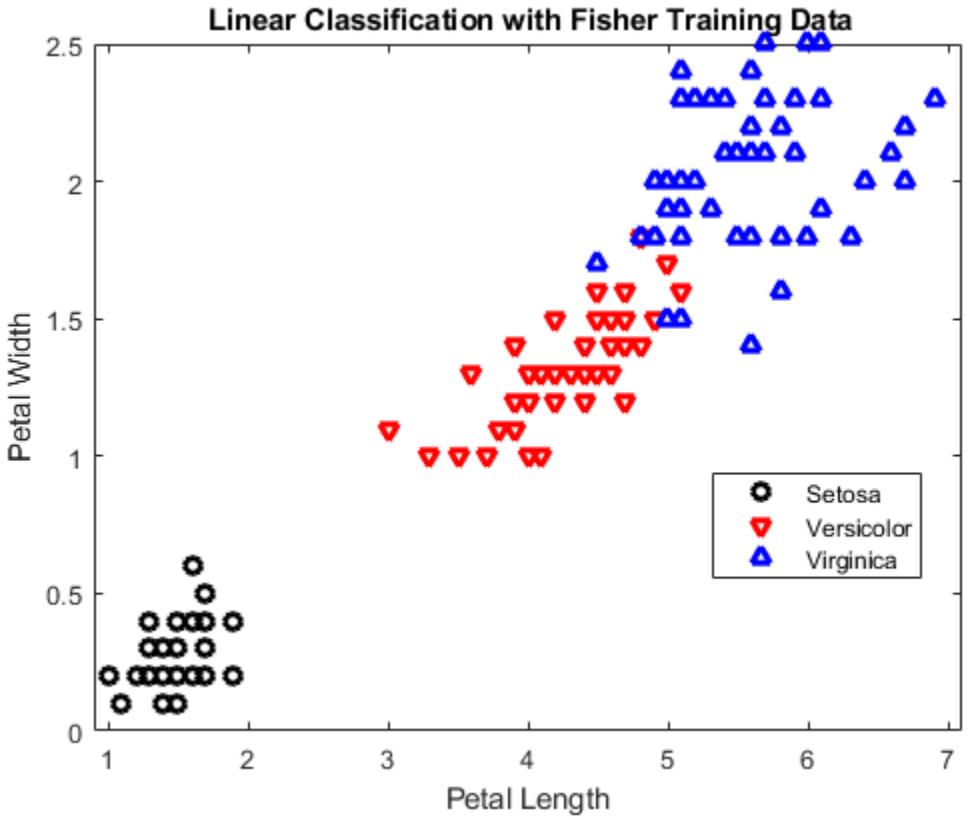


Create a quadratic discriminant classifier.

```
MdlQuadratic = fitcdiscr(X,species,'DiscrimType','quadratic');
```

Remove the linear boundaries from the plot.

```
delete(h2);
delete(h3);
```



Retrieve the coefficients for the quadratic boundary between the second and third classes.

```
MdlQuadratic.ClassNames([2 3])
```

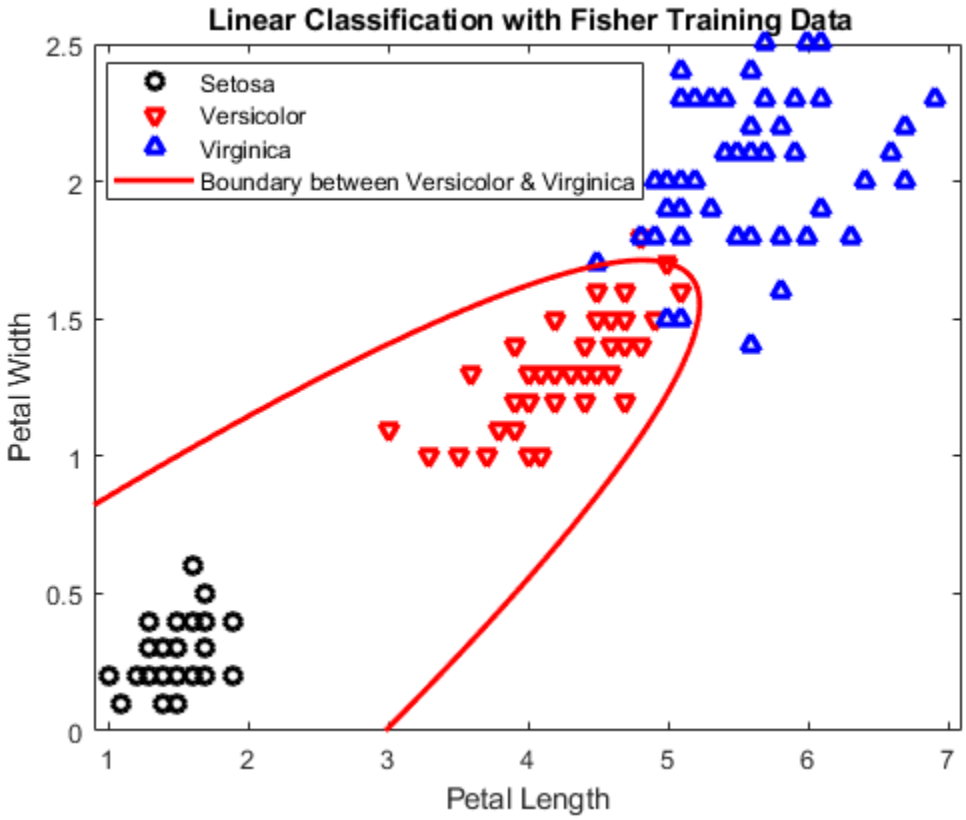
```
ans = 2x1 cell
    {'versicolor'}
    {'virginica' }
```

```
K = MdlQuadratic.Coeffs(2,3).Const;
L = MdlQuadratic.Coeffs(2,3).Linear;
Q = MdlQuadratic.Coeffs(2,3).Quadratic;
```

Plot the curve that separates the second and third classes

$$K + [x_1\ x_2]L + [x_1\ x_2]Q\begin{bmatrix}x_1\\x_2\end{bmatrix} = 0\,.$$

```
f = @(x1,x2) K + L(1)*x1 + L(2)*x2 + Q(1,1)*x1.^2 + ...
    (Q(1,2)+Q(2,1))*x1.*x2 + Q(2,2)*x2.^2;
h2 = fimplicit(f,[.9 7.1 0 2.5]);
h2.Color = 'r';
h2.LineWidth = 2;
h2.DisplayName = 'Boundary between Versicolor & Virginica';
```



Retrieve the coefficients for the quadratic boundary between the first and second classes.

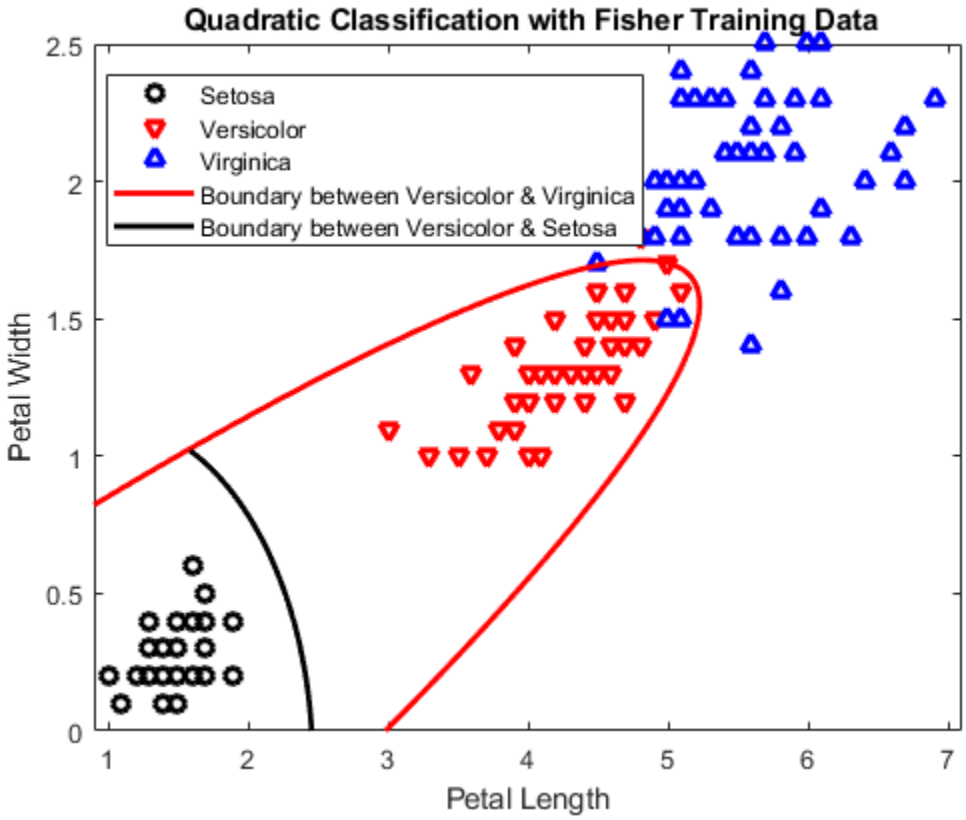
```
MdlQuadratic.ClassNames([1 2])
```

```
ans = 2x1 cell
    {'setosa'    }
    {'versicolor'}
```

```
K = MdlQuadratic.Coeffs(1,2).Const;
L = MdlQuadratic.Coeffs(1,2).Linear;
Q = MdlQuadratic.Coeffs(1,2).Quadratic;
```

Plot the curve that separates the first and second and classes.

```
f = @(x1,x2) K + L(1)*x1 + L(2)*x2 + Q(1,1)*x1.^2 + ...
    (Q(1,2)+Q(2,1))*x1.*x2 + Q(2,2)*x2.^2;
h3 = fimplicit(f,[.9 7.1 0 1.02]); % Plot the relevant portion of the curve.
h3.Color = 'k';
h3.LineWidth = 2;
h3.DisplayName = 'Boundary between Versicolor & Setosa';
axis([.9 7.1 0 2.5])
xlabel('Petal Length')
ylabel('Petal Width')
title('{\bf Quadratic Classification with Fisher Training Data}')
hold off
```



See Also

Functions
`fitcdiscr`

Objects
`ClassificationDiscriminant`

Related Examples

- “Discriminant Analysis Classification” on page 20-2

Improving Discriminant Analysis Models

In this section...

Deal with Singular Data

Discriminant analysis needs data sufficient to fit Gaussian models with invertible covariance matrices. If your data is not sufficient to fit such a model uniquely, `fitcdiscr` fails. This section shows methods for handling failures.

Tip To obtain a discriminant analysis classifier without failure, set the `DiscrimType` name-value pair to `'pseudoLinear'` or `'pseudoQuadratic'` in `fitcdiscr`.

“Pseudo” discriminants never fail, because they use the pseudoinverse of the covariance matrix Σ_k (see `pinv`).

Example: Singular Covariance Matrix

When the covariance matrix of the fitted classifier is singular, `fitcdiscr` can fail:

```
load popcorn
X = popcorn(:,[1 2]);
X(:,3) = 0; % a zero-variance column
Y = popcorn(:,3);
ppcrn = fitcdiscr(X,Y);

Error using ClassificationDiscriminant (line 635)
Predictor x3 has zero variance. Either exclude this predictor or set 'discrimType' to
'pseudoLinear' or 'diagLinear'.

Error in classreg.learning.FitTemplate/fit (line 243)
            obj = this.MakeFitObject(X,Y,W,this.ModelParameters,fitArgs{:});

Error in fitcdiscr (line 296)
            this = fit(temp,X,Y);
```

To proceed with linear discriminant analysis, use a `pseudoLinear` or `diagLinear` discriminant type:

```
ppcrn = fitcdiscr(X,Y,...
    'discrimType','pseudoLinear');
meanpredict = predict(ppcrn,mean(X))

meanpredict =
    3.5000
```

Choose a Discriminant Type

There are six types of discriminant analysis classifiers: linear and quadratic, with diagonal and pseudo variants of each type.

Tip To see if your covariance matrix is singular, set `discrimType` to `'linear'` or `'quadratic'`. If the matrix is singular, the `fitcdiscr` method fails for `'quadratic'`, and the `Gamma` property is nonzero for `'linear'`.

To obtain a quadratic classifier even when your covariance matrix is singular, set `DiscrimType` to `'pseudoQuadratic'` or `'diagQuadratic'`.

```
obj = fitcdiscr(X,Y,'DiscrimType','pseudoQuadratic') % or 'diagQuadratic'
```

Choose a classifier type by setting the `discrimType` name-value pair to one of:

- `'linear'` (default) — Estimate one covariance matrix for all classes.
- `'quadratic'` — Estimate one covariance matrix for each class.
- `'diagLinear'` — Use the diagonal of the `'linear'` covariance matrix, and use its pseudoinverse if necessary.
- `'diagQuadratic'` — Use the diagonals of the `'quadratic'` covariance matrices, and use their pseudoinverses if necessary.
- `'pseudoLinear'` — Use the pseudoinverse of the `'linear'` covariance matrix if necessary.
- `'pseudoQuadratic'` — Use the pseudoinverses of the `'quadratic'` covariance matrices if necessary.

`fitcdiscr` can fail for the `'linear'` and `'quadratic'` classifiers. When it fails, it returns an explanation, as shown in “Deal with Singular Data” on page 20-15.

`fitcdiscr` always succeeds with the diagonal and pseudo variants. For information about pseudoinverses, see `pinv`.

You can set the discriminant type using dot notation after constructing a classifier:

```
obj.DiscrimType = 'discrimType'
```

You can change between linear types or between quadratic types, but cannot change between a linear and a quadratic type.

Examine the Resubstitution Error and Confusion Matrix

The resubstitution error is the difference between the response training data and the predictions the classifier makes of the response based on the input training data. If the resubstitution error is high, you cannot expect the predictions of the classifier to be good. However, having low resubstitution error does not guarantee good predictions for new data. Resubstitution error is often an overly optimistic estimate of the predictive error on new data.

The confusion matrix shows how many errors, and which types, arise in resubstitution. When there are `K` classes, the confusion matrix `R` is a `K`-by-`K` matrix with

`R(i,j)` = the number of observations of class `i` that the classifier predicts to be of class `j`.

Example: Resubstitution Error of a Discriminant Analysis Classifier

Examine the resubstitution error of the default discriminant analysis classifier for the Fisher iris data:

```
load fisheriris
obj = fitcdiscr(meas,species);
resuberror = resubLoss(obj)

resuberror =
    0.0200
```

The resubstitution error is very low, meaning `obj` classifies nearly all the Fisher iris data correctly. The total number of misclassifications is:

```
resuberror * obj.NumObservations

ans =
    3.0000
```

To see the details of the three misclassifications, examine the confusion matrix:

```
R = confusionmat(obj.Y,resubPredict(obj))

R =
    50     0     0
     0    48     2
     0     1    49

obj.ClassNames

ans =
    'setosa'
    'versicolor'
    'virginica'
```

- `R(1,:) = [50 0 0]` means `obj` classifies all 50 setosa irises correctly.
- `R(2,:) = [0 48 2]` means `obj` classifies 48 versicolor irises correctly, and misclassifies two versicolor irises as virginica.
- `R(3,:) = [0 1 49]` means `obj` classifies 49 virginica irises correctly, and misclassifies one virginica iris as versicolor.

Cross Validation

Typically, discriminant analysis classifiers are robust and do not exhibit overtraining when the number of predictors is much less than the number of observations. Nevertheless, it is good practice to cross validate your classifier to ensure its stability.

Cross Validating a Discriminant Analysis Classifier

This example shows how to perform five-fold cross validation of a quadratic discriminant analysis classifier.

Load the sample data.

```
load fisheriris
```

Create a quadratic discriminant analysis classifier for the data.

```
quadisc = fitcdiscr(meas,species,'DiscrimType','quadratic');
```

Find the resubstitution error of the classifier.

```
qerror = resubLoss(quadisc)
```

```
qerror = 0.0200
```

The classifier does an excellent job. Nevertheless, resubstitution error can be an optimistic estimate of the error when classifying new data. So proceed to cross validation.

Create a cross-validation model.

```
cvmodel = crossval(quadisc,'kfold',5);
```

Find the cross-validation loss for the model, meaning the error of the out-of-fold observations.

```
cverror = kfoldLoss(cvmodel)
```

```
cverror = 0.0200
```

The cross-validated loss is as low as the original resubstitution loss. Therefore, you can have confidence that the classifier is reasonably accurate.

Change Costs and Priors

Sometimes you want to avoid certain misclassification errors more than others. For example, it might be better to have oversensitive cancer detection instead of undersensitive cancer detection. Oversensitive detection gives more false positives (unnecessary testing or treatment). Undersensitive detection gives more false negatives (preventable illnesses or deaths). The consequences of underdetection can be high. Therefore, you might want to set costs to reflect the consequences.

Similarly, the training data `Y` can have a distribution of classes that does not represent their true frequency. If you have a better estimate of the true frequency, you can include this knowledge in the classification `Prior` property.

Example: Setting Custom Misclassification Costs

Consider the Fisher iris data. Suppose that the cost of classifying a versicolor iris as virginica is 10 times as large as making any other classification error. Create a classifier from the data, then incorporate this cost and then view the resulting classifier.

1 Load the Fisher iris data and create a default (linear) classifier as in “Example: Resubstitution Error of a Discriminant Analysis Classifier” on page 20-16:

```
load fisheriris
obj = fitcdiscr(meas,species);
resuberror = resubLoss(obj)

resuberror =
    0.0200

R = confusionmat(obj.Y,resubPredict(obj))

R =
    50     0     0
     0    48     2
     0     1    49
```

```
obj.ClassNames

ans =
    'setosa'
    'versicolor'
    'virginica'
```

`R(2,:) = [0 48 2]` means `obj` classifies 48 versicolor irises correctly, and misclassifies two versicolor irises as virginica.

2 Change the cost matrix to make fewer mistakes in classifying versicolor irises as virginica:

```
obj.Cost(2,3) = 10;
R2 = confusionmat(obj.Y,resubPredict(obj))

R2 =
    50     0     0
     0    50     0
     0     7    43
```

`obj` now classifies all versicolor irises correctly, at the expense of increasing the number of misclassifications of virginica irises from 1 to 7.

Example: Setting Alternative Priors

Consider the Fisher iris data. There are 50 irises of each kind in the data. Suppose that, in a particular region, you have historical data that shows virginica are five times as prevalent as the other kinds. Create a classifier that incorporates this information.

1 Load the Fisher iris data and make a default (linear) classifier as in “Example: Resubstitution Error of a Discriminant Analysis Classifier” on page 20-16:

```
load fisheriris
obj = fitcdiscr(meas,species);
resuberror = resubLoss(obj)

resuberror =
    0.0200

R = confusionmat(obj.Y,resubPredict(obj))

R =
    50     0     0
     0    48     2
     0     1    49

obj.ClassNames

ans =
    'setosa'
    'versicolor'
    'virginica'
```

`R(3,:) = [0 1 49]` means `obj` classifies 49 virginica irises correctly, and misclassifies one virginica iris as versicolor.

2 Change the prior to match your historical data, and examine the confusion matrix of the new classifier:

```
obj.Prior = [1 1 5];
R2 = confusionmat(obj.Y,resubPredict(obj))

R2 =
    50     0     0
     0    46     4
     0     0    50
```

The new classifier classifies all virginica irises correctly, at the expense of increasing the number of misclassifications of versicolor irises from 2 to 4.

See Also

Functions
`crossval` | `cvshrink` | `fitcdiscr` | `logp` | `loss` | `predict` | `resubLoss`

Objects
`ClassificationDiscriminant`

Related Examples

- “Discriminant Analysis Classification” on page 20-2
- “Regularize Discriminant Analysis Classifier” on page 20-21

Regularize Discriminant Analysis Classifier

This example shows how to make a more robust and simpler model by trying to remove predictors without hurting the predictive power of the model. This is especially important when you have many predictors in your data. Linear discriminant analysis uses the two regularization parameters, “Gamma and Delta” on page 33-1124, to identify and remove redundant predictors. The `cvshrink` method helps identify appropriate settings for these parameters.

Load data and create a classifier.

Create a linear discriminant analysis classifier for the `ovariancancer` data. Set the `SaveMemory` and `FillCoeffs` name-value pair arguments to keep the resulting model reasonably small. For computational ease, this example uses a random subset of about one third of the predictors to train the classifier.

```
load ovariancancer
rng(1); % For reproducibility
numPred = size(obs,2);
obs = obs(:,randsample(numPred,ceil(numPred/3)));
Mdl = fitcdiscr(obs,grp,'SaveMemory','on','FillCoeffs','off');
```

Cross validate the classifier.

Use 25 levels of `Gamma` and 25 levels of `Delta` to search for good parameters. This search is time consuming. Set `Verbose` to `1` to view the progress.

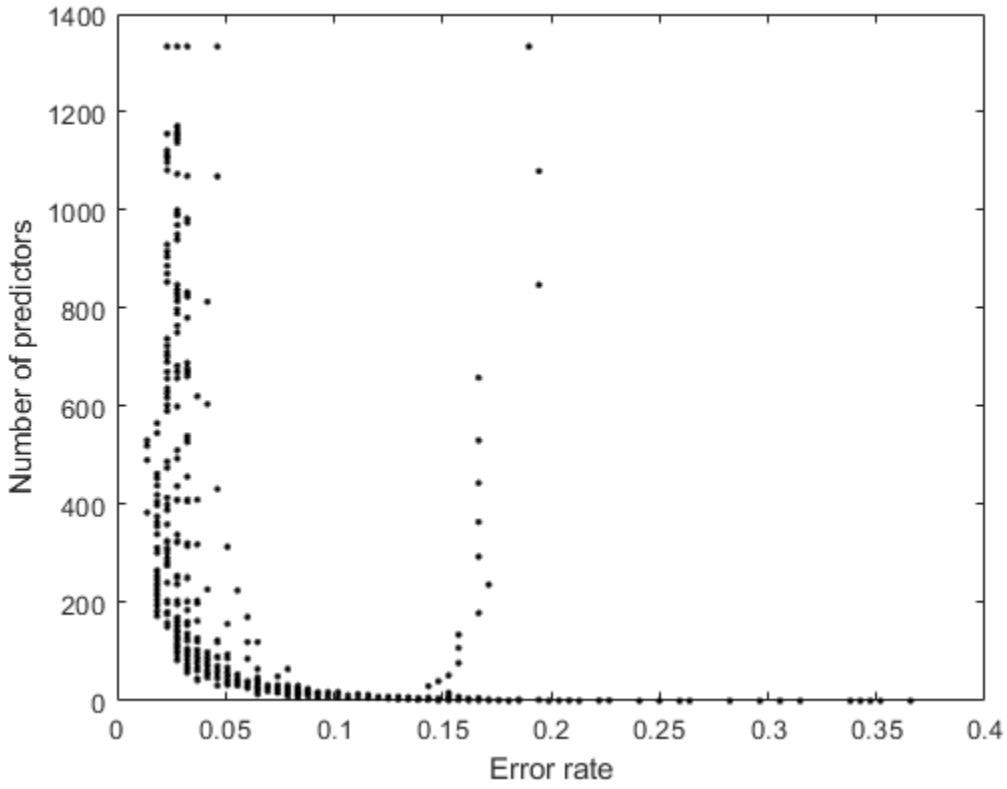
```
[err,gamma,delta,numpred] = cvshrink(Mdl,...
    'NumGamma',24,'NumDelta',24,'Verbose',1);
```

```
Done building cross-validated model.
Processing Gamma step 1 out of 25.
Processing Gamma step 2 out of 25.
Processing Gamma step 3 out of 25.
Processing Gamma step 4 out of 25.
Processing Gamma step 5 out of 25.
Processing Gamma step 6 out of 25.
Processing Gamma step 7 out of 25.
Processing Gamma step 8 out of 25.
Processing Gamma step 9 out of 25.
Processing Gamma step 10 out of 25.
Processing Gamma step 11 out of 25.
Processing Gamma step 12 out of 25.
Processing Gamma step 13 out of 25.
Processing Gamma step 14 out of 25.
Processing Gamma step 15 out of 25.
Processing Gamma step 16 out of 25.
Processing Gamma step 17 out of 25.
Processing Gamma step 18 out of 25.
Processing Gamma step 19 out of 25.
Processing Gamma step 20 out of 25.
Processing Gamma step 21 out of 25.
Processing Gamma step 22 out of 25.
Processing Gamma step 23 out of 25.
Processing Gamma step 24 out of 25.
Processing Gamma step 25 out of 25.
```

Examine the quality of the regularized classifiers.

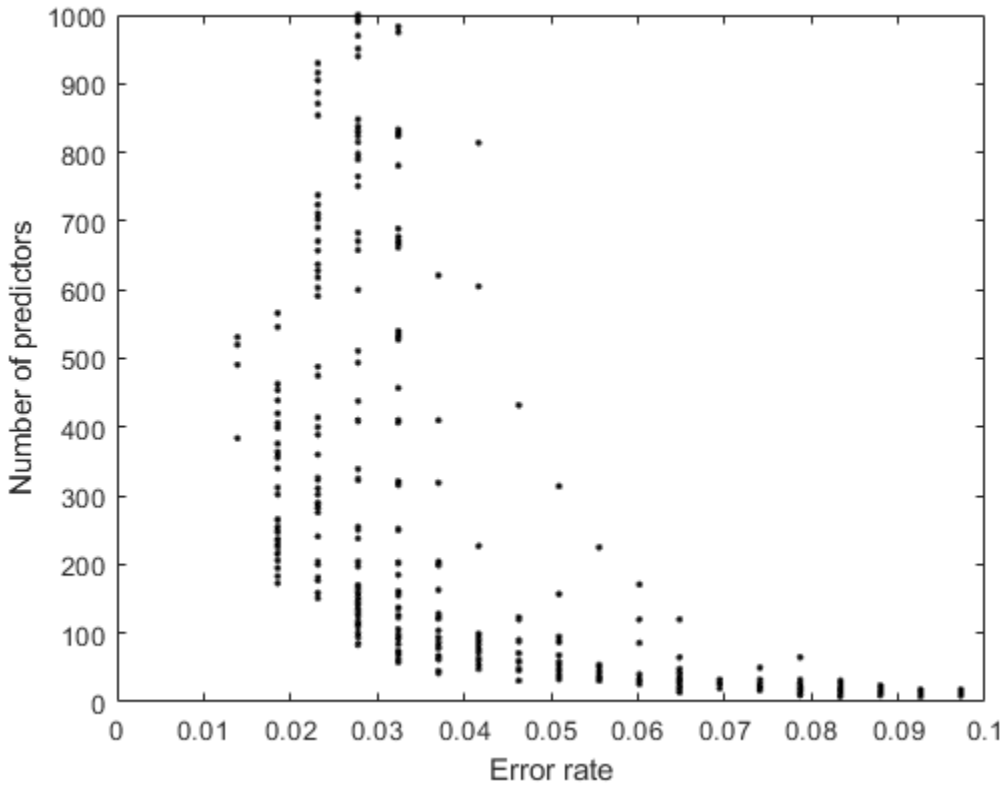
Plot the number of predictors against the error.

```
plot(err,numpred,'k.')
xlabel('Error rate')
ylabel('Number of predictors')
```



Examine the lower-left part of the plot more closely.

```
axis([0 .1 0 1000])
```



There is a clear tradeoff between lower number of predictors and lower error.

Choose an optimal tradeoff between model size and accuracy.

Multiple pairs of `Gamma` and `Delta` values produce about the same minimal error. Display the indices of these pairs and their values.

First, find the minimal error value.

```
minerr = min(min(err))
```

```
minerr = 0.0139
```

Find the subscripts of `err` producing minimal error.

```
[p,q] = find(err < minerr + 1e-4);
```

Convert from subscripts to linear indices.

```
idx = sub2ind(size(delta),p,q);
```

Display the `Gamma` and `Delta` values.

```
[gamma(p) delta(idx)]
```

```
ans = 4×2

    0.7202    0.1145
```

```
    0.7602    0.1131
    0.8001    0.1128
    0.8001    0.1410
```

These points have as few as 29% of the total predictors with nonzero coefficients in the model.

```
numpred(idx)/ceil(numPred/3)*100
```

```
ans = 4×1

   39.8051
   38.9805
   36.8066
   28.7856
```

To further lower the number of predictors, you must accept larger error rates. For example, to choose the `Gamma` and `Delta` that give the lowest error rate with 200 or fewer predictors.

```
low200 = min(min(err(numpred <= 200)));
lownum = min(min(numpred(err == low200)));
[low200 lownum]
```

```
ans = 1×2

    0.0185  173.0000
```

You need 173 predictors to achieve an error rate of 0.0185, and this is the lowest error rate among those that have 200 predictors or fewer.

Display the `Gamma` and `Delta` that achieve this error/number of predictors.

```
[r,s] = find((err == low200) & (numpred == lownum));
[gamma(r); delta(r,s)]
```

```
ans = 2×1

    0.6403
    0.2399
```

Set the regularization parameters.

To set the classifier with these values of `Gamma` and `Delta`, use dot notation.

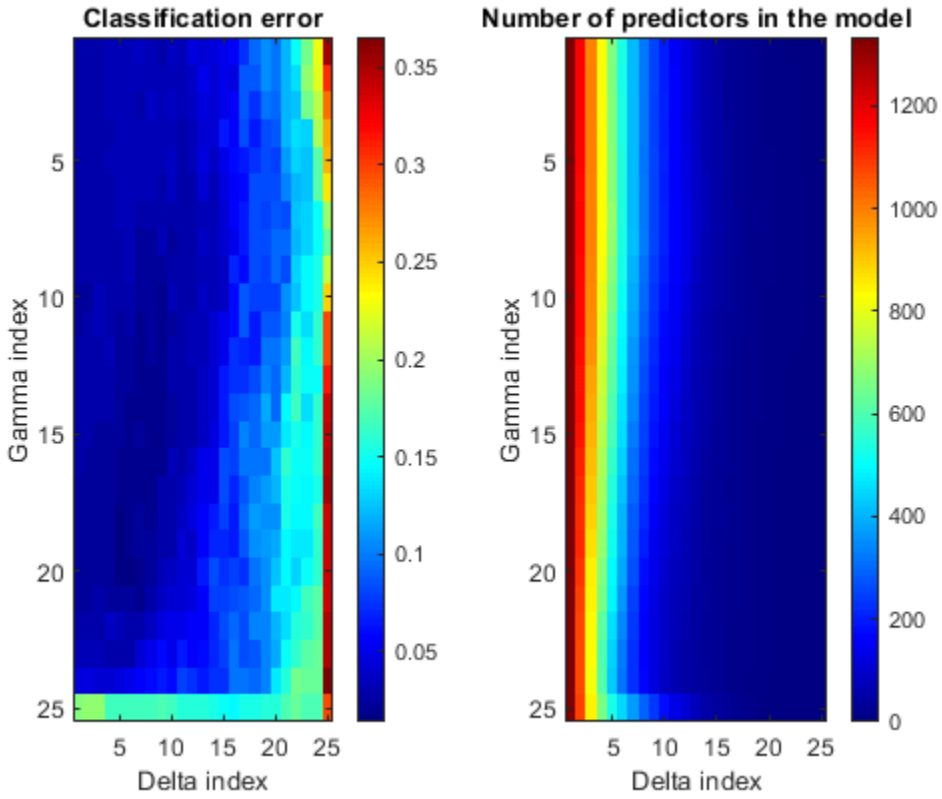
```
Mdl.Gamma = gamma(r);
Mdl.Delta = delta(r,s);
```

Heatmap plot

To compare the `cvshrink` calculation to that in Guo, Hastie, and Tibshirani [1], plot heatmaps of error and number of predictors against `Gamma` and the index of the `Delta` parameter. (The `Delta` parameter range depends on the value of the `Gamma` parameter. So to get a rectangular plot, use the `Delta` index, not the parameter itself.)

```
% Create the Delta index matrix
indx = repmat(1:size(delta,2),size(delta,1),1);
```

```
figure
subplot(1,2,1)
imagesc(err)
colorbar
colormap('jet')
title('Classification error')
xlabel('Delta index')
ylabel('Gamma index')

subplot(1,2,2)
imagesc(numpred)
colorbar
title('Number of predictors in the model')
xlabel('Delta index')
ylabel('Gamma index')
```



You see the best classification error when `Delta` is small, but fewest predictors when `Delta` is large.

References

[1] Guo, Y., T. Hastie, and R. Tibshirani. "Regularized Discriminant Analysis and Its Application in Microarray." *Biostatistics,* Vol. 8, No. 1, pp. 86–100, 2007.

See Also

Functions
`cvshrink` | `fitcdiscr`

Objects
`ClassificationDiscriminant`

Related Examples

- “Discriminant Analysis Classification” on page 20-2
- “Improving Discriminant Analysis Models” on page 20-15
- “Introduction to Feature Selection” on page 15-49
- “Interpret Machine Learning Models” on page 18-256

Examine the Gaussian Mixture Assumption

Discriminant analysis assumes that the data comes from a Gaussian mixture model (see “Creating Discriminant Analysis Model” on page 20-4). If the data appears to come from a Gaussian mixture model, you can expect discriminant analysis to be a good classifier. Furthermore, the default linear discriminant analysis assumes that all class covariance matrices are equal. This section shows methods to check these assumptions:

In this section...
“Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis” on page 20-27
“Q-Q Plot” on page 20-29
“Mardia Kurtosis Test of Multivariate Normality” on page 20-31

Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis

The Bartlett test (see Box [1]) checks equality of the covariance matrices of the various classes. If the covariance matrices are equal, the test indicates that linear discriminant analysis is appropriate. If not, consider using quadratic discriminant analysis, setting the `DiscrimType` name-value pair argument to `'quadratic'` in `fitcdiscr`.

The Bartlett test assumes normal (Gaussian) samples, where neither the means nor covariance matrices are known. To determine whether the covariances are equal, compute the following quantities:

- Sample covariance matrices per class Σ_i, $1 \le i \le k$, where k is the number of classes.
- Pooled-in covariance matrix Σ.
- Test statistic V:

  $$V = (n-k)\log(|\Sigma|) - \sum_{i=1}^{k}(n_i-1)\log(|\Sigma_i|)$$

  where n is the total number of observations, n_i is the number of observations in class i, and $|\Sigma|$ means the determinant of the matrix Σ.
- Asymptotically, as the number of observations in each class n_i becomes large, V is distributed approximately χ^2 with $kd(d+1)/2$ degrees of freedom, where d is the number of predictors (number of dimensions in the data).

The Bartlett test is to check whether V exceeds a given percentile of the χ^2 distribution with $kd(d+1)/2$ degrees of freedom. If it does, then reject the hypothesis that the covariances are equal.

Bartlett Test of Equal Covariance Matrices

Check whether the Fisher iris data is well modeled by a single Gaussian covariance, or whether it is better modeled as a Gaussian mixture by performing a Bartlett test of equal covariance matrices.

Load the `fisheriris` data set.

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
```

When all the class covariance matrices are equal, a linear discriminant analysis is appropriate.

Train a linear discriminant analysis model (the default type) using the Fisher iris data.

```
L = fitcdiscr(meas,species,'PredictorNames',prednames);
```

When the class covariance matrices are not equal, a quadratic discriminant analysis is appropriate.

Train a quadratic discriminant analysis model using the Fisher iris data and compute statistics

```
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
```

Store as variables the number of observations `N`, dimension of the data set `D`, number of classes `K`, and number of observations in each class `Nclass`.

```
[N,D] = size(meas)
```

```
N = 150
```

```
D = 4
```

```
K = numel(unique(species))
```

```
K = 3
```

```
Nclass = grpstats(meas(:,1),species,'numel')'
```

```
Nclass = 1×3

    50    50    50
```

Compute the test statistic `V`.

```
SigmaL = L.Sigma;
SigmaQ = Q.Sigma;
V = (N-K)*log(det(SigmaL));
for k=1:K
    V = V - (Nclass(k)-1)*log(det(SigmaQ(:,:,k)));
end
V
```

```
V = 146.6632
```

Compute the *p*-value.

```
nu = K*D*(D+1)/2;
pval1 = chi2cdf(V,nu,'upper')
```

```
pval1 = 2.6091e-17
```

Because `pval1` is smaller than 0.05, the Bartlett test rejects the hypothesis of equal covariance matrices. The result indicates to use quadratic discriminant analysis, as opposed to linear discriminant analysis.

Q-Q Plot

A Q-Q plot graphically shows whether an empirical distribution is close to a theoretical distribution. If the two are equal, the Q-Q plot lies on a 45° line. If not, the Q-Q plot strays from the 45° line.

Compare Q-Q Plots for Linear and Quadratic Discriminants

Analyze the Q-Q plots to check whether the Fisher iris data is better modeled by a single Gaussian covariance or as a Gaussian mixture.

Load the `fisheriris` data set.

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
```

When all the class covariance matrices are equal, a linear discriminant analysis is appropriate.

Train a linear discriminant analysis model.

```
L = fitcdiscr(meas,species,'PredictorNames',prednames);
```

When the class covariance matrices are not equal, a quadratic discriminant analysis is appropriate.

Train a quadratic discriminant analysis model using the Fisher iris data.

```
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
```

Compute the number of observations, dimension of the data set, and expected quantiles.

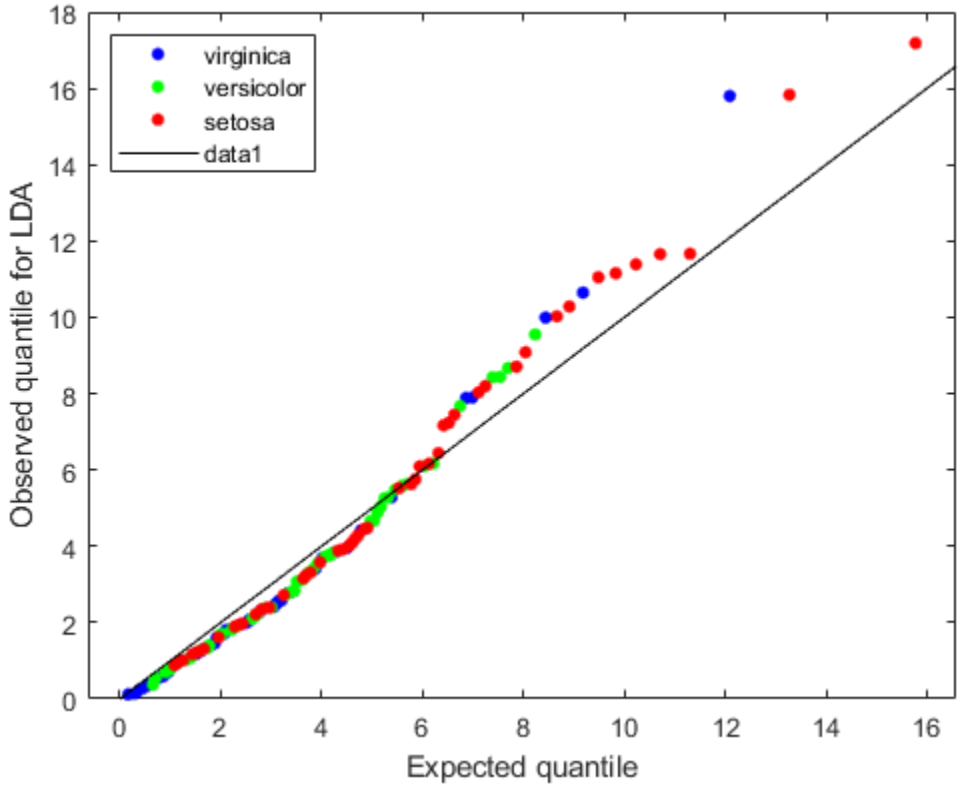
```
[N,D] = size(meas);
expQuant = chi2inv(((1:N)-0.5)/N,D);
```

Compute the observed quantiles for the linear discriminant model.

```
obsL = mahal(L,L.X,'ClassLabels',L.Y);
[obsL,sortedL] = sort(obsL);
```

Graph the Q-Q plot for the linear discriminant.

```
figure;
gscatter(expQuant,obsL,L.Y(sortedL),'bgr',[],[],'off');
legend('virginica','versicolor','setosa','Location','NW');
xlabel('Expected quantile');
ylabel('Observed quantile for LDA');
line([0 20],[0 20],'color','k');
```



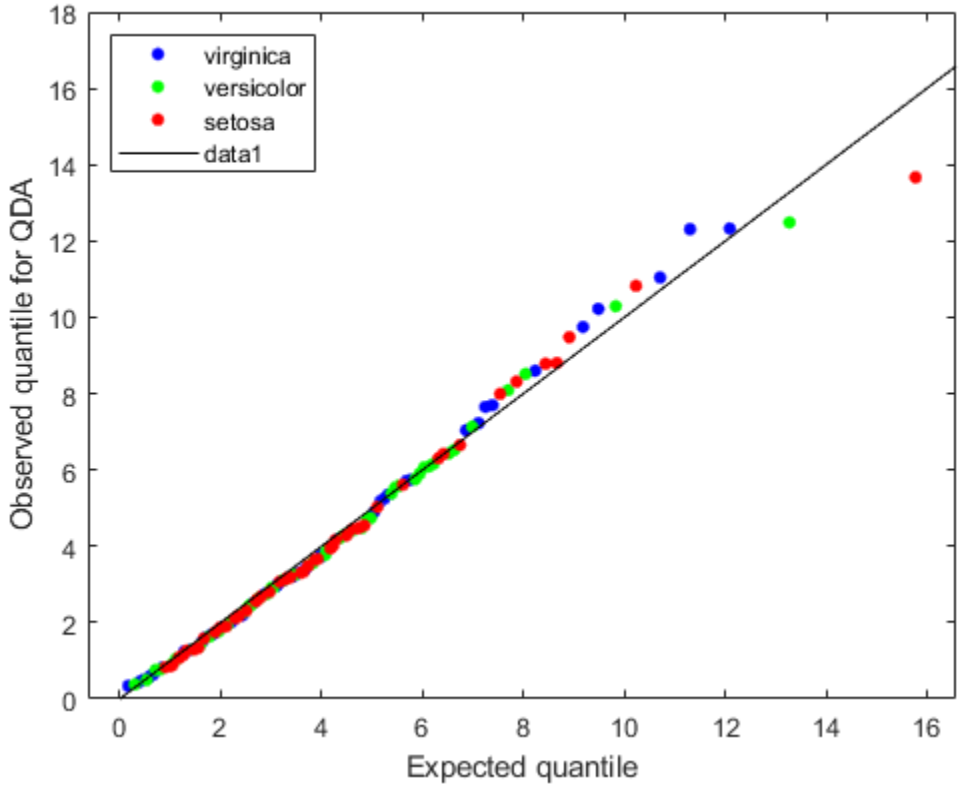
The expected and observed quantiles agree somewhat. The deviation of the plot from the 45° line upward indicates that the data has heavier tails than a normal distribution. The plot shows three possible outliers at the top: two observations from class `'setosa'` and one observation from class `'virginica'`.

Compute the observed quantiles for the quadratic discriminant model.

```
obsQ = mahal(Q,Q.X,'ClassLabels',Q.Y);
[obsQ,sortedQ] = sort(obsQ);
```

Graph the Q-Q plot for the quadratic discriminant.

```
figure;
gscatter(expQuant,obsQ,Q.Y(sortedQ),'bgr',[],[],'off');
legend('virginica','versicolor','setosa','Location','NW');
xlabel('Expected quantile');
ylabel('Observed quantile for QDA');
line([0 20],[0 20],'color','k');
```



The Q-Q plot for the quadratic discriminant shows a better agreement between the observed and expected quantiles. The plot shows only one possible outlier from class `'setosa'`. The Fisher iris data is better modeled as a Gaussian mixture with covariance matrices that are not required to be equal across classes.

Mardia Kurtosis Test of Multivariate Normality

The Mardia kurtosis test (see Mardia [2]) is an alternative to examining a Q-Q plot. It gives a numeric approach to deciding if data matches a Gaussian mixture model.

In the Mardia kurtosis test you compute M, the mean of the fourth power of the Mahalanobis distance of the data from the class means. If the data is normally distributed with a constant covariance matrix (and is thus suitable for linear discriminant analysis), M is asymptotically distributed as normal with mean $d(d + 2)$ and variance $8d(d + 2)/n$, where

- d is the number of predictors (number of dimensions in the data).
- n is the total number of observations.

The Mardia test is two sided: check whether M is close enough to $d(d + 2)$ with respect to a normal distribution of variance $8d(d + 2)/n$.

Mardia Kurtosis Test for Linear and Quadratic Discriminants

Perform a Mardia kurtosis tests to check whether the Fisher iris data is approximately normally distributed for both linear and quadratic discriminant analyses.

Load the `fisheriris` data set.

```
load fisheriris;
prednames = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
```

When all the class covariance matrices are equal, a linear discriminant analysis is appropriate.

Train a linear discriminant analysis model.

```
L = fitcdiscr(meas,species,'PredictorNames',prednames);
```

When the class covariance matrices are not equal, a quadratic discriminant analysis is appropriate.

Train a quadratic discriminant analysis model using the Fisher iris data.

```
Q = fitcdiscr(meas,species,'PredictorNames',prednames,'DiscrimType','quadratic');
```

Compute the mean and variance of the asymptotic distribution.

```
[N,D] = size(meas);
meanKurt = D*(D+2)
```

```
meanKurt = 24
```

```
varKurt = 8*D*(D+2)/N
```

```
varKurt = 1.2800
```

Compute the *p*-value for the Mardia kurtosis test on the linear discriminant model.

```
mahL = mahal(L,L.X,'ClassLabels',L.Y);
meanL = mean(mahL.^2);
[~,pvalL] = ztest(meanL,meanKurt,sqrt(varKurt))
```

```
pvalL = 0.0208
```

Because `pvalL` is smaller than 0.05, the Mardia kurtosis test rejects the hypothesis of the data being normally distributed with a constant covariance matrix.

Compute the *p*-value for the Mardia kurtosis test on the quadratic discriminant model.

```
mahQ = mahal(Q,Q.X,'ClassLabels',Q.Y);
meanQ = mean(mahQ.^2);
[~,pvalQ] = ztest(meanQ,meanKurt,sqrt(varKurt))
```

```
pvalQ = 0.7230
```

Because `pvalQ` is greater than 0.05, the data is consistent with the multivariate normal distribution.

The results indicate to use quadratic discriminant analysis, as opposed to linear discriminant analysis.

References

[1] Box, G. E. P. “A General Distribution Theory for a Class of Likelihood Criteria.” *Biometrika* 36, no. 3–4 (1949): 317–46. https://doi.org/10.1093/biomet/36.3-4.317.

[2] Mardia, K. V. “Measures of Multivariate Skewness and Kurtosis with Applications.” *Biometrika* 57, no. 3 (1970): 519–30. https://doi.org/10.1093/biomet/57.3.519.

See Also

Functions
`cvshrink` | `fitcdiscr`

Objects
`ClassificationDiscriminant` | `gmdistribution`

Related Examples

- “Discriminant Analysis Classification” on page 20-2
- “Bartlett Test of Equal Covariance Matrices for Linear Discriminant Analysis” on page 20-27
- “Compare Q-Q Plots for Linear and Quadratic Discriminants” on page 20-29

Naive Bayes

Naive Bayes Classification

The naive Bayes classifier is designed for use when predictors are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. It classifies data in two steps:

1. Training step: Using the training data, the method estimates the parameters of a probability distribution, assuming predictors are conditionally independent given the class.
2. Prediction step: For any unseen test data, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test data according the largest posterior probability.

The class-conditional independence assumption greatly simplifies the training step since you can estimate the one-dimensional class-conditional density for each predictor individually. While the class-conditional independence between predictors is not true in general, research shows that this optimistic assumption works well in practice. This assumption of class-conditional independence of the predictors allows the naive Bayes classifier to estimate the parameters required for accurate classification while using less training data than many other classifiers. This makes it particularly effective for data sets containing many predictors.

Supported Distributions

The training step in naive Bayes classification is based on estimating $P(X|Y)$, the probability or probability density of predictors `X` given class `Y`. The naive Bayes classification model `ClassificationNaiveBayes` and training function `fitcnb` provide support for normal (Gaussian), kernel, multinomial, and multivariate, multinomial predictor conditional distributions. To specify distributions for the predictors, use the `DistributionNames` name-value pair argument of `fitcnb`. You can specify one type of distribution for all predictors by supplying the character vector or string scalar corresponding to the distribution name, or specify different distributions for the predictors by supplying a length D string array or cell array of character vectors, where D is the number of predictors (that is, the number of columns of X).

Normal (Gaussian) Distribution

The `'normal'` distribution (specify using `'normal'`) is appropriate for predictors that have normal distributions in each class. For each predictor you model with a normal distribution, the naive Bayes classifier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class.

Kernel Distribution

The `'kernel'` distribution (specify using `'kernel'`) is appropriate for predictors that have a continuous distribution. It does not require a strong assumption such as a normal distribution and you can use it in cases where the distribution of a predictor may be skewed or have multiple peaks or modes. It requires more computing time and more memory than the normal distribution. For each predictor you model with a kernel distribution, the naive Bayes classifier computes a separate kernel density estimate for each class based on the training data for that class. By default the kernel is the normal kernel, and the classifier selects a width automatically for each class and predictor. The software supports specifying different kernels for each predictor, and different widths for each predictor or class.

Multivariate Multinomial Distribution

The multivariate, multinomial distribution (specify using `'mvmn'`) is appropriate for a predictor whose observations are categorical. Naive Bayes classifier construction using a multivariate multinomial predictor is described below. To illustrate the steps, consider an example where observations are labeled 0, 1, or 2, and a predictor the weather when the sample was conducted.

1. Record the distinct categories represented in the observations of the entire predictor. For example, the distinct categories (or predictor levels) might include sunny, rain, snow, and cloudy.
2. Separate the observations by response class. For example, segregate observations labeled 0 from observations labeled 1 and 2, and observations labeled 1 from observations labeled 2.
3. For each response class, fit a multinomial model using the category relative frequencies and total number of observations. For example, for observations labeled 0, the estimated probability it was sunny is $p_{sunny|0}$ = (number of sunny observations with label 0)/(number of observations with label 0), and similar for the other categories and response labels.

The class-conditional, multinomial random variables comprise a multivariate multinomial random variable.

Here are some other properties of naive Bayes classifiers that use multivariate multinomial.

- For each predictor you model with a multivariate multinomial distribution, the naive Bayes classifier:
  - Records a separate set of distinct predictor levels for each predictor
  - Computes a separate set of probabilities for the set of predictor levels for each class.
- The software supports modeling continuous predictors as multivariate multinomial. In this case, the predictor levels are the distinct occurrences of a measurement. This can lead a predictor having many predictor levels. It is good practice to discretize such predictors.

If an *observation* is a set of successes for various categories (represented by all of the predictors) out of a fixed number of independent trials, then specify that the predictors comprise a multinomial distribution. For details, see “Multinomial Distribution” on page 21-3.

Multinomial Distribution

The multinomial distribution (specify using `'DistributionNames','mn'`) is appropriate when, given the class, each *observation* is a multinomial random variable. That is, observation, or row, j of the predictor data X represents D categories, where x_{jd} is the number of successes for category (i.e., predictor) d in $n_j = \sum_{d=1}^{D} x_{jd}$ independent trials. The steps to train a naive Bayes classifier are outlined next.

1. For each class, fit a multinomial distribution for the predictors given the class by:
   a. Aggregating the weighted, category counts over all observations. Additionally, the software implements additive smoothing [1].
   b. Estimating the D category probabilities within each class using the aggregated category counts. These category probabilities compose the probability parameters of the multinomial distribution.
2. Let a new observation have a total count of m. Then, the naive Bayes classifier:

a Sets the total count parameter of each multinomial distribution to m

b For each class, estimates the class posterior probability using the estimated multinomial distributions

c Predicts the observation into the class corresponding to the highest posterior probability

Consider the so-called the bag-of-tokens model, where there is a bag containing a number of tokens of various types and proportions. Each predictor represents a distinct type of token in the bag, an observation is n independent draws (i.e., with replacement) of tokens from the bag, and the data is a vector of counts, where element d is the number of times token d appears.

A machine-learning application is the construction of an email spam classifier, where each predictor represents a word, character, or phrase (i.e., token), an observation is an email, and the data are counts of the tokens in the email. One predictor might count the number of exclamation points, another might count the number of times the word "money" appears, and another might count the number of times the recipient's name appears. This is a naive Bayes model under the further assumption that the total number of tokens (or the total document length) is independent of response class.

Other properties of naive Bayes classifiers that use multinomial observations include:

- Classification is based on the relative frequencies of the categories. If $n_j = 0$ for observation j, then classification is not possible for that observation.
- The predictors are not conditionally independent since they must sum to n_j.
- Naive Bayes is not appropriate when n_j provides information about the class. That is, this classifier requires that n_j is independent of the class.
- If you specify that the predictors are conditionally multinomial, then the software applies this specification to all predictors. In other words, you cannot include `'mn'` in a cell array when specifying `'DistributionNames'`.

If a *predictor* is categorical, i.e., is multinomial within a response class, then specify that it is multivariate multinomial. For details, see “Multivariate Multinomial Distribution” on page 21-3.

References

[1] Manning, C. D., P. Raghavan, and M. Schütze. *Introduction to Information Retrieval*, NY: Cambridge University Press, 2008.

See Also

Functions
`fitcnb` | `predict`

Objects
`ClassificationNaiveBayes`

Related Examples

- “Plot Posterior Classification Probabilities” on page 21-5
- “Visualize Decision Surfaces of Different Classifiers” on page 18-9
- “Classification” on page 17-7

Plot Posterior Classification Probabilities

This example shows how to visualize posterior classification probabilities predicted by a naive Bayes classification model.

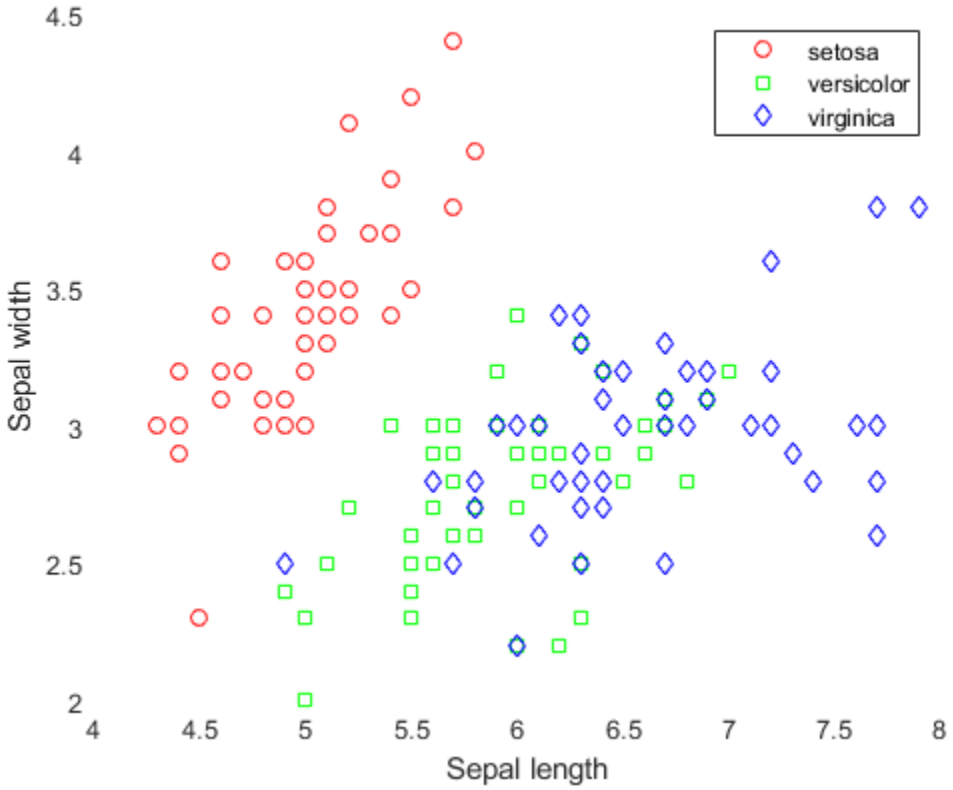
Load Fisher's iris data set.

```
load fisheriris
X = meas(:,1:2);
Y = species;
labels = unique(Y);
```

X is a numeric matrix that contains two petal measurements for 150 irises. Y is a cell array of character vectors that contains the corresponding iris species.

Visualize the data using a scatter plot. Group the variables by iris species.

```
figure;
gscatter(X(:,1), X(:,2), species,'rgb','osd');
xlabel('Sepal length');
ylabel('Sepal width');
```



Train a naive Bayes classifier.

```
mdl = fitcnb(X,Y);
```

`mdl` is a trained `ClassificationNaiveBayes` classifier.

Create a grid of points spanning the entire space within some bounds of the data. The data in `X(:,1)` ranges between 4.3 and 7.9. The data in `X(:,2)` ranges between 2 and 4.4.

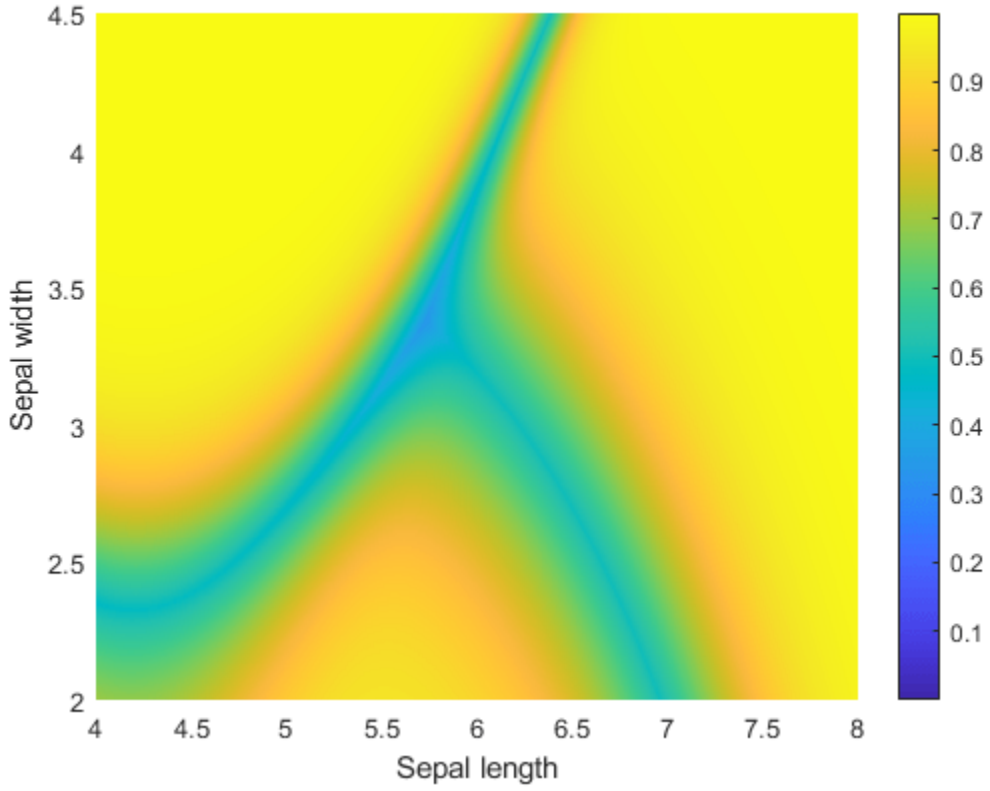
```
[xx1, xx2] = meshgrid(4:.01:8,2:.01:4.5);
XGrid = [xx1(:) xx2(:)];
```

Predict the iris species and posterior class probabilities of each observation in `XGrid` using `mdl`.

```
[predictedspecies,Posterior,~] = predict(mdl,XGrid);
```

Plot the posterior probability distribution for each species.

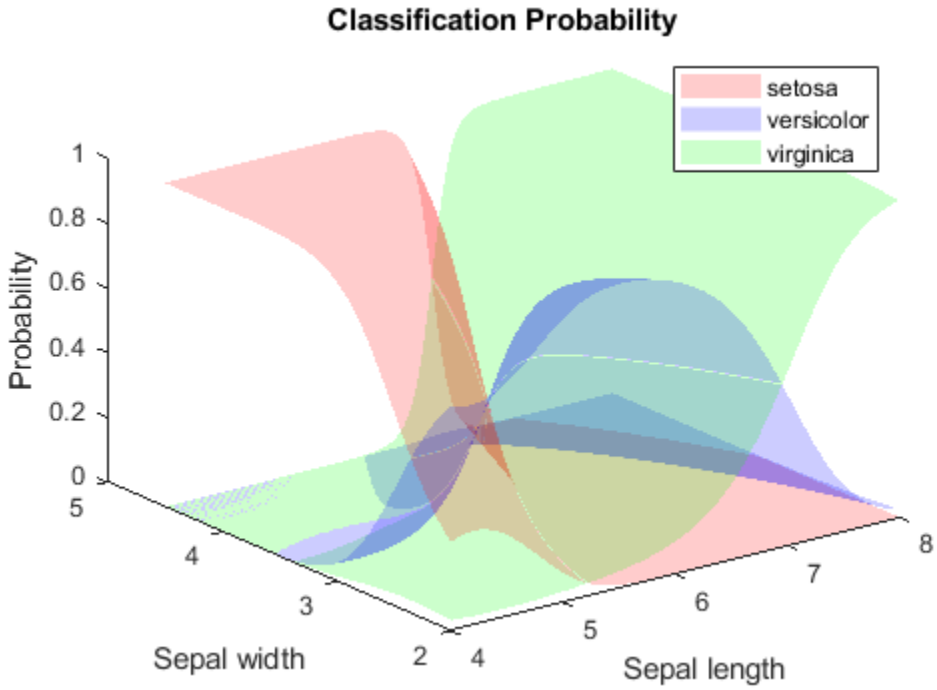
```
sz = size(xx1);
s = max(Posterior,[],2);

figure
hold on
surf(xx1,xx2,reshape(Posterior(:,1),sz),'EdgeColor','none')
surf(xx1,xx2,reshape(Posterior(:,2),sz),'EdgeColor','none')
surf(xx1,xx2,reshape(Posterior(:,3),sz),'EdgeColor','none')
xlabel('Sepal length');
ylabel('Sepal width');
colorbar
view(2)
hold off
```



The closer an observation gets to the decision surface, the less probable it is that the data belongs to a certain species.

Plot the classification probability distributions individually.

```
figure('Units','Normalized','Position',[0.25,0.55,0.4,0.35]);
hold on
surf(xx1,xx2,reshape(Posterior(:,1),sz),'FaceColor','red','EdgeColor','none')
surf(xx1,xx2,reshape(Posterior(:,2),sz),'FaceColor','blue','EdgeColor','none')
surf(xx1,xx2,reshape(Posterior(:,3),sz),'FaceColor','green','EdgeColor','none')
xlabel('Sepal length');
ylabel('Sepal width');
zlabel('Probability');
legend(labels)
title('Classification Probability')
alpha(0.2)
view(3)
hold off
```



See Also

Functions
`fitcnb` | `predict`

Objects
`ClassificationNaiveBayes`

Related Examples

- “Naive Bayes Classification” on page 21-2

22

Classification and Regression for High-Dimensional Data

Classification Learner

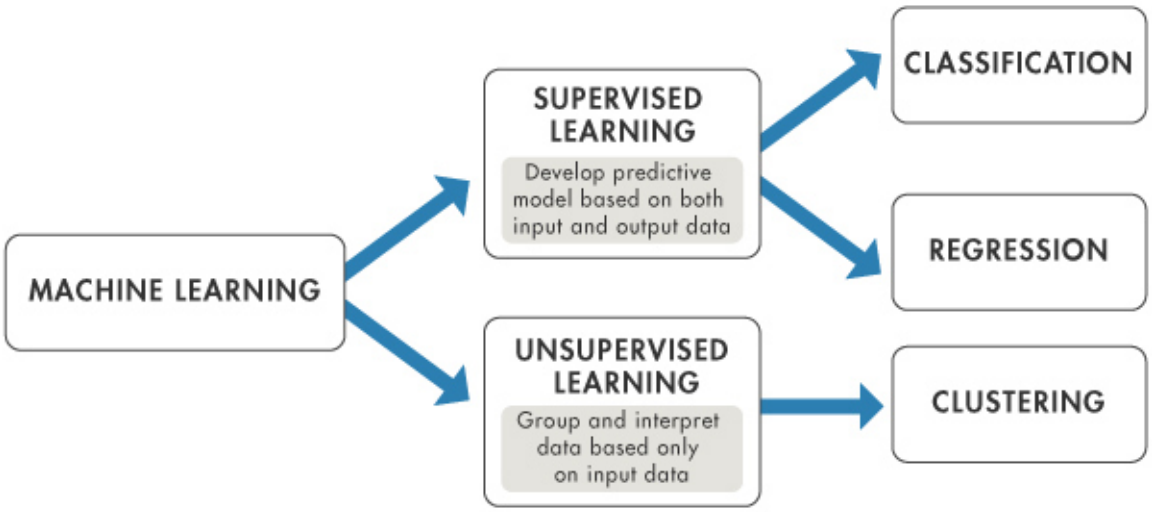
Machine Learning in MATLAB

In this section...

What Is Machine Learning?

Machine learning teaches computers to do what comes naturally to humans: learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases.

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.



The aim of supervised machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Supervised learning uses classification and regression techniques to develop predictive models.

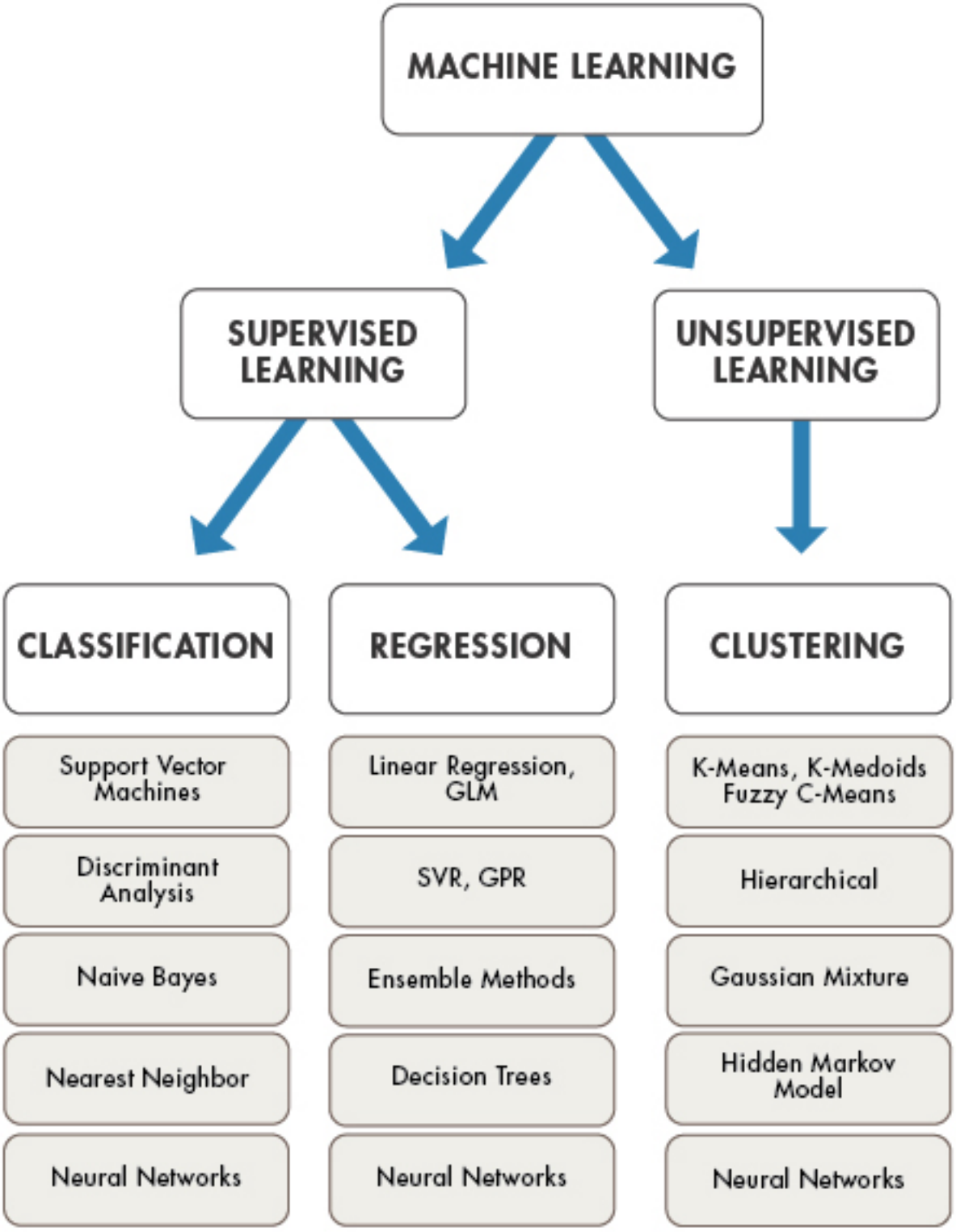
- Classification techniques predict categorical responses, for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, image and speech recognition, and credit scoring.

- Regression techniques predict continuous responses, for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading.

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses. Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for clustering include gene sequence analysis, market research, and object recognition.

Selecting the Right Algorithm

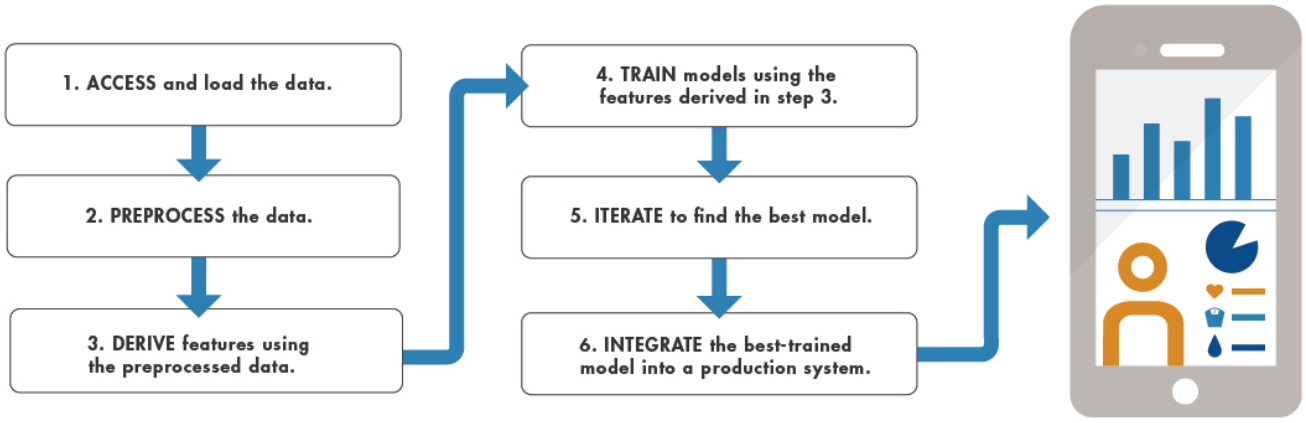
Choosing the right algorithm can seem overwhelming—there are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning. There is no best method or one size fits all. Finding the right algorithm is partly based on trial and error—even highly experienced data scientists cannot tell whether an algorithm will work without trying it out. Highly flexible models tend to overfit data by modeling minor variations that could be noise. Simple models are easier to interpret but might have lower accuracy. Therefore, choosing the right algorithm requires trading off one benefit against another, including model speed, accuracy, and complexity. Trial and error is at the core of machine learning—if one approach or algorithm does not work, you try another. MATLAB provides tools to help you try out a variety of machine learning models and choose the best.


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

To find MATLAB apps and functions to help you solve machine learning tasks, consult the following table. Some machine learning tasks are made easier by using apps, and others use command-line features.

Task	MATLAB Apps and Functions	Product	Learn More
Classification to predict categorical responses	Use the Classification Learner app to automatically train a selection of models and help you choose the best. You can generate MATLAB code to work with scripts. For more options, you can use the command-line interface.	Statistics and Machine Learning Toolbox	"Train Classification Models in Classification Learner App" on page 23-6 Classification Functions
Regression to predict continuous responses	Use the Regression Learner app to automatically train a selection of models and help you choose the best. You can generate MATLAB code to work with scripts and other function options. For more options, you can use the command-line interface.	Statistics and Machine Learning Toolbox	"Train Regression Models in Regression Learner App" on page 23-7 Regression Functions
Clustering	Use cluster analysis functions.	Statistics and Machine Learning Toolbox	"Cluster Analysis"
Computational finance tasks such as credit scoring	Use tools for modeling credit risk analysis.	Financial Toolbox™ and Risk Management Toolbox™	"Credit Risk" (Financial Toolbox)
Deep learning with neural networks for classification and regression	Use pretrained networks and functions to train convolutional neural networks.	Deep Learning Toolbox™	"Deep Learning in MATLAB" (Deep Learning Toolbox)
Facial recognition, motion detection, and object detection	Use deep learning tools for image processing and computer vision.	Deep Learning Toolbox and Computer Vision Toolbox™	"Recognition, Object Detection, and Semantic Segmentation" (Computer Vision Toolbox)

The following systematic machine learning workflow can help you tackle machine learning challenges. You can complete the entire workflow in MATLAB.



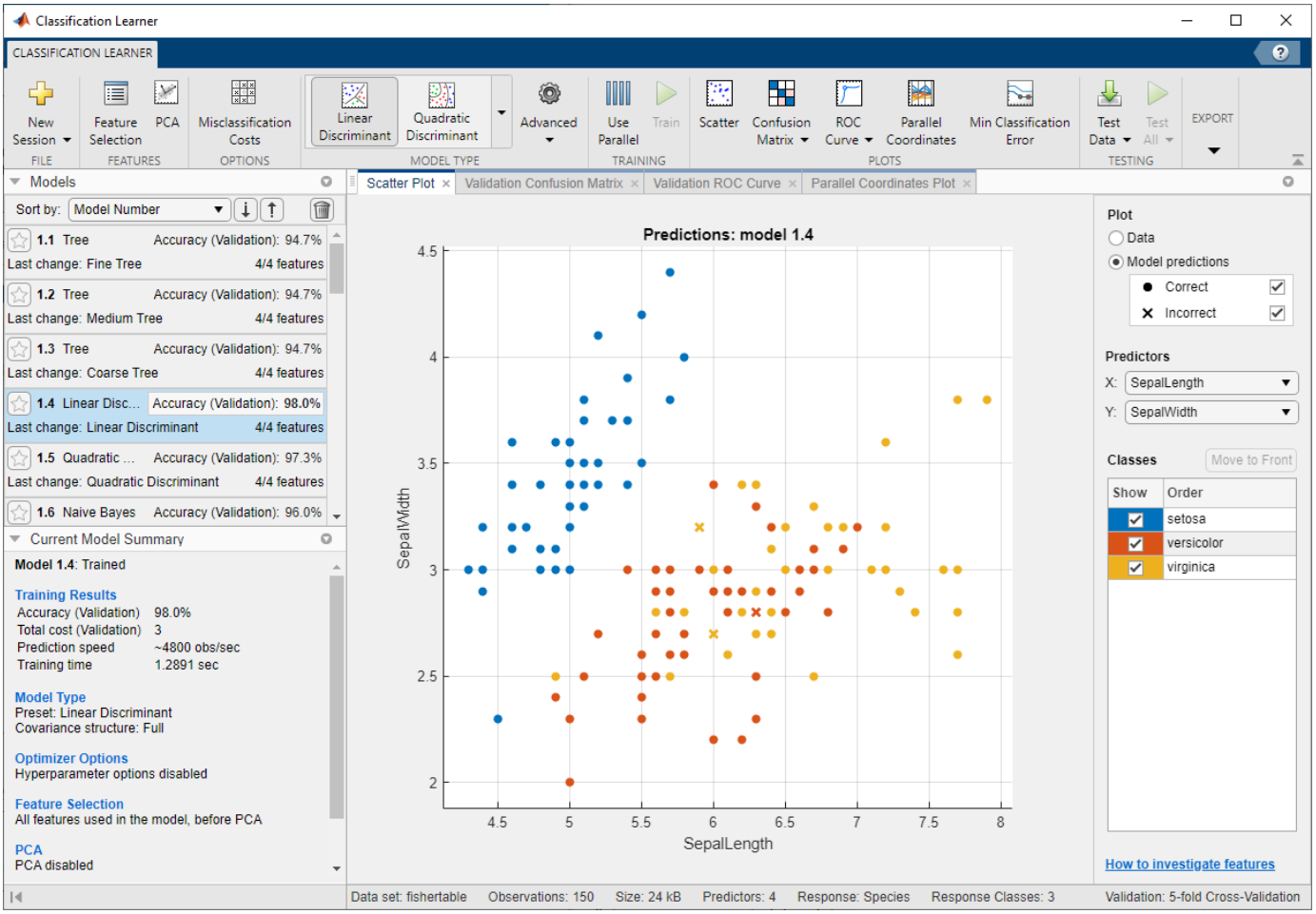
To integrate the best trained model into a production system, you can deploy Statistics and Machine Learning Toolbox machine learning models using MATLAB Compiler™. For many models, you can generate C-code for prediction using MATLAB Coder.

Train Classification Models in Classification Learner App

Use the Classification Learner app to train models to classify data using supervised machine learning. The app lets you explore supervised machine learning interactively using various classifiers.

- Automatically train a selection of models to help you choose the best model. Model types include decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensemble, and neural network classifiers.
- Explore your data, specify validation schemes, select features, and visualize results. By default, the app protects against overfitting by applying cross-validation. Alternatively, you can select holdout validation. Validation results help you choose the best model for your data. Plots and performance measures reflect the validated model results.
- Export models to the workspace to make predictions with new data. The app always trains a model on full data in addition to a model with the specified validation scheme, and the full model is the model you export.
- Generate MATLAB code from the app to create scripts, train with new data, work with huge data sets, or modify the code for further analysis.

To learn more, see “Train Classification Models in Classification Learner App” on page 23-10.



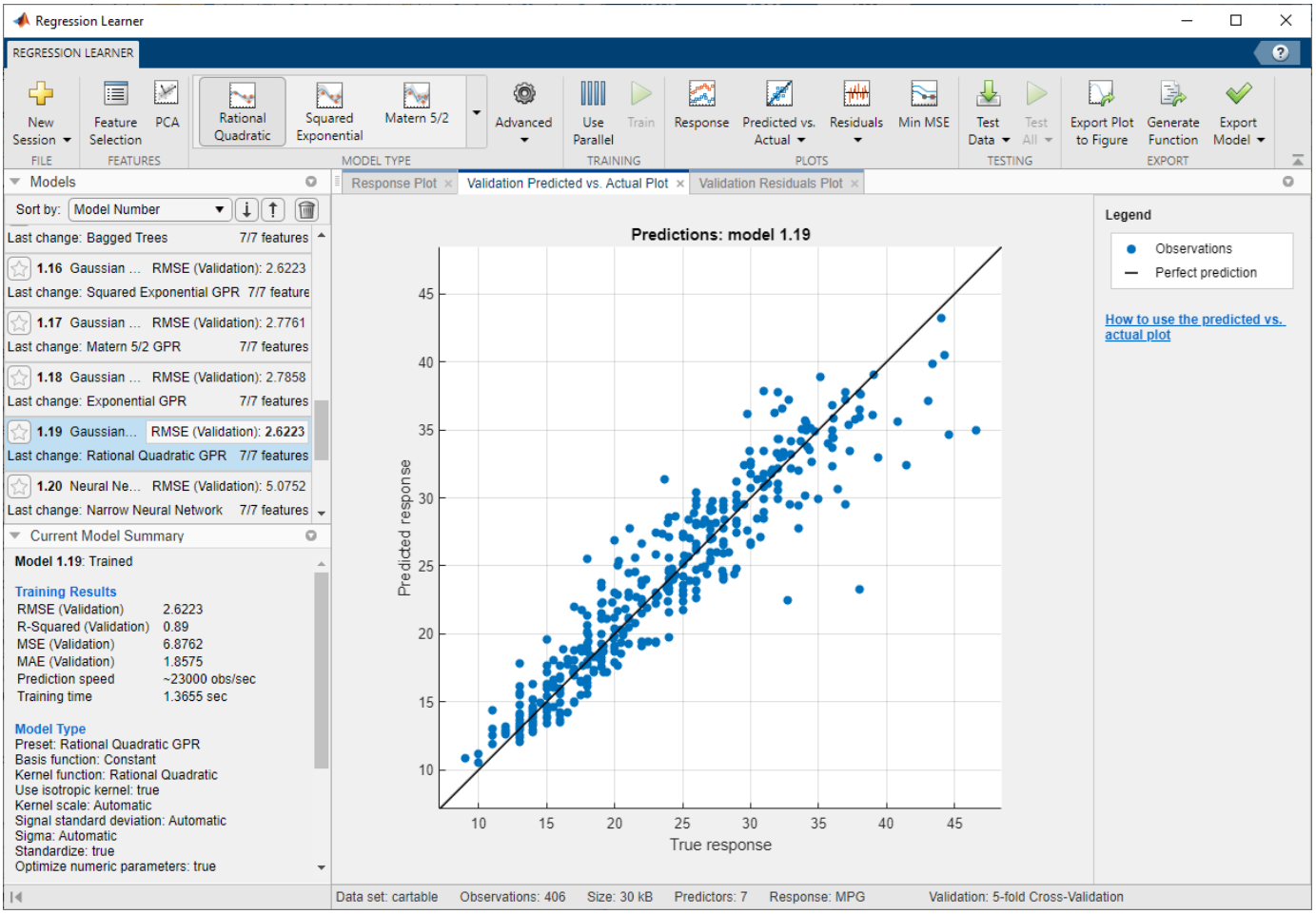
For more options, you can use the command-line interface. See “Classification”.

Train Regression Models in Regression Learner App

Use the Regression Learner app to train models to predict continuous data using supervised machine learning. The app lets you explore supervised machine learning interactively using various regression models.

- Automatically train a selection of models to help you choose the best model. Model types include linear regression models, regression trees, Gaussian process regression models, support vector machines, ensembles of regression trees, and neural network regression models.
- Explore your data, select features, and visualize results. Similar to Classification Learner, the Regression Learner applies cross-validation by default. The results and visualizations reflect the validated model. Use the results to choose the best model for your data.
- Export models to the workspace to make predictions with new data. The app always trains a model on full data in addition to a model with the specified validation scheme, and the full model is the model you export.
- Generate MATLAB code from the app to create scripts, train with new data, work with huge data sets, or modify the code for further analysis.

To learn more, see “Train Regression Models in Regression Learner App” on page 24-2.



For more options, you can use the command-line interface. See “Regression”.

Train Neural Networks for Deep Learning

Deep Learning Toolbox enables you to perform deep learning with convolutional neural networks for classification, regression, feature extraction, and transfer learning. The toolbox provides simple MATLAB commands for creating and interconnecting the layers of a deep neural network. Examples and pretrained networks make it easy to use MATLAB for deep learning, even without extensive knowledge of advanced computer vision algorithms or neural networks.

To learn more, see “Deep Learning in MATLAB” (Deep Learning Toolbox).

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Train Classification Models in Classification Learner App” on page 23-10
- “Cluster Analysis”

- “Credit Risk” (Financial Toolbox)
- “Recognition, Object Detection, and Semantic Segmentation” (Computer Vision Toolbox)
- “Deep Learning in MATLAB” (Deep Learning Toolbox)
- Machine Learning Made Easy (34 min 34 sec)

Train Classification Models in Classification Learner App

In this section...

You can use Classification Learner to train models of these classifiers: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensembles, and neural networks. In addition to training models, you can explore your data, select features, specify validation schemes, and evaluate results. You can export a model to the workspace to use the model with new data or generate MATLAB code to learn about programmatic classification.

Training a model in Classification Learner consists of two parts:

- Validated Model: Train a model with a validation scheme. By default, the app protects against overfitting by applying cross-validation. Alternatively, you can choose holdout validation. The validated model is visible in the app.
- Full Model: Train a model on full data without validation. The app trains this model simultaneously with the validated model. However, the model trained on full data is not visible in the app. When you choose a classifier to export to the workspace, Classification Learner exports the full model.

The app displays the results of the validated model. Diagnostic measures, such as model accuracy, and plots, such as a scatter plot or the confusion matrix chart, reflect the validated model results. You can automatically train a selection of or all classifiers, compare validation results, and choose the best model that works for your classification problem. When you choose a model to export to the workspace, Classification Learner exports the full model. Because Classification Learner creates a model object of the full model during training, you experience no lag time when you export the model. You can use the exported model to make predictions on new data.

To get started by training a selection of model types, see “Automated Classifier Training” on page 23-10. If you already know what classifier type you want to train, see “Manual Classifier Training” on page 23-13.

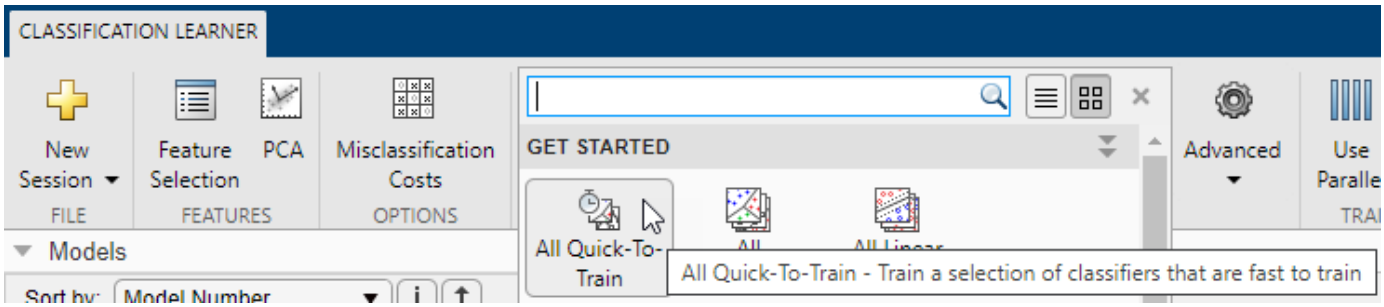
Automated Classifier Training

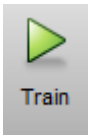
You can use Classification Learner to automatically train a selection of different classification models on your data.

- Get started by automatically training multiple models at once. You can quickly try a selection of models, then explore promising models interactively.
- If you already know what classifier type you want, train individual classifiers instead. See “Manual Classifier Training” on page 23-13.

1. On the **Apps** tab, in the **Machine Learning** group, click **Classification Learner**.
2. Click **New Session** and select data from the workspace or from file. Specify a response variable and variables to use as predictors. See “Select Data and Validation for Classification Problem” on page 23-18.

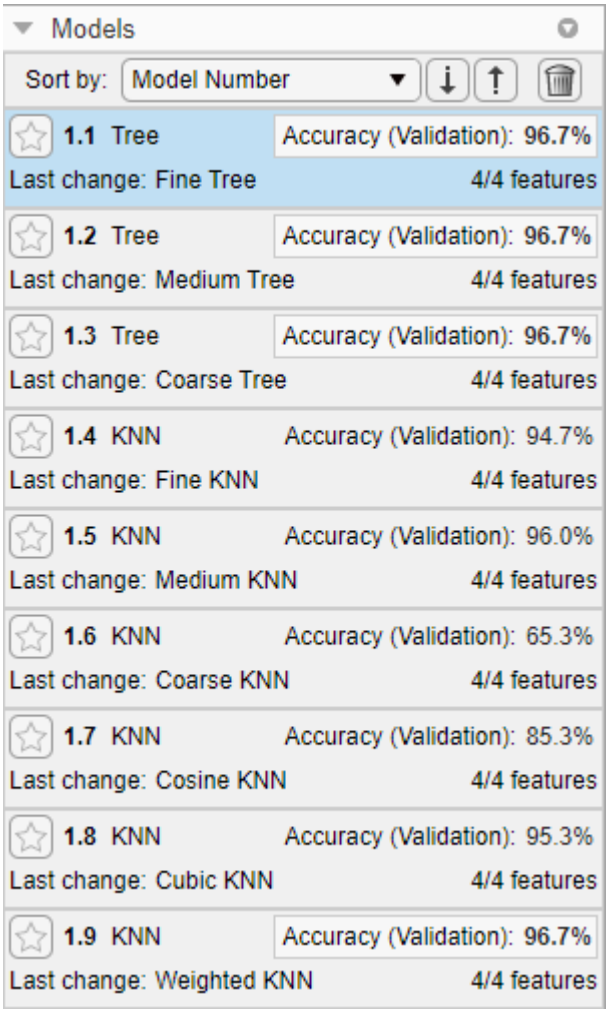
3 On the **Classification Learner** tab, in the **Model Type** section, click **All Quick-To-Train**. This option will train all the model presets available for your data set that are fast to fit.



4 Click **Train**. 

Note If you have Parallel Computing Toolbox, you can train the models in parallel. See “Parallel Classifier Training” on page 23-14.

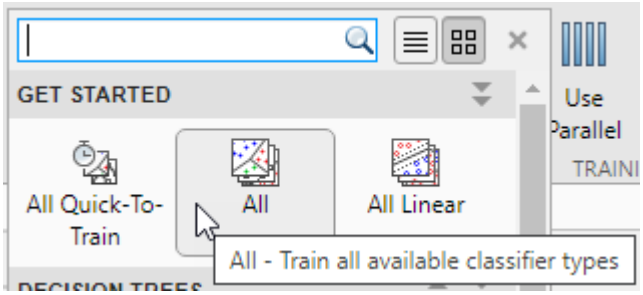
A selection of model types appears in the **Models** pane. When they finish training, the best percentage **Accuracy (Validation)** score is highlighted in a box.



5 Click models in the **Models** pane to explore results in the plots.

For next steps, see “Manual Classifier Training” on page 23-13 or “Compare and Improve Classification Models” on page 23-14.

6 To try all the nonoptimizable classifier model presets available for your data set, click **All**, then click **Train**.

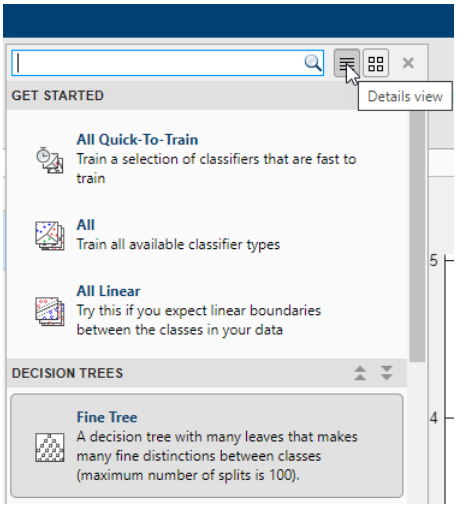


Manual Classifier Training

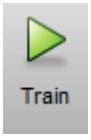
If you want to explore individual model types, or if you already know what classifier type you want, you can train classifiers one at a time, or a train a group of the same type.

1 Choose a classifier. On the **Classification Learner** tab, in the **Model Type** section, click a classifier type. To see all available classifier options, click the arrow on the far right of the **Model Type** section to expand the list of classifiers. The nonoptimizable model options in the **Model Type** gallery are preset starting points with different settings, suitable for a range of different classification problems.

   To read a description of each classifier, switch to the details view.



   For more information on each option, see “Choose Classifier Options” on page 23-22.

2 After selecting a classifier, click **Train**. 

   Repeat to try different classifiers.

   Tip Try decision trees and discriminants first. If the models are not accurate enough predicting the response, try other classifiers with higher flexibility. To avoid overfitting, look for a model of lower flexibility that provides sufficient accuracy.

3 If you want to try all nonoptimizable models of the same or different types, then select one of the **All** options in the **Model Type** gallery.

   Alternatively, if you want to automatically tune hyperparameters of a specific model type, select the corresponding **Optimizable** model and perform hyperparameter optimization. For more information, see “Hyperparameter Optimization in Classification Learner App” on page 23-54.

For next steps, see “Compare and Improve Classification Models” on page 23-14

Parallel Classifier Training

You can train models in parallel using Classification Learner if you have Parallel Computing Toolbox. Parallel training allows you to train multiple classifiers at once and continue working.

To control parallel training, toggle the **Use Parallel** button in the app toolstrip. The **Use Parallel** button is only available if you have Parallel Computing Toolbox.

1. The first time you click **Train** after clicking the **Use Parallel** button, you see a dialog box while the app opens a parallel pool of workers. After the pool opens, you can train multiple classifiers at once.
2. When classifiers are training in parallel, you see progress indicators on each training and queued model in the **Models** pane, and you can cancel individual models if you want. During training, you can examine results and plots from models, and initiate training of more classifiers.

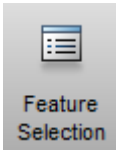
If you have Parallel Computing Toolbox, then parallel training is available in Classification Learner, and you do not need to set the `UseParallel` option of the `statset` function.

Note You cannot perform hyperparameter optimization in parallel. The app disables the **Use Parallel** button when you select an optimizable model. If you then select a nonoptimizable model, the button is off by default.

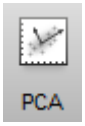
Compare and Improve Classification Models

1. Click models in the **Models** pane to explore the results in the plots. Compare model performance by inspecting results in the scatter plot and confusion matrix. Examine the **Accuracy (Validation)** score reported for each model.

   Additionally, you can compare the models by using the **Sort by** options in the **Models** pane. Delete any unwanted model by selecting the model and clicking the **Delete selected model** button in the upper right of the pane, or right-clicking the model and selecting **Delete model**.

   See “Assess Classifier Performance in Classification Learner” on page 23-65.
2. Select the best model in the **Models** pane and then try including and excluding different features in the model. Click **Feature Selection**. 

   Try the parallel coordinates plot to help you identify features to remove. See if you can improve the model by removing features with low predictive power. Specify predictors to include in the model, and train new models using the new options. Compare results among the models in the **Models** pane.

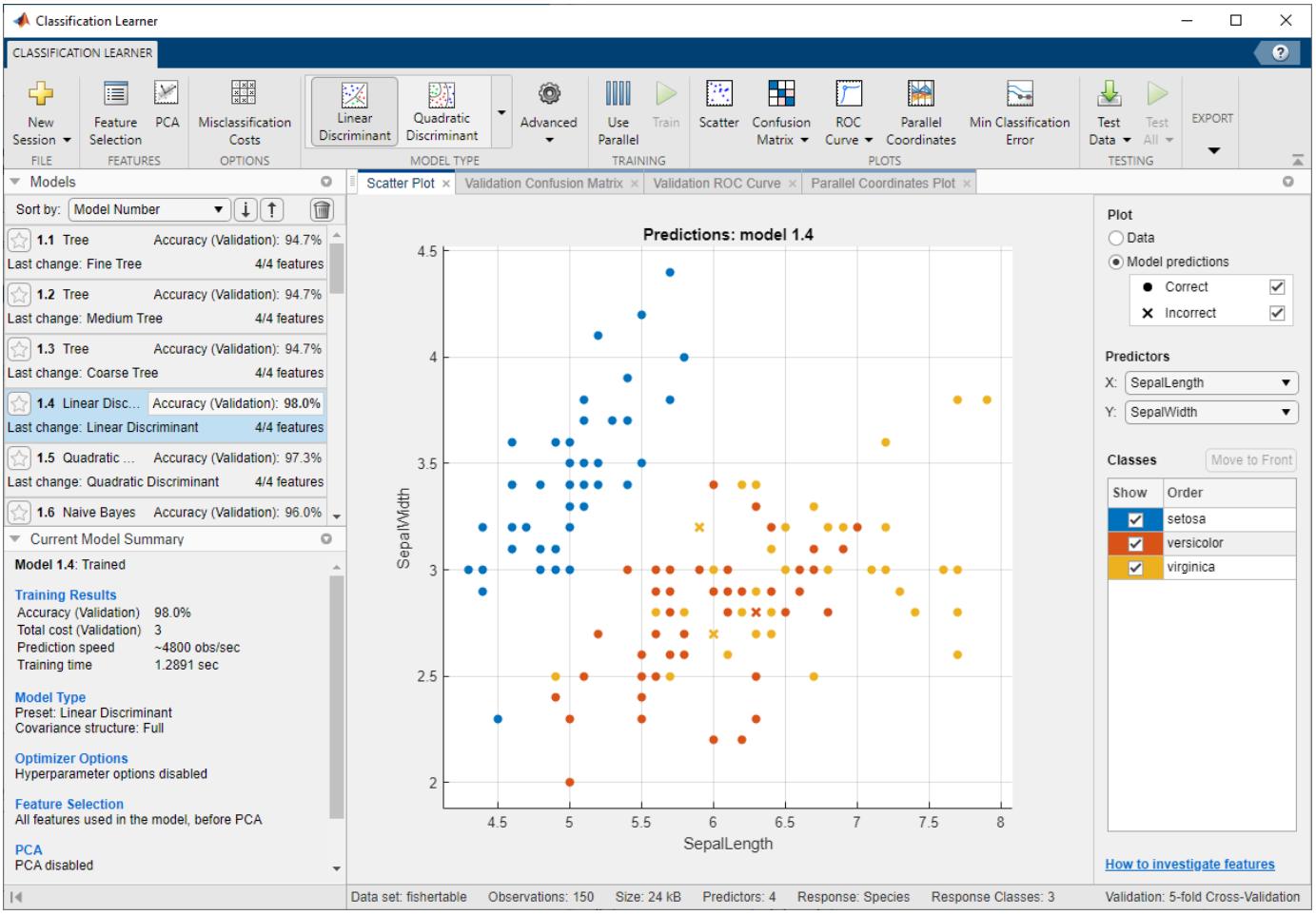
You can also try transforming features with PCA to reduce dimensionality. 

See “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42.

3. To improve the model further, you can try changing classifier parameter settings in the Advanced dialog box, and then train using the new options. To learn how to control model flexibility, see “Choose Classifier Options” on page 23-22. For information on how to tune model parameter settings automatically, see “Hyperparameter Optimization in Classification Learner App” on page 23-54.
4. If feature selection, PCA, or new parameter settings improve your model, try training **All** model types with the new settings. See if another model type does better with the new settings.

Tip To avoid overfitting, look for a model of lower flexibility that provides sufficient accuracy. For example, look for simple models such as decision trees and discriminants that are fast and easy to interpret. If the models are not accurate enough predicting the response, choose other classifiers with higher flexibility, such as ensembles. To learn about the model flexibility, see “Choose Classifier Options” on page 23-22.

The figure shows the app with a **Models** pane containing various classifier types.



Tip For a step-by-step example comparing different classifiers, see “Train Decision Trees Using Classification Learner App” on page 23-83.

For next steps, generate code to train the model with different data, or export trained models to the workspace to make predictions using new data. See “Export Classification Model to Predict New Data” on page 23-77.

See Also

Related Examples

- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

- “Train Decision Trees Using Classification Learner App” on page 23-83
- “Machine Learning in MATLAB” on page 23-2

Select Data and Validation for Classification Problem

In this section...

Select Data from Workspace

Tip In Classification Learner, tables are the easiest way to use your data, because they can contain numeric and label data. Use the Import Tool to bring your data into the MATLAB workspace as a table, or use the table functions to create a `table` from workspace variables. See “Tables”.

1. Load your data into the MATLAB workspace.

   Predictor and response variables can be numeric, categorical, string, or logical vectors, cell arrays of character vectors, or character arrays. Note: If your response variable is a string vector, then the predictions of the trained model form a cell array of character vectors.

   Combine the predictor data into one variable, either a table or a matrix. You can additionally combine your predictor data and response variable, or you can keep them separate.

   For example data sets, see “Example Data for Classification” on page 23-19.
2. On the **Apps** tab, click **Classification Learner**.
3. In Classification Learner, on the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.
4. In the New Session from Workspace dialog box, under **Data Set Variable**, select a table or matrix from the list of workspace variables.

   If you select a matrix, choose whether to use rows or columns for observations by clicking the option buttons.
5. Under **Response**, observe the default response variable. The app tries to select a suitable response variable from the data set variable and treats all other variables as predictors.

   If you want to use a different response variable, you can:

   - Use the list to select another variable from the data set variable.
   - Select a separate workspace variable by clicking the **From workspace** option button and then selecting a variable from the list.
6. Under **Predictors**, add or remove predictors using the check boxes. Add or remove all predictors by clicking **Add All** or **Remove All**. You can also add or remove multiple predictors by selecting them in the table, and then clicking **Add N** or **Remove N**, where **N** is the number of selected predictors. The **Add All** and **Remove All** buttons change to **Add N** and **Remove N** when you select multiple predictors.
7. To accept the default validation scheme and continue, click **Start Session**. The default validation option is 5-fold cross-validation, which protects against overfitting.

Tip If you have a large data set you might want to switch to holdout validation. To learn more, see “Choose Validation Scheme” on page 23-20.

Note If you prefer loading data into the app directly from the command line, you can specify the predictor data, response variable, and validation type to use in Classification Learner in the command line call to `classificationLearner`. For more information, see **Classification Learner**.

For next steps, see “Train Classification Models in Classification Learner App” on page 23-10.

Import Data from File

1. On the **Classification Learner** tab, in the **File** section, select **New Session** > **From File**.
2. Select a file type in the list, such as spreadsheets, text files, or comma separated values (`.csv`) files, or select **All Files** to browse for other file types such as `.dat`.

Example Data for Classification

To get started using Classification Learner, try the following example data sets.

Name	Size	Description
Fisher Iris	Number of predictors: 4 Number of observations: 150 Number of classes: 3 Response: species	Measurements from three species of iris. Try to classify the species. For a step-by-step example, see “Train Decision Trees Using Classification Learner App” on page 23-83.
	Create a table from the `.csv` file: `fishertable = readtable('fisheriris.csv');`	
Credit Rating	Number of predictors: 6 Number of observations: 3932 Number of classes: 7 Response: Rating	Financial ratios and industry sectors information for a list of corporate customers. The response variable consists of credit ratings (AAA, AA, A, BBB, BB, B, CCC) assigned by a rating agency.
	Create a table from the `CreditRating_Historical.dat` file: `creditrating = readtable('CreditRating_Historical.dat');`	
Cars	Number of predictors: 7 Number of observations: 100 Number of classes: 7 Response: Origin	Measurements of cars, in 1970, 1976, and 1982. Try to classify the country of origin.

	Create a table from variables in the `carsmall.mat` file:
	
	`load carsmall`
	`cartable = table(Acceleration, Cylinders, Displacement,...`
	`Horsepower, Model_Year, MPG, Weight, Origin);`	

Name	Size	Description
Arrhythmia	Number of predictors: 279 Number of observations: 452 Number of classes: 16 Response: Class (Y)	Patient information and response variables that indicate the presence and absence of cardiac arrhythmia. Misclassifying a patient as "normal" has more severe consequences than false positives classified as “has arrhythmia”.

	Create a table from the `.mat` file:
	
	`load arrhythmia`
	`Arrhythmia = array2table(X);`
	`Arrhythmia.Class = categorical(Y);`	

Ovarian Cancer	Number of predictors: 4000 Number of observations: 216 Number of classes: 2 Response: Group	Ovarian cancer data generated using the WCX2 protein array. Includes 95 controls and 121 ovarian cancers.

	Create a table from the `.mat` file:
	
	`load ovariancancer`
	`ovariancancer = array2table(obs);`
	`ovariancancer.Group = categorical(grp);`	

Ionosphere	Number of predictors: 34 Number of observations: 351 Number of classes: 2 Response: Group (Y)	Signals from a phased array of 16 high-frequency antennas. Good (“g”) returned radar signals are those showing evidence of some type of structure in the ionosphere. Bad (“b”) signals are those that pass through the ionosphere.

	Create a table from the `.mat` file:
	
	`load ionosphere`
	`ionosphere = array2table(X);`
	`ionosphere.Group = Y;`	

Choose Validation Scheme

Choose a validation method to examine the predictive accuracy of the fitted models. Validation estimates model performance on new data compared to the training data, and helps you choose the best model. Validation protects against overfitting. Choose a validation scheme before training any models, so that you can compare all the models in your session using the same validation scheme.

Tip Try the default validation scheme and click **Start Session** to continue. The default option is 5-fold cross-validation, which protects against overfitting.

If you have a large data set and training models takes too long using cross-validation, reimport your data and try the faster holdout validation instead.

- **Cross-Validation**: Select a number of folds (or divisions) to partition the data set.

  If you choose *k* folds, then the app:

1. Partitions the data into *k* disjoint sets or folds
2. For each validation fold:
   a. Trains a model using the training-fold observations (observations not in the validation fold)
   b. Assesses model performance using validation-fold data
3. Calculates the average validation error over all folds

This method gives a good estimate of the predictive accuracy of the final model trained with all the data. It requires multiple fits but makes efficient use of all the data, so it is recommended for small data sets.

- **Holdout Validation**: Select a percentage of the data to use as a validation set. The app trains a model on the training set and assesses its performance with the validation set. The model used for validation is based on only a portion of the data, so **Holdout Validation** is recommended only for large data sets. The final model is trained with the full data set.
- **Resubstitution Validation**: No protection against overfitting. The app uses all of the data for training and computes the error rate on the same data. Without any separate validation data, you get an unrealistic estimate of the model's performance on new data. That is, the training sample accuracy is likely to be unrealistically high, and the predictive accuracy is likely to be lower.

  To help you avoid overfitting to the training data, choose another validation scheme instead.

Note The validation scheme only affects the way that Classification Learner computes validation metrics. The final model is always trained using the full data set.

All the classification models you train after selecting data use the same validation scheme that you select in this dialog box. You can compare all the models in your session using the same validation scheme.

To change the validation selection and train new models, you can select data again, but you lose any trained models. The app warns you that importing data starts a new session. Save any trained models you want to keep to the workspace, and then import the data.

For next steps training models, see "Train Classification Models in Classification Learner App" on page 23-10.

See Also

Related Examples

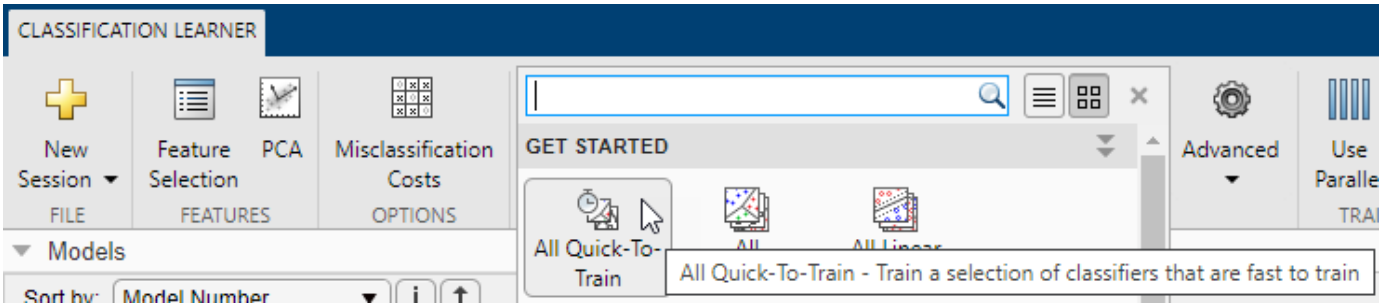
- "Train Classification Models in Classification Learner App" on page 23-10
- "Choose Classifier Options" on page 23-22
- "Feature Selection and Feature Transformation Using Classification Learner App" on page 23-42
- "Assess Classifier Performance in Classification Learner" on page 23-65
- "Export Classification Model to Predict New Data" on page 23-77
- "Train Decision Trees Using Classification Learner App" on page 23-83

Choose Classifier Options

In this section...

Choose a Classifier Type

You can use Classification Learner to automatically train a selection of different classification models on your data. Use automated training to quickly try a selection of model types, then explore promising models interactively. To get started, try these options first:



Get Started Classifier Buttons	Description
All Quick-To-Train	Try this first. The app will train all the model types available for your data set that are typically fast to fit.
All Linear	Try this if you expect linear boundaries between the classes in your data. This option fits only Linear SVM and Linear Discriminant.
All	Use this to train all available nonoptimizable model types. Trains every type regardless of any prior trained models. Can be time-consuming.

See “Automated Classifier Training” on page 23-10.

If you want to explore classifiers one at a time, or you already know what classifier type you want, you can select individual models or train a group of the same type. To see all available classifier options, on the **Classification Learner** tab, click the arrow in the **Model Type** section to expand the list of classifiers. The nonoptimizable model options in the **Model Type** gallery are preset starting points with different settings, suitable for a range of different classification problems. To use

optimizable model options and tune model hyperparameters automatically, see “Hyperparameter Optimization in Classification Learner App” on page 23-54.

For help choosing the best classifier type for your problem, see the table showing typical characteristics of different supervised learning algorithms. Use the table as a guide for your final choice of algorithms. Decide on the tradeoff you want in speed, memory usage, flexibility, and interpretability. The best classifier type depends on your data.

Tip To avoid overfitting, look for a model of lower flexibility that provides sufficient accuracy. For example, look for simple models such as decision trees and discriminants that are fast and easy to interpret. If the models are not accurate enough predicting the response, choose other classifiers with higher flexibility, such as ensembles. To control flexibility, see the details for each classifier type.

Characteristics of Classifier Types

Classifier	Prediction Speed	Memory Usage	Interpretability
“Decision Trees” on page 23-26 	Fast	Small	Easy
“Discriminant Analysis” on page 23-29 	Fast	Small for linear, large for quadratic	Easy
“Logistic Regression” on page 23-30 	Fast	Medium	Easy
“Naive Bayes Classifiers” on page 23-30 	Medium for simple distributions Slow for kernel distributions or high-dimensional data	Small for simple distributions Medium for kernel distributions or high-dimensional data	Easy
“Support Vector Machines” on page 23-31 	Medium for linear Slow for others	Medium for linear. All others: medium for multiclass, large for binary.	Easy for Linear SVM. Hard for all other kernel types.
“Nearest Neighbor Classifiers” on page 23-34 	Slow for cubic Medium for others	Medium	Hard
“Ensemble Classifiers” on page 23-37 	Fast to medium depending on choice of algorithm	Low to high depending on choice of algorithm	Hard
“Neural Network Classifiers” on page 23-39 	Fast	Medium	Hard

The tables on this page describe general characteristics of speed and memory usage for all the nonoptimizable preset classifiers. The classifiers were tested with various data sets (up to 7000 observations, 80 predictors, and 50 classes), and the results define the following groups:

Speed

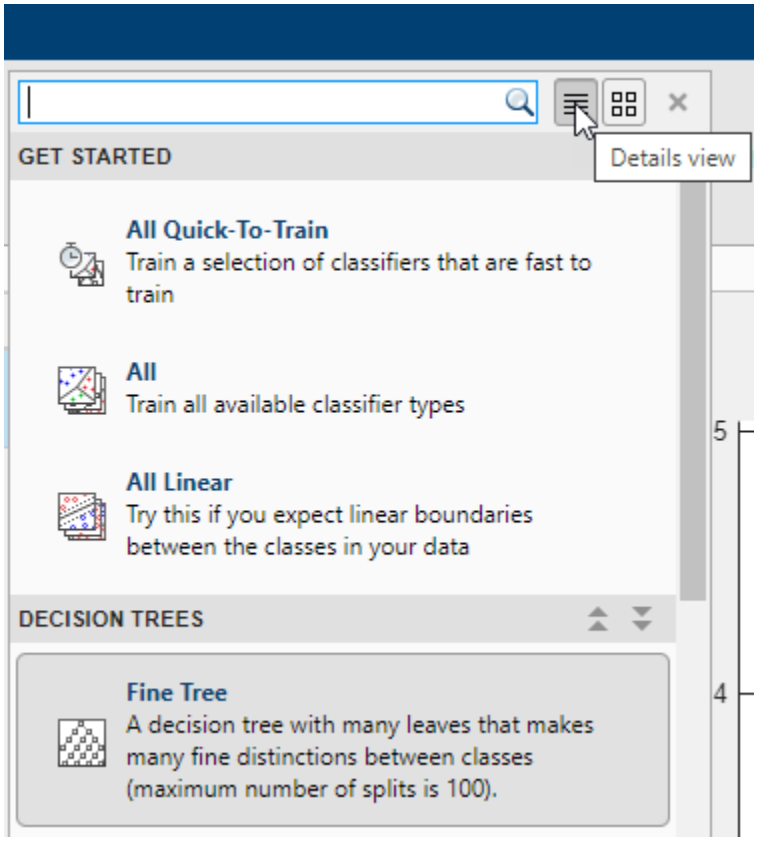
- Fast 0.01 second
- Medium 1 second
- Slow 100 seconds

Memory

- Small 1MB
- Medium 4MB
- Large 100MB

These tables provide a general guide. Your results depend on your data and the speed of your machine.

To read a description of each classifier in Classification Learner, switch to the details view.



Tip After you choose a classifier type (for example, decision trees), try training using each of the classifiers. The nonoptimizable options in the **Model Type** gallery are starting points with different settings. Try them all to see which option produces the best model with your data.

For workflow instructions, see “Train Classification Models in Classification Learner App” on page 23-10.

Categorical Predictor Support

In Classification Learner, the **Model Type** gallery shows as available the classifier types that support your selected data.

Classifier	All predictors numeric	All predictors categorical	Some categorical, some numeric
Decision Trees	Yes	Yes	Yes
Discriminant Analysis	Yes	No	No
Logistic Regression	Yes	Yes	Yes
Naive Bayes	Yes	Yes	Yes
SVM	Yes	Yes	Yes
Nearest Neighbor	Euclidean distance only	Hamming distance only	No
Ensembles	Yes	Yes, except Subspace Discriminant	Yes, except any Subspace
Neural Networks	Yes	Yes	Yes

Decision Trees

Decision trees are easy to interpret, fast for fitting and prediction, and low on memory usage, but they can have low predictive accuracy. Try to grow simpler trees to prevent overfitting. Control the depth with the **Maximum number of splits** setting.

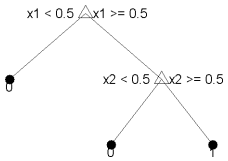
Tip Model flexibility increases with the **Maximum number of splits** setting.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Coarse Tree 	Fast	Small	Easy	Low Few leaves to make coarse distinctions between classes (maximum number of splits is 4).
Medium Tree 	Fast	Small	Easy	Medium Medium number of leaves for finer distinctions between classes (maximum number of splits is 20).
Fine Tree 	Fast	Small	Easy	High Many leaves to make many fine distinctions between classes (maximum number of splits is 100).

Tip In the **Model Type** gallery click **All Trees** to try each of the nonoptimizable decision tree options. Train them all to see which settings produce the best model with your data. Select the best

model in the **Models** pane. To try to improve your model, try feature selection, and then try changing some advanced options.

You train classification trees to predict responses to data. To predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response. Statistics and Machine Learning Toolbox trees are binary. Each step in a prediction involves checking the value of one predictor (variable). For example, here is a simple classification tree:

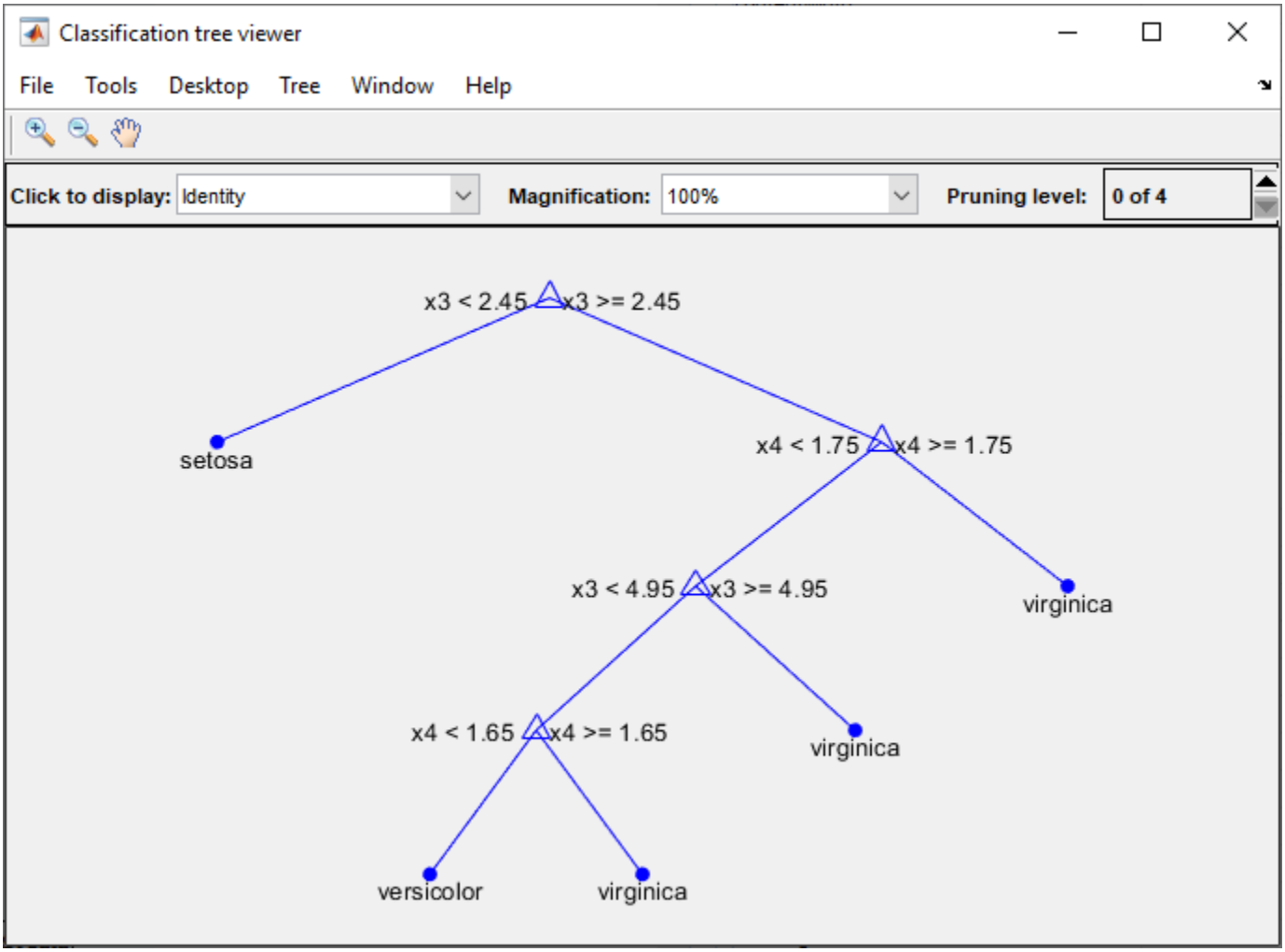


This tree predicts classifications based on two predictors, `x1` and `x2`. To predict, start at the top node. At each decision, check the values of the predictors to decide which branch to follow. When the branches reach a leaf node, the data is classified either as type `0` or `1`.

You can visualize your decision tree model by exporting the model from the app, and then entering:

```
view(trainedModel.ClassificationTree,'Mode','graph')
```

The figure shows an example fine tree trained with the `fisheriris` data.



Tip For an example, see “Train Decision Trees Using Classification Learner App” on page 23-83.

Advanced Tree Options

Classification trees in Classification Learner use the `fitctree` function. You can set these options:

- **Maximum number of splits**

  Specify the maximum number of splits or branch points to control the depth of your tree. When you grow a decision tree, consider its simplicity and predictive power. To change the number of splits, click the buttons or enter a positive integer value in the **Maximum number of splits** box.

  - A fine tree with many leaves is usually highly accurate on the training data. However, the tree might not show comparable accuracy on an independent test set. A leafy tree tends to overtrain, and its validation accuracy is often far lower than its training (or resubstitution) accuracy.

  - In contrast, a coarse tree does not attain high training accuracy. But a coarse tree can be more robust in that its training accuracy can approach that of a representative test set. Also, a coarse tree is easy to interpret.
- **Split criterion**

  Specify the split criterion measure for deciding when to split nodes. Try each of the three settings to see if they improve the model with your data.

  Split criterion options are `Gini's diversity index`, `Twoing rule`, or `Maximum deviance reduction` (also known as cross entropy).

  The classification tree tries to optimize to pure nodes containing only one class. Gini's diversity index (the default) and the deviance criterion measure node impurity. The twoing rule is a different measure for deciding how to split a node, where maximizing the twoing rule expression increases node purity.

  For details of these split criteria, see `ClassificationTree` “More About” on page 33-585.
- **Surrogate decision splits** — Only for missing data.

  Specify surrogate use for decision splits. If you have data with missing values, use surrogate splits to improve the accuracy of predictions.

  When you set **Surrogate decision splits** to `On`, the classification tree finds at most 10 surrogate splits at each branch node. To change the number, click the buttons or enter a positive integer value in the **Maximum surrogates per node** box.

  When you set **Surrogate decision splits** to `Find All`, the classification tree finds all surrogate splits at each branch node. The `Find All` setting can use considerable time and memory.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

Discriminant Analysis

Discriminant analysis is a popular first classification algorithm to try because it is fast, accurate and easy to interpret. Discriminant analysis is good for wide datasets.

Discriminant analysis assumes that different classes generate data based on different Gaussian distributions. To train a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Linear Discriminant 	Fast	Small	Easy	Low Creates linear boundaries between classes.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Quadratic Discriminant 	Fast	Large	Easy	Low Creates nonlinear boundaries between classes (ellipse, parabola or hyperbola).

Advanced Discriminant Options

Discriminant analysis in Classification Learner uses the `fitcdiscr` function. For both linear and quadratic discriminants, you can change the **Covariance structure** option. If you have predictors with zero variance or if any of the covariance matrices of your predictors are singular, training can fail using the default, `Full` covariance structure. If training fails, select the `Diagonal` covariance structure instead.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

Logistic Regression

If you have 2 classes, logistic regression is a popular simple classification algorithm to try because it is easy to interpret. The classifier models the class probabilities as a function of the linear combination of predictors.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Logistic Regression 	Fast	Medium	Easy	Low You cannot change any parameters to control model flexibility.

Logistic regression in Classification Learner uses the `fitglm` function. You cannot set any options for this classifier in the app.

Naive Bayes Classifiers

Naive Bayes classifiers are easy to interpret and useful for multiclass classification. The naive Bayes algorithm leverages Bayes theorem and makes the assumption that predictors are conditionally independent, given the class. Use these classifiers if this independence assumption is valid for predictors in your data. However, the algorithm still appears to work well when the independence assumption is not valid.

For kernel naive Bayes classifiers, you can control the kernel smoother type with the **Kernel Type** setting, and control the kernel smoothing density support with the **Support** setting.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Gaussian Naive Bayes 	Medium Slow for high-dimensional data	Small Medium for high-dimensional data	Easy	Low You cannot change any parameters to control model flexibility.
Kernel Naive Bayes 	Slow	Medium	Easy	Medium You can change settings for **Kernel Type** and **Support** to control how the classifier models predictor distributions.

Naive Bayes in Classification Learner uses the `fitcnb` function.

Advanced Naive Bayes Options

For kernel naive Bayes classifiers, you can set these options:

- **Kernel Type** — Specify the kernel smoother type. Try setting each of these options to see if they improve the model with your data.

  Kernel type options are `Gaussian`, `Box`, `Epanechnikov`, or `Triangle`.
- **Support** — Specify the kernel smoothing density support. Try setting each of these options to see if they improve the model with your data.

  Support options are `Unbounded` (all real values) or `Positive` (all positive real values).

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

For next steps training models, see “Train Classification Models in Classification Learner App” on page 23-10.

Support Vector Machines

In Classification Learner, you can train SVMs when your data has two or more classes.

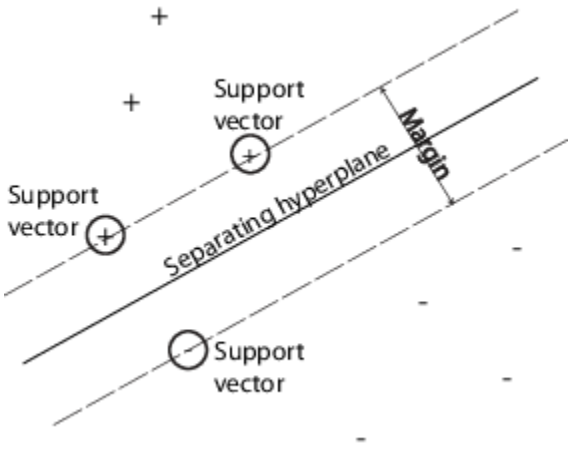
Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Linear SVM 	Binary: Fast Multiclass: Medium	Medium	Easy	Low Makes a simple linear separation between classes.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Quadratic SVM 	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	Medium
Cubic SVM 	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	Medium
Fine Gaussian SVM 	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	High — decreases with kernel scale setting. Makes finely detailed distinctions between classes, with kernel scale set to `sqrt(P)/4`.
Medium Gaussian SVM 	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	Medium Medium distinctions, with kernel scale set to `sqrt(P)`.
Coarse Gaussian SVM 	Binary: Fast Multiclass: Slow	Binary: Medium Multiclass: Large	Hard	Low Makes coarse distinctions between classes, with kernel scale set to `sqrt(P)*4`, where P is the number of predictors.

Tip Try training each of the nonoptimizable support vector machine options in the **Model Type** gallery. Train them all to see which settings produce the best model with your data. Select the best model in the **Models** pane. To try to improve your model, try feature selection, and then try changing some advanced options.

An SVM classifies data by finding the best hyperplane that separates data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyperplane that has no interior data points.

The support vectors are the data points that are closest to the separating hyperplane; these points are on the boundary of the slab. The following figure illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.



SVMs can also use a soft margin, meaning a hyperplane that separates many, but not all data points.

For an example, see “Train Support Vector Machines Using Classification Learner App” on page 23-98.

Advanced SVM Options

If you have exactly two classes, Classification Learner uses the `fitcsvm` function to train the classifier. If you have more than two classes, the app uses the `fitcecoc` function to reduce the multiclass classification problem to a set of binary classification subproblems, with one SVM learner for each subproblem. To examine the code for the binary and multiclass classifier types, you can generate code from your trained classifiers in the app.

You can set these options in the app:

- **Kernel function**

  Specify the Kernel function to compute the classifier.

  - Linear kernel, easiest to interpret
  - Gaussian or Radial Basis Function (RBF) kernel
  - Quadratic
  - Cubic
- **Box constraint level**

  Specify the box constraint to keep the allowable values of the Lagrange multipliers in a box, a bounded region.

  To tune your SVM classifier, try increasing the box constraint level. Click the buttons or enter a positive scalar value in the **Box constraint level** box. Increasing the box constraint level can decrease the number of support vectors, but also can increase training time.

  The Box Constraint parameter is the soft-margin penalty known as C in the primal equations, and is a hard “box” constraint in the dual equations.
- **Kernel scale mode**

  Specify manual kernel scaling if desired.

When you set **Kernel scale mode** to `Auto`, then the software uses a heuristic procedure to select the scale value. The heuristic procedure uses subsampling. Therefore, to reproduce results, set a random number seed using `rng` before training the classifier.

When you set **Kernel scale mode** to `Manual`, you can specify a value. Click the buttons or enter a positive scalar value in the **Manual kernel scale** box. The software divides all elements of the predictor matrix by the value of the kernel scale. Then, the software applies the appropriate kernel norm to compute the Gram matrix.

Tip Model flexibility decreases with the kernel scale setting.

- **Multiclass method**

  Only for data with 3 or more classes. This method reduces the multiclass classification problem to a set of binary classification subproblems, with one SVM learner for each subproblem. `One-vs-One` trains one learner for each pair of classes. It learns to distinguish one class from the other. `One-vs-All` trains one learner for each class. It learns to distinguish one class from all others.

- **Standardize data**

  Specify whether to scale each coordinate distance. If predictors have widely different scales, standardizing can improve the fit.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

Nearest Neighbor Classifiers

Nearest neighbor classifiers typically have good predictive accuracy in low dimensions, but might not in high dimensions. They have high memory usage, and are not easy to interpret.

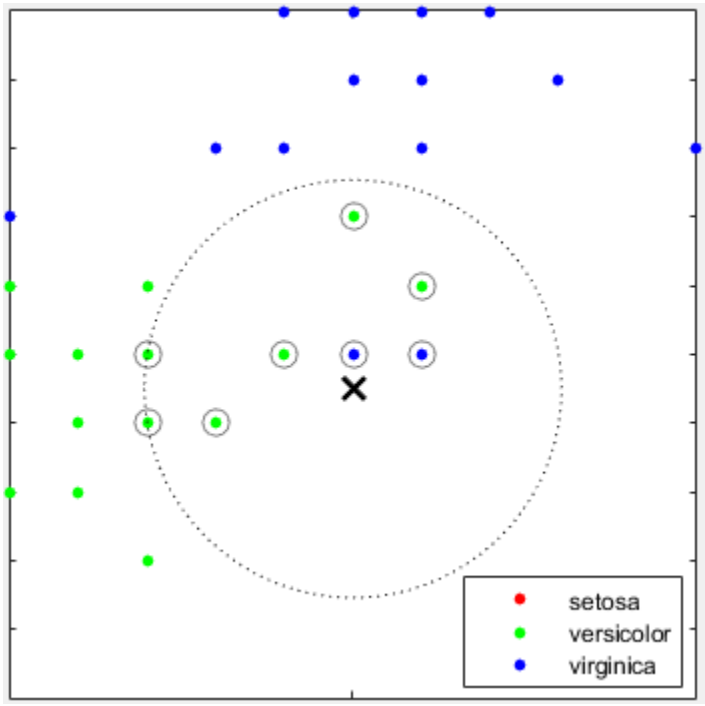
Tip Model flexibility decreases with the **Number of neighbors** setting.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Fine KNN 	Medium	Medium	Hard	Finely detailed distinctions between classes. The number of neighbors is set to 1.
Medium KNN 	Medium	Medium	Hard	Medium distinctions between classes. The number of neighbors is set to 10.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Coarse KNN 	Medium	Medium	Hard	Coarse distinctions between classes. The number of neighbors is set to 100.
Cosine KNN 	Medium	Medium	Hard	Medium distinctions between classes, using a Cosine distance metric. The number of neighbors is set to 10.
Cubic KNN 	Slow	Medium	Hard	Medium distinctions between classes, using a cubic distance metric. The number of neighbors is set to 10.
Weighted KNN 	Medium	Medium	Hard	Medium distinctions between classes, using a distance weight. The number of neighbors is set to 10.

Tip Try training each of the nonoptimizable nearest neighbor options in the **Model Type** gallery. Train them all to see which settings produce the best model with your data. Select the best model in the **Models** pane. To try to improve your model, try feature selection, and then (optionally) try changing some advanced options.

What is *k*-Nearest Neighbor classification? Categorizing query points based on their distance to points (or neighbors) in a training dataset can be a simple yet effective way of classifying new points. You can use various metrics to determine the distance. Given a set X of n points and a distance function, *k*-nearest neighbor (*k*NN) search lets you find the k closest points in X to a query point or set of points. *k*NN-based algorithms are widely used as benchmark machine learning rules.



For an example, see “Train Nearest Neighbor Classifiers Using Classification Learner App” on page 23-101.

Advanced KNN Options

Nearest Neighbor classifiers in Classification Learner use the `fitcknn` function. You can set these options:

- **Number of neighbors**

  Specify the number of nearest neighbors to find for classifying each point when predicting. Specify a fine (low number) or coarse classifier (high number) by changing the number of neighbors. For example, a fine KNN uses one neighbor, and a coarse KNN uses 100. Many neighbors can be time consuming to fit.
- **Distance metric**

  You can use various metrics to determine the distance to points. For definitions, see the class `ClassificationKNN`.
- **Distance weight**

  Specify the distance weighting function. You can choose `Equal` (no weights), `Inverse` (weight is 1/distance), or `Squared Inverse` (weight is $1/\text{distance}^2$).
- **Standardize data**

  Specify whether to scale each coordinate distance. If predictors have widely different scales, standardizing can improve the fit.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

Ensemble Classifiers

Ensemble classifiers meld results from many weak learners into one high-quality ensemble model. Qualities depend on the choice of algorithm.

Tip Model flexibility increases with the **Number of learners** setting.

All ensemble classifiers tend to be slow to fit because they often need many learners.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Ensemble Method	Model Flexibility
Boosted Trees 	Fast	Low	Hard	`AdaBoost`, with `Decision Tree` learners	Medium to high — increases with **Number of learners** or **Maximum number of splits** setting. **Tip** Boosted trees can usually do better than bagged, but might require parameter tuning and more learners
Bagged Trees 	Medium	High	Hard	Random forest `Bag`, with `Decision Tree` learners	High — increases with **Number of learners** setting. **Tip** Try this classifier first.
Subspace Discriminant 	Medium	Low	Hard	`Subspace`, with `Discriminant` learners	Medium — increases with **Number of learners** setting. Good for many predictors
Subspace KNN 	Medium	Medium	Hard	`Subspace`, with `Nearest Neighbor` learners	Medium — increases with **Number of learners** setting. Good for many predictors

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Ensemble Method	Model Flexibility
RUSBoost Trees 	Fast	Low	Hard	`RUSBoost`, with `Decision Tree` learners	Medium — increases with **Number of learners** or **Maximum number of splits** setting. Good for skewed data (with many more observations of 1 class)
GentleBoost or LogitBoost — not available in the **Model Type** gallery. If you have 2 class data, select manually.	Fast	Low	Hard	`GentleBoost` or `LogitBoost`, with `Decision Tree` learners Choose `Boosted Trees` and change to `GentleBoost` method.	Medium — increases with **Number of learners** or **Maximum number of splits** setting. For binary classification only

Bagged trees use Breiman's `'random forest'` algorithm. For reference, see Breiman, L. Random Forests. Machine Learning 45, pp. 5–32, 2001.

Tips

- Try bagged trees first. Boosted trees can usually do better but might require searching many parameter values, which is time-consuming.
- Try training each of the nonoptimizable ensemble classifier options in the **Model Type** gallery. Train them all to see which settings produce the best model with your data. Select the best model in the **Models** pane. To try to improve your model, try feature selection, PCA, and then (optionally) try changing some advanced options.
- For boosting ensemble methods, you can get fine detail with either deeper trees or larger numbers of shallow trees. As with single tree classifiers, deep trees can cause overfitting. You need to experiment to choose the best tree depth for the trees in the ensemble, in order to trade-off data fit with tree complexity. Use the **Number of learners** and **Maximum number of splits** settings.

For an example, see “Train Ensemble Classifiers Using Classification Learner App” on page 23-104.

Advanced Ensemble Options

Ensemble classifiers in Classification Learner use the `fitcensemble` function. You can set these options:

- For help choosing **Ensemble method** and **Learner type**, see the Ensemble table. Try the presets first.
- **Maximum number of splits**

  For boosting ensemble methods, specify the maximum number of splits or branch points to control the depth of your tree learners. Many branches tend to overfit, and simpler trees can be more

robust and easy to interpret. Experiment to choose the best tree depth for the trees in the ensemble.

- **Number of learners**

  Try changing the number of learners to see if you can improve the model. Many learners can produce high accuracy, but can be time consuming to fit. Start with a few dozen learners, and then inspect the performance. An ensemble with good predictive power can need a few hundred learners.

- **Learning rate**

  Specify the learning rate for shrinkage. If you set the learning rate to less than 1, the ensemble requires more learning iterations but often achieves better accuracy. 0.1 is a popular choice.

- **Subspace dimension**

  For subspace ensembles, specify the number of predictors to sample in each learner. The app chooses a random subset of the predictors for each learner. The subsets chosen by different learners are independent.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Classification Learner App” on page 23-54.

Neural Network Classifiers

Neural network models typically have good predictive accuracy and can be used for multiclass classification; however, they are not easy to interpret.

Model flexibility increases with the size and number of fully connected layers in the neural network.

Tip In the **Model Type** gallery, click **All Neural Networks**  to try each of the preset neural network options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Narrow Neural Network 	Fast	Medium	Hard	Medium — increases with **Size of each fully connected layer** setting
Medium Neural Network 	Fast	Medium	Hard	Medium — increases with **Size of each fully connected layer** setting

Classifier Type	Prediction Speed	Memory Usage	Interpretability	Model Flexibility
Wide Neural Network 	Fast	Medium	Hard	Medium — increases with **Size of each fully connected layer** setting
Bilayered Neural Network 	Fast	Medium	Hard	High — increases with **Size of each fully connected layer** setting
Trilayered Neural Network 	Fast	Medium	Hard	High — increases with **Size of each fully connected layer** setting

Each model is a feedforward, fully connected neural network for classification. The first fully connected layer of the neural network has a connection from the network input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer. The final fully connected layer and the subsequent softmax activation function produce the network's output, namely classification scores (posterior probabilities) and predicted labels. For more information, see “Neural Network Structure” on page 33-1823.

For an example, see “Train Neural Network Classifiers Using Classification Learner App” on page 23-117.

Advanced Neural Network Options

Neural network classifiers in Classification Learner use the `fitcnet` function. You can set these options:

- **Number of fully connected layers** — Specify the number of fully connected layers in the neural network, excluding the final fully connected layer for classification. You can choose a maximum of three fully connected layers.
- **Size of each fully connected layer** — Specify the size of each fully connected layer, excluding the final fully connected layer. If you choose to create a neural network with multiple fully connected layers, consider specifying layers with decreasing sizes.
- **Activation** — Specify the activation function for all fully connected layers, excluding the final fully connected layer. The activation function for the last fully connected layer is always softmax. Choose from the following activation functions: `ReLU`, `Tanh`, `Sigmoid`, and `None`.
- **Iteration limit** — Specify the maximum number of training iterations.
- **Regularization term strength** — Specify the ridge (L2) regularization penalty term.
- **Standardize data** — Specify whether to standardize the numeric predictors. If predictors have widely different scales, standardizing can improve the fit. Standardizing the data is highly recommended.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

Feature Selection and Feature Transformation Using Classification Learner App

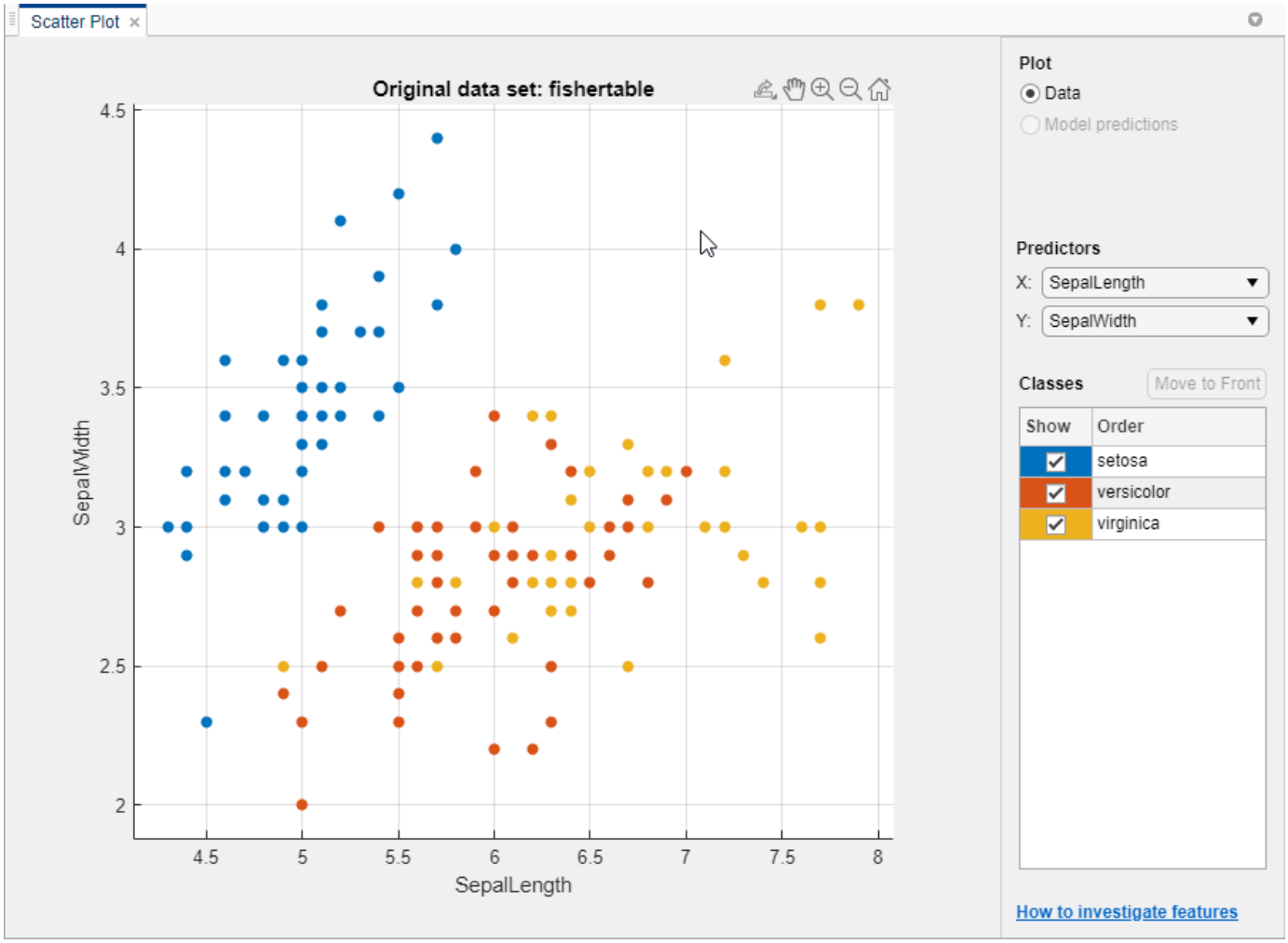
In this section...

Investigate Features in the Scatter Plot

In Classification Learner, try to identify predictors that separate classes well by plotting different pairs of predictors on the scatter plot. The plot can help you investigate features to include or exclude. You can visualize training data and misclassified points on the scatter plot.

Before you train a classifier, the scatter plot shows the data. If you have trained a classifier, the scatter plot shows model prediction results. Switch to plotting only the data by selecting **Data** in the **Plot** controls.

- Choose features to plot using the **X** and **Y** lists under **Predictors**.
- Look for predictors that separate classes well. For example, plotting the `fisheriris` data, you can see that sepal length and sepal width separate one of the classes well (`setosa`). You need to plot other predictors to see if you can separate the other two classes.



- Show or hide specific classes using the check boxes under **Show**.
- Change the stacking order of the plotted classes by selecting a class under **Classes** and then clicking **Move to Front**.
- Investigate finer details by zooming in and out and panning across the plot. To enable zooming or panning, hover the mouse over the scatter plot and click the corresponding button on the toolbar that appears above the top right of the plot.
- If you identify predictors that are not useful for separating out classes, then try using **Feature Selection** to remove them and train classifiers including only the most useful predictors.

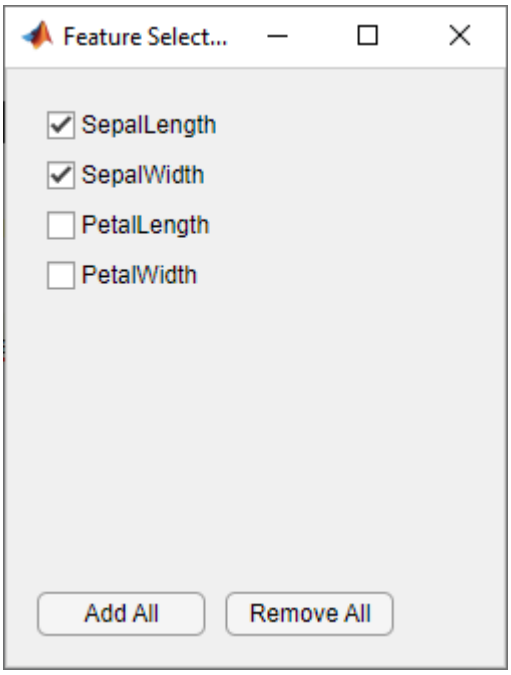
After you train a classifier, the scatter plot shows model prediction results. You can show or hide correct or incorrect results and visualize the results by class. See “Plot Classifier Results” on page 23-66.

You can export the scatter plots you create in the app to figures. See “Export Plots in Classification Learner App” on page 23-72.

Select Features to Include

In Classification Learner, you can specify different features (or predictors) to include in the model. See if you can improve models by removing features with low predictive power. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily without some predictors.

1. On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**.
2. In the Feature Selection dialog box, clear the check boxes for the predictors you want to exclude.



Tip You can close the Feature Selection dialog box, or move it. Your choices in the dialog box remain.

3. Click **Train** to train a new model using the new predictor options.
4. Observe the new model in the **Models** pane. The **Current Model Summary** pane displays how many predictors are excluded.
5. To check which predictors are included in a trained model, click the model in the **Models** pane and observe the check boxes in the Feature Selection dialog box.
6. You can try to improve the model by including different features in the model.

For an example using feature selection, see “Train Decision Trees Using Classification Learner App” on page 23-83.

Transform Features with PCA in Classification Learner

Use principal component analysis (PCA) to reduce the dimensionality of the predictor space. Reducing the dimensionality can create classification models in Classification Learner that help prevent overfitting. PCA linearly transforms predictors in order to remove redundant dimensions, and generates a new set of variables called principal components.

1. On the **Classification Learner** tab, in the **Features** section, select **PCA**.
2. In the Advanced PCA Options dialog box, select the **Enable PCA** check box.

   You can close the PCA dialog box, or move it. Your choices in the dialog box remain.
3. When you next click **Train**, the `pca` function transforms your selected features before training the classifier.
4. By default, PCA keeps only the components that explain 95% of the variance. In the PCA dialog box, you can change the percentage of variance to explain by selecting the **Explained variance** value. A higher value risks overfitting, while a lower value risks removing useful dimensions.
5. If you want to manually limit the number of PCA components, in the **Component reduction criterion** list, select `Specify number of components`. Select the **Number of numeric components** value. The number of components cannot be larger than the number of numeric predictors. PCA is not applied to categorical predictors.

Check PCA options for trained models in the **Current Model Summary** pane information. Check the explained variance percentages to decide whether to change the number of components. For example:

```
PCA is keeping enough components to explain 95% variance.
After training, 2 components were kept.
Explained variance per component (in order): 92.5%, 5.3%, 1.7%, 0.5%
```

To learn more about how Classification Learner applies PCA to your data, generate code for your trained classifier. For more information on PCA, see the `pca` function.

Investigate Features in the Parallel Coordinates Plot

To investigate features to include or exclude, use the parallel coordinates plot. You can visualize high-dimensional data on a single plot to see 2-D patterns. The plot can help you understand relationships between features and identify useful predictors for separating classes. You can visualize training data and misclassified points on the parallel coordinates plot. When you plot classifier results, misclassified points have dashed lines.

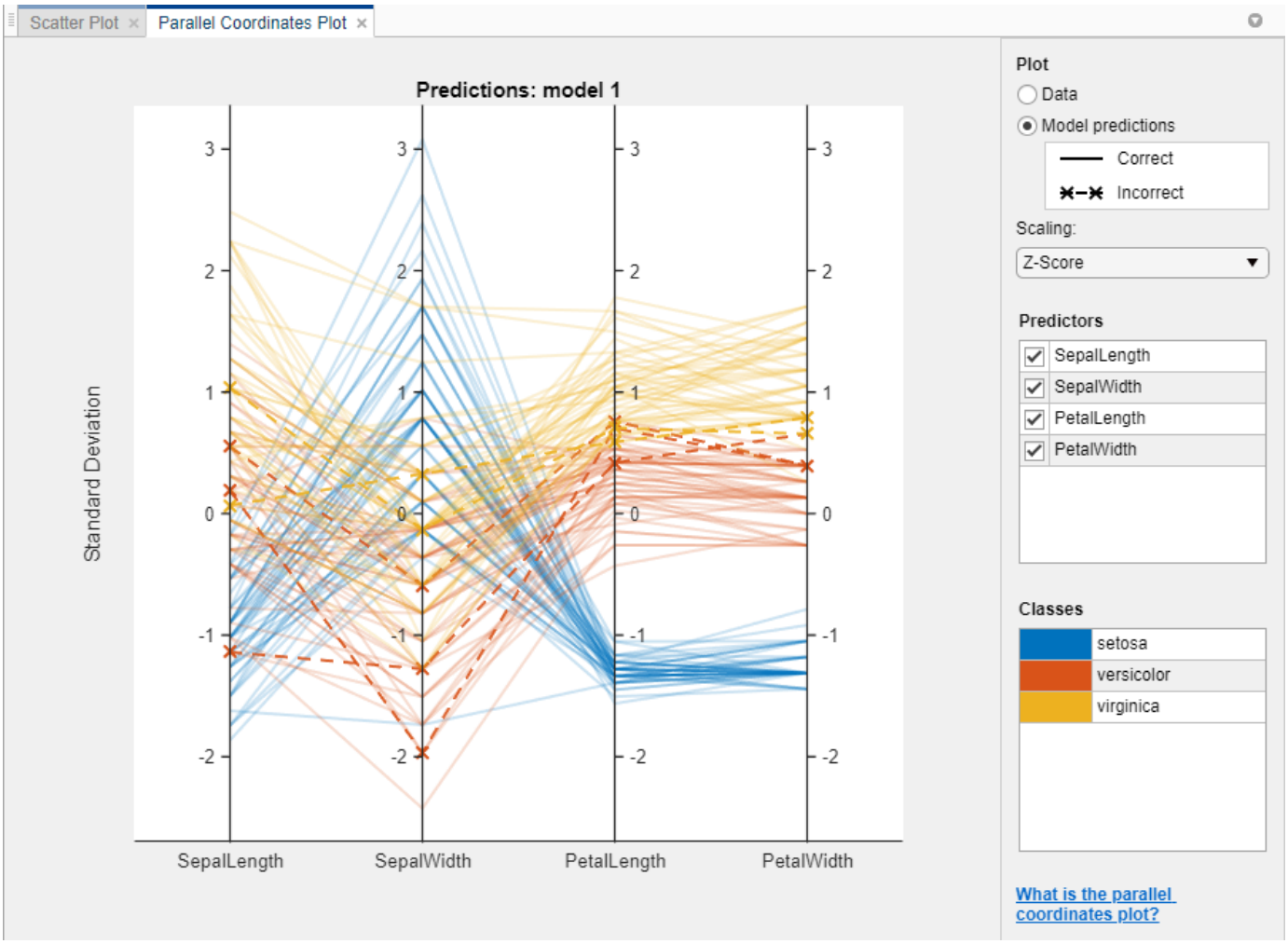
1. On the **Classification Learner** tab, in the **Plots** section, click **Parallel Coordinates**.
2. On the plot, drag the *X* tick labels to reorder the predictors. Changing the order can help you identify predictors that separate classes well.
3. To specify which predictors to plot, use the **Predictors** check boxes. A good practice is to plot a few predictors at a time. If your data has many predictors, the plot shows the first 10 predictors by default.
4. If the predictors have significantly different scales, scale the data for easier visualization. Try different options in the **Scaling** list:

   - `None` displays raw data along coordinate rulers that have the same minimum and maximum limits.
   - `Range` displays raw data along coordinate rulers that have independent minimum and maximum limits.
   - `Z-Score` displays z-scores (with a mean of 0 and a standard deviation of 1) along each coordinate ruler.
   - `Zero Mean` displays data centered to have a mean of 0 along each coordinate ruler.

- `Unit Variance` displays values scaled by standard deviation along each coordinate ruler.
- `L2 Norm` displays 2-norm values along each coordinate ruler.

5 If you identify predictors that are not useful for separating out classes, use **Feature Selection** to remove them and train classifiers including only the most useful predictors.

The plot of the `fisheriris` data shows the petal length and petal width features separate the classes best.



You can export the parallel coordinates plots you create in the app to figures. See “Export Plots in Classification Learner App” on page 23-72.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22

Misclassification Costs in Classification Learner App

In this section...

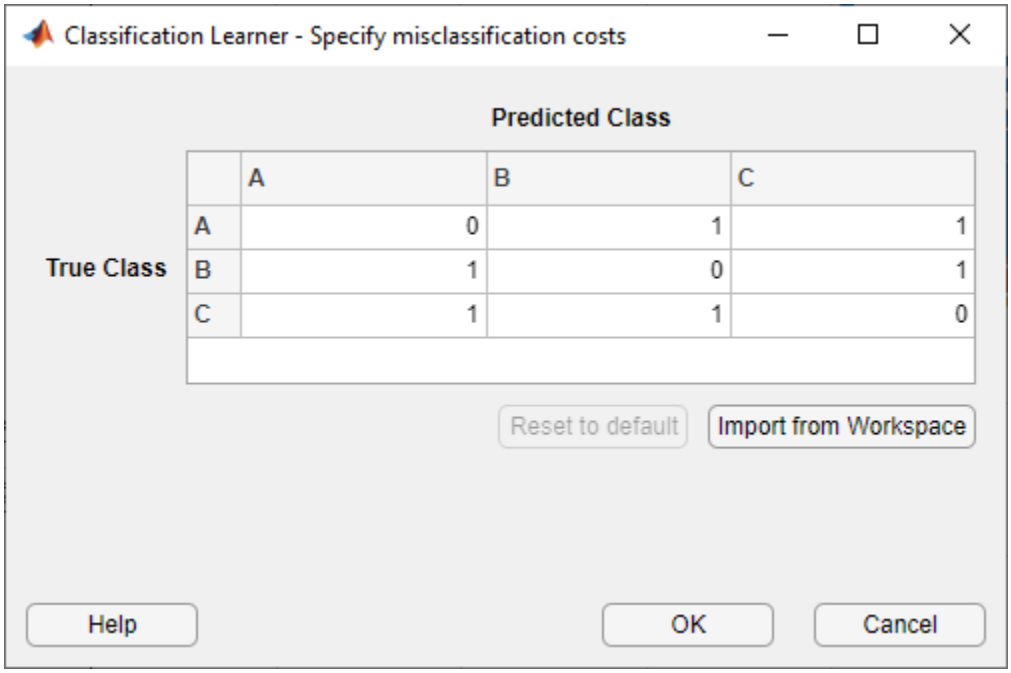
By default, the Classification Learner app creates models that assign the same penalty to all misclassifications during training. For a given observation, the app assigns a penalty of 0 if the observation is classified correctly and a penalty of 1 if the observation is classified incorrectly. In some cases, this assignment is inappropriate. For example, suppose you want to classify patients as either healthy or sick. The cost of misclassifying a sick person as healthy might be five times the cost of misclassifying a healthy person as sick. For cases where you know the cost of misclassifying observations of one class into another, and the costs vary across the classes, specify the misclassification costs before training your models.

Note Custom misclassification costs are not supported for logistic regression and neural network models.

Specify Misclassification Costs

In the Classification Learner app, in the **Options** section of the **Classification Learner** tab, select **Misclassification Costs**. The app opens a dialog box that shows the default misclassification costs (cost matrix) as a table with row and column labels determined by the classes in the response variable. The rows of the table correspond to the true classes, and the columns correspond to the predicted classes. You can interpret the cost matrix in this way: the entry in row i and column j is the cost of misclassifying ith class observations into the jth class. The diagonal entries of the cost matrix must be 0, and the off-diagonal entries must be nonnegative real numbers.

You can specify your own misclassification costs in two ways: by entering values directly into the table in the dialog box or by importing a workspace variable that contains the cost values.



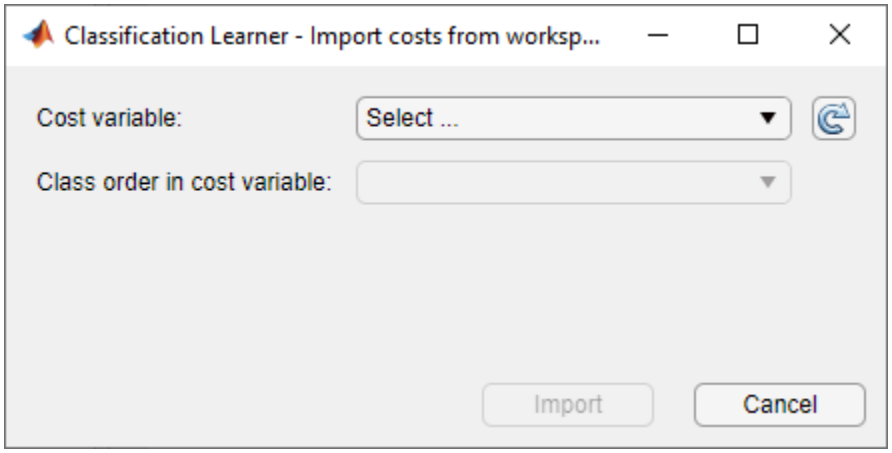
Note A scaled version of the cost matrix gives the same classification results (for example, confusion matrix and accuracy), but with a different total misclassification cost. That is, if `CostMat` is the misclassification cost matrix and `a` is a positive, real scalar, then a model trained with the cost matrix `a*CostMat` has the same confusion matrix as that model trained with `CostMat`.

Enter Costs Directly in Dialog Box

In the misclassification costs dialog box, double-click an entry in the table that you want to edit. Delete the value and type the correct misclassification cost for the entry. When you are done editing the table, click **OK** to save your changes.

Import Workspace Variable Containing Costs

In the misclassification costs dialog box, click **Import from Workspace**. The app opens a dialog box for importing costs from a variable in the MATLAB workspace.



From the **Cost variable** list, select the cost matrix or structure that contains the misclassification costs.

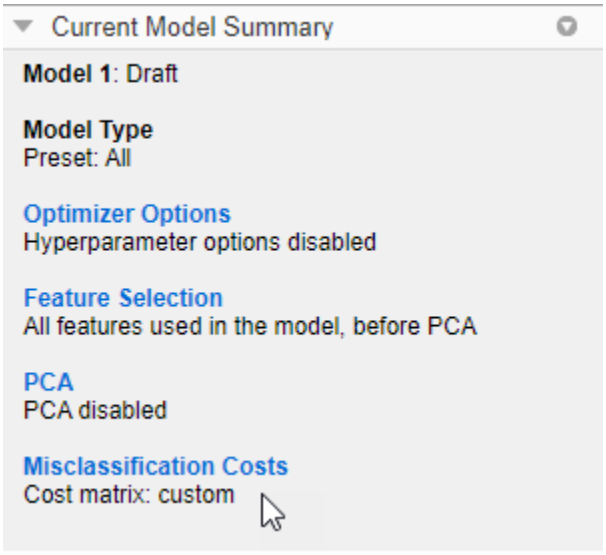
- Cost matrix – The matrix must contain the misclassification costs. The diagonal entries must be 0, and the off-diagonal entries must be nonnegative real numbers. By default, the app uses the class order shown in the previous misclassification costs dialog box to interpret the cost matrix values.

  To specify the order of the classes in the cost matrix, create a separate workspace variable containing the class names in the correct order. In the import dialog box, select the appropriate variable from the **Class order in cost variable** list. The workspace variable containing the class names must be a categorical vector, logical vector, numeric vector, string array, or cell array of character vectors. The class names must match (in spelling and capitalization) the class names in the response variable.
- Structure – The structure must contain the fields `ClassificationCosts` and `ClassNames` with these specifications:

  - `ClassificationCosts` – Matrix that contains misclassification costs.
  - `ClassNames` – Names of the classes. The order of the classes in `ClassNames` determines the order of the rows and columns of `ClassificationCosts`. The variable `ClassNames` must be a categorical vector, logical vector, numeric vector, string array, or cell array of character vectors. The class names must match (in spelling and capitalization) the class names in the response variable.

After specifying the cost variable and the class order in the cost variable, click **Import**. The app updates the table in the misclassification costs dialog box.

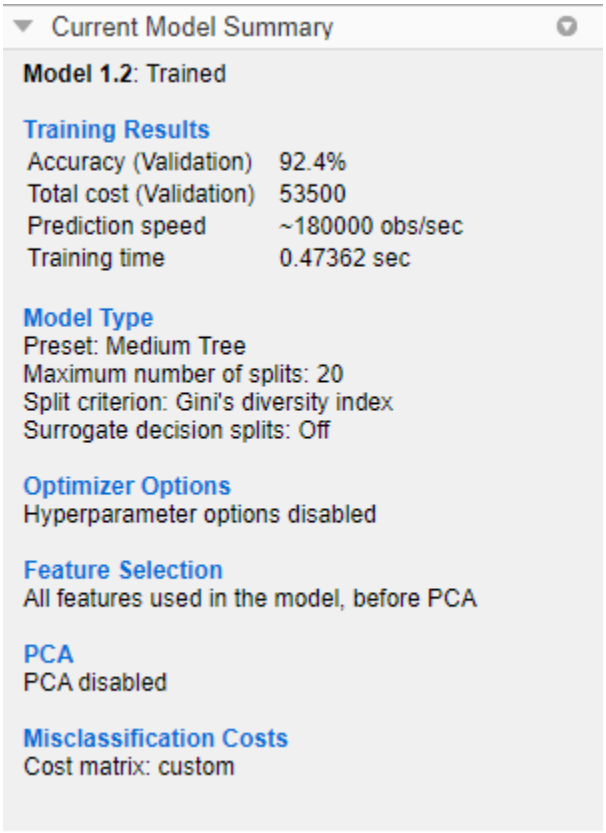
After you specify a cost matrix that differs from the default, the app updates the **Current Model Summary** pane for new models. In the **Current Model Summary** pane, under **Misclassification Costs**, the app lists the cost matrix as "custom". For models that use the default misclassification costs, the app lists the cost matrix as "default".



Assess Model Performance

After specifying misclassification costs, you can train and tune your models as usual. However, using custom misclassification costs can change how you assess the performance of a model. For example, instead of choosing the model with the best accuracy, choose a model that has good accuracy and a low total misclassification cost. The total misclassification cost for a model is `sum(CostMat.*ConfusionMat,'all')`, where `CostMat` is the misclassification cost matrix and `ConfusionMat` is the confusion matrix for the model. The confusion matrix shows how the model classifies observations in each class. See “Check Performance Per Class in the Confusion Matrix” on page 23-67.

To inspect the total misclassification cost of a trained model, select the model in the **Models** pane. In the **Current Model Summary** pane, look at the **Training Results** section. The total misclassification cost is listed below the accuracy of the model.



Misclassification Costs in Exported Model and Generated Code

After you train a model with custom misclassification costs and export it from the app, you can find the custom costs inside the exported model. For example, if you export a tree model as a structure named `trainedModel`, you can use the following code to access the cost matrix and the order of the classes in the matrix.

```
trainedModel.ClassificationTree.Cost
trainedModel.ClassificationTree.ClassNames
```

For ensemble and binary SVM models, the app uses the misclassification costs to adjust the model prior class probabilities. Therefore, the `Cost` property of the exported model is reset to the default cost matrix, but the `Prior` property is updated.

When you generate MATLAB code for a model trained with custom misclassification costs, the generated code includes a cost matrix that is passed to the training function through the `'Cost'` name-value pair argument.

See Also

Related Examples

- “Train and Compare Classifiers Using Misclassification Costs in Classification Learner App” on page 23-120

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Choose Classifier Options” on page 23-22
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

Hyperparameter Optimization in Classification Learner App

In this section...

After you choose a particular type of model to train, for example a decision tree or a support vector machine (SVM), you can tune your model by selecting different advanced options. For example, you can change the maximum number of splits for a decision tree or the box constraint of an SVM. Some of these options are internal parameters of the model, or hyperparameters, that can strongly affect its performance. Instead of manually selecting these options, you can use hyperparameter optimization within the Classification Learner app to automate the selection of hyperparameter values. For a given model type, the app tries different combinations of hyperparameter values by using an optimization scheme that seeks to minimize the model classification error, and returns a model with the optimized hyperparameters. You can use the resulting model as you would any other trained model.

Note Because hyperparameter optimization can lead to an overfitted model, the recommended approach is to create a separate test set before importing your data into the Classification Learner app. After you train your optimizable model, you can see how it performs on your test set. For an example, see “Train Classifier Using Hyperparameter Optimization in Classification Learner App” on page 23-128.

To perform hyperparameter optimization in Classification Learner, follow these steps:

1. Choose a model type and decide which hyperparameters to optimize. See “Select Hyperparameters to Optimize” on page 23-54.

   Note Hyperparameter optimization is not supported for logistic regression or neural network models.

2. (Optional) Specify how the optimization is performed. For more information, see “Optimization Options” on page 23-59.
3. Train your model. Use the “Minimum Classification Error Plot” on page 23-61 to track the optimization results.
4. Inspect your trained model. See “Optimization Results” on page 23-62.

Select Hyperparameters to Optimize

In the Classification Learner app, in the **Model Type** section of the **Classification Learner** tab, click the arrow to open the gallery. The gallery includes optimizable models that you can train using hyperparameter optimization.

After you select an optimizable model, you can choose which of its hyperparameters you want to optimize. In the **Model Type** section, select **Advanced > Advanced**. The app opens a dialog box in which you can select **Optimize** check boxes for the hyperparameters that you want to optimize.

Under **Values**, specify the fixed values for the hyperparameters that you do not want to optimize or that are not optimizable.

This table describes the hyperparameters that you can optimize for each type of model and the search range of each hyperparameter. It also includes the additional hyperparameters for which you can specify fixed values.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
Optimizable Tree 	• **Maximum number of splits** – The software searches among integers log-scaled in the range `[1,max(2,n–1)]`, where `n` is the number of observations. • **Split criterion** – The software searches among `Gini's diversity index`, `Twoing rule`, and `Maximum deviance reduction`.	• **Surrogate decision splits** • **Maximum surrogates per node**	For more information, see “Advanced Tree Options” on page 23-28.
Optimizable Discriminant 	• **Discriminant type** – The software searches among `Linear`, `Quadratic`, `Diagonal Linear`, and `Diagonal Quadratic`.		• The **Discriminant type** optimizable hyperparameter combines the preset model types (**Linear Discriminant** and **Quadratic Discriminant**) with the **Covariance structure** advanced option of the preset models. For more information, see “Advanced Discriminant Options” on page 23-30.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
Optimizable Naive Bayes 	• **Distribution names** – The software searches between `Gaussian` and `Kernel`. • **Kernel type** – The software searches among `Gaussian`, `Box`, `Epanechnikov`, and `Triangle`.	• **Support**	• The `Gaussian` value of the **Distribution names** optimizable hyperparameter specifies a **Gaussian Naive Bayes** model. Similarly, the `Kernel` **Distribution names** value specifies a **Kernel Naive Bayes** model. For more information, see “Advanced Naive Bayes Options” on page 23-31.
Optimizable SVM 	• **Kernel function** – The software searches among `Gaussian`, `Linear`, `Quadratic`, and `Cubic`. • **Box constraint level** – The software searches among positive values log-scaled in the range `[0.001,1000]`. • **Kernel scale** – The software searches among positive values log-scaled in the range `[0.001,1000]`. • **Multiclass method** – The software searches between `One-vs-One` and `One-vs-All`. • **Standardize data** – The software searches between `true` and `false`.		• The **Kernel scale** optimizable hyperparameter combines the **Kernel scale mode** and **Manual kernel scale** advanced options of the preset SVM models. • You can optimize the **Kernel scale** optimizable hyperparameter only when the **Kernel function** value is `Gaussian`. Unless you specify a value for **Kernel scale** by clearing the **Optimize** check box, the app uses the **Manual** value of `1` by default when the **Kernel function** has a value other than `Gaussian`. For more information, see “Advanced SVM Options” on page 23-33.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes

Optimizable KNN 	• **Number of neighbors** – The software searches among integers log-scaled in the range `[1,max(2,round(n/2))]`, where `n` is the number of observations.
	• **Distance metric** – The software searches among:
	  • `Euclidean`
	  • `City block`
	  • `Chebyshev`
	  • `Minkowski (cubic)`
	  • `Mahalanobis`
	  • `Cosine`
	  • `Correlation`
	  • `Spearman`
	  • `Hamming`
	  • `Jaccard`
	• **Distance weight** – The software searches among `Equal`, `Inverse`, and `Squared inverse`.
	• **Standardize** – The software searches between `true` and `false`.
			For more information, see “Advanced KNN Options” on page 23-36.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
Optimizable Ensemble 	• **Ensemble method** – The software searches among `AdaBoost`, `RUSBoost`, `LogitBoost`, `GentleBoost`, and `Bag`. • **Maximum number of splits** – The software searches among integers log-scaled in the range `[1,max(2,n–1)]`, where `n` is the number of observations. • **Number of learners** – The software searches among integers log-scaled in the range `[10,500]`. • **Learning rate** – The software searches among real values log-scaled in the range `[0.001,1]`. • **Number of predictors to sample** – The software searches among integers in the range `[1,max(2,p)]`, where `p` is the number of predictor variables.	• **Learner type**	• The `AdaBoost`, `LogitBoost`, and `GentleBoost` values of the **Ensemble method** optimizable hyperparameter specify a **Boosted Trees** model. Similarly, the `RUSBoost` **Ensemble method** value specifies an **RUSBoosted Trees** model, and the `Bag` **Ensemble method** value specifies a **Bagged Trees** model. • The `LogitBoost` and `GentleBoost` values are available only for binary classification. • The **Number of predictors to sample** optimizable hyperparameter is not available in the advanced options of the preset ensemble models. • You can optimize the **Number of predictors to sample** optimizable hyperparameter only when the **Ensemble method** value is `Bag`. Unless you specify a value for **Number of predictors to sample** by clearing the **Optimize** check box, the app uses the default value of **Select All** when the

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
			Ensemble method has a value other than `Bag`. For more information, see “Advanced Ensemble Options” on page 23-38.

Optimization Options

By default, the Classification Learner app performs hyperparameter tuning by using Bayesian optimization. The goal of Bayesian optimization, and optimization in general, is to find a point that minimizes an objective function. In the context of hyperparameter tuning in the app, a point is a set of hyperparameter values, and the objective function is the loss function, or the classification error. For more information on the basics of Bayesian optimization, see “Bayesian Optimization Workflow” on page 10-25.

You can specify how the hyperparameter tuning is performed. For example, you can change the optimization method to grid search or limit the training time. On the **Classification Learner** tab, in the **Model Type** section, select **Advanced > Optimizer Options**. The app opens a dialog box in which you can select optimization options.

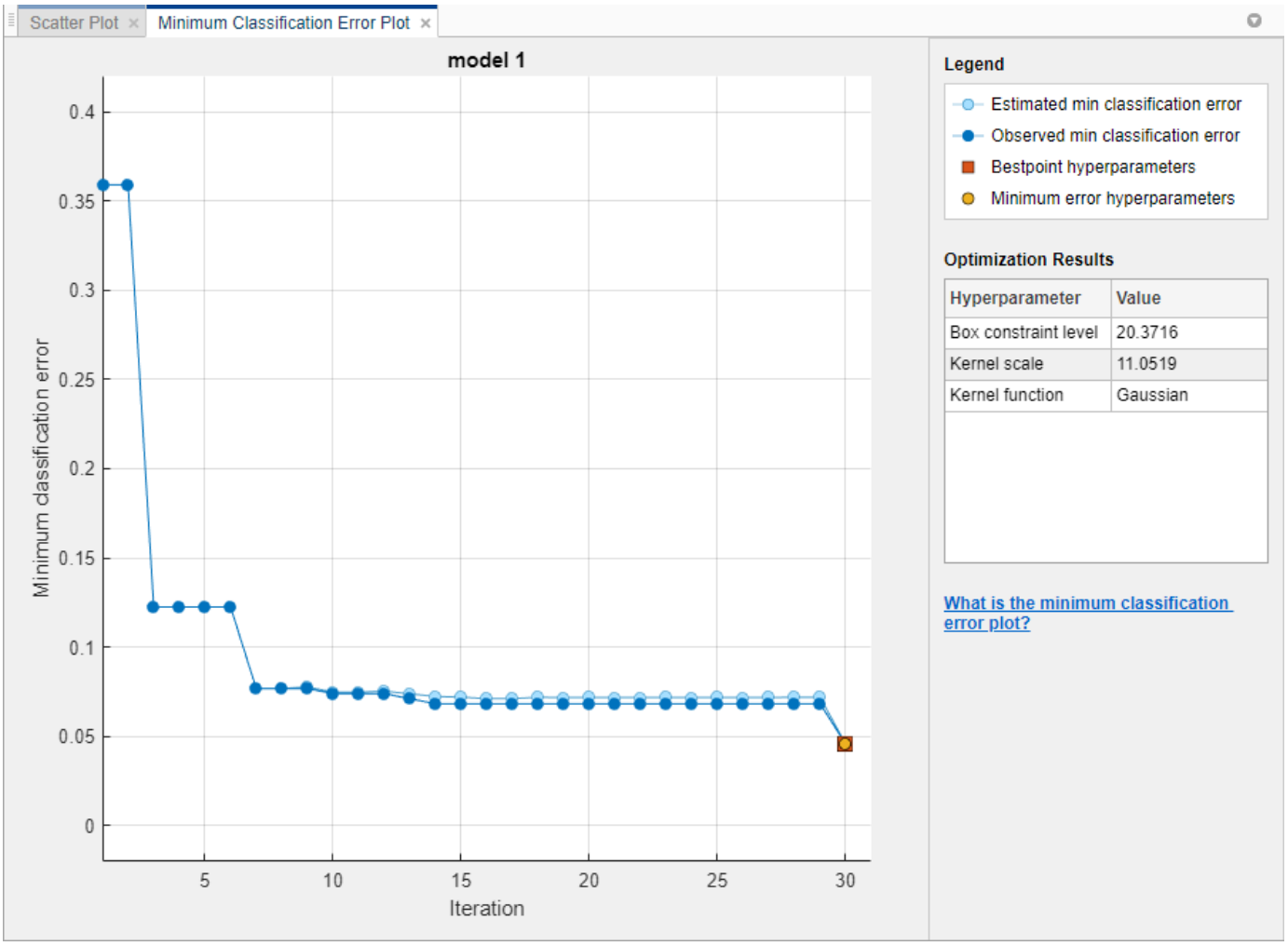
This table describes the available optimization options and their default values.

Option	Description
Optimizer	The optimizer values are: • `Bayesopt` (default) – Use Bayesian optimization. Internally, the app calls the `bayesopt` function. • `Grid search` – Use grid search with the number of values per dimension determined by the **Number of grid divisions** value. The app searches in a random order, using uniform sampling without replacement from the grid. • `Random search` – Search at random among points, where the number of points corresponds to the **Iterations** value.

Option	Description
Acquisition function	When the app performs Bayesian optimization for hyperparameter tuning, it uses the acquisition function to determine the next set of hyperparameter values to try. The acquisition function values are: • `Expected improvement per second plus` (default) • `Expected improvement` • `Expected improvement plus` • `Expected improvement per second` • `Lower confidence bound` • `Probability of improvement` For details on how these acquisition functions work in the context of Bayesian optimization, see “Acquisition Function Types” on page 10-3.
Iterations	Each iteration corresponds to a combination of hyperparameter values that the app tries. When you use Bayesian optimization or random search, specify a positive integer that sets the number of iterations. The default value is `30`. When you use grid search, the app ignores the **Iterations** value and evaluates the loss at every point in the entire grid. You can set a training time limit to stop the optimization process prematurely.
Training time limit	To set a training time limit, select this option and set the **Maximum training time in seconds** option. By default, the app does not have a training time limit.
Maximum training time in seconds	Set the training time limit in seconds as a positive real number. The default value is `300`. The run time can exceed the training time limit because this limit does not interrupt an iteration evaluation.
Number of grid divisions	When you use grid search, set a positive integer as the number of values the app tries for each numeric hyperparameter. The app ignores this value for categorical hyperparameters. The default value is `10`.

Minimum Classification Error Plot

After specifying which model hyperparameters to optimize and setting any additional optimization options (optional), train your optimizable model. On the **Classification Learner** tab, in the **Training** section, click **Train**. The app creates a **Minimum Classification Error Plot** that it updates as the optimization runs.



Note When you train an optimizable model, the app disables the **Use Parallel** button. After the training is complete, the app makes the button available again when you select a nonoptimizable model. The button is off by default.

The minimum classification error plot displays the following information:

- **Estimated minimum classification error** – Each light blue point corresponds to an estimate of the minimum classification error computed by the optimization process when considering all the sets of hyperparameter values tried so far, including the current iteration.

  The estimate is based on an upper confidence interval of the current classification error objective model, as mentioned in the **Bestpoint hyperparameters** description.

If you use grid search or random search to perform hyperparameter optimization, the app does not display these light blue points.

- **Observed minimum classification error** – Each dark blue point corresponds to the observed minimum classification error computed so far by the optimization process. For example, at the third iteration, the dark blue point corresponds to the minimum of the classification error observed in the first, second, and third iterations.
- **Bestpoint hyperparameters** – The red square indicates the iteration that corresponds to the optimized hyperparameters. You can find the values of the optimized hyperparameters listed in the upper right of the plot under **Optimization Results**.

  The optimized hyperparameters do not always provide the observed minimum classification error. When the app performs hyperparameter tuning by using Bayesian optimization (see “Optimization Options” on page 23-59 for a brief introduction), it chooses the set of hyperparameter values that minimizes an upper confidence interval of the classification error objective model, rather than the set that minimizes the classification error. For more information, see the `'Criterion','min-visited-upper-confidence-interval'` name-value pair argument of `bestPoint`.
- **Minimum error hyperparameters** – The yellow point indicates the iteration that corresponds to the hyperparameters that yield the observed minimum classification error.

  For more information, see the `'Criterion','min-observed'` name-value pair argument of `bestPoint`.

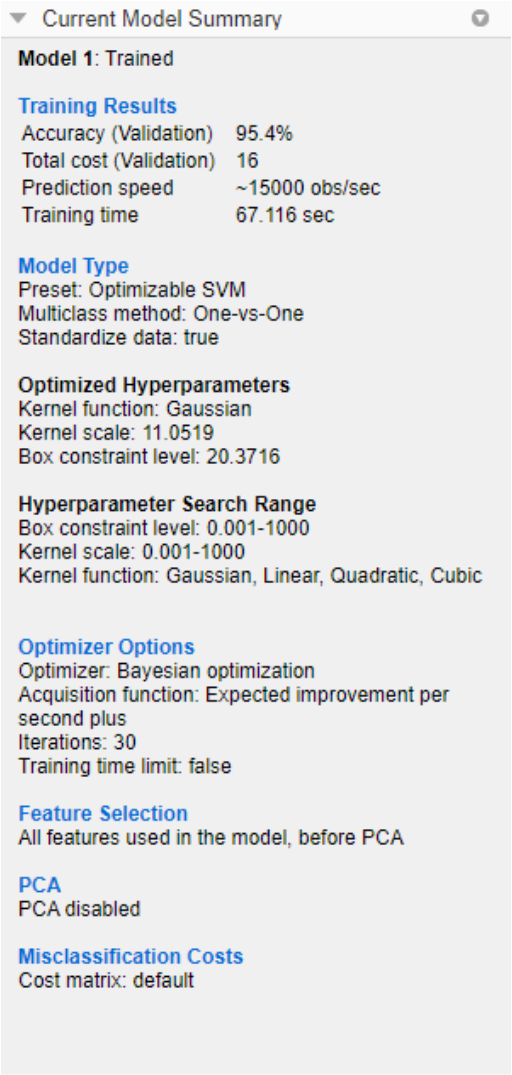
  If you use grid search to perform hyperparameter optimization, the **Bestpoint hyperparameters** and the **Minimum error hyperparameters** are the same.

Missing points in the plot correspond to `NaN` minimum classification error values.

Optimization Results

When the app finishes tuning model hyperparameters, it returns a model trained with the optimized hyperparameter values (**Bestpoint hyperparameters**). The model metrics, displayed plots, and exported model correspond to this trained model with fixed hyperparameter values.

To inspect the optimization results of a trained optimizable model, select the model in the **Models** pane and look at the **Current Model Summary** pane.



The **Current Model Summary** pane includes these sections:

- **Training Results** – Shows the performance of the optimizable model
- **Model Type** – Displays the type of optimizable model and lists any fixed hyperparameter values
- **Optimized Hyperparameters** – Lists the values of the optimized hyperparameters
- **Hyperparameter Search Range** – Displays the search ranges for the optimized hyperparameters
- **Optimizer Options** – Shows the selected optimizer options

When you perform hyperparameter tuning using Bayesian optimization and you export the resulting trained optimizable model to the workspace as a structure, the structure includes a `BayesianOptimization` object in the `HyperParameterOptimizationResult` field. The object contains the results of the optimization performed in the app.

When you generate MATLAB code from a trained optimizable model, the generated code uses the fixed and optimized hyperparameter values of the model to train on new data. The generated code does not include the optimization process. For information on how to perform Bayesian optimization when you use a fit function, see “Bayesian Optimization Using a Fit Function” on page 10-26.

See Also

Related Examples

- “Train Classifier Using Hyperparameter Optimization in Classification Learner App” on page 23-128
- “Bayesian Optimization Workflow” on page 10-25
- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

Assess Classifier Performance in Classification Learner

In this section...

After training classifiers in Classification Learner, you can compare models based on accuracy scores, visualize results by plotting class predictions, and check performance using confusion matrix and ROC curve.

- If you use *k*-fold cross-validation, then the app computes the accuracy scores using the observations in the *k* validation folds and reports the average cross-validation error. It also makes predictions on the observations in these validation folds and computes the confusion matrix and ROC curve based on these predictions.

  Note When you import data into the app, if you accept the defaults, the app automatically uses cross-validation. To learn more, see “Choose Validation Scheme” on page 23-20.

- If you use hold-out validation, the app computes the accuracy scores using the observations in the validation fold and makes predictions on these observations. The app also computes the confusion matrix and ROC curve based on these predictions.
- If you use resubstitution validation, the score is the resubstitution accuracy based on all the training data, and the predictions are resubstitution predictions.

Check Performance in the Models Pane

After training a model in Classification Learner, check the **Models** pane to see which model has the best overall accuracy in percent. The best **Accuracy (Validation)** score is highlighted in a box. This score is the validation accuracy. The validation accuracy score estimates a model's performance on new data compared to the training data. Use the score to help you choose the best model.

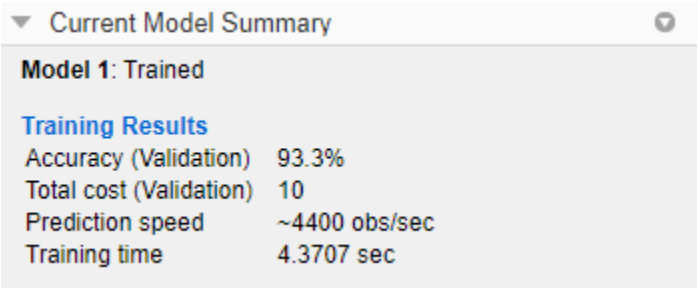
- For cross-validation, the score is the accuracy on all observations, counting each observation when it was in a held-out (validation) fold.
- For holdout validation, the score is the accuracy on the held-out observations.
- For resubstitution validation, the score is the resubstitution accuracy against all the training data observations.

The best overall score might not be the best model for your goal. A model with a slightly lower overall accuracy might be the best classifier for your goal. For example, false positives in a particular class might be important to you. You might want to exclude some predictors where data collection is expensive or difficult.

To find out how the classifier performed in each class, examine the confusion matrix.

View and Compare Model Metrics

You can view model metrics in the **Current Model Summary** pane and use these metrics to assess and compare models. The **Training Results** metrics are calculated on the validation set. The **Test Results** metrics, if displayed, are calculated on an imported test set. For more information, see “Evaluate Test Set Model Performance” on page 23-69.



To copy the information in the **Current Model Summary** pane, you can right-click into the pane and select **Copy text**.

Model Metrics

Metric	Description	Tip
Accuracy	Percentage of observations that are correctly classified	Look for larger accuracy values.
Total cost	Total misclassification cost	Look for smaller total cost values. Make sure the accuracy value is still large.

You can sort the models based on the different model metrics. To select a metric for model sorting, use the **Sort by** list at the top of the **Models** pane.

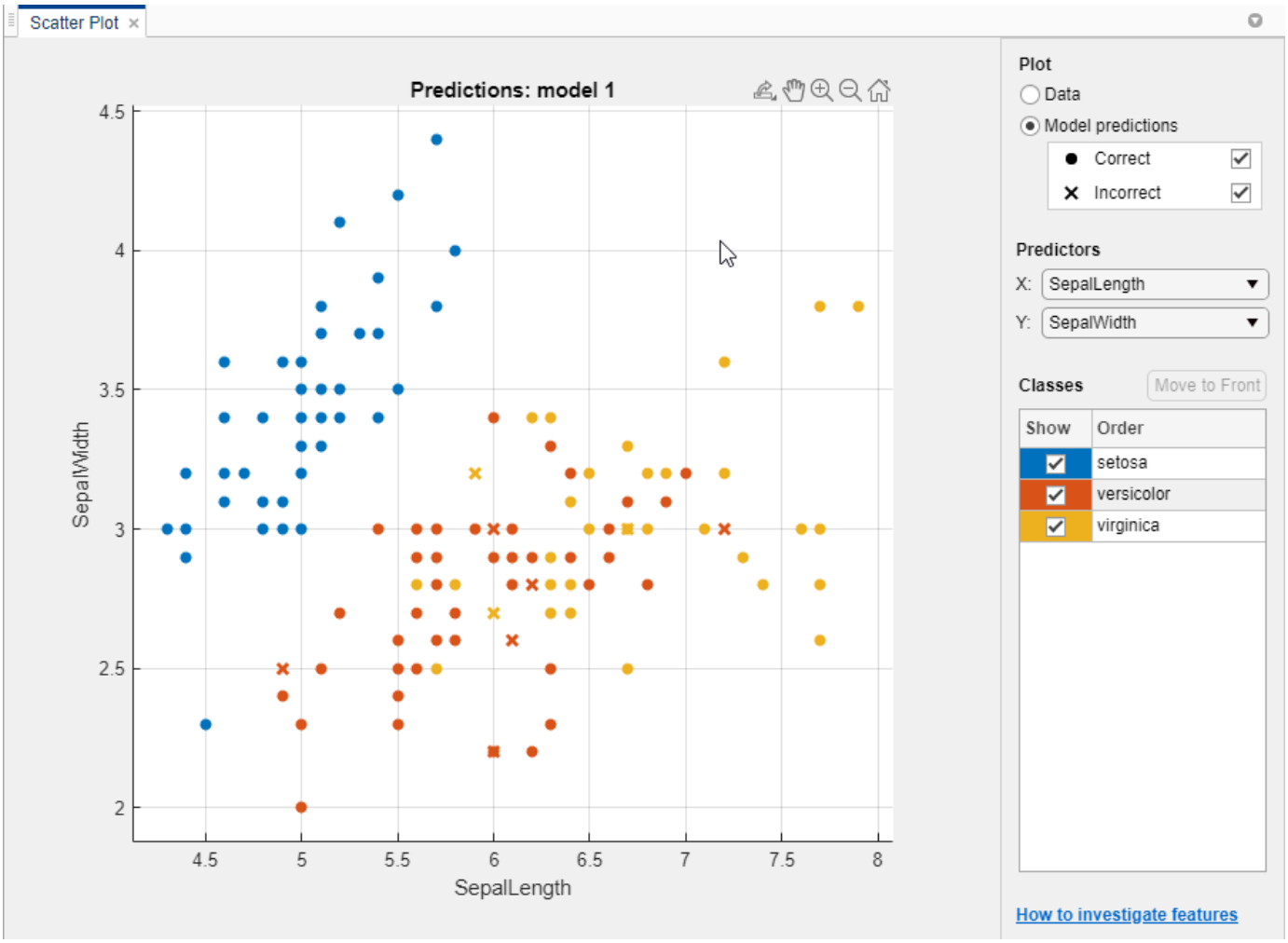
You can also delete unwanted models listed in the **Models** pane. Select the model you want to delete and click the **Delete selected model** button in the upper right of the pane, or right-click the model and select **Delete model**. You cannot delete the last remaining model in the **Models** pane.

Plot Classifier Results

In the scatter plot, view the classifier results. After you train a classifier, the scatter plot switches from displaying the data to showing model predictions. If you are using holdout or cross-validation, then these predictions are the predictions on the held-out (validation) observations. In other words, each prediction is obtained using a model that was trained without using the corresponding observation. To investigate your results, use the controls on the right. You can:

- Choose whether to plot model predictions or the data alone.
- Show or hide correct or incorrect results using the check boxes under **Model predictions**.
- Choose features to plot using the **X** and **Y** lists under **Predictors**.
- Visualize results by class by showing or hiding specific classes using the check boxes under **Show**.
- Change the stacking order of the plotted classes by selecting a class under **Classes** and then clicking **Move to Front**.

- Zoom in and out, or pan across the plot. To enable zooming or panning, hover the mouse over the scatter plot and click the corresponding button on the toolbar that appears above the top right of the plot.



See also “Investigate Features in the Scatter Plot” on page 23-42.

To export the scatter plots you create in the app to figures, see “Export Plots in Classification Learner App” on page 23-72.

Check Performance Per Class in the Confusion Matrix

Use the confusion matrix plot to understand how the currently selected classifier performed in each class. To view the confusion matrix after training a model, click **Confusion Matrix** and select **Validation Data** in the **Plots** section of the **Classification Learner** tab. The confusion matrix helps you identify the areas where the classifier has performed poorly.

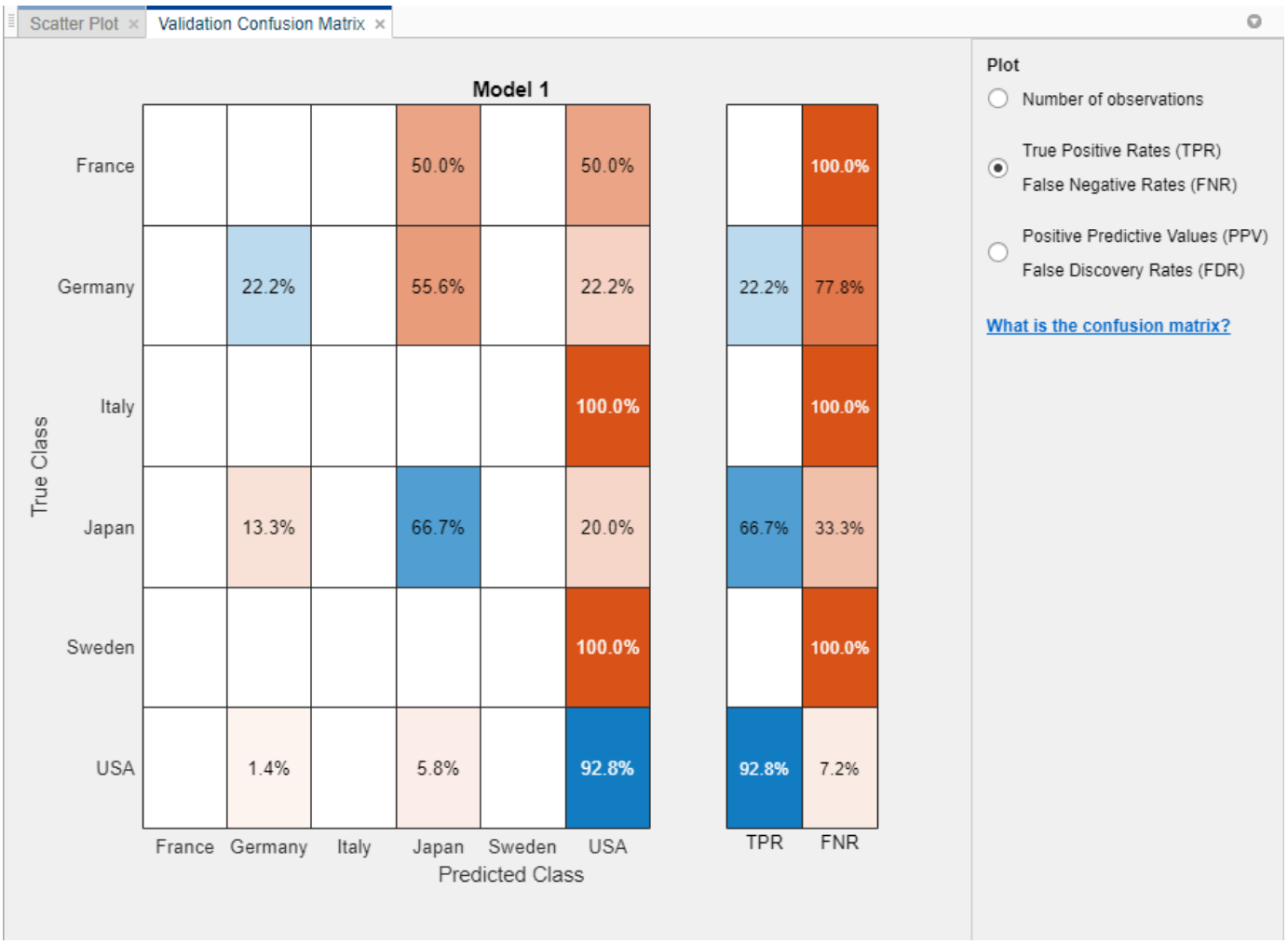
When you open the plot, the rows show the true class, and the columns show the predicted class. If you are using holdout or cross-validation, then the confusion matrix is calculated using the predictions on the held-out (validation) observations. The diagonal cells show where the true class

and predicted class match. If these diagonal cells are blue, the classifier has classified observations of this true class are classified correctly.

The default view shows the number of observations in each cell.

To see how the classifier performed per class, under **Plot**, select the **True Positive Rates (TPR)**, **False Negative Rates (FNR)** option. The TPR is the proportion of correctly classified observations per true class. The FNR is the proportion of incorrectly classified observations per true class. The plot shows summaries per true class in the last two columns on the right.

Tip Look for areas where the classifier performed poorly by examining cells off the diagonal that display high percentages and are orange. The higher the percentage, the darker the hue of the cell color. In these orange cells, the true class and the predicted class do not match. The data points are misclassified.



In this example, which uses the `carsmall` data set, the second row from the top shows all cars with the true class Germany. The columns show the predicted classes. 22.2% of the cars from Germany are correctly classified, so **22.2%** is the true positive rate for correctly classified points in this class, shown in the blue cell in the **TPR** column.

The other cars in the Germany row are misclassified: 55.6% of the cars are incorrectly classified as from Japan, and 22.2% are classified as from USA. The false negative rate for incorrectly classified points in this class is **77.8%**, shown in the orange cell in the **FNR** column.

If you want to see numbers of observations (cars, in this example) instead of percentages, under **Plot**, select **Number of observations**.

If false positives are important in your classification problem, plot results per predicted class (instead of true class) to investigate false discovery rates. To see results per predicted class, under **Plot**, select the **Positive Predictive Values (PPV)**, **False Discovery Rates (FDR)** option. The PPV is the proportion of correctly classified observations per predicted class. The FDR is the proportion of incorrectly classified observations per predicted class. With this option selected, the confusion matrix now includes summary rows below the table. Positive predictive values are shown in blue for the correctly predicted points in each class, and false discovery rates are shown in orange for the incorrectly predicted points in each class.

If you decide there are too many misclassified points in the classes of interest, try changing classifier settings or feature selection to search for a better model.

To export the confusion matrix plots you create in the app to figures, see “Export Plots in Classification Learner App” on page 23-72.

Check the ROC Curve

To view the ROC curve after training a model, on the **Classification Learner** tab, in the **Plots** section, click **ROC Curve** and select **Validation Data**. View the receiver operating characteristic (ROC) curve showing true and false positive rates. The ROC curve shows true positive rate versus false positive rate for the currently selected trained classifier. You can select different classes to plot.

The marker on the plot shows the performance of the currently selected classifier. The marker shows the values of the false positive rate (FPR) and the true positive rate (TPR) for the currently selected classifier. For example, a false positive rate (FPR) of 0.2 indicates that the current classifier assigns 20% of the observations incorrectly to the positive class. A true positive rate of 0.9 indicates that the current classifier assigns 90% of the observations correctly to the positive class.

A perfect result with no misclassified points is a right angle to the top left of the plot. A poor result that is no better than random is a line at 45 degrees. The **Area Under Curve** number is a measure of the overall quality of the classifier. Larger **Area Under Curve** values indicate better classifier performance. Compare classes and trained models to see if they perform differently in the ROC curve.

For more information, see `perfcurve`.

To export the ROC curve plots you create in the app to figures, see “Export Plots in Classification Learner App” on page 23-72.

Evaluate Test Set Model Performance

After training a model in Classification Learner, you can evaluate the model performance on a test set in the app. This process allows you to check whether the validation accuracy provides a good estimate for the model performance on new data.

1. Import a test data set into Classification Learner.

   - If the test data set is in the MATLAB workspace, then in the **Testing** section on the **Classification Learner** tab, click **Test Data** and select **From Workspace**.
   - If the test data set is in a file, then in the **Testing** section, click **Test Data** and select **From File**. Select a file type in the list, such as a spreadsheet, text file, or comma-separated values (`.csv`) file, or select **All Files** to browse for other file types such as `.dat`.

   In the Import Test Data dialog box, select the test data set from the **Test Data Set Variable** list. The test set must have the same variables as the predictors imported for training and validation. The unique values in the test response variable must be a subset of the classes in the full response variable.

2. Compute the test set metrics.

   - To compute test metrics for a single model, select the trained model in the **Models** pane. On the **Classification Learner** tab, in the **Testing** section, click **Test All** and select **Test Selected**.
   - To compute test metrics for all trained models, click **Test All** and select **Test All** in the **Testing** section.

   The app computes the test set performance of each model trained on the full data set, including training and validation data.

3. Compare the validation accuracy with the test accuracy.

   In the **Current Model Summary** pane, the app displays the validation metrics and test metrics under the **Training Results** section and **Test Results** section, respectively. You can check if the validation accuracy gives a good estimate for the test accuracy.

   You can also visualize the test results using plots.

   - Display a confusion matrix. In the **Plots** section on the **Classification Learner** tab, click **Confusion Matrix** and select **Test Data**.
   - Display an ROC curve. In the **Plots** section, click **ROC Curve** and select **Test Data**.

For an example, see “Check Classifier Performance Using Test Set in Classification Learner App” on page 23-137. For an example that uses test set metrics in a hyperparameter optimization workflow, see “Train Classifier Using Hyperparameter Optimization in Classification Learner App” on page 23-128.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Export Plots in Classification Learner App” on page 23-72
- “Export Classification Model to Predict New Data” on page 23-77

- “Train Decision Trees Using Classification Learner App” on page 23-83

Export Plots in Classification Learner App

After you create plots interactively in the Classification Learner app, you can export your app plots to MATLAB figures. You can then copy, save, or customize the new figures. Choose among the available plots: scatter plot on page 23-42, parallel coordinates plot on page 23-45, confusion matrix on page 23-67, ROC curve on page 23-69, and minimum classification error plot on page 23-61.

- Before exporting a plot, make sure the plot in the app displays the same data that you want in the new figure.
- On the **Classification Learner** tab, in the **Export** section, click **Export Plot to Figure**. The app creates a figure from the selected plot.
  - The new figure might not have the same interactivity options as the plot in the Classification Learner app. For example, data tips for the exported scatter plot show only X,Y values for the selected point, not the detailed information displayed in the app.
  - Additionally, the figure might have a different axes toolbar than the one in the app plot. For plots in Classification Learner, an axes toolbar appears above the top right of the plot. The buttons available on the toolbar depend on the contents of the plot. The toolbar can include buttons to export the plot as an image, add data tips, pan or zoom the data, and restore the view.

- Copy, save, or customize the new figure, which is displayed in the figure window.
  - To copy the figure, select **Edit > Copy Figure**. For more information, see “Copy Figure to Clipboard from Edit Menu”.
  - To save the figure, select **File > Save As**. Alternatively, you can follow the workflow described in “Customize Figure Before Saving”.
  - To customize the figure, click the Edit Plot button  on the figure toolbar. Right-click the section of the plot that you want to edit. You can change the listed properties, which might include **Color**, **Font**, **Line Style**, and other properties. Or, you can use the **Property Inspector** to change the figure properties.

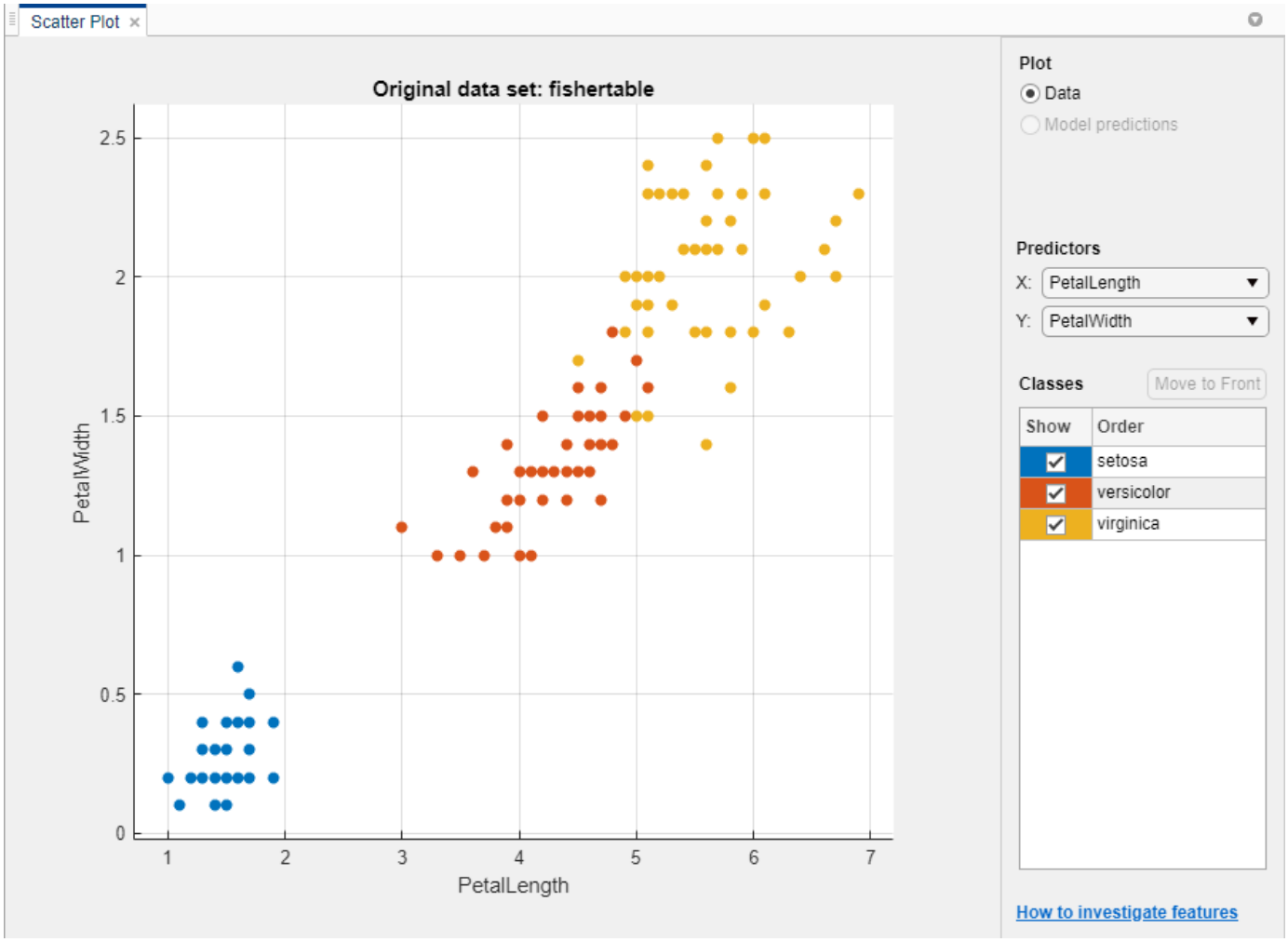
As an example, export a scatter plot in the app to a figure, customize the figure, and save the modified figure.

1. In the MATLAB Command Window, read the sample file `fisheriris.csv` into a table.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

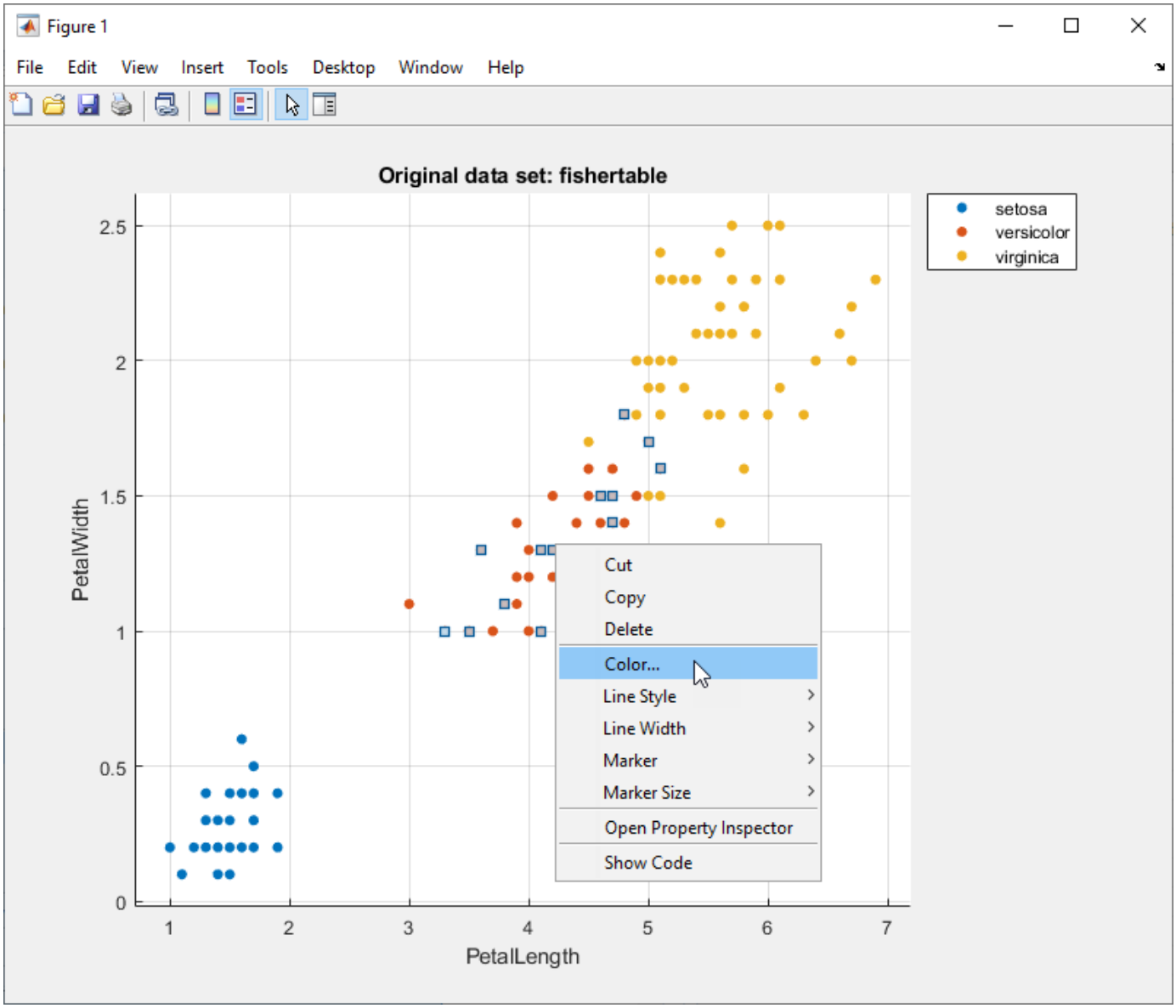
2. Click the **Apps** tab.
3. In the **Apps** section, click the arrow to open the gallery. Under **Machine Learning and Deep Learning**, click **Classification Learner**.
4. On the **Classification Learner** tab, in the **File** section, click **New Session**  .
5. In the New Session from Workspace dialog box, select the table `fishertable` from the **Workspace Variable** list.
6. Click **Start Session**. Classification Learner creates a scatter plot of the data by default.

7 Change the predictors in the scatter plot to `PetalLength` and `PetalWidth`.



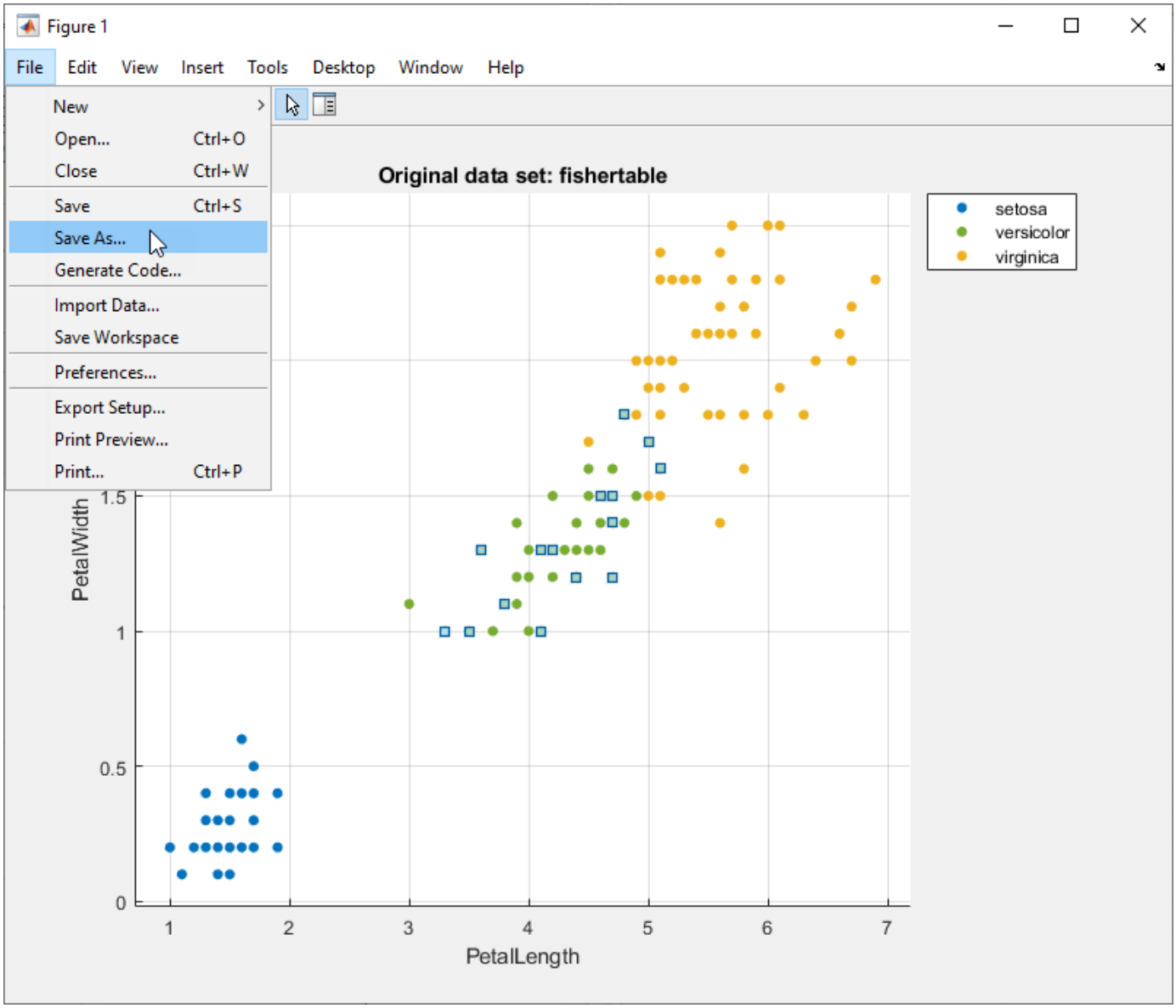
8 On the **Classification Learner** tab, in the **Export** section, click **Export Plot to Figure**.

9 In the new figure, click the Edit Plot button on the figure toolbar. Right-click the points in the plot corresponding to the versicolor irises. In the context menu, select **Color**.



10 In the Color dialog box, select a new color and click **OK**.

11 To save the figure, select **File** > **Save As**. Specify the saved file location, name, and type.



See Also

Related Examples

- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

Export Classification Model to Predict New Data

In this section...

Export the Model to the Workspace to Make Predictions for New Data

After you create classification models interactively in Classification Learner, you can export your best model to the workspace. You can then use the trained model to make predictions using new data.

Note The final model Classification Learner exports is always trained using the full data set. The validation scheme that you use only affects the way that the app computes validation metrics. You can use the validation metrics and various plots that visualize results to pick the best model for your classification problem.

Here are the steps for exporting a model to the MATLAB workspace:

1. In Classification Learner, select the model you want to export in the **Models** pane.
2. On the **Classification Learner** tab, in the **Export** section, click one of the export options:
   - If you want to include the data used for training the model, then select **Export Model**.

     You export the trained model to the workspace as a structure containing a classification object, such as a `ClassificationTree`, `ClassificationDiscriminant`, `ClassificationSVM`, `ClassificationNaiveBayes`, `ClassificationKNN`, `ClassificationEnsemble`, `ClassificationNeuralNetwork`, and so on.
   - If you do not want to include the training data, select **Export Compact Model**. This option exports the model with unnecessary data removed where possible. For some classifiers this is a compact classification object that does not include the training data (for example, `CompactClassificationTree`). You can use a compact classification object for making predictions of new data, but you can use fewer other methods with it.
3. In the Export Model dialog box, edit the name for your exported variable if you want, and then click **OK**. The default name for your exported model, `trainedModel`, increments every time you export to avoid overwriting your classifiers (for example, `trainedModel1`).

   The new variable (for example, `trainedModel`) appears in your workspace.

   The app displays information about the exported model in the command window. Read the message to learn how to make predictions with new data.

Make Predictions for New Data

After you export a model to the workspace from Classification Learner, or run the code generated from the app, you get a `trainedModel` structure that you can use to make predictions using new

data. The structure contains a classification object and a function for prediction. The structure allows you to make predictions for models that include principal component analysis (PCA).

1. To use the exported classifier to make predictions for new data, `T`, use the form:

   ```
   yfit = C.predictFcn(T)
   ```

   where `C` is the name of your variable (for example, `trainedModel`).

   Supply the data `T` with the same format and data type as the training data used in the app (table or matrix).

   - If you supply a table, ensure it contains the same predictor names as your training data. The `predictFcn` function ignores additional variables in tables. Variable formats and types must match the original training data.
   - If you supply a matrix, it must contain the same predictor columns or rows as your training data, in the same order and format. Do not include a response variable, any variables that you did not import in the app, or other unused variables.

   The output `yfit` contains a class prediction for each data point.
2. Examine the fields of the exported structure. For help making predictions, enter:

   ```
   C.HowToPredict
   ```

You can also extract the classification object from the exported structure for further analysis (for example, `trainedModel.ClassificationSVM`, `trainedModel.ClassificationTree`, and so on, depending on your model type). Be aware that if you used feature transformation such as PCA in the app, you will need to take account of this transformation by using the information in the PCA fields of the structure.

Generate MATLAB Code to Train the Model with New Data

After you create classification models interactively in Classification Learner, you can generate MATLAB code for your best model. You can then use the code to train the model with new data.

Generate MATLAB code to:

- Train on huge data sets. Explore models in the app trained on a subset of your data, then generate code to train a selected model on a larger data set
- Create scripts for training models without needing to learn syntax of the different functions
- Examine the code to learn how to train classifiers programmatically
- Modify the code for further analysis, for example to set options that you cannot change in the app
- Repeat your analysis on different data and automate training

1. In Classification Learner, in the **Models** pane, select the model you want to generate code for.
2. On the **Classification Learner** tab, in the **Export** section, click **Generate Function**.

   The app generates code from your session and displays the file in the MATLAB Editor. The file includes the predictors and response, the classifier training methods, and validation methods. Save the file.

3. To retrain your classifier model, call the function from the command line with your original data or new data as the input argument or arguments. New data must have the same shape as the original data.

   Copy the first line of the generated code, excluding the word `function`, and edit the `trainingData` input argument to reflect the variable name of your training data or new data. Similarly, edit the `responseData` input argument (if applicable).

   For example, to retrain a classifier trained with the `fishertable` data set, enter:

   ```
   [trainedModel,validationAccuracy] = trainClassifier(fishertable)
   ```

   The generated code returns a `trainedModel` structure that contains the same fields as the structure you create when you export a classifier from Classification Learner to the workspace.

4. If you want to automate training the same classifier with new data, or learn how to programmatically train classifiers, examine the generated code. The code shows you how to:

   - Process the data into the right shape
   - Train a classifier and specify all the classifier options
   - Perform cross-validation
   - Compute validation accuracy
   - Compute validation predictions and scores

   Note If you generate MATLAB code from a trained optimizable model, the generated code does not include the optimization process.

Generate C Code for Prediction

If you train one of the models in this table using Classification Learner, you can generate C code for prediction.

Model Type	Underlying Model Object
Decision Tree	`ClassificationTree` or `CompactClassificationTree`
Discriminant Analysis	`ClassificationDiscriminant` or `CompactClassificationDiscriminant`
Naive Bayes	`ClassificationNaiveBayes` or `CompactClassificationNaiveBayes`
Support Vector Machine	`ClassificationSVM` (binary), `CompactClassificationSVM` (binary), `ClassificationECOC` (multiclass), or `CompactClassificationECOC` (multiclass)
Nearest Neighbor	`ClassificationKNN`
Ensemble	`ClassificationEnsemble`, `CompactClassificationEnsemble`, or `ClassificationBaggedEnsemble`

Note You can generate C code for prediction using a logistic regression model. However, because the underlying model for logistic regression is a `GeneralizedLinearModel` or `CompactGeneralizedLinearModel` object, this process requires you to add extra lines of code in the prediction entry-point function to convert numeric predictions to class predictions. For an example, see “Code Generation for Logistic Regression Model Trained in Classification Learner” on page 32-144.

C code generation requires:

- MATLAB Coder license
- Appropriate model (binary or multiclass)

1. For example, train an SVM model in Classification Learner, and then export the model to the workspace.

   Find the underlying classification model object in the exported structure. Examine the fields of the structure to find the model object, for example, `C.ClassificationSVM`, where `C` is the name of your structure.

   The underlying model object depends on what type of SVM you trained (binary or multiclass) and whether you exported a compact model. The model object can be `ClassificationSVM`, `CompactClassificationSVM`, `ClassificationECOC`, or `CompactClassificationECOC`.

2. Use the function `saveLearnerForCoder` to prepare the model for code generation: `saveLearnerForCoder(Mdl,filename)`. For example:

   ```
   saveLearnerForCoder(C.ClassificationSVM,'mySVM')
   ```

3. Create a function that loads the saved model and makes predictions on new data. For example:

   ```
   function label = classifyX (X) %#codegen
   %CLASSIFYX Classify using SVM Model
   %  CLASSIFYX classifies the measurements in X
   %  using the SVM model in the file mySVM.mat, and then
   %  returns class labels in label.

   CompactMdl = loadLearnerForCoder('mySVM');
   label = predict(CompactMdl,X);
   end
   ```

4. Generate a MEX function from your function. For example:

   ```
   codegen classifyX.m -args {data}
   ```

   The `%#codegen` compilation directive indicates that the MATLAB code is intended for code generation. To ensure that the MEX function can use the same input, specify the data in the workspace as arguments to the function using the `-args` option. Specify `data` as a matrix containing only the predictor columns used to train the model.

5. Use the MEX function to make predictions. For example:

   ```
   labels = classifyX_mex(data);
   ```

If you used feature selection or PCA feature transformation in the app, then you need to take additional steps. If you used manual feature selection, supply the same columns in `X`. The `X` argument is the input to your function.

If you used PCA in the app, use the information in the PCA fields of the exported structure to take account of this transformation. It does not matter whether you imported a table or a matrix into the app, as long as X contains the matrix columns in the same order. Before generating code, follow these steps:

1. Save the PCACenters and PCACoefficients fields of the trained classifier structure, `C`, to file using the following command:

   ```
   save('pcaInfo.mat','-struct','C','PCACenters','PCACoefficients');
   ```

2. In your function file, include additional lines to perform the PCA transformation. Create a function that loads the saved model, performs PCA, and makes predictions on new data. For example:

   ```
   function label = classifyX (X) %#codegen
   %CLASSIFYX Classify using SVM Model
   %  CLASSIFYX classifies the measurements in X
   %  using the SVM model in the file mySVM.mat,
   %  and then returns class labels in label.
   % If you used manual feature selection in the app, ensure that X
   % contains only the columns you included in the model.

   CompactMdl = loadLearnerForCoder('mySVM');
   pcaInfo = coder.load('pcaInfo.mat','PCACenters','PCACoefficients');
   PCACenters = pcaInfo.PCACenters;
   PCACoefficients = pcaInfo.PCACoefficients;

   % Performs PCA transformation
   pcaTransformedX = bsxfun(@minus,X,PCACenters)*PCACoefficients;

   [label,scores] = predict(CompactMdl,pcaTransformedX);
   end
   ```

For a more detailed example, see “Code Generation and Classification Learner App” on page 32-31. For more information on the C code generation workflow and limitations, see “Code Generation”.

Deploy Predictions Using MATLAB Compiler

After you export a model to the workspace from Classification Learner, you can deploy it using MATLAB Compiler.

Suppose you export the trained model to MATLAB Workspace based on the instructions in “Export Model to Workspace” on page 24-54, with the name `trainedModel`. To deploy predictions, follow these steps.

- Save the `trainedModel` structure in a .mat file.

  ```
  save mymodel trainedModel
  ```

- Write the code to be compiled. This code must load the trained model and use it to make a prediction. It must also have a pragma, so the compiler recognizes that Statistics and Machine Learning Toolbox code is needed in the compiled application. This pragma could be any function in the toolbox.

  ```
  function ypred = mypredict(tbl)
  %#function fitctree
  load('mymodel.mat');
  ```

```
ypred = trainedModel.predictFcn(tbl);
end
```

- Compile as a standalone application.

```
mcc -m mypredict.m
```

See Also

Functions
`fitcdiscr` | `fitcecoc` | `fitcensemble` | `fitcknn` | `fitcnet` | `fitcsvm` | `fitctree` | `fitglm`

Classes
`ClassificationBaggedEnsemble` | `ClassificationDiscriminant` | `ClassificationECOC` | `ClassificationEnsemble` | `ClassificationKNN` | `ClassificationNaiveBayes` | `ClassificationNeuralNetwork` | `ClassificationSVM` | `ClassificationTree` | `CompactClassificationDiscriminant` | `CompactClassificationECOC` | `CompactClassificationEnsemble` | `CompactClassificationNaiveBayes` | `CompactClassificationNeuralNetwork` | `CompactClassificationSVM` | `CompactClassificationTree` | `GeneralizedLinearModel`

Related Examples

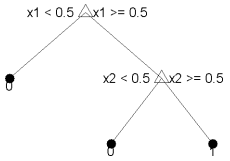
- “Train Classification Models in Classification Learner App” on page 23-10

Train Decision Trees Using Classification Learner App

This example shows how to create and compare various classification trees using Classification Learner, and export trained models to the workspace to make predictions for new data.

You can train classification trees to predict responses to data. To predict a response, follow the decisions in the tree from the root (beginning) node down to a leaf node. The leaf node contains the response.

Statistics and Machine Learning Toolbox trees are binary. Each step in a prediction involves checking the value of one predictor (variable). For example, here is a simple classification tree:

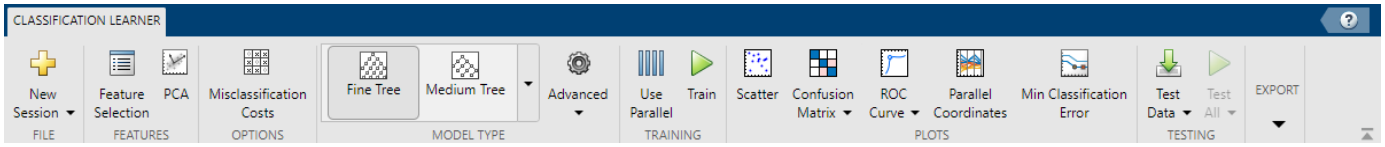


This tree predicts classifications based on two predictors, `x1` and `x2`. To predict, start at the top node. At each decision, check the values of the predictors to decide which branch to follow. When the branches reach a leaf node, the data is classified either as type `0` or `1`.

1. In MATLAB, load the `fisheriris` data set and create a table of measurement predictors (or features) using variables from the data set to use for a classification.

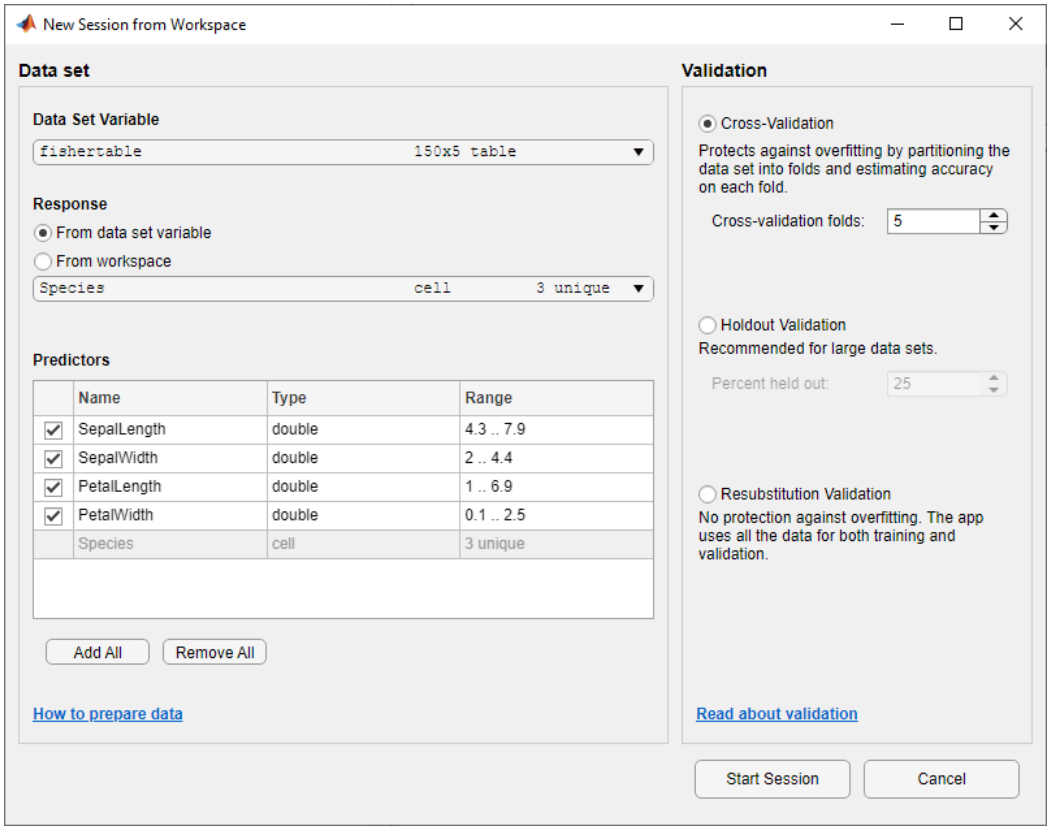
   ```
   fishertable = readtable('fisheriris.csv');
   ```

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.



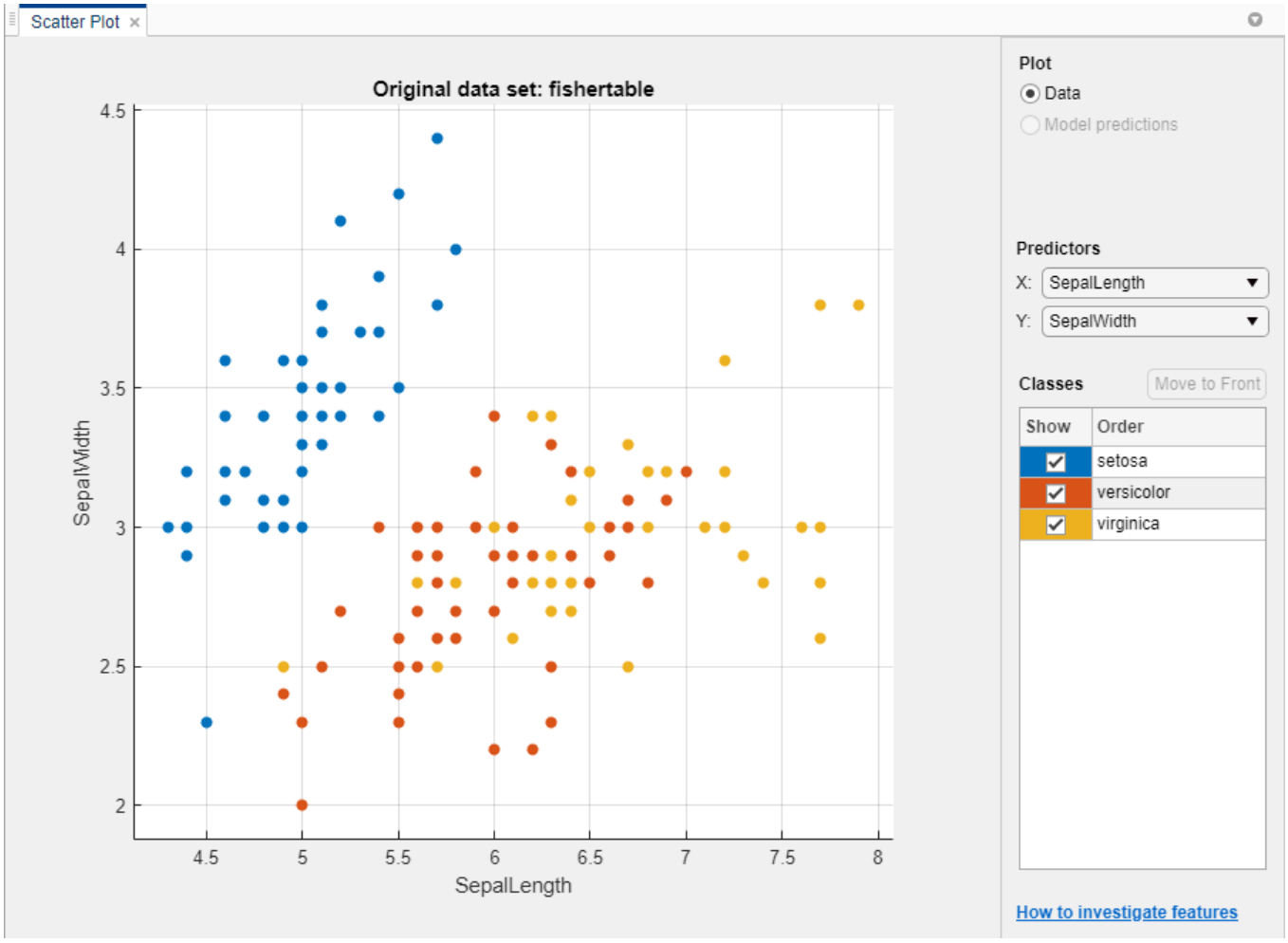
4. In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary).

   Observe that the app has selected response and predictor variables based on their data type. Petal and sepal length and width are predictors, and species is the response that you want to classify. For this example, do not change the selections.



5 To accept the default validation scheme and continue, click **Start Session**. The default validation option is cross-validation, to protect against overfitting.

Classification Learner creates a scatter plot of the data.



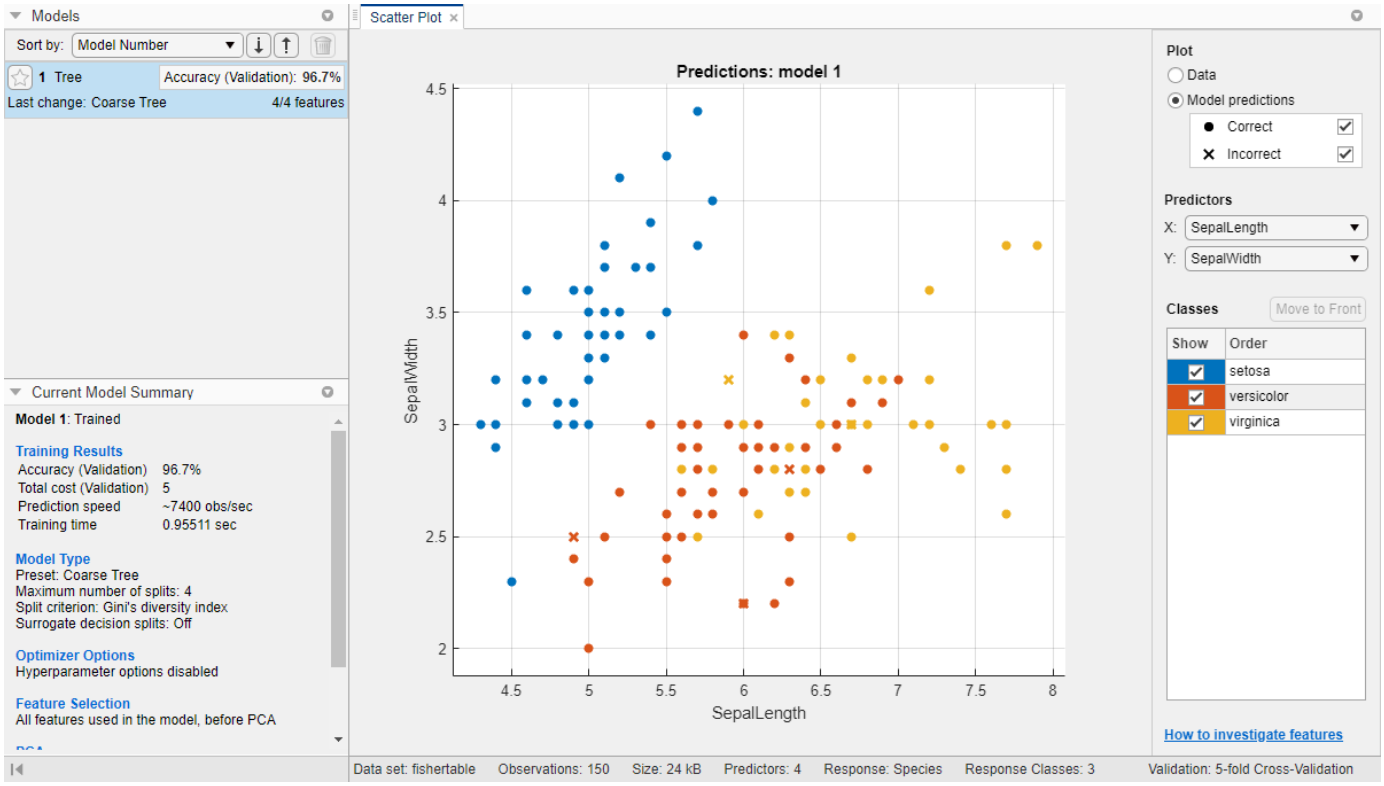
6 Use the scatter plot to investigate which variables are useful for predicting the response. Select different options on the **X** and **Y** lists under **Predictors** to visualize the distribution of species and measurements. Observe which variables separate the species colors most clearly.

   Observe that the `setosa` species (blue points) is easy to separate from the other two species with all four predictors. The `versicolor` and `virginica` species are much closer together in all predictor measurements, and overlap especially when you plot sepal length and width. `setosa` is easier to predict than the other two species.

7 To create a classification tree model, on the **Classification Learner** tab, in the **Model Type** section, click the down arrow to expand the gallery and click **Coarse Tree**. Then click **Train**.

   The app creates a simple classification tree, and plots the results.

   Observe the **Coarse Tree** model in the **Models** pane. Check the model validation score in the **Accuracy (Validation)** box. The model has performed well.



Note With validation, there is some randomness in the results, so your model validation score results can vary from those shown.

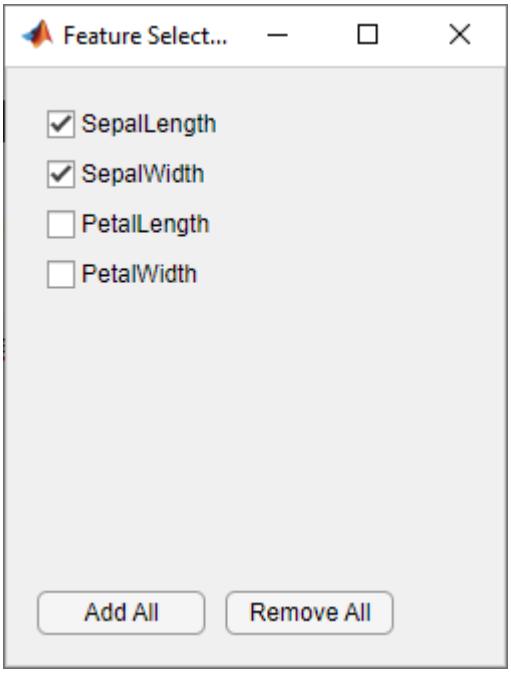
8. Examine the scatter plot. An X indicates misclassified points. The blue points (`setosa` species) are all correctly classified, but some of the other two species are misclassified. Under **Plot**, switch between the **Data** and **Model Predictions** options. Observe the color of the incorrect (X) points. Alternatively, while plotting model predictions, to view only the incorrect points, clear the **Correct** check box.
9. Train a different model to compare. Click **Medium Tree**, and then click **Train**.

   When you click **Train**, the app displays a new model in the **Models** pane.
10. Observe the **Medium Tree** model in the **Models** pane. The model validation score is no better than the coarse tree score. The app outlines in a box the **Accuracy (Validation)** score of the best model. Click each model in the **Models** pane to view and compare the results.
11. Examine the scatter plot for the **Medium Tree** model. The medium tree classifies as many points correctly as the previous coarse tree. You want to avoid overfitting, and the coarse tree performs well, so base all further models on the coarse tree.
12. Select **Coarse Tree** in the **Models** pane. To try to improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

    On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**.

    In the Feature Selection dialog box, clear the check boxes for **PetalLength** and **PetalWidth** to exclude them from the predictors. A new draft model appears in the **Models** pane with your new settings 2/4 features, based on the coarse tree.

Click **Train** to train a new tree model using the new predictor options.



13 Observe the third model in the **Models** pane. It is also a **Coarse Tree** model, trained using only 2 of 4 predictors. The app displays how many predictors are excluded. To check which predictors are included, click a model in the **Models** pane and observe the check boxes in the Feature Selection dialog box. The model with only sepal measurements has a much lower accuracy score than the petals-only model.

14 Train another model including only the petal measurements. Change the selections in the Feature Selection dialog box and click **Train**.

The model trained using only petal measurements performs comparably to the models containing all predictors. The models predict no better using all the measurements compared to only the petal measurements. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily without some predictors.

15 Repeat to train two more models including only the width measurements and then the length measurements. There is not much difference in score between several of the models.

16 Choose a best model among those of similar scores by examining the performance in each class. Select the coarse tree that includes all the predictors. To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. Use this plot to understand how the currently selected classifier performed in each class. View the matrix of true class and predicted class results.

Look for areas where the classifier performed poorly by examining cells off the diagonal that display high numbers and are red. In these red cells, the true class and the predicted class do not match. The data points are misclassified.

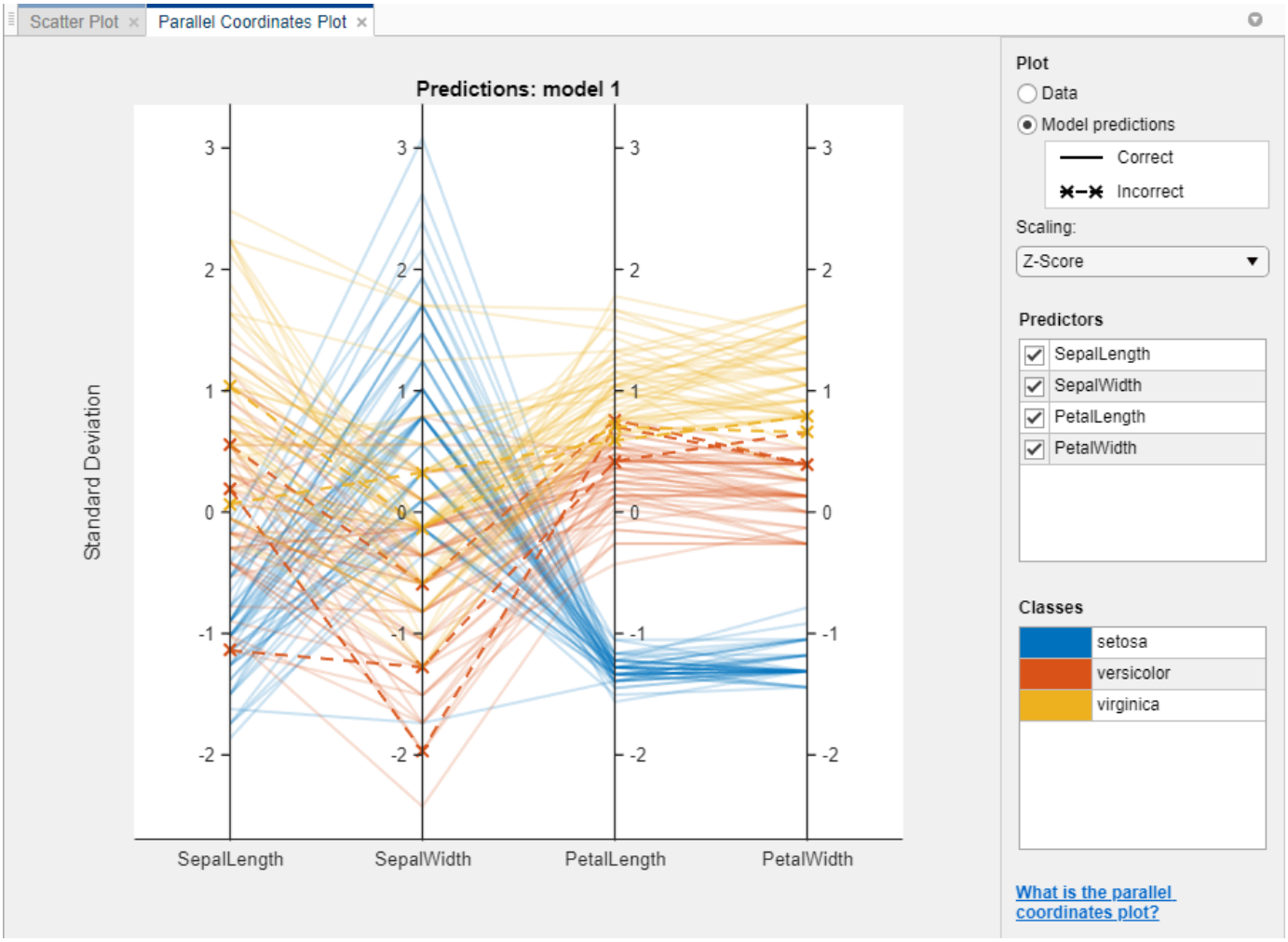
Note With validation, there is some randomness in the results, so your confusion matrix results can vary from those shown.

In this figure, examine the third cell in the middle row. In this cell, true class is `versicolor`, but the model misclassified the points as `virginica`. For this model, the cell shows 2 misclassified (your results can vary). To view percentages instead of numbers of observations, select the **True Positive Rates** option under **Plot** controls.

You can use this information to help you choose the best model for your goal. If false positives in this class are very important to your classification problem, then choose the best model at predicting this class. If false positives in this class are not very important, and models with fewer predictors do better in other classes, then choose a model to tradeoff some overall accuracy to exclude some predictors and make future data collection easier.

17 Compare the confusion matrix for each model in the **Models** pane. Check the Feature Selection dialog box to see which predictors are included in each model.

18 To investigate features to include or exclude, use the scatter plot and the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, click **Parallel Coordinates**. You can see that petal length and petal width are the features that separate the classes best.



19 To learn about model settings, choose a model in the **Models** pane and view the advanced settings. The nonoptimizable model options in the **Model Type** gallery are preset starting points, and you can change further settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Compare the simple and medium tree models in the **Models** pane, and observe the differences in the Advanced Tree Options dialog box. The **Maximum Number of Splits** setting controls tree depth.

To try to improve the coarse tree model further, try changing the **Maximum Number of Splits** setting, then train a new model by clicking **Train**.

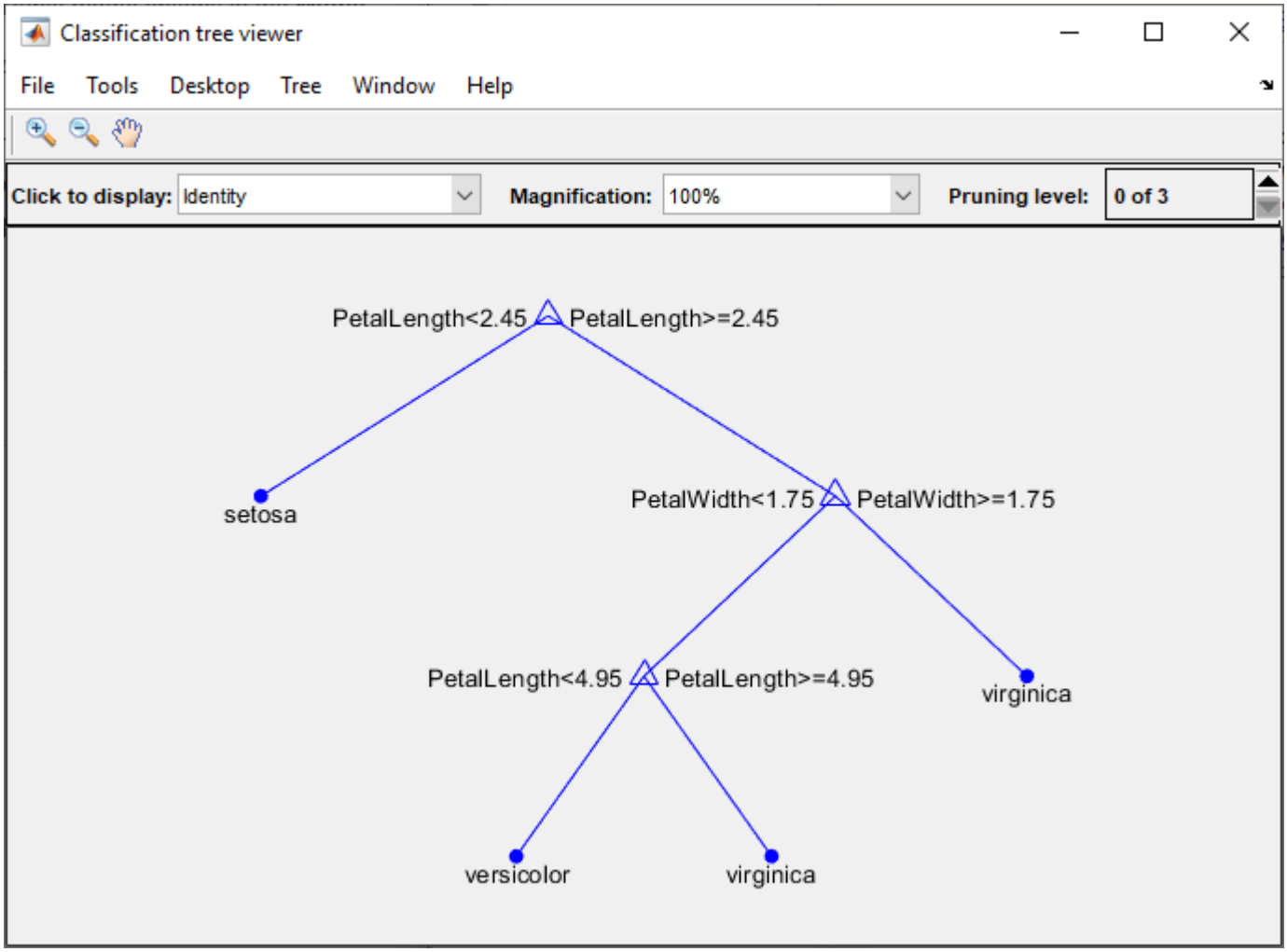
View the settings for the selected trained model in the **Current Model Summary** pane, or in the Advanced dialog box.

20 To export the best trained model to the workspace, on the **Classification Learner** tab, in the **Export** section, click **Export Model**. In the Export Model dialog box, click **OK** to accept the default variable name `trainedModel`.

Look in the command window to see information about the results.

21 To visualize your decision tree model, enter:

```
view(trainedModel.ClassificationTree,'Mode','graph')
```



22 You can use the exported classifier to make predictions on new data. For example, to make predictions for the `fishertable` data in your workspace, enter:

```
yfit = trainedModel.predictFcn(fishertable)
```

The output `yfit` contains a class prediction for each data point.

23 If you want to automate training the same classifier with new data, or learn how to programmatically train classifiers, you can generate code from the app. To generate code for the best trained model, on the **Classification Learner** tab, in the **Export** section, click **Generate Function**.

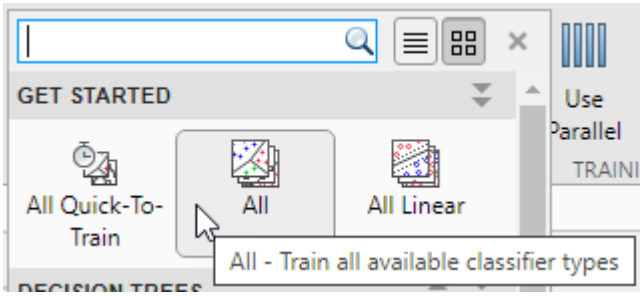
The app generates code from your model and displays the file in the MATLAB Editor. To learn more, see “Generate MATLAB Code to Train the Model with New Data” on page 23-78.

This example uses Fisher's 1936 iris data. The iris data contains measurements of flowers: the petal length, petal width, sepal length, and sepal width for specimens from three species. Train a classifier to predict the species based on the predictor measurements.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1. Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.
2. Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

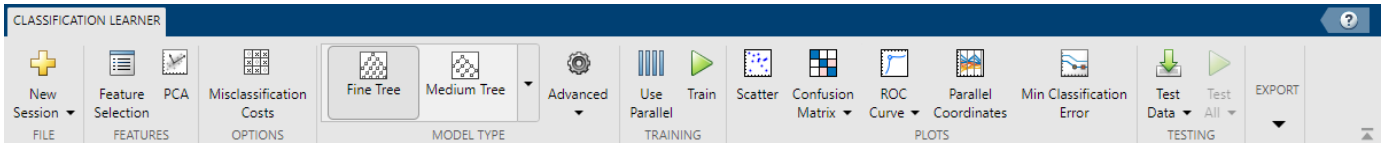
Train Discriminant Analysis Classifiers Using Classification Learner App

This example shows how to construct discriminant analysis classifiers in the Classification Learner app, using the `fisheriris` data set. You can use discriminant analysis with two or more classes in Classification Learner.

1. In MATLAB, load the `fisheriris` data set.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.

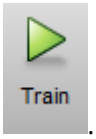


   In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary). Observe that the app has selected response and predictor variables based on their data type. Petal and sepal length and width are predictors, and species is the response that you want to classify. For this example, do not change the selections.

4. Click **Start Session**.

   Classification Learner creates a scatter plot of the data.

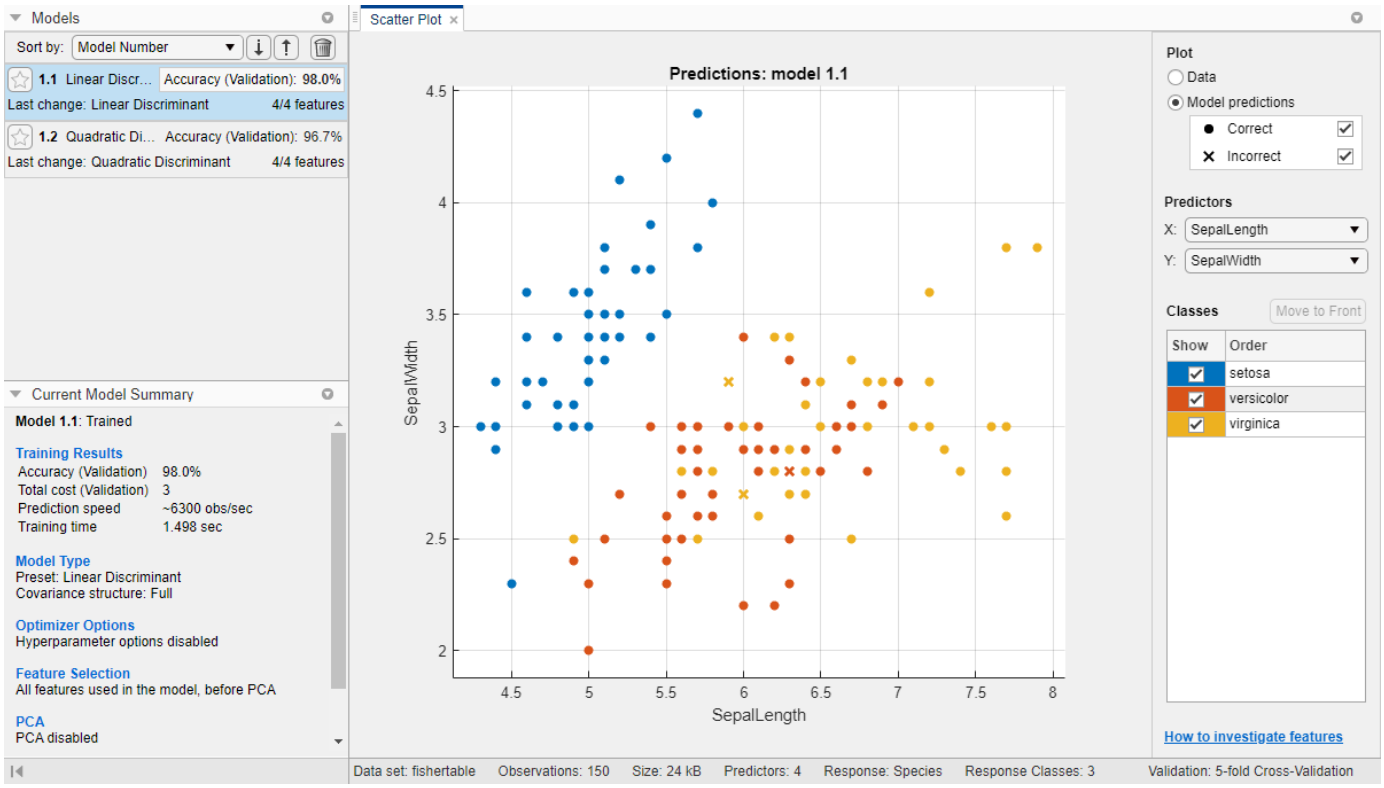
5. Use the scatter plot to visualize which variables are useful for predicting the response. Select different variables in the X- and Y-axis controls. Observe which variables separate the classes most clearly.
6. To train both nonoptimizable discriminant analysis classifiers, on the **Classification Learner** tab, in the **Model Type** section, click the down arrow to expand the list of classifiers, and under **Discriminant Analysis**, click **All Discriminants**.

Then click **Train** 

.

Tip If you have Parallel Computing Toolbox, you can train all the models (**All Discriminants**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Classification Learner trains one of each classification option in the gallery, linear and quadratic discriminants, and highlights the best score. The app outlines in a box the **Accuracy (Validation)** score of the best model.



7 Select a model in the **Models** pane to view the results. Examine the scatter plot for the trained model and try plotting different predictors. Misclassified points are shown as an X.

8 To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.

9 Select the other model in the **Models** pane to compare.

For information on the strengths of different model types, see “Discriminant Analysis” on page 23-29.

10 Choose the best model in the **Models** pane (the best score is highlighted in a box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, specify predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.

11 To investigate features to include or exclude, use the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.

12 Choose the best model in the **Models** pane. To try to improve the model further, try changing classifier settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Try changing a setting, then train the new model by clicking **Train**. For information on settings, see “Discriminant Analysis” on page 23-29.

13 To export the trained model to the workspace, select the Classification Learner tab and click **Export model**. See “Export Classification Model to Predict New Data” on page 23-77.

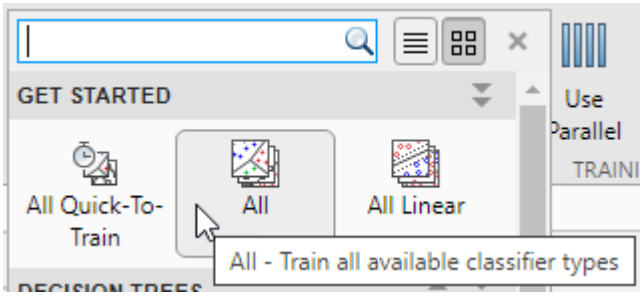
14 To examine the code for training this classifier, click **Generate Function**.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1 Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.

2 Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

Train Logistic Regression Classifiers Using Classification Learner App

This example shows how to construct logistic regression classifiers in the Classification Learner app, using the `ionosphere` data set that contains two classes. You can use logistic regression with two classes in Classification Learner. In the `ionosphere` data, the response variable is categorical with two levels: `g` represents good radar returns, and `b` represents bad radar returns.

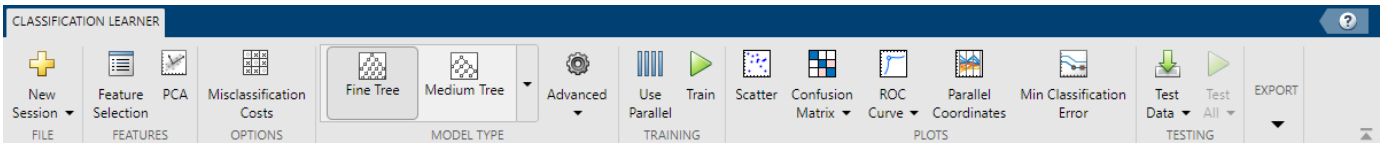
1. In MATLAB, load the `ionosphere` data set and define some variables from the data set to use for a classification.

   ```
   load ionosphere
   ionosphere = array2table(X);
   ionosphere.Group = Y;
   ```

   Alternatively, you can load the `ionosphere` data set and keep the `X` and `Y` data as separate variables.

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.

3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.



   In the New Session from Workspace dialog box, select the table `ionosphere` from the **Data Set Variable** list. Observe that the app has selected `Group` for the response variable, and the rest as predictors. `Group` has two levels.

   Alternatively, if you kept your predictor data `X` and response variable `Y` as two separate variables, you can first select the matrix `X` from the **Data Set Variable** list. Then, under **Response**, click the **From workspace** option button and select `Y` from the list. The `Y` variable is the same as the `Group` variable.

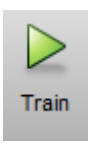
4. Click **Start Session**.

   Classification Learner creates a scatter plot of the data.

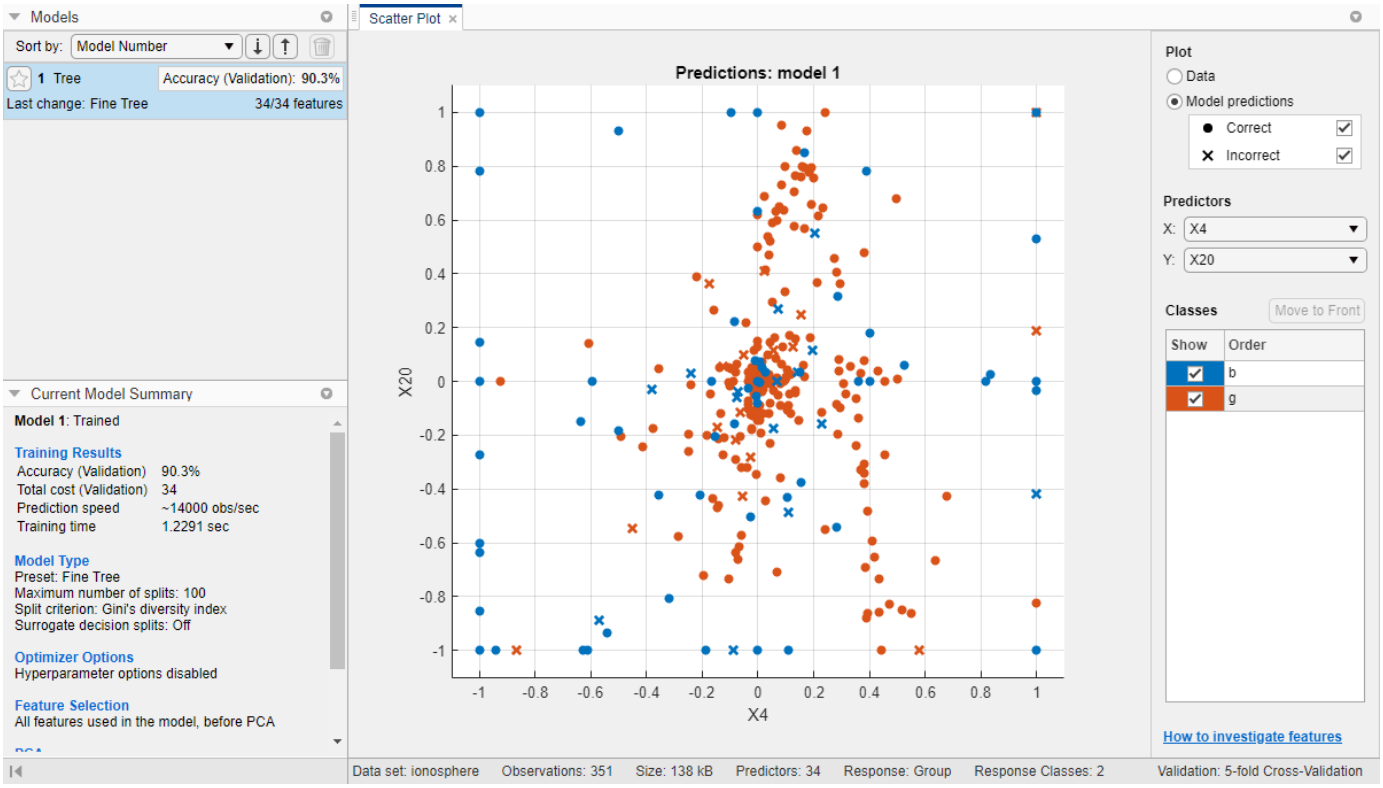
5. Use the scatter plot to visualize which variables are useful for predicting the response. Select different variables in the X- and Y-axis controls. Observe which variables separate the class colors most clearly.

6. To train the logistic regression classifier, on the **Classification Learner** tab, in the **Model Type** section, click the down arrow to expand the list of classifiers, and under **Logistic Regression Classifiers**, click **Logistic Regression**.

   Then click **Train**.



Classification Learner trains the model. The app outlines in a box the **Accuracy (Validation)** score of the best model (in this case, there is only one model).



7 Select the model in the **Models** pane to view the results. Examine the scatter plot for the trained model and try plotting different predictors. Misclassified points are shown as an X.

8 To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.

9 Choose the best model in the **Models** pane (the best score is highlighted in a box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

   On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, specify predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.

10 To investigate features to include or exclude, use the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.

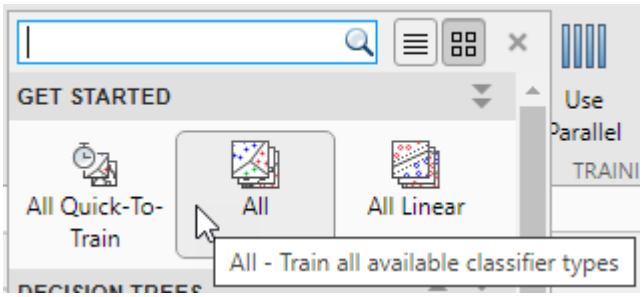
11 To export the trained model to the workspace, select the Classification Learner tab and click **Export model**. See “Export Classification Model to Predict New Data” on page 23-77.

12 To examine the code for training this classifier, click **Generate Function**.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1. Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.
2. Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Logistic Regression” on page 23-30
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

Train Support Vector Machines Using Classification Learner App

This example shows how to construct support vector machine (SVM) classifiers in the Classification Learner app, using the `ionosphere` data set that contains two classes. You can use a support vector machine (SVM) with two or more classes in Classification Learner. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of another class. In the `ionosphere` data, the response variable is categorical with two levels: `g` represents good radar returns, and `b` represents bad radar returns.

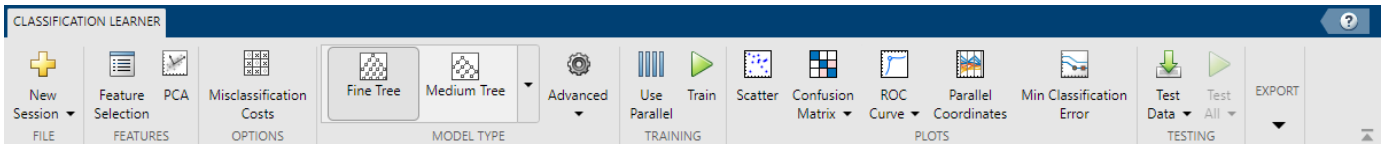
1. In MATLAB, load the `ionosphere` data set and define some variables from the data set to use for a classification.

   ```
   load ionosphere
   ionosphere = array2table(X);
   ionosphere.Group = Y;
   ```

   Alternatively, you can load the `ionosphere` data set and keep the `X` and `Y` data as separate variables.

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.

3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.



   In the New Session from Workspace dialog box, select the table `ionosphere` from the **Data Set Variable** list. Observe that the app has selected response and predictor variables based on their data type. The response variable `Group` has two levels. All the other variables are predictors.

   Alternatively, if you kept your predictor data `X` and response variable `Y` as two separate variables, you can first select the matrix `X` from the **Data Set Variable** list. Then, under **Response**, click the **From workspace** option button and select `Y` from the list. The `Y` variable is the same as the `Group` variable.

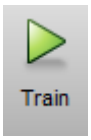
4. Click **Start Session**.

   Classification Learner creates a scatter plot of the data.

5. Use the scatter plot to visualize which variables are useful for predicting the response. Select different variables in the X- and Y-axis controls. Observe which variables separate the class colors most clearly.

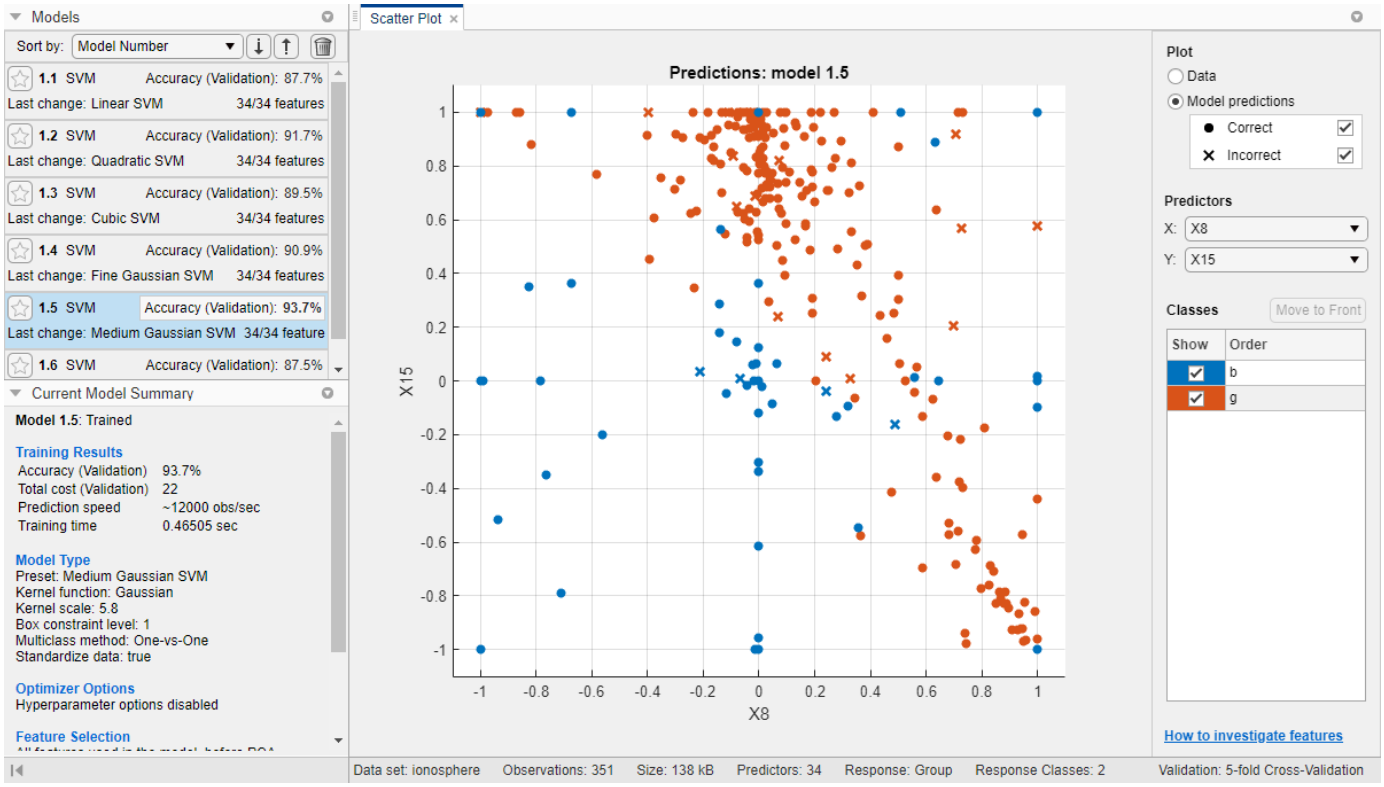
6. To create a selection of SVM models, on the **Classification Learner** tab, in the **Model Type** section, click the down arrow to expand the list of classifiers, and under **Support Vector Machines**, click **All SVMs**.

   Then click **Train**.



Tip If you have Parallel Computing Toolbox, you can train all the models (**All SVMs**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Classification Learner trains one of each nonoptimizable SVM classification option in the gallery, and highlights the best score. The app outlines in a box the **Accuracy (Validation)** score of the best model.



7 Select a model in the **Models** pane to view the results. Examine the scatter plot for the trained model and try plotting different predictors. Misclassified points are shown as an X.

8 To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.

9 Select the other models in the **Models** pane to compare.

10 Choose the best model (the best score is highlighted in a box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, specify predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.

11 To investigate features to include or exclude, use the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.

12 Choose the best model in the **Models** pane. To try to improve the model further, try changing SVM settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Try changing a setting, then train the new model by clicking **Train**. For information on settings, see “Support Vector Machines” on page 23-31.

13 To export the trained model to the workspace, select the Classification Learner tab and click **Export model**. See “Export Classification Model to Predict New Data” on page 23-77.

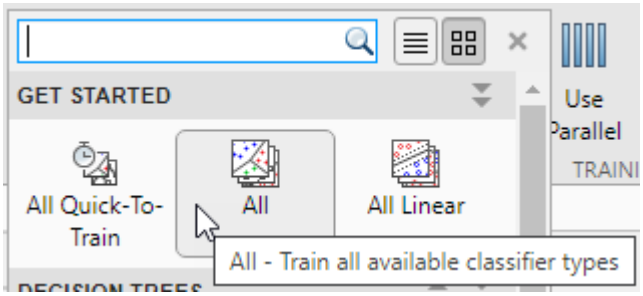
14 To examine the code for training this classifier, click **Generate Function**. For SVM models, see also “Generate C Code for Prediction” on page 23-79.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1 Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.

2 Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Generate C Code for Prediction” on page 23-79
- “Train Decision Trees Using Classification Learner App” on page 23-83

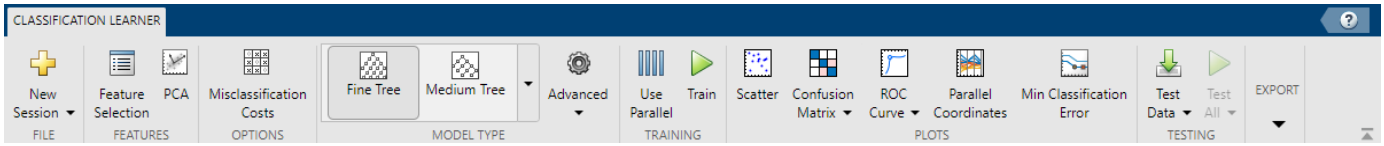
Train Nearest Neighbor Classifiers Using Classification Learner App

This example shows how to construct nearest neighbors classifiers in the Classification Learner app.

1. In MATLAB, load the `fisheriris` data set and define some variables from the data set to use for a classification.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.

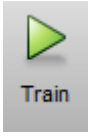


   In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary). Observe that the app has selected response and predictor variables based on their data type. Petal and sepal length and width are predictors, and species is the response that you want to classify. For this example, do not change the selections.

4. Click **Start Session**.

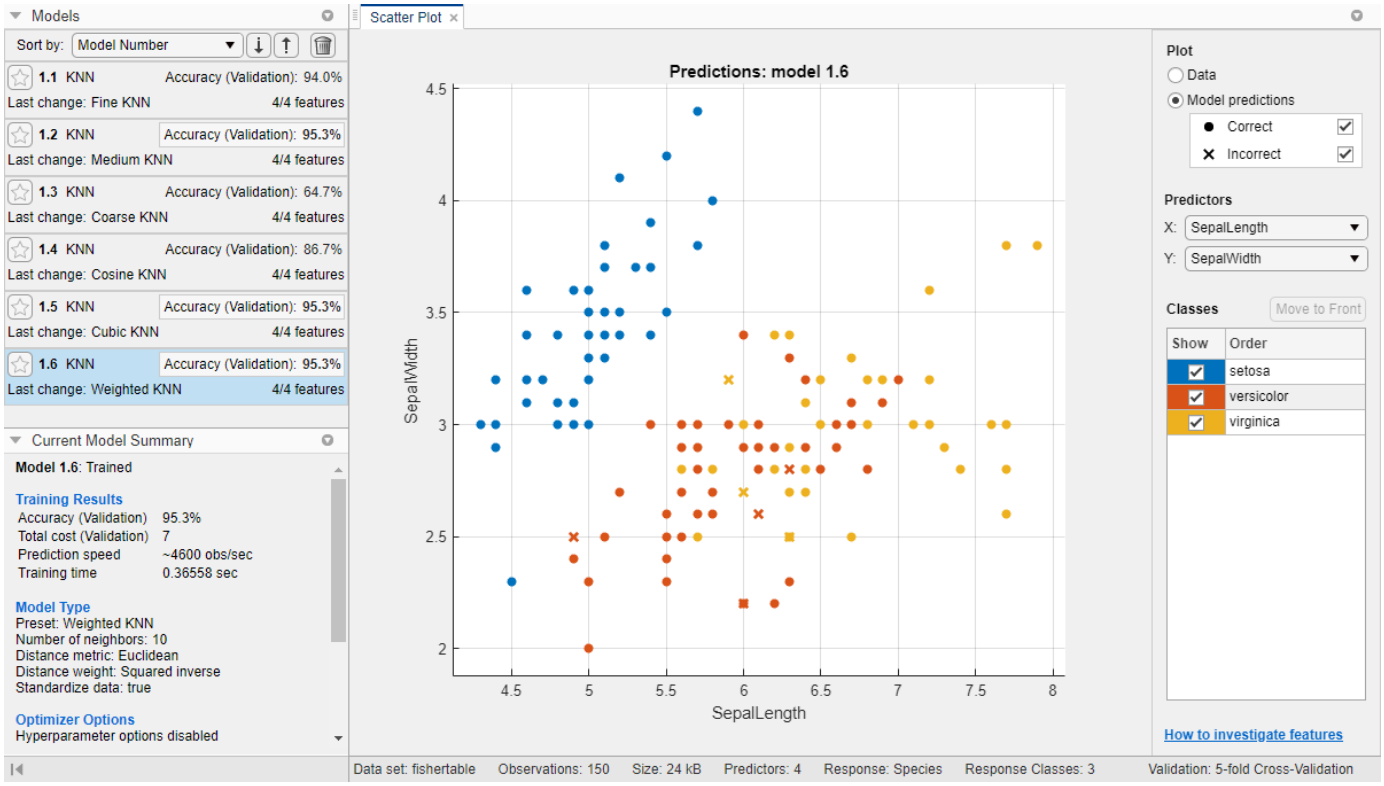
   The app creates a scatter plot of the data.

5. Use the scatter plot to investigate which variables are useful for predicting the response. To visualize the distribution of species and measurements, select different options on the **Variable on X axis** and **Variable on Y axis** menus. Observe which variables separate the species colors most clearly.
6. To create a selection of nearest neighbors models, on the **Classification Learner** tab, on the far right of the **Model Type** section, click the arrow to expand the list of classifiers, and under **Nearest Neighbor Classifiers**, click **All KNNs**.

7. In the **Training** section, click **Train**. 

Tip If you have Parallel Computing Toolbox, you can train all the models (**All KNNs**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Classification Learner trains one of each nonoptimizable nearest neighbor classification option in the gallery, and highlights the best score. The app outlines in a box the **Accuracy (Validation)** score of the best model.



8. Select a model in the **Models** pane to view the results. Examine the scatter plot for the trained model. An X indicates a misclassified point.
9. To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.
10. Select the other models in the **Models** pane to compare.
11. Choose the best model (the best score is highlighted in a box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

    On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, select predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.
12. To investigate features to include or exclude, use the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.
13. Choose the best model in the **Models** pane. To try to improve the model further, try changing settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Try changing a setting, and then train the new model by clicking **Train**. For information on settings and the strengths of different nearest neighbor model types, see “Nearest Neighbor Classifiers” on page 23-34.
14. To export the trained model to the workspace, in the **Export** section of the toolstrip, click **Export model**. See “Export Classification Model to Predict New Data” on page 23-77.

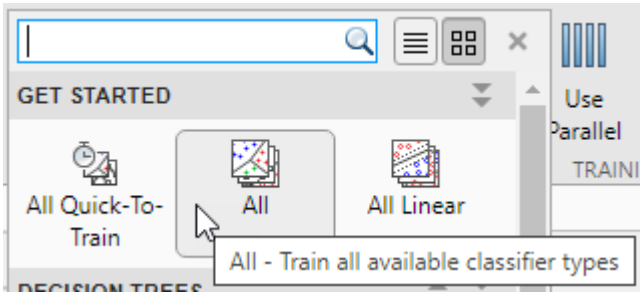
15 To examine the code for training this classifier, click **Generate Function**.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1 Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.

2 Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

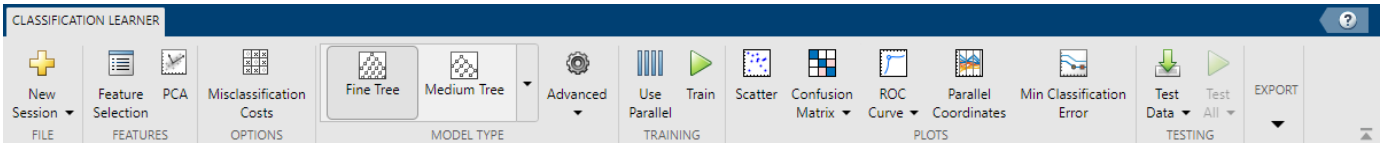
Train Ensemble Classifiers Using Classification Learner App

This example shows how to construct ensembles of classifiers in the Classification Learner app. Ensemble classifiers meld results from many weak learners into one high-quality ensemble predictor. Qualities depend on the choice of algorithm, but ensemble classifiers tend to be slow to fit because they often need many learners.

1. In MATLAB, load the `fisheriris` data set and define some variables from the data set to use for a classification.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, click **New Session > From Workspace**.

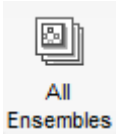


   In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary). Observe that the app has selected response and predictor variables based on their data type. Petal and sepal length and width are predictors. Species is the response that you want to classify. For this example, do not change the selections.

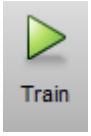
4. Click **Start Session**.

   Classification Learner creates a scatter plot of the data.

5. Use the scatter plot to investigate which variables are useful for predicting the response. Select different variables in the X- and Y-axis controls to visualize the distribution of species and measurements. Observe which variables separate the species colors most clearly.
6. To create a selection of ensemble models, on the **Classification Learner** tab, in the **Model Type** section, click the down arrow to expand the list of classifiers, then under **Ensemble Classifiers**, click **All Ensembles**.



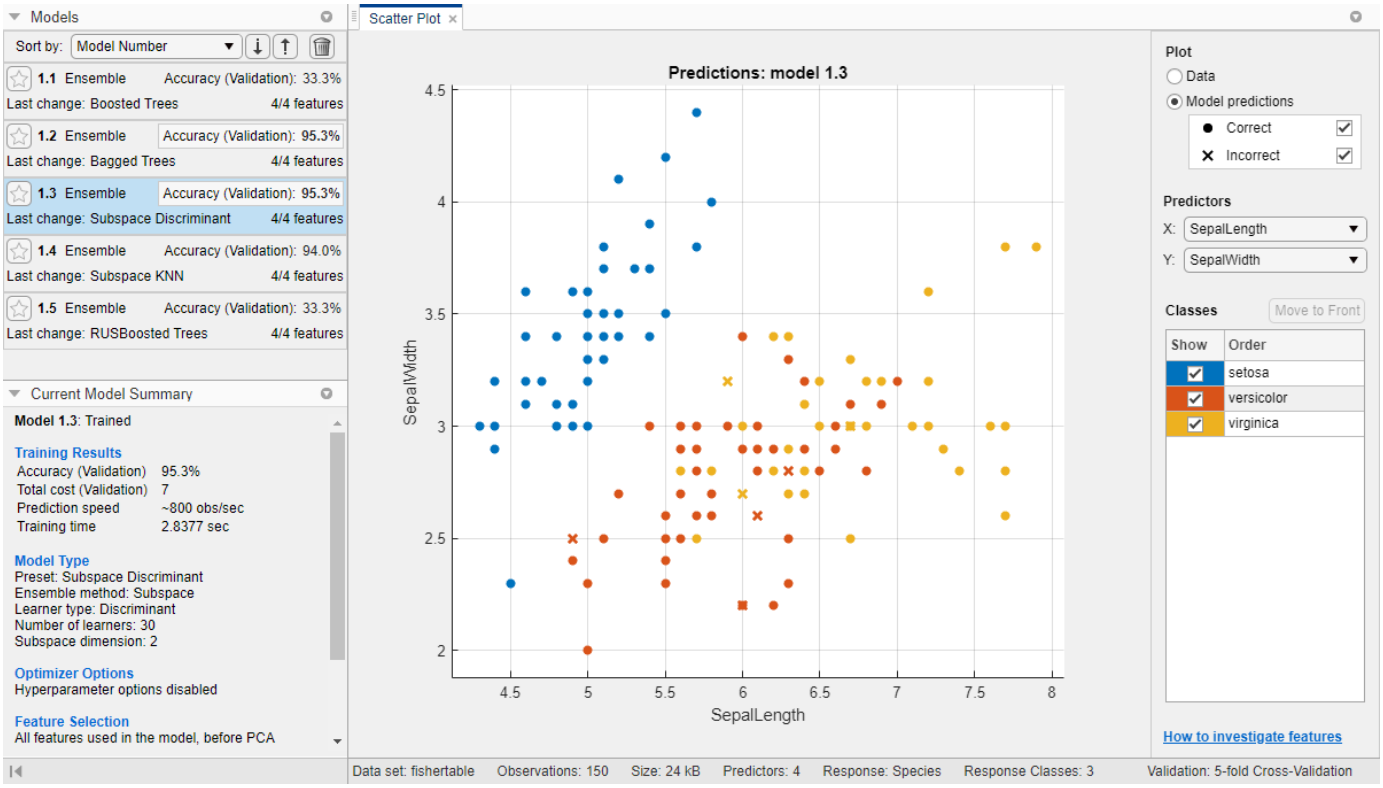
7. In the **Training** section, click **Train**.



Tip If you have Parallel Computing Toolbox, you can train all the models (**All Ensembles**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Classification Learner trains one of each nonoptimizable ensemble classification option in the gallery, and highlights the best score. The app outlines in a box the **Accuracy (Validation)** score of the best model.

8 Select a model in the **Models** pane to view the results. Examine the scatter plot for the trained model. Misclassified points are shown as an X.



9 To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.

10 Select the other models in the **Models** pane to compare.

11 Choose the best model (the best score is highlighted in the **Accuracy (Validation)** box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, specify predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.

12 To investigate features to include or exclude, use the scatter and parallel coordinates plots. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.

13 Choose the best model in the **Models** pane. To try to improve the model further, try changing settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Try changing a setting, then train the new model by clicking **Train**.

For information on the settings to try and the strengths of different ensemble model types, see “Ensemble Classifiers” on page 23-37.

14 To export the trained model to the workspace, select the Classification Learner tab and click **Export model**. See “Export Classification Model to Predict New Data” on page 23-77.

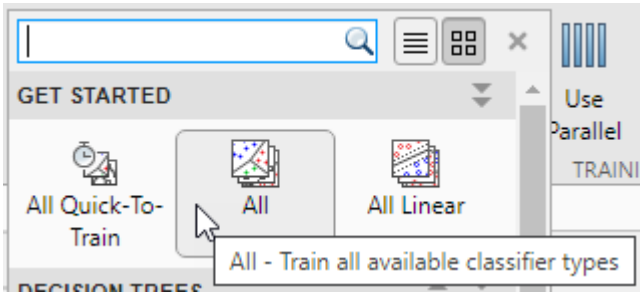
15 To examine the code for training this classifier, click **Generate Function**.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1 Click the arrow on the far right of the **Model Type** section to expand the list of classifiers.

2 Click **All**, then click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

Train Naive Bayes Classifiers Using Classification Learner App

This example shows how to create and compare different naive Bayes classifiers using the Classification Learner app, and export trained models to the workspace to make predictions for new data.

Naive Bayes classifiers leverage Bayes theorem and make the assumption that predictors are independent of one another within each class. However, the classifiers appear to work well even when the independence assumption is not valid. You can use naive Bayes with two or more classes in Classification Learner. The app allows you to train a Gaussian naive Bayes model or a kernel naive Bayes model individually or simultaneously.

This table lists the available naive Bayes models in Classification Learner and the probability distributions used by each model to fit predictors.

Model	Numerical Predictor	Categorical Predictor
Gaussian naive Bayes	Gaussian distribution (or normal distribution)	multivariate multinomial distribution
Kernel naive Bayes	Kernel distribution You can specify the kernel type and support. Classification Learner automatically determines the kernel width using the underlying `fitcnb` function.	multivariate multinomial distribution

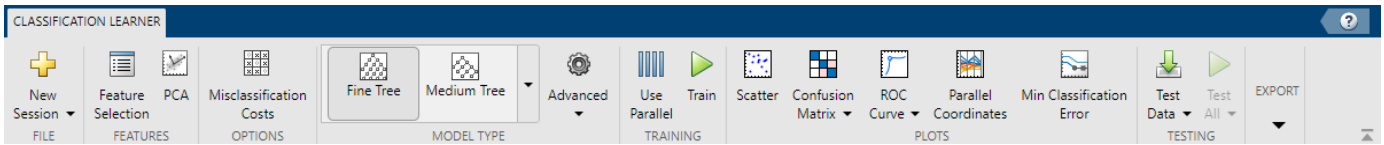
This example uses Fisher's iris data set, which contains measurements of flowers (petal length, petal width, sepal length, and sepal width) for specimens from three species. Train naive Bayes classifiers to predict the species based on the predictor measurements.

1. In the MATLAB Command Window, load the Fisher iris data set and create a table of measurement predictors (or features) using variables from the data set.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

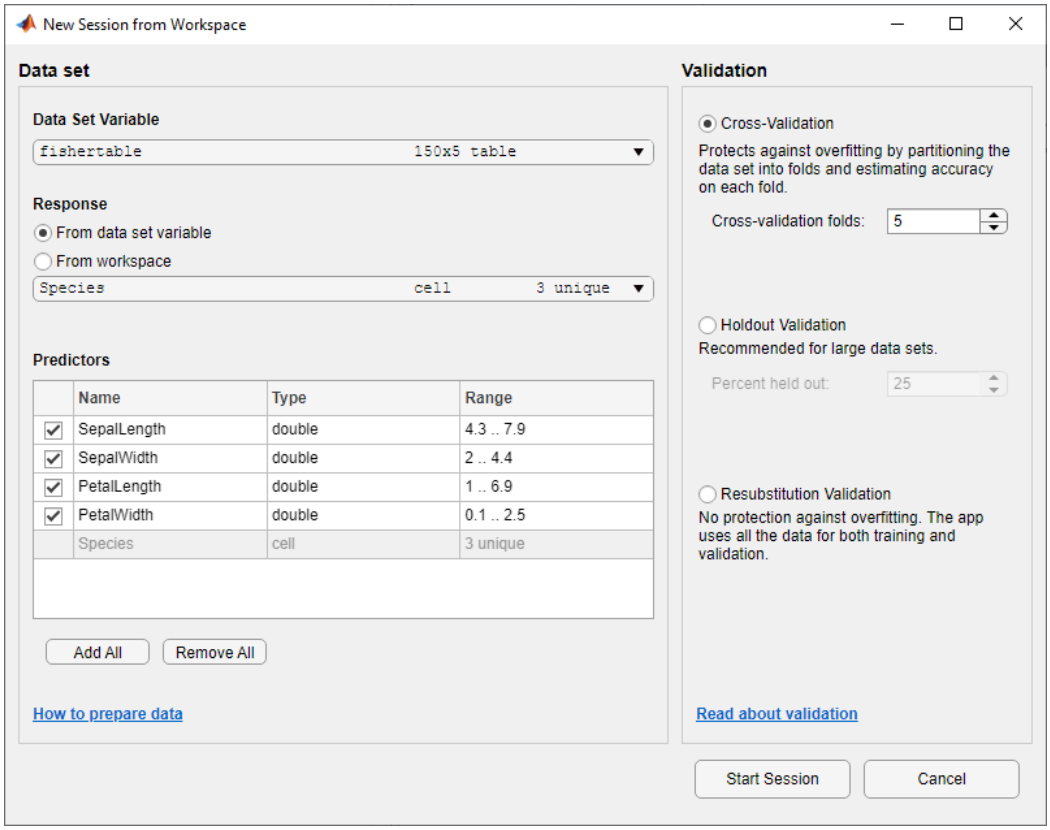
2. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.

3. On the **Classification Learner** tab, in the **File** section, select **New Session > From Workspace**.



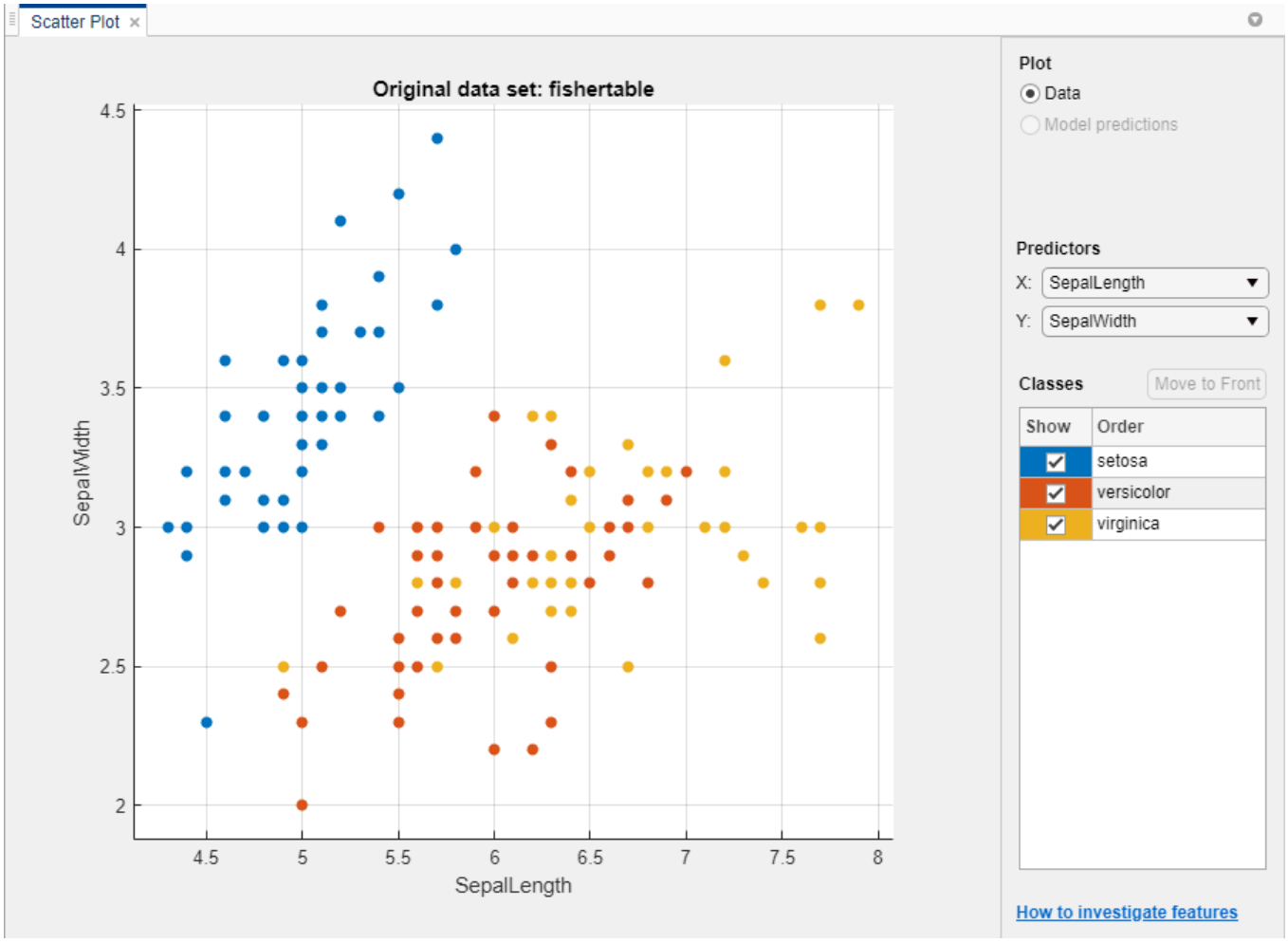
4. In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary).

   As shown in the dialog box, the app selects the response and predictor variables based on their data type. Petal and sepal length and width are predictors, and species is the response that you want to classify. For this example, do not change the selections.



5 To accept the default validation scheme and continue, click **Start Session**. The default validation option is cross-validation, to protect against overfitting.

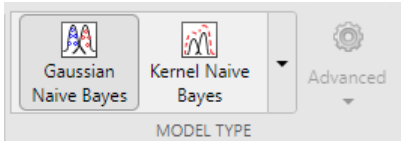
Classification Learner creates a scatter plot of the data.



6 Use the scatter plot to investigate which variables are useful for predicting the response. Select different options on the **X** and **Y** lists under **Predictors** to visualize the distribution of species and measurements. Observe which variables separate the species colors most clearly.

The `setosa` species (blue points) is easy to separate from the other two species with all four predictors. The `versicolor` and `virginica` species are much closer together in all predictor measurements and overlap, especially when you plot sepal length and width. `setosa` is easier to predict than the other two species.

7 Create a naive Bayes model. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Naive Bayes Classifiers** group, click **Gaussian Naive Bayes**. Note that Classification Learner disables the **Advanced** button in the **Model Type** section, because this type of model has no advanced settings.

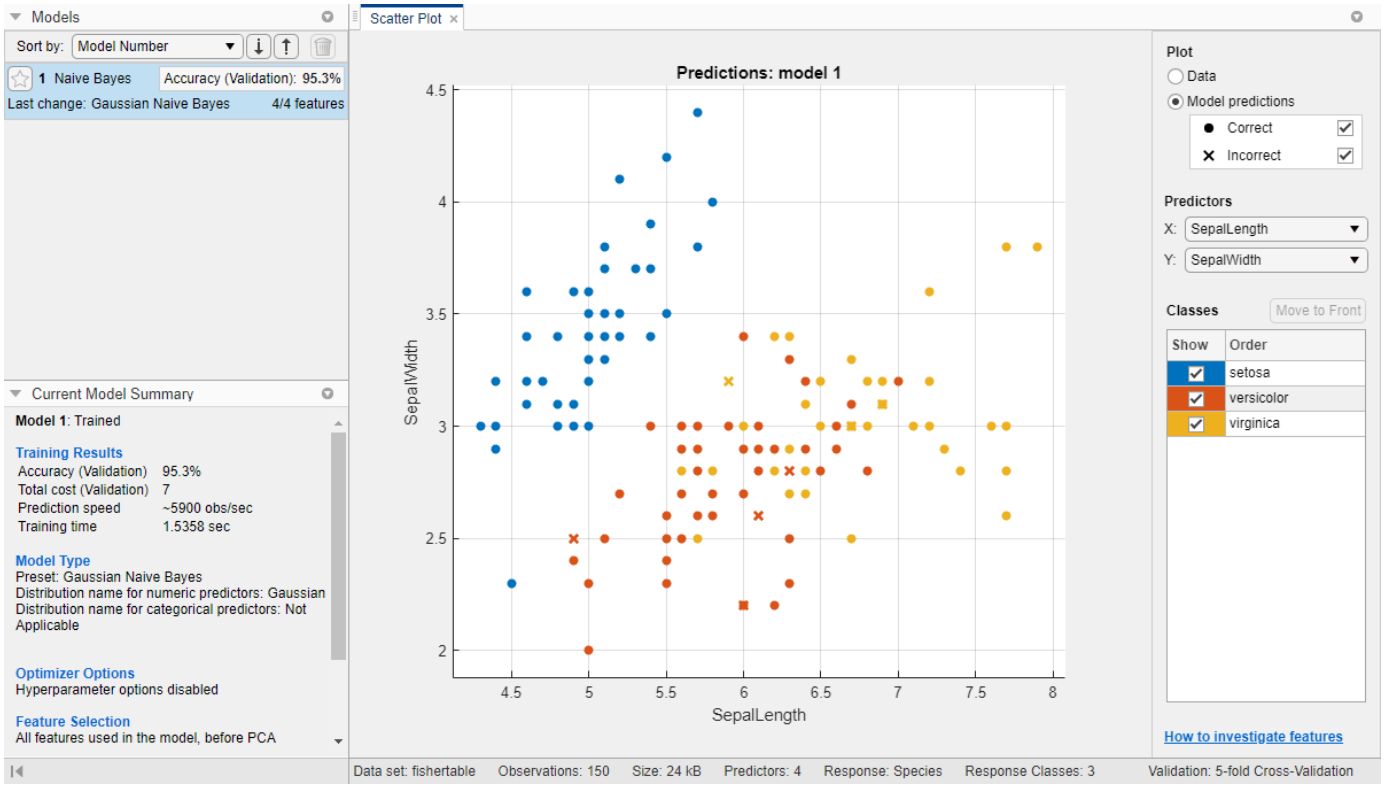


In the **Training** section, click **Train**.

The app creates a Gaussian naive Bayes model, and plots the results.

The app display the **Gaussian Naive Bayes** model in the **Models** pane. Check the model validation score in the **Accuracy (Validation)** box. The score shows that the model performs well.

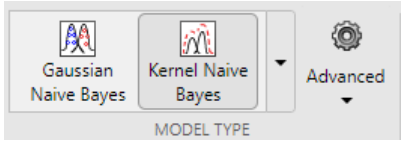
For the **Gaussian Naive Bayes** model, by default, the app models the distribution of numerical predictors using the Gaussian distribution, and models the distribution of categorical predictors using the multivariate multinomial distribution (MVMN).



Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

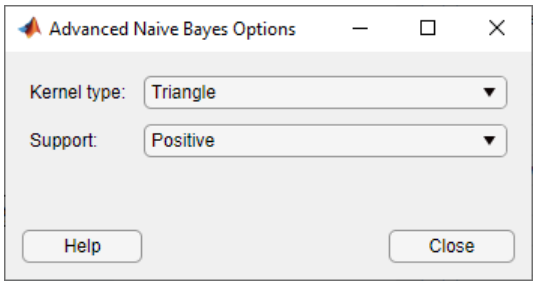
8 Examine the scatter plot. An X indicates misclassified points. The blue points (`setosa` species) are all correctly classified, but the other two species have misclassified points. Under **Plot**, switch between the **Data** and **Model predictions** options. Observe the color of the incorrect (X) points. Or, to view only the incorrect points, clear the **Correct** check box.

9 Train a kernel naive Bayes model for comparison. On the **Classification Learner** tab, in the **Model Type** section, click **Kernel Naive Bayes**. Note that Classification Learner enables the **Advanced** button, because this type of model has advanced settings.



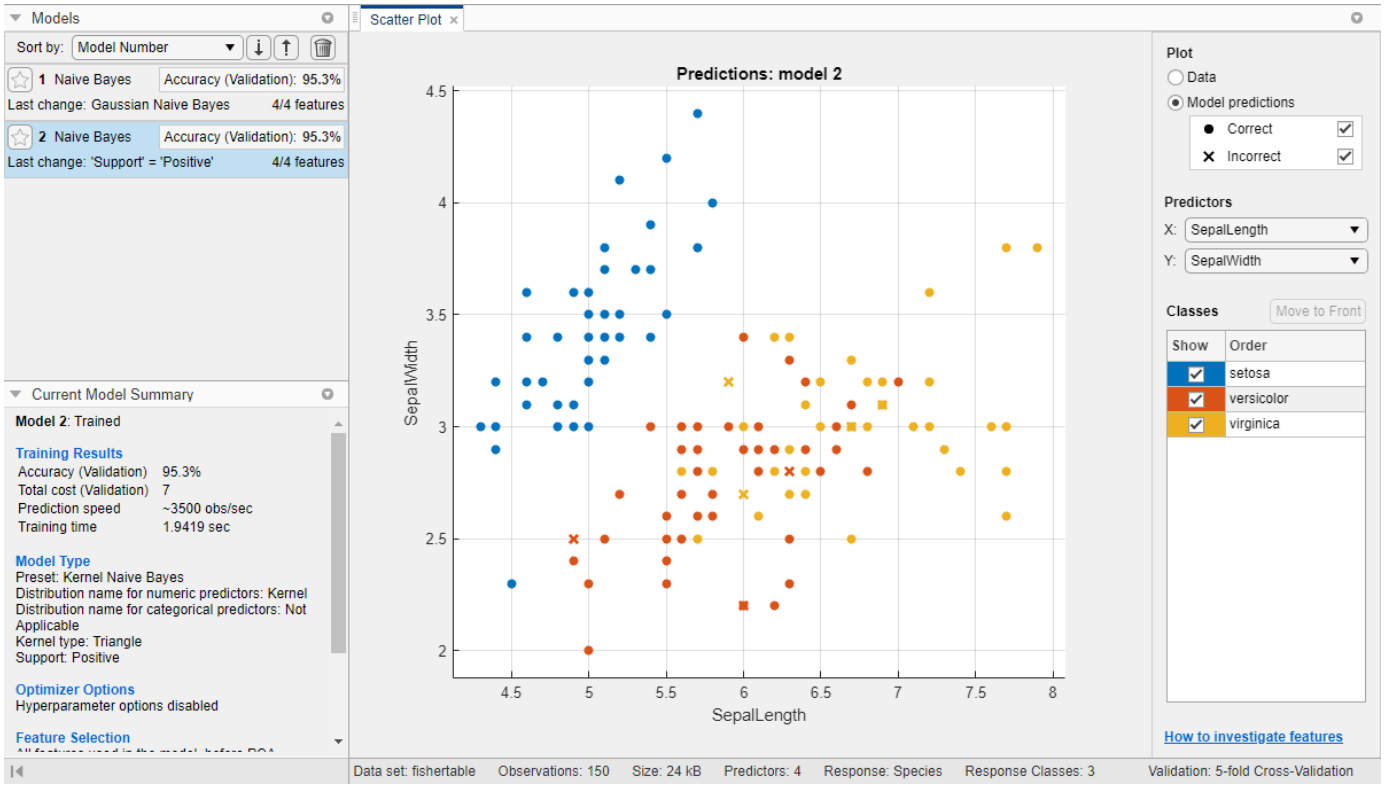
The app displays a draft kernel naive Bayes model in the **Models** pane.

In the **Model Type** section, click **Advanced** to change settings in the Advanced Naive Bayes Options dialog box. Select `Triangle` from the **Kernel Type** list, and select `Positive` from the **Support** list.



Note The settings in the Advanced Naive Bayes Options dialog box are available for continuous data only. Pointing to **Kernel Type** displays the tooltip "Specify Kernel smoothing function for continuous variables," and pointing to **Support** displays the tooltip "Specify Kernel smoothing density support for continuous variables."

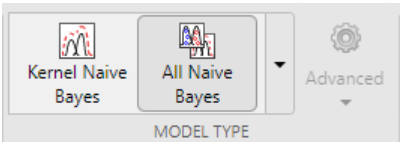
In the **Training** section, click **Train** to train the new model.



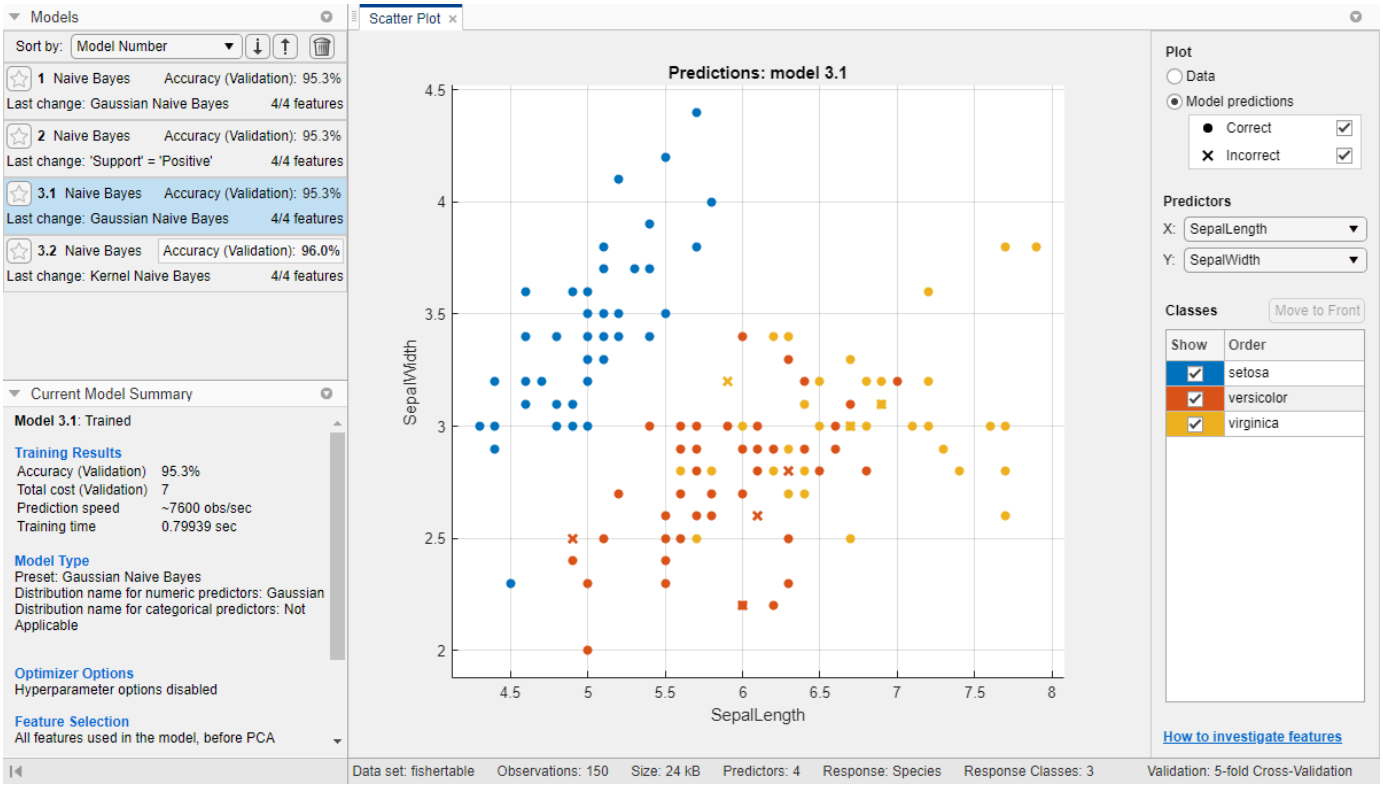
The **Models** pane now includes the new kernel naive Bayes model. Its model validation score is better than the score for the Gaussian naive Bayes model. The app highlights the **Accuracy (Validation)** score of the best model by outlining it in a box.

10 In the **Models** pane, click each model to view and compare the results.

11 Train a Gaussian naive Bayes model and a kernel naive Bayes model simultaneously. On the **Classification Learner** tab, in the **Model Type** section, click **All Naive Bayes**. Classification Learner disables the **Advanced** button. In the **Training** section, click **Train**.



The app trains one of each naive Bayes model type and highlights the **Accuracy (Validation)** score of the best model or models.



12 In the **Models** pane, click a model to view the results. Examine the scatter plot for the trained model and try plotting different predictors. Misclassified points appear as an X.

13 To inspect the accuracy of the predictions in each class, on the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. The app displays a matrix of true class and predicted class results.

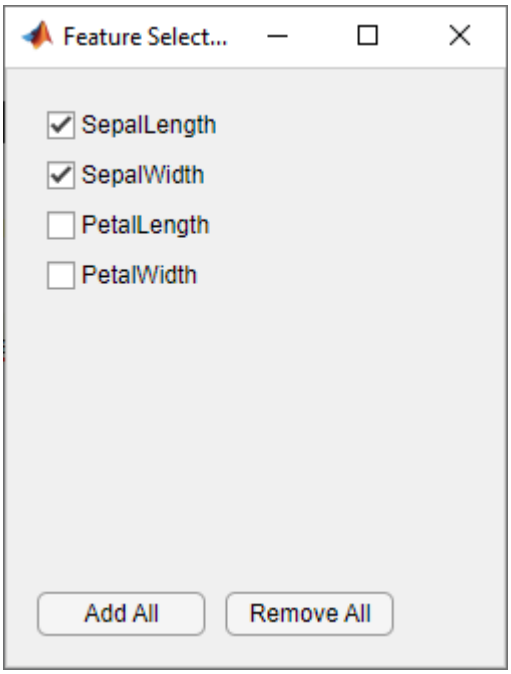
Note Validation introduces some randomness into the results. Your confusion matrix results can vary from the results shown in this example.

14 In the **Models** pane, click the other models and compare their results.

15 In the **Models** pane, click the model with the highest **Accuracy (Validation)** score. To improve the model, try modifying its features. For example, see if you can improve the model by removing features with low predictive power.

On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**.

In the Feature Selection dialog box, clear the check boxes for **PetalLength** and **PetalWidth** to exclude them from the predictors. A new draft model (model 4) appears in the **Models** pane with the new settings (2/4 features), based on the kernel naive Bayes model (model 3.2).



In the **Training** section, click **Train** to train a new kernel naive Bayes model using the new predictor options.

The **Models** pane now includes model 4. It is also a kernel naive Bayes model, trained using only 2 of 4 predictors.

16 To determine which predictors are included, click a model in the **Models** pane, then click **Feature Selection** in the **Features** section and note which check boxes are selected. The model with only sepal measurements (model 4) has a much lower **Accuracy (Validation)** score than the models containing all predictors.

17 Train another kernel naive Bayes model including only the petal measurements. Change the selections in the Feature Selection dialog box and click **Train**.

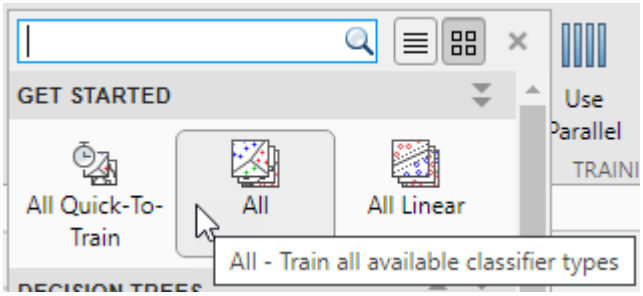
The model trained using only petal measurements (model 5) performs comparably to the model containing all predictors. The models predict no better using all the measurements compared to only the petal measurements. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily without some predictors.

18. To investigate features to include or exclude, use the parallel coordinates plot. On the **Classification Learner** tab, in the **Plots** section, click **Parallel Coordinates**.
19. In the **Models** pane, click the model with the highest **Accuracy (Validation)** score. To improve the model further, try changing naive Bayes settings (if available). On the **Classification Learner** tab, in the **Model Type** section, click **Advanced**. Recall that the **Advanced** button is enabled only for some models. Change a setting, then train the new model by clicking **Train**.
20. Export the trained model to the workspace. On the **Classification Learner** tab, in the **Export** section, select **Export Model > Export Model**. See “Export Classification Model to Predict New Data” on page 23-77.
21. Examine the code for training this classifier. In the **Export** section, click **Generate Function**.

Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To try all the nonoptimizable classifier model presets available for your data set:

1. Click the arrow on the **Model Type** section to open the gallery of classifiers.
2. In the **Get Started** group, click **All**, then click **Train** in the **Training** section.



For information about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Naive Bayes Classification” on page 21-2
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

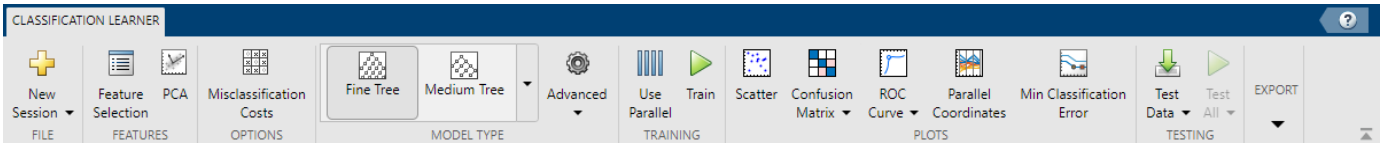
Train Neural Network Classifiers Using Classification Learner App

This example shows how to create and compare neural network classifiers in the Classification Learner app, and export trained models to the workspace to make predictions for new data.

1. In the MATLAB Command Window, load the `fisheriris` data set, and create a table from the variables in the data set to use for classification.

   ```
   fishertable = readtable('fisheriris.csv');
   ```

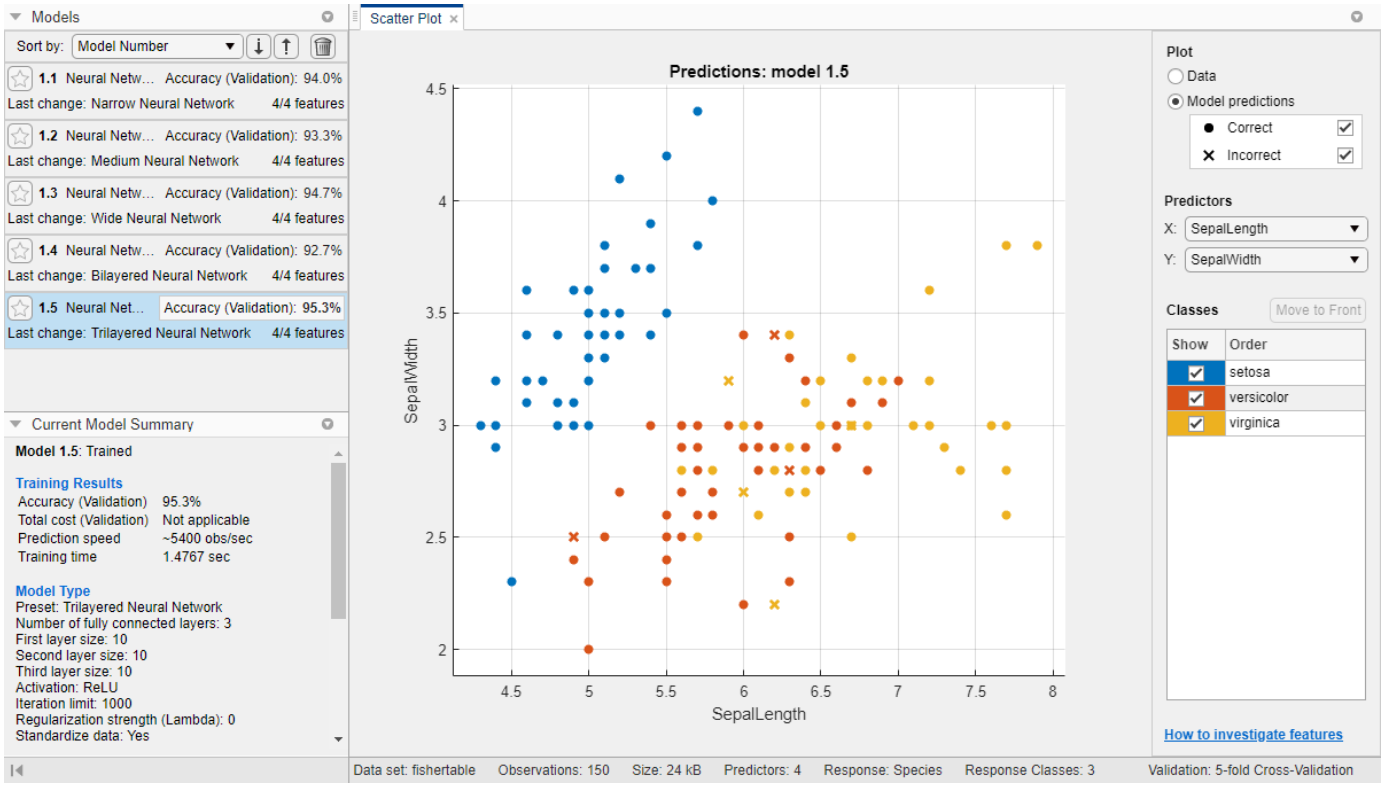
2. Click the **Apps** tab, and then click the **Show more** arrow on the right to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.



4. In the New Session from Workspace dialog box, select the table `fishertable` from the **Data Set Variable** list (if necessary). Observe that the app has selected response and predictor variables based on their data types. Petal and sepal length and width are predictors, and species is the response that you want to classify. For this example, do not change the selections.
5. To accept the default validation scheme and continue, click **Start Session**. The default validation option is 5-fold cross-validation, to protect against overfitting.

   Classification Learner creates a scatter plot of the data.
6. Use the scatter plot to investigate which variables are useful for predicting the response. Select different options in the **X** and **Y** lists under **Predictors** to visualize the distribution of species and measurements. Note which variables separate the species colors most clearly.
7. Create a selection of neural network models. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Neural Network Classifiers** group, click **All Neural Networks**.
8. In the **Training** section, click **Train**. Classification Learner trains one of each neural network classification option in the gallery. In the **Models** pane, the app outlines the **Accuracy (Validation)** score of the best model.

   Tip If you have Parallel Computing Toolbox, you can train all the models (**All Neural Networks**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.
9. Select a model in the **Models** pane to view the results. Examine the scatter plot for the trained model. Correctly classified points are marked with an O, and incorrectly classified points are marked with an X.



Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

10 Inspect the accuracy of the predictions in each class. On the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. View the matrix of true class and predicted class results.

11 Select the other models in the list for comparison.

12 Choose the best model in the **Models** pane (the best score is highlighted in the **Accuracy (Validation)** box). To improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power.

On the **Classification Learner** tab, in the **Features** section, click **Feature Selection**. In the Feature Selection dialog box, specify predictors to remove from the model, and click **Train** to train a new model using the new options. Compare results among the classifiers in the **Models** pane.

13 To investigate features to include or exclude, use the scatter and parallel coordinates plots. On the **Classification Learner** tab, in the **Plots** section, select **Parallel Coordinates**.

14 Choose the best model in the **Models** pane. To try to improve the model further, change its advanced settings. On the **Classification Learner** tab, in the **Model Type** section, click **Advanced** and select **Advanced**. Try changing some of the settings, like the sizes of the fully connected layers or the regularization strength, and then train the new model by clicking **Train**.

To learn more about neural network model settings, see “Neural Network Classifiers” on page 23-39.

15 You can export a full or compact version of the trained model to the workspace. On the **Classification Learner** tab, in the **Export** section, click **Export Model** and select either **Export Model** or **Export Compact Model**. See “Export Classification Model to Predict New Data” on page 23-77.

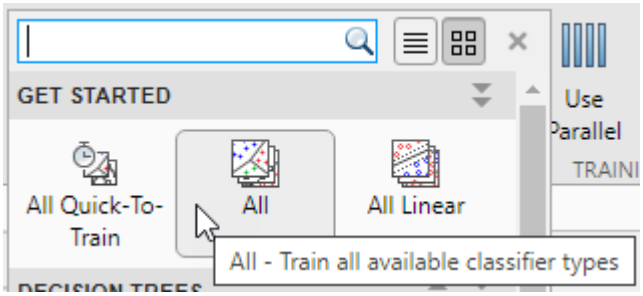
16 To examine the code for training this classifier, click **Generate Function** in the **Export** section.

Tip Use the same workflow to evaluate and compare the other classifier types you can train in Classification Learner.

To train all the nonoptimizable classifier model presets available for your data set:

1 On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery of models.

2 In the **Get Started** group, click **All**. Then, in the **Training** section, click **Train**.



To learn about other classifier types, see “Train Classification Models in Classification Learner App” on page 23-10.

See Also

Related Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Decision Trees Using Classification Learner App” on page 23-83

Train and Compare Classifiers Using Misclassification Costs in Classification Learner App

This example shows how to create and compare classifiers that use specified misclassification costs in the Classification Learner app. Specify the misclassification costs before training, and use the accuracy and total misclassification cost results to compare the trained models.

1 In the MATLAB Command Window, read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Combine all the `A` ratings into one rating. Do the same for the `B` and `C` ratings, so that the response variable has three distinct ratings. Among the three ratings, `A` is considered the best and `C` the worst.

```
creditrating = readtable('CreditRating_Historical.dat');
Rating = categorical(creditrating.Rating);
Rating = mergecats(Rating,{'AAA','AA','A'},'A');
Rating = mergecats(Rating,{'BBB','BB','B'},'B');
Rating = mergecats(Rating,{'CCC','CC','C'},'C');
creditrating.Rating = Rating;
```

2 Assume these are the costs associated with misclassifying the credit ratings of customers.

		Customer Predicted Rating		
		A	B	C
Customer True Rating	A	$0	$100	$200
	B	$500	$0	$100
	C	$1000	$500	$0

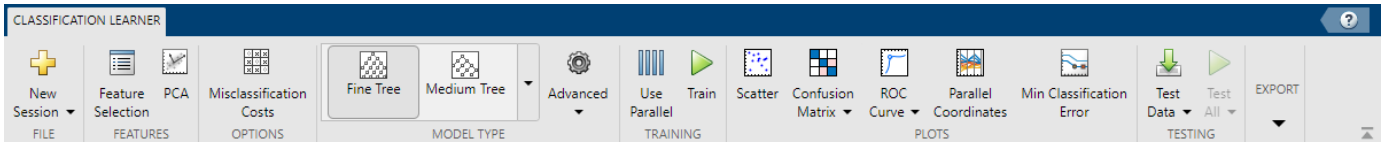
For example, the cost of misclassifying a `C` rating customer as an `A` rating customer is $1000. The costs indicate that classifying a customer with bad credit as a customer with good credit is more costly than classifying a customer with good credit as a customer with bad credit.

Create a matrix variable that contains the misclassification costs. Create another variable that specifies the class names and their order in the matrix variable.

```
ClassificationCosts = [0 100 200; 500 0 100; 1000 500 0];
ClassNames = categorical({'A','B','C'});
```

Tip Alternatively, you can specify misclassification costs directly inside the Classification Learner app. See “Specify Misclassification Costs” on page 23-48 for more information.

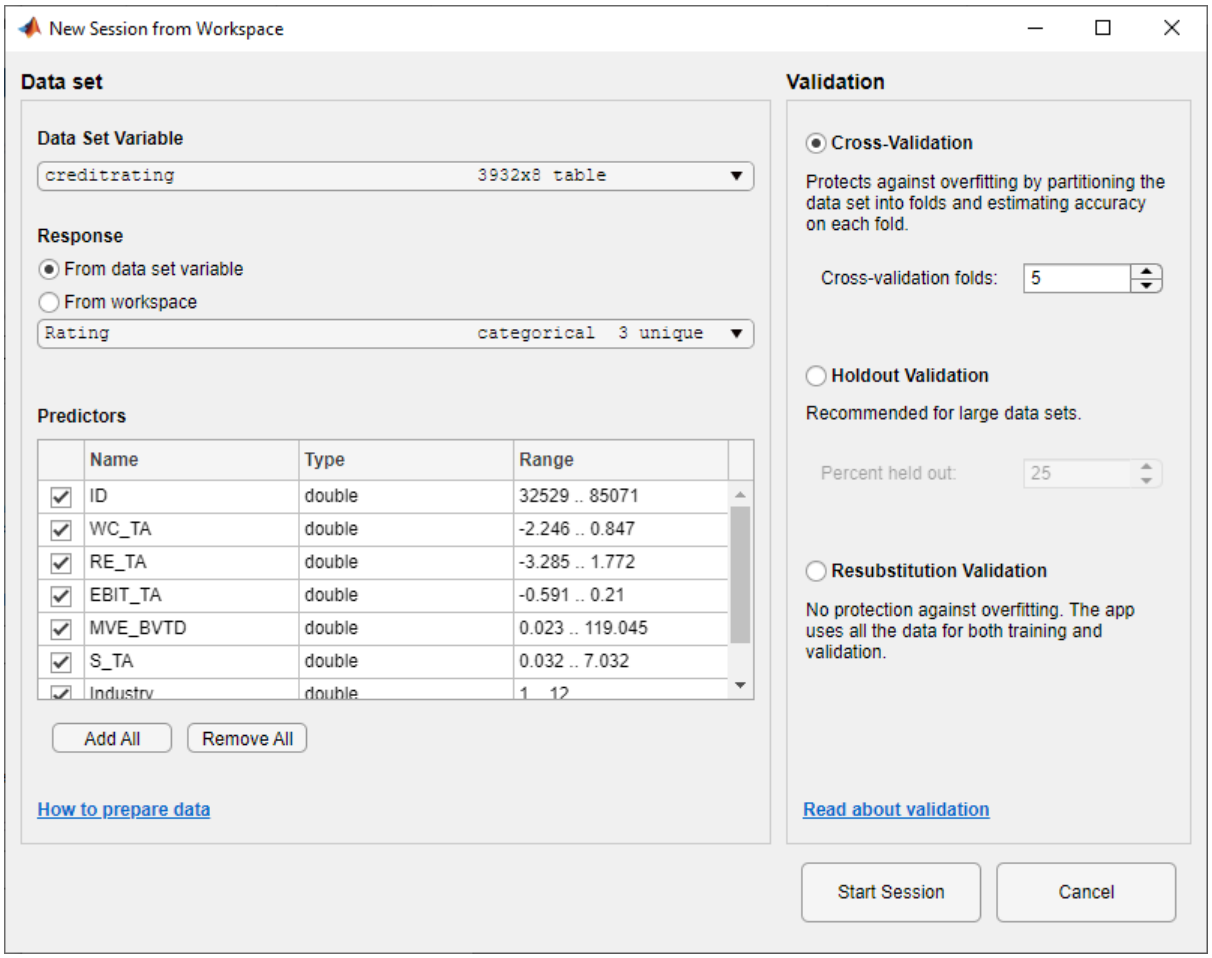
3 Open Classification Learner. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.



4 On the **Classification Learner** tab, in the **File** section, select **New Session > From Workspace**.

5 In the New Session from Workspace dialog box, select the table `creditrating` from the **Data Set Variable** list.

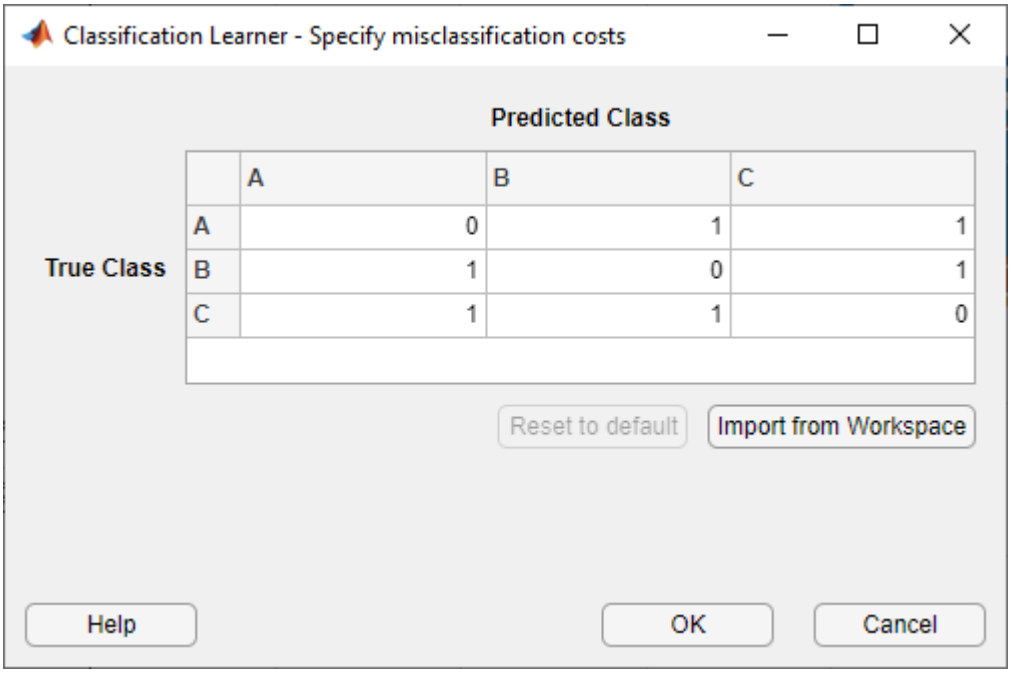
As shown in the dialog box, the app selects the response and predictor variables based on their data type. The default response variable is the `Rating` variable. The default validation option is cross-validation, to protect against overfitting. For this example, do not change the default settings.



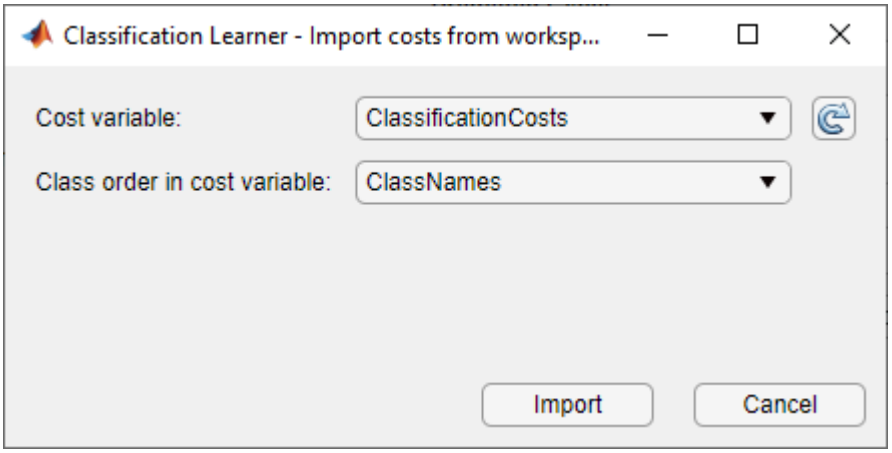
6 To accept the default settings, click **Start Session**.

7 Specify the misclassification costs. On the **Classification Learner** tab, in the **Options** section, click **Misclassification Costs**. The app opens a dialog box showing the default misclassification costs.

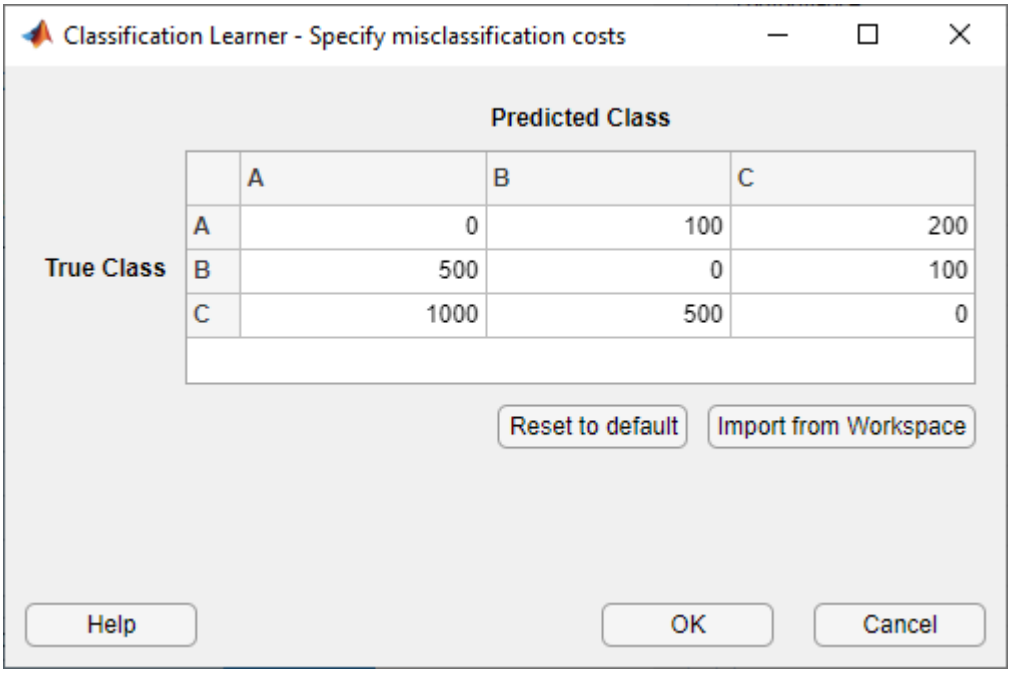
In the dialog box, click **Import from Workspace**.



In the import dialog box, select `ClassificationCosts` as the cost variable and `ClassNames` as the class order in the cost variable. Click **Import**.



The app updates the values in the misclassification costs dialog box. Click **OK** to save your changes.



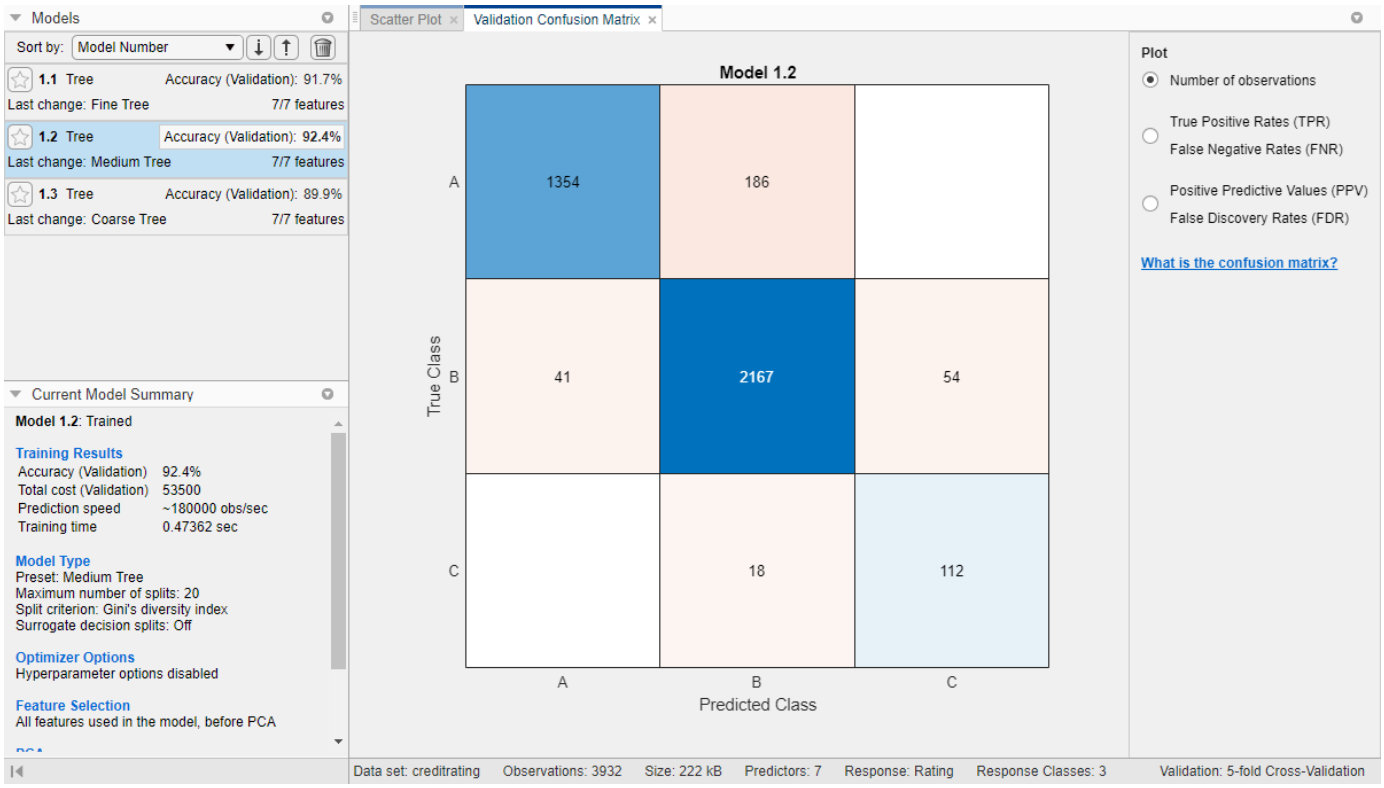
8 Train fine, medium, and coarse trees simultaneously. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Decision Trees** group, click **All Trees**. In the **Training** section, click **Train**. The app trains one of each tree model type and displays the models in the **Models** pane.

Tip If you have Parallel Computing Toolbox, you can train all the models (**All Trees**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

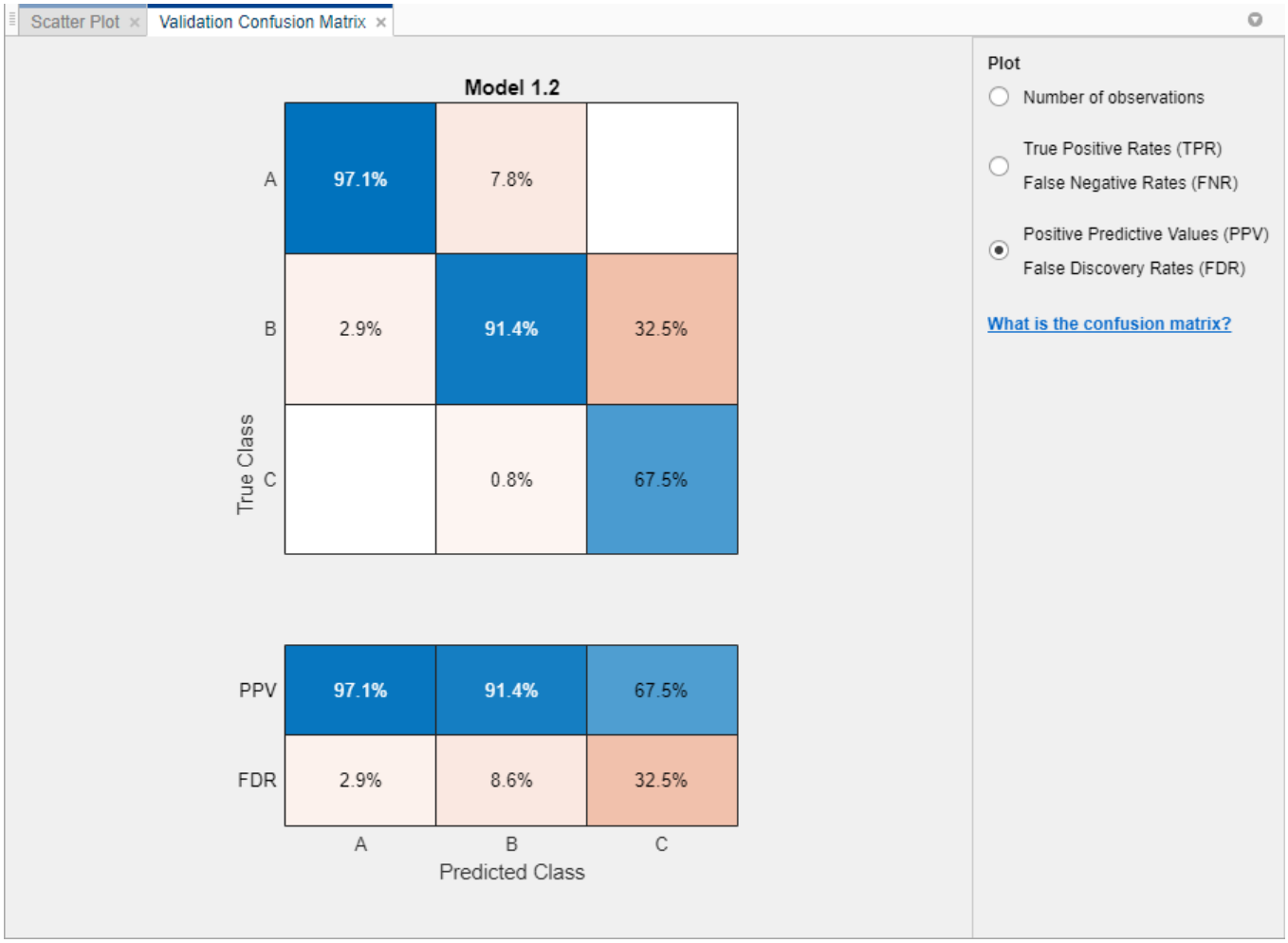
9 In the **Models** pane, click a model to view the results, which are displayed in the **Current Model Summary** pane. Each model has a validation accuracy score that indicates the percentage of correctly predicted responses. In the **Models** pane, the app highlights the highest **Accuracy (Validation)** score by outlining it in a box.

10 Inspect the accuracy of the predictions in each class. On the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**. The app displays a matrix of true class and predicted class results for the selected model (in this case, for the medium tree).

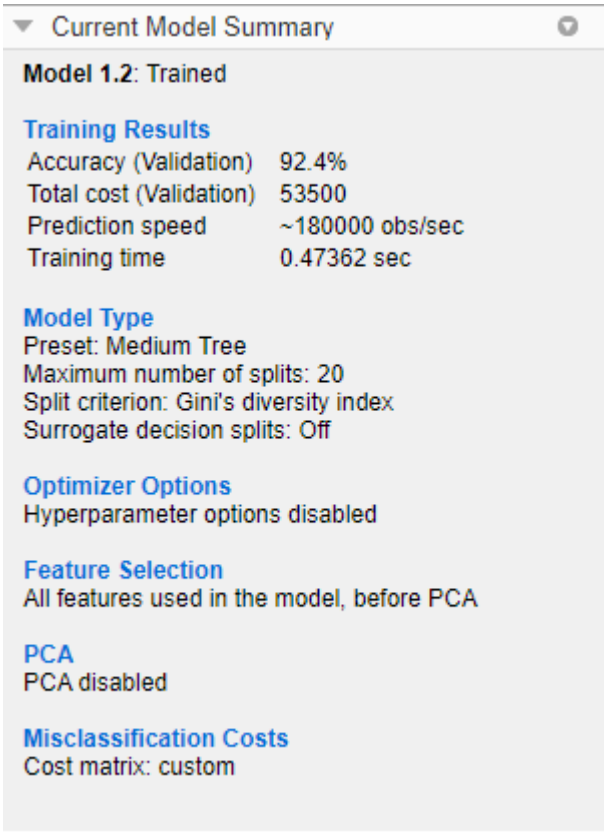


11 You can also plot results per predicted class to investigate false discovery rates. Under **Plot**, select the **Positive Predictive Values (PPV) False Discovery Rates (FDR)** option.

In the confusion matrix for the medium tree, the entries below the diagonal have small percentage values. These values indicate that the model tries to avoid assigning a credit rating that is higher than the true rating for a customer.



12 Compare the total misclassification costs of the tree models. To inspect the total misclassification cost of a model, select the model in the **Models** pane, and then view the **Training Results** section of the **Current Model Summary** pane. For example, the medium tree has these results.



In general, choose a model that has high accuracy and low total misclassification cost. In this example, the medium tree has the highest validation accuracy value and the lowest total misclassification cost of the three models.

You can perform feature selection and transformation or tune your model just as you do in the workflow without misclassification costs. However, always check the total misclassification cost of your model when assessing its performance. For differences in the exported model and exported code when you use misclassification costs, see “Misclassification Costs in Exported Model and Generated Code” on page 23-52.

See Also

Related Examples

- “Misclassification Costs in Classification Learner App” on page 23-48
- “Train Classification Models in Classification Learner App” on page 23-10

Train Classifier Using Hyperparameter Optimization in Classification Learner App

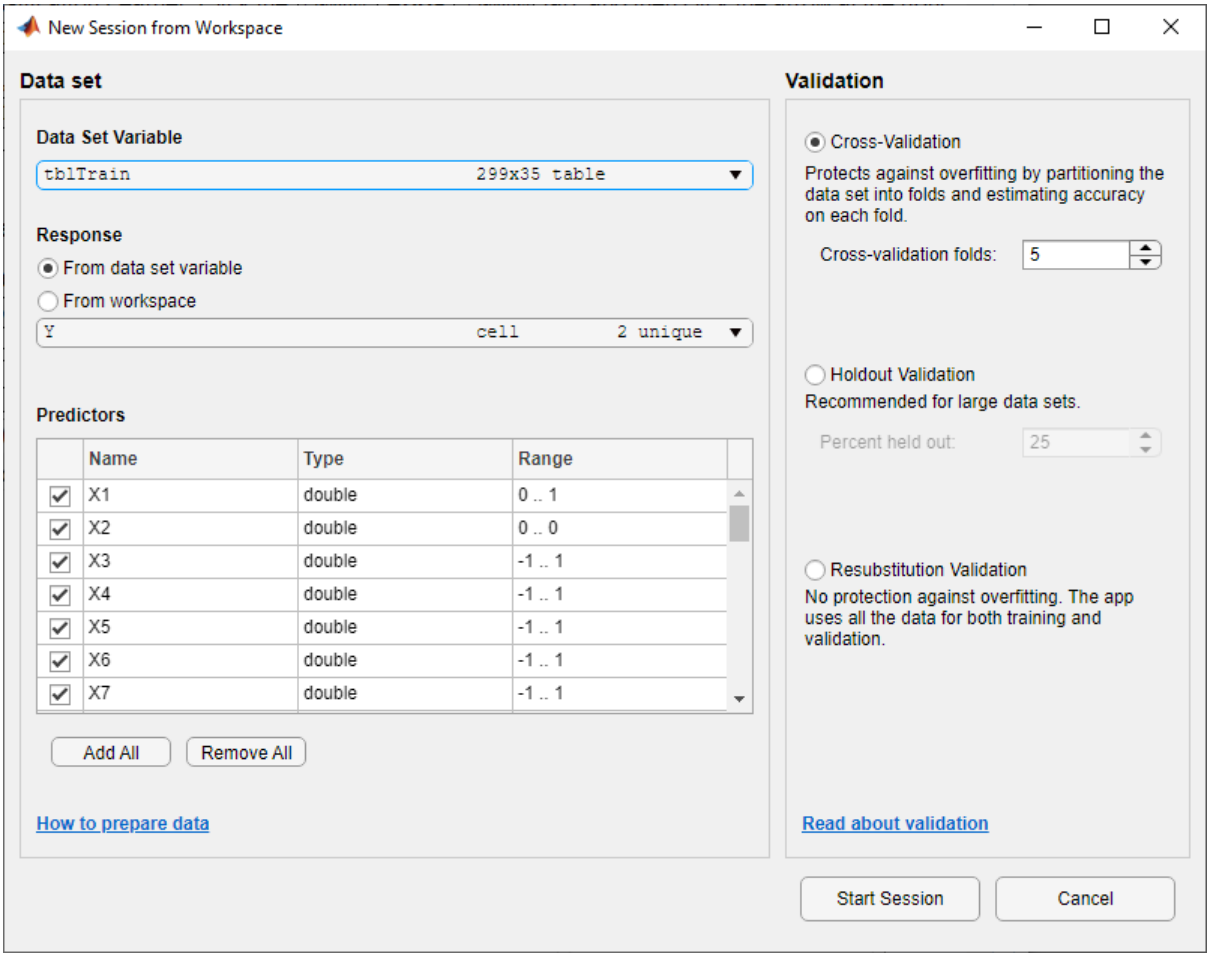
This example shows how to tune hyperparameters of a classification support vector machine (SVM) model by using hyperparameter optimization in the Classification Learner app. Compare the test set performance of the trained optimizable SVM to that of the best-performing preset SVM model.

1. In the MATLAB Command Window, load the `ionosphere` data set, and create a table containing the data. Separate the table into training and test sets.

   ```
   load ionosphere
   tbl = array2table(X);
   tbl.Y = Y;

   rng('default') % For reproducibility of the data split
   partition = cvpartition(Y,'Holdout',0.15);
   idxTrain = training(partition); % Indices for the training set
   tblTrain = tbl(idxTrain,:);
   tblTest = tbl(~idxTrain,:);
   ```

2. Open Classification Learner. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.
3. On the **Classification Learner** tab, in the **File** section, select **New Session > From Workspace**.
4. In the New Session from Workspace dialog box, select the `tblTrain` table from the **Data Set Variable** list.

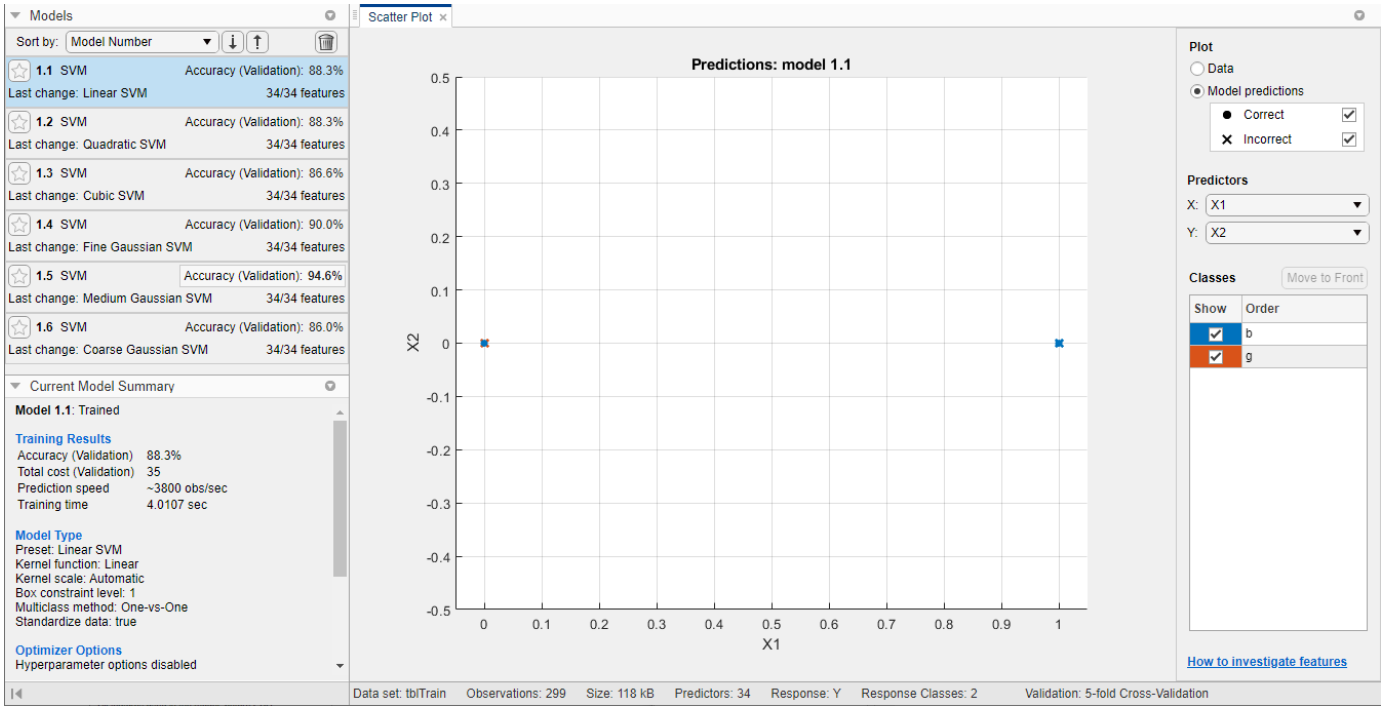
   As shown in the dialog box, the app selects the response and predictor variables. The default response variable is `Y`. The default validation option is 5-fold cross-validation, to protect against overfitting. For this example, do not change the default settings.



5 To accept the default options and continue, click **Start Session**.

6 Train all preset SVM models. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Support Vector Machines** group, click **All SVMs**. In the **Training** section, click **Train**. The app trains one of each SVM model type and displays the models in the **Models** pane.

Tip If you have Parallel Computing Toolbox, you can train all the SVM models (**All SVMs**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the SVM models simultaneously.

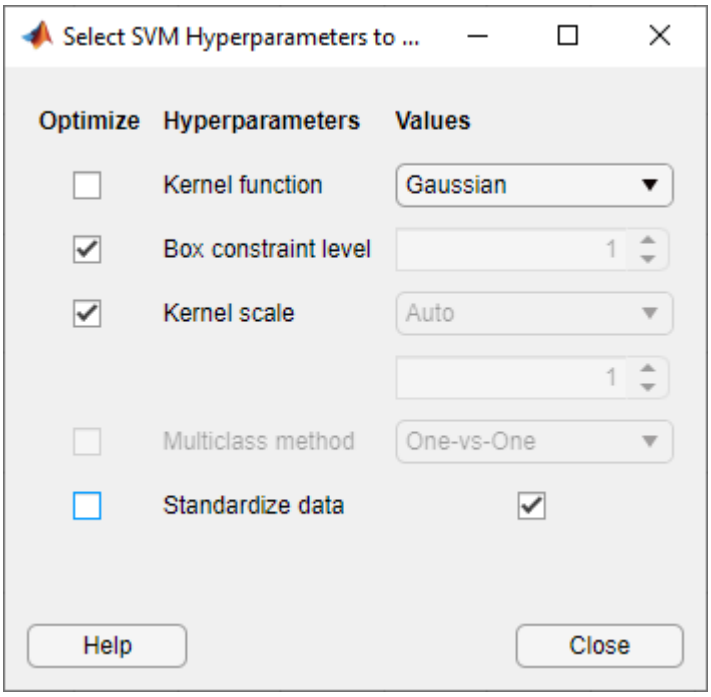


The app displays a scatter plot of the `ionosphere` data. Correctly classified points are marked with an O, and incorrectly classified points are marked with an X. The **Models** pane on the left shows the validation accuracy for each model.

Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

7. Select an optimizable SVM model to train. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Support Vector Machines** group, click **Optimizable SVM**. The app disables the **Use Parallel** button when you select an optimizable model.

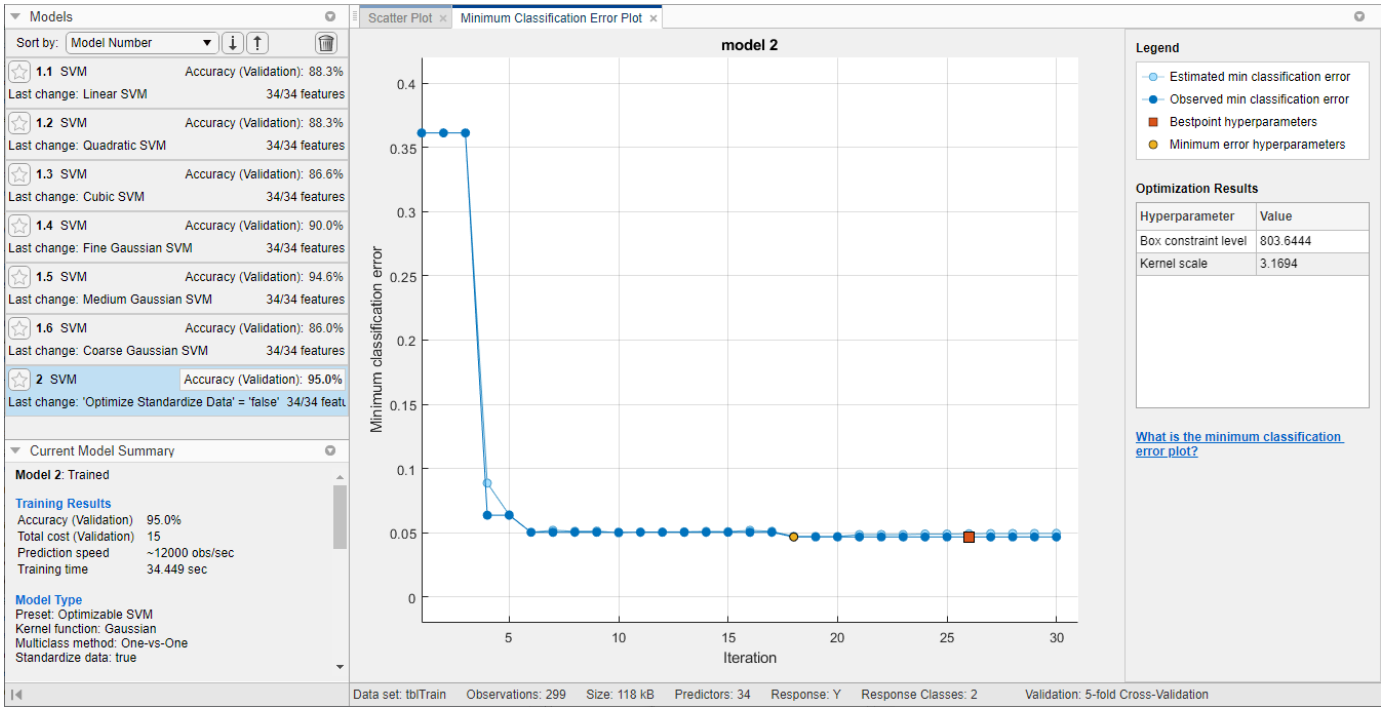
8. Select the model hyperparameters to optimize. In the **Model Type** section, select **Advanced > Advanced**. The app opens a dialog box in which you can select **Optimize** check boxes for the hyperparameters that you want to optimize. By default, all the check boxes for the available hyperparameters are selected. For this example, clear the **Optimize** check boxes for **Kernel function** and **Standardize data**. By default, the app disables the **Optimize** check box for **Kernel scale** whenever the kernel function has a fixed value other than `Gaussian`. Select a `Gaussian` kernel function, and select the **Optimize** check box for **Kernel scale**.



9 In the **Training** section, click **Train**.

10 The app displays a **Minimum Classification Error Plot** as it runs the optimization process. At each iteration, the app tries a different combination of hyperparameter values and updates the plot with the minimum validation classification error observed up to that iteration, indicated in dark blue. When the app completes the optimization process, it selects the set of optimized hyperparameters, indicated by a red square. For more information, see “Minimum Classification Error Plot” on page 23-61.

The app lists the optimized hyperparameters in both the **Optimization Results** section to the right of the plot and the **Optimized Hyperparameters** section of the **Current Model Summary** pane.



Note In general, the optimization results are not reproducible.

11 Compare the trained preset SVM models to the trained optimizable model. In the **Models** pane, the app highlights the highest **Accuracy (Validation)** by outlining it in a box. In this example, the trained optimizable SVM model outperforms the six preset models.

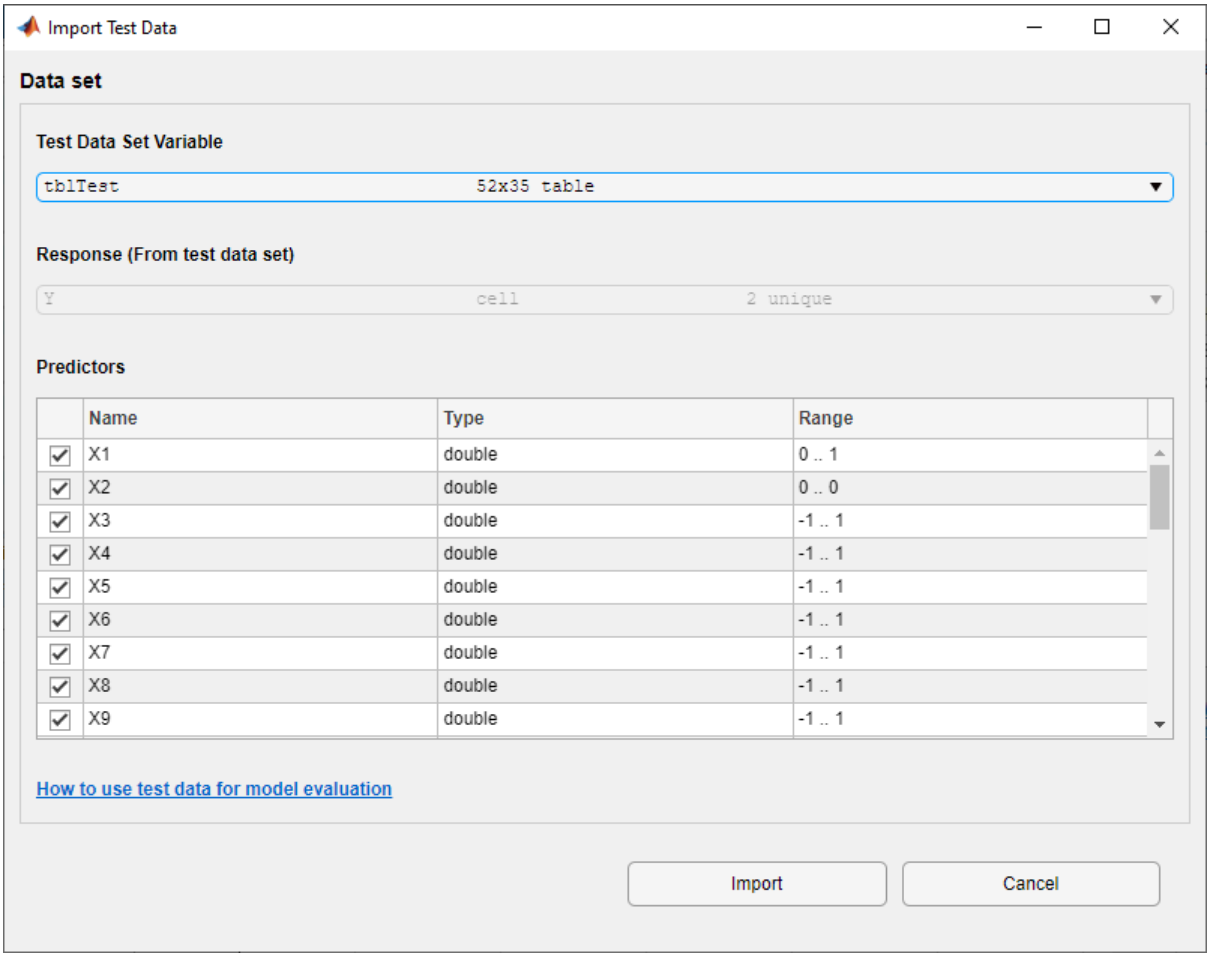
A trained optimizable model does not always have a higher accuracy than the trained preset models. If a trained optimizable model does not perform well, you can try to get better results by running the optimization for longer. In the **Model Type** section, select **Advanced > Optimizer Options**. In the dialog box, increase the **Iterations** value. For example, you can double-click the default value of `30` and enter a value of `60`.

12 Because hyperparameter tuning often leads to overfitted models, check the performance of the optimizable SVM model on a test set and compare it to the performance of the best preset SVM model. Begin by importing test data into the app.

On the **Classification Learner** tab, in the **Testing** section, select **Test Data > From Workspace**.

13 In the Import Test Data dialog box, select the `tblTest` table from the **Test Data Set Variable** list.

As shown in the dialog box, the app identifies the response and predictor variables.



14 Click **Import**.

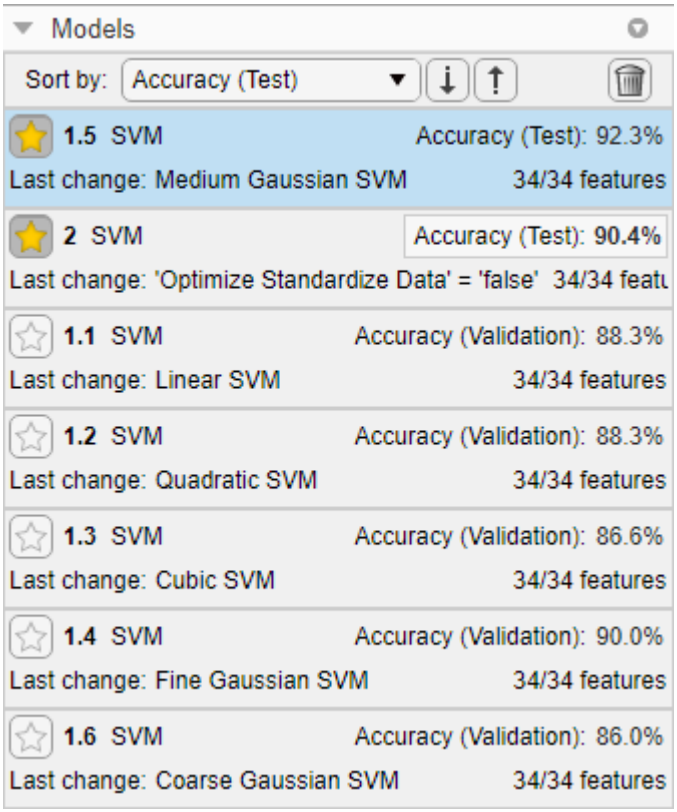
15 Compute the accuracy of the best preset model and the optimizable model on the `tblTest` data.

First, in the **Models** pane, click the star icons next to the **Medium Gaussian SVM** model and the **Optimizable SVM** model.

16 For each model, select the model in the **Models** pane, and then select **Test All > Test Selected** in the **Testing** section. The app computes the test set performance of the model trained on the full data set, including training and validation data.

17 Sort the models based on the test set accuracy. In the **Models** pane, open the **Sort by** list and select `Accuracy (Test)`.

In this example, the trained optimizable model does not perform as well as the trained preset model on the test set data.



18 In the **Plots** section on the **Classification Learner** tab, select **Confusion Matrix > Test Data**. Toggle between the medium Gaussian SVM model and the optimizable SVM model, and visually compare the two confusion matrices.


Scatter Plot
Minimum Classification Error Plot
Test Confusion Matrix
Model 1.5
15
3
1
33
b
g
True Class
b
g
Predicted Class

See Also

Related Examples

- “Hyperparameter Optimization in Classification Learner App” on page 23-54
- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Choose Classifier Options” on page 23-22
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Bayesian Optimization Workflow” on page 10-25

Check Classifier Performance Using Test Set in Classification Learner App

This example shows how to train multiple models in Classification Learner, and determine the best-performing models based on their validation accuracy. Check the test accuracy for the best-performing models trained on the full data set, including training and validation data.

1. In the MATLAB Command Window, load the `ionosphere` data set, and create a table containing the data. Separate the table into training and test sets.

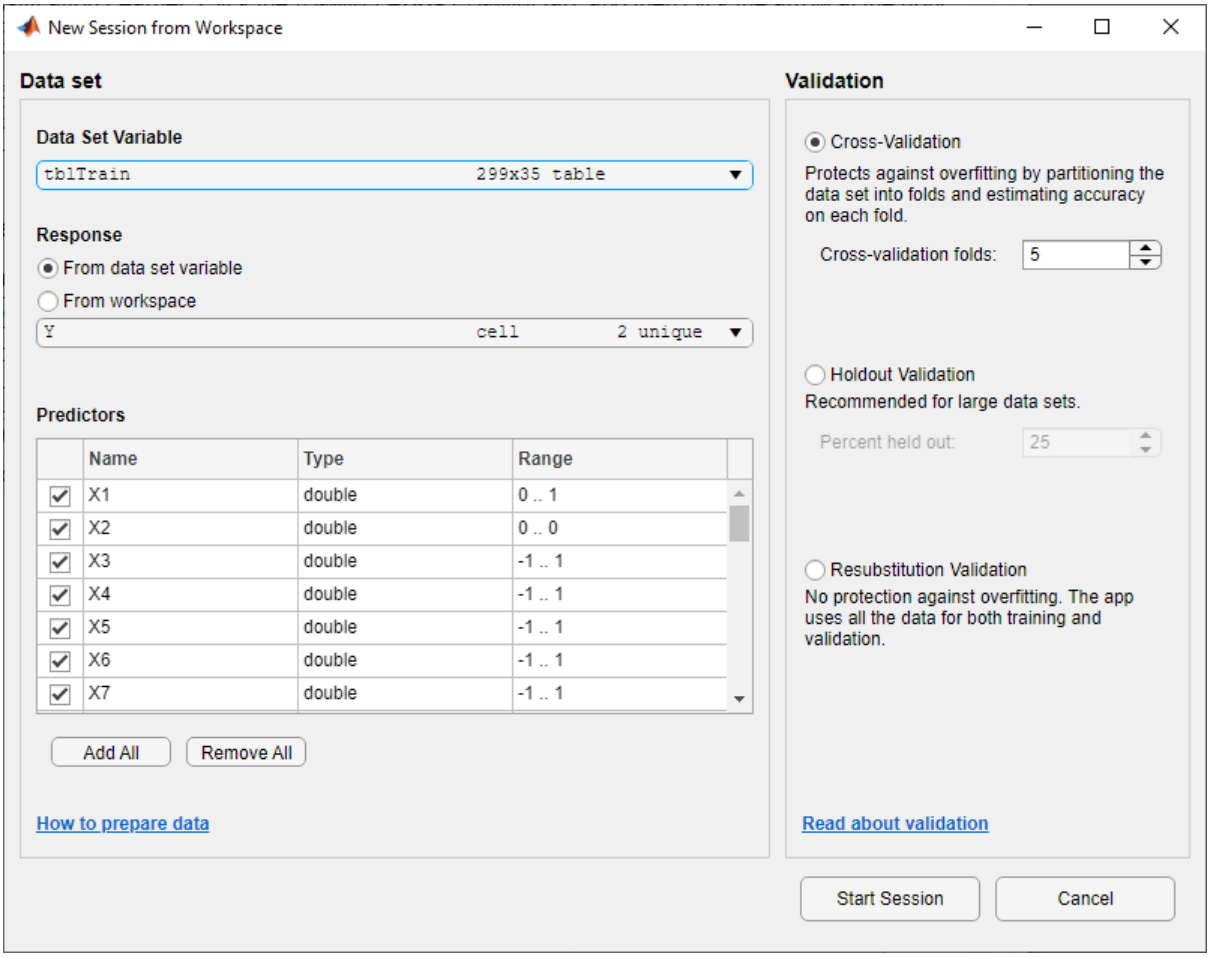
   ```
   load ionosphere
   tbl = array2table(X);
   tbl.Y = Y;

   rng('default') % For reproducibility of the data split
   partition = cvpartition(Y,'Holdout',0.15);
   idxTrain = training(partition); % Indices for the training set
   tblTrain = tbl(idxTrain,:);
   tblTest = tbl(~idxTrain,:);
   ```

2. Open Classification Learner. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.

3. On the **Classification Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.

4. In the New Session from Workspace dialog box, select the `tblTrain` table from the **Data Set Variable** list.

   As shown in the dialog box, the app selects the response and predictor variables. The default response variable is `Y`. To protect against overfitting, the default validation option is 5-fold cross-validation. For this example, do not change the default settings.



5 To accept the default options and continue, click **Start Session**.

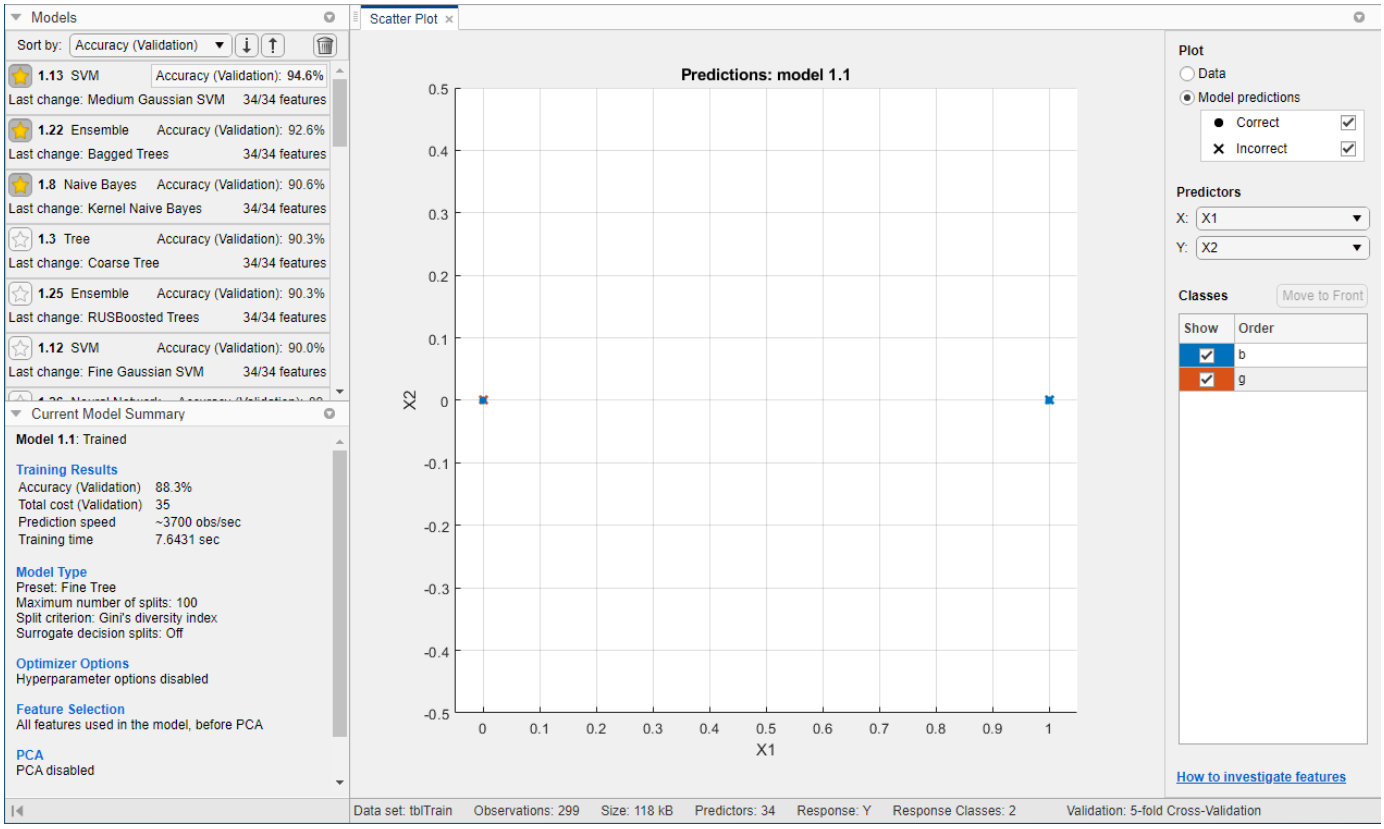
6 Train all preset models. On the **Classification Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Get Started** group, click **All**. In the **Training** section, click **Train**. The app trains one of each preset model type and displays the models in the **Models** pane.

Tip If you have Parallel Computing Toolbox, you can train all the models (**All**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

---

7 Sort the trained models based on the validation accuracy. In the **Models** pane, open the **Sort by** list and select `Accuracy (Validation)`.

8 In the **Models** pane, click the star icons next to the three models with the highest validation accuracy. The app highlights the highest validation accuracy by outlining it in a box. In this example, the trained **Medium Gaussian SVM** model has the highest validation accuracy.



The app displays a scatter plot of the `ionosphere` data. Correctly classified points are marked with an O, and incorrectly classified points are marked with an X. The **Models** pane on the left shows the validation accuracy for each model.

Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

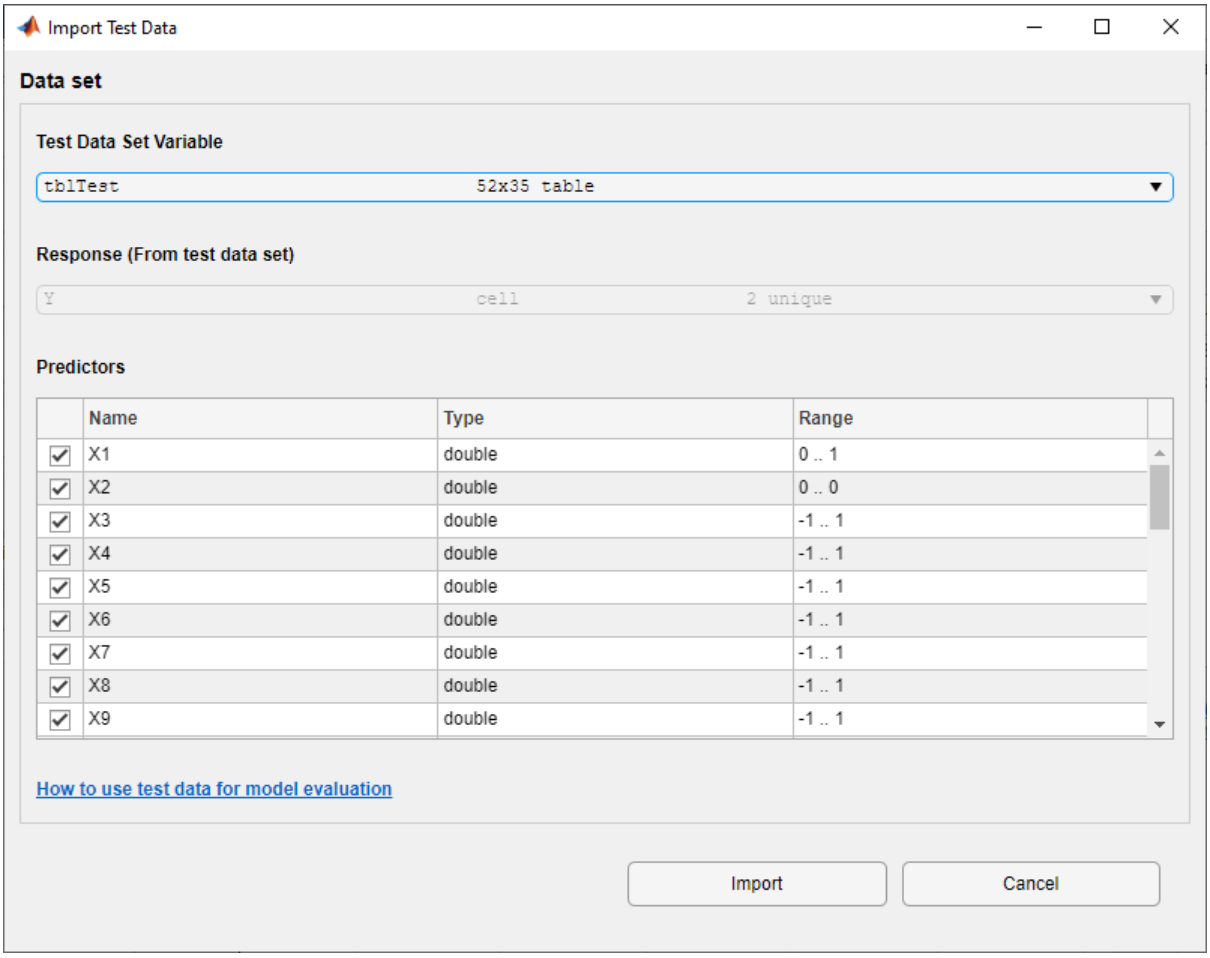
---

9 Check the test set performance of the best-performing models. Begin by importing test data into the app.

On the **Classification Learner** tab, in the **Testing** section, click **Test Data** and select **From Workspace**.

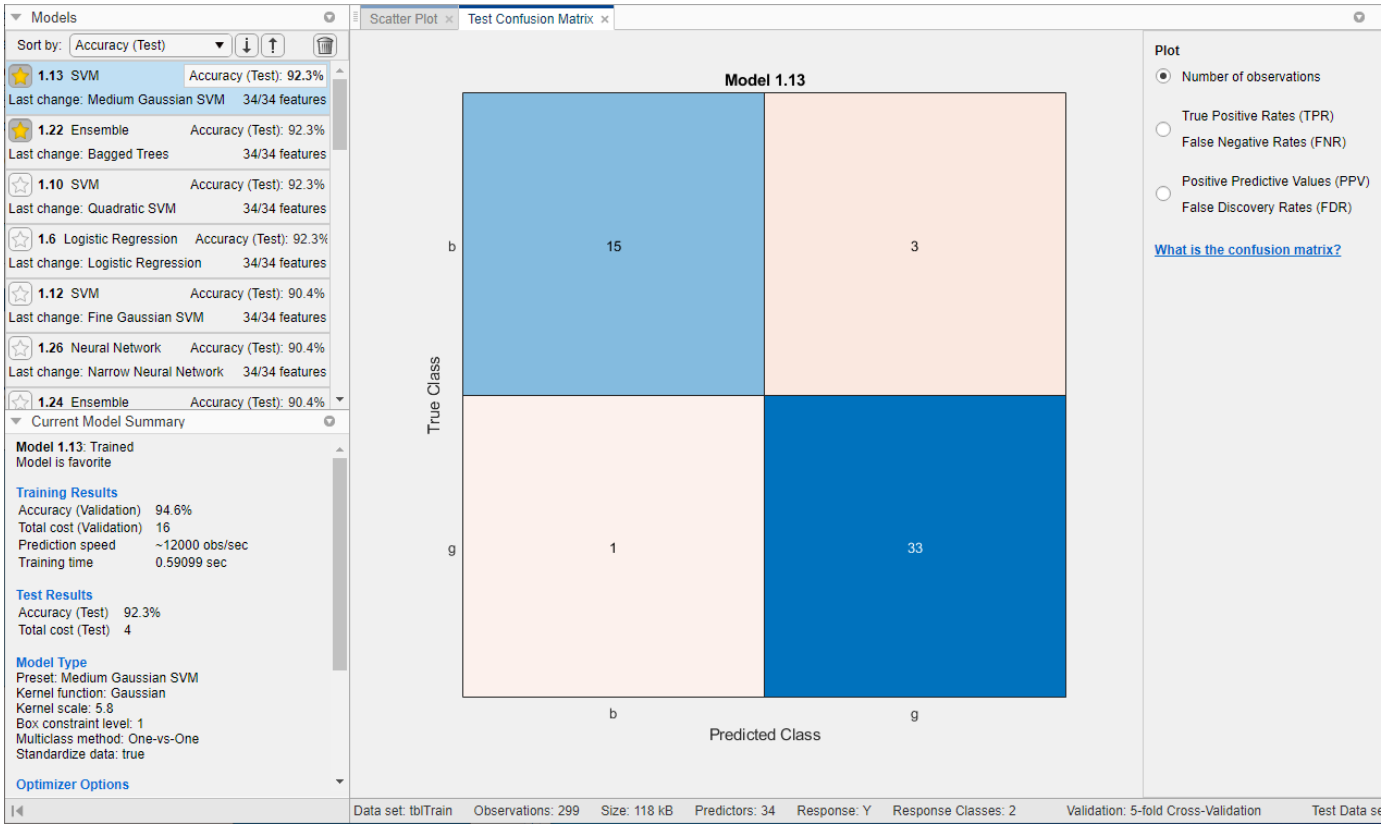
10 In the Import Test Data dialog box, select the `tblTest` table from the **Test Data Set Variable** list.

As shown in the dialog box, the app identifies the response and predictor variables.



11. Click **Import**.
12. Compute the accuracy of the best preset models on the `tblTest` data. For convenience, compute the test set accuracy for all models at once. On the **Classification Learner** tab, in the **Testing** section, click **Test All** and select **Test All**. The app computes the test set performance of the model trained on the full data set, including training and validation data.
13. Sort the models based on the test set accuracy. In the **Models** pane, open the **Sort by** list and select `Accuracy (Test)`. The app still outlines the metric for the model with the highest validation accuracy, despite displaying the test accuracy.
14. Visually check the test set performance of the models. On the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Test Data**. You can toggle between models to compare their performance.

    In this example, the trained **Medium Gaussian SVM** remains one of the best-performing models on the test set data.



15 Compare the validation and test accuracy for the trained **Medium Gaussian SVM** model. In the **Current Model Summary** pane, compare the **Accuracy (Validation)** value under **Training Results** to the **Accuracy (Test)** value under **Test Results**. In this example, the two values are close, which indicates that the validation accuracy is a good estimate of the test accuracy for this model.

See Also

Related Examples

- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77
- “Train Classifier Using Hyperparameter Optimization in Classification Learner App” on page 23-128

Regression Learner

Train Regression Models in Regression Learner App

In this section...

You can use Regression Learner to train regression models including linear regression models, regression trees, Gaussian process regression models, support vector machines, ensembles of regression trees, and neural network regression models. In addition to training models, you can explore your data, select features, specify validation schemes, and evaluate results. You can export a model to the workspace to use the model with new data or generate MATLAB code to learn about programmatic classification.

Training a model in Regression Learner consists of two parts:

- Validated Model: Training a model with a validation scheme. By default, the app protects against overfitting by applying cross-validation. Alternatively, you can choose holdout validation. The validated model is visible in the app.
- Full Model: Training a model on full data without validation. The app trains this model simultaneously with the validated model. However, the model trained on full data is not visible in the app. When you choose a regression model to export to the workspace, Regression Learner exports the full model.

The app displays the results of the validated model. Diagnostic measures, such as model accuracy, and plots, such as response plot or residuals plot reflect the validated model results. You can automatically train one or more regression models, compare validation results, and choose the best model that works for your regression problem. When you choose a model to export to the workspace, Regression Learner exports the full model. Because Regression Learner creates a model object of the full model during training, you experience no lag time when you export the model. You can use the exported model to make predictions on new data.

To get started by training a selection of model types, see “Automated Regression Model Training” on page 24-2. If you already know which regression model you want to train, see “Manual Regression Model Training” on page 24-4.

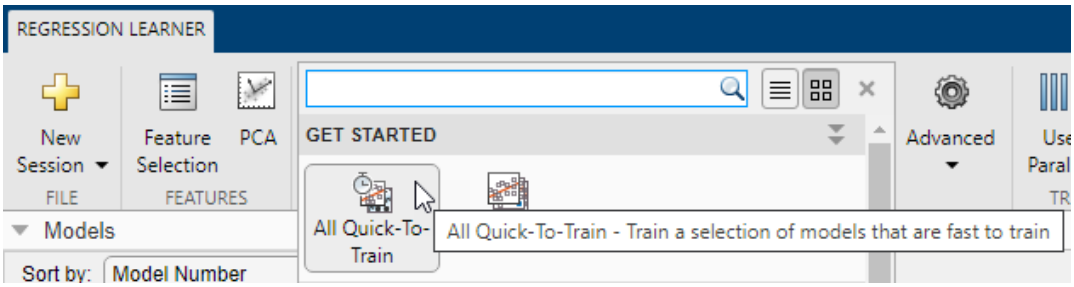
Automated Regression Model Training

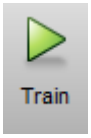
You can use Regression Learner to automatically train a selection of different regression models on your data.

- Get started by automatically training multiple models simultaneously. You can quickly try a selection of models, and then explore promising models interactively.
- If you already know what model type you want, then you can train individual models instead. See “Manual Regression Model Training” on page 24-4.

1. On the **Apps** tab, in the **Machine Learning** group, click **Regression Learner**.
2. Click **New Session** and select data from the workspace or from file. Specify a response variable and variables to use as predictors. See “Select Data and Validation for Regression Problem” on page 24-8.

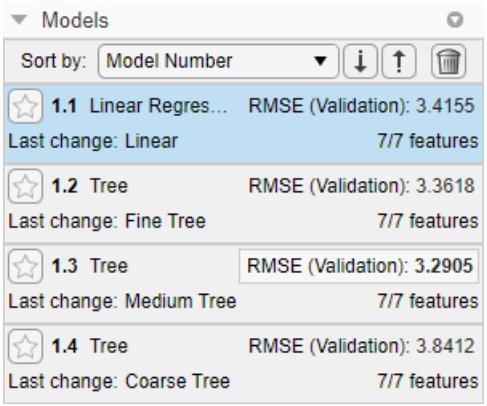
3 On the **Regression Learner** tab, in the **Model Type** section, click the arrow to expand the list of regression models. Select **All Quick-To-Train**  . This option trains all the model presets that are fast to fit.



4 Click **Train**. 

Note If you have Parallel Computing Toolbox, you can train the models in parallel. See “Parallel Regression Model Training” on page 24-5.

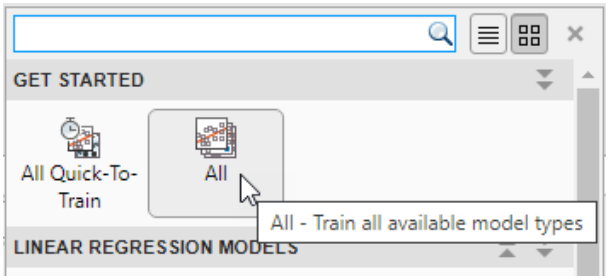
A selection of model types appears in the **Models** pane. When the models finish training, the best **RMSE (Validation)** score is highlighted in a box.



5 Click models in the **Models** pane to explore results in the plots.

For the next steps, see “Manual Regression Model Training” on page 24-4 or “Compare and Improve Regression Models” on page 24-5.

6 To try all the nonoptimizable model presets available, click **All**  , and then click **Train**.

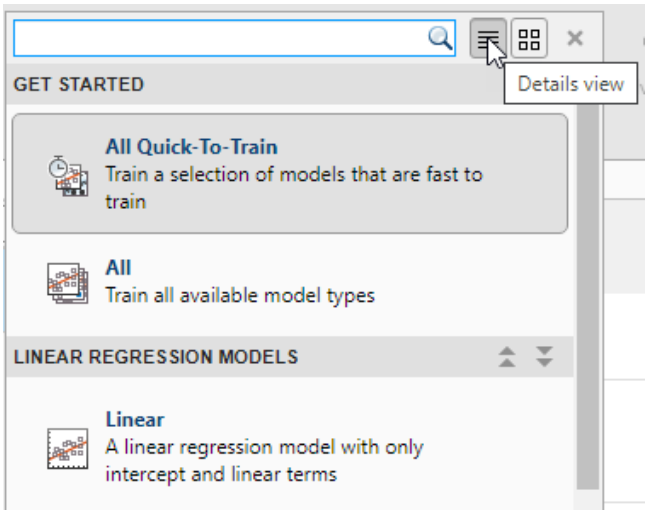


Manual Regression Model Training

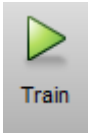
To explore individual model types, you can train models one at a time or train a group of models of the same type.

1 Choose a model type. On the **Regression Learner** tab, in the **Model Type** section, click a model type. To see all available model options, click the arrow in the **Model Type** section to expand the list of regression models. The nonoptimizable model options in the gallery are preset starting points with different settings, suitable for a range of different regression problems.

To read descriptions of the models, switch to the details view or hover the mouse over a button to display its tooltip.



For more information on each option, see “Choose Regression Model Options” on page 24-12.

2 After selecting a model, click **Train**. 

Repeat to explore different models.

Tip Select regression trees first. If your trained models do not predict the response accurately enough, then try other models with higher flexibility. To avoid overfitting, look for a less flexible model that provides sufficient accuracy.

3 If you want to try all nonoptimizable models of the same or different types, then select one of the **All** options in the gallery.

   Alternatively, if you want to automatically tune hyperparameters of a specific model type, select the corresponding **Optimizable** model and perform hyperparameter optimization. For more information, see “Hyperparameter Optimization in Regression Learner App” on page 24-30.

For next steps, see “Compare and Improve Regression Models” on page 24-5.

Parallel Regression Model Training

You can train models in parallel using Regression Learner if you have Parallel Computing Toolbox. Parallel training allows you to train multiple models simultaneously and continue working.

To control parallel training, toggle the **Use Parallel** button on the app toolstrip. The **Use Parallel** button is only available if you have Parallel Computing Toolbox.

1 The first time you click **Train** after clicking the **Use Parallel** button, you see a dialog box while the app opens a parallel pool of workers. After the pool opens, you can train multiple models at once.

2 When models are training in parallel, you see progress indicators on each training and queued model in the **Models** pane. If you want, you can cancel individual models. During training, you can examine results and plots from models, and initiate training of more models.

If you have Parallel Computing Toolbox, then parallel training is available in Regression Learner, and you do not need to set the `UseParallel` option of the `statset` function.

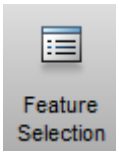
Note You cannot perform hyperparameter optimization in parallel. The app disables the **Use Parallel** button when you select an optimizable model. If you then select a nonoptimizable model, the button is off by default.

Compare and Improve Regression Models

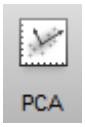
1 Click models in the **Models** pane to explore the results in the plots. Compare model performance by inspecting results in the plots. Examine the **RMSE (Validation)** score reported in the **Models** pane for each model.

   Additionally, you can compare the models by using the **Sort by** options in the **Models** pane. Delete any unwanted model by selecting the model and clicking the **Delete selected model** button in the upper right of the pane, or right-clicking the model and selecting **Delete model**.

   See “Assess Model Performance in Regression Learner” on page 24-42.

2. Select the best model in the **Models** pane and then try including and excluding different features in the model. Click **Feature Selection**. 

   Try the response plot to help you identify features to remove. See if you can improve the model by removing features with low predictive power. Specify predictors to include in the model, and train new models using the new options. Compare results among the models in the **Models** pane.

   You also can try transforming features with PCA to reduce dimensionality. 

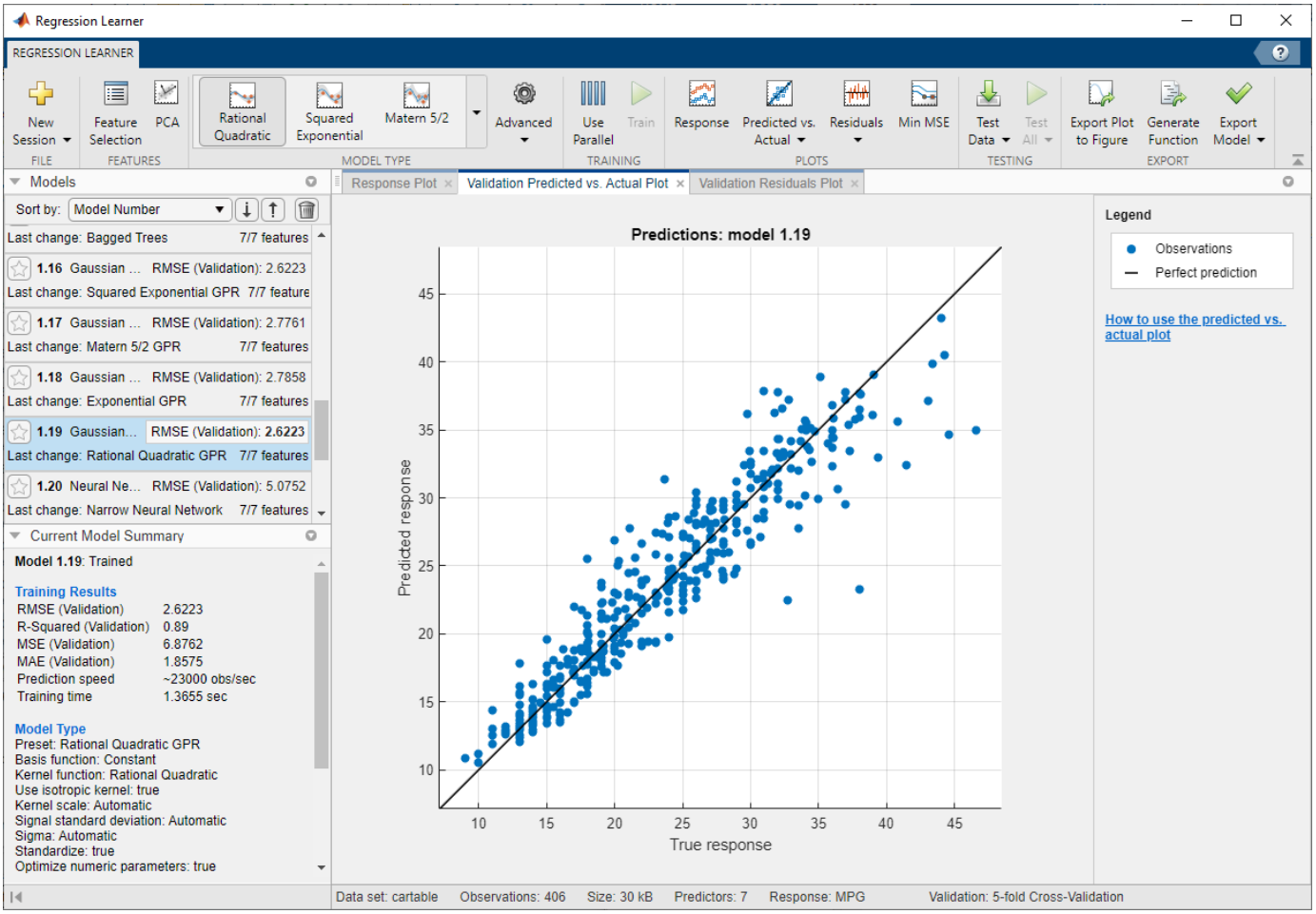
   See “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26.

3. Improve the model further by changing model parameter settings in the Advanced dialog box. Then, train using the new options. To learn how to control model flexibility, see “Choose Regression Model Options” on page 24-12. For information on how to tune model parameter settings automatically, see “Hyperparameter Optimization in Regression Learner App” on page 24-30.

   If feature selection, PCA, or new parameter settings improve your model, try training **All**  model types with the new settings. See if another model type does better with the new settings.

Tip To avoid overfitting, look for a less flexible model that provides sufficient accuracy. For example, look for simple models, such as regression trees that are fast and easy to interpret. If your models are not accurate enough, then try other models with higher flexibility, such as ensembles. To learn about the model flexibility, see “Choose Regression Model Options” on page 24-12.

This figure shows the app with a **Models** pane containing various regression model types.



Tip For a step-by-step example comparing different regression models, see “Train Regression Trees Using Regression Learner App” on page 24-60.

Next, you can generate code to train the model with different data or export trained models to the workspace to make predictions using new data. See “Export Regression Model to Predict New Data” on page 24-54.

See Also

Related Examples

- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54
- “Train Regression Trees Using Regression Learner App” on page 24-60

Select Data and Validation for Regression Problem

In this section...

Select Data from Workspace

Tip In Regression Learner, tables are the easiest way to work with your data, because they can contain numeric and label data. Use the Import Tool to bring your data into the MATLAB workspace as a table, or use the table functions to create a `table` from workspace variables. See “Tables”.

1. Load your data into the MATLAB workspace.

   Predictor variables can be numeric, categorical, string, or logical vectors, cell arrays of character vectors, or character arrays. The response variable must be a floating-point vector (single or double precision).

   Combine the predictor data into one variable, either a table or a matrix. You can additionally combine your predictor data and response variable, or you can keep them separate.

   For example data sets, see “Example Data for Regression” on page 24-9.
2. On the **Apps** tab, click **Regression Learner** to open the app.
3. On the **Regression Learner** tab, in the **File** section, click **New Session > From Workspace**.
4. In the New Session from Workspace dialog box, under **Data Set Variable**, select a table or matrix from the workspace variables.

   If you select a matrix, choose whether to use rows or columns for observations by clicking the option buttons.
5. Under **Response**, observe the default response variable. The app tries to select a suitable response variable from the data set variable and treats all other variables as predictors.

   If you want to use a different response variable, you can:

   - Use the list to select another variable from the data set variable.
   - Select a separate workspace variable by clicking the **From workspace** option button and then selecting a variable from the list.
6. Under **Predictors**, add or remove predictors using the check boxes. Add or remove all predictors by clicking **Add All** or **Remove All**. You can also add or remove multiple predictors by selecting them in the table, and then clicking **Add N** or **Remove N**, where **N** is the number of selected predictors. The **Add All** and **Remove All** buttons change to **Add N** and **Remove N** when you select multiple predictors.
7. Click **Start Session** to accept the default validation scheme and continue. The default validation option is 5-fold cross-validation, which protects against overfitting.

Tip If you have a large data set, you might want to switch to holdout validation. To learn more, see “Choose Validation Scheme” on page 24-10.

Note If you prefer loading data into the app directly from the command line, you can specify the predictor data, response variable, and validation type to use in Regression Learner in the command line call to `regressionLearner`. For more information, see **Regression Learner**.

For next steps, see “Train Regression Models in Regression Learner App” on page 24-2.

Import Data from File

1. On the **Regression Learner** tab, in the **File** section, select **New Session** > **From File**.
2. Select a file type in the list, such as spreadsheets, text files, or comma-separated values (`.csv`) files, or select **All Files** to browse for other file types such as `.dat`.

Example Data for Regression

To get started using Regression Learner, try these example data sets.

Name	Size	Description
Cars	Number of predictors: 7 Number of observations: 406 Response: `MPG` (miles per gallon)	Data on different car models, 1970–1982. Predict the fuel economy (in miles per gallon), or one of the other characteristics. For a step-by-step example, see “Train Regression Trees Using Regression Learner App” on page 24-60.

	Create a table from variables in the `carbig.mat` file:
	
	`load carbig`
	`cartable = table(Acceleration, Cylinders, Displacement,...`
	`Horsepower, Model_Year, Weight, Origin, MPG);`	

Abalone	Number of predictors: 8 Number of observations: 4177 Response: `Rings`	Measurements of abalone (a group of sea snails). Predict the age of abalones, which is closely related to the number of rings in their shells.

	Download the data from the UCI Machine Learning Repository and save it in your current folder. Read the data into a table and specify the variable names.
	
	`url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.`
	`websave('abalone.csv',url);`
	`varnames = {'Sex'; 'Length'; 'Diameter'; 'Height'; 'Whole_weight';...`
	`'Shucked_weight'; 'Viscera_weight'; 'Shell_weight'; 'Rings'};`
	`abalonetable = readtable('abalone.csv');`
	`abalonetable.Properties.VariableNames = varnames;`	

Hospital	Number of predictors: 5 Number of observations: 100 Response: `BloodPressure_2`	Simulated hospital data. Predict the blood pressure of patients.

Name	Size	Description

	Create a table from the `hospital` variable in the `hospital.mat` file:
	
	`load hospital.mat`
	`hospitaltable = dataset2table(hospital(:,2:end-1));`	

Choose Validation Scheme

Choose a validation method to examine the predictive accuracy of the fitted models. Validation estimates model performance on new data, and helps you choose the best model. Validation protects against overfitting. A model that is too flexible and suffers from overfitting has a worse validation accuracy. Choose a validation scheme before training any models so that you can compare all the models in your session using the same validation scheme.

Tip Try the default validation scheme and click **Start Session** to continue. The default option is 5-fold cross-validation, which protects against overfitting.

If you have a large data set and training the models takes too long using cross-validation, reimport your data and try the faster holdout validation instead.

- **Cross-Validation**: Select the number of folds (or divisions) to partition the data set.

  If you choose *k* folds, then the app:

  1. Partitions the data into *k* disjoint sets or folds
  2. For each validation fold:
     a. Trains a model using the training-fold observations (observations not in the validation fold)
     b. Assesses model performance using validation-fold data
  3. Calculates the average validation error over all folds

  This method gives a good estimate of the predictive accuracy of the final model trained using the full data set. The method requires multiple fits, but makes efficient use of all the data, so it works well for small data sets.
- **Holdout Validation**: Select a percentage of the data to use as a validation set. The app trains a model on the training set and assesses its performance with the validation set. The model used for validation is based on only a portion of the data, so holdout validation is appropriate only for large data sets. The final model is trained using the full data set.
- **Resubstitution Validation**: No protection against overfitting. The app uses all the data for training and computes the error rate on the same data. Without any separate validation data, you get an unrealistic estimate of the model's performance on new data. That is, the training sample accuracy is likely to be unrealistically high, and the predictive accuracy is likely to be lower.

  To help you avoid overfitting to the training data, choose another validation scheme instead.

Note The validation scheme only affects the way that Regression Learner computes validation metrics. The final model is always trained using the full data set.

All the models you train after selecting data use the same validation scheme that you select in this dialog box. You can compare all the models in your session using the same validation scheme.

To change the validation selection and train new models, you can select data again, but you lose any trained models. The app warns you that importing data starts a new session. Save any trained models you want to keep to the workspace, and then import the data.

For next steps training models, see “Train Regression Models in Regression Learner App” on page 24-2.

See Also

Related Examples

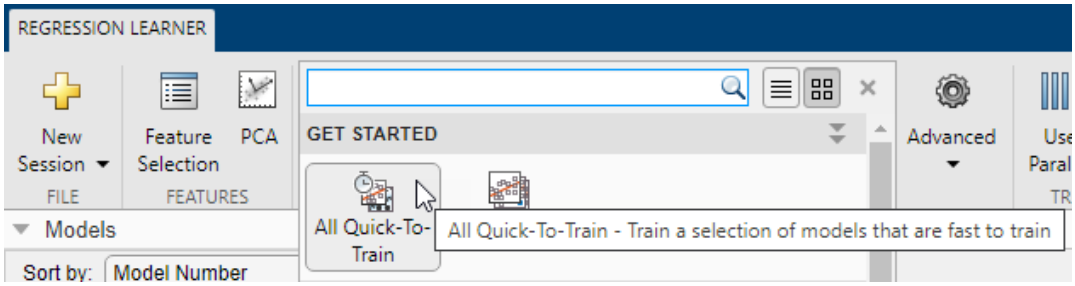
- “Train Regression Models in Regression Learner App” on page 24-2
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54
- “Train Regression Trees Using Regression Learner App” on page 24-60

Choose Regression Model Options

In this section...

Choose Regression Model Type

You can use the Regression Learner app to automatically train a selection of different models on your data. Use automated training to quickly try a selection of model types, and then explore promising models interactively. To get started, try these options first:



Get Started Regression Model Buttons	Description
All Quick-To-Train	Try the **All Quick-To-Train** button first. The app trains all model types that are typically quick to train.
All	Use the **All** button to train all available nonoptimizable model types. Trains every type regardless of any prior trained models. Can be time-consuming.

To learn more about automated model training, see “Automated Regression Model Training” on page 24-2.

If you want to explore models one at a time, or if you already know what model type you want, you can select individual models or train a group of the same type. To see all available regression model options, on the **Regression Learner** tab, click the arrow in the **Model Type** section to expand the list of regression models. The nonoptimizable model options in the gallery are preset starting points with different settings, suitable for a range of different regression problems. To use optimizable model options and tune model hyperparameters automatically, see “Hyperparameter Optimization in Regression Learner App” on page 24-30.

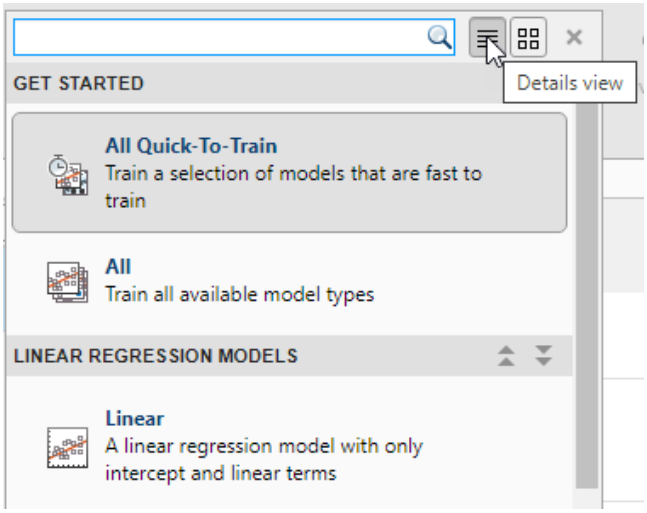
For help choosing the best model type for your problem, see the tables showing typical characteristics of different regression model types. Decide on the tradeoff you want in speed, flexibility, and interpretability. The best model type depends on your data.

Tip To avoid overfitting, look for a less flexible model that provides sufficient accuracy. For example, look for simple models such as regression trees that are fast and easy to interpret. If the models are not accurate enough predicting the response, choose other models with higher flexibility, such as ensembles. To control flexibility, see the details for each model type.

Characteristics of Regression Model Types

Regression Model Type	Interpretability
“Linear Regression Models” on page 24-14 	Easy
“Regression Trees” on page 24-16 	Easy
“Support Vector Machines” on page 24-18 	Easy for linear SVMs. Hard for other kernels.
“Gaussian Process Regression Models” on page 24-20 	Hard
“Ensembles of Trees” on page 24-22 	Hard
“Neural Networks” on page 24-24 	Hard

To read a description of each model in Regression Learner, switch to the details view in the list of all model presets.



Tip The nonoptimizable models in the **Model Type** gallery are preset starting points with different settings. After you choose a model type, such as regression trees, try training all the nonoptimizable presets to see which one produces the best model with your data.

For workflow instructions, see “Train Regression Models in Regression Learner App” on page 24-2.

Categorical Predictor Support

In Regression Learner, all model types support categorical predictors.

Tip If you have categorical predictors with many unique values, training linear models with interaction or quadratic terms and stepwise linear models can use a lot of memory. If the model fails to train, try removing these categorical predictors.

Linear Regression Models

Linear regression models have predictors that are linear in the model parameters, are easy to interpret, and are fast for making predictions. These characteristics make linear regression models popular models to try first. However, the highly constrained form of these models means that they often have low predictive accuracy. After fitting a linear regression model, try creating more flexible models, such as regression trees, and compare the results.

Tip In the **Model Type** gallery, click **All Linear** to try each of the linear regression options and see which settings produce the best model with your data. Select the best model in the **Models** pane and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Model Flexibility
Linear 	Easy	Very low
Interactions Linear 	Easy	Medium
Robust Linear 	Easy	Very low. Less sensitive to outliers, but can be slow to train.
Stepwise Linear 	Easy	Medium

Tip For a workflow example, see “Train Regression Trees Using Regression Learner App” on page 24-60.

Advanced Linear Regression Options

Regression Learner uses the `fitlm` function to train Linear, Interactions Linear, and Robust Linear models. The app uses the `stepwiselm` function to train Stepwise Linear models.

For Linear, Interactions Linear, and Robust Linear models you can set these options:

- **Terms**

  Specify which terms to use in the linear model. You can choose from:

  - `Linear`. A constant term and linear terms in the predictors
  - `Interactions`. A constant term, linear terms, and interaction terms between the predictors
  - `Pure Quadratic`. A constant term, linear terms, and terms that are purely quadratic in each of the predictors
  - `Quadratic`. A constant term, linear terms, and quadratic terms (including interactions)
- **Robust option**

  Specify whether to use a robust objective function and make your model less sensitive to outliers. With this option, the fitting method automatically assigns lower weights to data points that are more likely to be outliers.

Stepwise linear regression starts with an initial model and systematically adds and removes terms to the model based on the explanatory power of these incrementally larger and smaller models. For Stepwise Linear models, you can set these options:

- **Initial terms**

  Specify the terms that are included in the initial model of the stepwise procedure. You can choose from `Constant`, `Linear`, `Interactions`, `Pure Quadratic`, and `Quadratic`.
- **Upper bound on terms**

Specify the highest order of the terms that the stepwise procedure can add to the model. You can choose from `Linear`, `Interactions`, `Pure Quadratic`, and `Quadratic`.

- **Maximum number of steps**

  Specify the maximum number of different linear models that can be tried in the stepwise procedure. To speed up training, try reducing the maximum number of steps. Selecting a small maximum number of steps decreases your chances of finding a good model.

Tip If you have categorical predictors with many unique values, training linear models with interaction or quadratic terms and stepwise linear models can use a lot of memory. If the model fails to train, try removing these categorical predictors.

Regression Trees

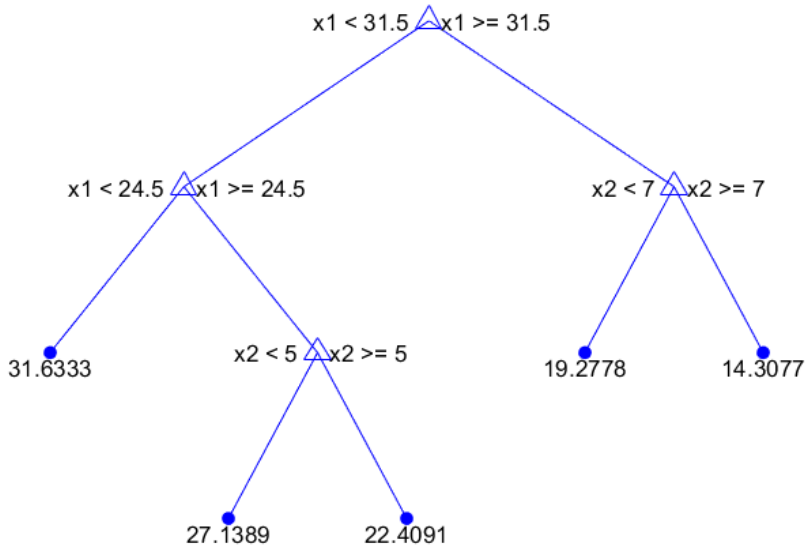
Regression trees are easy to interpret, fast for fitting and prediction, and low on memory usage. Try to grow smaller trees with fewer larger leaves to prevent overfitting. Control the leaf size with the **Minimum leaf size** setting.

Tip In the **Model Type** gallery, click **All Trees**  to try each of the nonoptimizable regression tree options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Model Flexibility
Fine Tree 	Easy	High Many small leaves for a highly flexible response function (Minimum leaf size is 4.)
Medium Tree 	Easy	Medium Medium-sized leaves for a less flexible response function (Minimum leaf size is 12.)
Coarse Tree 	Easy	Low Few large leaves for a coarse response function (Minimum leaf size is 36.)

To predict a response of a regression tree, follow the tree from the root (beginning) node down to a leaf node. The leaf node contains the value of the response.

Statistics and Machine Learning Toolbox trees are binary. Each step in a prediction involves checking the value of one predictor variable. For example, here is a simple regression tree



This tree predicts the response based on two predictors, `x1` and `x2`. To make a prediction, start at the top node. At each node, check the values of the predictors to decide which branch to follow. When the branches reach a leaf node, the response is set to the value corresponding to that node.

You can visualize your regression tree model by exporting the model from the app, and then entering:

```
view(trainedModel.RegressionTree,'Mode','graph')
```

Tip For a workflow example, see “Train Regression Trees Using Regression Learner App” on page 24-60.

Advanced Regression Tree Options

The Regression Learner app uses the `fitrtree` function to train regression trees. You can set these options:

- **Minimum leaf size**

  Specify the minimum number of training samples used to calculate the response of each leaf node. When you grow a regression tree, consider its simplicity and predictive power. To change the minimum leaf size, click the buttons or enter a positive integer value in the **Minimum leaf size** box.

  - A fine tree with many small leaves is usually highly accurate on the training data. However, the tree might not show comparable accuracy on an independent test set. A very leafy tree tends to overfit, and its validation accuracy is often far lower than its training (or resubstitution) accuracy.
  - In contrast, a coarse tree with fewer large leaves does not attain high training accuracy. But a coarse tree can be more robust in that its training accuracy can be near that of a representative test set.

Tip Decrease the **Minimum leaf size** to create a more flexible model.

- **Surrogate decision splits** — For missing data only.

  Specify surrogate use for decision splits. If you have data with missing values, use surrogate splits to improve the accuracy of predictions.

  When you set **Surrogate decision splits** to `On`, the regression tree finds at most 10 surrogate splits at each branch node. To change the number of surrogate splits, click the buttons or enter a positive integer value in the **Maximum surrogates per node** box.

  When you set **Surrogate decision splits** to `Find All`, the regression tree finds all surrogate splits at each branch node. The `Find All` setting can use considerable time and memory.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Regression Learner App” on page 24-30.

Support Vector Machines

You can train regression support vector machines (SVMs) in Regression Learner. Linear SVMs are easy to interpret, but can have low predictive accuracy. Nonlinear SVMs are more difficult to interpret, but can be more accurate.

Tip In the **Model Type** gallery, click **All SVMs**  to try each of the nonoptimizable SVM options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Model Flexibility
Linear SVM 	Easy	Low
Quadratic SVM 	Hard	Medium
Cubic SVM 	Hard	Medium
Fine Gaussian SVM 	Hard	High Allows rapid variations in the response function. Kernel scale is set to `sqrt(P)/4`, where `P` is the number of predictors.
Medium Gaussian SVM 	Hard	Medium Gives a less flexible response function. Kernel scale is set to `sqrt(P)`.

Regression Model Type	Interpretability	Model Flexibility
Coarse Gaussian SVM 	Hard	Low Gives a rigid response function. Kernel scale is set to `sqrt(P)*4`.

Statistics and Machine Learning Toolbox implements linear epsilon-insensitive SVM regression. This SVM ignores prediction errors that are less than some fixed number ε. The support vectors are the data points that have errors larger than ε. The function the SVM uses to predict new values depends only on the support vectors. To learn more about SVM regression, see Understanding Support Vector Machine Regression on page 25-2.

Tip For a workflow example, see "Train Regression Trees Using Regression Learner App" on page 24-60.

Advanced SVM Options

Regression Learner uses the `fitrsvm` function to train SVM regression models.

You can set these options in the app:

- **Kernel function**

  The kernel function determines the nonlinear transformation applied to the data before the SVM is trained. You can choose from:

  - `Gaussian` or Radial Basis Function (RBF) kernel
  - `Linear` kernel, easiest to interpret
  - `Quadratic` kernel
  - `Cubic` kernel
- **Box constraint mode**

  The box constraint controls the penalty imposed on observations with large residuals. A larger box constraint gives a more flexible model. A smaller value gives a more rigid model, less sensitive to overfitting.

  When **Box constraint mode** is set to `Auto`, the app uses a heuristic procedure to select the box constraint.

  Try to fine-tune your model by specifying the box constraint manually. Set **Box constraint mode** to `Manual` and specify a value. Change the value by clicking the buttons or entering a positive scalar value in the **Manual box constraint** box. The app automatically preselects a reasonable value for you. Try to increase or decrease this value slightly and see if this improves your model.

  Tip Increase the box constraint value to create a more flexible model.
- **Epsilon mode**

  Prediction errors that are smaller than the epsilon (ε) value are ignored and treated as equal to zero. A smaller epsilon value gives a more flexible model.

When **Epsilon mode** is set to `Auto`, the app uses a heuristic procedure to select the kernel scale.

Try to fine-tune your model by specifying the epsilon value manually. Set **Epsilon mode** to `Manual` and specify a value. Change the value by clicking the buttons or entering a positive scalar value in the **Manual epsilon** box. The app automatically preselects a reasonable value for you. Try to increase or decrease this value slightly and see if this improves your model.

Tip Decrease the epsilon value to create a more flexible model.

- **Kernel scale mode**

  The kernel scale controls the scale of the predictors on which the kernel varies significantly. A smaller kernel scale gives a more flexible model.

  When **Kernel scale mode** is set to `Auto`, the app uses a heuristic procedure to select the kernel scale.

  Try to fine-tune your model by specifying the kernel scale manually. Set **Kernel scale mode** to `Manual` and specify a value. Change the value by clicking the buttons or entering a positive scalar value in the **Manual kernel scale** box. The app automatically preselects a reasonable value for you. Try to increase or decrease this value slightly and see if this improves your model.

  Tip Decrease the kernel scale value to create a more flexible model.

- **Standardize**

  Standardizing the predictors transforms them so that they have mean 0 and standard deviation 1. Standardizing removes the dependence on arbitrary scales in the predictors and generally improves performance.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Regression Learner App” on page 24-30.

Gaussian Process Regression Models

You can train Gaussian process regression (GPR) models in Regression Learner. GPR models are often highly accurate, but can be difficult to interpret.

Tip In the **Model Type** gallery, click **All GPR Models**  to try each of the nonoptimizable GPR model options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Model Flexibility
Rational Quadratic 	Hard	Automatic

Regression Model Type	Interpretability	Model Flexibility
Squared Exponential 	Hard	Automatic
Matern 5/2 	Hard	Automatic
Exponential 	Hard	Automatic

In Gaussian process regression, the response is modeled using a probability distribution over a space of functions. The flexibility of the presets in the **Model Type** gallery is automatically chosen to give a small training error and, simultaneously, protection against overfitting. To learn more about Gaussian process regression, see Gaussian Process Regression Models on page 6-2.

Tip For a workflow example, see "Train Regression Trees Using Regression Learner App" on page 24-60.

Advanced Gaussian Process Regression Options

Regression Learner uses the `fitrgp` function to train GPR models.

You can set these options in the app:

- **Basis function**

  The basis function specifies the form of the prior mean function of the Gaussian process regression model. You can choose from `Zero`, `Constant`, and `Linear`. Try to choose a different basis function and see if this improves your model.

- **Kernel function**

  The kernel function determines the correlation in the response as a function of the distance between the predictor values. You can choose from `Rational Quadratic`, `Squared Exponential`, `Matern 5/2`, `Matern 3/2`, and `Exponential`.

  To learn more about kernel functions, see Kernel (Covariance) Function Options on page 6-6.

- **Use isotropic kernel**

  If you use an isotropic kernel, the correlation length scales are the same for all the predictors. With a nonisotropic kernel, each predictor variable has its own separate correlation length scale.

  Using a nonisotropic kernel can improve the accuracy of your model, but can make the model slow to fit.

  To learn more about nonisotropic kernels, see Kernel (Covariance) Function Options. on page 6-6

- **Kernel mode**

  You can manually specify *initial* values of the kernel parameters **Kernel scale** and **Signal standard deviation**. The signal standard deviation is the prior standard deviation of the response

values. By default the app locally optimizes the kernel parameters starting from the initial values. To use fixed kernel parameters, clear the **Optimize numeric parameters** check box in the advanced options.

When **Kernel scale mode** is set to `Auto`, the app uses a heuristic procedure to select the initial kernel parameters.

If you set **Kernel scale mode** to `Manual`, you can specify the initial values. Click the buttons or enter a positive scalar value in the **Kernel scale** box and the **Signal standard deviation** box.

If you clear the **Use isotropic kernel** check box, you cannot set initial kernel parameters manually.

- **Sigma mode**

  You can specify manually the *initial* value of the observation noise standard deviation **Sigma**. By default the app optimizes the observation noise standard deviation, starting from the initial value. To use fixed kernel parameters, clear the **Optimize numeric parameters** check box in the advanced options.

  When **Sigma mode** is set to `Auto`, the app uses a heuristic procedure to select the initial observation noise standard deviation.

  If you set **Sigma mode** to `Manual`, you can specify the initial values. Click the buttons or enter a positive scalar value in the **Sigma** box.

- **Standardize**

  Standardizing the predictors transforms them so that they have mean 0 and standard deviation 1. Standardizing removes the dependence on arbitrary scales in the predictors and generally improves performance.

- **Optimize numeric parameters**

  With this option, the app automatically optimizes numeric parameters of the GPR model. The optimized parameters are the coefficients of the **Basis function**, the kernel parameters **Kernel scale** and **Signal standard deviation**, and the observation noise standard deviation **Sigma**.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Regression Learner App” on page 24-30.

Ensembles of Trees

You can train ensembles of regression trees in Regression Learner. Ensemble models combine results from many weak learners into one high-quality ensemble model.

Tip In the **Model Type** gallery, click **All Ensembles**  to try each of the nonoptimizable ensemble options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Ensemble Method	Model Flexibility
Boosted Trees 	Hard	Least-squares boosting (`LSBoost`) with regression tree learners.	Medium to high
Bagged Trees 	Hard	Bootstrap aggregating or bagging, with regression tree learners.	High

Tip For a workflow example, see “Train Regression Trees Using Regression Learner App” on page 24-60.

Advanced Ensemble Options

Regression Learner uses the `fitrensemble` function to train ensemble models. You can set these options:

- **Minimum leaf size**

  Specify the minimum number of training samples used to calculate the response of each leaf node. When you grow a regression tree, consider its simplicity and predictive power. To change the minimum leaf size, click the buttons or enter a positive integer value in the **Minimum leaf size** box.

  - A fine tree with many small leaves is usually highly accurate on the training data. However, the tree might not show comparable accuracy on an independent test set. A very leafy tree tends to overfit, and its validation accuracy is often far lower than its training (or resubstitution) accuracy.
  - In contrast, a coarse tree with fewer large leaves does not attain high training accuracy. But a coarse tree can be more robust in that its training accuracy can be near that of a representative test set.

  Tip Decrease the **Minimum leaf size** to create a more flexible model.

- **Number of learners**

  Try changing the number of learners to see if you can improve the model. Many learners can produce high accuracy, but can be time consuming to fit.

  Tip Increase the **Number of learners** to create a more flexible model.

- **Learning rate**

  For boosted trees, specify the learning rate for shrinkage. If you set the learning rate to less than 1, the ensemble requires more learning iterations but often achieves better accuracy. 0.1 is a popular initial choice.

Alternatively, you can let the app choose some of these model options automatically by using hyperparameter optimization. See “Hyperparameter Optimization in Regression Learner App” on page 24-30.

Neural Networks

Neural network models typically have good predictive accuracy; however, they are not easy to interpret.

Model flexibility increases with the size and number of fully connected layers in the neural network.

Tip In the **Model Type** gallery, click **All Neural Networks**  to try each of the preset neural network options and see which settings produce the best model with your data. Select the best model in the **Models** pane, and try to improve that model by using feature selection and changing some advanced options.

Regression Model Type	Interpretability	Model Flexibility
Narrow Neural Network 	Hard	Medium — increases with **Size of each fully connected layer** setting
Medium Neural Network 	Hard	Medium — increases with **Size of each fully connected layer** setting
Wide Neural Network 	Hard	Medium — increases with **Size of each fully connected layer** setting
Bilayered Neural Network 	Hard	High — increases with **Size of each fully connected layer** setting
Trilayered Neural Network 	Hard	High — increases with **Size of each fully connected layer** setting

Each model is a feedforward, fully connected neural network for regression. The first fully connected layer of the neural network has a connection from the network input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer, excluding the last. The final fully connected layer produces the network's output, namely predicted response values. For more information, see “Neural Network Structure” on page 33-2239.

For an example, see “Train Regression Neural Networks Using Regression Learner App” on page 24-69.

Advanced Neural Network Options

Regression Learner uses the `fitrnet` function to train neural network models. You can set these options:

- **Number of fully connected layers** — Specify the number of fully connected layers in the neural network, excluding the final fully connected layer for regression. You can choose a maximum of three fully connected layers.
- **Size of each fully connected layer** — Specify the size of each fully connected layer, excluding the final fully connected layer. If you choose to create a neural network with multiple fully connected layers, consider specifying layers with decreasing sizes.
- **Activation** — Specify the activation function for all fully connected layers, excluding the final fully connected layer. Choose from the following activation functions: `ReLU`, `Tanh`, `Sigmoid`, and `None`.
- **Iteration limit** — Specify the maximum number of training iterations.
- **Regularization term strength** — Specify the ridge (L2) regularization penalty term.
- **Standardize data** — Specify whether to standardize the numeric predictors. If predictors have widely different scales, standardizing can improve the fit. Standardizing the data is highly recommended.

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54
- “Train Regression Trees Using Regression Learner App” on page 24-60

Feature Selection and Feature Transformation Using Regression Learner App

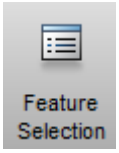
In this section...

Investigate Features in the Response Plot

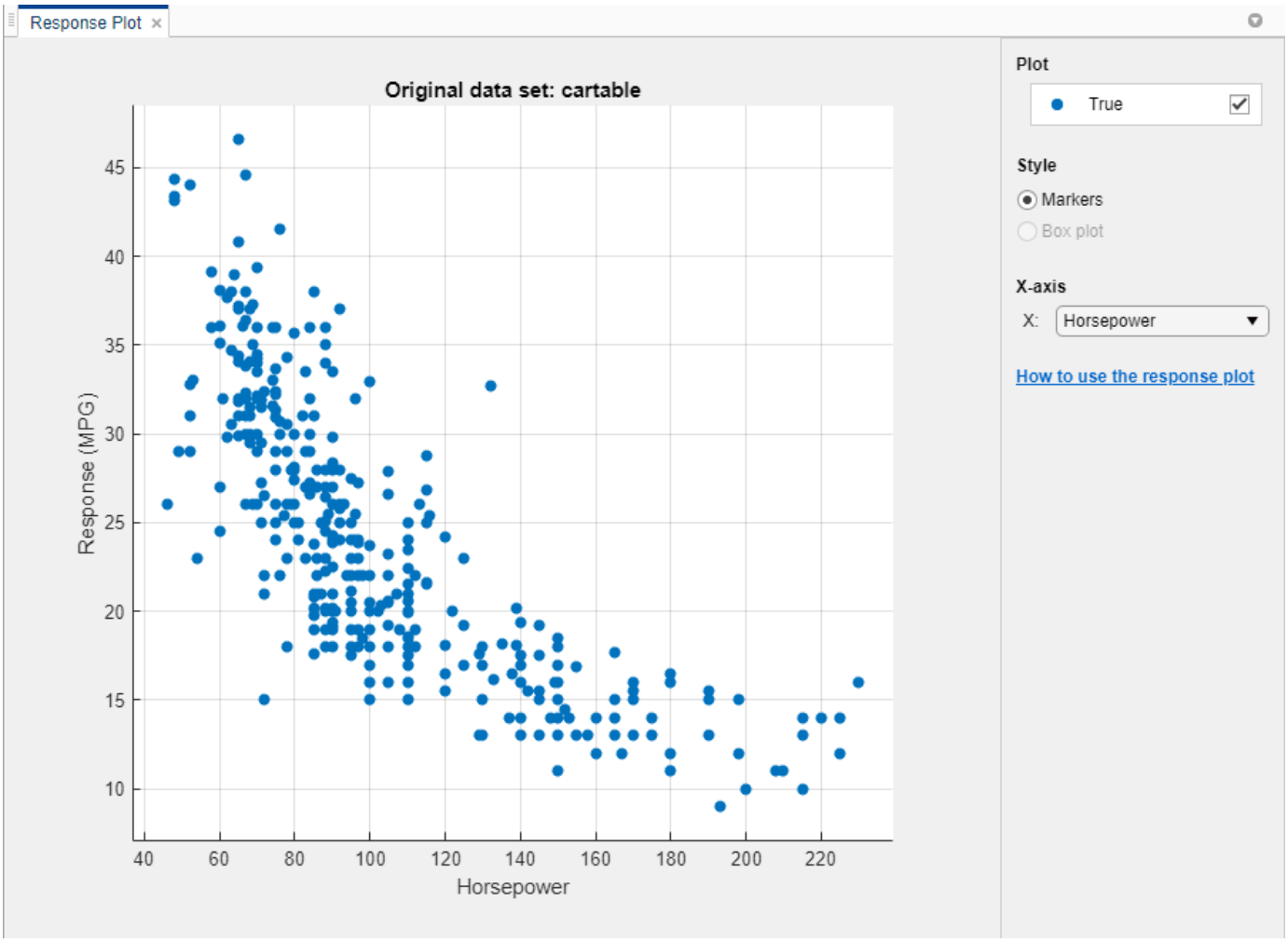
In Regression Learner, use the response plot to try to identify predictors that are useful for predicting the response. To visualize the relation between different predictors and the response, under **X-axis**, select different variables in the **X** list.

Before you train a regression model, the response plot shows the training data. If you have trained a regression model, then the response plot also shows the model predictions.

Observe which variables are associated most clearly with the response. When you plot the `carbig` data set, the predictor `Horsepower` shows a clear negative association with the response.

Look for features that do not seem to have any association with the response and use **Feature Selection** 

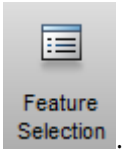
to remove those features from the set of used predictors.



You can export the response plots you create in the app to figures. See “Export Plots in Regression Learner App” on page 24-50.

Select Features to Include

In Regression Learner, you can specify different features (or predictors) to include in the model. See if you can improve models by removing features with low predictive power. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily with fewer predictors.

1. On the **Regression Learner** tab, in the **Features** section, click **Feature Selection** 

.
2. In the Feature Selection dialog box, clear the check boxes for the predictors you want to exclude.

Tip You can close the Feature Selection dialog box, or move it. The choices you make in the dialog box remain.

3. Click **Train** to train a new model using the new predictor options.
4. Observe the new model in the **Models** pane. The **Current Model Summary** pane displays how many predictors are excluded.
5. Check which predictors are included in a trained model. Click the model in the **Models** pane and look at the check boxes in the Feature Selection window.
6. Try to improve the model by including different features.

For an example using feature selection, see “Train Regression Trees Using Regression Learner App” on page 24-60.

Transform Features with PCA in Regression Learner

Use principal component analysis (PCA) to reduce the dimensionality of the predictor space. Reducing the dimensionality can create regression models in Regression Learner that help prevent overfitting. PCA linearly transforms predictors to remove redundant dimensions, and generates a new set of variables called principal components.

1. On the **Regression Learner** tab, in the **Features** section, select **PCA**.
2. In the Advanced PCA Options dialog box, select the **Enable PCA** check box.

   You can close the PCA dialog box, or move it. The choices you make in the dialog box remain.
3. Click **Train** again. The `pca` function transforms your selected features before training the model.

   By default, PCA keeps only the components that explain 95% of the variance. In the PCA dialog box, you can change the percentage of variance to explain by selecting the **Explained variance** value. A higher value risks overfitting, while a lower value risks removing useful dimensions.
4. Manually limit the number of PCA components. In the **Component reduction criterion** list, select `Specify number of components`. Select the **Number of numeric components** value. The number of components cannot be larger than the number of numeric predictors. PCA is not applied to categorical predictors.

You can check PCA Options for trained models in the **Current Model Summary** pane. For example:

```
PCA is keeping enough components to explain 95% variance.
After training, 2 components were kept.
Explained variance per component (in order): 92.5%, 5.3%, 1.7%, 0.5%
```

Check the explained variance percentages to decide whether to change the number of components.

To learn more about how Regression Learner applies PCA to your data, generate code for your trained regression model. For more information on PCA, see the `pca` function.

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Assess Model Performance in Regression Learner” on page 24-42

Hyperparameter Optimization in Regression Learner App

In this section...

After you choose a particular type of model to train, for example a decision tree or a support vector machine (SVM), you can tune your model by selecting different advanced options. For example, you can change the minimum leaf size of a decision tree or the box constraint of an SVM. Some of these options are internal parameters of the model, or hyperparameters, that can strongly affect its performance. Instead of manually selecting these options, you can use hyperparameter optimization within the Regression Learner app to automate the selection of hyperparameter values. For a given model type, the app tries different combinations of hyperparameter values by using an optimization scheme that seeks to minimize the model mean squared error (MSE), and returns a model with the optimized hyperparameters. You can use the resulting model as you would any other trained model.

Note Because hyperparameter optimization can lead to an overfitted model, the recommended approach is to create a separate test set before importing your data into the Regression Learner app. After you train your optimizable model, you can see how it performs on your test set. For an example, see “Train Regression Model Using Hyperparameter Optimization in Regression Learner App” on page 24-76.

To perform hyperparameter optimization in Regression Learner, follow these steps:

1. Choose a model type and decide which hyperparameters to optimize. See “Select Hyperparameters to Optimize” on page 24-30.

   Note Hyperparameter optimization is not supported for linear regression models and neural networks.

2. (Optional) Specify how the optimization is performed. For more information, see “Optimization Options” on page 24-36.
3. Train your model. Use the “Minimum MSE Plot” on page 24-37 to track the optimization results.
4. Inspect your trained model. See “Optimization Results” on page 24-39.

Select Hyperparameters to Optimize

In the Regression Learner app, in the **Model Type** section of the **Regression Learner** tab, click the arrow to open the gallery. The gallery includes optimizable models that you can train using hyperparameter optimization.

After you select an optimizable model, you can choose which of its hyperparameters you want to optimize. In the **Model Type** section, select **Advanced > Advanced**. The app opens a dialog box in which you can select **Optimize** check boxes for the hyperparameters that you want to optimize. Under **Values**, specify the fixed values for the hyperparameters that you do not want to optimize or that are not optimizable.

This table describes the hyperparameters that you can optimize for each type of model and the search range of each hyperparameter. It also includes the additional hyperparameters for which you can specify fixed values.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
Optimizable Tree 	• **Minimum leaf size** – The software searches among integers log-scaled in the range `[1,max(2,floor(n/2))]`, where `n` is the number of observations.	• **Surrogate decision splits** • **Maximum surrogates per node**	For more information, see “Advanced Regression Tree Options” on page 24-17.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
Optimizable SVM 	• **Kernel function** – The software searches among `Gaussian`, `Linear`, `Quadratic`, and `Cubic`. • **Box constraint** – The software searches among positive values log-scaled in the range `[0.001,1000]`. • **Kernel scale** – The software searches among positive values log-scaled in the range `[0.001,1000]`. • **Epsilon** – The software searches among positive values log-scaled in the range `[0.001,100]*iqr(Y)/1.349`, where `Y` is the response variable. • **Standardize data** – The software searches between `true` and `false`.		• The **Box constraint** optimizable hyperparameter combines the **Box constraint mode** and **Manual box constraint** advanced options of the preset SVM models. • The **Kernel scale** optimizable hyperparameter combines the **Kernel scale mode** and **Manual kernel scale** advanced options of the preset SVM models. • You can optimize the **Kernel scale** optimizable hyperparameter only when the **Kernel function** value is `Gaussian`. Unless you specify a value for **Kernel scale** by clearing the **Optimize** check box, the app uses the **Manual** value of `1` by default when the **Kernel function** has a value other than `Gaussian`. • The **Epsilon** optimizable hyperparameter combines the **Epsilon mode** and **Manual epsilon** advanced options of the preset SVM models.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
			For more information, see “Advanced SVM Options” on page 24-19.

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes

Optimizable GPR 	• **Basis function** – The software searches among `Zero`, `Constant`, and `Linear`.
	• **Kernel function** – The software searches among:
	  • `Nonisotropic Rational Quadratic`
	  • `Isotropic Rational Quadratic`
	  • `Nonisotropic Squared Exponential`
	  • `Isotropic Squared Exponential`
	  • `Nonisotropic Matern 5/2`
	  • `Isotropic Matern 5/2`
	  • `Nonisotropic Matern 3/2`
	  • `Isotropic Matern 3/2`
	  • `Nonisotropic Exponential`
	  • `Isotropic Exponential`
	• **Kernel scale** – The software searches among real values in the range `[0.001,1]*XMaxRange`, where `XMaxRange = max(max(X) – min(X))` and `X` is the predictor data.
	• **Sigma** – The software searches among real values in
		• **Signal standard deviation**
		• **Optimize numeric parameters**
			• The **Kernel function** optimizable hyperparameter combines the **Kernel function** and **Use isotropic kernel** advanced options of the preset Gaussian process models.
			• The **Kernel scale** optimizable hyperparameter combines the **Kernel mode** and **Kernel scale** advanced options of the preset Gaussian process models.
			• The **Sigma** optimizable hyperparameter combines the **Sigma mode** and **Sigma** advanced options of the preset Gaussian process models.
			• When you optimize the **Kernel scale** of isotropic kernel functions, only the kernel scale is optimized, not the signal standard deviation. You can either specify a **Signal standard deviation** value or use its default value.
			
			You cannot optimize the **Kernel scale** of nonisotropic kernel functions.
			
			For more information, see “Advanced Gaussian

Model	Optimizable Hyperparameters	Additional Hyperparameters	Notes
	the range `[0.0001,max(0.001,10*std(Y))]`, where `Y` is the response variable. • **Standardize** – The software searches between `true` and `false`.		Process Regression Options” on page 24-21.
Optimizable Ensemble 	• **Ensemble method** – The software searches among `Bag` and `LSBoost`. • **Minimum leaf size** – The software searches among integers log-scaled in the range `[1,max(2,floor(n/2))]`, where `n` is the number of observations. • **Number of learners** – The software searches among integers log-scaled in the range `[10,500]`. • **Learning rate** – The software searches among real values log-scaled in the range `[0.001,1]`. • **Number of predictors to sample** – The software searches among integers in the range `[1,max(2,p)]`, where `p` is the number of predictor variables.		• The `Bag` value of the **Ensemble method** optimizable hyperparameter specifies a **Bagged Trees** model. Similarly, the `LSBoost` **Ensemble method** value specifies a **Boosted Trees** model. • The **Number of predictors to sample** optimizable hyperparameter is not available in the advanced options of the preset ensemble models. For more information, see “Advanced Ensemble Options” on page 24-23.

Optimization Options

By default, the Regression Learner app performs hyperparameter tuning by using Bayesian optimization. The goal of Bayesian optimization, and optimization in general, is to find a point that minimizes an objective function. In the context of hyperparameter tuning in the app, a point is a set of hyperparameter values, and the objective function is the loss function, or the mean squared error (MSE). For more information on the basics of Bayesian optimization, see “Bayesian Optimization Workflow” on page 10-25.

You can specify how the hyperparameter tuning is performed. For example, you can change the optimization method to grid search or limit the training time. On the **Regression Learner** tab, in the **Model Type** section, select **Advanced > Optimizer Options**. The app opens a dialog box in which you can select optimization options.

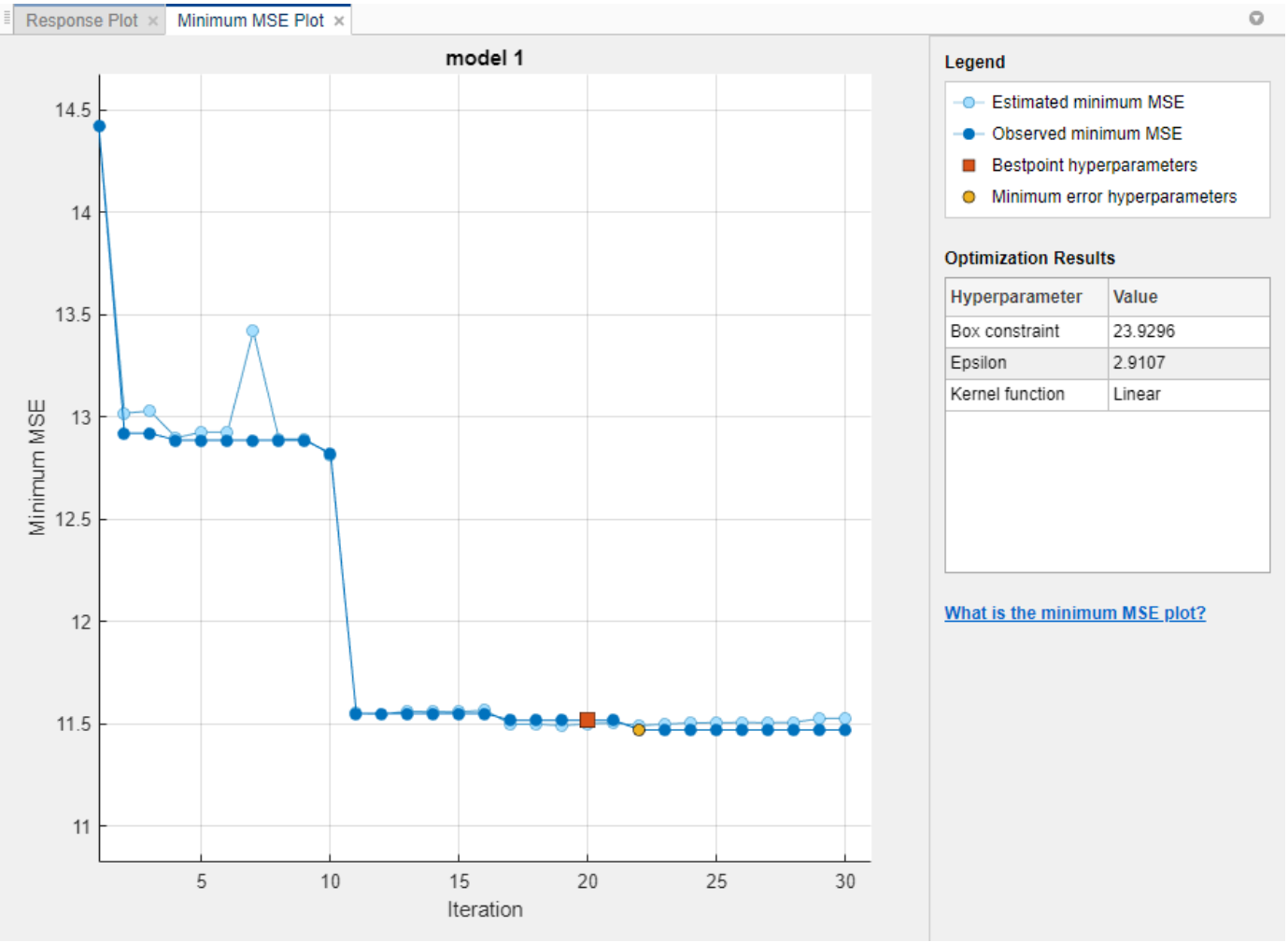
This table describes the available optimization options and their default values.

Option	Description
Optimizer	The optimizer values are: • `Bayesopt` (default) – Use Bayesian optimization. Internally, the app calls the `bayesopt` function. • `Grid search` – Use grid search with the number of values per dimension determined by the **Number of grid divisions** value. The app searches in a random order, using uniform sampling without replacement from the grid. • `Random search` – Search at random among points, where the number of points corresponds to the **Iterations** value.
Acquisition function	When the app performs Bayesian optimization for hyperparameter tuning, it uses the acquisition function to determine the next set of hyperparameter values to try. The acquisition function values are: • `Expected improvement per second plus` (default) • `Expected improvement` • `Expected improvement plus` • `Expected improvement per second` • `Lower confidence bound` • `Probability of improvement` For details on how these acquisition functions work in the context of Bayesian optimization, see “Acquisition Function Types” on page 10-3.

Option	Description
Iterations	Each iteration corresponds to a combination of hyperparameter values that the app tries. When you use Bayesian optimization or random search, specify a positive integer that sets the number of iterations. The default value is `30`. When you use grid search, the app ignores the **Iterations** value and evaluates the loss at every point in the entire grid. You can set a training time limit to stop the optimization process prematurely.
Training time limit	To set a training time limit, select this option and set the **Maximum training time in seconds** option. By default, the app does not have a training time limit.
Maximum training time in seconds	Set the training time limit in seconds as a positive real number. The default value is `300`. The run time can exceed the training time limit because this limit does not interrupt an iteration evaluation.
Number of grid divisions	When you use grid search, set a positive integer as the number of values the app tries for each numeric hyperparameter. The app ignores this value for categorical hyperparameters. The default value is `10`.

Minimum MSE Plot

After specifying which model hyperparameters to optimize and setting any additional optimization options (optional), train your optimizable model. On the **Regression Learner** tab, in the **Training** section, click **Train**. The app creates a **Minimum MSE Plot** that it updates as the optimization runs.



Note When you train an optimizable model, the app disables the **Use Parallel** button. After the training is complete, the app makes the button available again when you select a nonoptimizable model. The button is off by default.

The minimum mean squared error (MSE) plot displays the following information:

- **Estimated minimum MSE** – Each light blue point corresponds to an estimate of the minimum MSE computed by the optimization process when considering all the sets of hyperparameter values tried so far, including the current iteration.

  The estimate is based on an upper confidence interval of the current MSE objective model, as mentioned in the **Bestpoint hyperparameters** description.

  If you use grid search or random search to perform hyperparameter optimization, the app does not display these light blue points.
- **Observed minimum MSE** – Each dark blue point corresponds to the observed minimum MSE computed so far by the optimization process. For example, at the third iteration, the blue point corresponds to the minimum of the MSE observed in the first, second, and third iterations.

- **Bestpoint hyperparameters** – The red square indicates the iteration that corresponds to the optimized hyperparameters. You can find the values of the optimized hyperparameters listed in the upper right of the plot under **Optimization Results**.

  The optimized hyperparameters do not always provide the observed minimum MSE. When the app performs hyperparameter tuning by using Bayesian optimization (see “Optimization Options” on page 24-36 for a brief introduction), it chooses the set of hyperparameter values that minimizes an upper confidence interval of the MSE objective model, rather than the set that minimizes the MSE. For more information, see the `'Criterion','min-visited-upper-confidence-interval'` name-value pair argument of `bestPoint`.
- **Minimum error hyperparameters** – The yellow point indicates the iteration that corresponds to the hyperparameters that yield the observed minimum MSE.

  For more information, see the `'Criterion','min-observed'` name-value pair argument of `bestPoint`.

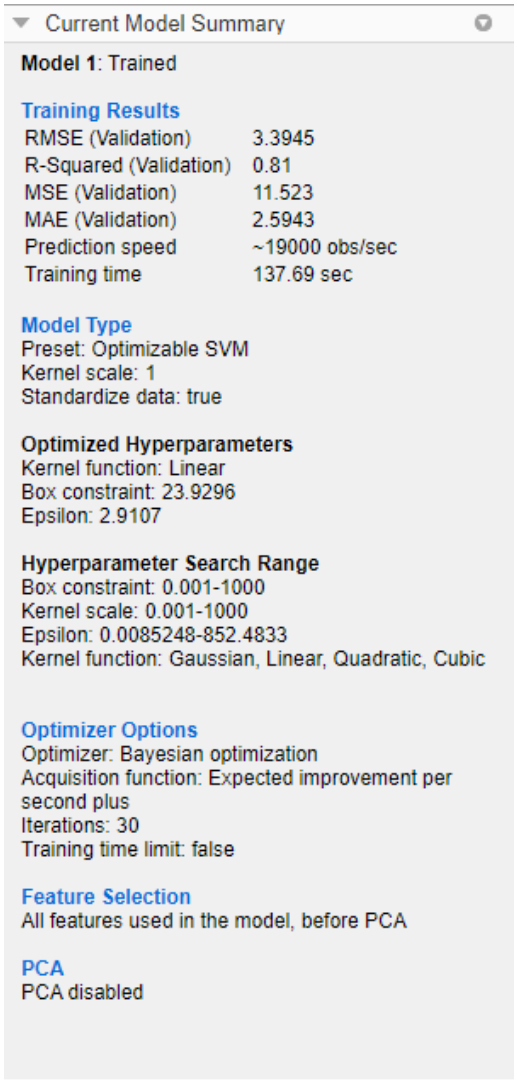
  If you use grid search to perform hyperparameter optimization, the **Bestpoint hyperparameters** and the **Minimum error hyperparameters** are the same.

Missing points in the plot correspond to `NaN` minimum MSE values.

Optimization Results

When the app finishes tuning model hyperparameters, it returns a model trained with the optimized hyperparameter values (**Bestpoint hyperparameters**). The model metrics, displayed plots, and exported model correspond to this trained model with fixed hyperparameter values.

To inspect the optimization results of a trained optimizable model, select the model in the **Models** pane and look at the **Current Model Summary** pane.



The **Current Model Summary** pane includes these sections:

- **Training Results** – Shows the performance of the optimizable model. See “View and Compare Model Statistics” on page 24-43
- **Model Type** – Displays the type of optimizable model and lists any fixed hyperparameter values
- **Optimized Hyperparameters** – Lists the values of the optimized hyperparameters
- **Hyperparameter Search Range** – Displays the search ranges for the optimized hyperparameters
- **Optimizer Options** – Shows the selected optimizer options

When you perform hyperparameter tuning using Bayesian optimization and you export a trained optimizable model to the workspace as a structure, the structure includes a `BayesianOptimization` object in the `HyperParameterOptimizationResult` field. The object contains the results of the optimization performed in the app.

When you generate MATLAB code from a trained optimizable model, the generated code uses the fixed and optimized hyperparameter values of the model to train on new data. The generated code

does not include the optimization process. For information on how to perform Bayesian optimization when you use a fit function, see “Bayesian Optimization Using a Fit Function” on page 10-26.

See Also

Related Examples

- “Train Regression Model Using Hyperparameter Optimization in Regression Learner App” on page 24-76
- “Bayesian Optimization Workflow” on page 10-25
- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54

Assess Model Performance in Regression Learner

In this section...

After training regression models in Regression Learner, you can compare models based on model statistics, visualize results in response plot, or by plotting actual versus predicted response, and evaluate models using the residual plot.

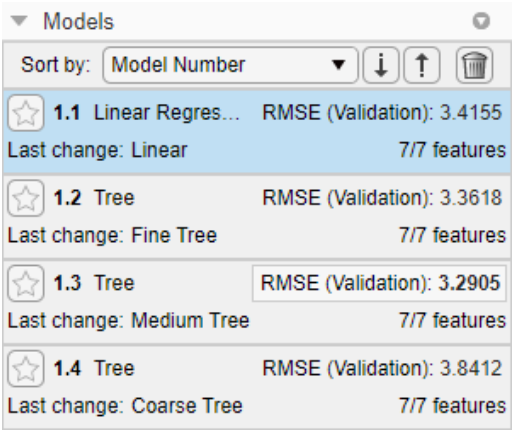
- If you use *k*-fold cross-validation, then the app computes the model statistics using the observations in the *k* validation folds and reports the average values. It makes predictions on the observations in the validation folds and the plots show these predictions. It also computes the residuals on the observations in the validation folds.

  Note When you import data into the app, if you accept the defaults, the app automatically uses cross-validation. To learn more, see “Choose Validation Scheme” on page 23-20.

- If you use holdout validation, the app computes the model statistics using the observations in the validation fold and makes predictions on these observations. The app uses these predictions in the plots and also computes the residuals based on the predictions.
- If you use resubstitution validation, the scores are resubstitution models statistics based on all the training data, and the predictions are resubstitution predictions.

Check Performance in Models Pane

After training a model in Regression Learner, check the **Models** pane to see which model has the best overall score. The best **RMSE (Validation)** is highlighted in a box. This score is the root mean square error (RMSE) on the validation set. The score estimates the performance of the trained model on new data. Use the score to help you choose the best model.

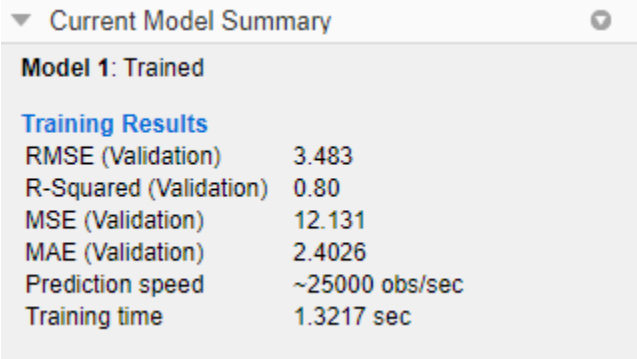


- For cross-validation, the score is the RMSE on all observations, counting each observation when it was in a held-out (validation) fold.
- For holdout validation, the score is the RMSE on the held-out observations.
- For resubstitution validation, the score is the resubstitution RMSE on all the training data.

The best overall score might not be the best model for your goal. Sometimes a model with slightly lower overall score is the better model for your goal. You want to avoid overfitting, and you might want to exclude some predictors where data collection is expensive or difficult.

View and Compare Model Statistics

You can view model statistics in the **Current Model Summary** pane and use these statistics to assess and compare models. The **Training Results** statistics are calculated on the validation set. The **Test Results** statistics, if displayed, are calculated on an imported test set. For more information, see “Evaluate Test Set Model Performance” on page 24-48.



To copy the information in the **Current Model Summary** pane, you can right-click into the pane and select **Copy text**.

Model Statistics

Statistic	Description	Tip
RMSE	Root mean square error. The RMSE is always positive and its units match the units of your response.	Look for smaller values of the RMSE.
R-Squared	Coefficient of determination. R-squared is always smaller than 1 and usually larger than 0. It compares the trained model with the model where the response is constant and equals the mean of the training response. If your model is worse than this constant model, then R-Squared is negative.	Look for an R-Squared close to 1.
MSE	Mean squared error. The MSE is the square of the RMSE.	Look for smaller values of the MSE.
MAE	Mean absolute error. The MAE is always positive and similar to the RMSE, but less sensitive to outliers.	Look for smaller values of the MAE.

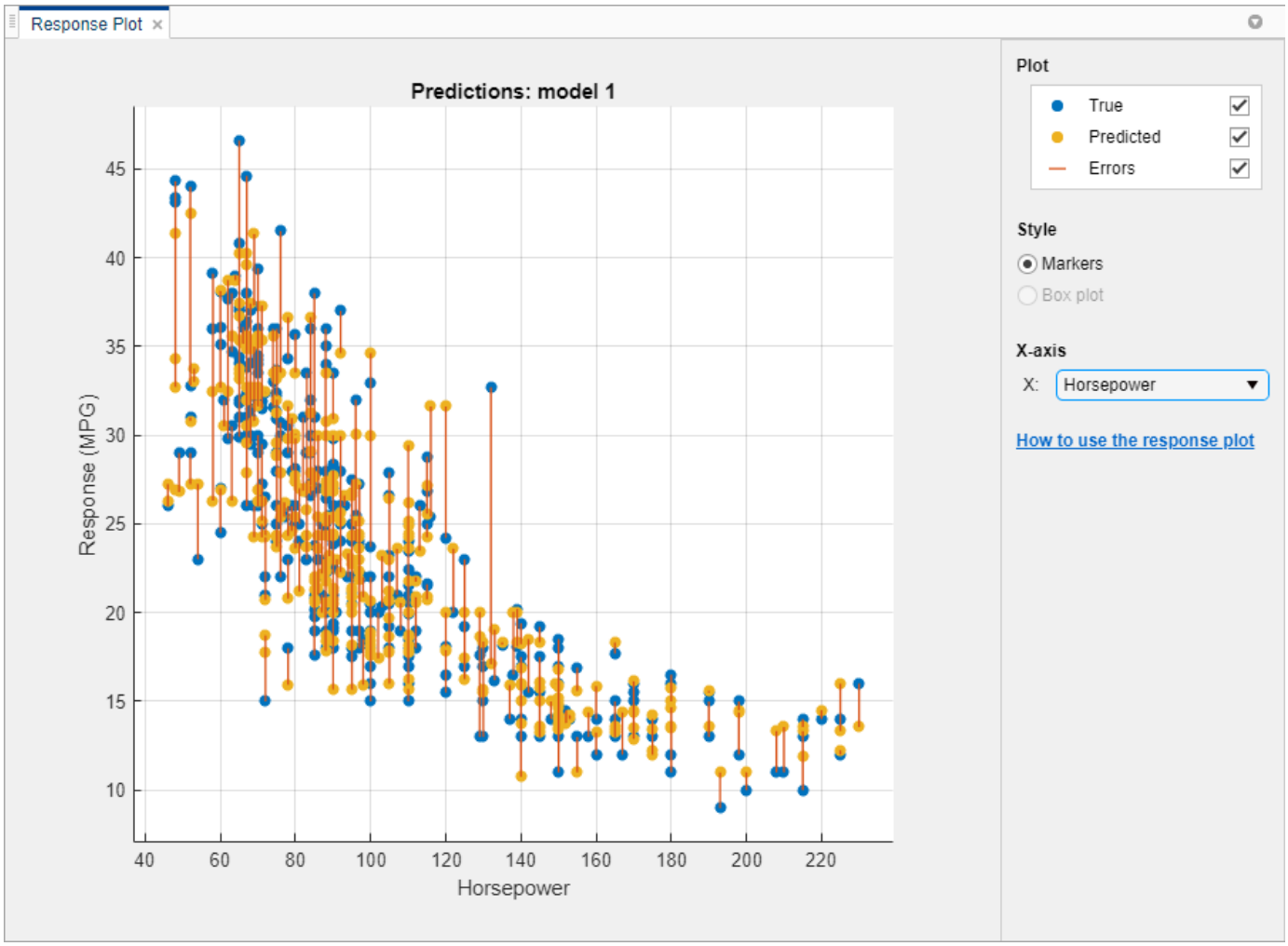
You can sort the models based on the different model statistics. To select a statistic for model sorting, use the **Sort by** list at the top of the **Models** pane.

You can also delete unwanted models listed in the **Models** pane. Select the model you want to delete and click the **Delete selected model** button in the upper right of the pane, or right-click the model and select **Delete model**. You cannot delete the last remaining model in the **Models** pane.

Explore Data and Results in Response Plot

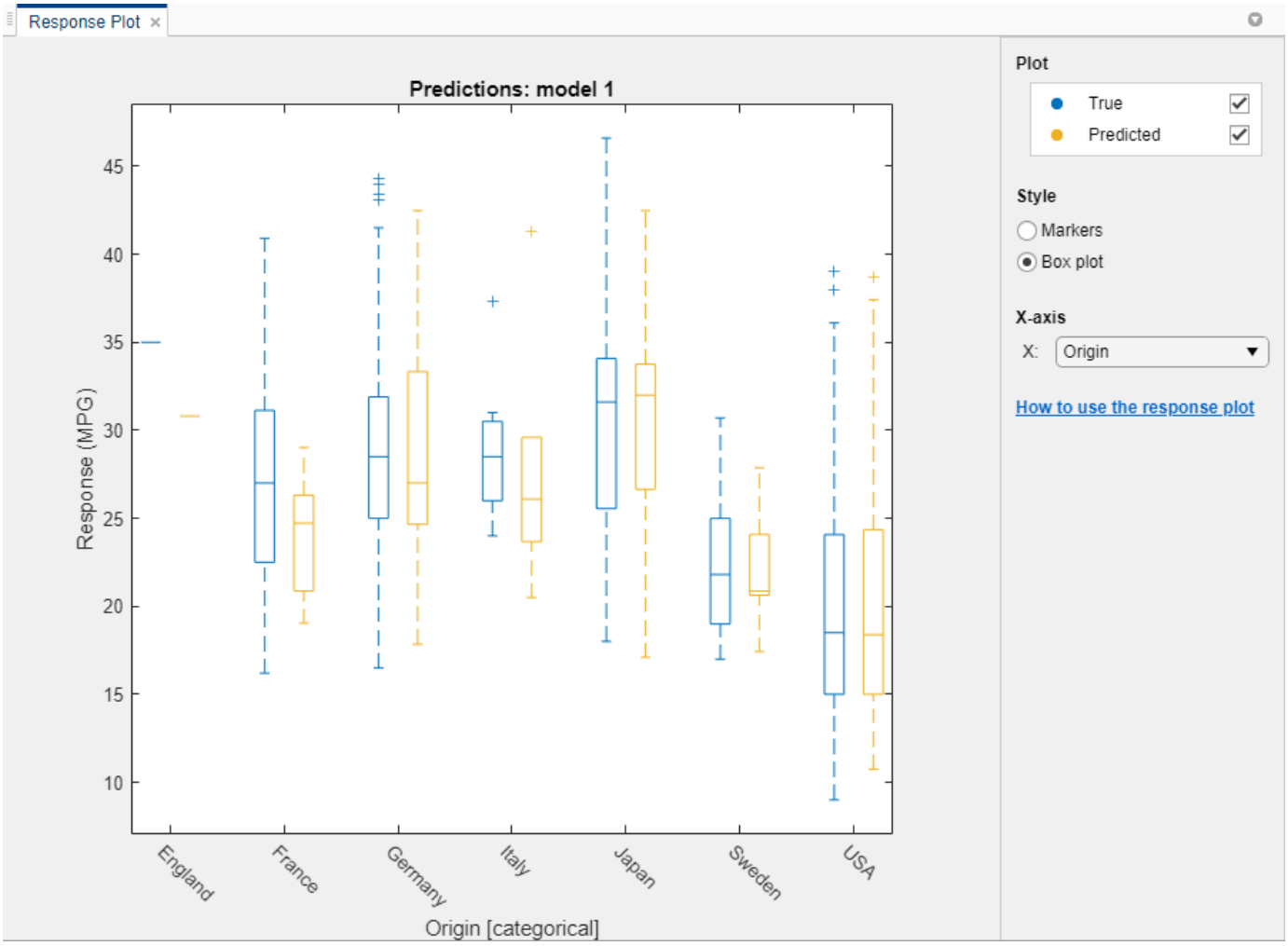
In the response plot, view the regression model results. After you train a regression model, the response plot displays the predicted response versus record number. If you are using holdout or cross-validation, then these predictions are the predictions on the held-out (validation) observations. In other words, each prediction is obtained using a model that was trained without using the corresponding observation. To investigate your results, use the controls on the right. You can:

- Plot predicted and/or true responses. Use the check boxes under **Plot** to make your selection.
- Show prediction errors, drawn as vertical lines between the predicted and true responses, by selecting the **Errors** check box.
- Choose the variable to plot on the *x*-axis under **X-axis**. You can choose either the record number or one of your predictor variables.



- Plot the response as markers, or as a box plot under **Style**. You can only select **Box plot** when the variable on the *x*-axis has few unique values.

A box plot displays the typical values of the response and any possible outliers. The central mark indicates the median, and the bottom and top edges of the box are the 25th and 75th percentiles, respectively. Vertical lines, called whiskers, extend from the boxes to the most extreme data points that are not considered outliers. The outliers are plotted individually using the `'+'` symbol. For more information about box plots, see `boxplot`.



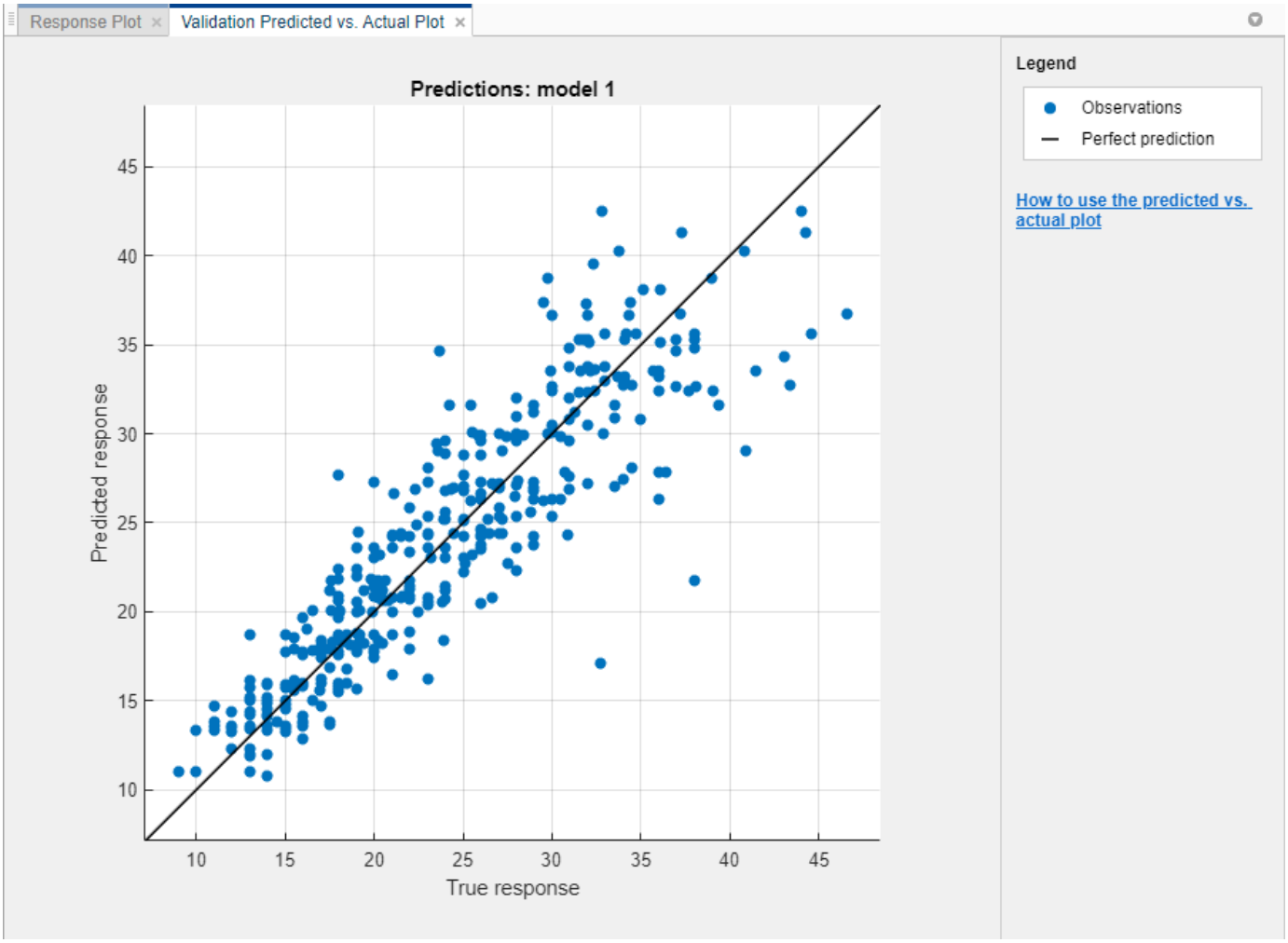
To export the response plots you create in the app to figures, see “Export Plots in Regression Learner App” on page 24-50.

Plot Predicted vs. Actual Response

Use the Predicted vs. Actual plot to check model performance. Use this plot to understand how well the regression model makes predictions for different response values. To view the Predicted vs. Actual plot after training a model, on the **Regression Learner** tab, in the **Plots** section, click **Predicted vs. Actual** and select **Validation Data**.

When you open the plot, the predicted response of your model is plotted against the actual, true response. A perfect regression model has a predicted response equal to the true response, so all the points lie on a diagonal line. The vertical distance from the line to any point is the error of the

prediction for that point. A good model has small errors, and so the predictions are scattered near the line.

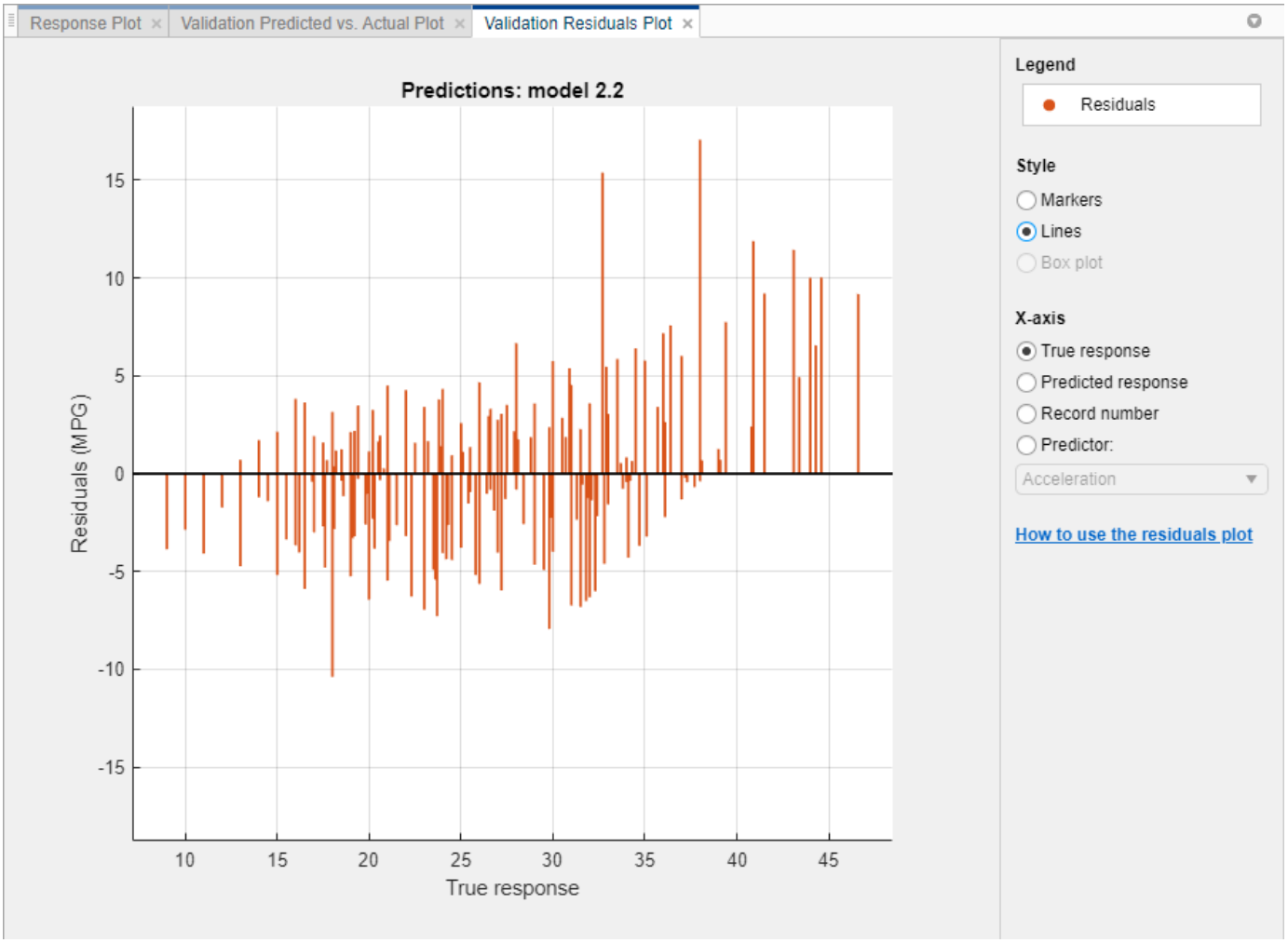


Usually a good model has points scattered roughly symmetrically around the diagonal line. If you can see any clear patterns in the plot, it is likely that you can improve your model. Try training a different model type or making your current model type more flexible using the **Advanced** options in the **Model Type** section. If you are unable to improve your model, it is possible that you need more data, or that you are missing an important predictor.

To export the Predicted vs. Actual plots you create in the app to figures, see “Export Plots in Regression Learner App” on page 24-50.

Evaluate Model Using Residuals Plot

Use the residuals plot to check model performance. To view the residuals plot after training a model, on the **Regression Learner** tab, in the **Plots** section, click **Residuals** and select **Validation Data**. The residuals plot displays the difference between the predicted and true responses. Choose the variable to plot on the *x*-axis under **X-axis**. Choose either the true response, predicted response, record number, or one of your predictors.



Usually a good model has residuals scattered roughly symmetrically around 0. If you can see any clear patterns in the residuals, it is likely that you can improve your model. Look for these patterns:

- Residuals are not symmetrically distributed around 0.
- Residuals change significantly in size from left to right in the plot.
- Outliers occur, that is, residuals that are much larger than the rest of the residuals.
- Clear, nonlinear pattern appears in the residuals.

Try training a different model type, or making your current model type more flexible using the **Advanced** options in the **Model Type** section. If you are unable to improve your model, it is possible that you need more data, or that you are missing an important predictor.

To export the residuals plots you create in the app to figures, see “Export Plots in Regression Learner App” on page 24-50.

Evaluate Test Set Model Performance

After training a model in Regression Learner, you can evaluate the model performance on a test set in the app. This process allows you to check whether the validation metrics provide good estimates for the model performance on new data.

1. Import a test data set into Regression Learner.

   - If the test data set is in the MATLAB workspace, then in the **Testing** section on the **Regression Learner** tab, click **Test Data** and select **From Workspace**.
   - If the test data set is in a file, then in the **Testing** section, click **Test Data** and select **From File**. Select a file type in the list, such as a spreadsheet, text file, or comma-separated values (`.csv`) file, or select **All Files** to browse for other file types such as `.dat`.

   In the Import Test Data dialog box, select the test data set from the **Test Data Set Variable** list. The test set must have the same variables as the predictors imported for training and validation.

2. Compute the test set metrics.

   - To compute test metrics for a single model, select the trained model in the **Models** pane. On the **Regression Learner** tab, in the **Testing** section, click **Test All** and select **Test Selected**.
   - To compute test metrics for all trained models, click **Test All** and select **Test All** in the **Testing** section.

   The app computes the test set performance of each model trained on the full data set, including training and validation data.

3. Compare the validation metrics with the test metrics.

   In the **Current Model Summary** pane, the app displays the validation metrics and test metrics under the **Training Results** section and **Test Results** section, respectively. You can check if the validation metrics give good estimates for the test metrics.

   You can also visualize the test results using plots.

   - Display a predicted vs. actual plot. In the **Plots** section on the **Regression Learner** tab, click **Predicted vs. Actual Plot** and select **Test Data**.
   - Display a residuals plot. In the **Plots** section, click **Residuals Plot** and select **Test Data**.

For an example, see “Check Model Performance Using Test Set in Regression Learner App” on page 24-83. For an example that uses test set metrics in a hyperparameter optimization workflow, see “Train Regression Model Using Hyperparameter Optimization in Regression Learner App” on page 24-76.

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26

- “Export Plots in Regression Learner App” on page 24-50
- “Export Regression Model to Predict New Data” on page 24-54
- “Train Regression Trees Using Regression Learner App” on page 24-60

Export Plots in Regression Learner App

After you create plots interactively in the Regression Learner app, you can export your app plots to MATLAB figures. You can then copy, save, or customize the new figures. Choose among the available plots: response plot on page 24-26, Predicted vs. Actual plot on page 24-45, residuals plot on page 24-46, and minimum MSE plot on page 24-37.

- Before exporting a plot, make sure the plot in the app displays the same data that you want in the new figure.
- On the **Regression Learner** tab, in the **Export** section, click **Export Plot to Figure**. The app creates a figure from the selected plot.
  - The new figure might not have the same interactivity options as the plot in the Regression Learner app.
  - Additionally, the figure might have a different axes toolbar than the one in the app plot. For plots in Regression Learner, an axes toolbar appears above the top right of the plot. The buttons available on the toolbar depend on the contents of the plot. The toolbar can include buttons to export the plot as an image, add data tips, pan or zoom the data, and restore the view.

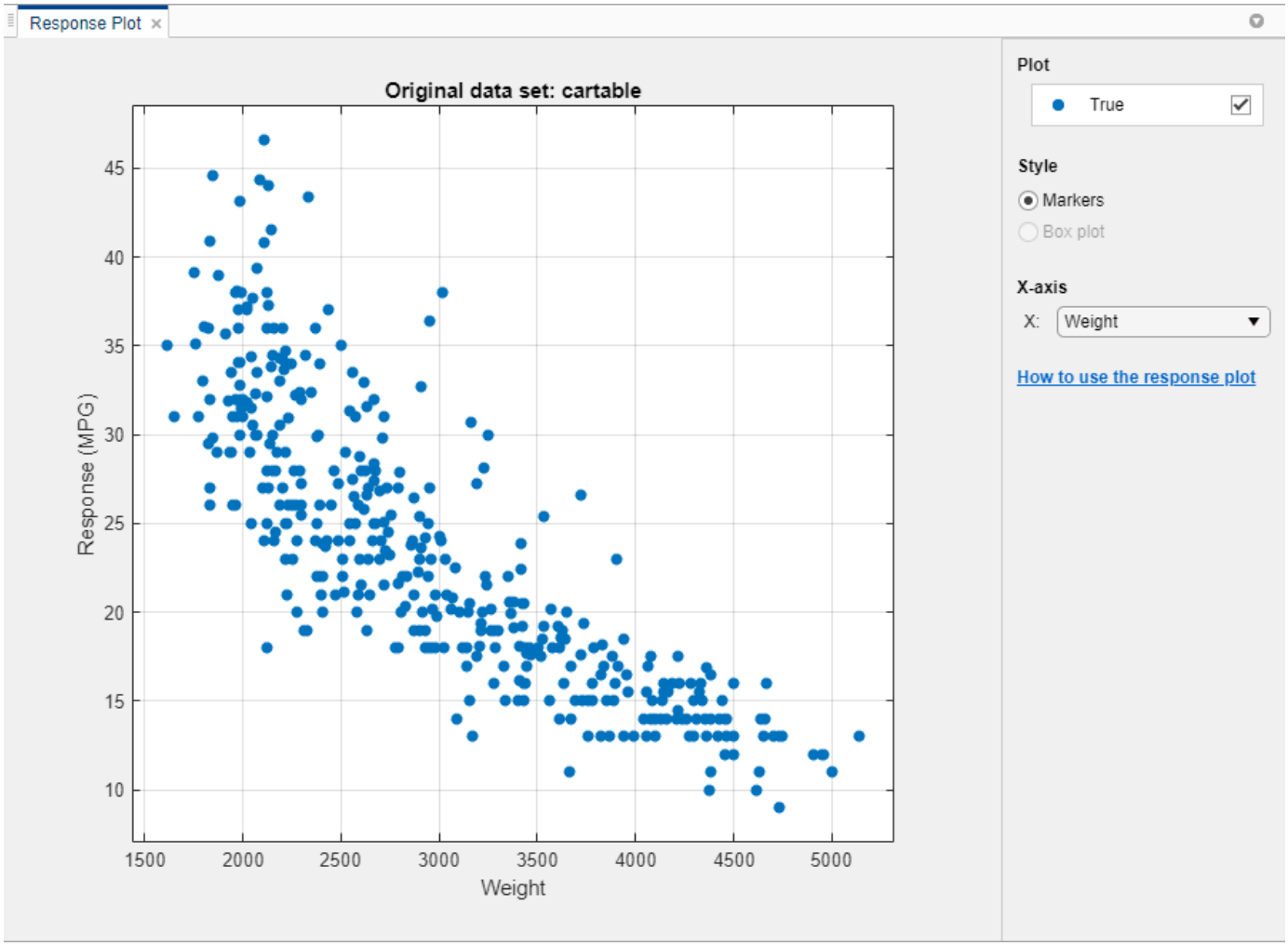
    
- Copy, save, or customize the new figure, which is displayed in the figure window.
  - To copy the figure, select **Edit > Copy Figure**. For more information, see “Copy Figure to Clipboard from Edit Menu”.
  - To save the figure, select **File > Save As**. Alternatively, you can follow the workflow described in “Customize Figure Before Saving”.
  - To customize the figure, click the Edit Plot button  on the figure toolbar. Right-click the section of the plot that you want to edit. You can change the listed properties, which might include **Color**, **Font**, **Line Style**, and other properties. Or, you can use the **Property Inspector** to change the figure properties.

As an example, export a response plot in the app to a figure, customize the figure, and save the modified figure.

1. In the MATLAB Command Window, load the `carbig` data set.

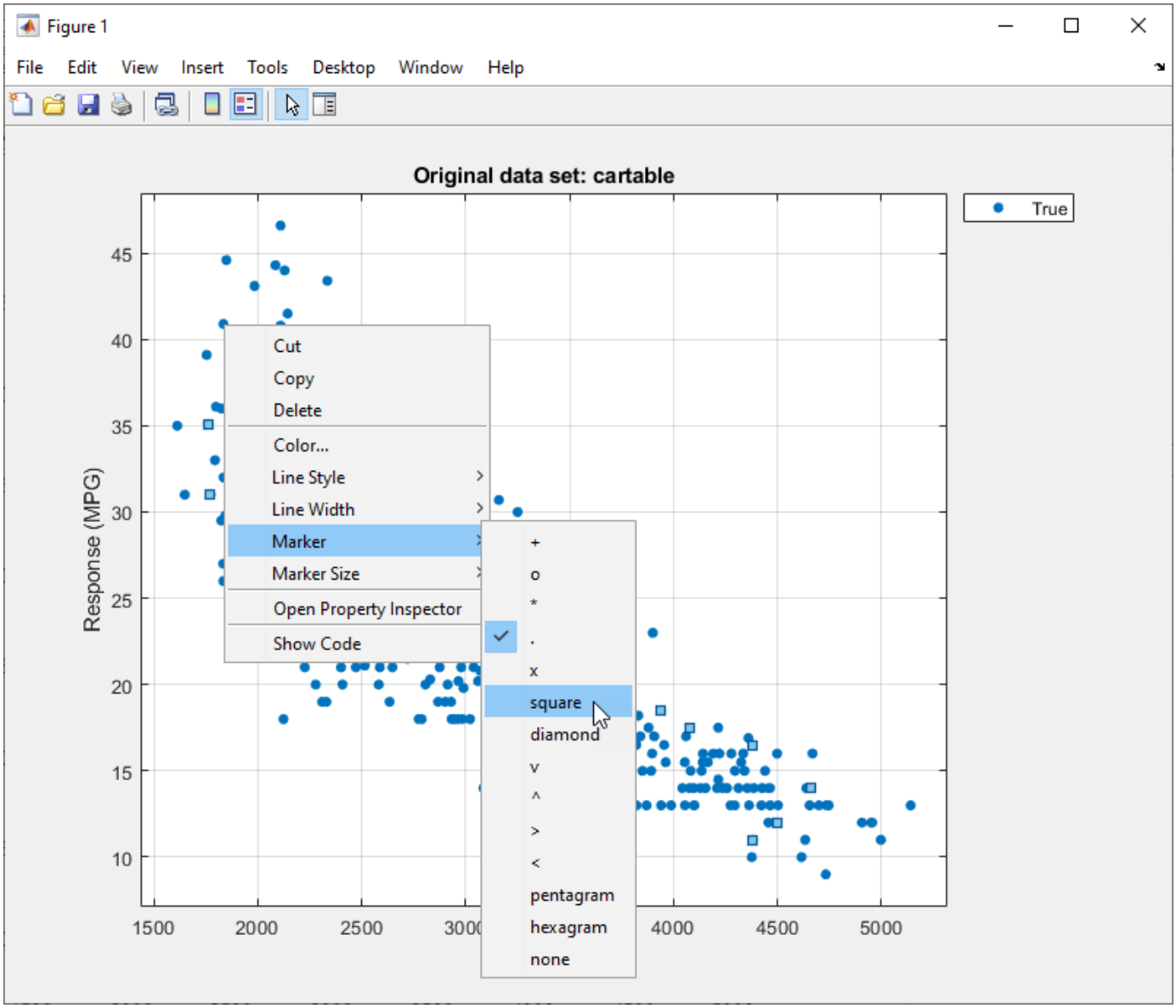
   ```
   load carbig
   cartable = table(Acceleration,Cylinders,Displacement, ...
       Horsepower,Model_Year,Weight,Origin,MPG);
   ```
2. Click the **Apps** tab.
3. In the **Apps** section, click the arrow to open the gallery. Under **Machine Learning and Deep Learning**, click **Regression Learner**.
4. On the **Regression Learner** tab, in the **File** section, click  .
5. In the New Session from Workspace dialog box, select the table `cartable` from the **Workspace Variable** list.
6. Click **Start Session**. Regression Learner creates a response plot of the data by default.

7 Change the x-axis data in the response plot to `Weight`.

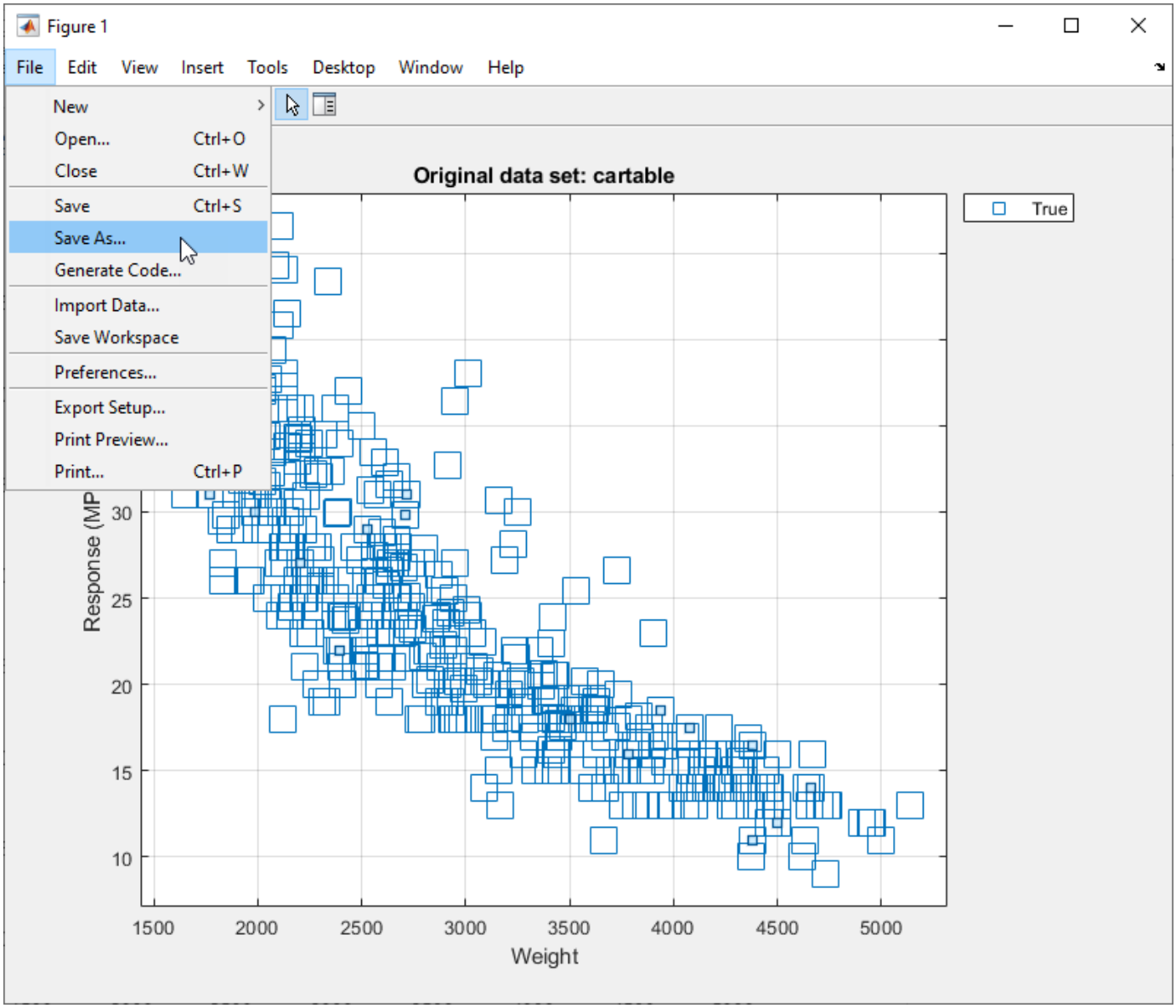


8 On the **Regression Learner** tab, in the **Export** section, click **Export Plot to Figure**.

9 In the new figure, click the Edit Plot button on the figure toolbar. Right-click the points in the plot. In the context menu, select **Marker > square**.



10 To save the figure, select **File** > **Save As**. Specify the saved file location, name, and type.



See Also

Related Examples

- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54

Export Regression Model to Predict New Data

In this section...

Export Model to Workspace

After you create regression models interactively in the Regression Learner app, you can export your best model to the workspace. Then you can use that trained model to make predictions using new data.

Note The final model Regression Learner exports is always trained using the full data set. The validation scheme that you use only affects the way that the app computes validation metrics. You can use the validation metrics and various plots that visualize results to pick the best model for your regression problem.

Here are the steps for exporting a model to the MATLAB workspace:

1. In the app, select the model you want to export in the **Models** pane.
2. On the **Regression Learner** tab, in the **Export** section, click one of the export options:
   - To include the data used for training the model, select **Export Model**.

     You export the trained model to the workspace as a structure containing a regression model object.
   - To exclude the training data, select **Export Compact Model**. This option exports the model with unnecessary data removed where possible. For some models this is a compact object that does not include the training data, but you can still use it for making predictions on new data.
3. In the Export Model dialog box, check the name of your exported variable, and edit it if you want. Then, click **OK**. The default name for your exported model, `trainedModel`, increments every time you export to avoid overwriting your models (for example, `trainedModel1`).

   The new variable (for example, `trainedModel`) appears in your workspace.

   The app displays information about the exported model in the command window. Read the message to learn how to make predictions with new data.

Make Predictions for New Data

After you export a model to the workspace from Regression Learner, or run the code generated from the app, you get a `trainedModel` structure that you can use to make predictions using new data. The structure contains a model object and a function for prediction. The structure enables you to make predictions for models that include principal component analysis (PCA).

1. Use the exported model to make predictions for new data, `T`:

   ```
   yfit = trainedModel.predictFcn(T)
   ```

   where `trainedModel` is the name of your exported variable.

   Supply the data `T` with the same format and data type as the training data used in the app (table or matrix).

   - If you supply a table, then ensure that it contains the same predictor names as your training data. The `predictFcn` ignores additional variables in tables. Variable formats and types must match the original training data.
   - If you supply a matrix, it must contain the same predictor columns or rows as your training data, in the same order and format. Do not include a response variable, any variables that you did not import in the app, or other unused variables.

   The output `yfit` contains a prediction for each data point.

2. Examine the fields of the exported structure. For help making predictions, enter:

   ```
   trainedModel.HowToPredict
   ```

You also can extract the model object from the exported structure for further analysis. If you use feature transformation such as PCA in the app, you must take into account this transformation by using the information in the PCA fields of the structure.

Generate MATLAB Code to Train Model with New Data

After you create regression models interactively in the Regression Learner app, you can generate MATLAB code for your best model. Then you can use the code to train the model with new data.

Generate MATLAB code to:

- Train on huge data sets. Explore models in the app trained on a subset of your data, and then generate code to train a selected model on a larger data set.
- Create scripts for training models without needing to learn syntax of the different functions.
- Examine the code to learn how to train models programmatically.
- Modify the code for further analysis, for example to set options that you cannot change in the app.
- Repeat your analysis on different data and automate training.

To generate code and use it to train a model with new data:

1. In the app, from the **Models** pane, select the model you want to generate code for.
2. On the **Regression Learner** tab, in the **Export** section, click **Generate Function**.

   The app generates code from your session and displays the file in the MATLAB Editor. The file includes the predictors and response, the model training methods, and the validation methods. Save the file.
3. To retrain your model, call the function from the command line with your original data or new data as the input argument or arguments. New data must have the same shape as the original data.

Copy the first line of the generated code, excluding the word `function`, and edit the `trainingData` input argument to reflect the variable name of your training data or new data. Similarly, edit the `responseData` input argument (if applicable).

For example, to retrain a regression model trained with the `cartable` data set, enter:

```
[trainedModel,validationRMSE] = trainRegressionModel(cartable)
```

The generated code returns a `trainedModel` structure that contains the same fields as the structure you create when you export a model from Regression Learner to the workspace.

If you want to automate training the same model with new data, or learn how to programmatically train models, examine the generated code. The code shows you how to:

- Process the data into the right shape.
- Train a model and specify all the model options.
- Perform cross-validation.
- Compute statistics.
- Compute validation predictions and scores.

Note If you generate MATLAB code from a trained optimizable model, the generated code does not include the optimization process.

Generate C Code for Prediction

If you train one of the models in this table using Regression Learner, you can generate C code for prediction.

Model Type	Underlying Model Object
Linear Regression	`LinearModel` or `CompactLinearModel`
Decision Tree	`RegressionTree` or `CompactRegressionTree`
Support Vector Machine	`RegressionSVM` or `CompactRegressionSVM`
Gaussian Process Regression	`RegressionGP` or `CompactRegressionGP`
Ensemble	`RegressionEnsemble`, `CompactRegressionEnsemble`, or `RegressionBaggedEnsemble`

C code generation requires:

- MATLAB Coder license
- Appropriate model

1 For example, train a tree model in Regression Learner, and then export the model to the workspace.

Find the underlying regression model object in the exported structure. Examine the fields of the structure to find the model object, for example, `S.RegressionTree`, where `S` is the name of your structure.

The underlying model object depends on whether you exported a compact model. The model object can be a `RegressionTree` or `CompactRegressionTree` object.

2. Use the function `saveLearnerForCoder` to prepare the model for code generation: `saveLearnerForCoder(Mdl,filename)`. For example:

```
saveLearnerForCoder(S.RegressionTree,'myTree')
```

3. Create a function that loads the saved model and makes predictions on new data. For example:

```
function yfit = predictY (X) %#codegen
%PREDICTY Predict responses using tree model
%  PREDICTY uses the measurements in X
%  and the tree model in the file myTree.mat, and then
%  returns predicted responses in yfit.

CompactMdl = loadLearnerForCoder('myTree');
yfit = predict(CompactMdl,X);
end
```

4. Generate a MEX function from your function. For example:

```
codegen predictY.m -args {data}
```

The `%#codegen` compilation directive indicates that the MATLAB code is intended for code generation. To ensure that the MEX function can use the same input, specify the data in the workspace as arguments to the function using the `-args` option. Specify `data` as a matrix containing only the predictor columns used to train the model.

5. Use the MEX function to make predictions. For example:

```
yfit = predictY_mex(data);
```

If you used feature selection or PCA feature transformation in the app, then you need to take additional steps. If you used manual feature selection, supply the same columns in `X`. The `X` argument is the input to your function.

If you used PCA in the app, use the information in the PCA fields of the exported structure to take account of this transformation. It does not matter whether you imported a table or a matrix into the app, as long as `X` contains the matrix columns in the same order. Before generating code, follow these steps:

1. Save the PCACenters and PCACoefficients fields of the trained regression structure, `S`, to file using the following command:

```
save('pcaInfo.mat','-struct','S','PCACenters','PCACoefficients');
```

2. In your function file, include additional lines to perform the PCA transformation. Create a function that loads the saved model, performs PCA, and makes predictions on new data. For example:

```
function yfit = predictY (X) %#codegen
%PREDICTY Predict responses using tree model
%  PREDICTY uses the measurements in X
%  and the tree model in the file myTree.mat,
%  and then returns predicted responses in yfit.
% If you used manual feature selection in the app, ensure that X
% contains only the columns you included in the model.

CompactMdl = loadLearnerForCoder('myTree');
```

```
pcaInfo = coder.load('pcaInfo.mat','PCACenters','PCACoefficients');
PCACenters = pcaInfo.PCACenters;
PCACoefficients = pcaInfo.PCACoefficients;

% Performs PCA transformation
pcaTransformedX = bsxfun(@minus,X,PCACenters)*PCACoefficients;

yfit = predict(CompactMdl,pcaTransformedX);
end
```

For more information on the C code generation workflow and limitations, see “Code Generation”. For examples, see `saveLearnerForCoder` and `loadLearnerForCoder`.

Deploy Predictions Using MATLAB Compiler

After you export a model to the workspace from Regression Learner, you can deploy it using MATLAB Compiler.

Suppose you export the trained model to MATLAB Workspace based on the instructions in “Export Model to Workspace” on page 24-54, with the name `trainedModel`. To deploy predictions, follow these steps.

- Save the `trainedModel` structure in a .mat file.

  ```
  save mymodel trainedModel
  ```

- Write the code to be compiled. This code must load the trained model and use it to make a prediction. It must also have a pragma, so the compiler recognizes that Statistics and Machine Learning Toolbox code is needed in the compiled application. This pragma could be any function in the toolbox.

  ```
  function ypred = mypredict(tbl)
  %#function fitrtree
  load('mymodel.mat');
  ypred = trainedModel.predictFcn(tbl);
  end
  ```

- Compile as a standalone application.

  ```
  mcc -m mypredict.m
  ```

See Also

Functions
`fitlm` | `fitrensemble` | `fitrgp` | `fitrnet` | `fitrsvm` | `fitrtree` | `stepwiselm`

Classes
`CompactLinearModel` | `CompactRegressionEnsemble` | `CompactRegressionGP` | `CompactRegressionNeuralNetwork` | `CompactRegressionSVM` | `CompactRegressionTree` | `LinearModel` | `RegressionEnsemble` | `RegressionGP` | `RegressionNeuralNetwork` | `RegressionSVM` | `RegressionTree`

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2

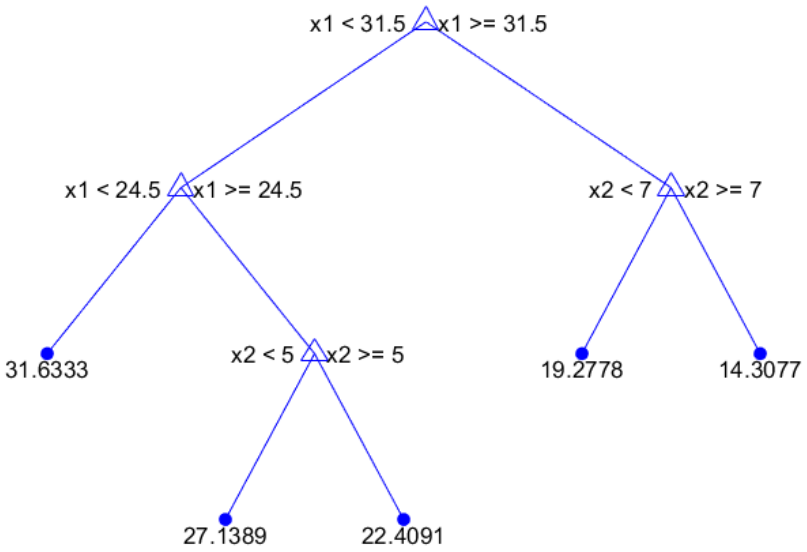
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Train Regression Trees Using Regression Learner App” on page 24-60

Train Regression Trees Using Regression Learner App

This example shows how to create and compare various regression trees using the Regression Learner app, and export trained models to the workspace to make predictions for new data.

You can train regression trees to predict responses to given input data. To predict the response of a regression tree, follow the tree from the root (beginning) node down to a leaf node. At each node, decide which branch to follow using the rule associated to that node. Continue until you arrive at a leaf node. The predicted response is the value associated to that leaf node.

Statistics and Machine Learning Toolbox trees are binary. Each step in a prediction involves checking the value of one predictor variable. For example, here is a simple regression tree:



This tree predicts the response based on two predictors, `x1` and `x2`. To predict, start at the top node. At each node, check the values of the predictors to decide which branch to follow. When the branches reach a leaf node, the response is set to the value corresponding to that node.

This example uses the `carbig` data set. This data set contains characteristics of different car models produced from 1970 through 1982, including:

- Acceleration
- Number of cylinders
- Engine displacement
- Engine power (Horsepower)
- Model year
- Weight
- Country of origin
- Miles per gallon (MPG)

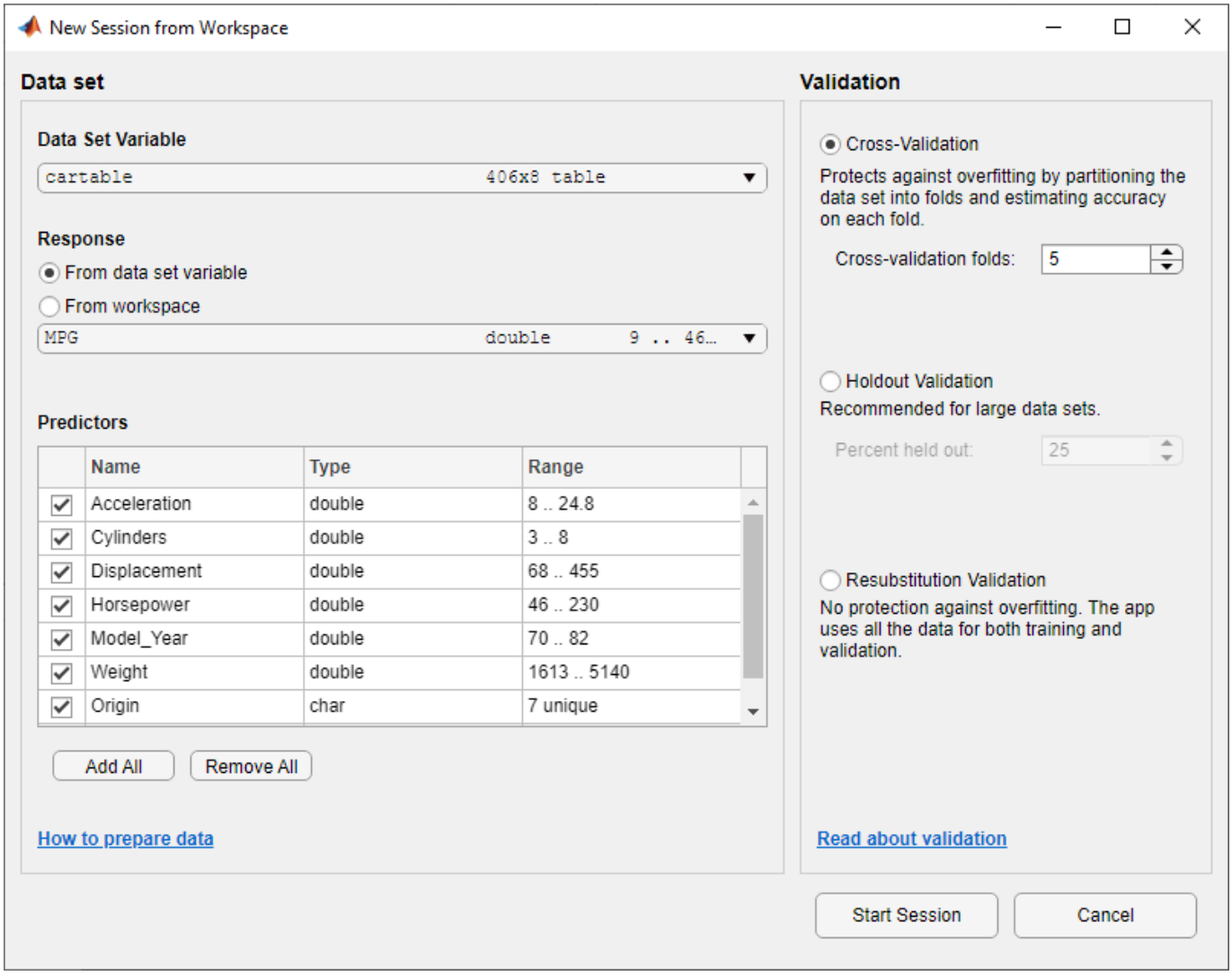
Train regression trees to predict the fuel economy in miles per gallon of a car model, given the other variables as inputs.

1. In MATLAB, load the `carbig` data set and create a table containing the different variables:

   ```
   load carbig
   cartable = table(Acceleration, Cylinders, Displacement,...
   Horsepower, Model_Year, Weight, Origin, MPG);
   ```

2. On the **Apps** tab, in the **Machine Learning and Deep Learning** group, click **Regression Learner**.
3. On the **Regression Learner** tab, in the **File** section, select **New Session > From Workspace**.
4. Under **Data Set Variable** in the New Session from Workspace dialog box, select `cartable` from the list of tables and matrices in your workspace.

   Observe that the app has preselected response and predictor variables. `MPG` is chosen as the response, and all the other variables as predictors. For this example, do not change the selections.

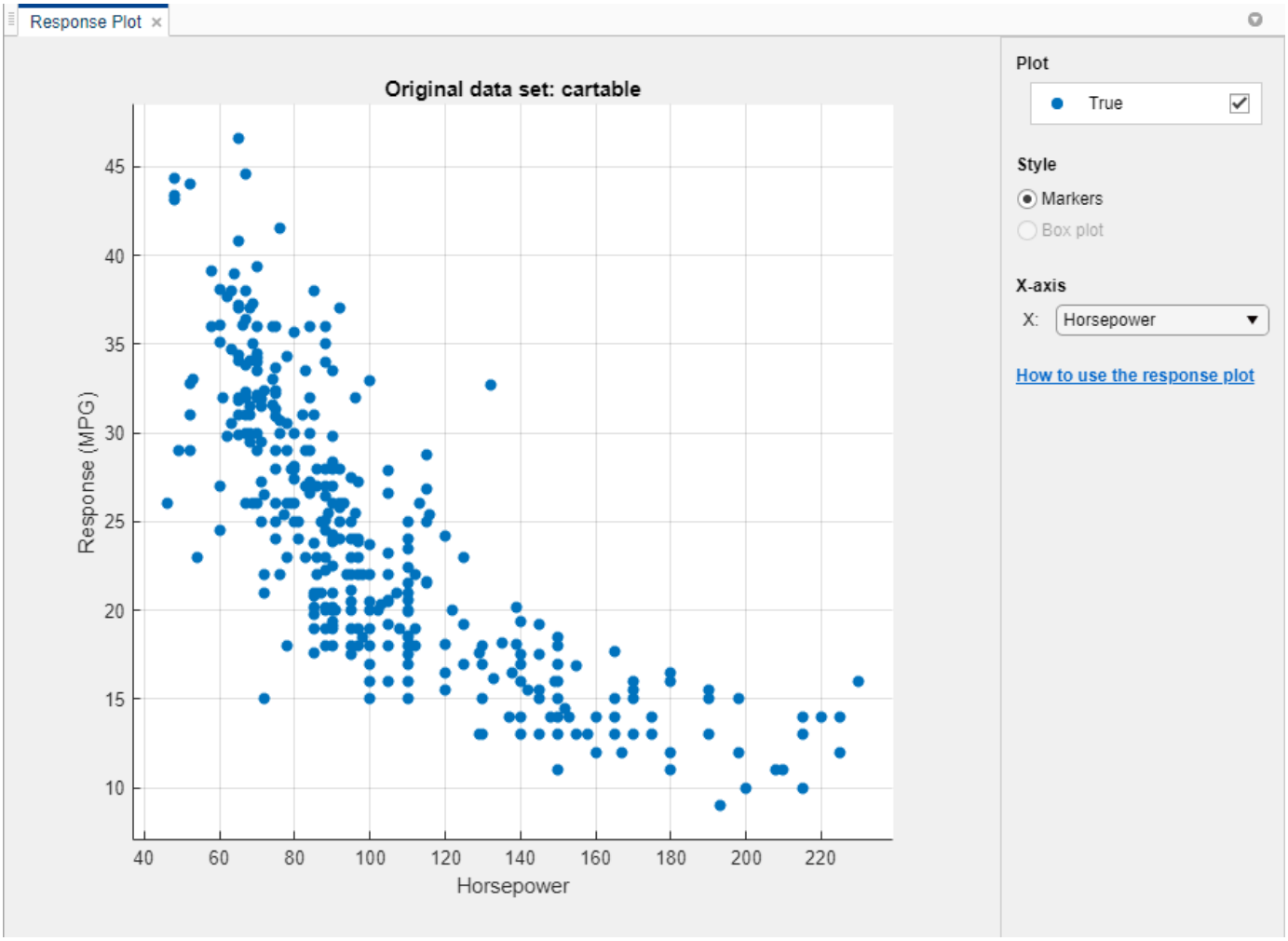


5 To accept the default validation scheme and continue, click **Start Session**. The default validation option is cross-validation, to protect against overfitting.

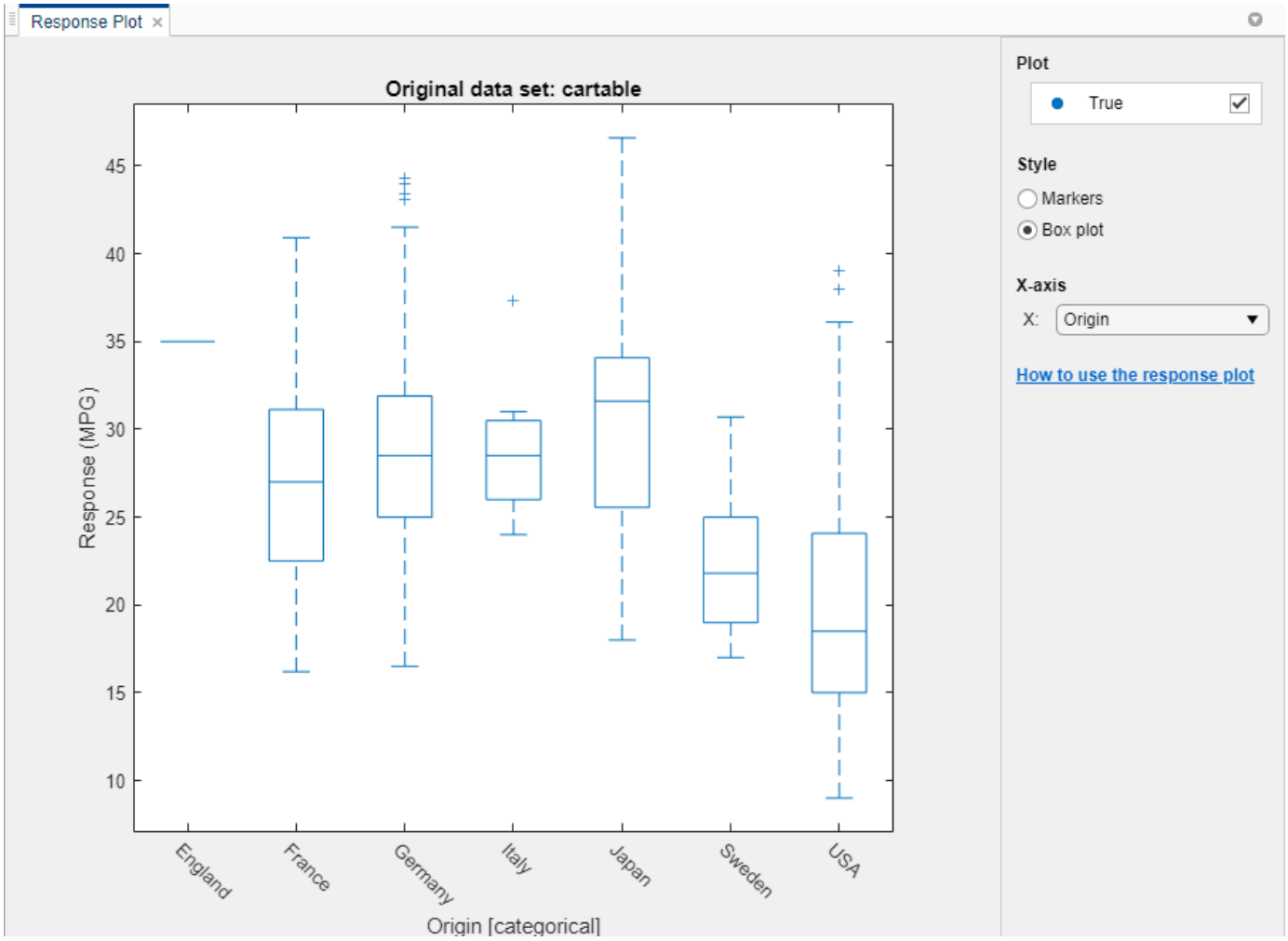
Regression Learner creates a plot of the response with the record number on the *x*-axis.

6 Use the response plot to investigate which variables are useful for predicting the response. To visualize the relation between different predictors and the response, select different variables in the **X** list under **X-axis**.

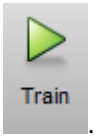
Observe which variables are correlated most clearly with the response. `Displacement`, `Horsepower`, and `Weight` all have a clearly visible impact on the response and all show a negative association with the response.



7 Select the variable `Origin` under **X-axis**. A box plot is automatically displayed. A box plot shows the typical values of the response and any possible outliers. The box plot is useful when plotting markers results in many points overlapping. To show a box plot when the variable on the *x*-axis has few unique values, under **Style**, select **Box plot**.



8 Create a selection of regression trees. On the **Regression Learner** tab, in the **Model Type** section, click **All Trees**  .

Then click **Train** 

.

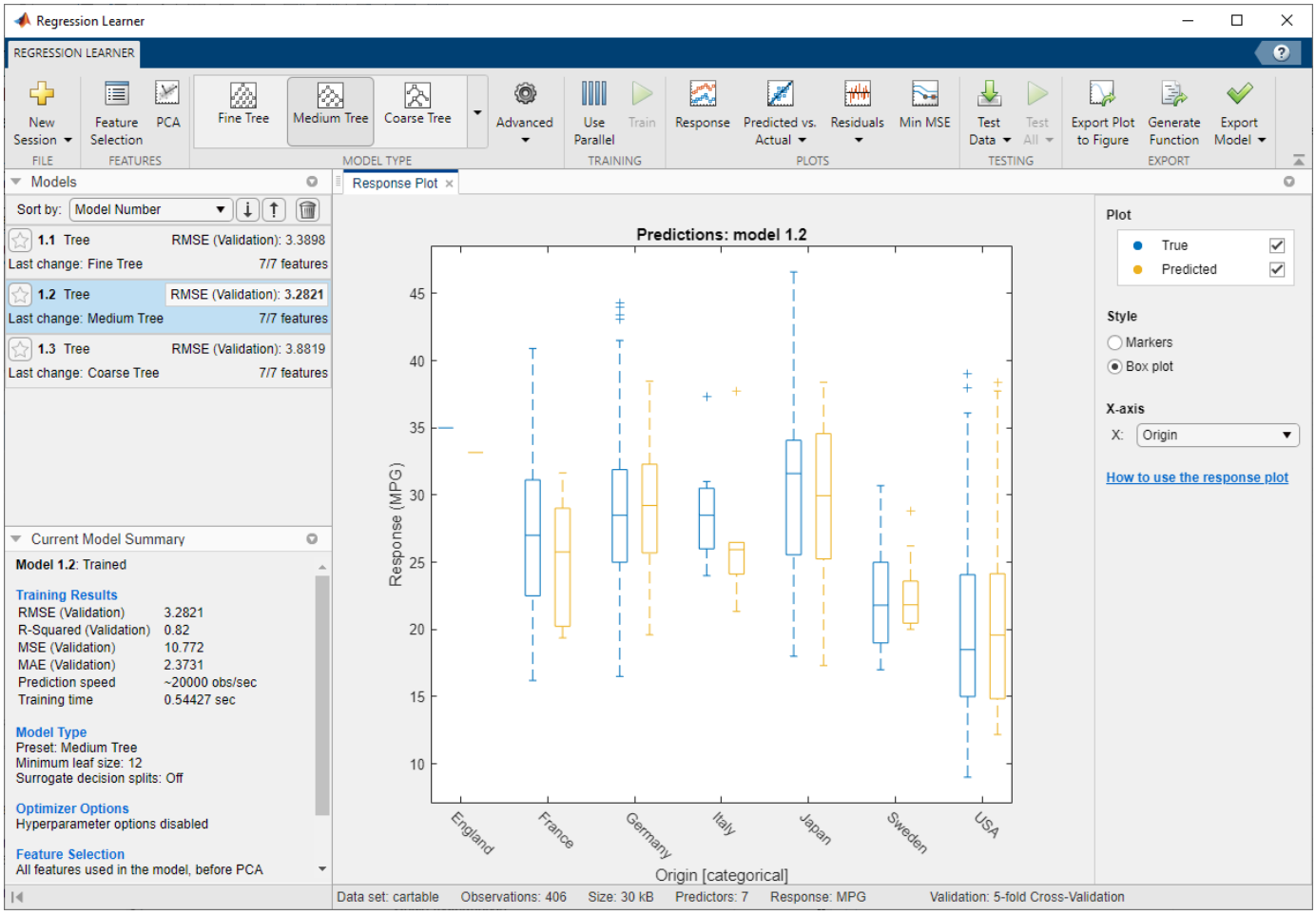
Tip If you have Parallel Computing Toolbox, you can train all the models (**All Trees**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

Regression Learner creates and trains three regression trees: a **Fine Tree**, a **Medium Tree**, and a **Coarse Tree**.

The three models appear in the **Models** pane. Check the **RMSE (Validation)** (validation root mean square error) of the models. The best score is highlighted in a box.

The **Fine Tree** and the **Medium Tree** have similar RMSEs, while the **Coarse Tree** is less accurate.

Regression Learner plots both the true training response and the predicted response of the currently selected model.

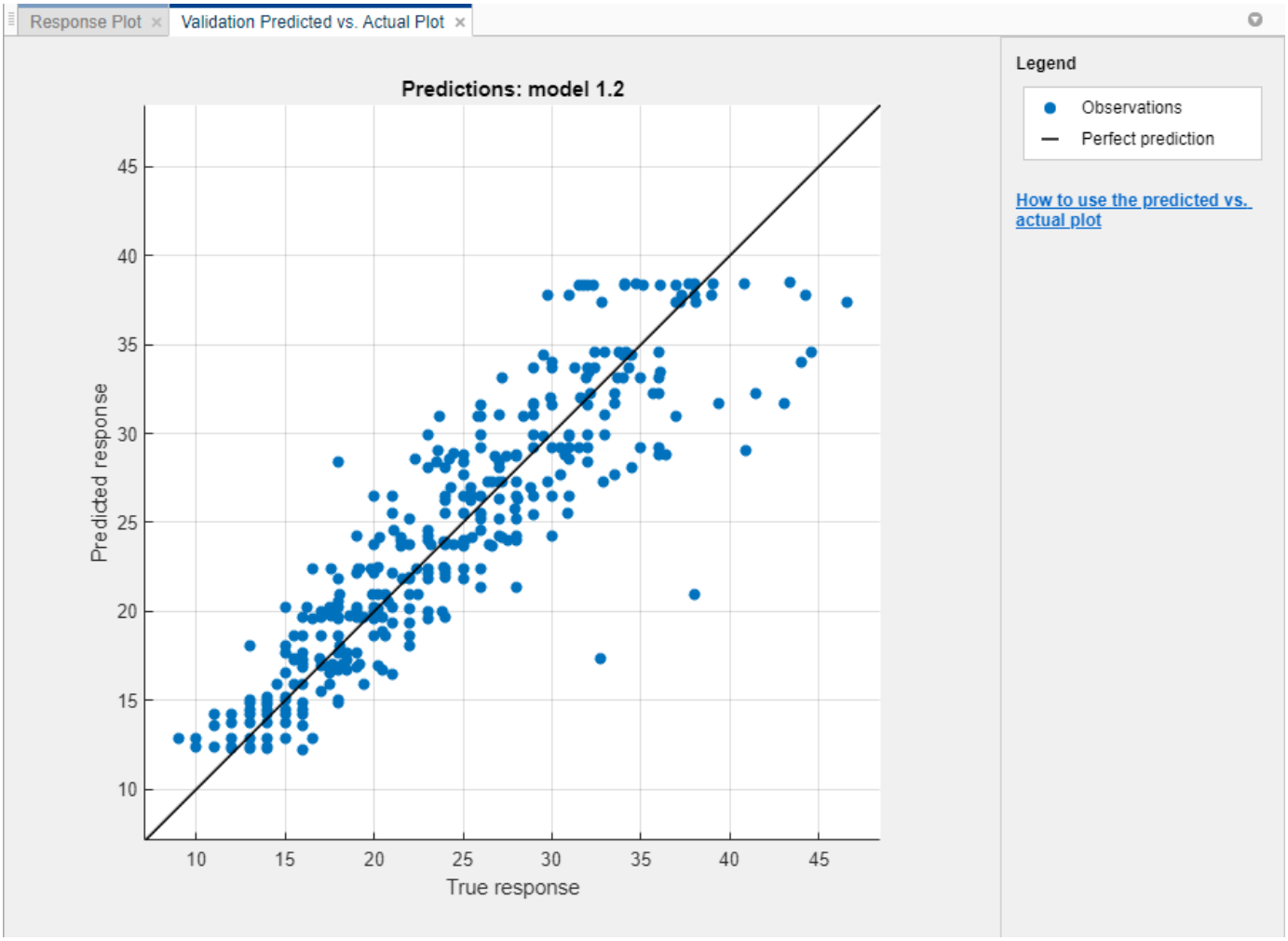


Note If you are using validation, there is some randomness in the results and so your model validation score can differ from the results shown.

9 Choose a model in the **Models** pane to view the results of that model. Under **X-axis**, select `Horsepower` and examine the response plot. Both the true and predicted responses are now plotted. Show the prediction errors, drawn as vertical lines between the predicted and true responses, by selecting the **Errors** check box.

10 See more details on the currently selected model in the **Current Model Summary** pane. Check and compare additional model characteristics, such as R-squared (coefficient of determination), MAE (mean absolute error), and prediction speed. To learn more, see “View and Compare Model Statistics” on page 24-43. In the **Current Model Summary** pane, you also can find details on the currently selected model type, such as options used for training the model.

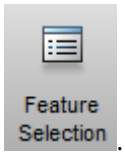
11 Plot the predicted response versus true response. On the **Regression Learner** tab, in the **Plots** section, click **Predicted vs. Actual** and select **Validation Data**. Use this plot to understand how well the regression model makes predictions for different response values.



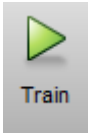
A perfect regression model has predicted response equal to true response, so all the points lie on a diagonal line. The vertical distance from the line to any point is the error of the prediction for that point. A good model has small errors, so the predictions are scattered near the line. Usually a good model has points scattered roughly symmetrically around the diagonal line. If you can see any clear patterns in the plot, it is likely that you can improve your model.

12 Select the other models in the **Models** pane and compare the predicted versus actual plots.

13 In the **Model Type** gallery, select **All Trees** again. To try to improve the model, try including different features in the model. See if you can improve the model by removing features with low predictive power. On the **Regression Learner** tab, in the **Features** section, click **Feature Selection**.



In the Feature Selection dialog box, clear the check boxes for **Acceleration** and **Cylinders** to exclude them from the predictors.

Click **Train** 

to train new regression trees using the new predictor settings.

14 Observe the new models in the **Models** pane. These models are the same regression trees as before, but trained using only five of seven predictors. The app displays how many predictors are used. To check which predictors are used, click a model in the **Models** pane and observe the check boxes in the Feature Selection dialog box.

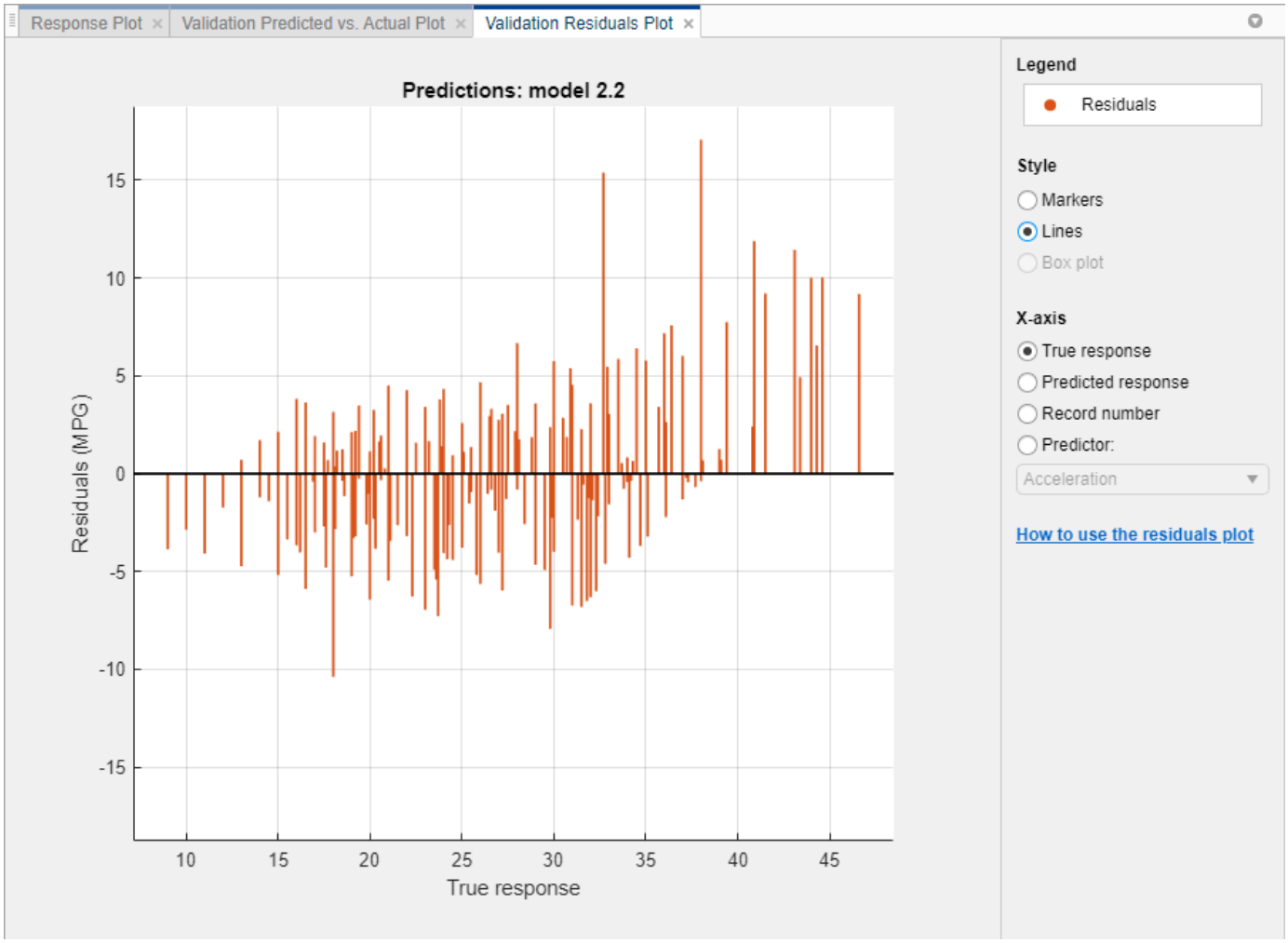
The models with the two features removed perform comparably to the models using all predictors. The models predict no better using all the predictors compared to using only a subset of the predictors. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily without some predictors.

15 Train the three regression tree presets using only `Horsepower` as predictor. Change the selections in the Feature Selection dialog box and click **Train**.

Using only the engine power as predictor results in models with lower accuracy. However, the models perform well given that they are using only a single predictor. With this simple one-dimensional predictor space, the coarse tree now performs as well as the medium and fine trees.

16 Select the best model in the **Models** pane and view the residuals plot. On the **Regression Learner** tab, in the **Plots** section, click **Residuals** and select **Validation Data**. The residuals plot displays the difference between the predicted and true responses. To display the residuals as a line graph, in the **Style** section, choose **Lines**.

Under **X-axis**, select the variable to plot on the *x*-axis. Choose either the true response, predicted response, record number, or one of your predictors.



Usually a good model has residuals scattered roughly symmetrically around 0. If you can see any clear patterns in the residuals, it is likely that you can improve your model.

17 To learn about model settings, choose the best model in the **Models** pane and view the advanced settings. The nonoptimizable model options in the **Model Type** gallery are preset starting points, and you can change additional settings. On the **Regression Learner** tab, in the **Model Type** section, click **Advanced**. Compare the different regression tree models in the **Models** pane, and observe the differences in the Advanced Regression Tree Options dialog box. The **Minimum leaf size** setting controls the size of the tree leaves, and through that the size and depth of the regression tree.

To try to improve the model further, change the **Minimum leaf size** setting to 8, and then train a new model by clicking **Train**.

View the settings for the selected trained model in the **Current Model Summary** pane or in the Advanced Regression Tree Options dialog box.

To learn more about regression tree settings, see “Regression Trees” on page 24-16.

18 Export the selected model to the workspace. On the **Regression Learner** tab, in the **Export** section, click **Export Model**. In the Export Model dialog box, click **OK** to accept the default variable name `trainedModel`.

To see information about the results, look in the command window.

19 Use the exported model to make predictions on new data. For example, to make predictions for the `cartable` data in your workspace, enter:

```
yfit = trainedModel.predictFcn(cartable)
```

The output `yfit` contains the predicted response for each data point.

20 If you want to automate training the same model with new data or learn how to programmatically train regression models, you can generate code from the app. To generate code for the best trained model, on the **Regression Learner** tab, in the **Export** section, click **Generate Function**.

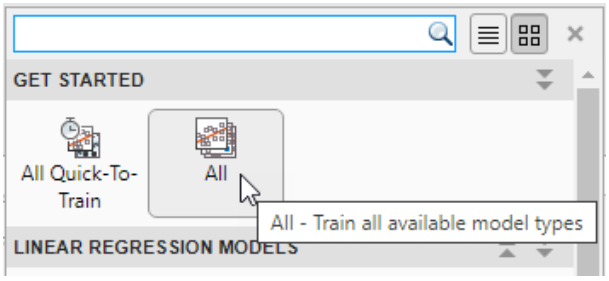
The app generates code from your model and displays the file in the MATLAB Editor. To learn more, see “Generate MATLAB Code to Train Model with New Data” on page 24-55.

Tip Use the same workflow as in this example to evaluate and compare the other regression model types you can train in Regression Learner.

Train all the nonoptimizable regression model presets available:

1 On the far right of the **Model Type** section, click the arrow to expand the list of regression models.

2 Click **All**  , and then click **Train**.



To learn about other regression model types, see “Train Regression Models in Regression Learner App” on page 24-2.

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54

Train Regression Neural Networks Using Regression Learner App

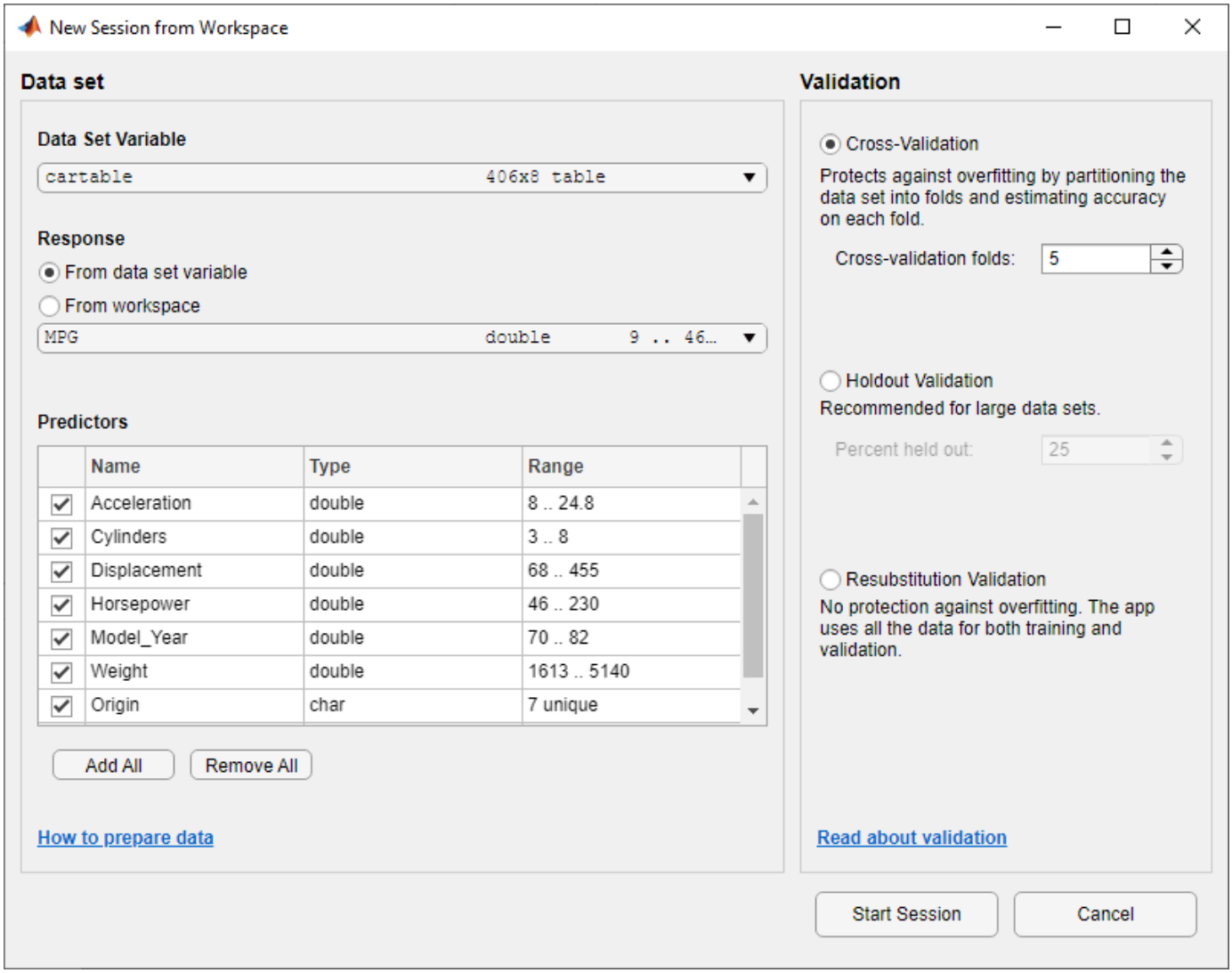
This example shows how to create and compare various regression neural network models using the Regression Learner app, and export trained models to the workspace to make predictions for new data.

1. In the MATLAB Command Window, load the `carbig` data set, and create a table containing the different variables.

   ```
   load carbig
   cartable = table(Acceleration,Cylinders,Displacement, ...
       Horsepower,Model_Year,Weight,Origin,MPG);
   ```

2. Click the **Apps** tab, and then click the **Show more** arrow on the right to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Regression Learner**.
3. On the **Regression Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.
4. In the New Session from Workspace dialog box, select the table `cartable` from the **Data Set Variable** list (if necessary).

   As shown in the dialog box, the app selects `MPG` as the response and the other variables as predictors. For this example, do not change the selections.



5 To accept the default validation scheme and continue, click **Start Session**. The default validation option is 5-fold cross-validation, to protect against overfitting.

   Regression Learner creates a plot of the response with the record number on the *x*-axis.

6 Use the response plot to investigate which variables are useful for predicting the response. To visualize the relation between different predictors and the response, select different variables in the **X** list under **X-axis**. Note which variables are correlated most clearly with the response.

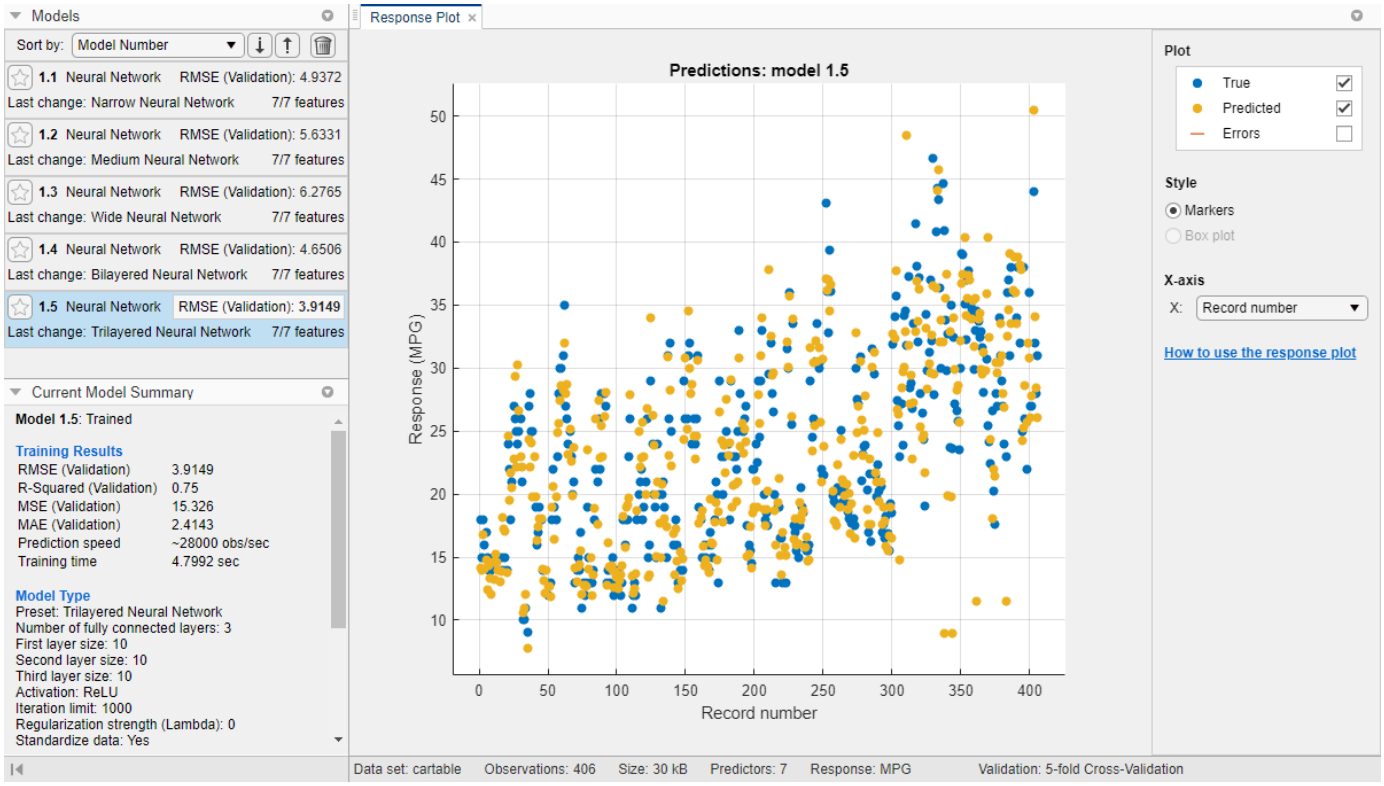
7 Create a selection of neural network models. On the **Regression Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Neural Networks** group, click **All Neural Networks**.

8 In the **Training** section, click **Train**. Regression Learner trains one of each neural network option in the gallery. In the **Models** pane, the app outlines the **RMSE (Validation)** (root mean squared error) of the best model.

   Tip If you have Parallel Computing Toolbox, you can train all the models (**All Neural Networks**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking

Train. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.

9 Select a model in the **Models** pane to view the results. Examine the response plot for the trained model. True responses are in blue, and predicted responses are in yellow.

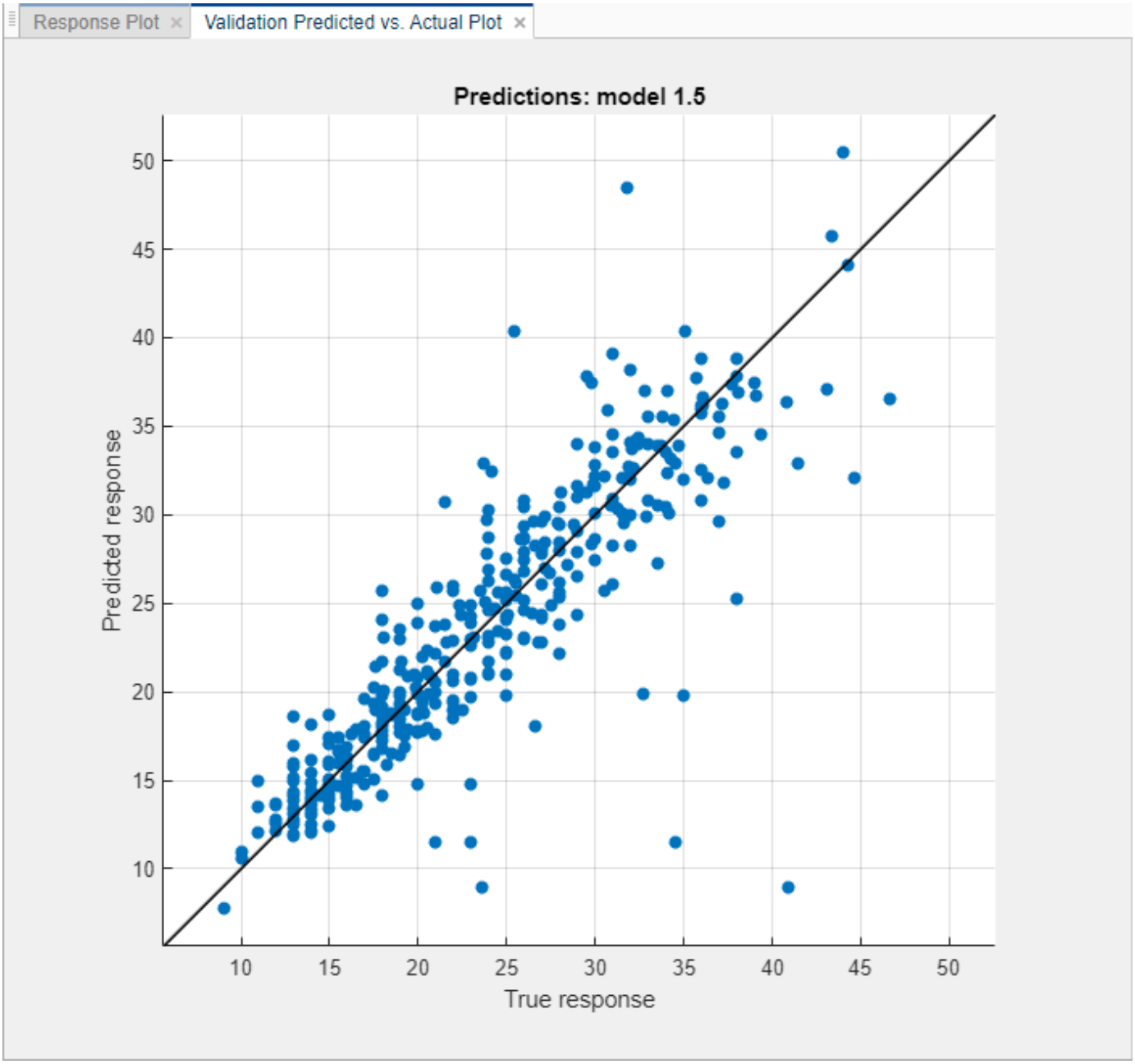


Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

10 Under **X-axis**, select `Horsepower` and examine the response plot. Both the true and predicted responses are now plotted. Show the prediction errors, drawn as vertical lines between the predicted and true responses, by selecting the **Errors** check box.

11 For more information on the currently selected model, consult the **Current Model Summary** pane. Check and compare additional model characteristics, such as R-squared (coefficient of determination), MAE (mean absolute error), and prediction speed. To learn more, see “View and Compare Model Statistics” on page 24-43. In the **Current Model Summary** pane, you can also find details on the currently selected model type, such as options used for training the model.

12 Plot the predicted response versus the true response. On the **Regression Learner** tab, in the **Plots** section, click **Predicted vs. Actual** and select **Validation Data**. Use this plot to determine how well the regression model makes predictions for different response values.



A perfect regression model has predicted responses equal to the true responses, so all the points lie on a diagonal line. The vertical distance from the line to any point is the error of the prediction for that point. A good model has small errors, so the predictions are scattered near the line. Typically, a good model has points scattered roughly symmetrically around the diagonal line. If you can see any clear patterns in the plot, you can most likely improve your model.

13 Select the other models in the **Models** pane and compare the predicted versus actual plots.

14 In the **Model Type** gallery, select **All Neural Networks** again. To try to improve the models, include different features. See if you can improve the models by removing features with low predictive power. On the **Regression Learner** tab, in the **Features** section, click **Feature Selection**.

In the Feature Selection dialog box, clear the check boxes for **Acceleration** and **Cylinders** to exclude them from the predictors.

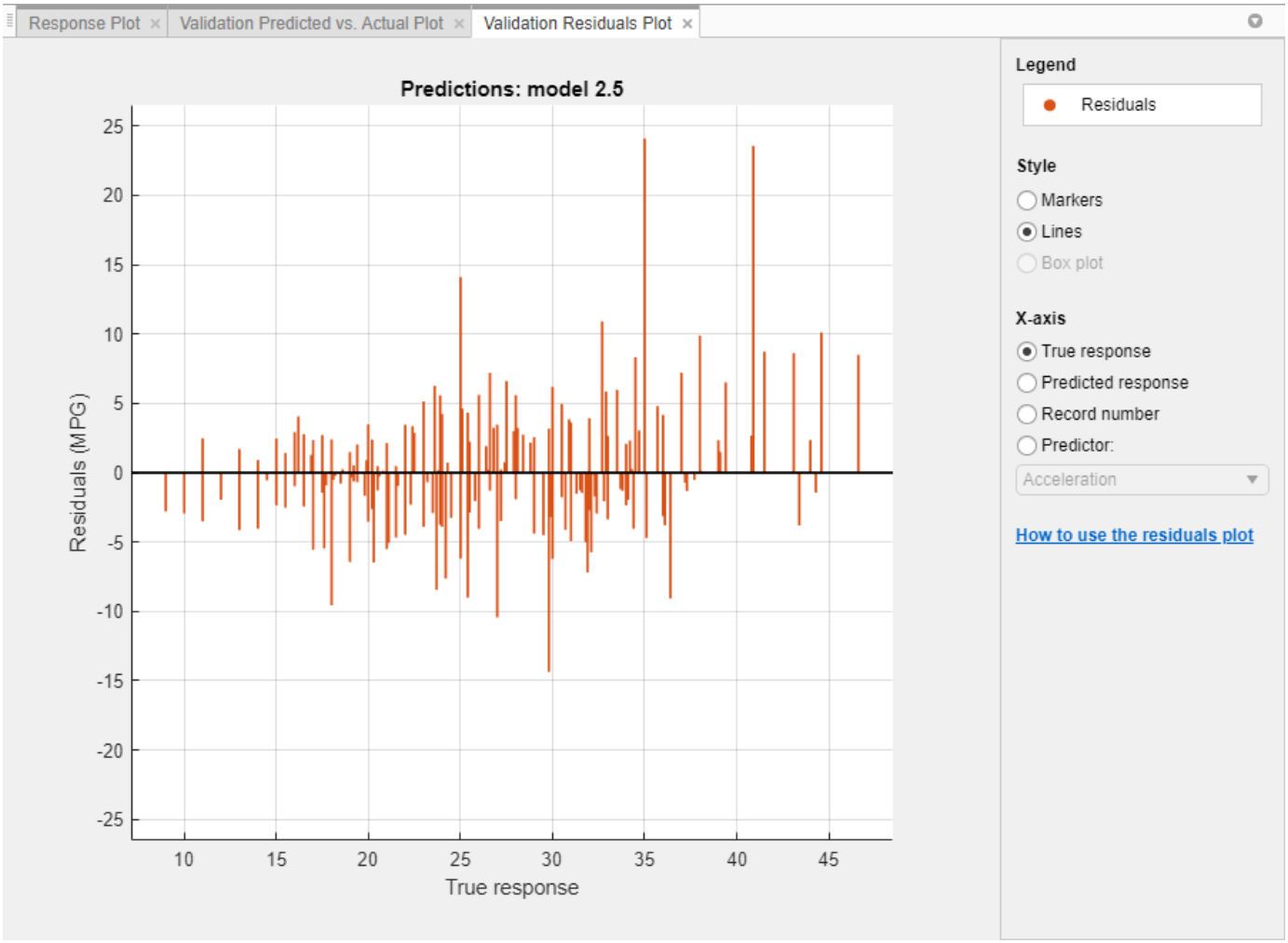
In the **Training** section, click **Train** to train the neural network models using the new predictor settings.

15 Observe the new models in the **Models** pane. These models are the same neural network models as before, but trained using only five of the seven predictors. For each model, the app displays how many predictors are used. To check which predictors are used, click a model in the **Models** pane and consult the **Feature Selection** section of the **Current Model Summary** pane.

The models with the two features removed perform comparably to the models with all predictors. The models predict no better using all the predictors compared to using only a subset of them. If data collection is expensive or difficult, you might prefer a model that performs satisfactorily without some predictors.

16 Select the best model in the **Models** pane and view the residuals plot. On the **Regression Learner** tab, in the **Plots** section, click **Residuals** and select **Validation Data**. The residuals plot displays the difference between the predicted and true responses. To display the residuals as a line graph, in the **Style** section, choose **Lines**.

Under **X-axis**, select the variable to plot on the *x*-axis. Choose the true response, predicted response, record number, or one of your predictors.



Typically, a good model has residuals scattered roughly symmetrically around 0. If you can see any clear patterns in the residuals, you can most likely improve your model.

17 You can try to further improve the best model in the **Models** pane by changing its advanced settings. On the **Regression Learner** tab, in the **Model Type** section, click **Advanced** and select **Advanced**. Try changing some of the settings, like the sizes of the fully connected layers or the regularization strength, and then train the new model by clicking **Train**.

To learn more about neural network model settings, see “Neural Networks” on page 24-24.

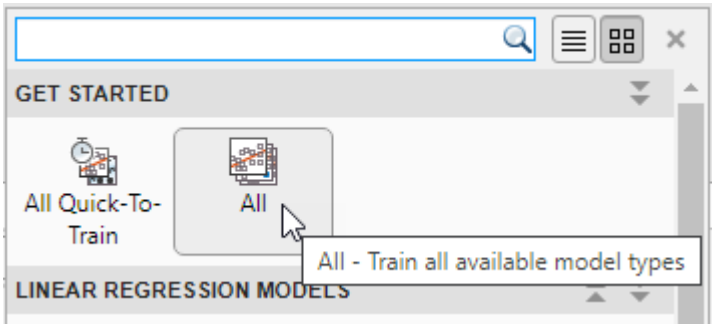
18 You can export a full or compact version of the trained model to the workspace. On the **Regression Learner** tab, in the **Export** section, click **Export Model** and select either **Export Model** or **Export Compact Model**. See “Export Regression Model to Predict New Data” on page 24-54.

19 To examine the code for training this model, click **Generate Function** in the **Export** section.

Tip Use the same workflow to evaluate and compare the other regression model types you can train in Regression Learner.

To train all the nonoptimizable regression model presets available for your data set:

1. On **Regression Learner** tab, in the **Model Type** section, click the arrow to open the gallery of regression models.
2. In the **Get Started** group, click **All**. Then, in the **Training** section, click **Train**.



To learn about other regression model types, see “Train Regression Models in Regression Learner App” on page 24-2.

See Also

Related Examples

- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Feature Selection and Feature Transformation Using Regression Learner App” on page 24-26
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54

Train Regression Model Using Hyperparameter Optimization in Regression Learner App

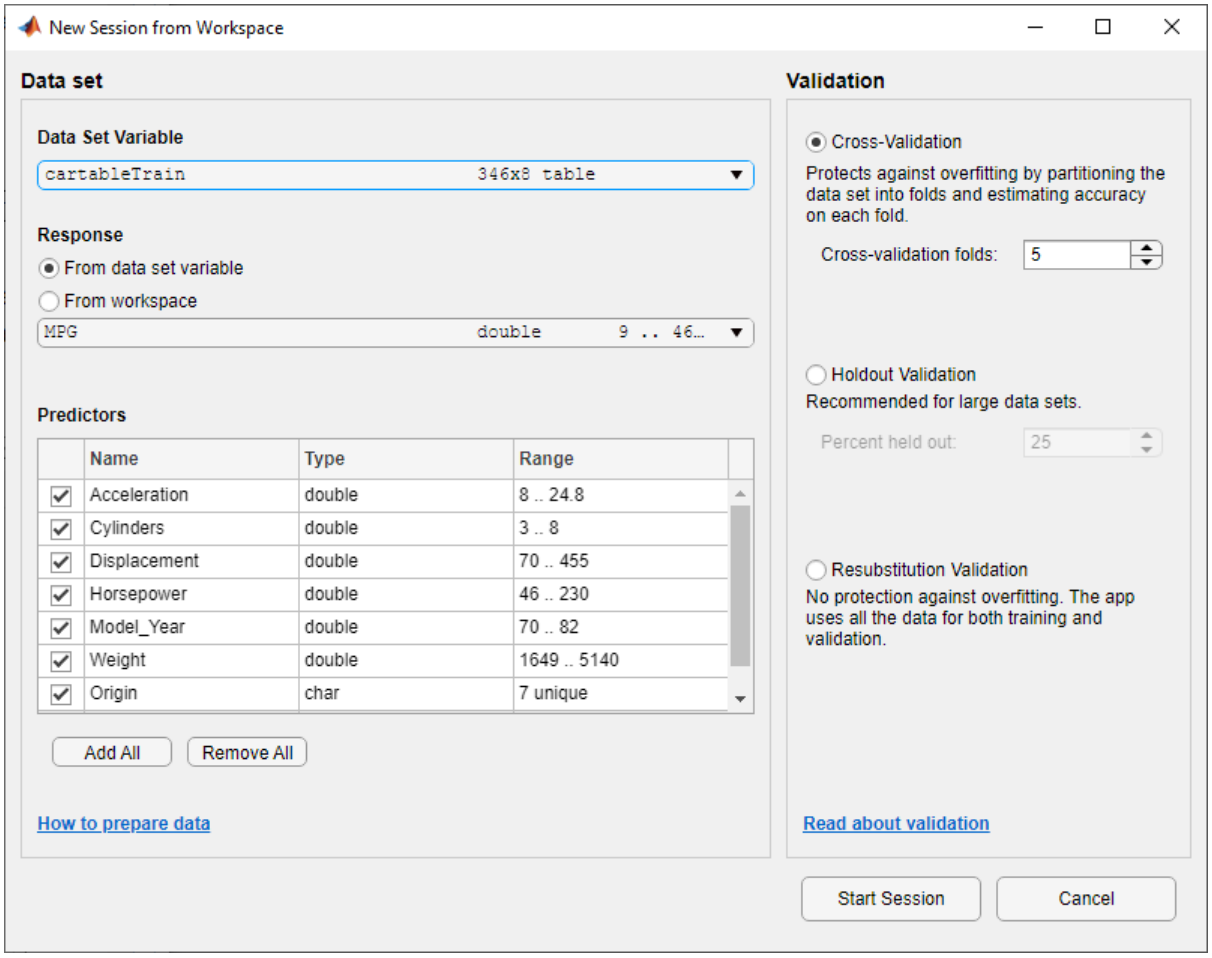
This example shows how to tune hyperparameters of a regression ensemble by using hyperparameter optimization in the Regression Learner app. Compare the test set performance of the trained optimizable ensemble to that of the best-performing preset ensemble model.

1. In the MATLAB Command Window, load the `carbig` data set, and create a table containing most of the variables. Separate the table into training and test sets.

   ```
   load carbig
   cartable = table(Acceleration,Cylinders,Displacement, ...
       Horsepower,Model_Year,Weight,Origin,MPG);

   rng('default') % For reproducibility of the data split
   n = length(MPG);
   partition = cvpartition(n,'Holdout',0.15);
   idxTrain = training(partition); % Indices for the training set
   cartableTrain = cartable(idxTrain,:);
   cartableTest = cartable(~idxTrain,:);
   ```

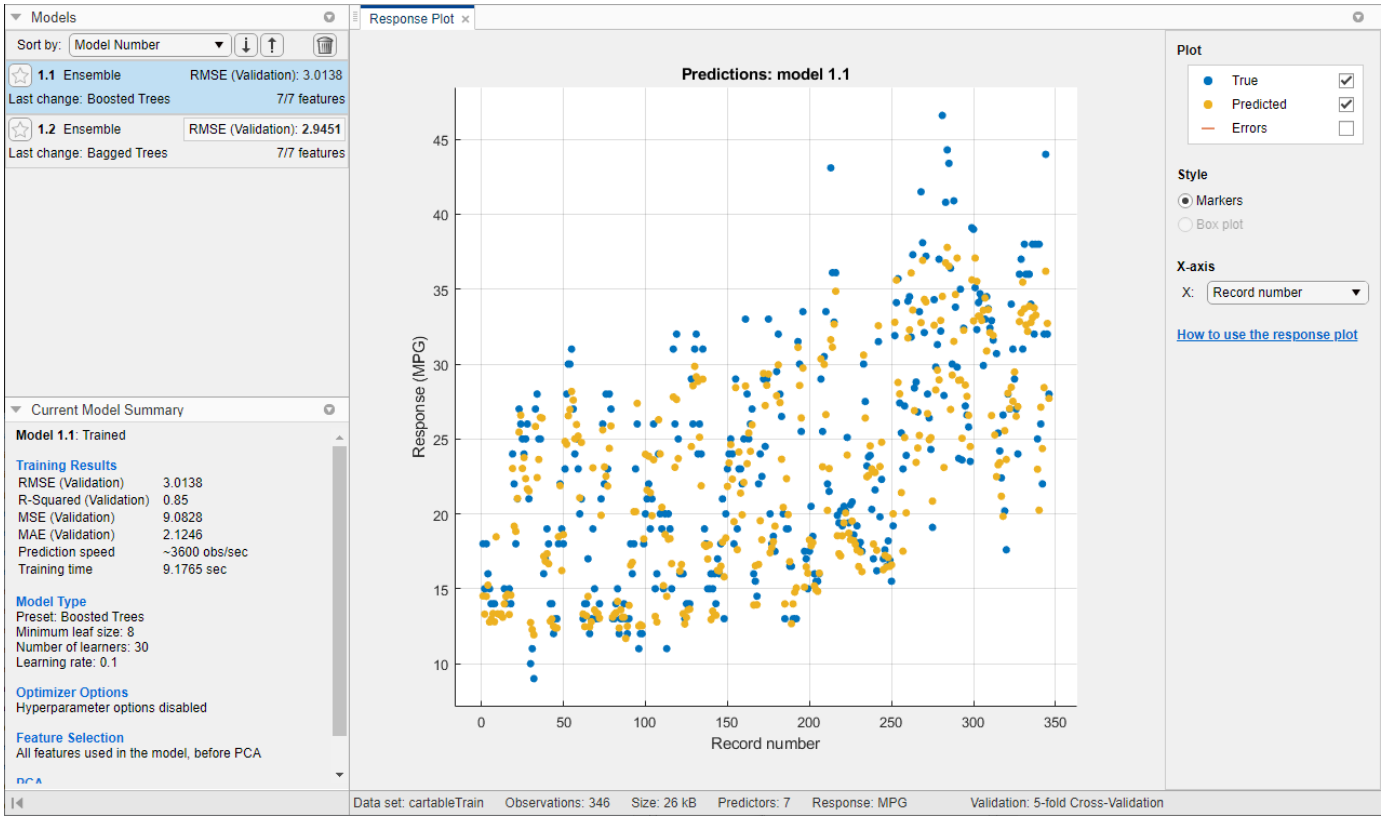
2. Open Regression Learner. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Regression Learner**.
3. On the **Regression Learner** tab, in the **File** section, select **New Session > From Workspace**.
4. In the New Session from Workspace dialog box, select the `cartableTrain` table from the **Data Set Variable** list.

   As shown in the dialog box, the app selects the response and predictor variables. The default response variable is `MPG`. The default validation option is 5-fold cross-validation, to protect against overfitting. For this example, do not change the default settings.



5. To accept the default options and continue, click **Start Session**.
6. Train all preset ensemble models. On the **Regression Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Ensembles of Trees** group, click **All Ensembles**. In the **Training** section, click **Train**. The app trains one of each ensemble model type and displays the models in the **Models** pane.

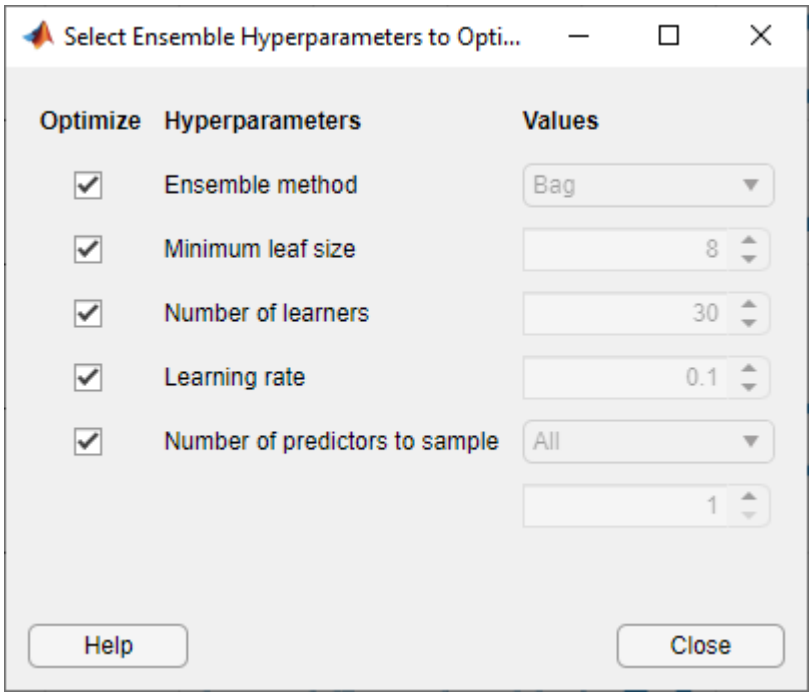
> **Tip** If you have Parallel Computing Toolbox, you can train all the ensemble models (**All Ensembles**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the ensemble models simultaneously.



The app displays a response plot of the car data. Blue points are true values, and yellow points are predicted values. The **Models** pane on the left shows the validation RMSE for each model.

Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

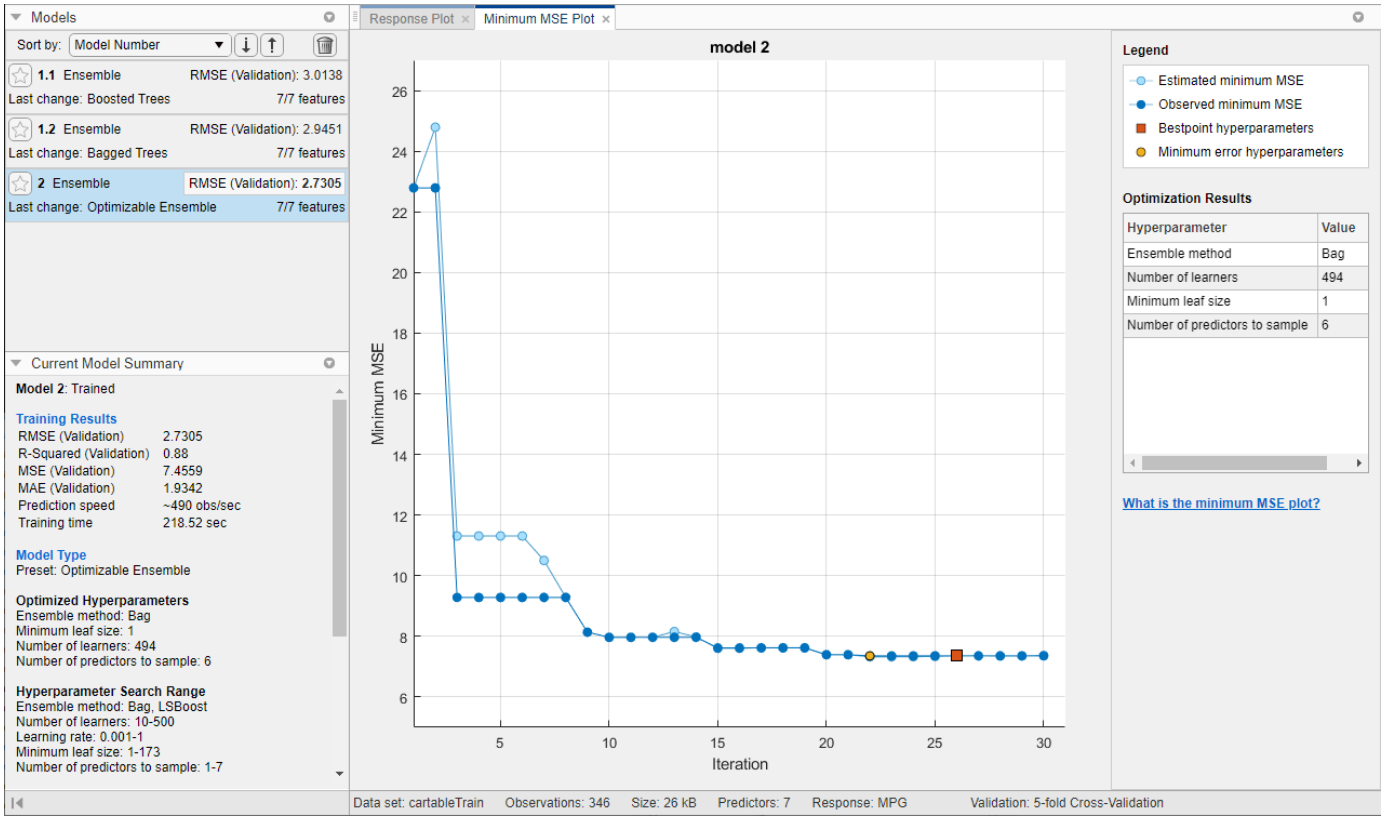
7. Select an optimizable ensemble model to train. On the **Regression Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Ensembles of Trees** group, click **Optimizable Ensemble**. The app disables the **Use Parallel** button when you select an optimizable model.
8. Select the model hyperparameters to optimize. In the **Model Type** section, select **Advanced > Advanced**. The app opens a dialog box in which you can select **Optimize** check boxes for the hyperparameters that you want to optimize. By default, all the check boxes are selected. For this example, accept the default selections, and close the dialog box.



9 In the **Training** section, click **Train**.

10 The app displays a **Minimum MSE Plot** as it runs the optimization process. At each iteration, the app tries a different combination of hyperparameter values and updates the plot with the minimum validation mean squared error (MSE) observed up to that iteration, indicated in dark blue. When the app completes the optimization process, it selects the set of optimized hyperparameters, indicated by a red square. For more information, see “Minimum MSE Plot” on page 24-37.

The app lists the optimized hyperparameters in both the **Optimization Results** section to the right of the plot and the **Optimized Hyperparameters** section of the **Current Model Summary** pane.



Note In general, the optimization results are not reproducible.

11 Compare the trained preset ensemble models to the trained optimizable model. In the **Models** pane, the app highlights the lowest **RMSE (Validation)** (validation root mean squared error) by outlining it in a box. In this example, the trained optimizable ensemble outperforms the two preset models.

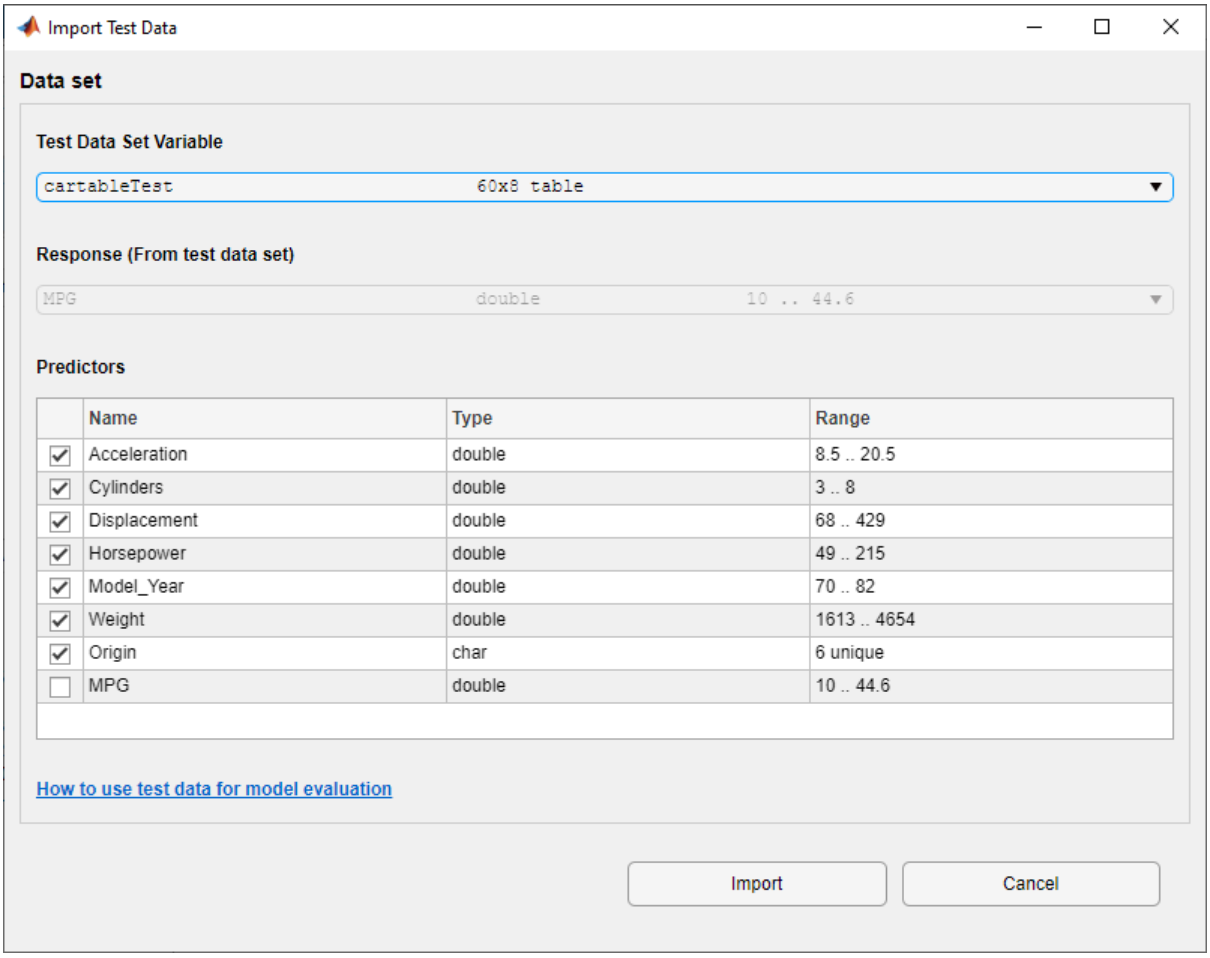
A trained optimizable model does not always have a lower RMSE than the trained preset models. If a trained optimizable model does not perform well, you can try to get better results by running the optimization for longer. In the **Model Type** section, select **Advanced > Optimizer Options**. In the dialog box, increase the **Iterations** value. For example, you can double-click the default value of `30` and enter a value of `60`.

12 Because hyperparameter tuning often leads to overfitted models, check the performance of the optimizable ensemble model on a test set and compare it to the performance of the best preset ensemble model. Begin by importing test data into the app.

On the **Regression Learner** tab, in the **Testing** section, select **Test Data > From Workspace**.

13 In the Import Test Data dialog box, select the `cartableTest` table from the **Test Data Set Variable** list.

As shown in the dialog box, the app identifies the response and predictor variables.



14 Click **Import**.

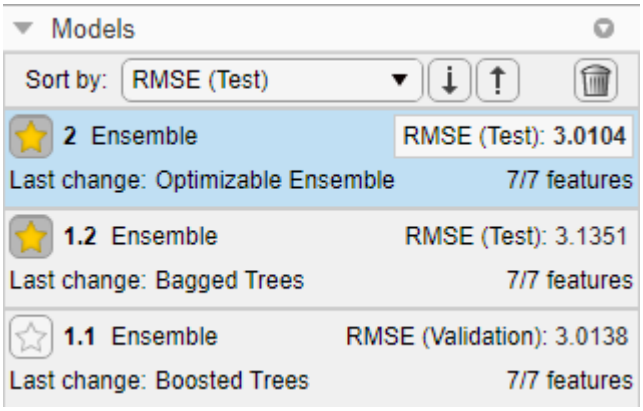
15 Compute the RMSE of the best preset model and the optimizable model on the `cartableTest` data.

First, in the **Models** pane, click the star icons next to the **Bagged Trees** model and the **Optimizable Ensemble** model.

16 For each model, select the model in the **Models** pane, and then select **Test All > Test Selected** in the **Testing** section. The app computes the test set performance of the model trained on the full data set, including training and validation data.

17 Sort the models based on the test set RMSE. In the **Models** pane, open the **Sort by** list and select `RMSE (Test)`.

In this example, the trained optimizable ensemble still outperforms the trained preset model on the test set data.



See Also

Related Examples

- “Hyperparameter Optimization in Regression Learner App” on page 24-30
- “Train Regression Models in Regression Learner App” on page 24-2
- “Select Data and Validation for Regression Problem” on page 24-8
- “Choose Regression Model Options” on page 24-12
- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54
- “Bayesian Optimization Workflow” on page 10-25

Check Model Performance Using Test Set in Regression Learner App

This example shows how to train multiple models in Regression Learner, and determine the best-performing models based on their validation metrics. Check the test metrics for the best-performing models trained on the full data set, including training and validation data.

1. In the MATLAB Command Window, load the `carbig` data set, and create a table containing most of the variables. Separate the table into training and test sets.

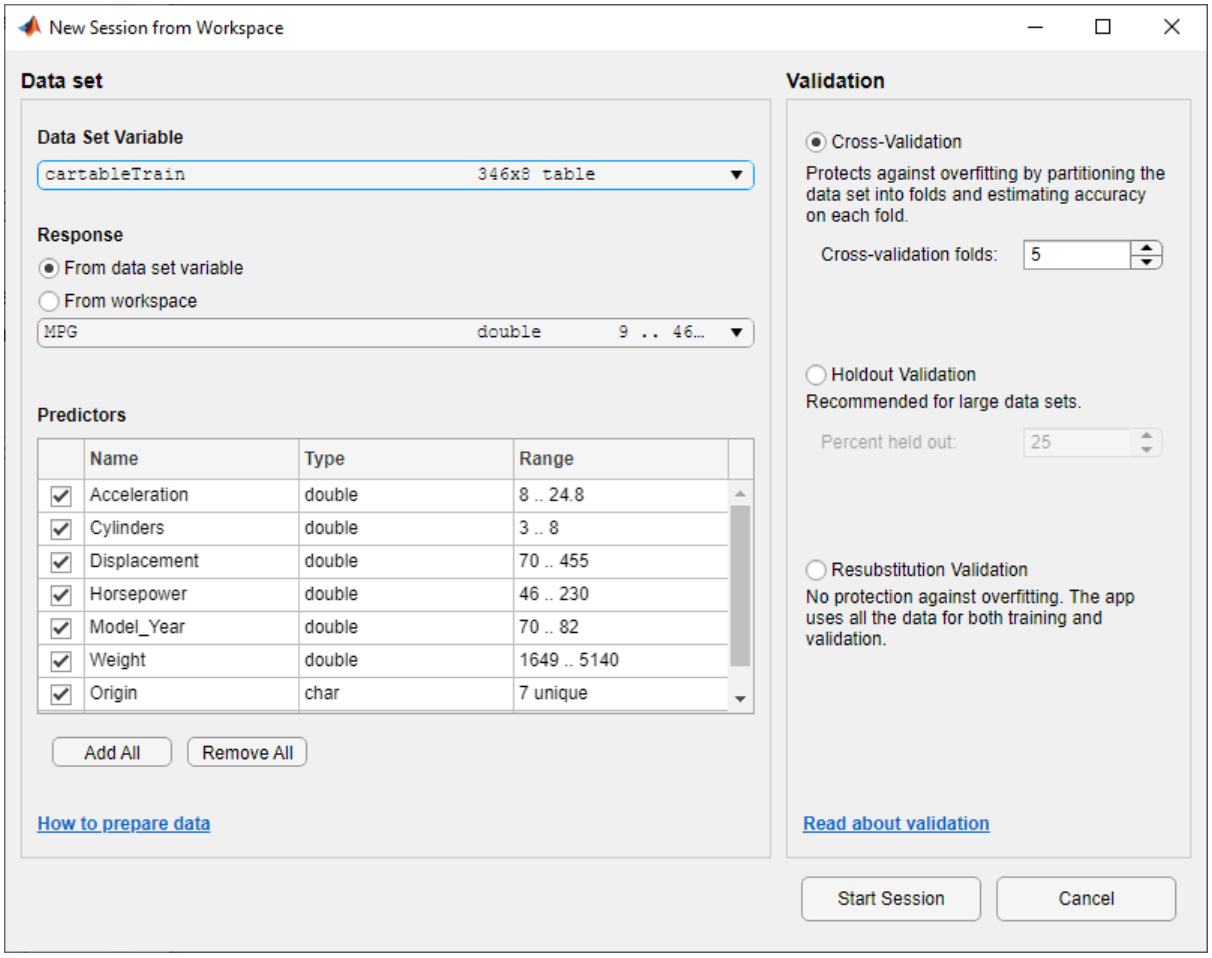
   ```
   load carbig
   cartable = table(Acceleration,Cylinders,Displacement, ...
       Horsepower,Model_Year,Weight,Origin,MPG);

   rng('default') % For reproducibility of the data split
   n = length(MPG);
   partition = cvpartition(n,'Holdout',0.15);
   idxTrain = training(partition); % Indices for the training set
   cartableTrain = cartable(idxTrain,:);
   cartableTest = cartable(~idxTrain,:);
   ```

2. Open Regression Learner. Click the **Apps** tab, and then click the arrow at the right of the **Apps** section to open the apps gallery. In the **Machine Learning and Deep Learning** group, click **Regression Learner**.

3. On the **Regression Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.

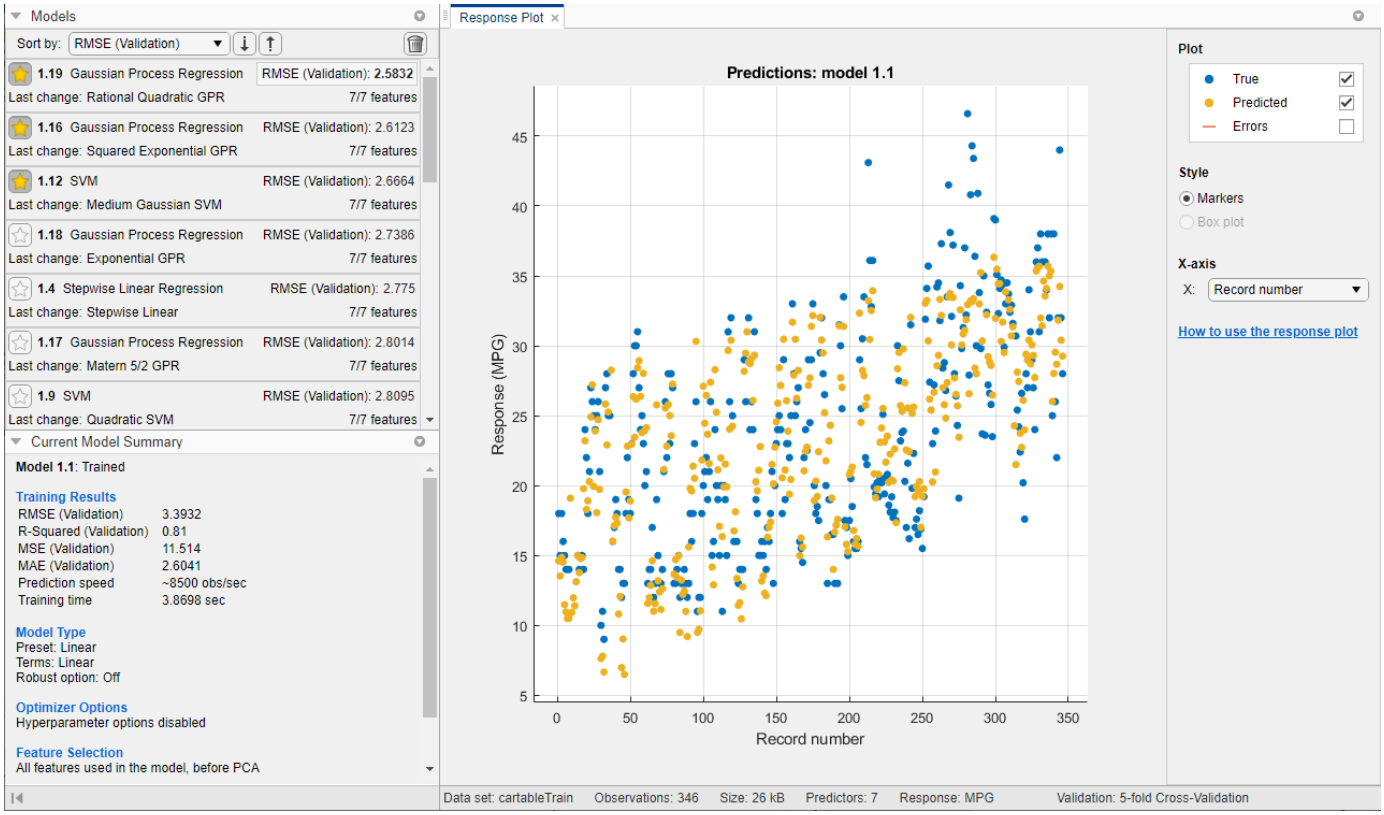
4. In the New Session from Workspace dialog box, select the `cartableTrain` table from the **Data Set Variable** list.

   As shown in the dialog box, the app selects the response and predictor variables. The default response variable is `MPG`. To protect against overfitting, the default validation option is 5-fold cross-validation. For this example, do not change the default settings.



5. To accept the default options and continue, click **Start Session**.
6. Train all preset models. On the **Regression Learner** tab, in the **Model Type** section, click the arrow to open the gallery. In the **Get Started** group, click **All**. In the **Training** section, click **Train**. The app trains one of each preset model type and displays the models in the **Models** pane.

   Tip If you have Parallel Computing Toolbox, you can train all the models (**All**) simultaneously by selecting the **Use Parallel** button in the **Training** section before clicking **Train**. After you click **Train**, the Opening Parallel Pool dialog box opens and remains open while the app opens a parallel pool of workers. During this time, you cannot interact with the software. After the pool opens, the app trains the models simultaneously.
7. Sort the trained models based on the validation root mean squared error (RMSE). In the **Models** pane, open the **Sort by** list and select `RMSE (Validation)`.
8. In the **Models** pane, click the star icons next to the three models with the lowest validation RMSE. The app highlights the lowest validation RMSE by outlining it in a box. In this example, the trained **Rational Quadratic GPR** model has the lowest validation RMSE.



The app displays a response plot of the car data. Blue points are true values, and yellow points are predicted values. The **Models** pane on the left shows the validation RMSE for each model.

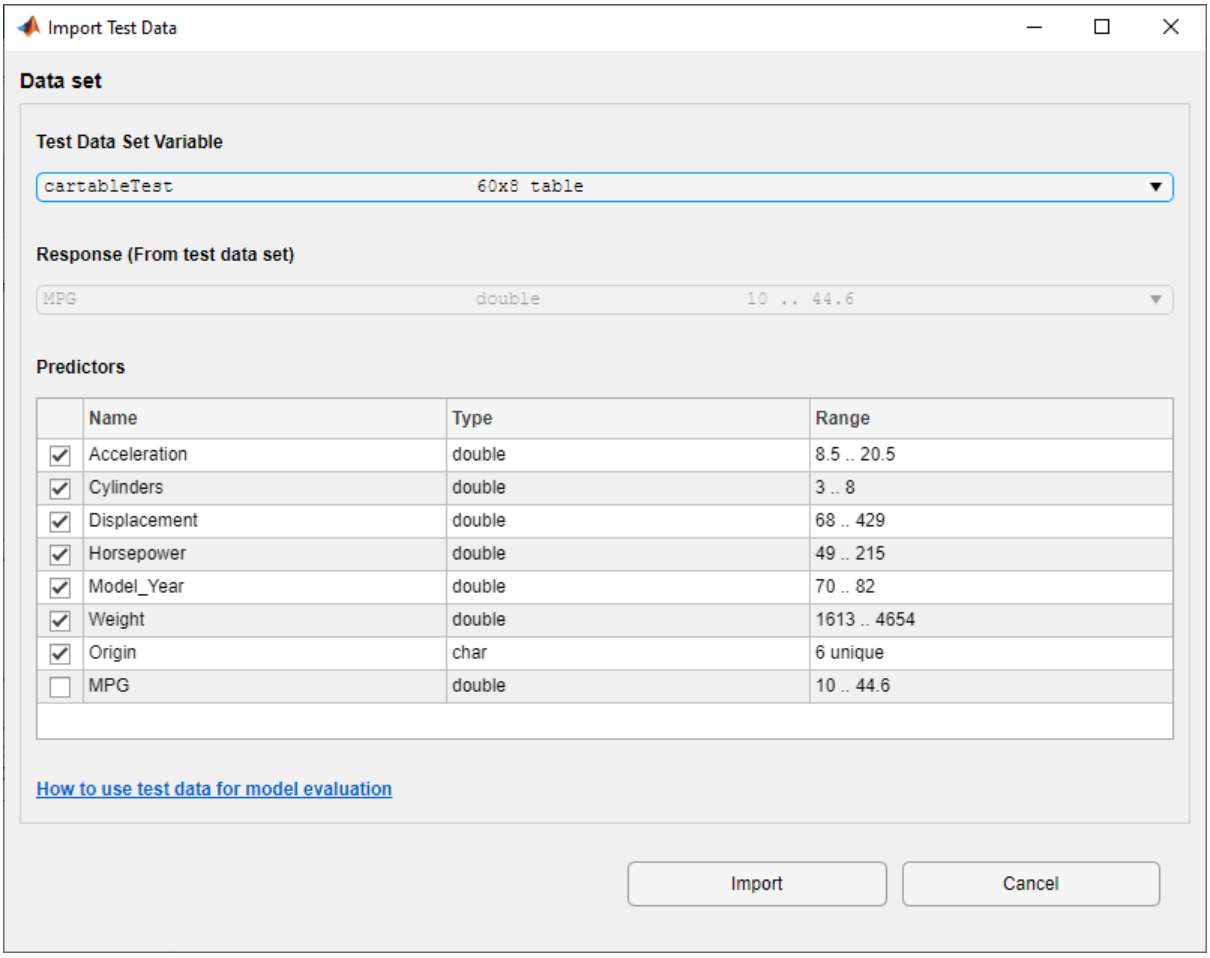
Note Validation introduces some randomness into the results. Your model validation results can vary from the results shown in this example.

9 Check the test set performance of the best-performing models. Begin by importing test data into the app.

On the **Regression Learner** tab, in the **Testing** section, click **Test Data** and select **From Workspace**.

10 In the Import Test Data dialog box, select the `cartableTest` table from the **Test Data Set Variable** list.

As shown in the dialog box, the app identifies the response and predictor variables.



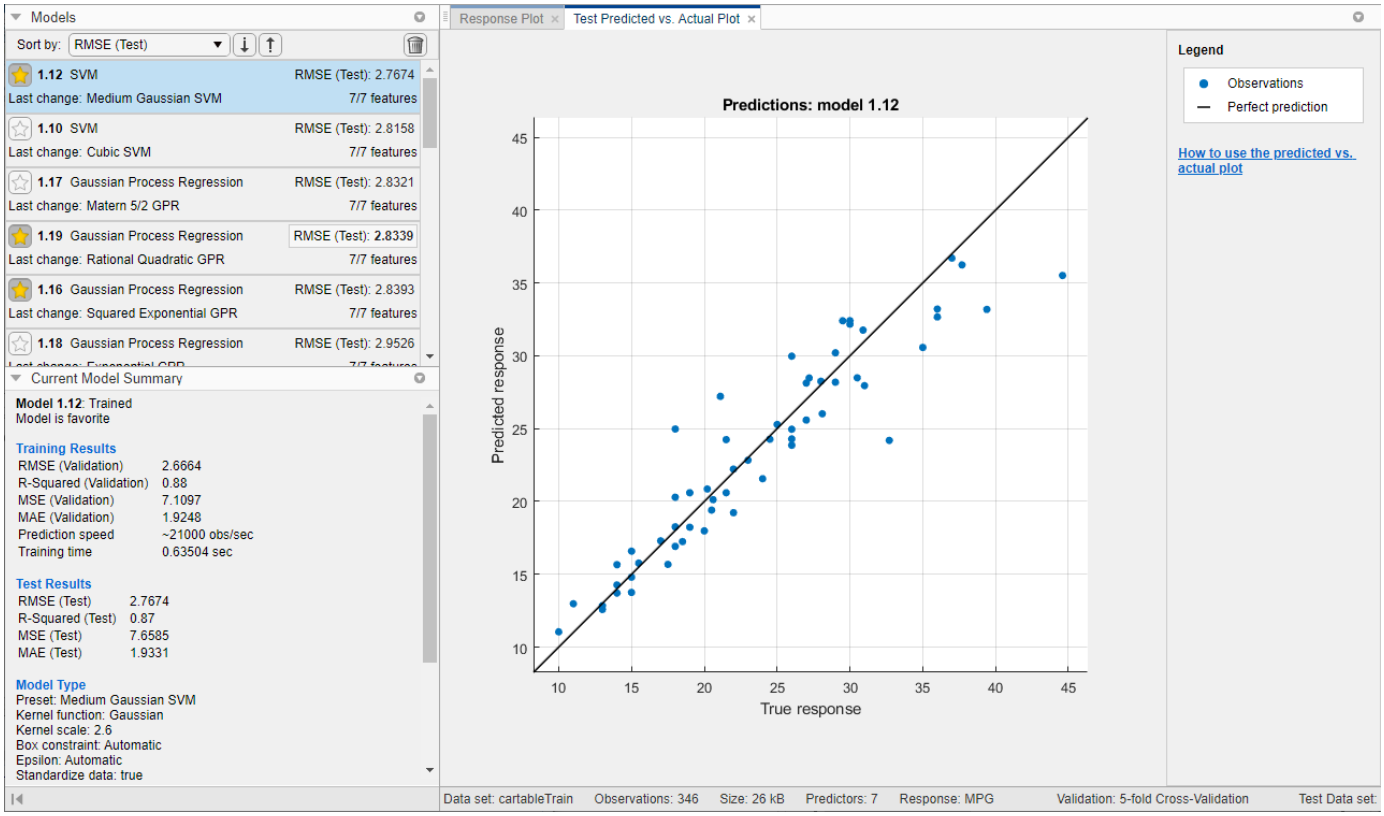
11 Click **Import**.

12 Compute the RMSE of the best preset models on the `cartableTest` data. For convenience, compute the test set RMSE for all models at once. On the **Regression Learner** tab, in the **Testing** section, click **Test All** and select **Test All**. The app computes the test set performance of the model trained on the full data set, including training and validation data.

13 Sort the models based on the test set RMSE. In the **Models** pane, open the **Sort by** list and select `RMSE (Test)`. The app still outlines the metric for the model with the lowest validation RMSE, despite displaying the test RMSE.

14 Visually check the test set performance of the models. On the **Regression Learner** tab, in the **Plots** section, click **Predicted vs. Actual** and select **Test Data**. You can toggle between models to compare their performance.

In this example, the trained **Medium Gaussian SVM** performs better on the test set data than the other two starred models.



15 Compare the validation and test RMSE for the trained **Medium Gaussian SVM** model. In the **Current Model Summary** pane, compare the **RMSE (Validation)** value under **Training Results** to the **RMSE (Test)** value under **Test Results**. In this example, the two values are close, which indicates that the validation RMSE is a good estimate of the test RMSE for this model.

See Also

Related Examples

- “Assess Model Performance in Regression Learner” on page 24-42
- “Export Regression Model to Predict New Data” on page 24-54
- “Train Regression Model Using Hyperparameter Optimization in Regression Learner App” on page 24-76

Support Vector Machines

Understanding Support Vector Machine Regression

In this section...

Mathematical Formulation of SVM Regression

Overview

Support vector machine (SVM) analysis is a popular machine learning tool for classification and regression, first identified by Vladimir Vapnik and his colleagues in 1992[5]. SVM regression is considered a nonparametric technique because it relies on kernel functions.

Statistics and Machine Learning Toolbox implements linear epsilon-insensitive SVM (ε-SVM) regression, which is also known as $L1$ loss. In ε-SVM regression, the set of training data includes predictor variables and observed response values. The goal is to find a function $f(x)$ that deviates from y_n by a value no greater than ε for each training point x, and at the same time is as flat as possible.

Linear SVM Regression: Primal Formula

Suppose we have a set of training data where x_n is a multivariate set of N observations with observed response values y_n.

To find the linear function

$$f(x) = x'\beta + b,$$

and ensure that it is as flat as possible, find $f(x)$ with the minimal norm value ($\beta'\beta$). This is formulated as a convex optimization problem to minimize

$$J(\beta) = \frac{1}{2}\beta'\beta$$

subject to all residuals having a value less than ε; or, in equation form:

$$\forall n: |y_n - (x_n'\beta + b)| \le \varepsilon \ .$$

It is possible that no such function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, introduce slack variables ξ_n and ξ^*_n for each point. This approach is similar to the “soft margin” concept in SVM classification, because the slack variables allow regression errors to exist up to the value of ξ_n and ξ^*_n, yet still satisfy the required conditions.

Including slack variables leads to the objective function, also known as the primal formula[5]:

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{n=1}^{N}(\xi_n + \xi_n^*),$$

subject to:

$$\forall n: y_n - (x_n'\beta + b) \le \varepsilon + \xi_n$$
$$\forall n: (x_n'\beta + b) - y_n \le \varepsilon + \xi_n^*$$
$$\forall n: \xi_n^* \ge 0$$
$$\forall n: \xi_n \ge 0 \ .$$

The constant C is the box constraint, a positive numeric value that controls the penalty imposed on observations that lie outside the epsilon margin (ε) and helps to prevent overfitting (regularization). This value determines the trade-off between the flatness of $f(x)$ and the amount up to which deviations larger than ε are tolerated.

The linear ε-insensitive loss function ignores errors that are within ε distance of the observed value by treating them as equal to zero. The loss is measured based on the distance between observed value y and the ε boundary. This is formally described by

$$L_\varepsilon = \begin{cases} 0 & \text{if } |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases}$$

Linear SVM Regression: Dual Formula

The optimization problem previously described is computationally simpler to solve in its Lagrange dual formulation. The solution to the dual problem provides a lower bound to the solution of the primal (minimization) problem. The optimal values of the primal and dual problems need not be equal, and the difference is called the “duality gap.” But when the problem is convex and satisfies a constraint qualification condition, the value of the optimal solution to the primal problem is given by the solution of the dual problem.

To obtain the dual formula, construct a Lagrangian function from the primal function by introducing nonnegative multipliers α_n and α^*_n for each observation x_n. This leads to the dual formula, where we minimize

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)x_i'x_j + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i)$$

subject to the constraints

$$\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0$$
$$\forall n: 0 \le \alpha_n \le C$$
$$\forall n: 0 \le \alpha_n^* \le C \ .$$

The β parameter can be completely described as a linear combination of the training observations using the equation

$$\beta = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)x_n \ .$$

The function used to predict new values depends only on the support vectors:

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)(x_n'x) + b \ . \qquad (25\text{-}1)$$

The Karush-Kuhn-Tucker (KKT) complementarity conditions are optimization constraints required to obtain optimal solutions. For linear SVM regression, these conditions are

$$\forall n: \alpha_n(\varepsilon + \xi_n - y_n + x_n'\beta + b) = 0$$
$$\forall n: \alpha_n^*(\varepsilon + \xi_n^* + y_n - x_n'\beta - b) = 0$$
$$\forall n: \xi_n(C - \alpha_n) = 0$$
$$\forall n: \xi_n^*(C - \alpha_n^*) = 0 \ .$$

These conditions indicate that all observations strictly inside the epsilon tube have Lagrange multipliers $\alpha_n = 0$ and $\alpha_n^* = 0$. If either α_n or α_n^* is not zero, then the corresponding observation is called a *support vector*.

The property Alpha of a trained SVM model stores the difference between two Lagrange multipliers of support vectors, $\alpha_n - \alpha_n^*$. The properties SupportVectors and Bias store x_n and b, respectively.

Nonlinear SVM Regression: Primal Formula

Some regression problems cannot adequately be described using a linear model. In such a case, the Lagrange dual formulation allows the previously-described technique to be extended to nonlinear functions.

Obtain a nonlinear SVM regression model by replacing the dot product $x_1'x_2$ with a nonlinear kernel function $G(x_1,x_2) = <\varphi(x_1),\varphi(x_2)>$, where $\varphi(x)$ is a transformation that maps x to a high-dimensional space. Statistics and Machine Learning Toolbox provides the following built-in positive semidefinite kernel functions.

Kernel Name	Kernel Function
Linear (dot product)	$G(x_j, x_k) = x_j'x_k$
Gaussian	$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right)$
Polynomial	$G(x_j, x_k) = (1 + x_j'x_k)^q$, where q is in the set {2,3,...}.

The *Gram matrix* is an n-by-n matrix that contains elements $g_{i,j} = G(x_i,x_j)$. Each element $g_{i,j}$ is equal to the inner product of the predictors as transformed by φ. However, we do not need to know φ, because we can use the kernel function to generate Gram matrix directly. Using this method, nonlinear SVM finds the optimal function $f(x)$ in the transformed predictor space.

Nonlinear SVM Regression: Dual Formula

The dual formula for nonlinear SVM regression replaces the inner product of the predictors ($x_i'x_j$) with the corresponding element of the Gram matrix ($g_{i,j}$).

Nonlinear SVM regression finds the coefficients that minimize

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)G\left(x_i, x_j\right) + \varepsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*)$$

subject to

$$\sum_{n=1}^{N}(\alpha_n - \alpha_n^*) = 0$$
$$\forall n: 0 \le \alpha_n \le C$$
$$\forall n: 0 \le \alpha_n^* \le C \ .$$

The function used to predict new values is equal to

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)G(x_n, x) + b \ . \tag{25-2}$$

The KKT complementarity conditions are

$$\forall n: \alpha_n(\varepsilon + \xi_n - y_n + f(x_n)) = 0$$
$$\forall n: \alpha_n^*(\varepsilon + \xi_n^* + y_n - f(x_n)) = 0$$
$$\forall n: \xi_n(C - \alpha_n) = 0$$
$$\forall n: \xi_n^*(C - \alpha_n^*) = 0 \ .$$

Solving the SVM Regression Optimization Problem

Solver Algorithms

The minimization problem can be expressed in standard quadratic programming form and solved using common quadratic programming techniques. However, it can be computationally expensive to use quadratic programming algorithms, especially since the Gram matrix may be too large to be stored in memory. Using a decomposition method instead can speed up the computation and avoid running out of memory.

Decomposition methods (also called *chunking and working set methods*) separate all observations into two disjoint sets: the working set and the remaining set. A decomposition method modifies only the elements in the working set in each iteration. Therefore, only some columns of the Gram matrix are needed in each iteration, which reduces the amount of storage needed for each iteration.

Sequential minimal optimization (SMO) is the most popular approach for solving SVM problems[4]. SMO performs a series of two-point optimizations. In each iteration, a working set of two points are chosen based on a selection rule that uses second-order information. Then the Lagrange multipliers for this working set are solved analytically using the approach described in [2] and [1].

In SVM regression, the gradient vector ∇L for the active set is updated after each iteration. The decomposed equation for the gradient vector is

$$(\nabla L)_n = \begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)G(x_i, x_n) + \varepsilon - y_n \ , & n \le N \\ -\sum_{i=1}^{N}(\alpha_i - \alpha_i^*)G(x_i, x_n) + \varepsilon + y_n \ , & n > N \end{cases} .$$

Iterative single data algorithm (ISDA) updates one Lagrange multiplier with each iteration[3]. ISDA is often conducted without the bias term b by adding a small positive constant a to the kernel function. Dropping b drops the sum constraint

$$\sum_{n=1}^{N}(\alpha_i - \alpha^*) = 0$$

in the dual equation. This allows us to update one Lagrange multiplier in each iteration, which makes it easier than SMO to remove outliers. ISDA selects the worst KKT violator among all the α_n and α_n^* values as the working set to be updated.

Convergence Criteria

Each of these solver algorithms iteratively computes until the specified convergence criterion is met. There are several options for convergence criteria:

- *Feasibility gap* — The feasibility gap is expressed as

  $$\Delta = \frac{J(\beta) + L(\alpha)}{J(\beta) + 1},$$

  where $J(\beta)$ is the primal objective and $L(\alpha)$ is the dual objective. After each iteration, the software evaluates the feasibility gap. If the feasibility gap is less than the value specified by `GapTolerance`, then the algorithm met the convergence criterion and the software returns a solution.
- *Gradient difference* — After each iteration, the software evaluates the gradient vector, ∇L. If the difference in gradient vector values for the current iteration and the previous iteration is less than the value specified by `DeltaGradientTolerance`, then the algorithm met the convergence criterion and the software returns a solution.
- *Largest KKT violation* — After each iteration, the software evaluates the KKT violation for all the α_n and α_n^* values. If the largest violation is less than the value specified by `KKTTolerance`, then the algorithm met the convergence criterion and the software returns a solution.

References

[1] Fan, R.E. , P.H. Chen, and C.J. Lin. "A Study on SMO-Type Decomposition Methods for Support Vector Machines." *IEEE Transactions on Neural Networks,* Vol. 17:893–908, 2006.

[2] Fan, R.E. , P.H. Chen, and C.J. Lin. "Working Set Selection Using Second Order Information for Training Support Vector Machines." *The Journal of Machine Learning Research,* Vol. 6:1871–1918, 2005.

[3] Huang, T.M., V. Kecman, and I. Kopriva. *Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-Supervised, and Unsupervised Learning*. Springer, New York, 2006.

[4] Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.* Technical Report MSR-TR-98–14, 1999.

[5] Vapnik, V. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

See Also

`RegressionSVM` | `fitrsvm` | `predict` | `resubPredict`

Related Examples

- “Train Linear Support Vector Machine Regression Model” on page 33-2339
- “Train Support Vector Machine Regression Model” on page 33-2341
- “Cross-Validate SVM Regression Model” on page 33-2342
- “Optimize SVM Regression” on page 33-2344

Incremental Learning

Incremental Learning Overview

In this section...

What Is Incremental Learning?

Incremental learning, or online learning, is a branch of machine learning that involves processing incoming data from a data stream—continuously and in real time—possibly given little to no knowledge of the distribution of the predictor variables, sample size, aspects of the prediction or objective function (including adequate tuning parameter values), and whether the observations have labels.

Incremental learning algorithms are flexible, efficient, and adaptive. The following characteristics distinguish incremental learning from traditional machine learning:

- An incremental model is fit to data quickly and efficiently, which means it can adapt, in real time, to changes (or drifts) in the data distribution.
- Because observation labels can be missing when corresponding predictor data is available, the algorithm must be able to generate predictions from the latest version of the model quickly, and defer training the model.
- Little information might be known about the population before incremental learning starts. Therefore, the algorithm can be run with a cold start. For example, for classification problems, the class names might not be known until after the model processes observations. When enough information is known before learning begins (for example, you have good estimates of linear model coefficients), you can specify such information to provide the model with a warm start.
- Because observations can arrive in a stream, the sample size is likely unknown and possibly large, which makes data storage inefficient or impossible. Therefore, the algorithm must process observations when they are available and before the system discards them. This incremental learning characteristic makes hyperparameter tuning difficult or impossible.

In traditional machine learning, a batch of labeled data is available to perform cross-validation to estimate the generalization error and tune hyperparameters, infer the predictor variable distribution, and fit the model. However, the resulting model must be retrained from the beginning if underlying distributions drift or the model degrades. Although performing cross-validation to tune hyperparameters is difficult in an incremental learning environment, incremental learning methods are flexible because they can adapt to distribution drift in real time, with predictive accuracy approaching that of a traditionally trained model as the model trains more data.

Suppose an incremental model is prepared to generate predictions and have its predictive performance measured. Given incoming chunks of observations, an incremental learning scheme processes data in real time and in any of the following ways, but usually in the specified order:

1. **Evaluate model**: Track the predictive performance of the model when true labels are available, either on the incoming data only, over a sliding window of observations, or over the entire history of the model used for incremental learning.
2. **Detect drift**: Check for structural breaks or distribution drift. For example, determine whether the distribution of any predictor variable has sufficiently changed.

3. **Train model**: Update the model by training it on the incoming observations, when true labels are available or when the current model has sufficiently degraded.
4. **Generate predictions**: Predict labels from the latest model.

This procedure is a special case of incremental learning, in which all incoming chunks are treated as test (holdout) sets. The procedure is called interleaved test-then-train or prequential evaluation [1].

If insufficient information exists for an incremental model to generate predictions, or you do not want to track the predictive performance of the model because it has not been trained enough, you can include an optional initial step to find adequate values for hyperparameters, for models that support one (estimation period), or an initial training period before model evaluation (metrics warm-up period).

As an example of an incremental learning problem, consider a smart thermostat that automatically sets a temperature given the ambient temperature, relative humidity, time of day, and other measurements, and can learn the user's indoor temperature preferences. Suppose the manufacturer prepared the device by embedding a known model that describes the average person's preferences given the measurements. After installation, the device collects data every minute, and adjusts the temperature to its presets. The thermostat adjusts the embedded model, or retrains itself, based on the user's actions or inactions with the device. This cycle can continue indefinitely. If the thermostat has limited disk space to store historical data, it needs to retrain itself in real time. If the manufacturer did not prepare the device with a known model, the device retrains itself more often.

Incremental Learning with MATLAB

Statistics and Machine Learning Toolbox functionalities enable you to implement incremental learning for classification or regression. Like other Statistics and Machine Learning Toolbox machine learning functionalities, the entry point into incremental learning is an incremental learning object, which you pass to functions with data to implement incremental learning. Unlike other machine learning functions, data is not required to create an incremental learning object. However, the incremental learning object specifies how to process incoming data, such as when to fit the model, measure performance metrics, or perform both actions, in addition to the parametric form of the model and problem-specific options.

Incremental Learning Model Objects

This table contains the available entry-point model objects for incremental learning with their supported machine learning objective, model type, and any information required upon creation.

Model Object	Objective	Model Type	Required Information
`incrementalClassificationLinear`	Binary classification	Linear SVM and logistic regression	None
`incrementalClassificationNaiveBayes`	Mutliclass classification	Naive Bayes with normal predictor conditional distributions	Maximum number of classes expected in the data during incremental learning or names of all expected classes
`incrementalRegressionLinear`	Regression	Linear	None

Properties of a incremental learning model object specify:

- Data characteristics, such as the number of predictor variables `NumPredictors` and their first and second moments.
- Model characteristics, such as, for linear models, the learner type `Learner`, linear coefficients `Beta`, and intercept `Bias`
- Training options, such as, for linear models, the objective solver `Solver` and solver-specific hyperparameters such as the ridge penalty `Lambda` for standard and average stochastic gradient descent (SGD and ASGD)
- Model performance evaluation characteristics and options, such as whether the model is warm `IsWarm`, which performance metrics to track `Metrics`, and the latest values of the performance metrics

Unlike when working with other machine learning model objects, you can create either model by directly calling the object and specifying property values of options using name-value arguments; you do not need to fit a model to data to create one. This feature is convenient when you have little information about the data or model before training it. Depending on your specifications, the software can enforce estimation and metrics warm-up periods, during which incremental fitting functions infer data characteristics and then train the model for performance evaluation. By default, for linear models, the software solves the objective function using the adaptive scale-invariant solver, which does not require tuning and is insensitive to the predictor variable scales [2].

Alternatively, you can convert a traditionally trained model to either model by using the `incrementalLearner` function. Convertible models include support vector machines (SVM) for binary classification and regression, naive Bayes classification, and linear regression models. For example, `incrementalLearner` converts a trained linear classification model of type `ClassificationLinear` to an `incrementalClassificationLinear` object. By default, the software considers converted models to be prepared for all aspects of incremental learning (converted models are warm). `incrementalLearner` carries over data characteristics (such as class names), fitted parameters, and options available for incremental learning from the traditionally trained model being converted. For example:

- For naive Bayes classification, `incrementalLearner` carries over all class names in the data expected during incremental learning, and the fitted moments of the conditional predictor distributions (`DistributionParameters`).
- For linear models, if the objective solver of the traditionally trained model is SGD, `incrementalLearner` sets the incremental learning solver to SGD.

Incremental Learning Functions

The incremental learning model object specifies all aspects of the incremental learning algorithm, from training and model evaluation preparation through training and model evaluation. To implement incremental learning, you pass the configured incremental learning model to an incremental fitting function or model evaluation function. Statistics and Machine Learning Toolbox incremental learning functions offer two workflows that are well suited for prequential learning. For simplicity, the following workflow descriptions assume that the model is prepared to evaluate the model performance (in other words, the model is warm).

- **Flexible workflow** — When a data chunk is available:

  1. Compute cumulative and window model performance metrics by passing the data and current model to the `updateMetrics` function. The data is treated as test (holdout) data because the model has not been trained on it yet. `updateMetrics` overwrites the model performance stored in the model with the new values. For linear models, see `updateMetrics` and, for naive Bayes classification models, see `updateMetrics`.

2. Optionally detect distribution drift or whether the model has degraded.
3. Train the model by passing the incoming data chunk and current model to the `fit` function. The `fit` function uses the specified solver to fit the model to the incoming data chunk, and overwrites the current coefficients and bias with the new estimates. For linear models, see `fit` and, for naive Bayes classification models, see `fit`.

The flexible workflow enables you to perform custom model and data quality assessments before deciding whether to train the model. All steps are optional, but call `updateMetrics` before `fit` when you plan to call both functions.

- **Succinct workflow** — When a data chunk is available, supply the incoming chunk and a configured incremental model to the `updateMetricsAndFit` function. `updateMetricsAndFit` calls `updateMetrics` immediately followed by `fit`. The succinct workflow enables you to implement incremental learning with prequential evaluation easily when you plan to track the model performance and train the model on all incoming data chunks. For linear models, see `updateMetricsAndFit` and, for naive Bayes classification models, see `updateMetricsAndFit`.

Once you create an incremental model object and choose a workflow to use, write a loop that implements incremental learning:

1. Read a chunk of observations from a data stream, when the chunk is available.
2. Implement the flexible or succinct workflow. To perform incremental learning properly, overwrite the input model with the output model. For example:

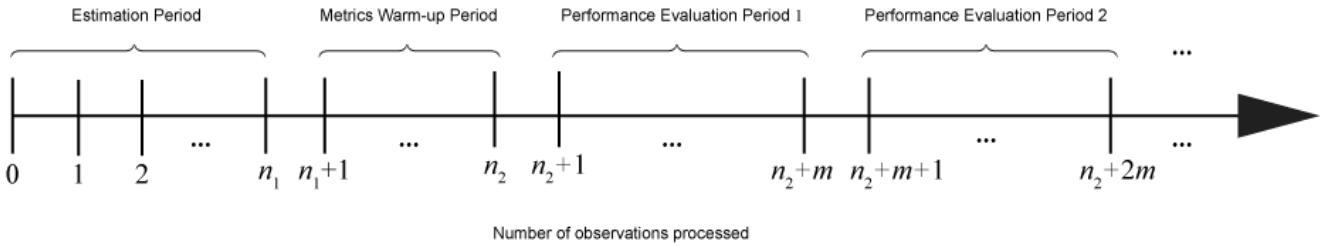
```
% Flexible workflow
Mdl = updateMetrics(Mdl,X,Y);
% Insert optional code
Mdl = fit(Mdl,X,Y);

% Succinct workflow
Mdl = updateMetricsAndFit(Mdl,X,Y);
```

   The model tracks its performance on incoming data incrementally using metrics measured since the beginning of training (cumulative) and over a specified window of consecutive observations (window). However, you can optionally compute the model loss on the incoming chunk, and then pass the incoming chunk and current model to the `loss` function. `loss` returns the scalar loss; it does not adjust the model. For linear models, see `loss` and, for naive Bayes classification models, see `loss`.

   Model configurations determine whether incremental learning functions train or evaluate model performance during each iteration. Configurations can change as the functions process data. For more details, see “Incremental Learning Periods” on page 26-6.
3. Optionally:
   - Generate predictions by passing the chunk and latest model to `predict`. For linear models, see `predict` and, for naive Bayes classification models, see `predict`.
   - If the model was fit to data, compute the resubstitution loss by passing the chunk and latest model to `loss`.
   - For naive Bayes classification models, the `logp` function enables you to detect outliers in real-time. The function returns the log unconditional probability density of the predictor variables at each observation in the chunk.

Incremental Learning Periods

Given incoming chunks of data, the actions performed by incremental learning functions depend on the current configuration or state of the model. This figure shows the periods (consecutive groups of observations) during which incremental learning functions perform particular actions.



This table describes the actions performed by incremental learning functions during each period.

Period	Associated Model Properties	Size (Number of Observations)	Actions
Estimation	`EstimationPeriod`, applies to linear classification and regression models only	n_1	When required, fitting functions choose values for hyperparameters based on estimation period observations. Actions include the following: • Estimate predictor moments `Mu` and `Sigma` for data standardization. • Adjust the learning rate `LearnRate` for SGD solvers according to the learning rate schedule `LearnRateSchedule`. • Estimate the SVM regression parameter ε `Epsilon`. • Store information buffers required for estimation. • Update corresponding properties at the end of the period.
Metrics Warm-up	`MetricsWarmupPeriod`	$n_2 - n_1$	When the property `IsWarm` is `false`, fitting functions perform the following actions: • Fit the model to the incoming chunk of data. • Update corresponding model properties, such as `Beta` or `DistributionParameters`, after fitting the model. • At the end of the period, the model is warm (the `IsWarm` property becomes `true`).

Period	Associated Model Properties	Size (Number of Observations)	Actions
Performance Evaluation *j*	`Metrics` and `MetricsWindowSize`	*m*	• At the start of Performance Evaluation Period 1, functions begin to track cumulative `Cumulative` and window `Window` metrics. `Window` is a vector of `NaN`s throughout this period. • Functions overwrite `Cumulative` metrics with the updated cumulative metric at each iteration. At the end of each Performance Evaluation Period, functions compute and overwrite `Window` metrics based on the last *m* observations. • Functions store information buffers required for computing model performance.

References

[1] Bifet, Albert, Ricard Gavaldá, Geoffrey Holmes, and Bernhard Pfahringer. Machine Learning for Data Streams with Practical Example in MOA. Cambridge, MA: The MIT Press, 2007.

[2] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

See Also

Objects

`incrementalClassificationLinear` | `incrementalClassificationNaiveBayes` | `incrementalRegressionLinear`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19
- “Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Configure Incremental Learning Model

An incremental learning model object completely specifies how functions implement incremental fitting and model performance evaluation. To configure (or prepare) an incremental learning model, create one by calling the object directly, or by converting a traditionally trained model to one of the objects. The following table lists the available objects, conversion functions, and supported learners for the corresponding machine learning objective.

Objective	Creation Approach	Function	Supported Learners
Binary classification	Call object	`incrementalClassificationLinear`	Linear SVM Logistic regression
	Convert model	`incrementalLearner` converts a linear SVM model (`ClassificationSVM` or `CompactClassificationSVM`) `incrementalLearner` converts a linear classification model (`ClassificationLinear`)	
Mutlicl ass classific ation	Call object	`incrementalClassificationNaiveBayes`	Naive Bayes classification with normally distributed predictor variables, conditioned on the class
	Convert model	`incrementalLearner` converts a full naive Bayes classification model (`ClassificationNaiveBayes`)	
Regress ion	Call object	`incrementalRegressionLinear`	Linear SVM regression Least squares
	Convert model	`incrementalLearner` converts a linear SVM regression model (`RegressionSVM` or `CompactRegressionSVM`) `incrementalLearner` converts a linear regression model (`RegressionLinear`)	

The creation approach you choose depends on the information you have and your preferences.

- **Call object**: Create an incremental model to your specifications by calling the object directly. This approach is flexible, enabling you to specify most options to suit your preferences, and the resulting model provides reasonable default values. However:

  - For linear classification or regression models, an estimation period might be required for your specifications.
  - For naive Bayes classification models, you must specify the maximum number of classes or all class names expected in the data during incremental learning.

  For more details, see “Call Object Directly” on page 26-9.
- **Convert model**: Convert a traditionally trained model to an incremental learner to initialize a model for incremental learning. `incrementalLearner` passes information that the traditionally trained model learned from the data. For linear models, passed information includes including optimized coefficient estimates, data characteristics, and applicable hyperparameter values. For naive Bayes classification models, passed information includes data characteristics, such as all

expected class names and fitted moments of the conditional predictor distributions. However, to convert a traditionally trained model, you must have a set of labeled data to which you can fit a model.

When you use `incrementalLearner`, you can specify all performance evaluation options and only training, model, and data options that are unknown during conversion. For more details, see "Convert Traditionally Trained Model" on page 26-12.

Regardless of the incremental model creation approach you use, consider these configurations:

- Model performance evaluation settings, such as the performance metrics to measure
- For linear models:
  - Model type, such as SVM
  - Coefficient initial values
  - Objective function solver, such as standard stochastic gradient descent (SGD)
  - Solver hyperparameter values, such as the learning rate of SGD solvers

Call Object Directly

Unlike when working with other machine learning model objects, you can create an incremental learning model by calling the corresponding object directly, without any knowledge about the data. For linear models, the only information required to create a model directly is the machine learning problem, either classification or regression, whereas naive Bayes classification models require the maximum number of classes expected in the data or all class names. For example, the following code creates a default incremental model for linear regression and a naive Bayes classification model for a data stream containing 5 classes.

```
MdlLR = incrementalRegressionLinear();
MdlNB = incrementalClassificationNaiveBayes('MaxNumClasses',5)
```

If you have information about the data to specify, or you want to configure model options or performance evaluation settings, use name-value arguments when you call the object. (All model properties are read-only; you cannot adjust them using dot notation.) For example, the following pseudocode creates an incremental logistic regression model for binary classification, initializes the linear model coefficients `Beta` and bias `Bias` (obtained from prior knowledge of the problem), and sets the performance metrics warm-up period to `500` observations.

```
Mdl = incrementalClassificationLinear('Learner','logistic',...
    'Beta',Beta,'Bias',Bias,'MetricsWarmupPeriod',500);
```

The following tables briefly describe notable options for the major aspects of incremental learning. For more details on all options, see `incrementalRegressionLinear`, `incrementalClassificationLinear`, or `incrementalClassificationNaiveBayes`.

Model Options and Data Properties

This table contains notable model options and data characteristics.

Model Type	Model Options and Data Properties	Description
Linear classification or regression	`'Beta'`	Linear coefficients that also serve as initial values for incremental fitting
	`'Bias'`	Model intercept that also serve as an initial value for incremental fitting
	`'Learner'`	Model type, such as linear SVM or least squares
Naive Bayes classification	`'Cost'`	Misclassification cost matrix
Classification	`'ClassNames'`	For classification, the expected class names in the observation labels

Training and Solver Options and Properties

This table contains notable training and solver options and properties.

Model Type	Training and Solver Options and Properties	Description
Linear classification or regression	`'EstimationPeriod'`	Pretraining estimation period
	`'Solver'`	Objective function optimization algorithm
	`'Standardize'`	Flag to standardize predictor data
	`'Lambda'`	Ridge penalty, a model hyperparameter that requires tuning for SGD optimization
	`'BatchSize'`	Mini-batch size, an SGD hyperparameter
	`'LearnRate'`	Learning rate, an SGD hyperparameter
	`'Mu'`	Read-only property containing predictor variable means
	`'Sigma'`	Read-only property containing predictor variable standard deviations
Naive Bayes classification	`'DistributionParameters'`	Fitted moments of each conditional predictor distribution given each class

For linear classification and regression models:

- The estimation period, specified by the number of observations in `EstimationPeriod`, occurs before training begins (see Incremental Learning Periods on page 26-6). During the estimation period, the incremental fitting function `fit` or `updateMetricsAndFit` computes quantities required for training when they are unknown. For example, if you set `'Standardize',true`, incremental learning functions require predictor means and standard deviations to standardize

the predictor data. Consequently, the incremental model requires a positive estimation period (the default is `1000`).

- The default solver is the adaptive scale-invariant solver `'scale-invariant'` [2], which is hyperparameter free and insensitive to the predictor variable scales; therefore, predictor data standardization is not required. You can specify standard or average SGD instead, `'sgd'` or `'asgd'`. However, SGD is sensitive to predictor variable scales and requires hyperparameter tuning, which can be difficult or impossible to do during incremental learning. If you plan to use an SGD solver, complete these steps:

  1. Obtain labeled data.
  2. Traditionally train a linear classification or regression model by calling `fitclinear` or `fitrlinear`, respectively. Specify the SGD solver you plan to use for incremental learning, cross-validate to determine an appropriate set of hyperparameters, and standardize the predictor data.
  3. Train the model on the entire sample using the specified hyperparameter set.
  4. Convert the resulting model to an incremental learner by using `incrementalLearner`.

Performance Evaluation Options and Properties

Performance evaluation properties and options enable you to configure how and when model performance is measured by the incremental learning function `updateMetrics` or `updateMetricsAndFit`. Regardless of the options you choose, first familiarize yourself with the incremental learning periods on page 26-6.

This table contains all performance evaluation options and properties.

Performance Evaluation Options and Properties	Description
`'Metrics'`	List of performance metrics or loss functions to measure incrementally
`'MetricsWarmupPeiod'`	Number of observations to which the incremental model must be fit before it tracks performance metrics
`'MetricsWindowSize'`	Number of observations to use to compute window performance metrics
`'IsWarm'`	Read-only property indicating whether the model is warm (measures performance metrics)
`'Metrics'`	Read-only property containing a table of tracked cumulative and window metrics

The metrics specified by the `'Metrics'` name-value form a table stored in the `Metrics` property of the model. For example, if you specify `'Metrics',["Metric1" "Metric2"]` when you create an incremental model `Mdl`, the `Metrics` property is

```
>> Mdl.Metrics

ans =

  2×2 table
                Cumulative    Window
                __________    ______
```

```
    Metric1        NaN         NaN
    Metric2        NaN         NaN
```

Specify a positive metrics warm-up period when you believe the model is of low quality and needs to be trained before the function `updateMetrics` or `updateMetricsAndFit` tracks performance metrics in the `Metrics` property. In this case, the `IsWarm` property is `false`, and you must pass the incoming data and model to the incremental fitting function `fit` or `updateMetricsAndFit`.

When the incremental fitting function processes enough data to satisfy the estimation (for linear models) and metrics warm-up periods, the `IsWarm` property becomes `true`, and you can measure the model performance on incoming data and optionally train the model. For naive Bayes classification models, incremental fitting functions must additionally fit the model to all expected classes to become warm.

When the model is warm, `updateMetrics` or `updateMetricsAndFit` tracks all specified metrics cumulatively (from the start of the evaluation) and within a window of observations specified by the `MetricsWindowSize` property. Cumulative metrics reflect the model performance over the entire incremental learning history; after Performance Evaluation Period 1 starts, cumulative metrics are independent of the evaluation period. Window metrics reflect the model performance only over the specified window size for each performance evaluation period.

Convert Traditionally Trained Model

`incrementalLearner` enables you to initialize an incremental model using information learned from a traditionally trained model. The converted model can generate predictions and it is warm, which means that incremental learning functions can measure model performance metrics from the start of the data stream. In other words, estimation and performance metrics warm-up periods are not required for incremental learning.

To convert a traditionally trained model to an incremental learner, pass the model and any options specified by name-value arguments to `incrementalLearner`. For example, the following pseudocode initializes an incremental classification model by using all information that a linear SVM model for binary classification has learned from a batch of data.

```
Mdl = fitcsvm(X,Y);
IncrementalMdl = incrementalLearner(Mdl,Name,Value);
```

`IncrementalMdl` is an incremental learner object associated with the machine learning objective.

Ease of incremental model creation and initialization is offset by decreased flexibility. The software assumes that fitted parameters, hyperparameter values, and data characteristics learned during traditional training are appropriate for incremental learning. Therefore, you cannot set corresponding learned or tuned options when you call `incrementalLearner`. This table lists notable read-only properties of `IncrementalMdl` that `incrementalLearner` transfers from `Mdl`, or that the function infers from other values.

Model Type	Property	Description
All	`NumPredictors`	Number of predictor variables. For models that dummy-code categorical predictor variables, `NumPredictors` is `numel(Mdl.ExpandedPredictorNames)`, and predictor variables expected during incremental learning correspond to the names. For more details, see “Dummy Variables” on page 2-48.
Classification	`ClassNames`	All class labels expected during incremental learning
	`Prior`	Prior class distribution
	`ScoreTransform`	A function to apply to classification scores. For example, if you configure an SVM model to compute posterior class probabilities, `ScoreTransform` (containing the score-to-posterior-probability function learned from the data) is transferred.
Regression	`Epsilon`	For an SVM learner, half the width of the epsilon-insensitive band
	`ResponseTransform`	A function to apply to predicted responses
Linear classification or regression	`Beta`	Linear model coefficients
	`Bias`	Model intercept
	`Learner`	Linear model type
	`Mu`	For an SVM model object, the predictor variable means
	`Sigma`	For an SVM model object, the predictor variable standard deviations
Naive Bayes classification	`DistributionNames`	A `NumPredictors` length cell vector with `'normal'` in each cell. If you convert a naive Bayes classification model containing at least one non-normal predictor, `incrementalLearner` issues an error.
	`DistributionParameters`	Fitted moments of each conditional predictor distribution given each class

Note

- The `NumTrainingObservations` property of `IncrementalMdl` does not include the observations used to train `Mdl`.
- If you specify `'Standardize',true` when you train `Mdl`, `IncrementalMdl` is configured to standardize predictors during incremental learning by default.

The following conditions apply when you convert a linear classification or regression model (`ClassificationLinear` and `RegressionLinear`, respectively):

- Incremental fitting functions support ridge (L2) regularization only.
- Incremental fitting functions support the specification of only one regularization value. Therefore, if you specify a regularization path (vector of regularization values) when you call `fitclinear` or `fitrlinear`, choose the model associated with one penalty by passing it to `selectModels`.
- If you solve the objective function by using standard or average SGD (`'sgd'` or `'asgd'` for the `'Solver'` name-value argument), these conditions apply when you call `incrementalLearner`:
  - `incrementalLearner` transfers the solver used to optimize `Mdl` to `IncrementalMdl`.
  - You can specify the adaptive scale-invariant solver `'scale-invariant'` instead, but you cannot specify a different SGD solver.
  - If you do not specify the adaptive scale-invariant solver, `incrementalLearner` transfers model and solver hyperparameter values to the incremental model object, such as the learning rate `LearnRate`, mini-batch size `BatchSize`, and ridge penalty `Lambda`. You cannot modify the transferred properties.

If you require more flexibility when you create an incremental model, you can call the object directly on page 26-9 and initialize the model by individually setting learned information using name-value arguments. The following pseudocode show two examples:

- Initialize an incremental classification model from the coefficients and class names learned by fitting a linear SVM model for binary classification to a batch of data `Xc` and `Yc`.
- Initialize an incremental regression model from the coefficients learned by fitting a linear model to a batch of data `Xr` and `Yr`.

```
% Linear Classification
Mdl = fitcsvm(Xc,Yc);
IncrementalMdl = incrementalClassificationLinear('Beta',Mdl.Beta,...
    'Bias',Mdl.Bias,'ClassNames',Mdl.ClassNames);

% Linear Regression
Mdl = fitlm(Xr,Yr);
Bias = Mdl.Coefficients.Estimate(1);
Beta = Mdl.Coefficients.Estimate(2:end);
IncrementalMdl = incrementalRegressionLinear('Learner','leastsquares',...
    'Bias',Bias,'Beta',Beta);
```

For naive Bayes classification models, you cannot specify the moments of the conditional probability distribution of each predictor variable `DistributionParameters`; they must be fitted to data, by `fit` or `fitcnb`

References

[1] Bifet, Albert, Ricard Gavaldá, Geoffrey Holmes, and Bernhard Pfahringer. Machine Learning for Data Streams with Practical Example in MOA. Cambridge, MA: The MIT Press, 2007.

[2] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

See Also

Objects
`incrementalClassificationLinear` | `incrementalClassificationNaiveBayes` | `incrementalRegressionLinear`

More About

- “Incremental Learning Overview” on page 26-2
- “Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19
- “Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Implement Incremental Learning for Linear Regression Using Succinct Workflow

This example shows how to use the succinct workflow to implement incremental learning for linear regression with prequential evaluation. Specifically, this example does the following:

1. Create a default incremental learning model for linear regression.
2. Simulate a data stream using a for loop, which feeds small chunks of observations to the incremental learning algorithm.
3. For each chunk, use `updateMetricsAndFit` to measure the model performance given the incoming data, and then fit the model to that data.

Create Default Model Object

Create a default incremental learning model for linear regression.

```
Mdl = incrementalRegressionLinear()
```

```
Mdl =
  incrementalRegressionLinear

               IsWarm: 0
              Metrics: [1x2 table]
    ResponseTransform: 'none'
                 Beta: [0x1 double]
                 Bias: 0
              Learner: 'svm'


  Properties, Methods
```

```
Mdl.EstimationPeriod
```

```
ans = 1000
```

`Mdl` is an `incrementalRegressionLinear` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations. The software sets the estimation period to 1000 because half the width of the epsilon insensitive band `Epsilon` is unknown. You can set `Epsilon` to a positive floating point scalar by using the `'Epsilon'` name-value pair argument. This action results in a default estimation period of 0.

Load Data

Load the robot arm data set.

```
load robotarm
```

For details on the data set, enter `Description` at the command line.

Implement Incremental Learning

Use the succinct workflow to update model performance metrics and fit the incremental model to the training data by calling the `updateMetricsAndFit` function. At each iteration:

- Process 50 observations to simulate a data stream.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the cumulative metrics, window metrics, and the first coefficient β_1 to see how they evolve during incremental learning.

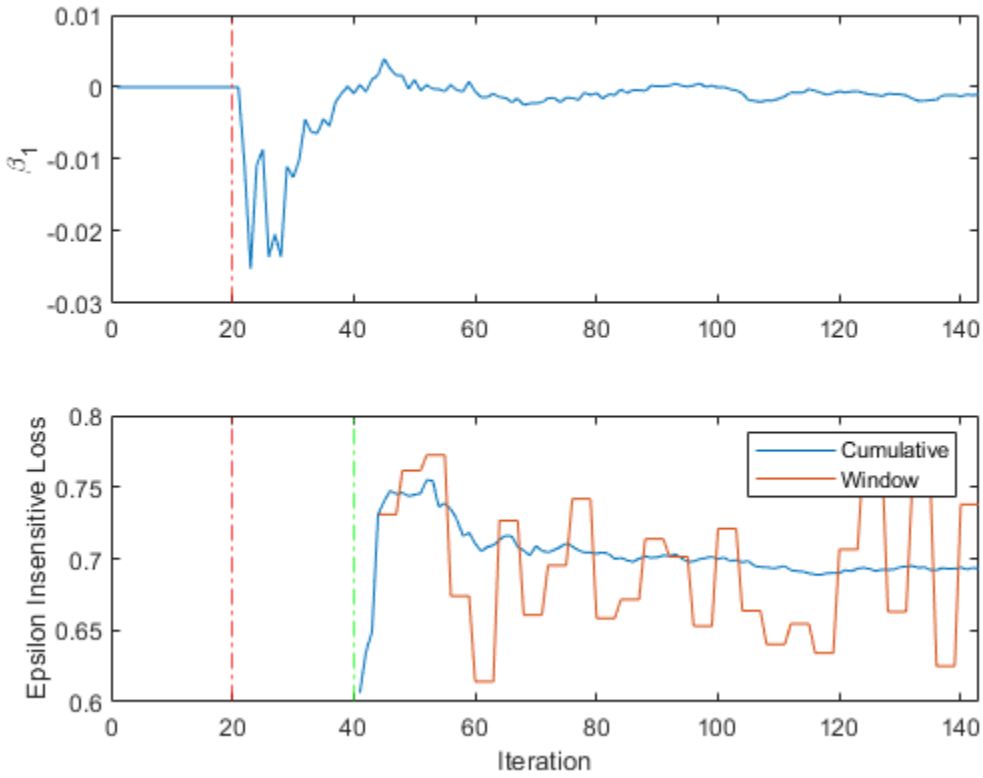
```
% Preallocation
n = numel(ytrain);
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,Xtrain(idx,:),ytrain(idx));
    ei{j,:} = Mdl.Metrics{"EpsilonInsensitiveLoss",:};
    beta1(j + 1) = Mdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

Inspect Model Evolution

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline((Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod)/numObsPerChunk,'g-.');
legend(h,ei.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- After the estimation period (first 20 iterations), fit β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 200 observations (4 iterations).

See Also

Objects
`incrementalRegressionLinear`

Functions
`updateMetricsAndFit`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Linear Regression Using Flexible Workflow” on page 26-22

Implement Incremental Learning for Classification Using Succinct Workflow

This example shows how to use the succinct workflow to implement incremental learning for binary classification with prequential evaluation. Specifically, this example does the following:

1. Create a default incremental learning model for binary classification.
2. Simulate a data stream using a for loop, which feeds small chunks of observations to the incremental learning algorithm.
3. For each chunk, use `updateMetricsAndFit` to measure the model performance given the incoming data, and then fit the model to that data.

Although this example treats the application as a binary classification problem, you can implement multiclass incremental learning using the naive Bayes algorithm instead by following this same workflow. See the `incrementalClassificationNaiveBayes` object.

Create Default Model Object

Create a default incremental learning model for binary classification.

```
Mdl = incrementalClassificationLinear()
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
              Beta: [0x1 double]
              Bias: 0
           Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Implement Incremental Learning

Use the succinct workflow to update model performance metrics and fit the incremental model to the training data by calling the `updateMetricsAndFit` function. At each iteration:

- Process 50 observations to simulate a data stream.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the cumulative metrics, the window metrics, and the first coefficient β_1 to see how they evolve during incremental learning.

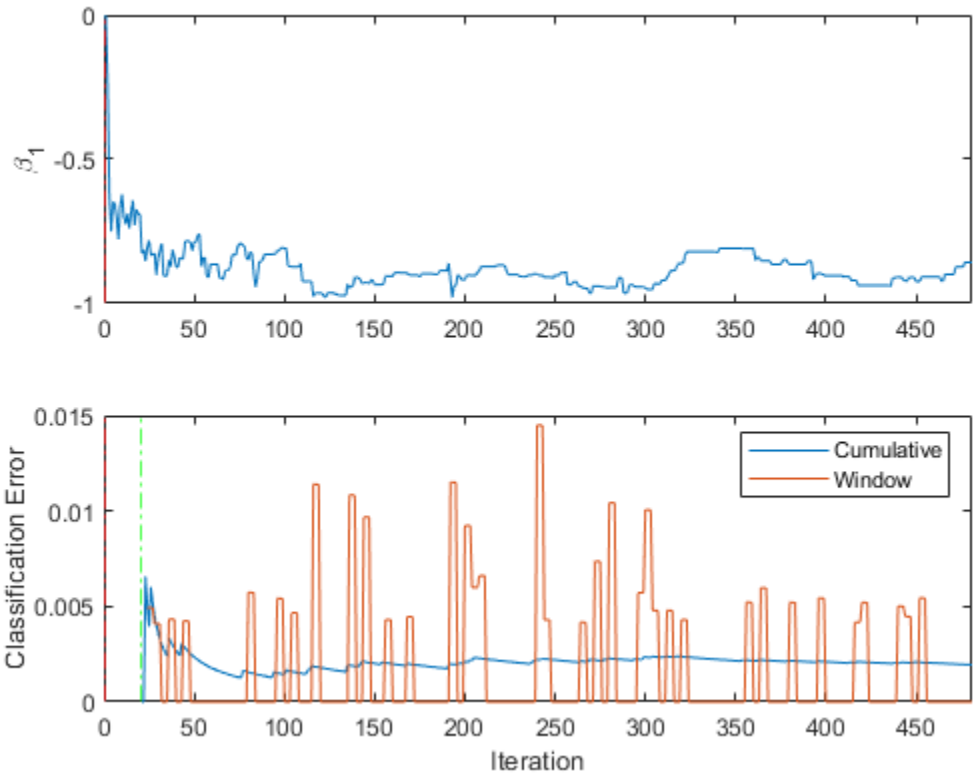
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    beta1(j + 1) = Mdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

Inspect Model Evolution

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline((Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod)/numObsPerChunk,'g-.');
legend(h,ce.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 200 observations (4 iterations).

See Also

Objects
`incrementalClassificationLinear`

Functions
`updateMetricsAndFit`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Implement Incremental Learning for Linear Regression Using Flexible Workflow

This example shows how to use the flexible workflow to implement incremental learning for linear regression with prequential evaluation. A traditionally trained model initializes the incremental model. Specifically, this example does the following:

1. Train a linear regression model on a subset of data.
2. Convert the traditionally trained model to an incremental learning model for linear regression.
3. Simulate a data stream using a for loop, which feeds small chunks of observations to the incremental learning algorithm.
4. For each chunk, use `updateMetrics` to measure the model performance given the incoming data, and then use `fit` to fit the model to that data.

Load and Preprocess Data

Load the 2015 NYC housing data set, and shuffle the data. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
rng(1); % For reproducibility
n = size(NYCHousing2015,1);
idxshuff = randsample(n,n);
NYCHousing2015 = NYCHousing2015(idxshuff,:);
```

Suppose that the data collected from Manhattan (`BOROUGH = 1`) was collected using a new method that doubles its quality. Create a weight variable that attributes `2` to observations collected from Manhattan, and `1` to all other observations.

```
NYCHousing2015.W = ones(n,1) + (NYCHousing2015.BOROUGH == 1);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data. Transpose the data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}]';
```

Train Linear Regression Model

Fit a linear regression model to a random sample of half the data. Specify that observations are oriented along the columns of the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitrlinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns')
```

```
TTMdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [313x1 double]
                 Bias: 0.1889
               Lambda: 2.1977e-05
              Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `RegressionLinear` model object representing a traditionally trained linear regression model.

Convert Trained Model

Convert the traditionally trained linear regression model to a linear regression model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalRegressionLinear

               IsWarm: 1
              Metrics: [1x2 table]
    ResponseTransform: 'none'
                 Beta: [313x1 double]
                 Bias: 0.1889
              Learner: 'svm'


  Properties, Methods
```

Implement Incremental Learning

Use the flexible workflow to update model performance metrics and fit the incremental model to the training data by calling the `updateMetrics` and `fit` functions separately. Simulate a data stream by processing 500 observations at a time. At each iteration:

1. Call `updateMetrics` to update the cumulative and window epsilon insensitive loss of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify that observations are oriented along the columns of the data.
2. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify that observations are oriented along the columns of the data.
3. Store the losses and last estimated coefficient β_{313}.

```
% Preallocation
numObsPerChunk = 500;
nchunk = floor(n/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta313 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,X(:,idx),Y(idx),'ObservationsIn','columns');
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    IncrementalMdl = fit(IncrementalMdl,X(:,idx),Y(idx),'ObservationsIn','columns');
    beta313(j) = IncrementalMdl.Beta(end);
end
```

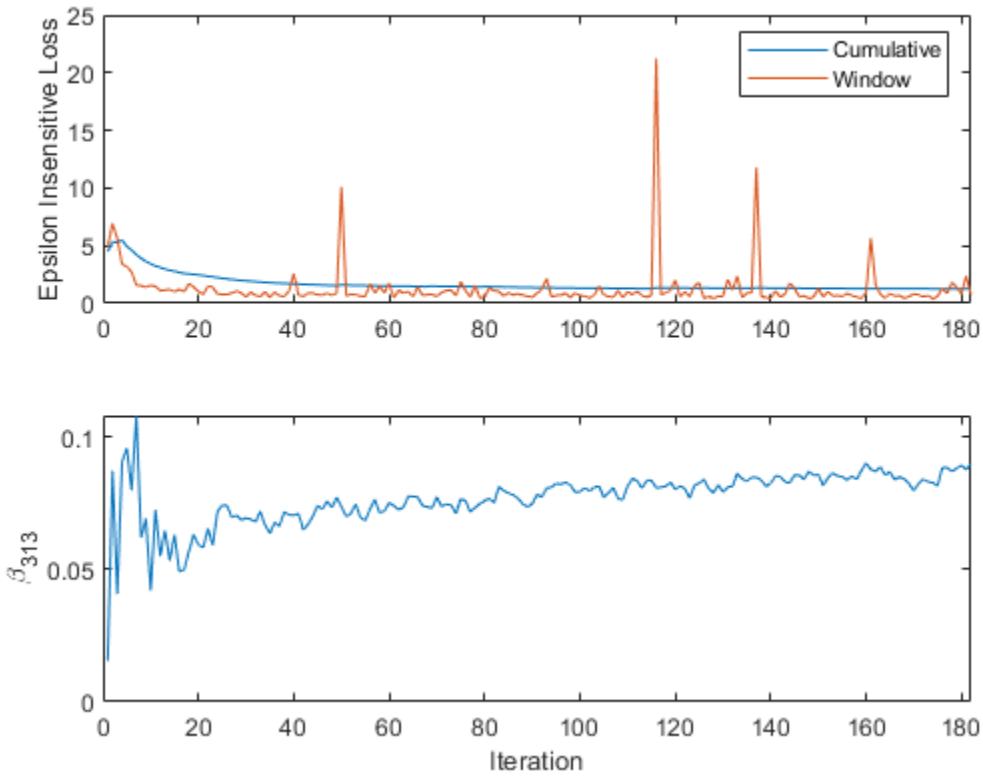
`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Inspect Model Evolution

Plot a trace plot of the performance metrics and estimated coefficient β_{313}.

```
figure;
subplot(2,1,1)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
legend(h,ei.Properties.VariableNames)
subplot(2,1,2)
plot(beta313)
ylabel('\beta_{313}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss gradually changes with each iteration (chunk of 500 observations), whereas the window loss jumps. Because the metrics window is 200 by default, `updateMetrics` measures the performance based on the latest 200 observations in each 500 observation chunk.

β_{313} changes abruptly at first and then just slightly as `fit` processes chunks of observations.

See Also

Objects
`incrementalRegressionLinear`

Functions
`fit` | `updateMetrics`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Linear Regression Using Succinct Workflow” on page 26-16

Implement Incremental Learning for Classification Using Flexible Workflow

This example shows how to use the flexible workflow to implement incremental learning for binary classification with prequential evaluation. A traditionally trained model initializes the incremental model. Specifically, this example does the following:

1. Train a linear model for binary classification on a subset of data.
2. Convert the traditionally trained model to an incremental learning model for binary classification.
3. Simulate a data stream using a for loop, which feeds small chunks of observations to the incremental learning algorithm.
4. For each chunk, use `updateMetrics` to measure the model performance given the incoming data, and then use `fit` to fit the model to that data.

Although this example treats the application as a binary classification problem, you can implement multiclass incremental learning using the naive Bayes algorithm instead by following this same workflow. See the `incrementalClassificationNaiveBayes` object.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data. Orient the observations of the predictor data in columns.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:)';
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Train Linear Model for Binary Classification

Fit a linear model for binary classification to a random sample of half the data. Specify that the observations are oriented along the columns of the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitclinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns')
```

```
TTMdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2999
            Lambda: 8.2967e-05
           Learner: 'svm'
```

```
  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear model for binary classification.

Convert Trained Model

Convert the traditionally trained classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

            IsWarm: 1
           Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2999
           Learner: 'svm'


  Properties, Methods
```

Implement Incremental Learning

Use the flexible workflow to update model performance metrics and fit the incremental model to the training data by calling the `updateMetrics` and `fit` functions separately. Simulate a data stream by processing 50 observations at a time. At each iteration:

1. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify that the observations are oriented in columns.
2. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify that the observations are oriented in columns.
3. Store the classification error and first estimated coefficient β_1.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];
Xil = X(:,idxil);
Yil = Y(idxil);

% Incremental fitting
```

```
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(:,idx),Yil(idx),...
        'ObservationsIn','columns');
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(:,idx),Yil(idx),'ObservationsIn','columns');
    beta1(j + 1) = IncrementalMdl.Beta(end);
end
```

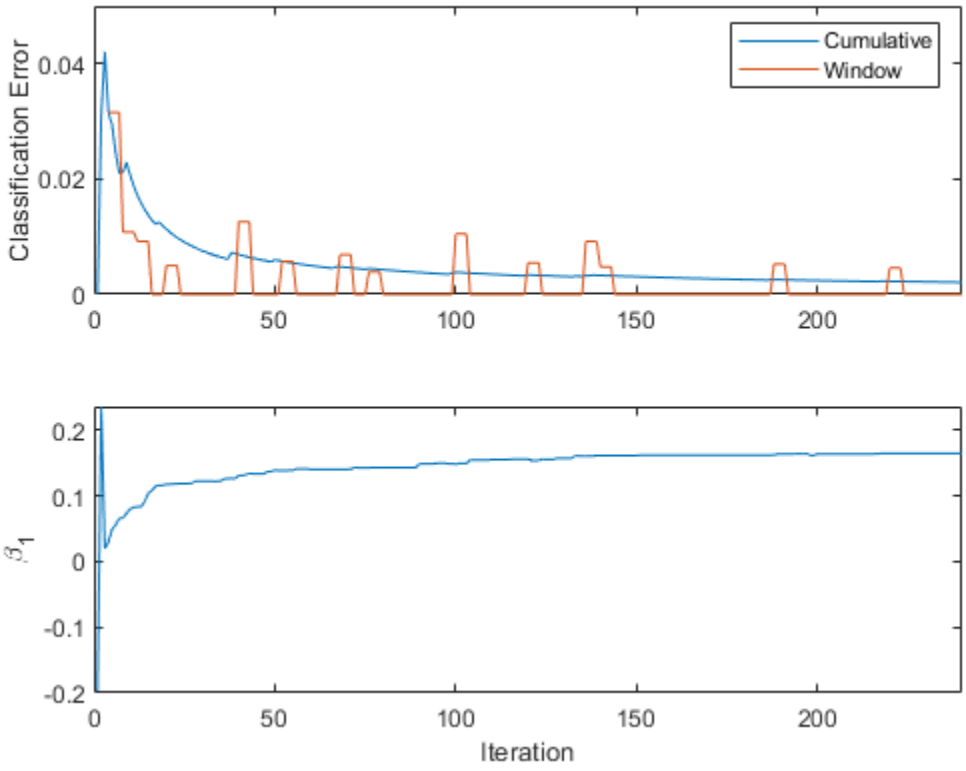
`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Inspect Model Evolution

Plot a trace plot of the performance metrics and estimated coefficient β_1.

```
figure;
subplot(2,1,1)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
legend(h,ce.Properties.VariableNames)
subplot(2,1,2)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable and decreases gradually, whereas the window loss jumps.

β_1 changes abruptly at first, and then gradually levels off as `fit` processes more chunks of observations.

See Also

Objects
`incrementalClassificationLinear`

Functions
`fit` | `updateMetrics`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19

Initialize Incremental Learning Model from SVM Regression Model Trained in Regression Learner

This example shows how to tune and train a linear SVM regression model using the **Regression Learner** app. Then, at the command line, initialize and train an incremental model for linear SVM regression using the information gained from training in the app.

Load and Preprocess Data

Load the 2015 NYC housing data set, and shuffle the data. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
rng(1); % For reproducibility
n = size(NYCHousing2015,1);
idxshuff = randsample(n,n);
NYCHousing2015 = NYCHousing2015(idxshuff,:);
```

For numerical stability, scale `SALEPRICE` by `1e6`.

```
NYCHousing2015.SALEPRICE = NYCHousing2015.SALEPRICE/1e6;
```

Consider training a linear SVM regression model to about 1% of the data, and reserving the remaining data for incremental learning.

Regression Learner supports categorical variables. However, because SVM models require dummy-coded categorical variables, and the `BUILDINGCLASSCATEGORY` and `NEIGHBORHOOD` variables contain many levels (some with low representation), the probability that a partition does not have all categories is high. Therefore, dummy-code all categorical variables. Concatenate the matrix of dummy variables to the rest of the numeric variables.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvars = splitvars(varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
      'InputVariables',catvars));
NYCHousing2015(:,catvars) = [];
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
NYCHousing2015 = [dumvars NYCHousing2015(:,idxnum)];
```

Randomly partition the data into 1% and 99% subsets by calling `cvpartition` and specifying a holdout (test) sample proportion of `0.99`. Create tables for the 1% and 99% partitions.

```
cvp = cvpartition(n,'HoldOut',0.99);
idxtt = cvp.training;
idxil = cvp.test;
NYCHousing2015tt = NYCHousing2015(idxtt,:);
NYCHousing2015il = NYCHousing2015(idxil,:);
```

Tune and Train Model Using Regression Learner

Open **Regression Learner** by entering `regressionLearner` at the command line.

```
regressionLearner
```

Alternatively, on the **Apps** tab, click the **Show more** arrow to open the apps gallery. Under **Machine Learning and Deep Learning**, click the app icon.

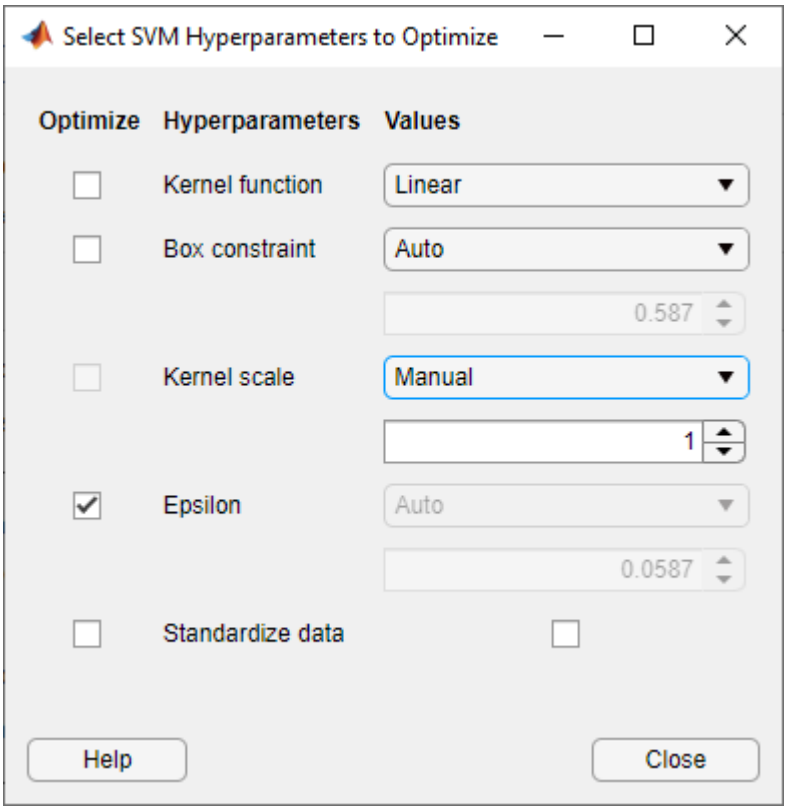
Choose the training data set and variables.

1. On the **Regression Learner** tab, in the **File** section, select **New Session**, and then select **From Workspace**.
2. In the **New Session from Workspace** dialog box, under **Data Set Variable**, select the data set **NYCHousing2015tt**.
3. Under **Response**, ensure the response variable **SALEPRICE** is selected.
4. Click **Start Session**.

The app implements 5-fold cross-validation by default.

Train a linear SVM regression model. Tune only the **Epsilon** hyperparameter by using Bayesian optimization.

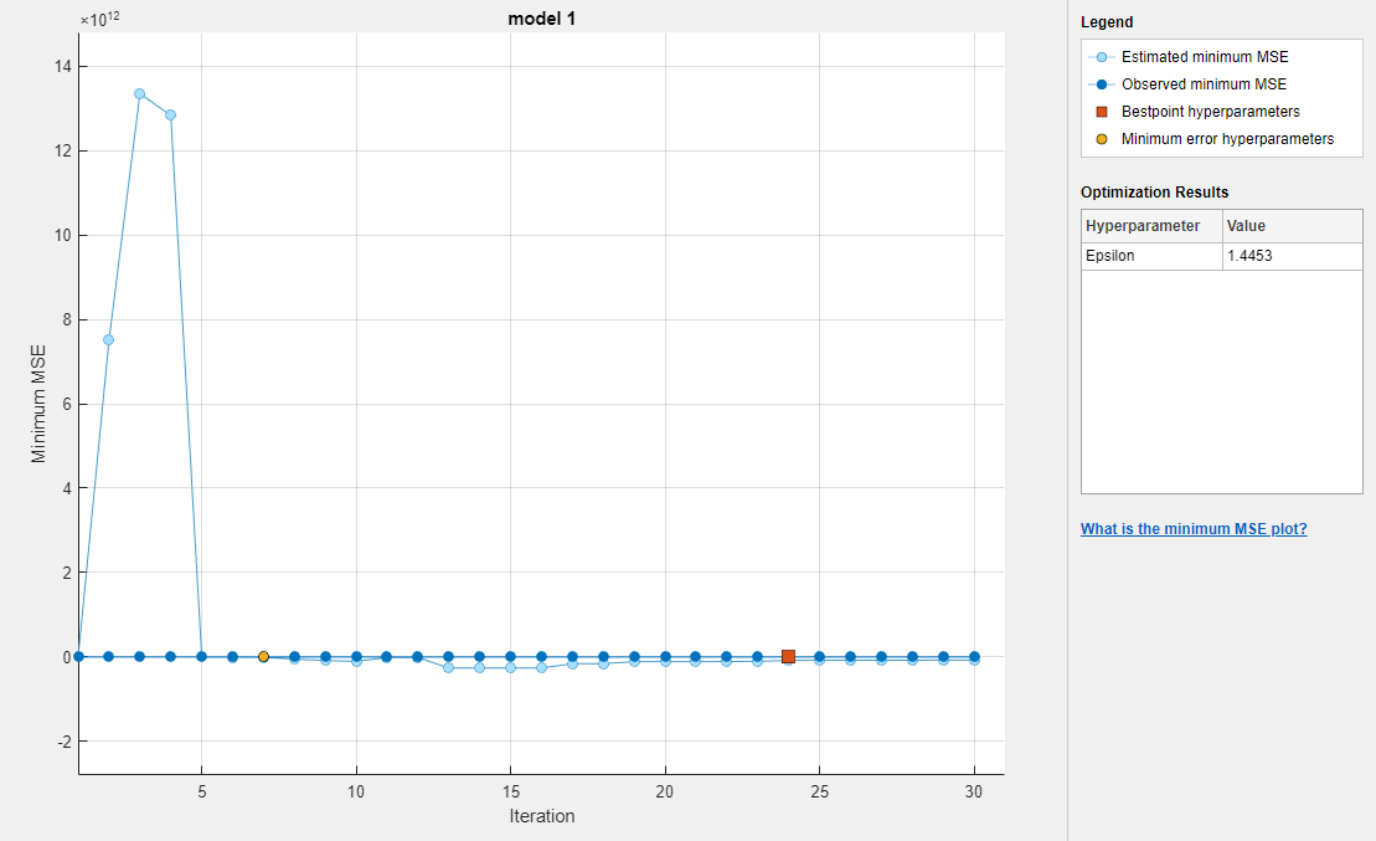
1. On the **Regression Learner** tab, in the **Model Type** section, click the **Show more** arrow to open the apps gallery. In the **Support Vector Machines** section, click **Optimizable SVM**.
2. On the **Regression Learner** tab, in the **Model Type** section, select **Advanced**, and then select **Advanced**.
3. In the **Select SVM Hyperparameters to Optimize** dialog box:
   a. Deselect the **Optimize** boxes for all options except **Epsilon**.
   b. Set the value of **Kernel scale** to `Manual` and `1`.
   c. Deselect the **Value** box **Standardize data**.



4. Close the dialog box.

5 On the **Regression Learner** tab, in the **Training** section, click **Train**.

The app shows a plot of the generalization minimum MSE of the model as optimization progresses. The app can take some time to optimize the algorithm.



Export the trained, optimized linear SVM regression model.

1 On the **Regression Learner** tab, in the **Export** section, select **Export Model**, and select **Export Model**.

2 In the **Export Model** dialog box, click **OK**.

The app passes the trained model, among other variables, in the structure array `trainedModel` to the workspace. Close **Regression Learner** .

Convert Exported Model to Incremental Model

At the command line, extract the trained SVM regression model from `trainedModel`.

```
Mdl = trainedModel.RegressionSVM;
```

Convert the model to an incremental model.

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl.Epsilon
```

```
IncrementalMdl =
```

```
  incrementalRegressionLinear

              IsWarm: 1
             Metrics: [1×2 table]
    ResponseTransform: 'none'
                Beta: [312×1 double]
                Bias: 10.1437
             Learner: 'svm'


  Properties, Methods


ans =

    1.4453
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object for incremental learning using a linear SVM regression model. `incrementalLearner` initializes `IncrementalMdl` using the coefficients and the optimized value of the `Epsilon` hyperparameter learned from `Mdl`. Therefore, you can predict responses by passing `IncrementalMdl` and data to `predict`. Also, the `IsWarm` property is `true`, which means that the incremental learning functions measure the model performance from the start of incremental learning.

Implement Incremental Learning

Because incremental learning functions accept floating-point matrices only, create matrices for the predictor and response data.

```
Xil = NYCHousing2015il{:,1:(end-1)};
Yil = NYCHousing2015il{:,end};
```

Perform incremental learning on the 99% data partition by using the `updateMetricsAndFit` function. Simulate a data stream by processing 500 observations at a time. At each iteration:

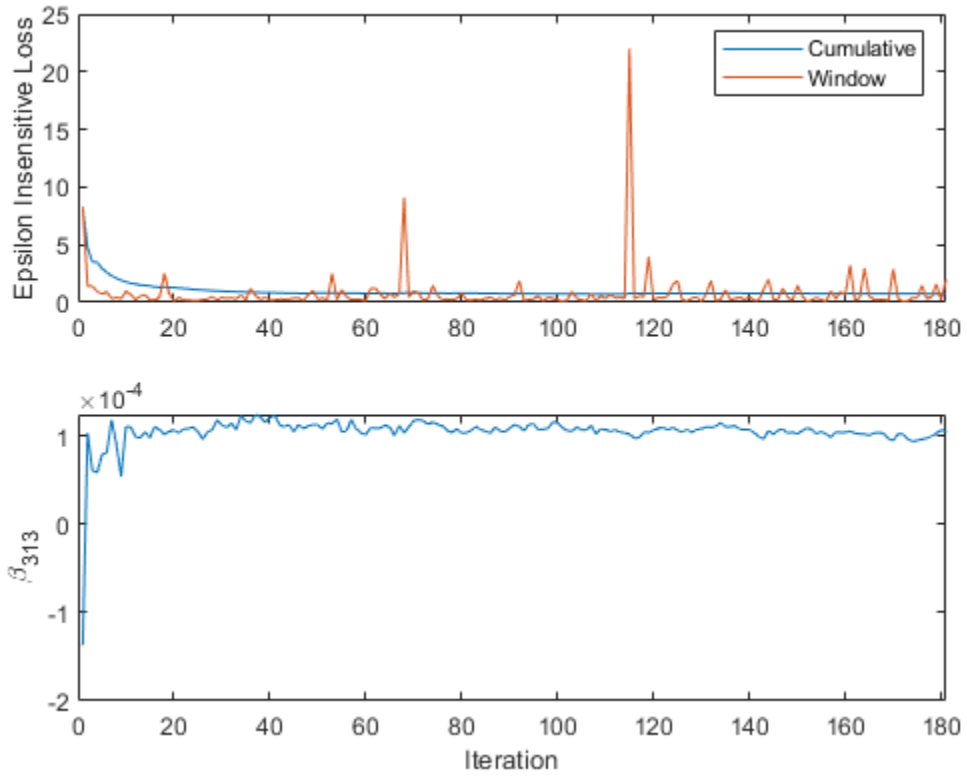
1. Call `updateMetricsAndFit` to update the cumulative and window epsilon insensitive loss of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model.
2. Store the losses and last estimated coefficient β_{313}.

```
% Preallocation
nil = sum(idxil);
numObsPerChunk = 500;
nchunk = floor(nil/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta313 = [IncrementalMdl.Beta(end); zeros(nchunk,1)];

% Incremental learning
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    beta313(j + 1) = IncrementalMdl.Beta(end);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

Plot a trace plot of the performance metrics and estimated coefficient β_{313}.

```
figure;
subplot(2,1,1)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
legend(h,ei.Properties.VariableNames)
subplot(2,1,2)
plot(beta313)
ylabel('\beta_{313}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss gradually changes with each iteration (chunk of 500 observations), whereas the window loss jumps. Because the metrics window is 200 by default, `updateMetricsAndFit` measures the performance based on the latest 200 observations in each 500 observation chunk.

β_{313} changes abruptly and then levels off as `updateMetricsAndFit` processes chunks of observations.

See Also

Apps
Regression Learner

Objects
`incrementalRegressionLinear`

Functions
`predict` | `updateMetricsAndFit`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Linear Regression Using Flexible Workflow” on page 26-22

Initialize Incremental Learning Model from Logistic Regression Model Trained in Classification Learner

This example shows how to train a logistic regression model using the **Classification Learner** app. Then, at the command line, initialize and train an incremental model for binary classification using the information gained from training in the app.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
actid = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by creating a categorical array that identifies whether the subject is moving (`actid > 2`).

```
moveidx = actid > 2;
Y = repmat("NotMoving",n,1);
Y(moveidx) = "Moving";
Y = categorical(Y);
```

Consider training a logistic regression model to about 1% of the data, and reserving the remaining data for incremental learning.

Randomly partition the data into 1% and 99% subsets by calling `cvpartition` and specifying a holdout (test) sample proportion of `0.99`. Create variables for the 1% and 99% partitions.

```
cvp = cvpartition(n,'HoldOut',0.99);
idxtt = cvp.training;
idxil = cvp.test;

Xtt = X(idxtt,:);
Xil = X(idxil,:);
Ytt = Y(idxtt);
Yil = Y(idxil);
```

Train Model Using Classification Learner

Open **Classification Learner** by entering `classificationLearner` at the command line.

```
classificationLearner
```

Alternatively, on the **Apps** tab, click the **Show more** arrow to open the apps gallery. Under **Machine Learning and Deep Learning**, click the app icon.

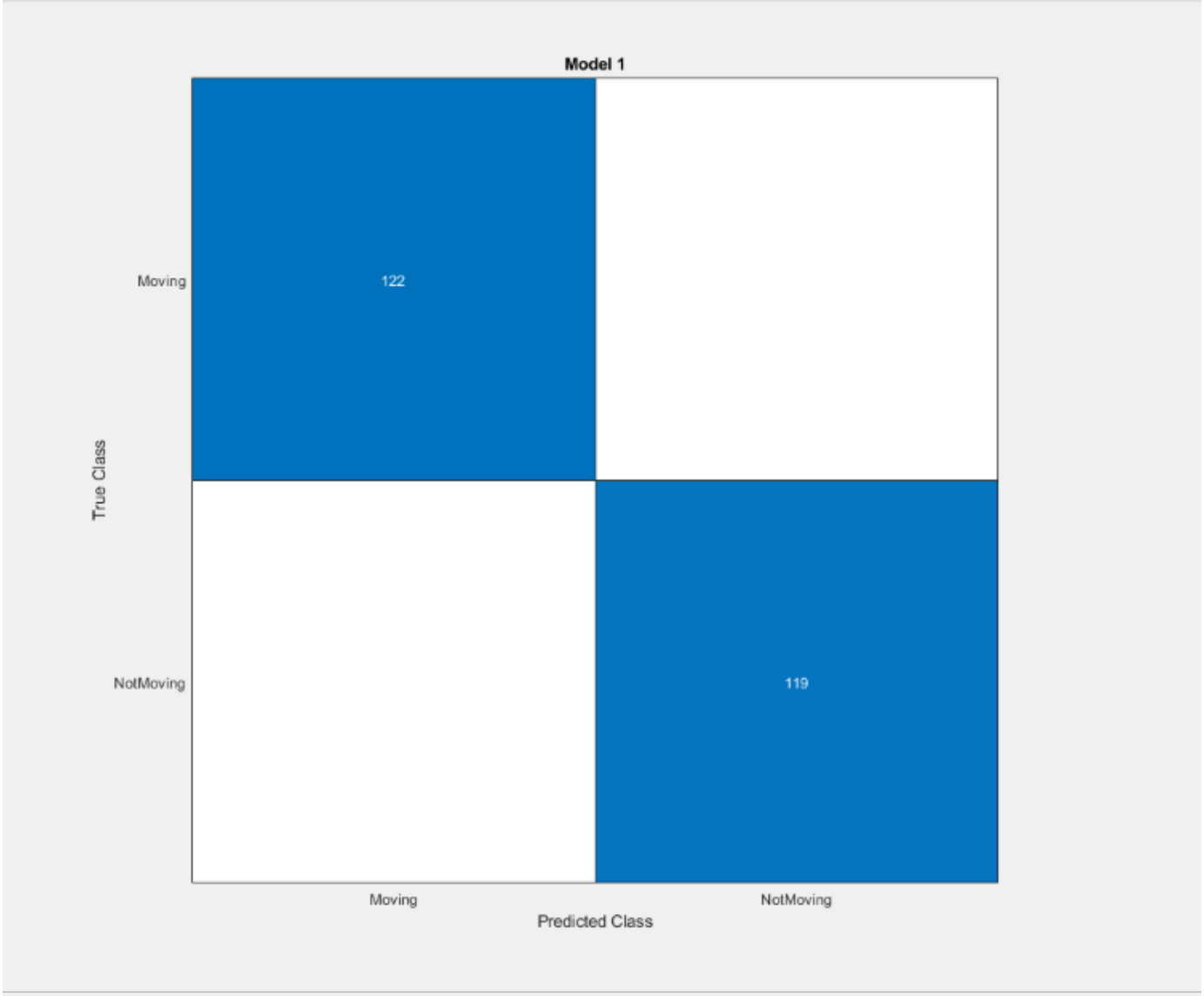
Choose the training data set and variables.

1. On the **Classification Learner** tab, in the **File** section, select **New Session > From Workspace**.

2. In the **New Session from Workspace** dialog box, under **Data Set Variable**, select the predictor variable **Xtt**.
3. Under **Response**, click **From workspace**; note that **Ytt** is selected automatically.
4. Under **Validation**, click **Resubstitution Validation**.
5. Click **Start Session**.

Train a logistic regression model.

1. On the **Classification Learner** tab, in the **Model Type** section, click the **Show more** arrow to open the gallery of models. In the **Logistic Regression Classifiers** section, click **Logistic Regression**.
2. On the **Classification Learner** tab, in the **Training** section, click **Train**.
3. When the app finishes training the model, plot a confusion matrix. On the **Classification Learner** tab, in the **Plots** section, click **Confusion Matrix** and select **Validation Data**.



The confusion matrix suggests that the model classifies in-sample observations well.

Export the trained logistic regression model.

1. On the **Classification Learner** tab, in the **Export** section, select **Export Model > Export Model**.
2. In the **Export Model** dialog box, click **OK**.

The app passes the trained model, among other variables, in the structure array `trainedModel` to the workspace. Close **Classification Learner** .

Initialize Incremental Model Using Exported Model

At the command line, extract the trained logistic regression model and the class names from `trainedModel`. The model is a `GeneralizedLinearModel` object. Because class names must match the data type of the response variable, convert the stored value to `categorical`.

```
Mdl = trainedModel.GeneralizedLinearModel;
ClassNames = categorical(trainedModel.ClassNames);
```

Extract the intercept and the coefficients from the model. The intercept is the first coefficient.

```
Bias = Mdl.Coefficients.Estimate(1);
Beta = Mdl.Coefficients.Estimate(2:end);
```

You cannot convert a `GeneralizedLinearModel` object to an incremental model directly. However, you can initialize an incremental model for binary classification by passing information learned from the app, such as estimated coefficients and class names.

Create an incremental model for binary classification directly. Specify the learner, intercept, coefficient estimates, and class names learned from **Classification Learner**. Because good initial values of coefficients exist and all class names are known, specify a metrics warm-up period of length `0`.

```
IncrementalMdl = incrementalClassificationLinear('Learner','logistic',...
    'Beta',Beta,'Bias',Bias,'ClassNames',ClassNames,...
    'MetricsWarmupPeriod',0)
```

```
IncrementalMdl =

  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1×2 table]
        ClassNames: [Moving    NotMoving]
    ScoreTransform: 'logit'
              Beta: [60×1 double]
              Bias: -471.7873
           Learner: 'logistic'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object for incremental learning using a logistic regression model. Because coefficients and all class names are specified, you can predict responses by passing `IncrementalMdl` and data to `predict`.

Implement Incremental Learning

Perform incremental learning on the 99% data partition by using the `updateMetricsAndFit` function. Simulate a data stream by processing 50 observations at a time. At each iteration:

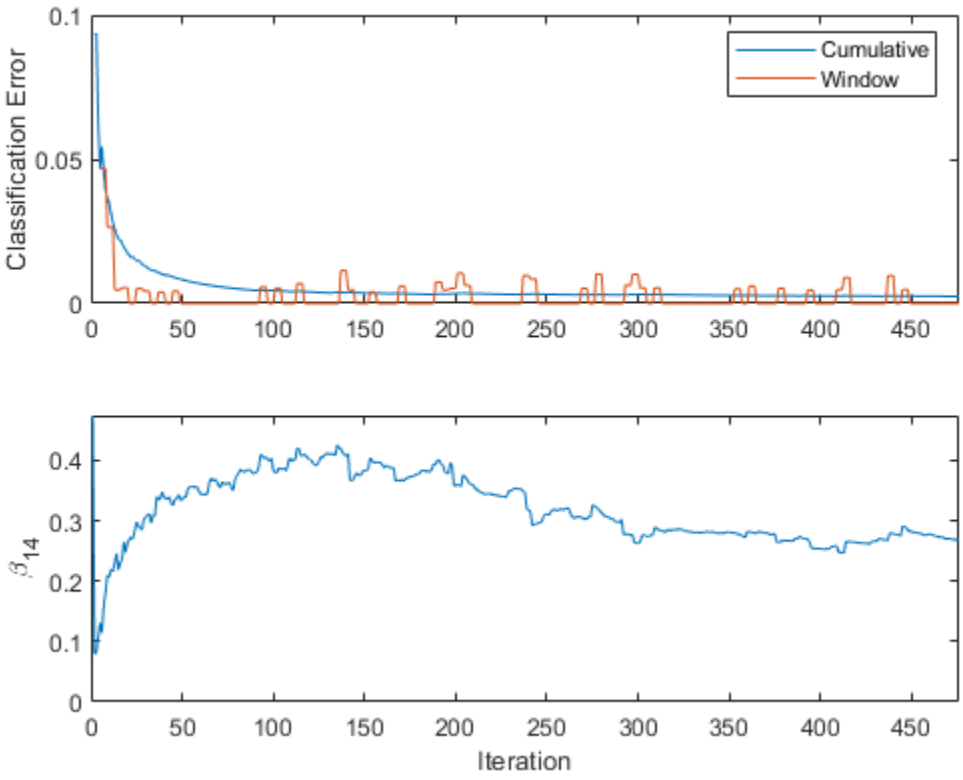
1. Call `updateMetricsAndFit` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model.
2. Store the losses and the estimated coefficient β_{14}.

```
% Preallocation
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta14 = [IncrementalMdl.Beta(14); zeros(nchunk,1)];

% Incremental learning
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    beta14(j + 1) = IncrementalMdl.Beta(14);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Plot a trace plot of the performance metrics and β_{14}.

```
figure;
subplot(2,1,1)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
legend(h,ce.Properties.VariableNames)
subplot(2,1,2)
plot(beta14)
ylabel('\beta_{14}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss gradually changes with each iteration (chunk of 50 observations), whereas the window loss jumps. Because the metrics window is 200 by default and `updateMetricsAndFit` measures the performance every four iterations.

β_{14} adapts to the data as `updateMetricsAndFit` processes chunks of observations.

See Also

Apps
Classification Learner

Objects
`incrementalClassificationLinear`

Functions
`predict` | `updateMetricsAndFit`

More About

- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Perform Conditional Training during Incremental Learning

This example shows how to train a naive Bayes mutliclass classification model for incremental learning only when the model performance is unsatisfactory.

The flexible incremental learning workflow enables you to train an incremental model on an incoming batch of data only when it is necessary (see “What Is Incremental Learning?” on page 26-2). For example, if the performance metrics of a model are satisfactory, then, to increase efficiency, you can skip training on incoming batches until the metrics become unsatisfactory.

Load Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Train Naive Bayes Classification Model

Configure a naive Bayes classification model for incremental learning by setting all the following:

- The maximum number of expected classes to 5
- The tracked performance metric to the misclassification error rate, which also includes minimal cost
- The metrics window size to 1000
- The metrics warmup period to 50

```
initobs = 50;
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5,'MetricsWindowSize',1000,...
    'Metrics','classiferror','MetricsWarmupPeriod',initobs);
```

Fit the configured model to the first 50 observations.

```
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs))
```

```
Mdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 1
                  Metrics: [2×2 table]
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
        DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

```
haveTrainedAllClasses = numel(unique(Y(1:initobs))) == 5
```

```
haveTrainedAllClasses = logical
   1
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object. The model is warm (`IsWarm` is `1`) because all the following conditions apply:

- The initial training data contains all expected classes (`haveTrainedAllClasses` is `true`).
- `Mdl` was fit to `Mdl.MetricsWarmupPeriod` observations.

Therefore, the model is prepared to generate predictions, and incremental learning functions measure performance metrics within the model.

Perform Incremetnal Learning with Conditional Training

Suppose that you want to train the model only when the most recent 1000 observations have a misclassification error greater than 5%.

Perform incremental learning, with conditional training, by following this procedure for each iteration:

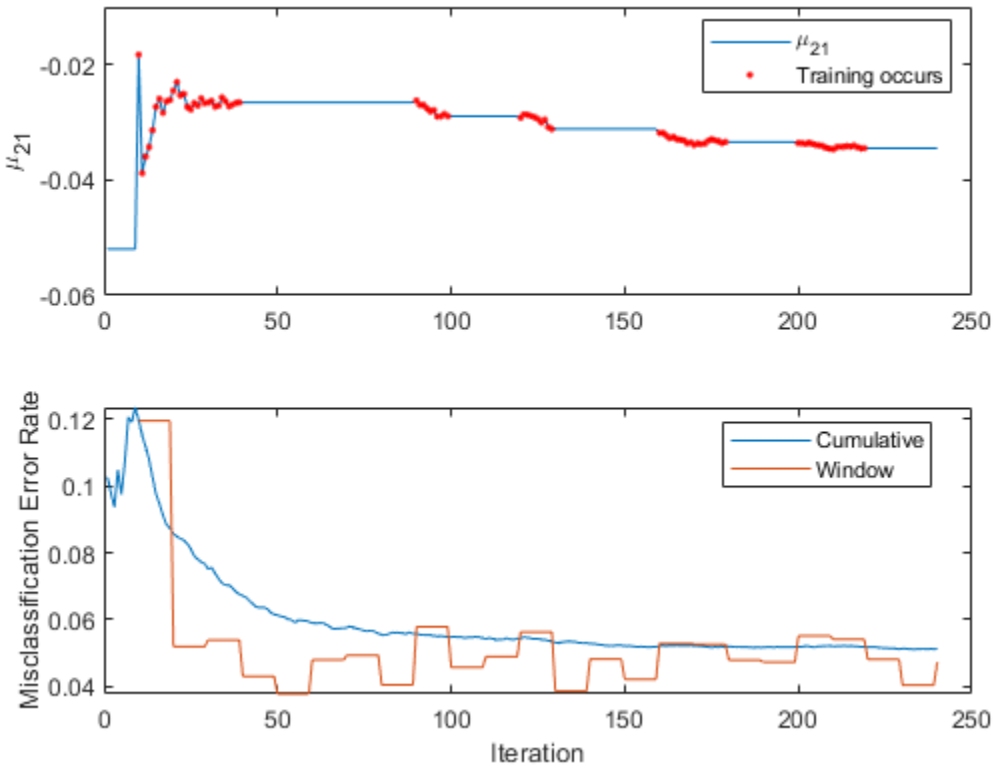
1. Simulate a data stream by processing a chunk of 100 observations at a time.
2. Update the model performance by passing the model and current chunk of data to `updateMetrics`. Overwrite the input model with the output model.
3. Fit the model to the chunk of data only when the misclassification error rate is greater than 0.05. Overwrite the input model with the output model when training occurs.
4. Store the misclassification error rate and the mean of the first predictor in the second class μ_{21} to see how they evolve during training.
5. Track when `fit` trains the model.

```
% Preallocation
numObsPerChunk = 100;
nchunk = floor((n - initobs)/numObsPerChunk);
mu21 = zeros(nchunk,1);
ce = array2table(nan(nchunk,2),'VariableNames',["Cumulative" "Window"]);
trained = false(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    if ce{j,"Window"} > 0.05
        Mdl = fit(Mdl,X(idx,:),Y(idx));
        trained(j) = true;
    end
    mu21(j) = Mdl.DistributionParameters{2,1}(1);
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the model performance and μ_{21} evolved during training, plot them on separate subplots. Identify periods during which the model was trained.

```
subplot(2,1,1)
plot(mu21)
hold on
plot(find(trained),mu21(trained),'r.')
ylabel('\mu_{21}')
legend('\mu_{21}','Training occurs','Location','best')
hold off
subplot(2,1,2)
plot(ce.Variables)
ylabel('Misclassification Error Rate')
xlabel('Iteration')
legend(ce.Properties.VariableNames,'Location','best')
```



The trace plot of μ_{21} shows periods of constant values, during which the model performance within the previous 1000 observation window is at most 0.05.

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `predict` | `updateMetrics`

More About

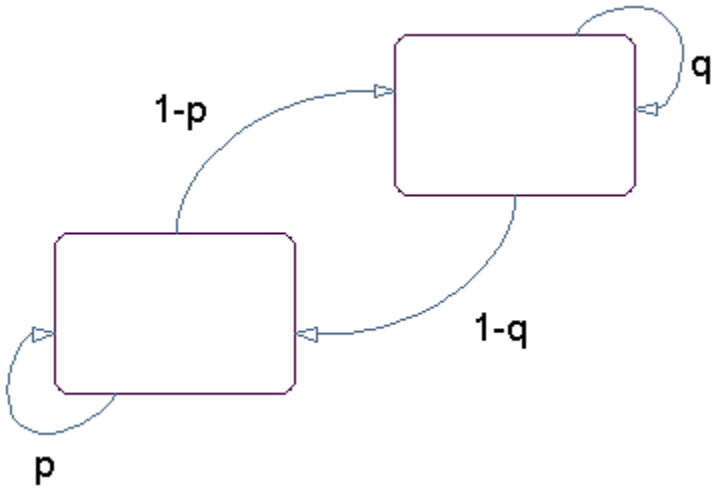
- “Configure Incremental Learning Model” on page 26-8
- “Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Markov Models

Markov Chains

Markov processes are examples of stochastic processes—processes that generate random sequences of outcomes or states according to certain probabilities. Markov processes are distinguished by being memoryless—their next state depends only on their current state, not on the history that led them there. Models of Markov processes are used in a wide variety of applications, from daily stock prices to the positions of genes in a chromosome.

A Markov model is given visual representation with a *state diagram*, such as the one below.



State Diagram for a Markov Model

The rectangles in the diagram represent the possible states of the process you are trying to model, and the arrows represent transitions between states. The label on each arrow represents the probability of that transition. At each step of the process, the model may generate an output, or emission, depending on which state it is in, and then make a transition to another state. An important characteristic of Markov models is that the next state depends only on the current state, and not on the history of transitions that lead to the current state.

For example, for a sequence of coin tosses the two states are heads and tails. The most recent coin toss determines the current state of the model and each subsequent toss determines the transition to the next state. If the coin is fair, the transition probabilities are all 1/2. The emission might simply be the current state. In more complicated models, random processes at each state will generate emissions. You could, for example, roll a die to determine the emission at any step.

Markov chains are mathematical descriptions of Markov models with a discrete set of states. Markov chains are characterized by:

- A set of states $\{1, 2, ..., M\}$
- An M-by-M transition matrix T whose i,j entry is the probability of a transition from state i to state j. The sum of the entries in each row of T must be 1, because this is the sum of the probabilities of making a transition from a given state to each of the other states.
- A set of possible outputs, or *emissions*, $\{s_1, s_2, ... , s_N\}$. By default, the set of emissions is $\{1, 2, ... , N\}$, where N is the number of possible emissions, but you can choose a different set of numbers or symbols.

- An M-by-N emission matrix E whose i,k entry gives the probability of emitting symbol s_k given that the model is in state i.

Markov chains begin in an initial state i_0 at step 0. The chain then transitions to state i_1 with probability T_{1i_1}, and emits an output s_{k_1} with probability $E_{i_1k_1}$. Consequently, the probability of observing the sequence of states $i_1i_2...i_r$ and the sequence of emissions $s_{k_1}s_{k_2}...s_{k_r}$ in the first r steps, is

$$T_{1i_1}E_{i_1k_1}T_{i_1i_2}E_{i_2k_2}...T_{i_{r-1}i_r}E_{i_rk}$$

See Also

Related Examples

- “Hidden Markov Models (HMM)” on page 27-4

Hidden Markov Models (HMM)

In this section...

Introduction to Hidden Markov Models (HMM)

A hidden Markov model (HMM) is one in which you observe a sequence of emissions, but do not know the sequence of states the model went through to generate the emissions. Analyses of hidden Markov models seek to recover the sequence of states from the observed data.

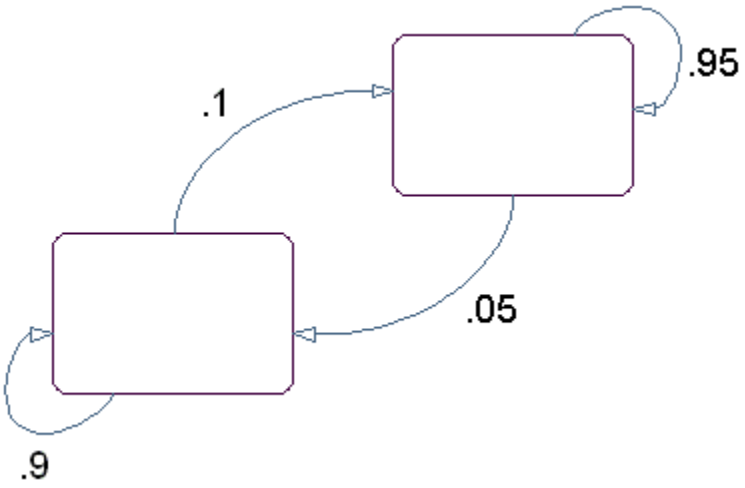
As an example, consider a Markov model with two states and six possible emissions. The model uses:

- A red die, having six sides, labeled 1 through 6.
- A green die, having twelve sides, five of which are labeled 2 through 6, while the remaining seven sides are labeled 1.
- A weighted red coin, for which the probability of heads is .9 and the probability of tails is .1.
- A weighted green coin, for which the probability of heads is .95 and the probability of tails is .05.

The model creates a sequence of numbers from the set {1, 2, 3, 4, 5, 6} with the following rules:

- Begin by rolling the red die and writing down the number that comes up, which is the emission.
- Toss the red coin and do one of the following:
  - If the result is heads, roll the red die and write down the result.
  - If the result is tails, roll the green die and write down the result.
- At each subsequent step, you flip the coin that has the same color as the die you rolled in the previous step. If the coin comes up heads, roll the same die as in the previous step. If the coin comes up tails, switch to the other die.

The state diagram for this model has two states, red and green, as shown in the following figure.



You determine the emission from a state by rolling the die with the same color as the state. You determine the transition to the next state by flipping the coin with the same color as the state.

The transition matrix is:

$$T = \begin{bmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{bmatrix}$$

The emissions matrix is:

$$E = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{7}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \end{bmatrix}$$

The model is not hidden because you know the sequence of states from the colors of the coins and dice. Suppose, however, that someone else is generating the emissions without showing you the dice or the coins. All you see is the sequence of emissions. If you start seeing more 1s than other numbers, you might suspect that the model is in the green state, but you cannot be sure because you cannot see the color of the die being rolled.

Hidden Markov models raise the following questions:

- Given a sequence of emissions, what is the most likely state path?
- Given a sequence of emissions, how can you estimate transition and emission probabilities of the model?
- What is the *forward probability* that the model generates a given sequence?
- What is the *posterior probability* that the model is in a particular state at any point in the sequence?

Analyzing Hidden Markov Models

Statistics and Machine Learning Toolbox functions related to hidden Markov models are:

- `hmmgenerate` — Generates a sequence of states and emissions from a Markov model
- `hmmestimate` — Calculates maximum likelihood estimates of transition and emission probabilities from a sequence of emissions and a known sequence of states
- `hmmtrain` — Calculates maximum likelihood estimates of transition and emission probabilities from a sequence of emissions
- `hmmviterbi` — Calculates the most probable state path for a hidden Markov model
- `hmmdecode` — Calculates the posterior state probabilities of a sequence of emissions

This section shows how to use these functions to analyze hidden Markov models.

Generating a Test Sequence

The following commands create the transition and emission matrices for the model described in the “Introduction to Hidden Markov Models (HMM)” on page 27-4:

```
TRANS = [.9 .1; .05 .95];

EMIS = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6;...
7/12, 1/12, 1/12, 1/12, 1/12, 1/12];
```

To generate a random sequence of states and emissions from the model, use `hmmgenerate`:

```
[seq,states] = hmmgenerate(1000,TRANS,EMIS);
```

The output `seq` is the sequence of emissions and the output `states` is the sequence of states.

`hmmgenerate` begins in state 1 at step 0, makes the transition to state i_1 at step 1, and returns i_1 as the first entry in `states`. To change the initial state, see “Changing the Initial State Distribution” on page 27-8.

Estimating the State Sequence

Given the transition and emission matrices `TRANS` and `EMIS`, the function `hmmviterbi` uses the Viterbi algorithm to compute the most likely sequence of states the model would go through to generate a given sequence `seq` of emissions:

```
likelystates = hmmviterbi(seq, TRANS, EMIS);
```

`likelystates` is a sequence the same length as `seq`.

To test the accuracy of `hmmviterbi`, compute the percentage of the actual sequence `states` that agrees with the sequence `likelystates`.

```
sum(states==likelystates)/1000
ans =
   0.8200
```

In this case, the most likely sequence of states agrees with the random sequence 82% of the time.

Estimating Transition and Emission Matrices

The functions `hmmestimate` and `hmmtrain` estimate the transition and emission matrices `TRANS` and `EMIS` given a sequence `seq` of emissions.

Using hmmestimate

The function `hmmestimate` requires that you know the sequence of states `states` that the model went through to generate `seq`.

The following takes the emission and state sequences and returns estimates of the transition and emission matrices:

```
[TRANS_EST, EMIS_EST] = hmmestimate(seq, states)
```

```
TRANS_EST =
0.8989    0.1011
0.0585    0.9415

EMIS_EST =
0.1721    0.1721    0.1749    0.1612    0.1803    0.1393
0.5836    0.0741    0.0804    0.0789    0.0726    0.1104
```

You can compare the outputs with the original transition and emission matrices, `TRANS` and `EMIS`:

```
TRANS
TRANS =
0.9000    0.1000
0.0500    0.9500

EMIS
EMIS =
0.1667    0.1667    0.1667    0.1667    0.1667    0.1667
0.5833    0.0833    0.0833    0.0833    0.0833    0.0833
```

Using hmmtrain

If you do not know the sequence of states `states`, but you have initial guesses for `TRANS` and `EMIS`, you can still estimate `TRANS` and `EMIS` using `hmmtrain`.

Suppose you have the following initial guesses for `TRANS` and `EMIS`.

```
TRANS_GUESS = [.85 .15; .1 .9];
EMIS_GUESS = [.17 .16 .17 .16 .17 .17;.6 .08 .08 .08 .08 08];
```

You estimate `TRANS` and `EMIS` as follows:

```
[TRANS_EST2, EMIS_EST2] = hmmtrain(seq, TRANS_GUESS, EMIS_GUESS)

TRANS_EST2 =
0.2286    0.7714
0.0032    0.9968

EMIS_EST2 =
0.1436    0.2348    0.1837    0.1963    0.2350    0.0066
0.4355    0.1089    0.1144    0.1082    0.1109    0.1220
```

`hmmtrain` uses an iterative algorithm that alters the matrices `TRANS_GUESS` and `EMIS_GUESS` so that at each step the adjusted matrices are more likely to generate the observed sequence, `seq`. The algorithm halts when the matrices in two successive iterations are within a small tolerance of each other.

If the algorithm fails to reach this tolerance within a maximum number of iterations, whose default value is `100`, the algorithm halts. In this case, `hmmtrain` returns the last values of `TRANS_EST` and `EMIS_EST` and issues a warning that the tolerance was not reached.

If the algorithm fails to reach the desired tolerance, increase the default value of the maximum number of iterations with the command:

```
hmmtrain(seq,TRANS_GUESS,EMIS_GUESS,'maxiterations',maxiter)
```

where `maxiter` is the maximum number of steps the algorithm executes.

Change the default value of the tolerance with the command:

```
hmmtrain(seq, TRANS_GUESS, EMIS_GUESS, 'tolerance', tol)
```

where `tol` is the desired value of the tolerance. Increasing the value of `tol` makes the algorithm halt sooner, but the results are less accurate.

Two factors reduce the reliability of the output matrices of `hmmtrain`:

- The algorithm converges to a local maximum that does not represent the true transition and emission matrices. If you suspect this, use different initial guesses for the matrices `TRANS_EST` and `EMIS_EST`.
- The sequence `seq` may be too short to properly train the matrices. If you suspect this, use a longer sequence for `seq`.

Estimating Posterior State Probabilities

The posterior state probabilities of an emission sequence `seq` are the conditional probabilities that the model is in a particular state when it generates a symbol in `seq`, given that `seq` is emitted. You compute the posterior state probabilities with `hmmdecode`:

```
PSTATES = hmmdecode(seq,TRANS,EMIS)
```

The output `PSTATES` is an M-by-L matrix, where M is the number of states and L is the length of `seq`. `PSTATES(i,j)` is the conditional probability that the model is in state `i` when it generates the `j`th symbol of `seq`, given that `seq` is emitted.

`hmmdecode` begins with the model in state 1 at step 0, prior to the first emission. `PSTATES(i,1)` is the probability that the model is in state i at the following step 1. To change the initial state, see “Changing the Initial State Distribution” on page 27-8.

To return the logarithm of the probability of the sequence `seq`, use the second output argument of `hmmdecode`:

```
[PSTATES,logpseq] = hmmdecode(seq,TRANS,EMIS)
```

The probability of a sequence tends to 0 as the length of the sequence increases, and the probability of a sufficiently long sequence becomes less than the smallest positive number your computer can represent. `hmmdecode` returns the logarithm of the probability to avoid this problem.

Changing the Initial State Distribution

By default, Statistics and Machine Learning Toolbox hidden Markov model functions begin in state 1. In other words, the distribution of initial states has all of its probability mass concentrated at state 1. To assign a different distribution of probabilities, $p = [p_1, p_2, ..., p_M]$, to the M initial states, do the following:

1. Create an M+1-by-M+1 augmented transition matrix, $\widehat{T}$ of the following form:

   $$\widehat{T} = \begin{bmatrix} 0 & p \\ 0 & T \end{bmatrix}$$

   where T is the true transition matrix. The first column of $\widehat{T}$ contains M+1 zeros. p must sum to 1.

2. Create an M+1-by-N augmented emission matrix, $\widehat{E}$, that has the following form:

$$\widehat{E} = \begin{bmatrix} 0 \\ E \end{bmatrix}$$

If the transition and emission matrices are `TRANS` and `EMIS`, respectively, you create the augmented matrices with the following commands:

```
TRANS_HAT = [0 p; zeros(size(TRANS,1),1) TRANS];

EMIS_HAT = [zeros(1,size(EMIS,2)); EMIS];
```

See Also

`hmmdecode` | `hmmestimate` | `hmmgenerate` | `hmmtrain` | `hmmviterbi`

More About

- “Markov Chains” on page 27-2

Design of Experiments

Design of Experiments

Passive data collection leads to a number of problems in statistical modeling. Observed changes in a response variable may be correlated with, but not caused by, observed changes in individual factors (process variables). Simultaneous changes in multiple factors may produce interactions that are difficult to separate into individual effects. Observations may be dependent, while a model of the data considers them to be independent.

Designed experiments address these problems. In a designed experiment, the data-producing process is actively manipulated to improve the quality of information and to eliminate redundant data. A common goal of all experimental designs is to collect data as parsimoniously as possible while providing sufficient information to accurately estimate model parameters.

For example, a simple model of a response y in an experiment with two controlled factors x_1 and x_2 might look like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Here ε includes both experimental error and the effects of any uncontrolled factors in the experiment. The terms $\beta_1 x_1$ and $\beta_2 x_2$ are main effects and the term $\beta_3 x_1 x_2$ is a two-way interaction effect. A designed experiment would systematically manipulate x_1 and x_2 while measuring y, with the objective of accurately estimating β_0, β_1, β_2, and β_3.

Full Factorial Designs

In this section...

Multilevel Designs

To systematically vary experimental factors, assign each factor a discrete set of levels. Full factorial designs measure response variables using every treatment (combination of the factor levels). A full factorial design for n factors with N_1, ..., N_n levels requires $N_1 \times ... \times N_n$ experimental runs—one for each treatment. While advantageous for separating individual effects, full factorial designs can make large demands on data collection.

As an example, suppose a machine shop has three machines and four operators. If the same operator always uses the same machine, it is impossible to determine if a machine or an operator is the cause of variation in production. By allowing every operator to use every machine, effects are separated. A full factorial list of treatments is generated by the function `fullfact`:

```
dFF = fullfact([3,4])
dFF =
     1     1
     2     1
     3     1
     1     2
     2     2
     3     2
     1     3
     2     3
     3     3
     1     4
     2     4
     3     4
```

Each of the 3×4 = 12 rows of `dFF` represent one machine/operator combination.

Two-Level Designs

Many experiments can be conducted with two-level factors, using two-level designs. For example, suppose the machine shop in the previous example always keeps the same operator on the same machine, but wants to measure production effects that depend on the composition of the day and night shifts. The function `ff2n` generates a full factorial list of treatments:

```
dFF2 = ff2n(4)
dFF2 =
     0     0     0     0
     0     0     0     1
     0     0     1     0
     0     0     1     1
     0     1     0     0
     0     1     0     1
     0     1     1     0
     0     1     1     1
```

```
1     0     0     0
1     0     0     1
1     0     1     0
1     0     1     1
1     1     0     0
1     1     0     1
1     1     1     0
1     1     1     1
```

Each of the $2^4 = 16$ rows of `dFF2` represent one schedule of operators for the day (`0`) and night (`1`) shifts.

Fractional Factorial Designs

In this section...

Introduction to Fractional Factorial Designs

Two-level designs are sufficient for evaluating many production processes. Factor levels of ±1 can indicate categorical factors, normalized factor extremes, or simply “up” and “down” from current factor settings. Experimenters evaluating process *changes* are interested primarily in the factor directions that lead to process improvement.

For experiments with many factors, two-level full factorial designs can lead to large amounts of data. For example, a two-level full factorial design with 10 factors requires $2^{10} = 1024$ runs. Often, however, individual factors or their interactions have no distinguishable effects on a response. This is especially true of higher order interactions. As a result, a well-designed experiment can use fewer runs for estimating model parameters.

Fractional factorial designs use a fraction of the runs required by full factorial designs. A subset of experimental treatments is selected based on an evaluation (or assumption) of which factors and interactions have the most significant effects. Once this selection is made, the experimental design must separate these effects. In particular, significant effects should not be confounded, that is, the measurement of one should not depend on the measurement of another.

Plackett-Burman Designs

Plackett-Burman designs are used when only main effects are considered significant. Two-level Plackett-Burman designs require a number of experimental runs that are a multiple of 4 rather than a power of 2. The function `hadamard` generates these designs:

```
dPB = hadamard(8)
dPB =
     1    1    1    1    1    1    1    1
     1   -1    1   -1    1   -1    1   -1
     1    1   -1   -1    1    1   -1   -1
     1   -1   -1    1    1   -1   -1    1
     1    1    1    1   -1   -1   -1   -1
     1   -1    1   -1   -1    1   -1    1
     1    1   -1   -1   -1   -1    1    1
     1   -1   -1    1   -1    1    1   -1
```

Binary factor levels are indicated by ±1. The design is for eight runs (the rows of `dPB`) manipulating seven two-level factors (the last seven columns of `dPB`). The number of runs is a fraction $8/2^7 = 0.0625$ of the runs required by a full factorial design. Economy is achieved at the expense of confounding main effects with any two-way interactions.

General Fractional Designs

At the cost of a larger fractional design, you can specify which interactions you wish to consider significant. A design of resolution *R* is one in which no *n*-factor interaction is confounded with any

other effect containing less than $R - n$ factors. Thus, a resolution III design does not confound main effects with one another but may confound them with two-way interactions (as in “Plackett-Burman Designs” on page 28-5), while a resolution IV design does not confound either main effects or two-way interactions but may confound two-way interactions with each other.

Specify general fractional factorial designs using a full factorial design for a selected subset of basic factors and generators for the remaining factors. Generators are products of the basic factors, giving the levels for the remaining factors. Use the function `fracfact` to generate these designs:

```
dfF = fracfact('a b c d bcd acd')
dfF =
    -1    -1    -1    -1    -1    -1
    -1    -1    -1     1     1     1
    -1    -1     1    -1     1     1
    -1    -1     1     1    -1    -1
    -1     1    -1    -1     1    -1
    -1     1    -1     1    -1     1
    -1     1     1    -1    -1     1
    -1     1     1     1     1    -1
     1    -1    -1    -1    -1     1
     1    -1    -1     1     1    -1
     1    -1     1    -1     1    -1
     1    -1     1     1    -1     1
     1     1    -1    -1     1     1
     1     1    -1     1    -1    -1
     1     1     1    -1    -1    -1
     1     1     1     1     1     1
```

This is a six-factor design in which four two-level basic factors (`a`, `b`, `c`, and `d` in the first four columns of `dfF`) are measured in every combination of levels, while the two remaining factors (in the last three columns of `dfF`) are measured only at levels defined by the generators `bcd` and `acd`, respectively. Levels in the generated columns are products of corresponding levels in the columns that make up the generator.

The challenge of creating a fractional factorial design is to choose basic factors and generators so that the design achieves a specified resolution in a specified number of runs. Use the function `fracfactgen` to find appropriate generators:

```
generators = fracfactgen('a b c d e f',4,4)
generators =
    'a'
    'b'
    'c'
    'd'
    'bcd'
    'acd'
```

These are generators for a six-factor design with factors `a` through `f`, using $2^4 = 16$ runs to achieve resolution IV. The `fracfactgen` function uses an efficient search algorithm to find generators that meet the requirements.

An optional output from `fracfact` displays the confounding pattern of the design:

```
[dfF,confounding] = fracfact(generators);
confounding
confounding =
    'Term'     'Generator'    'Confounding'
```

```
'X1'       'a'            'X1'
'X2'       'b'            'X2'
'X3'       'c'            'X3'
'X4'       'd'            'X4'
'X5'       'bcd'          'X5'
'X6'       'acd'          'X6'
'X1*X2'    'ab'           'X1*X2 + X5*X6'
'X1*X3'    'ac'           'X1*X3 + X4*X6'
'X1*X4'    'ad'           'X1*X4 + X3*X6'
'X1*X5'    'abcd'         'X1*X5 + X2*X6'
'X1*X6'    'cd'           'X1*X6 + X2*X5 + X3*X4'
'X2*X3'    'bc'           'X2*X3 + X4*X5'
'X2*X4'    'bd'           'X2*X4 + X3*X5'
'X2*X5'    'cd'           'X1*X6 + X2*X5 + X3*X4'
'X2*X6'    'abcd'         'X1*X5 + X2*X6'
'X3*X4'    'cd'           'X1*X6 + X2*X5 + X3*X4'
'X3*X5'    'bd'           'X2*X4 + X3*X5'
'X3*X6'    'ad'           'X1*X4 + X3*X6'
'X4*X5'    'bc'           'X2*X3 + X4*X5'
'X4*X6'    'ac'           'X1*X3 + X4*X6'
'X5*X6'    'ab'           'X1*X2 + X5*X6'
```

The confounding pattern shows that main effects are effectively separated by the design, but two-way interactions are confounded with various other two-way interactions.

Response Surface Designs

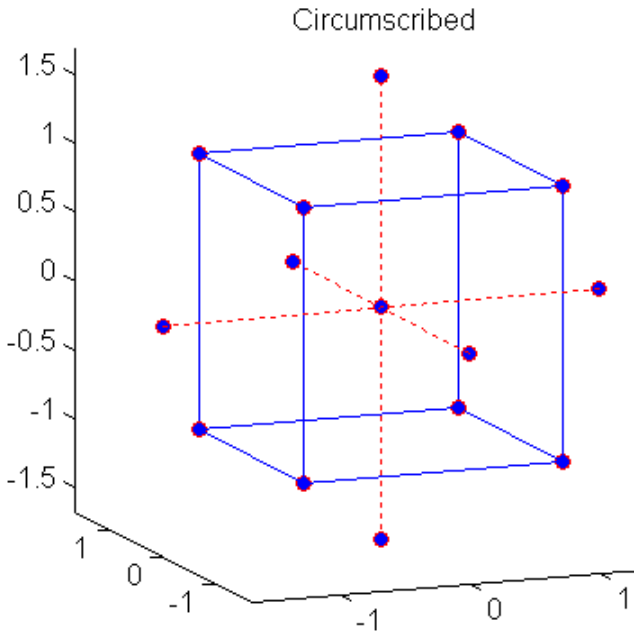
In this section...

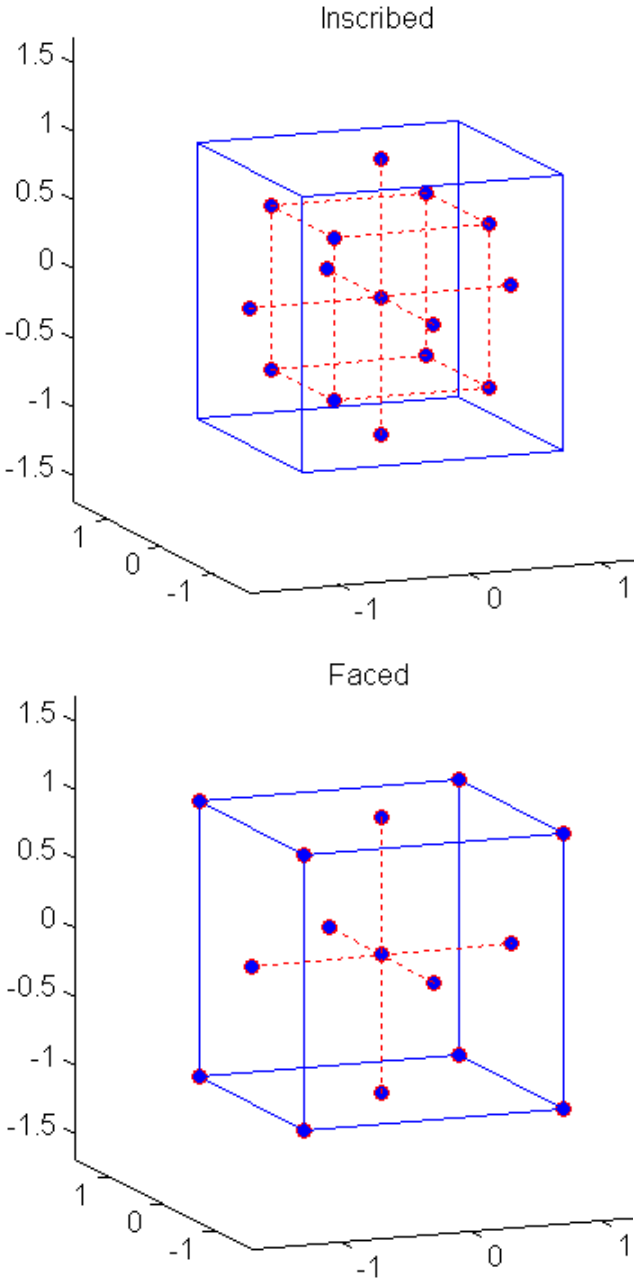
Introduction to Response Surface Designs

Quadratic response surfaces are simple models that provide a maximum or minimum without making additional assumptions about the form of the response. Quadratic models can be calibrated using full factorial designs with three or more levels for each factor, but these designs generally require more runs than necessary to accurately estimate model parameters. This section discusses designs for calibrating quadratic models that are much more efficient, using three or five levels for each factor, but not using all combinations of levels.

Central Composite Designs

Central composite designs (CCDs), also known as Box-Wilson designs, are appropriate for calibrating full quadratic models. There are three types of CCDs—circumscribed, inscribed, and faced—pictured below:





Each design consists of a factorial design (the corners of a cube) together with center and star points that allow for estimation of second-order effects. For a full quadratic model with n factors, CCDs have enough design points to estimate the $(n+2)(n+1)/2$ coefficients in a full quadratic model with n factors.

The type of CCD used (the position of the factorial and star points) is determined by the number of factors and by the desired properties of the design. The following table summarizes some important properties. A design is rotatable if the prediction variance depends only on the distance of the design point from the center of the design.

Design	Rotatable	Factor Levels	Uses Points Outside ±1	Accuracy of Estimates
Circumscribed (CCC)	Yes	5	Yes	Good over entire design space
Inscribed (CCI)	Yes	5	No	Good over central subset of design space
Faced (CCF)	No	3	No	Fair over entire design space; poor for pure quadratic coefficients

Generate CCDs with the function `ccdesign`:

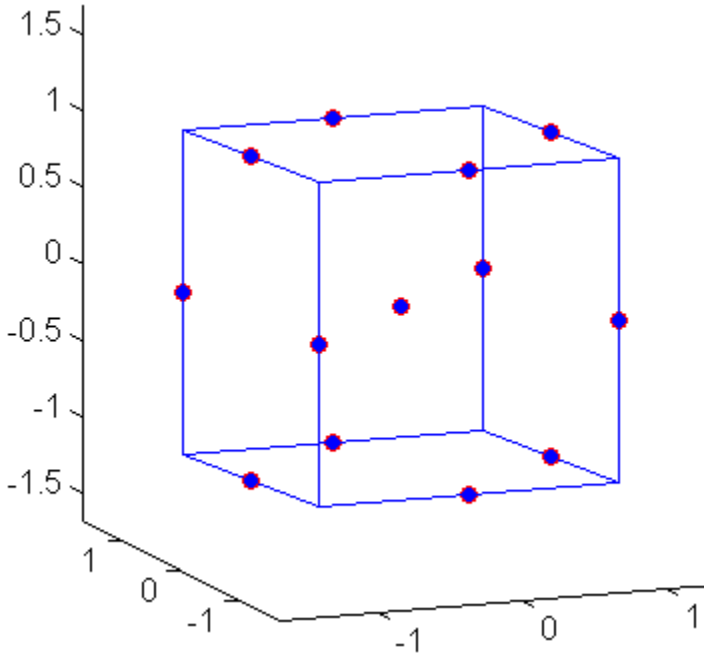
```
dCC = ccdesign(3,'type','circumscribed')
dCC =
   -1.0000   -1.0000   -1.0000
   -1.0000   -1.0000    1.0000
   -1.0000    1.0000   -1.0000
   -1.0000    1.0000    1.0000
    1.0000   -1.0000   -1.0000
    1.0000   -1.0000    1.0000
    1.0000    1.0000   -1.0000
    1.0000    1.0000    1.0000
   -1.6818         0         0
    1.6818         0         0
         0   -1.6818         0
         0    1.6818         0
         0         0   -1.6818
         0         0    1.6818
         0         0         0
         0         0         0
         0         0         0
         0         0         0
         0         0         0
         0         0         0
         0         0         0
         0         0         0
         0         0         0
```

The repeated center point runs allow for a more uniform estimate of the prediction variance over the entire design space.

Box-Behnken Designs

Like the designs described in “Central Composite Designs” on page 28-8, Box-Behnken designs are used to calibrate full quadratic models. Box-Behnken designs are rotatable and, for a small number of factors (four or less), require fewer runs than CCDs. By avoiding the corners of the design space, they allow experimenters to work around extreme factor combinations. Like an inscribed CCD, however, extremes are then poorly estimated.

The geometry of a Box-Behnken design is pictured in the following figure.



Design points are at the midpoints of edges of the design space and at the center, and do not contain an embedded factorial design.

Generate Box-Behnken designs with the function `bbdesign`:

```
dBB = bbdesign(3)
dBB =
    -1    -1     0
    -1     1     0
     1    -1     0
     1     1     0
    -1     0    -1
    -1     0     1
     1     0    -1
     1     0     1
     0    -1    -1
     0    -1     1
     0     1    -1
     0     1     1
     0     0     0
     0     0     0
     0     0     0
```

Again, the repeated center point runs allow for a more uniform estimate of the prediction variance over the entire design space.

D-Optimal Designs

In this section...

Introduction to D-Optimal Designs

Traditional experimental designs (“Full Factorial Designs” on page 28-3, “Fractional Factorial Designs” on page 28-5, and “Response Surface Designs” on page 28-8) are appropriate for calibrating linear models in experimental settings where factors are relatively unconstrained in the region of interest. In some cases, however, models are necessarily nonlinear. In other cases, certain treatments (combinations of factor levels) may be expensive or infeasible to measure. D-optimal designs are model-specific designs that address these limitations of traditional designs.

A D-optimal design is generated by an iterative search algorithm and seeks to minimize the covariance of the parameter estimates for a specified model. This is equivalent to maximizing the determinant $D = |X^TX|$, where X is the design matrix of model terms (the columns) evaluated at specific treatments in the design space (the rows). Unlike traditional designs, D-optimal designs do not require orthogonal design matrices, and as a result, parameter estimates may be correlated. Parameter estimates may also be locally, but not globally, D-optimal.

There are several Statistics and Machine Learning Toolbox functions for generating D-optimal designs:

Function	Description
`candexch`	Uses a row-exchange algorithm to generate a D-optimal design with a specified number of runs for a specified model and a specified candidate set. This is the second component of the algorithm used by `rowexch`.
`candgen`	Generates a candidate set for a specified model. This is the first component of the algorithm used by `rowexch`.
`cordexch`	Uses a coordinate-exchange algorithm to generate a D-optimal design with a specified number of runs for a specified model.
`daugment`	Uses a coordinate-exchange algorithm to augment an existing D-optimal design with additional runs to estimate additional model terms.
`dcovary`	Uses a coordinate-exchange algorithm to generate a D-optimal design with fixed covariate factors.
`rowexch`	Uses a row-exchange algorithm to generate a D-optimal design with a specified number of runs for a specified model. The algorithm calls `candgen` and then `candexch`. (Call `candexch` separately to specify a candidate set.)

The following sections explain how to use these functions to generate D-optimal designs.

Note The function `rsmdemo` generates simulated data for experimental settings specified by either the user or by a D-optimal design generated by `cordexch`. It uses the `rstool` interface to visualize response surface models fit to the data, and it uses the `nlintool` interface to visualize a nonlinear model fit to the data.

Generate D-Optimal Designs

Two Statistics and Machine Learning Toolbox algorithms generate D-optimal designs:

- The `cordexch` function uses a coordinate-exchange algorithm
- The `rowexch` function uses a row-exchange algorithm

Both `cordexch` and `rowexch` use iterative search algorithms. They operate by incrementally changing an initial design matrix X to increase $D = |X^TX|$ at each step. In both algorithms, there is randomness built into the selection of the initial design and into the choice of the incremental changes. As a result, both algorithms may return locally, but not globally, D-optimal designs. Run each algorithm multiple times and select the best result for your final design. Both functions have a `'tries'` parameter that automates this repetition and comparison.

At each step, the row-exchange algorithm exchanges an entire row of X with a row from a design matrix C evaluated at a candidate set of feasible treatments. The `rowexch` function automatically generates a C appropriate for a specified model, operating in two steps by calling the `candgen` and `candexch` functions in sequence. Provide your own C by calling `candexch` directly. In either case, if C is large, its static presence in memory can affect computation.

The coordinate-exchange algorithm, by contrast, does not use a candidate set. (Or rather, the candidate set is the entire design space.) At each step, the coordinate-exchange algorithm exchanges a single element of X with a new element evaluated at a neighboring point in design space. The absence of a candidate set reduces demands on memory, but the smaller scale of the search means that the coordinate-exchange algorithm is more likely to become trapped in a local minimum than the row-exchange algorithm.

For example, suppose you want a design to estimate the parameters in the following three-factor, seven-term interaction model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon$$

Use `cordexch` to generate a D-optimal design with seven runs:

```
nfactors = 3;
nruns = 7;
[dCE,X] = cordexch(nfactors,nruns,'interaction','tries',10)
dCE =
    -1     1     1
    -1    -1    -1
     1     1     1
    -1     1    -1
     1    -1     1
     1    -1    -1
    -1    -1     1
X =
     1    -1     1     1    -1    -1     1
     1    -1    -1    -1     1     1     1
     1     1     1     1     1     1     1
```

```
     1    -1     1    -1    -1     1    -1
     1     1    -1     1    -1     1    -1
     1     1    -1    -1    -1    -1     1
     1    -1    -1     1     1    -1    -1
```

Columns of the design matrix `X` are the model terms evaluated at each row of the design `dCE`. The terms appear in order from left to right:

1. Constant term
2. Linear terms (1, 2, 3)
3. Interaction terms (12, 13, 23)

Use `X` in a linear regression model fit to response data measured at the design points in `dCE`.

Use `rowexch` in a similar fashion to generate an equivalent design:

```
[dRE,X] = rowexch(nfactors,nruns,'interaction','tries',10)
dRE =
    -1    -1     1
     1    -1     1
     1    -1    -1
     1     1     1
    -1    -1    -1
    -1     1    -1
    -1     1     1
X =
     1    -1    -1     1     1    -1    -1
     1     1    -1     1    -1     1    -1
     1     1    -1    -1    -1    -1     1
     1     1     1     1     1     1     1
     1    -1    -1    -1     1     1     1
     1    -1     1    -1    -1     1    -1
     1    -1     1     1    -1    -1     1
```

Augment D-Optimal Designs

In practice, you may want to add runs to a completed experiment to learn more about a process and estimate additional model coefficients. The `daugment` function uses a coordinate-exchange algorithm to augment an existing D-optimal design.

For example, the following eight-run design is adequate for estimating main effects in a four-factor model:

```
dCEmain = cordexch(4,8)
dCEmain =
     1    -1    -1     1
    -1    -1     1     1
    -1     1    -1     1
     1     1     1    -1
     1     1     1     1
    -1     1    -1    -1
     1    -1    -1    -1
    -1    -1     1    -1
```

To estimate the six interaction terms in the model, augment the design with eight additional runs:

```
dCEinteraction = daugment(dCEmain,8,'interaction')
dCEinteraction =
     1    -1    -1     1
    -1    -1     1     1
    -1     1    -1     1
     1     1     1    -1
     1     1     1     1
    -1     1    -1    -1
     1    -1    -1    -1
    -1    -1     1    -1
    -1     1     1     1
    -1    -1    -1    -1
     1    -1     1    -1
     1     1    -1     1
    -1     1     1    -1
     1     1    -1    -1
     1    -1     1     1
     1     1     1    -1
```

The augmented design is full factorial, with the original eight runs in the first eight rows.

The `'start'` parameter of the `candexch` function provides the same functionality as `daugment`, but uses a row exchange algorithm rather than a coordinate-exchange algorithm.

Specify Fixed Covariate Factors

In many experimental settings, certain factors and their covariates are constrained to a fixed set of levels or combinations of levels. These cannot be varied when searching for an optimal design. The `dcovary` function allows you to specify fixed covariate factors in the coordinate exchange algorithm.

For example, suppose you want a design to estimate the parameters in a three-factor linear additive model, with eight runs that necessarily occur at different times. If the process experiences temporal linear drift, you may want to include the run time as a variable in the model. Produce the design as follows:

```
time = linspace(-1,1,8)';
[dCV,X] = dcovary(3,time,'linear')
dCV =
   -1.0000    1.0000    1.0000   -1.0000
    1.0000   -1.0000   -1.0000   -0.7143
   -1.0000   -1.0000   -1.0000   -0.4286
    1.0000   -1.0000    1.0000   -0.1429
    1.0000    1.0000   -1.0000    0.1429
   -1.0000    1.0000   -1.0000    0.4286
    1.0000    1.0000    1.0000    0.7143
   -1.0000   -1.0000    1.0000    1.0000
X =
    1.0000   -1.0000    1.0000    1.0000   -1.0000
    1.0000    1.0000   -1.0000   -1.0000   -0.7143
    1.0000   -1.0000   -1.0000   -1.0000   -0.4286
    1.0000    1.0000   -1.0000    1.0000   -0.1429
    1.0000    1.0000    1.0000   -1.0000    0.1429
    1.0000   -1.0000    1.0000   -1.0000    0.4286
    1.0000    1.0000    1.0000    1.0000    0.7143
    1.0000   -1.0000   -1.0000    1.0000    1.0000
```

The column vector `time` is a fixed factor, normalized to values between ±1. The number of rows in the fixed factor specifies the number of runs in the design. The resulting design `dCV` gives factor settings for the three controlled model factors at each time.

Specify Categorical Factors

Categorical factors take values in a discrete set of levels. Both `cordexch` and `rowexch` have a `'categorical'` parameter that allows you to specify the indices of categorical factors and a `'levels'` parameter that allows you to specify a number of levels for each factor.

For example, the following eight-run design is for a linear additive model with five factors in which the final factor is categorical with three levels:

```
dCEcat = cordexch(5,8,'linear','categorical',5,'levels',3)
dCEcat =
    -1    -1     1     1     2
    -1    -1    -1    -1     3
     1     1     1     1     3
     1     1    -1    -1     2
     1    -1    -1     1     3
    -1     1    -1     1     1
    -1     1     1    -1     3
     1    -1     1    -1     1
```

Specify Candidate Sets

The row-exchange algorithm exchanges rows of an initial design matrix X with rows from a design matrix C evaluated at a candidate set of feasible treatments. The `rowexch` function automatically generates a C appropriate for a specified model, operating in two steps by calling the `candgen` and `candexch` functions in sequence. Provide your own C by calling `candexch` directly.

For example, the following uses `rowexch` to generate a five-run design for a two-factor pure quadratic model using a candidate set that is produced internally:

```
dRE1 = rowexch(2,5,'purequadratic','tries',10)
dRE1 =
    -1     1
     0     0
     1    -1
     1     0
     1     1
```

The same thing can be done using `candgen` and `candexch` in sequence:

```
[dC,C] = candgen(2,'purequadratic') % Candidate set, C
dC =
    -1    -1
     0    -1
     1    -1
    -1     0
     0     0
     1     0
    -1     1
     0     1
     1     1
```

```
C =
     1    -1    -1     1     1
     1     0    -1     0     1
     1     1    -1     1     1
     1    -1     0     1     0
     1     0     0     0     0
     1     1     0     1     0
     1    -1     1     1     1
     1     0     1     0     1
     1     1     1     1     1
treatments = candexch(C,5,'tries',10) % D-opt subset
treatments =
     2
     1
     7
     3
     4
dRE2 = dC(treatments,:) % Display design
dRE2 =
     0    -1
    -1    -1
    -1     1
     1    -1
    -1     0
```

You can replace `C` in this example with a design matrix evaluated at your own candidate set. For example, suppose your experiment is constrained so that the two factors cannot have extreme settings simultaneously. The following produces a restricted candidate set:

```
constraint = sum(abs(dC),2) < 2; % Feasible treatments
my_dC = dC(constraint,:)
my_dC =
     0    -1
    -1     0
     0     0
     1     0
     0     1
```

Use the `x2fx` function to convert the candidate set to a design matrix:

```
my_C = x2fx(my_dC,'purequadratic')
my_C =
     1     0    -1     0     1
     1    -1     0     1     0
     1     0     0     0     0
     1     1     0     1     0
     1     0     1     0     1
```

Find the required design in the same manner:

```
my_treatments = candexch(my_C,5,'tries',10) % D-opt subset
my_treatments =
     2
     4
     5
     1
     3
my_dRE = my_dC(my_treatments,:) % Display design
```

```
my_dRE =
    -1     0
     1     0
     0     1
     0    -1
     0     0
```

Improve an Engine Cooling Fan Using Design for Six Sigma Techniques

This example shows how to improve the performance of an engine cooling fan through a Design for Six Sigma approach using Define, Measure, Analyze, Improve, and Control (DMAIC). The initial fan does not circulate enough air through the radiator to keep the engine cool during difficult conditions. First the example shows how to design an experiment to investigate the effect of three performance factors: fan distance from the radiator, blade-tip clearance, and blade pitch angle. It then shows how to estimate optimum values for each factor, resulting in a design that produces airflows beyond the goal of 875 ft^3 per minute using test data. Finally it shows how to use simulations to verify that the new design produces airflow according to the specifications in more than 99.999% of the fans manufactured. This example uses MATLAB, Statistics and Machine Learning Toolbox, and Optimization Toolbox.

Define the Problem

This example addresses an engine cooling fan design that is unable to pull enough air through the radiator to keep the engine cool during difficult conditions, such as stop-and-go traffic or hot weather). Suppose you estimate that you need airflow of at least 875 ft^3/min to keep the engine cool during difficult conditions. You need to evaluate the current design and develop an alternative design that can achieve the target airflow.

Assess Cooling Fan Performance

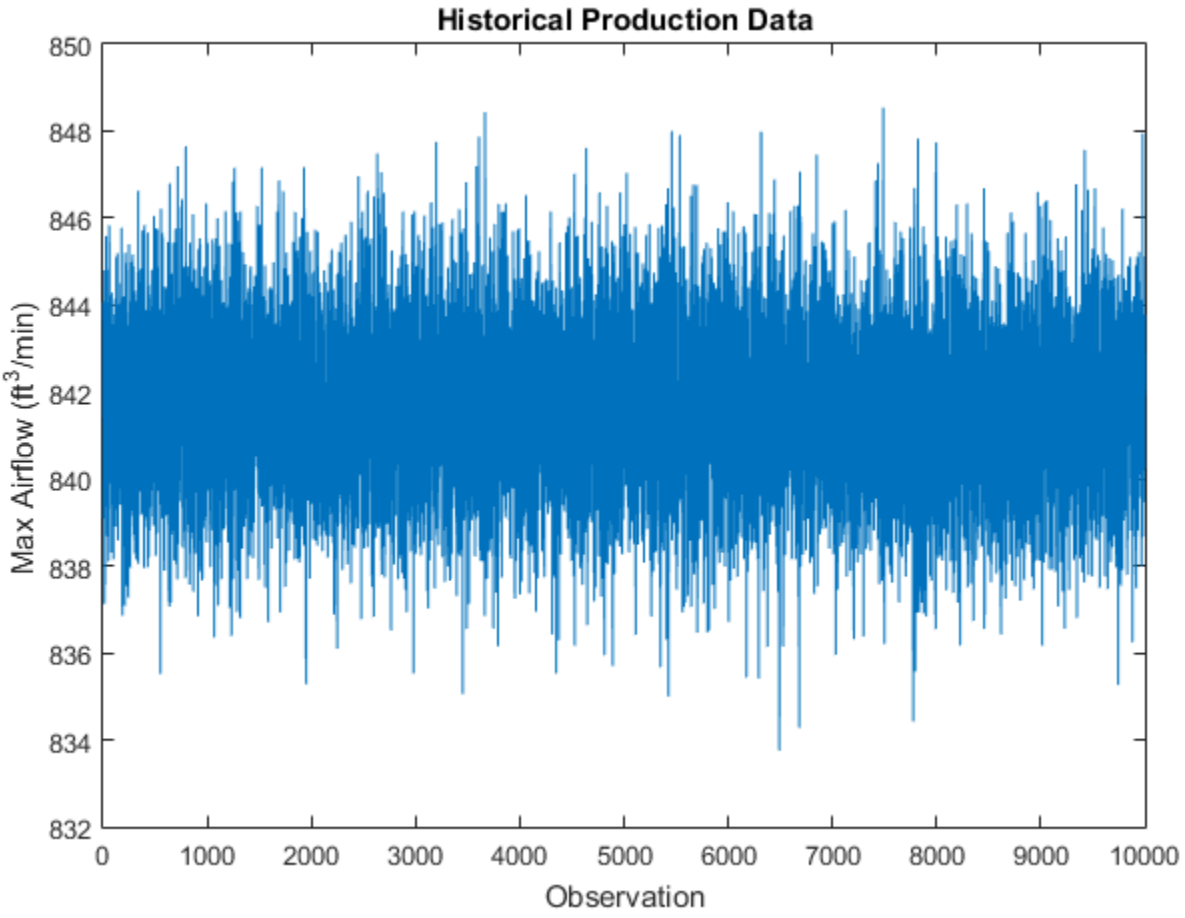
Load the sample data.

```
load(fullfile(matlabroot,'help/toolbox/stats/examples','OriginalFan.mat'))
```

The data consists of 10,000 measurements (historical production data) of the existing cooling fan performance.

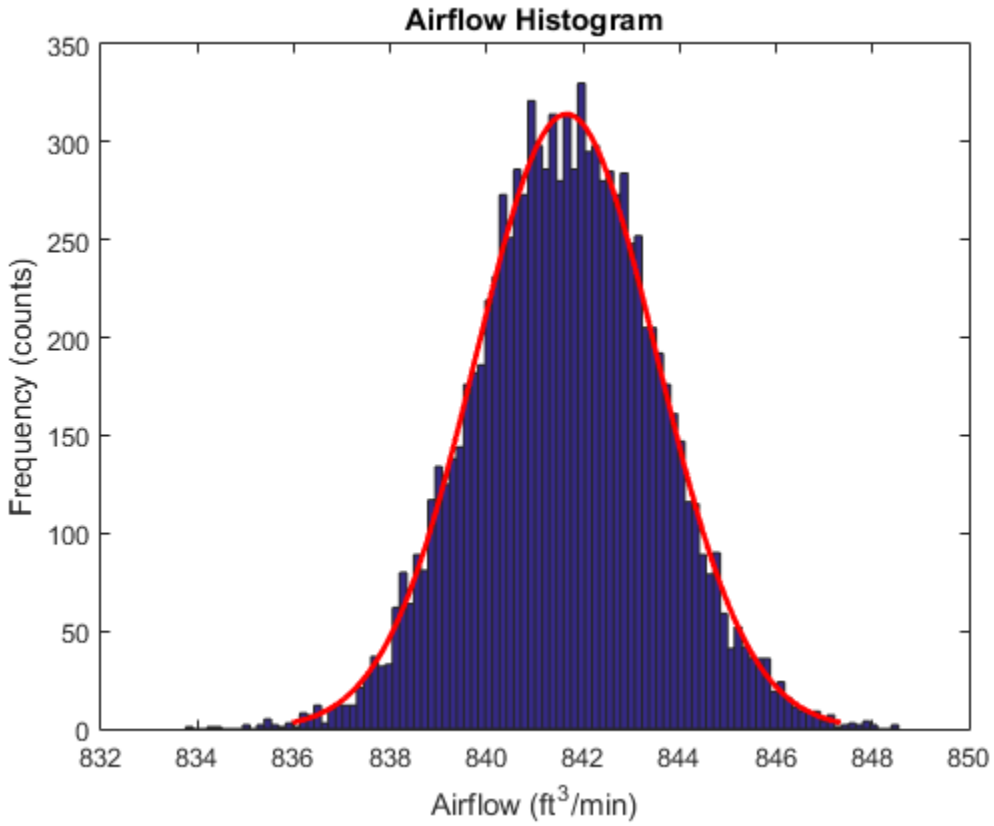
Plot the data to analyze the current fan's performance.

```
plot(originalfan)
xlabel('Observation')
ylabel('Max Airflow (ft^3/min)')
title('Historical Production Data')
```



The data is centered around 842 ft^3/min and most values fall within the range of about 8 ft^3/min. The plot does not tell much about the underlying distribution of data, however. Plot the histogram and fit a normal distribution to the data.

```
figure()
histfit(originalfan) % Plot histogram with normal distribution fit
format shortg
xlabel('Airflow (ft^3/min)')
ylabel('Frequency (counts)')
title('Airflow Histogram')
```




```
pd = fitdist(originalfan,'normal') % Fit normal distribution to data
```

```
pd =

  NormalDistribution

  Normal distribution
       mu = 841.652   [841.616, 841.689]
    sigma =  1.8768   [1.85114, 1.90318]
```

`fitdist` fits a normal distribution to data and estimates the parameters from data. The estimate for the mean airflow speed is 841.652 ft^3/min, and the 95% confidence interval for the mean airflow speed is (841.616, 841.689). This estimate makes it clear that the current fan is not close to the required 875 ft^3/min. There is need to improve the fan design to achieve the target airflow.

Determine Factors That Affect Fan Performance

Evaluate the factors that affect cooling fan performance using design of experiments (DOE). The response is the cooling fan airflow rate (ft^3/min). Suppose that the factors that you can modify and control are:

- Distance from radiator
- Pitch angle
- Blade tip clearance

In general, fluid systems have nonlinear behavior. Therefore, use a response surface design to estimate any nonlinear interactions among the factors. Generate the experimental runs for a Box-Behnken design on page 28-10 in coded (normalized) variables [-1, 0, +1].

```
CodedValue = bbdesign(3)
```

```
CodedValue =

    -1    -1     0
    -1     1     0
     1    -1     0
     1     1     0
    -1     0    -1
    -1     0     1
     1     0    -1
     1     0     1
     0    -1    -1
     0    -1     1
     0     1    -1
     0     1     1
     0     0     0
     0     0     0
     0     0     0
```

The first column is for the distance from radiator, the second column is for the pitch angle, and the third column is for the blade tip clearance. Suppose you want to test the effects of the variables at the following minimum and maximum values.

Distance from radiator: 1 to 1.5 inches
Pitch angle: 15 to 35 degrees
Blade tip clearance: 1 to 2 inches

Randomize the order of the runs, convert the coded design values to real-world units, and perform the experiment in the order specified.

```
runorder = randperm(15);     % Random permutation of the runs
bounds = [1 1.5;15 35;1 2];  % Min and max values for each factor

RealValue = zeros(size(CodedValue));
for i = 1:size(CodedValue,2) % Convert coded values to real-world units
    zmax = max(CodedValue(:,i));
    zmin = min(CodedValue(:,i));
    RealValue(:,i) = interp1([zmin zmax],bounds(i,:),CodedValue(:,i));
end
```

Suppose that at the end of the experiments, you collect the following response values in the variable `TestResult`.

```
TestResult = [837 864 829 856 880 879 872 874 834 833 860 859 874 876 875]';
```

Display the design values and the response.

```
disp({'Run Number','Distance','Pitch','Clearance','Airflow'})
disp(sortrows([runorder' RealValue TestResult]))
```

```
'Run Number'    'Distance'    'Pitch'    'Clearance'    'Airflow'

            1          1.5          35          1.5          856
```

```
 2         1.25          25          1.5          876
 3          1.5          25            1          872
 4         1.25          25          1.5          875
 5            1          35          1.5          864
 6         1.25          25          1.5          874
 7         1.25          15            2          833
 8          1.5          15          1.5          829
 9         1.25          15            1          834
10            1          15          1.5          837
11          1.5          25            2          874
12            1          25            1          880
13         1.25          35            1          860
14            1          25            2          879
15         1.25          35            2          859
```

Save the design values and the response in a `table`.

```
Expmt = table(runorder', CodedValue(:,1), CodedValue(:,2), CodedValue(:,3), ...
    TestResult,'VariableNames',{'RunNumber','D','P','C','Airflow'});
```

`D` stands for `Distance`, `P` stands for `Pitch`, and C stands for `Clearance`. Based on the experimental test results, the airflow rate is sensitive to the changing factors values. Also, four experimental runs meet or exceed the target airflow rate of 875 ft^3/min (runs 2, 4,12, and 14). However, it is not clear which, if any, of these runs is the optimal one. In addition, it is not obvious how robust the design is to variation in the factors. Create a model based on the current experimental data and use the model to estimate the optimal factor settings.

Improve the Cooling Fan Performance

The Box-Behnken design enables you to test for nonlinear (quadratic) effects. The form of the quadratic model is:

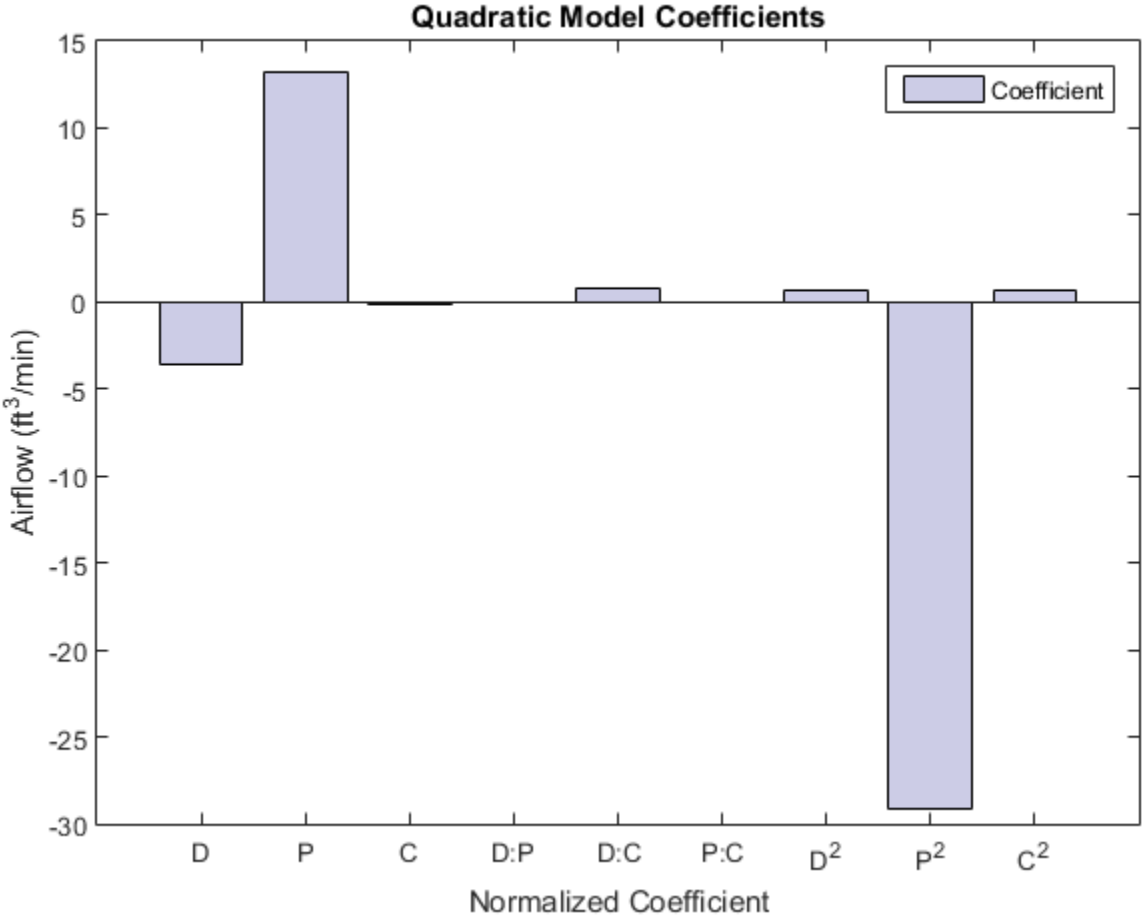
$$AF = \beta_0+\beta_1*Distance+\beta_2*Pitch+\beta_3*Clearance+\beta_4*Distance*Pitch$$
$$+\beta_5*Distance*Clearance+\beta_6*Pitch*Clearance+\beta_7*Distance^2$$
$$+\beta_8*Pitch^2+\beta_9*Clearance^2,$$

where AF is the airflow rate and B_i is the coefficient for the term i. Estimate the coefficients of this model using the `fitlm` function from Statistics and Machine Learning Toolbox.

```
mdl = fitlm(Expmt,'Airflow~D*P*C-D:P:C+D^2+P^2+C^2');
```

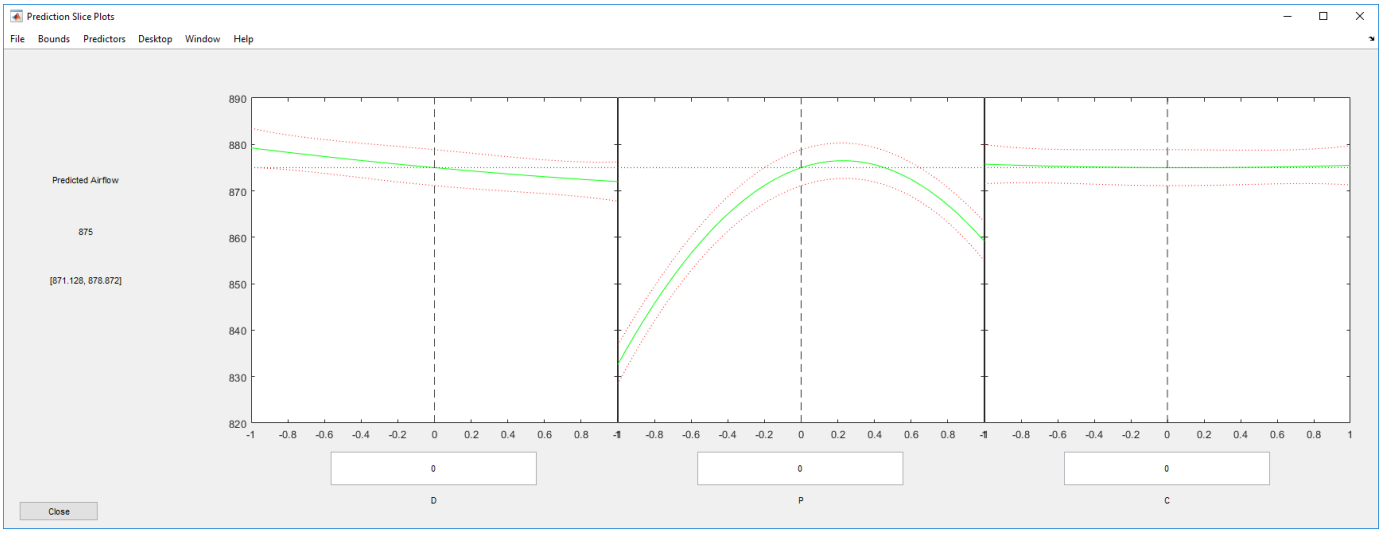
Display the magnitudes of the coefficients (for normalized values) in a bar chart.

```
figure()
h = bar(mdl.Coefficients.Estimate(2:10));
set(h,'facecolor',[0.8 0.8 0.9])
legend('Coefficient')
set(gcf,'units','normalized','position',[0.05 0.4 0.35 0.4])
set(gca,'xticklabel',mdl.CoefficientNames(2:10))
ylabel('Airflow (ft^3/min)')
xlabel('Normalized Coefficient')
title('Quadratic Model Coefficients')
```



The bar chart shows that *Pitch* and $Pitch^2$ are dominant factors. You can look at the relationship between multiple input variables and one output variable by generating a response surface plot. Use `plotSlice` to generate response surface plots for the model `mdl` interactively.

```
plotSlice(mdl)
```



The plot shows the nonlinear relationship of airflow with pitch. Move the blue dashed lines around and see the effect the different factors have on airflow. Although you can use `plotSlice` to determine the optimum factor settings, you can also use Optimization Toolbox to automate the task.

Find the optimal factor settings using the constrained optimization function `fmincon`.

Write the objective function.

```
f = @(x) -x2fx(x,'quadratic')*mdl.Coefficients.Estimate;
```

The objective function is a quadratic response surface fit to the data. Minimizing the negative airflow using `fmincon` is the same as maximizing the original objective function. The constraints are the upper and lower limits tested (in coded values). Set the initial starting point to be the center of the design of the experimental test matrix.

```
lb = [-1 -1 -1]; % Lower bound
ub = [1 1 1];    % Upper bound
x0 = [0 0 0];    % Starting point
[optfactors,fval] = fmincon(f,x0,[],[],[],[],lb,ub,[]); % Invoke the solver

Local minimum found that satisfies the constraints.

Optimization completed because the objective function is non-decreasing in
feasible directions, to within the default value of the function tolerance,
and constraints are satisfied to within the default value of the constraint tolerance.
```

Convert the results to a maximization problem and real-world units.

```
maxval = -fval;
maxloc = (optfactors + 1)';
bounds = [1 1.5;15 35;1 2];
maxloc=bounds(:,1)+maxloc .* ((bounds(:,2) - bounds(:,1))/2);
disp('Optimal Values:')
disp({'Distance','Pitch','Clearance','Airflow'})
disp([maxloc' maxval])

Optimal Values:
    'Distance'    'Pitch'    'Clearance'    'Airflow'
```

```
            1       27.275            1       882.26
```

The optimization result suggests placing the new fan one inch from the radiator, with a one-inch clearance between the tips of the fan blades and the shroud.

Because pitch angle has such a significant effect on airflow, perform additional analysis to verify that a 27.3 degree pitch angle is optimal.

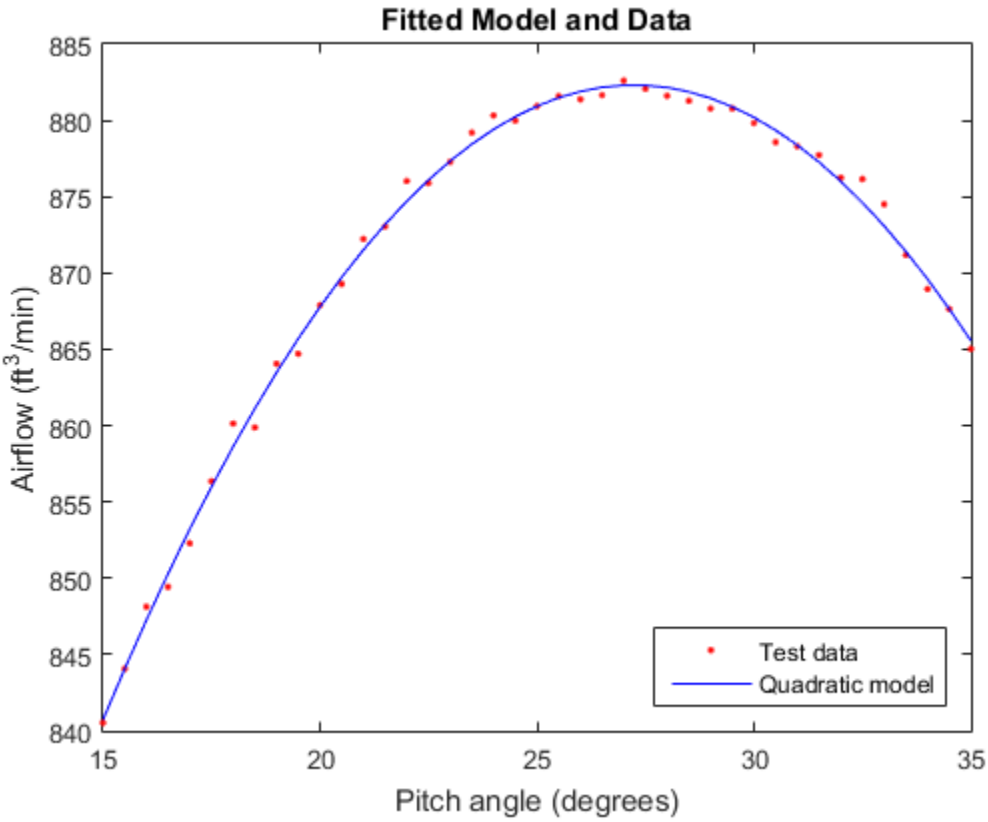
```
load(fullfile(matlabroot,'help/toolbox/stats/examples','AirflowData.mat'))
tbl = table(pitch,airflow);
mdl2 = fitlm(tbl,'airflow~pitch^2');
mdl2.Rsquared.Ordinary
```

```
ans =

      0.99632
```

The results show that a quadratic model explains the effect of pitch on the airflow well.

Plot the pitch angle against airflow and impose the fitted model.

```
figure()
plot(pitch,airflow,'.r')
hold on
ylim([840 885])
line(pitch,mdl2.Fitted,'color','b')
title('Fitted Model and Data')
xlabel('Pitch angle (degrees)')
ylabel('Airflow (ft^3/min)')
legend('Test data','Quadratic model','Location','se')
hold off
```



Find the pitch value that corresponds to the maximum airflow.

```
pitch(find(airflow==max(airflow)))
```

```
ans =

    27
```

The additional analysis confirms that a 27.3 degree pitch angle is optimal.

The improved cooling fan design meets the airflow requirements. You also have a model that approximates the fan performance well based on the factors you can modify in the design. Ensure that the fan performance is robust to variability in manufacturing and installation by performing a sensitivity analysis.

Sensitivity Analysis

Suppose that, based on historical experience, the manufacturing uncertainty is as follows.

Factor	Real Values	Coded Values
Distance from radiator	1.00 +/- 0.05 inch	1.00 +/- 0.20 inch
Blade pitch angle	27.3 +/- 0.25 degrees	0.227 +/- 0.028 degrees
Blade tip clearance	1.00 +/- 0.125 inch	-1.00 +/- 0.25 inch

Verify that these variations in factors will enable to maintain a robust design around the target airflow. The philosophy of Six Sigma targets a defect rate of no more than 3.4 per 1,000,000 fans. That is, the fans must hit the 875 ft^3/min target 99.999% of the time.

You can verify the design using Monte Carlo simulation. Generate 10,000 random numbers for three factors with the specified tolerance. First, set the state of the random number generators so results are consistent across different runs.

```
rng('default')
```

Perform the Monte Carlo simulation. Include a noise variable that is proportional to the noise in the fitted model, `mdl` (that is, the RMS error of the model). Because the model coefficients are in coded variables, you must generate `dist`, `pitch`, and `clearance` using the coded definition.

```
dist = random('normal',optfactors(1),0.20,[10000 1]);
pitch = random('normal',optfactors(2),0.028,[10000 1]);
clearance = random('normal',optfactors(3),0.25,[10000 1]);
noise = random('normal',0,mdl2.RMSE,[10000 1]);
```

Calculate airflow for 10,000 random factor combinations using the model.

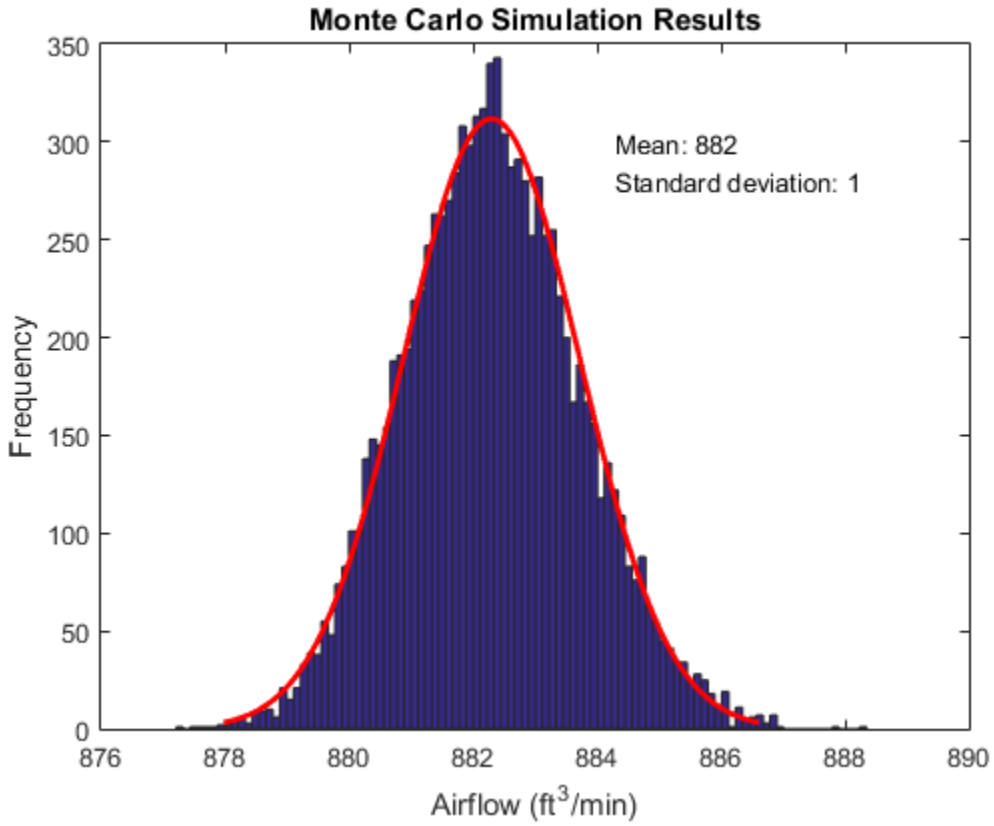
```
simfactor = [dist pitch clearance];
X = x2fx(simfactor,'quadratic');
```

Add noise to the model (the variation in the data that the model did not account for).

```
simflow = X*mdl.Coefficients.Estimate+noise;
```

Evaluate the variation in the model's predicted airflow using a histogram. To estimate the mean and standard deviation, fit a normal distribution to data.

```
pd = fitdist(simflow,'normal');
histfit(simflow)
hold on
text(pd.mu+2,300,['Mean: ' num2str(round(pd.mu))])
text(pd.mu+2,280,['Standard deviation: ' num2str(round(pd.sigma))])
hold off
xlabel('Airflow (ft^3/min)')
ylabel('Frequency')
title('Monte Carlo Simulation Results')
```



The results look promising. The average airflow is 882 ft^3/min and appears to be better than 875 ft^3/min for most of the data.

Determine the probability that the airflow is at 875 ft^3/min or below.

```
format long
pfail = cdf(pd,875)
pass = (1-pfail)*100
```

```
pfail =

     1.509289008603141e-07


pass =

  99.999984907109919
```

The design appears to achieve at least 875 ft^3/min of airflow 99.999% of the time.

Use the simulation results to estimate the process capability.

```
S = capability(simflow,[875.0 890])
pass = (1-S.Pl)*100
```

```
S =

       mu: 8.822982645666709e+02
```

```
    sigma: 1.424806876923940
        P: 0.999999816749816
       Pl: 1.509289008603141e-07
       Pu: 3.232128339675335e-08
       Cp: 1.754623760237126
      Cpl: 1.707427788957002
      Cpu: 1.801819731517250
      Cpk: 1.707427788957002


pass =

  99.9999849071099
```

The `Cp` value is 1.75. A process is considered high quality when `Cp` is greater than or equal to 1.6. The `Cpk` is similar to the `Cp` value, which indicates that the process is centered. Now implement this design. Monitor it to verify the design process and to ensure that the cooling fan delivers high-quality performance.

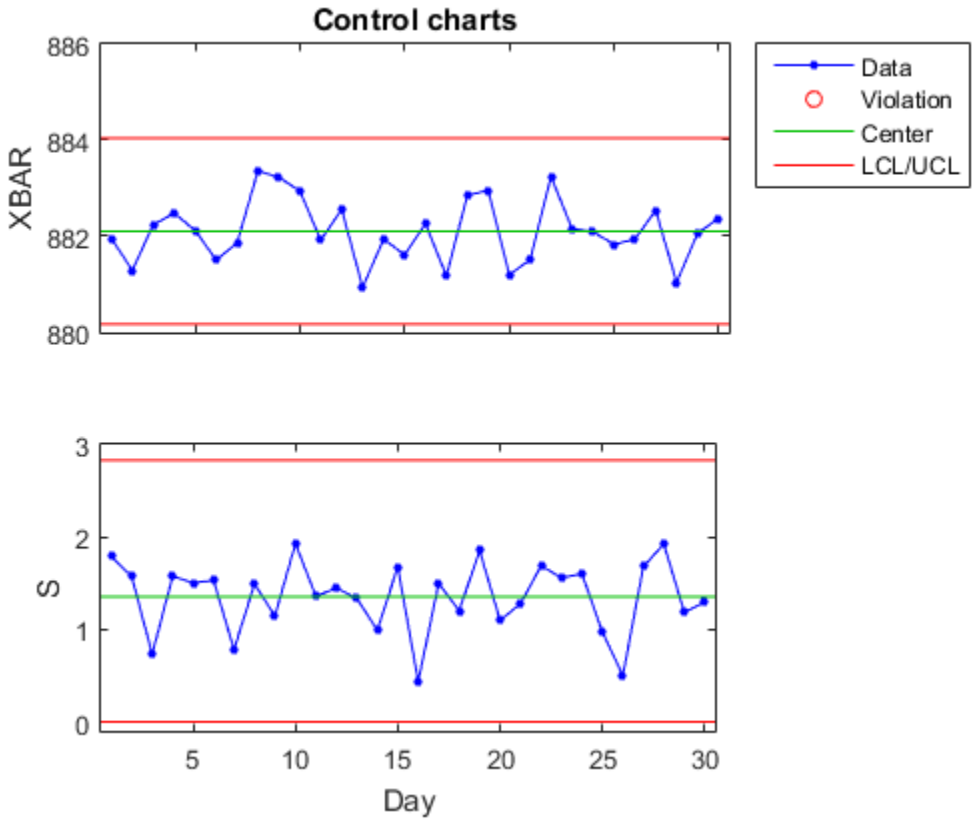
Control Manufacturing of the Improved Cooling Fan

You can monitor and evaluate the manufacturing and installation process of the new fan using control charts. Evaluate the first 30 days of production of the new cooling fan. Initially, five cooling fans per day were produced. First, load the sample data from the new process.

```
load(fullfile(matlabroot,'help/toolbox/stats/examples','spcdata.mat'))
```

Plot the *X*-bar and *S* charts.

```
figure()
controlchart(spcflow,'chart',{'xbar','s'}) % Reshape the data into daily sets
xlabel('Day')
```



According to the results, the manufacturing process is in statistical control, as indicated by the absence of violations of control limits or nonrandom patterns in the data over time. You can also run a capability analysis on the data to evaluate the process.

```
[row,col] = size(spcflow);
S2 = capability(reshape(spcflow,row*col,1),[875.0 890])
pass = (1-S.Pl)*100
```

```
S2 =

       mu: 8.821061141685465e+02
    sigma: 1.423887508874697
        P: 0.999999684316149
       Pl: 3.008932155898586e-07
       Pu: 1.479063578225176e-08
       Cp: 1.755756676295137
      Cpl: 1.663547652525458
      Cpu: 1.847965700064817
      Cpk: 1.663547652525458


pass =

  99.9999699106784
```

The `Cp` value of 1.755 is very similar to the estimated value of 1.73. The `Cpk` value of 1.66 is smaller than the `Cp` value. However, only a `Cpk` value less than 1.33, which indicates that the process shifted

significantly toward one of the process limits, is a concern. The process is well within the limits and it achieves the target airflow (875 ft^3/min) more than 99.999% of the time.

Statistical Process Control

Control Charts

A control chart displays measurements of process samples over time. The measurements are plotted together with user-defined *specification limits* and process-defined *control limits*. The process can then be compared with its specifications—to see if it is *in control* or *out of control*.

The chart is just a monitoring tool. Control activity might occur if the chart indicates an undesirable, systematic change in the process. The control chart is used to discover the variation, so that the process can be adjusted to reduce it.

Control charts are created with the `controlchart` function. Any of the following chart types may be specified:

- Xbar or mean
- Standard deviation
- Range
- Exponentially weighted moving average
- Individual observation
- Moving range of individual observations
- Moving average of individual observations
- Proportion defective
- Number of defectives
- Defects per unit
- Count of defects

Control rules are specified with the `controlrules` function. The following example illustrates how to use Western Electric rules to mark out of control measurements on an Xbar chart.

First load the sample data.

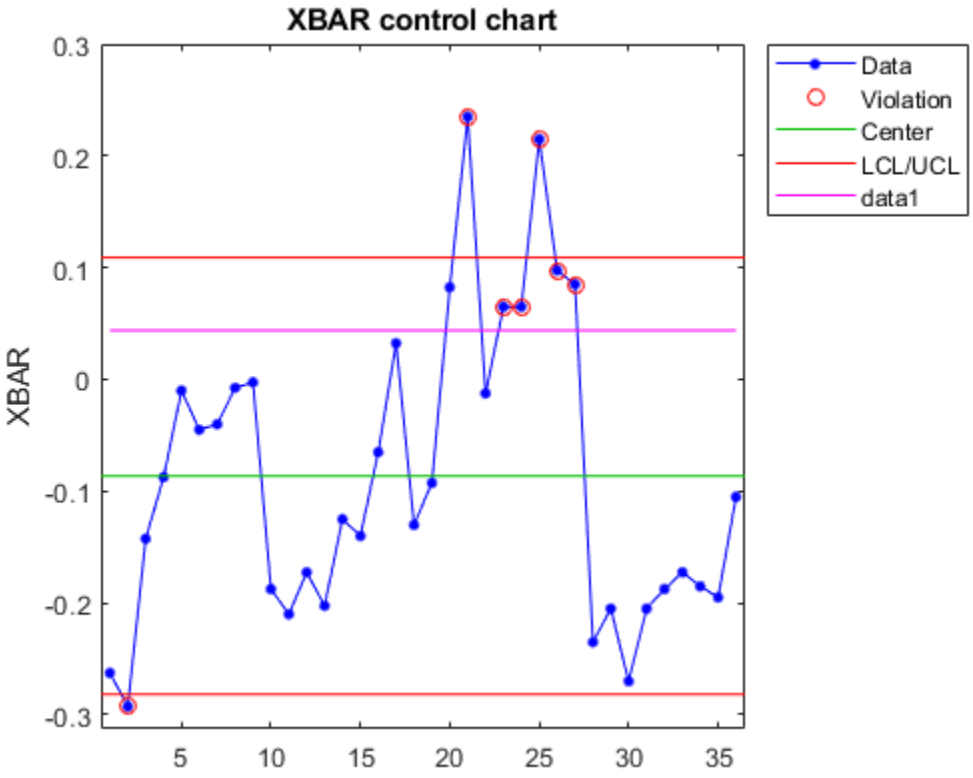
```
load parts
```

Construct the Xbar control chart using the Western Electric 2 rule (2 of 3 points at least 2 standard errors above the center line) to mark the out of control measurements.

```
st = controlchart(runout,'rules','we2');
```

For a better understanding of the Western Electric 2 rule, calculate and plot the 2 standard errors line on the chart.

```
x = st.mean;
cl = st.mu;
se = st.sigma./sqrt(st.n);
hold on
plot(cl+2*se,'m')
```



Identify the measurements that violate the control rule.

```
R = controlrules('we2',x,cl,se);
I = find(R)
```

```
I = 6×1

    21
    23
    24
    25
    26
    27
```

See Also

`controlchart` | `controlrules`

Related Examples

- “Capability Studies” on page 29-4

Capability Studies

Before going into production, many manufacturers run a *capability study* to determine if their process will run within specifications enough of the time. *Capability indices* produced by such a study are used to estimate expected percentages of defective parts.

Capability studies are conducted with the `capability` function. The following capability indices are produced:

- `mu` — Sample mean
- `sigma` — Sample standard deviation
- `P` — Estimated probability of being within the lower (`L`) and upper (`U`) specification limits
- `Pl` — Estimated probability of being below `L`
- `Pu` — Estimated probability of being above `U`
- `Cp` — `(U-L)/(6*sigma)`
- `Cpl` — `(mu-L)./(3.*sigma)`
- `Cpu` — `(U-mu)./(3.*sigma)`
- `Cpk` — `min(Cpl,Cpu)`

As an example, simulate a sample from a process with a mean of 3 and a standard deviation of 0.005:

```
rng default; % For reproducibility
data = normrnd(3,0.005,100,1);
```

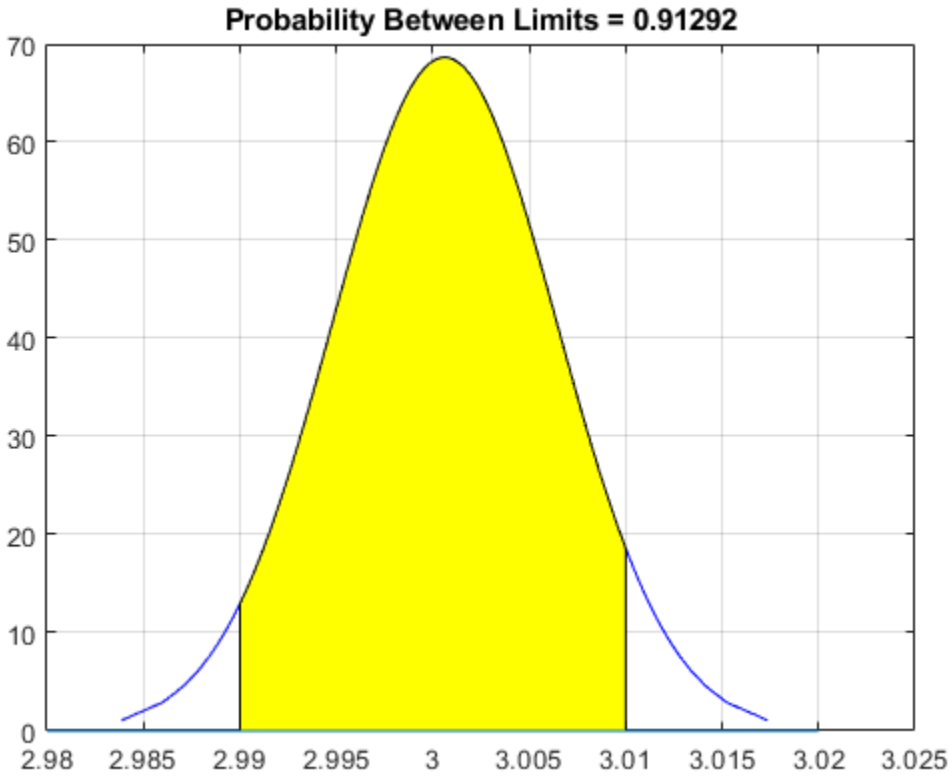
Compute capability indices if the process has an upper specification limit of 3.01 and a lower specification limit of 2.99:

```
S = capability(data,[2.99 3.01])
```

```
S = struct with fields:
       mu: 3.0006
    sigma: 0.0058
        P: 0.9129
       Pl: 0.0339
       Pu: 0.0532
       Cp: 0.5735
      Cpl: 0.6088
      Cpu: 0.5382
      Cpk: 0.5382
```

Visualize the specification and process widths:

```
capaplot(data,[2.99 3.01]);
grid on
```



See Also

`capability`

Related Examples

- “Control Charts” on page 29-2

Tall Arrays

Logistic Regression with Tall Arrays

This example shows how to use logistic regression and other techniques to perform data analysis on tall arrays. Tall arrays represent data that is too large to fit into computer memory.

Define Execution Environment

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Get Data into MATLAB

Create a datastore that references the folder location with the data. The data can be contained in a single file, a collection of files, or an entire folder. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Select a subset of the variables to work with, and include the name of the airline (`UniqueCarrier`) as a categorical variable. Create a tall table on top of the datastore.

```
ds = datastore('airlinesmall.csv');
ds.TreatAsMissing = 'NA';
ds.SelectedVariableNames = {'DayOfWeek','UniqueCarrier',...
    'ArrDelay','DepDelay','Distance'};
ds.SelectedFormats{2} = '%C';
tt = tall(ds);
tt.DayOfWeek = categorical(tt.DayOfWeek,1:7,...
    {'Sun','Mon','Tues','Wed','Thu','Fri','Sat'},'Ordinal',true)
```

```
tt =

  Mx5 tall table

    DayOfWeek    UniqueCarrier    ArrDelay    DepDelay    Distance
    _________    _____________    ________    ________    ________

        ?              ?              ?           ?           ?
        ?              ?              ?           ?           ?
        ?              ?              ?           ?           ?
        :              :              :           :           :
        :              :              :           :           :
```

Late Flights

Determine the flights that are late by 20 minutes or more by defining a logical variable that is true for a late flight. Add this variable to the tall table of data, noting that it is not yet evaluated. A preview of this variable includes the first few rows.

```
tt.LateFlight = tt.ArrDelay>=20
```

```
tt =

  Mx6 tall table

    DayOfWeek    UniqueCarrier    ArrDelay    DepDelay    Distance    LateFlight
```

```
    _________    ______________    ________    _________    ________    __________

        ?              ?             ?           ?           ?           ?
        ?              ?             ?           ?           ?           ?
        ?              ?             ?           ?           ?           ?
        :              :             :           :           :           :
        :              :             :           :           :           :
```

Calculate the mean of `LateFlight` to determine the overall proportion of late flights. Use `gather` to trigger evaluation of the tall array and bring the result into memory.

```
m = mean(tt.LateFlight)

m =

  tall double

    ?

m = gather(m)

Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 1.4 sec
- Pass 2 of 2: Completed in 1.6 sec
Evaluation completed in 3.9 sec

m = 0.1580
```

Late Flights by Carrier

Examine whether certain types of flights tend to be late. First, check to see if certain carriers are more likely to have late flights.

```
tt.LateFlight = double(tt.LateFlight);
late_by_carrier = gather(grpstats(tt,'UniqueCarrier','mean','DataVar','LateFlight'))

Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 3.6 sec
Evaluation completed in 4.6 sec

late_by_carrier=29×4 table
    GroupLabel    UniqueCarrier    GroupCount    mean_LateFlight
    __________    _____________    __________    _______________

    {'9E'    }       9E               521           0.13436
    {'AA'    }       AA             14930           0.16236
    {'AQ'    }       AQ               154          0.051948
    {'AS'    }       AS              2910           0.16014
    {'B6'    }       B6               806           0.23821
    {'CO'    }       CO              8138           0.16319
    {'DH'    }       DH               696           0.17672
    {'DL'    }       DL             16578           0.15261
    {'EA'    }       EA               920           0.15217
    {'EV'    }       EV              1699           0.21248
    {'F9'    }       F9               335           0.18209
    {'FL'    }       FL              1263           0.19952
    {'HA'    }       HA               273          0.047619
    {'HP'    }       HP              3660           0.13907
    {'ML (1)'}       ML (1)            69          0.043478
```

```
    {'MQ'    }       MQ               3962           0.18778
      ⋮
```

Carriers B6 and EV have higher proportions of late flights. Carriers AQ, ML(1), and HA have relatively few flights, but lower proportions of them are late.

Late Flights by Day of Week

Next, check to see if different days of the week tend to have later flights.

```
late_by_day = gather(grpstats(tt,'DayOfWeek','mean','DataVar','LateFlight'))
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.9 sec
Evaluation completed in 2.3 sec

late_by_day=7×4 table
    GroupLabel    DayOfWeek    GroupCount    mean_LateFlight
    __________    _________    __________    _______________

     {'Fri' }       Fri          15839           0.12899
     {'Mon' }       Mon          18077           0.14234
     {'Sat' }       Sat          16958           0.15603
     {'Sun' }       Sun          18019           0.15117
     {'Thu' }       Thu          18227           0.18418
     {'Tues'}       Tues         18163           0.15526
     {'Wed' }       Wed          18240           0.18399
```

Wednesdays and Thursdays have the highest proportion of late flights, while Fridays have the lowest proportion.

Late Flights by Distance

Check to see if longer or shorter flights tend to be late. First, look at the density of the flight distance for flights that are late, and compare that with flights that are on time.

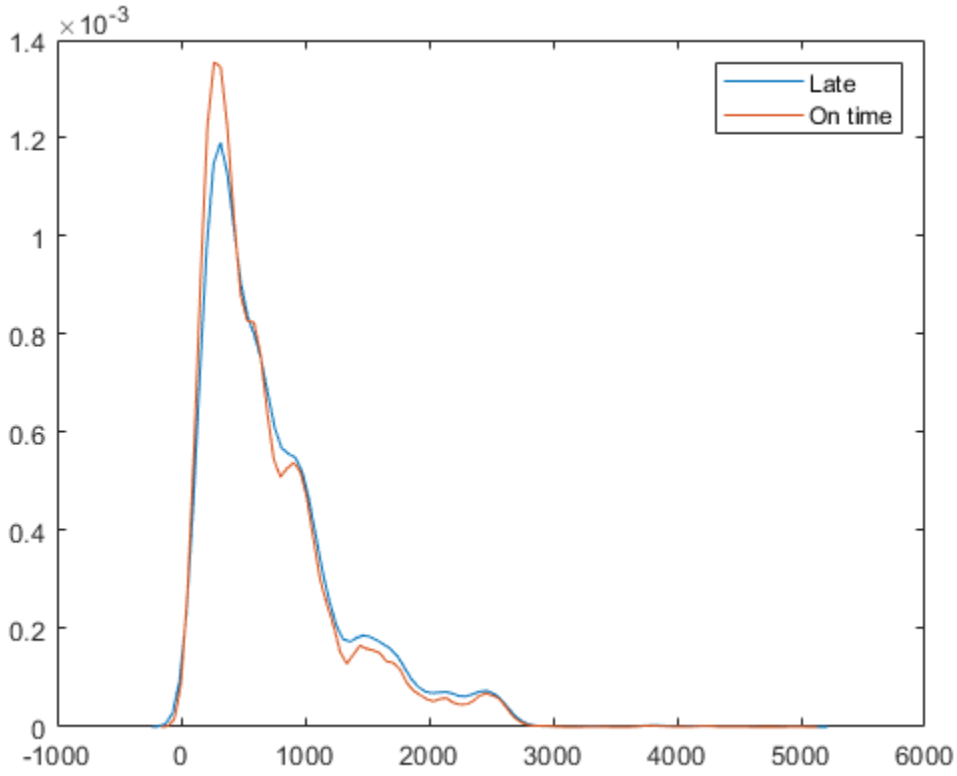
```
ksdensity(tt.Distance(tt.LateFlight==1))
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 1.7 sec
- Pass 2 of 2: Completed in 1.6 sec
Evaluation completed in 4 sec
```

```
hold on
ksdensity(tt.Distance(tt.LateFlight==0))
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 1.4 sec
- Pass 2 of 2: Completed in 1.7 sec
Evaluation completed in 3.5 sec
```

```
hold off
legend('Late','On time')
```



Flight distance does not make a dramatic difference in whether a flight is early or late. However, the density appears to be slightly higher for on-time flights at distances of about 400 miles. The density is also higher for late flights at distances of about 2000 miles. Calculate some simple descriptive statistics for the late and on-time flights.

```
late_by_distance = gather(grpstats(tt,'LateFlight',{'mean' 'std'},'DataVar','Distance'))
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.7 sec
Evaluation completed in 2.2 sec

late_by_distance=2×5 table
    GroupLabel    LateFlight    GroupCount    mean_Distance    std_Distance
    __________    __________    __________    _____________    ____________

      {'0'}           0          1.04e+05        693.14           544.75
      {'1'}           1             19519        750.24           574.12
```

Late flights are about 60 miles longer on average, although this value makes up only a small portion of the standard deviation of the distance values.

Logistic Regression Model

Build a model for the probability of a late flight, using both continuous variables (such as `Distance`) and categorical variables (such as `DayOfWeek`) to predict the probabilities. This model can help to determine if the previous results observed for each predictor individually also hold true when you consider them together.

```
glm = fitglm(tt,'LateFlight~Distance+DayOfWeek','Distribution','binomial')

Iteration [1]:      0% completed
Iteration [1]:     100% completed
Iteration [2]:      0% completed
Iteration [2]:     100% completed
Iteration [3]:      0% completed
Iteration [3]:     100% completed
Iteration [4]:      0% completed
Iteration [4]:     100% completed
Iteration [5]:      0% completed
Iteration [5]:     100% completed

glm =
Compact generalized linear regression model:
    logit(LateFlight) ~ 1 + DayOfWeek + Distance
    Distribution = Binomial

Estimated Coefficients:
                       Estimate         SE         tStat       pValue
                      __________    __________    _______    __________

    (Intercept)           -1.855      0.023052    -80.469            0
    DayOfWeek_Mon      -0.072603      0.029798    -2.4365      0.01483
    DayOfWeek_Tues      0.026909      0.029239    0.92029      0.35742
    DayOfWeek_Wed         0.2359      0.028276      8.343    7.2452e-17
    DayOfWeek_Thu        0.23569      0.028282     8.3338    7.8286e-17
    DayOfWeek_Fri       -0.19285      0.031583     -6.106    1.0213e-09
    DayOfWeek_Sat       0.033542      0.029702     1.1293       0.25879
    Distance          0.00018373    1.3507e-05     13.602    3.8741e-42


123319 observations, 123311 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 504, p-value = 8.74e-105
```

The model confirms that the previously observed conclusions hold true here as well:

- The Wednesday and Thursday coefficients are positive, indicating a higher probability of a late flight on those days. The Friday coefficient is negative, indicating a lower probability.
- The Distance coefficient is positive, indicating that longer flights have a higher probability of being late.

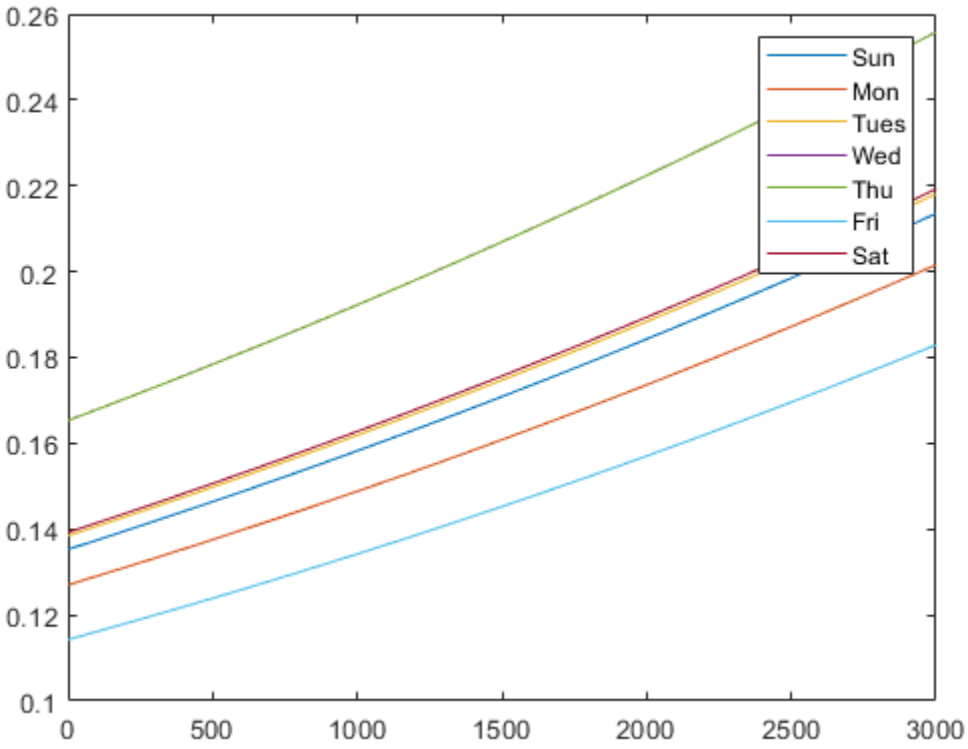
All of these coefficients have very small p-values. This is common with data sets that have many observations, since one can reliably estimate small effects with large amounts of data. In fact, the uncertainty in the model is larger than the uncertainty in the estimates for the parameters in the model.

Prediction with Model

Predict the probability of a late flight for each day of the week, and for distances ranging from 0 to 3000 miles. Create a table to hold the predictor values by indexing the first 100 rows in the original table `tt`.

```
x = gather(tt(1:100,{'Distance' 'DayOfWeek'}));
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.35 sec
Evaluation completed in 0.59 sec

x.Distance = linspace(0,3000)';
x.DayOfWeek(:) = 'Sun';
plot(x.Distance,predict(glm,x));

days = {'Sun' 'Mon' 'Tues' 'Wed' 'Thu' 'Fri' 'Sat'};
hold on
for j=2:length(days)
    x.DayOfWeek(:) = days{j};
    plot(x.Distance,predict(glm,x));
end
legend(days)
```



According to this model, a Wednesday or Thursday flight of 500 miles has the same probability of being late, about 18%, as a Friday flight of about 3000 miles.

Since these probabilities are all much less than 50%, the model is unlikely to predict that any given flight will be late using this information. Investigate the model more by focusing on the flights for which the model predicts a probability of 20% or more of being late, and compare that to the actual results.

```
C = gather(crosstab(tt.LateFlight,predict(glm,tt)>.20))
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec

C = 2×2

       99613        4391
       18394        1125
```

Among the flights predicted to have a 20% or higher probability of being late, about 20% were late `1125/(1125 + 4391)`. Among the remainder, less than 16% were late `18394/(18394 + 99613)`.

Bayesian Optimization with Tall Arrays

This example shows how to use Bayesian optimization to select optimal parameters for training a kernel classifier by using the `'OptimizeHyperparameters'` name-value argument. The sample data set `airlinesmall.csv` is a large data set that contains a tabular file of airline flight data. This example creates a tall table containing the data, and extracts class labels and predictor data from the tall table to run the optimization procedure.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Get Data into MATLAB®

Create a datastore that references the folder location with the data. The data can be contained in a single file, a collection of files, or an entire folder. For folders that contain a collection of files, you can specify the entire folder location, or use the wildcard character, `'*.csv'`, to include multiple files with the same file extension in the datastore. Select a subset of the variables to work with, and treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Create a tall table that contains the data in the datastore.

```
ds = datastore('airlinesmall.csv');
ds.SelectedVariableNames = {'Month','DayofMonth','DayOfWeek',...
                            'DepTime','ArrDelay','Distance','DepDelay'};
ds.TreatAsMissing = 'NA';
tt  = tall(ds) % Tall table
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

tt =

  M×7 tall table

    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    Distance    DepDelay
    _____    __________    _________    _______    ________    ________    ________

     10          21            3          642          8          308          12
     10          26            1         1021          8          296           1
     10          23            5         2055         21          480          20
     10          23            5         1332         13          296          12
     10          22            4          629          4          373          -1
     10          28            3         1446         59          308          63
     10           8            4          928          3          447          -2
     10          10            6          859         11          954          -1
      :           :            :            :          :            :           :
      :           :            :            :          :            :           :
```

Prepare Class Labels and Predictor Data

Determine the flights that are late by 10 minutes or more by defining a logical variable that is true for a late flight. This variable contains the class labels. A preview of this variable includes the first few rows.

```
Y = tt.DepDelay > 10 % Class labels
```

```
Y =

  M×1 tall logical array

   1
   0
   1
   1
   0
   1
   0
   0
   :
   :
```

Create a tall array for the predictor data.

```
X = tt{:,1:end-1} % Predictor data
```

```
X =

  M×6 tall double matrix

         10          21           3         642           8         308
         10          26           1        1021           8         296
         10          23           5        2055          21         480
         10          23           5        1332          13         296
         10          22           4         629           4         373
         10          28           3        1446          59         308
         10           8           4         928           3         447
         10          10           6         859          11         954
         :           :           :           :           :           :
         :           :           :           :           :           :
```

Remove rows in `X` and `Y` that contain missing data.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:end-1);
Y = R(:,end);
```

Perform Bayesian Optimization Using `OptimizeHyperparameters`

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value argument.

Standardize the predictor variables.

```
Z = zscore(X);
```

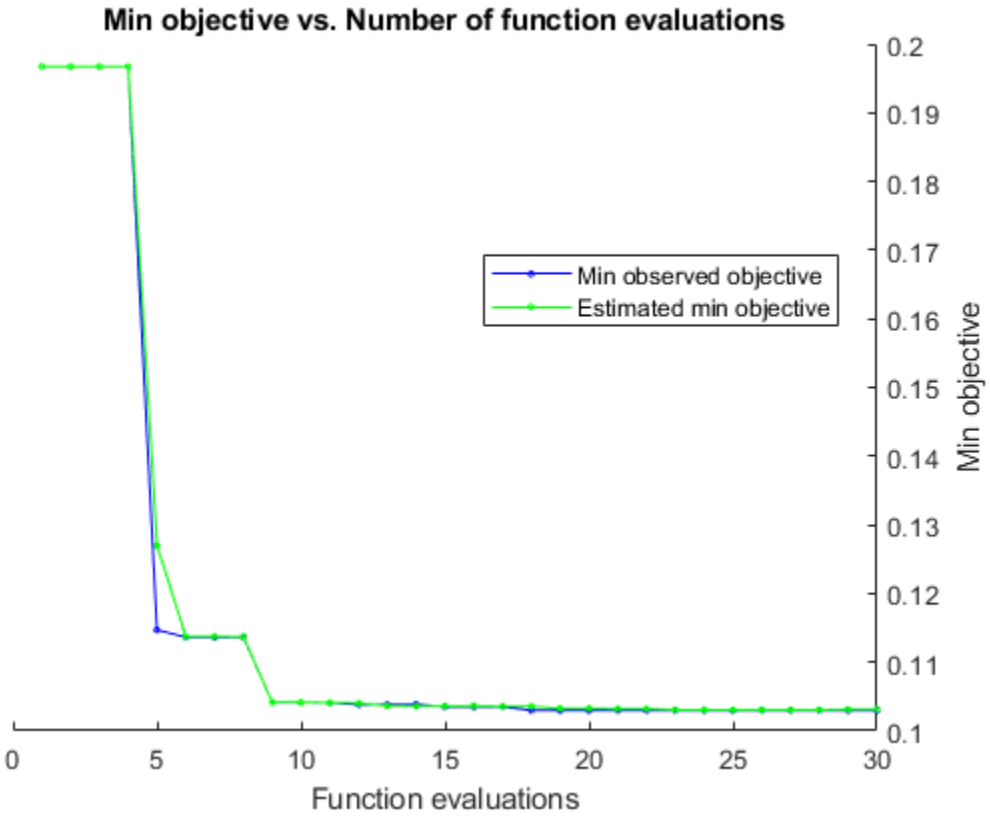
Find the optimal values for the `'KernelScale'` and `'Lambda'` name-value arguments that minimize the loss on the holdout validation set. By default, the software selects and reserves 20% of the data as validation data, and trains the model using the rest of the data. You can change the holdout fraction by using the `'HyperparameterOptimizationOptions'` name-value argument. For reproducibility, use the `'expected-improvement-plus'` acquisition function and set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

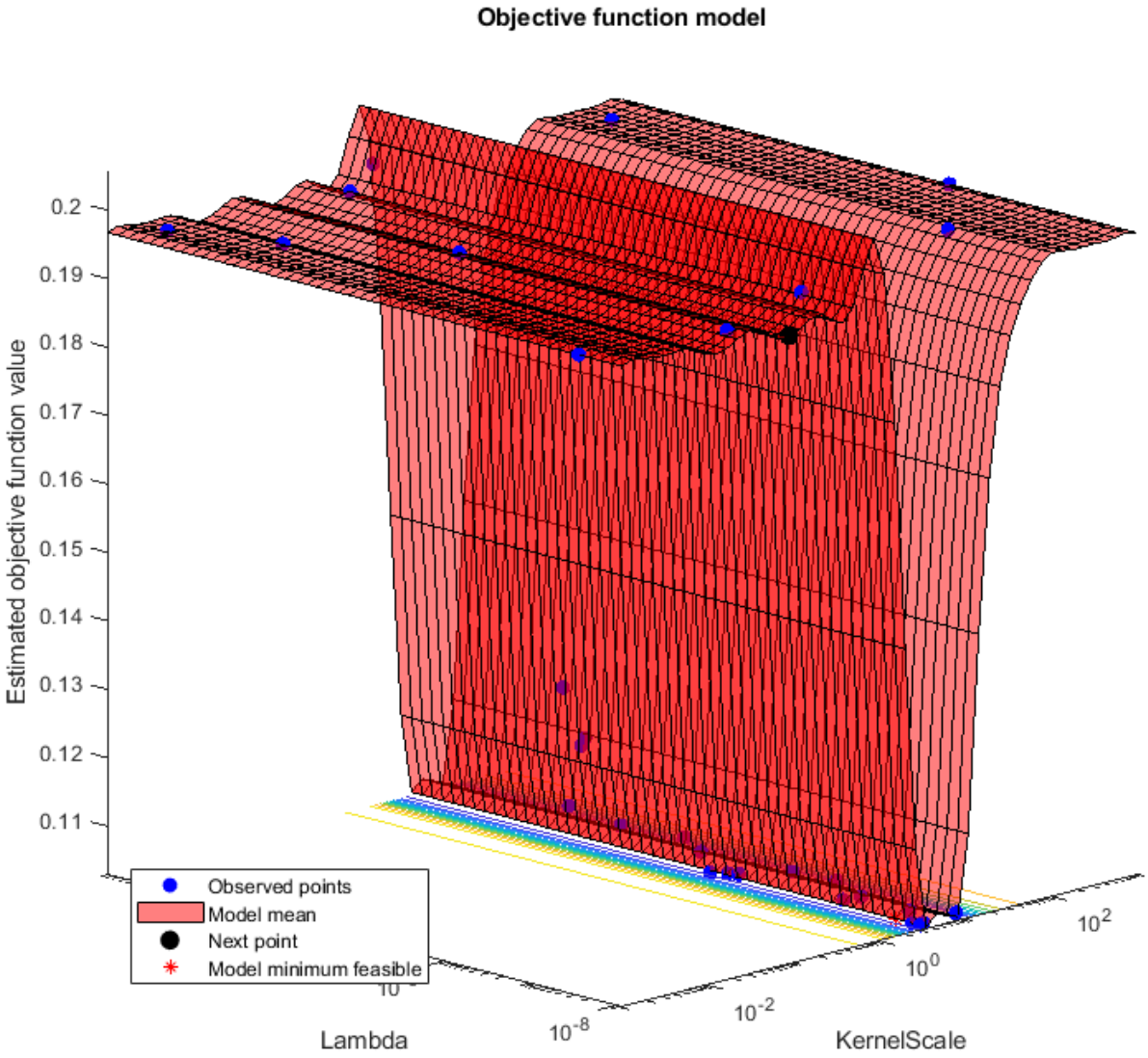
```
rng('default')
tallrng('default')
Mdl = fitckernel(Z,Y,'Verbose',0,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName','expected-improvement-p
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 7.1 sec
- Pass 2 of 2: Completed in 2.2 sec
Evaluation completed in 12 sec

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.6 sec
Evaluation completed in 1.8 sec
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  KernelScale |       La
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.19672 |      125.49 |     0.19672 |     0.19672 |       1.2297 |    0.008

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.93 sec
Evaluation completed in 1.1 sec
|    2 | Accept |     0.19672 |      53.653 |     0.19672 |     0.19672 |     0.039643 |   2.5756

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.5 sec
Evaluation completed in 1.6 sec
|    3 | Accept |     0.19672 |      52.453 |     0.19672 |     0.19672 |      0.02562 |   1.2555

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    4 | Accept |     0.19672 |      57.223 |     0.19672 |     0.19672 |       92.644 |   1.2056

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 1.5 sec
|    5 | Best   |     0.11469 |      89.981 |     0.11469 |     0.12698 |       11.173 |   0.0002

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.94 sec
Evaluation completed in 1.1 sec
|    6 | Best   |     0.11365 |      82.031 |     0.11365 |     0.11373 |       10.609 |   0.0002

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.92 sec
Evaluation completed in 1.1 sec
|    7 | Accept |     0.19672 |      50.604 |     0.11365 |     0.11373 |    0.0059498 |   0.0004

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|    8 | Accept |     0.12122 |      91.341 |     0.11365 |     0.11371 |        11.44 |   0.0004

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.97 sec
Evaluation completed in 1.1 sec
|    9 | Best   |     0.10417 |      42.696 |     0.10417 |     0.10417 |       8.0424 |   6.7998

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
```

```
Evaluation completed in 1 sec
|   10 | Accept |     0.10433 |      42.215 |     0.10417 |     0.10417 |       9.6694 |   1.4948

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1 sec
|   11 | Best   |     0.10409 |      41.618 |     0.10409 |     0.10411 |       6.2099 |   6.1093

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   12 | Best   |     0.10383 |      44.635 |     0.10383 |     0.10404 |       5.6767 |   7.6134

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|   13 | Accept |     0.10408 |      45.429 |     0.10383 |     0.10365 |       8.1769 |   8.5993

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|   14 | Accept |     0.10404 |      41.928 |     0.10383 |     0.10361 |       7.6191 |   6.4079

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.93 sec
Evaluation completed in 1.1 sec
|   15 | Best   |     0.10351 |      42.094 |     0.10351 |     0.10362 |       4.2987 |   9.2645

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   16 | Accept |     0.10404 |      44.684 |     0.10351 |     0.10362 |       4.8747 |   1.7838

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1 sec
|   17 | Accept |     0.10657 |      88.006 |     0.10351 |     0.10357 |       4.8239 |   0.0001

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   18 | Best   |     0.10299 |      41.303 |     0.10299 |     0.10358 |       3.5555 |   2.7165

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|   19 | Accept |     0.10366 |      41.301 |     0.10299 |     0.10324 |       3.8035 |   1.3542

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 0.99 sec
|   20 | Accept |     0.10337 |      41.345 |     0.10299 |     0.10323 |        3.806 |   1.8101

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  KernelScale |       La
|      | result |             | runtime     | (observed)  | (estim.)    |              |
```

```
|=====================================================================================================
|   21 | Accept |     0.10345 |      41.418 |     0.10299 |     0.10322 |       3.3655 |    9.082

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 0.98 sec
|   22 | Accept |     0.19672 |      60.129 |     0.10299 |     0.10322 |       999.62 |   1.2609

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   23 | Accept |     0.10315 |      41.133 |     0.10299 |     0.10306 |       3.6716 |   1.2445

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   24 | Accept |     0.19672 |      48.262 |     0.10299 |     0.10306 |    0.0010004 |   2.6214

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
|   25 | Accept |     0.19672 |      48.334 |     0.10299 |     0.10306 |      0.21865 |    0.002

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 0.98 sec
|   26 | Accept |     0.19672 |      60.229 |     0.10299 |     0.10306 |       299.92 |    0.003

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 0.99 sec
|   27 | Accept |     0.19672 |      48.361 |     0.10299 |     0.10306 |     0.002436 |    0.004

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.2 sec
Evaluation completed in 1.4 sec
|   28 | Accept |     0.19672 |      52.539 |     0.10299 |     0.10305 |      0.50559 |   3.3667

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   29 | Accept |     0.10354 |      43.957 |     0.10299 |     0.10313 |       3.7754 |   9.5626

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.93 sec
Evaluation completed in 1.1 sec
|   30 | Accept |     0.10405 |      41.388 |     0.10299 |     0.10315 |       8.9864 |   2.3136
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.2
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0
5
10
15
20
25
30




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 1677.1387 seconds
Total objective function evaluation time: 1645.7748

Best observed feasible point:
    KernelScale      Lambda
    ___________    __________

      3.5555       2.7165e-06

Observed objective function value = 0.10299
```

```
Estimated objective function value = 0.10332
Function evaluation time = 41.3029

Best estimated feasible point (according to models):
    KernelScale      Lambda
    ___________    __________

      3.6716       1.2445e-08

Estimated objective function value = 0.10315
Estimated function evaluation time = 42.3461

Mdl =
  ClassificationKernel
            PredictorNames: {'x1'  'x2'  'x3'  'x4'  'x5'  'x6'}
              ResponseName: 'Y'
                ClassNames: [0 1]
                   Learner: 'svm'
    NumExpansionDimensions: 256
               KernelScale: 3.6716
                    Lambda: 1.2445e-08
             BoxConstraint: 665.9442


  Properties, Methods
```

Perform Bayesian Optimization by Using bayesopt

Alternatively, you can use the `bayesopt` function to find the optimal values of hyperparameters.

Split the data set into training and test sets. Specify a 1/3 holdout sample for the test set.

```
rng('default') % For reproducibility
tallrng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',1/3);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition);         % Indices for the test set
```

Extract training and testing data and standardize the predictor data.

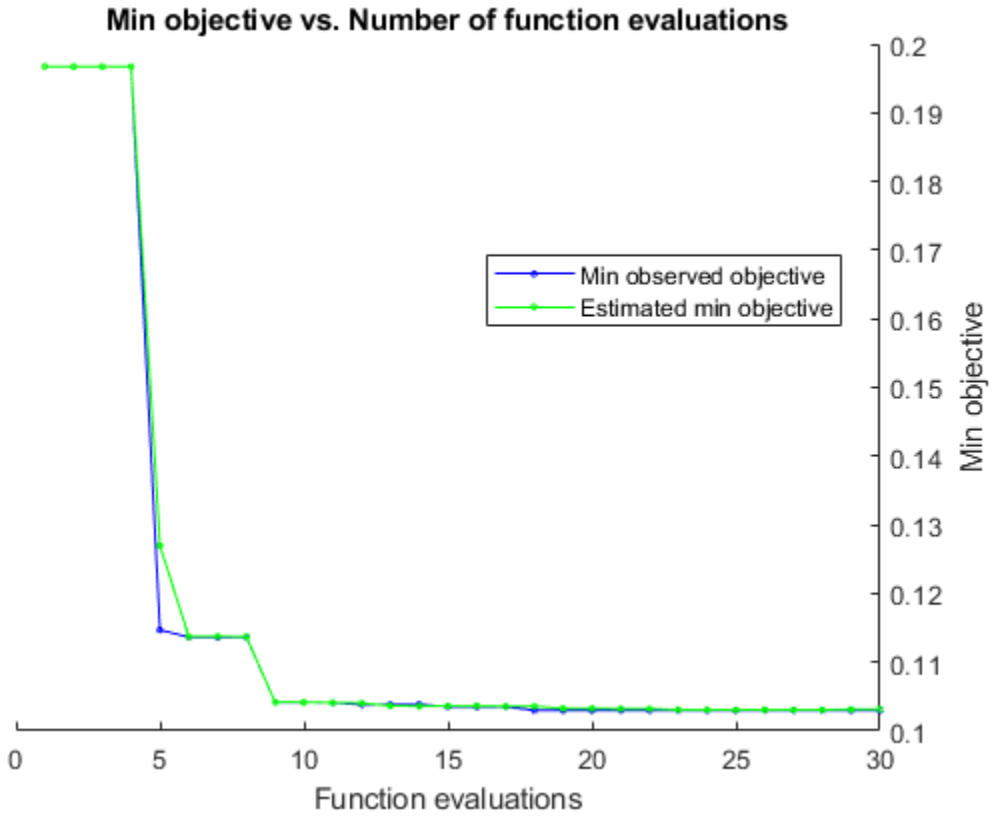
```
Ytrain = Y(trainingInds); % Training class labels
Xtrain = X(trainingInds,:);
[Ztrain,mu,stddev] = zscore(Xtrain); % Standardized training data

Ytest = Y(testInds); % Testing class labels
Xtest = X(testInds,:);
Ztest = (Xtest-mu)./stddev; % Standardized test data
```

Define the variables `sigma` and `lambda` to find the optimal values for the `'KernelScale'` and `'Lambda'` name-value arguments. Use `optimizableVariable` and specify a wide range for the variables because optimal values are unknown. Apply logarithmic transformation to the variables to search for the optimal values on a log scale.

```
N = gather(numel(Ytrain)); % Evaluate the length of the tall training array in memory
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: 0% complete
Evaluation 0% complete
```

```
- Pass 1 of 1: Completed in 0.95 sec
Evaluation 91% complete
```




```
Evaluation completed in 1.1 sec
```

```
sigma = optimizableVariable('sigma',[1e-3,1e3],'Transform','log');
lambda = optimizableVariable('lambda',[(1e-3)/N, (1e3)/N],'Transform','log');
```

Create the objective function for Bayesian optimization. The objective function takes in a table that contains the variables `sigma` and `lambda`, and then computes the classification loss value for the binary Gaussian kernel classification model trained using the `fitckernel` function. Set `'Verbose',0` within `fitckernel` to suppress the iterative display of diagnostic information.

```
minfn = @(z)gather(loss(fitckernel(Ztrain,Ytrain, ...
    'KernelScale',z.sigma,'Lambda',z.lambda,'Verbose',0), ...
    Ztest,Ytest));
```

Optimize the parameters `[sigma,lambda]` of the kernel classification model with respect to the classification loss by using `bayesopt`. By default, `bayesopt` displays iterative information about the optimization at the command line. For reproducibility, set the `AcquisitionFunctionName` option to `'expected-improvement-plus'`. The default acquisition function depends on run time and, therefore, can give varying results.

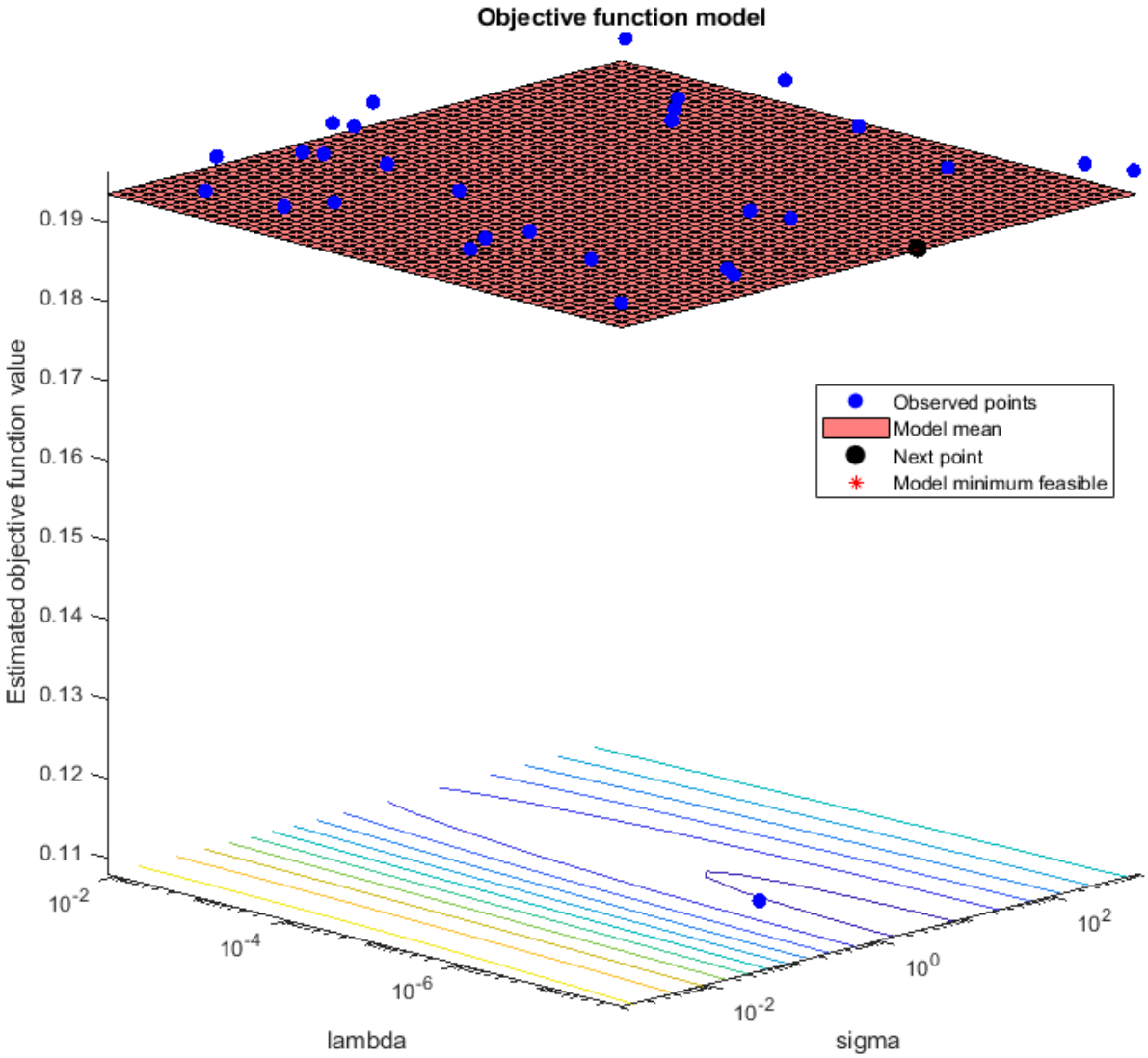
```
results = bayesopt(minfn,[sigma,lambda],'AcquisitionFunctionName','expected-improvement-plus')
```

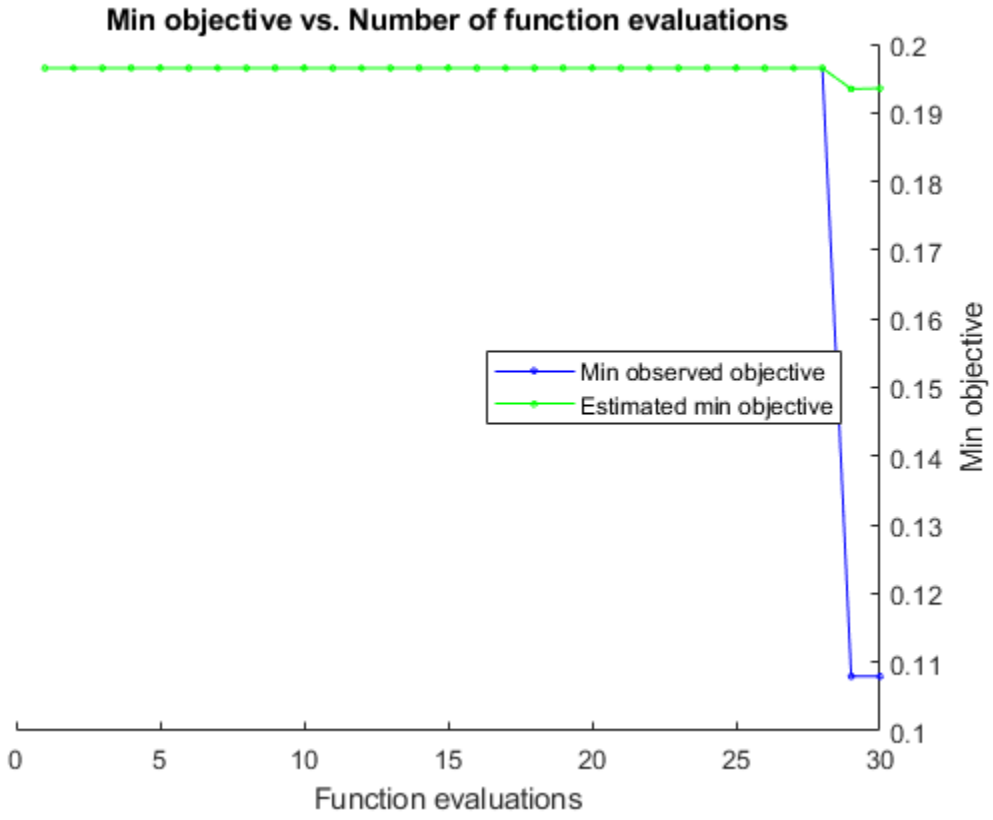
```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        sigma |       la
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.19651 |      84.526 |     0.19651 |     0.19651 |       1.2297 |     0.0
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    2 | Accept |     0.19651 |      112.57 |     0.19651 |     0.19651 |     0.039643 |   3.863
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|    3 | Accept |     0.19651 |      80.282 |     0.19651 |     0.19651 |      0.02562 |   1.883
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|    4 | Accept |     0.19651 |      52.306 |     0.19651 |     0.19651 |       92.644 |   1.808
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    5 | Accept |     0.19651 |      52.717 |     0.19651 |     0.19651 |       978.95 |   0.000
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    6 | Accept |     0.19651 |      90.336 |     0.19651 |     0.19651 |    0.0089609 |    0.00
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    7 | Accept |     0.19651 |      110.35 |     0.19651 |     0.19651 |    0.0010228 |    1.29
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|    8 | Accept |     0.19651 |      76.594 |     0.19651 |     0.19651 |      0.27475 |    0.00
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|    9 | Accept |     0.19651 |      77.641 |     0.19651 |     0.19651 |      0.81326 |   1.075
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   10 | Accept |     0.19651 |      100.21 |     0.19651 |     0.19651 |    0.0040507 |   0.000
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   11 | Accept |     0.19651 |      52.287 |     0.19651 |     0.19651 |       964.67 |   1.278
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   12 | Accept |     0.19651 |       107.7 |     0.19651 |     0.19651 |      0.24069 |    0.00
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   13 | Accept |     0.19651 |      52.092 |     0.19651 |     0.19651 |       974.15 |     0.0

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|   14 | Accept |     0.19651 |      92.184 |     0.19651 |     0.19651 |    0.0013246 |    0.00

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   15 | Accept |     0.19651 |      87.893 |     0.19651 |     0.19651 |    0.0067415 |   1.9074

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|   16 | Accept |     0.19651 |      110.46 |     0.19651 |     0.19651 |     0.020448 |    1.247

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   17 | Accept |     0.19651 |      104.12 |     0.19651 |     0.19651 |    0.0016556 |    0.000

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   18 | Accept |     0.19651 |      85.263 |     0.19651 |     0.19651 |    0.0047914 |   2.3289

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   19 | Accept |     0.19651 |      52.102 |     0.19651 |     0.19651 |       90.015 |   0.0002

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|   20 | Accept |     0.19651 |      82.238 |     0.19651 |     0.19651 |      0.68775 |   2.7178

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        sigma |       la
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|   21 | Accept |     0.19651 |      49.468 |     0.19651 |     0.19651 |       49.073 |   0.0001

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   22 | Accept |     0.19651 |      49.183 |     0.19651 |     0.19651 |       25.955 |   8.4946

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   23 | Accept |     0.19651 |      84.781 |     0.19651 |     0.19651 |     0.002241 |   1.6284

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
```

```
Evaluation completed in 1.2 sec
|   24 | Accept |     0.19651 |      90.023 |     0.19651 |     0.19651 |     0.060661 |   0.0004

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   25 | Accept |     0.19651 |      87.349 |     0.19651 |     0.19651 |     0.035771 |    0.002

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|   26 | Accept |     0.19651 |      49.932 |     0.19651 |     0.19651 |       713.45 |   3.5177

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   27 | Accept |     0.19651 |      87.169 |     0.19651 |     0.19651 |     0.012395 |   1.8186

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   28 | Accept |     0.19651 |       94.87 |     0.19651 |     0.19651 |     0.042872 |    0.001

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.2 sec
|   29 | Best   |     0.10795 |      37.932 |     0.10795 |     0.19346 |       1.5886 |   4.9128

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.1 sec
Evaluation completed in 1.3 sec
|   30 | Accept |     0.19651 |      52.241 |     0.10795 |     0.19356 |       236.64 |   5.0506
```


Objective function model
Estimated objective function value
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
Observed points
Model mean
Next point
Model minimum feasible
10^{-2}
10^{-4}
10^{-6}
lambda
10^{-2}
10^{0}
10^{2}
sigma




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 2455.5118 seconds
Total objective function evaluation time: 2346.8025

Best observed feasible point:
    sigma       lambda
    ______    __________

    1.5886    4.9128e-07

Observed objective function value = 0.10795
Estimated objective function value = 0.19356
Function evaluation time = 37.9317

Best estimated feasible point (according to models):
    sigma       lambda
    ______    __________

    1.5886    4.9128e-07

Estimated objective function value = 0.19356
Estimated function evaluation time = 66.1901

results =
  BayesianOptimization with properties:
```

```
                    ObjectiveFcn: @(z)gather(loss(fitckernel(Ztrain,Ytrain,'KernelScale',z.sigm
              VariableDescriptions: [1×2 optimizableVariable]
                           Options: [1×1 struct]
                      MinObjective: 0.1079
                   XAtMinObjective: [1×2 table]
             MinEstimatedObjective: 0.1936
          XAtMinEstimatedObjective: [1×2 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 2.4555e+03
                         NextPoint: [1×2 table]
                             XTrace: [30×2 table]
                    ObjectiveTrace: [30×1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30×1 cell}
    ObjectiveEvaluationTimeTrace: [30×1 double]
              IterationTimeTrace: [30×1 double]
                      ErrorTrace: [30×1 double]
                FeasibilityTrace: [30×1 logical]
     FeasibilityProbabilityTrace: [30×1 double]
             IndexOfMinimumTrace: [30×1 double]
          ObjectiveMinimumTrace: [30×1 double]
EstimatedObjectiveMinimumTrace: [30×1 double]
```

Return the best feasible point in the Bayesian model `results` by using the `bestPoint` function. Use the default criterion `min-visited-upper-confidence-interval`, which determines the best feasible point as the visited point that minimizes an upper confidence interval on the objective function value.

```
zbest = bestPoint(results)
```

```
zbest=1×2 table
    sigma       lambda
    ______    __________

    1.5886    4.9128e-07
```

The table `zbest` contains the optimal estimated values for the `'KernelScale'` and `'Lambda'` name-value arguments. You can specify these values when training a new optimized kernel classifier by using

```
Mdl = fitckernel(Ztrain,Ytrain,'KernelScale',zbest.sigma,'Lambda',zbest.lambda)
```

For tall arrays, the optimization procedure can take a long time. If the data set is too large to run the optimization procedure, you can try to optimize the parameters by using only partial data. Use the `datasample` function and specify `'Replace','false'` to sample data without replacement.

See Also

`bayesopt` | `bestPoint` | `cvpartition` | `datastore` | `fitckernel` | `gather` | `loss` | `optimizableVariable` | `tall`

Statistics and Machine Learning with Big Data Using Tall Arrays

This example shows how to perform statistical analysis and machine learning on out-of-memory data with MATLAB® and Statistics and Machine Learning Toolbox™.

Tall arrays and tables are designed for working with out-of-memory data. This type of data consists of a very large number of rows (observations) compared to a smaller number of columns (variables). Instead of writing specialized code that takes into account the huge size of the data, such as with MapReduce, you can use tall arrays to work with large data sets in a manner similar to in-memory MATLAB arrays. The fundamental difference is that tall arrays typically remain unevaluated until you request that the calculations be performed.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

This example works with a subset of data on a single computer to develop a linear regression model, and then it scales up to analyze all of the data set. You can scale up this analysis even further to:

- Work with data that cannot be read into memory
- Work with data distributed across clusters using MATLAB Parallel Server™
- Integrate with big data systems like Hadoop® and Spark®

Introduction to Machine Learning with Tall Arrays

Several unsupervised and supervised learning algorithms in Statistics and Machine Learning Toolbox are available to work with tall arrays to perform data mining and predictive modeling with out-of-memory data. These algorithms are appropriate for out-of-memory data and can include slight variations from the in-memory algorithms. Capabilities include:

- k-Means clustering
- Linear regression
- Generalized linear regression
- Logistic regression
- Discriminant analysis

The machine learning workflow for out-of-memory data in MATLAB is similar to in-memory data:

1. Preprocess
2. Explore
3. Develop model
4. Validate model
5. Scale up to larger data

This example follows a similar structure in developing a predictive model for airline delays. The data includes a large file of airline flight information from 1987 through 2008. The example goal is to predict the departure delay based on a number of variables.

Details on the fundamental aspects of tall arrays are included in the example “Analyze Big Data in MATLAB Using Tall Arrays”. This example extends the analysis to include machine learning with tall arrays.

Create Tall Table of Airline Data

A datastore is a repository for collections of data that are too large to fit in memory. You can create a datastore from a number of different file formats as the first step to create a tall array from an external data source.

Create a datastore for the sample file `airlinesmall.csv`. Select the variables of interest, treat `'NA'` values as missing data, and generate a preview table of the data.

```
ds = datastore(fullfile(matlabroot,'toolbox','matlab','demos','airlinesmall.csv'));
ds.SelectedVariableNames = {'Year','Month','DayofMonth','DayOfWeek',...
    'DepTime','ArrDelay','DepDelay','Distance'};
ds.TreatAsMissing = 'NA';
pre = preview(ds)
```

```
pre=8×8 table
    Year    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    DepDelay    Distance
    ____    _____    __________    _________    _______    ________    ________    ________

    1987     10          21            3          642          8          12          308
    1987     10          26            1         1021          8           1          296
    1987     10          23            5         2055         21          20          480
    1987     10          23            5         1332         13          12          296
    1987     10          22            4          629          4          -1          373
    1987     10          28            3         1446         59          63          308
    1987     10           8            4          928          3          -2          447
    1987     10          10            6          859         11          -1          954
```

Create a tall table backed by the datastore to facilitate working with the data. The underlying data type of a tall array depends on the type of datastore. In this case, the datastore is tabular text and returns a tall table. The display includes a preview of the data, with indication that the size is unknown.

```
tt = tall(ds)
```

```
tt =

  Mx8 tall table

    Year    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    DepDelay    Distance
    ____    _____    __________    _________    _______    ________    ________    ________

    1987     10          21            3          642          8          12          308
    1987     10          26            1         1021          8           1          296
    1987     10          23            5         2055         21          20          480
    1987     10          23            5         1332         13          12          296
    1987     10          22            4          629          4          -1          373
    1987     10          28            3         1446         59          63          308
    1987     10           8            4          928          3          -2          447
    1987     10          10            6          859         11          -1          954
     :        :           :            :            :          :           :           :
     :        :           :            :            :          :           :           :
```

Preprocess Data

This example aims to explore the time of day and day of week in more detail. Convert the day of week to categorical data with labels and determine the hour of day from the numeric departure time variable.

```
tt.DayOfWeek = categorical(tt.DayOfWeek,1:7,{'Sun','Mon','Tues',...
    'Wed','Thu','Fri','Sat'});
tt.Hr = discretize(tt.DepTime,0:100:2400,0:23)
```

```
tt =

  Mx9 tall table

    Year    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    DepDelay    Distance    H
    ____    _____    __________    _________    _______    ________    ________    ________    _

    1987     10          21          Tues         642          8          12          308        6
    1987     10          26          Sun         1021          8           1          296       10
    1987     10          23          Thu         2055         21          20          480       20
    1987     10          23          Thu         1332         13          12          296       13
    1987     10          22          Wed          629          4          -1          373        6
    1987     10          28          Tues        1446         59          63          308       14
    1987     10           8          Wed          928          3          -2          447        9
    1987     10          10          Fri          859         11          -1          954        8
     :        :           :            :           :          :           :           :         :
     :        :           :            :           :          :           :           :         :
```

Include only years after 2000 and ignore rows with missing data. Identify data of interest by logical condition.

```
idx = tt.Year >= 2000 & ...
    ~any(ismissing(tt),2);
tt = tt(idx,:);
```

Explore Data by Group

A number of exploratory functions are available for tall arrays. For example, the `grpstats` function calculates grouped statistics of tall arrays. Explore the data by determining the centrality and spread of the data with summary statistics grouped by day of week. Also, explore the correlation between the departure delay and arrival delay.

```
g = grpstats(tt(:,{'ArrDelay','DepDelay','DayOfWeek'}),'DayOfWeek',...
    {'mean','std','skewness','kurtosis'})
```

```
g =

  Mx11 tall table

    GroupLabel    DayOfWeek    GroupCount    mean_ArrDelay    std_ArrDelay    skewness_ArrDelay
    __________    _________    __________    _____________    ____________    _________________

        ?             ?            ?              ?               ?                 ?
        ?             ?            ?              ?               ?                 ?
        ?             ?            ?              ?               ?                 ?
        :             :            :              :               :                 :
        :             :            :              :               :                 :
```

```
C = corr(tt.DepDelay,tt.ArrDelay)
```

```
C =

  MxNx... tall array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

These commands produce more tall arrays. The commands are not executed until you explicitly gather the results into the workspace. The `gather` command triggers execution and attempts to minimize the number of passes required through the data to perform the calculations. `gather` requires that the resulting variables fit into memory.

```
[statsByDay,C] = gather(g,C)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 4.1 sec
Evaluation completed in 5.8 sec

statsByDay=7×11 table
    GroupLabel    DayOfWeek    GroupCount    mean_ArrDelay    std_ArrDelay    skewness_ArrDelay
    __________    _________    __________    _____________    ____________    _________________

     {'Fri' }       Fri           7339          4.1512           32.1             7.082
     {'Mon' }       Mon           8443          5.2487         32.453            4.5811
     {'Sat' }       Sat           8045           7.132         33.108            3.6457
     {'Sun' }       Sun           8570          7.7515         36.003            5.7943
     {'Thu' }       Thu           8601          10.053          36.18            4.1381
     {'Tues'}       Tues          8381          6.4786         32.322             4.374
     {'Wed' }       Wed           8489          9.3324         37.406            5.1638


C = 0.8966
```

The variables containing the results are now in-memory variables in the Workspace. Based on these calculations, variation occurs in the data and there is correlation between the delays that you can investigate further.

Explore the effect of day of week and hour of day and gain additional statistical information such as the standard error of the mean and the 95% confidence interval for the mean. You can pass the entire tall table and specify which variables to perform calculations on.

```
byDayHr = grpstats(tt,{'Hr','DayOfWeek'},...
    {'mean','sem','meanci'},'DataVar','DepDelay');
byDayHr = gather(byDayHr);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 3.2 sec
Evaluation completed in 4 sec
```

Due to the data partitioning of the tall array, the output might be unordered. Rearrange the data in memory for further exploration.

```
x = unstack(byDayHr(:,{'Hr','DayOfWeek','mean_DepDelay'}),...
    'mean_DepDelay','DayOfWeek');
x = sortrows(x)
```

```
x=24×8 table
    Hr      Sun        Mon        Tues       Wed        Thu        Fri        Sat
    __    _______    _______    _______    _______    _______    _______    _______

     0     38.519     71.914     39.656     34.667         90     25.536     65.579
     1     45.846     27.875       93.6     125.23     52.765     38.091     29.182
     2        NaN         39        102        NaN      78.25       -1.5        NaN
     3        NaN        NaN        NaN        NaN     -377.5       53.5        NaN
     4         -7     -6.2857         -7    -7.3333      -10.5         -5        NaN
     5    -2.2409     -3.7099    -4.0146    -3.9565    -3.5897    -3.5766    -4.1474
     6        0.4     -1.8909    -1.9802    -1.8304    -1.3578    0.84161    -2.2537
     7     3.4173    -0.47222    -0.18893    0.71546       0.08      1.069    -1.3221
     8     2.3759      1.4054      1.6745     2.2345     2.9668     1.6727    0.88213
     9     2.5325      1.6805      2.7656      2.683     5.6138     3.4838     2.5011
    10       6.37      5.2868      3.6822     7.5773     5.3372     6.9391     4.9979
    11     6.9946      4.9165      5.5639     5.5936     7.0435     4.8989     5.2839
    12      5.673      5.1193      5.7081     7.9178     7.5269     8.0625     7.4686
    13     8.0879      7.1017      5.0857     8.8082     8.2878     8.0675     6.2107
    14     9.5164      5.8343       7.416     9.5954     8.6667     6.0677      8.444
    15     8.1257      4.8802      7.4726     9.8674     10.235      7.167     8.6219
      ⋮
```

Visualize Data in Tall Arrays

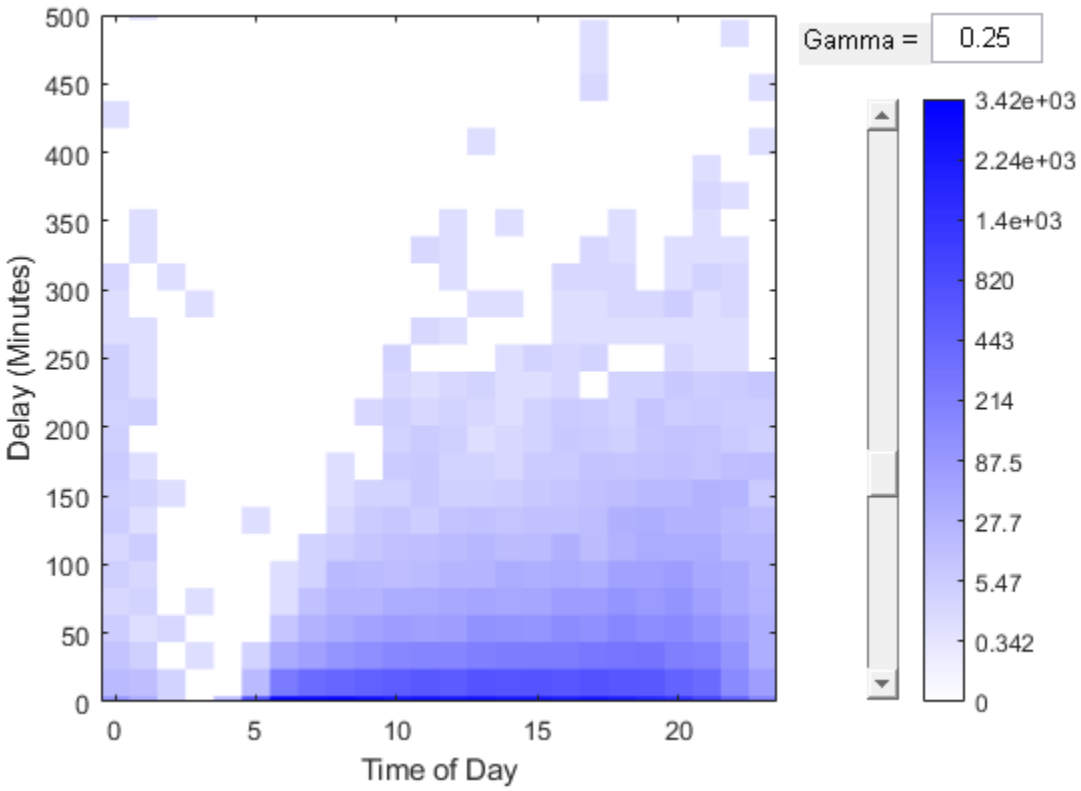
Currently, you can visualize tall array data using `histogram`, `histogram2`, `binScatterPlot`, and `ksdensity`. The visualizations all trigger execution, similar to calling the `gather` function.

Use `binScatterPlot` to examine the relationship between the `Hr` and `DepDelay` variables.

```
binScatterPlot(tt.Hr,tt.DepDelay,'Gamma',0.25)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.8 sec
Evaluation completed in 2.3 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.5 sec
Evaluation completed in 1.6 sec
```

```
ylim([0 500])
xlabel('Time of Day')
ylabel('Delay (Minutes)')
```



As noted in the output display, the visualizations often take two passes through the data: one to perform the binning, and one to perform the binned calculation and produce the visualization.

Split Data into Training and Validation Sets

To develop a machine learning model, it is useful to reserve part of the data to train and develop the model and another part of the data to test the model. A number of ways exist for you to split the data into training and validation sets.

Use `datasample` to obtain a random sampling of the data. Then use `cvpartition` to partition the data into test and training sets. To obtain nonstratified partitions, set a uniform grouping variable by multiplying the data samples by zero.

For reproducibility, set the seed of the random number generator using `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
tallrng('default')
data = datasample(tt,25000,'Replace',false);
groups = 0*data.DepDelay;
y = cvpartition(groups,'HoldOut',1/3);
dataTrain = data(training(y),:);
dataTest = data(test(y),:);
```

Fit Supervised Learning Model

Build a model to predict the departure delay based on several variables. The linear regression model function `fitlm` behaves similarly to the in-memory function. However, calculations with tall arrays result in a `CompactLinearModel`, which is more efficient for large data sets. Model fitting triggers execution because it is an iterative process.

```
model = fitlm(dataTrain,'ResponseVar','DepDelay')
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 1.6 sec
- Pass 2 of 2: Completed in 5.4 sec
Evaluation completed in 7.8 sec

model =
Compact linear regression model:
    DepDelay ~ [Linear formula with 9 terms in 8 predictors]

Estimated Coefficients:
                       Estimate         SE         tStat        pValue
                      __________    __________    ________    __________

    (Intercept)           30.715        75.873     0.40482       0.68562
    Year                -0.01585      0.037853    -0.41872       0.67543
    Month                0.03009      0.028097      1.0709       0.28421
    DayofMonth        -0.0094266      0.010903    -0.86457       0.38729
    DayOfWeek_Mon       -0.36333       0.35527     -1.0227       0.30648
    DayOfWeek_Tues       -0.2858       0.35245    -0.81091       0.41743
    DayOfWeek_Wed       -0.56082       0.35309     -1.5883       0.11224
    DayOfWeek_Thu       -0.25295       0.35239    -0.71782       0.47288
    DayOfWeek_Fri        0.91768       0.36625      2.5056      0.012234
    DayOfWeek_Sat        0.45668       0.35785      1.2762       0.20191
    DepTime            -0.011551     0.0053851      -2.145      0.031964
    ArrDelay              0.8081      0.002875      281.08             0
    Distance           0.0012881    0.00016887      7.6281    2.5106e-14
    Hr                    1.4058       0.53785      2.6138     0.0089613


Number of observations: 16667, Error degrees of freedom: 16653
Root Mean Squared Error: 12.4
R-squared: 0.834,  Adjusted R-Squared: 0.833
F-statistic vs. constant model: 6.41e+03, p-value = 0
```

Predict and Validate the Model

The display indicates fit information, as well as coefficients and associated coefficient statistics.

The `model` variable contains information about the fitted model as properties, which you can access using dot notation. Alternatively, double click the variable in the Workspace to explore the properties interactively.

```
model.Rsquared
```

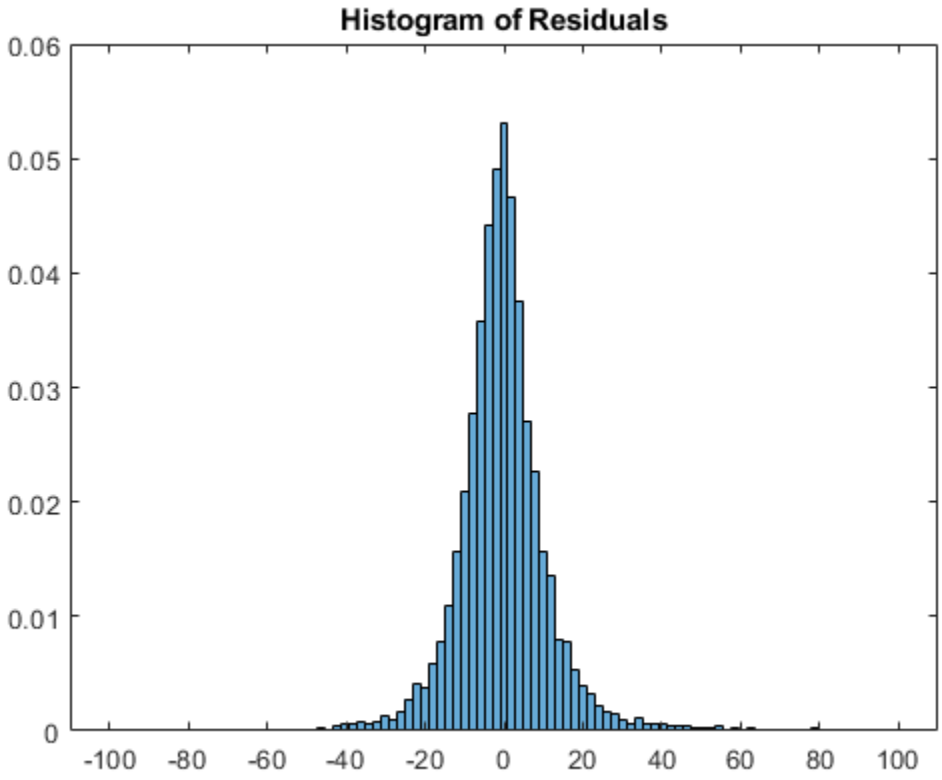
```
ans = struct with fields:
    Ordinary: 0.8335
    Adjusted: 0.8334
```

Predict new values based on the model, calculate the residuals, and visualize using a histogram. The `predict` function predicts new values for both tall and in-memory data.

```
pred = predict(model,dataTest);
err = pred - dataTest.DepDelay;
figure
histogram(err,'BinLimits',[-100 100],'Normalization','pdf')
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 3 sec
- Pass 2 of 2: Completed in 1.9 sec
Evaluation completed in 5.6 sec
```

```
title('Histogram of Residuals')
```



Assess and Adjust Model

Looking at the output p-values in the display, some variables might be unnecessary in the model. You can reduce the complexity of the model by removing these variables.

Examine the significance of the variables in the model more closely using `anova`.

```
a = anova(model)
```

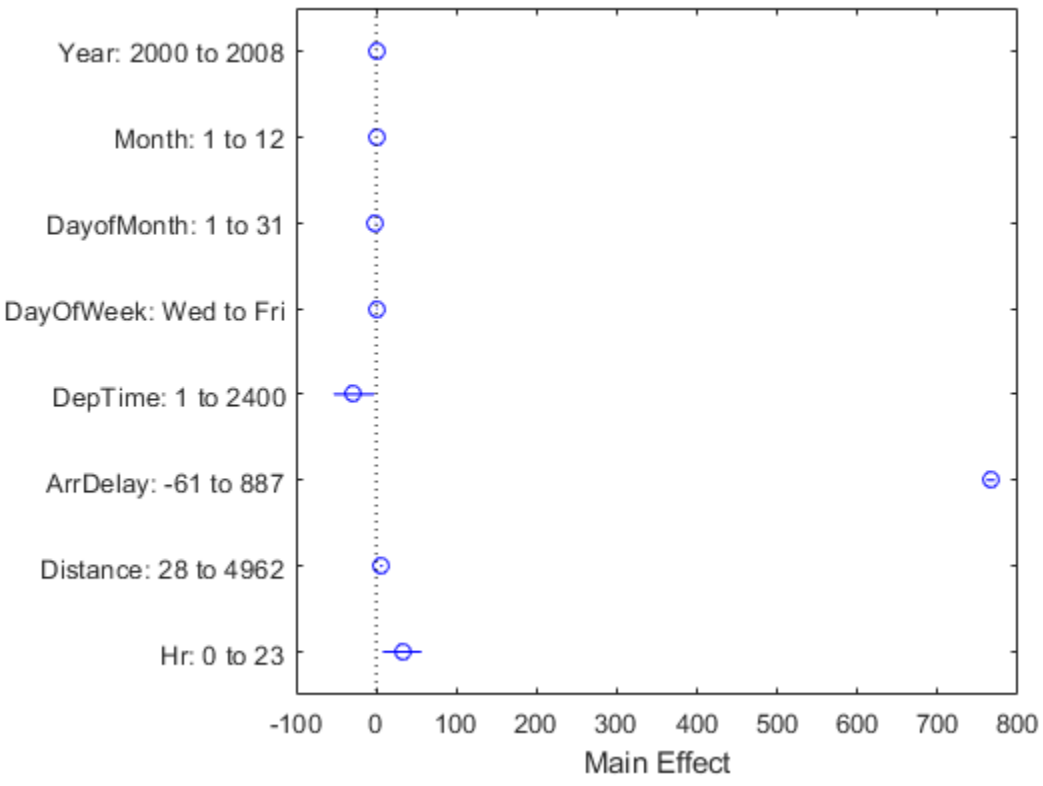
```
a=9×5 table
                SumSq        DF        MeanSq          F          pValue
              _________    _____    ___________    _______    ___________
```

```
Year              26.88        1         26.88    0.17533       0.67543
Month             175.84       1         175.84    1.1469       0.28421
DayofMonth        114.6        1         114.6    0.74749       0.38729
DayOfWeek         3691.4       6         615.23    4.0129     0.00050851
DepTime           705.42       1         705.42    4.6012       0.031964
ArrDelay       1.2112e+07      1      1.2112e+07    79004              0
Distance          8920.9       1         8920.9    58.188     2.5106e-14
Hr                1047.5       1         1047.5    6.8321      0.0089613
Error          2.5531e+06   16653        153.31
```

Based on the p-values, the variables `Year`, `Month`, and `DayOfMonth` are not significant to this model, so you can remove them without negatively affecting the model quality.

To explore these model parameters further, use interactive visualizations such as `plotSlice`, `plotInteractions`, and `plotEffects`. For example, use `plotEffects` to examine the estimated effect that each predictor variable has on the departure delay.

```
plotEffects(model)
```



Based on these calculations, `ArrDelay` is the main effect in the model (it is highly correlated to `DepDelay`). The other effects are observable, but have much less impact. In addition, `Hr` was determined from `DepTime`, so only one of these variables is necessary to the model.

Reduce the number of variables to exclude all date components, and then fit a new model.

```
model2 = fitlm(dataTrain,'DepDelay ~ DepTime + ArrDelay + Distance')
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 2.4 sec
Evaluation completed in 2.5 sec

model2 =
Compact linear regression model:
    DepDelay ~ 1 + DepTime + ArrDelay + Distance

Estimated Coefficients:
                   Estimate         SE         tStat       pValue
                   _________    __________    _______    __________

    (Intercept)      -1.4646       0.31696    -4.6207    3.8538e-06
    DepTime        0.0025087    0.00020401     12.297    1.3333e-34
    ArrDelay         0.80767     0.0028712      281.3             0
    Distance       0.0012981    0.00016886     7.6875    1.5838e-14


Number of observations: 16667, Error degrees of freedom: 16663
Root Mean Squared Error: 12.4
R-squared: 0.833,  Adjusted R-Squared: 0.833
F-statistic vs. constant model: 2.77e+04, p-value = 0
```

Model Development

Even with the model simplified, it can be useful to further adjust the relationships between the variables and include specific interactions. To experiment further, repeat this workflow with smaller tall arrays. For performance while tuning the model, you can consider working with a small extraction of in-memory data before scaling up to the entire tall array.

In this example, you can use functionality like stepwise regression, which is suited for iterative, in-memory model development. After tuning the model, you can scale up to use tall arrays.

Gather a subset of the data into the workspace and use `stepwiselm` to iteratively develop the model in memory.

```
subset = gather(dataTest);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.5 sec
Evaluation completed in 1.6 sec
```

```
sModel = stepwiselm(subset,'ResponseVar','DepDelay')
```

```
1. Adding ArrDelay, FStat = 42200.3016, pValue = 0
2. Adding DepTime, FStat = 51.7918, pValue = 6.70647e-13
3. Adding DepTime:ArrDelay, FStat = 42.4982, pValue = 7.48624e-11
4. Adding Distance, FStat = 15.4303, pValue = 8.62963e-05
5. Adding ArrDelay:Distance, FStat = 231.9012, pValue = 1.135326e-51
6. Adding DayOfWeek, FStat = 3.4704, pValue = 0.0019917
7. Adding DayOfWeek:ArrDelay, FStat = 26.334, pValue = 3.16911e-31
8. Adding DayOfWeek:DepTime, FStat = 2.1732, pValue = 0.042528

sModel =
Linear regression model:
    DepDelay ~ [Linear formula with 9 terms in 4 predictors]

Estimated Coefficients:
```

```
                              Estimate         SE         tStat        pValue
                             ___________    __________    ________    __________

    (Intercept)                 1.1799        1.0675      1.1053       0.26904
    DayOfWeek_Mon              -2.1377        1.4298     -1.4951       0.13493
    DayOfWeek_Tues             -4.2868        1.4683     -2.9196     0.0035137
    DayOfWeek_Wed              -1.6233         1.476     -1.0998       0.27145
    DayOfWeek_Thu             -0.74772        1.5226    -0.49109       0.62338
    DayOfWeek_Fri              -1.7618        1.5079     -1.1683        0.2427
    DayOfWeek_Sat              -2.1121        1.5214     -1.3882       0.16511
    DepTime                  7.5229e-05    0.00073613     0.10219        0.9186
    ArrDelay                     0.8671      0.013836      62.669             0
    Distance                  0.0015163    0.00023426      6.4728    1.0167e-10
    DayOfWeek_Mon:DepTime     0.0017633     0.0010106      1.7448      0.081056
    DayOfWeek_Tues:DepTime    0.0032578     0.0010331      3.1534     0.0016194
    DayOfWeek_Wed:DepTime    0.00097506      0.001044     0.93398       0.35034
    DayOfWeek_Thu:DepTime     0.0012517     0.0010694      1.1705       0.24184
    DayOfWeek_Fri:DepTime     0.0026464     0.0010711      2.4707      0.013504
    DayOfWeek_Sat:DepTime     0.0021477     0.0010646      2.0174      0.043689
    DayOfWeek_Mon:ArrDelay     -0.11023      0.014744     -7.4767     8.399e-14
    DayOfWeek_Tues:ArrDelay    -0.14589      0.014814     -9.8482    9.2943e-23
    DayOfWeek_Wed:ArrDelay    -0.041878      0.012849     -3.2593     0.0011215
    DayOfWeek_Thu:ArrDelay    -0.096741      0.013308     -7.2693    3.9414e-13
    DayOfWeek_Fri:ArrDelay    -0.077713      0.015462     -5.0259    5.1147e-07
    DayOfWeek_Sat:ArrDelay     -0.13669      0.014652      -9.329    1.3471e-20
    DepTime:ArrDelay          6.4148e-05    7.7372e-06      8.2909    1.3002e-16
    ArrDelay:Distance        -0.00010512    7.3888e-06     -14.227    2.1138e-45


Number of observations: 8333, Error degrees of freedom: 8309
Root Mean Squared Error: 12
R-squared: 0.845,  Adjusted R-Squared: 0.845
F-statistic vs. constant model: 1.97e+03, p-value = 0
```

The model that results from the stepwise fit includes interaction terms.

Now try to fit a model for the tall data by using `fitlm` with the formula returned by `stepwiselm`.

```
model3 = fitlm(dataTrain,sModel.Formula)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 2.2 sec
Evaluation completed in 2.3 sec

model3 =
Compact linear regression model:
    DepDelay ~ [Linear formula with 9 terms in 4 predictors]

Estimated Coefficients:
                              Estimate         SE         tStat        pValue
                             ___________    __________    ________    __________

    (Intercept)               -0.31595       0.74499     -0.4241        0.6715
    DayOfWeek_Mon             -0.64218        1.0473    -0.61316       0.53978
    DayOfWeek_Tues            -0.90163        1.0383    -0.86836       0.38521
    DayOfWeek_Wed              -1.0798        1.0417     -1.0365       0.29997
    DayOfWeek_Thu              -3.2765        1.0379      -3.157     0.0015967
    DayOfWeek_Fri              0.44193        1.0813     0.40869       0.68277
```

```
    DayOfWeek_Sat                1.1428       1.0777      1.0604       0.28899
    DepTime                   0.0014188    0.00051612      2.7489     0.0059853
    ArrDelay                    0.72526      0.011907      60.913             0
    Distance                  0.0014824    0.00017027      8.7059    3.4423e-18
    DayOfWeek_Mon:DepTime     0.00040994    0.00073548     0.55738       0.57728
    DayOfWeek_Tues:DepTime    0.00051826    0.00073645     0.70373       0.48161
    DayOfWeek_Wed:DepTime     0.00058426    0.00073695     0.79281        0.4279
    DayOfWeek_Thu:DepTime      0.0026229    0.00073649      3.5614    0.00036991
    DayOfWeek_Fri:DepTime      0.0002959    0.00077194     0.38332       0.70149
    DayOfWeek_Sat:DepTime     -0.00060921    0.00075776    -0.80396       0.42143
    DayOfWeek_Mon:ArrDelay     -0.034886      0.010435     -3.3432    0.00082993
    DayOfWeek_Tues:ArrDelay    -0.0073661      0.010113    -0.72837        0.4664
    DayOfWeek_Wed:ArrDelay      -0.028158     0.0099004     -2.8441     0.0044594
    DayOfWeek_Thu:ArrDelay      -0.061065      0.010381     -5.8821    4.1275e-09
    DayOfWeek_Fri:ArrDelay       0.052437      0.010927      4.7987    1.6111e-06
    DayOfWeek_Sat:ArrDelay       0.014205       0.01039      1.3671        0.1716
    DepTime:ArrDelay           7.2632e-05    5.3946e-06      13.464     4.196e-41
    ArrDelay:Distance         -2.4743e-05    4.6508e-06     -5.3203    1.0496e-07


Number of observations: 16667, Error degrees of freedom: 16643
Root Mean Squared Error: 12.3
R-squared: 0.837,  Adjusted R-Squared: 0.836
F-statistic vs. constant model: 3.7e+03, p-value = 0
```

You can repeat this process to continue to adjust the linear model. However, in this case, you should explore different types of regression that might be more appropriate for this data. For example, if you do not want to include the arrival delay, then this type of linear model is no longer appropriate. See “Logistic Regression with Tall Arrays” on page 30-2 for more information.

Scale to Spark

A key capability of tall arrays in MATLAB and Statistics and Machine Learning Toolbox is the connectivity to platforms such as Hadoop and Spark. You can even compile the code and run it on Spark using MATLAB Compiler™. See “Extend Tall Arrays with Other Products” for more information about using these products:

- Database Toolbox™
- Parallel Computing Toolbox™
- MATLAB® Parallel Server™
- MATLAB Compiler™

See Also

More About

- Function List (Tall Arrays)

Parallel Statistics

Quick Start Parallel Computing for Statistics and Machine Learning Toolbox

Note To use parallel computing as described in this chapter, you must have a Parallel Computing Toolbox license.

What Is Parallel Statistics Functionality?

You can use any of the Statistics and Machine Learning Toolbox functions with Parallel Computing Toolbox constructs such as `parfor` and `spmd`. However, some functions, such as those with interactive displays, can lose functionality in parallel. In particular, displays and interactive usage are not effective on workers (see “Vocabulary for Parallel Computation” on page 31-8).

Additionally, the following functions are enhanced to use parallel computing internally. These functions use `parfor` internally to parallelize calculations.

- `bayesopt`
- `bootci`
- `bootstrp`
- `candexch`
- `compare` of `LinearMixedModel`
- `cordexch`
- `crossval`
- `daugment`
- `dcovary`
- `fitcensemble`
- `fitrensemble`
- `jackknife`
- `kmeans`
- `kmedoids`
- `lasso`
- `lassoglm`
- `nnmf`
- `oobPermutedPredictorImportance` of `ClassificationBaggedEnsemble`
- `oobPermutedPredictorImportance` of `RegressionBaggedEnsemble`
- `perfcurve`
- `plotPartialDependence`
- `plsregress`
- `rowexch`
- `sequentialfs`
- `testckfold`

- `TreeBagger`
- `growTrees` of `TreeBagger`

The following functions for fitting multiclass models for support vector machines and other classifiers are also enhanced to use parallel computing internally.

- `fitcecoc`
- Methods of the class `ClassificationECOC`:
  - `resubEdge`
  - `resubLoss`
  - `resubMargin`
  - `resubPredict`
  - `crossval`
- Methods of the class `CompactClassificationECOC`
  - `edge`
  - `loss`
  - `margin`
  - `predict`
- Methods of the class `ClassificationPartitionedECOC`
  - `kfoldEdge`
  - `kfoldLoss`
  - `kfoldMargin`
  - `kfoldPredict`
- Methods of the class `ClassificationPartitionedLinearECOC`
  - `kfoldEdge`
  - `kfoldLoss`
  - `kfoldMargin`
  - `kfoldPredict`

The following functions perform hyperparameter optimization in parallel.

- `fitcdiscr`
- `fitcecoc`
- `fitctree`
- `fitclinear`
- `fitcknn`
- `fitrgp`
- `fitrtree`
- `fitcensemble`
- `fitcnb`
- `fitcsvm`

- `fitrensemble`
- `fitrlinear`
- `fitrsvm`
- `fitrkernel`
- `fitckernel`

This chapter gives the simplest way to use these enhanced functions in parallel. For more advanced topics, including the issues of reproducibility and nested `parfor` loops, see the other sections in this chapter.

For information on parallel statistical computing at the command line, enter

```
help parallelstats
```

How To Compute in Parallel

To have a function compute in parallel:

1. “Set Up a Parallel Environment” on page 31-4
2. “Set the UseParallel Option to true” on page 31-4
3. “Call the Function Using the Options Structure” on page 31-4

Set Up a Parallel Environment

To run a statistical computation in parallel, first set up a parallel environment.

Note Setting up a parallel environment can take several seconds.

For a multicore machine, enter the following at the MATLAB command line:

```
parpool(n)
```

n is the number of workers you want to use.

You can also run parallel code in MATLAB Online™. For details, see “Use Parallel Computing Toolbox with Cloud Center Cluster in MATLAB Online” (Parallel Computing Toolbox).

Set the UseParallel Option to true

Create an options structure with the `statset` function. To run in parallel, set the `UseParallel` option to `true`:

```
paroptions = statset('UseParallel',true);
```

Call the Function Using the Options Structure

Call your function with syntax that uses the options structure. For example:

```
% Run crossval in parallel
cvMse = crossval('mse',x,y,'predfun',regf,'Options',paroptions);

% Run bootstrp in parallel
```

```
sts = bootstrp(100,@(x)[mean(x) std(x)],y,'Options',paroptions);

% Run TreeBagger in parallel
b = TreeBagger(50,meas,spec,'OOBPred','on','Options',paroptions);
```

For more complete examples of parallel statistical functions, see “Use Parallel Processing for Regression TreeBagger Workflow” on page 31-6, “Implement Jackknife Using Parallel Computing” on page 31-17, “Implement Cross-Validation Using Parallel Computing” on page 31-18, and “Implement Bootstrap Using Parallel Computing” on page 31-20.

After you have finished computing in parallel, close the parallel environment:

```
delete mypool
```

Tip To save time, keep the pool open if you expect to compute in parallel again soon.

Use Parallel Processing for Regression TreeBagger Workflow

This example shows you how to:

- Use an ensemble of bagged regression trees to estimate feature importance.
- Improve computation speed by using parallel computing.

The sample data is a database of 1985 car imports with 205 observations, 25 predictors, and 1 response, which is insurance risk rating, or "symboling." The first 15 variables are numeric and the last 10 are categorical. The symboling index takes integer values from -3 to 3.

Load the sample data and separate it into predictor and response arrays.

```
load imports-85;
Y = X(:,1);
X = X(:,2:end);
```

Set up the parallel environment to use the default number of workers. The computer that created this example has six cores.

```
mypool = parpool
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

mypool =

 ProcessPool with properties:

            Connected: true
           NumWorkers: 6
              Cluster: local
        AttachedFiles: {}
    AutoAddClientPath: true
          IdleTimeout: 30 minutes (30 minutes remaining)
          SpmdEnabled: true
```

Set the options to use parallel processing.

```
paroptions = statset('UseParallel',true);
```

Estimate feature importance using leaf size `1` and `5000` trees in parallel. Time the function for comparison purposes.

```
tic
b = TreeBagger(5000,X,Y,'Method','r','OOBVarImp','on', ...
    'cat',16:25,'MinLeafSize',1,'Options',paroptions);
toc
```

```
Elapsed time is 9.873065 seconds.
```

Perform the same computation in serial for timing comparison.

```
tic
b = TreeBagger(5000,X,Y,'Method','r','OOBVarImp','on', ...
    'cat',16:25,'MinLeafSize',1);
toc
```

```
Elapsed time is 28.092654 seconds.
```

The results show that computing in parallel takes a fraction of the time it takes to compute serially. Note that the elapsed time can vary depending on your operating system.

See Also

`TreeBagger` | `parpool` | `statset`

Related Examples

- “Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
- “Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
- “Comparison of TreeBagger and Bagged Ensembles” on page 18-44

Concepts of Parallel Computing in Statistics and Machine Learning Toolbox

In this section...

Subtleties in Parallel Computing

There are two main subtleties in parallel computations:

- Nested parallel evaluations (see “No Nested parfor Loops” on page 31-11). Only the outermost `parfor` loop runs in parallel, the others run serially.
- Reproducible results when using random numbers (see “Reproducibility in Parallel Statistical Computations” on page 31-13). How can you get exactly the same results when repeatedly running a parallel computation that uses random numbers?

Vocabulary for Parallel Computation

- worker — An independent MATLAB session that runs code distributed by the client.
- client — The MATLAB session with which you interact, and that distributes jobs to workers.
- `parfor` — A Parallel Computing Toolbox function that distributes independent code segments to workers (see “Working with parfor” on page 31-11).
- random stream — A pseudorandom number generator, and the sequence of values it generates. MATLAB implements random streams with the `RandStream` class.
- reproducible computation — A computation that can be exactly replicated, even in the presence of random numbers (see “Reproducibility in Parallel Statistical Computations” on page 31-13).

When to Run Statistical Functions in Parallel

In this section...

Why Run in Parallel?

The main reason to run statistical computations in parallel is to gain speed, meaning to reduce the execution time of your program or functions. “Factors Affecting Speed” on page 31-9 discusses the main items affecting the speed of programs or functions. “Factors Affecting Results” on page 31-9 discusses details that can cause a parallel run to give different results than a serial run.

Note Some Statistics and Machine Learning Toolbox functions have built-in parallel computing capabilities. See Quick Start Parallel Computing for Statistics and Machine Learning Toolbox on page 31-2. You can also use any Statistics and Machine Learning Toolbox functions with Parallel Computing Toolbox functions such as `parfor` loops. To decide when to call functions in parallel, consider the factors affecting speed and results.

Factors Affecting Speed

Some factors that can affect the speed of execution of parallel processing are:

- Parallel environment setup. It takes time to run `parpool` to begin computing in parallel. If your computation is fast, the setup time can exceed any time saved by computing in parallel.
- Parallel overhead. There is overhead in communication and coordination when running in parallel. If function evaluations are fast, this overhead could be an appreciable part of the total computation time. Thus, solving a problem in parallel can be slower than solving the problem serially. For an example, see Improving Optimization Performance with Parallel Computing in MATLAB Digest, March 2009.
- No nested `parfor` loops. This is described in “Working with parfor” on page 31-11. `parfor` does not work in parallel when called from within another `parfor` loop. If you have programmed your custom functions to take advantage of parallel processing, the limitation of no nested `parfor` loops can cause a parallel function to run slower than expected.
- When executing serially, `parfor` loops run slightly slower than `for` loops.
- Passing parameters. Parameters are automatically passed to worker sessions during the execution of parallel computations. If there are many parameters, or they take a large amount of memory, passing parameters can slow the execution of your computation.
- Contention for resources: network and computing. If the pool of workers has low bandwidth or high latency, parallel computation can be slow.

Factors Affecting Results

Some factors can affect results when using parallel processing. You might need to adjust your code to run in parallel, for example, you need independent loops and the workers must be able to access the variables. Some important factors are:

- Persistent or global variables. If any functions use persistent or global variables, these variables can take different values on different worker processors. The body of a `parfor` loop cannot contain global or persistent variable declarations.
- Accessing external files. The order of computations is not guaranteed during parallel processing, so external files can be accessed in unpredictable order, leading to unpredictable results. Furthermore, if multiple processors try to read an external file simultaneously, the file can become locked, leading to a read error, and halting function execution.
- Noncomputational functions, such as `input`, `plot`, and `keyboard`, can behave badly when used in your custom functions. Do not use these functions in a `parfor` loop, because they can cause a worker to become nonresponsive, since it is waiting for input.
- `parfor` does not allow `break` or `return` statements.
- The random numbers you use can affect the results of your computations. See “Reproducibility in Parallel Statistical Computations” on page 31-13.

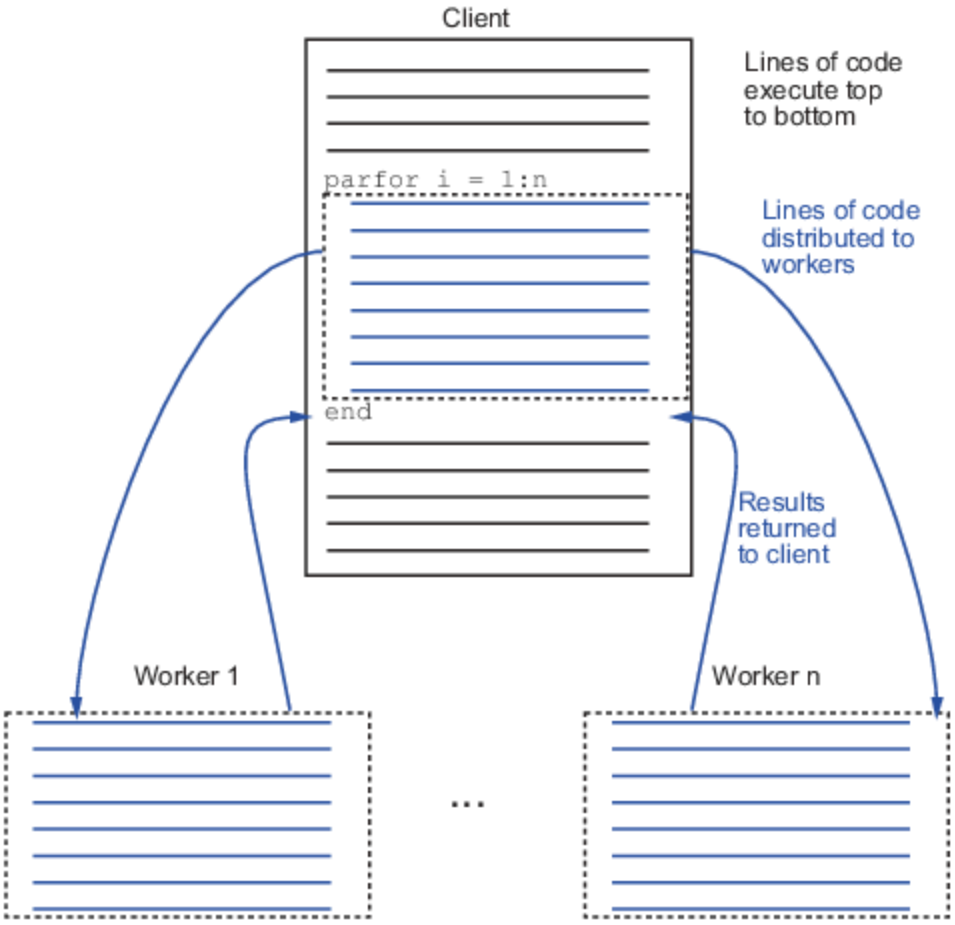
For advice on converting for loops to use `parfor`, see “Parallel for-Loops (parfor)” (Parallel Computing Toolbox).

Working with parfor

In this section...

How Statistical Functions Use parfor

`parfor` is a Parallel Computing Toolbox function similar to a `for` loop. Parallel statistical functions call `parfor` internally. `parfor` distributes computations to worker processors.



Characteristics of parfor

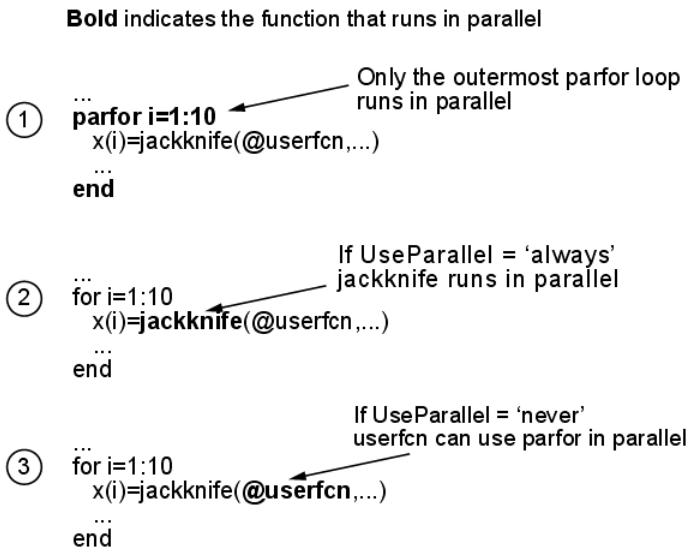
You might need to adjust your code to run in parallel, for example, you need independent loops and the workers must be able to access the variables. For advice on using `parfor`, see “Parallel for-Loops (parfor)” (Parallel Computing Toolbox).

No Nested parfor Loops

`parfor` does not work in parallel when called from within another `parfor` loop, or from an `spmd` block. Parallelization occurs only at the outermost level.

Suppose, for example, you want to apply `jackknife` to your function `userfcn`, which calls `parfor`, and you want to call `jackknife` in a loop. The following figure shows three cases:

1. The outermost loop is `parfor`. Only that loop runs in parallel.
2. The outermost `parfor` loop is in `jackknife`. Only `jackknife` runs in parallel.
3. The outermost `parfor` loop is in `userfcn`. `userfcn` uses `parfor` in parallel.



When `parfor` Runs in Parallel

For help converting nested loops to use `parfor`, see "Convert for-Loops Into parfor-Loops" (Parallel Computing Toolbox).

See also Quick Start Parallel Computing for Statistics and Machine Learning Toolbox on page 31-2.

Reproducibility in Parallel Statistical Computations

In this section...

Issues and Considerations in Reproducing Parallel Computations

A reproducible computation is one that gives the same results every time it runs. Reproducibility is important for:

- Debugging — To correct an anomalous result, you need to reproduce the result.
- Confidence — When you can reproduce results, you can investigate and understand them.
- Modifying existing code — When you change existing code, you want to ensure that you do not break anything.

Generally, you do not need to ensure reproducibility for your computation. Often, when you want reproducibility, the simplest technique is to run in serial instead of in parallel. In serial computation you can simply call the `rng` function as follows:

```
s = rng % Obtain the current state of the random stream
% run the statistical function
rng(s) % Reset the stream to the previous state
% run the statistical function again, obtain identical results
```

This section addresses the case when your function uses random numbers, and you want reproducible results in parallel. This section also addresses the case when you want the same results in parallel as in serial.

Running Reproducible Parallel Computations

To run a Statistics and Machine Learning Toolbox function reproducibly:

1. Set the `UseSubstreams` option to `true` using `statset`.
2. Set the `Streams` option to a type that supports substreams: `'mlfg6331_64'` or `'mrg32k3a'`. For information on these streams, see `RandStream.list`.
3. To compute in parallel, set the `UseParallel` option to `true`.
4. To fit an ensemble in parallel using `fitcensemble` or `fitrensemble`, create a tree template with the `'Reproducible'` name-value pair set to `true`:

   ```
   t = templateTree('Reproducible',true);
   ens = fitcensemble(X,Y,'Method','bag','Learners',t,...
       'Options',options);
   ```
5. Call the function with the options structure.
6. To reproduce the computation, reset the stream, then call the function again.

To understand why this technique gives reproducibility, see “How Substreams Enable Reproducible Parallel Computations” on page 31-14.

For example, to use the 'mlfg6331_64' stream for reproducible computation:

1. Create an appropriate options structure:

```
s = RandStream('mlfg6331_64');
options = statset('UseParallel',true, ...
    'Streams',s,'UseSubstreams',true);
```

2. Run your parallel computation. For instructions, see Quick Start Parallel Computing for Statistics and Machine Learning Toolbox on page 31-2.
3. Reset the random stream:

```
reset(s);
```

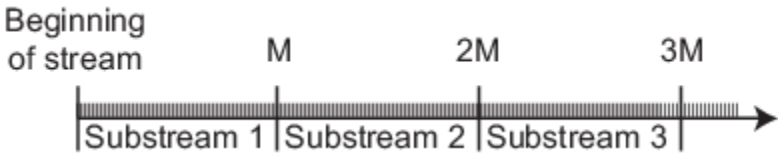
4. Rerun your parallel computation. You obtain identical results.

For examples of parallel computation run this reproducible way, see “Reproducible Parallel Bootstrap” on page 31-21 and “Train Classification Ensemble in Parallel” on page 18-109.

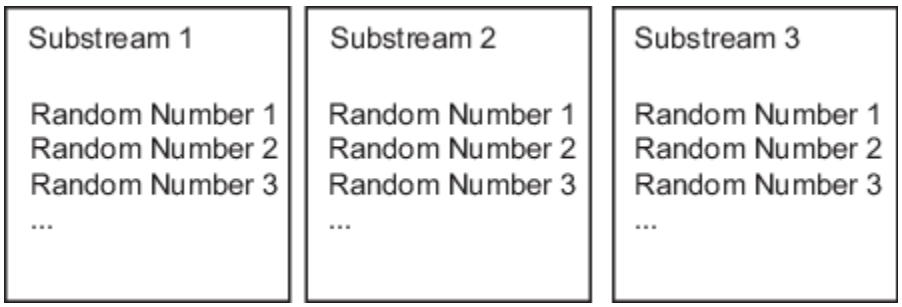
Parallel Statistical Computation Using Random Numbers

What Are Substreams?

A substream is a portion of a random stream that RandStream can access quickly. There is a number M such that for any positive integer k, RandStream can go to the kMth pseudorandom number in the stream. From that point, RandStream can generate the subsequent entries in the stream. Currently, RandStream has $M = 2^{72}$, about 5e21, or more.



The entries in different substreams have good statistical properties, similar to the properties of entries in a single stream: independence, and lack of *k*-way correlation at various lags. The substreams are so long that you can view the substreams as being independent streams, as in the following picture.



Two RandStream stream types support substreams: 'mlfg6331_64' and 'mrg32k3a'.

How Substreams Enable Reproducible Parallel Computations

When MATLAB performs computations in parallel with parfor, each worker receives loop iterations in an unpredictable order. Therefore, you cannot predict which worker gets which iteration, so cannot determine the random numbers associated with each iteration.

Substreams allow MATLAB to tie each iteration to a particular sequence of random numbers. `parfor` gives each iteration an index. The iteration uses the index as the substream number. Since the random numbers are associated with the iterations, not with the workers, the entire computation is reproducible.

To obtain reproducible results, simply reset the stream, and all the substreams generate identical random numbers when called again. This method succeeds when all the workers use the same stream, and the stream supports substreams. This concludes the discussion of how the procedure in “Running Reproducible Parallel Computations” on page 31-13 gives reproducible parallel results.

Random Numbers on the Client or Workers

A few functions generate random numbers on the client before distributing them to parallel workers. The workers do not use random numbers, so operate purely deterministically. For these functions, you can run a parallel computation reproducibly using any random stream type.

The functions that operate this way include:

- `crossval`
- `plsregress`
- `sequentialfs`

To obtain identical results, reset the random stream on the client, or the random stream you pass to the client. For example:

```
s = rng % Obtain the current state of the random stream
% run the statistical function
rng(s) % Reset the stream to the previous state
% run the statistical function again, obtain identical results
```

While this method enables you to run reproducibly in parallel, the results can differ from a serial computation. The reason for the difference is `parfor` loops run in reverse order from `for` loops. Therefore, a serial computation can generate random numbers in a different order than a parallel computation. For unequivocal reproducibility, use the technique in “Running Reproducible Parallel Computations” on page 31-13.

Distributing Streams Explicitly

For testing or comparison using particular random number algorithms, you must set the random number generators. How do you set these generators in parallel, or initialize streams on each worker in a particular way? Or you might want to run a computation using a different sequence of random numbers than any other you have run. How can you ensure the sequence you use is statistically independent?

Parallel Statistics and Machine Learning Toolbox functions allow you to set random streams on each worker explicitly. For information on *creating* multiple streams, enter `help RandStream/create` at the command line. To create four independent streams using the `'mrg32k3a'` generator:

```
s = RandStream.create('mrg32k3a','NumStreams',4,...
    'CellOutput',true);
```

Pass these streams to a statistical function using the `Streams` option. For example:

```
parpool(4) % if you have at least 4 cores
s = RandStream.create('mrg32k3a','NumStreams',4,...
```

```
    'CellOutput',true); % create 4 independent streams
paroptions = statset('UseParallel',true,...
    'Streams',s); % set the 4 different streams
x = [randn(700,1); 4 + 2*randn(300,1)];
latt = -4:0.01:12;
myfun = @(X) ksdensity(X,latt);
pdfestimate = myfun(x);
B = bootstrp(200,myfun,x,'Options',paroptions);
```

This method of distributing streams gives each worker a different stream for the computation. However, it does not allow for a reproducible computation, because the workers perform the 200 bootstraps in an unpredictable order. If you want to perform a reproducible computation, use substreams as described in “Running Reproducible Parallel Computations” on page 31-13.

If you set the `UseSubstreams` option to `true`, then set the `Streams` option to a single random stream of the type that supports substreams (`'mlfg6331_64'` or `'mrg32k3a'`). This setting gives reproducible computations.

Implement Jackknife Using Parallel Computing

This example is from the `jackknife` function reference page, but runs in parallel.

Generate a sample data of size 10000 from a normal distribution with mean 0 and standard deviation 5.

```
sigma = 5;
rng('default')
y = normrnd(0,sigma,10000,1);
```

Run `jackknife` in parallel to estimate the variance. To do this, use `statset` to create the options structure and set the `UseParallel` field to true.

```
opts = statset('UseParallel',true);
m = jackknife(@var,y,1,'Options',opts);
```

Compare the known bias formula with the jackknife bias estimate.

```
n = length(y);
bias = -sigma^2/n % Known bias formula
jbias = (n-1)*(mean(m)-var(y,1)) % jackknife bias estimate
```

```
Starting parallel pool (parpool) using the 'local' profile ...

Connected to the parallel pool (number of workers: 6).

bias =

   -0.0025


jbias =

   -0.0025
```

Compare how long it takes to compute in serial and in parallel.

```
tic;m = jackknife(@var,y,1);toc  % Serial computation
```

```
Elapsed time is 1.638026 seconds.
```

```
tic;m = jackknife(@var,y,1,'Options',opts);toc % Parallel computation
```

```
Elapsed time is 0.507961 seconds.
```

`jackknife` does not use random numbers, so gives the same results every time, whether run in parallel or serial.

Implement Cross-Validation Using Parallel Computing

In this section...

Simple Parallel Cross Validation

In this example, use `crossval` to compute a cross-validation estimate of mean-squared error for a regression model. Run the computations in parallel.

```
mypool = parpool()
Starting parpool using the 'local' profile ... connected to 2 workers.

mypool =

  Pool with properties:

    AttachedFiles: {0x1 cell}
       NumWorkers: 2
      IdleTimeout: 30
          Cluster: [1x1 parallel.cluster.Local]
     RequestQueue: [1x1 parallel.RequestQueue]
      SpmdEnabled: 1


opts = statset('UseParallel',true);

load('fisheriris');
y = meas(:,1);
X = [ones(size(y,1),1),meas(:,2:4)];
regf=@(XTRAIN,ytrain,XTEST)(XTEST*regress(ytrain,XTRAIN));

cvMse = crossval('mse',X,y,'Predfun',regf,'Options',opts)

cvMse =

    0.1028
```

This simple example is not a good candidate for parallel computation:

```
% How long to compute in serial?
tic;cvMse = crossval('mse',X,y,'Predfun',regf);toc
Elapsed time is 0.073438 seconds.

% How long to compute in parallel?
tic;cvMse = crossval('mse',X,y,'Predfun',regf,...
    'Options',opts);toc
Elapsed time is 0.289585 seconds.
```

Reproducible Parallel Cross Validation

To run `crossval` in parallel in a reproducible fashion, set the options and reset the random stream appropriately (see “Running Reproducible Parallel Computations” on page 31-13).

```
mypool = parpool()

Starting parpool using the 'local' profile ... connected to 2 workers.

mypool =

  Pool with properties:

    AttachedFiles: {0x1 cell}
       NumWorkers: 2
      IdleTimeout: 30
          Cluster: [1x1 parallel.cluster.Local]
     RequestQueue: [1x1 parallel.RequestQueue]
      SpmdEnabled: 1

s = RandStream('mlfg6331_64');
opts = statset('UseParallel',true,...
    'Streams',s,'UseSubstreams',true);

load('fisheriris');
y = meas(:,1);
X = [ones(size(y,1),1),meas(:,2:4)];
regf=@(XTRAIN,ytrain,XTEST)(XTEST*regress(ytrain,XTRAIN));

cvMse = crossval('mse',X,y,'Predfun',regf,'Options',opts)

cvMse =

    0.1020
```

Reset the stream:

```
reset(s)
cvMse = crossval('mse',X,y,'Predfun',regf,'Options',opts)

cvMse =

    0.1020
```

Implement Bootstrap Using Parallel Computing

In this section...

Bootstrap in Serial and Parallel

Here is an example timing a bootstrap in parallel versus in serial. The example generates data from a mixture of two Gaussians, constructs a nonparametric estimate of the resulting data, and uses a bootstrap to get a sense of the sampling variability.

1 Generate the data:

```
% Generate a random sample of size 1000,
% from a mixture of two Gaussian distributions
x = [randn(700,1); 4 + 2*randn(300,1)];
```

2 Construct a nonparametric estimate of the density from the data:

```
latt = -4:0.01:12;
myfun = @(X) ksdensity(X,latt);
pdfestimate = myfun(x);
```

3 Bootstrap the estimate to get a sense of its sampling variability. Run the bootstrap in serial for timing comparison.

```
tic;B = bootstrp(200,myfun,x);toc

Elapsed time is 10.878654 seconds.
```

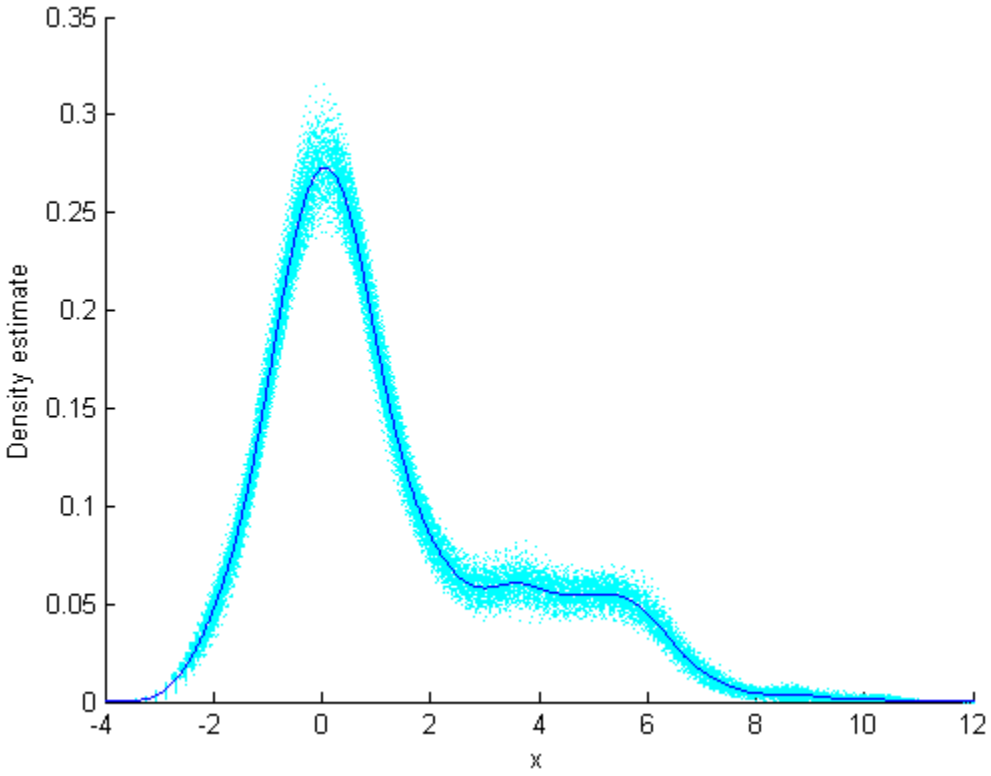
4 Run the bootstrap in parallel for timing comparison:

```
mypool = parpool()
Starting parpool using the 'local' profile ... connected to 2 workers.

mypool =

  Pool with properties:

    AttachedFiles: {0x1 cell}
       NumWorkers: 2
      IdleTimeout: 30
          Cluster: [1x1 parallel.cluster.Local]
     RequestQueue: [1x1 parallel.RequestQueue]
      SpmdEnabled: 1

opt = statset('UseParallel',true);
tic;B = bootstrp(200,myfun,x,'Options',opt);toc

Elapsed time is 6.304077 seconds.
```

Computing in parallel is nearly twice as fast as computing in serial for this example.

Overlay the `ksdensity` density estimate with the 200 bootstrapped estimates obtained in the parallel bootstrap. You can get a sense of how to assess the accuracy of the density estimate from this plot.

```
hold on
for i=1:size(B,1),
    plot(latt,B(i,:),'c:')
end
plot(latt,pdfestimate);
xlabel('x');ylabel('Density estimate')
```



Reproducible Parallel Bootstrap

To run the example in parallel in a reproducible fashion, set the options appropriately (see “Running Reproducible Parallel Computations” on page 31-13). First set up the problem and parallel environment as in “Bootstrap in Serial and Parallel” on page 31-20. Then set the options to use substreams along with a stream that supports substreams.

```
s = RandStream('mlfg6331_64'); % has substreams
opts = statset('UseParallel',true,...
    'Streams',s,'UseSubstreams',true);
B2 = bootstrp(200,myfun,x,'Options',opts);
```

To rerun the bootstrap and get the same result:

```
reset(s) % set the stream to initial state
B3 = bootstrp(200,myfun,x,'Options',opts);
isequal(B2,B3) % check if same results
```

```
ans =
     1
```

Code Generation

Introduction to Code Generation

In this section...

MATLAB Coder generates readable and portable C and C++ code from Statistics and Machine Learning Toolbox functions that support code generation. You can integrate the generated code into your projects as source code, static libraries, or dynamic libraries. You can also use the generated code within the MATLAB environment to accelerate computationally intensive portions of your MATLAB code.

Generating C/C++ code requires MATLAB Coder and has the following limitations:

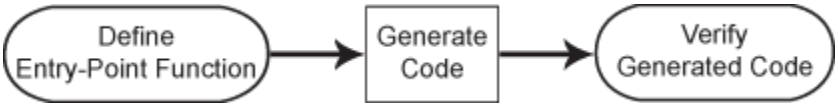
- You cannot call any function at the top level when generating code by using `codegen`. Instead, call the function within an *entry-point* function, and then generate code from the entry-point function. The entry-point function, also known as the *top-level* or *primary* function, is a function you define for code generation. All functions within the entry-point function must support code generation.
- The MATLAB Coder limitations also apply to Statistics and Machine Learning Toolbox for code generation. For details, see “MATLAB Language Features Supported for C/C++ Code Generation” (MATLAB Coder).
- Code generation in Statistics and Machine Learning Toolbox does not support sparse matrices.
- For the code generation usage notes and limitations for each function, see the Code Generation section on the function reference page.

For a list of Statistics and Machine Learning Toolbox functions that support code generation, see Function List (C/C++ Code Generation).

Code Generation Workflows

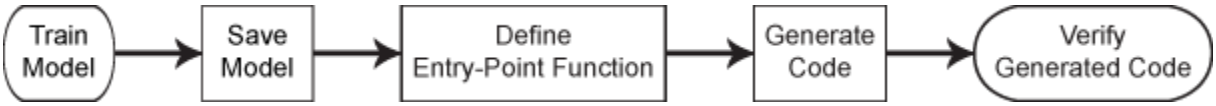
You can generate C/C++ code for the Statistics and Machine Learning Toolbox functions in several ways.

- General code generation workflow for functions that are not the object functions of machine learning models

  Define an *entry-point* function that calls the function that supports code generation, generate C/C++ code for the entry-point function by using `codegen`, and then verify the generated code. The entry-point function, also known as the *top-level* or *primary* function, is a function you define for code generation. Because you cannot call any function at the top level using `codegen`, you must define an entry-point function. All functions within the entry-point function must support code generation.

  For details, see “General Code Generation Workflow” on page 32-5.
- Code generation workflow for the object function of a machine learning model (including `predict`, `random`, `knnsearch`, `rangesearch`, and incremental learning object functions)



Save a trained model by using `saveLearnerForCoder`, and define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the object function. Then generate code for the entry-point function by using `codegen`, and verify the generated code. The input arguments of the entry-point function cannot be classification or regression model objects. Therefore, you need to work around this limitation by using `saveLearnerForCoder` and `loadLearnerForCoder`.

You can also generate single-precision C/C++ code for the prediction of machine learning models for classification and regression. For single-precision code generation, specify the name-value pair argument `'Datatype','single'` as an additional input to the `loadLearnerForCoder` function.

For details, see these examples

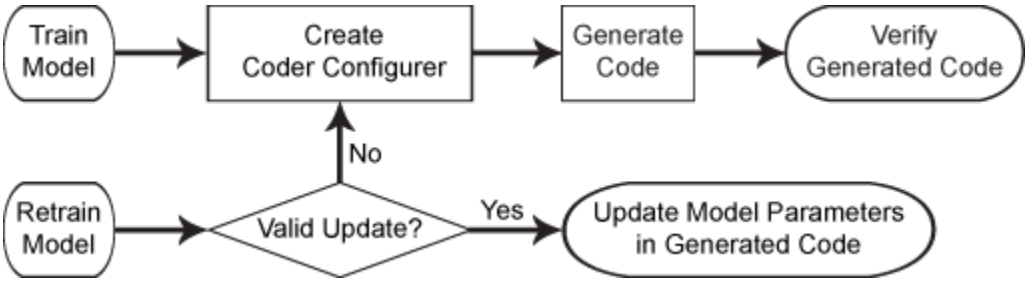
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Incremental Learning” on page 32-13
- “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
- “Code Generation for Nearest Neighbor Searcher” on page 32-19
- “Code Generation and Classification Learner App” on page 32-31
- “Specify Variable-Size Arguments for Code Generation” on page 32-45

You can also generate fixed-point C/C++ code for the prediction of a support vector machine (SVM) model, a decision tree model, and an ensemble of decision trees for classification and regression. This type of code generation requires Fixed-Point Designer™.

Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. You can also optimize the fixed-point data types before generating code.

For details, see “Fixed-Point Code Generation for Prediction of SVM” on page 32-87.

- Code generation workflow for the `predict` and `update` functions of a tree model, an SVM model, a linear model, or a multiclass error-correcting output codes (ECOC) classification model using SVM or linear binary learners



Create a coder configurer by using `learnerCoderConfigurer`, generate code by using `generateCode`, and then verify the generated code. You can configure code generation options and specify the coder attributes of the model parameters using object properties. After you retrain the model with new data or settings, you can update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code.

For details, see “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80.

Code Generation Applications

To integrate the prediction of a machine learning model into Simulink®, use a MATLAB Function block or the Simulink blocks in the Statistics and Machine Learning Toolbox library. For details, see these examples:

- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “Predict Responses Using RegressionSVM Predict Block” on page 32-115
- “Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111

Code generation for the Statistics and Machine Learning Toolbox functions also works with other toolboxes such as System object™ and Stateflow®, as described in these examples:

- “System Objects for Classification and Code Generation” on page 32-54
- “Predict Class Labels Using Stateflow” on page 32-62

For more applications of code generation, see these examples:

- “Code Generation for Image Classification” on page 32-103
- “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66

See Also

`codegen` | `generateLearnerDataTypeFcn` | `learnerCoderConfigurer` | `loadLearnerForCoder` | `saveLearnerForCoder`

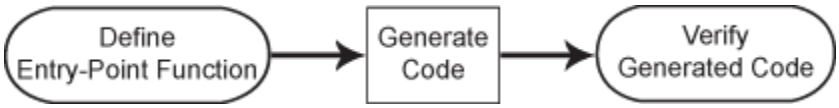
Related Examples

- “Get Started with MATLAB Coder” (MATLAB Coder)

General Code Generation Workflow

The general code generation workflow for the Statistics and Machine Learning Toolbox functions that are not the object functions of machine learning models is the same as the workflow described in MATLAB Coder. For details, see “Get Started with MATLAB Coder” (MATLAB Coder). To learn how to generate code for the object functions of machine learning models, see “Introduction to Code Generation” on page 32-2.

This example briefly explains the general code generation workflow as summarized in this flow chart:



Define Entry-Point Function

An *entry-point* function, also known as the *top-level* or *primary* function, is a function you define for code generation. Because you cannot call any function at the top level using `codegen`, you must define an entry-point function that calls code-generation-enabled functions, and generate C/C++ code for the entry-point function by using `codegen`. All functions within the entry-point function must support code generation.

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would cause errors during code generation. See “Check Code with the Code Analyzer” (MATLAB Coder).

For example, to generate code that estimates the interquartile range of a data set using `iqr`, define this function.

```
function r = iqrCodeGen(x) %#codegen
%IQRCODEGEN Estimate interquartile range
%   iqrCodeGen returns the interquartile range of the data x,
%   a single- or double-precision vector.
r = iqr(x);
end
```

You can allow for optional input arguments by specifying `varargin` as an input argument. For details, see “Code Generation for Variable Length Argument Lists” (MATLAB Coder) and “Specify Variable-Size Arguments for Code Generation” on page 32-45.

Generate Code

Set Up Compiler

To generate C/C++ code, you must have access to a compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. To view and change the default C compiler, enter:

```
mex -setup
```

For more details, see “Change Default Compiler”.

Generate Code Using codegen

After setting up your compiler, generate code for the entry-point function by using `codegen` or the MATLAB Coder app. To learn how to generate code using the MATLAB Coder app, see “Generate MEX Functions by Using the MATLAB Coder App” (MATLAB Coder).

To generate code at the command line, use `codegen`. Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Specify the data types and sizes of all inputs of the entry-point function when you call `codegen` by using the `-args` option.

- To specify the data type and exact input array size, pass a MATLAB expression that represents the set of values with a certain data type and array size. For example, to specify that the generated code from `iqrCodeGen.m` must accept a double-precision numeric column vector with 100 elements, enter:

  ```
  testX = randn(100,1);
  codegen iqrCodeGen -args {testX} -report
  ```

  The `-report` flag generates a code generation report. See “Code Generation Reports” (MATLAB Coder).
- To specify that at least one of the dimensions can have any length, use the `-args` option with `coder.typeof` as follows.

  ```
  -args {coder.typeof(example_value, size_vector, variable_dims)}
  ```

  The values of *`example_value`*, *`size_vector`*, and *`variable_dims`* specify the properties of the input array that the generated code can accept.

  - An input array has the same data type as the example values in *`example_value`*.
  - *`size_vector`* is the array size of an input array if the corresponding *`variable_dims`* value is `false`.
  - *`size_vector`* is the upper bound of the array size if the corresponding *`variable_dims`* value is `true`.
  - *`variable_dims`* specifies whether each dimension of the array has a variable size or a fixed size. A value of `true` (logical 1) means that the corresponding dimension has a variable size; a value of `false` (logical 0) means that the corresponding dimension has a fixed size.

  Specifying a variable-size input is convenient when you have data with an unknown number of observations at compile time. For example, to specify that the generated code from `iqrCodeGen.m` can accept a double-precision numeric column vector of any length, enter:

  ```
  testX = coder.typeof(0,[Inf,1],[1,0]);
  codegen iqrCodeGen -args {testX} -report
  ```

  `0` for the *`example_value`* value implies that the data type is `double` because `double` is the default numeric data type of MATLAB. `[Inf,1]` for the *`size_vector`* value and `[1,0]` for the *`variable_dims`* value imply that the size of the first dimension is variable and unbounded, and the size of the second dimension is fixed to be 1.

Note Specification of variable size inputs can affect performance. For details, see “Control Memory Allocation for Variable-Size Arrays” (MATLAB Coder).

- To specify a character array, such as supported name-value pair arguments, specify the character array as a constant using `coder.Constant`. For example, suppose that `'Name'` is a valid name-value pair argument for `iqrCodeGen.m`, and the corresponding value `value` is numeric. Then enter:

  ```
  codegen iqrCodeGen -args {testX,coder.Constant('Name'),value} -report
  ```

For more details, see “Generate C Code at the Command Line” (MATLAB Coder) and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

Build Type

MATLAB Coder can generate code for these types:

- MEX (MATLAB Executable) function
- Standalone C/C++ code
- Standalone C/C++ code compiled to a static library
- Standalone C/C++ code compiled to a dynamically linked library
- Standalone C/C++ code compiled to an executable

You can specify the build type using the `-config` option of `codegen`. For more details on setting code generation options, see “Configure Build Settings” (MATLAB Coder).

By default, `codegen` generates a MEX function. A MEX function is a C/C++ program that is executable from MATLAB. You can use a MEX function to accelerate MATLAB algorithms and to test the generated code for functionality and run-time issues. For details, see “MATLAB Algorithm Acceleration” (MATLAB Coder) and “Why Test MEX Functions in MATLAB?” (MATLAB Coder).

Code Generation Report

You can use the `-report` flag to produce a code generation report. This report helps you debug code generation issues and view the generated C/C++ code. For details, see “Code Generation Reports” (MATLAB Coder).

Verify Generated Code

Test a MEX function to verify that the generated code provides the same functionality as the original MATLAB code. To perform this test, run the MEX function using the same inputs that you used to run the original MATLAB code, and then compare the results. Running the MEX function in MATLAB before generating standalone code also enables you to detect and fix run-time errors that are much harder to diagnose in the generated standalone code. For more details, see “Why Test MEX Functions in MATLAB?” (MATLAB Coder).

Pass some data to verify whether `iqr`, `iqrCodeGen`, and `iqrCodeGen_mex` return the same interquartile range.

```
testX = randn(100,1);
r = iqr(testX);
r_entrypoint = iqrCodeGen(testX);
r_mex = iqrCodeGen_mex(testX);
```

Compare the outputs by using `isequal`.

```
isequal(r,r_entrypoint,r_mex)
```

`isequal` returns logical 1 (true) if all the inputs are equal.

You can also verify the MEX function using a test file and `coder.runTest`. For details, see “Testing Code Generated from MATLAB Code” (MATLAB Coder).

See Also

`codegen`

More About

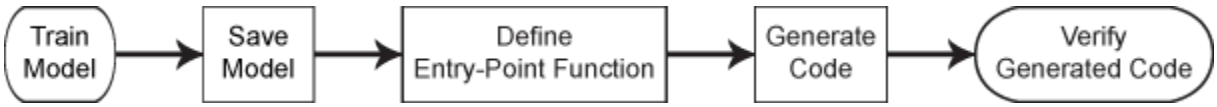
- “Introduction to Code Generation” on page 32-2
- “Code Generation for Probability Distribution Objects” on page 32-82
- Function List (C/C++ Code Generation)

Code Generation for Prediction of Machine Learning Model at Command Line

This example shows how to generate code for the prediction of classification and regression model objects at the command line. You can also generate code using the MATLAB® Coder™ app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details.

Certain classification and regression model objects have a `predict` or `random` function that supports code generation. Prediction using these object functions requires a trained classification or regression model object, but the `-args` option of `codegen` (MATLAB Coder) does not accept these objects. Work around this limitation by using `saveLearnerForCoder` and `loadLearnerForCoder` as described in this example.

This flow chart shows the code generation workflow for the object functions of classification and regression model objects.



After you train a model, save the trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the object function. Then generate code for the entry-point function by using `codegen`, and verify the generated code.

Train Classification Model

Train a classification model object equipped with a code-generation-enabled `predict` function. In this case, train a support vector machine (SVM) classification model.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,3:4);
Y = species(inds);
Mdl = fitcsvm(X,Y);
```

This step can include data preprocessing, feature selection, and optimizing the model using cross-validation, for example.

Save Model Using `saveLearnerForCoder`

Save the classification model to the file `SVMModel.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'SVMModel');
```

`saveLearnerForCoder` saves the classification model to the MATLAB binary file `SVMModel.mat` as a structure array in the current folder.

Define Entry-Point Function

An *entry-point* function, also known as the *top-level* or *primary* function, is a function you define for code generation. Because you cannot call any function at the top level using `codegen`, you must define an entry-point function that calls code-generation-enabled functions, and generate C/C++ code

for the entry-point function by using `codegen`. All functions within the entry-point function must support code generation.

Define an entry-point function that returns predicted labels for input predictor data. Within the function, load the trained classification model by using `loadLearnerForCoder`, and then pass the loaded model to `predict`. In this case, define the `predictLabelsSVM` function, which predicts labels using the SVM model `Mdl`.

```
type predictLabelsSVM.m % Display contents of predictLabelsSVM.m file
```

```
function label = predictLabelsSVM(x) %#codegen
%PREDICTLABELSSVM Label new observations using trained SVM model Mdl
%   predictLabelsSVM predicts the vector of labels label using
%   the saved SVM model Mdl and the predictor data x.
Mdl = loadLearnerForCoder('SVMModel');
label = predict(Mdl,x);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation. See “Check Code with the Code Analyzer” (MATLAB Coder).

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Set Up Compiler

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate Code Using `codegen`

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Specify the data types and sizes of all inputs of the entry-point function when you call `codegen` by using the `-args` option.

In this case, pass `X` as a value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`.

```
codegen predictLabelsSVM -args {X}
```

```
Code generation successful.
```

If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder)

Build Type

MATLAB Coder can generate code for the following build types:

- MEX (MATLAB Executable) function
- Standalone C/C++ code
- Standalone C/C++ code compiled to a static library
- Standalone C/C++ code compiled to a dynamically linked library
- Standalone C/C++ code compiled to an executable

You can specify the build type using the `-config` option of `codegen` (MATLAB Coder). For more details on setting code generation options, see the `-config` option of `codegen` (MATLAB Coder) and "Configure Build Settings" (MATLAB Coder).

By default, `codegen` generates a MEX function. A MEX function is a C/C++ program that is executable from MATLAB. You can use a MEX function to accelerate MATLAB algorithms and to test the generated code for functionality and run-time issues. For details, see "MATLAB Algorithm Acceleration" (MATLAB Coder) and "Why Test MEX Functions in MATLAB?" (MATLAB Coder).

Code Generation Report

You can use the `-report` flag to produce a code generation report. This report helps you debug code generation issues and view the generated C/C++ code. For details, see "Code Generation Reports" (MATLAB Coder).

Verify Generated Code

Test a MEX function to verify that the generated code provides the same functionality as the original MATLAB code. To perform this test, run the MEX function using the same inputs that you used to run the original MATLAB code, and then compare the results. Running the MEX function in MATLAB before generating standalone code also enables you to detect and fix run-time errors that are much harder to diagnose in the generated standalone code. For more details, see "Why Test MEX Functions in MATLAB?" (MATLAB Coder).

Pass some predictor data to verify whether `predict`, `predictLabelsSVM`, and the MEX function return the same labels.

```
labels1 = predict(Mdl,X);
labels2 = predictLabelsSVM(X);
labels3 = predictLabelsSVM_mex(X);
```

Compare the predicted labels by using `isequal`.

```
verifyMEX = isequal(labels1,labels2,labels3)
```

```
verifyMEX = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal. The comparison confirms that the `predict` function, `predictLabelsSVM` function, and MEX function return the same labels.

See Also

`codegen` | `learnerCoderConfigurer` | `loadLearnerForCoder` | `saveLearnerForCoder`

Related Examples

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
- “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80
- “Code Generation and Classification Learner App” on page 32-31
- “Specify Variable-Size Arguments for Code Generation” on page 32-45
- Function List (C/C++ Code Generation)

Code Generation for Incremental Learning

This example shows how to generate code that implements incremental learning for binary linear classification. To motivate its purpose, consider training a wearable device tasked to determine whether the wearer is idle or moving, based on sensory features the device reads.

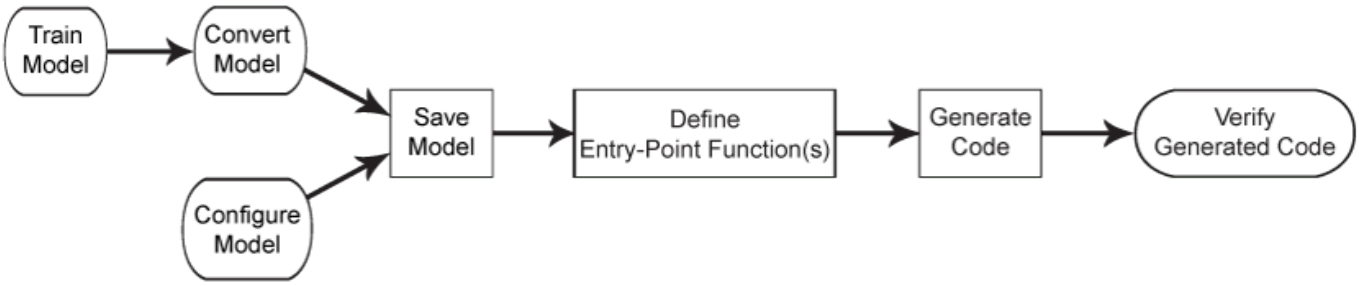
The generated code performs the following tasks, as defined in entry-point functions:

1. Load a configured incremental learning model template created at the command line.
2. Track performance metrics on the incoming batch of data from a data stream. This example tracks misclassification rate and hinge loss.
3. Update the model by fitting the incremental model to the batch of data.
4. Predicts labels for the batch of data.

This example generates code from the MATLAB® command line, but you can generate code using the MATLAB® Coder™ app instead. For more details, see “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22.

All incremental learning object functions for binary linear classification (and also linear regression) support code generation. To prepare code to generate for incremental learning, the object functions require an appropriately configured incremental learning model object, but the `-args` option of `codegen` (MATLAB Coder) does not accept these objects. To work around this limitation, use the `saveLearnerForCoder` and `loadLearnerForCoder` functions.

This flow chart shows the code generation workflows for the incremental learning object functions for linear models.



The flow chart suggests two distinct but merging workflows.

- The workflow beginning with **Train Model > Convert Model** requires data, in which case you can optionally perform feature selection or optimize the model by performing cross-validation before generating code for incremental learning.
- The workflow beginning with **Configure Model** does not require data. Instead, you must manually configure an incremental learning model object.

For details on the differences between the workflows, and for help deciding which one to use, see “Configure Incremental Learning Model” on page 26-8.

Regardless of the workflow you choose, the resulting incremental learning model must have all the following qualities:

- The `NumPredictors` property reflect the number of predictors in the predictor data during incremental learning.

- For classification, the `ClassNames` property must contain all class names expected during incremental learning.

If you choose the **Train Model > Convert Model** workflow and you fit the model to data containing all known classes, the model is configured for code generation.

After you prepare an incremental learning model, save the model object by using `saveLearnerForCoder`. Then, define an entry-point function that loads the saved model by using `loadLearnerForCoder`, and that performs incremental learning by calling the object functions. Alternatively, you can define multiple entry-point functions that perform the stages of incremental learning separately (this example uses this workflow). However, this workflow requires special treatment when an updated model object is an input to another entry-point function. For example, you write the following three entry-point functions:

- A function that accepts the current model and a batch of data, calls `updateMetrics`, and returns a model with updated performance metrics.
- A function that accepts the updated model and the batch of data, calls `fit`, and returns a model with updated coefficients.
- A function that accepts the further updated model and the batch of predictor data, calls `predict`, and returns predicted labels.

Finally, generate code for the entry-point functions by using `codegen`, and verify the generated code.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
p = size(feat,2);
idx = randsample(n,n);
X = feat(idx,:);
actid = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is idle (`actid` <= 2). Store the unique class names. Create categorical arrays.

```
classnames = categorical(["Idle" "NotIdle"]);
Y = repmat(classnames(1),n,1);
Y(actid > 2) = classnames(2);
```

Configure Incremental Learning Model

To generate code for incremental classification, you must appropriately configure a binary classification linear model for incremental learning `incrementalClassificationLinear`.

Create a binary classification (SVM) model for incremental learning. Fully configure the model for code generation by specify all expected class names and the number of predictor variables. Also, specify tracking the misclassification rate and hinge loss. For reproducibility, this example turns off observation shuffling for the scale-invariant solver.

```
metrics = ["classiferror" "hinge"];
IncrementalMdl = incrementalClassificationLinear('ClassNames',classnames,'NumPredictors',p,...
    'Shuffle',false,'Metrics',metrics)

IncrementalMdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [2x2 table]
        ClassNames: [Idle    NotIdle]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: 0
           Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incremenalClassificationLinear` model object configured for code generation. `Mdl` is *cold* (`Mdl.IsWarm` is `0`) because it has not processed data—the coefficients are `0`.

Alternatively, because the data is available, you can fit an SVM model to the data by using either `fitcsvm` or `fitclinear`, and then convert the resulting model to an incremental learning model by passing the model to `incrementalLearner`. The resulting model is *warm* because it has processed data—its coefficients are likely non-zero.

Save Model Using saveLearnerForCoder

Save the incremental learning model to the file `InitialMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(IncrementalMdl,'InitialMdl');
```

`saveLearnerForCoder` saves the incremental learning model to the MATLAB binary file `SVMClassIncrLearner.mat` as structure arrays in the current folder.

Define Entry-Point Functions

An *entry-point* function, also known as the *top-level* or *primary* function, is a function you define for code generation. Because you cannot call any function at the top level using `codegen`, you must define an entry-point function that calls code-generation-enabled functions, and generate C/C++ code for the entry-point function by using `codegen`. All functions within the entry-point function must support code generation.

Define four separate entry-point functions in your current folder that perform the following actions:

- `myInitialModelIncrLearn.m` — Load the saved model by using `loadLearnerForCoder`, and return a model of the same form for code generation. This entry-point function facilitates the use of a model, returned by an entry-point function, as an input to another entry-point function.
- `myUpdateMetricsIncrLearn.m` — Measure the performance of the current model on an incoming batch of data, and store the performance metrics in the model. The function accepts the current model, and predictor and response data, and returns an updated model.
- `myFitIncrLearn.m` — Fit the current model to the incoming batch of data, and store the updated coefficients in the model. The function accepts the current model, and predictor and response data, and returns an updated model.

- `myPredictIncrLearn.m` — Predicted labels for the incoming batch of data using the current model. The function accepts the current model and predictor data, and returns labels and class scores.

For more details on generating code for multiple entry-point functions, see “Generate Code for Multiple Entry-Point Functions” (MATLAB Coder).

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation. See “Check Code with the Code Analyzer” (MATLAB Coder).

Alternatively, you can access the functions in `mlr/examples/stats/main`, where `mlr` is the value of `matlabroot`.

Display the body of each function.

```
type myInitialModelIncrLearn.m
```

```
function incrementalModel = myInitialModelIncrLearn() %#codegen
% MYINITIALMODELINCRLEARN Load and return configured linear model for
% binary classification InitialMdl
    incrementalModel = loadLearnerForCoder('InitialMdl');
end
```

```
type myUpdateMetricsIncrLearn.m
```

```
function incrementalModel = myUpdateMetricsIncrLearn(incrementalModel,X,Y) %#codegen
% MYUPDATEMETRICSINCRLEARN Measure model performance metrics on new data
      incrementalModel = updateMetrics(incrementalModel,X,Y);
end
```

```
type myFitIncrLearn.m
```

```
function incrementalModel = myFitIncrLearn(incrementalModel,X,Y) %#codegen
% MYFITINCRLEARN Fit model to new data
      incrementalModel = fit(incrementalModel,X,Y);
end
```

```
type myPredictIncrLearn.m
```

```
function [labels,scores] = myPredictIncrLearn(incrementalModel,X) %#codegen
% MYPREDICTINCRLEARN Predict labels and classification scores on new data
      [labels,scores] = predict(incrementalModel,X);
end
```

Generate Code

Set Up Compiler

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Build Type

MATLAB Coder can generate code for the following build types:

- MEX (MATLAB Executable) function
- Standalone C/C++ code
- Standalone C/C++ code compiled to a static library
- Standalone C/C++ code compiled to a dynamically linked library
- Standalone C/C++ code compiled to an executable

You can specify the build type using the `-config` option of `codegen` (MATLAB Coder). For more details on setting code generation options, see the `-config` option of `codegen` (MATLAB Coder) and "Configure Build Settings" (MATLAB Coder).

By default, `codegen` generates a MEX function. A MEX function is a C/C++ program that is executable from MATLAB. You can use a MEX function to accelerate MATLAB algorithms and to test the generated code for functionality and run-time issues. For details, see "MATLAB Algorithm Acceleration" (MATLAB Coder) and "Why Test MEX Functions in MATLAB?" (MATLAB Coder).

Generate Code Using codegen

Because C and C++ are statically typed languages, you must specify the properties of all variables in the entry-point function at compile time. Specify all of the following:

- The data types of the data inputs of the entry-point functions by using `coder.typeof` (MATLAB Coder). Also, because the number of observations can vary from batch to batch, specify that the number of observations (first dimension) has variable size. For details, see "Specify Variable-Size Arguments for Code Generation" on page 32-45 and "Specify Properties of Entry-Point Function Inputs" (MATLAB Coder).
- Because several entry-point functions accept an incremental model object as a input, and operate on it, create a representation of the model object for code generation by using `coder.OutputType` (MATLAB Coder). For more details, see "Pass an Entry-Point Function Output as an Input" (MATLAB Coder).

```
predictorData = coder.typeof(X,[],[true false]);
responseData = coder.typeof(Y,[],true);
IncrMdlOutputType = coder.OutputType('myInitialModelIncrLearn');
```

Generate code for the entry-point functions using `codegen` (MATLAB Coder). For each entry-point function argument, use the `-args` flags to specify the coder representations of the variables. Specify the output MEX function name `myIncrLearn_mex`.

```
codegen -o myIncrLearn_mex ...
myInitialModelIncrLearn ...
myUpdateMetricsIncrLearn -args {IncrMdlOutputType,predictorData,responseData} ...
myFitIncrLearn -args {IncrMdlOutputType,predictorData,responseData} ...
myPredictIncrLearn –args {IncrMdlOutputType,predictorData} -report

Code generation successful: To view the report, open('codegen\mex\myIncrLearn_mex\html\report.ml
```

For help debugging code generation issues, view the generated C/C++ code by clicking `View report` (see "Code Generation Reports" (MATLAB Coder)).

Verify Generated Code

Test the MEX function to verify that the generated code provides the same functionality as the original MATLAB code. To perform this test, run the MEX function using the same inputs that you used to run the original MATLAB code, and then compare the results. Running the MEX function in

MATLAB before generating standalone code also enables you to detect and fix run-time errors that are much harder to diagnose in the generated standalone code. For more details, see “Why Test MEX Functions in MATLAB?” (MATLAB Coder).

Perform incremental learning by using the generated MEX functions and directly by using the object functions. Specify a batch

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
hinge = ce;
ceCG = ce;
hingeCG = ce;
IncrementalMdlCG = myIncrLearn_mex('myInitialModelIncrLearn');
scores = zeros(n,2);
scoresCG = zeros(n,2);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;

    IncrementalMdl = updateMetrics(IncrementalMdl,X(idx,:),Y(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    hinge{j,:} = IncrementalMdl.Metrics{"HingeLoss",:};
    IncrementalMdlCG = myIncrLearn_mex('myUpdateMetricsIncrLearn',IncrementalMdlCG,...
        X(idx,:),Y(idx));
    ceCG{j,:} = IncrementalMdlCG.Metrics{"ClassificationError",:};
    hingeCG{j,:} = IncrementalMdlCG.Metrics{"HingeLoss",:};

    IncrementalMdl = fit(IncrementalMdl,X(idx,:),Y(idx));
    IncrementalMdlCG = myIncrLearn_mex('myFitIncrLearn',IncrementalMdlCG,X(idx,:),Y(idx));

    [~,scores(idx,:)] = predict(IncrementalMdl,X(idx,:));
    [~,scoresCG(idx,:)] = myIncrLearn_mex('myPredictIncrLearn',IncrementalMdlCG,X(idx,:));
end
```

Compare the cumulative metrics and scores for classifying `Idle` returned by the object functions and MEX functions.

```
idx = all(~isnan(ce.Variables),2);
areCEsEqual = norm(ce.Cumulative(idx) - ceCG.Cumulative(idx))

areCEsEqual = 8.9904e-18

idx = all(~isnan(hinge.Variables),2);
areHingeLossesEqual = norm(hinge.Cumulative(idx) - hingeCG.Cumulative(idx))

areHingeLossesEqual = 9.5220e-17

areScoresEqual = norm(scores(:,1) - scoresCG(:,1))

areScoresEqual = 8.7996e-13
```

The differences between the returned quantities are negligible.

Code Generation for Nearest Neighbor Searcher

The object functions `knnsearch` and `rangesearch` of the nearest neighbor searcher objects, `ExhaustiveSearcher` and `KDTreeSearcher`, support code generation. This example shows how to generate code for finding the nearest neighbor using an exhaustive searcher object at the command line. The example shows two different ways to generate code, depending on the way you use the object: load the object by using `loadLearnerForCoder` in an entry-point function, and pass a compile-time constant object to the generated code.

Train Exhaustive Nearest Neighbor Searcher

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng('default');              % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Prepare an exhaustive nearest neighbor searcher using the training data. Specify the `'Distance'` and `'P'` name-value pair arguments to use the Minkowski distance with an exponent of 1 for finding the nearest neighbor.

```
Mdl = ExhaustiveSearcher(X,'Distance','minkowski','P',1);
```

Find the index of the training data (`X`) that is the nearest neighbor of each point in the query data (`Y`).

```
Idx = knnsearch(Mdl,Y);
```

Generate Code Using `saveLearnerForCoder` and `loadLearnerForCoder`

Generate code that loads an exhaustive searcher, takes query data as an input argument, and then finds the nearest neighbor.

Save the exhaustive searcher to a file using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'searcherModel')
```

`saveLearnerForCoder` saves the model to the MATLAB binary file `searcherModel.mat` as a structure array in the current folder.

Define the entry-point function `myknnsearch1` that takes query data as an input argument. Within the function, load the searcher object by using `loadLearnerForCoder`, and then pass the loaded model to `knnsearch`.

```
type myknnsearch1.m % Display contents of myknnsearch1.m file
```

```
function idx = myknnsearch1(Y) %#codegen
Mdl = loadLearnerForCoder('searcherModel');
idx = knnsearch(Mdl,Y);
end
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function files, `myknnsearch1.m`, `myknnsearch2.m`, and `myknnsearch3.m`.

Generate code for `myknnsearch1` by using `codegen` (MATLAB Coder). Specify the data type and dimension of the input argument by using `coder.typeof` (MATLAB Coder) so that the generated code accepts a variable-size array.

```
codegen myknnsearch1 -args {coder.typeof(Y,[Inf,4],[1,0])}
```

```
Code generation successful.
```

For a more detailed code generation example that uses `saveLearnerForCoder` and `loadLearnerForCoder`, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. For more details about specifying variable-size arguments, see “Specify Variable-Size Arguments for Code Generation” on page 32-45.

Pass the query data (`Y`) to verify that `myknnsearch1` and the MEX file return the same indices.

```
myIdx1 = myknnsearch1(Y);
myIdx1_mex = myknnsearch1_mex(Y);
```

Compare `myIdx1` and `myIdx1_mex` by using `isequal`.

```
verifyMEX1 = isequal(Idx,myIdx1,myIdx1_mex)
```

```
verifyMEX1 = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. This comparison confirms that `myknnsearch1` and the MEX file return the same results.

Generate Code with Constant Folded Model Object

Nearest neighbor searcher objects can be an input argument of a function you define for code generation. The `-args` option of `codegen` (MATLAB Coder) accept a compile-time constant searcher object.

Define the entry-point function `myknnsearch2` that takes both an exhaustive searcher model and query data as input arguments instead of loading the model in the function.

```
type myknnsearch2.m % Display contents of myknnsearch2.m file
```

```
function idx = myknnsearch2(Mdl,Y) %#codegen
idx = knnsearch(Mdl,Y);
end
```

To generate code that takes the model object as well as the query data, designate the model object as a compile-time constant by using `coder.Constant` (MATLAB Coder) and include the constant folded model object in the `-args` value of `codegen`.

```
codegen myknnsearch2 -args {coder.Constant(Mdl),coder.typeof(Y,[Inf,4],[1,0])}
```

```
Code generation successful.
```

The code generation workflow with a constant folded model object follows general code generation workflow. For details, see “General Code Generation Workflow” on page 32-5.

Verify that `myknnsearch2` and the MEX file return the same results.

```
myIdx2 = myknnsearch2(Mdl,Y);
myIdx2_mex = myknnsearch2_mex(Mdl,Y);
verifyMEX2 = isequal(Idx,myIdx2,myIdx2_mex)

verifyMEX2 = logical
   1
```

Generate Code with Name-Value Pair Arguments

Define the entry-point function `myknnsearch3` that takes a model object, query data, and name-value pair arguments. You can allow for optional name-value arguments by specifying varargin as an input argument. For details, see “Code Generation for Variable Length Argument Lists” (MATLAB Coder).

```
type myknnsearch3.m % Display contents of myknnsearch3.m file

function idx = myknnsearch3(Mdl,Y,varargin) %#codegen
idx = knnsearch(Mdl,Y,varargin{:});
end
```

To generate code that allows a user-defined exponent for the Minkowski distance, include `{coder.Constant('P'),0}` in the `-args` value of `codegen`. Use `coder.Constant` (MATLAB Coder) because the name of a name-value pair argument must be a compile-time constant.

```
codegen myknnsearch3 -args {coder.Constant(Mdl),coder.typeof(Y,[Inf,4],[1,0]),coder.Constant('P'

Code generation successful.
```

Verify that `myknnsearch3` and the MEX file return the same results.

```
newIdx = knnsearch(Mdl,Y,'P',2);
myIdx3 = myknnsearch3(Mdl,Y,'P',2);
myIdx3_mex = myknnsearch3_mex(Mdl,Y,'P',2);
verifyMEX3 = isequal(newIdx,myIdx3,myIdx3_mex)

verifyMEX3 = logical
   1
```

See Also

`ExhaustiveSearcher` | `KDTreeSearcher` | `codegen` | `knnsearch` | `loadLearnerForCoder` | `rangesearch` | `saveLearnerForCoder`

Related Examples

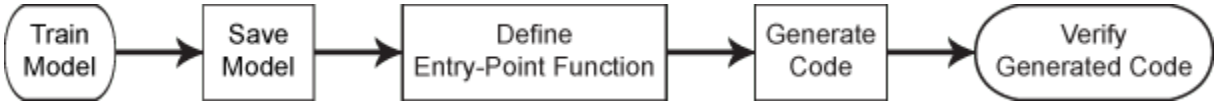
- “Introduction to Code Generation” on page 32-2
- “General Code Generation Workflow” on page 32-5
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Specify Variable-Size Arguments for Code Generation” on page 32-45

Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App

This example shows how to generate C/C++ code for the prediction of classification and regression model objects by using the MATLAB® Coder™ app. You can also generate code at the command line using `codegen` (MATLAB Coder). See “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9 for details.

Certain classification and regression model objects have a `predict` or `random` function that supports code generation. Prediction using these object functions requires a trained classification or regression model object, but an entry-point function for code generation cannot have these objects as input variables. Work around this limitation by using `saveLearnerForCoder` and `loadLearnerForCoder` as described in this example.

This flow chart shows the code generation workflow for the object functions of classification and regression model objects.



In this example, you train a classification ensemble model using *k*-nearest-neighbor weak learners and save the trained model by using `saveLearnerForCoder`. Then, define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the object function. Write a script to test the entry-point function. Finally, generate code by using the MATLAB Coder app and verify the generated code.

Train Classification Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a classification ensemble model with *k*-nearest-neighbor weak learners by using the random subspace method. For details of classifications that use a random subspace ensemble, see “Random Subspace Classification” on page 18-104.

```
rng('default')  % For reproducibility
learner = templateKNN('NumNeighbors',2);
Mdl = fitcensemble(X,Y,'Method','Subspace','NPredToSample',5, ...
    'Learners',learner,'NumLearningCycles',13);
```

Save Model Using `saveLearnerForCoder`

Save the trained ensemble model to a file named `knnEnsemble.mat` in your current folder.

```
saveLearnerForCoder(Mdl,'knnEnsemble')
```

`saveLearnerForCoder` makes the full classification model `Mdl` compact, and then saves it to the MATLAB binary file `knnEnsemble.mat` as a structure array in the current folder.

Define Entry-Point Function

An *entry-point* function, also known as the *top-level* or *primary* function, is a function you define for code generation. You must define an entry-point function that calls code-generation-enabled functions

and generate C/C++ code from the entry-point function. All functions within the entry-point function must support code generation.

In a new file in your current folder, define an entry-point function named `myknnEnsemblePredict` that does the following:

- Accept input data (`X`), the file name of the saved model (`fileName`), and valid name-value pair arguments of the `predict` function (`varargin`).
- Load a trained ensemble model by using `loadLearnerForCoder.`
- Predict labels and corresponding scores from the loaded model.

You can allow for optional name-value arguments by specifying varargin as an input argument. For details, see “Code Generation for Variable Length Argument Lists” (MATLAB Coder).

```
type myknnEnsemblePredict.m % Display the contents of myknnEnsemblePredict.m file.
```

```
function [label,score] = myknnEnsemblePredict(X,fileName,varargin) %#codegen
CompactMdl = loadLearnerForCoder(fileName);
[label,score] = predict(CompactMdl,X,varargin{:});
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation. See “Check Code with the Code Analyzer” (MATLAB Coder).

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB, then MATLAB opens the example folder. This folder includes the entry-point function file (`myknnEnsemblePredict.m`) and the test file (`test_myknnEnsemblePredict.m`, described later on).

Set Up Compiler

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Create Test File

Write a test script that calls the `myknnEnsemblePredict` function. In the test script, specify the input arguments and name-value pair arguments that you use in the generated code. You use this test script to define input types automatically when generating code using the MATLAB Coder app.

In this example, create the `test_myknnEnsemblePredict.m` file in your current folder, as shown.

```
type test_myknnEnsemblePredict.m % Display the contents of test_myknnEnsemblePredict.m file.
```

```
%% Load Sample data
load ionosphere

%% Test myknnEnsemblePredict
[label,score] = myknnEnsemblePredict(X,'knnEnsemble','Learners',1:13);
```

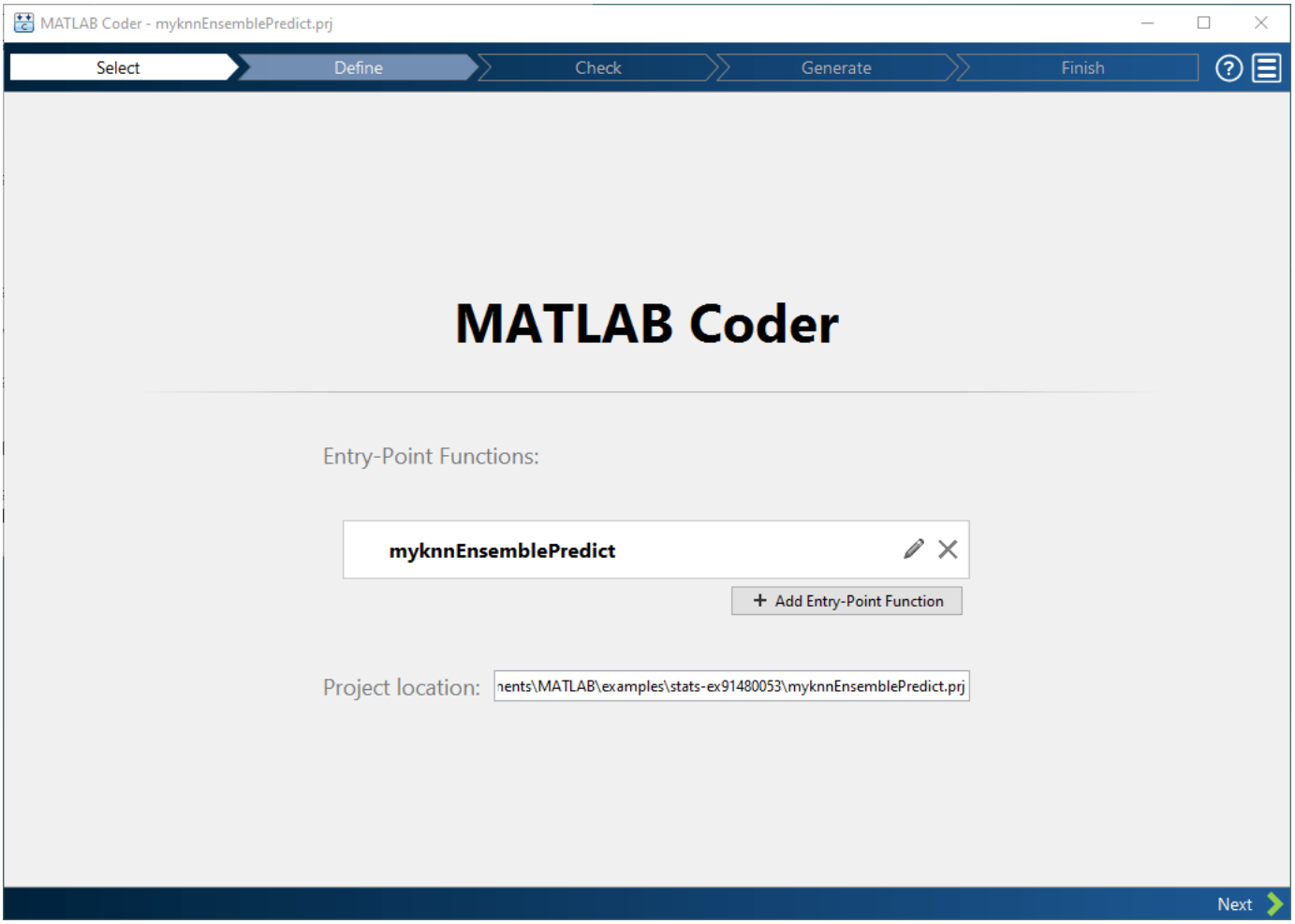
For details, see “Automatically Define Input Types by Using the App” (MATLAB Coder).

Generate Code Using MATLAB Coder App

The MATLAB Coder app generates C or C++ code from MATLAB code. The workflow-based user interface steps you through the code generation process. The following steps describe a brief workflow of the MATLAB Coder App. For more details, see MATLAB Coder (MATLAB Coder) and “Generate C Code by Using the MATLAB Coder App” (MATLAB Coder).

1. Open the MATLAB Coder App and Select the Entry-Point Function File.

On the **Apps** tab, in the **Apps** section, click the **Show more** arrow to open the apps gallery. Under **Code Generation**, click **MATLAB Coder**. The app opens the **Select Source Files** page. Enter or select the name of the entry-point function, `myknnEnsemblePredict`.

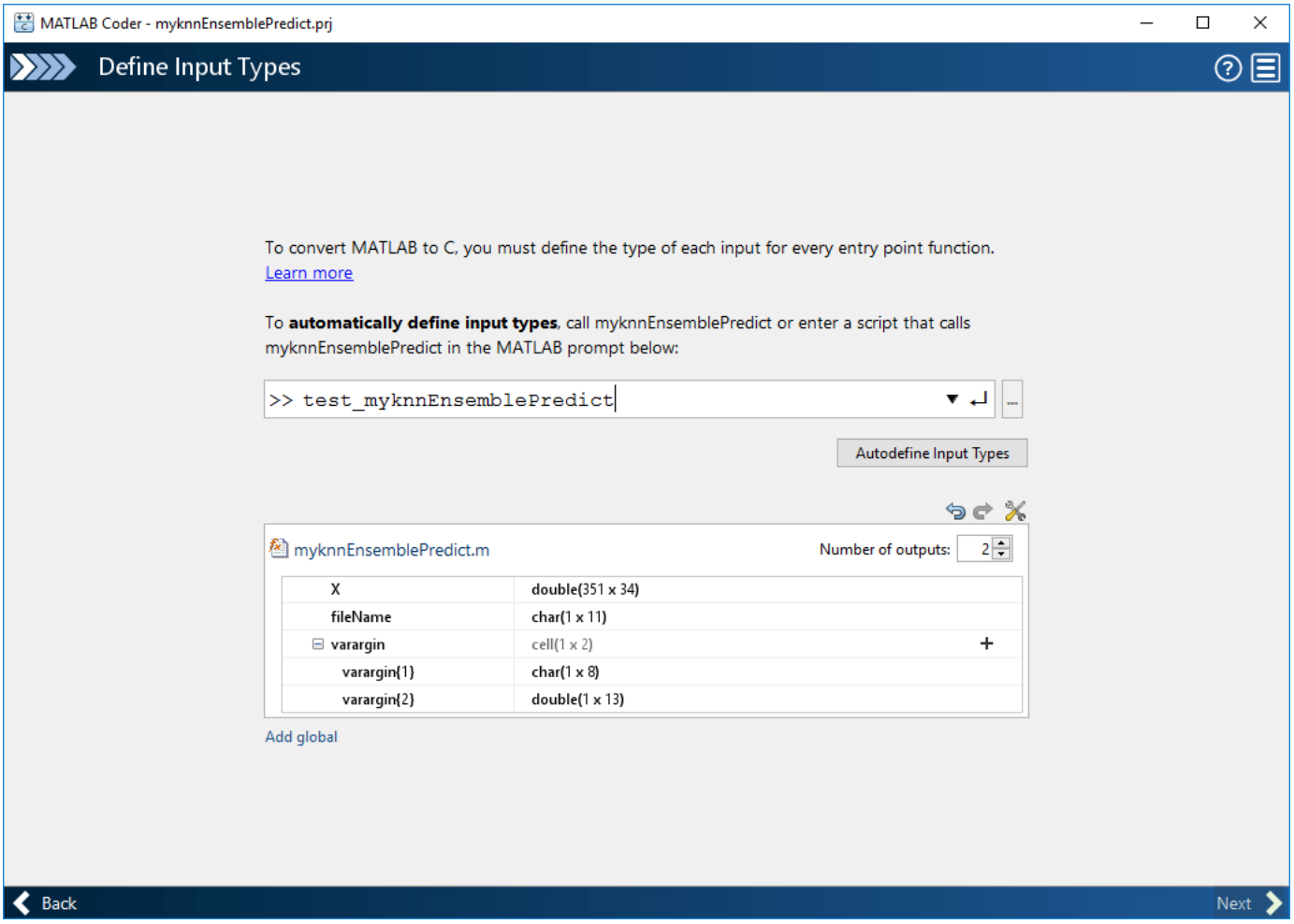


Click **Next** to go to the **Define Input Types** page.

2. Define Input Types

Because C uses static typing, MATLAB Coder must determine the properties of all variables in the MATLAB files at compile time. Therefore, you need to specify the properties of the entry-point function inputs.

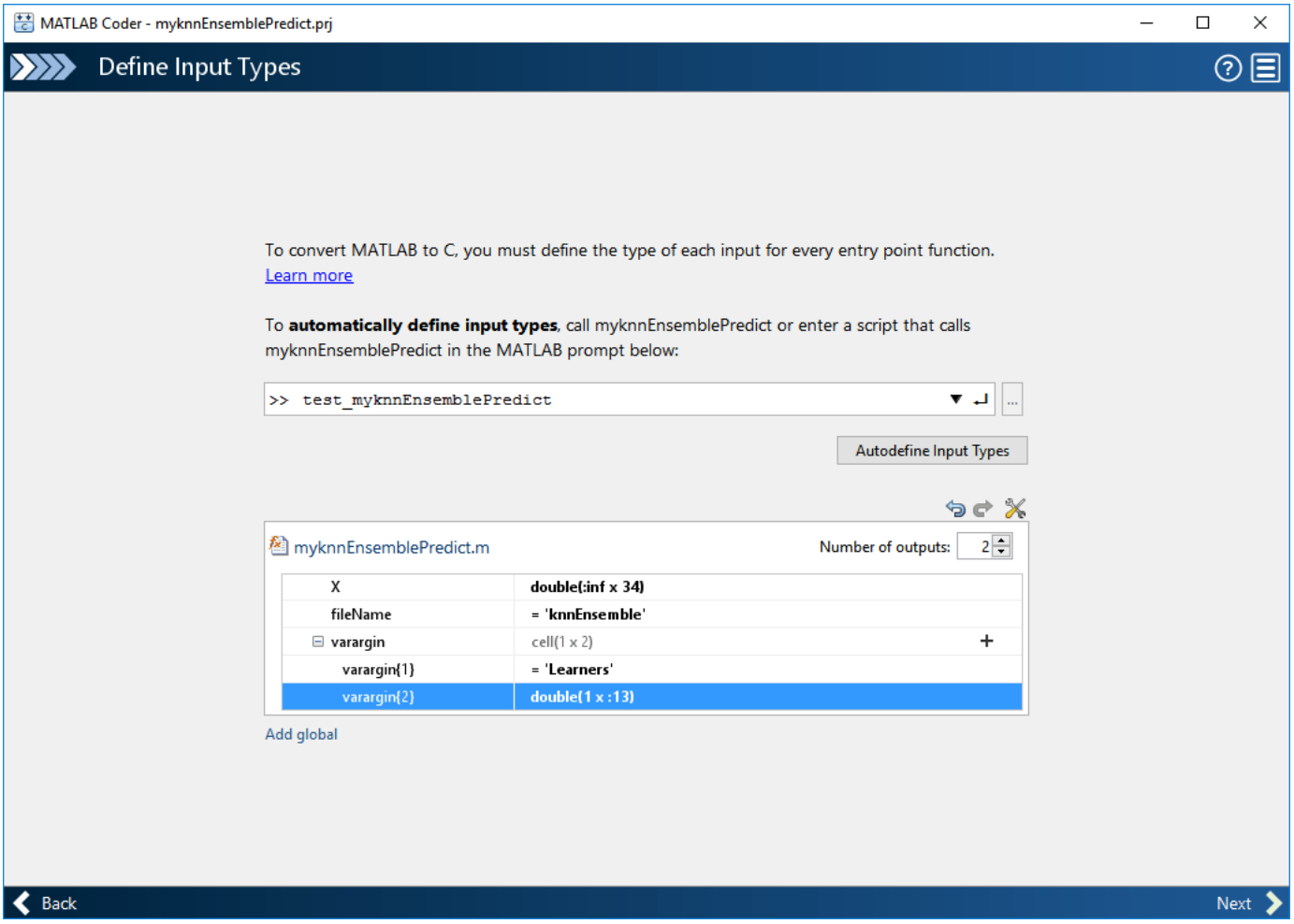
Enter or select the test script `test_myknnEnsemblePredict` and click **Autodefine Input Types**.



The MATLAB Coder app recognizes input types of the `myknnEnsemblePredict` function based on the test script.

Modify the input types:

- `X` — The app infers that input `X` is `double(351x34)`. The number of predictors must be fixed to be the same as the number of predictors in the trained model. However, you can have a different number of observations for prediction. If the number of observations is unknown, change `double(351x34)` to `double(:351x34)` or `double(:infx34)`. The setting `double(:351x34)` allows the number of observations up to 351, and the setting `double(:infx34)` allows an unbounded number of observations. In this example, specify `double(:infx34)` by clicking `351` and selecting `:inf`.
- `fileName` — Click **`char`**, select **Define Constant**, and type the file name with single quotes, `'knnEnsemble'`.
- `varargin{1}` — Names in name-value pair arguments must be compile-time constants. Click **`char`**, select **Define Constant**, and type `'Learners'`.
- `varargin{2}` — To allow user-defined indices up to 13 weak learners in the generated code, change `double(1x13)` to `double(1x:13)`.

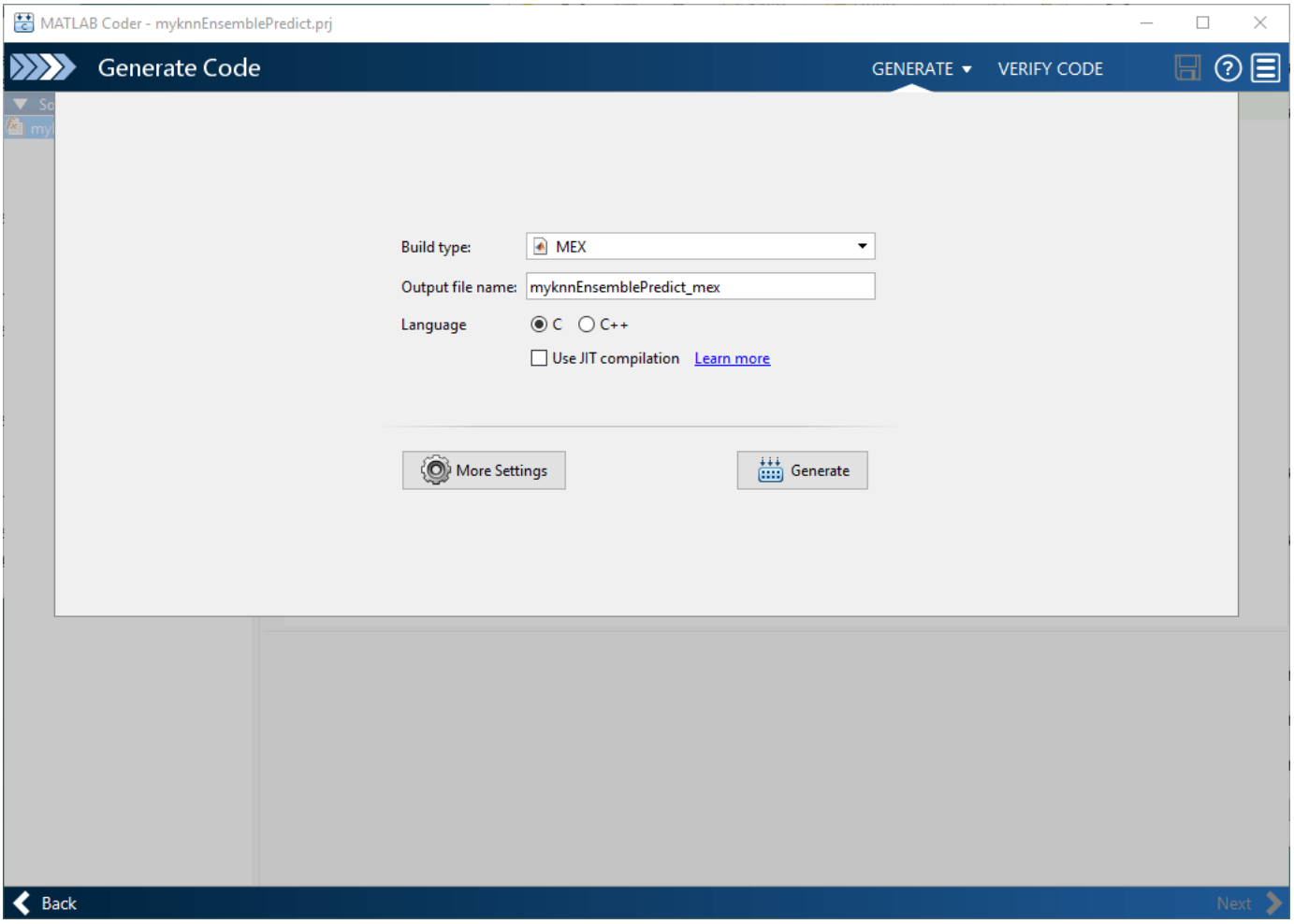


Click **Next** to go to the **Check for Run-Time Issues** page. This optional step generates a MEX file, runs the MEX function, and reports issues. Click **Next** to go to the **Generate Code** page.

3. Generate C Code

Set **Build type** to MEX and click **Generate**. The app generates a MEX function, `myknnEnsemblePredict_mex`. A MEX function is a C/C++ program that is executable from MATLAB. You can use a MEX function to accelerate MATLAB algorithms and to test the generated code for functionality and run-time issues. For details, see “MATLAB Algorithm Acceleration” (MATLAB Coder) and “Why Test MEX Functions in MATLAB?” (MATLAB Coder).

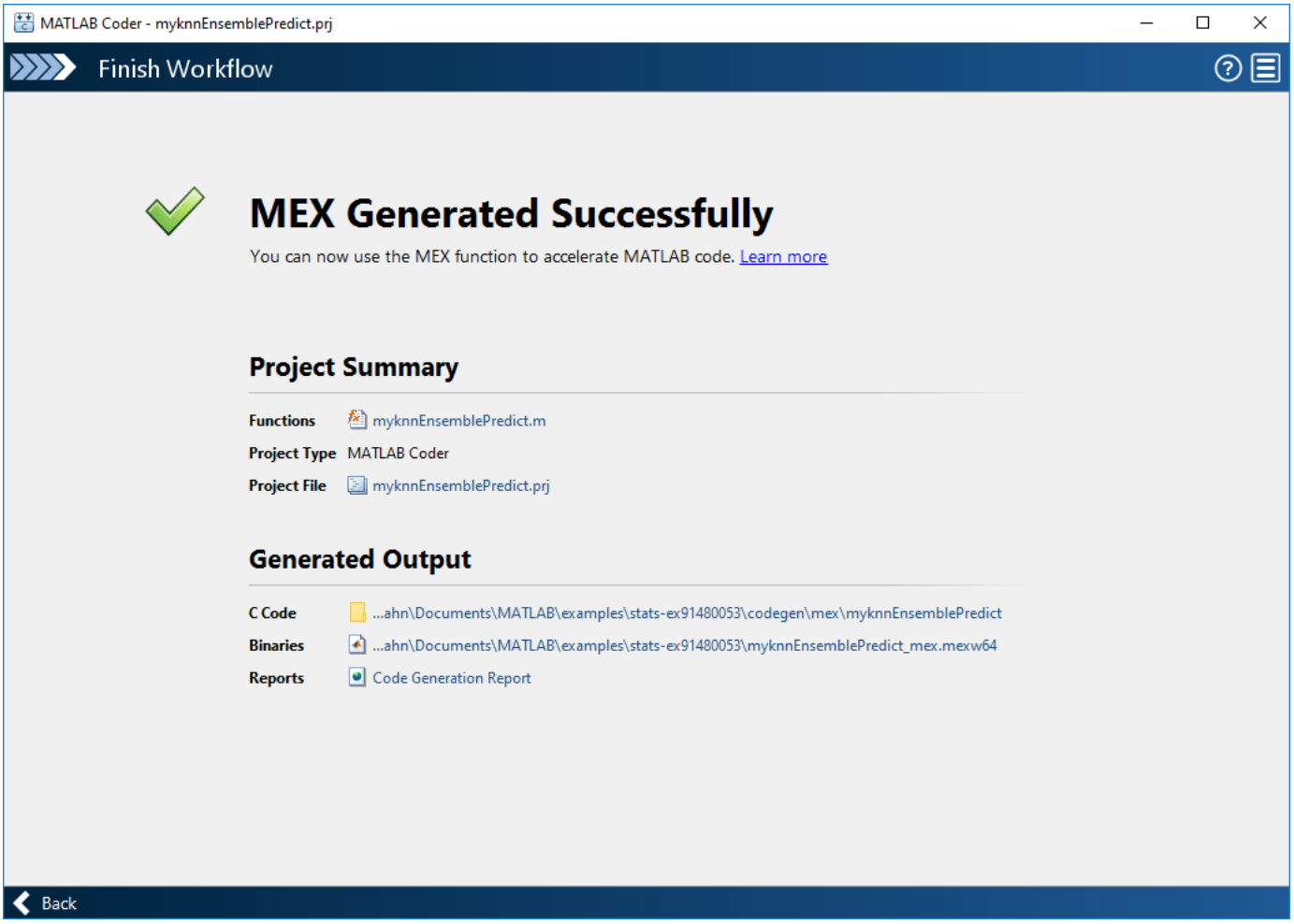
Depending on the specified build type, MATLAB Coder generates a MEX function or standalone C/C++ code compiled to a static library, dynamic linked library, or executable. For details on setting a build type, see “Configure Build Settings” (MATLAB Coder).



Click **Next** to go to the **Finish Workflow** page.

4. Review the Finish Workflow Page

The **Finish Workflow** page indicates that code generation succeeded. This page also provides a project summary and links to generated output.



Generate Code Using Script

You can convert a MATLAB Coder project to the equivalent script of MATLAB commands after you define input types. Then you run the script to generate code. For details, see “Convert MATLAB Coder Project to MATLAB Script” (MATLAB Coder).

On the MATLAB Coder app toolbar, click the **Open action menu** button: 

Select **Convert to script**, and then click **Save**. The app creates the file `myknnEnsemblePredict_script.m`, which reproduces the project in a configuration object and runs the `codegen` (MATLAB Coder) function.

Display the contents of the file `myknnEnsemblePredict_script.m`.

```
type myknnEnsemblePredict_script.m
```

```
% MYKNNENSEMBLEPREDICT_SCRIPT   Generate MEX-function myknnEnsemblePredict_mex
%  from myknnEnsemblePredict.
%
% Script generated from project 'myknnEnsemblePredict.prj' on 17-Nov-2017.
%
```

```
% See also CODER, CODER.CONFIG, CODER.TYPEOF, CODEGEN.

%% Create configuration object of class 'coder.MexCodeConfig'.
cfg = coder.config('mex');
cfg.GenerateReport = true;
cfg.ReportPotentialDifferences = false;

%% Define argument types for entry-point 'myknnEnsemblePredict'.
ARGS = cell(1,1);
ARGS{1} = cell(4,1);
ARGS{1}{1} = coder.typeof(0,[Inf  34],[1 0]);
ARGS{1}{2} = coder.Constant('knnEnsemble');
ARGS{1}{3} = coder.Constant('Learners');
ARGS{1}{4} = coder.typeof(0,[1 13],[0 1]);

%% Invoke MATLAB Coder.
codegen -config cfg myknnEnsemblePredict -args ARGS{1} -nargout 2
```

Run the script.

```
myknnEnsemblePredict_script
```

```
Code generation successful: To view the report, open('codegen\mex\myknnEnsemblePredict\html\repo
```

Verify Generated Code

Test a MEX function to verify that the generated code provides the same functionality as the original MATLAB code. To perform this test, run the MEX function using the same inputs that you used to run the original MATLAB code, and then compare the results. Running the MEX function in MATLAB before generating standalone code also enables you to detect and fix run-time errors that are much harder to diagnose in the generated standalone code. For more details, see “Why Test MEX Functions in MATLAB?” (MATLAB Coder).

Pass some predictor data to verify that `myknnEnsemblePredict` and the MEX function return the same results.

```
[label1,score1] = predict(Mdl,X,'Learners',1:10);
[label2,score2] = myknnEnsemblePredict(X,'knnEnsemble','Learners',1:10);
[label3,score3] = myknnEnsemblePredict_mex(X,'knnEnsemble','Learners',1:10);
```

Compare `label1`, `label2`, and `label3` by using `isequal`.

```
isequal(label1,label2,label3)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal.

The `score3` output from the MEX function might include round-off differences compared with the output from the `predict` function. In this case, compare `score1` and `score3`, allowing a small tolerance.

```
find(abs(score1-score3) > 1e-12)
```

```
ans =

  0x1 empty double column vector
```

`find` returns an empty vector if the element-wise absolute difference between `score1` and `score3` is not larger than the specified tolerance `1e-12`. The comparisons confirm that `myknnEnsemblePredict` and the MEX function return the same results.

See Also

`codegen` | `learnerCoderConfigurer` | `loadLearnerForCoder` | `saveLearnerForCoder`

More About

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80
- “Code Generation and Classification Learner App” on page 32-31
- “Specify Variable-Size Arguments for Code Generation” on page 32-45
- “Generate C Code by Using the MATLAB Coder App” (MATLAB Coder)
- Function List (C/C++ Code Generation)

Code Generation and Classification Learner App

Classification Learner is well suited for choosing and training classification models interactively, but it does not generate C/C++ code that labels data based on a trained model. The **Generate Function** button in the **Export** section of the Classification Learner app generates MATLAB code for training a model but does not generate C/C++ code. This example shows how to generate C code from a function that predicts labels using an exported classification model. The example builds a model that predicts the credit rating of a business given various financial ratios, according to these steps:

1. Use the credit rating data set in the file `CreditRating_Historical.dat`, which is included with Statistics and Machine Learning Toolbox.
2. Reduce the data dimensionality using principal component analysis (PCA).
3. Train a set of models that support code generation for label prediction.
4. Export the model with the minimum 5-fold, cross-validated classification accuracy.
5. Generate C code from an entry-point function that transforms the new predictor data and then predicts corresponding labels using the exported model.

Load Sample Data

Load sample data and import the data into the Classification Learner app. Review the data using scatter plots and remove unnecessary predictors.

Use `readtable` to load the historical credit rating data set in the file `CreditRating_Historical.dat` into a table.

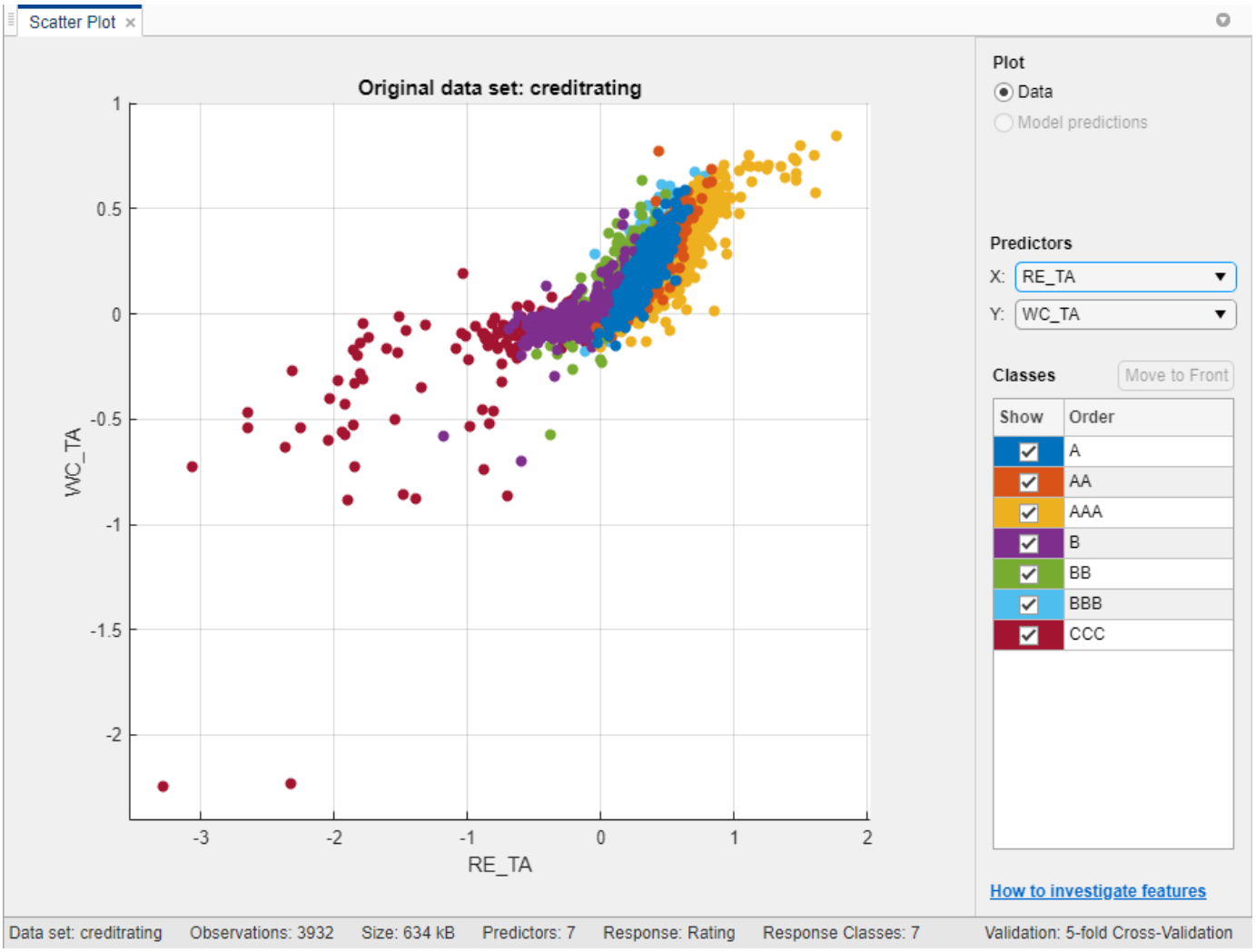
```
creditrating = readtable('CreditRating_Historical.dat');
```

On the **Apps** tab, click **Classification Learner**.

In Classification Learner, on the **Classification Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.

In the New Session from Workspace dialog box, select the table `creditrating`. All variables, except the one identified as the response, are double-precision numeric vectors. Click **Start Session** to compare classification models based on the 5-fold, cross-validated classification accuracy.

Classification Learner loads the data and plots a scatter plot of the variables `WC_TA` versus `ID`. Because identification numbers are not helpful to display in a plot, choose `RE_TA` for **X** under **Predictors**.



The scatter plot suggests that the two variables can separate the classes `AAA`, `BBB`, `BB`, and `CCC` fairly well. However, the observations corresponding to the remaining classes are mixed into these classes.

Identification numbers are not helpful for prediction. Therefore, in the **Features** section, click **Feature Selection** and then clear the **ID** check box. You can also remove unnecessary predictors from the beginning by using the check boxes in the New Session from Workspace dialog box. This example shows how to remove unused predictors for code generation when you have included all predictors.

Enable PCA

Enable PCA to reduce the data dimensionality.

In the **Features** section, click **PCA**, and then select **Enable PCA**. This action applies PCA to the predictor data, and then transforms the data before training the models. Classification Learner uses only components that collectively explain 95% of the variability.

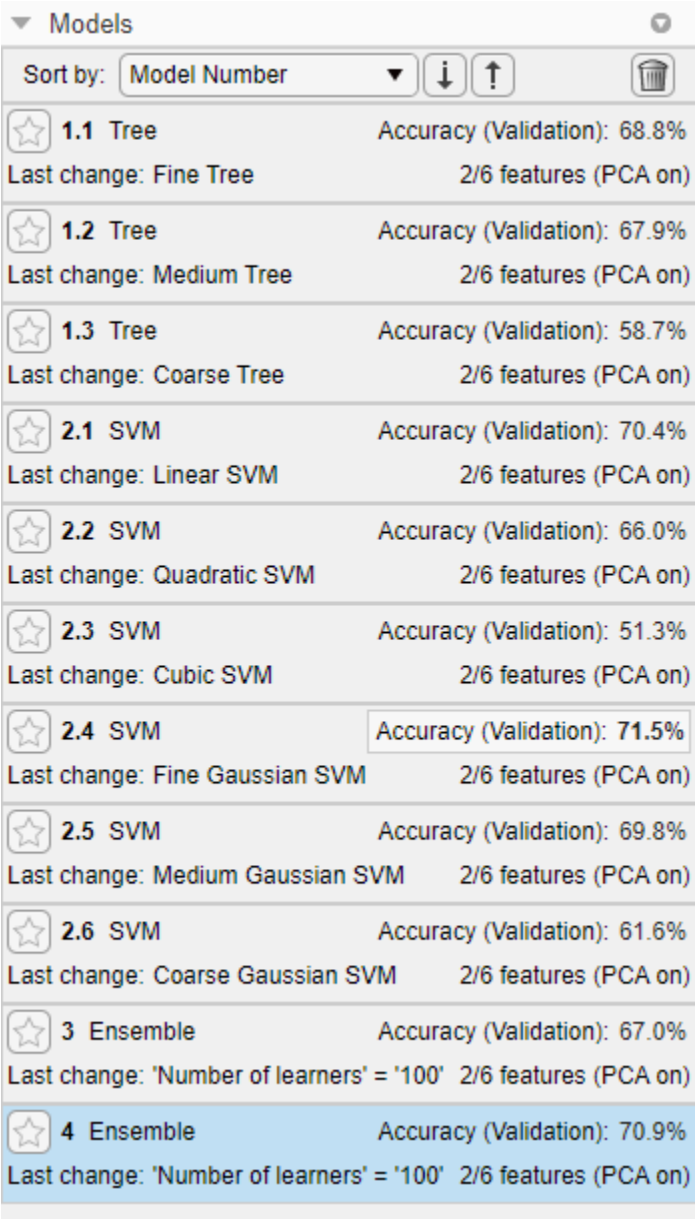
Train Models

Train a set of models that support code generation for label prediction. For a list of models in Classification Learner that support code generation, see “Generate C Code for Prediction” on page 23-79.

Select the following classification models and options, which support code generation for label prediction, and then perform cross-validation (for more details, see “Introduction to Code Generation” on page 32-2). To select each model, in the **Model Type** section, click the **Show more** arrow, and then click the model. After selecting a model and specifying any options, close any open menus, and then click **Train** in the **Training** section.

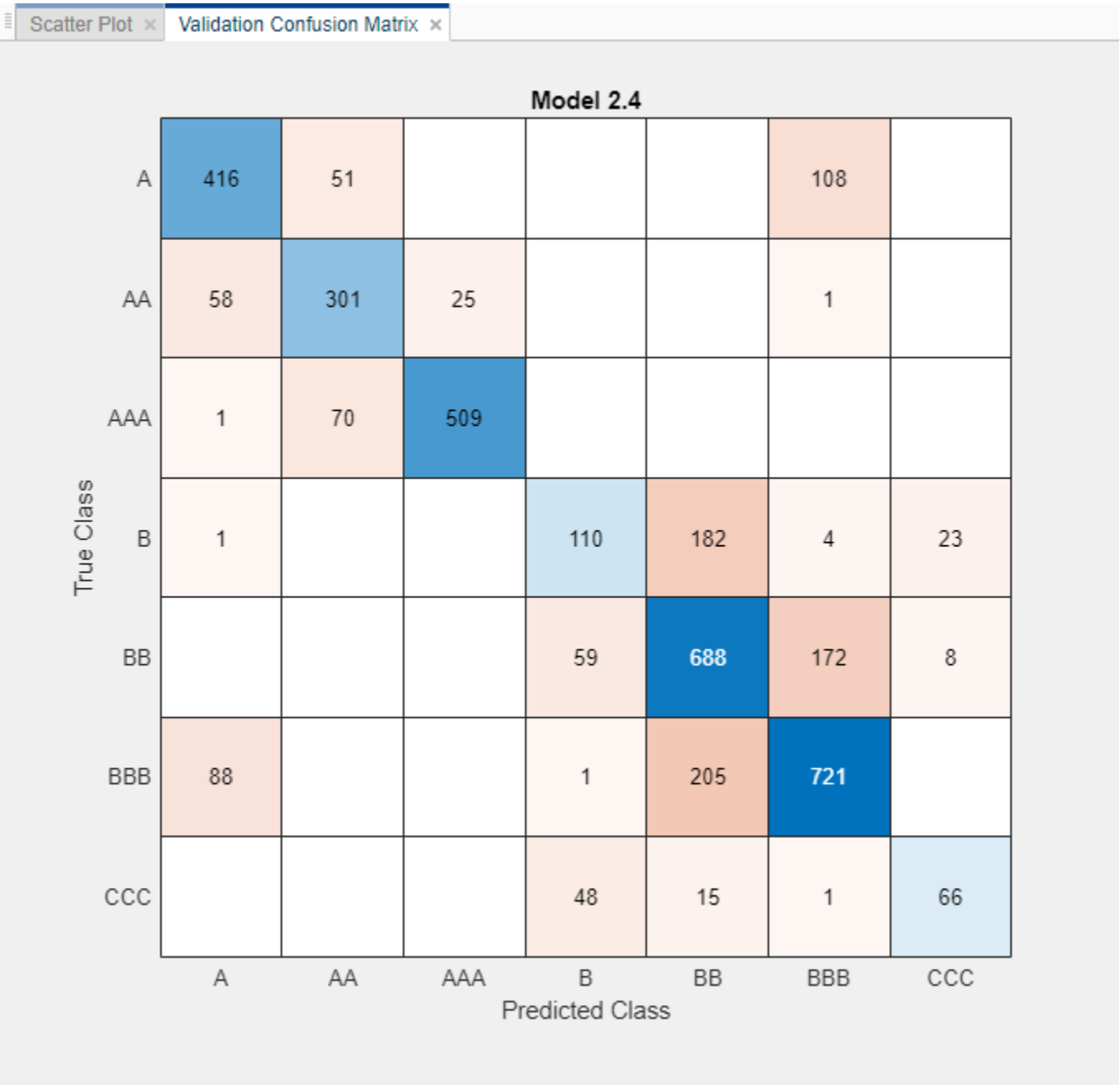
Models and Options to Select	Description
Under **Decision Trees**, select **All Trees**	Classification trees of various complexities
Under **Support Vector Machines**, select **All SVMs**	SVMs of various complexities and using various kernels. Complex SVMs require time to fit.
Under **Ensemble Classifiers**, select **Boosted Trees**. In the **Model Type** section, click **Advanced**. Reduce **Maximum number of splits** to **5** and increase **Number of learners** to **100**.	Boosted ensemble of classification trees
Under **Ensemble Classifiers**, select **Bagged Trees**. In the **Model Type** section, click **Advanced**. Increase **Maximum number of splits** to **50** and increase **Number of learners** to **100**.	Random forest of classification trees

After cross-validating each model type, the **Models** pane displays each model and its 5-fold, cross-validated classification accuracy, and highlights the model with the best accuracy.



Select the model that yields the maximum 5-fold, cross-validated classification accuracy, which is the error-correcting output codes (ECOC) model of Fine Gaussian SVM learners. With PCA enabled, Classification Learner uses two predictors out of six.

In the **Plots** section, click **Confusion Matrix** and select **Validation Data**.



The model does well distinguishing between A, B, and C classes. However, the model does not do as well distinguishing between particular levels within those groups, the lower B levels in particular.

Export Model to Workspace

Export the model to the MATLAB Workspace and save the model using `saveLearnerForCoder`.

In the **Export** section, click **Export Model**, and then select **Export Compact Model**. Click **OK** in the dialog box.

The structure `trainedModel` appears in the MATLAB Workspace. The field `ClassificationSVM` of `trainedModel` contains the compact model.

At the command line, save the compact model to a file called `ClassificationLearnerModel.mat` in your current folder.

```
saveLearnerForCoder(trainedModel.ClassificationSVM,'ClassificationLearnerModel')
```

Generate C Code for Prediction

Prediction using the object functions requires a trained model object, but the `-args` option of `codegen` does not accept such objects. Work around this limitation by using `saveLearnerForCoder` and `loadLearnerForCoder`. Save a trained model by using `saveLearnerForCoder`. Then, define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Finally, use `codegen` to generate code for the entry-point function.

Preprocess Data

Preprocess new data in the same way you preprocess the training data.

To preprocess, you need the following three model parameters:

- `removeVars` — Column vector of at most `p` elements identifying indices of variables to remove from the data, where `p` is the number of predictor variables in the raw data
- `pcaCenters` — Row vector of exactly `q` PCA centers
- `pcaCoefficients` — `q`-by-`r` matrix of PCA coefficients, where `r` is at most `q`

Specify the indices of predictor variables that you removed while selecting data using **Feature Selection** in Classification Learner. Extract the PCA statistics from `trainedModel`.

```
removeVars = 1;
pcaCenters = trainedModel.PCACenters;
pcaCoefficients = trainedModel.PCACoefficients;
```

Save the model parameters to a file named `ModelParameters.mat` in your current folder.

```
save('ModelParameters.mat','removeVars','pcaCenters','pcaCoefficients');
```

Define Entry-Point Function

An entry-point function is a function you define for code generation. Because you cannot call any function at the top level using `codegen`, you must define an entry-point function that calls code-generation-enabled functions, and then generate C/C++ code for the entry-point function by using `codegen`.

In your current folder, define a function named `mypredictCL.m` that:

- Accepts a numeric matrix (`X`) of raw observations containing the same predictor variables as the ones passed into Classification Learner
- Loads the classification model in `ClassificationLearnerModel.mat` and the model parameters in `ModelParameters.mat`
- Removes the predictor variables corresponding to the indices in `removeVars`
- Transforms the remaining predictor data using the PCA centers (`pcaCenters`) and coefficients (`pcaCoefficients`) estimated by Classification Learner

- Returns predicted labels using the model

```
function label = mypredictCL(X) %#codegen
%MYPREDICTCL Classify credit rating using model exported from
%Classification Learner
%   MYPREDICTCL loads trained classification model (SVM) and model
%   parameters (removeVars, pcaCenters, and pcaCoefficients), removes the
%   columns of the raw matrix of predictor data in X corresponding to the
%   indices in removeVars, transforms the resulting matrix using the PCA
%   centers in pcaCenters and PCA coefficients in pcaCoefficients, and then
%   uses the transformed data to classify credit ratings. X is a numeric
%   matrix with n rows and 7 columns. label is an n-by-1 cell array of
%   predicted labels.

% Load trained classification model and model parameters
SVM = loadLearnerForCoder('ClassificationLearnerModel');
data = coder.load('ModelParameters');
removeVars = data.removeVars;
pcaCenters = data.pcaCenters;
pcaCoefficients = data.pcaCoefficients;

% Remove unused predictor variables
keepvars = 1:size(X,2);
idx = ~ismember(keepvars,removeVars);
keepvars = keepvars(idx);
XwoID = X(:,keepvars);

% Transform predictors via PCA
Xpca = bsxfun(@minus,XwoID,pcaCenters)*pcaCoefficients;

% Generate label from SVM
label = predict(SVM,Xpca);
end
```

Generate Code

Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Specify variable-size arguments using `coder.typeof` and generate code using the arguments.

Create a double-precision matrix called `x` for code generation using `coder.typeof`. Specify that the number of rows of `x` is arbitrary, but that `x` must have `p` columns.

```
p = size(creditrating,2) - 1;
x = coder.typeof(0,[Inf,p],[1 0]);
```

For more details about specifying variable-size arguments, see “Specify Variable-Size Arguments for Code Generation” on page 32-45.

Generate a MEX function from `mypredictCL.m`. Use the `-args` option to specify `x` as an argument.

```
codegen mypredictCL -args x
```

`codegen` generates the MEX file `mypredictCL_mex.mexw64` in your current folder. The file extension depends on your platform.

Verify Generated Code

Verify that the MEX function returns the expected labels.

Remove the response variable from the original data set, and then randomly draw 15 observations.

```
rng('default'); % For reproducibility
m = 15;
testsampleT = datasample(creditrating(:,1:(end - 1)),m);
```

Predict corresponding labels by using `predictFcn` in the classification model trained by Classification Learner.

```
testLabels = trainedModel.predictFcn(testsampleT);
```

Convert the resulting table to a matrix.

```
testsample = table2array(testsampleT);
```

The columns of `testsample` correspond to the columns of the predictor data loaded by Classification Learner.

Pass the test data to `mypredictCL`. The function `mypredictCL` predicts corresponding labels by using `predict` and the classification model trained by Classification Learner.

```
testLabelsPredict = mypredictCL(testsample);
```

Predict corresponding labels by using the generated MEX function `mypredictCL_mex`.

```
testLabelsMEX = mypredictCL_mex(testsample);
```

Compare the sets of predictions.

```
isequal(testLabels,testLabelsMEX,testLabelsPredict)
```

```
ans =

  logical

   1
```

`isequal` returns logical 1 (true) if all the inputs are equal. `predictFcn`, `mypredictCL`, and the MEX function return the same values.

See Also

`codegen` | `coder.typeof` | `learnerCoderConfigurer` | `loadLearnerForCoder` | `saveLearnerForCoder`

Related Examples

- “Classification Learner App”
- “Predict Responses Using RegressionSVM Predict Block” on page 32-115

Predict Class Labels Using MATLAB Function Block

This example shows how to add a MATLAB® Function block to a Simulink® model for label prediction. The MATLAB Function block accepts streaming data, and predicts the label and classification score using a trained, support vector machine (SVM) classification model. For details on using the MATLAB Function block, see “Create Custom Functionality Using MATLAB Function Block” (Simulink).

Train Classification Model

This example uses the `ionosphere` data set, which contains radar-return qualities (`Y`) and predictor data (`X`). Radar returns are either of good quality (`'g'`) or of bad quality (`'b'`).

Load the `ionosphere` data set. Determine the sample size.

```
load ionosphere
n = numel(Y)
```

```
n = 351
```

The MATLAB Function block cannot return cell arrays. Convert the response variable to a logical vector whose elements are `1` if the radar returns are good, and `0` otherwise.

```
Y = strcmp(Y,'g');
```

Suppose that the radar returns are detected in sequence, and you have the first 300 observations, but you have not received the last 51 yet. Partition the data into present and future samples.

```
prsntX = X(1:300,:);
prsntY = Y(1:300);
ftrX = X(301:end,:);
ftrY = Y(301:end);
```

Train an SVM model using all, presently available data. Specify predictor data standardization.

```
Mdl = fitcsvm(prsntX,prsntY,'Standardize',true);
```

`Mdl` is a `ClassificationSVM` model.

Save Model Using `saveLearnerForCoder`

At the command line, you can use `Mdl` to make predictions for new observations. However, you cannot use `Mdl` as an input argument in a function meant for code generation.

Prepare `Mdl` to be loaded within the function using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'SVMIonosphere');
```

`saveLearnerForCoder` compacts `Mdl`, and then saves it in the MAT-file `SVMIonosphere.mat`.

Define MATLAB Function

Define a MATLAB function named `svmIonospherePredict.m` that predicts whether a radar return is of good quality. The function must:

- Include the code generation directive `%#codegen` somewhere in the function.

- Accept radar-return predictor data. The data must be commensurate with X except for the number of rows.
- Load `SVMIonosphere.mat` using `loadLearnerForCoder`.
- Return predicted labels and classification scores for predicting the quality of the radar return as good (that is, the positive-class score).

```
function [label,score] = svmIonospherePredict(X) %#codegen
%svmIonospherePredict Predict radar-return quality using SVM model
%   svmIonospherePredict predicts labels and estimates classification
%   scores of the radar returns in the numeric matrix of predictor data X
%   using the compact SVM model in the file SVMIonosphere.mat.  Rows of X
%   correspond to observations and columns to predictor variables.  label
%   is the predicted label and score is the confidence measure for
%   classifying the radar-return quality as good.
%
% Copyright 2016 The MathWorks Inc.
Mdl = loadLearnerForCoder('SVMIonosphere');
[label,bothscores] = predict(Mdl,X);
score = bothscores(:,2);
end
```

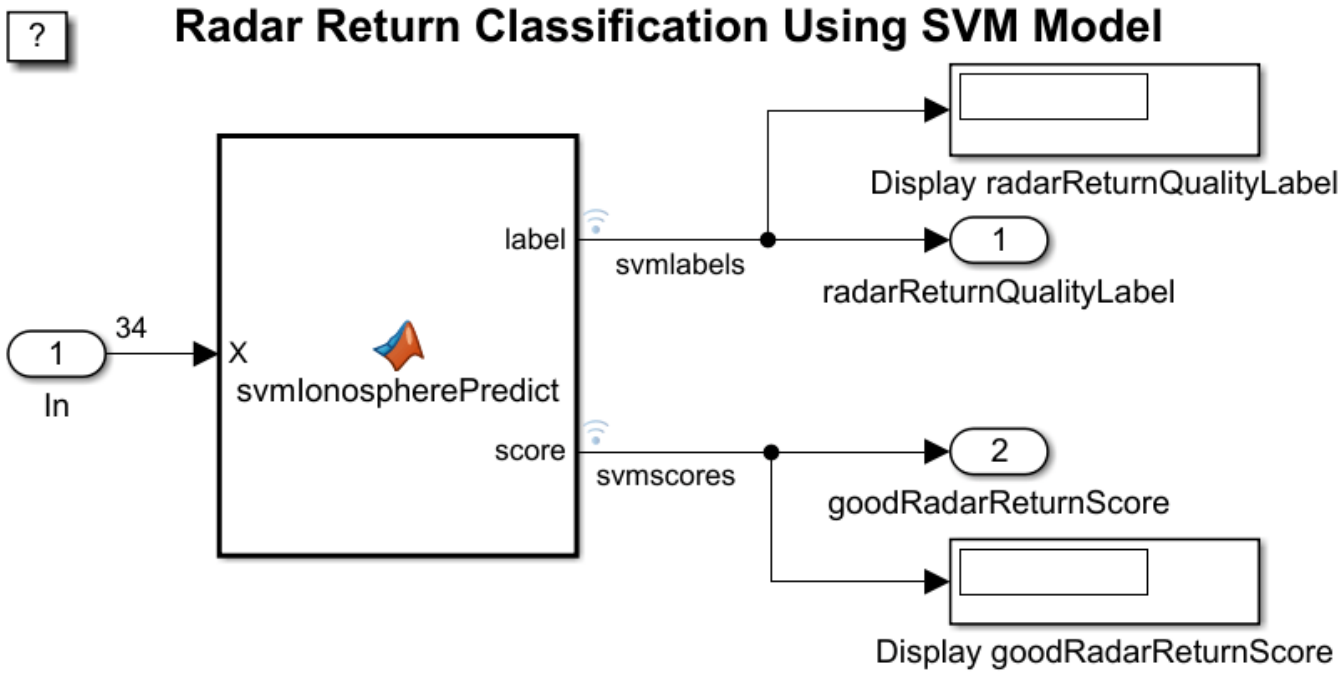
Note: If you click the button located in the upper-right section of this page and open this example in MATLAB, then MATLAB opens the example folder. This folder includes the entry-point function file.

Create Simulink Model

Create a Simulink model with the MATLAB Function block that dispatches to `svmIonospherePredict.m`.

This example provides the Simulink model `slexSVMIonospherePredictExample.slx`. Open the Simulink model.

```
SimMdlName = 'slexSVMIonospherePredictExample';
open_system(SimMdlName)
```



The figure displays the Simulink model. When the input node detects a radar return, it directs that observation into the MATLAB Function block that dispatches to `svmIonosphere Predict.m`. After predicting the label and score, the model returns these values to the workspace and displays the values within the model one at a time. When you load `slexSVMIonosphere PredictExample.slx`, MATLAB also loads the data set that it requires called `radarReturnInput`. However, this example shows how to construct the required data set.

The model expects to receive input data as a structure array called `radarReturnInput` containing these fields:

- `time` - The points in time at which the observations enter the model. In the example, the duration includes the integers from 0 though 50. The orientation must correspond to the observations in the predictor data. So, for this example, `time` must be a column vector.
- `signals` - A 1-by-1 structure array describing the input data, and containing the fields `values` and `dimensions`. `values` is a matrix of predictor data. `dimensions` is the number of predictor variables.

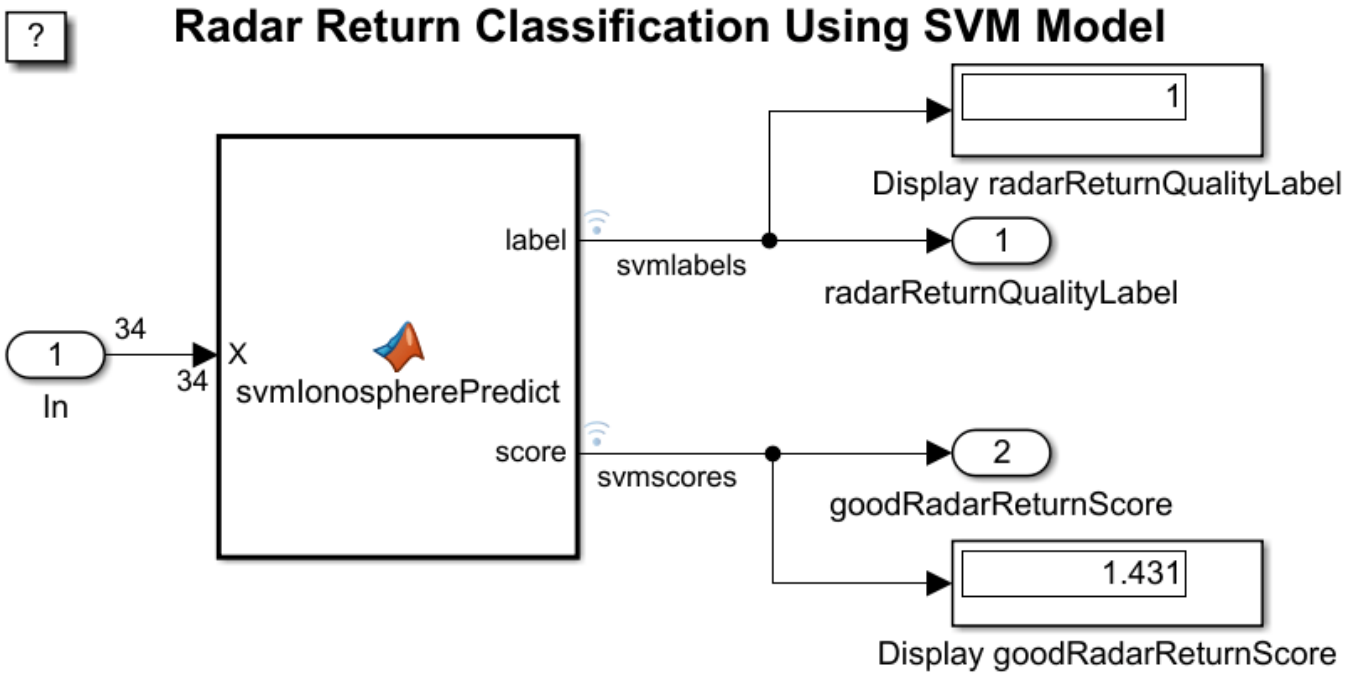
Create an appropriate structure array for future radar returns.

```
radarReturnInput.time = (0:50)';
radarReturnInput.signals(1).values = ftrX;
radarReturnInput.signals(1).dimensions = size(ftrX,2);
```

You can change the name from `radarReturnInput`, and then specify the new name in the model. However, Simulink expects the structure array to contain the described field names.

Simulate the model using the data held out of training, that is, the data in `radarReturnInput`.

```
sim(SimMdlName);
```



The figure shows the model after it processes all observations in `radarReturnInput` one at a time. The predicted label of `X(351,:)` is `1` and its positive-class score is `1.431`. The variables `tout`, `yout`, and `svmlogsout` appear in the workspace. `yout` and `svmlogsout` are `SimulinkData.Dataset` objects containing the predicted labels and scores. For more details, see "Data Format for Logged Simulation Data" (Simulink).

Extract the simulation data from the simulation log.

```
labelsSL = svmlogsout.getElement(1).Values.Data;
scoresSL = svmlogsout.getElement(2).Values.Data;
```

`labelsSL` is a 51-by-1 numeric vector of predicted labels. `labelsSL(j) = 1` means that the SVM model predicts that radar return `j` in the future sample is of good quality, and `0` means otherwise. `scoresSL` is a 51-by-1 numeric vector of positive-class scores, that is, signed distances from the decision boundary. Positive scores correspond to predicted labels of `1`, and negative scores correspond to predicted labels of `0`.

Predict labels and positive-class scores at the command line using `predict`.

```
[labelCMD,scoresCMD] = predict(Mdl,ftrX);
scoresCMD = scoresCMD(:,2);
```

`labelCMD` and `scoresCMD` are commensurate with `labelsSL` and `scoresSL`.

Compare the future-sample, positive-class scores returned by `slexSVMIonospherePredictExample` to those returned by calling `predict` at the command line.

```
err = sum((scoresCMD - scoresSL).^2);
err < eps
```

```
ans = logical
   1
```

The sum of squared deviations between the sets of scores is negligible.

If you also have a Simulink Coder™ license, then you can generate C code from `slexSVMIonospherePredictExample.slx` in Simulink or from the command line using `slbuild` (Simulink). For more details, see “Generate C Code for a Model” (Simulink Coder).

See Also

`learnerCoderConfigurer` | `loadLearnerForCoder` | `predict` | `saveLearnerForCoder` | `slbuild`

Related Examples

- “Predict Responses Using RegressionSVM Predict Block” on page 32-115
- “Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
- “Introduction to Code Generation” on page 32-2
- “Code Generation for Image Classification” on page 32-103
- “System Objects for Classification and Code Generation” on page 32-54
- “Predict Class Labels Using Stateflow” on page 32-62
- “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66

Specify Variable-Size Arguments for Code Generation

This example shows how to specify variable-size input arguments when you generate code for the object functions of classification and regression model objects. Variable-size data is data whose size might change at run time. Specifying variable-size input arguments is convenient when you have data with an unknown size at compile time. This example also describes how to include name-value pair arguments in an entry-point function and how to specify them when generating code.

For more detailed code generation workflow examples, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9 and “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22.

Train Classification Model

Load Fisher's iris data set. Convert the labels to a character matrix.

```
load fisheriris
species = char(species);
```

Train a classification tree using the entire data set.

```
Mdl = fitctree(meas,species);
```

`Mdl` is a `ClassificationTree` model.

Save Model Using `saveLearnerForCoder`

Save the trained classification tree to a file named `ClassTreeIris.mat` in your current folder by using `saveLearnerForCoder`.

```
MdlName = 'ClassTreeIris';
saveLearnerForCoder(Mdl,MdlName);
```

Define Entry-Point Function

In your current folder, define an entry-point function named `mypredictTree.m` that does the following:

- Accept measurements with columns corresponding to `meas` and accept valid name-value pair arguments.
- Load a trained classification tree by using `loadLearnerForCoder`.
- Predict labels and corresponding scores, node numbers, and class numbers from the loaded classification tree.

You can allow for optional name-value pair arguments by specifying varargin as an input argument. For details, see “Code Generation for Variable Length Argument Lists” (MATLAB Coder).

```
type mypredictTree.m  % Display contents of mypredictTree.m file
```

```
function [label,score,node,cnum] = mypredictTree(x,savedmdl,varargin) %#codegen
%MYPREDICTTREE Predict iris species using classification tree
%   MYPREDICTTREE predicts iris species for the n observations in the
%   n-by-4 matrix x using the classification tree stored in the MAT-file
%   whose name is in savedmdl, and then returns the predictions in the
%   array label. Each row of x contains the lengths and widths of the petal
%   and sepal of an iris (see the fisheriris data set). For other output
```

```
%   argument descriptions, see the predict reference page.
CompactMdl = loadLearnerForCoder(savedmdl);
[label,score,node,cnum] = predict(CompactMdl,x,varargin{:});
end
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Specify Variable-Size Arguments

Because C and C++ are statically typed languages, you must determine the properties of all variables in an entry-point function at compile time using the `-args` option of `codegen`.

Use `coder.Constant` (MATLAB Coder) to specify a compile-time constant input.

```
coder.Constant(v)
```

`coder.Constant(v)` creates a `coder.Constant` type variable whose values are constant, the same as `v`, during code generation.

Use `coder.typeof` (MATLAB Coder) to specify a variable-size input.

```
coder.typeof(example_value, size_vector, variable_dims)
```

The values of `example_value`, `size_vector`, and `variable_dims` specify the properties of the input array that the generated code can accept.

- An input array has the same data type as the example values in `example_value`.
- `size_vector` is the array size of an input array if the corresponding `variable_dims` value is `false`.
- `size_vector` is the upper bound of the array size if the corresponding `variable_dims` value is `true`.
- `variable_dims` specifies whether each dimension of the array has a variable size or a fixed size. A value of `true` (logical 1) means that the corresponding dimension has a variable size; a value of `false` (logical 0) means that the corresponding dimension has a fixed size.

The entry-point function `mypredictTree` accepts predictor data, the MAT-file name containing the trained model object, and optional name-value pair arguments. Suppose that you want to generate code that accepts a variable-size array for predictor data and the `'Subtrees'` name-value pair argument with a variable-size vector for its value. Then you have four input arguments: predictor data, the MAT-file name, and the name and value of the `'Subtrees'` name-value pair argument.

Define a 4-by-1 cell array and assign each input argument type of the entry-point function to each cell.

```
ARGS = cell(4,1);
```

For the first input, use `coder.typeof` to specify that the predictor data variable is double-precision with the same number of columns as the predictor data used in training the model, but that the number of observations (rows) is arbitrary.

```
p = numel(Mdl.PredictorNames);
ARGS{1} = coder.typeof(0,[Inf,p],[1,0]);
```

`0` for the `example_value` value implies that the data type is `double` because `double` is the default numeric data type of MATLAB. `[Inf,p]` for the `size_vector` value and `[1,0]` for the `variable_dims` value imply that the size of the first dimension is variable and unbounded, and the size of the second dimension is fixed to be `p`.

The second input is the MAT-file name, which must be a compile-time constant. Use `coder.Constant` to specify the type of the second input.

```
ARGS{2} = coder.Constant(MdlName);
```

The last two inputs are the name and value of the `'Subtrees'` name-value pair argument. Names of name-value pair arguments must be compile-time constants.

```
ARGS{3} = coder.Constant('Subtrees');
```

Use `coder.typeof` to specify that the value of `'Subtrees'` is a double-precision row vector and that the upper bound of the row vector size is `max(Mdl.PrunedList)`.

```
m = max(Mdl.PruneList);
ARGS{4} = coder.typeof(0,[1,m],[0,1]);
```

Again, `0` for the `example_value` value implies that the data type is `double` because `double` is the default numeric data type of MATLAB. `[1,m]` for the `size_vector` value and `[0,1]` for the `variable_dims` value imply that the size of the first dimension is fixed to be `1`, and the size of the second dimension is variable and its upper bound is `m`.

Generate Code Using `codegen`

Generate a MEX function from the entry-point function `mypredictTree` using the cell array `ARGS`, which includes input argument types for `mypredictTree`. Specify the input argument types using the `-args` option. Specify the number of output arguments in the generated entry-point function using the `-nargout` option. The generate code includes the specified number of output arguments in the order in which they occur in the entry-point function definition.

```
codegen mypredictTree -args ARGS -nargout 2
```

```
Code generation successful.
```

`codegen` generates the MEX function `mypredictTree_mex` with a platform-dependent extension in your current folder.

The `predict` function accepts single-precision values, double-precision values, and `'all'` for the `'SubTrees'` name-value pair argument. However, you can specify only double-precision values when you use the MEX function for prediction because the data type specified by `ARGS{4}` is double.

Verify Generated Code

Predict labels for a random selection of 15 values from the training data using the generated MEX function and the subtree at pruning level 1. Compare the labels from the MEX function with those predicted by `predict`.

```
rng('default'); % For reproducibility
Xnew = datasample(meas,15);
[labelMEX,scoreMEX] = mypredictTree_mex(Xnew,MdlName,'Subtrees',1);
[labelPREDICT,scorePREDICT] = predict(Mdl,Xnew,'Subtrees',1);
labelPREDICT
```

```
labelPREDICT = 15x10 char array
    'virginica '
    'virginica '
    'setosa    '
    'virginica '
    'versicolor'
    'setosa    '
    'setosa    '
    'versicolor'
    'virginica '
    'virginica '
    'setosa    '
    'virginica '
    'virginica '
    'versicolor'
    'virginica '
```

`labelMEX`

```
labelMEX = 15x1 cell
    {'virginica' }
    {'virginica' }
    {'setosa'    }
    {'virginica' }
    {'versicolor'}
    {'setosa'    }
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
    {'virginica' }
    {'setosa'    }
    {'virginica' }
    {'virginica' }
    {'versicolor'}
    {'virginica' }
```

The predicted labels are the same as the MEX function labels except for the data type. When the response data type is `char` and `codegen` cannot determine that the value of `Subtrees` is a scalar, then the output from the generated code is a cell array of character vectors.

For the comparison, you can convert `labelsPREDICT` to a cell array and use `isequal`.

```
cell_labelPREDICT = cellstr(labelPREDICT);
verifyLabel = isequal(labelMEX,cell_labelPREDICT)
```

```
verifyLabel = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal.

Compare the second outputs as well. `scoreMex` might include round-off differences compared with `scorePREDICT`. In this case, compare `scoreMEX` and `scorePREDICT`, allowing a small tolerance.

```
find(abs(scorePREDICT-scoreMEX) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

`find` returns an empty vector if the element-wise absolute difference between `scorePREDICT` and `scoreMEX` is not larger than the specified tolerance `1e-8`. The comparison confirms that `scorePREDICT` and `scoreMEX` are equal within the tolerance `1e−8`.

See Also

`codegen` | `coder.Constant` | `coder.typeof` | `learnerCoderConfigurer` | `loadLearnerForCoder` | `saveLearnerForCoder`

Related Examples

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
- “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80
- “Code Generation for Nearest Neighbor Searcher” on page 32-19
- “Code Generation and Classification Learner App” on page 32-31

Create Dummy Variables for Categorical Predictors and Generate C/C++ Code

This example shows how to generate code for classifying data using a support vector machine (SVM) model. Train the model using numeric and encoded categorical predictors. Use `dummyvar` to convert categorical predictors to numeric dummy variables before fitting an SVM classifier. When passing new data to your trained model, you must preprocess the data in a similar manner.

Alternatively, if a trained model identifies categorical predictors in the `CategoricalPredictors` property, then you do not need to create dummy variables manually to generate code. The software handles categorical predictors automatically. For an example, see “Generate Code to Classify Data in Table” on page 32-100.

Preprocess Data and Train SVM Classifier

Load the `patients` data set. Create a table using the `Diastolic` and `Systolic` numeric variables. Each row of the table corresponds to a different patient.

```
load patients
tbl = table(Diastolic,Systolic);
head(tbl)
```

```
ans=8×2 table
    Diastolic    Systolic
    _________    ________

       93           124
       77           109
       83           125
       75           117
       80           122
       70           121
       88           130
       82           115
```

Convert the `Gender` variable to a `categorical` variable. The order of the categories in `categoricalGender` is important because it determines the order of the columns in the predictor data. Use `dummyvar` to convert the categorical variable to a matrix of zeros and ones, where a `1` value in the `(i,j)`th entry indicates that the `i`th patient belongs to the `j`th category.

```
categoricalGender = categorical(Gender);
orderGender = categories(categoricalGender)
```

```
orderGender = 2x1 cell
    {'Female'}
    {'Male'  }
```

```
dummyGender = dummyvar(categoricalGender);
```

Note: The resulting `dummyGender` matrix is rank deficient. Depending on the type of model you train, this rank deficiency can be problematic. For example, when training linear models, remove the first column of the dummy variables.

Create a table that contains the dummy variable `dummyGender` with the corresponding variable headings. Combine this new table with `tbl`.

```
tblGender = array2table(dummyGender,'VariableNames',orderGender);
tbl = [tbl tblGender];
head(tbl)
```

```
ans=8×4 table
    Diastolic    Systolic    Female    Male
    _________    ________    ______    ____

       93           124         0        1
       77           109         0        1
       83           125         1        0
       75           117         1        0
       80           122         1        0
       70           121         1        0
       88           130         1        0
       82           115         0        1
```

Convert the `SelfAssessedHealthStatus` variable to a `categorical` variable. Note the order of the categories in `categoricalHealth`, and convert the variable to a numeric matrix using `dummyvar`.

```
categoricalHealth = categorical(SelfAssessedHealthStatus);
orderHealth = categories(categoricalHealth)
```

```
orderHealth = 4x1 cell
    {'Excellent'}
    {'Fair'     }
    {'Good'     }
    {'Poor'     }
```

```
dummyHealth = dummyvar(categoricalHealth);
```

Create a table that contains `dummyHealth` with the corresponding variable headings. Combine this new table with `tbl`.

```
tblHealth = array2table(dummyHealth,'VariableNames',orderHealth);
tbl = [tbl tblHealth];
head(tbl)
```

```
ans=8×8 table
    Diastolic    Systolic    Female    Male    Excellent    Fair    Good    Poor
    _________    ________    ______    ____    _________    ____    ____    ____

       93           124         0        1          1          0       0       0
       77           109         0        1          0          1       0       0
       83           125         1        0          0          0       1       0
       75           117         1        0          0          1       0       0
       80           122         1        0          0          0       1       0
       70           121         1        0          0          0       1       0
       88           130         1        0          0          0       1       0
       82           115         0        1          0          0       1       0
```

The third row of `tbl`, for example, corresponds to a patient with these characteristics: diastolic blood pressure of 83, systolic blood pressure of 125, female, and good self-assessed health status.

Because all the values in `tbl` are numeric, you can convert the table to a matrix `X`.

```
X = table2array(tbl);
```

Train an SVM classifier using `X`. Specify the `Smoker` variable as the response.

```
Y = Smoker;
Mdl = fitcsvm(X,Y);
```

Generate C/C++ Code

Generate code that loads the SVM classifier, takes new predictor data as an input argument, and then classifies the new data.

Save the SVM classifier to a file using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'SVMClassifier')
```

`saveLearnerForCoder` saves the classifier to the MATLAB® binary file `SVMClassifier.mat` as a structure array in the current folder.

Define the entry-point function `mySVMPredict`, which takes new predictor data as an input argument. Within the function, load the SVM classifier by using `loadLearnerForCoder`, and then pass the loaded classifier to `predict`.

```
function label = mySVMPredict(X) %#codegen
Mdl = loadLearnerForCoder('SVMClassifier');
label = predict(Mdl,X);
end
```

Generate code for `mySVMPredict` by using `codegen`. Specify the data type and dimensions of the new predictor data by using `coder.typeof` so that the generated code accepts a variable-size array.

```
codegen mySVMPredict -args {coder.typeof(X,[Inf 8],[1 0])}
```

```
Code generation successful.
```

Verify that `mySVMPredict` and the MEX file return the same results for the training data.

```
label = predict(Mdl,X);
mylabel = mySVMPredict(X);
mylabel_mex = mySVMPredict_mex(X);
verifyMEX = isequal(label,mylabel,mylabel_mex)
```

```
verifyMEX = logical
   1
```

Predict Labels for New Data

To predict labels for new data, you must first preprocess the new data. If you run the generated code in the MATLAB environment, you can follow the preprocessing steps described in this section. If you deploy the generated code outside the MATLAB environment, the preprocessing steps can differ. In either case, you must ensure that the new data has the same columns as the training data `X`.

In this example, take the third, fourth, and fifth patients in the `patients` data set. Preprocess the data for these patients so that the resulting numeric matrix matches the form of the training data.

Convert the categorical variables to dummy variables. Because the new observations might not include values from all categories, you need to specify the same categories as the ones used during training and maintain the same category order. In MATLAB, pass the ordered cell array of category names associated with the corresponding training data variable (in this example, `orderGender` for gender values and `orderHealth` for self-assessed health status values).

```
newcategoricalGender = categorical(Gender(3:5),orderGender);
newdummyGender = dummyvar(newcategoricalGender);

newcategoricalHealth = categorical(SelfAssessedHealthStatus(3:5),orderHealth);
newdummyHealth = dummyvar(newcategoricalHealth);
```

Combine all the new data into a numeric matrix.

```
newX = [Diastolic(3:5) Systolic(3:5) newdummyGender newdummyHealth]
```

```
newX = 3×8

    83   125     1     0     0     0     1     0
    75   117     1     0     0     1     0     0
    80   122     1     0     0     0     1     0
```

Note that `newX` corresponds exactly to the third, fourth, and fifth rows of the matrix `X`.

Verify that `mySVMPredict` and the MEX file return the same results for the new data.

```
newlabel = predict(Mdl,newX);
newmylabel = mySVMPredict(newX);
newmylabel_mex = mySVMPredict_mex(newX);
newverifyMEX = isequal(newlabel,newmylabel,newmylabel_mex)
```

```
newverifyMEX = logical
   1
```

See Also

`ClassificationSVM` | `categorical` | `codegen` | `coder.Constant` | `coder.typeof` | `dummyvar` | `loadLearnerForCoder` | `saveLearnerForCoder`

Related Examples

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
- “Code Generation and Classification Learner App” on page 32-31

System Objects for Classification and Code Generation

This example shows how to generate C code from a MATLAB® System object™ that classifies images of digits by using a trained classification model. This example also shows how to use the System object for classification in Simulink®. The benefit of using System objects over MATLAB function is that System objects are more appropriate for processing large amounts of streaming data. For more details, see “What Are System Objects?”.

This example is based on “Code Generation for Image Classification” on page 32-103, which is an alternative workflow to “Digit Classification Using HOG Features” (Computer Vision Toolbox).

Load Data

Load the `digitimages`.

```
load digitimages.mat
```

`images` is a 28-by-28-by-3000 array of `uint16` integers. Each page is a raster image of a digit. Each element is a pixel intensity. Corresponding labels are in the 3000-by-1 numeric vector `Y`. For more details, enter `Description` at the command line.

Store the number of observations and the number of predictor variables. Create a data partition that specifies to hold out 20% of the data. Extract training and test set indices from the data partition.

```
rng(1); % For reproducibility
n = size(images,3);
p = numel(images(:,:,1));
cvp = cvpartition(n,'Holdout',0.20);
idxTrn = training(cvp);
idxTest = test(cvp);
```

Rescale Data

Rescale the pixel intensities so that they range in the interval [0,1] within each image. Specifically, suppose p_{ij} is pixel intensity j within image i. For image i, rescale all of its pixel intensities by using this formula:

$$\hat{p}_{ij} = \frac{p_{ij} - \min_j(p_{ij})}{\max_j(p_{ij}) - \min_j(p_{ij})}.$$

```
X = double(images);

for i = 1:n
    minX = min(min(X(:,:,i)));
    maxX = max(max(X(:,:,i)));
    X(:,:,i) = (X(:,:,i) - minX)/(maxX - minX);
end
```

Reshape Data

For code generation, the predictor data for training must be in a table of numeric variables or a numeric matrix.

Reshape the data to a matrix such that predictor variables correspond to columns and images correspond to rows. Because `reshape` takes elements column-wise, transpose its result.

```
X = reshape(X,[p,n])';
```

Train and Optimize Classification Models

Cross-validate an ECOC model of SVM binary learners and a random forest based on the training observations. Use 5-fold cross-validation.

For the ECOC model, specify predictor standardization and optimize classification error over the ECOC coding design and the SVM box constraint. Explore all combinations of these values:

- For the ECOC coding design, use one-versus-one and one-versus-all.
- For the SVM box constraint, use three logarithmically spaced values from 0.1 to 100 each. For all models, store the 5-fold cross-validated misclassification rates.

```
coding = {'onevsone' 'onevsall'};
boxconstraint = logspace(-1,2,3);
cvLossECOC = nan(numel(coding),numel(boxconstraint)); % For preallocation

for i = 1:numel(coding)
    for j = 1:numel(boxconstraint)
        t = templateSVM('BoxConstraint',boxconstraint(j),'Standardize',true);
        CVMdl = fitcecoc(X(idxTrn,:),Y(idxTrn),'Learners',t,'KFold',5,...
            'Coding',coding{i});
        cvLossECOC(i,j) = kfoldLoss(CVMdl);
        fprintf('cvLossECOC = %f for model using %s coding and box constraint=%f\n',...
            cvLossECOC(i,j),coding{i},boxconstraint(j))
    end
end
```

```
cvLossECOC = 0.058333 for model using onevsone coding and box constraint=0.100000
cvLossECOC = 0.057083 for model using onevsone coding and box constraint=3.162278
cvLossECOC = 0.050000 for model using onevsone coding and box constraint=100.000000
cvLossECOC = 0.120417 for model using onevsall coding and box constraint=0.100000
cvLossECOC = 0.121667 for model using onevsall coding and box constraint=3.162278
cvLossECOC = 0.127917 for model using onevsall coding and box constraint=100.000000
```

For the random forest, vary the maximum number of splits by using the values in the sequence $\{3^2, 3^3, \ldots, 3^m\}$. m is such that 3^m is no greater than n - 1. To reproduce random predictor selections, specify `'Reproducible',true`.

```
n = size(X,1);
m = floor(log(n - 1)/log(3));
maxNumSplits = 3.^(2:m);
cvLossRF = nan(numel(maxNumSplits));
for i = 1:numel(maxNumSplits)
    t = templateTree('MaxNumSplits',maxNumSplits(i),'Reproducible',true);
    CVMdl = fitcensemble(X(idxTrn,:),Y(idxTrn),'Method','bag','Learners',t,...
        'KFold',5);
    cvLossRF(i) = kfoldLoss(CVMdl);
    fprintf('cvLossRF = %f for model using %d as the maximum number of splits\n',...
        cvLossRF(i),maxNumSplits(i))
end
```

```
cvLossRF = 0.319167 for model using 9 as the maximum number of splits
cvLossRF = 0.192917 for model using 27 as the maximum number of splits
cvLossRF = 0.066250 for model using 81 as the maximum number of splits
cvLossRF = 0.015000 for model using 243 as the maximum number of splits
```

```
cvLossRF = 0.013333 for model using 729 as the maximum number of splits
cvLossRF = 0.009583 for model using 2187 as the maximum number of splits
```

For each algorithm, determine the hyperparameter indices that yield the minimal misclassification rates.

```
minCVLossECOC = min(cvLossECOC(:))
```

```
minCVLossECOC = 0.0500
```

```
linIdx = find(cvLossECOC == minCVLossECOC,1);
[bestI,bestJ] = ind2sub(size(cvLossECOC),linIdx);
bestCoding = coding{bestI}
```

```
bestCoding =
'onevsone'
```

```
bestBoxConstraint = boxconstraint(bestJ)
```

```
bestBoxConstraint = 100
```

```
minCVLossRF = min(cvLossRF(:))
```

```
minCVLossRF = 0.0096
```

```
linIdx = find(cvLossRF == minCVLossRF,1);
[bestI,bestJ] = ind2sub(size(cvLossRF),linIdx);
bestMNS = maxNumSplits(bestI)
```

```
bestMNS = 2187
```

The random forest achieves a smaller cross-validated misclassification rate.

Train an ECOC model and a random forest using the training data. Supply the optimal hyperparameter combinations.

```
t = templateSVM('BoxConstraint',bestBoxConstraint,'Standardize',true);
MdlECOC = fitcecoc(X(idxTrn,:),Y(idxTrn),'Learners',t,'Coding',bestCoding);
t = templateTree('MaxNumSplits',bestMNS);
MdlRF = fitcensemble(X(idxTrn,:),Y(idxTrn),'Method','bag','Learners',t);
```

Create a variable for the test sample images and use the trained models to predict test sample labels.

```
testImages = X(idxTest,:);
testLabelsECOC = predict(MdlECOC,testImages);
testLabelsRF = predict(MdlRF,testImages);
```

Save Classification Model to Disk

`MdlECOC` and `MdlRF` are predictive classification models, but you must prepare them for code generation. Save `MdlECOC` and `MdlRF` to your present working folder using `saveLearnerForCoder`.

```
saveLearnerForCoder(MdlECOC,'DigitImagesECOC');
saveLearnerForCoder(MdlRF,'DigitImagesRF');
```

Create System Object for Prediction

Create two System objects, one for the ECOC model and the other for the random forest, that:

- Load the previously saved trained model by using `loadLearnerForCoder`.
- Make sequential predictions by the `step` method.
- Enforce no size changes to the input data.
- Enforce double-precision, scalar output.

```
type ECOCClassifier.m % Display contents of ECOCClassifier.m file
```

```
classdef ECOCClassifier < matlab.System
    % ECOCCLASSIFIER Predict image labels from trained ECOC model
    %
    % ECOCCLASSIFIER loads the trained ECOC model from
    % |'DigitImagesECOC.mat'|, and predicts labels for new observations
    % based on the trained model.  The ECOC model in
    % |'DigitImagesECOC.mat'| was cross-validated using the training data
    % in the sample data |digitimages.mat|.

    properties(Access = private)
        CompactMdl % The compacted, trained ECOC model
    end

    methods(Access = protected)

        function setupImpl(obj)
            % Load ECOC model from file
            obj.CompactMdl = loadLearnerForCoder('DigitImagesECOC');
        end

        function y = stepImpl(obj,u)
            y = predict(obj.CompactMdl,u);
        end

        function flag = isInputSizeMutableImpl(obj,index)
            % Return false if input size is not allowed to change while
            % system is running
            flag = false;
        end

        function dataout = getOutputDataTypeImpl(~)
            dataout = 'double';
        end

        function sizeout = getOutputSizeImpl(~)
            sizeout = [1 1];
        end
    end
end
```

```
type RFClassifier.m % Display contents of RFClassifier.m file
```

```
classdef RFClassifier < matlab.System
    % RFCLASSIFIER Predict image labels from trained random forest
    %
    % RFCLASSIFIER loads the trained random forest from
    % |'DigitImagesRF.mat'|, and predicts labels for new observations based
    % on the trained model.  The random forest in |'DigitImagesRF.mat'|
    % was cross-validated using the training data in the sample data
    % |digitimages.mat|.
```

```
    properties(Access = private)
        CompactMdl % The compacted, trained random forest
    end

    methods(Access = protected)

        function setupImpl(obj)
            % Load random forest from file
            obj.CompactMdl = loadLearnerForCoder('DigitImagesRF');
        end

        function y = stepImpl(obj,u)
            y = predict(obj.CompactMdl,u);
        end

        function flag = isInputSizeMutableImpl(obj,index)
            % Return false if input size is not allowed to change while
            % system is running
            flag = false;
        end

        function dataout = getOutputDataTypeImpl(~)
            dataout = 'double';
        end

        function sizeout = getOutputSizeImpl(~)
            sizeout = [1 1];
        end
    end
end
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the files used in this example.

For System object basic requirements, see “Define Basic System Objects”.

Define Prediction Functions for Code Generation

Define two MATLAB functions called `predictDigitECOCSO.m` and `predictDigitRFSO.m`. The functions:

- Include the code generation directive `%#codegen`.
- Accept image data commensurate with `X`.
- Predict labels using the `ECOCClassifier` and `RFClassifier` System objects, respectively.
- Return predicted labels.

```
type predictDigitECOCSO.m % Display contents of predictDigitECOCSO.m file
```

```
function label = predictDigitECOCSO(X) %#codegen
%PREDICTDIGITECOCSO Classify digit in image using ECOC Model System object
%   PREDICTDIGITECOCSO classifies the 28-by-28 images in the rows of X
%   using the compact ECOC model in the System object ECOCClassifier, and
%   then returns class labels in label.
classifier = ECOCClassifier;
```

```
label = step(classifier,X);
end
```

```
type predictDigitRFSO.m % Display contents of predictDigitRFSO.m file
```

```
function label = predictDigitRFSO(X) %#codegen
%PREDICTDIGITRFSO Classify digit in image using RF Model System object
%   PREDICTDIGITRFSO classifies the 28-by-28 images in the rows of X
%   using the compact random forest in the System object RFClassifier, and
%   then returns class labels in label.
classifier = RFClassifier;
label = step(classifier,X);
end
```

Compile MATLAB Function to MEX File

Compile the prediction function that achieves better test-sample accuracy to a MEX file by using `codegen`. Specify the test set images by using the `-args` argument.

```
if(minCVLossECOC <= minCVLossRF)
    codegen predictDigitECOCSO -args testImages
else
    codegen predictDigitRFSO -args testImages
end
```

```
Code generation successful.
```

Verify that the generated MEX file produces the same predictions as the MATLAB function.

```
if(minCVLossECOC <= minCVLossRF)
    mexLabels = predictDigitECOCSO_mex(testImages);
    verifyMEX = sum(mexLabels == testLabelsECOC) == numel(testLabelsECOC)
else
    mexLabels = predictDigitRFSO_mex(testImages);
    verifyMEX = sum(mexLabels == testLabelsRF) == numel(testLabelsRF)
end
```

```
verifyMEX = logical
   1
```

`verifyMEX` is `1`, which indicates that the predictions made by the generated MEX file and the corresponding MATLAB function are the same.

Predict Labels by Using System Objects in Simulink

Create a video file that displays the test-set images frame-by-frame.

```
v = VideoWriter('testImages.avi','Uncompressed AVI');
v.FrameRate = 1;
open(v);
dim = sqrt(p)*[1 1];
for j = 1:size(testImages,1)
    writeVideo(v,reshape(testImages(j,:),dim));
end
close(v);
```

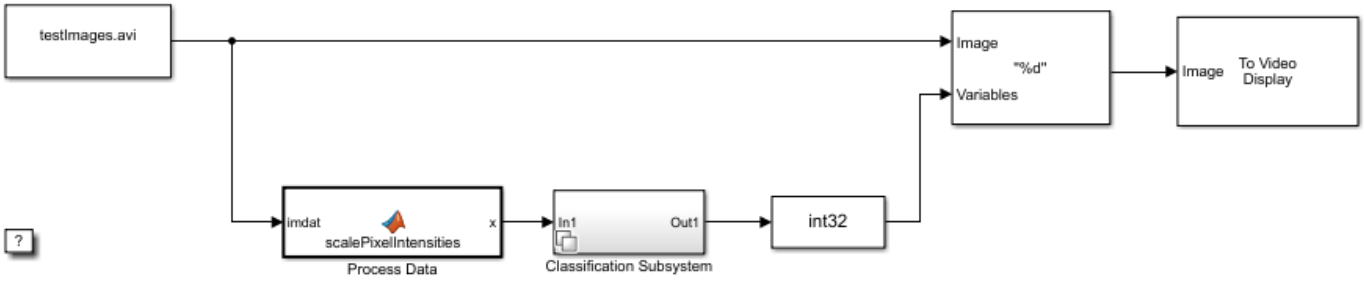
Define a function called `scalePixelIntensities.m` that converts RGB images to grayscale, and then scales the resulting pixel intensities so that their values are in the interval [0,1].

```
type scalePixelIntensities.m % Display contents of scalePixelIntensities.m file
```

```
function x = scalePixelIntensities(imdat)
%SCALEPIXELINTENSITIES Scales image pixel intensities
%   SCALEPIXELINTENSITIES scales the pixel intensities of the image such
%   that the result x is a row vector of values in the interval [0,1].
imdat = rgb2gray(imdat);

minimdat = min(min(imdat));
maximdat = max(max(imdat));
x = (imdat - minimdat)/(maximdat - minimdat);
end
```

Load the Simulink® model `slexClassifyAndDisplayDigitImages.slx`.

```
SimMdlName = 'slexClassifyAndDisplayDigitImages';
open_system(SimMdlName);
```



The figure displays the Simulink® model. At the beginning of simulation, the From Multimedia File block loads the video file of the test-set images. For each image in the video:

- The From Multimedia File block converts and outputs the image to a 28-by-28 matrix of pixel intensities.
- The Process Data block scales the pixel intensities by using `scalePixelIntensities.m`, and outputs a 1-by-784 vector of scaled intensities.
- The Classification Subsystem block predicts labels given the processed image data. The block chooses the System object that minimizes classification error. In this case, the block chooses the random forest. The block outputs a double-precision scalar label.
- The Data Type Conversion block converts the label to an `int32` scalar.
- The Insert Text block embeds the predicted label on the current frame.
- The To Video Display block displays the annotated frame.

Simulate the model.

```
sim(SimMdlName)
```

The model displays all 600 test-set images and its prediction quickly. The last image remains in the video display. You can generate predictions and display them with corresponding images one-by-one by clicking the **Step Forward** button instead.

If you also have a Simulink® Coder™ license, then you can generate C code from `slexClassifyAndDisplayDigitImages.slx` in Simulink® or from the command line using `slbuild` (Simulink). For more details, see “Generate C Code for a Model” (Simulink Coder).

See Also

`loadLearnerForCoder` | `predict` | `predict` | `saveLearnerForCoder`

Related Examples

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Image Classification” on page 32-103
- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “Predict Class Labels Using Stateflow” on page 32-62
- “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66
- “Digit Classification Using HOG Features” (Computer Vision Toolbox)

Predict Class Labels Using Stateflow

This example shows how to use a Stateflow® chart for label prediction. The example trains a discriminant analysis model for the Fisher iris data set by using `fitcdiscr`, and defines a function for code generation that loads the trained model and predicts labels for new data. The Stateflow chart in this example accepts streaming data and predicts labels using the function you define.

Fisher's iris data set, which is included in Statistics and Machine Learning Toolbox™, contains species (`species`) and measurements (`meas`) on sepal length, sepal width, petal length, and petal width for 150 iris specimens. The data set contains 50 specimens from each of three species: setosa, versicolor, and virginica.

Load the Fisher iris data set.

```
load fisheriris
```

Convert `species` to an index vector where 1, 2, and 3 correspond to setosa, versicolor, and virginica, respectively.

```
species = grp2idx(species);
```

Partition the data into a training set and a test set.

```
rng('default') % For reproducibility
idx1 = randperm(150,75)';
idx2 = setdiff((1:150)',idx1);
X = meas(idx1,:);
Y = species(idx1,:);
trainX = meas(idx2,:);
trainY = species(idx2,:);
```

Use `trainX` and `trainY` to train a model, and use `X` and `Y` to test the trained model.

Train a quadratic discriminant analysis model.

```
Mdl = fitcdiscr(trainX,trainY,'DiscrimType','quadratic');
```

`Mdl` is a `ClassificationDiscriminant` model. At the command line, you can use `Mdl` to make predictions for new observations. However, you cannot use `Mdl` as an input argument in a function for code generation. Prepare `Mdl` to be loaded within the function by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'DiscrIris');
```

`saveLearnerForCoder` compacts `Mdl` and saves it in the MAT-file `DiscrIris.mat`.

To display the predicted species in the display box of the Stateflow model, define an enumeration class by using a `classdef` block in the MATLAB® file `IrisSpecies.m`.

```
classdef IrisSpecies < Simulink.IntEnumType
  enumeration
    Setosa(1)
    Versicolor(2)
    Virginica(3)
  end
end
```

For details about enumerated data, see “Define Enumerated Data Types” (Stateflow).

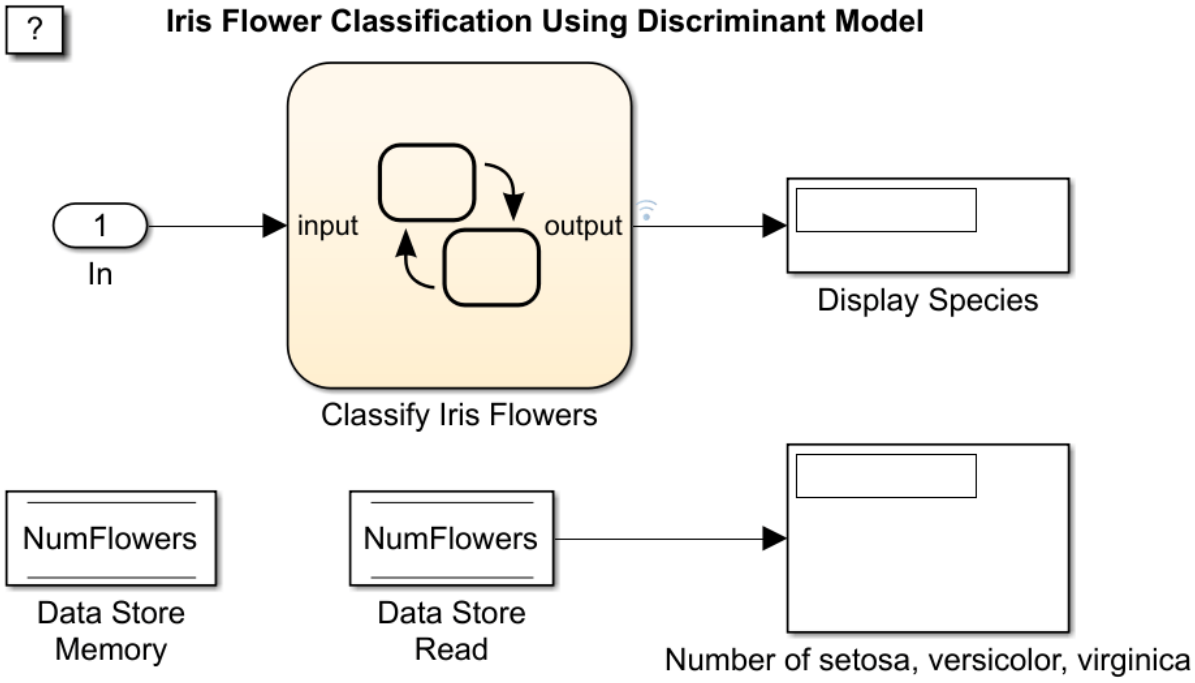
Define a function named `mypredict.m` that predicts the iris species from new measurement data by using the trained model. The function should:

- Include the code generation directive `%#codegen` somewhere in the function.
- Accept iris measurement data. The data must be consistent with X except for the number of rows.
- Load `DiscrIris.mat` using `loadLearnerForCoder`.
- Return predicted iris species.

```
function label  = mypredict(X) %#codegen
%MYPREDICT Predict species of iris flowers using discriminant model
%   mypredict predicts species of iris flowers using the compact
%   discriminant model in the file DiscrIris.mat. Rows of X correspond to
%   observations and columns correspond to predictor variables. label is
%   the predicted species.
mdl = loadLearnerForCoder('DiscrIris');
labelTemp = predict(mdl,X);
label = IrisSpecies(labelTemp);
end
```

Open the Simulink® model `sf_countflowers.slx`.

```
sfName = 'sf_countflowers';
open_system(sfName);
```



The figures display the Simulink model and the flow graph contained in the Stateflow chart. When the input node detects measurement data, it directs the data into the chart. The chart then predicts a species of iris flower and counts the number of flowers for each species. The chart returns the predicted species to the workspace and displays the species within the model, one at a time. The data store memory block `NumFlowers` stores the number of flowers for each species.

The chart expects to receive input data as a structure array called `fisheririsInput` containing these fields:

- `time` - The points in time at which the observations enter the model. In the example, the duration includes the integers from 0 through 74. The orientation of `time` must correspond to the observations in the predictor data. So, for this example, `time` must be a column vector.
- `signals` - A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`. The `values` field is a matrix of predictor data. The `dimensions` field is the number of predictor variables.

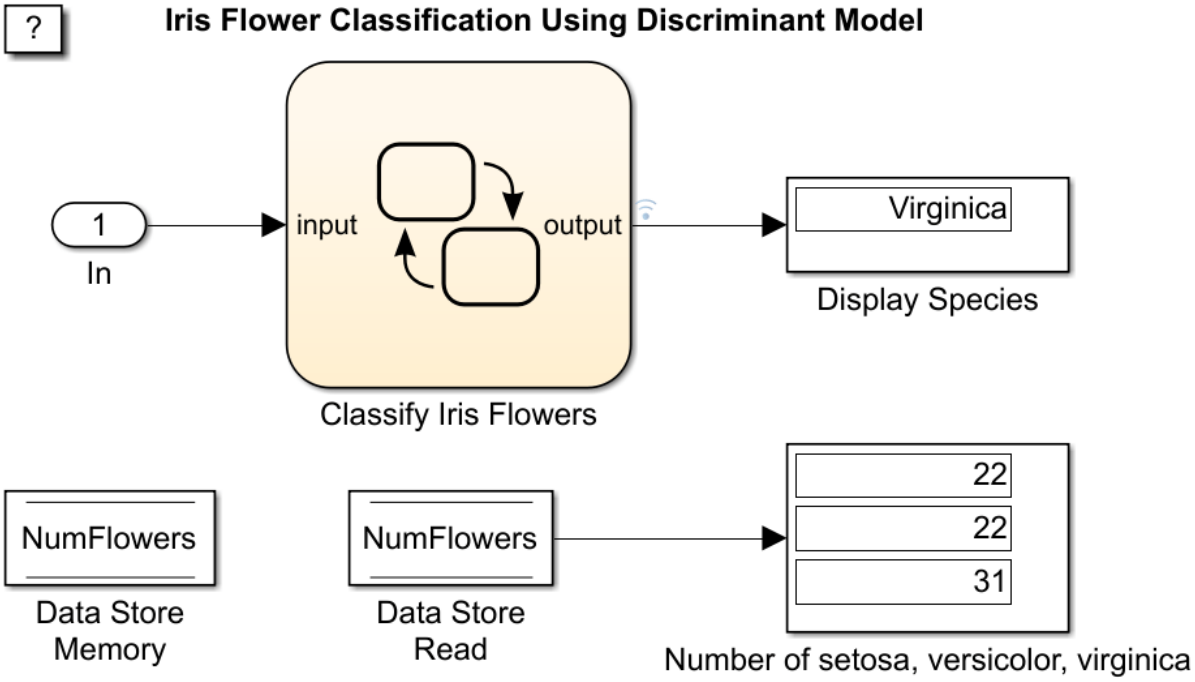
Create an appropriate structure array for iris flower measurements.

```
fisheririsInput.time = (0:74)';
fisheririsInput.signals.dimensions = 4;
fisheririsInput.signals.values = X;
```

You can change the name from `fisheririsInput`, and then specify the new name in the model. However, Stateflow expects the structure array to contain the described field names. For more details, see “Loading Data Structures to Root-Level Inputs” (Simulink).

Simulate the model.

```
sim(sfName)
```



The figure shows the model after it processes all observations in `fisheririsInput`, one at a time. The predicted species of `X(75,:)` is virginica. The number of setosa, versicolor, and virginica in `X` is 22, 22, and 31, respectively.

The variable `logsout` appears in the workspace. `logsout` is a `SimulinkData.Dataset` object containing the predicted species. Extract the predicted species data from the simulation log.

```
labelSF = logsout.getElement(1).Values.Data;
```

Predict species at the command line using `predict`.

```
labelCMD = predict(Mdl,X);
```

Compare the predicted species returned by `sf_countflowers` to those returned by calling `predict` at the command line.

```
isequal(labelCMD,labelSF)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. This comparison confirms that `sf_countflowers` returns the expected results.

If you also have a Simulink Coder™ license, then you can generate C code from `sf_countflowers.slx` in Simulink or from the command line using `rtwbuild` (Simulink Coder). For more details, see “Generate C Code for a Model” (Simulink Coder).

See Also

`loadLearnerForCoder` | `predict` | `saveLearnerForCoder` | `slbuild`

More About

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Image Classification” on page 32-103
- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “System Objects for Classification and Code Generation” on page 32-54
- “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66
- “Chart Programming” (Stateflow)

Human Activity Recognition Simulink Model for Smartphone Deployment

This example shows how to prepare a Simulink® model that classifies human activity based on smartphone sensor signals for code generation and smartphone deployment. The example provides two Simulink models that are ready for deployment to an Android™ device and an iOS device. After you install the required support package for a target device, train the classification model and deploy the Simulink model to the device.

Load Sample Data Set

Load the `humanactivity` data set.

```
load humanactivity
```

The `humanactivity` data set contains 24,075 observations of five different physical human activities: Sitting, Standing, Walking, Running, and Dancing. Each observation has 60 features extracted from acceleration data measured by smartphone accelerometer sensors. The data set contains the following variables:

- `actid` — Response vector containing the activity IDs in integers: 1, 2, 3, 4, and 5 representing Sitting, Standing, Walking, Running, and Dancing, respectively
- `actnames` — Activity names corresponding to the integer activity IDs
- `feat` — Feature matrix of 60 features for 24,075 observations
- `featlabels` — Labels of the 60 features

The Sensor HAR (human activity recognition) App [1] on page 32-0 was used to create the `humanactivity` data set. When measuring the raw acceleration data with this app, a person placed a smartphone in a pocket so that the smartphone was upside down and the screen faced toward the person. The software then calibrated the measured raw data accordingly and extracted the 60 features from the calibrated data. For details about the calibration and feature extraction, see [2] on page 32-0 and [3] on page 32-0 , respectively. The Simulink models described later also use the raw acceleration data and include blocks for calibration and feature extraction.

Prepare Data

This example uses 90% of the observations to train a model that classifies the five types of human activities and 10% of the observations to validate the trained model. Use `cvpartition` to specify a 10% holdout for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(actid,'Holdout',0.10);
trainingInds = training(Partition); % Indices for the training set
XTrain = feat(trainingInds,:);
YTrain = actid(trainingInds);
testInds = test(Partition); % Indices for the test set
XTest = feat(testInds,:);
YTest = actid(testInds);
```

Convert the feature matrix `XTrain` and the response vector `YTrain` into a table to load the training data set in the Classification Learner app.

```
tTrain = array2table([XTrain YTrain]);
```

Specify the variable name for each column of the table.

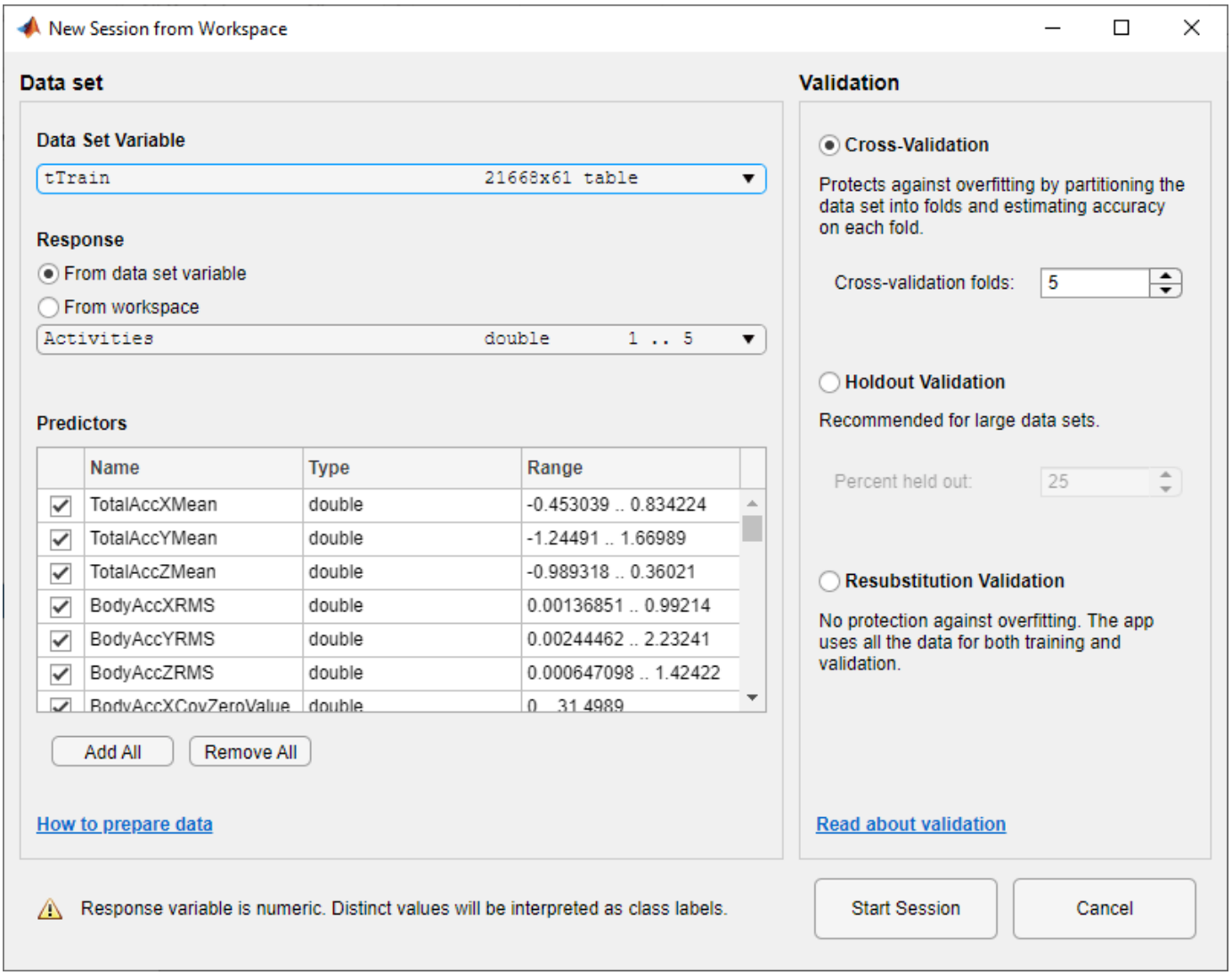
```
tTrain.Properties.VariableNames = [featlabels' 'Activities'];
```

Train Boosted Tree Ensemble Using Classification Learner App

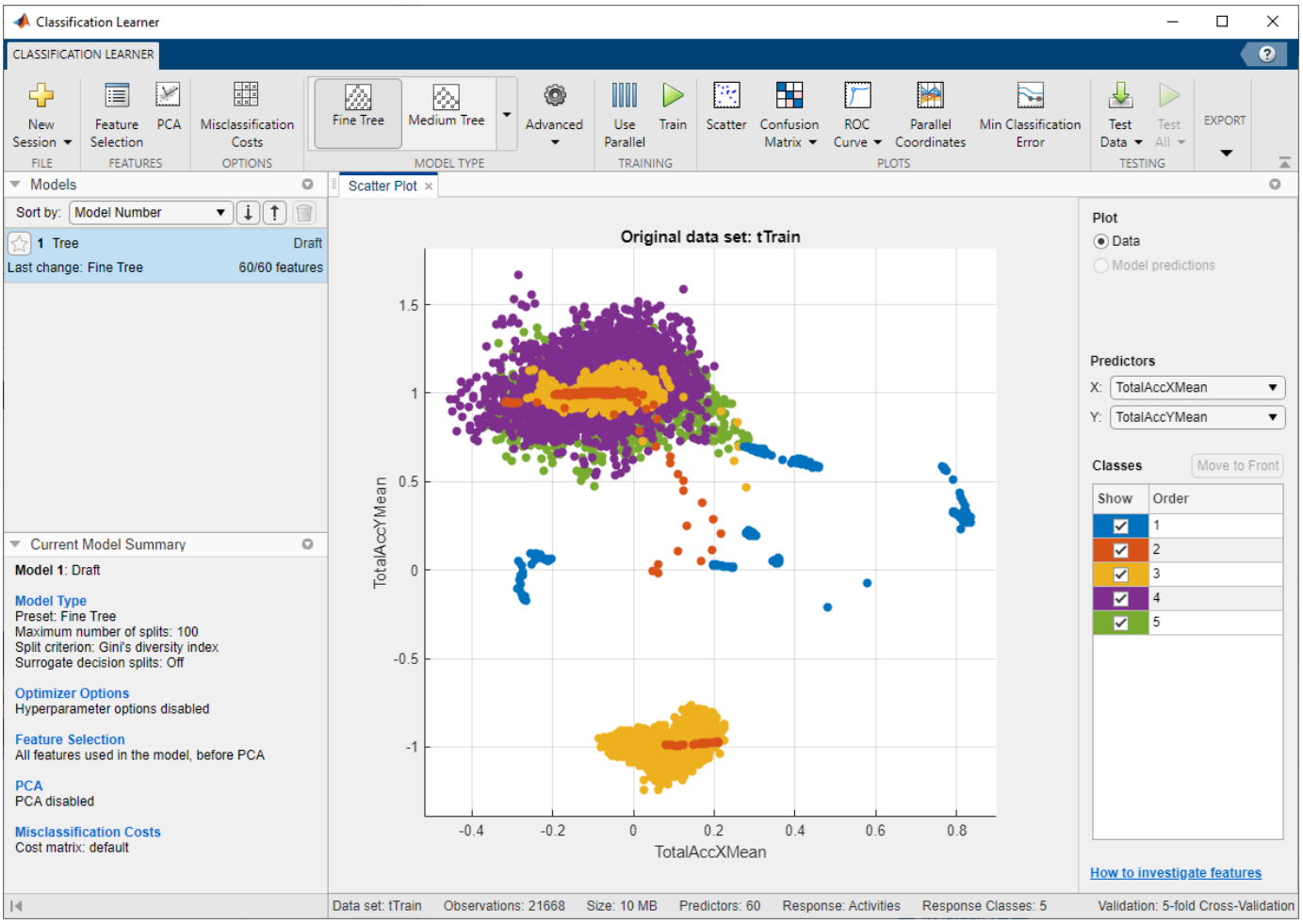
Train a classification model by using the Classification Learner app. To open the Classification Learner app, enter `classificationLearner` at the command line. Alternatively, click the **Apps** tab, and click the arrow at the right of the **Apps** section to open the gallery. Then, under **Machine Learning and Deep Learning**, click **Classification Learner**.

On the **Classification Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.

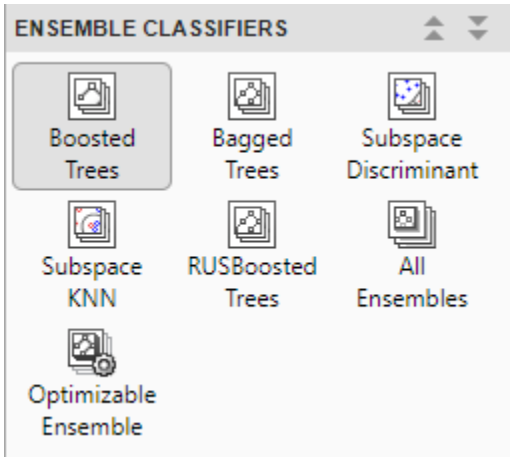
In the New Session from Workspace dialog box, click the arrow for **Data Set Variable**, and then select the table `tTrain`. Classification Learner detects the predictors and the response from the table.



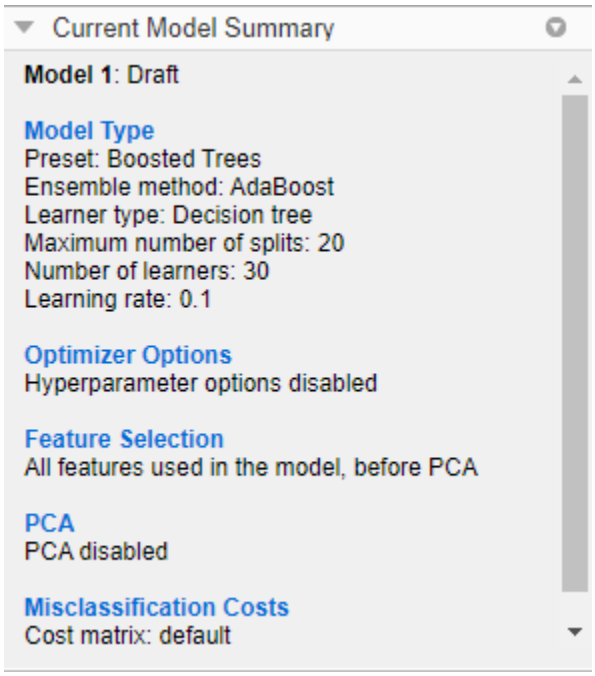
The default option is 5-fold cross-validation, which protects against overfitting. Click **Start Session**. Classification Learner loads the data set and plots a scatter plot of the first two features.



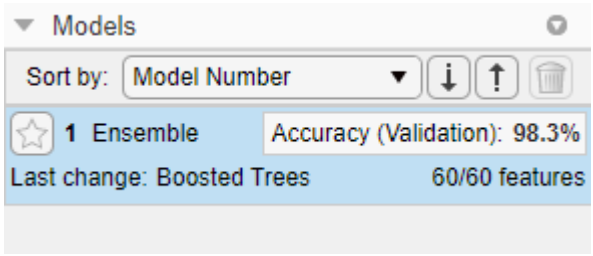
On the **Classification Learner** tab, click the arrow at the right of the **Model Type** section to open the gallery. Then, under **Ensemble Classifiers**, click **Boosted Trees**.



The **Current Model Summary** pane displays the default settings of the boosted tree ensemble model.



On the **Classification Learner** tab, in the **Training** section, click **Train**. When the training is complete, the **Models** pane displays the 5-fold, cross-validated classification accuracy.



On the **Classification Learner** tab, in the **Export** section, click **Export Model**, and then select **Export Compact Model**. Click **OK** in the dialog box. The structure `trainedModel` appears in the MATLAB® Workspace. The field `ClassificationEnsemble` of `trainedModel` contains the compact model. Extract the trained model from the structure.

```
classificationEnsemble = trainedModel.ClassificationEnsemble;
```

Train Boosted Tree Ensemble at Command Line

Alternatively, you can train the same classification model at the command line.

```
template = templateTree('MaxNumSplits',20,'Reproducible',true);
classificationEnsemble = fitcensemble(XTrain,YTrain, ...
    'Method','AdaBoostM2', ...
    'NumLearningCycles',30, ...
    'Learners',template, ...
    'LearnRate',0.1, ...
    'ClassNames',[1; 2; 3; 4; 5]);
```

Perform 5-fold cross-validation for `classificationEnsemble` and compute the validation accuracy.

```
partitionedModel = crossval(classificationEnsemble,'KFold',5);
validationAccuracy = 1-kfoldLoss(partitionedModel)
```

```
validationAccuracy = 0.9833
```

Evaluate Performance on Test Data

Evaluate performance on the test data set.

```
testAccuracy = 1-loss(classificationEnsemble,XTest,YTest)
```

```
testAccuracy = 0.9759
```

The trained model correctly classifies 97.59% of the human activities on the test data set. This result confirms that the trained model does not overfit to the training data set.

Note that the accuracy values can vary slightly depending on your operating system.

Save Trained Model

For code generation including a classification model object, use `saveLearnerForCoder` and `loadLearnerForCoder`.

Save the trained model by using `saveLearnerForCoder`.

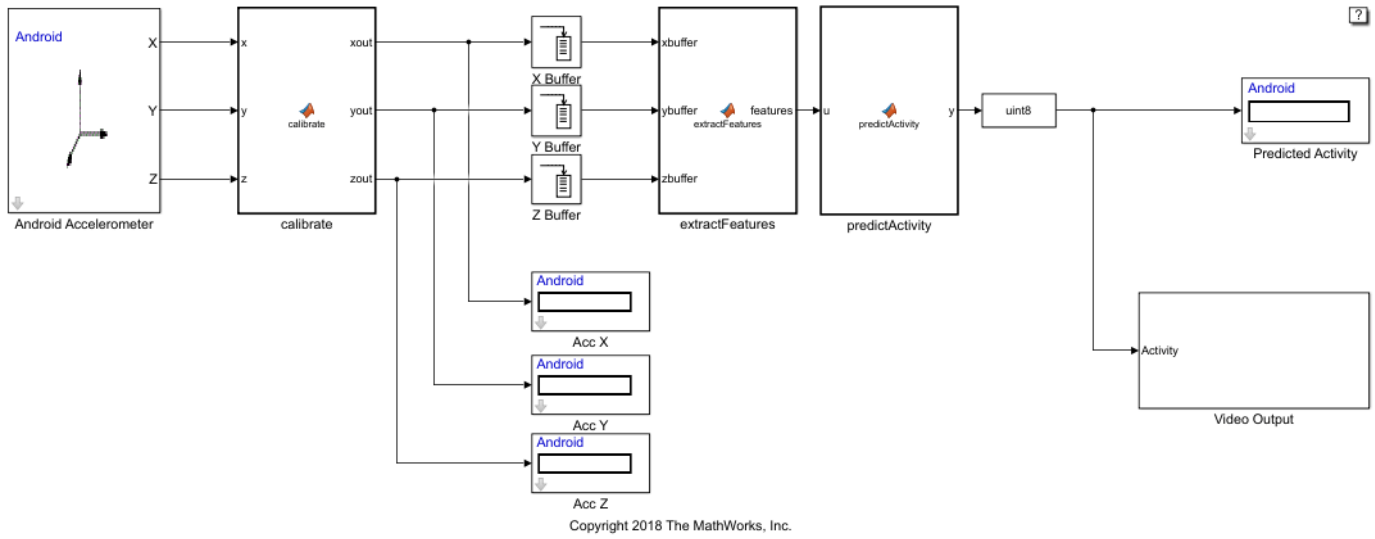
```
saveLearnerForCoder(classificationEnsemble,'EnsembleModel.mat');
```

The function block **predictActivity** in the Simulink models loads the trained model by using `loadLearnerForCoder` and uses the trained model to classify new data.

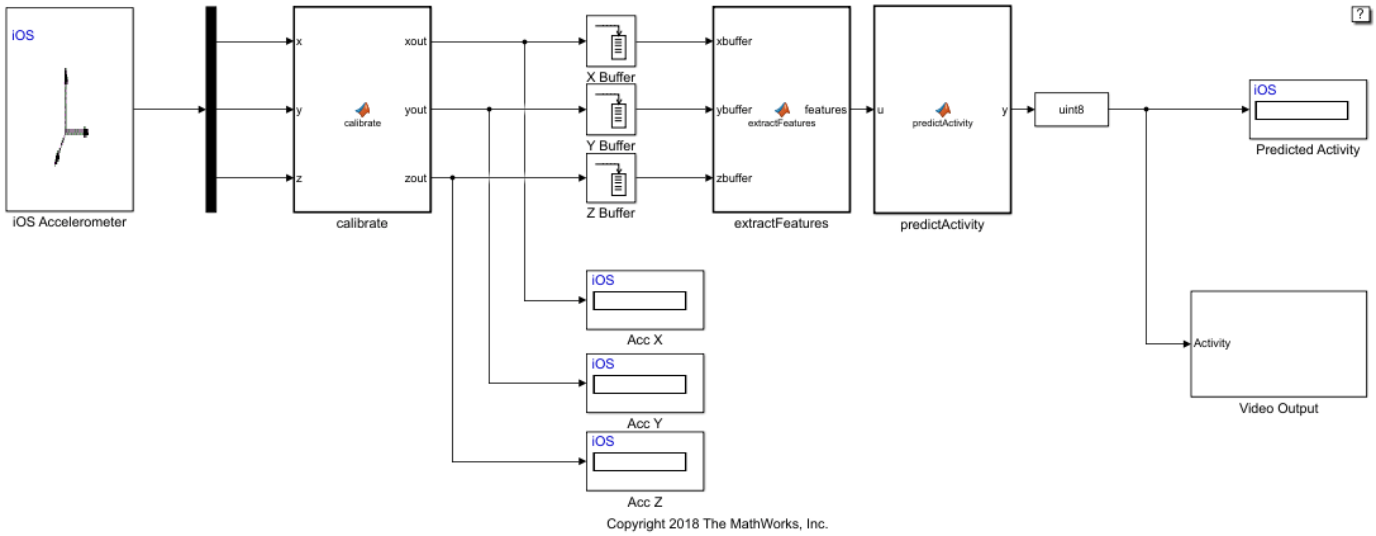
Deploy Simulink Model to Device

Now that you have prepared a classification model, you can open the Simulink model, depending on which type of smartphone you have, and deploy the model to your device. Note that the Simulink model requires the `EnsembleModel.mat` file and the calibration matrix file `slexHARAndroidCalibrationMatrix.mat` or `slexHARiOSCalibrationMatrix.mat`. If you click the button located in the upper-right section of this page and open this example in MATLAB, then MATLAB opens the example folder that includes these calibration matrix files.

Type `slexHARAndroidExample` to open the Simulink model for Android deployment.



Type `slexHARiOSExample` to open the Simulink model for iOS deployment. You can open the model on the Mac OS platform.



The two Simulink models classify human activity based on acceleration data measured by a smartphone sensor. The models include the following blocks:

- The **Accelerometer** block receives raw acceleration data from accelerometer sensors on the device.
- The **calibrate** block is a MATLAB Function block that calibrates the raw acceleration data. This block uses the calibration matrix in the `slexHARAndroidCalibrationMatrix.mat` file or the `slexHARiOSCalibrationMatrix.mat` file. If you click the button located in the upper-right section of this page and open this example in MATLAB, then MATLAB opens the example folder that includes these files.
- The display blocks **Acc X**, **Acc Y**, and **Acc Z** are connected to the **calibrate** block and display calibrated data points for each axis on the device.

- Each of the **Buffer** blocks, **X Buffer**, **Y Buffer**, and **Z Buffer**, buffers 32 samples of an accelerometer axis with 12 samples of overlap between buffered frames. After collecting 20 samples, each **Buffer** block joins the 20 samples with 12 samples from the previous frame and passes the total 32 samples to the **extractFeatures** block. Each **Buffer** block receives an input sample every 0.1 second and outputs a buffered frame including 32 samples every 2 seconds.
- The **extractFeatures** block is a MATLAB Function block that extracts 60 features from a buffered frame of 32 accelerometer samples. This function block uses DSP System Toolbox™ and Signal Processing Toolbox™.
- The **predictActivity** block is a MATLAB Function block that loads the trained model from the `EnsembleModel.mat` file by using `loadLearnerForCoder` and classifies the user activity using the extracted features. The output is an integer between 1 and 5, corresponding to Sitting, Standing, Walking, Running, and Dancing, respectively.
- The **Predicted Activity** block displays the classified user activity values on the device.
- The **Video Output** subsystem uses a multiport switch block to choose the corresponding user activity image data to display on the device. The **Convert to RGB** block decomposes the selected image into separate RGB vectors and passes the image to the **Activity Display** block.

To deploy the Simulink model to your device, follow the steps in “Run Model on Android Devices” (Simulink Support Package for Android Devices) or “Run Model on Apple iOS Devices” (Simulink Support Package for Apple iOS Devices). Run the model on your device, place the device in the same way as described earlier for collecting the training data, and try the five activities. The model displays the classified activity accordingly.

To ensure the accuracy of the model, you need to place your device in the same way as described for collecting the training data. If you want to place your device in a different location or orientation, then collect the data in your own way and use your data to train the classification model.

The accuracy of the model can be different from the accuracy of the test data set (`testaccuracy`), depending on the device. To improve the model, you can consider using additional sensors and updating the calibration matrix. Also, you can add another output block for audio feedback to the output subsystem using Audio Toolbox™. Use a ThingSpeak™ write block to publish classified

activities and acceleration data from your device to the Internet of Things. For details, see https://thingspeak.com/.

References

[1] El Helou, A. Sensor HAR recognition App. MathWorks File Exchange https://www.mathworks.com/matlabcentral/fileexchange/54138-sensor-har-recognition-app

[2] STMicroelectronics, AN4508 Application note. “Parameters and calibration of a low-g 3-axis accelerometer.” 2014.

[3] El Helou, A. Sensor Data Analytics. MathWorks File Exchange https://www.mathworks.com/matlabcentral/fileexchange/54139-sensor-data-analytics--french-webinar-code-

See Also

`fitcensemble` | `loadLearnerForCoder` | `predict` | `saveLearnerForCoder`

More About

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Image Classification” on page 32-103
- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “System Objects for Classification and Code Generation” on page 32-54
- “Predict Class Labels Using Stateflow” on page 32-62

Human Activity Recognition Simulink Model for Fixed-Point Deployment

This example shows how to prepare a Simulink® model that classifies human activity based on sensor signals for code generation and deployment to low-power hardware. The example provides a Simulink classification model that is ready for deployment to a BBC micro:bit device. First, download and install Simulink Coder™ Support Package for BBC micro:bit from the Add-On Explorer. Then, train the classification model and deploy the Simulink model to the target device.

Load Sample Data Set

Load the `humanactivity` data set.

```
load humanactivity
```

The `humanactivity` data set contains 24,075 observations of five physical human activities: Sitting, Standing, Walking, Running, and Dancing. Each observation has 60 features extracted from acceleration data measured by smartphone accelerometer sensors. The data set contains the following variables:

- `actid` — Response vector containing the activity IDs in integers: 1, 2, 3, 4, and 5 representing Sitting, Standing, Walking, Running, and Dancing, respectively
- `actnames` — Activity names corresponding to the integer activity IDs
- `feat` — Feature matrix of 60 features for 24,075 observations
- `featlabels` — Labels of the 60 features

The Sensor HAR (human activity recognition) App [1] on page 32-0 was used to create the `humanactivity` data set. When measuring the raw acceleration data with this app, a person placed a smartphone in a pocket so that the smartphone was upside down and the screen faced toward the person. The software then calibrated the measured raw data accordingly and extracted the 60 features from the calibrated data. For details about the calibration and feature extraction, see [2] on page 32-0 and [3] on page 32-0 , respectively. The Simulink models described later also use the raw acceleration data and include blocks for calibration and feature extraction.

To reduce the memory footprint for fixed-point deployment, specify to use only the first 15 features of the data set in the trained classifier.

```
feat = feat(:,1:15);
featlabels = featlabels(1:15);
```

Prepare Data

This example uses 90% of the observations to train a model that classifies the five types of human activities, and 10% of the observations to validate the trained model. Use `cvpartition` to specify a 10% holdout for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(actid,'Holdout',0.10);
```

Extract the training and test indices.

```
trainInds = training(Partition);
testInds = test(Partition);
```

Specify the training and test data sets.

```
XTrain = feat(trainInds,:);
YTrain = actid(trainInds);
XTest = feat(testInds,:);
YTest = actid(testInds);
```

Train Decision Tree at Command Line

Train a fitted binary classification decision tree using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. Also, specify a maximum of 20 branch nodes for the decision tree.

```
classificationTree = fitctree(XTrain,YTrain,...
    'ClassNames',[1;2;3;4;5],...
    'MaxNumSplits',20)
```

```
classificationTree =
  ClassificationTree
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 21668


  Properties, Methods
```

Perform 5-fold cross-validation for `classificationTree` and compute the validation accuracy.

```
partitionedModel = crossval(classificationTree,'KFold',5);
validationAccuracy = 1-kfoldLoss(partitionedModel)
```

```
validationAccuracy = 0.9700
```

Alternatively, you can train and cross-validate the same classification model using the Classification Learner app. For a similar example, see “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66.

Evaluate Performance on Test Data

Determine how well the algorithm generalizes by estimating the test sample classification error.

```
testAccuracy = 1-loss(classificationTree,XTest,YTest)
```

```
testAccuracy = 0.9617
```

The trained model correctly classifies 96.17% of the human activities on the test data set. This result confirms that the trained model does not overfit to the training data set.

Note that the accuracy values can vary slightly depending on your operating system.

Predict in Simulink Model

Now that you have prepared a classification model, you can open the Simulink model. You can import the trained classification object containing the decision tree `classificationTree` into a ClassificationTree Predict block. You can add this block from the Statistics and Machine Learning

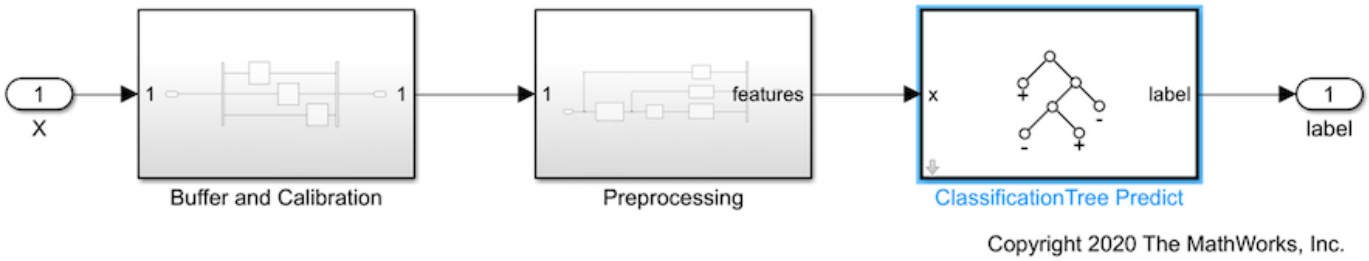
Toolbox™ library. For more information on how to create a model that includes a **ClassificationTree Predict** block, see “Predict Class Labels Using ClassificationTree Predict Block” on page 32-121. In this case, you will use the Simulink model `slexHARFixedPointExample` provided with this example.

Create a large set of accelerometer data `ts` to use as input to the Simulink model.

```
inData = load('rawAccData');
Xacc = inData.acc_data;
t = 0:size(Xacc,1)-1;
ts = timeseries(Xacc,t,'InterpretSingleRowDataAs3D',true);
numSteps = numel(t)-1;
```

Open the Simulink model `slexHARFixedPointExample` by entering the following at the command line. Note that the Simulink model includes callbacks that load necessary variables for the preprocessing subsystem into the base workspace.

```
slexHARFixedPointExample
```



The `slexHARFixedPointExample` model contains the following blocks:

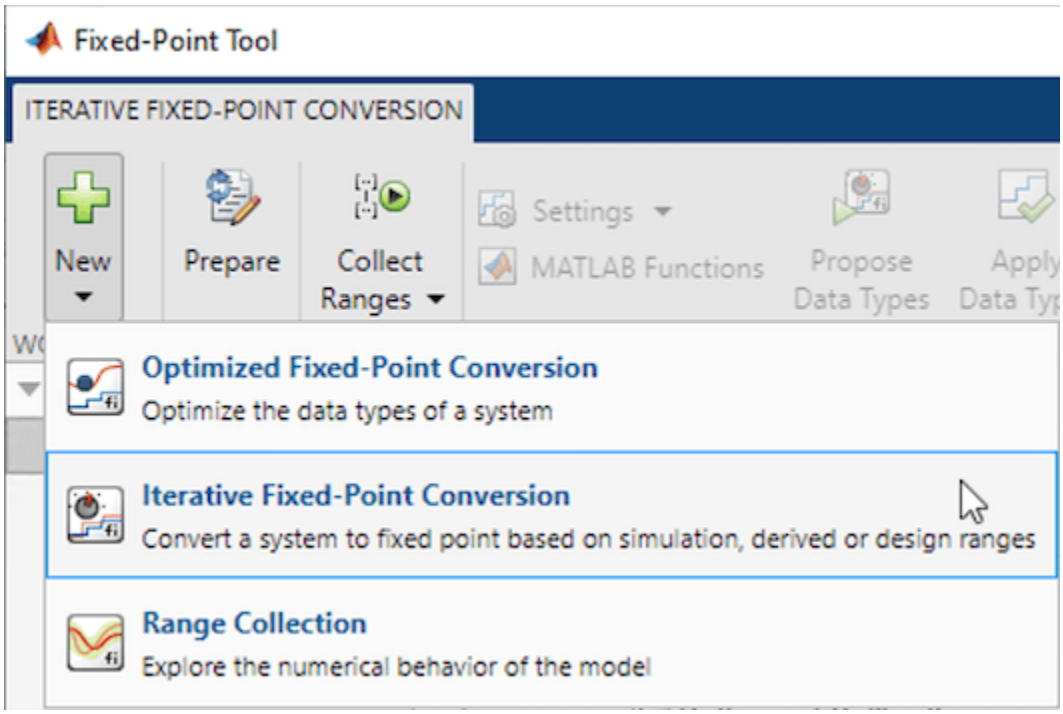
- The **X** block (input port) links the signal `ts` from the Workspace into the system.
- The **Buffer and Calibration** block contains three **Buffer** blocks: **X Buffer**, **Y Buffer**, and **Z Buffer**. Each of these blocks buffers 32 samples of an accelerometer axis with 12 samples of overlap between buffered frames. After collecting 20 samples, each **Buffer** block joins them with 12 samples from the previous frame and passes the total of 32 samples to the **Preprocessing** block. Each **Buffer** block receives an input sample every 0.1 second and outputs a buffered frame including 32 samples every 2 seconds.
- The **Preprocessing** block extracts 15 features from a buffered frame of 32 accelerometer samples. This subsystem block uses DSP System Toolbox™ and Signal Processing Toolbox™.
- The **ClassificationTree Predict** block is a library block from the Statistics and Machine Learning Toolbox library that classifies the human activities using the extracted features. The output is an integer between 1 and 5, corresponding to Sitting, Standing, Walking, Running, and Dancing, respectively.

Convert to Fixed-Point

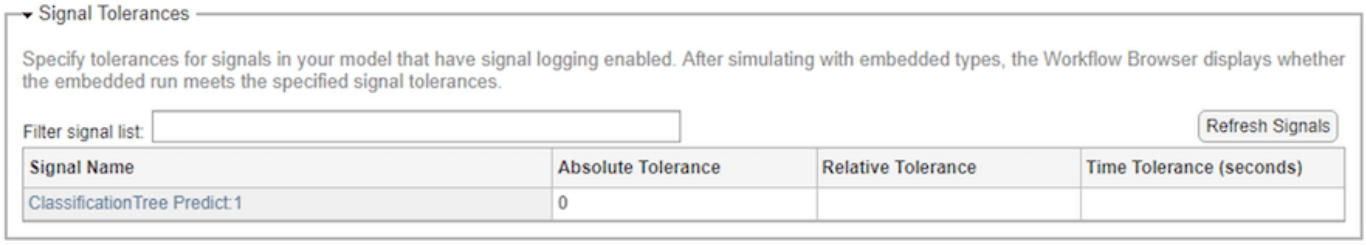
Convert the `slexHARFixedPointExample` model to the fixed-point model `slexHARFixedPointConvertedExample`. Then, deploy `slexHARFixedPointConvertedExample` to the BBC micro:bit board. The target device does not have a floating-point unit (FPU), and performs fixed-point calculations more efficiently than floating-point calculations.

In the `slexHARFixedPointExample` model, right-click the **Label** port and select **Log Selected Signals**. Then, open the **Fixed-Point Tool** app by selecting it from the apps gallery, available from

the **Apps** tab. In the Fixed-Point Tool, under **New** workflow, select `Iterative Fixed-Point Conversion`.



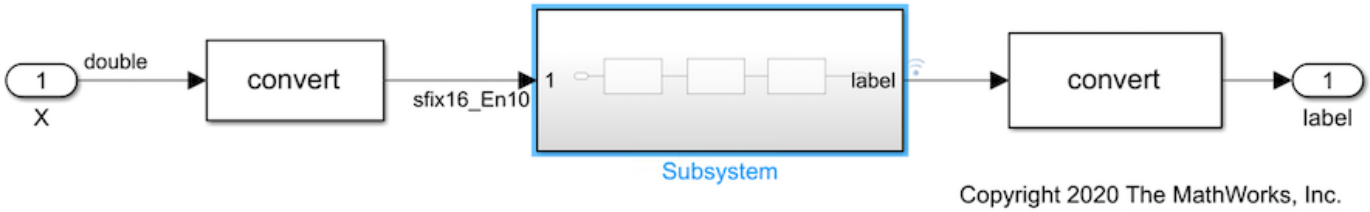
On the **Iterative Fixed-Point Conversion** tab, in the **Signal Tolerances** section, specify the acceptable level of tolerance (difference between the original value and the value of the new design) for the `label` signal. A recommended practice for classification models is to specify 0 absolute tolerance. With this setting, the labels returned by the fixed-point classification model must be the same as the labels returned by the floating-point model. (For regression models, the acceptable tolerance can be a nonzero user-specified number.)



Then, review the steps in the example “Convert Floating-Point Model to Fixed Point” (Fixed-Point Designer) to learn how to create a Simulink model that is converted to fixed-point.

Open the fixed-point Simulink model `slexHARFixedPointConvertedExample` by entering the following at the command line. The **Subsystem** block contains the **Buffer and Calibration**, **Preprocessing**, and **ClassificationTree Predict** blocks as shown earlier for the `slexHARFixedPointExample` model.

```
slexHARFixedPointConvertedExample
```

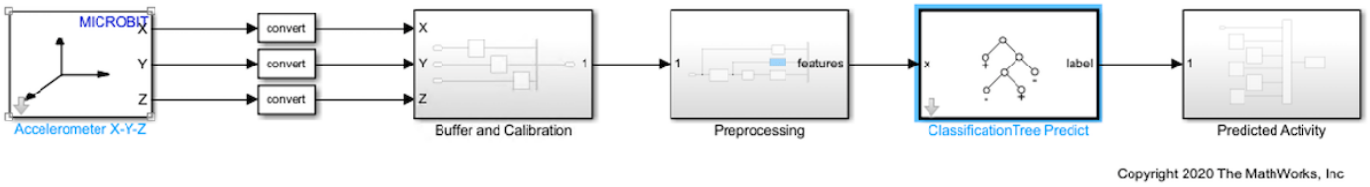


Alternatively, you can select the `Optimized Fixed-Point Conversion` workflow in the **Fixed-Point Tool** app or manually define the data type in the block dialog box. For more information, see "Configure Blocks with Fixed-Point Output" (Fixed-Point Designer).

Deploy to Hardware

Open the Simulink model for deployment to the BBC micro:bit device by entering the following at the command line. The `slexHARFixedPointDeployExample` model is converted to fixed-point, and has I/O blocks for the accelerometer and display ports on the target device.

```
slexHARFixedPointDeployExample
```



The Simulink model `slexHARFixedPointDeployExample` classifies the human activities based on acceleration data measured by a smartphone sensor. The model includes the following blocks:

- The **Accelerometer** block receives raw acceleration data from accelerometer sensors on the device.
- The **Buffer and Calibration**, **Preprocessing**, and **ClassificationTree Predict** blocks are the same as those shown earlier for the `slexHARFixedPointExample` model.
- The **Predicted Activity** block displays the classified human activity values on the 5x5 LED matrix of the BBC micro:bit device. The letters "S", "T", "W", "R", and "D" represent Sitting, Standing, Walking, Running, and Dancing, respectively.

To deploy the Simulink model to your device, follow the steps in "Getting Started with Simulink Coder Support Package for BBC micro:bit" (Simulink Coder Support Package for BBC micro:bit). Run the model on your device, place the device in the same way as described earlier for collecting the training data, and try the five activities. The model displays the classified activity accordingly.

To ensure the accuracy of the model, you must place your device in the same way as described for collecting the training data. If you want to place your device in a different location or at a different orientation, then collect the data in your own way and use your data to train the classification model.

The accuracy of the model can be different from the accuracy of the test data set (`testAccuracy`), depending on the device. To improve the model accuracy, consider using additional sensors, such as a gyroscope.

References

[1] El Helou, Amine. Sensor HAR Recognition App. MathWorks File Exchange https://www.mathworks.com/matlabcentral/fileexchange/54138-sensor-har-recognition-app

[2] STMicroelectronics, AN4508 Application note. “Parameters and calibration of a low-g 3-axis accelerometer.” 2014. https://www.st.com/resource/en/application_note/dm00119044-parameters-and-calibration-of-a-lowg-3axis-accelerometer-stmicroelectronics.pdf

[3] El Helou, Amine. Sensor Data Analytics. MathWorks File Exchange https://www.mathworks.com/matlabcentral/fileexchange/54139-sensor-data-analytics--french-webinar-code-

See Also

ClassificationTree Predict | `crossval` | `fitctree`

More About

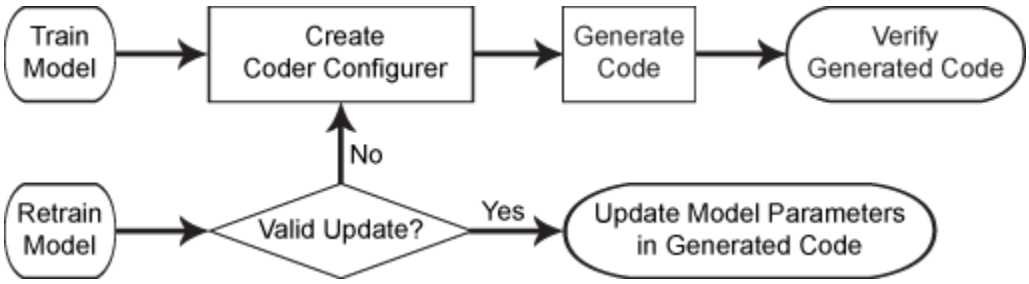
- “Introduction to Code Generation” on page 32-2
- “Predict Class Labels Using ClassificationTree Predict Block” on page 32-121
- “Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130
- “Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “System Objects for Classification and Code Generation” on page 32-54
- “Get Started with Fixed-Point Designer” (Fixed-Point Designer)

Code Generation for Prediction and Update Using Coder Configurer

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes of model parameters using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the model by using `generateCode`. This requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow for the `predict` and `update` functions using a coder configurer.



This table shows coder configurer objects corresponding to the supported machine learning models.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

For details and examples, see the reference pages for the coder configurer objects.

See Also

`generateCode` | `generateFiles` | `learnerCoderConfigurer` | `update` | `validatedUpdateInputs`

More About

- “Introduction to Code Generation” on page 32-2

- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
- “Code Generation and Classification Learner App” on page 32-31
- “Specify Variable-Size Arguments for Code Generation” on page 32-45

Code Generation for Probability Distribution Objects

This example shows how to generate code that fits a probability distribution to sample data and evaluates the fitted distribution.

First, define an entry-point function that uses `fitdist` to create a probability distribution object and uses its object functions to evaluate the fitted distribution. Then, generate code for the entry-point function by using `codegen` (MATLAB Coder). An entry-point function can have a probability distribution object as both an input argument and an output argument. Therefore, alternatively, you can define two entry-point functions, one for fitting a distribution and the other for evaluating the fitted distribution. The first entry-point function returns a fitted distribution, and the second entry-point function accepts the fitted distribution as an input argument. This example first describes the workflow with a single entry-point function, and then briefly describes the workflow with two entry-point functions.

`fitdist` supports code generation for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. The supported object functions of the fitted probability distribution objects, created by `fitdist`, are `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “General Code Generation Workflow” on page 32-5.

Define Entry-Point Function

Define an entry-point function named `myFitandEvaluate` that takes the sample data, distribution name, truncation limits of the distribution, and data values at which to evaluate the cumulative distribution function (cdf) and probability density function (pdf). Within the entry-point function, fit a probability distribution object to the sample data, truncate the distribution to the specified truncation limits, compute the mean of the truncated distribution, and compute the cdf and pdf values at the specified data values.

Display the contents of the `myFitandEvaluate.m` file.

```
type myFitandEvaluate.m
```

```
function [pd_truncated,st] = myFitandEvaluate(data,distname,truncation_limits,x) %#codegen
% Fit a probability distribution object to data.
pd = fitdist(data,distname);

% Truncate pd.
pd_truncated = truncate(pd,truncation_limits(1),truncation_limits(2));

% Compute the mean of the truncated pd.
mean_val = mean(pd_truncated);

% Compute the cdf and pdf, evaluated at x.
cdf_val = cdf(pd_truncated,x);
pdf_val = pdf(pd_truncated,x);

% Create a structure array containing the mean, cdf, and pdf values.
st = struct('mean', mean_val,'cdf',cdf_val,'pdf',pdf_val);
end
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function files for this example.

Generate Code

Specify the input argument types of `myFitandEvaluate` using a 4-by-1 cell array. Assign each input argument type of the entry-point function to each cell. Specify the data type and exact input array size by using an example value that represents the set of values with a certain data type and array size.

```
ARGS = cell(4,1);
ARGS{1} = ones(100,1);
ARGS{2} = coder.Constant('Exponential');
ARGS{3} = ones(1,2);
ARGS{4} = ones(10,1);
```

The second input of `myFitandEvaluate` is a distribution name, which is the second input argument of `fitdist`. This argument must be a compile-time constant. Therefore, you specify `ARGS{2}` by using `coder.Constant` (MATLAB Coder).

If you want to specify `ARGS{1}` and `ARGS{3}` as variable-size inputs, use `coder.typeof` (MATLAB Coder). For details, see “General Code Generation Workflow” on page 32-5.

Generate a MEX function from the entry-point function `myFitandEvaluate`. Specify the input argument types using the `-args` option and the cell array `ARGS`.

```
codegen myFitandEvaluate -args ARGS
```

```
Code generation successful.
```

`codegen` (MATLAB Coder) generates the MEX function `myFitandEvaluate_mex` with a platform-dependent extension in your current folder.

Verify Generated Code

Pass some data to verify whether `myFitandEvaluate` and `myFitandEvaluate_mex` return the same outputs.

```
rng('default') % For reproducibility
data = exprnd(1,[100,1]); % Exponential random numbers with mean parameter 1
distname = 'Exponential';
truncation_limits = [0,4];
x = (0:9)';
[pd_truncated,st] = myFitandEvaluate(data,distname,truncation_limits,x);
[pd_truncated_mex,st_mex] = myFitandEvaluate_mex(data,distname,truncation_limits,x);
```

Compare the probability distribution objects `pd_truncated` and `pd_truncated_mex`.

```
pd_truncated
```

```
pd_truncated =
  ExponentialDistribution

  Exponential distribution
    mu = 0.917049
  Truncated to the interval [0, 4]
```

```
pd_truncated_mex
```

```
pd_truncated_mex =
  ExponentialDistribution
```

```
  Exponential distribution
    mu = 0.917049
  Truncated to the interval [0, 4]
```

```
verifyMEX_pd = isequal(pd_truncated,pd_truncated_mex)
```

```
verifyMEX_pd = logical
   1
```

`isequal` returns logical 1 (`true`), which means `pd_truncated` and `pd_truncated_mex` are equal.

Compare the structure arrays that contain the mean, cdf, and pdf values.

```
verifyMEX_st = isequal(st,st_mex)
```

```
verifyMEX_st = logical
   1
```

The comparison confirms that `myFitandEvaluate` and `myFitandEvaluate_mex` return the same outputs. The generated code might not produce the same floating-point numerical results as MATLAB, as described in “Differences Between Generated Code and MATLAB Code” (MATLAB Coder). In that case, compare the values allowing a small tolerance.

Workflow with Two Entry-Point Functions

An entry-point function can have a probability distribution object as both an input argument and an output argument. Therefore, you can define two entry-point functions, one for fitting a distribution and the other for evaluating the fitted distribution. Then generate code for the two entry-point functions.

Define Entry-Point Functions

Define two entry-point functions. The first entry-point function `myFitDist` fits a probability distribution object to the sample data. The second entry-point function `myEvaluateDist` truncates the distribution, computes the mean of the truncated distribution, and computes the cdf and pdf values at the specified data values. `myEvaluateDist` takes the output of `myFitDist` as an input argument.

Display the contents of the `myFitDist.m` and `myEvaluateDist.m` files.

```
type myFitDist.m
```

```
function pd = myFitDist(data,dist) %#codegen
% Fit probability distribution object to data.
pd = fitdist(data,dist);
end
```

```
type myEvaluateDist.m
```

```
function [pd_truncated,st] = myEvaluateDist(pd,truncation_limits,x) %#codegen
% Truncate pd.
pd_truncated = truncate(pd,truncation_limits(1),truncation_limits(2));

% Compute the mean of the truncated pd.
```

```
mean_val = mean(pd_truncated);

% Compute the cdf and pdf, evaluated at x.
cdf_val = cdf(pd_truncated,x);
pdf_val = pdf(pd_truncated,x);

% Create a structure array containing the mean, cdf, and pdf values.
st = struct('mean', mean_val,'cdf',cdf_val,'pdf',pdf_val);
end
```

Generate Code

Specify the input argument types of `myFitDist` and `myEvaluateDist`.

```
ARGS_myFitDist = cell(2,1);
ARGS_myFitDist{1} = ones(100,1);
ARGS_myFitDist{2} = coder.Constant('Exponential');

ARGS_myEvaluateDist = cell(3,1);
ARGS_myEvaluateDist{1} = fitdist(exprnd(1,[100,1]),'Exponential');
ARGS_myEvaluateDist{2} = ones(1,2);
ARGS_myEvaluateDist{3} = ones(10,1);
```

If you do not need to generate a MEX function, then you can specify `ARGS_myEvaluateDist{1}` as `coder.OutputType('myFitdist')`, as described in "Pass an Entry-Point Function Output as an Input" (MATLAB Coder). You cannot use `coder.OutputType` (MATLAB Coder) when generating a MEX function, because the data type of the output from `myFitDist` does not match the data type of the input to `myEvaluateDist` in the generated MEX function.

Generate code for the two entry-point functions.

```
codegen -o myFitandEvaluate_mex2 myFitDist -args ARGS_myFitDist myEvaluateDist -args ARGS_myEvalu

Code generation successful.
```

`codegen` (MATLAB Coder) generates the MEX function `myFitandEvaluate_mex2`. For details about generating code for multiple entry-point functions, see "Generate Code for Multiple Entry-Point Functions" (MATLAB Coder).

Verify Generated Code

Verify the generated code.

```
rng('default')
data = exprnd(1,[100,1]);
distname = 'Exponential';
truncation_limits = [0,4];
x = (0:9)';
pd2 = myFitDist(data,distname);
[pd_truncated2,st2] = myEvaluateDist(pd2,truncation_limits,x);
pd_mex2 = myFitandEvaluate_mex2('myFitDist',data,distname);
[pd_truncated_mex2,st_mex2] = myFitandEvaluate_mex2('myEvaluateDist',pd_mex2,truncation_limits,x
verifyMEX_pd2 = isequal(pd2,pd_mex2)

verifyMEX_pd2 = logical
   1


verifyMEX_pd_truncated2 = isequal(pd_truncated2,pd_truncated_mex)
```

```
verifyMEX_pd_truncated2 = logical
   1
```

```
verifyMEX_st2 = isequal(st2,st_mex2)
```

```
verifyMEX_st2 = logical
   1
```

`isequal` returns logical 1 (`true`), which means that the entry-point functions and the corresponding MEX functions return the same outputs.

See Also

`BetaDistribution` | `ExponentialDistribution` | `ExtremeValueDistribution` | `LognormalDistribution` | `NormalDistribution` | `WeibullDistribution` | `codegen` | `fitdist`

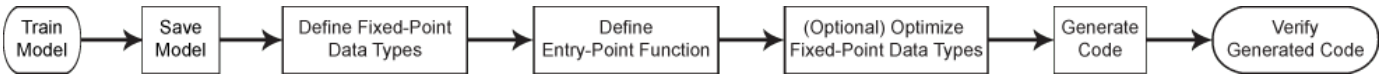
More About

- “Introduction to Code Generation” on page 32-2
- “General Code Generation Workflow” on page 32-5
- Function List (C/C++ Code Generation)

Fixed-Point Code Generation for Prediction of SVM

This example shows how to generate fixed-point C/C++ code for the prediction of a support vector machine (SVM) model. Compared to the general C/C++ code generation workflow, fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. You can also optimize the fixed-point data types before generating code.

This flow chart shows the fixed-point code generation workflow.



1. Train an SVM model.
2. Save the trained model by using `saveLearnerForCoder`.
3. Define the fixed-point data types of the variables required for prediction by using the data type function generated by `generateLearnerDataTypeFcn`.
4. Define an entry-point function that loads the model by using both `loadLearnerForCoder` and the structure, and then calls the `predict` function.
5. (Optional) Optimize the fixed-point data types.
6. Generate fixed-point C/C++ code.
7. Verify the generated code.

Step 5 is an optional step to improve the performance of the generated fixed-point code. To do so, repeat these two steps until you are satisfied with the code performance:

1. Record minimum and maximum values of the variables for prediction by using `buildInstrumentedMex` (Fixed-Point Designer).
2. View the instrumentation results using `showInstrumentationResults` (Fixed-Point Designer). Then, tune the fixed-point data types (if necessary) to prevent overflow and underflow, and to improve the precision of the fixed-point code.

In this workflow, you define the fixed-point data types by using the data type function generated from `generateLearnerDataTypeFcn`. Separating data types of the variables from the algorithm makes testing simpler. You can programmatically toggle data types between floating-point and fixed-point by using the input argument of the data type function. Also, this workflow is compatible with “Manual Fixed-Point Conversion Workflow” (Fixed-Point Designer).

Preprocess Data

Load the `census1994` data set. This data set consists of demographic data from the US Census Bureau used to predict whether an individual makes over $50,000 a year.

```
load census1994
```

Consider a model that predicts the salary category of employees given their age, working class, education level, capital gain and loss, and number of working hours per week. Extract the variables of interest and save them using a table.

```
tbl = adultdata(:,{'age','education_num','capital_gain','capital_loss','hours_per_week'});
```

Print a summary of the table.

```
summary(tbl)
```

```
Variables:

    age: 32561x1 double

        Values:

            Min          17
            Median       37
            Max          90

    education_num: 32561x1 double

        Values:

            Min           1
            Median       10
            Max          16

    capital_gain: 32561x1 double

        Values:

            Min            0
            Median         0
            Max        99999

    capital_loss: 32561x1 double

        Values:

            Min            0
            Median         0
            Max         4356

    hours_per_week: 32561x1 double

        Values:

            Min           1
            Median       40
            Max          99
```

The scales of the variables are not consistent. In this case, you can train a model using a standardized data set by specifying the `'Standardize'` name-value pair argument of `fitcsvm`. However, adding the operations for standardization to the fixed-point code can reduce precision and increase memory use. Instead, you can manually standardize the data set, as shown in this example. The example also describes how to check the memory use at the end.

Fixed-point code generation does not support tables or categorical arrays. So, define the predictor data `X` using a numeric matrix, and define the class labels `Y` using a logical vector. A logical vector uses memory most efficiently in a binary classification problem.

```
X = table2array(tbl);
Y = adultdata.salary == '<=50K';
```

Define the observation weights `w`.

```
w = adultdata.fnlwgt;
```

The memory use of a trained model increases as the number of support vectors in the model increases. To reduce the number of support vectors, you can increase the box constraint when training by using the `'BoxConstraint'` name-value pair argument or use a subsampled representative data set for training. Note that increasing the box constraint can lead to longer training times, and using a subsampled data set can reduce the accuracy of the trained model. In this example, you randomly sample 1000 observations from the data set and use the subsampled data for training.

```
rng('default') % For reproducibility
[X_sampled,idx] = datasample(X,1000,'Replace',false);
Y_sampled = Y(idx);
w_sampled = w(idx);
```

Find the weighted means and standard deviations by training the model using the `'Weight'` and `'Standardize'` name-value pair arguments.

```
tempMdl = fitcsvm(X_sampled,Y_sampled,'Weight',w_sampled,'KernelFunction','gaussian','Standardize
mu = tempMdl.Mu;
sigma = tempMdl.Sigma;
```

If you do not use the `'Cost'`, `'Prior'`, or `'Weight'` name-value pair argument for training, then you can find the mean and standard deviation values by using the `zscore` function.

```
[standardizedX_sampled,mu,sigma] = zscore(X_sampled);
```

Standardize the predictor data by using `mu` and `sigma`.

```
standardizedX = (X-mu)./sigma;
standardizedX_sampled = standardizedX(idx,:);
```

You can use a test data set to validate the trained model and to test an instrumented MEX function. Specify a test data set and standardize the test predictor data by using `mu` and `sigma`.

```
XTest = table2array(adulttest(:,{'age','education_num','capital_gain','capital_loss','hours_per_w
standardizedXTest = (XTest-mu)./sigma;
YTest = adulttest.salary == '<=50K';
```

Train Model

Train a binary SVM classification model.

```
Mdl = fitcsvm(standardizedX_sampled,Y_sampled,'Weight',w_sampled,'KernelFunction','gaussian');
```

`Mdl` is a `ClassificationSVM` model.

Compute the classification error for the training data set and the test data set.

```
loss(Mdl,standardizedX_sampled,Y_sampled)
```

```
ans = 0.1663
```

```
loss(Mdl,standardizedXTest,YTest)
```

```
ans = 0.1905
```

The SVM classifier misclassifies approximately 17% of the training data and 19% of the test data.

Save Model

Save the SVM classification model to the file `myMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'myMdl');
```

Define Fixed-Point Data Types

Use `generateLearnerDataTypeFcn` to generate a function that defines the fixed-point data types of the variables required for prediction of the SVM model. Use all available predictor data to obtain realistic ranges for the fixed-point data types.

```
generateLearnerDataTypeFcn('myMdl',[standardizedX; standardizedXTest])
```

`generateLearnerDataTypeFcn` generates the `myMdl_datatype` function. Display the contents of `myMdl_datatype.m` by using the `type` function.

```
type myMdl_datatype.m
```

```
function T = myMdl_datatype(dt)
%MYMDL_DATATYPE Define data types for fixed-point code generation
%
%   T = MYMDL_DATATYPE(DT) returns the data type structure T, which defines
%   data types for the variables required to generate fixed-point C/C++ code
%   for prediction of a machine learning model. Each field of T contains a
%   fixed-point object returned by fi. The input argument dt specifies the
%   DataType property of the fixed-point object. Specify dt as 'Fixed' (default)
%   for fixed-point code generation or specify dt as 'Double' to simulate
%   floating-point behavior of the fixed-point code.
%
%   Use the output structure T as both an input argument of an entry-point
%   function and the second input argument of loadLearnerForCoder within the
%   entry-point function. For more information, see loadLearnerForCoder.

%   File: myMdl_datatype.m
%   Statistics and Machine Learning Toolbox Version 12.1 (Release R2021a)
%   Generated by MATLAB, 25-Feb-2021 14:04:00

if nargin < 1
    dt = 'Fixed';
end

% Set fixed-point math settings
fm = fimath('RoundingMethod','Floor', ...
    'OverflowAction','Wrap', ...
    'ProductMode','FullPrecision', ...
    'MaxProductWordLength',128, ...
    'SumMode','FullPrecision', ...
    'MaxSumWordLength',128);

% Data type for predictor data
T.XDataType = fi([],true,16,11,fm,'DataType',dt);

% Data type for output score
T.ScoreDataType = fi([],true,16,14,fm,'DataType',dt);
```

```
% Internal variables
% Data type of the squared distance dist = (x-sv)^2 for the Gaussian kernel G(x,sv) = exp(-dist)
% where x is the predictor data for an observation and sv is a support vector
T.InnerProductDataType = fi([],true,16,6,fm,'DataType',dt);


end
```

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function file.

The `myMdl_datatype` function uses the default word length (16) and proposes the maximum fraction length to avoid overflows, based on the default word length (16) and safety margin (10%) for each variable.

Create a structure `T` that defines the fixed-point data types by using `myMdl_datatype`.

```
T = myMdl_datatype('Fixed')
```

```
T = struct with fields:
              XDataType: [0x0 embedded.fi]
           ScoreDataType: [0x0 embedded.fi]
    InnerProductDataType: [0x0 embedded.fi]
```

The structure `T` includes the fields for the named and internal variables required to run the `predict` function. Each field contains a fixed-point object, returned by `fi` (Fixed-Point Designer). For example, display the fixed-point data type properties of the predictor data.

```
T.XDataType
```

```
ans =

[]

          DataTypeMode: Fixed-point: binary point scaling
            Signedness: Signed
            WordLength: 16
        FractionLength: 11

        RoundingMethod: Floor
        OverflowAction: Wrap
           ProductMode: FullPrecision
  MaxProductWordLength: 128
               SumMode: FullPrecision
     MaxSumWordLength: 128
```

For more details about the generated function and the structure, see “Data Type Function” on page 33-2631.

Define Entry-Point Function

Define an entry-point function named `myFixedPointPredict` that does the following:

- Accept the predictor data `X` and the fixed-point data type structure `T`.
- Load a fixed-point version of a trained SVM classification model by using both `loadLearnerForCoder` and the structure `T`.

- Predict labels and scores using the loaded model.

```
function [label,score] = myFixedPointPredict(X,T) %#codegen
Mdl = loadLearnerForCoder('myMdl','DataType',T);
[label,score] = predict(Mdl,X);
end
```

(Optional) Optimize Fixed-Point Data Types

Optimize the fixed-point data types by using `buildInstrumentedMex` and `showInstrumentationResults`. Record minimum and maximum values of all named and internal variables for prediction by using `buildInstrumentedMex`. View the instrumentation results using `showInstrumentationResults`; then, based on the results, tune the fixed-point data type properties of the variables.

Specify Input Argument Types of Entry-Point Function

Specify the input argument types of `myFixedPointPredict` using a 2-by-1 cell array.

```
ARGS = cell(2,1);
```

The first input argument is the predictor data. The `XDataType` field of the structure `T` specifies the fixed-point data type of the predictor data. Convert `X` to the type specified in `T.XDataType` by using the `cast` (Fixed-Point Designer) function.

```
X_fx = cast(standardizedX,'like',T.XDataType);
```

The test data set does not have the same size as the training data set. Specify `ARGS{1}` by using `coder.typeof` (MATLAB Coder) so that the MEX function can take variable-size inputs.

```
ARGS{1} = coder.typeof(X_fx,size(standardizedX),[1,0]);
```

The second input argument is the structure `T`, which must be a compile-time constant. Use `coder.Constant` (MATLAB Coder) to specify `T` as a constant during code generation.

```
ARGS{2} = coder.Constant(T);
```

Create Instrumented MEX Function

Create an instrumented MEX function by using `buildInstrumentedMex` (Fixed-Point Designer).

- Specify the input argument types of the entry-point function by using the `-args` option.
- Specify the MEX function name by using the `-o` option.
- Compute a histogram by using the `-histogram` option.
- Allow full code generation support by using the `-coder` option.

```
buildInstrumentedMex myFixedPointPredict -args ARGS -o myFixedPointPredict_instrumented -histogra

Code generation successful.
```

Test Instrumented MEX Function

Run the instrumented MEX function to record instrumentation results.

```
[labels_fx1,scores_fx1] = myFixedPointPredict_instrumented(X_fx,T);
```

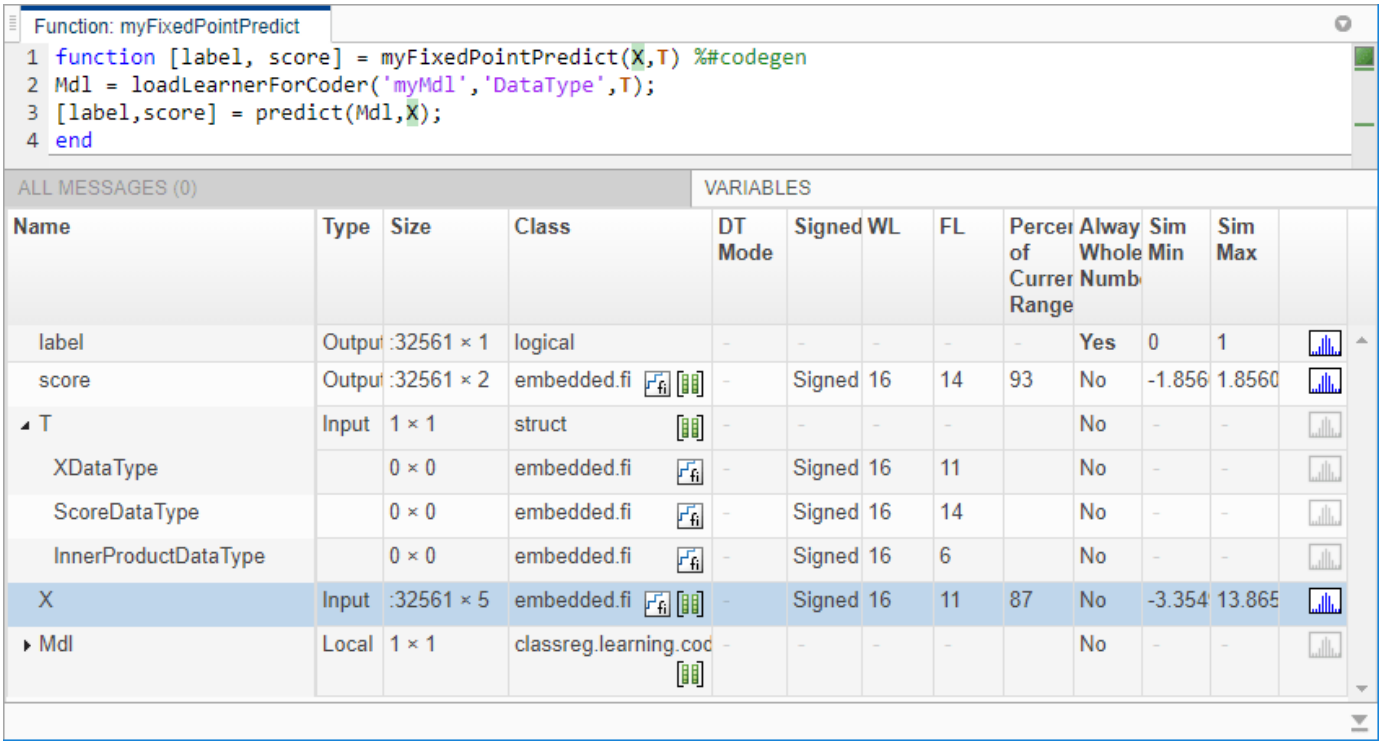
You can run the instrumented MEX function multiple times to record results from various test data sets. Run the instrumented MEX function using `standardizedXTest`.

```
Xtest_fx = cast(standardizedXTest,'like',T.XDataType);
[labels_fx1_test,scores_fx1_test] = myFixedPointPredict_instrumented(Xtest_fx,T);
```

View Results of Instrumented MEX Function

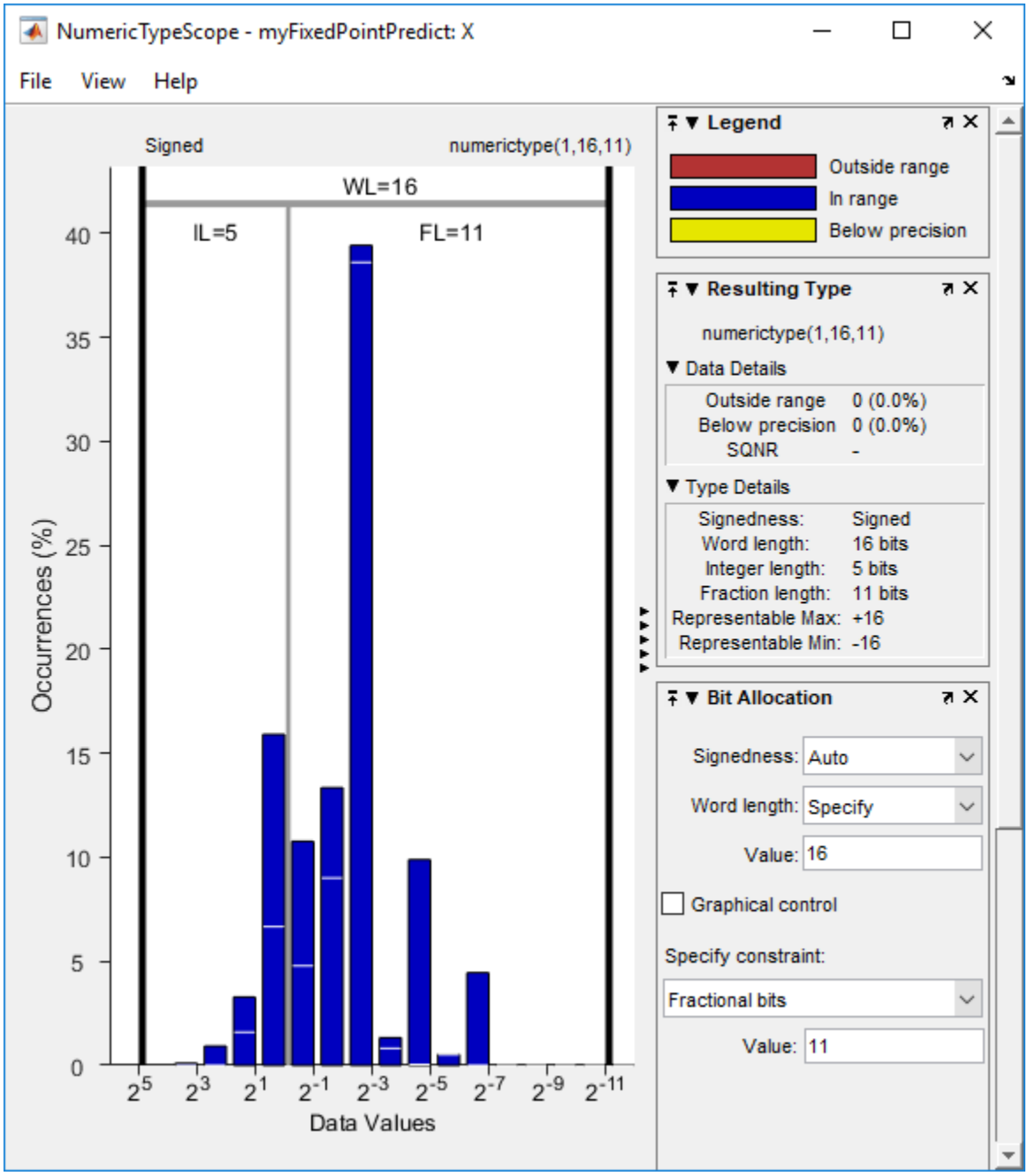
Call `showInstrumentationResults` (Fixed-Point Designer) to open a report containing the instrumentation results. View the simulation minimum and maximum values, proposed fraction length, percent of current range, and whole number status.

```
showInstrumentationResults('myFixedPointPredict_instrumented')
```



The proposed word lengths and fraction lengths in `X` are the same as those in `XDataType` in the structure `T`.

View the histogram for a variable by clicking  on the **Variables** tab.



The window contains the histogram and dialog panels with information about the variable. For information on this window, see the `NumericTypeScope` (Fixed-Point Designer) reference page.

Clear the results by using `clearInstrumentationResults` (Fixed-Point Designer).

```
clearInstrumentationResults('myFixedPointPredict_instrumented')
```

Verify Instrumented MEX Function

Compare the outputs from `predict` and `myFixedPointPredict_instrumented`.

```
[labels,scores] = predict(Mdl,standardizedX);
verify_labels1 = isequal(labels,labels_fx1)
```

```
verify_labels1 = logical
   0
```

`isequal` returns logical 1 (true) if `labels` and `labels_fx1` are equal. If the labels are not equal, you can compute the percentage of incorrectly classified labels as follows.

```
diff_labels1 = sum(strcmp(string(labels_fx1),string(labels))==0)/length(labels_fx1)*100
```

```
diff_labels1 = 0.1228
```

Find the maximum of the relative differences between the score outputs.

```
diff_scores1 = max(abs((scores_fx1.double(:,1)-scores(:,1))./scores(:,1)))
```

```
diff_scores1 = 174.0771
```

Tune Fixed-Point Data Types

You can tune the fixed-point data types if the recorded results show overflow or underflow, or if you want to improve the precision of the generated code. Modify the fixed-point data types by updating the `myMdl_datatype` function and creating a new structure, and then generate the code using the new structure. To update the `myMdl_datatype` function, you can manually modify the fixed-point data types in the function file (`myMdl_datatype.m`). Or, you can generate the function by using `generateLearnerDataTypeFcn` and specifying a longer word length, as shown in this example. For more details, see “Tips” on page 33-2632.

Generate a new data type function. Specify the word length 32 and the name `myMdl_datatype2` for the generated function.

```
generateLearnerDataTypeFcn('myMdl',[standardizedX; standardizedXTest],'WordLength',32,'OutputFun
```

Display the contents of `myMdl_datatype2.m`.

```
type myMdl_datatype2.m
```

```
function T = myMdl_datatype2(dt)
%MYMDL_DATATYPE2 Define data types for fixed-point code generation
%
%   T = MYMDL_DATATYPE2(DT) returns the data type structure T, which defines
%   data types for the variables required to generate fixed-point C/C++ code
%   for prediction of a machine learning model. Each field of T contains a
%   fixed-point object returned by fi. The input argument dt specifies the
%   DataType property of the fixed-point object. Specify dt as 'Fixed' (default)
%   for fixed-point code generation or specify dt as 'Double' to simulate
%   floating-point behavior of the fixed-point code.
%
%   Use the output structure T as both an input argument of an entry-point
%   function and the second input argument of loadLearnerForCoder within the
%   entry-point function. For more information, see loadLearnerForCoder.

%   File: myMdl_datatype2.m
%   Statistics and Machine Learning Toolbox Version 12.1 (Release R2021a)
%   Generated by MATLAB, 25-Feb-2021 14:05:34
```

```
if nargin < 1
    dt = 'Fixed';
end

% Set fixed-point math settings
fm = fimath('RoundingMethod','Floor', ...
    'OverflowAction','Wrap', ...
    'ProductMode','FullPrecision', ...
    'MaxProductWordLength',128, ...
    'SumMode','FullPrecision', ...
    'MaxSumWordLength',128);

% Data type for predictor data
T.XDataType = fi([],true,32,27,fm,'DataType',dt);

% Data type for output score
T.ScoreDataType = fi([],true,32,30,fm,'DataType',dt);

% Internal variables
% Data type of the squared distance dist = (x-sv)^2 for the Gaussian kernel G(x,sv) = exp(-dist)
% where x is the predictor data for an observation and sv is a support vector
T.InnerProductDataType = fi([],true,32,22,fm,'DataType',dt);


end
```

The `myMdl_datatype2` function specifies the word length 32 and proposes the maximum fraction length to avoid overflows.

Create a structure `T2` that defines the fixed-point data types by using `myMdl_datatype2`.

```
T2 = myMdl_datatype2('Fixed')
```

```
T2 = struct with fields:
               XDataType: [0x0 embedded.fi]
          ScoreDataType: [0x0 embedded.fi]
    InnerProductDataType: [0x0 embedded.fi]
```

Create a new instrumented MEX function, record the results, and view the results by using `buildInstrumentedMex` and `showInstrumentationResults`.

```
X_fx2 = cast(standardizedX,'like',T2.XDataType);
buildInstrumentedMex myFixedPointPredict -args {X_fx2,coder.Constant(T2)} -o myFixedPointPredict_
```

```
Code generation successful.
```

```
[labels_fx2,scores_fx2] = myFixedPointPredict_instrumented2(X_fx2,T2);
showInstrumentationResults('myFixedPointPredict_instrumented2')
```

Review the instrumentation report, and then clear the results.

```
clearInstrumentationResults('myFixedPointPredict_instrumented2')
```

Verify `myFixedPointPredict_instrumented2`.

```
verify_labels2 = isequal(labels,labels_fx2)
```

```
verify_labels2 = logical
   0


diff_labels2 = sum(strcmp(string(labels_fx2),string(labels))==0)/length(labels_fx2)*100

diff_labels2 = 0.0031

diff_scores2 = max(abs((scores_fx2.double(:,1)-scores(:,1))./scores(:,1)))

diff_scores2 = 4.5598
```

The percentage of incorrectly classified labels `diff_labels2` and the relative difference in score values `diff_scores2` are smaller than those from the previous MEX function generated using the default word length (16).

For more details about optimizing fixed-point data types by instrumenting MATLAB® code, see the reference pages `buildInstrumentedMex` (Fixed-Point Designer), `showInstrumentationResults` (Fixed-Point Designer), and `clearInstrumentationResults` (Fixed-Point Designer), and the example “Set Data Types Using Min/Max Instrumentation” (Fixed-Point Designer).

Generate Code

Generate code for the entry-point function using `codegen`. Instead of specifying a variable-size input for a predictor data set, specify a fixed-size input by using `coder.typeof`. If you know the size of the predictor data set that you pass to the generated code, then generating code for a fixed-size input is preferable for the simplicity of the code.

```
codegen myFixedPointPredict -args {coder.typeof(X_fx2,[1,5],[0,0]),coder.Constant(T2)}

Code generation successful.
```

`codegen` generates the MEX function `myFixedPointPredict_mex` with a platform-dependent extension.

Verify Generated Code

You can verify the `myFixedPointPredict_mex` function in the same way that you verify the instrumented MEX function. See the **Verify Instrumented MEX Function** section for details.

```
[labels_sampled,scores_sampled] = predict(Mdl,standardizedX_sampled);
n = size(standardizedX_sampled,1);
labels_fx = true(n,1);
scores_fx = zeros(n,2);
for i = 1:n
    [labels_fx(i),scores_fx(i,:)] = myFixedPointPredict_mex(X_fx2(idx(i),:),T2);
end
verify_labels = isequal(labels_sampled,labels_fx)

verify_labels = logical
   1


diff_labels = sum(strcmp(string(labels_fx),string(labels_sampled))==0)/length(labels_fx)*100

diff_labels = 0

diff_scores = max(abs((scores_fx(:,1)-scores_sampled(:,1))./scores_sampled(:,1)))
```

```
diff_scores = 0.0633
```

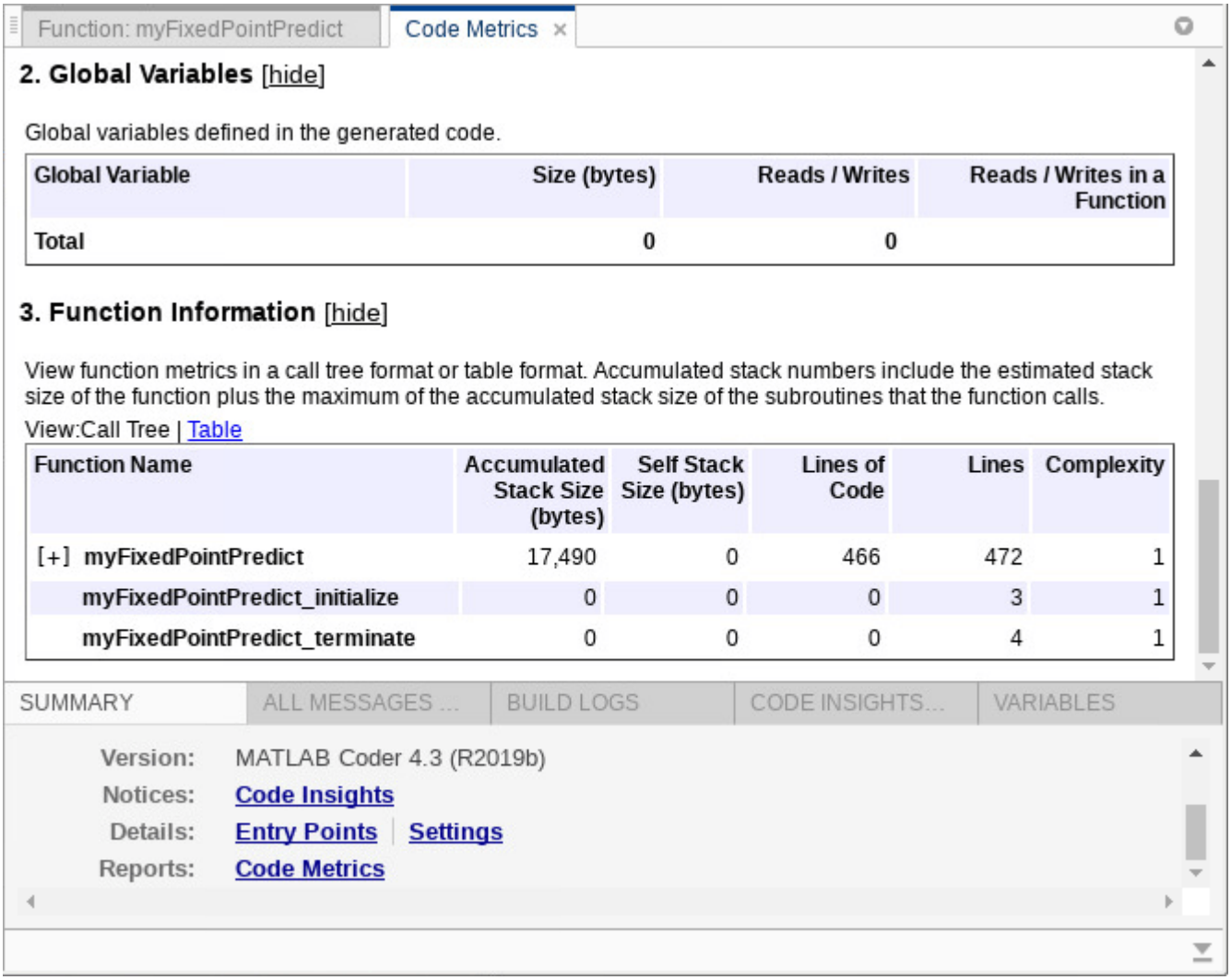
Memory Use

A good practice is to manually standardize predictor data before training a model. If you use the `'Standardize'` name-value pair argument instead, then the generated fixed-point code includes standardization operations, which can cause loss of precision and increased memory use.

If you generate a static library, you can find the memory use of the generated code by using a code generation report. Specify `-config:lib` to generate a static library, and use the `-report` option to generate a code generation report.

```
codegen myFixedPointPredict -args {coder.typeof(X_fx2,[1,5],[0,0]),coder.Constant(T2)} -o myFixed
```

On the **Summary** tab of the code generation report, click **Code Metrics**. The Function Information section shows the accumulated stack size.



To find the memory use of a model trained with `'Standardized','true'`, you can run the following code.

```
Mdl = fitcsvm(X_sampled,Y_sampled,'Weight',w_sampled,'KernelFunction','gaussian','Standardize',t
saveLearnerForCoder(Mdl,'myMdl');
generateLearnerDataTypeFcn('myMdl',[X; XTest],'WordLength',32,'OutputFunctionName','myMdl_standa
T3 = myMdl_standardize_datatype('Fixed');
X_fx3 = cast(X_sampled,'like',T3.XDataType);
codegen myFixedPointPredict -args {coder.typeof(X_fx3,[1,5],[0,0]),coder.Constant(T3)} -o myFixe
```

See Also

`buildInstrumentedMex` | `cast` | `clearInstrumentationResults` | `codegen` | `fi` | `generateLearnerDataTypeFcn` | `loadLearnerForCoder` | `saveLearnerForCoder` | `showInstrumentationResults`

More About

- “Fixed-Point Data Types” (Fixed-Point Designer)
- “Create Fixed-Point Data in MATLAB” (Fixed-Point Designer)
- “Set Data Types Using Min/Max Instrumentation” (Fixed-Point Designer)

Generate Code to Classify Data in Table

This example shows how to generate code for classifying numeric and categorical data in a table using a binary decision tree model. The trained model in this example identifies categorical predictors in the `CategoricalPredictors` property; therefore, the software handles categorical predictors automatically. You do not need to create dummy variables manually for categorical predictors to generate code.

In the general code generation workflow, you can train a classification or regression model on data in a table. You pass arrays (instead of a table) to your entry-point function for prediction, create a table inside the entry-point function, and then pass the table to `predict`. For more information on table support in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

Train Classification Model

Load the `patients` data set. Create a table that contains numeric predictors of type `single` and `double`, categorical predictors of type `categorical`, and the response variable `Smoker` of type `logical`. Each row of the table corresponds to a different patient.

```
load patients
Age = single(Age);
Weight = single(Weight);
Gender = categorical(Gender);
SelfAssessedHealthStatus = categorical(SelfAssessedHealthStatus);
Tbl = table(Age,Diastolic,Systolic,Weight,Gender,SelfAssessedHealthStatus,Smoker);
```

Train a classification tree using the data in `Tbl`.

```
Mdl = fitctree(Tbl,'Smoker')
```

```
Mdl =
  ClassificationTree
             PredictorNames: {1x6 cell}
               ResponseName: 'Smoker'
      CategoricalPredictors: [5 6]
                 ClassNames: [0 1]
             ScoreTransform: 'none'
            NumObservations: 100


  Properties, Methods
```

The `CategoricalPredictors` property value is `[5 6]`, which indicates that `Mdl` identifies the 5th and 6th predictors (`'Gender'` and `'SelfAssessedHealthStatus'`) as categorical predictors. To identify any other predictors as categorical predictors, you can specify them by using the `'CategoricalPredictors'` name-value argument.

Display the predictor names and their order in `Mdl`.

```
Mdl.PredictorNames
```

```
ans = 1x6 cell
  Columns 1 through 5
```

```
    {'Age'}    {'Diastolic'}    {'Systolic'}    {'Weight'}    {'Gender'}

  Column 6

    {'SelfAssessedHe...'}
```

Save Model

Save the tree classifier to a file using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'TreeModel');
```

`saveLearnerForCoder` saves the classifier to the MATLAB® binary file `TreeModel.mat` as a structure array in the current folder.

Define Entry-Point Function

Define the entry-point function `predictSmoker`, which takes predictor variables as input arguments. Within the function, load the tree classifier by using `loadLearnerForCoder`, create a table from the input arguments, and then pass the classifier and table to `predict`.

```
function [labels,scores] = predictSmoker(age,diastolic,systolic,weight,gender,selfAssessedHealthS
%PREDICTSMOKER Label new observations using a trained tree model
%   predictSmoker predicts whether patients are smokers (1) or nonsmokers
%   (0) based on their age, diastolic blood pressure, systolic blood
%   pressure, weight, gender, and self assessed health status. The function
%   also provides classification scores indicating the likelihood that a
%   predicted label comes from a particular class (smoker or nonsmoker).
mdl = loadLearnerForCoder('TreeModel');
varnames = mdl.PredictorNames;
tbl = table(age,diastolic,systolic,weight,gender,selfAssessedHealthStatus, ...
    'VariableNames',varnames);
[labels,scores] = predict(mdl,tbl);
end
```

When you create a table inside an entry-point function, you must specify the variable names (for example, by using the `'VariableNames'` name-value pair argument of `table`). If your table contains only predictor variables, and the predictors are in the same order as in the table used to train the model, then you can find the predictor variable names in `mdl.PredictorNames`.

Generate Code

Generate code for `predictSmoker` by using `codegen`. Specify the data type and dimensions of the predictor variable input arguments using `coder.typeof`.

- The first input argument of `coder.typeof` specifies the data type of the predictor.
- The second input argument specifies the upper bound on the number of rows (`Inf`) and columns (`1`) in the predictor.
- The third input argument specifies that the number of rows in the predictor can change at run time but the number of columns is fixed.

```
ARGS = cell(4,1);
ARGS{1} = coder.typeof(Age,[Inf 1],[1 0]);
ARGS{2} = coder.typeof(Diastolic,[Inf 1],[1 0]);
ARGS{3} = coder.typeof(Systolic,[Inf 1],[1 0]);
ARGS{4} = coder.typeof(Weight,[Inf 1],[1 0]);
```

```
ARGS{5} = coder.typeof(Gender,[Inf 1],[1 0]);
ARGS{6} = coder.typeof(SelfAssessedHealthStatus,[Inf 1],[1 0]);

codegen predictSmoker -args ARGS

Code generation successful.
```

`codegen` generates the MEX function `predictSmoker_mex` with a platform-dependent extension in your current folder.

Verify Generated Code

Verify that `predict`, `predictSmoker`, and the MEX file return the same results for a random sample of 20 patients.

```
rng('default') % For reproducibility
[newTbl,idx] = datasample(Tbl,20);

[labels1,scores1] = predict(Mdl,newTbl);
[labels2,scores2] = predictSmoker(Age(idx),Diastolic(idx),Systolic(idx),Weight(idx),Gender(idx),S
[labels3,scores3] = predictSmoker_mex(Age(idx),Diastolic(idx),Systolic(idx),Weight(idx),Gender(id

verifyMEXlabels = isequal(labels1,labels2,labels3)

verifyMEXlabels = logical
   1


verifyMEXscores = isequal(scores1,scores2,scores3)

verifyMEXscores = logical
   1
```

See Also

`codegen` | `coder.typeof` | `loadLearnerForCoder` | `saveLearnerForCoder`

More About

- “Introduction to Code Generation” on page 32-2
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Code Generation for Tables” (MATLAB Coder)
- “Table Limitations for Code Generation” (MATLAB Coder)

Code Generation for Image Classification

This example shows how to generate C code from a MATLAB® function that classifies images of digits using a trained classification model. This example demonstrates an alternative workflow to “Digit Classification Using HOG Features” (Computer Vision Toolbox). However, to support code generation in that example, you can follow the code generation steps in this example.

Automated image classification is an ubiquitous tool. For example, a trained classifier can be deployed to a drone to automatically identify anomalies on land in captured footage, or to a machine that scans handwritten zip codes on letters. In the latter example, after the machine finds the ZIP code and stores individual images of digits, the deployed classifier must guess which digits are in the images to reconstruct the ZIP code.

This example shows how to train and optimize a multiclass error-correcting output codes (ECOC) classification model to classify digits based on pixel intensities in raster images. The ECOC model contains binary support vector machine (SVM) learners. Then, this example shows how to generate C code that uses the trained model to classify new images. The data are synthetic images of warped digits of various fonts, which simulates handwritten digits.

Set Up Your C Compiler

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder™ locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Assumptions and Limitations

To generate C code, MATLAB Coder:

- Requires a properly configured compiler.
- Requires supported functions to be in a MATLAB function that you define. For the basic workflow, see “Introduction to Code Generation” on page 32-2.
- Forbids objects as input arguments of the defined function.

Concerning the last limitation, consider that:

- Trained classification models are objects
- MATLAB Coder supports `predict` to classify observations using trained models, but does not support fitting the model

To work around the code generation limitations for classification, train the classification model using MATLAB, then pass the resulting model object to `saveLearnerForCoder`. The `saveLearnerForCoder` function removes some properties that are not required for prediction, and then saves the trained model to disk as a structure array. Like the model, the structure array contains the information used to classify new observations.

After saving the model to disk, load the model in the MATLAB function by using `loadLearnerForCoder`. The `loadLearnerForCoder` function loads the saved structure array, and then reconstructs the model object. In the MATLAB function, to classify the observations, you can pass the model and predictor data set, which can be an input argument of the function, to `predict`.

Code Generation for Classification Workflow

Before deploying an image classifier onto a device:

1. Obtain a sufficient amount of labeled images.
2. Decide which features to extract from the images.
3. Train and optimize a classification model. This step includes choosing an appropriate algorithm and tuning hyperparameters, that is, model parameters not fit during training.
4. Save the model to disk by using `saveLearnerForCoder`.
5. Define a function for classifying new images. The function must load the model by using `loadLearnerForCoder`, and can return labels, such as classification scores.
6. Set up your C compiler.
7. Decide the environment in which to execute the generated code.
8. Generate C code for the function.

Load Data

Load the `digitimages` data set.

```
load digitimages
```

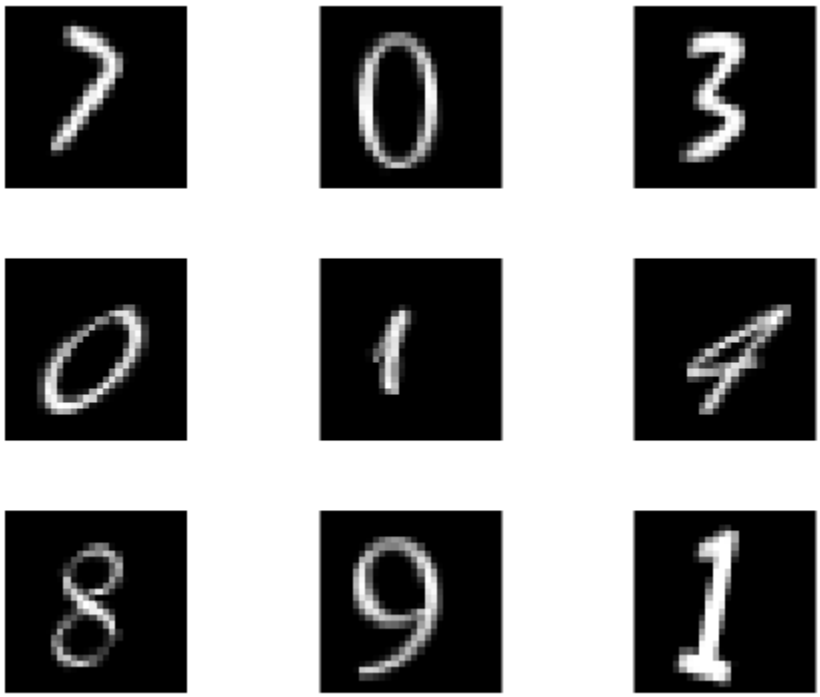
`images` is a 28-by-28-by-3000 array of `uint16` integers. Each page is a raster image of a digit. Each element is a pixel intensity. Corresponding labels are in the 3000-by-1 numeric vector `Y`. For more details, enter `Description` at the command line.

Store the number of observations and number of predictor variables. Create a data partition that specifies to hold out 20% of the data. Extract training and test set indices from the data partition.

```
rng(1) % For reproducibility
n = size(images,3);
p = numel(images(:,:,1));
cvp = cvpartition(n,'Holdout',0.20);
idxTrn = training(cvp);
idxTest = test(cvp);
```

Display nine random images from the data.

```
figure
for j = 1:9
    subplot(3,3,j)
    selectImage = datasample(images,1,3);
    imshow(selectImage,[])
end
```



Rescale Data

Because raw pixel intensities vary widely, you should normalize their values before training a classification model. Rescale the pixel intensities so that they range in the interval [0,1]. That is, suppose p_{ij} is pixel intensity j within image i. For image i, rescale all of its pixel intensities using this formula:

$$\widehat{p}_{ij} = \frac{p_{ij} - \min_j(p_{ij})}{\max_j(p_{ij}) - \min_j(p_{ij})}.$$

```
X = double(images);
for i = 1:n
    minX = min(min(X(:,:,i)));
    maxX = max(max(X(:,:,i)));
    X(:,:,i) = (X(:,:,i) - minX)/(maxX - minX);
end
```

Alternatively, if you have an Image Processing Toolbox™ license, then you can efficiently rescale pixel intensities of images to [0,1] by using `mat2gray`. For more details, see `mat2gray` (Image Processing Toolbox).

Reshape Data

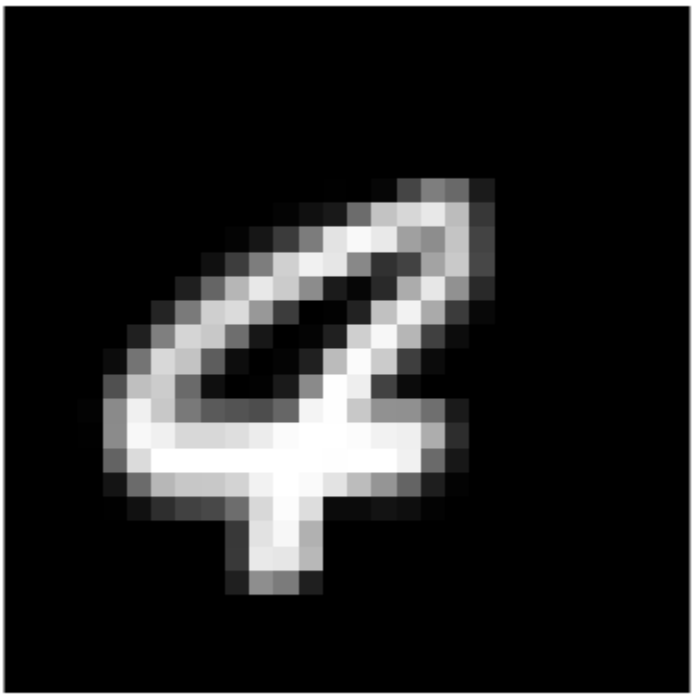
For code generation, the predictor data for training must be in a table of numeric variables or a numeric matrix.

Reshape the data to a matrix such that predictor variables (pixel intensities) correspond to columns, and images (observations) to rows. Because `reshape` takes elements column-wise, you must transpose its result.

```
X = reshape(X,[p,n])';
```

To ensure that preprocessing the data maintains the image, plot the first observation in X.

```
figure
imshow(reshape(X(1,:),sqrt(p)*[1 1]),[],'InitialMagnification','fit')
```



Extract Features

Computer Vision Toolbox™ offers several feature-extraction techniques for images. One such technique is the extraction of histogram of oriented gradient (HOG) features. To learn how to train an ECOC model using HOG features, see “Digit Classification Using HOG Features” (Computer Vision Toolbox). For details on other supported techniques, see “Local Feature Detection and Extraction” (Computer Vision Toolbox). This example uses the rescaled pixel intensities as predictor variables.

Train and Optimize Classification Model

Linear SVM models are often applied to image data sets for classification. However, SVM are binary classifiers, and there are 10 possible classes in the data set.

You can create a multiclass model of multiple binary SVM learners using `fitcecoc`. `fitcecoc` combines multiple binary learners using a coding design. By default, `fitcecoc` applies the one-versus-one design, which specifies training binary learners based on observations from all

combinations of pairs of classes. For example, in a problem with 10 classes, `fitcecoc` must train 45 binary SVM models.

In general, when you train a classification model, you should tune the hyperparameters until you achieve a satisfactory generalization error. That is, you should cross-validate models for particular sets of hyperparameters, and then compare the out-of-fold misclassification rates.

You can choose your own sets of hyperparameter values, or you can specify to implement Bayesian optimization. (For general details on Bayesian optimization, see “Bayesian Optimization Workflow” on page 10-25.) This example performs cross-validation over a chosen grid of values.

To cross-validate an ECOC model of SVM binary learners based on the training observations, use 5-fold cross-validation. Although the predictor values have the same range, to avoid numerical difficulties during training, standardize the predictors. Also, optimize the ECOC coding design and the SVM box constraint. Use all combinations of these values:

- For the ECOC coding design, use one-versus-one and one-versus-all.
- For the SVM box constraint, use three logarithmically-spaced values from 0.1 to 100 each.

For all models, store the 5-fold cross-validated misclassification rates.

```
coding = {'onevsone' 'onevsall'};
boxconstraint = logspace(-1,2,3);
cvLoss = nan(numel(coding),numel(boxconstraint)); % For preallocation

for i = 1:numel(coding)
    for j = 1:numel(boxconstraint)
        t = templateSVM('BoxConstraint',boxconstraint(j),'Standardize',true);
        CVMdl = fitcecoc(X(idxTrn,:),Y(idxTrn),'Learners',t,'KFold',5,...
            'Coding',coding{i});
        cvLoss(i,j) = kfoldLoss(CVMdl);
        fprintf('cvLoss = %f for model using %s coding and box constraint=%f\n',...
            cvLoss(i,j),coding{i},boxconstraint(j))
    end
end
```

```
cvLoss = 0.052083 for model using onevsone coding and box constraint=0.100000
cvLoss = 0.055000 for model using onevsone coding and box constraint=3.162278
cvLoss = 0.050000 for model using onevsone coding and box constraint=100.000000
cvLoss = 0.116667 for model using onevsall coding and box constraint=0.100000
cvLoss = 0.123750 for model using onevsall coding and box constraint=3.162278
cvLoss = 0.125000 for model using onevsall coding and box constraint=100.000000
```

Determine the hyperparameter indices that yield the minimal misclassification rate. Train an ECOC model using the training data. Standardize the training data and supply the observed, optimal hyperparameter combination.

```
minCVLoss = min(cvLoss(:))
```

```
minCVLoss = 0.0500
```

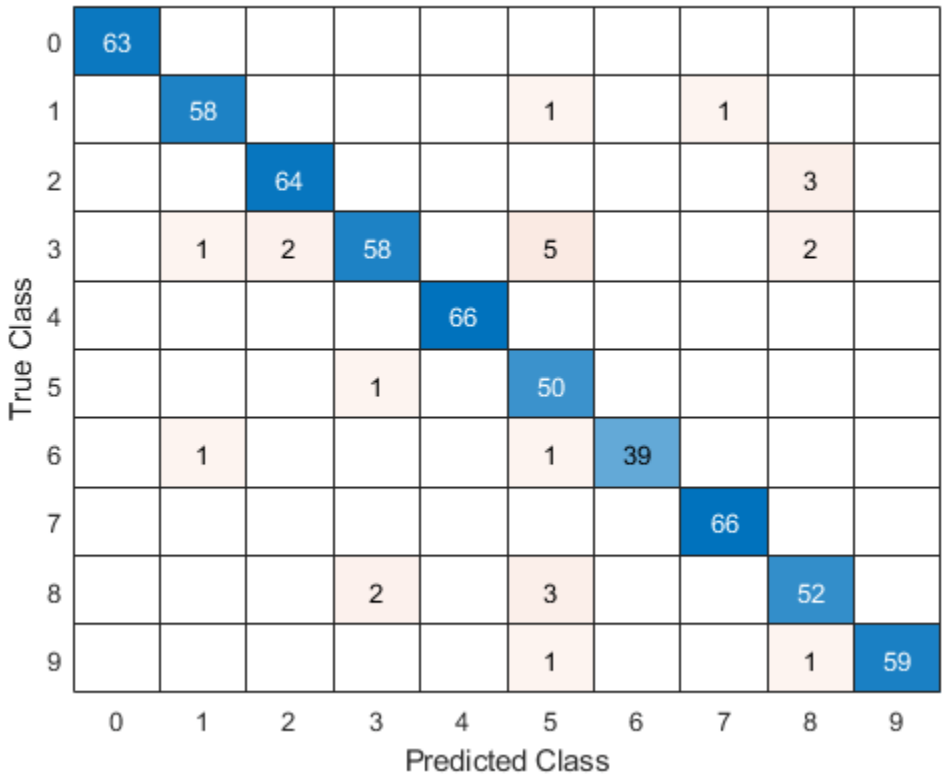
```
linIdx = find(cvLoss == minCVLoss);
[bestI,bestJ] = ind2sub(size(cvLoss),linIdx);
bestCoding = coding{bestI}
```

```
bestCoding =
'onevsone'
```

```
bestBoxConstraint = boxconstraint(bestJ)

bestBoxConstraint = 100

t = templateSVM('BoxConstraint',bestBoxConstraint,'Standardize',true);
Mdl = fitcecoc(X(idxTrn,:),Y(idxTrn),'Learners',t,'Coding',bestCoding);
```

Construct a confusion matrix for the test set images.

```
testImages = X(idxTest,:);
testLabels = predict(Mdl,testImages);
confusionMatrix = confusionchart(Y(idxTest),testLabels);
```



Diagonal and off-diagonal elements correspond to correctly and incorrectly classified observations, respectively. `Mdl` seems to correctly classify most images.

If you are satisfied with the performance of `Mdl`, then you can proceed to generate code for prediction. Otherwise, you can continue adjusting hyperparameters. For example, you can try training the SVM learners using different kernel functions.

Save Classification Model to Disk

`Mdl` is a predictive classification model, but you must prepare it for code generation. Save `Mdl` to your present working directory using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'DigitImagesECOC')
```

`saveLearnerForCoder` compacts `Mdl`, converts it to a structure array, and saves it in the MAT-file `DigitImagesECOC.mat`.

Define Prediction Function for Code Generation

Define an entry-point function named `predictDigitECOC.m` that does the following:

- Include the code generation directive `%#codegen` somewhere in the function.
- Accept image data commensurate with `X`.
- Load `DigitImagesECOC.mat` using `loadLearnerForCoder`.
- Return predicted labels.

```
type predictDigitECOC.m % Display contents of predictDigitECOC.m file
```

```
function label = predictDigitECOC(X) %#codegen
%PREDICTDIGITECOC Classify digit in image using ECOC Model
%   PREDICTDIGITECOC classifies the 28-by-28 images in the rows of X using
%   the compact ECOC model in the file DigitImagesECOC.mat, and then
%   returns class labels in label.
CompactMdl = loadLearnerForCoder('DigitImagesECOC.mat');
label = predict(CompactMdl,X);
end
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB, then MATLAB opens the example folder. This folder includes the entry-point function file.

Verify that the prediction function returns the same test set labels as `predict`.

```
pfLabels = predictDigitECOC(testImages);
verifyPF = isequal(pfLabels,testLabels)
```

```
verifyPF = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal. The `predictDigitECOC` yields the expected results.

Decide Which Environment to Execute Generated Code

Generated code can run:

- Inside the MATLAB environment as a C-MEX file
- Outside the MATLAB environment as a standalone executable
- Outside the MATLAB environment as a shared utility linked to another standalone executable

This example generates a MEX file to be run in the MATLAB environment. Generating such a MEX file allows you to test the generated code using MATLAB tools before deploying the function outside the MATLAB environment. In the MEX function, you can include code for verification, but not for code generation, by declaring the commands as extrinsic using `coder.extrinsic` (MATLAB Coder). Extrinsic commands can include functions that do not have code generation support. All extrinsic commands in the MEX function run in MATLAB, but `codegen` does not generate code for them.

If you plan to deploy the code outside the MATLAB environment, then you must generate a standalone executable. One way to specify your compiler choice is by using the `-config` option of

`codegen`. For example, to generate a static C executable, specify `-config:exe` when you call `codegen`. For more details on setting code generation options, see the `-config` option of `codegen` (MATLAB Coder).

Compile MATLAB Function to MEX File

Compile `predictDigitECOC.m` to a MEX file using `codegen`. Specify these options:

- `-report` — Generates a compilation report that identifies the original MATLAB code and the associated files that `codegen` creates during code generation.
- `-args` — MATLAB Coder requires that you specify the properties of all the function input arguments. One way to do this is to provide `codegen` with an example of input values. Consequently, MATLAB Coder infers the properties from the example values. Specify the test set images commensurate with `X`.

```
codegen predictDigitECOC -report -args {testImages}
```

```
Code generation successful: View report
```

`codegen` successfully generated the code for the prediction function. You can view the report by clicking the `View report` link or by entering `open('codegen/mex/predictDigitECOC/html/report.mldatx')` in the Command Window. If code generation is unsuccessful, then the report can help you debug.

`codegen` creates the directory `pwd/codegen/mex/predictDigitECOC`, where `pwd` is your present working directory. In the child directory, `codegen` generates, among other things, the MEX-file `predictDigitECOC_mex.mexw64`.

Verify that the MEX file returns the same labels as `predict`.

```
mexLabels = predictDigitECOC_mex(testImages);
verifyMEX = isequal(mexLabels,testLabels)
```

```
verifyMEX = logical
   1
```

`isequal` returns logical 1 (`true`), meaning that the MEX-file yields the expected results.

See Also

`codegen` | `loadLearnerForCoder` | `predict` | `saveLearnerForCoder`

Related Examples

- “Introduction to Code Generation” on page 32-2
- “Predict Class Labels Using MATLAB Function Block” on page 32-40
- “System Objects for Classification and Code Generation” on page 32-54
- “Predict Class Labels Using Stateflow” on page 32-62
- “Human Activity Recognition Simulink Model for Smartphone Deployment” on page 32-66
- “Digit Classification Using HOG Features” (Computer Vision Toolbox)

Predict Class Labels Using ClassificationSVM Predict Block

This example shows how to use the ClassificationSVM Predict block for label prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted class label and class score for the observation using the trained support vector machine (SVM) classification model.

Train Classification Model

This example uses the `ionosphere` data set, which contains radar return qualities (`Y`) and predictor data (`X`) of 34 variables. Radar returns are either of good quality (`'g'`) or bad quality (`'b'`).

Load the `ionosphere` data set. Determine the sample size.

```
load ionosphere
n = numel(Y)
```

```
n = 351
```

Suppose that the radar returns are detected in sequence, and you have the first 300 observations, but you have not received the last 51 yet. Partition the data into present and future samples.

```
prsntX = X(1:300,:);
prsntY = Y(1:300);
ftrX = X(301:end,:);
ftrY = Y(301:end);
```

Train an SVM model using all presently available data. Specify predictor data standardization.

```
svmMdl = fitcsvm(prsntX,prsntY,'Standardize',true);
```

`svmMdl` is a `ClassificationSVM` model.

Check the negative and positive class names by using the `ClassNames` property of `svmMdl`.

```
svmMdl.ClassNames
```

```
ans = 2x1 cell
    {'b'}
    {'g'}
```

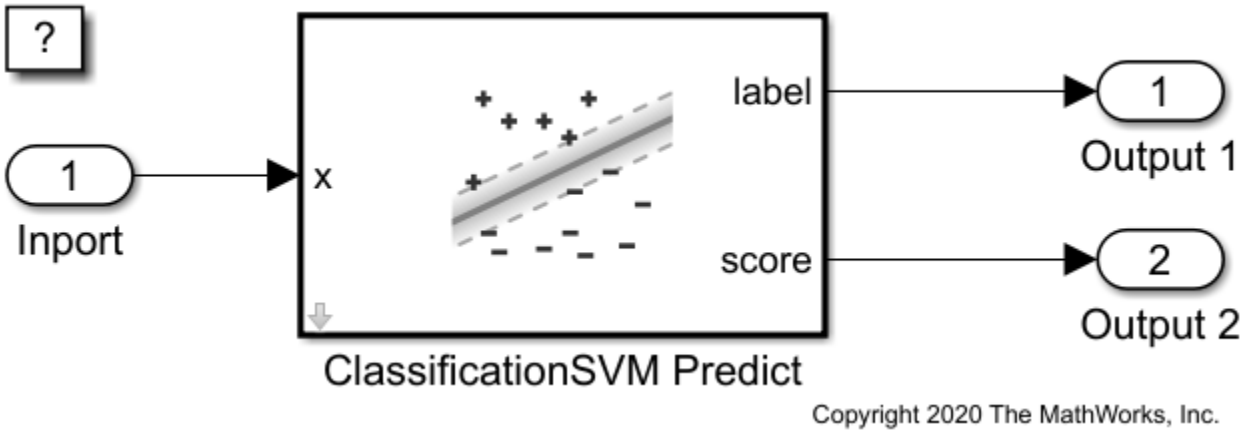
The negative class is `'b'`, and the positive class is `'g'`. The output values from the **score** port of the ClassificationSVM Predict block have the same order. The first and second elements correspond to the negative class and positive class scores, respectively.

Create Simulink Model

This example provides the Simulink model `slexIonosphereClassificationSVMPredictExample.slx`, which includes the ClassificationSVM Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexIonosphereClassificationSVMPredictExample.slx`.

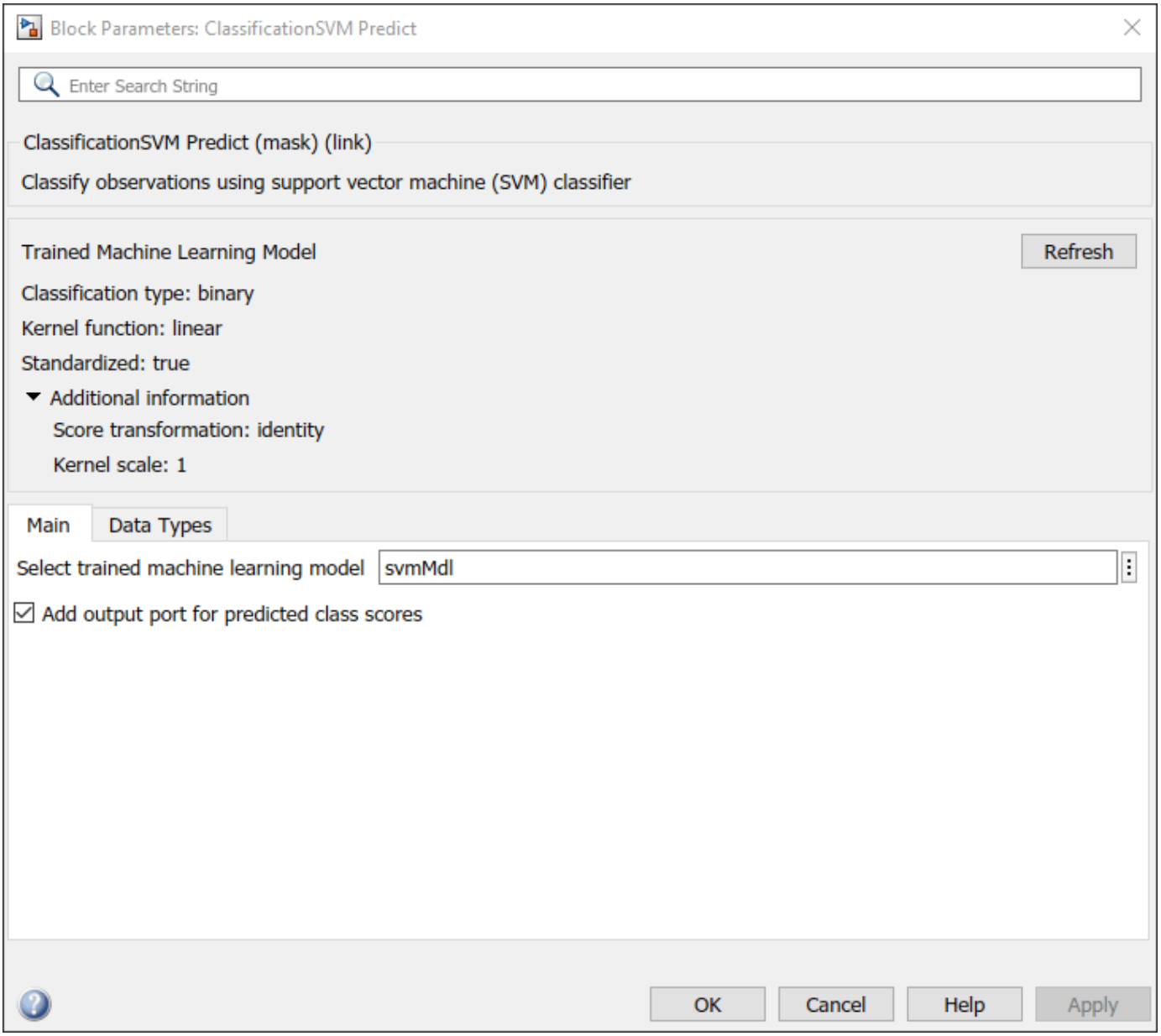
```
SimMdlName = 'slexIonosphereClassificationSVMPredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexIonosphereClassificationSVMPredictExample` includes code to load the sample data, train the SVM model, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the ClassificationSVM Predict block. Add the Inport and Outport blocks and connect them to the ClassificationSVM Predict block.

Double-click the ClassificationSVM Predict block to open the Block Parameters dialog box. Specify the **Select trained machine learning model** parameter as `svmMdl`, which is the name of a workspace variable that contains the trained SVM model. Click the **Refresh** button. The dialog box displays the options used to train the SVM model `svmMdl` under **Trained Machine Learning Model**. Select the **Add output port for predicted class scores** check box to add the second output port **score**.



The ClassificationSVM Predict block expects an observation containing 34 predictor values. Double-click the Inport block, and set the **Port dimensions** to 34 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. In this example, the duration includes the integers from 0 through 50. The orientation must correspond to the observations in the predictor data. So, in this case, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for future radar returns.

```
radarReturnInput.time = (0:50)';
radarReturnInput.signals(1).values = ftrX;
radarReturnInput.signals(1).dimensions = size(ftrX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `carsmallInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `radarReturnInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the ClassificationSVM Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport blocks.

See Also

ClassificationSVM Predict

Related Examples

- “Predict Class Labels Using ClassificationTree Predict Block” on page 32-121
- “Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Predict Responses Using RegressionSVM Predict Block

This example shows how to train a support vector machine (SVM) regression model using the Regression Learner app, and then use the RegressionSVM Predict block for response prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted response for the observation using the trained SVM regression model.

Train Regression Model in Regression Learner App

Train an SVM regression model by using hyperparameter optimization in the Regression Learner App.

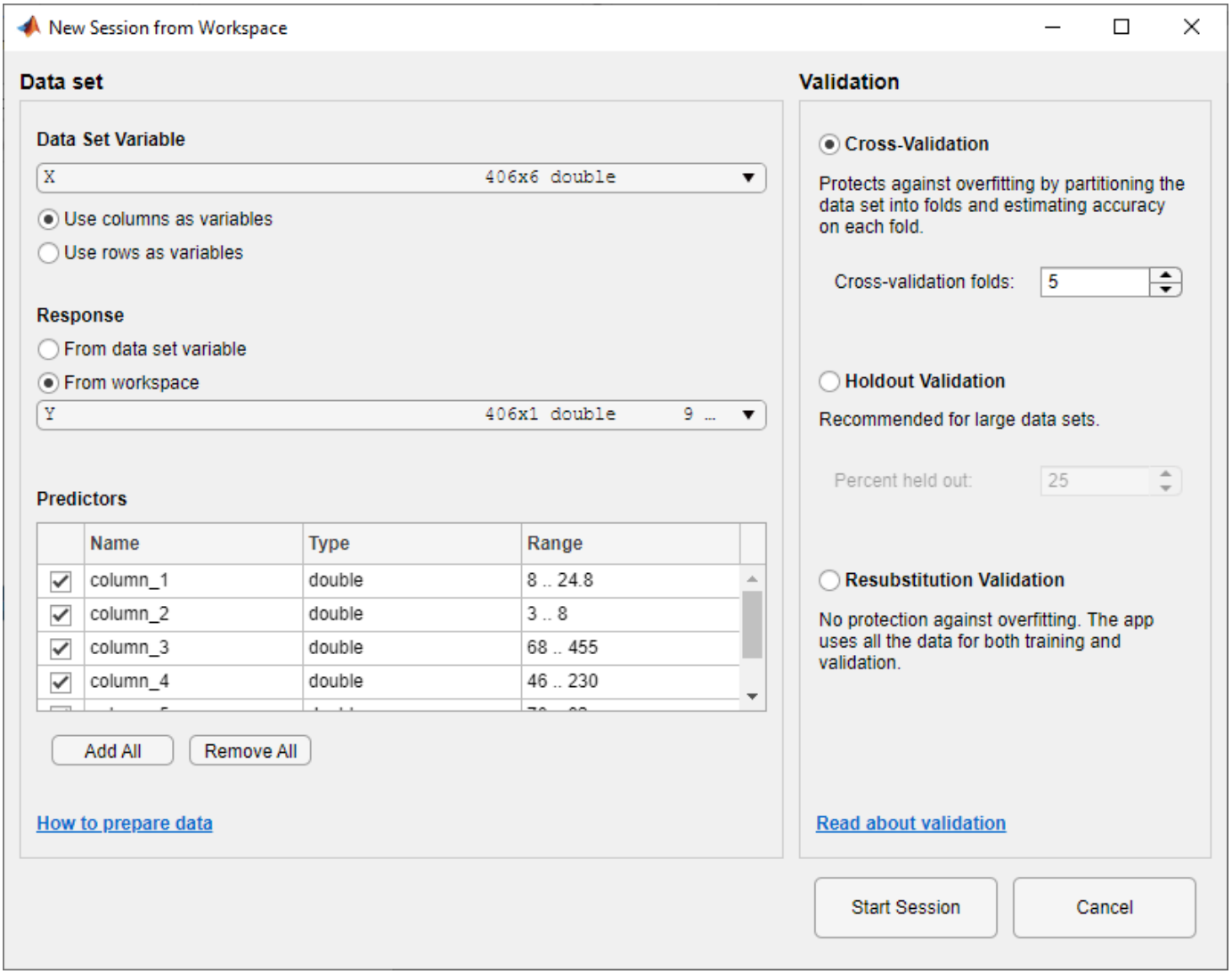
1. In the MATLAB® Command Window, load the `carbig` data set, and create a matrix containing most of the predictor variables and a vector of the response variable.

```
load carbig
X = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight];
Y = MPG;
```

2. Open Regression Learner. On the **Apps** tab, in the **Apps** section, click the **Show more** arrow to display the apps gallery. In the **Machine Learning and Deep Learning** group, click **Regression Learner**.

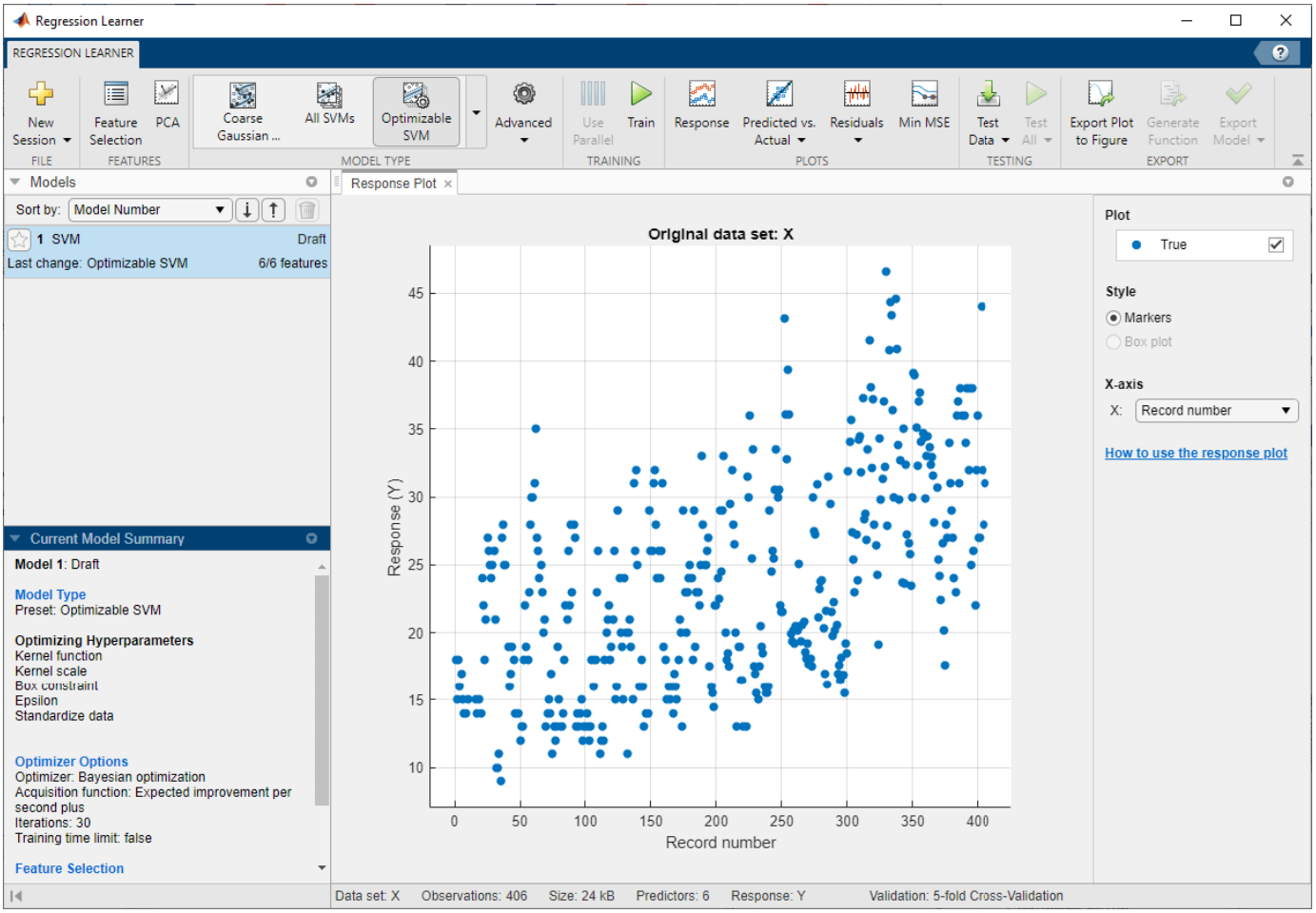
3. On the **Regression Learner** tab, in the **File** section, select **New Session** and select **From Workspace**..

4. In the New Session from Workspace dialog box, select the matrix X from the **Data Set Variable** list. Under **Response**, click the **From workspace** option button and select the vector Y from the workspace. The default validation option is 5-fold cross-validation, to protect against overfitting. For this example, do not change the default settings.



5. To accept the default options and continue, click **Start Session**.

6. Select an optimizable SVM model to train. On the **Regression Learner** tab, in the **Model Type** section, click the **Show more** arrow to open the gallery. In the **Support Vector Machines** group, click **Optimizable SVM**. The app disables the **Use Parallel** button when you select an optimizable model.



7. In the **Training** section, click **Train**. The app displays a **Minimum MSE Plot** as it runs the optimization process. At each iteration, the app tries a different combination of hyperparameter values and updates the plot with the minimum validation mean squared error (MSE) observed up to that iteration, indicated in dark blue. When the app completes the optimization process, it selects the set of optimized hyperparameters, indicated by a red square. For more information, see “Minimum MSE Plot” on page 24-37.

The app lists the optimized hyperparameters in both the **Optimization Results** section to the right of the plot and the **Optimized Hyperparameters** section of the **Current Model Summary** pane. In general, the optimization results are not reproducible.

8. Export the model to the MATLAB workspace. On the **Regression Learner** tab, in the **Export** section, click **Export Model** and select **Export Model**, then click **OK**. The default name for the exported model is `trainedModel`.

Alternatively, you can generate MATLAB code that trains a regression model with the same settings used to train the SVM model in the app. On the **Regression Learner** tab, in the **Export** section, click **Generate Function**. The app generates code from your session and displays the file in the MATLAB Editor. The file defines a function that accepts predictor and response variables, trains a regression model, and performs cross-validation. Change the function name to `trainRegressionSVMModel` and save the function file. Train an SVM model by using the `trainRegressionSVMModel` function.

```
trainedModel = trainRegressionSVMModel(X,Y);
```

9. Extract the trained SVM model from the `trainedModel` variable. `trainedModel` contains a `RegressionSVM` model object in the `RegressionSVM` field.

```
svmMdl = trainedModel.RegressionSVM;
```

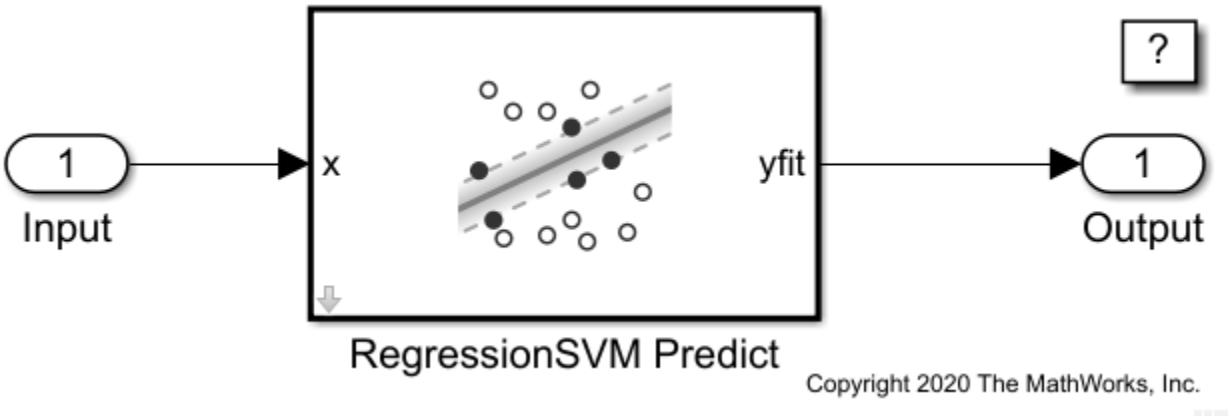
Because hyperparameter optimization can lead to an overfitted model, the recommended approach is to create a separate test set before importing your data into the Regression Learner app and see how the optimized model performs on your test set. For more details, see “Train Regression Model Using Hyperparameter Optimization in Regression Learner App” on page 24-76.

Create Simulink Model

This example provides the Simulink model `slexCarDataRegressionSVMPredictExample.slx`, which includes the RegressionSVM Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexCarDataRegressionSVMPredictExample.slx`.

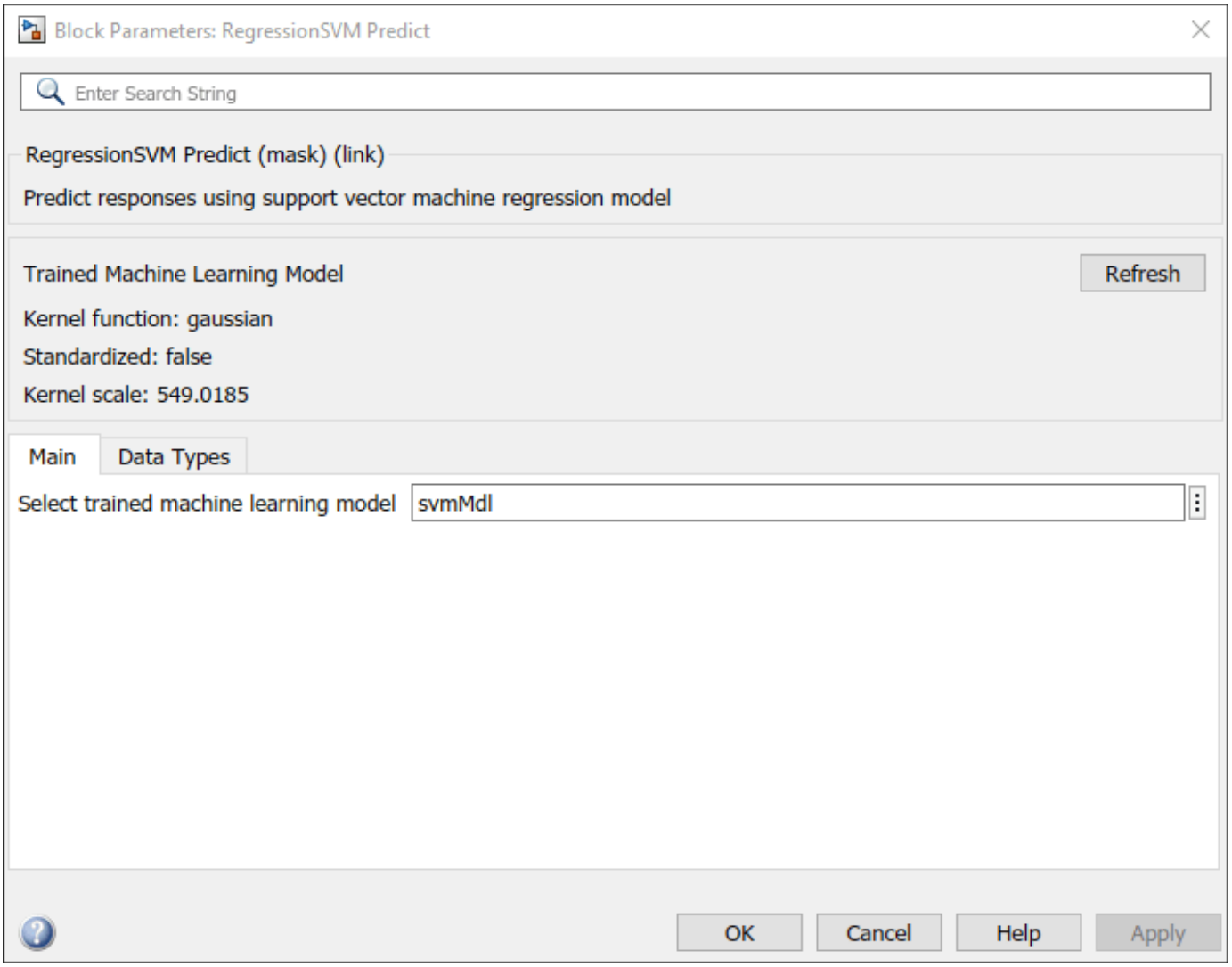
```
SimMdlName = 'slexCarDataRegressionSVMPredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexCarDataRegressionSVMPredictExample` includes code to load the sample data, train the SVM model, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the RegressionSVM Predict block. Add the Inport and Outport blocks and connect them to the RegressionSVM Predict block.

Double-click the RegressionSVM Predict block to open the Block Parameters dialog box. You can specify the name of a workspace variable that contains the trained SVM model. The default variable name is `svmMdl`. Click the **Refresh** button. The dialog box displays the options used to train the SVM model `svmMdl` under **Trained Machine Learning Model**.



The RegressionSVM Predict block expects an observation containing 6 predictor values. Double-click the Inport block, and set the **Port dimensions** to 6 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. The orientation must correspond to the observations in the predictor data. So, in this example, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for the `slexCarDataRegressionSVMPredictExample` model from the `carsmall` data set.

```
load carsmall
testX = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight];
```

```
testX = rmmissing(testX);
carsmallInput.time = (0:size(testX,1)-1)';
carsmallInput.signals(1).values = testX;
carsmallInput.signals(1).dimensions = size(testX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `carsmallInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `carsmallInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the RegressionSVM Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport block.

See Also

RegressionSVM Predict

Related Examples

- “Predict Responses Using RegressionTree Predict Block” on page 32-127
- “Predict Responses Using RegressionEnsemble Predict Block” on page 32-137
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Predict Class Labels Using ClassificationTree Predict Block

This example shows how to train a classification decision tree model using the Classification Learner app, and then use the ClassificationTree Predict block for label prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted class label and class score for the observation using the trained classification decision tree model.

Train Classification Model in Classification Learner App

Train a classification decision tree model by using hyperparameter optimization in the Classification Learner App.

1. In the MATLAB® Command Window, load the `ionosphere` data set, which contains radar return qualities (`Y`) and predictor data (`X`) of 34 variables. Radar returns are either of good quality (`'g'`) or bad quality (`'b'`).

Load the `ionosphere` data set. Determine the sample size.

```
load ionosphere
n = numel(Y)
```

```
n = 351
```

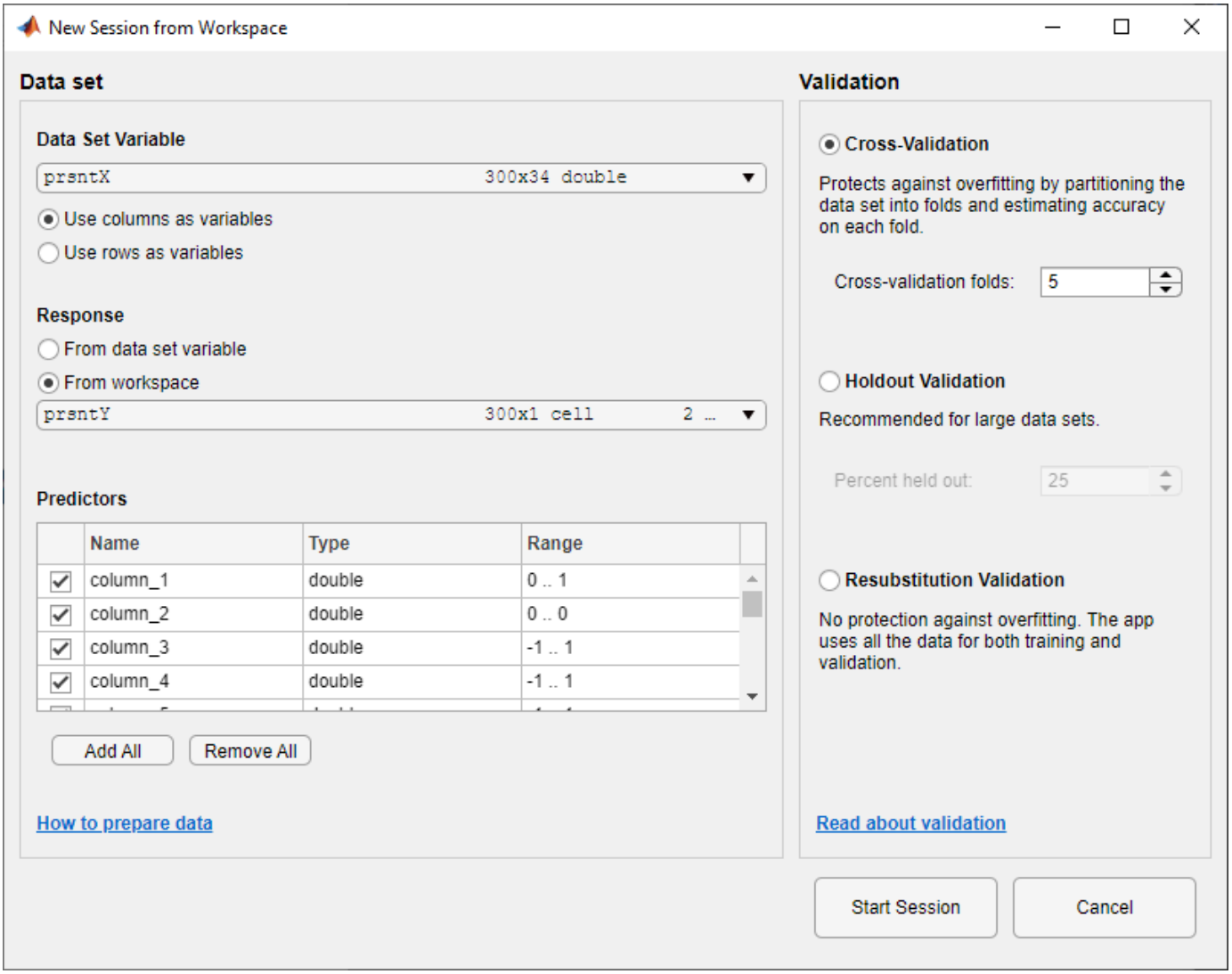
Suppose that the radar returns are detected in sequence, and you have the first 300 observations, but you have not received the last 51 yet. Partition the data into present and future samples.

```
prsntX = X(1:300,:);
prsntY = Y(1:300);
ftrX = X(301:end,:);
ftrY = Y(301:end);
```

2. Open Classification Learner. On the **Apps** tab, in the **Apps** section, click the **Show more** arrow to display the apps gallery. In the **Machine Learning and Deep Learning** group, click **Classification Learner**.

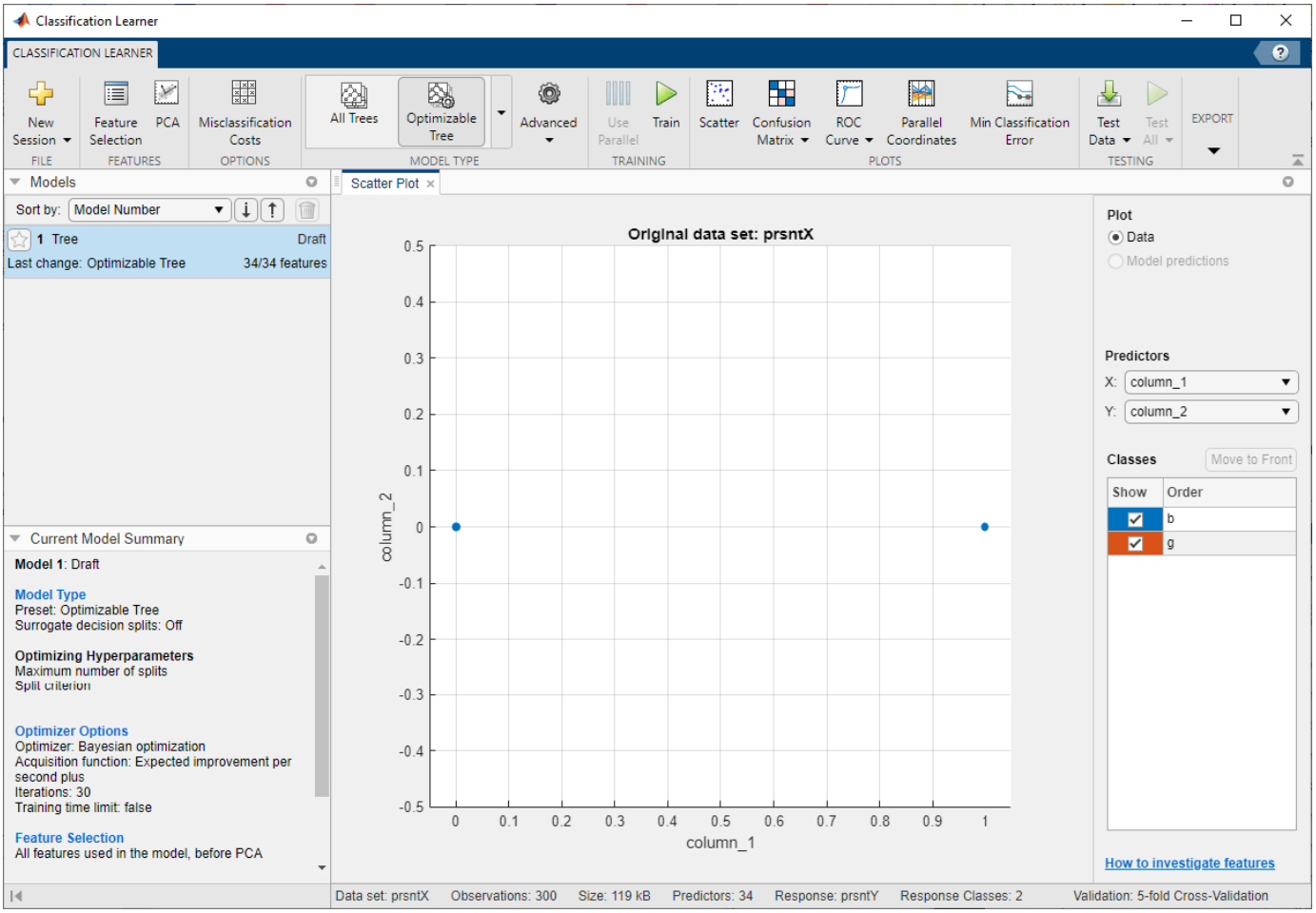
3. On the **Classification Learner** tab, in the **File** section, click **New Session** and select **From Workspace**.

4. In the New Session from Workspace dialog box, select the matrix `prsntX` from the **Data Set Variable** list. Under **Response**, click the **From workspace** option button and select the vector `prsntY` from the workspace. The default validation option is 5-fold cross-validation, to protect against overfitting. For this example, do not change the default settings.



5. To accept the default options and continue, click **Start Session**.

6. Select an optimizable tree model to train. On the **Classification Learner** tab, in the **Model Type** section, click the **Show more** arrow to open the gallery. In the **Decision Trees** group, click **Optimizable Tree**. The app disables the **Use Parallel** button when you select an optimizable model.



7. In the **Training** section, click **Train**. The app displays a **Minimum Classification Error Plot** as it runs the optimization process. At each iteration, the app tries a different combination of hyperparameter values and updates the plot with the minimum validation classification error observed up to that iteration, indicated in dark blue. When the app completes the optimization process, it selects the set of optimized hyperparameters, indicated by a red square. For more information, see “Minimum Classification Error Plot” on page 23-61.

The app lists the optimized hyperparameters in both the **Optimization Results** section to the right of the plot and the **Optimized Hyperparameters** section of the **Current Model Summary** pane. In general, the optimization results are not reproducible.

8. Export the model to the MATLAB workspace. On the **Classification Learner** tab, in the **Export** section, click **Export Model** and select **Export Model**, then click **OK**. The default name for the exported model is `trainedModel`.

Alternatively, you can generate MATLAB code that trains a classification model with the same settings used to train the model in the app. On the **Classification Learner** tab, in the **Export** section, click **Generate Function**. The app generates code from your session and displays the file in the MATLAB Editor. The file defines a function that accepts predictor and response variables, trains a classification model, and performs cross-validation. Change the function name to `trainClassificationTreeModel` and save the function file. Train a decision tree classification model by using the `trainClassificationTreeModel` function.

```
trainedModel = trainClassificationTreeModel(prsntX,prsntY);
```

9. Extract the trained model from the `trainedModel` variable. `trainedModel` contains a `ClassificationTree` model object in the `ClassificationTree` field.

```
treeMdl = trainedModel.ClassificationTree;
```

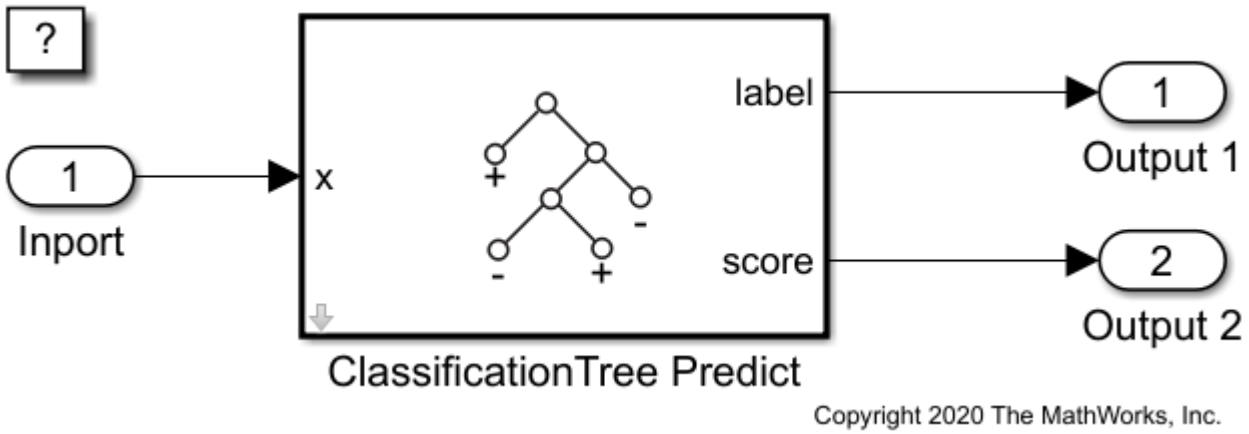
Because hyperparameter optimization can lead to an overfitted model, the recommended approach is to create a separate test set before importing your data into the Classification Learner app and see how the optimized model performs on your test set. For more details, see “Train Classifier Using Hyperparameter Optimization in Classification Learner App” on page 23-128.

Create Simulink Model

This example provides the Simulink model `slexIonosphereClassificationTreePredictExample.slx`, which includes the ClassificationTree Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexIonosphereClassificationTreePredictExample.slx`.

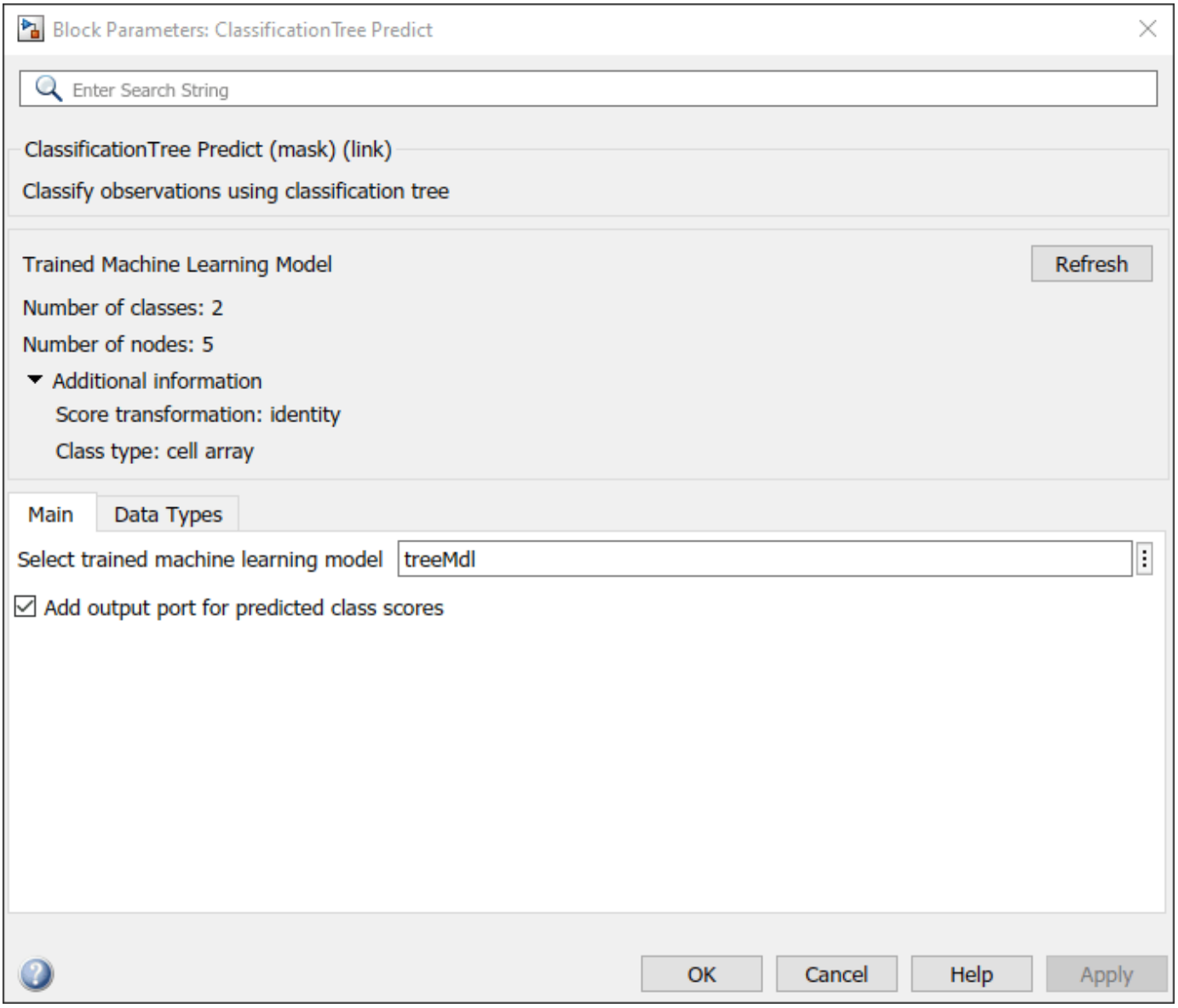
```
SimMdlName = 'slexIonosphereClassificationTreePredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexIonosphereClassificationTreePredictExample` includes code to load the sample data, train the model, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the ClassificationTree Predict block. Add the Inport and Outport blocks and connect them to the ClassificationTree Predict block.

Double-click the ClassificationTree Predict block to open the Block Parameters dialog box. You can specify the name of a workspace variable that contains the trained model. The default variable name is `treeMdl`. Click the **Refresh** button. The dialog box displays the options used to train the model `treeMdl` under **Trained Machine Learning Model**. Select the **Add output port for predicted class scores** check box to add the second output port **score**.



The ClassificationTree Predict block expects an observation containing 34 predictor values. Double-click the Inport block, and set the **Port dimensions** to 34 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. In this example, the duration includes the integers from 0 through 50. The orientation must correspond to the observations in the predictor data. So, in this case, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for future radar returns.

```
radarReturnInput.time = (0:50)';
radarReturnInput.signals(1).values = ftrX;
radarReturnInput.signals(1).dimensions = size(ftrX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `radarReturnInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `radarReturnInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the ClassificationTree Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport blocks.

See Also

ClassificationTree Predict

Related Examples

- “Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
- “Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Predict Responses Using RegressionTree Predict Block

This example shows how to use the RegressionTree Predict block for response prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted response for the observation using the trained regression tree model.

Train Regression Model

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a matrix containing the predictor variables and a vector of the response variable.

```
load carbig
X = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight];
Y = MPG;
```

Train a regression tree model.

```
treeMdl = fitrtree(X,Y);
```

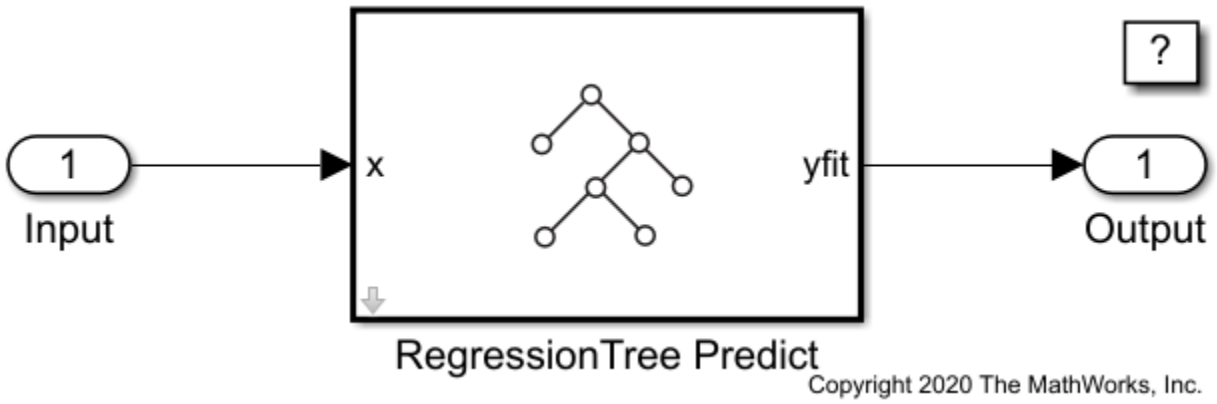
`treeMdl` is a `RegressionTree` model.

Create Simulink Model

This example provides the Simulink model `slexCarDataRegressionTreePredictExample.slx`, which includes the RegressionTree Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexCarDataRegressionTreePredictExample.slx`.

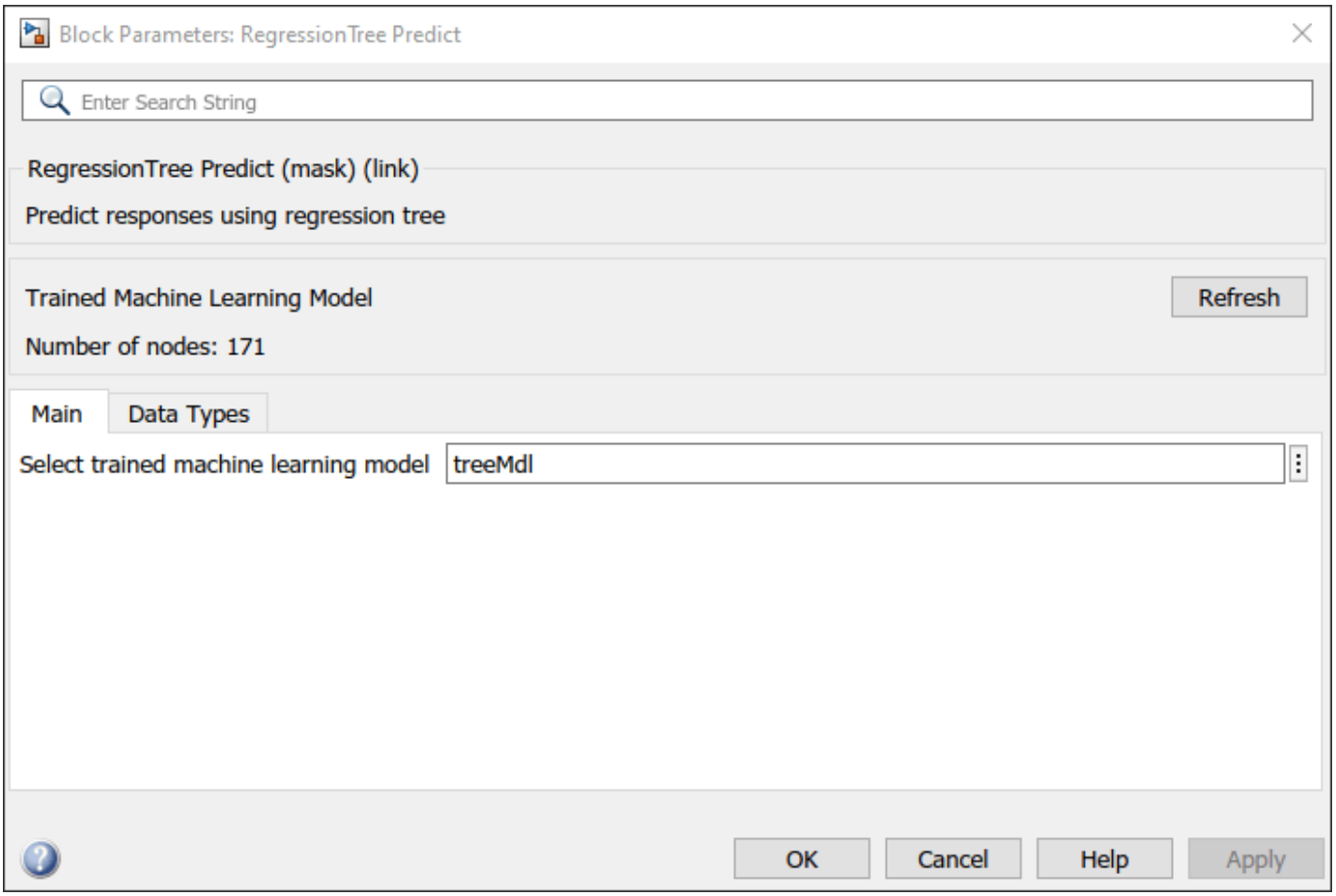
```
SimMdlName = 'slexCarDataRegressionTreePredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexCarDataRegressionTreePredictExample` includes code to load the sample data, train the tree model, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the RegressionTree Predict block. Add the Inport and Outport blocks and connect them to the RegressionTree Predict block.

Double-click the RegressionTree Predict block to open the Block Parameters dialog box. You can specify the name of a workspace variable that contains the trained tree model. The default variable name is `treeMdl`. Click the **Refresh** button. The dialog box displays the options used to train the tree model `treeMdl` under **Trained Machine Learning Model**.



The RegressionTree Predict block expects an observation containing 6 predictor values. Double-click the Inport block, and set the **Port dimensions** to 6 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. The orientation must correspond to the observations in the predictor data. So, in this example, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for the `slexCarDataRegressionTreePredictExample` model from the `carsmall` data set.

```
load carsmall
testX = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight];
testX = rmmissing(testX);
carsmallInput.time = (0:size(testX,1)-1)';
carsmallInput.signals(1).values = testX;
carsmallInput.signals(1).dimensions = size(testX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `carsmallInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `carsmallInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the RegressionTree Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport block.

See Also

RegressionTree Predict

Related Examples

- “Predict Responses Using RegressionSVM Predict Block” on page 32-115
- “Predict Responses Using RegressionEnsemble Predict Block” on page 32-137
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Predict Class Labels Using ClassificationEnsemble Predict Block

This example shows how to train an ensemble model with optimal hyperparameters, and then use the ClassificationEnsemble Predict block for label prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted class label and class score for the observation using the trained classification ensemble model.

Train Classification Model with Optimal Hyperparameters

Load the `CreditRating_Historical` data set. This data set contains customer IDs and their financial ratios, industry labels, and credit ratings. Determine the sample size.

```
tbl = readtable('CreditRating_Historical.dat');
n = numel(tbl)
```

```
n = 31456
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    ______

    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

`tbl.Industry` is a categorical variable for an industry label. When you train a model for the ClassificationEnsemble Predict block, you must preprocess categorical predictors by using the `dummyvar` function to include the categorical predictors in the model. You cannot use the `'CategoricalPredictors'` name-value argument. Create dummy variables for `tbl.Industry`.

```
d = dummyvar(tbl.Industry);
```

`dummyvar` creates dummy variables for each category of `tbl.Industry`. Determine the number of categories in `tbl.Industry` and the number of dummy variables in `d`.

```
unique(tbl.Industry)'
```

```
ans = 1×12

     1     2     3     4     5     6     7     8     9    10    11    12
```

```
size(d)
```

```
ans = 1×2

        3932          12
```

Create a numeric matrix for the predictor variables and a cell array for the response variable.

```
X = [table2array(tbl(:,2:6)) d];
Y = tbl.Rating;
```

X is a numeric matrix that contains 17 variables: the five financial ratios and the 12 dummy variables for the industry label. X does not use `tbl.ID` because the variable is not helpful in predicting credit ratings. Y is a cell array of character vectors that contains the corresponding credit ratings.

Suppose that you receive the data in sequence, and you have the first 3000 observations, but you have not received the last 932 yet. Partition the data into present and future samples.

```
prsntX = X(1:3000,:);
prsntY = Y(1:3000);
ftrX = X(3001:end,:);
ftrY = Y(3001:end);
```

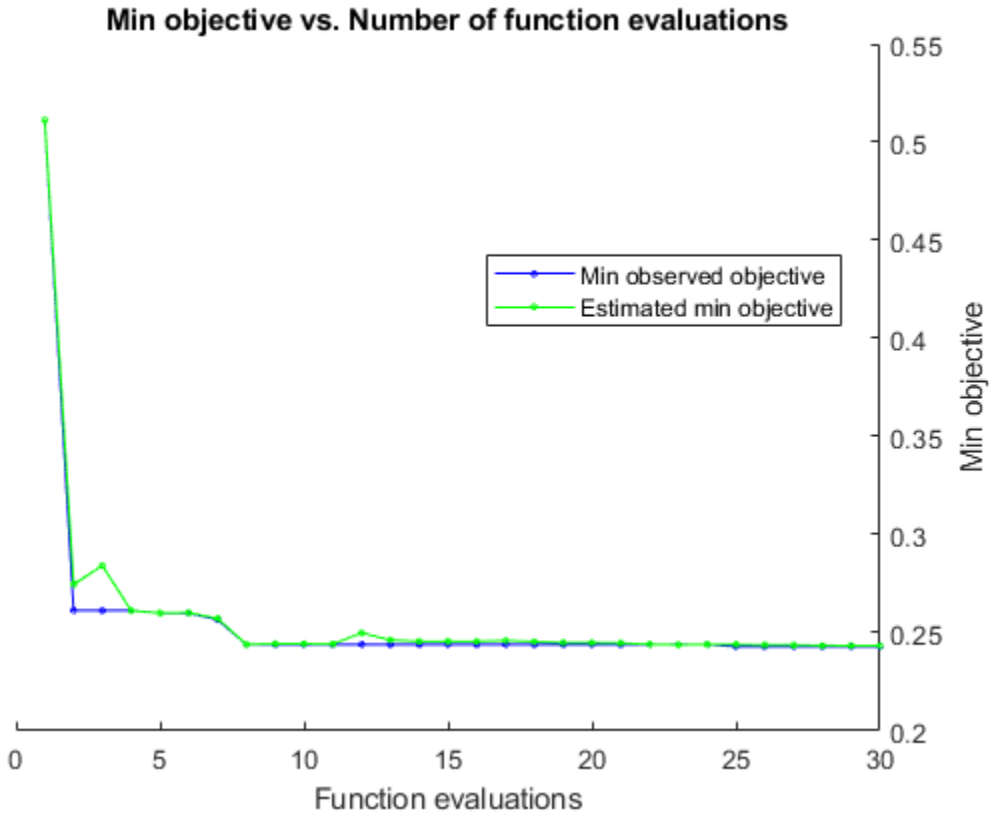
Train an ensemble using all presently available data `prsntX` and `prsntY` with these options:

- Specify `'OptimizeHyperparameters'` as `'auto'` to train an ensemble with optimal hyperparameters. The `'auto'` option finds optimal values for `'Method'`,`'NumLearningCycles'`, and `'LearnRate'` (for applicable methods) of `fitcensemble` and `'MinLeafSize'` of tree learners.
- For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function. Also, for reproducibility of the random forest algorithm, specify `'Reproducible'` as `true` for tree learners.
- Specify the order of the classes by using the `'ClassNames'` name-value argument. The output values from the **score** port of the ClassificationEnsemble Predict block have the same order.

```
rng('default')
t = templateTree('Reproducible',true);
ensMdl = fitcensemble(prsntX,prsntY, ...
    'ClassNames',{'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'}, ...
    'OptimizeHyperparameters','auto','Learners',t, ...
    'HyperparameterOptimizationOptions', ...
    struct('AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result |             | runtime     | (observed)  | (estim.)    |              | ycles
|=====================================================================================================
|    1 | Best   |     0.51133 |      13.652 |     0.51133 |     0.51133 |   AdaBoostM2 |
|    2 | Best   |     0.26133 |      18.827 |     0.26133 |     0.27463 |   AdaBoostM2 |
|    3 | Accept |     0.85133 |     0.76925 |     0.26133 |     0.28421 |     RUSBoost |
|    4 | Accept |       0.263 |     0.61254 |     0.26133 |     0.26124 |   AdaBoostM2 |
|    5 | Best   |        0.26 |      0.9413 |        0.26 |     0.26003 |          Bag |
|    6 | Accept |     0.28933 |      1.7101 |        0.26 |      0.2602 |          Bag |
|    7 | Best   |     0.25667 |      1.3583 |     0.25667 |     0.25726 |   AdaBoostM2 |
|    8 | Best   |       0.244 |      28.725 |       0.244 |     0.24406 |          Bag |
|    9 | Accept |       0.246 |        4.19 |       0.244 |     0.24435 |          Bag |
|   10 | Accept |     0.25533 |      1.3969 |       0.244 |     0.24437 |   AdaBoostM2 |
|   11 | Accept |     0.25733 |      1.5294 |       0.244 |      0.2442 |          Bag |
|   12 | Accept |     0.74267 |      16.444 |       0.244 |     0.24995 |          Bag |
|   13 | Accept |     0.28567 |      7.9382 |       0.244 |     0.24624 |     RUSBoost |
|   14 | Accept |       0.257 |      23.416 |       0.244 |     0.24559 |          Bag |
|   15 | Accept |     0.28433 |     0.71501 |       0.244 |     0.24557 |     RUSBoost |
|   16 | Accept |       0.267 |       17.82 |       0.244 |      0.2456 |   AdaBoostM2 |
|   17 | Accept |     0.24667 |      33.219 |       0.244 |     0.24601 |          Bag |
```

```
|   18 | Best   |       0.244 |      34.953 |       0.244 |      0.2454 |          Bag |
|   19 | Accept |     0.24467 |      31.568 |       0.244 |     0.24489 |          Bag |
|   20 | Accept |       0.259 |      19.187 |       0.244 |     0.24488 |   AdaBoostM2 |
|======================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result |             | runtime     | (observed)  | (estim.)    |              | ycles
|======================================================================================================
|   21 | Accept |     0.27733 |      19.735 |       0.244 |     0.24468 |     RUSBoost |
|   22 | Accept |       0.245 |      32.172 |       0.244 |      0.2441 |          Bag |
|   23 | Accept |       0.244 |      33.117 |       0.244 |     0.24388 |          Bag |
|   24 | Accept |       0.245 |       34.32 |       0.244 |     0.24406 |          Bag |
|   25 | Best   |       0.243 |      33.134 |       0.243 |     0.24394 |          Bag |
|   26 | Accept |     0.25733 |     0.55541 |       0.243 |     0.24371 |   AdaBoostM2 |
|   27 | Accept |       0.263 |     0.52438 |       0.243 |     0.24371 |   AdaBoostM2 |
|   28 | Accept |     0.24367 |      31.167 |       0.243 |     0.24344 |          Bag |
|   29 | Accept |       0.292 |      19.748 |       0.243 |     0.24342 |   AdaBoostM2 |
|   30 | Accept |       0.292 |      0.7854 |       0.243 |     0.24342 |     RUSBoost |
```




```
__________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 488.5833 seconds
Total objective function evaluation time: 464.2275

Best observed feasible point:
    Method    NumLearningCycles    LearnRate    MinLeafSize
```

```
    ______    _________________    _________    ___________

     Bag             499              NaN            5

Observed objective function value = 0.243
Estimated objective function value = 0.24342
Function evaluation time = 33.1343

Best estimated feasible point (according to models):
    Method    NumLearningCycles    LearnRate    MinLeafSize
    ______    _________________    _________    ___________

     Bag             499              NaN            5

Estimated objective function value = 0.24342
Estimated function evaluation time = 32.1002

ensMdl =
  ClassificationBaggedEnsemble
                         ResponseName: 'Y'
                CategoricalPredictors: []
                           ClassNames: {'AAA'  'AA'  'A'  'BBB'  'BB'  'B'  'CCC'}
                       ScoreTransform: 'none'
                      NumObservations: 3000
    HyperparameterOptimizationResults: [1×1 BayesianOptimization]
                             NumTrained: 499
                                 Method: 'Bag'
                          LearnerNames: {'Tree'}
                 ReasonForTermination: 'Terminated normally after completing the requested numbe
                              FitInfo: []
                   FitInfoDescription: 'None'
                            FResample: 1
                              Replace: 1
                     UseObsForLearner: [3000×499 logical]


  Properties, Methods
```

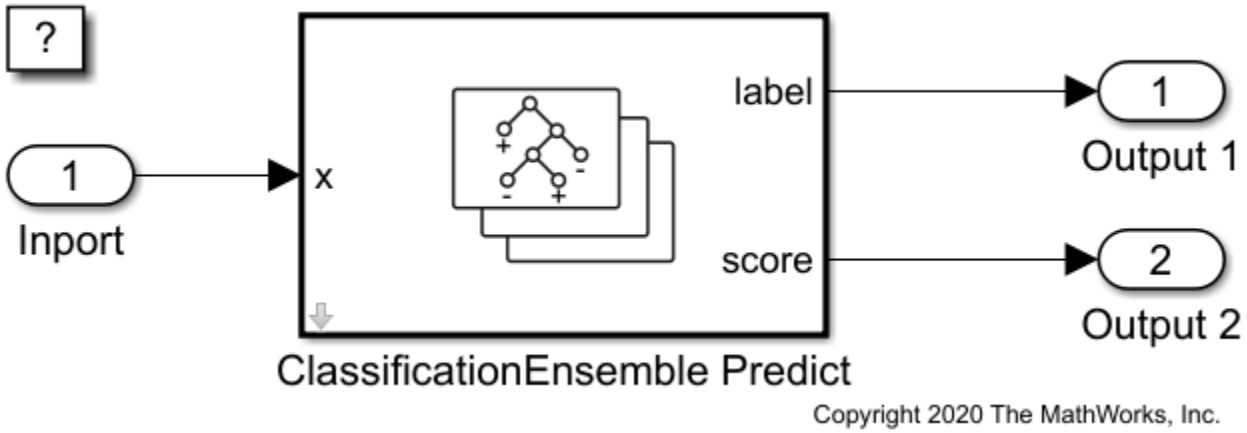
`fitcensemble` returns a `ClassificationBaggedEnsemble` object because the function finds the random forest algorithm (`'Bag'`) as the optimal method.

Create Simulink Model

This example provides the Simulink model `slexCreditRatingClassificationEnsemblePredictExample.slx`, which includes the ClassificationEnsemble Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexCreditRatingClassificationEnsemblePredictExample.slx`.

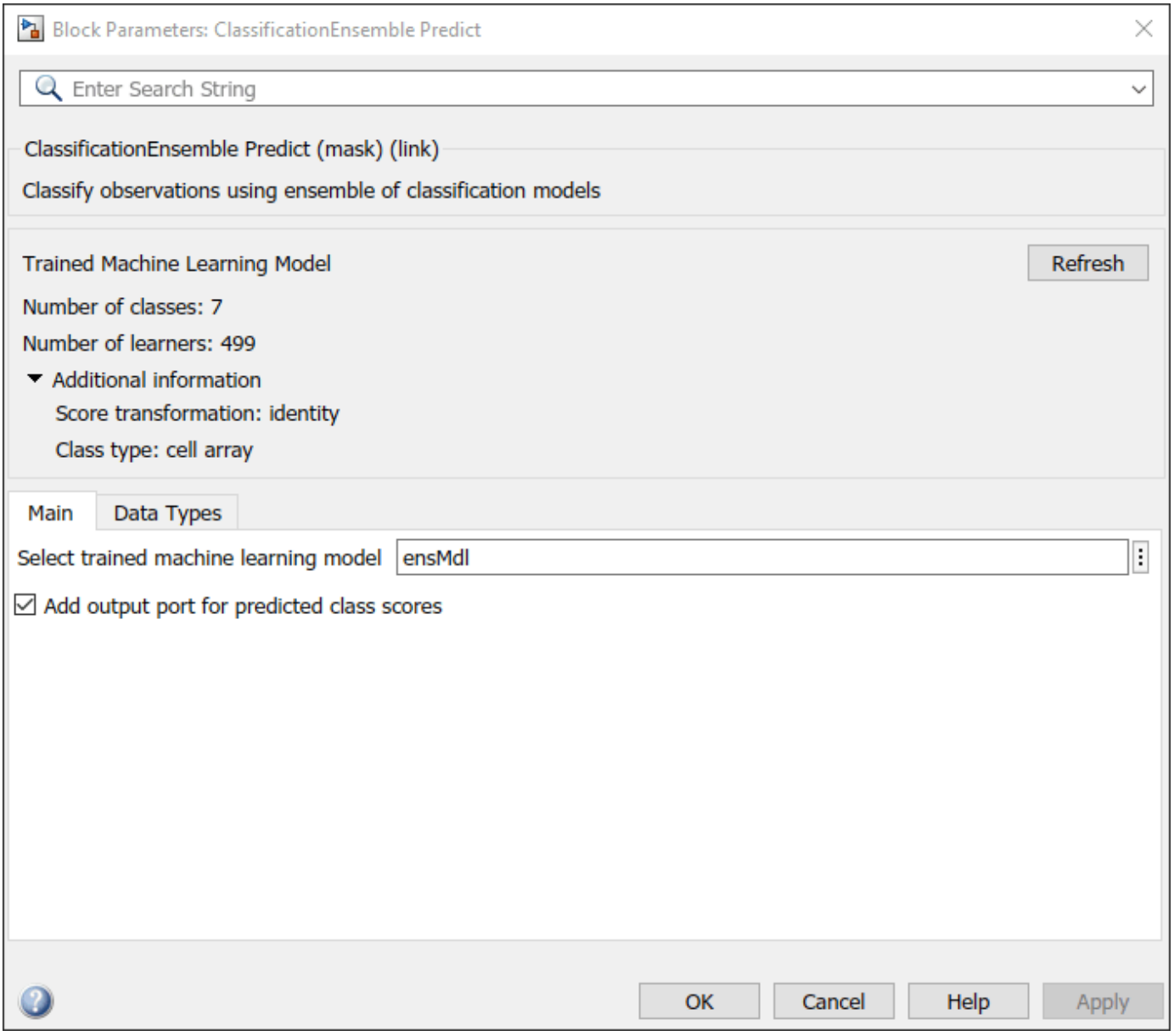
```
SimMdlName = 'slexCreditRatingClassificationEnsemblePredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexCreditRatingClassificationEnsemblePredictExample` includes code to load the sample data, train the model using the optimal hyperparameters, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the ClassificationEnsemble Predict block. Add the Inport and Outport blocks and connect them to the ClassificationEnsemble Predict block.

Double-click the ClassificationEnsemble Predict block to open the Block Parameters dialog box. Specify the **Select trained machine learning model** parameter as `ensMdl`, which is the name of a workspace variable that contains the trained model. Click the **Refresh** button. The dialog box displays the options used to train the model `ensMdl` under **Trained Machine Learning Model**. Select the **Add output port for predicted class scores** check box to add the second output port **score**.



The ClassificationEnsemble Predict block expects an observation containing 17 predictor values. Double-click the Inport block, and set the **Port dimensions** to 17 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. In this example, the duration includes the integers from 0 through 931. The orientation must correspond to the observations in the predictor data. So, in this case, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for future samples.

```
creditRatingInput.time = (0:931)';
creditRatingInput.signals(1).values = ftrX;
creditRatingInput.signals(1).dimensions = size(ftrX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `creditRatingInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `creditRatingInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the ClassificationEnsemble Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport blocks.

See Also

ClassificationEnsemble Predict

Related Examples

- “Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
- “Predict Class Labels Using ClassificationTree Predict Block” on page 32-121
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Predict Responses Using RegressionEnsemble Predict Block

This example shows how to train an ensemble model with optimal hyperparameters, and then use the RegressionEnsemble Predict block for response prediction in Simulink®. The block accepts an observation (predictor data) and returns the predicted response for the observation using the trained regression ensemble model.

Train Regression Model with Optimal Hyperparameters

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
whos
```

```
  Name                Size            Bytes  Class     Attributes

  Acceleration      406x1              3248  double
  Cylinders         406x1              3248  double
  Displacement      406x1              3248  double
  Horsepower        406x1              3248  double
  MPG               406x1              3248  double
  Mfg               406x13            10556  char
  Model             406x36            29232  char
  Model_Year        406x1              3248  double
  Origin            406x7              5684  char
  Weight            406x1              3248  double
  cyl4              406x5              4060  char
  org               406x7              5684  char
  when              406x5              4060  char
```

`Origin` is a categorical variable. When you train a model for the RegressionEnsemble Predict block, you must preprocess categorical predictors by using the `dummyvar` function to include the categorical predictors in the model. You cannot use the `'CategoricalPredictors'` name-value argument. Create dummy variables for `Origin`.

```
c_Origin = categorical(cellstr(Origin));
d_Origin = dummyvar(c_Origin);
```

`dummyvar` creates dummy variables for each category of `c_Origin`. Determine the number of categories in `c_Origin` and the number of dummy variables in `d_Origin`.

```
unique(cellstr(Origin))
```

```
ans = 7x1 cell
    {'England'}
    {'France' }
    {'Germany'}
    {'Italy'  }
    {'Japan'  }
    {'Sweden' }
    {'USA'    }
```

```
size(d_Origin)
```

```
ans = 1×2
```

```
   406     7
```

`dummyvar` creates dummy variables for each category of `Origin`.

Create a matrix containing six numeric predictor variables and the seven dummy variables for `Origin`. Also, create a vector of the response variable.

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,d_Origin];
Y = MPG;
```

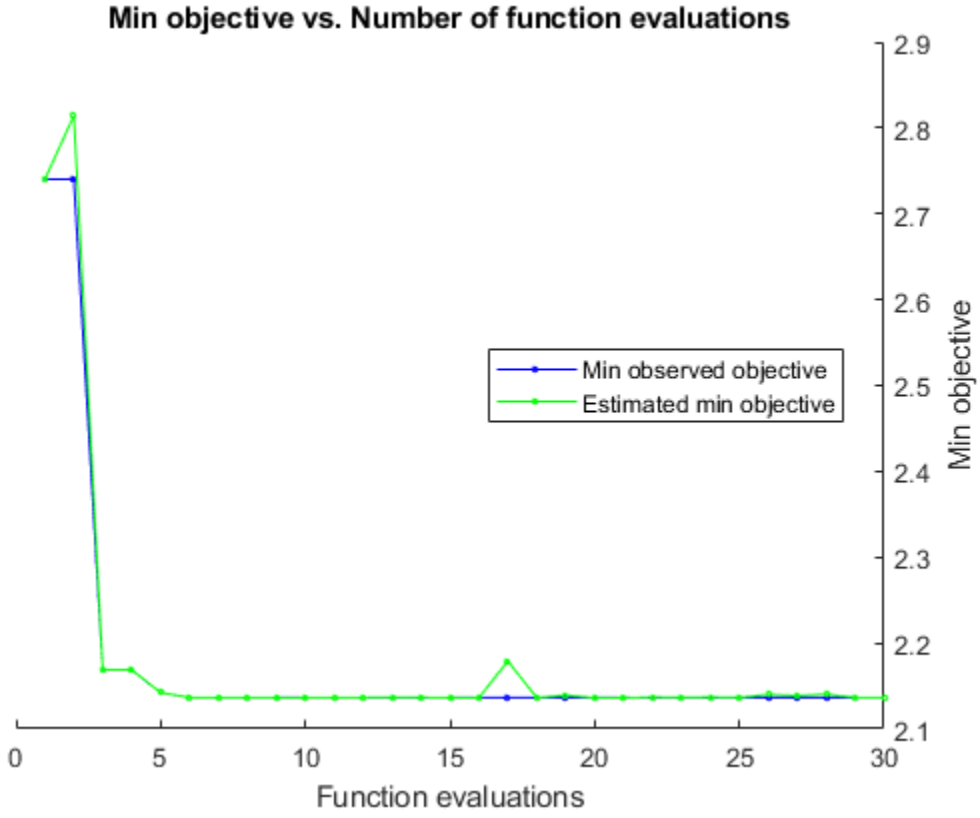
Train an ensemble using `X` and `Y` with these options:

- Specify `'OptimizeHyperparameters'` as `'auto'` to train an ensemble with optimal hyperparameters. The `'auto'` option finds optimal values for `'Method'`,`'NumLearningCycles'`, and `'LearnRate'` (for applicable methods) of `fitrensemble` and `'MinLeafSize'` of tree learners.
- For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function. Also, for reproducibility of the random forest algorithm, specify `'Reproducible'` as `true` for tree learners.

```
rng('default')
t = templateTree('Reproducible',true);
ensMdl = fitrensemble(X,Y,'Learners',t, ...
    'OptimizeHyperparameters','auto', ...
    'HyperparameterOptimizationOptions', ...
    struct('AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              | ycles
|=====================================================================================================
|    1 | Best   |      2.7403 |      10.979 |      2.7403 |      2.7403 |          Bag |
|    2 | Accept |      4.1317 |     0.73016 |      2.7403 |      2.8143 |          Bag |
|    3 | Best   |      2.1687 |      8.2239 |      2.1687 |      2.1689 |          Bag |
|    4 | Accept |      2.2747 |       2.226 |      2.1687 |      2.1688 |      LSBoost |
|    5 | Best   |      2.1421 |      5.2633 |      2.1421 |      2.1422 |          Bag |
|    6 | Best   |      2.1365 |      30.128 |      2.1365 |      2.1365 |          Bag |
|    7 | Accept |      2.4302 |      1.6932 |      2.1365 |      2.1365 |      LSBoost |
|    8 | Accept |      2.1813 |      23.393 |      2.1365 |      2.1365 |      LSBoost |
|    9 | Accept |      6.1992 |       4.788 |      2.1365 |      2.1363 |      LSBoost |
|   10 | Accept |      2.2119 |      24.082 |      2.1365 |      2.1363 |      LSBoost |
|   11 | Accept |      4.7782 |     0.87679 |      2.1365 |      2.1366 |      LSBoost |
|   12 | Accept |      2.3093 |      30.045 |      2.1365 |      2.1366 |      LSBoost |
|   13 | Accept |      4.1304 |      9.2596 |      2.1365 |      2.1366 |      LSBoost |
|   14 | Accept |       2.595 |      1.0223 |      2.1365 |      2.1367 |      LSBoost |
|   15 | Accept |      2.6643 |      1.6063 |      2.1365 |      2.1363 |      LSBoost |
|   16 | Accept |      2.2388 |     0.79398 |      2.1365 |      2.1363 |      LSBoost |
|   17 | Accept |      4.1304 |      1.1643 |      2.1365 |      2.1789 |      LSBoost |
|   18 | Accept |      2.3399 |      4.3167 |      2.1365 |      2.1363 |      LSBoost |
|   19 | Accept |      2.7734 |      6.3456 |      2.1365 |      2.1394 |      LSBoost |
|   20 | Accept |      2.3204 |      31.165 |      2.1365 |       2.136 |          Bag |
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              | ycles
|=====================================================================================================
|   21 | Accept |      2.2146 |      25.613 |      2.1365 |      2.1361 |      LSBoost |
```

```
|   22 | Accept |      4.3329 |     0.92594 |      2.1365 |       2.137 |      LSBoost |
|   23 | Accept |      2.6395 |     0.78224 |      2.1365 |      2.1366 |      LSBoost |
|   24 | Accept |      2.6223 |      27.668 |      2.1365 |       2.137 |      LSBoost |
|   25 | Accept |       2.197 |      13.263 |      2.1365 |      2.1366 |      LSBoost |
|   26 | Accept |      2.4544 |     0.63379 |      2.1365 |      2.1398 |      LSBoost |
|   27 | Accept |      2.1581 |      22.352 |      2.1365 |      2.1386 |      LSBoost |
|   28 | Accept |       2.187 |      28.205 |      2.1365 |      2.1402 |      LSBoost |
|   29 | Accept |      4.1304 |      23.919 |      2.1365 |      2.1366 |      LSBoost |
|   30 | Accept |      2.2396 |      29.916 |      2.1365 |      2.1366 |      LSBoost |
```




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 472.1618 seconds
Total objective function evaluation time: 371.3808

Best observed feasible point:
    Method    NumLearningCycles    LearnRate    MinLeafSize
    ______    _________________    _________    ___________

     Bag              500              NaN            1

Observed objective function value = 2.1365
Estimated objective function value = 2.1366
Function evaluation time = 30.1278
```

```
Best estimated feasible point (according to models):
    Method    NumLearningCycles    LearnRate    MinLeafSize
    ______    _________________    _________    ___________

     Bag             500              NaN            1

Estimated objective function value = 2.1366
Estimated function evaluation time = 31.2023

ensMdl =
  RegressionBaggedEnsemble
                         ResponseName: 'Y'
                CategoricalPredictors: []
                    ResponseTransform: 'none'
                      NumObservations: 398
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                           NumTrained: 500
                               Method: 'Bag'
                         LearnerNames: {'Tree'}
                 ReasonForTermination: 'Terminated normally after completing the requested numbe
                              FitInfo: []
                   FitInfoDescription: 'None'
                       Regularization: []
                            FResample: 1
                              Replace: 1
                     UseObsForLearner: [398x500 logical]


  Properties, Methods
```

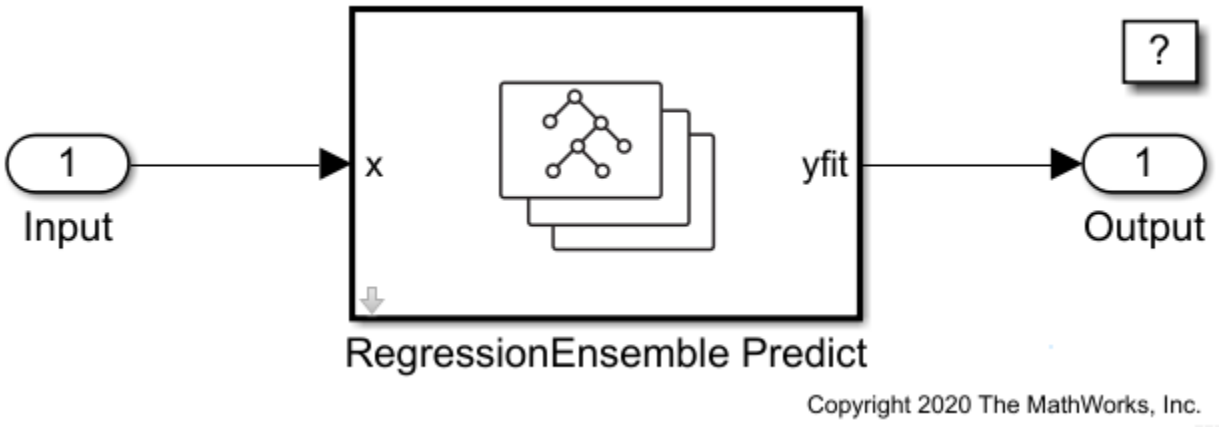
`fitrensemble` returns a `RegressionBaggedEnsemble` object because the function finds the random forest algorithm (`'Bag'`) as the optimal method.

Create Simulink Model

This example provides the Simulink model `slexCarDataRegressionEnsemblePredictExample.slx`, which includes the RegressionEnsemble Predict block. You can open the Simulink model or create a new model as described in this section.

Open the Simulink model `slexCarDataRegressionEnsemblePredictExample.slx`.

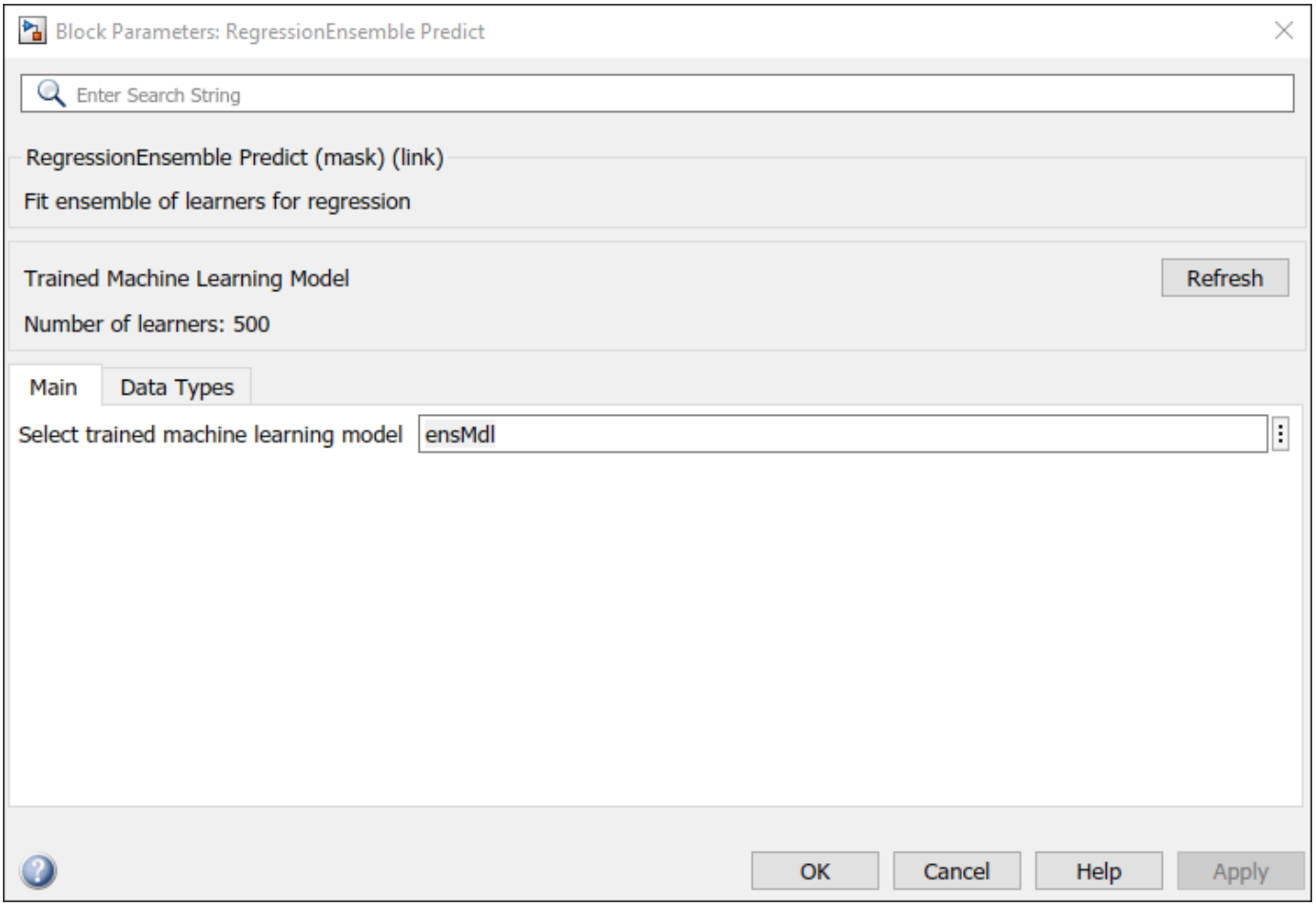
```
SimMdlName = 'slexCarDataRegressionEnsemblePredictExample';
open_system(SimMdlName)
```



The `PreLoadFcn` callback function of `slexCarDataRegressionEnsemblePredictExample` includes code to load the sample data, train the model using the optimal hyperparameters, and create an input signal for the Simulink model. If you open the Simulink model, then the software runs the code in `PreLoadFcn` before loading the Simulink model. To view the callback function, in the **Setup** section on the **Modeling** tab, click **Model Settings** and select **Model Properties**. Then, on the **Callbacks** tab, select the `PreLoadFcn` callback function in the **Model callbacks** pane.

To create a new Simulink model, open the **Blank Model** template and add the RegressionEnsemble Predict block. Add the Inport and Outport blocks and connect them to the RegressionEnsemble Predict block.

Double-click the RegressionEnsemble Predict block to open the Block Parameters dialog box. Specify the **Select trained machine learning model** parameter as `ensMdl`, which is the name of a workspace variable that contains the trained model. Click the **Refresh** button. The dialog box displays the options used to train the model `ensMdl` under **Trained Machine Learning Model**.



The RegressionEnsemble Predict block expects an observation containing 13 predictor values. Double-click the Inport block, and set the **Port dimensions** to 13 on the **Signal Attributes** tab.

Create an input signal in the form of a structure array for the Simulink model. The structure array must contain these fields:

- `time` — The points in time at which the observations enter the model. The orientation must correspond to the observations in the predictor data. So, in this example, `time` must be a column vector.
- `signals` — A 1-by-1 structure array describing the input data and containing the fields `values` and `dimensions`, where `values` is a matrix of predictor data, and `dimensions` is the number of predictor variables.

Create an appropriate structure array for the `slexCarDataRegressionEnsemblePredictExample` model from the `carsmall` data set. When you convert `Origin` in `carsmall` to the `categorical` data type array `c_Origin_small`, use `categories(c_Origin)` so that `c_Origin` and `c_Origin_small` have the same number of categories in the same order.

```
load carsmall
c_Origin_small = categorical(cellstr(Origin),categories(c_Origin));
d_Origin_small = dummyvar(c_Origin_small);
```

```
testX = [Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,d_Origin_small];
testX = rmmissing(testX);
carsmallInput.time = (0:size(testX,1)-1)';
carsmallInput.signals(1).values = testX;
carsmallInput.signals(1).dimensions = size(testX,2);
```

To import signal data from the workspace:

- Open the Configuration Parameters dialog box. On the **Modeling** tab, click **Model Settings**.
- In the **Data Import/Export** pane, select the **Input** check box and enter `carsmallInput` in the adjacent text box.
- In the **Solver** pane, under **Simulation time**, set **Stop time** to `carsmallInput.time(end)`. Under **Solver selection**, set **Type** to `Fixed-step`, and set **Solver** to `discrete (no continuous states)`.

For more details, see “Load Signal Data for Simulation” (Simulink).

Simulate the model.

```
sim(SimMdlName);
```

When the Inport block detects an observation, it directs the observation into the RegressionTree Predict block. You can use the Simulation Data Inspector (Simulink) to view the logged data of the Outport block.

See Also

RegressionEnsemble Predict

Related Examples

- “Predict Responses Using RegressionSVM Predict Block” on page 32-115
- “Predict Responses Using RegressionTree Predict Block” on page 32-127
- “Predict Class Labels Using MATLAB Function Block” on page 32-40

Code Generation for Logistic Regression Model Trained in Classification Learner

This example shows how to train a logistic regression model using Classification Learner, and then generate C code that predicts labels using the exported classification model.

Load Sample Data

Load sample data and import the data into the Classification Learner app.

Load the `patients` data set. Specify the predictor data `X`, consisting of `p` predictors, and the response variable `Y`.

```
load patients
X = [Age Diastolic Height Systolic Weight];
p = size(X,2);
Y = Gender;
```

On the **Apps** tab, click the **Show more** arrow at the right of the Apps section to display the gallery, and select **Classification Learner**. On the **Classification Learner** tab, in the **File** section, select **New Session > From Workspace**.

In the New Session from Workspace dialog box, under **Data Set Variable**, select `X` from the list of workspace variables. Under **Response**, click the **From workspace** option button and then select `Y` from the list. To accept the default validation scheme and continue, click **Start Session**. The default validation option is 5-fold cross-validation, to protect against overfitting.

By default, Classification Learner creates a scatter plot of the data.

Train Logistic Regression Model

Train a logistic regression model within the Classification Learner app.

On the **Classification Learner** tab, in the **Model Type** section, click the **Show more** arrow to display the gallery of classifiers. Under **Logistic Regression Classifiers**, click the **Logistic Regression** model. Click **Train** in the **Training** section. The app trains the model and displays its cross-validation accuracy score **Accuracy (Validation)**.

Export Model to Workspace

Export the model to the MATLAB® Workspace and save it using `saveLearnerForCoder`.

In the **Export** section, select **Export Model > Export Compact Model**. Click **OK** in the dialog box.

The structure `trainedModel` appears in the MATLAB Workspace. The field `GeneralizedLinearModel` of `trainedModel` contains the compact model.

Note: If you run this example with all supporting files, you can load the `trainedModel.mat` file at the command line rather than exporting the model. The `trainedModel` structure was created using the previous steps.

```
load('trainedModel.mat')
```

At the command line, save the compact model to a file named `myModel.mat` in your current folder.

```
saveLearnerForCoder(trainedModel.GeneralizedLinearModel,'myModel')
```

Additionally, save the names of the success, failure, and missing classes of the trained model.

```
classNames = {trainedModel.SuccessClass, ...
    trainedModel.FailureClass,trainedModel.MissingClass};
save('ModelParameters.mat','classNames');
```

Generate C Code for Prediction

Define the entry-point function for prediction, and generate code for the function by using `codegen`.

In your current folder, define a function named `classifyX.m` that does the following:

- Accepts a numeric matrix (`X`) of observations containing the same predictor variables as the ones used to train the logistic regression model
- Loads the classification model in `myModel.mat`
- Computes predicted probabilities using the model
- Converts the predicted probabilities to indices, where 1 indicates a success, 2 indicates a failure, and 3 indicates a missing value
- Loads the class names in `ModelParameters.mat`
- Returns predicted labels by indexing into the class names

```
function label = classifyX (X) %#codegen
%CLASSIFYX Classify using Logistic Regression Model
%  CLASSIFYX classifies the measurements in X
%  using the logistic regression model in the file myModel.mat,
%  and then returns class labels in label.

n = size(X,1);
label = coder.nullcopy(cell(n,1));

CompactMdl = loadLearnerForCoder('myModel');
probability = predict(CompactMdl,X);

index = ~isnan(probability).*((probability<0.5)+1) + isnan(probability)*3;

classInfo = coder.load('ModelParameters');
classNames = classInfo.classNames;

for i = 1:n
    label{i} = classNames{index(i)};
end
end
```

Note: If you create a logistic regression model in Classification Learner after using feature selection or principal component analysis (PCA), you must include additional lines of code in your entry-point function. For an example that shows these additional steps, see “Code Generation and Classification Learner App” on page 32-31.

Generate a MEX function from `classifyX.m`. Create a matrix `data` for code generation using `coder.typeof`. Specify that the number of rows in `data` is arbitrary, but that `data` must have `p` columns, where `p` is the number of predictors used to train the logistic regression model. Use the `-args` option to specify `data` as an argument.

```
data = coder.typeof(X,[Inf p],[1 0]);
codegen classifyX.m -args data
```

```
Code generation successful.
```

`codegen` generates the MEX file `classifyX_mex.mexw64` in your current folder. The file extension depends on your platform.

Verify that the MEX function returns the expected labels. Randomly draw 15 observations from `X`.

```
rng('default') % For reproducibility
testX = datasample(X,15);
```

Classify the observations by using the `predictFcn` function in the classification model trained in Classification Learner.

```
testLabels = trainedModel.predictFcn(testX);
```

Classify the observations by using the generated MEX function `classifyX_mex`.

```
testLabelsMEX = classifyX_mex(testX);
```

Compare the sets of predictions. `isequal` returns logical 1 (true) if `testLabels` and `testLabelsMEX` are equal.

```
isequal(testLabels,testLabelsMEX)
```

```
ans = logical
   1
```

`predictFcn` and the MEX function `classifyX_mex` return the same values.

See Also

`codegen` | `coder.typeof` | `fitglm` | `loadLearnerForCoder` | `predict` | `saveLearnerForCoder`

Related Examples

- “Code Generation and Classification Learner App” on page 32-31
- “Train Logistic Regression Classifiers Using Classification Learner App” on page 23-95
- “Export Classification Model to Predict New Data” on page 23-77
- “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
- “Classification Learner App”
- “Introduction to Code Generation” on page 32-2

Functions

addedvarplot

Create added variable plot using input data

Syntax

```
addedvarplot(X,y,num,inmodel)
addedvarplot(X,y,num,inmodel,stats)
addedvarplot(ax, ___ )
```

Description

`addedvarplot(X,y,num,inmodel)` displays an added variable plot using the predictive terms in `X`, the response values in `y`, the added term in column `num` of `X`, and the model with current terms specified by `inmodel`. `X` is an *n*-by-*p* matrix of *n* observations of *p* predictive terms. `y` is vector of *n* response values. `num` is a scalar index specifying the column of `X` with the term to be added. `inmodel` is a logical vector of *p* elements specifying the columns of `X` in the current model. By default, all elements of `inmodel` are `false`.

Note `addedvarplot` automatically includes a constant term in all models. Do not enter a column of 1s directly into `X`.

`addedvarplot(X,y,num,inmodel,stats)` uses the `stats` output from the `stepwisefit` function to improve the efficiency of repeated calls to `addedvarplot`. Otherwise, this syntax is equivalent to the previous syntax.

`addedvarplot(ax, ___ )` creates the plot in the axes specified by `ax` instead of the current axes (`gca`). The option `ax` can precede any of the input argument combinations in the previous syntaxes. For more information on creating an `Axes` object, see `axes` and `gca`.

Added variable plots are used to determine the unique effect of adding a new term to a multilinear model. The plot shows the relationship between the part of the response unexplained by terms already in the model and the part of the new term unexplained by terms already in the model. The "unexplained" parts are measured by the residuals of the respective regressions. A scatter of the residuals from the two regressions forms the added variable plot. In addition to the scatter of residuals, the plot produced by `addedvarplot` shows 95% confidence intervals on predictions from the fitted line. The slope of the fitted line is the coefficient that the new term would have if it were added to the model with terms `inmodel`. For more details, see "Added Variable Plot" on page 33-4579.

Added variable plots are sometimes known as partial regression leverage plots.

Examples

Create Added Variable Plot

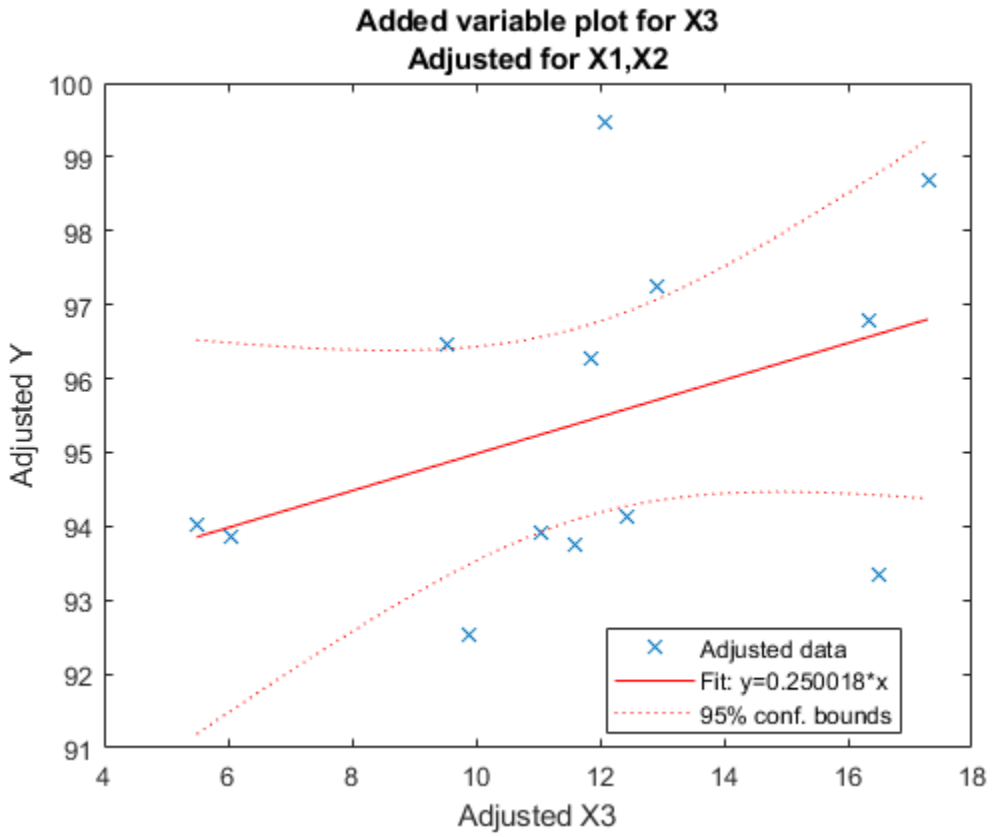
Load the data in `hald.mat`, which contains observations of the reaction to heat for various cement mixtures.

```
load hald
whos
```

```
  Name              Size            Bytes  Class     Attributes

  Description      22x58             2552  char
  hald             13x5               520  double
  heat             13x1               104  double
  ingredients      13x4               416  double
```

Create an added variable plot to investigate the effect of adding the third column of `ingredients` to a model that contains the first two columns.

```
inmodel = [true true false false];
addedvarplot(ingredients,heat,3,inmodel)
```



The wide scatter plot and the low slope of the fitted line are evidence against the statistical significance of adding the third column to the model.

Alternative Functionality

You can create a linear regression model object `LinearModel` by using `fitlm` or `stepwiselm` and use the object function `plotAdded` to create an added variable plot.

A `LinearModel` object provides the object properties and the object functions to investigate a fitted linear regression model. The object properties include information about coefficient estimates,

summary statistics, fitting method, and input data. Use the object functions to predict responses and to modify, evaluate, and visualize the linear regression model.

See Also

`plotAdded` | `stepwise` | `stepwisefit`

Introduced before R2006a

addK

Class: `clustering.evaluation.ClusterCriterion`
Package: `clustering.evaluation`

Evaluate additional numbers of clusters

Syntax

```
eva_out = addK(eva,klist)
```

Description

`eva_out = addK(eva,klist)` returns a clustering evaluation object `eva_out` that contains the evaluation data stored in the input object `eva`, plus additional evaluation data for the proposed number of clusters specified in `klist`.

Input Arguments

`eva` — Clustering evaluation data

clustering evaluation object

Clustering evaluation data, specified as a clustering evaluation object. Create a clustering evaluation object using `evalclusters`.

`klist` — Additional numbers of clusters to evaluate

vector of positive integer values

Additional numbers of clusters to evaluate, specified as a vector of positive integer values. If any values in `klist` overlap with clustering solutions already evaluated in the input object `eva`, then `addK` ignores the overlapping values.

Output Arguments

`eva_out` — Updated clustering evaluation data

clustering evaluation object

Updated clustering evaluation data, returned as a clustering evaluation object. `eva_out` contains data on the proposed clustering solutions included in the input clustering evaluation object `eva`, plus data on the additional proposed numbers of clusters specified in `klist`.

For all clustering evaluation object classes, `addK` updates the `InspectedK` and `CriterionValues` properties to include the proposed clustering solutions specified in `klist` and their corresponding criterion values. `addK` might also update the `OptimalK` and `OptimalY` properties to reflect the new optimal number of clusters and optimal clustering solution.

For certain cluster evaluation objects classes, `addK` might also update the following additional property values:

- For gap evaluation objects — `LogW`, `ExpectedLogW`, `StdLogW`, and `SE`

- For silhouette evaluation objects — `ClusterSilhouettes`

Examples

Evaluate Additional Numbers of Clusters

Create a clustering evaluation object using `evalclusters`, then use `addK` to evaluate additional numbers of clusters.

Load the sample data.

```
load fisheriris
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Cluster the flower measurement data using `kmeans`, and use the Calinski-Harabasz criterion to evaluate proposed solutions of one through five clusters.

```
eva = evalclusters(meas,'kmeans','calinski','klist',1:5)
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5]
    CriterionValues: [Inf 513.9245 561.6278 530.4871 456.1279]
           OptimalK: 1
```

The clustering evaluation object `eva` contains data on each proposed clustering solution. The returned value of `OptimalK` indicates that the optimal solution is three clusters.

Evaluate proposed solutions of 6 through 10 clusters using the same criteria. Add these evaluations to the original clustering evaluation object `eva`.

```
eva = addK(eva,6:10)
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6 7 8 9 10]
    CriterionValues: [1x10 double]
           OptimalK: 1
```

The updated values for `InspectedK` and `CriterionValues` show that `eva` now evaluates proposed solutions of 1 through 10 clusters. The `OptimalK` value still equals `3`, indicating that three clusters remain the optimal solution.

See Also

`CalinskiHarabaszEvaluation` | `DaviesBouldinEvaluation` | `GapEvaluation` | `SilhouetteEvaluation` | `evalclusters`

addlevels

(Not Recommended) Add levels to nominal or ordinal arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
B = addlevels(A,newlevels)
```

Description

`B = addlevels(A,newlevels)` adds new levels specified by `newlevels` to the `nominal` or `ordinal` array `A`. `addlevels` adds the new levels at the end of the list of possible levels in `A`, but does not modify the value of any element. `B` does not contain elements at the new levels.

Examples

Add Levels To A Nominal Array

Add levels for additional species to Fisher's iris data.

Create a nominal array of the existing species in Fisher's iris data.

```
load fisheriris
species = nominal(species);
getlevels(species)
```

```
ans = 1x3 nominal
     setosa      versicolor      virginica
```

Add two additional species.

```
species = addlevels(species,{'spuria','ruthenica'});
getlevels(species)
```

```
ans = 1x5 nominal
     setosa      versicolor      virginica      spuria      ruthenica
```

Even though there are new levels, there are no elements in `species` that are in these new levels.

```
sum(species=='spuria')
```

```
ans = 0
```

```
sum(species=='ruthenica')
```

```
ans = 0
```

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

newlevels — Levels to add

string array | cell array of character vectors | 2-D character matrix

Levels to add to the input `nominal` or `ordinal` array, specified as a string array, a cell array of character vectors, or a 2-D character matrix.

Data Types: `char` | `string` | `cell`

Output Arguments

B — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, returned as a `nominal` or `ordinal` array object.

See Also

`droplevels` | `mergelevels` | `nominal` | `ordinal` | `reorderlevels`

Topics

“Add and Drop Category Levels” on page 2-18

Introduced in R2007a

addInteractions

Add interaction terms to univariate generalized additive model (GAM)

Syntax

```
UpdatedMdl = addInteractions(Mdl,Interactions)
UpdatedMdl = addInteractions(Mdl,Interactions,Name,Value)
```

Description

`UpdatedMdl = addInteractions(Mdl,Interactions)` returns an updated model `UpdatedMdl` by adding the interaction terms in `Interactions` to the univariate generalized additive model `Mdl`. The model `Mdl` must contain only linear terms for predictors.

If you want to resume training for the existing terms in `Mdl`, use the `resume` function.

`UpdatedMdl = addInteractions(Mdl,Interactions,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'MaxPValue',0.05` specifies to include only the interaction terms whose *p*-values are not greater than 0.05.

Examples

Train GAM with Interaction Terms

Train a univariate GAM, which contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table that contains the predictor variables (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and the response variable (`MPG`).

```
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Train a univariate GAM that contains linear terms for predictors in `tbl`.

```
Mdl = fitrgam(tbl,'MPG');
```

Add the five most important interaction terms to the trained model.

```
UpdatedMdl = addInteractions(Mdl,5);
```

`Mdl` is a univariate GAM, and `UpdatedMdl` is an updated GAM that contains all the terms in `Mdl` and five additional interaction terms. Display the interaction terms in `UpdatedMdl`.

```
UpdatedMdl.Interactions
```

```
ans = 5×2
```

```
     2     3
     1     2
     3     4
     1     4
     1     3
```

Each row of the `Interactions` property represents one interaction term and contains the column indexes of the predictor variables for the interaction term. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitrgam` adds them to the model.

Specify Options for Interaction Terms

Train a univariate GAM, which contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function. Specify the `'MaxPValue'` name-value argument to add interaction terms whose *p*-values are not greater than the `'MaxPValue'` value.

Load Fisher's iris data set. Create a table that contains observations for versicolor and virginica.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
Tbl = array2table(meas(inds,:),'VariableNames',["x1","x2","x3","x4"]);
Tbl.Y = species(inds,:);
```

Train a univariate GAM that contains linear terms for predictors in `Tbl`.

```
Mdl = fitcgam(Tbl,'Y');
```

Add important interaction terms to the trained model `Mdl`. Specify `'all'` for the `Interactions` argument, and set the `'MaxPValue'` name-value argument to 0.05. Among all available interaction terms, `addInteractions` identifies those whose *p*-values are not greater than the `'MaxPValue'` value and adds them to the model. The default `'MaxPValue'` is 1 so that the function adds all specified interaction terms to the model.

```
UpdatedMdl = addInteractions(Mdl,'all','MaxPValue',0.05);
UpdatedMdl.Interactions
```

```
ans = 5×2

     3     4
     2     4
     1     4
     2     3
     1     3
```

`Mdl` is a univariate GAM, and `UpdatedMdl` is an updated GAM that contains all the terms in `Mdl` and five additional interaction terms. `UpdatedMdl` includes five of the six available pairs of interaction terms.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `RegressionGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `RegressionGAM` model object.

`Interactions` — Number of interaction terms or list of interaction terms

`0` | nonnegative integer | logical matrix | `'all'`

Number or list of interaction terms to include in the candidate set *S*, specified as a nonnegative integer scalar, a logical matrix, or `'all'`.

- Number of interaction terms, specified as a nonnegative integer — *S* includes the specified number of important interaction terms, selected based on the *p*-values of the terms.
- List of interaction terms, specified as a logical matrix — *S* includes the terms specified by a `t`-by-`p` logical matrix, where `t` is the number of interaction terms, and `p` is the number of predictors used to train the model. For example, `logical([1 1 0; 0 1 1])` represents two pairs of interaction terms: a pair of the first and second predictors, and a pair of the second and third predictors.

  If `addInteractions` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. That is, the column indexes of the logical matrix do not count the response and observation weight variables. The indexes also do not count any variables not used by the function.
- `'all'` — *S* includes all possible pairs of interaction terms, which is `p*(p − 1)/2` number of terms in total.

Among the interaction terms in *S*, the `addInteractions` function identifies those whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees. Use the default value (`'MaxPValue'`,1) to build interaction trees using all terms in *S*.

Data Types: `single` | `double` | `logical` | `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `addInteractions(Mdl,'all','MaxPValue',0.05,'Verbose',1,'NumPrints',10)` specifies to include all available interaction terms whose *p*-values are not greater than 0.05 and to display diagnostic messages every 10 iterations.

`InitialLearnRateForInteractions` — Initial learning rate of gradient boosting for interaction terms

`1` (default) | numeric scalar in (0,1]

Initial learning rate of gradient boosting for interaction terms, specified as a numeric scalar in the interval (0,1].

For each boosting iteration for interaction trees, `addInteractions` starts fitting with the initial learning rate. The function halves the learning rate until it finds a rate that improves the model fit.

Training a model using a small learning rate requires more learning iterations, but often achieves better accuracy.

For more details about gradient boosting, see “Gradient Boosting Algorithm” on page 33-15.

Example: `'InitialLearnRateForInteractions',0.1`

Data Types: `single` | `double`

`MaxNumSplitsPerInteraction` — Maximum number of decision splits per interaction tree
4 (default) | positive integer scalar

Maximum number of decision splits (or branch nodes) for each interaction tree (boosted tree for an interaction term), specified as a positive integer scalar.

Example: `'MaxNumSplitsPerInteraction',5`

Data Types: `single` | `double`

`MaxPValue` — Maximum *p*-value for detecting interaction terms
1 (default) | numeric scalar in [0,1]

Maximum *p*-value for detecting interaction terms, specified as a numeric scalar in the interval [0,1].

`addInteractions` first finds the candidate set *S* of interaction terms from the `Interactions` value. Then the function identifies the interaction terms whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees.

The default value (`'MaxPValue',1`) builds interaction trees for all interaction terms in the candidate set *S*.

For more details about detecting interaction terms, see “Interaction Term Detection” on page 33-15.

Example: `'MaxPValue',0.05`

Data Types: `single` | `double`

`NumPrint` — Number of iterations between diagnostic message printouts
`Mdl.ModelParameters.NumPrint` (default) | nonnegative integer scalar

Number of iterations between diagnostic message printouts, specified as a nonnegative integer scalar. This argument is valid only when you specify `'Verbose'` as 1.

If you specify `'Verbose',1` and `'NumPrint',numPrint`, then the software displays diagnostic messages every `numPrint` iterations in the Command Window.

The default value is `Mdl.ModelParameters.NumPrint`, which is the `NumPrint` value that you specify when creating the GAM object `Mdl`.

Example: `'NumPrint',500`

Data Types: `single` | `double`

`NumTreesPerInteraction` — Number of trees per interaction term
100 (default) | positive integer scalar

Number of trees per interaction term, specified as a positive integer scalar.

The `'NumTreesPerInteraction'` value is equivalent to the number of gradient boosting iterations for the interaction terms for predictors. For each iteration, `addInteractions` adds a set of interaction trees to the model, one tree for each interaction term. To learn about the gradient boosting algorithm, see “Gradient Boosting Algorithm” on page 33-15.

You can determine whether the fitted model has the specified number of trees by viewing the diagnostic message displayed when `'Verbose'` is 1 or 2, or by checking the `ReasonForTermination` property value of the model `Mdl`.

Example: `'NumTreesPerInteraction',500`

Data Types: `single` | `double`

`Verbose` — Verbosity level
`Mdl.ModelParameters.VerbosityLevel` (default) | `0` | `1` | `2`

Verbosity level, specified as `0`, `1`, or `2`. The `Verbose` value controls the amount of information that the software displays in the Command Window.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software displays no information.
`1`	The software displays diagnostic messages every `numPrint` iterations, where `numPrint` is the `'NumPrint'` value.
`2`	The software displays diagnostic messages at every iteration.

Each line of the diagnostic messages shows the information about each boosting iteration and includes the following columns:

- `Type` — Type of trained trees, `1D` (predictor trees, or boosted trees for linear terms for predictors) or `2D` (interaction trees, or boosted trees for interaction terms for predictors)
- `NumTrees` — Number of trees per linear term or interaction term that `addInteractions` added to the model so far
- `Deviance` — "Deviance" on page 33-15 of the model
- `RelTol` — Relative change of model predictions: $(\widehat{y}_k - \widehat{y}_{k-1})'(\widehat{y}_k - \widehat{y}_{k-1})/\widehat{y}_k'\widehat{y}_k$, where $\widehat{y}_k$ is a column vector of model predictions at iteration k
- `LearnRate` — Learning rate used for the current iteration

The default value is `Mdl.ModelParameters.VerbosityLevel`, which is the `Verbose` value that you specify when creating the GAM object `Mdl`.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`UpdatedMdl` — Updated generalized additive model
`ClassificationGAM` model object | `RegressionGAM` model object

Updated generalized additive model, returned as a `ClassificationGAM` or `RegressionGAM` model object. `UpdatedMdl` has the same object type as the input model `Mdl`.

To overwrite the input argument `Mdl`, assign the output of `addInteractions` to `Mdl`:

```
Mdl = addInteractions(Mdl,Interactions);
```

More About

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to the saturated model.

The deviance of a fitted model is twice the difference between the loglikelihoods of the model and the saturated model:

$$-2(\log L - \log L_s),$$

where L and L_s are the likelihoods of the fitted model and the saturated model, respectively. The saturated model is the model with the maximum number of parameters that you can estimate.

`addInteractions` uses the deviance to measure the goodness of model fit and finds a learning rate that reduces the deviance at each iteration. Specify `'Verbose'` as 1 or 2 to display the deviance and learning rate in the Command Window.

Algorithms

Gradient Boosting Algorithm

`addInteractions` adds sets of interaction trees (boosted trees for interaction terms for predictors) to a univariate generalized additive model by using a gradient boosting algorithm (“Least-Squares Boosting” on page 18-50 for regression and “Adaptive Logistic Regression” on page 18-48 for classification). The algorithm iterates for at most `'NumTreesPerInteraction'` times for interaction trees.

For each boosting iteration, `addInteractions` builds a set of interaction trees with the initial learning rate `'InitialLearnRateForInteractions'`.

- When building a set of trees, the function trains one tree at a time. It fits a tree to the residual that is the difference between the response (observed response values for regression or scores of observed classes for classification) and the aggregated prediction from all trees grown previously. To control the boosting learning speed, the function shrinks the tree by the learning rate and then adds the tree to the model and updates the residual.
  - Updated model = current model + (learning rate)·(new tree)
  - Updated residual = current residual – (learning rate)·(response explained by new tree)
- If adding the set of trees improves the model fit (that is, reduces the deviance of the fit), then `addInteractions` moves to the next iteration.
- Otherwise, `addInteractions` halves the learning rate and uses it to update the model and residual. The function continues to halve the learning rate until it finds a rate that improves the model fit.

  If the function cannot find such a learning rate for interaction trees, then it terminates the model fitting. You can determine why training stopped by checking the `ReasonForTermination` property of the trained model.

Interaction Term Detection

For each pairwise interaction term $x_i x_j$ (specified by `Interactions`), the software performs an F-test to examine whether the term is statistically significant.

To speed up the process, `addInteractions` bins numeric predictors into at most 8 equiprobable bins. The number of bins can be less than 8 if a predictor has fewer than 8 unique values. The *F*-test examines the null hypothesis that the bins created by x_i and x_j have equal responses versus the alternative that at least one bin has a different response value from the others. A small *p*-value indicates that differences are significant, which implies that the corresponding interaction term is significant and, therefore, including the term can improve the model fit.

`addInteractions` builds a set of interaction trees using the terms whose *p*-values are not greater than the `'MaxPValue'` value. You can use the default `'MaxPValue'` value `1` to build interaction trees using all terms specified by `Interactions`.

`addInteractions` adds interaction terms to the model in the order of importance based on the *p*-values. Use the `Interactions` property of the returned model to check the order of the interaction terms added to the model.

See Also

`ClassificationGAM` | `RegressionGAM` | `resume`

Topics

“Train Generalized Additive Model for Binary Classification” on page 12-77
“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

addlistener

Class: `qrandstream`

Add listener for event

Syntax

```
el = addlistener(hsource,'eventname',callback)
el = addlistener(hsource,property,'eventname',callback)
```

Description

`el = addlistener(hsource,'eventname',callback)` creates a listener for the event named `eventname`, the source of which is handle object `hsource`. If `hsource` is an array of source handles, the listener responds to the named event on any handle in the array. `callback` is a function handle that is invoked when the event is triggered.

`el = addlistener(hsource,property,'eventname',callback)` adds a listener for a property event. `eventname` must be `'PreGet'`, `'PostGet'`, `'PreSet'`, or `'PostSet'`. `property` must be either a property name or cell array of property names, or a `meta.property` or array of `meta.property`. The properties must belong to the class of `hsource`. If `hsource` is scalar, `property` can include dynamic properties.

For all forms, `addlistener` returns an `event.listener`. To remove a listener, delete the object returned by `addlistener`. For example, `delete(el)` calls the handle class delete method to remove the listener and delete it from the workspace.

See Also

`delete` | `dynamicprops` | `event.listener` | `events` | `meta.property` | `notify` | `qrandstream` | `reset`

anova

Class: `GeneralizedLinearMixedModel`

Analysis of variance for generalized linear mixed-effects model

Syntax

```
stats = anova(glme)
stats = anova(glme,Name,Value)
```

Description

`stats = anova(glme)` returns a table, `stats`, that contains the results of *F*-tests to determine if all coefficients representing each fixed-effects term in the generalized linear mixed-effects model `glme` are equal to 0.

`stats = anova(glme,Name,Value)` returns a table, `stats`, using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the method used to compute the approximate denominator degrees of freedom for the *F*-tests.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`DFMethod` — Method for computing approximate denominator degrees of freedom

`'residual'` (default) | `'none'`

Method for computing approximate denominator degrees of freedom to use in the *F*-test, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	All degrees of freedom are set to infinity.

The denominator degrees of freedom for the *F*-statistic correspond to the column `DF2` in the output structure `stats`.

Example: `'DFMethod','none'`

Output Arguments

`stats` — Results of *F*-tests for fixed-effects terms

table

Results of *F*-tests for fixed-effects terms, returned as a table with one row for each fixed-effects term in `glme` and the following columns.

Column Name	Description
`Term`	Name of the fixed-effects term
`FStat`	*F*-statistic for the term
`DF1`	Numerator degrees of freedom for the *F*-statistic
`DF2`	Denominator degrees of freedom for the *F*-statistic
`pValue`	*p*-value for the term

Each fixed-effects term is a continuous variable, a grouping variable, or an interaction between two or more continuous or grouping variables. For each fixed-effects term, `anova` performs an *F*-test (marginal test) to determine if all coefficients representing the fixed-effects term are equal to 0.

To perform tests for the type III hypothesis, when fitting the generalized linear mixed-effects model `fitglme`, you must use the `'effects'` contrasts for the `'DummyVarCoding'` name-value pair argument.

Examples

F-Tests for Fixed Effects

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defect}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log\mu_{ij} = \beta_0 + \beta_1\text{newprocess}_{ij} + \beta_2\text{time\_dev}_{ij} + \beta_3\text{temp\_dev}_{ij} + \beta_4\text{supplier\_C}_{ij} + \beta_5\text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company C or B, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects')
```

```
glme =
Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35
```

```
Fixed effects coefficients (95% CIs):
    Name                  Estimate     SE          tStat       DF    pValue
    {'(Intercept)'}          1.4689     0.15988      9.1875    94    9.8194e-15
    {'newprocess' }         -0.36766    0.17755     -2.0708    94      0.041122
    {'time_dev'   }        -0.094521    0.82849    -0.11409    94       0.90941
    {'temp_dev'   }         -0.28317     0.9617    -0.29444    94       0.76907
    {'supplier_C' }        -0.071868    0.078024     -0.9211    94       0.35936
    {'supplier_B' }         0.071072    0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
     -0.72019    -0.015134
      -1.7395       1.5505
      -2.1926       1.6263
     -0.22679     0.083051
    -0.082588      0.22473

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       0.31381

Group: Error
    Name                       Estimate
    {'sqrt(Dispersion)'}       1
```

Perform an *F*-test to determine if all fixed-effects coefficients are equal to 0.

```
stats = anova(glme)
```

```
stats =
    ANOVA marginal tests: DFMethod = 'residual'

    Term                  FStat       DF1    DF2    pValue
    {'(Intercept)'}          84.41    1      94     9.8194e-15
    {'newprocess' }        4.2881     1      94       0.041122
    {'time_dev'   }       0.013016    1      94        0.90941
    {'temp_dev'   }       0.086696    1      94        0.76907
    {'supplier'   }        0.59212    2      94         0.5552
```

The *p*-values for the intercept, `newprocess`, `time_dev`, and `temp_dev` are the same as in the coefficient table of the `glme` display. The small *p*-values for the intercept and `newprocess` indicate that these are significant predictors at the 5% significance level. The large *p*-values for `time_dev` and `temp_dev` indicate that these are not significant predictors at this level.

The *p*-value of 0.5552 for `supplier` measures the combined significance for both coefficients representing the categorical variable `supplier`. This includes the dummy variables `supplier_C` and `supplier_B` as shown in the coefficient table of the `glme` display. The large *p*-value indicates that `supplier` is not a significant predictor at the 5% significance level.

Tips

- For each fixed-effects term, `anova` performs an *F*-test (marginal test) to determine if all coefficients representing the fixed-effects term are equal to 0.

  When fitting a generalized linear mixed-effects (GLME) model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`):

  - If you specify the `'CovarianceMethod'` name-value pair argument as `'conditional'`, then the *F*-tests are conditional on the estimated covariance parameters.
  - If you specify the `'CovarianceMethod'` name-value pair as `'JointHessian'`, then the *F*-tests account for the uncertainty in estimation of covariance parameters.

  When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `anova` uses the fitted linear mixed effects model from the final pseudo likelihood iteration for inference on fixed effects.

See Also

`GeneralizedLinearMixedModel` | `coefCI` | `coefTest` | `disp` | `fitglme` | `fixedEffects`

addTerms

Add terms to generalized linear regression model

Syntax

```
NewMdl = addTerms(mdl,terms)
```

Description

`NewMdl = addTerms(mdl,terms)` returns a generalized linear regression model fitted using the input data and settings in `mdl` with the terms `terms` added.

Examples

Add Terms to Generalized Linear Regression Model

Create a generalized linear regression model using one predictor, and then add another predictor.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data. Include only the first predictor in the model.

```
mdl = fitglm(X,y,'y ~ x1','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     2.7784      0.014043    197.85      0
    x1              1.1732     0.0033653     348.6      0


100 observations, 98 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.25e+05, p-value = 0
```

Add the second predictor to the model.

```
mdl1 = addTerms(mdl,'x2')
```

```
mdl1 =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

`terms` — Terms to add to regression model

character vector or string scalar formula in Wilkinson notation | *t*-by-*p* terms matrix

Terms to add to the regression model `mdl`, specified as one of the following:

- Character vector or string scalar formula in "Wilkinson Notation" on page 33-25 representing one or more terms. The variable names in the formula must be valid MATLAB identifiers.
- Terms matrix `T` of size *t*-by-*p*, where *t* is the number of terms and *p* is the number of predictor variables in `mdl`. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

  For example, suppose `mdl` has three variables `A`, `B`, and `C` in that order. Each row of `T` represents one term:

  - `[0 0 0]` — Constant term or intercept
  - `[0 1 0]` — `B`; equivalently, `A^0 * B^1 * C^0`
  - `[1 0 1]` — `A*C`
  - `[2 0 0]` — `A^2`
  - `[0 1 2]` — `B*(C^2)`

`addTerms` treats a group of indicator variables for a categorical predictor as a single variable. Therefore, you cannot specify an indicator variable to add to the model. If you specify a categorical predictor to add to the model, `addTerms` adds a group of indicator variables for the predictor in one step.

Output Arguments

`NewMdl` — Generalized linear regression model with additional terms

`GeneralizedLinearModel` object

Generalized linear regression model with additional terms, returned as a `GeneralizedLinearModel` object. `NewMdl` is a newly fitted model that uses the input data and settings in `mdl` with additional terms specified in `terms`.

To overwrite the input argument `mdl`, assign the newly fitted model to `mdl`:

```
mdl = addTerms(mdl,terms);
```

More About

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- `+` means include the next variable.
- `−` means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- `*` defines an interaction and all lower-order terms.
- `^` raises the predictor to a power, exactly as in `*` repeated, so `^` includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
`1`	Constant (intercept) term
`x1^k`, where `k` is a positive integer	`x1`, $x1^2$, ..., $x1^k$
`x1 + x2`	`x1`, `x2`
`x1*x2`	`x1`, `x2`, `x1*x2`
`x1:x2`	`x1*x2` only
`−x2`	Do not include `x2`
`x1*x2 + x3`	`x1`, `x2`, `x3`, `x1*x2`
`x1 + x2 + x3 + x1:x2`	`x1`, `x2`, `x3`, `x1*x2`
`x1*x2*x3 − x1:x2:x3`	`x1`, `x2`, `x3`, `x1*x2`, `x1*x3`, `x2*x3`
`x1*(x2 + x3)`	`x1`, `x2`, `x3`, `x1*x2`, `x1*x3`

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- `addTerms` treats a categorical predictor as follows:
  - A model with a categorical predictor that has L levels (categories) includes L – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `addTerms` treats the group of L – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Alternative Functionality

- Use `stepwiseglm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `removeTerms` to remove specific terms from a model.
- Use `step` to optimally improve a model by adding or removing terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `removeTerms` | `step` | `stepwiseglm`

Topics

“Generalized Linear Models” on page 12-9

Introduced in R2012a

addTerms

Add terms to linear regression model

Syntax

```
NewMdl = addTerms(mdl,terms)
```

Description

`NewMdl = addTerms(mdl,terms)` returns a linear regression model fitted using the input data and settings in `mdl` with the terms `terms` added.

Examples

Add Terms to Linear Regression Model

Create a linear regression model of the `carsmall` data set without any interactions, and then add an interaction term.

Load the `carsmall` data set and create a model of the MPG as a function of weight and model year.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Weight + Year + Weight^2

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________

    (Intercept)        54.206        4.7117     11.505    2.6648e-19
    Weight          -0.016404     0.0031249    -5.2493    1.0283e-06
    Year_76            2.0887       0.71491     2.9215     0.0044137
    Year_82            8.1864       0.81531     10.041    2.6364e-16
    Weight^2       1.5573e-06    4.9454e-07      3.149     0.0022303


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.78
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 172, p-value = 5.52e-41
```

The model includes five terms, `Intercept`, `Weight`, `Year_76`, `Year_82`, and `Weight^2`, where `Year_76` and `Year_82` are indicator variables for the categorical variable `Year` that has three distinct values.

Add an interaction term between the `Year` and `Weight` variables to `mdl`.

```
terms = 'Year*Weight';
NewMdl = addTerms(mdl,terms)

NewMdl =
Linear regression model:
    MPG ~ 1 + Weight*Year + Weight^2

Estimated Coefficients:
                       Estimate          SE         tStat        pValue
                      ___________    __________    ________    __________

    (Intercept)           48.045         6.779      7.0874    3.3967e-10
    Weight              -0.012624     0.0041455     -3.0454     0.0030751
    Year_76               2.7768        3.0538     0.90931        0.3657
    Year_82               16.416        4.9802      3.2962     0.0014196
    Weight:Year_76    -0.00020693    0.00092403    -0.22394       0.82333
    Weight:Year_82     -0.0032574     0.0018919     -1.7217      0.088673
    Weight^2           1.0121e-06      6.12e-07      1.6538       0.10177


Number of observations: 94, Error degrees of freedom: 87
Root Mean Squared Error: 2.76
R-squared: 0.89,  Adjusted R-Squared: 0.882
F-statistic vs. constant model: 117, p-value = 1.88e-39
```

`NewMdl` includes two additional terms, `Weight*Year_76` and `Weight*Year_82`.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`terms` — Terms to add to regression model

character vector or string scalar formula in Wilkinson notation | *t*-by-*p* terms matrix

Terms to add to the regression model `mdl`, specified as one of the following:

- Character vector or string scalar formula in “Wilkinson Notation” on page 33-30 representing one or more terms. The variable names in the formula must be valid MATLAB identifiers.
- Terms matrix `T` of size *t*-by-*p*, where *t* is the number of terms and *p* is the number of predictor variables in `mdl`. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

  For example, suppose `mdl` has three variables `A`, `B`, and `C` in that order. Each row of `T` represents one term:

  - `[0 0 0]` — Constant term or intercept
  - `[0 1 0]` — `B`; equivalently, `A^0 * B^1 * C^0`
  - `[1 0 1]` — `A*C`
  - `[2 0 0]` — `A^2`
  - `[0 1 2]` — `B*(C^2)`

`addTerms` treats a group of indicator variables for a categorical predictor as a single variable. Therefore, you cannot specify an indicator variable to add to the model. If you specify a categorical predictor to add to the model, `addTerms` adds a group of indicator variables for the predictor in one step. See “Modify Linear Regression Model Using step” on page 33-5990 for an example that describes how to create indicator variables manually and treat each one as a separate variable.

Output Arguments

`NewMdl` — Linear regression model with additional terms

`LinearModel` object

Linear regression model with additional terms, returned as a `LinearModel` object. `NewMdl` is a newly fitted model that uses the input data and settings in `mdl` with additional terms specified in `terms`.

To overwrite the input argument `mdl`, assign the newly fitted model to `mdl`:

```
mdl = addTerms(mdl,terms);
```

More About

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.
- − means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
−x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 − x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3

Wilkinson Notation	Terms in Standard Notation
`x1*(x2 + x3)`	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- `addTerms` treats a categorical predictor as follows:
  - A model with a categorical predictor that has *L* levels (categories) includes *L* – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `addTerms` treats the group of *L* – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with *L* levels consist of the element-wise product of the *L* – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with *L* and *M* levels consist of the (*L* – 1)*(*M* – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Alternative Functionality

- Use `stepwiselm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `removeTerms` to remove specific terms from a model.
- Use `step` to optimally improve a model by adding or removing terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `removeTerms` | `step` | `stepwiselm`

Topics

“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9
“Stepwise Regression” on page 11-99

Introduced in R2012a

adtest

Anderson-Darling test

Syntax

```
h = adtest(x)
h = adtest(x,Name,Value)
[h,p] = adtest( ___ )
[h,p,adstat,cv] = adtest( ___ )
```

Description

`h = adtest(x)` returns a test decision for the null hypothesis that the data in vector `x` is from a population with a normal distribution, using the Anderson-Darling test on page 33-37. The alternative hypothesis is that `x` is not from a population with a normal distribution. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`h = adtest(x,Name,Value)` returns a test decision for the Anderson-Darling test with additional options specified by one or more name-value pair arguments. For example, you can specify a null distribution other than normal, or select an alternative method for calculating the *p*-value.

`[h,p] = adtest( ___ )` also returns the *p*-value, `p`, of the Anderson-Darling test, using any of the input arguments from the previous syntaxes.

`[h,p,adstat,cv] = adtest( ___ )` also returns the test statistic, `adstat`, and the critical value, `cv`, for the Anderson-Darling test.

Examples

Anderson-Darling Test for a Normal Distribution

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the exam grades come from a normal distribution. You do not need to specify values for the population parameters.

```
[h,p,adstat,cv] = adtest(x)
```

```
h = logical
   0


p = 0.1854

adstat = 0.5194

cv = 0.7470
```

The returned value of `h = 0` indicates that `adtest` fails to reject the null hypothesis at the default 5% significance level.

Anderson-Darling Test for Extreme Value Distribution

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the exam grades come from an extreme value distribution. You do not need to specify values for the population parameters.

```
[h,p] = adtest(x,'Distribution','ev')
```

```
h = logical
   0


p = 0.0714
```

The returned value of `h = 0` indicates that `adtest` fails to reject the null hypothesis at the default 5% significance level.

Anderson-Darling Test Using Specified Probability Distribution

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Create a normal probability distribution object with mean `mu = 75` and standard deviation `sigma = 10`.

```
dist = makedist('normal','mu',75,'sigma',10)
```

```
dist =
  NormalDistribution

  Normal distribution
       mu = 75
    sigma = 10
```

Test the null hypothesis that `x` comes from the hypothesized normal distribution.

```
[h,p] = adtest(x,'Distribution',dist)
```

```
h = logical
   0


p = 0.4687
```

The returned value of `h = 0` indicates that `adtest` fails to reject the null hypothesis at the default 5% significance level.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector. Missing observations in `x`, indicated by `NaN`, are ignored.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'MCTol',0.01` conducts the hypothesis test at the 1% significance level, and determines the p-value, `p`, using a Monte Carlo simulation with a maximum Monte Carlo standard error for `p` of 0.01.

Distribution — Hypothesized distribution

`'norm'` (default) | `'exp'` | `'ev'` | `'logn'` | `'weibull'` | probability distribution object

Hypothesized distribution of data vector `x`, specified as the comma-separated pair consisting of `'Distribution'` and one of the following.

`'norm'`	Normal distribution
`'exp'`	Exponential distribution
`'ev'`	Extreme value distribution
`'logn'`	Lognormal distribution
`'weibull'`	Weibull distribution

In this case, you do not need to specify population parameters. Instead, `adtest` estimates the distribution parameters from the sample data and tests `x` against a composite hypothesis that it comes from the selected distribution family with parameters unspecified.

Alternatively, you can specify any continuous probability distribution object for the null distribution. In this case, you must specify all the distribution parameters, and `adtest` tests `x` against a simple hypothesis that it comes from the given distribution with its specified parameters.

Example: `'Distribution','exp'`

Alpha — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`MCTol` — Maximum Monte Carlo standard error
positive scalar value

Maximum Monte Carlo standard error on page 33-38 for the *p*-value, `p`, specified as the comma-separated pair consisting of `'MCTol'` and a positive scalar value. If you use `MCTol`, `adtest` determines `p` using a Monte Carlo simulation, and the name-value pair argument `Asymptotic` must have the value `false`.

Example: `'MCTol',0.01`

Data Types: `single` | `double`

`Asymptotic` — Method for calculating *p*-value
`false` (default) | `true`

Method for calculating the *p*-value of the Anderson-Darling test, specified as the comma-separated pair consisting of `'Asymptotic'` and either `true` or `false`. If you specify `'true'`, `adtest` estimates the *p*-value using the limiting distribution of the Anderson-Darling test statistic. If you specify `false`, `adtest` calculates the *p*-value based on an analytical formula. For sample sizes greater than 120, the limiting distribution estimate is likely to be more accurate than the small sample size approximation method.

- If you specify a distribution family with unknown parameters for the `Distribution` name-value pair, `Asymptotic` must be `false`.
- If you use `MCTol` to calculate the *p*-value using a Monte Carlo simulation, `Asymptotic` must be `false`.

Example: `'Asymptotic',true`

Data Types: `logical`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as a logical value.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the Anderson-Darling test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. `p` is calculated using one of these methods:

- If the hypothesized distribution is a fully specified probability distribution object, `adtest` calculates `p` analytically. If `'Asymptotic'` is `true`, `adtest` uses the asymptotic distribution of the test statistic. If you specify a value for `'MCTol'`, `adtest` uses a Monte Carlo simulation.
- If the hypothesized distribution is specified as a distribution family with unknown parameters, `adtest` retrieves the critical value from a table and uses inverse interpolation to determine the *p*-value. If you specify a value for `'MCTol'`, `adtest` uses a Monte Carlo simulation.

`adstat` — Test statistic

scalar value

Test statistic for the Anderson-Darling test, returned as a scalar value.

- If the hypothesized distribution is a fully specified probability distribution object, `adtest` computes `adstat` using specified parameters.
- If the hypothesized distribution is specified as a distribution family with unknown parameters, `adtest` computes `adstat` using parameters estimated from the sample data.

`cv` — Critical value

scalar value

Critical value for the Anderson-Darling test at the significance level `Alpha`, returned as a scalar value. `adtest` determines `cv` by interpolating into a table based on the specified `Alpha` significance level.

More About

Anderson-Darling Test

The Anderson-Darling test is commonly used to test whether a data sample comes from a normal distribution. However, it can be used to test for another hypothesized distribution, even if you do not fully specify the distribution parameters. Instead, the test estimates any unknown parameters from the data sample.

The test statistic belongs to the family of quadratic empirical distribution function statistics, which measure the distance between the hypothesized distribution, $F(x)$ and the empirical cdf, $F_n(x)$ as

$$n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2w(x)dF(x),$$

over the ordered sample values $x_1 < x_2 < \ldots < x_n$, where $w(x)$ is a weight function and n is the number of data points in the sample.

The weight function for the Anderson-Darling test is

$$w(x) = [F(x)(1-F(x))]^{-1},$$

which places greater weight on the observations in the tails of the distribution, thus making the test more sensitive to outliers and better at detecting departure from normality in the tails of the distribution.

The Anderson-Darling test statistic is

$$A_n^2 = -n - \sum_{i=1}^{n}\frac{2i-1}{n}[\ln(F(X_i)) + \ln(1-F(X_{n+1-i}))],$$

where$\{X_1 < \ldots < X_n\}$ are the ordered sample data points and n is the number of data points in the sample.

In `adtest`, the decision to reject or not reject the null hypothesis is based on comparing the p-value for the hypothesis test with the specified significance level, not on comparing the test statistic with the critical value.

Monte Carlo Standard Error

The Monte Carlo standard error is the error due to simulating the *p*-value.

The Monte Carlo standard error is calculated as

$$SE = \sqrt{\frac{(\widehat{p})(1-\widehat{p})}{\text{mcreps}}},$$

where $\widehat{p}$ is the estimated *p*-value of the hypothesis test, and `mcreps` is the number of Monte Carlo replications performed.

`adtest` chooses the number of Monte Carlo replications, `mcreps`, large enough to make the Monte Carlo standard error for $\widehat{p}$ less than the value specified for `MCTol`.

See Also

`jbtest` | `kstest`

Introduced in R2013a

andrewsplot

Andrews plot

Syntax

```
andrewsplot(X)
andrewsplot(X,...,'Standardize',standopt)
andrewsplot(X,...,'Quantile',alpha)
andrewsplot(X,...,'Group',group)
andrewsplot(X,...,'PropName',PropVal,...)
andrewsplot(ax,X,...)
h = andrewsplot(X,...)
```

Description

`andrewsplot(X)` creates an Andrews plot of the multivariate data in the matrix `X`. The rows of `X` correspond to observations, the columns to variables. Andrews plots represent each observation by a function $f(t)$ of a continuous dummy variable t over the interval [0,1]. $f(t)$ is defined for the ith observation in `X` as

$$f(t) = X(i,1)/\sqrt{2} + X(i,2)\sin(2\pi t) + X(i,3)\cos(2\pi t) + \ldots$$

`andrewsplot` treats `NaN` values in `X` as missing values and ignores the corresponding rows.

`andrewsplot(X,...,'Standardize',standopt)` creates an Andrews plot where *`standopt`* is one of the following:

- `'on'` — scales each column of `X` to have mean `0` and standard deviation `1` before making the plot.
- `'PCA'` — creates an Andrews plot from the principal component scores of `X`, in order of decreasing eigenvalue. (See `pca`.)
- `'PCAStd'` — creates an Andrews plot using the standardized principal component scores. (See `pca`.)

`andrewsplot(X,...,'Quantile',alpha)` plots only the median and the `alpha` and (1 – `alpha`) quantiles of $f(t)$ at each value of t. This is useful if `X` contains many observations.

`andrewsplot(X,...,'Group',group)` plots the data in different groups with different colors. Groups are defined by `group`, a numeric array containing a group index for each observation. `group` can also be a categorical array, character matrix, string array, or cell array of character vectors containing a group name for each observation.

`andrewsplot(X,...,'PropName',PropVal,...)` sets optional `Line` object properties to the specified values for all `Line` objects created by `andrewsplot`. (See Line Properties.)

`andrewsplot(ax,X,...)` uses the plot axes specified in `ax`, an `Axes` object. (See `axes`.) Specify `ax` as the first input argument followed by any of the input argument combinations in the previous syntaxes.

`h = andrewsplot(X,...)` returns a column vector of handles to the `Line` objects created by `andrewsplot`, one handle per row of `X`. If you use the `'Quantile'` input parameter, `h` contains one

handle for each of the three `Line` objects created. If you use both the `'Quantile'` and the `'Group'` input parameters, `h` contains three handles for each group.

Examples

Create Andrews Plot to Visualize Grouped Data

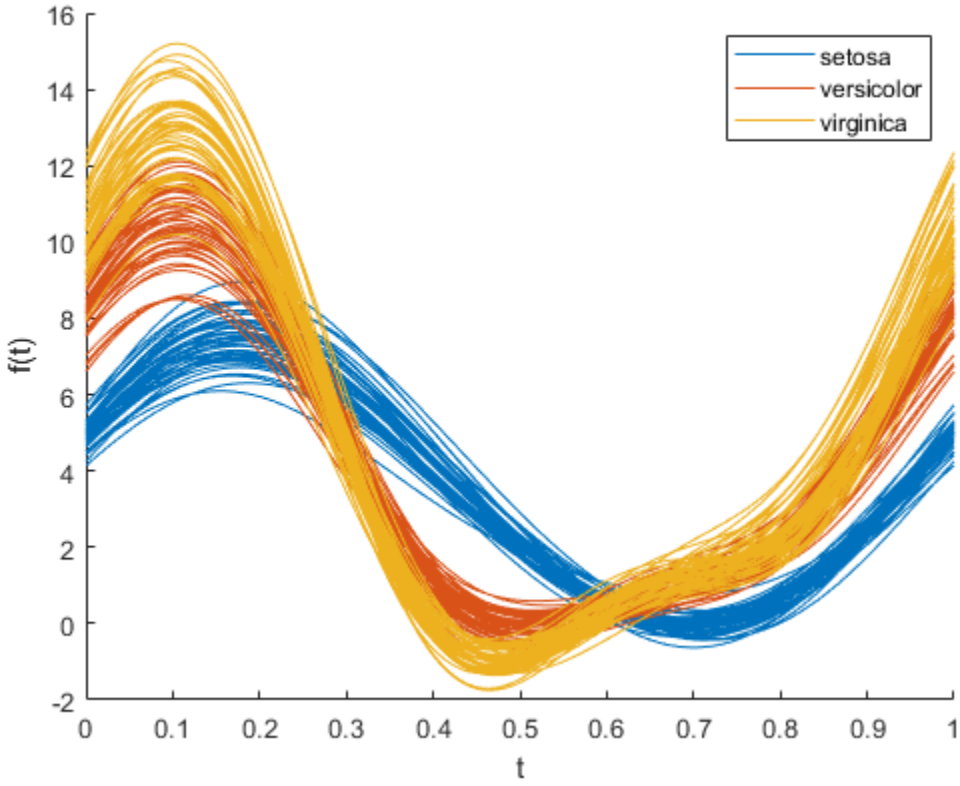
This example shows how to create an Andrews plot to visualize grouped sample data.

Load the sample data.

```
load fisheriris
```

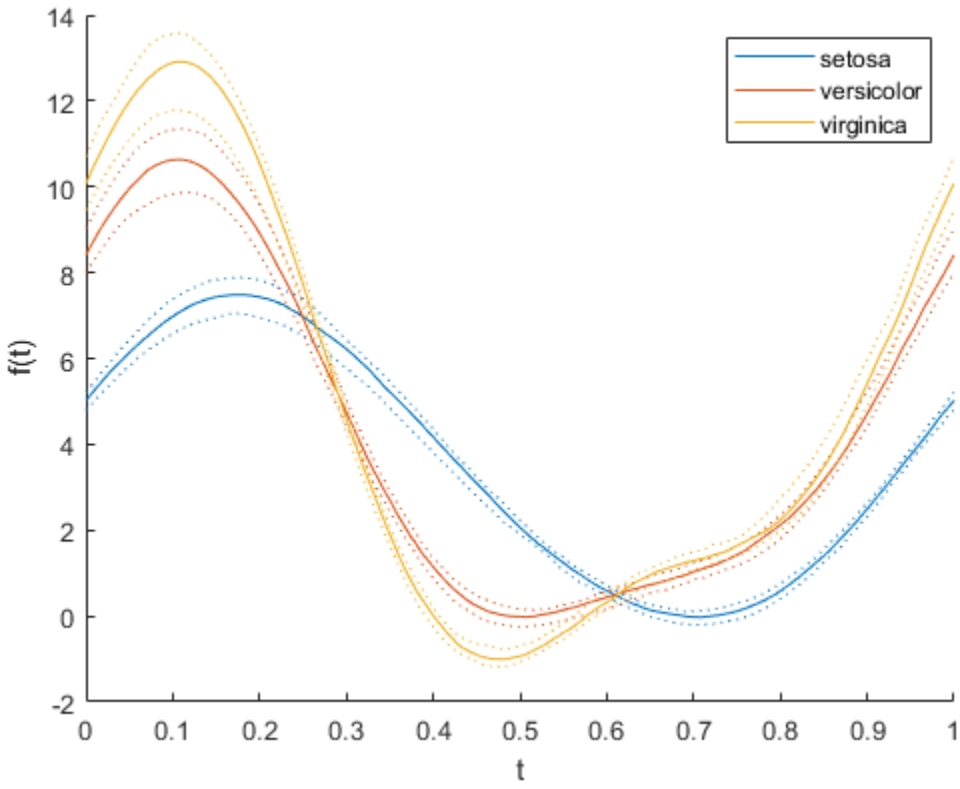
Create an Andrews plot, grouping the sample data by `species`.

```
andrewsplot(meas,'group',species)
```



Create a second, simplified Andrews plot that only displays the median and quartiles of each group.

```
andrewsplot(meas,'group',species,'quantile',.25)
```



See Also

`glyphplot` | `parallelcoords`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

anova

Package:

Analysis of variance for linear regression model

Syntax

```
tbl = anova(mdl)
tbl = anova(mdl,anovatype)
tbl = anova(mdl,'component',sstype)
```

Description

`tbl = anova(mdl)` returns a table with component ANOVA statistics.

`tbl = anova(mdl,anovatype)` returns ANOVA statistics of the specified type `anovatype`. For example, specify `anovatype` as `'component'`(default) to return a table with component ANOVA statistics, or specify `anovatype` as `'summary'` to return a table with summary ANOVA statistics.

`tbl = anova(mdl,'component',sstype)` computes component ANOVA statistics using the specified type of sum of squares.

Examples

Component ANOVA Table

Create a component ANOVA table from a linear regression model of the `hospital` data set.

Load the `hospital` data set and create a model of blood pressure as a function of age and gender.

```
load hospital
tbl = table(hospital.Age,hospital.Sex,hospital.BloodPressure(:,2), ...
    'VariableNames',{'Age','Sex','BloodPressure'});
tbl.Sex = categorical(tbl.Sex);
mdl = fitlm(tbl,'BloodPressure ~ Sex + Age^2')
```

```
mdl =
Linear regression model:
    BloodPressure ~ 1 + Age + Sex + Age^2

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   _________    ________    ________    _________

    (Intercept)       63.942      19.194      3.3314    0.0012275
    Age              0.90673      1.0442     0.86837      0.38736
    Sex_Male          3.0019      1.3765      2.1808     0.031643
    Age^2          -0.011275    0.013853    -0.81389      0.41772


Number of observations: 100, Error degrees of freedom: 96
```

```
Root Mean Squared Error: 6.83
R-squared: 0.0577,  Adjusted R-Squared: 0.0283
F-statistic vs. constant model: 1.96, p-value = 0.125
```

Create an ANOVA table of the model.

```
tbl = anova(mdl)
```

```
tbl=4×5 table
             SumSq     DF    MeanSq       F        pValue
             ______    __    ______    _______    ________

    Age      18.705     1    18.705    0.40055     0.52831
    Sex      222.09     1    222.09     4.7558    0.031643
    Age^2    30.934     1    30.934    0.66242     0.41772
    Error    4483.1    96    46.699
```

The table displays the following columns for each term except the constant (intercept) term:

- `SumSq` — Sum of squares explained by the term.
- `DF` — Degrees of freedom. In this example, `DF` is 1 for each term in the model and $n - p$ for the error term, where n is the number of observations and p is the number of coefficients (including the intercept) in the model. For example, the `DF` for the error term in this model is 100 – 4 = 96. If any variable in the model is a categorical variable, the `DF` for that variable is the number of indicator variables created for its categories (number of categories – 1).
- `MeanSq` — Mean square, defined by `MeanSq = SumSq/DF`. For example, the mean square of the error term, mean squared error (MSE), is 4.4831e+03/96 = 46.6991.
- `F` — F-statistic value to test the null hypothesis that the corresponding coefficient is zero, computed by `F = MeanSq/MSE`, where `MSE` is the mean squared error. When the null hypothesis is true, the F-statistic follows the F-distribution. The numerator degrees of freedom is the `DF` value for the corresponding term, and the denominator degrees of freedom is $n - p$. In this example, each F-statistic follows an $F_{(1,96)}$-distribution.
- `pValue` — p-value of the F-statistic value. For example, the p-value for `Age` is 0.5283, implying that `Age` is not significant at the 5% significance level given the other terms in the model.

Summary ANOVA Table

Create a summary ANOVA table from a linear regression model of the `hospital` data set.

Load the `hospital` data set and create a model of blood pressure as a function of age and gender.

```
load hospital
tbl = table(hospital.Age,hospital.Sex,hospital.BloodPressure(:,2), ...
    'VariableNames',{'Age','Sex','BloodPressure'});
tbl.Sex = categorical(tbl.Sex);
mdl = fitlm(tbl,'BloodPressure ~ Sex + Age^2')
```

```
mdl =
Linear regression model:
    BloodPressure ~ 1 + Age + Sex + Age^2

Estimated Coefficients:
```

```
                  Estimate        SE        tStat       pValue
                  _________    _________    _________    _________

    (Intercept)       63.942       19.194       3.3314    0.0012275
    Age              0.90673       1.0442      0.86837      0.38736
    Sex_Male          3.0019       1.3765       2.1808     0.031643
    Age^2          -0.011275    0.013853    -0.81389      0.41772


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 6.83
R-squared: 0.0577,  Adjusted R-Squared: 0.0283
F-statistic vs. constant model: 1.96, p-value = 0.125
```

Create a summary ANOVA table of the model.

```
tbl = anova(mdl,'summary')
```

```
tbl=7×5 table
                      SumSq     DF    MeanSq       F         pValue
                      ______    __    ______    _______    ________

    Total             4757.8    99    48.059
    Model             274.73     3    91.577      1.961     0.12501
    . Linear           243.8     2     121.9     2.6103    0.078726
    . Nonlinear       30.934     1    30.934    0.66242     0.41772
    Residual          4483.1    96    46.699
    . Lack of fit     1483.1    39    38.028    0.72253     0.85732
    . Pure error        3000    57    52.632
```

The table displays tests for groups of terms: `Total`, `Model`, and `Residual`.

- `Total` — This row shows the total sum of squares (`SumSq`), degrees of freedom (`DF`), and the mean squared error (`MeanSq`). Note that `MeanSq = SumSq/DF`.
- `Model` — This row includes `SumSq`, `DF`, `MeanSq`, *F*-statistic value (`F`), and *p*-value (`pValue`). Because this model includes a nonlinear term (`Age^2`), `anova` partitions the sum of squares (`SumSq`) of `Model` into two parts: `SumSq` explained by the linear terms (`Age` and `Sex`) and `SumSq` explained by the nonlinear term (`Age^2`). The corresponding *F*-statistic values are for testing the significance of the linear terms and the nonlinear term as separate groups. The nonlinear group consists of the `Age^2` term only, so it has the same *p*-value as the `Age^2` term in the “Component ANOVA Table” on page 33-42.
- `Residual` — This row includes `SumSq`, `DF`, `MeanSq`, `F`, and `pValue`. Because the data set includes replications, `anova` partitions the residual `SumSq` into the part for the replications (`Pure error`) and the rest (`Lack of fit`). To test the lack of fit, `anova` computes the *F*-statistic value by comparing the model residuals to the model-free variance estimate computed on the replications. The *F*-statistic value shows no evidence of lack of fit.

Linear Regression with Categorical Predictor

Fit a linear regression model that contains a categorical predictor. Reorder the categories of the categorical predictor to control the reference level in the model. Then, use `anova` to test the significance of the categorical variable.

Model with Categorical Predictor

Load the `carsmall` data set and create a linear regression model of `MPG` as a function of `Model_Year`. To treat the numeric vector `Model_Year` as a categorical variable, identify the predictor using the `'CategoricalVars'` name-value pair argument.

```
load carsmall
mdl = fitlm(Model_Year,MPG,'CategoricalVars',1,'VarNames',{'Model_Year','MPG'})
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                    Estimate      SE      tStat       pValue
                    ________    ______    ______    __________

    (Intercept)        17.69     1.0328    17.127    3.2371e-30
    Model_Year_76     3.8839     1.4059    2.7625     0.0069402
    Model_Year_82      14.02     1.4369    9.7571    8.2164e-16


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

The model formula in the display, `MPG ~ 1 + Model_Year`, corresponds to

$$\text{MPG} = \beta_0 + \beta_1 I_{\text{Year} = 76} + \beta_2 I_{\text{Year} = 82} + \epsilon,$$

where $I_{\text{Year} = 76}$ and $I_{\text{Year} = 82}$ are indicator variables whose value is one if the value of `Model_Year` is 76 and 82, respectively. The `Model_Year` variable includes three distinct values, which you can check by using the `unique` function.

```
unique(Model_Year)
```

```
ans = 3×1

    70
    76
    82
```

`fitlm` chooses the smallest value in `Model_Year` as a reference level (`'70'`) and creates two indicator variables $I_{\text{Year} = 76}$ and $I_{\text{Year} = 82}$. The model includes only two indicator variables because the design matrix becomes rank deficient if the model includes three indicator variables (one for each level) and an intercept term.

Model with Full Indicator Variables

You can interpret the model formula of `mdl` as a model that has three indicator variables without an intercept term:

$$y = \beta_0 I_{x_1 = 70} + (\beta_0 + \beta_1) I_{x_1 = 76} + (\beta_0 + \beta_2) I_{x_2 = 82} + \epsilon.$$

Alternatively, you can create a model that has three indicator variables without an intercept term by manually creating indicator variables and specifying the model formula.

```
temp_Year = dummyvar(categorical(Model_Year));
Model_Year_70 = temp_Year(:,1);
Model_Year_76 = temp_Year(:,2);
Model_Year_82 = temp_Year(:,3);
tbl = table(Model_Year_70,Model_Year_76,Model_Year_82,MPG);
mdl = fitlm(tbl,'MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82 - 1')
```

```
mdl =
Linear regression model:
    MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82

Estimated Coefficients:
                     Estimate      SE       tStat       pValue
                     ________    _______    ______    __________

    Model_Year_70      17.69      1.0328    17.127    3.2371e-30
    Model_Year_76     21.574     0.95387    22.617    4.0156e-39
    Model_Year_82      31.71     0.99896    31.743    5.2234e-51


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
```

Choose Reference Level in Model

You can choose a reference level by modifying the order of categories in a categorical variable. First, create a categorical variable `Year`.

```
Year = categorical(Model_Year);
```

Check the order of categories by using the `categories` function.

```
categories(Year)
```

```
ans = 3x1 cell
    {'70'}
    {'76'}
    {'82'}
```

If you use `Year` as a predictor variable, then `fitlm` chooses the first category `'70'` as a reference level. Reorder `Year` by using the `reordercats` function.

```
Year_reordered = reordercats(Year,{'76','70','82'});
categories(Year_reordered)
```

```
ans = 3x1 cell
    {'76'}
    {'70'}
    {'82'}
```

The first category of `Year_reordered` is `'76'`. Create a linear regression model of `MPG` as a function of `Year_reordered`.

```
mdl2 = fitlm(Year_reordered,MPG,'VarNames',{'Model_Year','MPG'})
```

```
mdl2 =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                    Estimate      SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)       21.574     0.95387     22.617    4.0156e-39
    Model_Year_70    -3.8839      1.4059    -2.7625     0.0069402
    Model_Year_82     10.136      1.3812     7.3385    8.7634e-11


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

`mdl2` uses `'76'` as a reference level and includes two indicator variables $I_{\text{Year = 70}}$ and $I_{\text{Year = 82}}$.

Evaluate Categorical Predictor

The model display of `mdl2` includes a *p*-value of each term to test whether or not the corresponding coefficient is equal to zero. Each *p*-value examines each indicator variable. To examine the categorical variable `Model_Year` as a group of indicator variables, use `anova`. Use the `'components'`(default) option to return a component ANOVA table that includes ANOVA statistics for each variable in the model except the constant term.

```
anova(mdl2,'components')
```

```
ans=2×5 table
                SumSq     DF    MeanSq      F        pValue
                ______    __    ______    _____    __________

    Model_Year    3190.1     2    1595.1    51.56    1.0694e-15
    Error         2815.2    91    30.936
```

The component ANOVA table includes the *p*-value of the `Model_Year` variable, which is smaller than the *p*-values of the indicator variables.

Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`anovatype` — ANOVA type

`'component'` (default) | `'summary'`

ANOVA type, specified as one of these values:

- `'component'` — `anova` returns the table `tbl` with ANOVA statistics for each variable in the model except the constant term.

- `'summary'` — `anova` returns the table `tbl` with summary ANOVA statistics for grouped variables and the model as a whole.

For details, see the `tbl` output argument description.

`sstype` — Sum of squares type
`'h'` (default) | 1 | 2 | 3

Sum of squares type for each term, specified as one of the values in this table.

Value	Description
1	Type 1 sum of squares — Reduction in residual sum of squares obtained by adding the term to a fit that already includes the preceding terms
2	Type 2 sum of squares — Reduction in residual sum of squares obtained by adding the term to a model that contains all other terms
3	Type 3 sum of squares — Reduction in residual sum of squares obtained by adding the term to a model that contains all other terms, but with their effects constrained to obey the usual “sigma restrictions” that make models estimable
`'h'`	Hierarchical model — Similar to Type 2, but uses both continuous and categorical factors to determine the hierarchy of terms

The sum of squares for any term is determined by comparing two models. For a model containing main effects but no interactions, the value of `sstype` influences the computations on unbalanced data only.

Suppose you are fitting a model with two factors and their interaction, and the terms appear in the order A, B, AB. Let $R(\cdot)$ represent the residual sum of squares for the model. So, $R(A, B, AB)$ is the residual sum of squares fitting the whole model, $R(A)$ is the residual sum of squares fitting the main effect of A only, and $R(1)$ is the residual sum of squares fitting the mean only. The three sum of squares types are as follows:

Term	Type 1 Sum of Squares	Type 2 Sum of Squares	Type 3 Sum of Squares
A	$R(1) - R(A)$	$R(B) - R(A, B)$	$R(B, AB) - R(A, B, AB)$
B	$R(A) - R(A, B)$	$R(A) - R(A, B)$	$R(A, AB) - R(A, B, AB)$
AB	$R(A, B) - R(A, B, AB)$	$R(A, B) - R(A, B, AB)$	$R(A, B) - R(A, B, AB)$

The models for Type 3 sum of squares have sigma restrictions imposed. This means, for example, that in fitting $R(B, AB)$, the array of AB effects is constrained to sum to 0 over A for each value of B, and over B for each value of A.

For Type 3 sum of squares:

- If `mdl` is a `CompactLinearModel` object and the regression model is nonhierarchical, `anova` returns an error.
- If `mdl` is a `LinearModel` object and the regression model is nonhierarchical, `anova` refits the model using effects coding whenever it needs to compute a Type 3 sum of squares.

- If the regression model in `mdl` is hierarchical, `anova` computes the results without refitting the model.

`sstype` applies only if `anovatype` is `'component'`.

Output Arguments

`tbl` — ANOVA summary statistics table

table

ANOVA summary statistics table, returned as a table.

The contents of `tbl` depend on the ANOVA type specified in `anovatype`.

- If `anovatype` is `'component'`, then `tbl` contains ANOVA statistics for each variable in the model except the constant (intercept) term. The table includes these columns for each variable:

Column	Description
`SumSq`	Sum of squares explained by the term, computed depending on `sstype`
`DF`	Degrees of freedom • `DF` of a numeric variable is 1. • `DF` of a categorical variable is the number of indicator variables created for the category (number of categories – 1). Note that `tbl` contains one row for each categorical variable instead of one row for each indicator variable as in the model display. Use `anova` to test a categorical variable as a group of indicator variables. • `DF` of an error term is $n – p$, where n is the number of observations and p is the number of coefficients in the model.
`MeanSq`	Mean square, defined by `MeanSq = SumSq/DF` `MeanSq` for the error term is the mean squared error (MSE).
`F`	F-statistic value to test the null hypothesis that the corresponding coefficient is zero, computed by `F = MeanSq/MSE` When the null hypothesis is true, the F-statistic follows the F-distribution. The numerator degrees of freedom is the `DF` value for the corresponding term, and the denominator degrees of freedom is $n – p$.
`pValue`	p-value of the F-statistic value

  For an example, see “Component ANOVA Table” on page 33-42.

- If `anovatype` is `'summary'`, then `tbl` contains summary statistics of grouped terms for each row. The table includes the same columns as `'component'` and these rows:

Row	Description
`Total`	Total statistics • `SumSq` — Total sum of squares, which is the sum of the squared deviations of the response around its mean • `DF` — Sum of degrees of freedom of `Model` and `Residual`
`Model`	Statistics for the model as a whole • `SumSq` — Model sum of squares, which is the sum of the squared deviations of the fitted value around the response mean. • `F` and `pValue` — These values provide a test of whether the model as a whole fits significantly better than a degenerate model consisting of only a constant term. If `mdl` includes only linear terms, then `anova` does not decompose `Model` into `Linear` and `NonLinear`.
`Linear`	Statistics for linear terms • `SumSq` — Sum of squares for linear terms, which is the difference between the model sum of squares and the sum of squares for nonlinear terms. • `F` and `pValue` — These values provide a test of whether the model with only linear terms fits better than a degenerate model consisting of only a constant term. `anova` uses the mean squared error that is based on the full model to compute this *F*-value, so the *F*-value obtained by dropping the nonlinear terms and repeating the test is not the same as the value in this row.
`Nonlinear`	Statistics for nonlinear terms • `SumSq` — Sum of squares for nonlinear (higher-order or interaction) terms, which is the increase in the residual sum of squares obtained by keeping only the linear terms and dropping all nonlinear terms. • `F` and `pValue` — These values provide a test of whether the full model fits significantly better than a smaller model consisting of only the linear terms.

Row	Description
`Residual`	Statistics for residuals • `SumSq` — Residual sum of squares, which is the sum of the squared residual values • `MeanSq` — Mean squared error, used to compute the *F*-statistic values for `Model`, `Linear`, and `NonLinear` If `mdl` is a full `LinearModel` object and the sample data contains replications (multiple observations sharing the same predictor values), then `anova` decomposes the residual sum of squares into a sum of squares for the replicated observations (`Lack of fit`) and the remaining sum of squares (`Pure error`).
`Lack of fit`	Lack-of-fit statistics • `SumSq` — Sum of squares due to lack of fit, which is the difference between the residual sum of squares and the replication sum of squares. • `F` and `pValue` — The *F*-statistic value is the ratio of lack-of-fit `MeanSq` to pure error `MeanSq`. The ratio provides a test of bias by measuring whether the variation of the residuals is larger than the variation of the replications. A low *p*-value implies that adding additional terms to the model can improve the fit.
`Pure error`	Statistics for pure error • `SumSq` — Replication sum of squares, obtained by finding the sets of points with identical predictor values, computing the sum of squared deviations around the mean within each set, and pooling the computed values • `MeanSq` — Model-free pure error variance estimate of the response

For an example, see “Summary ANOVA Table” on page 33-43.

Alternative Functionality

More complete ANOVA statistics are available in the `anova1`, `anova2`, and `anovan` functions.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `coefCI` | `coefTest` | `dwtest`

Topics

“F-statistic and t-statistic” on page 11-72
“Interpret Linear Regression Results” on page 11-50
“Linear Regression with Categorical Covariates” on page 2-52
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

anova

Class: `LinearMixedModel`

Analysis of variance for linear mixed-effects model

Syntax

```
stats = anova(lme)
stats = anova(lme,Name,Value)
```

Description

`stats = anova(lme)` returns the dataset array `stats` that includes the results of the *F*-tests for each fixed-effects term in the linear mixed-effects model `lme`.

`stats = anova(lme,Name,Value)` also returns the dataset array `stats` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`DFMethod` — Method for computing approximate degrees of freedom

`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing approximate degrees of freedom to use in the F-test, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

Output Arguments

`stats` — Results of *F*-tests for fixed-effects terms

dataset array

Results of *F*-tests for fixed-effects terms, returned as a dataset array with the following columns.

`Term`	Name of the fixed effects term
`Fstat`	*F*-statistic for the term
`DF1`	Numerator degrees of freedom for the *F*-statistic
`DF2`	Denominator degrees of freedom for the *F*-statistic
`pValue`	*p*-value of the test for the term

There is one row for each fixed-effects term. Each term is a continuous variable, a grouping variable, or an interaction between two or more continuous or grouping variables. For each fixed-effects term, `anova` performs an *F*-test (marginal test) to determine if all coefficients representing the fixed-effects term are 0. To perform tests for the type III hypothesis, you must use the `'effects'` contrasts while fitting the linear mixed-effects model.

Examples

F-Tests for Fixed Effects

Load the sample data.

```
load('shift.mat')
```

The data shows the deviations from the target quality characteristic measured from the products that five operators manufacture during three shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the deviation of the quality characteristics from the target value. This is simulated data.

`Shift` and `Operator` are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if performance significantly differs according to the time of the shift. Use the restricted maximum likelihood method and `'effects'` contrasts.

`'effects'` contrasts indicate that the coefficients sum to 0, and `fitlme` creates two contrast-coded variables in the fixed-effects design matrix, X1 and X2, where

$$Shift\_Evening = \begin{cases} 0, & if\ Morning \\ 1, & if\ Evening \\ -1, & if\ Night \end{cases}$$

and

$$Shift\_Morning = \begin{cases} 1, & if\ Morning \\ 0, & if\ Evening \\ -1, & if\ Night \end{cases}.$$

The model corresponds to

$$\text{MorningShift: } QCDev_{im} = \beta_0 + \beta_2 Shift\_Morning_i + b_{0m} + \varepsilon_{im}, \quad m = 1, 2, \ldots, 5,$$
$$\text{EveningShift: } QCDev_{im} = \beta_0 + \beta_1 Shift\_Evening_i + b_{0m} + \varepsilon_{im},$$
$$\text{NightShift: } \quad QCDev_{im} = \beta_0 - \beta_1 Shift\_Evening_i - \beta_2 Shift\_Morning_i + b_{0m} + \varepsilon_{im},$$

where $b \sim N(0, \sigma_b^2)$ and $\epsilon \sim N(0, \sigma^2)$.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)',...
'FitMethod','REML','DummyVarCoding','effects')
```

```
lme =
Linear mixed-effects model fit by REML

Model information:
    Number of observations             15
    Fixed effects coefficients          3
    Random effects coefficients         5
    Covariance parameters               2

Formula:
    QCDev ~ 1 + Shift + (1 | Operator)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    58.913    61.337    -24.456          48.913

Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE         tStat      DF    pValue
    {'(Intercept)'  }         3.6525    0.94109     3.8812    12    0.0021832
    {'Shift_Evening'}        -0.53293   0.31206    -1.7078    12      0.11339
    {'Shift_Morning'}        -0.91973   0.31206    -2.9473    12     0.012206


    Lower      Upper
     1.6021       5.703
    -1.2129     0.14699
    -1.5997    -0.23981

Random effects covariance parameters (95% CIs):
Group: Operator (5 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       2.0457


    Lower      Upper
    0.98207    4.2612

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.85462     0.52357    1.395
```

Perform an *F*-test to determine if all fixed-effects coefficients are 0.

```
anova(lme)
```

```
ans =
    ANOVA marginal tests: DFMethod = 'Residual'

    Term                   FStat     DF1    DF2    pValue
    {'(Intercept)'}        15.063    1      12     0.0021832
    {'Shift'      }        11.091    2      12     0.0018721
```

The *p*-value for the constant term, 0.0021832, is the same as in the coefficient table in the `lme` display. The *p*-value of 0.0018721 for `Shift` measures the combined significance for both coefficients representing `Shift`.

ANOVA for Fixed-Effects in LME Model

Load the sample data.

```
load('fertilizer.mat')
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type) and the plots within blocks (tomato types within soil types) independently. Use the `'effects'` contrasts when fitting the data for the type III sum of squares.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)',...
'DummyVarCoding','effects')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             60
    Fixed effects coefficients         20
    Random effects coefficients        18
    Covariance parameters               3

Formula:
    Yield ~ 1 + Tomato*Fertilizer + (1 | Soil) + (1 | Soil:Tomato)
```

```
Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    522.57    570.74    -238.29          476.57

Fixed effects coefficients (95% CIs):
    Name                                    Estimate    SE        tStat       DF
    {'(Intercept)'                 }           104.6    3.3008       31.69    40
    {'Tomato_Cherry'               }             1.4    5.9353     0.23588    40
    {'Tomato_Grape'                }          -7.7667    5.9353     -1.3085    40
    {'Tomato_Heirloom'             }          -11.183    5.9353     -1.8842    40
    {'Tomato_Plum'                 }           30.233    5.9353      5.0938    40
    {'Fertilizer_1'                }          -28.267    2.3475     -12.041    40
    {'Fertilizer_2'                }          -1.9333    2.3475    -0.82356    40
    {'Fertilizer_3'                }           10.733    2.3475      4.5722    40
    {'Tomato_Cherry:Fertilizer_1'  }         -0.73333    4.6951    -0.15619    40
    {'Tomato_Grape:Fertilizer_1'   }          -7.5667    4.6951     -1.6116    40
    {'Tomato_Heirloom:Fertilizer_1'}           5.1833    4.6951       1.104    40
    {'Tomato_Plum:Fertilizer_1'    }           2.7667    4.6951     0.58927    40
    {'Tomato_Cherry:Fertilizer_2'  }             7.6    4.6951      1.6187    40
    {'Tomato_Grape:Fertilizer_2'   }            -1.9    4.6951    -0.40468    40
    {'Tomato_Heirloom:Fertilizer_2'}           5.5167    4.6951       1.175    40
    {'Tomato_Plum:Fertilizer_2'    }            -3.9    4.6951    -0.83066    40
    {'Tomato_Cherry:Fertilizer_3'  }          -6.0667    4.6951     -1.2921    40
    {'Tomato_Grape:Fertilizer_3'   }           3.7667    4.6951     0.80226    40
    {'Tomato_Heirloom:Fertilizer_3'}           3.1833    4.6951     0.67802    40
    {'Tomato_Plum:Fertilizer_3'    }              1.1    4.6951     0.23429    40


    pValue        Lower      Upper
    5.9086e-30     97.929     111.27
       0.81473    -10.596     13.396
       0.19816    -19.762     4.2291
      0.066821    -23.179    0.81242
     8.777e-06     18.238     42.229
    7.0265e-15    -33.011    -23.522
       0.41507    -6.6779     2.8112
     4.577e-05     5.9888     15.478
       0.87667    -10.222     8.7558
       0.11491    -17.056     1.9224
       0.27619    -4.3058     14.672
       0.55899    -6.7224     12.256
       0.11337    -1.8891     17.089
       0.68787    -11.389     7.5891
       0.24695    -3.9724     15.006
        0.4111    -13.389     5.5891
       0.20373    -15.556     3.4224
       0.42714    -5.7224     13.256
       0.50167    -6.3058     12.672
       0.81596    -8.3891     10.589

Random effects covariance parameters (95% CIs):
Group: Soil (3 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        2.5028


    Lower      Upper
```

```
    0.02771    226.05

Group: Soil:Tomato (15 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        10.225


    Lower     Upper
    6.1497    17.001

Group: Error
    Name               Estimate    Lower     Upper
    {'Res Std'}        10.499      8.5389    12.908
```

Perform an analysis of variance to test for the fixed-effects.

```
anova(lme)
```

```
ans =
    ANOVA marginal tests: DFMethod = 'Residual'

    Term                         FStat     DF1    DF2    pValue
    {'(Intercept)'      }        1004.2     1     40     5.9086e-30
    {'Tomato'           }        7.1663     4     40     0.00018935
    {'Fertilizer'       }        58.833     3     40     1.0024e-14
    {'Tomato:Fertilizer'}        1.4182    12     40        0.19804
```

The *p*-value for the constant term, 5.9086e-30, is the same as in the coefficient table in the `lme` display. The *p*-values of 0.00018935, 1.0024e-14, and 0.19804 for `Tomato`, `Fertilizer`, and `Tomato:Fertilizer` represent the combined significance for all tomato coefficients, fertilizer coefficients, and coefficients representing the interaction between the tomato and fertilizer, respectively. The *p*-value of 0.19804 indicates that the interaction between tomato and fertilizer is not significant.

Satterthwaite Approximation for Degrees of Freedom

Load the sample data.

```
load('weight.mat')
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit the model using the `'effects'` contrasts.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)',...
        'DummyVarCoding','effects')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             120
    Fixed effects coefficients           9
    Random effects coefficients         40
    Covariance parameters                4

Formula:
    y ~ 1 + InitialWeight + Program*Week + (1 + Week | Subject)

Model fit statistics:
    AIC        BIC       LogLikelihood    Deviance
    -22.981    13.257    24.49            -48.981

Fixed effects coefficients (95% CIs):
    Name                       Estimate     SE           tStat       DF
    {'(Intercept)'   }          0.77122      0.24309      3.1725    111
    {'InitialWeight' }        0.0031879    0.0013814      2.3078    111
    {'Program_A'     }         -0.11017     0.080377     -1.3707    111
    {'Program_B'     }          0.25061      0.08045      3.1151    111
    {'Program_C'     }         -0.14344     0.080475     -1.7824    111
    {'Week'          }          0.19881     0.033727      5.8946    111
    {'Program_A:Week'}        -0.025607     0.058417    -0.43835    111
    {'Program_B:Week'}         0.013164     0.058417     0.22535    111
    {'Program_C:Week'}        0.0049357     0.058417    0.084492    111


    pValue        Lower         Upper
     0.0019549       0.28951       1.2529
      0.022863    0.00045067    0.0059252
       0.17323      -0.26945       0.0491
     0.0023402      0.091195      0.41003
      0.077424       -0.3029     0.016031
    4.1099e-08       0.13198      0.26564
       0.66198      -0.14136     0.090149
       0.82212      -0.10259      0.12892
       0.93282      -0.11082      0.12069

Random effects covariance parameters (95% CIs):
Group: Subject (20 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std' }        0.18407
    {'Week'       }        {'(Intercept)'}        {'corr'}        0.66841
    {'Week'       }        {'Week'       }        {'std' }        0.15033


    Lower      Upper
    0.12281    0.27587
    0.21077    0.88573
    0.11004    0.20537

Group: Error
    Name               Estimate    Lower       Upper
    {'Res Std'}        0.10261     0.087882    0.11981
```

The *p*-values 0.022863 and 4.1099e-08 indicate significant effects of the initial weights of the subjects and the time factor in the amount of weight lost. The weight loss of subjects who are in program B is significantly different relative to the weight loss of subjects that are in program A. The lower and upper limits of the covariance parameters for the random effects do not include zero, thus they are significant.

Perform an F-test that all fixed-effects coefficients are zero.

```
anova(lme)
```

```
ans =
    ANOVA marginal tests: DFMethod = 'Residual'

    Term                    FStat       DF1    DF2    pValue
    {'(Intercept)'  }         10.065    1      111     0.0019549
    {'InitialWeight'}          5.326    1      111      0.022863
    {'Program'      }          3.6798   3      111      0.014286
    {'Week'         }         34.747    1      111    4.1099e-08
    {'Program:Week' }        0.066648   3      111       0.97748
```

The *p*-values for the constant term, initial weight, and week are the same as in the coefficient table in the previous `lme` output display. The *p*-value of 0.014286 for `Program` represents the combined significance for all program coefficients. Similarly, the *p*-value for the interaction between program and week (`Program:Week`) measures the combined significance for all coefficients representing this interaction.

Now, use the Satterthwaite method to compute the degrees of freedom.

```
anova(lme,'DFMethod','satterthwaite')
```

```
ans =
    ANOVA marginal tests: DFMethod = 'Satterthwaite'

    Term                    FStat       DF1    DF2       pValue
    {'(Intercept)'  }         10.065    1      20.445      0.004695
    {'InitialWeight'}          5.326    1          20      0.031827
    {'Program'      }          3.6798   3       19.14      0.030233
    {'Week'         }         34.747    1          20    9.1346e-06
    {'Program:Week' }        0.066648   3          20       0.97697
```

The Satterthwaite method produces smaller denominator degrees of freedom and slightly larger *p*-values.

Tips

- For each fixed-effects term, `anova` performs an *F*-test (marginal test), that all coefficients representing the fixed-effects term are 0. To perform tests for type III hypotheses, you must set the `'DummyVarCoding'` name-value pair argument to `'effects'` contrasts while fitting your linear mixed-effects model.

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

anova1

One-way analysis of variance

Syntax

```
p = anova1(y)
p = anova1(y,group)
p = anova1(y,group,displayopt)
[p,tbl] = anova1( ___ )
[p,tbl,stats] = anova1( ___ )
```

Description

`p = anova1(y)` performs one-way ANOVA on page 9-3 for the sample data `y` and returns the *p*-value. `anova1` treats each column of `y` as a separate group. The function tests the hypothesis that the samples in the columns of `y` are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. The function also displays the box plot on page 33-71 for each group in `y` and the standard ANOVA table (`tbl`).

`p = anova1(y,group)` performs one-way ANOVA for the sample data `y`, grouped by `group`.

`p = anova1(y,group,displayopt)` enables the ANOVA table and box plot displays when `displayopt` is `'on'` (default) and suppresses the displays when `displayopt` is `'off'`.

`[p,tbl] = anova1( ___ )` returns the ANOVA table (including column and row labels) in the cell array `tbl` using any of the input argument combinations in the previous syntaxes. To copy a text version of the ANOVA table to the clipboard, select **Edit > Copy Text** from the ANOVA table figure.

`[p,tbl,stats] = anova1( ___ )` returns a structure, `stats`, which you can use to perform a multiple comparison test on page 9-18. A multiple comparison test enables you to determine which pairs of group means are significantly different. To perform this test, use `multcompare`, providing the `stats` structure as an input argument.

Examples

One-Way ANOVA

Create sample data matrix `y` with columns that are constants, plus random normal disturbances with mean 0 and standard deviation 1.

```
y = meshgrid(1:5);
rng default; % For reproducibility
y = y + normrnd(0,1,5,5)
```

```
y = 5×5

    1.5377    0.6923    1.6501    3.7950    5.6715
    2.8339    1.5664    6.0349    3.8759    3.7925
   -1.2588    2.3426    3.7254    5.4897    5.7172
```

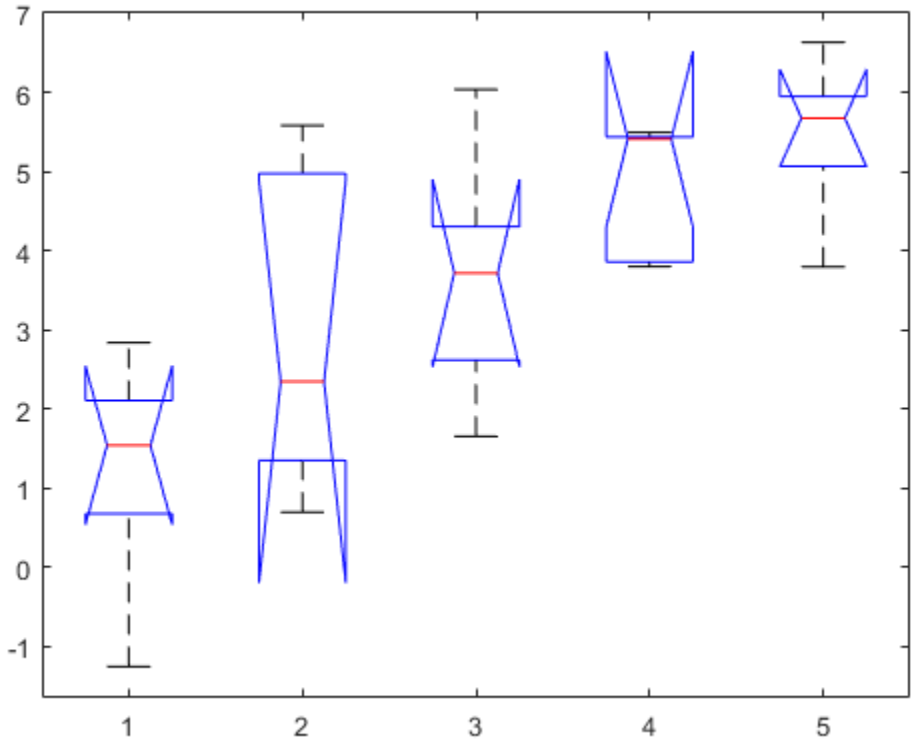
```
    1.8622    5.5784    2.9369    5.4090    6.6302
    1.3188    4.7694    3.7147    5.4172    5.4889
```

Perform one-way ANOVA.

```
p = anova1(y)
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	53.7238	4	13.4309	6.05	0.0023
Error	44.408	20	2.2204		
Total	98.1318	24			




```
p = 0.0023
```

The ANOVA table shows the between-groups variation (`Columns`) and within-groups variation (`Error`). `SS` is the sum of squares, and `df` is the degrees of freedom. The total degrees of freedom is total number of observations minus one, which is 25 - 1 = 24. The between-groups degrees of freedom is number of groups minus one, which is 5 - 1 = 4. The within-groups degrees of freedom is total degrees of freedom minus the between groups degrees of freedom, which is 24 - 4 = 20.

`MS` is the mean squared error, which is `SS/df` for each source of variation. The *F*-statistic is the ratio of the mean squared errors (13.4309/2.2204). The *p*-value is the probability that the test statistic can

take a value greater than the value of the computed test statistic, i.e., P(F > 6.05). The small *p*-value of 0.0023 indicates that differences between column means are significant.

Compare Beam Strength Using One-Way ANOVA

Input the sample data.

```
strength = [82 86 79 83 84 85 86 87 74 82 ...
            78 75 76 77 79 79 77 78 82 79];
alloy = {'st','st','st','st','st','st','st','st',...
         'al1','al1','al1','al1','al1','al1',...
         'al2','al2','al2','al2','al2','al2'};
```

The data are from a study of the strength of structural beams in Hogg (1987). The vector strength measures deflections of beams in thousandths of an inch under 3000 pounds of force. The vector alloy identifies each beam as steel (`'st'`), alloy 1 (`'al1'`), or alloy 2 (`'al2'`). Although alloy is sorted in this example, grouping variables do not need to be sorted.

Test the null hypothesis that the steel beams are equal in strength to the beams made of the two more expensive alloys. Turn the figure display off and return the ANOVA results in a cell array.

```
[p,tbl] = anova1(strength,alloy,'off')
```

```
p = 1.5264e-04

tbl=4×6 cell array
  Columns 1 through 5

    {'Source'}    {'SS'      }    {'df'}    {'MS'      }    {'F'       }
    {'Groups'}    {[184.8000]}    {[ 2]}    {[ 92.4000]}    {[ 15.4000]}
    {'Error' }    {[102.0000]}    {[17]}    {[  6.0000]}    {0x0 double}
    {'Total' }    {[286.8000]}    {[19]}    {0x0 double}    {0x0 double}

  Column 6

    {'Prob>F'    }
    {[1.5264e-04]}
    {0x0 double  }
    {0x0 double  }
```

The total degrees of freedom is total number of observations minus one, which is 20 − 1 = 19. The between-groups degrees of freedom is number of groups minus one, which is 3 − 1 = 2. The within-groups degrees of freedom is total degrees of freedom minus the between groups degrees of freedom, which is 19 − 2 = 17.

`MS` is the mean squared error, which is `SS/df` for each source of variation. The *F*-statistic is the ratio of the mean squared errors. The *p*-value is the probability that the test statistic can take a value greater than or equal to the value of the test statistic. The *p*-value of 1.5264e-04 suggests rejection of the null hypothesis.

You can retrieve the values in the ANOVA table by indexing into the cell array. Save the *F*-statistic value and the *p*-value in the new variables `Fstat` and `pvalue`.

```
Fstat = tbl{2,5}
```

```
Fstat = 15.4000

pvalue = tbl{2,6}

pvalue = 1.5264e-04
```

Multiple Comparisons for One-Way ANOVA

Input the sample data.

```
strength = [82 86 79 83 84 85 86 87 74 82 ...
            78 75 76 77 79 79 77 78 82 79];
alloy = {'st','st','st','st','st','st','st','st',...
         'al1','al1','al1','al1','al1','al1',...
         'al2','al2','al2','al2','al2','al2'};
```

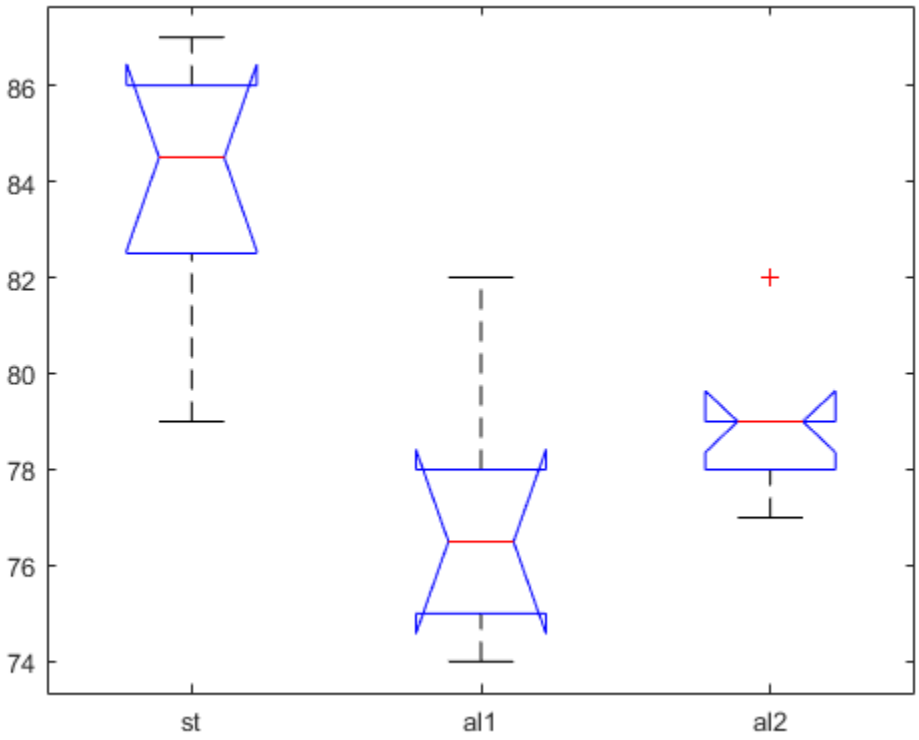
The data are from a study of the strength of structural beams in Hogg (1987). The vector strength measures deflections of beams in thousandths of an inch under 3000 pounds of force. The vector alloy identifies each beam as steel (`st`), alloy 1 (`al1`), or alloy 2 (`al2`). Although alloy is sorted in this example, grouping variables do not need to be sorted.

Perform one-way ANOVA using `anova1`. Return the structure `stats`, which contains the statistics `multcompare` needs for performing “Multiple Comparisons” on page 9-18.

```
[~,~,stats] = anova1(strength,alloy);
```

ANOVA Table

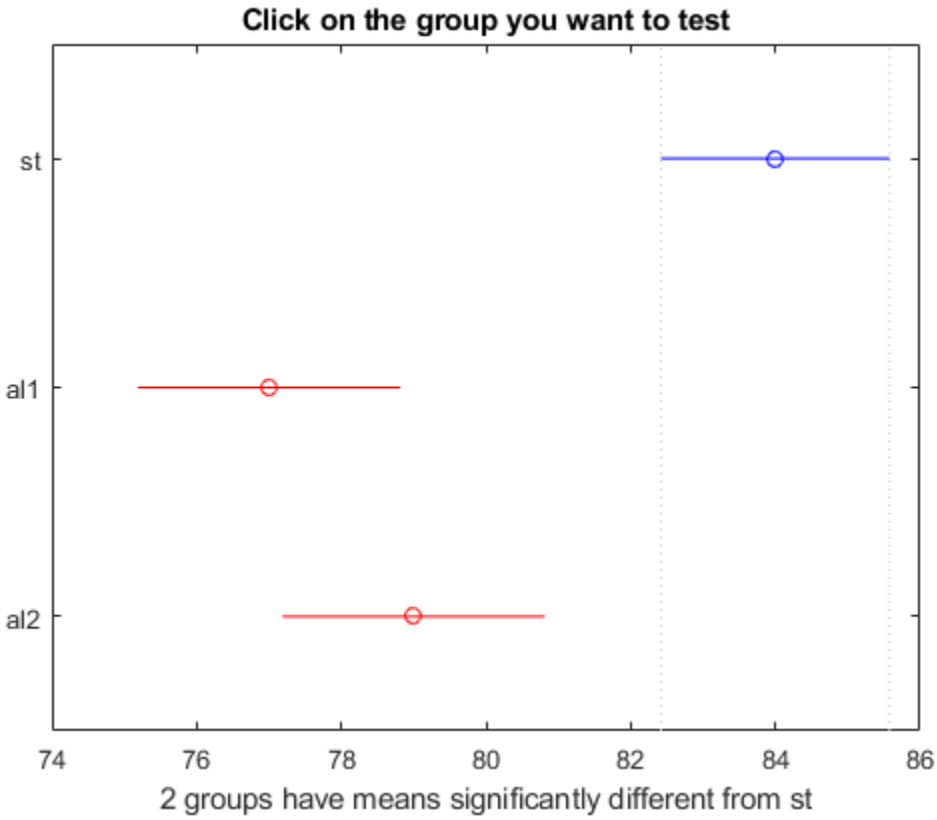
Source	SS	df	MS	F	Prob>F
Groups	184.8	2	92.4	15.4	0.0002
Error	102	17	6		
Total	286.8	19			



The small *p*-value of 0.0002 suggests that the strength of the beams is not the same.

Perform a multiple comparison of the mean strength of the beams.

```
[c,~,~,gnames] = multcompare(stats);
```



Display the comparison results with the corresponding group names.

```
[gnames(c(:,1)), gnames(c(:,2)), num2cell(c(:,3:6))]
```

```
ans=3×6 cell array
  Columns 1 through 5

    {'st' }    {'al1'}    {[ 3.6064]}    {[ 7]}    {[10.3936]}
    {'st' }    {'al2'}    {[ 1.6064]}    {[ 5]}    {[ 8.3936]}
    {'al1'}    {'al2'}    {[-5.6280]}    {[-2]}    {[ 1.6280]}

  Column 6

    {[1.6831e-04]}
    {[    0.0040]}
    {[    0.3560]}
```

The first two columns show the pair of groups that are compared. The fourth column shows the difference between the estimated group means. The third and fifth columns show the lower and upper limits for the 95% confidence intervals of the true difference of means. The sixth column shows the *p*-value for a hypothesis that the true difference of means for the corresponding groups is equal to zero.

The first two rows show that both comparisons involving the first group (steel) have confidence intervals that do not include zero. Because the corresponding *p*-values (1.6831e-04 and 0.0040, respectively) are small, those differences are significant.

The third row shows that the differences in strength between the two alloys is not significant. A 95% confidence interval for the difference is [-5.6,1.6], so you cannot reject the hypothesis that the true difference is zero. The corresponding *p*-value of 0.3560 in the sixth column confirms this result.

In the figure, the blue bar represents the comparison interval for mean material strength for steel. The red bars represent the comparison intervals for the mean material strength for alloy 1 and alloy 2. Neither of the red bars overlap with the blue bar, which indicates that the mean material strength for steel is significantly different from that of alloy 1 and alloy 2. To confirm the significant difference by clicking the bars that represent alloy 1 and 2.

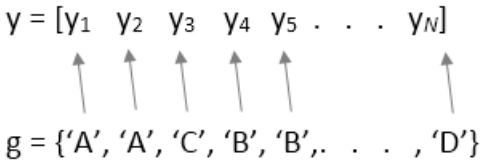
Input Arguments

y — sample data

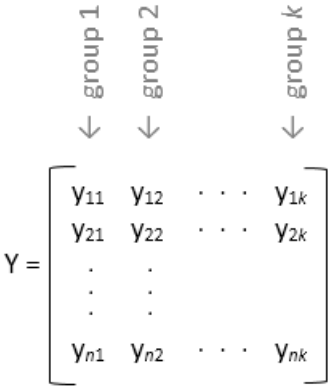
vector | matrix

Sample data, specified as a vector or matrix.

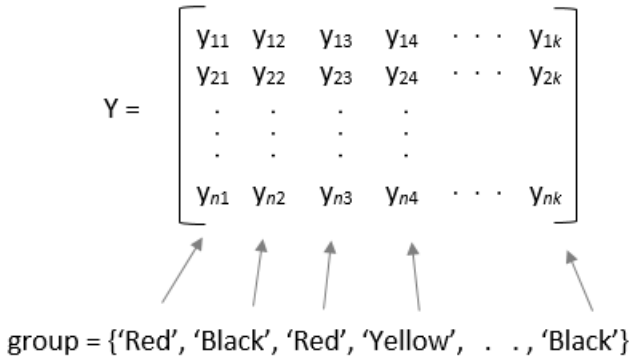
- If `y` is a vector, you must specify the `group` input argument. Each element in `group` represents a group name of the corresponding element in `y`. The `anova1` function treats the `y` values corresponding to the same value of `group` as part of the same group. Use this design when groups have different numbers of elements (unbalanced ANOVA).



- If `y` is a matrix and you do not specify `group`, then `anova1` treats each column of `y` as a separate group. In this design, the function evaluates whether the population means of the columns are equal. Use this design when each group has the same number of elements (balanced ANOVA).



- If `y` is a matrix and you specify `group`, then each element in `group` represents a group name for the corresponding column in `y`. The `anova1` function treats the columns that have the same group name as part of the same group.



Note `anova1` ignores any `NaN` values in `y`. Also, if `group` contains empty or `NaN` values, `anova1` ignores the corresponding observations in `y`. The `anova1` function performs balanced ANOVA if each group has the same number of observations after the function disregards empty or `NaN` values. Otherwise, `anova1` performs unbalanced ANOVA.

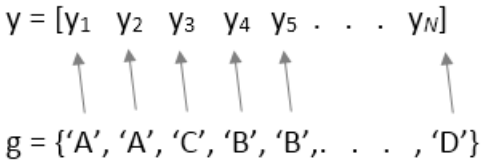
Data Types: `single` | `double`

`group` — Grouping variable

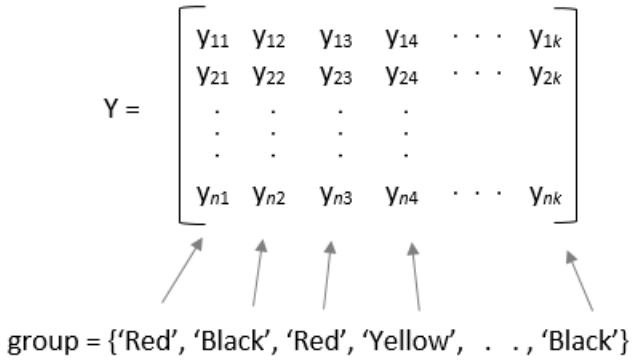
numeric vector | logical vector | categorical vector | character array | string array | cell array of character vectors

Grouping variable containing group names, specified as a numeric vector, logical vector, categorical vector, character array, string array, or cell array of character vectors.

- If `y` is a vector, then each element in `group` represents a group name of the corresponding element in `y`. The `anova1` function treats the `y` values corresponding to the same value of `group` as part of the same group.

  N is the total number of observations.
- If `y` is a matrix, then each element in `group` represents a group name for the corresponding column in `y`. The `anova1` function treats the columns of `y` that have the same group name as part of the same group.



If you do not want to specify group names for the matrix sample data `y`, enter an empty array (`[]`) or omit this argument. In this case, `anova1` treats each column of `y` as a separate group.

If `group` contains empty or `NaN` values, `anova1` ignores the corresponding observations in `y`.

For more information on grouping variables, see "Grouping Variables" on page 2-45.

Example: `'group',[1,2,1,3,1,...,3,1]` when `y` is a vector with observations categorized into groups 1, 2, and 3

Example: `'group',{'white','red','white','black','red'}` when `y` is a matrix with five columns categorized into groups red, white, and black

Data Types: `single` | `double` | `logical` | `categorical` | `char` | `string` | `cell`

`displayopt` — Indicator to display ANOVA table and box plot
`'on'` (default) | `'off'`

Indicator to display the ANOVA table and box plot, specified as `'on'` or `'off'`. When `displayopt` is `'off'`, `anova1` returns the output arguments, only. It does not display the standard ANOVA table and box plot.

Example: `p = anova(x,group,'off')`

Output Arguments

`p` — *p*-value for the *F*-test
scalar value

p-value for the *F*-test, returned as a scalar value. *p*-value is the probability that the *F*-statistic can take a value larger than the computed test-statistic value. `anova1` tests the null hypothesis that all group means are equal to each other against the alternative hypothesis that at least one group mean is different from the others. The function derives the *p*-value from the cdf of the *F*-distribution.

A *p*-value that is smaller than the significance level indicates that at least one of the sample means is significantly different from the others. Common significance levels are 0.05 or 0.01.

`tbl` — ANOVA table
cell array

ANOVA table, returned as a cell array. `tbl` has six columns.

Column	Definition
`source`	The source of the variability.
`SS`	The sum of squares due to each source.
`df`	The degrees of freedom associated with each source. Suppose N is the total number of observations and k is the number of groups. Then, $N - k$ is the within-groups degrees of freedom (`Error`), $k - 1$ is the between-groups degrees of freedom (`Columns`), and $N - 1$ is the total degrees of freedom. $N - 1 = (N - k) + (k - 1)$
`MS`	The mean squares for each source, which is the ratio `SS/df`.
`F`	F-statistic, which is the ratio of the mean squares.
`Prob>F`	The p-value, which is the probability that the F-statistic can take a value larger than the computed test-statistic value. `anova1` derives this probability from the cdf of F-distribution.

The rows of the ANOVA table show the variability in the data that is divided by the source.

Row	Definition
`Groups`	Variability due to the differences among the group means (variability *between* groups)
`Error`	Variability due to the differences between the data in each group and the group mean (variability *within* groups)
`Total`	Total variability

`stats` — Statistics for multiple comparison tests
structure

Statistics for multiple comparison tests on page 9-18, returned as a structure with the fields described in this table.

Field name	Definition
`gnames`	Names of the groups
`n`	Number of observations in each group
`source`	Source of the `stats` output
`means`	Estimated values of the means
`df`	Error (within-groups) degrees of freedom ($N - k$, where N is the total number of observations and k is the number of groups)
`s`	Square root of the mean squared error

More About

Box Plot

`anova1` returns a box plot of the observations for each group in `y`. Box plots provide a visual comparison of the group location parameters.

On each box, the central mark is the median (2nd quantile, q_2) and the edges of the box are the 25th and 75th percentiles (1st and 3rd quantiles, q_1 and q_3, respectively). The whiskers extend to the most extreme data points that are not considered outliers. The outliers are plotted individually using the `'+'` symbol. The extremes of the whiskers correspond to $q_3 + 1.5 \times (q_3 - q_1)$ and $q_1 - 1.5 \times (q_3 - q_1)$.

Box plots include notches for the comparison of the median values. Two medians are significantly different at the 5% significance level if their intervals, represented by notches, do not overlap. This test is different from the *F*-test that ANOVA performs; however, large differences in the center lines of the boxes correspond to a large *F*-statistic value and correspondingly a small *p*-value. The extremes of the notches correspond to $q_2 - 1.57(q_3 - q_1)/\text{sqrt}(n)$ and $q_2 + 1.57(q_3 - q_1)/\text{sqrt}(n)$, where n is the number of observations without any `NaN` values.

For more information about box plots, see `'Whisker'` and `'Notch'` of `boxplot`.

References

[1] Hogg, R. V., and J. Ledolter. *Engineering Statistics*. New York: MacMillan, 1987.

See Also

`anova2` | `anovan` | `boxplot` | `multcompare`

Topics

“Perform One-Way ANOVA” on page 9-5
“One-Way ANOVA” on page 9-3
“Multiple Comparisons” on page 9-18

Introduced before R2006a

anova2

Two-way analysis of variance

Syntax

```
p = anova2(y,reps)
p = anova2(y,reps,displayopt)
[p,tbl] = anova2( ___ )
[p,tbl,stats] = anova2( ___ )
```

Description

`anova2` performs two-way analysis of variance (ANOVA) with balanced designs. To perform two-way ANOVA with unbalanced designs, see `anovan`.

`p = anova2(y,reps)` returns the *p*-values for a balanced two-way ANOVA for comparing the means of two or more columns and two or more rows of the observations in `y`.

`reps` is the number of replicates for each combination of factor groups, which must be constant, indicating a balanced design. For unbalanced designs, use `anovan`. The `anova2` function tests the main effects for column and row factors and their interaction effect. To test the interaction effect, `reps` must be greater than 1.

`anova2` also displays the standard ANOVA table.

`p = anova2(y,reps,displayopt)` enables the ANOVA table display when `displayopt` is `'on'` (default) and suppresses the display when `displayopt` is `'off'`.

`[p,tbl] = anova2( ___ )` returns the ANOVA table (including column and row labels) in cell array `tbl`. To copy a text version of the ANOVA table to the clipboard, select **Edit > Copy Text** menu.

`[p,tbl,stats] = anova2( ___ )` returns a `stats` structure, which you can use to perform a multiple comparison test on page 9-18. A multiple comparison test enables you to determine which pairs of group means are significantly different. To perform this test, use `multcompare`, providing the `stats` structure as input.

Examples

Two-Way ANOVA

Load the sample data.

```
load popcorn
popcorn
```

```
popcorn = 6×3

    5.5000    4.5000    3.5000
    5.5000    4.5000    4.0000
    6.0000    4.0000    3.0000
```

```
    6.5000    5.0000    4.0000
    7.0000    5.5000    5.0000
    7.0000    5.0000    4.5000
```

The data is from a study of popcorn brands and popper types (Hogg 1987). The columns of the matrix `popcorn` are brands, Gourmet, National, and Generic, respectively. The rows are popper types, oil and air. In the study, researchers popped a batch of each brand three times with each popper, that is, the number of replications is 3. The first three rows correspond to the oil popper, and the last three rows correspond to the air popper. The response values are the yield in cups of popped popcorn.

Perform a two-way ANOVA. Save the ANOVA table in the cell array `tbl` for easy access to results.

```
[p,tbl] = anova2(popcorn,3);
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	15.75	2	7.875	56.7	0
Rows	4.5	1	4.5	32.4	0.0001
Interaction	0.0833	2	0.04167	0.3	0.7462
Error	1.6667	12	0.13889		
Total	22	17			

The column `Prob>F` shows the *p*-values for the three brands of popcorn (0.0000), the two popper types (0.0001), and the interaction between brand and popper type (0.7462). These values indicate that popcorn brand and popper type affect the yield of popcorn, but there is no evidence of an interaction effect of the two.

Display the cell array containing the ANOVA table.

```
tbl
```

```
tbl=6×6 cell array
  Columns 1 through 5

    {'Source'     }    {'SS'     }    {'df'}    {'MS'      }    {'F'       }
    {'Columns'    }    {[15.7500]}    {[ 2]}    {[  7.8750]}    {[ 56.7000]}
    {'Rows'       }    {[ 4.5000]}    {[ 1]}    {[  4.5000]}    {[ 32.4000]}
    {'Interaction'}    {[ 0.0833]}    {[ 2]}    {[  0.0417]}    {[  0.3000]}
    {'Error'      }    {[ 1.6667]}    {[12]}    {[  0.1389]}    {0x0 double}
    {'Total'      }    {[     22]}    {[17]}    {0x0 double}    {0x0 double}

  Column 6

    {'Prob>F'    }
    {[7.6790e-07]}
    {[1.0037e-04]}
    {[    0.7462]}
    {0x0 double  }
    {0x0 double  }
```

Store the *F*-statistic for the factors and factor interaction in separate variables.

```
Fbrands = tbl{2,5}
```

```
Fbrands = 56.7000
```

```
Fpoppertype = tbl{3,5}
```

```
Fpoppertype = 32.4000
```

```
Finteraction = tbl{4,5}
```

```
Finteraction = 0.3000
```

Multiple Comparisons for Two-Way ANOVA

Load the sample data.

```
load popcorn
popcorn
```

```
popcorn = 6×3

    5.5000    4.5000    3.5000
    5.5000    4.5000    4.0000
    6.0000    4.0000    3.0000
    6.5000    5.0000    4.0000
    7.0000    5.5000    5.0000
    7.0000    5.0000    4.5000
```

The data is from a study of popcorn brands and popper types (Hogg 1987). The columns of the matrix `popcorn` are brands (Gourmet, National, and Generic). The rows are popper types oil and air. In the study, researchers popped a batch of each brand three times with each popper. The values are the yield in cups of popped popcorn.

Perform a two-way ANOVA. Also compute the statistics that you need to perform a multiple comparison test on the main effects.

```
[~,~,stats] = anova2(popcorn,3,'off')
```

```
stats = struct with fields:
      source: 'anova2'
     sigmasq: 0.1389
    colmeans: [6.2500 4.7500 4]
        coln: 6
    rowmeans: [4.5000 5.5000]
        rown: 9
       inter: 1
        pval: 0.7462
          df: 12
```

The `stats` structure includes

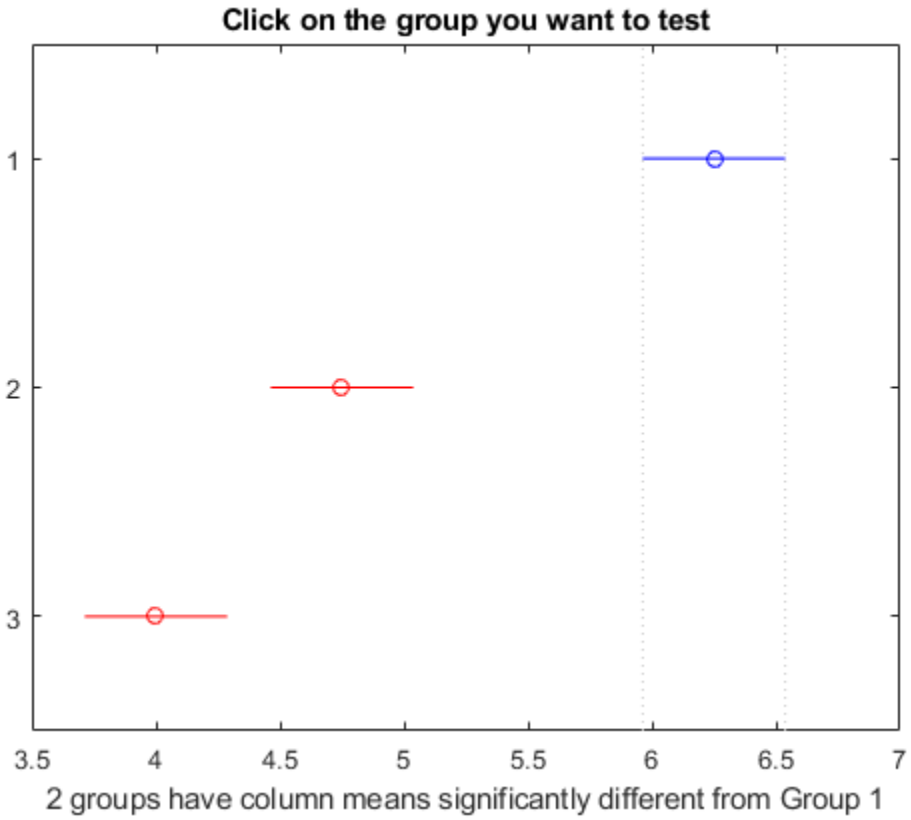
- The mean squared error (`sigmasq`)
- The estimates of the mean yield for each popcorn brand (`colmeans`)

- The number of observations for each popcorn brand (`coln`)
- The estimate of the mean yield for each popper type (`rowmeans`)
- The number of observations for each popper type (`rown`)
- The number of interactions (`inter`)
- The *p*-value that shows the significance level of the interaction term (`pval`)
- The error degrees of freedom (`df`).

Perform a multiple comparison test to see if the popcorn yield differs between pairs of popcorn brands (columns).

```
c = multcompare(stats)
```

```
Note: Your model includes an interaction term.  A test of main effects can be
difficult to interpret when the model includes interactions.
```




```
c = 3×6

    1.0000    2.0000    0.9260    1.5000    2.0740    0.0000
    1.0000    3.0000    1.6760    2.2500    2.8240    0.0000
    2.0000    3.0000    0.1760    0.7500    1.3240    0.0116
```

The first two columns of `c` show the groups that are compared. The fourth column shows the difference between the estimated group means. The third and fifth columns show the lower and upper limits for 95% confidence intervals for the true mean difference. The sixth column contains the *p*-

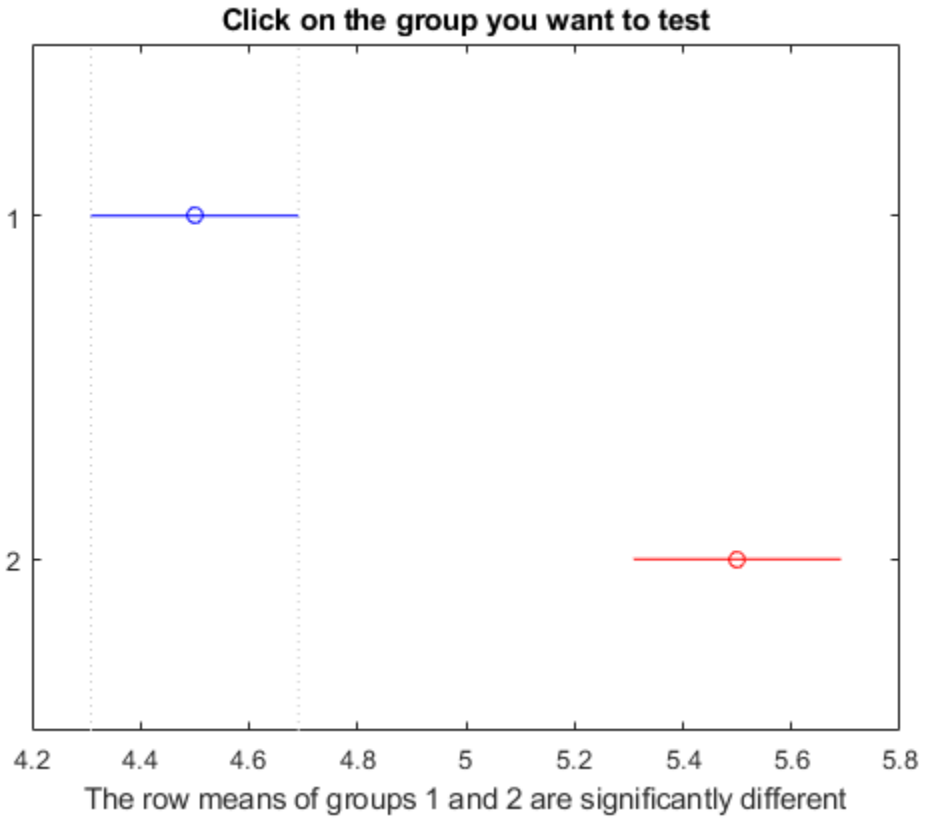
value for a hypothesis test that the corresponding mean difference is equal to zero. All *p*-values (0, 0, and 0.0116) are very small, which indicates that the popcorn yield differs across all three brands.

The figure shows the multiple comparison of the means. By default, the group 1 mean is highlighted and the comparison interval is in blue. Because the comparison intervals for the other two groups do not intersect with the intervals for the group 1 mean, they are highlighted in red. This lack of intersection indicates that both means are different than group 1 mean. Select other group means to confirm that all group means are significantly different from each other.

Perform a multiple comparison test to see the popcorn yield differs between the two popper types (rows).

```
c = multcompare(stats,'Estimate','row')
```

```
Note: Your model includes an interaction term.  A test of main effects can be
difficult to interpret when the model includes interactions.
```




```
c = 1×6

    1.0000    2.0000   -1.3828   -1.0000   -0.6172    0.0001
```

The small *p*-value of 0.0001 indicates that the popcorn yield differs between the two popper types (air and oil). The figure shows the same results. The disjoint comparison intervals indicate that the group means are significantly different from each other.

Input Arguments

y — Sample data

matrix

Sample data, specified as a matrix. The columns correspond to groups of one factor, and the rows correspond to the groups of the other factor and the replications. Replications are the measurements or observations for each combination of groups (levels) of the row and column factor. For example, in the following data the row factor *A* has three levels, column factor *B* has two levels, and there are two replications (`reps = 2`). The subscripts indicate row, column, and replication, respectively.

$$\begin{matrix} B=1 & B=2 \end{matrix}$$
$$\begin{bmatrix} y_{111} & y_{121} \\ y_{112} & y_{122} \\ y_{211} & y_{221} \\ y_{212} & y_{222} \\ y_{311} & y_{321} \\ y_{312} & y_{322} \end{bmatrix} \begin{matrix} \} A=1 \\ \} A=2 \\ \} A=3 \end{matrix}$$

Data Types: `single` | `double`

reps — Number of replications

1 (default) | an integer number

Number of replications for each combination of groups, specified as an integer number. For example, the following data has two replications (`reps = 2`) for each group combination of row factor *A* and column factor *B*.

$$\begin{matrix} B=1 & B=2 \end{matrix}$$
$$\begin{bmatrix} y_{111} & y_{121} \\ y_{112} & y_{122} \\ y_{211} & y_{221} \\ y_{212} & y_{222} \\ y_{311} & y_{321} \\ y_{312} & y_{322} \end{bmatrix} \begin{matrix} \} A=1 \\ \} A=2 \\ \} A=3 \end{matrix}$$

- When `reps` is `1` (default), `anova2` returns two *p*-values in vector `p`:
  - The *p*-value for the null hypothesis that all samples from factor *B* (i.e., all column samples in `y`) are drawn from the same population.
  - The *p*-value for the null hypothesis, that all samples from factor *A* (i.e., all row samples in `y`) are drawn from the same population.
- When `reps` is greater than `1`, `anova2` also returns the *p*-value for the null hypothesis that factors *A* and *B* have no interaction (i.e., the effects due to factors *A* and *B* are *additive*).

Example: `p = anova(y,3)` specifies that each combination of groups (levels) has three replications.

Data Types: `single` | `double`

`displayopt` — Indicator to display the ANOVA table
`'on'` (default) | `'off'`

Indicator to display the ANOVA table as a figure, specified as `'on'` or `'off'`.

Output Arguments

`p` — *p*-value
scalar value

p-value for the *F*-test, returned as a scalar value. A small *p*-value indicates that the results are statistically significant. Common significance levels are 0.05 or 0.01. For example:

- A sufficiently small *p*-value for the null hypothesis for group means of row factor *A* suggests that at least one row-sample mean is significantly different from the other row-sample means; i.e., there is a main effect due to factor *A*
- A sufficiently small *p*-value for the null hypothesis for group (level) means of column factor *B* suggests that at least one column-sample mean is significantly different from the other column-sample means; i.e., there is a main effect due to factor *B*.
- A sufficiently small *p*-value for combinations of groups (levels) of factors *A* and *B* suggests that there is an interaction between factors *A* and *B*.

`tbl` — ANOVA table
cell array

ANOVA table, returned as a cell array. `tbl` has six columns.

Column name	Definition
`source`	Source of the variability.
`SS`	Sum of squares due to each source.
`df`	Degrees of freedom associated with each source.
`MS`	Mean squares for each source, which is the ratio `SS/df`.
`F`	*F*-statistic, which is the ratio of the mean squares.
`Prob>F`	*p*-value, which is the probability that the *F*-statistic can take a value larger than the computed test-statistic value. `anova2` derives this probability from the cdf of the *F*-distribution.

The rows of the ANOVA table show the variability in the data, divided by the source into three or four parts, depending on the value of `reps`.

Row	Definition
`Columns`	Variability due to the differences among the column means
`Rows`	Variability due to the differences among the row means

Row	Definition
`Interaction`	Variability due to the interaction between rows and columns (if `reps` is greater than its default value of 1)
`Error`	Remaining variability not explained by any systematic source

Data Types: `cell`

`stats` — Statistics for multiple comparison test
structure

Statistics for multiple comparisons tests on page 9-18, returned as a structure. Use `multcompare` to perform multiple comparison tests, supplying `stats` as an input argument. `stats` has nine fields.

Field	Definition
`source`	Source of the `stats` output
`sigmasq`	Mean squared error
`colmeans`	Estimated values of the column means
`coln`	Number of observations for each group in columns
`rowmeans`	Estimated values of the row means
`rown`	Number of observations for each group in rows
`inter`	Number of interactions
`pval`	*p*-value for the interaction term
`df`	Error degrees of freedom (*reps* — 1)*r*c* where *reps* is the number of replications and *c* and *r* are the number of groups in factors, respectively.

Data Types: `struct`

References

[1] Hogg, R. V., and J. Ledolter. *Engineering Statistics*. New York: MacMillan, 1987.

See Also

`anova1` | `anovan` | `multcompare`

Topics
“Perform Two-Way ANOVA” on page 9-13
“Two-Way ANOVA” on page 9-11
“Multiple Comparisons” on page 9-18

Introduced before R2006a

anovan

N-way analysis of variance

Syntax

```
p = anovan(y,group)
p = anovan(y,group,Name,Value)
[p,tbl] = anovan( ___ )
[p,tbl,stats] = anovan( ___ )
[p,tbl,stats,terms] = anovan( ___ )
```

Description

`p = anovan(y,group)` returns a vector of *p*-values, one per term, for multiway (*n*-way) analysis of variance (ANOVA) for testing the effects of multiple factors on the mean of the vector `y`.

`anovan` also displays a figure showing the standard ANOVA table.

`p = anovan(y,group,Name,Value)` returns a vector of *p*-values for multiway (*n*-way) ANOVA using additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify which predictor variable is continuous, if any, or the type of sum of squares to use.

`[p,tbl] = anovan( ___ )` returns the ANOVA table (including factor labels) in cell array `tbl` for any of the input arguments specified in the previous syntaxes. Copy a text version of the ANOVA table to the clipboard by using the `Copy Text` item on the **Edit** menu.

`[p,tbl,stats] = anovan( ___ )` returns a `stats` structure that you can use to perform a multiple comparison test on page 9-18, which enables you to determine which pairs of group means are significantly different. You can perform such a test using the `multcompare` function by providing the `stats` structure as input.

`[p,tbl,stats,terms] = anovan( ___ )` returns the main and interaction terms used in the ANOVA computations in `terms`.

Examples

Three-Way ANOVA

Load the sample data.

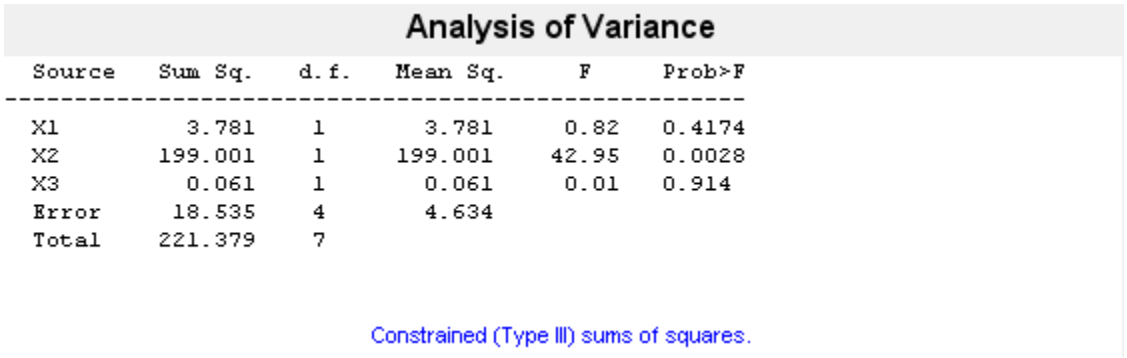
```
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';
g1 = [1 2 1 2 1 2 1 2];
g2 = {'hi';'hi';'lo';'lo';'hi';'hi';'lo';'lo'};
g3 = {'may';'may';'may';'may';'june';'june';'june';'june'};
```

`y` is the response vector and `g1`, `g2`, and `g3` are the grouping variables (factors). Each factor has two levels ,and every observation in `y` is identified by a combination of factor levels. For example,

observation `y(1)` is associated with level 1 of factor `g1`, level `'hi'` of factor `g2`, and level `'may'` of factor `g3`. Similarly, observation `y(6)` is associated with level 2 of factor `g1`, level `'hi'` of factor `g2`, and level `'june'` of factor `g3`.

Test if the response is the same for all factor levels.

```
p = anovan(y,{g1,g2,g3})
```



Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
X1	3.781	1	3.781	0.82	0.4174
X2	199.001	1	199.001	42.95	0.0028
X3	0.061	1	0.061	0.01	0.914
Error	18.535	4	4.634		
Total	221.379	7			

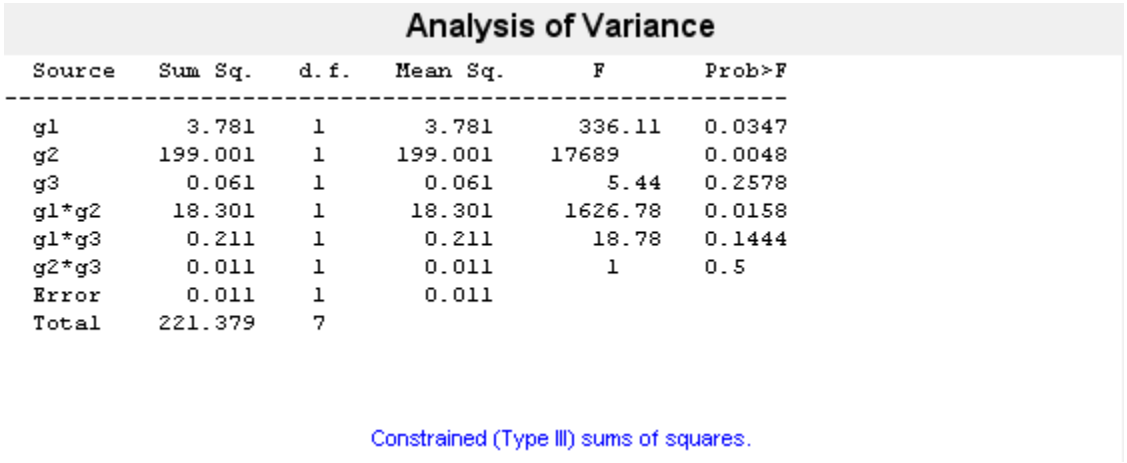
Constrained (Type III) sums of squares.

```
p = 3×1

    0.4174
    0.0028
    0.9140
```

In the ANOVA table, `X1`, `X2`, and `X3` correspond to the factors `g1`, `g2`, and `g3`, respectively. The *p*-value 0.4174 indicates that the mean responses for levels 1 and 2 of the factor `g1` are not significantly different. Similarly, the *p*-value 0.914 indicates that the mean responses for levels `'may'` and `'june'`, of the factor `g3` are not significantly different. However, the *p*-value 0.0028 is small enough to conclude that the mean responses are significantly different for the two levels, `'hi'` and `'lo'` of the factor `g2`. By default, `anovan` computes *p*-values just for the three main effects.

Test the two-factor interactions. This time specify the variable names.

```
p = anovan(y,{g1 g2 g3},'model','interaction','varnames',{'g1','g2','g3'})
```



Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
g1	3.781	1	3.781	336.11	0.0347
g2	199.001	1	199.001	17689	0.0048
g3	0.061	1	0.061	5.44	0.2578
g1*g2	18.301	1	18.301	1626.78	0.0158
g1*g3	0.211	1	0.211	18.78	0.1444
g2*g3	0.011	1	0.011	1	0.5
Error	0.011	1	0.011		
Total	221.379	7			

Constrained (Type III) sums of squares.

```
p = 6×1
```

```
    0.0347
    0.0048
    0.2578
    0.0158
    0.1444
    0.5000
```

The interaction terms are represented by `g1*g2`, `g1*g3`, and `g2*g3` in the ANOVA table. The first three entries of `p` are the *p*-values for the main effects. The last three entries are the *p*-values for the two-way interactions. The *p*-value of 0.0158 indicates that the interaction between `g1` and `g2` is significant. The *p*-values of 0.1444 and 0.5 indicate that the corresponding interactions are not significant.

Two-Way ANOVA for Unbalanced Design

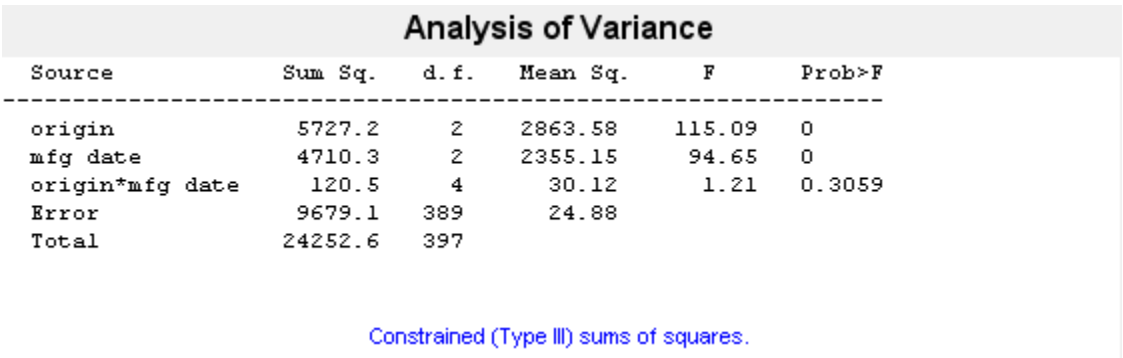
Load the sample data.

```
load carbig
```

The data has measurements on 406 cars. The variable `org` shows where the cars were made and `when` shows when in the year the cars were manufactured.

Study how the mileage depends on when and where the cars were made. Also include the two-way interactions in the model.

```
p = anovan(MPG,{org when},'model',2,'varnames',{'origin','mfg date'})
```




```
p = 3×1

    0.0000
    0.0000
    0.3059
```

The `'model',2` name-value pair argument represents the two-way interactions. The *p*-value for the interaction term, 0.3059, is not small, indicating little evidence that the effect of the time of manufacture (`mfg date`) depends on where the car was made (`origin`). The main effects of origin and manufacturing date, however, are significant, both *p*-values are 0.

Multiple Comparisons for Three-Way ANOVA

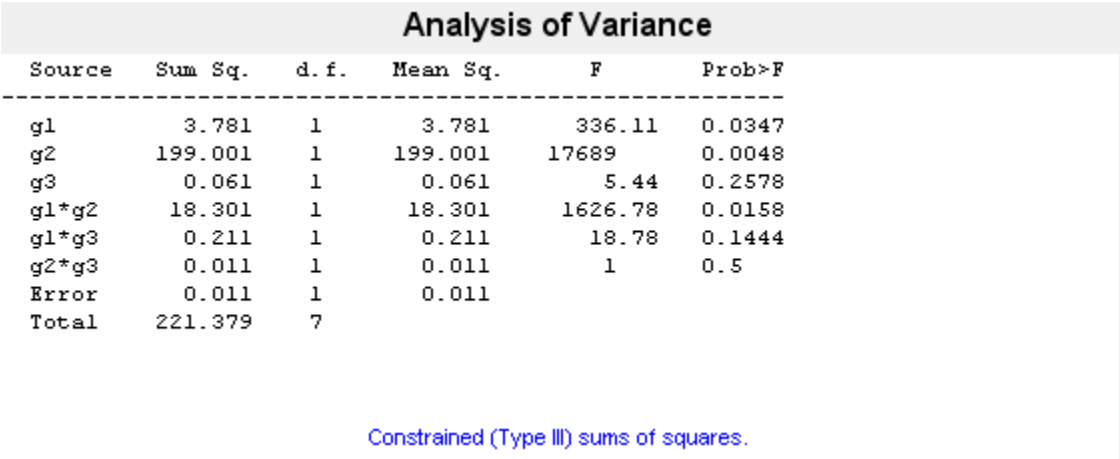
Load the sample data.

```
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';
g1 = [1 2 1 2 1 2 1 2];
g2 = {'hi';'hi';'lo';'lo';'hi';'hi';'lo';'lo'};
g3 = {'may';'may';'may';'may';'june';'june';'june';'june'};
```

`y` is the response vector and `g1`, `g2`, and `g3` are the grouping variables (factors). Each factor has two levels, and every observation in `y` is identified by a combination of factor levels. For example, observation `y(1)` is associated with level 1 of factor `g1`, level `'hi'` of factor `g2`, and level `'may'` of factor `g3`. Similarly, observation `y(6)` is associated with level 2 of factor `g1`, level `'hi'` of factor `g2`, and level `'june'` of factor `g3`.

Test if the response is the same for all factor levels. Also compute the statistics required for multiple comparison tests.

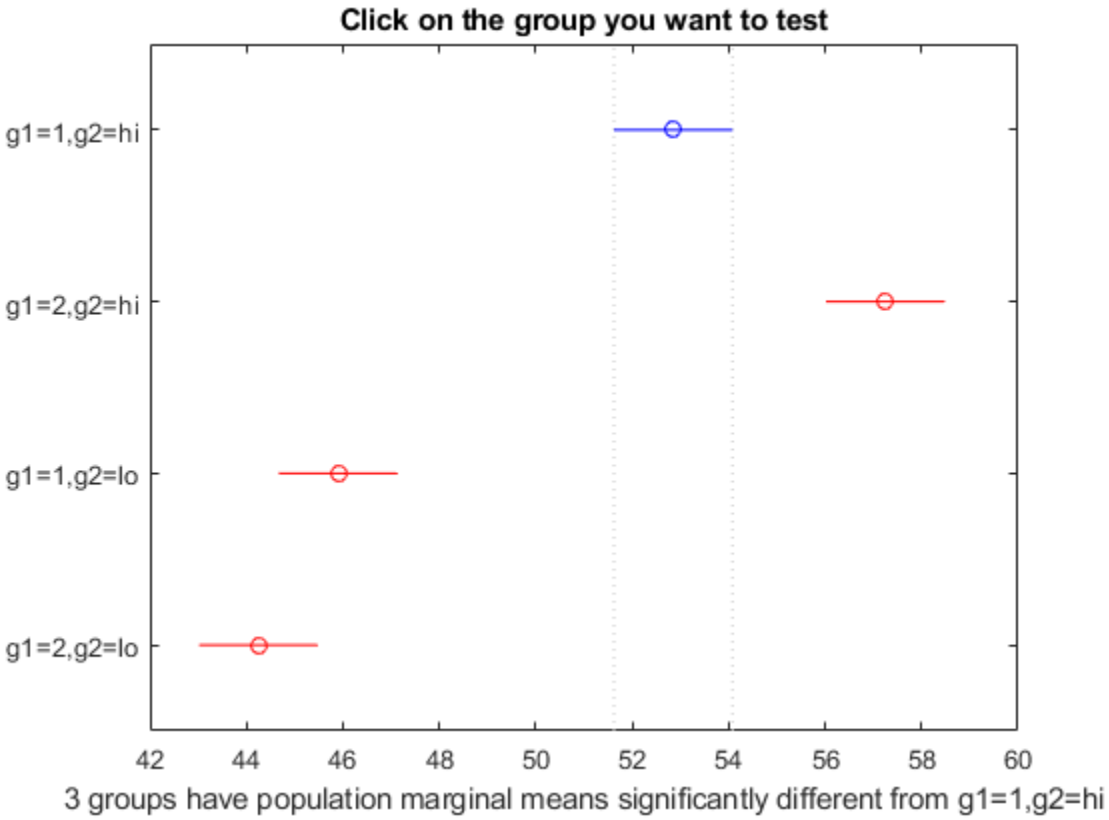
```
[~,~,stats] = anovan(y,{g1 g2 g3},'model','interaction',...
    'varnames',{'g1','g2','g3'});
```



The *p*-value of 0.2578 indicates that the mean responses for levels `'may'` and `'june'` of factor `g3` are not significantly different. The *p*-value of 0.0347 indicates that the mean responses for levels 1 and 2 of factor `g1` are significantly different. Similarly, the *p*-value of 0.0048 indicates that the mean responses for levels `'hi'` and `'lo'` of factor `g2` are significantly different.

Perform multiple comparison tests to find out which groups of the factors `g1` and `g2` are significantly different.

```
results = multcompare(stats,'Dimension',[1 2])
```




```
results = 6×6

    1.0000    2.0000   -6.8604   -4.4000   -1.9396    0.0272
    1.0000    3.0000    4.4896    6.9500    9.4104    0.0170
    1.0000    4.0000    6.1396    8.6000   11.0604    0.0136
    2.0000    3.0000    8.8896   11.3500   13.8104    0.0101
    2.0000    4.0000   10.5396   13.0000   15.4604    0.0087
    3.0000    4.0000   -0.8104    1.6500    4.1104    0.0737
```

`multcompare` compares the combinations of groups (levels) of the two grouping variables, `g1` and `g2`. In the `results` matrix, the number 1 corresponds to the combination of level `1` of `g1` and level `hi` of `g2`, the number 2 corresponds to the combination of level 2 of `g1` and level `hi` of `g2`. Similarly, the number 3 corresponds to the combination of level `1` of `g1` and level `lo` of `g2`, and the number 4 corresponds to the combination of level 2 of `g1` and level `lo` of `g2`. The last column of the matrix contains the *p*-values.

For example, the first row of the matrix shows that the combination of level `1` of `g1` and level `hi` of `g2` has the same mean response values as the combination of level 2 of `g1` and level `hi` of `g2`. The *p*-value corresponding to this test is 0.0280, which indicates that the mean responses are significantly different. You can also see this result in the figure. The blue bar shows the comparison interval for the mean response for the combination of level `1` of `g1` and level `hi` of `g2`. The red bars are the comparison intervals for the mean response for other group combinations. None of the red bars overlap with the blue bar, which means the mean response for the combination of level `1` of `g1` and level `hi` of `g2` is significantly different from the mean response for other group combinations.

You can test the other groups by clicking on the corresponding comparison interval for the group. The bar you click on turns to blue. The bars for the groups that are significantly different are red. The bars for the groups that are not significantly different are gray. For example, if you click on the comparison interval for the combination of level `1` of `g1` and level `lo` of `g2`, the comparison interval for the combination of level `2` of `g1` and level `lo` of `g2` overlaps, and is therefore gray. Conversely, the other comparison intervals are red, indicating significant difference.

Input Arguments

y — Sample data

numeric vector

Sample data, specified as a numeric vector.

Data Types: `single` | `double`

group — Grouping variables

cell array

Grouping variables, i.e. the factors and factor levels of the observations in `y`, specified as a cell array. Each of the cells in `group` contains a list of factor levels identifying the observations in `y` with respect to one of the factors. The list within each cell can be a categorical array, numeric vector, character matrix, string array, or single-column cell array of character vectors, and must have the same number of elements as `y`.

$$
\begin{array}{rcccccccc}
y & = & [& y_1, & y_2, & y_3, & y_4, & y_5, \cdots, & y_N &]' \\
 & & & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \\
g1 & = & \{ & 'A', & 'A', & 'C', & 'B', & 'B', \cdots, & 'D' & \} \\
g2 & = & [& 1 & 2 & 1 & 3 & 1 \cdots, & 2 &] \\
g3 & = & \{ & \text{'hi'}, & \text{'mid'}, & \text{'low'}, & \text{'mid'}, & \text{'hi'}, \cdots, & \text{'low'} & \}
\end{array}
$$

By default, `anovan` treats all grouping variables as fixed effects.

For example, in a study you want to investigate the effects of gender, school, and the education method on the academic success of elementary school students, then you can specify the grouping variables as follows.

Example: `{'Gender','School','Method'}`

Data Types: `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'alpha',0.01,'model','interaction','sstype',2` specifies `anovan` to compute the 99% confidence bounds and p-values for the main effects and two-way interactions using type II sum of squares.

alpha — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level for confidence bounds, specified as the comma-separated pair consisting of`'alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

Example: `'alpha',0.01` corresponds to 99% confidence intervals

Data Types: `single` | `double`

`continuous` — Indicator for continuous predictors
vector of indices

Indicator for continuous predictors, representing which grouping variables should be treated as continuous predictors rather than as categorical predictors, specified as the comma-separated pair consisting of `'continuous'` and a vector of indices.

For example, if there are three grouping variables and second one is continuous, then you can specify as follows.

Example: `'continuous',[2]`

Data Types: `single` | `double`

`display` — Indicator to display ANOVA table
`'on'` (default) | `'off'`

Indicator to display ANOVA table, specified as the comma-separated pair consisting of `'display'` and `'on'` or `'off'`. When `'display'` is `'off'`, `anovan` only returns the output arguments, and does not display the standard ANOVA table as a figure.

Example: `'display','off'`

`model` — Type of the model
`'linear'` (default) | `'interaction'` | `'full'` | integer value | terms matrix

Type of the model, specified as the comma-separated pair consisting of `'model'` and one of the following:

- `'linear'` — The default `'linear'` model computes only the *p*-values for the null hypotheses on the *N* main effects.
- `'interaction'` — The `'interaction'` model computes the *p*-values for null hypotheses on the *N* main effects and the $\binom{N}{2}$ two-factor interactions.
- `'full'` — The `'full'` model computes the *p*-values for null hypotheses on the *N* main effects and interactions at all levels.
- An integer — For an integer value of *k*, ($k \le N$) for model type, `anovan` computes all interaction levels through the *k*th level. For example, the value 3 means main effects plus two- and three-factor interactions. The values $k = 1$ and $k = 2$ are equivalent to the `'linear'` and `'interaction'` specifications, respectively. The value $k = N$ is equivalent to the `'full'` specification.
- Terms matrix — A matrix of term definitions having the same form as the input to the `x2fx` function. All entries must be `0` or `1` (no higher powers).

  For more precise control over the main and interaction terms that `anovan` computes, you can specify a matrix containing one row for each main or interaction term to include in the ANOVA model. Each row defines one term using a vector of *N* zeros and ones. The table below illustrates the coding for a 3-factor ANOVA for factors *A*, *B*, and *C*.

Matrix Row	ANOVA Term
`[1 0 0]`	Main term *A*
`[0 1 0]`	Main term *B*
`[0 0 1]`	Main term *C*
`[1 1 0]`	Interaction term *AB*
`[1 0 1]`	Interaction term *AC*
`[0 1 1]`	Interaction term *BC*
`[1 1 1]`	Interaction term *ABC*

For example, if there are three factors *A*, *B*, and *C*, and `'model',[0 1 0;0 0 1;0 1 1]`, then `anovan` tests for the main effects *B* and *C*, and the interaction effect *BC*, respectively.

A simple way to generate the terms matrix is to modify the `terms` output, which codes the terms in the current model using the format described above. If `anovan` returns `[0 1 0;0 0 1;0 1 1]` for `terms`, for example, and there is no significant interaction *BC*, then you can recompute ANOVA on just the main effects *B* and *C* by specifying `[0 1 0;0 0 1]` for `model`.

Example: `'model',[0 1 0;0 0 1;0 1 1]`

Example: `'model','interaction'`

Data Types: `char` | `string` | `single` | `double`

`nested` — Nesting relationships
matrix of 0’s and 1’s

Nesting relationships among the grouping variables, specified as the comma-separated pair consisting of `'nested'` and a matrix *M* of 0’s and 1’s, i.e.*M*(*i*,*j*) = 1 if variable *i* is nested in variable *j*.

You cannot specify nesting in a continuous variable.

For example, if there are two grouping variables District and School, where School is nested in District, then you can express this relationship as follows.

Example: `'nested',[0 0;1 0]`

Data Types: `single` | `double`

`random` — Indicator for random variables
vector of indices

Indicator for random variables, representing which grouping variables are random, specified as the comma-separated pair consisting of `'random'` and a vector of indices. By default, `anovan` treats all grouping variables as fixed.

`anovan` treats an interaction term as random if any of the variables in the interaction term is random.

Example: `'random',[3]`

Data Types: `single` | `double`

`sstype` — Type of sum of squares
3 (default) | 1 | 2 | `'h'`

Type of sum squares, specified as the comma-separated pair consisting of `'sstype'` and the following:

- 1 — Type I sum of squares. The reduction in residual sum of squares obtained by adding that term to a fit that already includes the terms listed before it.
- 2 — Type II sum of squares. The reduction in residual sum of squares obtained by adding that term to a model consisting of all other terms that do not contain the term in question.
- 3 — Type III sum of squares. The reduction in residual sum of squares obtained by adding that term to a model containing all other terms, but with their effects constrained to obey the usual "sigma restrictions" that make models estimable.
- `'h'` — Hierarchical model. Similar to type 2, but with continuous as well as categorical factors used to determine the hierarchy of terms.

The sum of squares for any term is determined by comparing two models. For a model containing main effects but no interactions, the value of `sstype` influences the computations on unbalanced data only.

Suppose you are fitting a model with two factors and their interaction, and the terms appear in the order *A*, *B*, *AB*. Let $R(\cdot)$ represent the residual sum of squares for the model. So, $R(A, B, AB)$ is the residual sum of squares fitting the whole model, $R(A)$ is the residual sum of squares fitting the main effect of *A* only, and $R(1)$ is the residual sum of squares fitting the mean only. The three sum of squares types are as follows:

Term	Type 1 Sum of Squares	Type 2 Sum of Squares	Type 3 Sum of Squares
A	$R(1) - R(A)$	$R(B) - R(A, B)$	$R(B, AB) - R(A, B, AB)$
B	$R(A) - R(A, B)$	$R(A) - R(A, B)$	$R(A, AB) - R(A, B, AB)$
AB	$R(A, B) - R(A, B, AB)$	$R(A, B) - R(A, B, AB)$	$R(A, B) - R(A, B, AB)$

The models for Type 3 sum of squares have sigma restrictions imposed. This means, for example, that in fitting $R(B, AB)$, the array of *AB* effects is constrained to sum to 0 over *A* for each value of *B*, and over *B* for each value of *A*.

Example: `'sstype','h'`

Data Types: `single` | `double` | `char` | `string`

`varnames` — Names of grouping variables
`X1,X2,...,XN` (default) | character matrix | string array | cell array of character vectors

Names of grouping variables, specified as the comma-separating pair consisting of `'varnames'` and a character matrix, a string array, or a cell array of character vectors.

Example: `'varnames',{'Gender','City'}`

Data Types: `char` | `string` | `cell`

Output Arguments

p — *p*-values
vector

p-values, returned as a vector.

Output vector `p` contains *p*-values for the null hypotheses on the *N* main effects and any interaction terms specified. Element `p(1)` contains the *p*-value for the null hypotheses that samples at all levels of factor *A* are drawn from the same population; element `p(2)` contains the *p*-value for the null hypotheses that samples at all levels of factor *B* are drawn from the same population; and so on.

For example, if there are three factors *A*, *B*, and *C*, and `'model',[0 1 0;0 0 1;0 1 1]`, then the output vector `p` contains the *p*-values for the null hypotheses on the main effects *B* and *C* and the interaction effect *BC*, respectively.

A sufficiently small *p*-value corresponding to a factor suggests that at least one group mean is significantly different from the other group means; that is, there is a main effect due to that factor. It is common to declare a result significant if the *p*-value is less than 0.05 or 0.01.

`tbl` — ANOVA table

cell array

ANOVA table, returned as a cell array. The ANOVA table has seven columns:

Column name	Definition
`source`	Source of the variability.
`SS`	Sum of squares due to each source.
`df`	Degrees of freedom associated with each source.
`MS`	Mean squares for each source, which is the ratio `SS/df`.
`Singular?`	Indication of whether the term is singular.
`F`	*F*-statistic, which is the ratio of the mean squares.
`Prob>F`	The *p*-values, which is the probability that the *F*-statistic can take a value larger than a computed test-statistic value. `anovan` derives these probabilities from the cdf of *F*-distribution.

The ANOVA table also contains the following columns if at least one of the grouping variables is specified as random using the name-value pair argument `random`:

Column name	Definition
`Type`	Type of each source; `'fixed'` for a fixed effect or `'random'` for a random effect.
`Expected MS`	Text representation of the expected value for the mean square. `Q(source)` represents a quadratic function of `source` and `V(source)` represents the variance of `source`.
`MS denom`	Denominator of the *F*-statistic.
`d.f. denom`	Degrees of freedom for the denominator of the *F*-statistic.
`Denom. defn.`	Text representation of the denominator of the *F*-statistic. `MS(source)` represents the mean square of `source`.

Column name	Definition
`Var. est.`	Variance component estimate.
`Var. lower bnd`	Lower bound of the 95% confidence interval for the variance component estimate.
`Var. upper bnd`	Upper bound of the 95% confidence interval for the variance component estimate.

`stats` — Statistics
structure

Statistics to use in a multiple comparison test on page 9-18 using the `multcompare` function, returned as a structure.

`anovan` evaluates the hypothesis that the different groups (levels) of a factor (or more generally, a term) have the same effect, against the alternative that they do not all have the same effect. Sometimes it is preferable to perform a test to determine which pairs of levels are significantly different, and which are not. Use the `multcompare` function to perform such tests by supplying the `stats` structure as input.

The `stats` structure contains the fields listed below, in addition to a number of other fields required for doing multiple comparisons using the `multcompare` function:

Field	Description
`coeffs`	Estimated coefficients
`coeffnames`	Name of term for each coefficient
`vars`	Matrix of grouping variable values for each term
`resid`	Residuals from the fitted model

The `stats` structure also contains the following fields if at least one of the grouping variables is specified as random using the name-value pair argument `random`:

Field	Description
`ems`	Expected mean squares
`denom`	Denominator definition
`rtnames`	Names of random terms
`varest`	Variance component estimates (one per random term)
`varci`	Confidence intervals for variance components

`terms` — Main and interaction terms
matrix

Main and interaction terms, returned as a matrix. The terms are encoded in the output matrix `terms` using the same format described above for input `model`. When you specify `model` itself in this format, the matrix returned in `terms` is identical.

References

[1] Dunn, O.J., and V.A. Clark. *Applied Statistics: Analysis of Variance and Regression*. New York: Wiley, 1974.

[2] Goodnight, J.H., and F.M. Speed. *Computing Expected Mean Squares*. Cary, NC: SAS Institute, 1978.

[3] Seber, G. A. F., and A. J. Lee. *Linear Regression Analysis*. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2003.

See Also

`anova1` | `anova2` | `fitrm` | `multcompare` | `ranova`

Topics

“Perform N-Way ANOVA” on page 9-27
“ANOVA with Random Effects” on page 9-33
“Multiple Comparisons” on page 9-18
“N-Way ANOVA” on page 9-25

Introduced before R2006a

anova

Class: RepeatedMeasuresModel

Analysis of variance for between-subject effects

Syntax

```
anovatbl = anova(rm)
anovatbl = anova(rm,'WithinModel',WM)
```

Description

`anovatbl = anova(rm)` returns the analysis of variance results for the repeated measures model `rm`.

`anovatbl = anova(rm,'WithinModel',WM)` returns the analysis of variance results it performs using the response or responses specified by the within-subject model `WM`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`WM` — Within-subject model

`'separatemeans'` (default) | `'orthogonalcontrasts'` | character vector or string scalar defining a model specification | *r*-by-*nc* matrix specifying *nc* contrasts

Within-subject model, specified as one of the following:

- `'separatemeans'` — The response is the average of the repeated measures (average across the within-subject model).
- `'orthogonalcontrasts'` — This is valid when the within-subject model has a single numeric factor *T*. Responses are the average, the slope of centered *T*, and, in general, all orthogonal contrasts for a polynomial up to *T*^(*p* – 1), where *p* is the number of rows in the within-subject model. `anova` multiplies `Y`, the response you use in the repeated measures model `rm` by the orthogonal contrasts, and uses the columns of the resulting product matrix as the responses.

  `anova` computes the orthogonal contrasts for *T* using the *Q* factor of a QR factorization on page 33-96 of the Vandermonde matrix on page 33-96.
- A character vector or string scalar that defines a model specification in the within-subject factors. Responses are defined by the terms in that model. `anova` multiplies *Y*, the response matrix you use in the repeated measures model `rm` by the terms of the model, and uses the columns of the result as the responses.

For example, if there is a Time factor and `'Time'` is the model specification, then `anova` uses two terms, the constant and the uncentered Time term. The default is `'1'` to perform on the average response.

- An *r*-by-*nc* matrix, *C*, specifying *nc* contrasts among the *r* repeated measures. If *Y* represents the matrix of repeated measures you use in the repeated measures model `rm`, then the output `tbl` contains a separate analysis of variance for each column of *Y***C*.

The `anova` table contains a separate univariate analysis of variance results for each response.

Example: `'WithinModel','Time'`

Example: `'WithinModel','orthogonalcontrasts'`

Output Arguments

`anovatbl` — Results of analysis of variance

table

Results of analysis of variance for between-subject effects, returned as a table. This includes all terms on the between-subjects model and the following columns.

Column Name	Definition
`Within`	Within-subject factors
`Between`	Between-subject factors
`SumSq`	Sum of squares
`DF`	Degrees of freedom
`MeanSq`	Mean squared error
`F`	*F*-statistic
`pValue`	*p*-value corresponding to the *F*-statistic

Examples

Analysis of Variance for Average Response

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform analysis of variance.

```
anova(rm)
```

```
ans=3×7 table
     Within     Between     SumSq     DF     MeanSq       F          pValue
    ________    ________    ______    ___    _______    ______    ___________

    Constant    constant    7201.7      1     7201.7     19650    2.0735e-158
    Constant    species     309.61      2      154.8    422.39     1.1517e-61
    Constant    Error       53.875    147    0.36649
```

There are 150 observations and 3 species. The degrees of freedom for species is 3 - 1 = 2, and for error it is 150 - 3 = 147. The small *p*-value of 1.1517e-61 indicates that the measurements differ significantly according to species.

Panel Data

Load the sample panel data.

```
load('panelData.mat');
```

The dataset array, `panelData`, contains yearly observations on eight cities for 6 years. The first variable, `Growth`, measures economic growth (the response variable). The second and third variables are city and year indicators, respectively. The last variable, `Employ`, measures employment (the predictor variable). This is simulated data.

Store the data in a table array and define city as a nominal variable.

```
t = table(panelData.Growth,panelData.City,panelData.Year,...
    'VariableNames',{'Growth','City','Year'});
```

Convert the data in a proper format to do repeated measures analysis.

```
t = unstack(t,'Growth','Year','NewDataVariableNames',...
    {'year1','year2','year3','year4','year5','year6'});
```

Add the mean employment level over the years as a predictor variable to the table `t`.

```
t(:,8) = table(grpstats(panelData.Employ,panelData.City));
t.Properties.VariableNames{'Var8'} = 'meanEmploy';
```

Define the within-subjects variable.

```
Year = [1 2 3 4 5 6]';
```

Fit a repeated measures model, where the growth figures over the 6 years are the responses and the mean employment is the predictor variable.

```
rm = fitrm(t,'year1-year6 ~ meanEmploy','WithinDesign',Year);
```

Perform analysis of variance.

```
anovatbl = anova(rm,'WithinModel',Year)
```

```
anovatbl=3×7 table
     Within       Between        SumSq       DF      MeanSq         F         pValue
    _________    __________    __________    __    __________    ________    _________

    Contrast1    constant          588.17    1         588.17    0.038495      0.85093
    Contrast1    meanEmploy    3.7064e+05    1     3.7064e+05      24.258    0.0026428
    Contrast1    Error              91675    6          15279
```

Longitudinal Data

Load the sample data.

```
load('longitudinalData.mat');
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (time = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual, and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Define a variable that stores gender information.

```
Gender = ['F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']';
```

Store the data in a proper table array format to do repeated measures analysis.

```
t = table(Gender,Y(:,1),Y(:,2),Y(:,3),Y(:,4),Y(:,5),...
'VariableNames',{'Gender','t0','t2','t4','t6','t8'});
```

Define the within-subjects variable.

```
Time = [0 2 4 6 8]';
```

Fit a repeated measures model, where blood levels are the responses and gender is the predictor variable.

```
rm = fitrm(t,'t0-t8 ~ Gender','WithinDesign',Time);
```

Perform analysis of variance.

```
anovatbl = anova(rm)
```

```
anovatbl=3×7 table
     Within     Between     SumSq     DF    MeanSq      F         pValue
    ________    ________    ______    __    ______    ______    __________

    Constant    constant     54702     1     54702    1079.2    1.1897e-14
    Constant    Gender      2251.7     1    2251.7    44.425    1.0693e-05
    Constant    Error        709.6    14    50.685
```

There are 2 genders and 16 observations, so the degrees of freedom for gender is (2 - 1) = 1 and for error it is (16 - 2)*(2 - 1) = 14. The small *p*-value of 1.0693e-05 indicates that there is a significant effect of gender on blood pressure.

Repeat analysis of variance using orthogonal contrasts.

```
anovatbl = anova(rm,'WithinModel','orthogonalcontrasts')
```

```
anovatbl=15×7 table
    Within      Between       SumSq       DF      MeanSq          F           pValue
    ________    ________    __________    __    __________    __________    __________

    Constant    constant         54702     1         54702        1079.2    1.1897e-14
    Constant    Gender          2251.7     1        2251.7        44.425    1.0693e-05
    Constant    Error            709.6    14        50.685
    Time        constant        310.83     1        310.83        31.023    6.9065e-05
    Time        Gender          13.341     1        13.341        1.3315       0.26785
    Time        Error           140.27    14        10.019
    Time^2      constant        565.42     1        565.42        98.901    1.0003e-07
    Time^2      Gender          1.4076     1        1.4076       0.24621       0.62746
    Time^2      Error           80.039    14        5.7171
    Time^3      constant        2.6127     1        2.6127        1.4318       0.25134
    Time^3      Gender       7.8853e-06    1    7.8853e-06    4.3214e-06       0.99837
    Time^3      Error           25.546    14        1.8247
    Time^4      constant        2.8404     1        2.8404       0.47924       0.50009
    Time^4      Gender          2.9016     1        2.9016       0.48956       0.49559
    Time^4      Error           82.977    14        5.9269
```

More About

Vandermonde Matrix

Vandermonde matrix is the matrix where columns are the powers of the vector a, that is, $V(i,j) = a(i)^{(n-j)}$, where n is the length of a.

QR Factorization

QR factorization of an m-by-n matrix A is the factorization that matrix into the product $A = Q*R$, where R is an m-by-n upper triangular matrix and Q is an m-by-m unitary matrix.

See Also

`fitrm` | `manova` | `qr` | `ranova` | `vander`

Topics

“Model Specification for Repeated Measures Models” on page 9-54

ansaribradley

Ansari-Bradley test

Syntax

```
h = ansaribradley(x,y)
h = ansaribradley(x,y,Name,Value)
[h,p] = ansaribradley( ___ )
[h,p,stats] = ansaribradley( ___ )
```

Description

`h = ansaribradley(x,y)` returns a test decision for the null hypothesis that the data in vectors `x` and `y` comes from the same distribution, using the Ansari-Bradley test on page 33-100. The alternative hypothesis is that the data in `x` and `y` comes from distributions with the same median and shape but different dispersions (e.g., variances). The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`h = ansaribradley(x,y,Name,Value)` returns a test decision for the Ansari-Bradley test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level, conduct a one-sided test, or use a normal approximation to calculate the value of the test statistic.

`[h,p] = ansaribradley( ___ )` also returns the *p*-value, `p`, of the test, using any of the input arguments in the previous syntaxes.

`[h,p,stats] = ansaribradley( ___ )` also returns the structure `stats` containing information about the test statistic.

Examples

Ansari-Bradley Test for Equal Variances

Load the sample data. Create data vectors of miles per gallon (`MPG`) measurements for the model years 1982 and 1976.

```
load carsmall
x = MPG(Model_Year==82);
y = MPG(Model_Year==76);
```

Test the null hypothesis that the miles per gallon measured in cars from 1982 and 1976 have equal variances.

```
[h,p,stats] = ansaribradley(x,y)
```

```
h = 0
```

```
p = 0.8426
```

```
stats = struct with fields:
        W: 526.9000
```

```
    Wstar: 0.1986
```

The returned value of `h = 0` indicates that `ansaribradley` does not reject the null hypothesis at the default 5% significance level.

Ansari-Bradley One-Sided Hypothesis Test

Load the sample data. Create data vectors of miles per gallon (`MPG`) measurements for the model years 1982 and 1976.

```
load carsmall
x = MPG(Model_Year==82);
y = MPG(Model_Year==76);
```

Test the null hypothesis that the miles per gallon measured in cars from 1982 and 1976 have equal variances, against the alternative hypothesis that the variance of cars from 1982 is greater than that of cars from 1976.

```
[h,p,stats] = ansaribradley(x,y,'Tail','right')
```

```
h = 0

p = 0.5787

stats = struct with fields:
        W: 526.9000
    Wstar: 0.1986
```

The returned value of `h = 0` indicates that `ansaribradley` does not reject the null hypothesis that the variance in miles per gallon is the same for the two model years, when the alternative is that the variance of cars from 1982 is greater than that of cars from 1976.

Input Arguments

x — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array.

- If `x` and `y` are specified as vectors, they do not need to be the same length.
- If `x` and `y` are specified as matrices, they must have the same number of columns. `ansaribradley` performs separate tests along each column and returns a vector of results.
- If `x` and `y` are specified as multidimensional arrays on page 33-101, `ansaribradley` works along the first nonsingleton dimension on page 33-101. `x` and `y` must have the same size along all remaining dimensions.

Data Types: `single` | `double`

y — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array.

- If `x` and `y` are specified as vectors, they do not need to be the same length.
- If `x` and `y` are specified as matrices, they must have the same number of columns. `ansaribradley` performs separate tests along each column and returns a vector of results.
- If `x` and `y` are specified as multidimensional arrays on page 33-101, `ansaribradley` works along the first nonsingleton dimension on page 33-101. `x` and `y` must have the same size along all remaining dimensions.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01` specifies a right-tailed hypothesis test at the 1% significance level.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension

first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix along which to test the means, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the column means, while `'Dim',2` tests the row means.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'left'` | `'right'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternative hypothesis that the dispersion parameters of `x` and `y` are not equal.
`'right'`	Test the alternative hypothesis that the dispersion parameter of `x` is greater than that of `y`.
`'left'`	Test the alternative hypothesis that the dispersion parameter of `x` is less than that of `y`.

Example: `'Tail','right'`

Method — Computation method
'exact' | 'approximate'

Computation method for the test statistic, specified as the comma-separated pair consisting of 'Method' and one of the following.

'exact'	Compute p using an exact calculation of the distribution of the test statistic W. This is the default if *n*, the total number of rows in x and y, is 25 or less. Note that *n* is computed before any NaN values (representing missing data) are removed.
'approximate'	Compute p using a normal approximation for the statistic W*. This is the default if *n*, the total number of rows in x and y, is greater than 25.

Example: 'Method','exact'

Output Arguments

h — Hypothesis test result
1 | 0

Hypothesis test result, returned as 1 or 0.

- If h = 1, this indicates the rejection of the null hypothesis at the Alpha significance level.
- If h = 0, this indicates a failure to reject the null hypothesis at the Alpha significance level.

p — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. p is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of p cast doubt on the validity of the null hypothesis.

stats — Test statistics
structure

Test statistics for the Ansari-Bradley test, returned as a structure containing:

- W — Value of the test statistic, which is the sum of the Ansari-Bradley ranks for the x sample.
- Wstar — Approximate normal statistic W*.

More About

Ansari-Bradley Test

The Ansari-Bradley test is a nonparametric alternative to the two-sample *F*-test of equal variances. It does not require the assumption that x and y come from normal distributions. The dispersion of a distribution is generally measured by its variance or standard deviation, but the Ansari-Bradley test can be used with samples from distributions that do not have finite variances.

This test requires that the samples have equal medians. Under that assumption, and if the distributions of the samples are continuous and identical, the test is independent of the distributions. If the samples do not have the same medians, the results can be misleading. In that case, Ansari and

Bradley recommend subtracting the median, but then the distribution of the resulting test under the null hypothesis is no longer independent of the common distribution of `x` and `y`. If you want to perform the tests with medians subtracted, you should subtract the medians from `x` and `y` before calling `ansaribradley`.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

See Also

`ttest2` | `vartest2` | `vartestn`

Introduced before R2006a

aoctool

Interactive analysis of covariance

Syntax

```
aoctool(x,y,group)
aoctool(x,y,group,alpha)
aoctool(x,y,group,alpha,xname,yname,gname)
aoctool(x,y,group,alpha,xname,yname,gname,displayopt)
aoctool(x,y,group,alpha,xname,yname,gname,displayopt,model)
h = aoctool(...)
[h,atab,ctab] = aoctool(...)
[h,atab,ctab,stats] = aoctool(...)
```

Description

`aoctool(x,y,group)` fits a separate line to the column vectors, `x` and `y`, for each group defined by the values in the array `group`. `group` may be a categorical variable, numeric vector, character array, string array, or cell array of character vectors. These types of models are known as one-way analysis of covariance (ANOCOVA) models. The output consists of three figures:

- An interactive graph of the data and prediction curves
- An ANOVA table
- A table of parameter estimates

You can use the figures to change models and to test different parts of the model. More information about interactive use of the `aoctool` function appears in “Analysis of Covariance Tool” on page 9-39.

`aoctool(x,y,group,alpha)` determines the confidence levels of the prediction intervals. The confidence level is `100(1-alpha)`%. The default value of `alpha` is 0.05.

`aoctool(x,y,group,alpha,xname,yname,gname)` specifies the name to use for the `x`, `y`, and `g` variables in the graph and tables. If you enter simple variable names for the `x`, `y`, and `g` arguments, the `aoctool` function uses those names. If you enter an expression for one of these arguments, you can specify a name to use in place of that expression by supplying these arguments. For example, if you enter `m(:,2)` as the `x` argument, you might choose to enter `'Col 2'` as the `xname` argument.

`aoctool(x,y,group,alpha,xname,yname,gname,displayopt)` enables the graph and table displays when *`displayopt`* is `'on'` (default) and suppresses those displays when *`displayopt`* is `'off'`.

`aoctool(x,y,group,alpha,xname,yname,gname,displayopt,model)` specifies the initial model to fit. The value of *`model`* can be any of the following:

- `'same mean'` — Fit a single mean, ignoring grouping
- `'separate means'` — Fit a separate mean to each group
- `'same line'` — Fit a single line, ignoring grouping
- `'parallel lines'` — Fit a separate line to each group, but constrain the lines to be parallel

- `'separate lines'` — Fit a separate line to each group, with no constraints

`h = aoctool(...)` returns a vector of handles to the line objects in the plot.

`[h,atab,ctab] = aoctool(...)` returns cell arrays containing the entries in ANOVA table (`atab`) and the table of coefficient estimates (`ctab`). (You can copy a text version of either table to the clipboard by using the `Copy Text` item on the **Edit** menu.)

`[h,atab,ctab,stats] = aoctool(...)` returns a `stats` structure that you can use to perform a follow-up multiple comparison test. The ANOVA table output includes tests of the hypotheses that the slopes or intercepts are all the same, against a general alternative that they are not all the same. Sometimes it is preferable to perform a test to determine which pairs of values are significantly different, and which are not. You can use the `multcompare` function to perform such tests by supplying the `stats` structure as input. You can test either the slopes, the intercepts, or population marginal means (the heights of the curves at the mean `x` value).

Examples

This example illustrates how to fit different models non-interactively. After loading the smaller car data set and fitting a separate-slopes model, you can examine the coefficient estimates.

```
load carsmall
[h,a,c,s] = aoctool(Weight,MPG,Model_Year,0.05,...
                    '','','','off','separate lines');
c(:,1:2)
ans =
  'Term'       'Estimate'
  'Intercept'  [45.97983716833132]
  ' 70'        [-8.58050531454973]
  ' 76'        [-3.89017396094922]
  ' 82'        [12.47067927549897]
  'Slope'      [-0.00780212907455]
  ' 70'        [ 0.00195840368824]
  ' 76'        [ 0.00113831038418]
  ' 82'        [-0.00309671407243]
```

Roughly speaking, the lines relating `MPG` to `Weight` have an intercept close to 45.98 and a slope close to -0.0078. Each group's coefficients are offset from these values somewhat. For instance, the intercept for the cars made in 1970 is 45.98-8.58 = 37.40.

Next, try a fit using parallel lines. (The ANOVA table shows that the parallel-lines fit is significantly worse than the separate-lines fit.)

```
[h,a,c,s] = aoctool(Weight,MPG,Model_Year,0.05,...
                    '','','','off','parallel lines');

c(:,1:2)

ans =

  'Term'       'Estimate'
  'Intercept'  [43.38984085130596]
  ' 70'        [-3.27948192983761]
  ' 76'        [-1.35036234809006]
  ' 82'        [ 4.62984427792768]
  'Slope'      [-0.00664751826198]
```

Again, there are different intercepts for each group, but this time the slopes are constrained to be the same.

See Also

`anova1` | `multcompare` | `polytool`

Introduced before R2006a

append

Class: `TreeBagger`

Append new trees to ensemble

Syntax

```
B = append(B,other)
```

Description

`B = append(B,other)` appends the trees from the `other` ensemble to those in `B`. This method checks for consistency of the `X` and `Y` properties of the two ensembles, as well as consistency of their compact objects and out-of-bag indices, before appending the trees. The output ensemble `B` takes training parameters such as `FBoot`, `Prior`, `Cost`, and `other` from the `B` input. There is no attempt to check if these training parameters are consistent between the two objects.

See Also

`combine`

barttest

Bartlett's test

Syntax

```
ndim = barttest(x,alpha)
[ndim,prob,chisquare] = barttest(x,alpha)
```

Description

`ndim = barttest(x,alpha)` returns the number of dimensions necessary to explain the nonrandom variation in the data matrix `x` at the `alpha` significance level.

`[ndim,prob,chisquare] = barttest(x,alpha)` also returns the significance values for the hypothesis tests `prob`, and the χ^2 values associated with the tests `chisquare`.

Examples

Determine Dimensions Needed to Explain Nonrandom Data Variation

Generate a 20-by-6 matrix of random numbers from a multivariate normal distribution with mean `mu = [0 0]` and covariance `sigma = [1 0.99; 0.99 1]`.

```
rng default  % for reproducibility
mu = [0 0];
sigma = [1 0.99; 0.99 1];
X = mvnrnd(mu,sigma,20);  % columns 1 and 2
X(:,3:4) = mvnrnd(mu,sigma,20);  % columns 3 and 4
X(:,5:6) = mvnrnd(mu,sigma,20);  % columns 5 and 6
```

Determine the number of dimensions necessary to explain the nonrandom variation in data matrix X. Report the significance values for the hypothesis tests.

```
[ndim, prob] = barttest(X,0.05)
```

```
ndim = 3

prob = 5×1

    0.0000
    0.0000
    0.0000
    0.5148
    0.3370
```

The returned value of `ndim` indicates that three dimensions are necessary to explain the nonrandom variation in X.

Input Arguments

`x` — Input data

matrix of scalar values

Input data, specified as a matrix of scalar values.

Data Types: `single` | `double`

`alpha` — Significance level

`0.05` (default) | scalar value in the range `(0,1)`

Significance level of the hypothesis test, specified as a scalar value in the range `(0,1)`.

Example: `0.1`

Data Types: `single` | `double`

Output Arguments

`ndim` — Number of dimensions

positive integer value

Number of dimensions, returned as a positive integer value. The dimension is determined by a series of hypothesis tests. The test for `ndim = 1` tests the hypothesis that the variances of the data values along each principal component are equal, the test for `ndim = 2` tests the hypothesis that the variances along the second through last components are equal, and so on. The null hypothesis is that the number of dimensions is equal to the number of the largest unequal eigenvalues of the covariance matrix of `x`.

`prob` — Significance value

vector of scalar values in the range `(0,1)`

Significance value for the hypothesis tests, returned as a vector of scalar values in the range `(0,1)`. Each element in `prob` corresponds to an element of `chisquare`.

`chisquare` — Test statistics

vector of scalar values

Test statistics for each dimension's hypothesis test, returned as a vector of scalar values.

See Also

Introduced before R2006a

BayesianOptimization

Bayesian optimization results

Description

A `BayesianOptimization` object contains the results of a Bayesian optimization. It is the output of `bayesopt` or a fit function that accepts the `OptimizeHyperparameters` name-value pair such as `fitcdiscr`. In addition, a `BayesianOptimization` object contains data for each iteration of `bayesopt` that can be accessed by a plot function or an output function.

Creation

Create a `BayesianOptimization` object using the `bayesopt` function or a fit function with the `OptimizeHyperparameters` name-value pair.

Properties

Problem Definition Properties

`ObjectiveFcn` — `ObjectiveFcn` argument used by `bayesopt`
function handle

This property is read-only.

`ObjectiveFcn` argument used by `bayesopt`, returned as a function handle.

- If you call `bayesopt` directly, `ObjectiveFcn` is the `bayesopt` objective function argument.
- If you call a fit function containing the `'OptimizeHyperparameters'` name-value pair argument, `ObjectiveFcn` is a function handle that returns the misclassification rate for classification or returns the logarithm of one plus the cross-validation loss for regression, measured by five-fold cross-validation.

Data Types: `function_handle`

`VariableDescriptions` — `VariableDescriptions` argument that `bayesopt` used
vector of `optimizableVariable` objects

This property is read-only.

`VariableDescriptions` argument that `bayesopt` used, returned as a vector of `optimizableVariable` objects.

- If you called `bayesopt` directly, `VariableDescriptions` is the `bayesopt` variable description argument.
- If you called a fit function with the `OptimizeHyperparameters` name-value pair, `VariableDescriptions` is the vector of hyperparameters.

`Options` — Options that `bayesopt` used

structure

This property is read-only.

Options that `bayesopt` used, returned as a structure.

- If you called `bayesopt` directly, `Options` is the options used in `bayesopt`, which are the name-value pairs See `bayesopt` “Input Arguments” on page 33-132.
- If you called a fit function with the `OptimizeHyperparameters` name-value pair, `Options` are the default `bayesopt` options, modified by the `HyperparameterOptimizationOptions` name-value pair.

`Options` is a read-only structure containing the following fields.

Option Name	Meaning
`AcquisitionFunctionName`	Acquisition function name. See “Acquisition Function Types” on page 10-3.
`IsObjectiveDeterministic`	`true` means the objective function is deterministic, `false` otherwise.
`ExplorationRatio`	Used only when `AcquisitionFunctionName` is `'expected-improvement-plus'` or `'expected-improvement-per-second-plus'`. See “Plus” on page 10-5.
`MaxObjectiveEvaluations`	Objective function evaluation limit.
`MaxTime`	Time limit.
`XConstraintFcn`	Deterministic constraints on variables. See “Deterministic Constraints — XConstraintFcn” on page 10-38.
`ConditionalVariableFcn`	Conditional variable constraints. See “Conditional Constraints — ConditionalVariableFcn” on page 10-39.
`NumCoupledConstraints`	Number of coupled constraints. See “Coupled Constraints” on page 10-40.
`CoupledConstraintTolerances`	Coupled constraint tolerances. See “Coupled Constraints” on page 10-40.
`AreCoupledConstraintsDeterministic`	Logical vector specifying whether each coupled constraint is deterministic.
`Verbose`	Command-line display level.
`OutputFcn`	Function called after each iteration. See “Bayesian Optimization Output Functions” on page 10-19.
`SaveVariableName`	Variable name for the `@assignInBase` output function.
`SaveFileName`	File name for the `@saveToFile` output function.

Option Name	Meaning
`PlotFcn`	Plot function called after each iteration. See “Bayesian Optimization Plot Functions” on page 10-11
`InitialX`	Points where `bayesopt` evaluated the objective function.
`InitialObjective`	Objective function values at `InitialX`.
`InitialConstraintViolations`	Coupled constraint function values at `InitialX`.
`InitialErrorValues`	Error values at `InitialX`.
`InitialObjectiveEvaluationTimes`	Objective function evaluation times at `InitialX`.
`InitialIterationTimes`	Time for each iteration, including objective function evaluation and other computations.

Data Types: `struct`

Solution Properties

`MinObjective` — Minimum observed value of objective function
real scalar

This property is read-only.

Minimum observed value of objective function, returned as a real scalar. When there are coupled constraints or evaluation errors, this value is the minimum over all observed points that are feasible according to the final constraint and Error models.

Data Types: `double`

`XAtMinObjective` — Observed point with minimum objective function value
1-by-`D` table

This property is read-only.

Observed point with minimum objective function value, returned as a `1`-by-`D` table, where `D` is the number of variables.

Data Types: `table`

`MinEstimatedObjective` — Minimum estimated value of objective function
real scalar

This property is read-only.

Minimum estimated value of objective function, returned as a real scalar. `MinEstimatedObjective` uses the final objective model.

`MinEstimatedObjective` is the same as the `CriterionValue` result of `bestPoint` with default criterion.

Data Types: `double`

`XAtMinEstimatedObjective` — Point with minimum estimated objective function value
1-by-`D` table

This property is read-only.

Point with minimum estimated objective function value, returned as a 1-by-D table, where D is the number of variables. `XAtMinEstimatedObjective` uses the final objective model.

Data Types: `table`

`NumObjectiveEvaluations` — Number of objective function evaluations
positive integer

This property is read-only.

Number of objective function evaluations, returned as a positive integer. This includes the initial evaluations to form a posterior model as well as evaluation during the optimization iterations.

Data Types: `double`

`TotalElapsedTime` — Total elapsed time of optimization in seconds
positive scalar

This property is read-only.

Total elapsed time of optimization in seconds, returned as a positive scalar.

Data Types: `double`

`NextPoint` — Next point to evaluate if optimization continues
1-by-D table

This property is read-only.

Next point to evaluate if optimization continues, returned as a 1-by-D table, where D is the number of variables.

Data Types: `table`

Trace Properties

`XTrace` — Points where the objective function was evaluated
T-by-D table

This property is read-only.

Points where the objective function was evaluated, returned as a T-by-D table, where T is the number of evaluation points and D is the number of variables.

Data Types: `table`

`ObjectiveTrace` — Objective function values
column vector of length T

This property is read-only.

Objective function values, returned as a column vector of length T, where T is the number of evaluation points. `ObjectiveTrace` contains the history of objective function evaluations.

Data Types: `double`

`ObjectiveEvaluationTimeTrace` — Objective function evaluation times
column vector of length `T`

This property is read-only.

Objective function evaluation times, returned as a column vector of length `T`, where `T` is the number of evaluation points. `ObjectiveEvaluationTimeTrace` includes the time in evaluating coupled constraints, because the objective function computes these constraints.

Data Types: `double`

`IterationTimeTrace` — Iteration times
column vector of length `T`

This property is read-only.

Iteration times, returned as a column vector of length `T`, where `T` is the number of evaluation points. `IterationTimeTrace` includes both objective function evaluation time and other overhead.

Data Types: `double`

`ConstraintsTrace` — Coupled constraint values
T-by-K array

This property is read-only.

Coupled constraint values, returned as a `T`-by-`K` array, where `T` is the number of evaluation points and `K` is the number of coupled constraints.

Data Types: `double`

`ErrorTrace` — Error indications
column vector of length `T` of `-1` or `1` entries

This property is read-only.

Error indications, returned as a column vector of length `T` of `-1` or `1` entries, where `T` is the number of evaluation points. Each `1` entry indicates that the objective function errored or returned `NaN` on the corresponding point in `XTrace`. Each `-1` entry indicates that the objective function value was computed.

Data Types: `double`

`FeasibilityTrace` — Feasibility indications
logical column vector of length `T`

This property is read-only.

Feasibility indications, returned as a logical column vector of length `T`, where `T` is the number of evaluation points. Each `1` entry indicates that the final constraint model predicts feasibility at the corresponding point in `XTrace`.

Data Types: `logical`

`FeasibilityProbabilityTrace` — Probability that evaluation point is feasible
column vector of length `T`

This property is read-only.

Probability that evaluation point is feasible, returned as a column vector of length `T`, where `T` is the number of evaluation points. The probabilities come from the final constraint model, including the error constraint model, on the corresponding points in `XTrace`.

Data Types: `double`

`IndexOfMinimumTrace` — Which evaluation gave minimum feasible objective
column vector of integer indices of length `T`

This property is read-only.

Which evaluation gave minimum feasible objective, returned as a column vector of integer indices of length `T`, where `T` is the number of evaluation points. Feasibility is determined with respect to the constraint models that existed at each iteration, including the error constraint model.

Data Types: `double`

`ObjectiveMinimumTrace` — Minimum observed objective
column vector of length `T`

This property is read-only.

Minimum observed objective, returned as a column vector of integer indices of length `T`, where `T` is the number of evaluation points.

Data Types: `double`

`EstimatedObjectiveMinimumTrace` — Minimum estimated objective
column vector of length `T`

This property is read-only.

Minimum estimated objective, returned as a column vector of integer indices of length `T`, where `T` is the number of evaluation points. The estimated objective at each iteration is determined with respect to the objective model that existed at that iteration.

Data Types: `double`

`UserDataTrace` — Auxiliary data from the objective function
cell array of length `T`

This property is read-only.

Auxiliary data from the objective function, returned as a cell array of length `T`, where `T` is the number of evaluation points. Each entry in the cell array is the `UserData` returned in the third output of the objective function.

Data Types: `cell`

Object Functions

bestPoint	Best point in a Bayesian optimization according to a criterion
plot	Plot Bayesian optimization results
predictConstraints	Predict coupled constraint violations at a set of points
predictError	Predict error value at a set of points
predictObjective	Predict objective function at a set of points
predictObjectiveEvaluationTime	Predict objective function run times at a set of points

resume	Resume a Bayesian optimization

Examples

Create a BayesianOptimization Object Using bayesopt

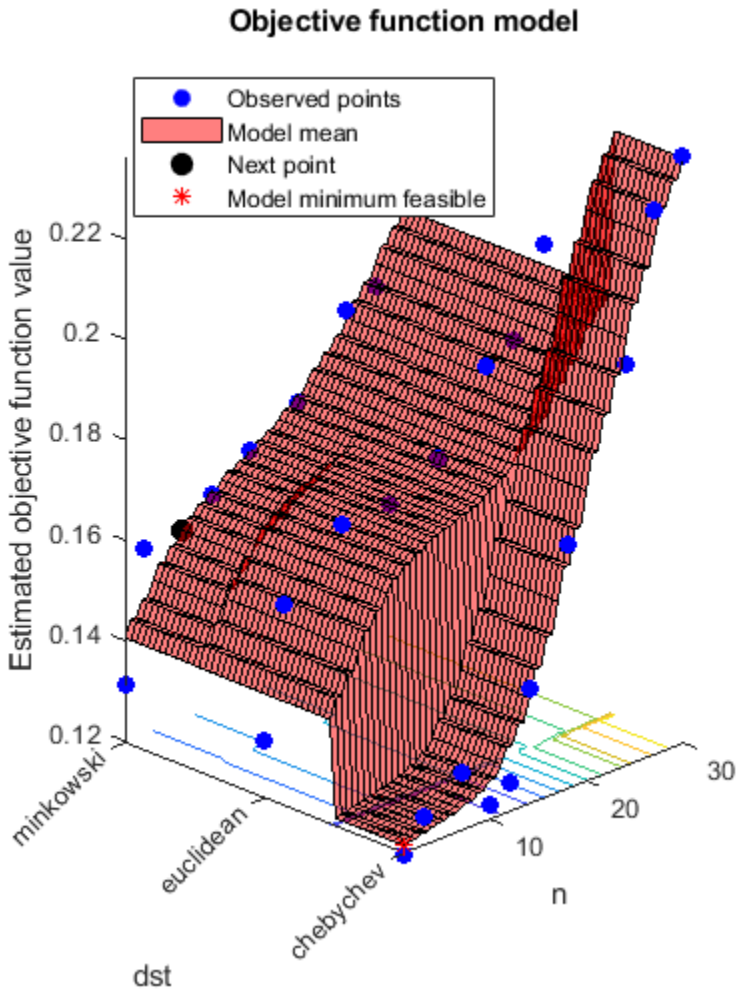
This example shows how to create a `BayesianOptimization` object by using `bayesopt` to minimize cross-validation loss.

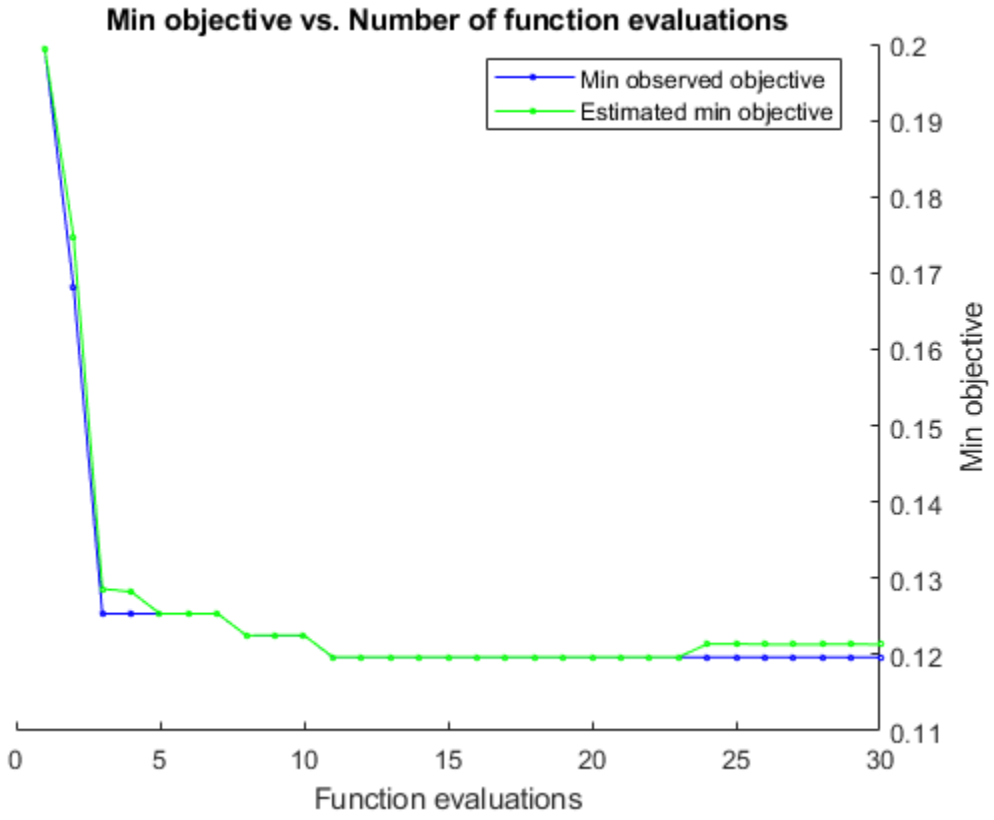
Optimize hyperparameters of a KNN classifier for the `ionosphere` data, that is, find KNN hyperparameters that minimize the cross-validation loss. Have `bayesopt` minimize over the following hyperparameters:

- Nearest-neighborhood sizes from 1 to 30
- Distance functions `'chebychev'`, `'euclidean'`, and `'minkowski'`.

For reproducibility, set the random seed, set the partition, and set the `AcquisitionFunctionName` option to `'expected-improvement-plus'`. To suppress iterative display, set `'Verbose'` to `0`. Pass the partition `c` and fitting data `X` and `Y` to the objective function `fun` by creating `fun` as an anonymous function that incorporates this data. See “Parameterizing Functions”.

```
load ionosphere
rng default
num = optimizableVariable('n',[1,30],'Type','integer');
dst = optimizableVariable('dst',{'chebychev','euclidean','minkowski'},'Type','categorical');
c = cvpartition(351,'Kfold',5);
fun = @(x)kfoldLoss(fitcknn(X,Y,'CVPartition',c,'NumNeighbors',x.n,...
    'Distance',char(x.dst),'NSMethod','exhaustive'));
results = bayesopt(fun,[num,dst],'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus')
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.22
0.2
0.18
0.16
0.14
0.12
minkowski
euclidean
chebychev
dst
10
20
30
n




```
results =
  BayesianOptimization with properties:

                      ObjectiveFcn: [function_handle]
              VariableDescriptions: [1x2 optimizableVariable]
                           Options: [1x1 struct]
                      MinObjective: 0.1197
                   XAtMinObjective: [1x2 table]
             MinEstimatedObjective: 0.1213
          XAtMinEstimatedObjective: [1x2 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 36.2488
                         NextPoint: [1x2 table]
                            XTrace: [30x2 table]
                    ObjectiveTrace: [30x1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30x1 cell}
       ObjectiveEvaluationTimeTrace: [30x1 double]
                 IterationTimeTrace: [30x1 double]
                         ErrorTrace: [30x1 double]
                  FeasibilityTrace: [30x1 logical]
       FeasibilityProbabilityTrace: [30x1 double]
              IndexOfMinimumTrace: [30x1 double]
            ObjectiveMinimumTrace: [30x1 double]
     EstimatedObjectiveMinimumTrace: [30x1 double]
```

Create a BayesianOptimization Object Using a Fit Function

This example shows how to minimize the cross-validation loss in the `ionosphere` data using Bayesian optimization of an SVM classifier.

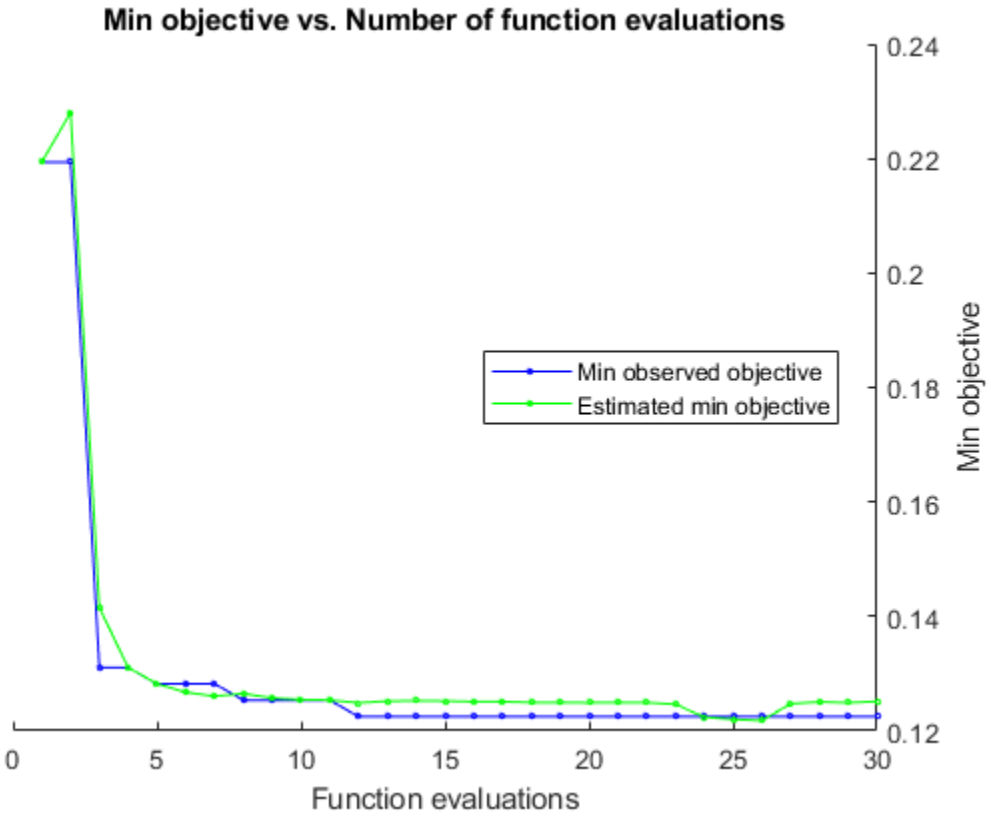
Load the data.

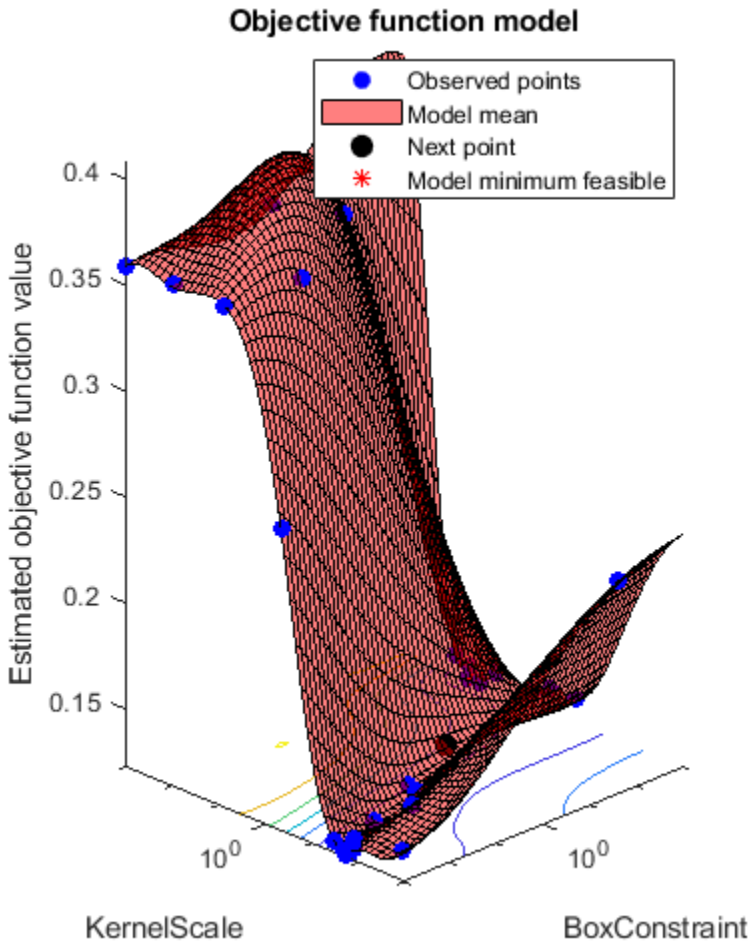
```
load ionosphere
```

Optimize the classification using the `'auto'` parameters.

```
rng default % For reproducibility
Mdl = fitcsvm(X,Y,'OptimizeHyperparameters','auto')
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.21937 |      16.012 |     0.21937 |     0.21937 |       64.836 |    0.001
|    2 | Accept |     0.35897 |     0.13391 |     0.21937 |     0.22807 |     0.036335 |        5
|    3 | Best   |     0.13105 |      5.8272 |     0.13105 |     0.14149 |    0.0022147 |    0.002
|    4 | Accept |     0.35897 |     0.12455 |     0.13105 |     0.13108 |       5.1259 |        9
|    5 | Best   |     0.12821 |     0.14718 |     0.12821 |     0.12823 |    0.0010294 |      0.0
|    6 | Accept |     0.12821 |     0.17625 |     0.12821 |     0.12673 |    0.0010607 |     0.01
|    7 | Accept |      0.1339 |      1.2937 |     0.12821 |     0.12605 |     0.027234 |     0.01
|    8 | Best   |     0.12536 |     0.17005 |     0.12536 |     0.12652 |    0.0010012 |     0.01
|    9 | Accept |     0.12536 |     0.15457 |     0.12536 |     0.12573 |    0.0010018 |     0.01
|   10 | Accept |     0.12536 |     0.15465 |     0.12536 |     0.12546 |    0.0010684 |     0.02
|   11 | Accept |     0.12536 |     0.15336 |     0.12536 |     0.12538 |    0.0010061 |     0.01
|   12 | Best   |     0.12251 |      0.1696 |     0.12251 |     0.12488 |    0.0011285 |     0.02
|   13 | Accept |     0.12821 |     0.18511 |     0.12251 |     0.12521 |    0.0017959 |     0.02
|   14 | Accept |     0.35897 |     0.13399 |     0.12251 |     0.12525 |       1.4879 |        9
|   15 | Accept |     0.35897 |     0.11378 |     0.12251 |     0.12522 |    0.0010867 |       97
|   16 | Accept |     0.26496 |     0.10713 |     0.12251 |     0.12511 |    0.0010017 |        0
|   17 | Accept |     0.35897 |     0.14043 |     0.12251 |      0.1251 |    0.0010157 |       99
|   18 | Accept |     0.12821 |      4.5411 |     0.12251 |     0.12494 |       997.24 |        1
|   19 | Accept |      0.1339 |      14.724 |     0.12251 |     0.12495 |       971.08 |      0.1
|   20 | Accept |      0.1339 |      11.238 |     0.12251 |     0.12491 |       949.69 |      0.6
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|   21 | Accept |     0.35897 |     0.14778 |     0.12251 |     0.12496 |       857.87 |        5
|   22 | Accept |     0.35897 |     0.15997 |     0.12251 |     0.12494 |    0.0010011 |        7
|   23 | Accept |     0.13105 |     0.14724 |     0.12251 |     0.12465 |        996.4 |        7
|   24 | Accept |     0.12251 |     0.18986 |     0.12251 |      0.1224 |       999.12 |       23
|   25 | Accept |     0.12251 |     0.15114 |     0.12251 |     0.12203 |       998.76 |       43
|   26 | Accept |     0.13105 |      0.2873 |     0.12251 |     0.12176 |       991.48 |       10
|   27 | Accept |     0.12821 |     0.15067 |     0.12251 |     0.12475 |       997.54 |       38
|   28 | Accept |     0.12536 |     0.20699 |     0.12251 |     0.12505 |     0.013305 |     0.00
|   29 | Accept |     0.13105 |     0.25356 |     0.12251 |     0.12491 |      0.13267 |      0.1
|   30 | Accept |     0.13105 |     0.40373 |     0.12251 |     0.12514 |      0.11138 |      0.0
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0
5
10
15
20
25
30




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 98.1515 seconds
Total objective function evaluation time: 57.7998

Best observed feasible point:
    BoxConstraint    KernelScale
    _____________    ___________

      0.0011285       0.020533

Observed objective function value = 0.12251
Estimated objective function value = 0.12528
Function evaluation time = 0.1696

Best estimated feasible point (according to models):
    BoxConstraint    KernelScale
    _____________    ___________
```

```
      0.0010684       0.020615

Estimated objective function value = 0.12514
Estimated function evaluation time = 0.15669

Mdl =
  ClassificationSVM
                         ResponseName: 'Y'
                CategoricalPredictors: []
                           ClassNames: {'b'  'g'}
                       ScoreTransform: 'none'
                      NumObservations: 351
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                                Alpha: [94x1 double]
                                 Bias: -5.4559
                     KernelParameters: [1x1 struct]
                       BoxConstraints: [351x1 double]
                      ConvergenceInfo: [1x1 struct]
                      IsSupportVector: [351x1 logical]
                               Solver: 'SMO'


  Properties, Methods
```

The fit achieved about 12% loss for the default 5-fold cross validation.

Examine the `BayesianOptimization` object that is returned in the `HyperparameterOptimizationResults` property of the returned model.

```
disp(Mdl.HyperparameterOptimizationResults)
```

```
  BayesianOptimization with properties:

                      ObjectiveFcn: @createObjFcn/inMemoryObjFcn
              VariableDescriptions: [5x1 optimizableVariable]
                           Options: [1x1 struct]
                      MinObjective: 0.1225
                   XAtMinObjective: [1x2 table]
             MinEstimatedObjective: 0.1251
          XAtMinEstimatedObjective: [1x2 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 98.1515
                         NextPoint: [1x2 table]
                            XTrace: [30x2 table]
                    ObjectiveTrace: [30x1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30x1 cell}
      ObjectiveEvaluationTimeTrace: [30x1 double]
                IterationTimeTrace: [30x1 double]
                        ErrorTrace: [30x1 double]
                  FeasibilityTrace: [30x1 logical]
       FeasibilityProbabilityTrace: [30x1 double]
                IndexOfMinimumTrace: [30x1 double]
              ObjectiveMinimumTrace: [30x1 double]
    EstimatedObjectiveMinimumTrace: [30x1 double]
```

See Also

`bayesopt` | `fitcdiscr` | `fitcecoc` | `fitcensemble` | `fitcknn` | `fitclinear` | `fitcnb` | `fitcsvm` | `fitctree` | `fitrensemble` | `fitrgp` | `fitrlinear` | `fitrsvm` | `fitrtree`

Topics

“Bayesian Optimization Workflow” on page 10-25

Introduced in R2016b

bayesopt

Select optimal machine learning hyperparameters using Bayesian optimization

Syntax

```
results = bayesopt(fun,vars)
results = bayesopt(fun,vars,Name,Value)
```

Description

`results = bayesopt(fun,vars)` attempts to find values of `vars` that minimize `fun(vars)`.

Note To include extra parameters in an objective function, see “Parameterizing Functions”.

`results = bayesopt(fun,vars,Name,Value)` modifies the optimization process according to the `Name,Value` arguments.

Examples

Create a BayesianOptimization Object Using bayesopt

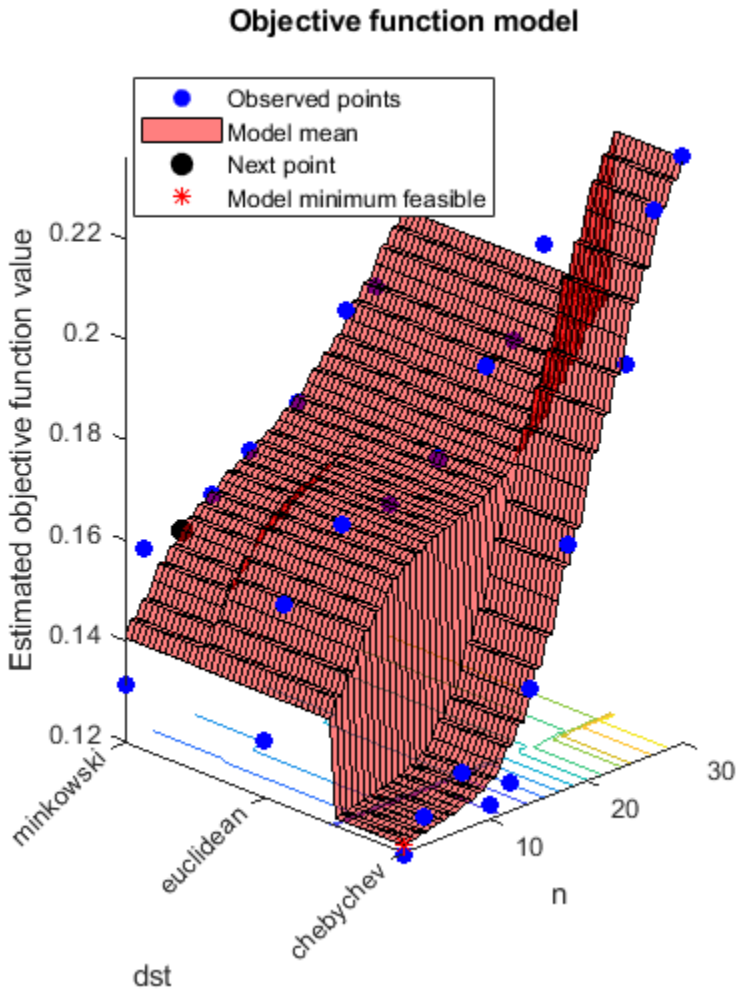
This example shows how to create a `BayesianOptimization` object by using `bayesopt` to minimize cross-validation loss.

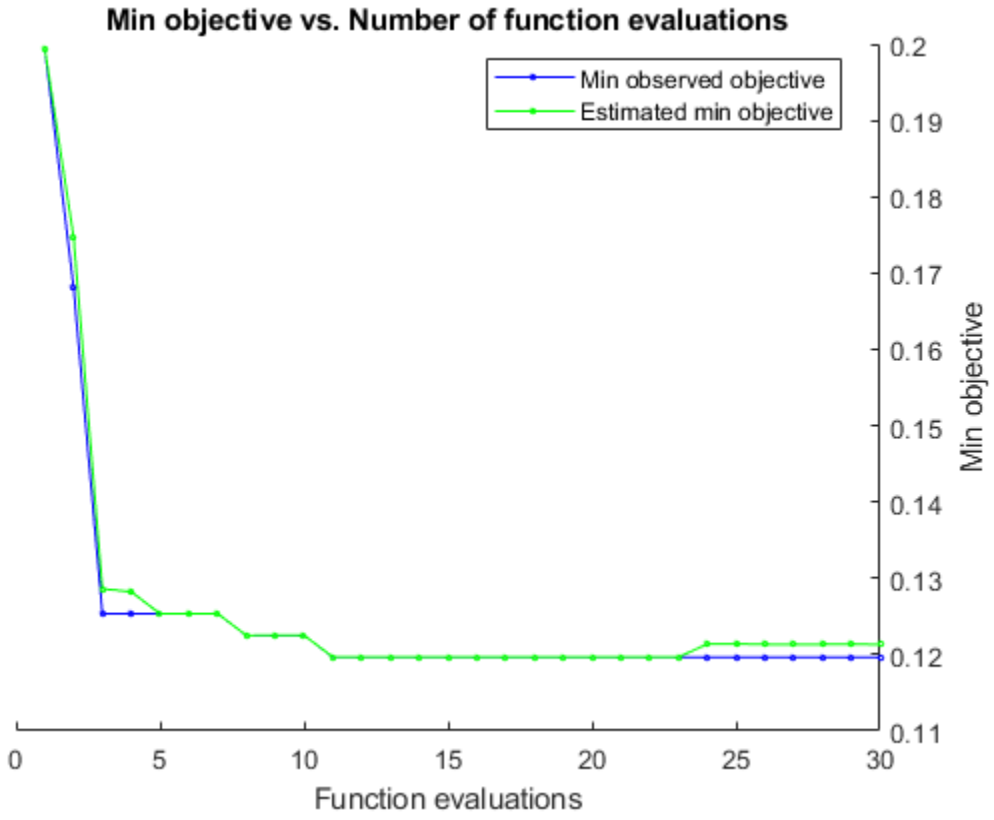
Optimize hyperparameters of a KNN classifier for the `ionosphere` data, that is, find KNN hyperparameters that minimize the cross-validation loss. Have `bayesopt` minimize over the following hyperparameters:

- Nearest-neighborhood sizes from 1 to 30
- Distance functions `'chebychev'`, `'euclidean'`, and `'minkowski'`.

For reproducibility, set the random seed, set the partition, and set the `AcquisitionFunctionName` option to `'expected-improvement-plus'`. To suppress iterative display, set `'Verbose'` to `0`. Pass the partition `c` and fitting data `X` and `Y` to the objective function `fun` by creating `fun` as an anonymous function that incorporates this data. See “Parameterizing Functions”.

```
load ionosphere
rng default
num = optimizableVariable('n',[1,30],'Type','integer');
dst = optimizableVariable('dst',{'chebychev','euclidean','minkowski'},'Type','categorical');
c = cvpartition(351,'Kfold',5);
fun = @(x)kfoldLoss(fitcknn(X,Y,'CVPartition',c,'NumNeighbors',x.n,...
    'Distance',char(x.dst),'NSMethod','exhaustive'));
results = bayesopt(fun,[num,dst],'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus')
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.22
0.2
0.18
0.16
0.14
0.12
minkowski
euclidean
chebychev
dst
10
20
30
n




```
results =
  BayesianOptimization with properties:

                      ObjectiveFcn: [function_handle]
              VariableDescriptions: [1x2 optimizableVariable]
                           Options: [1x1 struct]
                      MinObjective: 0.1197
                   XAtMinObjective: [1x2 table]
             MinEstimatedObjective: 0.1213
          XAtMinEstimatedObjective: [1x2 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 36.2488
                         NextPoint: [1x2 table]
                            XTrace: [30x2 table]
                    ObjectiveTrace: [30x1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30x1 cell}
       ObjectiveEvaluationTimeTrace: [30x1 double]
                 IterationTimeTrace: [30x1 double]
                         ErrorTrace: [30x1 double]
                  FeasibilityTrace: [30x1 logical]
       FeasibilityProbabilityTrace: [30x1 double]
              IndexOfMinimumTrace: [30x1 double]
            ObjectiveMinimumTrace: [30x1 double]
     EstimatedObjectiveMinimumTrace: [30x1 double]
```

Bayesian Optimization with Coupled Constraints

A coupled constraint is one that can be evaluated only by evaluating the objective function. In this case, the objective function is the cross-validated loss of an SVM model. The coupled constraint is that the number of support vectors is no more than 100. The model details are in “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45.

Create the data for classification.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
redpts = zeros(100,2);
grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
cdata = [grnpts;redpts];
grp = ones(200,1);
grp(101:200) = -1;
c = cvpartition(200,'KFold',10);
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
```

The objective function is the cross-validation loss of the SVM model for partition `c`. The coupled constraint is the number of support vectors minus 100.5. This ensures that 100 support vectors give a negative constraint value, but 101 support vectors give a positive value. The model has 200 data points, so the coupled constraint values range from -99.5 (there is always at least one support vector) to 99.5. Positive values mean the constraint is not satisfied.

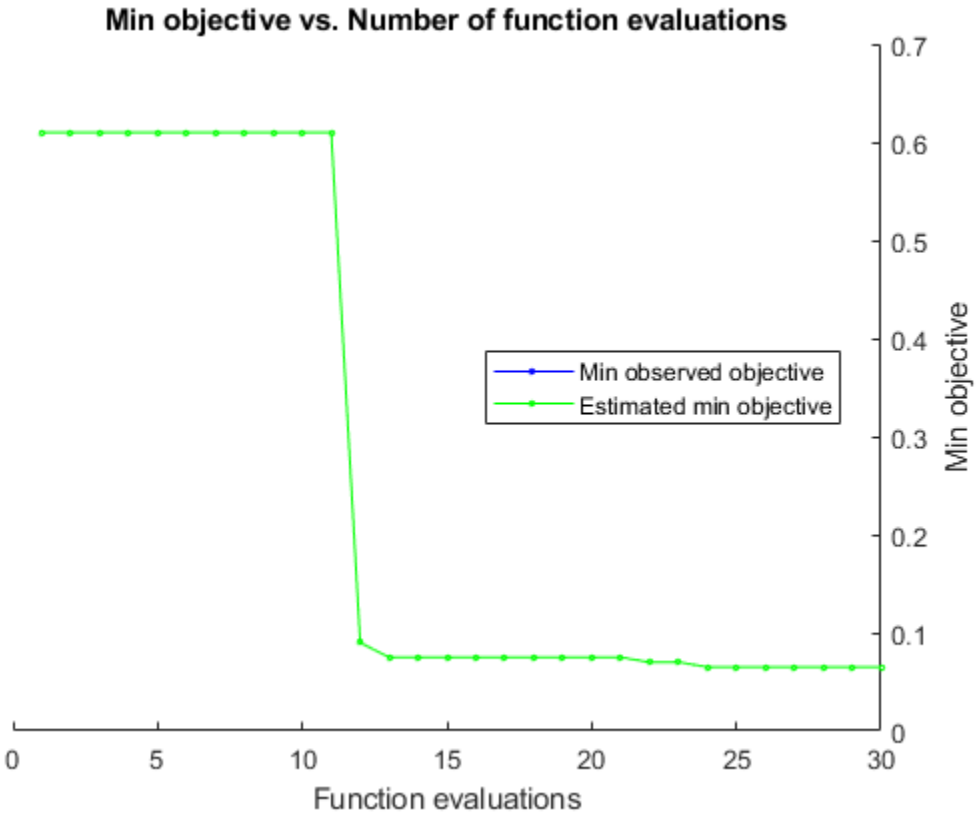
```
function [objective,constraint] = mysvmfun(x,cdata,grp,c)
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'BoxConstraint',x.box,...
    'KernelScale',x.sigma);
cvModel = crossval(SVMModel,'CVPartition',c);
objective = kfoldLoss(cvModel);
constraint = sum(SVMModel.IsSupportVector)-100.5;
```

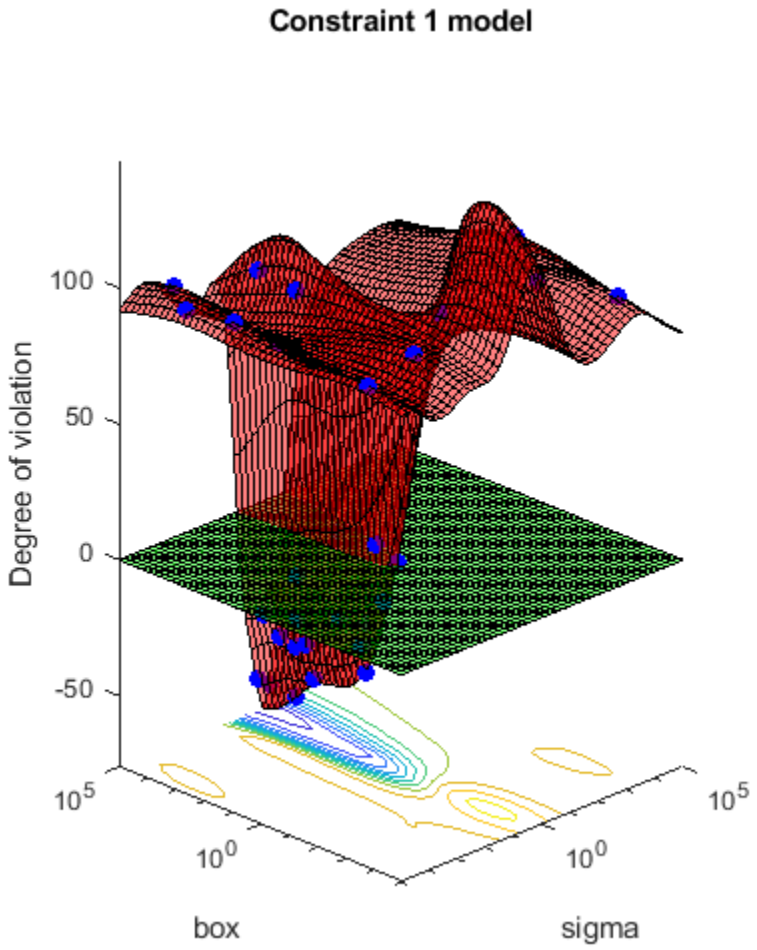
Pass the partition `c` and fitting data `cdata` and `grp` to the objective function `fun` by creating `fun` as an anonymous function that incorporates this data. See “Parameterizing Functions”.

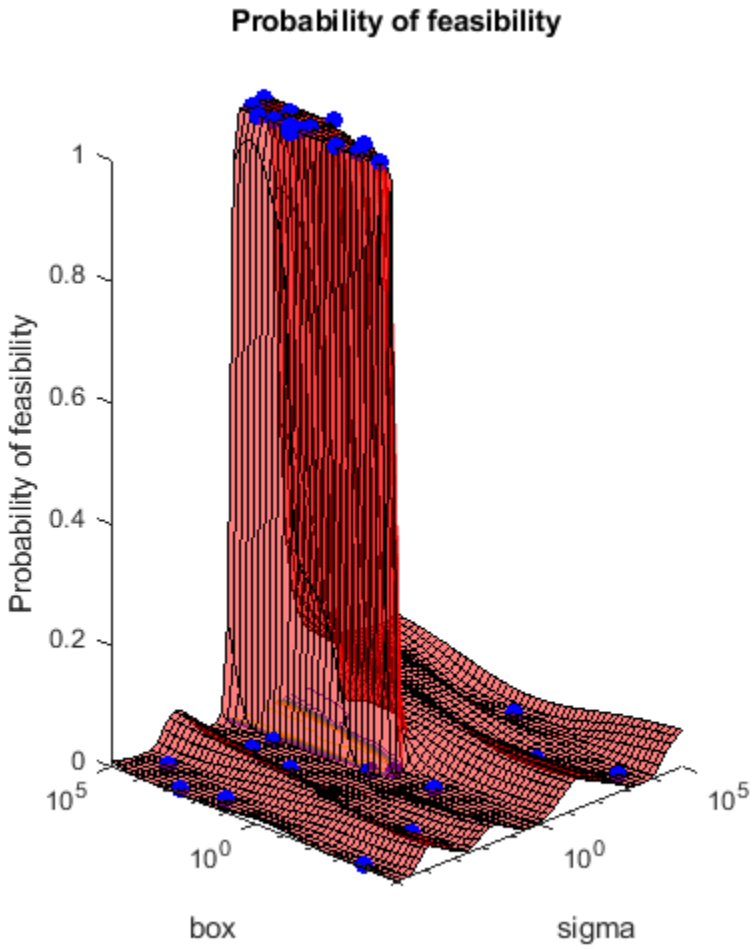
```
fun = @(x)mysvmfun(x,cdata,grp,c);
```

Set the `NumCoupledConstraints` to `1` so the optimizer knows that there is a coupled constraint. Set options to plot the constraint model.

```
results = bayesopt(fun,[sigma,box],'IsObjectiveDeterministic',true,...
    'NumCoupledConstraints',1,'PlotFcn',...
    {@plotMinObjective,@plotConstraintModels},...
    'AcquisitionFunctionName','expected-improvement-plus','Verbose',0);
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25
30


Constraint 1 model
Degree of violation
100
50
0
-50
10^5
10^0
box
10^5
10^0
sigma



Most points lead to an infeasible number of support vectors.

Parallel Bayesian Optimization

Improve the speed of a Bayesian optimization by using parallel objective function evaluation.

Prepare variables and the objective function for Bayesian optimization.

The objective function is the cross-validation error rate for the ionosphere data, a binary classification problem. Use `fitcsvm` as the classifier, with `BoxConstraint` and `KernelScale` as the parameters to optimize.

```
load ionosphere
box = optimizableVariable('box',[1e-4,1e3],'Transform','log');
kern = optimizableVariable('kern',[1e-4,1e3],'Transform','log');
vars = [box,kern];
fun = @(vars)kfoldLoss(fitcsvm(X,Y,'BoxConstraint',vars.box,'KernelScale',vars.kern,...
    'Kfold',5));
```

Search for the parameters that give the lowest cross-validation error by using parallel Bayesian optimization.

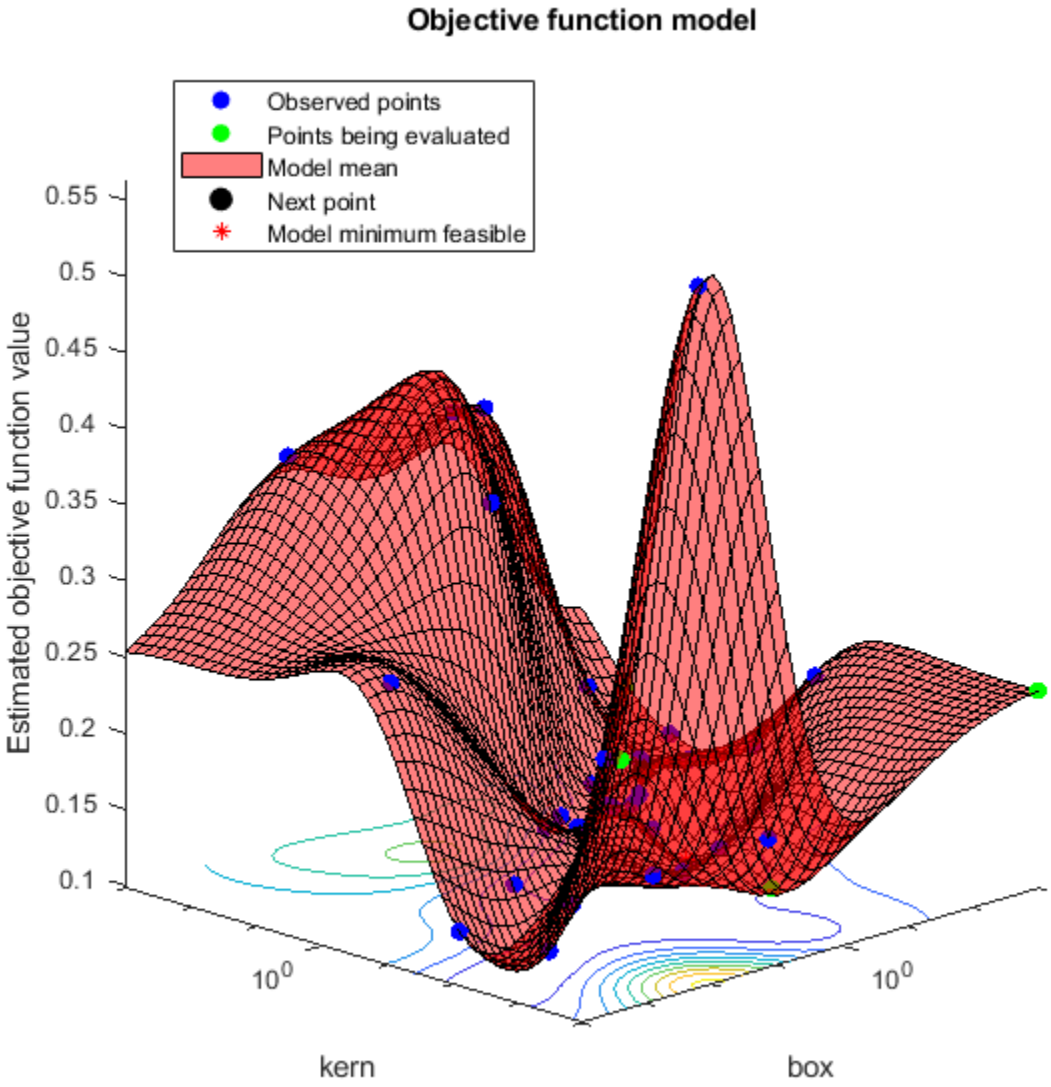
```
results = bayesopt(fun,vars,'UseParallel',true);
```

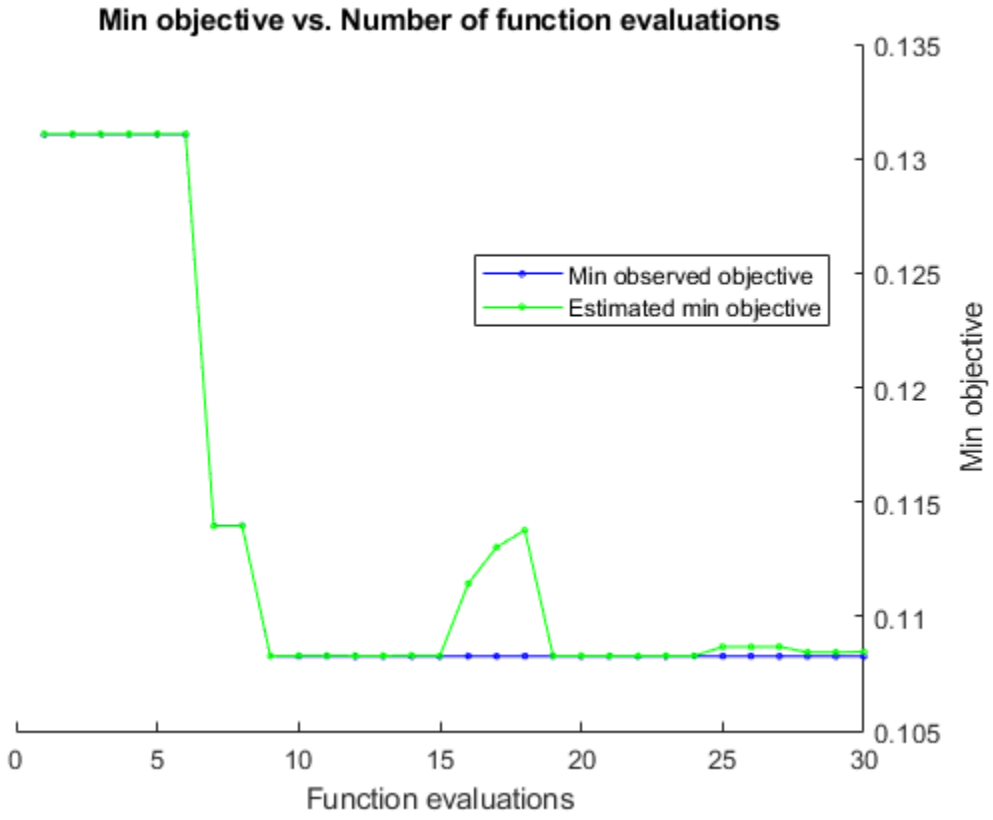
```
Copying objective function to workers...
Done copying objective function to workers.
```

Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	box
1	2	Accept	0.2735	0.56171	0.13105	0.13108	0.0002608
2	2	Accept	0.35897	0.4062	0.13105	0.13108	3.6999
3	2	Accept	0.13675	0.42727	0.13105	0.13108	0.33594
4	2	Accept	0.35897	0.4453	0.13105	0.13108	0.014127
5	2	Best	0.13105	0.45503	0.13105	0.13108	0.29713
6	6	Accept	0.35897	0.16605	0.13105	0.13108	8.1878
7	5	Best	0.11396	0.51146	0.11396	0.11395	8.7331
8	5	Accept	0.14245	0.24943	0.11396	0.11395	0.0020774
9	6	Best	0.10826	4.0711	0.10826	0.10827	0.0015925
10	6	Accept	0.25641	16.265	0.10826	0.10829	0.00057357
11	6	Accept	0.1339	15.581	0.10826	0.10829	1.4553
12	6	Accept	0.16809	19.585	0.10826	0.10828	0.26919
13	6	Accept	0.20513	18.637	0.10826	0.10828	369.59
14	6	Accept	0.12536	0.11382	0.10826	0.10829	5.7059
15	6	Accept	0.13675	2.63	0.10826	0.10828	984.19
16	6	Accept	0.12821	2.0743	0.10826	0.11144	0.0063411
17	6	Accept	0.1339	0.1939	0.10826	0.11302	0.00010225
18	6	Accept	0.12821	0.20933	0.10826	0.11376	7.7447
19	4	Accept	0.55556	17.564	0.10826	0.10828	0.0087593
20	4	Accept	0.1396	16.473	0.10826	0.10828	0.054844

Iter	Active workers	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	box
21	4	Accept	0.1339	0.17127	0.10826	0.10828	9.2668
22	4	Accept	0.12821	0.089065	0.10826	0.10828	12.265
23	4	Accept	0.12536	0.073586	0.10826	0.10828	1.3355
24	4	Accept	0.12821	0.08038	0.10826	0.10828	131.51
25	3	Accept	0.11111	10.687	0.10826	0.10867	1.4795
26	3	Accept	0.13675	0.18626	0.10826	0.10867	2.0513
27	6	Accept	0.12821	0.078559	0.10826	0.10868	980.04

```
|   28 |       5 | Accept |     0.33048 |    0.089844 |     0.10826 |     0.10843 |      0.41821
|   29 |       5 | Accept |     0.16239 |     0.12688 |     0.10826 |     0.10843 |       172.39

|   30 |       5 | Accept |     0.11966 |     0.14597 |     0.10826 |     0.10846 |       639.15
```






```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 48.2085 seconds.
Total objective function evaluation time: 128.3472

Best observed feasible point:
       box         kern
    _________    _________

    0.0015925    0.0050225

Observed objective function value = 0.10826
Estimated objective function value = 0.10846
Function evaluation time = 4.0711

Best estimated feasible point (according to models):
       box         kern
    _________    _________

    0.0015925    0.0050225

Estimated objective function value = 0.10846
Estimated function evaluation time = 2.8307
```

Return the best feasible point in the Bayesian model `results` by using the `bestPoint` function. Use the default criterion `min-visited-upper-confidence-interval`, which determines the best feasible point as the visited point that minimizes an upper confidence interval on the objective function value.

```
zbest = bestPoint(results)
```

```
zbest=1×2 table
       box          kern
    _________    _________

    0.0015925    0.0050225
```

The table `zbest` contains the optimal estimated values for the `'BoxConstraint'` and `'KernelScale'` name-value pair arguments. Use these values to train a new optimized classifier.

```
Mdl = fitcsvm(X,Y,'BoxConstraint',zbest.box,'KernelScale',zbest.kern);
```

Observe that the optimal parameters are in `Mdl`.

```
Mdl.BoxConstraints(1)
```

```
ans = 0.0016
```

```
Mdl.KernelParameters.Scale
```

```
ans = 0.0050
```

Input Arguments

`fun` — Objective function

function handle | `parallel.pool.Constant` whose `Value` is a function handle

Objective function, specified as a function handle or, when the `UseParallel` name-value pair is `true`, a `parallel.pool.Constant` whose `Value` is a function handle. Typically, `fun` returns a measure of loss (such as a misclassification error) for a machine learning model that has tunable hyperparameters to control its training. `fun` has these signatures:

```
objective = fun(x)
% or
[objective,constraints] = fun(x)
% or
[objective,constraints,UserData] = fun(x)
```

`fun` accepts `x`, a 1-by-`D` table of variable values, and returns `objective`, a real scalar representing the objective function value `fun(x)`.

Optionally, `fun` also returns:

- `constraints`, a real vector of coupled constraint violations. For a definition, see “Coupled Constraints” on page 33-140. `constraint(j) > 0` means constraint `j` is violated. `constraint(j) < 0` means constraint `j` is satisfied.
- `UserData`, an entity of any type (such as a scalar, matrix, structure, or object). For an example of a custom plot function that uses `UserData`, see “Create a Custom Plot Function” on page 10-12.

For details about using `parallel.pool.Constant` with `bayesopt`, see “Placing the Objective Function on Workers” on page 10-8.

Example: `@objfun`

Data Types: `function_handle`

`vars` — Variable descriptions

vector of `optimizableVariable` objects defining the hyperparameters to be tuned

Variable descriptions, specified as a vector of `optimizableVariable` objects defining the hyperparameters to be tuned.

Example: `[X1,X2]`, where `X1` and `X2` are `optimizableVariable` objects

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `results = bayesopt(fun,vars,'AcquisitionFunctionName','expected-improvement-plus')`

Algorithm Control

`AcquisitionFunctionName` — Function to choose next evaluation point

`'expected-improvement-per-second-plus'` (default) | `'expected-improvement'` | `'expected-improvement-plus'` | `'expected-improvement-per-second'` | `'lower-confidence-bound'` | `'probability-of-improvement'`

Function to choose next evaluation point, specified as one of the listed choices.

Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.

Example: `'AcquisitionFunctionName','expected-improvement-per-second'`

`IsObjectiveDeterministic` — Specify deterministic objective function

`false` (default) | `true`

Specify deterministic objective function, specified as `false` or `true`. If `fun` is stochastic (that is, `fun(x)` can return different values for the same `x`), then set `IsObjectiveDeterministic` to `false`. In this case, `bayesopt` estimates a noise level during optimization.

Example: `'IsObjectiveDeterministic',true`

Data Types: `logical`

`ExplorationRatio` — Propensity to explore

`0.5` (default) | positive real

Propensity to explore, specified as a positive real. Applies to the `'expected-improvement-plus'` and `'expected-improvement-per-second-plus'` acquisition functions. See “Plus” on page 10-5.

Example: `'ExplorationRatio',0.2`

Data Types: `double`

`GPActiveSetSize` — Fit Gaussian Process model to `GPActiveSetSize` or fewer points
`300` (default) | positive integer

Fit Gaussian Process model to `GPActiveSetSize` or fewer points, specified as a positive integer. When `bayesopt` has visited more than `GPActiveSetSize` points, subsequent iterations that use a GP model fit the model to `GPActiveSetSize` points. `bayesopt` chooses points uniformly at random without replacement among visited points. Using fewer points leads to faster GP model fitting, at the expense of possibly less accurate fitting.

Example: `'GPActiveSetSize',80`

Data Types: `double`

`UseParallel` — Compute in parallel
`false` (default) | `true`

Compute in parallel, specified as `false` (do not compute in parallel) or `true` (compute in parallel). Computing in parallel requires Parallel Computing Toolbox.

`bayesopt` performs parallel objective function evaluations concurrently on parallel workers. For algorithmic details, see “Parallel Bayesian Optimization” on page 10-7.

Example: `'UseParallel',true`

Data Types: `logical`

`ParallelMethod` — Imputation method for parallel worker objective function values
`'clipped-model-prediction'` (default) | `'model-prediction'` | `'max-observed'` | `'min-observed'`

Imputation method for parallel worker objective function values, specified as `'clipped-model-prediction'`, `'model-prediction'`, `'max-observed'`, or `'min-observed'`. To generate a new point to evaluate, `bayesopt` fits a Gaussian process to all points, including the points being evaluated on workers. To fit the process, `bayesopt` imputes objective function values for the points that are currently on workers. `ParallelMethod` specifies the method used for imputation.

- `'clipped-model-prediction'` — Impute the maximum of these quantities:
  - Mean Gaussian process prediction at the point `x`
  - Minimum observed objective function among feasible points visited
  - Minimum model prediction among all feasible points
- `'model-prediction'` — Impute the mean Gaussian process prediction at the point `x`.
- `'max-observed'` — Impute the maximum observed objective function value among feasible points.
- `'min-observed'` — Impute the minimum observed objective function value among feasible points.

Example: `'ParallelMethod','max-observed'`

`MinWorkerUtilization` — Tolerance on number of active parallel workers
`floor(0.8*Nworkers)` (default) | positive integer

Tolerance on the number of active parallel workers, specified as a positive integer. After `bayesopt` assigns a point to evaluate, and before it computes a new point to assign, it checks whether fewer

than `MinWorkerUtilization` workers are active. If so, `bayesopt` assigns random points within bounds to all available workers. Otherwise, `bayesopt` calculates the best point for one worker. `bayesopt` creates random points much faster than fitted points, so this behavior leads to higher utilization of workers, at the cost of possibly poorer points. For details, see “Parallel Bayesian Optimization” on page 10-7.

Example: `'MinWorkerUtilization',3`

Data Types: `double`

Starting and Stopping

`MaxObjectiveEvaluations` — Objective function evaluation limit

`30` (default) | positive integer

Objective function evaluation limit, specified as a positive integer.

Example: `'MaxObjectiveEvaluations',60`

Data Types: `double`

`MaxTime` — Time limit

`Inf` (default) | positive real

Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`.

Run time can exceed `MaxTime` because `bayesopt` does not interrupt function evaluations.

Example: `'MaxTime',3600`

Data Types: `double`

`NumSeedPoints` — Number of initial evaluation points

`4` (default) | positive integer

Number of initial evaluation points, specified as a positive integer. `bayesopt` chooses these points randomly within the variable bounds, according to the setting of the `Transform` setting for each variable (uniform for `'none'`, logarithmically spaced for `'log'`).

Example: `'NumSeedPoints',10`

Data Types: `double`

Constraints

`XConstraintFcn` — Deterministic constraints on variables

`[]` (default) | function handle

Deterministic constraints on variables, specified as a function handle.

For details, see “Deterministic Constraints — XConstraintFcn” on page 10-38.

Example: `'XConstraintFcn',@xconstraint`

Data Types: `function_handle`

`ConditionalVariableFcn` — Conditional variable constraints

`[]` (default) | function handle

Conditional variable constraints, specified as a function handle.

For details, see “Conditional Constraints — ConditionalVariableFcn” on page 10-39.

Example: `'ConditionalVariableFcn',@condfun`

Data Types: `function_handle`

`NumCoupledConstraints` — Number of coupled constraints
`0` (default) | positive integer

Number of coupled constraints, specified as a positive integer. For details, see “Coupled Constraints” on page 10-40.

Note `NumCoupledConstraints` is required when you have coupled constraints.

Example: `'NumCoupledConstraints',3`

Data Types: `double`

`AreCoupledConstraintsDeterministic` — Indication of whether coupled constraints are deterministic
`true` for all coupled constraints (default) | logical vector

Indication of whether coupled constraints are deterministic, specified as a logical vector of length `NumCoupledConstraints`. For details, see “Coupled Constraints” on page 10-40.

Example: `'AreCoupledConstraintsDeterministic',[true,false,true]`

Data Types: `logical`

Reports, Plots, and Halting

`Verbose` — Command-line display level
`1` (default) | `0` | `2`

Command-line display level, specified as `0`, `1`, or `2`.

- `0` — No command-line display.
- `1` — At each iteration, display the iteration number, result report (see the next paragraph), objective function model, objective function evaluation time, best (lowest) observed objective function value, best (lowest) estimated objective function value, and the observed constraint values (if any). When optimizing in parallel, the display also includes a column showing the number of active workers, counted after assigning a job to the next worker.

  The result report for each iteration is one of the following:

  - `Accept` — The objective function returns a finite value, and all constraints are satisfied.
  - `Best` — Constraints are satisfied, and the objective function returns the lowest value among feasible points.
  - `Error` — The objective function returns a value that is not a finite real scalar.
  - `Infeas` — At least one constraint is violated.
- `2` — Same as `1`, adding diagnostic information such as time to select the next point, model fitting time, indication that "plus" acquisition functions declare overexploiting, and parallel workers are being assigned to random points due to low parallel utilization.

Example: `'Verbose',2`

Data Types: `double`

`OutputFcn` — Function called after each iteration
`{}` (default) | function handle | cell array of function handles

Function called after each iteration, specified as a function handle or cell array of function handles. An output function can halt the solver, and can perform arbitrary calculations, including creating variables or plotting. Specify several output functions using a cell array of function handles.

There are two built-in output functions:

- `@assignInBase` — Constructs a `BayesianOptimization` instance at each iteration and assigns it to a variable in the base workspace. Choose a variable name using the `SaveVariableName` name-value pair.
- `@saveToFile` — Constructs a `BayesianOptimization` instance at each iteration and saves it to a file in the current folder. Choose a file name using the `SaveFileName` name-value pair.

You can write your own output functions. For details, see “Bayesian Optimization Output Functions” on page 10-19.

Example: `'OutputFcn',{@saveToFile @myOutputFunction}`

Data Types: `cell` | `function_handle`

`SaveFileName` — File name for the `@saveToFile` output function
`'BayesoptResults.mat'` (default) | character vector | string scalar

File name for the `@saveToFile` output function, specified as a character vector or string scalar. The file name can include a path, such as `'../optimizations/September2.mat'`.

Example: `'SaveFileName','September2.mat'`

Data Types: `char` | `string`

`SaveVariableName` — Variable name for the `@assignInBase` output function
`'BayesoptResults'` (default) | character vector | string scalar

Variable name for the `@assignInBase` output function, specified as a character vector or string scalar.

Example: `'SaveVariableName','September2Results'`

Data Types: `char` | `string`

`PlotFcn` — Plot function called after each iteration
`{@plotObjectiveModel,@plotMinObjective}` (default) | `'all'` | function handle | cell array of function handles

Plot function called after each iteration, specified as `'all'`, a function handle, or a cell array of function handles. A plot function can halt the solver, and can perform arbitrary calculations, including creating variables, in addition to plotting.

Specify no plot function as `[]`.

`'all'` calls all built-in plot functions. Specify several plot functions using a cell array of function handles.

The built-in plot functions appear in the following tables.

Model Plots — Apply When D ≤ 2	Description
`@plotAcquisitionFunction`	Plot the acquisition function surface.
`@plotConstraintModels`	Plot each constraint model surface. Negative values indicate feasible points. Also plot a *P*(feasible) surface. Also plot the error model, if it exists, which ranges from −1 to 1. Negative values mean that the model probably does not error, positive values mean that it probably does error. The model is: Plotted error = 2*Probability(error) – 1.
`@plotObjectiveEvaluationTimeModel`	Plot the objective function evaluation time model surface.
`@plotObjectiveModel`	Plot the `fun` model surface, the estimated location of the minimum, and the location of the next proposed point to evaluate. For one-dimensional problems, plot envelopes one credible interval above and below the mean function, and envelopes one noise standard deviation above and below the mean.

Trace Plots — Apply to All D	Description
`@plotObjective`	Plot each observed function value versus the number of function evaluations.
`@plotObjectiveEvaluationTime`	Plot each observed function evaluation run time versus the number of function evaluations.
`@plotMinObjective`	Plot the minimum observed and estimated function values versus the number of function evaluations.
`@plotElapsedTime`	Plot three curves: the total elapsed time of the optimization, the total function evaluation time, and the total modeling and point selection time, all versus the number of function evaluations.

You can write your own plot functions. For details, see “Bayesian Optimization Plot Functions” on page 10-11.

Note When there are coupled constraints, iterative display and plot functions can give counterintuitive results such as:

- A *minimum objective* plot can increase.
- The optimization can declare a problem infeasible even when it showed an earlier feasible point.

The reason for this behavior is that the decision about whether a point is feasible can change as the optimization progresses. `bayesopt` determines feasibility with respect to its constraint model, and this model changes as `bayesopt` evaluates points. So a “minimum objective” plot can increase when the minimal point is later deemed infeasible, and the iterative display can show a feasible point that is later deemed infeasible.

Example: `'PlotFcn','all'`

Data Types: `char` | `string` | `cell` | `function_handle`

Initialization

`InitialX` — Initial evaluation points

`NumSeedPoints`-by-`D` random initial points within bounds (default) | `N`-by-`D` table

Initial evaluation points, specified as an `N`-by-`D` table, where `N` is the number of evaluation points, and `D` is the number of variables.

Note If only `InitialX` is provided, it is interpreted as initial points to evaluate. The objective function is evaluated at `InitialX`.

If any other initialization parameters are also provided, `InitialX` is interpreted as prior function evaluation data. The objective function is not evaluated. Any missing values are set to `NaN`.

Data Types: `table`

`InitialObjective` — Objective values corresponding to `InitialX`

`[]` (default) | length-`N` vector

Objective values corresponding to `InitialX`, specified as a length-`N` vector, where `N` is the number of evaluation points.

Example: `'InitialObjective',[17;-3;-12.5]`

Data Types: `double`

`InitialConstraintViolations` — Constraint violations of coupled constraints

`[]` (default) | `N`-by-`K` matrix

Constraint violations of coupled constraints, specified as an `N`-by-`K` matrix, where `N` is the number of evaluation points and `K` is the number of coupled constraints. For details, see “Coupled Constraints” on page 10-40.

Data Types: `double`

`InitialErrorValues` — Errors for `InitialX`

`[]` (default) | length-`N` vector with entries `-1` or `1`

Errors for `InitialX`, specified as a length-`N` vector with entries `-1` or `1`, where `N` is the number of evaluation points. Specify `-1` for no error, and `1` for an error.

Example: `'InitialErrorValues',[-1,-1,-1,-1,1]`

Data Types: `double`

`InitialUserData` — Initial data corresponding to `InitialX`

`[]` (default) | length-`N` cell vector

Initial data corresponding to `InitialX`, specified as a length-`N` cell vector, where `N` is the number of evaluation points.

Example: `'InitialUserData',{2,3,-1}`

Data Types: `cell`

`InitialObjectiveEvaluationTimes` — Evaluation times of objective function at `InitialX`
`[]` (default) | length-`N` vector

Evaluation times of objective function at `InitialX`, specified as a length-`N` vector, where `N` is the number of evaluation points. Time is measured in seconds.

Data Types: `double`

`InitialIterationTimes` — Times for the first `N` iterations
`{}` (default) | length-`N` vector

Times for the first `N` iterations, specified as a length-`N` vector, where `N` is the number of evaluation points. Time is measured in seconds.

Data Types: `double`

Output Arguments

`results` — Bayesian optimization results
`BayesianOptimization` object

Bayesian optimization results, returned as a `BayesianOptimization` object.

More About

Coupled Constraints

Coupled constraints are those constraints whose value comes from the objective function calculation. See “Coupled Constraints” on page 10-40.

Tips

- Bayesian optimization is not reproducible if one of these conditions exists:
  - You specify an acquisition function whose name includes `per-second`, such as `'expected-improvement-per-second'`. The `per-second` modifier indicates that optimization depends on the run time of the objective function. For more details, see “Acquisition Function Types” on page 10-3.
  - You specify to run Bayesian optimization in parallel. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For more details, see “Parallel Bayesian Optimization” on page 10-7.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` name-value argument to `true` in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`BayesianOptimization` | `bestPoint` | `optimizableVariable`

Topics

“Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45
“Bayesian Optimization Algorithm” on page 10-2

Introduced in R2016b

bbdesign

Box-Behnken design

Syntax

```
dBB = bbdesign(n)
[dBB,blocks] = bbdesign(n)
[...] = bbdesign(n,param,val)
```

Description

`dBB = bbdesign(n)` generates a Box-Behnken design for `n` factors. `n` must be an integer `3` or larger. The output matrix `dBB` is *m*-by-`n`, where *m* is the number of runs in the design. Each row represents one run, with settings for all factors represented in the columns. Factor values are normalized so that the cube points take values between `-1` and `1`.

`[dBB,blocks] = bbdesign(n)` requests a blocked design. The output `blocks` is an *m*-by-1 vector of block numbers for each run. Blocks indicate runs that are to be measured under similar conditions to minimize the effect of inter-block differences on the parameter estimates.

`[...] = bbdesign(n,param,val)` specifies one or more optional parameter/value pairs for the design. The following table lists valid parameter/value pairs.

Parameter	Description	Values
`'center'`	Number of center points.	Integer. The default depends on `n`.
`'blocksize'`	Maximum number of points per block.	Integer. The default is `Inf`.

Examples

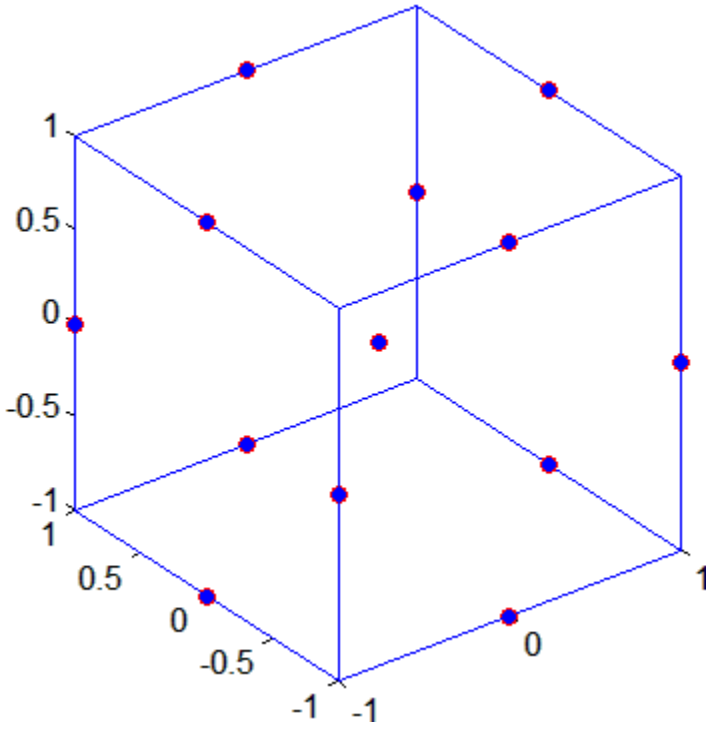
The following creates a 3-factor Box-Behnken design:

```
dBB = bbdesign(3)
dBB =
    -1    -1     0
    -1     1     0
     1    -1     0
     1     1     0
    -1     0    -1
    -1     0     1
     1     0    -1
     1     0     1
     0    -1    -1
     0    -1     1
     0     1    -1
     0     1     1
     0     0     0
     0     0     0
     0     0     0
```

The center point is run 3 times to allow for a more uniform estimate of the prediction variance over the entire design space.

Visualize the design as follows:

```
plot3(dBB(:,1),dBB(:,2),dBB(:,3),'ro',...
         'MarkerFaceColor','b')
X = [1 -1 -1 -1 1 -1 -1 -1 1 1 -1 -1; ...
     1 1 1 -1 1 1 1 -1 1 1 -1 -1];
Y = [-1 -1 1 -1 -1 -1 1 -1 1 -1 1 -1; ...
     1 -1 1 1 1 -1 1 1 1 -1 1 -1];
Z = [1 1 1 1 -1 -1 -1 -1 -1 -1 -1 -1; ...
     1 1 1 1 -1 -1 -1 -1 1 1 1 1];
line(X,Y,Z,'Color','b')
axis square equal
```



See Also

`ccdesign`

Introduced before R2006a

bestPoint

Best point in a Bayesian optimization according to a criterion

Syntax

```
x = bestPoint(results)
x = bestPoint(results,Name,Value)
[x,CriterionValue] = bestPoint( ___ )
[x,CriterionValue,iteration] = bestPoint( ___ )
```

Description

`x = bestPoint(results)` returns the best feasible point in the Bayesian model `results` according to the default criterion `'min-visited-upper-confidence-interval'`.

`x = bestPoint(results,Name,Value)` modifies the best point using name-value pairs.

`[x,CriterionValue] = bestPoint( ___ )`, for any previous syntax, also returns the value of the criterion at `x`.

`[x,CriterionValue,iteration] = bestPoint( ___ )` also returns the iteration number at which the best point was returned. Applies when the `Criterion` name-value pair is `'min-observed'`, `'min-visited-mean'`, or the default `'min-visited-upper-confidence-interval'`.

Examples

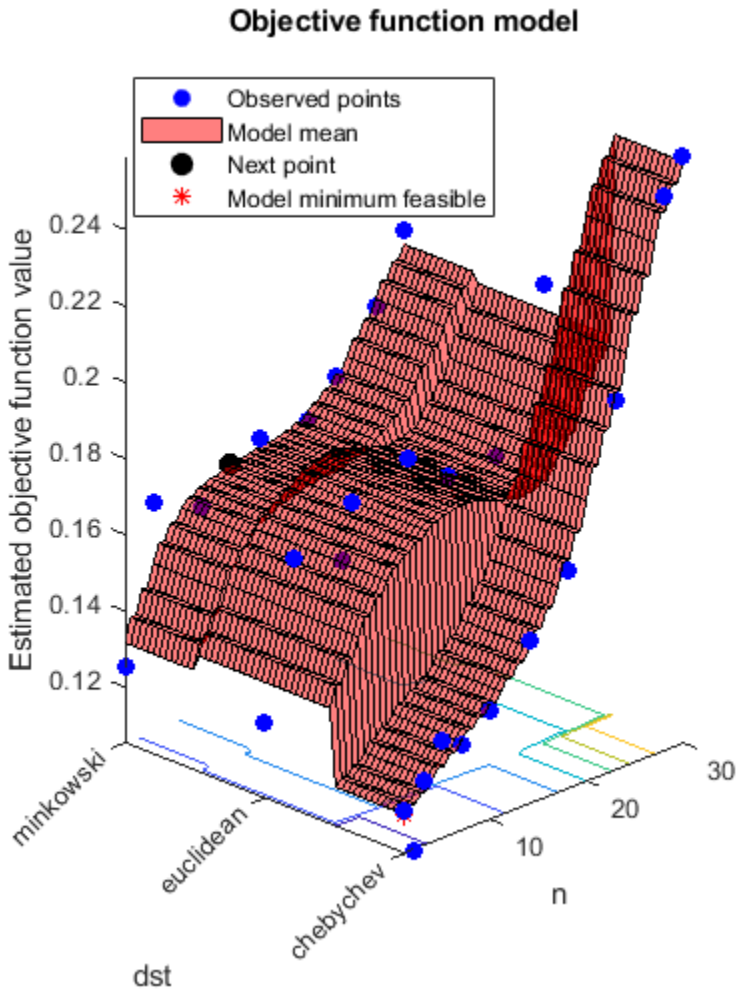
Best Point of an Optimized KNN Classifier

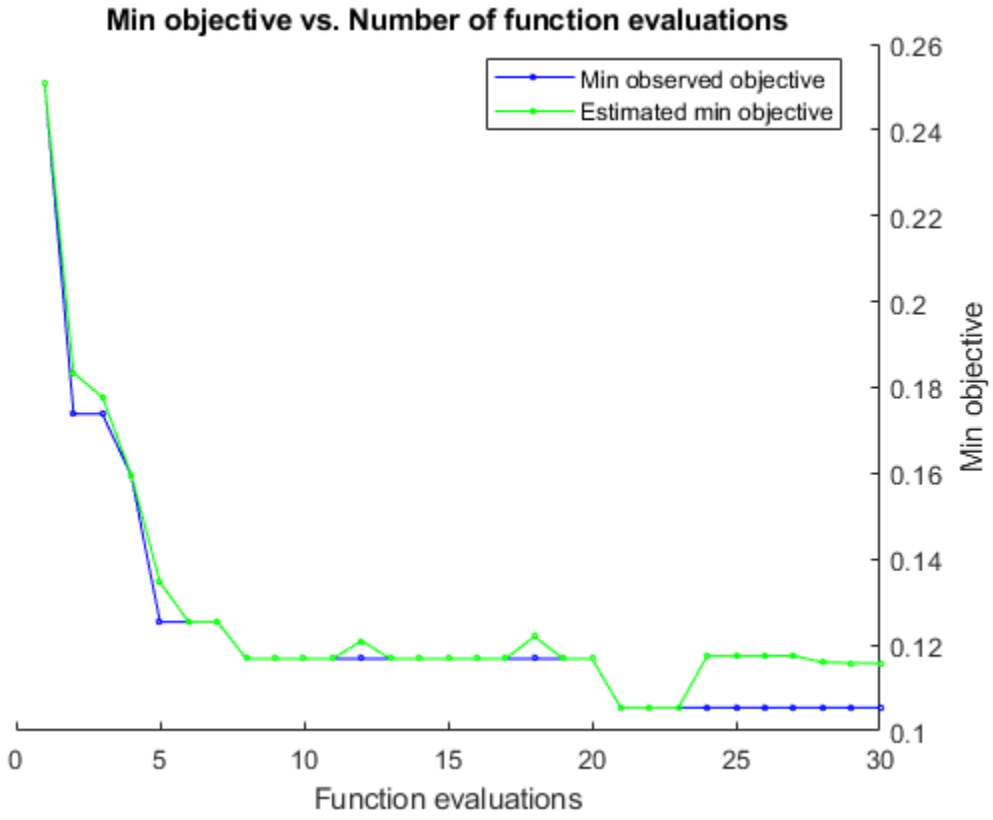
This example shows how to obtain the best point of an optimized classifier.

Optimize a KNN classifier for the `ionosphere` data, meaning find parameters that minimize the cross-validation loss. Minimize over nearest-neighborhood sizes from 1 to 30, and over the distance functions `'chebychev'`, `'euclidean'`, and `'minkowski'`.

For reproducibility, set the random seed, and set the `AcquisitionFunctionName` option to `'expected-improvement-plus'`.

```
load ionosphere
rng(11)
num = optimizableVariable('n',[1,30],'Type','integer');
dst = optimizableVariable('dst',{'chebychev','euclidean','minkowski'},'Type','categorical');
c = cvpartition(351,'Kfold',5);
fun = @(x)kfoldLoss(fitcknn(X,Y,'CVPartition',c,'NumNeighbors',x.n,...
    'Distance',char(x.dst),'NSMethod','exhaustive'));
results = bayesopt(fun,[num,dst],'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus');
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.24
0.22
0.2
0.18
0.16
0.14
0.12
minkowski
euclidean
chebychev
dst
10
20
30
n



Obtain the best point according to the default `'min-visited-upper-confidence-interval'` criterion.

```
x = bestPoint(results)
```

```
x=1×2 table
    n       dst
    _    _________

    1    chebychev
```

The lowest estimated cross-validation loss occurs for one nearest neighbor and `'chebychev'` distance.

Careful examination of the objective function model plot shows a point with two nearest neighbors and `'chebychev'` distance that has a lower objective function value. Find this point using a different criterion.

```
x = bestPoint(results,'Criterion','min-observed')
```

```
x=1×2 table
    n       dst
    _    _________

    2    chebychev
```

Also find the minimum observed objective function value, and the iteration number at which it was observed.

```
[x,CriterionValue,iteration] = bestPoint(results,'Criterion','min-observed')
```

```
x=1×2 table
    n       dst
    _    _________

    2    chebychev


CriterionValue = 0.1054

iteration = 21
```

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `x = bestPoint(results,'Criterion','min-observed')`

`Criterion` — Best point criterion

`'min-visited-upper-confidence-interval'` (default) | `'min-observed'` | `'min-mean'` | `'min-upper-confidence-interval'` | `'min-visited-mean'`

Best point criterion, specified as the comma-separated pair consisting of `'Criterion'` and a criterion name. The names are case-insensitive, do not require `-` characters, and require only enough characters to make the name uniquely distinguishable.

Criterion Name	Meaning
`'min-observed'`	`x` is the feasible point with minimum observed objective.
`'min-mean'`	`x` is the feasible point where the objective model mean is minimized.
`'min-upper-confidence-interval'`	`x` is the feasible point minimizing an upper confidence interval of the objective model. See `alpha`.
`'min-visited-mean'`	`x` is the feasible point where the objective model mean is minimized among the visited points.
`'min-visited-upper-confidence-interval'`	`x` is the feasible point minimizing an upper confidence interval of the objective model among the visited points. See `alpha`.

Example: `'Criterion','min-visited-mean'`

alpha — Probability that modeled objective mean exceeds CriterionValue
0.01 (default) | scalar between 0 and 1

Probability that the modeled objective mean exceeds CriterionValue, specified as the comma-separated pair consisting of 'alpha' and a scalar between 0 and 1. alpha relates to the 'min-upper-confidence-interval' and 'min-visited-upper-confidence-interval' Criterion values. The definition for the upper confidence interval is the value Y where

$P(\text{mean}_Q(\text{fun}(x)) > Y) = \text{alpha}$,

where fun is the objective function, and the mean is calculated with respect to the posterior distribution Q.

Example: 'alpha',0.05

Data Types: double

Output Arguments

x — Best point
1-by-D table

Best point, returned as a 1-by-D table, where D is the number of variables. The meaning of "best" is with respect to Criterion.

CriterionValue — Value of criterion
real scalar

Value of criterion, returned as a real scalar. The value depends on the setting of the Criterion name-value pair, which has a default value of 'min-visited-upper-confidence-interval'.

Criterion Name	Meaning
'min-observed'	Minimum observed objective.
'min-mean'	Minimum of model mean.
'min-upper-confidence-interval'	Value Y satisfying the equation $P(\text{mean}_Q(\text{fun}(x)) > Y) = \text{alpha}$.
'min-visited-mean'	Minimum of observed model mean.
'min-visited-upper-confidence-interval'	Value Y satisfying the equation $P(\text{mean}_Q(\text{fun}(x)) > Y) = \text{alpha}$ among observed points.

iteration — Iteration number at which best point was observed
positive integer

Iteration number at which best point was observed, returned as a positive integer. The best point is defined by CriterionValue.

See Also
BayesianOptimization | bayesopt

Introduced in R2016b

betacdf

Beta cumulative distribution function

Syntax

```
p = betacdf(x,a,b)
p = betacdf(x,a,b,'upper')
```

Description

`p = betacdf(x,a,b)` returns the beta cdf at each of the values in `x` using the corresponding parameters in `a` and `b`. `x`, `a`, and `b` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in `a` and `b` must all be positive, and the values in `x` must lie on the interval [0,1].

`p = betacdf(x,a,b,'upper')` returns the complement of the beta cdf at each of the values in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The beta cdf for a given value `x` and given pair of parameters `a` and `b` is

$$p = F(x|a,b) = \frac{1}{B(a,b)}\int_0^x t^{a-1}(1-t)^{b-1}dt$$

where $B(\cdot)$ is the Beta function.

Examples

Compute Beta Distribution CDF

Compute the cdf for a beta distribution with parameters `a = 2` and `b = 2`.

```
x = 0.1:0.2:0.9;
a = 2;
b = 2;
p = betacdf(x,a,b)
```

```
p = 1×5

    0.0280    0.2160    0.5000    0.7840    0.9720
```

```
a = [1 2 3];
p = betacdf(0.5,a,a)
```

```
p = 1×3

    0.5000    0.5000    0.5000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betafit` | `betainv` | `betalike` | `betapdf` | `betarnd` | `betastat` | `cdf`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betafit

Beta parameter estimates

Syntax

```
phat = betafit(data)
[phat,pci] = betafit(data,alpha)
```

Description

`phat = betafit(data)` computes the maximum likelihood estimates of the beta distribution parameters *a* and *b* from the data in the vector `data` and returns a column vector containing the *a* and *b* estimates, where the beta cdf is given by

$$F(x \mid a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

and $B(\cdot)$ is the Beta function. The elements of `data` must lie in the open interval (0, 1), where the beta distribution is defined. However, it is sometimes also necessary to fit a beta distribution to data that include exact zeros or ones. For such data, the beta likelihood function is unbounded, and standard maximum likelihood estimation is not possible. In that case, `betafit` maximizes a modified likelihood that incorporates the zeros or ones by treating them as if they were values that have been left-censored at `sqrt(realmin)` or right-censored at 1-`eps`/2, respectively.

`[phat,pci] = betafit(data,alpha)` returns confidence intervals on the *a* and *b* parameters in the 2-by-2 matrix `pci`. The first column of the matrix contains the lower and upper confidence bounds for parameter *a*, and the second column contains the confidence bounds for parameter *b*. The optional input argument `alpha` is a value in the range [0, 1] specifying the width of the confidence intervals. By default, `alpha` is `0.05`, which corresponds to 95% confidence intervals. The confidence intervals are based on a normal approximation for the distribution of the logs of the parameter estimates.

Examples

This example generates 100 beta distributed observations. The true *a* and *b* parameters are 4 and 3, respectively. Compare these to the values returned in `p` by the beta fit. Note that the columns of `ci` both bracket the true parameters.

```
data = betarnd(4,3,100,1);
[p,ci] = betafit(data,0.01)
p =
    5.5328    3.8097
ci =
    3.6538    2.6197
    8.3781    5.5402
```

References

[1] Hahn, Gerald J., and S. S. Shapiro. *Statistical Models in Engineering*. Hoboken, NJ: John Wiley & Sons, Inc., 1994, p. 95.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`betacdf` | `betainv` | `betalike` | `betapdf` | `betarnd` | `betastat` | `mle`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betainv

Beta inverse cumulative distribution function

Syntax

```
X = betainv(P,A,B)
```

Description

`X = betainv(P,A,B)` computes the inverse of the beta cdf with parameters specified by `A` and `B` for the corresponding probabilities in `P`. `P`, `A`, and `B` can be vectors, matrices, or multidimensional arrays that are all the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in `A` and `B` must all be positive, and the values in `P` must lie on the interval [0, 1].

The inverse beta cdf for a given probability p and a given pair of parameters a and b is

$$x = F^{-1}(p \mid a, b) = \left\{x : F(x \mid a, b) = p\right\}$$

where

$$p = F(x \mid a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1}(1-t)^{b-1} dt$$

and $B(\cdot)$ is the Beta function. Each element of output `X` is the value whose cumulative probability under the beta cdf defined by the corresponding parameters in `A` and `B` is specified by the corresponding value in `P`.

Examples

```
p = [0.01 0.5 0.99];
x = betainv(p,10,5)
x =
  0.3726  0.6742  0.8981
```

According to this result, for a beta cdf with $a = 10$ and $b = 5$, a value less than or equal to 0.3726 occurs with probability 0.01. Similarly, values less than or equal to 0.6742 and 0.8981 occur with respective probabilities 0.5 and 0.99.

Algorithms

The `betainv` function uses Newton's method with modifications to constrain steps to the allowable range for x, i.e., [0 1].

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betacdf` | `betafit` | `betainv` | `betalike` | `betapdf` | `betarnd` | `betastat` | `icdf`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betalike

Beta negative log-likelihood

Syntax

```
nlogL = betalike(params,data)
[nlogL,AVAR] = betalike(params,data)
```

Description

`nlogL = betalike(params,data)` returns the negative of the beta log-likelihood function for the beta parameters *a* and *b* specified in vector `params` and the observations specified in the column vector `data`.

The elements of `data` must lie in the open interval (0, 1), where the beta distribution is defined. However, it is sometimes also necessary to fit a beta distribution to data that include exact zeros or ones. For such data, the beta likelihood function is unbounded, and standard maximum likelihood estimation is not possible. In that case, `betalike` computes a modified likelihood that incorporates the zeros or ones by treating them as if they were values that have been left-censored at `sqrt(realmin)` or right-censored at 1-`eps`/2, respectively.

`[nlogL,AVAR] = betalike(params,data)` also returns `AVAR`, which is the asymptotic variance-covariance matrix of the parameter estimates if the values in `params` are the maximum likelihood estimates. `AVAR` is the inverse of Fisher's information matrix. The diagonal elements of `AVAR` are the asymptotic variances of their respective parameters.

`betalike` is a utility function for maximum likelihood estimation of the beta distribution. The likelihood assumes that all the elements in the data sample are mutually independent. Since `betalike` returns the negative beta log-likelihood function, minimizing `betalike` using `fminsearch` is the same as maximizing the likelihood.

Examples

This example continues the `betafit` example, which calculates estimates of the beta parameters for some randomly generated beta distributed data.

```
r = betarnd(4,3,100,1);
[nlogl,AVAR] = betalike(betafit(r),r)
nlogl =

  -27.5996


AVAR =

    0.2783    0.1316
    0.1316    0.0867
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betacdf` | `betafit` | `betainv` | `betapdf` | `betarnd` | `betastat`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betapdf

Beta probability density function

Syntax

```
Y = betapdf(X,A,B)
```

Description

`Y = betapdf(X,A,B)` computes the beta pdf at each of the values in `X` using the corresponding parameters in `A` and `B`. `X`, `A`, and `B` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions of the other inputs. The parameters in `A` and `B` must all be positive, and the values in `X` must lie on the interval `[0, 1]`.

The beta probability density function for a given value x and given pair of parameters a and b is

$$y = f(x \mid a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} I_{[0,1]}(x)$$

where $B(\cdot)$ is the Beta function. The uniform distribution on (0 1) is a degenerate case of the beta pdf where $a = 1$ and $b = 1$.

A *likelihood function* is the pdf viewed as a function of the parameters. Maximum likelihood estimators (MLEs) are the values of the parameters that maximize the likelihood function for a fixed value of x.

Examples

```
a = [0.5 1; 2 4]
a =
  0.5000  1.0000
  2.0000  4.0000
y = betapdf(0.5,a,a)
y =
  0.6366  1.0000
  1.5000  2.1875
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betacdf` | `betafit` | `betainv` | `betalike` | `betarnd` | `betastat` | `pdf`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betarnd

Beta random numbers

Syntax

```
R = betarnd(A,B)
R = betarnd(A,B,m,n,...)
R = betarnd(A,B,[m,n,...])
```

Description

`R = betarnd(A,B)` generates random numbers from the beta distribution with parameters specified by `A` and `B`. `A` and `B` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `R`. A scalar input for `A` or `B` is expanded to a constant array with the same dimensions as the other input.

`R = betarnd(A,B,m,n,...)` or `R = betarnd(A,B,[m,n,...])` generates an `m`-by-`n`-by-... array containing random numbers from the beta distribution with parameters `A` and `B`. `A` and `B` can each be scalars or arrays of the same size as `R`.

Examples

```
a = [1 1;2 2];
b = [1 2;1 2];

r = betarnd(a,b)
r =
  0.6987  0.6139
  0.9102  0.8067

r = betarnd(10,10,[1 5])
r =
  0.5974  0.4777  0.5538  0.5465  0.6327

r = betarnd(4,2,2,3)
r =
  0.3943  0.6101  0.5768
  0.5990  0.2760  0.5474
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.

- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betacdf` | `betafit` | `betainv` | `betalike` | `betapdf` | `betastat` | `random`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

betastat

Beta mean and variance

Syntax

```
[M,V] = betastat(A,B)
```

Description

`[M,V] = betastat(A,B)`, with `A>0` and `B>0`, returns the mean of and variance for the beta distribution with parameters specified by `A` and `B`. `A` and `B` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `M` and `V`. A scalar input for `A` or `B` is expanded to a constant array with the same dimensions as the other input.

The mean of the beta distribution with parameters a and b is $a/(a + b)$ and the variance is

$$\frac{ab}{(a+b+1)(a+b)^2}$$

Examples

If parameters a and b are equal, the mean is 1/2.

```
a = 1:6;
[m,v] = betastat(a,a)
m =
  0.5000  0.5000  0.5000  0.5000  0.5000  0.5000
v =
  0.0833  0.0500  0.0357  0.0278  0.0227  0.0192
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`betacdf` | `betafit` | `betainv` | `betalike` | `betapdf` | `betarnd`

Topics

“Beta Distribution” on page B-6

Introduced before R2006a

binocdf

Binomial cumulative distribution function

Syntax

```
y = binocdf(x,n,p)
y = binocdf(x,n,p,'upper')
```

Description

`y = binocdf(x,n,p)` computes a binomial cumulative distribution function at each of the values in `x` using the corresponding number of trials in `n` and the probability of success for each trial in `p`.

`x`, `n`, and `p` can be vectors, matrices, or multidimensional arrays of the same size. Alternatively, one or more arguments can be scalars. The `binocdf` function expands scalar inputs to constant arrays with the same dimensions as the other inputs.

`y = binocdf(x,n,p,'upper')` returns the complement of the binomial cumulative distribution function at each value in `x`, using an algorithm that computes the extreme upper tail probabilities more accurately than the default algorithm.

Examples

Compute and Plot Binomial Cumulative Distribution Function

Compute and plot the binomial cumulative distribution function for the specified range of integer values, number of trials, and probability of success for each trial.

A baseball team plays 100 games in a season and has a 50-50 chance of winning each game. Find the probability of the team winning more than 55 games in a season.

```
format long
1 - binocdf(55,100,0.5)
```

```
ans =
   0.135626512036917
```

Find the probability of the team winning between 50 and 55 games in a season.

```
binocdf(55,100,0.5) - binocdf(49,100,0.5)
```

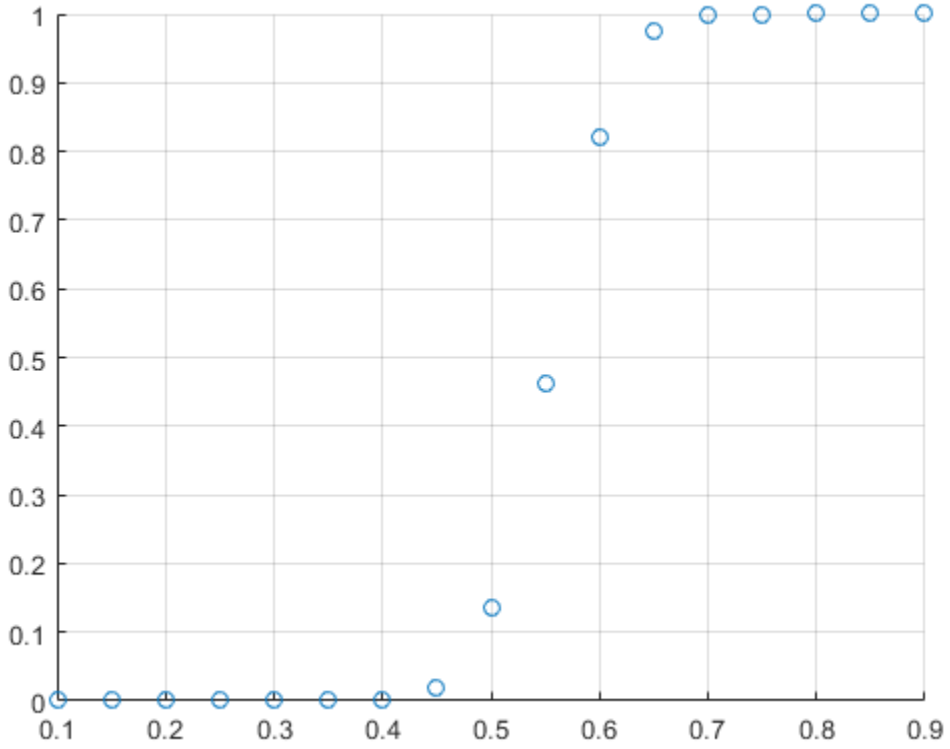
```
ans =
   0.404168106656672
```

Compute the probabilities of the team winning more than 55 games in a season if the chance of winning each game ranges from 10% to 90%.

```
chance = 0.1:0.05:0.9;
y = 1 - binocdf(55,100,chance);
```

Plot the results.

```
scatter(chance,y)
grid on
```



Compute Extreme Upper Tail Probabilities

Compute the complement of the binomial cumulative distribution function with more accurate upper tail probabilities.

A baseball team plays 100 games in a season and has a 50-50 chance of winning each game. Find the probability of the team winning more than 95 games in a season.

```
format long
1 - binocdf(95,100,0.5)
```

```
ans =
     0
```

This result shows that the probability is so close to 1 (within `eps`) that subtracting it from 1 gives 0. To approximate the extreme upper tail probabilities better, compute the complement of the binomial cumulative distribution function directly instead of computing the difference.

```
binocdf(95,100,0.5,'upper')
```

```
ans =
     3.224844447881779e-24
```

Alternatively, use the `binopdf` function to find the probabilities of the team winning 96, 97, 98, 99, and 100 games in a season. Find the sum of these probabilities by using the `sum` function.

```
sum(binopdf(96:100,100,0.5),'all')
```

```
ans =
     3.224844447881779e-24
```

Input Arguments

x — Values at which to evaluate binomial cdf

integer from interval `[0 n]` | array of integers from interval `[0 n]`

Values at which to evaluate the binomial cdf, specified as an integer or an array of integers. All values of `x` must belong to the interval `[0 n]`, where `n` is the number of trials.

Example: `[0 1 3 4]`

Data Types: `single` | `double`

n — Number of trials

positive integer | array of positive integers

Number of trials, specified as a positive integer or an array of positive integers.

Example: `[10 20 50 100]`

Data Types: `single` | `double`

p — Probability of success for each trial

scalar value from interval `[0 1]` | array of scalar values from interval `[0 1]`

Probability of success for each trial, specified as a scalar value or an array of scalar values. All values of `p` must belong to the interval `[0 1]`.

Example: `[0.01 0.1 0.5 0.7]`

Data Types: `single` | `double`

Output Arguments

y — Binomial cdf values

scalar value | array of scalar values

Binomial cdf values, returned as a scalar value or an array of scalar values. Each element in `y` is the binomial cdf value of the distribution evaluated at the corresponding element in `x`.

Data Types: `single` | `double`

More About

Binomial Cumulative Distribution Function

The binomial cumulative distribution function lets you obtain the probability of observing less than or equal to x successes in n trials, with the probability p of success on a single trial.

The binomial cumulative distribution function for a given value x and a given pair of parameters n and p is

$$y = F(x \mid n, p) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{(n-i)} I_{(0,1,\ldots,n)}(i).$$

The resulting value y is the probability of observing up to x successes in n independent trials, where the probability of success in any given trial is p. The indicator function $I_{(0,1,\ldots,n)}(i)$ ensures that x only adopts values of $0,1,\ldots,n$.

Alternative Functionality

- `binocdf` is a function specific to binomial distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, specify the probability distribution name and its parameters. Alternatively, create a `BinomialDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `binocdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`BinomialDistribution` | `binofit` | `binoinv` | `binopdf` | `binornd` | `binostat` | `cdf`

Topics

“Binomial Distribution” on page B-10

Introduced before R2006a

binofit

Binomial parameter estimates

Syntax

```
phat = binofit(x,n)
[phat,pci] = binofit(x,n)
[phat,pci] = binofit(x,n,alpha)
```

Description

`phat = binofit(x,n)` returns a maximum likelihood estimate of the probability of success in a given binomial trial based on the number of successes, `x`, observed in `n` independent trials. If `x = (x(1), x(2), ... x(k))` is a vector, `binofit` returns a vector of the same size as `x` whose ith entry is the parameter estimate for `x(i)`. All `k` estimates are independent of each other. If `n = (n(1), n(2), ..., n(k))` is a vector of the same size as `x`, the binomial fit, `binofit`, returns a vector whose ith entry is the parameter estimate based on the number of successes `x(i)` in `n(i)` independent trials. A scalar value for `x` or `n` is expanded to the same size as the other input.

`[phat,pci] = binofit(x,n)` returns the probability estimate, `phat`, and the 95% confidence intervals, `pci`. `binofit` uses the Clopper-Pearson method to calculate confidence intervals.

`[phat,pci] = binofit(x,n,alpha)` returns the `100(1 - alpha)`% confidence intervals. For example, `alpha = 0.01` yields 99% confidence intervals.

Note `binofit` behaves differently than other Statistics and Machine Learning Toolbox functions that compute parameter estimates, in that it returns independent estimates for each entry of `x`. By comparison, `expfit` returns a single parameter estimate based on all the entries of `x`.

Unlike most other distribution fitting functions, the `binofit` function treats its input `x` vector as a collection of measurements from separate samples. If you want to treat `x` as a single sample and compute a single parameter estimate for it, you can use `binofit(sum(x),sum(n))` when `n` is a vector, and `binofit(sum(X),N*length(X))` when `n` is a scalar.

Examples

This example generates a binomial sample of 100 elements, where the probability of success in a given trial is 0.6, and then estimates this probability from the outcomes in the sample.

```
r = binornd(100,0.6);
[phat,pci] = binofit(r,100)
phat =
  0.5800
pci =
  0.4771  0.6780
```

The 95% confidence interval, `pci`, contains the true value, 0.6.

References

[1] Johnson, N. L., S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. Hoboken, NJ: Wiley-Interscience, 1993.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`binocdf` | `binoinv` | `binopdf` | `binornd` | `binostat` | `mle`

Topics

“Binomial Distribution” on page B-10

Introduced before R2006a

binoinv

Binomial inverse cumulative distribution function

Syntax

```
X = binoinv(Y,N,P)
```

Description

`X = binoinv(Y,N,P)` returns the smallest integer `X` such that the binomial cdf evaluated at `X` is equal to or exceeds `Y`. You can think of `Y` as the probability of observing `X` successes in `N` independent trials where `P` is the probability of success in each trial. Each `X` is a positive integer less than or equal to `N`.

`Y`, `N`, and `P` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in `N` must be positive integers, and the values in both `P` and `Y` must lie on the interval [0 1].

Examples

If a baseball team has a 50-50 chance of winning any game, what is a reasonable range of games this team might win over a season of 162 games?

```
binoinv([0.05 0.95],162,0.5)
ans =
    71    91
```

This result means that in 90% of baseball seasons, a .500 team should win between 71 and 91 games.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`binocdf` | `binofit` | `binopdf` | `binornd` | `binostat` | `icdf`

Topics

“Binomial Distribution” on page B-10

Introduced before R2006a

binopdf

Binomial probability density function

Syntax

```
y = binopdf(x,n,p)
```

Description

`y = binopdf(x,n,p)` computes the binomial probability density function at each of the values in `x` using the corresponding number of trials in `n` and probability of success for each trial in `p`.

`x`, `n`, and `p` can be vectors, matrices, or multidimensional arrays of the same size. Alternatively, one or more arguments can be scalars. The `binopdf` function expands scalar inputs to constant arrays with the same dimensions as the other inputs.

Examples

Compute and Plot Binomial Probability Density Function

Compute and plot the binomial probability density function for the specified range of integer values, number of trials, and probability of success for each trial.

In one day, a quality assurance inspector tests 200 circuit boards. 2% of the boards have defects. Compute the probability that the inspector will find no defective boards on any given day.

```
binopdf(0,200,0.02)
```

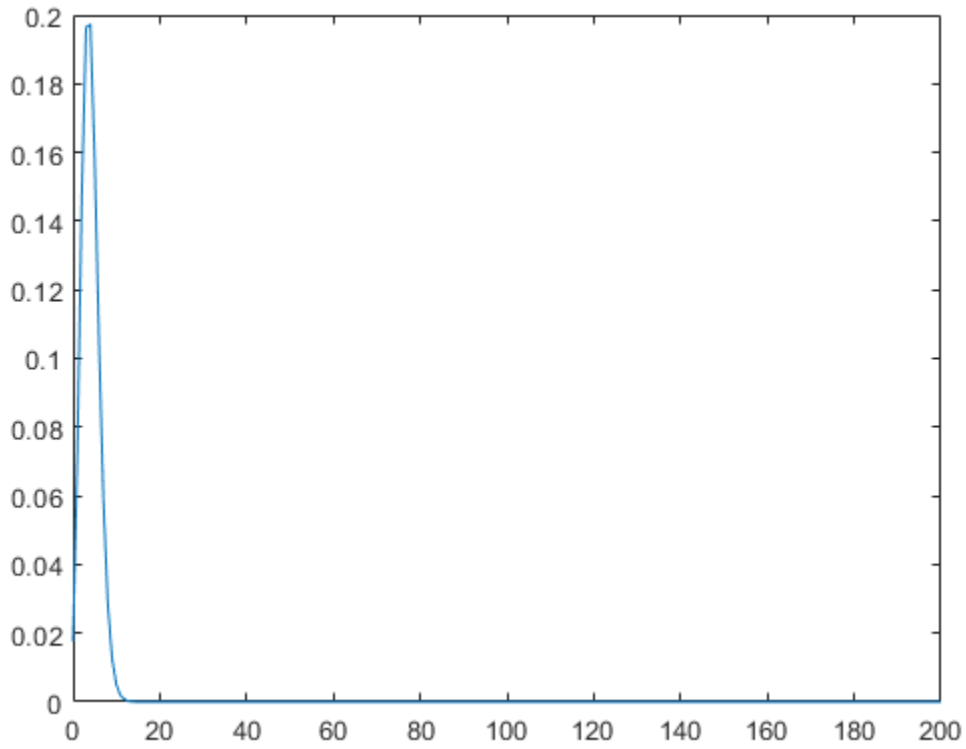
```
ans = 0.0176
```

Compute the binomial probability density function values at each value from 0 to 200. These values correspond to the probabilities that the inspector will find 0, 1, 2, ..., 200 defective boards on any given day.

```
defects = 0:200;
y = binopdf(defects,200,.02);
```

Plot the resulting binomial probability values.

```
plot(defects,y)
```



Compute the most likely number of defective boards that the inspector finds in a day.

```
[x,i] = max(y);
defects(i)
```

```
ans = 4
```

Input Arguments

x — Values at which to evaluate binomial pdf

integer from interval `[0 n]` | array of integers from interval `[0 n]`

Values at which to evaluate the binomial pdf, specified as an integer or an array of integers. All values of `x` must belong to the interval `[0 n]`, where `n` is the number of trials.

Example: `[0,1,3,4]`

Data Types: `single` | `double`

n — Number of trials

positive integer | array of positive integers

Number of trials, specified as a positive integer or an array of positive integers.

Example: `[10,20,50,100]`

Data Types: `single` | `double`

p — Probability of success for each trial
scalar value from interval `[0 1]` | array of scalar values from interval `[0 1]`

Probability of success for each trial, specified as a scalar value or an array of scalar values. All values of `p` must belong to the interval `[0 1]`.

Example: `[0.01,0.1,0.5,0.7]`

Data Types: `single` | `double`

Output Arguments

y — Binomial pdf values
scalar value | array of scalar values

Binomial pdf values, returned as a scalar value or array of scalar values. Each element in `y` is the binomial pdf value of the distribution evaluated at the corresponding element in `x`.

Data Types: `single` | `double`

More About

Binomial Probability Density Function

The binomial probability density function lets you obtain the probability of observing exactly x successes in n trials, with the probability p of success on a single trial.

The binomial probability density function for a given value x and given pair of parameters n and p is

$$y = f(x \mid n, p) = \binom{n}{x} p^x q^{(n-x)} I_{(0,1,\ldots,n)}(x)$$

where $q = 1 - p$. The resulting value y is the probability of observing exactly x successes in n independent trials, where the probability of success in any given trial is p. The indicator function $I_{(0,1,\ldots,n)}(x)$ ensures that x only adopts values of 0, 1, ..., n.

Alternative Functionality

- `binopdf` is a function specific to binomial distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, specify the probability distribution name and its parameters. Alternatively, create a `BinomialDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `binopdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`BinomialDistribution` | `binocdf` | `binofit` | `binoinv` | `binornd` | `binostat` | `pdf`

Topics

“Binomial Distribution” on page B-10

Introduced before R2006a

binornd

Random numbers from binomial distribution

Syntax

```
r = binornd(n,p)
r = binornd(n,p,sz1,...,szN)
r = binornd(n,p,sz)
```

Description

`r = binornd(n,p)` generates random numbers from the binomial distribution specified by the number of trials `n` and the probability of success for each trial `p`.

`n` and `p` can be vectors, matrices, or multidimensional arrays of the same size. Alternatively, one or more arguments can be scalars. The `binornd` function expands scalar inputs to constant arrays with the same dimensions as the other inputs. The function returns a vector, matrix, or multidimensional array `r` of the same size as `n` and `p`.

`r = binornd(n,p,sz1,...,szN)` generates an array of random numbers from the binomial distribution with the scalar parameters `n` and `p`, where `sz1,...,szN` indicates the size of each dimension.

`r = binornd(n,p,sz)` generates an array of random numbers from the binomial distribution with the scalar parameters `n` and `p`, where vector `sz` specifies `size(r)`.

Examples

Array of Random Numbers from Several Binomial Distributions

Generate an array of random numbers from the binomial distributions. For each distribution, you specify the number of trials and the probability of success for each trial.

Specify the numbers of trials.

```
n = 10:10:60
```

```
n = 1×6

    10    20    30    40    50    60
```

Specify the probabilities of success for each trial.

```
p = 1./n
```

```
p = 1×6

    0.1000    0.0500    0.0333    0.0250    0.0200    0.0167
```

Generate random numbers from the binomial distributions.

```
r = binornd(n,p)
```

```
r = 1×6

     0     1     1     0     1     1
```

Array of Random Numbers from One Binomial Distribution

Generate an array of random numbers from one binomial distribution. Here, the distribution parameters `n` and `p` are scalars.

Use the `binornd` function to generate random numbers from the binomial distribution with 100 trials, where the probability of success in each trial is 0.2. The function returns one number.

```
r_scalar = binornd(100,0.2)
```

```
r_scalar = 20
```

Generate a 2-by-3 array of random numbers from the same distribution by specifying the required array dimensions.

```
r_array = binornd(100,0.2,2,3)
```

```
r_array = 2×3

    18    23    20
    18    24    23
```

Alternatively, specify the required array dimensions as a vector.

```
r_array = binornd(100,0.2,[2 3])
```

```
r_array = 2×3

    21    21    20
    26    18    23
```

Input Arguments

n — Number of trials

positive integer | array of positive integers

Number of trials, specified as a positive integer or an array of positive integers.

Example: `[10 20 50 100]`

Data Types: `single` | `double`

p — Probability of success for each trial

scalar value | array of scalar values

Probability of success for each trial, specified as a scalar value or an array of scalar values. All values of `p` must belong to the interval `[0 1]`.

Example: `[0.01 0.1 0.5 0.7]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as separate arguments of integers. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the binomial probability distribution.

If either `n` or `p` is an array, then the specified dimensions `sz1,...,szN` must match the common dimensions of `n` and `p` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `binornd` ignores trailing dimensions with a size of 1. For example, `binornd(n,p,3,1,1,1)` produces a 3-by-1 vector of random numbers.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5 3 2]` generates a 5-by-3-by-2 array of random numbers from the binomial probability distribution.

If either `n` or `p` is an array, then the specified dimensions `sz` must match the common dimensions of `n` and `p` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value `[sz1]`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `binornd` ignores trailing dimensions with a size of 1. For example, `binornd(n,p,[3 1 1 1])` produces a 3-by-1 vector of random numbers.

Example: `[5 3 2]`

Data Types: `single` | `double`

Output Arguments

`r` — Random numbers from binomial distribution
scalar value | array of scalar values

Random numbers from the binomial distribution, returned as a scalar value or an array of scalar values.

Data Types: `single` | `double`

Alternative Functionality

- `binornd` is a function specific to binomial distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Alternatively, create a `BinomialDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `binornd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB in these two cases:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

Distributed Arrays

Partition large arrays across the combined memory of your cluster using Parallel Computing Toolbox™.

This function fully supports distributed arrays. For more information, see “Run MATLAB Functions with Distributed Arrays” (Parallel Computing Toolbox).

See Also

`BinomialDistribution` | `binocdf` | `binofit` | `binoinv` | `binopdf` | `binostat` | `random`

Topics

“Binomial Distribution” on page B-10

Introduced before R2006a

binostat

Binomial mean and variance

Syntax

```
[M,V] = binostat(N,P)
```

Description

`[M,V] = binostat(N,P)` returns the mean of and variance for the binomial distribution with parameters specified by the number of trials, `N`, and probability of success for each trial, `P`. `N` and `P` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `M` and `V`. A scalar input for `N` or `P` is expanded to a constant array with the same dimensions as the other input.

The mean of the binomial distribution with parameters n and p is np. The variance is npq, where $q = 1 – p$.

Examples

```
n = logspace(1,5,5)
n =
     10     100    1000    10000   100000

[m,v] = binostat(n,1./n)
m =
   1   1   1   1   1
v =
  0.9000  0.9900  0.9990  0.9999  1.0000

[m,v] = binostat(n,1/2)
m =
      5     50     500    5000    50000
v =
  1.0e+04 *
  0.0003  0.0025  0.0250  0.2500  2.5000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`binocdf` | `binofit` | `binoinv` | `binopdf` | `binornd`

Topics
“Binomial Distribution” on page B-10

Introduced before R2006a

binScatterPlot

Scatter plot of bins for tall arrays

Syntax

```
binScatterPlot(X,Y)
binScatterPlot(X,Y,nbins)
binScatterPlot(X,Y,Xedges,Yedges)
binScatterPlot(X,Y,Name,Value)
h = binScatterPlot( ___ )
```

Description

`binScatterPlot(X,Y)` creates a binned scatter plot of the data in `X` and `Y`. The `binScatterPlot` function uses an automatic binning algorithm that returns bins with a uniform area, chosen to cover the range of elements in `X` and `Y` and reveal the underlying shape of the distribution.

`binScatterPlot(X,Y,nbins)` specifies the number of bins to use in each dimension.

`binScatterPlot(X,Y,Xedges,Yedges)` specifies the edges of the bins in each dimension using the vectors `Xedges` and `Yedges`.

`binScatterPlot(X,Y,Name,Value)` specifies additional options with one or more name-value pair arguments using any of the previous syntaxes. For example, you can specify `'Color'` and a valid color option to change the color theme of the plot, or `'Gamma'` with a positive scalar to adjust the level of detail.

`h = binScatterPlot( ___ )` returns a `Histogram2` object. Use this object to inspect properties of the plot.

Examples

Binned Scatter Plot of Normally Distributed Random Data

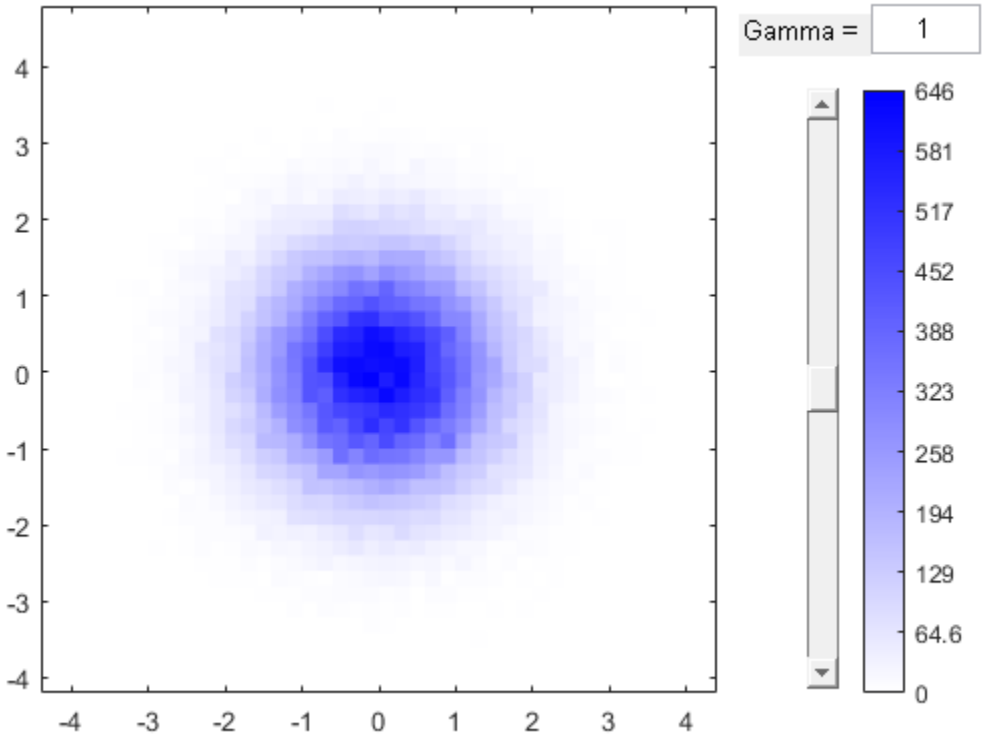
Create two tall vectors of random data. Create a binned scatter plot for the data.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)

X = tall(randn(1e5,1));
Y = tall(randn(1e5,1));
binScatterPlot(X,Y)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.53 sec
Evaluation completed in 1 sec
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.18 sec
Evaluation completed in 0.24 sec
```



The resulting figure contains a slider to adjust the level of detail in the image.

Specify Number of Scatter Plot Bins

Specify a scalar value as the third input argument to use the same number of bins in each dimension, or a two-element vector to use a different number of bins in each dimension.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

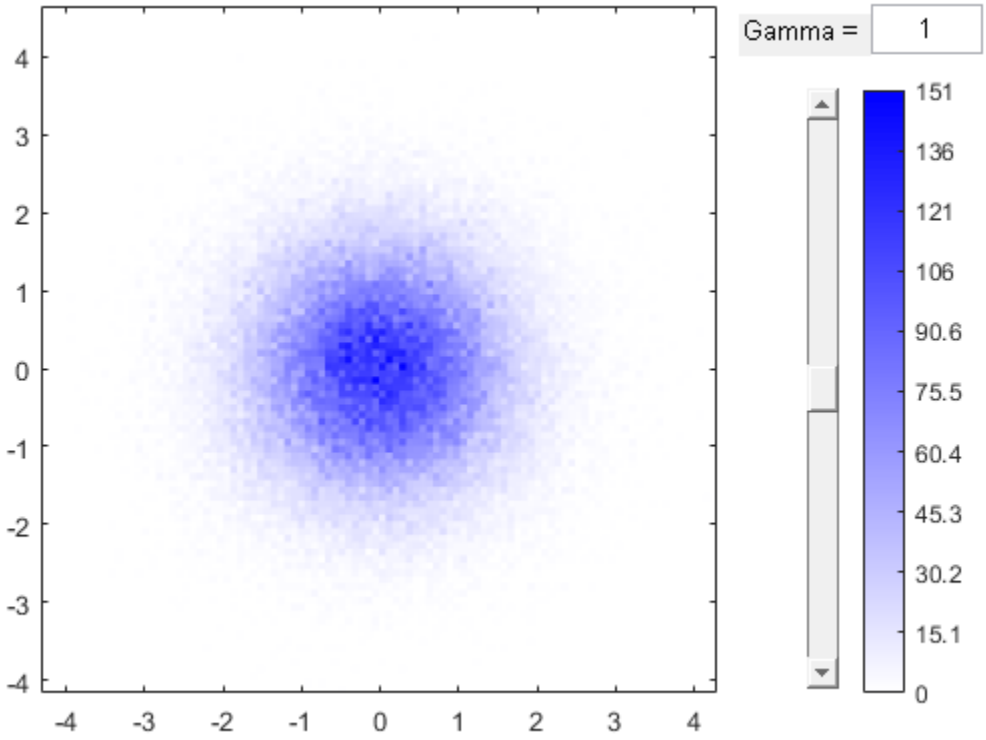
```
mapreducer(0)
```

Plot a binned scatter plot of random data sorted into 100 bins in each dimension.

```
X = tall(randn(1e5,1));
Y = tall(randn(1e5,1));
binScatterPlot(X,Y,100)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.35 sec
```

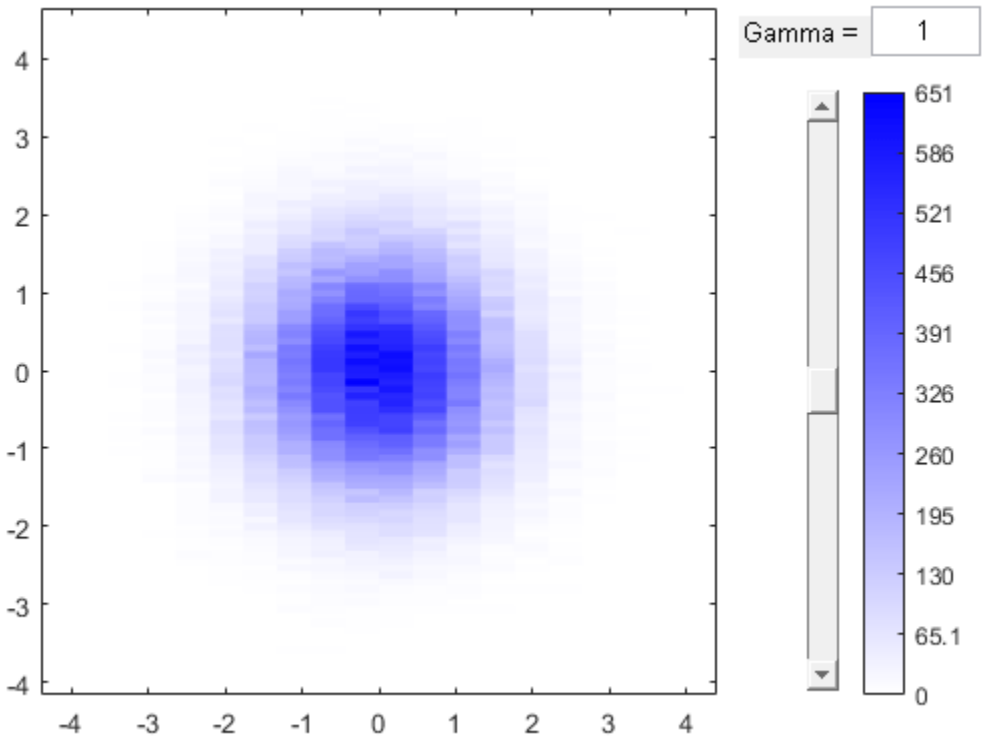
```
Evaluation completed in 0.55 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.17 sec
Evaluation completed in 0.26 sec
```



Use 20 bins in the *x*-dimension and continue to use 100 bins in the *y*-dimension.

```
binScatterPlot(X,Y,[20 100])
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.18 sec
Evaluation completed in 0.27 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.17 sec
Evaluation completed in 0.24 sec
```



Specify Bin Edges for Scatter Plot

Plot a binned scatter plot of random data with specific bin edges. Use bin edges of `Inf` and `-Inf` to capture outliers.

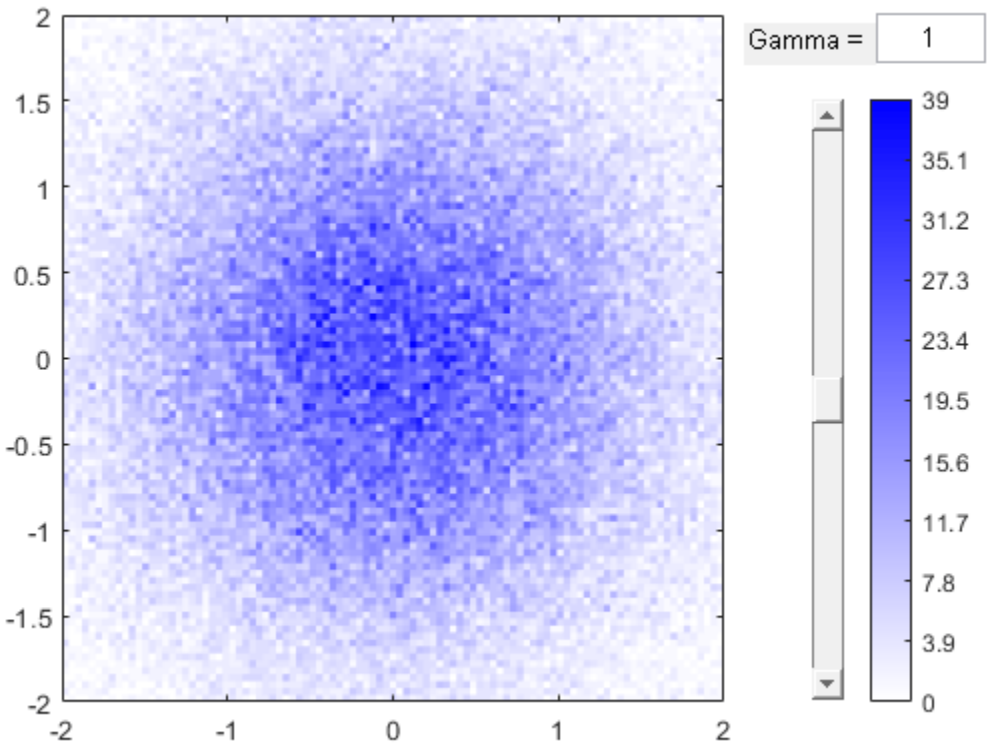
When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a binned scatter plot with 100 bin edges between `[-2 2]` in each dimension. The data outside the specified bin edges is not included in the plot.

```
X = tall(randn(1e5,1));
Y = tall(randn(1e5,1));
Xedges = linspace(-2,2);
Yedges = linspace(-2,2);
binScatterPlot(X,Y,Xedges,Yedges)
```

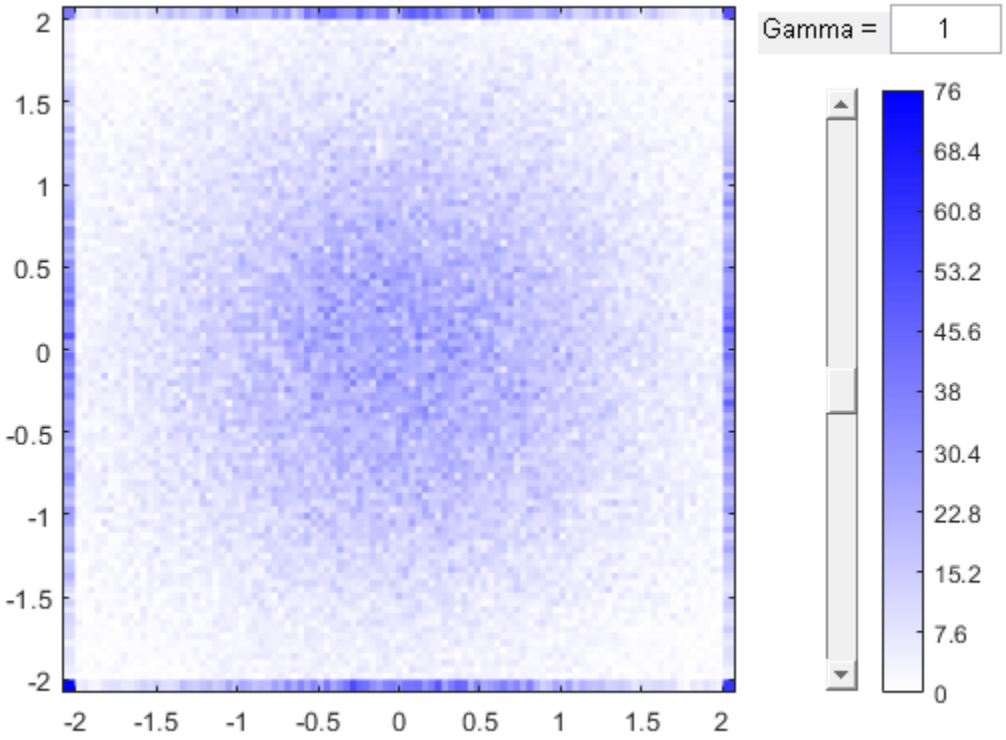
```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.2 sec
Evaluation completed in 1.6 sec
```



Use coarse bins extending to infinity on the edges of the plot to capture outliers.

```
Xedges = [-Inf linspace(-2,2) Inf];
Yedges = [-Inf linspace(-2,2) Inf];
binScatterPlot(X,Y,Xedges,Yedges)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.34 sec
Evaluation completed in 0.51 sec
```



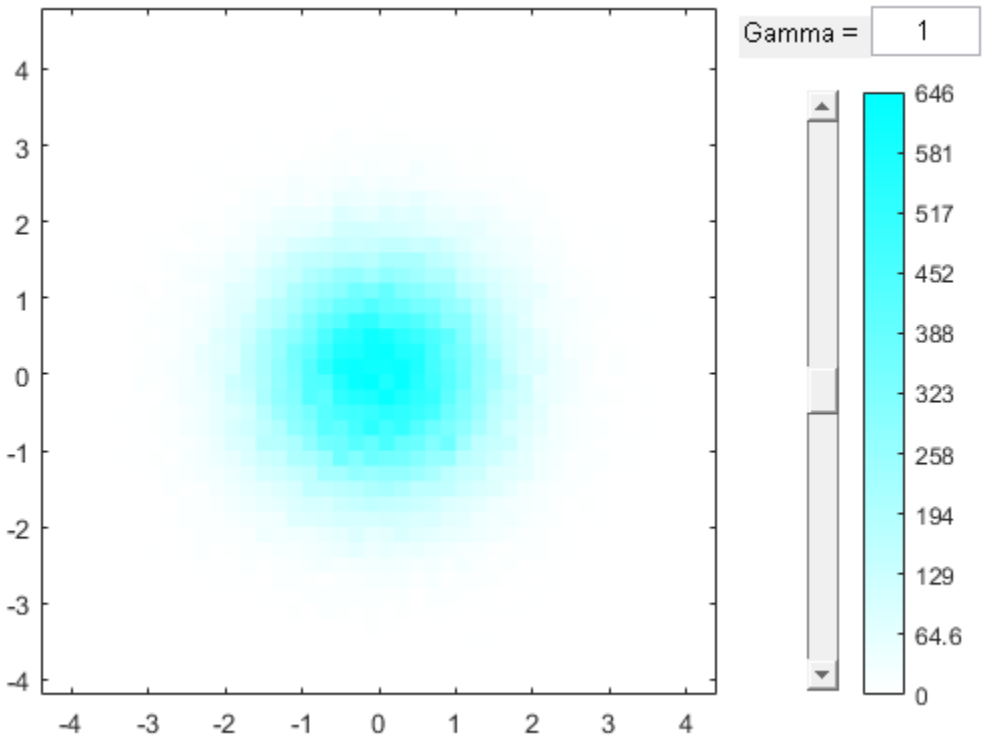
Adjust Plot Color Theme

Plot a binned scatter plot of random data, specifying `'Color'` as `'c'`.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)

X = tall(randn(1e5,1));
Y = tall(randn(1e5,1));
binScatterPlot(X,Y,'Color','c')
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.6 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.34 sec
Evaluation completed in 0.46 sec
```



Input Arguments

`X,Y` — Data to distribute among bins (as separate arguments)

tall vectors | tall matrices | tall multidimensional arrays

Data to distribute among bins, specified as separate arguments of tall vectors, matrices, or multidimensional arrays. `X` and `Y` must be the same size. If `X` and `Y` are not vectors, then `binScatterPlot` treats them as single column vectors, `X(:)` and `Y(:)`.

Corresponding elements in `X` and `Y` specify the *x* and *y* coordinates of 2-D data points, `[X(k),Y(k)]`. The underlying data types of `X` and `Y` can be different, but `binScatterPlot` concatenates these inputs into a single `N`-by-2 tall matrix of the dominant underlying data type.

`binScatterPlot` ignores all `NaN` values. Similarly, `binScatterPlot` ignores `Inf` and `-Inf` values, unless the bin edges explicitly specify `Inf` or `-Inf` as a bin edge.

Note If `X` or `Y` contain integers of type `int64` or `uint64` that are larger than `flintmax`, then it is recommended that you explicitly specify the bin edges.`binScatterPlot` automatically bins the input data using double precision, which lacks integer precision for numbers greater than `flintmax`.

Data Types: `single` | `double` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `logical`

nbins — Number of bins in each dimension
scalar | vector

Number of bins in each dimension, specified as a positive scalar integer or two-element vector of positive integers. If you do not specify `nbins`, then `binScatterPlot` automatically calculates how many bins to use based on the values in `X` and `Y`.

- If `nbins` is a scalar, then `binScatterPlot` uses that many bins in each dimension.
- If `nbins` is a vector, then `nbins(1)` specifies the number of bins in the *x*-dimension and `nbins(2)` specifies the number of bins in the *y*-dimension.

Example: `binScatterPlot(X,Y,20)` uses 20 bins in each dimension.

Example: `binScatterPlot(X,Y,[10 20])` uses 10 bins in the x-dimension and 20 bins in the y-dimension.

Xedges — Bin edges in *x*-dimension
vector

Bin edges in *x*-dimension, specified as a vector. `Xedges(1)` is the first edge of the first bin in the *x*-dimension, and `Xedges(end)` is the outer edge of the last bin.

The value `[X(k),Y(k)]` is in the `(i,j)`th bin if `Xedges(i)` ≤ `X(k)` < `Xedges(i+1)` **and** `Yedges(j)` ≤ `Y(k)` < `Yedges(j+1)`. The last bins in each dimension also include the last (outer) edge. For example, `[X(k),Y(k)]` falls into the `i`th bin in the last row if `Xedges(end-1)` ≤ `X(k)` ≤ `Xedges(end)` **and** `Yedges(i)` ≤ `Y(k)` < `Yedges(i+1)`.

Data Types: `single` | `double` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `logical`

Yedges — Bin edges in *y*-dimension
vector

Bin edges in *y*-dimension, specified as a vector. `Yedges(1)` is the first edge of the first bin in the *y*-dimension, and `Yedges(end)` is the outer edge of the last bin.

The value `[X(k),Y(k)]` is in the `(i,j)`th bin if `Xedges(i)` ≤ `X(k)` < `Xedges(i+1)` **and** `Yedges(j)` ≤ `Y(k)` < `Yedges(j+1)`. The last bins in each dimension also include the last (outer) edge. For example, `[X(k),Y(k)]` falls into the `i`th bin in the last row if `Xedges(end-1)` ≤ `X(k)` ≤ `Xedges(end)` **and** `Yedges(i)` ≤ `Y(k)` < `Yedges(i+1)`.

Data Types: `single` | `double` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `logical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `binScatterPlot(X,Y,'BinWidth',[5 10])`

BinMethod — Binning algorithm
`'auto'` (default) | `'scott'` | `'integers'`

Binning algorithm, specified as the comma-separated pair consisting of `'BinMethod'` and one of these values.

Value	Description
`'auto'`	The default `'auto'` algorithm uses a maximum of 100 bins and chooses a bin width to cover the data range and reveal the shape of the underlying distribution.
`'scott'`	Scott's rule is optimal if the data is close to being jointly normally distributed. This rule is appropriate for most other distributions, as well. It uses a bin size of `[3.5*std(X)*numel(X)^(-1/4), 3.5*std(Y)*numel(Y)^(-1/4)]`.
`'integers'`	The integer rule is useful with integer data, as it creates a bin for each integer. It uses a bin width of 1 and places bin edges halfway between integers. To avoid accidentally creating too many bins, you can use this rule to create a limit of 65536 bins (2^{16}). If the data range is greater than 65536, then the integer rule uses wider bins instead.

Note The `BinMethod` property of the resulting `Histogram2` object always has a value of `'manual'`.

`BinWidth` — Width of bins in each dimension

scalar | vector

Width of bins in each dimension, specified as the comma-separated pair consisting of `'BinWidth'` and a scalar or two-element vector of positive integers, `[xWidth yWidth]`. A scalar value indicates the same bin width for each dimension.

If you specify `BinWidth`, then `binScatterPlot` can use a maximum of 1024 bins (2^{10}) along each dimension. If instead the specified bin width requires more bins, then `binScatterPlot` uses a larger bin width corresponding to the maximum number of bins.

Example: `binScatterPlot(X,Y,'BinWidth',[5 10])` uses bins with size `5` in the `x`-dimension and size `10` in the `y`-dimension.

`Color` — Plot color theme

`'b'` (default) | `'y'` | `'m'` | `'c'` | `'r'` | `'g'` | `'k'`

Plot color theme, specified as the comma-separated pair consisting of `'Color'` and one of these options.

Option	Description
`'b'`	Blue
`'m'`	Magenta
`'c'`	Cyan
`'r'`	Red
`'g'`	Green

Option	Description
`'y'`	Yellow
`'k'`	Black

`Gamma` — Gamma correction
1 (default) | positive scalar

Gamma correction, specified as the comma-separated pair consisting of `'Gamma'` and a positive scalar. Use this option to adjust the brightness and color intensity to affect the amount of detail in the image.

- `gamma < 1` — As gamma decreases, the shading of bins with smaller bin counts becomes progressively darker, including more detail in the image.
- `gamma > 1` — As gamma increases, the shading of bins with smaller bin counts becomes progressively lighter, removing detail from the image.
- The default value of 1 does not apply any correction to the display.

`XBinLimits` — Bin limits in *x*-dimension
vector

Bin limits in *x*-dimension, specified as the comma-separated pair consisting of `'XBinLimits'` and a two-element vector, `[xbmin,xbmax]`. The vector indicates the first and last bin edges in the *x*-dimension.

`binScatterPlot` only plots data that falls within the bin limits inclusively, `Data(Data(:,1)>=xbmin & Data(:,1)<=xbmax)`.

`YBinLimits` — Bin limits in *y*-dimension
vector

Bin limits in *y*-dimension, specified as the comma-separated pair consisting of `'YBinLimits'` and a two-element vector, `[ybmin,ybmax]`. The vector indicates the first and last bin edges in the *y*-dimension.

`binScatterPlot` only plots data that falls within the bin limits inclusively, `Data(Data(:,2)>=ybmin & Data(:,2)<=ybmax)`.

Output Arguments

`h` — Binned scatter plot
`Histogram2` object

Binned scatter plot, returned as a `Histogram2` object. For more information, see Histogram2 Properties.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

Introduced in R2016b

biplot

Biplot

Syntax

```
biplot(coefs)
biplot(coefs,Name,Value)

biplot(ax, ___ )

h = biplot( ___ )
```

Description

`biplot(coefs)` creates a biplot of the coefficients in the matrix `coefs`. The biplot is 2-D if `coefs` has two columns or 3-D if it has three columns. The axes in the biplot represent the columns of `coefs`, and the vectors in the biplot represent the rows of `coefs` (the observed variables).

`biplot(coefs,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify `'Positive','true'` to restrict the biplot to the positive quadrant (in 2-D) or octant (in 3-D).

`biplot(ax, ___ )` uses the plot axes specified by the `Axes` object `ax`. Specify `ax` as the first input argument followed by any of the input argument combinations in the previous syntaxes.

`h = biplot( ___ )` returns a column vector of handles to the graphics objects created by `biplot`. Use `h` to query and modify properties of specific graphics objects. For more information, see Graphics Object Properties.

Examples

Biplot of Coefficients and Scores

Create a biplot of the first three principal component coefficients, the observations, and the observed variables for the `carsmall` data set.

Load the sample data.

```
load carsmall
```

Create a matrix consisting of the variables `Acceleration`, `Displacement`, `Horsepower`, `MPG`, and `Weight`. Delete rows in the matrix that have missing values.

```
X = [Acceleration Displacement Horsepower MPG Weight];
X = rmmissing(X);
```

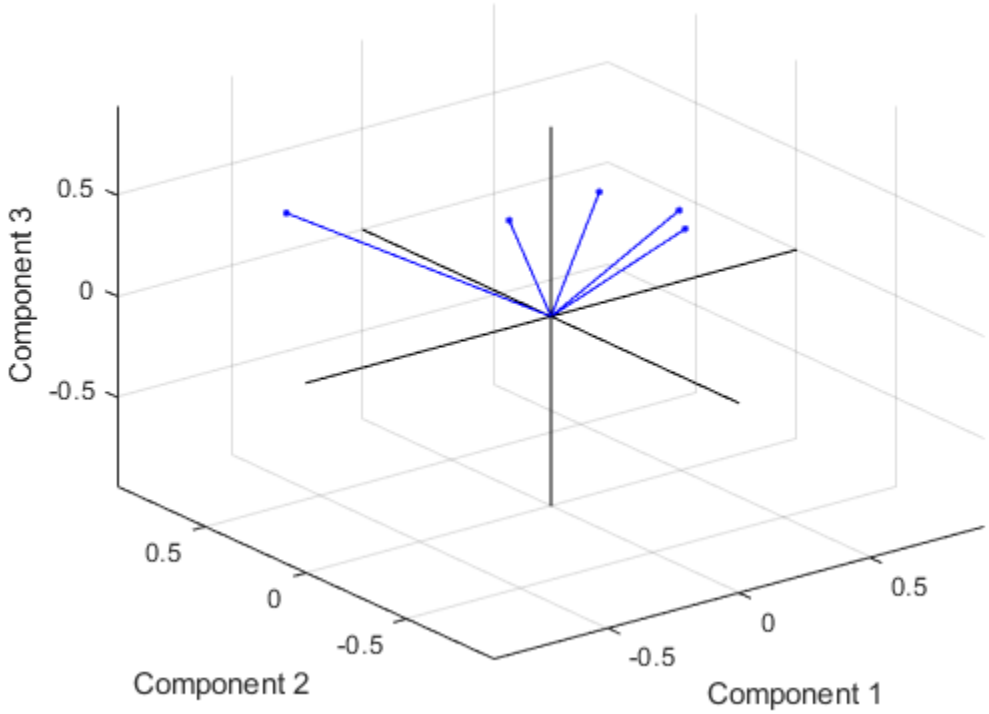
Standardize `X` and perform a principal component analysis.

```
Z = zscore(X); % Standardized data
[coefs,score] = pca(Z);
```

The 5-by-5 matrix `coefs` contains the principal component coefficients (one column for each principal component). The matrix `score` contains the principal component scores (the observations).

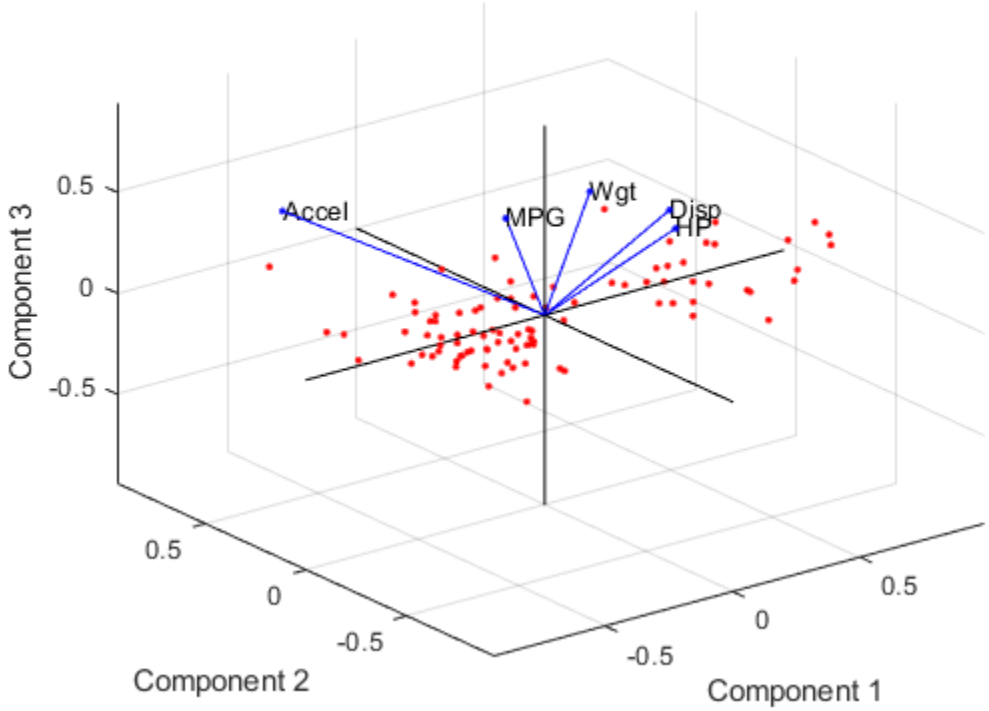
Create a biplot of the first three principal component coefficients. The axes of the biplot represent the columns of `coefs`, and the vectors in the biplot represent the rows of `coefs`.

```
biplot(coefs(:,1:3))
```



Create a more detailed biplot by labeling each variable and plotting the observations in the space of the first three principal components.

```
vbls = {'Accel','Disp','HP','MPG','Wgt'}; % Labels for the variables
biplot(coefs(:,1:3),'Scores',score(:,1:3),'VarLabels',vbls);
```



Specify Axes for Biplot

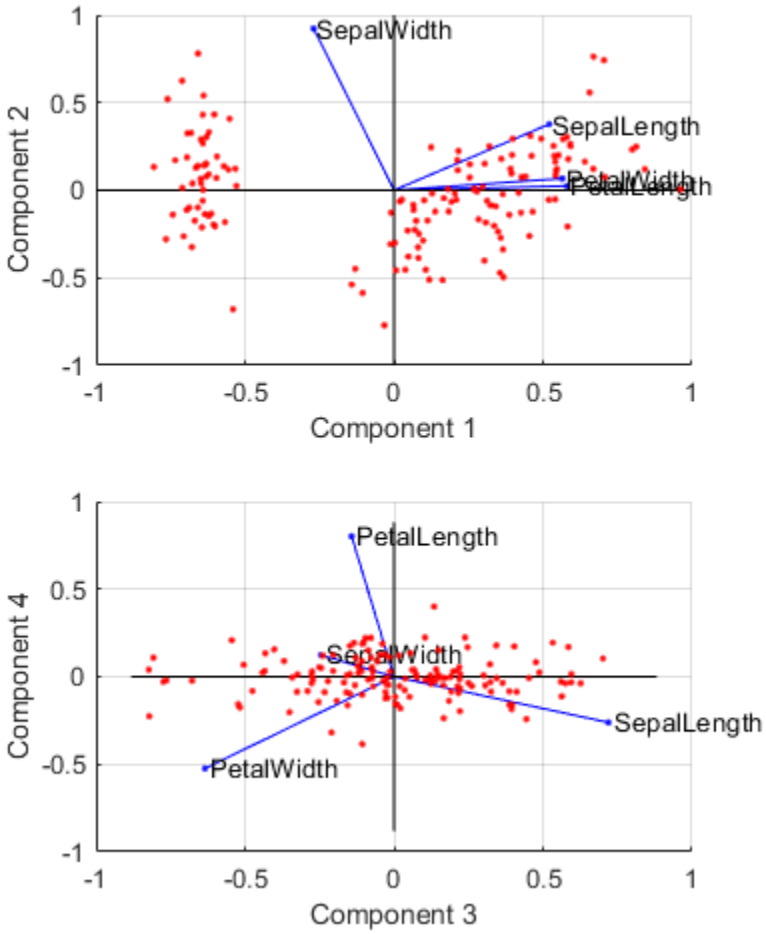
Load the `fisheriris` data set, standardize the flower measurements in `meas`, and perform a principal component analysis.

```
load fisheriris
Z = zscore(meas);
[coefs,scores] = pca(Z);
```

Create a figure with two subplots and return the `Axes` objects as `ax1` and `ax2`. Create a biplot in each set of axes by referring to the corresponding `Axes` object. In the top subplot, display a biplot using the first two principal components. In the bottom subplot, display a biplot using the third and fourth principal components. Specify the x-axis and y-axis limits by passing the corresponding `Axes` objects to the `xlim` and `ylim` functions. Change the x-axis and y-axis labels in the bottom plot by passing `ax2` to `xlabel` and `ylabel`.

```
figure('Units','normalized','Position',[0.3 0.3 0.3 0.5])
variables = {'SepalLength','SepalWidth','PetalLength','PetalWidth'};
ax1 = subplot(2,1,1); % Top subplot
biplot(ax1,coefs(:,1:2),'Scores',scores(:,1:2),'VarLabels',variables);
xlim(ax1,[-1 1])
ylim(ax1,[-1 1])

ax2 = subplot(2,1,2); % Bottom subplot
```

```
biplot(ax2,coefs(:,3:4),'Scores',scores(:,3:4),'VarLabels',variables);
xlim(ax2,[-1 1])
ylim(ax2,[-1 1])
xlabel(ax2,'Component 3')
ylabel(ax2,'Component 4')
```



Modify Biplot Properties

Control the appearance of a biplot by specifying supported line property names and values, and by using handles to the graphics objects created by `biplot`.

Load the sample data.

```
load carsmall
```

Create a matrix consisting of the variables `Acceleration`, `Displacement`, and `MPG`. Delete rows in the matrix that have missing values.

```
X = [Acceleration Displacement MPG];
X = rmmissing(X);
```

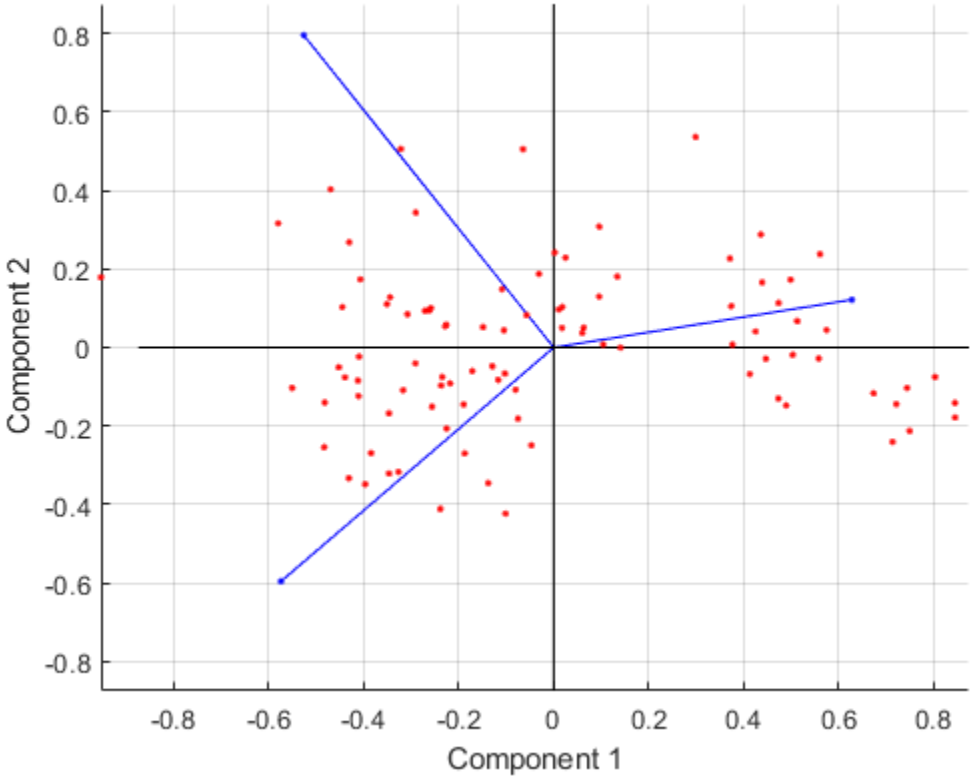
Standardize X and perform a principal component analysis.

```
Z = zscore(X); % Standardized data
[coefs,score] = pca(Z);
```

The 3-by-3 matrix `coefs` contains the principal component coefficients (one column for each principal component). The matrix `score` contains the principal component scores (the observations).

Create a biplot of the observations in the space of the first two principal components. Use the default properties for the biplot.

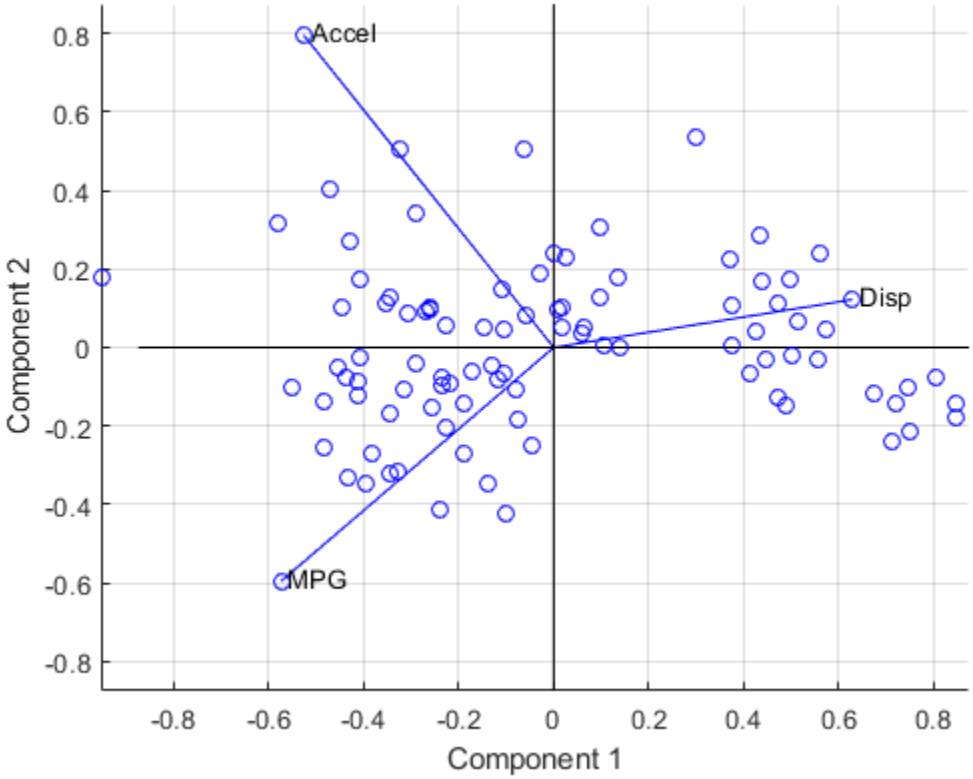
```
h = biplot(coefs(:,1:2),'Scores',score(:,1:2));
```



h is a vector of handles to graphics objects. You can modify the properties of the line objects returned by `biplot`.

Label the three variables for easy identification. Specify circles as the marker symbol and blue as the line color for all line objects.

```
vbls = {'Accel','Disp','MPG'}; % Array of variable labels
h1 = biplot(coefs(:,1:2),'Scores',score(:,1:2),...
    'Color','b','Marker','o','VarLabels',vbls);
```



`h1` is a vector of handles to graphics objects. View the first few elements of `h1`.

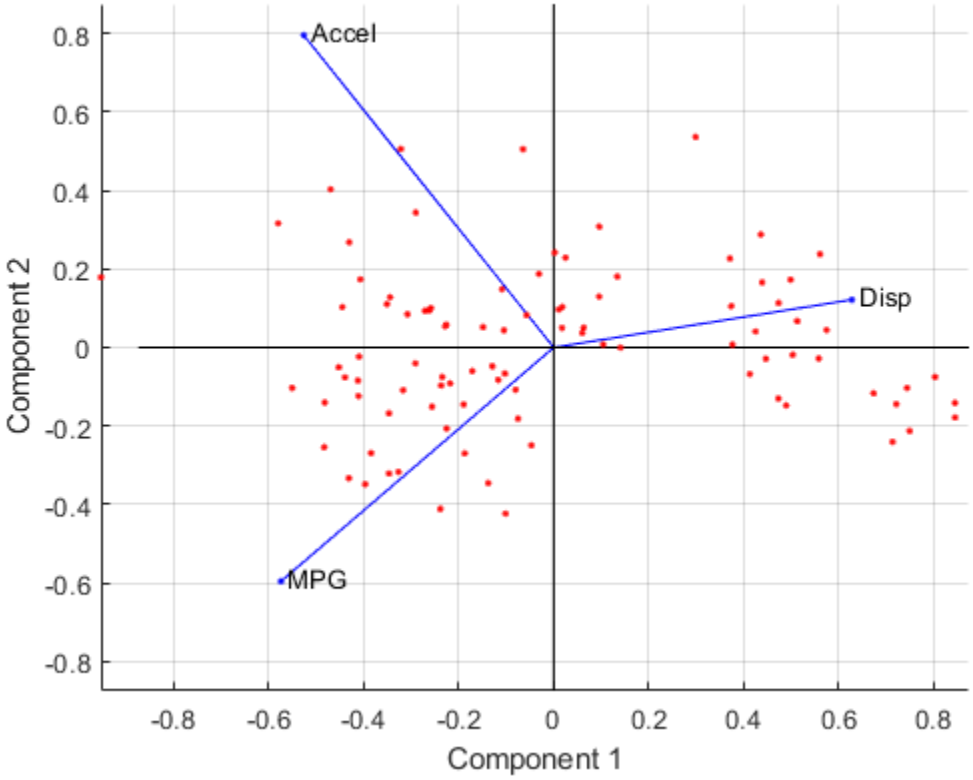
```
h1(1:10) % First ten object handles
```

```
ans =
  10x1 graphics array:

  Line    (varline)
  Line    (varline)
  Line    (varline)
  Line    (varmarker)
  Line    (varmarker)
  Line    (varmarker)
  Text    (varlabel)
  Text    (varlabel)
  Text    (varlabel)
  Line    (obsmarker)
```

The handles for the variable labels (`h1(7:9)`) are text. Therefore, the settings specified for the line properties do not affect these labels.

Create another biplot of the observations in the space of the first two principal components, and label the three variables for easy identification.

```
h2 = biplot(coefs(:,1:2),'Scores',score(:,1:2),'VarLabels',vbls);
```



`h2` is a vector of handles to graphics objects. View the first few elements of `h2`.

```
h2(1:10) % First ten object handles
```

```
ans =
  10x1 graphics array:

  Line    (varline)
  Line    (varline)
  Line    (varline)
  Line    (varmarker)
  Line    (varmarker)
  Line    (varmarker)
  Text    (varlabel)
  Text    (varlabel)
  Text    (varlabel)
  Line    (obsmarker)
```

`h2` contains 104 object handles.

- The first three handles (`h(1:3)`) correspond to line handles for the three variables.
- Handles `h(4:6)` correspond to marker handles for the three variables.
- Handles `h(7:9)` correspond to text handles for the three variables.
- The next 94 handles correspond to line handles for the observations.

- The last handle corresponds to a line handle for the axis lines.

Modify specific properties of the biplot by using handles to the graphics objects.

Change the line color of the variables (vectors).

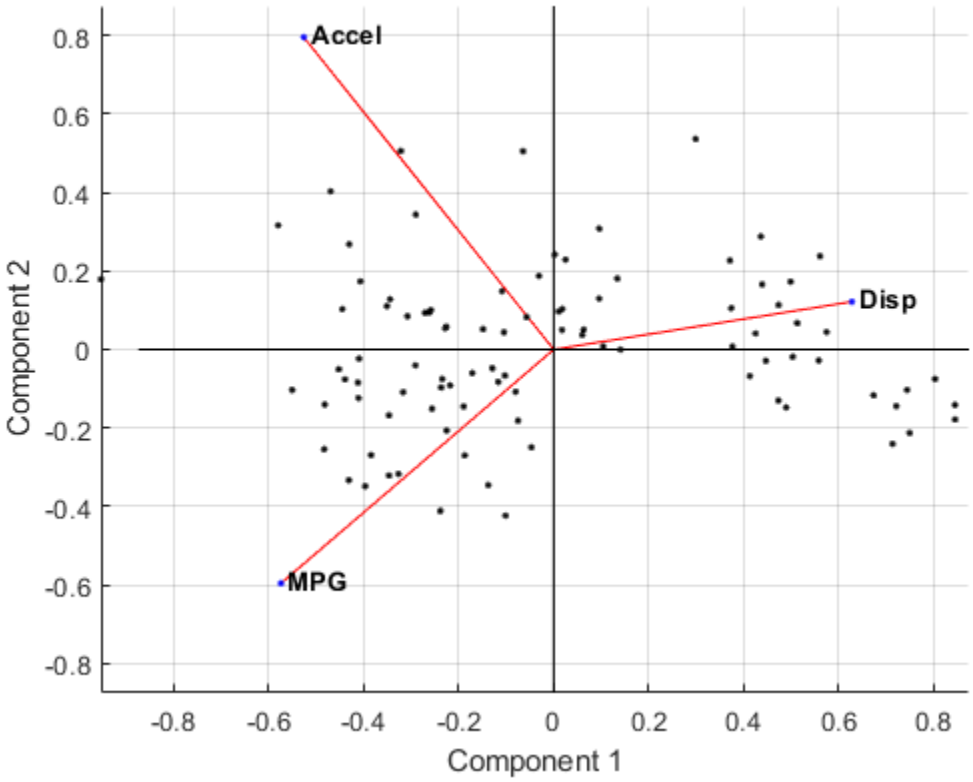
```
for k = 1:3
    h2(k).Color = 'r'; % Specify red as the line color
end
```

Modify the font of the variable labels.

```
for k = 7:9
    h2(k).FontWeight = 'bold';  % Specify bold font
end
```

Change the color of the observation markers.

```
for k = 10:103
    h2(k).MarkerEdgeColor = 'k';  % Specify black color for the observations
end
```



Input Arguments

`coefs` — Coefficients

matrix

Coefficients, specified as a matrix that has two or three columns. If `coefs` has two columns, then the biplot is 2-D; if `coefs` has three columns, then the biplot is 3-D. The columns of `coefs` usually contain principal component coefficients created with `pca` or `pcacov`, or factor loadings estimated with `factoran`. The axes in the biplot represent columns of `coefs`, and the vectors in the biplot represent rows of `coefs` (the observed variables).

Data Types: `single` | `double`

ax — Axes for plot
`Axes` object

Axes for the plot, specified as an `Axes` object. If you do not specify `ax`, then `biplot` creates the plot using the current axes. For more information on creating an `Axes` object, see `axes`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `biplot(coefs,'VarLabels',varlabels)` labels each vector (variable) with the text in the array `varlabels`.

`Scores` — Scores
matrix

Scores, specified as the comma-separated pair consisting of `'Scores'` and a matrix with the same number of columns as `coefs`. `Scores` usually contains principal component scores created with `pca` or factor scores estimated with `factoran`. The `biplot` function represents each row of `Scores` (the observations) as points and each row of `coefs` (the observed variables) as vectors.

Example: `'Scores',score(:,1:3)`

Data Types: `single` | `double`

`VarLabels` — Variable labels
character array | string array | cell array

Variable labels, specified as the comma-separated pair consisting of `'VarLabels'` and a character array, string array, or cell array. `biplot` labels each vector (observed variable) with the text in the array.

Example: `'VarLabels',varlabels`

Data Types: `char` | `string` | `cell`

`ObsLabels` — Observation labels
character array | string array | cell array

Observation labels, specified as the comma-separated pair consisting of `'ObsLabels'` and a character array, string array, or cell array. `biplot` uses the text in the array as observation names when displaying data cursors.

Example: `'ObsLabels',obslabels`

Data Types: `char` | `string` | `cell`

`Positive` — Indicator for plotting in positive coordinates
`false` (default) | `true`

Indicator for plotting in the positive coordinates, specified as the comma-separated pair consisting of `'Positive'` and one of these logical values.

Value	Description
`false`	Creates the biplot over the range `+/– max(coefs(:))` for all coordinates (default)
`true`	Restricts the biplot to the positive quadrant (in 2-D) or octant (in 3-D)

Example: `'Positive',true`

Data Types: `logical`

`PropertyName` — Property name
supported line property value

Property name, specified as the comma-separated pair consisting of a property name and its associated value for one or more supported Line Properties. These properties are the names and values for all primitive line graphics objects created by `biplot`. The specified property names control the appearance and behavior of the graphics objects.

Example: `'Marker','square','MarkerSize',10`

Output Arguments

`h` — Handles to graphics objects
column vector

Handles to the graphics objects created by `biplot`, returned as a column vector. The vector contains handles in this order:

1. Handles corresponding to variables (line handles first, followed by marker handles then text handles)
2. Handles corresponding to observations (marker handles first, followed by text handles)
3. Handles corresponding to the axis lines

You can use the handles to query and modify properties of specific graphics objects. See Graphics Object Handles and Graphics Arrays for more details.

Algorithms

A biplot allows you to visualize the magnitude and sign of each variable's contribution to the first two or three principal components, and to represent each observation in terms of those components. The `biplot` function:

- Imposes a sign convention, forcing the element with the largest magnitude in each column of `coefs` to be positive. This action flips some of the vectors in `coefs` to the opposite direction, but often makes the plot easier to read. Interpretation of the plot is unaffected, because changing the sign of a coefficient vector does not change its meaning.
- Scales the scores so that they fit on the plot. That is, the function divides each score by the maximum absolute value of all scores, and multiplies by the maximum coefficient length of `coefs`. Then `biplot` changes the sign of the score coordinates according to the sign convention for the coefficients.

See Also

`factoran` | `nnmf` | `pca` | `pcacov` | `rotatefactors` | `zscore`

Topics

Principal Component Analysis (PCA) on page 15-68
Line Properties
Graphics Arrays
Access Property Values
Graphics Object Properties

Introduced before R2006a

bootci

Bootstrap confidence interval

Syntax

```
ci = bootci(nboot,bootfun,d)
ci = bootci(nboot,bootfun,d1,...,dN)
ci = bootci(nboot,{bootfun,d},Name,Value)
ci = bootci(nboot,{bootfun,d1,...,dN},Name,Value)
[ci,bootstat] = bootci( ___ )
```

Description

`ci = bootci(nboot,bootfun,d)` computes a 95% bootstrap confidence interval for each statistic computed by the function `bootfun`. The `bootci` function uses `nboot` bootstrap samples in its computation, and creates each bootstrap sample by sampling with replacement from the rows of `d`.

`ci = bootci(nboot,bootfun,d1,...,dN)` creates each bootstrap sample by sampling with replacement from the rows of the nonscalar data arguments in `d1,...,dN`. These nonscalar arguments must have the same number of rows. The `bootci` function passes the samples of nonscalar data and the unchanged scalar data arguments in `d1,...,dN` to `bootfun`.

`ci = bootci(nboot,{bootfun,d},Name,Value)` specifies options using one or more name-value arguments. For example, you can change the type of confidence interval by specifying the `'Type'` name-value argument.

Note that you must pass the `bootfun` and `d` arguments to `bootci` as a single cell array.

`ci = bootci(nboot,{bootfun,d1,...,dN},Name,Value)` specifies options using one or more name-value arguments. For example, you can change the significance level of the confidence interval by specifying the `'Alpha'` name-value argument.

Note that you must pass the `bootfun` and `d1,...,dN` arguments to `bootci` as a single cell array.

`[ci,bootstat] = bootci( ___ )` also returns the bootstrapped statistic computed for each of the `nboot` bootstrap replicate samples, using any of the input argument combinations in the previous syntaxes. Each row of `bootstat` contains the results of applying `bootfun` to one bootstrap sample.

Examples

Bootstrap Confidence Interval

Compute the confidence interval for the capability index in statistical process control.

Generate 30 random numbers from the normal distribution with mean 1 and standard deviation 1.

```
rng('default') % For reproducibility
y = normrnd(1,1,30,1);
```

Specify the lower and upper specification limits of the process. Define the capability index.

```
LSL = -3;
USL = 3;
capable = @(x)(USL-LSL)./(6*std(x));
```

Compute the 95% confidence interval for the capability index by using 2000 bootstrap samples. By default, `bootci` uses the bias corrected and accelerated percentile method to construct the confidence interval.

```
ci = bootci(2000,capable,y)
```

```
ci = 2×1

    0.5937
    0.9900
```

Compute the studentized confidence interval for the capability index.

```
sci = bootci(2000,{capable,y},'Type','student')
```

```
sci = 2×1

    0.5193
    0.9930
```

Bootstrap Confidence Intervals for Nonlinear Regression Model Coefficients

Compute bootstrap confidence intervals for the coefficients of a nonlinear regression model. The technique used in this example involves bootstrapping the predictor and response values, and assumes that the predictor variable is random. For a technique that assumes the predictor variable is fixed and bootstraps the residuals, see “Bootstrap Confidence Intervals for Linear Regression Model Coefficients” on page 33-203.

Note: This example uses `nlinfit`, which is useful when you only need the coefficient estimates or residuals of a nonlinear regression model and you need to repeat fitting a model multiple times, as in the case of bootstrapping. If you need to investigate a fitted regression model further, create a nonlinear regression model object by using `fitnlm`. You can create confidence intervals for the coefficients of the resulting model by using the `coefCI` object function, although this function does not use bootstrapping.

Generate data from the nonlinear regression model $y = b_1 + b_2 \cdot exp(-b_3 x) + \epsilon$, where $b_1 = 1$, $b_2 = 3$, and $b_3 = 2$ are coefficients; the predictor variable x is exponentially distributed with mean 2; and the error term ϵ is normally distributed with mean 0 and standard deviation 0.1.

```
modelfun = @(b,x)(b(1)+b(2)*exp(-b(3)*x));

rng('default') % For reproducibility
b = [1;3;2];
x = exprnd(2,100,1);
y = modelfun(b,x) + normrnd(0,0.1,100,1);
```

Create a function handle for the nonlinear regression model that uses the initial values in `beta0`.

```
beta0 = [2;2;2];
beta = @(predictor,response)nlinfit(predictor,response,modelfun,beta0)
```

```
beta = function_handle with value:
    @(predictor,response)nlinfit(predictor,response,modelfun,beta0)
```

Compute the 95% bootstrap confidence intervals for the coefficients of the nonlinear regression model. Create the bootstrap samples from the generated data `x` and `y`.

```
ci = bootci(1000,beta,x,y)
```

```
ci = 2×3

    0.9821    2.9552    2.0180
    1.0410    3.1623    2.2695
```

The first two confidence intervals include the true coefficient values $b_1 = 1$ and $b_2 = 3$, respectively. However, the third confidence interval does not include the true coefficient value $b_3 = 2$.

Now compute the 99% bootstrap confidence intervals for the model coefficients.

```
newci = bootci(1000,{beta,x,y},'Alpha',0.01)
```

```
newci = 2×3

    0.9730    2.9112    1.9562
    1.0469    3.1876    2.3133
```

All three confidence intervals include the true coefficient values.

Bootstrap Confidence Intervals for Linear Regression Model Coefficients

Compute bootstrap confidence intervals for the coefficients of a linear regression model. The technique used in this example involves bootstrapping the residuals and assumes that the predictor variable is fixed. For a technique that assumes the predictor variable is random and bootstraps the predictor and response values, see “Bootstrap Confidence Intervals for Nonlinear Regression Model Coefficients” on page 33-202.

Note: This example uses `regress`, which is useful when you only need the coefficient estimates or residuals of a regression model and you need to repeat fitting a model multiple times, as in the case of bootstrapping. If you need to investigate a fitted regression model further, create a linear regression model object by using `fitlm`. You can create confidence intervals for the coefficients of the resulting model by using the `coefCI` object function, although this function does not use bootstrapping.

Load the sample data.

```
load hald
```

Perform a linear regression and compute the residuals.

```
x = [ones(size(heat)),ingredients];
y = heat;
b = regress(y,x);
yfit = x*b;
resid = y - yfit;
```

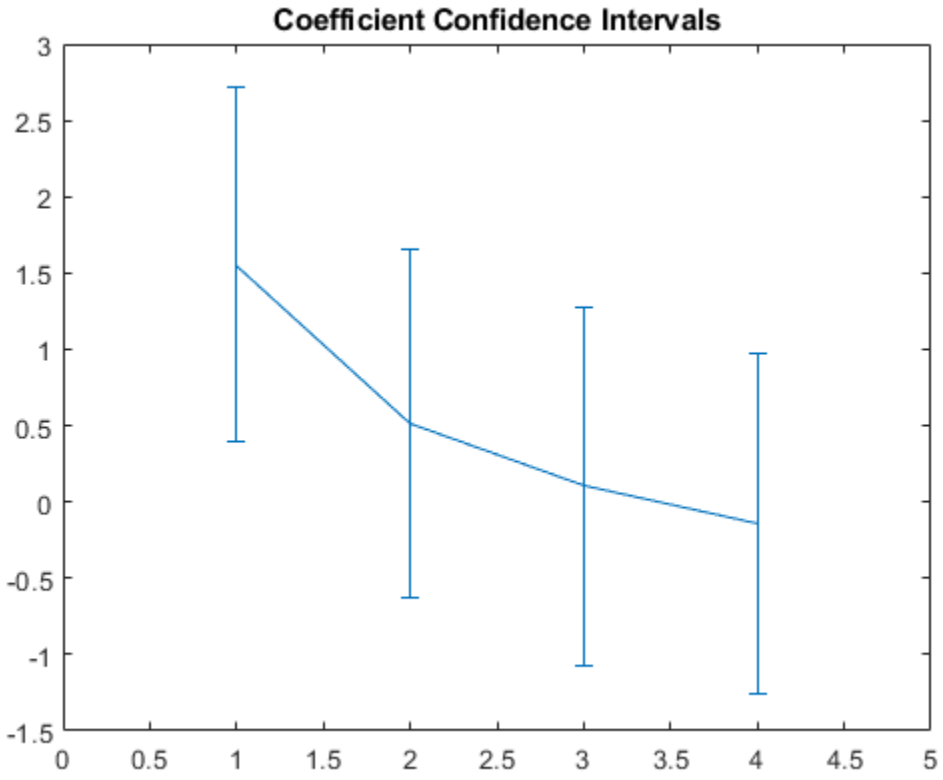
Compute the 95% bootstrap confidence intervals for the coefficients of the linear regression model. Create the bootstrap samples from the residuals. Use normal approximated intervals with bootstrapped bias and standard error by specifying `'Type','normal'`. You cannot use the default confidence interval type in this case.

```
ci = bootci(1000,{@(bootr)regress(yfit+bootr,x),resid}, ...
    'Type','normal')
```

```
ci = 2×5

  -47.7130    0.3916   -0.6298   -1.0697   -1.2604
  172.4899    2.7202    1.6495    1.2778    0.9704
```

Plot the estimated coefficients `b`, omitting the intercept term, and display error bars showing the coefficient confidence intervals.

```
slopes = b(2:end)';
lowerBarLengths = slopes-ci(1,2:end);
upperBarLengths = ci(2,2:end)-slopes;
errorbar(1:4,slopes,lowerBarLengths,upperBarLengths)
xlim([0 5])
title('Coefficient Confidence Intervals')
```



Only the first nonintercept coefficient is significantly different from 0.

Confidence Intervals for Multiple Statistics

Compute the mean and standard deviation of 100 bootstrap samples. Find the 95% confidence interval for each statistic.

Generate 100 random numbers from the exponential distribution with mean 5.

```
rng('default') % For reproducibility
y = exprnd(5,100,1);
```

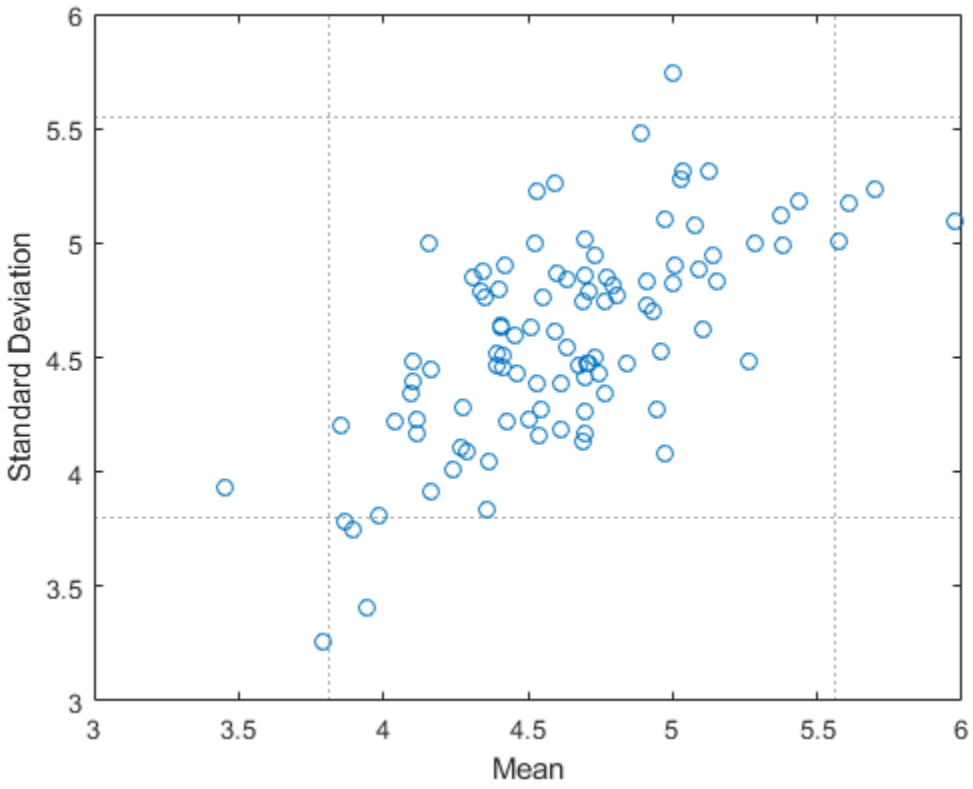
Draw 100 bootstrap samples from the vector `y`. For each bootstrap sample, compute the mean and standard deviation. Find the 95% bootstrap confidence interval for the mean and standard deviation.

```
[ci,bootstat] = bootci(100,@(x)[mean(x) std(x)],y);
```

`ci(:,1)` contains the lower and upper bounds of the mean confidence interval, and `c(:,2)` contains the lower and upper bounds of the standard deviation confidence interval. Each row of `bootstat` contains the mean and standard deviation of a bootstrap sample.

Plot the mean and standard deviation of each bootstrap sample as a point. Plot the lower and upper bounds of the mean confidence interval as dotted vertical lines, and plot the lower and upper bounds of the standard deviation confidence interval as dotted horizontal lines.

```
plot(bootstat(:,1),bootstat(:,2),'o')
xline(ci(1,1),':')
xline(ci(2,1),':')
yline(ci(1,2),':')
yline(ci(2,2),':')
xlabel('Mean')
ylabel('Standard Deviation')
```



Input Arguments

nboot — Number of bootstrap samples

positive integer scalar

Number of bootstrap samples to draw, specified as a positive integer scalar. To create each bootstrap sample, `bootci` randomly selects with replacement `n` out of the `n` rows of nonscalar data in `d` or `d1,...,dN`.

Example: `100`

Data Types: `single` | `double`

bootfun — Function to apply to each sample

function handle

Function to apply to each sample, specified as a function handle. The function can be a custom or built-in function. You must specify `bootfun` with the `@` symbol.

Example: `@mean`

Data Types: `function_handle`

d — Data to sample from

column vector | matrix

Data to sample from, specified as a column vector or matrix. The `n` rows of `d` correspond to observations. When you use multiple data input arguments `d1,...,dN`, you can specify some arguments as scalar values, but all nonscalar arguments must have the same number of rows.

If you use a single vector argument `d`, you can specify it as a row vector. `bootci` then samples from the elements of the vector.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `bootci(100,{@mean,1:6'},'Alpha',0.1)` specifies to draw 100 bootstrap samples from the values 1 through 6, take the mean of each sample, and then compute the 90% confidence interval for the sample mean.

Alpha — Significance level

`0.05` (default) | positive scalar in (0,1)

Significance level, specified as a positive scalar between 0 and 1. `bootci` computes the `100*(1-Alpha)` bootstrap confidence interval of each statistic defined by the function `bootfun`.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

Type — Confidence interval type

`'bca'` (default) | `'norm'` | `'per'` | `'cper'` | `'stud'` | ...

Confidence interval type, specified as one of the values in this table.

Value	Description
`'norm'` or `'normal'`	Normal approximated interval with bootstrapped bias and standard error [1]
`'per'` or `'percentile'`	Basic percentile method
`'cper'` or `'corrected percentile'`	Bias corrected percentile method [2]
`'bca'`	Bias corrected and accelerated percentile method [3], [4] — Involves a z_0 factor that is computed using the proportion of bootstrap values that are less than the original sample value. To produce reasonable results when the sample is lumpy, the software computes z_0 by including half of the bootstrap values that are the same as the original sample value.

Value	Description
`'stud'` or `'student'`	Studentized confidence interval [3]

Example: `'Type','student'`

`NBootStd` — Number of bootstrap samples for studentized standard error estimate
`100` (default) | positive integer scalar

Number of bootstrap samples for the studentized standard error estimate, specified as a positive integer scalar.

`bootci` computes the studentized bootstrap confidence interval of the statistic defined by the function `bootfun`, and estimates the standard error of the bootstrap statistics by using `NBootStd` bootstrap data samples.

Note To use this name-value argument, the `Type` value must be `'stud'` or `'student'`. Specify either `NBootStd` or `StdErr`, but not both.

Example: `'NBootStd',50`

Data Types: `single` | `double`

`StdErr` — Function used to compute studentized standard error estimate
function handle

Function used to compute the studentized standard error estimate, specified as a function handle.

`bootci` computes the studentized bootstrap confidence interval of the statistic defined by the function `bootfun`, and estimates the standard error of the bootstrap statistics by using the function `StdErr`. The `StdErr` function must take the same arguments as `bootfun` and return the standard error of the statistic computed by `bootfun`.

Note To use this name-value argument, the `Type` value must be `'stud'` or `'student'`. Specify either `NBootStd` or `StdErr`, but not both.

Example: `'StdErr',@std`

Data Types: `function_handle`

`Weights` — Observation weights
`ones(n,1)/n` (default) | nonnegative vector

Observation weights, specified as a nonnegative vector with at least one positive element. The number of elements in `Weights` must be equal to the number of rows `n` in the data `d` or `d1,...,dN`. To obtain one bootstrap sample, `bootci` randomly selects with replacement `n` out of `n` rows of data using these weights as multinomial sampling probabilities.

Data Types: `single` | `double`

`Options` — Options for computing iterations in parallel and setting random numbers
structure

Options for computing bootstrap iterations in parallel and setting random numbers during the bootstrap sampling, specified as a structure. Create the `Options` structure with `statset`. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute bootstrap iterations in parallel.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible fashion. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `bootci` uses the default stream or streams.

Note You need Parallel Computing Toolbox to run computations in parallel.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Output Arguments

`ci` — Confidence interval bounds

vector with two rows | matrix with two rows | multidimensional array with two rows

Confidence interval bounds, returned as a vector, matrix, or multidimensional array with two rows.

- If `bootfun` returns a scalar, then `ci` is a vector containing the lower and upper bounds of the confidence interval.
- If `bootfun` returns a vector of length *m*, then `ci` is a matrix of size 2-by-*m*, where `ci(1,:)` are lower bounds and `ci(2,:)` are upper bounds.
- If `bootfun` returns a multidimensional array, then `ci` is an array, where `ci(1,:,...)` is an array of lower bounds and `ci(2,:,...)` is an array of upper bounds.

`bootstat` — Bootstrap statistics

column vector | matrix

Bootstrap statistics, returned as a column vector or matrix with `nboot` rows. The `i`th row of `bootstat` corresponds to the results of applying `bootfun` to the `i`th bootstrap sample. If `bootfun`

returns a matrix or array, then the `bootci` function first converts this output to a row vector before storing it in `bootstat`.

References

[1] Davison, A. C., and D. V. Hinkley. *Bootstrap Methods and Their Applications.* Cambridge University Press, 1997.

[2] Efron, Bradley. *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: The Society for Industrial and Applied Mathematics, 1982.

[3] DiCiccio, Thomas J., and Bradley Efron. "Bootstrap Confidence Intervals." *Statistical Science* 11, no. 3 (1996): 189–228.

[4] Efron, Bradley, and Robert J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see "Run MATLAB Functions with Automatic Parallel Support" (Parallel Computing Toolbox).

See Also

`bootstrp` | `jackknife` | `parfor` | `randsample` | `statget` | `statset`

Topics

"Resampling Statistics" on page 3-14

Introduced in R2006a

bootstrp

Bootstrap sampling

Syntax

```
bootstat = bootstrp(nboot,bootfun,d)
bootstat = bootstrp(nboot,bootfun,d1,...,dN)
bootstat = bootstrp(___,Name,Value)
[bootstat,bootsam] = bootstrp(___)
```

Description

`bootstat = bootstrp(nboot,bootfun,d)` draws `nboot` bootstrap data samples from `d`, computes statistics on each sample using the function `bootfun`, and returns the results in `bootstat`. The `bootstrp` function creates each bootstrap sample by sampling with replacement from the rows of `d`. Each row of the output argument `bootstat` contains the results of applying `bootfun` to one bootstrap sample.

`bootstat = bootstrp(nboot,bootfun,d1,...,dN)` draws `nboot` bootstrap samples from the data in `d1,...,dN`. The nonscalar data arguments in `d1,...,dN` must have the same number of rows, `n`. The `bootstrp` function creates each bootstrap sample by sampling with replacement from the indices `1:n` and selecting the corresponding rows of the nonscalar `d1,...,dN`. The function passes the sample of nonscalar data and the unchanged scalar data arguments in `d1,...,dN` to `bootfun`.

`bootstat = bootstrp(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can add observation weights to your data or compute bootstrap iterations in parallel.

`[bootstat,bootsam] = bootstrp(___)` also returns `bootsam`, an `n`-by-`nboot` matrix of bootstrap sample indices, where `n` is the number of rows in the original, nonscalar data. Each column in `bootsam` corresponds to one bootstrap sample and contains the row indices of the values drawn from the nonscalar data to create that sample.

To get the bootstrap sample indices without applying a function to the samples, set `bootfun` to empty (`[]`).

Examples

Estimate Density of Bootstrapped Statistic

Estimate the kernel density of bootstrapped means.

Generate 100 random numbers from the exponential distribution with mean 5.

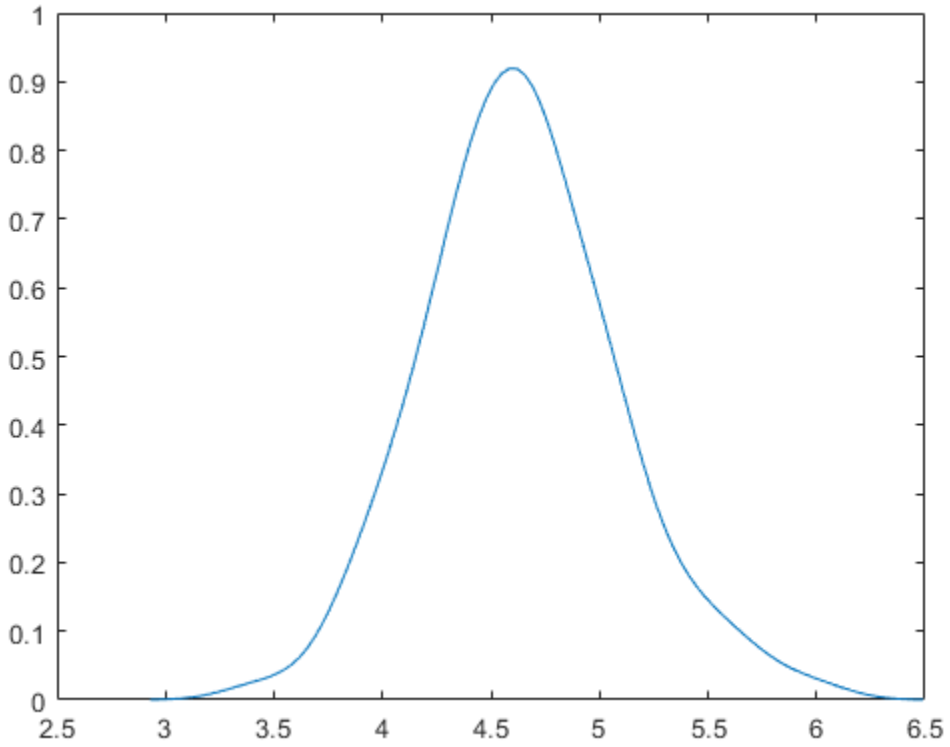
```
rng('default')  % For reproducibility
y = exprnd(5,100,1);
```

Compute a sample of 100 bootstrapped means of random samples taken from the vector `y`.

```
m = bootstrp(100,@mean,y);
```

Plot an estimate of the density of the bootstrapped means.

```
[fi,xi] = ksdensity(m);
plot(xi,fi)
```



Bootstrapping Multiple Statistics

Compute and plot the means and standard deviations of 100 bootstrap samples.

Generate 100 random numbers from the exponential distribution with mean 5.

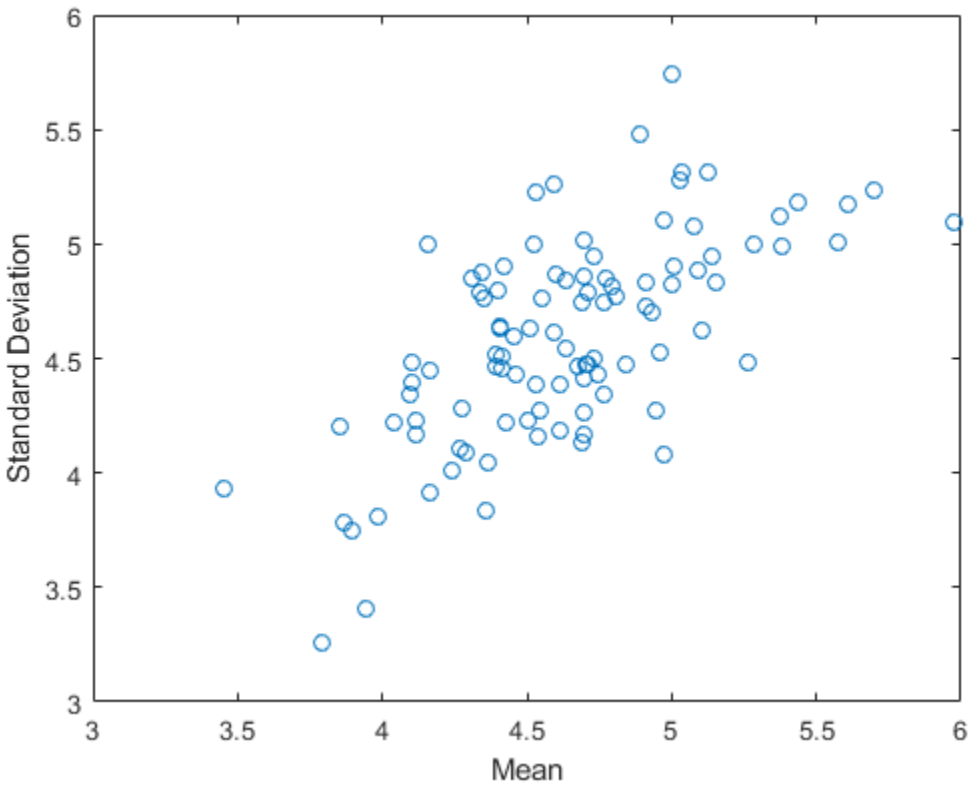
```
rng('default')  % For reproducibility
y = exprnd(5,100,1);
```

Compute a sample of 100 bootstrapped means and standard deviations of random samples taken from the vector `y`.

```
stats = bootstrp(100,@(x)[mean(x) std(x)],y);
```

Plot the bootstrap estimate pairs.

```
plot(stats(:,1),stats(:,2),'o')
xlabel('Mean')
ylabel('Standard Deviation')
```



Bootstrap Samples of Observations

Take bootstrap samples of patient data, compute the mean measurements for each data sample, and visualize the results.

Load the `patients` data set. Create the matrix `patientData` containing age, weight, and height measurements. Each row of `patientData` corresponds to one patient.

```
load patients
patientData = [Age Weight Height];
```

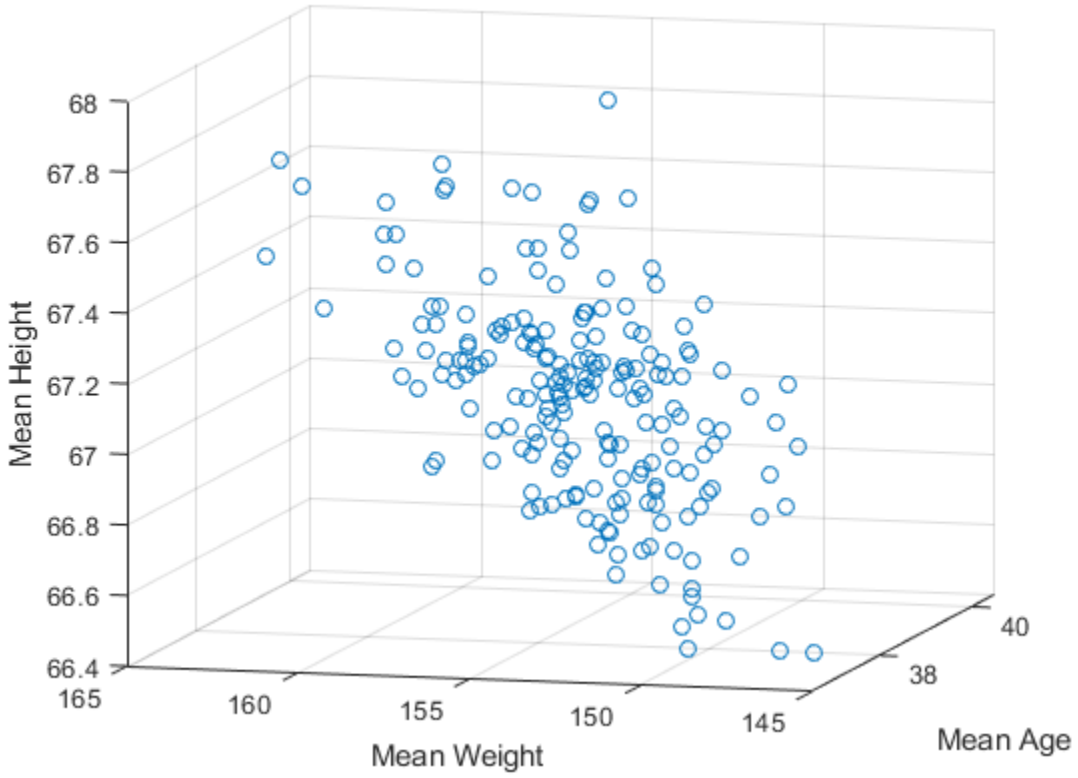
Create 200 bootstrap data samples from the data in `patientData`. To create each sample, randomly select with replacement 100 rows (that is, `size(patientData,1)`) from the rows in `patientData`. For each sample, calculate the mean age, weight, and height measurements. Each row of `bootstat` contains the three mean measurements for one bootstrap sample.

```
rng('default') % For reproducibility
bootstat = bootstrp(200,@mean,patientData);
```

Visualize the mean measurements for all 200 bootstrap data samples. Note that bootstrap samples with greater mean weights tend to have greater mean heights.

```
scatter3(bootstat(:,1),bootstat(:,2),bootstat(:,3))
xlabel('Mean Age')
```

```
ylabel('Mean Weight')
zlabel('Mean Height')

view([-75 10])
```



Bootstrapping Correlation Coefficient Standard Error

Compute a correlation coefficient standard error using bootstrap resampling of the sample data.

Load the `lawdata` data set, which contains the LSAT score and law school GPA for 15 students.

```
load lawdata
rng('default')  % For reproducibility
size(lsat)
```

```
ans = 1×2

    15     1
```

```
size(gpa)
```

```
ans = 1×2

    15     1
```

Create 1000 data samples by resampling the 15 data points, and compute the correlation between the two variables for each data sample.

```
[bootstat,bootsam] = bootstrp(1000,@corr,lsat,gpa);
```

Display the first 5 bootstrapped correlation coefficients.

```
bootstat(1:5,:)
```

```
ans = 5×1

    0.9874
    0.4918
    0.5459
    0.8458
    0.8959
```

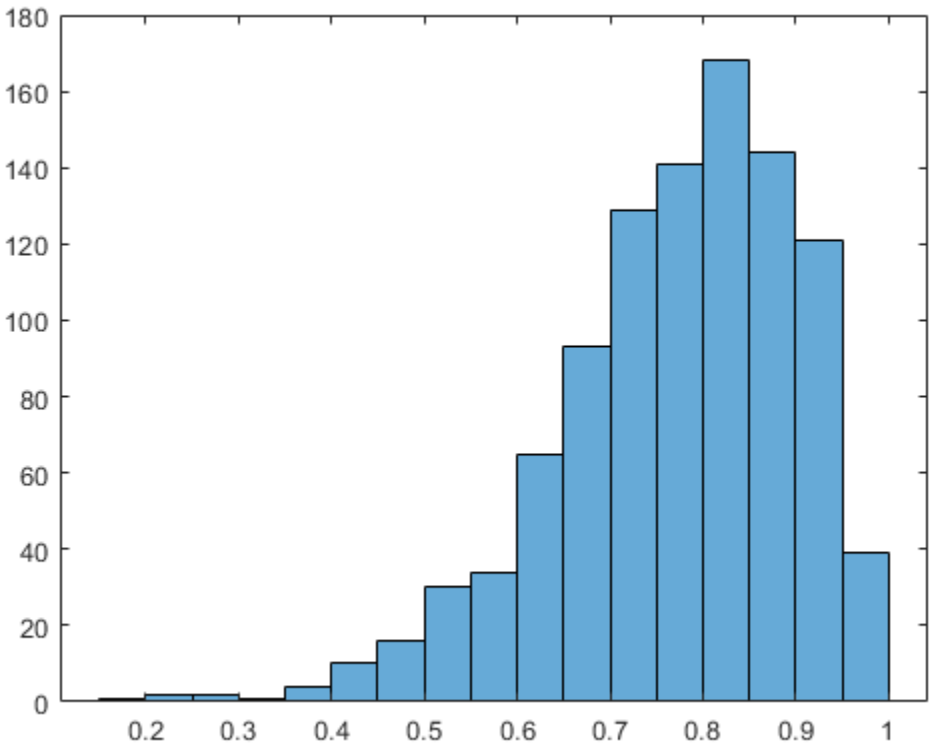
Display the indices of the data selected for the first 5 bootstrap samples.

```
bootsam(:,1:5)
```

```
ans = 15×5

    13     3    11     8    12
    14     7     1     7     4
     2    14     5    10     8
    14    12     1    11    11
    10    15     2    12    14
     2    10    13     5    15
     5     1    11    11     9
     9    13     5    10     3
    15    15    15     3     3
    15    11     1     2     4
     ⋮
```

Create a histogram that shows the variation of the correlation coefficient across all the bootstrap samples.

```
histogram(bootstat)
```



The sample minimum is positive, indicating that the relationship between LSAT score and GPA is not accidental.

Finally, compute a bootstrap standard of error for the estimated correlation coefficient.

```
se = std(bootstat)
```

```
se = 0.1285
```

Compare Bootstrap Samples with Different Observation Weights

Compare bootstrap samples with different observation weights. Create a custom function that computes statistics for each sample.

Create 50 bootstrap samples from the numbers 1 through 6. To create each sample, `bootstrp` randomly chooses with replacement from the numbers 1 through 6, six times. This process is similar to rolling a die six times. For each sample, the custom function `countfun` (shown at the end of this example) counts the number of 1s in the sample.

```
rng('default') %For reproducibility
counts = bootstrp(50,@countfun,(1:6)');
```

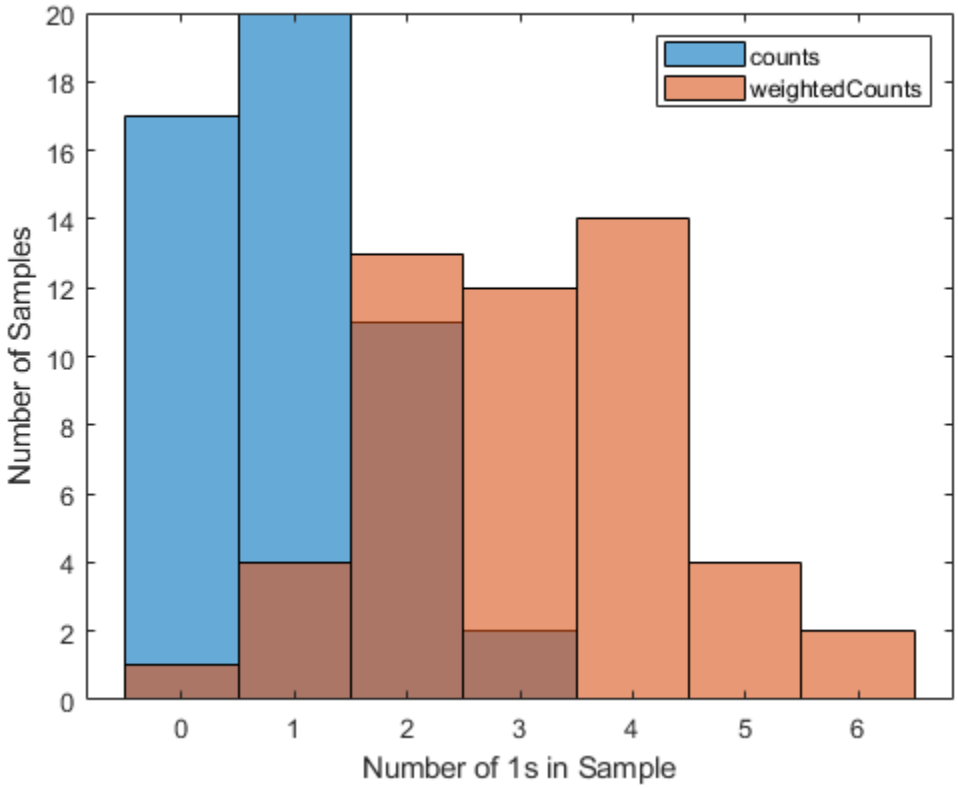
Note: If you use the live script file for this example, the `countfun` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Create 50 bootstrap samples from the numbers 1 through 6, but assign different weights to the numbers. Each time `bootstrp` randomly chooses from the numbers 1 through 6, the probability of choosing a 1 is `0.5`, the probability of choosing a 2 is `0.1`, and so on. Again, `countfun` counts the number of 1s in each sample.

```
weights = [0.5 0.1 0.1 0.1 0.1 0.1]';
weightedCounts = bootstrp(50,@countfun,(1:6)','Weights',weights);
```

Compare the two sets of bootstrap samples by using histograms.

```
histogram(counts)
hold on
histogram(weightedCounts)
legend
xlabel('Number of 1s in Sample')
ylabel('Number of Samples')
hold off
```



The two sets of bootstrap samples have different distributions; in particular, the samples in the second set tend to contain more 1s. For example, of the 50 samples in the first set, only two samples contain more than two 1s. By contrast, of the 50 samples in the second set (with observation weights), 12 + 14 + 4 + 2 = 32 samples contain more than two 1s.

This code creates the function `countfun`.

```
function numberofones = countfun(sample)
numberofones = sum(sample == 1);
end
```

Bootstrapping Regression Model

Estimate the standard errors for a coefficient vector in a linear regression by bootstrapping the residuals.

Note: This example uses `regress`, which is useful when you simply need the coefficient estimates or residuals of a regression model and you need to repeat fitting a model multiple times, as in the case of bootstrapping. If you need to investigate a fitted regression model further, create a linear regression model object by using `fitlm`.

Load the sample data.

```
load hald
```

Perform a linear regression, and compute the residuals.

```
x = [ones(size(heat)),ingredients];
y = heat;
b = regress(y,x);
yfit = x*b;
resid = y - yfit;
```

Estimate the standard errors by bootstrapping the residuals.

```
se = std(bootstrp(1000,@(bootr)regress(yfit+bootr,x),resid))
```

```
se = 1×5

   56.1752    0.5940    0.5815    0.5989    0.5691
```

Input Arguments

`nboot` — Number of bootstrap samples

positive integer scalar

Number of bootstrap samples to draw, specified as a positive integer scalar. To create each bootstrap sample, `bootstrp` randomly selects with replacement `n` out of the `n` rows of (nonscalar) data in `d` or `d1,...,dN`.

Example: `100`

Data Types: `single` | `double`

`bootfun` — Function to apply to each sample

function handle

Function to apply to each sample, specified as a function handle. The function can be a custom or built-in function. You must specify `bootfun` with the `@` symbol.

For an example that uses a custom function, see “Compare Bootstrap Samples with Different Observation Weights” on page 33-216.

Example: `@mean`

Data Types: `function_handle`

`d` — Data to sample from

column vector | matrix

Data to sample from, specified as a column vector or matrix. The `n` rows of `d` correspond to observations. When you use multiple data input arguments `d1,...,dN`, you can specify some arguments as scalar values, but all nonscalar arguments must have the same number of rows.

If you use a single vector argument `d`, you can specify it as a row vector. `bootstrp` then samples from the elements of the vector.

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `bootstrp(4,@mean,(1:2)','Weights',[0.4 0.6]')` specifies to draw four bootstrap samples from the values `1` and `2` and take the mean of each sample. For each draw, the probability of getting a `1` is `0.4`, and the probability of getting a `2` is `0.6`.

`Weights` — Observation weights

`ones(n,1)/n` (default) | nonnegative vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a nonnegative vector with at least one positive element. The number of elements in `Weights` must be equal to the number of rows `n` in the data `d` or `d1,...,dN`. To obtain one bootstrap sample, `bootstrp` randomly selects with replacement `n` out of `n` rows of data using the weights as multinomial sampling probabilities.

Data Types: `single` | `double`

`Options` — Options for computing in parallel and setting random numbers

structure

Options for computing bootstrap iterations in parallel and setting random numbers during the bootstrap sampling, specified as the comma-separated pair consisting of `'Options'` and a structure. Create the `Options` structure with `statset`. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute bootstrap iterations in parallel.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`

Field Name	Value	Default
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `bootstrp` uses the default stream or streams.

Note You need Parallel Computing Toolbox to run computations in parallel.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Output Arguments

`bootstat` — Bootstrap sample statistics

column vector | matrix

Bootstrap sample statistics, returned as a column vector or matrix with `nboot` rows. The `i`th row of `bootstat` corresponds to the results of applying `bootfun` to the `i`th bootstrap sample. If `bootfun` returns a matrix or array, then the `bootstrp` function first converts this output to a row vector before storing it in `bootstat`.

`bootsam` — Bootstrap sample indices

numeric matrix

Bootstrap sample indices, returned as an `n`-by-`nboot` numeric matrix, where `n` is the number of rows in the original, nonscalar data. Each column in `bootsam` corresponds to one bootstrap sample and contains the row indices of the values drawn from the nonscalar data to create that sample.

For example, if each data input argument in `d1,...,dN` contains 16 values, and `nboot = 4`, then `bootsam` is a 16-by-4 matrix. The first column contains the indices of the 16 values drawn from `d1,...,dN` for the first bootstrap sample, the second column contains the indices for the second bootstrap sample, and so on. The bootstrap indices are the same for all input data sets `d1,...,dN`.

Tips

- To get the bootstrap sample indices `bootsam` without applying a function to the samples, set `bootfun` to empty (`[]`).

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RandStream` | `bootci` | `histogram` | `ksdensity` | `parfor` | `random` | `randsample` | `statget` | `statset`

Topics

“Resampling Statistics” on page 3-14

Introduced before R2006a

boxplot

Visualize summary statistics with box plot

Syntax

```
boxplot(x)
boxplot(x,g)

boxplot(ax, ___ )

boxplot( ___ ,Name,Value)
```

Description

`boxplot(x)` creates a box plot of the data in `x`. If `x` is a vector, `boxplot` plots one box. If `x` is a matrix, `boxplot` plots one box for each column of `x`.

On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the `'+'` symbol.

`boxplot(x,g)` creates a box plot using one or more grouping variables contained in `g`. `boxplot` produces a separate box for each set of `x` values that share the same `g` value or values.

`boxplot(ax, ___ )` creates a box plot using the axes specified by the axes graphic object `ax`, using any of the previous syntaxes.

`boxplot( ___ ,Name,Value)` creates a box plot with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the box style or order.

Examples

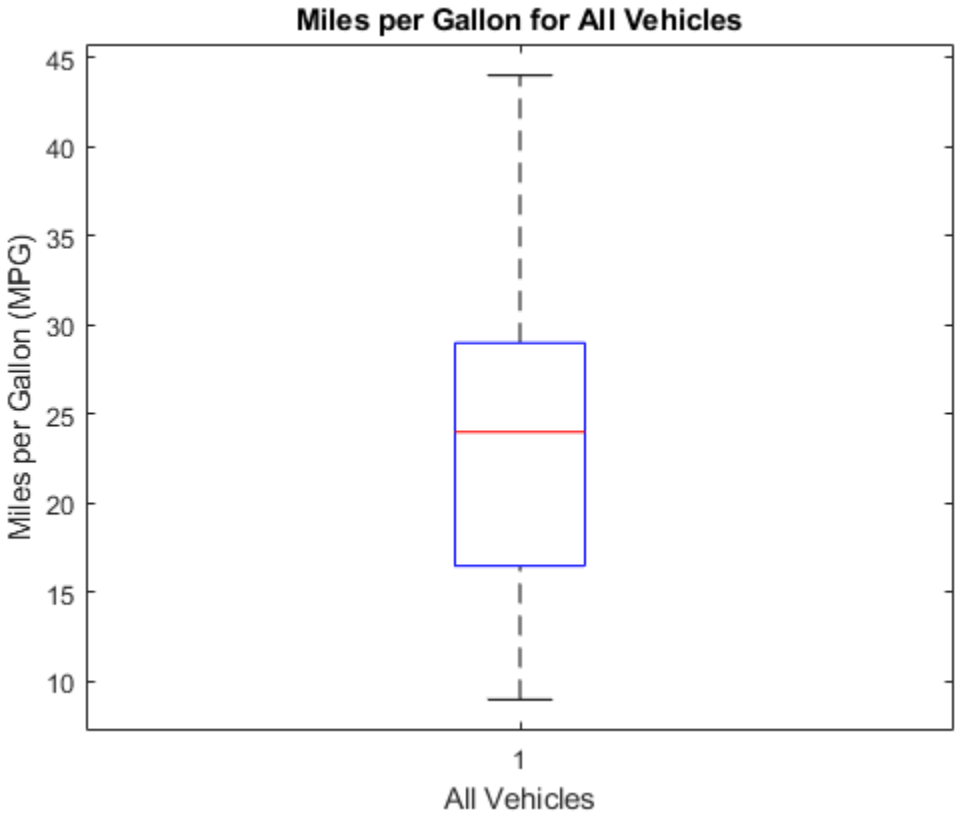
Create a Box Plot

Load the sample data.

```
load carsmall
```

Create a box plot of the miles per gallon (`MPG`) measurements. Add a title and label the axes.

```
boxplot(MPG)
xlabel('All Vehicles')
ylabel('Miles per Gallon (MPG)')
title('Miles per Gallon for All Vehicles')
```



The boxplot shows that the median miles per gallon for all vehicles in the sample data is approximately 24. The minimum value is about 9, and the maximum value is about 44.

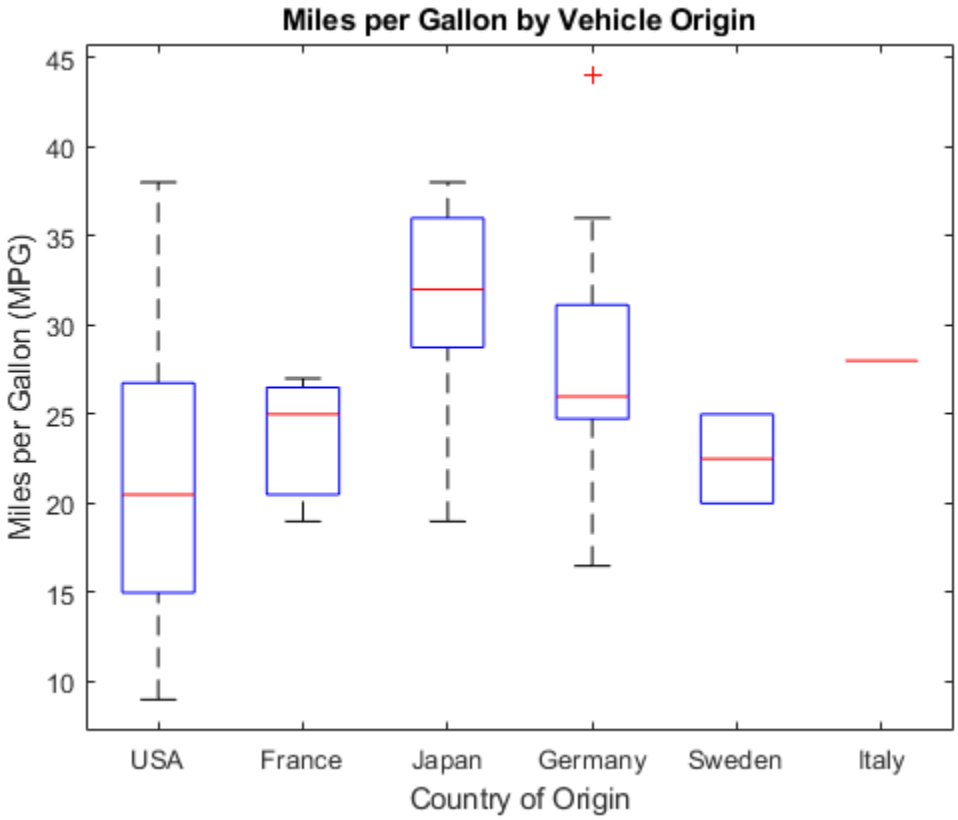
Create Box Plots for Grouped Data

Load the sample data.

```
load carsmall
```

Create a box plot of the miles per gallon (`MPG`) measurements from the sample data, grouped by the vehicles' country of origin (`Origin`). Add a title and label the axes.

```
boxplot(MPG,Origin)
title('Miles per Gallon by Vehicle Origin')
xlabel('Country of Origin')
ylabel('Miles per Gallon (MPG)')
```



Each box visually represents the MPG data for cars from the specified country. Italy's "box" appears as a single line because the sample data contains only one observation for this group.

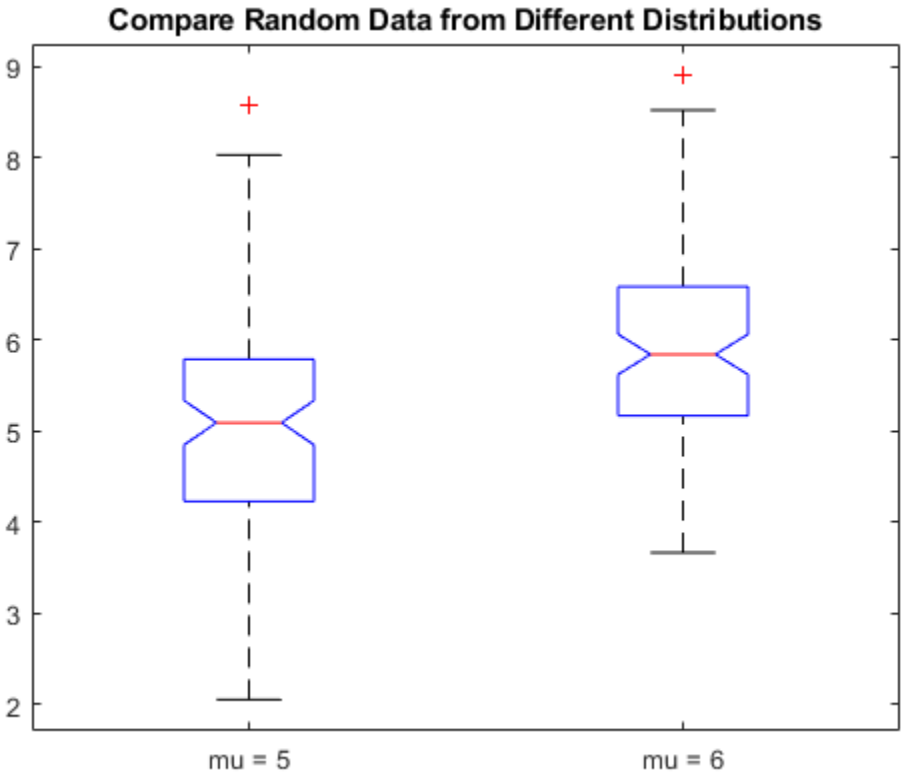
Create Notched Box Plots

Generate two sets of sample data. The first sample, `x1`, contains random numbers generated from a normal distribution with `mu = 5` and `sigma = 1`. The second sample, `x2`, contains random numbers generated from a normal distribution with `mu = 6` and `sigma = 1`.

```
rng default  % For reproducibility
x1 = normrnd(5,1,100,1);
x2 = normrnd(6,1,100,1);
```

Create notched box plots of `x1` and `x2`. Label each box with its corresponding `mu` value.

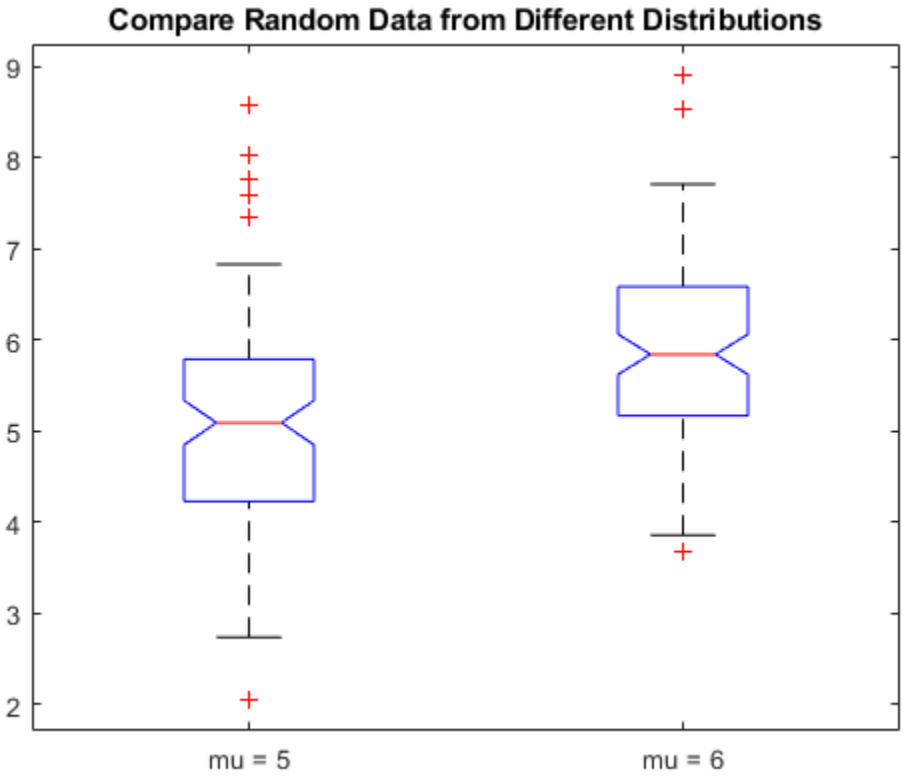
```
figure
boxplot([x1,x2],'Notch','on','Labels',{'mu = 5','mu = 6'})
title('Compare Random Data from Different Distributions')
```



The boxplot shows that the difference between the medians of the two groups is approximately 1. Since the notches in the box plot do not overlap, you can conclude, with 95% confidence, that the true medians do differ.

The following figure shows the box plot for the same data with the maximum whisker length specified as 1.0 times the interquartile range. Data points beyond the whiskers are displayed using +.

```
figure
boxplot([x1,x2],'Notch','on','Labels',{'mu = 5','mu = 6'},'Whisker',1)
title('Compare Random Data from Different Distributions')
```



With the smaller whiskers, `boxplot` displays more data points as outliers.

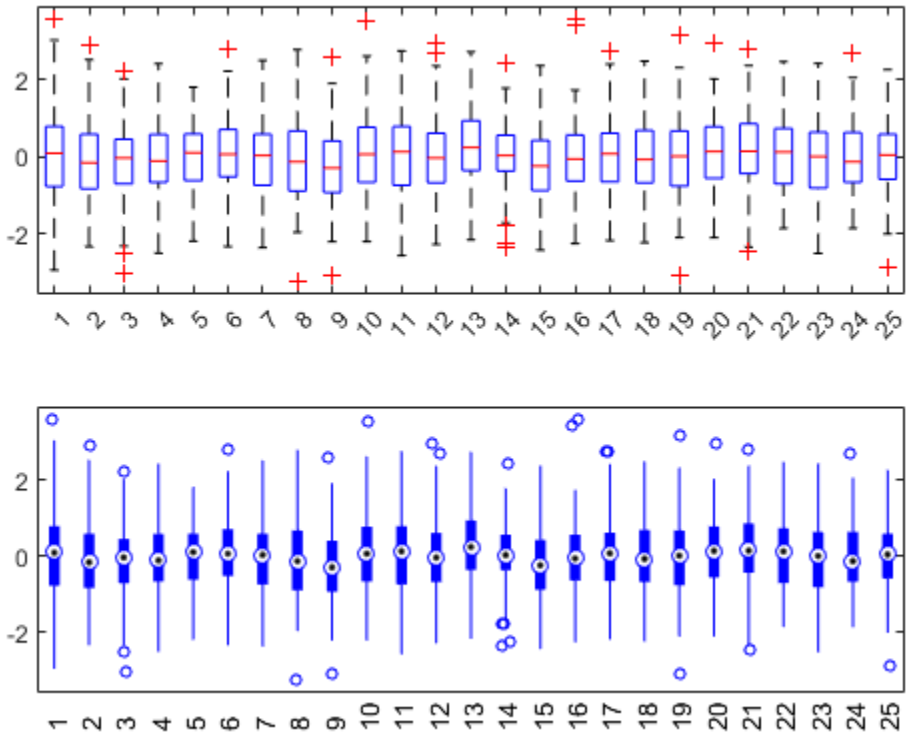
Create Compact Box Plots

Create a 100-by-25 matrix of random numbers generated from a standard normal distribution to use as sample data.

```
rng default  % For reproducibility
x = randn(100,25);
```

Create two box plots for the data in `x` on the same figure. Use the default formatting for the top plot, and compact formatting for the bottom plot.

```
figure

subplot(2,1,1)
boxplot(x)

subplot(2,1,2)
boxplot(x,'PlotStyle','compact')
```



Each plot presents the same data, but the compact formatting may improve readability for plots with many boxes.

Box Plots for Vectors of Varying Length

Create box plots for data vectors of varying length by using a grouping variable.

Randomly generate three column vectors of varying length: one of length `5`, one of length `10`, and one of length `15`. Combine the data into a single column vector of length `30`.

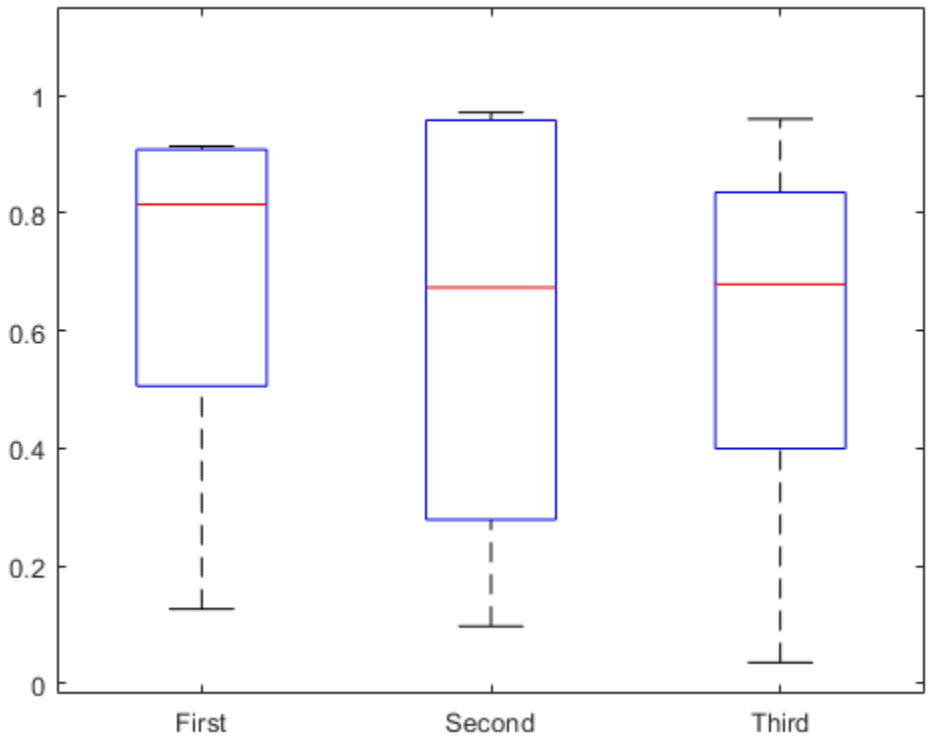
```
rng('default')  % For reproducibility
x1 = rand(5,1);
x2 = rand(10,1);
x3 = rand(15,1);
x = [x1; x2; x3];
```

Create a grouping variable that assigns the same value to rows that correspond to the same vector in `x`. For example, the first five rows of `g` have the same value, `First`, because the first five rows of `x` all come from the same vector, `x1`.

```
g1 = repmat({'First'},5,1);
g2 = repmat({'Second'},10,1);
g3 = repmat({'Third'},15,1);
g = [g1; g2; g3];
```

Create the box plots.

```
boxplot(x,g)
```



Input Arguments

x — Input data

numeric vector | numeric matrix

Input data, specified as a numeric vector or numeric matrix. If `x` is a vector, `boxplot` plots one box. If `x` is a matrix, `boxplot` plots one box for each column of `x`.

On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the `'+'` symbol.

Data Types: `single` | `double`

g — Grouping variables

numeric vector | character array | string array | cell array | categorical array

Grouping variables, specified as a numeric vector, character array, string array, cell array, or categorical array. You can specify multiple grouping variables in `g` by using a cell array of these variable types or a matrix. If you specify multiple grouping variables, they must all be the same length.

If `x` is a vector, then the grouping variables must contain one row for each element of `x`. If `x` is a matrix, then the grouping variables must contain one row for each column of `x`. Groups that contain a missing value (`NaN`), an empty character vector, an empty or `<missing>` string, or an `<undefined>` value in a grouping variable are omitted, and are not counted in the number of groups considered by other parameters.

By default, `boxplot` sorts character and string grouping variables in the order they initially appear in the data, categorical grouping variables by the order of their levels, and numeric grouping variables in numeric order. To control the order of groups, do one of the following:

- Use categorical variables in `g` and specify the order of their levels.
- Use the `'GroupOrder'` name-value pair argument.
- Pre-sort your data.

Data Types: `single` | `double` | `char` | `string` | `cell` | `categorical`

ax — Axes on which to plot
axes graphic object

Axes on which to plot, specified as an axes graphic object. If you do not specify `ax`, then `boxplot` creates the plot using the current axis. For more information on creating an axes graphic object, see `axes` and Axes.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Notch','on','Labels',{'mu = 5','mu = 6'}` creates a notched box plot and labels the two boxes `mu = 5` and `mu = 6`, from left to right

Box Appearance

BoxStyle — Box style
`'outline'` | `'filled'`

Box style, specified as the comma-separated pair consisting of `'BoxStyle'` and one of the following.

Name	Value
`'outline'`	Plot boxes using an unfilled box with dashed whiskers. This is the default if `'PlotStyle'` is `'traditional'`.
`'filled'`	Plot boxes using a narrow filled box with lines for whiskers. This is the default if `'PlotStyle'` is `'compact'`.

Example: `'BoxStyle','filled'`

Colors — Box colors
RGB triplet | character vector or string scalar of color names

Box colors, specified as the comma-separated pair consisting of `'Colors'` and an RGB triplet, character vector, or string scalar. An RGB triplet is a three-element row vector whose elements

specify the intensities of the red, green, and blue components of the color, respectively. Each intensity must be in the range [0,1].

The following table lists the available color characters and their equivalent RGB triplet values.

Long Name	Short Name	RGB Triplet
Yellow	`'y'`	`[1 1 0]`
Magenta	`'m'`	`[1 0 1]`
Cyan	`'c'`	`[0 1 1]`
Red	`'r'`	`[1 0 0]`
Green	`'g'`	`[0 1 0]`
Blue	`'b'`	`[0 0 1]`
White	`'w'`	`[1 1 1]`
Black	`'k'`	`[0 0 0]`

You can specify multiple colors either as a character vector or string scalar of color names (for example, `'rgbm'`) or a three-column matrix of RGB values. The sequence is replicated or truncated as required, so for example, `'rb'` gives boxes that alternate red and blue.

If you do not specify the name-value pair `'ColorGroup'`, then `boxplot` uses the same color scheme for all boxes. If you do specify `'ColorGroup'`, then the default is a modified `hsv` `colormap`.

Example: `'Colors','rgbm'`

`MedianStyle` — Median style
`'line'` | `'target'`

Median style, specified as the comma-separated pair consisting of `'MedianStyle'` and one of the following.

Name	Value
`'line'`	Draw a line to represent the median in each box. This is the default when `'PlotStyle'` is `'traditional'`.
`'target'`	Draw a black dot inside a white circle to represent the median in each box. This is the default when `'PlotStyle'` is `'compact'`.

Example: `'MedianStyle','target'`

`Notch` — Marker for comparison intervals
`'off'` (default) | `'on'` | `'marker'`

Marker for comparison intervals, specified as the comma-separated pair consisting of `'Notch'` and one of the following.

Name	Value
`'off'`	Omit comparison intervals from box display.

Name	Value
`'on'`	If `'PlotStyle'` is `'traditional'`, draw comparison intervals using notches. If `'PlotStyle'` is `'compact'`, draw comparison intervals using triangular markers.
`'marker'`	Draw comparison intervals using triangular markers.

Two medians are significantly different at the 5% significance level if their intervals do not overlap. `boxplot` represents interval endpoints using the extremes of the notches or the centers of the triangular markers. The notch extremes correspond to $q_2 - 1.57(q_3 - q_1)/\text{sqrt}(n)$ and $q_2 + 1.57(q_3 - q_1)/\text{sqrt}(n)$, where q_2 is the median (50th percentile), q_1 and q_3 are the 25th and 75th percentiles, respectively, and n is the number of observations without any `NaN` values. If the sample size is small, the notches might extend beyond the end of the box.

Example: `'Notch','on'`

`OutlierSize` — Marker size for outliers
positive numeric value

Marker size for outliers, specified as the comma-separated pair consisting of `'OutlierSize'` and a positive numeric value. The specified value represents the marker size in points.

If `'PlotStyle'` is `'traditional'`, then the default value for `OutlierSize` is 6. If `'PlotStyle'` is `'compact'`, then the default value for `OutlierSize` is 4.

Example: `'OutlierSize',8`

Data Types: `single` | `double`

`PlotStyle` — Plot style
`'traditional'` (default) | `'compact'`

Plot style, specified as the comma-separated pair consisting of `'PlotStyle'` and one of the following.

Name	Value
`'traditional'`	Plot boxes using a traditional box style.
`'compact'`	Plot boxes using a smaller box style designed for plots with many groups. This style changes the defaults for some other parameters.

Example: `'PlotStyle','compact'`

`Symbol` — Symbol and color for outliers
line specification

Symbol and color for outliers, specified as the comma-separated pair consisting of `'Symbol'` and a line specification. See the `LineSpec` parameter in `plot` for available line specifications.

If `'PlotStyle'` is `'traditional'`, then the default value is `'r+'`, which plots each outlier using a red `'+'` symbol.

If 'PlotStyle' is 'compact', then the default value is 'o', which plots each outlier using an 'o' symbol in the same color as the corresponding box.

If you omit the symbol, then the outliers appear invisible. If you omit the color, then the outliers appear in the same color as the box.

Example: 'kx'

Widths — Box width

numeric scalar | numeric vector

Box width, specified as the comma-separated pair consisting of 'Widths' and a numeric scalar or numeric vector. If the number of boxes is not equal to the number of width values specified, then the list of values is replicated or truncated as necessary.

This name-value pair argument does not alter the spacing between boxes. Therefore, if you specify a large value for 'Widths', the boxes might overlap.

The default box width is equal to half of the minimum separation between boxes, which is 0.5 when the 'Positions' name-value pair argument takes its default value.

Example: 'Widths',0.3

Data Types: single | double

Group Appearance

ColorGroup — Grouping variable for box color change

[] (default) | numeric vector | character array | string array | cell array | categorical array

Grouping variable for box color change, specified as the comma-separated pair consisting of 'ColorGroup' and a grouping variable. The grouping variable is a numeric vector, character array, string array, cell array, or categorical array. The box color changes when the specified grouping variable changes. The default value [] indicates that the box color does not change based on the group.

Data Types: single | double | char | string | cell | categorical

FactorDirection — Order of factors on plot

'data' (default) | 'list' | 'auto'

Order of factors on plot, specified as the comma-separated pair consisting of 'FactorDirection' and one of the following.

Name	Value
'data'	Factors appear with the first value next to the plot origin.
'list'	Factors appear left-to-right if on the x-axis, or top-to-bottom if on the y-axis.
'auto'	If the grouping variables are numeric, then boxplot uses 'data'. If the grouping variables are character arrays, string arrays, cell arrays, or categorical arrays, then boxplot uses 'list'.

`FullFactors` — Plot all group factors
`'off'` (default) | `'on'`

Plot all group factors, specified as the comma-separated pair consisting of `'FullFactors'` and either `'off'` or `'on'`. If `'off'`, then `boxplot` plots one box for each unique row of grouping variables. If `'on'`, then `boxplot` plots one box for each possible combination of grouping variable values, including combinations that do not appear in the data.

Example: `'FullFactors','on'`

`FactorGap` — Distance between different grouping factors
`[]` | positive numeric value | vector of positive numeric values | `'auto'`

Distance between different grouping factors, specified as the comma-separated pair consisting of `'FactorGap'` and a positive numeric value, a vector of positive numeric values, or `'auto'`. If you specify a vector, then the vector length must be less than or equal to the number of grouping variables.

`'FactorGap'` represents the distance of the gap between different factors of a grouping variable, expressed as a percentage of the width of the plot. For example, if you specify `[3,1]`, then the gap is three percent of the width of the plot between groups with different values of the first grouping variable, and one percent between groups with the same value of the first grouping variable but different values for the second.

If you specify `'auto'`, then `boxplot` selects a gap distance automatically. The value `[]` indicates no change in gap size between different factors.

If `'PlotStyle'` is `'traditional'`, then the default value for `FactorGap` is `[]`. If `'PlotStyle'` is `'compact'`, then the default value is `'auto'`.

Example: `'FactorGap',[3,1]`

Data Types: `single` | `double` | `char` | `string`

`FactorSeparator` — Separation between grouping factors
`[]` | positive integer | vector of positive integers | `'auto'`

Separation between grouping factors, specified as the comma-separated pair consisting of `'FactorSeparator'` and a positive integer or a vector of positive integers, or `'auto'`. If you specify a vector, then the length of the vector should be less than or equal to the number of grouping variables. The integer values must be in the range [1,*G*], where *G* is the number of grouping variables.

`'FactorSeparator'` specifies which factors should have their values separated by a grid line. For example, `[1,2]` adds a separator line when the first or second grouping variable changes value.

If `'PlotStyle'` is `'traditional'`, then the default value for `FactorSeparator` is `[]`. If `'PlotStyle'` is `'compact'`, then the default value is `'auto'`.

Example: `'FactorSeparator',[1,2]`

Data Types: `single` | `double` | `char` | `string`

`GroupOrder` — Plotting order of groups
`[]` (default) | string array | cell array

Plotting order of groups, specified as the comma-separated pair consisting of `'GroupOrder'` and a string array or cell array containing the names of the grouping variables. If you have multiple

grouping variables, separate values with a comma. You can also use categorical arrays as grouping variables to control the order of the boxes. The default value `[]` does not reorder the boxes.

Data Types: `string` | `cell`

Data Limits and Maximum Distances

`DataLim` — Extreme data limits

`[-Inf,Inf]` (default) | two-element numeric vector

Extreme data limits, specified as the comma-separated pair consisting of `'DataLim'` and a two-element numeric vector containing the lower and upper limits, respectively. The values specified for `'DataLim'` are used by `'ExtremeMode'` to determine which data points are extreme.

Data Types: `single` | `double`

`ExtremeMode` — Handling method for extreme data

`'clip'` (default) | `'compress'`

Handling method for extreme data, specified as the comma-separated pair consisting of `'ExtremeMode'` and one of the following.

Name	Value
`'clip'`	If any data values fall outside the limits specified by `'DataLim'`, then `boxplot` displays these values at `DataLim` on the plot.
`'compress'`	If any data values fall outside the limits specified by `'DataLim'` , then `boxplot` displays these values evenly distributed in a region just outside `DataLim`, retaining the relative order of the points.

If any data points lie outside the limit specified by `'DataLim'`, then the limit is marked with a dotted line. If any data points are compressed, then two gray lines mark the compression region. Values at `–Inf` or `Inf` can be clipped or compressed, but `NaN` values do not appear on the plot. Box notches are drawn to scale and may extend beyond the bounds if the median is inside the limit. Box notches are not drawn if the median is outside the limits.

Example: `'ExtremeMode','compress'`

`Jitter` — Maximum outlier displacement distance

numeric value

Maximum outlier displacement distance, specified as the comma-separated pair consisting of `'Jitter'` and a numeric value. `Jitter` is the maximum distance to displace outliers along the factor axis by a uniform random amount, in order to make duplicate points visible. If you specify `'Jitter'` equal to 1, then the jitter regions just touch between the closest adjacent groups.

If `'PlotStyle'` is `'traditional'`, then the default value for `Jitter` is 0. If `'PlotStyle'` is `'compact'`, then the default value is 0.5.

Example: `'Jitter',1`

Data Types: `single` | `double`

`Whisker` — Multiplier for maximum whisker length

1.5 (default) | positive numeric value

Multiplier for the maximum whisker length, specified as the comma-separated pair consisting of `'Whisker'` and a positive numeric value. The maximum whisker length is the product of `Whisker` and the interquartile range.

`boxplot` draws points as outliers if they are greater than $q_3 + w \times (q_3 - q_1)$ or less than $q_1 - w \times (q_3 - q_1)$, where w is the multiplier `Whisker`, and q_1 and q_3 are the 25th and 75th percentiles of the sample data, respectively.

The default value for `'Whisker'` corresponds to approximately +/–2.7σ and 99.3 percent coverage if the data are normally distributed. The plotted whisker extends to the *adjacent value*, which is the most extreme data value that is not an outlier.

Specify `'Whisker'` as 0 to give no whiskers and to make every point outside of q_1 and q_3 an outlier.

Example: `'Whisker',0`

Data Types: `single` | `double`

Plot Appearance

`Labels` — Box labels

character array | string array | cell array | numeric vector | numeric matrix

Box labels, specified as the comma-separated pair consisting of `'Labels'` and a character array, string array, cell array, or numeric vector containing the box label names. Specify one label per `x` value or one label per group. To specify multiple label variables, use a numeric matrix or a cell array containing any of the accepted data types.

To remove labels from a plot , use the following command: `set(gca,'XTickLabel',{' '})`.

Data Types: `char` | `string` | `cell` | `single` | `double`

`LabelOrientation` — Label orientation

`'inline'` | `'horizontal'`

Label orientation, specified as the comma-separated pair consisting of `'LabelOrientation'` and one of the following.

Name	Value
`'inline'`	Rotate box labels to be vertical. This is the default when `'PlotStyle'` is `'compact'`.
`'horizontal'`	Leave box labels horizontal. This is the default when `'PlotStyle'` is `'traditional'`.

If the labels are on the *y* axis, then both settings leave the labels horizontal.

Example: `'LabelOrientation','inline'`

`LabelVerbosity` — Labels to display on plot

`'all'` | `'minor'` | `'majorminor'`

Labels to display on plot, specified as the comma-separated pair consisting of `'LabelVerbosity'` and one of the following.

Name	Value
`'all'`	Display a label for every value of a grouping variable. This is the default when `'PlotStyle'` is `'traditional'`.
`'minor'`	For any grouping variable, display the value corresponding to box `j` only if that value differs from the value corresponding to box `(j − 1)`.
`'majorminor'`	For any grouping variable `g(:,i)`, display the value corresponding to box `j`, only if that value differs from the value of `g(:,i)` corresponding to box `(j − 1)`, or if the condition above holds for at least one of the grouping variables `g(:,1),…,g(:,i-1)`. This is the default when `'PlotStyle'` is `'compact'`.

Example: `'LabelVerbosity','minor'`

`Orientation` — Plot orientation
`'vertical'` (default) | `'horizontal'`

Plot orientation, specified as the comma-separated pair consisting of `'Orientation'` and one of the following.

Name	Value
`'vertical'`	Plot `x` on the *y*-axis.
`'horizontal'`	Plot `x` on the *x*-axis.

Example: `'horizontal'`

`Positions` — Box positions
numeric vector

Box positions, specified as the comma-separated pair consisting of `'Positions'` and a numeric vector containing one entry for each group or `x` value. The default is 1:*NumGroups*, where *NumGroups* is the number of groups.

Data Types: `single` | `double`

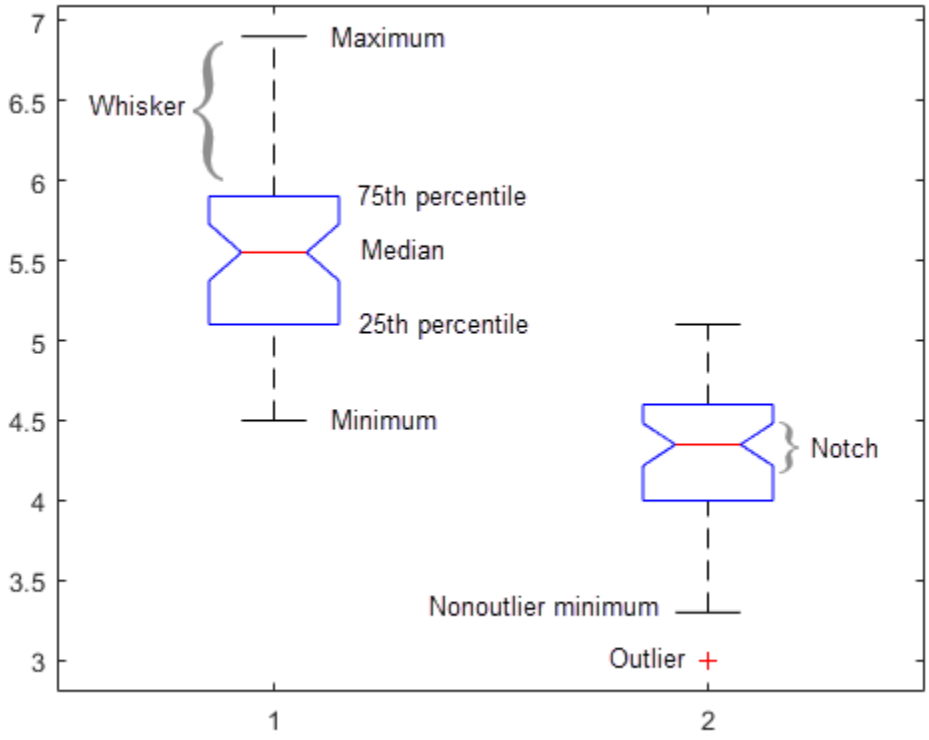
More About

Box Plot

A box plot provides a visualization of summary statistics for sample data and contains the following features:

- The bottom and top of each box are the 25th and 75th percentiles of the sample, respectively. The distance between the bottom and top of each box is the interquartile range.
- The red line in the middle of each box is the sample median. If the median is not centered in the box, the plot shows sample skewness.
- The whiskers are lines extending above and below each box. Whiskers go from the end of the interquartile range to the furthest observation within the whisker length (the *adjacent value*).

- Observations beyond the whisker length are marked as outliers. By default, an outlier is a value that is more than 1.5 times the interquartile range away from the bottom or top of the box. However, you can adjust this value by using additional input arguments. An outlier appears as a red + sign.
- Notches display the variability of the median between samples. The width of a notch is computed so that boxes whose notches do not overlap have different medians at the 5% significance level. The significance level is based on a normal distribution assumption, but comparisons of medians are reasonably robust for other distributions. Comparing box plot medians is like a visual hypothesis test, analogous to the *t* test used for means.



Tips

- `boxplot` creates a visual representation of the data, but does not return numeric values. To calculate the relevant summary statistics for the sample data, use the following functions:
  - `min` — Find the minimum value in the sample data.
  - `max` — Find the maximum value in the sample data.
  - `median` — Find the median value in the sample data.
  - `quantile` — Find the quantile values in the sample data. For example, to compute the 25th and 75th percentiles of `x`, specify `quantile(x,[0.25 0.75])`. For more information on how the percentiles are computed, see “Algorithms” on page 33-5103.
  - `iqr` — Find the interquartile range in the sample data.

  - `grpstats` — Calculate summary statistics for the sample data, organized by group.
- You can see data values and group names using the data cursor in the figure window. The cursor shows the original values of any points affected by the `datalim` parameter. You can label the group to which an outlier belongs using the `gname` function.
- To modify graphics properties of a box plot component, use `findobj` with the `Tag` property to find the component's handle. `Tag` values for box plot components depend on parameter settings, and are listed in the following table.

Parameter Settings	Tag Values
All settings	• `'Box'` • `'Outliers'`
When `'PlotStyle'` is `'traditional'`	• `'Median'` • `'Upper Whisker'` • `'Lower Whisker'` • `'Upper Adjacent Value'` • `'Lower Adjacent Value'`
When `'PlotStyle'` is `'compact'`	• `'Whisker'` • `'MedianOuter'` • `'MedianInner'`
When `'Notch'` is `'marker'`	• `'NotchLo'` • `'NotchHi'`

Alternative Functionality

You can also create a `BoxChart` object by using the `boxchart` function. Although `boxchart` does not include all the functionality of `boxplot`, it has some advantages. Unlike `boxplot`, the `boxchart` function:

- Allows for categorical rulers along the group axis
- Provides the option of a legend
- Works well with the `hold on` command
- Has an improved visual design that helps you see notches more easily

To control the appearance and behavior of the object, change the BoxChart Properties.

References

[1] McGill, R., J. W. Tukey, and W. A. Larsen. “Variations of Boxplots.” *The American Statistician*. Vol. 32, No. 1, 1978, pp. 12–16.

[2] Velleman, P.F., and D.C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis*. Pacific Grove, CA: Duxbury Press, 1981.

[3] Nelson, L. S. “Evaluating Overlapping Confidence Intervals.” *Journal of Quality Technology*. Vol. 21, 1989, pp. 140–141.

[4] Langford, E. “Quartiles in Elementary Statistics”, *Journal of Statistics Education*. Vol. 14, No. 3, 2006.

See Also

`anova1` | `grpstats` | `kruskalwallis` | `max` | `median` | `min` | `multcompare` | `quantile`

Topics

“Compare Grouped Data Using Box Plots” on page 4-4
“Grouping Variables” on page 2-45

Introduced before R2006a

boundary

Piecewise distribution boundaries

Syntax

```
[p,q] = boundary(pd)
[p,q] = boundary(pd,j)
```

Description

`[p,q] = boundary(pd)` returns the boundary points between segments in `pd`, the piecewise distribution. `p` is a vector of the cumulative probabilities at the boundaries, and `q` is a vector of the corresponding quantiles.

`[p,q] = boundary(pd,j)` returns boundary values of the `j`th boundary.

Examples

Boundaries in `paretotails` Object

Generate a sample data set and create a `paretotails` object by fitting a piecewise distribution with Pareto tails to the generated data. Find the boundary points between segments in a `paretotails` object by using the object function `boundary`.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and generalized Pareto distributions (GPDs) for the lower and upper 10% of the data set.

```
pd = paretotails(x,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.33251    (0 < p < 0.1): lower tail, GPD(-0.0063504,0.567017)
   -1.33251 < x < 1.80149  (0.1 < p < 0.9): interpolated empirical cdf
        1.80149 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.24874,3.00974)
```

Return the boundary values between the piecewise segments by using the `boundary` function.

```
[p,q] = boundary(pd)
```

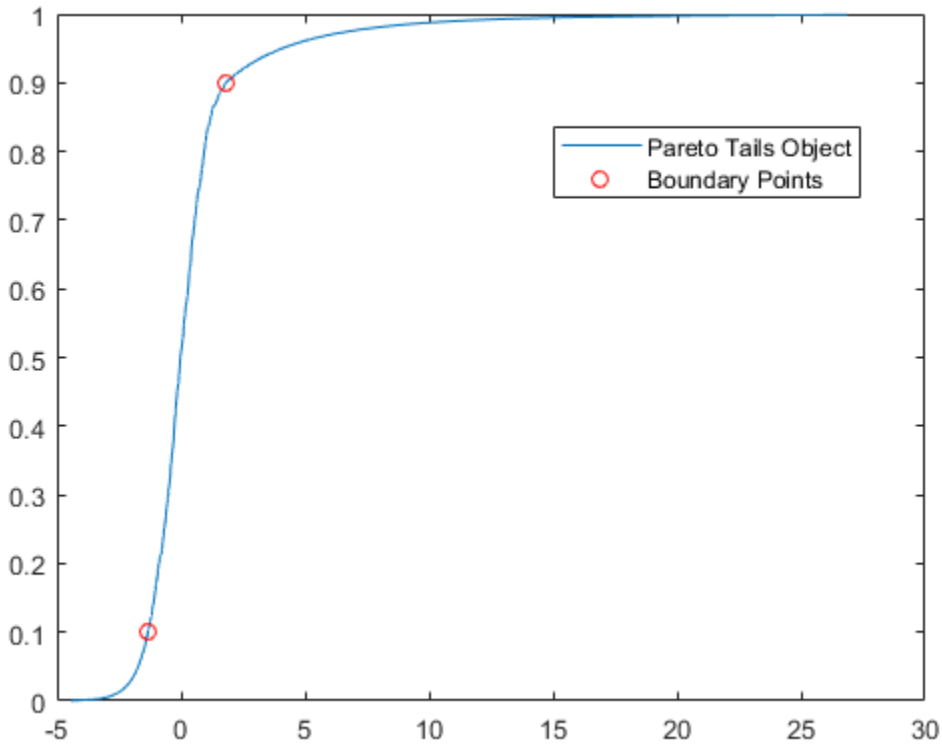
```
p = 2×1
```

```
    0.1000
    0.9000


q = 2×1

   -1.3325
    1.8015
```

The values in `p` are the cumulative probabilities at the boundaries, and the values in `q` are the corresponding quantiles.

Plot the cdf of the `paretotails` object and mark the boundary points on the figure.

```
xi = sort(x);
plot(xi,cdf(pd,xi))
hold on
plot(q,p,'ro')
legend('Pareto Tails Object','Boundary Points','Location','best')
hold off
```



Input Arguments

pd — Piecewise distribution with Pareto tails

`paretotails` object

Piecewise distribution with Pareto tails, specified as a `paretotails` object.

`j` — Boundary index
positive integer

Boundary index indicating which boundary to return, specified as a positive integer.

Data Types: `single` | `double`

Output Arguments

`p` — Cumulative probability at boundary
numeric vector of range `(0,1)` values

Cumulative probability at each boundary, returned as a numeric vector of range `(0,1)` values.

`q` — Quantile at boundary
numeric vector

Quantile at each boundary, returned as a numeric vector.

See Also

`lowerparams` | `nsegments` | `paretotails` | `segment` | `upperparams`

Topics
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

CalinskiHarabaszEvaluation

Package: `clustering.evaluation`
Superclasses: `ClusterCriterion`

Calinski-Harabasz criterion clustering evaluation object

Description

`CalinskiHarabaszEvaluation` is an object consisting of sample data, clustering data, and Calinski-Harabasz criterion values used to evaluate the optimal number of clusters. Create a Calinski-Harabasz criterion clustering evaluation object using `evalclusters`.

Construction

`eva = evalclusters(x,clust,'CalinskiHarabasz')` creates a Calinski-Harabasz criterion clustering evaluation object.

`eva = evalclusters(x,clust,'CalinskiHarabasz',Name,Value)` creates a Calinski-Harabasz criterion clustering evaluation object using additional options specified by one or more name-value pair arguments.

Input Arguments

x — Input data

matrix

Input data, specified as an *N*-by-*P* matrix. *N* is the number of observations, and *P* is the number of variables.

Data Types: `single` | `double`

clust — Clustering algorithm

`'kmeans'` | `'linkage'` | `'gmdistribution'` | matrix of clustering solutions | function handle

Clustering algorithm, specified as one of the following.

`'kmeans'`	Cluster the data in `x` using the `kmeans` clustering algorithm, with `'EmptyAction'` set to `'singleton'` and `'Replicates'` set to `5`.
`'linkage'`	Cluster the data in `x` using the `clusterdata` agglomerative clustering algorithm, with `'Linkage'` set to `'ward'`.
`'gmdistribution'`	Cluster the data in `x` using the `gmdistribution` Gaussian mixture distribution algorithm, with `'SharedCov'` set to `true` and `'Replicates'` set to `5`.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can specify a clustering algorithm using a function handle. The function must be of the form `C = clustfun(DATA,K)`, where `DATA` is the data to be clustered, and `K` is the number of clusters. The output of `clustfun` must be one of the following:

- A vector of integers representing the cluster index for each observation in `DATA`. There must be `K` unique values in this vector.
- A numeric *n*-by-*K* matrix of score for *n* observations and *K* classes. In this case, the cluster index for each observation is determined by taking the largest score value in each row.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can also specify `clust` as a *n*-by-*K* matrix containing the proposed clustering solutions. *n* is the number of observations in the sample data, and *K* is the number of proposed clustering solutions. Column *j* contains the cluster indices for each of the *N* points in the *j*th clustering solution.

Data Types: `single` | `double` | `char` | `string` | `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KList',[1:6]` specifies to test 1, 2, 3, 4, 5, and 6 clusters to find the optimal number.

`KList` — List of number of clusters to evaluate
vector

List of number of clusters to evaluate, specified as the comma-separated pair consisting of `'KList'` and a vector of positive integer values. You must specify `KList` when `clust` is a clustering algorithm name or a function handle. When `criterion` is `'gap'`, `clust` must be a character vector, a string scalar, or a function handle, and you must specify `KList`.

Example: `'KList',[1:6]`

Data Types: `single` | `double`

Properties

`ClusteringFunction`

Clustering algorithm used to cluster the input data, stored as a valid clustering algorithm name or function handle. If the clustering solutions are provided in the input, `ClusteringFunction` is empty.

`CriterionName`

Name of the criterion used for clustering evaluation, stored as a valid criterion name.

`CriterionValues`

Criterion values corresponding to each proposed number of clusters in `InspectedK`, stored as a vector of numerical values.

`InspectedK`

List of the number of proposed clusters for which to compute criterion values, stored as a vector of positive integer values.

`Missing`

Logical flag for excluded data, stored as a column vector of logical values. If `Missing` equals `true`, then the corresponding value in the data matrix `X` is not used in the clustering solution.

`NumObservations`

Number of observations in the data matrix `X`, minus the number of missing (`NaN`) values in `X`, stored as a positive integer value.

`OptimalK`

Optimal number of clusters, stored as a positive integer value.

`OptimalY`

Optimal clustering solution corresponding to `OptimalK`, stored as a column vector of positive integer values. If the clustering solutions are provided in the input, `OptimalY` is empty.

`X`

Data used for clustering, stored as a matrix of numerical values.

Methods

Inherited Methods

addK	Evaluate additional numbers of clusters
plot	Plot clustering evaluation object criterion values
compact	Compact clustering evaluation object

Examples

Evaluate the Clustering Solution Using Calinski-Harabasz Criterion

Evaluate the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion.

Load the sample data.

```
load fisheriris;
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Evaluate the optimal number of clusters using the Calinski-Harabasz criterion. Cluster the data using `kmeans`.

```
rng('default');  % For reproducibility
eva = evalclusters(meas,'kmeans','CalinskiHarabasz','KList',[1:6])
```

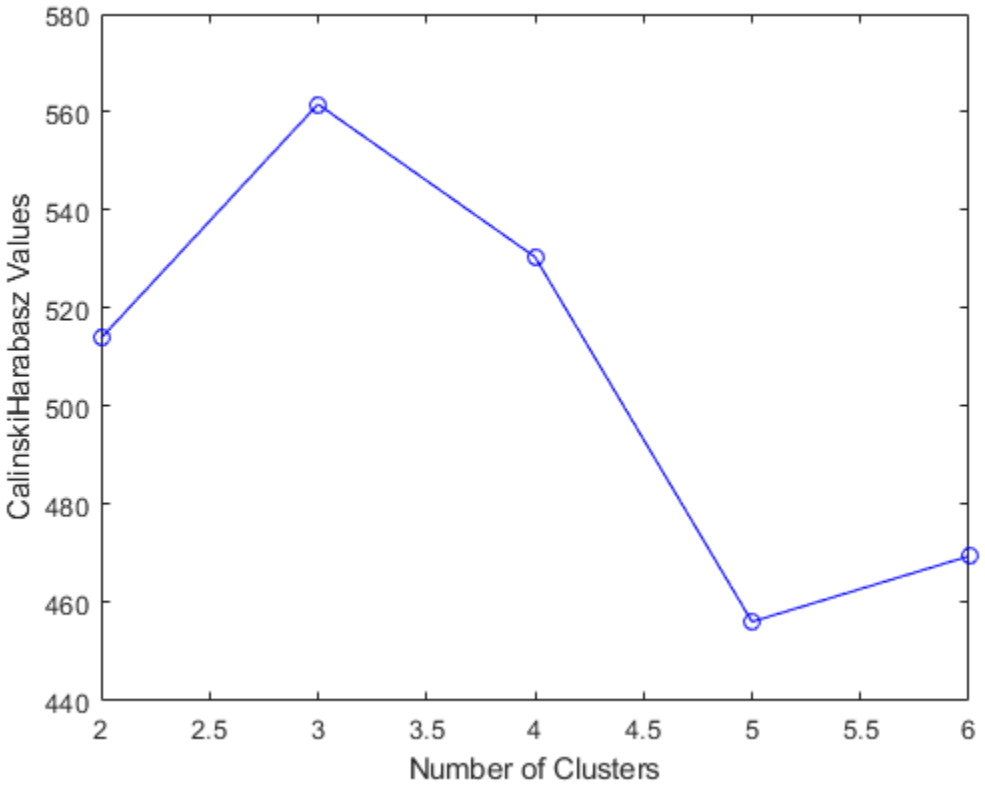
```
eva =
  CalinskiHarabaszEvaluation with properties:
```

```
    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [Inf 513.9245 561.6278 530.4871 456.1279 469.5068]
           OptimalK: 1
```

The `OptimalK` value indicates that, based on the Calinski-Harabasz criterion, the optimal number of clusters is three.

Plot the Calinski-Harabasz criterion values for each number of clusters tested.

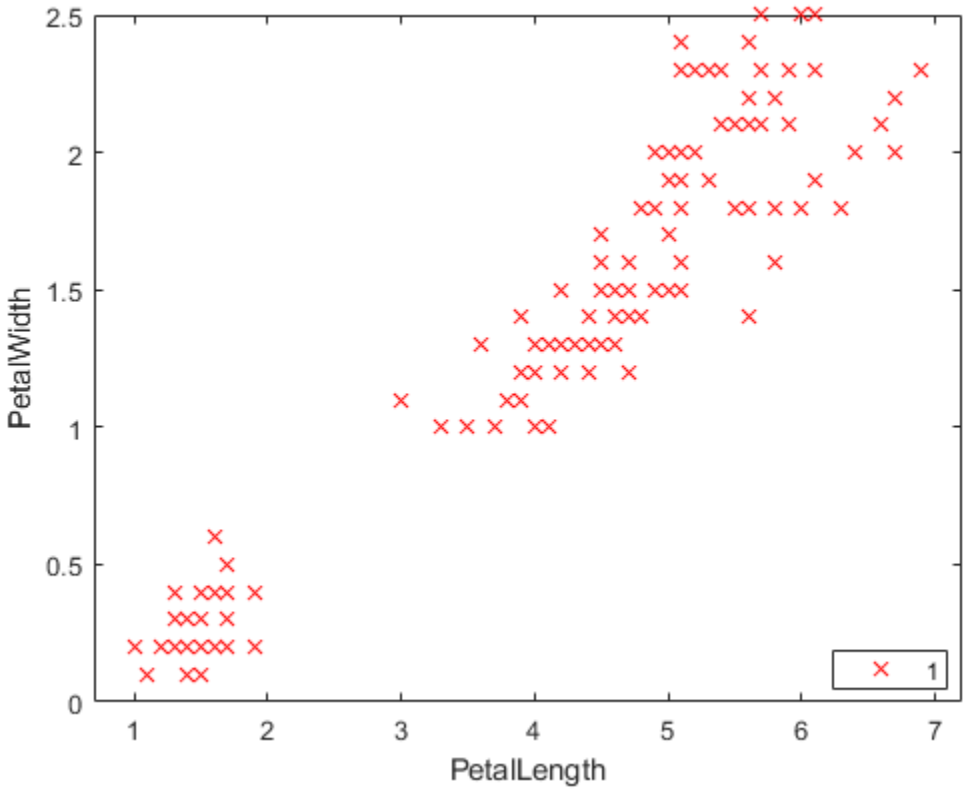
```
figure;
plot(eva);
```



The plot shows that the highest Calinski-Harabasz value occurs at three clusters, suggesting that the optimal number of clusters is three.

Create a grouped scatter plot to examine the relationship between petal length and width. Group the data by suggested clusters.

```
PetalLength = meas(:,3);
PetalWidth = meas(:,4);
ClusterGroup = eva.OptimalY;
figure;
gscatter(PetalLength,PetalWidth,ClusterGroup,'rbg','xod');
```



The plot shows cluster 3 in the lower-left corner, completely separated from the other two clusters. Cluster 3 contains flowers with the smallest petal widths and lengths. Cluster 1 is in the upper-right corner, and contains flowers with the largest petal widths and lengths. Cluster 2 is near the center of the plot, and contains flowers with measurements between these two extremes.

More About

Calinski-Harabasz Criterion

The Calinski-Harabasz criterion is sometimes called the variance ratio criterion (VRC). The Calinski-Harabasz index is defined as

$$VRC_k = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)},$$

where SS_B is the overall between-cluster variance, SS_W is the overall within-cluster variance, k is the number of clusters, and N is the number of observations.

The overall between-cluster variance SS_B is defined as

$$SS_B = \sum_{i=1}^{k} n_i \|m_i - m\|^2,$$

where k is the number of clusters, n_i is the number of observations in cluster i, m_i is the centroid of cluster i, m is the overall mean of the sample data, and $\|m_i - m\|$ is the L^2 norm (Euclidean distance) between the two vectors.

The overall within-cluster variance SS_W is defined as

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|^2,$$

where k is the number of clusters, x is a data point, c_i is the ith cluster, m_i is the centroid of cluster i, and $\|x - m_i\|$ is the L^2 norm (Euclidean distance) between the two vectors.

Well-defined clusters have a large between-cluster variance (SS_B) and a small within-cluster variance (SS_W). The larger the VRC_k ratio, the better the data partition. To determine the optimal number of clusters, maximize VRC_k with respect to k. The optimal number of clusters is the solution with the highest Calinski-Harabasz index value.

The Calinski-Harabasz criterion is best suited for k-means clustering solutions with squared Euclidean distances.

References

[1] Calinski, T., and J. Harabasz. "A dendrite method for cluster analysis." *Communications in Statistics*. Vol. 3, No. 1, 1974, pp. 1–27.

See Also

`DaviesBouldinEvaluation` | `GapEvaluation` | `SilhouetteEvaluation` | `evalclusters`

Topics
Class Attributes
Property Attributes

candexch

D-optimal design from candidate set using row exchanges

Syntax

```
rlist = candexch(C,nrows)
rlist = candexch(C,nrows,Name,Value)
```

Description

`rlist = candexch(C,nrows)` uses a row-exchange algorithm to select a *D*-optimal design from the candidate set `C`.

`rlist = candexch(C,nrows,Name,Value)` generates a *D*-optimal design with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`C`

`N`-by-`P` matrix containing the values of `P` model terms at each of `N` points.

Default:

`nrows`

The desired number of rows in the design.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`display`

When `'on'`, displays iteration number. Disable the display by setting to `'off'`.

Default: `'on'`, except when the `UseParallel` option is `true`

`init`

`nrows`-by-`P` matrix giving an initial design.

Default: A random subset of the rows of `C`

`maxiter`

Maximum number of iterations, a positive integer.

Default: `10`

`options`

A structure that specifies whether to run in parallel, and specifies the random stream or streams. This option requires Parallel Computing Toolbox.

Create the `options` structure with `statset`. Option fields:

- `UseParallel` — Set to `true` to compute in parallel. Default is `false`.
- `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. Default is `false`. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
- `Streams` — A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `candexch` uses the default stream or streams. If you choose to specify `Streams`, use a single object except in the case

  - `UseParallel` is `true`
  - `UseSubstreams` is `false`

  In that case, use a cell array the same size as the Parallel pool.

Default: `[]`

`start`

An `nobs`-by-*p* matrix of factor settings, specifying a set of `nobs` fixed design points to include in the design. `candexch` finds `nrows` additional rows to add to the `start` design. The parameter provides the same functionality as the `daugment` function, using a row-exchange algorithm rather than a coordinate-exchange algorithm.

Default: `[]`

`tries`

Number of times to try to generate a design from a new starting point. The algorithm uses random points for each try, except possibly the first.

Default: 1

Output Arguments

`rlist`

Vector of length `nrows` listing the selected rows.

Examples

This example shows how to generate a *D*-optimal design when there is a restriction on the candidate set, so the `rowexch` function isn't appropriate.

```
F = (fullfact([5 5 5])-1)/4; % factor settings in unit cube
T = sum(F,2)<=1.51;         % find rows matching a restriction
F = F(T,:);                 % take only those rows
C = [ones(size(F,1),1) F F.^2];
```

```
                            % compute model terms including
                            % a constant and all squared terms
R = candexch(C,12);         % find a D-optimal 12-point subset
X = F(R,:);                 % get factor settings
```

Algorithms

`candexch` selects a starting design `X` at random, and uses a row-exchange algorithm to iteratively replace rows of `X` by rows of `C` in an attempt to improve the determinant of `X'*X`.

Alternatives

The `rowexch` function also generates *D*-optimal designs using a row-exchange algorithm, but it automatically generates a candidate set that is appropriate for a specified model. The `daugment` function augments a set of fixed design points using a coordinate-exchange algorithm; the `'start'` parameter provides the same functionality using the row exchange algorithm.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`candgen` | `cordexch` | `daugment` | `rowexch` | `x2fx`

Topics

“Specify Candidate Sets” on page 28-16
“D-Optimal Designs” on page 28-12

Introduced before R2006a

candgen

Candidate set generation

Syntax

```
dC = candgen(nfactors,'model')
[dC,C] = candgen(nfactors,'model')
[...] = candgen(nfactors,'model','Name',value)
```

Description

`dC = candgen(nfactors,'model')` generates a candidate set `dC` of treatments appropriate for estimating the parameters in the *`model`* with `nfactors` factors. `dC` has `nfactors` columns and one row for each candidate treatment. *`model`* is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

Alternatively, *`model`* can be a matrix specifying polynomial terms of arbitrary order. In this case, *`model`* should have one column for each factor and one row for each term in the model. The entries in any row of *`model`* are powers for the factors in the columns. For example, if a model has factors `X1`, `X2`, and `X3`, then a row `[0 1 2]` in *`model`* specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in *`model`* specifies a constant term, which can be omitted.

`[dC,C] = candgen(nfactors,'model')` also returns the design matrix `C` evaluated at the treatments in `dC`. The order of the columns of `C` for a full quadratic model with *n* terms is:

1. The constant term
2. The linear terms in order 1, 2, ..., *n*
3. The interaction terms in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n* – 1, *n*)
4. The squared terms in order 1, 2, ..., *n*

Other models use a subset of these terms, in the same order.

Pass `C` to `candexch` to generate a *D*-optimal design using a coordinate-exchange algorithm.

`[...] = candgen(nfactors,'model','Name',value)` specifies one or more optional name/value pairs for the design. Valid parameters and their values are listed in the following table. Specify *`Name`* inside single quotes.

Name	Value
`bounds`	Lower and upper bounds for each factor, specified as a 2-by-`nfactors` matrix. Alternatively, this value can be a cell array containing `nfactors` elements, each element specifying the vector of allowable values for the corresponding factor.

Name	Value
`categorical`	Indices of categorical predictors.
`levels`	Vector of number of levels for each factor.

Note The `rowexch` function automatically generates a candidate set using `candgen`, and then creates a *D*-optimal design from that candidate set using `candexch`. Call `candexch` separately to specify your own candidate set to the row-exchange algorithm.

Examples

The following example uses `rowexch` to generate a five-run design for a two-factor pure quadratic model using a candidate set that is produced internally:

```
dRE1 = rowexch(2,5,'purequadratic','tries',10)
dRE1 =
    -1     1
     0     0
     1    -1
     1     0
     1     1
```

The same thing can be done using `candgen` and `candexch` in sequence:

```
[dC,C] = candgen(2,'purequadratic') % Candidate set, C
dC =
    -1    -1
     0    -1
     1    -1
    -1     0
     0     0
     1     0
    -1     1
     0     1
     1     1
C =
     1    -1    -1     1     1
     1     0    -1     0     1
     1     1    -1     1     1
     1    -1     0     1     0
     1     0     0     0     0
     1     1     0     1     0
     1    -1     1     1     1
     1     0     1     0     1
     1     1     1     1     1
treatments = candexch(C,5,'tries',10) % Find D-opt subset
treatments =
     2
     1
     7
     3
     4
dRE2 = dC(treatments,:) % Display design
dRE2 =
     0    -1
```

```
    -1    -1
    -1     1
     1    -1
    -1     0
```

See Also

candexch | rowexch

Introduced before R2006a

canoncorr

Canonical correlation

Syntax

```
[A,B] = canoncorr(X,Y)
[A,B,r] = canoncorr(X,Y)
[A,B,r,U,V] = canoncorr(X,Y)
[A,B,r,U,V,stats] = canoncorr(X,Y)
```

Description

`[A,B] = canoncorr(X,Y)` computes the sample canonical coefficients for the data matrices `X` and `Y`.

`[A,B,r] = canoncorr(X,Y)` also returns `r`, a vector of the sample canonical correlations.

`[A,B,r,U,V] = canoncorr(X,Y)` also returns `U` and `V`, matrices of the canonical scores for `X` and `Y`, respectively.

`[A,B,r,U,V,stats] = canoncorr(X,Y)` also returns `stats`, a structure containing information related to testing the sequence of hypotheses that the remaining correlations are all zero.

Examples

Compute Sample Canonical Correlation

Perform canonical correlation analysis for a sample data set.

The data set `carbig` contains measurements for 406 cars from the years 1970 to 1982.

Load the sample data.

```
load carbig;
data = [Displacement Horsepower Weight Acceleration MPG];
```

Define X as the matrix of displacement, horsepower, and weight observations, and `Y` as the matrix of acceleration and MPG observations. Omit rows with insufficient data.

```
nans = sum(isnan(data),2) > 0;
X = data(~nans,1:3);
Y = data(~nans,4:5);
```

Compute the sample canonical correlation.

```
[A,B,r,U,V] = canoncorr(X,Y);
```

View the output of `A` to determine the linear combinations of displacement, horsepower, and weight that make up the canonical variables of `X`.

```
A
```

```
A = 3×2

    0.0025    0.0048
    0.0202    0.0409
   -0.0000   -0.0027
```

`A(3,1)` is displayed as `-0.000` because it is very small. Display `A(3,1)` separately.

```
A(3,1)
```

```
ans = -2.4737e-05
```

The first canonical variable of `X` is `u1 = 0.0025*Disp + 0.0202*HP - 0.000025*Wgt`.

The second canonical variable of `X` is `u2 = 0.0048*Disp + 0.0409*HP - 0.0027*Wgt`.

View the output of `B` to determine the linear combinations of acceleration and MPG that make up the canonical variables of `Y`.

```
B
```

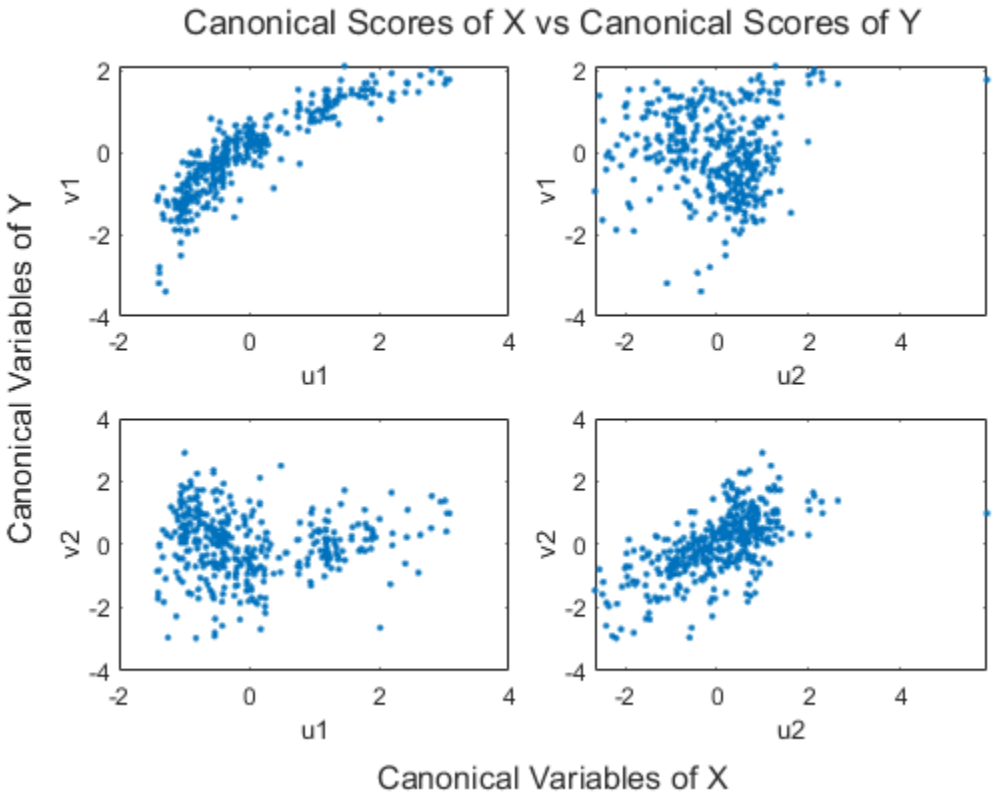
```
B = 2×2

   -0.1666   -0.3637
   -0.0916    0.1078
```

The first canonical variable of `Y` is `v1 = -0.1666*Accel - 0.0916*MPG`.

The second canonical variable of `Y` is `v2 = -0.3637*Accel + 0.1078*MPG`.

Plot the scores of the canonical variables of `X` and `Y` against each other.

```
t = tiledlayout(2,2);
title(t,'Canonical Scores of X vs Canonical Scores of Y')
xlabel(t,'Canonical Variables of X')
ylabel(t,'Canonical Variables of Y')
t.TileSpacing = 'compact';

nexttile
plot(U(:,1),V(:,1),'.')
xlabel('u1')
ylabel('v1')

nexttile
plot(U(:,2),V(:,1),'.')
xlabel('u2')
ylabel('v1')

nexttile
plot(U(:,1),V(:,2),'.')
xlabel('u1')
ylabel('v2')

nexttile
plot(U(:,2),V(:,2),'.')
xlabel('u2')
ylabel('v2')
```



The pairs of canonical variables $\{u_i, v_i\}$ are ordered from the strongest to weakest correlation, with all other pairs independent.

Return the correlation coefficient of the variables `u1` and `v1`.

```
r(1)
```

```
ans = 0.8782
```

Input Arguments

X — Input matrix

matrix

Input matrix, specified as an n-by-d_1 matrix. The rows of `X` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

Y — Input matrix

matrix

Input matrix, specified as an n-by-d_2 matrix where `X` is an n-by-d_1 matrix. The rows of `Y` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

Output Arguments

A — Sample canonical coefficients for X variables

matrix

Sample canonical coefficients for the variables in X, returned as a d_1-by-d matrix, where d = min(rank(X),rank(Y)).

The jth column of A contains the linear combination of variables that makes up the jth canonical variable for X.

If X is less than full rank, `canoncorr` gives a warning and returns zeros in the rows of A corresponding to dependent columns of X.

B — Sample canonical coefficients for Y variables

matrix

Sample canonical coefficients for the variables in Y, returned as a d_2-by-d matrix, where d = min(rank(X),rank(Y)).

The jth column of B contains the linear combination of variables that makes up the jth canonical variable for Y.

If Y is less than full rank, `canoncorr` gives a warning and returns zeros in the rows of B corresponding to dependent columns of Y.

r — Sample canonical correlations

vector

Sample canonical correlations, returned as a 1-by-d vector, where d = min(rank(X),rank(Y)).

The jth element of `r` is the correlation between the jth columns of U and V.

U — Canonical scores for the X variables

matrix

Canonical scores for the variables in X, returned as an n-by-d matrix, where X is an n-by-d_1 matrix and d = min(rank(X),rank(Y)).

V — Canonical scores for the Y variables

matrix

Canonical scores for the variables in Y, returned as an n-by-d matrix, where Y is an n-by-d_2 matrix and d = min(rank(X),rank(Y)).

stats — Hypothesis test information

structure

Hypothesis test information, returned as a structure. This information relates to the sequence of d null hypotheses $H_0^{(k)}$ that the (k+1)st through dth correlations are all zero for k=1,...,d-1, and d = min(rank(X),rank(Y)).

The fields of `stats` are 1-by-d vectors with elements corresponding to the values of k.

Field	Description
`Wilks`	Wilks' lambda (likelihood ratio) statistic
`df1`	Degrees of freedom for the chi-squared statistic, and the numerator degrees of freedom for the F statistic
`df2`	Denominator degrees of freedom for the F statistic
`F`	Rao's approximate F statistic for $H_0^{(k)}$
`pF`	Right-tail significance level for `F`
`chisq`	Bartlett's approximate chi-squared statistic for $H_0^{(k)}$ with Lawley's modification
`pChisq`	Right-tail significance level for `chisq`

`stats` has two other fields (`dfe` and `p`), which are equal to `df1` and `pChisq`, respectively, and exist for historical reasons.

Data Types: `struct`

More About

Canonical Correlation Analysis

The canonical scores of the data matrices X and Y are defined as

$$U_i = Xa_i$$
$$V_i = Yb_i$$

where a_i and b_i maximize the Pearson correlation coefficient $\rho(U_i,V_i)$ subject to being uncorrelated to all previous canonical scores and scaled so that U_i and V_i have zero mean and unit variance.

The canonical coefficients of X and Y are the matrices A and B with columns a_i and b_i, respectively.

The canonical variables of X and Y are the linear combinations of the columns of X and Y given by the canonical coefficients in A and B respectively.

The canonical correlations are the values $\rho(U_i,V_i)$ measuring the correlation of each pair of canonical variables of X and Y.

Algorithms

`canoncorr` computes `A`, `B`, and `r` using `qr` and `svd`. `canoncorr` computes `U` and `V` as `U = (X–mean(X))*A` and `V = (Y–mean(Y))*B`.

References

[1] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[2] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

See Also

`manova1` | `pca`

Introduced before R2006a

capability

Process capability indices

Syntax

```
S = capability(data,specs)
```

Description

`S = capability(data,specs)` estimates capability indices for measurements in `data` given the specifications in `specs`. `data` can be either a vector or a matrix of measurements. If `data` is a matrix, indices are computed for the columns. `specs` can be either a two-element vector of the form `[L,U]` containing lower and upper specification limits, or (if `data` is a matrix) a two-row matrix with the same number of columns as `data`. If there is no lower bound, use `-Inf` as the first element of `specs`. If there is no upper bound, use `Inf` as the second element of `specs`.

The output `S` is a structure with the following fields:

- `mu` — Sample mean
- `sigma` — Sample standard deviation
- `P` — Estimated probability of being within limits
- `Pl` — Estimated probability of being below `L`
- `Pu` — Estimated probability of being above `U`
- `Cp` — `(U-L)/(6*sigma)`
- `Cpl` — `(mu-L)./(3.*sigma)`
- `Cpu` — `(U-mu)./(3.*sigma)`
- `Cpk` — `min(Cpl,Cpu)`

Indices are computed under the assumption that data values are independent samples from a normal population with constant mean and variance.

Indices divide a “specification width” (between specification limits) by a “process width” (between control limits). Higher ratios indicate a process with fewer measurements outside of specification.

Examples

Compute Capability Indices

Simulate a sample from a process with a mean of 3 and a standard deviation of 0.005.

```
rng default; % for reproducibility
data = normrnd(3,0.005,100,1);
```

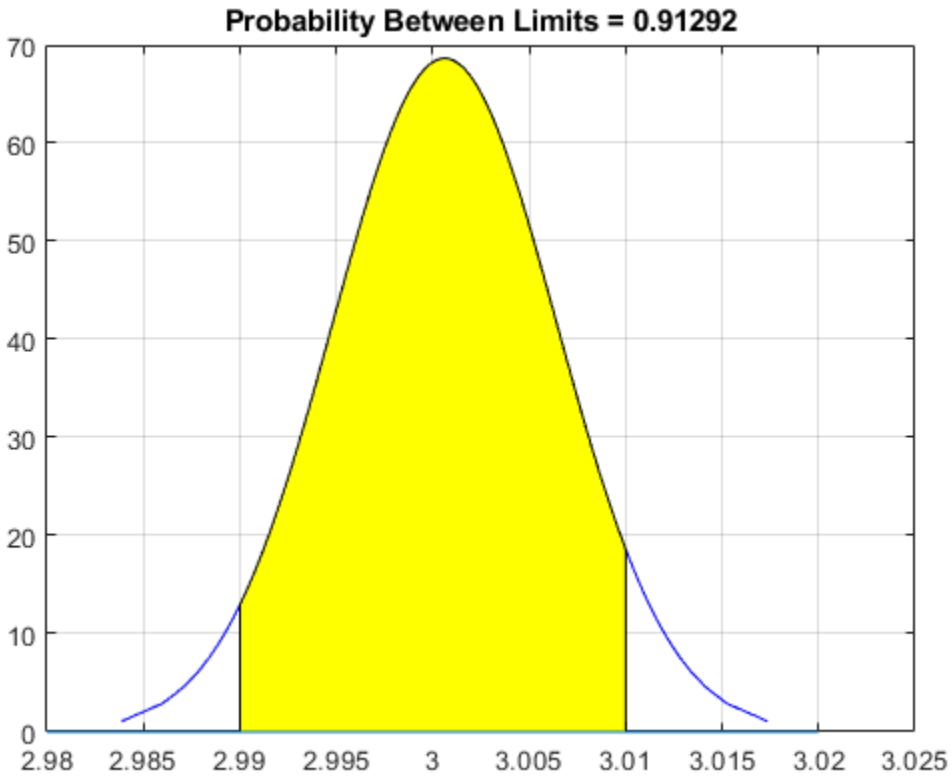
Compute capability indices if the process has an upper specification limit of 3.01 and a lower specification limit of 2.99.

```
S = capability(data,[2.99 3.01])
```

```
S = struct with fields:
       mu: 3.0006
    sigma: 0.0058
        P: 0.9129
       Pl: 0.0339
       Pu: 0.0532
       Cp: 0.5735
      Cpl: 0.6088
      Cpu: 0.5382
      Cpk: 0.5382
```

Visualize the specification and process widths.

```
capaplot(data,[2.99 3.01]);
grid on
```



References

[1] Montgomery, D. *Introduction to Statistical Quality Control*. Hoboken, NJ: John Wiley & Sons, 1991, pp. 369–374.

See Also

`capaplot` | `histfit`

Introduced in R2006b

capaplot

Process capability plot

Syntax

```
p = capaplot(data,specs)
[p,h] = capaplot(data,specs)
```

Description

`p = capaplot(data,specs)` estimates the mean of and variance for the observations in input vector `data`, and plots the pdf of the resulting T distribution. The observations in `data` are assumed to be normally distributed. The output, `p`, is the probability that a new observation from the estimated distribution will fall within the range specified by the two-element vector `specs`. The portion of the distribution between the lower and upper bounds specified in `specs` is shaded in the plot.

`[p,h] = capaplot(data,specs)` additionally returns handles to the plot elements in `h`.

`capaplot` treats `NaN` values in `data` as missing, and ignores them.

Examples

Create a Process Capability Plot

Randomly generate sample data from a normal process with a mean of 3 and a standard deviation of 0.005.

```
rng default; % For reproducibility
data = normrnd(3,0.005,100,1);
```

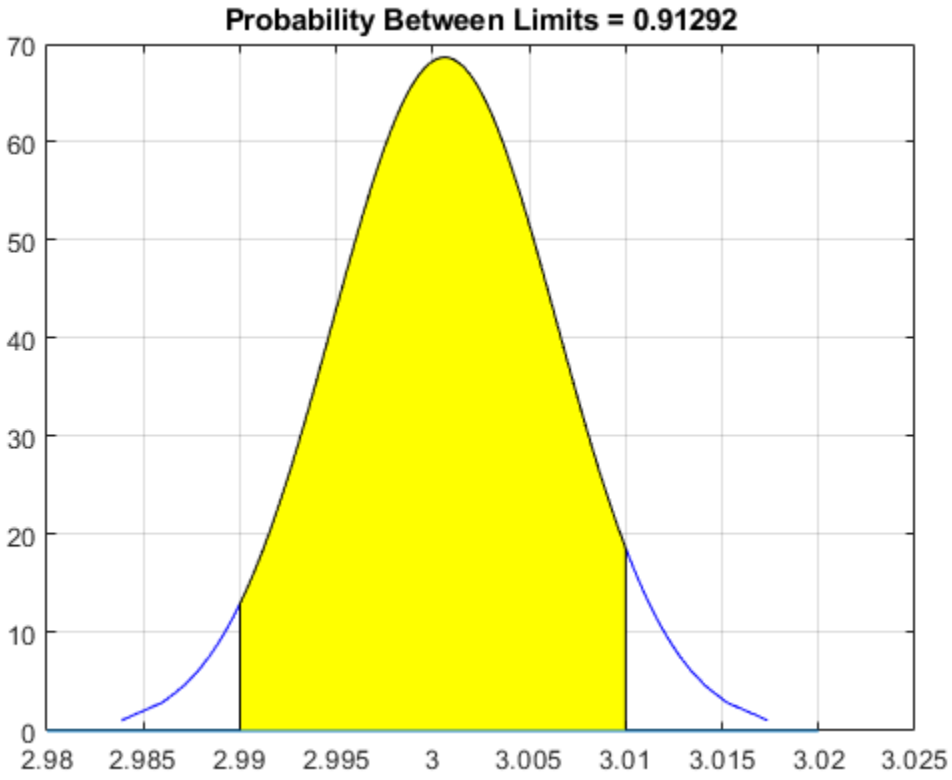
Compute capability indices if the process has an upper specification limit of 3.01 and a lower specification limit of 2.99.

```
S = capability(data,[2.99 3.01])
```

```
S = struct with fields:
       mu: 3.0006
    sigma: 0.0058
        P: 0.9129
       Pl: 0.0339
       Pu: 0.0532
       Cp: 0.5735
      Cpl: 0.6088
      Cpu: 0.5382
      Cpk: 0.5382
```

Visualize the specification and process widths.

```
capaplot(data,[2.99 3.01]);
grid on
```



See Also

`capability` | `histfit`

Introduced before R2006a

caseread

Read case names from file

Syntax

```
names = caseread(filename)
names = caseread
```

Description

`names = caseread(filename)` reads the contents of `filename` and returns a character array `names`. The `caseread` function treats each line of the file as a separate case name. Specify `filename` as either the name of a file in the current folder or the complete path name of a file.

`filename` can have one of the following file extensions:

- `.txt`, `.dat`, or `.csv` for delimited text files
- `.xls`, `.xlsm`, or `.xlsx` for Excel spreadsheet files

`names = caseread` opens the Select File to Open dialog box so that you can interactively select the file to read.

Examples

Write and Read Case Names

Create a character array of case names representing months.

```
months = char('January','February', ...
    'March','April','May');
```

Write the names to a file named `months.dat`. View the contents of the file by using the `type` function.

```
casewrite(months,'months.dat')
type months.dat
```

```
January
February
March
April
May
```

Read the names in the `months.dat` file.

```
names = caseread('months.dat')
```

```
names = 5x8 char array
    'January '
    'February'
    'March   '
```

```
    'April   '
    'May     '
```

Input Arguments

`filename` — Name of file to read

character vector | string scalar

Name of the file to read, specified as a character vector or string scalar.

Depending on the location of the file, `filename` has one of these forms.

Location of File	Form
Current folder or folder on the MATLAB path	Specify the name of the file in `filename`. **Example:** `'myTextFile.csv'`
Folder that is not the current folder or a folder on the MATLAB path	Specify the full or relative path name in `filename`. **Example:** `'C:\myFolder\myTextFile.csv'`

Example: `'months.dat'`

Data Types: `char` | `string`

Alternative Functionality

Instead of using `casewrite` and `caseread` with character arrays, consider using `writecell` and `readcell` with cell arrays. For example:

```
months = {'January';'February';'March';'April';'May'};
writecell(months,'months.dat')
names = readcell('months.dat')
```

```
names =

  5×1 cell array

    {'January' }
    {'February'}
    {'March'   }
    {'April'   }
    {'May'     }
```

See Also

`casewrite` | `gname` | `readcell` | `readtable` | `writecell` | `writetable`

Introduced before R2006a

casewrite

Write case names to file

Syntax

```
casewrite(strmat,filename)
casewrite(strmat)
```

Description

`casewrite(strmat,filename)` writes the contents of the character array or string column vector `strmat` to a file `filename`. Each row of `strmat` represents one case name, and `casewrite` writes each name to a separate line in `filename`. Specify `filename` as either a file name (to write the file to the current folder) or a complete path name (to write the file to a different folder).

`filename` can have one of the following file extensions:

- `.txt`, `.dat`, or `.csv` for delimited text files
- `.xls`, `.xlsm`, or `.xlsx` for Excel spreadsheet files

`casewrite(strmat)` opens the Select File to Write dialog box so that you can interactively specify the file to write.

Examples

Write and Read Case Names

Create a character array of case names representing months.

```
months = char('January','February', ...
    'March','April','May');
```

Write the names to a file named `months.dat`. View the contents of the file by using the `type` function.

```
casewrite(months,'months.dat')
type months.dat
```

```
January
February
March
April
May
```

Read the names in the `months.dat` file.

```
names = caseread('months.dat')
```

```
names = 5x8 char array
    'January '
    'February'
```

```
    'March   '
    'April   '
    'May     '
```

Input Arguments

`strmat` — Case names

character array | string column vector

Case names, specified as a character array or string column vector. Each row of `strmat` corresponds to a case name and becomes a line in `filename`.

Data Types: `char` | `string`

`filename` — Name of file to write

character vector | string scalar

Name of the file to write, specified as a character vector or string scalar.

Depending on the location you are writing to, `filename` has one of these forms.

Location of File	Form
Current folder	Specify the name of the file in `filename`. **Example:** `'myTextFile.csv'`
Folder that is different from the current folder	Specify the full or relative path name in `filename`. **Example:** `'C:\myFolder\myTextFile.csv'`

Example: `'months.dat'`

Data Types: `char` | `string`

Alternative Functionality

Instead of using `casewrite` and `caseread` with character arrays, consider using `writecell` and `readcell` with cell arrays. For example:

```
months = {'January';'February';'March';'April';'May'};
writecell(months,'months.dat')
names = readcell('months.dat')
```

```
names =

  5×1 cell array

    {'January' }
    {'February'}
    {'March'   }
    {'April'   }
    {'May'     }
```

See Also

caseread | gname | readcell | readtable | writecell | writetable

Introduced before R2006a

DaviesBouldinEvaluation

Package: `clustering.evaluation`
Superclasses: `ClusterCriterion`

Davies-Bouldin criterion clustering evaluation object

Description

`DaviesBouldinEvaluation` is an object consisting of sample data, clustering data, and Davies-Bouldin criterion values used to evaluate the optimal number of clusters. Create a Davies-Bouldin criterion clustering evaluation object using `evalclusters`.

Construction

`eva = evalclusters(x,clust,'DaviesBouldin')` creates a Davies-Bouldin criterion clustering evaluation object.

`eva = evalclusters(x,clust,'DaviesBouldin',Name,Value)` creates a Davies-Bouldin criterion clustering evaluation object using additional options specified by one or more name-value pair arguments.

Input Arguments

x — Input data

matrix

Input data, specified as an *N*-by-*P* matrix. *N* is the number of observations, and *P* is the number of variables.

Data Types: `single` | `double`

clust — Clustering algorithm

`'kmeans'` | `'linkage'` | `'gmdistribution'` | matrix of clustering solutions | function handle

Clustering algorithm, specified as one of the following.

`'kmeans'`	Cluster the data in `x` using the `kmeans` clustering algorithm, with `'EmptyAction'` set to `'singleton'` and `'Replicates'` set to `5`.
`'linkage'`	Cluster the data in `x` using the `clusterdata` agglomerative clustering algorithm, with `'Linkage'` set to `'ward'`.
`'gmdistribution'`	Cluster the data in `x` using the `gmdistribution` Gaussian mixture distribution algorithm, with `'SharedCov'` set to `true` and `'Replicates'` set to `5`.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can specify a clustering algorithm using a function handle. The function must be of the form `C = clustfun(DATA,K)`, where `DATA` is the data to be clustered, and `K` is the number of clusters. The output of `clustfun` must be one of the following:

- A vector of integers representing the cluster index for each observation in `DATA`. There must be `K` unique values in this vector.
- A numeric *n*-by-*K* matrix of score for *n* observations and *K* classes. In this case, the cluster index for each observation is determined by taking the largest score value in each row.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can also specify `clust` as a *n*-by-*K* matrix containing the proposed clustering solutions. *n* is the number of observations in the sample data, and *K* is the number of proposed clustering solutions. Column *j* contains the cluster indices for each of the *N* points in the *j*th clustering solution.

Data Types: `single` | `double` | `char` | `string` | `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KList',[1:5]` specifies to test 1, 2, 3, 4, and 5 clusters to find the optimal number.

`KList` — List of number of clusters to evaluate
vector

List of number of clusters to evaluate, specified as the comma-separated pair consisting of `'KList'` and a vector of positive integer values. You must specify `KList` when `clust` is a clustering algorithm name or a function handle. When `criterion` is `'gap'`, `clust` must be a character vector, a string scalar, or a function handle, and you must specify `KList`.

Example: `'KList',[1:6]`

Data Types: `single` | `double`

Properties

`ClusteringFunction`

Clustering algorithm used to cluster the input data, stored as a valid clustering algorithm name or function handle. If the clustering solutions are provided in the input, `ClusteringFunction` is empty.

`CriterionName`

Name of the criterion used for clustering evaluation, stored as a valid criterion name.

`CriterionValues`

Criterion values corresponding to each proposed number of clusters in `InspectedK`, stored as a vector of numerical values.

`InspectedK`

List of the number of proposed clusters for which to compute criterion values, stored as a vector of positive integer values.

`Missing`

Logical flag for excluded data, stored as a column vector of logical values. If `Missing` equals `true`, then the corresponding value in the data matrix `X` is not used in the clustering solution.

`NumObservations`

Number of observations in the data matrix `X`, minus the number of missing (`NaN`) values in `X`, stored as a positive integer value.

`OptimalK`

Optimal number of clusters, stored as a positive integer value.

`OptimalY`

Optimal clustering solution corresponding to `OptimalK`, stored as a column vector of positive integer values. If the clustering solutions are provided in the input, `OptimalY` is empty.

`X`

Data used for clustering, stored as a matrix of numerical values.

Methods

Inherited Methods

addK	Evaluate additional numbers of clusters
plot	Plot clustering evaluation object criterion values
compact	Compact clustering evaluation object

Examples

Evaluate the Clustering Solution Using Davies-Bouldin Criterion

Evaluate the optimal number of clusters using the Davies-Bouldin clustering evaluation criterion.

Generate sample data containing random numbers from three multivariate distributions with different parameter values.

```
rng('default');  % For reproducibility
mu1 = [2 2];
sigma1 = [0.9 -0.0255; -0.0255 0.9];

mu2 = [5 5];
sigma2 = [0.5 0 ; 0 0.3];

mu3 = [-2, -2];
sigma3 = [1 0 ; 0 0.9];

N = 200;

X = [mvnrnd(mu1,sigma1,N);...
```

```
    mvnrnd(mu2,sigma2,N);...
    mvnrnd(mu3,sigma3,N)];
```

Evaluate the optimal number of clusters using the Davies-Bouldin criterion. Cluster the data using `kmeans`.

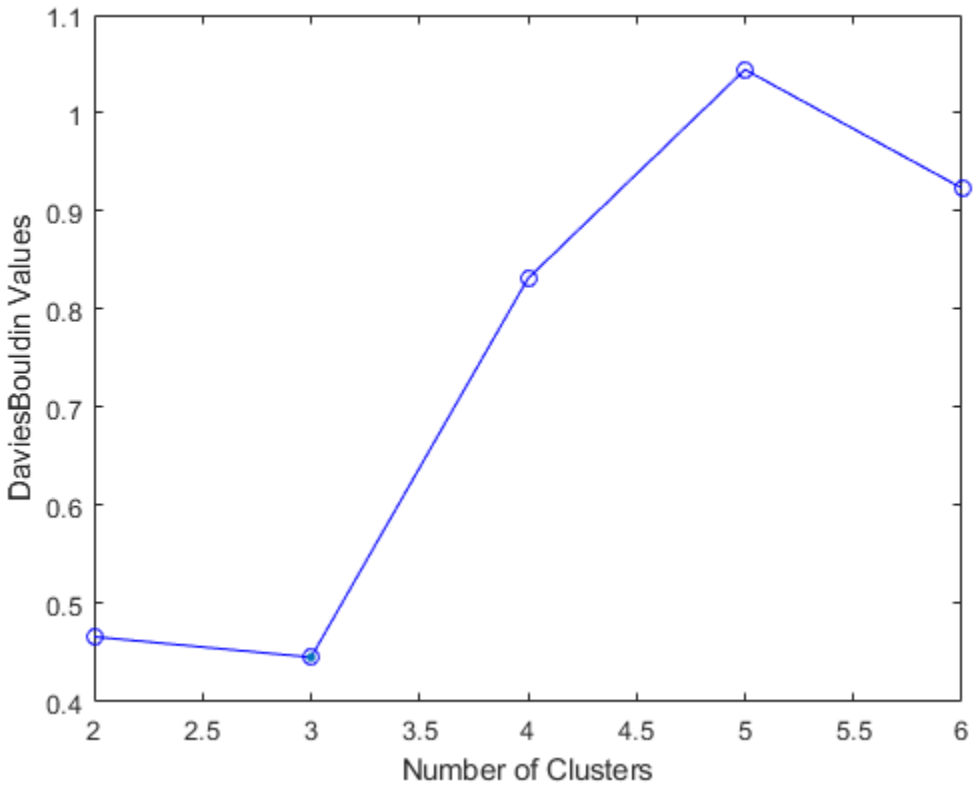
```
E = evalclusters(X,'kmeans','DaviesBouldin','klist',[1:6])
```

```
E =
  DaviesBouldinEvaluation with properties:

    NumObservations: 600
        InspectedK: [1 2 3 4 5 6]
    CriterionValues: [NaN 0.4663 0.4454 0.8316 1.0444 0.9236]
           OptimalK: 3
```

The `OptimalK` value indicates that, based on the Davies-Bouldin criterion, the optimal number of clusters is three.

Plot the Davies-Bouldin criterion values for each number of clusters tested.

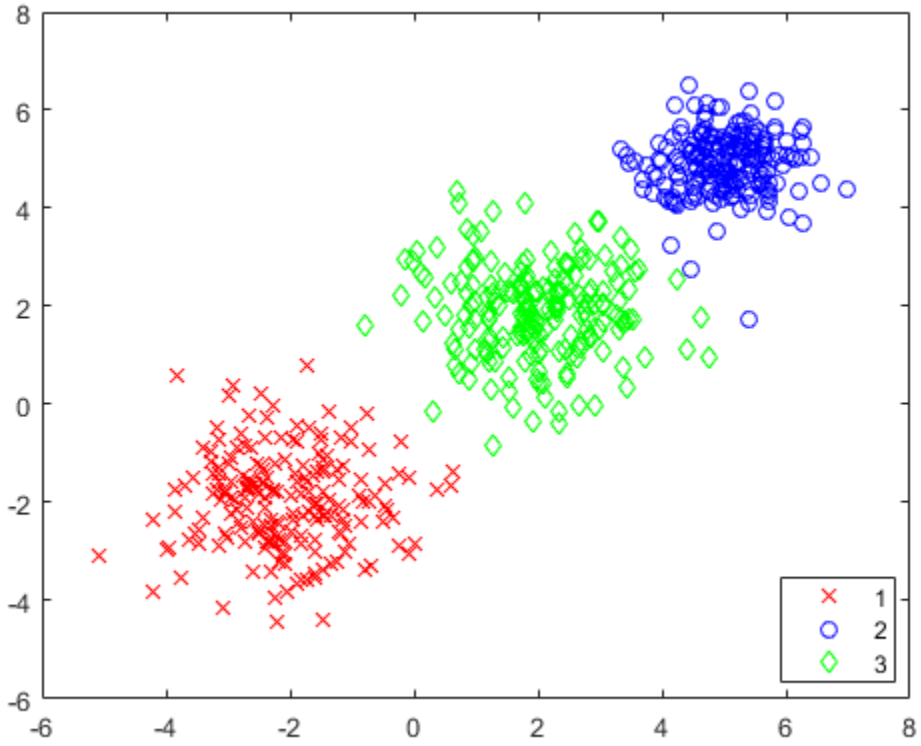
```
figure;
plot(E)
```



The plot shows that the lowest Davies-Bouldin value occurs at three clusters, suggesting that the optimal number of clusters is three.

Create a grouped scatter plot to visually examine the suggested clusters.

```
figure;
gscatter(X(:,1),X(:,2),E.OptimalY,'rbg','xod')
```



The plot shows three distinct clusters within the data: Cluster 1 is in the lower-left corner, cluster 2 is in the upper-right corner, and cluster 3 is near the center of the plot.

More About

Davies-Bouldin Criterion

The Davies-Bouldin criterion is based on a ratio of within-cluster and between-cluster distances. The Davies-Bouldin index is defined as

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\{D_{i,j}\},$$

where $D_{i,j}$ is the within-to-between cluster distance ratio for the ith and jth clusters. In mathematical terms,

$$D_{i,j} = \frac{\left(\bar{d}_i + \bar{d}_j\right)}{d_{i,j}}.$$

$\bar{d}_i$ is the average distance between each point in the ith cluster and the centroid of the ith cluster. $\bar{d}_j$ is the average distance between each point in the jth cluster and the centroid of the jth cluster. $d_{i,j}$ is the Euclidean distance between the centroids of the ith and jth clusters.

The maximum value of $D_{i,j}$ represents the worst-case within-to-between cluster ratio for cluster i. The optimal clustering solution has the smallest Davies-Bouldin index value.

References

[1] Davies, D. L., and D. W. Bouldin. “A Cluster Separation Measure.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. PAMI-1, No. 2, 1979, pp. 224–227.

See Also

`CalinskiHarabaszEvaluation` | `GapEvaluation` | `SilhouetteEvaluation` | `evalclusters`

Topics
Class Attributes
Property Attributes

cat

Class: `dataset`

(Not Recommended) Concatenate dataset arrays

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = cat(dim, ds1, ds2, ...)
```

Description

`ds = cat(dim, ds1, ds2, ...)` concatenates the dataset arrays `ds1, ds2, ...` along dimension `dim` by calling the `dataset/horzcat` or `dataset/vertcat` method. `dim` must be 1 or 2.

See Also

`horzcat` | `vertcat`

cdf

Cumulative distribution function for Gaussian mixture distribution

Syntax

```
y = cdf(gm,X)
```

Description

`y = cdf(gm,X)` returns the cumulative distribution function (cdf) of the Gaussian mixture distribution `gm`, evaluated at the values in `X`.

Examples

Compute cdf Values

Create a `gmdistribution` object and compute its cdf values.

Define the distribution parameters (means and covariances) of a two-component bivariate Gaussian mixture distribution.

```
mu = [1 2;-3 -5];
sigma = [1 1]; % shared diagonal covariance matrix
```

Create a `gmdistribution` object by using the `gmdistribution` function. By default, the function creates an equal proportion mixture.

```
gm = gmdistribution(mu,sigma)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:     1     2

Component 2:
Mixing proportion: 0.500000
Mean:    -3    -5
```

Compute the cdf values of `gm`.

```
X = [0 0;1 2;3 3;5 3];
cdf(gm,X)
```

```
ans = 4×1

    0.5011
    0.6250
    0.9111
    0.9207
```

Plot cdf

Create a `gmdistribution` object and plot its cdf.

Define the distribution parameters (means, covariances, and mixing proportions) of two bivariate Gaussian mixture components.

```
p = [0.4 0.6];               % Mixing proportions
mu = [1 2;-3 -5];            % Means
sigma = cat(3,[2 .5],[1 1])  % Covariances 1-by-2-by-2 array
```

```
sigma =
sigma(:,:,1) =

    2.0000    0.5000


sigma(:,:,2) =

     1     1
```

The `cat` function concatenates the covariances along the third array dimension. The defined covariance matrices are diagonal matrices. `sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.

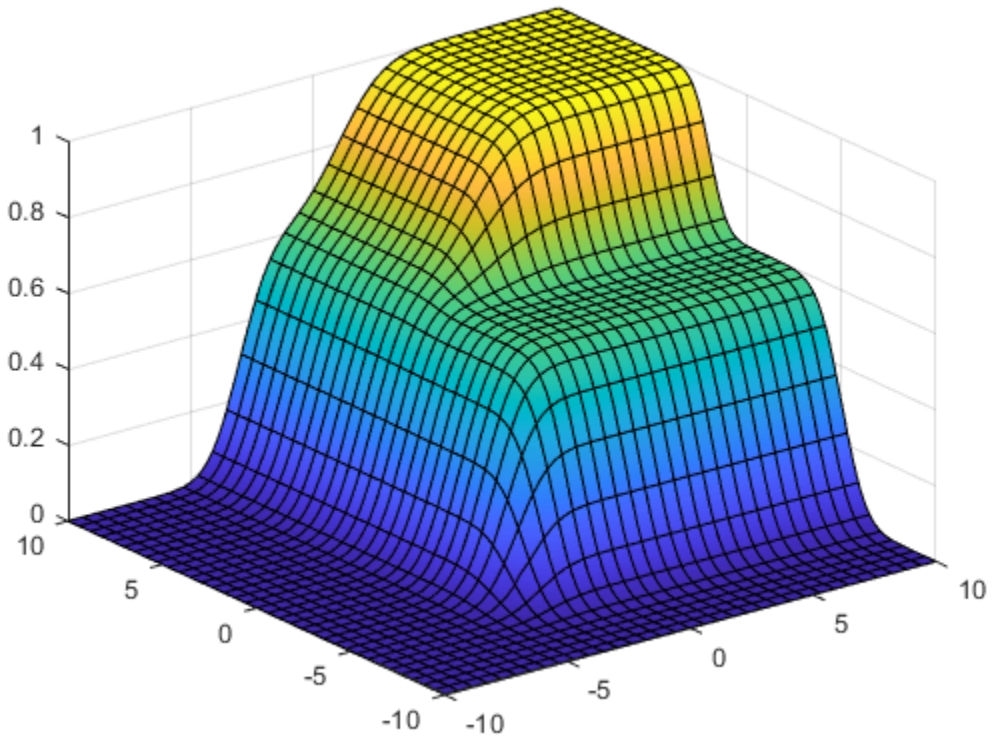
Create a `gmdistribution` object by using the `gmdistribution` function.

```
gm = gmdistribution(mu,sigma,p)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.400000
Mean:     1     2

Component 2:
Mixing proportion: 0.600000
Mean:    -3    -5
```

Plot the cdf of the Gaussian mixture distribution by using `fsurf`.

```
gmCDF = @(x,y) arrayfun(@(x0,y0) cdf(gm,[x0 y0]),x,y);
fsurf(gmCDF,[-10 10])
```



Input Arguments

gm — Gaussian mixture distribution
`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a `gmdistribution` object using `gmdistribution` or `fitgmdist`. Use the `gmdistribution` function to create a `gmdistribution` object by specifying the distribution parameters. Use the `fitgmdist` function to fit a `gmdistribution` model to data given a fixed number of components.

X — Values at which to evaluate cdf
n-by-*m* numeric matrix

Values at which to evaluate the cdf, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

Data Types: `single` | `double`

Output Arguments

y — cdf values
n-by-1 numeric vector

cdf values of the Gaussian mixture distribution `gm`, evaluated at `X`, returned as an *n*-by-1 numeric vector, where *n* is the number of observations in `X`.

See Also

`fitgmdist` | `gmdistribution` | `mvncdf` | `pdf` | `random`

Topics

“Create Gaussian Mixture Model” on page 5-112
“Fit Gaussian Mixture Model to Data” on page 5-115
“Simulate Data from Gaussian Mixture Model” on page 5-119
“Cluster Using Gaussian Mixture Model” on page 16-39

Introduced in R2007b

ccdesign

Central composite design

Syntax

```
dCC = ccdesign(n)
[dCC,blocks] = ccdesign(n)
[...] = ccdesign(n,'Name',value)
```

Description

`dCC = ccdesign(n)` generates a central composite design for `n` factors. `n` must be an integer `2` or larger. The output matrix `dCC` is *m*-by-`n`, where *m* is the number of runs in the design. Each row represents one run, with settings for all factors represented in the columns. Factor values are normalized so that the cube points take values between `-1` and `1`.

`[dCC,blocks] = ccdesign(n)` requests a blocked design. The output `blocks` is an *m*-by-1 vector of block numbers for each run. Blocks indicate runs that are to be measured under similar conditions to minimize the effect of inter-block differences on the parameter estimates.

`[...] = ccdesign(n,'Name',value)` specifies one or more optional name/value pairs for the design. Valid parameters and their values are listed in the following table. Specify *`Name`* in single quotes.

Parameter	Description	Values	Value Description
`center`	Number of center points.	Integer	Number of center points to include.
		`'uniform'`	Select number of center points to give uniform precision.
		`'orthogonal'`	Select number of center points to give an orthogonal design. This is the default.
`fraction`	Fraction of full-factorial cube, expressed as an exponent of 1/2.	`0`	Whole design. Default when $n \le 4$.
		`1`	1/2 fraction. Default when $4 < n \le 7$ or $n > 11$.
		`2`	1/4 fraction. Default when $7 < n \le 9$
		`3`	1/8 fraction. Default when $n = 10$.
		`4`	1/16 fraction. Default when $n = 11$.
`type`	Type of CCD.	`'circumscribed'`	Circumscribed (CCC). This is the default.

Parameter	Description	Values	Value Description
		`'inscribed'`	Inscribed (CCI).
		`'faced'`	Faced (CCF).
`blocksize`	Maximum number of points per block.	Integer	The default is `Inf`.

Examples

Two-Factor Central Composite Design

Create a 2-factor central composite design.

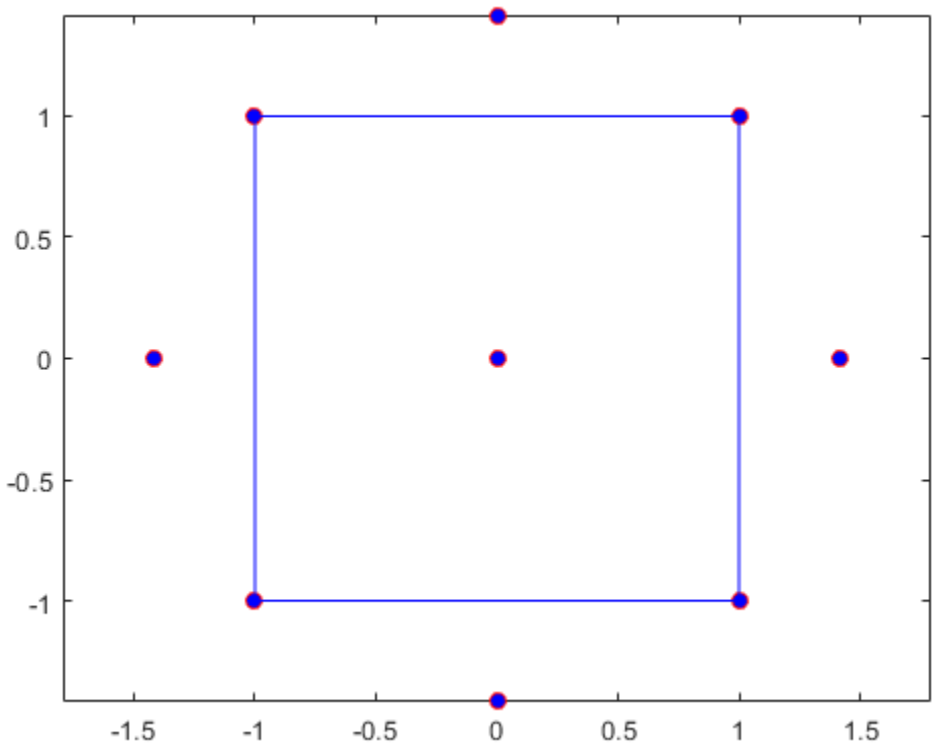
```
dCC = ccdesign(2,'type','circumscribed')
```

```
dCC = 16×2

   -1.0000   -1.0000
   -1.0000    1.0000
    1.0000   -1.0000
    1.0000    1.0000
   -1.4142         0
    1.4142         0
         0   -1.4142
         0    1.4142
         0         0
         0         0
    ⋮
```

The center point is run 8 times to reduce the correlations among the coefficient estimates.

Visualize the design.

```
plot(dCC(:,1),dCC(:,2),'ro','MarkerFaceColor','b')
X = [1 -1 -1 -1; 1 1 1 -1];
Y = [-1 -1 1 -1; 1 -1 1 1];
line(X,Y,'Color','b')
axis square equal
```



See Also

`bbdesign`

Introduced before R2006a

cdf

Package: `prob`

Cumulative distribution function

Syntax

```
y = cdf('name',x,A)
y = cdf('name',x,A,B)
y = cdf('name',x,A,B,C)
y = cdf('name',x,A,B,C,D)

y = cdf(pd,x)

y = cdf(___,'upper')
```

Description

`y = cdf('name',x,A)` returns the cumulative distribution function (cdf) for the one-parameter distribution family specified by `'name'` and the distribution parameter `A`, evaluated at the values in `x`.

`y = cdf('name',x,A,B)` returns the cdf for the two-parameter distribution family specified by `'name'` and the distribution parameters `A` and `B`, evaluated at the values in `x`.

`y = cdf('name',x,A,B,C)` returns the cdf for the three-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, and `C`, evaluated at the values in `x`.

`y = cdf('name',x,A,B,C,D)` returns the cdf for the four-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, `C`, and `D`, evaluated at the values in `x`.

`y = cdf(pd,x)` returns the cdf of the probability distribution object `pd`, evaluated at the values in `x`.

`y = cdf(___,'upper')` returns the complement of the cdf using an algorithm that more accurately computes the extreme upper-tail probabilities. `'upper'` can follow any of the input arguments in the previous syntaxes.

Examples

Compute Normal Distribution cdf

Create a standard normal distribution object with the mean, μ, equal to 0 and the standard deviation, σ, equal to 1.

```
mu = 0;
sigma = 1;
pd = makedist('Normal','mu',mu,'sigma',sigma);
```

Define the input vector x to contain the values at which to calculate the cdf.

```
x = [-2,-1,0,1,2];
```

Compute the cdf values for the standard normal distribution at the values in x.

```
y = cdf(pd,x)
```

```
y = 1×5

    0.0228    0.1587    0.5000    0.8413    0.9772
```

Each value in y corresponds to a value in the input vector x. For example, at the value x equal to 1, the corresponding cdf value y is equal to 0.8413.

Alternatively, you can compute the same cdf values without creating a probability distribution object. Use the `cdf` function, and specify a standard normal distribution using the same parameter values for μ and σ.

```
y2 = cdf('Normal',x,mu,sigma)
```

```
y2 = 1×5

    0.0228    0.1587    0.5000    0.8413    0.9772
```

The cdf values are the same as those computed using the probability distribution object.

Compute Poisson Distribution cdf

Create a Poisson distribution object with the rate parameter, λ, equal to 2.

```
lambda = 2;
pd = makedist('Poisson','lambda',lambda);
```

Define the input vector x to contain the values at which to calculate the cdf.

```
x = [0,1,2,3,4];
```

Compute the cdf values for the Poisson distribution at the values in x.

```
y = cdf(pd,x)
```

```
y = 1×5

    0.1353    0.4060    0.6767    0.8571    0.9473
```

Each value in y corresponds to a value in the input vector x. For example, at the value x equal to 3, the corresponding cdf value y is equal to 0.8571.

Alternatively, you can compute the same cdf values without creating a probability distribution object. Use the `cdf` function, and specify a Poisson distribution using the same value for the rate parameter, λ.

```
y2 = cdf('Poisson',x,lambda)
```

```
y2 = 1×5

    0.1353    0.4060    0.6767    0.8571    0.9473
```

The cdf values are the same as those computed using the probability distribution object.

Plot Standard Normal Distribution cdf

Create a standard normal distribution object.

```
pd = makedist('Normal')
```

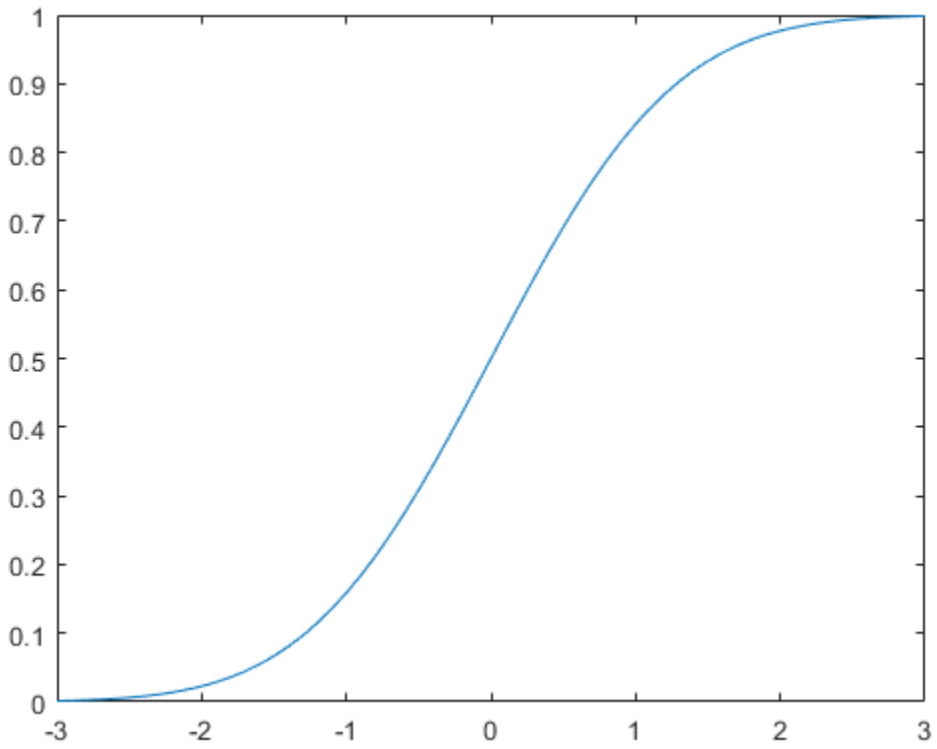
```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Specify the x values and compute the cdf.

```
x = -3:.1:3;
p = cdf(pd,x);
```

Plot the cdf of the standard normal distribution.

```
plot(x,p)
```



Plot Gamma Distribution cdf

Create three gamma distribution objects. The first uses the default parameter values. The second specifies `a = 1` and `b = 2`. The third specifies `a = 2` and `b = 1`.

```
pd_gamma = makedist('Gamma')

pd_gamma =
  GammaDistribution

  Gamma distribution
    a = 1
    b = 1


pd_12 = makedist('Gamma','a',1,'b',2)

pd_12 =
  GammaDistribution

  Gamma distribution
    a = 1
    b = 2


pd_21 = makedist('Gamma','a',2,'b',1)
```

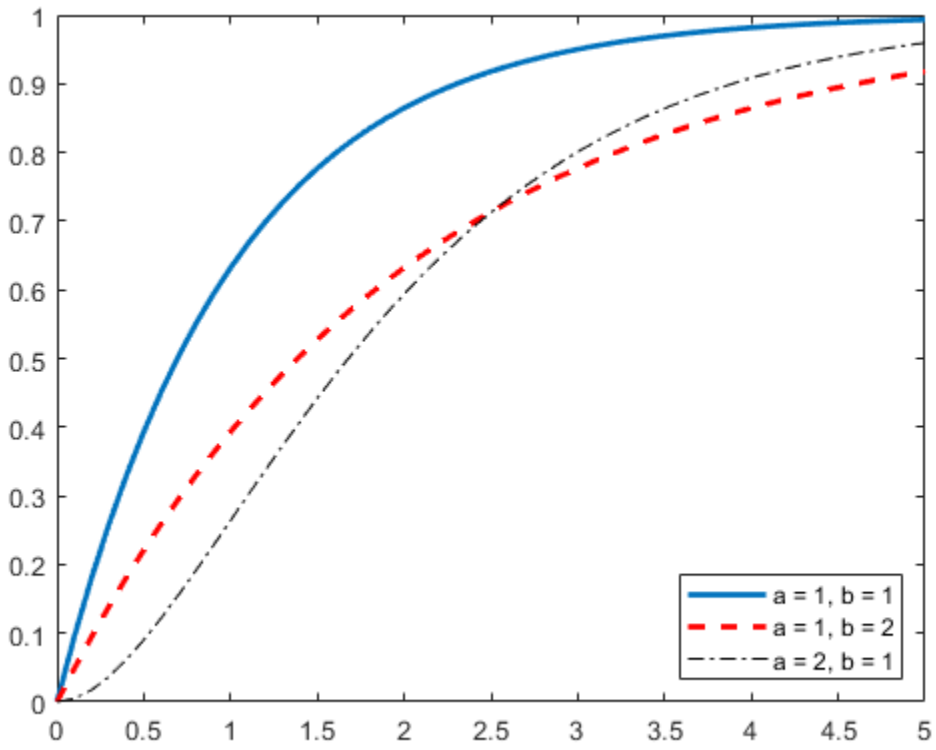
```
pd_21 =
  GammaDistribution

  Gamma distribution
    a = 2
    b = 1
```

Specify the x values and compute the cdf for each distribution.

```
x = 0:.1:5;
cdf_gamma = cdf(pd_gamma,x);
cdf_12 = cdf(pd_12,x);
cdf_21 = cdf(pd_21,x);
```

Create a plot to visualize how the cdf of the gamma distribution changes when you specify different values for the shape parameters a and b.

```
figure;
J = plot(x,cdf_gamma);
hold on;
K = plot(x,cdf_12,'r--');
L = plot(x,cdf_21,'k-.');
set(J,'LineWidth',2);
set(K,'LineWidth',2);
legend([J K L],'a = 1, b = 1','a = 1, b = 2','a = 2, b = 1','Location','southeast');
hold off;
```



Fit Pareto Tails to t Distribution and Compute the cdf

Fit Pareto tails to a *t* distribution at cumulative probabilities 0.1 and 0.9.

```
t = trnd(3,100,1);
obj = paretotails(t,0.1,0.9);
[p,q] = boundary(obj)
```

```
p = 2×1

    0.1000
    0.9000


q = 2×1

   -1.8487
    2.0766
```

Compute the cdf at the values in `q`.

```
cdf(obj,q)
```

```
ans = 2×1

    0.1000
    0.9000
```

Input Arguments

`'name'` — Probability distribution name

character vector or string scalar of probability distribution name

Probability distribution name, specified as one of the probability distribution names in this table.

`'name'`	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
`'Beta'`	“Beta Distribution” on page B-6	a first shape parameter	b second shape parameter	—	—
`'Binomial'`	“Binomial Distribution” on page B-10	n number of trials	p probability of success for each trial	—	—
`'BirnbaumSaunders'`	“Birnbaum-Saunders Distribution” on page B-18	β scale parameter	γ shape parameter	—	—
`'Burr'`	“Burr Type XII Distribution” on page B-19	α scale parameter	c first shape parameter	k second shape parameter	—
`'Chisquare'`	“Chi-Square Distribution” on page B-28	ν degrees of freedom	—	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Exponential'	"Exponential Distribution" on page B-33	μ mean	—	—	—
'Extreme Value'	"Extreme Value Distribution" on page B-40	μ location parameter	σ scale parameter	—	—
'F'	"F Distribution" on page B-45	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	—	—
'Gamma'	"Gamma Distribution" on page B-47	a shape parameter	b scale parameter	—	—
'Generalized Extreme Value'	"Generalized Extreme Value Distribution" on page B-55	k shape parameter	σ scale parameter	μ location parameter	—
'Generalized Pareto'	"Generalized Pareto Distribution" on page B-59	k tail index (shape) parameter	σ scale parameter	μ threshold (location) parameter	—
'Geometric'	"Geometric Distribution" on page B-63	p probability parameter	—	—	—
'HalfNormal'	"Half-Normal Distribution" on page B-68	μ location parameter	σ scale parameter	—	—
'Hypergeometric'	"Hypergeometric Distribution" on page B-73	m size of the population	k number of items with the desired characteristic in the population	n number of samples drawn	—
'InverseGaussian'	"Inverse Gaussian Distribution" on page B-75	μ scale parameter	λ shape parameter	—	—
'Logistic'	"Logistic Distribution" on page B-85	μ mean	σ scale parameter	—	—
'LogLogistic'	"Loglogistic Distribution" on page B-86	μ mean of logarithmic values	σ scale parameter of logarithmic values	—	—
'Lognormal'	"Lognormal Distribution" on page B-88	μ mean of logarithmic values	σ standard deviation of logarithmic values	—	—
'Nakagami'	"Nakagami Distribution" on page B-108	μ shape parameter	ω scale parameter	—	—
'Negative Binomial'	"Negative Binomial Distribution" on page B-109	r number of successes	p probability of success in a single trial	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Noncentral F'	“Noncentral F Distribution” on page B-115	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	δ noncentrality parameter	—
'Noncentral t'	“Noncentral t Distribution” on page B-117	ν degrees of freedom	δ noncentrality parameter	—	—
'Noncentral Chi-square'	“Noncentral Chi-Square Distribution” on page B-113	ν degrees of freedom	δ noncentrality parameter	—	—
'Normal'	“Normal Distribution” on page B-119	μ mean	σ standard deviation	—	—
'Poisson'	“Poisson Distribution” on page B-131	λ mean	—	—	—
'Rayleigh'	“Rayleigh Distribution” on page B-137	b scale parameter	—	—	—
'Rician'	“Rician Distribution” on page B-139	s noncentrality parameter	σ scale parameter	—	—
'Stable'	“Stable Distribution” on page B-140	α first shape parameter	β second shape parameter	γ scale parameter	δ location parameter
'T'	“Student's t Distribution” on page B-149	ν degrees of freedom	—	—	—
'tLocationScale'	“t Location-Scale Distribution” on page B-156	μ location parameter	σ scale parameter	ν shape parameter	—
'Uniform'	“Uniform Distribution (Continuous)” on page B-163	a lower endpoint (minimum)	b upper endpoint (maximum)	—	—
'Discrete Uniform'	“Uniform Distribution (Discrete)” on page B-168	n maximum observable value	—	—	—
'Weibull'	“Weibull Distribution” on page B-170	a scale parameter	b shape parameter	—	—

Example: 'Normal'

x — Values at which to evaluate cdf
scalar value | array of scalar values

Values at which to evaluate the cdf, specified as a scalar value or an array of scalar values.

If one or more of the input arguments x, A, B, C, and D are arrays, then the array sizes must be the same. In this case, cdf expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Example: `[0.1,0.25,0.5,0.75,0.9]`

Data Types: `single` | `double`

`A` — First probability distribution parameter
scalar value | array of scalar values

First probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `cdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`B` — Second probability distribution parameter
scalar value | array of scalar values

Second probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `cdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`C` — Third probability distribution parameter
scalar value | array of scalar values

Third probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `cdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`D` — Fourth probability distribution parameter
scalar value | array of scalar values

Fourth probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `cdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`pd` — Probability distribution
probability distribution object

Probability distribution, specified as a probability distribution object created with a function or app in this table.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.

Function or App	Description
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.
`paretotails`	Create a piecewise distribution object that has generalized Pareto distributions in the tails.

Output Arguments

y — cdf values

scalar value | array of scalar values

cdf values, returned as a scalar value or an array of scalar values. `y` is the same size as `x` after any necessary scalar expansion. Each element in `y` is the cdf value of the distribution, specified by the corresponding elements in the distribution parameters (`A`, `B`, `C`, and `D`) or the probability distribution object (`pd`), evaluated at the corresponding element in `x`.

Alternative Functionality

- `cdf` is a generic function that accepts either a distribution by its name `'name'` or a probability distribution object `pd`. It is faster to use a distribution-specific function, such as `normcdf` for the normal distribution and `binocdf` for the binomial distribution. For a list of distribution-specific functions, see “Supported Distributions” on page 5-14.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `'name'` must be a compile-time constant. For example, to use the normal distribution, include `coder.Constant('Normal')` in the `-args` value of `codegen`.
- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `ecdf` | `fitdist` | `icdf` | `makedist` | `mle` | `paretotails` | `pdf` | `random`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

cdfplot

Empirical cumulative distribution function (cdf) plot

Syntax

```
cdfplot(x)
h = cdfplot(x)
[h,stats] = cdfplot(x)
```

Description

`cdfplot(x)` creates an empirical cumulative distribution function (cdf) plot for the data in `x`. For a value t in `x`, the empirical cdf $F(t)$ is the proportion of the values in `x` less than or equal to t.

`h = cdfplot(x)` returns a handle of the empirical cdf plot line object. Use `h` to query or modify properties of the object after you create it. For a list of properties, see Line Properties.

`[h,stats] = cdfplot(x)` also returns a structure including summary statistics for the data in `x`.

Examples

Compare Empirical cdf to Theoretical cdf

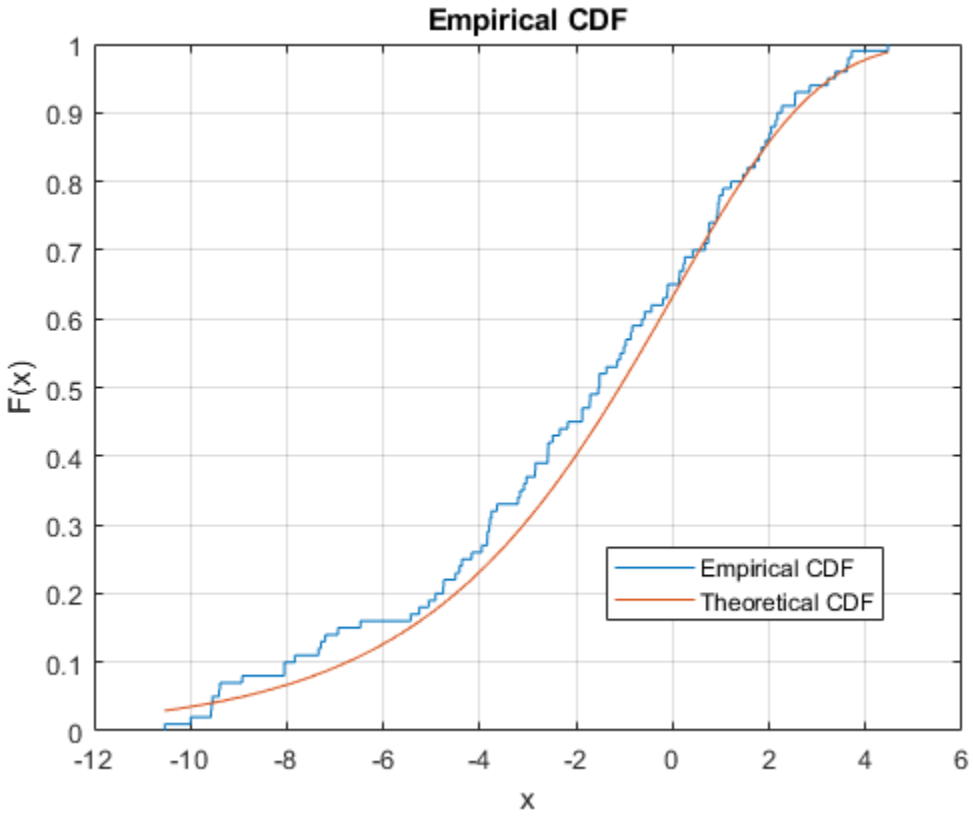
Plot the empirical cdf of a sample data set and compare it to the theoretical cdf of the underlying distribution of the sample data set. In practice, a theoretical cdf can be unknown.

Generate a random sample data set from the extreme value distribution with a location parameter of 0 and a scale parameter of 3.

```
rng('default')  % For reproducibility
y = evrnd(0,3,100,1);
```

Plot the empirical cdf of the sample data set and the theoretical cdf on the same figure.

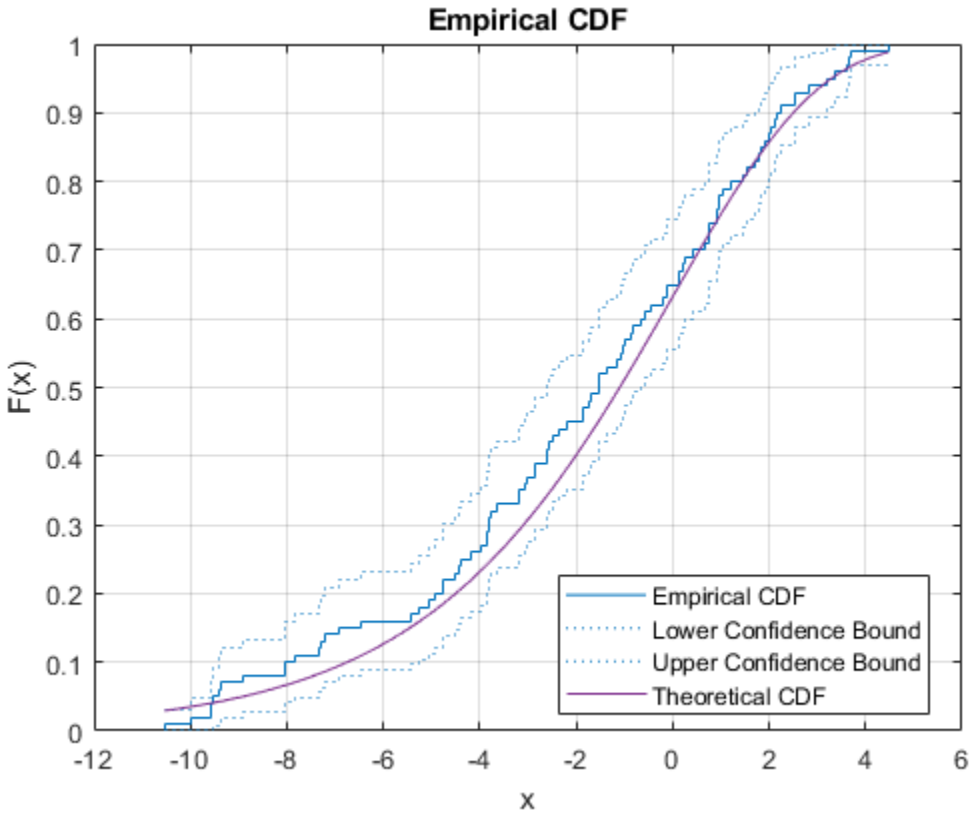
```
cdfplot(y)
hold on
x = linspace(min(y),max(y));
plot(x,evcdf(x,0,3))
legend('Empirical CDF','Theoretical CDF','Location','best')
hold off
```



The plot shows the similarity between the empirical cdf and the theoretical cdf.

Alternatively, you can use the `ecdf` function. The `ecdf` function also plots the 95% confidence intervals estimated by using Greenwood's Formula. For details, see “Greenwood’s Formula” on page 33-1279.

```
ecdf(y,'Bounds','on')
hold on
plot(x,evcdf(x,0,3))
grid on
title('Empirical CDF')
legend('Empirical CDF','Lower Confidence Bound','Upper Confidence Bound','Theoretical CDF','Loca
hold off
```



Test for Standard Normal Distribution

Perform the one-sample Kolmogorov-Smirnov test by using `kstest`. Confirm the test decision by visually comparing the empirical cumulative distribution function (cdf) to the standard normal cdf.

Load the `examgrades` data set. Create a vector containing the first column of the exam grade data.

```
load examgrades
test1 = grades(:,1);
```

Test the null hypothesis that the data comes from a normal distribution with a mean of 75 and a standard deviation of 10. Use these parameters to center and scale each element of the data vector, because `kstest` tests for a standard normal distribution by default.

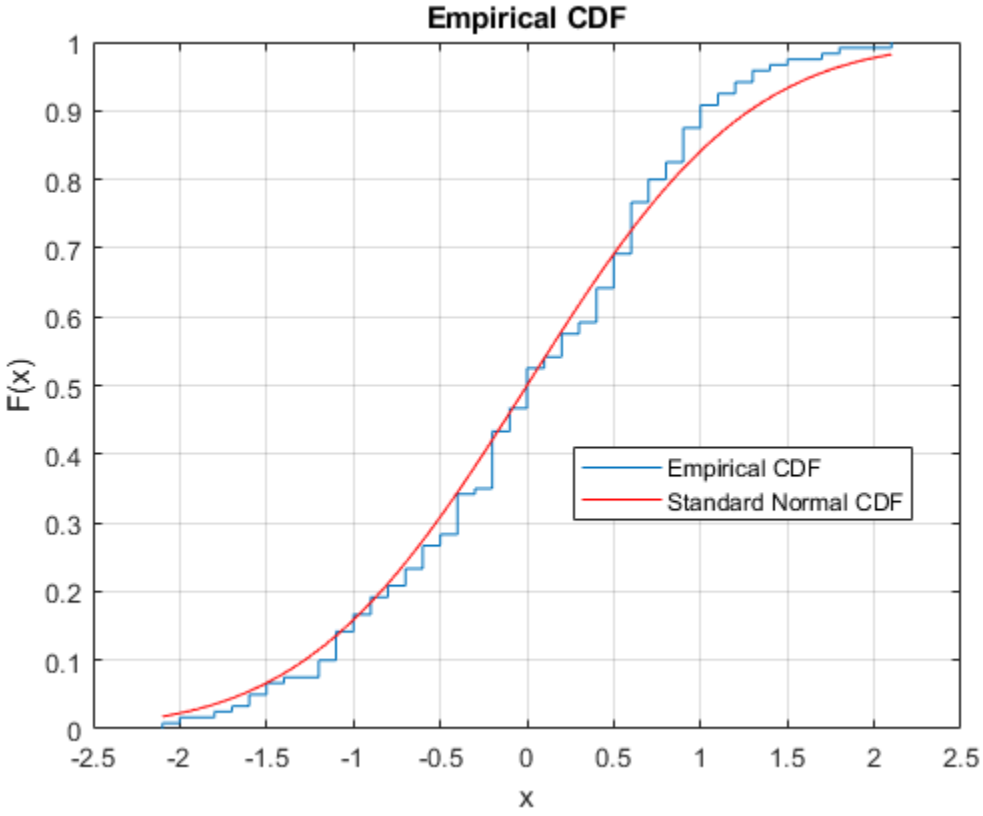
```
x = (test1-75)/10;
h = kstest(x)
```

```
h = logical
   0
```

The returned value of `h = 0` indicates that `kstest` fails to reject the null hypothesis at the default 5% significance level.

Plot the empirical cdf and the standard normal cdf for a visual comparison.

```
cdfplot(x)
hold on
x_values = linspace(min(x),max(x));
plot(x_values,normcdf(x_values,0,1),'r-')
legend('Empirical CDF','Standard Normal CDF','Location','best')
```



The figure shows the similarity between the empirical cdf of the centered and scaled data vector and the cdf of the standard normal distribution.

Input Arguments

x — Input data

numeric vector

Input data, specified as a numeric vector.

Data Types: `single` | `double`

Output Arguments

h — Handle of plot line object

chart line object

Handle of the empirical cdf plot line object, returned as a chart line object. Use `h` to query or modify properties of the object after you create it. For a list of properties, see Line Properties.

`stats` — Summary statistics
structure

Summary statistics for the data in `x`, returned as a structure with the following fields:

Field	Description
`min`	Minimum value
`max`	Maximum value
`mean`	Sample mean
`median`	Sample median (50th percentile)
`std`	Sample standard deviation

Tips

- `cdfplot` is useful for examining the distribution of a sample data set. You can overlay a theoretical cdf on the same plot of `cdfplot` to compare the empirical distribution of the sample to the theoretical distribution. For an example, see “Compare Empirical cdf to Theoretical cdf” on page 33-296.
- The `kstest`, `kstest2`, and `lillietest` functions compute test statistics derived from an empirical cdf. `cdfplot` is useful in helping you to understand the output from these functions. For an example, see “Test for Standard Normal Distribution” on page 33-298.

Alternative Functionality

You can use the `ecdf` function to find the empirical cdf values and create an empirical cdf plot. The `ecdf` function enables you to indicate censored data and compute the confidence bounds for the estimated cdf values.

See Also

`ecdf` | `ecdfhist` | `probplot` | `qqplot`

Topics
“Distribution Plots” on page 4-7
“Hypothesis Testing” on page 8-5

Introduced before R2006a

cell2dataset

(Not Recommended) Convert cell array to dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = cell2dataset(C)
ds = cell2dataset(C,Name,Value)
```

Description

`ds = cell2dataset(C)` converts a cell array to a `dataset` array.

`ds = cell2dataset(C,Name,Value)` performs the conversion using additional options specified by one or more `Name,Value` pair arguments.

Examples

Convert Cell Array to Dataset Array

Convert a cell array to a dataset array using the default options.

Create a cell array to convert.

```
C = {'Name','Gender','SystolicBP','DiastolicBP';
     'CLARK','M',124,93;
     'BROWN','F',122,80;
     'MARTIN','M',130,92}
```

```
C=4×4 cell array
    {'Name'  }    {'Gender'}    {'SystolicBP'}    {'DiastolicBP'}
    {'CLARK' }    {'M'     }    {[       124]}    {[         93]}
    {'BROWN' }    {'F'     }    {[       122]}    {[         80]}
    {'MARTIN'}    {'M'     }    {[       130]}    {[         92]}
```

Convert the cell array to a dataset array.

```
ds = cell2dataset(C)
```

```
ds =
    Name              Gender       SystolicBP    DiastolicBP
    {'CLARK' }        {'M'}        124           93
    {'BROWN' }        {'F'}        122           80
    {'MARTIN'}        {'M'}        130           92
```

The first row of `C` become the variable names in the output dataset array, `ds`.

Create a Dataset Array with Multicolumn Variables

Convert a cell array to a dataset array containing multicolumn variables.

Create a cell array to convert.

```
C = {'Name','Gender','SystolicBP','DiastolicBP';
     'CLARK','M',124,93;
     'BROWN','F',122,80;
     'MARTIN','M',130,92}
```

```
C=4×4 cell array
    {'Name'  }    {'Gender'}    {'SystolicBP'}    {'DiastolicBP'}
    {'CLARK' }    {'M'     }    {[       124]}    {[         93]}
    {'BROWN' }    {'F'     }    {[       122]}    {[         80]}
    {'MARTIN'}    {'M'     }    {[       130]}    {[         92]}
```

Convert the cell array to a dataset array, combining the systolic and diastolic blood pressure measurements into one variable named `BloodPressure`.

```
ds = cell2dataset(C,'NumCols',[1,1,2]);
ds.Properties.VarNames{3} = 'BloodPressure';
ds
```

```
ds =
    Name              Gender       BloodPressure
    {'CLARK' }        {'M'}        124          93
    {'BROWN' }        {'F'}        122          80
    {'MARTIN'}        {'M'}        130          92
```

The output dataset array has three observations and three variables.

Input Arguments

`C` — Input cell array

cell array

Input cell array to convert to a dataset array, specified as an *M*-by-*N* cell array. Each column of `C` becomes a variable in the output dataset array, `ds`. By default, `cell2dataset` assumes that the first row of `C` contains variable names.

Data Types: `cell` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ReadVarNames',false,'ReadObsNames',true` specifies that the first row of the cell array does not contain variable names, but the first column contains observation names.

ReadVarNames — Indicator for whether or not to read variable names
`true` (default) | `false`

Indicator for whether or not to read variable names from the first row of the input cell array, specified as the comma-separated pair consisting of `'ReadVarNames'` and either `true` or `false`. The default value is `true`, unless variable names are specified using the name-value pair argument `VarNames`. When `ReadVarNames` is `false`, `cell2dataset` creates default variable names if you do not provide any.

Example: `'ReadVarNames',false`

VarNames — Variable names for output dataset array
string array | cell array of character vectors

Variable names for the output dataset array, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors. You must provide a variable name for each variable in `ds`. The names must be valid MATLAB identifiers, and must be unique.

Example: `'VarNames',{'myVar1','myVar2','myVar3'}`

ReadObsNames — Indicator for whether or not to read observation names
`false` (default) | `true`

Indicator for whether or not to read observation names from the input cell array, specified as the comma-separated pair consisting of `'ReadObsNames'` and either `true` or `false`. When `ReadObsNames` has the value `true`, `cell2dataset` creates observation names in `ds` using the first column of `C`, and sets `ds.Properties.DimNames` equal to `{C{1,1},'Variables'}`.

Example: `'ReadObsNames',true`

ObsNames — Observation names for output dataset array
string array | cell array of character vectors

Observation names for the output dataset array, specified as the comma-separated pair consisting of `'ObsNames'` and a string array or cell array of character vectors. The names do not need to be valid MATLAB identifiers, but they must be unique.

NumCols — Number of columns for each variable
vector of nonnegative integers

Number of columns for each variable in `ds`, specified as the comma-separated pair consisting of `'NumCols'` and a vector of nonnegative integers. When the number of columns for a variable is greater than one, `cell2dataset` combines multiple columns in `C` into a single variable in `ds`. The vector you assign to `NumCols` must sum to `size(C,2)`, or `size(C,1)` of `ReadObsNames` is equal to `true`.

For example, to convert a cell array with eight columns into a dataset array with five variables, specify a vector with five elements that sum to eight, such as `'NumCols',[1,1,3,1,2]`.

Output Arguments

ds — Output dataset array
dataset array

Output dataset array, returned by default with a variable for each column of `C`, an observation for each row of `C` (except for the first row), and variable names corresponding to the first row of `C`.

- If you set `ReadVarNames` equal to `false` (or specify `VarNames`), then there is an observation in `ds` for each row of `C`, and `cell2dataset` creates default variable names (or uses the names in `VarNames`).
- If you set `ReadObsNames` equal to `true`, then `cell2dataset` uses the first column of `C` as observation names.
- If you specify `NumCols`, then the number of variables in `ds` is equal to the length of the specified vector of column numbers.

See Also

`dataset` | `dataset2cell` | `struct2dataset`

Topics

“Create a Dataset Array from Workspace Variables” on page 2-57
“Create a Dataset Array from a File” on page 2-62
“Dataset Arrays” on page 2-112

Introduced in R2012b

cellstr

Class: `dataset`

(Not Recommended) Create cell array of character vectors from dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = cellstr(A)
B = cellstr(A,VARS)
```

Description

`B = cellstr(A)` returns the contents of the dataset `A`, converted to a cell array of character vectors. The variables in the dataset must support the conversion and must have compatible sizes.

`B = cellstr(A,VARS)` returns the contents of the dataset variables specified by `VARS`. `VARS` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

See Also

`double` | `replacedata`

chi2cdf

Chi-square cumulative distribution function

Syntax

```
p = chi2cdf(x,nu)
p = chi2cdf(x,nu,'upper')
```

Description

`p = chi2cdf(x,nu)` returns the cumulative distribution function (cdf) of the chi-square distribution with degrees of freedom `nu`, evaluated at the values in `x`.

`p = chi2cdf(x,nu,'upper')` returns the complement of the cdf, evaluated at the values in `x` with degrees of freedom `nu`, using an algorithm that more accurately computes the extreme upper-tail probabilities than subtracting the lower tail value from 1.

Examples

Compute Chi-Square cdf

Compute the probability that an observation from the chi-square distribution with `5` degrees of freedom is in the interval `[0 3]`.

```
p1 = chi2cdf(3,5)
```

```
p1 = 0.3000
```

Compute the probability that an observation from the chi-square distributions with degrees of freedom `1` through `5` is in the interval `[0 3]`.

```
p2 = chi2cdf(3,1:5)
```

```
p2 = 1×5

    0.9167    0.7769    0.6084    0.4422    0.3000
```

The mean of the chi-square distribution is equal to the degrees of freedom. Compute the probability that an observation is in the interval `[0 nu]` for degrees of freedom `1` through `6`.

```
nu = 1:6;
x = nu;
p3 = chi2cdf(x,nu)
```

```
p3 = 1×6

    0.6827    0.6321    0.6084    0.5940    0.5841    0.5768
```

As the degrees of freedom increase, the probability that an observation from a chi-square distribution with degrees of freedom `nu` is less than the mean value approaches `0.5`.

Complementary cdf (Tail Distribution)

Determine the probability that an observation from the chi-square distribution with 3 degrees of freedom is in on the interval `[100 Inf]`.

```
p1 = 1 - chi2cdf(100,3)
```

```
p1 = 0
```

`chi2cdf(100,3)` is nearly `1`, so `p1` becomes `0`. Specify `'upper'` so that `chi2cdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = chi2cdf(100,3,'upper')
```

```
p2 = 1.5542e-21
```

Input Arguments

x — Values at which to evaluate cdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the cdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2cdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

nu — Degrees of freedom

positive scalar value | array of positive scalar values

Degrees of freedom for the chi-square distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2cdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

Output Arguments

p — cdf values

scalar value | array of scalar values

cdf values evaluated at the values in x, returned as a scalar value or an array of scalar values. p is the same size as x and nu after any necessary scalar expansion. Each element in p is the cdf value of the distribution specified by the corresponding element in nu, evaluated at the corresponding element in x.

More About

Chi-Square cdf

The chi-square distribution is a one-parameter family of curves. The parameter ν is the degrees of freedom.

The cdf of the chi-square distribution is

$$p = F(x|\nu) = \int_0^x \frac{t^{(\nu-2)/2}e^{-t/2}}{2^{\nu/2}\Gamma(\nu/2)}dt,$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result p is the probability that a single observation from the chi-square distribution with ν degrees of freedom falls in the interval $[0, x]$.

For more information, see “Chi-Square Distribution” on page B-28.

Alternative Functionality

- `chi2cdf` is a function specific to the chi-square distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `chi2cdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `chi2inv` | `chi2pdf` | `chi2rnd` | `chi2stat`

Topics
“Chi-Square Distribution” on page B-28

Introduced before R2006a

chi2gof

Chi-square goodness-of-fit test

Syntax

```
h = chi2gof(x)
h = chi2gof(x,Name,Value)
[h,p] = chi2gof(___)
[h,p,stats] = chi2gof(___)
```

Description

`h = chi2gof(x)` returns a test decision for the null hypothesis that the data in vector `x` comes from a normal distribution with a mean and variance estimated from `x`, using the chi-square goodness-of-fit test on page 33-315. The alternative hypothesis is that the data does not come from such a distribution. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = chi2gof(x,Name,Value)` returns a test decision for the chi-square goodness-of-fit test with additional options specified by one or more name-value pair arguments. For example, you can test for a distribution other than normal, or change the significance level of the test.

`[h,p] = chi2gof(___)` also returns the *p*-value `p` of the hypothesis test, using any of the input arguments from the previous syntaxes.

`[h,p,stats] = chi2gof(___)` also returns the structure `stats`, containing information about the test statistic.

Examples

Test for Normal Distribution

Create a standard normal probability distribution object. Generate a data vector `x` using random numbers from the distribution.

```
pd = makedist('Normal');
rng default;  % for reproducibility
x = random(pd,100,1);
```

Test the null hypothesis that the data in `x` comes from a population with a normal distribution.

```
h = chi2gof(x)
```

```
h = 0
```

The returned value `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the default 5% significance level.

Test Hypothesis at Different Significance Level

Create a standard normal probability distribution object. Generate a data vector `x` using random numbers from the distribution.

```
pd = makedist('Normal');
rng default;  % for reproducibility
x = random(pd,100,1);
```

Test the null hypothesis that the data in `x` comes from a population with a normal distribution at the 1% significance level.

```
[h,p] = chi2gof(x,'Alpha',0.01)
```

```
h = 0
```

```
p = 0.3775
```

The returned value `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the 1% significance level.

Test for Weibull Distribution Using Probability Distribution Object

Load the light bulb lifetime sample data.

```
load lightbulb
```

Create a vector from the first column of the data matrix, which contains the lifetime in hours of the light bulbs.

```
x = lightbulb(:,1);
```

Test the null hypothesis that the data in `x` comes from a population with a Weibull distribution. Use `fitdist` to create a probability distribution object with `A` and `B` parameters estimated from the data.

```
pd = fitdist(x,'Weibull');
h = chi2gof(x,'CDF',pd)
```

```
h = 1
```

The returned value `h = 1` indicates that `chi2gof` rejects the null hypothesis at the default 5% significance level.

Test for Poisson Distribution

Create six bins, numbered 0 through 5, to use for data pooling.

```
bins = 0:5;
```

Create a vector containing the observed counts for each bin and compute the total number of observations.

```
obsCounts = [6 16 10 12 4 2];
n = sum(obsCounts);
```

Fit a Poisson probability distribution object to the data and compute the expected count for each bin. Use the transpose operator `.'` to transform `bins` and `obsCounts` from row vectors to column vectors.

```
pd = fitdist(bins','Poisson','Frequency',obsCounts');
expCounts = n * pdf(pd,bins);
```

Test the null hypothesis that the data in `obsCounts` comes from a Poisson distribution with a lambda parameter equal to `lambdaHat`.

```
[h,p,st] = chi2gof(bins,'Ctrs',bins,...
                        'Frequency',obsCounts, ...
                        'Expected',expCounts,...
                        'NParams',1)
```

```
h = 0

p = 0.4654

st = struct with fields:
    chi2stat: 2.5550
          df: 3
       edges: [-0.5000 0.5000 1.5000 2.5000 3.5000 5.5000]
           O: [6 16 10 12 6]
           E: [7.0429 13.8041 13.5280 8.8383 6.0284]
```

The returned value `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the default 5% significance level. The vector `E` contains the expected counts for each bin under the null hypothesis, and `O` contains the observed counts for each bin.

Test for Normal Distribution Using Function Handle

Use the probability distribution function `normcdf` as a function handle in the chi-square goodness-of-fit test (`chi2gof`).

Test the null hypothesis that the sample data in the input vector `x` comes from a normal distribution with parameters μ and σ equal to the mean (`mean`) and standard deviation (`std`) of the sample data, respectively.

```
rng('default') % For reproducibility
x = normrnd(50,5,100,1);
h = chi2gof(x,'cdf',{@normcdf,mean(x),std(x)})
```

```
h = 0
```

The returned result `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the default 5% significance level.

Input Arguments

x — Sample data

vector

Sample data for the hypothesis test, specified as a vector.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NBins',8,'Alpha',0.01` pools the data into eight bins and conducts the hypothesis test at the 1% significance level.

`NBins` — Number of bins

`10` (default) | positive integer value

Number of bins to use for the data pooling, specified as the comma-separated pair consisting of `'NBins'` and a positive integer value. If you specify a value for `NBins`, do not specify a value for `Ctrs` or `Edges`.

Example: `'NBins',8`

Data Types: `single` | `double`

`Ctrs` — Bin centers

vector

Bin centers, specified as the comma-separated pair consisting of `'Ctrs'` and a vector of center values for each bin. If you specify a value for `Ctrs`, do not specify a value for `NBins` or `Edges`.

Example: `'Ctrs',[1 2 3 4 5]`

Data Types: `single` | `double`

`Edges` — Bin edges

vector

Bin edges, specified as the comma-separated pair consisting of `'Edges'` and a vector of edge values for each bin. If you specify a value for `Edges`, do not specify a value for `NBins` or `Ctrs`.

Example: `'Edges',[-2.5 -1.5 -0.5 0.5 1.5 2.5]`

Data Types: `single` | `double`

`CDF` — cdf of hypothesized distribution

probability distribution object | function handle | cell array

The cdf of the hypothesized distribution, specified as the comma-separated pair consisting of `'CDF'` and a probability distribution object, function handle, or cell array.

- If `CDF` is a probability distribution object, the degrees of freedom account for whether you estimate the parameters using `fitdist` or specify them using `makedist`.
- If `CDF` is a function handle, the distribution function must take `x` as its only argument.
- If `CDF` is a cell array, the first element must be a function handle, and the remaining elements must be parameter values, one per cell. The function must take `x` as its first argument, and the other parameters in the array as later arguments.

If you specify a value for `CDF`, do not specify a value for `Expected`.

Example: `'CDF',pd_object`

Data Types: `single` | `double`

`Expected` — Expected counts
vector of nonnegative values

Expected counts for each bin, specified as the comma-separated pair of `'Expected'` and a vector of nonnegative values. If `Expected` depends on estimated parameters, use `NParams` to ensure that `chi2gof` correctly calculates the degrees of freedom. If you specify a value for `Expected`, do not specify a value for `CDF`.

Example: `'Expected',[19.1446 18.3789 12.3224 8.2432 4.1378]`

Data Types: `single` | `double`

`NParams` — Number of estimated parameters
positive integer value

Number of estimated parameters used to describe the null distribution, specified as the comma-separated pair consisting of `'NParams'` and a positive integer value. This value adjusts the degrees of freedom of the test based on the number of estimated parameters used to compute the cdf or expected counts.

The default value for `NParams` depends on how you specify the null distribution:

- If you specify `CDF` as a probability distribution object, `NParams` is equal to the number of estimated parameters used to create the object.
- If you specify `CDF` as a function name or handle, the default value of `NParams` is `0`.
- If you specify `CDF` as a cell array, the default value of `NParams` is the number of parameters in the array.
- If you specify `Expected`, the default value of `NParams` is `0`.

Example: `'NParams',1`

Data Types: `single` | `double`

`EMin` — Minimum expected count per bin
`5` (default) | nonnegative integer value

Minimum expected count per bin, specified as the comma-separated pair consisting of `'EMin'` and a nonnegative integer value. If the bin at the extreme end of either tail has an expected value less than `EMin`, it is combined with a neighboring bin until the count in each extreme bin is at least `5`. If any interior bins have a count less than `5`, `chi2gof` displays a warning, but does not combine the interior bins. In that case, you should use fewer bins, or provide bin centers or edges, to increase the expected counts in all bins. Specify `EMin` as `0` to prevent the combining of bins.

Example: `'EMin',0`

Data Types: `single` | `double`

`Frequency` — Frequency
vector of nonnegative integer values

Frequency of data values, specified as the comma-separated pair consisting of `'Frequency'` and a vector of nonnegative integer values that is the same length as the vector `x`.

Example: `'Frequency',[20 16 13 10 8]`

Data Types: `single` | `double`

`Alpha` — Significance level
`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`stats` — Test statistics
structure

Test statistics, returned as a structure containing the following:

- `chi2stat` — Value of the test statistic.
- `df` — Degrees of freedom of the test.
- `edges` — Vector of bin edges after pooling.
- `O` — Vector of observed counts for each bin.
- `E` — Vector of expected counts for each bin.

More About

Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test determines if a data sample comes from a specified probability distribution, with parameters estimated from the data.

The test groups the data into bins, calculating the observed and expected counts for those bins, and computing the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{N} (O_i - E_i)^2 / E_i \,,$$

where O_i are the observed counts and E_i are the expected counts based on the hypothesized distribution. The test statistic has an approximate chi-square distribution when the counts are sufficiently large.

Algorithms

`chi2gof` compares the value of the test statistic to a chi-square distribution with degrees of freedom equal to *nbins* - 1 - *nparams*, where *nbins* is the number of bins used for the data pooling and *nparams* is the number of estimated parameters used to determine the expected counts. If there are not enough degrees of freedom to conduct the test, `chi2gof` returns the *p*-value as `NaN`.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`kstest` | `lillietest`

Topics

“Chi-Square Distribution” on page B-28

Introduced before R2006a

chi2inv

Chi-square inverse cumulative distribution function

Syntax

```
x = chi2inv(p,nu)
```

Description

`x = chi2inv(p,nu)` returns the inverse cumulative distribution function (icdf) of the chi-square distribution with degrees of freedom `nu`, evaluated at the probability values in `p`.

Examples

Compute Chi-Square icdf

Find the 95th percentile for the chi-square distribution with `10` degrees of freedom.

```
x = chi2inv(0.95,10)
```

```
x = 18.3070
```

If you generate random numbers from this chi-square distribution, you would observe numbers greater than `18.3` only `5%` of the time.

Median of Chi-Square Distributions

Compute the medians of the chi-square distributions with degrees of freedom one through six.

```
x = chi2inv(0.50,1:6)
```

```
x = 1×6

    0.4549    1.3863    2.3660    3.3567    4.3515    5.3481
```

Input Arguments

p — Probability values at which to evaluate icdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the icdf, specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

- To evaluate the icdf at multiple values, specify `p` using an array.

- To evaluate the icdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `p` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2inv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding probabilities in `p`.

Example: `[0.1,0.5,0.9]`

Data Types: `single` | `double`

`nu` — Degrees of freedom
positive scalar value | array of positive scalar values

Degrees of freedom for the chi-square distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `p` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2inv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding probabilities in `p`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

Output Arguments

`x` — icdf values
scalar value | array of scalar values

icdf values evaluated at the probabilities in `p`, returned as a scalar value or an array of scalar values. `x` is the same size as `p` and `nu` after any necessary scalar expansion. Each element in `x` is the icdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding probabilities in `p`.

More About

Chi-Square icdf

The chi-square distribution is a one-parameter family of curves. The parameter ν is the degrees of freedom.

The icdf of the chi-square distribution is

$$x = F^{-1}(p|\nu) = \left\{x: F(x|\nu) = p\right\},$$

where

$$p = F(x|\nu) = \int_0^x \frac{t^{(\nu-2)/2} e^{-t/2}}{2^{\nu/2}\Gamma(\nu/2)} dt,$$

ν is the degrees of freedom, and $\Gamma(\cdot)$ is the Gamma function. The result p is the probability that a single observation from the chi-square distribution with ν degrees of freedom falls in the interval [0, x].

For more information, see “Chi-Square Distribution” on page B-28.

Alternative Functionality

- `chi2inv` is a function specific to the chi-square distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use `icdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `chi2inv` is faster than the generic function `icdf`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`chi2cdf` | `chi2pdf` | `chi2rnd` | `chi2stat` | `icdf`

Topics

“Chi-Square Distribution” on page B-28

Introduced before R2006a

chi2pdf

Chi-square probability density function

Syntax

```
y = chi2pdf(x,nu)
```

Description

`y = chi2pdf(x,nu)` returns the probability density function (pdf) of the chi-square distribution with `nu` degrees of freedom, evaluated at the values in `x`.

Examples

Compute Chi-Square pdf

Compute the density of the observed value 2 in the chi-square distribution with 3 degrees of freedom.

```
y1 = chi2pdf(2,3)
```

```
y1 = 0.2076
```

Compute the density of the observed value 4 in the chi-square distributions with degrees of freedom 1 through 6.

```
y2 = chi2pdf(4,1:6)
```

```
y2 = 1×6

    0.0270    0.0677    0.1080    0.1353    0.1440    0.1353
```

The mean of the chi-square distribution is equal to the degrees of freedom. Compute the density of the mean for the chi-square distributions with degrees of freedom 1 through 6.

```
nu = 1:6;
x = nu;
y3 = chi2pdf(x,nu)
```

```
y3 = 1×6

    0.2420    0.1839    0.1542    0.1353    0.1220    0.1120
```

As the degrees of freedom increase, the density of the mean decreases.

Input Arguments

x — Values at which to evaluate pdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the pdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2pdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

`nu` — Degrees of freedom
positive scalar value | array of positive scalar values

Degrees of freedom for the chi-square distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `chi2pdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `p` is the same size as `x` and `nu` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

More About

Chi-Square pdf

The chi-square distribution is a one-parameter family of curves. The parameter ν is the degrees of freedom.

The pdf of the chi-square distribution is

$$y = f(x|\nu) = \frac{x^{(\nu-2)/2}e^{-x/2}}{2^{\frac{\nu}{2}}\Gamma(\nu/2)},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function.

For more information, see “Chi-Square Distribution” on page B-28.

Alternative Functionality

- `chi2pdf` is a function specific to the chi-square distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `chi2pdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`chi2cdf` | `chi2inv` | `chi2rnd` | `chi2stat` | `pdf`

Topics

“Chi-Square Distribution” on page B-28

Introduced before R2006a

chi2rnd

Chi-square random numbers

Syntax

```
r = chi2rnd(nu)
r = chi2rnd(nu,sz1,...,szN)
r = chi2rnd(nu,sz)
```

Description

`r = chi2rnd(nu)` generates a random number from the chi-square distribution with `nu` degrees of freedom.

`r = chi2rnd(nu,sz1,...,szN)` generates an array of random numbers from the chi-square distribution, where `sz1,...,szN` indicates the size of each dimension.

`r = chi2rnd(nu,sz)` generates an array of random numbers from the chi-square distribution, where vector `sz` specifies `size(r)`.

Examples

Generate Chi-Square Random Number

Generate a single random number from the chi-square distribution with `10` degrees of freedom.

```
nu = 10;
r = chi2rnd(nu)
```

```
r = 19.7102
```

Generate Chi-Square Random Numbers

Generate a 1-by-6 array of chi-square random numbers with `1` degree of freedom.

```
nu1 = ones(1,6);  % 1-by-6 array of ones
r1 = chi2rnd(nu1)
```

```
r1 = 1×6

    2.5368    0.2447    0.4314    2.0153    0.0418    4.3486
```

By default, `chi2rnd` generates an array that is the same size as `nu`.

If you specify `nu` as a scalar, `chi2rnd` expands `nu` into a constant array with dimensions specified by `sz1,...,szN`.

Generate a 2-by-6 array of chi-square random numbers, all with 3 degrees of freedom.

```
nu2 = 3;
sz1 = 2;
sz2 = 6;
r2 = chi2rnd(nu2,sz1,sz2)
```

```
r2 = 2×6

    0.5761    5.3582    1.0124    0.9851    1.0529    3.0765
    7.9240    1.7373    0.6291    7.0240    1.8496    2.2690
```

If you specify both `nu` and `sz` as arrays, then the dimensions specified by `sz` must match the dimension of `nu`.

Generate a 1-by-6 array of chi-square random numbers with 3 to 8 degrees of freedom.

```
nu3 = 3:8;
sz = [1 6];
r3 = chi2rnd(nu3,sz)
```

```
r3 = 1×6

    3.9690    7.0961    4.5651    2.4606   13.5038    8.8495
```

Input Arguments

`nu` — Degrees of freedom

positive scalar value | array of positive scalar values

Degrees of freedom for the chi-square distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `nu` using an array. Each element in `r` is the random number generated from the distribution specified by the corresponding element in `nu`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)

integers

Size of each dimension, specified as separate arguments of integers.

If `nu` is an array, then the specified dimensions `sz1,...,szN` must match the dimensions of `nu`. The default values of `sz1,...,szN` are the dimensions of `nu`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `chi2rnd` ignores trailing dimensions with a size of 1. For example, `chi2rnd(5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the distribution with five degrees of freedom.

Example: `2,3`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If `nu` is an array, then the specified dimensions `sz` must match the dimensions of `nu`. The default values of `sz` are the dimensions of `nu`.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `chi2rnd` ignores trailing dimensions with a size of 1. For example, `chi2rnd(5,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the distribution with five degrees of freedom.

Example: `[2 3]`

Data Types: `single` | `double`

Output Arguments

`r` — Chi-square random numbers
scalar value | array of scalar values

Chi-square random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding element in `nu`.

Alternative Functionality

- `chi2rnd` is a function specific to the chi-square distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Note that the distribution-specific function `chi2rnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`chi2cdf` | `chi2inv` | `chi2pdf` | `chi2stat` | `random`

Topics

“Chi-Square Distribution” on page B-28

Introduced before R2006a

chi2stat

Chi-square mean and variance

Syntax

```
[M,V] = chi2stat(NU)
```

Description

`[M,V] = chi2stat(NU)` returns the mean of and variance for the chi-square distribution with degrees of freedom parameters specified by `NU`. `NU` can be a vector, a matrix, or a multidimensional array. The degrees of freedom parameters in `NU` must be positive.

The mean of the chi-square distribution is ν, the degrees of freedom parameter, and the variance is 2ν.

Examples

```
nu = 1:10;
nu = nu'*nu;
[m,v] = chi2stat(nu)
m =
 1   2   3   4   5   6   7   8   9  10
 2   4   6   8  10  12  14  16  18  20
 3   6   9  12  15  18  21  24  27  30
 4   8  12  16  20  24  28  32  36  40
 5  10  15  20  25  30  35  40  45  50
 6  12  18  24  30  36  42  48  54  60
 7  14  21  28  35  42  49  56  63  70
 8  16  24  32  40  48  56  64  72  80
 9  18  27  36  45  54  63  72  81  90
10  20  30  40  50  60  70  80  90 100

v =
 2   4   6   8  10  12  14  16  18  20
 4   8  12  16  20  24  28  32  36  40
 6  12  18  24  30  36  42  48  54  60
 8  16  24  32  40  48  56  64  72  80
10  20  30  40  50  60  70  80  90 100
12  24  36  48  60  72  84  96 108 120
14  28  42  56  70  84  98 112 126 140
16  32  48  64  80  96 112 128 144 160
18  36  54  72  90 108 126 144 162 180
20  40  60  80 100 120 140 160 180 200
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`chi2cdf` | `chi2inv` | `chi2pdf` | `chi2rnd`

Topics

“Chi-Square Distribution” on page B-28

Introduced before R2006a

cholcov

Cholesky-like covariance decomposition

Syntax

```
T = cholcov(SIGMA)
[T,num] = cholcov(SIGMA)
[T,num] = cholcov(SIGMA,0)
```

Description

`T = cholcov(SIGMA)` computes `T` such that `SIGMA = T'*T`. `SIGMA` must be square, symmetric, and positive semi-definite. If `SIGMA` is positive definite, then `T` is the square, upper triangular Cholesky factor. If `SIGMA` is not positive definite, `T` is computed from an eigenvalue decomposition of `SIGMA`. `T` is not necessarily triangular or square in this case. Any eigenvectors whose corresponding eigenvalue is close to zero (within a small tolerance) are omitted. If any remaining eigenvalues are negative, `T` is empty.

`[T,num] = cholcov(SIGMA)` returns the number `num` of negative eigenvalues of `SIGMA`, and `T` is empty if `num` is positive. If `num` is zero, `SIGMA` is positive semi-definite. If `SIGMA` is not square and symmetric, `num` is `NaN` and `T` is empty.

`[T,num] = cholcov(SIGMA,0)` returns `num` equal to zero if `SIGMA` is positive definite, and `T` is the Cholesky factor. If `SIGMA` is not positive definite, `num` is a positive integer and `T` is empty. `[...] = cholcov(SIGMA,1)` is equivalent to `[...] = cholcov(SIGMA)`.

Examples

The following 4-by-4 covariance matrix is rank-deficient:

```
C1 = [2 1 1 2;1 2 1 2;1 1 2 2;2 2 2 3]
C1 =
     2     1     1     2
     1     2     1     2
     1     1     2     2
     2     2     2     3
rank(C1)
ans =
     3
```

Use `cholcov` to factor `C1`:

```
T = cholcov(C1)
T =
   -0.2113    0.7887   -0.5774         0
    0.7887   -0.2113   -0.5774         0
    1.1547    1.1547    1.1547    1.7321

C2 = T'*T
C2 =
    2.0000   1.0000   1.0000   2.0000
```

```
    1.0000   2.0000   1.0000   2.0000
    1.0000   1.0000   2.0000   2.0000
    2.0000   2.0000   2.0000   3.0000
```

Use `T` to generate random data with the specified covariance:

```
C3 = cov(randn(1e6,3)*T)
C3 =
    1.9973    0.9982    0.9995    1.9975
    0.9982    1.9962    0.9969    1.9956
    0.9995    0.9969    1.9980    1.9972
    1.9975    1.9956    1.9972    2.9951
```

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`chol` | `cov`

Introduced in R2007a

ClassificationBaggedEnsemble

Package: `classreg.learning.classif`
Superclasses: `ClassificationEnsemble`

Classification ensemble grown by resampling

Description

`ClassificationBaggedEnsemble` combines a set of trained weak learner models and data on which these learners were trained. It can predict ensemble response for new data by aggregating predictions from its weak learners.

Construction

Create a bagged classification ensemble object using `fitcensemble`. Set the name-value pair argument `'Method'` of `fitcensemble` to `'Bag'` to use bootstrap aggregation (bagging, for example, random forest).

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
```

```
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`CombineWeights`

Character vector describing how `ens` combines weak learner weights, either `'WeightedSum'` or `'WeightedAverage'`.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`FitInfo`

Numeric array of fit information. The `FitInfoDescription` property describes the content of this array.

`FitInfoDescription`

Character vector describing the meaning of the `FitInfo` array.

`FResample`

Numeric scalar between `0` and `1`. `FResample` is the fraction of training data `fitcensemble` resampled at random for every weak learner when constructing the ensemble.

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`

- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`Method`

Character vector describing the method that creates `ens`.

`ModelParameters`

Parameters used in training `ens`.

`NumTrained`

Number of trained weak learners in `ens`, a scalar.

`PredictorNames`

Cell array of names for the predictor variables, in the order in which they appear in `X`.

`ReasonForTermination`

Character vector describing the reason `fitcensemble` stopped adding weak learners to the ensemble.

`Replace`

Logical value indicating if the ensemble was trained with replacement (`true`) or without replacement (`false`).

`ResponseName`

Character vector with the name of the response variable `Y`.

`ScoreTransform`

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitctree`.

Add or change a `ScoreTransform` function using dot notation:

```
ens.ScoreTransform = 'function'
```

or

```
ens.ScoreTransform = @function
```

`Trained`

Trained learners, a cell array of compact classification models.

`TrainedWeights`

Numeric vector of trained weights for the weak learners in `ens`. `TrainedWeights` has `T` elements, where `T` is the number of weak learners in `learners`.

`UseObsForLearner`

Logical matrix of size `N`-by-`NumTrained`, where `N` is the number of observations in the training data and `NumTrained` is the number of trained weak learners. `UseObsForLearner(I,J)` is `true` if observation `I` was used for training learner `J`, and is `false` otherwise.

`W`

Scaled `weights`, a vector with length `n`, the number of rows in `X`. The sum of the elements of `W` is `1`.

`X`

Matrix or table of predictor values that trained the ensemble. Each column of `X` represents one variable, and each row represents one observation.

`Y`

A categorical array, cell array of character vectors, character array, logical vector, or a numeric vector with the same number of rows as `X`. Each row of `Y` represents the classification of the corresponding row of `X`.

Object Functions

compact	Compact classification ensemble
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross validate ensemble
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification error
margin	Classification margins
oobEdge	Out-of-bag classification edge
oobLoss	Out-of-bag classification error
oobMargin	Out-of-bag classification margins
oobPermutedPredictorImportance	Predictor importance estimates by permutation of out-of-bag predictor observations for random forest of classification trees
oobPredict	Predict out-of-bag response of ensemble
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using ensemble of classification models
predictorImportance	Estimates of predictor importance for classification ensemble of decision trees
removeLearners	Remove members of compact classification ensemble
resubEdge	Classification edge by resubstitution
resubLoss	Classification error by resubstitution
resubMargin	Classification margins by resubstitution
resubPredict	Classify observations in ensemble of classification models
resume	Resume training ensemble
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Bagged Ensemble of Classification Trees

Load the `ionosphere` data set.

```
load ionosphere
```

You can train a bagged ensemble of 100 classification trees using all measurements.

```
Mdl = fitcensemble(X,Y,'Method','Bag')
```

`fitcensemble` uses a default template tree object `templateTree()` as a weak learner when `'Method'` is `'Bag'`. In this example, for reproducibility, specify `'Reproducible',true` when you create a tree template object, and then use the object as a weak learner.

```
rng('default') % For reproducibility
t = templateTree('Reproducible',true); % For reproducibiliy of random predictor selections
Mdl = fitcensemble(X,Y,'Method','Bag','Learners',t)
```

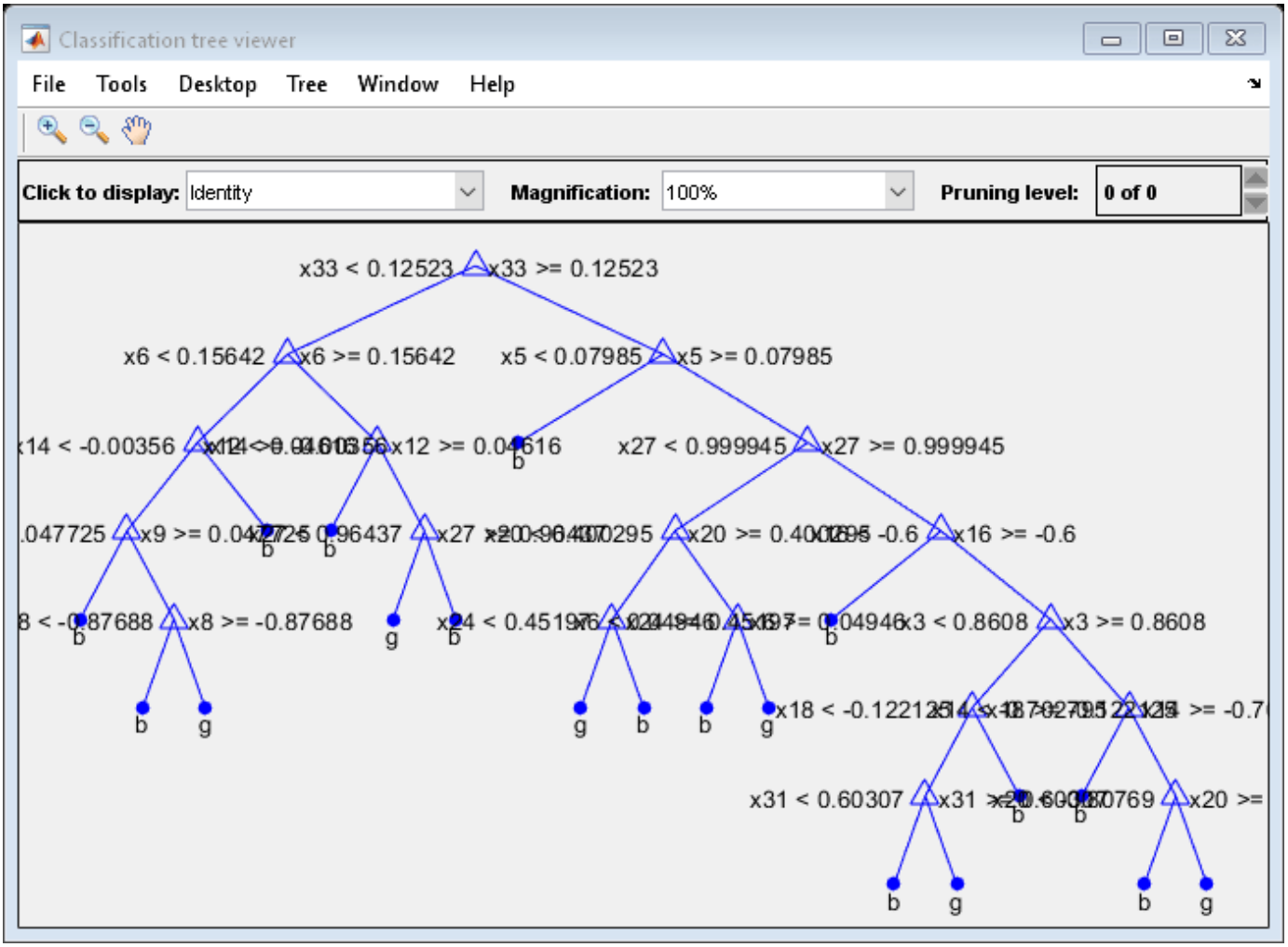
```
Mdl =
  ClassificationBaggedEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
                ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
                NumTrained: 100
                    Method: 'Bag'
              LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: []
       FitInfoDescription: 'None'
                 FResample: 1
                   Replace: 1
          UseObsForLearner: [351x100 logical]


  Properties, Methods
```

`Mdl` is a `ClassificationBaggedEnsemble` model object.

`Mdl.Trained` is the property that stores a 100-by-1 cell vector of the trained classification trees (`CompactClassificationTree` model objects) that compose the ensemble.

Plot a graph of the first trained classification tree.

```
view(Mdl.Trained{1},'Mode','graph')
```



By default, `fitcensemble` grows deep decision trees for bagged ensembles.

Estimate the in-sample misclassification rate.

```
L = resubLoss(Mdl)
```

```
L = 0
```

`L` is 0, which indicates that `Mdl` is perfect at classifying the training data.

Tip

For a bagged ensemble of classification trees, the `Trained` property of `ens` stores a cell vector of `ens.NumTrained` `CompactClassificationTree` model objects. For a textual or graphical display of tree *`t`* in the cell vector, enter

```
view(ens.Trained{t})
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the ClassificationEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitcensemble`, code generation limitations for the weak learners used in the ensemble also apply to the ensemble. For more details, see the Code Generation sections of `CompactClassificationDiscriminant` and `CompactClassificationTree`.
- For fixed-point code generation, you must train an ensemble using tree learners.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationEnsemble` | `compareHoldout` | `fitcensemble` | `fitctree` | `view`

Topics

“Framework for Ensemble Learning” on page 18-31

Introduced in R2011a

ClassificationECOC

Multiclass model for support vector machines (SVMs) and other classifiers

Description

`ClassificationECOC` is an error-correcting output codes (ECOC) classifier on page 33-350 for multiclass learning, where the classifier consists of multiple binary learners such as support vector machines (SVMs). Trained `ClassificationECOC` classifiers store training data, parameter values, prior probabilities, and coding matrices. Use these classifiers to perform tasks such as predicting labels or posterior probabilities for new data (see `predict`).

Creation

Create a `ClassificationECOC` object by using `fitcecoc`.

If you specify linear or kernel binary learners without specifying cross-validation options, then `fitcecoc` returns a `CompactClassificationECOC` object instead.

Properties

After you create a `ClassificationECOC` model object, you can use dot notation to access its properties. For an example, see “Train Multiclass Model Using SVM Learners” on page 33-344.

ECOC Properties

`BinaryLearners` — Trained binary learners
cell vector of model objects

Trained binary learners, specified as a cell vector of model objects. The number of binary learners depends on the number of classes in `Y` and the coding design.

The software trains `BinaryLearner{j}` according to the binary problem specified by `CodingMatrix(:,j)`. For example, for multiclass learning using SVM learners, each element of `BinaryLearners` is a `CompactClassificationSVM` classifier.

Data Types: `cell`

`BinaryLoss` — Binary learner loss function
`'binodeviance'` | `'exponential'` | `'hamming'` | `'hinge'` | `'linear'` | `'logit'` | `'quadratic'`

Binary learner loss function, specified as a character vector representing the loss function name.

If you train using binary learners that use different loss functions, then the software sets `BinaryLoss` to `'hamming'`. To potentially increase accuracy, specify a binary loss function other than the default during a prediction or loss computation by using the `'BinaryLoss'` name-value pair argument of `predict` or `loss`.

Data Types: `char`

`BinaryY` — Binary learner class labels

numeric matrix

Binary learner class labels, specified as a numeric matrix. `BinaryY` is a `NumObservations`-by-*L* matrix, where *L* is the number of binary learners (`length(Mdl.BinaryLearners)`).

Elements of `BinaryY` are `-1`, `0`, or `1`, and the value corresponds to a dichotomous class assignment. This table describes how learner `j` assigns observation `k` to a dichotomous class corresponding to the value of `BinaryY(k,j)`.

Value	Dichotomous Class Assignment
`-1`	Learner `j` assigns observation `k` to a negative class.
`0`	Before training, learner `j` removes observation `k` from the data set.
`1`	Learner `j` assigns observation `k` to a positive class.

Data Types: `double`

`BinEdges` — Bin edges for numeric predictors

cell array of numeric vectors | [ ]

This property is read-only.

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Data Types: `cell`

`CodingMatrix` — Class assignment codes
numeric matrix

Class assignment codes for the binary learners, specified as a numeric matrix. `CodingMatrix` is a *K*-by-*L* matrix, where *K* is the number of classes and *L* is the number of binary learners.

The elements of `CodingMatrix` are `–1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observations in class `i` to a dichotomous class corresponding to the value of `CodingMatrix(i,j)`.

Value	Dichotomous Class Assignment
`–1`	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

Data Types: `double` | `single` | `int8` | `int16` | `int32` | `int64`

`CodingName` — Coding design name
character vector

Coding design name, specified as a character vector. For more details, see “Coding Design” on page 33-351.

Data Types: `char`

`LearnerWeights` — Binary learner weights
numeric row vector

Binary learner weights, specified as a numeric row vector. The length of `LearnerWeights` is equal to the number of binary learners (`length(Mdl.BinaryLearners)`).

`LearnerWeights(j)` is the sum of the observation weights that binary learner `j` uses to train its classifier.

The software uses `LearnerWeights` to fit posterior probabilities by minimizing the Kullback-Leibler divergence. The software ignores `LearnerWeights` when it uses the quadratic programming method of estimating posterior probabilities.

Data Types: `double` | `single`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[ ]`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ModelParameters` — Parameter values
object

Parameter values, such as the name-value pair argument values, used to train the ECOC classifier, specified as an object. `ModelParameters` does not contain estimated parameters.

Access properties of `ModelParameters` using dot notation. For example, list the templates containing parameters of the binary learners by using `Mdl.ModelParameters.BinaryLearner`.

`NumObservations` — Number of observations
positive numeric scalar

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`PredictorNames` — Predictor names

cell array of character vectors

Predictor names in order of their appearance in the predictor data `X`, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns in `X`.

Data Types: `cell`

`Prior` — Prior class probabilities

numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as the number of classes in `ClassNames`, and the order of the elements corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`ResponseName` — Response variable name

character vector

Response variable name, specified as a character vector.

Data Types: `char`

`RowsUsed` — Rows used in fitting

`[]` (default) | logical vector

Rows of the original training data used in fitting the `ClassificationECOC` model, specified as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

`ScoreTransform` — Score transformation function to apply to predicted scores

`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$

Value	Description
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

`W` — Observation weights
numeric vector

Observation weights used to train the ECOC classifier, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

`X` — Unstandardized predictor data
numeric matrix | table

Unstandardized predictor data used to train the ECOC classifier, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

`Y` — Observed class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Observed class labels used to train the ECOC classifier, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements and has the same data type as the input argument `Y` of `fitcecoc`. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding row of `X`.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

Hyperparameter Optimization Properties

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table

This property is read-only.

Cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty if the `'OptimizeHyperparameters'` name-value pair argument is nonempty when you create the model. The value of `HyperparameterOptimizationResults` depends on the setting of the `Optimizer` field in the `HyperparameterOptimizationOptions` structure when you create the model.

Value of `Optimizer` Field	**Value of `HyperparameterOptimizationResults`**
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Object Functions

compact	Reduce size of multiclass error-correcting output codes (ECOC) model
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validate multiclass error-correcting output codes (ECOC) model
discardSupportVectors	Discard support vectors of linear SVM binary learners in ECOC model
edge	Classification edge for multiclass error-correcting output codes (ECOC) model
loss	Classification loss for multiclass error-correcting output codes (ECOC) model
margin	Classification margins for multiclass error-correcting output codes (ECOC) model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using multiclass error-correcting output codes (ECOC) model
resubEdge	Resubstitution classification edge for multiclass error-correcting output codes (ECOC) model
lime	Local interpretable model-agnostic explanations (LIME)
resubLoss	Resubstitution classification loss for multiclass error-correcting output codes (ECOC) model
resubMargin	Resubstitution classification margins for multiclass error-correcting output codes (ECOC) model
resubPredict	Classify observations in multiclass error-correcting output codes (ECOC) model
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Examples

Train Multiclass Model Using SVM Learners

Train a multiclass error-correcting output codes (ECOC) model using support vector machine (SVM) binary learners.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
```

Train a multiclass ECOC model using the default options.

```
Mdl = fitcecoc(X,Y)
```

```
Mdl =
  ClassificationECOC
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
           BinaryLearners: {3x1 cell}
               CodingName: 'onevsone'


  Properties, Methods
```

`Mdl` is a `ClassificationECOC` model. By default, `fitcecoc` uses SVM binary learners and a one-versus-one coding design. You can access `Mdl` properties using dot notation.

Display the class names and the coding design matrix.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
CodingMat = Mdl.CodingMatrix
```

```
CodingMat = 3×3

     1     1     0
    -1     0     1
     0    -1    -1
```

A one-versus-one coding design for three classes yields three binary learners. The columns of `CodingMat` correspond to the learners, and the rows correspond to the classes. The class order is the same as the order in `Mdl.ClassNames`. For example, `CodingMat(:,1)` is `[1; –1; 0]` and indicates that the software trains the first SVM binary learner using all observations classified as `'setosa'` and `'versicolor'`. Because `'setosa'` corresponds to `1`, it is the positive class; `'versicolor'` corresponds to `–1`, so it is the negative class.

You can access each binary learner using cell indexing and dot notation.

```
Mdl.BinaryLearners{1}   % The first binary learner

ans =
  CompactClassificationSVM
            ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [-1 1]
           ScoreTransform: 'none'
                     Beta: [4x1 double]
                     Bias: 1.4505
         KernelParameters: [1x1 struct]


  Properties, Methods
```

Compute the resubstitution classification error.

```
error = resubLoss(Mdl)

error = 0.0067
```

The classification error on the training data is small, but the classifier might be an overfitted model. You can cross-validate the classifier using `crossval` and compute the cross-validation classification error instead.

Inspect Binary Learner Properties of ECOC Classifier

Train an ECOC classifier using SVM binary learners. Then, access properties of the binary learners, such as estimated parameters, by using dot notation.

Load Fisher's iris data set. Specify the petal dimensions as the predictors and the species names as the response.

```
load fisheriris
X = meas(:,3:4);
Y = species;
```

Train an ECOC classifier using SVM binary learners and the default coding design (one-versus-one). Standardize the predictors and save the support vectors.

```
t = templateSVM('Standardize',true,'SaveSupportVectors',true);
predictorNames = {'petalLength','petalWidth'};
responseName = 'irisSpecies';
classNames = {'setosa','versicolor','virginica'}; % Specify class order
Mdl = fitcecoc(X,Y,'Learners',t,'ResponseName',responseName,...
    'PredictorNames',predictorNames,'ClassNames',classNames)

Mdl =
  ClassificationECOC
           PredictorNames: {'petalLength'  'petalWidth'}
             ResponseName: 'irisSpecies'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
```

```
           BinaryLearners: {3x1 cell}
               CodingName: 'onevsone'


  Properties, Methods
```

`t` is a template object that contains options for SVM classification. The function `fitcecoc` uses default values for the empty (`[]`) properties. `Mdl` is a `ClassificationECOC` classifier. You can access properties of `Mdl` using dot notation.

Display the class names and the coding design matrix.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl.CodingMatrix
```

```
ans = 3×3

     1     1     0
    -1     0     1
     0    -1    -1
```

The columns correspond to SVM binary learners, and the rows correspond to the distinct classes. The row order is the same as the order in the `ClassNames` property of `Mdl`. For each column:

- `1` indicates that `fitcecoc` trains the SVM using observations in the corresponding class as members of the positive group.
- `-1` indicates that `fitcecoc` trains the SVM using observations in the corresponding class as members of the negative group.
- `0` indicates that the SVM does not use observations in the corresponding class.

In the first SVM, for example, `fitcecoc` assigns all observations to `'setosa'` or `'versicolor'`, but not `'virginica'`.

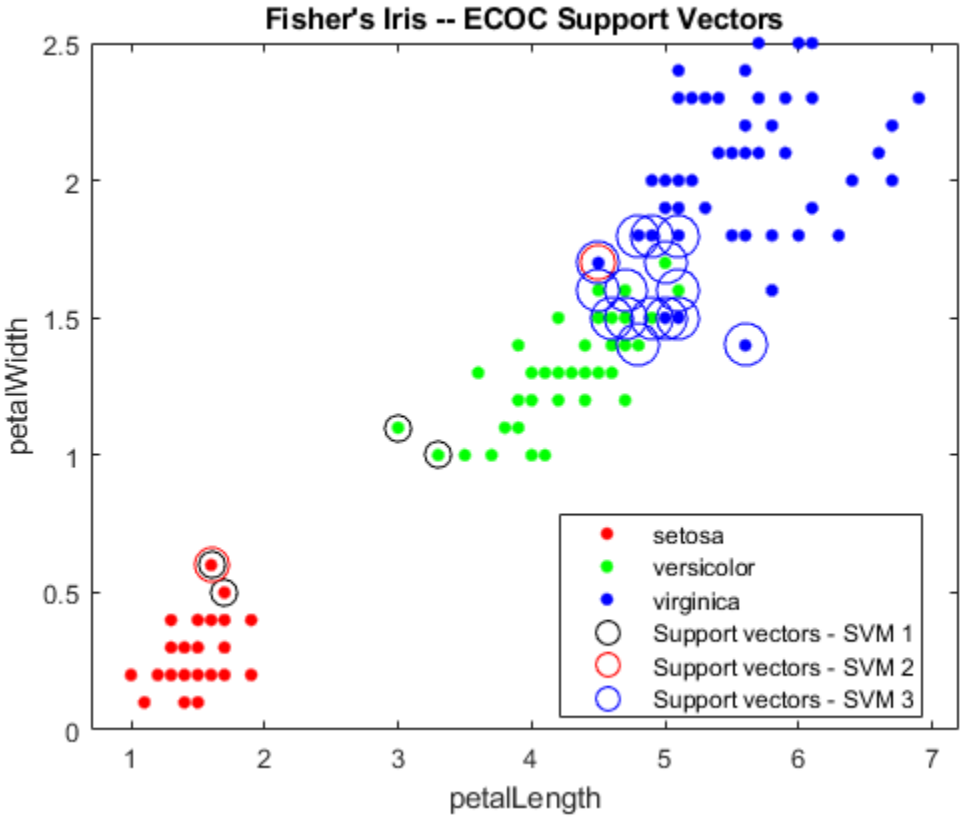
Access properties of the SVMs using cell subscripting and dot notation. Store the standardized support vectors of each SVM. Unstandardize the support vectors.

```
L = size(Mdl.CodingMatrix,2); % Number of SVMs
sv = cell(L,1); % Preallocate for support vector indices
for j = 1:L
    SVM = Mdl.BinaryLearners{j};
    sv{j} = SVM.SupportVectors;
    sv{j} = sv{j}.*SVM.Sigma + SVM.Mu;
end
```

`sv` is a cell array of matrices containing the unstandardized support vectors for the SVMs.

Plot the data, and identify the support vectors.

```
figure
gscatter(X(:,1),X(:,2),Y);
hold on
markers = {'ko','ro','bo'}; % Should be of length L
for j = 1:L
    svs = sv{j};
    plot(svs(:,1),svs(:,2),markers{j},...
        'MarkerSize',10 + (j - 1)*3);
end
title('Fisher''s Iris -- ECOC Support Vectors')
xlabel(predictorNames{1})
ylabel(predictorNames{2})
legend([classNames,{'Support vectors - SVM 1',...
    'Support vectors - SVM 2','Support vectors - SVM 3'}],...
    'Location','Best')
hold off
```



You can pass `Mdl` to these functions:

- `predict`, to classify new observations
- `resubLoss`, to estimate the classification error on the training data
- `crossval`, to perform 10-fold cross-validation

Cross-Validate ECOC Classifier

Cross-validate an ECOC classifier with SVM binary learners, and estimate the generalized classification error.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
rng(1); % For reproducibility
```

Create an SVM template, and standardize the predictors.

```
t = templateSVM('Standardize',true)
```

```
t =
Fit template for classification SVM.

                     Alpha: [0x1 double]
           BoxConstraint: []
               CacheSize: []
           CachingMethod: ''
              ClipAlphas: []
  DeltaGradientTolerance: []
                 Epsilon: []
           GapTolerance: []
           KKTTolerance: []
          IterationLimit: []
          KernelFunction: ''
             KernelScale: []
           KernelOffset: []
   KernelPolynomialOrder: []
                 NumPrint: []
                       Nu: []
         OutlierFraction: []
       RemoveDuplicates: []
        ShrinkagePeriod: []
                  Solver: ''
        StandardizeData: 1
     SaveSupportVectors: []
         VerbosityLevel: []
                Version: 2
                 Method: 'SVM'
                   Type: 'classification'
```

`t` is an SVM template. Most of the template object properties are empty. When training the ECOC classifier, the software sets the applicable properties to their default values.

Train the ECOC classifier, and specify the class order.

```
Mdl = fitcecoc(X,Y,'Learners',t,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a `ClassificationECOC` classifier. You can access its properties using dot notation.

Cross-validate `Mdl` using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

`CVMdl` is a `ClassificationPartitionedECOC` cross-validated ECOC classifier.

Estimate the generalized classification error.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.0400
```

The generalized classification error is 4%, which indicates that the ECOC classifier generalizes fairly well.

More About

Error-Correcting Output Codes Model

An error-correcting output codes (ECOC) model reduces the problem of classification with three or more classes to a set of binary classification problems.

ECOC classification requires a coding design, which determines the classes that the binary learners train on, and a decoding scheme, which determines how the results (predictions) of the binary classifiers are aggregated.

Assume the following:

- The classification problem has three classes.
- The coding design is one-versus-one. For three classes, this coding design is

	Learner 1	Learner 2	Learner 3
Class 1	1	1	0
Class 2	−1	0	1
Class 3	0	−1	−1

- The decoding scheme uses loss g.
- The learners are SVMs.

To build this classification model, the ECOC algorithm follows these steps.

1. Learner 1 trains on observations in Class 1 or Class 2, and treats Class 1 as the positive class and Class 2 as the negative class. The other learners are trained similarly.
2. Let M be the coding design matrix with elements m_{kl}, and s_l be the predicted classification score for the positive class of learner l. The algorithm assigns a new observation to the class ($\widehat{k}$) that minimizes the aggregation of the losses for the L binary learners.

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{l=1}^{L} |m_{kl}| g(m_{kl}, s_l)}{\sum_{l=1}^{L} |m_{kl}|}.$$

ECOC models can improve classification accuracy, compared to other multiclass models [2].

Coding Design

A coding design is a matrix where elements direct which classes are trained by each binary learner, that is, how the multiclass problem is reduced to a series of binary problems.

Each row of the coding design corresponds to a distinct class, and each column corresponds to a binary learner. In a ternary coding design, for a particular column (or binary learner):

- A row containing 1 directs the binary learner to group all observations in the corresponding class into a positive class.
- A row containing –1 directs the binary learner to group all observations in the corresponding class into a negative class.
- A row containing 0 directs the binary learner to ignore all observations in the corresponding class.

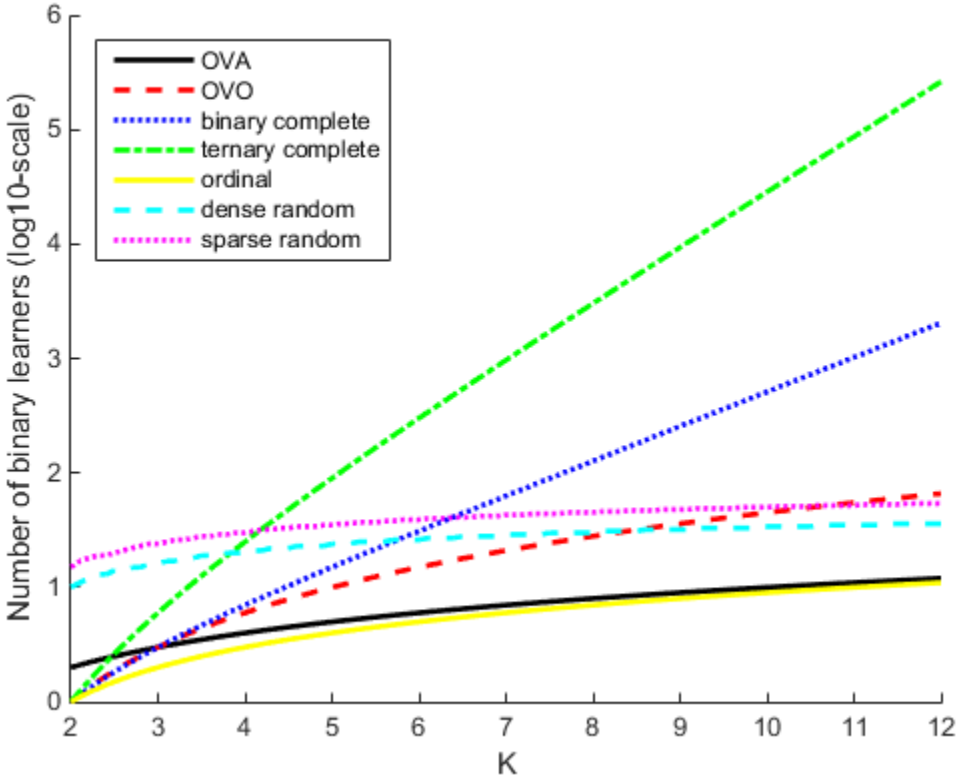
Coding design matrices with large, minimal, pairwise row distances based on the Hamming measure are optimal. For details on the pairwise row distance, see “Random Coding Design Matrices” on page 33-1646 and [4].

This table describes popular coding designs.

Coding Design	Description	Number of Learners	Minimal Pairwise Row Distance
one-versus-all (OVA)	For each binary learner, one class is positive and the rest are negative. This design exhausts all combinations of positive class assignments.	K	2
one-versus-one (OVO)	For each binary learner, one class is positive, another is negative, and the rest are ignored. This design exhausts all combinations of class pair assignments.	$K(K – 1)/2$	1
binary complete	This design partitions the classes into all binary combinations, and does not ignore any classes. That is, all class assignments are −1 and 1 with at least one positive class and one negative class in the assignment for each binary learner.	$2^{K-1} – 1$	2^{K-2}

Coding Design	Description	Number of Learners	Minimal Pairwise Row Distance
ternary complete	This design partitions the classes into all ternary combinations. That is, all class assignments are `0`, `−1`, and `1` with at least one positive class and one negative class in the assignment for each binary learner.	$(3^K - 2^{K+1} + 1)/2$	3^{K-2}
ordinal	For the first binary learner, the first class is negative and the rest are positive. For the second binary learner, the first two classes are negative and the rest are positive, and so on.	$K - 1$	1
dense random	For each binary learner, the software randomly assigns classes into positive or negative classes, with at least one of each type. For more details, see “Random Coding Design Matrices” on page 33-1646.	Random, but approximately $10 \log_2 K$	Variable
sparse random	For each binary learner, the software randomly assigns classes as positive or negative with probability 0.25 for each, and ignores classes with probability 0.5. For more details, see “Random Coding Design Matrices” on page 33-1646.	Random, but approximately $15 \log_2 K$	Variable

This plot compares the number of binary learners for the coding designs with increasing K.



Algorithms

Random Coding Design Matrices

For a given number of classes K, the software generates random coding design matrices as follows.

1. The software generates one of these matrices:
   a. Dense random — The software assigns 1 or –1 with equal probability to each element of the K-by-L_d coding design matrix, where $L_d \approx \lceil 10\log_2 K \rceil$.
   b. Sparse random — The software assigns 1 to each element of the K-by-L_s coding design matrix with probability 0.25, –1 with probability 0.25, and 0 with probability 0.5, where $L_s \approx \lceil 15\log_2 K \rceil$.
2. If a column does not contain at least one 1 and at least one –1, then the software removes that column.
3. For distinct columns u and v, if $u = v$ or $u = -v$, then the software removes v from the coding design matrix.

The software randomly generates 10,000 matrices by default, and retains the matrix with the largest, minimal, pairwise row distance based on the Hamming measure ([4]) given by

$$\Delta(k_1, k_2) = 0.5 \sum_{l=1}^{L} \left|m_{k_1 l}\right| \left|m_{k_2 l}\right| \left|m_{k_1 l} - m_{k_2 l}\right|,$$

where $m_{k,l}$ is an element of coding design matrix j.

Support Vector Storage

By default and for efficiency, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties for all linear SVM binary learners. `fitcecoc` lists `Beta`, rather than `Alpha`, in the model display.

To store `Alpha`, `SupportVectorLabels`, and `SupportVectors`, pass a linear SVM template that specifies storing support vectors to `fitcecoc`. For example, enter:

```
t = templateSVM('SaveSupportVectors',true)
Mdl = fitcecoc(X,Y,'Learners',t);
```

You can remove the support vectors and related values by passing the resulting `ClassificationECOC` model to `discardSupportVectors`.

Alternative Functionality

You can use these alternative algorithms to train a multiclass model:

- Classification ensembles—see `fitcensemble` and `ClassificationEnsemble`
- Classification trees—see `fitctree` and `ClassificationTree`
- Discriminant analysis classifiers—see `fitcdiscr` and `ClassificationDiscriminant`
- *k*-nearest neighbor classifiers—see `fitcknn` and `ClassificationKNN`
- Naive Bayes classifiers—see `fitcnb` and `ClassificationNaiveBayes`

References

[1] Fürnkranz, Johannes. “Round Robin Classification.” *Journal of Machine Learning Research*, Vol. 2, 2002, pp. 721–747.

[2] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recognition Letters*, Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- When you train an ECOC model by using `fitcecoc`, the following restrictions apply.
  - You cannot fit posterior probabilities by using the `'FitPosterior'` name-value pair argument.
  - All binary learners must be either SVM classifiers or linear classification models. For the `'Learners'` name-value pair argument, you can specify:
    - `'svm'` or `'linear'`

- An SVM template object or a cell array of such objects (see `templateSVM`)
- A linear classification model template object or a cell array of such objects (see `templateLinear`)

- When you generate code using a coder configurer for `predict` and `update`, the following additional restrictions apply for binary learners.

  - If you use a cell array of SVM template objects, the value of `'Standardize'` for SVM learners must be consistent. For example, if you specify `'Standardize',true` for one SVM learner, you must specify the same value for all SVM learners.
  - If you use a cell array of SVM template objects, and you use one SVM learner with a linear kernel (`'KernelFunction','linear'`) and another with a different type of kernel function, then you must specify `'SaveSupportVectors',true` for the learner with a linear kernel.

  For details, see `ClassificationECOCCoderConfigurer`. For information on name-value pair arguments that you cannot modify when you retrain a model, see “Tips” on page 33-6501.
- Code generation limitations for SVM classifiers and linear classification models also apply to ECOC classifiers, depending on the choice of binary learners. For more details, see “Code Generation” on page 33-807 of the `CompactClassificationSVM` class and “Code Generation” on page 33-419 of the `ClassificationLinear` class.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationPartitionedECOC` | `CompactClassificationECOC` | `fitcecoc` | `fitcsvm`

Introduced in R2014b

ClassificationECOCCoderConfigurer

Coder configurer for multiclass model using binary learners

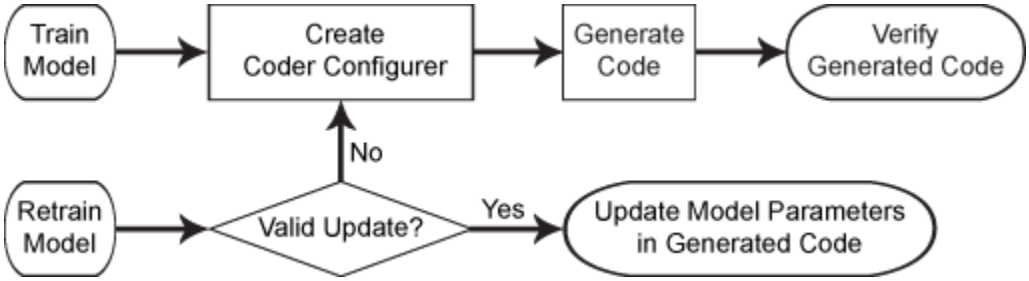
Description

A `ClassificationECOCCoderConfigurer` object is a coder configurer of a multiclass error-correcting output codes (ECOC) classification model (`ClassificationECOC` or `CompactClassificationECOC`) that uses support vector machine (SVM) or linear binary learners.

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes of model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the ECOC model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of a multiclass ECOC classification model, see the Code Generation sections of `CompactClassificationECOC`, `predict`, and `update`.

Creation

After training a multiclass ECOC classification model with SVM or linear binary learners by using `fitcecoc`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

`X` — Coder attributes of predictor data

`LearnerCoderInput` object

Coder attributes of predictor data to pass to the generated C/C++ code for the `predict` function of the ECOC classification model, specified as a `LearnerCoderInput` on page 33-371 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.

  - If the `Value` attribute of the `ObservationsIn` property for the `ClassificationECOCCoderConfigurer` is `'rows'`, then this `SizeVector` value is `[n p]`, where `n` corresponds to the number of observations and `p` corresponds to the number of predictors.
  - If the `Value` attribute of the `ObservationsIn` property for the `ClassificationECOCCoderConfigurer` is `'columns'`, then this `SizeVector` value is `[p n]`.

  To switch the elements of `SizeVector` (for example, to change `[n p]` to `[p n]`), modify the `Value` attribute of the `ObservationsIn` property for the `ClassificationECOCCoderConfigurer` accordingly. You cannot modify the `SizeVector` value directly.
- `VariableDimensions` — The default value is `[0 0]`, which indicates that the array size is fixed as specified in `SizeVector`.

  You can set this value to `[1 0]` if the `SizeVector` value is `[n p]` or to `[0 1]` if it is `[p n]`, which indicates that the array has variable-size rows and fixed-size columns. For example, `[1 0]` specifies that the first value of `SizeVector` (`n`) is the upper bound for the number of rows, and the second value of `SizeVector` (`p`) is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations (in rows) of three predictor variables (in columns), specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of X:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of X (number of observations) has a variable size and the second dimension of X (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

`BinaryLoss` — Coder attributes of binary learner loss function
`EnumeratedInput` object

Coder attributes of the binary learner loss function (`'BinaryLoss'` name-value pair argument of `predict`), specified as an `EnumeratedInput` on page 33-372 object.

The default attribute values of the `EnumeratedInput` object are based on the default values of the `predict` function:

- `Value` — Binary learner loss function, specified as one of the character vectors in `BuiltInOptions` or a character vector designating a custom function name. If the binary learners are SVMs or linear classification models of SVM learners, the default value is `'hinge'`. If the binary learners are linear classification models of logistic regression learners, the default value is `'quadratic'`.

  To use a custom option, define a custom function on the MATLAB search path, and specify `Value` as the name of the custom function.
- `SelectedOption` — This value is `'Built-in'`(default) or `'Custom'`. The software sets `SelectedOption` according to `Value`. This attribute is read-only.
- `BuiltInOptions` — Cell array of `'hamming'`, `'linear'`, `'quadratic'`, `'exponential'`, `'binodeviance'`, `'hinge'`, and `'logit'`. This attribute is read-only.
- `IsConstant` — This value must be `true`.
- `Tunability` — The default value is `false`. If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

`Decoding` — Coder attributes of decoding scheme
`EnumeratedInput` object

Coder attributes of the decoding scheme (`'Decoding'` name-value pair argument of `predict`), specified as an `EnumeratedInput` on page 33-372 object.

The default attribute values of the `EnumeratedInput` object are based on the default values of the `predict` function:

- `Value` — Decoding scheme value, specified as `'lossweighted'`(default), `'lossbased'`, or a `LearnerCoderInput` on page 33-371 object.

  If you set `IsConstant` to `false`, then the software changes `Value` to a `LearnerCoderInput` on page 33-371 object with these read-only coder attribute values:

  - `SizeVector` — `[1 12]`

- `VariableDimensions` — `[0 1]`
- `DataType` — `'char'`
- `Tunability` — 1

The input in the generated code is a variable-size, tunable character vector that is either `'lossweighted'` or `'lossbased'`.

- `SelectedOption` — This value is `'Built-in'`(default) or `'NonConstant'`. The software sets `SelectedOption` according to `Value`. This attribute is read-only.
- `BuiltInOptions` — Cell array of `'lossweighted'` and `'lossbased'`. This attribute is read-only.
- `IsConstant` — The default value is `true`. If you set this value to `false`, the software changes `Value` to a `LearnerCoderInput` object.
- `Tunability` — The default value is `false`. If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

`ObservationsIn` — Coder attributes of predictor data observation dimension
`EnumeratedInput` object

Coder attributes of the predictor data observation dimension (`'ObservationsIn'` name-value pair argument of `predict`), specified as an `EnumeratedInput` on page 33-372 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the `'ObservationsIn'` name-value pair argument determines the default values of the `EnumeratedInput` coder attributes:

- `Value` — The default value is the predictor data observation dimension you use when creating the coder configurer, specified as `'rows'` or `'columns'`. If you do not specify `'ObservationsIn'` when creating the coder configurer, the default value is `'rows'`.

  This value must be `'rows'` for a model that uses SVM binary learners.
- `SelectedOption` — This value is always `'Built-in'`. This attribute is read-only.
- `BuiltInOptions` — Cell array of `'rows'` and `'columns'`. This attribute is read-only.
- `IsConstant` — This value must be `true`.
- `Tunability` — The default value is `false` if you specify `'ObservationsIn','rows'` when creating the coder configurer, and `true` if you specify `'ObservationsIn','columns'`. If you set `Tunability` to `false`, the software sets `Value` to `'rows'`. If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

`NumOutputs` — Number of outputs in `predict`
1 (default) | 2 | 3

Number of output arguments to return from the generated C/C++ code for the `predict` function of the ECOC classification model, specified as 1, 2, or 3.

The output arguments of `predict` are, in order: `label` (predicted class labels), `NegLoss` (negated average binary losses), and `PBScore` (positive-class scores). `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function, where `n` is the `NumOutputs` value.

After creating the coder configurer `configurer`, you can specify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of an ECOC classification model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-371 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`BinaryLearners` — Coder attributes of trained binary learners
`ClassificationSVMCoderConfigurer` object | `ClassificationLinearCoderConfigurer` object

Coder attributes of the trained binary learners (`BinaryLearners` of an ECOC classification model), specified as a `ClassificationSVMCoderConfigurer` object (for SVM binary learners) or a `ClassificationLinearCoderConfigurer` object (for linear binary learners).

Use the `update` arguments of the SVM or linear coder configurer object to specify the coder attributes of all binary learners.

- For `ClassificationSVMCoderConfigurer`, the `update` arguments are `Alpha`, `Beta`, `Bias`, `Cost`, `Mu`, `Prior`, `Scale`, `Sigma`, `SupportVectorLabels`, and `SupportVectors`.
- For `ClassificationLinearCoderConfigurer`, the `update` arguments are `Beta`, `Bias`, `Cost`, and `Prior`.

For the configuration of `BinaryLearners`, the software uses only the `update` argument properties and ignores the other properties of the object.

When you train an ECOC model with SVM binary learners, each learner can have a different number of support vectors. Therefore, the software configures the default attribute values of the `LearnerCoderInput` on page 33-371 objects for `Alpha`, `SupportVectorLabels`, and `SupportVectors` to accommodate all binary learners, based on the input argument `Mdl` of `learnerCoderConfigurer`.

- `SizeVector`
  - This value is `[s 1]` for `Alpha` and `SupportVectorLabels`, where `s` is the largest number of support vectors in the binary learners.
  - This value is `[s p]` for `SupportVectors`, where `p` is the number of predictors.
- `VariableDimensions` — This value is `[0 0]` or `[1 0]`. If each learner has the same number of support vectors, the default value is `[0 0]`. Otherwise, this value must be `[1 0]`.

  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function, and the model stores the linear predictor coefficients (`Beta`) without the support vectors and related values, then this value must be `false`. Otherwise, this value must be `true`.

For details about the other `update` arguments, see `update` arguments on page 33-548 of `ClassificationSVMCoderConfigurer` and `update` arguments on page 33-423 of `ClassificationLinearCoderConfigurer`.

`Cost` — Coder attributes of misclassification cost
`LearnerCoderInput` object

Coder attributes of the misclassification cost (`Cost` of an ECOC classification model), specified as a `LearnerCoderInput` on page 33-371 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[c c]`, where `c` is the number of classes.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

`Prior` — Coder attributes of prior probabilities
`LearnerCoderInput` object

Coder attributes of the prior probabilities (`Prior` of an ECOC classification model), specified as a `LearnerCoderInput` on page 33-371 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 c]`, where `c` is the number of classes.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code
`'ClassificationECOCModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `ClassificationECOCCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level
`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — List of tunable input arguments of `predict`
cell array

This property is read-only.

List of tunable input arguments of the entry-point function `predict.m` for code generation, specified as a cell array. The cell array contains another cell array that includes `coder.PrimitiveType` objects and `coder.Constant` objects.

If you modify the coder attributes of `predict` arguments on page 33-357, then the software updates the corresponding objects accordingly. If you specify the `Tunability` attribute as `false`, then the software removes the corresponding objects from the `PredictInputs` list.

The cell array in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure. The structure includes a `coder.CellType` object for `BinaryLearners` and `coder.PrimitiveType` objects for `Cost` and `Prior`.

If you modify the coder attributes of `update` arguments on page 33-360, then the software updates the corresponding objects accordingly. If you specify the `Tunability` attribute as `false`, then the software removes the corresponding object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load Fisher's iris data set and train a multiclass ECOC model using SVM binary learners.

```
load fisheriris
X = meas;
Y = species;
Mdl = fitcecoc(X,Y);
```

`Mdl` is a `ClassificationECOC` object.

Create a coder configurer for the `ClassificationECOC` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

`configurer` is a `ClassificationECOCCoderConfigurer` object, which is a coder configurer of a `ClassificationECOC` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the ECOC classification model (`Mdl`) with default settings.

```
generateCode(configurer)

generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationECOCModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationECOCModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationECOCModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m

function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:00:25
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end

type update.m

function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:00:25
initialize('update',varargin{:});
end

type initialize.m

function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:00:25
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('ClassificationECOCModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: BinaryLearners
        %                       Prior
        %                       Cost
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
```

```
        end
end
end
```

Update Parameters of ECOC Classification Model in Generated Code

Train an error-correcting output codes (ECOC) model using SVM binary learners and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the ECOC model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using different settings, and update parameters in the generated code without regenerating the code.

Train Model

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create an SVM binary learner template to use a Gaussian kernel function and to standardize predictor data.

```
t = templateSVM('KernelFunction','gaussian','Standardize',true);
```

Train a multiclass ECOC model using the template `t`.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

`Mdl` is a `ClassificationECOC` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationECOC` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns the first two outputs of the `predict` function, which are the predicted labels and negated average binary losses.

```
configurer = learnerCoderConfigurer(Mdl,X,'NumOutputs',2)
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'
```

```
  Properties, Methods
```

`configurer` is a `ClassificationECOCCoderConfigurer` object, which is a coder configurer of a `ClassificationECOC` object. The display shows the tunable input arguments of `predict` and `update`: `X`, `BinaryLearners`, `Prior`, and `Cost`.

Specify Coder Attributes of Parameters

Specify the coder attributes of `predict` arguments (predictor data and the name-value pair arguments `'Decoding'` and `'BinaryLoss'`) and `update` arguments (support vectors of the SVM learners) so that you can use these arguments as the input arguments of `predict` and `update` in the generated code.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 4];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 4 predictors, so the second value of the `SizeVector` attribute must be 4 and the second value of the `VariableDimensions` attribute must be `false`.

Next, modify the coder attributes of `BinaryLoss` and `Decoding` to use the `'BinaryLoss'` and `'Decoding'` name-value pair arguments in the generated code. Display the coder attributes of `BinaryLoss`.

```
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

            Value: 'hinge'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 0
```

To use a nondefault value in the generated code, you must specify the value before generating the code. Specify the `Value` attribute of `BinaryLoss` as `'exponential'`.

```
configurer.BinaryLoss.Value = 'exponential';
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:
```

```
             Value: 'exponential'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 1
```

If you modify attribute values when `Tunability` is `false` (logical 0), the software sets the `Tunability` to `true` (logical 1).

Display the coder attributes of `Decoding`.

```
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: 'lossweighted'
    SelectedOption: 'Built-in'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 1
        Tunability: 0
```

Specify the `IsConstant` attribute of `Decoding` as `false` so that you can use all available values in `BuiltInOptions` in the generated code.

```
configurer.Decoding.IsConstant = false;
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: [1x1 LearnerCoderInput]
    SelectedOption: 'NonConstant'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 0
        Tunability: 1
```

The software changes the `Value` attribute of `Decoding` to a `LearnerCoderInput` object so that you can use both `'lossweighted'` and `'lossbased'` as the value of `'Decoding'`. Also, the software sets the `SelectedOption` to `'NonConstant'` and the `Tunability` to `true`.

Finally, modify the coder attributes of `SupportVectors` in `BinaryLearners`. Display the coder attributes of `SupportVectors`.

```
configurer.BinaryLearners.SupportVectors
```

```
ans =
  LearnerCoderInput with properties:

            SizeVector: [54 4]
    VariableDimensions: [1 0]
              DataType: 'double'
            Tunability: 1
```

The default value of `VariableDimensions` is `[true false]` because each learner has a different number of support vectors. If you retrain the ECOC model using new data or different settings, the number of support vectors in the SVM learners can vary. Therefore, increase the upper bound of the number of support vectors.

```
configurer.BinaryLearners.SupportVectors.SizeVector = [150 4];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrain
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Display the coder configurer.

```
configurer
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
        BinaryLoss: [1x1 EnumeratedInput]
          Decoding: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

The display now includes `BinaryLoss` and `Decoding` as well.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the ECOC classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationECOCModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationECOCModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationECOCModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument. Because you specified `'Decoding'` as a tunable input argument by changing the `IsConstant` attribute before generating the code, you also need to specify it in the call to the MEX function, even though `'lossweighted'` is the default value of `'Decoding'`.

```
[label,NegLoss] = predict(Mdl,X,'BinaryLoss','exponential');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
```

Compare `label` to `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`NegLoss_mex` might include round-off differences compared to `NegLoss`. In this case, compare `NegLoss_mex` to `NegLoss`, allowing a small tolerance.

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `NegLoss` and `NegLoss_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using a different setting. Specify `'KernelScale'` as `'auto'` so that the software selects an appropriate scale factor using a heuristic procedure.

```
t_new = templateSVM('KernelFunction','gaussian','Standardize',true,'KernelScale','auto');
retrainedMdl = fitcecoc(X,Y,'Learners',t_new);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationECOCModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` to the outputs from the `predict` function in the updated MEX function.

```
[label,NegLoss] = predict(retrainedMdl,X,'BinaryLoss','exponential','Decoding','lossbased');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `label` and `label_mex` are equal, and `NegLoss` and `NegLoss_mex` are equal within the tolerance.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array

Attribute Name	Description
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of the coefficients `Alpha` in `BinaryLearners` of the coder configurer `configurer`:

```
configurer.BinaryLearners.Alpha.SizeVector = [100 1];
configurer.BinaryLearners.Alpha.VariableDimensions = [1 0];
configurer.BinaryLearners.Alpha.DataType = 'double';
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

`EnumeratedInput` Object

A coder configurer uses an `EnumeratedInput` object to specify the coder attributes of `predict` input arguments that have a finite set of available values.

An `EnumeratedInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description

`Value`	Value of the `predict` argument in the generated code, specified as a character vector or a `LearnerCoderInput` on page 33-371 object.
	• Character vector in `BuiltInOptions` — You can specify one of the `BuiltInOptions` using either the option name or its index value. For example, to choose the first option, specify `Value` as either the first character vector in `BuiltInOptions` or `1`.
	• Character vector designating a custom function name — To use a custom option, define a custom function on the MATLAB search path, and specify `Value` as the name of the custom function.
	• `LearnerCoderInput` on page 33-371 object — If you set `IsConstant` to `false` (logical 0), then the software changes `Value` to a `LearnerCoderInput` on page 33-371 object with the following read-only coder attribute values. These values indicate that the input in the generated code is a variable-size, tunable character vector that is one of the available values in `BuiltInOptions`.
	  • `SizeVector` — `[1 c]`, indicating the upper bound of the array size, where `c` is the length of the longest available character vector in `Option`
	  • `VariableDimensions` — `[0 1]`, indicating that the array is a variable-size vector
	  • `DataType` — `'char'`
	  • `Tunability` — 1
	The default value of `Value` is consistent with the default value of the corresponding `predict` argument, which is one of the character vectors in `BuiltInOptions`.

`SelectedOption`	Status of the selected option, specified as `'Built-in'`, `'Custom'`, or `'NonConstant'`. The software sets `SelectedOption` according to `Value`: • `'Built-in'`(default) — When `Value` is one of the character vectors in `BuiltInOptions` • `'Custom'` — When `Value` is a character vector that is not in `BuiltInOptions` • `'NonConstant'` — When `Value` is a `LearnerCoderInput` on page 33-371 object This attribute is read-only.
`BuiltInOptions`	List of available character vectors for the corresponding `predict` argument, specified as a cell array. This attribute is read-only.

Attribute Name	Description
`IsConstant`	Indicator specifying whether or not the array value is a compile-time constant (`coder.Constant`) in the generated code, specified as `true` (logical 1, default) or `false` (logical 0). If you set this value to `false`, then the software changes `Value` to a `LearnerCoderInput` on page 33-371 object.
`Tunability`	Indicator specifying whether or not `predict` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0, default). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of `BinaryLoss` of the coder configurer `configurer`:

```
configurer.BinaryLoss.Value = 'linear';
```

See Also

`ClassificationECOC` | `ClassificationLinearCoderConfigurer` | `ClassificationSVMCoderConfigurer` | `CompactClassificationECOC` | `learnerCoderConfigurer` | `predict` | `update`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2019a

ClassificationDiscriminant class

Superclasses: `CompactClassificationDiscriminant`

Discriminant analysis classification

Description

A `ClassificationDiscriminant` object encapsulates a discriminant analysis classifier, which is a Gaussian mixture model for data generation. A `ClassificationDiscriminant` object can predict responses for new data using the `predict` method. The object contains the data used for training, so can compute resubstitution predictions.

Construction

Create a `ClassificationDiscriminant` object by using `fitcdiscr`.

Properties

`BetweenSigma`

`p`-by-`p` matrix, the between-class covariance, where `p` is the number of predictors.

`CategoricalPredictors`

Categorical predictor indices, which is always empty (`[]`) .

`ClassNames`

List of the elements in the training data `Y` with duplicates removed. `ClassNames` can be a categorical array, cell array of character vectors, character array, logical vector, or a numeric vector. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`Coeffs`

`k`-by-`k` structure of coefficient matrices, where `k` is the number of classes. `Coeffs(i,j)` contains coefficients of the linear or quadratic boundaries between classes `i` and `j`. Fields in `Coeffs(i,j)`:

- `DiscrimType`
- `Class1` — `ClassNames(i)`
- `Class2` — `ClassNames(j)`
- `Const` — A scalar
- `Linear` — A vector with `p` components, where `p` is the number of columns in `X`
- `Quadratic` — `p`-by-`p` matrix, exists for quadratic `DiscrimType`

The equation of the boundary between class `i` and class `j` is

`Const + Linear * x + x' * Quadratic * x = 0`,

where x is a column vector of length p.

If fitcdiscr had the FillCoeffs name-value pair set to 'off' when constructing the classifier, Coeffs is empty ([]).

Cost

Square matrix, where Cost(i,j) is the cost of classifying a point into class j if its true class is i (i.e., the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of Cost corresponds to the order of the classes in ClassNames. The number of rows and columns in Cost is the number of unique classes in the response.

Change a Cost matrix using dot notation: obj.Cost = costMatrix.

Delta

Value of the Delta threshold for a linear discriminant model, a nonnegative scalar. If a coefficient of obj has magnitude smaller than Delta, obj sets this coefficient to 0, and so you can eliminate the corresponding predictor from the model. Set Delta to a higher value to eliminate more predictors.

Delta must be 0 for quadratic discriminant models.

Change Delta using dot notation: obj.Delta = newDelta.

DeltaPredictor

Row vector of length equal to the number of predictors in obj. If DeltaPredictor(i) < Delta then coefficient i of the model is 0.

If obj is a quadratic discriminant model, all elements of DeltaPredictor are 0.

DiscrimType

Character vector specifying the discriminant type. One of:

- 'linear'
- 'quadratic'
- 'diagLinear'
- 'diagQuadratic'
- 'pseudoLinear'
- 'pseudoQuadratic'

Change DiscrimType using dot notation: obj.DiscrimType = newDiscrimType.

You can change between linear types, or between quadratic types, but cannot change between linear and quadratic types.

Gamma

Value of the Gamma regularization parameter, a scalar from 0 to 1. Change Gamma using dot notation: obj.Gamma = newGamma.

- If you set 1 for linear discriminant, the discriminant sets its type to 'diagLinear'.

- If you set a value between `MinGamma` and `1` for linear discriminant, the discriminant sets its type to `'linear'`.
- You cannot set values below the value of the `MinGamma` property.
- For quadratic discriminant, you can set either `0` (for `DiscrimType 'quadratic'`) or `1` (for `DiscrimType 'diagQuadratic'`).

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`LogDetSigma`

Logarithm of the determinant of the within-class covariance matrix. The type of `LogDetSigma` depends on the discriminant type:

- Scalar for linear discriminant analysis
- Vector of length `K` for quadratic discriminant analysis, where `K` is the number of classes

`MinGamma`

Nonnegative scalar, the minimal value of the Gamma parameter so that the correlation matrix is invertible. If the correlation matrix is not singular, `MinGamma` is `0`.

`ModelParameters`

Parameters used in training `obj`.

`Mu`

Class means, specified as a `K`-by-`p` matrix of scalar values class means of size. `K` is the number of classes, and `p` is the number of predictors. Each row of `Mu` represents the mean of the multivariate normal distribution of the corresponding class. The class indices are in the `ClassNames` attribute.

`NumObservations`

Number of observations in the training data, a numeric scalar. `NumObservations` can be less than the number of rows of input data `X` when there are missing values in `X` or response `Y`.

`PredictorNames`

Cell array of names for the predictor variables, in the order in which they appear in the training data `X`.

`Prior`

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`.

Add or change a `Prior` vector using dot notation: `obj.Prior = priorVector`.

`ResponseName`

Character vector describing the response variable `Y`.

`ScoreTransform`

Character vector representing a built-in transformation function, or a function handle for transforming scores. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitcdiscr`.

Implement dot notation to add or change a `ScoreTransform` function using one of the following:

- `cobj.ScoreTransform = '`*`function`*`'`
- `cobj.ScoreTransform = @`*`function`*

`Sigma`

Within-class covariance matrix or matrices. The dimensions depend on `DiscrimType`:

- `'linear'` (default) — Matrix of size `p`-by-`p`, where `p` is the number of predictors
- `'quadratic'` — Array of size `p`-by-`p`-by-`K`, where `K` is the number of classes
- `'diagLinear'` — Row vector of length `p`
- `'diagQuadratic'` — Array of size `1`-by-`p`-by-`K`
- `'pseudoLinear'` — Matrix of size `p`-by-`p`
- `'pseudoQuadratic'` — Array of size `p`-by-`p`-by-`K`

`W`

Scaled `weights`, a vector with length `n`, the number of rows in `X`.

`X`

Matrix of predictor values. Each column of `X` represents one predictor (variable), and each row represents one observation.

`Xcentered`

`X` data with class means subtracted. If `Y(i)` is of class `j`,

`Xcentered(i,:) = X(i,:) - Mu(j,:)`,

where `Mu` is the class mean property.

`Y`

A categorical array, cell array of character vectors, character array, logical vector, or a numeric vector with the same number of rows as `X`. Each row of `Y` represents the classification of the corresponding row of `X`.

Object Functions

compact	Compact discriminant analysis classifier
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validated discriminant analysis classifier
cvshrink	Cross-validate regularization of linear discriminant
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
logp	Log unconditional probability density for discriminant analysis classifier
loss	Classification error
mahal	Mahalanobis distance to class means
margin	Classification margins
nLinearCoeffs	Number of nonzero linear coefficients
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels using discriminant analysis classification model
resubEdge	Classification edge by resubstitution
resubLoss	Classification error by resubstitution
resubMargin	Classification margins by resubstitution
resubPredict	Predict resubstitution labels of discriminant analysis classification model
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Discriminant Analysis Model

Load Fisher's iris data set.

```
load fisheriris
```

Train a discriminant analysis model using the entire data set.

```
Mdl = fitcdiscr(meas,species)
```

```
Mdl =
  ClassificationDiscriminant
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
          ScoreTransform: 'none'
         NumObservations: 150
             DiscrimType: 'linear'
                      Mu: [3x4 double]
                  Coeffs: [3x3 struct]
```

```
  Properties, Methods
```

`Mdl` is a `ClassificationDiscriminant` model. To access its properties, use dot notation. For example, display the group means for each predictor.

```
Mdl.Mu
```

```
ans = 3×4

    5.0060    3.4280    1.4620    0.2460
    5.9360    2.7700    4.2600    1.3260
    6.5880    2.9740    5.5520    2.0260
```

To predict labels for new observations, pass `Mdl` and predictor data to `predict`.

More About

Discriminant Classification

The model for discriminant analysis is:

- Each class (`Y`) generates data (`X`) using a multivariate normal distribution. That is, the model assumes `X` has a Gaussian mixture distribution (`gmdistribution`).
  - For linear discriminant analysis, the model has the same covariance matrix for each class, only the means vary.
  - For quadratic discriminant analysis, both means and covariances of each class vary.

`predict` classifies so as to minimize the expected classification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{k = 1}^{K} \widehat{P}(k|x) C(y|k),$$

where

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(k|x)$ is the posterior probability on page 20-6 of class k for observation x.
- $C(y|k)$ is the cost on page 20-7 of classifying an observation as y when its true class is k.

For details, see “Prediction Using Discriminant Analysis Models” on page 20-6.

Regularization

Regularization is the process of finding a small set of predictors that yield an effective predictive model. For linear discriminant analysis, there are two parameters, γ and δ, that control regularization as follows. `cvshrink` helps you select appropriate values of the parameters.

Let Σ represent the covariance matrix of the data X, and let $\widehat{X}$ be the centered data (the data X minus the mean by class). Define

$$D = \text{diag}\left(\widehat{X}^T * \widehat{X}\right).$$

The regularized covariance matrix $\tilde{\Sigma}$ is

$$\tilde{\Sigma} = (1 - \gamma)\Sigma + \gamma D.$$

Whenever $\gamma \geq$ `MinGamma`, $\tilde{\Sigma}$ is nonsingular.

Let μ_k be the mean vector for those elements of X in class k, and let μ_0 be the global mean vector (the mean of the rows of X). Let C be the correlation matrix of the data X, and let $\tilde{C}$ be the regularized correlation matrix:

$$\tilde{C} = (1 - \gamma)C + \gamma I,$$

where I is the identity matrix.

The linear term in the regularized discriminant analysis classifier for a data point x is

$$(x - \mu_0)^T \tilde{\Sigma}^{-1}(\mu_k - \mu_0) = \left[(x - \mu_0)^T D^{-1/2}\right]\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right].$$

The parameter δ enters into this equation as a threshold on the final term in square brackets. Each component of the vector $\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right]$ is set to zero if it is smaller in magnitude than the threshold δ. Therefore, for class k, if component j is thresholded to zero, component j of x does not enter into the evaluation of the posterior probability.

The `DeltaPredictor` property is a vector related to this threshold. When $\delta \geq$ `DeltaPredictor(i)`, all classes k have

$$\left|\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right| \leq \delta.$$

Therefore, when $\delta \geq$ `DeltaPredictor(i)`, the regularized classifier does not use predictor `i`.

References

[1] Guo, Y., T. Hastie, and R. Tibshirani. "Regularized linear discriminant analysis and its application in microarrays." *Biostatistics,* Vol. 8, No. 1, pp. 86–100, 2007.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- When you train a discriminant analysis model by using `fitcdiscr` or create a compact discriminant analysis model by using `makecdiscr`, the value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`CompactClassificationDiscriminant` | `compareHoldout` | `fitcdiscr`

Topics

“Discriminant Analysis Classification” on page 20-2

Introduced in R2011b

ClassificationEnsemble

Package: `classreg.learning.classif`
Superclasses: `CompactClassificationEnsemble`

Ensemble classifier

Description

`ClassificationEnsemble` combines a set of trained weak learner models and data on which these learners were trained. It can predict ensemble response for new data by aggregating predictions from its weak learners. It stores data used for training, can compute resubstitution predictions, and can resume training if desired.

Construction

Create a classification ensemble object using `fitcensemble`.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`CombineWeights`

Character vector describing how `ens` combines weak learner weights, either `'WeightedSum'` or `'WeightedAverage'`.

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response. This property is read-only.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`FitInfo`

Numeric array of fit information. The `FitInfoDescription` property describes the content of this array.

`FitInfoDescription`

Character vector describing the meaning of the `FitInfo` array.

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`

- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`LearnerNames`

Cell array of character vectors with names of weak learners in the ensemble. The name of each learner appears just once. For example, if you have an ensemble of 100 trees, `LearnerNames` is `{'Tree'}`.

`Method`

Character vector describing the method that creates `ens`.

`ModelParameters`

Parameters used in training `ens`.

`NumObservations`

Numeric scalar containing the number of observations in the training data.

`NumTrained`

Number of trained weak learners in `ens`, a scalar.

`PredictorNames`

Cell array of names for the predictor variables, in the order in which they appear in `X`.

`Prior`

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`. The number of elements of `Prior` is the number of unique classes in the response. This property is read-only.

`ReasonForTermination`

Character vector describing the reason `fitcensemble` stopped adding weak learners to the ensemble.

`ResponseName`

Character vector with the name of the response variable `Y`.

`ScoreTransform`

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitctree`.

Add or change a `ScoreTransform` function using dot notation:

```
ens.ScoreTransform = 'function'
```

or

`ens.ScoreTransform = @function`

`Trained`

A cell vector of trained classification models.

- If `Method` is `'LogitBoost'` or `'GentleBoost'`, then `ClassificationEnsemble` stores trained learner `j` in the `CompactRegressionLearner` property of the object stored in `Trained{j}`. That is, to access trained learner `j`, use `ens.Trained{j}.CompactRegressionLearner`.
- Otherwise, cells of the cell vector contain the corresponding, compact classification models.

`TrainedWeights`

Numeric vector of trained weights for the weak learners in `ens`. `TrainedWeights` has `T` elements, where `T` is the number of weak learners in `learners`.

`UsePredForLearner`

Logical matrix of size `P`-by-`NumTrained`, where `P` is the number of predictors (columns) in the training data `X`. `UsePredForLearner(i,j)` is `true` when learner `j` uses predictor `i`, and is `false` otherwise. For each learner, the predictors have the same order as the columns in the training data `X`.

If the ensemble is not of type `Subspace`, all entries in `UsePredForLearner` are `true`.

`W`

Scaled `weights`, a vector with length `n`, the number of rows in `X`. The sum of the elements of `W` is `1`.

`X`

Matrix or table of predictor values that trained the ensemble. Each column of `X` represents one variable, and each row represents one observation.

`Y`

Numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X`.

Object Functions

compact	Compact classification ensemble
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross validate ensemble
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification error
margin	Classification margins
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using ensemble of classification models
predictorImportance	Estimates of predictor importance for classification ensemble of decision trees

removeLearners	Remove members of compact classification ensemble
resubEdge	Classification edge by resubstitution
resubLoss	Classification error by resubstitution
resubMargin	Classification margins by resubstitution
resubPredict	Classify observations in ensemble of classification models
resume	Resume training ensemble
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Boosted Classification Ensemble

Load the `ionosphere` data set.

```
load ionosphere
```

Train a boosted ensemble of 100 classification trees using all measurements and the `AdaBoostM1` method.

```
Mdl = fitcensemble(X,Y,'Method','AdaBoostM1')
```

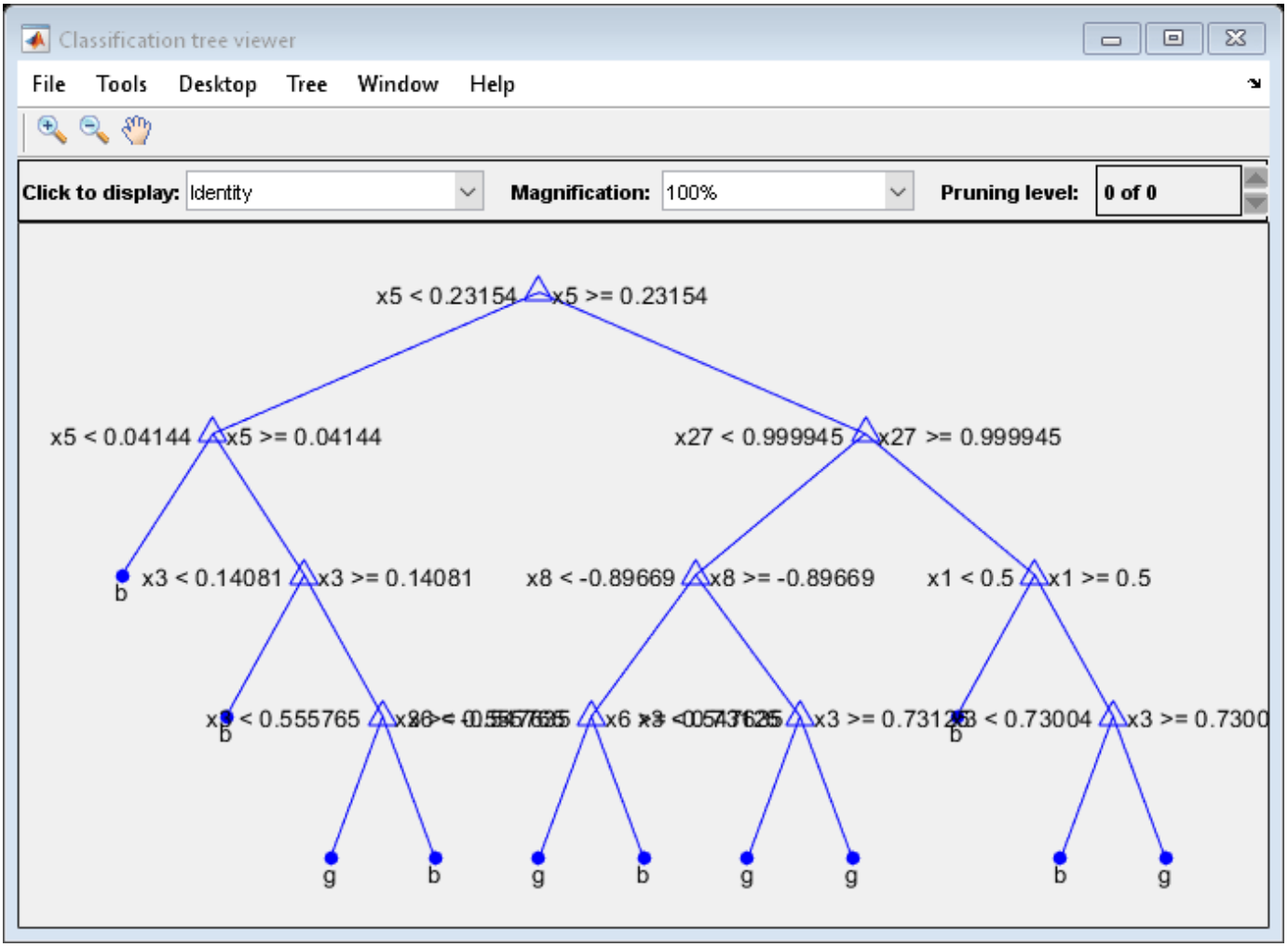
```
Mdl =
  ClassificationEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
              ClassNames: {'b'  'g'}
          ScoreTransform: 'none'
         NumObservations: 351
              NumTrained: 100
                  Method: 'AdaBoostM1'
            LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                 FitInfo: [100x1 double]
      FitInfoDescription: {2x1 cell}


  Properties, Methods
```

`Mdl` is a `ClassificationEnsemble` model object.

`Mdl.Trained` is the property that stores a 100-by-1 cell vector of the trained classification trees (`CompactClassificationTree` model objects) that compose the ensemble.

Plot a graph of the first trained classification tree.

```
view(Mdl.Trained{1},'Mode','graph')
```



By default, `fitcensemble` grows shallow trees for boosted ensembles of trees.

Predict the label of the mean of `X`.

```
predMeanX = predict(Mdl,mean(X))
```

```
predMeanX = 1x1 cell array
    {'g'}
```

Tip

For an ensemble of classification trees, the `Trained` property of `ens` stores an `ens.NumTrained`-by-1 cell vector of compact classification models. For a textual or graphical display of tree `t` in the cell vector, enter:

- `view(ens.Trained{t}.CompactRegressionLearner)` for ensembles aggregated using LogitBoost or GentleBoost.
- `view(ens.Trained{t})` for all other aggregation methods.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the ClassificationEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitcensemble`, code generation limitations for the weak learners used in the ensemble also apply to the ensemble. For more details, see the Code Generation sections of `ClassificationKNN`, `CompactClassificationDiscriminant`, and `CompactClassificationTree`.
- For fixed-point code generation, you must train an ensemble using tree learners.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationTree` | `CompactClassificationEnsemble` | `compareHoldout` | `fitcensemble` | `view`

Introduced in R2011a

ClassificationEnsemble Predict

Classify observations using ensemble of decision trees
Library: Statistics and Machine Learning Toolbox / Classification

Description

The ClassificationEnsemble Predict block classifies observations using an ensemble of decision trees (`ClassificationEnsemble`, `ClassificationBaggedEnsemble`, or `CompactClassificationEnsemble`) for multiclass classification.

Import a trained classification object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **label** returns a predicted class label for the observation. You can add an optional output port **score** that returns predicted class scores or posterior probabilities.

Ports

Input

x — Predictor data
row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the model specified by **`Select trained machine learning model`**.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

`label` — Predicted class label
scalar

Predicted class label, returned as a scalar. The predicted class is the class yielding the largest score.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point` | `enumerated`

`score` — Predicted class scores or posterior probabilities
row vector

Predicted class scores or posterior probabilities, returned as a row vector of size 1-by-*k*, where *k* is the number of classes in the tree model.

To check the order of the classes, use the `ClassNames` property of the tree model specified by **`Select trained machine learning model`**.

Dependencies

- To enable this port, select the check box for `Add output port for predicted class scores` on the **Main** tab of the Block Parameters dialog box.
- The definition and range of classification score values depend on the ensemble aggregation method. You can specify the ensemble aggregation method by using the `'Method'` name-value argument of `fitcensemble` when training the ensemble model. For details, see the “More About” on page 33-4832 section of the `predict` function reference page.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

`Select trained machine learning model` — Classification ensemble model

`ensMdl` (default) | `ClassificationEnsemble` object | `ClassificationBaggedEnsemble` object | `CompactClassificationEnsemble` object

Specify the name of a workspace variable that contains a `ClassificationEnsemble` object, `ClassificationBaggedEnsemble` object, or `CompactClassificationEnsemble` object.

When you train the model by using `fitcensemble`, the following restrictions apply:

- You must train an ensemble using tree weak learners.
- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ScoreTransform'` name-value argument cannot be `'invlogit'` or an anonymous function.
- You cannot use surrogate splits for tree weak learners, that is, the value of the `'Surrogate'` name-value argument must be `'off'` (default) when you define tree weak learners by using the `templateTree` function.

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `ClassificationEnsemble` object | `ClassificationBaggedEnsemble` object | `CompactClassificationEnsemble` object
Default: `'ensMdl'`

`Add output port for predicted class scores` — Add second output port for predicted class scores

`off` (default) | `on`

Select the check box to include the second output port **score** in the ClassificationEnsemble Predict block.

Programmatic Use

Block Parameter: `ShowOutputScore`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations
`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use

Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action
`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.

Action	Rationale	Impact on Overflows	Example
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use

Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type

`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use

Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Label data type` — Data type of label output

`Inherit: Inherit via back propagation` | `Inherit: auto` | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `Enum: <class name>` | `<data type expression>`

Specify the data type for the **label** output. The type can be inherited, specified as an enumerated data type, or expressed as a data type object such as `Simulink.NumericType`.

When you select an inherited option, the software behaves as follows:

- `Inherit: Inherit via back propagation` (default for numeric and logical labels) — Simulink automatically determines the **Label data type** of the block during data type propagation (see "Data Type Propagation" (Simulink)). In this case, the block uses the data type of a downstream block or signal object.
- `Inherit: auto` (default for nonnumeric labels) — The block uses an autodefined enumerated data type variable. For example, suppose the workspace variable name specified by `Select`

`trained machine learning model` is `myMdl`, and the class labels are `class 1` and `class 2`. Then, the corresponding **label** values are `myMdl_enumLabels.class_1` and `myMdl_enumLabels.class_2`. The block converts the class labels to valid MATLAB identifiers by using the `matlab.lang.makeValidName` function.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Dependencies

The supported data types depend on the labels used in the model specified by `Select trained machine learning model`.

- If the model uses numeric or logical labels, the supported data types are `Inherit: Inherit via back propagation` (default), `double`, `single`, `half`, `int8`, `uint8`, `int16`, `uint16`, `int32`, `uint32`, `int64`, `uint64`, `boolean`, fixed point, and a data type object.
- If the model uses nonnumeric labels, the supported data types are `Inherit: auto` (default), `Enum: <class name>`, and a data type object.

Programmatic Use

Block Parameter: `LabelDataTypeStr`
Type: character vector
Values: `'Inherit: Inherit via back propagation'` | `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'Enum: <class name>'` | `'<data type expression>'`
Default: `'Inherit: Inherit via back propagation'` (for numeric and logical labels) | `'Inherit: auto'` (for nonnumeric labels)

`Label minimum` — Minimum value of label output for range checking

[ ] (default) | scalar

Lower value of the **label** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label minimum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use

Block Parameter: `LabelOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Label maximum` — Maximum value of label output for range checking

`[]` (default) | scalar

Upper value of the **label** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label maximum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use

Block Parameter: `LabelOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score data type` — Data type of score output

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **score** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use
Block Parameter: `ScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Score minimum` — Minimum value of score output for range checking

`[ ]` (default) | scalar

Lower value of the **score** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score minimum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use
Block Parameter: `ScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score maximum` — Maximum value of score output for range checking

`[ ]` (default) | scalar

Upper value of the **score** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.

- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score maximum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `ScoreOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score data type` — Untransformed score data type

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `boolean` | `<data type expression>`

Specify the data type for the internal untransformed scores. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses a score transformation other than `'none'` (default, same as `'identity'`).

- If the model uses no score transformations (`'none'` or `'identity'`), then you can specify the score data type by using `Score data type`.
- If the model uses a score transformation other than `'none'` or `'identity'`, then you can specify the data type of untransformed raw scores by using this parameter and specify the data type of transformed scores by using `Score data type`.

You can change the score transformation option by specifying the `'ScoreTransform'` name-value argument during training, or by changing the `ScoreTransform` property after training.

Programmatic Use

Block Parameter: `RawScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Raw score minimum` — Minimum untransformed score for range checking

`[ ]` (default) | scalar

Lower value of the untransformed score range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score minimum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use

Block Parameter: `RawScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score maximum` — Maximum untransformed score for range checking

`[ ]` (default) | scalar

Upper value of the untransformed score range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score maximum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use

Block Parameter: `RawScoreOutMax`

Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Weak learner data type` — Data type of weak learner outputs
`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the outputs from weak learners. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use
Block Parameter: `WeakLearnerDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Weak learner minimum` — Minimum value of weak learner outputs for range checking
`[]` (default) | scalar

Lower value of the weak learner output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Weak learner minimum** parameter does not saturate or clip the actual weak learner output signals.

Programmatic Use
Block Parameter: `WeakLearnerOutMin`
Type: character vector

Values: `'[]'` | scalar
Default: `'[]'`

`Weak learner maximum` — Maximum value of weak learner outputs for range checking
`[]` (default) | scalar

Upper value of the weak learner output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Weak learner maximum** parameter does not saturate or clip the actual weak learner output signals.

Programmatic Use
Block Parameter: `WeakLearnerOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `enumerated` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of an ensemble of decision trees (`ClassificationEnsemble`, `ClassificationBaggedEnsemble`, or `CompactClassificationEnsemble`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the ClassificationEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion

Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks

ClassificationSVM Predict | ClassificationTree Predict | RegressionEnsemble Predict

Objects

`ClassificationBaggedEnsemble` | `ClassificationEnsemble` | `CompactClassificationEnsemble`

Functions

`fitcensemble` | `predict`

Topics

“Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
“Predict Class Labels Using ClassificationTree Predict Block” on page 32-121
“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2021a

ClassificationKNN

k-nearest neighbor classification

Description

`ClassificationKNN` is a nearest-neighbor classification model in which you can alter both the distance metric and the number of nearest neighbors. Because a `ClassificationKNN` classifier stores training data, you can use the model to compute resubstitution predictions. Alternatively, use the model to classify new observations using the `predict` method.

Creation

Create a `ClassificationKNN` model using `fitcknn`.

Properties

KNN Properties

`BreakTies` — Tie-breaking algorithm
`'smallest'` (default) | `'nearest'` | `'random'`

Tie-breaking algorithm used by `predict` when multiple classes have the same smallest cost, specified as one of the following:

- `'smallest'` — Use the smallest index among tied groups.
- `'nearest'` — Use the class with the nearest neighbor among tied groups.
- `'random'` — Use a random tiebreaker among tied groups.

By default, ties occur when multiple classes have the same number of nearest points among the *k* nearest neighbors. `BreakTies` applies when `IncludeTies` is `false`.

Change `BreakTies` using dot notation: `mdl.BreakTies = newBreakTies`.

`Distance` — Distance metric
`'cityblock'` | `'chebychev'` | `'correlation'` | `'cosine'` | `'euclidean'` | function handle | ...

Distance metric, specified as a character vector or a function handle. The values allowed depend on the `NSMethod` property.

NSMethod	Distance Metric Allowed
`'exhaustive'`	Any distance metric of `ExhaustiveSearcher`
`'kdtree'`	`'cityblock'`, `'chebychev'`, `'euclidean'`, or `'minkowski'`

The following table lists the `ExhaustiveSearcher` distance metrics.

Value	Description
`'cityblock'`	City block distance.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors).
`'euclidean'`	Euclidean distance.
`'hamming'`	Hamming distance, the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix `C`. The default value of `C` is the sample covariance matrix of `X`, as computed by `cov(X,'omitrows')`. To specify a different value for `C`, set the `DistParameter` property of `mdl` using dot notation.
`'minkowski'`	Minkowski distance. The default exponent is `2`. To specify a different exponent, set the `DistParameter` property of `mdl` using dot notation.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between `X` and a query point is scaled, meaning divided by a scale value `S`. The default value of `S` is the standard deviation computed from `X`, `S = std(X,'omitnan')`. To specify another value for `S`, set the `DistParameter` property of `mdl` using dot notation.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Distance function handle. *`distfun`* has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`N` vector containing one row of `X` or `Y`.
	• `ZJ` is an `M2`-by-`N` matrix containing multiple rows of `X` or `Y`.
	• `D2` is an `M2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.

For more information, see “Distance Metrics” on page 18-12.

Change `Distance` using dot notation: `mdl.Distance = newDistance`.

If `NSMethod` is `'kdtree'`, you can use dot notation to change `Distance` only for the metrics `'cityblock'`, `'chebychev'`, `'euclidean'`, and `'minkowski'`.

Data Types: `char` | `function_handle`

`DistanceWeight` — Distance weighting function

`'equal'` | `'inverse'` | `'squaredinverse'` | function handle

Distance weighting function, specified as one of the values in this table.

Value	Description
`'equal'`	No weighting
`'inverse'`	Weight is 1/distance
`'squaredinverse'`	Weight is 1/distance^2
`@`*`fcn`*	*`fcn`* is a function that accepts a matrix of nonnegative distances and returns a matrix of the same size containing nonnegative distance weights. For example, `'squaredinverse'` is equivalent to `@(d)d.^(-2)`.

Change `DistanceWeight` using dot notation: `mdl.DistanceWeight = newDistanceWeight`.

Data Types: `char` | `function_handle`

`DistParameter` — Parameter for distance metric

positive definite covariance matrix | positive scalar | vector of positive scale values

Parameter for the distance metric, specified as one of the values described in this table.

Distance Metric	Parameter
`'mahalanobis'`	Positive definite covariance matrix `C`
`'minkowski'`	Minkowski distance exponent, a positive scalar
`'seuclidean'`	Vector of positive scale values with length equal to the number of columns of `X`

For any other distance metric, the value of `DistParameter` must be `[]`.

You can alter `DistParameter` using dot notation: `mdl.DistParameter = newDistParameter`. However, if `Distance` is `'mahalanobis'` or `'seuclidean'`, then you cannot alter `DistParameter`.

Data Types: `single` | `double`

`IncludeTies` — Tie inclusion flag

`false` (default) | `true`

Tie inclusion flag indicating whether `predict` includes all the neighbors whose distance values are equal to the *k*th smallest distance, specified as `false` or `true`. If `IncludeTies` is `true`, `predict` includes all of these neighbors. Otherwise, `predict` uses exactly *k* neighbors (see the `BreakTies` property).

Change `IncludeTies` using dot notation: `mdl.IncludeTies = newIncludeTies`.

Data Types: `logical`

`NSMethod` — Nearest neighbor search method

`'kdtree'` | `'exhaustive'`

This property is read-only.

Nearest neighbor search method, specified as either `'kdtree'` or `'exhaustive'`.

- `'kdtree'` — Creates and uses a *K*d-tree to find nearest neighbors.
- `'exhaustive'` — Uses the exhaustive search algorithm. When predicting the class of a new point `xnew`, the software computes the distance values from all points in `X` to `xnew` to find nearest neighbors.

The default value is `'kdtree'` when `X` has `10` or fewer columns, `X` is not sparse, and the distance metric is a `'kdtree'` type. Otherwise, the default value is `'exhaustive'`.

`NumNeighbors` — Number of nearest neighbors
positive integer value

Number of nearest neighbors in `X` used to classify each point during prediction, specified as a positive integer value.

Change `NumNeighbors` using dot notation: `mdl.NumNeighbors = newNumNeighbors`.

Data Types: `single` | `double`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
`[]` | vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Names of classes in training data `Y`
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Names of the classes in the training data `Y` with duplicates removed, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as `Y`. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Cost of misclassification
square matrix

Cost of the misclassification of a point, specified as a square matrix. `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (that is, the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns in `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response.

By default, `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Change a `Cost` matrix using dot notation: `mdl.Cost = costMatrix`.

Data Types: `single` | `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ModelParameters` — Parameters used in training `ClassificationKNN`
object

This property is read-only.

Parameters used in training the `ClassificationKNN` model, specified as an object.

`Mu` — Predictor means
numeric vector

This property is read-only.

Predictor means, specified as a numeric vector of length `numel(PredictorNames)`.

If you do not standardize `mdl` when training the model using `fitcknn`, then `Mu` is empty (`[]`).

Data Types: `single` | `double`

`NumObservations` — Number of observations
positive integer scalar

This property is read-only.

Number of observations used in training the `ClassificationKNN` model, specified as a positive integer scalar. This number can be less than the number of rows in the training data because rows containing `NaN` values are not part of the fit.

Data Types: `double`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The variable names are in the same order in which they appear in the training data `X`.

Data Types: `cell`

`Prior` — Prior probabilities for each class
numeric vector

Prior probabilities for each class, specified as a numeric vector. The order of the elements in `Prior` corresponds to the order of the classes in `ClassNames`.

Add or change a `Prior` vector using dot notation: `mdl.Prior = priorVector`.

Data Types: `single` | `double`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`RowsUsed` — Rows used in fitting
`[]` | logical vector

This property is read-only.

Rows of the original training data used in fitting the `ClassificationKNN` model, specified as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as either a character vector or a function handle.

This table summarizes the available character vectors.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Change `ScoreTransform` using dot notation: `mdl.ScoreTransform = newScoreTransform`.

Data Types: `char` | `function_handle`

`Sigma` — Predictor standard deviations
numeric vector

This property is read-only.

Predictor standard deviations, specified as a numeric vector of length `numel(PredictorNames)`.

If you do not standardize the predictor variables during training, then `Sigma` is empty (`[]`).

Data Types: `single` | `double`

`W` — Observation weights
vector of nonnegative values

This property is read-only.

Observation weights, specified as a vector of nonnegative values with the same number of rows as `Y`. Each entry in `W` specifies the relative importance of the corresponding observation in `Y`.

Data Types: `single` | `double`

`X` — Unstandardized predictor data
numeric matrix

This property is read-only.

Unstandardized predictor data, specified as a numeric matrix. Each column of `X` represents one predictor (variable), and each row represents one observation.

Data Types: `single` | `double`

`Y` — Class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Class labels, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. Each value in `Y` is the observed class label for the corresponding row in `X`.

`Y` has the same data type as the data in `Y` used for training the model. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

Hyperparameter Optimization Properties

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table

This property is read-only.

Cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty when the `'OptimizeHyperparameters'` name-value pair argument is nonempty when you create the model using `fitcknn`. The value depends on the setting of the `'HyperparameterOptimizationOptions'` name-value pair argument when you create the model:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validate machine learning model
edge	Edge of k-nearest neighbor classifier
gather	Gather properties of machine learning model from GPU
lime	Local interpretable model-agnostic explanations (LIME)
loss	Loss of k-nearest neighbor classifier
margin	Margin of k-nearest neighbor classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels using k-nearest neighbor classification model
resubEdge	Resubstitution classification edge
resubLoss	Resubstitution classification loss
resubMargin	Resubstitution classification margin
resubPredict	Classify training data using trained classifier
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Examples

Train *k*-Nearest Neighbor Classifier

Train a *k*-nearest neighbor classifier for Fisher's iris data, where *k*, the number of nearest neighbors in the predictors, is 5.

Load Fisher's iris data.

```
load fisheriris
X = meas;
Y = species;
```

`X` is a numeric matrix that contains four petal measurements for 150 irises. `Y` is a cell array of character vectors that contains the corresponding iris species.

Train a 5-nearest neighbor classifier. Standardize the noncategorical predictor data.

```
Mdl = fitcknn(X,Y,'NumNeighbors',5,'Standardize',1)
```

```
Mdl =
  ClassificationKNN
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
          ScoreTransform: 'none'
         NumObservations: 150
                Distance: 'euclidean'
            NumNeighbors: 5
```

```
  Properties, Methods
```

`Mdl` is a trained `ClassificationKNN` classifier, and some of its properties appear in the Command Window.

To access the properties of `Mdl`, use dot notation.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl.Prior
```

```
ans = 1×3

    0.3333    0.3333    0.3333
```

`Mdl.Prior` contains the class prior probabilities, which you can specify using the `'Prior'` name-value pair argument in `fitcknn`. The order of the class prior probabilities corresponds to the order of the classes in `Mdl.ClassNames`. By default, the prior probabilities are the respective relative frequencies of the classes in the data.

You can also reset the prior probabilities after training. For example, set the prior probabilities to 0.5, 0.2, and 0.3, respectively.

```
Mdl.Prior = [0.5 0.2 0.3];
```

You can pass `Mdl` to `predict` to label new measurements or `crossval` to cross-validate the classifier.

Tips

- The `compact` function reduces the size of most classification models by removing the training data properties and any other properties that are not required to predict the labels of new observations. Because *k*-nearest neighbor classification models require all of the training data to predict labels, you cannot reduce the size of a `ClassificationKNN` model.

Alternative Functionality

`knnsearch` finds the *k*-nearest neighbors of points. `rangesearch` finds all the points within a fixed distance. You can use these functions for classification, as shown in “Classify Query Data” on page 18-18. If you want to perform classification, then using `ClassificationKNN` models can be more convenient because you can train a classifier in one step (using `fitcknn`) and classify in other steps (using `predict`). Alternatively, you can train a *k*-nearest neighbor classification model using one of the cross-validation options in the call to `fitcknn`. In this case, `fitcknn` returns a `ClassificationPartitionedModel` cross-validated model object.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- When you train a *k*-nearest neighbor classification model by using `fitcknn`, the following restrictions apply.
  - The value of the `'Distance'` name-value pair argument cannot be a custom distance function.
  - The value of the `'DistanceWeight'` name-value pair argument can be a custom distance weight function, but it cannot be an anonymous function.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions fully support GPU arrays:
  - `compareHoldout`
  - `edge`
  - `loss`
  - `margin`
  - `partialDependence`
  - `plotPartialDependence`
  - `predict`
- The following object functions support model objects fitted with GPU array input arguments:
  - `crossval`
  - `gather`
  - `resubEdge`
  - `resubLoss`
  - `resubMargin`
  - `resubPredict`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`fitcknn` | `predict`

Topics

“Construct KNN Classifier” on page 18-28

“Examine Quality of KNN Classifier” on page 18-28
“Predict Classification Using KNN Classifier” on page 18-29
“Modify KNN Classifier” on page 18-29
“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2012a

ClassificationLinear class

Linear model for binary classification of high-dimensional data

Description

`ClassificationLinear` is a trained linear model object for binary classification; the linear model is a support vector machine (SVM) or logistic regression model. `fitclinear` fits a `ClassificationLinear` model by minimizing the objective function using techniques that reduce computation time for high-dimensional data sets (e.g., stochastic gradient descent). The classification loss plus the regularization term compose the objective function.

Unlike other classification models, and for economical memory usage, `ClassificationLinear` model objects do not store the training data. However, they do store, for example, the estimated linear model coefficients, prior-class probabilities, and the regularization strength.

You can use trained `ClassificationLinear` models to predict labels or classification scores for new data. For details, see `predict`.

Construction

Create a `ClassificationLinear` object by using `fitclinear`.

Properties

Linear Classification Properties

Lambda — Regularization term strength

nonnegative scalar | vector of nonnegative values

Regularization term strength, specified as a nonnegative scalar or vector of nonnegative values.

Data Types: `double` | `single`

Learner — Linear classification model type

`'logistic'` | `'svm'`

Linear classification model type, specified as `'logistic'` or `'svm'`.

In this table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.
- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Loss Function	FittedLoss Value
`'logistic'`	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$	`'logit'`

Value	Algorithm	Loss Function	FittedLoss Value
'svm'	Support vector machine	Hinge: $\ell[y, f(x)] = \max [0, 1 - yf(x)]$	'hinge'

Beta — Linear coefficient estimates
numeric vector

Linear coefficient estimates, specified as a numeric vector with length equal to the number of predictors.

Data Types: double

Bias — Estimated bias term
numeric scalar

Estimated bias term or model intercept, specified as a numeric scalar.

Data Types: double

FittedLoss — Loss function used to fit linear model
'hinge' | 'logit'

This property is read-only.

Loss function used to fit the linear model, specified as 'hinge' or 'logit'.

Value	Algorithm	Loss Function	Learner Value
'hinge'	Support vector machine	Hinge: $\ell[y, f(x)] = \max [0, 1 - yf(x)]$	'svm'
'logit'	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log \{1 + \exp[-yf(x)]\}$	'logistic'

Regularization — Complexity penalty type
'lasso (L1)' | 'ridge (L2)'

Complexity penalty type, specified as 'lasso (L1)' or 'ridge (L2)'.

The software composes the objective function for minimization from the sum of the average loss function (see FittedLoss) and a regularization value from this table.

Value	Description
'lasso (L1)'	Lasso (L_1) penalty: $\lambda \sum_{j=1}^{p} \left\|\beta_j\right\|$
'ridge (L2)'	Ridge (L_2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

λ specifies the regularization term strength (see Lambda).

The software excludes the bias term (β_0) from the regularization penalty.

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`ModelParameters` — Parameters used for training model
structure

Parameters used for training the `ClassificationLinear` model, specified as a structure.

Access fields of `ModelParameters` using dot notation. For example, access the relative tolerance on the linear coefficients and the bias term by using `Mdl.ModelParameters.BetaTolerance`.

Data Types: `struct`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as classes in `ClassNames`, and the order of the elements corresponds to the elements of `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For linear classification models and before transformation, the predicted classification score for the observation x (row vector) is $f(x) = x\beta + b$, where β and b correspond to `Mdl.Beta` and `Mdl.Bias`, respectively.

To change the score transformation function to, for example, *`function`*, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1

Value	Description
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function, or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Object Functions

edge	Classification edge for linear classification models
incrementalLearner	Convert linear model for binary classification to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification loss for linear classification models
margin	Classification margins for linear classification models
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels for linear classification models
shapley	Shapley values
selectModels	Choose subset of regularized, binary linear classification models
update	Update model parameters for code generation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Linear Classification Model

Train a binary, linear classification model using support vector machines, dual SGD, and ridge regularization.

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels. There are more than two classes in the data.

Identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

Train the model using the entire data set. Determine how well the optimization algorithm fit the model to the data by extracting a fit summary.

```
rng(1); % For reproducibility
[Mdl,FitInfo] = fitclinear(X,Ystats)
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [34023x1 double]
              Bias: -1.0059
            Lambda: 3.1674e-05
           Learner: 'svm'


  Properties, Methods


FitInfo = struct with fields:
                    Lambda: 3.1674e-05
                 Objective: 5.3783e-04
                 PassLimit: 10
                 NumPasses: 10
                BatchLimit: []
             NumIterations: 238561
              GradientNorm: NaN
         GradientTolerance: 0
      RelativeChangeInBeta: 0.0562
             BetaTolerance: 1.0000e-04
             DeltaGradient: 1.4582
    DeltaGradientTolerance: 1
           TerminationCode: 0
         TerminationStatus: {'Iteration limit exceeded.'}
                     Alpha: [31572x1 double]
                   History: []
                   FitTime: 0.2014
                    Solver: {'dual'}
```

`Mdl` is a `ClassificationLinear` model. You can pass `Mdl` and the training or new data to `loss` to inspect the in-sample classification error. Or, you can pass `Mdl` and new predictor data to `predict` to predict class labels for new observations.

`FitInfo` is a structure array containing, among other things, the termination status (`TerminationStatus`) and how long the solver took to fit the model to the data (`FitTime`). It is good practice to use `FitInfo` to determine whether optimization-termination measurements are satisfactory. Because training time is small, you can try to retrain the model, but increase the number of passes through the data. This can improve measures like `DeltaGradient`.

Predict Class Labels Using Linear Classification Model

Load the NLP data set.

```
load nlpdata
n = size(X,1); % Number of observations
```

Identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Hold out 5% of the data.

```
rng(1); % For reproducibility
cvp = cvpartition(n,'Holdout',0.05)
```

```
cvp =
Hold-out cross validation partition
   NumObservations: 31572
       NumTestSets: 1
         TrainSize: 29994
          TestSize: 1578
```

`cvp` is a `CVPartition` object that defines the random partition of *n* data into training and test sets.

Train a binary, linear classification model using the training set that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. For faster training time, orient the predictor data matrix so that the observations are in columns.

```
idxTrain = training(cvp); % Extract training set indices
X = X';
Mdl = fitclinear(X(:,idxTrain),Ystats(idxTrain),'ObservationsIn','columns');
```

Predict observations and classification error for the hold out sample.

```
idxTest = test(cvp); % Extract test set indices
labels = predict(Mdl,X(:,idxTest),'ObservationsIn','columns');
L = loss(Mdl,X(:,idxTest),Ystats(idxTest),'ObservationsIn','columns')
```

```
L = 7.1753e-04
```

`Mdl` misclassifies fewer than 1% of the out-of-sample observations.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- When you train a linear classification model by using `fitclinear`, the following restrictions apply.
  - If the predictor data input argument value is a matrix, it must be a full, numeric matrix. Code generation does not support sparse data.
  - You can specify only one regularization strength, either `'auto'` or a nonnegative scalar for the `'Lambda'` name-value pair argument.

- The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.
- For code generation with a coder configurer, the following additional restrictions apply.
  - Categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`) are not supported. You cannot use the `'CategoricalPredictors'` name-value argument.To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.
  - Class labels with the `categorical` data type are not supported. Both the class label value in training data (`Tbl` or `Y`) and the value of the `'ClassNames'` name-value argument cannot be an array with the `categorical` data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationECOC` | `ClassificationKernel` | `ClassificationPartitionedLinear` | `ClassificationPartitionedLinearECOC` | `fitclinear` | `predict`

Introduced in R2016a

ClassificationLinearCoderConfigurer

Coder configurer for linear binary classification of high-dimensional data

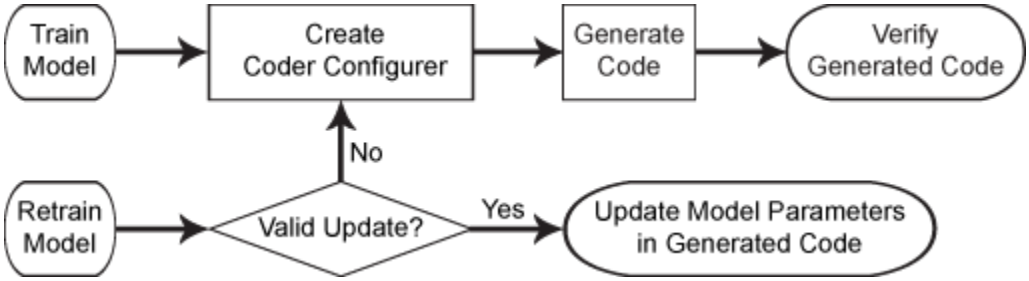
Description

A `ClassificationLinearCoderConfigurer` object is a coder configurer of a linear classification model (`ClassificationLinear`) used for binary classification of high-dimensional data.

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes of linear model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the linear classification model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the linear model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of a linear classification model, see the Code Generation sections of `ClassificationLinear`, `predict`, and `update`.

Creation

After training a linear classification model by using `fitclinear`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of the `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data
`LearnerCoderInput` object

Coder attributes of the predictor data to pass to the generated C/C++ code for the `predict` function of the linear classification model, specified as a `LearnerCoderInput` on page 33-433 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
  - If the `Value` attribute of the `ObservationsIn` property for the `ClassificationLinearCoderConfigurer` is `'rows'`, then this `SizeVector` value is `[n p]`, where `n` corresponds to the number of observations and `p` corresponds to the number of predictors.
  - If the `Value` attribute of the `ObservationsIn` property for the `ClassificationLinearCoderConfigurer` is `'columns'`, then this `SizeVector` value is `[p n]`.

  To switch the elements of `SizeVector` (for example, to change `[n p]` to `[p n]`), modify the `Value` attribute of the `ObservationsIn` property for the `ClassificationLinearCoderConfigurer` accordingly. You cannot modify the `SizeVector` value directly.
- `VariableDimensions` — The default value is `[0 0]`, which indicates that the array size is fixed as specified in `SizeVector`.

  You can set this value to `[1 0]` if the `SizeVector` value is `[n p]` or to `[0 1]` if it is `[p n]`, which indicates that the array has variable-size rows and fixed-size columns. For example, `[1 0]` specifies that the first value of `SizeVector` (`n`) is the upper bound for the number of rows, and the second value of `SizeVector` (`p`) is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations (in rows) of three predictor variables (in columns), specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of

observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

`ObservationsIn` — Coder attributes of predictor data observation dimension
`EnumeratedInput` object

Coder attributes of the predictor data observation dimension (`'ObservationsIn'` name-value pair argument of `predict`), specified as an `EnumeratedInput` on page 33-433 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the `'ObservationsIn'` name-value pair argument determines the default values of the `EnumeratedInput` coder attributes:

- `Value` — The default value is the predictor data observation dimension you use when creating the coder configurer, specified as `'rows'` or `'columns'`. If you do not specify `'ObservationsIn'` when creating the coder configurer, the default value is `'rows'`.
- `SelectedOption` — This value is always `'Built-in'`. This attribute is read-only.
- `BuiltInOptions` — Cell array of `'rows'` and `'columns'`. This attribute is read-only.
- `IsConstant` — This value must be `true`.
- `Tunability` — The default value is `false` if you specify `'ObservationsIn','rows'` when creating the coder configurer, and `true` if you specify `'ObservationsIn','columns'`. If you set `Tunability` to `false`, the software sets `Value` to `'rows'`. If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

`NumOutputs` — Number of outputs in `predict`
1 (default) | 2

Number of output arguments to return from the generated C/C++ code for the `predict` function of the linear classification model, specified as 1 or 2.

The output arguments of `predict` are `Label` (predicted class labels) and `Score` (classification scores), in that order. `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function, where `n` is the `NumOutputs` value.

After creating the coder configurer `configurer`, you can specify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of a linear classification model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the

parameters before generating code. Use a `LearnerCoderInput` on page 33-433 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Beta` — Coder attributes of linear predictor coefficients
`LearnerCoderInput` object

Coder attributes of the linear predictor coefficients (`Beta` of a linear classification model), specified as a `LearnerCoderInput` on page 33-433 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[p 1]`, where `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

`Bias` — Coder attributes of bias term
`LearnerCoderInput` object

Coder attributes of the bias term (`Bias` of a linear classification model), specified as a `LearnerCoderInput` on page 33-433 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

`Cost` — Coder attributes of misclassification cost
`LearnerCoderInput` object

Coder attributes of the misclassification cost (`Cost` of a linear classification model), specified as a `LearnerCoderInput` on page 33-433 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[2 2]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

`Prior` — Coder attributes of prior probabilities

`LearnerCoderInput` object

Coder attributes of the prior probabilities (`Prior` of a linear classification model), specified as a `LearnerCoderInput` on page 33-433 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 2]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code

`'ClassificationLinearModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `ClassificationLinearCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level

`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints,

then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — List of tunable input arguments of `predict`
cell array

This property is read-only.

List of tunable input arguments of the entry-point function `predict.m` for code generation, specified as a cell array. The cell array contains another cell array that includes `coder.PrimitiveType` objects and `coder.Constant` objects.

If you modify the coder attributes of `predict` arguments on page 33-421, then the software updates the corresponding objects accordingly. If you specify the `Tunability` attribute as `false`, then the software removes the corresponding objects from the `PredictInputs` list.

The cell array in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each `coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-423 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `ionosphere` data set, and train a binary linear classification model. Pass the transposed predictor matrix `Xnew` to `fitclinear`, and use the `'ObservationsIn'` name-value pair argument to specify that the columns of `Xnew` correspond to observations.

```
load ionosphere
Xnew = X';
Mdl = fitclinear(Xnew,Y,'ObservationsIn','columns');
```

`Mdl` is a `ClassificationLinear` object.

Create a coder configurer for the `ClassificationLinear` model by using `learnerCoderConfigurer`. Specify the predictor data `Xnew`, and use the `'ObservationsIn'`

name-value pair argument to specify the observation dimension of `Xnew`. The `learnerCoderConfigurer` function uses these input arguments to configure the coder attributes of the corresponding input arguments of `predict`.

```
configurer = learnerCoderConfigurer(Mdl,Xnew,'ObservationsIn','columns')
```

```
configurer =
  ClassificationLinearCoderConfigurer with properties:

   Update Inputs:
              Beta: [1x1 LearnerCoderInput]
              Bias: [1x1 LearnerCoderInput]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
    ObservationsIn: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'ClassificationLinearModel'


  Properties, Methods
```

`configurer` is a `ClassificationLinearCoderConfigurer` object, which is a coder configurer of a `ClassificationLinear` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the linear classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationLinearModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationLinearModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationLinearModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:29
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:29
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:29
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('ClassificationLinearModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Beta
        %                       Bias
        %                       Prior
        %                       Cost
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X, ObservationsIn
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of Linear Classification Model in Generated Code

Train a linear classification model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the linear model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`). Train a binary linear classification model using half of the

observations. Transpose the predictor data, and use the `'ObservationsIn'` name-value pair argument to specify that the columns of `XTrain` correspond to observations.

```
load ionosphere

rng('default') % For reproducibility
n = length(Y);
c = cvpartition(Y,'HoldOut',0.5);
idxTrain = training(c,1);
XTrain = X(idxTrain,:)';
YTrain = Y(idxTrain);

Mdl = fitclinear(XTrain,YTrain,'ObservationsIn','columns');
```

`Mdl` is a `ClassificationLinear` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationLinear` model by using `learnerCoderConfigurer`. Specify the predictor data `XTrain`, and use the `'ObservationsIn'` name-value pair argument to specify the observation dimension of `XTrain`. The `learnerCoderConfigurer` function uses these input arguments to configure the coder attributes of the corresponding input arguments of `predict`. Also, set the number of outputs to 2 so that the generated code returns predicted labels and scores.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'ObservationsIn','columns','NumOutputs',2);
```

`configurer` is a `ClassificationLinearCoderConfigurer` object, which is a coder configurer of a `ClassificationLinear` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the linear classification model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of the predictor data that you want to pass to the generated code.

Specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [34 Inf];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   0   1
```

The size of the first dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 34 predictors, the value of the `SizeVector` attribute must be 34 and the value of the `VariableDimensions` attribute must be `0`.

The size of the second dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor

data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The order of the dimensions in `SizeVector` and `VariableDimensions` depends on the coder attributes of `ObservationsIn`.

```
configurer.ObservationsIn
```

```
ans =
  EnumeratedInput with properties:

             Value: 'columns'
    SelectedOption: 'Built-in'
    BuiltInOptions: {'rows'  'columns'}
        IsConstant: 1
        Tunability: 1
```

When the `Value` attribute of the `ObservationsIn` property is `'columns'`, the first dimension of the `SizeVector` and `VariableDimensions` attributes of `X` corresponds to the number of predictors, and the second dimension corresponds to the number of observations. When the `Value` attribute of `ObservationsIn` is `'rows'`, the order of the dimensions is switched.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the linear classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationLinearModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationLinearModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationLinearModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score] = predict(Mdl,XTrain,'ObservationsIn','columns');
[label_mex,score_mex] = ClassificationLinearModel('predict',XTrain,'ObservationsIn','columns');
```

Compare `label` and `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

Compare `score` and `score_mex`.

```
max(abs(score-score_mex),[],'all')
```

```
ans = 0
```

In general, `score_mex` might include round-off differences compared to `score`. In this case, the comparison confirms that `score` and `score_mex` are equal.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitclinear(X',Y,'ObservationsIn','columns');
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationLinearModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score] = predict(retrainedMdl,X','ObservationsIn','columns');
[label_mex,score_mex] = ClassificationLinearModel('predict',X','ObservationsIn','columns');
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
max(abs(score-score_mex),[],'all')
```

```
ans = 0
```

The comparison confirms that `label` and `label_mex` are equal, and that the score values are equal.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the data type of the bias term `Bias` of the coder configurer `configurer`:

```
configurer.Bias.DataType = 'single';
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

`EnumeratedInput` Object

A coder configurer uses an `EnumeratedInput` object to specify the coder attributes of `predict` input arguments that have a finite set of available values.

An `EnumeratedInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description

`Value`	Value of the `predict` argument in the generated code, specified as a character vector or a `LearnerCoderInput` on page 33-433 object.
	• Character vector in `BuiltInOptions` — You can specify one of the `BuiltInOptions` using either the option name or its index value. For example, to choose the first option, specify `Value` as either the first character vector in `BuiltInOptions` or `1`.
	• Character vector designating a custom function name — To use a custom option, define a custom function on the MATLAB search path, and specify `Value` as the name of the custom function.
	• `LearnerCoderInput` on page 33-433 object — If you set `IsConstant` to `false` (logical 0), then the software changes `Value` to a `LearnerCoderInput` on page 33-433 object with the following read-only coder attribute values. These values indicate that the input in the generated code is a variable-size, tunable character vector that is one of the available values in `BuiltInOptions`.
	  • `SizeVector` — `[1 c]`, indicating the upper bound of the array size, where `c` is the length of the longest available character vector in `Option`
	  • `VariableDimensions` — `[0 1]`, indicating that the array is a variable-size vector
	  • `DataType` — `'char'`
	  • `Tunability` — 1
	The default value of `Value` is consistent with the default value of the corresponding `predict` argument, which is one of the character vectors in `BuiltInOptions`.

`SelectedOption`	Status of the selected option, specified as `'Built-in'`, `'Custom'`, or `'NonConstant'`. The software sets `SelectedOption` according to `Value`: • `'Built-in'`(default) — When `Value` is one of the character vectors in `BuiltInOptions` • `'Custom'` — When `Value` is a character vector that is not in `BuiltInOptions` • `'NonConstant'` — When `Value` is a `LearnerCoderInput` on page 33-433 object This attribute is read-only.
`BuiltInOptions`	List of available character vectors for the corresponding `predict` argument, specified as a cell array. This attribute is read-only.

Attribute Name	Description
`IsConstant`	Indicator specifying whether or not the array value is a compile-time constant (`coder.Constant`) in the generated code, specified as `true` (logical 1, default) or `false` (logical 0). If you set this value to `false`, then the software changes `Value` to a `LearnerCoderInput` on page 33-433 object.
`Tunability`	Indicator specifying whether or not `predict` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0, default). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of `ObservationsIn` of the coder configurer `configurer`:

```
configurer.ObservationsIn.Value = 'columns';
```

See Also

`ClassificationECOCCoderConfigurer` | `ClassificationLinear` | `learnerCoderConfigurer` | `predict` | `update`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2019b

ClassificationNaiveBayes

Naive Bayes classification for multiclass classification

Description

`ClassificationNaiveBayes` is a "Naive Bayes" on page 33-445 classifier for multiclass learning. Trained `ClassificationNaiveBayes` classifiers store the training data, parameter values, data distribution, and prior probabilities. Use these classifiers to perform tasks such as estimating resubstitution predictions (see `resubPredict`) and predicting labels or posterior probabilities for new data (see `predict`).

Creation

Create a `ClassificationNaiveBayes` object by using `fitcnb`.

Properties

Predictor Properties

`PredictorNames` — Predictor names
cell array of character vectors

This property is read-only.

Predictor names, specified as a cell array of character vectors. The order of the elements in `PredictorNames` corresponds to the order in which the predictor names appear in the training data `X`.

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

If the model uses dummy variable encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`CategoricalPredictors` — Categorical predictor indices
`[]` | vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`CategoricalLevels` — Multivariate multinomial levels

cell array

This property is read-only.

Multivariate multinomial levels, specified as a cell array. The length of `CategoricalLevels` is equal to the number of predictors (`size(X,2)`).

The cells of `CategoricalLevels` correspond to predictors that you specify as `'mvmn'` during training, that is, they have a multivariate multinomial distribution. Cells that do not correspond to a multivariate multinomial distribution are empty (`[]`).

If predictor *j* is multivariate multinomial, then `CategoricalLevels{`*j*`}` is a list of all distinct values of predictor *j* in the sample. `NaN`s are removed from `unique(X(:,j))`.

`X` — Unstandardized predictors

numeric matrix

This property is read-only.

Unstandardized predictors used to train the naive Bayes classifier, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one variable. The software excludes observations containing at least one missing value, and removes corresponding elements from `Y`.

Predictor Distribution Properties

`DistributionNames` — Predictor distributions

`'normal'` (default) | `'kernel'` | `'mn'` | `'mvmn'` | cell array of character vectors

This property is read-only.

Predictor distributions, specified as a character vector or cell array of character vectors. `fitcnb` uses the predictor distributions to model the predictors. This table lists the available distributions.

Value	Description
`'kernel'`	Kernel smoothing density estimate
`'mn'`	Multinomial distribution. If you specify `mn`, then all features are components of a multinomial distribution. Therefore, you cannot include `'mn'` as an element of a string array or a cell array of character vectors. For details, see “Estimated Probability for Multinomial Distribution” on page 33-446.
`'mvmn'`	Multivariate multinomial distribution. For details, see “Estimated Probability for Multivariate Multinomial Distribution” on page 33-446.
`'normal'`	Normal (Gaussian) distribution

If `DistributionNames` is a 1-by-*P* cell array of character vectors, then `fitcnb` models the feature *j* using the distribution in element *j* of the cell array.

Example: `'mn'`

Example: `{'kernel','normal','kernel'}`

Data Types: `char` | `string` | `cell`

`DistributionParameters` — Distribution parameter estimates

cell array

This property is read-only.

Distribution parameter estimates, specified as a cell array. `DistributionParameters` is a *K*-by-*D* cell array, where cell (*k*,*d*) contains the distribution parameter estimates for instances of predictor *d* in class *k*. The order of the rows corresponds to the order of the classes in the property `ClassNames`, and the order of the predictors corresponds to the order of the columns of `X`.

If class *k* has no observations for predictor *j*, then the `Distribution{k,j}` is empty (`[]`).

The elements of `DistributionParameters` depend on the distributions of the predictors. This table describes the values in `DistributionParameters{k,j}`.

Distribution of Predictor *j*	Value of Cell Array for Predictor `j` and Class `k`
`kernel`	A `KernelDistribution` model. Display properties using cell indexing and dot notation. For example, to display the estimated bandwidth of the kernel density for predictor 2 in the third class, use `Mdl.DistributionParameters{3,2}.BandWidth`.
`mn`	A scalar representing the probability that token *j* appears in class *k*. For details, see "Estimated Probability for Multinomial Distribution" on page 33-446.
`mvmn`	A numeric vector containing the probabilities for each possible level of predictor *j* in class *k*. The software orders the probabilities by the sorted order of all unique levels of predictor *j* (stored in the property `CategoricalLevels`). For more details, see "Estimated Probability for Multivariate Multinomial Distribution" on page 33-446.
`normal`	A 2-by-1 numeric vector. The first element is the sample mean and the second element is the sample standard deviation.

`Kernel` — Kernel smoother type

`'normal'` (default) | `'box'` | cell array | ...

This property is read-only.

Kernel smoother type, specified as the name of a kernel or a cell array of kernel names. The length of `Kernel` is equal to the number of predictors (`size(X,2)`). `Kernel{j}` corresponds to predictor *j* and contains a character vector describing the type of kernel smoother. If a cell is empty (`[]`), then `fitcnb` did not fit a kernel distribution to the corresponding predictor.

This table describes the supported kernel smoother types. $I\{u\}$ denotes the indicator function.

Value	Kernel	Formula
`'box'`	Box (uniform)	$f(x) = 0.5I\{\|x\| \le 1\}$

Value	Kernel	Formula
`'epanechnikov'`	Epanechnikov	$f(x) = 0.75\left(1 - x^2\right)I\{\|x\| \le 1\}$
`'normal'`	Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-0.5x^2\right)$
`'triangle'`	Triangular	$f(x) = (1 - \|x\|)I\{\|x\| \le 1\}$

Example: `'box'`

Example: `{'epanechnikov','normal'}`

Data Types: `char` | `string` | `cell`

`Support` — Kernel smoother density support
cell array

This property is read-only.

Kernel smoother density support, specified as a cell array. The length of `Support` is equal to the number of predictors (`size(X,2)`). The cells represent the regions to which `fitcnb` applies the kernel density. If a cell is empty (`[]`), then `fitcnb` did not fit a kernel distribution to the corresponding predictor.

This table describes the supported options.

Value	Description
1-by-2 numeric row vector	The density support applies to the specified bounds, for example `[L,U]`, where `L` and `U` are the finite lower and upper bounds, respectively.
`'positive'`	The density support applies to all positive real values.
`'unbounded'`	The density support applies to all real values.

`Width` — Kernel smoother window width
numeric matrix

This property is read-only.

Kernel smoother window width, specified as a numeric matrix. `Width` is a K-by-P matrix, where K is the number of classes in the data, and P is the number of predictors (`size(X,2)`).

`Width(k,j)` is the kernel smoother window width for the kernel smoothing density of predictor j within class k. `NaN`s in column j indicate that `fitcnb` did not fit predictor j using a kernel density.

Response Properties

`ClassNames` — Unique class names
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class names used in the training model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`ClassNames` has the same data type as `Y`, and has *K* elements (or rows) for character arrays. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `categorical` | `char` | `string` | `logical` | `double` | `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Class labels used to train the naive Bayes classifier, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. Each row of `Y` represents the observed classification of the corresponding row of `X`.

`Y` has the same data type as the data in `Y` used for training the model. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Training Properties

`ModelParameters` — Parameter values used to train model
object

This property is read-only.

Parameter values used to train the `ClassificationNaiveBayes` model, specified as an object. `ModelParameters` contains parameter values such as the name-value pair argument values used to train the naive Bayes classifier.

Access the properties of `ModelParameters` by using dot notation. For example, access the kernel support using `Mdl.ModelParameters.Support`.

`NumObservations` — Number of training observations
numeric scalar

This property is read-only.

Number of training observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

`Prior` — Prior probabilities
numeric vector

Prior probabilities, specified as a numeric vector. The order of the elements in `Prior` corresponds to the elements of `Mdl.ClassNames`.

`fitcnb` normalizes the prior probabilities you set using the `'Prior'` name-value pair argument, so that `sum(Prior) = 1`.

The value of `Prior` does not affect the best-fitting model. Therefore, you can reset `Prior` after training `Mdl` using dot notation.

Example: `Mdl.Prior = [0.2 0.8]`

Data Types: `double` | `single`

`W` — Observation weights

vector of nonnegative values

This property is read-only.

Observation weights, specified as a vector of nonnegative values with the same number of rows as `Y`. Each entry in `W` specifies the relative importance of the corresponding observation in `Y`. `fitcnb` normalizes the value you set for the `'Weights'` name-value pair argument, so that the weights within a particular class sum to the prior probability for that class.

Classifier Properties

`Cost` — Misclassification cost

square matrix

Misclassification cost, specified as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

The misclassification cost matrix must have zeros on the diagonal.

The value of `Cost` does not influence training. You can reset `Cost` after training `Mdl` using dot notation.

Example: `Mdl.Cost = [0 0.5 ; 1 0]`

Data Types: `double` | `single`

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters

`BayesianOptimization` object | table

This property is read-only.

Cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty if the `'OptimizeHyperparameters'` name-value pair argument is nonempty when you create the model. The value of `HyperparameterOptimizationResults` depends on the setting of the `Optimizer` field in the `HyperparameterOptimizationOptions` structure when you create the model.

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`ScoreTransform` — Classification score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Classification score transformation, specified as a character vector or function handle. This table summarizes the available character vectors.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transformation. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `Mdl.ScoreTransform = 'logit'`

Data Types: `char` | `string` | `function handle`

Object Functions

compact	Reduce size of machine learning model
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validate machine learning model
edge	Classification edge for naive Bayes classifier
incrementalLearner	Convert naive Bayes classification model to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
logp	Log unconditional probability density for naive Bayes classifier
loss	Classification loss for naive Bayes classifier
margin	Classification margins for naive Bayes classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using naive Bayes classifier
resubEdge	Resubstitution classification edge
resubLoss	Resubstitution classification loss
resubMargin	Resubstitution classification margin
resubPredict	Classify training data using trained classifier
shapley	Shapley values

testckfold	Compare accuracies of two classification models by repeated cross-validation

Examples

Train Naive Bayes Classifier

Create a naive Bayes classifier for Fisher's iris data set. Then, specify prior probabilities after training the classifier.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. `fitcnb` assumes each predictor is independent and fits each predictor using a normal distribution by default.

```
Mdl = fitcnb(X,Y)
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier. Some of the `Mdl` properties appear in the Command Window.

Display the properties of `Mdl` using dot notation. For example, display the class names and prior probabilities.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl.Prior
```

```
ans = 1×3

    0.3333    0.3333    0.3333
```

The order of the class prior probabilities in `Mdl.Prior` corresponds to the order of the classes in `Mdl.ClassNames`. By default, the prior probabilities are the respective relative frequencies of the classes in the data. Alternatively, you can set the prior probabilities when calling `fitcnb` by using the `'Prior'` name-value pair argument.

Set the prior probabilities after training the classifier by using dot notation. For example, set the prior probabilities to 0.5, 0.2, and 0.3, respectively.

```
Mdl.Prior = [0.5 0.2 0.3];
```

You can now use this trained classifier to perform additional tasks. For example, you can label new measurements using `predict` or cross-validate the classifier using `crossval`.

Train and Cross-Validate Naive Bayes Classifier

Train and cross-validate a naive Bayes classifier. `fitcnb` implements 10-fold cross-validation by default. Then, estimate the cross-validated classification error.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
rng('default')  % for reproducibility
```

Train and cross-validate a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
CVMdl = fitcnb(X,Y,'ClassNames',{'b','g'},'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedModel
    CrossValidatedModel: 'NaiveBayes'
         PredictorNames: {1x32 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedModel` cross-validated, naive Bayes classifier. Alternatively, you can cross-validate a trained `ClassificationNaiveBayes` model by passing it to `crossval`.

Display the first training fold of `CVMdl` using dot notation.

```
CVMdl.Trained{1}
```

```
ans =
  CompactClassificationNaiveBayes
```

```
          ResponseName: 'Y'
 CategoricalPredictors: []
            ClassNames: {'b'  'g'}
        ScoreTransform: 'none'
     DistributionNames: {1x32 cell}
DistributionParameters: {2x32 cell}


  Properties, Methods
```

Each fold is a `CompactClassificationNaiveBayes` model trained on 90% of the data.

Full and compact naive Bayes models are not used for predicting on new data. Instead, use them to estimate the generalization error by passing `CVMdl` to `kfoldLoss`.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.1852
```

On average, the generalization error is approximately 19%.

You can specify a different conditional distribution for the predictors, or tune the conditional distribution parameters to reduce the generalization error.

More About

Bag-of-Tokens Model

In the bag-of-tokens model, the value of predictor j is the nonnegative number of occurrences of token j in the observation. The number of categories (bins) in the multinomial model is the number of distinct tokens (number of predictors).

Naive Bayes

Naive Bayes is a classification algorithm that applies density estimation to the data.

The algorithm leverages Bayes theorem, and (naively) assumes that the predictors are conditionally independent, given the class. Although the assumption is usually violated in practice, naive Bayes classifiers tend to yield posterior distributions that are robust to biased class density estimates, particularly where the posterior is 0.5 (the decision boundary) [1].

Naive Bayes classifiers assign observations to the most probable class (in other words, the maximum a posteriori decision rule). Explicitly, the algorithm takes these steps:

1. Estimate the densities of the predictors within each class.
2. Model posterior probabilities according to Bayes rule. That is, for all $k = 1,...,K$,

$$\widehat{P}(Y = k \mid X_1, .., X_P) = \frac{\pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)},$$

where:

- Y is the random variable corresponding to the class index of an observation.
- $X_1,...,X_P$ are the random predictors of an observation.
- $\pi(Y = k)$ is the prior probability that a class index is k.

3 Classify an observation by estimating the posterior probability for each class, and then assign the observation to the class yielding the maximum posterior probability.

If the predictors compose a multinomial distribution, then the posterior probability $\widehat{P}(Y = k|X_1, .., X_P) \propto \pi(Y = k)P_{mn}(X_1, ..., X_P|Y = k)$, where $P_{mn}(X_1, ..., X_P|Y = k)$ is the probability mass function of a multinomial distribution.

Algorithms

Estimated Probability for Multinomial Distribution

If you specify `'DistributionNames','mn'` when training `Mdl` using `fitcnb`, then the software fits a multinomial distribution using the "Bag-of-Tokens Model" on page 33-445. The software stores the probability that token j appears in class k in the property `DistributionParameters{k,j}`. With additive smoothing [2], the estimated probability is

$$P(\text{token } j \,|\, \text{class } k) = \frac{1 + c_{j|k}}{P + c_k},$$

where:

- $c_{j|k} = n_k \dfrac{\sum_{i:\, y_i \in \text{class } k} x_{ij} w_i}{\sum_{i:\, y_i \in \text{class } k} w_i}$, which is the weighted number of occurrences of token j in class k.
- n_k is the number of observations in class k.
- w_i is the weight for observation i. The software normalizes weights within a class so that they sum to the prior probability for that class.
- $c_k = \sum_{j=1}^{P} c_{j|k}$, which is the total weighted number of occurrences of all tokens in class k.

Estimated Probability for Multivariate Multinomial Distribution

If you specify `'DistributionNames','mvmn'` when training `Mdl` using `fitcnb`, then the software takes these steps:

1 For each predictor, the software collects a list of the unique levels, stores the sorted list in `CategoricalLevels`, and considers each level a bin. Each combination of predictor and class is a separate, independent multinomial random variable.

2 For predictor j in class k, the software counts instances of each categorical level using the list stored in `CategoricalLevels{j}`.

3 The software stores the probability that predictor j in class k has level L in the property `DistributionParameters{k,j}`, for all levels in `CategoricalLevels{j}`. With additive smoothing [2], the estimated probability is

$$P(\text{predictor } j = L \,|\, \text{class } k) = \frac{1 + m_{j|k}(L)}{m_j + m_k},$$

where:

- $m_{j|k}(L) = n_k \frac{\sum_{i:y_i \in \text{class } k} I\{x_{ij} = L\} w_i}{\sum_{i:y_i \in \text{class } k} w_i}$, which is the weighted number of observations for which predictor j equals L in class k.
- n_k is the number of observations in class k.
- $I\{x_{ij} = L\} = 1$ if $x_{ij} = L$, and 0 otherwise.
- w_i is the weight for observation i. The software normalizes weights within a class so that they sum to the prior probability for that class.
- m_j is the number of distinct levels in predictor j.
- m_k is the weighted number of observations in class k.

References

[1] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7.

[2] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*, NY: Cambridge University Press, 2008.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- When you train a naive Bayes model by using `fitcnb`, the following restrictions apply.
  - The value of the `'DistributionNames'` name-value pair argument cannot contain `'mn'`.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`CompactClassificationNaiveBayes` | `fitcnb` | `loss` | `predict`

Topics

“Naive Bayes Classification” on page 21-2
“Grouping Variables” on page 2-45

Introduced in R2014b

ClassificationNeuralNetwork

Neural network model for classification

Description

A `ClassificationNeuralNetwork` object is a trained, feedforward, and fully connected neural network for classification. The first fully connected layer of the neural network has a connection from the network input (predictor data `X`), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix (`LayerWeights`) and then adds a bias vector (`LayerBiases`). An activation function follows each fully connected layer (`Activations` and `OutputLayerActivation`). The final fully connected layer and the subsequent softmax activation function produce the network's output, namely classification scores (posterior probabilities) and predicted labels. For more information, see “Neural Network Structure” on page 33-1823.

Creation

Create a `ClassificationNeuralNetwork` object by using `fitcnet`.

Properties

Neural Network Properties

`LayerSizes` — Sizes of fully connected layers
positive integer vector

This property is read-only.

Sizes of the fully connected layers in the neural network model, returned as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer. This layer always has *K* outputs, where *K* is the number of classes in `Y`.

Data Types: `single` | `double`

`LayerWeights` — Learned layer weights
cell array

This property is read-only.

Learned layer weights for the fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer weights for the *i*th fully connected layer. For example, `Mdl.LayerWeights{1}` returns the weights for the first fully connected layer of the model `Mdl`.

`LayerWeights` includes the weights for the final fully connected layer.

Data Types: `cell`

`LayerBiases` — Learned layer biases

cell array

This property is read-only.

Learned layer biases for the fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer biases for the *i*th fully connected layer. For example, `Mdl.LayerBiases{1}` returns the biases for the first fully connected layer of the model `Mdl`.

`LayerBiases` includes the biases for the final fully connected layer.

Data Types: `cell`

`Activations` — Activation functions for fully connected layers

`'relu'` | `'tanh'` | `'sigmoid'` | `'none'` | cell array of character vectors

This property is read-only.

Activation functions for the fully connected layers of the neural network model, returned as a character vector or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If `Activations` contains only one activation function, then it is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer. The activation function for the final fully connected layer is always softmax (`OutputLayerActivation`).
- If `Activations` is an array of activation functions, then the *i*th element is the activation function for the *i*th layer of the neural network model.

Data Types: `char` | `cell`

`OutputLayerActivation` — Activation function for final fully connected layer

`'softmax'`

This property is read-only.

Activation function for the final fully connected layer, returned as `'softmax'`. The function takes each input x_i and returns the following, where K is the number of classes in the response variable:

$$f(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)}.$$

The results correspond to the predicted classification scores (or posterior probabilities).

`ModelParameters` — Parameter values used to train model
`NeuralNetworkParams` object

This property is read-only.

Parameter values used to train the `ClassificationNeuralNetwork` model, returned as a `NeuralNetworkParams` object. `ModelParameters` contains parameter values such as the name-value arguments used to train the neural network classifier.

Access the properties of `ModelParameters` by using dot notation. For example, access the function used to initialize the fully connected layer weights of a model `Mdl` by using `Mdl.ModelParameters.LayerWeightsInitializer`.

Convergence Control Properties

`ConvergenceInfo` — Convergence information
structure array

This property is read-only.

Convergence information, returned as a structure array.

Field	Description
`Iterations`	Number of training iterations used to train the neural network model
`TrainingLoss`	Training cross-entropy loss for the returned model, or `resubLoss(Mdl,'LossFun','crossentropy')` for model `Mdl`
`Gradient`	Gradient of the loss function with respect to the weights and biases at the iteration corresponding to the returned model
`Step`	Step size at the iteration corresponding to the returned model
`Time`	Total time spent across all iterations (in seconds)
`ValidationLoss`	Validation cross-entropy loss for the returned model
`ValidationChecks`	Maximum number of times in a row that the validation loss was greater than or equal to the minimum validation loss
`ConvergenceCriterion`	Criterion for convergence

Field	Description
`History`	See `TrainingHistory`

Data Types: `struct`

`TrainingHistory` — Training history
table

This property is read-only.

Training history, returned as a table.

Column	Description
`Iteration`	Training iteration
`TrainingLoss`	Training cross-entropy loss for the model at this iteration
`Gradient`	Gradient of the loss function with respect to the weights and biases at this iteration
`Step`	Step size at this iteration
`Time`	Time spent during this iteration (in seconds)
`ValidationLoss`	Validation cross-entropy loss for the model at this iteration
`ValidationChecks`	Running total of times that the validation loss is greater than or equal to the minimum validation loss

Data Types: `table`

`Solver` — Solver used to train neural network model
`'LBFGS'`

This property is read-only.

Solver used to train the neural network model, returned as `'LBFGS'`. To create a `ClassificationNeuralNetwork` model, `fitcnet` uses a limited-memory Broyden-Flecter-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) as its loss function minimization technique, where the software minimizes the cross-entropy loss.

Predictor Properties

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, returned as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | [ ]

This property is read-only.

Categorical predictor indices, returned as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, returned as a cell array of character vectors. If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`X` — Unstandardized predictors
numeric matrix | table

This property is read-only.

Unstandardized predictors used to train the neural network model, returned as a numeric matrix or table. `X` retains its original orientation, with observations in rows or columns depending on the value of the `ObservationsIn` name-value argument in the call to `fitcnet`.

Data Types: `single` | `double` | `table`

Response Properties

`ClassNames` — Unique class names
numeric vector | categorical vector | logical vector | character array | cell array of character vectors

This property is read-only.

Unique class names used in training, returned as a numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, returned as a character vector.

Data Types: `char`

`Y` — Class labels
numeric vector | categorical vector | logical vector | character array | cell array of character vectors

This property is read-only.

Class labels used to train the model, returned as a numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. `Y` has the same data type as the response variable used to train the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the classification of the corresponding observation in `X`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `cell`

Other Data Properties

`NumObservations` — Number of observations

positive numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, returned as a positive numeric scalar.

Data Types: `double`

`RowsUsed` — Rows used in fitting

`[]` | logical vector

This property is read-only.

Rows of the original training data used in fitting the model, returned as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

`W` — Observation weights

numeric vector

This property is read-only.

Observation weights used to train the model, returned as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `Weights` name-value argument so that the elements of `W` within a particular class sum up to the prior probability of that class.

Data Types: `single` | `double`

Other Classification Properties

`Cost` — Misclassification cost

numeric square matrix

This property is read-only.

Misclassification cost, returned as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The cost matrix always has this form: `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`Prior` — Prior probabilities

numeric vector

This property is read-only.

Prior probabilities for each class, returned as a numeric vector. The order of the elements of `Prior` corresponds to the elements of `ClassNames`.

Data Types: `double`

`ScoreTransform` — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

Object Functions

compact Reduce size of machine learning model

compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validate machine learning model
edge	Classification edge for neural network classifier
loss	Classification loss for neural network classifier
margin	Classification margins for neural network classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using neural network classifier
resubEdge	Resubstitution classification edge
resubLoss	Resubstitution classification loss
resubMargin	Resubstitution classification margin
resubPredict	Classify training data using trained classifier

Examples

Train Neural Network Classifier

Train a neural network classifier, and assess the performance of the classifier on a test set.

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable("CreditRating_Historical.dat");
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____     _____     _______    ________    _____    ________    _______

    62394     0.013     0.104     0.036       0.447      0.142        3       {'BB' }
    48608     0.232     0.335     0.062       1.969      0.281        8       {'A'  }
    42444     0.311     0.367     0.074       1.935      0.366        1       {'A'  }
    48631     0.194     0.263     0.062       1.017      0.228        4       {'BBB'}
    43768     0.121     0.413     0.057       3.647      0.466       12       {'AAA'}
    39255    -0.117    -0.799      0.01       0.179      0.082        4       {'CCC'}
    62236     0.087     0.158     0.049       0.816      0.324        2       {'BBB'}
    39354     0.005     0.181     0.034       2.597      0.388        7       {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,"ID");
creditrating.Industry = categorical(creditrating.Industry);
```

Convert the `Rating` response variable to an ordinal `categorical` variable.

```
creditrating.Rating = categorical(creditrating.Rating, ...
    ["AAA","AA","A","BBB","BB","B","CCC"],"Ordinal",true);
```

Partition the data into training and test sets. Use approximately 80% of the observations to train a neural network model, and 20% of the observations to test the performance of the trained model on new data. Use `cvpartition` to partition the data.

```
rng("default") % For reproducibility of the partition
c = cvpartition(creditrating.Rating,"Holdout",0.20);
trainingIndices = training(c); % Indices for the training set
testIndices = test(c); % Indices for the test set
creditTrain = creditrating(trainingIndices,:);
creditTest = creditrating(testIndices,:);
```

Train a neural network classifier by passing the training data `creditTrain` to the `fitcnet` function.

```
Mdl = fitcnet(creditTrain,"Rating")
```

```
Mdl =
  ClassificationNeuralNetwork
           PredictorNames: {'WC_TA'  'RE_TA'  'EBIT_TA'  'MVE_BVTD'  'S_TA'  'Industry'}
             ResponseName: 'Rating'
    CategoricalPredictors: 6
               ClassNames: [AAA    AA    A    BBB    BB    B    CCC]
          ScoreTransform: 'none'
         NumObservations: 3146
              LayerSizes: 10
             Activations: 'relu'
    OutputLayerActivation: 'softmax'
                  Solver: 'LBFGS'
        ConvergenceInfo: [1×1 struct]
        TrainingHistory: [1000×7 table]


  Properties, Methods
```

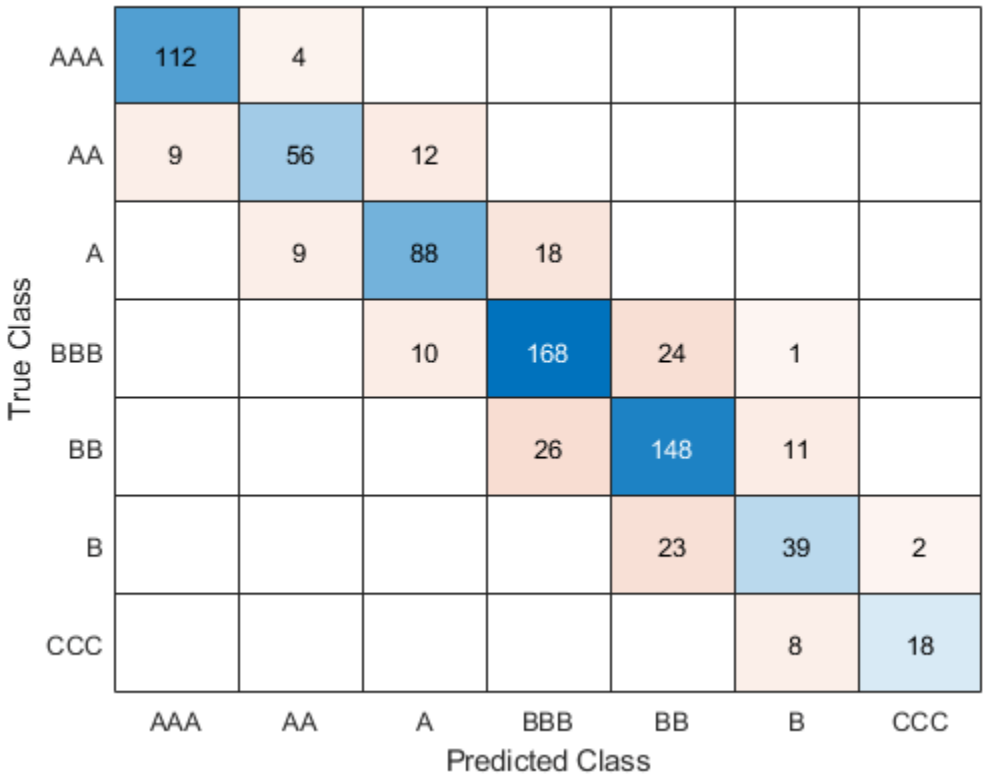
`Mdl` is a trained `ClassificationNeuralNetwork` classifier. You can use dot notation to access the properties of `Mdl`. For example, you can specify `Mdl.TrainingHistory` to get more information about the training history of the neural network model.

Evaluate the performance of the classifier on the test set by computing the test set classification error. Visualize the results by using a confusion matrix.

```
testAccuracy = 1 - loss(Mdl,creditTest,"Rating", ...
    "LossFun","classiferror")
```

```
testAccuracy = 0.8003
```

```
confusionchart(creditTest.Rating,predict(Mdl,creditTest))
```



Specify Neural Network Classifier Architecture

Specify the structure of a neural network classifier, including the size of the fully connected layers.

Load the `ionosphere` data set, which includes radar signal data. `X` contains the predictor data, and `Y` is the response variable, whose values represent either good ("g") or bad ("b") radar signals.

```
load ionosphere
```

Separate the data into training data (`XTrain` and `YTrain`) and test data (`XTest` and `YTest`) by using a stratified holdout partition. Reserve approximately 30% of the observations for testing, and use the rest of the observations for training.

```
rng("default") % For reproducibility of the partition
cvp = cvpartition(Y,"Holdout",0.3);
XTrain = X(training(cvp),:);
YTrain = Y(training(cvp));
XTest = X(test(cvp),:);
YTest = Y(test(cvp));
```

Train a neural network classifier. Specify to have 35 outputs in the first fully connected layer and 20 outputs in the second fully connected layer. By default, both layers use a rectified linear unit (ReLU) activation function. You can change the activation functions for the fully connected layers by using the `Activations` name-value argument.

```
Mdl = fitcnet(XTrain,YTrain, ...
    "LayerSizes",[35 20])

Mdl =
  ClassificationNeuralNetwork
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 246
               LayerSizes: [35 20]
              Activations: 'relu'
    OutputLayerActivation: 'softmax'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [47×7 table]


  Properties, Methods
```

Access the weights and biases for the fully connected layers of the trained classifier by using the `LayerWeights` and `LayerBiases` properties of `Mdl`. The first two elements of each property correspond to the values for the first two fully connected layers, and the third element corresponds to the values for the final fully connected layer with a softmax activation function for classification. For example, display the weights and biases for the second fully connected layer.

```
Mdl.LayerWeights{2}

ans = 20×35

    0.0481    0.2501   -0.1535   -0.0934    0.0760   -0.0579   -0.2465    1.0411    0.3712   -1.2
   -0.9489   -1.8343    0.5510   -0.5751   -0.8726    0.8815    0.0203   -1.6379    2.0315    1.7
   -0.1910    0.0246   -0.3511    0.0097    0.3160   -0.0693    0.2270   -0.0783   -0.1626   -0.3
   -0.0415   -0.0059   -0.0753   -0.1477   -0.1621   -0.1762    0.2164    0.1710   -0.0610   -0.1
    1.1848    1.6142   -0.1352    0.5774    0.5491    0.0103    0.0209    0.7219   -0.8643   -0.5
    0.2486   -0.2920   -0.0004    0.2806    0.2987   -0.2709    0.1473   -0.2580   -0.0499   -0.0
   -0.0516    0.0640    0.1824   -0.0675   -0.2065   -0.0052   -0.1682   -0.1520    0.0060    0.0
   -0.6192   -0.7804   -0.0506   -0.4205   -0.2584   -0.2020   -0.0008    0.0534    1.0185   -0.0
    0.5049   -0.1362   -0.2218    0.1637   -0.1282   -0.1008    0.1445    0.4527   -0.4887    0.0
    1.1109   -0.0466    0.4044    0.6366    0.1863    0.5660    0.2839    0.8793   -0.5497    0.0
      ⋮


Mdl.LayerBiases{2}

ans = 20×1

    0.6147
    0.1891
   -0.2767
   -0.2977
    1.3655
    0.0347
    0.1509
   -0.4839
   -0.3960
    0.9248
```

```
      ⋮
```

The final fully connected layer has two outputs, one for each class in the response variable. The number of layer outputs corresponds to the first dimension of the layer weights and layer biases.

```
size(Mdl.LayerWeights{end})
```

```
ans = 1×2

     2    20
```

```
size(Mdl.LayerBiases{end})
```

```
ans = 1×2

     2     1
```

To estimate the performance of the trained classifier, compute the test set classification error for `Mdl`.

```
testError = loss(Mdl,XTest,YTest, ...
    "LossFun","classiferror")
```

```
testError = 0.0774
```

```
accuracy = 1 - testError
```

```
accuracy = 0.9226
```

`Mdl` accurately classifies approximately 92% of the observations in the test set.

See Also

`ClassificationPartitionedModel` | `CompactClassificationNeuralNetwork` | `edge` | `fitcnet` | `loss` | `margin` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

ClassificationPartitionedECOC

Cross-validated multiclass ECOC model for support vector machines (SVMs) and other classifiers

Description

`ClassificationPartitionedECOC` is a set of error-correcting output codes (ECOC) models trained on cross-validated folds. Estimate the quality of the cross-validated classification by using one or more "kfold" functions: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, `kfoldEdge`, and `kfoldfun`.

Every "kfold" method uses models trained on training-fold (in-fold) observations to predict the response for validation-fold (out-of-fold) observations. For example, suppose you cross-validate using five folds. In this case, the software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (roughly 4/5 of the data), and the validation fold contains the other group (roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) by using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) by using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third, fourth, and fifth models.

If you validate by using `kfoldPredict`, the software computes predictions for the observations in group *i* by using the *i*th model. In short, the software estimates a response for every observation by using the model trained without that observation.

Creation

You can create a `ClassificationPartitionedECOC` model in two ways:

- Create a cross-validated ECOC model from an ECOC model by using the `crossval` object function.
- Create a cross-validated ECOC model by using the `fitcecoc` function and specifying one of the name-value pair arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
character vector

Cross-validated model name, specified as a character vector.

For example, `'ECOC'` specifies a cross-validated ECOC model.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

Cross-validation parameter values, specified as an object. The parameter values correspond to the name-value pair argument values used to cross-validate the ECOC classifier. `ModelParameters` does not contain estimated parameters.

You can access the properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations
positive numeric scalar

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Compact classifiers trained on cross-validation folds
cell array of `CompactClassificationECOC` models

Compact classifiers trained on cross-validation folds, specified as a cell array of `CompactClassificationECOC` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

`X` — Unstandardized predictor data
numeric matrix | table

Unstandardized predictor data used to cross-validate the classifier, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

Y — Observed class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Observed class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements and has the same data type as the input argument `Y` that you pass to `fitcecoc` to cross-validate the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding row of `X`.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

ECOC Properties

BinaryLoss — Binary learner loss function

`'binodeviance'` | `'exponential'` | `'hamming'` | `'hinge'` | `'linear'` | `'logit'` | `'quadratic'`

Binary learner loss function, specified as a character vector representing the loss function name.

If you train using binary learners that use different loss functions, then the software sets `BinaryLoss` to `'hamming'`. To potentially increase accuracy, specify a binary loss function other than the default during a prediction or loss computation by using the `'BinaryLoss'` name-value pair argument of `kfoldPredict` or `kfoldLoss`.

Data Types: `char`

BinaryY — Binary learner class labels

numeric matrix | `[]`

Binary learner class labels, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across all folds, then `BinaryY` is a `NumObservations`-by-*L* matrix, where *L* is the number of binary learners (`size(CodingMatrix,2)`).

  The elements of `BinaryY` are `–1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observation `k` to a dichotomous class corresponding to the value of `BinaryY(k,j)`.

Value	Dichotomous Class Assignment
`–1`	Learner `j` assigns observation `k` to a negative class.
`0`	Before training, learner `j` removes observation `k` from the data set.
`1`	Learner `j` assigns observation `k` to a positive class.

- If the coding matrix varies across folds, then `BinaryY` is empty (`[]`).

Data Types: `double`

CodingMatrix — Codes specifying class assignments

numeric matrix | `[]`

Codes specifying class assignments for the binary learners, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across all folds, then `CodingMatrix` is a *K*-by-*L* matrix, where *K* is the number of classes and *L* is the number of binary learners.

  The elements of `CodingMatrix` are `-1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observations in class `i` to a dichotomous class corresponding to the value of `CodingMatrix(i,j)`.

Value	Dichotomous Class Assignment
`-1`	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

- If the coding matrix varies across folds, then `CodingMatrix` is empty (`[]`). You can obtain the coding matrix for each fold by using the `Trained` property. For example, `CVMdl.Trained{1}.CodingMatrix` is the coding matrix in the first fold of the cross-validated ECOC model `CVMdl`.

Data Types: `double` | `single` | `int8` | `int16` | `int32` | `int64`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has *K* rows and columns, where *K* is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data `X`, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns in `X`.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as the number of classes in `ClassNames`, and the order of the elements corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function to apply to predicted scores
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$

Value	Description
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  `Mdl.ScoreTransform = @function;`

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

Object Functions

kfoldEdge	Classification edge for cross-validated ECOC model
kfoldLoss	Classification loss for cross-validated ECOC model
kfoldMargin	Classification margins for cross-validated ECOC model
kfoldPredict	Classify observations in cross-validated ECOC model
kfoldfun	Cross-validate function using cross-validated ECOC model

Examples

Cross-Validate ECOC Classifier

Cross-validate an ECOC classifier with SVM binary learners, and estimate the generalized classification error.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
rng(1); % For reproducibility
```

Create an SVM template, and standardize the predictors.

```
t = templateSVM('Standardize',true)
```

```
t =
Fit template for classification SVM.

                    Alpha: [0x1 double]
              BoxConstraint: []
                  CacheSize: []
              CachingMethod: ''
                 ClipAlphas: []
    DeltaGradientTolerance: []
                    Epsilon: []
              GapTolerance: []
              KKTTolerance: []
            IterationLimit: []
            KernelFunction: ''
               KernelScale: []
```

```
          KernelOffset: []
KernelPolynomialOrder: []
              NumPrint: []
                    Nu: []
       OutlierFraction: []
     RemoveDuplicates: []
      ShrinkagePeriod: []
               Solver: ''
      StandardizeData: 1
   SaveSupportVectors: []
       VerbosityLevel: []
              Version: 2
               Method: 'SVM'
                 Type: 'classification'
```

`t` is an SVM template. Most of the template object properties are empty. When training the ECOC classifier, the software sets the applicable properties to their default values.

Train the ECOC classifier, and specify the class order.

```
Mdl = fitcecoc(X,Y,'Learners',t,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a `ClassificationECOC` classifier. You can access its properties using dot notation.

Cross-validate `Mdl` using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

`CVMdl` is a `ClassificationPartitionedECOC` cross-validated ECOC classifier.

Estimate the generalized classification error.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.0400
```

The generalized classification error is 4%, which indicates that the ECOC classifier generalizes fairly well.

Speed Up Training ECOC Classifiers Using Binning and Parallel Computing

Train a one-versus-all ECOC classifier using a `GentleBoost` ensemble of decision trees with surrogate splits. To speed up training, bin numeric predictors and use parallel computing. Binning is valid only when `fitcecoc` uses a tree learner. After training, estimate the classification error using 10-fold cross-validation. Note that parallel computing requires Parallel Computing Toolbox™.

Load Sample Data

Load and inspect the `arrhythmia` data set.

```
load arrhythmia
[n,p] = size(X)
```

```
n = 452
```

```
p = 279

isLabels = unique(Y);
nLabels = numel(isLabels)

nLabels = 13

tabulate(categorical(Y))

  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

The data set contains `279` predictors, and the sample size of `452` is relatively small. Of the 16 distinct labels, only 13 are represented in the response (`Y`). Each label describes various degrees of arrhythmia, and 54.20% of the observations are in class `1`.

Train One-Versus-All ECOC Classifier

Create an ensemble template. You must specify at least three arguments: a method, a number of learners, and the type of learner. For this example, specify `'GentleBoost'` for the method, `100` for the number of learners, and a decision tree template that uses surrogate splits because there are missing observations.

```
tTree = templateTree('surrogate','on');
tEnsemble = templateEnsemble('GentleBoost',100,tTree);
```

`tEnsemble` is a template object. Most of its properties are empty, but the software fills them with their default values during training.

Train a one-versus-all ECOC classifier using the ensembles of decision trees as binary learners. To speed up training, use binning and parallel computing.

- Binning (`'NumBins',50`) — When you have a large training data set, you can speed up training (a potential decrease in accuracy) by using the `'NumBins'` name-value pair argument. This argument is valid only when `fitcecoc` uses a tree learner. If you specify the `'NumBins'` value, then the software bins every numeric predictor into a specified number of equiprobable bins, and then grows trees on the bin indices instead of the original data. You can try `'NumBins',50` first, and then change the `'NumBins'` value depending on the accuracy and training speed.
- Parallel computing (`'Options',statset('UseParallel',true)`) — With a Parallel Computing Toolbox license, you can speed up the computation by using parallel computing, which sends each binary learner to a worker in the pool. The number of workers depends on your system configuration. When you use decision trees for binary learners, `fitcecoc` parallelizes training using Intel® Threading Building Blocks (TBB) for dual-core systems and above. Therefore, specifying the `'UseParallel'` option is not helpful on a single computer. Use this option on a cluster.

Additionally, specify that the prior probabilities are $1/K$, where $K = 13$ is the number of distinct classes.

```
options = statset('UseParallel',true);
Mdl = fitcecoc(X,Y,'Coding','onevsall','Learners',tEnsemble,...
                'Prior','uniform','NumBins',50,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

`Mdl` is a `ClassificationECOC` model.

Cross-Validation

Cross-validate the ECOC classifier using 10-fold cross-validation.

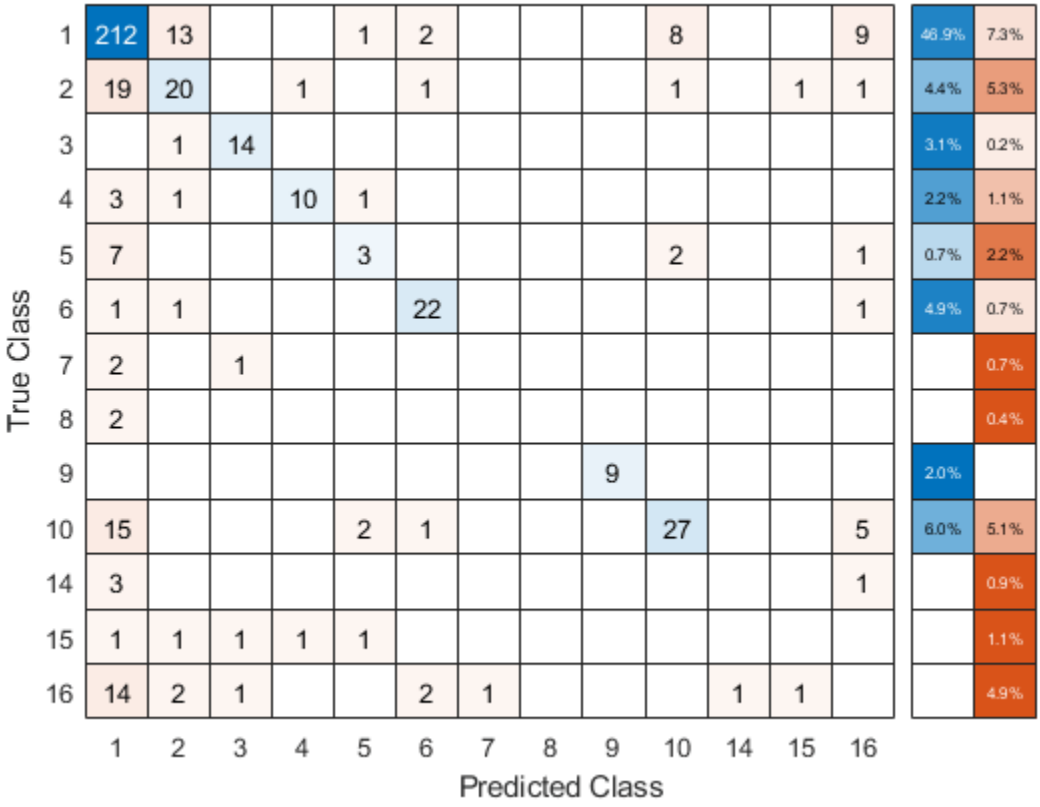
```
CVMdl = crossval(Mdl,'Options',options);
```

```
Warning: One or more folds do not contain points from all the groups.
```

`CVMdl` is a `ClassificationPartitionedECOC` model. The warning indicates that some classes are not represented while the software trains at least one fold. Therefore, those folds cannot predict labels for the missing classes. You can inspect the results of a fold using cell indexing and dot notation. For example, access the results of the first fold by entering `CVMdl.Trained{1}`.

Use the cross-validated ECOC classifier to predict validation-fold labels. You can compute the confusion matrix by using `confusionchart`. Move and resize the chart by changing the inner position property to ensure that the percentages appear in the row summary.

```
oofLabel = kfoldPredict(CVMdl,'Options',options);
ConfMat = confusionchart(Y,oofLabel,'RowSummary','total-normalized');
ConfMat.InnerPosition = [0.10 0.12 0.85 0.85];
```



Reproduce Binned Data

Reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

```
X = Mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = Mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are `0` for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `crossval` | `cvpartition` | `fitcecoc`

Introduced in R2014b

ClassificationPartitionedEnsemble

Package: `classreg.learning.partition`
Superclasses: `ClassificationPartitionedModel`

Cross-validated classification ensemble

Description

`ClassificationPartitionedEnsemble` is a set of classification ensembles trained on cross-validated folds. Estimate the quality of classification by cross validation using one or more “kfold” methods: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, `kfoldEdge`, and `kfoldfun`.

Every “kfold” method uses models trained on in-fold observations to predict response for out-of-fold observations. For example, suppose you cross validate using five folds. In this case, every training fold contains roughly 4/5 of the data and every test fold contains roughly 1/5 of the data. The first model stored in `Trained{1}` was trained on `X` and `Y` with the first 1/5 excluded, the second model stored in `Trained{2}` was trained on `X` and `Y` with the second 1/5 excluded, and so on. When you call `kfoldPredict`, it computes predictions for the first 1/5 of the data using the first model, for the second 1/5 of data using the second model, and so on. In short, response for every observation is computed by `kfoldPredict` using the model trained without this observation.

Construction

`cvens = crossval(ens)` creates a cross-validated ensemble from `ens`, a classification ensemble. For syntax details, see the `crossval` method reference page.

`cvens = fitcensemble(X,Y,Name,Value)` creates a cross-validated ensemble when `Name` is one of `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`. For syntax details, see the `fitcensemble` function reference page.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
```

```
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a numeric vector, vector of categorical variables, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`Combiner`

Cell array of combiners across all folds.

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response. This property is read-only.

`CrossValidatedModel`

Name of the cross-validated model, a character vector.

`KFold`

Number of folds used in a cross-validated ensemble, a positive integer.

`ModelParameters`

Object holding parameters of `cvens`.

`NumObservations`

Number of data points used in training the ensemble, a positive integer.

`NumTrainedPerFold`

Number of weak learners used in training each fold of the ensemble, a positive integer.

`Partition`

Partition of class `cvpartition` used in creating the cross-validated ensemble.

`PredictorNames`

Cell array of names for the predictor variables, in the order in which they appear in `X`.

`Prior`

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`. The number of elements of `Prior` is the number of unique classes in the response. This property is read-only.

`ResponseName`

Name of the response variable `Y`, a character vector.

`ScoreTransform`

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitctree`.

Add or change a `ScoreTransform` function using dot notation:

```
ens.ScoreTransform = 'function'
```

or

```
ens.ScoreTransform = @function
```

`Trainable`

Cell array of ensembles trained on cross-validation folds. Every ensemble is full, meaning it contains its training data and weights.

`Trained`

Cell array of compact ensembles trained on cross-validation folds.

`W`

Scaled `weights`, a vector with length `n`, the number of rows in `X`.

`X`

A matrix or table of predictor values. Each column of `X` represents one variable, and each row represents one observation.

Y

Numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. Each row of Y is the response to the data in the corresponding row of X.

Object Functions

kfoldEdge	Classification edge for cross-validated classification model
kfoldLoss	Classification loss for cross-validated classification model
kfoldMargin	Classification margins for cross-validated classification model
kfoldPredict	Classify observations in cross-validated classification model
kfoldfun	Cross-validate function for classification
resume	Resume training learners on cross-validation folds

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Evaluate K-Fold Cross-Validation Error for Classification Ensemble

Evaluate the k-fold cross-validation error for a classification ensemble that models the Fisher iris data.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Create a cross-validated ensemble from `ens` and find the k-fold cross-validation error.

```
rng(10,'twister') % For reproducibility
cvens = crossval(ens);
L = kfoldLoss(cvens)
```

```
L = 0.0533
```

See Also

`ClassificationEnsemble` | `ClassificationPartitionedModel` | `RegressionPartitionedEnsemble` | `fitctree`

ClassificationPartitionedGAM

Cross-validated generalized additive model (GAM) for classification

Description

`ClassificationPartitionedGAM` is a set of generalized additive models trained on cross-validated folds. Estimate the quality of the cross-validated classification by using one or more *kfold* functions: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, `kfoldEdge`, and `kfoldfun`.

Every *kfold* object function uses models trained on training-fold (in-fold) observations to predict the response for validation-fold (out-of-fold) observations. For example, suppose you cross-validate using five folds. The software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (roughly 4/5 of the data), and the validation fold contains the other group (roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) by using the observations in the last four groups, and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) by using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar manner for the third, fourth, and fifth models.

If you validate by using `kfoldPredict`, the software computes predictions for the observations in group *i* by using the *i*th model. In short, the software estimates a response for every observation by using the model trained without that observation.

Creation

You can create a `ClassificationPartitionedGAM` model in two ways:

- Create a cross-validated model from a GAM object `ClassificationGAM` by using the `crossval` object function.
- Create a cross-validated model by using the `fitcgam` function and specifying one of the name-value arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
`'GAM'`

This property is read-only.

Cross-validated model name, specified as `'GAM'`.

`KFold` — Number of cross-validated folds
positive integer

This property is read-only.

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

This property is read-only.

Cross-validation parameter values, specified as an object. The parameter values correspond to the values of the name-value arguments used to cross-validate the generalized additive model. `ModelParameters` does not contain estimated parameters.

You can access the properties of `ModelParameters` using dot notation.

`Partition` — Data partition
`cvpartition` model

This property is read-only.

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Compact classifiers trained on cross-validation folds
cell array of `CompactClassificationGAM` models

This property is read-only.

Compact classifiers trained on cross-validation folds, specified as a cell array of `CompactClassificationGAM` model objects. `Trained` has *k* cells, where *k* is the number of folds.

Data Types: `cell`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Cost` — Misclassification costs

2-by-2 numeric matrix

Misclassification costs, specified as a 2-by-2 numeric matrix.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

The software uses the `Cost` value for prediction, but not training. You can change the value by using dot notation.

Example: `Mdl.Cost = C;`

Data Types: `double`

`NumObservations` — Number of observations

numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

Data Types: `double`

`PredictorNames` — Predictor variable names

cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`Prior` — Prior class probabilities

numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector with two elements. The order of the elements corresponds to the order of the elements in `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

This property determines the output score computation for object functions such as `kfoldPredict`, `kfoldMargin`, and `kfoldEdge`. Use `'logit'` to compute posterior probabilities, and use `'none'` to compute the logit of posterior probabilities.

Data Types: `char` | `function_handle`

`W` — Observation weights

numeric vector

This property is read-only.

Observation weights used to train the model, specified as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `'Weights'` name-value argument so that the elements of `W` within a particular class sum up to the prior probability of that class.

Data Types: `double`

X — Predictors

numeric matrix | table

This property is read-only.

Predictors used to cross-validate the model, specified as a numeric matrix or table.

Each row of X corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

Y — Class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. Y has the same data type as the response variable used to train the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of Y represents the observed classification of the corresponding row of X.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

Object Functions

kfoldPredict	Classify observations in cross-validated classification model
kfoldLoss	Classification loss for cross-validated classification model
kfoldMargin	Classification margins for cross-validated classification model
kfoldEdge	Classification edge for cross-validated classification model
kfoldfun	Cross-validate function for classification

Examples

Create Cross-Validated GAM Using `fitcgam`

Train a cross-validated GAM with 10 folds, which is the default cross-validation option, by using `fitcgam`. Then, use `kfoldPredict` to predict class labels for validation-fold observations using a model trained on training-fold observations.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`.

```
rng('default') % For reproducibility
CVMdl = fitcgam(X,Y,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedGAM
    CrossValidatedModel: 'GAM'
         PredictorNames: {1x34 cell}
```

```
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
     NumTrainedPerFold: [1x1 struct]
            ClassNames: {'b'  'g'}
       ScoreTransform: 'logit'


  Properties, Methods
```

The `fitcgam` function creates a `ClassificationPartitionedGAM` model object `CVMdl` with 10 folds. During cross-validation, the software completes these steps:

1. Randomly partition the data into 10 sets.
2. For each set, reserve the set as validation data, and train the model using the other 9 sets.
3. Store the 10 compact, trained models in a 10-by-1 cell vector in the `Trained` property of the cross-validated model object `ClassificationPartitionedGAM`.

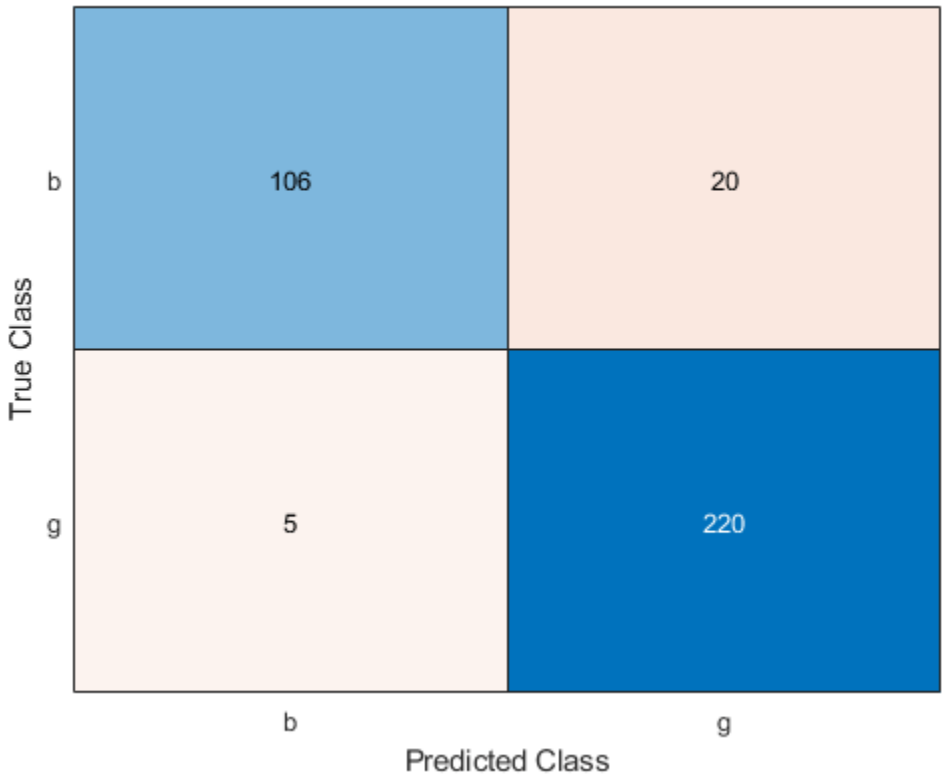
You can override the default cross-validation setting by using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument.

Classify the observations in `X` by using `kfoldPredict`. The function predicts class labels for every observation using the model trained without that observation.

```
label = kfoldPredict(CVMdl);
```

Create a confusion matrix to compare the true classes of the observations to their predicted labels.

```
C = confusionchart(Y,label);
```



Compute the classification error.

```
L = kfoldLoss(CVMdl)
```

```
L = 0.0712
```

The average misclassification rate over 10 folds is about 7%.

Create Cross-Validated GAM Using `crossval`

Train a GAM by using `fitcgam`, and create a cross-validated GAM by using `crossval` and the holdout option. Then, use `kfoldPredict` to predict responses for validation-fold observations using a model trained on training-fold observations.

Load the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

```
load census1994
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. To reduce the running time for this example, subsample 500 training observations from `adultdata` by using the `datasample` function.

```
rng('default')
NumSamples = 5e2;
adultdata = datasample(adultdata,NumSamples,'Replace',false);
```

Train a GAM that contains both linear and interaction terms for predictors. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
Mdl = fitcgam(adultdata,'salary','Interactions','all','MaxPValue',0.05);
```

`Mdl` is a `ClassificationGAM` model object.

Cross-validate the model by specifying a 30% holdout sample.

```
CVMdl = crossval(Mdl,'Holdout',0.3)
```

```
CVMdl =
  ClassificationPartitionedGAM
      CrossValidatedModel: 'GAM'
           PredictorNames: {1x14 cell}
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
             ResponseName: 'salary'
          NumObservations: 500
                    KFold: 1
                Partition: [1x1 cvpartition]
        NumTrainedPerFold: [1x1 struct]
               ClassNames: [<=50K    >50K]
           ScoreTransform: 'logit'


  Properties, Methods
```

The `crossval` function creates a `ClassificationPartitionedGAM` model object `CVMdl` with the holdout option. During cross-validation, the software completes these steps:

1. Randomly select and reserve 30% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model object `ClassificationPartitionedGAM`.

You can choose a different cross-validation setting by using the `'CrossVal'`, `'CVPartition'`, `'KFold'`, or `'Leaveout'` name-value argument.

Classify the validation-fold observations by using `kfoldPredict`. The function predicts class labels for the validation-fold observations by using the model trained on the training-fold observations. The function assigns the most frequently predicted label to the training-fold observations.

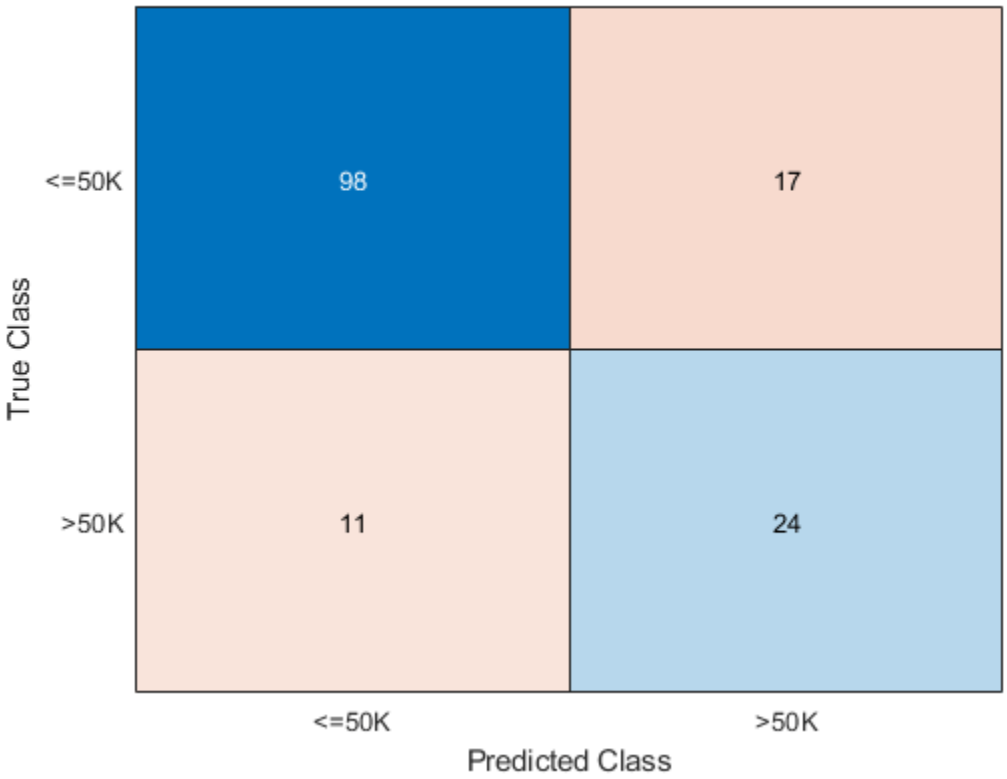
```
[labels,scores] = kfoldPredict(CVMdl);
```

Find the validation-fold observations. `kfoldPredict` returns 0 scores for both classes for the training-fold observations. Therefore, you can identify the validation-fold observations by finding the observations whose scores are all zeros.

```
idx = find(sum(abs(scores),2)~=0);
```

Create a confusion matrix to compare the true classes of the observations to their predicted labels, and compute the classification error for the validation-fold observations.

```
C = confusionchart(adultdata.salary(idx),labels(idx));
```




```
L = kfoldLoss(CVMdl)
```

```
L = 0.1867
```

Find Optimal Number of Trees for GAM Using `kfoldLoss`

Train a cross-validated generalized additive model (GAM) with 10 folds. Then, use `kfoldLoss` to compute cumulative cross-validation classification errors (misclassification rate in decimal). Use the errors to determine the optimal number of trees per predictor (linear term for predictor) and the optimal number of trees per interaction term.

Alternatively, you can find optimal values of `fitcgam` name-value arguments by using the `bayesopt` function. For an example, see “Optimize Cross-Validated GAM Using bayesopt” on page 33-1693.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
rng('default') % For reproducibility
CVMdl = fitcgam(X,Y,'CrossVal','on','Interactions','all','MaxPValue',0.05);
```

If you specify `'Mode'` as `'cumulative'` for `kfoldLoss`, then the function returns cumulative errors, which are the average errors across all folds obtained using the same number of trees for each fold. Display the number of trees for each fold.

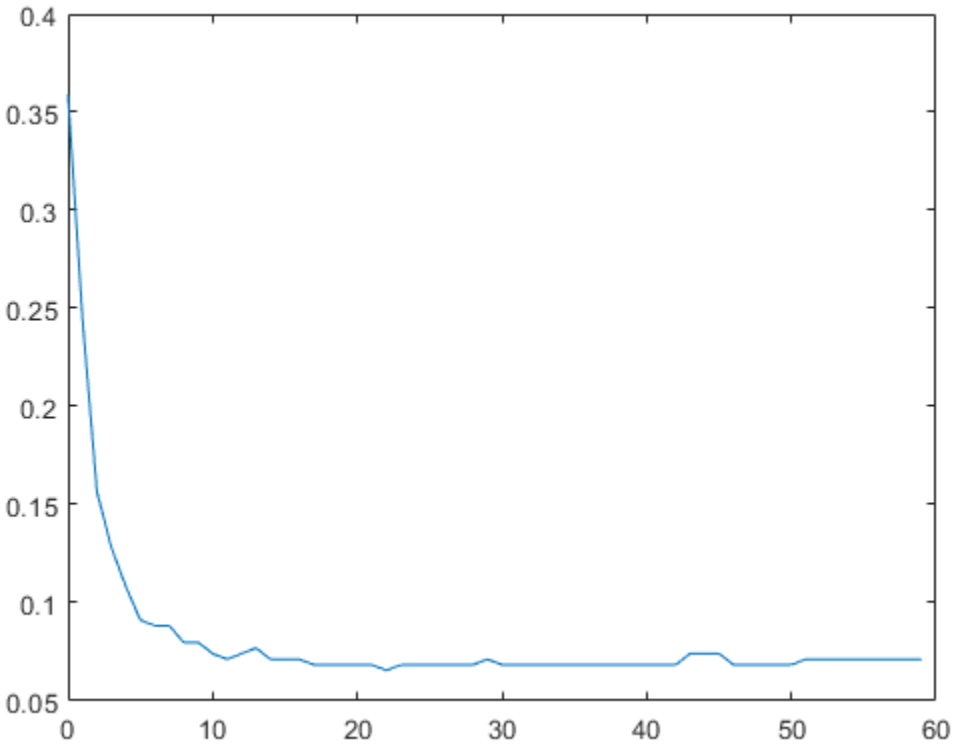
```
CVMdl.NumTrainedPerFold
```

```
ans = struct with fields:
      PredictorTrees: [65 64 59 61 60 66 65 62 64 61]
    InteractionTrees: [1 2 2 2 2 1 2 2 2 2]
```

`kfoldLoss` can compute cumulative errors using up to 59 predictor trees and one interaction tree.

Plot the cumulative, 10-fold cross-validated, classification error (misclassification rate in decimal). Specify `'IncludeInteractions'` as `false` to exclude interaction terms from the computation.

```
L_noInteractions = kfoldLoss(CVMdl,'Mode','cumulative','IncludeInteractions',false);
figure
plot(0:min(CVMdl.NumTrainedPerFold.PredictorTrees),L_noInteractions)
```



The first element of `L_noInteractions` is the average error over all folds obtained using only the intercept (constant) term. The (`J+1`)th element of `L_noInteractions` is the average error obtained using the intercept term and the first `J` predictor trees per linear term. Plotting the cumulative loss allows you to monitor how the error changes as the number of predictor trees in GAM increases.

Find the minimum error and the number of predictor trees used to achieve the minimum error.

```
[M,I] = min(L_noInteractions)
```

```
M = 0.0655
```

```
I = 23
```

The GAM achieves the minimum error when it includes 22 predictor trees.

Compute the cumulative classification error using both linear terms and interaction terms.

```
L = kfoldLoss(CVMdl,'Mode','cumulative')
```

```
L = 2×1

    0.0712
    0.0712
```

The first element of `L` is the average error over all folds obtained using the intercept (constant) term and all predictor trees per linear term. The second element of `L` is the average error obtained using the intercept term, all predictor trees per linear term, and one interaction tree per interaction term. The error does not decrease when interaction terms are added.

If you are satisfied with the error when the number of predictor trees is 22, you can create a predictive model by training the univariate GAM again and specifying `'NumTreesPerPredictor',22` without cross-validation.

See Also

`ClassificationGAM` | `crossval`

Topics

“Train Generalized Additive Model for Binary Classification” on page 12-77

Introduced in R2021a

ClassificationPartitionedKernel

Cross-validated, binary kernel classification model

Description

`ClassificationPartitionedKernel` is a binary kernel classification model, trained on cross-validated folds. You can estimate the quality of classification, or how well the kernel classification model generalizes, using one or more "kfold" functions: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, and `kfoldEdge`.

Every "kfold" method uses models trained on training-fold (in-fold) observations to predict the response for validation-fold (out-of-fold) observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (that is, roughly 4/5 of the data) and the validation fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) by using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third, fourth, and fifth models.

If you validate by using `kfoldPredict`, the software computes predictions for the observations in group *i* by using the *i*th model. In short, the software estimates a response for every observation by using the model trained without that observation.

Note `ClassificationPartitionedKernel` model objects do not store the predictor data set.

Creation

You can create a `ClassificationPartitionedKernel` model by training a classification kernel model using `fitckernel` and specifying one of these name-value pair arguments: `'Crossval'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
character vector

This property is read-only.

Cross-validated model name, specified as a character vector.

For example, `'Kernel'` specifies a cross-validated kernel model.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer scalar

This property is read-only.

Number of cross-validated folds, specified as a positive integer scalar.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

This property is read-only.

Cross-validation parameter values, specified as an object. The parameter values correspond to the name-value pair argument values used to cross-validate the kernel classifier. `ModelParameters` does not contain estimated parameters.

You can access the properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations
positive numeric scalar

This property is read-only.

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

This property is read-only.

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Kernel classifiers trained on cross-validation folds
cell array of `ClassificationKernel` models

This property is read-only.

Kernel classifiers trained on cross-validation folds, specified as a cell array of `ClassificationKernel` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

Y — Observed class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Observed class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements and has the same data type as the input argument `Y` that you pass to `fitckernel` to cross-validate the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding row of `X`.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the observed class labels property `Y` and determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs

square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`PredictorNames` — Predictor names

cell array of character vectors

This property is read-only.

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns used as predictor variables in the training data `X` or `Tbl`.

Data Types: `cell`

`Prior` — Prior class probabilities

numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as there are classes in `ClassNames`, and the order of the elements corresponds to the elements of `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function

`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For a kernel classification model `Mdl`, and before the score transformation, the predicted classification score for the observation x (row vector) is $f(x) = T(x)\beta + b$.

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

To change the `CVMdl` score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value from the table.

  ```
  CVMdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)

Value	Description
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  CVMdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Object Functions

kfoldEdge	Classification edge for cross-validated kernel classification model
kfoldLoss	Classification loss for cross-validated kernel classification model
kfoldMargin	Classification margins for cross-validated kernel classification model
kfoldPredict	Classify observations in cross-validated kernel classification model

Examples

Cross-Validate Kernel Classification Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
rng('default') % For reproducibility
```

Cross-validate a binary kernel classification model. By default, the software uses 10-fold cross-validation.

```
CVMdl = fitckernel(X,Y,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

```
numel(CVMdl.Trained)
```

```
ans = 10
```

`CVMdl` is a `ClassificationPartitionedKernel` model. Because `fitckernel` implements 10-fold cross-validation, `CVMdl` contains 10 `ClassificationKernel` models that the software trains on training-fold (in-fold) observations.

Estimate the cross-validated classification error.

```
kfoldLoss(CVMdl)
```

```
ans = 0.0940
```

The classification error rate is approximately 9%.

See Also

`ClassificationKernel` | `fitckernel`

Introduced in R2018b

ClassificationPartitionedKernelECOC

Cross-validated kernel error-correcting output codes (ECOC) model for multiclass classification

Description

`ClassificationPartitionedKernelECOC` is an error-correcting output codes (ECOC) model composed of kernel classification models, trained on cross-validated folds. Estimate the quality of the classification by cross-validation using one or more "kfold" functions: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, and `kfoldEdge`.

Every "kfold" method uses models trained on training-fold (in-fold) observations to predict the response for validation-fold (out-of-fold) observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (that is, roughly 4/5 of the data) and the validation fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) by using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third, fourth, and fifth models.

If you validate by using `kfoldPredict`, the software computes predictions for the observations in group *i* by using the *i*th model. In short, the software estimates a response for every observation by using the model trained without that observation.

Note `ClassificationPartitionedKernelECOC` model objects do not store the predictor data set.

Creation

You can create a `ClassificationPartitionedKernelECOC` model by training an ECOC model using `fitcecoc` and specifying these name-value pair arguments:

- `'Learners'`– Set the value to `'kernel'`, a template object returned by `templateKernel`, or a cell array of such template objects.
- One of the arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

For more details, see `fitcecoc`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
character vector

This property is read-only.

Cross-validated model name, specified as a character vector.

For example, `'KernelECOC'` specifies a cross-validated kernel ECOC model.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer scalar

This property is read-only.

Number of cross-validated folds, specified as a positive integer scalar.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

This property is read-only.

Cross-validation parameter values, specified as an object. The parameter values correspond to the name-value pair argument values used to cross-validate the ECOC classifier. `ModelParameters` does not contain estimated parameters.

You can access the properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations
positive numeric scalar

This property is read-only.

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

This property is read-only.

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Compact classifiers trained on cross-validation folds
cell array of `CompactClassificationECOC` models

This property is read-only.

Compact classifiers trained on cross-validation folds, specified as a cell array of `CompactClassificationECOC` models. `Trained` has *k* cells, where *k* is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

`Y` — Observed class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Observed class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements and has the same data type as the input argument `Y` that you pass to `fitcecoc` to cross-validate the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding row of the predictor data.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

ECOC Properties

`BinaryLoss` — Binary learner loss function
`'binodeviance'` | `'exponential'` | `'hamming'` | `'hinge'` | `'linear'` | `'logit'` | `'quadratic'`

This property is read-only.

Binary learner loss function, specified as a character vector representing the loss function name.

By default, if all binary learners are kernel classification models using SVM, then `BinaryLoss` is `'hinge'`. If all binary learners are kernel classification models using logistic regression, then `BinaryLoss` is `'quadratic'`. To potentially increase accuracy, specify a binary loss function other than the default during a prediction or loss computation by using the `'BinaryLoss'` name-value pair argument of `kfoldPredict` or `kfoldLoss`.

Data Types: `char`

`BinaryY` — Binary learner class labels
numeric matrix | `[]`

This property is read-only.

Binary learner class labels, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across all folds, then `BinaryY` is a `NumObservations`-by-*L* matrix, where *L* is the number of binary learners (`size(CodingMatrix,2)`).

The elements of `BinaryY` are `-1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observation `k` to a dichotomous class corresponding to the value of `BinaryY(k,j)`.

Value	Dichotomous Class Assignment
`-1`	Learner `j` assigns observation `k` to a negative class.
`0`	Before training, learner `j` removes observation `k` from the data set.
`1`	Learner `j` assigns observation `k` to a positive class.

- If the coding matrix varies across folds, then `BinaryY` is empty (`[]`).

Data Types: `double`

`CodingMatrix` — Codes specifying class assignments

numeric matrix | `[]`

This property is read-only.

Codes specifying class assignments for the binary learners, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across all folds, then `CodingMatrix` is a *K*-by-*L* matrix, where *K* is the number of classes and *L* is the number of binary learners.

  The elements of `CodingMatrix` are `-1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observations in class `i` to a dichotomous class corresponding to the value of `CodingMatrix(i,j)`.

Value	Dichotomous Class Assignment
`-1`	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

- If the coding matrix varies across folds, then `CodingMatrix` is empty (`[]`). You can obtain the coding matrix for each fold by using the `Trained` property. For example, `CVMdl.Trained{1}.CodingMatrix` is the coding matrix in the first fold of the cross-validated ECOC model `CVMdl`.

Data Types: `double` | `single` | `int8` | `int16` | `int32` | `int64`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the observed class labels property `Y` and determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`PredictorNames` — Predictor names
cell array of character vectors

This property is read-only.

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns used as predictor variables in the training data `X` or `Tbl`.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as there are classes in `ClassNames`, and the order of the elements corresponds to the elements of `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For a kernel classification model `Mdl`, and before the score transformation, the predicted classification score for the observation x (row vector) is $f(x) = T(x)\beta + b$.

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

To change the `CVMdl` score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value from the table.

  ```
  CVMdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  CVMdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Object Functions

kfoldEdge	Classification edge for cross-validated kernel ECOC model
kfoldLoss	Classification loss for cross-validated kernel ECOC model
kfoldMargin	Classification margins for cross-validated kernel ECOC model
kfoldPredict	Classify observations in cross-validated kernel ECOC model

Examples

Cross-Validate Multiclass Kernel Classification Model

Create a cross-validated, multiclass kernel ECOC classification model using `fitcecoc`.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate a multiclass kernel ECOC classification model that can identify the species of a flower based on the flower's measurements.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` cross-validated model. `fitcecoc` implements 10-fold cross-validation by default. Therefore, `CVMdl.Trained` contains a 10-by-1 cell array of ten `CompactClassificationECOC` models, one for each fold. Each compact ECOC model is composed of binary kernel classification models.

Estimate the classification error by passing `CVMdl` to `kfoldLoss`.

```
error = kfoldLoss(CVMdl)
```

```
error = 0.0333
```

The estimated classification error is about 3% misclassified observations.

To change default options when training ECOC models composed of kernel classification models, create a kernel classification model template using `templateKernel`, and then pass the template to `fitcecoc`.

See Also

`ClassificationKernel` | `CompactClassificationECOC` | `fitcecoc` | `fitckernel`

Introduced in R2018b

ClassificationPartitionedLinear

Package: `classreg.learning.partition`
Superclasses: `ClassificationPartitionedModel`

Cross-validated linear model for binary classification of high-dimensional data

Description

`ClassificationPartitionedLinear` is a set of linear classification models trained on cross-validated folds. To obtain a cross-validated, linear classification model, use `fitclinear` and specify one of the cross-validation options. You can estimate the quality of classification, or how well the linear classification model generalizes, using one or more of these "kfold" methods: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, and `kfoldEdge`.

Every "kfold" method uses models trained on in-fold observations to predict the response for out-of-fold observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five roughly equally sized groups. The training fold contains four of the groups (that is, roughly 4/5 of the data) and the test fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model, which is stored in `CVMdl.Trained{2}`, using the observations in the first group and last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third through fifth models.

If you validate by calling `kfoldPredict`, it computes predictions for the observations in group 1 using the first model, group 2 for the second model, and so on. In short, the software estimates a response for every observation using the model trained without that observation.

Note `ClassificationPartitionedLinear` model objects do not store the predictor data set.

Construction

`CVMdl = fitclinear(X,Y,Name,Value)` creates a cross-validated, linear classification model when `Name` is either `'CrossVal'`, `'CVPartition'`, `'Holdout'`, or `'KFold'`. For more details, see `fitclinear`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name

character vector

Cross-validated model name, specified as a character vector.

For example, `'Linear'` specifies a cross-validated linear model for binary classification or regression.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

Cross-validation parameter values, e.g., the name-value pair argument values used to cross-validate the linear model, specified as an object. `ModelParameters` does not contain estimated parameters.

Access properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations
positive numeric scalar

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Linear classification models trained on cross-validation folds
cell array of `ClassificationLinear` model objects

Linear classification models trained on cross-validation folds, specified as a cell array of `ClassificationLinear` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes `W` so that the weights for observations within a particular class sum up to the prior probability of that class.

Data Types: `single` | `double`

`Y` — Observed class labels
categorical array | character array | logical vector | vector of numeric values | cell array of character vectors

Observed class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements, and

is the same data type as the input argument `Y` that you passed to `fitclinear` to cross-validate the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding observation in the predictor data.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs

square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`PredictorNames` — Predictor names

cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`Prior` — Prior class probabilities

numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as classes in `ClassNames`, and the order of the elements corresponds to the elements of `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For linear classification models and before transformation, the predicted classification score for the observation x (row vector) is $f(x) = x\beta + b$, where β and b correspond to `Mdl.Beta` and `Mdl.Bias`, respectively.

To change the score transformation function to, for example, *`function`*, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function, or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Methods

kfoldEdge	Classification edge for observations not used for training
kfoldLoss	Classification loss for observations not used in training
kfoldMargin	Classification margins for observations not used in training
kfoldPredict	Predict labels for observations not used for training

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Create Cross-Validated Binary Linear Classification Model

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels. There are more than two classes in the data.

Identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Cross-validate a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinear` cross-validated model. Because `fitclinear` implements 10-fold cross-validation by default, `CVMdl.Trained` contains ten `ClassificationLinear` models that contain the results of training linear classification models for each of the folds.

Estimate labels for out-of-fold observations and estimate the generalization error by passing `CVMdl` to `kfoldPredict` and `kfoldLoss`, respectively.

```
oofLabels = kfoldPredict(CVMdl);
ge = kfoldLoss(CVMdl)
```

```
ge = 7.6017e-04
```

The estimated generalization error is less than 0.1% misclassified observations.

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, implement 5-fold cross-validation.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Estimate the coefficients using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
X = X';
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns','KFold',5,...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

```
numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `ClassificationPartitionedLinear` model. Because `fitclinear` implements 5-fold cross-validation, `CVMdl` contains 5 `ClassificationLinear` models that the software trains on each fold.

Display the first trained linear classification model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl1` is a `ClassificationLinear` model object. `fitclinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 11 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated classification error.

```
ce = kfoldLoss(CVMdl);
```

Because there are 11 regularization strengths, `ce` is a 1-by-11 vector of classification error rates.

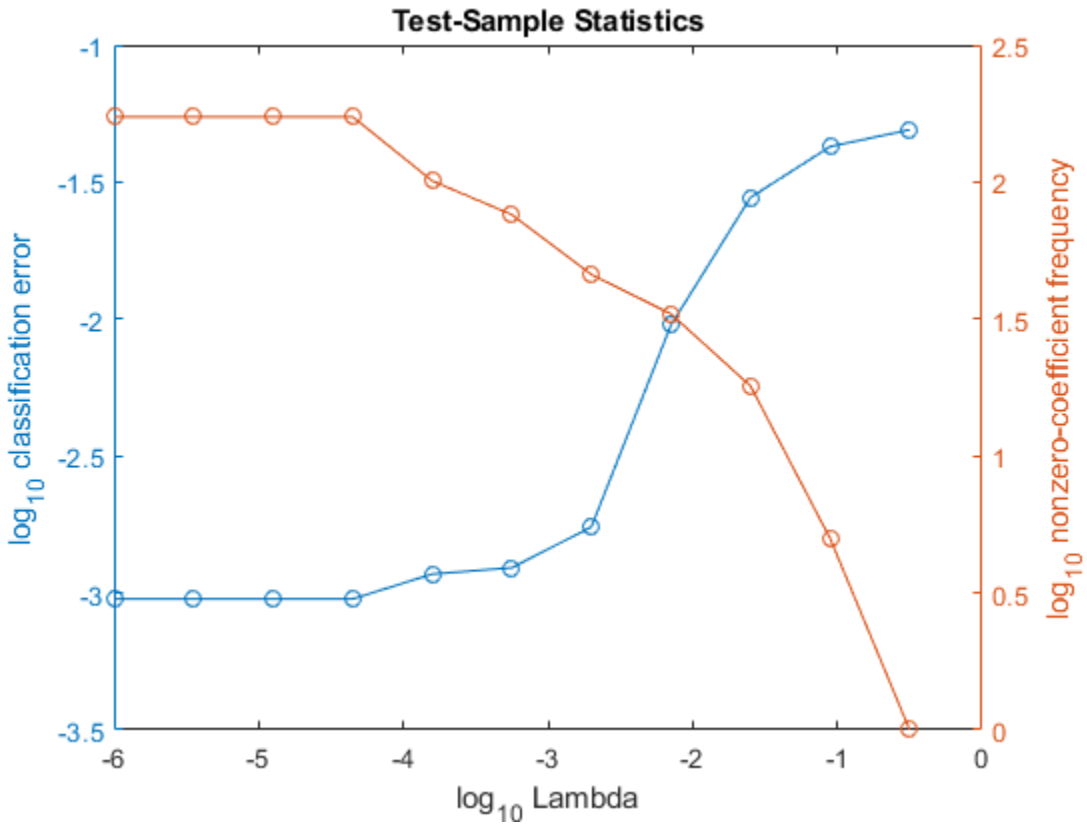
Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated, classification error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(ce),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} classification error')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
```

```
title('Test-Sample Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low classification error. In this case, a value between 10^{-4} to 10^{-1} should suffice.

```
idxFinal = 7;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

See Also

`ClassificationLinear` | `fitclinear` | `kfoldLoss` | `kfoldPredict`

Introduced in R2016a

ClassificationPartitionedLinearECOC

Package: `classreg.learning.partition`
Superclasses: `ClassificationPartitionedModel`

Cross-validated linear error-correcting output codes model for multiclass classification of high-dimensional data

Description

`ClassificationPartitionedLinearECOC` is a set of error-correcting output codes (ECOC) models composed of linear classification models, trained on cross-validated folds. Estimate the quality of classification by cross-validation using one or more "kfold" functions: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, and `kfoldEdge`.

Every "kfold" method uses models trained on in-fold observations to predict the response for out-of-fold observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five roughly equal-sized groups. The training fold contains four of the groups (that is, roughly 4/5 of the data) and the test fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows.

1. The software trains the first model (stored in `CVMdl.Trained{1}`) using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third, fourth, and fifth models.

If you validate by calling `kfoldPredict`, it computes predictions for the observations in group 1 using the first model, group 2 for the second model, and so on. In short, the software estimates a response for every observation using the model trained without that observation.

Note `ClassificationPartitionedLinearECOC` model objects do not store the predictor data set.

Construction

`CVMdl = fitcecoc(X,Y,'Learners',t,Name,Value)` returns a cross-validated, linear ECOC model when:

- `t` is `'Linear'` or a template object returned by `templateLinear`.
- `Name` is one of `'CrossVal'`, `'CVPartition'`, `'Holdout'`, or `'KFold'`.

For more details, see `fitcecoc`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name

character vector

Cross-validated model name, specified as a character vector.

For example, `'ECOC'` specifies a cross-validated ECOC model.

Data Types: `char`

`KFold` — Number of cross-validated folds

positive integer

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values

object

Cross-validation parameter values, e.g., the name-value pair argument values used to cross-validate the ECOC classifier, specified as an object. `ModelParameters` does not contain estimated parameters.

Access properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations

positive numeric scalar

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition

`cvpartition` model

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Compact classifiers trained on cross-validation folds

cell array of `CompactClassificationECOC` models

Compact classifiers trained on cross-validation folds, specified as a cell array of `CompactClassificationECOC` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights

numeric vector

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

Y — Observed class labels
categorical array | character array | logical vector | vector of numeric values | cell array of character vectors

Observed class labels used to cross-validate the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has `NumObservations` elements, and is the same data type as the input argument `Y` that you passed to `fitcecoc` to cross-validate the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the observation in the predictor data.

Data Types: `char` | `cell` | `categorical` | `logical` | `single` | `double`

ECOC Properties

`BinaryLoss` — Binary learner loss function
`'binodeviance'` | `'exponential'` | `'hamming'` | `'hinge'` | `'linear'` | `'logit'` | `'quadratic'`

Binary learner loss function, specified as a character vector representing the loss function name.

If you train using binary learners that use different loss functions, then the software sets `BinaryLoss` to `'hamming'`. To potentially increase accuracy, specify a binary loss function other than the default during a prediction or loss computation by using the `'BinaryLoss'` name-value pair argument of `predict` or `loss`.

Data Types: `char`

`BinaryY` — Binary learner class labels
numeric matrix | `[]`

Binary learner class labels, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across folds, then `BinaryY` is a `NumObservations`-by-*L* matrix, where *L* is the number of binary learners (`size(CodingMatrix,2)`).

  Elements of `BinaryY` are `-1`, `0`, or `1`, and the value corresponds to a dichotomous class assignment. This table describes how learner `j` assigns observation `k` to a dichotomous class corresponding to the value of `BinaryY(k,j)`.

Value	Dichotomous Class Assignment
–1	Learner `j` assigns observation `k` to a negative class.
`0`	Before training, learner `j` removes observation `k` from the data set.
`1`	Learner `j` assigns observation `k` to a positive class.

- If the coding matrix varies across folds, then `BinaryY` is empty (`[]`).

Data Types: `double`

`CodingMatrix` — Codes specifying class assignments
numeric matrix | `[]`

Codes specifying class assignments for the binary learners, specified as a numeric matrix or `[]`.

- If the coding matrix is the same across folds, then `CodingMatrix` is a *K*-by-*L* matrix. *K* is the number of classes and *L* is the number of binary learners.

  Elements of `CodingMatrix` are `-1`, `0`, or `1`, and the value corresponds to a dichotomous class assignment. This table describes how learner `j` assigns observations in class `i` to a dichotomous class corresponding to the value of `CodingMatrix(i,j)`.

Value	Dichotomous Class Assignment
–1	Learner `j` assigns observations in class `i` to a negative class.
0	Before training, learner `j` removes observations in class `i` from the data set.
1	Learner `j` assigns observations in class `i` to a positive class.

- If the coding matrix varies across folds, then `CodingMatrix` is empty (`[]`). Obtain the coding matrix for each fold using the `Trained` property. For example, `CVMdl.Trained{1}.CodingMatrix` is the coding matrix in the first fold of the cross-validated ECOC model `CVMdl`.

Data Types: `double` | `single` | `int8` | `int16` | `int32` | `int64`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs

square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has *K* rows and columns, where *K* is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as the number of classes in `ClassNames`, and the order of the elements corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For linear classification models and before transformation, the predicted classification score for the observation x (row vector) is $f(x) = x\beta + b$, where β and b correspond to `Mdl.Beta` and `Mdl.Bias`, respectively.

To change the score transformation function to, for example, *`function`*, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$

Value	Description
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function, or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Methods

kfoldEdge	Classification edge for observations not used for training
kfoldLoss	Classification loss for observations not used in training
kfoldMargin	Classification margins for observations not used in training
kfoldPredict	Predict labels for observations not used for training

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Create Cross-Validated Multiclass Linear Classification Model

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels.

Cross-validate a multiclass, linear classification model that can identify which MATLAB® toolbox a documentation web page is from based on counts of words on the page.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners','linear','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedLinearECOC
    CrossValidatedModel: 'LinearECOC'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: [1x13 categorical]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` cross-validated model. Because `fitcecoc` implements 10-fold cross-validation by default, `CVMdl.Trained` contains a 10-by-1 cell vector of ten `CompactClassificationECOC` models that contain the results of training ECOC models composed of binary, linear classification models for each of the folds.

Estimate labels for out-of-fold observations and estimate the generalization error by passing `CVMdl` to `kfoldPredict` and `kfoldLoss`, respectively.

```
oofLabels = kfoldPredict(CVMdl);
ge = kfoldLoss(CVMdl)
```

```
ge = 0.0958
```

The estimated generalization error is about 10% misclassified observations.

To improve generalization error, try specifying another solver, such as LBFGS. To change default options when training ECOC models composed of linear classification models, create a linear classification model template using `templateLinear`, and then pass the template to `fitcecoc`.

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for an ECOC model composed of linear classification models that use logistic regression learners, implement 5-fold cross-validation.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

For simplicity, use the label 'others' for all observations in `Y` that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-7} through 10^{-2}.

```
Lambda = logspace(-7,-2,11);
```

Create a linear classification model template that specifies to use logistic regression learners, use lasso penalties with strengths in `Lambda`, train using SpaRSA, and lower the tolerance on the gradient of the objective function to `1e-8`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns.

```
X = X';
rng(10); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns','KFold',5);
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model.

Dissect `CVMdl`, and each model within it.

```
numECOCModels = numel(CVMdl.Trained)
```

```
numECOCModels = 5
```

```
ECOCMdl1 = CVMdl.Trained{1}
```

```
ECOCMdl1 =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [comm    dsp    simulink    others]
    ScoreTransform: 'none'
    BinaryLearners: {6×1 cell}
      CodingMatrix: [4×6 double]


  Properties, Methods
```

```
numCLModels = numel(ECOCMdl1.BinaryLearners)
```

```
numCLModels = 6
```

```
CLMdl1 = ECOCMdl1.BinaryLearners{1}
```

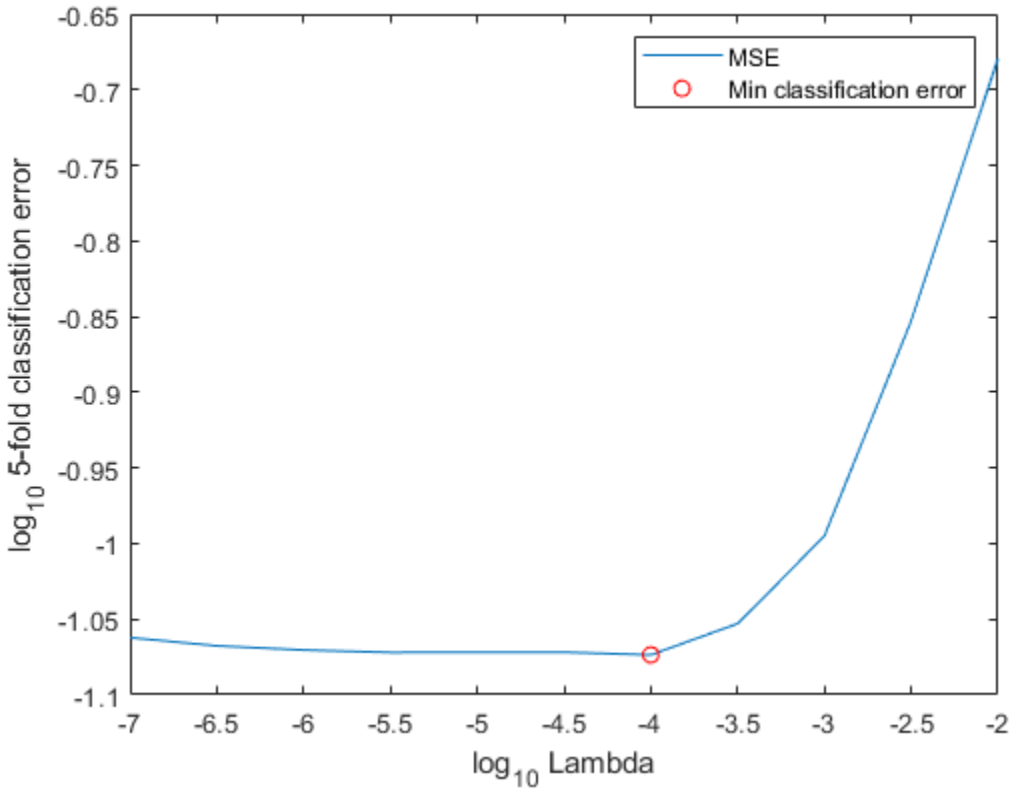
```
CLMdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [-1 1]
    ScoreTransform: 'logit'
              Beta: [34023×11 double]
              Bias: [-0.3169 -0.3169 -0.3168 -0.3168 -0.3168 -0.3167 -0.1725 -0.0805 -0.1762 -0.1
            Lambda: [1.0000e-07 3.1623e-07 1.0000e-06 3.1623e-06 1.0000e-05 3.1623e-05 1.0000e-04
           Learner: 'logistic'


  Properties, Methods
```

Because `fitcecoc` implements 5-fold cross-validation, `CVMdl` contains a 5-by-1 cell array of `CompactClassificationECOC` models that the software trains on each fold. The `BinaryLearners` property of each `CompactClassificationECOC` model contains the `ClassificationLinear`

models. The number of `ClassificationLinear` models within each compact ECOC model depends on the number of distinct labels and coding design. Because `Lambda` is a sequence of regularization strengths, you can think of `CLMdl1` as 11 models, one for each regularization strength in `Lambda`.

Determine how well the models generalize by plotting the averages of the 5-fold classification error for each regularization strength. Identify the regularization strength that minimizes the generalization error over the grid.

```
ce = kfoldLoss(CVMdl);
figure;
plot(log10(Lambda),log10(ce))
[~,minCEIdx] = min(ce);
minLambda = Lambda(minCEIdx);
hold on
plot(log10(minLambda),log10(ce(minCEIdx)),'ro');
ylabel('log_{10} 5-fold classification error')
xlabel('log_{10} Lambda')
legend('MSE','Min classification error')
hold off
```



Train an ECOC model composed of linear classification model using the entire data set, and specify the minimal regularization strength.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',minLambda,'GradientTolerance',1e-8);
MdlFinal = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns');
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

See Also

`ClassificationECOC` | `ClassificationLinear` | `fitcecoc` | `fitclinear` | `kfoldLoss` | `kfoldPredict`

Introduced in R2016a

ClassificationPartitionedModel

Package: `classreg.learning.partition`

Cross-validated classification model

Description

`ClassificationPartitionedModel` is a set of classification models trained on cross-validated folds. Estimate the quality of classification by cross validation using one or more "kfold" methods: `kfoldPredict`, `kfoldLoss`, `kfoldMargin`, `kfoldEdge`, and `kfoldfun`.

Every "kfold" method uses models trained on in-fold observations to predict the response for out-of-fold observations. For example, suppose you cross validate using five folds. In this case, the software randomly assigns each observation into five roughly equally sized groups. The training fold contains four of the groups (i.e., roughly 4/5 of the data) and the test fold contains the other group (i.e., roughly 1/5 of the data). In this case, cross validation proceeds as follows:

- The software trains the first model (stored in `CVMdl.Trained{1}`) using the observations in the last four groups and reserves the observations in the first group for validation.
- The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and last three groups, and reserves the observations in the second group for validation.
- The software proceeds in a similar fashion for the third to fifth models.

If you validate by calling `kfoldPredict`, it computes predictions for the observations in group 1 using the first model, group 2 for the second model, and so on. In short, the software estimates a response for every observation using the model trained without that observation.

Construction

`CVMdl = crossval(Mdl)` creates a cross-validated classification model from a classification model (`Mdl`).

Alternatively:

- `CVDiscrMdl = fitcdiscr(X,Y,Name,Value)`
- `CVKNNMdl = fitcknn(X,Y,Name,Value)`
- `CVNetMdl = fitcnet(X,Y,Name,Value)`
- `CVNBMdl = fitcnb(X,Y,Name,Value)`
- `CVSVMMdl = fitcsvm(X,Y,Name,Value)`
- `CVTreeMdl = fitctree(X,Y,Name,Value)`

create a cross-validated model when `Name` is either `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`. For syntax details, see `fitcdiscr`, `fitcknn`, `fitcnet`, `fitcnb`, `fitcsvm`, and `fitctree`.

Input Arguments

Mdl

A classification model, specified as one of the following:

- A classification tree trained using `fitctree`
- A discriminant analysis classifier trained using `fitcdiscr`
- A neural network classifier trained using `fitcnet`
- A naive Bayes classifier trained using `fitcnb`
- A nearest-neighbor classifier trained using `fitcknn`
- A support vector machine classifier trained using `fitcsvm`

Properties

BinEdges

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

CategoricalPredictors

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

If `Mdl` is a trained discriminant analysis classifier, then `CategoricalPredictors` is always empty (`[]`).

ClassNames

Unique class labels used in training the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors.

Cost

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response.

If `CVModel` is a cross-validated `ClassificationDiscriminant`, `ClassificationKNN`, or `ClassificationNaiveBayes` model, then you can change its cost matrix to e.g., `CostMatrix`, using dot notation.

```
CVModel.Cost = CostMatrix;
```

CrossValidatedModel

Name of the cross-validated model, which is a character vector.

KFold

Number of folds used in cross-validated model, which is a positive integer.

ModelParameters

Object holding parameters of `CVModel`.

NumObservations

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

Partition

The partition of class `CVPartition` used in creating the cross-validated model.

PredictorNames

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Prior

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`.

If `CVModel` is a cross-validated `ClassificationDiscriminant` or `ClassificationNaiveBayes` model, then you can change its vector of priors to e.g., `priorVector`, using dot notation.

```
CVModel.Prior = priorVector;
```

`ResponseName`

Response variable name, specified as a character vector.

`ScoreTransform`

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

`Trained`

The trained learners, which is a cell array of compact classification models.

`W`

The scaled `weights`, which is a vector with length `n`, the number of observations in `X`.

X

A matrix or table of predictor values.

Y

Categorical or character array, logical or numeric vector, or cell array of character vectors specifying the class labels for each observation. Each entry of Y is the response value of the corresponding observation in X.

Object Functions

kfoldEdge	Classification edge for cross-validated classification model
kfoldLoss	Classification loss for cross-validated classification model
kfoldMargin	Classification margins for cross-validated classification model
kfoldPredict	Classify observations in cross-validated classification model
kfoldfun	Cross-validate function for classification

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Evaluate the Classification Error of a Classification Tree Classifier

Evaluate the *k*-fold cross-validation error for a classification tree model.

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification tree using default options.

```
Mdl = fitctree(meas,species);
```

Cross validate the classification tree model.

```
CVMdl = crossval(Mdl);
```

Estimate the 10-fold cross-validation loss.

```
L = kfoldLoss(CVMdl)
```

```
L = 0.0533
```

Estimate Posterior Probabilities for Test Samples

Estimate positive class posterior probabilities for the test set of an SVM algorithm.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Specify a 20% holdout sample. It is good practice to standardize the predictors and specify the class order.

```
rng(1) % For reproducibility
CVSVMModel = fitcsvm(X,Y,'Holdout',0.2,'Standardize',true,...
    'ClassNames',{'b','g'});
```

`CVSVMModel` is a trained `ClassificationPartitionedModel` cross-validated classifier.

Estimate the optimal score function for mapping observation scores to posterior probabilities of an observation being classified as `'g'`.

```
ScoreCVSVMModel = fitSVMPosterior(CVSVMModel);
```

`ScoreSVMModel` is a trained `ClassificationPartitionedModel` cross-validated classifier containing the optimal score transformation function estimated from the training data.

Estimate the out-of-sample positive class posterior probabilities. Display the results for the first 10 out-of-sample observations.

```
[~,OOSPostProbs] = kfoldPredict(ScoreCVSVMModel);
indx = ~isnan(OOSPostProbs(:,2));
hoObs = find(indx); % Holdout observation numbers
OOSPostProbs = [hoObs, OOSPostProbs(indx,2)];
table(OOSPostProbs(1:10,1),OOSPostProbs(1:10,2),...
    'VariableNames',{'ObservationIndex','PosteriorProbability'})
```

```
ans=10×2 table
    ObservationIndex    PosteriorProbability
    ________________    ____________________

           6                    0.17381
           7                    0.89639
           8                  0.0076613
           9                    0.91602
          16                   0.026722
          22                  4.6114e-06
          23                      0.9024
          24                  2.4137e-06
          38                  0.00042705
          41                     0.86427
```

Tips

To estimate posterior probabilities of trained, cross-validated SVM classifiers, use `fitSVMPosterior`.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions support model objects fitted with GPU array input arguments:
  - `kfoldEdge`
  - `kfoldLoss`
  - `kfoldMargin`
  - `kfoldPredict`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationKNN` | `ClassificationNaiveBayes` | `ClassificationNeuralNetwork` | `CompactClassificationDiscriminant` | `CompactClassificationSVM` | `CompactClassificationTree` | `fitSVMPosterior` | `fitcdiscr` | `fitcknn` | `fitcnb` | `fitcnet` | `fitcsvm` | `fitctree`

Topics

“Cross Validating a Discriminant Analysis Classifier” on page 20-17

ClassificationSVM

Support vector machine (SVM) for one-class and binary classification

Description

`ClassificationSVM` is a support vector machine (SVM) classifier on page 33-540 for one-class and two-class learning. Trained `ClassificationSVM` classifiers store training data, parameter values, prior probabilities, support vectors, and algorithmic implementation information. Use these classifiers to perform tasks such as fitting a score-to-posterior-probability transformation function (see `fitPosterior`) and predicting labels for new data (see `predict`).

Creation

Create a `ClassificationSVM` object by using `fitcsvm`.

Properties

SVM Properties

`Alpha` — Trained classifier coefficients

numeric vector

This property is read-only.

Trained classifier coefficients, specified as an *s*-by-1 numeric vector. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`.

`Alpha` contains the trained classifier coefficients from the dual problem, that is, the estimated Lagrange multipliers. If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `Alpha` contains one coefficient corresponding to the entire set. That is, MATLAB attributes a nonzero coefficient to one observation from the set of duplicates and a coefficient of `0` to all other duplicate observations in the set.

Data Types: `single` | `double`

`Beta` — Linear predictor coefficients

numeric vector

This property is read-only.

Linear predictor coefficients, specified as a numeric vector. The length of `Beta` is equal to the number of predictors used to train the model.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. `Beta` stores one value for each predictor variable, including the dummy variables. For example, if there are three predictors, one of which is a categorical variable with three levels, then `Beta` is a numeric vector containing five values.

If `KernelParameters.Function` is `'linear'`, then the classification score for the observation x is

$$f(x) = (x/s)'\beta + b\,.$$

`Mdl` stores β, b, and s in the properties `Beta`, `Bias`, and `KernelParameters.Scale`, respectively.

To estimate classification scores manually, you must first apply any transformations to the predictor data that were applied during training. Specifically, if you specify `'Standardize',true` when using `fitcsvm`, then you must standardize the predictor data manually by using the mean `Mdl.Mu` and standard deviation `Mdl.Sigma`, and then divide the result by the kernel scale in `Mdl.KernelParameters.Scale`.

All SVM functions, such as `resubPredict` and `predict`, apply any required transformation before estimation.

If `KernelParameters.Function` is not `'linear'`, then `Beta` is empty (`[]`).

Data Types: `single` | `double`

`Bias` — Bias term

scalar

This property is read-only.

Bias term, specified as a scalar.

Data Types: `single` | `double`

`BoxConstraints` — Box constraints

numeric vector

This property is read-only.

Box constraints, specified as a numeric vector of n-by-1 box constraints on page 33-539. n is the number of observations in the training data (see the `NumObservations` property).

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations, MATLAB sums the box constraints and then attributes the sum to one observation. MATLAB attributes the box constraints of `0` to all other observations in the set.

Data Types: `single` | `double`

`CacheInfo` — Caching information

structure array

This property is read-only.

Caching information, specified as a structure array. The caching information contains the fields described in this table.

Field	Description
Size	The cache size (in MB) that the software reserves to train the SVM classifier. For details, see `'CacheSize'`.

Field	Description
Algorithm	The caching algorithm that the software uses during optimization. Currently, the only available caching algorithm is `Queue`. You cannot set the caching algorithm.

Display the fields of `CacheInfo` by using dot notation. For example, `Mdl.CacheInfo.Size` displays the value of the cache size.

Data Types: `struct`

`IsSupportVector` — Support vector indicator
logical vector

This property is read-only.

Support vector indicator, specified as an *n*-by-1 logical vector that flags whether a corresponding observation in the predictor data matrix is a “Support Vector” on page 33-540. *n* is the number of observations in the training data (see `NumObservations`).

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `IsSupportVector` flags only one observation as a support vector.

Data Types: `logical`

`KernelParameters` — Kernel parameters
structure array

This property is read-only.

Kernel parameters, specified as a structure array. The kernel parameters property contains the fields listed in this table.

Field	Description
Function	Kernel function used to compute the elements of the Gram matrix on page 33-1862. For details, see `'KernelFunction'`.
Scale	Kernel scale parameter used to scale all elements of the predictor data on which the model is trained. For details, see `'KernelScale'`.

To display the values of `KernelParameters`, use dot notation. For example, `Mdl.KernelParameters.Scale` displays the kernel scale parameter value.

The software accepts `KernelParameters` as inputs and does not modify them.

Data Types: `struct`

`Nu` — One-class learning parameter
positive scalar

This property is read-only.

One-class learning on page 33-539 parameter ν, specified as a positive scalar.

Data Types: `single` | `double`

`OutlierFraction` — Proportion of outliers
numeric scalar

This property is read-only.

Proportion of outliers in the training data, specified as a numeric scalar.

Data Types: `double`

`Solver` — Optimization routine
`'ISDA'` | `'L1QP'` | `'SMO'`

This property is read-only.

Optimization routine used to train the SVM classifier, specified as `'ISDA'`, `'L1QP'`, or `'SMO'`. For more details, see `'Solver'`.

`SupportVectorLabels` — Support vector class labels
s-by-1 numeric vector

This property is read-only.

Support vector class labels, specified as an *s*-by-1 numeric vector. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`.

A value of `+1` in `SupportVectorLabels` indicates that the corresponding support vector is in the positive class (`ClassNames{2}`). A value of `–1` indicates that the corresponding support vector is in the negative class (`ClassNames{1}`).

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `SupportVectorLabels` contains one unique support vector label.

Data Types: `single` | `double`

`SupportVectors` — Support vectors
s-by-*p* numeric matrix

This property is read-only.

Support vectors in the trained classifier, specified as an *s*-by-*p* numeric matrix. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`, and *p* is the number of predictor variables in the predictor data.

`SupportVectors` contains rows of the predictor data `X` that MATLAB considers to be support vectors. If you specify `'Standardize',true` when training the SVM classifier using `fitcsvm`, then `SupportVectors` contains the standardized rows of `X`.

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `SupportVectors` contains one unique support vector.

Data Types: `single` | `double`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Cost` — Misclassification cost

numeric square matrix

This property is read-only.

Misclassification cost, specified as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`.

During training, the software updates the prior probabilities by incorporating the penalties described in the cost matrix.

- For two-class learning, `Cost` always has this form: `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.
- For one-class learning, `Cost = 0`.

For more details, see Algorithms on page 33-541.

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names

cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

If the model uses dummy variable encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: cell

Gradient — Training data gradient values
numeric vector

This property is read-only.

Training data gradient values, specified as a numeric vector. The length of Gradient is equal to the number of observations (NumObservations).

Data Types: single | double

ModelParameters — Parameters used to train model
object

This property is read-only.

Parameters used to train the ClassificationSVM model, specified as an object. ModelParameters contains parameter values such as the name-value pair argument values used to train the SVM classifier. ModelParameters does not contain estimated parameters.

Access the properties of ModelParameters by using dot notation. For example, access the initial values for estimating Alpha by using Mdl.ModelParameters.Alpha.

Mu — Predictor means
numeric vector | []

This property is read-only.

Predictor means, specified as a numeric vector. If you specify 'Standardize',1 or 'Standardize',true when you train an SVM classifier using fitcsvm, then the length of Mu is equal to the number of predictors.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. Mu stores one value for each predictor variable, including the dummy variables. However, MATLAB does not standardize the columns that contain categorical variables.

If you set 'Standardize',false when you train the SVM classifier using fitcsvm, then Mu is an empty vector ([]).

Data Types: single | double

NumObservations — Number of observations
numeric scalar

This property is read-only.

Number of observations in the training data stored in X and Y, specified as a numeric scalar.

Data Types: double

PredictorNames — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`Prior` — Prior probabilities

numeric vector

This property is read-only.

Prior probabilities for each class, specified as a numeric vector. The order of the elements of `Prior` corresponds to the elements of `Mdl.ClassNames`.

For two-class learning, if you specify a cost matrix, then the software updates the prior probabilities by incorporating the penalties described in the cost matrix.

For more details, see Algorithms on page 33-541.

Data Types: `single` | `double`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`RowsUsed` — Rows used in fitting

`[]` | logical vector

This property is read-only.

Rows of the original training data used in fitting the `ClassificationSVM` model, specified as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

`ScoreTransform` — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$

Value	Description
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  `Mdl.ScoreTransform = @`*`function`*`;`

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

`Sigma` — Predictor standard deviations
`[]` (default) | numeric vector

This property is read-only.

Predictor standard deviations, specified as a numeric vector.

If you specify `'Standardize',true` when you train the SVM classifier using `fitcsvm`, then the length of `Sigma` is equal to the number of predictor variables.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. `Sigma` stores one value for each predictor variable, including the dummy variables. However, MATLAB does not standardize the columns that contain categorical variables.

If you set `'Standardize',false` when you train the SVM classifier using `fitcsvm`, then `Sigma` is an empty vector (`[]`).

Data Types: `single` | `double`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to train the SVM classifier, specified as an *n*-by-1 numeric vector. *n* is the number of observations (see `NumObservations`).

`fitcsvm` normalizes the observation weights specified in the `'Weights'` name-value pair argument so that the elements of `W` within a particular class sum up to the prior probability of that class.

Data Types: `single` | `double`

X — Unstandardized predictors

numeric matrix | table

This property is read-only.

Unstandardized predictors used to train the SVM classifier, specified as a numeric matrix or table.

Each row of X corresponds to one observation, and each column corresponds to one variable.

MATLAB excludes observations containing at least one missing value, and removes corresponding elements from Y.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Class labels used to train the SVM classifier, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. Y is the same data type as the input argument Y of `fitcsvm`. (The software treats string arrays as cell arrays of character vectors.)

Each row of Y represents the observed classification of the corresponding row of X.

MATLAB excludes elements containing missing values, and removes corresponding observations from X.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

Convergence Control Properties

`ConvergenceInfo` — Convergence information

structure array

This property is read-only.

Convergence information, specified as a structure array.

Field	Description
`Converged`	Logical flag indicating whether the algorithm converged (`1` indicates convergence).
`ReasonForConvergence`	Character vector indicating the criterion the software uses to detect convergence.
`Gap`	Scalar feasibility gap between the dual and primal objective functions.
`GapTolerance`	Scalar feasibility gap tolerance. Set this tolerance, for example to `1e-2`, by using the name-value pair argument `'GapTolerance',1e-2` of `fitcsvm`.

Field	Description
`DeltaGradient`	Scalar-attained gradient difference between upper and lower violators
`DeltaGradientTolerance`	Scalar tolerance for the gradient difference between upper and lower violators. Set this tolerance, for example to `1e-2`, by using the name-value pair argument `'DeltaGradientTolerance',1e-2` of `fitcsvm`.
`LargestKKTViolation`	Maximal scalar Karush-Kuhn-Tucker (KKT) violation value.
`KKTTolerance`	Scalar tolerance for the largest KKT violation. Set this tolerance, for example, to `1e-3`, by using the name-value pair argument `'KKTTolerance',1e-3` of `fitcsvm`.
`History`	Structure array containing convergence information at set optimization iterations. The fields are: • `NumIterations`: numeric vector of iteration indices for which the software records convergence information • `Gap`: numeric vector of `Gap` values at the iterations • `DeltaGradient`: numeric vector of `DeltaGradient` values at the iterations • `LargestKKTViolation`: numeric vector of `LargestKKTViolation` values at the iterations • `NumSupportVectors`: numeric vector indicating the number of support vectors at the iterations • `Objective`: numeric vector of `Objective` values at the iterations
`Objective`	Scalar value of the dual objective function.

Data Types: `struct`

`NumIterations` — Number of iterations
positive integer

This property is read-only.

Number of iterations required by the optimization routine to attain convergence, specified as a positive integer.

To set the limit on the number of iterations to `1000`, for example, specify `'IterationLimit',1000` when you train the SVM classifier using `fitcsvm`.

Data Types: `double`

`ShrinkagePeriod` — Number of iterations between reductions of active set

nonnegative integer

This property is read-only.

Number of iterations between reductions of the active set, specified as a nonnegative integer.

To set the shrinkage period to `1000`, for example, specify `'ShrinkagePeriod',1000` when you train the SVM classifier using `fitcsvm`.

Data Types: `single` | `double`

Hyperparameter Optimization Properties

`HyperparameterOptimizationResults` — Description of cross-validation optimization of hyperparameters

`BayesianOptimization` object | table

This property is read-only.

Description of the cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty when the `'OptimizeHyperparameters'` name-value pair argument of `fitcsvm` is nonempty at creation. The value of `HyperparameterOptimizationResults` depends on the setting of the `Optimizer` field in the `HyperparameterOptimizationOptions` structure of `fitcsvm` at creation, as described in this table.

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Object Functions

compact	Reduce size of machine learning model
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validate machine learning model
discardSupportVectors	Discard support vectors for linear support vector machine (SVM) classifier
edge	Find classification edge for support vector machine (SVM) classifier
fitPosterior	Fit posterior probabilities for support vector machine (SVM) classifier
incrementalLearner	Convert binary classification support vector machine (SVM) model to incremental learner
loss	Find classification error for support vector machine (SVM) classifier
margin	Find classification margins for support vector machine (SVM) classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using support vector machine (SVM) classifier
resubEdge	Resubstitution classification edge
lime	Local interpretable model-agnostic explanations (LIME)

resubLoss	Resubstitution classification loss
resubMargin	Resubstitution classification margin
resubPredict	Classify training data using trained classifier
resume	Resume training support vector machine (SVM) classifier
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation

Examples

Train SVM Classifier

Load Fisher's iris data set. Remove the sepal lengths and widths and all observed setosa irises.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,3:4);
y = species(inds);
```

Train an SVM classifier using the processed data set.

```
SVMModel = fitcsvm(X,y)
```

```
SVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
              ClassNames: {'versicolor'  'virginica'}
          ScoreTransform: 'none'
         NumObservations: 100
                   Alpha: [24x1 double]
                    Bias: -14.4149
        KernelParameters: [1x1 struct]
          BoxConstraints: [100x1 double]
         ConvergenceInfo: [1x1 struct]
         IsSupportVector: [100x1 logical]
                  Solver: 'SMO'


  Properties, Methods
```

`SVMModel` is a trained `ClassificationSVM` classifier. Display the properties of `SVMModel`. For example, to determine the class order, use dot notation.

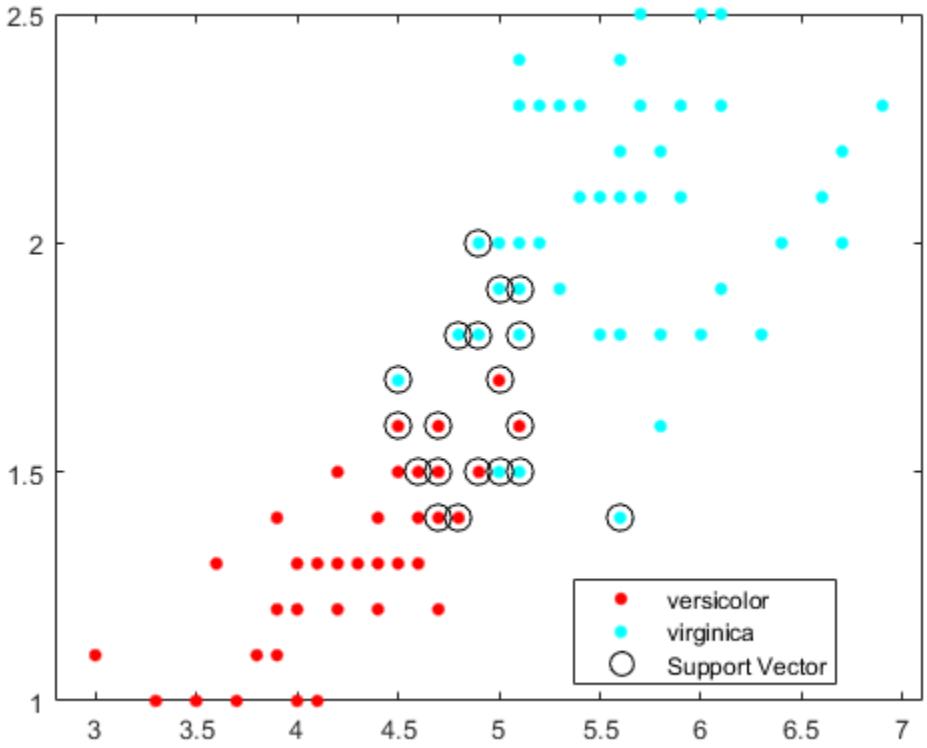
```
classOrder = SVMModel.ClassNames
```

```
classOrder = 2x1 cell
    {'versicolor'}
    {'virginica' }
```

The first class (`'versicolor'`) is the negative class, and the second (`'virginica'`) is the positive class. You can change the class order during training by using the `'ClassNames'` name-value pair argument.

Plot a scatter diagram of the data and circle the support vectors.

```
sv = SVMModel.SupportVectors;
figure
gscatter(X(:,1),X(:,2),y)
hold on
plot(sv(:,1),sv(:,2),'ko','MarkerSize',10)
legend('versicolor','virginica','Support Vector')
hold off
```



The support vectors are observations that occur on or beyond their estimated class boundaries.

You can adjust the boundaries (and, therefore, the number of support vectors) by setting a box constraint during training using the `'BoxConstraint'` name-value pair argument.

Train and Cross-Validate SVM Classifier

Load the `ionosphere` data set.

```
load ionosphere
```

Train and cross-validate an SVM classifier. Standardize the predictor data and specify the order of the classes.

```
rng(1);  % For reproducibility
CVSVMModel = fitcsvm(X,Y,'Standardize',true,...
    'ClassNames',{'b','g'},'CrossVal','on')

CVSVMModel =
  ClassificationPartitionedModel
    CrossValidatedModel: 'SVM'
         PredictorNames: {1x34 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVSVMModel` is a `ClassificationPartitionedModel` cross-validated SVM classifier. By default, the software implements 10-fold cross-validation.

Alternatively, you can cross-validate a trained `ClassificationSVM` classifier by passing it to `crossval`.

Inspect one of the trained folds using dot notation.

```
CVSVMModel.Trained{1}
```

```
ans =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
                    Alpha: [78x1 double]
                     Bias: -0.2209
         KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           SupportVectors: [78x34 double]
      SupportVectorLabels: [78x1 double]


  Properties, Methods
```

Each fold is a `CompactClassificationSVM` classifier trained on 90% of the data.

Estimate the generalization error.

```
genError = kfoldLoss(CVSVMModel)
```

```
genError = 0.1168
```

On average, the generalization error is approximately 12%.

More About

Box Constraint

A box constraint is a parameter that controls the maximum penalty imposed on margin-violating observations, which helps to prevent overfitting (regularization).

If you increase the box constraint, then the SVM classifier assigns fewer support vectors. However, increasing the box constraint can lead to longer training times.

Gram Matrix

The Gram matrix of a set of n vectors $\{x_1,..,x_n; x_j \in R^p\}$ is an n-by-n matrix with element (j,k) defined as $G(x_j,x_k) = <\phi(x_j),\phi(x_k)>$, an inner product of the transformed predictors using the kernel function ϕ.

For nonlinear SVM, the algorithm forms a Gram matrix using the rows of the predictor data X. The dual formalization replaces the inner product of the observations in X with corresponding elements of the resulting Gram matrix (called the “kernel trick”). Consequently, nonlinear SVM operates in the transformed predictor space to find a separating hyperplane.

Karush-Kuhn-Tucker Complementarity Conditions

KKT complementarity conditions are optimization constraints required for optimal nonlinear programming solutions.

In SVM, the KKT complementarity conditions are

$$\begin{cases} \alpha_j\left[y_j f\left(x_j\right) - 1 + \xi_j\right] = 0 \\ \xi_j\left(C - \alpha_j\right) = 0 \end{cases}$$

for all $j = 1,...,n$, where $f\left(x_j\right) = \phi\left(x_j\right)'\beta + b$, ϕ is a kernel function (see Gram matrix on page 33-1862), and ξ_j is a slack variable. If the classes are perfectly separable, then $\xi_j = 0$ for all $j = 1,...,n$.

One-Class Learning

One-class learning, or unsupervised SVM, aims to separate data from the origin in the high-dimensional predictor space (not the original predictor space), and is an algorithm used for outlier detection.

The algorithm resembles that of SVM for binary classification on page 33-1863. The objective is to minimize the dual expression

$$0.5\sum_{jk}\alpha_j\alpha_k G(x_j, x_k)$$

with respect to $\alpha_1, ..., \alpha_n$, subject to

$$\sum \alpha_j = n\nu$$

and $0 \le \alpha_j \le 1$ for all $j = 1,...,n$. The value of $G(x_j,x_k)$ is in element (j,k) of the Gram matrix on page 33-1862.

A small value of ν leads to fewer support vectors and, therefore, a smooth, crude decision boundary. A large value of ν leads to more support vectors and, therefore, a curvy, flexible decision boundary. The

optimal value of ν should be large enough to capture the data complexity and small enough to avoid overtraining. Also, $0 < \nu \le 1$.

For more details, see [5].

Support Vector

Support vectors are observations corresponding to strictly positive estimates of $\alpha_1,...,\alpha_n$.

SVM classifiers that yield fewer support vectors for a given training set are preferred.

Support Vector Machines for Binary Classification

The SVM binary classification algorithm searches for an optimal hyperplane that separates the data into two classes. For separable classes, the optimal hyperplane maximizes a margin (space that does not contain any observations) surrounding itself, which creates boundaries for the positive and negative classes. For inseparable classes, the objective is the same, but the algorithm imposes a penalty on the length of the margin for every observation that is on the wrong side of its class boundary.

The linear SVM score function is

$$f(x) = x'\beta + b,$$

where:

- x is an observation (corresponding to a row of `X`).
- The vector β contains the coefficients that define an orthogonal vector to the hyperplane (corresponding to `Mdl.Beta`). For separable data, the optimal margin length is $2/\|\beta\|$.
- b is the bias term (corresponding to `Mdl.Bias`).

The root of $f(x)$ for particular coefficients defines a hyperplane. For a particular hyperplane, $f(z)$ is the distance from point z to the hyperplane.

The algorithm searches for the maximum margin length, while keeping observations in the positive ($y = 1$) and negative ($y = –1$) classes separate.

- For separable classes, the objective is to minimize $\|\beta\|$ with respect to the β and b subject to $y_j f(x_j) \ge 1$, for all $j = 1,..,n$. This is the primal formalization for separable classes.
- For inseparable classes, the algorithm uses slack variables (ξ_j) to penalize the objective function for observations that cross the margin boundary for their class. $\xi_j = 0$ for observations that do not cross the margin boundary for their class, otherwise $\xi_j \ge 0$.

  The objective is to minimize $0.5\|\beta\|^2 + C\sum \xi_j$ with respect to the β, b, and ξ_j subject to $y_j f(x_j) \ge 1 - \xi_j$ and $\xi_j \ge 0$ for all $j = 1,..,n$, and for a positive scalar box constraint on page 33-1862 C. This is the primal formalization for inseparable classes.

The algorithm uses the Lagrange multipliers method to optimize the objective, which introduces n coefficients $\alpha_1,...,\alpha_n$ (corresponding to `Mdl.Alpha`). The dual formalizations for linear SVM are as follows:

- For separable classes, minimize

$$0.5\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\alpha_k y_j y_k x_j' x_k - \sum_{j=1}^{n}\alpha_j$$

with respect to $\alpha_1,...,\alpha_n$, subject to $\sum\alpha_j y_j = 0$, $\alpha_j \geq 0$ for all $j = 1,...,n$, and Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862.

- For inseparable classes, the objective is the same as for separable classes, except for the additional condition $0 \leq \alpha_j \leq C$ for all $j = 1,..,n$.

The resulting score function is

$$\widehat{f}(x) = \sum_{j=1}^{n}\widehat{\alpha}_j y_j x' x_j + \widehat{b}\ .$$

$\widehat{b}$ is the estimate of the bias and $\widehat{\alpha}_j$ is the jth estimate of the vector $\widehat{\alpha}$, $j = 1,...,n$. Written this way, the score function is free of the estimate of β as a result of the primal formalization.

The SVM algorithm classifies a new observation z using $\text{sign}\left(\widehat{f}(z)\right)$.

In some cases, a nonlinear boundary separates the classes. Nonlinear SVM works in a transformed predictor space to find an optimal, separating hyperplane.

The dual formalization for nonlinear SVM is

$$0.5\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\alpha_k y_j y_k G(x_j, x_k) - \sum_{j=1}^{n}\alpha_j$$

with respect to $\alpha_1,...,\alpha_n$, subject to $\sum\alpha_j y_j = 0$, $0 \leq \alpha_j \leq C$ for all $j = 1,..,n$, and the KKT complementarity conditions. $G(x_k,x_j)$ are elements of the Gram matrix on page 33-1862. The resulting score function is

$$\widehat{f}(x) = \sum_{j=1}^{n}\widehat{\alpha}_j y_j G(x, x_j) + \widehat{b}\ .$$

For more details, see Understanding Support Vector Machines on page 18-150, [1], and [3].

Algorithms

- For the mathematical formulation of the SVM binary classification algorithm, see “Support Vector Machines for Binary Classification” on page 33-1863 and “Understanding Support Vector Machines” on page 18-150.
- `NaN`, `<undefined>`, empty character vector (`''`), empty string (`""`), and `<missing>` values indicate missing values. `fitcsvm` removes entire rows of data corresponding to a missing response. When computing total weights (see the next bullets), `fitcsvm` ignores any weight corresponding to an observation with at least one missing predictor. This action can lead to unbalanced prior probabilities in balanced-class problems. Consequently, observation box constraints might not equal `BoxConstraint`.
- `fitcsvm` removes observations that have zero weight or prior probability.

- For two-class learning, if you specify the cost matrix C (see `Cost`), then the software updates the class prior probabilities p (see `Prior`) to p_c by incorporating the penalties described in C.

  Specifically, `fitcsvm` completes these steps:

  1. Compute $p_c^* = p'\mathrm{C}$.
  2. Normalize p_c^* so that the updated prior probabilities sum to 1.

  $$p_c = \frac{1}{\sum_{j=1}^{K} p_{c,j}^*} p_c^* .$$

  K is the number of classes.

  3. Reset the cost matrix to the default

  $$\mathrm{C} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

  4. Remove observations from the training data corresponding to classes with zero prior probability.

- For two-class learning, `fitcsvm` normalizes all observation weights (see `Weights`) to sum to 1. The function then renormalizes the normalized weights to sum up to the updated prior probability of the class to which the observation belongs. That is, the total weight for observation j in class k is

  $$w_j^* = \frac{w_j}{\sum_{\forall j \in \text{Class } k} w_j} p_{c,k} .$$

  w_j is the normalized weight for observation j; $p_{c,k}$ is the updated prior probability of class k (see previous bullet).

- For two-class learning, `fitcsvm` assigns a box constraint to each observation in the training data. The formula for the box constraint of observation j is

  $$C_j = nC_0 w_j^* .$$

  n is the training sample size, C_0 is the initial box constraint (see the `'BoxConstraint'` name-value pair argument), and w_j^* is the total weight of observation j (see previous bullet).

- If you set `'Standardize',true` and the `'Cost'`, `'Prior'`, or `'Weights'` name-value pair argument, then `fitcsvm` standardizes the predictors using their corresponding weighted means and weighted standard deviations. That is, `fitcsvm` standardizes predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*} .$$

  $$\mu_j^* = \frac{1}{\sum_k w_k^*} \sum_k w_k^* x_{jk} .$$

  x_{jk} is observation k (row) of predictor j (column).

$$\left(\sigma_j^*\right)^2 = \frac{v_1}{v_1^2 - v_2}\sum_k w_k^*\left(x_{jk} - \mu_j^*\right)^2 .$$

$$v_1 = \sum_j w_j^* .$$

$$v_2 = \sum_j \left(w_j^*\right)^2 .$$

- Assume that `p` is the proportion of outliers that you expect in the training data, and that you set `'OutlierFraction',p`.
  - For one-class learning, the software trains the bias term such that 100`p`% of the observations in the training data have negative scores.
  - The software implements robust learning for two-class learning. In other words, the software attempts to remove 100`p`% of the observations when the optimization algorithm converges. The removed observations correspond to gradients that are large in magnitude.
- If your predictor data contains categorical variables, then the software generally uses full dummy encoding for these variables. The software creates one dummy variable for each level of each categorical variable.
  - The `PredictorNames` property stores one element for each of the original predictor variable names. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `PredictorNames` is a 1-by-3 cell array of character vectors containing the original names of the predictor variables.
  - The `ExpandedPredictorNames` property stores one element for each of the predictor variables, including the dummy variables. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `ExpandedPredictorNames` is a 1-by-5 cell array of character vectors containing the names of the predictor variables and the new dummy variables.
  - Similarly, the `Beta` property stores one beta coefficient for each predictor, including the dummy variables.
  - The `SupportVectors` property stores the predictor values for the support vectors, including the dummy variables. For example, assume that there are *m* support vectors and three predictors, one of which is a categorical variable with three levels. Then `SupportVectors` is an *m*-by-5 matrix.
  - The `X` property stores the training data as originally input and does not include the dummy variables. When the input is a table, `X` contains only the columns used as predictors.
- For predictors specified in a table, if any of the variables contain ordered (ordinal) categories, the software uses ordinal encoding for these variables.
  - For a variable with *k* ordered levels, the software creates *k* – 1 dummy variables. The *j*th dummy variable is –1 for levels up to *j*, and +1 for levels *j* + 1 through *k*.
  - The names of the dummy variables stored in the `ExpandedPredictorNames` property indicate the first level with the value +1. The software stores *k* – 1 additional predictor names for the dummy variables, including the names of levels 2, 3, ..., *k*.
- All solvers implement *L*1 soft-margin minimization.
- For one-class learning, the software estimates the Lagrange multipliers, $\alpha_1,\ldots,\alpha_n$, such that

$$\sum_{j=1}^{n} \alpha_j = n\nu .$$

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[2] Scholkopf, B., J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. “Estimating the Support of a High-Dimensional Distribution.” *Neural Comput.*, Vol. 13, Number 7, 2001, pp. 1443–1471.

[3] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[4] Scholkopf, B., and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press, 2002.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of an SVM classification model into Simulink, you can use the ClassificationSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an SVM model by using `fitcsvm`, the following restrictions apply.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function. For generating code that predicts posterior probabilities given new observations, pass a trained SVM model to `fitPosterior` or `fitSVMPosterior`. The `ScoreTransform` property of the returned model contains an anonymous function that represents the score-to-posterior-probability function and is configured for code generation.
  - For fixed-point code generation, the value of the `'ScoreTransform'` name-value pair argument cannot be `'invlogit'`. Also, the value of the `'KernelFunction'` name-value pair argument must be `'gaussian'`, `'linear'`, or `'polynomial'`.
  - For fixed-point code generation and code generation with a coder configurer, the following additional restrictions apply.
    - Categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`) are not supported. You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.
    - Class labels with the `categorical` data type are not supported. Both the class label value in training data (`Tbl` or `Y`) and the value of the `'ClassNames'` name-value argument cannot be an array with the `categorical` data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationPartitionedModel` | `CompactClassificationSVM` | `fitcsvm`

Topics

Using Support Vector Machines on page 18-154
Understanding Support Vector Machines on page 18-150

Introduced in R2014a

ClassificationSVMCoderConfigurer

Coder configurer for support vector machine (SVM) for one-class and binary classification

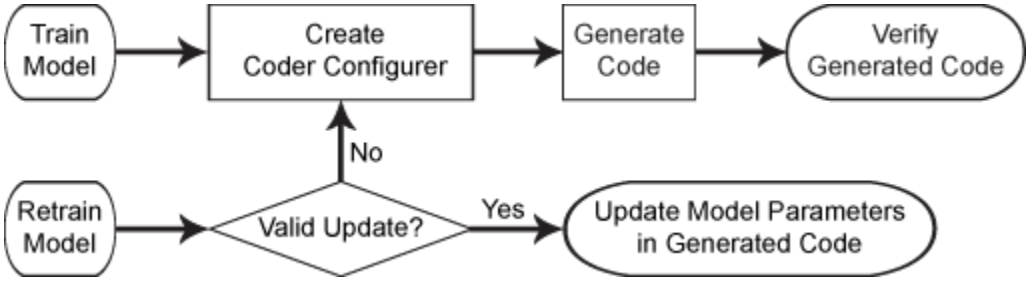
Description

A `ClassificationSVMCoderConfigurer` object is a coder configurer of an SVM classification model (`ClassificationSVM` or `CompactClassificationSVM`).

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes of SVM model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the SVM classification model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the SVM model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of an SVM classification model, see the Code Generation sections of `CompactClassificationSVM`, `predict`, and `update`.

Creation

After training an SVM classification model by using `fitcsvm`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data

`LearnerCoderInput` object

Coder attributes of predictor data to pass to the generated C/C++ code for the `predict` function of the SVM classification model, specified as a `LearnerCoderInput` on page 33-559 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations of three predictor variables, specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

NumOutputs — Number of outputs in predict

1 (default) | 2

Number of output arguments to return from the generated C/C++ code for the `predict` function of the SVM classification model, specified as 1 or 2.

The output arguments of `predict` are `label` (predicted class labels) and `score` (scores or posterior probabilities) in the order of listed. `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function, where `n` is the `NumOutputs` value.

After creating the coder configurer `configurer`, you can specify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of an SVM classification model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-559 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Alpha` — Coder attributes of trained classifier coefficients

`LearnerCoderInput` object

Coder attributes of the trained classifier coefficients (`Alpha` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[s,1]`, where `s` is the number of support vectors in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function, and the model stores the linear predictor coefficients (`Beta`) without the support vectors and related values, then this value must be `false`. Otherwise, this value must be `true`.

`Beta` — Coder attributes of linear predictor coefficients

`LearnerCoderInput` object

Coder attributes of the linear predictor coefficients (`Beta` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[p 1]`, where `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function, and the model stores the linear predictor coefficients (`Beta`) without the support vectors and related values, then this value must be `true`. Otherwise, this value must be `false`.

`Bias` — Coder attributes of bias term
`LearnerCoderInput` object

Coder attributes of the bias term (`Bias` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

`Cost` — Coder attributes of misclassification cost
`LearnerCoderInput` object

Coder attributes of the misclassification cost (`Cost` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — For binary classification, this value must be `[2 2]`. For one-class classification, this value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — For binary classification, the default value is `true`. For one-class classification, this value must be `false`.

`Mu` — Coder attributes of predictor means
`LearnerCoderInput` object

Coder attributes of the predictor means (`Mu` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, this value must be `[1,p]`, where `p` is the number of predictors in `Mdl`. Otherwise, this value must be `[0,0]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, the default value is `true`. Otherwise, this value must be `false`.

`Prior` — Coder attributes of prior probabilities
`LearnerCoderInput` object

Coder attributes of the prior probabilities (`Prior` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — For binary classification, this value must be `[1 2]`. For one-class classification, this value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — For binary classification, the default value is `true`. For one-class classification, this value must be `false`.

`Scale` — Coder attributes of kernel scale parameter
`LearnerCoderInput` object

Coder attributes of the kernel scale parameter (`KernelParameters.Scale` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

`Sigma` — Coder attributes of predictor standard deviations
`LearnerCoderInput` object

Coder attributes of the predictor standard deviations (`Sigma` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, this value must be `[1,p]`, where `p` is the number of predictors in `Mdl`. Otherwise, this value must be `[0,0]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, the default value is `true`. Otherwise, this value must be `false`.

`SupportVectorLabels` — Coder attributes of support vector class labels
`LearnerCoderInput` object

Coder attributes of the support vector class labels (`SupportVectorLabels` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[s,1]`, where `s` is the number of support vectors in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function, and the model stores the linear predictor coefficients (`Beta`) without the support vectors and related values, then this value must be `false`. Otherwise, this value must be `true`.

`SupportVectors` — Coder attributes of support vectors
`LearnerCoderInput` object

Coder attributes of the support vectors (`SupportVectors` of an SVM classification model), specified as a `LearnerCoderInput` on page 33-559 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[s,p]`, where `s` is the number of support vectors, and `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.

- `Tunability` — If you train a model with a linear kernel function, and the model stores the linear predictor coefficients (`Beta`) without the support vectors and related values, then this value must be `false`. Otherwise, this value must be `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code

`'ClassificationSVMModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `ClassificationSVMCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level

`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — Input argument of `predict`
cell array of a `coder.PrimitiveType` object

This property is read-only.

Input argument of the entry-point function `predict.m` for code generation, specified as a cell array of a `coder.PrimitiveType` object. The `coder.PrimitiveType` object includes the coder attributes of the predictor data stored in the `X` property.

If you modify the coder attributes of the predictor data, then the software updates the `coder.PrimitiveType` object accordingly.

The `coder.PrimitiveType` object in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each

`coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-548 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `ionosphere` data set and train a binary SVM classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` object.

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  ClassificationSVMCoderConfigurer with properties:

   Update Inputs:
                  Alpha: [1x1 LearnerCoderInput]
          SupportVectors: [1x1 LearnerCoderInput]
    SupportVectorLabels: [1x1 LearnerCoderInput]
                  Scale: [1x1 LearnerCoderInput]
                   Bias: [1x1 LearnerCoderInput]
                  Prior: [1x1 LearnerCoderInput]
                   Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                      X: [1x1 LearnerCoderInput]
```

```
   Code Generation Parameters:
             NumOutputs: 1
          OutputFileName: 'ClassificationSVMModel'


  Properties, Methods
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the SVM classification model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationSVMModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationSVMModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:56:19
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:56:19
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:56:19
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('ClassificationSVMModel.mat');
```

```
end
switch(command)
    case 'update'
        % Update struct fields: Alpha
        %                       SupportVectors
        %                       SupportVectorLabels
        %                       Scale
        %                       Bias
        %                       Prior
        %                       Cost
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of SVM Classification Model in Generated Code

Train a SVM model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`). Train a binary SVM classification model using the first 50 observations.

```
load ionosphere
Mdl = fitcsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `ClassificationSVM` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted labels and scores.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:),'NumOutputs',2);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM classification model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 34];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 34 predictors, so the value of the `SizeVector` attribute must be 34 and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 34];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
VariableDimensions attribute for SupportVectorLabels has been modified to satisfy configuration c
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM classification model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `ClassificationSVMModel` for the two entry-point functions in the `codegen\mex\ClassificationSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score] = predict(Mdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
```

Compare `label` and `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`score_mex` might include round-off differences compared with `score`. In this case, compare `score_mex` and `score`, allowing a small tolerance.

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `score` and `score_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitcsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score] = predict(retrainedMdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `labels` and `labels_mex` are equal, and the score values are equal within the tolerance.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array

Attribute Name	Description
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of the coefficients `Alpha` of the coder configurer `configurer` as follows:

```
configurer.Alpha.SizeVector = [100 1];
configurer.Alpha.VariableDimensions = [1 0];
configurer.Alpha.DataType = 'double';
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

See Also

`ClassificationECOCCoderConfigurer` | `ClassificationSVM` | `CompactClassificationSVM` | `learnerCoderConfigurer` | `predict` | `update`

Topics

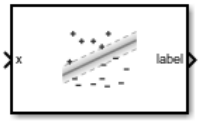
“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

ClassificationSVM Predict

Classify observations using support vector machine (SVM) classifier for one-class and binary classification

Library: Statistics and Machine Learning Toolbox / Classification



Description

The ClassificationSVM Predict block classifies observations using an SVM classification object (`ClassificationSVM` or `CompactClassificationSVM`) for one-class and two-class (binary) classification.

Import a trained SVM classification object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **label** returns a predicted class label for the observation. You can add an optional output port **score** that returns predicted class scores or posterior probabilities.

Ports

Input

x — Predictor data

row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the SVM model specified by **`Select trained machine learning model`**.
- If you set `'Standardize',true` in `fitcsvm` when training the SVM model, then the ClassificationSVM Predict block standardizes the values of **x** using the means and standard deviations in the `Mu` and `Sigma` properties (respectively) of the SVM model.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

label — Predicted class label

scalar

Predicted class label, returned as a scalar.

Dependencies

- For one-class learning, **label** is the value representing the positive class.

- For two-class learning, **label** is the class yielding the largest score or the largest posterior probability.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point` | `enumerated`

`score` — Predicted class scores or posterior probabilities

scalar | 1-by-2 vector

Predicted class scores on page 33-572 or posterior probabilities on page 33-572, returned as a scalar for one-class learning or a 1-by-2 vector for two-class learning.

- For one-class learning, **score** is the classification score of the positive class. You cannot obtain posterior probabilities for one-class learning.
- For two-class learning, **score** is a 1-by-2 vector.
  - The first and second element of **score** correspond to the classification scores of the negative class (`svmMdl.ClassNames(1)`) and the positive class (`svmMdl.ClassNames(2)`), respectively, where `svmMdl` is the SVM model specified by **`Select trained machine learning model`**. You can use the `ClassNames` property of `svmMdl` to check the negative and positive class names.
  - If you fit the optimal score-to-posterior-probability transformation function using `fitPosterior` or `fitSVMPosterior`, then **score** contains class posterior probabilities. Otherwise, **score** contains class scores.

Dependencies

To enable this port, select the check box for `Add output port for predicted class scores` on the **Main** tab of the Block Parameters dialog box.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

`Select trained machine learning model` — SVM classification model

`svmMdl` (default) | `ClassificationSVM` object | `CompactClassificationSVM` object

Specify the name of a workspace variable that contains a `ClassificationSVM` object or `CompactClassificationSVM` object.

When you train the SVM model by using `fitcsvm`, the following restrictions apply:

- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ScoreTransform'` name-value argument cannot be `'invlogit'` or an anonymous function. For a block that predicts posterior probabilities given new observations, pass a trained SVM model to `fitPosterior` or `fitSVMPosterior`.

- The value of the `'KernelFunction'` name-value argument must be `'gaussian'` (same as `'rbf'`, default for one-class learning), `'linear'` (default for two-class learning), or `'polynomial'`.

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `ClassificationSVM` object | `CompactClassificationSVM` object
Default: `'svmMdl'`

`Add output port for predicted class scores` — Add second output port for predicted class scores

`off` (default) | `on`

Select the check box to include the second output port **score** in the ClassificationSVM Predict block.

Programmatic Use

Block Parameter: `ShowOutputScore`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations

`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use

Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action

`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use
Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type
`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use
Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Label data type` — Data type of label output

`Inherit: Inherit via back propagation` | `Inherit: auto` | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `Enum: <class name>` | `<data type expression>`

Specify the data type for the **label** output. The type can be inherited, specified as an enumerated data type, or expressed as a data type object such as `Simulink.NumericType`.

When you select an inherited option, the software behaves as follows:

- `Inherit: Inherit via back propagation` (default for numeric and logical labels) — Simulink automatically determines the **Label data type** of the block during data type propagation (see "Data Type Propagation" (Simulink)). In this case, the block uses the data type of a downstream block or signal object.
- `Inherit: auto` (default for nonnumeric labels) — The block uses an autodefined enumerated data type variable. For example, suppose the workspace variable name specified by `Select trained machine learning model` is `myMdl`, and the class labels are `class 1` and `class 2`. Then, the corresponding **label** values are `myMdl_enumLabels.class_1` and `myMdl_enumLabels.class_2`. The block converts the class labels to valid MATLAB identifiers by using the `matlab.lang.makeValidName` function.

For more information about data types, see "Control Signal Data Types" (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see "Specify Data Types Using Data Type Assistant" (Simulink).

Dependencies

The supported data types depend on the labels used in the model specified by `Select trained machine learning model`.

- If the model uses numeric or logical labels, the supported data types are `Inherit: Inherit via back propagation` (default), `double`, `single`, `half`, `int8`, `uint8`, `int16`, `uint16`, `int32`, `uint32`, `int64`, `uint64`, `boolean`, fixed point, and a data type object.
- If the model uses nonnumeric labels, the supported data types are `Inherit: auto` (default), `Enum: <class name>`, and a data type object.

Programmatic Use

Block Parameter: `LabelDataTypeStr`
Type: character vector
Values: `'Inherit: Inherit via back propagation'` | `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'Enum: <class name>'` | `'<data type expression>'`
Default: `'Inherit: Inherit via back propagation'` (for numeric and logical labels) | `'Inherit: auto'` (for nonnumeric labels)

`Label minimum` — Minimum value of label output for range checking

`[ ]` (default) | scalar

Lower value of the **label** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label minimum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use

Block Parameter: `LabelOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Label maximum` — Maximum value of label output for range checking

`[]` (default) | scalar

Upper value of the **label** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label maximum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use
Block Parameter: `LabelOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score data type` — Data type of score output

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **score** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button 

to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use
Block Parameter: `ScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Score minimum` — Minimum value of score output for range checking

`[]` (default) | scalar

Lower value of the **score** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score minimum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use
Block Parameter: `ScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score maximum` — Maximum value of score output for range checking

`[]` (default) | scalar

Upper value of the **score** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score maximum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use
Block Parameter: `ScoreOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score data type` — Untransformed score data type

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `boolean` | `<data type expression>`

Specify the data type for the internal untransformed scores. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button 

to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses a score transformation other than `'none'` (default, same as `'identity'`).

- If the model uses no score transformations (`'none'` or `'identity'`), then you can specify the score data type by using `Score data type`.
- If the model uses a score transformation other than `'none'` or `'identity'`, then you can specify the data type of untransformed raw scores by using this parameter and specify the data type of transformed scores by using `Score data type`.

You can change the score transformation option by specifying the `'ScoreTransform'` name-value argument during training, or by changing the `ScoreTransform` property after training.

Programmatic Use
Block Parameter: `RawScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Raw score minimum` — Minimum untransformed score for range checking

`[]` (default) | scalar

Lower value of the untransformed score range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score minimum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use
Block Parameter: `RawScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score maximum` — Maximum untransformed score for range checking

`[]` (default) | scalar

Upper value of the untransformed score range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.

- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score maximum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use

Block Parameter: `RawScoreOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Kernel data type` — Kernel computation data type

`double` (default) | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `int64` | `uint64` | `uint32` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type of a parameter for kernel computation. The type can be specified directly or expressed as a data type object such as `Simulink.NumericType`.

The **Kernel data type** parameter specifies the data type of a different parameter depending on the type of kernel function of the specified SVM model. You specify the `'KernelFunction'` name-value argument when training the SVM model.

`'KernelFunction'` value	Data Type
`'gaussian'` or `'rbf'`	**Kernel data type** specifies the data type of the squared distance $D^2 = \|x - s\|^2$ for the Gaussian kernel $G(x,s) = \exp\left(-D^2\right)$, where x is the predictor data for an observation and s is a support vector.
`'linear'`	**Kernel data type** specifies the data type for the output of the linear kernel function $G(x,s) = xs'$, where x is the predictor data for an observation and s is a support vector.
`'polynomial'`	**Kernel data type** specifies the data type for the output of the polynomial kernel function $G(x,s) = (1 + xs')^p$, where x is the predictor data for an observation, s is a support vector, and p is a polynomial kernel function order.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use
Block Parameter: `KernelDataTypeStr`
Type: character vector
Values: `'double'|'single'|'half'|'int8'|'uint8'|'int16'|'uint16'|'int32'|'uint32'|'uint64'|'int64'|'boolean'|'fixdt(1,16)'|'fixdt(1,16,0)'|'fixdt(1,16,2^0,0)'|'<data type expression>'`
Default: `'double'`

`Kernel minimum` — Minimum kernel computation value for range checking

`[]` (default) | scalar

Lower value of the kernel computation internal variable range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Kernel minimum** parameter does not saturate or clip the actual kernel computation value signal.

Programmatic Use
Block Parameter: `KernelOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Kernel maximum` — Maximum kernel computation value for range checking

`[]` (default) | scalar

Upper value of the kernel computation internal variable range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Kernel maximum** parameter does not saturate or clip the actual kernel computation value signal.

Programmatic Use

Block Parameter: `KernelOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `enumerated` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

More About

Classification Score

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

The positive class classification score $f(x)$ is the trained SVM classification function. $f(x)$ is also the numerical predicted response for x, or the score for predicting x into the positive class.

$$f(x) = \sum_{j=1}^{n} \alpha_j y_j G(x_j, x) + b,$$

where $(\alpha_1, ..., \alpha_n, b)$ are the estimated SVM parameters, $G(x_j, x)$ is the dot product in the predictor space between x and the support vectors, and the sum includes the training set observations. The negative class classification score for x, or the score for predicting x into the negative class, is $-f(x)$.

If $G(x_j,x) = x_j'x$ (the linear kernel), then the score function reduces to

$$f(x) = (x/s)'\beta + b\,.$$

s is the kernel scale and β is the vector of fitted linear coefficients.

For more details, see “Understanding Support Vector Machines” on page 18-150.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For SVM, the posterior probability is a function of the score $P(s)$ that observation j is in class $k = \{-1,1\}$.

- For separable classes, the posterior probability is the step function

$$P(s_j) = \begin{cases} 0; & s < \max_{y_k = -1} s_k \\ \pi; & \max_{y_k = -1} s_k \le s_j \le \min_{y_k = +1} s_k \\ 1; & s_j > \min_{y_k = +1} s_k \end{cases},$$

  where:

  - s_j is the score of observation j.
  - +1 and –1 denote the positive and negative classes, respectively.
  - π is the prior probability that an observation is in the positive class.
- For inseparable classes, the posterior probability is the sigmoid function

$$P(s_j) = \frac{1}{1 + \exp(As_j + B)},$$

  where the parameters A and B are the slope and intercept parameters, respectively.

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Tips

- If you are using a linear SVM model and it has many support vectors, then prediction (classifying observations) can be slow. To efficiently classify observations based on a linear SVM model, remove the support vectors from the `ClassificationSVM` or `CompactClassificationSVM` object by using `discardSupportVectors`.

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of an SVM classification object (`ClassificationSVM` or `CompactClassificationSVM`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the ClassificationSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Compatibility Considerations

Default value of Label data type is `Inherit: Inherit via back propagation` for numeric and logical labels and `Inherit: auto` for nonnumeric labels
Behavior changed in R2021a

Starting in R2021a, the default data type value and the supported data types of the **Label data type** parameter depend on the labels used in the model specified by `Select trained machine learning model`. The default value is `Inherit: Inherit via back propagation` for numeric and logical labels, and `Inherit: auto` for nonnumeric labels.

If you specified **Label data type** as `Inherit: Inherit via back propagation` for nonnumeric labels or `Inherit: Inherit from 'Constant value'`, then change the value to `Inherit: auto`.

Default value of Score data type and Raw score data type is `Inherit: auto`
Behavior changed in R2021a

Starting in R2021a, the default value of the parameters **Score data type** and **Raw score data type** is `Inherit: auto`.

Specify Kernel data type as a data type name or data type object
Behavior changed in R2021a

Starting in R2021a, the **Kernel data type** parameter does not support inherited options. You can specify **Kernel data type** as a supported data type name or data type object.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion
Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks
ClassificationEnsemble Predict | ClassificationTree Predict | RegressionSVM Predict

Objects
`ClassificationSVM` | `CompactClassificationSVM`

Functions
`fitcsvm` | `predict`

Topics
“Predict Class Labels Using ClassificationTree Predict Block” on page 32-121
“Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130
“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2020b

ClassificationTree class

Superclasses: `CompactClassificationTree`

Binary decision tree for multiclass classification

Description

A `ClassificationTree` object represents a decision tree with binary splits for classification. An object of this class can predict responses for new data using the `predict` method. The object contains the data used for training, so it can also compute resubstitution predictions.

Construction

Create a `ClassificationTree` object by using `fitctree`.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CategoricalSplit`

An *n*-by-2 cell array, where *n* is the number of categorical splits in `tree`. Each row in `CategoricalSplits` gives left and right values for a categorical split. For each branch node with categorical split `j` based on a categorical predictor variable `z`, the left child is chosen if `z` is in `CategoricalSplits(j,1)` and the right child is chosen if `z` is in `CategoricalSplits(j,2)`. The splits are in the same order as nodes of the tree. Find the nodes for these splits by selecting `'categorical'` cuts from top to bottom in the `CutType` property.

`Children`

An *n*-by-2 array containing the numbers of the child nodes for each node in `tree`, where *n* is the number of nodes. Leaf nodes have child node `0`.

`ClassCount`

An *n*-by-*k* array of class counts for the nodes in `tree`, where *n* is the number of nodes and *k* is the number of classes. For any node number `i`, the class counts `ClassCount(i,:)` are counts of observations (from the data used in fitting the tree) from each class satisfying the conditions for node `i`.

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a categorical array, cell array of character vectors, character array, logical vector, or a numeric vector. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`ClassProbability`

An *n*-by-*k* array of class probabilities for the nodes in `tree`, where *n* is the number of nodes and *k* is the number of classes. For any node number `i`, the class probabilities `ClassProbability(i,:)` are the estimated probabilities for each class for a point satisfying the conditions for node `i`.

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response. This property is read-only.

`CutCategories`

An *n*-by-2 cell array of the categories used at branches in `tree`, where *n* is the number of nodes. For each branch node `i` based on a categorical predictor variable `X`, the left child is chosen if `X` is among

the categories listed in `CutCategories{i,1}`, and the right child is chosen if `X` is among those listed in `CutCategories{i,2}`. Both columns of `CutCategories` are empty for branch nodes based on continuous predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPoint`

An *n*-element vector of the values used as cut points in `tree`, where *n* is the number of nodes. For each branch node `i` based on a continuous predictor variable `X`, the left child is chosen if `X<CutPoint(i)` and the right child is chosen if `X>=CutPoint(i)`. `CutPoint` is `NaN` for branch nodes based on categorical predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutType`

An *n*-element cell array indicating the type of cut at each node in `tree`, where *n* is the number of nodes. For each node `i`, `CutType{i}` is:

- `'continuous'` — If the cut is defined in the form `X < v` for a variable `X` and cut point `v`.
- `'categorical'` — If the cut is defined by whether a variable `X` takes a value in a set of categories.
- `''` — If `i` is a leaf node.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictor`

An *n*-element cell array of the names of the variables used for branching in each node in `tree`, where *n* is the number of nodes. These variables are sometimes known as *cut variables*. For leaf nodes, `CutPredictor` contains an empty character vector.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictorIndex`

An *n*-element array of numeric indices for the variables used for branching in each node in `tree`, where *n* is the number of nodes. For more information, see `CutPredictor`.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty

when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`IsBranchNode`

An *n*-element logical vector that is `true` for each branch node and `false` for each leaf node of `tree`.

`ModelParameters`

Parameters used in training `tree`. To display all parameter values, enter `tree.ModelParameters`. To access a particular parameter, use dot notation.

`NumObservations`

Number of observations in the training data, a numeric scalar. `NumObservations` can be less than the number of rows of input data `X` when there are missing values in `X` or response `Y`.

`NodeClass`

An *n*-element cell array with the names of the most probable classes in each node of `tree`, where *n* is the number of nodes in the tree. Every element of this array is a character vector equal to one of the class names in `ClassNames`.

`NodeError`

An *n*-element vector of the errors of the nodes in `tree`, where *n* is the number of nodes. `NodeError(i)` is the misclassification probability for node `i`.

`NodeProbability`

An *n*-element vector of the probabilities of the nodes in `tree`, where *n* is the number of nodes. The probability of a node is computed as the proportion of observations from the original data that satisfy the conditions for the node. This proportion is adjusted for any prior probabilities assigned to each class.

`NodeRisk`

An *n*-element vector of the risk of the nodes in the tree, where *n* is the number of nodes. The risk for each node is the measure of impurity (Gini index or deviance) for this node weighted by the node probability. If the tree is grown by twoing, the risk for each node is zero.

`NodeSize`

An *n*-element vector of the sizes of the nodes in `tree`, where *n* is the number of nodes. The size of a node is defined as the number of observations from the data used to create the tree that satisfy the conditions for the node.

`NumNodes`

The number of nodes in `tree`.

`Parent`

An *n*-element vector containing the number of the parent node for each node in `tree`, where *n* is the number of nodes. The parent of the root node is `0`.

`PredictorNames`

Cell array of character vectors containing the predictor names, in the order which they appear in `X`.

`Prior`

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`. The number of elements of `Prior` is the number of unique classes in the response. This property is read-only.

`PruneAlpha`

Numeric vector with one element per pruning level. If the pruning level ranges from 0 to *M*, then `PruneAlpha` has *M* + 1 elements sorted in ascending order. `PruneAlpha(1)` is for pruning level 0 (no pruning), `PruneAlpha(2)` is for pruning level 1, and so on.

`PruneList`

An *n*-element numeric vector with the pruning levels in each node of `tree`, where *n* is the number of nodes. The pruning levels range from 0 (no pruning) to *M*, where *M* is the distance between the deepest leaf and the root node.

`ResponseName`

A character vector that specifies the name of the response variable (`Y`).

`RowsUsed`

An *n*-element logical vector indicating which rows of the original predictor data (`X`) were used in fitting. If the software uses all rows of `X`, then `RowsUsed` is an empty array (`[]`).

`ScoreTransform`

Function handle for transforming predicted classification scores, or character vector representing a built-in transformation function.

`none` means no transformation, or `@(x)x`.

To change the score transformation function to, for example, *`function`*, use dot notation.

- For available functions (see `fitctree`), enter

  ```
  Mdl.ScoreTransform = 'function';
  ```

- You can set a function handle for an available function, or a function you define yourself by entering

  ```
  tree.ScoreTransform = @function;
  ```

`SurrogateCutCategories`

An *n*-element cell array of the categories used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutCategories{k}` is a cell array. The length of

`SurrogateCutCategories{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutCategories{k}` is either an empty character vector for a continuous surrogate predictor, or is a two-element cell array with categories for a categorical surrogate predictor. The first element of this two-element cell array lists categories assigned to the left child by this surrogate split, and the second element of this two-element cell array lists categories assigned to the right child by this surrogate split. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutCategories` contains an empty cell.

`SurrogateCutFlip`

An *n*-element cell array of the numeric cut assignments used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutFlip{k}` is a numeric vector. The length of `SurrogateCutFlip{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutFlip{k}` is either zero for a categorical surrogate predictor, or a numeric cut assignment for a continuous surrogate predictor. The numeric cut assignment can be either –1 or +1. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and the cut assignment for this surrogate split is +1, or if $Z\geq C$ and the cut assignment for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and the cut assignment for this surrogate split is +1, or if $Z<C$ and the cut assignment for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutFlip` contains an empty array.

`SurrogateCutPoint`

An *n*-element cell array of the numeric values used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutPoint{k}` is a numeric vector. The length of `SurrogateCutPoint{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutPoint{k}` is either `NaN` for a categorical surrogate predictor, or a numeric cut for a continuous surrogate predictor. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z<C$ and `SurrogateCutFlip` for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables returned by `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPoint` contains an empty cell.

`SurrogateCutType`

An *n*-element cell array indicating types of surrogate splits at each node in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutType{k}` is a cell array with the types of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutType` contains an empty cell. A surrogate split type can be either `'continuous'` if the cut is defined in the form `Z<V` for a variable `Z` and cut point `V` or `'categorical'` if the cut is defined by whether `Z` takes a value in a set of categories.

`SurrogateCutPredictor`

An *n*-element cell array of the names of the variables used for surrogate splits in each node in `tree`, where *n* is the number of nodes in `tree`. Every element of `SurrogateCutPredictor` is a cell array with the names of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPredictor` contains an empty cell.

`SurrogatePredictorAssociation`

An *n*-element cell array of the predictive measures of association for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogatePredictorAssociation{k}` is a numeric vector. The length of `SurrogatePredictorAssociation{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogatePredictorAssociation{k}` gives the predictive measure of association between the optimal split and this surrogate split. The order of the surrogate split variables at each node is the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogatePredictorAssociation` contains an empty cell.

`W`

The scaled `weights`, a vector with length *n*, the number of rows in `X`.

`X`

A matrix or table of predictor values. Each column of `X` represents one variable, and each row represents one observation.

`Y`

A categorical array, cell array of character vectors, character array, logical vector, or a numeric vector. Each row of `Y` represents the classification of the corresponding row of `X`.

Object Functions

compact	Compact tree
compareHoldout	Compare accuracies of two classification models using new data
crossval	Cross-validated decision tree
cvloss	Classification error by cross validation
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification error
margin	Classification margins
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels using classification tree
predictorImportance	Estimates of predictor importance for classification tree
prune	Produce sequence of classification subtrees by pruning
resubEdge	Classification edge by resubstitution
resubLoss	Classification error by resubstitution
resubMargin	Classification margins by resubstitution
resubPredict	Predict resubstitution labels of classification tree

surrogateAssociation	Mean predictive measure of association for surrogate splits in classification tree
shapley	Shapley values
testckfold	Compare accuracies of two classification models by repeated cross-validation
view	View classification tree

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Grow a Classification Tree

Grow a classification tree using the `ionosphere` data set.

```
load ionosphere
tc = fitctree(X,Y)
```

```
tc =
  ClassificationTree
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351


  Properties, Methods
```

Control Tree Depth

You can control the depth of the trees using the `MaxNumSplits`, `MinLeafSize`, or `MinParentSize` name-value pair parameters. `fitctree` grows deep decision trees by default. You can grow shallower trees to reduce model complexity or computation time.

Load the `ionosphere` data set.

```
load ionosphere
```

The default values of the tree depth controllers for growing classification trees are:

- `n - 1` for `MaxNumSplits`. `n` is the training sample size.
- `1` for `MinLeafSize`.
- `10` for `MinParentSize`.

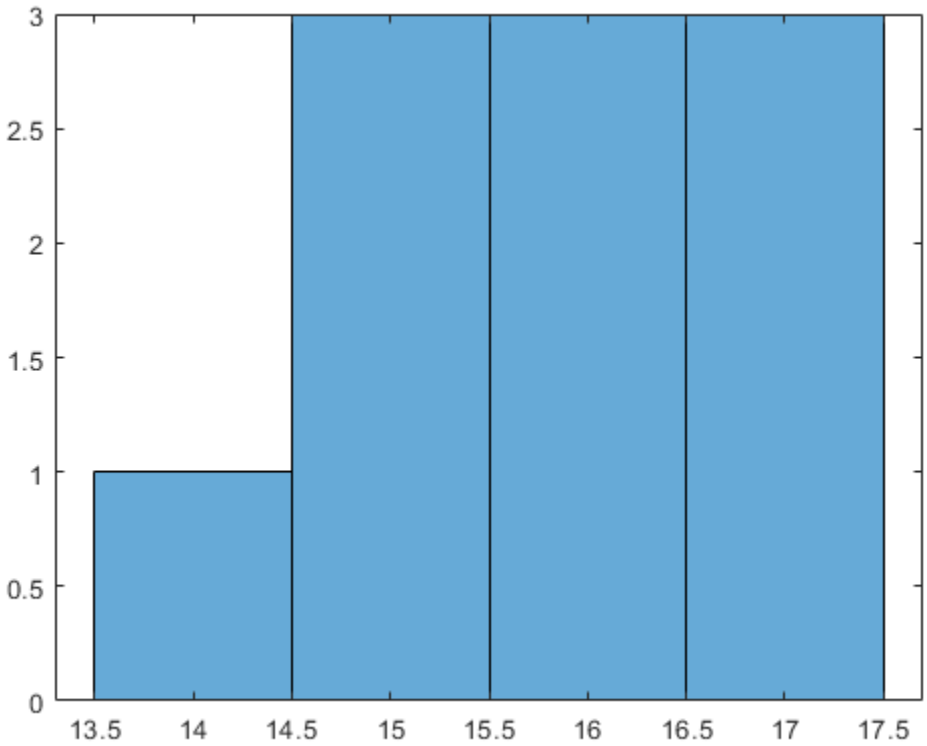
These default values tend to grow deep trees for large training sample sizes.

Train a classification tree using the default values for tree depth control. Cross-validate the model by using 10-fold cross-validation.

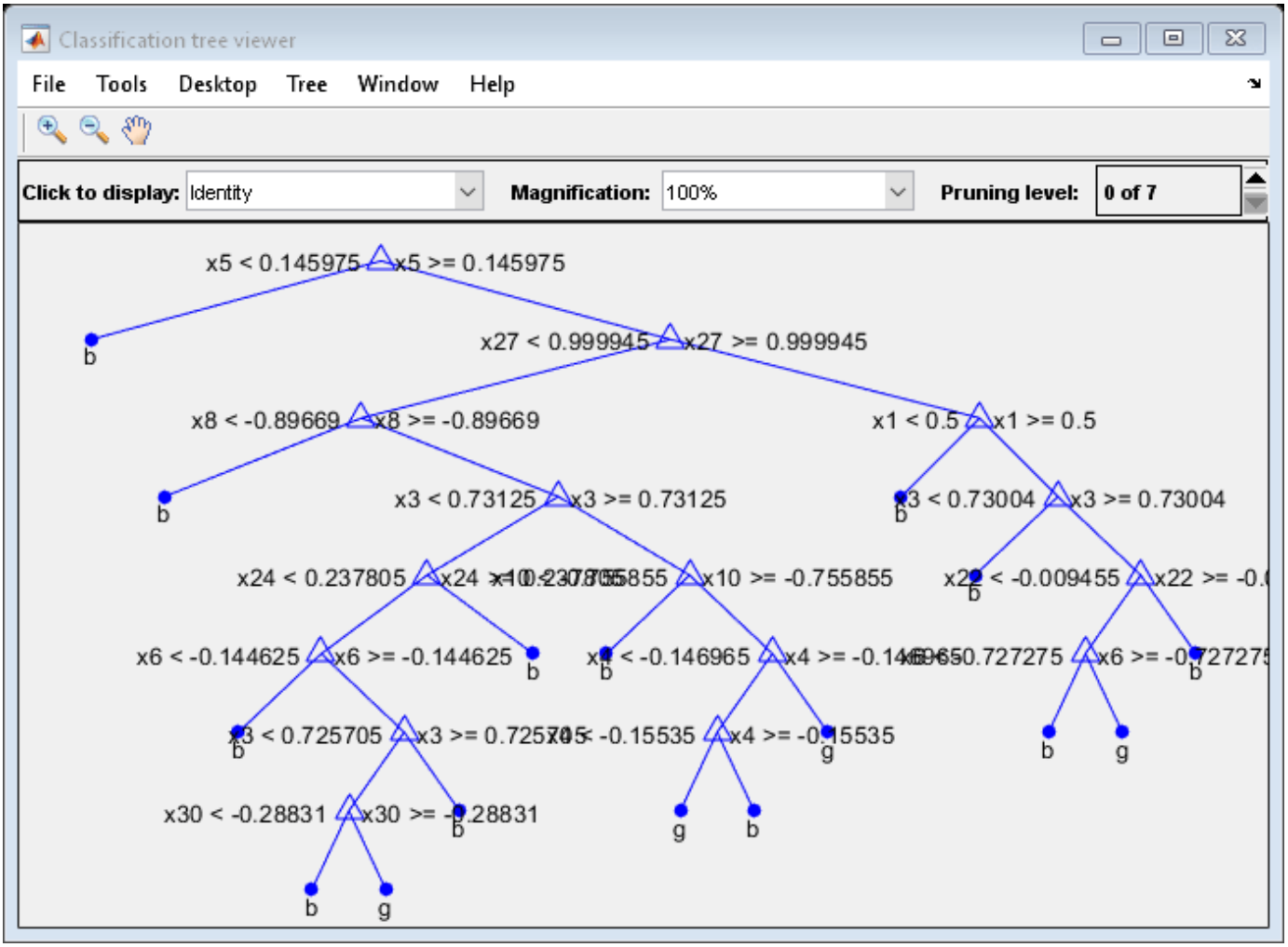
```
rng(1); % For reproducibility
MdlDefault = fitctree(X,Y,'CrossVal','on');
```

Draw a histogram of the number of imposed splits on the trees. Also, view one of the trees.

```
numBranches = @(x)sum(x.IsBranch);
mdlDefaultNumSplits = cellfun(numBranches, MdlDefault.Trained);

figure;
histogram(mdlDefaultNumSplits)
```



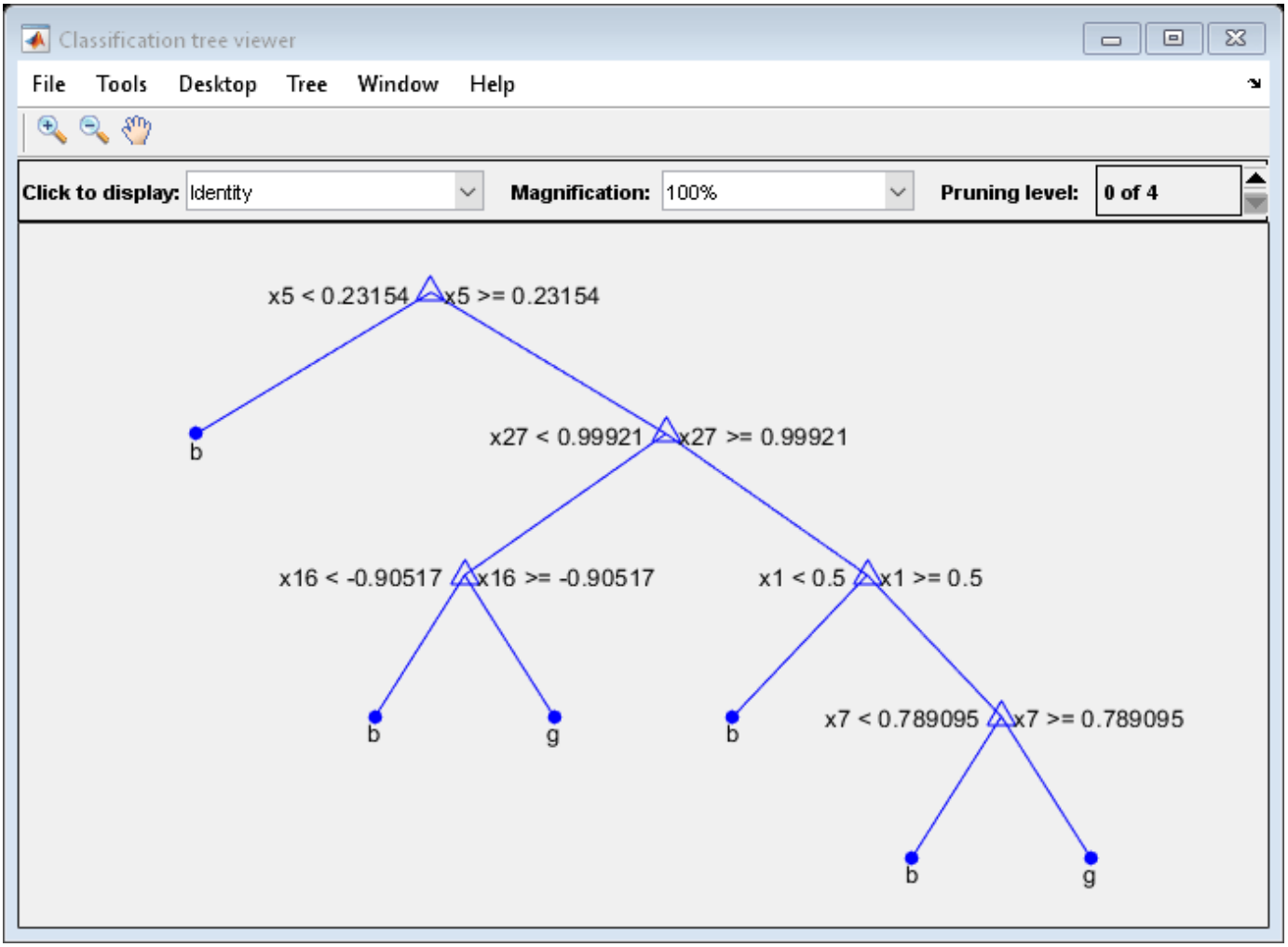

```
view(MdlDefault.Trained{1},'Mode','graph')
```



The average number of splits is around 15.

Suppose that you want a classification tree that is not as complex (deep) as the ones trained using the default number of splits. Train another classification tree, but set the maximum number of splits at 7, which is about half the mean number of splits from the default classification tree. Cross-validate the model by using 10-fold cross-validation.

```
Mdl7 = fitctree(X,Y,'MaxNumSplits',7,'CrossVal','on');
view(Mdl7.Trained{1},'Mode','graph')
```



Compare the cross-validation classification errors of the models.

```
classErrorDefault = kfoldLoss(MdlDefault)
```

```
classErrorDefault = 0.1168
```

```
classError7 = kfoldLoss(Mdl7)
```

```
classError7 = 0.1311
```

`Mdl7` is much less complex and performs only slightly worse than `MdlDefault`.

More About

Impurity and Node Error

A decision tree splits nodes based on either impurity or node error.

Impurity means one of several things, depending on your choice of the `SplitCriterion` name-value pair argument:

- Gini's Diversity Index (`gdi`) — The Gini index of a node is

  $$1 - \sum_i p^2(i),$$

  where the sum is over the classes i at the node, and $p(i)$ is the observed fraction of classes with class i that reach the node. A node with just one class (a pure node) has Gini index `0`; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.
- Deviance (`'deviance'`) — With $p(i)$ defined the same as for the Gini index, the deviance of a node is

  $$-\sum_i p(i)\log_2 p(i)\,.$$

  A pure node has deviance `0`; otherwise, the deviance is positive.
- Twoing rule (`'twoing'`) — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class i in the left child node after a split, and $R(i)$ denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

  $$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2,$$

  where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.
- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $1 – p(j)$.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of a classification tree model into Simulink, you can use the ClassificationTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train a classification tree using `fitctree`, the following restrictions apply.

- The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function. For fixed-point code generation, the `'ScoreTransform'` value cannot be `'invlogit'`.
- You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value pair argument must be `'off'`.
- For fixed-point code generation and code generation with a coder configurer, the following additional restrictions apply.
  - Categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`) are not supported. You cannot use the `'CategoricalPredictors'` name-value argument.To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.
  - Class labels with the `categorical` data type are not supported. Both the class label value in training data (`Tbl` or `Y`) and the value of the `'ClassNames'` name-value argument cannot be an array with the `categorical` data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationEnsemble` | `CompactClassificationTree` | `RegressionTree` | `compareHoldout` | `fitctree` | `predict`

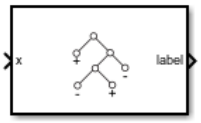
Topics

“Decision Trees” on page 19-2

Introduced in R2011a

ClassificationTree Predict

Classify observations using decision tree classifier
Library: Statistics and Machine Learning Toolbox / Classification



Description

The ClassificationTree Predict block classifies observations using a classification tree object (`ClassificationTree` or `CompactClassificationTree`) for multiclass classification.

Import a trained classification object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **label** returns a predicted class label for the observation. You can add an optional output port **score** that returns predicted class scores or posterior probabilities.

Ports

Input

x — Predictor data
row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the model specified by **`Select trained machine learning model`**.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

`label` — Predicted class label
scalar

Predicted class label, returned as a scalar. The predicted class is the class that minimizes the expected classification cost. For more details, see the “More About” on page 33-4868 section of the `predict` function reference page.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point` | `enumerated`

`score` — Predicted class scores or posterior probabilities
row vector

Predicted class scores or posterior probabilities, returned as a row vector of size 1-by-*k*, where *k* is the number of classes in the tree model.

The classification score of a leaf node is the posterior probability of the classification at the node. The posterior probability of the classification at a node is the number of training observations that lead to the node with the classification, divided by the number of training observations that lead to the node.

To check the order of the classes, use the `ClassNames` property of the tree model specified by **`Select trained machine learning model`**.

Dependencies

To enable this port, select the check box for `Add output port for predicted class scores` on the **Main** tab of the Block Parameters dialog box.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

`Select trained machine learning model` — Classification tree model

`treeMdl` (default) | `ClassificationTree` object | `CompactClassificationTree` object

Specify the name of a workspace variable that contains a `ClassificationTree` object or `CompactClassificationTree` object.

When you train the model by using `fitctree`, the following restrictions apply:

- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ScoreTransform'` name-value argument cannot be `'invlogit'` or an anonymous function.
- You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value argument must be `'off'` (default).

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `ClassificationTree` object | `CompactClassificationTree` object
Default: `'treeMdl'`

`Add output port for predicted class scores` — Add second output port for predicted class scores

`off` (default) | `on`

Select the check box to include the second output port **score** in the ClassificationTree Predict block.

Programmatic Use

Block Parameter: `ShowOutputScore`

Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations
`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use
Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action
`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.

Action	Rationale	Impact on Overflows	Example
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use

Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type

`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use

Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Label data type` — Data type of label output

`Inherit: Inherit via back propagation` | `Inherit: auto` | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `Enum: <class name>` | `<data type expression>`

Specify the data type for the **label** output. The type can be inherited, specified as an enumerated data type, or expressed as a data type object such as `Simulink.NumericType`.

When you select an inherited option, the software behaves as follows:

- `Inherit: Inherit via back propagation` (default for numeric and logical labels) — Simulink automatically determines the **Label data type** of the block during data type propagation (see "Data Type Propagation" (Simulink)). In this case, the block uses the data type of a downstream block or signal object.
- `Inherit: auto` (default for nonnumeric labels) — The block uses an autodefined enumerated data type variable. For example, suppose the workspace variable name specified by `Select`

`trained machine learning model` is `myMdl`, and the class labels are `class 1` and `class 2`. Then, the corresponding **label** values are `myMdl_enumLabels.class_1` and `myMdl_enumLabels.class_2`. The block converts the class labels to valid MATLAB identifiers by using the `matlab.lang.makeValidName` function.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Dependencies

The supported data types depend on the labels used in the model specified by `Select trained machine learning model`.

- If the model uses numeric or logical labels, the supported data types are `Inherit: Inherit via back propagation` (default), `double`, `single`, `half`, `int8`, `uint8`, `int16`, `uint16`, `int32`, `uint32`, `int64`, `uint64`, `boolean`, fixed point, and a data type object.
- If the model uses nonnumeric labels, the supported data types are `Inherit: auto` (default), `Enum: <class name>`, and a data type object.

Programmatic Use

Block Parameter: `LabelDataTypeStr`
Type: character vector
Values: `'Inherit: Inherit via back propagation'` | `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'Enum: <class name>'` | `'<data type expression>'`
Default: `'Inherit: Inherit via back propagation'` (for numeric and logical labels) | `'Inherit: auto'` (for nonnumeric labels)

`Label minimum` — Minimum value of label output for range checking

`[ ]` (default) | scalar

Lower value of the **label** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label minimum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use

Block Parameter: `LabelOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Label maximum` — Maximum value of label output for range checking

`[]` (default) | scalar

Upper value of the **label** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Label maximum** parameter does not saturate or clip the actual **label** output signal. Use the Saturation block instead.

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses numeric labels.

Programmatic Use

Block Parameter: `LabelOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score data type` — Data type of score output

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **score** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use

Block Parameter: `ScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Score minimum` — Minimum value of score output for range checking

[ ] (default) | scalar

Lower value of the **score** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score minimum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `ScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Score maximum` — Maximum value of score output for range checking

[ ] (default) | scalar

Upper value of the **score** output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.

- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Score maximum** parameter does not saturate or clip the actual **score** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `ScoreOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score data type` — Untransformed score data type

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `boolean` | `<data type expression>`

Specify the data type for the internal untransformed scores. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button 

to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Dependencies

You can specify this parameter only if the model specified by `Select trained machine learning model` uses a score transformation other than `'none'` (default, same as `'identity'`).

- If the model uses no score transformations (`'none'` or `'identity'`), then you can specify the score data type by using `Score data type`.
- If the model uses a score transformation other than `'none'` or `'identity'`, then you can specify the data type of untransformed raw scores by using this parameter and specify the data type of transformed scores by using `Score data type`.

You can change the score transformation option by specifying the `'ScoreTransform'` name-value argument during training, or by changing the `ScoreTransform` property after training.

Programmatic Use

Block Parameter: `RawScoreDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Raw score minimum` — Minimum untransformed score for range checking

`[ ]` (default) | scalar

Lower value of the untransformed score range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score minimum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use

Block Parameter: `RawScoreOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Raw score maximum` — Maximum untransformed score for range checking

`[ ]` (default) | scalar

Upper value of the untransformed score range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Raw score maximum** parameter does not saturate or clip the actual untransformed score signal.

Programmatic Use

Block Parameter: `RawScoreOutMax`

Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `enumerated` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of a classification tree object (`ClassificationTree` or `CompactClassificationTree`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the ClassificationTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion

Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks

ClassificationEnsemble Predict | ClassificationSVM Predict | RegressionTree Predict

Objects

`ClassificationTree` | `CompactClassificationTree`

Functions
`fitctree` | `predict`

Topics
“Predict Class Labels Using ClassificationSVM Predict Block” on page 32-111
“Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130
“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2021a

ClassificationTreeCoderConfigurer

Coder configurer of binary decision tree model for multiclass classification

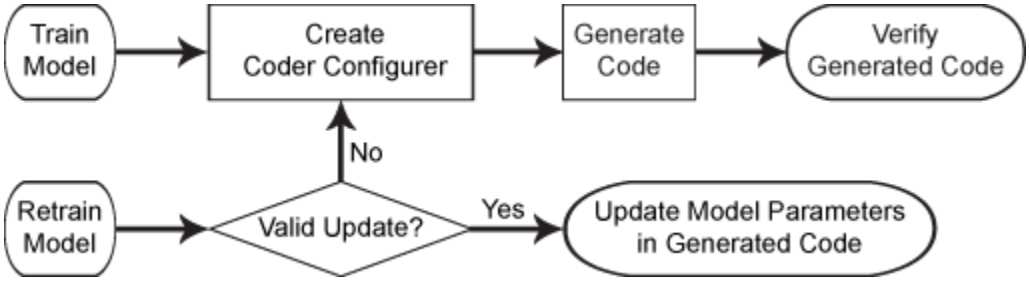
Description

A `ClassificationTreeCoderConfigurer` object is a coder configurer of a binary decision tree model for multiclass classification (`ClassificationTree` or `CompactClassificationTree`).

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes of the tree model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the classification tree model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the tree model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of a classification tree model, see the Code Generation sections of `CompactClassificationTree`, `predict`, and `update`.

Creation

After training a classification tree model by using `fitctree`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of the `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data
`LearnerCoderInput` object

Coder attributes of the predictor data to pass to the generated C/C++ code for the `predict` function of the classification tree model, specified as a `LearnerCoderInput` on page 33-612 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations of three predictor variables, specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

NumOutputs — Number of outputs in `predict`
1 (default) | 2 | 3 | 4

Number of output arguments to return from the generated C/C++ code for the `predict` function of the classification tree model, specified as 1, 2, 3, or 4.

The output arguments of `predict` are `label` (predicted class labels), `score` (posterior probabilities), `node` (node numbers for predicted classes), and `cnum` (class numbers of predicted labels), in that order. `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function, where `n` is the `NumOutputs` value.

After creating the coder configurer `configurer`, you can specify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of a classification tree model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-612 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Children` — Coder attributes of child nodes for each node

`LearnerCoderInput` object

Coder attributes of the child nodes for each node in the tree (`Children` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 2]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `ClassProbability`, `CutPoint`, and `CutPredictorIndex`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`ClassProbability` — Coder attributes of class probabilities for each node

`LearnerCoderInput` object

Coder attributes of the class probabilities for each node in the tree (`ClassProbability` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd c]`, where `nd` is the number of nodes in `Mdl` and `c` is the number of classes.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `CutPoint`, and `CutPredictorIndex`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`Cost` — Coder attributes of misclassification cost
`LearnerCoderInput` object

Coder attributes of the misclassification cost (`Cost` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[c c]`, where `c` is the number of classes.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

`CutPoint` — Coder attributes of cut point for each node
`LearnerCoderInput` object

Coder attributes of the cut point for each node in the tree (`CutPoint` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 1]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.

  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `ClassProbability`, and `CutPredictorIndex`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`CutPredictorIndex` — Coder attributes of cut predictor index for each node
`LearnerCoderInput` object

Coder attributes of the cut predictor index for each node in the tree (`CutPredictorIndex` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 1]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `ClassProbability`, and `CutPoint`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`Prior` — Coder attributes of prior probabilities
`LearnerCoderInput` object

Coder attributes of the prior probabilities (`Prior` of a classification tree model), specified as a `LearnerCoderInput` on page 33-612 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 c]`, where `c` is the number of classes.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.

- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code
`'ClassificationTreeModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `ClassificationTreeCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level
`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — Input argument of `predict`
cell array of a `coder.PrimitiveType` object

This property is read-only.

Input argument of the entry-point function `predict.m` for code generation, specified as a cell array of a `coder.PrimitiveType` object. The `coder.PrimitiveType` object includes the coder attributes of the predictor data stored in the `X` property.

If you modify the coder attributes of the predictor data, then the software updates the `coder.PrimitiveType` object accordingly.

The `coder.PrimitiveType` object in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each

`coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-601 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `fisheriris` data set, which contains flower data, and train a decision tree model.

```
load fisheriris
X = meas;
Y = species;
Mdl = fitctree(X,Y);
```

`Mdl` is a `ClassificationTree` object.

Create a coder configurer for the `ClassificationTree` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  ClassificationTreeCoderConfigurer with properties:

   Update Inputs:
             Children: [1x1 LearnerCoderInput]
    ClassProbability: [1x1 LearnerCoderInput]
             CutPoint: [1x1 LearnerCoderInput]
     CutPredictorIndex: [1x1 LearnerCoderInput]
                 Prior: [1x1 LearnerCoderInput]
                  Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                     X: [1x1 LearnerCoderInput]
```

```
   Code Generation Parameters:
           NumOutputs: 1
       OutputFileName: 'ClassificationTreeModel'


  Properties, Methods
```

`configurer` is a `ClassificationTreeCoderConfigurer` object, which is a coder configurer of a `ClassificationTree` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the classification tree model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationTreeModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:51
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:51
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 13:59:51
coder.inline('always')
persistent model
if isempty(model)
```

```
    model = loadLearnerForCoder('ClassificationTreeModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Children
        %                       ClassProbability
        %                       CutPoint
        %                       CutPredictorIndex
        %                       Prior
        %                       Cost
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of Classification Tree Model in Generated Code

Train a decision tree for multiclass classification using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Load the `fisheriris` data set, which contains flower data. This data set has four predictors: the sepal length, sepal width, petal length, and petal width of the flowers. The response variable contains the flower species names: setosa, versicolor, and virginica. Train a classification tree model using half of the observations.

```
load fisheriris
X = meas;
Y = species;

rng('default') % For reproducibility
n = length(Y);
c = cvpartition(Y,'HoldOut',0.5);
idxTrain = training(c,1);
XTrain = X(idxTrain,:);
YTrain = Y(idxTrain);

Mdl = fitctree(XTrain,YTrain);
```

`Mdl` is a `ClassificationTree` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationTree` model by using `learnerCoderConfigurer`. Specify the predictor data. The `learnerCoderConfigurer` function uses the input `XTrain` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 4 so that the generated code returns predicted labels, scores, node numbers, and class numbers.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'NumOutputs',4);
```

`configurer` is a `ClassificationTreeCoderConfigurer` object, which is a coder configurer of a `ClassificationTree` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the classification tree model parameters so that you can update the parameters in the generated code after retraining the model.

First, specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 4];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

The size of the first dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 4 predictors, the value of the `SizeVector` attribute must be 4 and the value of the `VariableDimensions` attribute must be `0`.

If you retrain the tree model using new data or different settings, the number of nodes in the tree can vary. Therefore, specify the first dimension of the `SizeVector` attribute of one of these properties so that you can update the number of nodes in the generated code: `Children`, `ClassProbability`, `CutPoint`, or `CutPredictorIndex`. The software then modifies the other properties automatically.

For example, set the first value of the `SizeVector` attribute of the `CutPoint` property to `Inf`. The software modifies the `SizeVector` and `VariableDimensions` attributes of `Children`, `ClassProbability`, and `CutPredictorIndex` to match the new upper bound on the number of nodes in the tree. Additionally, the first value of the `VariableDimensions` attribute of `CutPoint` changes to `1`.

```
configurer.CutPoint.SizeVector = [Inf 1];
```

```
SizeVector attribute for Children has been modified to satisfy configuration constraints.
SizeVector attribute for CutPredictorIndex has been modified to satisfy configuration constraints
```

```
VariableDimensions attribute for Children has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPredictorIndex has been modified to satisfy configuration co
SizeVector attribute for ClassProbability has been modified to satisfy configuration constraints
VariableDimensions attribute for ClassProbability has been modified to satisfy configuration cons
```

```
configurer.CutPoint.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the classification tree model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationTreeModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same output arguments. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score,node,cnum] = predict(Mdl,XTrain);
[label_mex,score_mex,node_mex,cnum_mex] = ClassificationTreeModel('predict',XTrain);
```

Compare `label` and `label_mex` by using `isequal`. Similarly, compare `node` to `node_mex` and `cnum` to `cnum_mex`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
isequal(node,node_mex)
```

```
ans = logical
   1
```

```
isequal(cnum,cnum_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the input arguments are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels, node numbers, and class numbers.

Compare `score` and `score_mex`.

```
max(abs(score-score_mex),[],'all')
```

```
ans = 0
```

In general, `score_mex` might include round-off differences compared to `score`. In this case, the comparison confirms that `score` and `score_mex` are equal.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitctree(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationTreeModel('update',params)
```

Verify Generated Code

Compare the output arguments from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score,node,cnum] = predict(retrainedMdl,X);
[label_mex,score_mex,node_mex,cnum_mex] = ClassificationTreeModel('predict',X);
```

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
isequal(node,node_mex)
```

```
ans = logical
   1
```

```
isequal(cnum,cnum_mex)
```

```
ans = logical
   1
```

```
max(abs(score-score_mex),[],'all')
```

```
ans = 0
```

The comparison confirms that the labels, node numbers, class numbers, and scores are equal.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of the `CutPoint` property of the coder configurer `configurer`:

```
configurer.CutPoint.SizeVector = [20 1];
configurer.CutPoint.VariableDimensions = [1 0];
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

See Also

`ClassificationTree` | `CompactClassificationTree` | `learnerCoderConfigurer` | `predict` | `update`

Topics
“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2019b

classify

Discriminant analysis

Syntax

```
class = classify(sample,training,group)
class = classify(sample,training,group,'type')
class = classify(sample,training,group,'type',prior)
[class,err] = classify(...)
[class,err,POSTERIOR] = classify(...)
[class,err,POSTERIOR,logp] = classify(...)
[class,err,POSTERIOR,logp,coeff] = classify(...)
```

Description

`class = classify(sample,training,group)` classifies each row of the data in `sample` into one of the groups in `training`. `sample` and `training` must be matrices with the same number of columns. `group` is a grouping variable for `training`. Its unique values define groups; each element defines the group to which the corresponding row of `training` belongs. `group` can be a categorical variable, a numeric vector, a character array, a string array, or a cell array of character vectors. `training` and `group` must have the same number of rows. `classify` treats `<undefined>` values, `NaN`s, empty character vectors, empty strings, and `<missing>` string values in `group` as missing data values, and ignores the corresponding rows of `training`. The output `class` indicates the group to which each row of `sample` has been assigned, and is of the same type as `group`.

`class = classify(sample,training,group,'type')` allows you to specify the type of discriminant function. Specify *`type`* inside single quotes. *`type`* is one of:

- `linear` — Fits a multivariate normal density to each group, with a pooled estimate of covariance. This is the default.
- `diaglinear` — Similar to `linear`, but with a diagonal covariance matrix estimate (naive Bayes classifiers).
- `quadratic` — Fits multivariate normal densities with covariance estimates stratified by group.
- `diagquadratic` — Similar to `quadratic`, but with a diagonal covariance matrix estimate (naive Bayes classifiers).
- `mahalanobis` — Uses Mahalanobis distances with stratified covariance estimates.

`class = classify(sample,training,group,'type',prior)` allows you to specify prior probabilities for the groups. *`prior`* is one of:

- A numeric vector the same length as the number of unique values in `group` (or the number of levels defined for `group`, if `group` is categorical). If `group` is numeric or categorical, the order of *`prior`* must correspond to the ordered values in `group`. Otherwise, the order of *`prior`* must correspond to the order of the first occurrence of the values in `group`.
- A 1-by-1 structure with fields:
  - `prob` — A numeric vector.

- `group` — Of the same type as `group`, containing unique values indicating the groups to which the elements of `prob` correspond.

  As a structure, *`prior`* can contain groups that do not appear in `group`. This can be useful if `training` is a subset a larger training set. `classify` ignores any groups that appear in the structure but not in the `group` array.
- The character vector or string scalar `'empirical'`, indicating that group prior probabilities should be estimated from the group relative frequencies in `training`.

`prior` defaults to a numeric vector of equal probabilities, i.e., a uniform distribution. *`prior`* is not used for discrimination by Mahalanobis distance, except for error rate calculation.

`[class,err] = classify(...)` also returns an estimate `err` of the misclassification error rate based on the `training` data. `classify` returns the apparent error rate, i.e., the percentage of observations in `training` that are misclassified, weighted by the prior probabilities for the groups.

`[class,err,POSTERIOR] = classify(...)` also returns a matrix `POSTERIOR` of estimates of the posterior probabilities that the *j*th training group was the source of the *i*th sample observation, i.e., *Pr*(*group j*|*obs i*). `POSTERIOR` is not computed for Mahalanobis discrimination.

`[class,err,POSTERIOR,logp] = classify(...)` also returns a vector `logp` containing estimates of the logarithms of the unconditional predictive probability density of the sample observations, $p(obs\ i) = \sum p(obs\ i|group\ j)Pr(group\ j)$ over all groups. `logp` is not computed for Mahalanobis discrimination.

`[class,err,POSTERIOR,logp,coeff] = classify(...)` also returns a structure array `coeff` containing coefficients of the boundary curves between pairs of groups. Each element `coeff(I,J)` contains information for comparing group `I` to group `J` in the following fields:

- `type` — Type of discriminant function, from the *`type`* input.
- `name1` — Name of the first group.
- `name2` — Name of the second group.
- `const` — Constant term of the boundary equation (K)
- `linear` — Linear coefficients of the boundary equation (L)
- `quadratic` — Quadratic coefficient matrix of the boundary equation (Q)

For the `linear` and `diaglinear` types, the `quadratic` field is absent, and a row `x` from the `sample` array is classified into group `I` rather than group `J` if `0 < K+x*L`. For the other types, `x` is classified into group `I` if `0 < K+x*L+x*Q*x'`.

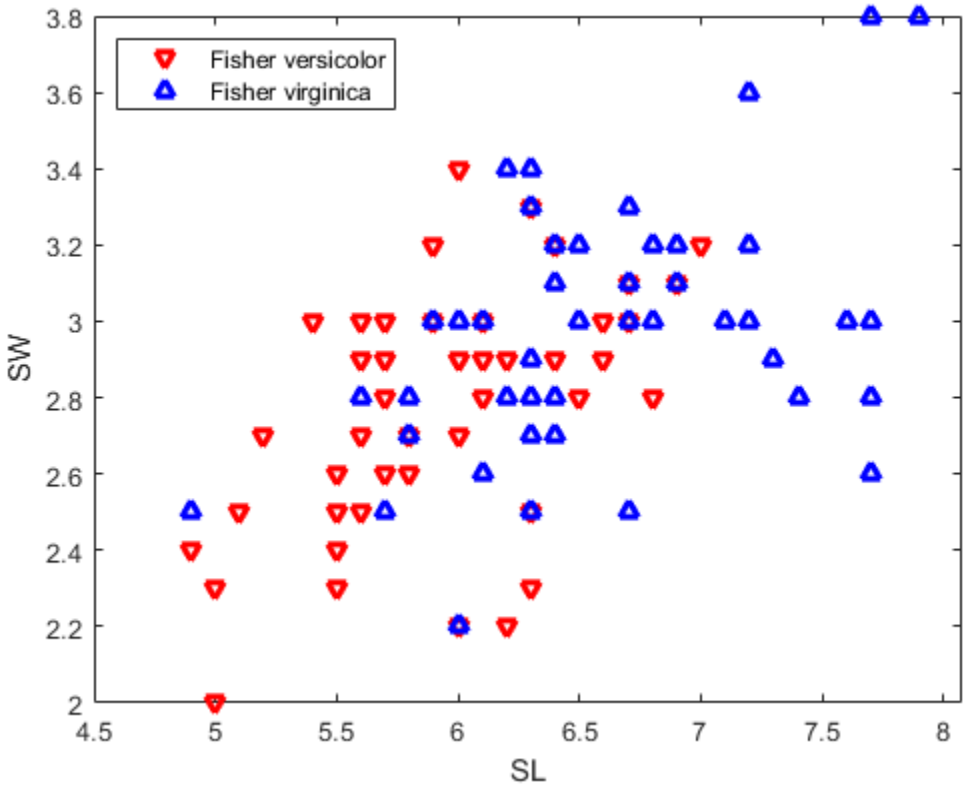
Examples

Classify Using Discriminant Analysis

For training data, use Fisher's sepal measurements for iris versicolor and virginica:

```
load fisheriris
SL = meas(51:end,1);
SW = meas(51:end,2);
group = species(51:end);
h1 = gscatter(SL,SW,group,'rb','v^',[],'off');
```

```
set(h1,'LineWidth',2)
legend('Fisher versicolor','Fisher virginica',...
       'Location','NW')
```

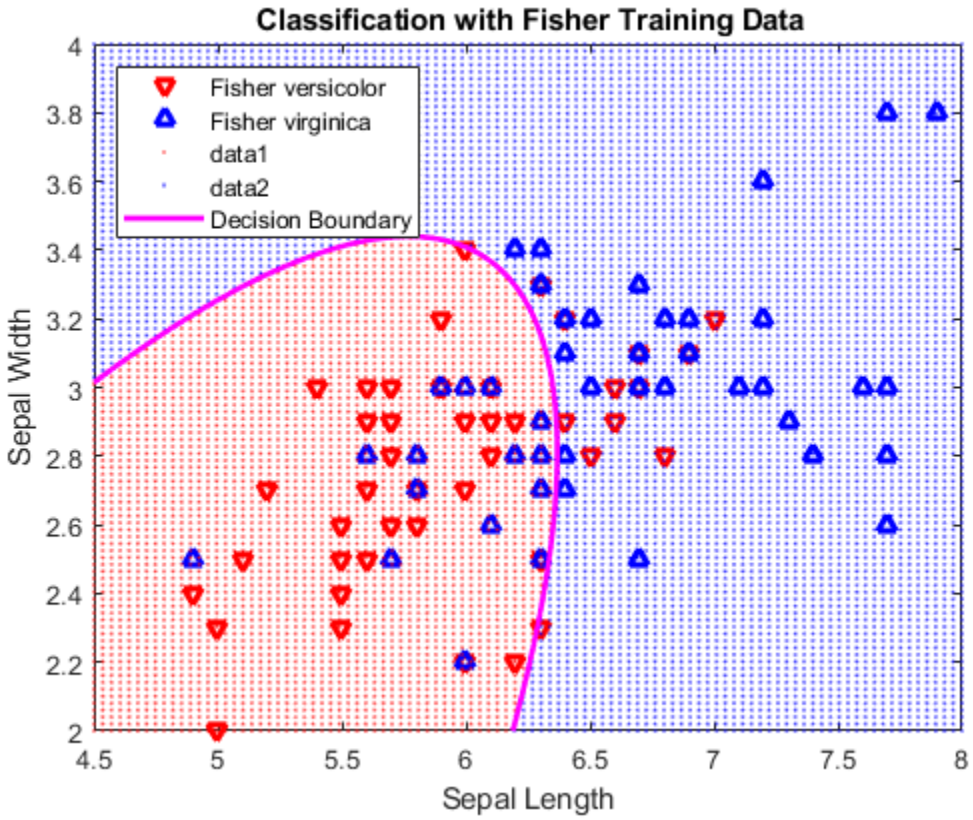


Classify a grid of measurements on the same scale:

```
[X,Y] = meshgrid(linspace(4.5,8),linspace(2,4));
X = X(:); Y = Y(:);
[C,err,P,logp,coeff] = classify([X Y],[SL SW],...
                                 group,'Quadratic');
```

Visualize the classification:

```
hold on;
gscatter(X,Y,C,'rb','.',1,'off');
K = coeff(1,2).const;
L = coeff(1,2).linear;
Q = coeff(1,2).quadratic;
% Function to compute K + L*v + v'*Q*v for multiple vectors
% v=[x;y]. Accepts x and y as scalars or column vectors.
f = @(x,y) K + L(1)*x + L(2)*y + Q(1,1)*x.*x + (Q(1,2)+Q(2,1))*x.*y + Q(2,2)*y.*y;
h2 = fimplicit(f,[4.5 8 2 4]);
set(h2,'Color','m','LineWidth',2,'DisplayName','Decision Boundary')
axis([4.5 8 2 4])
xlabel('Sepal Length')
ylabel('Sepal Width')
title('{\bf Classification with Fisher Training Data}')
```



Alternative Functionality

The `fitcdiscr` function also performs discriminant analysis. You can train a classifier by using the `fitcdiscr` function and predict labels of new data by using the `predict` function. The `fitcdiscr` supports cross-validation and hyperparameter optimization and does not require you to fit the classifier every time you make a new prediction or you change prior probabilities.

References

[1] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[2] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

See Also

`ClassificationDiscriminant` | `fitcdiscr` | `fitcnb` | `fitctree` | `mahal` | `predict`

Topics

“Discriminant Analysis Classification” on page 20-2
“Grouping Variables” on page 2-45

Introduced before R2006a

cluster

Construct agglomerative clusters from linkages

Syntax

```
T = cluster(Z,'Cutoff',C)
T = cluster(Z,'Cutoff',C,'Depth',D)
T = cluster(Z,'Cutoff',C,'Criterion',criterion)

T = cluster(Z,'MaxClust',N)
```

Description

`T = cluster(Z,'Cutoff',C)` defines clusters from an agglomerative hierarchical cluster tree `Z`. The input `Z` is the output of the `linkage` function for an input data matrix `X`. `cluster` cuts `Z` into clusters, using `C` as a threshold for the inconsistency coefficients (or `inconsistent` values) of nodes in the tree. The output `T` contains cluster assignments of each observation (row of `X`).

`T = cluster(Z,'Cutoff',C,'Depth',D)` evaluates `inconsistent` values by looking to a depth `D` below each node.

`T = cluster(Z,'Cutoff',C,'Criterion',criterion)` uses either `'inconsistent'` (default) or `'distance'` as the `criterion` for defining clusters. `criterion` must be less than `C` for `cluster` to define clusters.

`T = cluster(Z,'MaxClust',N)` defines a maximum of `N` clusters using `'distance'` as the criterion for defining clusters.

Examples

Define Clusters by Specifying Depth

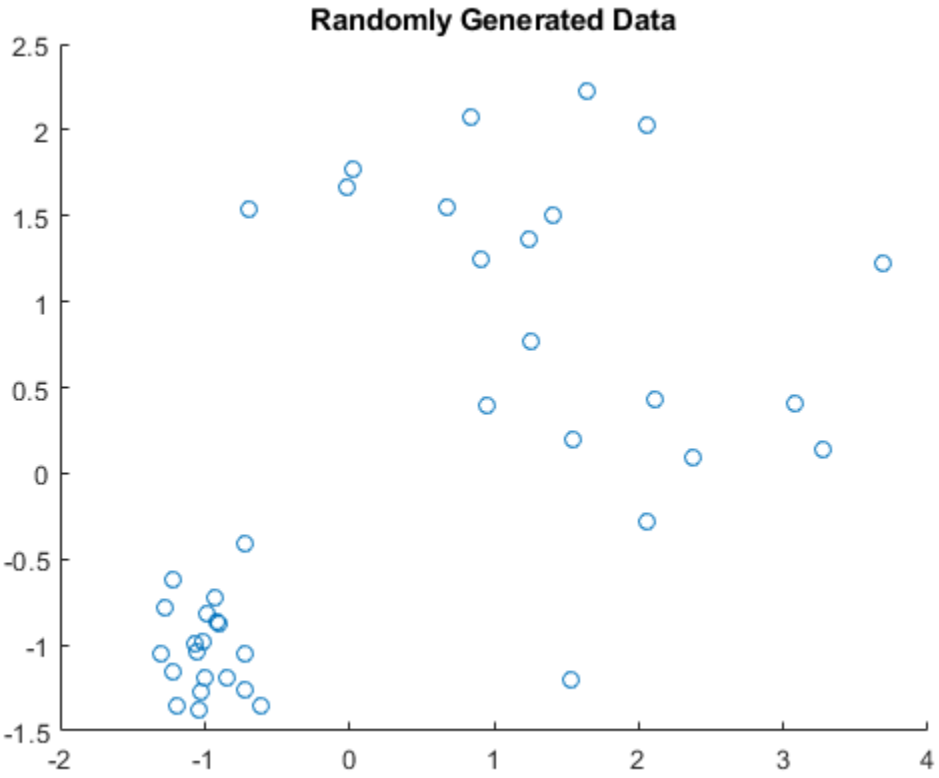
Perform agglomerative clustering on randomly generated data by evaluating inconsistent values to a depth of four below each node.

Randomly generate the sample data.

```
rng('default'); % For reproducibility
X = [(randn(20,2)*0.75)+1;
    (randn(20,2)*0.25)-1];
```

Create a scatter plot of the data.

```
scatter(X(:,1),X(:,2));
title('Randomly Generated Data');
```

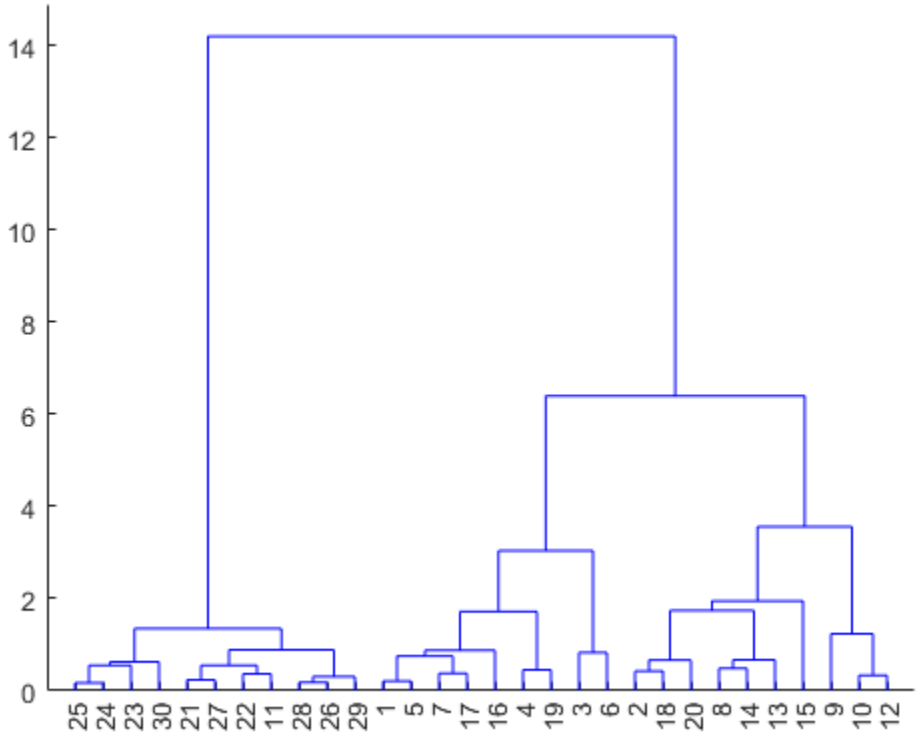


Create a hierarchical cluster tree using the `ward` linkage method.

```
Z = linkage(X,'ward');
```

Create a dendrogram plot of the data.

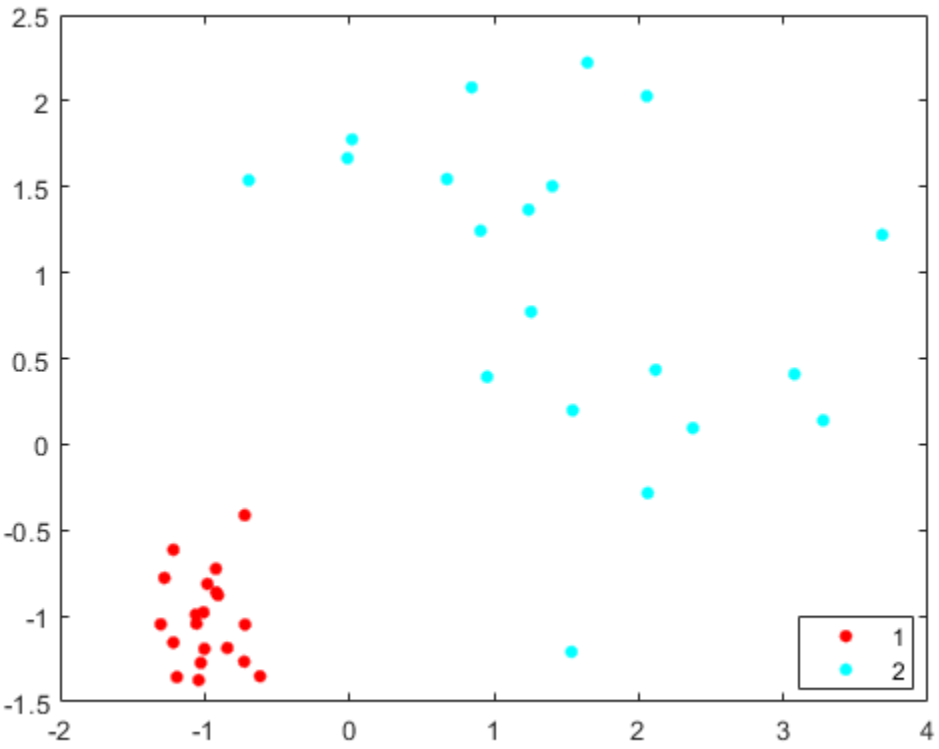
```
dendrogram(Z)
```



The scatter plot and the dendrogram plot seem to show two clusters in the data.

Cluster the data using a threshold of 3 for the inconsistency coefficient and looking to a depth of 4 below each node. Plot the resulting clusters.

```
T = cluster(Z,'cutoff',3,'Depth',4);
gscatter(X(:,1),X(:,2),T)
```



`cluster` identifies two clusters in the data.

Cluster Data Using Distance Criterion

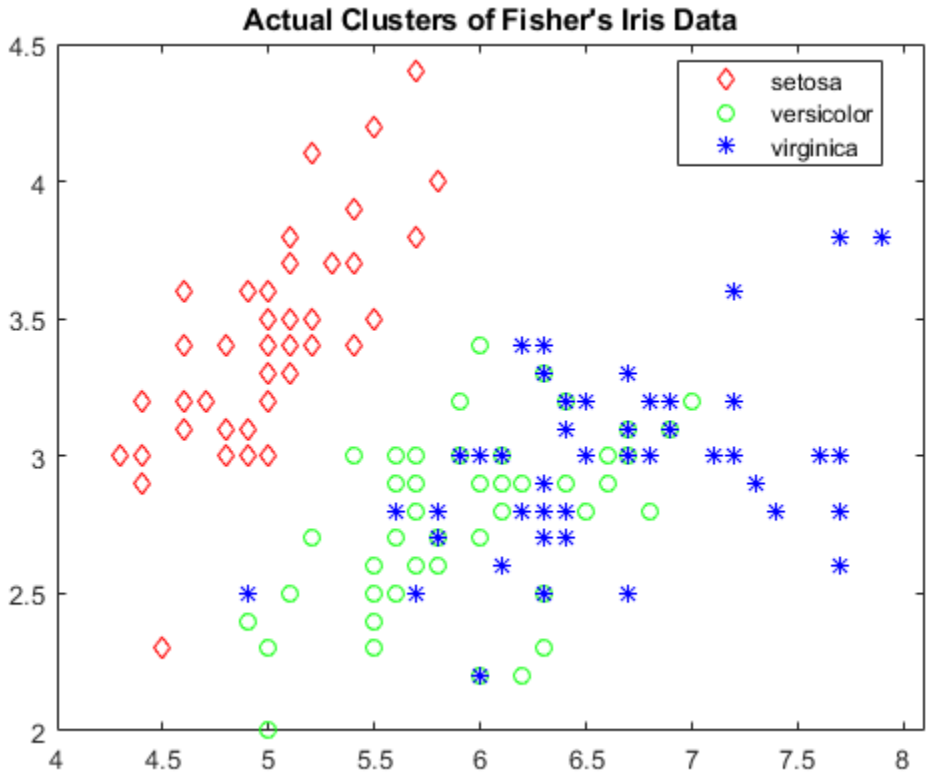
Perform agglomerative clustering on the `fisheriris` data set using `'distance'` as the criterion for defining clusters. Visualize the cluster assignments of the data.

Load the `fisheriris` data set.

```
load fisheriris
```

Visualize a 2-D scatter plot of the data using species as the grouping variable. Specify marker colors and marker symbols for the three different species.

```
gscatter(meas(:,1),meas(:,2),species,'rgb','do*')
title("Actual Clusters of Fisher's Iris Data")
```



Create a hierarchical cluster tree using the `'average'` method and the `'chebychev'` metric.

```
Z = linkage(meas,'average','chebychev');
```

Cluster the data using a threshold of 1.5 for the `'distance'` criterion.

```
T = cluster(Z,'cutoff',1.5,'Criterion','distance')
```

```
T = 150×1

     2
     2
     2
     2
     2
     2
     2
     2
     2
     2
     ⋮
```

`T` contains numbers that correspond to the cluster assignments. Find the number of classes that `cluster` identifies.

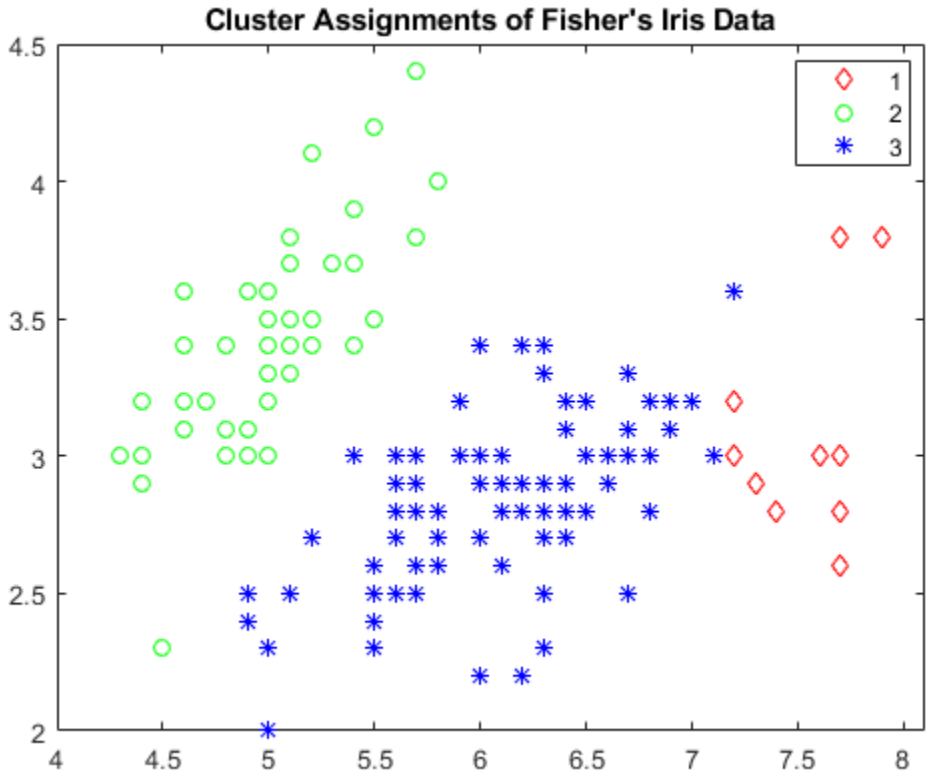
```
length(unique(T))
```

```
ans = 3
```

`cluster` identifies three classes for the specified values of `cutoff` and `Criterion`.

Visualize a 2-D scatter plot of the clustering results using `T` as the grouping variable. Specify marker colors and marker symbols for the three different classes.

```
gscatter(meas(:,1),meas(:,2),T,'rgb','do*')
title("Cluster Assignments of Fisher's Iris Data")
```



Clustering correctly identifies the setosa class (class 2) as belonging to a distinct cluster, but poorly distinguishes between the versicolor and virginica classes (classes 1 and 3, respectively). Note that the scatter plot labels the classes using the numbers contained in `T`.

Compare Cluster Assignments to Classes

Find a maximum of three clusters in the `fisheriris` data set and compare cluster assignments of the flowers to their known classification.

Load the sample data.

```
load fisheriris
```

Create a hierarchical cluster tree using the `'average'` method and the `'chebychev'` metric.

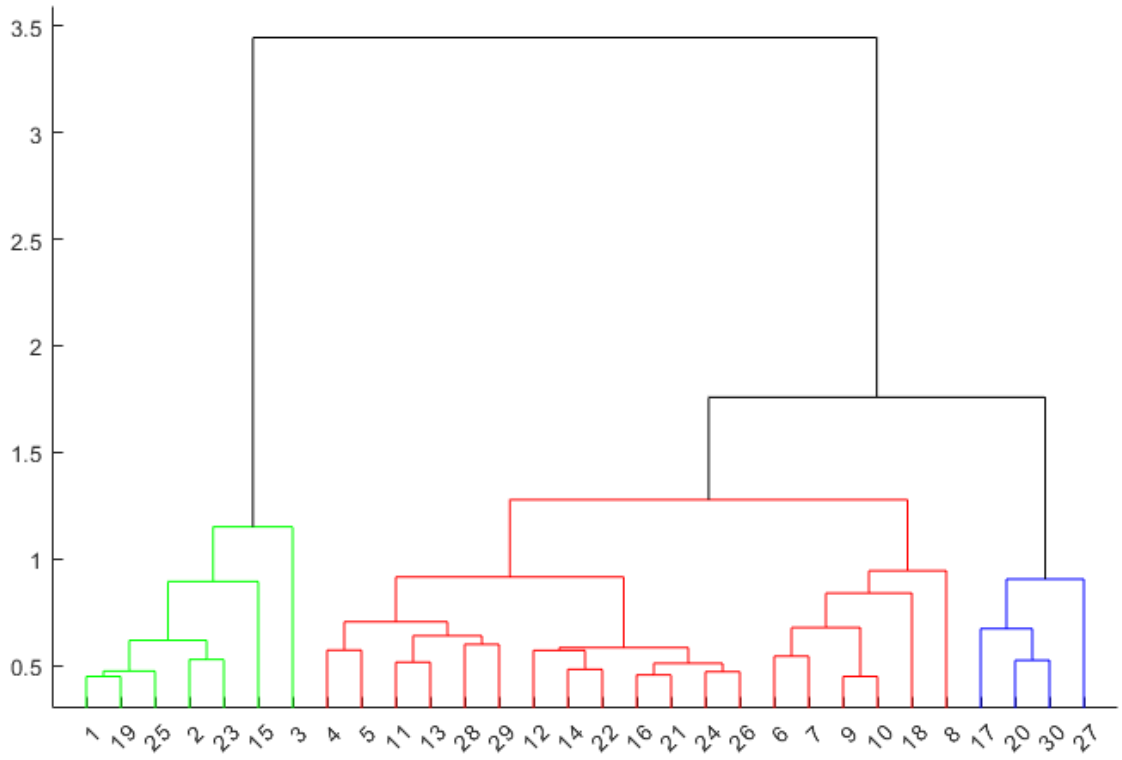
```
Z = linkage(meas,'average','chebychev');
```

Find a maximum of three clusters in the data.

```
T = cluster(Z,'maxclust',3);
```

Create a dendrogram plot of `Z`. To see the three clusters, use `'ColorThreshold'` with a cutoff halfway between the third-from-last and second-from-last linkages.

```
cutoff = median([Z(end-2,3) Z(end-1,3)]);
dendrogram(Z,'ColorThreshold',cutoff)
```



Display the last two rows of `Z` to see how the three clusters are combined into one. `linkage` combines the 293rd (blue) cluster with the 297th (red) cluster to form the 298th cluster with a linkage of `1.7583`. `linkage` then combines the 296th (green) cluster with the 298th cluster.

```
lastTwo = Z(end-1:end,:)
```

```
lastTwo = 2×3

  293.0000  297.0000    1.7583
  296.0000  298.0000    3.4445
```

See how the cluster assignments correspond to the three species. For example, one of the clusters contains `50` flowers of the second species and `40` flowers of the third species.

```
crosstab(T,species)
```

```
ans = 3×3

     0     0    10
     0    50    40
    50     0     0
```

Cluster Data and Plot Result

Randomly generate sample data with 20,000 observations.

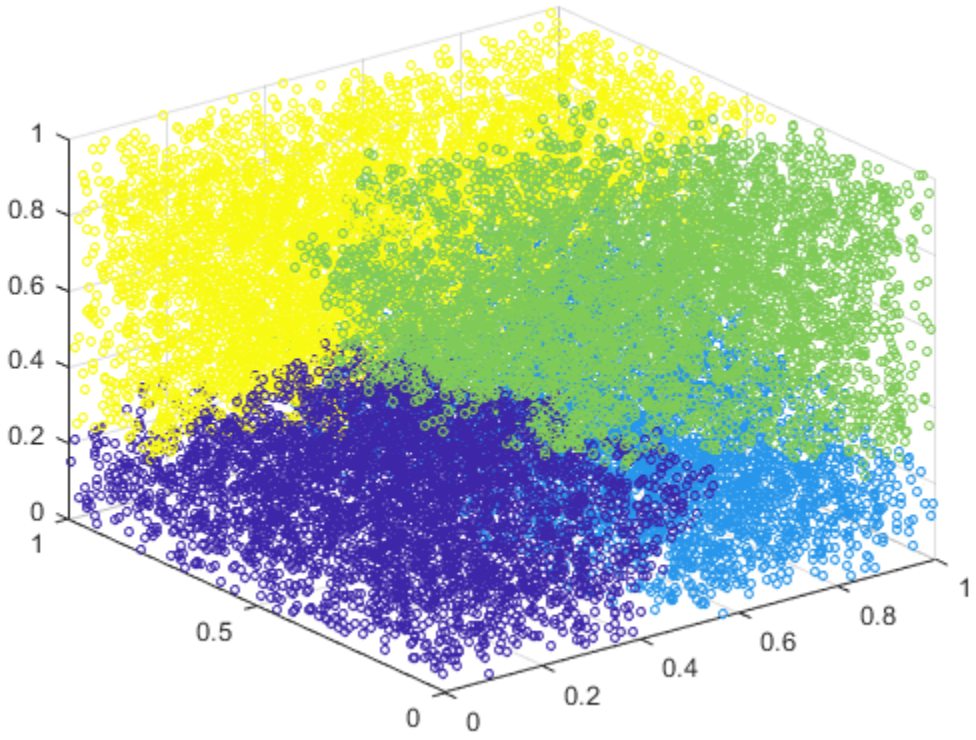
```
rng('default') % For reproducibility
X = rand(20000,3);
```

Create a hierarchical cluster tree using the `ward` linkage method. In this case, the `'SaveMemory'` option of the `clusterdata` function is set to `'on'` by default. In general, specify the best value for `'SaveMemory'` based on the dimensions of `X` and the available memory.

```
Z = linkage(X,'ward');
```

Cluster the data into a maximum of four groups and plot the result.

```
c = cluster(Z,'Maxclust',4);
scatter3(X(:,1),X(:,2),X(:,3),10,c)
```



`cluster` identifies four groups in the data.

Input Arguments

`Z` — Agglomerative hierarchical cluster tree

numeric matrix

Agglomerative hierarchical cluster tree that is the output of the `linkage` function, specified as a numeric matrix. For an input data matrix `X` with *m* rows (or observations), `linkage` returns an (*m* – 1)-by-3 matrix `Z`. For an explanation of how `linkage` creates the cluster tree, see `Z`.

Example: `Z = linkage(X)`, where `X` is an input data matrix

Data Types: `single` | `double`

`C` — Threshold for defining clusters

positive scalar | vector of positive scalars

Threshold for defining clusters, specified as a positive scalar or a vector of positive scalars. `cluster` uses `C` as a threshold for either the heights or the inconsistency coefficients of nodes, depending on the `criterion` for defining clusters in a hierarchical cluster tree.

- If the criterion for defining clusters is `'distance'`, then `cluster` groups all leaves at or below a node into a cluster, provided that the height of the node is less than `C`.
- If the criterion for defining clusters is `'inconsistent'`, then the `inconsistent` values of a node and all its subnodes must be less than `C` for `cluster` to group them into a cluster. `cluster` begins from the root of the cluster tree `Z` and steps down through the tree until it encounters a node whose `inconsistent` value is less than the threshold `C`, and whose subnodes (or descendants) have inconsistent values less than `C`. Then `cluster` groups all leaves at or below the node into a cluster (or a singleton if the node itself is a leaf). `cluster` follows every branch in the tree until all leaf nodes are in clusters.

Example: `cluster(Z,'Cutoff',0.5)`

Data Types: `single` | `double`

`D` — Depth for computing inconsistent values

2 (default) | numeric scalar

Depth for computing inconsistent values, specified as a numeric scalar. `cluster` evaluates inconsistent values by looking to a depth `D` below each node.

Example: `cluster(Z,'Cutoff',0.5,'Depth',3)`

Data Types: `single` | `double`

`criterion` — Criterion for defining clusters

`'inconsistent'` (default) | `'distance'`

Criterion for defining clusters, specified as `'inconsistent'` or `'distance'`.

If the criterion for defining clusters is `'distance'`, then `cluster` groups all leaves at or below a node into a cluster (or a singleton if the node itself is a leaf), provided that the height of the node is less than `C`. The height of a node in a tree represents the distance between the two subnodes that are merged at that node. Specifying `'distance'` results in clusters that correspond to a horizontal slice of the `dendrogram` plot of `Z`.

If the criterion for defining clusters is `'inconsistent'`, then `cluster` groups a node and all its subnodes into a cluster, provided that the inconsistency coefficients (or `inconsistent` values) of the node and subnodes are less than `C`. Specifying `'inconsistent'` is equivalent to `cluster(Z,'Cutoff',C)`.

Example: `cluster(Z,'Cutoff',0.5,'Criterion','distance')`

Data Types: `char` | `string`

`N` — Maximum number of clusters
positive integer | vector of positive integers

Maximum number of clusters to form, specified as a positive integer or a vector of positive integers. `cluster` constructs a maximum of `N` clusters, using `'distance'` as the criterion for defining clusters. The height of each node in the tree represents the distance between the two subnodes merged at that node. `cluster` finds the smallest height at which a horizontal cut through the tree will leave `N` or fewer clusters. See Specify Arbitrary Clusters on page 16-16 for more details.

Example: `cluster(Z,'MaxClust',5)`

Data Types: `single` | `double`

Output Arguments

`T` — Cluster assignment
numeric vector | numeric matrix

Cluster assignment, returned as a numeric vector or matrix. For the (*m* – 1)-by-3 hierarchical cluster tree `Z` (the output of `linkage` given input `X`), `T` contains the cluster assignments of the *m* rows (observations) of `X`.

The size of `T` depends on the corresponding size of `C` or `N`.

- If `C` is a positive scalar, then `T` is a vector of length *m*.
- If `N` is a positive integer, then `T` is a vector of length *m*.
- If `C` is a length *l* vector of positive scalars, then `T` is an *m*-by-*l* matrix with one column per value in `C`.
- If `N` is a length *l* vector of positive integers, then `T` is an *m*-by-*l* matrix with one column per value in `N`.

Alternative Functionality

If you have an input data matrix `X`, you can use `clusterdata` to perform agglomerative clustering and return cluster indices for each observation (row) in `X`. The `clusterdata` function performs all the necessary steps for you, so you do not need to execute the `pdist`, `linkage`, and `cluster` functions separately.

See Also

`clusterdata` | `cophenet` | `dendrogram` | `inconsistent` | `kmeans` | `linkage` | `pdist`

Topics
“Hierarchical Clustering” on page 16-6

Introduced before R2006a

cluster

Construct clusters from Gaussian mixture distribution

Syntax

```
idx = cluster(gm,X)
[idx,nlogL] = cluster(gm,X)
[idx,nlogL,P] = cluster(gm,X)
[idx,nlogL,P,logpdf] = cluster(gm,X)
[idx,nlogL,P,logpdf,d2] = cluster(gm,X)
```

Description

`idx = cluster(gm,X)` partitions the data in `X` into *k* clusters determined by the *k* Gaussian mixture components in `gm`. The value in `idx(i)` is the cluster index of observation `i` and indicates the component with the largest posterior probability given the observation `i`.

`[idx,nlogL] = cluster(gm,X)` also returns the negative loglikelihood of the Gaussian mixture model `gm` given the data `X`.

`[idx,nlogL,P] = cluster(gm,X)` also returns the posterior probabilities of each Gaussian mixture component in `gm` given each observation in `X`.

`[idx,nlogL,P,logpdf] = cluster(gm,X)` also returns a logarithm of the estimated probability density function (pdf) evaluated at each observation in `X`.

`[idx,nlogL,P,logpdf,d2] = cluster(gm,X)` also returns the squared Mahalanobis distance of each observation in `X` to each Gaussian mixture component in `gm`.

Examples

Cluster Data

Generate random variates that follow a mixture of two bivariate Gaussian distributions by using the `mvnrnd` function. Fit a Gaussian mixture model (GMM) to the generated data by using the `fitgmdist` function. Then, use the `cluster` function to partition the data into two clusters determined by the fitted GMM components.

Define the distribution parameters (means and covariances) of two bivariate Gaussian mixture components.

```
mu1 = [2 2];          % Mean of the 1st component
sigma1 = [2 0; 0 1];  % Covariance of the 1st component
mu2 = [-2 -1];        % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Covariance of the 2nd component
```

Generate an equal number of random variates from each component, and combine the two sets of random variates.

```
rng('default') % For reproducibility
r1 = mvnrnd(mu1,sigma1,1000);
r2 = mvnrnd(mu2,sigma2,1000);
X = [r1; r2];
```

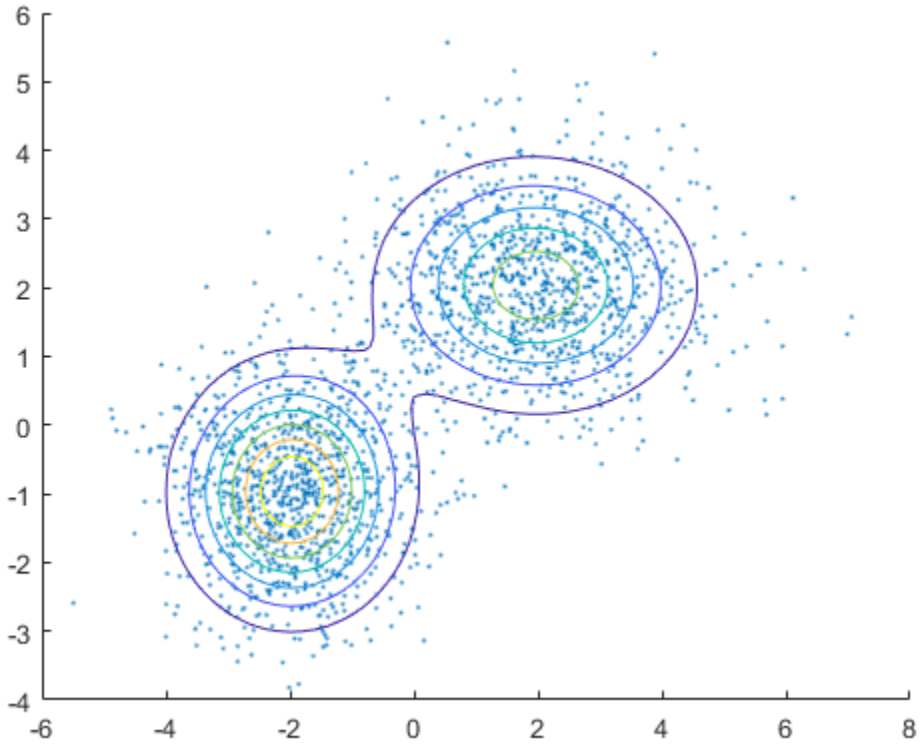
The combined data set X contains random variates following a mixture of two bivariate Gaussian distribution.

Fit a two-component GMM to X.

```
gm = fitgmdist(X,2);
```

Plot X by using `scatter`. Visualize the fitted model `gm` by using `pdf` and `fcontour`.

```
figure
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fcontour(gmPDF,[-6 8 -4 6])
```

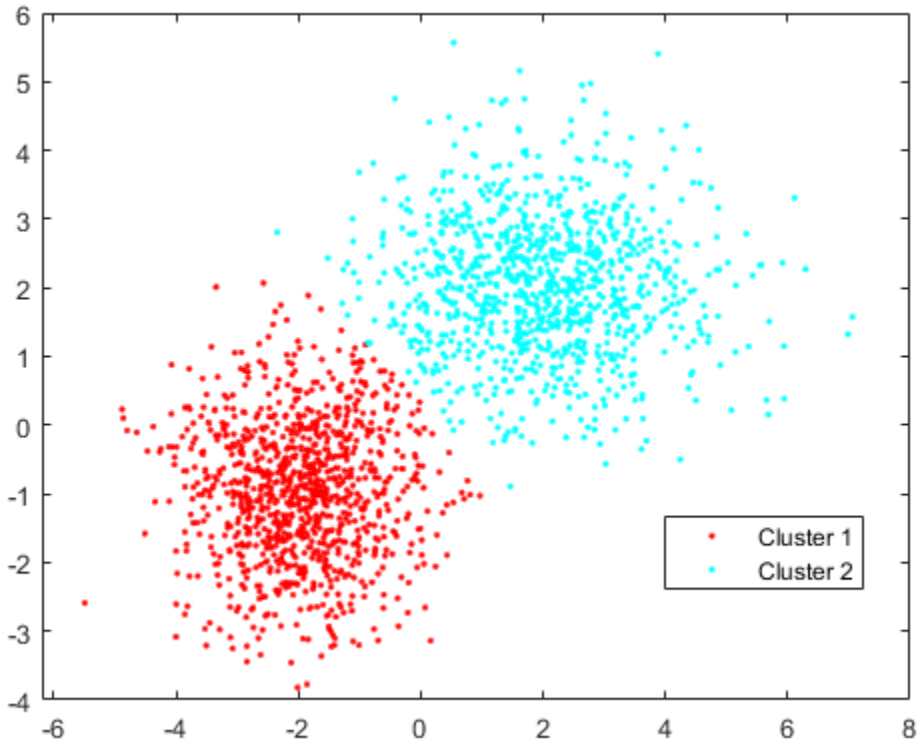


Partition the data into clusters by passing the fitted GMM and the data to `cluster`.

```
idx = cluster(gm,X);
```

Use `gscatter` to create a scatter plot grouped by `idx`.

```
figure;
gscatter(X(:,1),X(:,2),idx);
legend('Cluster 1','Cluster 2','Location','best');
```



Input Arguments

gm — Gaussian mixture distribution

`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a `gmdistribution` object using `gmdistribution` or `fitgmdist`. Use the `gmdistribution` function to create a `gmdistribution` object by specifying the distribution parameters. Use the `fitgmdist` function to fit a `gmdistribution` model to data given a fixed number of components.

X — Data

n-by-*m* numeric matrix

Data, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

To provide meaningful clustering results, `X` must come from the same population as the data used to create `gm`.

If a row of `X` contains `NaNs`, then `cluster` excludes the row from the computation. The corresponding value in `idx`, `P`, `logpdf`, and `d2` is `NaN`.

Data Types: `single` | `double`

Output Arguments

`idx` — Cluster index

n-by-1 positive integer vector

Cluster index, returned as an *n*-by-1 positive integer vector, where *n* is the number of observations in `X`.

`idx(i)` is the cluster index of observation `i` and indicates the Gaussian mixture component with the largest posterior probability given the observation `i`.

`nlogL` — Negative loglikelihood

numeric value

Negative loglikelihood value of the Gaussian mixture model `gm` given the data `X`, returned as a numeric value.

`P` — Posterior probability

n-by-*k* numeric vector

Posterior probability of each Gaussian mixture component in `gm` given each observation in `X`, returned as an *n*-by-*k* numeric vector, where *n* is the number of observations in `X` and *k* is the number of mixture components in `gm`.

`P(i,j)` is the posterior probability of the `j`th Gaussian mixture component given observation `i`, Probability(component `j` | observation `i`).

`logpdf` — Logarithm of estimated pdf

n-by-1 numeric vector

Logarithm of the estimated pdf, evaluated at each observation in `X`, returned as an *n*-by-1 numeric vector, where *n* is the number of observations in `X`.

`logpdf(i)` is the logarithm of the estimated pdf at observation `i`. The `cluster` function computes the estimated pdf by using the likelihood of each component given each observation and the component probabilities.

$$\text{logpdf}(i) = \log \sum_{j=1}^{k} L(C_j \big| O_i)\text{P}(\text{C}_j),$$

where $L(C_j|O_j)$ is the likelihood of component `j` given observation `i`, and $P(C_j)$ is the probability of component `j`. The `cluster` function computes the likelihood term by using the multivariate normal pdf of the `j`th Gaussian mixture component evaluated at observation `i`. The component probabilities are the mixing proportions of mixture components, the `ComponentProportion` property of `gm`.

`d2` — Squared Mahalanobis distance

n-by-*k* numeric matrix

Squared Mahalanobis distance of each observation in `X` to each Gaussian mixture component in `gm`, returned as an *n*-by-*k* numeric matrix, where *n* is the number of observations in `X` and *k* is the number of mixture components in `gm`.

`d2(i,j)` is the squared distance of observation `i` to the `j`th Gaussian mixture component.

See Also

`fitgmdist` | `gmdistribution` | `mahal` | `posterior`

Topics

"Cluster Using Gaussian Mixture Model" on page 16-39
"Cluster Gaussian Mixture Data Using Hard Clustering" on page 16-46
"Cluster Gaussian Mixture Data Using Soft Clustering" on page 16-52

Introduced in R2007b

ClusterCriterion

Package: `clustering.evaluation`

Clustering evaluation object

Description

Create a clustering evaluation object using `evalclusters`.

Properties

`ClusteringFunction`

Clustering algorithm used to cluster the input data, stored as a valid clustering algorithm name or function handle. If the clustering solutions are provided in the input, `ClusteringFunction` is empty.

`CriterionName`

Name of the criterion used for clustering evaluation, stored as a valid criterion name.

`CriterionValues`

Criterion values corresponding to each proposed number of clusters in `InspectedK`, stored as a vector of numerical values.

`InspectedK`

List of the number of proposed clusters for which to compute criterion values, stored as a vector of positive integer values.

`Missing`

Logical flag for excluded data, stored as a column vector of logical values. If `Missing` equals `true`, then the corresponding value in the data matrix `x` is not used in the clustering solution.

`NumObservations`

Number of observations in the data matrix `X`, minus the number of missing (`NaN`) values in `X`, stored as a positive integer value.

`OptimalK`

Optimal number of clusters, stored as a positive integer value.

`OptimalY`

Optimal clustering solution corresponding to `OptimalK`, stored as a column vector of positive integer values. If the clustering solutions are provided in the input, `OptimalY` is empty.

X

Data used for clustering, stored as a matrix of numerical values.

Methods

addK	Evaluate additional numbers of clusters
plot	Plot clustering evaluation object criterion values
compact	Compact clustering evaluation object

See Also

`CalinskiHarabaszEvaluation` | `DaviesBouldinEvaluation` | `GapEvaluation` | `SilhouetteEvaluation` | `evalclusters`

Topics
Class Attributes
Property Attributes

clusterdata

Construct agglomerative clusters from data

Syntax

```
T = clusterdata(X,cutoff)

T = clusterdata(X,Name,Value)
```

Description

`T = clusterdata(X,cutoff)` returns cluster indices for each observation (row) of an input data matrix `X`, given a threshold `cutoff` for cutting an agglomerative hierarchical tree that the `linkage` function generates from `X`.

`clusterdata` supports agglomerative clustering and incorporates the `pdist`, `linkage`, and `cluster` functions, which you can use separately for more detailed analysis. See Algorithm Description on page 16-6 for more details.

`T = clusterdata(X,Name,Value)` specifies clustering options using one or more name-value pair arguments. You must specify either `Cutoff` or `MaxClust`. For example, specify `'MaxClust',5` to find a maximum of five clusters.

Examples

Find Limited Number of Clusters from Sample Data

Find and visualize a maximum of three clusters in a randomly generated data set using two different approaches:

1. Specify a value for the `cutoff` input argument.
2. Specify a value for the `'MaxClust'` name-value pair argument.

Create a sample data set consisting of randomly generated data from three standard uniform distributions.

```
rng('default');  % For reproducibility
X = [gallery('uniformdata',[10 3],12); ...
    gallery('uniformdata',[10 3],13)+1.2; ...
    gallery('uniformdata',[10 3],14)+2.5];
y = [ones(10,1);2*(ones(10,1));3*(ones(10,1))]; % Actual classes
```

Create a scatter plot of the data.

```
scatter3(X(:,1),X(:,2),X(:,3),100,y,'filled')
title('Randomly Generated Data in Three Clusters');
```

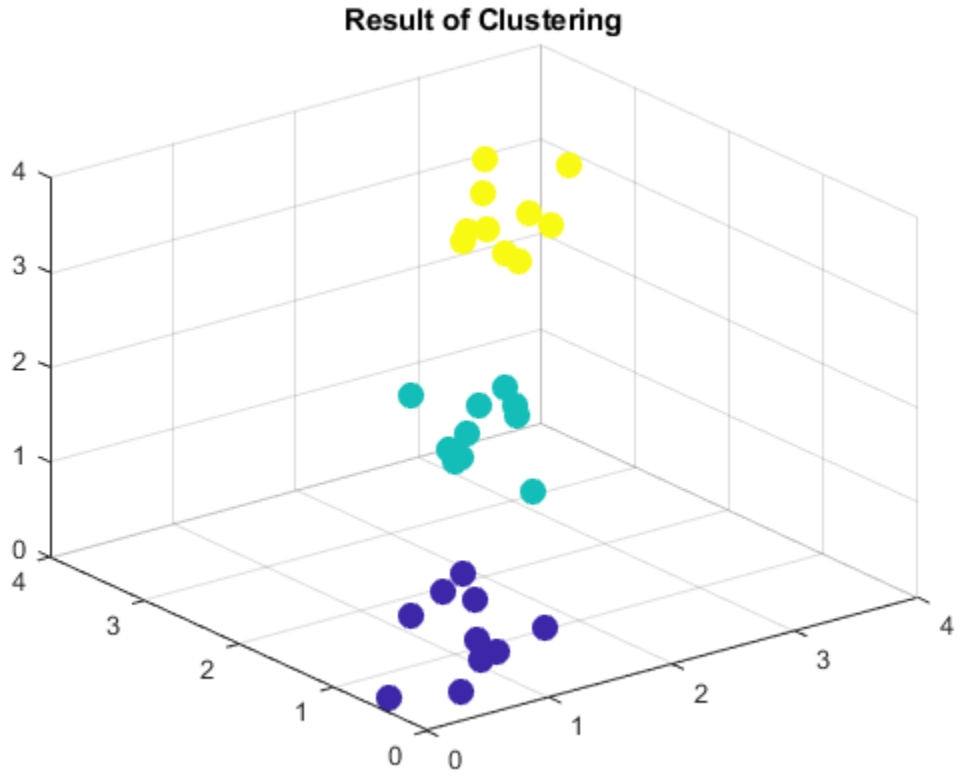
Find a maximum of three clusters in the data by specifying the value 3 for the `cutoff` input argument.

```
T1 = clusterdata(X,3);
```

Because the value of `cutoff` is greater than 2, `clusterdata` interprets `cutoff` as the maximum number of clusters.

Plot the data with the resulting cluster assignments.

```
scatter3(X(:,1),X(:,2),X(:,3),100,T1,'filled')
title('Result of Clustering');
```

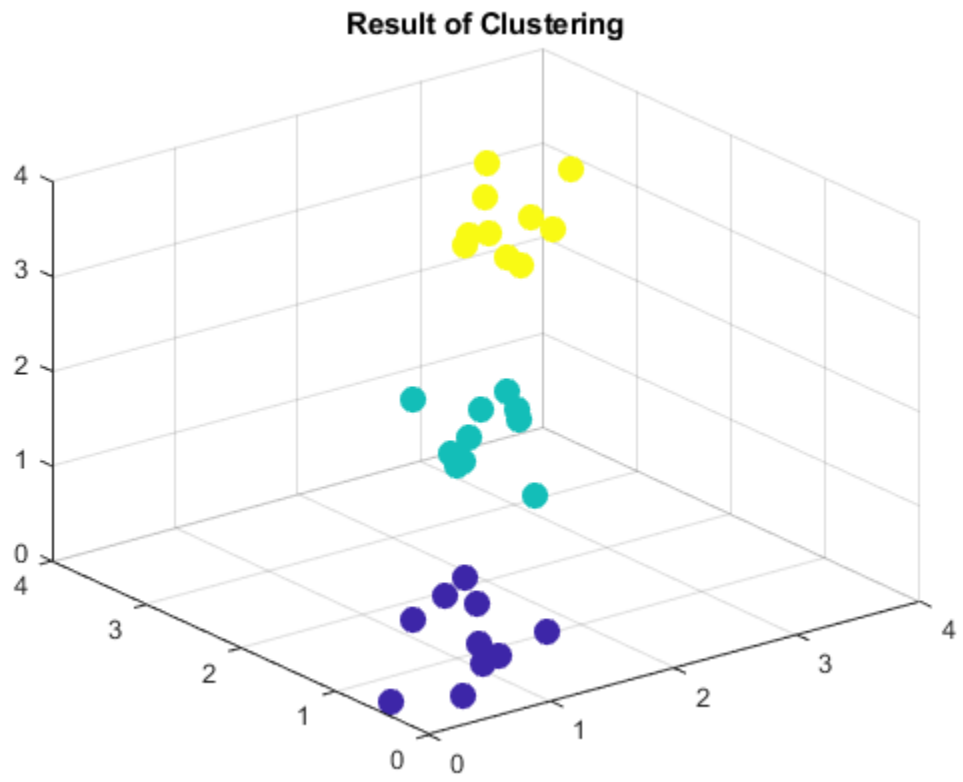


Find a maximum of three clusters by specifying the value 3 for the `'MaxClust'` name-value pair argument.

```
T2 = clusterdata(X,'Maxclust',3);
```

Plot the data with the resulting cluster assignments.

```
scatter3(X(:,1),X(:,2),X(:,3),100,T2,'filled')
title('Result of Clustering');
```



Using both approaches, `clusterdata` identifies the three distinct clusters in the data.

Create and Cluster Hierarchical Tree

Create a hierarchical cluster tree and find clusters in one step. Visualize the clusters using a 3-D scatter plot.

Create a 20,000-by-3 matrix of sample data generated from the standard uniform distribution.

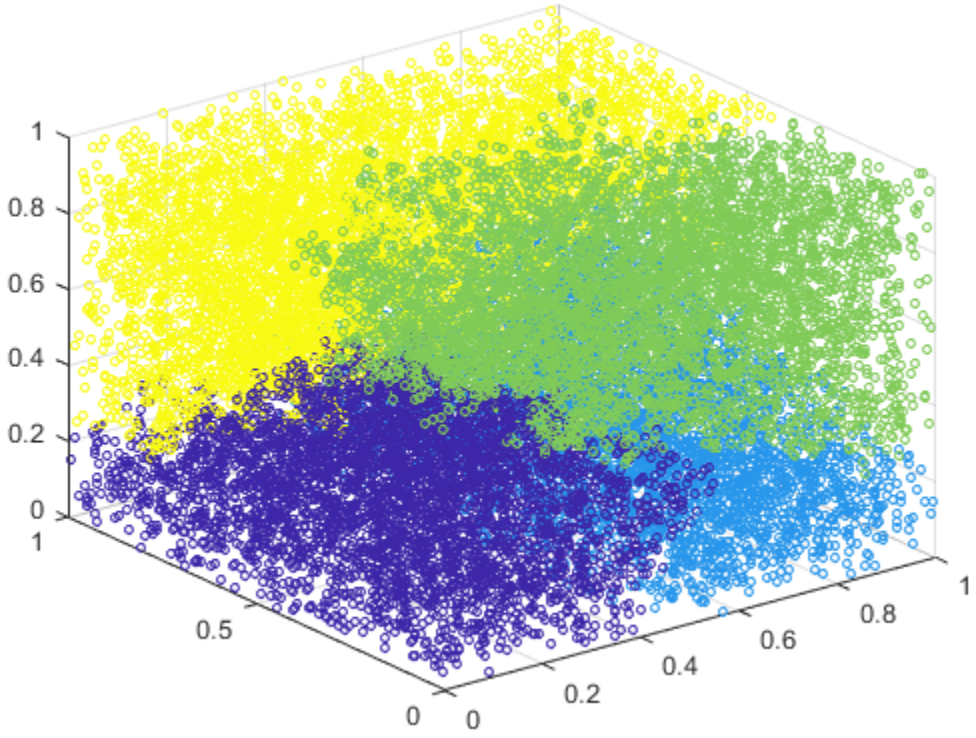
```
rng('default');  % For reproducibility
X = rand(20000,3);
```

Find a maximum of four clusters in a hierarchical cluster tree created using the `ward` linkage method. Specify `'SaveMemory'` as `'on'` to construct clusters without computing the distance matrix. Otherwise, you can receive an out-of-memory error if your machine does not have enough memory to hold the distance matrix.

```
T = clusterdata(X,'Linkage','ward','SaveMemory','on','Maxclust',4);
```

Plot the data with each cluster shown in a different color.

```
scatter3(X(:,1),X(:,2),X(:,3),10,T)
```



`clusterdata` identifies four clusters in the data.

Input Arguments

X — Input data

numeric matrix

Input data, specified as a numeric matrix with two or more rows. The rows represent observations, and the columns represent categories or dimensions.

Data Types: `single` | `double`

`cutoff` — Threshold for cutting the hierarchical tree

positive scalar between `0` and `2` | positive integer ≥ `2`

Threshold for cutting the hierarchical tree defined by `linkage`, specified as a positive scalar between `0` and `2` or a positive integer ≥ `2`. `clusterdata` behaves differently depending on the value specified for `cutoff`.

- If `0 < cutoff < 2`, then `clusterdata` forms clusters when `inconsistent` values are greater than `cutoff`.
- If `cutoff` is an integer ≥ `2`, then `clusterdata` forms a maximum of `cutoff` clusters.

When you specify `cutoff`, you cannot specify any name-value pair arguments.

Example: `clusterdata(X,3)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `clusterdata(X,'Linkage','ward','MaxClust',3)` specifies creating a maximum of three clusters of `X` using Ward linkage.

`Criterion` — Criterion for defining clusters

`'inconsistent'` | `'distance'`

Criterion for defining clusters in a hierarchical cluster tree, specified as the comma-separated pair consisting of `'Criterion'` and either `'inconsistent'` or `'distance'`. When you specify `'Criterion'`, you must also specify a value for `MaxClust` or `Cutoff`.

Example: `clusterdata(X,'Criterion','distance','Cutoff',.5)`

Data Types: `char` | `string`

`Cutoff` — Cutoff for inconsistent or distance criterion

positive scalar

Cutoff for inconsistent or distance criterion, specified as the comma-separated pair consisting of `'Cutoff'` and a positive scalar. `clusterdata` uses `Cutoff` as a threshold for either the heights or the inconsistency coefficients of nodes, depending on the value of `Criterion`. If you specify a value for `'Cutoff'` without specifying the criterion for defining clusters, then `clusterdata` uses the `'inconsistent'` criterion by default.

- If `'Criterion'` is `'distance'`, then `clusterdata` groups all leaves at or below a node into a cluster, provided that the height of the node is less than `Cutoff`.
- If `'Criterion'` is `'inconsistent'`, then the `inconsistent` values of a node and all its subnodes must be less than `Cutoff` for `clusterdata` to group them into a cluster.

You must specify either `Cutoff` or `MaxClust`.

Example: `clusterdata(X,'Cutoff',0.2)`

Data Types: `single` | `double`

`Depth` — Depth for computing inconsistent values

numeric scalar

Depth for computing inconsistent values, specified as the comma-separated pair consisting of `'Depth'` and a numeric scalar. `clusterdata` evaluates inconsistent values by looking to the specified depth below each node in the hierarchical cluster tree. When you specify `'Depth'`, you must also specify a value for `MaxClust` or `Cutoff`.

Example: `clusterdata(X,'Depth',3,'Cutoff',0.5)`

Data Types: `single` | `double`

`Distance` — Distance metric
`'euclidean'` (default) | `'squaredeuclidean'` | `'seuclidean'` | `'mahalanobis'` | function handle | ...

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and any distance metric accepted by the `pdist` function, as descried in the following table. When you specify `'Distance'`, you must also specify a value for `MaxClust` or `Cutoff`.

Metric	Description
`'euclidean'`	Euclidean distance (default)
`'squaredeuclidean'`	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`
`'cityblock'`	City block distance
`'minkowski'`	Minkowski distance. The default exponent is 2. To use a different exponent `P`, specify `P` after `'minkowski'`, where `P` is a positive scalar value: `'minkowski',P`.
`'chebychev'`	Chebychev distance (maximum coordinate difference)
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, using a built-in `distance` is generally faster than using a function handle.

For more information, see “Distance Metrics” on page 33-4495.

Example: `clusterdata(X,'Distance','minkowski','MaxClust',4)`

Data Types: `char` | `string` | `function_handle`

`Linkage` — Algorithm for computing the distance between clusters
`'average'` | `'centroid'` | `'complete'` | `'median'` | `'single'` | `'ward'` | `'weighted'`

Algorithm for computing distance between clusters, specified as the comma-separated pair consisting of `'Linkage'` and any algorithm accepted by the `linkage` function, as described in the following table. When you specify `'Linkage'`, you must also specify a value for `MaxClust` or `Cutoff`.

Algorithm	Description
`'average'`	Unweighted average distance (UPGMA)
`'centroid'`	Centroid distance (UPGMC), appropriate for Euclidean distances only
`'complete'`	Farthest distance
`'median'`	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
`'single'`	Shortest distance
`'ward'`	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
`'weighted'`	Weighted average distance (WPGMA)

For more information, see “Linkages” on page 33-3554.

Example: `clusterdata(X,'Linkage','median','MaxClust',4)`

Data Types: `char` | `string`

`MaxClust` — Maximum number of clusters
positive integer

Maximum number of clusters to form, specified as the comma-separated pair consisting of `'MaxClust'` and a positive integer.

You must specify either `Cutoff` or `MaxClust`.

Example: `clusterdata(X,'MaxClust',4)`

Data Types: `single` | `double`

`SaveMemory` — Option for saving memory
`'on'` | `'off'`

Option for saving memory, specified as the comma-separated pair consisting of `'SaveMemory'` and either `'on'` or `'off'`. When you specify `'SaveMemory'`, you must also specify a value for `MaxClust` or `Cutoff`.

The `'on'` setting causes `clusterdata` to construct clusters without computing the distance matrix. The `'on'` setting applies when both of these conditions are satisfied:

- `Linkage` is `'centroid'`, `'median'`, or `'ward'`.

- `Distance` is `'euclidean'` (default).

When these two conditions apply, the default value for `'SaveMemory'` is `'on'` if `X` has 20 columns or fewer, or if the computer does not have enough memory to store the distance matrix. Otherwise, the default value for `'SaveMemory'` is `'off'`.

When `'SaveMemory'` is `'on'`, the `linkage` run time is proportional to the number of dimensions (number of columns of `X`). When `'SaveMemory'` is `'off'`, the `linkage` memory requirement is proportional to N^2, where `N` is the number of observations. Choosing the best (least-time) setting for `'SaveMemory'` depends on the problem dimensions, number of observations, and available memory. The default `'SaveMemory'` setting is a rough approximation of an optimal setting.

Example: `'SaveMemory','on'`

Data Types: `char` | `string`

Output Arguments

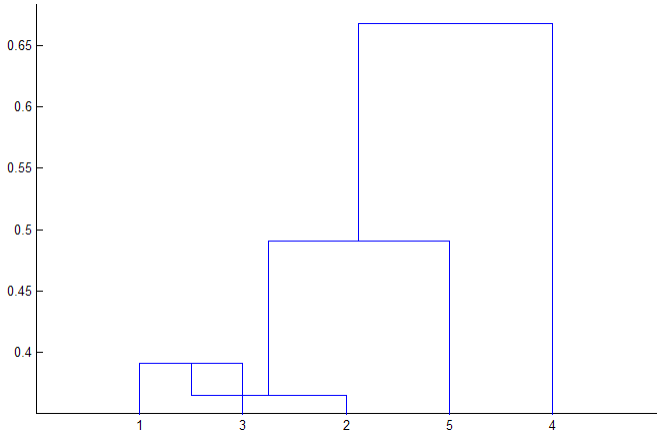
T — Cluster indices

numeric column vector

Cluster indices, returned as a numeric column vector. `T` has as many rows as `X`, and each row of `T` indicates the cluster assignment of the corresponding observation in `X`.

Tips

- If `'Linkage'` is `'centroid'` or `'median'`, then `linkage` can produce a cluster tree that is not monotonic. This result occurs when the distance from the union of two clusters, *r* and *s*, to a third cluster is less than the distance between *r* and *s*. In this case, in a dendrogram drawn with the default orientation, the path from a leaf to the root node takes some downward steps. To avoid this result, specify another value for `'Linkage'`. The following image shows a nonmonotonic cluster tree.



In this case, cluster 1 and cluster 3 are joined into a new cluster, while the distance between this new cluster and cluster 2 is less than the distance between cluster 1 and cluster 3.

Algorithms

If you specify a value `c` for the `cutoff` input argument, then `T = clusterdata(X,c)` performs the following steps:

1. Create a vector of the Euclidean distance between pairs of observations in `X` by using `pdist`.

   ```
   Y = pdist(X,'euclidean')
   ```

2. Create an agglomerative hierarchical cluster tree from `Y` by using `linkage` with the `'single'` method for computing the shortest distance between clusters.

   ```
   Z = linkage(Y,'single')
   ```

3. If `0 < c < 2`, use `cluster` to define clusters from `Z` when inconsistent values are less than `c`.

   ```
   T = cluster(Z,'Cutoff',c)
   ```

4. If `c` is an integer value ≥ 2, use `cluster` to find a maximum of `c` clusters from `Z`.

   ```
   T = cluster(Z,'MaxClust',c)
   ```

Alternative Functionality

If you have a hierarchical cluster tree `Z` (the output of the `linkage` function for the input data matrix `X`), you can use `cluster` to perform agglomerative clustering on `Z` and return the cluster assignment for each observation (row) in `X`.

See Also

`cluster` | `dendrogram` | `inconsistent` | `kmeans` | `linkage` | `pdist`

Topics

“Hierarchical Clustering” on page 16-6

Introduced before R2006a

cmdscale

Classical multidimensional scaling

Syntax

```
Y = cmdscale(D)
[Y,e] = cmdscale(D)
[Y,e] = cmdscale(D,p)
```

Description

`Y = cmdscale(D)` takes an `n`-by-`n` distance matrix `D`, and returns an `n`-by-`p` configuration matrix `Y`. Rows of `Y` are the coordinates of `n` points in `p`-dimensional space for some `p < n`. When `D` is a Euclidean distance matrix, the distances between those points are given by `D`. `p` is the dimension of the smallest space in which the `n` points whose inter-point distances are given by `D` can be embedded.

`[Y,e] = cmdscale(D)` also returns the eigenvalues of `Y*Y'`. When `D` is Euclidean, the first `p` elements of `e` are positive, the rest zero. If the first `k` elements of `e` are much larger than the remaining `(n-k)`, then you can use the first `k` columns of `Y` as `k`-dimensional points whose inter-point distances approximate `D`. This can provide a useful dimension reduction for visualization, e.g., for `k = 2`.

`D` need not be a Euclidean distance matrix. If it is non-Euclidean or a more general dissimilarity matrix, then some elements of `e` are negative, and `cmdscale` chooses `p` as the number of positive eigenvalues. In this case, the reduction to `p` or fewer dimensions provides a reasonable approximation to `D` only if the negative elements of `e` are small in magnitude.

`[Y,e] = cmdscale(D,p)` also accepts a positive integer `p` between 1 and `n`. `p` specifies the dimensionality of the desired embedding `Y`. If a `p` dimensional embedding is possible, then `Y` will be of size `n`-by-`p` and `e` will be of size `p`-by-1. If only a `q` dimensional embedding with `q < p` is possible, then `Y` will be of size `n`-by-`q` and `e` will be of size `p`-by-1. Specifying `p` may reduce the computational burden when `n` is very large.

You can specify `D` as either a full dissimilarity matrix, or in upper triangle vector form such as is output by `pdist`. A full dissimilarity matrix must be real and symmetric, and have zeros along the diagonal and positive elements everywhere else. A dissimilarity matrix in upper triangle form must have real, positive entries. You can also specify `D` as a full similarity matrix, with ones along the diagonal and all other elements less than one. `cmdscale` transforms a similarity matrix to a dissimilarity matrix in such a way that distances between the points returned in `Y` equal or approximate `sqrt(1-D)`. To use a different transformation, you must transform the similarities prior to calling `cmdscale`.

Examples

Construct a Map Using Multidimensional Scaling

This example shows how to construct a map of 10 US cities based on the distances between those cities, using `cmdscale`.

First, create the distance matrix and pass it to `cmdscale`. In this example, `D` is a full distance matrix: it is square and symmetric, has positive entries off the diagonal, and has zeros on the diagonal.

```
cities = ...
{'Atl','Chi','Den','Hou','LA','Mia','NYC','SF','Sea','WDC'};
D = [    0  587 1212  701 1936  604  748 2139 2182   543;
       587    0  920  940 1745 1188  713 1858 1737   597;
      1212  920    0  879  831 1726 1631  949 1021  1494;
       701  940  879    0 1374  968 1420 1645 1891  1220;
      1936 1745  831 1374    0 2339 2451  347  959  2300;
       604 1188 1726  968 2339    0 1092 2594 2734   923;
       748  713 1631 1420 2451 1092    0 2571 2408   205;
      2139 1858  949 1645  347 2594 2571    0  678  2442;
      2182 1737 1021 1891  959 2734 2408  678    0  2329;
       543  597 1494 1220 2300  923  205 2442 2329     0];
[Y,eigvals] = cmdscale(D);
```

Next, look at the eigenvalues returned by `cmdscale`. Some of these are negative, indicating that the original distances are not Euclidean. This is because of the curvature of the earth.

```
format short g
[eigvals eigvals/max(abs(eigvals))]
```

```
ans = 10×2

   9.5821e+06           1
   1.6868e+06     0.17604
       8157.3   0.0008513
       1432.9  0.00014954
       508.67  5.3085e-05
       25.143   2.624e-06
   4.6705e-10  4.8741e-17
       -897.7 -9.3685e-05
      -5467.6  -0.0005706
       -35479  -0.0037026
```

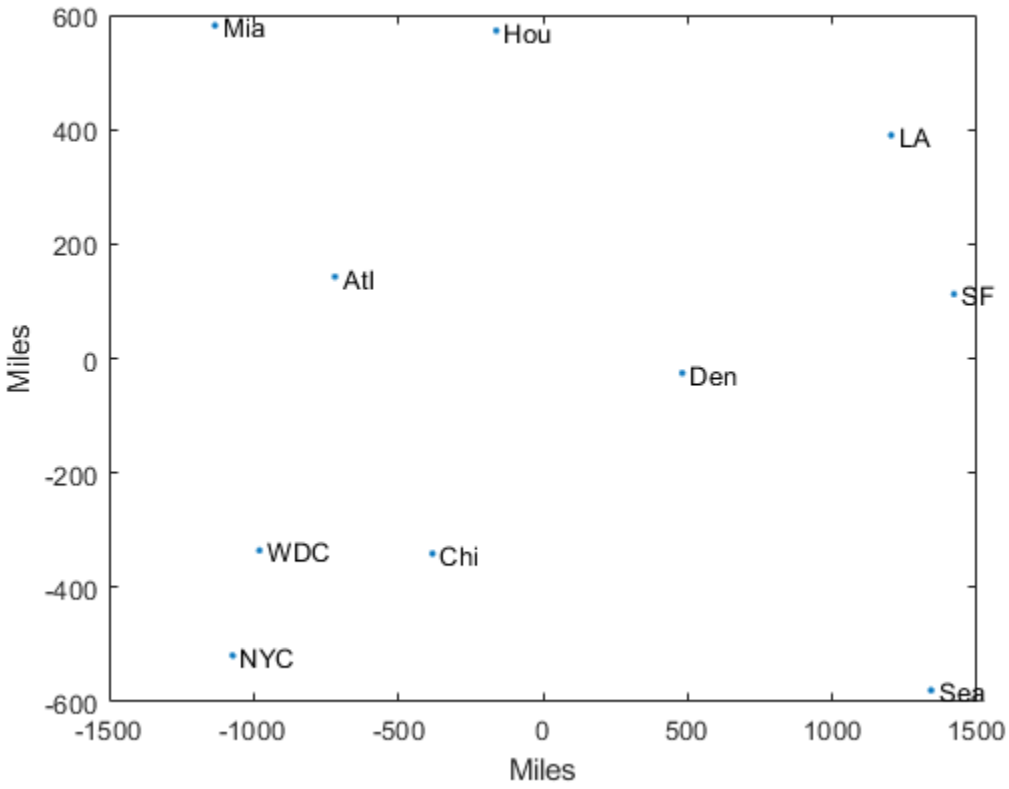
However, in this case, the two largest positive eigenvalues are much larger in magnitude than the remaining eigenvalues. So, despite the negative eigenvalues, the first two coordinates of `Y` are sufficient for a reasonable reproduction of `D`.

```
Dtriu = D(find(tril(ones(10),-1)))';
maxrelerr = max(abs(Dtriu-pdist(Y(:,1:2))))./max(Dtriu)
```

```
maxrelerr =
    0.0075371
```

Here is a plot of the reconstructed city locations as a map. The orientation of the reconstruction is arbitrary.

```
plot(Y(:,1),Y(:,2),'.')
text(Y(:,1)+25,Y(:,2),cities)
xlabel('Miles')
ylabel('Miles')
```



Evaluate Reconstructions Using Different Distance Metrics

Determine how the quality of reconstruction varies when you reduce points to distances using different metrics.

Generate ten points in 4-D space that are close to 3-D space. Take a linear transformation of the points so that their transformed values are close to a 3-D subspace that does not align with the coordinate axes.

```
rng default   % Set the seed for reproducibility
A = [normrnd(0,1,10,3) normrnd(0,0.1,10,1)];
B = randn(4,4);
X = A*B;
```

Reduce the points in `X` to distances by using the Euclidean metric. Find a configuration `Y` with the inter-point distances.

```
D = pdist(X,'euclidean');
Y = cmdscale(D);
```

Compare the quality of the reconstructions when using 2, 3, or 4 dimensions. The small `maxerr3` value indicates that the first 3 dimensions provide a good reconstruction.

```
maxerr2 = max(abs(pdist(X)-pdist(Y(:,1:2))))
```

```
maxerr2 = 0.1631

maxerr3 = max(abs(pdist(X)-pdist(Y(:,1:3))))

maxerr3 = 0.0187

maxerr4 = max(abs(pdist(X)-pdist(Y)))

maxerr4 = 1.3545e-14
```

Reduce the points in `X` to distances by using the `'cityblock'` metric. Find a configuration `Y` with the inter-point distances.

```
D = pdist(X,'cityblock');
[Y,e] = cmdscale(D);
```

Evaluate the quality of the reconstruction. `e` contains at least one negative element of large magnitude, which might account for the poor quality of the reconstruction.

```
maxerr = max(abs(pdist(X)-pdist(Y)))

maxerr = 9.0488

min(e)

ans = -5.6586
```

References

[1] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

See Also

`mdscale` | `pdist` | `procrustes`

Introduced before R2006a

coefci

Confidence interval for Cox proportional hazards model coefficients

Syntax

```
ci = coefci(coxMdl)
ci = coefci(coxMdl,level)
```

Description

`ci = coefci(coxMdl)` returns a 95% confidence interval for the coefficients of a trained Cox proportional hazards model.

`ci = coefci(coxMdl,level)` returns a 100(1 – `level`)% confidence interval for the coefficients.

Examples

Cox Model Confidence Interval

Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

```
load lightbulb
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3))
```

```
coxMdl =
Cox Proportional Hazards regression model:

           Beta       SE      zStat       pValue
          ______    ______    ______    __________

    X1    4.7262    1.0372    4.5568    5.1936e-06
```

Find a 95% confidence interval for the returned `Beta` estimate.

```
ci = coefci(coxMdl)
```

```
ci = 1×2

    2.6934    6.7590
```

Find a 99% confidence interval for the `Beta` estimate.

```
ci99 = coefci(coxMdl,0.01)
```

```
ci99 = 1×2

    2.0546    7.3978
```

Confidence Intervals for Multiple Predictors

Find confidence intervals for predictors of the `readmissiontimes` data set. The response variable is `ReadmissionTime`, which shows the readmission times for 100 patients. The predictor variables are `Age`, `Sex`, `Weight`, and `Smoker`, the smoking status of each patient. A 1 indicates the patient is a smoker, and a 0 indicates the patient does not smoke. The column vector `Censored` contains the censorship information for each patient, where 1 indicates censored data, and 0 indicates the exact readmission times are observed. (This data is simulated.)

Load the data.

```
load readmissiontimes
```

Use all four predictors for fitting a model.

```
X = [Age Sex Weight Smoker];
```

Fit the model using the censoring information.

```
coxMdl = fitcox(X,ReadmissionTime,'censoring',Censored);
```

View the point estimates for the `Age`, `Sex`, `Weight`, and `Smoker` coefficients.

```
coxMdl.Coefficients.Beta
```

```
ans = 4×1

    0.0184
   -0.0676
    0.0343
    0.8172
```

Find 95% confidence intervals for these estimates.

```
ci = coefci(coxMdl)
```

```
ci = 4×2

   -0.0139    0.0506
   -1.6488    1.5136
    0.0042    0.0644
    0.2767    1.3576
```

The Sex coefficient (second row) has a large confidence interval, and the first two coefficients bracket the value 0. Therefore, you cannot reject the hypothesis that the Age and Sex predictors are zero.

Input Arguments

coxMdl — Fitted Cox proportional hazards model

CoxModel object

Fitted Cox proportional hazards model, specified as a CoxModel object. Create coxMdl using fitcox.

level — Level of significance for confidence interval

0.05 (default) | positive number less than 1

Level of significance for the confidence interval, specified as a positive number less than 1. The resulting percentage is 100(1 – level)%. For example, for a 99% confidence interval, specify level as 0.01.

Example: 0.01

Data Types: double

Output Arguments

ci — Confidence interval

real two-column matrix

Confidence interval, returned as a real two-column matrix. Each row of the matrix is a confidence interval for the corresponding predictor. The probability that the true predictor coefficient lies in its confidence interval is 100(1 – level)%. For example, the default value of level is 0.05, so with no level specified, the probability that each predictor lies in its row of ci is 95%.

See Also

CoxModel | fitcox | linhyptest

Introduced in R2021a

coefCI

Package:

Confidence intervals of coefficient estimates of generalized linear regression model

Syntax

```
ci = coefCI(mdl)
ci = coefCI(mdl,alpha)
```

Description

`ci = coefCI(mdl)` returns 95% confidence intervals for the coefficients in `mdl`.

`ci = coefCI(mdl,alpha)` returns confidence intervals using the confidence level 1 – `alpha`.

Examples

Find Confidence Intervals for Model Coefficients

Find the confidence intervals for the coefficients of a fitted generalized linear regression model.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Find 95% (default) confidence intervals for the coefficients of the model.

```
ci = coefCI(mdl)
```

```
ci = 3×2

    0.9966    1.0844
    0.9901    1.0035
    1.9744    1.9996
```

Find 99% confidence intervals for the coefficients.

```
alpha = 0.01;
ci = coefCI(mdl,alpha)
```

```
ci = 3×2

    0.9824    1.0986
    0.9880    1.0056
    1.9703    2.0036
```

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

`alpha` — Significance level

0.05 (default) | numeric value in the range [0,1]

Significance level for the confidence interval, specified as a numeric value in the range [0,1]. The confidence level of `ci` is equal to 100(1 – `alpha`)%. `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

`ci` — Confidence intervals

numeric matrix

Confidence intervals, returned as a *k*-by-2 numeric matrix, where *k* is the number of coefficients. The *j*th row of `ci` is the confidence interval of the *j*th coefficient of `mdl`. The name of coefficient *j* is stored in the `CoefficientNames` property of `mdl`.

Data Types: `single` | `double`

More About

Confidence Interval

The coefficient confidence intervals provide a measure of precision for regression coefficient estimates.

A 100(1 – α)% confidence interval gives the range that the corresponding regression coefficient will be in with 100(1 – α)% confidence, meaning that 100(1 – α)% of the intervals resulting from repeated experimentation will contain the true value of the coefficient.

The software finds confidence intervals using the Wald method. The 100*(1 – α)% confidence intervals for regression coefficients are

$$b_i \pm t_{(1-\alpha/2,\, n-p)} SE(b_i),$$

where b_i is the coefficient estimate, $SE(b_i)$ is the standard error of the coefficient estimate, and $t_{(1-\alpha/2,n-p)}$ is the 100(1 – α/2) percentile of t-distribution with $n – p$ degrees of freedom. n is the number of observations and p is the number of regression coefficients.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `coefTest` | `devianceTest`

Topics

“Generalized Linear Model Workflow” on page 12-28

Introduced in R2012a

coefCI

Class: `GeneralizedLinearMixedModel`

Confidence intervals for coefficients of generalized linear mixed-effects model

Syntax

```
feCI = coefCI(glme)
feCI = coefCI(glme,Name,Value)
[feCI,reCI] = coefCI(___)
```

Description

`feCI = coefCI(glme)` returns the 95% confidence intervals for the fixed-effects coefficients in the generalized linear mixed-effects model `glme`.

`feCI = coefCI(glme,Name,Value)` returns the confidence intervals using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify a different confidence level or the method used to compute the approximate degrees of freedom.

`[feCI,reCI] = coefCI(___)` also returns the confidence intervals for the random-effects coefficients using any of the previous syntaxes.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range [0,1]

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range [0,1]. For a value α, the confidence level is 100 × (1 – α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`DFMethod` — Method for computing approximate degrees of freedom

`'residual'` (default) | `'none'`

Method for computing approximate degrees of freedom, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom value is assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	The degrees of freedom is set to infinity.

Example: `'DFMethod','none'`

Output Arguments

`feCI` — Fixed-effects confidence intervals

p-by-2 matrix

Fixed-effects confidence intervals, returned as a *p*-by-2 matrix. `feCI` contains the confidence limits that correspond to the *p*-by-1 fixed-effects vector returned by the `fixedEffects` method. The first column of `feCI` contains the lower confidence limits and the second column contains the upper confidence limits.

When fitting a GLME model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`):

- If you specify the `'CovarianceMethod'` name-value pair argument as `'conditional'`, then the confidence intervals are conditional on the estimated covariance parameters.
- If you specify the `'CovarianceMethod'` name-value pair argument as `'JointHessian'`, then the confidence intervals account for the uncertainty in the estimated covariance parameters.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `coefci` uses the fitted linear mixed effects model from the final pseudo likelihood iteration to compute confidence intervals on the fixed effects.

`reCI` — Random-effects confidence intervals

q-by-2 matrix

Random-effects confidence intervals, returned as a *q*-by-2 matrix. `reCI` contains the confidence limits corresponding to the *q*-by-1 random-effects vector `B` returned by the `randomEffects` method. The first column of `reCI` contains the lower confidence limits, and the second column contains the upper confidence limits.

When fitting a GLME model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`), `coefCI` computes the confidence intervals using the conditional mean squared error of prediction (CMSEP) approach conditional on the estimated covariance parameters and the observed response. Alternatively, you can interpret the confidence intervals from `coefCI` as approximate Bayesian credible intervals conditional on the estimated covariance parameters and the observed response.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `coefci` uses the fitted linear mixed effects model from the final pseudo likelihood iteration to compute confidence intervals on the random effects.

Examples

95% Confidence Intervals for Fixed Effects

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.

- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Use `fixedEffects` to display the estimates and names of the fixed-effects coefficients in `glme`.

```
[beta,betanames] = fixedEffects(glme)
```

```
beta = 6×1

    1.4689
   -0.3677
   -0.0945
   -0.2832
   -0.0719
    0.0711


betanames=6×1 table
         Name
    ______________

    {'(Intercept)'}
    {'newprocess' }
    {'time_dev'   }
    {'temp_dev'   }
    {'supplier_C' }
    {'supplier_B' }
```

Each row of `beta` contains the estimated value for the coefficient named in the corresponding row of `betanames`. For example, the value –0.0945 in row 3 of `beta` is the estimated coefficient for the predictor variable `time_dev`.

Compute the 95% confidence intervals for the fixed-effects coefficients.

```
feCI = coefCI(glme)
```

```
feCI = 6×2

    1.1515    1.7864
   -0.7202   -0.0151
   -1.7395    1.5505
   -2.1926    1.6263
   -0.2268    0.0831
   -0.0826    0.2247
```

Column 1 of `feCI` contains the lower bound of the 95% confidence interval. Column 2 contains the upper bound. Row 1 corresponds to the intercept term. Rows 2, 3, and 4 correspond to `newprocess`, `time_dev`, and `temp_dev`, respectively. Rows 5 and 6 correspond to the indicator variables `supplier_C` and `supplier_B`, respectively. For example, the 95% confidence interval for the coefficient for `time_dev` is [-1.7395 , 1.5505]. Some of the confidence intervals include 0, which

indicates that those predictors are not significant at the 5% significance level. To obtain specific *p*-values for each fixed-effects term, use `fixedEffects`. To test significance for entire terms, use `anova`.

99% Confidence Intervals for Random Effects

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).

- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Use `randomEffects` to compute and display the estimates of the empirical Bayes predictors (EBPs) for the random effects associated with `factory`.

```
[B,Bnames] = randomEffects(glme)
```

```
B = 20×1

    0.2913
    0.1542
   -0.2633
   -0.4257
    0.5453
   -0.1069
    0.3040
   -0.1653
   -0.1458
   -0.0816
      ⋮


Bnames=20×3 table
       Group       Level          Name
    ___________    ______    _______________

    {'factory'}    {'1' }    {'(Intercept)'}
    {'factory'}    {'2' }    {'(Intercept)'}
    {'factory'}    {'3' }    {'(Intercept)'}
    {'factory'}    {'4' }    {'(Intercept)'}
    {'factory'}    {'5' }    {'(Intercept)'}
    {'factory'}    {'6' }    {'(Intercept)'}
    {'factory'}    {'7' }    {'(Intercept)'}
    {'factory'}    {'8' }    {'(Intercept)'}
    {'factory'}    {'9' }    {'(Intercept)'}
    {'factory'}    {'10'}    {'(Intercept)'}
    {'factory'}    {'11'}    {'(Intercept)'}
    {'factory'}    {'12'}    {'(Intercept)'}
    {'factory'}    {'13'}    {'(Intercept)'}
    {'factory'}    {'14'}    {'(Intercept)'}
    {'factory'}    {'15'}    {'(Intercept)'}
    {'factory'}    {'16'}    {'(Intercept)'}
      ⋮
```

Each row of `B` contains the estimated EBPs for the random-effects coefficient named in the corresponding row of `Bnames`. For example, the value -0.2633 in row 3 of `B` is the estimated coefficient of `'(Intercept)'` for level `'3'` of `factory`.

Compute the 99% confidence intervals of the EBPs for the random effects.

```
[feCI,reCI] = coefCI(glme,'Alpha',0.01);
reCI
```

```
reCI = 20×2

   -0.2125    0.7951
   -0.3510    0.6595
   -0.8219    0.2954
   -0.9953    0.1440
    0.0730    1.0176
   -0.6362    0.4224
   -0.1796    0.7877
   -0.7044    0.3738
   -0.6795    0.3880
   -0.6142    0.4509
      ⋮
```

Column 1 of `reCI` contains the lower bound of the 99% confidence interval. Column 2 contains the upper bound. Each row corresponds to a level of `factory`, in the order shown in `Bnames`. For example, row 3 corresponds to the coefficient of `'(Intercept)'` for level `'3'` of `factory`, which has a 99% confidence interval of [-0.8219 , 0.2954]. For additional statistics related to each random-effects term, use `randomEffects`.

References

[1] Booth, J.G., and J.P. Hobert. “Standard Errors of Prediction in Generalized Linear Mixed Models.” *Journal of the American Statistical Association*. Vol. 93, 1998, pp. 262–272.

See Also

`GeneralizedLinearMixedModel` | `anova` | `coefTest` | `covarianceParameters` | `fixedEffects` | `randomEffects`

coefCI

Package:

Confidence intervals of coefficient estimates of linear regression model

Syntax

```
ci = coefCI(mdl)
ci = coefCI(mdl,alpha)
```

Description

`ci = coefCI(mdl)` returns 95% confidence intervals for the coefficients in `mdl`.

`ci = coefCI(mdl,alpha)` returns confidence intervals using the confidence level `1 – alpha`.

Examples

Find Confidence Intervals for Model Coefficients

Fit a linear regression model and obtain the default 95% confidence intervals for the resulting model coefficients.

Load the `carbig` data set and create a table in which the `Origin` predictor is categorical.

```
load carbig
Origin = categorical(cellstr(Origin));
tbl = table(Horsepower,Weight,MPG,Origin);
```

Fit a linear regression model. Specify `Horsepower`, `Weight`, and `Origin` as predictor variables, and specify `MPG` as the response variable.

```
modelspec = 'MPG ~ 1 + Horsepower + Weight + Origin';
mdl = fitlm(tbl,modelspec);
```

View the names of the coefficients.

```
mdl.CoefficientNames
```

```
ans = 1x9 cell
  Columns 1 through 4

    {'(Intercept)'}    {'Horsepower'}    {'Weight'}    {'Origin_France'}

  Columns 5 through 7

    {'Origin_Germany'}    {'Origin_Italy'}    {'Origin_Japan'}

  Columns 8 through 9

    {'Origin_Sweden'}    {'Origin_USA'}
```

Find confidence intervals for the coefficients of the model.

```
ci = coefCI(mdl)
```

```
ci = 9×2

   43.3611   59.9390
   -0.0748   -0.0315
   -0.0059   -0.0037
  -17.3623   -0.3477
  -15.7503    0.7434
  -17.2091    0.0613
  -14.5106    1.8738
  -18.5820   -1.5036
  -17.3114   -0.9642
```

Specify Confidence Level

Fit a linear regression model and obtain the confidence intervals for the resulting model coefficients using a specified confidence level.

Load the `carbig` data set and create a table in which the `Origin` predictor is categorical.

```
load carbig
Origin = categorical(cellstr(Origin));
tbl = table(Horsepower,Weight,MPG,Origin);
```

Fit a linear regression model. Specify `Horsepower`, `Weight`, and `Origin` as predictor variables, and specify `MPG` as the response variable.

```
modelspec = 'MPG ~ 1 + Horsepower + Weight + Origin';
mdl = fitlm(tbl,modelspec);
```

Find 99% confidence intervals for the coefficients.

```
ci = coefCI(mdl,.01)
```

```
ci = 9×2

   40.7365   62.5635
   -0.0816   -0.0246
   -0.0062   -0.0034
  -20.0560    2.3459
  -18.3615    3.3546
  -19.9433    2.7955
  -17.1045    4.4676
  -21.2858    1.2002
  -19.8995    1.6238
```

The confidence intervals are wider than the default 95% confidence intervals in “Find Confidence Intervals for Model Coefficients” on page 33-664.

Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`alpha` — Significance level

0.05 (default) | numeric value in the range [0,1]

Significance level for the confidence interval, specified as a numeric value in the range [0,1]. The confidence level of `ci` is equal to 100(1 – `alpha`)%. `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

`ci` — Confidence intervals

numeric matrix

Confidence intervals, returned as a *k*-by-2 numeric matrix, where *k* is the number of coefficients. The *j*th row of `ci` is the confidence interval of the *j*th coefficient of `mdl`. The name of coefficient *j* is stored in the `CoefficientNames` property of `mdl`.

Data Types: `single` | `double`

More About

Confidence Interval

The coefficient confidence intervals provide a measure of precision for regression coefficient estimates.

A 100(1 – α)% confidence interval gives the range that the corresponding regression coefficient will be in with 100(1 – α)% confidence, meaning that 100(1 – α)% of the intervals resulting from repeated experimentation will contain the true value of the coefficient.

The software finds confidence intervals using the Wald method. The 100*(1 – α)% confidence intervals for regression coefficients are

$$b_i \pm t_{(1-\alpha/2,\, n-p)} SE(b_i),$$

where b_i is the coefficient estimate, $SE(b_i)$ is the standard error of the coefficient estimate, and $t_{(1-\alpha/2,n-p)}$ is the 100(1 – α/2) percentile of *t*-distribution with $n - p$ degrees of freedom. n is the number of observations and p is the number of regression coefficients.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `anova` | `coefTest` | `dwtest`

Topics

“Coefficient Standard Errors and Confidence Intervals” on page 11-58
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

coefCI

Class: `LinearMixedModel`

Confidence intervals for coefficients of linear mixed-effects model

Syntax

```
feCI = coefCI(lme)
feCI = coefCI(lme,Name,Value)
[feCI,reCI] = coefCI( ___ )
```

Description

`feCI = coefCI(lme)` returns the 95% confidence intervals for the fixed-effects coefficients in the linear mixed-effects model `lme`.

`feCI = coefCI(lme,Name,Value)` returns the 95% confidence intervals for the fixed-effects coefficients in the linear mixed-effects model `lme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the confidence level or method to compute the degrees of freedom.

`[feCI,reCI] = coefCI( ___ )` also returns the 95% confidence intervals for the random-effects coefficients in the linear mixed-effects model `lme`.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

DFMethod — Method for computing approximate degrees of freedom

`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing approximate degrees of freedom for confidence interval computation, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

Output Arguments

feCI — Fixed-effects confidence intervals

p-by-2 matrix

Fixed-effects confidence intervals, returned as a *p*-by-2 matrix. `feCI` contains the confidence limits that correspond to the *p* fixed-effects estimates in the vector `beta` returned by the `fixedEffects` method. The first column of `feCI` has the lower confidence limits and the second column has the upper confidence limits.

reCI — Random-effects confidence intervals

q-by-2 matrix

Random-effects confidence intervals, returned as a *q*-by-2 matrix. `reCI` contains the confidence limits corresponding to the *q* random-effects estimates in the vector `B` returned by the `randomEffects` method. The first column of `reCI` has the lower confidence limits and the second column has the upper confidence limits.

Examples

95% Confidence Intervals for Fixed-Effects Coefficients

Load the sample data.

```
load('weight.mat')
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight, Program, Subject,Week, y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

Compute the fixed-effects coefficient estimates.

```
fe = fixedEffects(lme)
```

```
fe = 9×1

    0.6610
    0.0032
    0.3608
   -0.0333
    0.1132
    0.1732
    0.0388
    0.0305
    0.0331
```

The first estimate, 0.6610, corresponds to the constant term. The second row, 0.0032, and the third row, 0.3608, are estimates for the coefficient of initial weight and week, respectively. Rows four to six correspond to the indicator variables for programs B-D, and the last three rows correspond to the interaction of programs B-D and week.

Compute the 95% confidence intervals for the fixed-effects coefficients.

```
fecI = coefCI(lme)
```

```
fecI = 9×2

    0.1480    1.1741
    0.0005    0.0059
    0.1004    0.6211
   -0.2932    0.2267
   -0.1471    0.3734
    0.0395    0.3069
   -0.1503    0.2278
   -0.1585    0.2196
   -0.1559    0.2221
```

Some confidence intervals include 0. To obtain specific *p*-values for each fixed-effects term, use the `fixedEffects` method. To test for entire terms use the `anova` method.

Confidence Intervals with Specified Options

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration and horsepower, and a potentially correlated random effect for intercept and acceleration grouped by model year. First, store the data in a table.

```
tbl = table(Acceleration,Horsepower,Model_Year,MPG);
```

Fit the model.

```
lme = fitlme(tbl, 'MPG ~ Acceleration + Horsepower + (Acceleration|Model_Year)');
```

Compute the fixed-effects coefficient estimates.

```
fe = fixedEffects(lme)
```

```
fe = 3×1

   50.1325
   -0.5833
   -0.1695
```

Compute the 99% confidence intervals for fixed-effects coefficients using the residuals method to determine the degrees of freedom. This is the default method.

```
feCI = coefCI(lme,'Alpha',0.01)
```

```
feCI = 3×2

   44.2690   55.9961
   -0.9300   -0.2365
   -0.1883   -0.1507
```

Compute the 99% confidence intervals for fixed-effects coefficients using the Satterthwaite approximation to compute the degrees of freedom.

```
feCI = coefCI(lme,'Alpha',0.01,'DFMethod','satterthwaite')
```

```
feCI = 3×2

   44.0949   56.1701
   -0.9640   -0.2025
   -0.1884   -0.1507
```

The Satterthwaite approximation produces similar confidence intervals than the residual method.

Compute Confidence Intervals for Random Effects

Load the sample data.

```
load('shift.mat')
```

The data shows the deviations from the target quality characteristic measured from the products that five operators manufacture during three shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the deviation of the quality characteristics from the target value. This is simulated data.

`Shift` and `Operator` are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if there is significant difference in the performance according to the time of the shift.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)');
```

Compute the estimate of the BLUPs for random effects.

```
randomEffects(lme)
```

```
ans = 5×1

    0.5775
    1.1757
   -2.1715
    2.3655
   -1.9472
```

Compute the 95% confidence intervals for random effects.

```
[~,reCI] = coefCI(lme)
```

```
reCI = 5×2

   -1.3916    2.5467
   -0.7934    3.1449
   -4.1407   -0.2024
    0.3964    4.3347
   -3.9164    0.0219
```

Compute the 99% confidence intervals for random effects using the residuals method to determine the degrees of freedom. This is the default method.

```
[~,reCI] = coefCI(lme,'Alpha',0.01)
```

```
reCI = 5×2

   -2.1831    3.3382
   -1.5849    3.9364
   -4.9322    0.5891
   -0.3951    5.1261
   -4.7079    0.8134
```

Compute the 99% confidence intervals for random effects using the Satterthwaite approximation to determine the degrees of freedom.

```
[~,reCI] = coefCI(lme,'Alpha',0.01,'DFMethod','satterthwaite')
```

```
reCI = 5×2

   -2.6840    3.8390
   -2.0858    4.4372
   -5.4330    1.0900
   -0.8960    5.6270
```

```
   -5.2087    1.3142
```

The Satterthwaite approximation might produce smaller `DF` values than the residual method. That is why these confidence intervals are larger than the previous ones computed using the residual method.

See Also

`LinearMixedModel` | `coefTest` | `fixedEffects` | `randomEffects`

coefCI

Class: `NonLinearModel`

Confidence intervals of coefficient estimates of nonlinear regression model

Syntax

```
ci = coefCI(mdl)
ci = coefCI(mdl,alpha)
```

Description

`ci = coefCI(mdl)` returns confidence intervals for the coefficients in `mdl`.

`ci = coefCI(mdl,alpha)` returns confidence intervals with confidence level `1 - alpha`.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`alpha`

Scalar from 0 to 1, the probability that the confidence interval does not contain the true value.

Default: 0.05

Output Arguments

`ci`

k-by-2 matrix of confidence intervals. The `j`th row of `ci` is the confidence interval of coefficient `j` of `mdl`. The name of coefficient `j` of `mdl` is in *`mdl`*`.CoefNames`.

Examples

Default Confidence Intervals

Create a nonlinear model for auto mileage based on the `carbig` data. Then obtain confidence intervals for the resulting model coefficients.

Load the data and create a nonlinear model.

```
load carbig
ds = dataset(Horsepower,Weight,MPG);
modelfun = @(b,x)b(1) + b(2)*x(:,1) + ...
    b(3)*x(:,2) + b(4)*x(:,1).*x(:,2);
```

```
beta0 = [1 1 1 1];
mdl = fitnlm(ds,modelfun,beta0)

mdl =
Nonlinear regression model:
    MPG ~ b1 + b2*Horsepower + b3*Weight + b4*Horsepower*Weight

Estimated Coefficients:
           Estimate         SE         tStat       pValue
          __________    __________    _______    __________

    b1        63.558        2.3429     27.127    1.2343e-91
    b2      -0.25084      0.027279    -9.1952    2.3226e-18
    b3     -0.010772    0.00077381    -13.921    5.1372e-36
    b4    5.3554e-05    6.6491e-06     8.0542    9.9336e-15


Number of observations: 392, Error degrees of freedom: 388
Root Mean Squared Error: 3.93
R-Squared: 0.748,  Adjusted R-Squared 0.746
F-statistic vs. constant model: 385, p-value = 7.26e-116
```

All the coefficients have extremely small *p*-values. This means a confidence interval around the coefficients will not contain the point `0`, unless the confidence level is very high.

Find 95% confidence intervals for the coefficients of the model.

```
ci = coefCI(mdl)

ci = 4×2

   58.9515   68.1644
   -0.3045   -0.1972
   -0.0123   -0.0093
    0.0000    0.0001
```

The confidence interval for `b4` seems to contain `0`. Examine it in more detail.

```
ci(4,:)

ans = 1×2
10^-4 ×

    0.4048    0.6663
```

As expected, the confidence interval does not contain the point `0`.

More About

Confidence Interval

The coefficient confidence intervals provide a measure of precision for regression coefficient estimates.

A 100(1 – α)% confidence interval gives the range that the corresponding regression coefficient will be in with 100(1 – α)% confidence, meaning that 100(1 – α)% of the intervals resulting from repeated experimentation will contain the true value of the coefficient.

The software finds confidence intervals using the Wald method. The 100*(1 – α)% confidence intervals for regression coefficients are

$$b_i \pm t_{(1-\alpha/2,\, n-p)} SE(b_i),$$

where b_i is the coefficient estimate, $SE(b_i)$ is the standard error of the coefficient estimate, and $t_{(1-\alpha/2,n-p)}$ is the 100(1 – α/2) percentile of t-distribution with n – p degrees of freedom. n is the number of observations and p is the number of regression coefficients.

See Also

`NonLinearModel`

Topics

“Nonlinear Regression Workflow” on page 13-12
“Nonlinear Regression” on page 13-2

coefTest

Package:

Linear hypothesis test on generalized linear regression model coefficients

Syntax

```
p = coefTest(mdl)
p = coefTest(mdl,H)
p = coefTest(mdl,H,C)
[p,F] = coefTest(___)
[p,F,r] = coefTest(___)
```

Description

`p = coefTest(mdl)` computes the *p*-value for an *F* test that all coefficient estimates in `mdl`, except the intercept term, are zero.

`p = coefTest(mdl,H)` performs an *F*-test that $H \times B = 0$, where B represents the coefficient vector. Use `H` to specify the coefficients to include in the *F*-test.

`p = coefTest(mdl,H,C)` performs an *F*-test that $H \times B = C$.

`[p,F] = coefTest(___)` also returns the *F*-test statistic `F` using any of the input argument combinations in previous syntaxes.

`[p,F,r] = coefTest(___)` also returns the numerator degrees of freedom `r` for the test.

Examples

Test Significance of Generalized Linear Regression Model

Fit a generalized linear regression model, and test the coefficients of the fitted model to see if they differ from zero.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
```

```
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Test whether the fitted model has coefficients that differ significantly from zero.

```
p = coefTest(mdl)
```

```
p = 4.1131e-153
```

The small *p*-value indicates that the model fits significantly better than a degenerate model consisting of only an intercept term.

Test Significance of Generalized Linear Regression Model Coefficient

Fit a generalized linear regression model, and test the significance of a specified coefficient in the fitted model.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0
```

```
100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Test the significance of the `x1` coefficient. According to the model display, `x1` is the second predictor. Specify the coefficient by using a numeric index vector.

```
p = coefTest(mdl,[0 1 0])
```

```
p = 2.8681e-145
```

The returned *p*-value indicates that `x1` is statistically significant in the fitted model.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

`H` — Hypothesis matrix

numeric index matrix

Hypothesis matrix, specified as an `r`-by-`s` numeric index matrix, where `r` is the number of coefficients to include in an *F*-test, and `s` is the total number of coefficients.

- If you specify `H`, then the output `p` is the *p*-value for an *F*-test that $H \times B = 0$, where B represents the coefficient vector.
- If you specify `H` and `C`, then the output `p` is the *p*-value for an *F*-test that $H \times B = C$.

Example: `[1 0 0 0 0]` tests the first coefficient among five coefficients.

Data Types: `single` | `double`

`C` — Hypothesized value

numeric vector

Hypothesized value for testing the null hypothesis, specified as a numeric vector with the same number of rows as `H`.

If you specify `H` and `C`, then the output `p` is the *p*-value for an *F*-test that $H \times B = C$, where B represents the coefficient vector.

Data Types: `single` | `double`

Output Arguments

`p` — *p*-value for *F*-test

numeric value in the range [0,1]

p-value for the *F*-test, returned as a numeric value in the range [0,1].

F — Value of test statistic for *F*-test
numeric value

Value of the test statistic for the *F*-test, returned as a numeric value.

r — Numerator degrees of freedom for *F*-test
positive integer

Numerator degrees of freedom for the *F*-test, returned as a positive integer. The *F*-statistic has `r` degrees of freedom in the numerator and `mdl.DFE` degrees of freedom in the denominator.

Algorithms

The *p*-value, *F*-statistic, and numerator degrees of freedom are valid under these assumptions:

- The data comes from a model represented by the formula in the `Formula` property of the fitted model.
- The observations are independent, conditional on the predictor values.

Under these assumptions, let β represent the (unknown) coefficient vector of the linear regression. Suppose H is a full-rank matrix of size r-by-s, where r is the number of coefficients to include in an F-test, and s is the total number of coefficients. Let c be a column vector with r rows. The following is a test statistic for the hypothesis that $H\beta = c$:

$$F = \left(H\widehat{\beta} - c\right)'(HVH')^{-1}\left(H\widehat{\beta} - c\right).$$

Here $\widehat{\beta}$ is the estimate of the coefficient vector β, stored in the `Coefficients` property, and V is the estimated covariance of the coefficient estimates, stored in the `CoefficientCovariance` property. When the hypothesis is true, the test statistic F has an "F Distribution" on page B-45 with r and u degrees of freedom, where u is the degrees of freedom for error, stored in the `DFE` property.

Alternative Functionality

The values of commonly used test statistics are available in the `Coefficients` property of a fitted model.

Extended Capabilities

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `coefCI` | `devianceTest` | `linhyptest`

Topics
“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2012a

coefTest

Class: `GeneralizedLinearMixedModel`

Hypothesis test on fixed and random effects of generalized linear mixed-effects model

Syntax

```
pVal = coefTest(glme)
pVal = coefTest(glme,H)
pVal = coefTest(glme,H,C)
pVal = coefTest(glme,H,C,Name,Value)
[pVal,F,DF1,DF2] = coefTest( ___ )
```

Description

`pVal = coefTest(glme)` returns the *p*-value of an *F*-test of the null hypothesis that all fixed-effects coefficients of the generalized linear mixed-effects model `glme`, except for the intercept, are equal to 0.

`pVal = coefTest(glme,H)` returns the *p*-value of an *F*-test using a specified contrast matrix, `H`. The null hypothesis is H_0: $H\beta = 0$, where β is the fixed-effects vector.

`pVal = coefTest(glme,H,C)` returns the *p*-value for an *F*-test using the hypothesized value, `C`. The null hypothesis is H_0: $H\beta = C$, where β is the fixed-effects vector.

`pVal = coefTest(glme,H,C,Name,Value)` returns the *p*-value for an *F*-test on the fixed- and/or random-effects coefficients of the generalized linear mixed-effects model `glme`, with additional options specified by one or more name-value pair arguments. For example, you can specify the method to compute the approximate denominator degrees of freedom for the *F*-test.

`[pVal,F,DF1,DF2] = coefTest( ___ )` also returns the *F*-statistic, `F`, and the numerator and denominator degrees of freedom for `F`, respectively `DF1` and `DF2`, using any of the previous syntaxes.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`H` — Fixed-effects contrasts

m-by-*p* matrix

Fixed-effects contrasts, specified as an *m*-by-*p* matrix, where *p* is the number of fixed-effects coefficients in `glme`. Each row of `H` represents one contrast. The columns of `H` (left to right) correspond to the rows of the *p*-by-1 fixed-effects vector `beta` (top to bottom) whose estimate is returned by the `fixedEffects` method.

Data Types: `single` | `double`

`C` — Hypothesized value

m-by-1 vector

Hypothesized value for testing the null hypothesis $H\beta = C$, specified as an *m*-by-1 vector. Here, β is the vector of fixed-effects whose estimate is returned by `fixedEffects`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`DFMethod` — Method for computing approximate degrees of freedom

`'residual'` (default) | `'none'`

Method for computing approximate degrees of freedom, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom value is assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	The degrees of freedom is set to infinity.

Example: `'DFMethod','none'`

`REContrast` — Random-effects contrasts

m-by-*q* matrix

Random-effects contrasts, specified as the comma-separated pair consisting of `'REContrast'` and an *m*-by-*q* matrix, where *q* is the number of random effects parameters in `glme`. The columns of the matrix (left to right) correspond to the rows of the *q*-by-1 random-effects vector `B` (top to bottom), whose estimate is returned by the `randomEffects` method.

Data Types: `single` | `double`

Output Arguments

`pVal` — *p*-value

scalar value

p-value for the *F*-test on the fixed- and/or random-effects coefficients of the generalized linear mixed-effects model `glme`, returned as a scalar value.

When fitting a GLME model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`), `coefTest` uses an approximation of the conditional mean squared error of prediction (CMSEP) of the estimated linear combination of fixed- and random-effects to compute *p*-values. This accounts for the uncertainty in the fixed-effects estimates, but not for the uncertainty in the covariance parameter estimates. For tests on fixed effects only, if you specify the `'CovarianceMethod'` name-value pair argument in `fitglme` as `'JointHessian'`, then `coefTest` accounts for the uncertainty in the estimation of covariance parameters.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `coefTest` bases the inference on the fitted linear mixed effects model from the final pseudo likelihood iteration.

`F` — *F*-statistic
scalar value

F-statistic, returned as a scalar value.

`DF1` — Numerator degrees of freedom for `F`
scalar value

Numerator degrees of freedom for the *F*-statistic `F`, returned as a scalar value.

- If you test the null hypothesis H_0: $H\beta = 0$ or H_0: $H\beta = C$, then `DF1` is equal to the number of linearly independent rows in `H`.
- If you test the null hypothesis H_0: $H\beta + KB = C$, then `DF1` is equal to the number of linearly independent rows in `[H,K]`.

`DF2` — Denominator degrees of freedom for `F`
scalar value

Denominator degrees of freedom for the *F*-statistic `F`, returned as a scalar value. The value of `DF2` depends on the option specified by the `'DFMethod'` name-value pair argument.

Examples

Test the Significance of Coefficients

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects intercept grouped by `factory`, to

account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, $newprocess_{ij}$ indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Test if there is any significant difference between supplier C and supplier B.

```
H = [0,0,0,0,1,-1];
```

```
[pVal,F,DF1,DF2] = coefTest(glme,H)
```

```
pVal = 0.2793
```

```
F = 1.1842
```

```
DF1 = 1
```

```
DF2 = 94
```

The large p-value indicates that there is no significant difference between supplier C and supplier B at the 5% significance level. Here, `coefTest` also returns the F-statistic, the numerator degrees of freedom, and the approximate denominator degrees of freedom.

Test if there is any significant difference between supplier A and supplier B.

If you specify the `'DummyVarCoding'` name-value pair argument as `'effects'` when fitting the model using `fitglme`, then

$$\beta_A + \beta_B + \beta_C = 0,$$

where β_A, β_B, and β_C correspond to suppliers A, B, and C, respectively. β_A is the effect of A minus the average effect of A, B, and C. To determine the contrast matrix corresponding to a test between supplier A and supplier B,

$$\beta_B - \beta_A = \beta_B - (-\beta_B - \beta_C) = 2\beta_B + \beta_C .$$

From the output of `disp(glme)`, column 5 of the contrast matrix corresponds to β_C, and column 6 corresponds to β_B. Therefore, the contrast matrix for this test is specified as `H = [0,0,0,0,1,2]`.

```
H = [0,0,0,0,1,2];

[pVal,F,DF1,DF2] = coefTest(glme,H)

pVal = 0.6177

F = 0.2508

DF1 = 1

DF2 = 94
```

The large *p*-value indicates that there is no significant difference between supplier A and supplier B at the 5% significance level.

References

[1] Booth, J.G., and J.P. Hobert. “Standard Errors of Prediction in Generalized Linear Mixed Models.” *Journal of the American Statistical Association*, Vol. 93, 1998, pp. 262–272.

See Also

`GeneralizedLinearMixedModel` | `anova` | `coefCI` | `covarianceParameters` | `fixedEffects` | `randomEffects`

coefTest

Package:

Linear hypothesis test on linear regression model coefficients

Syntax

```
p = coefTest(mdl)
p = coefTest(mdl,H)
p = coefTest(mdl,H,C)
[p,F] = coefTest( ___ )
[p,F,r] = coefTest( ___ )
```

Description

`p = coefTest(mdl)` computes the *p*-value for an *F*-test that all coefficient estimates in `mdl`, except for the intercept term, are zero.

`p = coefTest(mdl,H)` performs an *F*-test that $H \times B = 0$, where B represents the coefficient vector. Use `H` to specify the coefficients to include in the *F*-test.

`p = coefTest(mdl,H,C)` performs an *F*-test that $H \times B = C$.

`[p,F] = coefTest( ___ )` also returns the *F*-test statistic `F` using any of the input argument combinations in previous syntaxes.

`[p,F,r] = coefTest( ___ )` also returns the numerator degrees of freedom `r` for the test.

Examples

Test Significance of Linear Regression Model

Fit a linear regression model and test the coefficients of the fitted model to see if they are zero.

Load the `carsmall` data set and create a table in which the `Model_Year` predictor is categorical.

```
load carsmall
Model_Year = categorical(Model_Year);
tbl = table(MPG,Weight,Model_Year);
```

Fit a linear regression model of mileage as a function of the weight, weight squared, and model year.

```
mdl = fitlm(tbl,'MPG ~ Model_Year + Weight^2')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Weight + Model_Year + Weight^2

Estimated Coefficients:
                       Estimate         SE         tStat       pValue
```

```
                    ___________    ___________    _______    ___________

    (Intercept)           54.206         4.7117     11.505    2.6648e-19
    Weight             -0.016404      0.0031249    -5.2493    1.0283e-06
    Model_Year_76         2.0887        0.71491     2.9215     0.0044137
    Model_Year_82         8.1864        0.81531     10.041    2.6364e-16
    Weight^2           1.5573e-06    4.9454e-07      3.149     0.0022303


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.78
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 172, p-value = 5.52e-41
```

The last line of the model display shows the *F*-statistic value of the regression model and the corresponding *p*-value. The small *p*-value indicates that the model fits significantly better than a degenerate model consisting of only an intercept term. You can return these two values by using `coefTest`.

```
[p,F] = coefTest(mdl)
```

```
p = 5.5208e-41
```

```
F = 171.8844
```

Test Significance of Linear Model Coefficient

Fit a linear regression model and test the significance of a specified coefficient in the fitted model by using `coefTest`. You can also use `anova` to test the significance of each predictor in the model.

Load the `carsmall` data set and create a table in which the `Model_Year` predictor is categorical.

```
load carsmall
Model_Year = categorical(Model_Year);
tbl = table(MPG,Acceleration,Weight,Model_Year);
```

Fit a linear regression model of mileage as a function of the weight, weight squared, and model year.

```
mdl = fitlm(tbl,'MPG ~ Acceleration + Model_Year + Weight')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Acceleration + Weight + Model_Year

Estimated Coefficients:
                       Estimate         SE         tStat        pValue
                    ___________    ___________    ________    ___________

    (Intercept)           40.523         2.5293      16.021    5.8302e-28
    Acceleration       -0.023438        0.11353    -0.20644       0.83692
    Weight             -0.0066799    0.00045796     -14.586    2.5314e-25
    Model_Year_76         1.9898        0.80696      2.4657      0.015591
    Model_Year_82         7.9661        0.89745      8.8763    6.7725e-14


Number of observations: 94, Error degrees of freedom: 89
```

```
Root Mean Squared Error: 2.93
R-squared: 0.873,  Adjusted R-Squared: 0.867
F-statistic vs. constant model: 153, p-value = 5.86e-39
```

The model display includes the *p*-value for the *t*-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero.

You can examine the significance of the coefficient using `coefTest`. For example, test the significance of the `Acceleration` coefficient. According to the model display, `Acceleration` is the second predictor. Specify the coefficient by using a numeric index vector.

```
[p_Acceleration,F_Acceleration,r_Acceleration] = coefTest(mdl,[0 1 0 0 0])
```

```
p_Acceleration = 0.8369

F_Acceleration = 0.0426

r_Acceleration = 1
```

`p_Acceleration` is the *p*-value corresponding to the *F*-statistic value `F_Acceleration`, and `r_Acceleration` is the numerator degrees of freedom for the *F*-test. The returned *p*-value indicates that `Acceleration` is not statistically significant in the fitted model. Note that `p_Acceleration` is equal to the *p*-value of *t*-statistic (`tStat`) in the model display, and `F_Acceleration` is the square of `tStat`.

Test the significance of the categorical predictor `Model_Year`. Instead of testing `Model_Year_76` and `Model_Year_82` separately, you can perform a single test for the categorical predictor `Model_Year`. Specify `Model_Year_76` and `Model_Year_82` by using a numeric index matrix.

```
[p_Model_Year,F_Model_Year,r_Model_Year] = coefTest(mdl,[0 0 0 1 0; 0 0 0 0 1])
```

```
p_Model_Year = 2.7408e-14

F_Model_Year = 45.2691

r_Model_Year = 2
```

The returned *p*-value indicates that `Model_Year` is statistically significant in the fitted model.

You can also return these values by using `anova`.

```
anova(mdl)
```

```
ans=4×5 table
                    SumSq     DF    MeanSq        F          pValue
                   _______    __    _______    ________    __________

    Acceleration    0.36613     1    0.36613    0.042618       0.83692
    Weight           1827.7     1     1827.7      212.75    2.5314e-25
    Model_Year       777.81     2      388.9      45.269    2.7408e-14
    Error            764.59    89      8.591
```

Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`H` — Hypothesis matrix
numeric index matrix

Hypothesis matrix, specified as an `r`-by-`s` numeric index matrix, where `r` is the number of coefficients to include in an *F*-test, and `s` is the total number of coefficients.

- If you specify `H`, then the output `p` is the *p*-value for an *F*-test that $H \times B = 0$, where B represents the coefficient vector.
- If you specify `H` and `C`, then the output `p` is the *p*-value for an *F*-test that $H \times B = C$.

Example: `[1 0 0 0 0]` tests the first coefficient among five coefficients.

Data Types: `single` | `double`

`C` — Hypothesized value
numeric vector

Hypothesized value for testing the null hypothesis, specified as a numeric vector with the same number of rows as `H`.

If you specify `H` and `C`, then the output `p` is the *p*-value for an *F*-test that $H \times B = C$, where B represents the coefficient vector.

Data Types: `single` | `double`

Output Arguments

`p` — *p*-value for *F*-test
numeric value in the range [0,1]

p-value for the *F*-test, returned as a numeric value in the range [0,1].

`F` — Value of test statistic for *F*-test
numeric value

Value of the test statistic for the *F*-test, returned as a numeric value.

`r` — Numerator degrees of freedom for *F*-test
positive integer

Numerator degrees of freedom for the *F*-test, returned as a positive integer. The *F*-statistic has `r` degrees of freedom in the numerator and `mdl.DFE` degrees of freedom in the denominator.

Algorithms

The *p*-value, *F*-statistic, and numerator degrees of freedom are valid under these assumptions:

- The data comes from a model represented by the formula in the `Formula` property of the fitted model.
- The observations are independent, conditional on the predictor values.

Under these assumptions, let β represent the (unknown) coefficient vector of the linear regression. Suppose H is a full-rank matrix of size r-by-s, where r is the number of coefficients to include in an F-test, and s is the total number of coefficients. Let c be a column vector with r rows. The following is a test statistic for the hypothesis that $H\beta = c$:

$$F = \left(H\widehat{\beta} - c\right)'(HVH')^{-1}\left(H\widehat{\beta} - c\right).$$

Here $\widehat{\beta}$ is the estimate of the coefficient vector β, stored in the `Coefficients` property, and V is the estimated covariance of the coefficient estimates, stored in the `CoefficientCovariance` property. When the hypothesis is true, the test statistic F has an "F Distribution" on page B-45 with r and u degrees of freedom, where u is the degrees of freedom for error, stored in the `DFE` property.

Alternative Functionality

- The values of commonly used test statistics are available in the `Coefficients` property of a fitted model.
- `anova` provides tests for each model predictor and groups of predictors.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `anova` | `coefCI` | `dwtest` | `linhyptest`

Topics

"F-statistic and t-statistic" on page 11-72
"Interpret Linear Regression Results" on page 11-50
"Linear Regression Workflow" on page 11-35
"Linear Regression" on page 11-9

Introduced in R2012a

coefTest

Class: `LinearMixedModel`

Hypothesis test on fixed and random effects of linear mixed-effects model

Syntax

```
pVal = coefTest(lme)
pVal = coefTest(lme,H)
pVal = coefTest(lme,H,C)
pVal = coefTest(lme,H,C,Name,Value)
[pVal,F,DF1,DF2] = coefTest( ___ )
```

Description

`pVal = coefTest(lme)` returns the *p*-value for an *F*-test that all fixed-effects coefficients except for the intercept are 0.

`pVal = coefTest(lme,H)` returns the *p*-value for an *F*-test on fixed-effects coefficients of linear mixed-effects model `lme`, using the contrast matrix `H`. It tests the null hypothesis that H_0: $H\beta = 0$, where β is the fixed-effects vector.

`pVal = coefTest(lme,H,C)` returns the *p*-value for an *F*-test on fixed-effects coefficients of the linear mixed-effects model `lme`, using the contrast matrix `H`. It tests the null hypothesis that H_0: $H\beta = C$, where β is the fixed-effects vector.

`pVal = coefTest(lme,H,C,Name,Value)` returns the *p*-value for an *F*-test on the fixed- and/or random-effects coefficients of the linear mixed-effects model `lme`, with additional options specified by one or more name-value pair arguments. For example, `'REContrast',K` tells `coefTest` to test the null hypothesis that H_0: $H\beta + KB = C$, where β is the fixed-effects vector and B is the random-effects vector.

`[pVal,F,DF1,DF2] = coefTest( ___ )` also returns the *F*-statistic `F`, and the numerator and denominator degrees of freedom for `F`, respectively `DF1` and `DF2`.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`H` — Fixed-effects contrasts

m-by-*p* matrix

Fixed-effects contrasts, specified as an *m*-by-*p* matrix, where *p* is the number of fixed-effects coefficients in `lme`. Each row of `H` represents one contrast. The columns of `H` (left to right) correspond to the rows of the *p*-by-1 fixed-effects vector `beta` (top to bottom), returned by the `fixedEffects` method.

Data Types: `single` | `double`

`C` — Hypothesized value
m-by-1 vector

Hypothesized value for testing the null hypothesis `H*beta = C`, specified as an *m*-by-1 matrix. Here, `beta` is the vector of fixed-effects estimates returned by the `fixedEffects` method.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`DFMethod` — Method for computing approximate denominator degrees of freedom
`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing the approximate denominator degrees of freedom for the *F*-test, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

`REContrast` — Random-effects contrasts
m-by-*q* matrix

Random-effects contrasts, specified as the comma-separated pair consisting of `'REContrast'` and an *m*-by-*q* matrix `K`, where *q* is the number of random effects parameters in `lme`. The columns of `K` (left to right) correspond to the rows of the random-effects best linear unbiased predictor vector `B` (top to bottom), returned by the `randomEffects` method.

Data Types: `single` | `double`

Output Arguments

`pVal` — *p*-value
scalar value

p-value for the *F*-test on the fixed and/or random-effects coefficients of the linear mixed-effects model `lme`, returned as a scalar value.

`F` — *F*-statistic
scalar value

F-statistic, returned as a scalar value.

DF1 — Numerator degrees of freedom for F
scalar value

Numerator degrees of freedom for F, returned as a scalar value.

- If you test the null hypothesis H_0: $H\beta = 0$, or H_0: $H\beta = C$, then DF1 is equal to the number of linearly independent rows in H.
- If you test the null hypothesis H_0: $H\beta + KB = C$, then DF1 is equal to the number of linearly independent rows in [H,K].

DF2 — Denominator degrees of freedom for F
scalar value

Denominator degrees of freedom for F, returned as a scalar value. The value of DF2 depends on the option you select for DFMethod.

Examples

Test Fixed-Effects Coefficients for Categorical Data

Load the sample data.

```
load('shift.mat')
```

The data shows the absolute deviations from the target quality characteristic measured from the products that five operators manufacture during three different shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the absolute deviation of the quality characteristics from the target value. This is simulated data.

Shift and Operator are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if there is significant difference in the performance according to the time of the shift.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             15
    Fixed effects coefficients          3
    Random effects coefficients         5
    Covariance parameters               2

Formula:
    QCDev ~ 1 + Shift + (1 | Operator)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    59.012    62.552    -24.506          49.012
```

```
Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE         tStat       DF    pValue
    {'(Intercept)'  }        3.1196     0.88681      3.5178    12    0.0042407
    {'Shift_Morning'}       -0.3868     0.48344    -0.80009    12      0.43921
    {'Shift_Night'  }        1.9856     0.48344      4.1072    12    0.0014535


    Lower      Upper
     1.1874     5.0518
    -1.4401    0.66653
    0.93227     3.0389

Random effects covariance parameters (95% CIs):
Group: Operator (5 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        1.8297


    Lower      Upper
    0.94915    3.5272

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.76439     0.49315    1.1848
```

Test if all fixed-effects coefficients except for the intercept are 0.

```
pVal = coefTest(lme)
```

```
pVal = 7.5956e-04
```

The small *p*-value indicates that not all fixed-effects coefficients are 0.

Test the significance of the `Shift` term using a contrast matrix.

```
H = [0 1 0; 0 0 1];
pVal = coefTest(lme,H)
```

```
pVal = 7.5956e-04
```

Test the significance of the `Shift` term using the `anova` method.

```
anova(lme)
```

```
ans =
    ANOVA marginal tests: DFMethod = 'Residual'

    Term                   FStat     DF1    DF2    pValue
    {'(Intercept)'}        12.375    1      12      0.0042407
    {'Shift'      }        13.864    2      12     0.00075956
```

The *p*-value for `Shift`, 0.00075956, is the same as the *p*-value of the previous hypothesis test.

Test if there is any difference between the evening and morning shifts.

```
pVal = coefTest(lme,[0 1 -1])
```

```
pVal = 3.6147e-04
```

This small *p*-value indicates that the performance of the operators are not the same in the morning and the evening shifts.

Hypothesis Tests for Fixed-Effects Coefficients

Load the sample data.

```
load('weight.mat')
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             120
    Fixed effects coefficients           9
    Random effects coefficients         40
    Covariance parameters                4

Formula:
    y ~ 1 + InitialWeight + Program*Week + (1 + Week | Subject)

Model fit statistics:
    AIC        BIC       LogLikelihood    Deviance
    -22.981    13.257    24.49            -48.981

Fixed effects coefficients (95% CIs):
    Name                      Estimate     SE           tStat       DF
    {'(Intercept)'   }          0.66105      0.25892      2.5531    111
    {'InitialWeight' }        0.0031879    0.0013814      2.3078    111
    {'Program_B'     }          0.36079      0.13139       2.746    111
    {'Program_C'     }        -0.033263      0.13117    -0.25358    111
    {'Program_D'     }          0.11317      0.13132     0.86175    111
    {'Week'          }           0.1732     0.067454      2.5677    111
    {'Program_B:Week'}         0.038771     0.095394     0.40644    111
    {'Program_C:Week'}         0.030543     0.095394     0.32018    111
    {'Program_D:Week'}         0.033114     0.095394     0.34713    111
```

```
    pValue       Lower         Upper
     0.012034       0.14798       1.1741
     0.022863    0.00045067    0.0059252
    0.0070394       0.10044      0.62113
      0.80029      -0.29319      0.22666
      0.39068      -0.14706       0.3734
     0.011567      0.039536      0.30686
      0.68521      -0.15026       0.2278
      0.74944      -0.15849      0.21957
      0.72915      -0.15592      0.22214

Random effects covariance parameters (95% CIs):
Group: Subject (20 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std' }       0.18407
    {'Week'       }        {'(Intercept)'}        {'corr'}       0.66841
    {'Week'       }        {'Week'       }        {'std' }       0.15033


    Lower      Upper
    0.12281    0.27587
    0.21076    0.88573
    0.11004    0.20537

Group: Error
    Name               Estimate    Lower       Upper
    {'Res Std'}        0.10261     0.087882    0.11981
```

Test for the significance of the interaction between `Program` and `Week`.

```
H = [0 0 0 0 0 0 1 0 0;
     0 0 0 0 0 0 0 1 0;
     0 0 0 0 0 0 0 0 1];
pVal = coefTest(lme,H)
```

```
pVal = 0.9775
```

The high *p*-value indicates that the interaction between `Program` and `Week` is not statistically significant.

Now, test whether all coefficients involving `Program` are 0.

```
H = [0 0 1 0 0 0 0 0 0;
     0 0 0 1 0 0 0 0 0;
     0 0 0 0 1 0 0 0 0;
     0 0 0 0 0 0 1 0 0;
     0 0 0 0 0 0 0 1 0;
     0 0 0 0 0 0 0 0 1];
C = [0;0;0;0;0;0];
pVal = coefTest(lme,H,C)
```

```
pVal = 0.0274
```

The *p*-value of 0.0274 indicates that not all coefficients involving `Program` are zero.

Hypothesis Tests for Fixed- and Random-Effects Coefficients

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the CDC).

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses and region as the predictor variable, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the response variable, `FluRate`, the nominal variable, `Region`, that shows which region each estimate is from, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate',...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Fit a linear mixed-effects model with fixed effects for the region and a random intercept that varies by `Date`.

```
lme = fitlme(flu2,'FluRate ~ 1 + Region + (1|Date)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             468
    Fixed effects coefficients           9
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    FluRate ~ 1 + Region + (1 | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    318.71    364.35    -148.36          296.71

Fixed effects coefficients (95% CIs):
    Name                        Estimate    SE          tStat      DF
    {'(Intercept)'     }          1.2233    0.096678     12.654    459
    {'Region_MidAtl'   }        0.010192    0.052221    0.19518    459
    {'Region_ENCentral'}        0.051923    0.052221     0.9943    459
    {'Region_WNCentral'}         0.23687    0.052221     4.5359    459
    {'Region_SAtl'     }        0.075481    0.052221     1.4454    459
    {'Region_ESCentral'}         0.33917    0.052221      6.495    459
    {'Region_WSCentral'}           0.069    0.052221     1.3213    459
    {'Region_Mtn'      }        0.046673    0.052221    0.89377    459
    {'Region_Pac'      }        -0.16013    0.052221    -3.0665    459


    pValue        Lower        Upper
     1.085e-31       1.0334       1.4133
       0.84534    -0.092429      0.11281
        0.3206    -0.050698      0.15454
    7.3324e-06      0.13424      0.33949
```

```
       0.14902     -0.02714       0.1781
    2.1623e-10      0.23655      0.44179
       0.18705    -0.033621      0.17162
       0.37191    -0.055948      0.14929
     0.0022936     -0.26276    -0.057514

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       0.6443


    Lower     Upper
    0.5297    0.78368

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.26627     0.24878    0.285
```

Test the hypothesis that the random effects-term for week 10/9/2005 is zero.

```
[~,~,STATS] = randomEffects(lme); % Compute the random-effects statistics (STATS)
STATS.Level = nominal(STATS.Level);
K = zeros(length(STATS),1);
K(STATS.Level == '10/9/2005') = 1;
pVal = coefTest(lme,[0 0 0 0 0 0 0 0 0],0,'REContrast',K')
```

```
pVal = 0.1692
```

Refit the model this time with a random intercept and slope.

```
lme = fitlme(flu2,'FluRate ~ 1 + Region + (1 + Region|Date)');
```

Test the hypothesis that the combined coefficient of region `WNCentral` for week 10/9/2005 is zero.

```
[~,~,STATS] = randomEffects(lme); STATS.Level = nominal(STATS.Level);
K = zeros(length(STATS),1);
K(STATS.Level == '10/9/2005' & flu2.Region == 'WNCentral') = 1;
pVal = coefTest(lme,[0 0 0 1 0 0 0 0 0],0,'REContrast',K')
```

```
pVal = 1.1939e-12
```

Also return the *F*-statistic with the numerator and denominator degrees of freedom.

```
[pVal,F,DF1,DF2] = coefTest(lme,[0 0 0 1 0 0 0 0 0],0,'REContrast',K')
```

```
pVal = 1.1939e-12

F = 53.4396

DF1 = 1

DF2 = 459
```

Repeat the test using the Satterthwaite approximation for the denominator degrees of freedom.

```
[pVal,F,DF1,DF2] = coefTest(lme,[0 0 0 1 0 0 0 0 0],0,'REContrast',K',...
            'DFMethod','satterthwaite')
```

```
pVal = NaN

F = 53.4396

DF1 = 1

DF2 = 0
```

See Also

LinearMixedModel | anova | coefCI

coefTest

Class: `NonLinearModel`

Linear hypothesis test on nonlinear regression model coefficients

Syntax

```
p = coefTest(mdl)
p = coefTest(mdl,H)
p = coefTest(mdl,H,C)
[p,F] = coefTest(mdl,...)
[p,F,r] = coefTest(mdl,...)
```

Description

`p = coefTest(mdl)` computes the *p*-value for an *F* test that all coefficient estimates in `mdl` are zero.

`p = coefTest(mdl,H)` performs an *F* test that `H*B = 0`, where `B` represents the coefficient vector.

`p = coefTest(mdl,H,C)` performs an *F* test that `H*B = C`.

`[p,F] = coefTest(mdl,...)` returns the *F* test statistic.

`[p,F,r] = coefTest(mdl,...)` returns the numerator degrees of freedom for the test.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`H`

Numeric matrix having one column for each coefficient in the model. When `H` is an input, the output `p` is the *p*-value for an *F* test that `H*B = 0`, where `B` represents the coefficient vector.

`C`

Numeric vector with the same number of rows as `H`. When `C` is an input, the output `p` is the *p*-value for an *F* test that `H*B = C`, where `B` represents the coefficient vector.

Output Arguments

`p`

p-value of the *F* test (see “More About” on page 33-702).

`F`

Value of the test statistic for the *F* test (see “More About” on page 33-702).

`r`

Numerator degrees of freedom for the *F* test (see "More About" on page 33-702). The *F* statistic has `r` degrees of freedom in the numerator and `mdl.DFE` degrees of freedom in the denominator.

Examples

Test Nonlinear Regression Model Coefficients

Make a nonlinear model of mileage as a function of the weight from the `carsmall` data set. Test the coefficients to see if all should be zero.

Create an exponential model of car mileage as a function of weight from the `carsmall` data. Scale the weight by a factor of 1000 so all the variables are roughly equal in size.

```
load carsmall
X = Weight;
y = MPG;
modelfun = 'y ~ b1 + b2*exp(-b3*x/1000)';
beta0 = [1 1 1];
mdl = fitnlm(X,y,modelfun,beta0);
```

Test the model for significant differences from a constant model.

```
p = coefTest(mdl)
```

```
p = 1.3708e-36
```

There is no doubt that the model contains nonzero terms.

More About

Test Statistics

The *p*-value, *F* statistic, and numerator degrees of freedom are valid under these assumptions:

- The data comes from a normal distribution.
- The entries are independent.

Suppose these assumptions hold. Let β represent the unknown coefficient vector of the linear regression. Suppose H is a full-rank matrix of size r-by-s, where s is the number of terms in β. Let c be a vector the same size as β. The following is a test statistic for the hypothesis that $H\beta = c$:

$$F = \left(H\widehat{\beta} - c\right)'(HVH')^{-1}\left(H\widehat{\beta} - c\right).$$

Here $\widehat{\beta}$ is the estimate of the coefficient vector β in `mdl.Coefs`, and V is the estimated covariance of the coefficient estimates in `mdl.CoefCov`. When the hypothesis is true, the test statistic F has an "F Distribution" on page B-45 with r and u degrees of freedom.

Alternatives

The values of commonly used test statistics are available in the *`mdl`*`.Coefficients` table.

See Also

`NonLinearModel`

Topics

“Nonlinear Regression” on page 13-2

coeftest

Class: `RepeatedMeasuresModel`

Linear hypothesis test on coefficients of repeated measures model

Syntax

```
tbl = coeftest(rm,A,C,D)
```

Description

`tbl = coeftest(rm,A,C,D)` returns a table `tbl` containing the multivariate analysis of variance (manova) for the repeated measures model `rm`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`A` — Specification representing between-subjects model

a-by-*p* matrix

Specification representing the between-subjects model, specified as an *a*-by-*p* numeric matrix, with rank $a \le p$.

Data Types: `single` | `double`

`C` — Specification representing within-subjects hypothesis

r-by-*c* matrix

Specification representing the within-subjects (within time) hypotheses, specified as an *r*-by-*c* numeric matrix, with rank $c \le r \le n - p$.

Data Types: `single` | `double`

`D` — Hypothesized value

0 (default) | scalar value | *a*-by-*c* matrix

Hypothesized value, specified as a scalar value or an *a*-by-*c* matrix.

Data Types: `single` | `double`

Output Arguments

`tbl` — Results of multivariate analysis of variance

table

Results of multivariate analysis of variance for the repeated measures model `rm`, returned as a table containing the following columns.

`Statistic`	Type of test statistic used
`Value`	Value of the corresponding test statistic
`F`	*F*-statistic value
`RSquare`	Measure of variance explained
`df1`	Numerator degrees of freedom for the *F*-statistic
`df2`	Denominator degrees of freedom for the *F*-statistic
`pValue`	*p*-value associated with the test statistic value

Examples

Test Coefficients for First and Last Repeated Measures

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Test that the coefficients of all terms in the between-subjects model are the same for the first and last repeated measurement variable.

```
coeftest(rm,eye(8),[1 0 0 0 0 0 0 -1]')
```

```
ans=4×7 table
    Statistic     Value       F       RSquare    df1    df2    pValue
    _________    _______    ______    _______    ___    ___    _______

    Pillai        0.3355    1.3884    0.3355      8     22     0.25567
    Wilks         0.6645    1.3884    0.3355      8     22     0.25567
    Hotelling    0.50488    1.3884    0.3355      8     22     0.25567
    Roy          0.50488    1.3884    0.3355      8     22     0.25567
```

The *p*-value of 0.25567 indicates that there is not enough statistical evidence to conclude that the coefficients of all terms in the between-subjects model for the first and last repeated measures variable are different.

Tips

- This test is defined as `A*B*C = D`, where `B` is the matrix of coefficients in the repeated measures model. `A` and `C` are numeric matrices of the proper size for this multiplication. `D` is a scalar or numeric matrix of the proper size. The default is `D = 0`.

See Also

`fitrm` | `manova`

combine

Class: `CompactTreeBagger`

Combine two ensembles

Syntax

```
B1 = combine(B1,B2)
```

Description

`B1 = combine(B1,B2)` appends decision trees from ensemble `B2` to those stored in `B1` and returns ensemble `B1`. This method requires that the class and variable names be identical in both ensembles.

See Also

`append`

combnk

(Not recommended) Enumeration of combinations

Note `combnk` is not recommended. Use the MATLAB® function `nchoosek` instead. For more information, see “Compatibility Considerations”.

Syntax

```
C = combnk(v,k)
```

Description

`C = combnk(v,k)` returns a matrix containing all possible combinations of the elements of vector `v` taken `k` at a time. Matrix `C` has `k` columns and `n!/((n – k)! k!)` rows, where `n` is the number of observations in `v`.

Examples

Combinations of Four Characters

Create a character array of every four-letter combination of the characters in the word `'tendril'`.

```
C = combnk('tendril',4);
```

`C` is a 35-by-4 character array.

Display the last five combinations in the list.

```
last5 = C(31:35,:)
```

```
last5 = 5x4 char array
    'tedr'
    'tenl'
    'teni'
    'tenr'
    'tend'
```

Combinations of Elements from a Numeric Vector

List all two-number combinations of the numbers one through four.

```
C = combnk(1:4,2)
```

```
C = 6×2

     3     4
```

```
     2     4
     2     3
     1     4
     1     3
     1     2
```

Because `1:4` is a vector of doubles, `C` is a matrix of doubles.

Input Arguments

v — Set of all elements

vector

Set of all elements, specified as a vector.

Example: `[1 2 3 4 5]`

Example: `'abcd'`

Data Types: `single` | `double` | `logical` | `char`

k — Number of selected choices

nonnegative integer scalar

Number of elements to select, specified as a nonnegative integer scalar. `k` can be any numeric type, but must be real.

There are no restrictions on combining inputs of different types for `combnk(v,k)`.

Example: `3`

Data Types: `single` | `double`

Output Arguments

C — All combinations

matrix

All combinations of `v`, returned as a matrix of the same type as `v`. `C` has `k` columns and `n!/((n - k)! k!)` rows, where `n` is the number of observations in `v`.

Each row of `C` contains a combination of `k` items selected from `v`. The elements in each row of `C` are listed in the same order as they appear in `v`.

If `k` is larger than `n`, then `C` is an empty matrix.

Limitations

`combnk` is practical only for situations where `v` has fewer than 15 observations.

Compatibility Considerations

`combnk` is not recommended

Not recommended starting in R2020b

`combnk` is not recommended. Use the MATLAB function `nchoosek` instead. There are no plans to remove `combnk`.

To update your code, change instances of the function name `combnk` to `nchoosek`. You do not need to change the input arguments. For example, use `C = nchoosek(v,k)`. The output `C` contains all possible combinations of the elements of vector `v` taken `k` at a time. Note that `C` from `nchoosek` can have a different order compared to the output from `combnk`.

The `nchoosek` function has several advantages over the `combnk` function.

- `nchoosek` also returns the binomial coefficient when the first input argument is a scalar value.
- `nchoosek` has extended functionality using MATLAB Coder.
- `nchoosek` is faster than `combnk`.

See Also

`nchoosek` | `perms` | `randperm`

Introduced before R2006a

compact

Reduce size of machine learning model

Syntax

```
CompactMdl = compact(Mdl)
```

Description

`CompactMdl = compact(Mdl)` returns a compact model (`CompactMdl`), the compact version of the trained machine learning model `Mdl`.

`CompactMdl` does not contain the training data, whereas `Mdl` contains the training data in its `X` and `Y` properties. Therefore, although you can predict class labels using `CompactMdl`, you cannot perform tasks such as cross-validation with the compact model.

Examples

Reduce Size of Naive Bayes Classifier

Reduce the size of a full naive Bayes classifier by removing the training data. Full naive Bayes classifiers hold the training data. You can use a compact naive Bayes classifier to improve memory efficiency.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'b','g'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
        DistributionNames: {1x32 cell}
   DistributionParameters: {2x32 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Reduce the size of the naive Bayes classifier.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: 'none'
        DistributionNames: {1x32 cell}
   DistributionParameters: {2x32 cell}


  Properties, Methods
```

`CMdl` is a trained `CompactClassificationNaiveBayes` classifier.

Display the amount of memory used by each classifier.

```
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class

  CMdl      1x1               15060  classreg.learning.classif.CompactClassificationNaiveBayes
  Mdl       1x1              111174  ClassificationNaiveBayes
```

The full naive Bayes classifier (`Mdl`) is more than seven times larger than the compact naive Bayes classifier (`CMdl`).

To label new observations efficiently, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CMdl` and new predictor values to `predict`.

Reduce Size of SVM Classifier

Reduce the size of a full support vector machine (SVM) classifier by removing the training data. Full SVM classifiers (that is, `ClassificationSVM` classifiers) hold the training data. To improve efficiency, use a smaller classifier.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Standardize the predictor data and specify the order of the classes.

```
SVMModel = fitcsvm(X,Y,'Standardize',true,...
    'ClassNames',{'b','g'})
```

```
SVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: 'none'
         NumObservations: 351
```

```
                    Alpha: [90x1 double]
                     Bias: -0.1343
          KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
            BoxConstraints: [351x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [351x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`SVMModel` is a `ClassificationSVM` classifier.

Reduce the size of the SVM classifier.

```
CompactSVMModel = compact(SVMModel)
```

```
CompactSVMModel =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
                    Alpha: [90x1 double]
                     Bias: -0.1343
          KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           SupportVectors: [90x34 double]
      SupportVectorLabels: [90x1 double]


  Properties, Methods
```

`CompactSVMModel` is a `CompactClassificationSVM` classifier.

Display the amount of memory used by each classifier.

```
whos('SVMModel','CompactSVMModel')
```

```
  Name                Size             Bytes  Class

  CompactSVMModel     1x1              31058  classreg.learning.classif.CompactClassificationSVM
  SVMModel            1x1             141148  ClassificationSVM
```

The full SVM classifier (`SVMModel`) is more than four times larger than the compact SVM classifier (`CompactSVMModel`).

To label new observations efficiently, you can remove `SVMModel` from the MATLAB® Workspace, and then pass `CompactSVMModel` and new predictor values to `predict`.

To further reduce the size of the compact SVM classifier, use the `discardSupportVectors` function to discard support vectors.

Reduce Size of Generalized Additive Model

Reduce the size of a full generalized additive model (GAM) for regression by removing the training data. Full models hold the training data. You can use a compact model to improve memory efficiency.

Load the `carbig` data set.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a GAM using `X` and `Y`.

```
Mdl = fitrgam(X,Y)
```

```
Mdl =
  RegressionGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                Intercept: 26.9442
          NumObservations: 398


  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object.

Reduce the size of the model.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactRegressionGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                Intercept: 26.9442


  Properties, Methods
```

`CMdl` is a `CompactRegressionGAM` model object.

Display the amount of memory used by each regression model.

```
whos('Mdl','CMdl')
```

```
  Name      Size             Bytes  Class                                    Attributes
```

```
  CMdl      1x1             578154  classreg.learning.regr.CompactRegressionGAM
  Mdl       1x1             611947  RegressionGAM
```

The full model (`Mdl`) is larger than the compact model (`CMdl`).

To efficiently predict responses for new observations, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CMdl` and new predictor values to `predict`.

Input Arguments

`Mdl` — Machine learning model

full regression model object | full classification model object

Machine learning model, specified as a full regression or classification model object, as given in the following tables of supported models.

Regression Model Object

Model	Full Regression Model Object
Generalized additive model	`RegressionGAM`
Neural network model	`RegressionNeuralNetwork`

Classification Model Object

Model	Full Classification Model Object
Generalized additive model	`ClassificationGAM`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

Output Arguments

`CompactMdl` — Compact machine learning model

compact regression model object | compact classification model object

Compact machine learning model, returned as one of the compact model objects in the following tables, depending on the input model `Mdl`.

Regression Model Object

Model	Full Model (`Mdl`)	Compact Model (`CompactMdl`)
Generalized additive model	`RegressionGAM`	`CompactRegressionGAM`
Neural network model	`RegressionNeuralNetwork`	`CompactRegressionNeuralNetwork`

Classification Model Object

Model	Full Model (`Mdl`)	Compact Model (`CompactMdl`)
Generalized additive model	`ClassificationGAM`	`CompactClassificationGAM`
Naive Bayes model	`ClassificationNaiveBayes`	`CompactClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`	`CompactClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`CompactClassificationSVM`

See Also

Introduced in R2014a

compact

Class: `ClassificationDiscriminant`

Compact discriminant analysis classifier

Syntax

```
cobj = compact(obj)
```

Description

`cobj = compact(obj)` creates a compact version of `obj`.

Input Arguments

`obj`

Discriminant analysis classifier created using `fitcdiscr`.

Output Arguments

`cobj`

Compact classifier. `cobj` has class `CompactClassificationDiscriminant`. You can predict classifications using `cobj` exactly as you can using `obj`. However, since `cobj` does not contain training data, you cannot perform some actions, such as cross validation.

Examples

Compare the size of the discriminant analysis classifier for Fisher's iris data to the compact version of the classifier:

```
load fisheriris
fullobj = fitcdiscr(meas,species);
cobj = compact(fullobj);
b = whos('fullobj'); % b.bytes = size of fullobj
c = whos('cobj'); % c.bytes = size of cobj
[b.bytes c.bytes] % shows cobj uses 60% of the memory

ans =
       18578       11498
```

See Also

`ClassificationDiscriminant` | `fitcdiscr`

Topics

“Discriminant Analysis Classification” on page 20-2

compact

Reduce size of multiclass error-correcting output codes (ECOC) model

Syntax

```
CompactMdl = compact(Mdl)
```

Description

`CompactMdl = compact(Mdl)` returns a compact multiclass error-correcting output codes (ECOC) model (`CompactMdl`), the compact version of the trained ECOC model `Mdl`. `CompactMdl` is a `CompactClassificationECOC` object.

`CompactMdl` does not contain the training data, whereas `Mdl` contains the training data in its `X` and `Y` properties. Therefore, although you can predict class labels using `CompactMdl`, you cannot perform tasks such as cross-validation with the compact ECOC model.

Examples

Reduce Size of Full ECOC Model

Reduce the size of a full ECOC model by removing the training data. Full ECOC models (`ClassificationECOC` models) hold the training data. To improve efficiency, use a smaller classifier.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
```

Train an ECOC model using SVM binary classifiers. Standardize the predictor data using an SVM template `t`, and specify the order of the classes. During training, the software uses default values for empty options in `t`.

```
t = templateSVM('Standardize',true);
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`Mdl` is a `ClassificationECOC` model.

Reduce the size of the ECOC model.

```
CompactMdl = compact(Mdl)
```

```
CompactMdl =
  CompactClassificationECOC
            ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [setosa    versicolor    virginica]
          ScoreTransform: 'none'
```

```
           BinaryLearners: {3x1 cell}
             CodingMatrix: [3x3 double]


  Properties, Methods
```

`CompactMdl` is a `CompactClassificationECOC` model. `CompactMdl` does not store all of the properties that `Mdl` stores. In particular, it does not store the training data.

Display the amount of memory each classifier uses.

```
whos('CompactMdl','Mdl')
```

```
  Name            Size            Bytes  Class                                                  A

  CompactMdl      1x1             15116  classreg.learning.classif.CompactClassificationECOC
  Mdl             1x1             28357  ClassificationECOC
```

The full ECOC model (`Mdl`) is approximately double the size of the compact ECOC model (`CompactMdl`).

To label new observations efficiently, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CompactMdl` and new predictor values to `predict`.

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `fitcecoc` | `predict`

Introduced in R2014b

compact

Compact classification ensemble

Syntax

```
cens = compact(ens)
```

Description

`cens = compact(ens)` creates a compact version of `ens`. You can predict classifications using `cens` exactly as you can using `ens`. However, since `cens` does not contain training data, you cannot perform some actions, such as cross validation.

Input Arguments

ens

A classification ensemble created with `fitcensemble`.

Output Arguments

cens

A compact classification ensemble. `cens` has class `CompactClassificationEnsemble`.

Examples

View Size of Compact Classification Ensemble

Compare the size of a classification ensemble for the Fisher iris data to the compact version of the ensemble.

Load the Fisher iris data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Create a compact version of `ens` and compare ensemble sizes.

```
cens = compact(ens);
b = whos('ens');  % b.bytes = size of ens
c = whos('cens'); % c.bytes = size of cens
[b.bytes c.bytes] % Shows cens uses less memory
```

```
ans = 1×2
```

```
447180      406462
```

The compact version of the ensemble uses less memory than the full ensemble. Note that the ensemble sizes can vary slightly, depending on your operating system.

See Also

`ClassificationTree` | `fitcensemble`

Topics

“Framework for Ensemble Learning” on page 18-31

compact

Class: `ClassificationTree`

Compact tree

Syntax

```
ctree = compact(tree)
```

Description

`ctree = compact(tree)` creates a compact version of `tree`.

Input Arguments

tree

A classification tree created using `fitctree`.

Output Arguments

ctree

A compact decision tree. `ctree` has class `CompactClassificationTree`. You can predict classifications using `ctree` exactly as you can using `tree`. However, since `ctree` does not contain training data, you cannot perform some actions, such as cross validation.

Examples

Create a Compact Classification Tree

Compare the size of the classification tree for Fisher's iris data to the compact version of the tree.

```
load fisheriris
fulltree = fitctree(meas,species);
ctree = compact(fulltree);
b = whos('fulltree'); % b.bytes = size of fulltree
c = whos('ctree'); % c.bytes = size of ctree
[b.bytes c.bytes] % shows ctree uses half the memory
```

```
ans = 1×2

       11762        5097
```

See Also

`ClassificationTree` | `CompactClassificationTree` | `fitctree` | `predict`

compact

Compact linear regression model

Syntax

```
compactMdl = compact(mdl)
```

Description

`compactMdl = compact(mdl)` returns the compact linear regression model `compactMdl`, which is the compact version of the full, fitted linear regression model `mdl`.

Examples

Compact Linear Regression Model

Fit a linear regression model to data and reduce the size of a full, fitted linear regression model by discarding the sample data and some information related to the fitting process.

Load the `largedata4reg` data set, which contains 15,000 observations and 45 predictor variables.

```
load largedata4reg
```

Fit a linear regression model to the data.

```
mdl = fitlm(X,Y);
```

Compact the model.

```
compactMdl = compact(mdl);
```

The compact model discards the original sample data and some information related to the fitting process.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans = 1×2

       81537    11408528
```

The compact model consumes less memory than the full model.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

Output Arguments

`compactMdl` — Compact linear regression model

`CompactLinearModel` object

Compact linear regression model, returned as a `CompactLinearModel` object.

A `CompactLinearModel` object consumes less memory than a `LinearModel` object because a compact model does not store the input data used to fit the model or information related to the fitting process. You can still use a compact model to predict responses using new input data, but some `LinearModel` object functions do not work with a compact model.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `fitlm` | `stepwiselm`

Topics

“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2016a

compact

Compact generalized linear regression model

Syntax

```
compactMdl = compact(mdl)
```

Description

`compactMdl = compact(mdl)` returns the compact generalized linear regression model `compactMdl`, which is the compact version of the full, fitted generalized linear regression model `mdl`.

Examples

Compact Generalized Linear Regression Model

Fit a generalized linear regression model to data and reduce the size of a full, fitted model by discarding the sample data and some information related to the fitting process.

Load the `largedata4reg` data set, which contains 15,000 observations and 45 predictor variables.

```
load largedata4reg
```

Fit a generalized linear regression model to the data using the first 15 predictor variables.

```
mdl = fitglm(X(:,1:15),Y);
```

Compact the model.

```
compactMdl = compact(mdl);
```

The compact model discards the original sample data and some information related to the fitting process, so it uses less memory than the full model.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans = 1×2

       15517     4382500
```

The compact model consumes less memory than the full model.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

Output Arguments

`compactMdl` — Compact generalized linear regression model

`CompactGeneralizedLinearModel` object

Compact generalized linear regression model, returned as a `CompactGeneralizedLinearModel` object.

A `CompactGeneralizedLinearModel` object consumes less memory than a `GeneralizedLinearModel` object because a compact model does not store the input data used to fit the model or information related to the fitting process. You can still use a compact model to predict responses using new input data, but some `GeneralizedLinearModel` object functions that require the input data do not work with a compact model.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `fitglm` | `stepwiseglm`

Introduced in R2016b

compact

Create compact regression ensemble

Syntax

```
cens = compact(ens)
```

Description

`cens = compact(ens)` creates a compact version of `ens`. You can predict regressions using `cens` exactly as you can using `ens`. However, since `cens` does not contain training data, you cannot perform some actions, such as cross validation.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`.

Output Arguments

`cens`

A compact regression ensemble. `cens` is of class `CompactRegressionEnsemble`.

Examples

View Size of Compact Regression Ensemble

Compare the size of a regression ensemble for the `carsmall` data to the size of the compact version of the ensemble.

Load the `carsmall` data set and select acceleration, number of cylinders, displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Acceleration Cylinders Displacement Horsepower Weight];
```

Train an ensemble of regression trees.

```
ens = fitrensemble(X,MPG);
```

Create a compact version of `ens` and compare ensemble sizes.

```
cens = compact(ens);
b = whos('ens');
c = whos('cens');
[b.bytes c.bytes]  % b.bytes = size of ens and c.bytes = size of cens
```

```
ans = 1×2
```

```
493630      461479
```

The compact ensemble uses less memory.

See Also

`CompactRegressionEnsemble` | `RegressionEnsemble`

compact

Class: `RegressionGP`

Create compact Gaussian process regression model

Syntax

```
cgprMdl = compact(gprMdl)
```

Description

`cgprMdl = compact(gprMdl)` returns a compact version of the trained Gaussian process regression (GPR) model, `gprMdl`.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` object.

Output Arguments

`cgprMdl` — Compact Gaussian process regression model

`CompactRegressionGP` object

Compact Gaussian process regression model, returned as a `CompactRegressionGP` object.

Examples

Compute Predictions and Regression Loss for Test Data

Generate example training data.

```
rng(1) % For reproducibility
n = 100000;
X = linspace(0,1,n)';
X = [X,X.^2];
y = 1 + X*[1;2] + sin(20*X*[1;-2]) + 0.2*randn(n,1);
```

Train a GPR model using the subset of regressors (`'sr'`) approximation method and predict using the subset of data (`'sd'`) method. Use 50 points in the active set and sparse greedy matrix approximation (`'sgma'`) method for active set selection. Because the scales of the first and second predictors are different, it is good practice to standardize the data.

```
gprMdl = fitrgp(X,y,'KernelFunction','squaredExponential','FitMethod',...
'sr','PredictMethod','sd','Basis','none','ActiveSetSize',50,...
'ActiveSetMethod','sgma','Standardize',1,'KernelParameters',[1;1]);
```

`fitrgp` accepts any combination of fitting, prediction, and active set selection methods. In some cases it might not be possible to compute the standard deviations of the predicted responses, hence the prediction intervals. See “Tips” on page 33-731. And, in some cases, using the exact method might be expensive because of the size of the training data.

Create a compact GPR object.

```
cgprMdl = compact(gprMdl);
```

Generate the test data.

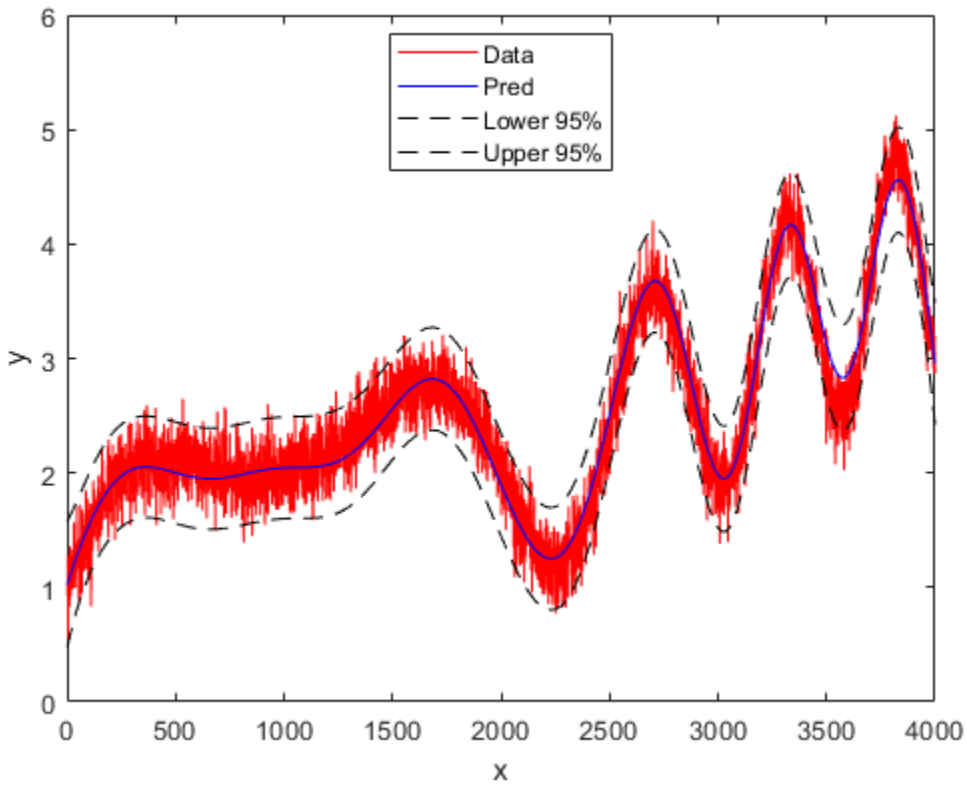
```
n = 4000;
Xnew = linspace(0,1,n)';
Xnew = [Xnew,Xnew.^2];
ynew = 1 + Xnew*[1;2] + sin(20*Xnew*[1;-2]) + 0.2*randn(n,1);
```

Use the compact object to predict the response in test data and the prediction intervals.

```
[ypred,~,yci] = predict(cgprMdl,Xnew);
```

Plot the true response, predicted response, and prediction intervals.

```
figure;
plot(ynew,'r');
hold on;
plot(ypred,'b')
plot(yci(:,1),'k--');
plot(yci(:,2),'k--');
legend('Data','Pred','Lower 95%','Upper 95%','Location','Best');
xlabel('x');
ylabel('y');
hold off
```



Compute the mean squared error loss on the test data using the trained GPR model.

```
L = loss(cgprMdl,Xnew,ynew)
```

```
L = 0.0497
```

Tips

- The compact object does not contain the training data that `fitrgp` uses in training the GPR model, `gprMdl`.
- You can use the compact object to make predictions (see `predict`) or compute the regression error (see `loss`).
- Computation of standard deviations, `ysd`, and prediction intervals, `yint`, is not supported when `PredictMethod` is `'bcd'`.
- If `gprMdl` is a `CompactRegressionGP` object, you cannot compute standard deviations, `ysd`, or prediction intervals, `yint`, for `PredictMethod` equal to `'sr'` or `'fic'`. To compute `ysd` and `yint` for `PredictMethod` equal to `'sr'` or `'fic'`, use the full regression (`RegressionGP`) object.

See Also

`CompactRegressionGP` | `fitrgp` | `loss` | `predict`

Introduced in R2015b

compact

Class: `RegressionSVM`

Compact support vector machine regression model

Syntax

```
compactMdl = compact(mdl)
```

Description

`compactMdl = compact(mdl)` returns a compact support vector machine (SVM) regression model, `compactMdl`, which is the compact version of the full, trained SVM regression model `mdl`.

`compactMdl` does not contain the training data, whereas `mdl` contains the training data in its properties `mdl.X` and `mdl.Y`.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Output Arguments

`compactMdl` — Compact SVM regression model

`CompactRegressionSVM` model

Compact SVM regression model, returned as a `CompactRegressionSVM` model.

Predict response values using `compactMdl` exactly as you would using `mdl`. However, since `compactMdl` does not contain training data, you cannot perform certain tasks, such as cross validation.

Examples

Compact an SVM Regression Model

This example shows how to reduce the size of a full, trained SVM regression model by discarding the training data and some information related to the training process.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model using a Gaussian kernel function and an automatic kernel scale. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale','auto','Standardize',true)
```

```
mdl =

  RegressionSVM
            PredictorNames: {1x8 cell}
              ResponseName: 'Var9'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635x1 double]
                     Bias: 10.8144
          KernelParameters: [1x1 struct]
                       Mu: [1x10 double]
                    Sigma: [1x10 double]
           NumObservations: 4177
            BoxConstraints: [4177x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [4177x1 logical]
                    Solver: 'SMO'


  Properties, Methods
```

Compact the model.

```
compactMdl = compact(mdl)
```

```
compactMdl =

  classreg.learning.regr.CompactRegressionSVM
            PredictorNames: {1x8 cell}
              ResponseName: 'Var9'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635x1 double]
                     Bias: 10.8144
          KernelParameters: [1x1 struct]
                       Mu: [1x10 double]
                    Sigma: [1x10 double]
             SupportVectors: [3635x10 double]


  Properties, Methods
```

The compacted model discards the training data and some information related to the training process.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans =

      323793      775968
```

The compacted model consumes about half the memory of the full model.

Reduce Memory Consumption of SVM Regression Model

This example shows how to reduce the memory consumption of a full, trained SVM regression model by compacting the model and discarding the support vectors.

Load the `carsmall` sample data.

```
load carsmall
rng default  % for reproducibility
```

Train a linear SVM regression model using `Weight` as the predictor variable and `MPG` as the response variable. Standardize the data.

```
mdl = fitrsvm(Weight,MPG,'Standardize',true);
```

Note that `MPG` contains several `NaN` values. When training a model, `fitrsvm` will remove rows that contain `NaN` values from both the predictor and response data. As a result, the trained model uses only 94 of the 100 total observations contained in the sample data.

Compact the regression model to discard the training data and some information related to the training process.

```
compactMdl = compact(mdl);
```

`compactMdl` is a `CompactRegressionSVM` model that has the same parameters, support vectors, and related estimates as `mdl`, but no longer stores the training data.

Discard the support vectors and related estimates for the compacted model.

```
mdlOut = discardSupportVectors(compactMdl);
```

`mdlOut` is a `CompactRegressionSVM` model that has the same parameters as `mdl` and `compactMdl`, but no longer stores the support vectors and related estimates.

Compare the sizes of the three SVM regression models, `compactMdl`, `mdl`, and `mdlOut`.

```
vars = whos('compactMdl','mdl','mdlOut');
[vars(1).bytes,vars(2).bytes,vars(3).bytes]
```

```
ans =

        3601       13727        2305
```

The compacted model `compactMdl` consumes 3601 bytes of memory, while the full model `mdl` consumes 13727 bytes of memory. The model `mdlOut`, which also discards the support vectors, consumes 2305 bytes of memory.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

See Also

`CompactRegressionSVM` | `RegressionSVM` | `fitrsvm`

Introduced in R2015b

compact

Class: `RegressionTree`

Compact regression tree

Syntax

```
ctree = compact(tree)
```

Description

`ctree = compact(tree)` creates a compact version of `tree`.

Input Arguments

`tree`

A regression tree created using `fitrtree`.

Output Arguments

`ctree`

A compact regression tree. `ctree` has class `CompactRegressionTree`. You can predict regressions using `ctree` exactly as you can using `tree`. However, since `ctree` does not contain training data, you cannot perform some actions, such as cross validation.

Examples

Reduce Memory Consumption of Regression Tree Model

Compare the size of a full regression tree model to the compacted model.

Load the `carsmall` data set. Consider `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as predictor variables.

```
load carsmall
X = [Acceleration Cylinders Displacement Horsepower Weight];
```

Grow a regression tree using the entire data set.

```
Mdl = fitrtree(X,MPG)
```

```
Mdl =
  RegressionTree
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94
```

```
  Properties, Methods
```

`Mdl` is a `RegressionTree` model. It is a full model, that is, it stores information such as the predictor and response data `fitrtree` used in training. For a properties list of full regression tree models, see `RegressionTree`.

Create a compact version of the full regression tree. That is, one that contains enough information to make predictions only.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactRegressionTree
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'


  Properties, Methods
```

`CMdl` is a `CompactRegressionTree` model. For a properties list of compact regression tree models, see `CompactRegressionTree`.

Inspect the amounts of memory that the full and compact regression trees consume.

```
mdlInfo = whos('Mdl');
cMdlInfo = whos('CMdl');
[mdlInfo.bytes cMdlInfo.bytes]
```

```
ans = 1×2

       12401        6898
```

```
cMdlInfo.bytes/mdlInfo.bytes
```

```
ans = 0.5562
```

In this case, the compact regression tree model consumes about 25% less memory than the full model consumes.

See Also

`CompactRegressionTree` | `RegressionTree` | `fitrtree` | `predict`

compact

Class: `TreeBagger`

Compact ensemble of decision trees

Description

`CMdl = compact(Mdl)` creates a compact version of `Mdl`, a `TreeBagger` model object. You can predict regressions using `CMdl` exactly as you can using `Mdl`. However, since `CMdl` does not contain training data, you cannot perform some actions, such as make out-of-bag predictions using `oobPredict`.

Input Arguments

`Mdl`

A regression ensemble created with `TreeBagger`.

Output Arguments

`CMdl`

A compact regression ensemble. `CMdl` is of class `CompactTreeBagger`.

Examples

Reduce Size of Bag of Trees

Create a compact bag of trees for efficiently making predictions on new data.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a bag of 100 classification trees using all measurements and the `AdaBoostM1` method.

```
Mdl = TreeBagger(100,X,Y,'Method','classification')
```

```
Mdl =
  TreeBagger
Ensemble with 100 bagged decision trees:
                    Training X:              [351x34]
                    Training Y:               [351x1]
                        Method:       classification
                NumPredictors:                   34
       NumPredictorsToSample:                    6
                   MinLeafSize:                   1
                 InBagFraction:                   1
       SampleWithReplacement:                     1
        ComputeOOBPrediction:                     0
```

```
 ComputeOOBPredictorImportance:                    0
                    Proximity:                   []
                   ClassNames:             'b'             'g'

  Properties, Methods
```

`Mdl` is a `TreeBagger` model object that contains the training data, among other things.

Create a compact version of `Mdl`.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactTreeBagger
Ensemble with 100 bagged decision trees:
              Method:       classification
       NumPredictors:                   34
          ClassNames: 'b' 'g'

  Properties, Methods
```

`CMdl` is a `CompactTreeBagger` model object. `CMdl` is almost the same as `Mdl`. One exception is that it does not store the training data.

Compare the amounts of space consumed by `Mdl` and `CMdl`.

```
mdlInfo = whos('Mdl');
cMdlInfo = whos('CMdl');
[mdlInfo.bytes cMdlInfo.bytes]
```

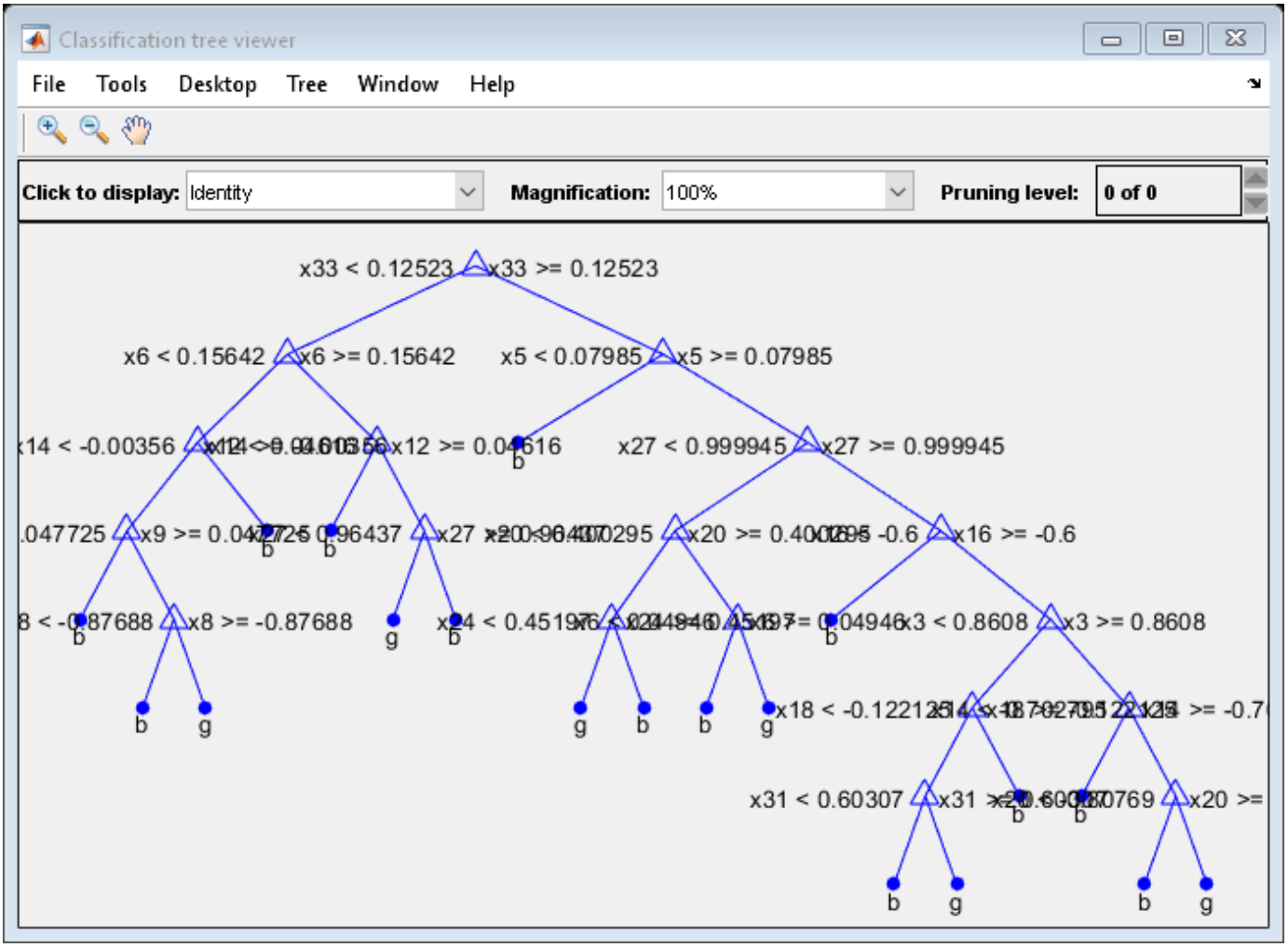
```
ans = 1×2

     1115742      976936
```

`Mdl` consumes more space than `CMdl`.

`CMdl.Trees` stores the trained classification trees (`CompactClassificationTree` model objects) that compose `Mdl`.

Display a graph of the first tree in the compact model.

```
view(CMdl.Trees{1},'Mode','graph');
```



By default, `TreeBagger` grows deep trees.

Predict the label of the mean of `X` using the compact ensemble.

```
predMeanX = predict(CMdl,mean(X))
```

```
predMeanX = 1x1 cell array
    {'g'}
```

See Also

`CompactTreeBagger` | `error` | `predict`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

CompactClassificationDiscriminant

Package: `classreg.learning.classif`

Compact discriminant analysis class

Description

A `CompactClassificationDiscriminant` object is a compact version of a discriminant analysis classifier. The compact version does not include the data for training the classifier. Therefore, you cannot perform some tasks with a compact classifier, such as cross validation. Use a compact classifier for making predictions (classifications) of new data.

Construction

`cobj = compact(obj)` constructs a compact classifier from a full classifier.

`cobj = makecdiscr(Mu,Sigma)` constructs a compact discriminant analysis classifier from the class means `Mu` and covariance matrix `Sigma`. For syntax details, see `makecdiscr`.

Input Arguments

`obj`

Discriminant analysis classifier, created using `fitcdiscr`.

Properties

`BetweenSigma`

p-by-p matrix, the between-class covariance, where p is the number of predictors.

`CategoricalPredictors`

Categorical predictor indices, which is always empty (`[]`) .

`ClassNames`

List of the elements in the training data Y with duplicates removed. `ClassNames` can be a categorical array, cell array of character vectors, character array, logical vector, or a numeric vector. `ClassNames` has the same data type as the data in the argument Y. (The software treats string arrays as cell arrays of character vectors.)

`Coeffs`

k-by-k structure of coefficient matrices, where k is the number of classes. `Coeffs(i,j)` contains coefficients of the linear or quadratic boundaries between classes `i` and `j`. Fields in `Coeffs(i,j)`:

- `DiscrimType`
- `Class1` — `ClassNames(i)`

- `Class2` — `ClassNames(j)`
- `Const` — A scalar
- `Linear` — A vector with `p` components, where `p` is the number of columns in `X`
- `Quadratic` — `p`-by-`p` matrix, exists for quadratic `DiscrimType`

The equation of the boundary between class `i` and class `j` is

`Const + Linear * x + x' * Quadratic * x = 0`,

where `x` is a column vector of length `p`.

If `fitcdiscr` had the `FillCoeffs` name-value pair set to `'off'` when constructing the classifier, `Coeffs` is empty (`[]`).

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response.

Change a `Cost` matrix using dot notation: `obj.Cost = costMatrix`.

`Delta`

Value of the Delta threshold for a linear discriminant model, a nonnegative scalar. If a coefficient of `obj` has magnitude smaller than `Delta`, `obj` sets this coefficient to `0`, and so you can eliminate the corresponding predictor from the model. Set `Delta` to a higher value to eliminate more predictors.

`Delta` must be `0` for quadratic discriminant models.

Change `Delta` using dot notation: `obj.Delta = newDelta`.

`DeltaPredictor`

Row vector of length equal to the number of predictors in `obj`. If `DeltaPredictor(i) < Delta` then coefficient `i` of the model is `0`.

If `obj` is a quadratic discriminant model, all elements of `DeltaPredictor` are `0`.

`DiscrimType`

Character vector specifying the discriminant type. One of:

- `'linear'`
- `'quadratic'`
- `'diagLinear'`
- `'diagQuadratic'`
- `'pseudoLinear'`
- `'pseudoQuadratic'`

Change `DiscrimType` using dot notation: `obj.DiscrimType = newDiscrimType`.

You can change between linear types, or between quadratic types, but cannot change between linear and quadratic types.

`Gamma`

Value of the Gamma regularization parameter, a scalar from `0` to `1`. Change `Gamma` using dot notation: `obj.Gamma = newGamma`.

- If you set `1` for linear discriminant, the discriminant sets its type to `'diagLinear'`.
- If you set a value between `MinGamma` and `1` for linear discriminant, the discriminant sets its type to `'linear'`.
- You cannot set values below the value of the `MinGamma` property.
- For quadratic discriminant, you can set either `0` (for `DiscrimType 'quadratic'`) or `1` (for `DiscrimType 'diagQuadratic'`).

`LogDetSigma`

Logarithm of the determinant of the within-class covariance matrix. The type of `LogDetSigma` depends on the discriminant type:

- Scalar for linear discriminant analysis
- Vector of length `K` for quadratic discriminant analysis, where `K` is the number of classes

`MinGamma`

Nonnegative scalar, the minimal value of the Gamma parameter so that the correlation matrix is invertible. If the correlation matrix is not singular, `MinGamma` is `0`.

`Mu`

Class means, specified as a `K`-by-`p` matrix of scalar values class means of size. `K` is the number of classes, and `p` is the number of predictors. Each row of `Mu` represents the mean of the multivariate normal distribution of the corresponding class. The class indices are in the `ClassNames` attribute.

`PredictorNames`

Cell array of names for the predictor variables, in the order in which they appear in the training data `X`.

`Prior`

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`.

Add or change a `Prior` vector using dot notation: `obj.Prior = priorVector`.

`ResponseName`

Character vector describing the response variable `Y`.

`ScoreTransform`

Character vector representing a built-in transformation function, or a function handle for transforming scores. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list

of built-in transformation functions and the syntax of custom transformation functions, see `fitcdiscr`.

Implement dot notation to add or change a `ScoreTransform` function using one of the following:

- `cobj.ScoreTransform = '`*`function`*`'`
- `cobj.ScoreTransform = @`*`function`*

`Sigma`

Within-class covariance matrix or matrices. The dimensions depend on `DiscrimType`:

- `'linear'` (default) — Matrix of size `p`-by-`p`, where `p` is the number of predictors
- `'quadratic'` — Array of size `p`-by-`p`-by-`K`, where `K` is the number of classes
- `'diagLinear'` — Row vector of length `p`
- `'diagQuadratic'` — Array of size `1`-by-`p`-by-`K`
- `'pseudoLinear'` — Matrix of size `p`-by-`p`
- `'pseudoQuadratic'` — Array of size `p`-by-`p`-by-`K`

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
logp	Log unconditional probability density for discriminant analysis classifier
loss	Classification error
mahal	Mahalanobis distance to class means
margin	Classification margins
nLinearCoeffs	Number of nonzero linear coefficients
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels using discriminant analysis classification model
shapley	Shapley values

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Construct a Compact Discriminant Analysis Classifier

Load the sample data.

```
load fisheriris
```

Construct a discriminant analysis classifier for the sample data.

```
fullobj = fitcdiscr(meas,species);
```

Construct a compact discriminant analysis classifier, and compare its size to that of the full classifier.

```
cobj = compact(fullobj);
b = whos('fullobj'); % b.bytes = size of fullobj
c = whos('cobj'); % c.bytes = size of cobj
[b.bytes c.bytes] % shows cobj uses 60% of the memory
```

```
ans = 1×2

       18291       11678
```

The compact classifier is smaller than the full classifier.

Construct Classifier Using Means and Covariances

Construct a compact discriminant analysis classifier from the means and covariances of the Fisher iris data.

```
load fisheriris
mu(1,:) = mean(meas(1:50,:));
mu(2,:) = mean(meas(51:100,:));
mu(3,:) = mean(meas(101:150,:));

mm1 = repmat(mu(1,:),50,1);
mm2 = repmat(mu(2,:),50,1);
mm3 = repmat(mu(3,:),50,1);
cc = meas;
cc(1:50,:) = cc(1:50,:) - mm1;
cc(51:100,:) = cc(51:100,:) - mm2;
cc(101:150,:) = cc(101:150,:) - mm3;
sigstar = cc' * cc / 147;
cpct = makecdiscr(mu,sigstar,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

More About

Discriminant Classification

The model for discriminant analysis is:

- Each class (Y) generates data (X) using a multivariate normal distribution. That is, the model assumes X has a Gaussian mixture distribution (`gmdistribution`).
  - For linear discriminant analysis, the model has the same covariance matrix for each class, only the means vary.
  - For quadratic discriminant analysis, both means and covariances of each class vary.

`predict` classifies so as to minimize the expected classification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{k = 1}^{K} \widehat{P}(k|x) C(y|k),$$

where

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(k|x)$ is the posterior probability on page 20-6 of class k for observation x.
- $C(y|k)$ is the cost on page 20-7 of classifying an observation as y when its true class is k.

For details, see “Prediction Using Discriminant Analysis Models” on page 20-6.

Regularization

Regularization is the process of finding a small set of predictors that yield an effective predictive model. For linear discriminant analysis, there are two parameters, γ and δ, that control regularization as follows. `cvshrink` helps you select appropriate values of the parameters.

Let Σ represent the covariance matrix of the data X, and let $\widehat{X}$ be the centered data (the data X minus the mean by class). Define

$$D = \text{diag}\left(\widehat{X}^T * \widehat{X}\right).$$

The regularized covariance matrix $\tilde{\Sigma}$ is

$$\tilde{\Sigma} = (1-\gamma)\Sigma + \gamma D\,.$$

Whenever $\gamma \geq$ `MinGamma`, $\tilde{\Sigma}$ is nonsingular.

Let μ_k be the mean vector for those elements of X in class k, and let μ_0 be the global mean vector (the mean of the rows of X). Let C be the correlation matrix of the data X, and let $\tilde{C}$ be the regularized correlation matrix:

$$\tilde{C} = (1-\gamma)C + \gamma I,$$

where I is the identity matrix.

The linear term in the regularized discriminant analysis classifier for a data point x is

$$(x-\mu_0)^T\tilde{\Sigma}^{-1}(\mu_k-\mu_0) = \left[(x-\mu_0)^T D^{-1/2}\right]\left[\tilde{C}^{-1}D^{-1/2}(\mu_k-\mu_0)\right].$$

The parameter δ enters into this equation as a threshold on the final term in square brackets. Each component of the vector $\left[\tilde{C}^{-1}D^{-1/2}(\mu_k-\mu_0)\right]$ is set to zero if it is smaller in magnitude than the threshold δ. Therefore, for class k, if component j is thresholded to zero, component j of x does not enter into the evaluation of the posterior probability.

The `DeltaPredictor` property is a vector related to this threshold. When $\delta \geq$ `DeltaPredictor(i)`, all classes k have

$$\left|\tilde{C}^{-1}D^{-1/2}(\mu_k-\mu_0)\right| \leq \delta\,.$$

Therefore, when $\delta \geq$ `DeltaPredictor(i)`, the regularized classifier does not use predictor `i`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- When you train a discriminant analysis model by using `fitcdiscr` or create a compact discriminant analysis model by using `makecdiscr`, the value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationDiscriminant` | `compact` | `compareHoldout` | `fitcdiscr` | `makecdiscr` | `predict`

Topics

“Discriminant Analysis Classification” on page 20-2

Introduced in R2011b

CompactClassificationECOC

Compact multiclass model for support vector machines (SVMs) and other classifiers

Description

`CompactClassificationECOC` is a compact version of the multiclass error-correcting output codes (ECOC) model. The compact classifier does not include the data used for training the multiclass ECOC model. Therefore, you cannot perform certain tasks, such as cross-validation, using the compact classifier. Use a compact multiclass ECOC model for tasks such as classifying new data (`predict`).

Creation

You can create a `CompactClassificationECOC` model in two ways:

- Create a compact ECOC model from a trained `ClassificationECOC` model by using the `compact` object function.
- Create a compact ECOC model by using the `fitcecoc` function and specifying the `'Learners'` name-value pair argument as `'linear'`, `'kernel'`, a `templateLinear` or `templateKernel` object, or a cell array of such objects.

Properties

After you create a `CompactClassificationECOC` model object, you can use dot notation to access its properties. For an example, see “Train and Cross-Validate ECOC Classifier” on page 33-754.

ECOC Properties

`BinaryLearners` — Trained binary learners
cell vector of model objects

Trained binary learners, specified as a cell vector of model objects. The number of binary learners depends on the number of classes in `Y` and the coding design.

The software trains `BinaryLearner{j}` according to the binary problem specified by `CodingMatrix(:,j)`. For example, for multiclass learning using SVM learners, each element of `BinaryLearners` is a `CompactClassificationSVM` classifier.

Data Types: `cell`

`BinaryLoss` — Binary learner loss function
`'binodeviance'` | `'exponential'` | `'hamming'` | `'hinge'` | `'linear'` | `'logit'` | `'quadratic'`

Binary learner loss function, specified as a character vector representing the loss function name.

If you train using binary learners that use different loss functions, then the software sets `BinaryLoss` to `'hamming'`. To potentially increase accuracy, specify a binary loss function other than the default during a prediction or loss computation by using the `'BinaryLoss'` name-value pair argument of `predict` or `loss`.

Data Types: `char`

`CodingMatrix` — Class assignment codes
numeric matrix

Class assignment codes for the binary learners, specified as a numeric matrix. `CodingMatrix` is a *K*-by-*L* matrix, where *K* is the number of classes and *L* is the number of binary learners.

The elements of `CodingMatrix` are `–1`, `0`, or `1`, and the values correspond to dichotomous class assignments. This table describes how learner `j` assigns observations in class `i` to a dichotomous class corresponding to the value of `CodingMatrix(i,j)`.

Value	Dichotomous Class Assignment
`–1`	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

Data Types: `double` | `single` | `int8` | `int16` | `int32` | `int64`

`LearnerWeights` — Binary learner weights
numeric row vector

Binary learner weights, specified as a numeric row vector. The length of `LearnerWeights` is equal to the number of binary learners (`length(Mdl.BinaryLearners)`).

`LearnerWeights(j)` is the sum of the observation weights that binary learner `j` uses to train its classifier.

The software uses `LearnerWeights` to fit posterior probabilities by minimizing the Kullback-Leibler divergence. The software ignores `LearnerWeights` when it uses the quadratic programming method of estimating posterior probabilities.

Data Types: `double` | `single`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`.

(The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has *K* rows and columns, where *K* is the number of classes.

`Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as the number of classes in `ClassNames`, and the order of the elements corresponds to the order of the classes in `ClassNames`.

`fitcecoc` incorporates misclassification costs differently among different types of binary learners.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function to apply to predicted scores
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value in the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
discardSupportVectors	Discard support vectors of linear SVM binary learners in ECOC model
edge	Classification edge for multiclass error-correcting output codes (ECOC) model
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification loss for multiclass error-correcting output codes (ECOC) model
margin	Classification margins for multiclass error-correcting output codes (ECOC) model
partialDependence	Compute partial dependence

plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using multiclass error-correcting output codes (ECOC) model
shapley	Shapley values
selectModels	Choose subset of multiclass ECOC models composed of binary ClassificationLinear learners
update	Update model parameters for code generation

Examples

Reduce Size of Full ECOC Model

Reduce the size of a full ECOC model by removing the training data. Full ECOC models (`ClassificationECOC` models) hold the training data. To improve efficiency, use a smaller classifier.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
```

Train an ECOC model using SVM binary classifiers. Standardize the predictor data using an SVM template `t`, and specify the order of the classes. During training, the software uses default values for empty options in `t`.

```
t = templateSVM('Standardize',true);
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`Mdl` is a `ClassificationECOC` model.

Reduce the size of the ECOC model.

```
CompactMdl = compact(Mdl)
```

```
CompactMdl =
  CompactClassificationECOC
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [setosa    versicolor    virginica]
           ScoreTransform: 'none'
           BinaryLearners: {3x1 cell}
             CodingMatrix: [3x3 double]


  Properties, Methods
```

`CompactMdl` is a `CompactClassificationECOC` model. `CompactMdl` does not store all of the properties that `Mdl` stores. In particular, it does not store the training data.

Display the amount of memory each classifier uses.

```
whos('CompactMdl','Mdl')
```

```
  Name            Size            Bytes  Class                                                A

  CompactMdl      1x1             15116  classreg.learning.classif.CompactClassificationECOC
  Mdl             1x1             28357  ClassificationECOC
```

The full ECOC model (`Mdl`) is approximately double the size of the compact ECOC model (`CompactMdl`).

To label new observations efficiently, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CompactMdl` and new predictor values to `predict`.

Train and Cross-Validate ECOC Classifier

Train and cross-validate an ECOC classifier using different binary learners and the one-versus-all coding design.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`. Determine the class names and the number of classes.

```
load fisheriris
X = meas;
Y = species;
classNames = unique(species(~strcmp(species,''))) % Remove empty classes
```

```
classNames = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
K = numel(classNames) % Number of classes
```

```
K = 3
```

You can use `classNames` to specify the order of the classes during training.

For a one-versus-all coding design, this example has `K = 3` binary learners. Specify templates for the binary learners such that:

- Binary learner 1 and 2 are naive Bayes classifiers. By default, each predictor is conditionally, normally distributed given its label.
- Binary learner 3 is an SVM classifier. Specify to use the Gaussian kernel.

```
rng(1);  % For reproducibility
tNB = templateNaiveBayes();
tSVM = templateSVM('KernelFunction','gaussian');
tLearners = {tNB tNB tSVM};
```

`tNB` and `tSVM` are template objects for naive Bayes and SVM learning, respectively. The objects indicate which options to use during training. Most of their properties are empty, except those specified by name-value pair arguments. During training, the software fills in the empty properties with their default values.

Train and cross-validate an ECOC classifier using the binary learner templates and the one-versus-all coding design. Specify the order of the classes. By default, naive Bayes classifiers use posterior

probabilities as scores, whereas SVM classifiers use distances from the decision boundary. Therefore, to aggregate the binary learners, you must specify to fit posterior probabilities.

```
CVMdl = fitcecoc(X,Y,'ClassNames',classNames,'CrossVal','on',...
    'Learners',tLearners,'FitPosterior',true);
```

`CVMdl` is a `ClassificationPartitionedECOC` cross-validated model. By default, the software implements 10-fold cross-validation. The scores across the binary learners have the same form (that is, they are posterior probabilities), so the software can aggregate the results of the binary classifications properly.

Inspect one of the trained folds using dot notation.

```
CVMdl.Trained{1}
```

```
ans =
  CompactClassificationECOC
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
           BinaryLearners: {3x1 cell}
             CodingMatrix: [3x3 double]


  Properties, Methods
```

Each fold is a `CompactClassificationECOC` model trained on 90% of the data.

You can access the results of the binary learners using dot notation and cell indexing. Display the trained SVM classifier (the third binary learner) in the first fold.

```
CVMdl.Trained{1}.BinaryLearners{3}
```

```
ans =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [-1 1]
           ScoreTransform: '@(S)sigmoid(S,-4.016735e+00,-3.243073e-01)'
                    Alpha: [33x1 double]
                     Bias: -0.1346
         KernelParameters: [1x1 struct]
           SupportVectors: [33x4 double]
      SupportVectorLabels: [33x1 double]


  Properties, Methods
```

Estimate the generalization error.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.0333
```

On average, the generalization error is approximately 3%.

Algorithms

Random Coding Design Matrices

For a given number of classes K, the software generates random coding design matrices as follows.

1. The software generates one of these matrices:
   a. Dense random — The software assigns 1 or –1 with equal probability to each element of the K-by-L_d coding design matrix, where $L_d \approx \lceil 10\log_2 K \rceil$.
   b. Sparse random — The software assigns 1 to each element of the K-by-L_s coding design matrix with probability 0.25, –1 with probability 0.25, and 0 with probability 0.5, where $L_s \approx \lceil 15\log_2 K \rceil$.
2. If a column does not contain at least one 1 and at least one –1, then the software removes that column.
3. For distinct columns u and v, if $u = v$ or $u = -v$, then the software removes v from the coding design matrix.

The software randomly generates 10,000 matrices by default, and retains the matrix with the largest, minimal, pairwise row distance based on the Hamming measure ([4]) given by

$$\Delta(k_1, k_2) = 0.5 \sum_{l=1}^{L} \left|m_{k_1 l}\right|\left|m_{k_2 l}\right|\left|m_{k_1 l} - m_{k_2 l}\right|,$$

where $m_{k_j l}$ is an element of coding design matrix j.

Support Vector Storage

By default and for efficiency, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties for all linear SVM binary learners. `fitcecoc` lists `Beta`, rather than `Alpha`, in the model display.

To store `Alpha`, `SupportVectorLabels`, and `SupportVectors`, pass a linear SVM template that specifies storing support vectors to `fitcecoc`. For example, enter:

```
t = templateSVM('SaveSupportVectors',true)
Mdl = fitcecoc(X,Y,'Learners',t);
```

You can remove the support vectors and related values by passing the resulting `ClassificationECOC` model to `discardSupportVectors`.

References

[1] Fürnkranz, Johannes. “Round Robin Classification.” *Journal of Machine Learning Research*, Vol. 2, 2002, pp. 721–747.

[2] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recognition Letters*, Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- When you train an ECOC model by using `fitcecoc`, the following restrictions apply.
  - You cannot fit posterior probabilities by using the `'FitPosterior'` name-value pair argument.
  - All binary learners must be either SVM classifiers or linear classification models. For the `'Learners'` name-value pair argument, you can specify:
    - `'svm'` or `'linear'`
    - An SVM template object or a cell array of such objects (see `templateSVM`)
    - A linear classification model template object or a cell array of such objects (see `templateLinear`)
  - When you generate code using a coder configurer for `predict` and `update`, the following additional restrictions apply for binary learners.
    - If you use a cell array of SVM template objects, the value of `'Standardize'` for SVM learners must be consistent. For example, if you specify `'Standardize',true` for one SVM learner, you must specify the same value for all SVM learners.
    - If you use a cell array of SVM template objects, and you use one SVM learner with a linear kernel (`'KernelFunction','linear'`) and another with a different type of kernel function, then you must specify `'SaveSupportVectors',true` for the learner with a linear kernel.

    For details, see `ClassificationECOCCoderConfigurer`. For information on name-value pair arguments that you cannot modify when you retrain a model, see “Tips” on page 33-6501.
  - Code generation limitations for SVM classifiers and linear classification models also apply to ECOC classifiers, depending on the choice of binary learners. For more details, see “Code Generation” on page 33-807 of the `CompactClassificationSVM` class and “Code Generation” on page 33-419 of the `ClassificationLinear` class.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `ClassificationPartitionedKernelECOC` | `ClassificationPartitionedLinearECOC` | `compact` | `fitcecoc`

Introduced in R2014b

CompactClassificationEnsemble

Package: `classreg.learning.classif`

Compact classification ensemble class

Description

Compact version of a classification ensemble (of class `ClassificationEnsemble`). The compact version does not include the data for training the classification ensemble. Therefore, you cannot perform some tasks with a compact classification ensemble, such as cross validation. Use a compact classification ensemble for making predictions (classifications) of new data.

Construction

`ens = compact(fullEns)` constructs a compact decision ensemble from a full decision ensemble.

Input Arguments

`fullEns`

A classification ensemble created by `fitcensemble`.

Properties

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a numeric vector, vector of categorical variables, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

`CombineWeights`

Character vector describing how `ens` combines weak learner weights, either `'WeightedSum'` or `'WeightedAverage'`.

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response. This property is read-only.

ExpandedPredictorNames

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

NumTrained

Number of trained weak learners in `ens`, a scalar.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in `X`.

Prior

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`. The number of elements of `Prior` is the number of unique classes in the response. This property is read-only.

ResponseName

Character vector with the name of the response variable `Y`.

ScoreTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitctree`.

Add or change a `ScoreTransform` function using dot notation:

```
ens.ScoreTransform = 'function'
```

or

```
ens.ScoreTransform = @function
```

Trained

A cell vector of trained classification models.

- If `Method` is `'LogitBoost'` or `'GentleBoost'`, then `CompactClassificationEnsemble` stores trained learner j in the `CompactRegressionLearner` property of the object stored in `Trained{`j`}`. That is, to access trained learner j, use `ens.Trained{`j`}.CompactRegressionLearner`.
- Otherwise, cells of the cell vector contain the corresponding, compact classification models.

TrainedWeights

Numeric vector of trained weights for the weak learners in `ens`. `TrainedWeights` has `T` elements, where `T` is the number of weak learners in `learners`.

UsePredForLearner

Logical matrix of size P-by-NumTrained, where P is the number of predictors (columns) in the training data X. UsePredForLearner(i,j) is true when learner j uses predictor i, and is false otherwise. For each learner, the predictors have the same order as the columns in the training data X.

If the ensemble is not of type Subspace, all entries in UsePredForLearner are true.

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification error
margin	Classification margins
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using ensemble of classification models
predictorImportance	Estimates of predictor importance for classification ensemble of decision trees
removeLearners	Remove members of compact classification ensemble
shapley	Shapley values

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Reduce Size of Classification Ensemble

Create a compact classification ensemble for efficiently making predictions on new data.

Load the ionosphere data set.

```
load ionosphere
```

Train a boosted ensemble of 100 classification trees using all measurements and the AdaBoostM1 method.

```
Mdl = fitcensemble(X,Y,'Method','AdaBoostM1')
```

```
Mdl =
  ClassificationEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
               NumTrained: 100
                   Method: 'AdaBoostM1'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
```

```
                FitInfo: [100x1 double]
      FitInfoDescription: {2x1 cell}


  Properties, Methods
```

`Mdl` is a `ClassificationEnsemble` model object that contains the training data, among other things.

Create a compact version of `Mdl`.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactClassificationEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
               NumTrained: 100


  Properties, Methods
```

`CMdl` is a `CompactClassificationEnsemble` model object. `CMdl` is almost the same as `Mdl`. One exception is that `CMdl` does not store the training data.

Compare the amounts of space consumed by `Mdl` and `CMdl`.

```
mdlInfo = whos('Mdl');
cMdlInfo = whos('CMdl');
[mdlInfo.bytes cMdlInfo.bytes]
```

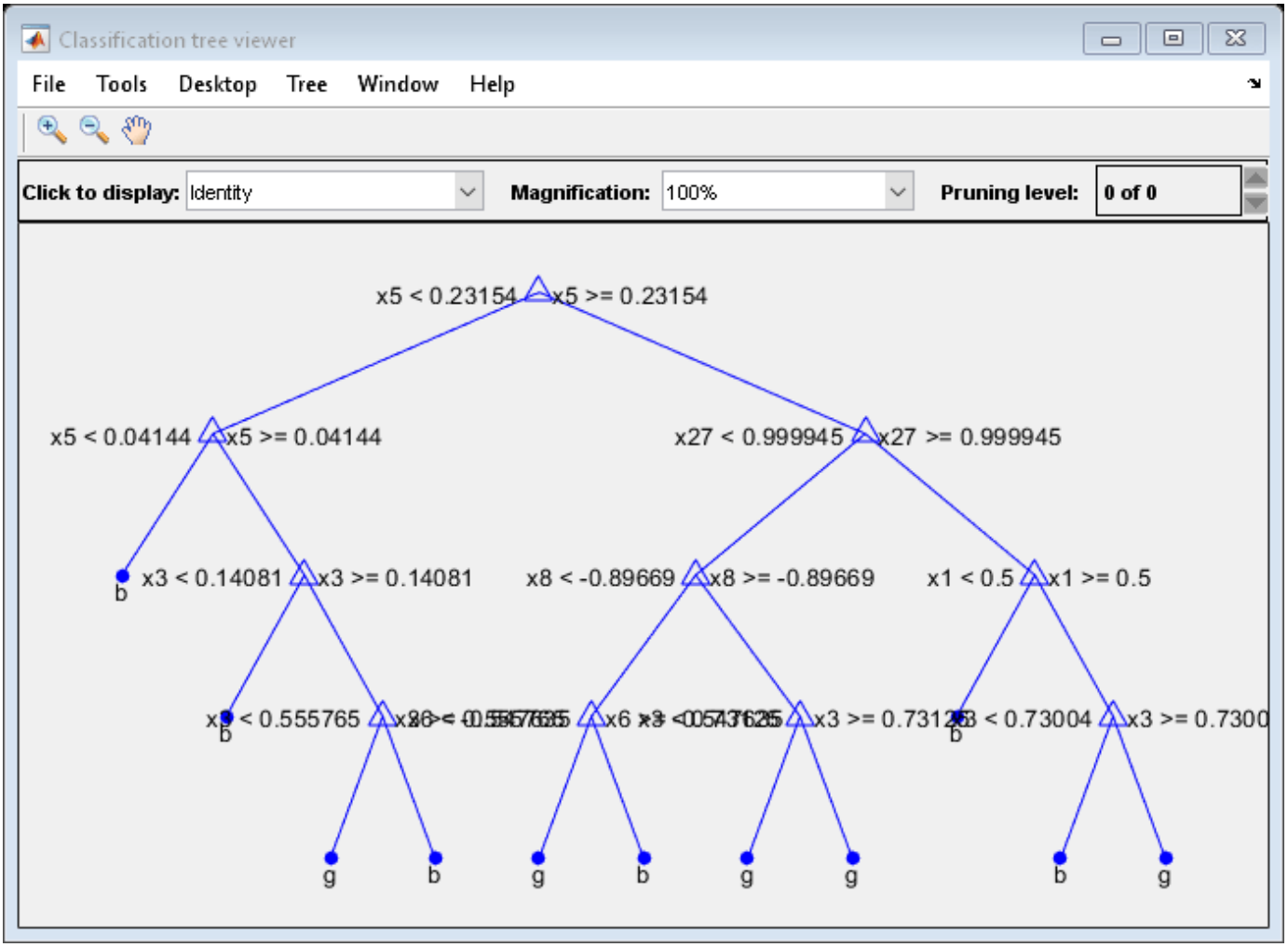
```
ans = 1×2

      878146      631686
```

`Mdl` consumes more space than `CMdl`.

`CMdl.Trained` stores the trained classification trees (`CompactClassificationTree` model objects) that compose `Mdl`.

Display a graph of the first tree in the compact ensemble.

```
view(CMdl.Trained{1},'Mode','graph');
```



By default, `fitcensemble` grows shallow trees for boosted ensembles of trees.

Predict the label of the mean of `X` using the compact ensemble.

```
predMeanX = predict(CMdl,mean(X))
```

```
predMeanX = 1x1 cell array
    {'g'}
```

Tip

For an ensemble of classification trees, the `Trained` property of `ens` stores an `ens.NumTrained`-by-1 cell vector of compact classification models. For a textual or graphical display of tree `t` in the cell vector, enter:

- `view(ens.Trained{t}.CompactRegressionLearner)` for ensembles aggregated using LogitBoost or GentleBoost.
- `view(ens.Trained{t})` for all other aggregation methods.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the ClassificationEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitcensemble`, code generation limitations for the weak learners used in the ensemble also apply to the ensemble. For more details, see the Code Generation sections of `ClassificationKNN`, `CompactClassificationDiscriminant`, and `CompactClassificationTree`.
- For fixed-point code generation, you must train an ensemble using tree learners.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationEnsemble` | `compact` | `compareHoldout` | `fitcensemble` | `fitctree` | `predict` | `view`

Introduced in R2011a

ClassificationGAM

Generalized additive model (GAM) for binary classification

Description

A `ClassificationGAM` object is a generalized additive model on page 33-775 (GAM) object for binary classification. It is an interpretable model that explains class scores (the logit of class probabilities) using a sum of univariate and bivariate shape functions.

You can classify new observations by using the `predict` function, and plot the effect of each shape function on the prediction (class score) for an observation by using the `plotLocalEffects` function. For the full list of object functions for `ClassificationGAM`, see "Object Functions" on page 33-769.

Creation

Create a `ClassificationGAM` object by using `fitcgam`. You can specify both linear terms and interaction terms for predictors to include univariate shape functions (predictor trees) and bivariate shape functions (interaction trees) in a trained model, respectively.

You can update a trained model by using `resume` or `addInteractions`.

- The `resume` function resumes training for the existing terms in a model.
- The `addInteractions` function adds interaction terms to a model that contains only linear terms.

Properties

GAM Properties

`BinEdges` — Bin edges for numeric predictors
cell array of numeric vectors | `[ ]`

This property is read-only.

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
```

```
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Data Types: `cell`

`Interactions` — Interaction term indices
two-column matrix of positive integers | `[]`

This property is read-only.

Interaction term indices, specified as a `t`-by-2 matrix of positive integers, where `t` is the number of interaction terms in the model. Each row of the matrix represents one interaction term and contains the column indexes of the predictor data `X` for the interaction term. If the model does not include an interaction term, then this property is empty (`[]`).

The software adds interaction terms to the model in the order of importance based on the *p*-values. Use this property to check the order of the interaction terms added to the model.

Data Types: `double`

`Intercept` — Intercept term of model
numeric scalar

This property is read-only.

Intercept (constant) term of the model, which is the sum of the intercept terms in the predictor trees and interaction trees, specified as a numeric scalar.

Data Types: `single` | `double`

`ModelParameters` — Parameters used to train model
model parameter object

This property is read-only.

Parameters used to train the model, specified as a model parameter object. `ModelParameters` contains parameter values such as those for the name-value arguments used to train the model. `ModelParameters` does not contain estimated parameters.

Access the fields of `ModelParameters` by using dot notation. For example, access the maximum number of decision splits per interaction tree by using `Mdl.ModelParameters.MaxNumSplitsPerInteraction`.

`PairDetectionBinEdges` — Bin edges for interaction term detection
cell array of numeric vectors

This property is read-only.

Bin edges for interaction term detection for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

To speed up the interaction term detection process, the software bins numeric predictors into at most 8 equiprobable bins. The number of bins can be less than 8 if a predictor has fewer than 8 unique values.

Data Types: `cell`

`ReasonForTermination` — Reason training stops
structure

This property is read-only.

Reason training the model stops, specified as a structure with two fields, `PredictorTrees` and `InteractionTrees`.

Use this property to check if the model contains the specified number of trees for each linear term (`'NumTreesPerPredictor'`) and for each interaction term (`'NumTreesPerInteraction'`). If the `fitcgam` function terminates training before adding the specified number of trees, this property contains the reason for the termination.

Data Types: `struct`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Cost` — Misclassification costs
2-by-2 numeric matrix

Misclassification costs, specified as a 2-by-2 numeric matrix.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

The software uses the `Cost` value for prediction, but not training. You can change the value by using dot notation.

Example: `Mdl.Cost = C;`

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

`ExpandedPredictorNames` is the same as `PredictorNames` for a generalized additive model.

Data Types: `cell`

`NumObservations` — Number of observations
numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

Data Types: `double`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector with two elements. The order of the elements corresponds to the order of the elements in `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

RowsUsed — Rows used in fitting

`[]` | logical vector

This property is read-only.

Rows of the original training data used in fitting the `ClassificationGAM` model, specified as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

ScoreTransform — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

This property determines the output score computation for object functions such as `predict`, `margin`, and `edge`. Use `'logit'` to compute posterior probabilities, and use `'none'` to compute the logit of posterior probabilities.

Data Types: `char` | `function_handle`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to train the model, specified as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `'Weights'` name-value argument so that the elements of `W` within a particular class sum up to the prior probability of that class.

Data Types: `double`

`X` — Predictors
numeric matrix | table

This property is read-only.

Predictors used to train the model, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

`Y` — Class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Class labels used to train the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `Y` has the same data type as the response variable used to train the model. (The software treats string arrays as cell arrays of character vectors.)

Each row of `Y` represents the observed classification of the corresponding row of `X`.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

Object Functions

Create `CompactClassificationGAM`
compact Reduce size of machine learning model

Create `ClassificationPartitionedGAM`
crossval Cross-validate machine learning model

Update GAM
addInteractions Add interaction terms to univariate generalized additive model (GAM)

resume	Resume training of generalized additive model (GAM)

Interpret Prediction

lime	Local interpretable model-agnostic explanations (LIME)
partialDependence	Compute partial dependence
plotLocalEffects	Plot local effects of terms in generalized additive model (GAM)
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
shapley	Shapley values

Assess Predictive Performance on New Observations

predict	Classify observations using generalized additive model (GAM)
loss	Classification loss for generalized additive model (GAM)
margin	Classification margins for generalized additive model (GAM)
edge	Classification edge for generalized additive model (GAM)

Assess Predictive Performance on Training Data

resubPredict	Classify training data using trained classifier
resubLoss	Resubstitution classification loss
resubMargin	Resubstitution classification margin
resubEdge	Resubstitution classification edge

Compare Accuracies

compareHoldout	Compare accuracies of two classification models using new data
testckfold	Compare accuracies of two classification models by repeated cross-validation

Examples

Train Generalized Additive Model

Train a univariate generalized additive model, which contains linear terms for predictors. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a univariate GAM that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

```
Mdl = fitcgam(X,Y)
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 2.2715
          NumObservations: 351
```

```
  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. The model display shows a partial list of the model properties. To view the full list of properties, double-click the variable name `Mdl` in the Workspace. The Variables editor opens for `Mdl`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the class order of `Mdl`.

```
classOrder = Mdl.ClassNames
```

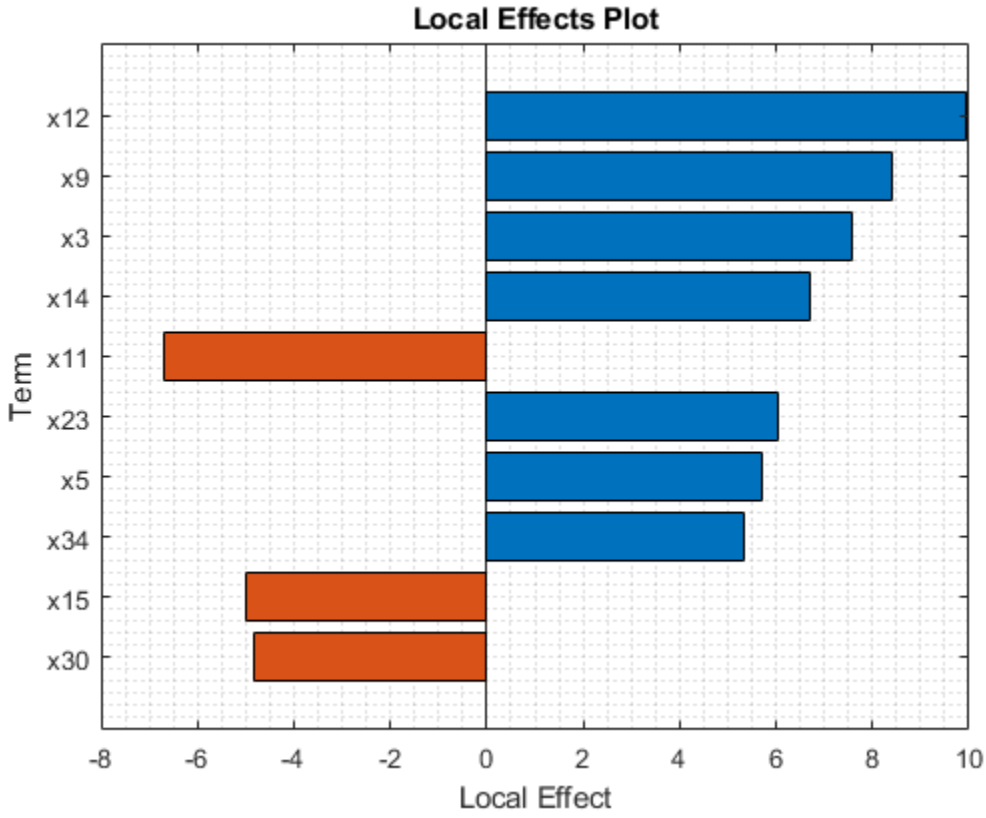
```
classOrder = 2x1 cell
    {'b'}
    {'g'}
```

Classify the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction.

```
label = predict(Mdl,X(1,:))
```

```
label = 1x1 cell array
    {'g'}
```

```
plotLocalEffects(Mdl,X(1,:))
```



The `predict` function classifies the first observation `X(1,:)` as `'g'`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the 10 most important terms on

the prediction. Each local effect value shows the contribution of each term to the classification score for `'g'`, which is the logit of the posterior probability that the classification is `'g'` for the observation.

Train GAM with Interaction Terms

Train a generalized additive model that contains linear and interaction terms for predictors in three different ways:

- Specify the interaction terms using the `formula` input argument.
- Specify the `'Interactions'` name-value argument.
- Build a model with linear terms first and add interaction terms to the model by using the `addInteractions` function.

Load Fisher's iris data set. Create a table that contains observations for versicolor and virginica.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
tbl = array2table(meas(inds,:),'VariableNames',["x1","x2","x3","x4"]);
tbl.Y = species(inds,:);
```

Specify `formula`

Train a GAM that contains the four linear terms (`x1`, `x2`, `x3`, and `x4`) and two interaction terms (`x1*x2` and `x2*x3`). Specify the terms using a formula in the form `'Y ~ terms'`.

```
Mdl1 = fitcgam(tbl,'Y ~ x1 + x2 + x3 + x4 + x1:x2 + x2:x3');
```

The function adds interaction terms to the model in the order of importance. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitcgam` adds them to the model. Display the `Interactions` property.

```
Mdl1.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term.

Specify `'Interactions'`

Pass the training data (`tbl`) and the name of the response variable in `tbl` to `fitcgam`, so that the function includes the linear terms for all the other variables as predictors. Specify the `'Interactions'` name-value argument using a logical matrix to include the two interaction terms, `x1*x2` and `x2*x3`.

```
Mdl2 = fitcgam(tbl,'Y','Interactions',logical([1 1 0 0; 0 1 1 0]));
Mdl2.Interactions
```

```
ans = 2×2
```

```
     2     3
     1     2
```

You can also specify `'Interactions'` as the number of interaction terms or as `'all'` to include all available interaction terms. Among the specified interaction terms, `fitcgam` identifies those whose *p*-values are not greater than the `'MaxPValue'` value and adds them to the model. The default `'MaxPValue'` is 1 so that the function adds all specified interaction terms to the model.

Specify `'Interactions','all'` and set the `'MaxPValue'` name-value argument to 0.01.

```
Mdl3 = fitcgam(tbl,'Y','Interactions','all','MaxPValue',0.01);
Mdl3.Interactions
```

```
ans = 5×2

     3     4
     2     4
     1     4
     2     3
     1     3
```

`Mdl3` includes five of the six available pairs of interaction terms.

Use `addInteractions` Function

Train a univariate GAM that contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function. Specify the second input argument of `addInteractions` in the same way you specify the `'Interactions'` name-value argument of `fitcgam`. You can specify the list of interaction terms using a logical matrix, the number of interaction terms, or `'all'`.

Specify the number of interaction terms as 5 to add the five most important interaction terms to the trained model.

```
Mdl4 = fitcgam(tbl,'Y');
UpdatedMdl4 = addInteractions(Mdl4,5);
UpdatedMdl4.Interactions
```

```
ans = 5×2

     3     4
     2     4
     1     4
     2     3
     1     3
```

`Mdl4` is a univariate GAM, and `UpdatedMdl4` is an updated GAM that contains all the terms in `Mdl4` and five additional interaction terms.

Resume Training Predictor Trees in GAM

Train a univariate classification GAM (which contains only linear terms) for a small number of iterations. After training the model for more iterations, compare the resubstitution loss.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a univariate GAM that identifies whether the radar return is bad (`'b'`) or good (`'g'`). Specify the number of trees per linear term as 2. `fitcgam` iterates the boosting algorithm for the specified number of iterations. For each boosting iteration, the function adds one tree per linear term. Specify `'Verbose'` as 2 to display diagnostic messages at every iteration.

```
Mdl = fitcgam(X,Y,'NumTreesPerPredictor',2,'Verbose',2);
```

```
|=============================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=============================================================|
|    1D|         0|      486.59|      -     |      -     |
|    1D|         1|      166.71|         Inf|           1|
|    1D|         2|      78.336|     0.58205|           1|
```

To check whether `fitcgam` trains the specified number of trees, display the `ReasonForTermination` property of the trained model and view the displayed message.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: ''
```

Compute the classification loss for the training data.

```
resubLoss(Mdl)
```

```
ans = 0.0142
```

Resume training the model for another 100 iterations. Because `Mdl` contains only linear terms, the `resume` function resumes training for the linear terms and adds more trees for them (predictor trees). Specify `'Verbose'` and `'NumPrint'` to display diagnostic messages at every 10 iterations.

```
UpdatedMdl = resume(Mdl,100,'Verbose',1,'NumPrint',10);
```

```
|=============================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=============================================================|
|    1D|         0|      78.336|      -     |      -     |
|    1D|         1|      38.364|     0.17429|           1|
|    1D|        10|     0.16311|    0.011894|           1|
|    1D|        20|  0.00035693|   0.0025178|           1|
|    1D|        30|  8.1191e-07|   0.0011006|           1|
|    1D|        40|  1.7978e-09|  0.00074607|           1|
|    1D|        50|  3.6113e-12|  0.00034404|           1|
|    1D|        60|  1.7497e-13|  0.00016541|           1|
```

```
UpdatedMdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Unable to improve the model fit.'
    InteractionTrees: ''
```

`resume` terminates training when adding more trees does not improve the deviance of the model fit.

Compute the classification loss using the updated model.

```
resubLoss(UpdatedMdl)
```

```
ans = 0
```

The classification loss decreases after `resume` updates the model with more iterations.

More About

Generalized Additive Model (GAM) for Binary Classification

A generalized additive model (GAM) is an interpretable model that explains class scores (the logit of class probabilities) using a sum of univariate and bivariate shape functions of predictors.

`fitcgam` uses a boosted tree as a shape function for each predictor and, optionally, each pair of predictors; therefore, the function can capture a nonlinear relation between a predictor and the response variable. Because contributions of individual shape functions to the prediction (classification score) are well separated, the model is easy to interpret.

The standard GAM uses a univariate shape function for each predictor.

$$y \sim Binomial(n, \mu)$$

$$g(\mu) = \log\frac{\mu}{1-\mu} = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where y is a response variable that follows the binomial distribution with the probability of success (probability of positive class) μ in n observations. $g(\mu)$ is a logit link function, and c is an intercept (constant) term. $f_i(x_i)$ is a univariate shape function for the ith predictor, which is a boosted tree for a linear term for the predictor (predictor tree).

You can include interactions between predictors in a model by adding bivariate shape functions of important interaction terms to the model.

$$g(\mu) = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \sum_{i,\, j \in \{1, 2, \cdots, p\}} f_{ij}(x_i x_j),$$

where $f_{ij}(x_i x_j)$ is a bivariate shape function for the ith and jth predictors, which is a boosted tree for an interaction term for the predictors (interaction tree).

`fitcgam` finds important interaction terms based on the p-values of F-tests. For details, see “Interaction Term Detection” on page 33-1709.

References

[1] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible Models for Classification and Regression." *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’12).* Beijing, China: ACM Press, 2012, pp. 150–158.

[2] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate Intelligible Models with Pairwise Interactions." *Proceedings of the 19th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining (KDD '13) Chicago, Illinois, USA: ACM Press, 2013, pp. 623–631.

See Also

`ClassificationPartitionedGAM` | `CompactClassificationGAM` | `addInteractions` | `fitcgam` | `resume`

Topics

"Train Generalized Additive Model for Binary Classification" on page 12-77

Introduced in R2021a

CompactClassificationNaiveBayes

Compact naive Bayes classifier for multiclass classification

Description

`CompactClassificationNaiveBayes` is a compact version of the naive Bayes classifier. The compact classifier does not include the data used for training the naive Bayes classifier. Therefore, you cannot perform some tasks, such as cross-validation, using the compact classifier. Use a compact naive Bayes classifier for tasks such as predicting the labels of the data.

Creation

Create a `CompactClassificationNaiveBayes` model from a full, trained `ClassificationNaiveBayes` classifier by using `compact`.

Properties

Predictor Properties

`PredictorNames` — Predictor names
cell array of character vectors

This property is read-only.

Predictor names, specified as a cell array of character vectors. The order of the elements in `PredictorNames` corresponds to the order in which the predictor names appear in the training data `X`.

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

If the model uses dummy variable encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`CategoricalPredictors` — Categorical predictor indices
`[]` | vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`CategoricalLevels` — Multivariate multinomial levels

cell array

This property is read-only.

Multivariate multinomial levels, specified as a cell array. The length of `CategoricalLevels` is equal to the number of predictors (`size(X,2)`).

The cells of `CategoricalLevels` correspond to predictors that you specify as `'mvmn'` during training, that is, they have a multivariate multinomial distribution. Cells that do not correspond to a multivariate multinomial distribution are empty (`[]`).

If predictor *j* is multivariate multinomial, then `CategoricalLevels{`*j*`}` is a list of all distinct values of predictor *j* in the sample. `NaN`s are removed from `unique(X(:,j))`.

Predictor Distribution Properties

`DistributionNames` — Predictor distributions

`'normal'` (default) | `'kernel'` | `'mn'` | `'mvmn'` | cell array of character vectors

This property is read-only.

Predictor distributions, specified as a character vector or cell array of character vectors. `fitcnb` uses the predictor distributions to model the predictors. This table lists the available distributions.

Value	Description
`'kernel'`	Kernel smoothing density estimate
`'mn'`	Multinomial distribution. If you specify `mn`, then all features are components of a multinomial distribution. Therefore, you cannot include `'mn'` as an element of a string array or a cell array of character vectors. For details, see “Estimated Probability for Multinomial Distribution” on page 33-446.
`'mvmn'`	Multivariate multinomial distribution. For details, see “Estimated Probability for Multivariate Multinomial Distribution” on page 33-446.
`'normal'`	Normal (Gaussian) distribution

If `DistributionNames` is a 1-by-*P* cell array of character vectors, then `fitcnb` models the feature *j* using the distribution in element *j* of the cell array.

Example: `'mn'`

Example: `{'kernel','normal','kernel'}`

Data Types: `char` | `string` | `cell`

`DistributionParameters` — Distribution parameter estimates

cell array

This property is read-only.

Distribution parameter estimates, specified as a cell array. `DistributionParameters` is a *K*-by-*D* cell array, where cell (*k*,*d*) contains the distribution parameter estimates for instances of predictor *d*

in class k. The order of the rows corresponds to the order of the classes in the property `ClassNames`, and the order of the predictors corresponds to the order of the columns of `X`.

If class k has no observations for predictor j, then the `Distribution{k,j}` is empty (`[]`).

The elements of `DistributionParameters` depend on the distributions of the predictors. This table describes the values in `DistributionParameters{k,j}`.

Distribution of Predictor *j*	Value of Cell Array for Predictor `j` and Class `k`
`kernel`	A `KernelDistribution` model. Display properties using cell indexing and dot notation. For example, to display the estimated bandwidth of the kernel density for predictor 2 in the third class, use `Mdl.DistributionParameters{3,2}.BandWidth`.
`mn`	A scalar representing the probability that token j appears in class k. For details, see "Estimated Probability for Multinomial Distribution" on page 33-446.
`mvmn`	A numeric vector containing the probabilities for each possible level of predictor j in class k. The software orders the probabilities by the sorted order of all unique levels of predictor j (stored in the property `CategoricalLevels`). For more details, see "Estimated Probability for Multivariate Multinomial Distribution" on page 33-446.
`normal`	A 2-by-1 numeric vector. The first element is the sample mean and the second element is the sample standard deviation.

`Kernel` — Kernel smoother type

`'normal'` (default) | `'box'` | cell array | ...

This property is read-only.

Kernel smoother type, specified as the name of a kernel or a cell array of kernel names. The length of `Kernel` is equal to the number of predictors (`size(X,2)`). `Kernel{j}` corresponds to predictor j and contains a character vector describing the type of kernel smoother. If a cell is empty (`[]`), then `fitcnb` did not fit a kernel distribution to the corresponding predictor.

This table describes the supported kernel smoother types. $I\{u\}$ denotes the indicator function.

Value	Kernel	Formula
`'box'`	Box (uniform)	$f(x) = 0.5I\{\|x\| \le 1\}$
`'epanechnikov'`	Epanechnikov	$f(x) = 0.75\left(1 - x^2\right)I\{\|x\| \le 1\}$
`'normal'`	Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-0.5x^2\right)$
`'triangle'`	Triangular	$f(x) = (1 - \|x\|)I\{\|x\| \le 1\}$

Example: `'box'`

Example: `{'epanechnikov','normal'}`

Data Types: `char` | `string` | `cell`

`Support` — Kernel smoother density support

cell array

This property is read-only.

Kernel smoother density support, specified as a cell array. The length of `Support` is equal to the number of predictors (`size(X,2)`). The cells represent the regions to which `fitcnb` applies the kernel density. If a cell is empty (`[]`), then `fitcnb` did not fit a kernel distribution to the corresponding predictor.

This table describes the supported options.

Value	Description
1-by-2 numeric row vector	The density support applies to the specified bounds, for example `[L,U]`, where `L` and `U` are the finite lower and upper bounds, respectively.
`'positive'`	The density support applies to all positive real values.
`'unbounded'`	The density support applies to all real values.

`Width` — Kernel smoother window width

numeric matrix

This property is read-only.

Kernel smoother window width, specified as a numeric matrix. `Width` is a *K*-by-*P* matrix, where *K* is the number of classes in the data, and *P* is the number of predictors (`size(X,2)`).

`Width(`*k*`,`*j*`)` is the kernel smoother window width for the kernel smoothing density of predictor *j* within class *k*. `NaN`s in column *j* indicate that `fitcnb` did not fit predictor *j* using a kernel density.

Response Properties

`ClassNames` — Unique class names

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class names used in the training model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`ClassNames` has the same data type as `Y`, and has *K* elements (or rows) for character arrays. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `categorical` | `char` | `string` | `logical` | `double` | `cell`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char` | `string`

Training Properties

`Prior` — Prior probabilities
numeric vector

Prior probabilities, specified as a numeric vector. The order of the elements in `Prior` corresponds to the elements of `Mdl.ClassNames`.

`fitcnb` normalizes the prior probabilities you set using the `'Prior'` name-value pair argument, so that `sum(Prior) = 1`.

The value of `Prior` does not affect the best-fitting model. Therefore, you can reset `Prior` after training `Mdl` using dot notation.

Example: `Mdl.Prior = [0.2 0.8]`

Data Types: `double` | `single`

Classifier Properties

`Cost` — Misclassification cost
square matrix

Misclassification cost, specified as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

The misclassification cost matrix must have zeros on the diagonal.

The value of `Cost` does not influence training. You can reset `Cost` after training `Mdl` using dot notation.

Example: `Mdl.Cost = [0 0.5 ; 1 0]`

Data Types: `double` | `single`

`ScoreTransform` — Classification score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Classification score transformation, specified as a character vector or function handle. This table summarizes the available character vectors.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$

Value	Description
'symmetric'	$2x - 1$
'symmetricismax'	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
'symmetriclogit'	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transformation. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `Mdl.ScoreTransform = 'logit'`

Data Types: `char` | `string` | `function handle`

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
edge	Classification edge for naive Bayes classifier
lime	Local interpretable model-agnostic explanations (LIME)
logp	Log unconditional probability density for naive Bayes classifier
loss	Classification loss for naive Bayes classifier
margin	Classification margins for naive Bayes classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using naive Bayes classifier
shapley	Shapley values

Examples

Reduce Size of Naive Bayes Classifier

Reduce the size of a full naive Bayes classifier by removing the training data. Full naive Bayes classifiers hold the training data. You can use a compact naive Bayes classifier to improve memory efficiency.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'b','g'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
```

```
          NumObservations: 351
        DistributionNames: {1x32 cell}
    DistributionParameters: {2x32 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Reduce the size of the naive Bayes classifier.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
        DistributionNames: {1x32 cell}
    DistributionParameters: {2x32 cell}


  Properties, Methods
```

`CMdl` is a trained `CompactClassificationNaiveBayes` classifier.

Display the amount of memory used by each classifier.

```
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class

  CMdl      1x1               15060  classreg.learning.classif.CompactClassificationNaiveBayes
  Mdl       1x1              111174  ClassificationNaiveBayes
```

The full naive Bayes classifier (`Mdl`) is more than seven times larger than the compact naive Bayes classifier (`CMdl`).

To label new observations efficiently, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CMdl` and new predictor values to `predict`.

Train and Cross-Validate Naive Bayes Classifier

Train and cross-validate a naive Bayes classifier. `fitcnb` implements 10-fold cross-validation by default. Then, estimate the cross-validated classification error.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
rng('default')  % for reproducibility
```

Train and cross-validate a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
CVMdl = fitcnb(X,Y,'ClassNames',{'b','g'},'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedModel
    CrossValidatedModel: 'NaiveBayes'
         PredictorNames: {1x32 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedModel` cross-validated, naive Bayes classifier. Alternatively, you can cross-validate a trained `ClassificationNaiveBayes` model by passing it to `crossval`.

Display the first training fold of `CVMdl` using dot notation.

```
CVMdl.Trained{1}
```

```
ans =
  CompactClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
         DistributionNames: {1x32 cell}
    DistributionParameters: {2x32 cell}


  Properties, Methods
```

Each fold is a `CompactClassificationNaiveBayes` model trained on 90% of the data.

Full and compact naive Bayes models are not used for predicting on new data. Instead, use them to estimate the generalization error by passing `CVMdl` to `kfoldLoss`.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.1852
```

On average, the generalization error is approximately 19%.

You can specify a different conditional distribution for the predictors, or tune the conditional distribution parameters to reduce the generalization error.

More About

Bag-of-Tokens Model

In the bag-of-tokens model, the value of predictor j is the nonnegative number of occurrences of token j in the observation. The number of categories (bins) in the multinomial model is the number of distinct tokens (number of predictors).

Naive Bayes

Naive Bayes is a classification algorithm that applies density estimation to the data.

The algorithm leverages Bayes theorem, and (naively) assumes that the predictors are conditionally independent, given the class. Although the assumption is usually violated in practice, naive Bayes classifiers tend to yield posterior distributions that are robust to biased class density estimates, particularly where the posterior is 0.5 (the decision boundary) [1].

Naive Bayes classifiers assign observations to the most probable class (in other words, the maximum a posteriori decision rule). Explicitly, the algorithm takes these steps:

1. Estimate the densities of the predictors within each class.
2. Model posterior probabilities according to Bayes rule. That is, for all $k = 1,...,K$,

$$\widehat{P}(Y = k \,|\, X_1, .., X_P) = \frac{\pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)},$$

   where:

   - Y is the random variable corresponding to the class index of an observation.
   - $X_1,...,X_P$ are the random predictors of an observation.
   - $\pi(Y = k)$ is the prior probability that a class index is k.
3. Classify an observation by estimating the posterior probability for each class, and then assign the observation to the class yielding the maximum posterior probability.

If the predictors compose a multinomial distribution, then the posterior probability $\widehat{P}(Y = k \,|\, X_1, .., X_P) \propto \pi(Y = k) P_{mn}(X_1, ..., X_P \,|\, Y = k)$, where $P_{mn}(X_1, ..., X_P \,|\, Y = k)$ is the probability mass function of a multinomial distribution.

Algorithms

Estimated Probability for Multinomial Distribution

If you specify `'DistributionNames','mn'` when training `Mdl` using `fitcnb`, then the software fits a multinomial distribution using the “Bag-of-Tokens Model” on page 33-445. The software stores the probability that token j appears in class k in the property `DistributionParameters{k,j}`. With additive smoothing [2], the estimated probability is

$$P(\text{token } j \,|\, \text{class } k) = \frac{1 + c_{j|k}}{P + c_k},$$

where:

- $c_{j|k} = n_k \frac{\sum_{i:y_i \in \text{class } k} x_{ij} w_i}{\sum_{i:y_i \in \text{class } k} w_i}$, which is the weighted number of occurrences of token j in class k.
- n_k is the number of observations in class k.
- w_i is the weight for observation i. The software normalizes weights within a class so that they sum to the prior probability for that class.
- $c_k = \sum_{j=1}^{P} c_{j|k}$, which is the total weighted number of occurrences of all tokens in class k.

Estimated Probability for Multivariate Multinomial Distribution

If you specify `'DistributionNames','mvmn'` when training `Mdl` using `fitcnb`, then the software takes these steps:

1. For each predictor, the software collects a list of the unique levels, stores the sorted list in `CategoricalLevels`, and considers each level a bin. Each combination of predictor and class is a separate, independent multinomial random variable.
2. For predictor j in class k, the software counts instances of each categorical level using the list stored in `CategoricalLevels{`j`}`.
3. The software stores the probability that predictor j in class k has level L in the property `DistributionParameters{`k`,`j`}`, for all levels in `CategoricalLevels{`j`}`. With additive smoothing [2], the estimated probability is

   $$P(\text{predictor } j = L \,|\, \text{class } k) = \frac{1 + m_{j|k}(L)}{m_j + m_k},$$

   where:

   - $m_{j|k}(L) = n_k \frac{\sum_{i:y_i \in \text{class } k} I\{x_{ij} = L\} w_i}{\sum_{i:y_i \in \text{class } k} w_i}$, which is the weighted number of observations for which predictor j equals L in class k.
   - n_k is the number of observations in class k.
   - $I\{x_{ij} = L\} = 1$ if $x_{ij} = L$, and 0 otherwise.
   - w_i is the weight for observation i. The software normalizes weights within a class so that they sum to the prior probability for that class.
   - m_j is the number of distinct levels in predictor j.
   - m_k is the weighted number of observations in class k.

References

[1] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7.

[2] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*, NY: Cambridge University Press, 2008.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- When you train a naive Bayes model by using `fitcnb`, the following restrictions apply.
  - The value of the `'DistributionNames'` name-value pair argument cannot contain `'mn'`.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationNaiveBayes` | `fitcnb` | `loss` | `predict`

Topics

“Naive Bayes Classification” on page 21-2
“Grouping Variables” on page 2-45

Introduced in R2014b

CompactClassificationNeuralNetwork

Compact neural network model for classification

Description

`CompactClassificationNeuralNetwork` is a compact version of a `ClassificationNeuralNetwork` model object. The compact model does not include the data used for training the classifier. Therefore, you cannot perform some tasks, such as cross-validation, using the compact model. Use a compact model for tasks such as predicting the labels of new data.

Creation

Create a `CompactClassificationNeuralNetwork` object from a full `ClassificationNeuralNetwork` model object by using `compact`.

Properties

Neural Network Properties

`LayerSizes` — Sizes of fully connected layers
positive integer vector

This property is read-only.

Sizes of the fully connected layers in the neural network model, returned as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer. This layer always has *K* outputs, where *K* is the number of classes in the response variable.

Data Types: `single` | `double`

`LayerWeights` — Learned layer weights
cell array

This property is read-only.

Learned layer weights for the fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer weights for the *i*th fully connected layer. For example, `Mdl.LayerWeights{1}` returns the weights for the first fully connected layer of the model `Mdl`.

`LayerWeights` includes the weights for the final fully connected layer.

Data Types: `cell`

`LayerBiases` — Learned layer biases
cell array

This property is read-only.

Learned layer biases for the fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer biases for the *i*th fully connected layer. For example, `Mdl.LayerBiases{1}` returns the biases for the first fully connected layer of the model `Mdl`.

`LayerBiases` includes the biases for the final fully connected layer.

Data Types: `cell`

`Activations` — Activation functions for fully connected layers

`'relu'` | `'tanh'` | `'sigmoid'` | `'none'` | cell array of character vectors

This property is read-only.

Activation functions for the fully connected layers of the neural network model, returned as a character vector or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If `Activations` contains only one activation function, then it is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer. The activation function for the final fully connected layer is always softmax (`OutputLayerActivation`).
- If `Activations` is an array of activation functions, then the *i*th element is the activation function for the *i*th layer of the neural network model.

Data Types: `char` | `cell`

`OutputLayerActivation` — Activation function for final fully connected layer

`'softmax'`

This property is read-only.

Activation function for the final fully connected layer, returned as `'softmax'`. The function takes each input x_i and returns the following, where K is the number of classes in the response variable:

$$f(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)}.$$

The results correspond to the predicted classification scores (or posterior probabilities).

Data Properties

`PredictorNames` — Predictor variable names

cell array of character vectors

This property is read-only.

Predictor variable names, returned as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, returned as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names

cell array of character vectors

This property is read-only.

Expanded predictor names, returned as a cell array of character vectors. If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ClassNames` — Unique class names

numeric vector | categorical vector | logical vector | character array | cell array of character vectors

This property is read-only.

Unique class names used in training, returned as a numeric vector, categorical vector, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the class labels in the response variable used to train the model. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `cell`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, returned as a character vector.

Data Types: `char`

Other Classification Properties

`Cost` — Misclassification cost

numeric square matrix

This property is read-only.

Misclassification cost, returned as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. The cost matrix always has this form: `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`Prior` — Prior probabilities

numeric vector

This property is read-only.

Prior probabilities for each class, returned as a numeric vector. The order of the elements of `Prior` corresponds to the elements of `ClassNames`.

Data Types: `double`

`ScoreTransform` — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)

Value	Description
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  `Mdl.ScoreTransform = @`*`function`*`;`

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
edge	Classification edge for neural network classifier
loss	Classification loss for neural network classifier
margin	Classification margins for neural network classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using neural network classifier

Examples

Reduce Size of Neural Network Classifier

Reduce the size of a full neural network classifier by removing the training data from the model. You can use a compact model to improve memory efficiency.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Train a neural network classifier using the data. Specify the `Smoker` column of `tbl` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitcnet(tbl,"Smoker","Standardize",true)
```

```
Mdl =
  ClassificationNeuralNetwork
             PredictorNames: {'Diastolic'  'Systolic'  'Gender'  'Height'  'Weight'  'Age'}
               ResponseName: 'Smoker'
    CategoricalPredictors: 3
```

```
               ClassNames: [0 1]
           ScoreTransform: 'none'
          NumObservations: 100
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'softmax'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [36×7 table]


  Properties, Methods
```

`Mdl` is a full `ClassificationNeuralNetwork` model object.

Reduce the size of the model by using `compact`.

```
compactMdl = compact(Mdl)
```

```
compactMdl =
  CompactClassificationNeuralNetwork
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'softmax'


  Properties, Methods
```

`compactMdl` is a `CompactClassificationNeuralNetwork` model object. `compactMdl` contains fewer properties than the full model `Mdl`.

Display the amount of memory used by each neural network model.

```
whos("Mdl","compactMdl")
```

```
  Name            Size            Bytes  Class

  Mdl             1x1             18836  ClassificationNeuralNetwork
  compactMdl      1x1              6663  classreg.learning.classif.CompactClassificationNeuralNet
```

The full model is larger than the compact model.

See Also

`ClassificationNeuralNetwork` | `ClassificationPartitionedModel` | `compact` | `edge` | `fitcnet` | `loss` | `margin` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

CompactClassificationGAM

Compact generalized additive model (GAM) for binary classification

Description

`CompactClassificationGAM` is a compact version of a `ClassificationGAM` model object (GAM for binary classification). The compact model does not include the data used for training the classifier. Therefore, you cannot perform some tasks, such as cross-validation, using the compact model. Use a compact model for tasks such as predicting the labels of new data.

Creation

Create a `CompactClassificationGAM` object from a full `ClassificationGAM` model object by using `compact`.

Properties

GAM Properties

`Interactions` — Interaction term indices
two-column matrix of positive integers | `[]`

This property is read-only.

Interaction term indices, specified as a `t`-by-2 matrix of positive integers, where `t` is the number of interaction terms in the model. Each row of the matrix represents one interaction term and contains the column indexes of the predictor data `X` for the interaction term. If the model does not include an interaction term, then this property is empty (`[]`).

The software adds interaction terms to the model in the order of importance based on the *p*-values. Use this property to check the order of the interaction terms added to the model.

Data Types: `double`

`Intercept` — Intercept term of model
numeric scalar

This property is read-only.

Intercept (constant) term of the model, which is the sum of the intercept terms in the predictor trees and interaction trees, specified as a numeric scalar.

Data Types: `single` | `double`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Cost` — Misclassification costs

2-by-2 numeric matrix

Misclassification costs, specified as a 2-by-2 numeric matrix.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

The software uses the `Cost` value for prediction, but not training. You can change the value by using dot notation.

Example: `Mdl.Cost = C;`

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names

cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

`ExpandedPredictorNames` is the same as `PredictorNames` for a generalized additive model.

Data Types: `cell`

`PredictorNames` — Predictor variable names

cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`Prior` — Prior class probabilities

numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector with two elements. The order of the elements corresponds to the order of the elements in `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation
character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

This property determines the output score computation for object functions such as `predict`, `margin`, and `edge`. Use `'logit'` to compute posterior probabilities, and use `'none'` to compute the logit of posterior probabilities.

Data Types: `char` | `function_handle`

Object Functions

Interpret Prediction

lime	Local interpretable model-agnostic explanations (LIME)
partialDependence	Compute partial dependence
plotLocalEffects	Plot local effects of terms in generalized additive model (GAM)
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
shapley	Shapley values

Assess Predictive Performance on New Observations

predict	Classify observations using generalized additive model (GAM)
loss	Classification loss for generalized additive model (GAM)
margin	Classification margins for generalized additive model (GAM)
edge	Classification edge for generalized additive model (GAM)

Compare Accuracies

compareHoldout	Compare accuracies of two classification models using new data

Examples

Reduce Size of Generalized Additive Model

Reduce the size of a full generalized additive model (GAM) by removing the training data. Full models hold the training data. You can use a compact model to improve memory efficiency.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(X,Y,'ClassNames',{'b','g'})
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 2.2715
          NumObservations: 351
```

```
  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object.

Reduce the size of the classifier.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 2.2715


  Properties, Methods
```

`CMdl` is a `CompactClassificationGAM` model object.

Display the amount of memory used by each classifier.

```
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class                                                  Attri

  CMdl      1x1             1030010  classreg.learning.classif.CompactClassificationGAM
  Mdl       1x1             1230986  ClassificationGAM
```

The full classifier (`Mdl`) is larger than the compact classifier (`CMdl`).

To efficiently label new observations, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CMdl` and new predictor values to `predict`.

See Also

`ClassificationGAM` | `compact`

Topics

“Train Generalized Additive Model for Binary Classification” on page 12-77

Introduced in R2021a

CompactClassificationSVM

Compact support vector machine (SVM) for one-class and binary classification

Description

`CompactClassificationSVM` is a compact version of the support vector machine (SVM) classifier. The compact classifier does not include the data used for training the SVM classifier. Therefore, you cannot perform some tasks, such as cross-validation, using the compact classifier. Use a compact SVM classifier for tasks such as predicting the labels of new data.

Creation

Create a `CompactClassificationSVM` model from a full, trained `ClassificationSVM` classifier by using `compact`.

Properties

SVM Properties

`Alpha` — Trained classifier coefficients
numeric vector

This property is read-only.

Trained classifier coefficients, specified as an *s*-by-1 numeric vector. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`.

`Alpha` contains the trained classifier coefficients from the dual problem, that is, the estimated Lagrange multipliers. If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `Alpha` contains one coefficient corresponding to the entire set. That is, MATLAB attributes a nonzero coefficient to one observation from the set of duplicates and a coefficient of `0` to all other duplicate observations in the set.

Data Types: `single` | `double`

`Beta` — Linear predictor coefficients
numeric vector

This property is read-only.

Linear predictor coefficients, specified as a numeric vector. The length of `Beta` is equal to the number of predictors used to train the model.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. `Beta` stores one value for each predictor variable, including the dummy variables. For example, if there are three predictors, one of which is a categorical variable with three levels, then `Beta` is a numeric vector containing five values.

If `KernelParameters.Function` is `'linear'`, then the classification score for the observation x is

$$f(x) = (x/s)'\beta + b\,.$$

`Mdl` stores β, b, and s in the properties `Beta`, `Bias`, and `KernelParameters.Scale`, respectively.

To estimate classification scores manually, you must first apply any transformations to the predictor data that were applied during training. Specifically, if you specify `'Standardize',true` when using `fitcsvm`, then you must standardize the predictor data manually by using the mean `Mdl.Mu` and standard deviation `Mdl.Sigma`, and then divide the result by the kernel scale in `Mdl.KernelParameters.Scale`.

All SVM functions, such as `resubPredict` and `predict`, apply any required transformation before estimation.

If `KernelParameters.Function` is not `'linear'`, then `Beta` is empty (`[]`).

Data Types: `single` | `double`

`Bias` — Bias term
scalar

This property is read-only.

Bias term, specified as a scalar.

Data Types: `single` | `double`

`KernelParameters` — Kernel parameters
structure array

This property is read-only.

Kernel parameters, specified as a structure array. The kernel parameters property contains the fields listed in this table.

Field	Description
Function	Kernel function used to compute the elements of the Gram matrix on page 33-1862. For details, see `'KernelFunction'`.
Scale	Kernel scale parameter used to scale all elements of the predictor data on which the model is trained. For details, see `'KernelScale'`.

To display the values of `KernelParameters`, use dot notation. For example, `Mdl.KernelParameters.Scale` displays the kernel scale parameter value.

The software accepts `KernelParameters` as inputs and does not modify them.

Data Types: `struct`

`SupportVectorLabels` — Support vector class labels
s-by-1 numeric vector

This property is read-only.

Support vector class labels, specified as an *s*-by-1 numeric vector. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`.

A value of `+1` in `SupportVectorLabels` indicates that the corresponding support vector is in the positive class (`ClassNames{2}`). A value of `–1` indicates that the corresponding support vector is in the negative class (`ClassNames{1}`).

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `SupportVectorLabels` contains one unique support vector label.

Data Types: `single` | `double`

`SupportVectors` — Support vectors
s-by-*p* numeric matrix

This property is read-only.

Support vectors in the trained classifier, specified as an *s*-by-*p* numeric matrix. *s* is the number of support vectors in the trained classifier, `sum(Mdl.IsSupportVector)`, and *p* is the number of predictor variables in the predictor data.

`SupportVectors` contains rows of the predictor data `X` that MATLAB considers to be support vectors. If you specify `'Standardize',true` when training the SVM classifier using `fitcsvm`, then `SupportVectors` contains the standardized rows of `X`.

If you remove duplicates by using the `RemoveDuplicates` name-value pair argument of `fitcsvm`, then for a given set of duplicate observations that are support vectors, `SupportVectors` contains one unique support vector.

Data Types: `single` | `double`

Other Classification Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Cost` — Misclassification cost
numeric square matrix

This property is read-only.

Misclassification cost, specified as a numeric square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`.

During training, the software updates the prior probabilities by incorporating the penalties described in the cost matrix.

- For two-class learning, `Cost` always has this form: `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. The rows correspond to the true class and the columns correspond to the predicted class. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.
- For one-class learning, `Cost = 0`.

For more details, see Algorithms on page 33-541.

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

If the model uses dummy variable encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`Mu` — Predictor means
numeric vector | `[]`

This property is read-only.

Predictor means, specified as a numeric vector. If you specify `'Standardize',1` or `'Standardize',true` when you train an SVM classifier using `fitcsvm`, then the length of `Mu` is equal to the number of predictors.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. `Mu` stores one value for each predictor variable, including the dummy variables. However, MATLAB does not standardize the columns that contain categorical variables.

If you set `'Standardize',false` when you train the SVM classifier using `fitcsvm`, then `Mu` is an empty vector (`[]`).

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`Prior` — Prior probabilities

numeric vector

This property is read-only.

Prior probabilities for each class, specified as a numeric vector. The order of the elements of `Prior` corresponds to the elements of `Mdl.ClassNames`.

For two-class learning, if you specify a cost matrix, then the software updates the prior probabilities by incorporating the penalties described in the cost matrix.

For more details, see Algorithms on page 33-541.

Data Types: `single` | `double`

`ScoreTransform` — Score transformation

character vector | function handle

Score transformation, specified as a character vector or function handle. `ScoreTransform` represents a built-in transformation function or a function handle for transforming predicted classification scores.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter a character vector.

  ```
  Mdl.ScoreTransform = 'function';
  ```

  This table describes the available built-in functions.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Data Types: `char` | `function_handle`

`Sigma` — Predictor standard deviations
`[]` (default) | numeric vector

This property is read-only.

Predictor standard deviations, specified as a numeric vector.

If you specify `'Standardize',true` when you train the SVM classifier using `fitcsvm`, then the length of `Sigma` is equal to the number of predictor variables.

MATLAB expands categorical variables in the predictor data using full dummy encoding. That is, MATLAB creates one dummy variable for each level of each categorical variable. `Sigma` stores one value for each predictor variable, including the dummy variables. However, MATLAB does not standardize the columns that contain categorical variables.

If you set `'Standardize',false` when you train the SVM classifier using `fitcsvm`, then `Sigma` is an empty vector (`[]`).

Data Types: `single` | `double`

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
discardSupportVectors	Discard support vectors for linear support vector machine (SVM) classifier
edge	Find classification edge for support vector machine (SVM) classifier
fitPosterior	Fit posterior probabilities for compact support vector machine (SVM) classifier
incrementalLearner	Convert binary classification support vector machine (SVM) model to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
loss	Find classification error for support vector machine (SVM) classifier
margin	Find classification margins for support vector machine (SVM) classifier
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Classify observations using support vector machine (SVM) classifier
shapley	Shapley values
update	Update model parameters for code generation

Examples

Reduce Size of SVM Classifier

Reduce the size of a full support vector machine (SVM) classifier by removing the training data. Full SVM classifiers (that is, `ClassificationSVM` classifiers) hold the training data. To improve efficiency, use a smaller classifier.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Standardize the predictor data and specify the order of the classes.

```
SVMModel = fitcsvm(X,Y,'Standardize',true,...
    'ClassNames',{'b','g'})

SVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
                    Alpha: [90x1 double]
                     Bias: -0.1343
         KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           BoxConstraints: [351x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [351x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`SVMModel` is a `ClassificationSVM` classifier.

Reduce the size of the SVM classifier.

```
CompactSVMModel = compact(SVMModel)

CompactSVMModel =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
                    Alpha: [90x1 double]
                     Bias: -0.1343
         KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           SupportVectors: [90x34 double]
      SupportVectorLabels: [90x1 double]


  Properties, Methods
```

`CompactSVMModel` is a `CompactClassificationSVM` classifier.

Display the amount of memory used by each classifier.

```
whos('SVMModel','CompactSVMModel')

  Name                 Size              Bytes  Class

  CompactSVMModel      1x1               31058  classreg.learning.classif.CompactClassificationSVM
  SVMModel             1x1              141148  ClassificationSVM
```

The full SVM classifier (`SVMModel`) is more than four times larger than the compact SVM classifier (`CompactSVMModel`).

To label new observations efficiently, you can remove `SVMModel` from the MATLAB® Workspace, and then pass `CompactSVMModel` and new predictor values to `predict`.

To further reduce the size of the compact SVM classifier, use the `discardSupportVectors` function to discard support vectors.

Train and Cross-Validate SVM Classifier

Load the `ionosphere` data set.

```
load ionosphere
```

Train and cross-validate an SVM classifier. Standardize the predictor data and specify the order of the classes.

```
rng(1);  % For reproducibility
CVSVMModel = fitcsvm(X,Y,'Standardize',true,...
    'ClassNames',{'b','g'},'CrossVal','on')
```

```
CVSVMModel =
  ClassificationPartitionedModel
    CrossValidatedModel: 'SVM'
         PredictorNames: {1x34 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVSVMModel` is a `ClassificationPartitionedModel` cross-validated SVM classifier. By default, the software implements 10-fold cross-validation.

Alternatively, you can cross-validate a trained `ClassificationSVM` classifier by passing it to `crossval`.

Inspect one of the trained folds using dot notation.

```
CVSVMModel.Trained{1}
```

```
ans =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
                    Alpha: [78x1 double]
                     Bias: -0.2209
```

```
         KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           SupportVectors: [78x34 double]
      SupportVectorLabels: [78x1 double]


  Properties, Methods
```

Each fold is a `CompactClassificationSVM` classifier trained on 90% of the data.

Estimate the generalization error.

```
genError = kfoldLoss(CVSVMModel)
```

```
genError = 0.1168
```

On average, the generalization error is approximately 12%.

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[2] Scholkopf, B., J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. “Estimating the Support of a High-Dimensional Distribution.” *Neural Computation*. Vol. 13, Number 7, 2001, pp. 1443–1471.

[3] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[4] Scholkopf, B., and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press, 2002.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of an SVM classification model into Simulink, you can use the ClassificationSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an SVM model by using `fitcsvm`, the following restrictions apply.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function. For generating code that predicts posterior probabilities given new observations, pass a trained SVM model to `fitPosterior` or `fitSVMPosterior`. The `ScoreTransform` property of the returned model contains an anonymous function that represents the score-to-posterior-probability function and is configured for code generation.

- For fixed-point code generation, the value of the `'ScoreTransform'` name-value pair argument cannot be `'invlogit'`. Also, the value of the `'KernelFunction'` name-value pair argument must be `'gaussian'`, `'linear'`, or `'polynomial'`.
- For fixed-point code generation and code generation with a coder configurer, the following additional restrictions apply.
  - Categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`) are not supported. You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.
  - Class labels with the `categorical` data type are not supported. Both the class label value in training data (`Tbl` or `Y`) and the value of the `'ClassNames'` name-value argument cannot be an array with the `categorical` data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationSVM` | `compact` | `discardSupportVectors` | `fitcsvm`

Topics

Using Support Vector Machines on page 18-154
Understanding Support Vector Machines on page 18-150

Introduced in R2014a

CompactClassificationTree

Package: `classreg.learning.classif`

Compact classification tree

Description

Compact version of a classification tree (of class `ClassificationTree`). The compact version does not include the data for training the classification tree. Therefore, you cannot perform some tasks with a compact classification tree, such as cross validation. Use a compact classification tree for making predictions (classifications) of new data.

Construction

`ctree = compact(tree)` constructs a compact decision tree from a full decision tree.

Input Arguments

`tree`

A decision tree constructed using `fitctree`.

Properties

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CategoricalSplits`

An *n*-by-2 cell array, where *n* is the number of categorical splits in `tree`. Each row in `CategoricalSplits` gives left and right values for a categorical split. For each branch node with categorical split `j` based on a categorical predictor variable `z`, the left child is chosen if `z` is in `CategoricalSplits(j,1)` and the right child is chosen if `z` is in `CategoricalSplits(j,2)`. The splits are in the same order as nodes of the tree. Find the nodes for these splits by selecting `'categorical'` cuts from top to bottom in the `CutType` property.

`Children`

An *n*-by-2 array containing the numbers of the child nodes for each node in `tree`, where *n* is the number of nodes. Leaf nodes have child node `0`.

`ClassCount`

An *n*-by-*k* array of class counts for the nodes in `tree`, where *n* is the number of nodes and *k* is the number of classes. For any node number `i`, the class counts `ClassCount(i,:)` are counts of observations (from the data used in fitting the tree) from each class satisfying the conditions for node `i`.

`ClassNames`

List of the elements in `Y` with duplicates removed. `ClassNames` can be a numeric vector, vector of categorical variables, logical vector, character array, or cell array of character vectors. `ClassNames` has the same data type as the data in the argument `Y`. (The software treats string arrays as cell arrays of character vectors.)

If the value of a property has at least one dimension of length *k*, then `ClassNames` indicates the order of the elements along that dimension (e.g., `Cost` and `Prior`).

`ClassProbability`

An *n*-by-*k* array of class probabilities for the nodes in `tree`, where *n* is the number of nodes and *k* is the number of classes. For any node number `i`, the class probabilities `ClassProbability(i,:)` are the estimated probabilities for each class for a point satisfying the conditions for node `i`.

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response. This property is read-only.

`CutCategories`

An *n*-by-2 cell array of the categories used at branches in `tree`, where *n* is the number of nodes. For each branch node `i` based on a categorical predictor variable `x`, the left child is chosen if `x` is among the categories listed in `CutCategories{i,1}`, and the right child is chosen if `x` is among those listed in `CutCategories{i,2}`. Both columns of `CutCategories` are empty for branch nodes based on continuous predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPoint`

An *n*-element vector of the values used as cut points in `tree`, where *n* is the number of nodes. For each branch node `i` based on a continuous predictor variable `x`, the left child is chosen if `x<CutPoint(i)` and the right child is chosen if `x>=CutPoint(i)`. `CutPoint` is `NaN` for branch nodes based on categorical predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutType`

An *n*-element cell array indicating the type of cut at each node in `tree`, where *n* is the number of nodes. For each node `i`, `CutType{i}` is:

- `'continuous'` — If the cut is defined in the form `x < v` for a variable `x` and cut point `v`.
- `'categorical'` — If the cut is defined by whether a variable `x` takes a value in a set of categories.
- `''` — If `i` is a leaf node.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictor`

An *n*-element cell array of the names of the variables used for branching in each node in `tree`, where *n* is the number of nodes. These variables are sometimes known as *cut variables*. For leaf nodes, `CutPredictor` contains an empty character vector.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictorIndex`

An *n*-element array of numeric indices for the variables used for branching in each node in `tree`, where *n* is the number of nodes. For more information, see `CutPredictor`.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`IsBranchNode`

An *n*-element logical vector that is `true` for each branch node and `false` for each leaf node of `tree`.

`NodeClass`

An *n*-element cell array with the names of the most probable classes in each node of `tree`, where *n* is the number of nodes in the tree. Every element of this array is a character vector equal to one of the class names in `ClassNames`.

`NodeError`

An *n*-element vector of the errors of the nodes in `tree`, where *n* is the number of nodes. `NodeError(i)` is the misclassification probability for node `i`.

`NodeProbability`

An *n*-element vector of the probabilities of the nodes in `tree`, where *n* is the number of nodes. The probability of a node is computed as the proportion of observations from the original data that satisfy the conditions for the node. This proportion is adjusted for any prior probabilities assigned to each class.

`NodeRisk`

An *n*-element vector of the risk of the nodes in the tree, where *n* is the number of nodes. The risk for each node is the measure of impurity (Gini index or deviance) for this node weighted by the node probability. If the tree is grown by twoing, the risk for each node is zero.

NodeSize

An *n*-element vector of the sizes of the nodes in `tree`, where *n* is the number of nodes. The size of a node is defined as the number of observations from the data used to create the tree that satisfy the conditions for the node.

NumNodes

The number of nodes in `tree`.

Parent

An *n*-element vector containing the number of the parent node for each node in `tree`, where *n* is the number of nodes. The parent of the root node is `0`.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in `X`.

Prior

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`. The number of elements of `Prior` is the number of unique classes in the response. This property is read-only.

PruneAlpha

Numeric vector with one element per pruning level. If the pruning level ranges from 0 to *M*, then `PruneAlpha` has *M* + 1 elements sorted in ascending order. `PruneAlpha(1)` is for pruning level 0 (no pruning), `PruneAlpha(2)` is for pruning level 1, and so on.

PruneList

An *n*-element numeric vector with the pruning levels in each node of `tree`, where *n* is the number of nodes. The pruning levels range from 0 (no pruning) to *M*, where *M* is the distance between the deepest leaf and the root node.

ResponseName

Character vector describing the response variable `Y`.

ScoreTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`. For a list of built-in transformation functions and the syntax of custom transformation functions, see `fitctree`.

Add or change a `ScoreTransform` function using dot notation:

```
ctree.ScoreTransform = 'function'
or
ctree.ScoreTransform = @function
```

SurrogateCutCategories

An *n*-element cell array of the categories used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutCategories{k}` is a cell array. The length of

`SurrogateCutCategories{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutCategories{k}` is either an empty character vector for a continuous surrogate predictor, or is a two-element cell array with categories for a categorical surrogate predictor. The first element of this two-element cell array lists categories assigned to the left child by this surrogate split and the second element of this two-element cell array lists categories assigned to the right child by this surrogate split. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutVar`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutCategories` contains an empty cell.

`SurrogateCutFlip`

An *n*-element cell array of the numeric cut assignments used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrSurrogateCutFlip{k}` is a numeric vector. The length of `SurrogateCutFlip{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutFlip{k}` is either zero for a categorical surrogate predictor, or a numeric cut assignment for a continuous surrogate predictor. The numeric cut assignment can be either –1 or +1. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and the cut assignment for this surrogate split is +1, or if $Z\geq C$ and the cut assignment for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and the cut assignment for this surrogate split is +1, or if $Z<C$ and the cut assignment for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutFlip` contains an empty array.

`SurrogateCutPoint`

An *n*-element cell array of the numeric values used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutPoint{k}` is a numeric vector. The length of `SurrogateCutPoint{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutPoint{k}` is either `NaN` for a categorical surrogate predictor, or a numeric cut for a continuous surrogate predictor. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z<C$ and `SurrogateCutFlip` for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables returned by `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPoint` contains an empty cell.

`SurrogateCutType`

An *n*-element cell array indicating types of surrogate splits at each node in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutType{k}` is a cell array with the types of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutType` contains an empty cell. A surrogate split type can be either `'continuous'` if the cut is defined in the form `Z<V` for a variable `Z` and cut point `V` or `'categorical'` if the cut is defined by whether `Z` takes a value in a set of categories.

SurrogateCutPredictor

An *n*-element cell array of the names of the variables used for surrogate splits in each node in `tree`, where *n* is the number of nodes in `tree`. Every element of `SurrogateCutPredictor` is a cell array with the names of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPredictor` contains an empty cell.

SurrogatePredictorAssociation

An *n*-element cell array of the predictive measures of association for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogatePredictorAssociation{k}` is a numeric vector. The length of `SurrogatePredictorAssociation{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogatePredictorAssociation{k}` gives the predictive measure of association between the optimal split and this surrogate split. The order of the surrogate split variables at each node is the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogatePredictorAssociation` contains an empty cell.

Object Functions

compareHoldout	Compare accuracies of two classification models using new data
edge	Classification edge
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification error
margin	Classification margins
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels using classification tree
predictorImportance	Estimates of predictor importance for classification tree
shapley	Shapley values
surrogateAssociation	Mean predictive measure of association for surrogate splits in classification tree
update	Update model parameters for code generation
view	View classification tree

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Construct a Compact Classification Tree

Construct a compact classification tree for the Fisher iris data.

```
load fisheriris
tree = fitctree(meas,species);
ctree = compact(tree);
```

Compare the size of the resulting tree to that of the original tree.

```
t = whos('tree'); % t.bytes = size of tree in bytes
c = whos('ctree'); % c.bytes = size of ctree in bytes
[c.bytes t.bytes]
```

```
ans = 1×2

        5097       11762
```

The compact tree is smaller than the original tree.

More About

Impurity and Node Error

A decision tree splits nodes based on either impurity or node error.

Impurity means one of several things, depending on your choice of the `SplitCriterion` name-value pair argument:

- Gini's Diversity Index (`gdi`) — The Gini index of a node is

  $$1 - \sum_i p^2(i),$$

  where the sum is over the classes i at the node, and $p(i)$ is the observed fraction of classes with class i that reach the node. A node with just one class (a pure node) has Gini index `0`; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.
- Deviance (`'deviance'`) — With $p(i)$ defined the same as for the Gini index, the deviance of a node is

  $$-\sum_i p(i)\log_2 p(i)\,.$$

  A pure node has deviance `0`; otherwise, the deviance is positive.
- Twoing rule (`'twoing'`) — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class i in the left child node after a split, and $R(i)$ denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

  $$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2,$$

  where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.
- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $1 – p(j)$.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of a classification tree model into Simulink, you can use the ClassificationTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train a classification tree using `fitctree`, the following restrictions apply.
  - The value of the `'ScoreTransform'` name-value pair argument cannot be an anonymous function. For fixed-point code generation, the `'ScoreTransform'` value cannot be `'invlogit'`.
  - You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value pair argument must be `'off'`.
  - For fixed-point code generation and code generation with a coder configurer, the following additional restrictions apply.
    - Categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`) are not supported. You cannot use the `'CategoricalPredictors'` name-value argument.To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.
    - Class labels with the `categorical` data type are not supported. Both the class label value in training data (`Tbl` or `Y`) and the value of the `'ClassNames'` name-value argument cannot be an array with the `categorical` data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationTree` | `compact` | `compareHoldout` | `fitctree`

Introduced in R2011a

CompactLinearModel

Compact linear regression model

Description

`CompactLinearModel` is a compact version of a full linear regression model object `LinearModel`. Because a compact model does not store the input data used to fit the model or information related to the fitting process, a `CompactLinearModel` object consumes less memory than a `LinearModel` object. You can still use a compact model to predict responses using new input data, but some `LinearModel` object functions do not work with a compact model.

Creation

Create a `CompactLinearModel` model from a full, trained `LinearModel` model by using `compact`.

Properties

Coefficient Estimates

`CoefficientCovariance` — Covariance matrix of coefficient estimates
numeric matrix

This property is read-only.

Covariance matrix of coefficient estimates, specified as a *p*-by-*p* matrix of numeric values. *p* is the number of coefficients in the fitted model.

For details, see “Coefficient Standard Errors and Confidence Intervals” on page 11-58.

Data Types: `single` | `double`

`CoefficientNames` — Coefficient names
cell array of character vectors

This property is read-only.

Coefficient names, specified as a cell array of character vectors, each containing the name of the corresponding term.

Data Types: `cell`

`Coefficients` — Coefficient values
table

This property is read-only.

Coefficient values, specified as a table. `Coefficients` contains one row for each coefficient and these columns:

- `Estimate` — Estimated coefficient value
- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is zero
- `pValue` — *p*-value for the *t*-statistic

Use `anova` (only for a linear regression model) or `coefTest` to perform other tests on the coefficients. Use `coefCI` to find the confidence intervals of the coefficient estimates.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the estimated coefficient vector in the model `mdl`:

```
beta = mdl.Coefficients.Estimate
```

Data Types: `table`

`NumCoefficients` — Number of model coefficients
positive integer

This property is read-only.

Number of model coefficients, specified as a positive integer. `NumCoefficients` includes coefficients that are set to zero when the model terms are rank deficient.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated coefficients
positive integer

This property is read-only.

Number of estimated coefficients in the model, specified as a positive integer. `NumEstimatedCoefficients` does not include coefficients that are set to zero when the model terms are rank deficient. `NumEstimatedCoefficients` is the degrees of freedom for regression.

Data Types: `double`

Summary Statistics

`DFE` — Degrees of freedom for error
positive integer

This property is read-only.

Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients, specified as a positive integer.

Data Types: `double`

`LogLikelihood` — Loglikelihood
numeric value

This property is read-only.

Loglikelihood of response values, specified as a numeric value, based on the assumption that each response value follows a normal distribution. The mean of the normal distribution is the fitted (predicted) response value, and the variance is the `MSE`.

Data Types: `single` | `double`

`ModelCriterion` — Criterion for model comparison

structure

This property is read-only.

Criterion for model comparison, specified as a structure with these fields:

- `AIC` — Akaike information criterion. `AIC = –2*logL + 2*m`, where `logL` is the loglikelihood and `m` is the number of estimated parameters.
- `AICc` — Akaike information criterion corrected for the sample size. `AICc = AIC + (2*m*(m + 1))/(n – m – 1)`, where `n` is the number of observations.
- `BIC` — Bayesian information criterion. `BIC = –2*logL + m*log(n)`.
- `CAIC` — Consistent Akaike information criterion. `CAIC = –2*logL + m*(log(n) + 1)`.

Information criteria are model selection tools that you can use to compare multiple models fit to the same data. These criteria are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). Different information criteria are distinguished by the form of the penalty.

When you compare multiple models, the model with the lowest information criterion value is the best-fitting model. The best-fitting model can vary depending on the criterion used for model comparison.

To obtain any of the criterion values as a scalar, index into the property using dot notation. For example, obtain the AIC value `aic` in the model `mdl`:

```
aic = mdl.ModelCriterion.AIC
```

Data Types: `struct`

`MSE` — Mean squared error

numeric value

This property is read-only.

Mean squared error (residuals), specified as a numeric value.

$$MSE = SSE / DFE,$$

where *MSE* is the mean squared error, *SSE* is the sum of squared errors, and *DFE* is the degrees of freedom.

Data Types: `single` | `double`

`RMSE` — Root mean squared error

numeric value

This property is read-only.

Root mean squared error (residuals), specified as a numeric value.

$$RMSE = \mathrm{sqrt}(MSE),$$

where *RMSE* is the root mean squared error and *MSE* is the mean squared error.

Data Types: `single` | `double`

Rsquared — R-squared value for model
structure

This property is read-only.

R-squared value for the model, specified as a structure with two fields:

- `Ordinary` — Ordinary (unadjusted) R-squared
- `Adjusted` — R-squared adjusted for the number of coefficients

The R-squared value is the proportion of the total sum of squares explained by the model. The ordinary R-squared value relates to the `SSR` and `SST` properties:

```
Rsquared = SSR/SST,
```

where `SST` is the total sum of squares, and `SSR` is the regression sum of squares.

For details, see “Coefficient of Determination (R-Squared)” on page 11-61.

To obtain either of these values as a scalar, index into the property using dot notation. For example, obtain the adjusted R-squared value in the model `mdl`:

```
r2 = mdl.Rsquared.Adjusted
```

Data Types: `struct`

SSE — Sum of squared errors
numeric value

This property is read-only.

Sum of squared errors (residuals), specified as a numeric value.

The Pythagorean theorem implies

```
SST = SSE + SSR,
```

where `SST` is the total sum of squares, `SSE` is the sum of squared errors, and `SSR` is the regression sum of squares.

Data Types: `single` | `double`

SSR — Regression sum of squares
numeric value

This property is read-only.

Regression sum of squares, specified as a numeric value. The regression sum of squares is equal to the sum of squared deviations of the fitted values from their mean.

The Pythagorean theorem implies

```
SST = SSE + SSR,
```

where `SST` is the total sum of squares, `SSE` is the sum of squared errors, and `SSR` is the regression sum of squares.

Data Types: `single` | `double`

SST — Total sum of squares
numeric value

This property is read-only.

Total sum of squares, specified as a numeric value. The total sum of squares is equal to the sum of squared deviations of the response vector `y` from the `mean(y)`.

The Pythagorean theorem implies

`SST = SSE + SSR`,

where `SST` is the total sum of squares, `SSE` is the sum of squared errors, and `SSR` is the regression sum of squares.

Data Types: `single` | `double`

Fitting Method

`Robust` — Robust fit information

structure

This property is read-only.

Robust fit information, specified as a structure with the fields described in this table.

Field	Description
`WgtFun`	Robust weighting function, such as `'bisquare'` (see `'RobustOpts'`)
`Tune`	Tuning constant. This field is empty (`[]`) if `WgtFun` is `'ols'` or if `WgtFun` is a function handle for a custom weight function with the default tuning constant 1.
`Weights`	Vector of weights used in the final iteration of robust fit. This field is empty for a `CompactLinearModel` object.

This structure is empty unless you fit the model using robust regression.

Data Types: `struct`

Input Data

`Formula` — Model information

`LinearFormula` object

This property is read-only.

Model information, specified as a `LinearFormula` object.

Display the formula of the fitted model `mdl` using dot notation:

```
mdl.Formula
```

`NumObservations` — Number of observations

positive integer

This property is read-only.

Number of observations the fitting function used in fitting, specified as a positive integer. `NumObservations` is the number of observations supplied in the original table, dataset, or matrix, minus any excluded rows (set with the `'Exclude'` name-value pair argument) or rows with missing values.

Data Types: `double`

`NumPredictors` — Number of predictor variables
positive integer

This property is read-only.

Number of predictor variables used to fit the model, specified as a positive integer.

Data Types: `double`

`NumVariables` — Number of variables
positive integer

This property is read-only.

Number of variables in the input data, specified as a positive integer. `NumVariables` is the number of variables in the original table or dataset, or the total number of columns in the predictor matrix and response vector.

`NumVariables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `double`

`PredictorNames` — Names of predictors used to fit model
cell array of character vectors

This property is read-only.

Names of predictors used to fit the model, specified as a cell array of character vectors.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`VariableInfo` — Information about variables
table

This property is read-only.

Information about variables contained in `Variables`, specified as a table with one row for each variable and the columns described in this table.

Column	Description
`Class`	Variable class, specified as a cell array of character vectors, such as `'double'` and `'categorical'`

Column	Description
`Range`	Variable range, specified as a cell array of vectors • Continuous variable — Two-element vector `[min,max]`, the minimum and maximum values • Categorical variable — Vector of distinct variable values
`InModel`	Indicator of which variables are in the fitted model, specified as a logical vector. The value is `true` if the model includes the variable.
`IsCategorical`	Indicator of categorical variables, specified as a logical vector. The value is `true` if the variable is categorical.

`VariableInfo` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

`VariableNames` — Names of variables
cell array of character vectors

This property is read-only.

Names of variables, specified as a cell array of character vectors.

- If the fit is based on a table or dataset, this property provides the names of the variables in the table or dataset.
- If the fit is based on a predictor matrix and response vector, `VariableNames` contains the values specified by the `'VarNames'` name-value pair argument of the fitting method. The default value of `'VarNames'` is `{'x1','x2',...,'xn','y'}`.

`VariableNames` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `cell`

Object Functions

Predict Responses

feval	Predict responses of linear regression model using one input for each predictor
predict	Predict responses of linear regression model
random	Simulate responses with random noise for linear regression model

Evaluate Linear Model

anova	Analysis of variance for linear regression model
coefCI	Confidence intervals of coefficient estimates of linear regression model
coefTest	Linear hypothesis test on linear regression model coefficients
partialDependence	Compute partial dependence

Visualize Linear Model

plotEffects	Plot main effects of predictors in linear regression model

plotInteraction	Plot interaction effects of two predictors in linear regression model
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotSlice	Plot of slices through fitted linear regression surface

Gather Properties of Linear Model

gather	Gather properties of machine learning model from GPU

Examples

Compact Linear Regression Model

Fit a linear regression model to data and reduce the size of a full, fitted linear regression model by discarding the sample data and some information related to the fitting process.

Load the `largedata4reg` data set, which contains 15,000 observations and 45 predictor variables.

```
load largedata4reg
```

Fit a linear regression model to the data.

```
mdl = fitlm(X,Y);
```

Compact the model.

```
compactMdl = compact(mdl);
```

The compact model discards the original sample data and some information related to the fitting process.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans = 1×2

       81537    11408528
```

The compact model consumes less memory than the full model.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `random` functions support code generation.

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions fully support GPU arrays:
  - `feval`
  - `predict`
  - `random`
  - `partialDependence`
  - `plotPartialDependence`
- The following object functions support model objects fitted with GPU array input arguments:
  - `anova`
  - `coefCI`
  - `coefTest`
  - `plotEffects`
  - `plotInteraction`
  - `plotSlice`
  - `gather`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `compact` | `fitlm` | `stepwiselm`

Topics

“Linear Regression” on page 11-9

Introduced in R2016a

CompactGeneralizedLinearModel

Compact generalized linear regression model class

Description

`CompactGeneralizedLinearModel` is a compact version of a full generalized linear regression model object `GeneralizedLinearModel`. Because a compact model does not store the input data used to fit the model or information related to the fitting process, a `CompactGeneralizedLinearModel` object consumes less memory than a `GeneralizedLinearModel` object. You can still use a compact model to predict responses using new input data, but some `GeneralizedLinearModel` object functions do not work with a compact model.

Creation

Create a `CompactGeneralizedLinearModel` model from a full, trained `GeneralizedLinearModel` model by using `compact`.

`fitglm` returns `CompactGeneralizedLinearModel` when you work with tall arrays, and returns `GeneralizedLinearModel` when you work with in-memory tables and arrays.

Properties

Coefficient Estimates

`CoefficientCovariance` — Covariance matrix of coefficient estimates
numeric matrix

This property is read-only.

Covariance matrix of coefficient estimates, specified as a *p*-by-*p* matrix of numeric values. *p* is the number of coefficients in the fitted model.

For details, see “Coefficient Standard Errors and Confidence Intervals” on page 11-58.

Data Types: `single` | `double`

`CoefficientNames` — Coefficient names
cell array of character vectors

This property is read-only.

Coefficient names, specified as a cell array of character vectors, each containing the name of the corresponding term.

Data Types: `cell`

`Coefficients` — Coefficient values
table

This property is read-only.

Coefficient values, specified as a table. `Coefficients` contains one row for each coefficient and these columns:

- `Estimate` — Estimated coefficient value
- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is zero
- `pValue` — *p*-value for the *t*-statistic

Use `anova` (only for a linear regression model) or `coefTest` to perform other tests on the coefficients. Use `coefCI` to find the confidence intervals of the coefficient estimates.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the estimated coefficient vector in the model `mdl`:

```
beta = mdl.Coefficients.Estimate
```

Data Types: `table`

`NumCoefficients` — Number of model coefficients
positive integer

This property is read-only.

Number of model coefficients, specified as a positive integer. `NumCoefficients` includes coefficients that are set to zero when the model terms are rank deficient.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated coefficients
positive integer

This property is read-only.

Number of estimated coefficients in the model, specified as a positive integer. `NumEstimatedCoefficients` does not include coefficients that are set to zero when the model terms are rank deficient. `NumEstimatedCoefficients` is the degrees of freedom for regression.

Data Types: `double`

Summary Statistics

`Deviance` — Deviance of fit
numeric value

This property is read-only.

Deviance of the fit, specified as a numeric value. The deviance is useful for comparing two models when one model is a special case of the other model. The difference between the deviance of the two models has a chi-square distribution with degrees of freedom equal to the difference in the number of estimated parameters between the two models. For more information, see “Deviance” on page 33-834.

Data Types: `single` | `double`

DFE — Degrees of freedom for error
positive integer

This property is read-only.

Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients, specified as a positive integer.

Data Types: `double`

Dispersion — Scale factor of variance of response
numeric scalar

This property is read-only.

Scale factor of the variance of the response, specified as a numeric scalar.

If the `'DispersionFlag'` name-value pair argument of `fitglm` or `stepwiseglm` is `true`, then the function estimates the `Dispersion` scale factor in computing the variance of the response. The variance of the response equals the theoretical variance multiplied by the scale factor.

For example, the variance function for the binomial distribution is $p(1–p)/n$, where p is the probability parameter and n is the sample size parameter. If `Dispersion` is near `1`, the variance of the data appears to agree with the theoretical variance of the binomial distribution. If `Dispersion` is larger than `1`, the data set is “overdispersed” relative to the binomial distribution.

Data Types: `double`

DispersionEstimated — Flag to indicate use of dispersion scale factor
logical value

This property is read-only.

Flag to indicate whether `fitglm` used the `Dispersion` scale factor to compute standard errors for the coefficients in `Coefficients.SE`, specified as a logical value. If `DispersionEstimated` is `false`, `fitglm` used the theoretical value of the variance.

- `DispersionEstimated` can be `false` only for the binomial and Poisson distributions.
- Set `DispersionEstimated` by setting the `'DispersionFlag'` name-value pair argument of `fitglm` or `stepwiseglm`.

Data Types: `logical`

LogLikelihood — Loglikelihood
numeric value

This property is read-only.

Loglikelihood of the model distribution at the response values, specified as a numeric value. The mean is fitted from the model, and other parameters are estimated as part of the model fit.

Data Types: `single` | `double`

ModelCriterion — Criterion for model comparison
structure

This property is read-only.

Criterion for model comparison, specified as a structure with these fields:

- `AIC` — Akaike information criterion. `AIC = –2*logL + 2*m`, where `logL` is the loglikelihood and `m` is the number of estimated parameters.
- `AICc` — Akaike information criterion corrected for the sample size. `AICc = AIC + (2*m*(m + 1))/(n – m – 1)`, where `n` is the number of observations.
- `BIC` — Bayesian information criterion. `BIC = –2*logL + m*log(n)`.
- `CAIC` — Consistent Akaike information criterion. `CAIC = –2*logL + m*(log(n) + 1)`.

Information criteria are model selection tools that you can use to compare multiple models fit to the same data. These criteria are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). Different information criteria are distinguished by the form of the penalty.

When you compare multiple models, the model with the lowest information criterion value is the best-fitting model. The best-fitting model can vary depending on the criterion used for model comparison.

To obtain any of the criterion values as a scalar, index into the property using dot notation. For example, obtain the AIC value `aic` in the model `mdl`:

```
aic = mdl.ModelCriterion.AIC
```

Data Types: `struct`

Rsquared — R-squared value for model
structure

This property is read-only.

R-squared value for the model, specified as a structure with five fields.

Field	Description	Equation
`Ordinary`	Ordinary (unadjusted) R-squared	$R^2_{\text{Ordinary}} = 1 - \frac{\text{SSE}}{\text{SST}}$ `SSE` is the sum of squared errors, and `SST` is the total sum of squared deviations of the response vector from the mean of the response vector.
`Adjusted`	R-squared adjusted for the number of coefficients	$R^2_{\text{Adjusted}} = 1 - \frac{\text{SSE}}{\text{SST}} \cdot \frac{N-1}{\text{DFE}}$ N is the number of observations (`NumObservations`), and `DFE` is the degrees of freedom for the error (residuals).
`LLR`	Loglikelihood ratio	$R^2_{\text{LLR}} = 1 - \frac{L}{L_0}$ L is the loglikelihood of the fitted model (`LogLikelihood`), and L_0 is the loglikelihood of a model that includes only a constant term. R^2_{LLR} is the McFadden pseudo R-squared value [1] for logistic regression models.

Field	Description	Equation
`Deviance`	Deviance R-squared	$R^2_{\text{Deviance}} = 1 - \frac{D}{D_0}$ D is the deviance of the fitted model (`Deviance`), and D_0 is the deviance of a model that includes only a constant term.
`AdjGeneralized`	Adjusted generalized R-squared	$R^2_{\text{AdjGeneralized}} = \frac{1 - \exp\left(\frac{2(L_0 - L)}{N}\right)}{1 - \exp\left(\frac{2L_0}{N}\right)}$ $R^2_{\text{AdjGeneralized}}$ is the Nagelkerke adjustment [2] to a formula proposed by Maddala [3], Cox and Snell [4], and Magee [5] for logistic regression models.

To obtain any of these values as a scalar, index into the property using dot notation. For example, to obtain the adjusted R-squared value in the model `mdl`, enter:

```
r2 = mdl.Rsquared.Adjusted
```

Data Types: `struct`

SSE — Sum of squared errors
numeric value

This property is read-only.

Sum of squared errors (residuals), specified as a numeric value.

Data Types: `single` | `double`

SSR — Regression sum of squares
numeric value

This property is read-only.

Regression sum of squares, specified as a numeric value. The regression sum of squares is equal to the sum of squared deviations of the fitted values from their mean.

Data Types: `single` | `double`

SST — Total sum of squares
numeric value

This property is read-only.

Total sum of squares, specified as a numeric value. The total sum of squares is equal to the sum of squared deviations of the response vector `y` from the `mean(y)`.

Data Types: `single` | `double`

Input Data

`Distribution` — Generalized distribution information

structure

This property is read-only.

Generalized distribution information, specified as a structure with the fields described in this table.

Field	Description
`Name`	Name of the distribution: `'normal'`, `'binomial'`, `'poisson'`, `'gamma'`, or `'inverse gaussian'`
`DevianceFunction`	Function that computes the components of the deviance as a function of the fitted parameter values and the response values
`VarianceFunction`	Function that computes the theoretical variance for the distribution as a function of the fitted parameter values. When `DispersionEstimated` is `true`, the software multiplies the variance function by `Dispersion` in the computation of the coefficient standard errors.

Data Types: `struct`

`Formula` — Model information

`LinearFormula` object

This property is read-only.

Model information, specified as a `LinearFormula` object.

Display the formula of the fitted model `mdl` using dot notation:

```
mdl.Formula
```

`Link` — Link function

structure

This property is read-only.

Link function, specified as a structure with the fields described in this table.

Field	Description
`Name`	Name of the link function, specified as a character vector. If you specify the link function using a function handle, then `Name` is `''`.
`Link`	Function f that defines the link function, specified as a function handle
`Derivative`	Derivative of f, specified as a function handle
`Inverse`	Inverse of f, specified as a function handle

The link function is a function f that links the distribution parameter μ to the fitted linear combination Xb of the predictors:

$$f(\mu) = Xb.$$

Data Types: `struct`

`NumObservations` — Number of observations
positive integer

This property is read-only.

Number of observations the fitting function used in fitting, specified as a positive integer. `NumObservations` is the number of observations supplied in the original table, dataset, or matrix, minus any excluded rows (set with the `'Exclude'` name-value pair argument) or rows with missing values.

Data Types: `double`

`NumPredictors` — Number of predictor variables
positive integer

This property is read-only.

Number of predictor variables used to fit the model, specified as a positive integer.

Data Types: `double`

`NumVariables` — Number of variables
positive integer

This property is read-only.

Number of variables in the input data, specified as a positive integer. `NumVariables` is the number of variables in the original table or dataset, or the total number of columns in the predictor matrix and response vector.

`NumVariables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `double`

`PredictorNames` — Names of predictors used to fit model
cell array of character vectors

This property is read-only.

Names of predictors used to fit the model, specified as a cell array of character vectors.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`VariableInfo` — Information about variables
table

This property is read-only.

Information about variables contained in `Variables`, specified as a table with one row for each variable and the columns described in this table.

Column	Description
`Class`	Variable class, specified as a cell array of character vectors, such as `'double'` and `'categorical'`
`Range`	Variable range, specified as a cell array of vectors • Continuous variable — Two-element vector `[min,max]`, the minimum and maximum values • Categorical variable — Vector of distinct variable values
`InModel`	Indicator of which variables are in the fitted model, specified as a logical vector. The value is `true` if the model includes the variable.
`IsCategorical`	Indicator of categorical variables, specified as a logical vector. The value is `true` if the variable is categorical.

`VariableInfo` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

`VariableNames` — Names of variables
cell array of character vectors

This property is read-only.

Names of variables, specified as a cell array of character vectors.

- If the fit is based on a table or dataset, this property provides the names of the variables in the table or dataset.
- If the fit is based on a predictor matrix and response vector, `VariableNames` contains the values specified by the `'VarNames'` name-value pair argument of the fitting method. The default value of `'VarNames'` is `{'x1','x2',...,'xn','y'}`.

`VariableNames` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `cell`

Object Functions

Predict Responses

feval	Predict responses of generalized linear regression model using one input for each predictor
predict	Predict responses of generalized linear regression model
random	Simulate responses with random noise for generalized linear regression model

Evaluate Generalized Linear Model

coefCI	Confidence intervals of coefficient estimates of generalized linear regression model
coefTest	Linear hypothesis test on generalized linear regression model coefficients

devianceTest	Analysis of deviance for generalized linear regression model
partialDependence	Compute partial dependence

Visualize Generalized Linear Model and Summary Statistics

plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotSlice	Plot of slices through fitted generalized linear regression surface

Gather Properties of Generalized Linear Model

gather	Gather properties of machine learning model from GPU

Examples

Compact Generalized Linear Regression Model

Fit a generalized linear regression model to data and reduce the size of a full, fitted model by discarding the sample data and some information related to the fitting process.

Load the `largedata4reg` data set, which contains 15,000 observations and 45 predictor variables.

```
load largedata4reg
```

Fit a generalized linear regression model to the data using the first 15 predictor variables.

```
mdl = fitglm(X(:,1:15),Y);
```

Compact the model.

```
compactMdl = compact(mdl);
```

The compact model discards the original sample data and some information related to the fitting process, so it uses less memory than the full model.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans = 1×2

       15517     4382500
```

The compact model consumes less memory than the full model.

More About

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to a saturated model.

Deviance of a model M_1 is twice the difference between the loglikelihood of the model M_1 and the saturated model M_s. A saturated model is a model with the maximum number of parameters that you can estimate.

For example, if you have n observations (y_i, $i = 1, 2, ..., n$) with potentially different values for $X_i^T\beta$, then you can define a saturated model with n parameters. Let $L(b,y)$ denote the maximum value of the likelihood function for a model with the parameters b. Then the deviance of the model M_1 is

$$-2(\log L(b_1, y) - \log L(b_S, y)),$$

where b_1 and b_s contain the estimated parameters for the model M_1 and the saturated model, respectively. The deviance has a chi-square distribution with $n – p$ degrees of freedom, where n is the number of parameters in the saturated model and p is the number of parameters in the model M_1.

Assume you have two different generalized linear regression models M_1 and M_2, and M_1 has a subset of the terms in M_2. You can assess the fit of the models by comparing the deviances D_1 and D_2 of the two models. The difference of the deviances is

$$\begin{aligned} D = D_2 - D_1 &= -2(\log L(b_2, y) - \log L(b_S, y)) + 2(\log L(b_1, y) - \log L(b_S, y)) \\ &= -2(\log L(b_2, y) - \log L(b_1, y)). \end{aligned}$$

Asymptotically, the difference D has a chi-square distribution with degrees of freedom v equal to the difference in the number of parameters estimated in M_1 and M_2. You can obtain the p-value for this test by using `1 – chi2cdf(D,v)`.

Typically, you examine D using a model M_2 with a constant term and no predictors. Therefore, D has a chi-square distribution with $p – 1$ degrees of freedom. If the dispersion is estimated, the difference divided by the estimated dispersion has an F distribution with $p – 1$ numerator degrees of freedom and $n – p$ denominator degrees of freedom.

References

[1] McFadden, Daniel. "Conditional logit analysis of qualitative choice behavior." in *Frontiers in Econometrics*, edited by P. Zarembka,105–42. New York: Academic Press, 1974.

[2] Nagelkerke, N. J. D. "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78, no. 3 (1991): 691–92.

[3] Maddala, Gangadharrao S. *Limited-Dependent and Qualitative Variables in Econometrics.* Econometric Society Monographs. New York, NY: Cambridge University Press, 1983.

[4] Cox, D. R., and E. J. Snell. *Analysis of Binary Data*. 2nd ed. Monographs on Statistics and Applied Probability 32. London; New York: Chapman and Hall, 1989.

[5] Magee, Lonnie. "R 2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests." *The American Statistician* 44, no. 3 (August 1990): 250–53.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `random` functions support code generation.
- When you fit a model by using `fitglm` or `stepwiseglm`, you cannot specify `Link`, `Derivative`, and `Inverse` fields of the `'Link'` name-value pair argument as anonymous functions. That is, you cannot generate code using a generalized linear model that was created using anonymous functions for links. Instead, define functions for link components.

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions fully support GPU arrays:
  - `feval`
  - `predict`
  - `random`
  - `partialDependence`
  - `plotPartialDependence`
- The following object functions support model objects fitted with GPU array input arguments:
  - `coefCI`
  - `coefTest`
  - `devianceTest`
  - `plotSlice`
  - `gather`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`GeneralizedLinearModel` | `compact` | `fitglm` | `stepwiseglm`

Topics
“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2016b

CompactRegressionEnsemble

Package: `classreg.learning.regr`

Compact regression ensemble class

Description

Compact version of a regression ensemble (of class `RegressionEnsemble`). The compact version does not include the data for training the regression ensemble. Therefore, you cannot perform some tasks with a compact regression ensemble, such as cross validation. Use a compact regression ensemble for making predictions (regressions) of new data.

Construction

`ens = compact(fullEns)` constructs a compact decision ensemble from a full decision ensemble.

Input Arguments

`fullEns`

A regression ensemble created by `fitrensemble`.

Properties

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CombineWeights`

A character vector describing how the ensemble combines learner predictions.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`NumTrained`

Number of trained learners in the ensemble, a positive scalar.

`PredictorNames`

A cell array of names for the predictor variables, in the order in which they appear in `X`.

ResponseName

A character vector with the name of the response variable `Y`.

ResponseTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`.

Add or change a `ResponseTransform` function using dot notation:

```
ens.ResponseTransform = @function
```

Trained

The trained learners, a cell array of compact regression models.

TrainedWeights

A numeric vector of weights the ensemble assigns to its learners. The ensemble computes predicted response by aggregating weighted predictions from its learners.

Object Functions

lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using ensemble of regression models
predictorImportance	Estimates of predictor importance for regression ensemble
removeLearners	Remove members of compact regression ensemble
shapley	Shapley values

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Reduce Size of Regression Ensemble

Create a compact regression ensemble for efficiently making predictions on new data.

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

Train a boosted ensemble of 100 regression trees using the `LSBoost`. Specify that `Cylinders` is a categorical variable.

```
Mdl = fitrensemble(X,Y,'PredictorNames',{'W','C'},...
    'CategoricalPredictors',2)

Mdl =
  RegressionEnsemble
           PredictorNames: {'W'  'C'}
             ResponseName: 'Y'
    CategoricalPredictors: 2
        ResponseTransform: 'none'
          NumObservations: 94
                NumTrained: 100
                    Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

`Mdl` is a `RegressionEnsemble` model object that contains the training data, among other things.

Create a compact version of `Mdl`.

```
CMdl = compact(Mdl)

CMdl =
  CompactRegressionEnsemble
           PredictorNames: {'W'  'C'}
             ResponseName: 'Y'
    CategoricalPredictors: 2
        ResponseTransform: 'none'
                NumTrained: 100


  Properties, Methods
```

`CMdl` is a `CompactRegressionEnsemble` model object. `CMdl` is almost the same as `Mdl`. One exception is that `CMdl` does not store the training data.

Compare the amounts of space consumed by `Mdl` and `CMdl`.

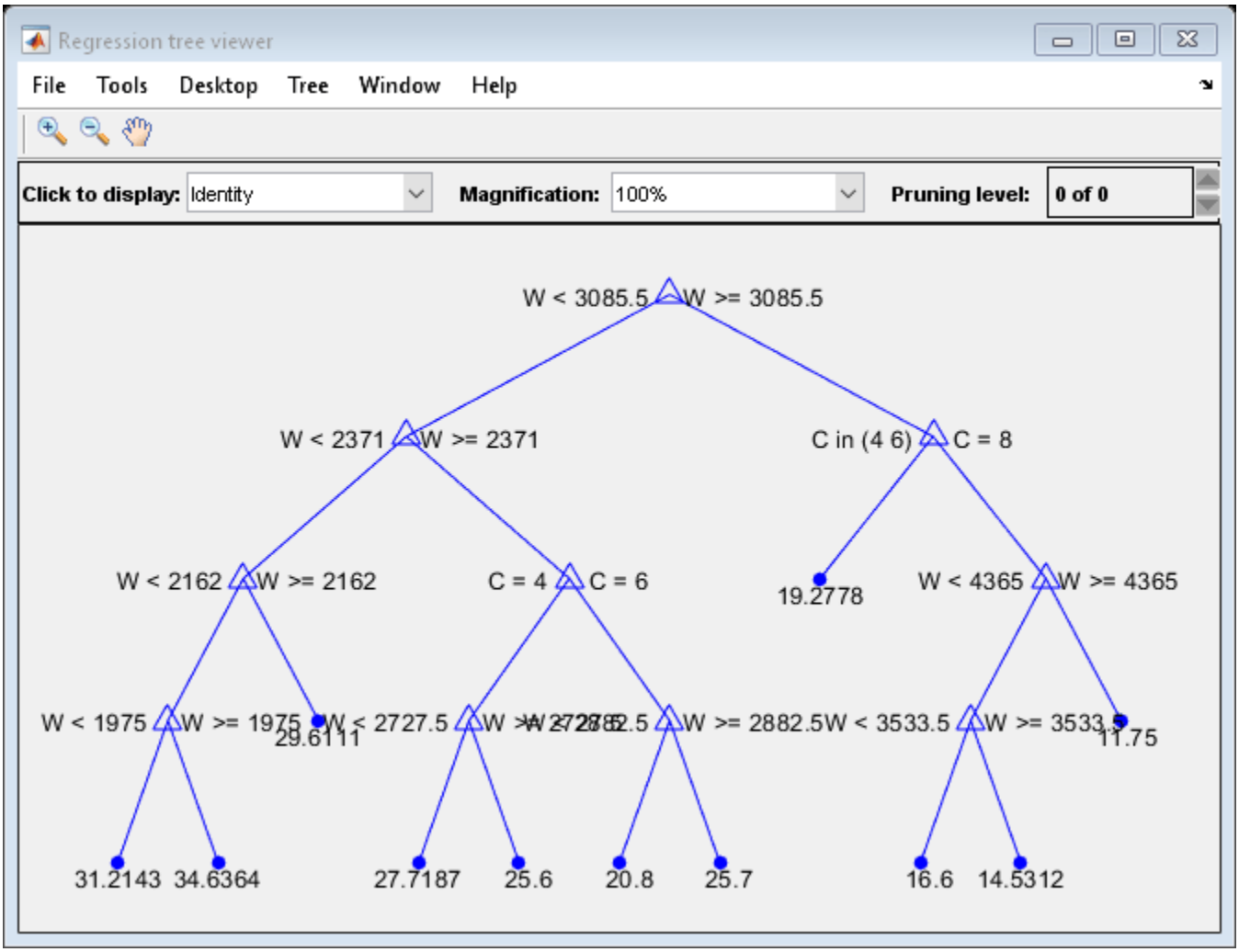
```
mdlInfo = whos('Mdl');
cMdlInfo = whos('CMdl');
[mdlInfo.bytes cMdlInfo.bytes]

ans = 1×2

      524616      499348
```

`Mdl` consumes more space than `CMdl`.

`CMdl.Trained` stores the trained regression trees (`CompactRegresionTree` model objects) that compose `Mdl`.

Display a graph of the first tree in the compact ensemble.

```
view(CMdl.Trained{1},'Mode','graph');
```



By default, `fitrensemble` grows shallow trees for boosted ensembles of trees.

Predict the fuel economy of a typical car using the compact ensemble.

```
typicalX = [mean(X(:,1)) mode(X(:,2))];
predMeanX = predict(CMdl,typicalX)
```

```
predMeanX = 26.2520
```

Tip

For a compact ensemble of regression trees, the `Trained` property of `ens` stores a cell vector of `ens.NumTrained` `CompactRegressionTree` model objects. For a textual or graphical display of tree *`t`* in the cell vector, enter

```
view(ens.Trained{t})
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the RegressionEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitrensemble`, code generation limitations for regression trees also apply to ensembles of regression trees. For more details, see "Code Generation" on page 33-868 of the `CompactRegressionTree` class.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`RegressionEnsemble` | `compact` | `fitrensemble` | `predict` | `templateTree` | `view`

Introduced in R2011a

CompactRegressionGAM

Compact generalized additive model (GAM) for regression

Description

`CompactRegressionGAM` is a compact version of a `RegressionGAM` model object (GAM for regression). The compact model does not include the data used for training the model. Therefore, you cannot perform some tasks, such as cross-validation, using the compact model. Use a compact model for tasks such as predicting the responses of new data.

Creation

Create a `CompactRegressionGAM` object from a full `RegressionGAM` model object by using `compact`.

Properties

GAM Properties

`Interactions` — Interaction term indices
two-column matrix of positive integers | `[]`

This property is read-only.

Interaction term indices, specified as a `t`-by-2 matrix of positive integers, where `t` is the number of interaction terms in the model. Each row of the matrix represents one interaction term and contains the column indexes of the predictor data `X` for the interaction term. If the model does not include an interaction term, then this property is empty (`[]`).

The software adds interaction terms to the model in the order of importance based on the *p*-values. Use this property to check the order of the interaction terms added to the model.

Data Types: `double`

`Intercept` — Intercept term of model
numeric scalar

This property is read-only.

Intercept (constant) term of the model, which is the sum of the intercept terms in the predictor trees and interaction trees, specified as a numeric scalar.

Data Types: `single` | `double`

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

`ExpandedPredictorNames` is the same as `PredictorNames` for a generalized additive model.

Data Types: `cell`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

Object Functions

Interpret Prediction

lime	Local interpretable model-agnostic explanations (LIME)
partialDependence	Compute partial dependence
plotLocalEffects	Plot local effects of terms in generalized additive model (GAM)

plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
shapley	Shapley values

Assess Predictive Performance on New Observations

predict	Predict responses using generalized additive model (GAM)
loss	Regression loss for generalized additive model (GAM)

Examples

Reduce Size of Generalized Additive Model

Reduce the size of a full generalized additive model (GAM) for regression by removing the training data. Full models hold the training data. You can use a compact model to improve memory efficiency.

Load the `carbig` data set.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a GAM using `X` and `Y`.

```
Mdl = fitrgam(X,Y)
```

```
Mdl =
  RegressionGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                Intercept: 26.9442
          NumObservations: 398


  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object.

Reduce the size of the model.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactRegressionGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                Intercept: 26.9442
```

```
  Properties, Methods
```

`CMdl` is a `CompactRegressionGAM` model object.

Display the amount of memory used by each regression model.

```
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class                                          Attributes

  CMdl      1x1              578154  classreg.learning.regr.CompactRegressionGAM
  Mdl       1x1              611947  RegressionGAM
```

The full model (`Mdl`) is larger than the compact model (`CMdl`).

To efficiently predict responses for new observations, you can remove `Mdl` from the MATLAB® Workspace, and then pass `CMdl` and new predictor values to `predict`.

See Also

`RegressionGAM` | `compact`

Topics

“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

CompactRegressionGP

Package: `classreg.learning.regr`

Compact Gaussian process regression model class

Description

`CompactRegressionGP` is a compact Gaussian process regression (GPR) model. The compact model consumes less memory than a full model, because it does not include the data used for training the GPR model.

Because the compact model does not include the training data, you cannot perform some tasks, such as cross-validation, using the compact model. However, you can use the compact model for making predictions or calculate regression loss for new data (use `predict` and `loss`).

Construction

`compactMdl = compact(gprMdl)` returns a compact GPR model, `compactMdl`, from a full, trained GPR model, `gprMdl`. For more information, see `compact`.

Input Arguments

`gprMdl` — Full, trained Gaussian process regression model

`RegressionGP` model

Full, trained Gaussian process regression model, specified as a `RegressionGP` model, returned by `fitrgp`.

Properties

Fitting

`FitMethod` — Method used to estimate the parameters

`'none'` | `'exact'` | `'sd'` | `'sr'` | `'fic'`

Method used to estimate the basis function coefficients, β; noise standard deviation, σ; and kernel parameters, θ, of the GPR model, stored as a character vector. It can be one of the following.

Fit Method	Description
`'none'`	No estimation. `fitrgp` uses the initial parameter values as the parameter values.
`'exact'`	Exact Gaussian process regression.
`'sd'`	Subset of data points approximation.
`'sr'`	Subset of regressors approximation.
`'fic'`	Fully independent conditional approximation.

`BasisFunction` — Explicit basis function

`'none'` | `'constant'` | `'linear'` | `'pureQuadratic'` | function handle

Explicit basis function used in the GPR model, stored as a character vector or a function handle. It can be one of the following. If n is the number of observations, the basis function adds the term $\boldsymbol{H}*\boldsymbol{\beta}$ to the model, where $\boldsymbol{H}$ is the basis matrix and $\boldsymbol{\beta}$ is a p-by-1 vector of basis coefficients.

Explicit Basis	Basis Matrix
`'none'`	Empty matrix.
`'constant'`	$H = 1$ (n-by-1 vector of 1s, where n is the number of observations)
`'linear'`	$H = [1, X]$
`'pureQuadratic'`	$H = [1, X, X_2]$, where $X_2 = \begin{bmatrix} x_{11}^2 & x_{12}^2 & \cdots & x_{1d}^2 \\ x_{21}^2 & x_{22}^2 & \cdots & x_{2d}^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 & x_{n2}^2 & \cdots & x_{nd}^2 \end{bmatrix}$.
Function handle	Function handle, `hfcn`, that `fitrgp` calls as: $H = hfcn(X)$, where X is an n-by-d matrix of predictors and H is an n-by-p matrix of basis functions.

Data Types: `char` | `function_handle`

`CategoricalPredictors` — Indices of categorical predictors

vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`Beta` — Estimated coefficients

vector

Estimated coefficients for the explicit basis functions, stored as a vector. You can define the explicit basis function by using the `BasisFunction` name-value pair argument in `fitrgp`.

Data Types: `double`

`Sigma` — Estimated noise standard deviation

scalar value

Estimated noise standard deviation of the GPR model, stored as a scalar value.

Data Types: `double`

`ModelParameters` — Parameters used for training

`GPParams` object

Parameters used for training the GPR model, stored as a `GPParams` object.

Kernel Function

`KernelFunction` — Form of the covariance function

`'squaredExponential'` | `'matern32'` | `'matern52'` | `'ardsquaredexponential'` | `'ardmatern32'` | `'ardmatern52'` | function handle

Form of the covariance function used in the GPR model, stored as a character vector containing the name of the built-in covariance function or a function handle. It can be one of the following.

Function	Description
`'squaredexponential'`	Squared exponential kernel.
`'matern32'`	Matern kernel with parameter 3/2.
`'matern52'`	Matern kernel with parameter 5/2.
`'ardsquaredexponential'`	Squared exponential kernel with a separate length scale per predictor.
`'ardmatern32'`	Matern kernel with parameter 3/2 and a separate length scale per predictor.
`'ardmatern52'`	Matern kernel with parameter 5/2 and a separate length scale per predictor.
Function handle	A function handle that `fitrgp` can call like this: `Kmn = kfcn(Xm,Xn,theta)` where `Xm` is an *m*-by-*d* matrix, `Xn` is an *n*-by-*d* matrix and `Kmn` is an *m*-by-*n* matrix of kernel products such that `Kmn`(*i*,*j*) is the kernel product between `Xm`(*i*,:) and `Xn`(*j*,:). `theta` is the *r*-by-1 unconstrained parameter vector for `kfcn`.

Data Types: `char` | `function_handle`

`KernelInformation` — Information about the parameters of the kernel function

structure

Information about the parameters of the kernel function used in the GPR model, stored as a structure with the following fields.

Field Name	Description
`Name`	Name of the kernel function
`KernelParameters`	Vector of the estimated kernel parameters
`KernelParameterNames`	Names associated with the elements of `KernelParameters`.

Data Types: `struct`

Prediction

`PredictMethod` — Method used to make predictions

`'exact'` | `'bcd'` | `'sd'` | `'sr'` | `'fic'`

Method that `predict` uses to make predictions from the GPR model, stored as a character vector. It can be one of the following.

`PredictMethod`	Description
`'exact'`	Exact Gaussian process regression
`'bcd'`	Block Coordinate Descent
`'sd'`	Subset of Data points approximation
`'sr'`	Subset of Regressors approximation
`'fic'`	Fully Independent Conditional approximation

`Alpha` — Weights

numeric vector

Weights used to make predictions from the trained GPR model, stored as a numeric vector. `predict` computes the predictions for a new predictor matrix `Xnew` by using the product

$$K(X_{new}, A) * \alpha \ .$$

$K(X_{new}, A)$ is the matrix of kernel products between X_{new} and active set vector $\boldsymbol{A}$ and α is a vector of weights.

Data Types: `double`

`ResponseTransform` — Transformation applied to predicted response

`'none'` (default)

Transformation applied to the predicted response, stored as a character vector describing how the response values predicted by the model are transformed. In `RegressionGP`, `ResponseTransform` is `'none'` by default, and `RegressionGP` does not use `ResponseTransform` when making predictions.

Active Set Selection

`ActiveSetVectors` — Subset of training data

matrix

Subset of training data used to make predictions from the GPR model, stored as a matrix.

`predict` computes the predictions for a new predictor matrix `Xnew` by using the product

$$K(X_{new}, A) * \alpha \ .$$

$K(X_{new}, A)$ is the matrix of kernel products between X_{new} and active set vector $\boldsymbol{A}$ and α is a vector of weights.

`ActiveSetVectors` is equal to the training data `X` for exact GPR fitting and a subset of the training data `X` for sparse GPR methods. When there are categorical predictors in the model, `ActiveSetVectors` contains dummy variables for the corresponding predictors.

Data Types: `double`

`ActiveSetMethod` — Method used to select the active set
`'sgma'` | `'entropy'` | `'likelihood'` | `'random'`

Method used to select the active set for sparse methods (`'sd'`,`'sr'`, or `'fic'`), stored as a character vector. It can be one of the following.

`ActiveSetMethod`	Description
`'sgma'`	Sparse greedy matrix approximation
`'entropy'`	Differential entropy-based selection
`'likelihood'`	Subset of regressors log likelihood-based selection
`'random'`	Random selection

The selected active set is used in parameter estimation or prediction, depending on the choice of `FitMethod` and `PredictMethod` in `fitrgp`.

`ActiveSetSize` — Size of the active set
integer value

Size of the active set for sparse methods (`'sd'`,`'sr'`, or `'fic'`), stored as an integer value.

Data Types: `double`

Object Functions

lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error for Gaussian process regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict response of Gaussian process regression model
shapley	Shapley values

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`RegressionGP` | `compact` | `fitrgp`

Topics
Class Attributes
Property Attributes

Introduced in R2015b

CompactRegressionNeuralNetwork

Compact neural network model for regression

Description

`CompactRegresionNeuralNetwork` is a compact version of a `RegressionNeuralNetwork` model object. The compact model does not include the data used for training the regression model. Therefore, you cannot perform some tasks, such as cross-validation, using the compact model. Use a compact model for tasks such as predicting the response values of new data.

Creation

Create a `CompactRegressionNeuralNetwork` object from a full `RegressionNeuralNetwork` model object by using `compact`.

Properties

Neural Network Properties

`LayerSizes` — Sizes of fully connected layers
positive integer vector

This property is read-only.

Sizes of the fully connected layers in the neural network model, returned as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer. This layer always has one output.

Data Types: `single` | `double`

`LayerWeights` — Learned layer weights
cell array

This property is read-only.

Learned layer weights for fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer weights for the *i*th fully connected layer. For example, `Mdl.LayerWeights{1}` returns the weights for the first fully connected layer of the model `Mdl`.

`LayerWeights` includes the weights for the final fully connected layer.

Data Types: `cell`

`LayerBiases` — Learned layer biases
cell array

This property is read-only.

Learned layer biases for fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer biases for the *i*th fully connected layer. For example, `Mdl.LayerBiases{1}` returns the biases for the first fully connected layer of the model `Mdl`.

`LayerBiases` includes the biases for the final fully connected layer.

Data Types: `cell`

`Activations` — Activation functions for fully connected layers

`'relu'` | `'tanh'` | `'sigmoid'` | `'none'` | cell array of character vectors

This property is read-only.

Activation functions for the fully connected layers of the neural network model, returned as a character vector or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If `Activations` contains only one activation function, then it is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer, which does not have an activation function (`OutputLayerActivation`).
- If `Activations` is an array of activation functions, then the *i*th element is the activation function for the *i*th layer of the neural network model.

Data Types: `char` | `cell`

`OutputLayerActivation` — Activation function for final fully connected layer

`'none'`

This property is read-only.

Activation function for final fully connected layer, returned as `'none'`.

Data Properties

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, returned as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, returned as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, returned as a cell array of character vectors. If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, returned as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'`

This property is read-only.

Response transformation function, returned as `'none'`. The software does not transform the raw response values.

Object Functions

loss	Loss for regression neural network
partialDependence	Compute partial dependence

plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using regression neural network

Examples

Reduce Size of Regression Neural Network Model

Reduce the size of a full regression neural network model by removing the training data from the model. You can use a compact model to improve memory efficiency.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Systolic` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Age,Diastolic,Gender,Height,Smoker,Weight,Systolic);
```

Train a regression neural network model using the data. Specify the `Systolic` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitrnet(tbl,"Systolic","Standardize",true)
```

```
Mdl =
  RegressionNeuralNetwork
           PredictorNames: {'Age'  'Diastolic'  'Gender'  'Height'  'Smoker'  'Weight'}
             ResponseName: 'Systolic'
    CategoricalPredictors: [3 5]
        ResponseTransform: 'none'
          NumObservations: 100
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'linear'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [1000×7 table]


  Properties, Methods
```

`Mdl` is a full `RegressionNeuralNetwork` model object.

Reduce the size of the model by using `compact`.

```
compactMdl = compact(Mdl)
```

```
compactMdl =
  CompactRegressionNeuralNetwork
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'linear'


  Properties, Methods
```

`compactMdl` is a `CompactRegressionNeuralNetwork` model object. `compactMdl` contains fewer properties than the full model `Mdl`.

Display the amount of memory used by each neural network model.

```
whos("Mdl","compactMdl")
```

```
  Name            Size            Bytes  Class

  Mdl             1x1             72818  RegressionNeuralNetwork
  compactMdl      1x1              5995  classreg.learning.regr.CompactRegressionNeuralNetwork
```

The full model is larger than the compact model.

See Also

`RegressionNeuralNetwork` | `RegressionPartitionedModel` | `compact` | `fitrnet` | `loss` | `predict`

Topics

“Assess Regression Neural Network Performance” on page 18-184

Introduced in R2021a

CompactRegressionSVM

Package: `classreg.learning.regr`

Compact support vector machine regression model

Description

`CompactRegressionSVM` is a compact support vector machine (SVM) regression model. It consumes less memory than a full, trained support vector machine model (`RegressionSVM` model) because it does not store the data used to train the model.

Because the compact model does not store the training data, you cannot use it to perform certain tasks, such as cross validation. However, you can use a compact SVM regression model to predict responses using new input data.

Construction

`compactMdl = compact(mdl)` returns a compact SVM regression model `compactMdl` from a full, trained SVM regression model, `mdl`. For more information, see `compact`.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Properties

`Alpha` — Dual problem coefficients

vector of numeric values

Dual problem coefficients, specified as a vector of numeric values. `Alpha` contains *m* elements, where *m* is the number of support vectors in the trained SVM regression model. The dual problem introduces two Lagrange multipliers for each support vector. The values of `Alpha` are the differences between the two estimated Lagrange multipliers for the support vectors. For more details, see “Understanding Support Vector Machine Regression” on page 25-2.

If you specified to remove duplicates using `RemoveDuplicates`, then, for a particular set of duplicate observations that are support vectors, `Alpha` contains one coefficient corresponding to the entire set. That is, MATLAB attributes a nonzero coefficient to one observation from the set of duplicates and a coefficient of `0` to all other duplicate observations in the set.

Data Types: `single` | `double`

`Beta` — Primal linear problem coefficients

vector of numeric values | `'[]'`

Primal linear problem coefficients, stored as a numeric vector of length *p*, where *p* is the number of predictors in the SVM regression model.

The values in `Beta` are the linear coefficients for the primal optimization problem.

If the model is obtained using a kernel function other than `'linear'`, this property is empty (`'[]'`).

The `predict` method computes predicted response values for the model as `YFIT = (X/S)×Beta + Bias`, where`S` is the value of the kernel scale stored in the `KernelParameters.Scale` property.

Data Types: `double`

`Bias` — Bias term
scalar value

Bias term in the SVM regression model, stored as a scalar value.

Data Types: `double`

`CategoricalPredictors` — Indices of categorical predictors
vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`KernelParameters` — Kernel function parameters
structure

Kernel function parameters, stored as a structure with the following fields.

Field	Description
`Function`	Kernel function name (a character vector).
`Scale`	Numeric scale factor used to divide predictor values.

You can specify values for `KernelParameters.Function` and `KernelParameters.Scale` by using the `KernelFunction` and `KernelScale` name-value pair arguments in `fitrsvm`, respectively.

Data Types: `struct`

`Mu` — Predictor means
vector of numeric values | `'[]'`

Predictor means, stored as a vector of numeric values.

If the training data is standardized, then `Mu` is a numeric vector of length *p*, where *p* is the number of predictors used to train the model. In this case, the `predict` method centers predictor matrix `X` by subtracting the corresponding element of `Mu` from each column.

If the training data is not standardized, then `Mu` is empty (`'[]'`).

Data Types: `single` | `double`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names, stored as a cell array of character vectors containing the name of each predictor in the order in which they appear in `X`. `PredictorNames` has a length equal to the number of columns in `X`.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

Response variable name, stored as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

`Sigma` — Predictor standard deviations
vector of numeric values | `'[]'`

Predictor standard deviations, stored as a vector of numeric values.

If the training data is standardized, then `Sigma` is a numeric vector of length *p*, where *p* is the number of predictors used to train the model. In this case, the `predict` method scales the predictor matrix `X` by dividing each column by the corresponding element of `Sigma`, after centering each element using `Mu`.

If the training data is not standardized, then `Sigma` is empty (`'[]'`).

Data Types: `single` | `double`

`SupportVectors` — Support vectors
matrix of numeric values

Support vectors, stored as an *m*-by-*p* matrix of numeric values. *m* is the number of support vectors (`sum(Mdl.IsSupportVector)`), and *p* is the number of predictors in `X`.

If you specified to remove duplicates using `RemoveDuplicates`, then for a given set of duplicate observations that are support vectors, `SupportVectors` contains one unique support vector.

Data Types: `single` | `double`

Object Functions

discardSupportVectors	Discard support vectors
incrementalLearner	Convert support vector machine (SVM) regression model to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error for support vector machine regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using support vector machine regression model
shapley	Shapley values
update	Update model parameters for code generation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Compact an SVM Regression Model

This example shows how to reduce the size of a full, trained SVM regression model by discarding the training data and some information related to the training process.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model using a Gaussian kernel function and an automatic kernel scale. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale','auto','Standardize',true)
```

```
mdl =

  RegressionSVM
           PredictorNames: {1x8 cell}
             ResponseName: 'Var9'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635x1 double]
```

```
                    Bias: 10.8144
          KernelParameters: [1x1 struct]
                      Mu: [1x10 double]
                   Sigma: [1x10 double]
           NumObservations: 4177
            BoxConstraints: [4177x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [4177x1 logical]
                    Solver: 'SMO'


  Properties, Methods
```

Compact the model.

```
compactMdl = compact(mdl)
```

```
compactMdl =

  classreg.learning.regr.CompactRegressionSVM
            PredictorNames: {1x8 cell}
              ResponseName: 'Var9'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635x1 double]
                     Bias: 10.8144
          KernelParameters: [1x1 struct]
                      Mu: [1x10 double]
                   Sigma: [1x10 double]
            SupportVectors: [3635x10 double]


  Properties, Methods
```

The compacted model discards the training data and some information related to the training process.

Compare the size of the full model `mdl` and the compact model `compactMdl`.

```
vars = whos('compactMdl','mdl');
[vars(1).bytes,vars(2).bytes]
```

```
ans =

      323793      775968
```

The compacted model consumes about half the memory of the full model.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of an SVM regression model into Simulink, you can use the RegressionSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an SVM regression model by using `fitrsvm`, the following restrictions apply.
  - The value of the `'ResponseTransform'` name-value pair argument must be `'none'` (default).
  - For fixed-point code generation, the value of the `'KernelFunction'` name-value pair argument must be `'gaussian'`, `'linear'`, or `'polynomial'`.
  - Fixed-point code generation and code generation with a coder configurer do not support categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`RegressionSVM` | `compact` | `fitrsvm` | `update`

Introduced in R2015b

CompactRegressionTree

Package: `classreg.learning.regr`

Compact regression tree

Description

Compact version of a regression tree (of class `RegressionTree`). The compact version does not include the data for training the regression tree. Therefore, you cannot perform some tasks with a compact regression tree, such as cross validation. Use a compact regression tree for making predictions (regressions) of new data.

Construction

`ctree = compact(tree)` constructs a compact decision tree from a full decision tree.

Input Arguments

`tree` — Full, trained regression tree
`RegressionTree` object

Full, trained regression tree, specified as a `RegressionTree` object constructed by `fitrtree`.

Properties

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CategoricalSplits`

An *n*-by-2 cell array, where *n* is the number of categorical splits in `tree`. Each row in `CategoricalSplits` gives left and right values for a categorical split. For each branch node with categorical split `j` based on a categorical predictor variable `z`, the left child is chosen if `z` is in `CategoricalSplits(j,1)` and the right child is chosen if `z` is in `CategoricalSplits(j,2)`. The splits are in the same order as nodes of the tree. Nodes for these splits can be found by running `cuttype` and selecting `'categorical'` cuts from top to bottom.

`Children`

An *n*-by-2 array containing the numbers of the child nodes for each node in `tree`, where *n* is the number of nodes. Leaf nodes have child node `0`.

`CutCategories`

An *n*-by-2 cell array of the categories used at branches in `tree`, where *n* is the number of nodes. For each branch node `i` based on a categorical predictor variable `x`, the left child is chosen if `x` is among the categories listed in `CutCategories{i,1}`, and the right child is chosen if `x` is among those

listed in `CutCategories{i,2}`. Both columns of `CutCategories` are empty for branch nodes based on continuous predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPoint`

An *n*-element vector of the values used as cut points in `tree`, where *n* is the number of nodes. For each branch node `i` based on a continuous predictor variable `x`, the left child is chosen if `CutPoint<v(i)` and the right child is chosen if `x>=CutPoint(i)`. `CutPoint` is `NaN` for branch nodes based on categorical predictors and for leaf nodes.

`CutType`

An *n*-element cell array indicating the type of cut at each node in `tree`, where *n* is the number of nodes. For each node `i`, `CutType{i}` is:

- `'continuous'` — If the cut is defined in the form `x < v` for a variable `x` and cut point `v`.
- `'categorical'` — If the cut is defined by whether a variable `x` takes a value in a set of categories.
- `''` — If `i` is a leaf node.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictor`

An *n*-element cell array of the names of the variables used for branching in each node in `tree`, where *n* is the number of nodes. These variables are sometimes known as *cut variables*. For leaf nodes, `CutPredictor` contains an empty character vector.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPredictorIndex`

An *n*-element array of numeric indices for the variables used for branching in each node in `tree`, where *n* is the number of nodes. For more information, see `CutPredictor`.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`IsBranchNode`

An *n*-element logical vector `ib` that is `true` for each branch node and `false` for each leaf node of `tree`.

`NodeError`

An *n*-element vector `e` of the errors of the nodes in `tree`, where *n* is the number of nodes. `e(i)` is the misclassification probability for node `i`.

`NodeMean`

An *n*-element numeric array with mean values in each node of `tree`, where *n* is the number of nodes in the tree. Every element in `NodeMean` is the average of the true `Y` values over all observations in the node.

`NodeProbability`

An *n*-element vector `p` of the probabilities of the nodes in `tree`, where *n* is the number of nodes. The probability of a node is computed as the proportion of observations from the original data that satisfy the conditions for the node. This proportion is adjusted for any prior probabilities assigned to each class.

`NodeRisk`

An *n*-element vector of the risk of the nodes in the tree, where *n* is the number of nodes. The risk for each node is the node error weighted by the node probability.

`NodeSize`

An *n*-element vector `sizes` of the sizes of the nodes in `tree`, where *n* is the number of nodes. The size of a node is defined as the number of observations from the data used to create the tree that satisfy the conditions for the node.

`NumNodes`

The number of nodes `n` in `tree`.

`Parent`

An *n*-element vector `p` containing the number of the parent node for each node in `tree`, where *n* is the number of nodes. The parent of the root node is `0`.

`PredictorNames`

A cell array of names for the predictor variables, in the order in which they appear in `X`.

`PruneAlpha`

Numeric vector with one element per pruning level. If the pruning level ranges from 0 to *M*, then `PruneAlpha` has *M* + 1 elements sorted in ascending order. `PruneAlpha(1)` is for pruning level 0 (no pruning), `PruneAlpha(2)` is for pruning level 1, and so on.

`PruneList`

An *n*-element numeric vector with the pruning levels in each node of `tree`, where *n* is the number of nodes. The pruning levels range from 0 (no pruning) to *M*, where *M* is the distance between the deepest leaf and the root node.

ResponseName

Name of the response variable `Y`, a character vector.

ResponseTransform

Function handle for transforming the raw response values (mean squared error). The function handle must accept a matrix of response values and return a matrix of the same size. The default `'none'` means `@(x)x`, or no transformation.

Add or change a `ResponseTransform` function using dot notation:

```
ctree.ResponseTransform = @function
```

SurrogateCutCategories

An *n*-element cell array of the categories used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutCategories{k}` is a cell array. The length of `SurrogateCutCategories{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutCategories{k}` is either an empty character vector for a continuous surrogate predictor, or is a two-element cell array with categories for a categorical surrogate predictor. The first element of this two-element cell array lists categories assigned to the left child by this surrogate split, and the second element of this two-element cell array lists categories assigned to the right child by this surrogate split. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutCategories` contains an empty cell.

SurrogateCutFlip

An *n*-element cell array of the numeric cut assignments used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutFlip{k}` is a numeric vector. The length of `SurrogateCutFlip{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutFlip{k}` is either zero for a categorical surrogate predictor, or a numeric cut assignment for a continuous surrogate predictor. The numeric cut assignment can be either –1 or +1. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and the cut assignment for this surrogate split is +1, or if $Z\geq C$ and the cut assignment for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and the cut assignment for this surrogate split is +1, or if $Z<C$ and the cut assignment for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutFlip` contains an empty array.

SurrogateCutPoint

An *n*-element cell array of the numeric values used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutPoint{k}` is a numeric vector. The length of `SurrogateCutPoint{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutPoint{k}` is either `NaN` for a categorical surrogate predictor, or a numeric cut for a continuous surrogate predictor. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is –1. Similarly, the right child is chosen if $Z\geq C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z<C$ and `SurrogateCutFlip` for this surrogate split is –1. The order of the surrogate split variables at each

node is matched to the order of variables returned by `SurrogateCutVar`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPoint` contains an empty cell.

`SurrogateCutType`

An *n*-element cell array indicating types of surrogate splits at each node in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutType{k}` is a cell array with the types of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutType` contains an empty cell. A surrogate split type can be either `'continuous'` if the cut is defined in the form `Z<V` for a variable `Z` and cut point `V` or `'categorical'` if the cut is defined by whether `Z` takes a value in a set of categories.

`SurrogateCutPredictor`

An *n*-element cell array of the names of the variables used for surrogate splits in each node in `tree`, where *n* is the number of nodes in `tree`. Every element of `SurrogateCutPredictor` is a cell array with the names of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutPredictor` contains an empty cell.

`SurrogatePredictorAssociation`

An *n*-element cell array of the predictive measures of association for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogatePredictorAssociation{k}` is a numeric vector. The length of `SurrogatePredictorAssociation{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogatePredictorAssociation{k}` gives the predictive measure of association between the optimal split and this surrogate split. The order of the surrogate split variables at each node is the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogatePredictorAssociation` contains an empty cell.

Object Functions

lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using regression tree
predictorImportance	Estimates of predictor importance for regression tree
shapley	Shapley values
surrogateAssociation	Mean predictive measure of association for surrogate splits in regression tree
update	Update model parameters for code generation
view	View regression tree

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Construct and Compact a Regression Tree

Load the sample data.

```
load carsmall
```

Construct a regression tree for the sample data.

```
tree = fitrtree([Weight, Cylinders],MPG,...
    'MinParentSize',20,...
    'PredictorNames',{'W','C'});
```

Make a compact version of the tree.

```
ctree = compact(tree);
```

Compare the size of the compact tree to that of the full tree.

```
t = whos('tree'); % t.bytes = size of tree in bytes
c = whos('ctree'); % c.bytes = size of ctree in bytes
[c.bytes t.bytes]
```

```
ans = 1×2

        4311        7558
```

The compact tree is smaller than the full tree.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of a regression tree model into Simulink, you can use the RegressionTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train a regression tree model by using `fitrtree`, the following restrictions apply.
  - The value of the `'ResponseTransform'` name-value pair argument must be `'none'` (default).
  - You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value pair argument must be `'off'`.
  - Fixed-point code generation and code generation with a coder configurer do not support categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`RegressionTree` | `compact` | `fitrtree`

Introduced in R2011a

CompactTreeBagger class

Compact ensemble of decision trees grown by bootstrap aggregation

Description

`CompactTreeBagger` class is a lightweight class that contains the trees grown using `TreeBagger`. `CompactTreeBagger` does not preserve any information about how `TreeBagger` grew the decision trees. It does not contain the input data used for growing trees, nor does it contain training parameters such as minimal leaf size or number of variables sampled for each decision split at random. You can only use `CompactTreeBagger` for predicting the response of the trained ensemble given new data `X`, and other related functions.

`CompactTreeBagger` lets you save the trained ensemble to disk, or use it in any other way, while discarding training data and various parameters of the training configuration irrelevant for predicting response of the fully grown ensemble. This reduces storage and memory requirements, especially for ensembles trained on large data sets.

Construction

CompactTreeBagger	Create `CompactTreeBagger` object

`CMdl = compact(Mdl)` creates a compact version of `Mdl`, a `TreeBagger` model object. You can predict regressions using `CMdl` exactly as you can using `Mdl`. However, since `CMdl` does not contain training data, you cannot perform some actions, such as make out-of-bag predictions using `oobPredict`.

Object Functions

combine	Combine two ensembles
error	Error (misclassification probability or MSE)
margin	Classification margin
mdsprox	Multidimensional scaling of proximity matrix
meanMargin	Mean classification margin
outlierMeasure	Outlier measure for data
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using ensemble of bagged decision trees
proximity	Proximity matrix for data
setDefaultYfit	Set default value for predict

Properties

`ClassNames`

The `ClassNames` property is a cell array containing the class names for the response variable `Y` supplied to `TreeBagger`. This property is empty for regression trees.

`DefaultYfit`

The `DefaultYfit` property controls what predicted value `CompactTreeBagger` returns when no prediction is possible, for example when the `predict` method needs to predict for an observation which has only false values in the matrix supplied through `'useifort'` argument.

For classification, you can set this property to either `''` or `'MostPopular'`. If you choose `'MostPopular'` (default), the property value becomes the name of the most probable class in the training data.

For regression, you can set this property to any numeric scalar. The default is the mean of the response for the training data.

`DeltaCriterionDecisionSplit`

The `DeltaCriterionDecisionSplit` property is a numeric array of size 1-by-`Nvars` of changes in the split criterion summed over splits on each variable, averaged across the entire ensemble of grown trees.

`Method`

The `Method` property is `'classification'` for classification ensembles and `'regression'` for regression ensembles.

`NumPredictorSplit`

The `NumPredictorSplit` property is a numeric array of size 1-by-*Nvars*, where every element gives a number of splits on this predictor summed over all trees.

`NumTrees`

The `NumTrees` property is a scalar equal to the number of decision trees in the ensemble.

`PredictorNames`

The `PredictorNames` property is a cell array containing the names of the predictor variables (features). These names are taken from the optional `'names'` parameter that supplied to `TreeBagger`. The default names are `'x1'`, `'x2'`, etc.

`SurrogateAssociation`

The `SurrogateAssociation` property is a matrix of size *Nvars*-by-*Nvars* with predictive measures of variable association, averaged across the entire ensemble of grown trees. If you grew the ensemble setting `'surrogate'` to `'on'`, this matrix for each tree is filled with predictive measures of association averaged over the surrogate splits. If you grew the ensemble setting `'surrogate'` to `'off'` (default), `SurrogateAssociation` is diagonal.

`Trees`

The `Trees` property is a cell array of size `NumTrees`-by-1 containing the trees in the ensemble.

Examples

Reduce Size of Bag of Trees

Create a compact bag of trees for efficiently making predictions on new data.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a bag of 100 classification trees using all measurements and the `AdaBoostM1` method.

```
Mdl = TreeBagger(100,X,Y,'Method','classification')
```

```
Mdl =
  TreeBagger
Ensemble with 100 bagged decision trees:
                    Training X:              [351x34]
                    Training Y:               [351x1]
                        Method:       classification
                NumPredictors:                   34
          NumPredictorsToSample:                    6
                     MinLeafSize:                    1
                  InBagFraction:                     1
          SampleWithReplacement:                     1
           ComputeOOBPrediction:                     0
 ComputeOOBPredictorImportance:                      0
                      Proximity:                     []
                    ClassNames:             'b'             'g'

  Properties, Methods
```

`Mdl` is a `TreeBagger` model object that contains the training data, among other things.

Create a compact version of `Mdl`.

```
CMdl = compact(Mdl)
```

```
CMdl =
  CompactTreeBagger
Ensemble with 100 bagged decision trees:
              Method:       classification
       NumPredictors:                   34
          ClassNames: 'b' 'g'

  Properties, Methods
```

`CMdl` is a `CompactTreeBagger` model object. `CMdl` is almost the same as `Mdl`. One exception is that it does not store the training data.

Compare the amounts of space consumed by `Mdl` and `CMdl`.

```
mdlInfo = whos('Mdl');
cMdlInfo = whos('CMdl');
[mdlInfo.bytes cMdlInfo.bytes]
```

```
ans = 1×2
```

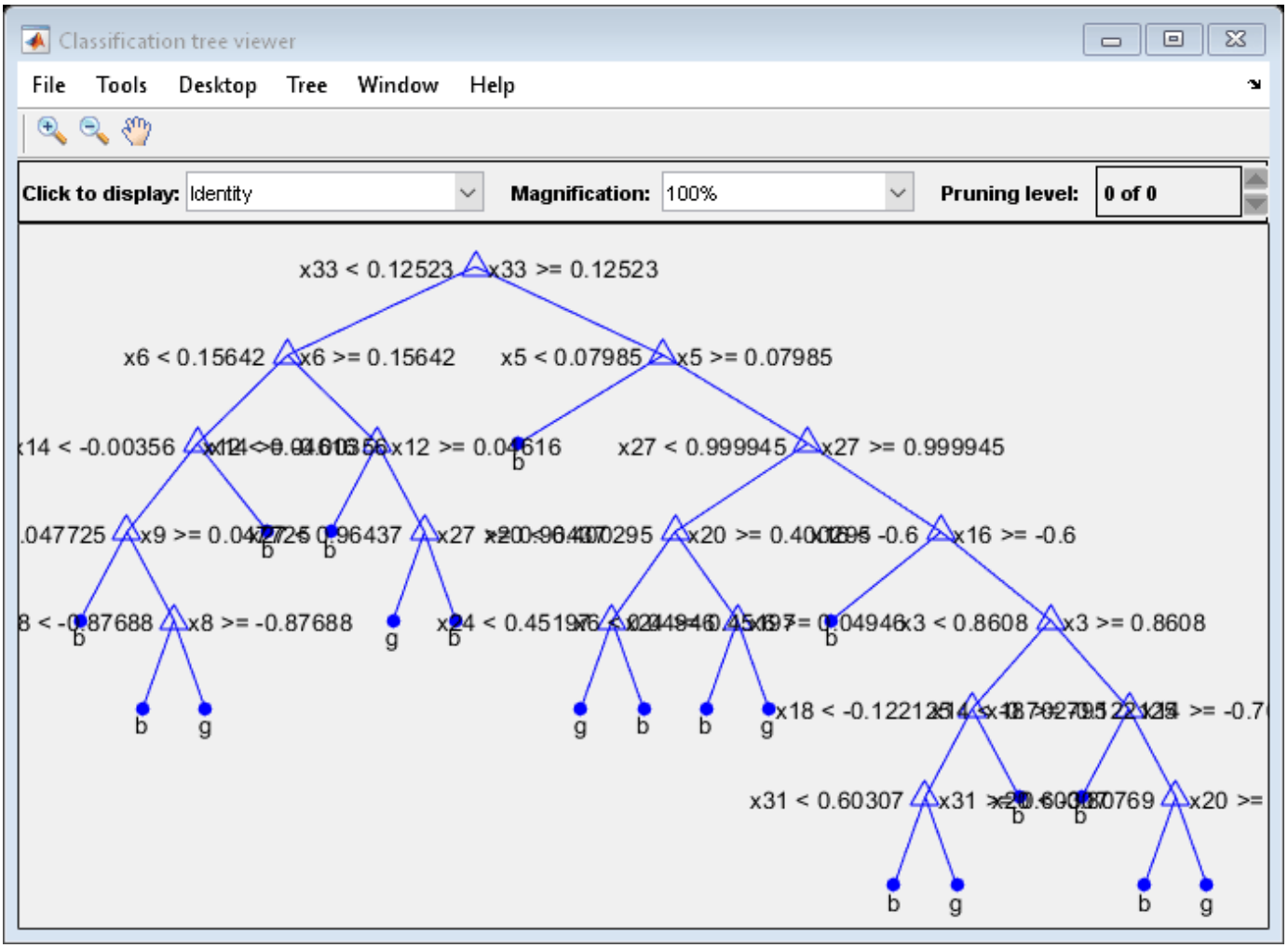
```
      1115742       976936
```

`Mdl` consumes more space than `CMdl`.

`CMdl.Trees` stores the trained classification trees (`CompactClassificationTree` model objects) that compose `Mdl`.

Display a graph of the first tree in the compact model.

```
view(CMdl.Trees{1},'Mode','graph');
```



By default, `TreeBagger` grows deep trees.

Predict the label of the mean of `X` using the compact ensemble.

```
predMeanX = predict(CMdl,mean(X))
```

```
predMeanX = 1x1 cell array
    {'g'}
```

Copy Semantics

Value. To learn how this affects your use of the class, see Comparing Handle and Value Classes in the MATLAB Object-Oriented Programming documentation.

Tip

The `Trees` property of `CMdl` stores a cell vector of `CMdl.NumTrees` `CompactClassificationTree` or `CompactRegressionTree` model objects. For a textual or graphical display of tree *`t`* in the cell vector, enter

```
view(CMdl.Trees{t})
```

See Also

`ClassificationTree` | `RegressionTree` | `TreeBagger` | `compact` | `error` | `predict` | `view` | `view`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
“Framework for Ensemble Learning” on page 18-31
“Decision Trees” on page 19-2
“Grouping Variables” on page 2-45

CompactTreeBagger

Class: `CompactTreeBagger`

Create `CompactTreeBagger` object

Description

When you use the `TreeBagger` constructor to grow trees, it creates a `CompactTreeBagger` object. You can obtain the compact object from the full `TreeBagger` object using the `TreeBagger/compact` method. You do not create an instance of `CompactTreeBagger` directly.

See Also

`TreeBagger`

Topics

“Grouping Variables” on page 2-45
“Framework for Ensemble Learning” on page 18-31

Introduced in R2009a

compare

Class: `GeneralizedLinearMixedModel`

Compare generalized linear mixed-effects models

Syntax

```
results = compare(glme,altglme)
results = compare(glme,altglme,Name,Value)
```

Description

`results = compare(glme,altglme)` returns the results of a likelihood ratio test on page 33-879 that compares the generalized linear mixed-effects models `glme` and `altglme`. To conduct a valid likelihood ratio test, both models must use the same response vector in the fit, and `glme` must be nested in `altglme`. Always input the smaller model first, and the larger model second.

`compare` tests the following null and alternate hypotheses:

- H_0: Observed response vector is generated by `glme`.
- H_1: Observed response vector is generated by model `altglme`.

`results = compare(glme,altglme,Name,Value)` returns the results of a likelihood ratio test using additional options specified by one or more `Name,Value` pair arguments. For example, you can check if the first input model, `glme`, is nested in the second input model, `altglme`.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

You can create a `GeneralizedLinearMixedModel` object by fitting a generalized linear mixed-effects model to your sample data using `fitglme`. To conduct a valid likelihood ratio test on two models that have response distributions other than normal, you must fit both models using the `'ApproximateLaplace'` or `'Laplace'` fit method. Models with response distributions other than normal that are fitted using `'MPL'` or `'REMPL'` cannot be compared using a likelihood ratio test.

`altglme` — Alternative generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Alternative generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. `altglme` be must fit to the same response vector as `glme`, but with different model specifications. `glme` must be nested in `altglme`, such that you can obtain `glme` from `altglme` by setting some of the model parameters of `altglme` to fixed values such as 0.

You can create a `GeneralizedLinearMixedModel` object by fitting a generalized linear mixed-effects model to your sample data using `fitglme`. To conduct a valid likelihood ratio test on two

models that have response distributions other than normal, you must fit both models using the `'ApproximateLaplace'` or `'Laplace'` fit method. Models with response distributions other than normal that are fitted using `'MPL'` or `'REMPL'` cannot be compared using a likelihood ratio test.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CheckNesting` — Indicator to check nesting between two models
`true` (default) | `false`

Indicator to check nesting on page 33-880 between two models, specified as the comma-separated pair consisting of `'CheckNesting'` and either `true` or `false`. If `'CheckNesting'` is `true`, then `compare` checks if the smaller model `glme` is nested in the larger model `altglme`. If the nesting requirements are not satisfied, then `compare` returns an error. If `'CheckNesting'` is `false`, then `compare` does not perform this check.

Example: `'CheckNesting',true`

Output Arguments

`results` — Results of likelihood ratio test
table

Results of the likelihood ratio test, returned as a table with two rows. The first row is for `glme`, and the second row is for `altglme`. The columns of `results` contain the following.

Column Name	Description
`Model`	Name of the model
`DF`	Degrees of freedom
`AIC`	Akaike information criterion for the model
`BIC`	Bayesian information criterion for the model
`LogLik`	Maximized log likelihood for the model
`LRStat`	Likelihood ratio test statistic for comparing `altglme` and `glme`
`deltaDF`	`DF` for `altglme` minus `DF` for `glme`
`pValue`	*p*-value for the likelihood ratio test

Examples

Compare Mixed-Effects Models

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to

decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a fixed-effects-only model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Specify the response distribution as Poisson, the link function as log, and the fit method as Laplace. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

```
FEglme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier','Distribution',
```

Fit a second model that uses the same fixed-effects predictors, response distribution, link function, and fit method. This time, include a random-effects intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Compare the two models using a theoretical likelihood ratio test. Specify `'CheckNesting'` as `true`, so `compare` returns a warning if the nesting requirements are not satisfied.

```
results = compare(FEglme,glme,'CheckNesting',true)
```

```
results =
    Theoretical Likelihood Ratio Test

    Model     DF    AIC       BIC       LogLik     LRStat    deltaDF
    FEglme    6     431.02    446.65    -209.51
    glme      7     416.35    434.58    -201.17    16.672    1


    pValue

    4.4435e-05
```

Since `compare` did not return an error, the nesting requirements are satisfied. The small *p*-value indicates that `compare` rejects the null hypothesis that the observed response vector is generated by the model `FEglme`, and instead accepts the alternate model `glme`. The smaller `AIC` and `BIC` values for `glme` also support the conclusion that `glme` provides a better fitting model for the response.

More About

Likelihood Ratio Test

A *likelihood ratio test* compares the specifications of two nested models by assessing the significance of restrictions to an extended model with unrestricted parameters. Under the null hypothesis H_0, the likelihood ratio test statistic has an approximate chi-squared reference distribution with degrees of freedom `deltaDF`.

When comparing two models, `compare` computes the *p*-value for the likelihood ratio test by comparing the observed likelihood ratio test statistic with this chi-squared reference distribution. A small *p*-value leads to a rejection of H_0 in favor of H_1, and acceptance of the alternate model `altglme`. On the other hand, a large *p*-value indicates that we cannot reject H_0, and reflects insufficient evidence to accept the model `altglme`.

The *p*-values obtained using the likelihood ratio test can be conservative when testing for the presence or absence of random-effects terms, and anti-conservative when testing for the presence or absence of fixed-effects terms. Instead, use the `fixedEffects` or `coefTest` methods to test for fixed effects.

To conduct a valid likelihood ratio test on GLME models, both models must be fitted using a Laplace or approximate Laplace fit method. Models fitted using a maximum pseudo likelihood (MPL) or restricted maximum pseudo likelihood (REMPL) method cannot be compared using a likelihood ratio test. When comparing models fitted using MPL, the maximized log likelihood of the pseudodata from the final pseudo likelihood iteration is used in the likelihood ratio test. If you compare models with non-normal distributions fitted using MPL, then `compare` gives a warning that the likelihood ratio test is using maximized log likelihood of pseudodata from the final pseudo likelihood iteration. To use the true maximized log likelihood in the likelihood ratio test, fit both `glme` and `altglme` using approximate Laplace or Laplace prior to model comparison.

Nesting Requirements

To conduct a valid likelihood ratio test, `glme` must be nested in `altglme`. The `'CheckNesting',true` name-value pair argument checks the following requirements, and returns an error if any are not satisfied:

- You must fit both models (`glme` and `altglme`) using the `'ApproximateLaplace'` or `'Laplace'` fit method. You cannot compare GLME models fitted using `'MPL'` or `'REMPL'` using a likelihood ratio test.
- You must fit both models using the same response vector, response distribution, and link function.
- The smaller model (`glme`) must be nested within the larger model (`altglme`), such that you can obtain `glme` from `altglme` by setting some of the model parameters of `altglme` to fixed values such as 0.
- The maximized log likelihood of the larger model (`altglme`) must be greater than or equal to the maximized log likelihood of the smaller model (`glme`).
- The weight vectors used to fit `glme` and `altglme` must be identical.
- The random-effects design matrix of the larger model (`altglme`) must contain the random-effects design matrix of the smaller model (`glme`).
- The fixed-effects design matrix of the larger model (`altglme`) must contain the fixed-effects design matrix of the smaller model (`glme`).

Akaike and Bayesian Information Criteria

The *Akaike information criterion* (AIC) is $AIC = -2\log L_M + 2(param)$.

$\log L_M$ depends on the method used to fit the model.

- If you use `'Laplace'` or `'ApproximateLaplace'`, then $\log L_M$ is the maximized log likelihood.
- If you use `'MPL'`, then $\log L_M$ is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration.
- If you use `'REMPL'`, then $\log L_M$ is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.

param is the total number of parameters estimated in the model. For most GLME models, *param* is equal to $nc + p + 1$, where nc is the total number of parameters in the random-effects covariance, excluding the residual variance, and p is the number of fixed-effects coefficients. However, if the dispersion parameter is fixed at 1.0 for binomial or Poisson distributions, then *param* is equal to $(nc + p)$.

The *Bayesian information criterion* (BIC) is $BIC = -2*\log L_M + \ln(n_{eff})(param)$.

$\log L_M$ depends on the method used to fit the model.

- If you use `'Laplace'` or `'ApproximateLaplace'`, then $\log L_M$ is the maximized log likelihood.
- If you use `'MPL'`, then $\log L_M$ is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration.
- If you use `'REMPL'`, then $\log L_M$ is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.

n_{eff} is the effective number of observations.

- If you use `'MPL'`, `'Laplace'`, or `'ApproximateLaplace'`, then $n_{eff} = n$, where n is the number of observations.
- If you use `'REMPL'`, then $n_{eff} = n - p$.

param is the total number of parameters estimated in the model. For most GLME models, *param* is equal to $nc + p + 1$, where nc is the total number of parameters in the random-effects covariance, excluding the residual variance, and p is the number of fixed-effects coefficients. However, if the dispersion parameter is fixed at 1.0 for binomial or Poisson distributions, then *param* is equal to ($nc + p$).

A lower value of deviance indicates a better fit. As the value of deviance decreases, both AIC and BIC tend to decrease. Both AIC and BIC also include penalty terms based on the number of parameters estimated, p. So, when the number of parameters increase, the values of AIC and BIC tend to increase as well. When comparing different models, the model with the lowest AIC or BIC value is considered as the best fitting model.

For models fitted using `'MPL'` and `'REMPL'`, AIC and BIC are based on the log likelihood (or restricted log likelihood) of pseudo data from the final pseudo likelihood iteration. Therefore, a direct comparison of AIC and BIC values between models fitted using `'MPL'` and `'REMPL'` is not appropriate.

See Also

`GeneralizedLinearMixedModel` | `covarianceParameters` | `fixedEffects` | `randomEffects`

compare

Class: `LinearMixedModel`

Compare linear mixed-effects models

Syntax

```
results = compare(lme,altlme)
results = compare( ___ ,Name,Value)

[results,siminfo] = compare(lme,altlme,'NSim',nsim)
[results,siminfo] = compare( ___ ,Name,Value)
```

Description

`results = compare(lme,altlme)` returns the results of a likelihood ratio test on page 33-890 that compares the linear mixed-effects models `lme` and `altlme`. Both models must use the same response vector in the fit and `lme` must be nested in `altlme` for a valid theoretical likelihood ratio test. Always input the smaller model first, and the larger model second.

`compare` tests the following null and alternate hypotheses:

H_0: Observed response vector is generated by `lme`.

H_1: Observed response vector is generated by model `altlme`.

It is recommended that you fit `lme` and `altlme` using the maximum likelihood (ML) method prior to model comparison. If you use the restricted maximum likelihood (REML) method, then both models must have the same fixed-effects design matrix.

To test for fixed effects, use `compare` with the simulated likelihood ratio test on page 33-891 when `lme` and `altlme` are fit using ML or use the `fixedEffects`, `anova`, or `coefTest` methods.

`results = compare( ___ ,Name,Value)` also returns the results of a likelihood ratio test that compares linear mixed-effects models `lme` and `altlme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can check if the first input model is nested in the second input model.

`[results,siminfo] = compare(lme,altlme,'NSim',nsim)` returns the results of a simulated likelihood ratio test that compares linear mixed-effects models `lme` and `altlme`.

You can fit `lme` and `altlme` using ML or REML. Also, `lme` does not have to be nested in `altlme`. If you use the restricted maximum likelihood (REML) method to fit the models, then both models must have the same fixed-effects design matrix.

`[results,siminfo] = compare( ___ ,Name,Value)` also returns the results of a simulated likelihood ratio test that compares linear mixed-effects models `lme` and `altlme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can change the options for performing the simulated likelihood ratio test, or change the confidence level of the confidence interval for the *p*-value.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`altlme` — Alternative linear mixed-effects model

`LinearMixedModel` object

Alternative linear mixed-effects model fit to the same response vector but with different model specifications, specified as a `LinearMixedModel` object. `lme` must be nested in `altlme`, that is, `lme` should be obtained from `altlme` by setting some parameters to fixed values, such as 0. You can create a linear mixed-effects object using `fitlme` or `fitlmematrix`.

`nsim` — Number of replications for simulations

positive integer number

Number of replications for simulations in the simulated likelihood ratio test, specified as a positive integer number. You must specify `nsim` to do a simulated likelihood ratio test.

Example: `'NSim',1000`

Data Types: `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Options` — Options for performing simulated likelihood ratio test

structure

Options for performing the simulated likelihood ratio test in parallel, specified as the comma-separated pair consisting of `'Options'`, and a structure created by `statset('LinearMixedModel')`.

These options require Parallel Computing Toolbox.

`compare` uses the following fields.

`'UseParallel'`	• `False` for serial computation. Default. • `True` for parallel computation. You need Parallel Computing Toolbox for parallel computation.
`'UseSubstreams'`	• `False` for not using a separate substream of the random number generator for each iteration. Default. • `True` for using a separate substream of the random number generator for each iteration. You can only use this option with random stream types that support substreams.

`'Streams'`	• If `'UseSubstreams'` is `True`, then `'Streams'` must be a single random number stream, or a scalar cell array containing a single stream.
	• If `'UseSubstreams'` is `False` and
	  • `'UseParallel'` is `False`, then `'Streams'` must be a single random number stream, or a scalar cell array containing a single stream.
	  • `'UseParallel'` is `True`, then `'Streams'` must be equal to the number of processors used. If a parallel pool is open, then the `'Streams'` is the same length as the size of the parallel pool. If `'UseParallel'` is `True`, a parallel pool might open up for you. But since `'Streams'` must be equal to the number of processors used, it is best to open a pool explicitly using the `parpool` command, before calling `compare` with the `'UseParallel','True'` option.

For information on parallel statistical computing at the command line, enter

```
help parallelstats
```

Data Types: `struct`

`CheckNesting` — Indicator to check nesting between two models
`false` (default) | `true`

Indicator to check nesting on page 33-891 between two models, specified as the comma-separated pair consisting of `'CheckNesting'` and one of the following.

`false`	Default. No checks.
`true`	`compare` checks if the smaller model `lme` is nested in the bigger model `altlme`.

`lme` must be nested in the alternate model `altlme` for a valid theoretical likelihood ratio test on page 33-890. `compare` returns an error message if the nesting requirements are not satisfied.

Although valid for both tests, the nesting requirements are weaker for the simulated likelihood ratio test on page 33-891.

Example: `'CheckNesting',true`

Data Types: `single` | `double`

Output Arguments

`results` — Results of likelihood ratio test or simulated likelihood ratio test

dataset array

Results of the likelihood ratio test or simulated likelihood ratio test, returned as a dataset array with two rows. The first row is for `lme`, and the second row is for `altlme`. The columns of `results` depend on whether the test is a likelihood ratio or a simulated likelihood ratio test.

- If you use the likelihood ratio test on page 33-890, then `results` contains the following columns.

`Model`	Name of the model
`DF`	Degrees of freedom, that is, the number of free parameters in the model
`AIC`	Akaike information criterion for the model
`BIC`	Bayesian information criterion for the model
`LogLik`	Maximized log likelihood for the model
`LRStat`	Likelihood ratio test statistic for comparing `altlme` versus `lme`
`deltaDF`	`DF` for `altlme` minus `DF` for `lme`
`pValue`	*p*-value for the likelihood ratio test

- If you use the simulated likelihood ratio test on page 33-891, then `results` contains the following columns.

`Model`	Name of the model
`DF`	Degrees of freedom, that is, the number of free parameters in the model
`LogLik`	Maximized log likelihood for the model
`LRStat`	Likelihood ratio test statistic for comparing `altlme` versus `lme`
`pValue`	*p*-value for the likelihood ratio test
`Lower`	Lower limit of the confidence interval for `pValue`
`Upper`	Upper limit of the confidence interval for `pValue`

`siminfo` — Simulation output

structure

Simulation output, returned as a structure with the following fields.

`nsim`	Value set for `nsim`.
`alpha`	Value set for `'Alpha'`.
`pValueSim`	Simulation-based *p*-value.
`pValueSimCI`	Confidence interval for `pValueSim`. The first element of the vector is the lower limit and the second element of the vector contains the upper limit.
`deltaDF`	The number of free parameters in `altlme` minus the number of free parameters in `lme`. `DF` for `altlme` minus `DF` for `lme`.
`TH0`	A vector of simulated likelihood ratio test statistics under the null hypothesis that the model `lme` generated the observed response vector `y`.

Examples

Test for Random Effects

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the CDC).

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses and region as the predictor variable, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the response variable, `FluRate`, the nominal variable, `Region`, that shows which region each estimate is from, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate',...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Fit a linear mixed-effects model, with a varying intercept and varying slope for each region, grouped by `Date`.

```
altlme = fitlme(flu2,'FluRate ~ 1 + Region + (1 + Region|Date)');
```

Fit a linear mixed-effects model with fixed effects for the region and a random intercept that varies by `Date`.

```
lme = fitlme(flu2,'FluRate ~ 1 + Region + (1|Date)');
```

Compare the two models. Also check if `lme2` is nested in `lme`.

```
compare(lme,altlme,'CheckNesting',true)
```

```
ans =
    Theoretical Likelihood Ratio Test
```

```
    Model     DF    AIC        BIC        LogLik     LRStat    deltaDF    pValue
    lme       11     318.71     364.35    -148.36
    altlme    55    -305.51    -77.346     207.76    712.22    44         0
```

The small *p*-value of 0 indicates that model `altlme` is significantly better than the simpler model `lme`.

Test for Fixed and Random Effects

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type) and the plots within blocks (tomato types within soil types) independently.

```
lmeBig = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Refit the model after removing the interaction term `Tomato:Fertilizer` and the random-effects term `(1 | Soil)`.

```
lmeSmall = fitlme(ds,'Yield ~ Fertilizer + Tomato + (1|Soil:Tomato)');
```

Compare the two models using the simulated likelihood ratio test with 1000 replications. You must use this test to test for both fixed- and random-effect terms. Note that both models are fit using the default fitting method, ML. That's why, there is no restriction on the fixed-effects design matrices. If you use restricted maximum likelihood (REML) method, both models must have identical fixed-effects design matrices.

```
[table,siminfo] = compare(lmeSmall,lmeBig,'nsim',1000)
```

```
table =
    Simulated Likelihood Ratio Test: Nsim = 1000, Alpha = 0.05

    Model       DF    AIC       BIC       LogLik     LRStat    pValue
    lmeSmall    10    511.06       532    -245.53
    lmeBig      23    522.57    570.74    -238.29    14.491    0.57343


    Lower      Upper
```

```
    0.54211    0.60431


siminfo = struct with fields:
           nsim: 1000
          alpha: 0.0500
      pvalueSim: 0.5734
    pvalueSimCI: [0.5421 0.6043]
        deltaDF: 13
            TH0: [1000x1 double]
```

The high *p*-value suggests that the larger model, `lme` is not significantly better than the smaller model, `lme2`. The smaller values of "Akaike and Bayesian Information Criteria" on page 33-892 for `lme2` also support this.

Models with Correlated and Uncorrelated Random Effects

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and the cylinders, and potentially correlated random effects for intercept and acceleration grouped by model year.

First, prepare the design matrices.

```
X = [ones(406,1) Acceleration Horsepower];
Z = [ones(406,1) Acceleration];
Model_Year = nominal(Model_Year);
G = Model_Year;
```

Now, fit the model using `fitlmematrix` with the defined design matrices and grouping variables.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'}},'RandomEffectGroups',{'Model_Year'});
```

Refit the model with uncorrelated random effects for intercept and acceleration. First prepare the random effects design and the random effects grouping variables.

```
Z = {ones(406,1),Acceleration};
G = {Model_Year,Model_Year};

lme2 = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept'},{'Acceleration'}},'RandomEffectGroups',...
{'Model_Year','Model_Year'});
```

Compare `lme` and `lme2` using the simulated likelihood ratio test.

```
compare(lme2,lme,'CheckNesting',true,'NSim',1000)
```

```
ans =


    SIMULATED LIKELIHOOD RATIO TEST: NSIM = 1000, ALPHA = 0.05

    Model    DF    AIC       BIC       LogLik     LRStat    pValue      Lower
    lme2     6     2194.5    2218.3    -1091.3
    lme      7     2193.5    2221.3    -1089.7    3.0323    0.095904    0.078373


    Upper

    0.11585
```

The high p-value indicates that `lme2` is not a significantly better fit than `lme`.

Simulated Likelihood Ratio Test Using Parallel Computing

Load the sample data.

```
load('fertilizer.mat')
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a table called `tbl`, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
tbl = dataset2table(fertilizer);
tbl.Tomato = categorical(tbl.Tomato);
tbl.Soil = categorical(tbl.Soil);
tbl.Fertilizer = categorical(tbl.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(tbl,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Refit the model after removing the interaction term `Tomato:Fertilizer` and the random-effects term `(1|Soil)`.

```
lme2 = fitlme(tbl,'Yield ~ Fertilizer + Tomato + (1|Soil:Tomato)');
```

Create the options structure for `LinearMixedModel`.

```
opt = statset('LinearMixedModel')
```

```
opt = struct with fields:
          Display: 'off'
      MaxFunEvals: []
```

```
          MaxIter: 10000
           TolBnd: []
           TolFun: 1.0000e-06
       TolTypeFun: []
             TolX: 1.0000e-12
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

Change the options for parallel testing.

```
opt.UseParallel = true;
```

Start a parallel environment.

```
mypool = parpool();
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Compare `lme2` and `lme` using the simulated likelihood ratio test with 1000 replications and parallel computing.

```
compare(lme2,lme,'nsim',1000,'Options',opt)
```

```
ans =
    Simulated Likelihood Ratio Test: Nsim = 1000, Alpha = 0.05

    Model    DF    AIC       BIC       LogLik     LRStat    pValue     Lower    Upper
    lme2     10    511.06       532    -245.53
    lme      23    522.57    570.74    -238.29    14.491    0.53447    0.503    0.56573
```

The high *p*-value suggests that the larger model, `lme` is not significantly better than the smaller model, `lme2`. The smaller values of `AIC` and `BIC` for `lme2` also support this.

More About

Likelihood Ratio Test

Under the null hypothesis H_0, the observed likelihood ratio test statistic has an approximate chi-squared reference distribution with degrees of freedom `deltaDF`. When comparing two models, `compare` computes the *p*-value for the likelihood ratio test by comparing the observed likelihood ratio test statistic with this chi-squared reference distribution.

The *p*-values obtained using the likelihood ratio test can be conservative when testing for the presence or absence of random-effects terms and anticonservative when testing for the presence or absence of fixed-effects terms. Hence, use the `fixedEffects`, `anova`, or `coefTest` method or the simulated likelihood ratio test while testing for fixed effects.

Simulated Likelihood Ratio Test

To perform the simulated likelihood ratio test, `compare` first generates the reference distribution of the likelihood ratio test statistic under the null hypothesis. Then, it assesses the statistical significance of the alternate model by comparing the observed likelihood ratio test statistic to this reference distribution.

`compare` produces the simulated reference distribution of the likelihood ratio test statistic under the null hypothesis as follows:

- Generate random data `ysim` from the fitted model `lme`.
- Fit the model specified in `lme` and alternate model `altlme` to the simulated data `ysim`.
- Calculate the likelihood ratio test statistic using results from step 2 and store the value.
- Repeat step 1 to 3 `nsim` times.

Then, `compare` computes the *p*-value for the simulated likelihood ratio test by comparing the observed likelihood ratio test statistic with the simulated reference distribution. The *p*-value estimate is the ratio of the number of times the simulated likelihood ratio test statistic is equal to or exceeds the observed value plus one, to the number of replications plus one.

Suppose the observed likelihood ratio statistic is T, and the simulated reference distribution is stored in vector T_{H_0}. Then,

$$p-value=\frac{\left[\sum_{j=1}^{nsim} I\left(T_{H_0}(j)\geq T\right)\right]+1}{nsim+1}.$$

To account for the uncertainty in the simulated reference distribution, `compare` computes a 100*(1 – α)% confidence interval for the true *p*-value.

You can use the simulated likelihood ratio test to compare arbitrary linear mixed-effects models. That is, when you are using the simulated likelihood ratio test, `lme` does not have to be nested within `altlme`, and you can fit `lme` and `altlme` using either maximum likelihood (ML) or restricted maximum likelihood (REML) methods. If you use the restricted maximum likelihood (REML) method to fit the models, then both models must have the same fixed-effects design matrix.

Nesting Requirements

The `'CheckNesting','True'` name-value pair argument checks the following requirements.

For a simulated likelihood ratio test:

- You must use the same method to fit both models (`lme` and `altlme`). `compare` cannot compare a model fit using ML to a model fit using REML.
- You must fit both models to the same response vector.
- If you use REML to fit `lme` and `altlme`, then both models must have the same fixed-effects design matrix.

- The maximized log likelihood or restricted log likelihood of the bigger model (`altlme`) must be greater than or equal to that of the smaller model (`lme`).

For a theoretical test, `'CheckNesting','True'` checks all the requirements listed for a simulated likelihood ratio test and the following:

- Weight vectors you use to fit `lme` and `altlme` must be identical.
- If you use ML to fit `lme` and `altlme`, the fixed-effects design matrix of the bigger model (`altlme`) must contain that of the smaller model (`lme`).
- The random-effects design matrix of the bigger model (`altlme`) must contain that of the smaller model (`lme`).

Akaike and Bayesian Information Criteria

Akaike information criterion (AIC) is $AIC = -2*\log L_M + 2*(nc + p + 1)$, where $\log L_M$ is the maximized log likelihood (or maximized restricted log likelihood) of the model, and $nc + p + 1$ is the number of parameters estimated in the model. p is the number of fixed-effects coefficients, and nc is the total number of parameters in the random-effects covariance excluding the residual variance.

Bayesian information criterion (BIC) is $BIC = -2*\log L_M + \ln(n_{eff})*(nc + p + 1)$, where $\log L_M$ is the maximized log likelihood (or maximized restricted log likelihood) of the model, n_{eff} is the effective number of observations, and $(nc + p + 1)$ is the number of parameters estimated in the model.

- If the fitting method is maximum likelihood (ML), then $n_{eff} = n$, where n is the number of observations.
- If the fitting method is restricted maximum likelihood (REML), then $n_{eff} = n - p$.

A lower value of deviance indicates a better fit. As the value of deviance decreases, both AIC and BIC tend to decrease. Both AIC and BIC also include penalty terms based on the number of parameters estimated, p. So, when the number of parameters increase, the values of AIC and BIC tend to increase as well. When comparing different models, the model with the lowest AIC or BIC value is considered as the best fitting model.

Deviance

`LinearMixedModel` computes the deviance of model M as minus two times the loglikelihood of that model. Let L_M denote the maximum value of the likelihood function for model M. Then, the deviance of model M is

$$-2*\log L_M .$$

A lower value of deviance indicates a better fit. Suppose M_1 and M_2 are two different models, where M_1 is nested in M_2. Then, the fit of the models can be assessed by comparing the deviances Dev_1 and Dev_2 of these models. The difference of the deviances is

$$Dev = Dev_1 - Dev_2 = 2(\log LM_2 - \log LM_1) .$$

Usually, the asymptotic distribution of this difference has a chi-square distribution with degrees of freedom v equal to the number of parameters that are estimated in one model but fixed (typically at 0) in the other. That is, it is equal to the difference in the number of parameters estimated in M_1 and M_2. You can get the p-value for this test using `1 - chi2cdf(Dev,V)`, where $Dev = Dev_2 - Dev_1$.

However, in mixed-effects models, when some variance components fall on the boundary of the parameter space, the asymptotic distribution of this difference is more complicated. For example, consider the hypotheses

H_0: $D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix}$, D is a q-by-q symmetric positive semidefinite matrix.

H_1: D is a (q+1)-by-(q+1) symmetric positive semidefinite matrix.

That is, H_1 states that the last row and column of D are different from zero. Here, the bigger model M_2 has $q + 1$ parameters and the smaller model M_1 has q parameters. And *Dev* has a 50:50 mixture of ${\chi^2}_q$ and ${\chi^2}_{(q+1)}$ distributions (Stram and Lee, 1994).

References

[1] Hox, J. *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Inc., 2002.

[2] Stram D. O. and J. W. Lee. “Variance components testing in the longitudinal mixed-effects model”. *Biometrics*, Vol. 50, 4, 1994, pp. 1171–1177.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`LinearMixedModel` | `anova` | `covarianceParameters` | `fitlme` | `fitlmematrix` | `fixedEffects` | `randomEffects`

compareHoldout

Package:

Compare accuracies of two classification models using new data

Syntax

```
h = compareHoldout(C1,C2,T1,T2,ResponseVarName)

h = compareHoldout(C1,C2,T1,T2,Y)
h = compareHoldout(C1,C2,X1,X2,Y)

h = compareHoldout( ___ ,Name,Value)
[h,p,e1,e2] = compareHoldout( ___ )
```

Description

`compareHoldout` statistically assesses the accuracies of two classification models. The function first compares their predicted labels against the true labels, and then it detects whether the difference between the misclassification rates is statistically significant.

You can determine whether the accuracies of the classification models differ or whether one model performs better than another. `compareHoldout` can conduct several McNemar test on page 33-909 variations, including the asymptotic test, the exact-conditional test, and the mid-*p*-value test. For cost-sensitive assessment on page 33-908, available tests include a chi-square test (requires Optimization Toolbox) and a likelihood ratio test.

`h = compareHoldout(C1,C2,T1,T2,ResponseVarName)` returns the test decision from testing the null hypothesis that the trained classification models `C1` and `C2` have equal accuracy for predicting the true class labels in the `ResponseVarName` variable. The alternative hypothesis is that the labels have unequal accuracy.

The first classification model `C1` uses the predictor data in `T1`, and the second classification model `C2` uses the predictor data in `T2`. The tables `T1` and `T2` must contain the same response variable but can contain different sets of predictors. By default, the software conducts the mid-*p*-value McNemar test to compare the accuracies.

`h = 1` indicates rejecting the null hypothesis at the 5% significance level. `h = 0` indicates not rejecting the null hypothesis at the 5% level.

The following are examples of tests you can conduct:

- Compare the accuracies of a simple classification model and a model that is more complex by passing the same set of predictor data (that is, `T1 = T2`).
- Compare the accuracies of two potentially different models using two potentially different sets of predictors.
- Perform various types of Feature Selection on page 15-49. For example, you can compare the accuracy of a model trained using a set of predictors to the accuracy of one trained on a subset or different set of those predictors. You can choose the set of predictors arbitrarily, or use a feature selection technique such as PCA or sequential feature selection (see `pca` and `sequentialfs`).

`h = compareHoldout(C1,C2,T1,T2,Y)` returns the test decision from testing the null hypothesis that the trained classification models `C1` and `C2` have equal accuracy for predicting the true class labels `Y`. The alternative hypothesis is that the labels have unequal accuracy.

The first classification model `C1` uses the predictor data `T1`, and the second classification model `C2` uses the predictor data `T2`. By default, the software conducts the mid-*p*-value McNemar test to compare the accuracies.

`h = compareHoldout(C1,C2,X1,X2,Y)` returns the test decision from testing the null hypothesis that the trained classification models `C1` and `C2` have equal accuracy for predicting the true class labels `Y`. The alternative hypothesis is that the labels have unequal accuracy.

The first classification model `C1` uses the predictor data `X1`, and the second classification model `C2` uses the predictor data `X2`. By default, the software conducts the mid-*p*-value McNemar test to compare the accuracies.

`h = compareHoldout( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input argument combinations in previous syntaxes. For example, you can specify the type of alternative hypothesis, specify the type of test, and supply a cost matrix.

`[h,p,e1,e2] = compareHoldout( ___ )` returns the *p*-value for the hypothesis test (`p`) and the respective classification losses on page 33-912 of each set of predicted class labels (`e1` and `e2`) using any of the input arguments in the previous syntaxes.

Examples

Compare Accuracies of Full and Reduced Classification Models

Train two *k*-nearest neighbor classifiers, one using a subset of the predictors used for the other. Conduct a statistical test comparing the accuracies of the two models on a test set.

Load the `carsmall` data set.

```
load carsmall
```

Create two tables of input data, where the second table excludes the predictor `Acceleration`. Specify `Model_Year` as the response variable.

```
T1 = table(Acceleration,Displacement,Horsepower,MPG,Model_Year);
T2 = T1(:,2:end);
```

Create a partition that splits the data into training and test sets. Keep 30% of the data for testing.

```
rng(1)  % For reproducibility
CVP = cvpartition(Model_Year,'holdout',0.3);
idxTrain = training(CVP);   % Training-set indices
idxTest = test(CVP);    % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Train the `ClassificationKNN` models using the `T1` and `T2` data.

```
C1 = fitcknn(T1(idxTrain,:),'Model_Year');
C2 = fitcknn(T2(idxTrain,:),'Model_Year');
```

`C1` and `C2` are trained `ClassificationKNN` models.

Test whether the two models have equal predictive accuracies on the test set.

```
h = compareHoldout(C1,C2,T1(idxTest,:),T2(idxTest,:),'Model_Year')
```

```
h = logical
   0
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Compare Accuracies of Two Different Classification Models

Train two classification models using different algorithms. Conduct a statistical test comparing the misclassification rates of the two models on a test set.

Load the `ionosphere` data set.

```
load ionosphere
```

Create a partition that evenly splits the data into training and test sets.

```
rng(1)                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);           % Training-set indices
idxTest = test(CVP);                % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Train an SVM model and an ensemble of 100 bagged classification trees. For the SVM model, specify to use the radial basis function kernel and a heuristic procedure to determine the kernel scale.

```
C1 = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true, ...
    'KernelFunction','RBF','KernelScale','auto');
t = templateTree('Reproducible',true);  % For reproducibility of random predictor selections
C2 = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','Bag', ...
    'Learners',t);
```

`C1` is a trained `ClassificationSVM` model. `C2` is a trained `ClassificationBaggedEnsemble` model.

Test whether the two models have equal predictive accuracies. Use the same test-set predictor data for each model.

```
h = compareHoldout(C1,C2,X(idxTest,:),X(idxTest,:),Y(idxTest))
```

```
h = logical
   0
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Compare Classification Model to More Complex Model

Train two classification models using the same algorithm, but adjust a hyperparameter to make the algorithm more complex. Conduct a statistical test to assess whether the simpler model has better accuracy on test data than the more complex model.

Load the `ionosphere` data set.

```
load ionosphere;
```

Create a partition that evenly splits the data into training and test sets.

```
rng(1);                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);            % Training-set indices
idxTest = test(CVP);                 % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Train two SVM models: one that uses a linear kernel (the default for binary classification) and one that uses the radial basis function kernel. Use the default kernel scale of 1.

```
C1 = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true);
C2 = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true,...
    'KernelFunction','RBF');
```

`C1` and `C2` are trained `ClassificationSVM` models.

Test the null hypothesis that the simpler model (`C1`) is at most as accurate as the more complex model (`C2`). Because the test-set size is large, conduct the asymptotic McNemar test, and compare the results with the mid-*p*-value test (the cost-insensitive testing default). Request to return *p*-values and misclassification rates.

```
Asymp = zeros(4,1); % Preallocation
MidP = zeros(4,1);

[Asymp(1),Asymp(2),Asymp(3),Asymp(4)] = compareHoldout(C1,C2,...
    X(idxTest,:),X(idxTest,:),Y(idxTest),'Alternative','greater',...
    'Test','asymptotic');
[MidP(1),MidP(2),MidP(3),MidP(4)] = compareHoldout(C1,C2,...
    X(idxTest,:),X(idxTest,:),Y(idxTest),'Alternative','greater');
table(Asymp,MidP,'RowNames',{'h' 'p' 'e1' 'e2'})
```

```
ans=4×2 table
            Asymp          MidP
          __________    __________

    h              1             1
    p     7.2801e-09    2.7649e-10
    e1       0.13714       0.13714
    e2       0.33143       0.33143
```

The *p*-value is close to zero for both tests, providing strong evidence to reject the null hypothesis that the simpler model is less accurate than the more complex model. No matter what test you specify, `compareHoldout` returns the same type of misclassification measure for both models.

Conduct Cost-Sensitive Comparison of Two Classification Models

For data sets with imbalanced class representations, or for data sets with imbalanced false-positive and false-negative costs, you can statistically compare the predictive performance of two classification models by including a cost matrix in the analysis.

Load the `arrhythmia` data set. Determine the class representations in the data.

```
load arrhythmia;
Y = categorical(Y);
tabulate(Y);
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

There are 16 classes, however some are not represented in the data set (for example, class 13). Most observations are classified as not having arrhythmia (class 1). The data set is highly discrete with imbalanced classes.

Combine all observations with arrhythmia (classes 2 through 15) into one class. Remove those observations with unknown arrhythmia status (class 16) from the data set.

```
idx = (Y ~= '16');
Y = Y(idx);
X = X(idx,:);
Y(Y ~= '1') = 'WithArrhythmia';
Y(Y == '1') = 'NoArrhythmia';
Y = removecats(Y);
```

Create a partition that evenly splits the data into training and test sets.

```
rng(1);                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);            % Training-set indices
idxTest = test(CVP);                 % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Create a cost matrix such that misclassifying a patient with arrhythmia into the "no arrhythmia" class is five times worse than misclassifying a patient without arrhythmia into the arrhythmia class. Classifying correctly incurs no cost. The rows indicate the true class and the columns indicate the predicted class. When you conduct a cost-sensitive analysis, a good practice is to specify the order of the classes.

```
cost = [0 1;5 0];
ClassNames = {'NoArrhythmia','WithArrhythmia'};
```

Train two boosting ensembles of 50 classification trees, one that uses AdaBoostM1 and another that uses LogitBoost. Because the data set contains missing values, specify to use surrogate splits. Train the models using the cost matrix.

```
t = templateTree('Surrogate','on');
numTrees = 50;
C1 = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','AdaBoostM1', ...
    'NumLearningCycles',numTrees,'Learners',t, ...
    'Cost',cost,'ClassNames',ClassNames);
C2 = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','LogitBoost', ...
    'NumLearningCycles',numTrees,'Learners',t, ...
    'Cost',cost,'ClassNames',ClassNames);
```

`C1` and `C2` are trained `ClassificationEnsemble` models.

Test whether the AdaBoostM1 ensemble (`C1`) and the LogitBoost ensemble (`C2`) have equal predictive accuracy. Supply the cost matrix. Conduct the asymptotic, likelihood ratio, cost-sensitive test (the default when you pass in a cost matrix). Request to return *p*-values and misclassification costs.

```
[h,p,e1,e2] = compareHoldout(C1,C2,X(idxTest,:),X(idxTest,:),Y(idxTest),...
    'Cost',cost)
```

```
h = logical
   0


p = 0.3334

e1 = 0.5581

e2 = 0.4698
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Select Features Using Statistical Accuracy Comparison

Reduce classification model complexity by selecting a subset of predictor variables (features) from a larger set. Then, statistically compare the out-of-sample accuracy between the two models.

Load the `ionosphere` data set.

```
load ionosphere;
```

Create a partition that evenly splits the data into training and test sets.

```
rng(1);  % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP); % Training-set indices
idxTest = test(CVP);      % Test-set indices
```

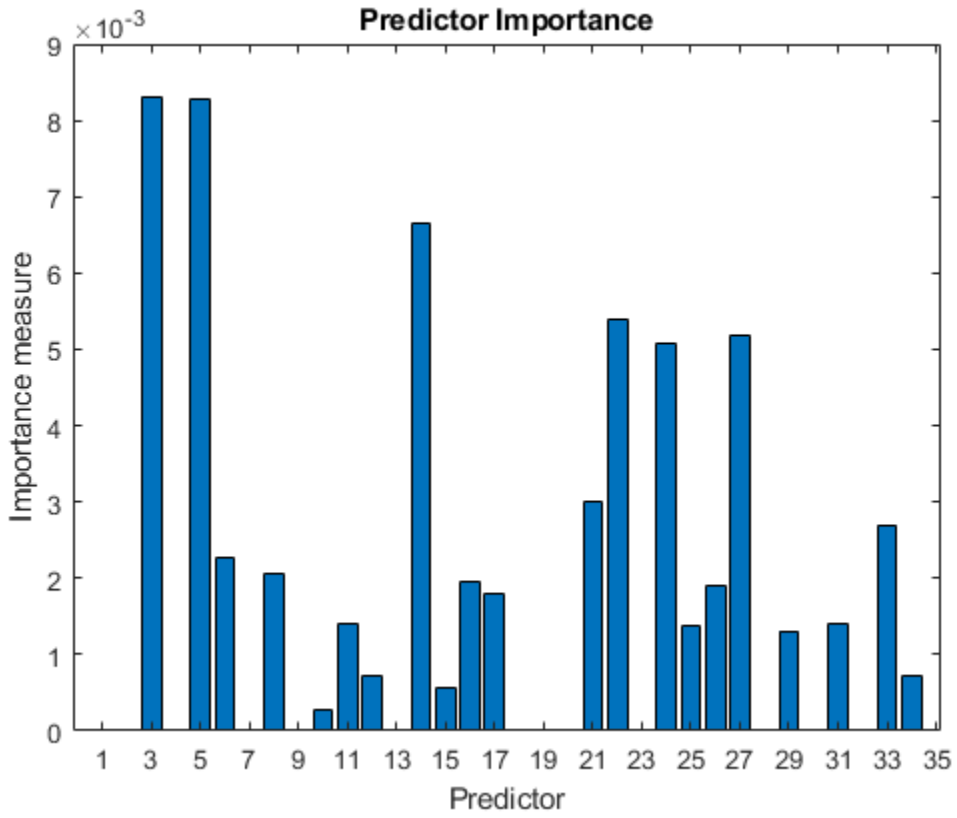
`CVP` is a cross-validation partition object that specifies the training and test sets.

Train an ensemble of 100 boosted classification trees using AdaBoostM1 and the entire set of predictors. Inspect the importance measure for each predictor.

```
t = templateTree('MaxNumSplits',1); % Weak-learner template tree object
C2 = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','AdaBoostM1',...
    'Learners',t);
predImp = predictorImportance(C2);

figure;
bar(predImp);
h = gca;
h.XTick = 1:2:h.XLim(2)
```

```
h =
  Axes with properties:

            XLim: [-0.2000 35.2000]
            YLim: [0 0.0090]
          XScale: 'linear'
          YScale: 'linear'
    GridLineStyle: '-'
         Position: [0.1300 0.1100 0.7750 0.8150]
            Units: 'normalized'

  Show all properties
```

```
title('Predictor Importance');
xlabel('Predictor');
ylabel('Importance measure');
```



Identify the top five predictors in terms of their importance.

```
[~,idxSort] = sort(predImp,'descend');
idx5 = idxSort(1:5);
```

Train another ensemble of 100 boosted classification trees using AdaBoostM1 and the five predictors with the greatest importance.

```
C1 = fitcensemble(X(idxTrain,idx5),Y(idxTrain),'Method','AdaBoostM1',...
    'Learners',t);
```

Test whether the two models have equal predictive accuracies. Specify the reduced test-set predictor data for `C1` and the full test-set predictor data for `C2`.

```
[h,p,e1,e2] = compareHoldout(C1,C2,X(idxTest,idx5),X(idxTest,:),Y(idxTest))
```

```
h = logical
   0


p = 0.7744

e1 = 0.0914

e2 = 0.0857
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies. This result favors the simpler ensemble, `C1`.

Input Arguments

C1 — First trained classification model

trained classification model object | trained, compact classification model object

First trained classification model, specified as any trained classification model object or compact classification model object described in this table.

Trained Model Type	Model Object	Returned By
Classification tree	`ClassificationTree`	`fitctree`
Discriminant analysis	`ClassificationDiscriminant`	`fitcdiscr`
Ensemble of bagged classification models	`ClassificationBaggedEnsemble`	`fitcensemble`
Ensemble of classification models	`ClassificationEnsemble`	`fitcensemble`
Multiclass, error-correcting output codes (ECOC)	`ClassificationECOC`	`fitcecoc`
Generalized additive model (GAM)	`ClassificationGAM`	`fitcgam`
*k*NN	`ClassificationKNN`	`fitcknn`
Naive Bayes	`ClassificationNaiveBayes`	`fitcnb`
Neural network	`ClassificationNeuralNetwork` (with observations in rows)	`fitcnet`
Support vector machine (SVM)	`ClassificationSVM`	`fitcsvm`
Compact discriminant analysis	`CompactClassificationDiscriminant`	`compact`
Compact ECOC	`CompactClassificationECOC`	`compact`
Compact ensemble of classification models	`CompactClassificationEnsemble`	`compact`
Compact GAM	`CompactClassificationGAM`	`compact`
Compact naive Bayes	`CompactClassificationNaiveBayes`	`compact`
Compact neural network	`CompactClassificationNeuralNetwork`	`compact`
Compact SVM	`CompactClassificationSVM`	`compact`
Compact classification tree	`CompactClassificationTree`	`compact`

C2 — Second trained classification model

trained classification model object | trained, compact classification model object

Second trained classification model, specified as any trained classification model object or compact classification model object that is a valid choice for `C1`.

T1 — Test-set predictor data for first classification model

table

Test-set predictor data for the first classification model, `C1`, specified as a table. Each row of `T1` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `T1` can contain an additional column for the response variable. `T1` must contain all the predictors used to train `C1`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`T1` and `T2` must have the same number of rows and the same response values. If `T1` and `T2` contain the response variable used to train `C1` and `C2`, then you do not need to specify `ResponseVarName` or `Y`.

Data Types: `table`

`T2` — Test-set predictor data for second classification model
table

Test-set predictor data for the second classification model, `C2`, specified as a table. Each row of `T2` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `T2` can contain an additional column for the response variable. `T2` must contain all the predictors used to train `C2`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`T1` and `T2` must have the same number of rows and the same response values. If `T1` and `T2` contain the response variable used to train `C1` and `C2`, then you do not need to specify `ResponseVarName` or `Y`.

Data Types: `table`

`X1` — Test-set predictor data for first classification model
numeric matrix

Test-set predictor data for the first classification model, `C1`, specified as a numeric matrix.

Each row of `X1` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a predictor or feature). The variables used to train `C1` must compose `X1`.

The number of rows in `X1` and `X2` must equal the number of rows in `Y`.

Data Types: `double` | `single`

`X2` — Test-set predictor data for second classification model
numeric matrix

Test-set predictor data for the second classification model, `C2`, specified as a numeric matrix.

Each row of `X2` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a predictor or feature). The variables used to train `C2` must compose `X2`.

The number of rows in `X2` and `X1` must equal the number of rows in `Y`.

Data Types: `double` | `single`

`ResponseVarName` — Response variable name
name of a variable in `T1` and `T2`

Response variable name, specified as the name of a variable in `T1` and `T2`. If `T1` and `T2` contain the response variable used to train `C1` and `C2`, then you do not need to specify `ResponseVarName`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable is stored as `T1.Response`, then specify it as `'Response'`. Otherwise, the software treats all columns of `T1` and `T2`, including `Response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Y — True class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

True class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

When you specify `Y`, `compareHoldout` treats all variables in the matrices `X1` and `X2` or the tables `T1` and `T2` as predictor variables.

If `Y` is a character array, then each element must correspond to one row of the array.

The number of rows in the predictor data must equal the number of rows in `Y`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Note `NaN`s, `<undefined>` values, empty character vectors (`''`), empty strings (`""`), and `<missing>` values indicate missing values. `compareHoldout` removes missing values in `Y` and the corresponding rows of `X1` and `X2`. Additionally, `compareHoldout` predicts classes whether `X1` and `X2` have missing observations.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `compareHoldout(C1,C2,X1,X2,Y,'Alternative','greater','Test','asymptotic','Cost',[0 2;1 0])` tests whether the first set of predicted class labels is more accurate than the second set, conducts the asymptotic McNemar test, and penalizes misclassifying observations with the true label `ClassNames{1}` twice as much as misclassifying observations with the true label `ClassNames{2}`.

`Alpha` — Hypothesis test significance level
`0.05` (default) | scalar value in the interval (0,1)

Hypothesis test significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the interval (0,1).

Example: `'Alpha',0.1`

Data Types: `single` | `double`

`Alternative` — Alternative hypothesis to assess
`'unequal'` (default) | `'greater'` | `'less'`

Alternative hypothesis to assess, specified as the comma-separated pair consisting of `'Alternative'` and one of the values listed in this table.

Value	Alternative Hypothesis
`'unequal'` (default)	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` and `C2` applied to `X2` have unequal accuracies.
`'greater'`	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` is more accurate than `C2` applied to `X2`.
`'less'`	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` is less accurate than `C2` applied to `X2`.

Example: `'Alternative','greater'`

`ClassNames` — Class names
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class names, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. You must set `ClassNames` using the data type of `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of any input argument dimension that corresponds to class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost`.
- Select a subset of classes for testing. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train and test models using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default is the set of all distinct class names in `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure array.

- If you specify the square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- If you specify the structure `S`, then `S` must have two fields:

- `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`. You can use this field to specify the order of the classes.
- `S.ClassificationCosts`, which contains the cost matrix, with rows and columns ordered as in `S.ClassNames`.

If you specify `Cost`, then `compareHoldout` cannot conduct one-sided, exact, or mid-*p* tests. You must also specify `'Alternative','unequal','Test','asymptotic'`. For cost-sensitive testing options, see the `CostTest` name-value pair argument.

A best practice is to supply the same cost matrix used to train the classification models.

The default is `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `single` | `double` | `struct`

`CostTest` — Cost-sensitive test type
`'likelihood'` (default) | `'chisquare'`

Cost-sensitive test type, specified as the comma-separated pair consisting of `'CostTest'` and `'chisquare'` or `'likelihood'`. If you do not specify a cost matrix using the `Cost` name-value pair argument, `compareHoldout` ignores `CostTest`.

This table summarizes the available options for cost-sensitive testing.

Value	Asymptotic Test Type	Requirements
`'chisquare'`	Chi-square test	Optimization Toolbox to implement `quadprog`
`'likelihood'`	Likelihood ratio test	None

For more details, see “Cost-Sensitive Testing” on page 33-908.

Example: `'CostTest','chisquare'`

`Test` — Test to conduct
`'asymptotic'` | `'exact'` | `'midp'`

Test to conduct, specified as the comma-separated pair consisting of `'Test'` and `'asymptotic'`, `'exact'`, or `'midp'`.

This table summarizes the available options for cost-insensitive testing.

Value	Description
`'asymptotic'`	Asymptotic McNemar test
`'exact'`	Exact-conditional McNemar test
`'midp'` (default)	Mid-*p*-value McNemar test

For more details, see “McNemar Tests” on page 33-909.

For cost-sensitive testing, `Test` must be `'asymptotic'`. When you specify the `Cost` name-value pair argument and choose a cost-sensitive test using the `CostTest` name-value pair argument, `'asymptotic'` is the default.

Example: `'Test','asymptotic'`

Output Arguments

h — Hypothesis test result

`1` | `0`

Hypothesis test result, returned as a logical value.

`h = 1` indicates the rejection of the null hypothesis at the `Alpha` significance level.

`h = 0` indicates failure to reject the null hypothesis at the `Alpha` significance level.

Data Types: `logical`

p — *p*-value

scalar in the interval [0,1]

p-value of the test, returned as a scalar in the interval [0,1]. `p` is the probability that a random test statistic is at least as extreme as the observed test statistic, given that the null hypothesis is true.

`compareHoldout` estimates `p` using the distribution of the test statistic, which varies with the type of test. For details on test statistics derived from the available variants of the McNemar test, see “McNemar Tests” on page 33-909. For details on test statistics derived from cost-sensitive tests, see “Cost-Sensitive Testing” on page 33-908.

Data Types: `double`

e1 — Classification loss

numeric scalar

Classification loss, returned as a numeric scalar. `e1` summarizes the accuracy of the first set of class labels predicting the true class labels (`Y`). `compareHoldout` applies the first test-set predictor data (`X1`) to the first classification model (`C1`) to estimate the first set of class labels. Then, the function compares the estimated labels to `Y` to obtain the classification loss.

For cost-insensitive testing, `e1` is the misclassification rate. That is, `e1` is the proportion of misclassified observations, which is a scalar in the interval [0,1].

For cost-sensitive testing, `e1` is the misclassification cost. That is, `e1` is the weighted average of the misclassification costs, in which the weights are the respective estimated proportions of misclassified observations.

For more information, see “Classification Loss” on page 33-912.

Data Types: `double`

e2 — Classification loss

numeric scalar

Classification loss, returned as a numeric scalar. `e2` summarizes the accuracy of the second set of class labels predicting the true class labels (`Y`). `compareHoldout` applies the second test-set predictor data (`X2`) to the second classification model (`C2`) to estimate the second set of class labels. Then, the function compares the estimated labels to `Y` to obtain the classification loss.

For cost-insensitive testing, `e2` is the misclassification rate. That is, `e2` is the proportion of misclassified observations, which is a scalar in the interval [0,1].

For cost-sensitive testing, `e2` is the misclassification cost. That is, `e2` is the weighted average of the misclassification costs, in which the weights are the respective estimated proportions of misclassified observations.

For more information, see “Classification Loss” on page 33-912.

Data Types: `double`

Limitations

- `compareHoldout` does not compare ECOC models composed of linear or kernel classification models (that is, `ClassificationLinear` or `ClassificationKernel` model objects). To compare `ClassificationECOC` models composed of linear or kernel classification models, use `testcholdout` instead.
- Similarly, `compareHoldout` does not compare `ClassificationLinear` or `ClassificationKernel` model objects. To compare these models, use `testcholdout` instead.

More About

Cost-Sensitive Testing

Conduct cost-sensitive testing when the cost of misclassification is imbalanced. By conducting a cost-sensitive analysis, you can account for the cost imbalance when you train the classification models and when you statistically compare them.

If the cost of misclassification is imbalanced, then the misclassification rate tends to be a poorly performing classification loss. Use misclassification cost instead to compare classification models.

Misclassification costs are often imbalanced in applications. For example, consider classifying subjects based on a set of predictors into two categories: healthy and sick. Misclassifying a sick subject as healthy poses a danger to the subject's life. However, misclassifying a healthy subject as sick typically causes some inconvenience, but does not pose significant danger. In this situation, you assign misclassification costs such that misclassifying a sick subject as healthy is more costly than misclassifying a healthy subject as sick.

The definitions that follow summarize the cost-sensitive tests. In the definitions:

- n_{ijk} and $\widehat{\pi}_{ijk}$ are the number and estimated proportion of test-sample observations with the following characteristics. k is the true class, i is the label assigned by the first classification model, and j is the label assigned by the second classification model. The unknown true value of $\widehat{\pi}_{ijk}$ is π_{ijk}. The test-set sample size is $\sum_{i,j,k} n_{ijk} = n_{test}$. Additionally, $\sum_{i,j,k} \pi_{ijk} = \sum_{i,j,k} \widehat{\pi}_{ijk} = 1$.
- c_{ij} is the relative cost of assigning label j to an observation with true class i. $c_{ii} = 0$, $c_{ij} \geq 0$, and, for at least one (i,j) pair, $c_{ij} > 0$.
- All subscripts take on integer values from 1 through K, which is the number of classes.
- The expected difference in the misclassification costs of the two classification models is

$$\delta = \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \left(c_{ki} - c_{kj}\right) \pi_{ijk} .$$

- The hypothesis test is

  $$H_0: \delta = 0$$
  $$H_1: \delta \neq 0.$$

  The available cost-sensitive tests are appropriate for two-tailed testing.

Available asymptotic tests that address imbalanced costs are a chi-square test and a likelihood ratio test.

- Chi-square test — The chi-square test statistic is based on the Pearson and Neyman chi-square test statistics, but with a Laplace correction factor to account for any $n_{ijk} = 0$. The test statistic is

  $$t^*_{\chi^2} = \sum_{i \neq j} \sum_{k} \frac{\left(n_{ijk} + 1 - \left(n_{test} + K^3\right)\widehat{\pi}^{(1)}_{ijk}\right)^2}{n_{ijk} + 1}.$$

  If $1 - F_{\chi^2}\left(t^*_{\chi^2}; 1\right) < \alpha$, then reject H_0.

  - $\widehat{\pi}^{(1)}_{ijk}$ are estimated by minimizing $t^*_{\chi^2}$ under the constraint that $\delta = 0$.
  - $F_{\chi^2}(x; 1)$ is the χ^2 cdf with one degree of freedom evaluated at x.
- Likelihood ratio test — The likelihood ratio test is based on N_{ijk}, which are binomial random variables with sample size n_{test} and success probability π_{ijk}. The random variables represent the random number of observations with: true class k, label i assigned by the first classification model, and label j assigned by the second classification model. Jointly, the distribution of the random variables is multinomial.

  The test statistic is

  $$t^*_{LRT} = 2\log\left[\frac{P\left(\bigcap_{i,j,k} N_{ijk} = n_{ijk}; n_{test}, \widehat{\pi}_{ijk} = \widehat{\pi}^{(2)}_{ijk}\right)}{P\left(\bigcap_{i,j,k} N_{ijk} = n_{ijk}; n_{test}, \widehat{\pi}_{ijk} = \widehat{\pi}^{(3)}_{ijk}\right)}\right].$$

  If $1 - F_{\chi^2}\left(t^*_{LRT}; 1\right) < \alpha$, then reject H_0.

  - $\widehat{\pi}^{(2)}_{ijk} = \frac{n_{ijk}}{n_{test}}$ is the unrestricted MLE of π_{ijk}.
  - $\widehat{\pi}^{(3)}_{ijk} = \frac{n_{ijk}}{n_{test} + \lambda(c_{ki} - c_{kj})}$ is the MLE under the null hypothesis that $\delta = 0$. λ is the solution to

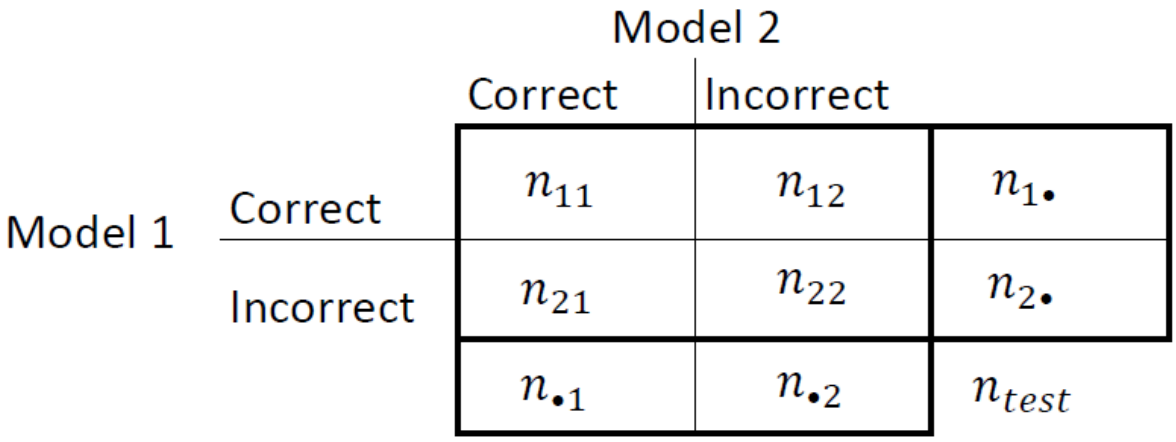
    $$\sum_{i,j,k} \frac{n_{ijk}(c_{ki} - c_{kj})}{n_{test} + \lambda(c_{ki} - c_{kj})} = 0.$$
  - $F_{\chi^2}(x; 1)$ is the χ^2 cdf with one degree of freedom evaluated at x.

McNemar Tests

McNemar Tests are hypothesis tests that compare two population proportions while addressing the issues resulting from two dependent, matched-pair samples.

One way to compare the predictive accuracies of two classification models is:

1. Partition the data into training and test sets.
2. Train both classification models using the training set.
3. Predict class labels using the test set.
4. Summarize the results in a two-by-two table similar to this figure.



n_{ii} are the number of concordant pairs, that is, the number of observations that both models classify the same way (correctly or incorrectly). n_{ij}, $i \neq j$, are the number of discordant pairs, that is, the number of observations that models classify differently (correctly or incorrectly).

The misclassification rates for Models 1 and 2 are $\widehat{\pi}_{2\bullet} = n_{2\bullet}/n$ and $\widehat{\pi}_{\bullet 2} = n_{\bullet 2}/n$, respectively. A two-sided test for comparing the accuracy of the two models is

$$H_0: \pi_{\bullet 2} = \pi_{2\bullet}$$
$$H_1: \pi_{\bullet 2} \neq \pi_{2\bullet}.$$

The null hypothesis suggests that the population exhibits marginal homogeneity, which reduces the null hypothesis to $H_0: \pi_{12} = \pi_{21}$. Also, under the null hypothesis, $N_{12} \sim \text{Binomial}(n_{12} + n_{21}, 0.5)$ [1].

These facts are the basis for the available McNemar test variants: the asymptotic, exact-conditional, and mid-p-value McNemar tests. The definitions that follow summarize the available variants.

- Asymptotic — The asymptotic McNemar test statistics and rejection regions (for significance level α) are:
  - For one-sided tests, the test statistic is

$$t^*_{a1} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

    If $1 - \Phi\left(\left|t^*_1\right|\right) < \alpha$, where Φ is the standard Gaussian cdf, then reject H_0.
  - For two-sided tests, the test statistic is

$$t_{a2}^{*} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

If $1 - F_{\chi^2}(t_2^{*}; m) < \alpha$, where $F_{\chi^2}(x; m)$ is the χ_m^2 cdf evaluated at x, then reject H_0.

The asymptotic test requires large-sample theory, specifically, the Gaussian approximation to the binomial distribution.

- The total number of discordant pairs, $n_d = n_{12} + n_{21}$, must be greater than 10 ([1], Ch. 10.1.4).
- In general, asymptotic tests do not guarantee nominal coverage. The observed probability of falsely rejecting the null hypothesis can exceed α, as suggested in simulation studies in [18]. However, the asymptotic McNemar test performs well in terms of statistical power.

- Exact-Conditional — The exact-conditional McNemar test statistics and rejection regions (for significance level α) are ([36], [38]):

  - For one-sided tests, the test statistic is

  $$t_1^{*} = n_{12}.$$

  If $F_{\text{Bin}}(t_1^{*}; n_d, 0.5) < \alpha$, where $F_{\text{Bin}}(x; n, p)$ is the binomial cdf with sample size n and success probability p evaluated at x, then reject H_0.
  - For two-sided tests, the test statistic is

  $$t_2^{*} = \min(n_{12}, n_{21}).$$

  If $F_{\text{Bin}}(t_2^{*}; n_d, 0.5) < \alpha/2$, then reject H_0.

  The exact-conditional test always attains nominal coverage. Simulation studies in [18] suggest that the test is conservative, and then show that the test lacks statistical power compared to other variants. For small or highly discrete test samples, consider using the mid-p-value test ([1], Ch. 3.6.3).
- Mid-p-value test — The mid-p-value McNemar test statistics and rejection regions (for significance level α) are ([32]):

  - For one-sided tests, the test statistic is

  $$t_1^{*} = n_{12}.$$

  If $F_{\text{Bin}}(t_1^{*} - 1; n_{12} + n_{21}, 0.5) + 0.5 f_{\text{Bin}}(t_1^{*}; n_{12} + n_{21}, 0.5) < \alpha$, where $F_{\text{Bin}}(x; n, p)$ and $f_{\text{Bin}}(x; n, p)$ are the binomial cdf and pdf, respectively, with sample size n and success probability p evaluated at x, then reject H_0.
  - For two-sided tests, the test statistic is

  $$t_2^{*} = \min(n_{12}, n_{21}).$$

  If $F_{\text{Bin}}(t_2^{*} - 1; n_{12} + n_{21} - 1, 0.5) + 0.5 f_{\text{Bin}}(t_2^{*}; n_{12} + n_{21}, 0.5) < \alpha/2$, then reject H_0.

  The mid-p-value test addresses the over-conservative behavior of the exact-conditional test. The simulation studies in [18] demonstrate that this test attains nominal coverage, and has good statistical power.

Classification Loss

Classification losses indicate the accuracy of a classification model or set of predicted labels. Two classification losses are the misclassification rate and cost.

`compareHoldout` returns the classification losses (see `e1` and `e2`) under the alternative hypothesis (that is, the unrestricted classification losses). n_{ijk} is the number of test-sample observations with: true class k, label i assigned by the first classification model, and label j assigned by the second classification model. The corresponding estimated proportion is $\widehat{\pi}_{ijk} = \frac{n_{ijk}}{n_{test}}$. The test-set sample size is $\sum_{i,j,k} n_{ijk} = n_{test}$. The indices are taken from 1 through K, the number of classes.

- The misclassification rate, or classification error, is a scalar in the interval [0,1] representing the proportion of misclassified observations. That is, the misclassification rate for the first classification model is

  $$e_1 = \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{i \neq k} \widehat{\pi}_{ijk}.$$

  For the misclassification rate of the second classification model (e_2), switch the indices i and j in the formula.

  Classification accuracy decreases as the misclassification rate increases to 1.
- The misclassification cost is a nonnegative scalar that is a measure of classification quality relative to the values of the specified cost matrix. Its interpretation depends on the specified costs of misclassification. The misclassification cost is the weighted average of the costs of misclassification (specified in a cost matrix, C) in which the weights are the respective estimated proportions of misclassified observations. The misclassification cost for the first classification model is

  $$e_1 = \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{i \neq k} \widehat{\pi}_{ijk} c_{ki},$$

  where c_{kj} is the cost of classifying an observation into class j if its true class is k. For the misclassification cost of the second classification model (e_2), switch the indices i and j in the formula.

  In general, for a fixed cost matrix, classification accuracy decreases as the misclassification cost increases.

Tips

- One way to perform cost-insensitive feature selection is:

  1. Train the first classification model (`C1`) using the full predictor set.
  2. Train the second classification model (`C2`) using the reduced predictor set.
  3. Specify `X1` as the full test-set predictor data and `X2` as the reduced test-set predictor data.
  4. Enter `compareHoldout(C1,C2,X1,X2,Y,'Alternative','less')`. If `compareHoldout` returns `1`, then there is enough evidence to suggest that the classification model that uses fewer predictors performs better than the model that uses the full predictor set.

Alternatively, you can assess whether there is a significant difference between the accuracies of the two models. To perform this assessment, remove the `'Alternative','less'` specification in step 4. `compareHoldout` conducts a two-sided test, and `h = 0` indicates that there is not enough evidence to suggest a difference in the accuracy of the two models.

- Cost-sensitive tests perform numerical optimization, which requires additional computational resources. The likelihood ratio test conducts numerical optimization indirectly by finding the root of a Lagrange multiplier in an interval. For some data sets, if the root lies close to the boundaries of the interval, then the method can fail. Therefore, if you have an Optimization Toolbox license, consider conducting the cost-sensitive chi-square test instead. For more details, see `CostTest` and "Cost-Sensitive Testing" on page 33-908.

Alternative Functionality

To directly compare the accuracy of two sets of class labels in predicting a set of true class labels, use `testcholdout`.

References

[1] Agresti, A. *Categorical Data Analysis*, 2nd Ed. John Wiley & Sons, Inc.: Hoboken, NJ, 2002.

[2] Fagerlan, M.W., S. Lydersen, and P. Laake. "The McNemar Test for Binary Matched-Pairs Data: Mid-p and Asymptotic Are Better Than Exact Conditional." *BMC Medical Research Methodology*. Vol. 13, 2013, pp. 1–8.

[3] Lancaster, H.O. "Significance Tests in Discrete Distributions." *JASA*, Vol. 56, Number 294, 1961, pp. 223–234.

[4] McNemar, Q. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages." *Psychometrika*, Vol. 12, Number 2, 1947, pp. 153–157.

[5] Mosteller, F. "Some Statistical Problems in Measuring the Subjective Response to Drugs." *Biometrics*, Vol. 8, Number 3, 1952, pp. 220–226.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function fully supports GPU arrays for a trained classification model specified as a `ClassificationKNN`, `ClassificationSVM`, or `CompactClassificationSVM` object.
- This function supports `ClassificationKNN`, `ClassificationSVM`, and `CompactClassificationSVM` objects fitted with GPU array input arguments.
- `compareHoldout` executes on a GPU in these cases only:
  - Either or both of the input arguments `X1` and `X2` are GPU arrays.
  - Either or both of the input arguments `T1` and `T2` contain `gpuArray` elements.
  - Either or both of the input arguments `C1` and `C2` were fitted with GPU array input arguments.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`testcholdout` | `testckfold`

Topics

“Hypothesis Tests”

Introduced in R2015a

confusionchart

Create confusion matrix chart for classification problem

Syntax

```
confusionchart(trueLabels,predictedLabels)
confusionchart(m)
confusionchart(m,classLabels)
confusionchart(parent, ___ )
confusionchart( ___ ,Name,Value)
cm = confusionchart( ___ )
```

Description

`confusionchart(trueLabels,predictedLabels)` creates a confusion matrix chart from true labels `trueLabels` and predicted labels `predictedLabels` and returns a `ConfusionMatrixChart` object. The rows of the confusion matrix correspond to the true class and the columns correspond to the predicted class. Diagonal and off-diagonal cells correspond to correctly and incorrectly classified observations, respectively. Use `cm` to modify the confusion matrix chart after it is created. For a list of properties, see ConfusionMatrixChart Properties.

`confusionchart(m)` creates a confusion matrix chart from the numeric confusion matrix `m`. Use this syntax if you already have a numeric confusion matrix in the workspace.

`confusionchart(m,classLabels)` specifies class labels that appear along the *x*-axis and *y*-axis. Use this syntax if you already have a numeric confusion matrix and class labels in the workspace.

`confusionchart(parent, ___ )` creates the confusion chart in the figure, panel, or tab specified by `parent`.

`confusionchart( ___ ,Name,Value)` specifies additional `ConfusionMatrixChart` properties using one or more name-value pair arguments. Specify the properties after all other input arguments. For a list of properties, see ConfusionMatrixChart Properties.

`cm = confusionchart( ___ )` returns the `ConfusionMatrixChart` object. Use `cm` to modify properties of the chart after creating it. For a list of properties, see ConfusionMatrixChart Properties.

Examples

Create Confusion Matrix Chart

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

`X` is a numeric matrix that contains four petal measurements for 150 irises. `Y` is a cell array of character vectors that contains the corresponding iris species.

Train a *k*-nearest neighbor (KNN) classifier, where the number of nearest neighbors in the predictors (*k*) is 5. A good practice is to standardize numeric predictor data.

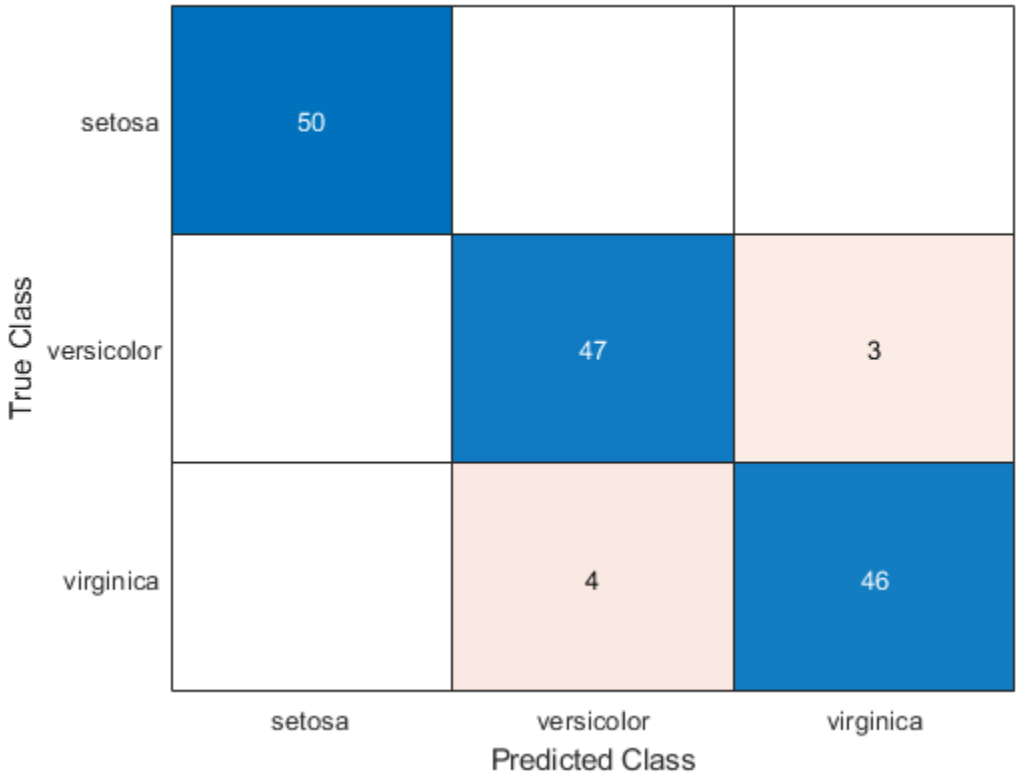
```
Mdl = fitcknn(X,Y,'NumNeighbors',5,'Standardize',1);
```

Predict the labels of the training data.

```
predictedY = resubPredict(Mdl);
```

Create a confusion matrix chart from the true labels `Y` and the predicted labels `predictedY`.

```
cm = confusionchart(Y,predictedY);
```



The confusion matrix displays the total number of observations in each cell. The rows of the confusion matrix correspond to the true class, and the columns correspond to the predicted class. Diagonal and off-diagonal cells correspond to correctly and incorrectly classified observations, respectively.

By default, `confusionchart` sorts the classes into their natural order as defined by `sort`. In this example, the class labels are character vectors, so `confusionchart` sorts the classes alphabetically. Use `sortClasses` to sort the classes by a specified order or by the confusion matrix values.

The `NormalizedValues` property contains the values of the confusion matrix. Display these values using dot notation.

```
cm.NormalizedValues
```

```
ans = 3×3
```

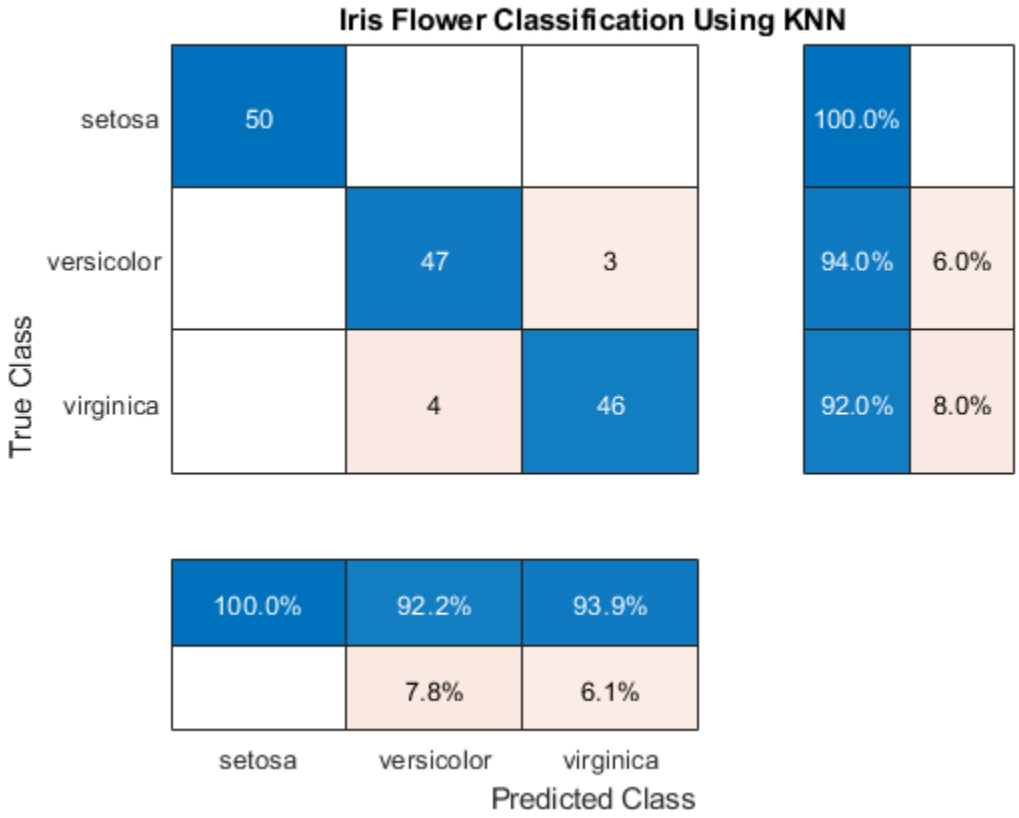
```
    50     0     0
     0    47     3
     0     4    46
```

Modify the appearance and behavior of the confusion matrix chart by changing property values. Add a title.

```
cm.Title = 'Iris Flower Classification Using KNN';
```

Add column and row summaries.

```
cm.RowSummary = 'row-normalized';
cm.ColumnSummary = 'column-normalized';
```



A row-normalized row summary displays the percentages of correctly and incorrectly classified observations for each true class. A column-normalized column summary displays the percentages of correctly and incorrectly classified observations for each predicted class.

Sort Classes by Precision or Recall

Create a confusion matrix chart and sort the classes of the chart according to the class-wise true positive rate (recall) or the class-wise positive predictive value (precision).

Load and inspect the `arrhythmia` data set.

```
load arrhythmia
isLabels = unique(Y);
nLabels = numel(isLabels)

nLabels = 13

tabulate(categorical(Y))

  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

The data contains 16 distinct labels that describe various degrees of arrhythmia, but the response (`Y`) includes only 13 distinct labels.

Train a classification tree and predict the resubstitution response of the tree.

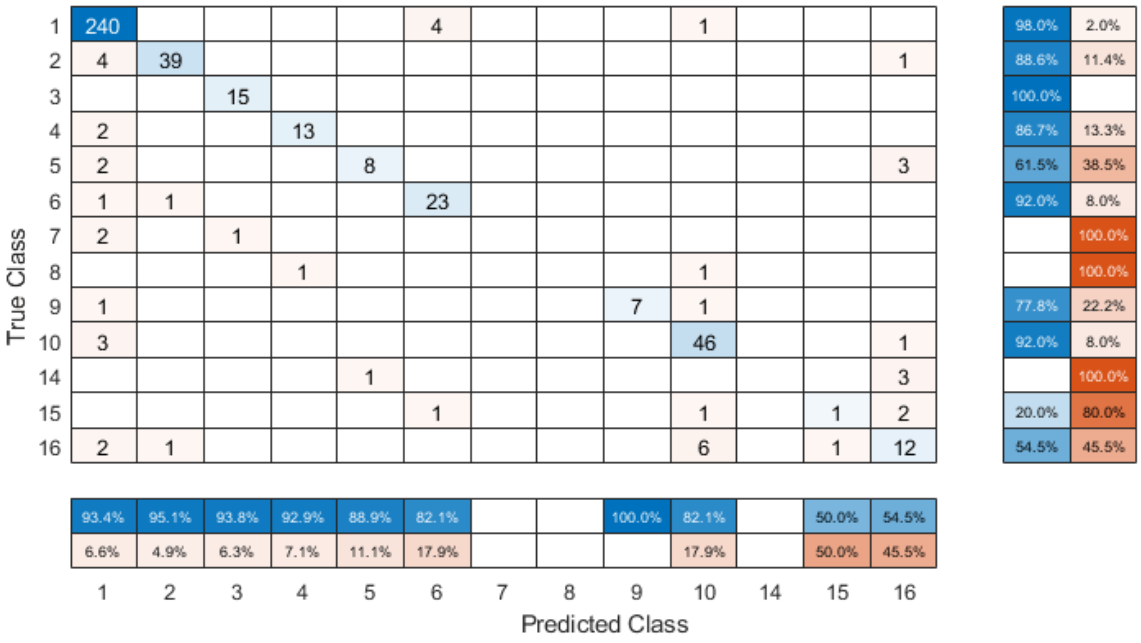
```
Mdl = fitctree(X,Y);
predictedY = resubPredict(Mdl);
```

Create a confusion matrix chart from the true labels `Y` and the predicted labels `predictedY`. Specify `'RowSummary'` as `'row-normalized'` to display the true positive rates and false positive rates in the row summary. Also, specify `'ColumnSummary'` as `'column-normalized'` to display the positive predictive values and false discovery rates in the column summary.

```
fig = figure;
cm = confusionchart(Y,predictedY,'RowSummary','row-normalized','ColumnSummary','column-normalized
```

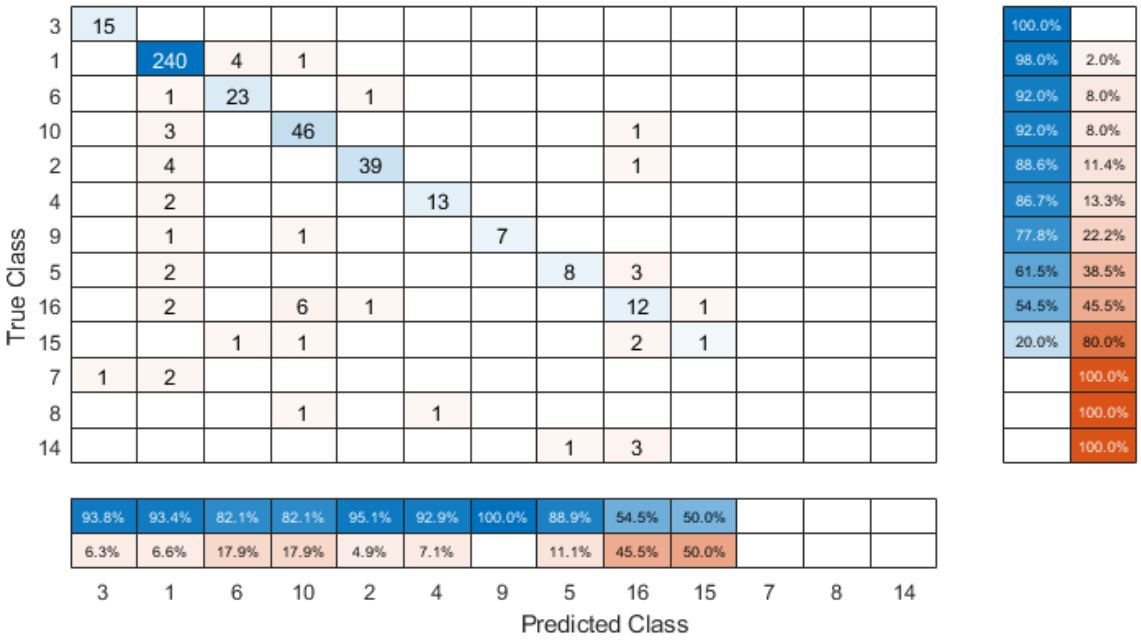
Resize the container of the confusion chart so percentages appear in the row summary.

```
fig_Position = fig.Position;
fig_Position(3) = fig_Position(3)*1.5;
fig.Position = fig_Position;
```



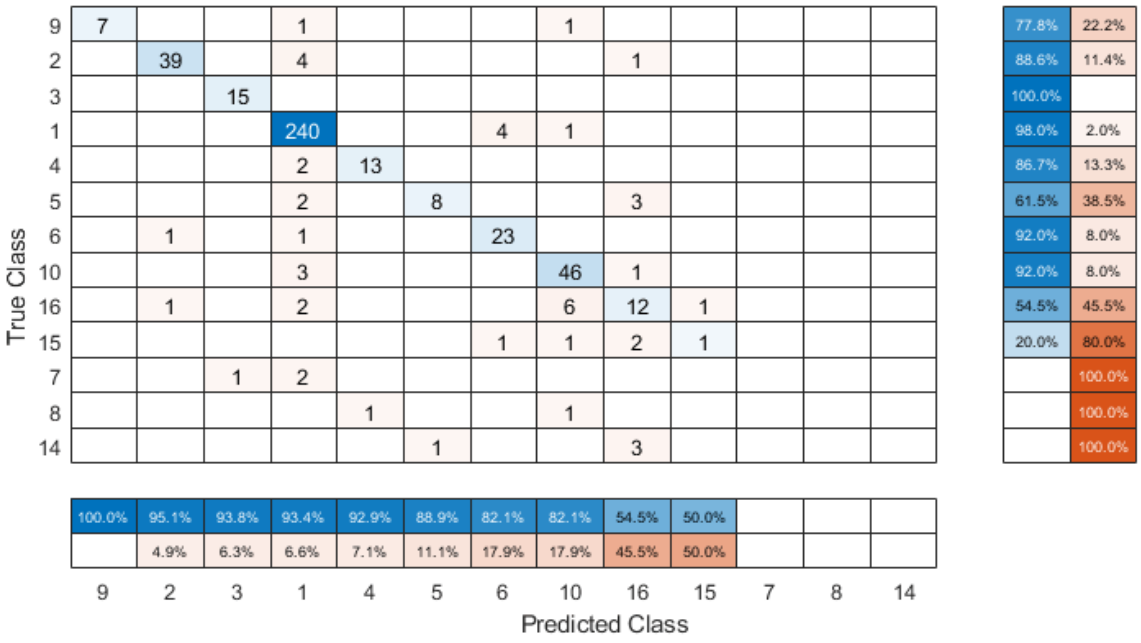
To sort the confusion matrix according to the true positive rate, normalize the cell values across each row by setting the `Normalization` property to `'row-normalized'` and then use `sortClasses`. After sorting, reset the `Normalization` property back to `'absolute'` to display the total number of observations in each cell.

```
cm.Normalization = 'row-normalized';
sortClasses(cm,'descending-diagonal')
cm.Normalization = 'absolute';
```



To sort the confusion matrix according to the positive predictive value, normalize the cell values across each column by setting the `Normalization` property to `'column-normalized'` and then use `sortClasses`. After sorting, reset the `Normalization` property back to `'absolute'` to display the total number of observations in each cell.

```
cm.Normalization = 'column-normalized';
sortClasses(cm,'descending-diagonal')
cm.Normalization = 'absolute';
```



Confusion Matrix for Classification Using Tall Arrays

Perform classification on a tall array of the Fisher iris data set. Compute a confusion matrix chart for the known and predicted tall labels by using the `confusionchart` function.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Load Fisher's iris data set.

```
load fisheriris
```

Convert the in-memory arrays `meas` and `species` to tall arrays.

```
tx = tall(meas);
ty = tall(species);
```

Find the number of observations in the tall array.

```
numObs = gather(length(ty));   % gather collects tall array into memory
```

Set the seeds of the random number generators using `rng` and `tallrng` for reproducibility, and randomly select training samples. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
```

```
numTrain = floor(numObs/2);
[txTrain,trIdx] = datasample(tx,numTrain,'Replace',false);
tyTrain = ty(trIdx);
```

Fit a decision tree classifier model on the training samples.

```
mdl = fitctree(txTrain,tyTrain);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 2: Completed in 1.3 sec
- Pass 2 of 2: Completed in 1 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 0.68 sec
- Pass 2 of 4: Completed in 0.83 sec
- Pass 3 of 4: Completed in 0.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 0.31 sec
- Pass 2 of 4: Completed in 0.28 sec
- Pass 3 of 4: Completed in 0.36 sec
- Pass 4 of 4: Completed in 0.42 sec
Evaluation completed in 1.8 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 0.28 sec
- Pass 2 of 4: Completed in 0.29 sec
- Pass 3 of 4: Completed in 0.31 sec
- Pass 4 of 4: Completed in 0.62 sec
Evaluation completed in 1.8 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 0.33 sec
- Pass 2 of 4: Completed in 0.27 sec
- Pass 3 of 4: Completed in 0.31 sec
- Pass 4 of 4: Completed in 0.31 sec
Evaluation completed in 1.6 sec
```

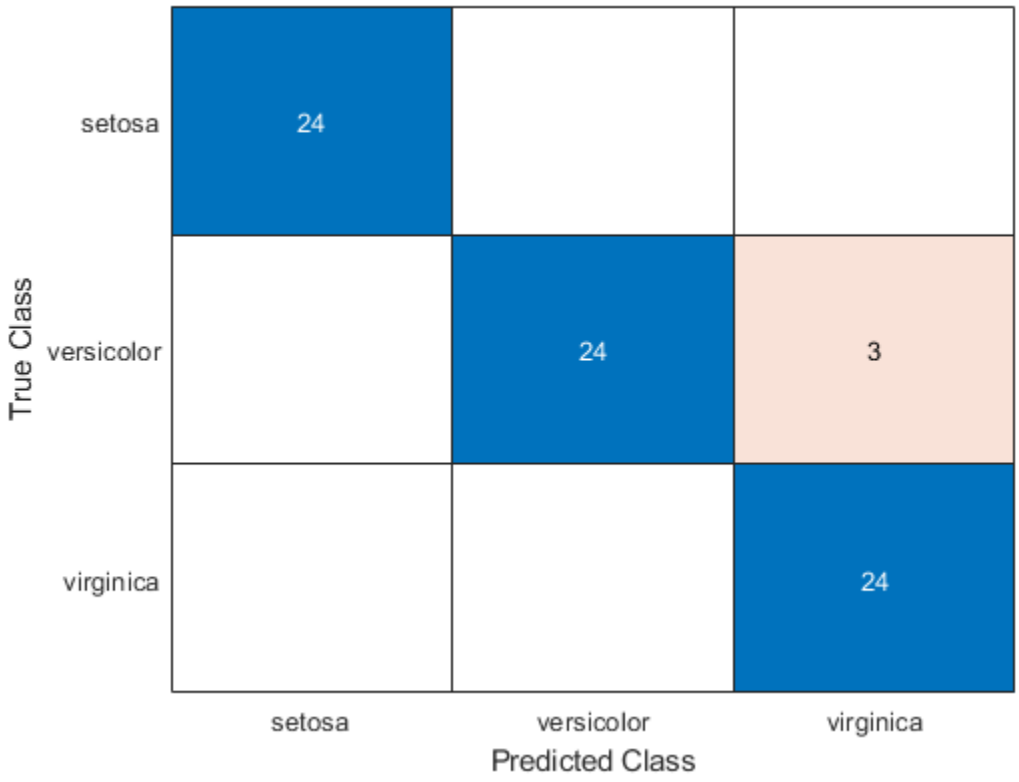
Predict labels for the test samples by using the trained model.

```
txTest = tx(~trIdx,:);
label = predict(mdl,txTest);
```

Create the confusion matrix chart for the resulting classification.

```
tyTest = ty(~trIdx);
cm = confusionchart(tyTest,label)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.2 sec
Evaluation completed in 0.47 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.27 sec
Evaluation completed in 0.49 sec
```




```
cm =
  ConfusionMatrixChart with properties:

    NormalizedValues: [3x3 double]
         ClassLabels: {3x1 cell}

  Show all properties
```

The confusion matrix chart shows that three measurements in the versicolor class are misclassified. All the measurements belonging to setosa and virginica are classified correctly.

Input Arguments

`trueLabels` — True labels of classification problem

categorical vector | numeric vector | string vector | character array | cell array of character vectors | logical vector

True labels of classification problem, specified as a categorical vector, numeric vector, string vector, character array, cell array of character vectors, or logical vector. If `trueLabels` is a vector, then each element corresponds to one observation. If `trueLabels` is a character array, then it must be two-dimensional with each row corresponding to the label of one observation.

`predictedLabels` — Predicted labels of classification problem
categorical vector | numeric vector | string vector | character array | cell array of character vectors | logical vector

Predicted labels of classification problem, specified as a categorical vector, numeric vector, string vector, character array, cell array of character vectors, or logical vector. If `predictedLabels` is a vector, then each element corresponds to one observation. If `predictedLabels` is a character array, then it must be two-dimensional with each row corresponding to the label of one observation.

`m` — Confusion matrix
matrix

Confusion matrix, specified as a matrix. `m` must be square and its elements must be positive integers. The element `m(i,j)` is the number of times an observation of the `i`th true class was predicted to be of the `j`th class. Each colored cell of the confusion matrix chart corresponds to one element of the confusion matrix `m`.

`classLabels` — Class labels
categorical vector | numeric vector | string vector | character array | cell array of character vectors | logical vector

Class labels of the confusion matrix chart, specified as a categorical vector, numeric vector, string vector, character array, cell array of character vectors, or logical vector. If `classLabels` is a vector, then it must have the same number of elements as the confusion matrix has rows and columns. If `classLabels` is a character array, then it must be two-dimensional with each row corresponding to the label of one class.

`parent` — Parent container
`Figure` object | `Panel` object | `Tab` object | `TiledChartLayout` object | `GridLayout` object

Parent container, specified as a `Figure`, `Panel`, `Tab`, `TiledChartLayout`, or `GridLayout` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `cm = confusionchart(trueLabels,predictedLabels,'Title','My Title Text','ColumnSummary','column-normalized')`

Note The properties listed here are only a subset. For a complete list, see ConfusionMatrixChart Properties.

`Title` — Title
`''` (default) | character vector | string scalar

Title of the confusion matrix chart, specified as a character vector or string scalar.

Example: `cm = confusionchart(__,'Title','My Title Text')`

Example: `cm.Title = 'My Title Text'`

`ColumnSummary` — Column summary
`'off'` (default) | `'absolute'` | `'column-normalized'` | `'total-normalized'`

Column summary of the confusion matrix chart, specified as one of the following:

Option	Description
`'off'`	Do not display a column summary.
`'absolute'`	Display the total number of correctly and incorrectly classified observations for each predicted class.
`'column-normalized'`	Display the number of correctly and incorrectly classified observations for each predicted class as percentages of the number of observations of the corresponding predicted class. The percentages of correctly classified observations can be thought of as class-wise precisions (or positive predictive values).
`'total-normalized'`	Display the number of correctly and incorrectly classified observations for each predicted class as percentages of the total number of observations.

Example: `cm = confusionchart(__,'ColumnSummary','column-normalized')`

Example: `cm.ColumnSummary = 'column-normalized'`

`RowSummary` — Row summary

`'off'` (default) | `'absolute'` | `'row-normalized'` | `'total-normalized'`

Row summary of the confusion matrix chart, specified as one of the following:

Option	Description
`'off'`	Do not display a row summary.
`'absolute'`	Display the total number of correctly and incorrectly classified observations for each true class.
`'row-normalized'`	Display the number of correctly and incorrectly classified observations for each true class as percentages of the number of observations of the corresponding true class. The percentages of correctly classified observations can be thought of as class-wise recalls (or true positive rates).
`'total-normalized'`	Display the number of correctly and incorrectly classified observations for each true class as percentages of the total number of observations.

Example: `cm = confusionchart(__,'RowSummary','row-normalized')`

Example: `cm.RowSummary = 'row-normalized'`

`Normalization` — Normalization of cell values

`'absolute'` (default) | `'column-normalized'` | `'row-normalized'` | `'total-normalized'`

Normalization of cell values, specified as one of the following:

Option	Description
`'absolute'`	Display the total number of observations in each cell.
`'column-normalized'`	Normalize each cell value by the number of observations that has the same predicted class.
`'row-normalized'`	Normalize each cell value by the number of observations that has the same true class.
`'total-normalized'`	Normalize each cell value by the total number of observations.

Modifying the normalization of cell values also affects the colors of the cells.

Example: `cm = confusionchart(__,'Normalization','total-normalized')`

Example: `cm.Normalization = 'total-normalized'`

Output Arguments

`cm` — Confusion matrix chart object

`ConfusionMatrixChart` object

`ConfusionMatrixChart` object, which is a standalone visualization on page 33-926. Use `cm` to set properties of the confusion matrix chart after creating it.

Limitations

- MATLAB code generation is not supported for `ConfusionMatrixChart` objects.

More About

Standalone Visualization

A standalone visualization is a chart designed for a special purpose that works independently from other charts. Unlike other charts such as `plot` and `surf`, a standalone visualization has a preconfigured axes object built into it, and some customizations are not available. A standalone visualization also has these characteristics:

- It cannot be combined with other graphics elements, such as lines, patches, or surfaces. Thus, the `hold` command is not supported.
- The `gca` function can return the chart object as the current axes.
- You can pass the chart object to many MATLAB functions that accept an axes object as an input argument. For example, you can pass the chart object to the `title` function.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

Functions
`categorical` | `confusionmat` | `sortClasses`

Properties
ConfusionMatrixChart Properties

Introduced in R2018b

ConfusionMatrixChart Properties

Confusion matrix chart appearance and behavior

Description

`ConfusionMatrixChart` properties control the appearance and behavior of a `ConfusionMatrixChart` object. By changing property values, you can modify certain aspects of the confusion matrix chart. For example, you can add a title:

```
cm = confusionchart([1 3 5; 2 4 6; 11 7 3]);
cm.Title = 'My Confusion Matrix Title';
```

Properties

Labels

`Title` — Title
`''` (default) | character vector | string scalar

Title of the confusion matrix chart, specified as a character vector or string scalar.

Example: `cm = confusionchart(__,'Title','My Title Text')`

Example: `cm.Title = 'My Title Text'`

`XLabel` — Label for *x*-axis
`'Predicted class'` (default) | string scalar | character vector

Label for the *x*-axis, specified as a string scalar or character vector.

Example: `cm = confusionchart(__,'XLabel','My Label')`

Example: `cm.XLabel = 'My Label'`

`YLabel` — Label for *y*-axis
`'True class'` (default) | string scalar | character vector

Label for the *x*-axis, specified as a string scalar or character vector.

Example: `cm = confusionchart(__,'YLabel','My Label')`

Example: `cm.YLabel = 'My Label'`

`ClassLabels` — Class labels
categorical vector | numeric vector | string vector | character array | cell array of character vectors | logical vector

This property is read-only.

Class labels of the confusion matrix chart, stored as a categorical vector, numeric vector, string vector, character array, cell array of character vectors, or logical vector.

Row and Column Summaries

ColumnSummary — Column summary

`'off'` (default) | `'absolute'` | `'column-normalized'` | `'total-normalized'`

Column summary of the confusion matrix chart, specified as one of the following:

Option	Description
`'off'`	Do not display a column summary.
`'absolute'`	Display the total number of correctly and incorrectly classified observations for each predicted class.
`'column-normalized'`	Display the number of correctly and incorrectly classified observations for each predicted class as percentages of the number of observations of the corresponding predicted class. The percentages of correctly classified observations can be thought of as class-wise precisions (or positive predictive values).
`'total-normalized'`	Display the number of correctly and incorrectly classified observations for each predicted class as percentages of the total number of observations.

Example: `cm = confusionchart(__,'ColumnSummary','column-normalized')`

Example: `cm.ColumnSummary = 'column-normalized'`

RowSummary — Row summary

`'off'` (default) | `'absolute'` | `'row-normalized'` | `'total-normalized'`

Row summary of the confusion matrix chart, specified as one of the following:

Option	Description
`'off'`	Do not display a row summary.
`'absolute'`	Display the total number of correctly and incorrectly classified observations for each true class.
`'row-normalized'`	Display the number of correctly and incorrectly classified observations for each true class as percentages of the number of observations of the corresponding true class. The percentages of correctly classified observations can be thought of as class-wise recalls (or true positive rates).
`'total-normalized'`	Display the number of correctly and incorrectly classified observations for each true class as percentages of the total number of observations.

Example: `cm = confusionchart(__,'RowSummary','row-normalized')`

Example: `cm.RowSummary = 'row-normalized'`

Data

NormalizedValues — Values of the confusion matrix

numeric matrix

This property is read-only.

Values of the confusion matrix, stored as a numeric matrix. This property equals the values of the confusion matrix normalized using the method of the `Normalization` property. The software recalculates the normalized values of the confusion matrix each time you modify the `Normalization` property.

Normalization — Normalization of cell values

`'absolute'` (default) | `'column-normalized'` | `'row-normalized'` | `'total-normalized'`

Normalization of cell values, specified as one of the following:

Option	Description
`'absolute'`	Display the total number of observations in each cell.
`'column-normalized'`	Normalize each cell value by the number of observations that has the same predicted class.
`'row-normalized'`	Normalize each cell value by the number of observations that has the same true class.
`'total-normalized'`	Normalize each cell value by the total number of observations.

Modifying the normalization of cell values also affects the colors of the cells.

Example: `cm = confusionchart(__,'Normalization','total-normalized')`

Example: `cm.Normalization = 'total-normalized'`

Color and Styling

GridVisible — State of grid visibility

`'on'` (default) | on/off logical value

State of grid visibility, specified as `'on'` or `'off'`, or as numeric or logical `1` (`true`) or `0` (`false`). A value of `'on'` is equivalent to true, and `'off'` is equivalent to `false`. Thus, you can use the value of this property as a logical value. The value is stored as an on/off logical value of type `matlab.lang.OnOffSwitchState`.

- `'on'` — Display grid lines between the chart cells.
- `'off'` — Do not display grid lines between the chart cells.

Example: `cm = confusionchart(__,'GridVisible','off')`

Example: `cm.GridVisible = 'off'`

DiagonalColor — Color for diagonal cells

`[0 0.4471 0.7412]` (default) | RGB triplet | hexadecimal color code | `'r'` | `'g'` | `'b'` | ...

Color for diagonal cells, specified as an RGB triplet, a hexadecimal color code, a color name, or a short name. The color of each diagonal cell is proportional to the cell value and the `DiagonalColor`

property, normalized to the largest cell value of the confusion matrix chart. Cells with positive values are colored with a minimum amount of color, proportional to the `DiagonalColor` property.

RGB triplets and hexadecimal color codes are useful for specifying custom colors.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

The software chooses an appropriate text color for cell labels automatically, depending on the color of the chart cells.

Example: `cm = confusionchart(__,'DiagonalColor','blue')`

Example: `cm.DiagonalColor = 'blue'`

`OffDiagonalColor` — Color for off-diagonal cells

`[0.8510 0.3255 0.0980]` (default) | RGB triplet | hexadecimal color code | `'r'` | `'g'` | `'b'` | ...

Color for off-diagonal cells, specified as an RGB triplet, a hexadecimal color code, a color name, or a short name. The color of each diagonal cell is proportional to the cell value and the `OffDiagonalColor` property, normalized to the largest cell value of the confusion matrix chart. Cells with positive values are colored with a minimum amount of color, proportional to the `OffDiagonalColor` property.

RGB triplets and hexadecimal color codes are useful for specifying custom colors.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

The software chooses an appropriate text color for cell labels automatically, depending on the color of the chart cells.

Example: `cm = confusionchart(__,'OffDiagonalColor','blue')`

Example: `cm.OffDiagonalColor = 'blue'`

`FontColor` — Text color for title, axis labels, and class labels

`[0.1500 0.1500 0.1500]` (default) | RGB triplet | hexadecimal color code | `'r'` | `'g'` | `'b'` | ...

Text color for title, axis labels, and class labels, specified as an RGB triplet, a hexadecimal color code, a color name, or a short name.

RGB triplets and hexadecimal color codes are useful for specifying custom colors.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

The software chooses an appropriate text color for cell labels automatically, depending on the color of the chart cells.

Example: `cm = confusionchart(__,'FontColor','blue')`

Example: `cm.FontColor = 'blue'`

Font

FontName — Font name
system supported font name

Font name, specified as a system supported font name. The default font depends on the specific operating system and locale.

Example: `cm = confusionchart(__,'FontName','Cambria')`

Example: `cm.FontName = 'Cambria'`

FontSize — Font size
positive scalar

Font size used for the title, axis labels, class labels, and cell labels, specified as a positive scalar. The default font depends on the specific operating system and locale.

The title and axis labels use a slightly larger font size (scaled up by 10%). If there is not enough room to display the cell labels within the cells, then the cell labels use a smaller font size. If the cell labels become too small, then they are hidden.

Example: `cm = confusionchart(__,'FontSize',12)`

Example: `cm.FontSize = 12`

Position

PositionConstraint — Position to hold constant
`'outerposition'` | `'innerposition'`

Position property to hold constant when adding, removing, or changing decorations, specified as one of the following values:

- `'outerposition'` — The `OuterPosition` property remains constant when you add, remove, or change decorations such as a title or an axis label. If any positional adjustments are needed, MATLAB adjusts the `InnerPosition` property.
- `'innerposition'` — The `InnerPosition` property remains constant when you add, remove, or change decorations such as a title or an axis label. If any positional adjustments are needed, MATLAB adjusts the `OuterPosition` property.

Note Setting this property has no effect when the parent container is a `TiledChartLayout`.

OuterPosition — Outer size and position
`[0 0 1 1]` (default) | four-element vector

Outer size and position within the parent container (a figure, panel, or tab), specified as a four-element vector of the form `[left bottom width height]`. The outer position includes the title, axis labels, and class labels.

- The `left` and `bottom` elements define the distance from the lower left corner of the container to the lower left corner of the chart.
- The `width` and `height` elements are the chart dimensions, which include the chart cells, plus a margin for the surrounding text.

The default value of `[0 0 1 1]` is the whole interior of the container.

By default, the values are normalized to the container. To change the units, set the `Units` property.

Example: `cm = confusionchart(__,'OuterPosition',[0.1 0.1 0.8 0.8])`

Example: `cm.OuterPosition = [0.1 0.1 0.8 0.8]`

`InnerPosition` — Inner size and position

`[0.1300 0.1100 0.7750 0.8150]` (default) | four-element vector

Inner size and position of the chart within the parent container (a figure, panel, or tab) returned as a four-element vector of the form `[left bottom width height]`. The inner position does not include the title, axis labels, or class labels.

- The `left` and `bottom` elements define the distance from the lower left corner of the container to the lower left corner of the chart.
- The `width` and `height` elements are the chart dimensions, which include only the chart cells.

Example: `cm = confusionchart(__,'InnerPosition',[0.1 0.1 0.8 0.8])`

Example: `cm.InnerPosition = [0.1 0.1 0.8 0.8]`

`Position` — Inner size and position

four-element vector

Inner size and position of the chart within the parent container (a figure, panel, or tab) returned as a four-element vector of the form `[left bottom width height]`. This property is equivalent to the `InnerPosition` property.

`Units` — Position units

`'normalized'` (default) | `'inches'` | `'centimeters'` | `'points'` | `'pixels'` | `'characters'`

Position units, specified as one of these values:

`Units`	Description
`'normalized'`	Normalized with respect to the container, which is typically the figure or a panel. The lower left corner of the container maps to `(0,0)`, and the upper right corner maps to `(1,1)`.
`'inches'`	Inches.
`'centimeters'`	Centimeters.
`'characters'`	Based on the default uicontrol font of the graphics root object: • Character width = width of letter `x`. • Character height = distance between the baselines of two lines of text.
`'points'`	Typography points. One point equals 1/72 inch.

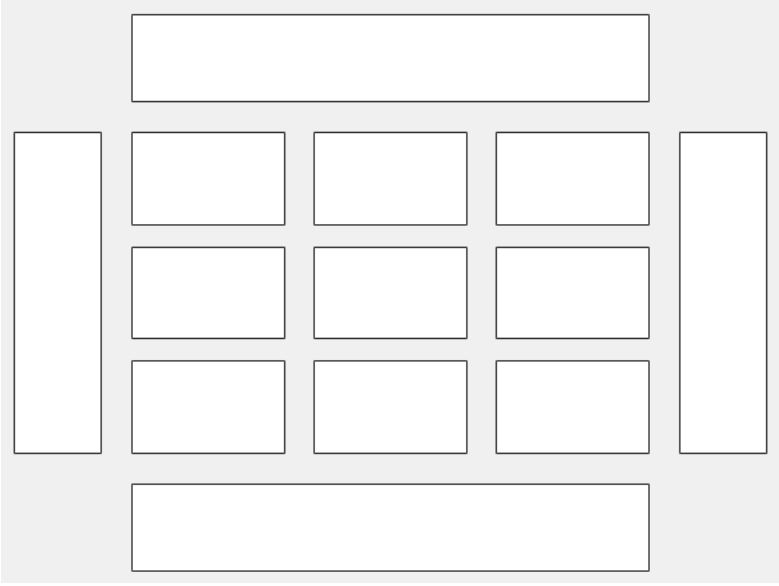
Units	Description
'pixels'	Pixels. Starting in R2015b, distances in pixels are independent of your system resolution on Windows® and Macintosh systems: • On Windows systems, a pixel is 1/96th of an inch. • On Macintosh systems, a pixel is 1/72nd of an inch. On Linux® systems, the size of a pixel is determined by your system resolution.

When specifying the units as a name-value pair during object creation, you must set the `Units` property before specifying the properties that you want to use these units for, such as `OuterPosition`.

Layout — Layout options
empty `LayoutOptions` array (default) | `TiledChartLayoutOptions` object | `GridLayoutOptions` object

Layout options, specified as a `TiledChartLayoutOptions` or `GridLayoutOptions` object. This property is useful when the chart is either in a tiled chart layout or a grid layout.

To position the chart within the grid of a tiled chart layout, set the `Tile` and `TileSpan` properties on the `TiledChartLayoutOptions` object. For example, consider a 3-by-3 tiled chart layout. The layout has a grid of tiles in the center, and four tiles along the outer edges. In practice, the grid is invisible and the outer tiles do not take up space until you populate them with axes or charts.



This code places the chart `c` in the third tile of the grid..

```
c.Layout.Tile = 3;
```

To make the chart span multiple tiles, specify the `TileSpan` property as a two-element vector. For example, this chart spans 2 rows and 3 columns of tiles.

```
c.Layout.TileSpan = [2 3];
```

To place the chart in one of the surrounding tiles, specify the `Tile` property as `'north'`, `'south'`, `'east'`, or `'west'`. For example, setting the value to `'east'` places the chart in the tile to the right of the grid.

```
c.Layout.Tile = 'east';
```

To place the chart into a layout within an app, specify this property as a `GridLayoutOptions` object. For more information about working with grid layouts in apps, see `uigridlayout`.

If the chart is not a child of either a tiled chart layout or a grid layout (for example, if it is a child of a figure or panel) then this property is empty and has no effect.

`Visible` — State of visibility

`'on'` (default) | on/off logical value

State of visibility, specified as `'on'` or `'off'`, or as numeric or logical `1` (`true`) or `0` (`false`). A value of `'on'` is equivalent to true, and `'off'` is equivalent to `false`. Thus, you can use the value of this property as a logical value. The value is stored as an on/off logical value of type `matlab.lang.OnOffSwitchState`.

- `'on'` — Display the chart.
- `'off'` — Hide the chart without deleting it. You still can access the properties of an invisible chart.

Parent/Child

`Parent` — Parent container

`Figure` object | `Panel` object | `Tab` object | `TiledChartLayout` object | `GridLayout` object

Parent container, specified as a `Figure`, `Panel`, `Tab`, `TiledChartLayout`, or `GridLayout` object.

`HandleVisibility` — Visibility of object handle

`'on'` (default) | `'off'` | `'callback'`

Visibility of the chart object handle in the `Children` property of the parent, specified as one of these values:

- `'on'` — Object handle is always visible.
- `'off'` — Object handle is invisible at all times. This option is useful for preventing unintended changes to the UI by another function. To temporarily hide the handle during the execution of that function, set the `HandleVisibility` to `'off'`.
- `'callback'` — Object handle is visible from within callbacks or functions invoked by callbacks, but not from within functions invoked from the command line. This option blocks access to the object at the command line, but allows callback functions to access it.

If the object is not listed in the `Children` property of the parent, then functions that obtain object handles by searching the object hierarchy or querying handle properties cannot return it. This includes `get`, `findobj`, `gca`, `gcf`, `gco`, `newplot`, `cla`, `clf`, and `close`.

Hidden object handles are still valid. Set the root `ShowHiddenHandles` property to `'on'` to list all object handles, regardless of their `HandleVisibility` property setting.

See Also

Functions
`categorical` | `confusionchart` | `sortClasses`

Introduced in R2018b

confusionmat

Compute confusion matrix for classification problem

Syntax

```
C = confusionmat(group,grouphat)
C = confusionmat(group,grouphat,'Order',grouporder)
[C,order] = confusionmat( ___ )
```

Description

`C = confusionmat(group,grouphat)` returns the confusion matrix `C` determined by the known and predicted groups in `group` and `grouphat`, respectively.

`C = confusionmat(group,grouphat,'Order',grouporder)` uses `grouporder` to order the rows and columns of `C`.

`[C,order] = confusionmat( ___ )` also returns the order of the rows and columns of `C` in the variable `order` using any of the input arguments in previous syntaxes.

Examples

Display Confusion Matrix

Display the confusion matrix for data with two misclassifications and one missing classification.

Create vectors for the known groups and the predicted groups.

```
g1 = [3 2 2 3 1 1]';    % Known groups
g2 = [4 2 3 NaN 1 1]';    % Predicted groups
```

Return the confusion matrix.

```
C = confusionmat(g1,g2)
```

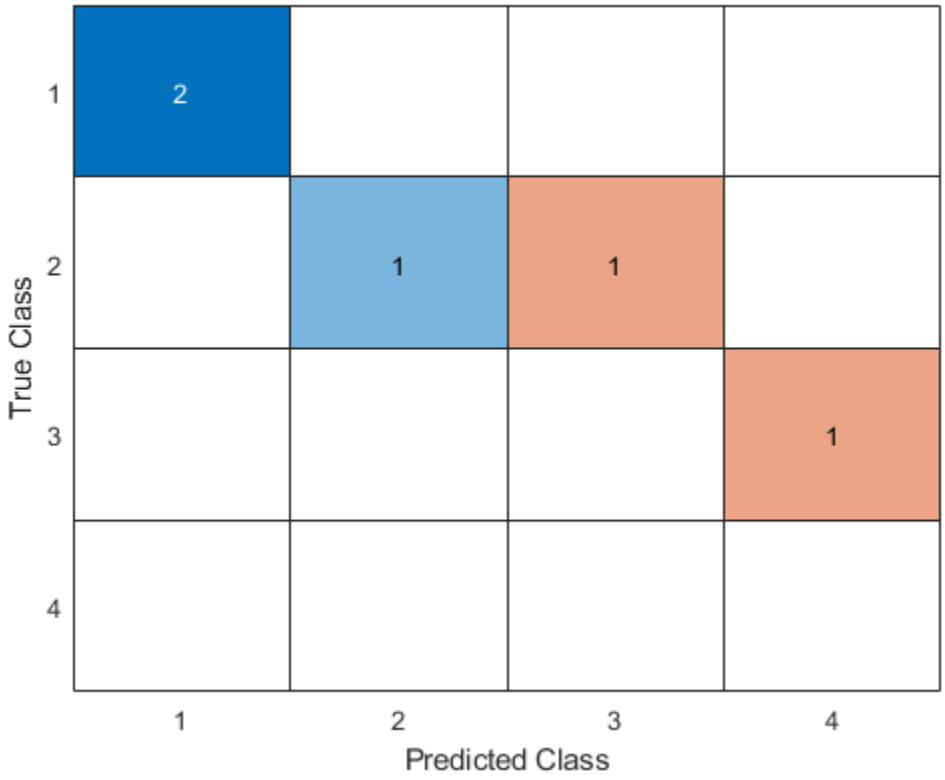
```
C = 4×4

     2     0     0     0
     0     1     1     0
     0     0     0     1
     0     0     0     0
```

The indices of the rows and columns of the confusion matrix `C` are identical and arranged by default in the sorted order of `[g1;g2]`, that is, `(1,2,3,4)`.

The confusion matrix shows that the two data points known to be in group 1 are classified correctly. For group 2, one of the data points is misclassified into group 3. Also, one of the data points known to be in group 3 is misclassified into group 4. `confusionmat` treats the `NaN` value in the grouping variable `g2` as a missing value and does not include it in the rows and columns of `C`.

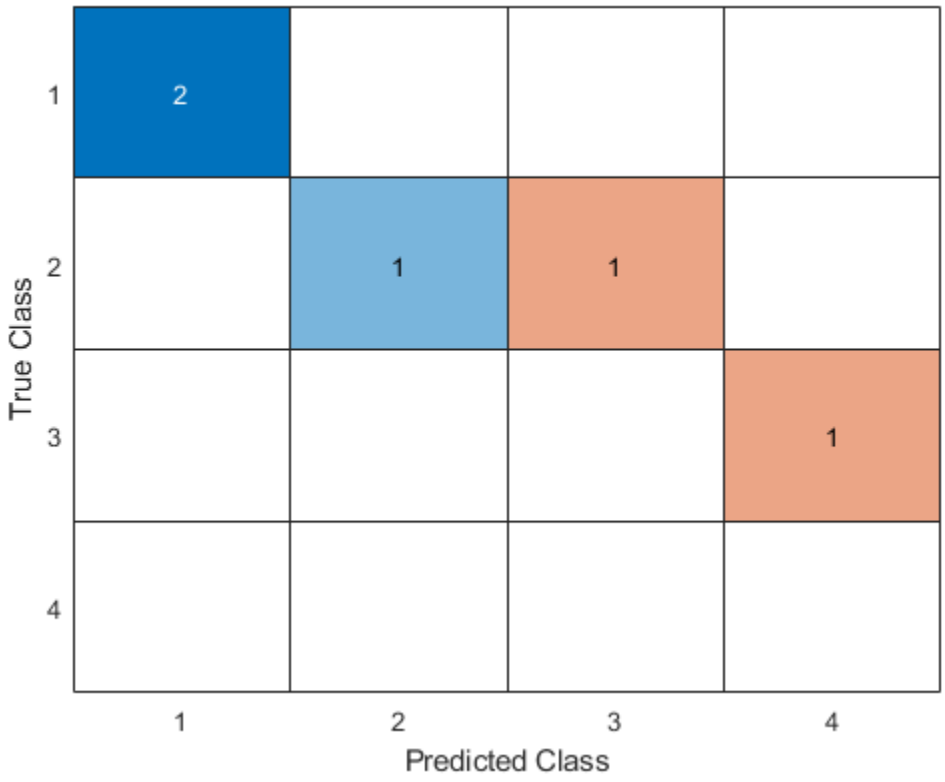
Plot the confusion matrix as a confusion matrix chart by using `confusionchart`.

```
confusionchart(C)
```



You do not need to calculate the confusion matrix first and then plot it. Instead, plot a confusion matrix chart directly from the true and predicted labels by using `confusionchart`.

```
cm = confusionchart(g1,g2)
```




```
cm =
  ConfusionMatrixChart with properties:

    NormalizedValues: [4x4 double]
         ClassLabels: [4x1 double]

  Show all properties
```

The `ConfusionMatrixChart` object stores the numeric confusion matrix in the `NormalizedValues` property and the classes in the `ClassLabels` property. Display these properties using dot notation.

```
cm.NormalizedValues
```

```
ans = 4×4

     2     0     0     0
     0     1     1     0
     0     0     0     1
     0     0     0     0
```

```
cm.ClassLabels
```

```
ans = 4×1

     1
     2
```

```
     3
     4
```

Specify Group Order of Confusion Matrix

Display the confusion matrix for data with two misclassifications and one missing classification, and specify the group order.

Create vectors for the known groups and the predicted groups.

```
g1 = [3 2 2 3 1 1]';    % Known groups
g2 = [4 2 3 NaN 1 1]';    % Predicted groups
```

Specify the group order and return the confusion matrix.

```
C = confusionmat(g1,g2,'Order',[4 3 2 1])
```

```
C = 4×4

     0     0     0     0
     1     0     0     0
     0     1     1     0
     0     0     0     2
```

The indices of the rows and columns of the confusion matrix `C` are identical and arranged in the order specified by the group order, that is, `(4,3,2,1)`.

The second row of the confusion matrix `C` shows that one of the data points known to be in group 3 is misclassified into group 4. The third row of `C` shows that one of the data points belonging to group 2 is misclassified into group 3, and the fourth row shows that the two data points known to be in group 1 are classified correctly. `confusionmat` treats the `NaN` value in the grouping variable `g2` as a missing value and does not include it in the rows and columns of `C`.

Confusion Matrix for Classification

Perform classification on a sample of the `fisheriris` data set and display the confusion matrix for the resulting classification.

Load Fisher's iris data set.

```
load fisheriris
```

Randomize the measurements and groups in the data.

```
rng(0,'twister'); % For reproducibility
numObs = length(species);
p = randperm(numObs);
meas = meas(p,:);
species = species(p);
```

Train a discriminant analysis classifier by using measurements in the first half of the data.

```
half = floor(numObs/2);
training = meas(1:half,:);
trainingSpecies = species(1:half);
Mdl = fitcdiscr(training,trainingSpecies);
```

Predict labels for the measurements in the second half of the data by using the trained classifier.

```
sample = meas(half+1:end,:);
grouphat = predict(Mdl,sample);
```

Specify the group order and display the confusion matrix for the resulting classification.

```
group = species(half+1:end);
[C,order] = confusionmat(group,grouphat,'Order',{'setosa','versicolor','virginica'})
```

```
C = 3×3

    29     0     0
     0    22     2
     0     0    22


order = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

The confusion matrix shows that the measurements belonging to setosa and virginica are classified correctly, while two of the measurements belonging to versicolor are misclassified as virginica. The output `order` contains the order of the rows and columns of the confusion matrix in the sequence specified by the group order `{'setosa','versicolor','virginica'}`.

Confusion Matrix for Classification Using Tall Arrays

Perform classification on a tall array of the `fisheriris` data set, compute a confusion matrix for the known and predicted tall labels by using the `confusionmat` function, and plot the confusion matrix by using the `confusionchart` function.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Load Fisher's iris data set.

```
load fisheriris
```

Convert the in-memory arrays `meas` and `species` to tall arrays.

```
tx = tall(meas);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
ty = tall(species);
```

Find the number of observations in the tall array.

```
numObs = gather(length(ty));   % gather collects tall array into memory
```

Set the seeds of the random number generators using `rng` and `tallrng` for reproducibility, and randomly select training samples. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
numTrain = floor(numObs/2);
[txTrain,trIdx] = datasample(tx,numTrain,'Replace',false);
tyTrain = ty(trIdx);
```

Fit a decision tree classifier model on the training samples.

```
mdl = fitctree(txTrain,tyTrain);
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 3.9 sec
- Pass 2 of 2: Completed in 1.5 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.88 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 4 sec
- Pass 4 of 4: Completed in 2.7 sec
Evaluation completed in 11 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 3 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 7.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 3.1 sec
- Pass 4 of 4: Completed in 2.5 sec
Evaluation completed in 8.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.42 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 3 sec
- Pass 4 of 4: Completed in 2.1 sec
Evaluation completed in 7.6 sec
```

Predict labels for the test samples by using the trained model.

```
txTest = tx(~trIdx,:);
label = predict(mdl,txTest);
```

Compute the confusion matrix for the resulting classification.

```
tyTest = ty(~trIdx);
[C,order] = confusionmat(tyTest,label)
```

```
C =
```

```
  M×N×... tall array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :

Preview deferred. Learn more.


order =

  M×N×... tall array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :

Preview deferred. Learn more.
```

Use the `gather` function to perform the deferred calculation and return the result of `confusionmat` in memory.

```
gather(C)
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.9 sec
Evaluation completed in 2.3 sec

ans = 3×3

    20     0     0
     1    30     2
     0     0    22
```

```
gather(order)
```

```
Evaluating tall expression using the Parallel Pool 'local':
Evaluation completed in 0.032 sec

ans = 3×1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

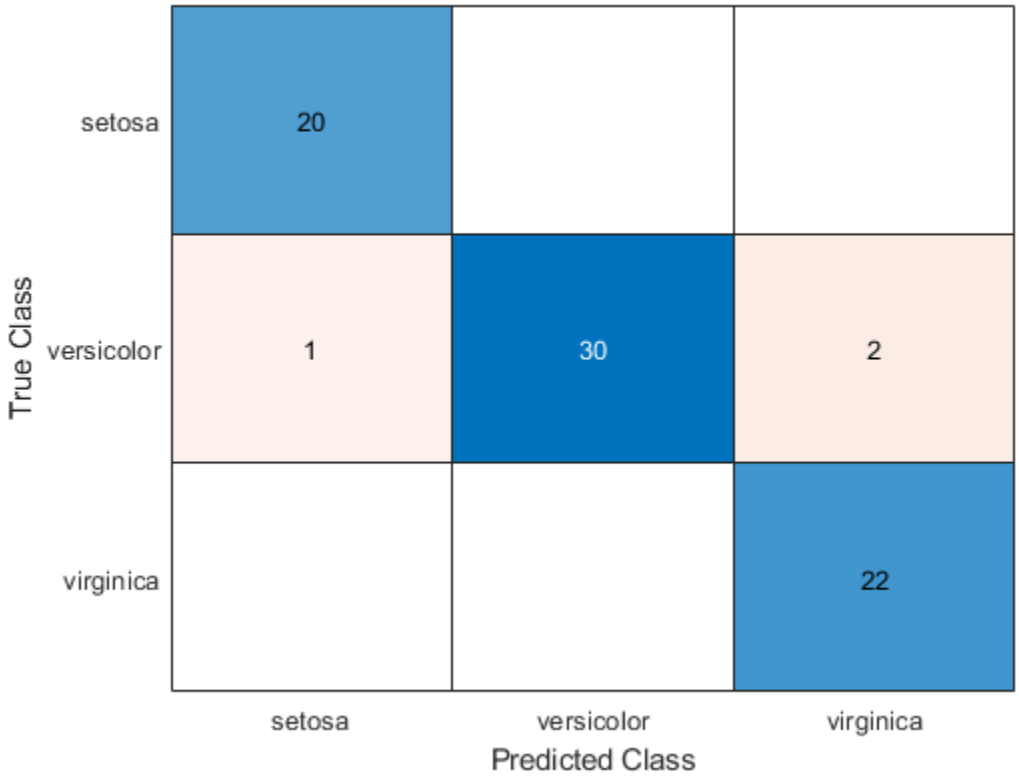
The confusion matrix shows that three measurements in the versicolor class are misclassified. All the measurements belonging to setosa and virginica are classified correctly.

To compute and plot the confusion matrix, use `confusionchart` on page 33-915 instead.

```
cm = confusionchart(tyTest,label)
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.34 sec
```

```
Evaluation completed in 0.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.48 sec
Evaluation completed in 0.67 sec
```




```
cm =
  ConfusionMatrixChart with properties:

    NormalizedValues: [3×3 double]
         ClassLabels: {3×1 cell}

  Show all properties
```

Input Arguments

`group` — Known groups

numeric vector | logical vector | character array | string array | cell array of character vectors | categorical vector

Known groups for categorizing observations, specified as a numeric vector, logical vector, character array, string array, cell array of character vectors, or categorical vector.

`group` is a grouping variable of the same type as `grouphat`. The `group` argument must have the same number of observations as `grouphat`, as described in “Grouping Variables” on page 2-45. The

`confusionmat` function treats character arrays and string arrays as cell arrays of character vectors. Additionally, `confusionmat` treats `NaN`, empty, and `'undefined'` values in `group` as missing values and does not count them as distinct groups or categories.

Example: `{'Male','Female','Female','Male','Female'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`grouphat` — Predicted groups
numeric vector | logical vector | character array | string array | cell array of character vectors | categorical vector

Predicted groups for categorizing observations, specified as a numeric vector, logical vector, character array, string array, cell array of character vectors, or categorical vector.

`grouphat` is a grouping variable of the same type as `group`. The `grouphat` argument must have the same number of observations as `group`, as described in "Grouping Variables" on page 2-45. The `confusionmat` function treats character arrays and string arrays as cell arrays of character vectors. Additionally, `confusionmat` treats `NaN`, empty, and `'undefined'` values in `grouphat` as missing values and does not count them as distinct groups or categories.

Example: `[1 0 0 1 0]`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`grouporder` — Group order
numeric vector | logical vector | character array | string array | cell array of character vectors | categorical vector

Group order, specified as a numeric vector, logical vector, character array, string array, cell array of character vectors, or categorical vector.

`grouporder` is a grouping variable containing all the distinct elements in `group` and `grouphat`. Specify `grouporder` to define the order of the rows and columns of `C`. If `grouporder` contains elements that are not in `group` or `grouphat`, the corresponding entries in `C` are `0`.

By default, the group order depends on the data type of `s = [group;grouphat]`:

- For numeric and logical vectors, the order is the sorted order of `s`.
- For categorical vectors, the order is the order returned by `categories(s)`.
- For other data types, the order is the order of first appearance in `s`.

Example: `'order',{'setosa','versicolor','virginica'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Output Arguments

`C` — Confusion matrix
matrix

Confusion matrix, returned as a square matrix with size equal to the total number of distinct elements in the `group` and `grouphat` arguments. `C(i,j)` is the count of observations known to be in group `i` but predicted to be in group `j`.

The rows and columns of `C` have identical ordering of the same group indices. By default, the group order depends on the data type of `s = [group;grouphat]`:

- For numeric and logical vectors, the order is the sorted order of `s`.
- For categorical vectors, the order is the order returned by `categories(s)`.
- For other data types, the order is the order of first appearance in `s`.

To change the order, specify `grouporder`,

The `confusionmat` function treats `NaN`, empty, and `'undefined'` values in the grouping variables as missing values and does not include them in the rows and columns of `C`.

`order` — Order of rows and columns
numeric vector | logical vector | categorical vector | cell array of character vectors

Order of rows and columns in `C`, returned as a numeric vector, logical vector, categorical vector, or cell array of character vectors. If `group` and `grouphat` are character arrays, string arrays, or cell arrays of character vectors, then the variable `order` is a cell array of character vectors. Otherwise, `order` is of the same type as `group` and `grouphat`.

Alternative Functionality

- Use `confusionchart` to calculate and plot a confusion matrix. Additionally, `confusionchart` displays summary statistics about your data and sorts the classes of the confusion matrix according to the class-wise precision (positive predictive value), class-wise recall (true positive rate), or total number of correctly classified observations.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also
`categories` | `confusionchart` | `crosstab`

Topics
“Grouping Variables” on page 2-45

Introduced in R2008b

controlchart

Shewhart control charts

Syntax

```
controlchart(X)
controlchart(x,group)
controlchart(X,group)
[stats,plotdata] = controlchart(x,[group])
controlchart(x,group,'name',value)
```

Description

`controlchart(X)` produces an xbar chart of the measurements in matrix `X`. Each row of `X` is considered to be a subgroup of measurements containing replicate observations taken at the same time. The rows should be in time order. If `X` is a time series object, the time samples should contain replicate observations.

The chart plots the means of the subgroups in time order, a center line (`CL`) at the average of the means, and upper and lower control limits (`UCL`, `LCL`) at three standard errors from the center line. The standard error is the estimated process standard deviation divided by the square root of the subgroup size. Process standard deviation is estimated from the average of the subgroup standard deviations. Out of control measurements are marked as violations and drawn with a red circle. Data cursor mode is enabled, so clicking any data point displays information about that point.

`controlchart(x,group)` accepts a grouping variable `group` for a vector of measurements `x`. `group` is a categorical variable, numeric vector, character vector, string array, or cell array of character vectors the same length as `x`. Consecutive measurements `x(n)` sharing the same value of `group(n)` for $1 \le n \le$ `length(x)` are defined to be a subgroup. Subgroups can have different numbers of observations.

`controlchart(X,group)` accepts a grouping variable `group` for a matrix of measurements in `X`. In this case, `group` is only used to label the time axis; it does not change the default grouping by rows.

`[stats,plotdata] = controlchart(x,[group])` returns a structure `stats` of subgroup statistics and parameter estimates, and a structure `plotdata` of plotted values. `plotdata` contains one record for each chart.

The fields in `stats` and `plotdata` depend on the chart type.

The fields in `stats` are selected from the following:

- `mean` — Subgroup means
- `std` — Subgroup standard deviations
- `range` — Subgroup ranges
- `n` — Subgroup size, or total inspection size or area
- `i` — Individual data values
- `ma` — Moving averages

- `mr` — Moving ranges
- `count` — Count of defects or defective items
- `mu` — Estimated process mean
- `sigma` — Estimated process standard deviation
- `p` — Estimated proportion defective
- `m` — Estimated mean defects per unit

The fields in `plotdata` are the following:

- `pts` — Plotted point values
- `cl` — Center line
- `lcl` — Lower control limit
- `ucl` — Upper control limit
- `se` — Standard error of plotted point
- `n` — Subgroup size
- `ooc` — Logical that is true for points that are out of control

`controlchart(x,group,'`*`name`*`',`*`value`*`)` specifies one or more of the following optional parameter name/value pairs, with *`name`* in single quotes:

- `charttype` — The name of a chart type chosen from among the following:
  - `'xbar'` — Xbar or mean
  - `'s'` — Standard deviation
  - `'r'` — Range
  - `'ewma'` — Exponentially weighted moving average
  - `'i'` — Individual observation
  - `'mr'` — Moving range of individual observations
  - `'ma'` — Moving average of individual observations
  - `'p'` — Proportion defective
  - `'np'` — Number of defectives
  - `'u'` — Defects per unit
  - `'c'` — Count of defects

  Alternatively, a parameter can be a string array or cell array listing multiple compatible chart types. There are four sets of compatible types:

  - `'xbar'`, `'s'`, `'r'`, and `'ewma'`
  - `'i'`, `'mr'`, and `'ma'`
  - `'p'` and `'np'`
  - `'u'` and `'c'`
- `display` — Either `'on'` (default) to display the control chart, or `'off'` to omit the display
- `label` — A character vector, string array, or cell array of character vectors, one per subgroup. This label is displayed as part of the data cursor for a point on the plot.

- `lambda` — A parameter between 0 and 1 controlling how much the current prediction is influenced by past observations in an EWMA plot. Higher values of `'lambda'` give less weight to past observations and more weight to the current observation. The default is 0.4.
- `limits` — A three-element vector specifying the values of the lower control limit, center line, and upper control limits. Default is to estimate the center line and to compute control limits based on the estimated value of sigma. Not permitted if there are multiple chart types.
- `mean` — Value for the process mean, or an empty value (default) to estimate the mean from `X`. This is the `p` parameter for `p` and `np` charts, the mean defects per unit for `u` and `c` charts, and the normal `mu` parameter for other charts.
- `nsigma` — The number of sigma multiples from the center line to a control limit. Default is 3.
- `parent` — The handle of the axes to receive the control chart plot. Default is to create axes in a new figure. Not permitted if there are multiple chart types.
- `rules` — The name of a control rule, or a string array or cell array containing multiple control rule names. These rules, together with the control limits, determine if a point is marked as out of control. The default is to apply no control rules, and to use only the control limits to decide if a point is out of control. See `controlrules` for more information. Control rules are applied to charts that measure the process level (`xbar`, `i`, `c`, `u`, `p`, and `np`) rather than the variability (`r`, `s`), and they are not applied to charts based on moving statistics (`ma`, `mr`, `ewma`).
- `sigma` — Either a value for sigma, or a method of estimating sigma chosen from among `'std'` (the default) to use the average within-subgroup standard deviation, `'range'` to use the average subgroup range, and `'variance'` to use the square root of the pooled variance. When creating `i`, `mr`, or `ma` charts for data not in subgroups, the estimate is always based on a moving range.
- `specs` — A vector specifying specification limits. Typically this is a two-element vector of lower and upper specification limits. Since specification limits typically apply to individual measurements, this parameter is primarily suitable for `i` charts. These limits are not plotted on `r`, `s`, or `mr` charts.
- `unit` — The total number of inspected items for `p` and `np` charts, and the size of the inspected unit for `u` and `c` charts. In both cases `X` must be the count of the number of defects or defectives found. Default is 1 for `u` and `c` charts. This argument is required (no default) for `p` and `np` charts.
- `width` — The width of the window used for computing the moving ranges and averages in `mr` and `ma` charts, and for computing the sigma estimate in `i`, `mr`, and `ma` charts. Default is 5.

Examples

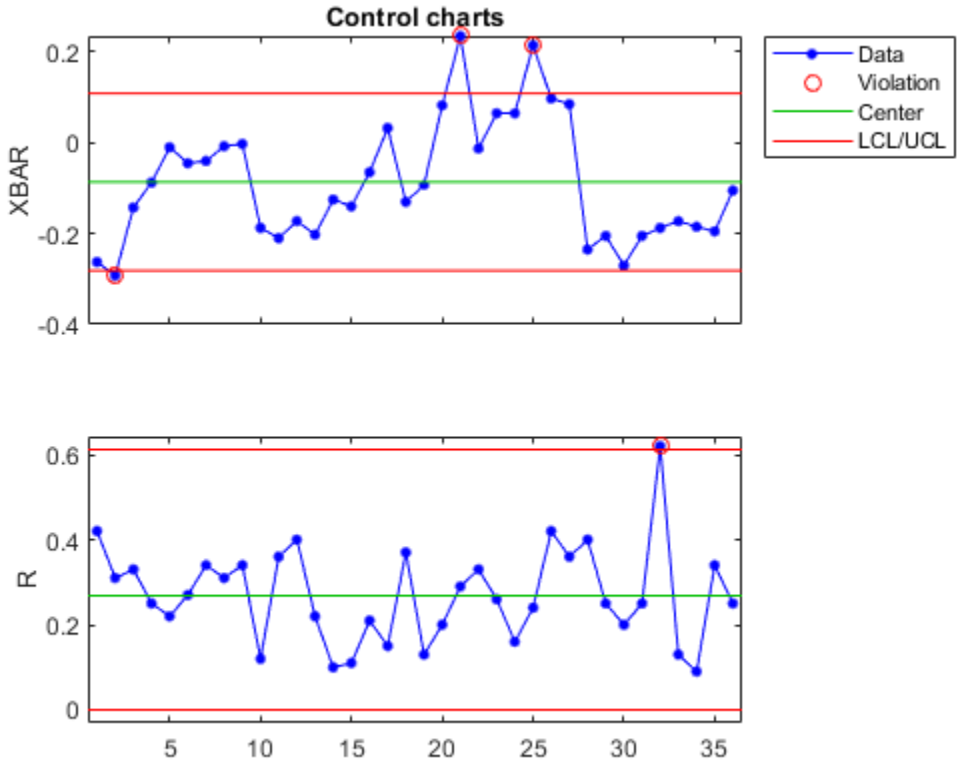
XBar and R Charts

Load the sample data.

```
load parts
```

Create xbar and r control charts for the data.

```
st = controlchart(runout,'charttype',{'xbar' 'r'});
```



Display the process mean and standard deviation.

```
fprintf('Parameter estimates:  mu = %g, sigma = %g\n',st.mu,st.sigma);
```

```
Parameter estimates:  mu = -0.0863889, sigma = 0.130215
```

See Also

`controlrules`

Topics

“Grouping Variables” on page 2-45

Introduced in R2006b

controlrules

Western Electric and Nelson control rules

Syntax

```
R = controlrules('rules',x,cl,se)
[R,RULES] = controlrules('rules',x,cl,se)
```

Description

`R = controlrules('rules',x,cl,se)` determines which points in the vector `x` violate the control rules in *`rules`*. `cl` is a vector of center-line values. `se` is a vector of standard errors. (Typically, control limits on a control chart are at the values `cl – 3*se` and `cl + 3*se`.) `rules` is the name of a control rule, or a string array or cell array containing multiple control rule names. If `x` has *n* values and *`rules`* contains *m* rules, then `R` is an *n*-by-*m* logical array, with `R(i,j)` assigned the value `1` if point `i` violates rule `j`, `0` if it does not.

The following are accepted values for *`rules`* (specified inside single quotes):

- `we1` — 1 point above `cl + 3*se`
- `we2` — 2 of 3 above `cl + 2*se`
- `we3` — 4 of 5 above `cl + se`
- `we4` — 8 of 8 above `cl`
- `we5` — 1 below `cl − 3*se`
- `we6` — 2 of 3 below `cl − 2*se`
- `we7` — 4 of 5 below `cl − se`
- `we8` — 8 of 8 below `cl`
- `we9` — 15 of 15 between `cl − se` and `cl + se`
- `we10` — 8 of 8 below `cl − se` or above `cl + se`
- `n1` — 1 point below `cl − 3*se` or above `cl + 3*se`
- `n2` — 9 of 9 on the same side of `cl`
- `n3` — 6 of 6 increasing or decreasing
- `n4` — 14 alternating up/down
- `n5` — 2 of 3 below `cl − 2*se` or above `cl + 2*se`, same side
- `n6` — 4 of 5 below `cl − se` or above `cl + se`, same side
- `n7` — 15 of 15 between `cl − se` and `cl + se`
- `n8` — 8 of 8 below `cl − se` or above `cl + se`, either side
- `we` — All Western Electric rules
- `n` — All Nelson rules

For multi-point rules, a rule violation at point `i` indicates that the set of points ending at point `i` triggered the rule. Point `i` is considered to have violated the rule only if it is one of the points violating the rule's condition.

Any points with `NaN` as their `x`, `cl`, or `se` values are not considered to have violated rules, and are not counted in the rules for other points.

Control rules can be specified in the `controlchart` function as values for the `'rules'` parameter.

`[R,RULES] = controlrules('rules',x,cl,se)` returns a cell array of text `RULES` listing the rules applied.

Examples

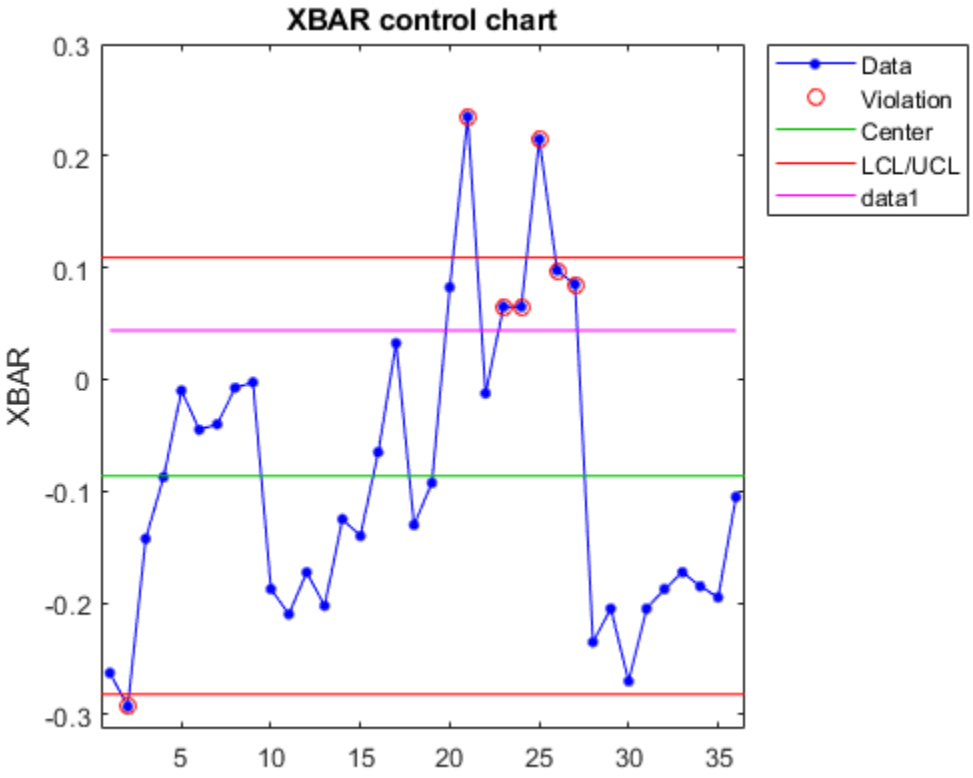
Use Western Electric Control Rule

Load the sample data.

```
load parts;
```

Create an Xbar chart using the `we2` rule to mark out of control measurements.

```
st = controlchart(runout,'rules','we2');
x = st.mean;
cl = st.mu;
se = st.sigma./sqrt(st.n);
hold on
plot(cl+2*se,'m')
```



You can see the out of control points marked with a red circle.

Use `controlrules` to identify the measurements that violate the control rule.

```
R = controlrules('we2',x,cl,se);
I = find(R)
```

```
I = 6×1

    21
    23
    24
    25
    26
    27
```

See Also

`controlchart`

Introduced in R2006b

cophenet

Cophenetic correlation coefficient

Syntax

```
c = cophenet(Z,Y)
[c,d] = cophenet(Z,Y)
```

Description

`c = cophenet(Z,Y)` computes the cophenetic correlation coefficient for the hierarchical cluster tree represented by `Z`. `Z` is the output of the `linkage` function. `Y` contains the distances or dissimilarities used to construct `Z`, as output by the `pdist` function. `Z` is a matrix of size (*m*–1)-by-3, with distance information in the third column. `Y` is a vector of size *m**(*m*–1)/2.

`[c,d] = cophenet(Z,Y)` returns the cophenetic distances `d` in the same lower triangular distance vector format as `Y`.

The cophenetic correlation for a cluster tree is defined as the linear correlation coefficient between the cophenetic distances obtained from the tree, and the original distances (or dissimilarities) used to construct the tree. Thus, it is a measure of how faithfully the tree represents the dissimilarities among observations.

The cophenetic distance between two observations is represented in a dendrogram by the height of the link at which those two observations are first joined. That height is the distance between the two subclusters that are merged by that link.

The output value, `c`, is the cophenetic correlation coefficient. The magnitude of this value should be very close to 1 for a high-quality solution. This measure can be used to compare alternative cluster solutions obtained using different algorithms.

The cophenetic correlation between `Z(:,3)` and `Y` is defined as

$$c = \frac{\sum_{i<j}(Y_{ij}-y)(Z_{ij}-z)}{\sqrt{\sum_{i<j}(Y_{ij}-y)^2\sum_{i<j}(Z_{ij}-z)^2}}$$

where:

- Y_{ij} is the distance between objects i and j in `Y`.
- Z_{ij} is the cophenetic distance between objects i and j, from `Z(:,3)`.
- y and z are the average of `Y` and `Z(:,3)`, respectively.

Examples

```
X = [rand(10,3); rand(10,3)+1; rand(10,3)+2];
Y = pdist(X);
Z = linkage(Y,'average');

% Compute Spearman's rank correlation between the
```

```
% dissimilarities and the cophenetic distances
[c,D] = cophenet(Z,Y);
r = corr(Y',D','type','spearman')
r =
   0.8279
```

See Also

cluster | dendrogram | inconsistent | linkage | pdist | squareform

Introduced before R2006a

copulacdf

Copula cumulative distribution function

Syntax

```
y = copulacdf('Gaussian',u,rho)

y = copulacdf('t',u,rho,nu)

y = copulacdf(family,u,alpha)
```

Description

`y = copulacdf('Gaussian',u,rho)` returns the cumulative probability of the Gaussian copula, with linear correlation parameters `rho` evaluated at the points in `u`.

`y = copulacdf('t',u,rho,nu)` returns the cumulative probability of the *t* copula, with linear correlation parameters, `rho`, and degrees of freedom parameter `nu` evaluated at the points in `u`.

`y = copulacdf(family,u,alpha)` returns the cumulative probability of the bivariate Archimedean copula of the type specified by `family`, with scalar parameter `alpha` evaluated at the points in `u`.

Examples

Compute the Clayton Copula cdf

Define two 10-by-10 matrices containing the values at which to compute the cdf.

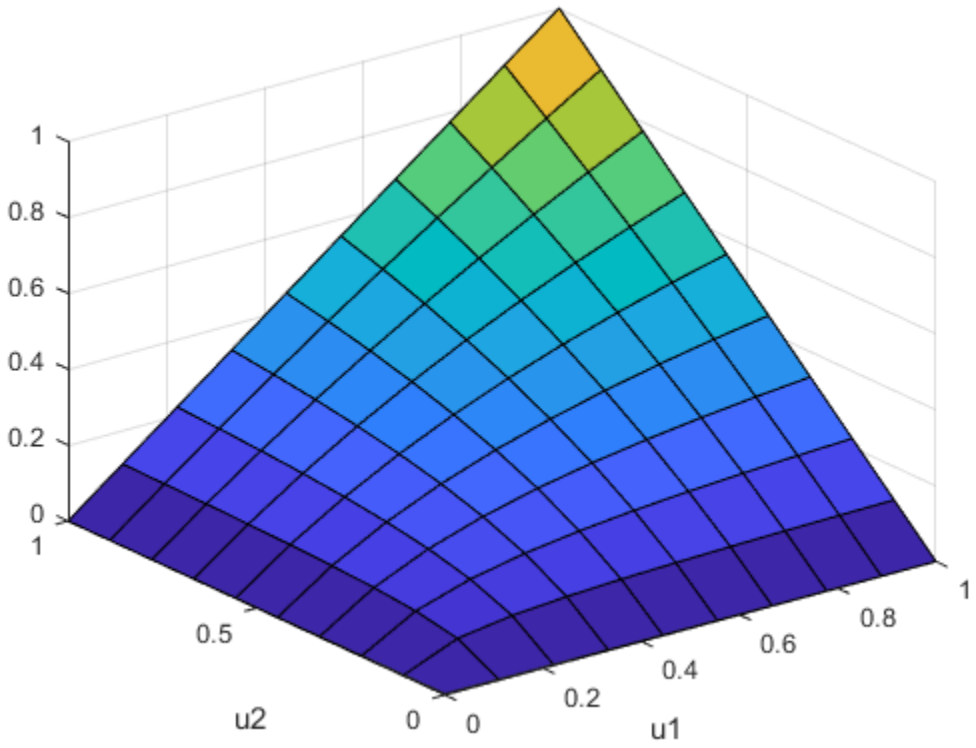
```
u = linspace(0,1,10);
[u1,u2] = meshgrid(u,u);
```

Compute the cdf of a Clayton copula that has an alpha parameter equal to 1, at the values in `u`.

```
y = copulacdf('Clayton',[u1(:),u2(:)],1);
```

Plot the cdf as a surface, and label the axes.

```
surf(u1,u2,reshape(y,10,10))
xlabel('u1')
ylabel('u2')
```



Input Arguments

u — Values at which to evaluate cdf

matrix of scalar values in the range [0,1]

Values at which to evaluate the cdf, specified as a matrix of scalar values in the range [0,1]. If `u` is an *n*-by-*p* matrix, then its values represent *n* points in the *p*-dimensional unit hypercube. If `u` is an *n*-by-2 matrix, then its values represent *n* points in the unit square.

If you specify a bivariate Archimedean copula type (`'Clayton'`, `'Frank'`, or `'Gumbel'`), then `u` must be an *n*-by-2 matrix.

Data Types: `single` | `double`

rho — Linear correlation parameters

scalar values | matrix of scalar values

Linear correlation parameters for the copula, specified as a scalar value or matrix of scalar values.

- If `u` is an *n*-by-*p* matrix, then `rho` is a *p*-by-*p* correlation matrix.
- If `u` is an *n*-by-2 matrix, then `rho` can be a scalar correlation coefficient.

Data Types: `single` | `double`

nu — Degrees of freedom
positive integer value

Degrees of freedom for the *t* copula, specified as a positive integer value.

Data Types: `single` | `double`

family — Bivariate Archimedean copula family
`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

alpha — Bivariate Archimedean copula parameter
scalar value

Bivariate Archimedean copula parameter, specified as a scalar value. Permitted values for `alpha` depend on the specified copula family.

Copula Family	Permitted Alpha Values
`'Clayton'`	[0,∞)
`'Frank'`	(-∞,∞)
`'Gumbel'`	[1,∞)

Data Types: `single` | `double`

Output Arguments

y — Cumulative distribution function
vector of scalar values

Cumulative distribution function of the copula, evaluated at the points in `u`, returned as a vector of scalar values.

See Also

`copulaparam` | `copulapdf` | `copularnd` | `copulastat`

Topics
“Generate Correlated Data Using Rank Correlation” on page 5-108
“Copulas: Generate Correlated Samples” on page 5-121

Introduced in R2006a

copulafit

Fit copula to data

Syntax

```
rhohat = copulafit('Gaussian',u)

[rhohat,nuhat] = copulafit('t',u)
[rhohat,nuhat,nuci] = copulafit('t',u)

paramhat = copulafit(family,u)
[paramhat,paramci] = copulafit(family,u)

___ = copulafit(___,Name,Value)
```

Description

`rhohat = copulafit('Gaussian',u)` returns an estimate, `rhohat`, of the matrix of linear correlation parameters for a Gaussian copula, given the data in `u`.

`[rhohat,nuhat] = copulafit('t',u)` returns an estimate, `rhohat`, of the matrix of linear correlation parameters for a *t* copula, and an estimate of the degrees of freedom parameter, `nuhat`, given the data in `u`.

`[rhohat,nuhat,nuci] = copulafit('t',u)` also returns an approximate 95% confidence interval, `nuci`, for the degrees of freedom estimated in `nuhat`.

`paramhat = copulafit(family,u)` returns an estimate, `paramhat`, of the copula parameter for a bivariate Archimedean copula of the type specified by `family`, given the data in `u`.

`[paramhat,paramci] = copulafit(family,u)` also returns an approximate 95% confidence interval, `paramci`, for the copula parameter estimated in `paramhat`.

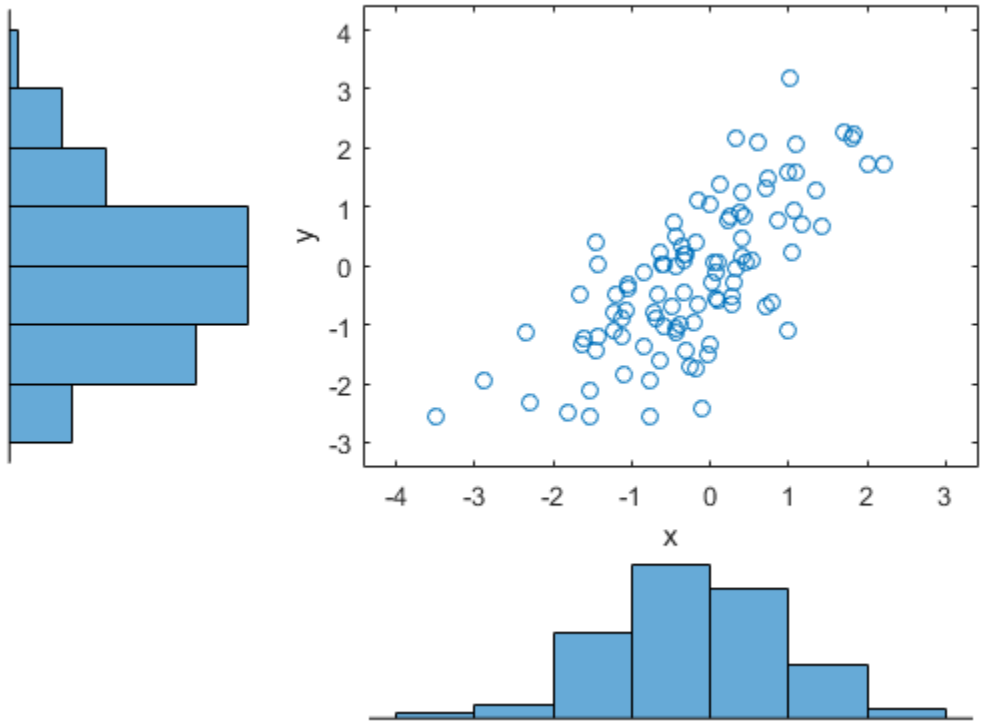
`___ = copulafit(___,Name,Value)` returns any of the previous syntaxes, with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence interval to compute, or specify control parameters for the iterative parameter estimation algorithm using a options structure.

Examples

Fit a *t* Copula to Data

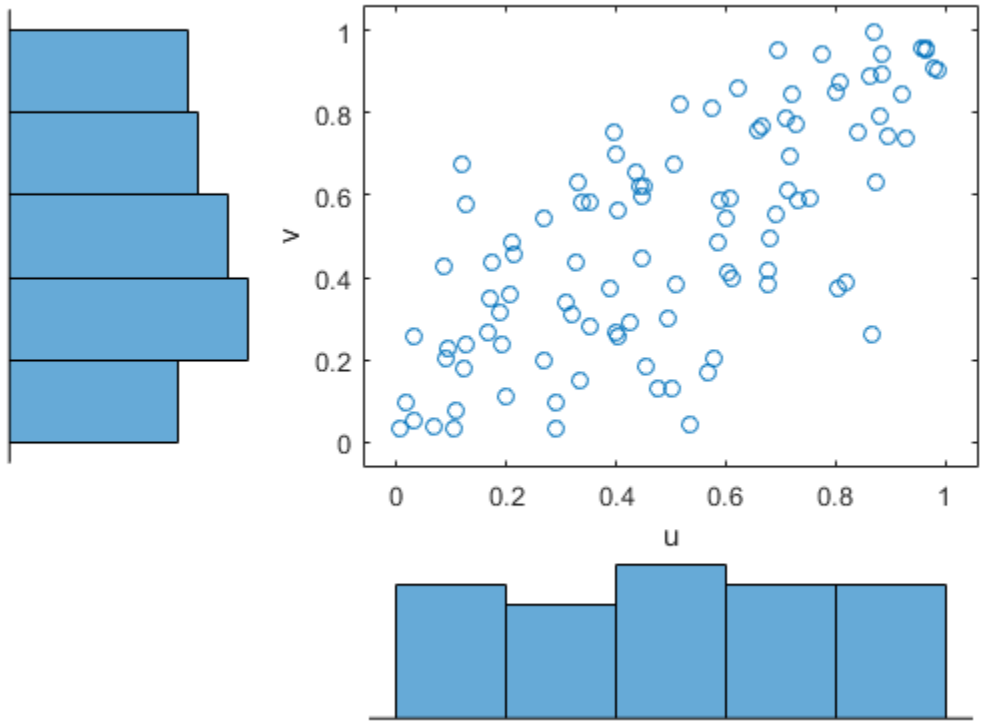
Load and plot simulated stock return data.

```
load stockreturns
x = stocks(:,1);
y = stocks(:,2);

figure;
scatterhist(x,y)
```



Transform the data to the copula scale (unit square) using a kernel estimator of the cumulative distribution function.

```
u = ksdensity(x,x,'function','cdf');
v = ksdensity(y,y,'function','cdf');

figure;
scatterhist(u,v)
xlabel('u')
ylabel('v')
```



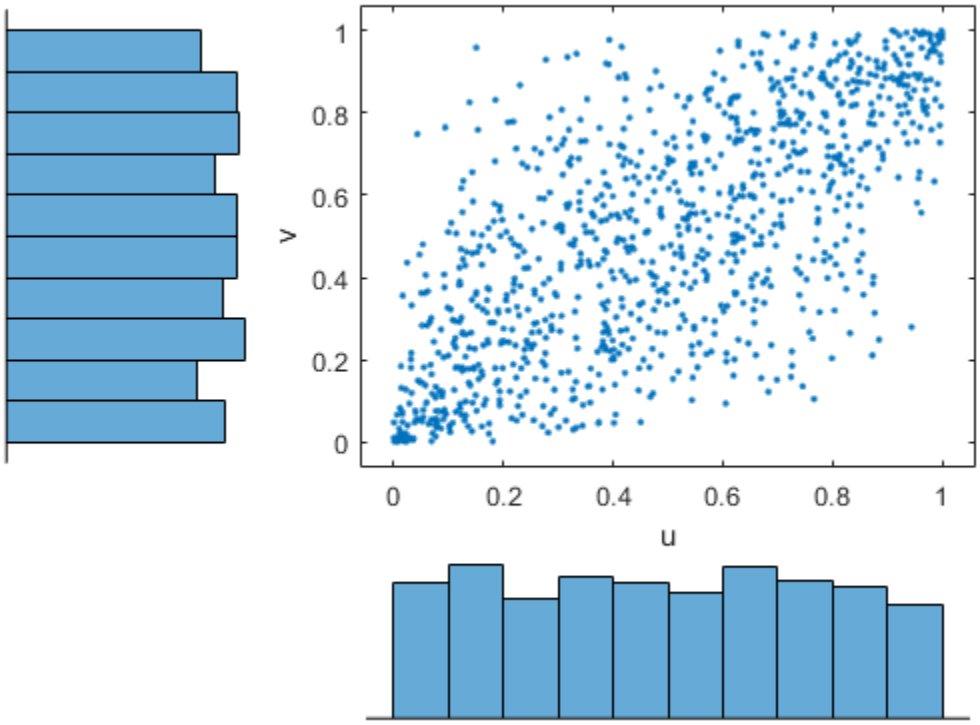
Fit a *t* copula to the data.

```
rng default  % For reproducibility
[Rho,nu] = copulafit('t',[u v],'Method','ApproximateML')
```

```
Rho =

    1.0000    0.7220
    0.7220    1.0000


nu =

   3.4516e+06
```

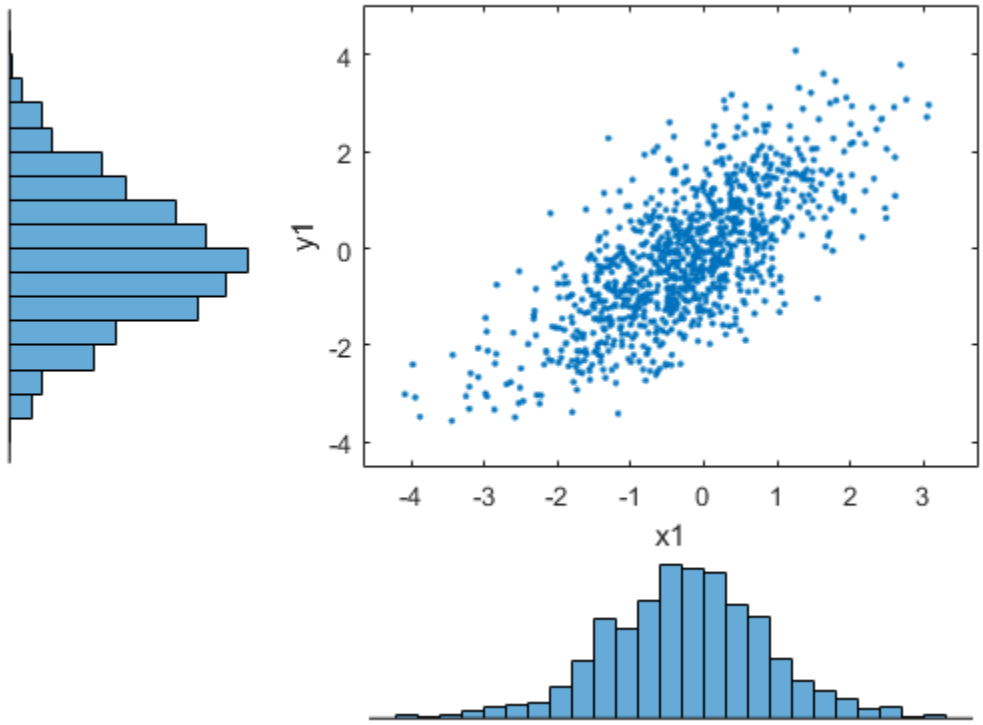
Generate a random sample from the *t* copula.

```
r = copularnd('t',Rho,nu,1000);
u1 = r(:,1);
v1 = r(:,2);

figure;
scatterhist(u1,v1)
xlabel('u')
ylabel('v')
set(get(gca,'children'),'marker','.')
```



Transform the random sample back to the original scale of the data.

```
x1 = ksdensity(x,u1,'function','icdf');
y1 = ksdensity(y,v1,'function','icdf');

figure;
scatterhist(x1,y1)
set(get(gca,'children'),'marker','.')
```



Input Arguments

`u` — Copula values

matrix of scalar values in the range (0,1)

Copula values, specified as a matrix of scalar values in the range (0,1). If `u` is an *n*-by-*p* matrix, then its values represent *n* points in the *p*-dimensional unit hypercube. If `u` is an *n*-by-2 matrix, then its values represent *n* points in the unit square.

If you specify a bivariate Archimedean copula type (`'Clayton'`, `'Frank'`, or `'Gumbel'`), then `u` must be an *n*-by-2 matrix.

Data Types: `single` | `double`

`family` — Bivariate Archimedean copula family

`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'Method','ApproximateML'` computes 99% confidence intervals for the estimated copula parameter and uses an approximation method to fit the copula.

`Alpha` — Significance level for confidence intervals
0.05 (default) | scalar value in the range (0,1)

Significance level for confidence intervals, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). `copulafit` returns approximate 100 × (1–Alpha)% confidence intervals.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Method` — Method for fitting *t* copula
`'ML'` (default) | `'ApproximateML'`

Method for fitting *t* copula, specified as the comma-separated pair consisting of `'Method'` and either `'ML'` or `'ApproximateML'`.

If you specify `'ApproximateML'`, then `copulafit` fits a *t* copula for large samples by maximizing an objective function that approximates the profile log likelihood for the degrees of freedom parameter [1]. This method can be significantly faster than maximum likelihood (`'ML'`), but the estimates and confidence limits may not be accurate for small to moderate sample sizes.

Example: `'Method','ApproximateML'`

`Options` — Control parameter specifications
structure

Control parameter specifications, specified as the comma-separated pair consisting of `'Options'` and an options structure created by `statset`. To see the fields and default values used by `copulafit`, type `statset('copulafit')` at the command prompt.

This name-value pair is not applicable when you specify the copula type as `'Gaussian'`.

Data Types: `struct`

Output Arguments

`rhohat` — Estimated correlation parameters for the fitted Gaussian copula
matrix of scalar values

Estimated correlation parameters for the fitted Gaussian copula, given the data in `u`, returned as a matrix of scalar values.

`nuhat` — Estimated degrees of freedom parameter for the fitted *t* copula
scalar value

Estimated degrees of freedom parameter for the fitted *t* copula, returned as a scalar value.

`nuci` — Approximate confidence interval for the degrees of freedom parameter
1-by-2 matrix of scalar values

Approximate confidence interval for the degrees of freedom parameter, returned as a 1-by-2 matrix of scalar values. The first column contains the lower boundary, and the second column contains the upper boundary. By default, `copulafit` returns the approximate 95% confidence interval. You can specify a different confidence interval using the `'Alpha'` name-value pair.

`paramhat` — Estimated copula parameter for the fitted Archimedean copula
scalar value

Estimated copula parameter for the fitted Archimedean copula, returned as a scalar value.

`paramci` — Approximate confidence interval for the copula parameter
1-by-2 matrix of scalar values

Approximate confidence interval for the copula parameter, returned as a 1-by-2 matrix of scalar values. The first column contains the lower boundary, and the second column contains the upper boundary. By default, `copulafit` returns the approximate 95% confidence interval. You can specify a different confidence interval using the `'Alpha'` name-value pair.

Algorithms

By default, `copulafit` uses maximum likelihood to fit a copula to `u`. When `u` contains data transformed to the unit hypercube by parametric estimates of their marginal cumulative distribution functions, this is known as the *Inference Functions for Margins (IFM)* method. When `u` contains data transformed by the empirical cdf (see `ecdf`), this is known as *Canonical Maximum Likelihood (CML)*.

References

[1] Bouyé, E., V. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli. “Copulas for Finance: A Reading Guide and Some Applications.” Working Paper. Groupe de Recherche Opérationnelle, Crédit Lyonnais, Paris, 2000.

See Also

`copulacdf` | `copulaparam` | `copulapdf` | `copularnd` | `copulastat`

Topics
“Generate Correlated Data Using Rank Correlation” on page 5-108
“Copulas: Generate Correlated Samples” on page 5-121

Introduced in R2007b

copulaparam

Copula parameters as function of rank correlation

Syntax

```
rho = copulaparam('Gaussian',r)

rho = copulaparam('t',r,nu)

alpha = copulaparam(family,r)

___ = copulaparam(___,Name,Value)
```

Description

`rho = copulaparam('Gaussian',r)` returns the linear correlation parameters, `rho`, that correspond to a Gaussian copula with Kendall's rank correlation, `r`.

`rho = copulaparam('t',r,nu)` returns the linear correlation parameters, `rho`, that correspond to a *t* copula with Kendall's rank correlation, `r`, and degrees of freedom, `nu`.

`alpha = copulaparam(family,r)` returns the copula parameter, `alpha`, that corresponds to a bivariate Archimedean copula of the type specified by `family`, with Kendall's rank correlation, `r`.

`___ = copulaparam(___,Name,Value)` returns the correlation parameter using any of the previous syntaxes, with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify whether the input rank correlation value is Spearman's *rho* or Kendall's *tau*.

Examples

Generate Correlated Data Using the Inverse cdf

Generate correlated random data from a beta distribution using a bivariate Gaussian copula with Kendall's *tau* rank correlation equal to -0.5.

Compute the linear correlation parameter from the rank correlation value.

```
rng default  % For reproducibility
tau = -0.5;
rho = copulaparam('Gaussian',tau)
```

```
rho = -0.7071
```

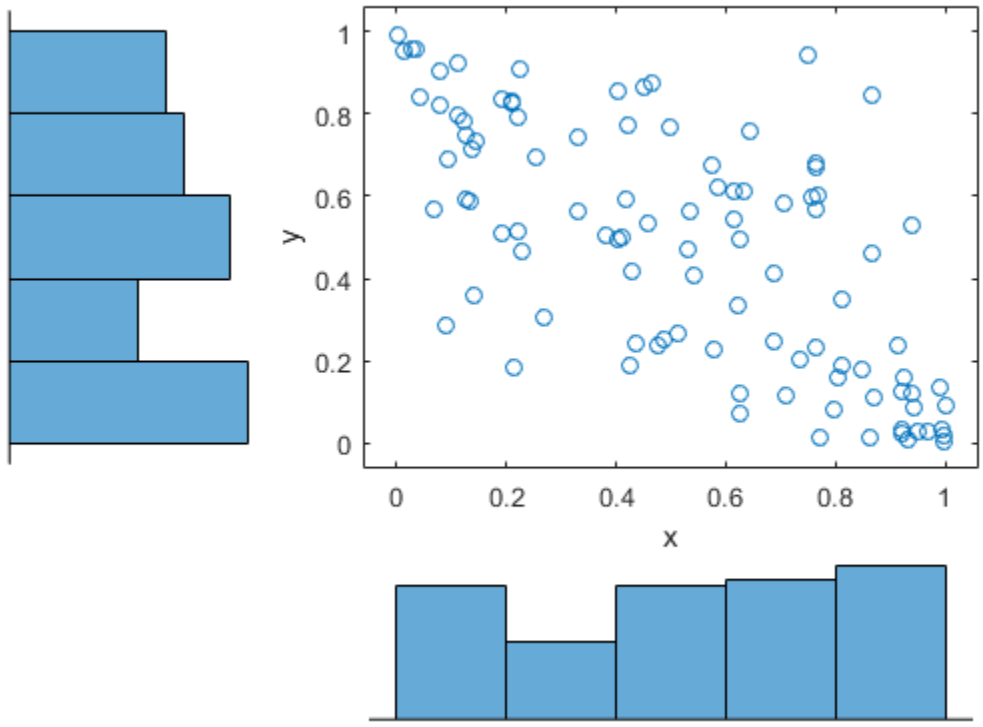
Use a Gaussian copula to generate a two-column matrix of dependent random values.

```
u = copularnd('gaussian',rho,100);
```

Each column contains 100 random values between 0 and 1, inclusive, sampled from a continuous uniform distribution.

Create a `scatterhist` plot to visualize the random numbers generated using the copula.

```
figure
scatterhist(u(:,1),u(:,2))
```



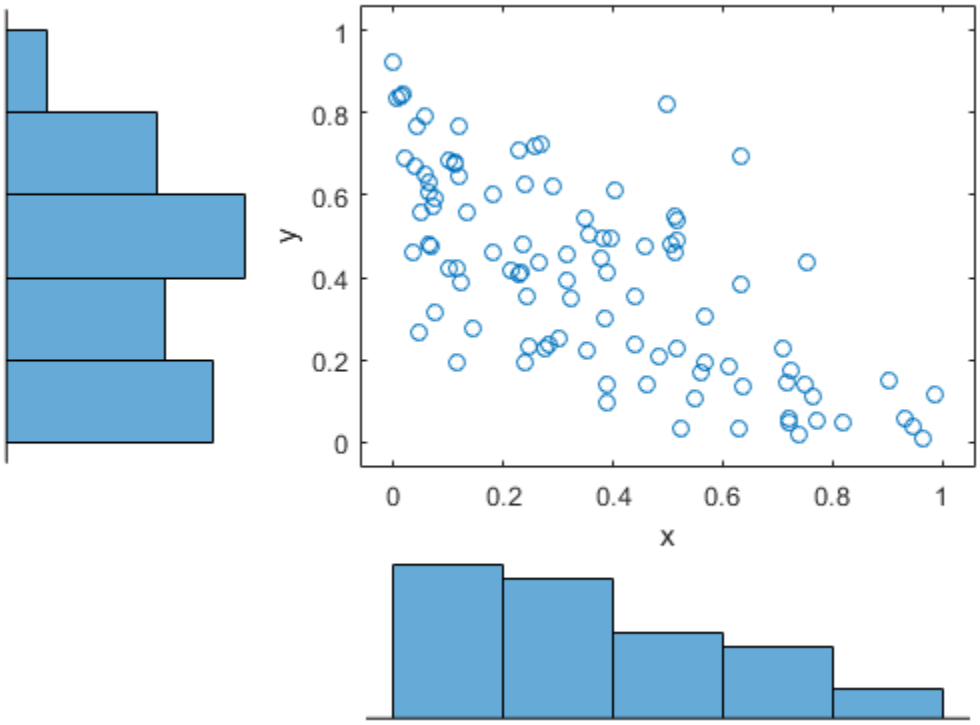
The histograms show that the data in each column of the copula has a marginal uniform distribution. The scatterplot shows that the data in the two columns is negatively correlated.

Use the inverse cdf function `betainv` to transform each column of the uniform marginal distributions into random numbers from a beta distribution. In the first column, the first shape parameter A is equal to 1, and a second shape parameter B is equal to 2. In the second column, the first shape parameter A is equal to 1.5, and a second shape parameter B is equal to 2.

```
b = [betainv(u(:,1),1,2), betainv(u(:,2),1.5,2)];
```

Create a `scatterhist` plot to visualize the correlated beta distribution data.

```
figure
scatterhist(b(:,1),b(:,2))
```



The histograms show the marginal beta distributions for each variable. The scatterplot shows the negative correlation.

Verify that the sample has a rank correlation approximately equal to the initial value for Kendall's *tau*.

```
tau_sample = corr(b,'type','kendall')
```

```
tau_sample = 2×2

    1.0000   -0.5135
   -0.5135    1.0000
```

The sample rank correlation of -0.5135 is approximately equal to the -0.5 initial value for *tau*.

Input Arguments

`r` — Copula rank correlation

scalar value | matrix of scalar values

Copula rank correlation, returned as a scalar value or matrix of scalar values.

- If `r` is a scalar correlation coefficient, then `rho` is a scalar correlation coefficient corresponding to a bivariate copula.
- If `r` is a *p*-by-*p* correlation matrix, then `rho` is a *p*-by-*p* correlation matrix.

If the copula is specified as one of the bivariate Archimedean copula types (`'Clayton'`, `'Frank'`, or `'Gumbel'`), then `r` is a scalar value.

`nu` — Degrees of freedom
positive integer value

Degrees of freedom for the *t* copula, specified as a positive integer value.

Data Types: `single` | `double`

`family` — Bivariate Archimedean copula family
`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'type','Spearman'` computes Spearman's rank correlation.

`type` — Type of rank correlation
`'Kendall'` (default) | `'Spearman'`

Type of rank correlation, specified as the comma-separated pair consisting of `'type'` and one of the following.

- `'Kendall'` — Indicates that the input value for `r` is a Kendall's *tau* correlation value
- `'Spearman'` — Indicates that the input value for `r` is a Spearman's *rho* rank correlation value

`copulaparam` uses an approximation to Spearman's rank correlation for copula families that do not have an existing analytic formula. The approximation is based on a smooth fit to values computed at discrete values of the copula parameters. For a *t* copula, the approximation is accurate for degrees of freedom larger than 0.05.

Example: `'type','Spearman'`

Output Arguments

`rho` — Linear correlation parameter
scalar value | matrix of scalar values

Linear correlation parameter, returned as a scalar value or matrix of scalar values.

- If `r` is a scalar correlation coefficient, then `rho` is a scalar correlation coefficient corresponding to a bivariate copula.
- If `r` is a *p*-by-*p* correlation matrix, then `rho` is a *p*-by-*p* correlation matrix.

alpha — Bivariate Archimedean copula parameter
scalar value

Bivariate Archimedean copula parameter, returned as a scalar value. Permitted values for `alpha` depend on the specified copula family.

Copula Family	Permitted Alpha Values
`'Clayton'`	[0,∞)
`'Frank'`	(-∞,∞)
`'Gumbel'`	[1,∞)

Data Types: `single` | `double`

See Also

`copulacdf` | `copulafit` | `copulapdf` | `copularnd` | `copulastat` | `ecdf`

Topics

“Generate Correlated Data Using Rank Correlation” on page 5-108
“Copulas: Generate Correlated Samples” on page 5-121

Introduced in R2006a

copulapdf

Copula probability density function

Syntax

```
y = copulapdf('Gaussian',u,rho)

y = copulapdf('t',u,rho,nu)

y = copulapdf(family,u,alpha)
```

Description

`y = copulapdf('Gaussian',u,rho)` returns the probability density of the Gaussian copula with linear correlation parameters, `rho`, evaluated at the points in `u`.

`y = copulapdf('t',u,rho,nu)` returns the probability density of the *t* copula with linear correlation parameters, `rho`, and degrees of freedom parameter, `nu`, evaluated at the points in `u`.

`y = copulapdf(family,u,alpha)` returns the probability density of the bivariate Archimedean copula of the type specified by `family`, with scalar parameter, `alpha`, evaluated at the points in `u`.

Examples

Compute the Clayton Copula pdf

Define two 10-by-10 matrices containing the values at which to compute the pdf.

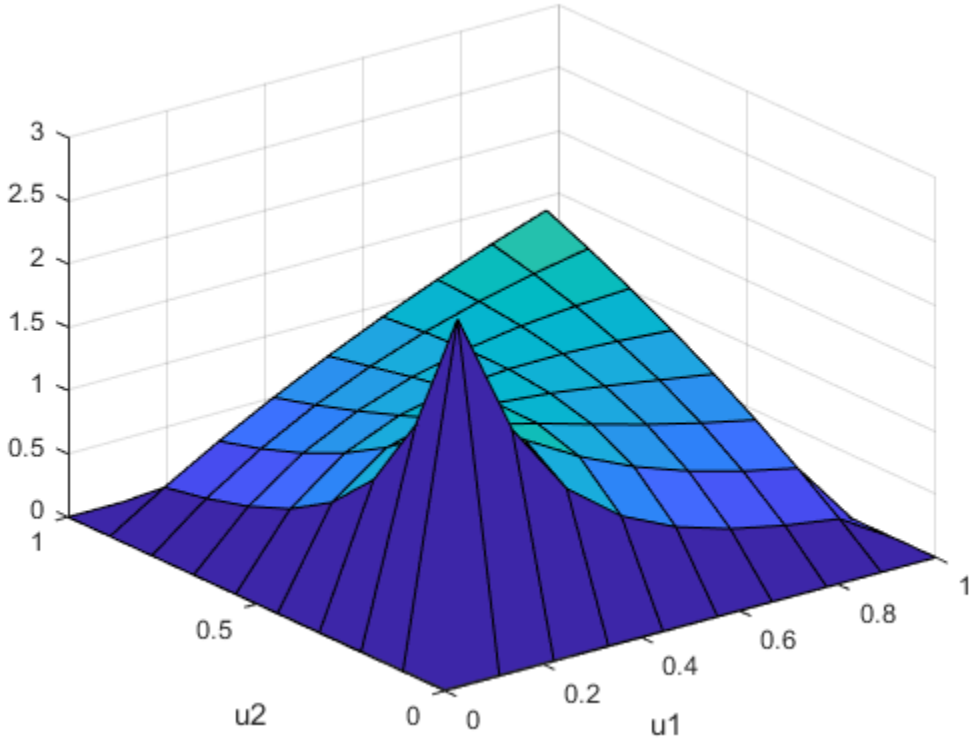
```
u = linspace(0,1,10);
[u1,u2] = meshgrid(u,u);
```

Compute the pdf of a Clayton copula that has an alpha parameter equal to 1, at the values in `u`.

```
y = copulapdf('Clayton',[u1(:),u2(:)],1);
```

Plot the pdf as a surface, and label the axes.

```
surf(u1,u2,reshape(y,10,10))
xlabel('u1')
ylabel('u2')
```



Input Arguments

u — Values at which to evaluate pdf

matrix of scalar values in the range [0,1]

Values at which to evaluate the pdf, specified as a matrix of scalar values in the range [0,1]. If `u` is an *n*-by-*p* matrix, then its values represent *n* points in the *p*-dimensional unit hypercube. If `u` is an *n*-by-2 matrix, then its values represent *n* points in the unit square.

If you specify a bivariate Archimedean copula type (`'Clayton'`, `'Frank'`, or `'Gumbel'`), then `u` must be an *n*-by-2 matrix.

Data Types: `single` | `double`

rho — Linear correlation parameters

scalar values | matrix of scalar values

Linear correlation parameters for the copula, specified as a scalar value or matrix of scalar values.

- If `u` is an *n*-by-*p* matrix, then `rho` is a *p*-by-*p* correlation matrix.
- If `u` is an *n*-by-2 matrix, then `rho` can be a scalar correlation coefficient.

Data Types: `single` | `double`

`nu` — Degrees of freedom
positive integer value

Degrees of freedom for the *t* copula, specified as a positive integer value.

Data Types: `single` | `double`

`family` — Bivariate Archimedean copula family
`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

`alpha` — Bivariate Archimedean copula parameter
scalar value

Bivariate Archimedean copula parameter, specified as a scalar value. Permitted values for `alpha` depend on the specified copula family.

Copula Family	Permitted Alpha Values
`'Clayton'`	[0,∞)
`'Frank'`	(-∞,∞)
`'Gumbel'`	[1,∞)

Data Types: `single` | `double`

Output Arguments

`y` — Probability density function
vector of scalar values

Probability density function, evaluated at the values in `u`, returned as a vector of scalar values.

See Also

`copulacdf` | `copulaparam` | `copularnd` | `copulastat`

Topics
“Generate Correlated Data Using Rank Correlation” on page 5-108
“Copulas: Generate Correlated Samples” on page 5-121

Introduced in R2006a

copulastat

Copula rank correlation

Syntax

```
r = copulastat('Gaussian',rho)

r = copulastat('t',rho,nu)

r = copulastat(family,alpha)

r = copulastat( ___ ,Name,Value)
```

Description

`r = copulastat('Gaussian',rho)` returns the Kendall's rank correlation, `r`, that corresponds to a Gaussian copula with linear correlation parameters `rho`.

`r = copulastat('t',rho,nu)` returns the Kendall's rank correlation, `r`, that corresponds to a *t* copula with linear correlation parameters, `rho`, and degrees of freedom parameter, `nu`.

`r = copulastat(family,alpha)` returns the Kendall's rank correlation, `r`, that corresponds to a bivariate Archimedean copula that has the type specified by `family` and scalar parameter `alpha`.

`r = copulastat( ___ ,Name,Value)` returns the copula rank correlation with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can return Spearman's *rho* rank correlation.

Examples

Compute the Gaussian Copula Rank Correlation

Compute the rank correlation for a Gaussian copula with the specified linear correlation parameter `rho`.

```
rho = -.7071;
tau = copulastat('gaussian',rho)
```

```
tau = -0.5000
```

Use the copula to generate dependent random values from a beta distribution that has parameters a and b equal to 2.

```
rng default  % For reproducibility
u = copularnd('gaussian',rho,100);
b = betainv(u,2,2);
```

Verify that the sample has a rank correlation approximately equal to `tau`.

```
tau_sample = corr(b,'type','k')
```

```
tau_sample = 2×2

    1.0000   -0.5135
   -0.5135    1.0000
```

Input Arguments

`rho` — Linear correlation parameters

scalar values | matrix of scalar values

Linear correlation parameters for the copula, specified as a scalar value or matrix of scalar values.

- If `rho` is a scalar correlation coefficient, then `r` is a scalar correlation coefficient corresponding to a bivariate copula.
- If `rho` is a *p*-by-*p* correlation matrix, then `r` is a *p*-by-*p* correlation matrix.

Data Types: `single` | `double`

`nu` — Degrees of freedom

positive integer value

Degrees of freedom for the *t* copula, specified as a positive integer value.

Data Types: `single` | `double`

`family` — Bivariate Archimedean copula family

`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

`alpha` — Bivariate Archimedean copula parameter

scalar value

Bivariate Archimedean copula parameter, specified as a scalar value. Permitted values for `alpha` depend on the specified copula family.

Copula Family	Permitted Alpha Values
`'Clayton'`	[0,∞)
`'Frank'`	(-∞,∞)
`'Gumbel'`	[1,∞)

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'type','Spearman'` computes Spearman's rank correlation.

`type` — Type of rank correlation

`'Kendall'` (default) | `'Spearman'`

Type of rank correlation, specified as the comma-separated pair consisting of `'type'` and one of the following.

- `'Kendall'` — Compute Kendall's *tau*.
- `'Spearman'` — Compute Spearman's *rho* (rank correlation).

`copulastat` uses an approximation to Spearman's rank correlation for copula families that do not have an existing analytic formula. The approximation is based on a smooth fit to values computed at discrete values of the copula parameters. For a *t* copula, the approximation is accurate for degrees of freedom larger than 0.05.

Example: `'type','Spearman'`

Output Arguments

`r` — Copula rank correlation

scalar value | matrix of scalar values

Copula rank correlation, returned as a scalar value or matrix of scalar values.

- If `rho` is a scalar correlation coefficient, then `r` is a scalar correlation coefficient corresponding to a bivariate copula.
- If `rho` is a *p*-by-*p* correlation matrix, then `r` is a *p*-by-*p* correlation matrix.

See Also

`copulacdf` | `copulaparam` | `copulapdf` | `copularnd`

Topics

"Generate Correlated Data Using Rank Correlation" on page 5-108
"Copulas: Generate Correlated Samples" on page 5-121

Introduced in R2006a

copularnd

Copula random numbers

Syntax

```
u = copularnd('Gaussian',rho,n)

u = copularnd('t',rho,nu,n)

u = copularnd(family,alpha,n)
```

Description

`u = copularnd('Gaussian',rho,n)` returns `n` random vectors generated from a Gaussian copula with linear correlation parameters `rho`.

`u = copularnd('t',rho,nu,n)` returns `n` random vectors generated from a *t* copula with linear correlation parameters `rho` and degrees of freedom `nu`.

`u = copularnd(family,alpha,n)` returns `n` random vectors generated from a bivariate Archimedean copula that has the type specified by `family` and the scalar parameter `alpha`.

Examples

Generate Correlated Data Using the Inverse cdf

Generate correlated random data from a beta distribution using a bivariate Gaussian copula with Kendall's *tau* rank correlation equal to -0.5.

Compute the linear correlation parameter from the rank correlation value.

```
rng default  % For reproducibility
tau = -0.5;
rho = copulaparam('Gaussian',tau)
```

```
rho = -0.7071
```

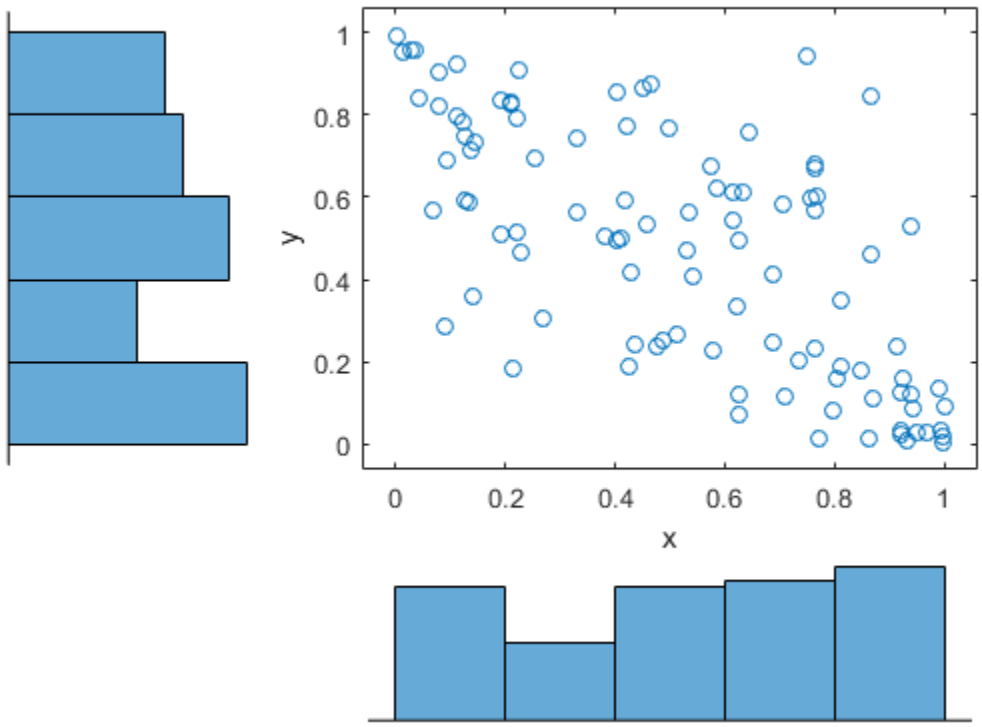
Use a Gaussian copula to generate a two-column matrix of dependent random values.

```
u = copularnd('gaussian',rho,100);
```

Each column contains 100 random values between 0 and 1, inclusive, sampled from a continuous uniform distribution.

Create a `scatterhist` plot to visualize the random numbers generated using the copula.

```
figure
scatterhist(u(:,1),u(:,2))
```



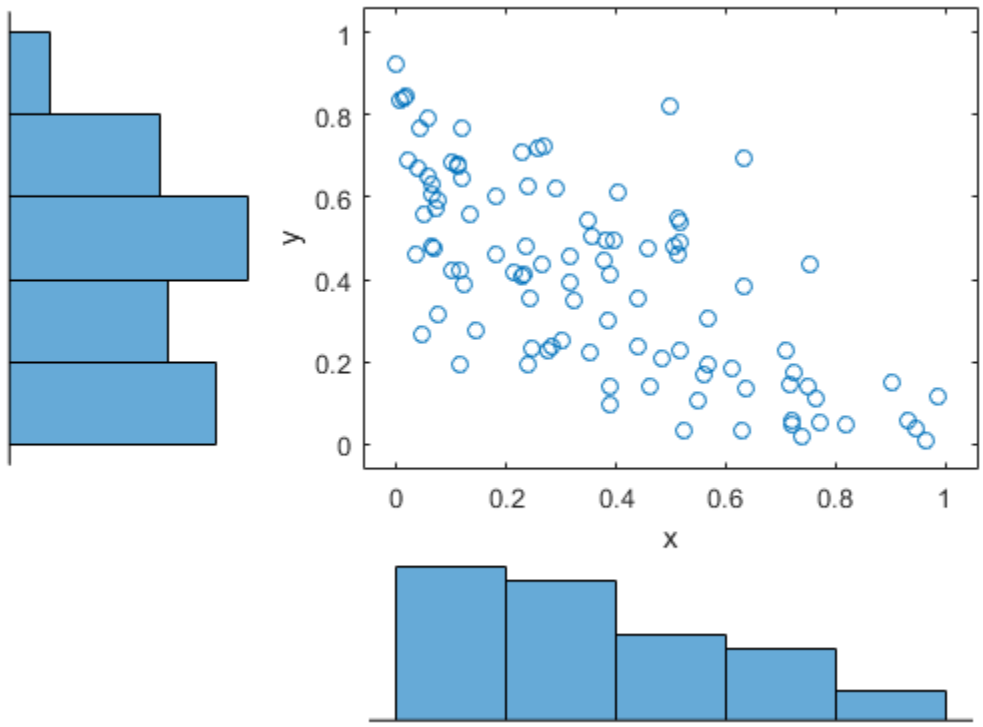
The histograms show that the data in each column of the copula has a marginal uniform distribution. The scatterplot shows that the data in the two columns is negatively correlated.

Use the inverse cdf function `betainv` to transform each column of the uniform marginal distributions into random numbers from a beta distribution. In the first column, the first shape parameter A is equal to 1, and a second shape parameter B is equal to 2. In the second column, the first shape parameter A is equal to 1.5, and a second shape parameter B is equal to 2.

```
b = [betainv(u(:,1),1,2), betainv(u(:,2),1.5,2)];
```

Create a `scatterhist` plot to visualize the correlated beta distribution data.

```
figure
scatterhist(b(:,1),b(:,2))
```



The histograms show the marginal beta distributions for each variable. The scatterplot shows the negative correlation.

Verify that the sample has a rank correlation approximately equal to the initial value for Kendall's *tau*.

```
tau_sample = corr(b,'type','kendall')
```

```
tau_sample = 2×2

    1.0000   -0.5135
   -0.5135    1.0000
```

The sample rank correlation of -0.5135 is approximately equal to the -0.5 initial value for *tau*.

Input Arguments

`rho` — Linear correlation parameters

scalar values | matrix of scalar values

Linear correlation parameters for the copula, specified as a scalar value or matrix of scalar values.

- If `rho` is a *p*-by-*p* correlation matrix, then the output argument `u` is an *n*-by-*p* matrix.
- If `rho` is a scalar correlation coefficient, then the output argument `u` is an *n*-by-2 matrix.

Data Types: `single` | `double`

n — Number of random vectors to return
positive scalar value

Number of random vectors to return, specified as a positive scalar value.

- If you specify the copula type as `'Gaussian'` or `'t'`, and `rho` is a *p*-by-*p* correlation matrix, then `u` is an *n*-by-*p* matrix.
- If you specify the copula type as `'Gaussian'` or `'t'`, and `rho` is a scalar correlation coefficient, then `u` is an *n*-by-2 matrix.
- If you specify the copula type as `'Clayton'`, `'Frank'`, or `'Gumbel'`, then `u` is an *n*-by-2 matrix.

Data Types: `single` | `double`

nu — Degrees of freedom
positive integer value

Degrees of freedom for the *t* copula, specified as a positive integer value.

Data Types: `single` | `double`

family — Bivariate Archimedean copula family
`'Clayton'` | `'Frank'` | `'Gumbel'`

Bivariate Archimedean copula family, specified as one of the following.

`'Clayton'`	Clayton copula
`'Frank'`	Frank copula
`'Gumbel'`	Gumbel copula

alpha — Bivariate Archimedean copula parameter
scalar value

Bivariate Archimedean copula parameter, specified as a scalar value. Permitted values for `alpha` depend on the specified copula family.

Copula Family	Permitted Alpha Values
`'Clayton'`	[0,∞)
`'Frank'`	(-∞,∞)
`'Gumbel'`	[1,∞)

Data Types: `single` | `double`

Output Arguments

u — Copula random numbers
matrix of scalar values

Copula random numbers, returned as a matrix of scalar values. Each column of `u` is a sample from a `Uniform(0,1)` marginal distribution.

- If you specify the copula type as `'Gaussian'` or `'t'`, and `rho` is a *p*-by-*p* correlation matrix, then `u` is an *n*-by-*p* matrix.

- If you specify the copula type as `'Gaussian'` or `'t'`, and `rho` is a scalar correlation coefficient, then `u` is an *n*-by-2 matrix.
- If you specify the copula type as `'Clayton'`, `'Frank'`, or `'Gumbel'`, then `u` is an *n*-by-2 matrix.

See Also

`copulacdf` | `copulaparam` | `copulapdf` | `copulastat`

Topics

“Generate Correlated Data Using Rank Correlation” on page 5-108
“Copulas: Generate Correlated Samples” on page 5-121

Introduced in R2006a

cordexch

Coordinate exchange

Syntax

```
dCE = cordexch(nfactors,nruns)
[dCE,X] = cordexch(nfactors,nruns)
[dCE,X] = cordexch(nfactors,nruns,'model')
[dCE,X] = cordexch(...,'name',value)
```

Description

`dCE = cordexch(nfactors,nruns)` uses a coordinate-exchange algorithm to generate a *D*-optimal design `dCE` with `nruns` runs (the rows of `dCE`) for a linear additive model with `nfactors` factors (the columns of `dCE`). The model includes a constant term.

`[dCE,X] = cordexch(nfactors,nruns)` also returns the associated design matrix `X`, whose columns are the model terms evaluated at each treatment (row) of `dCE`.

`[dCE,X] = cordexch(nfactors,nruns,'model')` uses the linear regression model specified in `model`. `model` is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

The order of the columns of `X` for a full quadratic model with *n* terms is:

1. The constant term
2. The linear terms in order 1, 2, ..., *n*
3. The interaction terms in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n* – 1, *n*)
4. The squared terms in order 1, 2, ..., *n*

Other models use a subset of these terms, in the same order.

Alternatively, `model` can be a matrix specifying polynomial terms of arbitrary order. In this case, `model` should have one column for each factor and one row for each term in the model. The entries in any row of `model` are powers for the factors in the columns. For example, if a model has factors `X1`, `X2`, and `X3`, then a row `[0 1 2]` in `model` specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in `model` specifies a constant term, which can be omitted.

`[dCE,X] = cordexch(...,'name',value)` specifies one or more optional name/value pairs for the design. Valid parameters and their values are listed in the following table. Specify `name` inside single quotes.

name	Value
`bounds`	Lower and upper bounds for each factor, specified as a 2-by-`nfactors` matrix. Alternatively, this value can be a cell array containing `nfactors` elements, each element specifying the vector of allowable values for the corresponding factor.
`categorical`	Indices of categorical predictors.
`display`	Either `'on'` or `'off'` to control display of the iteration counter. The default is `'on'`.
`excludefun`	Handle to a function that excludes undesirable runs. If the function is *f*, it must support the syntax $b = f(S)$, where S is a matrix of treatments with `nfactors` columns and b is a vector of Boolean values with the same number of rows as S. $b(i)$ is true if the method should exclude ith row S.
`init`	Initial design as a `nruns`-by-`nfactors` matrix. The default is a randomly selected set of points.
`levels`	Vector of number of levels for each factor. Not used when `bounds` is specified as a cell array.
`maxiter`	Maximum number of iterations. The default is `10`.
`tries`	Number of times to try to generate a design from a new starting point. The algorithm uses random points for each try, except possibly the first. The default is `1`.

`options`	A structure that specifies whether to run in parallel, and specifies the random stream or streams. This option requires Parallel Computing Toolbox.
	
	Create the `options` structure with `statset`. Structure fields:
	
	• `UseParallel` — Set to `true` to compute in parallel. Default is `false`.
	• `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. Default is `false`. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
	• `Streams` — A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `cordexch` uses the default stream or streams. If you choose to specify `Streams`, use a single object except in the case
	
	  • `UseParallel` is `true`
	  • `UseSubstreams` is `false`
	
	  In that case, use a cell array the same size as the Parallel pool.

Examples

Suppose you want a design to estimate the parameters in the following three-factor, seven-term interaction model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon$$

Use `cordexch` to generate a *D*-optimal design with seven runs:

```
nfactors = 3;
nruns = 7;
```

```
[dCE,X] = cordexch(nfactors,nruns,'interaction','tries',10)
dCE =
    -1     1     1
    -1    -1    -1
     1     1     1
    -1     1    -1
     1    -1     1
     1    -1    -1
    -1    -1     1
X =
     1    -1     1     1    -1    -1     1
     1    -1    -1    -1     1     1     1
     1     1     1     1     1     1     1
     1    -1     1    -1    -1     1    -1
     1     1    -1     1    -1     1    -1
     1     1    -1    -1    -1    -1     1
     1    -1    -1     1     1    -1    -1
```

Columns of the design matrix `X` are the model terms evaluated at each row of the design `dCE`. The terms appear in order from left to right: constant term, linear terms (1, 2, 3), interaction terms (12, 13, 23). Use `X` to fit the model, as described in “Linear Regression” on page 11-9, to response data measured at the design points in `dCE`.

Algorithms

Both `cordexch` and `rowexch` use iterative search algorithms. They operate by incrementally changing an initial design matrix X to increase $D = |X^TX|$ at each step. In both algorithms, there is randomness built into the selection of the initial design and into the choice of the incremental changes. As a result, both algorithms may return locally, but not globally, D-optimal designs. Run each algorithm multiple times and select the best result for your final design. Both functions have a `'tries'` parameter that automates this repetition and comparison.

Unlike the row-exchange algorithm used by `rowexch`, `cordexch` does not use a candidate set. (Or rather, the candidate set is the entire design space.) At each step, the coordinate-exchange algorithm exchanges a single element of X with a new element evaluated at a neighboring point in design space. The absence of a candidate set reduces demands on memory, but the smaller scale of the search means that the coordinate-exchange algorithm is more likely to become trapped in a local minimum.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`daugment` | `dcovary` | `rowexch`

Introduced before R2006a

corr

Linear or rank correlation

Syntax

```
rho = corr(X)
rho = corr(X,Y)
[rho,pval] = corr(X,Y)
[rho,pval] = corr( ___ ,Name,Value)
```

Description

`rho = corr(X)` returns a matrix of the pairwise linear correlation coefficient between each pair of columns in the input matrix `X`.

`rho = corr(X,Y)` returns a matrix of the pairwise correlation coefficient between each pair of columns in the input matrices `X` and `Y`.

`[rho,pval] = corr(X,Y)` also returns `pval`, a matrix of *p*-values for testing the hypothesis of no correlation against the alternative hypothesis of a nonzero correlation.

`[rho,pval] = corr( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input arguments in the previous syntaxes. For example, `'Type','Kendall'` specifies computing Kendall's tau correlation coefficient.

Examples

Find Correlation Between Two Matrices

Find the correlation between two matrices and compare it to the correlation between two column vectors.

Generate sample data.

```
rng('default')
X = randn(30,4);
Y = randn(30,4);
```

Introduce correlation between column two of the matrix `X` and column four of the matrix `Y`.

```
Y(:,4) = Y(:,4)+X(:,2);
```

Calculate the correlation between columns of `X` and `Y`.

```
[rho,pval] = corr(X,Y)
```

```
rho = 4×4

   -0.1686   -0.0363    0.2278    0.3245
    0.3022    0.0332   -0.0866    0.7653
```

```
   -0.3632   -0.0987   -0.0200   -0.3693
   -0.1365   -0.1804    0.0853    0.0279


pval = 4×4

    0.3731    0.8489    0.2260    0.0802
    0.1045    0.8619    0.6491    0.0000
    0.0485    0.6039    0.9166    0.0446
    0.4721    0.3400    0.6539    0.8837
```

As expected, the correlation coefficient between column two of `X` and column four of `Y`, `rho(2,4)`, is the highest, and it represents a high positive correlation between the two columns. The corresponding *p*-value, `pval(2,4)`, is zero to the four digits shown. Because the *p*-value is less than the significance level of `0.05`, it indicates rejection of the hypothesis that no correlation exists between the two columns.

Calculate the correlation between `X` and `Y` using `corrcoef`.

```
[r,p] = corrcoef(X,Y)
```

```
r = 2×2

    1.0000   -0.0329
   -0.0329    1.0000


p = 2×2

    1.0000    0.7213
    0.7213    1.0000
```

The MATLAB® function `corrcoef`, unlike the `corr` function, converts the input matrices `X` and `Y` into column vectors, `X(:)` and `Y(:)`, before computing the correlation between them. Therefore, the introduction of correlation between column two of matrix `X` and column four of matrix `Y` no longer exists, because those two columns are in different sections of the converted column vectors.

The value of the off-diagonal elements of `r`, which represents the correlation coefficient between `X` and `Y`, is low. This value indicates little to no correlation between `X` and `Y`. Likewise, the value of the off-diagonal elements of `p`, which represents the *p*-value, is much higher than the significance level of `0.05`. This value indicates that not enough evidence exists to reject the hypothesis of no correlation between `X` and `Y`.

Test Alternative Hypotheses for Correlation

Test alternative hypotheses for positive, negative, and nonzero correlation between the columns of two matrices. Compare values of the correlation coefficient and *p*-value in each case.

Generate sample data.

```
rng('default')
X = randn(50,4);
Y = randn(50,4);
```

Introduce positive correlation between column one of the matrix `X` and column four of the matrix `Y`.

```
Y(:,4) = Y(:,4)+0.7*X(:,1);
```

Introduce negative correlation between column two of `X` and column two of `Y`.

```
Y(:,2) = Y(:,2)-2*X(:,2);
```

Test the alternative hypothesis that the correlation is greater than zero.

```
[rho,pval] = corr(X,Y,'Tail','right')
```

```
rho = 4×4

    0.0627   -0.1438   -0.0035    0.7060
   -0.1197   -0.8600   -0.0440    0.1984
   -0.1119    0.2210   -0.3433    0.1070
   -0.3526   -0.2224    0.1023    0.0374


pval = 4×4

    0.3327    0.8405    0.5097    0.0000
    0.7962    1.0000    0.6192    0.0836
    0.7803    0.0615    0.9927    0.2298
    0.9940    0.9397    0.2398    0.3982
```

As expected, the correlation coefficient between column one of `X` and column four of `Y`, `rho(1,4)`, has the highest positive value, representing a high positive correlation between the two columns. The corresponding *p*-value, `pval(1,4)`, is zero to the four digits shown, which is lower than the significance level of `0.05`. These results indicate rejection of the null hypothesis that no correlation exists between the two columns and lead to the conclusion that the correlation is greater than zero.

Test the alternative hypothesis that the correlation is less than zero.

```
[rho,pval] = corr(X,Y,'Tail','left')
```

```
rho = 4×4

    0.0627   -0.1438   -0.0035    0.7060
   -0.1197   -0.8600   -0.0440    0.1984
   -0.1119    0.2210   -0.3433    0.1070
   -0.3526   -0.2224    0.1023    0.0374


pval = 4×4

    0.6673    0.1595    0.4903    1.0000
    0.2038    0.0000    0.3808    0.9164
    0.2197    0.9385    0.0073    0.7702
    0.0060    0.0603    0.7602    0.6018
```

As expected, the correlation coefficient between column two of `X` and column two of `Y`, `rho(2,2)`, has the negative number with the largest absolute value (`-0.86`), representing a high negative correlation between the two columns. The corresponding *p*-value, `pval(2,2)`, is zero to the four digits shown, which is lower than the significance level of `0.05`. Again, these results indicate rejection of the null hypothesis and lead to the conclusion that the correlation is less than zero.

Test the alternative hypothesis that the correlation is not zero.

```
[rho,pval] = corr(X,Y)
```

```
rho = 4×4

    0.0627   -0.1438   -0.0035    0.7060
   -0.1197   -0.8600   -0.0440    0.1984
   -0.1119    0.2210   -0.3433    0.1070
   -0.3526   -0.2224    0.1023    0.0374


pval = 4×4

    0.6654    0.3190    0.9807    0.0000
    0.4075    0.0000    0.7615    0.1673
    0.4393    0.1231    0.0147    0.4595
    0.0120    0.1206    0.4797    0.7964
```

The *p*-values, `pval(1,4)` and `pval(2,2)`, are both zero to the four digits shown. Because the *p*-values are lower than the significance level of `0.05`, the correlation coefficients `rho(1,4)` and `rho(2,2)` are significantly different from zero. Therefore, the null hypothesis is rejected; the correlation is not zero.

Input Arguments

X — Input matrix

matrix

Input matrix, specified as an *n*-by-*k* matrix. The rows of `X` correspond to observations, and the columns correspond to variables.

Example: `X = randn(10,5)`

Data Types: `single` | `double`

Y — Input matrix

matrix

Input matrix, specified as an *n*-by-k_2 matrix when `X` is specified as an *n*-by-k_1 matrix. The rows of `Y` correspond to observations, and the columns correspond to variables.

Example: `Y = randn(20,7)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `corr(X,Y,'Type','Kendall','Rows','complete')` returns Kendall's tau correlation coefficient using only the rows that contain no missing values.

Type — Type of correlation
'Pearson' (default) | 'Kendall' | 'Spearman'

Type of correlation, specified as the comma-separated pair consisting of 'Type' and one of these values.

Value	Description
'Pearson'	"Pearson's Linear Correlation Coefficient" on page 33-993
'Kendall'	"Kendall's Tau Coefficient" on page 33-993
'Spearman'	"Spearman's Rho" on page 33-994

corr computes the *p*-values for Pearson's correlation using a Student's *t* distribution for a transformation of the correlation. This correlation is exact when X and Y come from a normal distribution. corr computes the *p*-values for Kendall's tau and Spearman's rho using either the exact permutation distributions (for small sample sizes) or large-sample approximations.

Example: 'Type','Spearman'

Rows — Rows to use in computation
'all' (default) | 'complete' | 'pairwise'

Rows to use in computation, specified as the comma-separated pair consisting of 'Rows' and one of these values.

Value	Description
'all'	Use all rows of the input regardless of missing values (NaNs).
'complete'	Use only rows of the input with no missing values.
'pairwise'	Compute rho(i,j) using rows with no missing values in column i or j.

The 'complete' value, unlike the 'pairwise' value, always produces a positive definite or positive semidefinite rho. Also, the 'complete' value generally uses fewer observations to estimate rho when rows of the input (X or Y) contain missing values.

Example: 'Rows','pairwise'

Tail — Alternative hypothesis
'both' (default) | 'right' | 'left'

Alternative hypothesis, specified as the comma-separated pair consisting of 'Tail' and one of the values in the table. 'Tail' specifies the alternative hypothesis against which to compute *p*-values for testing the hypothesis of no correlation.

Value	Description
'both'	Test the alternative hypothesis that the correlation is not 0.
'right'	Test the alternative hypothesis that the correlation is greater than 0
'left'	Test the alternative hypothesis that the correlation is less than 0.

corr computes the *p*-values for the two-tailed test by doubling the more significant of the two one-tailed *p*-values.

Example: 'Tail','left'

Output Arguments

`rho` — Pairwise linear correlation coefficient

matrix

Pairwise linear correlation coefficient, returned as a matrix.

- If you input only a matrix `X`, `rho` is a symmetric k-by-k matrix, where k is the number of columns in `X`. The entry `rho(a,b)` is the pairwise linear correlation coefficient between column a and column b in `X`.
- If you input matrices `X` and `Y`, `rho` is a k_1-by-k_2 matrix, where k_1 and k_2 are the number of columns in `X` and `Y`, respectively. The entry `rho(a,b)` is the pairwise linear correlation coefficient between column a in `X` and column b in `Y`.

`pval` — *p*-values

matrix

p-values, returned as a matrix. Each element of `pval` is the *p*-value for the corresponding element of `rho`.

If `pval(a,b)` is small (less than `0.05`), then the correlation `rho(a,b)` is significantly different from zero.

More About

Pearson's Linear Correlation Coefficient

Pearson's linear correlation coefficient is the most commonly used linear correlation coefficient. For column X_a in matrix X and column Y_b in matrix Y, having means $\bar{X}_a = \sum_{i=1}^{n} (X_{a,i})/n$, and $\bar{Y}_b = \sum_{j=1}^{n} (X_{b,j})/n$, Pearson's linear correlation coefficient *rho(a,b)* is defined as:

$$rho(a,b) = \frac{\sum_{i=1}^{n} (X_{a,i} - \bar{X}_a)(Y_{b,i} - \bar{Y}_b)}{\left\{ \sum_{i=1}^{n} (X_{a,i} - \bar{X}_a)^2 \sum_{j=1}^{n} (Y_{b,j} - \bar{Y}_b)^2 \right\}^{1/2}},$$

where n is the length of each column.

Values of the correlation coefficient can range from `−1` to `+1`. A value of `−1` indicates perfect negative correlation, while a value of `+1` indicates perfect positive correlation. A value of `0` indicates no correlation between the columns.

Kendall's Tau Coefficient

Kendall's tau is based on counting the number of (i,j) pairs, for $i<j$, that are concordant—that is, for which $X_{a,i} - X_{a,j}$ and $Y_{b,i} - Y_{b,j}$ have the same sign. The equation for Kendall's tau includes an adjustment for ties in the normalizing constant and is often referred to as tau-b.

For column X_a in matrix X and column Y_b in matrix Y, Kendall's tau coefficient is defined as:

$$\tau = \frac{2K}{n(n-1)},$$

where $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j})$, and

$$\xi^*(X_{a,i}, X_{a,j}, Y_{b,i}, Y_{b,j}) = \begin{cases} 1 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) > 0 \\ 0 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) = 0 \\ -1 & \text{if } (X_{a,i} - X_{a,j})(Y_{b,i} - Y_{b,j}) < 0 \end{cases}.$$

Values of the correlation coefficient can range from `-1` to `+1`. A value of `-1` indicates that one column ranking is the reverse of the other, while a value of `+1` indicates that the two rankings are the same. A value of `0` indicates no relationship between the columns.

Spearman's Rho

Spearman's rho is equivalent to “Pearson's Linear Correlation Coefficient” on page 33-993 applied to the rankings of the columns X_a and Y_b.

If all the ranks in each column are distinct, the equation simplifies to:

$$rho(a,b) = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

where d is the difference between the ranks of the two columns, and n is the length of each column.

Tips

The difference between `corr(X,Y)` and the MATLAB function `corrcoef(X,Y)` is that `corrcoef(X,Y)` returns a matrix of correlation coefficients for two column vectors `X` and `Y`. If `X` and `Y` are not column vectors, `corrcoef(X,Y)` converts them to column vectors.

References

[1] Gibbons, J.D. *Nonparametric Statistical Inference*. 2nd ed. M. Dekker, 1985.

[2] Hollander, M., and D.A. Wolfe. *Nonparametric Statistical Methods*. Wiley, 1973.

[3] Kendall, M.G. *Rank Correlation Methods*. Griffin, 1970.

[4] Best, D.J., and D.E. Roberts. "Algorithm AS 89: The Upper Tail Probabilities of Spearman's rho." *Applied Statistics*, 24:377-379.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with the limitation:

Only the `'Pearson'` type is supported.

For more information, see “Tall Arrays for Out-of-Memory Data”.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`corrcoef` | `corrcov` | `partialcorr` | `tiedrank`

Introduced before R2006a

corrcov

Convert covariance matrix to correlation matrix

Syntax

```
R = corrcov(C)
[R,sigma] = corrcov(C)
```

Description

`R = corrcov(C)` returns the correlation matrix `R` corresponding to the covariance matrix `C`.

`[R,sigma] = corrcov(C)` also returns `sigma`, a vector of standard deviations.

Examples

Compare Correlation Matrices Obtained by Two Different Methods

Compare the correlation matrix obtained by applying `corrcov` on a covariance matrix with the correlation matrix obtained by direct computation using `corrcoef` on an input matrix.

Load the `hospital` data set and create a matrix containing the `Weight` and `BloodPressure` measurements. Note that `hospital.BloodPressure` has two columns of data.

```
load hospital
X = [hospital.Weight hospital.BloodPressure];
```

Compute the covariance matrix.

```
C = cov(X)
```

```
C = 3×3

  706.0404   27.7879   41.0202
   27.7879   45.0622   23.8194
   41.0202   23.8194   48.0590
```

Compute the correlation matrix from the covariance matrix by using `corrcov`.

```
R1 = corrcov(C)
```

```
R1 = 3×3

    1.0000    0.1558    0.2227
    0.1558    1.0000    0.5118
    0.2227    0.5118    1.0000
```

Compute the correlation matrix directly by using `corrcoef`, and then compare `R1` with `R2`.

```
R2 = corrcoef(X)
```

```
R2 = 3×3

    1.0000    0.1558    0.2227
    0.1558    1.0000    0.5118
    0.2227    0.5118    1.0000
```

The correlation matrices `R1` and `R2` are the same.

Find Standard Deviations from Covariance Matrix

Find the vector of standard deviations from the covariance matrix, and show the relationship between the standard deviations and the covariance matrix.

Load the `hospital` data set and create a matrix containing the `Weight`, `BloodPressure`, and `Age` measurements. Note that `hospital.BloodPressure` has two columns of data.

```
load hospital
X = [hospital.Weight hospital.BloodPressure hospital.Age];
```

Compute the covariance matrix of `X`.

```
C = cov(X)
```

```
C = 4×4

  706.0404   27.7879   41.0202   17.5152
   27.7879   45.0622   23.8194    6.4966
   41.0202   23.8194   48.0590    4.0315
   17.5152    6.4966    4.0315   52.0622
```

`C` is square, symmetric, and positive semidefinite. The diagonal elements of `C` are the variances of the four variables in `X`.

Compute the correlation matrix and standard deviations of `X` from the covariance matrix `C`.

```
[R,s1] = corrcov(C)
```

```
R = 4×4

    1.0000    0.1558    0.2227    0.0914
    0.1558    1.0000    0.5118    0.1341
    0.2227    0.5118    1.0000    0.0806
    0.0914    0.1341    0.0806    1.0000


s1 = 4×1

   26.5714
    6.7128
    6.9325
    7.2154
```

Compute the square root of the diagonal elements in `C`, and then compare `s1` with `s2`.

```
s2 = sqrt(diag(C))

s2 = 4×1

   26.5714
    6.7128
    6.9325
    7.2154
```

`s1` and `s2` are equal and correspond to the standard deviation of the variables in `X`.

Input Arguments

`C` — Covariance matrix

matrix

Covariance on page 33-998 matrix, specified as a square, symmetric, and positive semidefinite matrix.

For a matrix X that has N observations (rows) and n random variables (columns), `C` is an n-by-n matrix. The n diagonal elements of `C` are the variances on page 33-999 of the n random variables in X, and a zero diagonal element in `C` indicates a constant variable in X.

Data Types: `single` | `double`

Output Arguments

`R` — Correlation matrix

matrix

Correlation matrix, returned as a matrix that corresponds to the covariance matrix `C`.

Data Types: `single` | `double`

`sigma` — Standard deviations

vector

Standard deviations, returned as an n-by-1 vector.

The elements of `sigma` are the standard deviations of the variables in X, the N-by-n matrix that produces `C`. Row `i` in `sigma` corresponds to the standard deviation of column `i` in X.

Data Types: `single` | `double`

More About

Covariance

For two random variable vectors A and B, the covariance is defined as

$$\mathrm{cov}(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(A_i - \mu_A)^*(B_i - \mu_B)$$

where N is the length of each column, μ_A and μ_B are the mean values of A and B, respectively, and * denotes the complex conjugate.

The *covariance matrix* of two random variables is the matrix of pairwise covariance calculations between each variable,

$$C = \begin{pmatrix} \text{cov}(A,A) & \text{cov}(A,B) \\ \text{cov}(B,A) & \text{cov}(B,B) \end{pmatrix}.$$

For a matrix X, in which each column is a random variable composed of observations, the covariance matrix is the pairwise covariance calculation between each column combination. In other words, $C(i,j) = \text{cov}(X(:,i), X(:,j))$.

Variance

For a random variable vector A composed of N scalar observations, the variance is defined as

$$V = \frac{1}{N-1}\sum_{i=1}^{N} |A_i - \mu|^2$$

where μ is the mean of A,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i .$$

Some definitions of variance use a normalization factor of N instead of $N–1$, but the mean always has the normalization factor N.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cholcov` | `corr` | `corrcoef` | `cov`

Introduced in R2007b

covarianceParameters

Class: `GeneralizedLinearMixedModel`

Extract covariance parameters of generalized linear mixed-effects model

Syntax

```
psi = covarianceParameters(glme)
[psi,dispersion] = covarianceParameters(glme)
[psi,dispersion,stats] = covarianceParameters(glme)
[___] = covarianceParameters(glme,Name,Value)
```

Description

`psi = covarianceParameters(glme)` returns the estimated prior covariance parameters of random-effects predictors in the generalized linear mixed-effects model `glme`.

`[psi,dispersion] = covarianceParameters(glme)` also returns an estimate of the dispersion parameter.

`[psi,dispersion,stats] = covarianceParameters(glme)` also returns a cell array `stats` containing the covariance parameter estimates and related statistics.

`[___] = covarianceParameters(glme,Name,Value)` returns any of the above output arguments using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence level for the confidence limits of covariance parameters.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range [0,1]

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range [0,1]. For a value α, the confidence level is 100 × (1 – α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

Output Arguments

`psi` — Estimated prior covariance parameters

cell array

Estimated prior covariance parameters for the random-effects predictors, returned as a cell array of length R, where R is the number of grouping variables used in the model. `psi{r}` contains the covariance matrix of random effects associated with grouping variable g_r, where $r = 1, 2, ..., R$, The order of grouping variables in `psi` is the same as the order entered when fitting the model. For more information on grouping variables, see “Grouping Variables” on page 2-45.

`dispersion` — Dispersion parameter

scalar value

Dispersion parameter, returned as a scalar value.

`stats` — Covariance parameter estimates and related statistics

cell array

Covariance parameter estimates and related statistics, returned as a cell array of length (R + 1), where R is the number of grouping variables used in the model. The first R cells of `stats` each contain a dataset array with the following columns.

Column Name	Description
`Group`	Grouping variable name
`Name1`	Name of the first predictor variable
`Name2`	Name of the second predictor variable
`Type`	If `Name1` and `Name2` are the same, then `Type` is `std` (standard deviation). If `Name1` and `Name2` are different, then `Type` is `corr` (correlation).
`Estimate`	If `Name1` and `Name2` are the same, then `Estimate` is the standard deviation of the random effect associated with predictor `Name1` or `Name2`. If `Name1` and `Name2` are different, then `Estimate` is the correlation between the random effects associated with predictors `Name1` and `Name2`.
`Lower`	Lower limit of the confidence interval for the covariance parameter
`Upper`	Upper limit of the confidence interval for the covariance parameter

Cell R + 1 contains related statistics for the dispersion parameter.

It is recommended that the presence or absence of covariance parameters in `glme` be tested using the `compare` method, which uses a likelihood ratio test.

When fitting a GLME model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`), `covarianceParameters` derives the confidence intervals in `stats` based on a Laplace approximation to the log likelihood of the generalized linear mixed-effects model.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `covarianceParameters` derives the confidence intervals in `stats` based on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

Examples

Obtain Estimated Covariance Parameters

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Compute and display the estimate of the prior covariance parameter for the random-effects predictor.

```
[psi,dispersion,stats] = covarianceParameters(glme);
psi{1}
```

```
ans = 0.0985
```

`psi{1}` is an estimate of the prior covariance matrix of the first grouping variable. In this example, there is only one grouping variable (`factory`), so `psi{1}` is an estimate of σ_b^2.

Display the dispersion parameter.

```
dispersion
```

```
dispersion = 1
```

Display the estimated standard deviation of the random effect associated with the predictor. The first cell of `stats` contains statistics for `factory`, while the second cell contains statistics for the dispersion parameter.

```
stats{1}
```

```
ans =
    Covariance Type: Isotropic

    Group      Name1                Name2                Type
    factory    {'(Intercept)'}      {'(Intercept)'}      {'std'}


    Estimate    Lower      Upper
    0.31381     0.19253    0.51148
```

The estimated standard deviation of the random effect associated with the predictor is 0.31381. The 95% confidence interval is [0.19253 , 0.51148]. Because the confidence interval does not contain 0, the random intercept is significant at the 5% significance level.

See Also

`GeneralizedLinearMixedModel` | `compare` | `fitglme` | `fixedEffects` | `randomEffects`

covarianceParameters

Class: `LinearMixedModel`

Extract covariance parameters of linear mixed-effects model

Syntax

```
psi = covarianceParameters(lme)
[psi,mse] = covarianceParameters(lme)
[psi,mse,stats] = covarianceParameters(lme)
[psi,mse,stats] = covarianceParameters(lme,Name,Value)
```

Description

`psi = covarianceParameters(lme)` returns the estimated covariance parameters that parameterize the prior covariance of random effects.

`[psi,mse] = covarianceParameters(lme)` also returns an estimate of the residual variance.

`[psi,mse,stats] = covarianceParameters(lme)` also returns a cell array, `stats`, containing the covariance parameters and related statistics.

`[psi,mse,stats] = covarianceParameters(lme,Name,Value)` returns the covariance parameters and related statistics in `stats` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the confidence level for the confidence limits of covariance parameters.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

Output Arguments

`psi` — Estimate of covariance parameters

cell array

Estimate of covariance parameters that parameterize the prior covariance of the random effects, returned as a cell array of length R, such that `psi{r}` contains the covariance matrix of random effects associated with grouping variable g_r, $r = 1, 2, \ldots, R$. The order of grouping variables is the same order you enter when you fit the model.

`mse` — Residual variance estimate

scalar value

Residual variance estimate, returned as a scalar value.

`stats` — Covariance parameter estimates and related statistics

cell array

Covariance parameter estimates and related statistics, returned as a cell array of length $(R + 1)$ containing dataset arrays with the following columns.

`Group`	Grouping variable name
`Name1`	Name of the first predictor variable
`Name2`	Name of the second predictor variable
`Type`	`std` (standard deviation), if `Name1` and `Name2` are the same `corr` (correlation), if `Name1` and `Name2` are different
`Estimate`	Standard deviation of the random effect associated with predictor `Name1` or `Name2`, if `Name1` and `Name2` are the same Correlation between the random effects associated with predictors `Name1` and `Name2`, if `Name1` and `Name2` are different
`Lower`	Lower limit of a 95% confidence interval for the covariance parameter
`Upper`	Upper limit of a 95% confidence interval for the covariance parameter

`stats{r}` is a dataset array containing statistics on covariance parameters for the rth grouping variable, $r = 1, 2, \ldots, R$. `stats{R+1}` contains statistics on the residual standard deviation. The dataset array for the residual error has the fields `Group`, `Name`, `Estimate`, `Lower`, and `Upper`.

Examples

Two Random-Effects Terms for Intercept

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` is the fixed-effects variable, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently. This model corresponds to

$$y_{ijk} = \beta_0 + \sum_{j=2}^{5} \beta_{2j} I[T]_{ij} + b_{0jk}(S*T)_{jk} + \epsilon_{ijk},$$

where $i = 1, 2, ..., 60$ corresponds to the observations, $j = 2, ..., 5$ corresponds to the tomato types, and $k = 1, 2, 3$ corresponds to the blocks (soil). S_k represents the k th soil type, and $(S*T)_{jk}$ represents the j th tomato type nested in the k th soil type. $I[T]_{ij}$ is the dummy variable representing the level j of the tomato type.

The random effects and observation error have the following prior distributions: $b_{0k} \sim N(0, \sigma_S^2)$, $b_{0jk} \sim N(0, \sigma_{S*T}^2)$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

```
lme = fitlme(ds,'Yield ~ Fertilizer + (1|Soil) + (1|Soil:Tomato)');
```

Compute the covariance parameter estimates (estimates of σ_S^2 and σ_{S*T}^2) of the random-effects terms.

```
psi = covarianceParameters(lme)
```

```
psi=2×1 cell array
    {[3.1731e-17]}
    {[  352.8481]}
```

Compute the residual variance (σ^2).

```
[~,mse] = covarianceParameters(lme)
```

```
mse = 151.9007
```

Potentially Correlated Random-Effects Terms

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a dataset array. Define `Subject` and `Program` as categorical variables.

```
ds = dataset(InitialWeight,Program,Subject,Week,y);
ds.Subject = nominal(ds.Subject);
ds.Program = nominal(ds.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

For `'reference'` dummy variable coding, `fitlme` uses Program A as reference and creates the necessary dummy variables $I[\,.\,]$. This model corresponds to

$$\begin{aligned} y_{im} &= \beta_0 + \beta_1 IW_i + \beta_2 \mathrm{Week}_i + \beta_3 I[PB]_I + \beta_4 I[PC]_i \\ &\quad + \beta_5 I[PD]_i + b_{0m} + b_{1m}\mathrm{Week}_{im} + \epsilon_{im} \end{aligned}$$

where i corresponds to the observation number, $i = 1, 2, \ldots, 120$, and m corresponds to the subject number, $m = 1, 2, \ldots, 20$. β_j are the fixed-effects coefficients, $j = 0, 1, \ldots, 8$, and b_{0m} and b_{1m} are random effects. IW stands for initial weight and $I[\,.\,]$ is a dummy variable representing a type of program. For example, $I[PB]_i$ is the dummy variable representing Program B.

The random effects and observation error have the following prior distributions:

$$\begin{pmatrix} b_{0m} \\ b_{1m} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{pmatrix}\right)$$

and

$$\epsilon_{im} \sim N(0, \sigma^2)\,.$$

```
lme = fitlme(ds,'y ~ InitialWeight + Program + (Week|Subject)');
```

Compute the estimates of covariance parameters for the random effects.

```
[psi,mse,stats] = covarianceParameters(lme)
```

```
psi = 1x1 cell array
    {2x2 double}


mse = 0.0105

stats=2×1 cell array
    {3x7 classreg.regr.lmeutils.titleddataset}
    {1x5 classreg.regr.lmeutils.titleddataset}
```

`mse` is the estimated residual variance. It is the estimate for σ^2.

To see the covariance parameters estimates for the random-effects terms (σ_0^2, σ_1^2, and $\sigma_{0,1}$), index into `psi`.

```
psi{1}
```

```
ans = 2×2

    0.0572    0.0490
    0.0490    0.0624
```

The estimate of the variance of the random effects term for the intercept, σ_0^2, is 0.0572. The estimate of the variance of the random effects term for week, σ_1^2, is 0.0624. The estimate for the covariance of the random effects terms for the intercept and week, $\sigma_{0,1}$, is 0.0490.

`stats` is a 2-by-1 cell array. The first cell of `stats` contains the confidence intervals for the standard deviation of the random effects and the correlation between the random effects for intercept and week. To display them, index into `stats`.

```
stats{1}
```

```
ans =
    Covariance Type: FullCholesky

    Group      Name1                Name2                Type
    Subject    {'(Intercept)'}      {'(Intercept)'}      {'std' }
    Subject    {'Week'       }      {'(Intercept)'}      {'corr'}
    Subject    {'Week'       }      {'Week'       }      {'std' }


    Estimate    Lower      Upper
    0.23927     0.14364    0.39854
    0.81971     0.38662    0.95658
     0.2497     0.18303    0.34067
```

The display shows the name of the grouping parameter (`Group`), the random-effects variables (`Name1`, `Name2`), the type of the covariance parameters (`Type`), the estimate (`Estimate`) for each parameter, and the 95% confidence intervals for the parameters (`Lower`, `Upper`). The estimates in this table are related to the estimates in `psi` as follows.

The standard deviation of the random-effects term for intercept is 0.23927 = sqrt(0.0527). Likewise, the standard deviation of the random effects term for week is 0.2497 = sqrt(0.0624). Finally, the correlation between the random-effects terms of intercept and week is 0.81971 = 0.0490/(0.23927*0.2497).

Note that this display also shows which covariance pattern you use when fitting the model. In this case, the covariance pattern is `FullCholesky`. To change the covariance pattern for the random-effects terms, you must use the `'CovariancePattern'` name-value pair argument when fitting the model.

The second cell of `stats` includes similar statistics for the residual standard deviation. Display the contents of the second cell.

```
stats{2}
```

```
ans =
    Group    Name               Estimate    Lower       Upper
    Error    {'Res Std'}        0.10261     0.087882    0.11981
```

The estimate for residual standard deviation is the square root of `mse`, 0.10261 = sqrt(0.0105).

Two Grouping Variables

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration and weight, a potentially correlated random effect for intercept and acceleration grouped by model year, and an independent random effect for weight, grouped by the origin of the car. This model corresponds to

$$\mathrm{MPG}_{imk} = \beta_0 + \beta_1 \mathrm{Acc}_i + \beta_2 \mathrm{Weight}_i + b_{10m} + b_{11m}\mathrm{Acc}_i + b_{21k}\mathrm{Weight}_i + \epsilon_{imk}$$

where $m = 1, 2, \ldots, 13$ represents the levels for the variable `Model_Year`, and $k = 1, 2, \ldots, 8$ represents the levels for the variable `Origin`. MPG_{imk} is the miles per gallon for the ith observation,| m| th model year, and|k| th origin that correspond to the ith observation. The random-effects terms and the observation error have the following prior distributions:

$$b_{1m} = \begin{pmatrix} b_{10m} \\ b_{11m} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_{10}^2 & \sigma_{10,11} \\ \sigma_{10,11} & \sigma_{11}^2 \end{pmatrix}\right),$$

$$b_{2k} \sim N\left(0, \sigma_2^2\right),$$

$$\epsilon_{imk} \sim N\left(0, \sigma^2\right).$$

Here, the random-effects term b_{1m} represents the first random effect at level m of the first grouping variable. The random-effects term b_{10m} corresponds to the first random effects term (1), for the intercept (0), at the m th level (m) of the first grouping variable. Likewise b_{11m} is the level m for the first predictor (1) in the first random-effects term (1).

Similarly, b_{2k} stands for the second random effects-term at level k of the second grouping variable.

σ_{10}^2 is the variance of the random-effects term for the intercept, σ_{11}^2 is the variance of the random effects term for the predictor acceleration, and $\sigma_{10,11}$ is the covariance of the random-effects terms for the intercept and the predictor acceleration. σ_2^2 is the variance of the second random-effects term, and σ^2 is the residual variance.

First, prepare the design matrices for fitting the linear mixed-effects model.

```
X = [ones(406,1) Acceleration Weight];
Z = {[ones(406,1) Acceleration],[Weight]};
```

```
Model_Year = nominal(Model_Year);
Origin = nominal(Origin);
G = {Model_Year,Origin};
```

Fit the model using the design matrices.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Weight'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'},{'Weight'}},'RandomEffectGroups',{'Model_Year','Origin'});
```

Compute the estimates of covariance parameters for the random effects.

```
[psi,mse,stats] = covarianceParameters(lme)
```

```
psi=2×1 cell array
    {2x2 double  }
    {[6.6397e-08]}


mse = 9.1177

stats=3×1 cell array
    {3x7 classreg.regr.lmeutils.titleddataset}
    {1x7 classreg.regr.lmeutils.titleddataset}
    {1x5 classreg.regr.lmeutils.titleddataset}
```

The residual variance `mse` is 9.0755. `psi` is a 2-by-1 cell array, and `stats` is a 3-by-1 cell array. To see the contents, you must index into these cell arrays.

First, index into the first cell of `psi`.

```
psi{1}
```

```
ans = 2×2

    5.9642   -0.6579
   -0.6579    0.0959
```

The first cell of `psi` contains the covariance parameters for the correlated random effects for intercept σ_{10}^2 as 8.5160, and for acceleration σ_{11}^2 as 0.1087. The estimate for the covariance of the random-effects terms for the intercept and acceleration $\sigma_{10,11}$ is -0.8387.

Now, index into the second cell of `psi`.

```
psi{2}
```

```
ans = 6.6397e-08
```

The second cell of `psi` contains the estimate for the variance of the random-effects term for weight σ_2^2.

Index into the first cell of `stats`.

```
stats{1}
```

```
ans =
    Covariance Type: FullCholesky
```

```
    Group         Name1                  Name2                  Type
    Model_Year    {'Intercept'   }       {'Intercept'   }       {'std' }
    Model_Year    {'Acceleration'}       {'Intercept'   }       {'corr'}
    Model_Year    {'Acceleration'}       {'Acceleration'}       {'std' }


    Estimate    Lower       Upper
      2.4422     0.72825      8.1898
    -0.87002    -0.98726    -0.14036
     0.30962       0.173     0.55415
```

This table shows the standard deviation estimates for the random-effects terms for intercept and acceleration. Note that the standard deviations estimates are the square roots of the diagonal elements in the first cell of `psi`. Specifically, 2.9182 = sqrt(8.5160) and 0.32968 = sqrt(0.1087). The correlation is a function of the covariance of intercept and acceleration, and the standard deviations of intercept and acceleration. The covariance of intercept and acceleration is the off-diagonal value in the first cell of psi, -0.8387. So, the correlation is -.8387/(0.32968*2.92182) = -0.87.

The grouping variable for intercept and acceleration is `Model_Year`.

Index into the second cell of `stats`.

```
stats{2}
```

```
ans =
    Covariance Type: FullCholesky

    Group     Name1             Name2             Type           Estimate
    Origin    {'Weight'}        {'Weight'}        {'std'}        0.00025768


    Lower       Upper
    9.07e-05    0.00073205
```

The second cell of `stats` has the standard deviation estimate and the 95% confidence limits for the standard deviation of the random-effects term for `Weight`. The grouping variable is `Origin`.

Index into the third cell of `stats`.

```
stats{3}
```

```
ans =
    Group    Name               Estimate    Lower     Upper
    Error    {'Res Std'}        3.0196      2.8083    3.2467
```

The third cell of `stats` contains the estimate for residual standard deviation and the 95% confidence limits. The estimate for residual standard deviation is the square root of `mse`, sqrt(9.0755) = 3.0126.

Construct 99% confidence intervals for the covariance parameters.

```
[~,~,stats] = covarianceParameters(lme,'Alpha',0.01);
stats{1}
```

```
ans =
    Covariance Type: FullCholesky
```

```
    Group         Name1                  Name2                  Type
    Model_Year    {'Intercept'   }       {'Intercept'   }       {'std' }
    Model_Year    {'Acceleration'}       {'Intercept'   }       {'corr'}
    Model_Year    {'Acceleration'}       {'Acceleration'}       {'std' }


    Estimate    Lower       Upper
      2.4422     0.49792     11.978
    -0.87002    -0.99396    0.22908
     0.30962     0.14408    0.66537
```

```
stats{2}
```

```
ans =
    Covariance Type: FullCholesky

    Group     Name1             Name2             Type           Estimate
    Origin    {'Weight'}        {'Weight'}        {'std'}        0.00025768


    Lower         Upper
    6.5331e-05    0.0010163
```

```
stats{3}
```

```
ans =
    Group    Name               Estimate    Lower    Upper
    Error    {'Res Std'}        3.0196      2.745    3.3216
```

See Also

LinearMixedModel | compare | fixedEffects | randomEffects

CoxModel

Cox proportional hazards model

Description

A Cox proportional hazards model relates to lifetime or failure time data. The basic Cox model includes a hazard function $h_0(t)$ and model coefficients b such that, for predictor `X`, the hazard rate at time t is

$$h(X_i,t) = h_0(t)\exp\left[\sum_{j=1}^{p} x_{ij}b_j\right],$$

where the b coefficients do not depend on time. The creation function `fitcox` infers both the model coefficients b and the hazard rate $h_0(t)$, and stores them as properties in the resulting `CoxModel` object.

The full Cox model includes extensions to the basic model, such as hazards with respect to different baselines or the inclusion of stratification variables. See “Extension of Cox Proportional Hazards Model” on page 14-27.

Creation

Create a `CoxModel` object using `fitcox`.

Properties

`Baseline` — Baseline hazard
`mean(X)` (default) | real scalar

Baseline hazard specified when model was fitted, specified as a real scalar. The Cox model is a relative hazard model, so it requires a baseline at which to compare hazards of given data, relative to the baseline. The default is `mean(X)` (and the mean within each stratification for stratified models), so the hazard rate at `X` is `h(t)*exp((X-mean(X))*b)`. Enter `0` to compute the baseline relative to `0`, so the hazard rate at `X` is `h(t)*exp(X*b)`. Changing the baseline does not affect the coefficients.

Data Types: `double`

`CoefficientCovariance` — Covariance matrix for coefficient estimates
square matrix

Covariance matrix for coefficient estimates, specified as a square matrix with the number of rows equal to the number of predictors.

Data Types: `double`

`Coefficients` — Coefficients and related statistics
table

Coefficients and related statistics, specified as a table with four columns:

- `Beta` — Coefficient estimate
- `SE` — Standard error of the coefficient estimate
- `zStat` — z statistic
- `pValue` — *p*-value for the coefficient (compared to a zero Beta)

Each row of the table corresponds to one predictor.

To obtain any of these columns as a vector, index into the property using dot notation. For example, in the `coxMdl` object, the estimated coefficient vector is

```
beta = coxMdl.Coefficients.Beta
```

To perform other tests on the coefficients, use `linhyptest`.

Data Types: `table`

`Formula` — Representation of the model used in fit
formula in Wilkinson notation

Representation of the model used in the fit, specified as a formula in Wilkinson notation. See “Wilkinson Notation” on page 11-91. For example, to include several predictors, use

```
'X ~ a + b + … + c'
```

where each of the variables `a`, `b`, `c` represents one predictor. These variables are column names for the table `X`.

`Hazard` — Estimated baseline cumulative hazard
matrix double

Estimated baseline cumulative hazard, specified as a matrix double. The cumulative hazard is evaluated at time points defined in training.

`Hazard` has at least two columns. The first column contains the time values, and the rest of the columns contain the cumulative hazard at each listed time.

- For nonstratified models, `Hazard` has two columns.
- For stratified models, `Hazard` has an additional column for each unique combination of the stratification levels. Distinct time values in `Hazard(:,1)` for each stratification are separated by `0` entries in `Hazard(:,2)`. A stratified model is a model trained using the `'Stratification'` name-value argument.

Theoretically, the cumulative hazard at time t is –log(1 – cdf(t)). The empirical cumulative hazard is

$$\widehat{H_0}(t) = \sum_{t_i \le t} \widehat{h_0}(t_i) = \sum_{t_i \le t} \frac{1}{\sum_{j \in R_i} \exp(\beta \cdot x_j)},$$

where R_i is the risk set at time t_i, meaning the observations that are at risk of failing. See “Partial Likelihood Function” on page 14-27.

Data Types: `double`

`IsStratified` — Indication that model is trained with stratified data
logical

Indication that the model is trained with stratified data, specified as a logical value (`true` or `false`).

Data Types: `logical`

`LikelihoodRatioTestPValue` — *P*-value indicating model is significant relative to null model
real scalar

P-value indicating if the model is significant relative to the null model, specified as a real scalar.

This property contains the *p*-value of performing the likelihood ratio test against the null model, that is, a model with all coefficients equal to 0.

The likelihood ratio test compares the likelihood function of the data at the coefficient estimates, and at all the coefficients being 0. The comparison yields a test statistic that can be used to determine if the trained model is significant, relative to a model with all coefficients equal to 0. The null hypothesis is that there is no difference between the null model and the trained model, so a significant *p*-value implies the trained model is significant.

Data Types: `double`

`LogLikelihood` — Log of likelihood function at coefficient estimates
real scalar

Log of the likelihood function at the coefficient estimates, specified as a real scalar.

Data Types: `double`

`NumPredictors` — Number of predictors
positive integer

Number of predictors (coefficients) in the model, specified as a positive integer.

Data Types: `double`

`PredictorNames` — Names of predictors
cell array of character vectors

Names of the predictors used to fit the model, specified as a cell array of character vectors. If the model is trained on data in a table, the predictor names are the names of the table columns. Otherwise, the predictor names are `X1`, `X2`, and so on.

Data Types: `cell`

`ProportionalHazardsPValue` — *P*-value indicating covariates satisfies the proportional hazards assumption
real vector

P-value indicating if each covariate satisfies the proportional hazards assumption, specified as a real vector, with one entry for each predictor.

The Cox model relies on the assumption of proportional hazards, that is, for any two data points X1 and X2, hazard(X1)/hazard(X2) is constant. This assumption might be violated if the predictors depend on time. For example, if a predictor corresponds to age, it generally becomes more hazardous as age increases.

The test of this assumption uses the scaled Schoenfeld residuals and was derived by Grambsch and Therneau in [1].

The null hypothesis is that each coefficient satisfies the proportional hazards assumption. A significant *p*-value implies that a specific coefficient violates the proportional hazards assumption. The test is performed on each coefficient, so this property is a vector with as many elements as the number of coefficients.

Data Types: `double`

`ProportionalHazardsPValueGlobal` — *P*-value indicating model satisfies proportional hazards assumption

real scalar

P-value indicating if the whole model satisfies the proportional hazards assumption, specified as a real scalar.

The Cox model relies on the assumption of proportional hazards, that is, for any two data points X1 and X2, hazard(X1)/hazard(X2) is constant. This assumption might be violated if the predictors depend on time. For example, if a predictor corresponds to age, it generally becomes more hazardous as age increases.

The test of this assumption uses the scaled Schoenfeld residuals and was derived by Grambsch and Therneau in [1].

The null hypothesis is that the model, as a whole, satisfies the proportional hazards assumption. A significant *p*-value implies the whole model does not satisfy the proportional hazards assumption.

Data Types: `double`

`Residuals` — Residuals of various types

table

Residuals of various types, specified as a table with seven columns, one for each residual:

- `'CoxSnell'` — The Cox-Snell residuals for an observation `X(i)` are defined as the cumulative hazard at time `i` (`cumHazard(i)`) multiplied by the hazard of `X(i)`: `csres(i) = cumHazard(i) * exp(X(i) * Beta)`. Beta is the fitted Beta vector stored in `Coefficients`.
- `'Deviance'` — The deviance residual is defined using the martingale residual as follows: `D(i) = sign(M(i))*sqrt(-2*[M(i) + delta(i)*log(delta(i)-M(i)))]`, where `D(i)` is the `i`th deviance residual, `M(i)` is the `i`th martingale residual, and `delta(i)` indicates if the data point `i` died or not.
- `'Martingale'` — The martingale residual for a point `X(i)` is `delta(i) - CoxSnell(i)`, where `delta(i)` indicates if `X(i)` died, and `CoxSnell(i)` is the Cox-Snell residual at `i`. The martingale residual can be viewed as the difference between the true number of deaths for `X(i)` minus the expected number of deaths based on the model.
- `'Schoenfeld'` — The Schoenfeld residuals are defined as: `scres(i,j) = X(i,j) - M(Beta,i,j)`, where `X(i,j)` is the `j`th element of observation `i`, and `M(Beta,i,j)` is the expected value of `X(i,j)`, given the number of living observations left at time `i`. The Schoenfeld residuals can be viewed as the difference between a true dead observation at time `i` and how the model expects a dead observation to look at time `i`, given the remaining living observations. The residuals are calculated for each covariate, so they have as many columns as the number of learned parameters. The residuals are valid only for times and observations at which there were deaths. For any censored observations, the corresponding residual is `NaN`.
- `'ScaledSchoenfeld'` — The scaled Schoenfeld residuals are the Schoenfeld residuals scaled by the variance of the learned coefficients. Like the Schoenfeld residuals, the scaled residuals are not

defined for observations and times at which there were no deaths; a residual at such a point is `NaN`.

- `'Score'` — The score residuals are defined as: `scores(i,t) = integral_{0}^{t}(X(i,u) – Xbar(u)) dScres(i,u)`, where `Schres(i)` is the Schoenfeld residual at observation `i`, and `Xbar` is the mean of the observations still alive at time `u`. The residuals are calculated for each covariate, so they have as many columns as the number of learned parameters.
- `'ScaledScore'` — The scaled score residuals are the score residuals scaled by the covariance of the fitted coefficients.

`Residuals` has the same number of rows as the training data.

Data Types: `table`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector. For models where the response value is in a table, the response variable name is the name of the relevant table column. Otherwise, `ResponseName` is `'T'`.

Data Types: `char`

`StandardError` — Standard errors of coefficient estimates
real vector

Standard errors of coefficient estimates, specified as a real vector. `StandardError` is the square root of the diagonal of the `CoefficientCovariance` matrix.

Data Types: `double`

`Stratification` — Array of unique combinations of input stratification
numeric array | string array | cell array of strings | categorical array | cell array

Array of unique combinations of input stratification during training, specified as one of the following data types.

- Numeric array — All stratification variables are numeric.
- String array — All variables are strings.
- Cell array of strings — All variables are cell strings.
- Categorical array — All variables are categorical.
- Cell array — Variables are mixed types.

Given some data `X` and `T`, the following table shows examples of what `Stratification` contains.

Input Data	Example	Resulting Stratification
Double	`mdl = fitcox(X,T,'Stratification',[1 2; 1 2; 2 2; 2 2]);`	`[1 2; 2 2]`
String	`mdl = fitcox(X,T,'Stratification', ["cat";"dog";"dog";"bird"]);`	`["cat"; "dog"; "bird"]`
Cell string	`mdl = fitcox(X,T,'Stratification', {'cat';'dog';'dog';'bird'});`	`{'cat';'dog';'bird'}`

Input Data	Example	Resulting Stratification
Categorical	`mdl = fitcox(X,T,'Stratification',categorical([1;2;3;4]));`	`categorical([1;2;3;4])`

Mixed types	`data = table(X,T,[1;2;1;3], {'cat';'cat';'dog';'dog'},'VariableNames', {'X','T','S1','S2');`
	
	`mdl = fitcox(data,'T','Stratification', {'S1','S2'});`
		`{1 'cat'; 2 'cat'; 1 'dog'; 3 'dog'}`

Data Types: `double` | `char` | `string` | `cell` | `categorical`

`VariableInfo` — Information about fitting data
table

Information about fitting data, specified as a table with four columns:

- `Class` — The class of the predictor.
- `Range` — The minimum and maximum of the predictor if it is not categorical, or the list of all the categories if the predictor is categorical.
- `InModel` — A logical indicating if the predictor is used in the model. The response variable is not in the model. Predictor variables used for training are in the model.
- `IsCategorical` — A logical indicating if the predictor was treated as categorical during training.

If the model has no categorical predictors, and no formula was used to fit the model, the number of rows of `VariableInfo` is the number of model predictors. Otherwise, the number of rows is the same as the number of elements in `PredictorNames`.

Data Types: `table`

Object Functions

coefci	Confidence interval for Cox proportional hazards model coefficients
hazardratio	Estimate Cox model hazard relative to baseline
linhyptest	Linear hypothesis tests on Cox model coefficients
plotSurvival	Plot survival function of Cox proportional hazards model
survival	Calculate survival of Cox proportional hazards model

Examples

Estimate Cox Proportional Hazard Regression

Weibull random variables with the same shape parameter have proportional hazard rates; see “Weibull Distribution” on page B-170. The hazard rate with scale parameter a and shape parameter b at time t is

$$\frac{b}{a^b}t^{b-1}.$$

Generate pseudorandom samples from the Weibull distribution with scale parameters 1, 5, and 1/3, and with the same shape parameter B.

```
rng default % For reproducibility
B = 2;
A = ones(100,1);
data1 = wblrnd(A,B);
A2 = 5*A;
data2 = wblrnd(A2,B);
A3 = A/3;
data3 = wblrnd(A3,B);
```

Create a table of data. The predictors are the three variable types, 1, 2, or 3.

```
predictors = categorical([A;2*A;3*A]);
data = table(predictors,[data1;data2;data3],'VariableNames',["Predictors" "Times"]);
```

Fit a Cox regression to the data.

```
mdl = fitcox(data,"Times")
```

```
mdl =
Cox Proportional Hazards regression model:

                    Beta        SE        zStat       pValue
                   _______    _______    _______    __________

    Predictors_2    -3.5834    0.33187    -10.798    3.5299e-27
    Predictors_3     2.1668    0.20802     10.416    2.0899e-25


```

```
rates = exp(mdl.Coefficients.Beta)
```

```
rates = 2×1

    0.0278
    8.7301
```

Fit Cox Proportional Hazards Model to Lifetime Data

Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Load the `lightbulb` data set.

```
load lightbulb
```

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

```
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3))
```

```
coxMdl =
Cox Proportional Hazards regression model:

           Beta       SE      zStat       pValue
          _______    _______    _______    __________

    X1    4.7262    1.0372    4.5568    5.1936e-06
```

Find the hazard rate of incandescent bulbs compared to fluorescent bulbs by evaluating exp(Beta).

```
hr = exp(coxMdl.Coefficients.Beta)
```

```
hr = 112.8646
```

The estimate of the hazard ratio is e^{Beta} = 112.8646, which means that the estimated hazard for the incandescent bulbs is 112.86 times the hazard for the fluorescent bulbs. The small value of `coxMdl.Coefficients.pValue` indicates there is a negligible chance that the two types of light bulbs have identical hazard rates, which would mean `Beta` = 0.

References

[1] Grambsch, Patricia M., and Terry M. Therneau. *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals*. Biometrika, vol. 81, no. 3, 1994, pp. 515–526. JSTOR, `https://www.jstor.org/stable/2337123`.

See Also

`fitcox`

Topics

“Cox Proportional Hazards Model” on page 14-26
“Cox Proportional Hazards Model Object” on page 14-39

Introduced in R2021a

coxphfit

Cox proportional hazards regression

Syntax

```
b = coxphfit(X,T)
b = coxphfit(X,T,Name,Value)
[b,logl,H,stats] = coxphfit( ___ )
```

Description

`b = coxphfit(X,T)` returns a *p*-by-1 vector, `b`, of coefficient estimates for a Cox proportional hazards regression on page 33-1028 of the observed responses `T` on the predictors `X`, where `T` is either an *n*-by-1 vector or an *n*-by-2 matrix, and `X` is an *n*-by-*p* matrix.

The model does not include a constant term, and `X` cannot contain a column of 1s.

`b = coxphfit(X,T,Name,Value)` returns a vector of coefficient estimates, with additional options specified by one or more `Name,Value` pair arguments.

`[b,logl,H,stats] = coxphfit( ___ )` also returns the loglikelihood, `logl`, a structure, `stats`, that contains additional statistics, and a two-column matrix, `H`, that contains the `T` values in the first column and the estimated baseline cumulative hazard, in the second column. You can use any of the input arguments in the previous syntaxes.

Examples

Use Cox Proportional Hazards Regression to Model Lifetime of Light Bulbs

Load the sample data.

```
load('lightbulb.mat');
```

The first column of the light bulb data has the lifetime (in hours) of two different types of bulbs. The second column has the binary variable indicating whether the bulb is fluorescent or incandescent. 0 indicates that the bulb is fluorescent, and 1 indicates that it is incandescent. The third column contains the censorship information, where 0 indicates the bulb was observed until failure, and 1 indicates the bulb was censored.

Fit a Cox proportional hazards model for the lifetime of the light bulbs, also accounting for censoring. The predictor variable is the type of bulb.

```
b = coxphfit(lightbulb(:,2),lightbulb(:,1), ...
'Censoring',lightbulb(:,3))
```

```
b = 4.7262
```

The estimate of the hazard ratio is e^b = 112.8646. This means that the hazard for the incandescent bulbs is 112.86 times the hazard for the fluorescent bulbs.

Change Algorithm Parameters for Cox Proportional Hazards Model

Load the sample data.

```
load('lightbulb.mat');
```

The first column of the data has the lifetime (in hours) of two types of bulbs. The second column has the binary variable indicating whether the bulb is fluorescent or incandescent. 1 indicates that the bulb is fluorescent and 0 indicates that it is incandescent. The third column contains the censorship information, where 0 indicates the bulb is observed until failure, and 1 indicates the item (bulb) is censored.

Fit a Cox proportional hazards model, also accounting for censoring. The predictor variable is the type of bulb.

```
b = coxphfit(lightbulb(:,2),lightbulb(:,1),...
'Censoring',lightbulb(:,3))
```

```
b = 4.7262
```

Display the default control parameters for the algorithm `coxphfit` uses to estimate the coefficients.

```
statset('coxphfit')
```

```
ans = struct with fields:
          Display: 'off'
      MaxFunEvals: 200
          MaxIter: 100
           TolBnd: 1.0000e-06
           TolFun: 1.0000e-08
       TolTypeFun: []
             TolX: 1.0000e-08
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

Save the options under a different name and change how the results will be displayed and the maximum number of iterations, `Display` and `MaxIter`.

```
coxphopt = statset('coxphfit');
coxphopt.Display = 'final';
coxphopt.MaxIter = 50;
```

Run `coxphfit` with the new algorithm parameters.

```
b = coxphfit(lightbulb(:,2),lightbulb(:,1),...
'Censoring',lightbulb(:,3),'Options',coxphopt)
```

```
Successful convergence: Norm of gradient less than OPTIONS.TolFun
```

```
b = 4.7262
```

`coxphfit` displays a report on the final iteration. Changing the maximum number of iterations did not affect the coefficient estimate.

Fit and Compare Cox and Weibull Survivor Functions

Generate Weibull data depending on predictor `X`.

```
rng('default') % for reproducibility
X = 4*rand(100,1);
A = 50*exp(-0.5*X);
B = 2;
y = wblrnd(A,B);
```

The response values are generated from a Weibull distribution with a shape parameter depending on the predictor variable `X` and a scale parameter of 2.

Fit a Cox proportional hazards model.

```
[b,logL,H,stats] = coxphfit(X,y);
[b logL]
```

```
ans = 1×2

    0.9409 -331.1479
```

The coefficient estimate is 0.9409 and the log likelihood value is –331.1479.

Request the model statistics.

```
stats
```

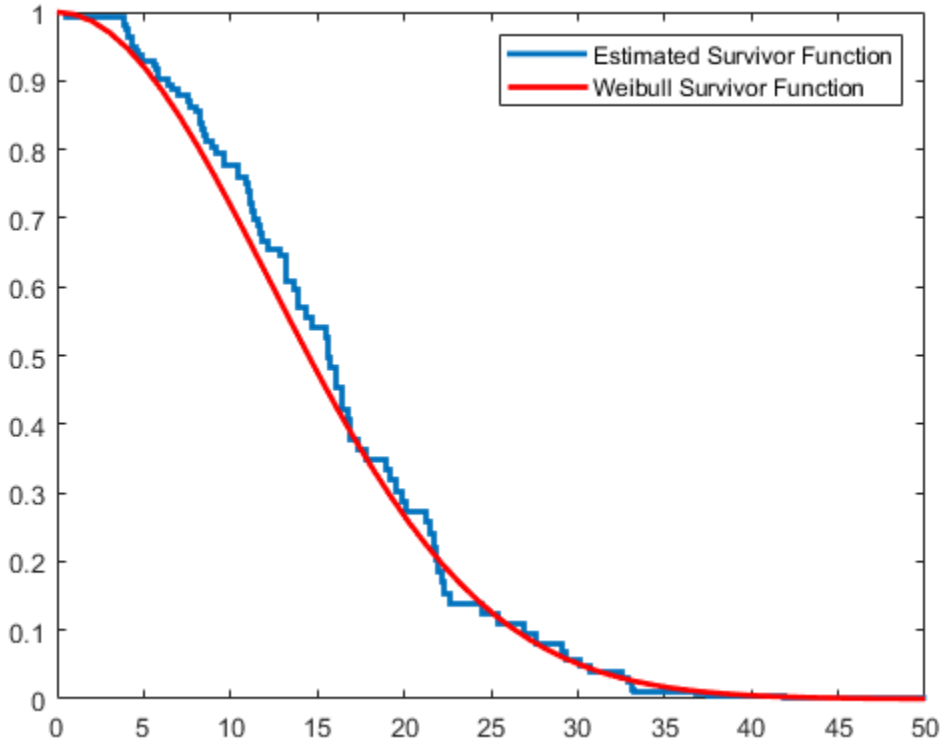
```
stats = struct with fields:
                    covb: 0.0158
                    beta: 0.9409
                      se: 0.1256
                       z: 7.4889
                       p: 6.9462e-14
                   csres: [100x1 double]
                  devres: [100x1 double]
                 martres: [100x1 double]
                  schres: [100x1 double]
                 sschres: [100x1 double]
                  scores: [100x1 double]
                 sscores: [100x1 double]
    LikelihoodRatioTestP: 6.6613e-16
```

The covariance matrix of the coefficient estimates, `covb`, contains only one value, which is equal to the variance of the coefficient estimate in this example. The coefficient estimate, `beta`, is the same as

`b` and is equal to 0.9409. The standard error of the coefficient estimate, `se`, is 0.1256, which is the square root of the variance 0.0158. The *z*-statistic, `z`, is `beta/se` = 0.9409/0.1256 = 7.4880. The p-value, `p`, indicates that the effect of `X` is significant.

Plot the Cox estimate of the baseline survivor function together with the known Weibull function.

```
stairs(H(:,1),exp(-H(:,2)),'LineWidth',2)
xx = linspace(0,100);
line(xx,1-wblcdf(xx,50*exp(-0.5*mean(X)),B),'color','r','LineWidth',2)
xlim([0,50])
legend('Estimated Survivor Function','Weibull Survivor Function')
```



The fitted model gives a close estimate to the survivor function of the actual distribution.

Input Arguments

X — Observations on predictor variables

matrix

Observations on predictor variables, specified as an *n*-by-*p* matrix of *p* predictors for each of *n* observations.

The model does not include a constant term, thus `X` cannot contain a column of 1s.

If `X`, `T`, or the value of `'Frequency'` or `'Strata'` contain `NaN` values, then `coxphfit` removes rows with `NaN` values from all data when fitting a Cox model.

Data Types: `double`

`T` — Time-to-event data
vector | two-column matrix

Time-to-event data, specified as an *n*-by-1 vector or a two-column matrix.

- When T is an *n*-by-1 vector, it represents the event time of right-censored time-to-event data.
- When T is an *n*-by-2 matrix, each row represents the risk interval (start,stop] in the counting process format for time-dependent covariates. The first column is the start time and the second column is the stop time. For an example, see “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35.

If `X`, `T`, or the value of `'Frequency'` or `'Strata'` contain `NaN` values, then `coxphfit` removes rows with `NaN` values from all data when fitting a Cox model.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Baseline',0,'Censoring',censoreddata,'Frequency',freq` specifies that `coxphfit` calculates the baseline hazard rate relative to 0, considering the censoring information in the vector `censoreddata`, and the frequency of observations on `T` and `X` given in the vector `freq`.

`B0` — Coefficient initial values
`0.01/std(X)` (default) | numeric vector

Coefficient initial values, specified as the comma-separated value consisting of `'B0'` and a numeric vector.

Data Types: `double`

`Baseline` — `X` values at which to compute the baseline hazard
`mean(X)` (default) | scalar value

`X` values at which to compute the baseline hazard, specified as the comma-separated pair consisting of `'Baseline'` and a scalar value.

The default is `mean(X)`, so the hazard rate at `X` is `h(t)*exp((X-mean(X))*b)`. Enter `0` to compute the baseline relative to 0, so the hazard rate at `X` is `h(t)*exp(X*b)`. Changing the baseline does not affect the coefficient estimates, but the hazard ratio changes.

Example: `'Baseline',0`

Data Types: `double`

`Censoring` — Indicator for censoring
array of 0s (default) | array of 0s and 1s

Indicator for censoring, specified as the comma-separated pair consisting of `'Censoring'` and a Boolean array of the same size as `T`. Use 1 for observations that are right censored and 0 for observations that are fully observed. The default is all observations are fully observed. For an example, see “Cox Proportional Hazards Model for Censored Data” on page 14-31.

Example: `'Censoring',cens`

Data Types: `logical`

`Frequency` — Frequency or weights of observations
array of 1s (default) | vector of nonnegative scalar values

Frequency or weights of observations, specified as the comma-separated pair consisting of `'Frequency'` and an array that is the same size as `T` containing nonnegative scalar values. The array can contain integer values corresponding to frequencies of observations or nonnegative values corresponding to observation weights.

If `X`, `T`, or the value of `'Frequency'` or `'Strata'` contain `NaN` values, then `coxphfit` removes rows with `NaN` values from all data when fitting a Cox model.

The default is 1 per row of `X` and `T`.

Example: `'Frequency',w`

Data Types: `double`

`Strata` — Stratification variables
`[]` (default) | matrix of real values

Stratification variables, specified as the comma-separated pair consisting of a matrix of real values. The matrix must have the same number of rows as `T`, with each row corresponding to an observation.

If `X`, `T`, or the value of `'Frequency'` or `'Strata'` contain `NaN` values, then `coxphfit` removes rows with `NaN` values from all data when fitting a Cox model.

The default, `[]`, is no stratification variable.

Example: `'Strata',Gender`

Data Types: `single` | `double`

`Ties` — Method to handle tied failure times
`'breslow'` (default) | `'efron'`

Method to handle tied failure times, specified as the comma-separated pair consisting of `'Ties'` and either `'breslow'` (Breslow's method) or `'efron'` (Efron's method).

Example: `'Ties','efron'`

`Options` — Algorithm control parameters
structure

Algorithm control parameters for the iterative algorithm used to estimate `b`, specified as the comma-separated pair consisting of `'Options'` and a structure. A call to `statset` creates this argument. For parameter names and default values, type `statset('coxphfit')`. You can set the options under a new name and use that in the name-value pair argument.

Example: `'Options',statset('coxphfit')`

Output Arguments

`b` — Coefficient estimates
vector

Coefficient estimates for a Cox proportional hazards regression on page 33-1028, returned as a *p*-by-1 vector.

`logl` — Loglikelihood
scalar

Loglikelihood of the fitted model, returned as a scalar.

You can use log likelihood values to compare different models and assess the significance of effects of terms in the model.

`H` — Estimated baseline cumulative hazard
two-column matrix | (2+*k*) column matrix

Estimated baseline cumulative hazard rate evaluated at `T` values, returned as one of the following.

- If the model is unstratified, then `H` is a two-column matrix. The first column of the matrix contains `T` values, and the second column contains cumulative hazard rate estimates.
- If the model is stratified, then `H` is a (2+*k*) column matrix, where the last *k* columns correspond to the stratification variables using the `Strata` name-value pair argument.

`stats` — Coefficient statistics
structure

Coefficient statistics, returned as a structure that contains the following fields.

`beta`	Coefficient estimates (same as `b`)
`se`	Standard errors of coefficient estimates, `b`
`z`	*z*-statistics for `b` (that is, `b` divided by standard error)
`p`	*p*-values for `b`
`covb`	Estimated covariance matrix for `b`
`csres`	Cox-Snell residuals
`devres`	Deviance residuals
`martres`	Martingale residuals
`schres`	Schoenfeld residuals
`sschres`	Scaled Schoenfeld residuals
`scores`	Score residuals
`sscores`	Scaled score residuals

`coxphfit` returns the Cox-Snell, martingale, and deviance residuals as a column vector with one row per observation. It returns the Schoenfeld, scaled Schoenfeld, score, and scaled score residuals as matrices of the same size as X. Schoenfeld and scaled Schoenfeld residuals of censored data are `NaN`s.

More About

Cox Proportional Hazards Regression

Cox proportional hazards regression is a semiparametric method for adjusting survival rate estimates to remove the effect of confounding variables and to quantify the effect of predictor variables. The

method represents the effects of explanatory and confounding variables as a multiplier of a common baseline hazard function, $h_0(t)$.

For a baseline relative to 0, this model corresponds to

$$h(X_i, t) = h_0(t)\exp\left[\sum_{j=1}^{p} x_{ij}b_j\right],$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ is the predictor variable for the ith subject, $h(X_i,t)$ is the hazard rate at time t for X_i, and $h_0(t)$ is the baseline hazard rate function. The baseline hazard function is the nonparametric part of the Cox proportional hazards regression function, whereas the impact of the predictor variables is a loglinear regression. The assumption is that the baseline hazard function depends on time, t, but the predictor variables do not depend on time. See “Cox Proportional Hazards Model” on page 14-26 for details, including the extensions for stratification and time-dependent variables, tied events, and observation weights.

References

[1] Cox, D.R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[3] Kleinbaum, D. G., and M. Klein. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2005.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- `X` can be a single- or double-precision matrix and can be variable-size.
- The value of the `'Ties'` name-value pair argument must be a compile-time constant. For example, to use Efron's method to handle tied failure times, include `{coder.Constant('Ties'),coder.Constant('efron')}` in the `-args` value of `codegen`.
- Names in name-value pair arguments must be compile-time constants.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`ecdf` | `statset` | `wblfit`

Topics

“Hazard and Survivor Functions for Different Groups” on page 14-16
“Survivor Functions for Two Groups” on page 14-22
“Cox Proportional Hazards Model for Censored Data” on page 14-31
“Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35
“What Is Survival Analysis?” on page 14-2

“Kaplan-Meier Method” on page 14-10
“Cox Proportional Hazards Model” on page 14-26

Introduced before R2006a

createns

Create nearest neighbor searcher object

Syntax

```
NS = createns(X)
NS = createns(X,Name,Value)
```

Description

`NS = createns(X)` creates either an `ExhaustiveSearcher` or `KDTreeSearcher` model object using the *n*-by-*K* numeric matrix of the training data `X`.

`NS = createns(X,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify `NSMethod` to determine which type of object to create.

Examples

Train Default Exhaustive Nearest Neighbor Searcher

Load Fisher's iris data set.

```
load fisheriris
X = meas;
[n,k] = size(X)
```

```
n = 150
```

```
k = 4
```

`X` has 150 observations and 4 predictors.

Prepare an exhaustive nearest neighbor searcher using the entire data set as training data.

```
Mdl1 = ExhaustiveSearcher(X)
```

```
Mdl1 =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl1` is an `ExhaustiveSearcher` model object, and its properties appear in the Command Window. The object contains information about the trained algorithm, such as the distance metric. You can alter property values using dot notation.

Alternatively, you can prepare an exhaustive nearest neighbor searcher by using `createns` and specifying `'exhaustive'` as the search method.

```
Mdl2 = createns(X,'NSMethod','exhaustive')
```

```
Mdl2 =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl2` is also an `ExhaustiveSearcher` model object, and it is equivalent to `Mdl1`.

To search `X` for the nearest neighbors to a batch of query data, pass the `ExhaustiveSearcher` model object and the query data to `knnsearch` or `rangesearch`.

Grow Default *K*d-Tree

Grow a four-dimensional *K*d-tree that uses the Euclidean distance.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
[n,k] = size(X)
```

```
n = 150
```

```
k = 4
```

`X` has 150 observations and 4 predictors.

Grow a four-dimensional *K*d-tree using the entire data set as training data.

```
Mdl1 = KDTreeSearcher(X)
```

```
Mdl1 =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl1` is a `KDTreeSearcher` model object, and its properties appear in the Command Window. The object contains information about the grown four-dimensional *K*d-tree, such as the distance metric. You can alter property values using dot notation.

Alternatively, you can grow a *K*d-tree by using `createns`.

```
Mdl2 = createns(X)
```

```
Mdl2 =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
```

```
                X: [150x4 double]
```

`Mdl2` is also a `KDTreeSearcher` model object, and it is equivalent to `Mdl1`. Because `X` has four columns and the default distance metric is Euclidean, `createns` creates a `KDTreeSearcher` model by default.

To find the nearest neighbors in `X` to a batch of query data, pass the `KDTreeSearcher` model object and the query data to `knnsearch` or `rangesearch`.

Grow *K*d-Tree Using Minkowski Distance Metric

Grow a *K*d-tree that uses the Minkowski distance with an exponent of five.

Load Fisher's iris data set. Create a variable for the petal dimensions.

```
load fisheriris
X = meas(:,3:4);
```

Grow a *K*d-tree. Specify the Minkowski distance with an exponent of five.

```
Mdl = createns(X,'Distance','minkowski','P',5)
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'minkowski'
    DistParameter: 5
                X: [150x2 double]
```

Because `X` has two columns and the distance metric is Minkowski, `createns` creates a `KDTreeSearcher` model object by default.

Search for Nearest Neighbors of Query Data Using Mahalanobis Distance

Create an exhaustive searcher object by using the `createns` function. Pass the object and query data to the `knnsearch` function to find *k*-nearest neighbors.

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng('default');             % For reproducibility
n = size(meas,1);           % Sample size
qIdx = randsample(n,5);     % Indices of query data
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Prepare an exhaustive nearest neighbor searcher using the training data. Specify the Mahalanobis distance for finding nearest neighbors.

```
Mdl = createns(X,'Distance','mahalanobis')

Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'mahalanobis'
    DistParameter: [4x4 double]
                X: [145x4 double]
```

Because the distance metric is Mahalanobis, `createns` creates an `ExhaustiveSearcher` model object by default.

The software uses the covariance matrix of the predictors (columns) in the training data for computing the Mahalanobis distance. To display this value, use `Mdl.DistParameter`.

```
Mdl.DistParameter

ans = 4×4

    0.6547   -0.0368    1.2320    0.5026
   -0.0368    0.1914   -0.3227   -0.1193
    1.2320   -0.3227    3.0671    1.2842
    0.5026   -0.1193    1.2842    0.5800
```

Find the indices of the training data (`Mdl.X`) that are the two nearest neighbors of each point in the query data (`Y`).

```
IdxNN = knnsearch(Mdl,Y,'K',2)

IdxNN = 5×2

     5     6
    98    95
   104   128
   135    65
   102   115
```

Each row of `IdxNN` corresponds to a query data observation. The column order corresponds to the order of the nearest neighbors with respect to ascending distance. For example, based on the Mahalanobis metric, the second nearest neighbor of `Y(3,:)` is `X(128,:)`.

Input Arguments

X — Training data

numeric matrix

Training data, specified as a numeric matrix. `X` has *n* rows, each corresponding to an observation (that is, an instance or example), and *K* columns, each corresponding to a predictor (that is, a feature).

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `NS = createns(X,'Distance','mahalanobis')` creates an `ExhaustiveSearcher` model object that uses the Mahalanobis distance metric when searching for nearest neighbors.

For Exhaustive and Kd-Tree Nearest Neighbor Searchers

`NSMethod` — Nearest neighbor search method
`'kdtree'` | `'exhaustive'`

Nearest neighbor search method used to define the type of object created, specified as the comma-separated pair consisting of `'NSMethod'` and `'kdtree'` or `'exhaustive'`.

- `'kdtree'` — `createns` creates a `KDTreeSearcher` model object using the *K*d-tree algorithm.
- `'exhaustive'` — `createns` creates an `ExhaustiveSearcher` model object using the exhaustive search algorithm.

The default value is `'kdtree'` when these three conditions are true:

- The number of columns of `X` (K) is less than or equal to 10 (that is, $K \le 10$).
- `X` is not sparse.
- `Distance` is `'euclidean'`, `'cityblock'`, `'chebychev'`, or `'minkowski'`.

Otherwise, the default value is `'exhaustive'`.

Example: `'NSMethod','exhaustive'`

`Distance` — Distance metric
`'euclidean'` (default) | character vector or string scalar of distance metric name | custom distance function

Distance metric used when you call `knnsearch` or `rangesearch` to find nearest neighbors for future query points, specified as the comma-separated pair consisting of `'Distance'` and a character vector or string scalar of distance metric name or function handle.

For both types of nearest neighbor searchers, `createns` supports these distance metrics.

Value	Description
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cityblock'`	City block distance.
`'euclidean'`	Euclidean distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value pair argument.

If `createns` uses the exhaustive search algorithm (`'NSMethod'` is `'exhaustive'`), then `createns` also supports these distance metrics.

Value	Description
'correlation'	One minus the sample linear correlation between observations (treated as sequences of values)
'cosine'	One minus the cosine of the included angle between observations (treated as row vectors)
'hamming'	Hamming distance, which is the percentage of coordinates that differ
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ
'mahalanobis'	Mahalanobis distance
'seuclidean'	Standardized Euclidean distance
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

If `createns` uses the exhaustive search algorithm (`'NSMethod'` is `'exhaustive'`), then you can also specify a function handle for a custom distance metric by using `@` (for example, `@distfun`). A custom distance function must:

- Have the form `function D2 = distfun(ZI,ZJ)`.
- Take as arguments:
  - A 1-by-*K* vector `ZI` containing a single row from `X` or from the query points `Y`, where *K* is the number of columns in `X`.
  - An *m*-by-*K* matrix `ZJ` containing multiple rows of `X` or `Y`, where *m* is a positive integer.
- Return an *m*-by-1 vector of distances `D2`, where `D2(`*j*`)` is the distance between the observations `ZI` and `ZJ(`*j*`,:)`.

For more details, see “Distance Metrics” on page 18-12.

Example: `'Distance','minkowski'`

P — Exponent for Minkowski distance metric

2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

For Exhaustive Nearest Neighbor Searchers

Cov — Covariance matrix for Mahalanobis distance metric

`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a *K*-by-*K* positive definite matrix, where *K* is the number of columns in `X`. This argument is valid only if `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

`Scale` — Scale parameter value for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector of length K, where K is the number of columns in `X`. The software scales each difference between the training and query data using the corresponding element of `Scale`. This argument is valid only if `'Distance'` is `'seuclidean'`.

Example: `'Scale',quantile(X,0.75) - quantile(X,0.25)`

Data Types: `single` | `double`

For Nearest Neighbor Searchers Using Kd-Tree

`BucketSize` — Maximum number of data points in each leaf node
`50` (default) | positive integer

Maximum number of data points in each leaf node of the Kd-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer.

This argument is valid only when you create a `KDTreeSearcher` model object.

Example: `'BucketSize',10`

Data Types: `single` | `double`

Output Arguments

`NS` — Nearest neighbor searcher
`ExhaustiveSearcher` model object | `KDTreeSearcher` model object

Nearest neighbor searcher, returned as an `ExhaustiveSearcher` model object or a `KDTreeSearcher` model object.

Once you create a nearest neighbor searcher model object, you can find the neighboring points in the training data to the query data by performing a nearest neighbor search using `knnsearch` or a radius search using `rangesearch`.

See Also

`ExhaustiveSearcher` | `KDTreeSearcher` | `knnsearch` | `rangesearch`

Topics
“k-Nearest Neighbor Search and Radius Search” on page 18-14

Introduced in R2010a

crosstab

Cross-tabulation

Syntax

```
tbl = crosstab(x1,x2)
tbl = crosstab(x1,...,xn)
[tbl,chi2,p] = crosstab( ___ )
[tbl,chi2,p,labels] = crosstab( ___ )
```

Description

`tbl = crosstab(x1,x2)` returns a cross-tabulation, `tbl`, of two vectors of the same length, `x1` and `x2`.

`tbl = crosstab(x1,...,xn)` returns a multi-dimensional cross-tabulation, `tbl`, of data for multiple input vectors, `x1`, `x2`, ..., `xn`.

`[tbl,chi2,p] = crosstab( ___ )` also returns the chi-square statistic, `chi2`, and its *p*-value, `p`, for a test that `tbl` is independent in each dimension. You can use any of the previous syntaxes.

`[tbl,chi2,p,labels] = crosstab( ___ )` also returns a cell array, `labels`, which contains one column of labels for each input argument, `x1` ... `xn`.

Examples

Cross-Tabulate Two Data Vectors

Create two sample data vectors, containing three and four distinct values, respectively.

```
x = [1 1 2 3 1];
y = [1 2 5 3 1];
```

Cross-tabulate `x` and `y`.

```
table = crosstab(x,y)
```

```
table = 3×4

     2     1     0     0
     0     0     0     1
     0     0     1     0
```

The rows in `table` correspond to the three distinct values in `x`, and the columns correspond to the four distinct values in `y`.

Cross-Tabulate Independent Data Vectors

Generate two independent vectors, `x1` and `x2`, each containing 50 discrete uniform random numbers in the range `1:3`.

```
rng default;  % for reproducibility
x1 = unidrnd(3,50,1);
x2 = unidrnd(3,50,1);
```

Cross-tabulate `x1` and `x2`.

```
[table,chi2,p] = crosstab(x1,x2)
```

```
table = 3×3

     1     6     7
     5     5     2
    11     7     6


chi2 = 7.5449

p = 0.1097
```

The returned `p` value of `0.1097` indicates that, at the 5% significance level, `crosstab` fails to reject the null hypothesis that `table` is independent in each dimension.

Cross-Tabulate Grouped Data

Load the sample data, which contains measurements of large model cars during the years 1970-1982.

```
load carbig
```

Cross-tabulate the data of four-cylinder cars (`cyl4`) based on model year (`when`) and country of origin (`org`).

```
[table,chi2,p,labels] = crosstab(cyl4,when,org);
```

Use `labels` to determine the index location in `table` for the number of four-cylinder cars made in the USA during the late period of the data.

```
labels
```

```
labels=3×3 cell array
    {'Other'   }    {'Early'}    {'USA'   }
    {'Four'    }    {'Mid'  }    {'Europe'}
    {0x0 double}    {'Late' }    {'Japan' }
```

The first column of `labels` corresponds to the data in `cyl4`, and indicates that row 2 of `table` contains data on cars with four cylinders. The second column of `labels` corresponds to the data in `when`, and indicates that column 3 of `table` contains data on cars made during the late period. The third column of `labels` corresponds to the data in `org`, and indicates that location 1 of the third dimension of `table` contains data on cars made in the USA.

Therefore, `table(2,3,1)` contains the number of four-cylinder cars made in the USA during the late period.

```
table(2,3,1)
```

```
ans = 38
```

The data contains 38 four-cylinder cars made in the USA during the late period.

Generate and Visualize Contingency Table

Create a contingency table from data, and visualize the table in a heatmap chart.

Load the hospital data.

```
load hospital
```

The `hospital` dataset array contains data on 100 hospital patients, including last name, gender, age, weight, smoking status, and systolic and diastolic blood pressure measurements.

Convert the dataset array to a MATLAB® table.

```
Tbl = dataset2table(hospital);
```

Determine whether smoking status is independent of gender by creating a 2-by-2 contingency table of smokers and nonsmokers, grouped by gender.

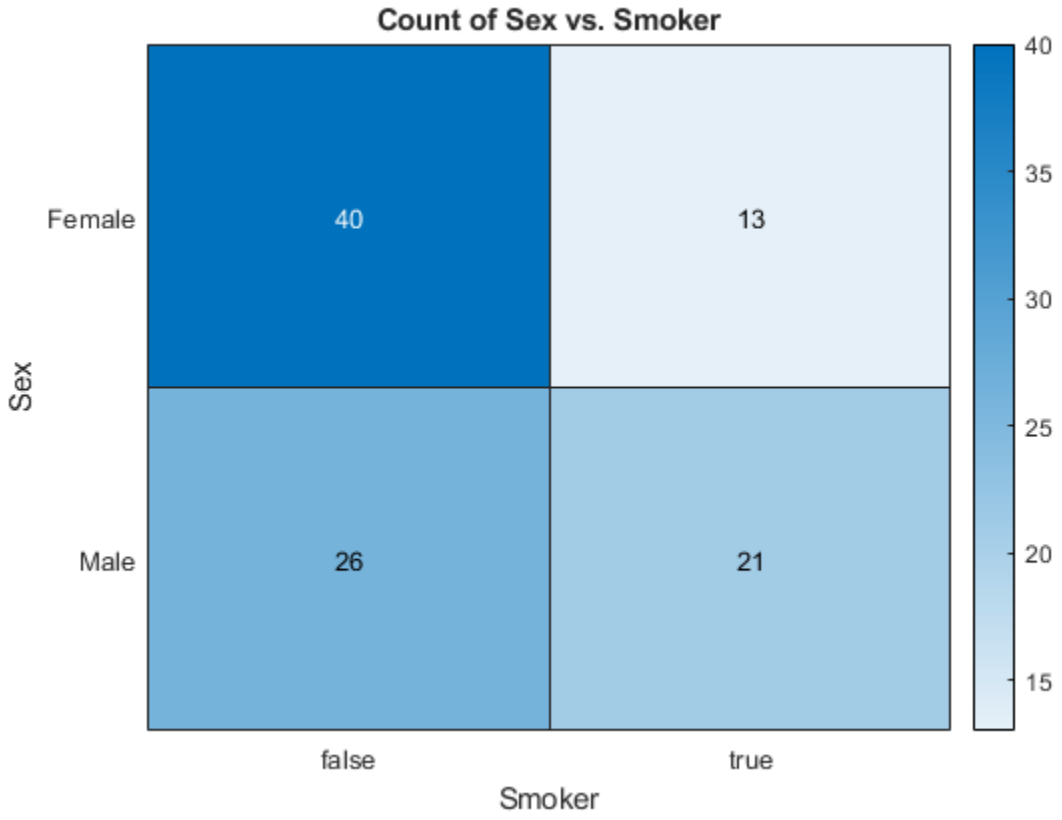
```
[conttbl,chi2,p,labels] = crosstab(Tbl.Sex,Tbl.Smoker)
```

```
conttbl = 2×2

    40    13
    26    21


chi2 = 4.5083

p = 0.0337

labels = 2x2 cell
    {'Female'}    {'0'}
    {'Male'  }    {'1'}
```

The rows of the resulting contingency table `conttbl` correspond to patient gender, with row 1 containing data for females and row 2 containing data for males. The columns correspond to patient smoking status, with column 1 containing data for nonsmokers and column 2 containing data for smokers. The returned result `chi2 = 4.5083` is the value of the chi-squared test statistic for a Pearson's chi-squared test of independence. The *p*-value for the test `p = 0.0337` suggests, at a 5% level of significance, rejection of the null hypothesis that gender and smoking status are independent.

Visualize the contingency table in a heatmap. Plot smoking status on the *x*-axis and gender on the *y*-axis.

```
heatmap(Tbl,'Smoker','Sex')
```



Input Arguments

x1 — Input vector

vector of grouping variables

Input vector, specified as a vector of grouping variables. All input vectors, including `x1`, `x2`, ..., `xn`, must be the same length.

Data Types: `single` | `double` | `char` | `string` | `logical` | `categorical`

x2 — Input vector

vector of grouping variables

Input vector, specified as a vector of grouping variables. All input vectors, including `x1`, `x2`, ..., `xn`, must be the same length.

Data Types: `single` | `double` | `char` | `string` | `logical` | `categorical`

x1,...,xn — Input vectors

vectors of grouping variables

Input vectors, specified as vectors of grouping variables. If you use this syntax to specify more than two input vectors, then `crosstab` generates a multi-dimensional cross-tabulation table. All input vectors, including `x1`, `x2`, ..., `xn`, must be the same length.

Data Types: `single` | `double` | `char` | `string` | `logical` | `categorical`

Output Arguments

`tbl` — Cross-tabulation table
matrix of integer values

Cross-tabulation table, returned as a matrix of integer values.

If you specify two input vectors, `x1` and `x2`, then `tbl` is an *m*-by-*n* matrix, where *m* is the number of distinct values in `x1` and *n* is the number of distinct values in `x2`.

If you specify three or more input vectors, then `tbl(i,j,...,n)` is a count of indices where `grp2idx(x1)` is `i`, `grp2idx(x2)` is `j`, `grp2idx(x3)` is `k`, and so on.

`chi2` — Chi-square statistic
positive scalar value

Chi-square statistic, returned as a positive scalar value. The null hypothesis is that the proportion in any entry of `tbl` is the product of the proportions in each dimension.

`p` — *p*-Value
scalar value in the range `[0,1]`

p-value for the chi-square test statistic, returned as a scalar value in the range `[0,1]`. `crosstab` tests that `tbl` is independent in each dimension.

`labels` — Data labels
cell array

Data labels, returned as a cell array. The entries in the first column are labels for the rows of `tbl`, the entries in the second column are labels for the columns, and so on, for a multi-dimensional `tbl`.

Algorithms

- `crosstab` uses `grp2idx` to assign a positive integer to each distinct value. `tbl(i,j)` is a count of indices where `grp2idx(x1)` is `i` and `grp2idx(x2)` is `j`. The numerical order of `grp2idx(x1)` and `grp2idx(x2)` order rows and columns of `tbl`, respectively.

  In this case, the returned value of `tbl(i,j,...,n)` is a count of indices where `grp2idx(x1)` is `i`, `grp2idx(x2)` is `j`, `grp2idx(x3)` is `k`, and so on.
- `crosstab` computes the *p*-value of the chi-square test statistic using a formula that is asymptotically valid for a large sample size. The approximation is less accurate for small samples or samples with uneven marginal distributions. If your sample includes only two variables and each has two levels, you can use `fishertest` instead. This function performs Fisher's exact test, which does not depend on large-sample distribution assumptions.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with the limitation:

The fourth output, `labels`, is returned as a cell array containing `M` unevaluated tall cell arrays, where `M` is the number of input grouping variables. Each unevaluated tall cell array, `labels{j}`, contains the labels for one grouping variable.

For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`fishertest` | `grp2idx` | `heatmap` | `tabulate`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

crossval

Estimate loss using cross-validation

Syntax

```
err = crossval(criterion,X,y,'Predfun',predfun)
err = crossval(criterion,X1,...,XN,y,'Predfun',predfun)

values = crossval(fun,X)
values = crossval(fun,X1,...,XN)

___ = crossval( ___ ,Name,Value)
```

Description

`err = crossval(criterion,X,y,'Predfun',predfun)` returns a 10-fold cross-validation error estimate for the function `predfun` based on the specified `criterion`, either `'mse'` (mean squared error) or `'msc'` (misclassification rate). The rows of `X` and `y` correspond to observations, and the columns of `X` correspond to predictor variables.

For more information, see “General Cross-Validation Steps for predfun” on page 33-1059.

`err = crossval(criterion,X1,...,XN,y,'Predfun',predfun)` returns a 10-fold cross-validation error estimate for `predfun` by using the predictor variables `X1` through `XN` and the response variable `y`.

`values = crossval(fun,X)` performs 10-fold cross-validation for the function `fun`, applied to the data in `X`. The rows of `X` correspond to observations, and the columns of `X` correspond to variables.

For more information, see “General Cross-Validation Steps for fun” on page 33-1059.

`values = crossval(fun,X1,...,XN)` performs 10-fold cross-validation for the function `fun`, applied to the data in `X1,...,XN`. Every data set, `X1` through `XN`, must have the same number of observations and, therefore, the same number of rows.

`___ = crossval( ___ ,Name,Value)` specifies cross-validation options using one or more name-value pair arguments in addition to any of the input argument combinations and output arguments in previous syntaxes. For example, `'KFold',5` specifies to perform 5-fold cross-validation.

Examples

Compute Mean Squared Error Using Cross-Validation

Compute the mean squared error of a regression model by using 10-fold cross-validation.

Load the `carsmall` data set. Put the acceleration, horsepower, weight, and miles per gallon (MPG) values into the matrix `data`. Remove any rows that contain `NaN` values.

```
load carsmall
data = [Acceleration Horsepower Weight MPG];
data(any(isnan(data),2),:) = [];
```

Specify the last column of `data`, which corresponds to `MPG`, as the response variable `y`. Specify the other columns of `data` as the predictor data `X`. Add a column of ones to `X` when your regression function uses `regress`, as in this example.

Note: `regress` is useful when you simply need the coefficient estimates or residuals of a regression model. If you need to investigate a fitted regression model further, create a linear regression model object by using `fitlm`. For an example that uses `fitlm` and `crossval`, see “Compute Mean Absolute Error Using Cross-Validation” on page 33-1048.

```
y = data(:,4);
X = [ones(length(y),1) data(:,1:3)];
```

Create the custom function `regf` (shown at the end of this example). This function fits a regression model to training data and then computes predicted values on a test set.

Note: If you use the live script file for this example, the `regf` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Compute the default 10-fold cross-validation mean squared error for the regression model with predictor data `X` and response variable `y`.

```
rng('default') % For reproducibility
cvMSE = crossval('mse',X,y,'Predfun',@regf)
```

```
cvMSE = 17.5399
```

This code creates the function `regf`.

```
function yfit = regf(Xtrain,ytrain,Xtest)
b = regress(ytrain,Xtrain);
yfit = Xtest*b;
end
```

Compute Misclassification Error Using Logistic Regression Model and Cross-Validation

Compute the misclassification error of a logistic regression model trained on numeric and categorical predictor data by using 10-fold cross-validation.

Load the `patients` data set. Specify the numeric variables `Diastolic` and `Systolic` and the categorical variable `Gender` as predictors, and specify `Smoker` as the response variable.

```
load patients
X1 = Diastolic;
X2 = categorical(Gender);
X3 = Systolic;
y = Smoker;
```

Create the custom function `classf` (shown at the end of this example). This function fits a logistic regression model to training data and then classifies test data.

Note: If you use the live script file for this example, the `classf` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Compute the 10-fold cross-validation misclassification error for the model with predictor data `X1`, `X2`, and `X3` and response variable `y`. Specify `'Stratify',y` to ensure that training and test sets have roughly the same proportion of smokers.

```
rng('default') % For reproducibility
err = crossval('mcr',X1,X2,X3,y,'Predfun',@classf,'Stratify',y)
```

```
err = 0.1100
```

This code creates the function `classf`.

```
function pred = classf(X1train,X2train,X3train,ytrain,X1test,X2test,X3test)
Xtrain = table(X1train,X2train,X3train,ytrain, ...
    'VariableNames',{'Diastolic','Gender','Systolic','Smoker'});
Xtest = table(X1test,X2test,X3test, ...
    'VariableNames',{'Diastolic','Gender','Systolic'});
modelspec = 'Smoker ~ Diastolic + Gender + Systolic';
mdl = fitglm(Xtrain,modelspec,'Distribution','binomial');
yfit = predict(mdl,Xtest);
pred = (yfit > 0.5);
end
```

Determine Number of Clusters Using Cross-Validation

For a given number of clusters, compute the cross-validated sum of squared distances between observations and their nearest cluster center. Compare the results for one through ten clusters.

Load the `fisheriris` data set. `X` is the matrix `meas`, which contains flower measurements for 150 different flowers.

```
load fisheriris
X = meas;
```

Create the custom function `clustf` (shown at the end of this example). This function performs the following steps:

1. Standardize the training data.
2. Separate the training data into `k` clusters.
3. Transform the test data using the training data mean and standard deviation.
4. Compute the distance from each test data point to the nearest cluster center, or centroid.
5. Compute the sum of the squares of the distances.

Note: If you use the live script file for this example, the `clustf` function is already included at the end of the file. Otherwise, you need to create the function at the end of your .m file or add it as a file on the MATLAB® path.

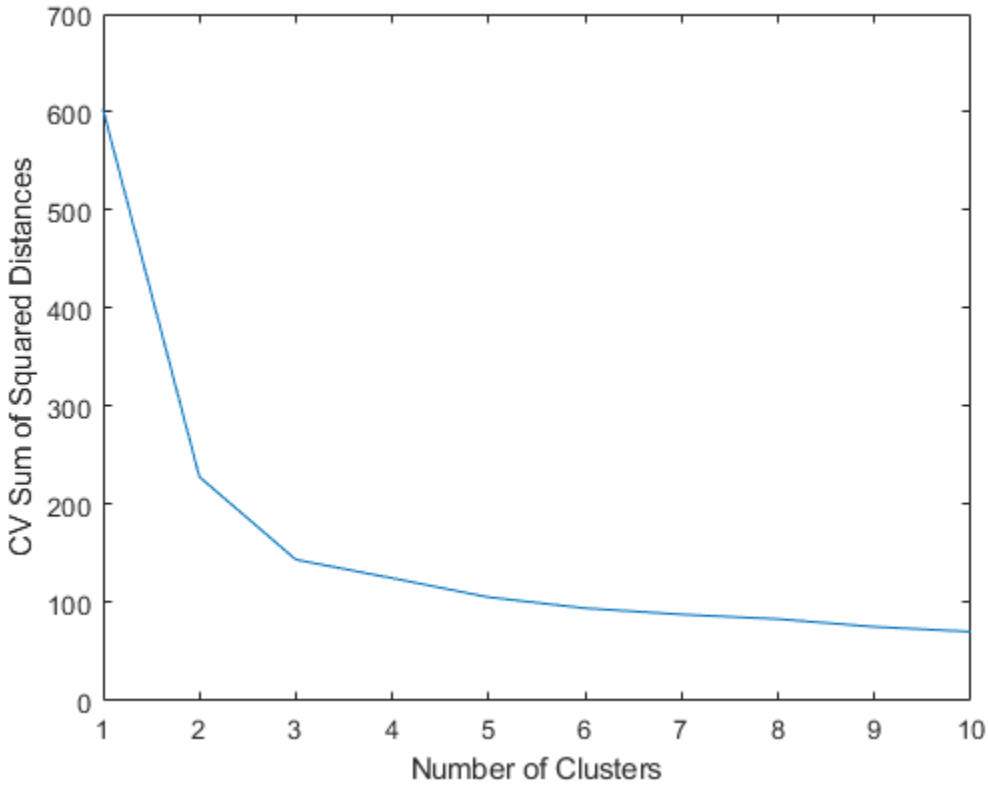
Create a `for` loop that specifies the number of clusters `k` for each iteration. For each fixed number of clusters, pass the corresponding `clustf` function to `crossval`. Because `crossval` performs 10-fold

cross-validation by default, the software computes 10 sums of squared distances, one for each partition of training and test data. Take the sum of those values; the result is the cross-validated sum of squared distances for the given number of clusters.

```
rng('default') % For reproducibility
cvdist = zeros(5,1);
for k = 1:10
    fun = @(Xtrain,Xtest)clustf(Xtrain,Xtest,k);
    distances = crossval(fun,X);
    cvdist(k) = sum(distances);
end
```

Plot the cross-validated sum of squared distances for each number of clusters.

```
plot(cvdist)
xlabel('Number of Clusters')
ylabel('CV Sum of Squared Distances')
```



In general, when determining how many clusters to use, consider the greatest number of clusters that corresponds to a significant decrease in the cross-validated sum of squared distances. For this example, using two or three clusters seems appropriate, but using more than three clusters does not.

This code creates the function `clustf`.

```
function distances = clustf(Xtrain,Xtest,k)
[Ztrain,Zmean,Zstd] = zscore(Xtrain);
[~,C] = kmeans(Ztrain,k); % Creates k clusters
```

```
Ztest = (Xtest-Zmean)./Zstd;
d = pdist2(C,Ztest,'euclidean','Smallest',1);
distances = sum(d.^2);
end
```

Compute Mean Absolute Error Using Cross-Validation

Compute the mean absolute error of a regression model by using 10-fold cross-validation.

Load the `carsmall` data set. Specify the `Acceleration` and `Displacement` variables as predictors and the `Weight` variable as the response.

```
load carsmall
X1 = Acceleration;
X2 = Displacement;
y = Weight;
```

Create the custom function `regf` (shown at the end of this example). This function fits a regression model to training data and then computes predicted car weights on a test set. The function compares the predicted car weight values to the true values, and then computes the mean absolute error (MAE) and the MAE adjusted to the range of the test set car weights.

Note: If you use the live script file for this example, the `regf` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

By default, `crossval` performs 10-fold cross-validation. For each of the 10 training and test set partitions of the data in `X1`, `X2`, and `y`, compute the MAE and adjusted MAE values using the `regf` function. Find the mean MAE and mean adjusted MAE.

```
rng('default') % For reproducibility
values = crossval(@regf,X1,X2,y)
```

```
values = 10×2

  319.2261    0.1132
  342.3722    0.1240
  214.3735    0.0902
  174.7247    0.1128
  189.4835    0.0832
  249.4359    0.1003
  194.4210    0.0845
  348.7437    0.1700
  283.1761    0.1187
  210.7444    0.1325
```

```
mean(values)
```

```
ans = 1×2

  252.6701    0.1129
```

This code creates the function `regf`.

```
function errors = regf(X1train,X2train,ytrain,X1test,X2test,ytest)
tbltrain = table(X1train,X2train,ytrain, ...
    'VariableNames',{'Acceleration','Displacement','Weight'});
tbltest = table(X1test,X2test,ytest, ...
    'VariableNames',{'Acceleration','Displacement','Weight'});
mdl = fitlm(tbltrain,'Weight ~ Acceleration + Displacement');
yfit = predict(mdl,tbltest);
MAE = mean(abs(yfit-tbltest.Weight));
adjMAE = MAE/range(tbltest.Weight);
errors = [MAE adjMAE];
end
```

Compute Misclassification Error Using PCA and Cross-Validation

Compute the misclassification error of a classification tree by using principal component analysis (PCA) and 5-fold cross-validation.

Load the `fisheriris` data set. The `meas` matrix contains flower measurements for 150 different flowers. The `species` variable lists the species for each flower.

```
load fisheriris
```

Create the custom function `classf` (shown at the end of this example). This function fits a classification tree to training data and then classifies test data. Use PCA inside the function to reduce the number of predictors used to create the tree model.

Note: If you use the live script file for this example, the `classf` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Create a `cvpartition` object for stratified 5-fold cross-validation. By default, `cvpartition` ensures that training and test sets have roughly the same proportions of flower species.

```
rng('default') % For reproducibility
cvp = cvpartition(species,'KFold',5);
```

Compute the 5-fold cross-validation misclassification error for the classification tree with predictor data `meas` and response variable `species`.

```
cvError = crossval('mcr',meas,species,'Predfun',@classf,'Partition',cvp)
```

```
cvError = 0.1067
```

This code creates the function `classf`.

```
function yfit = classf(Xtrain,ytrain,Xtest)

% Standardize the training predictor data. Then, find the
% principal components for the standardized training predictor
% data.
[Ztrain,Zmean,Zstd] = zscore(Xtrain);
[coeff,scoreTrain,~,~,explained,mu] = pca(Ztrain);

% Find the lowest number of principal components that account
% for at least 95% of the variability.
```

```
n = find(cumsum(explained)>=95,1);

% Find the n principal component scores for the standardized
% training predictor data. Train a classification tree model
% using only these scores.
scoreTrain95 = scoreTrain(:,1:n);
mdl = fitctree(scoreTrain95,ytrain);

% Find the n principal component scores for the transformed
% test data. Classify the test data.
Ztest = (Xtest-Zmean)./Zstd;
scoreTest95 = (Ztest-mu)*coeff(:,1:n);
yfit = predict(mdl,scoreTest95);

end
```

Create Confusion Matrix Using Cross-Validation

Create a confusion matrix from the 10-fold cross-validation results of a discriminant analysis model.

Note: Use `classify` when training speed is a concern. Otherwise, use `fitcdiscr` to create a discriminant analysis model. For an example that shows the same workflow as this example, but uses `fitcdiscr`, see “Create Confusion Matrix Using Cross-Validation Predictions” on page 33-3304.

Load the `fisheriris` data set. `X` contains flower measurements for 150 different flowers, and `y` lists the species for each flower. Create a variable `order` that specifies the order of the flower species.

```
load fisheriris
X = meas;
y = species;
order = unique(y)
```

```
order = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

Create a function handle named `func` for a function that completes the following steps:

- Take in training data (`Xtrain` and `ytrain`) and test data (`Xtest` and `ytest`).
- Use the training data to create a discriminant analysis model that classifies new data (`Xtest`). Create this model and classify new data by using the `classify` function.
- Compare the true test data classes (`ytest`) to the predicted test data values, and create a confusion matrix of the results by using the `confusionmat` function. Specify the class order by using `'Order',order`.

```
func = @(Xtrain,ytrain,Xtest,ytest)confusionmat(ytest, ...
    classify(Xtest,Xtrain,ytrain),'Order',order);
```

Create a `cvpartition` object for stratified 10-fold cross-validation. By default, `cvpartition` ensures that training and test sets have roughly the same proportions of flower species.

```
rng('default') % For reproducibility
cvp = cvpartition(y,'Kfold',10);
```

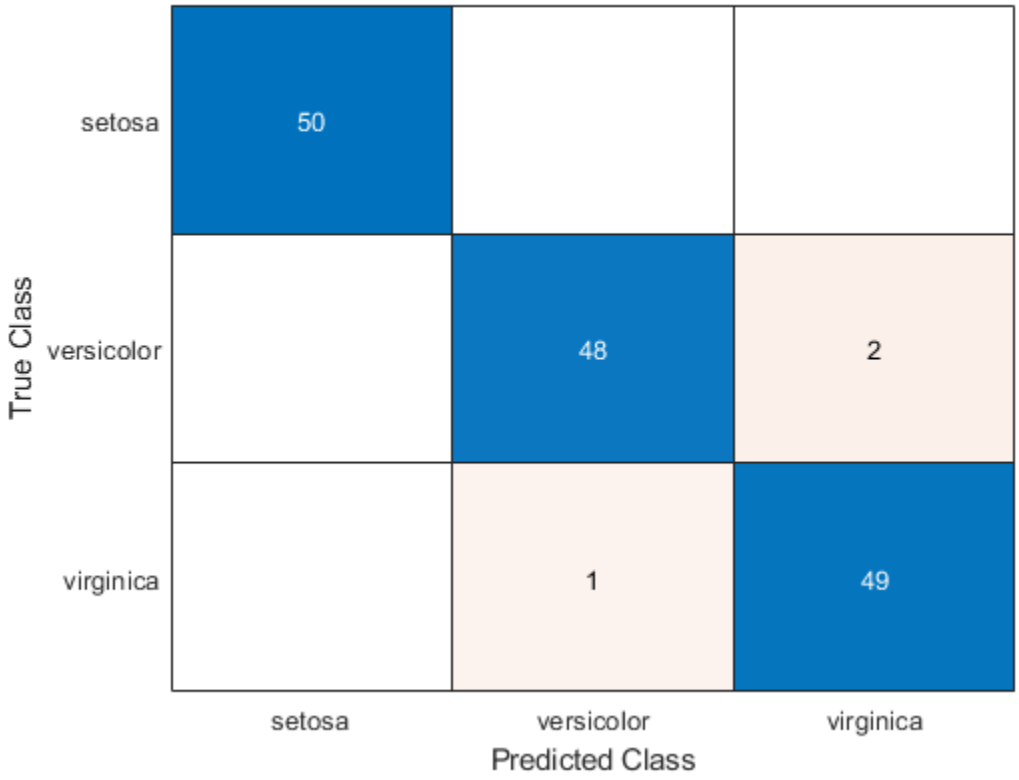
Compute the 10 test set confusion matrices for each partition of the predictor data `X` and response variable `y`. Each row of `confMat` corresponds to the confusion matrix results for one test set. Aggregate the results and create the final confusion matrix `cvMat`.

```
confMat = crossval(func,X,y,'Partition',cvp);
cvMat = reshape(sum(confMat),3,3)
```

```
cvMat = 3×3

    50     0     0
     0    48     2
     0     1    49
```

Plot the confusion matrix as a confusion matrix chart by using `confusionchart`.

```
confusionchart(cvMat,order)
```



Input Arguments

`criterion` — Type of error estimate

`'mse'` | `'mcr'`

Type of error estimate, specified as either `'mse'` or `'mcr'`.

Value	Description
`'mse'`	Mean squared error (MSE) — Appropriate for regression algorithms only
`'mcr'`	Misclassification rate, or proportion of misclassified observations — Appropriate for classification algorithms only

`X` — Data set
column vector | matrix | array

Data set, specified as a column vector, matrix, or array. The rows of `X` correspond to observations, and the columns of `X` generally correspond to variables. If you pass multiple data sets `X1,...,XN` to `crossval`, then all data sets must have the same number of rows.

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`y` — Response data
column vector | character array

Response data, specified as a column vector or character array. The rows of `y` correspond to observations, and `y` must have the same number of rows as the predictor data `X` or `X1,...,XN`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`predfun` — Prediction function
function handle

Prediction function, specified as a function handle. You must create this function as an anonymous function, a function defined at the end of the .m or .mlx file containing the rest of your code, or a file on the MATLAB path.

This table describes the required function syntax, given the type of predictor data passed to `crossval`.

Value	Predictor Data	Function Syntax

`@myfunction`	`X`	`function yfit = myfunction(Xtrain,ytrain,Xtest)`
		`% Calculate predicted response`
		`...`
		`end`
		
		• `Xtrain` — Subset of the observations in `X` used as training predictor data. The function uses `Xtrain` and `ytrain` to construct a classification or regression model.
		• `ytrain` — Subset of the responses in `y` used as training response data. The rows of `ytrain` correspond to the same observations in the rows of `Xtrain`. The function uses `Xtrain` and `ytrain` to construct a classification or regression model.
		• `Xtest` — Subset of the observations in `X` used as test predictor data. The function uses `Xtest` and the model trained on `Xtrain` and `ytrain` to compute the predicted values `yfit`.
		• `yfit` — Set of predicted values for observations in `Xtest`. The `yfit` values form a column vector with the same number of rows as `Xtest`.

Value	Predictor Data	Function Syntax

`@myfunction`	`X1,...,XN`	`function yfit = myfunction(X1train,...,XNtrain,ytrain`
		`% Calculate predicted response`
		`...`
		`end`
		• `X1train,...,XNtrain` — Subsets of the predictor data in `X1,...,XN`, respectively, that are used as training predictor data. The rows of `X1train,...,XNtrain` correspond to the same observations. The function uses `X1train,...,XNtrain` and `ytrain` to construct a classification or regression model.
		• `ytrain` — Subset of the responses in `y` used as training response data. The rows of `ytrain` correspond to the same observations in the rows of `X1train,...,XNtrain`. The function uses `X1train,...,XNtrain` and `ytrain` to construct a classification or regression model.
		• `X1test,...,XNtest` — Subsets of the observations in `X1,...,XN`, respectively, that are used as test predictor data. The rows of `X1test,...,XNtest` correspond to the same observations. The function uses `X1test,...,XNtest` and the model trained on `X1train,...,XNtrain` and `ytrain` to compute the predicted values `yfit`.
		• `yfit` — Set of predicted values for observations in `X1test,...,XNtest`. The `yfit` values form a column vector with the same number of rows as `X1test,...,XNtest`.

Example: `@(Xtrain,ytrain,Xtest)(Xtest*regress(ytrain,Xtrain));`

Data Types: `function_handle`

`fun` — Function to cross-validate
function handle

Function to cross-validate, specified as a function handle. You must create this function as an anonymous function, a function defined at the end of the .m or .mlx file containing the rest of your code, or a file on the MATLAB path.

This table describes the required function syntax, given the type of data passed to `crossval`.

Value	Data	Function Syntax

`@myfunction`	`X`	`function value = myfunction(Xtrain,Xtest)`
		`% Calculation of value`
		`...`
		`end`
		• `Xtrain` — Subset of the observations in `X` used as training data. The function uses `Xtrain` to construct a model.
		• `Xtest` — Subset of the observations in `X` used as test data. The function uses `Xtest` and the model trained on `Xtrain` to compute `value`.
		• `value` — Quantity or variable. In most cases, `value` is a numeric scalar representing a loss estimate. `value` can also be an array, provided that the array size is the same for each partition of training and test data. If you want to return a variable `output` that can change size depending on the data partition, set `value` to be the cell scalar `{output}` instead.

`@myfunction`	`X1,...,XN`	`function value = myfunction(X1train,...,XNtrain,X1tes`
		`% Calculation of value`
		`...`
		`end`
		• `X1train,...,XNtrain` — Subsets of the data in `X1,...,XN`, respectively, that are used as training data. The rows of `X1train,...,XNtrain` correspond to the same observations. The function uses `X1train,...,XNtrain` to construct a model.
		• `X1test,...,XNtest` — Subsets of the data in `X1,...,XN`, respectively, that are used as test data. The rows of `X1test,...,XNtest` correspond to the same observations. The function uses `X1test,...,XNtest` and the model trained on `X1train,...,XNtrain` to compute `value`.
		• `value` — Quantity or variable. In most cases, `value` is a numeric scalar representing a loss estimate. `value` can also be an array, provided that the array size is the same for each partition of training and test data. If you want to return a variable `output` that can change size depending on the data partition, set `value` to be the cell scalar `{output}` instead.

Data Types: `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example:
`crossval('mcr',meas,species,'Predfun',@classf,'KFold',5,'Stratify',species)` specifies to compute the stratified 5-fold cross-validation misclassification rate for the `classf` function with predictor data `meas` and response variable `species`.

`Holdout` — Fraction or number of observations used for holdout validation

`[]` (default) | scalar value in the range (0,1) | positive integer scalar

Fraction or number of observations used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range (0,1) or a positive integer scalar.

- If the `Holdout` value `p` is a scalar in the range (0,1), then `crossval` randomly selects and reserves approximately `p*100`% of the observations as test data.
- If the `Holdout` value `p` is a positive integer scalar, then `crossval` randomly selects and reserves `p` observations as test data.

In either case, `crossval` then trains the model specified by either `fun` or `predfun` using the rest of the data. Finally, the function uses the test data along with the trained model to compute either `values` or `err`.

You can use only one of these four name-value pair arguments: `Holdout`, `KFold`, `Leaveout`, and `Partition`.

Example: `'Holdout',0.3`

Example: `'Holdout',50`

Data Types: `single` | `double`

`KFold` — Number of folds

`10` (default) | positive integer scalar greater than 1

Number of folds for k-fold cross-validation, specified as the comma-separated pair consisting of `'KFold'` and a positive integer scalar greater than 1.

If you specify `'KFold',k`, then `crossval` randomly partitions the data into `k` sets. For each set, the function reserves the set as test data, and trains the model specified by either `fun` or `predfun` using the other `k` – 1 sets. `crossval` then uses the test data along with the trained model to compute either `values` or `err`.

You can use only one of these four name-value pair arguments: `Holdout`, `KFold`, `Leaveout`, and `Partition`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation

`[]` (default) | `1`

Leave-one-out cross-validation, specified as the comma-separated pair consisting of `'Leaveout'` and `1`.

If you specify `'Leaveout',1`, then for each observation, `crossval` reserves the observation as test data, and trains the model specified by either `fun` or `predfun` using the other observations. The function then uses the test observation along with the trained model to compute either `values` or `err`.

You can use only one of these four name-value pair arguments: `Holdout`, `KFold`, `Leaveout`, and `Partition`.

Example: `'Leaveout',1`

Data Types: `single` | `double`

`MCReps` — Number of Monte Carlo repetitions
1 (default) | positive integer scalar

Number of Monte Carlo repetitions for validation, specified as the comma-separated pair consisting of `'MCReps'` and a positive integer scalar. If the first input of `crossval` is `'mse'` or `'mcr'` (see `criterion`), then `crossval` returns the mean MSE or misclassification rate across all Monte Carlo repetitions. Otherwise, `crossval` concatenates the values from all Monte Carlo repetitions along the first dimension.

If you specify both `Partition` and `MCReps`, then the first Monte Carlo repetition uses the partition information in the `cvpartition` object, and the software calls the `repartition` object function to generate new partitions for each of the remaining repetitions.

Example: `'MCReps',5`

Data Types: `single` | `double`

`Partition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as the comma-separated pair consisting of `'Partition'` and a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and test sets.

When you use `crossval`, you cannot specify both `Partition` and `Stratify`. Instead, directly specify a stratified partition when you create the `cvpartition` partition object.

You can use only one of these four name-value pair arguments: `Holdout`, `KFold`, `Leaveout`, and `Partition`.

`Stratify` — Variable specifying groups used for stratification
column vector

Variable specifying the groups used for stratification, specified as the comma-separated pair consisting of `'Stratify'` and a column vector with the same number of rows as the data `X` or `X1,...,XN`.

When you specify `Stratify`, both the training and test sets have roughly the same class proportions as in the `Stratify` vector. The software treats `NaN`s, empty character vectors, empty strings, `<missing>` values, and `<undefined>` values in `Stratify` as missing data values, and ignores the corresponding rows of the data.

A good practice is to use stratification when you use cross-validation with classification algorithms. Otherwise, some test sets might not include observations for all classes.

When you use `crossval`, you cannot specify both `Partition` and `Stratify`. Instead, directly specify a stratified partition when you create the `cvpartition` partition object.

Data Types: `single` | `double` | `logical` | `string` | `cell` | `categorical`

`Options` — Options for running in parallel and setting random streams
structure

Options for running computations in parallel and setting random streams, specified as a structure. Create the `Options` structure with `statset`. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to run computations in parallel.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or a cell array consisting of one such object.	If you do not specify `Streams`, then `crossval` uses the default stream.

Note You need Parallel Computing Toolbox to run computations in parallel.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Output Arguments

`err` — Mean squared error or misclassification rate
numeric scalar

Mean squared error or misclassification rate, returned as a numeric scalar. The type of error depends on the `criterion` value.

`values` — Loss values
column vector | matrix

Loss values, returned as a column vector or matrix. Each row of `values` corresponds to the output of `fun` for one partition of training and test data.

If the output returned by `fun` is multidimensional, then `crossval` reshapes the output and fits it into one row of `values`. For an example, see “Create Confusion Matrix Using Cross-Validation” on page 33-1050.

Tips

- A good practice is to use stratification (see `Stratify`) when you use cross-validation with classification algorithms. Otherwise, some test sets might not include observations for all classes.

Algorithms

General Cross-Validation Steps for `predfun`

When you use `predfun`, the `crossval` function typically performs 10-fold cross-validation as follows:

1. Split the observations in the predictor data `X` and the response variable `y` into 10 groups, each of which has approximately the same number of observations.
2. Use the last nine groups of observations to train a model as specified in `predfun`. Use the first group of observations as test data, pass the test predictor data to the trained model, and compute predicted values as specified in `predfun`. Compute the error specified by `criterion`.
3. Use the first group and the last eight groups of observations to train a model as specified in `predfun`. Use the second group of observations as test data, pass the test data to the trained model, and compute predicted values as specified in `predfun`. Compute the error specified by `criterion`.
4. Proceed in a similar manner until each group of observations is used as test data exactly once.
5. Return the mean error estimate as the scalar `err`.

General Cross-Validation Steps for `fun`

When you use `fun`, the `crossval` function typically performs 10-fold cross-validation as follows:

1. Split the data in `X` into 10 groups, each of which has approximately the same number of observations.
2. Use the last nine groups of data to train a model as specified in `fun`. Use the first group of data as a test set, pass the test set to the trained model, and compute some value (for example, loss) as specified in `fun`.
3. Use the first group and the last eight groups of data to train a model as specified in `fun`. Use the second group of data as a test set, pass the test set to the trained model, and compute some value as specified in `fun`.
4. Proceed in a similar manner until each group of data is used as a test set exactly once.
5. Return the 10 computed values as the vector `values`.

Alternative Functionality

Many classification and regression functions allow you to perform cross-validation directly.

- When you use fit functions such as `fitcsvm`, `fitctree`, and `fitrtree`, you can specify cross-validation options by using name-value pair arguments. Alternatively, you can first create models with these fit functions and then create a partitioned object by using the `crossval` object function. Use the `kfoldLoss` and `kfoldPredict` object functions to compute the loss and

predicted values for the partitioned object. For more information, see `ClassificationPartitionedModel` and `RegressionPartitionedModel`.

- You can also specify cross-validation options when you perform lasso or elastic net regularization using `lasso` and `lassoglm`.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`classify` | `confusionmat` | `cvpartition` | `kmeans` | `pca` | `regress`

Topics

“Selecting Features for Classifying High-dimensional Data” on page 15-171

Introduced in R2008a

crossval

Cross-validate machine learning model

Syntax

```
CVMdl = crossval(Mdl)
CVMdl = crossval(Mdl,Name,Value)
```

Description

`CVMdl = crossval(Mdl)` returns a cross-validated (partitioned) machine learning model (`CVMdl`) from a trained model (`Mdl`). By default, `crossval` uses 10-fold cross-validation on the training data.

`CVMdl = crossval(Mdl,Name,Value)` sets an additional cross-validation option. You can specify only one name-value argument. For example, you can specify the number of folds or a holdout sample proportion.

Examples

Cross-Validate SVM Classifier

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
rng(1); % For reproducibility
```

Train a support vector machine (SVM) classifier. Standardize the predictor data and specify the order of the classes.

```
SVMModel = fitcsvm(X,Y,'Standardize',true,'ClassNames',{'b','g'});
```

`SVMModel` is a trained `ClassificationSVM` classifier. `'b'` is the negative class and `'g'` is the positive class.

Cross-validate the classifier using 10-fold cross-validation.

```
CVSVMModel = crossval(SVMModel)
```

```
CVSVMModel =
  ClassificationPartitionedModel
    CrossValidatedModel: 'SVM'
         PredictorNames: {1x34 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'
```

```
  Properties, Methods
```

`CVSVMModel` is a `ClassificationPartitionedModel` cross-validated classifier. During cross-validation, the software completes these steps:

1. Randomly partition the data into 10 sets of equal size.
2. Train an SVM classifier on nine of the sets.
3. Repeat steps 1 and 2 $k = 10$ times. The software leaves out one partition each time and trains on the other nine partitions.
4. Combine generalization statistics for each fold.

Display the first model in `CVSVMModel.Trained`.

```
FirstModel = CVSVMModel.Trained{1}
```

```
FirstModel =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
                    Alpha: [78x1 double]
                     Bias: -0.2209
         KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
           SupportVectors: [78x34 double]
      SupportVectorLabels: [78x1 double]


  Properties, Methods
```

`FirstModel` is the first of the 10 trained classifiers. It is a `CompactClassificationSVM` classifier.

You can estimate the generalization error by passing `CVSVMModel` to `kfoldLoss`.

Specify Holdout Sample Proportion for Naive Bayes Cross-Validation

Specify a holdout sample proportion for cross-validation. By default, `crossval` uses 10-fold cross-validation to cross-validate a naive Bayes classifier. However, you have several other options for cross-validation. For example, you can specify a different number of folds or a holdout sample proportion.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Remove the first two predictors for stability.

```
X = X(:,3:end);
rng('default'); % For reproducibility
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `'b'` is the negative class and `'g'` is the positive class. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'b','g'});
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Cross-validate the classifier by specifying a 30% holdout sample.

```
CVMdl = crossval(Mdl,'Holdout',0.3)
```

```
CVMdl =
  ClassificationPartitionedModel
    CrossValidatedModel: 'NaiveBayes'
         PredictorNames: {1x32 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedModel` cross-validated, naive Bayes classifier.

Display the properties of the classifier trained using 70% of the data.

```
TrainedModel = CVMdl.Trained{1}
```

```
TrainedModel =
  CompactClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
         DistributionNames: {1x32 cell}
    DistributionParameters: {2x32 cell}


  Properties, Methods
```

`TrainedModel` is a `CompactClassificationNaiveBayes` classifier.

Estimate the generalization error by passing `CVMdl` to `kfoldloss`.

```
kfoldLoss(CVMdl)
```

```
ans = 0.2095
```

The out-of-sample misclassification error is approximately 21%.

Reduce the generalization error by choosing the five most important predictors.

```
idx = fscmrmr(X,Y);
Xnew = X(:,idx(1:5));
```

Train a naive Bayes classifier for the new predictor.

```
Mdlnew = fitcnb(Xnew,Y,'ClassNames',{'b','g'});
```

Cross-validate the new classifier by specifying a 30% holdout sample, and estimate the generalization error.

```
CVMdlnew = crossval(Mdlnew,'Holdout',0.3);
kfoldLoss(CVMdlnew)
```

```
ans = 0.1429
```

The out-of-sample misclassification error is reduced from approximately 21% to approximately 14%.

Create Cross-Validated Regression GAM Using `crossval`

Train a regression generalized additive model (GAM) by using `fitrgam`, and create a cross-validated GAM by using `crossval` and the holdout option. Then, use `kfoldPredict` to predict responses for validation-fold observations using a model trained on training-fold observations.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Train a GAM that contains linear terms for predictors.

```
Mdl = fitrgam(tbl,'Systolic');
```

`Mdl` is a `RegressionGAM` model object.

Cross-validate the model by specifying a 30% holdout sample.

```
rng('default') % For reproducibility
CVMdl = crossval(Mdl,'Holdout',0.3)
```

```
CVMdl =
  RegressionPartitionedGAM
      CrossValidatedModel: 'GAM'
           PredictorNames: {1x6 cell}
    CategoricalPredictors: [3 5 6]
             ResponseName: 'Systolic'
          NumObservations: 100
                    KFold: 1
                Partition: [1x1 cvpartition]
        NumTrainedPerFold: [1x1 struct]
        ResponseTransform: 'none'
```

```
  Properties, Methods
```

The `crossval` function creates a `RegressionPartitionedGAM` model object `CVMdl` with the holdout option. During cross-validation, the software completes these steps:

1. Randomly select and reserve 30% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model object `RegressionPartitionedGAM`.

You can choose a different cross-validation setting by using the `'CrossVal'`, `'CVPartition'`, `'KFold'`, or `'Leaveout'` name-value argument.

Predict responses for the validation-fold observations by using `kfoldPredict`. The function predicts responses for the validation-fold observations by using the model trained on the training-fold observations. The function assigns `NaN` to the training-fold observations.

```
yFit = kfoldPredict(CVMdl);
```

Find the validation-fold observation indexes, and create a table containing the observation index, observed response values, and predicted response values. Display the first eight rows of the table.

```
idx = find(~isnan(yFit));
t = table(idx,tbl.Systolic(idx),yFit(idx), ...
    'VariableNames',{'Obseraction Index','Observed Value','Predicted Value'});
head(t)
```

```
ans=8×3 table
    Obseraction Index    Observed Value    Predicted Value
    _________________    ______________    _______________

            1                124              130.22
            6                121              124.38
            7                130              125.26
           12                115              117.05
           20                125              121.82
           22                123              116.99
           23                114                 107
           24                128              122.52
```

Compute the regression error (mean squared error) for the validation-fold observations.

```
L = kfoldLoss(CVMdl)
```

```
L = 43.8715
```

Input Arguments

`Mdl` — Machine learning model

full regression model object | full classification model object

Machine learning model, specified as a full regression or classification model object, as given in the following tables of supported models.

Regression Model Object

Model	Full Regression Model Object
Generalized additive model	`RegressionGAM`
Neural network model	`RegressionNeuralNetwork`

Classification Model Object

Model	Full Classification Model Object
Generalized additive model	`ClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `crossval(Mdl,'KFold',3)` specifies using three folds in a cross-validated model.

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

You can specify only one of these four name-value arguments: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model. If `Mdl` does not have a corresponding compact object, then `Trained` contains a full object.

You can specify only one of these four name-value arguments: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model. If `Mdl` does not have a corresponding compact object, then `Trained` contains a full object.

You can specify only one of these four name-value arguments: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model. If `Mdl` does not have a corresponding compact object, then `Trained` contains a full object.

You can specify only one of these four name-value arguments: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'Leaveout','on'`

Output Arguments

`CVMdl` — Cross-validated machine learning model
cross-validated (partitioned) model object

Cross-validated machine learning model, returned as one of the cross-validated (partitioned) model objects in the following tables, depending on the input model `Mdl`.

Regression Model Object

Model	Regression Model (`Mdl`)	Cross-Validated Model (`CompactMdl`)
Generalized additive model	`RegressionGAM`	`RegressionPartitionedGAM`
Neural network model	`RegressionNeuralNetwork`	`RegressionPartitionedModel`

Classification Model Object

Model	Classification Model (`Mdl`)	Cross-Validated Model (`CompactMdl`)
Generalized additive model	`ClassificationGAM`	`ClassificationPartitionedGAM`
k-nearest neighbor model	`ClassificationKNN`	`ClassificationPartitionedModel`
Naive Bayes model	`ClassificationNaiveBayes`	`ClassificationPartitionedModel`
Neural network model	`ClassificationNeuralNetwork`	`ClassificationPartitionedModel`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`ClassificationPartitionedModel`

Tips

- Assess the predictive performance of `Mdl` on cross-validated data by using the *kfold* functions and properties of `CVMdl`, such as `kfoldPredict` and `kfoldLoss`.
- Return a partitioned classifier with stratified partitioning by using the name-value argument `'KFold'` or `'Holdout'`.
- Create a `cvpartition` object `cvp` using `cvp = cvpartition(n,'KFold',k)`. Return a partitioned classifier with nonstratified partitioning by using the name-value argument `'CVPartition',cvp`.

Alternative Functionality

Instead of training a model and then cross-validating it, you can create a cross-validated model directly by using a fitting function and specifying one of these name-value argument: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'Leaveout'`, or `'KFold'`.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports `ClassificationKNN` and `ClassificationSVM` objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

Introduced in R2012a

crossval

Class: `ClassificationDiscriminant`

Cross-validated discriminant analysis classifier

Syntax

```
cvmodel = crossval(obj)
cvmodel = crossval(obj,Name,Value)
```

Description

`cvmodel = crossval(obj)` creates a partitioned model from `obj`, a fitted discriminant analysis classifier. By default, `crossval` uses 10-fold cross validation on the training data to create `cvmodel`.

`cvmodel = crossval(obj,Name,Value)` creates a partitioned model with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier, produced using `fitcdiscr`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CVPartition`

Object of class `cvpartition`, created by the `cvpartition` function. `crossval` splits the data into subsets with `cvpartition`.

Use only one of these options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Default: `[]`

`Holdout`

Holdout validation tests the specified fraction of the data, and uses the rest of the data for training. Specify a numeric scalar from `0` to `1`. Use only one of these options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

`KFold`

Number of folds to use in a cross-validated classifier, a positive integer value greater than 1.

Use only one of these options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Default: 10

`Leaveout`

Set to `'on'` for leave-one-out cross validation.

Use only one of these options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Examples

Create a classification model for the Fisher iris data, and then create a cross-validation model. Evaluate the quality the model using `kfoldLoss`.

```
load fisheriris
obj = fitcdiscr(meas,species);
cvmodel = crossval(obj);
L = kfoldLoss(cvmodel)

L =
    0.0200
```

Tips

- Assess the predictive performance of `obj` on cross-validated data using the “kfold” methods and properties of `cvmodel`, such as `kfoldLoss`.

Alternatives

You can create a cross-validation classifier directly from the data, instead of creating a discriminant analysis classifier followed by a cross-validation classifier. To do so, include one of these options in `fitcdiscr`: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

See Also

`crossval` | `fitcdiscr` | `kfoldEdge` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict` | `kfoldfun`

Topics

“Discriminant Analysis Classification” on page 20-2

crossval

Cross-validate multiclass error-correcting output codes (ECOC) model

Syntax

```
CVMdl = crossval(Mdl)
CVMdl = crossval(Mdl,Name,Value)
```

Description

`CVMdl = crossval(Mdl)` returns a cross-validated (partitioned) multiclass error-correcting output codes (ECOC) model (`CVMdl`) from a trained ECOC model (`Mdl`). By default, `crossval` uses 10-fold cross-validation on the training data to create `CVMdl`, a `ClassificationPartitionedECOC` model.

`CVMdl = crossval(Mdl,Name,Value)` returns a partitioned ECOC model with additional options specified by one or more name-value pair arguments. For example, you can specify the number of folds or a holdout sample proportion.

Examples

Cross-Validate ECOC Classifier

Cross-validate an ECOC classifier with SVM binary learners, and estimate the generalized classification error.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
rng(1); % For reproducibility
```

Create an SVM template, and standardize the predictors.

```
t = templateSVM('Standardize',true)
```

```
t =
Fit template for classification SVM.

                    Alpha: [0x1 double]
            BoxConstraint: []
                CacheSize: []
            CachingMethod: ''
               ClipAlphas: []
    DeltaGradientTolerance: []
                  Epsilon: []
             GapTolerance: []
             KKTTolerance: []
           IterationLimit: []
           KernelFunction: ''
              KernelScale: []
```

```
          KernelOffset: []
KernelPolynomialOrder: []
              NumPrint: []
                    Nu: []
       OutlierFraction: []
     RemoveDuplicates: []
      ShrinkagePeriod: []
               Solver: ''
      StandardizeData: 1
   SaveSupportVectors: []
       VerbosityLevel: []
              Version: 2
               Method: 'SVM'
                 Type: 'classification'
```

`t` is an SVM template. Most of the template object properties are empty. When training the ECOC classifier, the software sets the applicable properties to their default values.

Train the ECOC classifier, and specify the class order.

```
Mdl = fitcecoc(X,Y,'Learners',t,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a `ClassificationECOC` classifier. You can access its properties using dot notation.

Cross-validate `Mdl` using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

`CVMdl` is a `ClassificationPartitionedECOC` cross-validated ECOC classifier.

Estimate the generalized classification error.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.0400
```

The generalized classification error is 4%, which indicates that the ECOC classifier generalizes fairly well.

Cross-Validate ECOC Classifier Using Parallel Computing

Consider the `arrhythmia` data set. This data set contains 16 classes, 13 of which are represented in the data. The first class indicates that the subject does not have arrhythmia, and the last class indicates that the arrhythmia state of the subject is not recorded. The other classes are ordinal levels indicating the severity of arrhythmia.

Train an ECOC classifier with a custom coding design specified by the description of the classes.

Load the `arrhythmia` data set. Convert `Y` to a `categorical` variable, and determine the number of classes.

```
load arrhythmia
Y = categorical(Y);
K = unique(Y); % Number of distinct classes
```

Construct a coding matrix that describes the nature of the classes.

```
OrdMat = designecoc(11,'ordinal');
nOrdMat = size(OrdMat);
class1VSOrd = [1; -ones(11,1); 0];
class1VSClass16 = [1; zeros(11,1); -1];
OrdVSClass16 = [0; ones(11,1); -1];
Coding = [class1VSOrd class1VSClass16 OrdVSClass16,...
    [zeros(1,nOrdMat(2)); OrdMat; zeros(1,nOrdMat(2))]];
```

Train an ECOC classifier using the custom coding design (`Coding`) and parallel computing. Specify an ensemble of 50 classification trees boosted using GentleBoost.

```
t = templateEnsemble('GentleBoost',50,'Tree');
options = statset('UseParallel',true);
Mdl = fitcecoc(X,Y,'Coding',Coding,'Learners',t,'Options',options);

Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

`Mdl` is a `ClassificationECOC` model. You can access its properties using dot notation.

Cross-validate `Mdl` using 8-fold cross-validation and parallel computing.

```
rng(1); % For reproducibility
CVMdl = crossval(Mdl,'Options',options,'KFold',8);

Warning: One or more folds do not contain points from all the groups.
```

Because some classes have low relative frequency, some folds do not train using observations from those classes. `CVMdl` is a `ClassificationPartitionedECOC` cross-validated ECOC model.

Estimate the generalization error using parallel computing.

```
error = kfoldLoss(CVMdl,'Options',options)

error = 0.3208
```

The cross-validated classification error is 32%, which indicates that this model does not generalize well. To improve the model, try training using a different boosting method, such as RobustBoost, or a different algorithm, such as SVM.

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `crossval(Mdl,'KFold',3)` specifies using three folds in a cross-validated model.

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

Tips

- Assess the predictive performance of `Mdl` on cross-validated data using the "kfold" methods and properties of `CVMdl`, such as `kfoldLoss`.

Alternative Functionality

Instead of training an ECOC model and then cross-validating it, you can create a cross-validated ECOC model directly by using `fitcecoc` and specifying one of these name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'Leaveout'`, or `'KFold'`.

Extended Capabilities

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `CompactClassificationECOC` | `cvpartition` | `fitcecoc` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

crossval

Cross validate ensemble

Syntax

```
cvens = crossval(ens)
cvens = crossval(ens,Name,Value)
```

Description

`cvens = crossval(ens)` creates a cross-validated ensemble from `ens`, a classification ensemble. Default is 10-fold cross validation.

`cvens = crossval(ens,Name,Value)` creates a cross-validated ensemble with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A classification ensemble created with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`cvpartition`

A partition of class `cvpartition`. Sets the partition for cross validation.

Use no more than one of the name-value pairs `cvpartition`, `holdout`, `kfold`, or `leaveout`.

`holdout`

Holdout validation tests the specified fraction of the data, and uses the rest of the data for training. Specify a numeric scalar from `0` to `1`. You can only use one of these four options at a time for creating a cross-validated tree: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`kfold`

Number of folds for cross validation, a numeric positive scalar greater than 1.

Use no more than one of the name-value pairs `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`leaveout`

If `'on'`, use leave-one-out cross validation.

Use no more than one of the name-value pairs `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`nprint`

Printout frequency, a positive integer scalar. Use this parameter to observe the training of cross-validation folds.

Default: `'off'`, meaning no printout

Output Arguments

`cvens`

A cross-validated classification ensemble of class `ClassificationPartitionedEnsemble`.

Examples

Cross-Validate Classification Ensemble

Create a cross-validated classification model for the Fisher iris data, and assess its quality using the `kfoldLoss` method.

Load the Fisher iris data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Create a cross-validated ensemble from `ens` and find the classification error averaged over all folds.

```
rng(10,'twister') % For reproducibility
cvens = crossval(ens);
L = kfoldLoss(cvens)
```

```
L = 0.0533
```

Alternatives

You can create a cross-validation ensemble directly from the data, instead of creating an ensemble followed by a cross-validation ensemble. To do so, include one of these five options in `fitcensemble`: `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

See Also

`ClassificationPartitionedEnsemble` | `cvpartition`

crossval

Class: `ClassificationTree`

Cross-validated decision tree

Syntax

```
cvmodel = crossval(model)
cvmodel = crossval(model,Name,Value)
```

Description

`cvmodel = crossval(model)` creates a partitioned model from `model`, a fitted classification tree. By default, `crossval` uses 10-fold cross validation on the training data to create `cvmodel`.

`cvmodel = crossval(model,Name,Value)` creates a partitioned model with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`model`

A classification model, produced using `fitctree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` object

Cross-validation partition, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` object created by the `cvpartition` function. `crossval` splits the data into subsets with `cvpartition`.

Use only one of these four options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range `(0,1)`

Fraction of the data used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range `(0,1)`.

Use only one of these four options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'Holdout',0.3`

Data Types: `single` | `double`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1.

Use only one of these four options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'KFold',3`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. Leave-one-out is a special case of `'KFold'` in which the number of folds equals the number of observations.

Use only one of these four options at a time: `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Example: `'Leaveout','on'`

Output Arguments

`cvmodel` — Partitioned model
`ClassificationPartionedModel` object

Partitioned model, returned as a `ClassificationPartitionedModel` object.

Examples

Create a Cross-Validation Model

Create a classification model for the ionosphere data, then create a cross-validation model. Evaluate the quality the model using `kfoldLoss`.

```
load ionosphere
tree = fitctree(X,Y);
cvmodel = crossval(tree);
L = kfoldLoss(cvmodel)
```

```
L = 0.1083
```

Tips

- Assess the predictive performance of `model` on cross-validated data using the “kfold” methods and properties of `cvmodel`, such as `kfoldLoss`.

Alternatives

You can create a cross-validation tree directly from the data, instead of creating a decision tree followed by a cross-validation tree. To do so, include one of these five options in `fitctree`: `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

See Also

`crossval` | `fitctree`

crossval

Cross validate ensemble

Syntax

```
cvens = crossval(ens)
cvens = crossval(ens,Name,Value)
```

Description

`cvens = crossval(ens)` creates a cross-validated ensemble from `ens`, a regression ensemble. Default is 10-fold cross validation.

`cvens = crossval(ens,Name,Value)` creates a cross-validated ensemble with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`cvpartition`

A partition of class `cvpartition`. Sets the partition for cross validation.

Use no more than one of the name-value pairs `cvpartition`, `holdout`, `kfold`, and `leaveout`.

`holdout`

Holdout validation tests the specified fraction of the data, and uses the rest of the data for training. Specify a numeric scalar from `0` to `1`. You can only use one of these four options at a time for creating a cross-validated tree: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`kfold`

Number of folds for cross validation, a positive integer value greater than 1.

Use no more than one of the name-value pairs `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`leaveout`

If `'on'`, use leave-one-out cross-validation.

Use no more than one of the name-value pairs `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`nprint`

Printout frequency, a positive integer scalar. Use this parameter to observe the training of cross-validation folds.

Default: `'off'`, meaning no printout

Output Arguments

`cvens`

A cross-validated classification ensemble of class `RegressionPartitionedEnsemble`.

Examples

Create Cross-Validated Regression Model

Create a cross-validated regression model for the `carsmall` data, and evaluate its quality using the `kfoldLoss` method.

Load the `carsmall` data set and select acceleration, displacement, horsepower, and vehicle weight as predictors.

```
load carsmall;
X = [Acceleration Displacement Horsepower Weight];
```

Train a regression ensemble.

```
rens = fitrensemble(X,MPG);
```

Create a cross-validated ensemble from `rens` and find the cross-validation loss.

```
rng(10,'twister') % For reproducibility
cvens = crossval(rens);
L = kfoldLoss(cvens)
```

```
L = 30.3471
```

Alternatives

You can create a cross-validation ensemble directly from the data, instead of creating an ensemble followed by a cross-validation ensemble. To do so, include one of these five options in `fitrensemble`: `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

See Also

`RegressionPartitionedEnsemble` | `cvpartition` | `kfoldLoss`

crossval

Class: `RegressionGP`

Cross-validate Gaussian process regression model

Syntax

```
cvMdl = crossval(gprMdl)
cvmdl = crossval(gprMdl,Name,Value)
```

Description

`cvMdl = crossval(gprMdl)` returns the partitioned model, `cvMdl`, built from the Gaussian process regression (GPR) model, `gprMdl`, using 10-fold cross validation.

`cvmdl` is a `RegressionPartitionedModel` object, and `gprMdl` is a `RegressionGP` (full) object.

`cvmdl = crossval(gprMdl,Name,Value)` returns the partitioned model, `cvmdl`, with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the number of folds or the fraction of the data to use for testing.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` (full) object. You cannot call `crossval` on a compact regression object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CVPartition` — Random partition for a *k*-fold cross validation

`cvpartition` object

Random partition for a *k*-fold cross validation, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` object.

Example: `'CVPartition',cvp` uses the random partition defined by `cvp`.

If you specify `CVPartition`, then you cannot specify `Holdout`, `KFold`, or `LeaveOut`.

`Holdout` — Fraction of data to use for testing

scalar value in the range from 0 to 1

Fraction of the data to use for testing in holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range from 0 to 1. If you specify `'Holdout',p`, then `crossval`:

1. Randomly reserves *p**100% of the data as validation data, and trains the model using the rest of the data
2. Stores the compact, trained model in `cvgprMdl.Trained`.

Example: `'Holdout', 0.3` uses 30% of the data for testing and 70% of the data for training.

If you specify `Holdout`, then you cannot specify `CVPartition`, `KFold`, or `Leaveout`.

Data Types: `single` | `double`

`KFold` — Number of folds
10 (default) | positive integer value greater than 1

Number of folds to use in cross-validated GPR model, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1. `Kfold` must be greater than 1. If you specify `'Kfold',k` then `crossval`:
1. Randomly partitions the data into *k* sets.
2. For each set, reserves the set as test data, and trains the model using the other *k* – 1 sets.
3. Stores the *k* compact, trained models in the cells of a *k*-by-1 cell array in `cvgprMdl.Trained`.

Example: `'KFold',5` uses 5 folds in cross-validation. That is, for each fold, it uses that fold as test data, and trains the model on the remaining 4 folds.

If you specify `KFold`, then you cannot specify `CVPartition`, `Holdout`, or `Leaveout`.

Data Types: `single` | `double`

`Leaveout` — Indicator for leave-one-out cross-validation
`'off'` (default) | `'on'`

Indicator for leave-one-out cross-validation, specified as the comma-separated pair consisting of `'LeaveOut'` and either `'on'` or `'off'`. If you specify `'Leaveout','on'`, then, for each of the *n* observations, `crossval`:
1. Reserves the observation as test data, and trains the model using the other *n* – 1 observations.
2. Stores the *n* compact, trained models in the cells of a *n*-by-1 cell array in `cvgprMdl.Trained`.

Example: `'Leaveout','on'`

If you specify `Leaveout`, then you cannot specify `CVPartition`, `Holdout`, or `KFold`.

Output Arguments

`cvgprMdl` — Partitioned Gaussian process regression model
`RegressionPartitionedModel` object

Partitioned Gaussian process regression model, returned as a `RegressionPartitionedModel` object.

Examples

Partition Data for Cross-Validation

Download the housing data [1], from the UCI Machine Learning Repository [4].

The dataset has 506 observations. The first 13 columns contain the predictor values and the last column contains the response values. The goal is to predict the median value of owner-occupied homes in suburban Boston as a function of 13 predictors.

Load the data and define the response vector and the predictor matrix.

```
load('housing.data');
X = housing(:,1:13);
y = housing(:,end);
```

Fit a GPR model using the squared exponential kernel function with separate length scale for each predictor. Standardize the predictor variables.

```
gprMdl = fitrgp(X,y,'KernelFunction','ardsquaredexponential','Standardize',1);
```

Create a cross-validation partition for data using predictor 4 as a grouping variable.

```
rng('default') % For reproducibility
cvp = cvpartition(X(:,4),'kfold',10);
```

Create a 10-fold cross-validated model using the partitioned data in `cvp`.

```
cvgprMdl = crossval(gprMdl,'CVPartition',cvp);
```

Compute the regression loss for in-fold observations using models trained on out-of-fold observations.

```
L = kfoldLoss(cvgprMdl)
```

```
L =

    9.5299
```

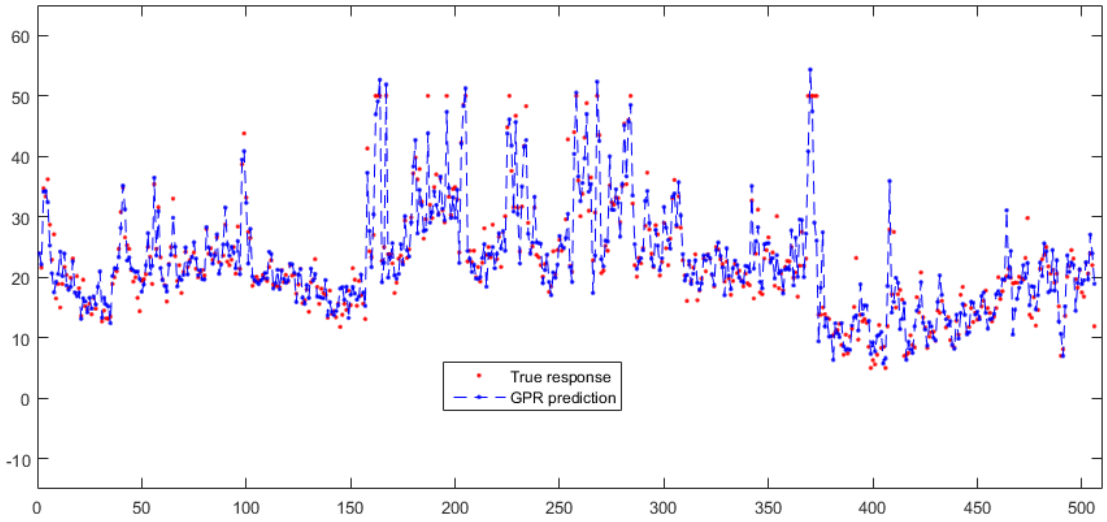
Predict the response for in-fold observations, i.e. observations not used for training.

```
ypred = kfoldPredict(cvgprMdl);
```

For every fold, `kfoldPredict` predicts responses for observations in that fold using the models trained on out-of-fold observations.

Plot the actual responses and prediction data.

```
plot(y,'r.');
hold on;
plot(ypred,'b--.');
axis([0 510 -15 65]);
legend('True response','GPR prediction','Location','Best');
hold off;
```



Train GPR Model Using 4-Fold Cross Validation

Download the abalone data [2], [3], from the UCI Machine Learning Repository [4] and save it in your current directory with the name `abalone.data`.

Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
```

The dataset has 4177 observations. The goal is to predict the age of abalone from 8 physical measurements.

Fit a GPR model using the subset of regressors (`sr`) method for parameter estimation and fully independent conditional (`fic`) method for prediction. Standardize the predictors and use a squared exponential kernel function with a separate length scale for each predictor.

```
gprMdl = fitrgp(tbl,tbl(:,end),'KernelFunction','ardsquaredexponential',...
      'FitMethod','sr','PredictMethod','fic','Standardize',1);
```

Cross-validate the model using 4-fold cross validation. This partitions the data into 4 sets. For each set, `fitrgp` uses that set (25% of the data) as the test data, and trains the model on the remaining 3 sets (75% of the data).

```
rng('default') % For reproducibility
cvgprMdl = crossval(gprMdl,'KFold',4);
```

Compute the loss over individual folds.

```
L = kfoldLoss(cvgprMdl,'mode','individual')
```

```
L =

    4.3669
    4.6896
```

```
    4.0565
    4.3162
```

Compute the average cross-validated loss on over all folds. The default is the mean squared error.

```
L2 = kfoldLoss(cvgprMdl)
```

```
L2 =

    4.3573
```

This is equal to the mean loss over individual folds.

```
mse = mean(L)
```

```
mse =

    4.3573
```

Tips

- You can only use one of the name-value pair arguments at a time.
- You cannot compute the prediction intervals for a cross-validated model.

Alternatives

Alternatively, you can train a cross-validated model using the related name-value pair arguments in `fitrgp`.

If you supply a custom `'ActiveSet'` in the call to `fitrgp`, then you cannot cross validate the GPR model.

References

[1] Harrison, D. and D.L., Rubinfeld. "Hedonic prices and the demand for clean air." *J. Environ. Economics & Management*. Vol.5, 1978, pp. 81-102.

[2] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[3] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[4] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

See Also

`RegressionGP` | `RegressionPartitionedModel` | `fitrgp` | `kfoldLoss` | `kfoldPredict`

Topics

“Train GPR Model Using Cross-Validation” on page 33-2148

Introduced in R2015b

crossval

Class: `RegressionSVM`

Cross-validated support vector machine regression model

Syntax

```
CVMdl = crossval(mdl)
CVMdl = crossval(mdl,Name,Value)
```

Description

`CVMdl = crossval(mdl)` returns a cross-validated (partitioned) support vector machine regression model, `CVMdl`, from a trained SVM regression model, `mdl`.

`CVMdl = crossval(mdl,Name,Value)` returns a cross-validated model with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Output Arguments

`CVMdl` — Cross-validated SVM regression model
`RegressionPartitionedSVM` model

Cross-validated SVM regression model, returned as a `RegressionPartitionedSVM` model.

Examples

Train Cross-Validated SVM Regression Model Using `crossval`

This example shows how to train a cross-validated SVM regression model using `crossval`.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current folder with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone and determine its age using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with a kernel scale equal to 2.2. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale',2.2,'Standardize',true);
```

`mdl` is a trained `RegressionSVM` regression model.

Cross validate the model using 10-fold cross validation.

```
CVMdl = crossval(mdl)
```

```
CVMdl =

  classreg.learning.partition.RegressionPartitionedSVM
      CrossValidatedModel: 'SVM'
           PredictorNames: {1x8 cell}
    CategoricalPredictors: 1
             ResponseName: 'Var9'
          NumObservations: 4177
                    KFold: 10
                Partition: [1x1 cvpartition]
        ResponseTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedSVM` cross-validated regression model. The software:

1. Randomly partitions the data into 10 equally sized sets.

2. Trains an SVM regression model on nine of the 10 sets.

3. Repeats steps 1 and 2 $k = 10$ times. It leaves out one of the partitions each time, and trains on the other nine partitions.

4. Combines generalization statistics for each fold.

Calculate the resubstitution loss for the cross-validated model.

```
loss = kfoldLoss(CVMdl)
```

```
loss =

    4.5712
```

Specify Cross-Validation Holdout Proportion for SVM Regression

This example shows how to specify a holdout proportion for training a cross-validated SVM regression model.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current folder with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone and determine its age using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with an automatic kernel scale. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale','auto','Standardize',true);
```

`mdl` is a trained `RegressionSVM` regression model.

Cross validate the regression model by specifying a 10% holdout sample.

```
CVMdl = crossval(mdl,'Holdout',0.1)
```

```
CVMdl =

  classreg.learning.partition.RegressionPartitionedSVM
      CrossValidatedModel: 'SVM'
           PredictorNames: {1x8 cell}
    CategoricalPredictors: 1
             ResponseName: 'Var9'
          NumObservations: 4177
                    KFold: 1
                Partition: [1x1 cvpartition]
        ResponseTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedSVM` model object.

Calculate the resubstitution loss for the cross-validated model.

```
loss = kfoldLoss(CVMdl)
```

```
loss =

    5.2499
```

Alternatives

Instead of training an SVM regression model and then cross-validating it, you can create a cross-validated model directly by using `fitrsvm` and specifying any of these name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'Leaveout'`, or `'KFold'`.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

See Also

`CompactRegressionSVM` | `RegressionPartitionedSVM` | `RegressionSVM` | `fitrsvm` | `kfoldLoss` | `kfoldPredict`

Introduced in R2015b

crossval

Class: `RegressionTree`

Cross-validated decision tree

Syntax

```
cvmodel = crossval(model)
cvmodel = crossval(model,Name,Value)
```

Description

`cvmodel = crossval(model)` creates a partitioned model from `model`, a fitted regression tree. By default, `crossval` uses 10-fold cross validation on the training data to create `cvmodel`.

`cvmodel = crossval(model,Name,Value)` creates a partitioned model with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`model`

A regression model, produced using `fitrtree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`CVPartition`

Object of class `cvpartition`, created by the `cvpartition` function. `crossval` splits the data into subsets with `cvpartition`.

Use only one of these four options at a time: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Default: `[]`

`Holdout`

Holdout validation tests the specified fraction of the data, and uses the rest of the data for training. Specify a numeric scalar from `0` to `1`. You can only use one of these four options at a time for creating a cross-validated tree: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

`KFold`

Number of folds to use in a cross-validated tree, a positive integer value greater than 1.

Use only one of these four options at a time: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Default: 10

`Leaveout`

Set to `'on'` for leave-one-out cross-validation.

Output Arguments

`cvmodel`

A partitioned model of class `RegressionPartitionedModel`.

Examples

Cross-Validate Regression Tree

Load the `carsmall` data set. Consider `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as predictor variables.

```
load carsmall
X = [Acceleration Displacement Horsepower Weight];
```

Grow a regression tree using the entire data set.

```
Mdl = fitrtree(X,MPG);
```

`Mdl` is a `RegressionTree` model.

Cross-validate the regression tree using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

`CVMdl` is a `RegressionPartitionedModel` cross-validated model. `crossval` stores the ten trained, compact regression trees in the `Trained` property of `CVMdl`.

Display the compact regression tree that `crossval` trained using all observations except those in the first fold.

```
CVMdl.Trained{1}
```

```
ans =
  CompactRegressionTree
             PredictorNames: {'x1'  'x2'  'x3'  'x4'}
               ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'


  Properties, Methods
```

Estimate the generalization error of `Mdl` by computing the 10-fold cross-validated mean-squared error.

```
L = kfoldLoss(CVMdl)
```

```
L = 23.5706
```

Tips

- Assess the predictive performance of `model` on cross-validated data using the “kfold” methods and properties of `cvmodel`, such as `kfoldLoss`.

Alternatives

You can create a cross-validation tree directly from the data, instead of creating a decision tree followed by a cross-validation tree. To do so, include one of these five options in `fitrtree`: `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

See Also

`crossval` | `fitrtree`

cvloss

Class: `ClassificationTree`

Classification error by cross validation

Syntax

```
E = cvloss(tree)
[E,SE] = cvloss(tree)
[E,SE,Nleaf] = cvloss(tree)
[E,SE,Nleaf,BestLevel] = cvloss(tree)
[___] = cvloss(tree,Name,Value)
```

Description

`E = cvloss(tree)` returns the cross-validated classification error (loss) for `tree`, a classification tree. The `cvloss` method uses stratified partitioning to create cross-validated sets. That is, for each fold, each partition of the data has roughly the same class proportions as in the data used to train `tree`.

`[E,SE] = cvloss(tree)` returns the standard error of `E`.

`[E,SE,Nleaf] = cvloss(tree)` returns the number of leaves of `tree`.

`[E,SE,Nleaf,BestLevel] = cvloss(tree)` returns the optimal pruning level for `tree`.

`[___] = cvloss(tree,Name,Value)` cross validates with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree` — Trained classification tree

`ClassificationTree` model object

Trained classification tree, specified as a `ClassificationTree` model object produced by `fitctree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level

0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `cvloss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`cvloss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as one of the following values:

- `'se'` — `cvloss` uses the smallest tree whose cost is within one standard error of the minimum cost.
- `'min'` — `cvloss` uses the minimal cost tree.

Example: `'TreeSize','min'`

`KFold` — Number of cross-validation samples
10 (default) | positive integer value greater than 1

Number of cross-validation samples, specified as a positive integer value greater than 1.

Example: `'KFold',8`

Output Arguments

`E` — Cross-validation classification error
numeric vector | scalar value

Cross-validation classification error (loss), returned as a vector or scalar depending on the setting of the `Subtrees` name-value pair.

`SE` — Standard error
numeric vector | scalar value

Standard error of `E`, returned as a vector or scalar depending on the setting of the `Subtrees` name-value pair.

`Nleaf` — Number of leaf nodes
numeric vector | scalar value

Number of leaf nodes in `tree`, returned as a vector or scalar depending on the setting of the `Subtrees` name-value pair. Leaf nodes are terminal nodes, which give classifications, not splits.

`BestLevel` — Best pruning level

scalar value

Best pruning level, returned as a scalar value. By default, a scalar representing the largest pruning level that achieves a value of `E` within `SE` of the minimum error. If you set `TreeSize` to `'min'`, `BestLevel` is the smallest value in `Subtrees`.

Examples

Compute the Cross-Validation Error

Compute the cross-validation error for a default classification tree.

Load the `ionosphere` data set.

```
load ionosphere
```

Grow a classification tree using the entire data set.

```
Mdl = fitctree(X,Y);
```

Compute the cross-validation error.

```
rng(1); % For reproducibility
E = cvloss(Mdl)
```

```
E = 0.1168
```

`E` is the 10-fold misclassification error.

Find the Best Pruning Level Using Cross Validation

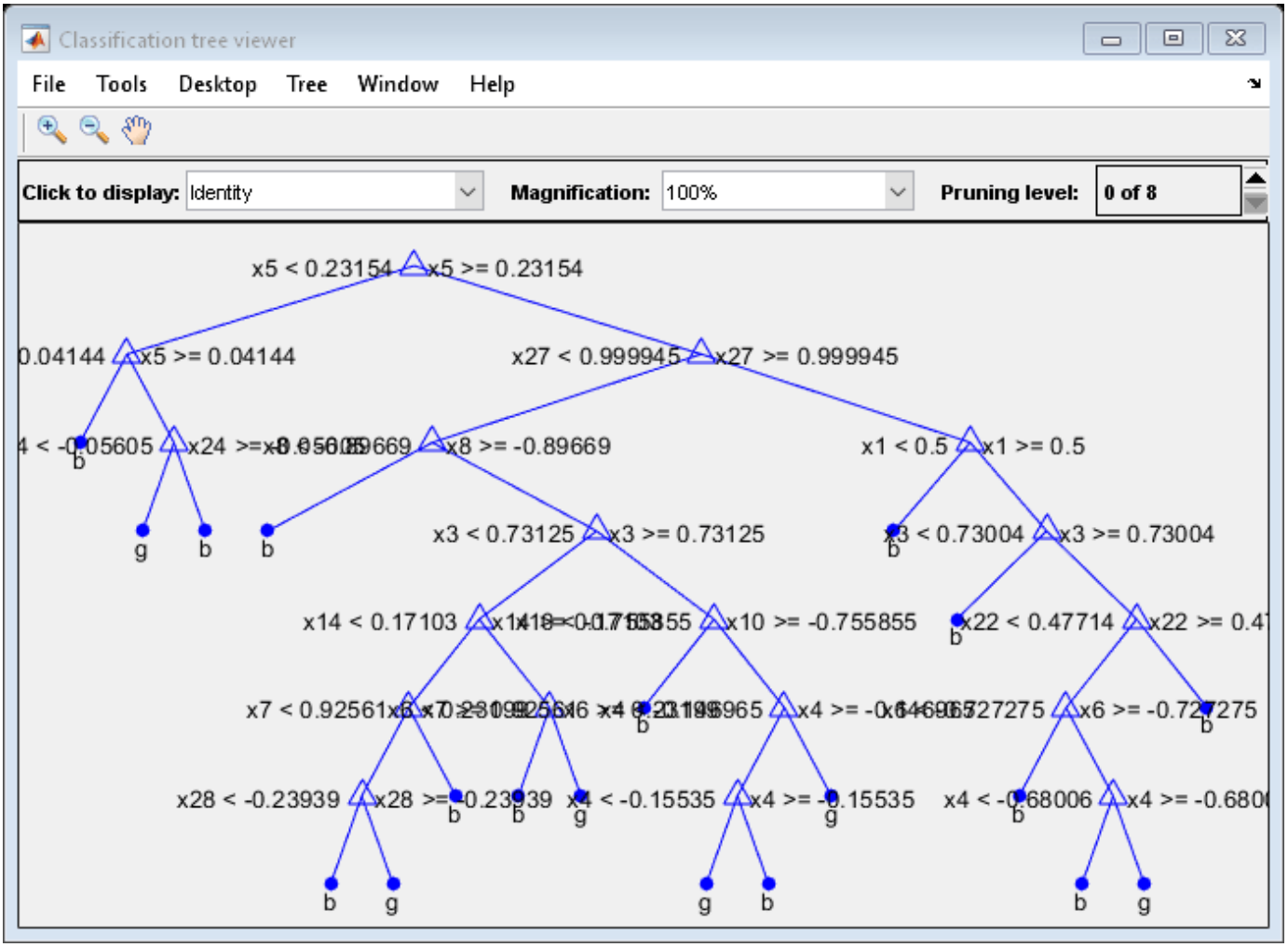
Apply *k*-fold cross validation to find the best level to prune a classification tree for all of its subtrees.

Load the `ionosphere` data set.

```
load ionosphere
```

Grow a classification tree using the entire data set. View the resulting tree.

```
Mdl = fitctree(X,Y);
view(Mdl,'Mode','graph')
```



Compute the 5-fold cross-validation error for each subtree except for the highest pruning level. Specify to return the best pruning level over all subtrees.

```
rng(1); % For reproducibility
m = max(Mdl.PruneList) - 1
```

```
m = 7
```

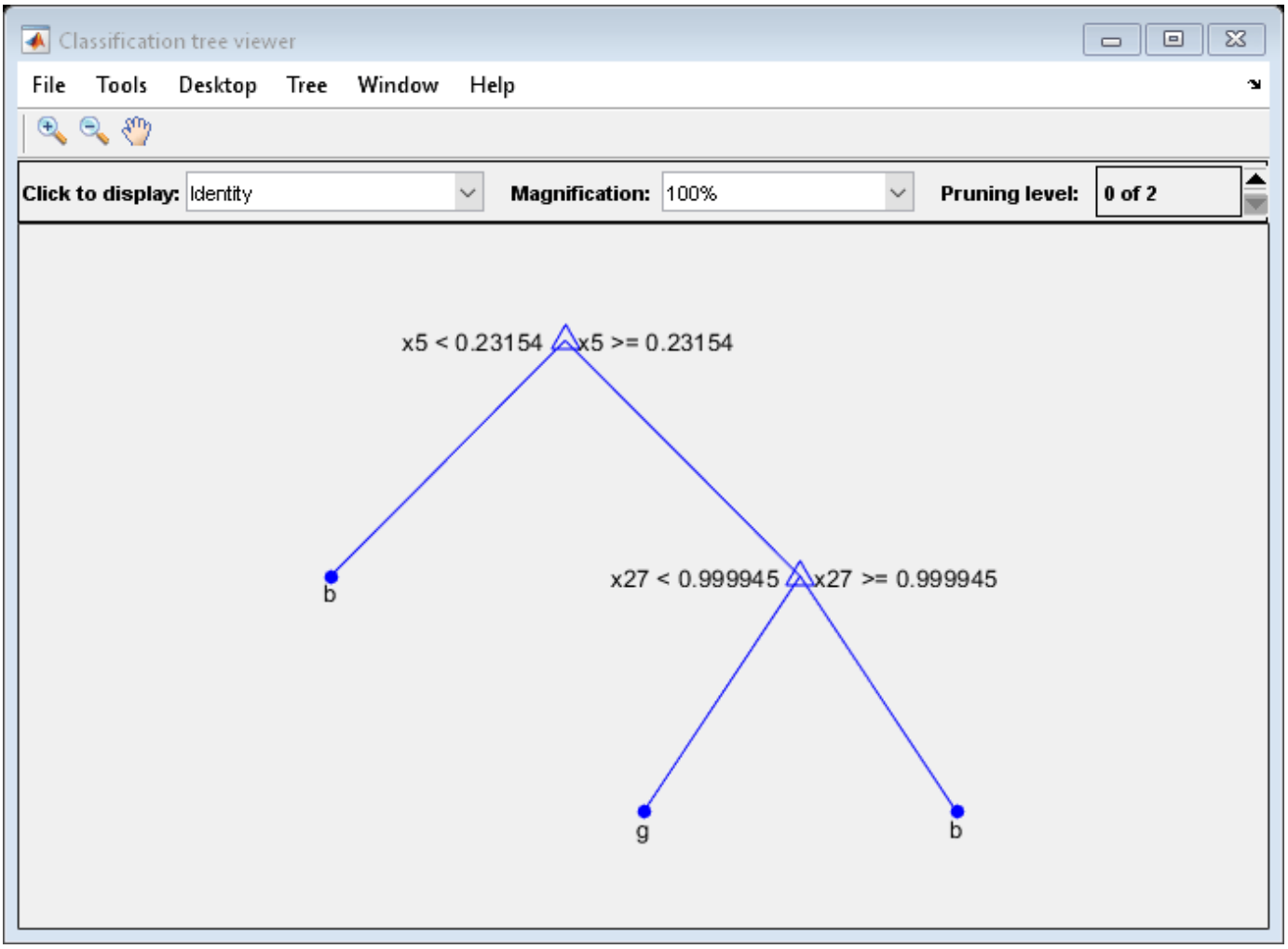
```
[E,~,~,bestLevel] = cvloss(Mdl,'SubTrees',0:m,'KFold',5)
```

```
E = 8×1

    0.1282
    0.1254
    0.1225
    0.1282
    0.1282
    0.1197
    0.0997
    0.1738


bestLevel = 6
```

Of the 7 pruning levels, the best pruning level is 6.

Prune the tree to the best level. View the resulting tree.

```
MdlPrune = prune(Mdl,'Level',bestLevel);
view(MdlPrune,'Mode','graph')
```



Alternatives

You can construct a cross-validated tree model with `crossval`, and call `kfoldLoss` instead of `cvloss`. If you are going to examine the cross-validated tree more than once, then the alternative can save time.

However, unlike `cvloss`, `kfoldLoss` does not return SE,`Nleaf`, or `BestLevel`. `kfoldLoss` also does not allow you to examine any error other than the classification error.

See Also

`crossval` | `fitctree` | `kfoldLoss` | `loss`

cvloss

Class: `RegressionTree`

Regression error by cross validation

Syntax

```
E = cvloss(tree)
[E,SE] = cvloss(tree)
[E,SE,Nleaf] = cvloss(tree)
[E,SE,Nleaf,BestLevel] = cvloss(tree)
[E,...] = cvloss(tree,Name,Value)
```

Description

`E = cvloss(tree)` returns the cross-validated regression error (loss) for `tree`, a regression tree.

`[E,SE] = cvloss(tree)` returns the standard error of `E`.

`[E,SE,Nleaf] = cvloss(tree)` returns the number of leaves (terminal nodes) in `tree`.

`[E,SE,Nleaf,BestLevel] = cvloss(tree)` returns the optimal pruning level for `tree`.

`[E,...] = cvloss(tree,Name,Value)` cross validates with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree` — Trained regression tree

`RegressionTree` object

Trained regression tree, specified as a `RegressionTree` object constructed using `fitrtree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level

`0` (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `cvloss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`cvloss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as the comma-separated pair consisting of `'TreeSize'` and one of the following:

- `'se'` — `cvloss` uses the smallest tree whose cost is within one standard error of the minimum cost.
- `'min'` — `cvloss` uses the minimal cost tree.

`KFold` — Number of folds
10 (default) | positive integer value greater than 1

Number of folds to use in a cross-validated tree, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1.

Example: `'KFold',8`

Output Arguments

`E` — Mean squared error
scalar value | numeric vector

Cross-validation mean squared error (loss), returned as a vector or scalar depending on the setting of the `Subtrees` name-value pair.

`SE` — Standard error
scalar value | numeric vector

Standard error of `E`, returned as vector or scalar depending on the setting of the `Subtrees` name-value pair.

`Nleaf` — Number of leaf nodes
vector of integer values

Number of leaf nodes in `tree`, returned as a vector or scalar depending on the setting of the `Subtrees` name-value pair. Leaf nodes are terminal nodes, which give responses, not splits.

`BestLevel` — Best pruning level
scalar value

Best pruning level as defined in the `TreeSize` name-value pair, returned as a scalar whose value depends on `TreeSize`:

- If `TreeSize` is `'se'`, then `BestLevel` is the largest pruning level that achieves a value of `E` within `SE` of the minimum error.
- If `TreeSize` is `'min'`, then `BestLevel` is the smallest value in `Subtrees`.

Examples

Compute the Cross-Validation Error

Compute the cross-validation error for a default regression tree.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using the entire data set.

```
Mdl = fitrtree(X,MPG);
```

Compute the cross-validation error.

```
rng(1); % For reproducibility
E = cvloss(Mdl)
```

```
E = 27.6976
```

`E` is the 10-fold weighted, average MSE (weighted by number of test observations in the folds).

Find the Best Pruning Level Using Cross Validation

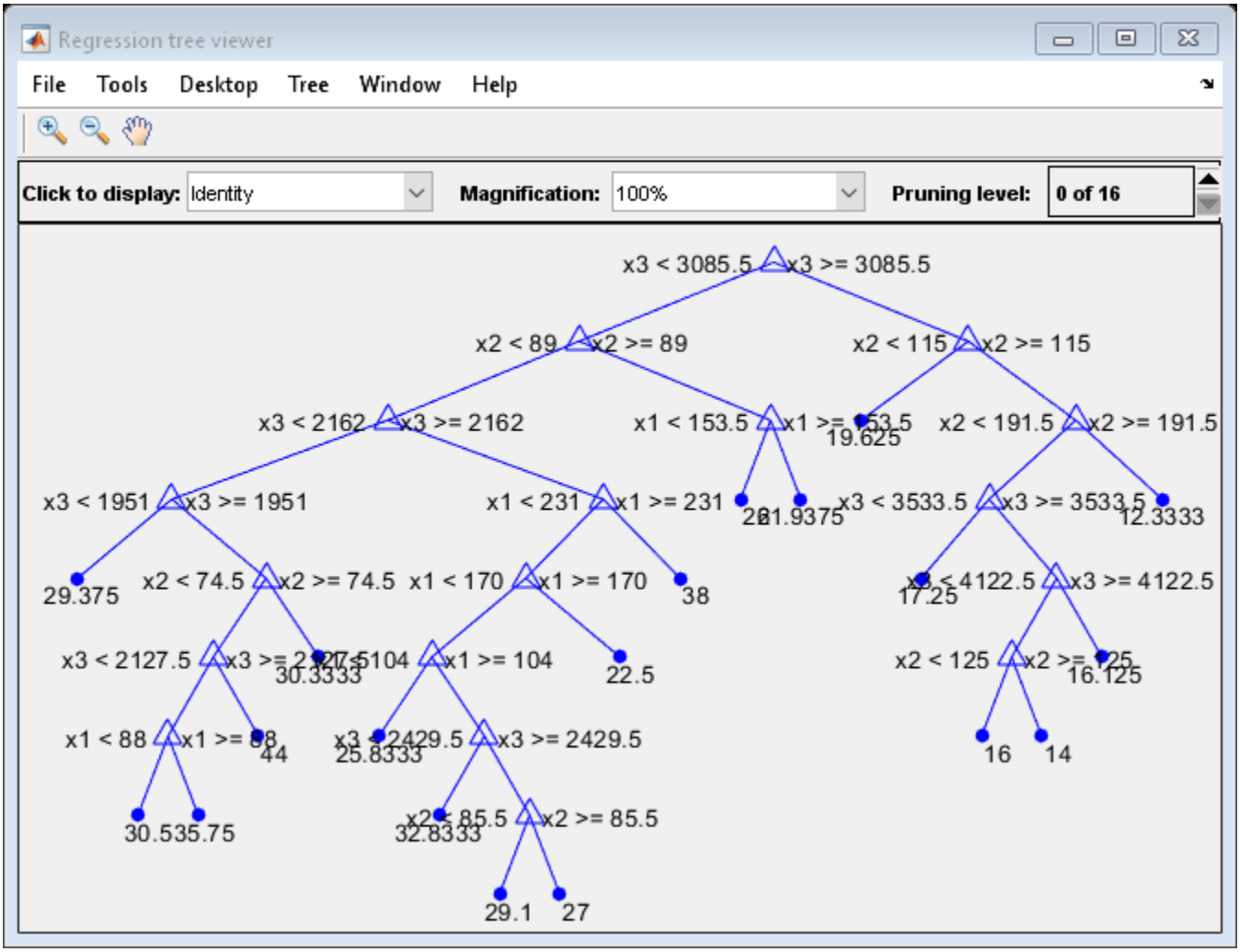
Apply *k*-fold cross validation to find the best level to prune a regression tree for all of its subtrees.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using the entire data set. View the resulting tree.

```
Mdl = fitrtree(X,MPG);
view(Mdl,'Mode','graph')
```



Compute the 5-fold cross-validation error for each subtree except for the first two lowest and highest pruning level. Specify to return the best pruning level over all subtrees.

```
rng(1); % For reproducibility
m = max(Mdl.PruneList) - 1
```

```
m = 15
```

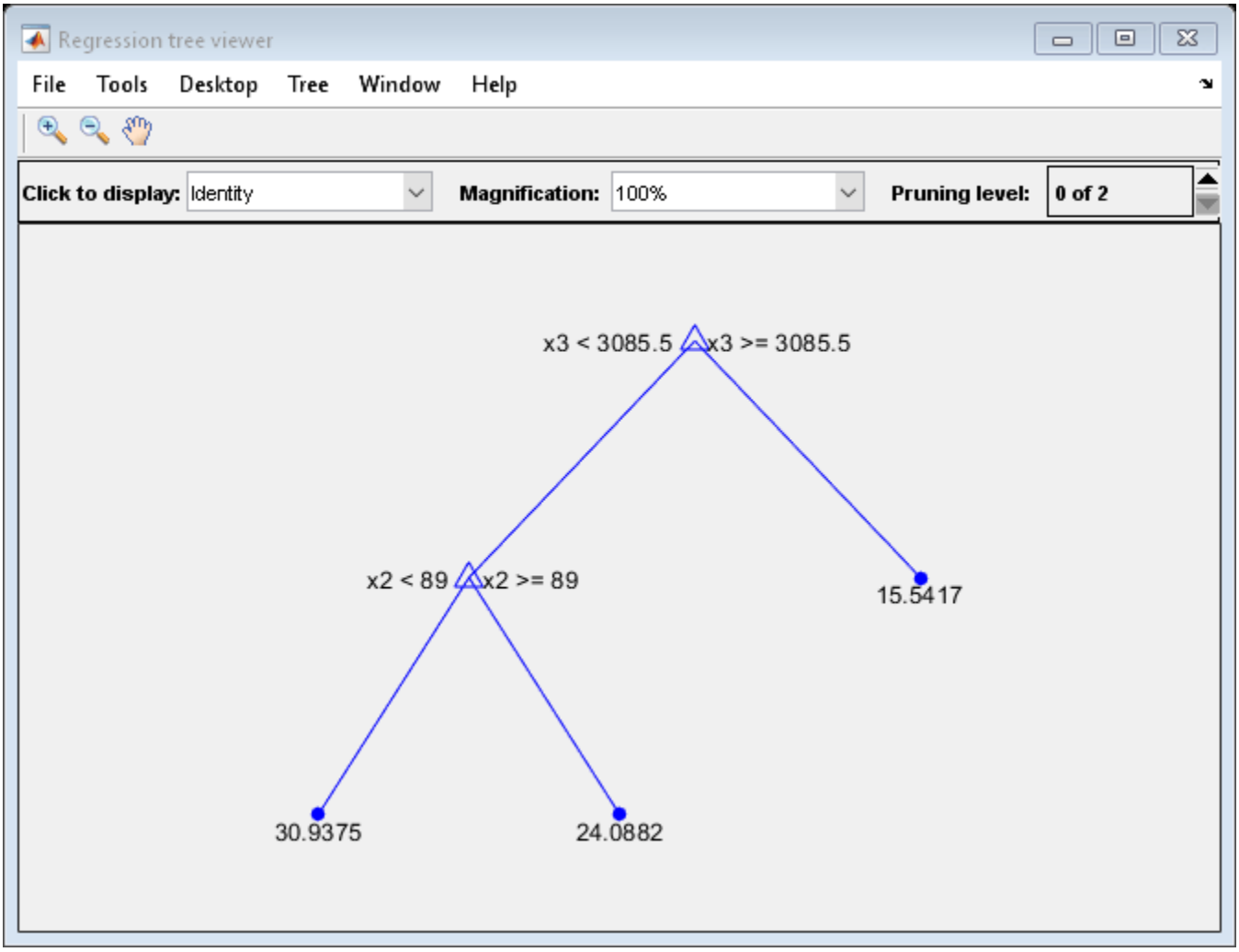
```
[~,~,~,bestLevel] = cvloss(Mdl,'SubTrees',2:m,'KFold',5)
```

```
bestLevel = 14
```

Of the 15 pruning levels, the best pruning level is 14.

Prune the tree to the best level. View the resulting tree.

```
MdlPrune = prune(Mdl,'Level',bestLevel);
view(MdlPrune,'Mode','graph')
```



Alternatives

You can construct a cross-validated tree model with `crossval`, and call `kfoldLoss` instead of `cvloss`. If you are going to examine the cross-validated tree more than once, then the alternative can save time.

However, unlike `cvloss`, `kfoldLoss` does not return `SE`, `Nleaf`, or `BestLevel`.

See Also

`crossval` | `fitrtree` | `kfoldLoss` | `loss`

cvpartition

Partition data for cross-validation

Description

`cvpartition` defines a random partition on a data set. Use this partition to define training and test sets for validating a statistical model using cross-validation. Use `training` to extract the training indices and `test` to extract the test indices for cross-validation. Use `repartition` to define a new random partition of the same type as a given `cvpartition` object.

Creation

Syntax

```
c = cvpartition(n,'KFold',k)
c = cvpartition(n,'Holdout',p)

c = cvpartition(group,'KFold',k)
c = cvpartition(group,'KFold',k,'Stratify',stratifyOption)
c = cvpartition(group,'Holdout',p)
c = cvpartition(group,'Holdout',p,'Stratify',stratifyOption)

c = cvpartition(n,'Leaveout')
c = cvpartition(n,'Resubstitution')
```

Description

`c = cvpartition(n,'KFold',k)` returns a `cvpartition` object `c` that defines a random nonstratified partition for `k`-fold cross-validation on `n` observations. The partition randomly divides the observations into `k` disjoint subsamples, or folds, each of which has approximately the same number of observations.

`c = cvpartition(n,'Holdout',p)` creates a random nonstratified partition for holdout validation on `n` observations. This partition divides the observations into a training set and a test, or holdout, set.

`c = cvpartition(group,'KFold',k)` creates a random partition for stratified `k`-fold cross-validation. Each subsample, or fold, has approximately the same number of observations and contains approximately the same class proportions as in `group`.

When you specify `group` as the first input argument, `cvpartition` discards rows of observations corresponding to missing values in `group`.

`c = cvpartition(group,'KFold',k,'Stratify',stratifyOption)` returns a `cvpartition` object `c` that defines a random partition for `k`-fold cross-validation. If you specify `'Stratify',false`, then `cvpartition` ignores the class information in `group` and creates a nonstratified random partition. Otherwise, the function implements stratification by default.

`c = cvpartition(group,'Holdout',p)` randomly partitions observations into a training set and a test, or holdout, set with stratification, using the class information in `group`. Both the training and test sets have approximately the same class proportions as in `group`.

`c = cvpartition(group,'Holdout',p,'Stratify',stratifyOption)` returns an object `c` that defines a random partition into a training set and a test, or holdout, set. If you specify `'Stratify',false`, then `cvpartition` creates a nonstratified random partition. Otherwise, the function implements stratification by default.

`c = cvpartition(n,'Leaveout')` creates a random partition for leave-one-out cross-validation on `n` observations. Leave-one-out is a special case of `'KFold'` in which the number of folds equals the number of observations.

`c = cvpartition(n,'Resubstitution')` creates an object `c` that does not partition the data. Both the training set and the test set contain all of the original `n` observations.

Input Arguments

n — Number of observations
positive integer scalar

Number of observations in the sample data, specified as a positive integer scalar.

Example: `100`

Data Types: `single` | `double`

k — Number of folds
`10` (default) | positive integer scalar

Number of folds in the partition, specified as a positive integer scalar. `k` must be smaller than the total number of observations.

Example: `5`

Data Types: `single` | `double`

p — Fraction or number of observations in test set
`0.1` (default) | scalar in the range (0,1) | integer scalar in the range [1,*n*)

Fraction or number of observations in the test set used for holdout validation, specified as a scalar in the range (0,1) or an integer scalar in the range [1,*n*), where *n* is the total number of observations.

- If `p` is a scalar in the range (0,1), then `cvpartition` randomly selects approximately `p`**n* observations for the test set.
- If `p` is an integer scalar in the range [1,*n*), then `cvpartition` randomly selects `p` observations for the test set.

Example: `0.2`

Example: `50`

Data Types: `single` | `double`

group — Grouping variable for stratification
numeric vector | logical vector | categorical array | character array | string array | cell array of character vectors

Grouping variable for stratification, specified as a numeric or logical vector, a categorical, character, or string array, or a cell array of character vectors indicating the class of each observation. `cvpartition` creates a partition from the observations in `group`.

Data Types: `single` | `double` | `logical` | `categorical` | `char` | `string` | `cell`

`stratifyOption` — Indicator for stratification
`true` | `false`

Indicator for stratification, specified as `true` or `false`.

- If the first input argument to `cvpartition` is `group`, then `cvpartition` implements stratification by default (`'Stratify',true`). For a nonstratified random partition, specify `'Stratify',false`.
- If the first input argument to `cvpartition` is `n`, then `cvpartition` always creates a nonstratified random partition (`'Stratify',false`). In this case, you cannot specify `'Stratify',true`.

Data Types: `logical`

Properties

`NumObservations` — Number of observations
positive integer scalar

This property is read-only.

Number of observations, including observations with missing `group` values, specified as a positive integer scalar.

Data Types: `double`

`NumTestSets` — Total number of test sets
number of folds | 1

This property is read-only.

Total number of test sets in the partition, specified as the number of folds when the partition type is `'kfold'` or `'leaveout'`, and 1 when the partition type is `'holdout'` or `'resubstitution'`.

Data Types: `double`

`TestSize` — Size of each test set
positive integer vector | positive integer scalar

This property is read-only.

Size of each test set, specified as a positive integer vector when the partition type is `'kfold'` or `'leaveout'`, and a positive integer scalar when the partition type is `'holdout'` or `'resubstitution'`.

Data Types: `double`

`TrainSize` — Size of each training set
positive integer vector | positive integer scalar

This property is read-only.

Size of each training set, specified as a positive integer vector when the partition type is `'kfold'` or `'leaveout'`, and a positive integer scalar when the partition type is `'holdout'` or `'resubstitution'`.

Data Types: `double`

Type — Type of validation partition
`'kfold'` | `'holdout'` | `'leaveout'` | `'resubstitution'`

This property is read-only.

Type of validation partition, specified as `'kfold'`, `'holdout'`, `'leaveout'`, or `'resubstitution'`.

Object Functions

repartition	Repartition data for cross-validation
test	Test indices for cross-validation
training	Training indices for cross-validation

Examples

Estimate Accuracy of Classifying New Data by Using Cross-Validation Error

Use the cross-validation misclassification error to estimate how a model will perform on new data.

Load the `ionosphere` data set. Create a table containing the predictor data `X` and the response variable `Y`.

```
load ionosphere
tbl = array2table(X);
tbl.Y = Y;
```

Use a random nonstratified partition `hpartition` to split the data into training data (`tblTrain`) and a reserved data set (`tblNew`). Reserve approximately 30 percent of the data.

```
rng('default') % For reproducibility
n = length(tbl.Y);
hpartition = cvpartition(n,'Holdout',0.3); % Nonstratified partition
idxTrain = training(hpartition);
tblTrain = tbl(idxTrain,:);
idxNew = test(hpartition);
tblNew = tbl(idxNew,:);
```

Train a support vector machine (SVM) classification model using the training data `tblTrain`. Calculate the misclassification error and the classification accuracy on the training data.

```
Mdl = fitcsvm(tblTrain,'Y');
trainError = resubLoss(Mdl)
```

```
trainError = 0.0569
```

```
trainAccuracy = 1-trainError
```

```
trainAccuracy = 0.9431
```

Typically, the misclassification error on the training data is not a good estimate of how a model will perform on new data because it can underestimate the misclassification rate on new data. A better estimate is the cross-validation error.

Create a partitioned model `cvMdl`. Compute the 10-fold cross-validation misclassification error and classification accuracy. By default, `crossval` ensures that the class proportions in each fold remain approximately the same as the class proportions in the response variable `tblTrain.Y`.

```
cvMdl = crossval(Mdl); % Performs stratified 10-fold cross-validation
cvtrainError = kfoldLoss(cvMdl)
```

```
cvtrainError = 0.1220
```

```
cvtrainAccuracy = 1-cvtrainError
```

```
cvtrainAccuracy = 0.8780
```

Notice that the cross-validation error `cvtrainError` is greater than the resubstitution error `trainError`.

Classify the new data in `tblNew` using the trained SVM model. Compare the classification accuracy on the new data to the accuracy estimates `trainAccuracy` and `cvtrainAccuracy`.

```
newError = loss(Mdl,tblNew,'Y');
newAccuracy = 1-newError
```

```
newAccuracy = 0.8700
```

The cross-validation error gives a better estimate of the model performance on new data than the resubstitution error.

Find Misclassification Rates Using K-Fold Cross-Validation

Use the same stratified partition for 5-fold cross-validation to compute the misclassification rates of two models.

Load the `fisheriris` data set. The matrix `meas` contains flower measurements for 150 different flowers. The variable `species` lists the species for each flower.

```
load fisheriris
```

Create a random partition for stratified 5-fold cross-validation. The training and test sets have approximately the same proportions of flower species as `species`.

```
rng('default') % For reproducibility
c = cvpartition(species,'KFold',5);
```

Create a partitioned discriminant analysis model and a partitioned classification tree model by using `c`.

```
discrCVModel = fitcdiscr(meas,species,'CVPartition',c);
treeCVModel = fitctree(meas,species,'CVPartition',c);
```

Compute the misclassification rates of the two partitioned models.

```
discrRate = kfoldLoss(discrCVModel)
```

```
discrRate = 0.0200
```

```
treeRate = kfoldLoss(treeCVModel)
```

```
treeRate = 0.0333
```

The discriminant analysis model has a smaller cross-validation misclassification rate.

Create Nonstratified Partition

Observe the test set (fold) class proportions in a 5-fold nonstratified partition of the `fisheriris` data. The class proportions differ across the folds.

Load the `fisheriris` data set. The `species` variable contains the species name (class) for each flower (observation). Convert `species` to a `categorical` variable.

```
load fisheriris
species = categorical(species);
```

Find the number of observations in each class. Notice that the three classes occur in equal proportion.

```
C = categories(species) % Class names
```

```
C = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
numClasses = size(C,1);
n = countcats(species) % Number of observations in each class
```

```
n = 3×1

    50
    50
    50
```

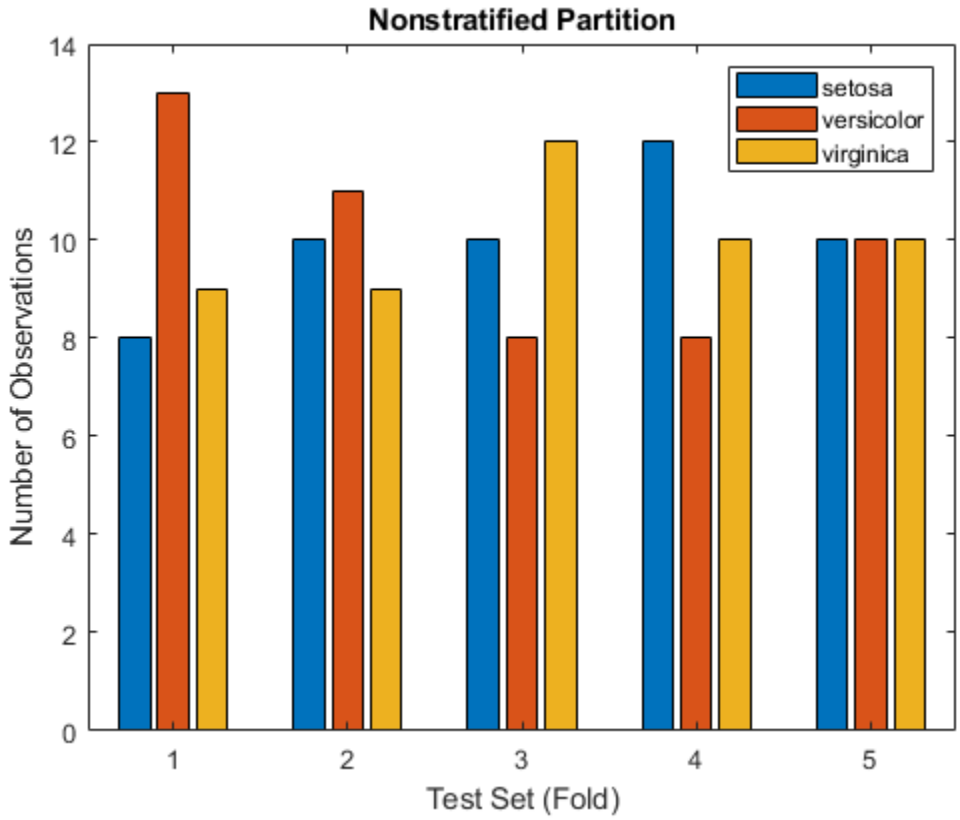
Create a random nonstratified 5-fold partition.

```
rng('default') % For reproducibility
cv = cvpartition(species,'KFold',5,'Stratify',false)
```

```
cv =
K-fold cross validation partition
   NumObservations: 150
       NumTestSets: 5
         TrainSize: 120  120  120  120  120
          TestSize: 30  30  30  30  30
```

Show that the three classes do not occur in equal proportion in each of the five test sets, or folds. Use a for-loop to update the `nTestData` matrix so that each entry `nTestData(i,j)` corresponds to the number of observations in test set `i` and class `C(j)`. Create a bar chart from the data in `nTestData`.

```
numFolds = cv.NumTestSets;
nTestData = zeros(numFolds,numClasses);
for i = 1:numFolds
    testClasses = species(cv.test(i));
    nCounts = countcats(testClasses); % Number of test set observations in each class
    nTestData(i,:) = nCounts';
end

bar(nTestData)
xlabel('Test Set (Fold)')
ylabel('Number of Observations')
title('Nonstratified Partition')
legend(C)
```



Notice that the class proportions vary in some of the test sets. For example, the first test set contains 8 setosa, 13 versicolor, and 9 virginica flowers, rather than 10 flowers per species. Because `cv` is a random nonstratified partition of the `fisheriris` data, the class proportions in each test set (fold) are not guaranteed to be equal to the class proportions in `species`. That is, the classes do not always occur equally in each test set, as they do in `species`.

Create Nonstratified and Stratified Holdout Partitions for Tall Array

Create a nonstratified holdout partition and a stratified holdout partition for a tall array. For the two holdout sets, compare the number of observations in each class.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (the default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a numeric vector of two classes, where class `1` and class `2` occur in the ratio `1:10`.

```
group = [ones(20,1);2*ones(200,1)]
```

```
group = 220×1

     1
     1
     1
     1
     1
     1
     1
     1
     1
     1
     ⋮
```

Create a tall array from `group`.

```
tgroup = tall(group)
```

```
tgroup =

  220x1 tall double column vector

     1
     1
     1
     1
     1
     1
     1
     1
     :
     :
```

`Holdout` is the only `cvpartition` option that is supported for tall arrays. Create a random nonstratified holdout partition.

```
CV0 = cvpartition(tgroup,'Holdout',1/4,'Stratify',false)
```

```
CV0 =
Hold-out cross validation partition
   NumObservations: [1x1 tall]
       NumTestSets: 1
         TrainSize: [1x1 tall]
          TestSize: [1x1 tall]
```

Return the result of `CV0.test` to memory by using the `gather` function.

```
testIdx0 = gather(CV0.test);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.78 sec
Evaluation completed in 1.1 sec
```

Find the number of times each class occurs in the test, or holdout, set.

```
accumarray(group(testIdx0),1) % Number of observations per class in the holdout set
```

```
ans = 2×1

     5
    51
```

`cvpartition` produces randomness in the results, so your number of observations in each class can vary from those shown.

Because `CV0` is a nonstratified partition, class `1` observations and class `2` observations in the holdout set are not guaranteed to occur in the same ratio as in `tgroup`. However, because of the inherent randomness in `cvpartition`, you can sometimes obtain a holdout set in which the classes occur in the same ratio as in `tgroup`, even though you specify `'Stratify',false`. Because the training set is the complement of the holdout set, excluding any `NaN` or missing observations, you can obtain a similar result for the training set.

Return the result of `CV0.training` to memory.

```
trainIdx0 = gather(CV0.training);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.3 sec
Evaluation completed in 0.38 sec
```

Find the number of times each class occurs in the training set.

```
accumarray(group(trainIdx0),1) % Number of observations per class in the training set
```

```
ans = 2×1

    15
   149
```

The classes in the nonstratified training set are not guaranteed to occur in the same ratio as in `tgroup`.

Create a random stratified holdout partition.

```
CV1 = cvpartition(tgroup,'Holdout',1/4)
```

```
CV1 =
Hold-out cross validation partition
   NumObservations: [1x1 tall]
       NumTestSets: 1
         TrainSize: [1x1 tall]
          TestSize: [1x1 tall]
```

Return the result of `CV1.test` to memory.

```
testIdx1 = gather(CV1.test);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 0.16 sec
Evaluation completed in 0.21 sec
```

Find the number of times each class occurs in the test, or holdout, set.

```
accumarray(group(testIdx1),1) % Number of observations per class in the holdout set
```

```
ans = 2×1

     5
    51
```

In the case of the stratified holdout partition, the class ratio in the holdout set and the class ratio in `tgroup` are the same `(1:10)`.

Find Influential Observations Using Leave-One-Out Partition

Create a random partition of data for leave-one-out cross-validation. Compute and compare training set means. A repetition with a significantly different mean suggests the presence of an influential observation.

Create a data set `X` that contains one value that is much greater than the others.

```
X = [1 2 3 4 5 6 7 8 9 20]';
```

Create a `cvpartition` object that has 10 observations and 10 repetitions of training and test data. For each repetition, `cvpartition` selects one observation to remove from the training set and reserve for the test set.

```
c = cvpartition(10,'Leaveout')
```

```
c =
Leave-one-out cross validation partition
   NumObservations: 10
       NumTestSets: 10
         TrainSize: 9  9  9  9  9  9  9  9  9  9
          TestSize: 1  1  1  1  1  1  1  1  1  1
```

Apply the leave-one-out partition to `X`, and take the mean of the training observations for each repetition by using `crossval`.

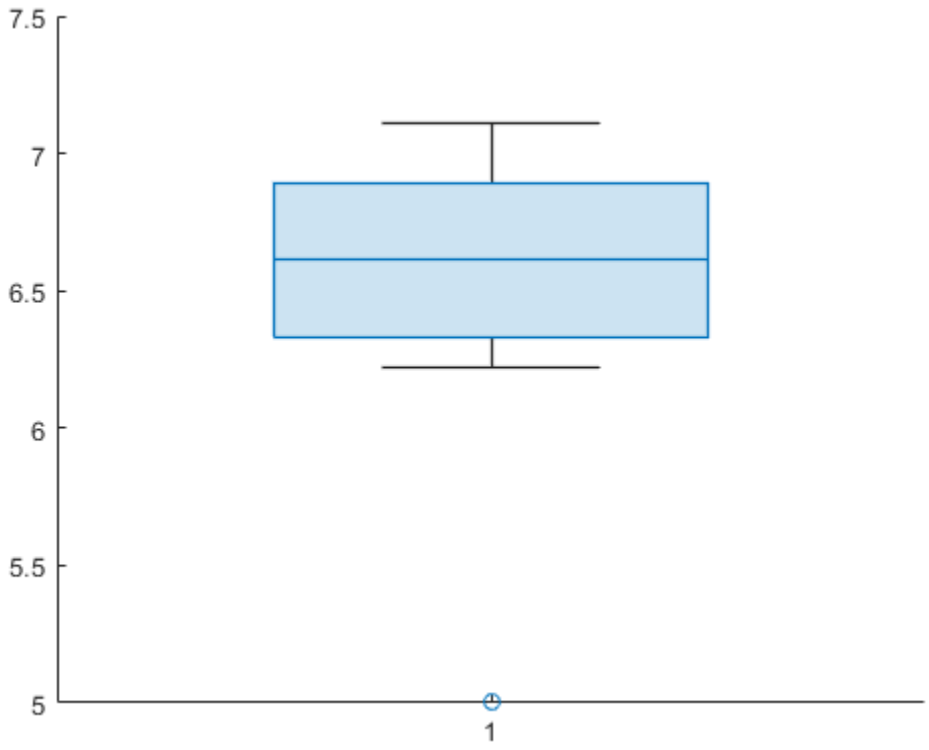
```
values = crossval(@(Xtrain,Xtest)mean(Xtrain),X,'Partition',c)
```

```
values = 10×1

    6.5556
    6.4444
    7.0000
    6.3333
    6.6667
    7.1111
    6.8889
```

```
    6.7778
    6.2222
    5.0000
```

View the distribution of the training set means using a box chart (or box plot). The plot displays one outlier.

```
boxchart(values)
```



Find the repetition corresponding to the outlier value. For that repetition, find the observation in the test set.

```
[~,repetitionIdx] = min(values)
```

```
repetitionIdx = 10
```

```
observationIdx = test(c,repetitionIdx);
influentialObservation = X(observationIdx)
```

```
influentialObservation = 20
```

Training sets that contain the observation have substantially different means from the mean of the training set without the observation. This significant change in mean suggests that the value of `20` in `X` is an influential observation.

Tips

- If you specify `group` as the first input argument to `cvpartition`, then the function discards rows of observations corresponding to missing values in `group`.
- If you specify `group` as the first input argument to `cvpartition`, then the function implements stratification by default. You can specify `'Stratify',false` to create a nonstratified random partition.
- You can specify `'Stratify',true` only when the first input argument to `cvpartition` is `group`.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

The `cvpartition` function supports tall arrays for out-of-memory data with some limitations.

- When you use `cvpartition` with tall arrays, the first input argument must be a grouping variable, `tGroup`. If you specify a tall scalar as the first input argument, `cvpartition` gives an error.
- `cvpartition` supports only `Holdout` cross-validation for tall arrays; for example, `c = cvpartition(tGroup,'Holdout',p)`. By default, `cvpartition` randomly partitions observations into a training set and a test set with stratification, using the class information in `tGroup`. The parameter `p` is a scalar such that `0 < p < 1`.
- To create nonstratified `Holdout` partitions, specify the value of the `'Stratify'` name-value pair argument as `false`; for example, `c = cvpartition(tGroup,'Holdout',p,'Stratify',false)`.

For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`crossval` | `repartition` | `test` | `training`

Topics

“Grouping Variables” on page 2-45

Introduced in R2008a

cvshrink

Class: `ClassificationDiscriminant`

Cross-validate regularization of linear discriminant

Syntax

```
err = cvshrink(obj)
[err,gamma] = cvshrink(obj)
[err,gamma,delta] = cvshrink(obj)
[err,gamma,delta,numpred] = cvshrink(obj)
[err,...] = cvshrink(obj,Name,Value)
```

Description

`err = cvshrink(obj)` returns a vector of cross-validated classification error values for differing values of the regularization parameter Gamma.

`[err,gamma] = cvshrink(obj)` also returns the vector of Gamma values.

`[err,gamma,delta] = cvshrink(obj)` also returns the vector of Delta values.

`[err,gamma,delta,numpred] = cvshrink(obj)` returns the vector of number of nonzero predictors for each setting of the parameters Gamma and Delta.

`[err,...] = cvshrink(obj,Name,Value)` cross validates with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier, produced using `fitcdiscr`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`delta`

- Scalar `delta` — `cvshrink` uses this value of `delta` with every value of `gamma` for regularization.
- Row vector `delta` — For each `i` and `j`, `cvshrink` uses `delta(j)` with `gamma(i)` for regularization.
- Matrix `delta` — The number of rows of `delta` must equal the number of elements in `gamma`. For each `i` and `j`, `cvshrink` uses `delta(i,j)` with `gamma(i)` for regularization.

Default: `0`

gamma

Vector of Gamma values for cross-validation.

Default: `0:0.1:1`

NumDelta

Number of Delta intervals for cross-validation. For every value of Gamma, `cvshrink` cross-validates the discriminant using `NumDelta + 1` values of Delta, uniformly spaced from zero to the maximal Delta at which all predictors are eliminated for this value of Gamma. If you set `delta`, `cvshrink` ignores `NumDelta`.

Default: `0`

NumGamma

Number of Gamma intervals for cross-validation. `cvshrink` cross-validates the discriminant using `NumGamma + 1` values of Gamma, uniformly spaced from `MinGamma` to `1`. If you set `gamma`, `cvshrink` ignores `NumGamma`.

Default: `10`

verbose

Verbosity level, an integer from `0` to `2`. Higher values give more progress messages.

Default: `0`

Output Arguments

err

Numeric vector or matrix of errors. `err` is the misclassification error rate, meaning the average fraction of misclassified data over all folds.

- If `delta` is a scalar (default), `err(i)` is the misclassification error rate for `obj` regularized with `gamma(i)`.
- If `delta` is a vector, `err(i,j)` is the misclassification error rate for `obj` regularized with `gamma(i)` and `delta(j)`.
- If `delta` is a matrix, `err(i,j)` is the misclassification error rate for `obj` regularized with `gamma(i)` and `delta(i,j)`.

gamma

Vector of Gamma values used for regularization. See “Gamma and Delta” on page 33-1124.

delta

Vector or matrix of Delta values used for regularization. See “Gamma and Delta” on page 33-1124.

- If you give a scalar for the `delta` name-value pair, the output `delta` is a row vector the same size as `gamma`, with entries equal to the input scalar.

- If you give a row vector for the `delta` name-value pair, the output `delta` is a matrix with the same number of columns as the row vector, and with the number of rows equal to the number of elements of `gamma`. The output `delta(i,j)` is equal to the input `delta(j)`.
- If you give a matrix for the `delta` name-value pair, the output `delta` is the same as the input matrix. The number of rows of `delta` must equal the number of elements in `gamma`.

`numpred`

Numeric vector or matrix containing the number of predictors in the model at various regularizations. `numpred` has the same size as `err`.

- If `delta` is a scalar (default), `numpred(i)` is the number of predictors for `obj` regularized with `gamma(i)` and `delta`.
- If `delta` is a vector, `numpred(i,j)` is the number of predictors for `obj` regularized with `gamma(i)` and `delta(j)`.
- If `delta` is a matrix, `numpred(i,j)` is the number of predictors for `obj` regularized with `gamma(i)` and `delta(i,j)`.

Examples

Regularize Data with Many Predictors

Regularize a discriminant analysis classifier, and view the tradeoff between the number of predictors in the model and the classification accuracy.

Create a linear discriminant analysis classifier for the `ovariancancer` data. Set the `SaveMemory` and `FillCoeffs` options to keep the resulting model reasonably small.

```
load ovariancancer
obj = fitcdiscr(obs,grp,...
    'SaveMemory','on','FillCoeffs','off');
```

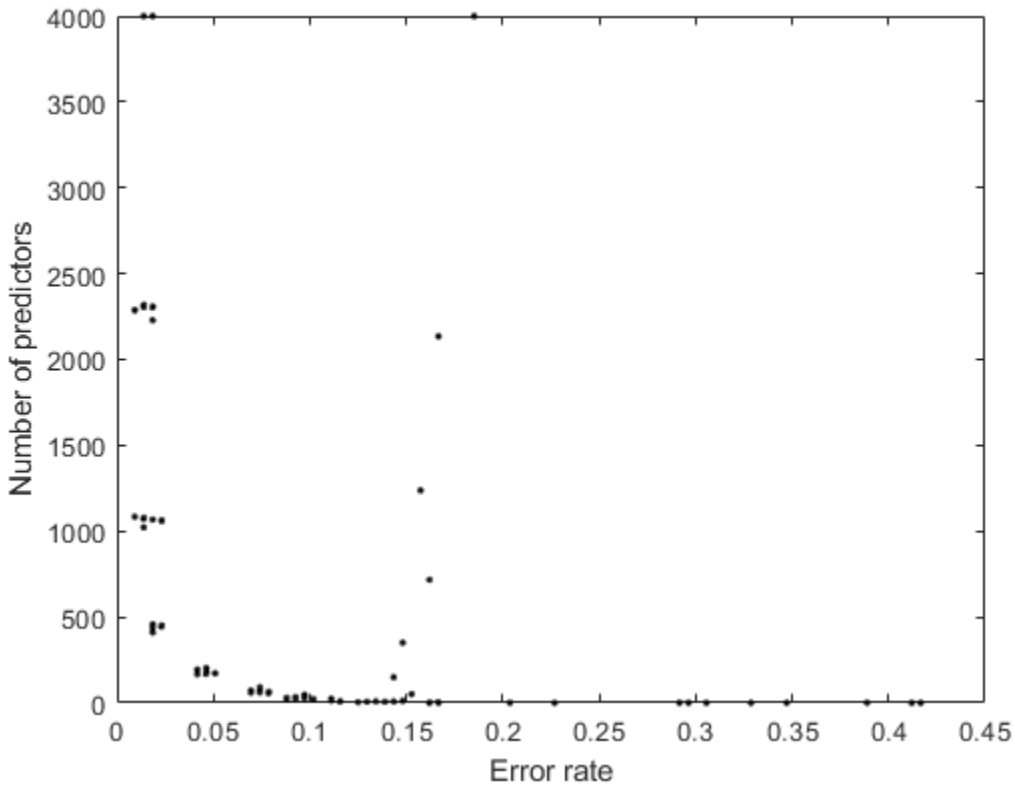
Use 10 levels of `Gamma` and 10 levels of `Delta` to search for good parameters. This search is time-consuming. Set `Verbose` to `1` to view the progress.

```
rng('default') % for reproducibility
[err,gamma,delta,numpred] = cvshrink(obj,...
    'NumGamma',9,'NumDelta',9,'Verbose',1);
```

```
Done building cross-validated model.
Processing Gamma step 1 out of 10.
Processing Gamma step 2 out of 10.
Processing Gamma step 3 out of 10.
Processing Gamma step 4 out of 10.
Processing Gamma step 5 out of 10.
Processing Gamma step 6 out of 10.
Processing Gamma step 7 out of 10.
Processing Gamma step 8 out of 10.
Processing Gamma step 9 out of 10.
Processing Gamma step 10 out of 10.
```

Plot the classification error rate against the number of predictors.

```
plot(err,numpred,'k.')
xlabel('Error rate');
ylabel('Number of predictors');
```



More About

Gamma and Delta

Regularization is the process of finding a small set of predictors that yield an effective predictive model. For linear discriminant analysis, there are two parameters, γ and δ, that control regularization as follows. `cvshrink` helps you select appropriate values of the parameters.

Let Σ represent the covariance matrix of the data X, and let $\widehat{X}$ be the centered data (the data X minus the mean by class). Define

$$D = \mathrm{diag}\left(\widehat{X}^T * \widehat{X}\right).$$

The regularized covariance matrix $\tilde{\Sigma}$ is

$$\tilde{\Sigma} = (1 - \gamma)\Sigma + \gamma D.$$

Whenever $\gamma \geq$ `MinGamma`, $\tilde{\Sigma}$ is nonsingular.

Let μ_k be the mean vector for those elements of X in class k, and let μ_0 be the global mean vector (the mean of the rows of X). Let C be the correlation matrix of the data X, and let $\tilde{C}$ be the regularized correlation matrix:

$$\tilde{C} = (1 - \gamma)C + \gamma I,$$

where I is the identity matrix.

The linear term in the regularized discriminant analysis classifier for a data point x is

$$(x - \mu_0)^T \tilde{\Sigma}^{-1}(\mu_k - \mu_0) = \left[(x - \mu_0)^T D^{-1/2}\right]\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right].$$

The parameter δ enters into this equation as a threshold on the final term in square brackets. Each component of the vector $\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right]$ is set to zero if it is smaller in magnitude than the threshold δ. Therefore, for class k, if component j is thresholded to zero, component j of x does not enter into the evaluation of the posterior probability.

The `DeltaPredictor` property is a vector related to this threshold. When $\delta \geq$ `DeltaPredictor(i)`, all classes k have

$$\left|\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right| \leq \delta\,.$$

Therefore, when $\delta \geq$ `DeltaPredictor(i)`, the regularized classifier does not use predictor `i`.

Tips

- Examine the `err` and `numpred` outputs to see the tradeoff between cross-validated error and number of predictors. When you find a satisfactory point, set the corresponding `gamma` and `delta` properties in the model using dot notation. For example, if `(i,j)` is the location of the satisfactory point, set

  ```
  obj.Gamma = gamma(i);
  obj.Delta = delta(i,j);
  ```

See Also

`ClassificationDiscriminant` | `fitcdiscr`

Topics

“Regularize Discriminant Analysis Classifier” on page 20-21
“Discriminant Analysis Classification” on page 20-2

cvshrink

Cross validate shrinking (pruning) ensemble

Syntax

```
vals = cvshrink(ens)
[vals,nlearn] = cvshrink(ens)
[vals,nlearn] = cvshrink(ens,Name,Value)
```

Description

`vals = cvshrink(ens)` returns an L-by-T matrix with cross-validated values of the mean squared error. L is the number of `lambda` values in the `ens.Regularization` structure. T is the number of `threshold` values on weak learner weights. If `ens` does not have a `Regularization` property filled in by the `regularize` method, pass a `lambda` name-value pair.

`[vals,nlearn] = cvshrink(ens)` returns an L-by-T matrix of the mean number of learners in the cross-validated ensemble.

`[vals,nlearn] = cvshrink(ens,Name,Value)` cross validates with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression ensemble, created with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`cvpartition`

A partition created with `cvpartition` to use in a cross-validated tree. You can only use one of these four options at a time: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`holdout`

Holdout validation tests the specified fraction of the data, and uses the rest of the data for training. Specify a numeric scalar from `0` to `1`. You can only use one of these four options at a time for creating a cross-validated tree: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`kfold`

Number of folds to use in a cross-validated tree, a positive integer. If you do not supply a cross-validation method, `cvshrink` uses 10-fold cross validation. You can only use one of these four options at a time: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

Default: `10`

`lambda`

Vector of nonnegative regularization parameter values for lasso. If empty, `cvshrink` does not perform cross validation.

Default: `[]`

`leaveout`

Use leave-one-out cross validation by setting to `'on'`. You can only use one of these four options at a time: `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.

`threshold`

Numeric vector with lower cutoffs on weights for weak learners. `cvshrink` discards learners with weights below `threshold` in its cross-validation calculation.

Default: `0`

Output Arguments

`vals`

L-by-T matrix with cross-validated values of the mean squared error. L is the number of values of the regularization parameter `'lambda'`, and T is the number of `'threshold'` values on weak learner weights.

`nlearn`

L-by-T matrix with cross-validated values of the mean number of learners in the cross-validated ensemble.L is the number of values of the regularization parameter `'lambda'`, and T is the number of `'threshold'` values on weak learner weights.

Examples

Cross-Validate Regression Ensemble

Create a regression ensemble for predicting mileage from the `carsmall` data. Cross-validate the ensemble.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

You can train an ensemble of bagged regression trees.

```
ens = fitrensemble(X,Y,'Method','Bag')
```

`fircensemble` uses a default template tree object `templateTree()` as a weak learner when `'Method'` is `'Bag'`. In this example, for reproducibility, specify `'Reproducible',true` when you create a tree template object, and then use the object as a weak learner.

```
rng('default') % For reproducibility
t = templateTree('Reproducible',true); % For reproducibiliy of random predictor selections
ens = fitrensemble(X,MPG,'Method','Bag','Learners',t);
```

Specify values for `lambda` and `threshold`. Use these values to cross-validate the ensemble.

```
[vals,nlearn] = cvshrink(ens,'lambda',[.01 .1 1],'threshold',[0 .01 .1])
```

```
vals = 3×3

   18.9150   19.0092  128.5935
   18.9099   18.9504  128.8449
   19.0328   18.9636  116.8500


nlearn = 3×3

   13.7000   11.6000    4.1000
   13.7000   11.7000    4.1000
   13.9000   11.6000    4.1000
```

Clearly, setting a threshold of `0.1` leads to unacceptable errors, while a threshold of `0.01` gives similar errors to a threshold of `0`. The mean number of learners with a threshold of `0.01` is about `11.4`, whereas the mean number is about `13.8` when the threshold is `0`.

See Also

`regularize` | `shrink`

datasample

Randomly sample from data, with or without replacement

Syntax

```
y = datasample(data,k)
y = datasample(data,k,dim)
y = datasample( ___ ,Name,Value)
y = datasample(s, ___ )
[y,idx] = datasample( ___ )
```

Description

`y = datasample(data,k)` returns *k* observations sampled uniformly at random, with replacement, from the data in `data`.

`y = datasample(data,k,dim)` returns a sample taken along dimension `dim` of `data`.

`y = datasample( ___ ,Name,Value)` returns a sample for any of the input arguments in the previous syntaxes, with additional options specified by one or more name-value pair arguments. For example, `'Replace',false` specifies sampling without replacement.

`y = datasample(s, ___ )` uses the random number stream `s` to generate random numbers. The option `s` can precede any of the input arguments in the previous syntaxes.

`[y,idx] = datasample( ___ )` also returns an index vector indicating which values `datasample` sampled from `data` using any of the input arguments in the previous syntaxes.

Examples

Sample Unique Values from Vector

Create the random number stream for reproducibility.

```
s = RandStream('mlfg6331_64');
```

Draw five unique values from the integers `1` to `10`.

```
y = datasample(s,1:10,5,'Replace',false)
```

```
y = 1×5

     9     8     3     6     2
```

Generate Random Characters for Specified Probabilities

Create the random number stream for reproducibility.

```
s = RandStream('mlfg6331_64');
```

Generate 48 random characters from the sequence ACGT per specified probabilities.

```
seq = datasample(s,'ACGT',48,'Weights',[0.15 0.35 0.35 0.15])
```

```
seq =
'GGCGGCGCAAGGCGCCGGACCTGGCTGCACGCCGTTCCCTGCTACTCG'
```

Select Random Subset of Matrix Columns

Set the random seed for reproducibility of the results.

```
rng(10,'twister')
```

Generate a matrix with 10 rows and 1000 columns.

```
X = randn(10,1000);
```

Create the random number stream for reproducibility within `datasample`.

```
s = RandStream('mlfg6331_64');
```

Randomly select five unique columns from X.

```
Y = datasample(s,X,5,2,'Replace',false)
```

```
Y = 10×5

    0.4317   -0.3327    0.9112   -2.3244    0.9559
    0.6977   -0.7422    0.4578   -1.3745   -0.8634
   -0.8543   -0.3105    0.9836   -0.6434   -0.4457
    0.1686    0.6609   -0.0553   -0.1202   -1.3699
   -1.7649   -1.1607   -0.3513   -1.5533    0.0597
   -0.3821    0.5696   -1.6264   -0.2104   -1.5486
   -1.6844    0.7148   -0.6876   -0.4447   -1.4615
   -0.4170    1.3696    1.1874   -0.9901    0.5875
   -0.2410    1.4703   -2.5003   -1.1321   -1.8451
    0.6212    1.4118   -0.4518    0.8697    0.8093
```

Create a Bootstrap Replicate Data Set

Resample observations from a dataset array to create a bootstrap replicate data set. See “Bootstrap Resampling” on page 3-14 for more information about bootstrapping.

Load the sample data set.

```
load hospital
```

Create a data set that has the same size as the `hospital` data set and contains random samples chosen with replacement from the `hospital` data set.

```
y = datasample(hospital,size(hospital,1));
```

Sample in Parallel from Two Data Vectors

Select samples from data based on indices of a sample chosen from another vector.

Generate two random vectors.

```
x1 = randn(100,1);
x2 = randn(100,1);
```

Select a sample of `10` elements from vector `x1`, and return the indices of the sample in vector `idx`.

```
[y1,idx] = datasample(x1,10);
```

Select a sample of `10` elements from vector `x2` using the indices in vector `idx`.

```
y2 = x2(idx);
```

Input Arguments

`data` — Input data

vector | matrix | multidimensional array | table | dataset array

Input data from which to sample, specified as a vector, matrix, multidimensional array, table, or dataset array. By default, `datasample` samples from the first nonsingleton dimension of `data`. For example, if `data` is a matrix, then `datasample` samples from the rows. Change this behavior with the `dim` input argument.

Data Types: `single` | `double` | `logical` | `char` | `string` | `table`

`k` — Number of samples

positive integer

Number of samples, specified as a positive integer.

Example: `datasample(data,100)` returns 100 observations sampled uniformly and at random from the data in `data`.

Data Types: `single` | `double`

`dim` — Dimension to sample

1 (default) | positive integer

Dimension to sample, specified as a positive integer. For example, if `data` is a matrix and `dim` is `2`, `y` contains a selection of columns in `data`. If `data` is a table or dataset array and `dim` is `2`, `y` contains a selection of variables in `data`. Use `dim` to ensure sampling along a specific dimension regardless of whether `data` is a vector, matrix, or *N*-dimensional array.

Data Types: `single` | `double`

`s` — Random number stream

global stream (default) | `RandStream`

Random number stream, specified as the global stream or `RandStream`. For example, `s = RandStream('mlfg6331_64')` creates a random number stream that uses the multiplicative lagged

Fibonacci generator algorithm. For details, see “Creating and Controlling a Random Number Stream”.

The `rng` function provides a simple way to control the global stream. For example, `rng(seed)` seeds the random number generator using the nonnegative integer seed. For details, see “Managing the Global Stream Using RandStream”.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Replace',false,'Weights',ones(datasize,1)` samples without replacement and with probability proportional to the elements of `Weights`, where `datasize` is the size of the dimension being sampled.

`Replace` — Indicator for sampling with replacement
true (default) | false

Indicator for sampling with replacement, specified as the comma-separated pair consisting of `'Replace'` and either `true` or `false`.

Sample with replacement if `'Replace'` is `true`, or without replacement if `'Replace'` is `false`. If `'Replace'` is `false`, then `k` must not be larger than the size of the dimension being sampled. For example, if `data = [1 3 Inf; 2 4 5]` and `y = datasample(data,k,'Replace',false)`, then `k` cannot be larger than 2.

Data Types: `logical`

`Weights` — Sampling weights
`ones(datasize,1)` (default) | vector of nonnegative numeric values

Sampling weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of nonnegative numeric values. The vector is of size `datasize`, where `datasize` is the size of the dimension being sampled. The vector must have at least one positive value and cannot contain `NaN` values. The `datasample` function samples with probability proportional to the elements of `'Weights'`.

Example: `'Weights',[0.1 0.5 0.35 0.46]`

Data Types: `single` | `double`

Output Arguments

`y` — Sample
vector | matrix | multidimensional array | table | dataset array

Sample, returned as a vector, matrix, multidimensional array, table, or dataset array.

- If `data` is a vector, then `y` is a vector containing `k` elements selected from `data`.
- If `data` is a matrix and `dim = 1`, then `y` is a matrix containing `k` rows selected from `data`. Or, if `dim = 2`, then `y` is a matrix containing `k` columns selected from `data`.

- If `data` is an *N*-dimensional array and `dim = 1`, then `y` is an *N*-dimensional array of samples taken along the first nonsingleton dimension of `data`. Or, if you specify a value for the `dim` name-value pair argument, `datasample` samples along the dimension `dim`.
- If `data` is a table and `dim = 1`, then `y` is a table containing `k` rows selected from `data`. Or, if `dim = 2`, then `y` is a table containing `k` variables selected from `data`.
- If `data` is a dataset array and `dim = 1`, then `y` is a dataset array containing `k` rows selected from `data`. Or, if `dim = 2`, then `y` is a dataset array containing `k` variables selected from `data`.

If the input `data` contains missing observations that are represented as `NaN` values, `datasample` samples from the entire input, including the `NaN` values. For example, `y = datasample([NaN 6 14],2)` can return `y = NaN 14`.

When the sample is taken with replacement (default), `y` can contain repeated observations from `data`. Set the `Replace` name-value pair argument to `false` to sample without replacement.

`idx` — Indices
vector

Indices, returned as a vector indicating which elements `datasample` chooses from `data` to create `y`. For example:

- If `data` is a vector, then `y = data(idx)`.
- If `data` is a matrix and `dim = 1`, then `y = data(idx,:)`.
- If `data` is a matrix and `dim = 2`, then `y = data(:,idx)`.

Tips

- To sample random integers with replacement from a range, use `randi`.
- To sample random integers without replacement, use `randperm` or `datasample`.
- To randomly sample from data, with or without replacement, use `datasample`.

Algorithms

`datasample` uses `randperm`, `rand`, or `randi` to generate random values. Therefore, `datasample` changes the state of the MATLAB global random number generator. Control the random number generator using `rng`.

For selecting weighted samples without replacement, `datasample` uses the algorithm of Wong and Easton [1].

Alternative Functionality

You can use `randi` or `randperm` to generate indices for random sampling with or without replacement, respectively. However, `datasample` can be more convenient to use because it samples directly from your data. `datasample` also allows weighted sampling.

References

[1] Wong, C. K. and M. C. Easton. "An Efficient Method for Weighted Sampling Without Replacement." *SIAM Journal of Computing* 9(1), pp. 111–113, 1980.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- `datasample` is useful as a precursor to plotting and fitting a random subset of a large data set. Sampling a large data set preserves trends in the data without requiring the use of all the data points. If the sample is small enough to fit in memory, then you can apply plotting and fitting functions that do not directly support tall arrays.
- `datasample` supports sampling only along the first dimension of the data.
- For tall arrays, `datasample` does not support sampling with replacement. You must specify `'Replace',false`, for example, `datasample(data,k,'Replace',false)`.
- The value of `'Weights'` must be a numeric tall array of the same height as `data`.
- For the syntax `[Y,idx] = datasample(___)`, the output `idx` is a tall logical vector of the same height as `data`. The vector indicates whether each data point is included in the sample.
- If you specify a random number stream, then the underlying generator must support multiple streams and substreams. If you do not specify a random number stream, then `datasample` uses the stream controlled by `tallrng`.

For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`RandStream` | `rand` | `randi` | `randperm` | `rng` | `tallrng`

Introduced in R2011b

dataset class

(Not Recommended) Arrays for statistical data

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Description

Dataset arrays are used to collect heterogeneous data and metadata including variable and observation names into a single container variable. Dataset arrays are suitable for storing column-oriented or tabular data that are often stored as columns in a text file or in a spreadsheet, and can accommodate variables of different types, sizes, units, etc.

Dataset arrays can contain different kinds of variables, including numeric, logical, character, string, categorical, and cell. However, a dataset array is a different class than the variables that it contains. For example, even a dataset array that contains only variables that are double arrays cannot be operated on as if it were itself a double array. However, using dot subscripting, you can operate on variable in a dataset array as if it were a workspace variable.

You can subscript dataset arrays using parentheses much like ordinary numeric arrays, but in addition to numeric and logical indices, you can use variable and observation names as indices.

Construction

Use the `dataset` constructor to create a dataset array from variables in the MATLAB workspace. You can also create a dataset array by reading data from a text or spreadsheet file. You can access each variable in a dataset array much like fields in a structure, using dot subscripting. See the following section for a list of operations available for dataset arrays.

dataset	(Not Recommended) Construct dataset array

Methods

cat	(Not Recommended) Concatenate dataset arrays
cellstr	(Not Recommended) Create cell array of character vectors from dataset array
dataset2cell	(Not Recommended) Convert dataset array to cell array
dataset2struct	(Not Recommended) Convert dataset array to structure
datasetfun	(Not Recommended) Apply function to dataset array variables
disp	(Not Recommended) Display dataset array
display	(Not Recommended) Display dataset array
double	(Not Recommended) Convert dataset variables to double array
end	(Not Recommended) Last index in indexing expression for dataset array
export	(Not Recommended) Write dataset array to file
get	(Not Recommended) Access dataset array properties
horzcat	(Not Recommended) Horizontal concatenation for dataset arrays
intersect	(Not Recommended) Set intersection for dataset array observations
isempty	(Not Recommended) True for empty dataset array
ismember	(Not Recommended) Dataset array elements that are members of set
ismissing	(Not Recommended) Find dataset array elements with missing values
join	(Not Recommended) Merge dataset array observations
length	(Not Recommended) Length of dataset array
ndims	(Not Recommended) Number of dimensions of dataset array
numel	(Not Recommended) Number of elements in dataset array
replacedata	(Not Recommended) Replace dataset variables
replaceWithMissing	(Not Recommended) Insert missing data indicators into a dataset array
set	(Not Recommended) Set and display dataset array properties
setdiff	(Not Recommended) Set difference for dataset array observations
setxor	(Not Recommended) Set exclusive or for dataset array observations
single	(Not Recommended) Convert dataset variables to single array
size	(Not Recommended) Size of dataset array
sortrows	(Not Recommended) Sort rows of dataset array
stack	(Not Recommended) Stack dataset array from multiple variables into single variable
subsasgn	(Not Recommended) Subscripted assignment to dataset array
subsref	(Not Recommended) Subscripted reference for dataset array
summary	(Not Recommended) Print summary of dataset array
union	(Not Recommended) Set union for dataset array observations
unique	(Not Recommended) Unique observations in dataset array
unstack	(Not Recommended) Unstack dataset array from single variable into multiple variables
vertcat	(Not Recommended) Vertical concatenation for dataset arrays

Properties

A dataset array `D` has properties that store metadata (information about your data). Access or assign to a property using `P = D.Properties.PropName` or `D.Properties.PropName = P`, where `PropName` is one of the following:

`Description`

`Description` is a character vector describing the dataset array. The default is an empty character vector.

`DimNames`

A two-element cell array of character vectors giving the names of the two dimensions of the dataset array. The default is `{'Observations' 'Variables'}`.

`ObsNames`

A cell array of nonempty, distinct character vectors giving the names of the observations in the dataset array. This property may be empty, but if not empty, the number of character vectors must equal the number of observations.

`Units`

A cell array of character vectors giving the units of the variables in the dataset array. This property may be empty, but if not empty, the number of character vectors must equal the number of variables. Any individual character vector may be empty for a variable that does not have units defined. The default is an empty cell array.

`UserData`

Any variable containing additional information to be associated with the dataset array. The default is an empty array.

`VarDescription`

A cell array of character vectors giving the descriptions of the variables in the dataset array. This property may be empty, but if not empty, the number of character vectors must equal the number of variables. Any individual character vector may be empty for a variable that does not have a description defined. The default is an empty cell array.

`VarNames`

A cell array of nonempty, distinct character vectors giving the names of the variables in the dataset array. The number of character vectors must equal the number of variables. The default is the cell array of names for the variables used to create the data set.

Copy Semantics

Value. To learn how this affects your use of the class, see Comparing Handle and Value Classes in the MATLAB Object-Oriented Programming documentation.

Examples

Load a dataset array from a .mat file and create some simple subsets:

```
load hospital
h1 = hospital(1:10,:)
h2 = hospital(:,{'LastName' 'Age' 'Sex' 'Smoker'})

% Access and modify metadata
hospital.Properties.Description
hospital.Properties.VarNames{4} = 'Wgt'

% Create a new dataset variable from an existing one
hospital.AtRisk = hospital.Smoker | (hospital.Age > 40)

% Use individual variables to explore the data
boxplot(hospital.Age,hospital.Sex)
h3 = hospital(hospital.Age<30,...
   {'LastName' 'Age' 'Sex' 'Smoker'})

% Sort the observations based on two variables
h4 = sortrows(hospital,{'Sex','Age'})
```

See Also

`tdfread` | `textscan` | `xlsread`

Topics

“Dataset Arrays” on page 2-112

dataset

Class: `dataset`

(Not Recommended) Construct dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
A = dataset(varspec,'ParamName',Value)
A = dataset('File',filename,'ParamName',Value)
A = dataset('XLSFile',filename,'ParamName',Value)
A = dataset('XPTFile',xptfilename,'ParamName',Value)
```

Description

`A = dataset(varspec,'ParamName',Value)` creates dataset array `A` using the workspace variable input method *`varspec`* and one or more optional name/value pairs (see Parameter Name/Value Pairs).

The input method *`varspec`* can be one or more of the following:

- *`VAR`* — a workspace variable. `dataset` uses the workspace name for the variable name in `A`. To include multiple variables, specify *`VAR_1`*,*`VAR_2`*,...,*`VAR_N`*. Variables can be arrays of any size, but all variables must have the same number of rows. `VAR` can also be an expression. In this case, `dataset` creates a default name automatically.
- {*`VAR`*,*`name`*} — a workspace variable, *`VAR`* and a variable name, *`name`* . `dataset` uses *`name`* as the variable name. To include multiple variables and names, specify {*`VAR_1`*,*`name_1`*}, {*`VAR_2`*,*`name_2`*},..., {*`VAR_N`*,*`name_N`*}.
- {*`VAR`*,*`name_1`*,`...`,*`name_m`*} — an *m*-columned workspace variable, *`VAR`*. `dataset` uses the names *`name_1`*, `...`, *`name_m`* as variable names. You must include a name for every column in `VAR`. Each column becomes a separate variable in `A`.

You can combine these input methods to include as many variables and names as needed. Names must be valid, unique MATLAB identifiers. For example input combinations, see Examples. For optional name/value pairs see Inputs.

To convert numeric arrays, cell arrays, structure arrays, or tables to dataset arrays, you can also use (respectively):

- `mat2dataset`
- `cell2dataset`
- `struct2dataset`
- `table2dataset`

Note Dataset arrays may contain built-in types or array objects as variables. Array objects must implement each of the following:

- Standard MATLAB parenthesis indexing of the form `var(i,...)`, where `i` is a numeric or logical vector corresponding to rows of the variable
- A `size` method with a `dim` argument
- A `vertcat` method

`A = dataset('File',filename,'ParamName',Value)` creates dataset array `A` from column-oriented data in the text file specified by `filename`. Variables in `A` are of type `double` if data in the corresponding column of the file, following the column header, are entirely numeric; otherwise the variables in `A` are cell arrays of character vectors. `dataset` converts empty fields to either `NaN` (for a numeric variable) or the empty character vector (for a character-valued variable). `dataset` ignores insignificant white space in the file. You cannot specify both a file and workspace variables as input. See Name/Value Pairs for more information.

`A = dataset('XLSFile',filename,'ParamName',Value)` creates dataset array `A` from column-oriented data in the Excel spreadsheet specified by `filename`. Variables in `A` are of type `double` if data in the corresponding column of the spreadsheet, following the column header, are entirely numeric; otherwise the variables in `A` are cell arrays of character vectors. See Name/Value Pairs for more information.

`A = dataset('XPTFile',xptfilename,'ParamName',Value)` creates a dataset array from a SAS® XPORT format file. Variable names from the XPORT format file are preserved. Numeric data types in the XPORT format file are preserved but all other data types are converted to cell arrays of character vectors. The XPORT format allows for 28 missing data types. `dataset` represents these in the file by an upper case letter, `'.'` or `'_'`. `dataset` converts all missing data to `NaN` values in `A`. See Name/Value Pairs for more information.

Parameter Name/Value Pairs

Specify one or more of the following name/value pairs when constructing a dataset:

`VarNames`

A string array or cell array `{name_1,...,name_m}` naming the `m` variables in `A` with the specified variable names. Names must be valid, unique MATLAB identifiers. The number of names must equal the number of variables in `A`. You cannot use the `VarNames` parameter if you provide names for individual variables using `{VAR,name}` pairs. To specify `VarNames` when using a file as input, set `ReadVarNames` to `false`.

`ObsNames`

A string array or cell array `{name_1,...,name_n}` naming the `n` observations in `A` with the specified observation names. The names need not be valid MATLAB identifiers, but must be unique. The number of names must equal the number of observations (rows) in `A`. To specify `ObsNames` when using a file as input, set `ReadObsNames` to `false`.

Name/value pairs available when using text files as inputs:

`Delimiter`

A character vector or string scalar indicating the character separating columns in the file. Values are

- `'\t'` (tab, the default when no `format` is specified)
- `' '` (space, the default when a `format` is specified)
- `','` (comma)
- `';'` (semicolon)
- `'|'` (bar)

`Format`

A format character vector or string scalar, as accepted by `textscan`. `dataset` reads the file using `textscan`, and creates variables in `A` according to the conversion specifiers in the format character vector or string scalar. You may also provide any name/value pairs accepted by `textscan`. Using the `Format` parameter is much faster for large files. If `ReadObsNames` is `true`, then `format` should include a format specifier for the first column of the file.

`HeaderLines`

Numeric value indicating the number of lines to skip at the beginning of a file.

Default: `0`

`TreatAsEmpty`

Specifies characters to treat as the empty character vector in a numeric column. Values may be a character array, a string array, or a cell array of character vectors. The parameter applies only to numeric columns in the file; `dataset` does not accept numeric literals such as `'-99'`.

Name/value pairs available when using text files or Excel spreadsheets as inputs:

`ReadVarNames`

A logical value indicating whether (`true`) or not (`false`) to read variable names from the first row of the file. The default is `true`. If `ReadVarNames` is `true`, variable names in the column headers of the file or range (if using an Excel spreadsheet) cannot be empty.

`ReadObsNames`

A logical value indicating whether (`true`) or not (`false`) to read observation names from the first column of the file or range (if using an Excel spreadsheet). The default is `false`. If `ReadObsNames` and `ReadVarNames` are both `true`, `dataset` saves the header of the first column in the file or range as the name of the first dimension in `A.Properties.DimNames`.

When reading from an `XPT` format file, the `ReadObsNames` parameter name/value pair determines whether or not to try to use the first variable in the file as observation names. Specify as a logical value (default `false`). If the contents of the first variable are not valid observation names then `dataset` reads the variable into a variable of the dataset array and does not set the observation names.

Name/value pairs available when using Excel spreadsheets as input:

Sheet

A positive scalar value of type `double` indicating the sheet number, or a quoted sheet name.

Range

A character vector or string scalar of the form `'C1:C2'` where `C1` and `C2` are the names of cells at opposing corners of a rectangular region to be read, as for `xlsread`. By default, the rectangular region extends to the right-most column containing data. If the spreadsheet contains empty columns between columns of data, or if the spreadsheet contains figures or other non-tabular information, specify a range that contains only data.

Examples

Create a dataset array from workspace variables, including observation names:

```
load cereal
cereal = dataset(Calories,Protein,Fat,Sodium,Fiber,Carbo,...
   Sugars,'ObsNames',Name)
cereal.Properties.VarDescription = Variables(4:10,2);
```

Create a dataset array from a single, multi-columned workspace variable, designating variable names for each column:

```
load cities
categories = cellstr(categories);
cities = dataset({ratings,categories{:}},...
   'ObsNames',cellstr(names))
```

Load data from a text or spreadsheet file

```
patients = dataset('File','hospital.dat',...
   'Delimiter',',','ReadObsNames',true)
patients2 = dataset('XLSFile','hospital.xls',...
   'ReadObsNames',true)
```

1 Load patient data from the CSV file `hospital.dat` and store the information in a `dataset` array with observation names given by the first column in the data (patient identification):

```
patients = dataset('file','hospital.dat', ...
             'format','%s%s%s%f%f%f%f%f%f%f%f', ...
             'Delimiter',',','ReadObsNames',true);
```

You can also load the data without specifying a format. `dataset` will automatically create `dataset` variables that are either `double` arrays or cell arrays of character vectors, depending on the contents of the file:

```
patients = dataset('file','hospital.dat',...
                   'delimiter',',',...
                   'ReadObsNames',true);
```

2 Make the {0,1}-valued variable `smoke` nominal, and change the labels to `'No'` and `'Yes'`:

```
patients.smoke = nominal(patients.smoke,{'No','Yes'});
```

3 Add new levels to `smoke` as placeholders for more detailed histories of smokers:

```
patients.smoke = addlevels(patients.smoke,...
                 {'0-5 Years','5-10 Years','LongTerm'});
```

4 Assuming the nonsmokers have never smoked, relabel the `'No'` level:

```
patients.smoke = setlabels(patients.smoke,'Never','No');
```

5 Drop the undifferentiated `'Yes'` level from `smoke`:

```
patients.smoke = droplevels(patients.smoke,'Yes');

Warning: OLDLEVELS contains categorical levels that
were present in A, caused some array elements to have
undefined levels.
```

Note that smokers now have an undefined level.

6 Set each smoker to one of the new levels, by observation name:

```
patients.smoke('YPL-320') = '5-10 Years';
```

See Also

`cell2dataset` | `mat2dataset` | `struct2dataset` | `tdfread` | `textscan` | `xlsread`

Topics

“Create a Dataset Array from Workspace Variables” on page 2-57
“Create a Dataset Array from a File” on page 2-62
“Dataset Arrays in the Variables Editor” on page 2-101
“Dataset Arrays” on page 2-112

Introduced in R2007a

dataset2cell

Class: `dataset`

(Not Recommended) Convert dataset array to cell array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = dataset2cell(D)
```

Description

`C = dataset2cell(D)` converts the dataset array `D` to a cell array `C`. Each variable of `D` becomes a column in `C`. If `D` is an `M`-by-`N` array, then `C` is `(M+1)`-by-`N`, with the variable names of `D` in the first row. If `D` contains observation names, then `C` is `(M+1)`-by-`(N+1)`, with the observation names in the first column.

See Also

`cell2dataset` | `dataset` | `export`

Topics

“Dataset Arrays” on page 2-112

dataset2struct

Class: `dataset`

(Not Recommended) Convert dataset array to structure

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
S = dataset2struct(D)
S = dataset2struct(D,'AsScalar',true)
```

Description

`S = dataset2struct(D)` converts a dataset array to a structure array. Each variable of `D` becomes a field in `S`. If `D` is an *M*-by-*N* dataset array, then `S` is *M*-by-1 and has *N* fields. If `D` contains observation names, then `S` contains those names in the additional field `ObsNames`.

`S = dataset2struct(D,'AsScalar',true)` converts a dataset array to a scalar structure. Each variable of `D` becomes a field in `S`. If `D` is an *M*-by-*N* dataset array, then `S` has *N* fields, each of which as *M* rows. If `D` contains observation names, then `S` contains those names in the additional field `ObsNames`.

Input Arguments

`D`

M-by-*N* dataset array.

Output Arguments

`S`

M-by-1 structure array, with *N* fields. If the input dataset array contains observation names, then `S` has an additional field `ObsNames`.

If you specify `'AsScalar',true`, then `S` is a scalar structure, with *N* fields, each with *M* rows.

Examples

Convert Dataset Array to Structure Array

Load sample dataset array.

```
load('hospital')
```

Create a dataset array, D, that has only a subset of the observations and variables.

```
D = hospital(1:8,{'LastName','Sex','Age'});
size(D)
```

```
ans = 1×2

     8     3
```

The dataset array D has 8 observations and 3 variables.

Convert D to a structure array.

```
S = dataset2struct(D)
```

```
S=8×1 struct array with fields:
    ObsNames
    LastName
    Sex
    Age
```

The structure is 8x1, corresponding to the 8 observations in the dataset array. S also has the field ObsNames, since D had observation names.

Display the field data for the first element of S.

```
S(1)
```

```
ans = struct with fields:
    ObsNames: 'YPL-320'
    LastName: 'SMITH'
         Sex: Male
         Age: 38
```

This information corresponds to the first observation (row) of the dataset array.

Convert Dataset Array to Scalar Structure

Load sample dataset array.

```
load('hospital')
```

Create a dataset array, D, that has only a subset of the observations and variables.

```
D = hospital(1:8,{'LastName','Sex','Age'});
size(D)
```

```
ans = 1×2

     8     3
```

The dataset array D has 8 observations and 3 variables.

Convert `D` to a scalar structure array.

```
S = dataset2struct(D,'AsScalar',true)
```

```
S = struct with fields:
    ObsNames: {8x1 cell}
    LastName: {8x1 cell}
         Sex: [8x1 nominal]
         Age: [8x1 double]
```

The data in the fields of the scalar structure is `8x1`, corresponding to the 8 observations in the dataset array. `S` also has the field `ObsNames`, since `D` had observation names.

Display the data for the field `LastName`.

```
S.LastName
```

```
ans = 8x1 cell
    {'SMITH'   }
    {'JOHNSON' }
    {'WILLIAMS'}
    {'JONES'   }
    {'BROWN'   }
    {'DAVIS'   }
    {'MILLER'  }
    {'WILSON'  }
```

The structure field `LastName` contains all of the data that was in the original dataset array variable, `LastName`.

See Also

`dataset` | `dataset2cell` | `struct2dataset`

Topics

“Dataset Arrays” on page 2-112

dataset2table

Convert dataset array to table

Syntax

```
t = dataset2table(ds)
```

Description

`t = dataset2table(ds)` converts a dataset array to a table.

Examples

Convert Dataset Array to Table

Load the sample data, which contains nutritional information for 77 cereals.

```
load cereal;
```

Create a dataset array containing the calorie, protein, fat, and name data for the first five cereals. Label the variables.

```
Calories = Calories(1:5);
Protein = Protein(1:5);
Fat = Fat(1:5);
Name = Name(1:5);

cereal = dataset(Calories,Protein,Fat,'ObsNames',Name)
```

```
cereal =
                              Calories    Protein    Fat
    100% Bran                  70         4          1
    100% Natural Bran         120         3          5
    All-Bran                   70         4          1
    All-Bran with Extra Fiber  50         4          0
    Almond Delight            110         2          2
```

```
cereal.Properties.VarDescription = Variables(4:6,2);
```

Convert the dataset array to a table.

```
t = dataset2table(cereal)
```

```
t=5×3 table
                              Calories    Protein    Fat
                              ________    _______    ___

    100% Bran                    70          4        1
    100% Natural Bran           120          3        5
    All-Bran                     70          4        1
    All-Bran with Extra Fiber    50          4        0
```

```
    Almond Delight                  110          2        2
```

Input Arguments

ds — Input dataset array

dataset array

Input dataset array to convert to a table, specified as a dataset array. Each variable in `ds` becomes a variable in the output table `t`.

Output Arguments

t — Output table

table

Output table, returned as a table. The table can store metadata such as descriptions, variable units, variable names, and row names. For more information, see “Tables”.

See Also

`dataset` | `table`

Topics

“Dataset Arrays” on page 2-112
“Tables”

Introduced in R2013b

datasetfun

Class: `dataset`

(Not Recommended) Apply function to dataset array variables

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
b = datasetfun(fun,A)
[b,c,...] = datasetfun(fun,A)
[b,...] = datasetfun(fun,A,...,'UniformOutput',false)
[b,...] = datasetfun(fun,A,...,'DatasetOutput',true)
[b,...] = datasetfun(fun,A,...,'DataVars',vars)
[b,...] = datasetfun(fun,A,...,'ObsNames',obsnames)
[b,...] = datasetfun(fun,A,...,'ErrorHandler',efun)
```

Description

`b = datasetfun(fun,A)` applies the function specified by `fun` to each variable of the dataset array `A`, and returns the results in the vector `b`. The *i*th element of `b` is equal to `fun` applied to the *i*th dataset variable of `A`. `fun` is a function handle to a function that takes one input argument and returns a scalar value. `fun` must return values of the same class each time it is called, and `datasetfun` concatenates them into the vector `b`. The outputs from `fun` must be one of the following types: numeric, logical, character, structure, or cell.

To apply functions that return results that are nonscalar or of different sizes and types, use the `'UniformOutput'` or `'DatasetOutput'` parameters described below.

Do not rely on the order in which `datasetfun` computes the elements of `b`, which is unspecified.

If `fun` is bound to more than one built-in function or file, (that is, if it represents a set of overloaded functions), `datasetfun` follows MATLAB dispatching rules in calling the function. (See “Function Precedence Order”.)

`[b,c,...] = datasetfun(fun,A)`, where `fun` is a function handle to a function that returns multiple outputs, returns vectors `b`, `c`, ..., each corresponding to one of the output arguments of `fun`. `datasetfun` calls `fun` each time with as many outputs as there are in the call to `datasetfun`. `fun` may return output arguments having different classes, but the class of each output must be the same each time `fun` is called.

`[b,...] = datasetfun(fun,A,...,'UniformOutput',false)` allows you to specify a function `fun` that returns values of different sizes or types. `datasetfun` returns a cell array (or multiple cell arrays), where the *i*th cell contains the value of `fun` applied to the *i*th dataset variable of `A`. Setting `'UniformOutput'` to `true` is equivalent to the default behavior.

`[b,...] = datasetfun(fun,A,...,'DatasetOutput',true)` specifies that the output(s) of `fun` are returned as variables in a dataset array (or multiple dataset arrays). `fun` must return values

with the same number of rows each time it is called, but it may return values of any type. The variables in the output dataset array(s) have the same names as the variables in the input. Setting `'DatasetOutput'` to `false` (the default) specifies that the type of the output(s) from `datasetfun` is determined by `'UniformOutput'`.

`[b,...] = datasetfun(fun,A,...,'DataVars',vars)` allows you to apply `fun` only to the dataset variables in `A` specified by `vars`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

`[b,...] = datasetfun(fun,A,...,'ObsNames',obsnames)` specifies observation names for the dataset output when `'DatasetOutput'` is `true`.

`[b,...] = datasetfun(fun,A,...,'ErrorHandler',efun)`, where `efun` is a function handle, specifies the MATLAB function to call if the call to `fun` fails. The error-handling function is called with the following input arguments:

- A structure with the fields `identifier`, `message`, and `index`, respectively containing the identifier of the error that occurred, the text of the error message, and the linear index into the input array(s) at which the error occurred
- The set of input arguments at which the call to the function failed

The error-handling function should either re-throw an error, or return the same number of outputs as `fun`. These outputs are then returned as the outputs of `datasetfun`. If `'UniformOutput'` is true, the outputs of the error handler must also be scalars of the same type as the outputs of `fun`. For example, the following code could be saved in a file as the error-handling function:

```
function [A,B] = errorFunc(S,varargin)

warning(S.identifier,S.message);
A = NaN;
B = NaN;
```

If an error-handling function is not specified, the error from the call to `fun` is rethrown.

Examples

Work With Datasets Using Function Handles

Use function handles to compute the mean and plot a histogram of selected variables in a dataset array.

Load the sample data.

```
load hospital
```

Use `datasetfun` to compute the means of the `Weight` and `BloodPressure` variables, and store the results in a dataset array.

```
stats = datasetfun(@mean,hospital,...
        'DataVars',{'Weight','BloodPressure'},...
        'UniformOutput',false)
```

```
stats=1×2 cell array
    {[154]}    {[122.7800 82.9600]}
```

The variable `BloodPressure` contains two columns: One for the systolic measurement, and one for the diastolic measurement.

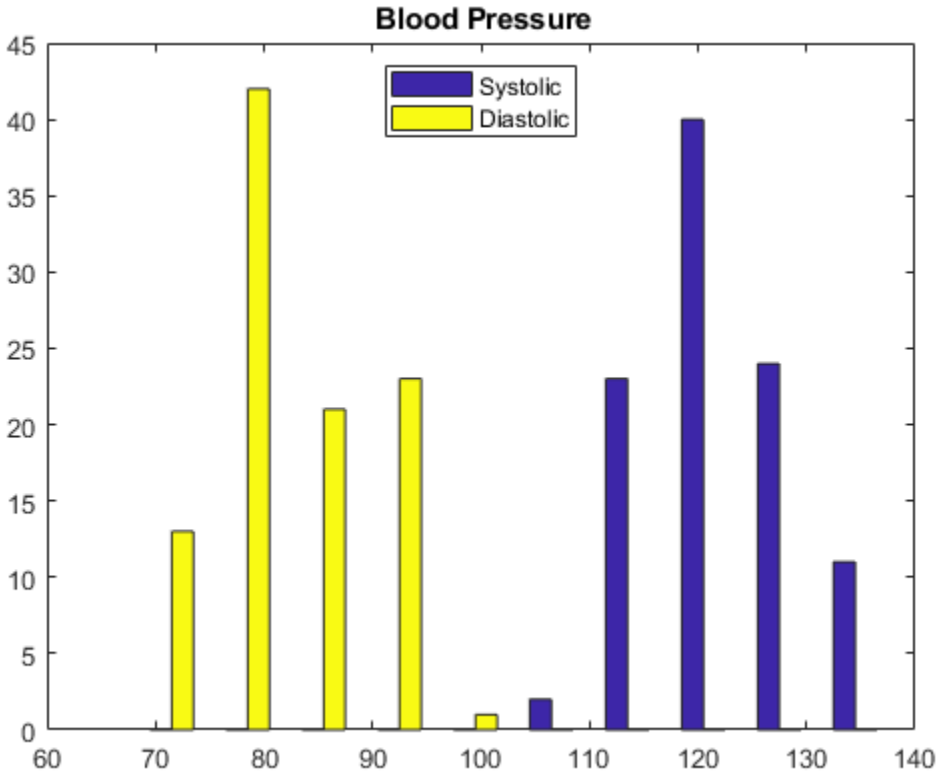
Display the mean of the blood pressure variable.

```
stats{2}
```

```
ans = 1×2

  122.7800   82.9600
```

Plot a histogram of the blood pressure variable.

```
datasetfun(@hist,hospital,...
           'DataVars','BloodPressure',...
           'UniformOutput',false);
title('{\bf Blood Pressure}')
legend('Systolic','Diastolic','Location','N')
```



See Also

`grpstats`

daugment

D-optimal augmentation

Syntax

```
dCE2 = daugment(dCE,mruns)
[dCE2,X] = daugment(dCE,mruns)
[dCE2,X] = daugment(dCE,mruns,model)
[dCE2,X] = daugment(...,param1,val1,param2,val2,...)
```

Description

`dCE2 = daugment(dCE,mruns)` uses a coordinate-exchange algorithm to *D*-optimally add `mruns` runs to an existing experimental design `dCE` for a linear additive model.

`[dCE2,X] = daugment(dCE,mruns)` also returns the design matrix `X` associated with the augmented design.

`[dCE2,X] = daugment(dCE,mruns,model)` uses the linear regression model specified in `model`. `model` is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

The order of the columns of `X` for a full quadratic model with *n* terms is:

1. The constant term
2. The linear terms in order 1, 2, ..., *n*
3. The interaction terms in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n* – 1, *n*)
4. The squared terms in order 1, 2, ..., *n*

Other models use a subset of these terms, in the same order.

Alternatively, `model` can be a matrix specifying polynomial terms of arbitrary order. In this case, `model` should have one column for each factor and one row for each term in the model. The entries in any row of `model` are powers for the factors in the columns. For example, if a model has factors `X1`, `X2`, and `X3`, then a row `[0 1 2]` in `model` specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in `model` specifies a constant term, which can be omitted.

`[dCE2,X] = daugment(...,param1,val1,param2,val2,...)` specifies additional parameter/value pairs for the design. Valid parameters and their values are listed in the following table.

Parameter	Value
'bounds'	Lower and upper bounds for each factor, specified as a 2-by-nfactors matrix, where nfactors is the number of factors. Alternatively, this value can be a cell array containing nfactors elements, each element specifying the vector of allowable values for the corresponding factor.
'categorical'	Indices of categorical predictors.
'display'	Either 'on' or 'off' to control display of the iteration counter. The default is 'on'.
'excludefun'	Handle to a function that excludes undesirable runs. If the function is *f*, it must support the syntax $b = f(S)$, where S is a matrix of treatments with nfactors columns, where nfactors is the number of factors, and b is a vector of Boolean values with the same number of rows as S. $b(i)$ is true if the ith row S should be excluded.
'init'	Initial design as an mruns-by-nfactors matrix, where nfactors is the number of factors. The default is a randomly selected set of points.
'levels'	Vector of number of levels for each factor.
'maxiter'	Maximum number of iterations. The default is 10.

'options'	The value is a structure that contains options specifying whether to compute multiple tries in parallel, and specifying how to use random numbers when generating the starting points for the tries. Create the options structure with statset. Applicable statset parameters are:
	• 'UseParallel' — If true and if Parallel Computing Toolbox is installed, compute in parallel. If the Parallel Computing Toolbox is not installed, or 'UseParallel',false, then computation occurs in serial mode. Default is false, meaning serial computation.
	• UseSubstreams — Set to true to compute in parallel in a reproducible fashion. Default is false. To compute reproducibly, set Streams to a type allowing substreams: 'mlfg6331_64' or 'mrg32k3a'.
	• Streams — A RandStream object or cell array of such objects. If you do not specify Streams, daugment uses the default stream or streams. If you choose to specify Streams, use a single object except in the case
	  • UseParallel is true
	  • UseSubstreams is false
	  In that case, use a cell array the same size as the Parallel pool.

'tries'	Number of times to try to generate a design from a new starting point. The algorithm uses random points for each try, except possibly the first. The default is 1.

Note The daugment function augments an existing design using a coordinate-exchange algorithm; the 'start' parameter of the candexch function provides the same functionality using a row-exchange algorithm.

Examples

The following eight-run design is adequate for estimating main effects in a four-factor model:

```
dCEmain = cordexch(4,8)
dCEmain =
     1    -1    -1     1
    -1    -1     1     1
    -1     1    -1     1
     1     1     1    -1
     1     1     1     1
    -1     1    -1    -1
     1    -1    -1    -1
    -1    -1     1    -1
```

To estimate the six interaction terms in the model, augment the design with eight additional runs:

```
dCEinteraction = daugment(dCEmain,8,'interaction')
dCEinteraction =
     1    -1    -1     1
    -1    -1     1     1
    -1     1    -1     1
     1     1     1    -1
     1     1     1     1
    -1     1    -1    -1
     1    -1    -1    -1
    -1    -1     1    -1
    -1     1     1     1
    -1    -1    -1    -1
     1    -1     1    -1
     1     1    -1     1
    -1     1     1    -1
     1     1    -1    -1
     1    -1     1     1
     1     1     1    -1
```

The augmented design is full factorial, with the original eight runs in the first eight rows.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`candexch` | `cordexch` | `dcovary`

Introduced before R2006a

dbscan

Density-based spatial clustering of applications with noise (DBSCAN)

Syntax

```
idx = dbscan(X,epsilon,minpts)
idx = dbscan(X,epsilon,minpts,Name,Value)

idx = dbscan(D,epsilon,minpts,'Distance','precomputed')

[idx,corepts] = dbscan( ___ )
```

Description

`idx = dbscan(X,epsilon,minpts)` partitions observations in the *n*-by-*p* data matrix `X` into clusters using the DBSCAN algorithm (see Algorithms on page 33-1168). `dbscan` clusters the observations (or points) based on a threshold for a neighborhood search radius `epsilon` and a minimum number of neighbors `minpts` required to identify a core point. The function returns an *n*-by-1 vector (`idx`) containing cluster indices of each observation.

`idx = dbscan(X,epsilon,minpts,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify `'Distance','minkowski','P',3` to use the Minkowski distance metric with an exponent of three in the DBSCAN algorithm.

`idx = dbscan(D,epsilon,minpts,'Distance','precomputed')` returns a vector of cluster indices for the precomputed pairwise distances `D` between observations. `D` can be the output of `pdist` or `pdist2`, or a more general dissimilarity vector or matrix conforming to the output format of `pdist` or `pdist2`, respectively.

`[idx,corepts] = dbscan( ___ )` also returns a logical vector `corepts` that contains the core points identified by `dbscan`, using any of the input argument combinations in the previous syntaxes.

Examples

Perform DBSCAN on Input Data

Cluster a 2-D circular data set using DBSCAN with the default Euclidean distance metric. Also, compare the results of clustering the data set using DBSCAN and *k*-Means clustering with the squared Euclidean distance metric.

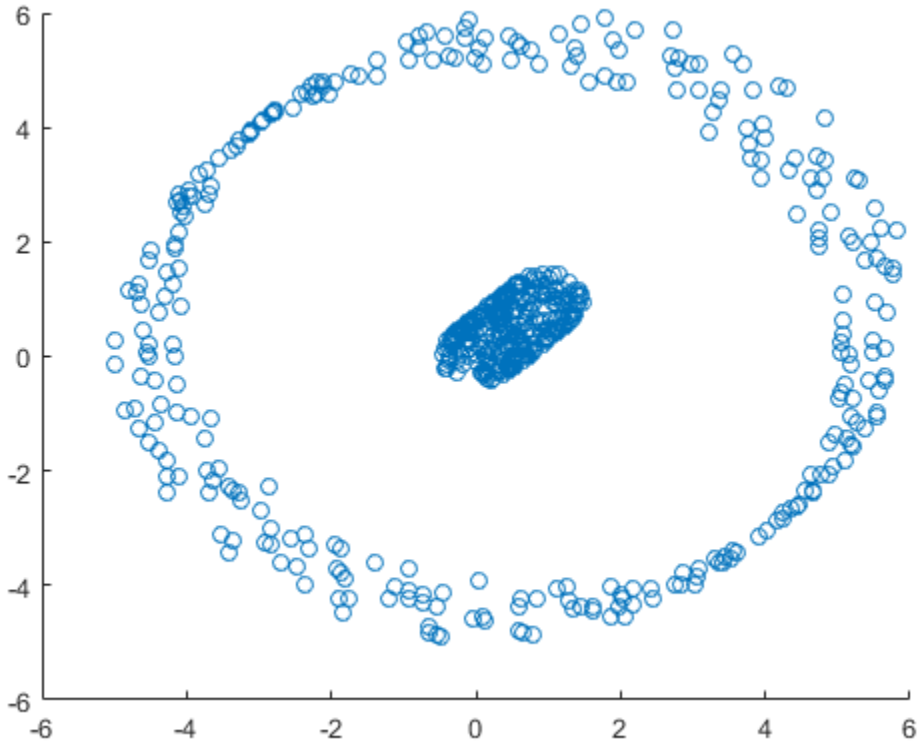
Generate synthetic data that contains two noisy circles.

```
rng('default') % For reproducibility

% Parameters for data generation
N = 300;  % Size of each cluster
r1 = 0.5; % Radius of first circle
r2 = 5;   % Radius of second circle
theta = linspace(0,2*pi,N)';
```

```
X1 = r1*[cos(theta),sin(theta)]+ rand(N,1);
X2 = r2*[cos(theta),sin(theta)]+ rand(N,1);
X = [X1;X2]; % Noisy 2-D circular data set
```

Visualize the data set.

```
scatter(X(:,1),X(:,2))
```



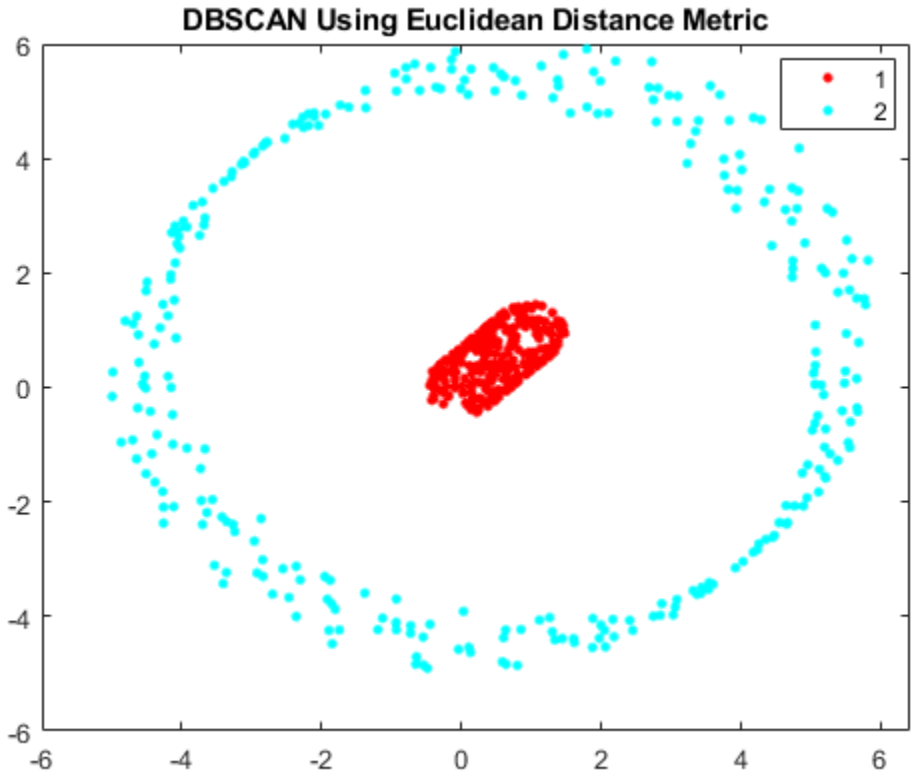
The plot shows that the data set contains two distinct clusters.

Perform DBSCAN clustering on the data. Specify an `epsilon` value of 1 and a `minpts` value of 5.

```
idx = dbscan(X,1,5); % The default distance metric is Euclidean distance
```

Visualize the clustering.

```
gscatter(X(:,1),X(:,2),idx);
title('DBSCAN Using Euclidean Distance Metric')
```



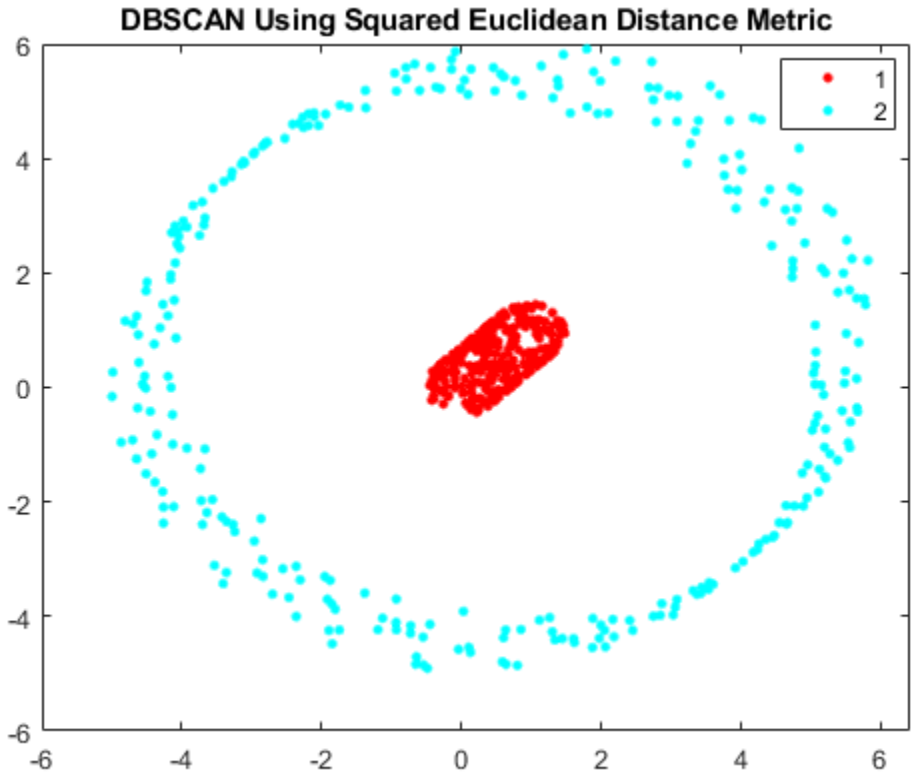
Using the Euclidean distance metric, DBSCAN correctly identifies the two clusters in the data set.

Perform DBSCAN clustering using the squared Euclidean distance metric. Specify an `epsilon` value of 1 and a `minpts` value of 5.

```
idx2 = dbscan(X,1,5,'Distance','squaredeuclidean');
```

Visualize the clustering.

```
gscatter(X(:,1),X(:,2),idx2);
title('DBSCAN Using Squared Euclidean Distance Metric')
```



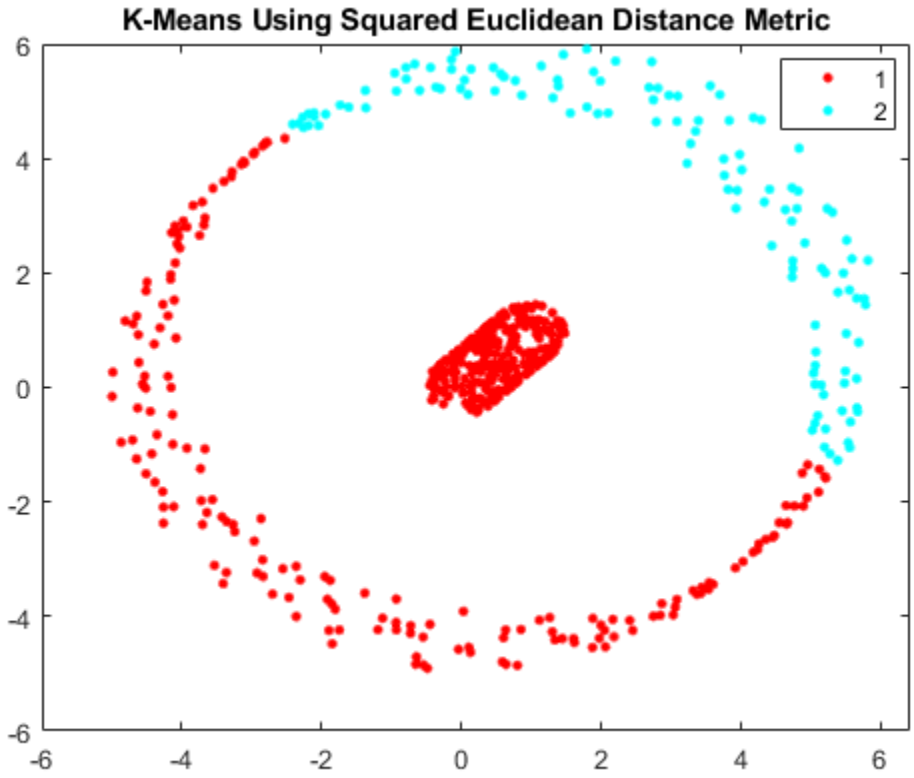
Using the squared Euclidean distance metric, DBSCAN correctly identifies the two clusters in the data set.

Perform *k*-Means clustering using the squared Euclidean distance metric. Specify $k = 2$ clusters.

```
kidx = kmeans(X,2); % The default distance metric is squared Euclidean distance
```

Visualize the clustering.

```
gscatter(X(:,1),X(:,2),kidx);
title('K-Means Using Squared Euclidean Distance Metric')
```



Using the squared Euclidean distance metric, *k*-Means clustering fails to correctly identify the two clusters in the data set.

Perform DBSCAN on Pairwise Distances

Perform DBSCAN clustering using a matrix of pairwise distances between observations as input to the `dbscan` function, and find the number of outliers and core points. The data set is a Lidar scan, stored as a collection of 3-D points, that contains the coordinates of objects surrounding a vehicle.

Load the *x*, *y*, *z* coordinates of the objects.

```
load('lidar_subset.mat')
loc = lidar_subset;
```

To highlight the environment around the vehicle, set the region of interest to span 20 meters to the left and right of the vehicle, 20 meters in front and back of the vehicle, and the area above the surface of the road.

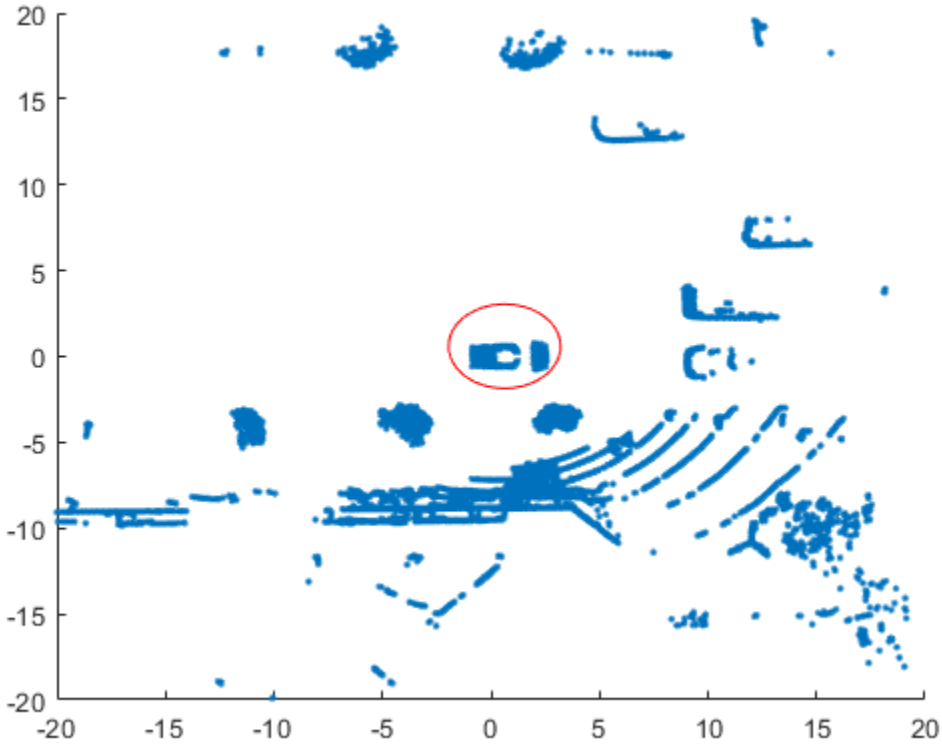
```
xBound = 20; % in meters
yBound = 20; % in meters
zLowerBound = 0; % in meters
```

Crop the data to contain only points within the specified region.

```
indices = loc(:,1) <= xBound & loc(:,1) >= -xBound ...
    & loc(:,2) <= yBound & loc(:,2) >= -yBound ...
    & loc(:,3) > zLowerBound;
loc = loc(indices,:);
```

Visualize the data as a 2-D scatter plot. Annotate the plot to highlight the vehicle.

```
scatter(loc(:,1),loc(:,2),'.');
annotation('ellipse',[0.48 0.48 .1 .1],'Color','red')
```



The center of the set of points (circled in red) contains the roof and hood of the vehicle. All other points are obstacles.

Precompute a matrix of pairwise distances `D` between observations by using the `pdist2` function.

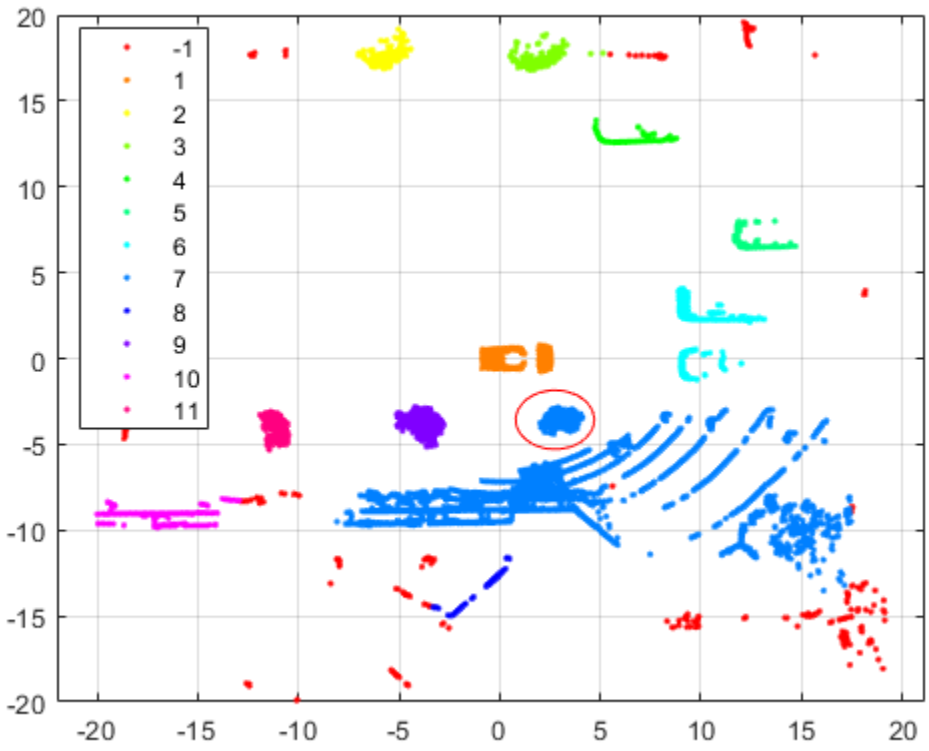
```
D = pdist2(loc,loc);
```

Cluster the data by using `dbscan` with the pairwise distances. Specify an `epsilon` value of 2 and a `minpts` value of 50.

```
[idx, corepts] = dbscan(D,2,50,'Distance','precomputed');
```

Visualize the results and annotate the figure to highlight a specific cluster.

```
gscatter(loc(:,1),loc(:,2),idx);
annotation('ellipse',[0.54 0.41 .07 .07],'Color','red')
grid
```



As shown in the scatter plot, `dbscan` identifies 11 clusters and places the vehicle in a separate cluster.

`dbscan` assigns the group of points circled in red (and centered around `(3,−4)`) to the same cluster (group 7) as the group of points in the southeast quadrant of the plot. The expectation is that these groups should be in separate clusters. You can try using a smaller value of `epsilon` to split up large clusters and further partition the points.

The function also identifies some outliers (an `idx` value of `−1`) in the data. Find the number of points that `dbscan` identifies as outliers.

```
sum(idx == -1)
```

```
ans = 412
```

`dbscan` identifies 412 outliers out of 19,070 observations.

Find the number of points that `dbscan` identifies as core points. A `corepts` value of `1` indicates a core point.

```
sum(corepts == 1)
```

```
ans = 18446
```

`dbscan` identifies 18,446 observations as core points.

See “Determine Values for DBSCAN Parameters” on page 16-20 for a more extensive example.

Input Arguments

X — Input data

numeric matrix

Input data, specified as an *n*-by-*p* numeric matrix. The rows of `X` correspond to observations (or points), and the columns correspond to variables.

Data Types: `single` | `double`

D — Pairwise distances

numeric row vector | numeric square matrix | logical row vector | logical square matrix

Pairwise distances between observations, specified as a numeric row vector that is the output of `pdist`, numeric square matrix that is the output of `pdist2`, logical row vector, or logical square matrix. `D` can also be a more general dissimilarity vector or matrix that conforms to the output format of `pdist` or `pdist2`, respectively.

For the aforementioned specifications, the following table describes the formats that `D` can take, given an input matrix `X` that has *n* observations (rows) and *p* dimensions (columns).

Specification	Format
Numeric row vector (output of `pdist(X)`)	• A row vector of length $n(n-1)/2$, corresponding to pairs of observations in `X` • Distances arranged in the order (2,1), (3,1), ..., (*n*,1), (3,2), ..., (*n*,2), ..., (*n*,*n* – 1))
Numeric square matrix (output of `pdist2(X,X)`)	• An *n*-by-*n* matrix, where `D(i,j)` is the distance between observations `i` and `j` in `X` • A symmetric matrix having diagonal elements equal to zero
Logical row vector	• A row vector of length $n(n-1)/2$, corresponding to pairs of observations in `X` • A logical row vector with elements indicating distances that are less than or equal to `epsilon` • Elements of `D` arranged in the order (2,1), (3,1), ..., (*n*,1), (3,2), ..., (*n*,2), ..., (*n*,*n* – 1))
Logical square matrix	• An *n*-by-*n* matrix, where `D(i,j)` indicates the distance between observations `i` and `j` in `X` that are less than or equal to `epsilon`

Note If `D` is a logical vector or matrix, then the value of `epsilon` must be empty; for example, `dbscan(D,[],5,'Distance','precomputed')`.

Data Types: `single` | `double` | `logical`

epsilon — Epsilon neighborhood

numeric scalar | `[]`

Epsilon neighborhood of a point, specified as a numeric scalar that defines a neighborhood search radius around the point. If the epsilon neighborhood of a point contains at least `minpts` neighbors, then `dbscan` identifies the point as a core point.

The value of `epsilon` must be empty (`[]`) when `D` is a logical vector or matrix.

Example: `dbscan(X,2.5,10)`

Example: `dbscan(D,[],5,'Distance','precomputed')`, for a logical matrix or vector `D`

Data Types: `single` | `double`

`minpts` — Minimum number of neighbors required for core point
positive integer

Minimum number of neighbors required for a core point, specified as a positive integer. The epsilon neighborhood of a core point in a cluster must contain at least `minpts` neighbors, whereas the epsilon neighborhood of a border point can contain fewer neighbors than `minpts`.

Example: `dbscan(X,2.5,5)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `dbscan(D,2.5,5,'Distance','precomputed')` specifies DBSCAN clustering using a precomputed matrix of pairwise distances `D` between observations, an epsilon neighborhood of `2.5`, and a minimum of `5` neighbors.

`Distance` — Distance metric
character vector | string scalar | function handle

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle, as described in this table.

Value	Description
`'precomputed'`	Precomputed distances. You must specify this option if the first input to `dbscan` is a vector or matrix of pairwise distances `D`.
`'euclidean'`	Euclidean distance (default)
`'squaredeuclidean'`	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`. Use `Scale` to specify another value for `S`.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use `Cov` to specify another value for `C`, where the matrix `C` is symmetric and positive definite.
`'cityblock'`	City block distance

Value	Description
`'minkowski'`	Minkowski distance. The default exponent is 2. Use `P` to specify a different exponent, where `P` is a positive scalar value.
`'chebychev'`	Chebychev distance (maximum coordinate difference)
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For definitions, see Distance Metrics on page 33-4495.

When you use the `'seuclidean'`, `'minkowski'`, or `'mahalanobis'` distance metric, you can specify the additional name-value pair argument `'Scale'`, `'P'`, or `'Cov'`, respectively, to control the distance metrics.

Example: `dbscan(X,2.5,5,'Distance','minkowski','P',3)` specifies an epsilon neighborhood of `2.5`, a minimum of `5` neighbors to grow a cluster, and use of the Minkowski distance metric with an exponent of `3` when performing the clustering algorithm.

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar.

This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

`Cov` — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | numeric matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a symmetric, positive definite, numeric matrix.

This argument is valid only if `'Distance'` is `'mahalanobis'`.

Data Types: `single` | `double`

`Scale` — Scaling factors for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | numeric vector of positive values

Scaling factors for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a numeric vector of positive values.

Each dimension (column) of `X` has a corresponding value in `'Scale'`; therefore, `'Scale'` is of length *p* (the number of columns in `X`). For each dimension of `X`, `dbscan` uses the corresponding value in `'Scale'` to standardize the difference between `X` and a query point.

This argument is valid only if `'Distance'` is `'seuclidean'`.

Data Types: `single` | `double`

Output Arguments

`idx` — Cluster indices
numeric column vector

Cluster indices, returned as a numeric column vector. `idx` has *n* rows, and each row of `idx` indicates the cluster assignment of the corresponding observation in `X`. An index equal to `−1` indicates an outlier (or noise point).

Note Cluster assignment using the DBSCAN algorithm is dependent on the order of observations. Therefore, shuffling the rows of `X` can lead to different cluster assignments for the observations. For more details, see Algorithms on page 33-1168.

Data Types: `double`

`corepts` — Indicator for core points
logical vector

Indicator for core points, returned as an *n*-by-*1* logical vector indicating the indices of the core points identified by `dbscan`. A value of `1` in any row of `corepts` indicates that the corresponding observation in `X` is a core point. Otherwise, `corepts` has a value of `0` for rows corresponding to observations that are not core points.

Data Types: `logical`

More About

Core Points

Core points in a cluster are points that have at least a minimum number of neighbors (`minpts`) in their epsilon neighborhood (`epsilon`). Each cluster must contain at least one core point.

Border Points

Border points in a cluster are points that have fewer than the required minimum number of neighbors for a core point (`minpts`) in their epsilon neighborhood (`epsilon`). Generally, the epsilon neighborhood of a border point contains significantly fewer points than the epsilon neighborhood of a core point.

Noise Points

Noise points are outliers that do not belong to any cluster.

Tips

- For improved speed when iterating over many values of `epsilon`, consider passing in `D` as the input to `dbscan`. This approach prevents the function from having to compute the distances at every point of the iteration.
- If you use `pdist2` to precompute `D`, do not specify the `'Smallest'` or `'Largest'` name-value pair arguments of `pdist2` to select or sort columns of `D`. Selecting fewer than n distances results in an error, because `dbscan` expects `D` to be a square matrix. Sorting the distances in each column of `D` leads to a loss in the interpretation of `D` and can give meaningless results when used in the `dbscan` function.
- For efficient memory usage, consider passing in `D` as a logical matrix rather than a numeric matrix to `dbscan` when `D` is large. By default, MATLAB stores each value in a numeric matrix using 8 bytes (64 bits), and each value in a logical matrix using 1 byte (8 bits).
- To select a value for `minpts`, consider a value greater than or equal to the number of dimensions of the input data plus one [1]. For example, for an n-by-p matrix `X`, set `'minpts'` equal to $p+1$ or greater.
- One possible strategy for selecting a value for `epsilon` is to generate a k-distance graph for `X`. For each point in `X`, find the distance to the kth nearest point, and plot sorted points against this distance. Generally, the graph contains a knee. The distance that corresponds to the knee is typically a good choice for `epsilon`, because it is the region where points start tailing off into outlier (noise) territory [1].

Algorithms

- DBSCAN is a density-based clustering algorithm that is designed to discover clusters and noise in data. The algorithm identifies three kinds of points: core points, border points, and noise points [1]. For specified values of `epsilon` and `minpts`, the `dbscan` function implements the algorithm as follows:

  1. From the input data set `X`, select the first unlabeled observation x_1 as the current point, and initialize the first cluster label C to 1.
  2. Find the set of points within the epsilon neighborhood `epsilon` of the current point. These points are the neighbors.

a If the number of neighbors is less than `minpts`, then label the current point as a noise point (or an outlier). Go to step 4.

Note `dbscan` can reassign noise points to clusters if the noise points later satisfy the constraints set by `epsilon` and `minpts` from some other point in `X`. This process of reassigning points happens for border points of a cluster.

b Otherwise, label the current point as a core point belonging to cluster *C*.

3 Iterate over each neighbor (new current point) and repeat step 2 until no new neighbors are found that can be labeled as belonging to the current cluster *C*.

4 Select the next unlabeled point in `X` as the current point, and increase the cluster count by 1.

5 Repeat steps 2–4 until all points in `X` are labeled.

- If two clusters have varying densities and are close to each other, that is, the distance between two border points (one from each cluster) is less than `epsilon`, then `dbscan` can merge the two clusters into one.
- Every valid cluster might not contain at least `minpts` observations. For example, `dbscan` can identify a border point belonging to two clusters that are close to each other. In such a situation, the algorithm assigns the border point to the first discovered cluster. As a result, the second cluster is still a valid cluster, but it can have fewer than `minpts` observations.

References

[1] Ester, M., H.-P. Kriegel, J. Sander, and X. Xiaowei. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In *Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining*, 226-231. Portland, OR: AAAI Press, 1996.

See Also

`clusterdata` | `kmeans` | `kmedoids` | `linkage` | `pdist` | `pdist2`

Topics

“Introduction to DBSCAN” on page 16-19

Introduced in R2019a

dcovary

D-optimal design with fixed covariates

Syntax

```
dCV = dcovary(nfactors,fixed)
[dCV,X] = dcovary(nfactors,fixed)
[dCV,X] = dcovary(nfactors,fixed,model)
[dCV,X] = daugment(...,param1,val1,param2,val2,...)
```

Description

`dCV = dcovary(nfactors,fixed)` uses a coordinate-exchange algorithm to generate a *D*-optimal design for a linear additive model with `nfactors` factors, subject to the constraint that the model include the fixed covariate factors in `fixed`. The number of runs in the design is the number of rows in `fixed`. The design `dCV` augments `fixed` with initial columns for treatments of the model terms.

`[dCV,X] = dcovary(nfactors,fixed)` also returns the design matrix `X` associated with the design.

`[dCV,X] = dcovary(nfactors,fixed,model)` uses the linear regression model specified in `model`. `model` is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

The order of the columns of `X` for a full quadratic model with *n* terms is:

1. The constant term
2. The linear terms in order 1, 2, ..., *n*
3. The interaction terms in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n* – 1, *n*)
4. The squared terms in order 1, 2, ..., *n*

Other models use a subset of these terms, in the same order.

Alternatively, `model` can be a matrix specifying polynomial terms of arbitrary order. In this case, `model` should have one column for each factor and one row for each term in the model. The entries in any row of `model` are powers for the factors in the columns. For example, if a model has factors `X1`, `X2`, and `X3`, then a row `[0 1 2]` in `model` specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in `model` specifies a constant term, which can be omitted.

`[dCV,X] = daugment(...,param1,val1,param2,val2,...)` specifies additional parameter/value pairs for the design. Valid parameters and their values are listed in the following table.

Parameter	Value
`'bounds'`	Lower and upper bounds for each factor, specified as a 2-by-`nfactors` matrix. Alternatively, this value can be a cell array containing `nfactors` elements, each element specifying the vector of allowable values for the corresponding factor.
`'categorical'`	Indices of categorical predictors.
`'display'`	Either `'on'` or `'off'` to control display of the iteration counter. The default is `'on'`.
`'excludefun'`	Handle to a function that excludes undesirable runs. If the function is *f*, it must support the syntax $b = f(S)$, where S is a matrix of treatments with `nfactors` columns and b is a vector of Boolean values with the same number of rows as S. $b(i)$ is true if the ith row S should be excluded.
`'init'`	Initial design as an `mruns`-by-`nfactors` matrix. The default is a randomly selected set of points.
`'levels'`	Vector of number of levels for each factor.
`'maxiter'`	Maximum number of iterations. The default is `10`.

`'options'`	The value is a structure that contains options specifying whether to compute multiple tries in parallel, and specifying how to use random numbers when generating the starting points for the tries. Create the options structure with `statset`. Applicable `statset` parameters are:
	• `'UseParallel'` — If `true` and if Parallel Computing Toolbox is installed, compute in parallel. If the Parallel Computing Toolbox is not installed, then computation occurs in serial mode. Default is `false`, meaning serial computation.
	• `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. Default is `false`. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
	• `Streams` — A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `dcovary` uses the default stream or streams. If you choose to specify `Streams`, use a single object except in the case
	  • `UseParallel` is `true`
	  • `UseSubstreams` is `false`
	  In that case, use a cell array the same size as the Parallel pool.

`'tries'`	Number of times to try to generate a design from a new starting point. The algorithm uses random points for each try, except possibly the first. The default is `1`.

Examples

Example 1

Suppose you want a design to estimate the parameters in a three-factor linear additive model, with eight runs that necessarily occur at different times. If the process experiences temporal linear drift, you may want to include the run time as a variable in the model. Produce the design as follows:

```
time = linspace(-1,1,8)';
[dCV1,X] = dcovary(3,time,'linear')
dCV1 =
   -1.0000    1.0000    1.0000   -1.0000
    1.0000   -1.0000   -1.0000   -0.7143
   -1.0000   -1.0000   -1.0000   -0.4286
    1.0000   -1.0000    1.0000   -0.1429
    1.0000    1.0000   -1.0000    0.1429
   -1.0000    1.0000   -1.0000    0.4286
    1.0000    1.0000    1.0000    0.7143
   -1.0000   -1.0000    1.0000    1.0000
X =
    1.0000   -1.0000    1.0000    1.0000   -1.0000
    1.0000    1.0000   -1.0000   -1.0000   -0.7143
    1.0000   -1.0000   -1.0000   -1.0000   -0.4286
    1.0000    1.0000   -1.0000    1.0000   -0.1429
    1.0000    1.0000    1.0000   -1.0000    0.1429
    1.0000   -1.0000    1.0000   -1.0000    0.4286
    1.0000    1.0000    1.0000    1.0000    0.7143
    1.0000   -1.0000   -1.0000    1.0000    1.0000
```

The column vector `time` is a fixed factor, normalized to values between `±1`. The number of rows in the fixed factor specifies the number of runs in the design. The resulting design `dCV` gives factor settings for the three controlled model factors at each time.

Example 2

The following example uses the `dummyvar` function to block an eight-run experiment into 4 blocks of size 2 for estimating a linear additive model with two factors:

```
fixed = dummyvar([1 1 2 2 3 3 4 4]);
dCV2 = dcovary(2,fixed(:,1:3),'linear')
dCV2 =
   1   1   1   0   0
  -1  -1   1   0   0
  -1   1   0   1   0
   1  -1   0   1   0
   1   1   0   0   1
  -1  -1   0   0   1
  -1   1   0   0   0
   1  -1   0   0   0
```

The first two columns of `dCV2` contain the settings for the two factors; the last three columns are dummy variable codings for the four blocks.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

cordexch | daugment | dummyvar

Introduced before R2006a

delete

Class: `qrandstream`

Delete handle object

Syntax

```
delete(h)
```

Description

`delete(h)` deletes the handle object `h`, where `h` is a scalar handle. The `delete` method deletes a handle object but does not clear the handle from the workspace. A deleted handle is no longer valid.

See Also

`clear` | `isvalid` | `qrandstream`

dendrogram

Dendrogram plot

Syntax

```
dendrogram(tree)
dendrogram(tree,Name,Value)

dendrogram(tree,P)
dendrogram(tree,P,Name,Value)

H = dendrogram( ___ )
[H,T,outperm] = dendrogram( ___ )
```

Description

`dendrogram(tree)` generates a dendrogram plot of the hierarchical binary cluster tree. A dendrogram consists of many *U*-shaped lines that connect data points in a hierarchical tree. The height of each *U* represents the distance between the two data points being connected.

- If there are 30 or fewer data points in the original data set, then each leaf in the dendrogram corresponds to one data point.
- If there are more than 30 data points, then `dendrogram` collapses lower branches so that there are 30 leaf nodes. As a result, some leaves in the plot correspond to more than one data point.

`dendrogram(tree,Name,Value)` uses additional options specified by one or more name-value pair arguments.

`dendrogram(tree,P)` generates a dendrogram plot with no more than `P` leaf nodes. If there are more than `P` data points in the original data set, then `dendrogram` collapses the lower branches of the tree. As a result, some leaves in the plot correspond to more than one data point.

`dendrogram(tree,P,Name,Value)` uses additional options specified by one or more name-value pair arguments.

`H = dendrogram( ___ )` generates a dendrogram plot and returns a vector of line handles. You can use any of the input arguments from the previous syntaxes.

`[H,T,outperm] = dendrogram( ___ )` also returns a vector containing the leaf node number for each object in the original data set, `T`, and a vector giving the order of the node labels of the leaves as shown in the dendrogram, `outperm`.

- It is useful to return `T` when the number of leaf nodes, `P`, is less than the total number of data points, so that some leaf nodes in the display correspond to multiple data points.
- The order of the node labels given in `outperm` is from left to right for a horizontal dendrogram, and from bottom to top for a vertical dendrogram.

Examples

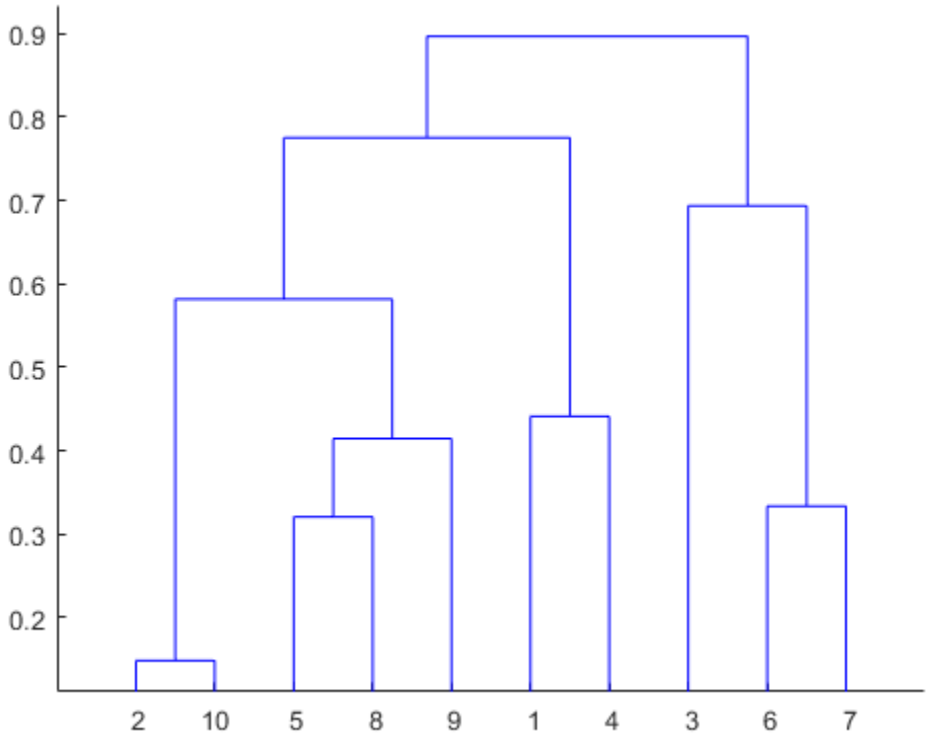
Plot Dendrogram

Generate sample data.

```
rng('default') % For reproducibility
X = rand(10,3);
```

Create a hierarchical binary cluster tree using `linkage`. Then, plot the dendrogram using the default options.

```
tree = linkage(X,'average');

figure()
dendrogram(tree)
```



Specify Dendrogram Leaf Node Order

Generate sample data.

```
rng('default') % For reproducibility
X = rand(10,3);
```

Create a hierarchical binary cluster tree using `linkage`.

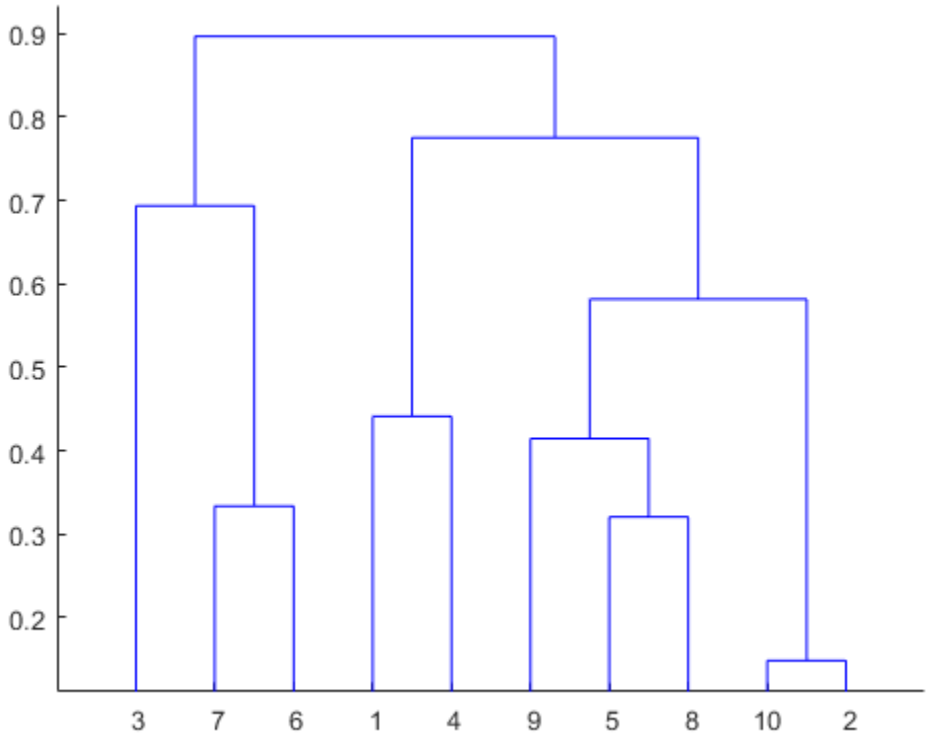
```
tree = linkage(X,'average');
```

```
D = pdist(X);
leafOrder = optimalleaforder(tree,D)
```

```
leafOrder = 1×10

     3     7     6     1     4     9     5     8    10     2
```

Plot the dendrogram using an optimal leaf order.

```
figure()
dendrogram(tree,'Reorder',leafOrder)
```



The order of the leaf nodes in the dendrogram plot corresponds - from left to right - to the permutation in `leafOrder`.

Specify Number of Nodes in Dendrogram Plot

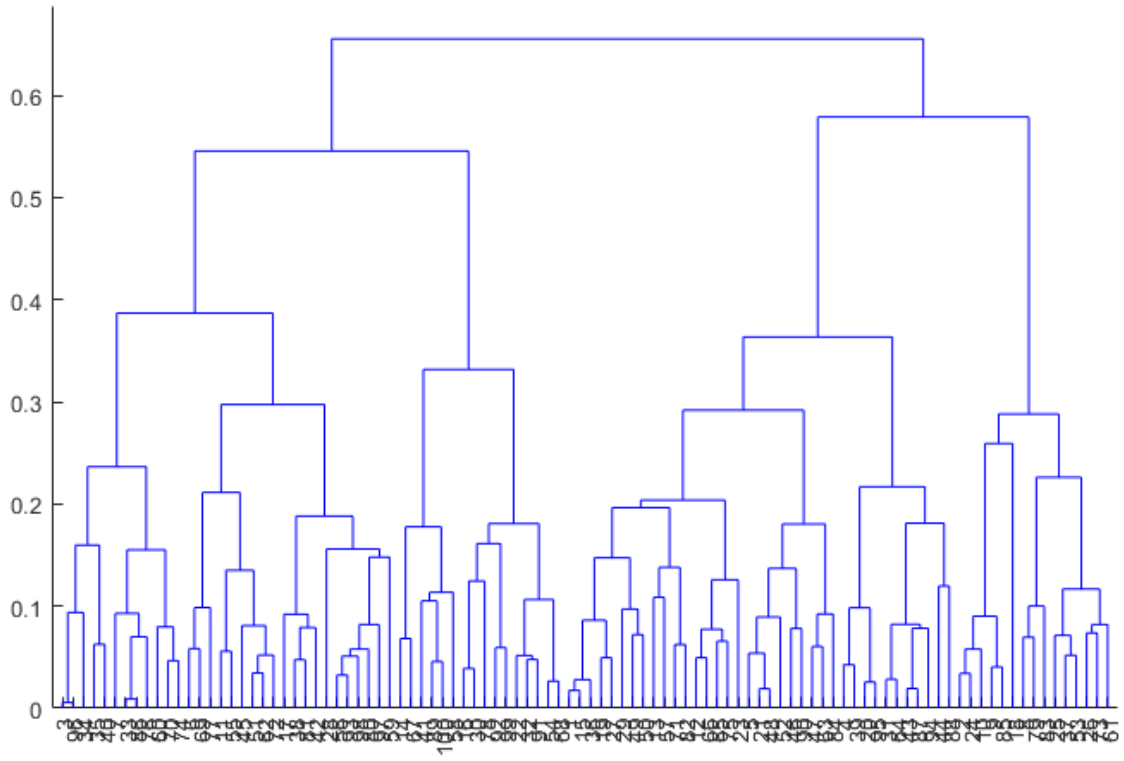
Generate sample data.

```
rng('default') % For reproducibility
X = rand(100,2);
```

There are 100 data points in the original data set, `X`.

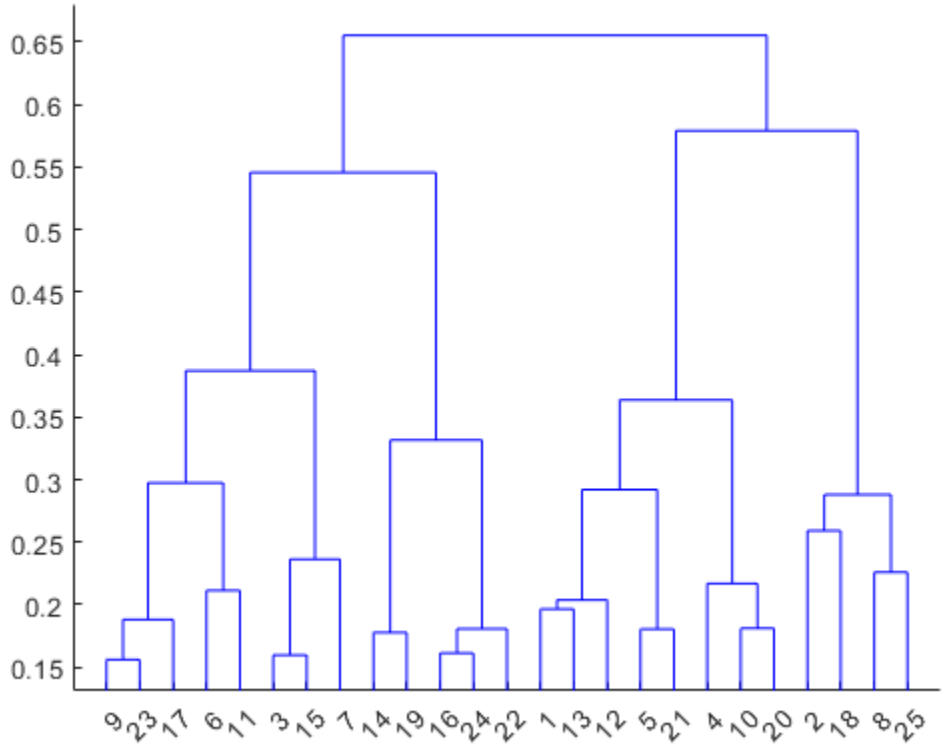
Create a hierarchical binary cluster tree using `linkage`. Then, plot the dendrogram for the complete tree (100 leaf nodes) by setting the input argument `P` equal to `0`.

```
tree = linkage(X,'average');
dendrogram(tree,0)
```



Now, plot the dendrogram with only 25 leaf nodes. Return the mapping of the original data points to the leaf nodes shown in the plot.

```
figure
[~,T] = dendrogram(tree,25);
```



List the original data points that are in leaf node 7 of the dendrogram plot.

```
find(T==7)
```

```
ans = 7×1

     7
    33
    60
    70
    74
    76
    86
```

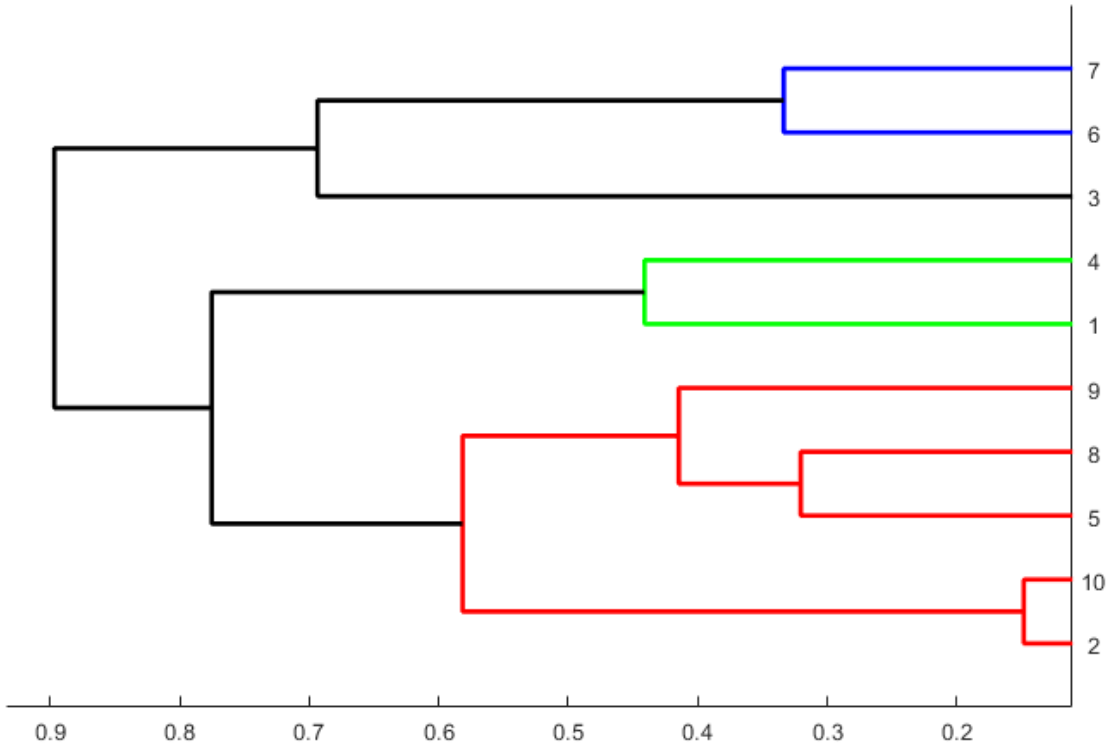
Change Dendrogram Orientation and Line Width

Generate sample data.

```
rng('default') % For reproducibility
X = rand(10,3);
```

Create a hierarchical binary cluster tree using `linkage`. Then, plot the dendrogram with a vertical orientation, using the default color threshold. Return handles to the lines so you can change the dendrogram line widths.

```
tree = linkage(X,'average');
H = dendrogram(tree,'Orientation','left','ColorThreshold','default');
set(H,'LineWidth',2)
```



Input Arguments

`tree` — Hierarchical binary cluster tree

matrix returned by `linkage`

Hierarchical binary cluster tree, specified as an (M – 1)-by-3 matrix that you generate using `linkage`, where M is the number of data points in the original data set.

`P` — Maximum number of leaf nodes

`30` (default) | positive integer value

Maximum number of leaf nodes to include in the dendrogram plot, specified as a positive integer value.

- If there are `P` or fewer data points in the original data set, then each leaf in the dendrogram corresponds to one data point.
- If there are more than `P` data points, then `dendrogram` collapses lower branches so that there are `P` leaf nodes. As a result, some leaves in the plot correspond to more than one data point.

If you do not specify `P`, then `dendrogram` uses 30 as the maximum number of leaf nodes. To display the complete tree, set `P` equal to `0`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Orientation','left','Reorder',myOrder` specifies a vertical dendrogram with leaves in the order specified by `myOrder`.

`Reorder` — Order of leaf nodes

vector

Order of leaf nodes in the dendrogram plot, specified as the comma-separated pair consisting of `'Reorder'` and a vector giving the order of nodes in the complete tree. The order vector must be a permutation of the vector `1:M`, where *M* is the number of data points in the original data set. Specify the order from left to right for horizontal dendrograms, and from bottom to top for vertical dendrograms.

If *M* is greater than the number of leaf nodes in the dendrogram plot, `P` (by default, `P` is 30), then you can only specify a permutation vector that does not separate the groups of leaves that correspond to collapsed nodes.

Data Types: `single` | `double`

`CheckCrossing` — Indicator for whether to check for crossing branches

`true` (default) | `false`

Indicator for whether to check for crossing branches in the dendrogram plot, specified as the comma-separated pair consisting of `'CheckCrossing'` and either `true` or `false`. This option is only useful when you specify a value for `Reorder`.

When `CheckCrossing` has the value `true`, `dendrogram` issues a warning if the order of the leaf nodes causes crossing branches in the plot. If the dendrogram plot does not show a complete tree (because the number of data points in the original data set is greater than `P`), `dendrogram` only issues a warning when the order of the leaf nodes causes branch to cross in the dendrogram as shown in the plot. That is, there is no warning if the order causes crossing branches in the complete tree but not in the dendrogram as shown in the plot.

Data Types: `logical`

`ColorThreshold` — Threshold for unique colors

`'default'` | scalar value in the range `(0,max(tree(:,3)))`

Threshold for unique colors in the dendrogram plot, specified as the comma-separated pair consisting of `'ColorThreshold'` and either `'default'` or a scalar value in the range `(0,max(tree(:,3)))`. If `ColorThreshold` has the value *T*, then `dendrogram` assigns a unique color to each group of nodes in the dendrogram whose linkage is less than *T*.

- If `ColorThreshold` has the value `'default'`, then the threshold, *T*, is 70% of the maximum linkage, `0.7*max(tree(:,3))`.

- If you do not specify a value for `ColorThreshold`, or if you specify a threshold outside the range `(0,max(tree(:,3)))`, then `dendrogram` uses only one color for the dendrogram plot.

`Orientation` — Orientation of dendrogram
`'top'` (default) | `'bottom'` | `'left'` | `'right'`

Orientation of the dendrogram in the figure window, specified as the comma-separated pair consisting of `'Orientation'` and one of these values:

`'top'`	Top to bottom
`'bottom'`	Bottom to top
`'left'`	Left to right
`'right'`	Right to left

`Labels` — Label for each data point
character array | string array | cell array of character vectors

Label for each data point in the original data set, specified as the comma-separated pair consisting of `'Labels'` and a character array, string array or cell array of character vectors. `dendrogram` labels any leaves in the dendrogram plot containing a single data point with that data point's label.

Output Arguments

`H` — Handles to lines
vector

Handles to lines in the dendrogram plot, returned as a vector.

`T` — Leaf node numbers
column vector

Leaf node numbers for each data point in the original data set, returned as a column vector of length *M*, where *M* is the number of data points in the original data set.

When there are fewer than `P` data points in the original data (`P` is 30, by default), all data points are displayed in the dendrogram, with each node containing a single data point. In this case, `T` is the identity map, `T = (1:M)'`.

`T` is useful when `P` is less than the total number of data points. That is, when some leaf nodes in the dendrogram display correspond to multiple data points. For example, to find out which data points are contained in leaf node `k` of the dendrogram plot, use `find(T==k)`.

`outperm` — Permutation of node labels
vector

Permutation of the node labels of the leaves of the dendrogram as shown in the plot, returned as a row vector. `outperm` gives the order from left to right for a horizontal dendrogram, and from bottom to top for a vertical dendrogram. If there are `P` leaves in the dendrogram plot, `outperm` is a permutation of the vector `1:P`.

See Also
`cluster` | `clusterdata` | `cophenet` | `inconsistent` | `linkage` | `pdist` | `silhouette`

Introduced before R2006a

describe

Describe generated features

Syntax

```
describe(Transformer)
describe(Transformer,Index)
Info = describe( ___ )
```

Description

`describe(Transformer)` prints the description of the features generated by `Transformer`. Create the `FeatureTransformer` object `Transformer` by using the `gencfeatures` function.

`describe(Transformer,Index)` prints the description of the features identified by `Index`.

`Info = describe( ___ )` returns the feature descriptions in a table. Row names of `Info` correspond to the names of the features.

Examples

Generate and Inspect Features

Generate features from a table of predictor data by using `gencfeatures`. Inspect the generated features by using the `describe` object function.

Read power outage data into the workspace as a table. Remove observations with missing values, and display the first few rows of the table.

```
outages = readtable("outages.csv");
Tbl = rmmissing(outages);
head(Tbl)
```

```
ans=8×6 table
       Region           OutageTime        Loss     Customers     RestorationTime            Cause
    ____________    ________________    ______    __________    ________________    _____________

    {'SouthWest'}    2002-02-01 12:18    458.98    1.8202e+06    2002-02-07 16:50    {'winter sto
    {'SouthEast'}    2003-02-07 21:15     289.4    1.4294e+05    2003-02-17 08:14    {'winter sto
    {'West'     }    2004-04-06 05:44    434.81    3.4037e+05    2004-04-06 06:10    {'equipment
    {'MidWest'  }    2002-03-16 06:18    186.44    2.1275e+05    2002-03-18 23:23    {'severe sto
    {'West'     }    2003-06-18 02:49         0             0    2003-06-18 10:54    {'attack'
    {'NorthEast'}    2003-07-16 16:23    239.93         49434    2003-07-17 01:12    {'fire'
    {'MidWest'  }    2004-09-27 11:09    286.72         66104    2004-09-27 16:37    {'equipment
    {'SouthEast'}    2004-09-05 17:48    73.387         36073    2004-09-05 20:46    {'equipment
```

Some of the variables, such as `OutageTime` and `RestorationTime`, have data types that are not supported by classifier training functions like `fitcensemble`.

Generate 25 features from the predictors in `Tbl` that can be used to train a bagged ensemble. Specify the `Region` table variable as the response.

```
Transformer = gencfeatures(Tbl,"Region",25,"TargetLearner","bag")
```

```
Transformer =
  FeatureTransformer with properties:

                     Type: 'classification'
            TargetLearner: 'bag'
    NumEngineeredFeatures: 22
      NumOriginalFeatures: 3
         TotalNumFeatures: 25
```

The `Transformer` object contains the information about the generated features and the transformations used to create them.

To better understand the generated features, use the `describe` object function.

```
Info = describe(Transformer)
```

```
Info=25×4 table
                                 Type        IsOriginal          InputVariables                  
                              ___________    __________    ___________________________________    

    Loss                      Numeric          true        Loss                                "" 
    Customers                 Numeric          true        Customers                           "" 
    c(Cause)                  Categorical      true        Cause                               "Va
    RestorationTime-OutageTime Numeric         false       OutageTime, RestorationTime         "El
    sdn(OutageTime)           Numeric          false       OutageTime                          "Se
    woe3(c(Cause))            Numeric          false       Cause                               "Va
    doy(OutageTime)           Numeric          false       OutageTime                          "Da
    year(OutageTime)          Numeric          false       OutageTime                          "Ye
    kmd1                      Numeric          false       Loss, Customers                     "Eu
    kmd5                      Numeric          false       Loss, Customers                     "Eu
    quarter(OutageTime)       Numeric          false       OutageTime                          "Qu
    woe2(c(Cause))            Numeric          false       Cause                               "Va
    year(RestorationTime)     Numeric          false       RestorationTime                     "Ye
    month(OutageTime)         Numeric          false       OutageTime                          "Mo
    Loss.*Customers           Numeric          false       Loss, Customers                     "Lo
    tods(OutageTime)          Numeric          false       OutageTime                          "Ti
      ⋮
```

The `Info` table indicates the following:

- The first three generated features are original to `Tbl`, although the software converts the original `Cause` variable to a categorical variable `c(Cause)`.
- The `OutageTime` and `RestorationTime` variables are not included as generated features because they are `datetime` variables, which cannot be used to train a bagged ensemble model. However, the software derives many of the generated features from these variables, such as the fourth feature `RestorationTime-OutageTime`.
- Some generated features are a combination of multiple transformations. For example, the software generates the sixth feature `woe3(c(Cause))` by converting the `Cause` variable to a categorical variable and then calculating the Weight of Evidence values for the resulting variable.

Input Arguments

`Transformer` — Feature transformer

`FeatureTransformer` object

Feature transformer, specified as a `FeatureTransformer` object.

`Index` — Features to describe

numeric vector | logical vector | string array | cell array of character vectors

Features to describe, specified as a numeric or logical vector indicating the position of the features, or a string array or cell array of character vectors indicating the names of the features.

Example: `1:12`

Data Types: `single` | `double` | `logical` | `string` | `cell`

Output Arguments

`Info` — Feature descriptions

table

Feature descriptions, returned as a table. Each row corresponds to a generated feature, and each column provides the following information.

Column Name	Description
`Type`	Indicates the data type of the feature, either `numeric` or `categorical`
`IsOriginal`	Indicates whether the feature is an original feature (`true`) or an engineered feature (`false`)
`InputVariables`	Indicates the original features used to generate the feature
`Transformations`	Describes the transformations used to generate the feature, in the order they are applied — For more information, see “Feature Transformations” on page 33-1186.

Algorithms

Feature Transformations

This table provides additional information on some of the more complex feature transformation descriptions in `Info.Transformations`.

Sample Feature Name	Sample Transformation Description in `Info`	Additional Information
`eb4(Variable)`	`Equal-width binning (number of bins = 4)`	The software splits the `Variable` values into 4 bins of equal width. The resulting feature is a categorical variable.

Sample Feature Name	Sample Transformation Description in `Info`	Additional Information
`fenc(Variable)`	`Frequency encoding (number of levels = 10)`	The software calculates the frequency of the `10` categories (or levels) in `Variable`. In the resulting feature, the software replaces each categorical value with the corresponding category frequency, creating a numeric variable.
`kmc1`	`Centroid encoding (component #1) (kmeans clustering with k = 10)`	The software uses *k*-means clustering to assign each observation to one of `10` clusters. Each row in the resulting feature corresponds to an observation and is the `1`st component of the cluster centroid associated with that observation. The resulting feature is a numeric variable.
`kmd4`	`Euclidean distance to centroid 4 (kmeans clustering with k = 10)`	The software uses *k*-means clustering to assign each observation to one of `10` clusters. Each row in the resulting feature is the Euclidean distance from the corresponding observation to the centroid of the `4`th cluster. The resulting feature is a numeric variable.
`kmi`	`Cluster index encoding (kmeans clustering with k = 10)`	The software uses *k*-means clustering to assign each observation to one of `10` clusters. Each row in the resulting feature is the cluster index for the corresponding observation. The resulting feature is a categorical variable.
`q50(Variable)`	`Equiprobable binning (number of bins = 50)`	The software splits the `Variable` values into `50` bins of equal probability. The resulting feature is a categorical variable.
`woe5(Variable)`	`Weight of Evidence (positive class = Class5)`	The software performs the following steps to create the resulting feature: • Calculate how many total observations have `Class5` as a response (a) and how many have a different response (b). • For each category in `Variable`, determine how many observations in that category have `Class5` as a response (c) and how many have a different response (d). • For each category, compute the Weight of Evidence (WoE) as $\ln\left(\frac{(c+0.5)/a}{(d+0.5)/b}\right)$. • Replace each categorical value with the corresponding WoE, creating a numeric variable.

See Also

`FeatureTransformer` | `gencfeatures` | `transform`

Topics

“Automated Feature Engineering for Classification” on page 18-190

Introduced in R2021a

designecoc

Coding matrix for reducing error-correcting output code to binary

Syntax

```
M = designecoc(K,name)
M = designecoc(K,name,Name,Value)
```

Description

`M = designecoc(K,name)` returns the coding matrix `M` that reduces the error-correcting output code (ECOC) design specified by `name` and `K` classes to a binary problem. `M` has `K` rows and L columns, with each row corresponding to a class and each column corresponding to a binary learner. `name` and `K` determine the value of L.

You can view or customize `M`, and then specify it as the coding matrix for training an ECOC multiclass classifier using `fitcecoc`.

`M = designecoc(K,name,Name,Value)` returns the coding matrix with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the number of trials when generating a dense or sparse, random coding matrix.

Examples

Train ECOC Classifiers Using a Custom Coding Design

Consider the `arrhythmia` data set. There are 16 classes in the study, 13 of which are represented in the data. The first class indicates that the subject did not have arrhythmia, and the last class indicates that the subject's arrhythmia state was not recorded. Suppose that the other classes are ordinal levels indicating the severity of arrhythmia. Train an ECOC classifier using a custom coding design specified by the description of the classes.

Load the `arrhythmia` data set.

```
load arrhythmia
K = 13; % Number of distinct classes
```

Construct a coding matrix that describes the nature of the classes.

```
OrdMat = designecoc(11,'ordinal');
nOM = size(OrdMat);
class1VSOrd = [1; -ones(11,1); 0];
class1VSClass16 = [1; zeros(11,1); -1];
OrdVSClass16 = [0; ones(11,1); -1];
Coding = [class1VSOrd class1VSClass16 OrdVSClass16,...
    [zeros(1,nOM(2)); OrdMat; zeros(1,nOM(2))]]
```

```
Coding = 13×13
```

```
     1     1     0     0     0     0     0     0     0     0     0     0     0
    -1     0     1    -1    -1    -1    -1    -1    -1    -1    -1    -1    -1
    -1     0     1     1    -1    -1    -1    -1    -1    -1    -1    -1    -1
    -1     0     1     1     1    -1    -1    -1    -1    -1    -1    -1    -1
    -1     0     1     1     1     1    -1    -1    -1    -1    -1    -1    -1
    -1     0     1     1     1     1     1    -1    -1    -1    -1    -1    -1
    -1     0     1     1     1     1     1     1    -1    -1    -1    -1    -1
    -1     0     1     1     1     1     1     1     1    -1    -1    -1    -1
    -1     0     1     1     1     1     1     1     1     1    -1    -1    -1
    -1     0     1     1     1     1     1     1     1     1     1    -1    -1
      ⋮
```

Train an ECOC classifier using the custom coding design `Coding` and specify that the binary learners are decision trees.

```
Mdl = fitcecoc(X,Y,'Coding',Coding,'Learner','Tree');
```

Estimate the in-sample classification error.

```
genErr = resubLoss(Mdl)
```

```
genErr = 0.1460
```

Choose Among Several Random Coding Designs

If you request a random coding matrix by specifying `sparserandom` or `denserandom`, then, by default, `designecoc` generates 10,000 random matrices. Then, it chooses the matrix with the largest, minimal, pair-wise row distances based on the Hamming measure. You can specify to generate more matrices to increase the chance of obtaining a better one, or you can generate several coding matrices, and then see which performs best.

Load the `arrhythmia` data set. Reserve the observations classified into class 16 (i.e., those that do not have an arrhythmia classification) as new data.

```
load arrhythmia
oosIdx = Y == 16;
isIdx = ~oosIdx;
Y = categorical(Y(isIdx));
tabulate(Y)
```

```
  Value    Count   Percent
      1      245     56.98%
      2       44     10.23%
      3       15      3.49%
      4       15      3.49%
      5       13      3.02%
      6       25      5.81%
      7        3      0.70%
      8        2      0.47%
      9        9      2.09%
     10       50     11.63%
     14        4      0.93%
     15        5      1.16%
```

```
K = numel(unique(Y));
```

Generate four random coding design matrices such that the first two are dense and the second two are sparse. Specify to find the best out of 20,000 variates.

```
rng(1); % For reproducibility

Coding = cell(4,1); % Preallocate for coding matrices
CodingTypes = {'denserandom','denserandom','sparserandom','sparserandom'};
for j = 1:4;
    Coding{j} = designecoc(K,CodingTypes{j},'NumTrials',2e4);
end
```

`Coding` is a 4-by-1 cell array, where each cell is a coding design matrix. The matrices have `K` rows, but the number of columns (i.e., binary learners) might vary.

Train and cross validate ECOC classifiers using the 15-fold cross validation. Specify that each ECOC classifier be trained using a classification tree, and the random coding matrix stored in `Coding`.

```
Mdl = cell(4,1); % Preallocate for the ECOC classifiers
for j = 1:4;
    Mdl{j} = fitcecoc(X(isIdx,:),Y,'Learners','tree',...
        'Coding',Coding{j},'KFold',15);
end
```

```
Warning: One or more of the unique class values in GROUP is not present in one or more folds. Fo

Warning: One or more of the unique class values in GROUP is not present in one or more folds. Fo

Warning: One or more of the unique class values in GROUP is not present in one or more folds. Fo

Warning: One or more of the unique class values in GROUP is not present in one or more folds. Fo
```

`Mdl` is a 4-by-1 cell array of `ClassificationPartitionedECOC` models. Several classes have low relative frequency in the data, and so there is a chance that, during cross validation, some in-sample folds will not train using observations from those classes.

Estimate the 15-fold classification error for each classifier.

```
genErr = nan(4,1);
for j = 1:4;
    genErr(j) = kfoldLoss(Mdl{j});
end

genErr
```

```
genErr = 4×1

    0.2256
    0.2116
    0.2186
    0.2186
```

Though the generalization error is still high, the best performing model, based solely on the out-of-sample classification error, is the model that used the coding design `Coding{3}`.

You can try to improve the generalization error by tuning some parameters of the binary learners. For example, you can specify to use the twoing rule or deviance for the split criterion, rather than the default Gini's diversity index. You might also specify to use surrogate splits since there are missing values in the data.

Input Arguments

K — Number of classes

positive integer

Number of classes, specified as a positive integer.

K specifies the number of rows of the coding matrix M.

Data Types: single | double

name — Coding design name

'binarycomplete' | 'denserandom' | 'onevsall' | 'onevsone' | 'sparserandom' | ...

Coding design name, specified as a value in the following table. The table summarizes the coding schemes.

Value	Number of Binary Learners	Description
'allpairs' and 'onevsone'	$K(K-1)/2$	For each binary learner, one class is positive, another is negative, and the software ignores the rest. This design exhausts all combinations of class pair assignments.
'binarycomplete'	$2^{(K-1)}-1$	This design partitions the classes into all binary combinations, and does not ignore any classes. For each binary learner, all class assignments are -1 and 1 with at least one positive and negative class in the assignment.
'denserandom'	Random, but approximately $10\log_2 K$	For each binary learner, the software randomly assigns classes into positive or negative classes, with at least one of each type. For more details, see “Random Coding Design Matrices” on page 33-1646.
'onevsall'	K	For each binary learner, one class is positive and the rest are negative. This design exhausts all combinations of positive class assignments.

Value	Number of Binary Learners	Description
`'ordinal'`	K – 1	For the first binary learner, the first class is negative, and the rest positive. For the second binary learner, the first two classes are negative, the rest positive, and so on.
`'sparserandom'`	Random, but approximately 15 $\log_2 K$	For each binary learner, the software randomly assigns classes as positive or negative with probability 0.25 for each, and ignores classes with probability 0.5. For more details, see “Random Coding Design Matrices” on page 33-1646.
`'ternarycomplete'`	$\left(3^K - 2^{(K+1)} + 1\right)/2$	This design partitions the classes into all ternary combinations. All class assignments are `0`, `-1`, and `1` with at least one positive and one negative class in the assignment.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumTrials',1000` specifies to generate `1000` random matrices.

`NumTrials` — Number of random coding matrices to generate

`10000` (default) | positive integer

Number of random coding matrices to generate, specified as the comma-separated pair consisting of `'NumTrials'` and a positive integer.

The software:

- Generates `NumTrials` matrices, and selects the one with the maximal, pair-wise row distance.
- Ignores `NumTrials` for all values of `name` except `'denserandom'` and `'sparserandom'`.

Example: `'NumTrials',1000`

Data Types: `single` | `double`

Output Arguments

`M` — Coding matrix

numeric matrix

Coding matrix that reduces an ECOC scheme to binary, returned as a numeric matrix. `M` has `K` rows and L columns, where L is the number of binary learners. Each row corresponds to a class and each column corresponds to a binary learner.

The elements of `M` are `-1`, `0`, or `1`, and the value corresponds to a dichotomous class assignment. This table describes the meaning of `M(i,j)`, that is, the class that learner `j` assigns to observations in class `i`.

Value	Dichotomous Class Assignment
–1	Learner `j` assigns observations in class `i` to a negative class.
0	Before training, learner `j` removes observations in class `i` from the data set.
1	Learner `j` assigns observations in class `i` to a positive class.

The binary learners for designs `denserandom`, `binarycomplete`, and `onevsall` do not assign `0` to observations in any class.

Tips

- The number of binary learners grows with the number of classes. For a problem with many classes, the `binarycomplete` and `ternarycomplete` coding designs are not efficient. However:
  - If $K \leq 4$, then use `ternarycomplete` coding design rather than `sparserandom`.
  - If $K \leq 5$, then use `binarycomplete` coding design rather than `denserandom`.

  You can display the coding design matrix of a trained ECOC classifier by entering `Mdl.CodingMatrix` into the Command Window.
- You should form a coding matrix using intimate knowledge of the application, and taking into account computational constraints. If you have sufficient computational power and time, then try several coding matrices and choose the one with the best performance (e.g., check the confusion matrices for each model using `confusionchart`).
- Leave-one-out cross-validation (`Leaveout`) is inefficient for data sets with many observations. Instead, use k-fold cross-validation (`KFold`).

Algorithms

Custom Coding Design Matrices

Custom coding matrices must have a certain form. The software validates custom coding matrices by ensuring:

- Every element is -1, 0, or 1.
- Every column contains as least one -1 and one 1.
- For all distinct column vectors u and v, $u \neq v$ and $u \neq -v$.
- All rows vectors are unique.
- The matrix can separate any two classes. That is, you can travel from any row to any other row following these rules:

- You can move vertically from 1 to -1 or -1 to 1.
- You can move horizontally from a nonzero element to another nonzero element.
- You can use a column of the matrix for a vertical move only once.

If it is not possible to move from row i to row j using these rules, then classes i and j cannot be separated by the design. For example, in the coding design

$$\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$$

classes 1 and 2 cannot be separated from classes 3 and 4 (that is, you cannot move horizontally from the -1 in row 2 to column 2 since there is a 0 in that position). Therefore, the software rejects this coding design.

Random Coding Design Matrices

For a given number of classes K, the software generates random coding design matrices as follows.

1. The software generates one of these matrices:
   a. Dense random — The software assigns 1 or –1 with equal probability to each element of the K-by-L_d coding design matrix, where $L_d \approx \lceil 10\log_2 K \rceil$.
   b. Sparse random — The software assigns 1 to each element of the K-by-L_s coding design matrix with probability 0.25, –1 with probability 0.25, and 0 with probability 0.5, where $L_s \approx \lceil 15\log_2 K \rceil$.
2. If a column does not contain at least one 1 and at least one –1, then the software removes that column.
3. For distinct columns u and v, if $u = v$ or $u = -v$, then the software removes v from the coding design matrix.

The software randomly generates 10,000 matrices by default, and retains the matrix with the largest, minimal, pairwise row distance based on the Hamming measure ([4]) given by

$$\Delta(k_1, k_2) = 0.5 \sum_{l=1}^{L} \left|m_{k_1 l}\right| \left|m_{k_2 l}\right| \left|m_{k_1 l} - m_{k_2 l}\right|,$$

where $m_{k_j l}$ is an element of coding design matrix j.

References

[1] Fürnkranz, Johannes. “Round Robin Classification.” *J. Mach. Learn. Res.*, Vol. 2, 2002, pp. 721–747.

[2] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recog. Lett.*, Vol. 30, Issue 3, 2009, pp. 285–297.

See Also

`ClassificationECOC` | `fitcecoc`

Introduced in R2014b

devianceTest

Package:

Analysis of deviance for generalized linear regression model

Syntax

```
tbl = devianceTest(mdl)
```

Description

`tbl = devianceTest(mdl)` returns an analysis of deviance table for the generalized linear regression model `mdl`. The table `tbl` gives the result of a test that determines whether the model `mdl` fits significantly better than a constant model.

Examples

Perform Deviance Test

Perform a deviance test on a generalized linear regression model.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Test whether the model differs from a constant in a statistically significant way.

```
tbl = devianceTest(mdl)
```

```
tbl=2×4 table
                            Deviance     DFE     chi2Stat     pValue
                           __________    ___    __________    ______

    log(y) ~ 1             2.9544e+05    99
    log(y) ~ 1 + x1 + x2         107.4    97     2.9533e+05       0
```

The small *p*-value indicates that the model significantly differs from a constant. Note that the model display of `mdl` includes the statistics shown in the second row of the table.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

Output Arguments

`tbl` — Analysis of deviance summary statistics

table

Analysis of deviance summary statistics, returned as a table.

`tbl` contains analysis of deviance statistics for both a constant model and the model `mdl`. The table includes these columns for each model.

Column	Description
`Deviance`	Deviance is twice the difference between the loglikelihoods of the corresponding model (`mdl` or constant) and the saturated model. For more information, see Deviance on page 33-1199.
`DFE`	Degrees of freedom for the error (residuals), equal to $n - p$, where n is the number of observations, and p is the number of estimated coefficients
`chi2Stat`	*F*-statistic or chi-squared statistic, depending on whether the dispersion is estimated (*F*-statistic) or not (chi-squared statistic) • *F*-statistic is the difference between the deviance of the constant model and the deviance of the full model, divided by the estimated dispersion. • Chi-squared statistic is the difference between the deviance of the constant model and the deviance of the full model.

Column	Description
`pValue`	*p*-value associated with the test: chi-squared statistic with $p - 1$ degrees of freedom, or *F*-statistic with $p - 1$ numerator degrees of freedom and `DFE` denominator degrees of freedom, where *p* is the number of estimated coefficients

More About

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to a saturated model.

Deviance of a model M_1 is twice the difference between the loglikelihood of the model M_1 and the saturated model M_s. A saturated model is a model with the maximum number of parameters that you can estimate.

For example, if you have n observations (y_i, $i = 1, 2, \ldots, n$) with potentially different values for $X_i^T\beta$, then you can define a saturated model with n parameters. Let $L(b,y)$ denote the maximum value of the likelihood function for a model with the parameters b. Then the deviance of the model M_1 is

$$-2(\log L(b_1, y) - \log L(b_S, y)),$$

where b_1 and b_s contain the estimated parameters for the model M_1 and the saturated model, respectively. The deviance has a chi-square distribution with $n - p$ degrees of freedom, where n is the number of parameters in the saturated model and p is the number of parameters in the model M_1.

Assume you have two different generalized linear regression models M_1 and M_2, and M_1 has a subset of the terms in M_2. You can assess the fit of the models by comparing the deviances D_1 and D_2 of the two models. The difference of the deviances is

$$\begin{aligned} D = D_2 - D_1 &= -2(\log L(b_2, y) - \log L(b_S, y)) + 2(\log L(b_1, y) - \log L(b_S, y)) \\ &= -2(\log L(b_2, y) - \log L(b_1, y)). \end{aligned}$$

Asymptotically, the difference D has a chi-square distribution with degrees of freedom v equal to the difference in the number of parameters estimated in M_1 and M_2. You can obtain the *p*-value for this test by using `1 - chi2cdf(D,v)`.

Typically, you examine D using a model M_2 with a constant term and no predictors. Therefore, D has a chi-square distribution with $p - 1$ degrees of freedom. If the dispersion is estimated, the difference divided by the estimated dispersion has an F distribution with $p - 1$ numerator degrees of freedom and $n - p$ denominator degrees of freedom.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `coefTest`

Topics

“Generalized Linear Models” on page 12-9

Introduced in R2012a

designMatrix

Class: `GeneralizedLinearMixedModel`

Fixed- and random-effects design matrices

Syntax

```
D = designMatrix(glme)
D = designMatrix(glme,'Fixed')

D = designMatrix(glme,'Random')
Dsub = designMatrix(glme,'Random',gnumbers)
[Dsub,gnames] = designMatrix(glme,'Random',gnumbers)
```

Description

`D = designMatrix(glme)` or `D = designMatrix(glme,'Fixed')` returns the fixed-effects design matrix for the generalized linear mixed-effects model `glme`.

`D = designMatrix(glme,'Random')` returns the random-effects design matrix for the generalized linear mixed-effects model `glme`.

`Dsub = designMatrix(glme,'Random',gnumbers)` returns a subset of the random-effects design matrix for the generalized linear mixed-effects model `glme` that corresponds to the grouping variables indicated by `gnumbers`.

`[Dsub,gnames] = designMatrix(glme,'Random',gnumbers)` also returns the grouping variable names that correspond to `gnumbers`.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`gnumbers` — Grouping variable numbers

array of integer values

Grouping variable numbers, specified as an array of integer values containing elements in the range [1,*R*], where *R* is the length of the cell array that contains the grouping variables for the generalized linear mixed-effects model `glme`.

For example, you can specify the grouping variables g_1, g_3, and g_r as `[1,3,r]`.

Data Types: `single` | `double`

Output Arguments

D — Design matrix

matrix

Design matrix of a generalized linear mixed-effects model `glme` returned as one of the following:

- Fixed-effects design matrix — *n*-by-*p* matrix consisting of the fixed-effects design matrix of `glme`, where *n* is the number of observations and *p* is the number of fixed-effects terms. The order of fixed-effects terms in `D` matches the order of terms in the `CoefficientNames` property of the `GeneralizedLinearMixedModel` object `glme`.
- Random-effects design matrix — *n*-by-*k* matrix, consisting of the random-effects design matrix of `glme`. Here, *k* is equal to `length(B)`, where `B` is the random-effects coefficients vector of generalized linear mixed-effects model `glme`. The random-effects design matrix is returned as a sparse matrix. For more information, see “Sparse Matrices”.

  If `glme` has *R* grouping variables g_1, g_2, ..., g_R, with levels m_1, m_2, ..., m_R, respectively, and if q_1, q_2, ..., q_R are the lengths of the random-effects vectors that are associated with g_1, g_2, ..., g_R, respectively, then `B` is a column vector of length $q_1*m_1 + q_2*m_2 + ... + q_R*m_R$.

  `B` is made by concatenating the empirical Bayes predictors of random effects vectors corresponding to each level of each grouping variable as $[g_1level_1; g_1level_2; ...; g_1level_{m_1}; g_2level_1; g_2level_2; ...; g_2level_{m_2}; ...; g_Rlevel_1; g_Rlevel_2; ...; g_Rlevel_{m_R}]'$.

Data Types: `single` | `double`

Dsub — Submatrix of random-effects design matrix

matrix

Submatrix of random-effects design matrix that corresponds to the grouping variables specified by `gnumbers`, returned as an *n*-by-*k* matrix, where *k* is length of the column vector `Bsub`.

`Bsub` contains the concatenated empirical Bayes predictors of random-effects vectors, corresponding to each level of the grouping variables, specified by `gnumbers`.

If, for example, `gnumbers` is `[1,3,r]`, this corresponds to the grouping variables g_1, g_3, and g_r. Then, `Bsub` contains the empirical Bayes predictors of random-effects vectors corresponding to each level of the grouping variables g_1, g_3, and g_r, such as

$[g_1level_1; g_1level_2; ...; g_1level_{m_1}; g_3level_1; g_3level_2; ...; g_3level_{m_3}; g_rlevel_1; g_rlevel_2; ...; g_rlevel_{m_r}]'$.

Thus, `Dsub*Bsub` represents the contribution of all random effects corresponding to grouping variables g_1, g_3, and g_r to the response of `glme`.

If `gnumbers` is empty, then `Dsub` is the full random-effects design matrix.

Data Types: `single` | `double`

gnames — Names of grouping variables

k-by-1 cell array

Names of grouping variables corresponding to the integers in `gnumbers` if the design type is `'Random'`, returned as a *k*-by-1 cell array. If the design type is `'Fixed'`, then `gnames` is an empty matrix `[]`.

Data Types: `cell`

Examples

Obtain Fixed- and Random-Effects Design Matrices

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij}).$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.

- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess$_{ij}$, time_dev$_{ij}$, and temp_dev$_{ij}$ are the measurements for each variable that correspond to factory i during batch j. For example, newprocess$_{ij}$ indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C$_{ij}$ and supplier_B$_{ij}$ are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Extract the fixed-effects design matrix and display rows 1 through 10.

```
Dfe = designMatrix(glme,'Fixed');
disp(Dfe(1:10,:))
```

```
    1.0000         0    0.1834    0.2259    1.0000         0
    1.0000         0    0.3035    0.0725         0    1.0000
    1.0000         0    0.0717    0.1630    1.0000         0
    1.0000         0    0.1069    0.0809   -1.0000   -1.0000
    1.0000         0    0.0241    0.0319    1.0000         0
    1.0000         0    0.1214    0.1114         0    1.0000
    1.0000         0    0.0033    0.0553    1.0000         0
    1.0000         0    0.2350    0.0616    1.0000         0
    1.0000         0    0.0488    0.0177         0    1.0000
    1.0000         0    0.1148    0.0105    1.0000         0
```

Column 1 of the fixed-effects design matrix `Dfe` contains the constant term. Column 2, 3, and 4 contain the `newprocess`, `time_dev`, and `temp_dev` terms, respectively. Columns 5 and 6 contain dummy variables for `supplier_C` and `supplier_B`, respectively.

Extract the random-effects design matrix and display rows 1 through 10.

```
Dre = designMatrix(glme,'Random');
disp(Dre(1:10,:))
```

```
   (1,1)        1
   (2,1)        1
   (3,1)        1
   (4,1)        1
   (5,1)        1
   (6,2)        1
   (7,2)        1
   (8,2)        1
   (9,2)        1
  (10,2)        1
```

Convert the sparse matrix `Dre` to a full matrix and display rows 1 through 10.

```
full(Dre(1:10,:))
```

```
ans = 10×20
```

```
     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0
     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0
     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0
     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0
     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0
```

Each column corresponds to a level of the grouping variable `factory`.

See Also

`GeneralizedLinearMixedModel` | `fitglme` | `fitted` | `residuals` | `response`

designMatrix

Class: `LinearMixedModel`

Fixed- and random-effects design matrices

Syntax

```
D = designMatrix(lme)
D = designMatrix(lme,'Fixed')

D = designMatrix(lme,'Random')
Dsub = designMatrix(lme,'Random',gnumbers)
[Dsub,gnames] = designMatrix(lme,'Random',gnumbers)
```

Description

`D = designMatrix(lme)` or `D = designMatrix(lme,'Fixed')` returns the fixed-effects design matrix for the linear mixed-effects model `lme`.

`D = designMatrix(lme,'Random')` returns the random-effects design matrix for the linear mixed-effects model `lme`.

`Dsub = designMatrix(lme,'Random',gnumbers)` returns a subset of the random-effects design matrix for the linear mixed-effects model `lme`corresponding to the grouping variables indicated by the integers in `gnumbers`.

`[Dsub,gnames] = designMatrix(lme,'Random',gnumbers)` also returns the grouping variable names corresponding to the integers in `gnumbers`.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`gnumbers` — Grouping variable numbers

integer array

Grouping variable numbers, specified as an integer array, where R is the length of the cell array that contains the grouping variables for the linear mixed-effects model `lme`.

For example, you can specify the grouping variables g_1, g_3, and g_r as follows.

Example: `[1,3,r]`

Data Types: `double` | `single`

Output Arguments

`D` — Design matrix

matrix

Design matrix of a linear mixed-effects model `lme` returned as one of the following:

- Fixed-effects design matrix — *n*-by-*p* matrix consisting of the fixed-effects design of `lme`, where *n* is the number of observations and *p* is the number of fixed-effects terms. The order of fixed-effects terms in `D` matches the order of terms in the `CoefficientNames` property of the `LinearMixedModel` object `lme`.
- Random-effects design matrix — *n*-by-*k* matrix, consisting of the random-effects design matrix of `lme`. Here, *k* is equal to `length(B)`, where `B` is the random-effects coefficients vector of linear mixed-effects model `lme`.

  If `lme` has *R* grouping variables g_1, g_2, ..., g_R, with levels m_1, m_2, ..., m_R, respectively, and if q_1, q_2, ..., q_R are the lengths of the random-effects vectors that are associated with g_1, g_2, ..., g_R, respectively, then `B` is a column vector of length $q_1*m_1 + q_2*m_2 + ... + q_R*m_R$.

  `B` is made by concatenating the best linear unbiased predictors of random-effects vectors corresponding to each level of each grouping variable as `[`g_1`level`$_1$`; `g_1`level`$_2$`; ...; `g_1`level`$_{m_1}$`; `g_2`level`$_1$`; `g_2`level`$_2$`; ...; `g_2`level`$_{m_2}$`; ...; `g_R`level`$_1$`; `g_R`level`$_2$`; ...; `g_R`level`$_{m_R}$`]'`.

Data Types: `single` | `double`

`Dsub` — Submatrix of random-effects design matrix

matrix

Submatrix of random-effects design matrix corresponding to the grouping variables indicated by the integers in `gnumbers`, returned as an *n*-by-*k* matrix, where *k* is length of the column vector `Bsub`.

`Bsub` contains the concatenated best linear unbiased predictors (BLUPs) of random-effects vectors, corresponding to each level of the grouping variables, specified by `gnumbers`.

If, for example, `gnumbers` is `[1,3,r]`, this corresponds to the grouping variables g_1, g_3, and g_r. Then, `Bsub` contains the concatenated BLUPs of random-effects vectors corresponding to each level of the grouping variables g_1, g_3, and g_r, such as

`[`g_1`level`$_1$`; `g_1`level`$_2$`; ...; `g_1`level`$_{m_1}$`; `g_3`level`$_1$`; `g_3`level`$_2$`; ...; `g_3`level`$_{m_3}$`; `g_r`level`$_1$`; `g_r`level`$_2$`; ...; `g_r`level`$_{m_r}$`]'`.

Thus, `Dsub*Bsub` represents the contribution of all random effects corresponding to grouping variables g_1, g_3, and g_r to the response of `lme`.

If `gnumbers` is empty, then `Dsub` is the full random-effects design matrix.

Data Types: `single` | `double`

`gnames` — Names of grouping variables

k-by-1 cell array

Names of grouping variables corresponding to the integers in `gnumbers` if the design type is `'Random'`, returned as a *k*-by-1 cell array. If the design type is `'Fixed'`, then `gnames` is an empty matrix `[]`.

Data Types: `cell`

Examples

Display Fixed- and Random-Effects Design Matrices

Load the sample data.

```
load('shift.mat');
```

The data shows the deviations from the target quality characteristic measured from the products that 5 operators manufacture during three different shifts, morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the deviation of the quality characteristics from the target value. This is simulated data.

`Shift` and `Operator` are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if performance significantly differs according to the time of the shift.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)');
```

Display the fixed-effects design matrix.

```
designMatrix(lme)
```

```
ans = 15×3

     1     1     0
     1     0     0
     1     0     1
     1     1     0
     1     0     0
     1     0     1
     1     1     0
     1     0     0
     1     0     1
     1     1     0
      ⋮
```

The column of 1s represents the constant term in the model. `fitlme` takes the evening shift as the reference group and creates two dummy variables to represent the morning and night shifts, respectively.

Display the random-effects design matrix.

```
designMatrix(lme,'random')
```

```
ans =
   (1,1)        1
   (2,1)        1
```

```
  (3,1)        1
  (4,2)        1
  (5,2)        1
  (6,2)        1
  (7,3)        1
  (8,3)        1
  (9,3)        1
 (10,4)        1
 (11,4)        1
 (12,4)        1
 (13,5)        1
 (14,5)        1
 (15,5)        1
```

The first number, `i`, in the (`i`,|j|) indices corresponds to the observation number, and|j| corresponds to the level of the grouping variable, `Operator`, i.e., the operator number.

Show the full display of the random-effects design matrix.

```
full(designMatrix(lme,'random'))
```

```
ans = 15×5

     1     0     0     0     0
     1     0     0     0     0
     1     0     0     0     0
     0     1     0     0     0
     0     1     0     0     0
     0     1     0     0     0
     0     0     1     0     0
     0     0     1     0     0
     0     0     1     0     0
     0     0     0     1     0
       ⋮
```

Each column corresponds to a level of the grouping variable, `Operator`.

Random-Effects Design Matrix of Multiple Grouping Variables

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
```

```
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Store and examine the full random-effects design matrix.

```
D = full(designMatrix(lme,'random'));
```

The first three columns of matrix `D` contain the indicator variables `fitlme` creates for the three levels (`Loamy`, `Silty`, `Sandy`, respectively) of the first grouping variable, `Soil`. The next 15 columns contain the indicator variables created for the second grouping variable, `Tomato` nested under `Soil`. These are basically the elementwise products of the dummy variables representing the levels of `Soil` (`Loamy`, `Silty`, and `Sandy`, respectively) and the levels of `Tomato` (`Cherry`, `Grape`, `Heirloom`, `Plum`, `Vine`, respectively).

Subset of the Random-Effects Design Matrix

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Compute the random-effects design matrix for the second grouping variable, and display the first 12 rows.

```
[Dsub,gname]  = designMatrix(lme,'random',2);
full(Dsub(1:12,:))
```

```
ans = 12×15
```

```
     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0
     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0
     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0
     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0
   ⋮
```

`Dsub` contains the dummy variables created for the second grouping variable, that is, tomato nested under soil. These are the elementwise products of the dummy variables representing the levels of `Soil` (`Loamy`, `Silty`, `Sandy`, respectively) and the levels of `Tomato` (`Cherry`, `Grape`, `Heirloom`, `Plum`, `Vine`, respectively).

Display the name of the grouping variable.

```
gname
```

```
gname = 1x1 cell array
    {'Soil:Tomato'}
```

See Also

`LinearMixedModel` | `fitlmematrix` | `fitted`

dfittool

Open Distribution Fitter app

Note The `distributionFitter` function was introduced in R2017a as a replacement for the `dfittool` function. Both functions continue to work to start the Distribution Fitter app.

Syntax

```
dfittool
dfittool(y)
dfittool(y,cens)
dfittool(y,cens,freq)
dfittool(y,cens,freq,dsname)
```

Description

This page contains programmatic syntax information for the Distribution Fitter app. For general usage information, see **Distribution Fitter**.

`dfittool` opens the Distribution Fitter app, or brings focus to the app if it is already open.

`dfittool(y)` opens the Distribution Fitter app populated with the data specified by the vector `y`.

`dfittool(y,cens)` uses the vector `cens` to specify whether each observation in `y` is censored.

`dfittool(y,cens,freq)` uses the vector `freq` to specify the frequency of each element of `y`.

`dfittool(y,cens,freq,dsname)` creates a data set with the name `dsname`, using the data vector, `y`, censoring indicator, `cens`, and frequency vector, `freq`.

Examples

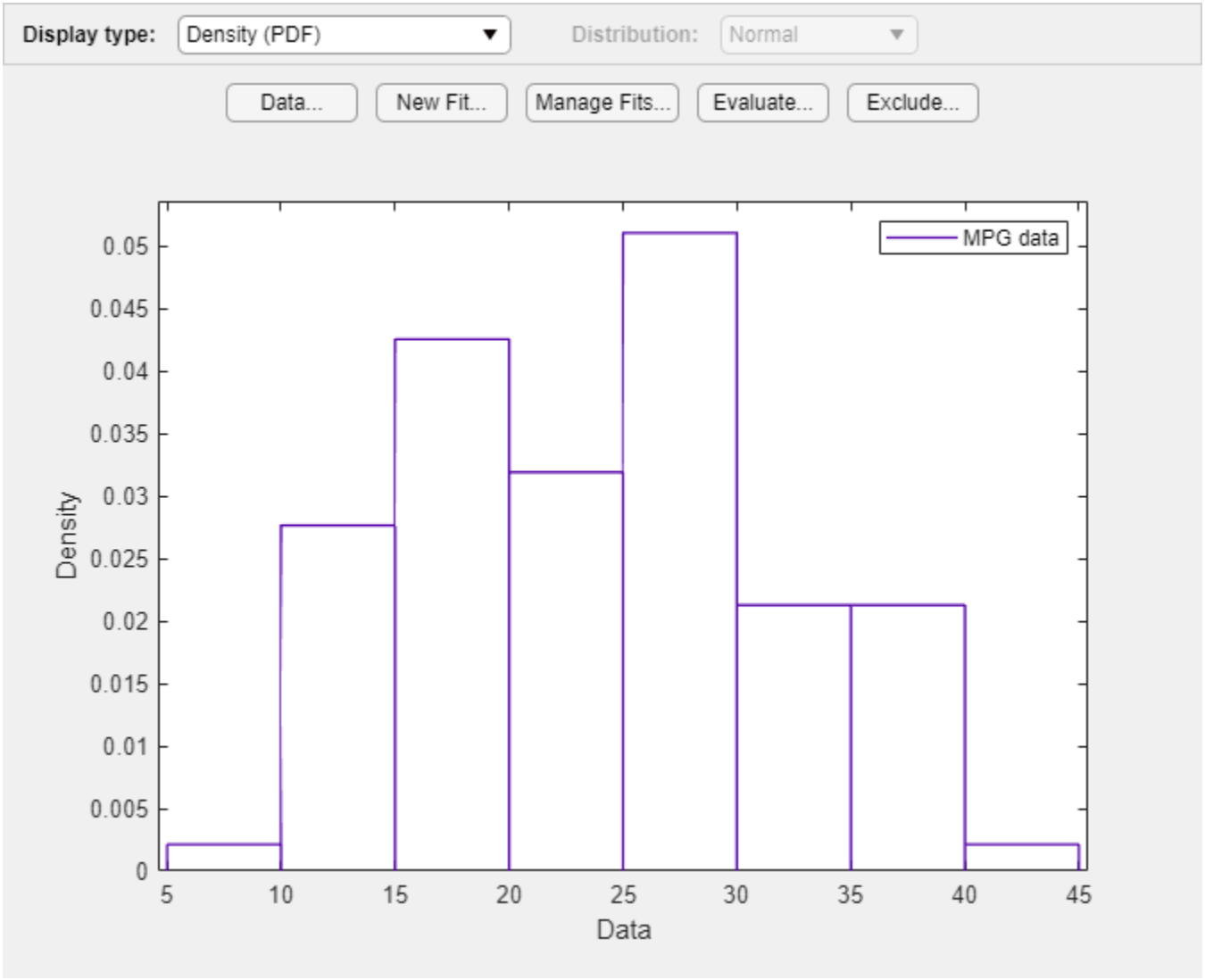
Open Distribution Fitter App with Existing Data

Load the `carsmall` sample data.

```
load carsmall
```

Open the Distribution Fitter app using the `MPG` miles per gallon data.

```
distributionFitter(MPG)
```



The Distribution Fitter app opens, populated with the `MPG` data, and displays the density (PDF) plot. You can use the app to display different plots and fit distributions to this data.

Open Distribution Fitter App with Censoring Data

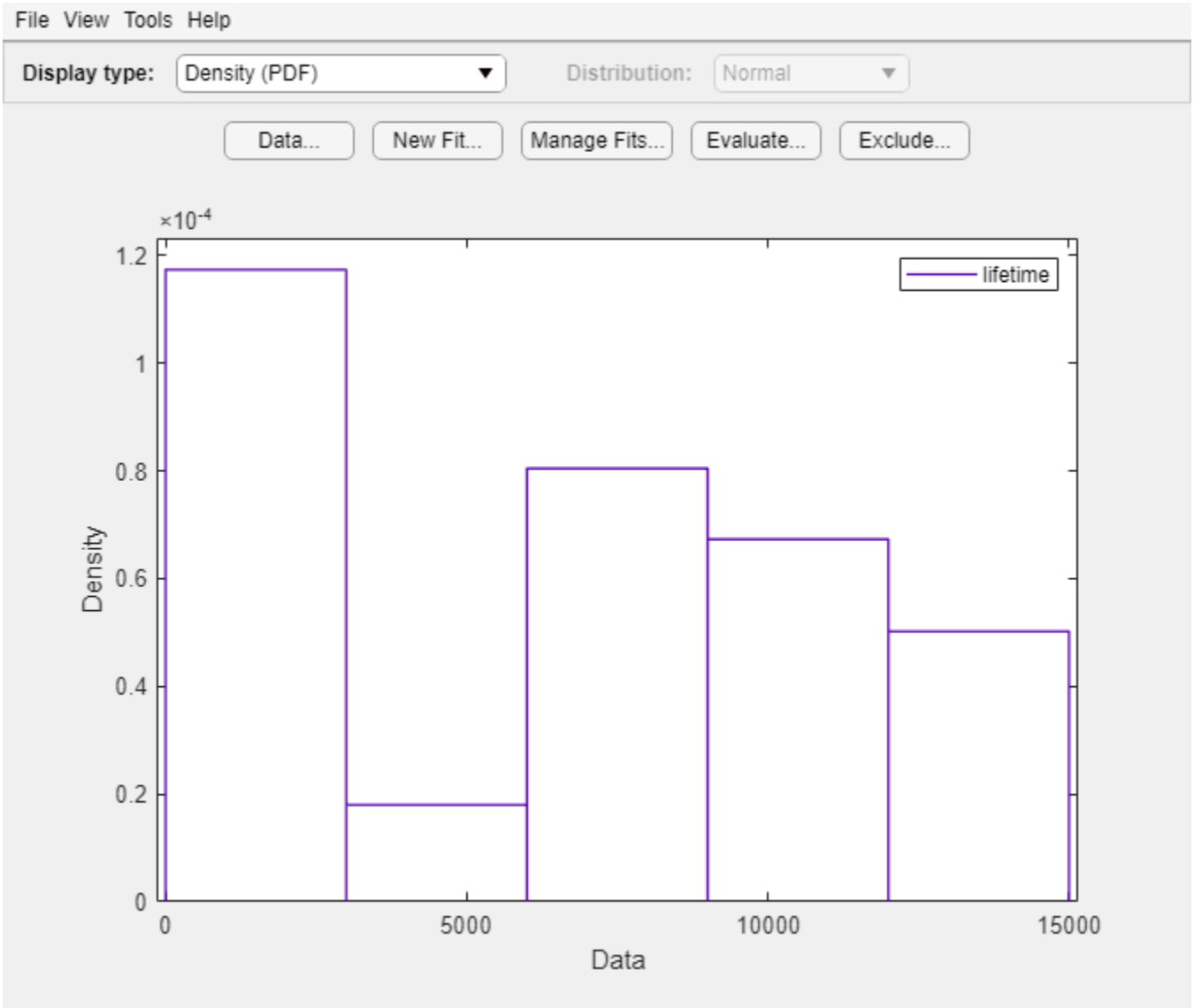
Load the sample data.

```
load lightbulb.mat
```

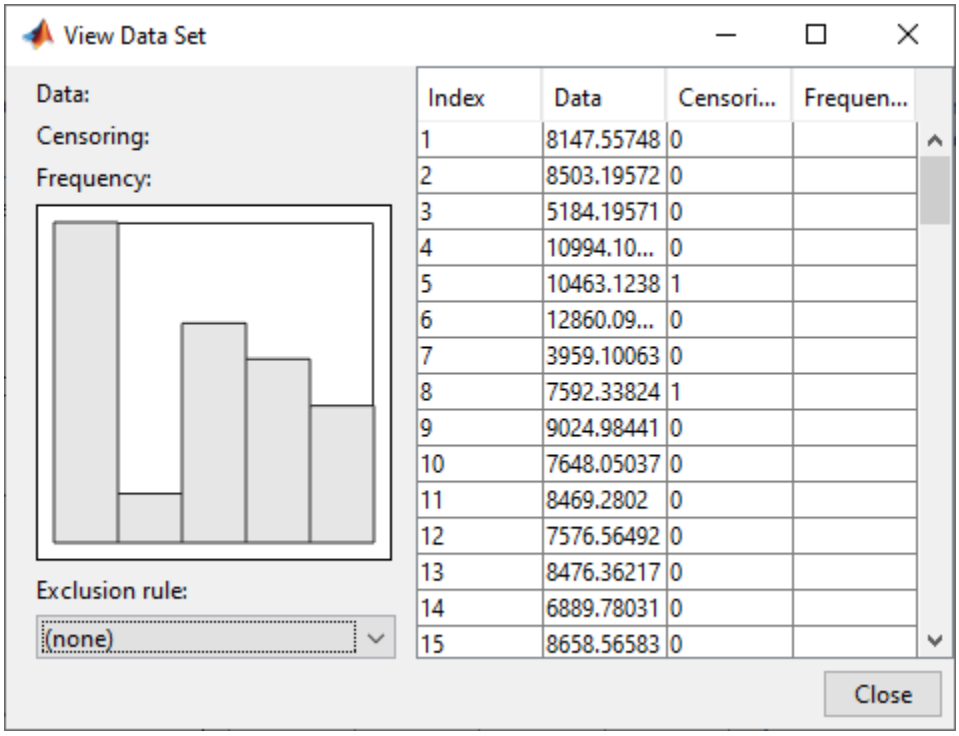
The first column of the data contains the lifetime (in hours) of two types of light bulbs. The second column contains information about the type of light bulb. 1 indicates fluorescent bulbs, and 0 indicates the incandescent bulb. The third column contains censoring information. 1 indicates censored data, and 0 indicates the exact failure time. This is simulated data.

Open the Distribution Fitter app using the first column of `lightbulb` as the input data, and the third column as the censoring data. Name the data `lifetime`.

```
distributionFitter(lightbulb(:,1),lightbulb(:,3),[],'lifetime')
```



To open the Data dialog box, click **Data**. In the **Manage data sets** pane, click to highlight the `lifetime` data set row. Finally, to open the View Data Set dialog, click **View**. The lifetime data appears in the second column and the corresponding censoring indicator appears in the third column.



Input Arguments

y — Input data

array of scalar values | variable representing an array of scalar values

Input data, specified as an array of scalar values or a variable representing an array of such values.

Data Types: `single` | `double`

cens — Censoring indicator

`zeros(n)` (default) | vector of 0 and 1 values

Censoring indicator, specified as a vector of 0 and 1 values. The length of `cens` must be equal to the length of `y`. If `y(j)` is censored, then `(cens(j)==1)`. If `y(j)` is not censored, then `(cens(j)==0)`. If `cens` is omitted or empty, then no `y` values are censored.

If you have frequency data (`freq`) but not censoring data (`cens`), then you must specify empty brackets (`[]`) for `cens`.

Data Types: `single` | `double`

freq — Frequency data

`ones(n)` (default) | vector of scalar values

Frequency data, specified as a vector of scalar values. The length of `freq` must be equal to the length of `y`. If `freq` is omitted or empty, then all `y` values have a frequency of 1.

If you have frequency data (`freq`) but not censoring data (`cens`), then you must specify empty brackets (`[]`) for `cens`.

Data Types: `single` | `double`

`dsname` — Data set name
character vector | string scalar

Data set name, specified as a character vector enclosed in single quotes or a string scalar enclosed in double quotes.

If you want to specify a data set name, but do not have censoring data (`cens`) or frequency data (`freq`), then you must specify empty brackets (`[]`) for both `freq` and `cens`.

Example: `'MyData'`

Data Types: `char` | `string`

See Also

Distribution Fitter | `distributionFitter` | `fitdist` | `makedist`

Topics

“Fit a Distribution Using the Distribution Fitter App” on page 5-71
“Model Data Using the Distribution Fitter App” on page 5-51
“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

discardSupportVectors

Package:

Discard support vectors of linear SVM binary learners in ECOC model

Syntax

```
Mdl = discardSupportVectors(MdlSV)
```

Description

`Mdl = discardSupportVectors(MdlSV)` returns a trained multiclass error-correcting output codes (ECOC) model (`Mdl`) from the trained multiclass ECOC model (`MdlSV`), which contains at least one linear `CompactClassificationSVM` binary learner. Both `Mdl` and `MdlSV` are objects of the same type, either `ClassificationECOC` objects or `CompactClassificationECOC` objects.

`Mdl` has these characteristics:

- The `Alpha`, `SupportVectors`, and `SupportVectorLabels` properties of all the linear SVM binary learners on page 33-1219 are empty (`[]`).
- If you display any linear SVM binary learners stored in the cell array of trained models `Mdl.BinaryLearners`, the software lists the `Beta` property instead of `Alpha`.

Examples

Retain and Discard Support Vectors of SVM Binary Learners

When you train an ECOC model with linear SVM binary learners, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties of the binary learners by default. You can choose instead to retain the support vectors and related values, and then discard them from the model later.

Load Fisher's iris data set.

```
load fisheriris
rng(1); % For reproducibility
```

Train an ECOC model using the entire data set. Specify retaining the support vectors by passing in the appropriate SVM template.

```
t = templateSVM('SaveSupportVectors',true);
MdlSV = fitcecoc(meas,species,'Learners',t);
```

`MdlSV` is a trained `ClassificationECOC` model with linear SVM binary learners. By default, `fitcecoc` implements a one-versus-one coding design, which requires three binary learners for three-class learning.

Access the estimated α (alpha) values using dot notation.

```
alpha = cell(3,1);
alpha{1} = MdlSV.BinaryLearners{1}.Alpha;
```

```
alpha{2} = MdlSV.BinaryLearners{2}.Alpha;
alpha{3} = MdlSV.BinaryLearners{3}.Alpha;
alpha
```

```
alpha=3×1 cell array
    { 3x1 double}
    { 3x1 double}
    {23x1 double}
```

`alpha` is a 3-by-1 cell array that stores the estimated values of α.

Discard the support vectors and related values from the ECOC model.

```
Mdl = discardSupportVectors(MdlSV);
```

`Mdl` is similar to `MdlSV`, except that the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties of all the linear SVM binary learners are empty (`[]`).

```
areAllEmpty = @(x)isempty([x.Alpha x.SupportVectors x.SupportVectorLabels]);
cellfun(areAllEmpty,Mdl.BinaryLearners)
```

```
ans = 3x1 logical array

   1
   1
   1
```

Compare the sizes of the two ECOC models.

```
vars = whos('Mdl','MdlSV');
100*(1 - vars(1).bytes/vars(2).bytes)
```

```
ans = 4.7075
```

`Mdl` is about 5% smaller than `MdlSV`.

Reduce your memory usage by compacting `Mdl` and then clearing `Mdl` and `MdlSV` from the workspace.

```
CompactMdl = compact(Mdl);
clear Mdl MdlSV;
```

Predict the label for a random row of the training data using the more efficient SVM model.

```
idx = randsample(size(meas,1),1)
```

```
idx = 63
```

```
predictedLabel = predict(CompactMdl,meas(idx,:))
```

```
predictedLabel = 1x1 cell array
    {'versicolor'}
```

```
trueLabel = species(idx)
```

```
trueLabel = 1x1 cell array
    {'versicolor'}
```

Input Arguments

`MdlSV` — Full or compact, trained multiclass ECOC model

`ClassificationECOC` model | `CompactClassificationECOC` model

Full or compact, trained multiclass ECOC model containing at least one linear SVM binary learner, specified as a `ClassificationECOC` or `CompactClassificationECOC` model.

More About

Linear SVM Binary Learner

In the context of this page, a linear support vector machine (SVM) binary learner is a binary SVM classifier created using a linear kernel function. If the `j`th binary learner in an ECOC model `Mdl` is a linear SVM binary learner, then `Mdl.BinaryLearners{j}` is a `CompactClassificationSVM` object, where `Mdl.BinaryLearners{j}.KernelParameters.Function` is `'linear'`.

Tips

- By default and for efficiency, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties for all linear SVM binary learners. `fitcecoc` lists `Beta`, rather than `Alpha`, in the model display.

  To store `Alpha`, `SupportVectorLabels`, and `SupportVectors`, pass a linear SVM template that specifies storing support vectors to `fitcecoc`. For example, enter:

  ```
  t = templateSVM('SaveSupportVectors',true)
  Mdl = fitcecoc(X,Y,'Learners',t);
  ```

  You can remove the support vectors and related values by passing the resulting `ClassificationECOC` model to `discardSupportVectors`.

Algorithms

`predict` and `resubPredict` estimate SVM scores $f(x)$ for each linear SVM binary learner in an ECOC model using

$$f(x) = x'\beta + b\,.$$

β is the `Beta` property and b is the `Bias` property of the binary learners. You can access these properties for each linear SVM binary learner in the cell array `Mdl.BinaryLearners`. For more details on the SVM score calculation, see “Support Vector Machines for Binary Classification” on page 33-1863.

See Also

`ClassificationECOC` | `ClassificationSVM` | `CompactClassificationECOC` | `discardSupportVectors` | `fitcecoc` | `fitcsvm` | `templateSVM`

Introduced in R2015a

discardSupportVectors

Package: `classreg.learning.classif`

Discard support vectors for linear support vector machine (SVM) classifier

Syntax

```
Mdl = discardSupportVectors(MdlSV)
```

Description

`Mdl = discardSupportVectors(MdlSV)` returns the trained, linear support vector machine (SVM) model `Mdl`. Both `Mdl` and the trained, linear SVM model `MdlSV` are the same type of object. That is, they both are either `ClassificationSVM` objects or `CompactClassificationSVM` objects. However, `Mdl` and `MdlSV` differ in the following ways:

- The `Alpha`, `SupportVectors`, and `SupportVectorLabels` properties are empty (`[]`) in `Mdl`.
- If you display `Mdl`, the software lists the `Beta` property instead of `Alpha`.

Examples

Discard Support Vectors

Create a linear SVM model that is more memory-efficient by discarding support vectors and other related parameters.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a linear SVM model using the entire data set.

```
MdlSV = fitcsvm(X,Y)
```

```
MdlSV =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: 'none'
         NumObservations: 351
                   Alpha: [103x1 double]
                    Bias: -3.8828
        KernelParameters: [1x1 struct]
          BoxConstraints: [351x1 double]
         ConvergenceInfo: [1x1 struct]
         IsSupportVector: [351x1 logical]
                  Solver: 'SMO'
```

```
  Properties, Methods
```

Display the number of support vectors in `MdlSV`.

```
numSV = size(MdlSV.SupportVectors,1)
```

```
numSV = 103
```

Display the number of predictor variables in `X`.

```
p = size(X,2)
```

```
p = 34
```

By default, `fitcsvm` trains a linear SVM model for two-class learning. The software lists `Alpha` in the display. The model includes 103 support vectors and 34 predictors. If you discard the support vectors, the resulting model consumes less memory.

Discard the support vectors and other related parameters.

```
Mdl = discardSupportVectors(MdlSV)
```

```
Mdl =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
                     Beta: [34x1 double]
                     Bias: -3.8828
         KernelParameters: [1x1 struct]
           BoxConstraints: [351x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [351x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

Display the coefficients in `Mdl`.

```
Mdl.Alpha
```

```
ans =

     []
```

Display the support vectors in `Mdl`.

```
Mdl.SupportVectors
```

```
ans =

     []
```

Display the support vector class labels in `Mdl`.

```
Mdl.SupportVectorLabels

ans =

     []
```

The software lists `Beta` in the display instead of `Alpha`. The `Alpha`, `SupportVectors`, and `SupportVectorLabels` properties are empty.

Compare the sizes of the models.

```
vars = whos('MdlSV','Mdl');
100*(1 - vars(1).bytes/vars(2).bytes)

ans = 20.5503
```

`Mdl` is about 20% smaller than `MdlSV`.

Remove `MdlSV` from the workspace.

```
clear MdlSV
```

Reduce Memory Consumption of SVM Models

Compact an SVM model by discarding the stored support vectors and other related estimates. Predict the label for a row of the training data by using the compacted model.

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM model using the default options.

```
MdlSV = fitcsvm(X,Y);
```

`MdlSV` is a `ClassificationSVM` model containing nonempty values for its `Alpha`, `SupportVectors`, and `SupportVectorLabels` properties.

Reduce the size of the SVM model by discarding the training data, support vectors, and related estimates.

```
CMdlSV = compact(MdlSV);               % Discard training data
CMdl = discardSupportVectors(CMdlSV);  % Discard support vectors
```

`CMdl` is a `CompactClassificationSVM` model.

Compare the sizes of the SVM models `MdlSV` and `CMdl`.

```
vars = whos('MdlSV','CMdl');
100*(1 - vars(1).bytes/vars(2).bytes)

ans = 96.8174
```

The compacted model `CMdl` consumes much less memory than the full model.

Predict the label for a random row of the training data by using `CMdl`. The `predict` function accepts compacted SVM models, and, for linear SVM models, does not require the `Alpha`, `SupportVectors`, and `SupportVectorLabels` properties to predict labels for new observations.

```
idx = randsample(size(X,1),1)
```

```
idx = 147
```

```
predictedLabel = predict(CMdl,X(idx,:))
```

```
predictedLabel = 1x1 cell array
    {'b'}
```

```
trueLabel = Y(idx)
```

```
trueLabel = 1x1 cell array
    {'b'}
```

Input Arguments

`MdlSV` — Trained, linear SVM model

`ClassificationSVM` model | `CompactClassificationSVM` model

Trained, linear SVM model, specified as a `ClassificationSVM` or `CompactClassificationSVM` model.

If the field `MdlSV.KernelParameters.Function` is not `'linear'` (that is, `MdlSV` is not a linear SVM model), the software returns an error.

Tips

- For a trained, linear SVM model, the `SupportVectors` property is an n_{sv}-by-p matrix. n_{sv} is the number of support vectors (at most the training sample size) and p is the number of predictors, or features. The `Alpha` and `SupportVectorLabels` properties are vectors with n_{sv} elements. These properties can be large for complex data sets containing many observations or examples. The `Beta` property is a vector with p elements.
- If the trained SVM model has many support vectors, use `discardSupportVectors` to reduce the amount of space consumed by the trained, linear SVM model. You can display the size of the support vector matrix by entering `size(MdlSV.SupportVectors)`.

Algorithms

`predict` and `resubPredict` estimate SVM scores $f(x)$, and subsequently label and estimate posterior probabilities using

$$f(x) = x'\beta + b\,.$$

β is `Mdl.Beta` and b is `Mdl.Bias`, that is, the `Beta` and `Bias` properties of `Mdl`, respectively. For more details, see “Support Vector Machines for Binary Classification” on page 33-1863.

See Also

`ClassificationECOC` | `ClassificationSVM` | `CompactClassificationSVM` | `discardSupportVectors` | `fitcsvm` | `templateSVM`

Introduced in R2015a

discardSupportVectors

Discard support vectors

Syntax

```
mdlOut = discardSupportVectors(mdl)
```

Description

`mdlOut = discardSupportVectors(mdl)` returns the trained, linear support vector machine (SVM) regression model `mdlOut`, which is similar to the trained, linear SVM regression model `mdl`, except:

- The `Alpha` and `SupportVectors` properties are empty (`[]`).
- If you display `mdlOut`, the software lists the `Beta` property instead of the `Alpha` property.

Input Arguments

`mdl` — Trained, linear SVM regression model

`RegressionSVM` model | `CompactRegressionSVM` model

Trained, linear SVM regression model, specified as a `RegressionSVM` or `CompactRegressionSVM` model.

If you train the model using a kernel function that is not linear (i.e., if the field `mdl.KernelFunction` is something other than `'linear'`), the software returns an error. You can only discard support vectors for linear models.

Output Arguments

`mdlOut` — Trained, linear SVM regression model

`RegressionSVM` model | `CompactRegressionSVM` model

Trained, linear SVM regression model, returned as a `RegressionSVM` or `CompactRegressionSVM` model. `mdlOut` is the same type as `mdl`.

After discarding the support vectors, the properties `Alpha` and `SupportVectors` are empty (`[]`). The software lists the property `Beta` in its display, and does not list the property `Alpha`. The `predict` and `resubPredict` methods compute predicted responses using the coefficients stored in the `Beta` property.

Examples

Discard Support Vectors for SVM Regression Model

This model shows how to reduce the disk space used by a trained, linear SVM regression model by discarding the support vectors and other related parameters.

Load the `carsmall` data set. Specify `Horsepower` and `Weight` as the predictor variables (`X`), and `MPG` as the response variable (`Y`).

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
```

Train a linear SVM regression model, standardizing the data. Display the number of support vectors.

```
mdl = fitrsvm(X,Y,'Standardize',true)
numSV = size(mdl.SupportVectors,1)
```

```
mdl =

  RegressionSVM
            PredictorNames: {'x1'  'x2'}
              ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                    Alpha: [77x1 double]
                     Bias: 22.9131
          KernelParameters: [1x1 struct]
                       Mu: [109.3441 2.9625e+03]
                    Sigma: [45.3545 805.9668]
           NumObservations: 93
            BoxConstraints: [93x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [93x1 logical]
                    Solver: 'SMO'


  Properties, Methods


numSV =

    77
```

By default, `fitrsvm` trains a linear SVM regression model. The software lists `Alpha` in the display. The model has 77 support vectors.

Note that the predictor and response variables contain several `NaN` values. When training a model, `fitrsvm` will remove rows that contain `NaN` values from both the predictor and response data. As a result, the trained model uses only 93 of the 100 total observations contained in the sample data.

Discard the support vectors and other related parameters.

```
mdlOut = discardSupportVectors(mdl)
mdlOut.Alpha
mdlOut.SupportVectors
```

```
mdlOut =

  RegressionSVM
            PredictorNames: {'x1'  'x2'}
              ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
```

```
                    Beta: [2x1 double]
                    Bias: 22.9131
          KernelParameters: [1x1 struct]
                      Mu: [109.3441 2.9625e+03]
                   Sigma: [45.3545 805.9668]
           NumObservations: 93
            BoxConstraints: [93x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [93x1 logical]
                    Solver: 'SMO'


  Properties, Methods


ans =

     []


ans =

     []
```

The software lists `Beta` in the display instead of `Alpha`. The `Alpha` and `SupportVectors` properties are empty.

Compare the sizes of the models.

```
vars = whos('mdl','mdlOut');
[vars(1).bytes,vars(2).bytes]
```

```
ans =

       15004       13156
```

`mdlOut` consumes less memory than `mdl` because it does not store the support vectors.

Reduce Memory Consumption of SVM Regression Model

This example shows how to reduce the memory consumption of a full, trained SVM regression model by compacting the model and discarding the support vectors.

Load the `carsmall` sample data.

```
load carsmall
rng default  % for reproducibility
```

Train a linear SVM regression model using `Weight` as the predictor variable and `MPG` as the response variable. Standardize the data.

```
mdl = fitrsvm(Weight,MPG,'Standardize',true);
```

Note that `MPG` contains several `NaN` values. When training a model, `fitrsvm` will remove rows that contain `NaN` values from both the predictor and response data. As a result, the trained model uses only 94 of the 100 total observations contained in the sample data.

Compact the regression model to discard the training data and some information related to the training process.

```
compactMdl = compact(mdl);
```

`compactMdl` is a `CompactRegressionSVM` model that has the same parameters, support vectors, and related estimates as `mdl`, but no longer stores the training data.

Discard the support vectors and related estimates for the compacted model.

```
mdlOut = discardSupportVectors(compactMdl);
```

`mdlOut` is a `CompactRegressionSVM` model that has the same parameters as `mdl` and `compactMdl`, but no longer stores the support vectors and related estimates.

Compare the sizes of the three SVM regression models, `compactMdl`, `mdl`, and `mdlOut`.

```
vars = whos('compactMdl','mdl','mdlOut');
[vars(1).bytes,vars(2).bytes,vars(3).bytes]
```

```
ans =

        3601       13727        2305
```

The compacted model `compactMdl` consumes 3601 bytes of memory, while the full model `mdl` consumes 13727 bytes of memory. The model `mdlOut`, which also discards the support vectors, consumes 2305 bytes of memory.

Tips

For a trained, linear SVM regression model, the `SupportVectors` property is an n_{sv}-by-p matrix. n_{sv} is the number of support vectors (at most the training sample size) and p is the number of predictor variables. If any of the predictors are categorical, then p includes the number of dummy variables necessary to account for all of the categorical predictor levels. The `Alpha` property is a vector with n_{sv} elements.

The `SupportVectors` and `Alpha` properties can be large for complex data sets that contain many observations or examples. However, the `Beta` property is a vector with p elements, which may be considerably smaller. You can use a trained SVM regression model to predict response values even if you discard the support vectors because the `predict` and `resubPredict` methods use `Beta` to compute the predicted responses.

If the trained, linear SVM regression model has many support vectors, use `discardSupportVectors` to reduce the amount of disk space that the trained, linear SVM regression model consumes. You can display the size of the support vector matrix by entering `size(mdlIn.SupportVectors)`.

Algorithms

The `predict` and `resubPredict` estimate response values using the formula

$$f(x) = \left(\frac{X}{S}\right)\beta + \beta_0 ,$$

where:

- β is the Beta value, stored as `mdl.Beta`.
- β_0 is the bias value, stored as `mdl.Bias`.
- `X` is the training data.
- `S` is the kernel scale value, stored as `mdl.KernelParameters.Scale`.

In this way, the software can use the value of `mdl.Beta` to make predictions even after discarding the support vectors.

See Also

`CompactRegressionSVM` | `RegressionSVM` | `fitrsvm` | `predict` | `resubPredict`

Introduced in R2015b

disp

Class: `dataset`

(Not Recommended) Display dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
disp(ds)
```

Description

`disp(ds)` prints the dataset array `ds`, including variable names and observation names (if present), without printing the dataset name. In all other ways it's the same as leaving the semicolon off an expression.

For numeric or categorical variables that are 2-D and have three or fewer columns, `disp` prints the actual data using either short g, long g, or bank format, depending on the current command line setting. Otherwise, `disp` prints the size and type of each dataset element.

For character variables that are 2-D and 10 or fewer characters wide, `disp` prints quoted text. Otherwise, `disp` prints the size and type of each dataset element.

For cell variables that are 2-D and have three or fewer columns, `disp` prints the contents of each cell (or its size and type if too large). Otherwise, `disp` prints the size of each dataset element.

For time series variables, `disp` prints columns for both the time and the data. If the variable is 2-D and has three or fewer columns, `disp` prints the actual data Otherwise, `disp` prints the size and type of each dataset element.

For other types of variables, `disp` prints the size and type of each dataset element.

See Also

`dataset` | `display` | `format`

disp

Class: `GeneralizedLinearMixedModel`

Display generalized linear mixed-effects model

Syntax

```
disp(glme)
```

Description

`disp(glme)` displays fitted generalized linear mixed-effects model `glme`.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Examples

Display a Generalized Linear Mixed-Effects Model

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by

`factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1\text{newprocess}_{ij} + \beta_2\text{time\_dev}_{ij} + \beta_3\text{temp\_dev}_{ij} + \beta_4\text{supplier\_C}_{ij} + \beta_5\text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Display the model.

```
disp(glme)
```

```
Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35

Fixed effects coefficients (95% CIs):
    Name                   Estimate     SE          tStat       DF    pValue
```

```
    {'(Intercept)'}          1.4689     0.15988      9.1875    94    9.8194e-15
    {'newprocess' }         -0.36766    0.17755     -2.0708    94      0.041122
    {'time_dev'   }        -0.094521    0.82849    -0.11409    94       0.90941
    {'temp_dev'   }         -0.28317     0.9617    -0.29444    94       0.76907
    {'supplier_C' }        -0.071868   0.078024     -0.9211    94       0.35936
    {'supplier_B' }         0.071072    0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
     -0.72019    -0.015134
      -1.7395       1.5505
      -2.1926       1.6263
     -0.22679     0.083051
    -0.082588      0.22473

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1                Name2                Type           Estimate
    {'(Intercept)'}      {'(Intercept)'}      {'std'}        0.31381

Group: Error
    Name                       Estimate
    {'sqrt(Dispersion)'}       1
```

The `Model information` table displays the total number of observations in the sample data (100), the number of fixed- and random-effects coefficients (6 and 20, respectively), and the number of covariance parameters (1). It also indicates that the response variable has a `Poisson` distribution, the link function is `Log`, and the fit method is `Laplace`.

`Formula` indicates the model specification using Wilkinson's notation.

The `Model fit statistics` table displays statistics used to assess the goodness of fit of the model. This includes the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`) values, log likelihood (`LogLikelihood`), and deviance (`Deviance`) values.

The `Fixed effects coefficients` table indicates that `fitglme` returned 95% confidence intervals. It contains one row for each fixed-effects predictor, and each column contains statistics corresponding to that predictor. Column 1 (`Name`) contains the name of each fixed-effects coefficient, column 2 (`Estimate`) contains its estimated value, and column 3 (`SE`) contains the standard error of the coefficient. Column 4 (`tStat`) contains the *t*-statistic for a hypothesis test that the coefficient is equal to 0. Column 5 (`DF`) and column 6 (`pValue`) contain the degrees of freedom and *p*-value that correspond to the *t*-statistic, respectively. The last two columns (`Lower` and `Upper`) display the lower and upper limits, respectively, of the 95% confidence interval for each fixed-effects coefficient.

`Random effects covariance parameters` displays a table for each grouping variable (here, only `factory`), including its total number of levels (20), and the type and estimate of the covariance parameter. Here, `std` indicates that `fitglme` returns the standard deviation of the random effect associated with the factory predictor, which has an estimated value of 0.31381. It also displays a table containing the error parameter type (here, the square root of the dispersion parameter), and its estimated value of 1.

The standard display generated by `fitglme` does not provide confidence intervals for the random-effects parameters. To compute and display these values, use `covarianceParameters`.

More About

Akaike and Bayesian Information Criteria

The *Akaike information criterion* (AIC) is $AIC = -2\log L_M + 2(param)$.

$\log L_M$ depends on the method used to fit the model.

- If you use `'Laplace'` or `'ApproximateLaplace'`, then $\log L_M$ is the maximized log likelihood.
- If you use `'MPL'`, then $\log L_M$ is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration.
- If you use `'REMPL'`, then $\log L_M$ is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.

param is the total number of parameters estimated in the model. For most GLME models, *param* is equal to $nc + p + 1$, where nc is the total number of parameters in the random-effects covariance, excluding the residual variance, and p is the number of fixed-effects coefficients. However, if the dispersion parameter is fixed at 1.0 for binomial or Poisson distributions, then *param* is equal to $(nc + p)$.

The *Bayesian information criterion* (BIC) is $BIC = -2*\log L_M + \ln(n_{eff})(param)$.

$\log L_M$ depends on the method used to fit the model.

- If you use `'Laplace'` or `'ApproximateLaplace'`, then $\log L_M$ is the maximized log likelihood.
- If you use `'MPL'`, then $\log L_M$ is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration.
- If you use `'REMPL'`, then $\log L_M$ is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.

n_{eff} is the effective number of observations.

- If you use `'MPL'`, `'Laplace'`, or `'ApproximateLaplace'`, then $n_{eff} = n$, where n is the number of observations.
- If you use `'REMPL'`, then $n_{eff} = n - p$.

param is the total number of parameters estimated in the model. For most GLME models, *param* is equal to $nc + p + 1$, where nc is the total number of parameters in the random-effects covariance, excluding the residual variance, and p is the number of fixed-effects coefficients. However, if the dispersion parameter is fixed at 1.0 for binomial or Poisson distributions, then *param* is equal to $(nc + p)$.

A lower value of deviance indicates a better fit. As the value of deviance decreases, both AIC and BIC tend to decrease. Both AIC and BIC also include penalty terms based on the number of parameters estimated, p. So, when the number of parameters increase, the values of AIC and BIC tend to increase as well. When comparing different models, the model with the lowest AIC or BIC value is considered as the best fitting model.

For models fitted using `'MPL'` and `'REMPL'`, AIC and BIC are based on the log likelihood (or restricted log likelihood) of pseudo data from the final pseudo likelihood iteration. Therefore, a direct comparison of AIC and BIC values between models fitted using `'MPL'` and `'REMPL'` is not appropriate.

See Also

`GeneralizedLinearMixedModel` | `covarianceParameters` | `fitglme`

disp

Class: `LinearMixedModel`

Display linear mixed-effects model

Syntax

```
display(lme)
```

Description

`display(lme)` displays the fitted linear mixed-effects model `lme`.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Examples

Randomized Block Design

Load the sample data.

```
load('shift.mat');
```

The dataset array shows the absolute deviations from the target quality characteristic measured from the products that five operators manufacture during three shifts, morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the absolute deviation of the quality characteristics from the target value. This is simulated data.

`Shift` and `Operator` are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if performance significantly differs according to the time of the shift.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)');
```

Display the model.

```
disp(lme)
```

```
Linear mixed-effects model fit by ML
```

```
Model information:
    Number of observations              15
    Fixed effects coefficients           3
    Random effects coefficients          5
    Covariance parameters                2

Formula:
    QCDev ~ 1 + Shift + (1 | Operator)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    59.012    62.552    -24.506          49.012

Fixed effects coefficients (95% CIs):
    Name                      Estimate    SE         tStat       DF    pValue
    {'(Intercept)'  }          3.1196     0.88681      3.5178    12    0.0042407
    {'Shift_Morning'}         -0.3868     0.48344    -0.80009    12      0.43921
    {'Shift_Night'  }          1.9856     0.48344      4.1072    12    0.0014535


    Lower      Upper
     1.1874     5.0518
    -1.4401    0.66653
    0.93227     3.0389

Random effects covariance parameters (95% CIs):
Group: Operator (5 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        1.8297


    Lower      Upper
    0.94915    3.5272

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.76439     0.49315    1.1848
```

This display includes the model performance statistics, “Akaike and Bayesian Information Criteria” on page 33-1239, “Akaike and Bayesian Information Criteria” on page 33-1239, loglikelihood, and “Deviance” on page 33-1239.

The fixed-effects coefficients table includes the names and estimates of the coefficients in the first two columns. The third column `SE` shows the standard errors of the coefficients. The column `tStat` includes the *t*-statistic values that correspond to each coefficient. `DF` is the residual degrees of freedom, and the `pValue` is the *p*-value that corresponds to the corresponding *t*-statistic value. The columns `Lower` and `Upper` display the lower and upper limits of a 95% confidence interval for each fixed-effects coefficient.

The first table for the random effects shows the types and the estimates of the random effects covariance parameters, with the lower and upper limits of a 95% confidence interval for each parameter. The display also shows the name of the grouping variable, operator, and the total number of levels, 5.

The second table for the random effects shows the estimate of the observation error, with the lower and upper limits of a 95% confidence interval.

More About

Akaike and Bayesian Information Criteria

Akaike information criterion (AIC) is $AIC = -2*\log L_M + 2*(nc + p + 1)$, where $\log L_M$ is the maximized log likelihood (or maximized restricted log likelihood) of the model, and $nc + p + 1$ is the number of parameters estimated in the model. p is the number of fixed-effects coefficients, and nc is the total number of parameters in the random-effects covariance excluding the residual variance.

Bayesian information criterion (BIC) is $BIC = -2*\log L_M + \ln(n_{eff})*(nc + p + 1)$, where $\log L_M$ is the maximized log likelihood (or maximized restricted log likelihood) of the model, n_{eff} is the effective number of observations, and $(nc + p + 1)$ is the number of parameters estimated in the model.

- If the fitting method is maximum likelihood (ML), then $n_{eff} = n$, where n is the number of observations.
- If the fitting method is restricted maximum likelihood (REML), then $n_{eff} = n - p$.

A lower value of deviance indicates a better fit. As the value of deviance decreases, both AIC and BIC tend to decrease. Both AIC and BIC also include penalty terms based on the number of parameters estimated, p. So, when the number of parameters increase, the values of AIC and BIC tend to increase as well. When comparing different models, the model with the lowest AIC or BIC value is considered as the best fitting model.

Deviance

`LinearMixedModel` computes the deviance of model M as minus two times the loglikelihood of that model. Let L_M denote the maximum value of the likelihood function for model M. Then, the deviance of model M is

$$-2*\log L_M .$$

A lower value of deviance indicates a better fit. Suppose M_1 and M_2 are two different models, where M_1 is nested in M_2. Then, the fit of the models can be assessed by comparing the deviances Dev_1 and Dev_2 of these models. The difference of the deviances is

$$Dev = Dev_1 - Dev_2 = 2(\log LM_2 - \log LM_1) .$$

Usually, the asymptotic distribution of this difference has a chi-square distribution with degrees of freedom v equal to the number of parameters that are estimated in one model but fixed (typically at 0) in the other. That is, it is equal to the difference in the number of parameters estimated in M_1 and M_2. You can get the p-value for this test using `1 - chi2cdf(Dev,V)`, where $Dev = Dev_2 - Dev_1$.

However, in mixed-effects models, when some variance components fall on the boundary of the parameter space, the asymptotic distribution of this difference is more complicated. For example, consider the hypotheses

H_0: $D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix}$, D is a q-by-q symmetric positive semidefinite matrix.

H_1: D is a $(q+1)$-by-$(q+1)$ symmetric positive semidefinite matrix.

That is, H_1 states that the last row and column of D are different from zero. Here, the bigger model M_2 has $q + 1$ parameters and the smaller model M_1 has q parameters. And Dev has a 50:50 mixture of ${\chi^2}_q$ and ${\chi^2}_{(q+1)}$ distributions (Stram and Lee, 1994).

References

[1] Hox, J. *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Inc., 2002.

[2] Stram D. O. and J. W. Lee. “Variance components testing in the longitudinal mixed-effects model”. *Biometrics*, Vol. 50, 4, 1994, pp. 1171–1177.

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix`

disp

Class: `NonLinearModel`

Display nonlinear regression model

Syntax

```
disp(mdl)
```

Description

`disp(mdl)` displays the `mdl` nonlinear model at the command line.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

Examples

Display a Nonlinear Regression Model

Create and display a nonlinear regression model.

Load the `reaction` data, and specify both a model function and starting values for the iterations.

```
load reaction
modelfun = 'rate~(b1*x2-x3/b5)/(1+b2*x1+b3*x2+b4*x3)';
beta0 = [1 .05 .02 .1 2];
```

Create a model of the data.

```
mdl = fitnlm(reactants,rate,modelfun,beta0);
```

Display the model.

```
disp(mdl)
```

```
Nonlinear regression model:
    rate ~ (b1*x2 - x3/b5)/(1 + b2*x1 + b3*x2 + b4*x3)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1      1.2526     0.86701    1.4447    0.18654
    b2    0.062776    0.043561    1.4411    0.18753
    b3    0.040048    0.030885    1.2967    0.23089
    b4     0.11242    0.075157    1.4957    0.17309
```

```
    b5      1.1914     0.83671    1.4239     0.1923

Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 0.193
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 3.91e+03, p-value = 2.54e-13
```

Alternatives

Enter *mdl* at the command line to obtain a display, where *mdl* is the name of your model.

See Also

NonLinearModel

Topics

“Nonlinear Regression” on page 13-2

disp

Class: `qrandstream`

Display `qrandstream` object

Syntax

```
disp(q)
```

Description

`disp(q)` displays the quasi-random stream `q`, without printing the variable name. `disp` prints the type and number of dimensions in the stream, and follows it with the list of point set properties.

See Also

`qrandstream`

display

Class: `dataset`

(Not Recommended) Display dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
display(ds)
```

Description

`display(ds)` prints the dataset array `ds`, including variable names and observation names (if present). `dataset` calls`display` when a you do not use a semicolon to terminate a statement

For numeric or categorical variables that are 2-D and have three or fewer columns, `display` prints the actual data. Otherwise, `display` prints the size and type of each dataset element.

For character variables that are 2-D and 10 or fewer characters wide, `display` prints quoted text. Otherwise, `display` prints the size and type of each dataset element.

For cell variables that are 2-D and have three or fewer columns, `display` prints the contents of each cell (or its size and type if too large). Otherwise, `display` prints the size of each dataset element.

For time series variables, `display` prints columns for both the time and the data. If the variable is 2-D and has three or fewer columns, `display` prints the actual data. Otherwise, `display` prints the size and type of each dataset element.

For other types of variables, `display` prints the size and type of each dataset element.

See Also

`dataset` | `display` | `format`

distributionFitter

Open Distribution Fitter app

Syntax

```
distributionFitter
distributionFitter(y)
distributionFitter(y,cens)
distributionFitter(y,cens,freq)
distributionFitter(y,cens,freq,dsname)
```

Description

This page contains programmatic syntax information for the Distribution Fitter app. For general usage information, see **Distribution Fitter**.

`distributionFitter` opens the Distribution Fitter app, or brings focus to the app if it is already open.

`distributionFitter(y)` opens the Distribution Fitter app populated with the data specified by the vector `y`.

`distributionFitter(y,cens)` uses the vector `cens` to specify whether each observation in `y` is censored.

`distributionFitter(y,cens,freq)` uses the vector `freq` to specify the frequency of each element of `y`.

`distributionFitter(y,cens,freq,dsname)` creates a data set with the name `dsname`, using the data vector, `y`, censoring indicator, `cens`, and frequency vector, `freq`.

Examples

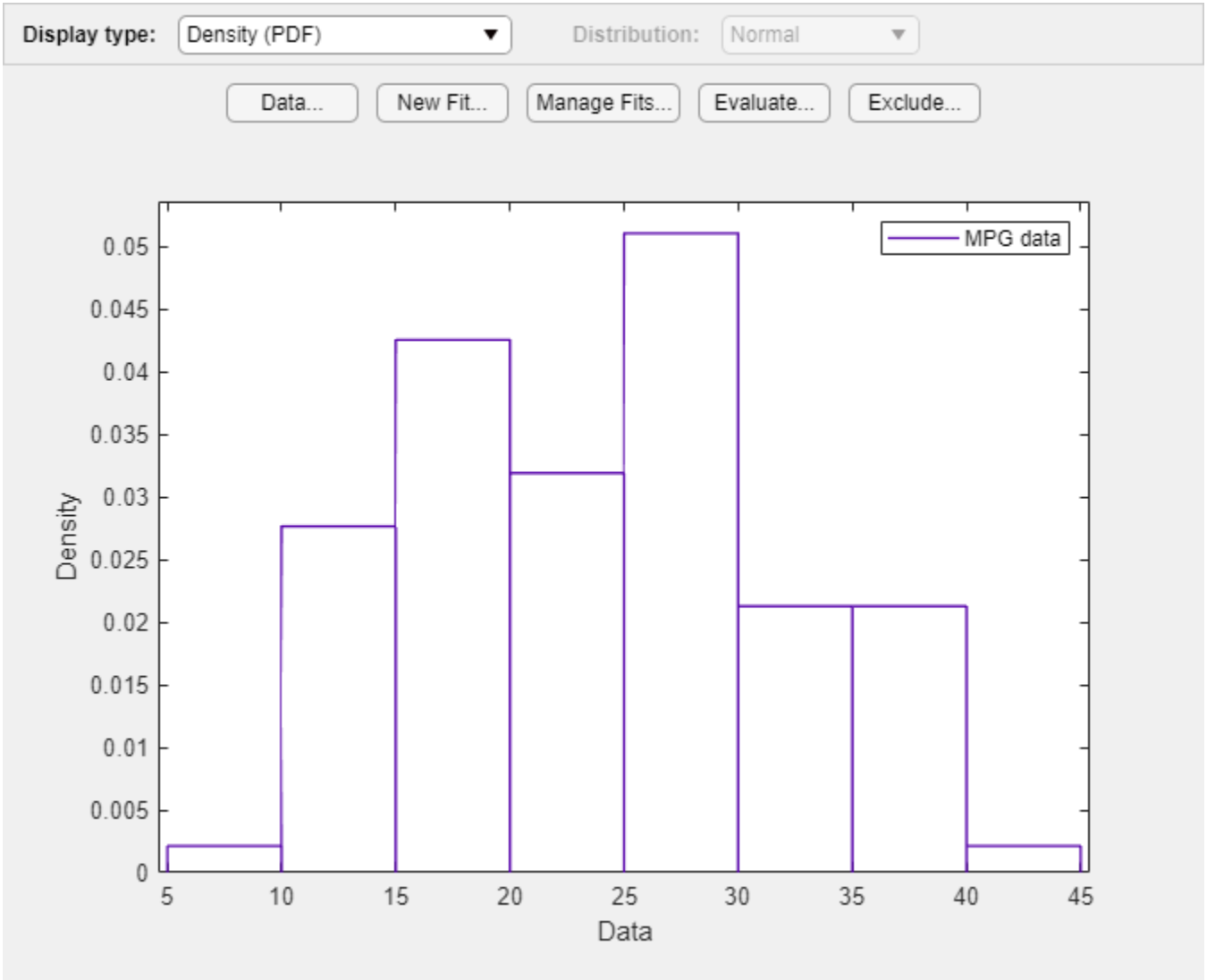
Open Distribution Fitter App with Existing Data

Load the `carsmall` sample data.

```
load carsmall
```

Open the Distribution Fitter app using the `MPG` miles per gallon data.

```
distributionFitter(MPG)
```



The Distribution Fitter app opens, populated with the `MPG` data, and displays the density (PDF) plot. You can use the app to display different plots and fit distributions to this data.

Open Distribution Fitter App with Censoring Data

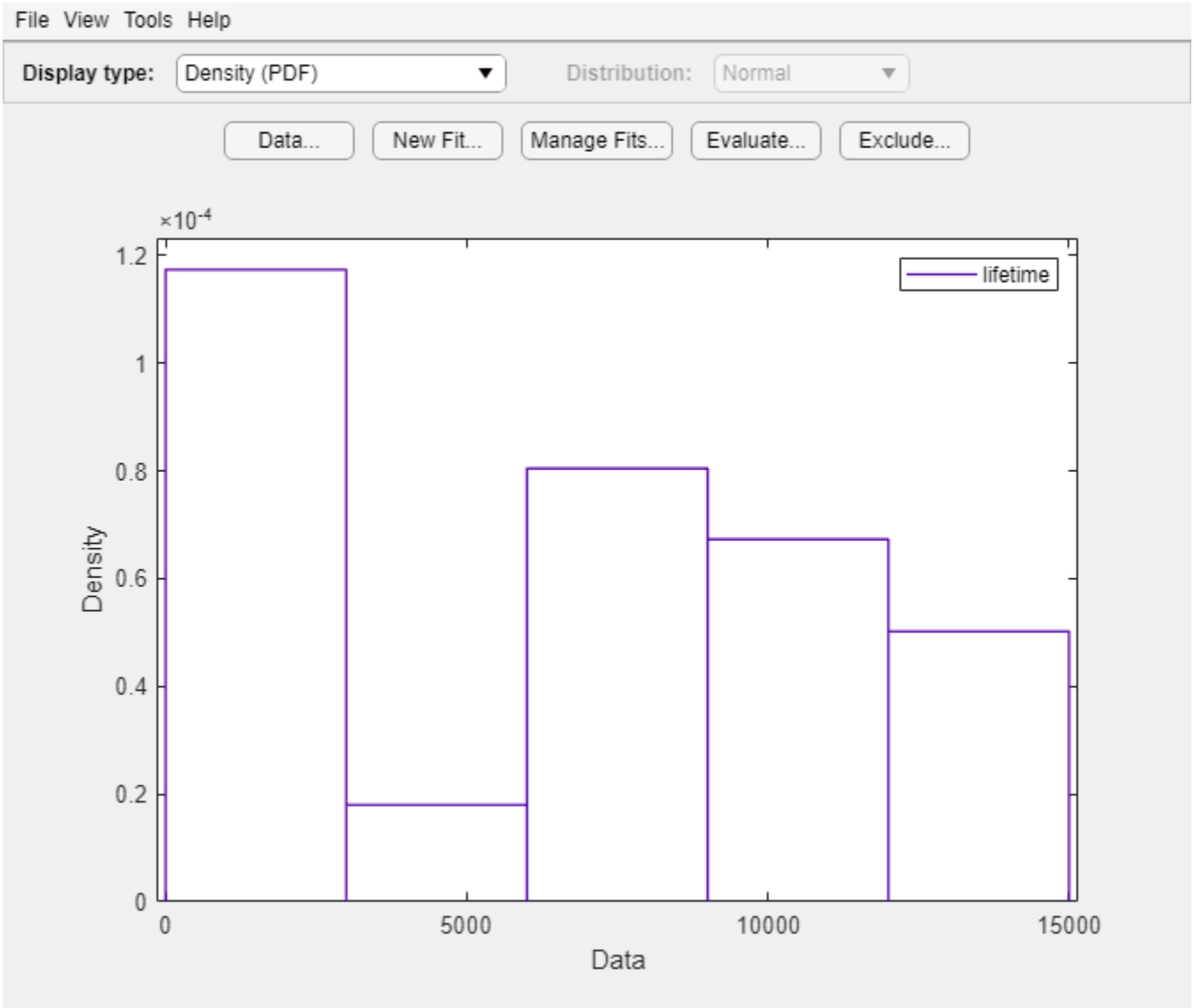
Load the sample data.

```
load lightbulb.mat
```

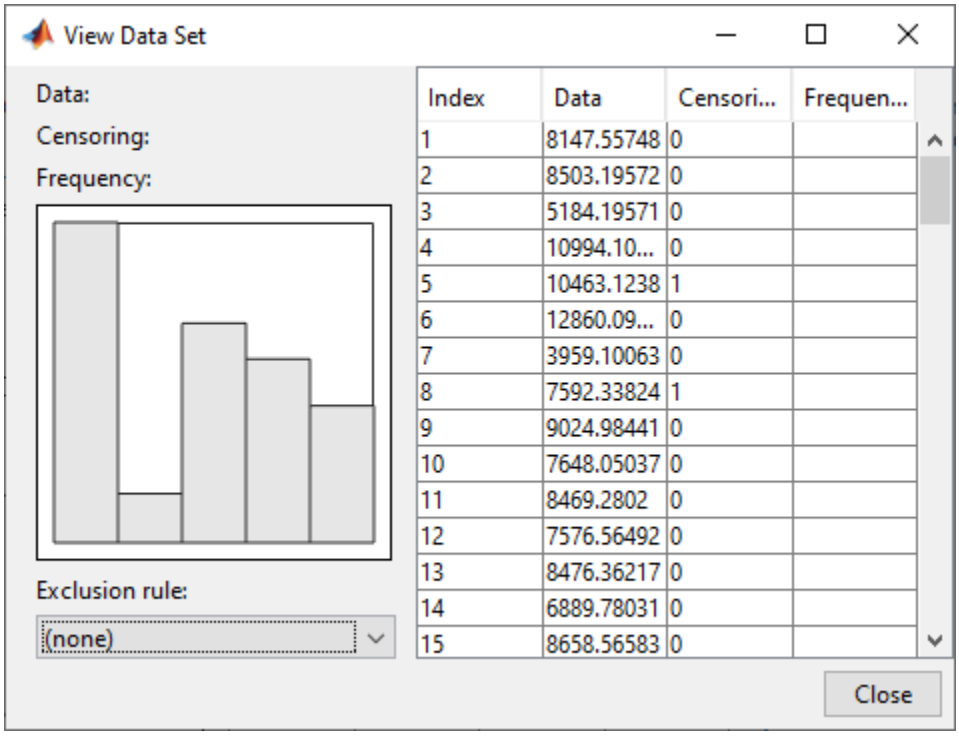
The first column of the data contains the lifetime (in hours) of two types of light bulbs. The second column contains information about the type of light bulb. 1 indicates fluorescent bulbs, and 0 indicates the incandescent bulb. The third column contains censoring information. 1 indicates censored data, and 0 indicates the exact failure time. This is simulated data.

Open the Distribution Fitter app using the first column of `lightbulb` as the input data, and the third column as the censoring data. Name the data `lifetime`.

```
distributionFitter(lightbulb(:,1),lightbulb(:,3),[],'lifetime')
```



To open the Data dialog box, click **Data**. In the **Manage data sets** pane, click to highlight the `lifetime` data set row. Finally, to open the View Data Set dialog, click **View**. The lifetime data appears in the second column and the corresponding censoring indicator appears in the third column.



Input Arguments

y — Input data

array of scalar values | variable representing an array of scalar values

Input data, specified as an array of scalar values or a variable representing an array of such values.

Data Types: `single` | `double`

cens — Censoring indicator

`zeros(n)` (default) | vector of 0 and 1 values

Censoring indicator, specified as a vector of 0 and 1 values. The length of `cens` must be equal to the length of `y`. If `y(j)` is censored, then `(cens(j)==1)`. If `y(j)` is not censored, then `(cens(j)==0)`. If `cens` is omitted or empty, then no `y` values are censored.

If you have frequency data (`freq`) but not censoring data (`cens`), then you must specify empty brackets (`[]`) for `cens`.

Data Types: `single` | `double`

freq — Frequency data

`ones(n)` (default) | vector of scalar values

Frequency data, specified as a vector of scalar values. The length of `freq` must be equal to the length of `y`. If `freq` is omitted or empty, then all `y` values have a frequency of 1.

If you have frequency data (`freq`) but not censoring data (`cens`), then you must specify empty brackets (`[]`) for `cens`.

Data Types: `single` | `double`

`dsname` — Data set name
character vector | string scalar

Data set name, specified as a character vector enclosed in single quotes or a string scalar enclosed in double quotes.

If you want to specify a data set name, but do not have censoring data (`cens`) or frequency data (`freq`), then you must specify empty brackets (`[]`) for both `freq` and `cens`.

Example: `'MyData'`

Data Types: `char` | `string`

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

"Fit a Distribution Using the Distribution Fitter App" on page 5-71
"Model Data Using the Distribution Fitter App" on page 5-51
"Working with Probability Distributions" on page 5-3
"Supported Distributions" on page 5-14

Introduced in R2017a

Probability Distribution Function

Interactive density and distribution plots

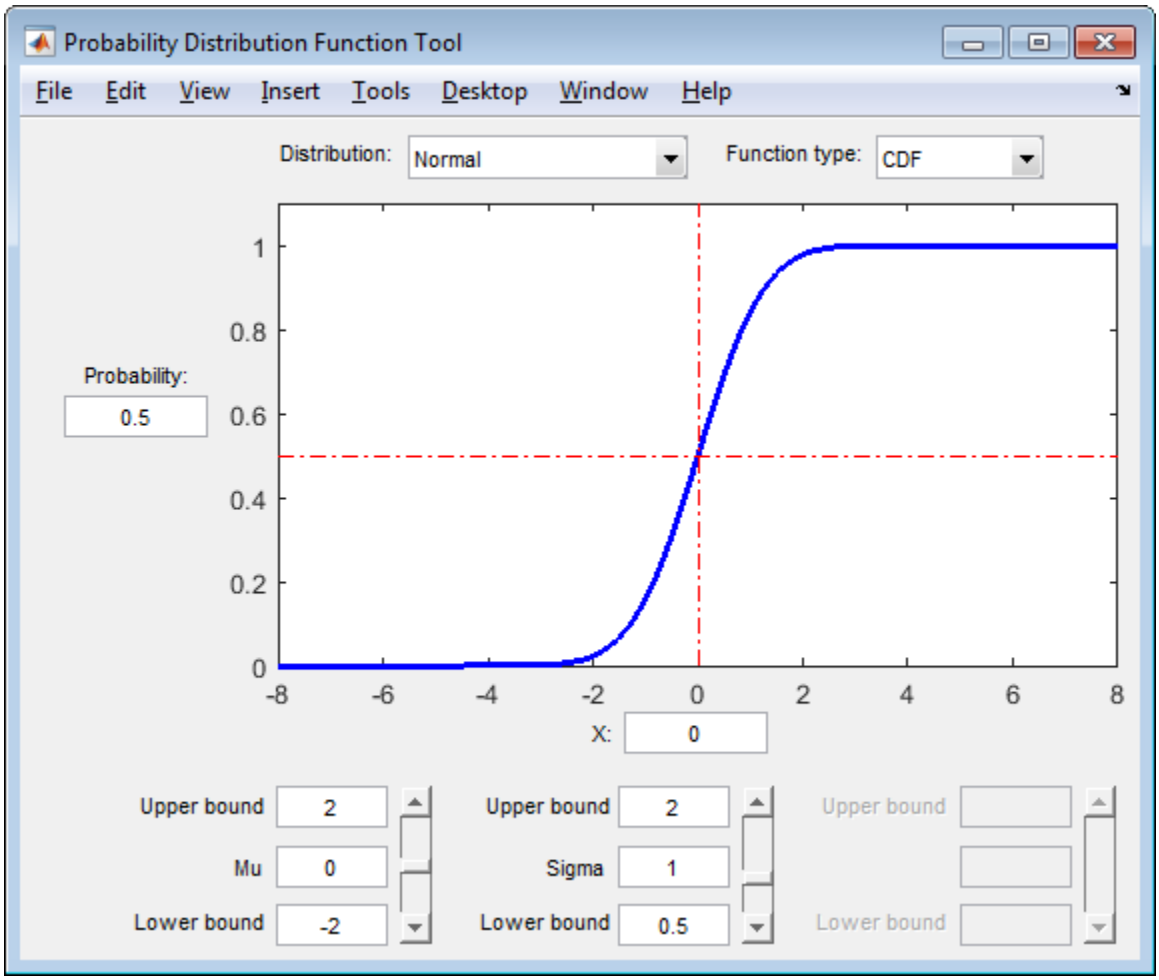
Description

The Probability Distribution Function user interface creates an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution. Explore the effects of changing parameter values on the shape of the plot, either by specifying parameter values or using interactive sliders.

Required Products

- MATLAB
- Statistics and Machine Learning Toolbox

Note: `disttool` does not provide printing, code generating, or data importing functionality in MATLAB Online.



Open the Probability Distribution Function App

- At the command prompt, enter `disttool`.

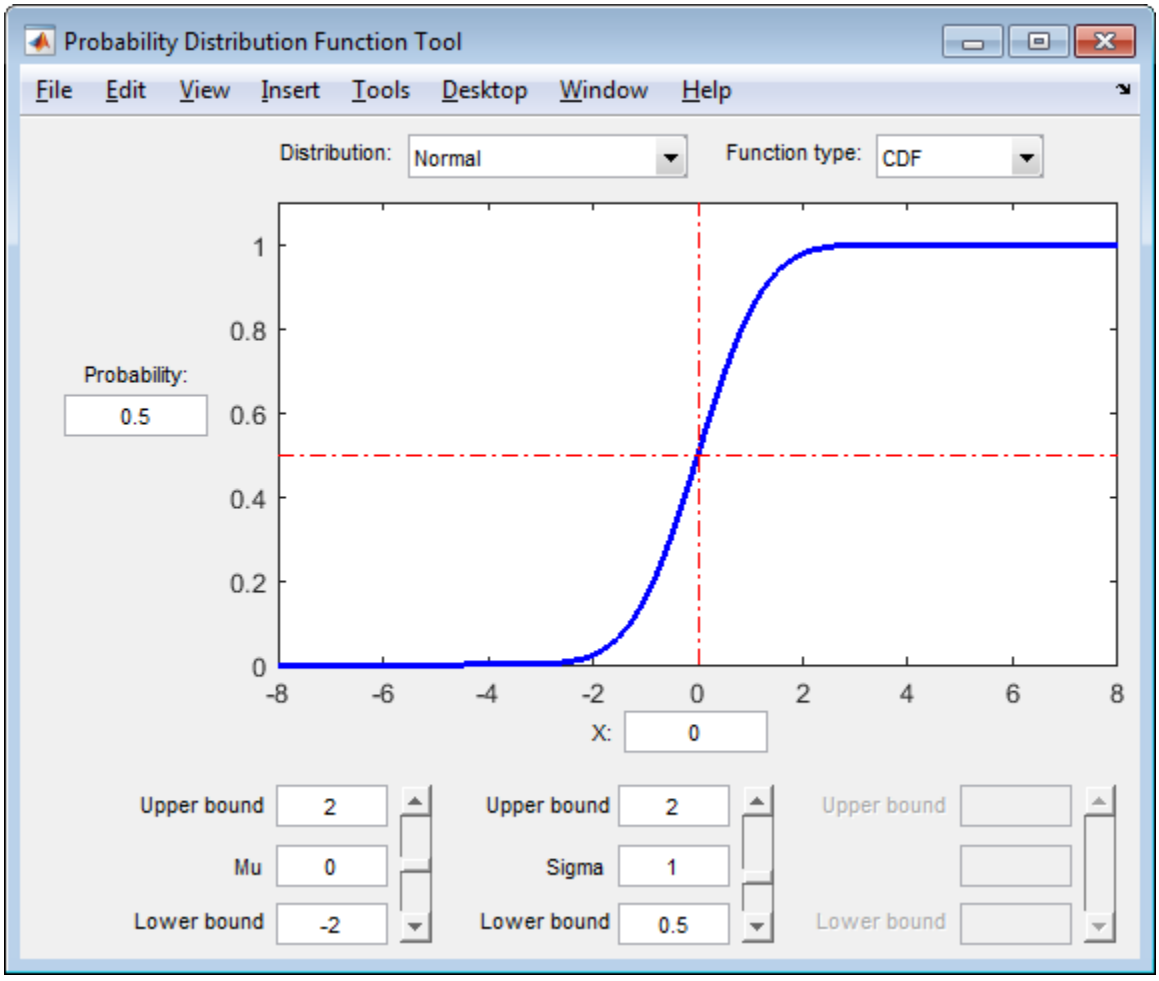
Examples

Explore the Probability Distribution Function User Interface

This example shows how to use the Probability Distribution Function user interface to explore the shape of cdf and pdf plots for different probability distributions and parameter values.

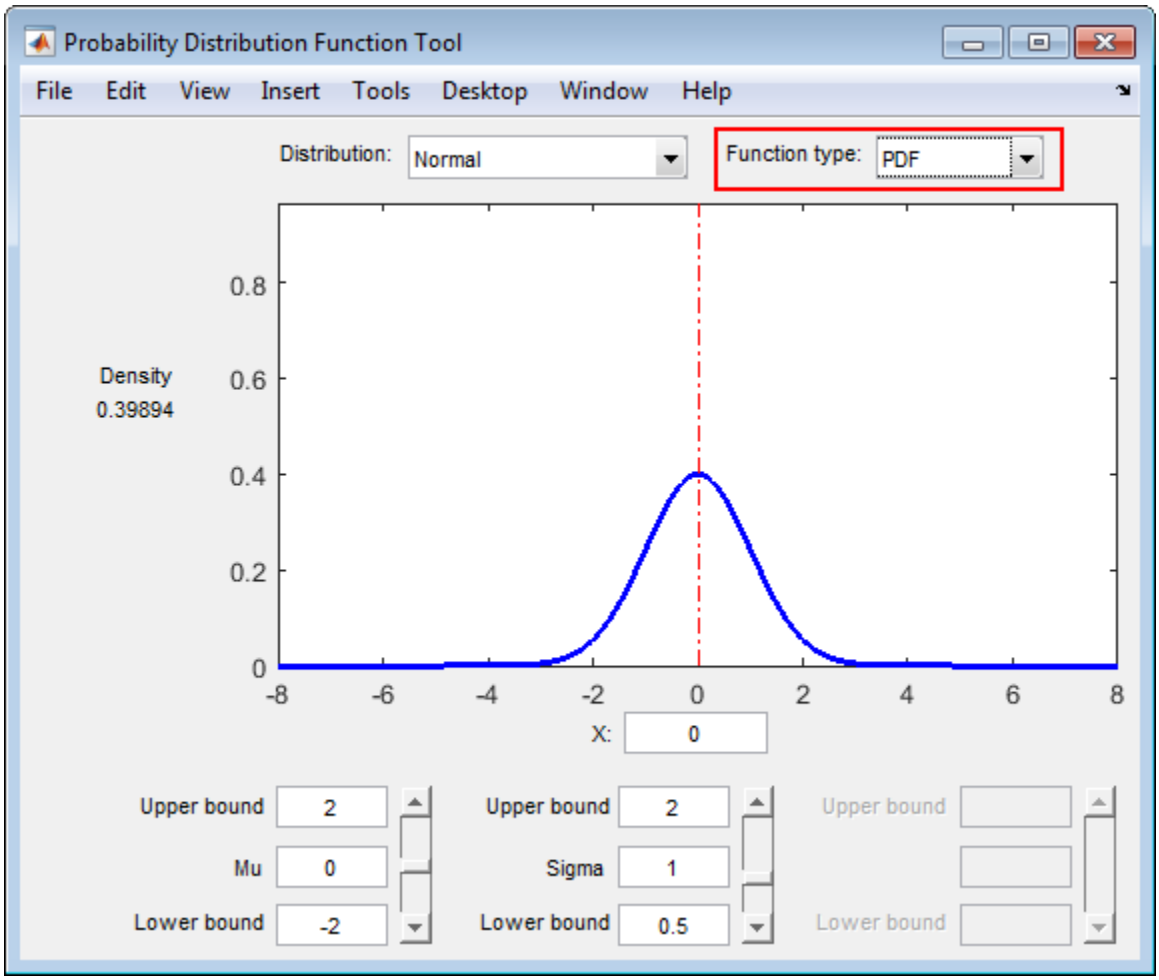
Open the Probability Distribution Function user interface.

```
disttool
```

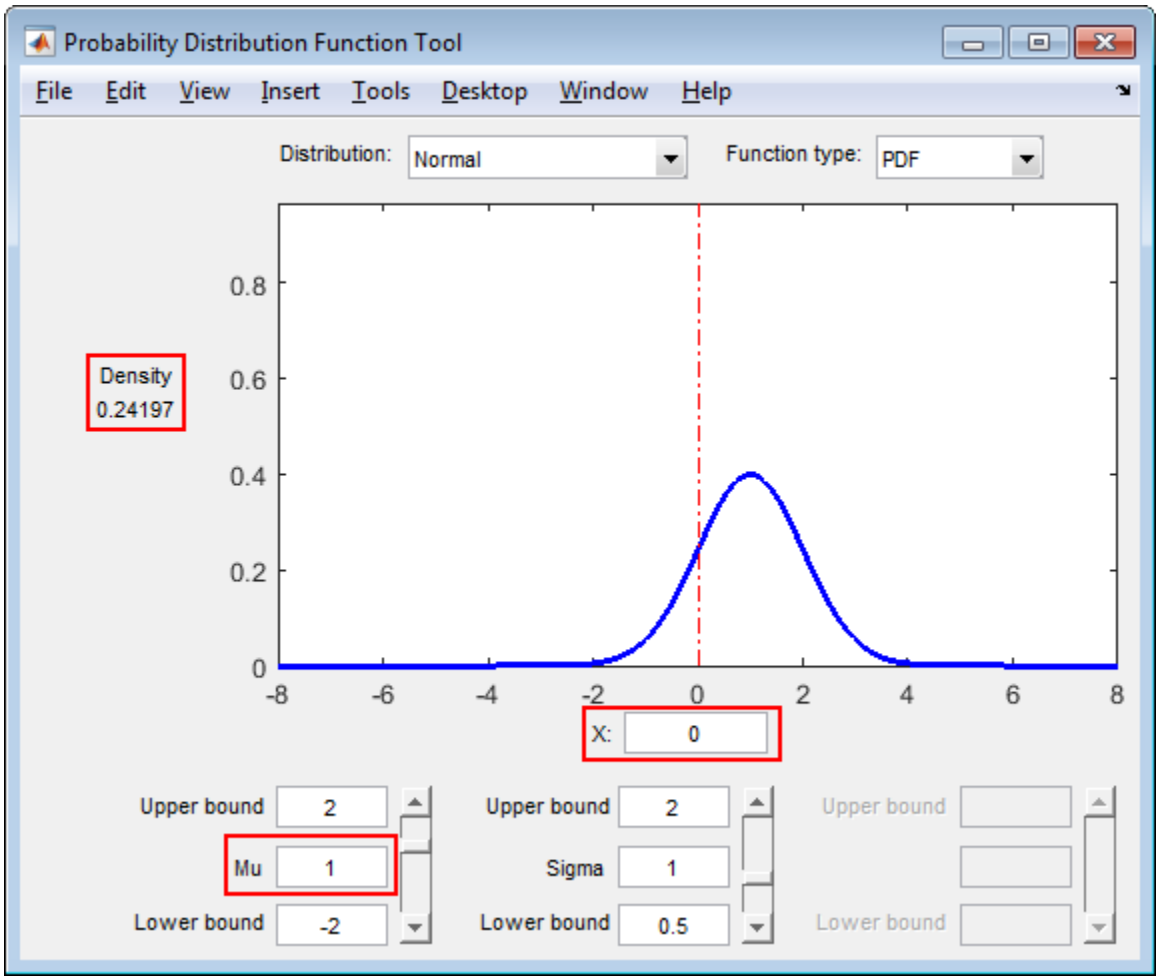


The interface opens with a plot of the cdf of the Normal distribution. The initial parameter settings are `Mu = 0` and `Sigma = 1`.

Select `PDF` from the **Function type** drop-down menu to plot the pdf of the Normal distribution using the same parameter values.

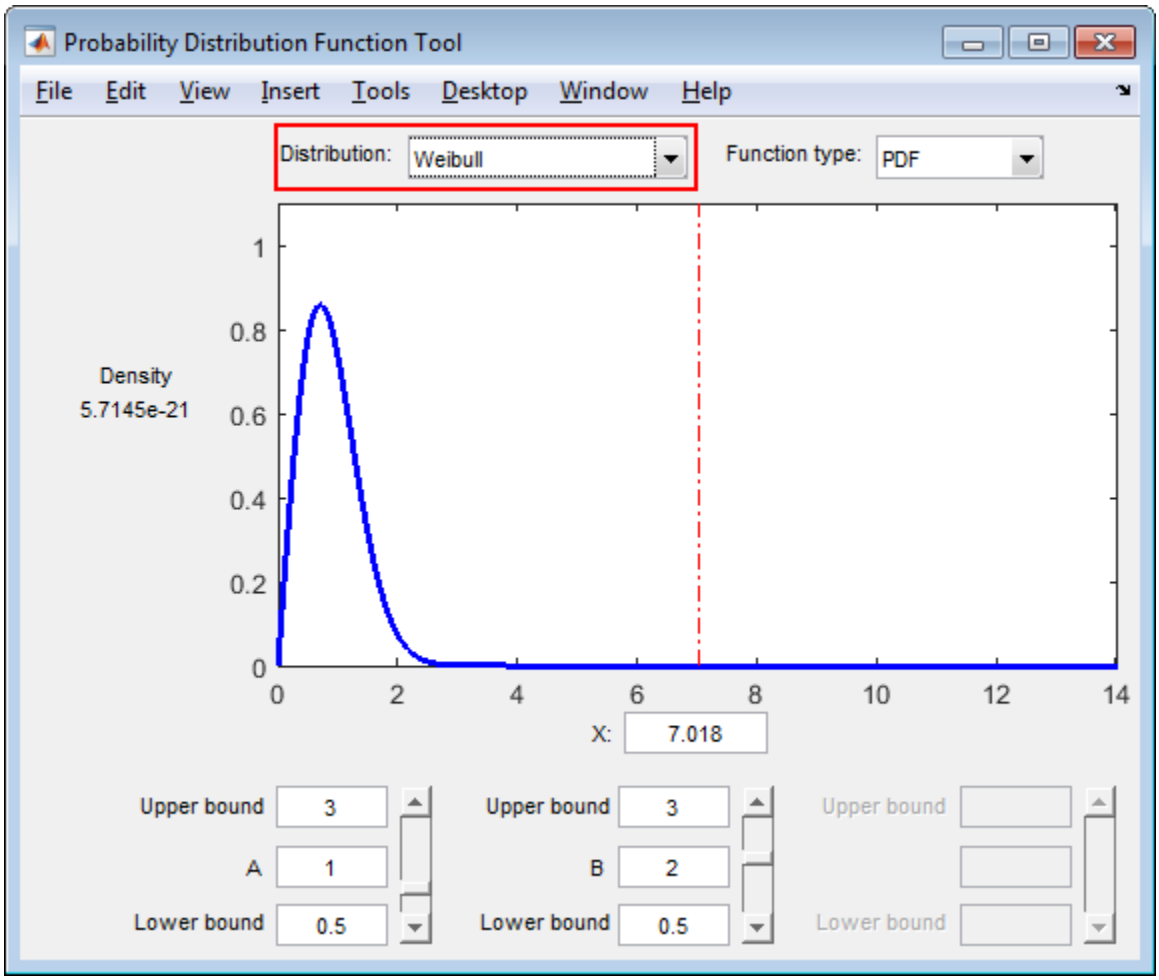


Change the value of the location parameter `Mu` to 1.



As the parameter values change, the shape of the plot also changes. Also, the value of X remains the same, but the density value changes because of the new parameter value.

Use the **Distribution** drop-down menu to change the distribution type from `Normal` to **Weibull**.



The shape of the plot changes, along with the names and values of the parameters.

Parameters

`Distribution` — Probability distribution
`Normal` (default) | `Exponential` | `Poisson` | `Weibull` | `...`

Specify the probability distribution to explore by selecting a distribution name from the drop-down list. The drop-down list includes approximately 25 probability distribution options, including `Normal`, `Exponential`, `Poisson`, `Weibull`, and more.

`Function type` — Probability distribution function type
`CDF` (default) | `PDF`

Specify the probability distribution function type as `CDF` (cumulative distribution function) or `PDF` (probability density function) by selecting the function name from the drop-down list.

`Probability` — Cumulative distribution function value
numeric value in the range [0,1]

Specify the cumulative distribution function (cdf) value of interest as a numeric value in the range [0,1]. The corresponding random variable value appears in the **X** field below the plot. Alternatively, you can specify a value for **X**, and the **Probability** value will update automatically.

This option only appears if **Function type** is `CDF`. If **Function type** is `PDF`, then the probability density at the specified **X** value displays to the left of the plot.

`X` — Random variable
numeric value

Specify the random variable of interest as a numeric value. If the **Function type** is `CDF`, then the corresponding cumulative distribution function (cdf) value appears in the **Probability** field to the left of the plot. Alternatively, you can specify a value for **Probability**, and the **X** value will update automatically. If the **Function type** is `PDF`, then the corresponding probability density value appears to the left of the plot.

`Parameters` — Parameter boundaries and values
numeric value

Specify the parameter boundaries and values as numeric values. Each column contains a field for the upper bound, value, and lower bound of one parameter. The name and number of available parameters changes based on the distribution specified in the **Distribution** drop-down list. For example, if you select the `Normal` distribution, then `disttool` enables two columns: One column for the **Mu** parameter and one column for the **Sigma** parameter. If you select the **Exponential** distribution, then `disttool` enables one column for the **Mu** parameter.

Tips

To change the value of **X** (on the *y*-axis), or **Probability** or **Density** (on the *x*-axis):

- Type the values of interest into the **Probability** or **X** fields;
- Click on the point of interest on the plot; or
- Click and drag the reference lines across the plot.

See Also

Functions
`distributionFitter` | `fitdist` | `makedist`

Introduced before R2006a

double

Class: `dataset`

(Not Recommended) Convert dataset variables to double array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
b = double(A)
b = double(a,vars)
```

Description

`b = double(A)` returns the contents of the dataset `A`, converted to one double array. The classes of the variables in the dataset must support the conversion.

`b = double(a,vars)` returns the contents of the dataset variables specified by `vars`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

See Also

`dataset` | `replacedata` | `single`

droplevels

(Not Recommended) Drop levels from a nominal or ordinal array

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
B = droplevels(A)
B = droplevels(A,oldlevels)
```

Description

`B = droplevels(A)` drops unused levels from the nominal or ordinal array `A`. The array `B` has the same size, type, and values as `A`, but has a list of potential levels that includes only those present in some element of `A`.

`B = droplevels(A,oldlevels)` removes the specified levels `oldlevels` from `A`.

`droplevels` removes levels, but does not remove elements. Elements of `B` that correspond to elements of `A` having levels in `oldlevels` all have an undefined level.

Examples

Drop Levels From an Ordinal Array

Bin patient ages into ordinal levels corresponding to 10-year intervals.

```
load hospital
edges = 0:10:100;
labels = strcat(num2str((0:10:90)','%d'),{'s'});
A = ordinal(hospital.Age,labels,[],edges);
getlabels(A)
```

```
ans = 1x10 cell
  Columns 1 through 7

    {'0s'}    {'10s'}    {'20s'}    {'30s'}    {'40s'}    {'50s'}    {'60s'}

  Columns 8 through 10

    {'70s'}    {'80s'}    {'90s'}
```

Drop any levels that have no patients in them.

```
A = droplevels(A);
getlabels(A)
```

```
ans = 1x4 cell
    {'20s'}    {'30s'}    {'40s'}    {'50s'}
```

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

oldlevels — Levels to remove

string array | cell array of character vectors | 2-D character matrix

Levels to remove from the `nominal` or `ordinal` array, specified as a string array, a cell array of character vectors, or a 2-D character matrix.

Data Types: `char` | `string` | `cell`

Output Arguments

B — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, returned as a `nominal` or `ordinal` array object.

See Also

`addlevels` | `mergelevels` | `nominal` | `ordinal` | `reorderlevels`

Topics

“Add and Drop Category Levels” on page 2-18

Introduced in R2007a

dummyvar

Create dummy variables

Syntax

```
D = dummyvar(group)
```

Description

`D = dummyvar(group)` returns a matrix `D` containing zeros and ones, whose columns are dummy variables for the grouping variables on page 2-45 in `group`. Each column of `group` is a single grouping variable, with values indicating category levels. The rows of `group` represent observations across all variables.

Examples

Create Dummy Variables from Categorical Grouping Variable

Create a column vector of categorical data specifying color types.

```
Colors = {'Red';'Blue';'Green';'Red';'Green';'Blue'};
Colors = categorical(Colors);
```

Create dummy variables for each color type.

```
D = dummyvar(Colors)
```

```
D = 6×3

     0     0     1
     1     0     0
     0     1     0
     0     0     1
     0     1     0
     1     0     0
```

The columns in `D` correspond to the levels in `Colors`. For example, the first column of `dummyvar` corresponds to the first level, `'Blue'`, in `Colors`.

Display the category levels of `Colors`.

```
categories(Colors)
```

```
ans = 3x1 cell
    {'Blue' }
    {'Green'}
    {'Red'  }
```

Create Dummy Variables from Numeric Grouping Variables

Create a matrix `group` of data containing the effects of two machines and three operators on a process.

```
machine = [1 1 1 1 2 2 2 2]';
operator = [1 2 3 1 2 3 1 2]';
group = [machine operator]
```

```
group = 8×2

     1     1
     1     2
     1     3
     1     1
     2     2
     2     3
     2     1
     2     2
```

Create dummy variables of the data in `group`.

```
D = dummyvar(group)
```

```
D = 8×5

     1     0     1     0     0
     1     0     0     1     0
     1     0     0     0     1
     1     0     1     0     0
     0     1     0     1     0
     0     1     0     0     1
     0     1     1     0     0
     0     1     0     1     0
```

The first two columns of `D` represent observations of machine 1 and machine 2, respectively. The remaining columns represent observations of the three operators.

Create Dummy Variables from Multiple Grouping Variables

Create a cell array of phone types and a numeric vector of area codes.

```
phone = {'mobile';'landline';'mobile';'mobile';'mobile';'landline';'landline'};
codes = [802 802 603 603 802 603 802]';
```

Because the area code data has two levels (rather than 802 levels corresponding to the integers `1:802`), convert `codes` to a categorical vector.

```
newcodes = categorical(codes);
```

Combine the `phone` and `newcodes` grouping variables into the cell array `group`.

```
group = {phone,newcodes};
```

Create dummy variables for the groups in `group`.

```
D = dummyvar(group)
```

```
D = 7×4

     1     0     0     1
     0     1     0     1
     1     0     1     0
     1     0     1     0
     1     0     0     1
     0     1     1     0
     0     1     0     1
```

The first two columns of `D` correspond to the phone types, and the last two columns correspond to the area codes.

Input Arguments

`group` — Grouping variables

positive integer vector | categorical column vector | cell array | positive integer matrix

Grouping variables, specified as a positive integer vector or categorical column vector representing levels within a single variable, a cell array containing one or more grouping variables on page 2-45, or a positive integer matrix representing levels within multiple variables.

If `group` is a categorical vector, then the groups and their order match the output of the `categories` function applied to `group`. If `group` is a numeric vector, then `dummyvar` assumes that the groups and their order are `1:max(group)`. In this respect, `dummyvar` treats a numeric grouping variable differently from `grp2idx`. For information on the order of groups within grouping variables, see "Grouping Variables" on page 2-45.

Example: `[2 1 1 1 2 3 3 2]'`

Example: `{Origin,Cylinders}`

Data Types: `single` | `double` | `categorical` | `cell`

Output Arguments

`D` — Dummy variables

numeric matrix

Dummy variables, returned as an *n*-by-*s* numeric matrix, where *n* is the number of rows of `group` and *s* is the sum of the number of levels in each column of `group`. From left to right, the columns of `D` are dummy variables created from the first column of `group`, followed by dummy variables created from the second column of `group`, and so on.

Data Types: `single` | `double`

Tips

- Use dummy variables in regression analysis and ANOVA to indicate values of categorical predictors.

- `dummyvar` treats `NaN` values and undefined categorical levels in `group` as missing data and returns `NaN` values in `D`.
- If a column of ones is introduced in the matrix `D`, then the resulting matrix `X = [ones(size(D,1),1) D]` is rank deficient. If `group` has multiple columns, then the matrix `D` itself is rank deficient because dummy variables produced from any column of `group` always sum to a column of ones. Regression and ANOVA calculations often address this issue by eliminating one dummy variable (implicitly setting the coefficients for dropped columns to zero) from each group of dummy variables produced by a column of `group`.
- If `group` is a numeric vector with levels that do not correspond exactly to the integers `1:max(group)`, first convert the data to a categorical vector by using `categorical`. You can then pass the result to `dummyvar`. For an example, see “Create Dummy Variables from Multiple Grouping Variables” on page 33-1260.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`anova1` | `categories` | `grp2idx` | `regress`

Topics

“Grouping Variables” on page 2-45
“Dummy Variables” on page 2-48
“Linear Regression with Categorical Covariates” on page 2-52

Introduced before R2006a

dwtest

Durbin-Watson test with residual inputs

Syntax

```
p = dwtest(r,x)
p = dwtest(r,x,Name,Value)
[p,d] = dwtest( ___ )
```

Description

`p = dwtest(r,x)` returns the *p*-value for the Durbin-Watson test on page 33-1265 of the null hypothesis that the residuals from a linear regression are uncorrelated. The alternative hypothesis is that there is autocorrelation among the residuals.

`p = dwtest(r,x,Name,Value)` returns the *p*-value for the Durbin-Watson test with additional options specified by one or more name-value pair arguments. For example, you can conduct a one-sided test or calculate the *p*-value using a normal approximation.

`[p,d] = dwtest( ___ )` also returns the Durbin-Watson test statistic, `d`, using any of the input arguments from the previous syntaxes.

Examples

Test Residuals For Correlation

Load the sample census data.

```
load census
```

Create a design matrix using the census date (`cdate`) as the predictor. Add a column of `1` values to include a constant term.

```
n = length(cdate);
x = [ones(n,1),cdate];
```

Fit a linear regression to the data.

```
[b,bint,r] = regress(pop,x);
```

Test the null hypothesis that there is no autocorrelation among the residuals, `r`.

```
[p,d] = dwtest(r,x)
```

```
p = 3.6190e-15
```

```
d = 0.1308
```

The returned value `p = 3.6190e-15` indicates rejection of the null hypothesis at the 5% significance level.

One-Sided Hypothesis Test

Load the sample census data.

```
load census
```

Create a design matrix using the census date (`cdate`) as the predictor. Add a column of `1` values to include a constant term.

```
n = length(cdate);
x = [ones(n,1),cdate];
```

Fit a linear regression to the data.

```
[b,bint,r] = regress(pop,x);
```

Test the null hypothesis that there is no autocorrelation among regression residuals, against the alternative hypothesis that the autocorrelation is greater than zero.

```
[p,d] = dwtest(r,x,'Tail','right')
```

```
p = 1.8095e-15
```

```
d = 0.1308
```

The returned value `p = 1.8095e-15` indicates rejection of the null hypothesis at the 5% significance level, in favor of the alternative hypothesis that the autocorrelation among residuals is greater than zero.

Input Arguments

x — Design matrix

matrix

Design matrix for a linear regression, specified as a matrix. Include a column of `1` values in the design matrix so the model contains a constant term.

Data Types: `single` | `double`

r — Regression residuals

vector

Regression residuals, specified as a vector. Obtain `r` by performing a linear regression using a function such as `regress`, or by using the backslash operator.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Method','approximate'` specifies a right-tailed hypothesis test and calculates the p-value using a normal approximation.

`Method` — Algorithm for computing *p*-value
`'exact'` | `'approximate'`

Algorithm for computing the *p*-value, specified as the comma-separated pair consisting of `'Method'` and one of these values:

`'exact'`	Calculate an exact *p*-value using the Pan algorithm [2]. This is the default if the sample size is less than 400.
`'approximate'`	Calculate the *p*-value using a normal approximation [1]. This is the default if the sample size is 400 or larger.

Example: `'Method','exact'`

`Tail` — Type of alternative hypothesis
`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternate hypothesis that autocorrelation among the residuals is not zero.
`'right'`	Test the alternative hypothesis that autocorrelation among the residuals is greater than zero.
`'left'`	Test the alternative hypothesis that autocorrelation among the residuals is less than zero.

Example: `'Tail','right'`

Output Arguments

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`d` — Test statistic
nonnegative scalar value

Test statistic of the hypothesis test, returned as a nonnegative scalar value.

More About

Durbin-Watson Test

The Durbin-Watson test tests the null hypothesis that linear regression residuals of time series data are uncorrelated, against the alternative hypothesis that autocorrelation exists.

The test statistic for the Durbin-Watson test is

$$DW = \frac{\sum_{i=1}^{n-1} (r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2},$$

where r_i is the ith raw residual, and n is the number of observations.

The p-value of the Durbin-Watson test is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. A significantly small p-value casts doubt on the validity of the null hypothesis and indicates autocorrelation among residuals.

Alternative Functionality

- You can create a linear regression model object by using `fitlm` or `stepwiselm` and use the object function `dwtest` to perform the Durbin-Watson test.

  A `LinearModel` object provides the object properties and the object functions to investigate a fitted linear regression model. The object properties include information about coefficient estimates, summary statistics, fitting method, and input data. Use the object functions to predict responses and to modify, evaluate, and visualize the linear regression model.

References

[1] Durbin, J., and G. S. Watson. "Testing for Serial Correlation in Least Squares Regression I." *Biometrika* 37, pp. 409–428, 1950.

[2] Farebrother, R. W. Pan's "Procedure for the Tail Probabilities of the Durbin-Watson Statistic." *Applied Statistics* 29, pp. 224–227, 1980.

See Also

`dwtest` | `fitlm` | `regress`

Introduced in R2006a

dwtest

Durbin-Watson test with linear regression model object

Syntax

```
p = dwtest(mdl)
p = dwtest(mdl,method)
p = dwtest(mdl,method,tail)
[p,DW] = dwtest( ___ )
```

Description

`p = dwtest(mdl)` returns the *p*-value of the "Durbin-Watson Test" on page 33-1268 on the residuals of the linear regression model `mdl`. The null hypothesis is that the residuals are uncorrelated, and the alternative hypothesis is that the residuals are autocorrelated.

`p = dwtest(mdl,method)` specifies the algorithm for computing the *p*-value.

`p = dwtest(mdl,method,tail)` specifies the alternative hypothesis.

`[p,DW] = dwtest( ___ )` also returns the Durbin-Watson statistic using any of the input argument combinations in the previous syntaxes.

Examples

Test Residuals for Autocorrelation

Determine whether a fitted linear regression model has autocorrelated residuals.

Load the `census` data set and create a linear regression model.

```
load census
mdl = fitlm(cdate,pop);
```

Find the *p*-value of the Durbin-Watson autocorrelation test.

```
p = dwtest(mdl)
```

```
p = 3.6190e-15
```

The small *p*-value indicates that the residuals are autocorrelated.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`method` — Algorithm for computing *p*-value
`'exact'` | `'approximate'`

Algorithm for computing the *p*-value, specified as one of these values:

- `'exact'` — Calculate an exact *p*-value using Pan's algorithm [2].
- `'approximate'` — Calculate the *p*-value using a normal approximation [1].

The default is `'exact'` when the sample size is less than `400`, and `'approximate'` otherwise.

`tail` — Type of alternative hypothesis
`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to test, specified as one of these values:

Value	Alternative Hypothesis
`'both'`	Serial correlation is not 0.
`'right'`	Serial correlation is greater than 0 (right-tailed test).
`'left'`	Serial correlation is less than 0 (left-tailed test).

`dwtest` tests whether `mdl` has no serial correlation, against the specified alternative hypothesis.

Output Arguments

`p` — *p*-value of test
numeric value

p-value of the test, returned as a numeric value. `dwtest` tests whether the residuals are uncorrelated, against the alternative that autocorrelation exists among the residuals. A small *p*-value indicates that the residuals are autocorrelated.

`DW` — Durbin-Watson statistic
nonnegative numeric value

Durbin-Watson statistic value, returned as a nonnegative numeric value.

More About

Durbin-Watson Test

The Durbin-Watson test tests the null hypothesis that linear regression residuals of time series data are uncorrelated, against the alternative hypothesis that autocorrelation exists.

The test statistic for the Durbin-Watson test is

$$DW = \frac{\sum_{i=1}^{n-1} (r_{i+1} - r_i)^2}{\sum_{i=1}^{n} r_i^2},$$

where r_i is the ith raw residual, and n is the number of observations.

The *p*-value of the Durbin-Watson test is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. A significantly small *p*-value casts doubt on the validity of the null hypothesis and indicates autocorrelation among residuals.

References

[1] Durbin, J., and G. S. Watson. "Testing for Serial Correlation in Least Squares Regression I." *Biometrika* 37, pp. 409–428, 1950.

[2] Farebrother, R. W. Pan's "Procedure for the Tail Probabilities of the Durbin-Watson Statistic." *Applied Statistics* 29, pp. 224–227, 1980.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `anova` | `coefCI` | `coefTest`

Topics

“Durbin-Watson Test” on page 11-70
“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2012a

ecdf

Empirical cumulative distribution function

Syntax

```
[f,x] = ecdf(y)
[f,x] = ecdf(y,Name,Value)
[f,x,flo,fup] = ecdf(___)

ecdf(___)
ecdf(ax,___)
```

Description

`[f,x] = ecdf(y)` returns the empirical cumulative distribution function (cdf), `f`, evaluated at the points in `x`, using the data in the vector `y`.

In survival and reliability analysis, this empirical cdf is called the Kaplan-Meier estimate. And the data might correspond to survival or failure times.

`[f,x] = ecdf(y,Name,Value)` returns the empirical function values, `f`, evaluated at the points in `x`, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the type of function to evaluate or which data is censored.

`[f,x,flo,fup] = ecdf(___)` also returns the 95% lower and upper confidence bounds for the evaluated function values. You can use any of the input arguments in the previous syntaxes.

`ecdf` computes the confidence bounds using Greenwood's formula on page 33-1279. They are not simultaneous confidence bounds.

`ecdf(___)` draws a stairstep graph of the evaluated function by using the `stairs` function. Specify `'Bounds','on'` to include the confidence bounds in the graph.

`ecdf(ax,___)` plots on the axes specified by `ax` instead of the current axes (`gca`).

Examples

Compute Empirical Cumulative Distribution Function

Compute the Kaplan-Meier estimate of the cumulative distribution function (cdf) for simulated survival data.

Generate survival data from a Weibull distribution with parameters 3 and 1.

```
rng('default')  % for reproducibility
failuretime = random('wbl',3,1,15,1);
```

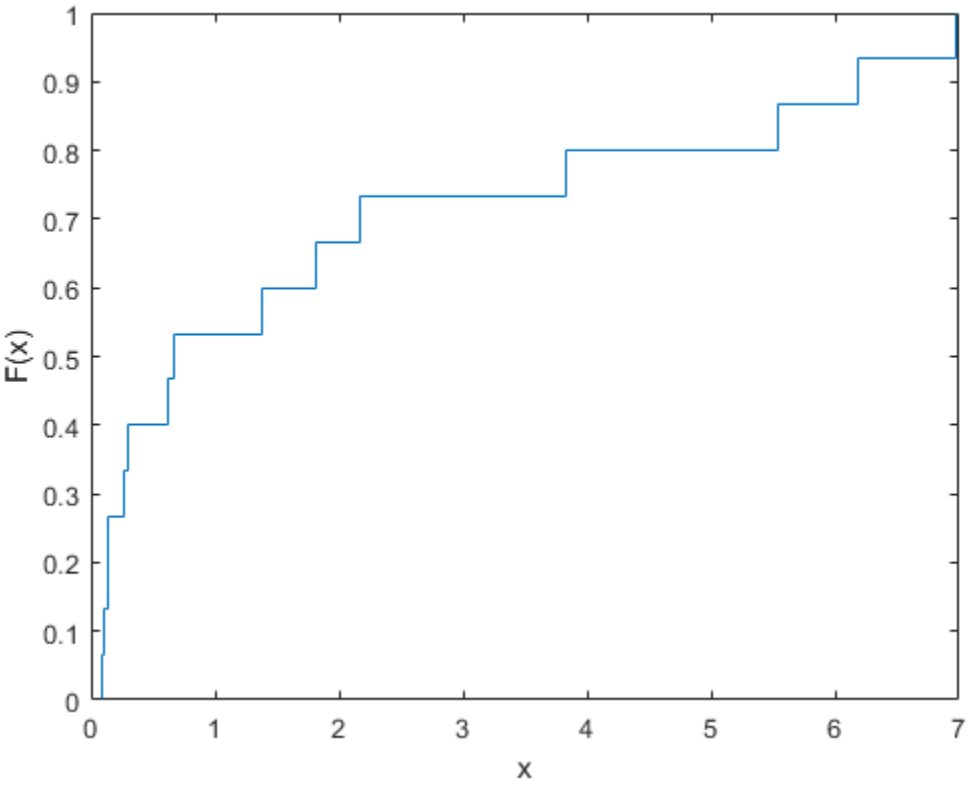
Compute the Kaplan-Meier estimate of the cdf for survival data.

```
[f,x] = ecdf(failuretime);
[f,x]
```

```
ans = 16×2

         0    0.0895
    0.0667    0.0895
    0.1333    0.1072
    0.2000    0.1303
    0.2667    0.1313
    0.3333    0.2718
    0.4000    0.2968
    0.4667    0.6147
    0.5333    0.6684
    0.6000    1.3749
       ⋮
```

Plot the estimated cdf.

```
ecdf(failuretime)
```



Empirical Hazard Function of Right-Censored Data

Compute and plot the hazard function of simulated right-censored survival data.

Generate failure times from a Birnbaum-Saunders distribution.

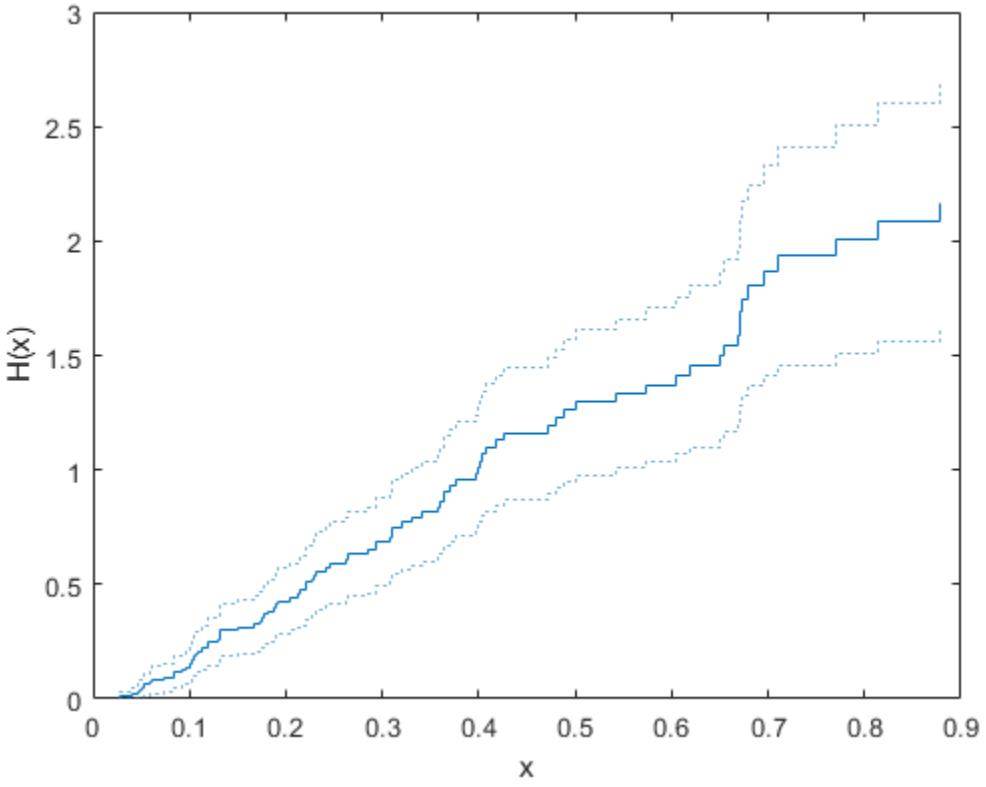
```
rng('default')  % For reproducibility
failuretime = random('birnbaumsaunders',0.3,1,100,1);
```

Assuming that the end of the study is at time 0.9, generate a logical array that indicates simulated failure times that are larger than 0.9 as censored data, and store this information in a vector.

```
T = 0.9;
cens = (failuretime>T);
```

Plot the empirical hazard function for the data.

```
ecdf(failuretime,'Function','cumulative hazard', ...
    'Censoring',cens,'Bounds','on');
```



Compare Empirical Cumulative Distribution Function (CDF) with Known CDF

Generate right-censored survival data and compare the empirical cumulative distribution function (cdf) with the known cdf.

Generate failure times from an exponential distribution with mean failure time of 15.

```
rng('default')  % For reproducibility
y = exprnd(15,75,1);
```

Generate drop-out times from an exponential distribution with mean failure time of 30.

```
d = exprnd(30,75,1);
```

Generate the observed failure times. They are the minimum of the generated failure times and the drop-out times.

```
t = min(y,d);
```

Create a logical array that indicates generated failure times that are larger than the drop-out times. The data for which this is true are censored.

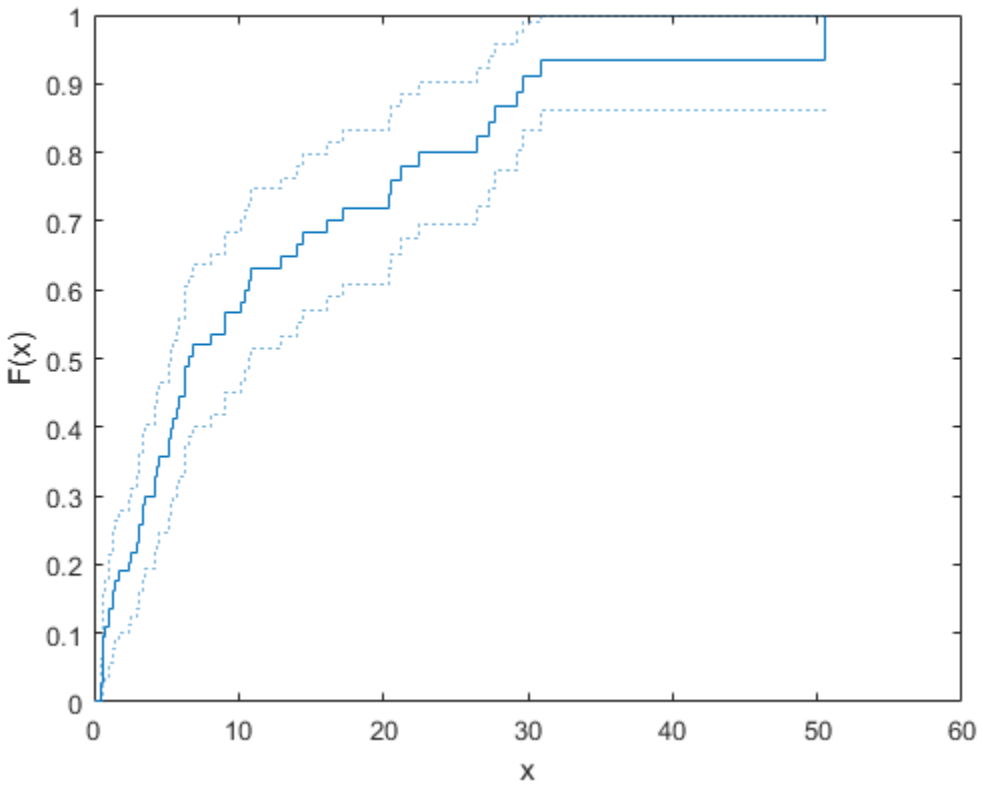
```
censored = (y>d);
```

Compute the empirical cdf and confidence bounds.

```
[f,x,flo,fup] = ecdf(t,'Censoring',censored);
```

Plot the cdf and confidence bounds.

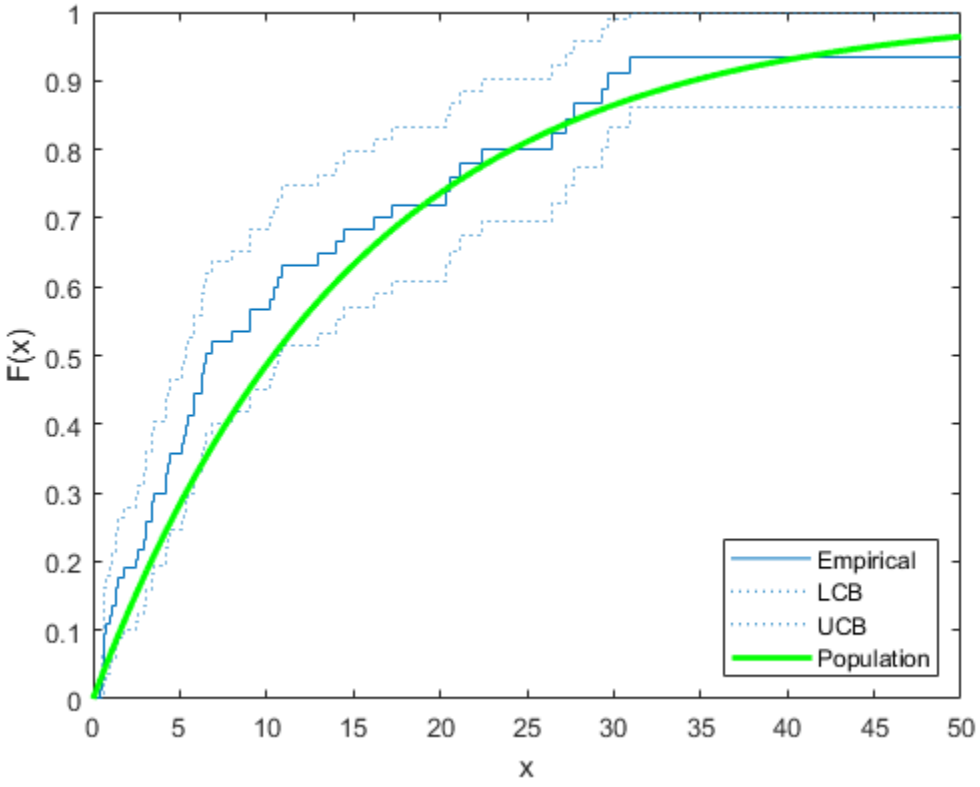
```
figure()
ecdf(t,'Censoring',censored,'Bounds','on');
hold on
```



Superimpose a plot of the known population cdf.

```
xx = 0:.1:max(t);
yy = 1-exp(-xx/15);
```

```
plot(xx,yy,'g-','LineWidth',2)
axis([0 50 0 1])
legend('Empirical','LCB','UCB','Population', ...
    'Location','southeast')
hold off
```



Empirical Survivor Function with Confidence Bounds

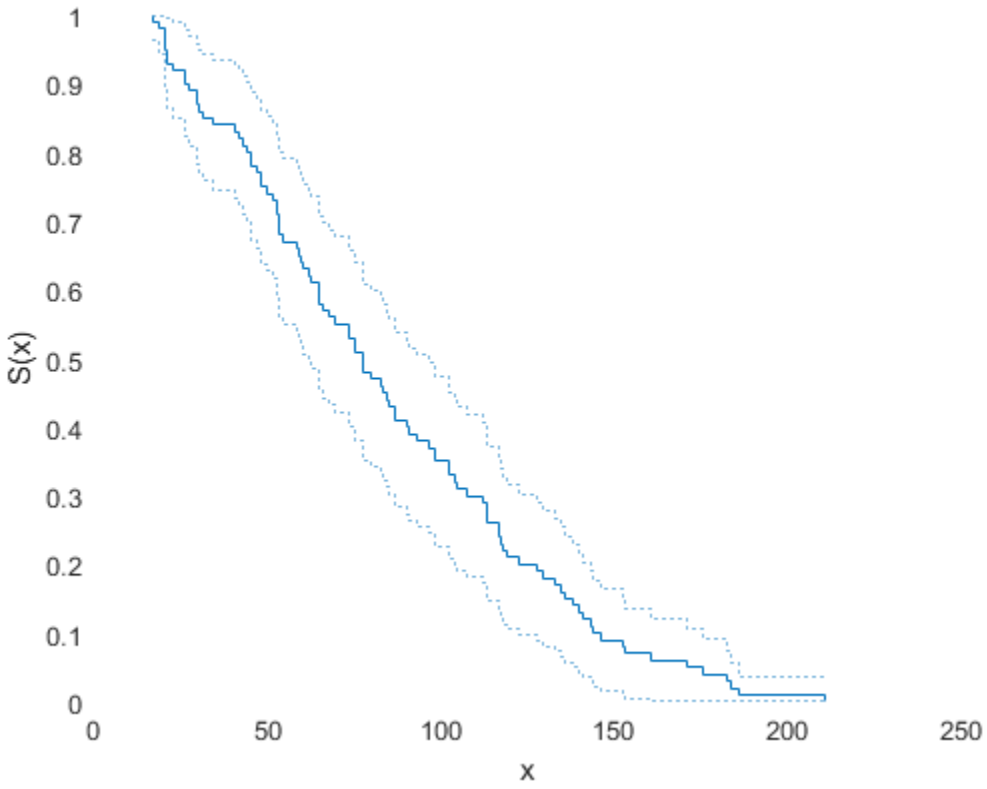
Generate survival data and plot the empirical survivor function with 99% confidence bounds.

Generate lifetime data from a Weibull distribution with parameters 100 and 2.

```
rng('default')  % For reproducibility
R = wblrnd(100,2,100,1);
```

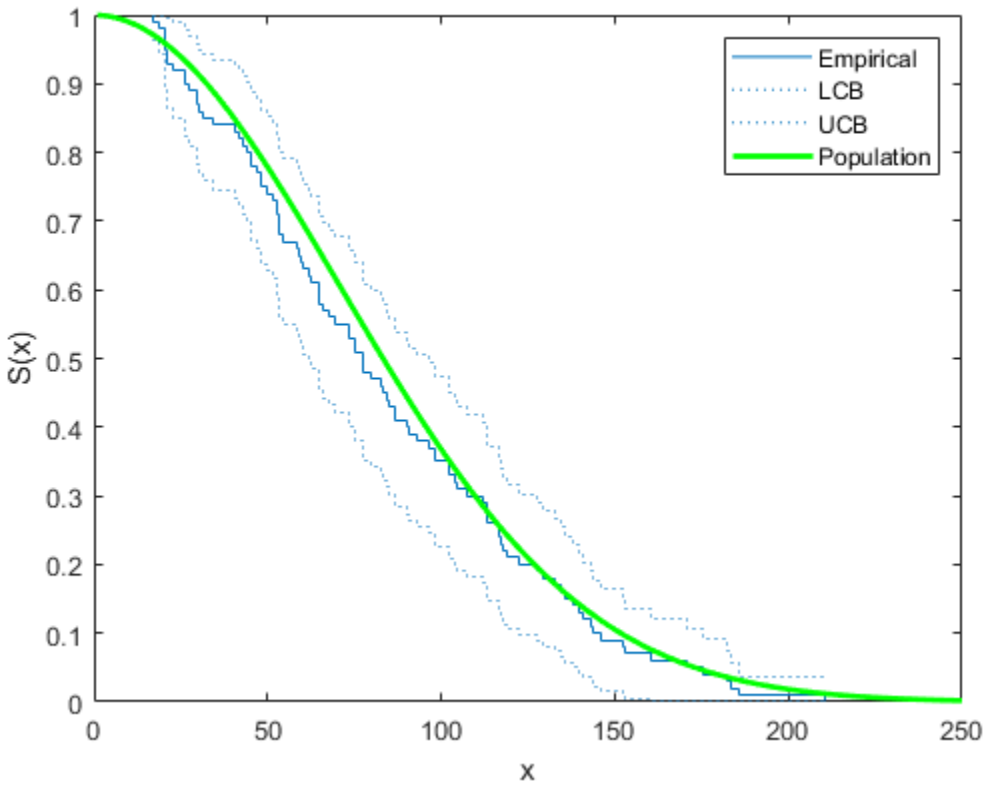
Plot the survivor function for the data with 99% confidence bounds.

```
ecdf(R,'Function','survivor','Alpha',0.01,'Bounds','on')
hold on
```



Fit the Weibull survivor function.

```
x = 1:1:250;
wblsurv = 1-cdf('weibull',x,100,2);
plot(x,wblsurv,'g-','LineWidth',2)
legend('Empirical','LCB','UCB','Population', ...
    'Location','northeast')
```



The survivor function based on the actual distribution is within the confidence bounds.

Input Arguments

y — Input data

vector

Input data, specified as a vector. For example, in survival or reliability analysis, data might be survival or failure times for each item or individual.

`ecdf` ignores `NaN` values in `y`. Additionally, any `NaN` values in the censoring vector (`'Censoring'`) or frequency vector (`'Frequency'`) cause `ecdf` to ignore the corresponding values in `y`.

Data Types: `single` | `double`

ax — Axes handle

handle

Axes handle for the figure `ecdf` plots to, specified as a handle.

For instance, if `h` is a handle for a figure, then `ecdf` can plot to that figure as follows.

Example: `ecdf(h,x)`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Censoring',c,'Function','cumulative hazard','Alpha',0.025,'Bounds','on'` specifies that `ecdf` returns the cumulative hazard function and plots the 97.5% confidence bounds, accounting for the censored data specified by vector `c`.

`Censoring` — Indicator of censored data

array of 0s (default) | vector of 0s and 1s

Indicator of censored data, specified as the comma-separated pair including `'Censoring'` and a Boolean array of the same size as `x`. Enter `1` for observations that are right-censored and `0` for observations that are fully observed. Default is all observations are fully observed.

`ecdf` ignores any `NaN` values in this censoring vector. Additionally, any `NaN` values in `y` or the frequency vector (`'Frequency'`) cause `ecdf` to ignore the corresponding values in the censoring vector.

Example: If vector `cdata` stores the censored data information, enter `'Censoring',cdata`.

Data Types: `logical`

`Frequency` — Frequency of observations

array of 1s (default) | vector of nonnegative scalars

Frequency of observations, specified as the comma-separated pair consisting of `'Frequency'` and a vector containing nonnegative integer counts. This vector is the same size as the vector `x`. The `j`th element of this vector gives the number of times the `j`th element of `x` was observed. Default is one observation per element of `x`.

`ecdf` ignores any `NaN` values in this frequency vector. Additionally, any `NaN` values in `y` or the censoring vector (`'Censoring'`) cause `ecdf` to ignore the corresponding values in the frequency vector.

Example: If `failurefreq` is a vector of frequencies, enter `'Frequency',failurefreq`

Data Types: `single` | `double`

`Alpha` — Significance level

0.05 (default) | scalar value in the range (0,1)

Significance level for the confidence interval of the evaluated function, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value between in the range (0,1). Default is 0.05 for 95% confidence. For a given value `alpha`, the confidence level is `100(1-alpha)`%.

For instance, for a 99% confidence interval, you can specify the alpha value as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Function` — Type of function returned

`'cdf'` (default) | `'survivor'` | `'cumulative hazard'`

Type of function that `ecdf` evaluates and returns, specified as the comma-separated pair consisting of `'Function'` and one of the following.

`'cdf'`	Default. Cumulative distribution function.
`'survivor'`	Survivor function.
`'cumulative hazard'`	Cumulative hazard function.

Example: `'Function','cumulative hazard'`

Bounds — Indicator for including bounds
`'off'` (default) | `'on'`

Indicator for including bounds, specified as the comma-separated pair consisting of `'Bounds'` and one of the following.

`'off'`	Default. Specify to omit bounds.
`'on'`	Specify to include bounds.

Note This name-value argument is used only for plotting.

Example: `'Bounds','on'`

Output Arguments

`f` — Function values
column vector

Function values evaluated at the points in `x`, returned as a column vector.

`x` — Sorted observed points
column vector

Sorted observed points in the data vector `y`, returned as a column vector.

`ecdf` sorts `y`, removes duplicate values in the sorted `y`, and saves the results to the output `x`. The output `x` includes the minimum value of `y` as its first two values. These two values are useful for plotting the outputs of `ecdf` using the `stairs` function.

`flo` — Lower confidence bound
column vector

Lower confidence bound for the evaluated function, returned as a column vector. `ecdf` computes the confidence bounds using Greenwood's formula on page 33-1279. They are not simultaneous confidence bounds.

`fup` — Upper confidence bound
column vector

Upper confidence bound for the evaluated function, returned as a column vector. `ecdf` computes the confidence bounds using Greenwood's formula on page 33-1279. They are not simultaneous confidence bounds.

More About

Greenwood's Formula

Approximation for the variance of Kaplan-Meier estimator.

The variance estimate is given by

$$V(S(t)) = S^2(t) \sum_{t_i \le T} \frac{d_i}{r_i(r_i - d_i)},$$

where r_i is the number at risk at time t_i, and d_i is the number of failures at time t_i.

References

[1] Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2003.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Plotting is not supported.
- Names in name-value pair arguments must be compile-time constants.
- Values in the `'Function'` and `'Bounds'` name-value pair arguments must also be compile-time constants. For example, to use the `'Function','survivor'` name-value pair argument in the generated code, include `{coder.Constant('Function'),coder.Constant('survivor')}` in the `-args` value of `codegen`.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`cdfplot` | `ecdfhist`

Topics

"Hazard and Survivor Functions for Different Groups" on page 14-16
"Survivor Functions for Two Groups" on page 14-22
"What Is Survival Analysis?" on page 14-2
"Kaplan-Meier Method" on page 14-10

Introduced before R2006a

ecdfhist

Histogram based on empirical cumulative distribution function

Syntax

```
[n,c] = ecdfhist(f,x)
[n,c] = ecdfhist(f,x,m)

n = ecdfhist(f,x,centers)

ecdfhist(___)
```

Description

`[n,c] = ecdfhist(f,x)` returns the heights, `n`, of histogram bars for 10 equally spaced bins and the position of the bin centers, `c`.

`ecdfhist` computes the bar heights from the increases in the empirical cumulative distribution function, `f`, at evaluation points, `x`. It normalizes the bar heights so that the area of the histogram is equal to 1. In contrast, `histogram` produces bars with heights representing bin counts.

`[n,c] = ecdfhist(f,x,m)` returns the histogram bars using `m` bins.

`n = ecdfhist(f,x,centers)` returns the heights of the histogram bars with bin centers specified by `centers`.

`ecdfhist(___)` plots the histogram bars.

Examples

Return Histogram Bar Heights and Bin Centers

Compute the histogram bar heights based on the empirical cumulative distribution function.

Generate failure times from a Birnbaum-Saunders distribution.

```
rng('default') % for reproducibility
failuretime = random('birnbaumsaunders',0.3,1,100,1);
```

Assuming that the end of the study is at time 0.9, mark the generated failure times that are larger than 0.9 as censored data and store that information in a vector.

```
T = 0.9;
cens = (failuretime>T);
```

Compute the empirical cumulative distribution function for the data.

```
[f,x] = ecdf(failuretime,'censoring',cens);
```

Now, find the bar heights of the histogram using the cumulative distribution function estimate.

```
[n,c] = ecdfhist(f,x);
[n' c']

ans = 10×2

    2.3529    0.0715
    1.7647    0.1565
    1.4117    0.2415
    1.5294    0.3265
    1.0588    0.4115
    0.4706    0.4965
    0.4706    0.5815
    0.9412    0.6665
    0.2353    0.7515
    0.2353    0.8365
```

Return Bar Heights and Bin Centers for a Given Number of Bins

Compute the bar heights for six bins using the empirical cumulative distribution function and also return the bin centers.

Generate failure times from a Birnbaum-Saunders distribution.

```
rng('default') % for reproducibility
failuretime = random('birnbaumsaunders',0.3,1,100,1);
```

Assuming that the end of the study is at time 0.9, mark the generated failure times that are larger than 0.9 as censored data and store that information in a vector.

```
T = 0.9;
cens = (failuretime>T);
```

First, compute the empirical cumulative distribution function for the data.

```
[f,x] = ecdf(failuretime,'censoring',cens);
```

Now, estimate the histogram with six bins using the cumulative distribution function estimate.

```
[n,c] = ecdfhist(f,x,6);
[n' c']

ans = 6×2

    1.9764    0.0998
    1.7647    0.2415
    1.1294    0.3831
    0.4235    0.5248
    0.7764    0.6665
    0.2118    0.8081
```

Draw Histogram for Given Bin Centers

Draw the histogram of the empirical cumulative distribution histogram for specified bin centers.

Generate failure times from a Birnbaum-Saunders distribution.

```
rng default;  % For reproducibility
failuretime = random('birnbaumsaunders',0.3,1,100,1);
```

Assuming that the end of the study is at time 0.9, mark the generated failure times that are larger than 0.9 as censored data and store that information in a vector.

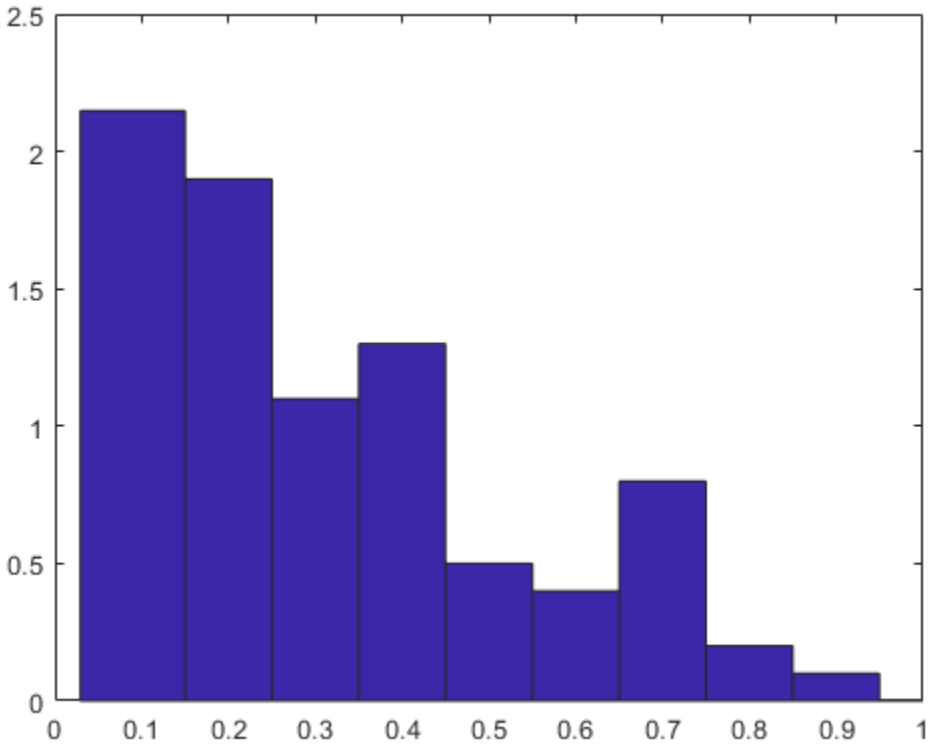
```
T = 0.9;
cens = (failuretime>T);
```

Define bin centers.

```
centers = 0.1:0.1:1;
```

Compute the empirical cumulative distribution function for the data and draw the histogram for specified bin centers.

```
[f,x] = ecdf(failuretime,'censoring',cens);
ecdfhist(f,x,centers)
axis([0 1 0 2.5])
```



Compare Histogram with Known Probability Distribution Function

Generate right-censored survival data and compare the histogram from cumulative distribution function with the known probability distribution function.

Generate failure times from an exponential distribution with mean failure time of 15.

```
rng default; % For reproducibility
y = exprnd(15,75,1);
```

Generate drop-out times from an exponential distribution with mean failure time of 30.

```
d = exprnd(30,75,1);
```

Record the minimum of these times as the observed failure times.

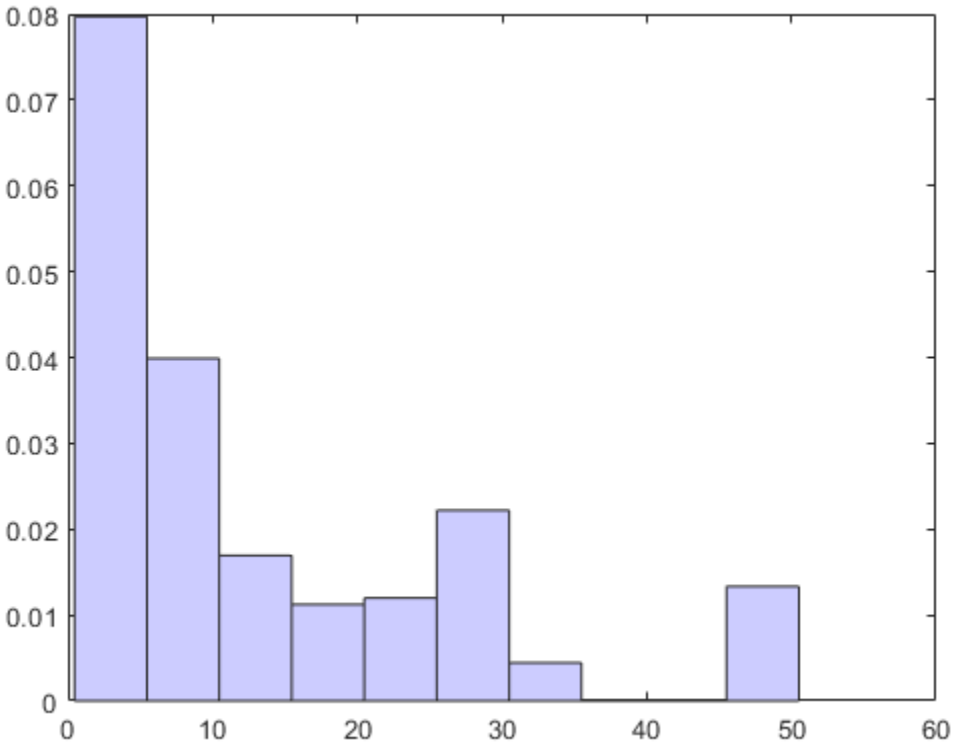
```
t = min(y,d);
```

Generate censoring by finding the generated failure times that are greater than the drop-out times.

```
censored = (y>d);
```

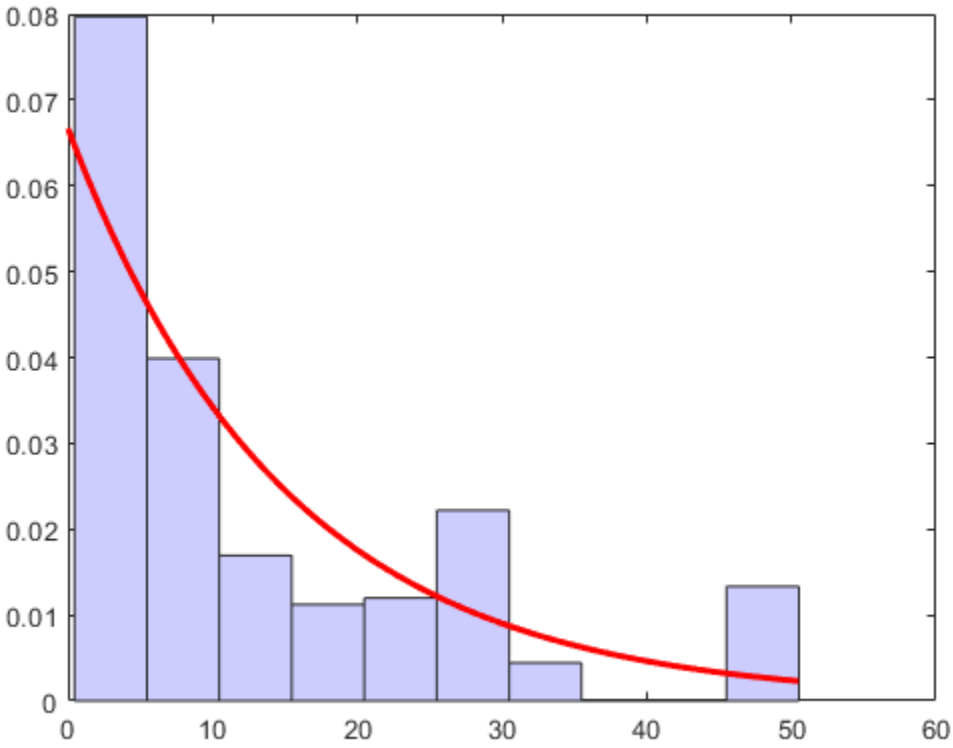
Calculate the empirical cdf and plot a histogram using the empirical cumulative distribution function.

```
[f,x] = ecdf(t,'censoring',censored);
ecdfhist(f,x)
h = findobj(gca,'Type','patch');
h.FaceColor = [.8 .8 1];
hold on
```



Superimpose a plot of the known population pdf.

```
xx = 0:.1:max(t);
yy = exp(-xx/15)/15;
plot(xx,yy,'r-','LineWidth',2)
hold off
```



Input Arguments

`f` — Empirical cdf values

vector

Empirical cdf values at given evaluation points, `x`, specified as a vector.

For instance, you can use `ecdf` to obtain the empirical cdf values and enter them in `ecdfhist` as follows.

Example: `[f,x] = ecdf(failure); ecdfhist(f,x);`

Data Types: `single` | `double`

`x` — Evaluation points

vector

Evaluation points at which empirical cdf values, `f`, are calculated, specified as a vector.

For instance, you can use `ecdf` to obtain the empirical cdf values and enter them in `ecdfhist` as follows.

Example: `[f,x] = ecdf(failure); ecdfhist(f,x);`

Data Types: `single` | `double`

`m` — Number of bins
scalar

Number of bins, specified as a scalar.

For instance, you can draw a histogram with 8 bins as follows.

Example: `ecdfhist(f,x,8)`

Data Types: `single` | `double`

`centers` — Center points of bins
vector

Center points of bins, specified as a vector.

Example: `centers = 2:2:10; ecdfhist(f,x,centers);`

Data Types: `single` | `double`

Output Arguments

`n` — Heights of histogram bars
row vector

Heights of histogram bars `ecdfhist` calculates based on the empirical cdf values, returned as a row vector.

`c` — Position of bin centers
row vector

Position of bin centers, returned as a row vector.

See Also

`ecdf` | `histc` | `histogram`

Topics
“Nonparametric and Empirical Probability Distributions” on page 5-30

Introduced before R2006a

edge

Package:

Classification edge for generalized additive model (GAM)

Syntax

```
e = edge(Mdl,Tbl,ResponseVarName)
e = edge(Mdl,Tbl,Y)
e = edge(Mdl,X,Y)
e = edge(___,Name,Value)
```

Description

`e = edge(Mdl,Tbl,ResponseVarName)` returns the "Classification Edge" on page 33-1294 (`e`) for the generalized additive model `Mdl` using the predictor data in `Tbl` and the true class labels in `Tbl.ResponseVarName`.

`e = edge(Mdl,Tbl,Y)` uses the predictor data in table `Tbl` and the true class labels in `Y`.

`e = edge(Mdl,X,Y)` uses the predictor data in matrix `X` and the true class labels in `Y`.

`e = edge(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify observation weights and whether to include interaction terms in computations.

Examples

Estimate Test Sample Classification Margins and Edge

Estimate the test sample classification margins and edge of a generalized additive model. The test sample margins are the observed true class scores minus the false class scores, and the test sample edge is the mean of the margins.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains two sepal and two petal measurements for versicolor and virginica irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
X = meas(inds,:);
Y = species(inds,:);
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(XTrain,YTrain,'ClassNames',{'versicolor','virginica'});
```

`Mdl` is a `ClassificationGAM` model object.

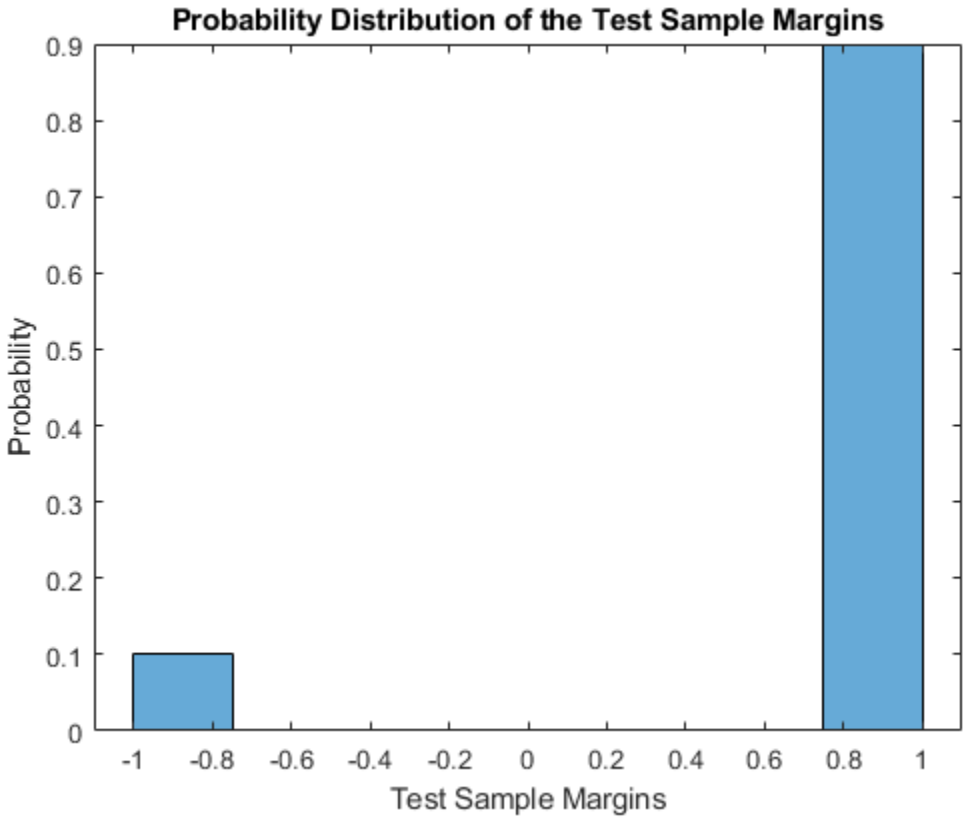
Estimate the test sample classification margins and edge.

```
m = margin(Mdl,XTest,YTest);
e = edge(Mdl,XTest,YTest)
```

```
e = 0.8000
```

Display the histogram of the test sample classification margins.

```
histogram(m,length(unique(m)),'Normalization','probability')
xlabel('Test Sample Margins')
ylabel('Probability')
title('Probability Distribution of the Test Sample Margins')
```



Estimate Test Sample Weighted Edge

Estimate the test sample weighted edge (the weighted average of margins) of a generalized additive model.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains two sepal and two petal measurements for versicolor and virginica irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
idx1 = strcmp(species,'versicolor') | strcmp(species,'virginica');
X = meas(idx1,:);
Y = species(idx1,:);
```

Suppose that the quality of some measurements is lower because they were measured with older technology. To simulate this effect, add noise to a random subset of 20 measurements.

```
rng('default') % For reproducibility
idx2 = randperm(size(X,1),20);
X(idx2,:) = X(idx2,:) + 2*randn(20,size(X,2));
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(XTrain,YTrain,'ClassNames',{'versicolor','virginica'});
```

`Mdl` is a `ClassificationGAM` model object.

Estimate the test sample edge.

```
e = edge(Mdl,XTest,YTest)
```

```
e = 0.8000
```

The average margin is approximately 0.80.

One way to reduce the effect of the noisy measurements is to assign them less weight than the other observations. Define a weight vector that gives the higher quality observations twice the weight of the other observations.

```
n = size(X,1);
weights = ones(size(X,1),1);
weights(idx2) = 0.5;
weightsTrain = weights(trainInds);
weightsTest = weights(testInds);
```

Train a GAM using the predictors `XTrain`, class labels `YTrain`, and weights `weightsTrain`.

```
Mdl_W = fitcgam(XTrain,YTrain,'Weights',weightsTrain,...
    'ClassNames',{'versicolor','virginica'});
```

Estimate the test sample weighted edge using the weighting scheme.

```
e_W = edge(Mdl_W,XTest,YTest,'Weights',weightsTest)
```

```
e_W = 0.8770
```

The weighted average margin is approximately 0.88. This result indicates that, on average, the labels from weighted classifier labels have higher confidence.

Compare GAMs by Examining Test Sample Margins and Edge

Compare a GAM with linear terms to a GAM with both linear and interaction terms by examining the test sample margins and edge. Based solely on this comparison, the classifier with the highest margins and edge is the best model.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'Holdout',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM that contains both linear and interaction terms for predictors. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
Mdl = fitcgam(XTrain,YTrain,'Interactions','all','MaxPValue',0.05)
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 3.0398
             Interactions: [561x2 double]
          NumObservations: 246


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. `Mdl` includes all available interaction terms.

Estimate the test sample margins and edge for `Mdl`.

```
M = margin(Mdl,XTest,YTest);
E = edge(Mdl,XTest,YTest)
```

```
E = 0.7848
```

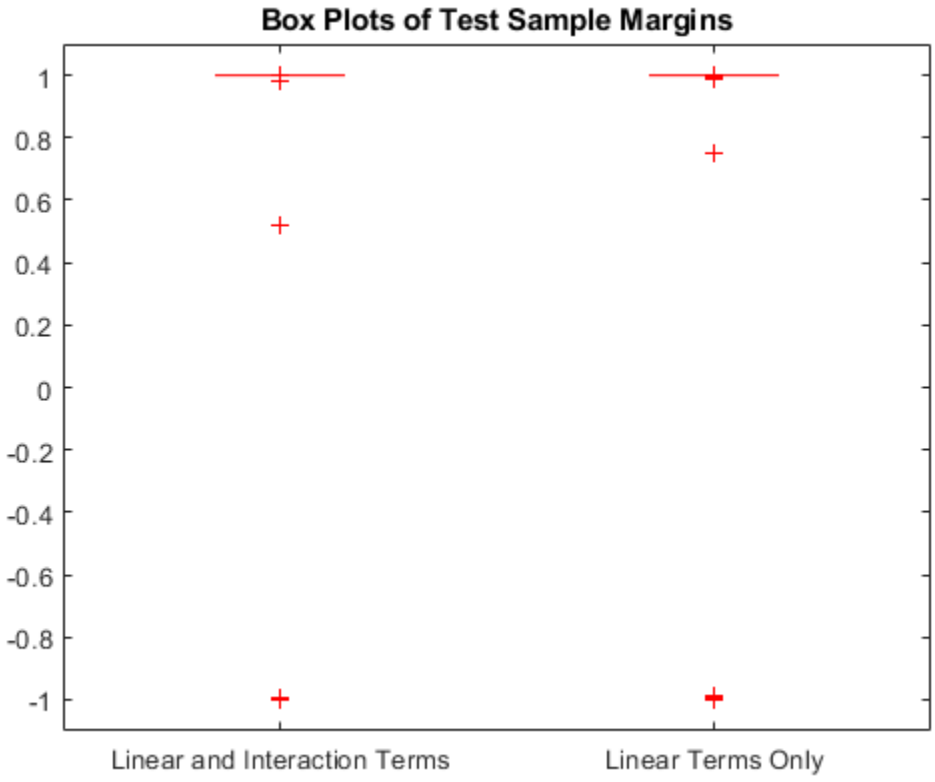
Estimate the test sample margins and edge for `Mdl` without including interaction terms.

```
M_nointeractions = margin(Mdl,XTest,YTest,'IncludeInteractions',false);
E_nointeractions = edge(Mdl,XTest,YTest,'IncludeInteractions',false)
```

```
E_nointeractions = 0.7871
```

Display the distributions of the margins using box plots.

```
boxplot([M M_nointeractions],'Labels',{'Linear and Interaction Terms','Linear Terms Only'})
title('Box Plots of Test Sample Margins')
```



The margins `M` and `M_nointeractions` have a similar distribution, but the test sample edge of the classifier with only linear terms is larger. Classifiers that yield relatively large margins are preferred.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `CompactClassificationGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `CompactClassificationGAM` model object.

- If you trained `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table (`Tbl`).
- If you trained `Mdl` using sample data contained in a matrix, then the input data for `edge` must also be in a matrix (`X`).

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`Tbl` must contain all the predictors used to train `Mdl`. Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights.

- The response variable must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.) If the response variable in `Tbl` has the same name as the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.
- The weight values must be a numeric vector. You must specify the observation weights in `Tbl` by using `'Weights'`.

If you trained `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X` or `Tbl`.

`Y` must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

If you trained `Mdl` using sample data contained in a matrix, then the input data for `edge` must also be in a matrix.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'IncludeInteractions',false,'Weights',w` specifies to exclude interaction terms from the model and to use the observation weights `w`.

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `'IncludeInteractions'` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored in `Tbl.W`, then specify it as `'W'`.

`edge` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

See Also
`loss` | `margin` | `predict` | `resubEdge`

Topics
“Train Generalized Additive Model for Binary Classification” on page 12-77

Introduced in R2021a

edge

Edge of *k*-nearest neighbor classifier

Syntax

```
E = edge(mdl,tbl,ResponseVarName)
E = edge(mdl,tbl,Y)
E = edge(mdl,X,Y)
E = edge( ___ ,'Weights',weights)
```

Description

`E = edge(mdl,tbl,ResponseVarName)` returns the classification edge for `mdl` with data `tbl` and classification `tbl.ResponseVarName`. If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName`.

The classification edge (`E`) is a scalar value that represents the mean of the classification margins on page 33-1297.

`E = edge(mdl,tbl,Y)` returns the classification edge for `mdl` with data `tbl` and classification `Y`.

`E = edge(mdl,X,Y)` returns the classification edge for `mdl` with data `X` and classification `Y`.

`E = edge( ___ ,'Weights',weights)` computes the edge with additional observation weights `weights`, using any of the input arguments in the previous syntaxes.

Examples

Edge Calculation

Create a *k*-nearest neighbor classifier for the Fisher iris data, where k = 5.

Load the Fisher iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create a classifier for five nearest neighbors.

```
mdl = fitcknn(X,Y,'NumNeighbors',5);
```

Examine the edge of the classifier for minimum, mean, and maximum observations classified as `'setosa'`, `'versicolor'`, and `'virginica'`, respectively.

```
NewX = [min(X);mean(X);max(X)];
Y = {'setosa';'versicolor';'virginica'};
E = edge(mdl,NewX,Y)
```

```
E = 1
```

All five nearest neighbors of each NewX point classify as the corresponding Y entry.

Input Arguments

mdl — *k*-nearest neighbor classifier model

ClassificationKNN object

k-nearest neighbor classifier model, specified as a ClassificationKNN object.

tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of tbl corresponds to one observation, and each column corresponds to one predictor variable. Optionally, tbl can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If tbl contains the response variable used to train mdl, then you do not need to specify ResponseVarName or Y.

If you train mdl using sample data contained in a table, then the input data for edge must also be in a table.

Data Types: table

ResponseVarName — Response variable name

name of a variable in tbl

Response variable name, specified as the name of a variable in tbl. If tbl contains the response variable used to train mdl, then you do not need to specify ResponseVarName.

You must specify ResponseVarName as a character vector or string scalar. For example, if the response variable is stored as tbl.response, then specify it as 'response'. Otherwise, the software treats all columns of tbl, including tbl.response, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: char | string

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of X represents one observation, and each column represents one variable.

Data Types: single | double

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. Each row of Y represents the classification of the corresponding row of X.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `tbl`

Observation weights, specified as a numeric vector or the name of a variable in `tbl`.

If you specify `weights` as a numeric vector, then the size of `weights` must be equal to the number of rows in `X` or `tbl`.

If you specify `weights` as the name of a variable in `tbl`, then the name must be a character vector or string scalar. For example, if the weights are stored as `tbl.w`, then specify `weights` as `'w'`. Otherwise, the software treats all columns of `tbl`, including `tbl.w`, as predictors.

If you specify `weights`, then the `edge` function weights the observation in each row of `X` or `tbl` with the corresponding weight in `weights`.

Example: `'Weights','w'`

Data Types: `single` | `double` | `char` | `string`

More About

Margin

The classification margin for each observation is the difference between the classification score for the true class and the maximal classification score for the false classes.

The classification margins form a column vector with the same number of rows as `X` or `tbl`.

Score

The score of a classification is the posterior probability of the classification. The posterior probability is the number of neighbors with that classification divided by the number of neighbors. For a more detailed definition that includes weights and prior probabilities, see “Posterior Probability” on page 33-4786.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.
- `edge` executes on a GPU in these cases only:
  - The input argument `X` is a `gpuArray`.

- The input argument `tbl` contains `gpuArray` elements.
- The input argument `mdl` was fitted with GPU array input arguments.

See Also

`ClassificationKNN` | `fitcknn` | `loss` | `margin`

Topics

“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2012a

edge

Class: `ClassificationLinear`

Classification edge for linear classification models

Syntax

```
e = edge(Mdl,X,Y)

e = edge(Mdl,Tbl,ResponseVarName)
e = edge(Mdl,Tbl,Y)

e = edge( ___ ,Name,Value)
```

Description

`e = edge(Mdl,X,Y)` returns the classification edges on page 33-1306 for the binary, linear classification model `Mdl` using predictor data in `X` and corresponding class labels in `Y`. `e` contains a classification edge for each regularization strength in `Mdl`.

`e = edge(Mdl,Tbl,ResponseVarName)` returns the classification edges for the trained linear classifier `Mdl` using the predictor data in `Tbl` and the class labels in `Tbl.ResponseVarName`.

`e = edge(Mdl,Tbl,Y)` returns the classification edges for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

`e = edge( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that columns in the predictor data correspond to observations or supply observation weights.

Input Arguments

`Mdl` — Binary, linear classification model

`ClassificationLinear` model object

Binary, linear classification model, specified as a `ClassificationLinear` model object. You can create a `ClassificationLinear` model object using `fitclinear`.

`X` — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix. This orientation of `X` indicates that rows correspond to individual observations, and columns correspond to individual predictor variables.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table.

ResponseVarName — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of observations in `X` or `Tbl`.
- If you specify `Weights` as the name of a variable in `Tbl`, then the name must be a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, then for each regularization strength, `edge` computes the weighted classification edge on page 33-1306 and normalizes weights to sum up to the value of the prior probability in the respective class.

Data Types: `double` | `single`

Output Arguments

`e` — Classification edges
numeric scalar | numeric row vector

Classification edges on page 33-1306, returned as a numeric scalar or row vector.

`e` is the same size as `Mdl.Lambda`. `e(j)` is the classification edge of the linear classification model trained using the regularization strength `Mdl.Lambda(j)`.

Examples

Estimate Test-Sample Edge

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. Specify to holdout 30% of the observations. Optimize the objective function using SpaRSA.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','Holdout',0.30);
CMdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `ClassificationLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Estimate the training- and test-sample edges.

```
eTrain = edge(CMdl,X(trainIdx,:),Ystats(trainIdx))
```

```
eTrain = 15.6660
```

```
eTest = edge(CMdl,X(testIdx,:),Ystats(testIdx))
```

```
eTest = 15.4767
```

Feature Selection Using Test-Sample Edges

One way to perform feature selection is to compare test-sample edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages. For quicker execution time, orient the predictor data so that individual observations correspond to columns.

```
Ystats = Y == 'stats';
X = X';
rng(1); % For reproducibility
```

Create a data partition which holds out 30% of the observations for testing.

```
Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition); % Test-set indices
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

`Partition` is a `cvpartition` object that defines the data set partition.

Randomly choose half of the predictor variables.

```
p = size(X,1); % Number of predictors
idxPart = randsample(p,ceil(0.5*p));
```

Train two binary, linear classification models: one that uses the all of the predictors and one that uses half of the predictors. Optimize the objective function using SpaRSA, and indicate that observations correspond to columns.

```
CVMdl = fitclinear(X,Ystats,'CVPartition',Partition,'Solver','sparsa',...
    'ObservationsIn','columns');
PCVMdl = fitclinear(X(idxPart,:),Ystats,'CVPartition',Partition,'Solver','sparsa',...
    'ObservationsIn','columns');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinear` models.

Extract the trained `ClassificationLinear` models from the cross-validated models.

```
CMdl = CVMdl.Trained{1};
PCMdl = PCVMdl.Trained{1};
```

Estimate the test sample edge for each classifier.

```
fullEdge = edge(CMdl,XTest,YTest,'ObservationsIn','columns')
```

```
fullEdge = 15.4767
```

```
partEdge = edge(PCMdl,XTest(idxPart,:),YTest,'ObservationsIn','columns')
```

```
partEdge = 13.4458
```

Based on the test-sample edges, the classifier that uses all of the predictors is the better model.

Find Good Lasso Penalty Using Edge

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare test-sample edges.

Load the NLP data set. Preprocess the data as in “Feature Selection Using Test-Sample Edges” on page 33-1302.

```
load nlpdata
Ystats = Y == 'stats';
X = X';

Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition);
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,11);
```

Train binary, linear classification models that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'CVPartition',Partition,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract the trained linear classification model.

```
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the test-sample edges.

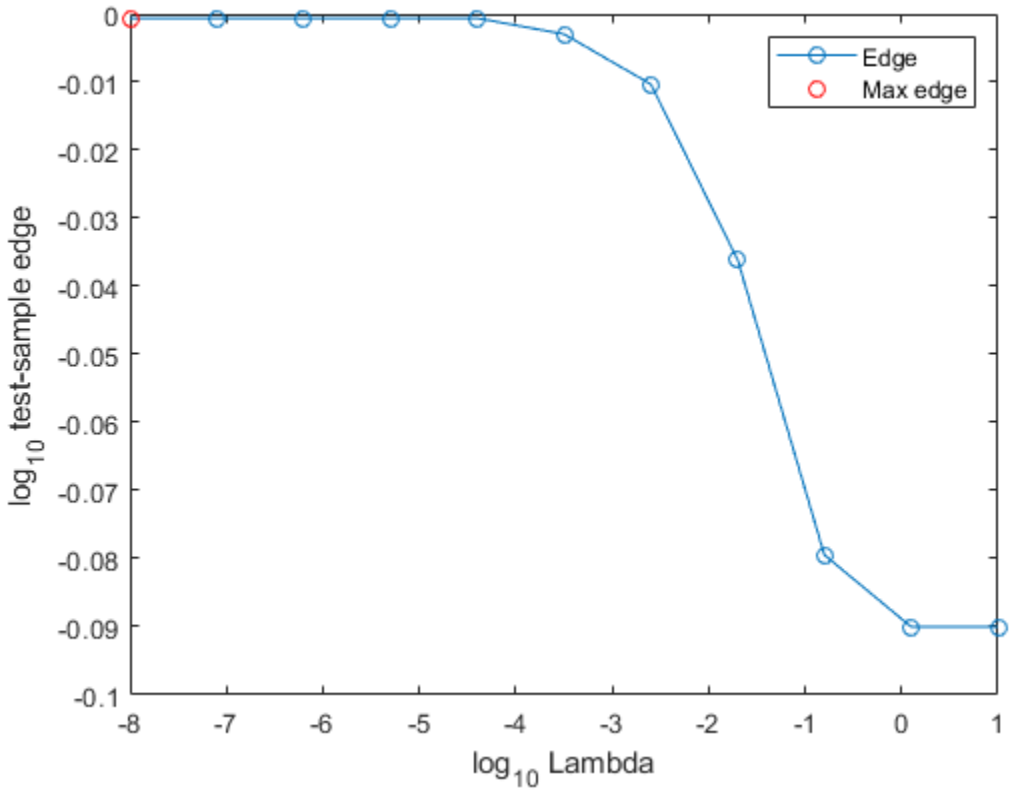
```
e = edge(Mdl,X(:,testIdx),Ystats(testIdx),'ObservationsIn','columns')
```

```
e = 1×11

    0.9986    0.9986    0.9986    0.9986    0.9986    0.9932    0.9766    0.9203    0.8328    0.8
```

Because there are 11 regularization strengths, `e` is a 1-by-11 vector of edges.

Plot the test-sample edges for each regularization strength. Identify the regularization strength that maximizes the edges over the grid.

```
figure;
plot(log10(Lambda),log10(e),'-o')
[~, maxEIdx] = max(e);
maxLambda = Lambda(maxEIdx);
hold on
plot(log10(maxLambda),log10(e(maxEIdx)),'ro');
ylabel('log_{10} test-sample edge')
xlabel('log_{10} Lambda')
legend('Edge','Max edge')
hold off
```



Several values of `Lambda` yield similarly high edges. Higher values of lambda lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the edge starts decreasing.

```
LambdaFinal = Lambda(5);
```

Train a linear classification model using the entire data set and specify the regularization strength yielding the maximal edge.

```
MdlFinal = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',LambdaFinal);
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Algorithms

By default, observation weights are prior class probabilities. If you supply weights using `Weights`, then the software normalizes them to sum to the prior probabilities in the respective classes. The software uses the normalized weights to estimate the weighted edge.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `edge` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationLinear` | `fitclinear` | `margin` | `predict`

Introduced in R2016a

edge

Classification edge

Syntax

```
E = edge(obj,X,Y)
E = edge(obj,X,Y,Name,Value)
```

Description

`E = edge(obj,X,Y)` returns the classification edge for `obj` with data `X` and classification `Y`.

`E = edge(obj,X,Y,Name,Value)` computes the edge with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier of class `ClassificationDiscriminant` or `CompactClassificationDiscriminant`, typically constructed with `fitcdiscr`.

`X`

Matrix where each row represents an observation, and each column represents a predictor. The number of columns in `X` must equal the number of predictors in `obj`.

`Y`

Class labels, with the same data type as exists in `obj`. The number of elements of `Y` must equal the number of rows of `X`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`weights`

Observation weights, a numeric vector of length `size(X,1)`. If you supply weights, `edge` computes the weighted classification edge.

Default: `ones(size(X,1),1)`

Output Arguments

`E`

Edge, a scalar representing the weighted average value of the margin.

Examples

Compute the classification edge and margin for the Fisher iris data, trained on its first two columns of data, and view the last 10 entries:

```
load fisheriris
X = meas(:,1:2);
obj = fitcdiscr(X,species);
E = edge(obj,X,species)

E =
    0.4980

M = margin(obj,X,species);
M(end-10:end)

ans =
    0.6551
    0.4838
    0.6551
   -0.5127
    0.5659
    0.4611
    0.4949
    0.1024
    0.2787
   -0.1439
   -0.4444
```

The classifier trained on all the data is better:

```
obj = fitcdiscr(meas,species);
E = edge(obj,meas,species)

E =
    0.9454

M = margin(obj,meas,species);
M(end-10:end)

ans =
    0.9983
    1.0000
    0.9991
    0.9978
    1.0000
    1.0000
    0.9999
    0.9882
    0.9937
    1.0000
    0.9649
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are class prior probabilities. If you supply additional weights, those weights are normalized to sum to the prior probabilities in the respective classes, and are then used to compute the weighted average.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes.

The classification margin is a column vector with the same number of rows as in the matrix `X`. A high value of margin indicates a more reliable prediction than a low value.

Score (discriminant analysis)

For discriminant analysis, the score of a classification is the posterior probability of the classification. For the definition of posterior probability in discriminant analysis, see “Posterior Probability” on page 20-6.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also
`ClassificationDiscriminant` | `fitcdiscr` | `loss` | `margin` | `predict`

Topics
“Discriminant Analysis Classification” on page 20-2

edge

Package:

Classification edge for multiclass error-correcting output codes (ECOC) model

Syntax

```
e = edge(Mdl,tbl,ResponseVarName)
e = edge(Mdl,tbl,Y)

e = edge(Mdl,X,Y)

e = edge(___,Name,Value)
```

Description

`e = edge(Mdl,tbl,ResponseVarName)` returns the classification edge on page 33-1318 (`e`) for the trained multiclass error-correcting output codes (ECOC) classifier `Mdl` using the predictor data in table `tbl` and the class labels in `tbl.ResponseVarName`.

`e = edge(Mdl,tbl,Y)` returns the classification edge for the classifier `Mdl` using the predictor data in table `tbl` and the class labels in vector `Y`.

`e = edge(Mdl,X,Y)` returns the classification edge (`e`) for the classifier `Mdl` using the predictor data in matrix `X` and the class labels in vector `Y`.

`e = edge(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify a decoding scheme, binary learner loss function, and verbosity level.

Examples

Test-Sample Edge of ECOC Model

Compute the test-sample classification edge of an ECOC model with SVM binary classifiers.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Specify a 30% holdout sample for testing, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.30,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training data.

Compute the test-sample edge.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
e = edge(Mdl,XTest,YTest)
```

```
e = 0.4574
```

The average of the test-sample margins is approximately 0.46.

Mean of Test-Sample Weighted Margins of ECOC Model

Compute the mean of the test-sample weighted margins of an ECOC model.

Suppose that the observations in a data set are measured sequentially, and that the last 75 observations have better quality due to a technology upgrade. Incorporate this advancement by giving the better quality observations more weight than the other observations.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Define a weight vector that assigns twice as much weight to the better quality observations.

```
n = size(X,1);
weights = [ones(n-75,1);2*ones(75,1)];
```

Train an ECOC model using SVM binary classifiers. Specify a 30% holdout sample and the weighting scheme. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.30,'Weights',weights,...
    'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a trained `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` classifier that the software trained using the training data.

Compute the test-sample weighted edge using the weighting scheme.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
wTest = weights(testInds,:);
e = edge(Mdl,XTest,YTest,'Weights',wTest)
```

```
e = 0.4797
```

The average weighted margin of the test sample is approximately 0.48.

Select ECOC Model Features by Comparing Test-Sample Edges

Perform feature selection by comparing test-sample edges from multiple models. Based solely on this comparison, the classifier with the greatest edge is the best classifier.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Partition the data set into training and test sets. Specify a 30% holdout sample for testing.

```
Partition = cvpartition(Y,'Holdout',0.30);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`Partition` defines the data set partition.

Define these two data sets:

- `fullX` contains all predictors.
- `partX` contains the petal dimensions only.

```
fullX = X;
partX = X(:,3:4);
```

Train an ECOC model using SVM binary classifiers for each predictor set. Specify the partition definition, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
fullPMdl = fitcecoc(fullX,Y,'CVPartition',Partition,'Learners',t,...
    'ClassNames',classOrder);
partPMdl = fitcecoc(partX,Y,'CVPartition',Partition,'Learners',t,...
    'ClassNames',classOrder);
fullMdl = fullPMdl.Trained{1};
partMdl = partPMdl.Trained{1};
```

`fullPMdl` and `partPMdl` are `ClassificationPartitionedECOC` models. Each model has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the corresponding training set.

Calculate the test-sample edge for each classifier.

```
fullEdge = edge(fullMdl,XTest,YTest)
```

```
fullEdge = 0.4574
```

```
partEdge = edge(partMdl,XTest(:,3:4),YTest)
```

```
partEdge = 0.4839
```

`partMdl` yields an edge value comparable to the value for the more complex model `fullMdl`.

Input Arguments

`Mdl` — Full or compact multiclass ECOC model

`ClassificationECOC` model object | `CompactClassificationECOC` model object

Full or compact multiclass ECOC model, specified as a `ClassificationECOC` or `CompactClassificationECOC` model object.

To create a full or compact ECOC model, see `ClassificationECOC` or `CompactClassificationECOC`.

`tbl` — Sample data

table

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain additional columns for the response variable and observation weights. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you train `Mdl` using sample data contained in a `table`, then the input data for `edge` must also be in a table.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `tbl.y`, then specify `ResponseVarName` as `'y'`. Otherwise, the software treats all columns of `tbl`, including `tbl.y`, as predictors.

The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one variable. The variables in the columns of `X` must be the same as the variables that trained the classifier `Mdl`.

The number of rows in `X` must equal the number of rows in `Y`.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `double` | `single`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `Y` must have the same data type as `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The number of rows in `Y` must equal the number of rows in `tbl` or `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `edge(Mdl,X,Y,'BinaryLoss','exponential','Decoding','lossbased')` specifies an exponential binary learner loss function and a loss-based decoding scheme for aggregating the binary losses.

BinaryLoss — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$

Value	Description	Score Domain	$g(y_j,s_j)$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_j s_j)/2$
`'hamming'`	Hamming	$[0,1]$ or $(-\infty,\infty)$	$[1 - \text{sign}(y_j s_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_j s_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_j s_j)/2$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_j s_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	$[0,1]$	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-1318.

Example: `'Decoding','lossbased'`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`. `Mdl.BinaryLearners` must contain `ClassificationLinear` models.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, you can experience a significant reduction in execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `tbl`. If you supply weights, `edge` computes the weighted classification edge on page 33-1318.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of observations in `X` or `tbl`. The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

If you specify `Weights` as the name of a variable in `tbl`, you must do so as a character vector or string scalar. For example, if the weights are stored as `tbl.w`, then specify `Weights` as `'w'`. Otherwise, the software treats all columns of `tbl`, including `tbl.w`, as predictors.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

e — Classification edge

numeric scalar | numeric vector

Classification edge on page 33-1318, returned as a numeric scalar or vector. `e` represents the weighted mean of the classification margins on page 33-1318.

If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `e` is a 1-by-L vector, where L is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`). The value `e(j)` is the edge for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.

Otherwise, `e` is a scalar value.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.

- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Tips

- To compare the margins or edges of several ECOC classifiers, use template objects to specify a common score transform function among the classifiers during training.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `edge` does not support tall `table` data when `Mdl` contains kernel or linear binary learners.

For more information, see “Tall Arrays”.

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `fitcecoc` | `loss` | `margin` | `predict` | `resubEdge`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

edge

Classification edge

Syntax

```
E = edge(ens,tbl,ResponseVarName)
E = edge(ens,tbl,Y)
E = edge(ens,X,Y)
E = edge( ___ ,Name,Value)
```

Description

`E = edge(ens,tbl,ResponseVarName)` returns the classification edge for `ens` with data `tbl` and classification `tbl.ResponseVarName`.

`E = edge(ens,tbl,Y)` returns the classification edge for `ens` with data `tbl` and classification `Y`.

`E = edge(ens,X,Y)` returns the classification edge for `ens` with data `X` and classification `Y`.

`E = edge( ___ ,Name,Value)` computes the edge with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

Input Arguments

`ens`

A classification ensemble constructed with `fitcensemble`, or a compact classification ensemble constructed with `compact`.

`tbl`

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all of the predictors used to train the model. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `ens` using sample data contained in a `table`, then the input data for this method must also be in a table.

`ResponseVarName`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `tbl`, including `Y`, as predictors when training the model.

`X`

A matrix where each row represents an observation, and each column represents a predictor. The number of columns in `X` must equal the number of predictors in `ens`.

If you trained `ens` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

`Y`

Class labels of observations in `tbl` or `X`. `Y` should be of the same type as the classification used to train `ens`, and its number of elements should equal the number of rows of `tbl` or `X`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `edge` uses only these learners for calculating loss.

Default: `1:NumTrained`

`mode`

Meaning of the output `E`:

- `'ensemble'` — `E` is a scalar value, the edge for the entire ensemble.
- `'individual'` — `E` is a vector with one element per trained learner.
- `'cumulative'` — `E` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

`UseObsForLearner`

A logical matrix of size `N`-by-`T`, where:

- `N` is the number of rows of `X`.
- `T` is the number of weak learners in `ens`.

When `UseObsForLearner(i,j)` is `true`, learner `j` is used in predicting the class of row `i` of `X`.

Default: `true(N,T)`

`weights`

Observation weights, a numeric vector of length `size(X,1)`. If you supply weights, `edge` computes weighted classification edge.

Default: `ones(size(X,1),1)`

Output Arguments

E

The classification edge, a vector or scalar depending on the setting of the `mode` name-value pair. Classification edge is weighted average classification margin.

Examples

Find Classification Edge of Training Data

Find the classification edge for some of the data used to train a boosted ensemble classifier.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an ensemble of 100 boosted classification trees using AdaBoostM1.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
```

Find the classification edge for the last few rows.

```
E = edge(ens,X(end-10:end,:),Y(end-10:end))
```

```
E = 8.3310
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `X`.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `ens.Prior`. If you supply weights in the `weights` name-value pair, those weights are used instead of class probabilities.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`edge` | `margin`

edge

Classification edge for naive Bayes classifier

Syntax

```
e = edge(Mdl,tbl,ResponseVarName)
e = edge(Mdl,tbl,Y)

e = edge(Mdl,X,Y)

e = edge(___,'Weights',Weights)
```

Description

`e = edge(Mdl,tbl,ResponseVarName)` returns the "Classification Edge" on page 33-1330 (`e`) for the naive Bayes classifier `Mdl` using the predictor data in table `tbl` and the class labels in `tbl.ResponseVarName`.

The classification edge (`e`) is a scalar value that represents the weighted mean of the "Classification Margins" on page 33-1330.

`e = edge(Mdl,tbl,Y)` returns the classification edge for `Mdl` using the predictor data in table `tbl` and the class labels in vector `Y`.

`e = edge(Mdl,X,Y)` returns the classification edge for `Mdl` using the predictor data in matrix `X` and the class labels in `Y`.

`e = edge(___,'Weights',Weights)` returns the classification edge with additional observation weights supplied in `Weights` using any of the input argument combinations in the previous syntaxes.

Examples

Estimate Test Sample Edge of Naive Bayes Classifier

Estimate the test sample edge (the classification margin average) of a naive Bayes classifier. The test sample edge is the average test sample difference between the estimated posterior probability for the predicted class and the posterior probability for the class with the next lowest posterior probability.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 105
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Estimate the test sample edge.

```
e = edge(Mdl,XTest,YTest)
```

```
e = 0.8658
```

The margin average is approximately `0.87`. This result suggests that the classifier labels predictors with high confidence.

Estimate Test Sample Weighted Edge of Naive Bayes Classifier

Estimate the test sample weighted edge (the weighted margin average) of a naive Bayes classifier. The test sample edge is the average test sample difference between the estimated posterior probability for the predicted class and the posterior probability for the class with the next lowest posterior probability. The weighted sample edge estimates the margin average when the software assigns a weight to each observation.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Suppose that some of the measurements are lower quality because they were measured with older technology. To simulate this effect, add noise to a random subset of 20 measurements.

```
idx = randperm(size(X,1),20);
X(idx,:) = X(idx,:) + 2*randn(20,size(X,2));
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Estimate the test sample edge.

```
e = edge(Mdl,XTest,YTest)
```

```
e = 0.5920
```

The average margin is approximately 0.59.

One way to reduce the effect of the noisy measurements is to assign them less weight than the other observations. Define a weight vector that gives the better quality observations twice the weight of the other observations.

```
n = size(X,1);
weights = ones(size(X,1),1);
weights(idx) = 0.5;
weightsTrain = weights(trainInds);
weightsTest = weights(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain`, class labels `YTrain`, and weights `weightsTrain`.

```
Mdl_W = fitcnb(XTrain,YTrain,'Weights',weightsTrain,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl_W` is a trained `ClassificationNaiveBayes` classifier.

Estimate the test sample weighted edge using the weighting scheme.

```
e_W = edge(Mdl_W,XTest,YTest,'Weights',weightsTest)
```

```
e_W = 0.6816
```

The weighted average margin is approximately 0.69. This result indicates that, on average, the weighted classifier labels predictors with higher confidence than the noise corrupted predictors.

Select Naive Bayes Classifier Features by Comparing Test Sample Edges

The classifier edge measures the average of the classifier margins. One way to perform feature selection is to compare test sample edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'Holdout',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Define these two training data sets:

- `fullXTrain` contains all predictors.
- `partXTrain` contains the 10 most important predictors.

```
fullXTrain = XTrain;
idx = fscmrmr(XTrain,YTrain);
partXTrain = XTrain(:,idx(1:10));
```

Train a naive Bayes classifier for each predictor set.

```
fullMdl = fitcnb(fullXTrain,YTrain);
partMdl = fitcnb(partXTrain,YTrain);
```

`fullMdl` and `partMdl` are trained `ClassificationNaiveBayes` classifiers.

Estimate the test sample edge for each classifier.

```
fullEdge = edge(fullMdl,XTest,YTest)
```

```
fullEdge = 0.5831
```

```
partEdge = edge(partMdl,XTest(:,idx(1:10)),YTest)
```

```
partEdge = 0.7593
```

The test sample edge of the classifier using the 10 most important predictors is larger.

Input Arguments

`Mdl` — Naive Bayes classification model

`ClassificationNaiveBayes` model object | `CompactClassificationNaiveBayes` model object

Naive Bayes classification model, specified as a `ClassificationNaiveBayes` model object or `CompactClassificationNaiveBayes` model object returned by `fitcnb` or `compact`, respectively.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed. Optionally, `tbl` can contain additional columns for the response variable and observation weights.

If you train `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table.

`ResponseVarName` — Response variable name

name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `y` is stored as `tbl.y`, then specify it as `'y'`. Otherwise, the software treats all columns of `tbl`, including `y`, as predictors.

If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of X corresponds to one observation (also known as an *instance* or *example*), and each column corresponds to one variable (also known as a *feature*). The variables in the columns of X must be the same as the variables that trained the Mdl classifier.

The length of Y and the number of rows of X must be equal.

Data Types: double | single

Y — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. Y must have the same data type as Mdl.ClassNames. (The software treats string arrays as cell arrays of character vectors.)

The length of Y must be equal to the number of rows of tbl or X.

Data Types: categorical | char | string | logical | single | double | cell

Weights — Observation weights
ones(size(X,1),1) (default) | numeric vector | name of a variable in tbl

Observation weights, specified as a numeric vector or the name of a variable in tbl. The software weighs the observations in each row of X or tbl with the corresponding weights in Weights.

If you specify Weights as a numeric vector, then the size of Weights must be equal to the number of rows of X or tbl.

If you specify Weights as the name of a variable in tbl, then the name must be a character vector or string scalar. For example, if the weights are stored as tbl.w, then specify Weights as 'w'. Otherwise, the software treats all columns of tbl, including tbl.w, as predictors.

Data Types: double | char | string

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

If you supply weights, then the software normalizes them to sum to the prior probability of their respective class. The software uses the normalized weights to compute the weighted mean.

When choosing among multiple classifiers to perform a task such as feature section, choose the classifier that yields the highest edge.

Classification Margins

The classification margin for each observation is the difference between the score for the true class and the maximal score for the false classes. Margins provide a classification confidence measure; among multiple classifiers, those that yield larger margins (on the same scale) are better.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For naive Bayes, the posterior probability that a classification is k for a given observation $(x_1,...,x_P)$ is

$$\widehat{P}(Y = k \,|\, x_1, .., x_P) = \frac{P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)}{P(X_1, ..., X_P)},$$

where:

- $P(X_1, ..., X_P \,|\, y = k)$ is the conditional joint density of the predictors given they are in class k. `Mdl.DistributionNames` stores the distribution names of the predictors.
- $\pi(Y = k)$ is the class prior probability distribution. `Mdl.Prior` stores the prior distribution.
- $P(X_1, .., X_P)$ is the joint density of the predictors. The classes are discrete, so $P(X_1, ..., X_P) = \sum_{k=1}^{K} P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)$.

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Classification Score

The naive Bayes score is the class posterior probability given the observation.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`ClassificationNaiveBayes` | `CompactClassificationNaiveBayes` | `fitcnb` | `loss` | `margin` | `predict` | `resubEdge` | `resubLoss`

Topics

“Naive Bayes Classification” on page 21-2

Introduced in R2014b

edge

Package:

Classification edge for neural network classifier

Syntax

```
e = edge(Mdl,Tbl,ResponseVarName)
e = edge(Mdl,Tbl,Y)

e = edge(Mdl,X,Y)

e = edge( ___ ,Name,Value)
```

Description

`e = edge(Mdl,Tbl,ResponseVarName)` returns the classification edge on page 33-1338 for the trained neural network classifier `Mdl` using the predictor data in table `Tbl` and the class labels in the `ResponseVarName` table variable.

`e` is returned as a scalar value that represents the mean of the classification margins.

`e = edge(Mdl,Tbl,Y)` returns the classification edge for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

`e = edge(Mdl,X,Y)` returns the classification edge for the trained neural network classifier `Mdl` using the predictor data `X` and the corresponding class labels in `Y`.

`e = edge( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that columns in the predictor data correspond to observations or supply observation weights.

Examples

Test Set Classification Edge of Neural Network

Calculate the test set classification edge of a neural network classifier.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(tbl.Smoker,"Holdout",0.30);
```

```
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a neural network classifier using the training set. Specify the `Smoker` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitcnet(tblTrain,"Smoker", ...
    "Standardize",true);
```

Calculate the test set classification edge.

```
e = edge(Mdl,tblTest,"Smoker")
```

```
e = 0.8657
```

The mean of the classification margins is close to 1, which indicates that the model performs well overall.

Select Features to Include in Neural Network Classifier

Perform feature selection by comparing test set classification margins, edges, errors, and predictions. Compare the test set metrics for a model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(fishertable.Species,"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

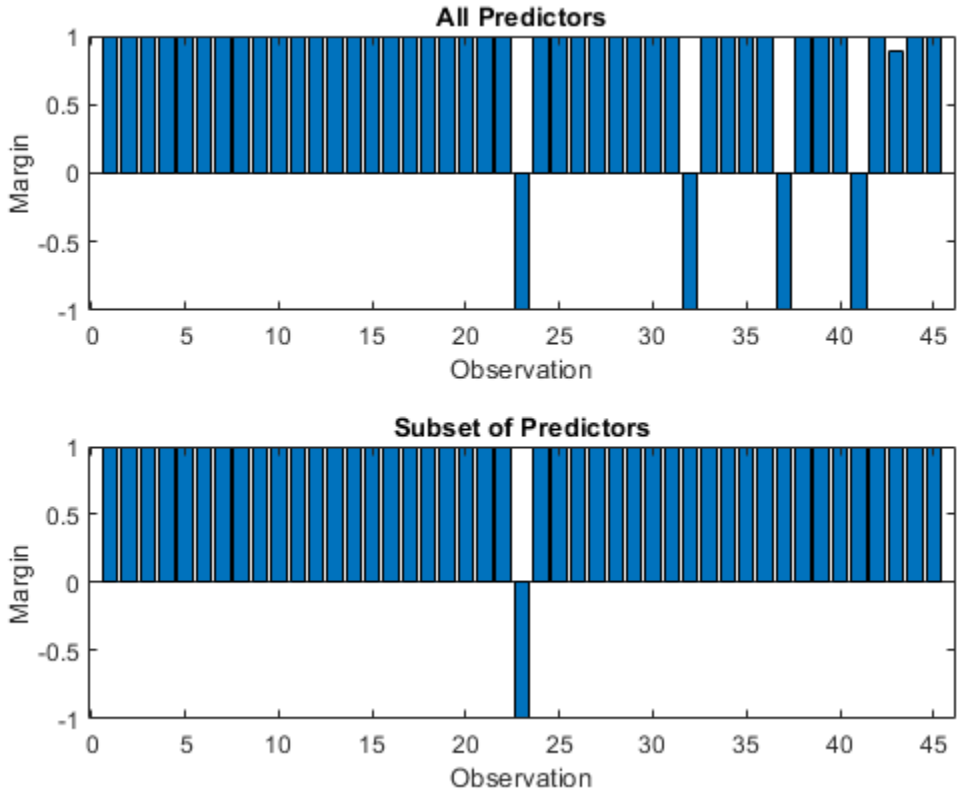
Train one neural network classifier using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `Species` as the response variable, and standardize the predictors.

```
allMdl = fitcnet(trainTbl,"Species","Standardize",true);
subsetMdl = fitcnet(trainTbl,"Species ~ SepalLength + SepalWidth + PetalLength", ...
    "Standardize",true);
```

Calculate the test set classification margins for the two models. Because the test set includes only 45 observations, display the margins using bar graphs.

For each observation, the classification margin is the difference between the classification score for the true class and the maximal score for the false classes. Because neural network classifiers return classification scores that are posterior probabilities, margin values close to 1 indicate confident classifications and negative margin values indicate misclassifications.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allMargins = margin(allMdl,testTbl);
bar(ax1,allMargins)
xlabel(ax1,"Observation")
ylabel(ax1,"Margin")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
subsetMargins = margin(subsetMdl,testTbl);
bar(ax2,subsetMargins)
xlabel(ax2,"Observation")
ylabel(ax2,"Margin")
title(ax2,"Subset of Predictors")
```



Compare the test set classification edge, or mean of the classification margins, of the two models.

```
allEdge = edge(allMdl,testTbl)
```

```
allEdge = 0.8198
```

```
subsetEdge = edge(subsetMdl,testTbl)
```

```
subsetEdge = 0.9556
```

Based on the test set classification margins and edges, the model trained on a subset of the predictors seems to outperform the model trained on all the predictors.

Compare the test set classification error of the two models.

```
allError = loss(allMdl,testTbl);
allAccuracy = 1-allError
```

```
allAccuracy = 0.9111
```

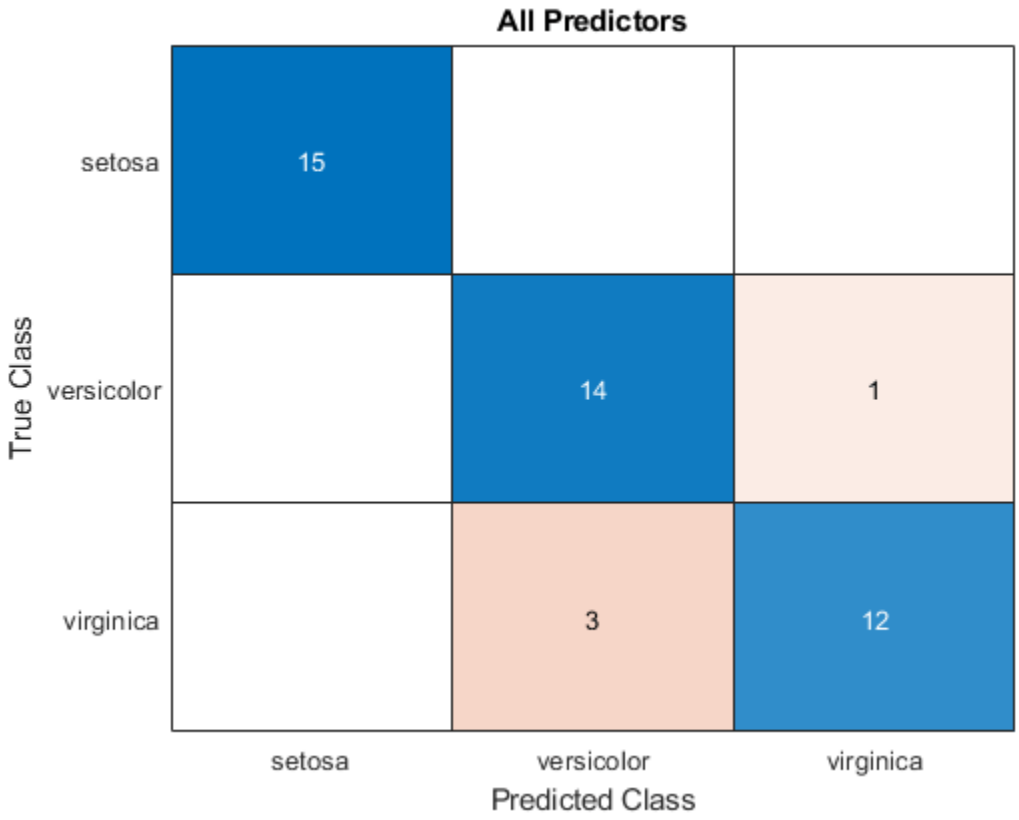
```
subsetError = loss(subsetMdl,testTbl);
subsetAccuracy = 1-subsetError
```

```
subsetAccuracy = 0.9778
```

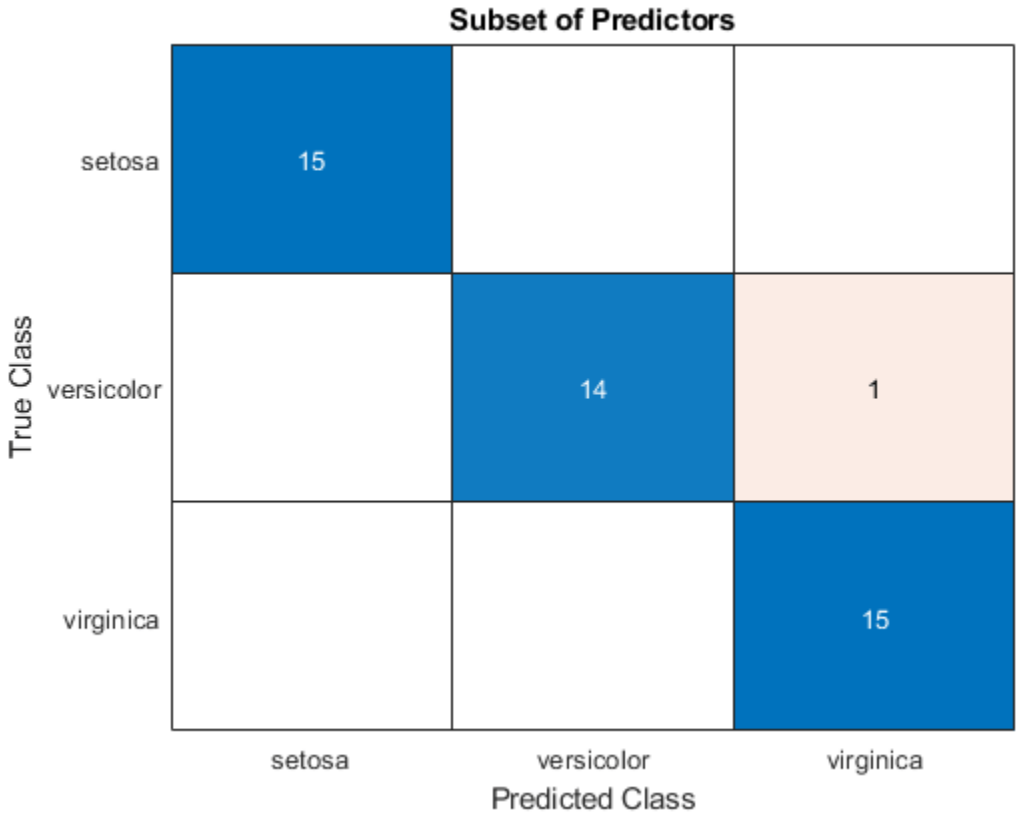
Again, the model trained using only a subset of the predictors seems to perform better than the model trained using all the predictors.

Visualize the test set classification results using confusion matrices.

```
allLabels = predict(allMdl,testTbl);
figure
confusionchart(testTbl.Species,allLabels)
title("All Predictors")
```




```
subsetLabels = predict(subsetMdl,testTbl);
figure
confusionchart(testTbl.Species,subsetLabels)
title("Subset of Predictors")
```



The model trained using all the predictors misclassifies four of the test set observations. The model trained using a subset of the predictors misclassifies only one of the test set observations.

Given the test set performance of the two models, consider using the model trained using all the predictors except `PetalWidth`.

Input Arguments

`Mdl` — Trained neural network classifier

`ClassificationNeuralNetwork` model object | `CompactClassificationNeuralNetwork` model object

Trained neural network classifier, specified as a `ClassificationNeuralNetwork` model object or `CompactClassificationNeuralNetwork` model object returned by `fitcnet` or `compact`, respectively.

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain an additional column for the response variable. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.
- If you trained `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table.
- If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. By default, `edge` assumes that each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `edge(Mdl,Tbl,"Response","Weights","W")` specifies to use the `Response` and `W` variables in the table `Tbl` as the class labels and observation weights, respectively.

`ObservationsIn` — Predictor data observation dimension

`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights

nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

If you supply weights, then `edge` computes the weighted classification edge and normalizes weights to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

More About

Classification Edge

The classification edge is the mean of the classification margins, or the weighted mean of the classification margins when you specify `Weights`.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

See Also

`ClassificationNeuralNetwork` | `CompactClassificationNeuralNetwork` | `fitcnet` | `loss` | `margin` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

edge

Package: `classreg.learning.classif`

Find classification edge for support vector machine (SVM) classifier

Syntax

```
e = edge(SVMModel,TBL,ResponseVarName)
e = edge(SVMModel,TBL,Y)

e = edge(SVMModel,X,Y)

e = edge( ___ ,'Weights',weights)
```

Description

`e = edge(SVMModel,TBL,ResponseVarName)` returns the classification edge on page 33-1344 (`e`) for the support vector machine (SVM) classifier `SVMModel` using the predictor data in table `TBL` and the class labels in `TBL.ResponseVarName`.

The classification edge (`e`) is a scalar value that represents the weighted mean of the classification margins on page 33-1344.

`e = edge(SVMModel,TBL,Y)` returns the classification edge on page 33-1344 (`e`) for the SVM classifier `SVMModel` using the predictor data in table `TBL` and the class labels in `Y`.

`e = edge(SVMModel,X,Y)` returns the classification edge for `SVMModel` using the predictor data in matrix `X` and the class labels in `Y`.

`e = edge( ___ ,'Weights',weights)` computes the classification edge for the observation weights supplied in `weights` using any of the input arguments in the previous syntaxes.

Examples

Estimate Test Sample Edge of SVM Classifiers

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify a 15% holdout sample for testing, standardize the data, and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Holdout',0.15,'ClassNames',{'b','g'},...
    'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1}; % Extract trained, compact classifier
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`CVSVMModel` is a `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Estimate the test sample edge.

```
e = edge(CompactSVMModel,XTest,YTest)
```

```
e = 5.0765
```

The margin average of the test sample is approximately 5.

Estimate Test Sample Weighted Margin Mean of SVM Classifiers

Suppose that the observations in a data set are measured sequentially, and that the last 150 observations have better quality due to a technology upgrade. Incorporate this advancement by weighing the better quality observations more than the other observations.

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Define a weight vector that weighs the better quality observations two times the other observations.

```
n = size(X,1);
weights = [ones(n-150,1);2*ones(150,1)];
```

Train an SVM classifier. Specify the weighting scheme and a 15% holdout sample for testing. Also, standardize the data and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Weights',weights,'Holdout',0.15,...
    'ClassNames',{'b','g'},'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1};
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
wTest = weights(testInds,:);
```

`CVSVMModel` is a trained `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Estimate the test sample weighted edge using the weighting scheme.

```
e = edge(CompactSVMModel,XTest,YTest,'Weights',wTest)
```

```
e = 4.8341
```

The weighted average margin of the test sample is approximately 5.

Select SVM Classifier Features by Comparing Test Sample Edges

Perform feature selection by comparing test sample edges from multiple models. Based solely on this comparison, the classifier with the highest edge is the best classifier.

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Partition the data set into training and test sets. Specify a 15% holdout sample for testing.

```
Partition = cvpartition(Y,'Holdout',0.15);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`Partition` defines the data set partition.

Define these two data sets:

- `fullX` contains all predictors (except the removed column of 0s).
- `partX` contains the last 20 predictors.

```
fullX = X;
partX = X(:,end-20:end);
```

Train SVM classifiers for each predictor set. Specify the partition definition.

```
FullCVSVMModel = fitcsvm(fullX,Y,'CVPartition',Partition);
PartCVSVMModel = fitcsvm(partX,Y,'CVPartition',Partition);
FCSVMModel = FullCVSVMModel.Trained{1};
PCSVMModel = PartCVSVMModel.Trained{1};
```

`FullCVSVMModel` and `PartCVSVMModel` are `ClassificationPartitionedModel` classifiers. They contain the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Estimate the test sample edge for each classifier.

```
fullEdge = edge(FCSVMModel,XTest,YTest)
```

```
fullEdge = 2.8321
```

```
partEdge = edge(PCSVMModel,XTest(:,end-20:end),YTest)
```

```
partEdge = 1.5541
```

The edge for the classifier trained on the complete data set is greater, suggesting that the classifier trained with all the predictors is better.

Input Arguments

`SVMModel` — SVM classification model

`ClassificationSVM` model object | `CompactClassificationSVM` model object

SVM classification model, specified as a `ClassificationSVM` model object or `CompactClassificationSVM` model object returned by `fitcsvm` or `compact`, respectively.

`TBL` — Sample data
table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all of the predictors used to train `SVMModel`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName` or `Y`.

If you trained `SVMModel` using sample data contained in a table, then the input data for `edge` must also be in a table.

If you set `'Standardize',true` in `fitcsvm` when training `SVMModel`, then the software standardizes the columns of the predictor data using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `table`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature). The variables in the columns of `X` must be the same as the variables that trained the `SVMModel` classifier.

The length of `Y` and the number of rows in `X` must be equal.

If you set `'Standardize',true` in `fitcsvm` to train `SVMModel`, then the software standardizes the columns of `X` using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `double` | `single`

`ResponseVarName` — Response variable name
name of variable in `TBL`

Response variable name, specified as the name of a variable in `TBL`. If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `TBL.Response`, then specify `ResponseVarName` as `'Response'`. Otherwise, the software treats all columns of `TBL`, including `TBL.Response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Y — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. `Y` must be the same as the data type of `SVMModel.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The length of `Y` must equal the number of rows in `TBL` or the number of rows in `X`.

`weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `TBL`

Observation weights, specified as a numeric vector or the name of a variable in `TBL`.

If you specify `weights` as a numeric vector, then the size of `weights` must be equal to the number of rows in `X` or `TBL`.

If you specify `weights` as the name of a variable in `TBL`, you must do so as a character vector or string scalar. For example, if the weights are stored as `TBL.W`, then specify `weights` as `'W'`. Otherwise, the software treats all columns of `TBL`, including `TBL.W`, as predictors.

If you supply weights, `edge` computes the weighted classification edge on page 33-1344. The software weights the observations in each row of `X` or `TBL` with the corresponding weight in `weights`.

Example: `'Weights','W'`

Data Types: `single` | `double` | `char` | `string`

More About

Classification Edge

The edge is the weighted mean of the classification margins.

The weights are the prior class probabilities. If you supply weights, then the software normalizes them to sum to the prior probabilities in the respective classes. The software uses the renormalized weights to compute the weighted mean.

One way to choose among multiple classifiers, for example, to perform feature selection, is to choose the classifier that yields the highest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

The positive class classification score $f(x)$ is the trained SVM classification function. $f(x)$ is also the numerical predicted response for x, or the score for predicting x into the positive class.

$$f(x) = \sum_{j=1}^{n} \alpha_j y_j G(x_j, x) + b,$$

where $(\alpha_1, ..., \alpha_n, b)$ are the estimated SVM parameters, $G(x_j, x)$ is the dot product in the predictor space between x and the support vectors, and the sum includes the training set observations. The negative class classification score for x, or the score for predicting x into the negative class, is $-f(x)$.

If $G(x_j,x) = x_j'x$ (the linear kernel), then the score function reduces to

$$f(x) = (x/s)'\beta + b\,.$$

s is the kernel scale and β is the vector of fitted linear coefficients.

For more details, see “Understanding Support Vector Machines” on page 18-150.

Algorithms

For binary classification, the software defines the margin for observation j, m_j, as

$$m_j = 2y_j f(x_j),$$

where $y_j \in \{-1,1\}$, and $f(x_j)$ is the predicted score of observation j for the positive class. However, $m_j = y_j f(x_j)$ is commonly used to define the margin.

References

[1] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`ClassificationSVM` | `CompactClassificationSVM` | `fitcsvm` | `kfoldEdge` | `loss` | `margin` | `predict` | `resubEdge`

Introduced in R2014a

edge

Classification edge

Syntax

```
E = edge(tree,TBL,ResponseVarName)
E = edge(tree,X,Y)
E = edge( ___ ,Name,Value)
```

Description

`E = edge(tree,TBL,ResponseVarName)` returns the classification edge for `tree` with data `TBL` and classification `TBL.ResponseVarName`.

`E = edge(tree,X,Y)` returns the classification edge for `tree` with data `X` and classification `Y`.

`E = edge( ___ ,Name,Value)` computes the edge with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can specify observation weights.

Input Arguments

`tree` — Trained classification tree

`ClassificationTree` model object | `CompactClassificationTree` model object

Trained classification tree, specified as a `ClassificationTree` or `CompactClassificationTree` model object. That is, `tree` is a trained classification model returned by `fitctree` or `compact`.

`TBL` — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all the predictors used to train `tree`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `tree`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `tree` using sample data contained in a `table`, then the input data for this method must also be in a table.

Data Types: `table`

`X` — Data to classify

numeric matrix

Data to classify, specified as a numeric matrix. Each row of `X` represents one observation, and each column represents one predictor. `X` must have the same number of columns as the data used to train `tree`. `X` must have the same number of rows as the number of elements in `Y`.

Data Types: `single` | `double`

`ResponseVarName` — Response variable name
name of a variable in `TBL`

Response variable name, specified as the name of a variable in `TBL`. If `TBL` contains the response variable used to train `tree`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `TBL.Response`, then specify it as `'Response'`. Otherwise, the software treats all columns of `TBL`, including `TBL.ResponseVarName`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `Y` must be of the same type as the classification used to train `tree`, and its number of elements must equal the number of rows of `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | name of a variable in `TBL` | numeric vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `TBL`.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `TBL`.

If you specify `Weights` as the name of a variable in `TBL`, you must do so as a character vector or string scalar. For example, if the weights are stored as `TBL.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `TBL`, including `TBL.W`, as predictors.

If you supply weights, `edge` computes the weighted classification edge on page 33-1344. The software weights the observations in each row of `X` or `TBL` with the corresponding weight in `Weights`.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

E — Classification edge

scalar value

Classification edge, returned as a scalar representing the weighted average value of the margin.

Examples

Compute the classification margin and edge for the Fisher iris data, trained on its first two columns of data, and view the last 10 entries:

```
load fisheriris
X = meas(:,1:2);
tree = fitctree(X,species);
E = edge(tree,X,species)

E =
    0.6299

M = margin(tree,X,species);
M(end-10:end)

ans =
    0.1111
    0.1111
    0.1111
   -0.2857
    0.6364
    0.6364
    0.1111
    0.7500
    1.0000
    0.6364
    0.2000
```

The classification tree trained on all the data is better.

```
tree = fitctree(meas,species);
E = edge(tree,meas,species)

E =
    0.9384

M = margin(tree,meas,species);
M(end-10:end)

ans =
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
```

```
    0.9565
    0.9565
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as the matrix X.

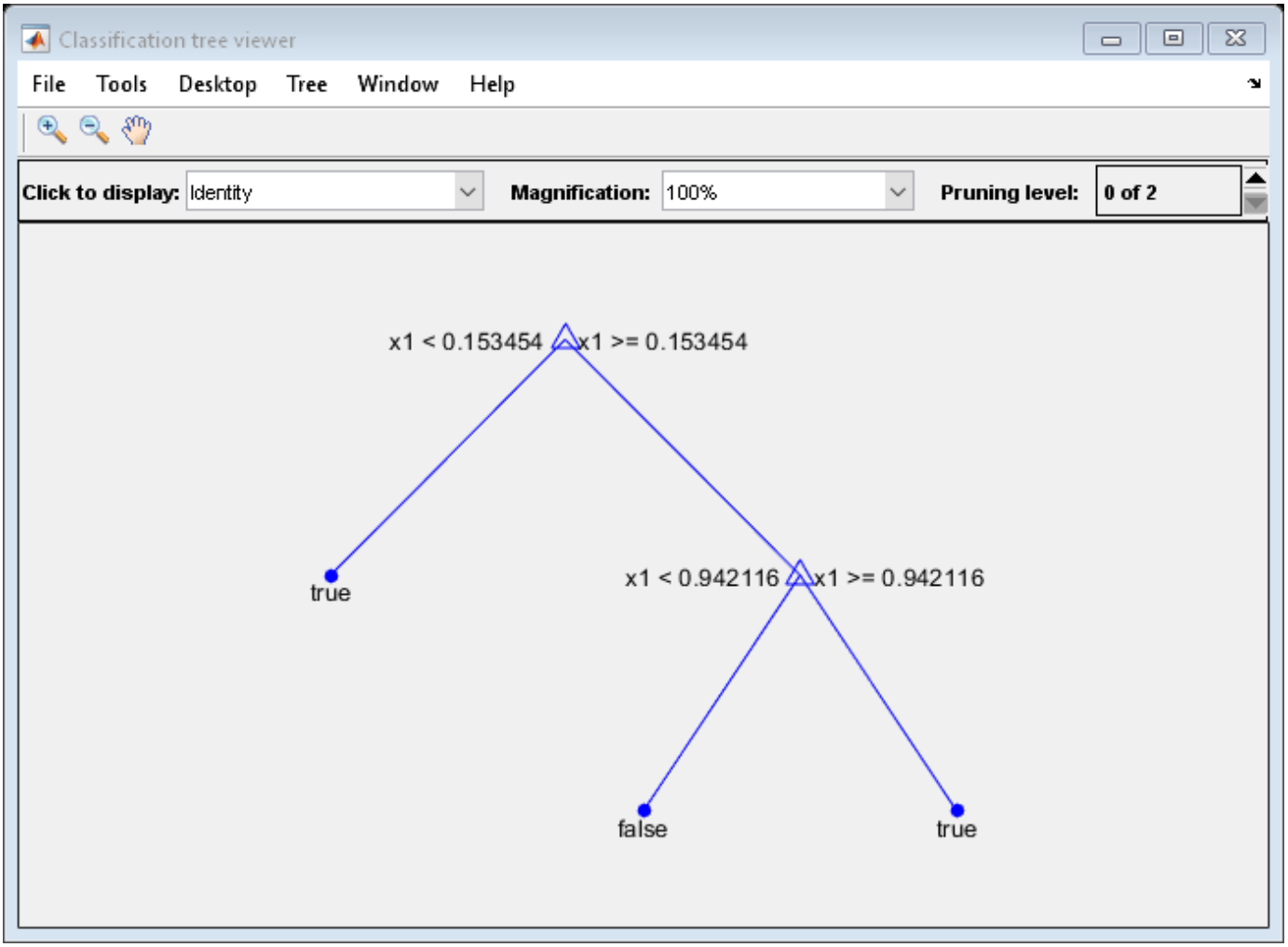
Score (tree)

For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor X as `true` when $X < 0.15$ or $X > 0.95$, and X is false otherwise.

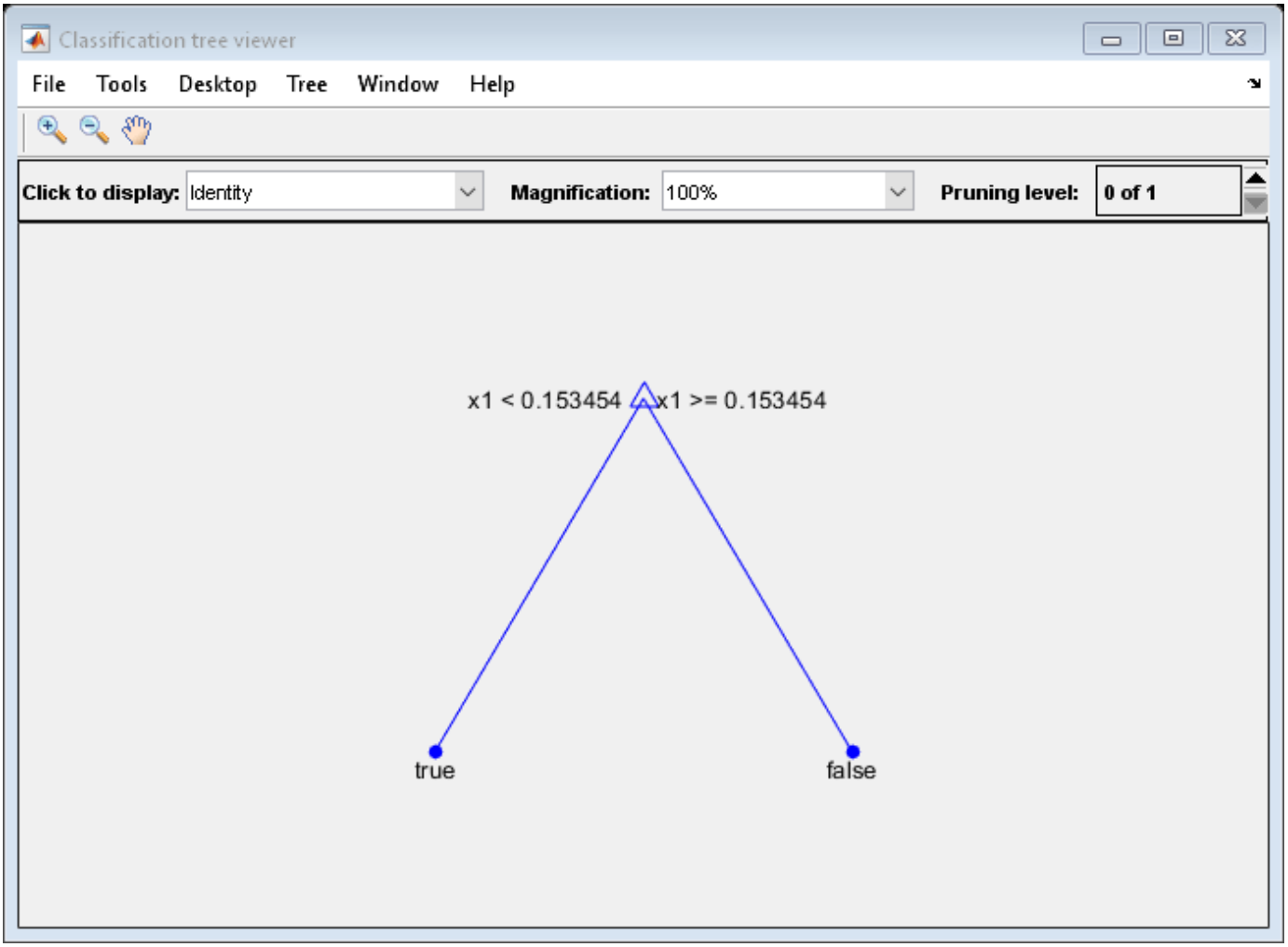
Generate 100 random points and classify them:

```
rng(0,'twister') % for reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `tree.Prior`. If you supply weights in the `weights` name-value pair, those weights are normalized to sum to the prior probabilities in the respective classes, and are then used to compute the weighted average.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`fitctree` | `loss` | `margin` | `predict`

end

Class: `dataset`

(Not Recommended) Last index in indexing expression for dataset array

---

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
end(A,k,n)
```

Description

`end(A,k,n)` is called for indexing expressions involving the dataset `A` when `end` is part of the `k`-th index out of `n` indices. For example, the expression `A(end-1,:)` calls `A`'s `end` method with `end(A,1,2)`.

See Also

`size`

epsilon

Class: `RepeatedMeasuresModel`

Epsilon adjustment for repeated measures anova

Syntax

```
tbl = epsilon(rm)
tbl = epsilon(rm,C)
```

Description

`tbl = epsilon(rm)` returns the epsilon adjustment factors for repeated measures model `rm`.

`tbl = epsilon(rm,C)` returns the epsilon adjustment factors for the test based on the contrast matrix `C`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`C` — Contrasts

matrix

Contrasts, specified as a matrix. The default value of `C` is the Q factor in a QR decomposition of the matrix M, where M is defined so that Y*M is the difference between all successive pairs of columns of the repeated measures matrix Y.

Data Types: `single` | `double`

Output Arguments

`tbl` — Epsilon adjustment factors

table

Epsilon adjustment factors for the repeated measures model `rm`, returned as a table. `tbl` contains four different adjustments for `epsilon`.

Correction	Definition
`Uncorrected`	No adjustments, epsilon = 1
`Greenhouse-Geisser`	Greenhouse-Geisser approximation
`Huynh-Feldt`	Huynh-Feldt approximation

Correction	Definition
`Lower bound`	Lower bound on the *p*-value

For details, see “Compound Symmetry Assumption and Epsilon Corrections” on page 9-55.

Data Types: `table`

Examples

Epsilon Corrections for Repeated Measures ANOVA

Load the sample data.

```
load fisheriris
```

The column vector, `species` consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform repeated measures analysis of variance.

```
ranovatbl = ranova(rm)
```

```
ranovatbl=3×8 table
                              SumSq     DF      MeanSq       F          pValue          pValueGG
                              ______    ___    ________    ______    ____________    ____________

    (Intercept):Measurements    1656.3     3      552.09    6873.3              0    9.4491e-279
    species:Measurements        282.47     6      47.078     586.1    1.4271e-206    4.9313e-150
    Error(Measurements)         35.423    441    0.080324
```

`ranova` computes the last three *p*-values using Greenhouse-Geisser, Huynh-Feldt, and lower bound corrections, respectively.

Display the epsilon correction values.

```
epsilon(rm)
```

```
ans=1×4 table
    Uncorrected    GreenhouseGeisser    HuynhFeldt    LowerBound
    ___________    _________________    __________    __________

         1              0.75179           0.76409       0.33333
```

You can check the compound symmetry (sphericity) assumption using the `mauchly` method.

Tips

- The `mauchly` method tests for sphericity.
- The `ranova` method contains *p*-values based on each epsilon value.

Algorithms

`ranova` computes the regular *p*-value (in the `pValue` column of the `rmanova` table) using the *F*-statistic cumulative distribution function:

p-value = 1 – fcdf(F,v_1,v_2).

When the compound symmetry assumption is not satisfied, `ranova` uses a correction factor epsilon, ε, to compute the corrected *p*-values as follows:

p-value_corrected = 1 – fcdf(F,ε*v_1,ε*v_2).

The `epsilon` method returns the epsilon adjustment values.

See Also

`fitrm` | `mauchly` | `ranova`

Topics

“Compound Symmetry Assumption and Epsilon Corrections” on page 9-55
“Mauchly’s Test of Sphericity” on page 9-57

evcdf

Extreme value cumulative distribution function

Syntax

```
p = evcdf(x,mu,sigma)
[p,plo,pup] = evcdf(x,mu,sigma,pcov,alpha)
[p,plo,pup] = evcdf( ___ ,'upper')
```

Description

`p = evcdf(x,mu,sigma)` returns the cumulative distribution function (cdf) for the type 1 extreme value distribution, with location parameter `mu` and scale parameter `sigma`, at each of the values in `x`. `x`, `mu`, and `sigma` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other inputs. The default values for `mu` and `sigma` are `0` and `1`, respectively.

`[p,plo,pup] = evcdf(x,mu,sigma,pcov,alpha)` returns confidence bounds for `p` when the input parameters `mu` and `sigma` are estimates. `pcov` is a 2-by-2 covariance matrix of the estimated parameters. `alpha` has a default value of `0.05`, and specifies `100(1 - alpha)`% confidence bounds. `plo` and `pup` are arrays of the same size as `p`, containing the lower and upper confidence bounds.

`[p,plo,pup] = evcdf( ___ ,'upper')` returns the complement of the type 1 extreme value distribution cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities. You can use the `'upper'` argument with any of the previous syntaxes.

The function `evcdf` computes confidence bounds for `P` using a normal approximation to the distribution of the estimate

$$\frac{X-\widehat{\mu}}{\widehat{\sigma}}$$

and then transforming those bounds to the scale of the output `P`. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pcov` from large samples, but in smaller samples other methods of computing the confidence bounds might be more accurate.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating `X` and subtracting the resulting distribution values from `1`. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `evfit` | `evinv` | `evlike` | `evpdf` | `evrnd` | `evstat`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

evfit

Extreme value parameter estimates

Syntax

```
parmhat = evfit(data)
[parmhat,parmci] = evfit(data)
[parmhat,parmci] = evfit(data,alpha)
[...] = evfit(data,alpha,censoring)
[...] = evfit(data,alpha,censoring,freq)
[...] = evfit(data,alpha,censoring,freq,options)
```

Description

`parmhat = evfit(data)` returns maximum likelihood estimates of the parameters of the type 1 extreme value distribution given the sample data in `data`. The sample data `data` must be a double-precision vector. `parmhat(1)` is the location parameter μ, and `parmhat(2)` is the scale parameter σ.

`[parmhat,parmci] = evfit(data)` returns 95% confidence intervals for the parameter estimates on the μ and σ parameters in the 2-by-2 matrix `parmci`. The first column of the matrix of the extreme value fit contains the lower and upper confidence bounds for the parameter μ, and the second column contains the confidence bounds for the parameter σ.

`[parmhat,parmci] = evfit(data,alpha)` returns 100(1 - `alpha`)% confidence intervals for the parameter estimates, where `alpha` is a value in the range `[0 1]` specifying the width of the confidence intervals. By default, `alpha` is `0.05`, which corresponds to 95% confidence intervals.

`[...] = evfit(data,alpha,censoring)` accepts a Boolean vector, `censoring`, of the same size as `data`, which is `1` for observations that are right-censored and `0` for observations that are observed exactly.

`[...] = evfit(data,alpha,censoring,freq)` accepts a frequency vector, `freq` of the same size as `data`. Typically, `freq` contains integer frequencies for the corresponding elements in `data`, but can contain any nonnegative values. Pass in `[]` for `alpha`, `censoring`, or `freq` to use their default values.

`[...] = evfit(data,alpha,censoring,freq,options)` accepts a structure, `options`, that specifies control parameters for the iterative algorithm the function uses to compute maximum likelihood estimates. You can create `options` using the function `statset`. Enter `statset('evfit')` to see the names and default values of the parameters that `evfit` accepts in the `options` structure. See the reference page for `statset` for more information about these options.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating `X`. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`evcdf` | `evinv` | `evlike` | `evpdf` | `evrnd` | `evstat` | `mle`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

evinv

Extreme value inverse cumulative distribution function

Syntax

```
X = evinv(P,mu,sigma)
[X,XLO,XUP] = evinv(P,mu,sigma,pcov,alpha)
```

Description

`X = evinv(P,mu,sigma)` returns the inverse cumulative distribution function (cdf) for a type 1 extreme value distribution with location parameter `mu` and scale parameter `sigma`, evaluated at the values in `P`. `P`, `mu`, and `sigma` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other inputs. The default values for `mu` and `sigma` are `0` and `1`, respectively.

`[X,XLO,XUP] = evinv(P,mu,sigma,pcov,alpha)` produces confidence bounds for `X` when the input parameters `mu` and `sigma` are estimates. `pcov` is the covariance matrix of the estimated parameters. `alpha` is a scalar that specifies 100(1 – `alpha`)% confidence bounds for the estimated parameters, and has a default value of 0.05. `XLO` and `XUP` are arrays of the same size as `X` containing the lower and upper confidence bounds.

The function `evinv` computes confidence bounds for `P` using a normal approximation to the distribution of the estimate

$$\hat{\mu} + \hat{\sigma} q$$

where q is the `P`th quantile from an extreme value distribution with parameters $\mu = 0$ and $\sigma = 1$. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pcov` from large samples, but in smaller samples other methods of computing the confidence bounds might be more accurate.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating `X`. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

evcdf | evfit | evlike | evpdf | evrnd | evstat | icdf

Introduced before R2006a

eq

Class: `qrandstream`

Test handle equality

Syntax

```
h1 == h2
tf = eq(h1, h2)
```

Description

`h1 == h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise equality result. If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = eq(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`ge` | `gt` | `le` | `lt` | `ne` | `qrandstream`

error

Class: `CompactTreeBagger`

Error (misclassification probability or MSE)

Syntax

```
err = error(B,TBLnew,Ynew)
err = error(B,Xnew,Ynew)
err = error(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
err = error(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`err = error(B,TBLnew,Ynew)` computes the misclassification probability for classification trees or mean squared error (MSE) for regression trees for each tree, for the predictors contained in the table `TBLnew` given true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`err = error(B,Xnew,Ynew)` computes the misclassification probability for classification trees or mean squared error (MSE) for regression trees for each tree, the for predictors contained in the matrix `Xnew` given true response `Ynew`. If you trained `B` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

For classification, `Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector, or logical vector. For regression, `Y` must be a numeric vector. `err` is a vector with one error measure for each of the `NTrees` trees in the ensemble `B`.

`err = error(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `err = error(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	Character vector or string scalar indicating how the method computes errors. If set to `'cumulative'` (default), `error` computes cumulative errors and `err` is a vector of length `NTrees`, where the first element gives error from `trees(1)`, second element gives error from`trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `err` is a vector of length `NTrees`, where each element is an error from each tree in the ensemble. If set to `'ensemble'`, `err` is a scalar showing the cumulative error for the entire ensemble.
`'Weights'`	Vector of observation weights to use for error averaging. By default the weight of every observation is 1. The length of this vector must be equal to the number of rows in `X`.

`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives error from `trees(1)`, the second element gives error from `trees(1:2)` etc.
`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. The method uses these weights to combine output from the specified trees by taking a weighted average instead of the simple non-weighted majority vote. You cannot use this argument in the `'individual'` mode.
`'UseInstanceForTree'`	Logical matrix of size `Nobs`-by-`NTrees` indicating which trees should be used to make predictions for each observation. By default the method uses all trees for all observations.

Algorithms

When estimating the ensemble error:

- Using the `'Mode'` name-value pair argument, you can specify to return the error any of these three ways:
  - The error for individual trees in the ensemble
  - The cumulative error over all trees
  - The error for the entire ensemble
- Using the `'Trees'` name-value pair argument, you can specify which trees to use in the ensemble error calculations.
- Using the `'UseInstanceForTree'` name-value pair argument, you can specify which observations in the input data (`X` and `Y`) to use in the ensemble error calculation for each selected tree.
- Using the `'Weights'` name-value pair argument, you can attribute each *observation* with a weight. For the formulae that follow, w_j is the weight of observation j.
- Using the `'TreeWeights'` name-value pair argument, you can attribute each *tree* with a weight.

For regression problems, `error` estimates the weighted MSE of the ensemble of bagged regression trees for predicting `Y` given `X` using selected trees and observations.

1. `error` predicts responses for selected observations in `X` using the selected regression trees in the ensemble.
2. The MSE estimate depends on the value of `'Mode'`.
   - If you specify `'Mode','Individual'`, then the weighted MSE for tree t is

$$\mathrm{MSE}_t = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j \left(y_j - \hat{y}_{tj}\right)^2 .$$

$\widehat{y}_{tj}$ is the predicted response of observation j from selected regression tree t. `error` sets any unselected observations within a selected tree to the weighted sample average of the observed, training data responses.

- If you specify `'Mode','Cumulative'`, then the weighted MSE is a vector of size T^* containing cumulative, weighted MSEs over the $T^* \le T$ selected trees. `error` follows these steps to estimate MSE_t^*, the cumulative, weighted MSE using the first t selected trees.

  a For selected observation j, $j = 1,\ldots,n$, `error` estimates $\widehat{y}_{\mathrm{bag},tj}$, the weighted average of the predictions among the first t selected trees (for details, see `predict`). For this computation, `error` uses the tree weights.

  b `error` estimates the cumulative, weighted MSE through tree t.

$$\mathrm{MSE}_t^* = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j \left(y_j - \widehat{y}_{\mathrm{bag},tj}\right)^2 .$$

  `error` sets observations that are unselected for all selected trees to the weighted sample average of the observed, training data responses.

- If you specify `'Mode','Ensemble'`, then the weighted MSE is the last element of the cumulative, weighted MSE vector.

For classification problems, `error` estimates the weighted misclassification rate of the ensemble of bagged classification trees for predicting `Y` given `X` using selected trees and observations.

- If you specify `'Mode','Individual'`, then the weighted misclassification rate for tree t is

$$e_t = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j I\left(y_j \ne \widehat{y}_{tj}\right) .$$

  $\widehat{y}_{tj}$ is the predicted class for selected observation j using from selected classification tree t. `error` sets any unselected observations within a selected tree to the predicted, weighted, most popular class over all training responses. If there are multiple most popular classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular.

- If you specify `'Mode','Cumulative'` then the weighted misclassification rate is a vector of size T^* containing cumulative, weighted misclassification rates over the $T^* \le T$ selected trees. `error` follows these steps to estimate e_t^*, the cumulative, weighted misclassification rate using the first t selected trees.

  1 For selected observation j, $j = 1,\ldots,n$, `error` estimates $\widehat{y}_{\mathrm{bag},tj}$, the weighted, most popular class among the first t selected trees (for details, see `predict`). For this computation, `error` uses the tree weights.

  2 `error` estimates the cumulative, weighted misclassification rate through tree t.

$$e_t^* = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j I\left(y_j \ne \widehat{y}_{\mathrm{bag},tj}\right) .$$

  `error` sets any observations that are unselected for all selected trees to the predicted, weighted, most popular class over all training responses. If there are multiple most popular

classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular.

- If you specify `'Mode','Ensemble'`, then the weighted misclassification rate is the last element of the cumulative, weighted misclassification rate vector.

See Also

`CompactTreeBagger` | `predict`

Topics

"Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger" on page 18-113
"Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger" on page 18-124

error

Class: `TreeBagger`

Error (misclassification probability or MSE)

Syntax

```
err = error(B,TBLnew,Ynew)
err = error(B,Xnew,Ynew)
err = error(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
err = error(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`err = error(B,TBLnew,Ynew)` computes the misclassification probability for classification trees or mean squared error (MSE) for regression trees for each tree, for the predictors contained in the table `TBLnew` given true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`err = error(B,Xnew,Ynew)` computes the misclassification probability for classification trees or mean squared error (MSE) for regression trees for each tree, the for predictors contained in the matrix `Xnew` given true response `Ynew`. If you trained `B` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

For classification, `Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector or logical vector. For regression, `Y` must be a numeric vector. `err` is a vector with one error measure for each of the `NTrees` trees in the ensemble `B`.

`err = error(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `err = error(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	Character vector or string scalar indicating how the method computes errors. If set to `'cumulative'` (default), `error` computes cumulative errors and `err` is a vector of length `NTrees`, where the first element gives error from `trees(1)`, second element gives error from`trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `err` is a vector of length `NTrees`, where each element is an error from each tree in the ensemble. If set to `'ensemble'`, `err` is a scalar showing the cumulative error for the entire ensemble.
`'Weights'`	Vector of observation weights to use for error averaging. By default the weight of every observation is 1. The length of this vector must be equal to the number of rows in `X`.

`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives error from `trees(1)`, the second element gives error from `trees(1:2)` etc.
`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. The method uses these weights to combine output from the specified trees by taking a weighted average instead of the simple non-weighted majority vote. You cannot use this argument in the `'individual'` mode.
`'UseInstanceForTree'`	Logical matrix of size `Nobs`-by-`NTrees` indicating which trees should be used to make predictions for each observation. By default the method uses all trees for all observations.

Algorithms

When estimating the ensemble error:

- Using the `'Mode'` name-value pair argument, you can specify to return the error any of these three ways:
  - The error for individual trees in the ensemble
  - The cumulative error over all trees
  - The error for the entire ensemble
- Using the `'Trees'` name-value pair argument, you can specify which trees to use in the ensemble error calculations.
- Using the `'UseInstanceForTree'` name-value pair argument, you can specify which observations in the input data (`X` and `Y`) to use in the ensemble error calculation for each selected tree.
- Using the `'Weights'` name-value pair argument, you can attribute each *observation* with a weight. For the formulae that follow, w_j is the weight of observation j.
- Using the `'TreeWeights'` name-value pair argument, you can attribute each *tree* with a weight.

For regression problems, `error` estimates the weighted MSE of the ensemble of bagged regression trees for predicting `Y` given `X` using selected trees and observations.

1. `error` predicts responses for selected observations in `X` using the selected regression trees in the ensemble.
2. The MSE estimate depends on the value of `'Mode'`.
   - If you specify `'Mode','Individual'`, then the weighted MSE for tree t is

$$\mathrm{MSE}_t = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j \left(y_j - \hat{y}_{tj}\right)^2 .$$

$\widehat{y}_{tj}$ is the predicted response of observation j from selected regression tree t. `error` sets any unselected observations within a selected tree to the weighted sample average of the observed, training data responses.

- If you specify `'Mode','Cumulative'`, then the weighted MSE is a vector of size T^* containing cumulative, weighted MSEs over the $T^* \le T$ selected trees. `error` follows these steps to estimate MSE_t^*, the cumulative, weighted MSE using the first t selected trees.

  a For selected observation j, $j = 1,\ldots,n$, `error` estimates $\widehat{y}_{\mathrm{bag},tj}$, the weighted average of the predictions among the first t selected trees (for details, see `predict`). For this computation, `error` uses the tree weights.

  b `error` estimates the cumulative, weighted MSE through tree t.

$$\mathrm{MSE}_t^* = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j \left(y_j - \widehat{y}_{\mathrm{bag},tj}\right)^2 .$$

  `error` sets observations that are unselected for all selected trees to the weighted sample average of the observed, training data responses.

- If you specify `'Mode','Ensemble'`, then the weighted MSE is the last element of the cumulative, weighted MSE vector.

For classification problems, `error` estimates the weighted misclassification rate of the ensemble of bagged classification trees for predicting `Y` given `X` using selected trees and observations.

- If you specify `'Mode','Individual'`, then the weighted misclassification rate for tree t is

$$e_t = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j I\left(y_j \ne \widehat{y}_{tj}\right) .$$

  $\widehat{y}_{tj}$ is the predicted class for selected observation j using from selected classification tree t. `error` sets any unselected observations within a selected tree to the predicted, weighted, most popular class over all training responses. If there are multiple most popular classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular.

- If you specify `'Mode','Cumulative'` then the weighted misclassification rate is a vector of size T^* containing cumulative, weighted misclassification rates over the $T^* \le T$ selected trees. `error` follows these steps to estimate e_t^*, the cumulative, weighted misclassification rate using the first t selected trees.

  1 For selected observation j, $j = 1,\ldots,n$, `error` estimates $\widehat{y}_{\mathrm{bag},tj}$, the weighted, most popular class among the first t selected trees (for details, see `predict`). For this computation, `error` uses the tree weights.

  2 `error` estimates the cumulative, weighted misclassification rate through tree t.

$$e_t^* = \frac{1}{\sum_{j=1}^{n} w_j} \sum_{j=1}^{n} w_j I\left(y_j \ne \widehat{y}_{\mathrm{bag},tj}\right) .$$

  `error` sets any observations that are unselected for all selected trees to the predicted, weighted, most popular class over all training responses. If there are multiple most popular

classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular.

- If you specify `'Mode','Ensemble'`, then the weighted misclassification rate is the last element of the cumulative, weighted misclassification rate vector.

See Also

`TreeBagger` | `compact` | `error` | `oobError` | `oobPredict` | `predict` | `quantileError`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

evalclusters

Evaluate clustering solutions

Syntax

```
eva = evalclusters(x,clust,criterion)
eva = evalclusters(x,clust,criterion,Name,Value)
```

Description

`eva = evalclusters(x,clust,criterion)` creates a clustering evaluation object containing data used to evaluate the optimal number of data clusters.

`eva = evalclusters(x,clust,criterion,Name,Value)` creates a clustering evaluation object using additional options specified by one or more name-value pair arguments.

Examples

Evaluate the Clustering Solution Using Calinski-Harabasz Criterion

Evaluate the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion.

Load the sample data.

```
load fisheriris;
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Evaluate the optimal number of clusters using the Calinski-Harabasz criterion. Cluster the data using `kmeans`.

```
rng('default');  % For reproducibility
eva = evalclusters(meas,'kmeans','CalinskiHarabasz','KList',[1:6])
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [Inf 513.9245 561.6278 530.4871 456.1279 469.5068]
           OptimalK: 1
```

The `OptimalK` value indicates that, based on the Calinski-Harabasz criterion, the optimal number of clusters is three.

Evaluate a Matrix of Clustering Solutions

Use an input matrix of proposed clustering solutions to evaluate the optimal number of clusters.

Load the sample data.

```
load fisheriris;
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Use `kmeans` to create an input matrix of proposed clustering solutions for the sepal length measurements, using 1, 2, 3, 4, 5, and 6 clusters.

```
clust = zeros(size(meas,1),6);
for i=1:6
clust(:,i) = kmeans(meas,i,'emptyaction','singleton',...
        'replicate',5);
end
```

Each row of `clust` corresponds to one sepal length measurement. Each of the six columns corresponds to a clustering solution containing 1 to 6 clusters.

Evaluate the optimal number of clusters using the Calinski-Harabasz criterion.

```
eva = evalclusters(meas,clust,'CalinskiHarabasz')
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [NaN 513.9245 561.6278 530.4871 456.1279 469.5068]
           OptimalK: 3
```

The `OptimalK` value indicates that, based on the Calinski-Harabasz criterion, the optimal number of clusters is three.

Specify Clustering Algorithm with a Function Handle

Use a function handle to specify the clustering algorithm, then evaluate the optimal number of clusters.

Load the sample data.

```
load fisheriris;
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Use a function handle to specify the clustering algorithm.

```
myfunc = @(X,K)(kmeans(X, K, 'emptyaction','singleton',...
    'replicate',5));
```

Evaluate the optimal number of clusters for the sepal length data using the Calinski-Harabasz criterion.

```
eva = evalclusters(meas,myfunc,'CalinskiHarabasz',...
    'klist',[1:6])
```

```
eva =
  CalinskiHarabaszEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [NaN 513.9245 561.6278 530.4871 456.1279 469.5068]
           OptimalK: 3
```

The `OptimalK` value indicates that, based on the Calinski-Harabasz criterion, the optimal number of clusters is three.

Input Arguments

x — Input data

matrix

Input data, specified as an *N*-by-*P* matrix. *N* is the number of observations, and *P* is the number of variables.

Data Types: `single` | `double`

clust — Clustering algorithm

`'kmeans'` | `'linkage'` | `'gmdistribution'` | matrix of clustering solutions | function handle

Clustering algorithm, specified as one of the following.

`'kmeans'`	Cluster the data in `x` using the `kmeans` clustering algorithm, with `'EmptyAction'` set to `'singleton'` and `'Replicates'` set to `5`.
`'linkage'`	Cluster the data in `x` using the `clusterdata` agglomerative clustering algorithm, with `'Linkage'` set to `'ward'`.
`'gmdistribution'`	Cluster the data in `x` using the `gmdistribution` Gaussian mixture distribution algorithm, with `'SharedCov'` set to `true` and `'Replicates'` set to `5`.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can specify a clustering algorithm using a function handle. The function must be of the form `C = clustfun(DATA,K)`, where `DATA` is the data to be clustered, and `K` is the number of clusters. The output of `clustfun` must be one of the following:

- A vector of integers representing the cluster index for each observation in `DATA`. There must be `K` unique values in this vector.
- A numeric *n*-by-*K* matrix of score for *n* observations and *K* classes. In this case, the cluster index for each observation is determined by taking the largest score value in each row.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can also specify `clust` as a *n*-by-*K* matrix containing the proposed clustering solutions. *n* is the number of

observations in the sample data, and *K* is the number of proposed clustering solutions. Column *j* contains the cluster indices for each of the *N* points in the *j*th clustering solution.

Data Types: `single` | `double` | `char` | `string` | `function_handle`

`criterion` — Clustering evaluation criterion

`'CalinskiHarabasz'` | `'DaviesBouldin'` | `'gap'` | `'silhouette'`

Clustering evaluation criterion, specified as one of the following.

`'CalinskiHarabasz'`	Create a `CalinskiHarabaszEvaluation` clustering evaluation object containing Calinski-Harabasz index values.
`'DaviesBouldin'`	Create a `DaviesBouldinEvaluation` cluster evaluation object containing Davies-Bouldin index values.
`'gap'`	Create a `GapEvaluation` cluster evaluation object containing gap criterion values.
`'silhouette'`	Create a `SilhouetteEvaluation` cluster evaluation object containing silhouette values.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KList',[1:5],'Distance','cityblock'` specifies to test 1, 2, 3, 4, and 5 clusters using the city block distance metric.

For All Criteria

`KList` — List of number of clusters to evaluate

vector

List of number of clusters to evaluate, specified as the comma-separated pair consisting of `'KList'` and a vector of positive integer values. You must specify `KList` when `clust` is a clustering algorithm name or a function handle. When `criterion` is `'gap'`, `clust` must be a character vector, a string scalar, or a function handle, and you must specify `KList`.

Example: `'KList',[1:6]`

Data Types: `single` | `double`

For Silhouette and Gap

`Distance` — Distance metric

`'sqEuclidean'` (default) | `'Euclidean'` | `'cityblock'` | vector | function | ...

Distance metric used for computing the criterion values, specified as the comma-separated pair consisting of `'Distance'` and one of the following.

`'sqEuclidean'`	Squared Euclidean distance

`'Euclidean'`	Euclidean distance. This option is not valid for the `kmeans` clustering algorithm.
`'cityblock'`	Sum of absolute differences
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)
`'Hamming'`	Percentage of coordinates that differ. This option is only valid for the `Silhouette` criterion.
`'Jaccard'`	Percentage of nonzero coordinates that differ. This option is only valid for the `Silhouette` criterion.

For detailed information about each distance metric, see `pdist`.

You can also specify a function for the distance metric using a function handle. The distance function must be of the form `d2 = distfun(XI,XJ)`, where `XI` is a 1-by-*n* vector corresponding to a single row of the input matrix `X`, and `XJ` is an m_2-by-*n* matrix corresponding to multiple rows of `X`. `distfun` must return an m_2-by-1 vector of distances `d2`, whose *k*th element is the distance between `XI` and `XJ(k,:)`.

`Distance` only accepts a function handle if the clustering algorithm `clust` accepts a function handle as the distance metric. For example, the `kmeans` clustering algorithm does not accept a function handle as the distance metric. Therefore, if you use the `kmeans` algorithm and then specify a function handle for `Distance`, the software errors.

- If `criterion` is `'silhouette'`, you can also specify `Distance` as the output vector created by the function `pdist`.
- When `clust` is `'kmeans'` or `'gmdistribution'`, `evalclusters` uses the distance metric specified for `Distance` to cluster the data.
- If `clust` is `'linkage'`, and `Distance` is either `'sqEuclidean'` or `'Euclidean'`, then the clustering algorithm uses the Euclidean distance and Ward linkage.
- If `clust` is `'linkage'` and `Distance` is any other metric, then the clustering algorithm uses the specified distance metric and average linkage.
- In all other cases, the distance metric specified for `Distance` must match the distance metric used in the clustering algorithm to obtain meaningful results.

Example: `'Distance','Euclidean'`

Data Types: `single` | `double` | `char` | `string` | `function_handle`

For Silhouette Only

`ClusterPriors` — Prior probabilities for each cluster

`'empirical'` (default) | `'equal'`

Prior probabilities for each cluster, specified as the comma-separated pair consisting of `'ClusterPriors'` and one of the following.

`'empirical'`	Compute the overall silhouette value for the clustering solution by averaging the silhouette values for all points. Each cluster contributes to the overall silhouette value proportionally to its size.
`'equal'`	Compute the overall silhouette value for the clustering solution by averaging the silhouette values for all points within each cluster, and then averaging those values across all clusters. Each cluster contributes equally to the overall silhouette value, regardless of its size.

Example: `'ClusterPriors','empirical'`

For Gap Only

B — Number of reference data sets

`100` (default) | positive integer value

Number of reference data sets generated from the reference distribution `ReferenceDistribution`, specified as the comma-separated pair consisting of `'B'` and a positive integer value.

Example: `'B',150`

Data Types: `single` | `double`

ReferenceDistribution — Reference data generation method

`'PCA'` (default) | `'uniform'`

Reference data generation method, specified as the comma-separated pair consisting of `'ReferenceDistributions'` and one of the following.

`'PCA'`	Generate reference data from a uniform distribution over a box aligned with the principal components of the data matrix `X`.
`'uniform'`	Generate reference data uniformly over the range of each feature in the data matrix `X`.

Example: `'ReferenceDistribution','uniform'`

SearchMethod — Method for selecting optimal number of clusters

`'globalMaxSE'` (default) | `'firstMaxSE'`

Method for selecting the optimal number of clusters, specified as the comma-separated pair consisting of `'SearchMethod'` and one of the following.

`'globalMaxSE'`	Evaluate each proposed number of clusters in `KList` and select the smallest number of clusters satisfying $$\text{Gap}(K) \geq GAPMAX - \text{SE}(GAPMAX),$$ where K is the number of clusters, $\text{Gap}(K)$ is the gap value for the clustering solution with K clusters, $GAPMAX$ is the largest gap value, and $\text{SE}(GAPMAX)$ is the standard error corresponding to the largest gap value.

`'firstMaxSE'`	Evaluate each proposed number of clusters in `KList` and select the smallest number of clusters satisfying $\text{Gap}(K) \geq \text{Gap}(K+1) - \text{SE}(K+1)$, where K is the number of clusters, $\text{Gap}(K)$ is the gap value for the clustering solution with K clusters, and $\text{SE}(K+1)$ is the standard error of the clustering solution with $K+1$ clusters.

Example: `'SearchMethod','globalMaxSE'`

Output Arguments

eva — Clustering evaluation data

clustering evaluation object

Clustering evaluation data, returned as a clustering evaluation object.

See Also

`CalinskiHarabaszEvaluation` | `DaviesBouldinEvaluation` | `GapEvaluation` | `SilhouetteEvaluation`

Topics

“Cluster Using Gaussian Mixture Model” on page 16-39
“k-Means Clustering” on page 16-33
“Hierarchical Clustering” on page 16-6

Introduced in R2013b

evlike

Extreme value negative log-likelihood

Syntax

```
nlogL = evlike(params,data)
[nlogL,AVAR] = evlike(params,data)
[...] = evlike(params,data,censoring)
[...] = evlike(params,data,censoring,freq)
```

Description

`nlogL = evlike(params,data)` returns the negative of the log-likelihood for the type 1 extreme value distribution. `params(1)` is the tail location parameter, `mu`, and `params(2)` is the scale parameter, `sigma`. `nlogL` is a scalar.

`[nlogL,AVAR] = evlike(params,data)` returns the inverse of Fisher's information matrix, `AVAR`. If the input parameter values in `params` are the maximum likelihood estimates, the diagonal elements of `AVAR` are their asymptotic variances. `AVAR` is based on the observed Fisher's information, not the expected information.

`[...] = evlike(params,data,censoring)` accepts a Boolean vector of the same size as `data`, which is 1 for observations that are right-censored and 0 for observations that are observed exactly.

`[...] = evlike(params,data,censoring,freq)` accepts a frequency vector of the same size as `data`. `freq` typically contains integer frequencies for the corresponding elements in `data`, but can contain any nonnegative values. Pass in `[]` for `censoring` to use its default value.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating `data`. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`evcdf` | `evfit` | `evinv` | `evpdf` | `evrnd` | `evstat`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

evpdf

Extreme value probability density function

Syntax

```
Y = evpdf(X,mu,sigma)
```

Description

`Y = evpdf(X,mu,sigma)` returns the pdf of the type 1 extreme value distribution with location parameter `mu` and scale parameter `sigma`, evaluated at the values in `X`. `X`, `mu`, and `sigma` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other inputs. The default values for `mu` and `sigma` are `0` and `1`, respectively.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating `X`. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`evcdf` | `evfit` | `evinv` | `evlike` | `evrnd` | `evstat` | `pdf`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

evrnd

Extreme value random numbers

Syntax

```
R = evrnd(mu,sigma)
R = evrnd(mu,sigma,m,n,...)
R = evrnd(mu,sigma,[m,n,...])
```

Description

`R = evrnd(mu,sigma)` generates random numbers from the extreme value distribution with parameters specified by location parameter `mu` and scale parameter `sigma`. `mu` and `sigma` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of R. A scalar input for `mu` or `sigma` is expanded to a constant array with the same dimensions as the other input.

`R = evrnd(mu,sigma,m,n,...)` or `R = evrnd(mu,sigma,[m,n,...])` generates an `m`-by-`n`-by-... array containing random numbers from the extreme value distribution with parameters `mu` and `sigma`. `mu` and `sigma` can each be scalars or arrays of the same size as `R`.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima by negating R. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`evcdf` | `evfit` | `evinv` | `evlike` | `evpdf` | `evstat` | `random`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

evstat

Extreme value mean and variance

Syntax

```
[M,V] = evstat(mu,sigma)
```

Description

`[M,V] = evstat(mu,sigma)` returns the mean of and variance for the type 1 extreme value distribution with location parameter `mu` and scale parameter `sigma`. `mu` and `sigma` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other input. The default values for `mu` and `sigma` are `0` and `1`, respectively.

The type 1 extreme value distribution is also known as the Gumbel distribution. The version used here is suitable for modeling minima; the mirror image of this distribution can be used to model maxima. See “Extreme Value Distribution” on page B-40 for more details. If x has a Weibull distribution, then $X = \log(x)$ has the type 1 extreme value distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`evcdf` | `evfit` | `evinv` | `evlike` | `evpdf` | `evrnd`

Topics

“Extreme Value Distribution” on page B-40

Introduced before R2006a

expcdf

Exponential cumulative distribution function

Syntax

```
p = expcdf(x)
p = expcdf(x,mu)

[p,pLo,pUp] = expcdf(x,mu,pCov)
[p,pLo,pUp] = expcdf(x,mu,pCov,alpha)

___ = expcdf( ___ ,'upper')
```

Description

`p = expcdf(x)` returns the cumulative distribution function (cdf) of the standard exponential distribution, evaluated at the values in `x`.

`p = expcdf(x,mu)` returns the cdf of the exponential distribution with mean `mu`, evaluated at the values in `x`.

`[p,pLo,pUp] = expcdf(x,mu,pCov)` also returns the 95% confidence interval [`pLo`,`pUp`] of `p` when `mu` is an estimate with variance `pCov`.

`[p,pLo,pUp] = expcdf(x,mu,pCov,alpha)` specifies the confidence level for the confidence interval [`pLo` `pUp`] to be `100(1–alpha)`%.

`___ = expcdf( ___ ,'upper')` returns the complement of the cdf, evaluated at the values in `x`, using an algorithm that more accurately computes the extreme upper-tail probabilities than subtracting the lower tail value from 1. `'upper'` can follow any of the input argument combinations in the previous syntaxes.

Examples

Standard Exponential Distribution cdf

Compute the probability that an observation in the standard exponential distribution falls in the interval `[1 2]`.

```
p = expcdf([1 2]);
p(2) - p(1)
```

```
ans = 0.2325
```

Compute Exponential cdf

The median of the exponential distribution is `μ*log(2)`.

Confirm the median by computing the cdf of `μ*log(2)` for several different choices of μ.

```
mu = 10:10:60;
p = expcdf(log(2)*mu,mu)
```

```
p = 1×6

    0.5000    0.5000    0.5000    0.5000    0.5000    0.5000
```

The cdf of the mean is always equal to `1-1/e (~0.6321)`.

Confirm the result by computing the exponential cdf of the mean for means one through six.

```
mu = 1:6;
x = mu;
p = expcdf(x,mu)
```

```
p = 1×6

    0.6321    0.6321    0.6321    0.6321    0.6321    0.6321
```

Confidence Interval of Exponential cdf Value

Find a confidence interval estimating the probability that an observation is in the interval `[0 1]` using exponentially distributed data.

Generate a sample of `1000` random numbers drawn from the exponential distribution with mean 5.

```
rng('default') % For reproducibility
x = exprnd(5,1000,1);
```

Estimate the mean with a confidence interval.

```
[muhat,muci] = expfit(x)
```

```
muhat = 5.0129
```

```
muci = 2×1

    4.7161
    5.3387
```

Estimate the variance of the mean estimate.

```
[~,nCov] = explike(muhat,x)
```

```
nCov = 0.0251
```

Create the confidence interval estimating the probability an observation is in the interval `[0 1]`.

```
[p,pLo,pUp] = expcdf(1,muhat,nCov);
pCi = [pLo; pUp]
```

```
pCi = 2×1
```

```
    0.1710
    0.1912
```

`expcdf` calculates the confidence interval using a normal approximation for the distribution of the log estimate of the mean. Compute a more accurate confidence interval for `p` by evaluating `expcdf` on the confidence interval `muci`.

```
pCi2 = expcdf(1,muci)
```

```
pCi2 = 2×1

    0.1911
    0.1708
```

The bounds `pCi2` are reversed because a lower mean makes the event more likely and a higher mean makes the event less likely.

Complementary cdf (Tail Distribution)

Determine the probability that an observation from the exponential distribution with mean `1` is in the interval `[50 Inf]`.

```
p1 = 1 - expcdf(50,1)
```

```
p1 = 0
```

`expcdf(50,1)` is nearly `1`, so `p1` becomes `0`. Specify `'upper'` so that `expcdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = expcdf(50,1,'upper')
```

```
p2 = 1.9287e-22
```

Input Arguments

x — Values at which to evaluate cdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the cdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `mu` using an array.

If either or both of the input arguments `x` and `mu` are arrays, then the array sizes must be the same. In this case, `expcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding element in `mu`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

mu — Mean
1 (default) | positive scalar value | array of positive scalar values

Mean of the exponential distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the cdf at multiple values, specify x using an array.
- To evaluate the cdfs of multiple distributions, specify mu using an array.

If either or both of the input arguments x and mu are arrays, then the array sizes must be the same. In this case, expcdf expands each scalar input into a constant array of the same size as the array inputs. Each element in p is the cdf value of the distribution specified by the corresponding element in mu, evaluated at the corresponding element in x.

Example: [1 2 3 5]

Data Types: single | double

pCov — Variance of Mean Estimate
positive scalar value

Variance of the estimate of mu, specified as a positive scalar value.

You can estimate mu from data by using expfit or mle. You can then estimate the variance of mu by using explike. The resulting confidence interval bounds are based on a normal approximation for the distribution of the log of the mu estimate. You can get a more accurate set of bounds by applying expcdf to the confidence interval returned by expfit. For an example, see “Confidence Interval of Exponential cdf Value” on page 33-1386.

Example: 0.10

Data Types: single | double

alpha — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is 100(1–alpha)%, where alpha is the probability that the confidence interval does not contain the true value.

Example: 0.01

Data Types: single | double

Output Arguments

p — cdf values
scalar value | array of scalar values

cdf values evaluated at x, returned as a scalar value or an array of scalar values. p is the same size as x and mu after any necessary scalar expansion. Each element in p is the cdf value of the distribution specified by the corresponding element in mu, evaluated at the corresponding element in x.

pLo — Lower confidence bound for p
scalar value | array of scalar values

Lower confidence bound for `p`, returned as a scalar value or an array of scalar values. `pLo` has the same size as `p`.

`pUp` — Upper confidence bound for `p`
scalar value | array of scalar values

Upper confidence bound for `p`, returned as a scalar value or an array of scalar values. `pUp` has the same size as `p`.

More About

Exponential cdf

The exponential distribution is a one-parameter family of curves. The parameter μ is the mean.

The cdf of the exponential distribution is

$$p = F(x \mid u) = \int_0^x \frac{1}{\mu} e^{\frac{-t}{\mu}} dt = 1 - e^{\frac{-x}{\mu}}.$$

The result p is the probability that a single observation from the exponential distribution with mean μ falls in the interval $[0, x]$. A common alternative parameterization of the exponential distribution is to use λ defined as the mean number of events in an interval as opposed to μ, which is the mean wait time for an event to occur. λ and μ are reciprocals.

For more information, see “Exponential Distribution” on page B-33.

Alternative Functionality

- `expcdf` is a function specific to the exponential distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, create an `ExponentialDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `expcdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`ExponentialDistribution` | `cdf` | `expfit` | `expinv` | `explike` | `exppdf` | `exprnd` | `expstat`

Topics
“Exponential Distribution” on page B-33

Introduced before R2006a

expfit

Exponential parameter estimates

Syntax

```
muhat = expfit(data)
[muhat,muci] = expfit(data)
[muhat,muci] = expfit(data,alpha)
[...] = expfit(data,alpha,censoring)
[...] = expfit(data,alpha,censoring,freq)
```

Description

`muhat = expfit(data)` estimates the mean of exponentially distributed sample data in the vector `data`.

`[muhat,muci] = expfit(data)` also returns the 95% confidence interval for the mean parameter estimates in `muci`. The first row of `muci` is the lower bound of the confidence interval, and the second row is the upper bound.

`[muhat,muci] = expfit(data,alpha)` returns the 100(1–`alpha`)% confidence interval for the parameter estimate `muhat`, where `alpha` is a value in the range `[0 1]` specifying the width of the confidence interval. By default, `alpha` is `0.05`, which corresponds to the 95% confidence interval.

`[...] = expfit(data,alpha,censoring)` accepts a Boolean vector, `censoring`, of the same size as `data`, which is 1 for observations that are right-censored and 0 for observations that are observed exactly. `data` must be a vector in order to pass in the argument `censoring`.

`[...] = expfit(data,alpha,censoring,freq)` accepts a frequency vector, `freq` of the same size as `data`. Typically, `freq` contains integer frequencies for the corresponding elements in `data`, but can contain any nonnegative values. Pass in `[]` for `alpha`, `censoring`, or `freq` to use their default values.

Examples

The following estimates the mean `mu` of exponentially distributed data, and returns a 95% confidence interval for the estimate:

```
mu = 3;
data = exprnd(mu,100,1); % Simulated data

[muhat,muci] = expfit(data)
muhat =
    2.7511
muci =
    2.2826
    3.3813
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`expcdf` | `expinv` | `explike` | `exppdf` | `exprnd` | `expstat` | `mle`

Introduced before R2006a

ExhaustiveSearcher

Create exhaustive nearest neighbor searcher

Description

`ExhaustiveSearcher` model objects store the training data, distance metric, and parameter values of the distance metric for an exhaustive nearest neighbor search. The exhaustive search algorithm finds the distance from each query observation to all n observations in the training data, which is an n-by-K numeric matrix.

Once you create an `ExhaustiveSearcher` model object, find neighboring points in the training data to the query data by performing a nearest neighbor search using `knnsearch` or a radius search using `rangesearch`. The exhaustive search algorithm is more efficient than the *K*d-tree algorithm when K is large (that is, $K > 10$), and it is more flexible than the *K*d-tree algorithm with respect to distance metric choices. The `ExhaustiveSearcher` model object also supports sparse data.

Creation

Use either the `createns` function or the `ExhaustiveSearcher` function (described here) to create an `ExhaustiveSearcher` object. Both functions use the same syntax except that the `createns` function has the `'NSMethod'` name-value pair argument, which you use to choose the nearest neighbor search method. The `createns` function also creates a `KDTreeSearcher` object. Specify `'NSMethod','exhaustive'` to create an `ExhaustiveSearcher` object. The default is `'exhaustive'` if $K > 10$, the training data is sparse, or the distance metric is not the Euclidean, city block, Chebychev, or Minkowski.

Syntax

```
Mdl = ExhaustiveSearcher(X)
Mdl = ExhaustiveSearcher(X,Name,Value)
```

Description

`Mdl = ExhaustiveSearcher(X)` creates an exhaustive nearest neighbor searcher object (`Mdl`) using the n-by-K numeric matrix of training data (`X`).

`Mdl = ExhaustiveSearcher(X,Name,Value)` specifies additional options using one or more name-value pair arguments. You can specify the distance metric and set the distance metric parameter (`DistParameter`) property. For example, `ExhaustiveSearcher(X,'Distance','chebychev')` creates an exhaustive nearest neighbor searcher object that uses the Chebychev distance. To specify `DistParameter`, use the `Cov`, `P`, or `Scale` name-value pair argument.

Input Arguments

`X` — Training data
numeric matrix

Training data that prepares the exhaustive searcher algorithm, specified as a numeric matrix. `X` has *n* rows, each corresponding to an observation (that is, an instance or example), and *K* columns, each corresponding to a predictor (that is, a feature).

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distance','mahalanobis','Cov',eye(3)` specifies to use the Mahalanobis distance when searching for nearest neighbors and a 3-by-3 identity matrix for the covariance matrix in the Mahalanobis distance metric.

`Distance` — Distance metric

`'euclidean'` (default) | character vector | string scalar | custom distance function

Distance metric used when you call `knnsearch` or `rangesearch` to find nearest neighbors for future query points, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle.

This table describes the supported distance metrics specified as character vectors or string scalars.

Value	Description
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cityblock'`	City block distance.
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as row vectors).
`'euclidean'`	Euclidean distance.
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value pair argument.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix. To change the value of the covariance matrix, use the `'Cov'` name-value pair argument.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between rows in `Mdl.X` and the query matrix is scaled by dividing by the corresponding element of the standard deviation computed from `Mdl.X`. To specify another scaling, use the `'Scale'` name-value pair argument.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

For more details, see “Distance Metrics” on page 18-12.

You can specify a function handle for a custom distance metric by using `@` (for example, `@distfun`). A custom distance function must:

- Have the form `function D2 = distfun(ZI, ZJ)`.
- Take as arguments:
  - A 1-by-*K* vector `ZI` containing a single row from `X` or from the query points `Y`, where *K* is the number of columns in `X`.
  - An *m*-by-*K* matrix `ZJ` containing multiple rows of `X` or `Y`, where *m* is a positive integer.
- Return an *m*-by-1 vector of distance `D2`, where `D2(j)` is the distance between the observations `ZI` and `ZJ(j,:)`.

The software does not use the distance metric for creating an `ExhaustiveSearcher` model object, so you can alter the distance metric by using dot notation after creating the object.

Example: `'Distance','mahalanobis'`

`Cov` — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a *K*-by-*K* positive definite matrix, where *K* is the number of columns in `X`. This argument is valid only if `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

`P` — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

`Scale` — Scale parameter value for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector of length *K*, where *K* is the number of columns in `X`. The software scales each difference between the training and query data using the corresponding element of `Scale`. This argument is valid only if `'Distance'` is `'seuclidean'`.

Example: `'Scale',quantile(X,0.75) - quantile(X,0.25)`

Data Types: `single` | `double`

Properties

`X` — Training data
numeric matrix

This property is read-only.

Training data that prepares the exhaustive searcher algorithm, specified as a numeric matrix. `X` has *n* rows, each corresponding to an observation (that is, an instance or example), and *K* columns, each corresponding to a predictor (that is, a feature).

The input argument `X` of `createns` or `ExhaustiveSearcher` sets this property.

Data Types: `single` | `double`

`Distance` — Distance metric
character vector | string scalar | custom distance function

Distance metric used when you call `knnsearch` or `rangesearch` to find nearest neighbors for future query points, specified as a character vector or string scalar (`'chebychev'`, `'cityblock'`, `'correlation'`, `'cosine'`, `'euclidean'`, `'hamming'`, `'jaccard'`, `'minkowski'`, `'mahalanobis'`, `'seuclidean'`, or `'spearman'`), or a function handle.

The `'Distance'` name-value pair argument of `createns` or `ExhaustiveSearcher` sets this property.

The software does not use the distance metric for creating an `ExhaustiveSearcher` model object, so you can alter it by using dot notation.

`DistParameter` — Distance metric parameter values
`[]` | positive scalar

Distance metric parameter values, specified as empty (`[]`) or a positive scalar.

This table describes the distance parameters of the supported distance metrics.

Distance Metric	Parameter Description
`'mahalanobis'`	A positive definite matrix representing the covariance matrix used for computing the Mahalanobis distance. By default, the software sets the covariance using `cov(Mdl.X,'omitrows')`. The `'Cov'` name-value pair argument of `createns` or `ExhaustiveSearcher` sets this property. You can alter `DistParameter` by using dot notation, for example, `Mdl.DistParameter = CovNew`, where `CovNew` is a *K*-by-*K* positive definite numeric matrix.
`'minkowski'`	A positive scalar indicating the exponent of the Minkowski distance. By default, the exponent is `2`. The `'P'` name-value pair argument of `createns` or `ExhaustiveSearcher` sets this property. You can alter `DistParameter` by using dot notation, for example, `Mdl.DistParameter = PNew`, where `PNew` is a positive scalar.

Distance Metric	Parameter Description
`'seuclidean'`	A positive numeric vector indicating the values used by the software to scale the predictors when computing the standardized Euclidean distance. By default, the software: **1** Estimates the standard deviation of each predictor (column) of `X` using `scale = std(Mdl.X,'omitnan')` **2** Scales each coordinate difference between the rows in `X` and the query matrix by dividing by the corresponding element of `scale` The `'Scale'` name-value pair argument of `createns` or `ExhaustiveSearcher` sets this property. You can alter `DistParameter` by using dot notation, for example, `Mdl.DistParameter = sNew`, where `sNew` is a K-dimensional positive numeric vector.

If `Mdl.Distance` is not one of the parameters listed in this table, then `Mdl.DistParameter` is `[]`, which means that the specified distance metric formula has no parameters.

Data Types: `single` | `double`

Object Functions

knnsearch Find k-nearest neighbors using searcher object
rangesearch Find all neighbors within specified distance using searcher object

Examples

Train Default Exhaustive Nearest Neighbor Searcher

Load Fisher's iris data set.

```
load fisheriris
X = meas;
[n,k] = size(X)
```

```
n = 150
```

```
k = 4
```

`X` has 150 observations and 4 predictors.

Prepare an exhaustive nearest neighbor searcher using the entire data set as training data.

```
Mdl1 = ExhaustiveSearcher(X)
```

```
Mdl1 =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl1` is an `ExhaustiveSearcher` model object, and its properties appear in the Command Window. The object contains information about the trained algorithm, such as the distance metric. You can alter property values using dot notation.

Alternatively, you can prepare an exhaustive nearest neighbor searcher by using `createns` and specifying `'exhaustive'` as the search method.

```
Mdl2 = createns(X,'NSMethod','exhaustive')
```

```
Mdl2 =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl2` is also an `ExhaustiveSearcher` model object, and it is equivalent to `Mdl1`.

To search `X` for the nearest neighbors to a batch of query data, pass the `ExhaustiveSearcher` model object and the query data to `knnsearch` or `rangesearch`.

Specify the Mahalanobis Distance for Nearest Neighbor Search

Load Fisher's iris data set. Focus on the petal dimensions.

```
load fisheriris
X = meas(:,[3 4]); % Predictors
```

Prepare an exhaustive nearest neighbor searcher. Specify the Mahalanobis distance metric.

```
Mdl = createns(X,'Distance','mahalanobis')
```

```
Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'mahalanobis'
    DistParameter: [2x2 double]
                X: [150x2 double]
```

Because the distance metric is Mahalanobis, `createns` creates an `ExhaustiveSearcher` model object by default.

Access properties of `Mdl` by using dot notation. For example, use `Mdl.DistParameter` to access the Mahalanobis covariance parameter.

```
Mdl.DistParameter
```

```
ans = 2×2

    3.1163    1.2956
    1.2956    0.5810
```

You can pass query data and `Mdl` to:

- `knnsearch` to find indices and distances of nearest neighbors
- `rangesearch` to find indices of all nearest neighbors within a distance that you specify

Alter Properties of ExhaustiveSearcher Model

Create an `ExhaustiveSearcher` model object and alter the `Distance` property by using dot notation.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
```

Train a default exhaustive searcher algorithm using the entire data set as training data.

```
Mdl = ExhaustiveSearcher(X)
```

```
Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

Specify that the neighbor searcher use the Mahalanobis metric to compute the distances between the training and query data.

```
Mdl.Distance = 'mahalanobis'
```

```
Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'mahalanobis'
    DistParameter: [4x4 double]
                X: [150x4 double]
```

You can pass `Mdl` and the query data to either `knnsearch` or `rangesearch` to find the nearest neighbors to the points in the query data based on the Mahalanobis distance.

Search for Nearest Neighbors of Query Data Using Mahalanobis Distance

Create an exhaustive searcher object by using the `createns` function. Pass the object and query data to the `knnsearch` function to find k-nearest neighbors.

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng('default');             % For reproducibility
n = size(meas,1);           % Sample size
qIdx = randsample(n,5);     % Indices of query data
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Prepare an exhaustive nearest neighbor searcher using the training data. Specify the Mahalanobis distance for finding nearest neighbors.

```
Mdl = createns(X,'Distance','mahalanobis')
```

```
Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'mahalanobis'
    DistParameter: [4x4 double]
                X: [145x4 double]
```

Because the distance metric is Mahalanobis, `createns` creates an `ExhaustiveSearcher` model object by default.

The software uses the covariance matrix of the predictors (columns) in the training data for computing the Mahalanobis distance. To display this value, use `Mdl.DistParameter`.

```
Mdl.DistParameter
```

```
ans = 4×4

    0.6547   -0.0368    1.2320    0.5026
   -0.0368    0.1914   -0.3227   -0.1193
    1.2320   -0.3227    3.0671    1.2842
    0.5026   -0.1193    1.2842    0.5800
```

Find the indices of the training data (`Mdl.X`) that are the two nearest neighbors of each point in the query data (`Y`).

```
IdxNN = knnsearch(Mdl,Y,'K',2)
```

```
IdxNN = 5×2

     5     6
    98    95
   104   128
   135    65
   102   115
```

Each row of `IdxNN` corresponds to a query data observation. The column order corresponds to the order of the nearest neighbors with respect to ascending distance. For example, based on the Mahalanobis metric, the second nearest neighbor of `Y(3,:)` is `X(128,:)`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `knnsearch` and `rangesearch` functions support code generation.
- When you train an `ExhaustiveSearcher` model object, the value of the `'Distance'` name-value pair argument cannot be a custom distance function.

For more information, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Nearest Neighbor Searcher” on page 32-19.

See Also

`KDTreeSearcher` | `createns`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2010a

expinv

Exponential inverse cumulative distribution function

Syntax

```
x = expinv(p)
x = expinv(p,mu)

[x,xLo,xUp] = expinv(p,mu,pCov)
[x,xLo,xUp] = expinv(p,mu,pCov,alpha)
```

Description

`x = expinv(p)` returns the inverse cumulative distribution function (icdf) of the standard exponential distribution, evaluated at the values in `p`.

`x = expinv(p,mu)` returns the icdf of the exponential distribution with mean `mu`, evaluated at the values in `p`.

`[x,xLo,xUp] = expinv(p,mu,pCov)` also returns the 95% confidence interval [`xLo`,`xUp`] of `x` when `mu` is an estimate with variance `pCov`.

`[x,xLo,xUp] = expinv(p,mu,pCov,alpha)` specifies the confidence level for the confidence interval [`xLo` `xUp`] to be `100(1–alpha)`%.

Examples

Compute Exponential icdf

Assume that the lifetime of light bulbs are exponentially distributed with a mean of `700` hours. Find the median lifetime using `expinv`.

```
expinv(0.50,700)
```

```
ans = 485.2030
```

Half of the light bulbs will burn out within the first 485 hours of use.

Confidence Interval of Exponential icdf Value

Find a confidence interval estimating the median using exponentially distributed data.

Generate a sample of `1000` exponentially distributed random numbers with mean 5.

```
rng('default') % For reproducibility
x = exprnd(5,100,1);
```

Estimate the mean with a confidence interval.

```
[muhat,muci] = expfit(x)
```

```
muhat = 4.5852
```

```
muci = 2×1

    3.8043
    5.6355
```

Estimate the variance of the mean estimate.

```
[~,pCov] = explike(muhat,x)
```

```
pCov = 0.2102
```

Create a confidence interval for the median.

```
[x,xLo,xUp] = expinv(0.5,muhat,pCov);
xCi = [xLo; xUp]
```

```
xCi = 2×1

    2.6126
    3.8664
```

Alternatively, compute a more accurate confidence interval for `x` by evaluating `expinv` on the confidence interval `muci`.

```
xCi2 = expinv(0.5,muci)
```

```
xCi2 = 2×1

    2.6369
    3.9062
```

Input Arguments

p — Probability values at which to evaluate icdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the icdf, specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `mu` using an array.

If either or both of the input arguments `p` and `mu` are arrays, then the array sizes must be the same. In this case, `expinv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding element in `mu`, evaluated at the corresponding element in `p`.

Example: `[0.1,0.5,0.9]`

Data Types: `single` | `double`

mu — Mean
1 (default) | positive scalar value | array of positive scalar values

Mean of the exponential distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the icdf at multiple values, specify p using an array.
- To evaluate the icdfs of multiple distributions, specify mu using an array.

If either or both of the input arguments p and mu are arrays, then the array sizes must be the same. In this case, expinv expands each scalar input into a constant array of the same size as the array inputs. Each element in x is the icdf value of the distribution specified by the corresponding element in mu, evaluated at the corresponding element in p.

Example: [1 2 3 5]

Data Types: single | double

pCov — Variance of mean estimate
positive scalar value

Variance of the estimate of mu, specified as a positive scalar.

You can estimate mu from data by using expfit. You can then estimate the variance of mu by using explike. The resulting confidence interval bounds are based on a normal approximation for the distribution of the log of the mu estimate. You can get a more accurate set of bounds by applying expinv to the confidence interval returned by expfit. For an example, see “Confidence Interval of Exponential icdf Value” on page 33-1402.

Example: 0.10

Data Types: single | double

alpha — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is 100(1–alpha)%, where alpha is the probability that the confidence interval does not contain the true value.

Example: 0.01

Data Types: single | double

Output Arguments

x — icdf values
scalar value | array of scalar values

icdf values evaluated at the probability values in p, returned as a scalar value or an array of scalar values. x is the same size as p and mu after any necessary scalar expansion. Each element in x is the icdf value of the distribution specified by the corresponding element in mu, evaluated at the corresponding element in p.

xLo — Lower confidence bound for x
scalar value | array of scalar values

Lower confidence bound for `x`, returned as a scalar value or an array of scalar values. `xLo` has the same size as `x`.

xUp — Upper confidence bound for x
scalar value | array of scalar values

Upper confidence bound for `x`, returned as a scalar value or an array of scalar values. `xUp` has the same size as `x`.

More About

Exponential icdf

The exponential distribution is a one-parameter family of curves. The parameter μ is the mean.

The icdf of the exponential distribution is

$$x = F^{-1}(p \mid \mu) = -\mu \ln(1-p).$$

The result x is the value such that an observation from an exponential distribution with parameter μ will falls in the range $[0,x]$ with probability p. A common alternative parameterization of the exponential distribution is to use λ defined as the mean number of events in an interval as opposed to μ, which is the mean wait time for an event to occur. λ and μ are reciprocals.

For more information, see “Exponential Distribution” on page B-33.

Alternative Functionality

- `expinv` is a function specific to the exponential distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use `icdf`, create an `ExponentialDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `expinv` is faster than the generic function `icdf`.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`ExponentialDistribution` | `expcdf` | `expfit` | `explike` | `exppdf` | `exprnd` | `expstat` | `icdf`

Topics
“Exponential Distribution” on page B-33

Introduced before R2006a

explike

Exponential negative log-likelihood

Syntax

```
nlogL = explike(param,data)
[nlogL,avar] = explike(param,data)
[...] = explike(param,data,censoring)
[...] = explike(param,data,censoring,freq)
```

Description

`nlogL = explike(param,data)` returns the negative of the log-likelihood for the exponential distribution. `param` is the mean parameter, `mu`. `nlogL` is a scalar.

`[nlogL,avar] = explike(param,data)` returns the inverse of Fisher's information, `avar`, a scalar. If the input parameter value in `param` is the maximum likelihood estimate, `avar` is its asymptotic variance. `avar` is based on the observed Fisher's information, not the expected information.

`[...] = explike(param,data,censoring)` accepts a Boolean vector, `censoring`, of the same size as `data`, which is `1` for observations that are right-censored and `0` for observations that are observed exactly.

`[...] = explike(param,data,censoring,freq)` accepts a frequency vector, `freq`, of the same size as `data`. The vector `freq` typically contains integer frequencies for the corresponding elements in `data`, but can contain any nonnegative values. Pass in `[]` for `censoring` to use its default value.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`expcdf` | `expfit` | `expinv` | `exppdf` | `exprnd` | `expstat`

Topics

“Exponential Distribution” on page B-33

Introduced before R2006a

export

Class: `dataset`

(Not Recommended) Write dataset array to file

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
export(DS,'file',filename)
export(DS)
export(DS,'file',filename,'Delimiter',delim)
export(DS,'XLSfile',filename)
export(DS,'XPTFile',filename)
export(DS,...,'WriteVarNames',false)
export(DS,...,'WriteObsNames',false)
```

Description

`export(DS,'file',filename)` writes the dataset array `DS` to a tab-delimited text file, including variable names and observation names, if present. If the observation names exist, the name in the first column of the first line of the file is the first dimension name for the dataset (by default, `'Observations'`). `export` overwrites any existing file named `filename`.

`export(DS)` writes to a text file whose default name is the name of the dataset array `DS` appended by `'.txt'`. If `export` cannot construct the file name from the dataset array input, it writes to the file `'dataset.txt'`. `export` overwrites any existing file.

`export(DS,'file',filename,'Delimiter',delim)` writes the dataset array `DS` to a text file using the delimiter `delim`. `delim` must be one of the following:

- `' '` or `'space'`
- `'\t'` or `'tab'`
- `','` or `'comma'`
- `';'` or `'semi'`
- `'|'` or `'bar'`

`export(DS,'XLSfile',filename)` writes the dataset array `DS` to a Microsoft® Excel spreadsheet file, including variable names and observation names (if present). You can specify the `'Sheet'` and `'Range'` parameter name/value pairs, with parameter values as accepted by the `xlsread` function. Since `export` uses the `xlswrite` function internally, this syntax is only compatible with Microsoft Excel for Windows, and does not work on a Mac. For more information, see `xlswrite`.

`export(DS,'XPTFile',filename)` writes the dataset array `DS` to a `SAS XPORT` format file. When writing to an `XPORT` format file, variables must be scalar valued. `export` saves observation names to a variable called `obsnames`, unless the `WriteObsNames` parameter described below is `false`. The

XPORT format restricts the length of variable names to eight characters; longer variable names are truncated.

`export(DS,...,'WriteVarNames',false)` does not write the variable names to the text file. `export(DS,...,'WriteVarNames',true)` is the default, writing the names as column headings in the first line of the file.

`export(DS,...,'WriteObsNames',false)` does not write the observation names to the text file. `export(DS,...,'WriteObsNames',true)` is the default, writing the names as the first column of the file.

In some cases, `export` creates a text file that does not represent `A` exactly, as described below. If you use `dataset` to read the file back into MATLAB, the new dataset array may not have exactly the same contents as the original dataset array. Save `A` as a MAT-file if you need to import it again as a dataset array.

`export` writes out numeric variables using `long g format`, and categorical or character variables as unquoted text.

For non-character variables with more than one column, `export` writes out multiple delimiter-separated fields on each line, and constructs suitable column headings for the first line of the file.

`export` writes out variables that have more than two dimensions as a single empty field in each line of the file.

For cell-valued variables, `export` writes out the contents of each cell only when the cell contains a single row, and writes out a single empty field otherwise.

In some cases, `export` creates a file that cannot be read back into MATLAB using `dataset`. Writing a dataset array that contains a cell-valued variable whose cell contents are not scalars results in a mismatch in the file between the number of fields on each line and the number of column headings on the first line. Writing a dataset array that contains a cell-valued variable whose cell contents are not all the same length results in a different number of fields on each line in the file. Therefore, if you might need to import a dataset array again, save it as a `.mat` file.

Examples

Move data between external text files and dataset arrays in the MATLAB workspace:

```
A = dataset('file','sat2.dat','delimiter',',')
A =
    Test                  Gender          Score
    'Verbal'              'Male'          470
    'Verbal'              'Female'        530
    'Quantitative'        'Male'          520
    'Quantitative'        'Female'        480

export(A(A.Score > 500,:),'file','HighScores.txt')

B = dataset('file','HighScores.txt','delimiter','\t')
B =
    Test                  Gender          Score
    'Verbal'              'Female'        530
    'Quantitative'        'Male'          520
```

See Also

`dataset`

exppdf

Exponential probability density function

Syntax

```
y = exppdf(x)
y = exppdf(x,mu)
```

Description

`y = exppdf(x)` returns the probability density function (pdf) of the standard exponential distribution, evaluated at the values in `x`.

`y = exppdf(x,mu)` returns the pdf of the exponential distribution with mean `mu`, evaluated at the values in `x`.

Examples

Compute Exponential pdf

Compute the density of the observed value 5 in the standard exponential distribution.

```
y1 = exppdf(5)
```

```
y1 = 0.0067
```

Compute the density of the observed value 5 in the exponential distributions specified by means 1 through 5.

```
y2 = exppdf(5,1:5)
```

```
y2 = 1×5

    0.0067    0.0410    0.0630    0.0716    0.0736
```

Compute the density of the observed values 1 through 5 in the exponential distributions specified by means 1 through 5, respectively.

```
y3 = exppdf(1:5,1:5)
```

```
y3 = 1×5

    0.3679    0.1839    0.1226    0.0920    0.0736
```

Input Arguments

x — Values at which to evaluate pdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the pdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `mu` using an array.

If either or both of the input arguments `x` and `mu` are arrays, then the array sizes must be the same. In this case, `exppdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `mu`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

`mu` — mean
1 (default) | positive scalar value | array of positive scalar values

Mean of the exponential distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `mu` using an array.

If either or both of the input arguments `x` and `mu` are arrays, then the array sizes must be the same. In this case, `exppdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `mu`, evaluated at the corresponding element in `x`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x` and `mu` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `mu`, evaluated at the corresponding element in `x`.

More About

Exponential pdf

The exponential distribution is a one-parameter family of curves. The parameter μ is the mean.

The pdf of the exponential distribution is

$$y = f(x \mid \mu) = \frac{1}{\mu} e^{\frac{-x}{\mu}}.$$

A common alternative parameterization of the exponential distribution is to use λ defined as the mean number of events in an interval as opposed to μ, which is the mean wait time for an event to occur. λ and μ are reciprocals.

For more information, see “Exponential Distribution” on page B-33.

Alternative Functionality

- `exppdf` is a function specific to the exponential distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create an `ExponentialDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `exppdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ExponentialDistribution` | `expcdf` | `expfit` | `expinv` | `explike` | `exprnd` | `expstat` | `pdf`

Topics

“Exponential Distribution” on page B-33

Introduced before R2006a

exprnd

Exponential random numbers

Syntax

```
r = exprnd(mu)
r = exprnd(mu,sz1,...,szN)
r = exprnd(mu,sz)
```

Description

`r = exprnd(mu)` generates a random number from the exponential distribution with mean `mu`.

`r = exprnd(mu,sz1,...,szN)` generates an array of random numbers from the exponential distribution, where `sz1,...,szN` indicates the size of each dimension.

`r = exprnd(mu,sz)` generates an array of random numbers from the exponential distribution, where vector `sz` specifies `size(r)`.

Examples

Generate Exponential Random Number

Generate a single random number from the exponential distribution with mean 5.

```
r = exprnd(5)
```

```
r = 1.0245
```

Generate Array of Exponential Random Numbers

Generate a 1-by-6 array of exponential random numbers with unit mean.

```
mu1 = ones(1,6); % 1-by-6 array of ones
r1 = exprnd(mu1)
```

```
r1 = 1×6

    0.2049    0.0989    2.0637    0.0906    0.4583    2.3275
```

By default, `exprnd` generates an array that is the same size as `mu`.

If you specify `mu` as a scalar, then `exprnd` expands it into a constant array with dimensions specified by `sz1,...,szn`.

Generate a 2-by-6 array of exponential random numbers with mean 3.

```
mu2 = 3;
sz1 = 2;
sz2 = 6;
r2 = exprnd(mu2,sz1,sz2)
```

```
r2 = 2×6

    3.8350    0.1303    5.5428    0.1313    0.6684    2.5899
    1.8106    0.1072    0.0895    2.1685    5.8582    0.2641
```

If you specify both `mu` and `sz1,...,szn` as arrays, then the dimensions specified by `sz1,...,szn` must match the dimension of `mu`.

Generate a 1-by-6 array of exponential random numbers with means 5 through 10.

```
mu3 = 5:10;
sz = [1 6];
r3 = exprnd(mu3,sz)
```

```
r3 = 1×6

    1.1647    0.2481    2.9539   26.6582    1.4719    0.6829
```

Input Arguments

`mu` — mean

1 (default) | positive scalar value | array of positive scalar values

Mean of the exponential distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `mu` using an array. Each element in `r` is the random number generated from the distribution specified by the corresponding element in `mu`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)

integers

Size of each dimension, specified as separate arguments of integers.

If `mu` is an array, then the specified dimensions `sz1,...,szN` must match the dimensions of `mu`. The default values of `sz1,...,szN` are the dimensions of `mu`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `exprnd` ignores trailing dimensions with a size of 1. For example, `exprnd(4,3,1,1,1)` produces a 3-by-1 vector of random numbers from the distribution with mean 4.

Example: `2,4`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If `mu` is an array, then the specified dimensions `sz` must match the dimensions of `mu`. The default values of `sz` are the dimensions of `mu`.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `exprnd` ignores trailing dimensions with a size of 1. For example, `exprnd(4,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the distribution with mean 4.

Example: `[2 4]`

Data Types: `single` | `double`

Output Arguments

`r` — Exponential random numbers
nonnegative scalar value | array of nonnegative scalar values

Exponential random numbers, returned as a nonnegative scalar value or an array of nonnegative scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding element in `mu`.

Alternative Functionality

- `exprnd` is a function specific to the exponential distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, create an `ExponentialDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `exprnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ExponentialDistribution` | `expcdf` | `expfit` | `expinv` | `explike` | `exppdf` | `expstat` | `random`

Topics

“Exponential Distribution” on page B-33

Introduced before R2006a

expstat

Exponential mean and variance

Syntax

```
[m,v] = expstat(mu)
```

Description

`[m,v] = expstat(mu)` returns the mean of and variance for the exponential distribution with parameters `mu`. `mu` can be a vectors, matrix, or multidimensional array. The mean of the exponential distribution is μ, and the variance is μ^2.

Examples

```
[m,v] = expstat([1 10 100 1000])
m =
      1     10     100     1000
v =
      1     100    10000   1000000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`expcdf` | `expfit` | `expinv` | `explike` | `exppdf` | `exprnd` | `expstat`

Topics

“Exponential Distribution” on page B-33

Introduced before R2006a

factoran

Factor analysis

Syntax

```
lambda = factoran(X,m)
[lambda,psi] = factoran(X,m)
[lambda,psi,T] = factoran(X,m)
[lambda,psi,T,stats] = factoran(X,m)
[lambda,psi,T,stats,F] = factoran(X,m)
___  = factoran(X,m,Name,Value)
```

Description

`factoran` computes the maximum likelihood estimate (MLE) of the factor loadings matrix Λ in the factor analysis model

$$x = \mu + \Lambda f + e$$

where x is a vector of observed variables, μ is a constant vector of means, Λ is a constant `d`-by-`m` matrix of factor loadings, f is a vector of independent, standardized common factors, and e is a vector of independent specific factors. x, μ, and e each has length `d`. f has length `m`.

Alternatively, the factor analysis model can be specified as

$$\mathrm{cov}(x) = \Lambda\Lambda^T + \Psi$$

where $\Psi = \mathrm{cov}(e)$ is a `d`-by-`d` diagonal matrix of specific variances.

For the uses of `factoran` and its relation to `pca`, see “Perform Factor Analysis on Exam Grades” on page 15-180.

`lambda = factoran(X,m)` returns the factor loadings matrix `lambda` for the data matrix `X` with `m` common factors.

`[lambda,psi] = factoran(X,m)` also returns maximum likelihood estimates of the specific variances.

`[lambda,psi,T] = factoran(X,m)` also returns the `m`-by-`m` factor loadings rotation matrix `T`.

`[lambda,psi,T,stats] = factoran(X,m)` also returns the structure `stats` containing information relating to the null hypothesis H_0 that the number of common factors is `m`.

`[lambda,psi,T,stats,F] = factoran(X,m)` also returns predictions of the common factors (factor scores).

`___  = factoran(X,m,Name,Value)` modifies the model fit and outputs using one or more name-value pair arguments, for any output arguments in the previous syntaxes. For example, you can specify that the `X` data is a covariance matrix.

Examples

Factor Analysis of Artificial Data

Create some pseudorandom raw data.

```
rng default % For reproducibility
n = 100;
X1 = 5 + 3*rand(n,1); % Factor 1
X2 = 20 - 5*rand(n,1); % Factor 2
```

Create six data vectors from the raw data, and add random noise.

```
Y1 = 2*X1 + 3*X2 + randn(n,1);
Y2 = 4*X1 + X2 + 2*randn(n,1);
Y3 = X1 - X2 + 3*randn(n,1);
Y4 = -2*X1 + 4*X2 + 4*randn(n,1);
Y5 = 3*(X1 + X2) + 5*randn(n,1);
Y6 = X1 - X2/2 + 6*randn(n,1);
```

Create a data matrix from the data vectors.

```
X = [Y1,Y2,Y3,Y4,Y5,Y6];
```

Extract the two factors from the noisy data matrix `X` using `factoran`. Display the outputs.

```
m = 2;
[lambda,psi,T,stats,F] = factoran(X,m);
disp(lambda)
```

```
    0.8666    0.4828
    0.8688   -0.0998
   -0.0131   -0.5412
    0.2150    0.8458
    0.7040    0.2678
   -0.0806   -0.2883
```

```
disp(psi)
```

```
    0.0159
    0.2352
    0.7070
    0.2385
    0.4327
    0.9104
```

```
disp(T)
```

```
    0.8728    0.4880
    0.4880   -0.8728
```

```
disp(stats)
```

```
    loglike: -0.0531
        dfe: 4
      chisq: 5.0335
          p: 0.2839
```

```
disp(F(1:10,:))
```

```
    1.8845   -0.6568
   -0.1714   -0.8113
   -1.0534    2.0743
    1.0390   -1.1784
    0.4309    0.9907
   -1.1823    0.6570
   -0.2129    1.1898
   -0.0844   -0.7421
    0.5854   -1.1379
    0.8279   -1.9624
```

View the correlation matrix of the data.

```
corrX = corr(X)
```

```
corrX = 6×6

    1.0000    0.7047   -0.2710    0.5947    0.7391   -0.2126
    0.7047    1.0000    0.0203    0.1032    0.5876    0.0289
   -0.2710    0.0203    1.0000   -0.4793   -0.1495    0.1450
    0.5947    0.1032   -0.4793    1.0000    0.3752   -0.2134
    0.7391    0.5876   -0.1495    0.3752    1.0000   -0.2030
   -0.2126    0.0289    0.1450   -0.2134   -0.2030    1.0000
```

Compare `corrX` to its corresponding values returned by `factoran`, `lambda*lambda' + diag(psi)`.

```
C0 = lambda*lambda' + diag(psi)
```

```
C0 = 6×6

    1.0000    0.7047   -0.2726    0.5946    0.7394   -0.2091
    0.7047    1.0000    0.0426    0.1023    0.5849   -0.0413
   -0.2726    0.0426    1.0000   -0.4605   -0.1542    0.1571
    0.5946    0.1023   -0.4605    1.0000    0.3779   -0.2611
    0.7394    0.5849   -0.1542    0.3779    1.0000   -0.1340
   -0.2091   -0.0413    0.1571   -0.2611   -0.1340    1.0000
```

`factoran` obtains `lambda` and `psi` that correspond closely to the correlation matrix of the original data.

View the results without using rotation.

```
[lambda,psi,T,stats,F] = factoran(X,m,'Rotate','none');
disp(lambda)
```

```
    0.9920    0.0015
    0.7096    0.5111
   -0.2755    0.4659
    0.6004   -0.6333
    0.7452    0.1098
   -0.2111    0.2123
```

```
disp(psi)
```

```
    0.0159
    0.2352
```

```
    0.7070
    0.2385
    0.4327
    0.9104
```

```
disp(T)
```

```
     1     0
     0     1
```

```
disp(stats)
```

```
    loglike: -0.0531
        dfe: 4
      chisq: 5.0335
          p: 0.2839
```

```
disp(F(1:10,:))
```

```
    1.3243    1.4929
   -0.5456    0.6245
    0.0928   -2.3246
    0.3318    1.5356
    0.8596   -0.6544
   -0.7114   -1.1504
    0.3947   -1.1424
   -0.4358    0.6065
   -0.0444    1.2789
   -0.2350    2.1169
```

Compute the factors using only the covariance matrix of X.

```
X2 = cov(X);
[lambda2,psi2,T2,stats2] = factoran(X2,m,'Xtype','covariance','Nobs',n)
```

```
lambda2 = 6×2

    0.8666    0.4828
    0.8688   -0.0998
   -0.0131   -0.5412
    0.2150    0.8458
    0.7040    0.2678
   -0.0806   -0.2883


psi2 = 6×1

    0.0159
    0.2352
    0.7070
    0.2385
    0.4327
    0.9104


T2 = 2×2

    0.8728    0.4880
    0.4880   -0.8728
```

```
stats2 = struct with fields:
    loglike: -0.0531
        dfe: 4
      chisq: 5.0335
          p: 0.2839
```

The results are the same as with the raw data, except `factoran` cannot compute the factor scores matrix `F` for covariance data.

Estimate and Plot Factor Loadings

Load the sample data.

```
load carbig
```

Define the variable matrix.

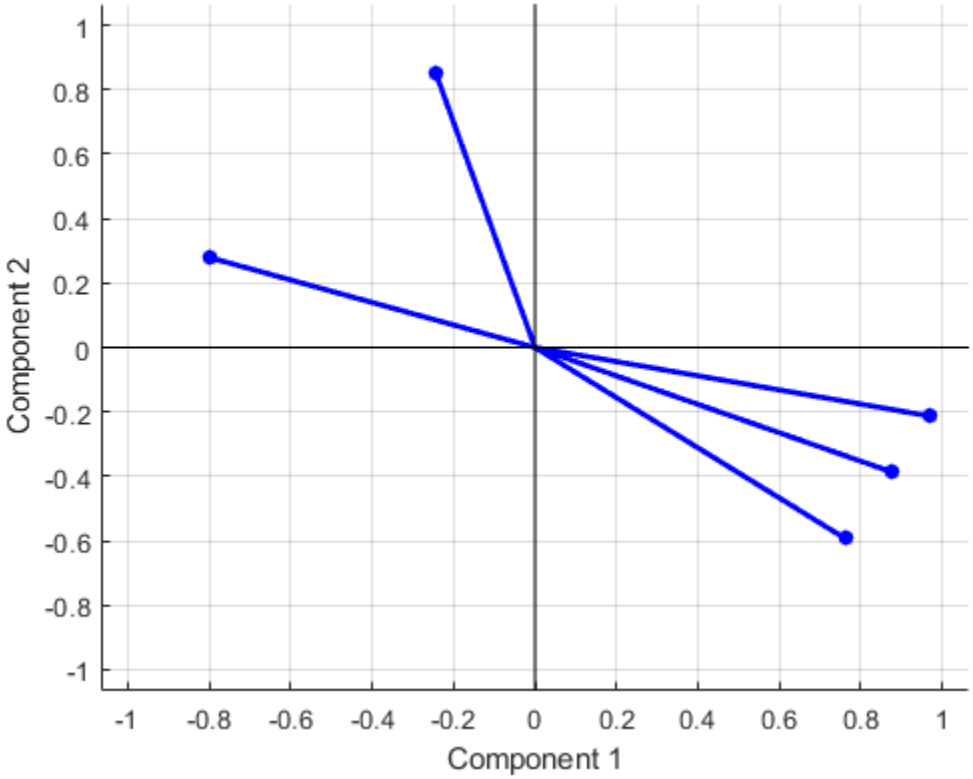
```
X = [Acceleration Displacement Horsepower MPG Weight];
X = X(all(~isnan(X),2),:);
```

Estimate the factor loadings using a minimum mean squared error prediction for a factor analysis with two common factors.

```
[Lambda,Psi,T,stats,F] = factoran(X,2,'Scores','regression');
inv(T'*T);   % Estimated correlation matrix of F, == eye(2)
Lambda*Lambda' + diag(Psi); % Estimated correlation matrix
Lambda*inv(T);              % Unrotate the loadings
F*T';                       % Unrotate the factor scores
```

Create a biplot of two factors.

```
biplot(Lambda,'LineWidth',2,'MarkerSize',20)
```



Estimate the factor loadings using the covariance (or correlation) matrix.

```
[Lambda,Psi,T] = factoran(cov(X),2,'Xtype','covariance')
```

```
Lambda = 5×2

   -0.2432   -0.8500
    0.8773    0.3871
    0.7618    0.5930
   -0.7978   -0.2786
    0.9692    0.2129


Psi = 5×1

    0.2184
    0.0804
    0.0680
    0.2859
    0.0152


T = 2×2

    0.9476    0.3195
    0.3195   -0.9476
```

(You could instead use `corrcoef(X)` instead of `cov(X)` to create the data for `factoran`.) Although the estimates are the same, the use of a covariance matrix rather than raw data prevents you from requesting the scores or significance level.

Use promax rotation.

```
[Lambda,Psi,T,stats,F] = factoran(X,2,'Rotate','promax','power',4);
inv(T'*T)                            % Estimated correlation of F, no longer eye(2)
```

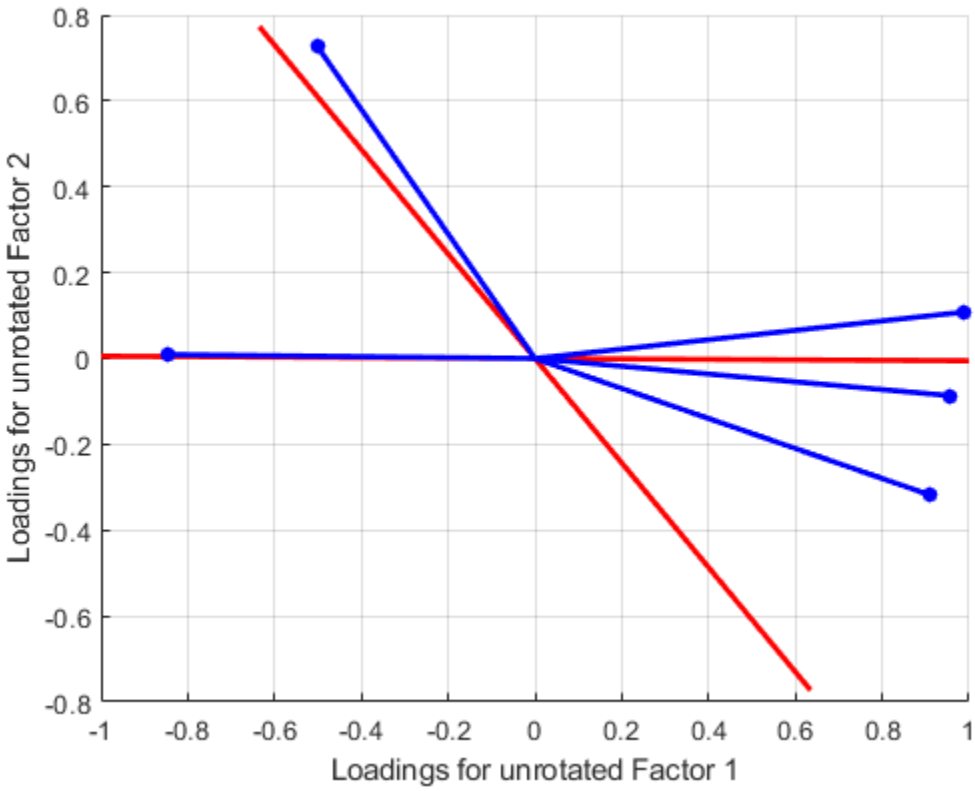
```
ans = 2×2

    1.0000   -0.6391
   -0.6391    1.0000
```

```
Lambda*inv(T'*T)*Lambda'+diag(Psi)   % Estimated correlation of X
```

```
ans = 5×5

    1.0000   -0.5424   -0.6893    0.4309   -0.4167
   -0.5424    1.0000    0.8979   -0.8078    0.9328
   -0.6893    0.8979    1.0000   -0.7730    0.8647
    0.4309   -0.8078   -0.7730    1.0000   -0.8326
   -0.4167    0.9328    0.8647   -0.8326    1.0000
```

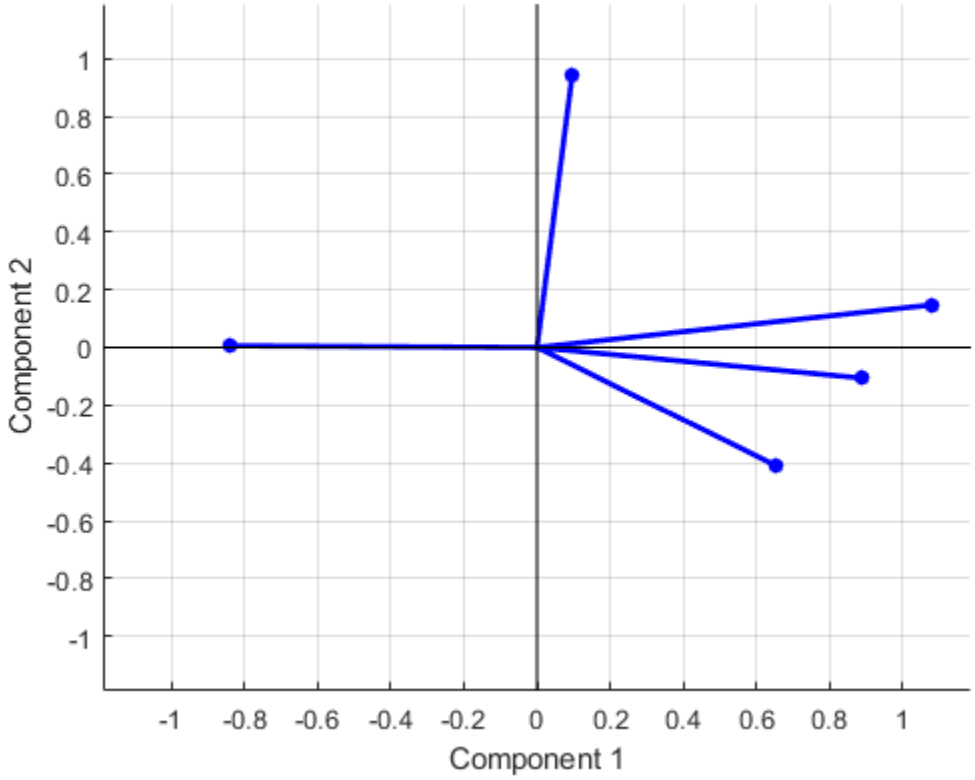
Plot the unrotated variables with oblique axes superimposed.

```
invT = inv(T);
Lambda0 = Lambda*invT;
figure()
line([-invT(1,1) invT(1,1) NaN -invT(2,1) invT(2,1)], ...
     [-invT(1,2) invT(1,2) NaN -invT(2,2) invT(2,2)], ...
     'Color','r','LineWidth',2)
grid on
hold on
biplot(Lambda0,'LineWidth',2,'MarkerSize',20)
xlabel('Loadings for unrotated Factor 1')
ylabel('Loadings for unrotated Factor 2')
```



Plot the rotated variables against the oblique axes.

```
figure()
biplot(Lambda,'LineWidth',2,'MarkerSize',20)
```



Input Arguments

`X` — Data

matrix

Data, specified as an `n`-by-`d` matrix, where each row is an observation of `d` variables.

Data Types: `double`

`m` — Number of common factors

positive integer

Number of common factors, specified as a positive integer.

Example: 3

Data Types: `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `lambda = factoran(X,m,'Start',10,'Scores','Thomson')` specifies to use a starting point for specific variances of 10 and the Thomson method for predicting factor scores.

Xtype — Input data type
'data' (default) | 'covariance'

Input data type of X, specified as the comma-separated pair consisting of 'Xtype' and one of the following:

- 'data' — X is raw data.
- 'covariance' — X is a positive definite covariance or correlation matrix.

Example: 'Xtype','covariance'

Data Types: char | string

Scores — Method for predicting factor scores
'wls' or the equivalent 'Bartlett' (default) | 'regression' or the equivalent 'Thomson'

Method for predicting factor scores, specified as the comma-separated pair consisting of 'Scores' and one of the following:

- 'wls' or the equivalent 'Bartlett' — Weighted least-squares estimate treating F as fixed
- 'regression' or the equivalent 'Thomson' — Minimum mean squared error prediction that is equivalent to a ridge regression

Example: 'Scores','regression'

Data Types: char | string

Start — Starting point for specific variances psi in maximum likelihood optimization
'Rsquared' (default) | 'random' | positive integer | matrix with d rows

Starting point for the specific variances psi in the maximum likelihood optimization, specified as the comma-separated pair consisting of 'Start' and one of the following:

- 'Rsquared' — Chooses the starting vector as a scale factor times diag(inv(corrcoef(X))) (default). For examples, see Jöreskog [2].
- 'random' — Chooses d uniformly distributed values on the interval [0,1].
- Positive integer — Performs the given number of maximum likelihood fits, each initialized in the same way as 'random'. factoran returns the fit with the highest likelihood.
- Matrix with d rows — Performs one maximum likelihood fit for each column of the specified matrix. factoran initializes the ith optimization with the values from the ith column.

Example: 'Start',5

Data Types: double | char | string

Rotate — Method used to rotate factor loadings and scores
'varimax' (default) | 'none' | 'quartimax' | 'equamax' | 'parsimax' | 'orthomax' | 'promax' | 'procrustes' | 'pattern' | function handle

Method used to rotate factor loadings and scores, specified as the comma-separated pair consisting of 'Rotate' and one of the values in the following table. You can control the rotation by specifying additional name-value pair arguments of the rotatefactors function, as described in the table. For details, see rotatefactors.

Value	Description
`'none'`	Performs no rotation
`'equamax'`	Special case of the `'orthomax'` rotation. Use the `'normalize'`, `'reltol'`, and `'maxit'` arguments to control the details of the rotation.
`'orthomax'`	Orthogonal rotation that maximizes a criterion based on the variance of the loadings. Use the `'coeff'`, `'normalize'`, `'reltol'`, and `'maxit'` arguments to control the details of the rotation.
`'parsimax'`	Special case of the orthomax rotation. Use the `'normalize'`, `'reltol'`, and `'maxit'` arguments to control the details of the rotation.
`'pattern'`	Performs either an oblique rotation (the default) or an orthogonal rotation to best match a specified pattern matrix. Use the `'type'` argument to choose the type of rotation. Use the `'target'` argument to specify the pattern matrix.
`'procrustes'`	Performs either an oblique rotation (the default) or an orthogonal rotation to best match a specified target matrix in the least-squares sense. Use the `'type'` argument to choose the type of rotation. Use the `'target'` argument to specify the target matrix.
`'promax'`	Performs an oblique procrustes rotation to a target matrix determined by `factoran` as a function of an orthomax solution. Use the `'power'` argument to specify the exponent for creating the target matrix. Because `'promax'` uses `'orthomax'` internally, you can also specify the arguments that apply to `'orthomax'`.
`'quartimax'`	Special case of the `'orthomax'` rotation. Use the `'normalize'`, `'reltol'`, and `'maxit'` arguments to control the details of the rotation.
`'varimax'`	Special case of the `'orthomax'` rotation (default). Use the `'normalize'`, `'reltol'`, and `'maxit'` arguments to control the details of the rotation.
function handle	Function handle to a rotation function of the form `[B,T] = myrotation(A,...)` where `A` is a `d`-by-`m` matrix of unrotated factor loadings, `B` is a `d`-by-`m` matrix of rotated loadings, and `T` is the corresponding `m`-by-`m` rotation matrix. Use the `factoran` argument `'UserArgs'` to pass additional arguments to this rotation function. See “User-Defined Rotation Function” on page 33-1432.

Example: `[lambda,psi,T] = factoran(X,m,'Rotate','promax','power',5,'maxit',100)`

Data Types: `char` | `string` | `function_handle`

`Delta` — Lower bound for `psi` during maximum likelihood optimization
`0.005` (default) | scalar between 0 and 1

Lower bound for the `psi` argument during maximum likelihood optimization, specified as the comma-separated pair consisting of `'Delta'` and a scalar value between 0 and 1 (0 < `Delta` < 1).

Example: `0.02`

Data Types: `double`

`OptimOpts` — Options for maximum likelihood optimization
`[]` (default) | structure created by `statset`

Options for the maximum likelihood optimization, specified as the comma-separated pair consisting of `'OptimOpts'` and a structure created by `statset`. You can enter `statset('factoran')` for the list of options, which are also described in the following table.

Field Name (`statset` argument)	Meaning	Value {default}
`'Display'`	Amount of information displayed by the algorithm	• `{'off'}` — Displays no information • `'final'` — Displays the final output • `'iter'` — Displays iterative output to the command window for some functions; otherwise displays the final output
`MaxFunEvals`	Maximum number of objective function evaluations allowed	Positive integer, `{400}`
`MaxIter`	Maximum number of iterations allowed	Positive integer, `{100}`
`TolFun`	Termination tolerance for the objective function value. The solver stops when successive function values are less than `TolFun` apart.	Positive scalar, `{1e-8}`
`TolX`	Termination tolerance for the parameters. The solver stops when successive parameter values are less than `TolX` apart.	Positive scalar, `{1e-8}`

Example: `statset('Display','iter')`

Data Types: `struct`

`Nobs` — Number of observations used to estimate `X`
positive integer

Number of observations used to estimate `X`, specified as the comma-separated pair consisting of `'Nobs'` and a positive integer. `Nobs` applies only when `Xtype` is `'covariance'`. Specifying `'Nobs'` enables you to obtain the `stats` output structure fields `chisq` and `p`.

Example: 50

Data Types: `double`

Output Arguments

`lambda` — Factor loadings

matrix

Factor loadings, returned as a `d`-by-`m` matrix. `d` is the number of columns of the data matrix `X`, and `m` is the second input argument of `factoran`.

The `(i,j)`th element of `lambda` is the coefficient, or loading, of the `j`th factor for the `i`th variable. By default, `factoran` calls the function `rotatefactors` to rotate the estimated factor loadings using the `'varimax'` option. For information about rotation, see “Rotation of Factor Loadings and Scores” on page 33-1432.

`psi` — Specific variances

vector

Specific variances, returned as a `d`-by-`1` vector. `d` is the number of columns of the data matrix `X`. The entries of `psi` are maximum likelihood estimates.

`T` — Factor loadings rotation

matrix

Factor loadings rotation, returned as an `m`-by-`m` matrix. `m` is the second input argument of `factoran`. For information about rotation, see “Rotation of Factor Loadings and Scores” on page 33-1432.

`stats` — Information about common factors

structure

Information about the common factors, returned as a structure. `stats` contains information relating to the null hypothesis H_0 that the number of common factors is `m`.

`stats` contains the following fields.

Field	Description
`loglike`	Maximized loglikelihood value
`dfe`	Error degrees of freedom = `((d-m)^2 - (d+m))/2`
`chisq`	Approximate chi-squared statistic for the null hypothesis
`p`	Right-tail significance level for the null hypothesis

`factoran` does not compute the `chisq` and `p` fields unless `dfe` is positive and all the specific variance estimates in `psi` are positive (see “Heywood Case” on page 33-1432). If `X` is a covariance matrix and you want `factoran` to compute the `chisq` and `p` fields, then you must also specify the `'Nobs'` name-value pair argument.

`F` — Factor scores

matrix

Factor scores, also called predictions of the common factors, returned as an `n`-by-`m` matrix. `n` is the number of rows in the data matrix `X`, and `m` is the second input argument of `factoran`.

Note If `X` is a covariance matrix (`Xtype = 'covariance'`), `factoran` cannot compute `F`.

`factoran` rotates `F` using the same criterion as for `lambda`. For information about rotation, see “Rotation of Factor Loadings and Scores” on page 33-1432.

More About

Heywood Case

If elements of `psi` are equal to the value of the `Delta` parameter (that is, they are essentially zero), the fit is known as a Heywood case, and interpretation of the resulting estimates is problematic. In particular, there can be multiple local maxima of the likelihood, each with different estimates of the loadings and the specific variances. Heywood cases can indicate overfitting (`m` is too large), but can also be the result of underfitting.

Rotation of Factor Loadings and Scores

Unless you explicitly specify no rotation using the `'Rotate'` name-value pair argument, `factoran` rotates the estimated factor loadings `lambda` and the factor scores `F`. The output matrix `T` is used to rotate the loadings, that is, `lambda = lambda0*T`, where `lambda0` is the initial (unrotated) MLE of the loadings. `T` is an orthogonal matrix for orthogonal rotations, and the identity matrix for no rotation. The inverse of `T` is known as the primary axis rotation matrix, whereas `T` itself is related to the reference axis rotation matrix. For orthogonal rotations, the two are identical.

`factoran` computes factor scores that have been rotated by `inv(T')`, that is, `F = F0 * inv(T')`, where `F0` contains the unrotated predictions. The estimated covariance of `F` is `inv(T'*T)`, which is the identity matrix for orthogonal or no rotation. Rotation of factor loadings and scores is an attempt to create a structure that is easier to interpret in the loadings matrix after maximum likelihood estimation.

User-Defined Rotation Function

The syntax for passing additional arguments to a user-defined rotation function is:

```
[Lambda,Psi,T] = ...
     factoran(X,2,'Rotate',@myrotation,'UserArgs',1,'two');
```

References

[1] Harman, Harry Horace. *Modern Factor Analysis*. 3rd Ed. Chicago: University of Chicago Press, 1976.

[2] Jöreskog, K. G. “Some Contributions to Maximum Likelihood Factor Analysis.” *Psychometrika* 32, no. 4 (December 1967): 443–82. https://doi.org/10.1007/BF02289658

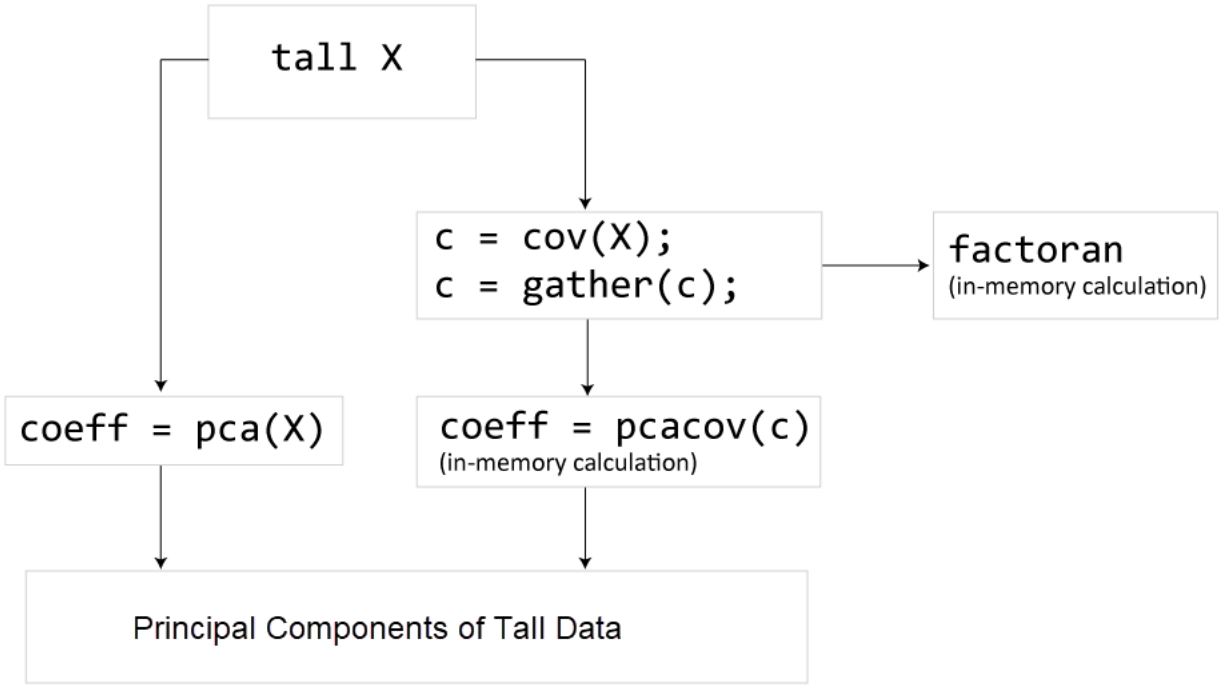
[3] Lawley, D. N., and A. E. Maxwell. *Factor Analysis as a Statistical Method*. 2nd Ed. New York: American Elsevier Publishing Co., 1971.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

`pcacov` and `factoran` do not work directly on tall arrays. Instead, use `C = gather(cov(X))` to compute the covariance matrix of a tall array. Then, you can use `pcacov` or `factoran` to work on the in-memory covariance matrix. Alternatively, you can use `pca` directly on a tall array.



For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`biplot` | `pca` | `pcacov` | `procrustes` | `rotatefactors` | `statset`

Topics

“Perform Factor Analysis on Exam Grades” on page 15-180
“Analyze Stock Prices Using Factor Analysis” on page 15-79

Introduced before R2006a

fcdf

F cumulative distribution function

Syntax

```
p = fcdf(x,v1,v2)
p = fcdf(x,v1,v2,'upper')
```

Description

`p = fcdf(x,v1,v2)` computes the *F* cdf at each of the values in `x` using the corresponding numerator degrees of freedom `v1` and denominator degrees of freedom `v2`. `x`, `v1`, and `v2` can be vectors, matrices, or multidimensional arrays that are all the same size. A scalar input is expanded to a constant matrix with the same dimensions as the other inputs. `v1` and `v2` parameters must contain real positive values, and the values in `x` must lie on the interval `[0 Inf]`.

`p = fcdf(x,v1,v2,'upper')` returns the complement of the *F* cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The *F* cdf is

$$p = F(x|\nu_1,\nu_2) = \int_0^x \frac{\Gamma\left[\frac{(\nu_1+\nu_2)}{2}\right]}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{t^{\frac{\nu_1-2}{2}}}{\left[1+\left(\frac{\nu_1}{\nu_2}\right)t\right]^{\frac{\nu_1+\nu_2}{2}}}dt$$

The result, *p*, is the probability that a single observation from an *F* distribution with parameters ν_1 and ν_2 will fall in the interval [0 *x*].

Examples

Compute F Distribution CDF

The following illustrates a useful mathematical identity for the *F* distribution.

```
nu1 = 1:5;
nu2 = 6:10;
x = 2:6;

F1 = fcdf(x,nu1,nu2)
```

```
F1 = 1×5

    0.7930    0.8854    0.9481    0.9788    0.9919
```

```
F2 = 1 - fcdf(1./x,nu2,nu1)
```

```
F2 = 1×5
```

```
    0.7930    0.8854    0.9481    0.9788    0.9919
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `finv` | `fpdf` | `frnd` | `fstat`

Topics

“F Distribution” on page B-45

Introduced before R2006a

FeatureTransformer

Generated feature transformations

Description

A `FeatureTransformer` object contains information about the feature transformations generated from a training data set. To better understand the generated features, you can use the `describe` object function. To apply the same training set feature transformations to a test set, you can use the `transform` object function.

Creation

Create a `FeatureTransformer` object by using the `gencfeatures` function.

Properties

Type — Type of model
`'classification'`

This property is read-only.

Type of model, specified as `'classification'`.

`TargetLearner` — Expected learner type
`'linear'` | `'bag'`

This property is read-only.

Expected learner type, specified as `'linear'` or `'bag'`. The software creates and selects new features assuming that they will be used to train a linear classifier or a bagged ensemble classifier, respectively.

`NumEngineeredFeatures` — Number of engineered features
nonnegative scalar

This property is read-only.

Number of engineered features generated by `gencfeatures` and stored in `FeatureTransformer`, returned as a nonnegative scalar.

Data Types: `double`

`NumOriginalFeatures` — Number of original features
nonnegative scalar

This property is read-only.

Number of original features kept by `gencfeatures` and stored in `FeatureTransformer`, returned as a nonnegative scalar.

Data Types: `double`

`TotalNumFeatures` — Total number of features
nonnegative scalar

This property is read-only.

Total number of features stored in `FeatureTransformer`, returned as a nonnegative scalar. `TotalNumFeatures` equals the sum of `NumEngineeredFeatures` and `NumOriginalFeatures`.

Data Types: `double`

Object Functions

describe	Describe generated features
transform	Transform new data using generated features

Examples

Generate and Inspect Features

Generate features from a table of predictor data by using `gencfeatures`. Inspect the generated features by using the `describe` object function.

Read power outage data into the workspace as a table. Remove observations with missing values, and display the first few rows of the table.

```
outages = readtable("outages.csv");
Tbl = rmmissing(outages);
head(Tbl)
```

```
ans=8×6 table
       Region           OutageTime        Loss     Customers     RestorationTime            Cause
    ____________    ________________    ______    __________    ________________    _____________

    {'SouthWest'}    2002-02-01 12:18    458.98    1.8202e+06    2002-02-07 16:50    {'winter sto
    {'SouthEast'}    2003-02-07 21:15     289.4    1.4294e+05    2003-02-17 08:14    {'winter sto
    {'West'     }    2004-04-06 05:44    434.81    3.4037e+05    2004-04-06 06:10    {'equipment
    {'MidWest'  }    2002-03-16 06:18    186.44    2.1275e+05    2002-03-18 23:23    {'severe sto
    {'West'     }    2003-06-18 02:49         0             0    2003-06-18 10:54    {'attack'
    {'NorthEast'}    2003-07-16 16:23    239.93         49434    2003-07-17 01:12    {'fire'
    {'MidWest'  }    2004-09-27 11:09    286.72         66104    2004-09-27 16:37    {'equipment
    {'SouthEast'}    2004-09-05 17:48    73.387         36073    2004-09-05 20:46    {'equipment
```

Some of the variables, such as `OutageTime` and `RestorationTime`, have data types that are not supported by classifier training functions like `fitcensemble`.

Generate 25 features from the predictors in `Tbl` that can be used to train a bagged ensemble. Specify the `Region` table variable as the response.

```
Transformer = gencfeatures(Tbl,"Region",25,"TargetLearner","bag")
```

```
Transformer =
  FeatureTransformer with properties:
```

```
                    Type: 'classification'
           TargetLearner: 'bag'
    NumEngineeredFeatures: 22
      NumOriginalFeatures: 3
         TotalNumFeatures: 25
```

The `Transformer` object contains the information about the generated features and the transformations used to create them.

To better understand the generated features, use the `describe` object function.

```
Info = describe(Transformer)
```

```
Info=25×4 table
                                 Type        IsOriginal          InputVariables                  
                              ___________    __________    ____________________________    ___

    Loss                      Numeric          true        Loss                            ""
    Customers                 Numeric          true        Customers                       ""
    c(Cause)                  Categorical      true        Cause                           "Va
    RestorationTime-OutageTime Numeric         false       OutageTime, RestorationTime     "E
    sdn(OutageTime)           Numeric          false       OutageTime                      "Se
    woe3(c(Cause))            Numeric          false       Cause                           "Va
    doy(OutageTime)           Numeric          false       OutageTime                      "Da
    year(OutageTime)          Numeric          false       OutageTime                      "Ye
    kmd1                      Numeric          false       Loss, Customers                 "Eu
    kmd5                      Numeric          false       Loss, Customers                 "Eu
    quarter(OutageTime)       Numeric          false       OutageTime                      "Qu
    woe2(c(Cause))            Numeric          false       Cause                           "Va
    year(RestorationTime)     Numeric          false       RestorationTime                 "Ye
    month(OutageTime)         Numeric          false       OutageTime                      "Mo
    Loss.*Customers           Numeric          false       Loss, Customers                 "Lo
    tods(OutageTime)          Numeric          false       OutageTime                      "Ti
      ⋮
```

The `Info` table indicates the following:

- The first three generated features are original to `Tbl`, although the software converts the original `Cause` variable to a categorical variable `c(Cause)`.
- The `OutageTime` and `RestorationTime` variables are not included as generated features because they are `datetime` variables, which cannot be used to train a bagged ensemble model. However, the software derives many of the generated features from these variables, such as the fourth feature `RestorationTime-OutageTime`.
- Some generated features are a combination of multiple transformations. For example, the software generates the sixth feature `woe3(c(Cause))` by converting the `Cause` variable to a categorical variable and then calculating the Weight of Evidence values for the resulting variable.

Train Model Using Subset of Generated Features

Train a linear classifier using only the numeric generated features returned by `gencfeatures`.

Load the `patients` data set. Create a table from a subset of the variables.

```
load patients
Tbl = table(Age,Diastolic,Height,SelfAssessedHealthStatus, ...
    Smoker,Systolic,Weight,Gender);
```

Partition the data into training and test sets. Use approximately 70% of the observations as training data, and 30% of the observations as test data. Partition the data using `cvpartition`.

```
rng("default")
c = cvpartition(Tbl.Gender,"Holdout",0.30);
TrainTbl = Tbl(training(c),:);
TestTbl = Tbl(test(c),:);
```

Use the training data to generate 25 new features. Specify the minimum redundancy maximum relevance (MRMR) feature selection method for selecting new features.

```
Transformer = gencfeatures(TrainTbl,"Gender",25, ...
    "FeatureSelectionMethod","mrmr")
```

```
Transformer =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'linear'
    NumEngineeredFeatures: 24
      NumOriginalFeatures: 1
         TotalNumFeatures: 25
```

Inspect the generated features.

```
Info = describe(Transformer)
```

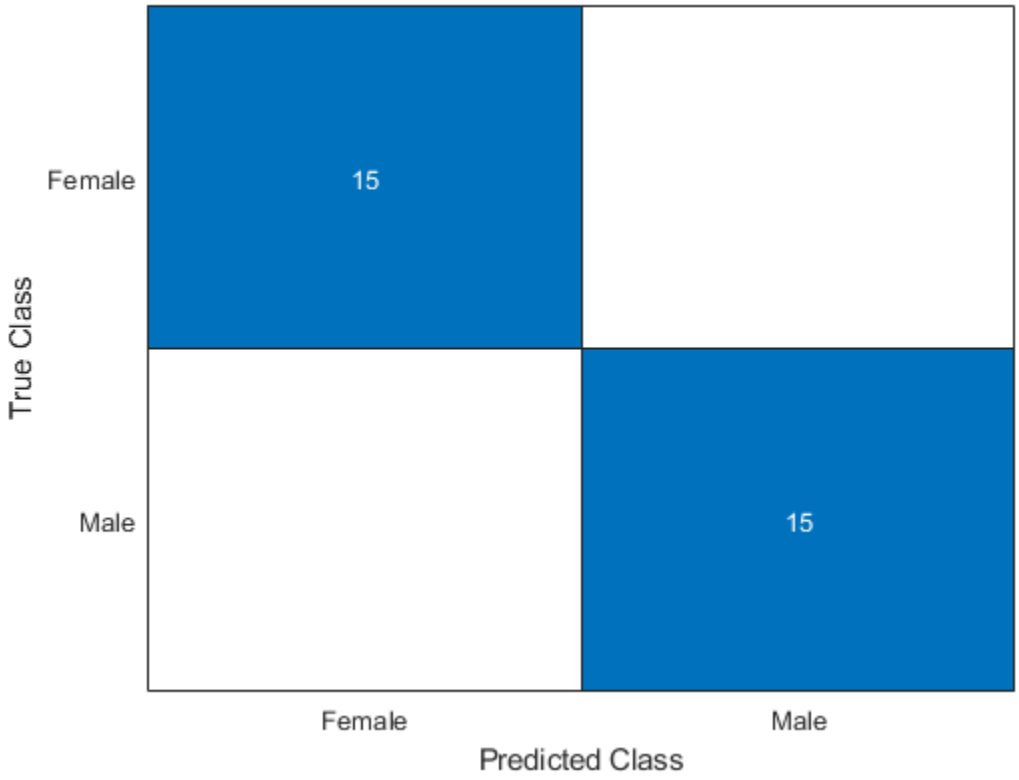
```
Info=25×4 table
                                  Type        IsOriginal        InputVariables
                                ___________    __________    ____________________________

    c(SelfAssessedHealthStatus)    Categorical      true        SelfAssessedHealthStatus    "Vari
    eb5(Weight)                    Categorical      false       Weight                      "Equa
    zsc(sqrt(Systolic))            Numeric          false       Systolic                    "sqrt
    zsc(sin(Systolic))             Numeric          false       Systolic                    "sin(
    zsc(Systolic./Weight)          Numeric          false       Systolic, Weight            "Syst
    zsc(Age+Weight)                Numeric          false       Age, Weight                 "Age+
    zsc(Age./Weight)               Numeric          false       Age, Weight                 "Age.
    zsc(Diastolic.*Weight)         Numeric          false       Diastolic, Weight           "Dias
    q6(Height)                     Categorical      false       Height                      "Equi
    zsc(Systolic+Weight)           Numeric          false       Systolic, Weight            "Syst
    zsc(Diastolic-Weight)          Numeric          false       Diastolic, Weight           "Dias
    zsc(Age-Weight)                Numeric          false       Age, Weight                 "Age-
    zsc(Height./Weight)            Numeric          false       Height, Weight              "Heig
    zsc(Height.*Weight)            Numeric          false       Height, Weight              "Heig
    zsc(Diastolic+Weight)          Numeric          false       Diastolic, Weight           "Dias
    zsc(Age.*Weight)               Numeric          false       Age, Weight                 "Age.
      ⋮
```

Transform the training and test sets, but retain only the numeric predictors.

```
numericIdx = (Info.Type == "Numeric");
NewTrainTbl = transform(Transformer,TrainTbl,numericIdx);
NewTestTbl = transform(Transformer,TestTbl,numericIdx);
```

Train a linear model using the transformed training data. Visualize the accuracy of the model's test set predictions by using a confusion matrix.

```
Mdl = fitclinear(NewTrainTbl,TrainTbl.Gender);
testLabels = predict(Mdl,NewTestTbl);
confusionchart(TestTbl.Gender,testLabels)
```



See Also

`describe` | `fitcensemble` | `fitclinear` | `gencfeatures` | `transform`

Topics

“Automated Feature Engineering for Classification” on page 18-190

Introduced in R2021a

feval

Package:

Predict responses of generalized linear regression model using one input for each predictor

Syntax

```
ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)
```

Description

`ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)` returns the predicted response of `mdl` to the new input predictors `Xnew1,Xnew2,...,Xnewn`.

Examples

Predict Response Values

Create a generalized linear regression model, and plot its responses to a range of input data.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

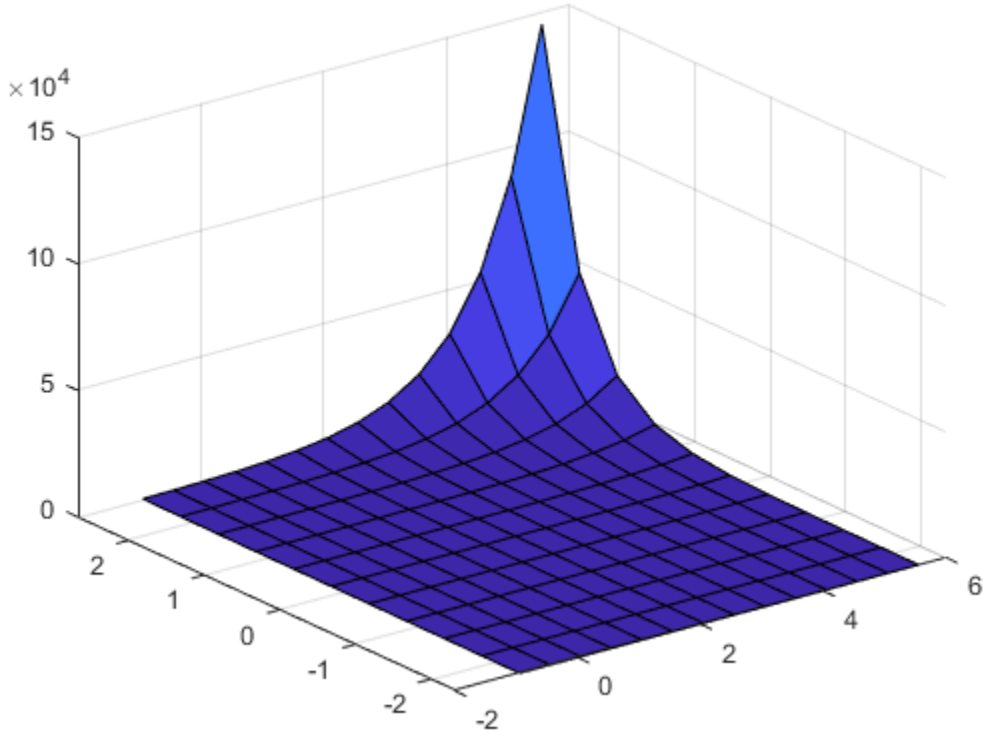
```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

Generate a range of values for `X(:,1)` and `X(:,2)`, and plot the predictions at the values.

```
[Xtest1,Xtest2] = meshgrid(min(X(:,1)):.5:max(X(:,1)),min(X(:,2)):.5:max(X(:,2)));
Z = feval(mdl,Xtest1,Xtest2);
surf(Xtest1,Xtest2,Z)
```



Input Arguments

`mdl` — Generalized linear regression model
`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

`Xnew1,Xnew2,...,Xnewn` — New predictor input values
vector | matrix | table | dataset array

New predictor values, specified as a vector, matrix, table, or dataset array.

- If you pass multiple inputs `Xnew1,Xnew2,...,Xnewn` and each includes observations for one predictor variable, then each input must be a vector. Each vector must have the same size. If you specify a predictor variable as a scalar, then `feval` expands the scalar argument into a constant vector of the same size as the other arguments.
- If you pass a single input `Xnew1`, then `Xnew1` must be a table, dataset array, or matrix.
  - If `Xnew1` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
  - If `Xnew1` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew1` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating

`mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Output Arguments

`ypred` — Predicted response values

numeric vector

Predicted response values at `Xnew1,Xnew2,...,Xnewn`, returned as a numeric vector.

For a binomial model, `feval` uses 1 as the `BinomialSize` parameter, so the values in `ypred` are predicted probabilities. To return the numbers of successes in the trials, use the `predict` function and specify the number of trials by using the `'BinomialSize'` name-value pair argument.

For a model with an offset, `feval` uses 0 as the offset value. To specify the offset value used when you fit a model, use the `predict` function and the `'Offset'` name-value pair argument.

Tips

- A regression object is, mathematically, a function that estimates the relationship between the response and predictors. The `feval` function enables an object to behave like a function in MATLAB. You can pass `feval` to another function that accepts a function input, such as `fminsearch` and `integral`.
- `feval` can be simpler to use with a model created from a table or dataset array. When you have new predictor data, you can pass it to `feval` without creating a table or matrix.

Alternative Functionality

- `predict` gives the same predictions as `feval` if you use the default values for the `'Offset'` and `'BinomialSize'` name-value pair arguments of `predict`. The prediction values can be different if you specify other values for these arguments. The `predict` function also returns confidence intervals on its predictions. Note that the `predict` function accepts a single input argument containing all predictor variables, rather than multiple input arguments with one input for each predictor variable.
- `random` predicts responses with added noise.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `predict` | `random`

Topics
“feval” on page 12-24
“Generalized Linear Models” on page 12-9

Introduced in R2012a

feval

Package:

Predict responses of linear regression model using one input for each predictor

Syntax

```
ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)
```

Description

`ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)` returns the predicted response of `mdl` to the new input predictors `Xnew1,Xnew2,...,Xnewn`.

Examples

Plot Different Categorical Levels

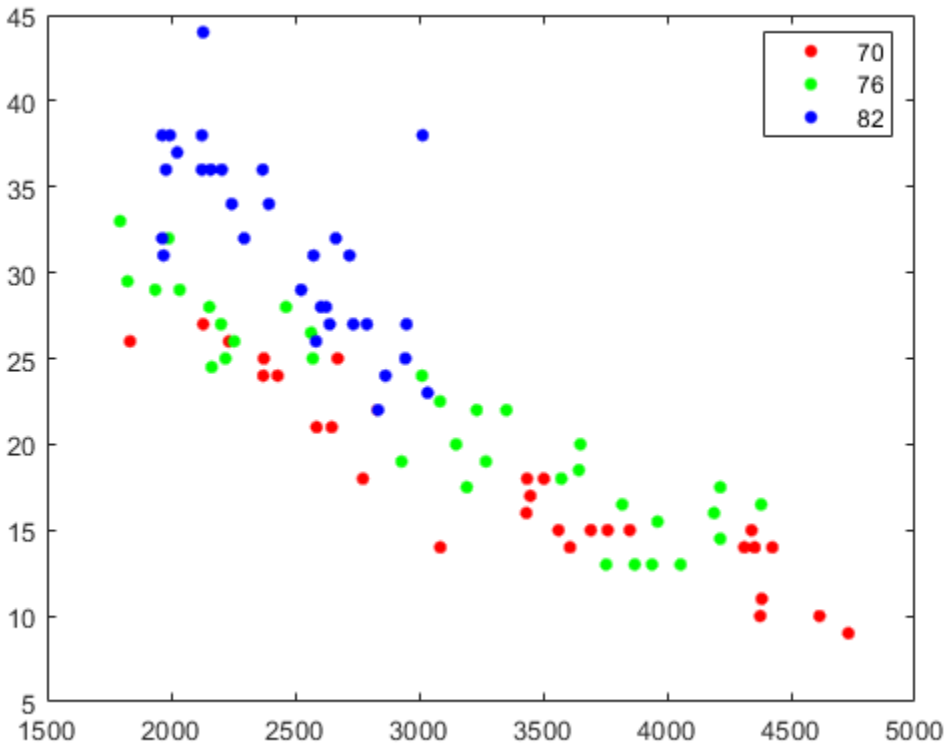
Fit a mileage model to the `carsmall` data set, including the `Year` categorical predictor. Superimpose fitted curves on a scatter plot of the data.

Load the data set and fit the model.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

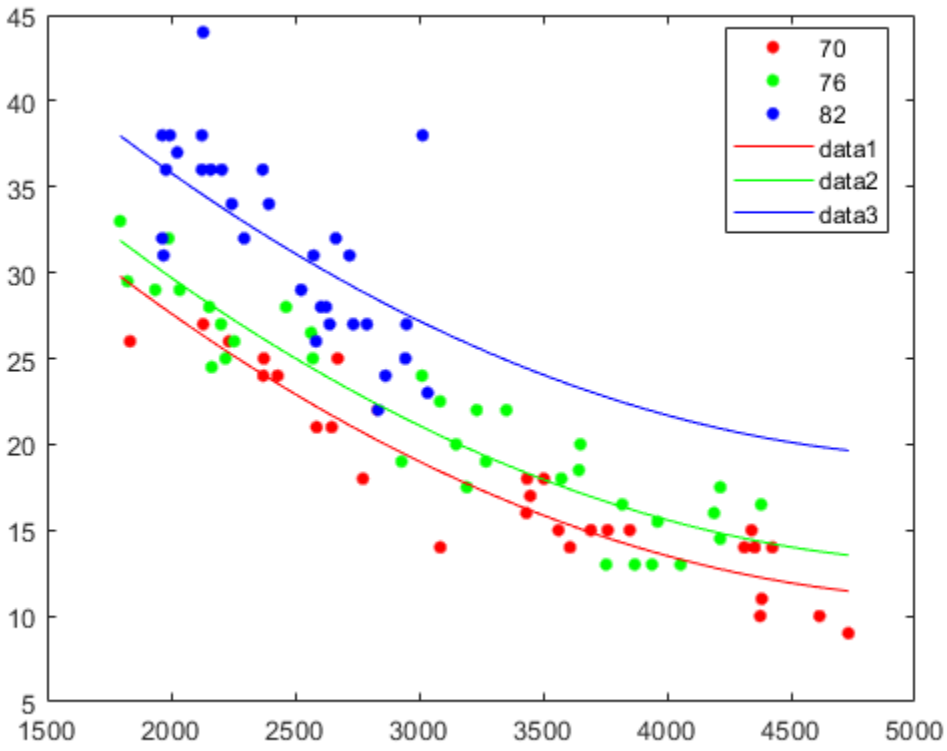
Create a scatter plot of `MPG` versus `Weight`, grouped by `Year`.

```
gscatter(tbl.Weight,tbl.MPG,tbl.Year);
```



Plot curves of the model predictions for the various years and weights by using `feval`.

```
w = linspace(min(tbl.Weight),max(tbl.Weight))';
line(w,feval(mdl,w,'70'),'Color','r')
line(w,feval(mdl,w,'76'),'Color','g')
line(w,feval(mdl,w,'82'),'Color','b')
```



Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`Xnew1,Xnew2,...,Xnewn` — New predictor input values

vector | matrix | table | dataset array

New predictor values, specified as a vector, matrix, table, or dataset array.

- If you pass multiple inputs `Xnew1,Xnew2,...,Xnewn` and each includes observations for one predictor variable, then each input must be a vector. Each vector must have the same size. If you specify a predictor variable as a scalar, then `feval` expands the scalar argument into a constant vector of the same size as the other arguments.
- If you pass a single input `Xnew1`, then `Xnew1` must be a table, dataset array, or matrix.
  - If `Xnew1` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
  - If `Xnew1` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew1` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating

`mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Output Arguments

`ypred` — Predicted response values

numeric vector

Predicted response values at `Xnew1,Xnew2,...,Xnewn`, returned as a numeric vector.

Tips

- A regression object is, mathematically, a function that estimates the relationship between the response and predictors. The `feval` function enables an object to behave like a function in MATLAB. You can pass `feval` to another function that accepts a function input, such as `fminsearch` and `integral`.
- `feval` can be simpler to use with a model created from a table or dataset array. When you have new predictor data, you can pass it to `feval` without creating a table or matrix.

Alternative Functionality

- `predict` gives the same predictions as `feval` by using a single input argument containing all predictor variables, rather than multiple input arguments with one input for each predictor variable. `predict` also gives confidence intervals on its predictions.
- `random` predicts responses with added noise.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactLinearModel` | `LinearModel` | `predict` | `random`

Topics

“Predict or Simulate Responses to New Data” on page 11-31
“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2012a

feval

Class: `NonLinearModel`

Evaluate nonlinear regression model prediction

Syntax

```
ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)
```

Description

`ypred = feval(mdl,Xnew1,Xnew2,...,Xnewn)` returns the predicted response of `mdl` to the input `[Xnew1,Xnew2,...,Xnewn]`.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`Xnew1,Xnew2,...,Xnewn`

Predictor components. `Xnewi` can be one of:

- Scalar
- Vector
- Array

Each nonscalar component must have the same size (number of elements in each dimension).

If you pass just one `Xnew` array, `Xnew` can be a table, dataset array, or an array of doubles, where each column of the array represents one predictor.

Output Arguments

`ypred`

Predicted mean values at `Xnew`. `ypred` is the same size as each component of `Xnew`.

Examples

Predict a Nonlinear Model from a Table

Create a nonlinear model for auto mileage based on the `carbig` data. Predict the mileage of an average car.

Load the data and create a nonlinear model.

```
load carbig
tbl = table(Horsepower,Weight,MPG);
modelfun = @(b,x)b(1) + b(2)*x(:,1).^b(3) + ...
    b(4)*x(:,2).^b(5);
beta0 = [-50 500 -1 500 -1];
mdl = fitnlm(tbl,modelfun,beta0);
```

Find the predicted mileage of an average car. The data contains some missing (`NaN`) observations, so compute the mean using `mean` with the `'omitnan'` option.

```
Xnew = mean([Horsepower Weight],'omitnan');
MPGnew = feval(mdl,Xnew)
```

```
MPGnew = 21.8073
```

Alternatives

`predict` gives the same predictions, but uses a single input array with one observation in each row, rather than one component in each input argument. `predict` also gives confidence intervals on its predictions.

`random` predicts with added noise.

See Also

`NonLinearModel` | `predict` | `random`

Topics

“Predict or Simulate Responses Using a Nonlinear Model” on page 13-9
“Nonlinear Regression” on page 13-2

ff2n

Two-level full factorial design

Syntax

```
dFF2 = ff2n(n)
```

Description

`dFF2 = ff2n(n)` gives factor settings `dFF2` for a two-level full factorial design with `n` factors. `dFF2` is *m*-by-`n`, where *m* is the number of treatments in the full-factorial design. Each row of `dFF2` corresponds to a single treatment. Each column contains the settings for a single factor, with values of `0` and `1` for the two levels.

Examples

```
dFF2 = ff2n(3)
dFF2 =
   0   0   0
   0   0   1
   0   1   0
   0   1   1
   1   0   0
   1   0   1
   1   1   0
   1   1   1
```

See Also

`fullfact`

Introduced before R2006a

fillprox

Class: `TreeBagger`

Proximity matrix for training data

Syntax

```
B = fillprox(B)
B = fillprox(B,'param1',val1,'param2',val2,...)
```

Description

`B = fillprox(B)` computes a proximity matrix for the training data and stores it in the Properties field of B.

`B = fillprox(B,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Trees'`	Either `'all'` or a vector of indices of the trees in the ensemble to be used in computing the proximity matrix. Default is `'all'`.
`'NumPrint'`	Number of training cycles (grown trees) after which `TreeBagger` displays a diagnostic message showing training progress. Default is no diagnostic messages.

See Also

`outlierMeasure` | `proximity`

findobj

Class: `qrandstream`

Find objects matching specified conditions

Syntax

```
hm = findobj(h, 'conditions')
```

Description

The `findobj` method of the `handle` class follows the same syntax as the MATLAB `findobj` command, except that the first argument must be an array of handles to objects.

`hm = findobj(h, 'conditions')` searches the handle object array `h` and returns an array of handle objects matching the specified conditions. Only the public members of the objects of `h` are considered when evaluating the conditions.

See Also

`findobj` | `qrandstream`

findprop

Class: `qrandstream`

Find property of MATLAB handle object

Syntax

```
p = findprop(h,'propname')
```

Description

`p = findprop(h,'propname')` finds and returns the `meta.property` object associated with property name `propname` of scalar handle object `h`. `propname` must be a character vector. It can be the name of a property defined by the class of `h` or a dynamic property added to scalar object `h`.

If no property named `propname` exists for object `h`, an empty `meta.property` array is returned.

See Also

`dynamicprops` | `findobj` | `meta.property` | `qrandstream`

finv

F inverse cumulative distribution function

Syntax

```
X = finv(P,V1,V2)
```

Description

`X = finv(P,V1,V2)` computes the inverse of the *F* cdf with numerator degrees of freedom `V1` and denominator degrees of freedom `V2` for the corresponding probabilities in `P`. `P`, `V1`, and `V2` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. `V1` and `V2` parameters must contain real positive values, and the values in `P` must lie on the interval [0 1].

The *F* inverse function is defined in terms of the *F* cdf as

$$x = F^{-1}(p|\nu_1,\nu_2) = \{x : F(x|\nu_1,\nu_2) = p\}$$

where

$$p = F(x\left|\nu_1,\nu_2\right.) = \int_0^x \frac{\Gamma\left[\frac{(\nu_1+\nu_2)}{2}\right]}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{t^{\frac{\nu_1-2}{2}}}{\left[1+\left(\frac{\nu_1}{\nu_2}\right)t\right]^{\frac{\nu_1+\nu_2}{2}}} dt$$

Examples

Find a value that should exceed 95% of the samples from an *F* distribution with 5 degrees of freedom in the numerator and 10 degrees of freedom in the denominator.

```
x = finv(0.95,5,10)
x =
  3.3258
```

You would observe values greater than 3.3258 only 5% of the time by chance.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`fcdf` | `fpdf` | `frnd` | `fstat` | `icdf`

Topics

“F Distribution” on page B-45

Introduced before R2006a

fishertest

Fisher’s exact test

Syntax

```
h = fishertest(x)
[h,p,stats] = fishertest(x)
[___] = fishertest(x,Name,Value)
```

Description

`h = fishertest(x)` returns a test decision for Fisher’s exact test of the null hypothesis that there are no nonrandom associations between the two categorical variables in `x`, against the alternative that there is a nonrandom association. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`[h,p,stats] = fishertest(x)` also returns the significance level `p` of the test and a structure `stats` containing additional test results, including the odds ratio and its asymptotic confidence interval.

`[___] = fishertest(x,Name,Value)` returns a test decision using additional options specified by one or more name-value pair arguments. For example, you can change the significance level of the test or conduct a one-sided test.

Examples

Conduct Fisher's Exact Test

In a small survey, a researcher asked 17 individuals if they received a flu shot this year, and whether they caught the flu this winter. The results indicate that, of the nine people who did not receive a flu shot, three got the flu and six did not. Of the eight people who received a flu shot, one got the flu and seven did not.

Create a 2-by-2 contingency table containing the survey data. Row 1 contains data for the individuals who did not receive a flu shot, and row 2 contains data for the individuals who received a flu shot. Column 1 contains the number of individuals who got the flu, and column 2 contains the number of individuals who did not.

```
x = table([3;1],[6;7],'VariableNames',{'Flu','NoFlu'},'RowNames',{'NoShot','Shot'})
```

```
x=2×2 table
              Flu    NoFlu
              ___    _____

    NoShot     3       6
    Shot       1       7
```

Use Fisher's exact test to determine if there is a nonrandom association between receiving a flu shot and getting the flu.

```
h = fishertest(x)
```

```
h = logical
   0
```

The returned test decision `h = 0` indicates that `fishertest` does not reject the null hypothesis of no nonrandom association between the categorical variables at the default 5% significance level. Therefore, based on the test results, individuals who do not get a flu shot do not have different odds of getting the flu than those who got the flu shot.

Conduct a One-Sided Fisher's Exact Test

In a small survey, a researcher asked 17 individuals if they received a flu shot this year, and whether they caught the flu. The results indicate that, of the nine people who did not receive a flu shot, three got the flu and six did not. Of the eight people who received a flu shot, one got the flu and seven did not.

```
x = [3,6;1,7];
```

Use a right-tailed Fisher's exact test to determine if the odds of getting the flu is higher for individuals who did not receive a flu shot than for individuals who did. Conduct the test at the 1% significance level.

```
[h,p,stats] = fishertest(x,'Tail','right','Alpha',0.01)
```

```
h = logical
   0


p = 0.3353

stats = struct with fields:
             OddsRatio: 3.5000
    ConfidenceInterval: [0.1289 95.0408]
```

The returned test decision `h = 0` indicates that `fishertest` does not reject the null hypothesis of no nonrandom association between the categorical variables at the 1% significance level. Since this is a right-tailed hypothesis test, the conclusion is that individuals who do not get a flu shot do not have greater odds of getting the flu than those who got the flu shot.

Generate a Contingency Table Using `crosstab`

Load the hospital data.

```
load hospital
```

The `hospital` dataset array contains data on 100 hospital patients, including last name, gender, age, weight, smoking status, and systolic and diastolic blood pressure measurements.

To determine if smoking status is independent of gender, use `crosstab` to create a 2-by-2 contingency table of smokers and nonsmokers, grouped by gender.

```
[tbl,chi2,p,labels] = crosstab(hospital.Sex,hospital.Smoker)
```

```
tbl = 2×2

    40    13
    26    21


chi2 = 4.5083

p = 0.0337

labels = 2x2 cell
    {'Female'}    {'0'}
    {'Male'  }    {'1'}
```

The rows of the resulting contingency table `tbl` correspond to the patient's gender, with row 1 containing data for females and row 2 containing data for males. The columns correspond to the patient's smoking status, with column 1 containing data for nonsmokers and column 2 containing data for smokers. The returned result `chi2 = 4.5083` is the value of the chi-squared test statistic for a chi-squared test of independence. The returned value `p = 0.0337` is an approximate *p*-value based on the chi-squared distribution.

Use the contingency table generated by `crosstab` to perform Fisher's exact test on the data.

```
[h,p,stats] = fishertest(tbl)
```

```
h = logical
   1


p = 0.0375

stats = struct with fields:
             OddsRatio: 2.4852
    ConfidenceInterval: [1.0624 5.8135]
```

The result `h = 1` indicates that `fishertest` rejects the null hypothesis of nonassociation between smoking status and gender at the 5% significance level. In other words, there is an association between gender and smoking status. The odds ratio indicates that the male patients have about 2.5 times greater odds of being smokers than the female patients.

The returned *p*-value of the test, `p = 0.0375`, is close to, but not exactly the same as, the result obtained by `crosstab`. This is because `fishertest` computes an exact *p*-value using the sample data, while `crosstab` uses a chi-squared approximation to compute the *p*-value.

Input Arguments

x — Contingency table

2-by-2 matrix of nonnegative integer values | 2-by-2 table of nonnegative integer values

Contingency table, specified as a 2-by-2 matrix or table containing nonnegative integer values. A contingency table contains the frequency distribution of the variables in the sample data. You can use `crosstab` to generate a contingency table from sample data.

Example: `[4,0;0,4]`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'Tail','right'` specifies a right-tailed hypothesis test at the 1% significance level.

`Alpha` — Significance level

0.05 (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Two-tailed test. The alternative hypothesis is that there is a nonrandom association between the two variables in `x`, and the odds ratio is not equal to 1.
`'right'`	Right-tailed test. The alternative hypothesis is that the odds ratio is greater than 1.
`'left'`	Left-tailed test. The alternative hypothesis is that the odds ratio is less than 1.

Example: `'Tail','right'`

Output Arguments

`h` — Hypothesis test result

`1` | `0`

Hypothesis test result, returned as a logical value.

- If `h` is `1`, then `fishertest` rejects the null hypothesis at the `Alpha` significance level.
- If `h` is `0`, then `fishertest` fails to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value

scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`stats` — Test data
structure

Test data, returned as a structure with the following fields:

- `OddsRatio` — A measure of association between the two variables.
- `ConfidenceInterval` — Asymptotic confidence interval for the odds ratio. If any of the cell frequencies in `x` are 0, then `fishertest` does not compute a confidence interval and instead displays `[-Inf Inf]`.

More About

Fisher's Exact Test

Fisher's exact test is a nonparametric statistical test used to test the null hypothesis that no nonrandom associations exist between two categorical variables, against the alternative that there is a nonrandom association between the variables.

Fisher's exact test provides an alternative to the chi-squared test for small samples, or samples with very uneven marginal distributions. Unlike the chi-squared test, Fisher's exact test does not depend on large-sample distribution assumptions, and instead calculates an exact *p*-value based on the sample data. Although Fisher's exact test is valid for samples of any size, it is not recommended for large samples because it is computationally intensive. If all of the frequency counts in the contingency table are greater than or equal to `1e7`, then `fishertest` errors. For contingency tables that contain large count values or are well-balanced, use `crosstab` or `chi2gof` instead.

`fishertest` accepts a 2-by-2 contingency table as input, and computes the *p*-value of the test as follows:

1. Calculate the sums for each row, column, and total number of observations in the contingency table.
2. Using a multivariate generalization of the hypergeometric probability function, calculate the conditional probability of observing the exact result in the contingency table if the null hypothesis were true, given its row and column sums. The conditional probability is

   $$P_{cutoff} = \frac{(R_1!R_2!)(C_1!C_2!)}{N!\prod_{i,j} n_{ij}!},$$

   where R_1 and R_2 are the row sums, C_1 and C_2 are the column sums, N is the total number of observations in the contingency table, and n_{ij} is the value in the ith row and jth column of the table.
3. Find all possible matrices of nonnegative integers consistent with the row and column sums. For each matrix, calculate the associated conditional probability using the equation for P_{cutoff}.
4. Use these values to calculate the *p*-value of the test, based on the alternative hypothesis of interest.

   - For a two-sided test, sum all of the conditional probabilities less than or equal to P_{cutoff} for the observed contingency table. This represents the probability of observing a result as extreme

as, or more extreme than, the actual outcome if the null hypothesis were true. Small *p*-values cast doubt on the validity of the null hypothesis, in favor of the alternative hypothesis of association between the variables.

- For a left-sided test, sum the conditional probabilities of all the matrices with a (1,1) cell frequency less than or equal to n_{11}.
- For a right-sided test, sum the conditional probabilities of all the matrices with a (1,1) cell frequency greater than or equal to n_{11} in the observed contingency table.

The odds ratio is

$$OR = \frac{n_{11}n_{22}}{n_{21}n_{12}} .$$

The null hypothesis of conditional independence is equivalent to the hypothesis that the odds ratio equals 1. The left-sided alternative is equivalent to an odds ratio less than 1, and the right-sided alternative is equivalent to an odds ratio greater than 1.

The asymptotic 100(1 – α)% confidence interval for the odds ratio is

$$CI = \left[\exp\left(L - \Phi^{-1}\left(\frac{1-\alpha}{2}\right)SE\right),\ \exp\left(L + \Phi^{-1}\left(\frac{1-\alpha}{2}\right)SE\right)\right],$$

where L is the log odds ratio, $\Phi^{-1}(\bullet)$ is the inverse of the normal inverse cumulative distribution function, and SE is the standard error for the log odds ratio. If the 100(1 – α)% confidence interval does not contain the value 1, then the association is significant at the α significance level. If any of the four cell frequencies are 0, then `fishertest` does not compute the confidence interval and instead displays `[-Inf Inf]`.

`fishertest` only accepts 2-by-2 contingency tables as input. To test the independence of categorical variables with more than two levels, use the chi-squared test provided by `crosstab`.

See Also

`chi2gof` | `crosstab`

Introduced in R2014b

fit

Fit simple model of local interpretable model-agnostic explanations (LIME)

Syntax

```
newresults = fit(results,queryPoint,numImportantPredictors)
newresults = fit(results,queryPoint,numImportantPredictors,Name,Value)
```

Description

`newresults = fit(results,queryPoint,numImportantPredictors)` fits a new simple model for the specified query point (`queryPoint`) by using the specified number or predictors (`numImportantPredictors`). The function returns a `lime` object `newresults` that contains the new simple model.

`fit` uses the simple model options that you specify when you create the `lime` object `results`. You can change the options using the name-value pair arguments of the `fit` function.

`newresults = fit(results,queryPoint,numImportantPredictors,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify `'SimpleModelType','tree'` to fit a decision tree model.

Examples

Explain Prediction with Linear Simple Model

Train a regression model and create a `lime` object that uses a linear simple model. When you create a `lime` object, if you do not specify a query point and the number of important predictors, then the software generates samples of a synthetic data set but does not fit a simple model. Use the object function `fit` to fit a simple model for a query point. Then display the coefficients of the fitted linear simple model by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Create a table of predictor variables by removing the response variable from `tbl`.

```
tblX = removevars(tbl,'MPG');
```

Train a blackbox model of `MPG` by using the `fitrkernel` function.

```
rng('default') % For reproducibility
mdl = fitrkernel(tblX,tbl.MPG,'CategoricalPredictors',[2 5]);
```

Create a `lime` object. Specify a predictor data set because `mdl` does not contain predictor data.

```
results = lime(mdl,tblX)
```

```
results =
  lime with properties:

             BlackboxModel: [1x1 RegressionKernel]
              DataLocality: 'global'
     CategoricalPredictors: [2 5]
                      Type: 'regression'
                         X: [392x6 table]
                QueryPoint: []
    NumImportantPredictors: []
          NumSyntheticData: 5000
             SyntheticData: [5000x6 table]
                    Fitted: [5000x1 double]
               SimpleModel: []
       ImportantPredictors: []
            BlackboxFitted: []
          SimpleModelFitted: []
```

`results` contains the generated synthetic data set. The `SimpleModel` property is empty (`[]`).

Fit a linear simple model for the first observation in `tblX`. Specify the number of important predictors to find as 3.

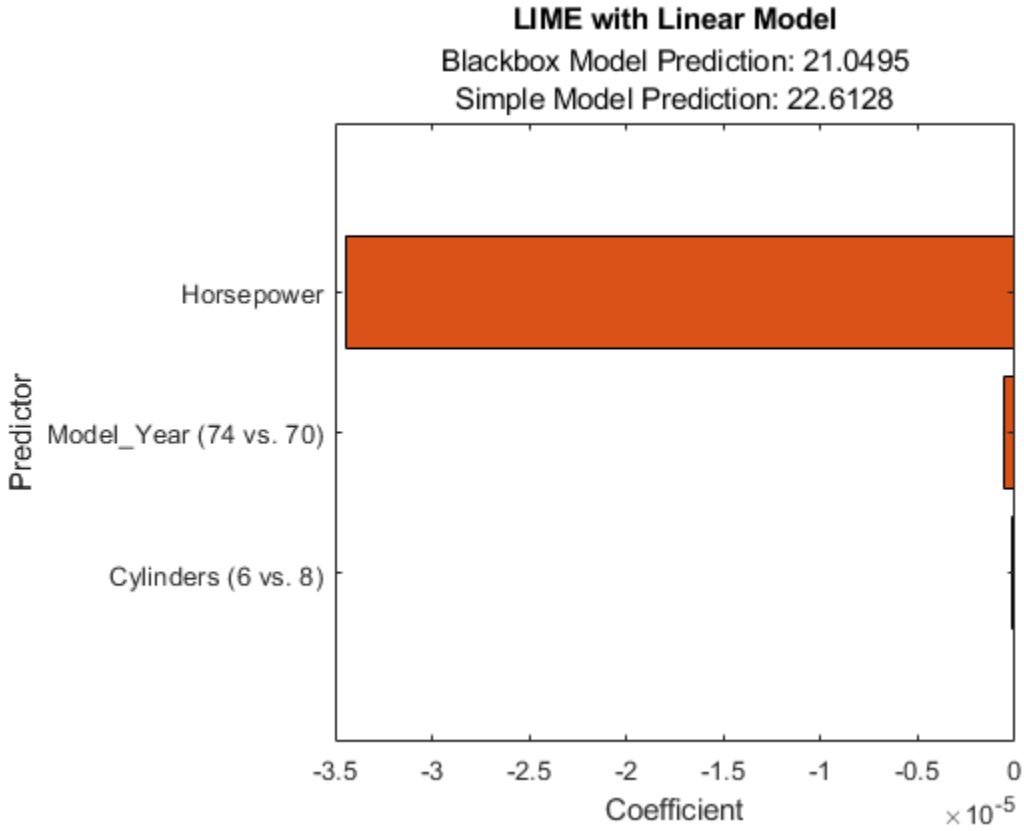
```
queryPoint = tblX(1,:)
```

```
queryPoint=1×6 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight
    ____________    _________    ____________    __________    __________    ______

         12             8            307            130            70          3504
```

```
results = fit(results,queryPoint,3);
```

Plot the `lime` object `results` by using the object function `plot`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The plot displays two predictions for the query point, which correspond to the “BlackboxFitted” on page 33-0 property and the “SimpleModelFitted” on page 33-0 property of `results`.

The horizontal bar graph shows the coefficient values of the simple model, sorted by their absolute values. LIME finds `Horsepower`, `Model_Year`, and `Cylinders` as important predictors for the query point.

`Model_Year` and `Cylinders` are categorical predictors that have multiple categories. For a linear simple model, the software creates one less dummy variable than the number of categories for each categorical predictor. The bar graph displays only the most important dummy variable. You can check the coefficients of the other dummy variables using the `SimpleModel` property of `results`. Display the sorted coefficient values, including all categorical dummy variables.

```
[~,I] = sort(abs(results.SimpleModel.Beta),'descend');
table(results.SimpleModel.ExpandedPredictorNames(I)',results.SimpleModel.Beta(I), ...
    'VariableNames',{'Exteded Predictor Name','Coefficient'})
```

```
ans=17×2 table
      Exteded Predictor Name      Coefficient
    __________________________    ___________

    {'Horsepower'            }    -3.4485e-05
    {'Model_Year (74 vs. 70)'}    -6.1279e-07
    {'Model_Year (80 vs. 70)'}     -4.015e-07
    {'Model_Year (81 vs. 70)'}     3.4176e-07
    {'Model_Year (82 vs. 70)'}    -2.2483e-07
    {'Cylinders (6 vs. 8)'   }    -1.9024e-07
```

```
    {'Model_Year (76 vs. 70)'}     1.8136e-07
    {'Cylinders (5 vs. 8)'   }     1.7461e-07
    {'Model_Year (71 vs. 70)'}      1.558e-07
    {'Model_Year (75 vs. 70)'}     1.5456e-07
    {'Model_Year (77 vs. 70)'}      1.521e-07
    {'Model_Year (78 vs. 70)'}     1.4272e-07
    {'Model_Year (72 vs. 70)'}     6.7001e-08
    {'Model_Year (73 vs. 70)'}     4.7214e-08
    {'Cylinders (4 vs. 8)'   }     4.5118e-08
    {'Model_Year (79 vs. 70)'}    -2.2598e-08
      ⋮
```

Fit Simple Models for Multiple Query Points

Train a classification model and create a `lime` object that uses a decision tree simple model. Fit multiple models for multiple query points.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Create a table of predictor variables by removing the columns of customer IDs and ratings from `tbl`.

```
tblX = removevars(tbl,["ID","Rating"]);
```

Train a blackbox model of credit ratings by using the `fitcecoc` function.

```
blackbox = fitcecoc(tblX,tbl.Rating,'CategoricalPredictors','Industry')
```

```
blackbox =
  ClassificationECOC
           PredictorNames: {1x6 cell}
             ResponseName: 'Y'
    CategoricalPredictors: 6
               ClassNames: {'A'  'AA'  'AAA'  'B'  'BB'  'BBB'  'CCC'}
           ScoreTransform: 'none'
           BinaryLearners: {21x1 cell}
               CodingName: 'onevsone'


  Properties, Methods
```

Create a `lime` object with the `blackbox` model.

```
rng('default') % For reproducibility
results = lime(blackbox);
```

Find two query points whose true rating values are AAA and B, respectively.

```
queryPoint(1,:) = tblX(find(strcmp(tbl.Rating,'AAA'),1),:);
queryPoint(2,:) = tblX(find(strcmp(tbl.Rating,'B'),1),:)
```

```
queryPoint=2×6 table
    WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry
```

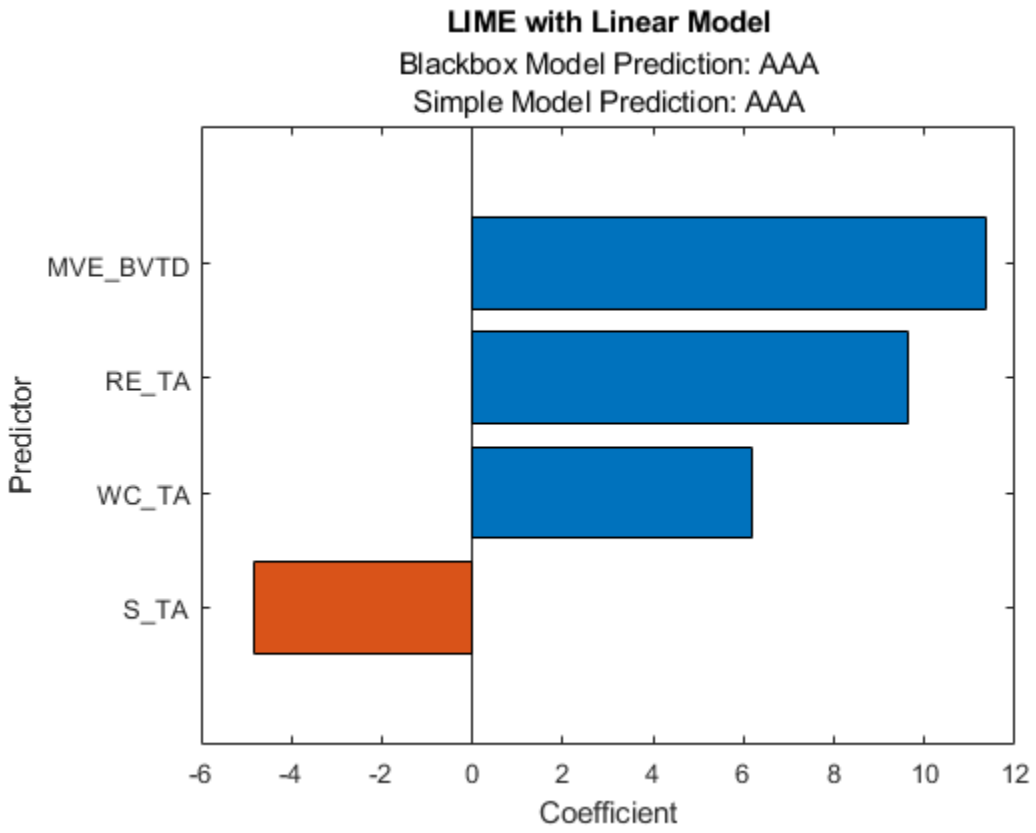
```
    _____    _____    _______    ________    _____    ________

    0.121    0.413     0.057      3.647      0.466       12
    0.019    0.009     0.042      0.257      0.119        1
```

Fit a linear simple model for the first query point. Set the number of important predictors to 4.

```
newresults1 = fit(results,queryPoint(1,:),4);
```

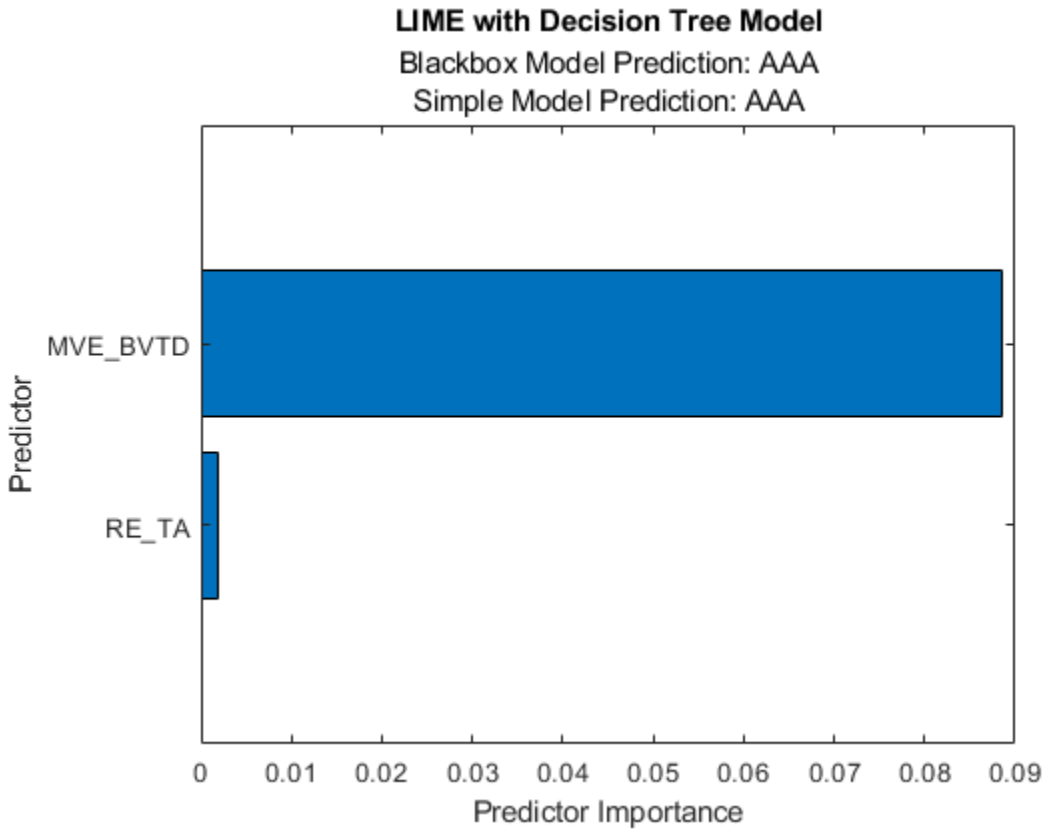
Plot the LIME results `newresults1` for the first query point. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f1 = plot(newresults1);
f1.CurrentAxes.TickLabelInterpreter = 'none';
```



Fit a linear decision tree model for the first query point.

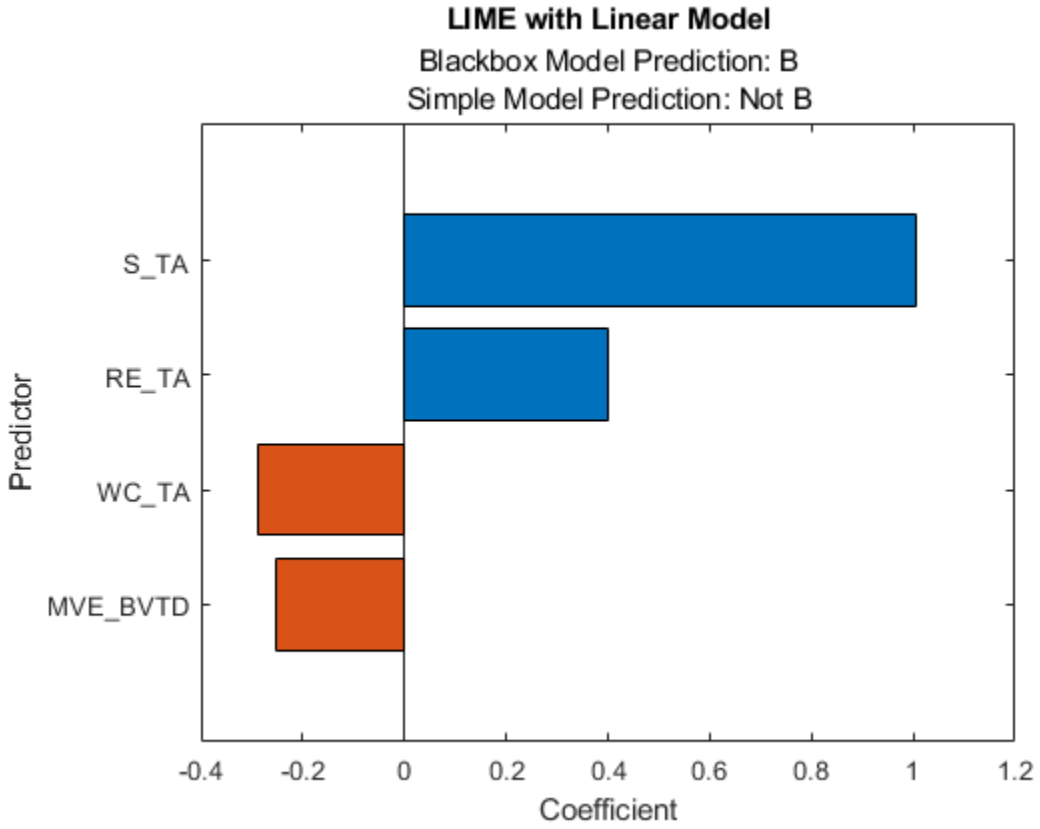
```
newresults2 = fit(results,queryPoint(1,:),6,'SimpleModelType','tree');
f2 = plot(newresults2);
f2.CurrentAxes.TickLabelInterpreter = 'none';
```



The simple models in `newresults1` and `newresults2` both find `MVE_BVTD` and `RE_TA` as important predictors.

Fit a linear simple model for the second query point, and plot the LIME results for the second query point.

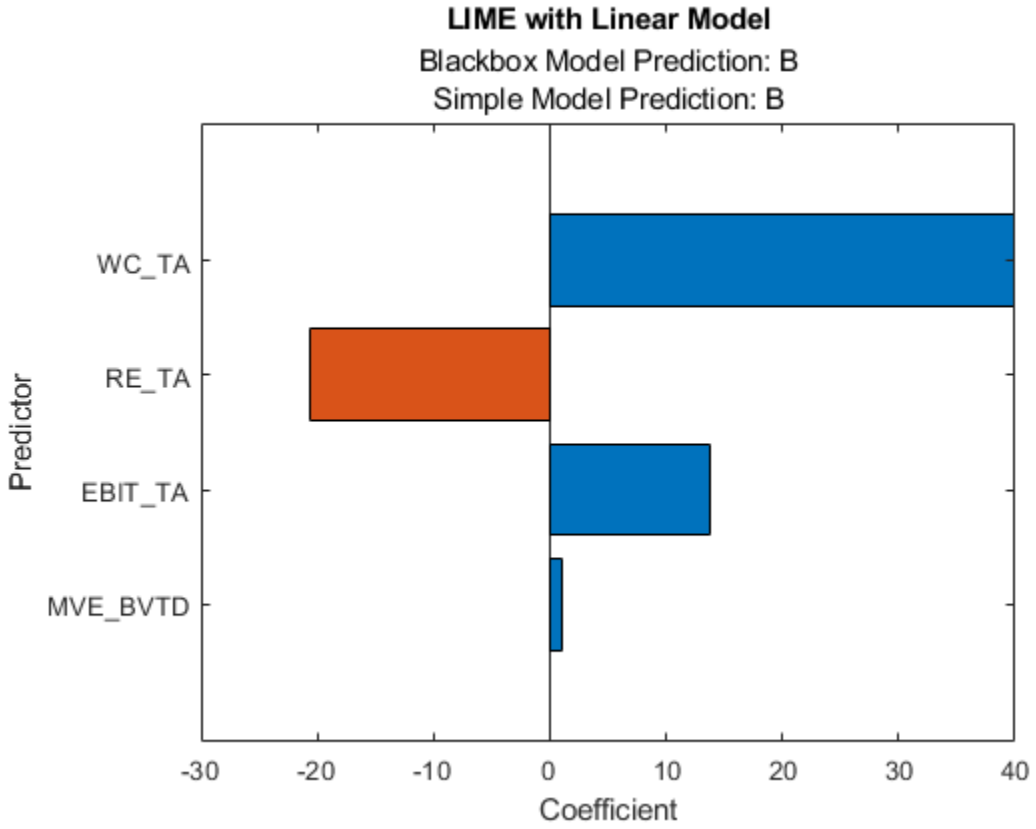
```
newresults3 = fit(results,queryPoint(2,:),4);
f3 = plot(newresults3);
f3.CurrentAxes.TickLabelInterpreter = 'none';
```



The prediction from the `blackbox` model is `B`, but the prediction from the simple model is not `B`. When the two predictions are not the same, you can specify a smaller `'KernelWidth'` value. The software fits a simple model using weights that are more focused on the samples near the query point. If a query point is an outlier or is located near a decision boundary, then the two prediction values can be different, even if you specify a small `'KernelWidth'` value. In such a case, you can change other name-value pair arguments. For example, you can generate a local synthetic data set (specify `'DataLocality'` of `lime` as `'local'`) for the query point and increase the number of samples (`'NumSyntheticData'` of `lime` or `fit`) in the synthetic data set. You can also use a different distance metric (`'Distance'` of `lime` or `fit`).

Fit a linear simple model with a small `'KernelWidth'` value.

```
newresults4 = fit(results,queryPoint(2,:),4,'KernelWidth',0.01);
f4 = plot(newresults4);
f4.CurrentAxes.TickLabelInterpreter = 'none';
```



The credit ratings for the first and second query points are `AAA` and `B`, respectively. The simple models in `newresults1` and `newresults4` both find `MVE_BVTD`, `RE_TA`, and `WC_TA` as important predictors. However, their coefficient values are different. The plots show that these predictors act differently depending on the credit ratings.

Input Arguments

`results` — LIME results

`lime` object

LIME results, specified as a `lime` object.

`queryPoint` — Query point

row vector of numeric values | single-row table

Query point around which the `fit` function fits the simple model, specified as a row vector of numeric values or a single-row table. The `queryPoint` value must have the same data type and the same number of columns as the predictor data (`results.X` or `results.SyntheticData`) in the `lime` object `results`.

`queryPoint` must not contain missing values.

Data Types: `single` | `double` | `table`

`numImportantPredictors` — Number of important predictors to use in simple model
positive integer scalar value

Number of important predictors to use in the simple model, specified as a positive integer scalar value.

- If `'SimpleModelType'` is `'linear'`, then the software selects the specified number of important predictors and fits a linear model of the selected predictors.
- If `'SimpleModelType'` is `'tree'`, then the software specifies the maximum number of decision splits (or branch nodes) as the number of important predictors so that the fitted decision tree uses at most the specified number of predictors.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumSyntheticData',2000,'SimpleModelType','tree'` sets the number of samples to generate for the synthetic data set to 2000 and specifies the simple model type as a decision tree.

`Cov` — Covariance matrix for Mahalanobis distance metric
positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a *K*-by-*K* positive definite matrix, where *K* is the number of predictors.

This argument is valid only if `'Distance'` is `'mahalanobis'`.

The default value is the `'Cov'` value that you specify when creating the `lime` object `results`. The default `'Cov'` value of `lime` is `cov(PD,'omitrows')`, where `PD` is the predictor data or synthetic predictor data. If you do not specify the `'Cov'` value, then the software uses different covariance matrices when computing the distances for both the predictor data and the synthetic predictor data.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

`Distance` — Distance metric
character vector | string scalar | function handle

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle.

- If the predictor data includes only continuous variables, then `fit` supports these distance metrics.

Value	Description
`'euclidean'`	Euclidean distance.

Value	Description
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(PD,'omitnan')`, where `PD` is the predictor data or synthetic predictor data. To specify different scaling, use the `'Scale'` name-value argument.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `PD`, `C = cov(PD,'omitrows')`. To change the value of the covariance matrix, use the `'Cov'` name-value argument.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value argument.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`t` vector containing a single observation.
	• `ZJ` is an `s`-by-`t` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `s`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance metric instead of a function handle.

- If the predictor data includes both continuous and categorical variables, then `fit` supports these distance metrics.

Value	Description
`'goodall3'`	Modified Goodall distance
`'ofd'`	Occurrence frequency distance

For definitions, see “Distance Metrics” on page 33-1475.

The default value is the `'Distance'` value that you specify when creating the `lime` object `results`. The default `'Distance'` value of `lime` is `'euclidean'` if the predictor data includes only

continuous variables, or `'goodall3'` if the predictor data includes both continuous and categorical variables.

Example: `'Distance','ofd'`

Data Types: `char` | `string` | `function_handle`

`KernelWidth` — Kernel width
numeric scalar value

Kernel width of the squared exponential (or Gaussian) kernel function, specified as the comma-separated pair consisting of `'KernelWidth'` and a numeric scalar value.

The `fit` function computes distances between the query point and the samples in the synthetic predictor data set, and then converts the distances to weights by using the squared exponential kernel function. If you lower the `'KernelWidth'` value, then `fit` uses weights that are more focused on the samples near the query point. For details, see “LIME” on page 33-1477.

The default value is the `'KernelWidth'` value that you specify when creating the `lime` object `results`. The default `'KernelWidth'` value of `lime` is 0.75.

Example: `'KernelWidth',0.5`

Data Types: `single` | `double`

`NumNeighbors` — Number of neighbors of query point
positive integer scalar value

Number of neighbors of the query point, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer scalar value. This argument is valid only when the `DataLocality` property of `results` is `'local'`.

The `fit` function estimates the distribution parameters of the predictor data using the specified number of nearest neighbors of the query point. Then the function generates synthetic predictor data using the estimated distribution.

If you specify a value larger than the number of observations in the predictor data set (`results.X`) in the `lime` object `results`, then `fit` uses all observations.

The default value is the `'NumNeighbors'` value that you specify when creating the `lime` object `results`. The default `'NumNeighbors'` value of `lime` is 1500.

Example: `'NumNeighbors',2000`

Data Types: `single` | `double`

`NumSyntheticData` — Number of samples to generate for synthetic data set
`results.NumSyntheticData` (default) | positive integer scalar value

Number of samples to generate for the synthetic data set, specified as the comma-separated pair consisting of `'NumSyntheticData'` and a positive integer scalar value. This argument is valid only when the `DataLocality` property of `results` is `'local'`.

The default value is the `NumSyntheticData` property value of `results`.

Example: `'NumSyntheticData',2500`

Data Types: `single` | `double`

P — Exponent for Minkowski distance metric
positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar.

This argument is valid only if `'Distance'` is `'minkowski'`.

The default value is the `'P'` value that you specify when creating the `lime` object `results`. The default `'P'` value of `lime` is 2.

Example: `'P',3`

Data Types: `single` | `double`

Scale — Scale parameter value for standardized Euclidean distance metric
nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector of length *K*, where *K* is the number of predictors.

This argument is valid only if `'Distance'` is `'seuclidean'`.

The default value is the `'Scale'` value that you specify when creating the `lime` object `results`. The default `'Scale'` value of `lime` is `std(PD,'omitnan')`, where `PD` is the predictor data or synthetic predictor data. If you do not specify the `'Scale'` value, then the software uses different scale parameters when computing the distances for both the predictor data and the synthetic predictor data.

Example: `'Scale',quantile(X,0.75) - quantile(X,0.25)`

Data Types: `single` | `double`

SimpleModelType — Type of simple model
`'linear'` | `'tree'`

Type of the simple model, specified as the comma-separated pair consisting of `'SimpleModelType'` and `'linear'` or `'tree'`.

- `'linear'` — The software fits a linear model by using `fitrlinear` for regression or `fitclinear` for classification.
- `'tree'` — The software fits a decision tree model by using `fitrtree` for regression or `fitctree` for classification.

The default value is the `'SimpleModelType'` value that you specify when creating the `lime` object `results`. The default `'SimpleModelType'` value of `lime` is `'linear'`.

Example: `'SimpleModelType','tree'`

Data Types: `char` | `string`

Output Arguments

newresults — LIME results
`lime` object

LIME results, returned as a `lime` object. `newresults` contains the new simple model.

To overwrite the input argument `results`, assign the output of `fit` to `results`:

```
results = fit(results,queryPoint,numImportantPredictors);
```

More About

Distance Metrics

A distance metric is a function that defines a distance between two observations. `fit` supports various distance metrics for continuous variables and a mix of continuous and categorical variables.

- Distance metrics for continuous variables

  Given an m_x-by-n data matrix X, which is treated as m_x (1-by-n) row vectors x_1, x_2, ..., x_{m_x}, and an m_y-by-n data matrix Y, which is treated as m_y (1-by-n) row vectors y_1, y_2, ...,y_{m_y}, the various distances between the vector x_s and y_t are defined as follows:

  - Euclidean distance

    $$d_{st}^2 = (x_s - y_t)(x_s - y_t)'.$$

    The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.
  - Standardized Euclidean distance

    $$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)',$$

    where V is the n-by-n diagonal matrix whose jth diagonal element is $(S(j))^2$, where S is a vector of scaling factors for each dimension.
  - Mahalanobis distance

    $$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)',$$

    where C is the covariance matrix.
  - City block distance

    $$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right|.$$

    The city block distance is a special case of the Minkowski distance, where $p = 1$.
  - Minkowski distance

    $$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right|^p}.$$

    For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.
  - Chebychev distance

$$d_{st} = \max_j\{|x_{sj} - y_{tj}|\}.$$

The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- Cosine distance

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right).$$

- Correlation distance

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}},$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

and

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}.$$

- Spearman distance

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$, as computed by `tiedrank`.
  - r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, that is, $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.
  - $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.
  - $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

- Distance metrics for a mix of continuous and categorical variables

  - Modified Goodall distance

    This distance is a variant of the Goodall distance, which assigns a small distance if the matching values are infrequent regardless of the frequencies of the other values. For mismatches, the distance contribution of the predictor is 1/(number of variables).

  - Occurrence frequency distance

    For a match, the occurrence frequency distance assigns zero distance. For a mismatch, the occurrence frequency distance assigns a higher distance on a less frequent value and a lower distance on a more frequent value.

Algorithms

LIME

To explain a prediction of a machine learning model using LIME [1], the software generates a synthetic data set and fits a simple interpretable model to the synthetic data set by using `lime` and `fit`, as described in steps 1–5.

- If you specify the `queryPoint` and `numImportantPredictors` values of `lime`, then the `lime` function performs all steps.
- If you do not specify `queryPoint` and `numImportantPredictors` and specify `'DataLocality'` as `'global'` (default), then the `lime` function generates a synthetic data set (steps 1–2), and the `fit` function fits a simple model (steps 3–5).
- If you do not specify `queryPoint` and `numImportantPredictors` and specify `'DataLocality'` as `'local'`, then the `fit` function performs all steps.

The `lime` and `fit` functions perform these steps:

1. Generate a synthetic predictor data set X_s using a multivariate normal distribution for continuous variables and a multinomial distribution for each categorical variable. You can specify the number of samples to generate by using the `'NumSyntheticData'` name-value argument.

   - If `'DataLocality'` is `'global'` (default), then the software estimates the distribution parameters from the whole predictor data set (`X` or predictor data in `blackbox`).
   - If `'DataLocality'` is `'local'`, then the software estimates the distribution parameters using the k-nearest neighbors of the query point, where k is the `'NumNeighbors'` value. You can specify a distance metric to find the nearest neighbors by using the `'Distance'` name-value argument.

   The software ignores missing values in the predictor data set when estimating the distribution parameters.

   Alternatively, you can provide a pregenerated, custom synthetic predictor data set by using the `customSyntheticData` input argument of `lime`.
2. Compute the predictions Y_s for the synthetic data set X_s. The predictions are predicted responses for regression or classified labels for classification. The software uses the `predict` function of the `blackbox` model to compute the predictions. If you specify `blackbox` as a function handle, then the software computes the predictions by using the function handle.
3. Compute the distances d between the query point and the samples in the synthetic predictor data set using the distance metric specified by `'Distance'`.
4. Compute the weight values w_q of the samples in the synthetic predictor data set with respect to the query point q using the squared exponential (or Gaussian) kernel function

   $$w_q(x_s) = \exp\left(-\frac{1}{2}\left(\frac{d(x_s, q)}{\sqrt{p}\sigma}\right)^2\right).$$

   - x_s is a sample in the synthetic predictor data set X_s.
   - $d(x_s,q)$ is the distance between the sample x_s and the query point q.
   - p is the number of predictors in X_s.
   - σ is the kernel width, which you can specify by using the `'KernelWidth'` name-value argument. The default `'KernelWidth'` value is 0.75.

The weight value at the query point is 1, and then it converges to zero as the distance value increases. The `'KernelWidth'` value controls how fast the weight value converges to zero. The lower the `'KernelWidth'` value, the faster the weight value converges to zero. Therefore, the algorithm gives more weight to samples near the query point. Because this algorithm uses such weight values, the selected important predictors and fitted simple model effectively explain the predictions for the synthetic data locally, around the query point.

5 Fit a simple model.

- If `'SimpleModelType'` is `'linear'` (default), then the software selects important predictors and fits a linear model of the selected important predictors.
  - Select n important predictors ($\tilde{X}_s$) by using the group orthogonal matching pursuit (OMP) algorithm [2][3], where n is the `numImportantPredictors` value. This algorithm uses the synthetic predictor data set (X_s), predictions (Y_s), and weight values (w_q).
  - Fit a linear model of the selected important predictors ($\tilde{X}_s$) to the predictions (Y_s) using the weight values (w_q). The software uses `fitrlinear` for regression or `fitclinear` for classification. For a multiclass model, the software uses the one-versus-all scheme to construct a binary classification problem. The positive class is the predicted class for the query point from the `blackbox` model, and the negative class refers to the other classes.
- If `'SimpleModelType'` is `'tree'`, then the software fits a decision tree model by using `fitrtree` for regression or `fitctree` for classification. The software specifies the maximum number of decision splits (or branch nodes) as the number of important predictors so that the fitted decision tree uses at most the specified number of predictors.

References

[1] Ribeiro, Marco Tulio, S. Singh, and C. Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. San Francisco, California: ACM, 2016.

[2] Świrszcz, Grzegorz, Naoki Abe, and Aurélie C. Lozano. "Grouped Orthogonal Matching Pursuit for Variable Selection and Prediction." *Advances in Neural Information Processing Systems* (2009): 1150–58.

[3] Lozano, Aurélie C., Grzegorz Świrszcz, and Naoki Abe. "Group Orthogonal Matching Pursuit for Logistic Regression." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (2011): 452–60.

See Also

`lime` | `plot`

Topics

“Interpret Machine Learning Models” on page 18-256

Introduced in R2020b

fit

Train linear model for incremental learning

Syntax

```
Mdl = fit(Mdl,X,Y)
Mdl = fit(Mdl,X,Y,Name,Value)
```

Description

The `fit` function fits a configured incremental learning model for linear regression (`incrementalRegressionLinear` object) or linear binary classification (`incrementalClassificationLinear` object) to streaming data. To additionally track performance metrics using the data as it arrives, use `updateMetricsAndFit` instead.

To fit or cross-validate a regression or classification model to an entire batch of data at once, see the other machine learning models in “Regression” or “Classification”.

`Mdl = fit(Mdl,X,Y)` returns an incremental learning model `Mdl`, which represents the input incremental learning model `Mdl` trained using the predictor and response data, `X` and `Y` respectively. Specifically, `fit` implements the following procedure:

1. Initialize the solver with the configurations and linear model coefficient and bias estimates of the input incremental learning model `Mdl`.
2. Fit the model to the data, and store the updated coefficient estimates and configurations in the output model `Mdl`.

The input and output models are the same data type.

`Mdl = fit(Mdl,X,Y,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify that the columns of the predictor data matrix correspond to observations, and set observation weights.

Examples

Incrementally Train Model

Create a default incremental linear SVM model for binary classification. Specify an estimation period of 5000 observations and the SGD solver.

```
Mdl = incrementalClassificationLinear('EstimationPeriod',5000,'Solver','sgd')
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
```

```
               Beta: [0x1 double]
               Bias: 0
            Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Fit the incremental model to the training data, in chunks of 50 observations at a time by using the `fit` function. At each iteration:

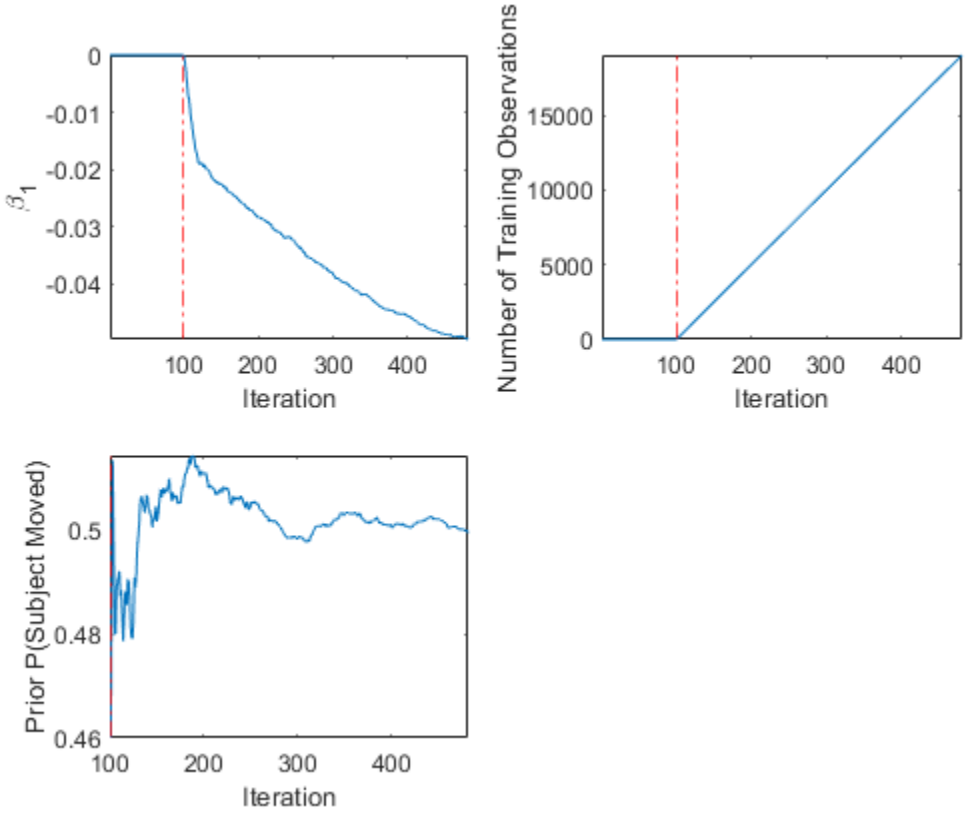
- Simulate a data stream by processing 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the number of training observations, and the prior probability of whether the subject moved (`Y = true`) to see how they evolve during incremental training.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
beta1 = zeros(nchunk,1);
numtrainobs = zeros(nchunk,1);
priormoved = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = fit(Mdl,X(idx,:),Y(idx));
    beta1(j) = Mdl.Beta(1);
    numtrainobs(j) = Mdl.NumTrainingObservations;
    priormoved(j) = Mdl.Prior(Mdl.ClassNames == true);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

To see how the parameters evolved during incremental learning, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(beta1)
ylabel('\beta_1')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
axis tight
subplot(2,2,2)
plot(numtrainobs);
ylabel('Number of Training Observations')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
axis tight
subplot(2,2,3)
plot(priormoved);
ylabel('Prior P(Subject Moved)')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
axis tight
```



The plot suggests that `fit` does not fit the model to the data or update the parameters until after the estimation period.

Specify Orientation of Observations and Observation Weights

Train a linear model for binary classification by using `fitclinear`, convert it to an incremental learner, track its performance, and fit it to streaming data. Orient the observations in columns, and specify observation weights.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data. Orient the observations of the predictor data in columns.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:)';
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Suppose that the data collected when the subject was not moving (`Y = false`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from a still subject, and 1 to a moving subject.

```
W = ones(n,1) + ~Y;
```

Train Linear Model for Binary Classification

Fit a linear model for binary classification to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitclinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns',...
    'Weights',W(idxtt))
```

```
TTMdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.1107
            Lambda: 8.2967e-05
           Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear model for binary classification.

Convert Trained Model

Convert the traditionally trained classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

            IsWarm: 1
           Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.1107
           Learner: 'svm'


  Properties, Methods
```

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. At each iteration:

1. Simulate a data stream by processing 50 observations at a time.
2. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify that the observations are oriented in columns, and specify the observation weights.
3. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify that the observations are oriented in columns, and specify the observation weights.
4. Store the classification error and first estimated coefficient β_1.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];
Xil = X(:,idxil);
Yil = Y(idxil);
Wil = W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(:,idx),Yil(idx),...
        'ObservationsIn','columns','Weights',Wil(idx));
```

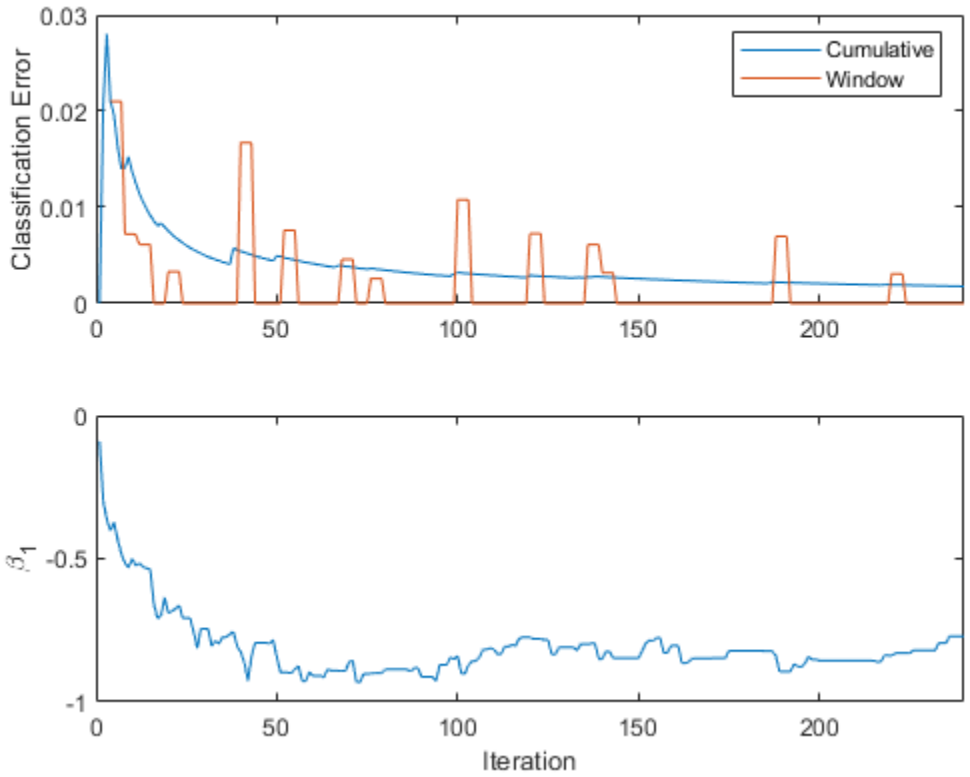
```
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(:,idx),Yil(idx),'ObservationsIn','columns',...
        'Weights',Wil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Plot a trace plot of the performance metrics and estimated coefficient β_1.

```
figure;
subplot(2,1,1)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
legend(h,ce.Properties.VariableNames)
subplot(2,1,2)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable and gradually decreases, whereas the window loss jumps.

β_1 changes gradually, then levels off, as `fit` processes more chunks.

Perform Conditional Training

Incrementally train a linear regression model only when its performance degrades.

Load and shuffle the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015

rng(1) % For reproducibility
n = size(NYCHousing2015,1);
shuffidx = randsample(n,n);
NYCHousing2015 = NYCHousing2015(shuffidx,:);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}];
```

Configure a linear regression model for incremental learning so that it does not have an estimation or metrics warm-up period. Specify a metrics window size of 1000. Fit the configured model to the first 100 observations.

```
Mdl = incrementalRegressionLinear('EstimationPeriod',0,'MetricsWarmupPeriod',0,'MetricsWindowSize
numObsPerChunk = 100;
Mdl = fit(Mdl,X(1:numObsPerChunk,:),Y(1:numObsPerChunk));
```

`Mdl` is an `incrementalRegressionLinear` model object.

Perform incremental learning, with conditional fitting, by following this procedure for each iteration:

- Simulate a data stream by processing a chunk of 100 observations at a time.
- Update the model performance by computing the epsilon insensitive loss, within a 200 observation window.
- Fit the model to the chunk of data only when the loss more than doubles from the minimum loss experienced.
- When tracking performance and fitting, overwrite the previous incremental model.

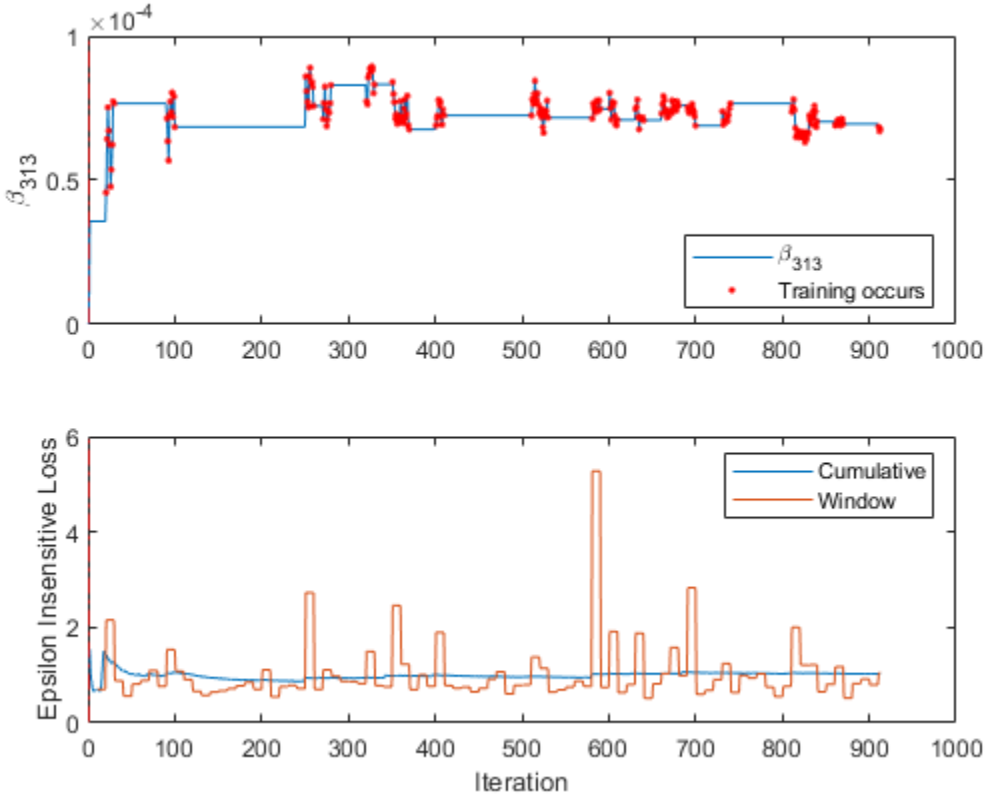
- Store the epsilon insensitive loss and β_{313} to see how the loss and coefficient evolve during training.
- Track when `fit` trains the model.

```
% Preallocation
n = numel(Y) - numObsPerChunk;
nchunk = floor(n/numObsPerChunk);
beta313 = zeros(nchunk,1);
ei = array2table(nan(nchunk,2),'VariableNames',["Cumulative" "Window"]);
trained = false(nchunk,1);

% Incremental fitting
for j = 2:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ei{j,:} = Mdl.Metrics{"EpsilonInsensitiveLoss",:};
    minei = min(ei{:,2});
    pdiffloss = (ei{j,2} - minei)/minei*100;
    if pdiffloss > 100
        Mdl = fit(Mdl,X(idx,:),Y(idx));
        trained(j) = true;
    end
    beta313(j) = Mdl.Beta(end);
end
```

`Mdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

To see how the model performance and β_{313} evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(beta313)
hold on
plot(find(trained),beta313(trained),'r.')
ylabel('\beta_{313}')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
legend('\beta_{313}','Training occurs','Location','southeast')
hold off
subplot(2,1,2)
plot(ei.Variables)
ylabel('Epsilon Insensitive Loss')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
legend(ei.Properties.VariableNames)
```



The trace plot of β_{313} shows periods of constant values, during which the loss did not double from the minimum experienced.

Input Arguments

`Mdl` — Incremental learning model to fit to streaming data

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Incremental learning model to fit to streaming data, specified as an `incrementalClassificationLinear` or `incrementalRegressionLinear` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

`X` — Chunk of predictor data

floating-point matrix

Chunk of predictor data to which the model is fit, specified as a floating-point matrix of n observations and `Mdl.NumPredictors` predictor variables. The value of the `'ObservationsIn'` name-value pair argument determines the orientation of the variables and observations.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`j`)` is the label of observation j (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `fit` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `fit` issues an error.
- `fit` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

Y — Chunk of labels

categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels to which the model is fit, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors for classification problems; or a floating-point vector for regression problems.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

For classification problems:

- `fit` supports binary classification only.
- When the `ClassNames` property of the input model `Mdl` is nonempty, the following conditions apply:
  - If `Y` contains a label that is not a member of `Mdl.ClassNames`, `fit` issues an error.
  - The data type of `Y` and `Mdl.ClassNames` must be the same.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Note

- If an observation (predictor or label) or weight `Weight` contains at least one missing (`NaN`) value, `fit` ignores the observation. Consequently, `fit` uses fewer than *n* observations to compute the model performance.
- The chunk size *n* and the stochastic gradient descent (SGD) hyperparameter batch size (`Mdl.BatchSize`) can be different values. If *n* < `Mdl.BatchSize`, `fit` uses the *n* available observations when it applies SGD.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ObservationsIn','columns','Weights',W` specifies that the columns of the predictor matrix correspond to observations, and the vector `W` contains observation weights to apply during incremental learning.

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`.

Data Types: `char` | `string`

`Weights` — Chunk of observation weights
floating-point vector of positive values

Chunk of observation weights, specified as the comma-separated pair consisting of `'Weights'` and a floating-point vector of positive values. `fit` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(`*n*`,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-1490.

Data Types: `double` | `single`

Output Arguments

`Mdl` — Updated incremental learning model
`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Updated incremental learning model, returned as an incremental learning model object of the same data type as the input model `Mdl`, either `incrementalClassificationLinear` or `incrementalRegressionLinear`.

If `Mdl.EstimationPeriod` > 0, the incremental fitting functions `updateMetricsAndFit` and `fit` estimate hyperparameters using the first `Mdl.EstimationPeriod` observations passed to either function; they do not train the input model to that data. However, if an incoming chunk of *n* observations is greater than or equal to the number of observations remaining in the estimation period *m*, `fit` estimates hyperparameters using the first *n* – *m* observations, and fits the input model to the remaining *m* observations. Consequently, the software updates the `Beta` and `Bias` properties, hyperparameter properties, and record keeping properties such as `NumTrainingObservations`.

For classification problems, if the `ClassNames` property of the input model `Mdl` is an empty array, `fit` sets the `ClassNames` property of the output model `Mdl` to `unique(Y)`.

Tips

- Unlike traditional training, incremental learning might not have a separate test (holdout) set. Therefore, to treat each incoming chunk of data as a test set, pass the incremental model and each incoming chunk to `updateMetrics` before training the model on the same data.

Algorithms

Observation Weights

For classification problems, if the prior class probability distribution is known (in other words, the prior distribution is not empirical), `fit` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that observation weights are the respective prior class probabilities by default.

For regression problems or if the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `fit`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `fit` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `fit` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for fit, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `fit`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For usage notes and limitations of the model object, see `incrementalClassificationLinear` or `incrementalRegressionLinear`.
`X`	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `Y`. • The number of predictor variables must equal to `Mdl.NumPredictors`. • `X` must be `single` or `double`.
`Y`	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `X`. • For classification problems, all labels in `Y` must be represented in `Mdl.ClassNames`. • `Y` and `Mdl.ClassNames` must have the same data type.

- The following restrictions apply:

- If you configure `Mdl` to shuffle data (`Mdl.Shuffle` is `true`, or `Mdl.Solver` is `'sgd'` or `'asgd'`), the `fit` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB. Therefore, the fitted coefficients computed in MATLAB and the generated code might not be equal.
- Use a homogeneous data type for all floating-point input arguments and object properties, specifically, either `single` or `double`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Objects
`incrementalClassificationLinear` | `incrementalRegressionLinear`

Functions
`predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2020b

fit

Train naive Bayes classification model for incremental learning

Syntax

```
Mdl = fit(Mdl,X,Y)
Mdl = fit(Mdl,X,Y,'Weights',Weights)
```

Description

The `fit` function fits a configured naive Bayes classification model for incremental learning (`incrementalClassificationNaiveBayes` object) to streaming data. To additionally track performance metrics using the data as it arrives, use `updateMetricsAndFit` instead.

To fit or cross-validate a naive Bayes classification model to an entire batch of data at once, see `fitcnb`.

`Mdl = fit(Mdl,X,Y)` returns a naive Bayes classification model for incremental learning `Mdl`, which represents the input naive Bayes classification model for incremental learning `Mdl` trained using the predictor and response data, `X` and `Y` respectively. Specifically, `fit` updates the conditional posterior distribution of the predictor variables given the data.

`Mdl = fit(Mdl,X,Y,'Weights',Weights)` specifies observation weights `Weights`.

Examples

Incrementally Train Model Without Little Prior Information

This example shows how to fit an incremental naive Bayes learner when you know only the expected maximum number of classes in the data.

Create a incremental naive Bayes model. Specify that the maximum number of expected classes is 5.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5)
```

```
Mdl =
  incrementalClassificationNaiveBayes

                  IsWarm: 0
                 Metrics: [1×2 table]
              ClassNames: [1×0 double]
          ScoreTransform: 'none'
       DistributionNames: 'normal'
  DistributionParameters: {}


  Properties, Methods
```

`Mdl` is an `incrementalClassificationNaiveBayes` model. All its properties are read-only. `Mdl` can encounter at most 5 unique classes. By default, the prior class distribution `Mdl.Prior` is empirical, which means the software updates the prior distribution as it encounters labels.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Fit the incremental model to the training data, in chunks of 50 observations at a time by using the `fit` function. At each iteration:

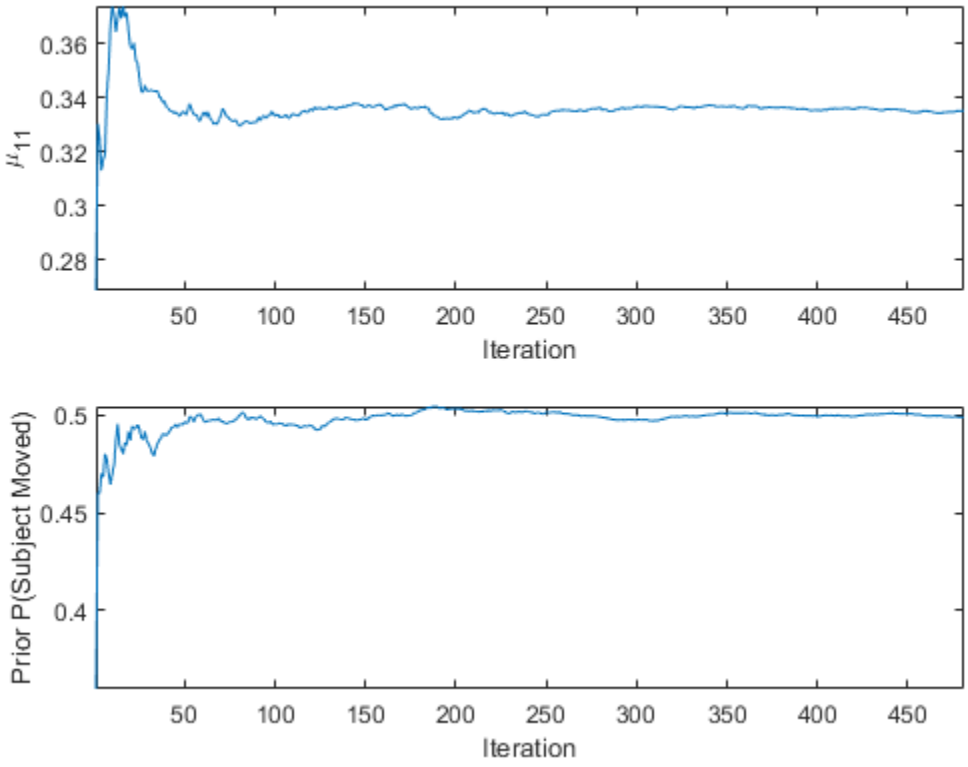
- Simulate a data stream by processing 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the mean of the first predictor in the first class μ_{11} and the prior probability that the subject is moving (`Y` > 2) to see how they evolve during incremental training.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
mu11 = zeros(nchunk,1);
priormoved = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = fit(Mdl,X(idx,:),Y(idx));
    mu11(j) = Mdl.DistributionParameters{1,1}(1);
    priormoved(j) = sum(Mdl.Prior(Mdl.ClassNames > 2));
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the parameters evolved during incremental learning, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(mu11)
ylabel('\mu_{11}')
xlabel('Iteration')
axis tight
subplot(2,1,2)
plot(priormoved);
ylabel('Prior P(Subject Moved)')
xlabel('Iteration')
axis tight
```



`fit` updates the posterior mean of the predictor distribution as it processes each chunk. Because the prior class distribution is empirical, π(subject is moving) changes as `fit` processes each chunk.

Specify All Class Names Before Fitting

This example shows how to fit an incremental naive Bayes learner when you know all the class names in the data.

Consider training a device to predict whether a subject is sitting, standing, walking, running, or dancing based on biometric data measured on the subject, and you know the class names map 1 through 5 to an activity. Also, suppose that the researchers plan to expose the device to each class uniformly.

Create an incremental naive Bayes learner for multiclass learning. Specify the class names and a uniform prior class distribution.

```
classnames = 1:5;
Mdl = incrementalClassificationNaiveBayes('ClassNames',classnames,'Prior','uniform')
```

```
Mdl =
  incrementalClassificationNaiveBayes

                  IsWarm: 0
                 Metrics: [1×2 table]
```

```
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
        DistributionNames: 'normal'
    DistributionParameters: {5×0 cell}


  Properties, Methods
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object. All its properties are read-only. During training, observed labels must be in `Mdl.ClassNames`.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Fit the incremental model to the training data by using the `fit` function. Simulate a data stream by processing chunks of 50 observations at a time. At each iteration:

- Process 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the mean of the first predictor in the first class μ_{11} and the prior probability that the subject is moving (`Y` > 2) to see how they evolve during incremental training.

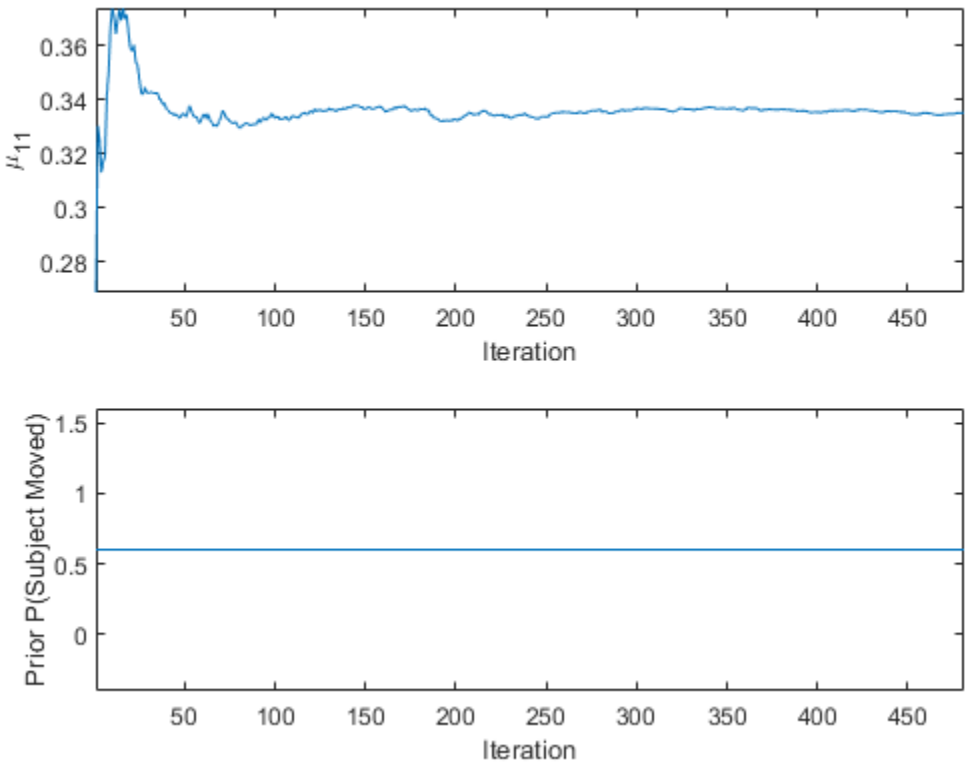
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
mu11 = zeros(nchunk,1);
priormoved = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = fit(Mdl,X(idx,:),Y(idx));
    mu11(j) = Mdl.DistributionParameters{1,1}(1);
    priormoved(j) = sum(Mdl.Prior(Mdl.ClassNames > 2));
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the parameters evolved during incremental learning, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(mu11)
```

```
ylabel('\mu_{11}')
xlabel('Iteration')
axis tight
subplot(2,1,2)
plot(priormoved);
ylabel('Prior P(Subject Moved)')
xlabel('Iteration')
axis tight
```



`fit` updates the posterior mean of the predictor distribution as it processes each chunk. Because the prior class distribution is specified as uniform, π(subject is moving) = 0.6 and does not change as `fit` processes each chunk.

Specify Observation Weights

Train a naive Bayes classification model by using `fitcnb`, convert it to an incremental learner, track its performance on streaming data, and then fit it to the data. Specify observation weights.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
```

```
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Suppose that the data collected when the subject was not moving (`Y <= 2`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from a still subject, and 1 to a moving subject.

```
W = ones(n,1) + ~Y;
```

Train Naive Bayes Classification Model

Fit a naive Bayes classification model to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitcnb(X(idxtt,:),Y(idxtt),'Weights',W(idxtt))
```

```
TTMdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 12053
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained naive Bayes classification model.

Convert Trained Model

Convert the traditionally trained model to a naive Bayes classification for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                    IsWarm: 1
                   Metrics: [1×2 table]
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model. Because class names are specified in `Mdl.ClassNames`, labels encountered during incremental learning must be in `Mdl.ClassNames`.

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. At each iteration:

1. Simulate a data stream by processing 50 observations at a time.
2. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify the observation weights.
3. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify the observation weights.
4. Store the minimal cost.

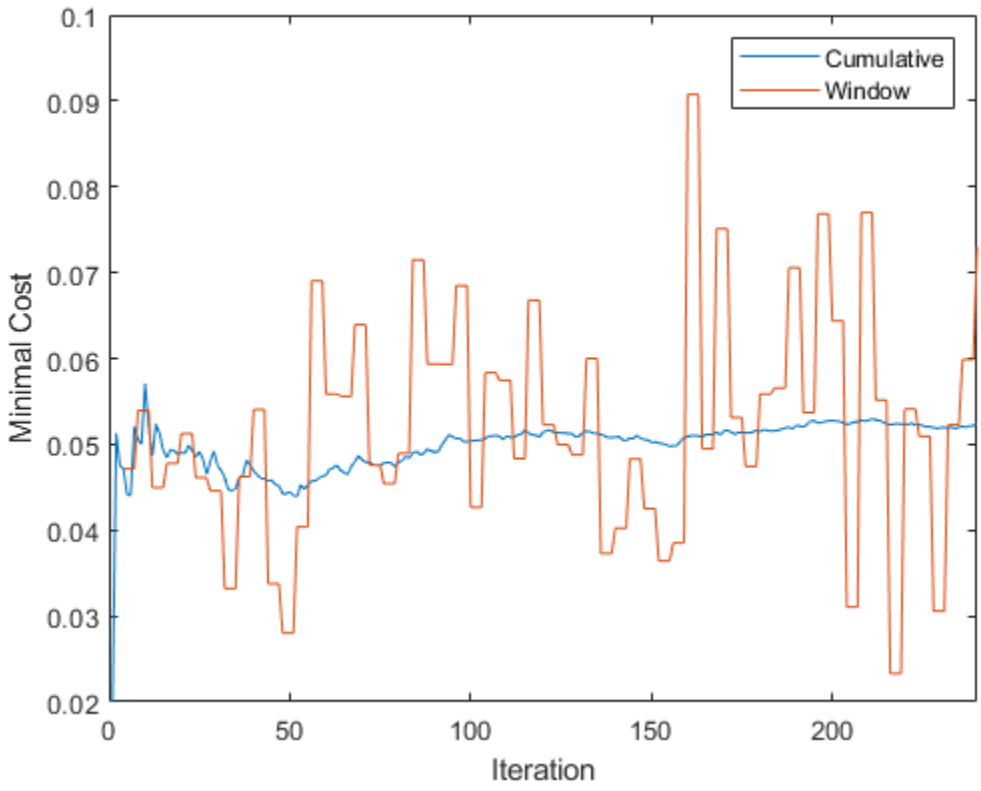
```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
Xil = X(idxil,:);
Yil = Y(idxil);
Wil = W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(idx,:),Yil(idx),...
        'Weights',Wil(idx));
    mc{j,:} = IncrementalMdl.Metrics{"MinimalCost",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx),'Weights',Wil(idx));
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Plot a trace plot of the performance metrics.

```
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
legend(h,mc.Properties.VariableNames)
xlabel('Iteration')
```



The cumulative loss gradually stabilizes, whereas the window loss jumps.

Perform Conditional Training

Incrementally train a naive Bayes classification model only when its performance degrades.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Configure a naive Bayes classification model for incremental learning so that the maximum number of expected classes is 5, the tracked performance metric includes the misclassification error rate, and the metrics window size of 1000. Fit the configured model to the first 1000 observations.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5,'MetricsWindowSize',1000,...
    'Metrics','classiferror');
initobs = 1000;
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs));
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object.

Perform incremental learning, with conditional fitting, by following this procedure for each iteration:

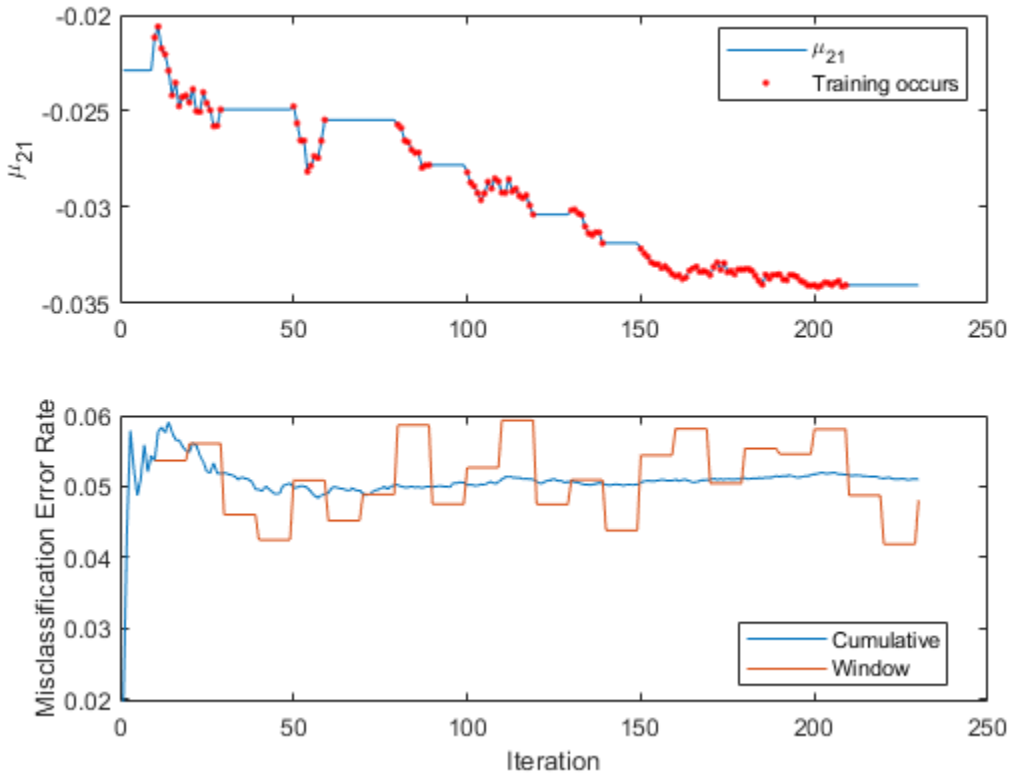
- Simulate a data stream by processing a chunk of 100 observations at a time.
- Update the model performance on the incoming chunk of data.
- Fit the model to the chunk of data only when the misclassification error rate is greater than 0.05.
- When tracking performance and fitting, overwrite the previous incremental model.
- Store the misclassification error rate and the mean of the first predictor in the second class μ_{21} to see how they evolve during training.
- Track when `fit` trains the model.

```
% Preallocation
numObsPerChunk = 100;
nchunk = floor((n - initobs)/numObsPerChunk);
mu21 = zeros(nchunk,1);
ce = array2table(nan(nchunk,2),'VariableNames',["Cumulative" "Window"]);
trained = false(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    if ce{j,2} > 0.05
        Mdl = fit(Mdl,X(idx,:),Y(idx));
        trained(j) = true;
    end
    mu21(j) = Mdl.DistributionParameters{2,1}(1);
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the model performance and μ_{21} evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(mu21)
hold on
plot(find(trained),mu21(trained),'r.')
ylabel('\mu_{21}')
legend('\mu_{21}','Training occurs','Location','best')
hold off
subplot(2,1,2)
plot(ce.Variables)
ylabel('Misclassification Error Rate')
xlabel('Iteration')
legend(ce.Properties.VariableNames,'Location','best')
```



The trace plot of μ_{21} shows periods of constant values, during which the loss within the previous 1000 observation window is at most 0.05.

Input Arguments

`Mdl` — Naive Bayes classification model for incremental learning to fit to streaming data

`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning to fit to streaming data, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

`X` — Chunk of predictor data

floating-point matrix

Chunk of predictor data to which the model is fit, specified as an *n*-by-`Mdl.NumPredictors` floating point matrix.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `fit` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `fit` issues an error.
- `fit` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`Y` — Chunk of labels
categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels to which the model is fit, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*`j`*`)` is the label of observation *j* (row or column) in `X`. `fit` issues an error when at least one of the conditions is met:

- `Y` contains a newly encountered label and the maximum number of classes has been reached previously (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`).
- The data types of `Y` and `Mdl.ClassNames` are different.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

`Weights` — Chunk of observation weights
floating-point vector of positive values

Chunk of observation weights, specified as a floating-point vector of positive values. `fit` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(`*`n`*`,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-1503.

Data Types: `double` | `single`

Note

If an observation (predictor or label) or weight contains at least one missing (`NaN`) value, `fit` ignores the observation. Consequently, `fit` uses fewer than *n* observations to compute the model performance.

Output Arguments

`Mdl` — Updated naive Bayes classification model for incremental learning

`incrementalClassificationNaiveBayes` model object

Updated naive Bayes classification model for incremental learning, returned as an incremental learning model object of the same data type as the input model `Mdl`, an `incrementalClassificationNaiveBayes` object.

If the `ClassNames` property of the input model `Mdl` is an empty array, `fit` sets the `ClassNames` property of the output model `Mdl` to `unique(Y)`. If the maximum number of classes is not reached, `fit` appends to `Mdl.ClassNames` any newly encountered labels in `Y`.

Tips

- Unlike traditional training, incremental learning might not have a separate test (holdout) set. Therefore, to treat each incoming chunk of data as a test set, pass the incremental model and each incoming chunk to `updateMetrics` before training the model on the same data.

Algorithms

Observation Weights

For each conditional predictor distribution, `fit` computes the weighted average and standard deviation.

If the prior class probability distribution is known (in other words, the prior distribution is not empirical), `fit` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that the default observation weights are the respective prior class probabilities.

If the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `fit`.

See Also

Objects

`incrementalClassificationNaiveBayes`

Functions

`predict` | `updateMetrics` | `updateMetricsAndFit`

Topics

“Incremental Learning Overview” on page 26-2
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2021a

fit

Compute Shapley values for query point

Syntax

```
newExplainer = fit(explainer,queryPoint)
newExplainer = fit(explainer,queryPoint,Name,Value)
```

Description

`newExplainer = fit(explainer,queryPoint)` computes the Shapley values for the specified query point (`queryPoint`) and stores the computed Shapley values in the `ShapleyValues` property of `newExplainer`. The `shapley` object `explainer` contains a machine learning model and the options for computing Shapley values.

`fit` uses the Shapley value computation options that you specify when you create `explainer`. You can change the options using the name-value arguments of the `fit` function. The function returns a `shapley` object `newExplainer` that contains the newly computed Shapley values.

`newExplainer = fit(explainer,queryPoint,Name,Value)` specifies additional options using one or more name-value arguments. For example, specify `'UseParallel',true` to compute Shapley values in parallel.

Examples

Create shapley Object and Compute Shapley Values Using fit

Train a regression model and create a `shapley` object. When you create a `shapley` object, if you do not specify a query point, then the software does not compute Shapley values. Use the object function `fit` to compute the Shapley values for the specified query point. Then create a bar graph of the Shapley values by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Train a blackbox model of `MPG` by using the `fitrkernel` function

```
rng('default') % For reproducibility
mdl = fitrkernel(tbl,'MPG','CategoricalPredictors',[2 5]);
```

Create a `shapley` object. Specify the data set `tbl`, because `mdl` does not contain training data.

```
explainer = shapley(mdl,tbl)
```

```
explainer =
  shapley with properties:

            BlackboxModel: [1x1 RegressionKernel]
               QueryPoint: []
           BlackboxFitted: []
            ShapleyValues: []
               NumSubsets: 64
                        X: [392x7 table]
    CategoricalPredictors: [2 5]
                   Method: 'interventional-kernel'
```

`explainer` stores the training data `tbl` in the `X` property.

Compute the Shapley values of all predictor variables for the first observation in `tbl`.

```
queryPoint = tbl(1,:)
```

```
queryPoint=1×7 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight    MPG
    ____________    _________    ____________    __________    __________    ______    ___

         12             8            307            130            70          3504     18
```

```
explainer = fit(explainer,queryPoint);
```

For a regression model, `shapley` computes Shapley values using the predicted response, and stores them in the `ShapleyValues` property. Display the values in the `ShapleyValues` property.

```
explainer.ShapleyValues
```

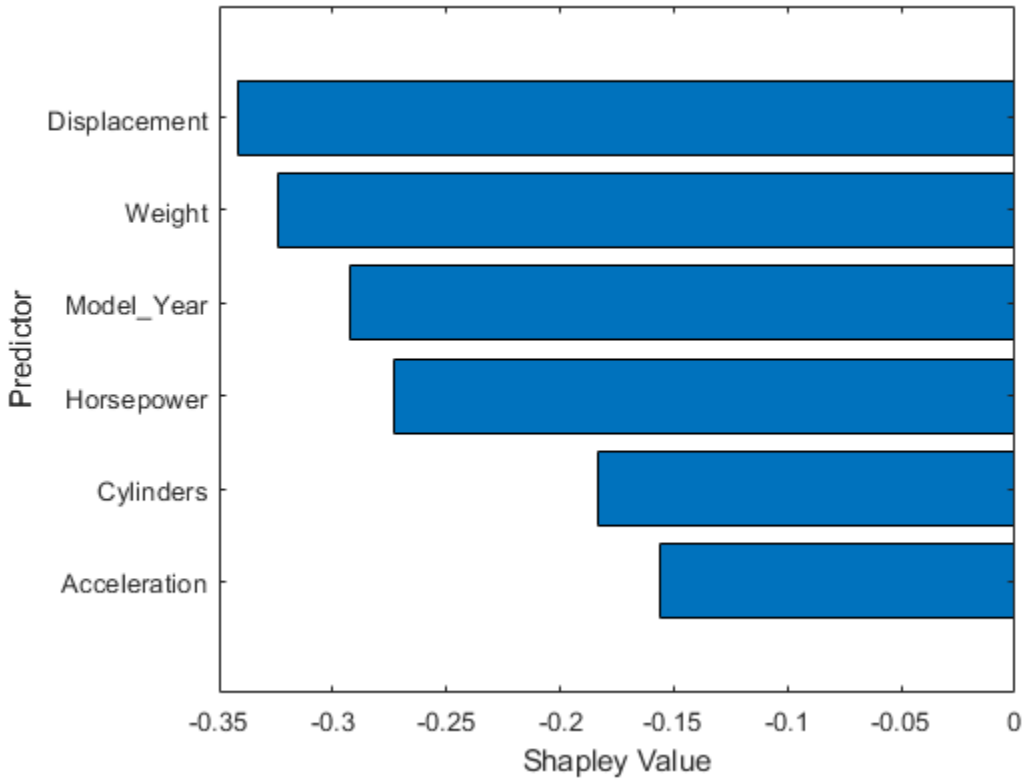
```
ans=6×2 table
      Predictor       ShapleyValue
    ________________    ____________

    "Acceleration"        -0.1561
    "Cylinders"          -0.18306
    "Displacement"       -0.34203
    "Horsepower"         -0.27291
    "Model_Year"          -0.2926
    "Weight"             -0.32402
```

Display the predicted response for the query point, and plot the Shapley values for the query point by using the `plot` function. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
explainer.BlackboxFitted
```

```
ans = 21.0495
```

```
f = figure;
plot(explainer)
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The horizontal bar graph shows the Shapley values for all variables, sorted by their absolute values. Each Shapley value explains the deviation of the prediction for the query point from the average, due to the corresponding variable.

Compute Shapley Values for Multiple Query Points

Train a classification model and create a `shapley` object. Then compute the Shapley values for multiple query points.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Train a blackbox model of credit ratings by using the `fitcecoc` function. Use the variables from the second through seventh columns in `tbl` as the predictor variables.

```
blackbox = fitcecoc(tbl,'Rating', ...
    'PredictorNames',tbl.Properties.VariableNames(2:7), ...
    'CategoricalPredictors','Industry');
```

Create a `shapley` object with the `blackbox` model. For faster computation, subsample 25% of the observations from `tbl` with stratification and use the samples to compute the Shapley values. Specify to use the extension to the kernelSHAP algorithm.

```
rng('default') % For reproducibility
c = cvpartition(tbl.Rating,'Holdout',0.25);
tbl_s = tbl(test(c),:);
explainer = shapley(blackbox,tbl_s,'Method','conditional-kernel');
```

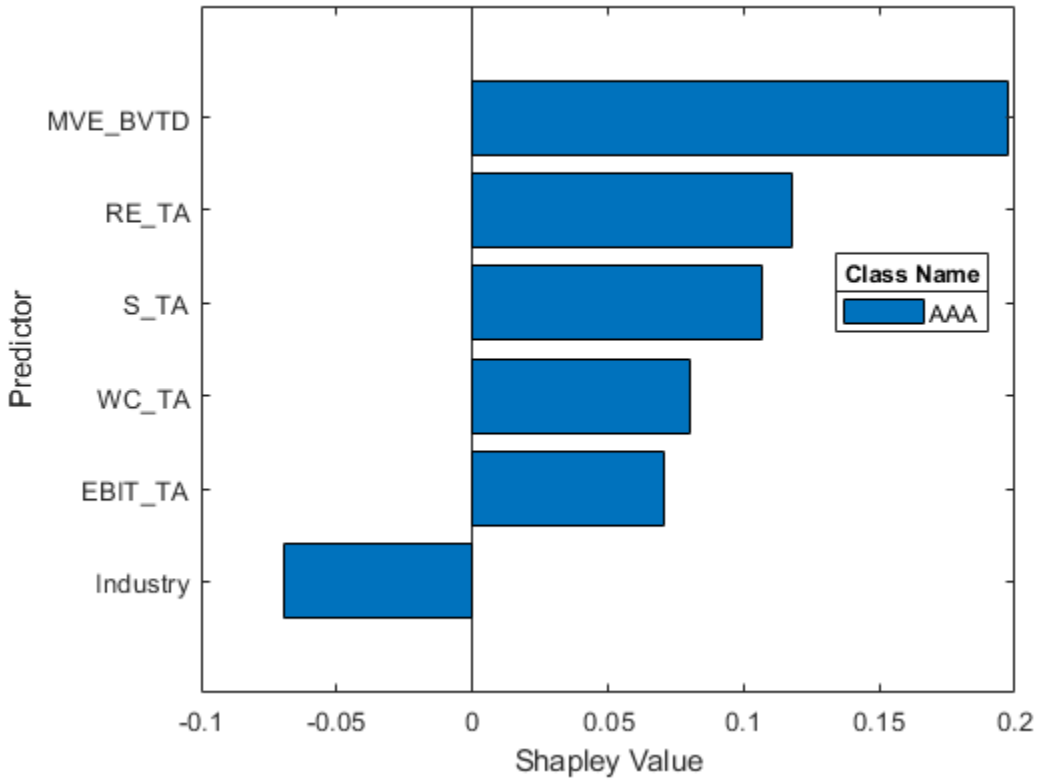
Find two query points whose true rating values are AAA and B, respectively.

```
queryPoint(1,:) = tbl_s(find(strcmp(tbl_s.Rating,'AAA'),1),:);
queryPoint(2,:) = tbl_s(find(strcmp(tbl_s.Rating,'B'),1),:)
```

```
queryPoint=2×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____     _____     _______    ________    _____    ________    _______

    58258     0.511     0.869     0.106       8.538      0.732       2        {'AAA'}
    82367    -0.078    -0.042     0.011       0.262      0.167       7        {'B'  }
```

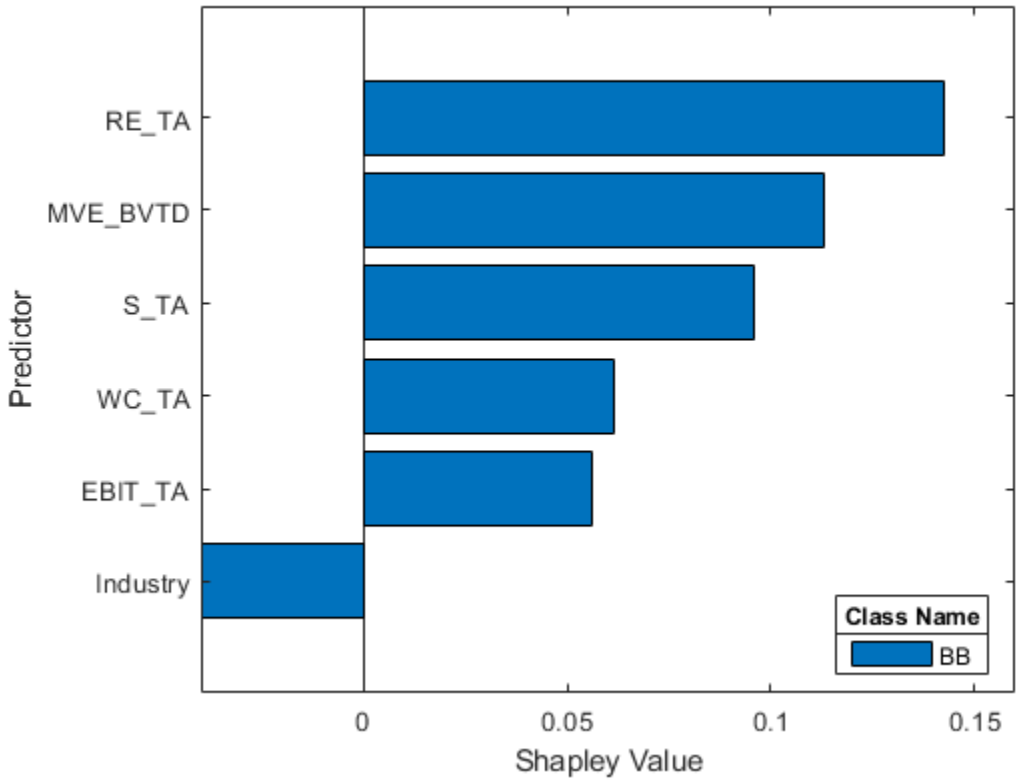
Compute and plot the Shapley values for the first query point. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
explainer1 = fit(explainer,queryPoint(1,:));
f1 = figure;
plot(explainer1)
f1.CurrentAxes.TickLabelInterpreter = 'none';
```



Compute and plot the Shapley values for the second query point.

```
explainer2 = fit(explainer,queryPoint(2,:));
f2 = figure;
plot(explainer2)
f2.CurrentAxes.TickLabelInterpreter = 'none';
```



The true rating for the second query point is `B`, but the predicted rating is `BB`. The plot shows the Shapley values for the predicted rating.

`explainer1` and `explainer2` include the Shapley values for the first query point and second query point, respectively.

Input Arguments

`explainer` — Object explaining blackbox model

`shapley` object

Object explaining the blackbox model, specified as a `shapley` object.

`queryPoint` — Query point

row vector of numeric values | single-row table

Query point at which `fit` explains a prediction, specified as a row vector of numeric values or a single-row table.

- For a row vector of numeric values:

- The variables that makes up the columns of `queryPoint` must have the same order as the predictor data `X` in `explainer`.
- If the predictor data `explainer.X` is a table, then `queryPoint` can be a numeric vector if the table contains all numeric variables.

- For a single-row table:
  - If the predictor data `explainer.X` is a table, then all predictor variables in `queryPoint` must have the same variable names and data types as those in `explainer.X`. However, the column order of `queryPoint` does not need to correspond to the column order of `explainer.X`.
  - If the predictor data `explainer.X` is a numeric matrix, then the predictor names in `explainer.BlackboxModel.PredictorNames` and the corresponding predictor variable names in `queryPoint` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `queryPoint` must be numeric vectors.
  - `queryPoint` can contain additional variables (response variables, observation weights, and so on), but `fit` ignores them.
  - `fit` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

If `queryPoint` contains `NaN`s for continuous predictors and `'Method'` is `'conditional-kernel'`, then the Shapley values (`ShapleyValues`) in the returned object are `NaN`s. Otherwise, `fit` handles `NaN` values in the same way as `explainer.BlackboxModel` (the `predict` object function of `explainer.BlackboxModel` or a function handle specified by `blackbox`).

Example: `explainer.X(1,:)` specifies the query point as the first observation of the predictor data `X` in `explainer`.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fit(explainer,q,'Method','conditional-kernel','UseParallel',true)` computes the Shapley values for the query point `q` using the extension to the kernelSHAP algorithm, and executes the computation in parallel.

`MaxNumSubsets` — Maximum number of predictor subsets

`explainer.NumSubsets` (default) | positive integer

Maximum number of predictor subsets to use for Shapley value computation, specified as a positive integer.

For details on how `fit` chooses the subsets to use, see “Complexity of Computing Shapley Values” on page 18-277.

Example: `'MaxNumSubsets',100`

Data Types: `single` | `double`

`Method` — Shapley value computation algorithm

`explainer.Method` (default) | `'interventional-kernel'` | `'conditional-kernel'`

Shapley value computation algorithm, specified as `'interventional-kernel'` or `'conditional-kernel'`.

- `'interventional-kernel'` — `fit` uses the kernelSHAP algorithm [1] with an interventional value function.
- `'conditional-kernel'` — `fit` uses the extension to the kernelSHAP algorithm [2] with a conditional value function.

For details about these algorithms, see “Shapley Value Computation Algorithms” on page 18-272.

Example: `'Method','conditional-kernel'`

Data Types: `char` | `string`

`UseParallel` — Flag to run in parallel
`false` (default) | `true`

Flag to run in parallel, specified as `true` or `false`. If you specify `'UseParallel',true`, the `fit` function executes for-loop iterations in parallel by using `parfor`. This option requires Parallel Computing Toolbox.

Example: `'UseParallel',true`

Data Types: `logical`

Output Arguments

`newExplainer` — Object explaining blackbox model
`shapley` object

Object explaining the blackbox model, returned as a `shapley` object. The `ShapleyValues` property of the object contains the computed Shapley values.

To overwrite the input argument `explainer`, assign the output of `fit` to `explainer`:

```
explainer = fit(explainer,queryPoint);
```

More About

Shapley Values

In game theory, the Shapley value of a player is the average marginal contribution of the player in a cooperative game. In the context of machine learning prediction, the Shapley value of a feature for a query point explains the contribution of the feature to a prediction (response for regression or score of each class for classification) at the specified query point.

The Shapley value corresponds to the deviation of the prediction for the query point from the average prediction, due to the feature. For a query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average.

For more details, see “Shapley Values for Machine Learning Model” on page 18-272.

References

[1] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[2] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` name-value argument to `true` in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`plot` | `shapley`

Topics

“Shapley Values for Machine Learning Model” on page 18-272
“Interpret Machine Learning Models” on page 18-256

Introduced in R2021a

fitcauto

Automatically select classification model with optimized hyperparameters

Syntax

```
Mdl = fitcauto(Tbl,ResponseVarName)
Mdl = fitcauto(Tbl,formula)
Mdl = fitcauto(Tbl,Y)

Mdl = fitcauto(X,Y)

Mdl = fitcauto(___,Name,Value)
[Mdl,OptimizationResults] = fitcauto(___)
```

Description

Given predictor and response data, `fitcauto` automatically tries a selection of classification model types with different hyperparameter values. The function uses Bayesian optimization to select models and their hyperparameter values, and computes the cross-validation classification error for each model. After the optimization is complete, `fitcauto` returns the model, trained on the entire data set, that is expected to best classify new data. You can use the `predict` and `loss` object functions of the returned model to classify new data and compute the test set classification error, respectively.

Use `fitcauto` when you are uncertain which classifier types best suit your data. For information on alternative methods for tuning hyperparameters of classification models, see “Alternative Functionality” on page 33-1571.

`Mdl = fitcauto(Tbl,ResponseVarName)` returns a classification model `Mdl` with tuned hyperparameters. The table `Tbl` contains the predictor variables and the response variable, where `ResponseVarName` is the name of the response variable.

`Mdl = fitcauto(Tbl,formula)` uses `formula` to specify the response variable and the predictor variables to consider among the variables in `Tbl`.

`Mdl = fitcauto(Tbl,Y)` uses the predictor variables in table `Tbl` and the class labels in vector `Y`.

`Mdl = fitcauto(X,Y)` uses the predictor variables in matrix `X` and the class labels in vector `Y`.

`Mdl = fitcauto(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, use the `HyperparameterOptimizationOptions` name-value pair argument to specify how the Bayesian optimization is performed.

`[Mdl,OptimizationResults] = fitcauto(___)` additionally returns `OptimizationResults`, a `BayesianOptimization` object containing the results of the model selection and hyperparameter tuning process.

Examples

Automatically Select Classifier Using Table Data

Use `fitcauto` to automatically select a classification model with optimized hyperparameters, given predictor and response data stored in a table.

Load Data

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Categorize the cars based on whether they were made in the USA.

```
Origin = categorical(cellstr(Origin));
Origin = mergecats(Origin,{'France','Japan','Germany', ...
    'Sweden','Italy','England'},'NotUSA');
```

Create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `Origin`.

```
cars = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,MPG,Weight,Origin);
```

Partition Data

Partition the data into training and test sets. Use approximately 80% of the observations for the model selection and hyperparameter tuning process, and 20% of the observations to test the performance of the final model returned by `fitcauto`. Use `cvpartition` to partition the data.

```
rng('default') % For reproducibility of the data partition
c = cvpartition(Origin,'Holdout',0.2);
trainingIdx = training(c); % Training set indices
carsTrain = cars(trainingIdx,:);
testIdx = test(c); % Test set indices
carsTest = cars(testIdx,:);
```

Run `fitcauto`

Pass the training data to `fitcauto`. By default, `fitcauto` determines appropriate model types to try, uses Bayesian optimization to find good hyperparameter values, and returns a trained model `Mdl` with the best expected performance. Additionally, `fitcauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-1568.

Expect this process to take some time. To speed up the optimization process, consider specifying to run the optimization in parallel, if you have a Parallel Computing Toolbox™ license. To do so, pass `'HyperparameterOptimizationOptions',struct('UseParallel',true)` to `fitcauto` as a name-value pair argument.

```
Mdl = fitcauto(carsTrain,'Origin');
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the

Learner types to explore: ensemble, knn, nb, svm, tree
Total iterations (MaxObjectiveEvaluations): 150
Total time (MaxTime): Inf

|=====================================================================================================
```

```
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|    1 | Best   |    0.14154 |            10.563 |         0.14154 |         0.14154 |     ensemb
|    2 | Accept |    0.18269 |           0.57392 |         0.14154 |         0.14154 |          t
|    3 | Accept |    0.23397 |            0.1264 |         0.14154 |         0.14154 |          t
|    4 | Accept |    0.16308 |            12.867 |         0.14154 |         0.15468 |     ensemb
|    5 | Accept |    0.20833 |             0.124 |         0.14154 |         0.15468 |          t
|    6 | Accept |    0.22115 |          0.079641 |         0.14154 |         0.15468 |          t
|    7 | Accept |    0.16923 |           0.20013 |         0.14154 |         0.15468 |         tr
|    8 | Accept |    0.37179 |           0.59222 |         0.14154 |         0.15468 |          s
|    9 | Accept |    0.37179 |           0.11659 |         0.14154 |         0.15468 |          s
|   10 | Accept |    0.24615 |           0.98386 |         0.14154 |         0.15468 |
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   11 | Accept |    0.16923 |          0.079996 |         0.14154 |         0.15468 |         tr
|   12 | Accept |    0.26923 |           0.10923 |         0.14154 |         0.15468 |          s
|   13 | Best   |    0.12923 |           0.11568 |         0.12923 |         0.15468 |         tr
|   14 | Accept |    0.21154 |          0.084406 |         0.12923 |         0.15468 |          t
|   15 | Accept |    0.14154 |          0.080022 |         0.12923 |         0.15294 |         tr
|   16 | Accept |    0.14769 |          0.092395 |         0.12923 |         0.15097 |         tr
|   17 | Accept |    0.14154 |            10.869 |         0.12923 |         0.14872 |     ensemb
|   18 | Accept |    0.37179 |           0.12386 |         0.12923 |         0.14872 |          s
|   19 | Accept |    0.22769 |           0.15545 |         0.12923 |         0.14872 |
|   20 | Accept |    0.22115 |          0.070813 |         0.12923 |         0.14872 |          t
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
```

```
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   21 | Accept |    0.37179 |           0.11553 |         0.12923 |         0.14872 |
|      |        |            |                   |                 |                 |

|   22 | Accept |    0.12923 |          0.080362 |         0.12923 |         0.14194 |

|   23 | Best   |    0.10154 |          0.079656 |         0.10154 |         0.13213 |

|   24 | Accept |    0.22769 |            0.2529 |         0.10154 |         0.13213 |
|      |        |            |                   |                 |                 |

|   25 | Accept |    0.11385 |          0.080085 |         0.10154 |          0.1289 |

|   26 | Accept |    0.13782 |          0.092228 |         0.10154 |          0.1289 |
|      |        |            |                   |                 |                 |

|   27 | Accept |    0.22769 |          0.073346 |         0.10154 |          0.1289 |
|      |        |            |                   |                 |                 |

|   28 | Accept |    0.21795 |          0.074914 |         0.10154 |          0.1289 |

|   29 | Accept |    0.24308 |           0.27621 |         0.10154 |          0.1289 |
|      |        |            |                   |                 |                 |

|   30 | Accept |    0.16308 |            12.328 |         0.10154 |          0.1289 |     ensemb
|      |        |            |                   |                 |                 |
|      |        |            |                   |                 |                 |

|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   31 | Accept |    0.24308 |           0.22334 |         0.10154 |          0.1289 |
|      |        |            |                   |                 |                 |

|   32 | Accept |    0.22115 |          0.066711 |         0.10154 |          0.1289 |

|   33 | Accept |    0.13846 |          0.079934 |         0.10154 |         0.12465 |

|   34 | Accept |    0.21474 |          0.085438 |         0.10154 |         0.12465 |

|   35 | Accept |    0.16615 |             10.05 |         0.10154 |         0.12465 |     ensemb
|      |        |            |                   |                 |                 |
|      |        |            |                   |                 |                 |

|   36 | Accept |    0.14154 |            12.866 |         0.10154 |         0.12465 |     ensemb
|      |        |            |                   |                 |                 |
|      |        |            |                   |                 |                 |

|   37 | Accept |    0.22769 |          0.070251 |         0.10154 |         0.12465 |
|      |        |            |                   |                 |                 |

|   38 | Accept |    0.37179 |          0.077924 |         0.10154 |         0.12465 |
|      |        |            |                   |                 |                 |

|   39 | Accept |    0.16923 |          0.068411 |         0.10154 |         0.12552 |

|   40 | Accept |    0.20833 |          0.064563 |         0.10154 |         0.12552 |

|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
```

```
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   41 | Accept |    0.16308 |            12.932 |         0.10154 |         0.12552 |     ensemb
|      |        |            |                   |                 |                 |
|   42 | Accept |    0.22462 |           0.24365 |         0.10154 |         0.12552 |
|      |        |            |                   |                 |                 |
|   43 | Accept |    0.20308 |            11.027 |         0.10154 |         0.12552 |     ensemb
|      |        |            |                   |                 |                 |
|   44 | Accept |    0.16923 |          0.069279 |         0.10154 |         0.12291 |          tr
|   45 | Accept |    0.22769 |          0.078716 |         0.10154 |         0.12291 |
|      |        |            |                   |                 |                 |
|   46 | Accept |    0.22769 |          0.068458 |         0.10154 |         0.12291 |
|      |        |            |                   |                 |                 |
|   47 | Accept |    0.16615 |            9.9849 |         0.10154 |         0.12291 |     ensemb
|      |        |            |                   |                 |                 |
|   48 | Accept |    0.14769 |            14.541 |         0.10154 |         0.12291 |     ensemb
|      |        |            |                   |                 |                 |
|   49 | Accept |    0.23077 |           0.21379 |         0.10154 |         0.12291 |
|      |        |            |                   |                 |                 |
|   50 | Accept |    0.37179 |          0.091937 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   51 | Accept |    0.21474 |          0.098808 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|   52 | Accept |    0.37179 |           0.10415 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|   53 | Accept |    0.25846 |            0.2444 |         0.10154 |         0.12291 |
|      |        |            |                   |                 |                 |
|   54 | Accept |    0.21154 |          0.074835 |         0.10154 |         0.12291 |          k
|   55 | Accept |    0.13846 |            12.173 |         0.10154 |         0.12291 |     ensemb
|      |        |            |                   |                 |                 |
|   56 | Accept |    0.36538 |           0.10958 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|   57 | Accept |    0.16615 |            11.482 |         0.10154 |         0.12291 |     ensemb
|      |        |            |                   |                 |                 |
```

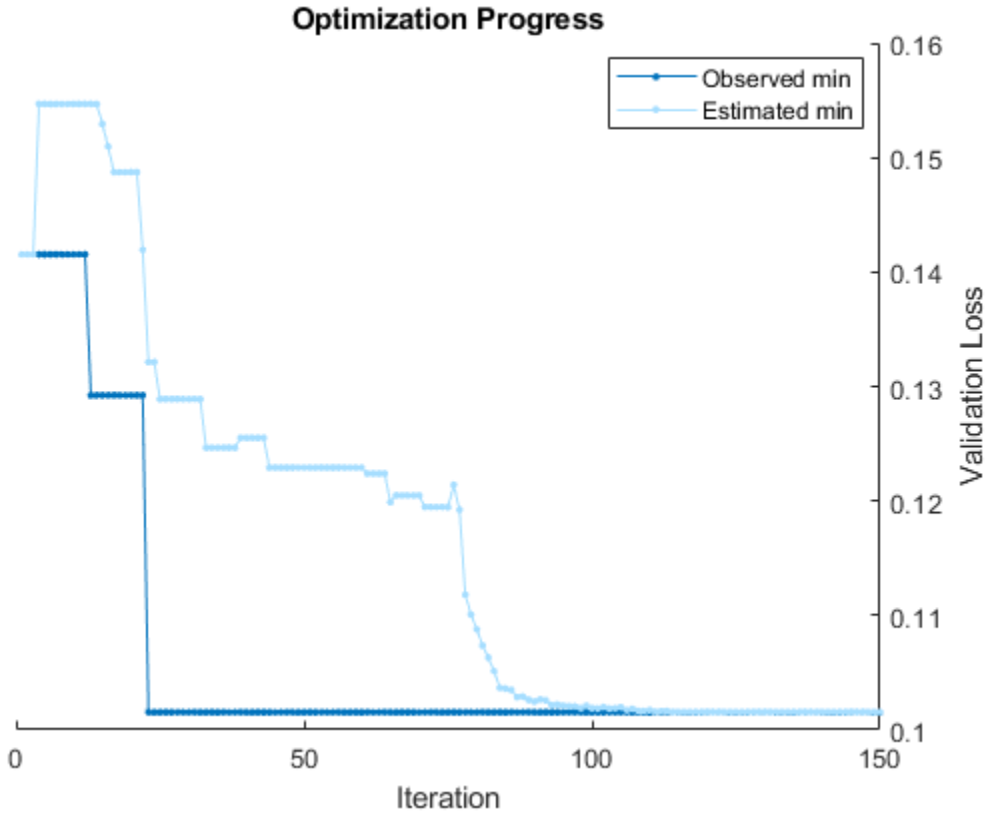
```
|   58 | Accept |    0.37179 |          0.095419 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|   59 | Accept |    0.37179 |          0.097469 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|   60 | Accept |    0.37179 |           0.11284 |         0.10154 |         0.12291 |          s
|      |        |            |                   |                 |                 |
|======|========|============|===================|=================|=================|=========
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======|========|============|===================|=================|=================|=========
|   61 | Accept |    0.11692 |          0.077515 |         0.10154 |         0.12239 |         tr
|   62 | Accept |    0.29167 |          0.091617 |         0.10154 |         0.12239 |          s
|      |        |            |                   |                 |                 |
|   63 | Accept |    0.21795 |          0.067171 |         0.10154 |         0.12239 |          k
|   64 | Accept |    0.18269 |          0.062887 |         0.10154 |         0.12239 |          k
|   65 | Accept |    0.12923 |          0.075704 |         0.10154 |         0.11989 |         tr
|   66 | Accept |    0.16923 |          0.065889 |         0.10154 |         0.12048 |         tr
|   67 | Accept |     0.1891 |          0.068215 |         0.10154 |         0.12048 |          k
|   68 | Accept |    0.13231 |            14.135 |         0.10154 |         0.12048 |     ensemb
|      |        |            |                   |                 |                 |
|   69 | Accept |    0.22769 |          0.060902 |         0.10154 |         0.12048 |
|      |        |            |                   |                 |                 |
|   70 | Accept |    0.37231 |           0.24511 |         0.10154 |         0.12048 |
|      |        |            |                   |                 |                 |
|======|========|============|===================|=================|=================|=========
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======|========|============|===================|=================|=================|=========
|   71 | Accept |    0.16923 |          0.069135 |         0.10154 |         0.11947 |         tr
|   72 | Accept |    0.22769 |          0.060552 |         0.10154 |         0.11947 |
|      |        |            |                   |                 |                 |
|   73 | Accept |    0.16308 |            10.133 |         0.10154 |         0.11947 |     ensemb
|      |        |            |                   |                 |                 |
|   74 | Accept |    0.13231 |            13.055 |         0.10154 |         0.11947 |     ensemb
|      |        |            |                   |                 |                 |
|   75 | Accept |    0.21474 |          0.069312 |         0.10154 |         0.11947 |          k
|   76 | Accept |    0.13846 |          0.083714 |         0.10154 |          0.1214 |         tr
|   77 | Accept |       0.12 |          0.069041 |         0.10154 |         0.11923 |         tr
|   78 | Accept |    0.10154 |          0.077399 |         0.10154 |          0.1118 |         tr
```

```
|   79 | Accept |       0.12 |          0.076269 |         0.10154 |         0.11007 |         tr
|   80 | Accept |    0.10154 |           0.09325 |         0.10154 |         0.10878 |         tr
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   81 | Accept |    0.10154 |          0.074541 |         0.10154 |         0.10737 |         tr
|   82 | Accept |       0.12 |          0.069929 |         0.10154 |          0.1063 |         tr
|   83 | Accept |    0.10154 |          0.072591 |         0.10154 |         0.10514 |         tr
|   84 | Accept |    0.10154 |          0.071254 |         0.10154 |         0.10366 |         tr
|   85 | Accept |    0.10154 |          0.077378 |         0.10154 |         0.10361 |         tr
|   86 | Accept |    0.10154 |          0.070643 |         0.10154 |         0.10348 |         tr
|   87 | Accept |       0.12 |          0.070551 |         0.10154 |         0.10286 |         tr
|   88 | Accept |       0.12 |          0.078438 |         0.10154 |          0.1029 |         tr
|   89 | Accept |    0.10154 |          0.072996 |         0.10154 |         0.10262 |         tr
|   90 | Accept |    0.10154 |          0.078162 |         0.10154 |         0.10246 |         tr
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|   91 | Accept |    0.10154 |           0.07038 |         0.10154 |         0.10267 |         tr
|   92 | Accept |    0.10154 |           0.07752 |         0.10154 |         0.10257 |         tr
|   93 | Accept |    0.10154 |          0.076155 |         0.10154 |         0.10217 |         tr
|   94 | Accept |    0.10154 |          0.075983 |         0.10154 |         0.10221 |         tr
|   95 | Accept |    0.10154 |           0.07407 |         0.10154 |         0.10211 |         tr
|   96 | Accept |    0.10154 |          0.080633 |         0.10154 |         0.10207 |         tr
|   97 | Accept |    0.10154 |          0.086164 |         0.10154 |         0.10205 |         tr
|   98 | Accept |    0.10154 |          0.080264 |         0.10154 |         0.10191 |         tr
|   99 | Accept |    0.12308 |          0.076015 |         0.10154 |          0.1021 |         tr
|  100 | Accept |    0.10154 |          0.074579 |         0.10154 |          0.1019 |         tr
|======================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================
|  101 | Accept |    0.10154 |          0.072077 |         0.10154 |         0.10186 |         tr
|  102 | Accept |    0.10154 |          0.071287 |         0.10154 |         0.10199 |         tr
|  103 | Accept |    0.10154 |          0.080003 |         0.10154 |         0.10186 |         tr
|  104 | Accept |       0.12 |           0.07462 |         0.10154 |         0.10189 |         tr
```

```
|  105 | Accept |    0.10154 |          0.077244 |         0.10154 |         0.10198 |         tr
|  106 | Accept |       0.12 |          0.080302 |         0.10154 |         0.10173 |         tr
|  107 | Accept |    0.10154 |          0.071858 |         0.10154 |         0.10183 |         tr
|  108 | Accept |    0.10154 |          0.076013 |         0.10154 |         0.10166 |         tr
|  109 | Accept |    0.10154 |          0.079106 |         0.10154 |         0.10164 |         tr
|  110 | Accept |    0.10154 |          0.075807 |         0.10154 |         0.10172 |         tr

|=====================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================
|  111 | Accept |    0.10154 |          0.076903 |         0.10154 |         0.10159 |         tr
|  112 | Accept |    0.10154 |          0.077409 |         0.10154 |         0.10163 |         tr
|  113 | Accept |    0.10154 |          0.069569 |         0.10154 |         0.10165 |         tr
|  114 | Accept |    0.12308 |          0.076744 |         0.10154 |         0.10156 |         tr
|  115 | Accept |    0.19551 |          0.097443 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  116 | Accept |    0.26603 |          0.099474 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  117 | Accept |    0.37179 |          0.077589 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  118 | Accept |    0.15385 |           0.08281 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  119 | Accept |    0.20833 |          0.074271 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  120 | Accept |    0.17949 |          0.081488 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|=====================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================
|  121 | Accept |    0.22115 |          0.076901 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  122 | Accept |    0.17308 |          0.076171 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  123 | Accept |      0.125 |           0.09187 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  124 | Accept |    0.13462 |           0.15746 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
|  125 | Accept |    0.13462 |          0.093366 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
```

```
|  126 | Accept |    0.22436 |           0.10124 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  127 | Accept |    0.15705 |            0.1133 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  128 | Accept |    0.16346 |           0.10388 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  129 | Accept |    0.13782 |          0.087555 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  130 | Accept |    0.12821 |          0.093318 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|=====================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================
|  131 | Accept |    0.13462 |           0.08648 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  132 | Accept |    0.14103 |          0.090667 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  133 | Accept |    0.12179 |          0.089208 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  134 | Accept |    0.21474 |          0.082634 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  135 | Accept |      0.125 |          0.088947 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  136 | Accept |    0.13141 |           0.07954 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  137 | Accept |    0.13782 |          0.091167 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  138 | Accept |      0.125 |          0.087338 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  139 | Accept |      0.125 |          0.080567 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  140 | Accept |    0.16667 |          0.083852 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|=====================================================================================================
| Iter | Eval   | Validation | Time for training | Observed min    | Estimated min   | Learner
|      | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================
|  141 | Accept |      0.125 |          0.089769 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  142 | Accept |    0.21474 |            0.1003 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |

|  143 | Accept |    0.13782 |            0.0833 |         0.10154 |         0.10156 |          s
|      |        |            |                   |                 |                 |
```

```
|  144 | Accept |    0.13462 |           0.10562 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  145 | Accept |    0.14103 |          0.076631 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  146 | Accept |    0.15385 |          0.082538 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  147 | Accept |    0.14423 |          0.077025 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  148 | Accept |    0.13782 |          0.089308 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  149 | Accept |    0.14103 |           0.08348 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
|  150 | Accept |    0.13782 |          0.079744 |         0.10154 |         0.10156 |
|      |        |            |                   |                 |                 |
```




```
__________________________________________________________
Optimization completed.
Total iterations: 150
Total elapsed time: 835.2167 seconds
Total time for training and validation: 193.4979 seconds

Best observed learner is a tree model with:
    MinLeafSize:             5
```

```
Observed validation loss: 0.10154
Time for training and validation: 0.079656 seconds

Best estimated learner (returned model) is a tree model with:
    MinLeafSize:              5
Estimated validation loss: 0.10156
Estimated time for training and validation: 0.076205 seconds

Documentation for fitcauto display
```

The final model returned by `fitcauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`carsTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

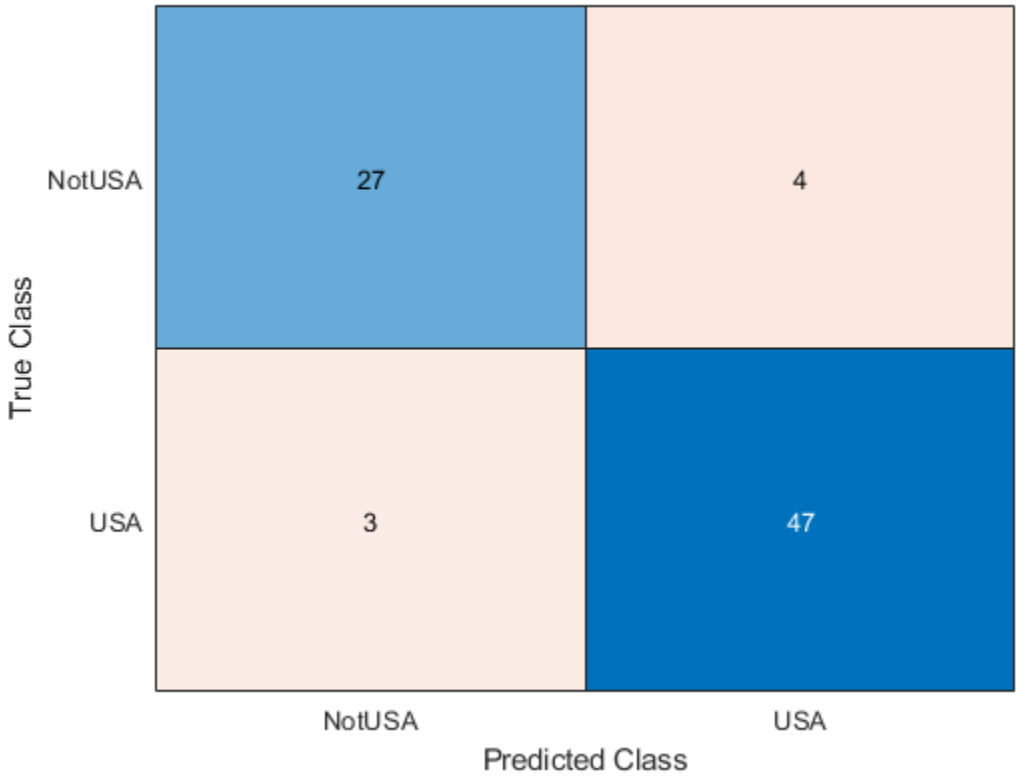
Evaluate Test Set Performance

Evaluate the performance of the model on the test set.

```
testAccuracy = 1 - loss(Mdl,carsTest,'Origin')
```

```
testAccuracy = 0.9143
```

```
confusionchart(carsTest.Origin,predict(Mdl,carsTest))
```



Automatically Select Classifier Using Matrix Data

Use `fitcauto` to automatically select a classification model with optimized hyperparameters, given predictor and response data stored in separate variables.

Load Data

Load the `humanactivity` data set. This data set contains 24,075 observations of five physical human activities: Sitting (1), Standing (2), Walking (3), Running (4), and Dancing (5). Each observation has 60 features extracted from acceleration data measured by smartphone accelerometer sensors. The variable `feat` contains the predictor data matrix of the 60 features for the 24,075 observations, and the response variable `actid` contains the activity IDs for the observations as integers.

```
load humanactivity
```

Partition Data

Partition the data into training and test sets. Use 90% of the observations to select a model, and 10% of the observations to validate the final model returned by `fitcauto`. Use `cvpartition` to reserve 10% of the observations for testing.

```
rng('default') % For reproducibility of the partition
c = cvpartition(actid,'Holdout',0.10);
trainingIndices = training(c); % Indices for the training set
XTrain = feat(trainingIndices,:);
YTrain = actid(trainingIndices);
testIndices = test(c); % Indices for the test set
XTest = feat(testIndices,:);
YTest = actid(testIndices);
```

Run `fitcauto`

Pass the training data to `fitcauto`. By default, `fitcauto` determines appropriate model (or learner) types to try, uses Bayesian optimization to find good hyperparameter values for those models, and returns a trained model with the best expected performance. Specify to run the optimization in parallel (requires Parallel Computing Toolbox™). Return a second output `OptimizationResults` that contains the details of the Bayesian optimization.

Expect this model selection process to take some time. By default, `fitcauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-1568.

```
options = struct('UseParallel',true);
[Mdl,OptimizationResults] = fitcauto(XTrain,YTrain,'HyperparameterOptimizationOptions',options);
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the

Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
Copying objective function to workers...
Done copying objective function to workers.

Learner types to explore: ensemble, knn, nb, svm, tree
Total iterations (MaxObjectiveEvaluations): 150
Total time (MaxTime): Inf

|=====================================================================================================
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
```

```
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|    1 |       6 | Best   |    0.28088 |            48.752 |         0.28088 |         0.28088 |
|    2 |       6 | Best   |   0.036459 |            51.455 |        0.036459 |        0.036459 |
|    3 |       6 | Best   |   0.025845 |            5.1379 |        0.025845 |        0.025845 |
|    4 |       6 | Best   |   0.006415 |            60.999 |        0.006415 |        0.021738 |
|    5 |       6 | Accept |   0.025845 |            4.8823 |        0.006415 |        0.021738 |
|    6 |       6 | Accept |   0.017768 |            5.4601 |        0.006415 |        0.021738 |
|    7 |       6 | Accept |   0.050212 |            1.1024 |        0.006415 |        0.021738 |
|    8 |       6 | Accept |   0.050212 |           0.64183 |        0.006415 |        0.021738 |
|    9 |       6 | Accept |   0.019568 |            151.32 |        0.006415 |        0.021152 |
|   10 |       6 | Accept |   0.026537 |            165.42 |        0.006415 |        0.022035 |
|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   11 |       6 | Accept |    0.59166 |            29.107 |        0.006415 |        0.022035 |
|   12 |       6 | Accept |   0.021645 |            50.626 |        0.006415 |        0.022122 |
|   13 |       6 | Accept |   0.043567 |            24.583 |        0.006415 |        0.022122 |
|   14 |       6 | Accept |   0.028844 |            22.503 |        0.006415 |        0.022122 |
|   15 |       6 | Accept |    0.04389 |            157.12 |        0.006415 |        0.022122 |
|   16 |       6 | Accept |   0.024598 |            20.721 |        0.006415 |        0.022122 |
|   17 |       6 | Accept |    0.03009 |            20.969 |        0.006415 |        0.022122 |
```

```
|   18 |       6 | Accept |   0.016753 |            6.5065 |        0.006415 |        0.021547 |

|   19 |       6 | Accept |   0.040059 |            4.3496 |        0.006415 |        0.022122 |

|   20 |       6 | Accept |   0.060319 |            2.2673 |        0.006415 |        0.022122 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   21 |       6 | Accept |   0.050212 |           0.97596 |        0.006415 |        0.022122 |
|      |         |        |            |                   |                 |                 |

|   22 |       6 | Accept |   0.036552 |            20.775 |        0.006415 |        0.022122 |

|   23 |       6 | Accept |   0.050212 |           0.55594 |        0.006415 |        0.022122 |
|      |         |        |            |                   |                 |                 |

|   24 |       6 | Accept |    0.11076 |            36.229 |        0.006415 |        0.022122 |

|   25 |       6 | Accept |    0.27884 |            66.582 |        0.006415 |        0.024532 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |

|   26 |       6 | Accept |    0.58127 |            31.861 |        0.006415 |        0.024532 |
|      |         |        |            |                   |                 |                 |

|   27 |       6 | Accept |    0.01583 |            5.7511 |        0.006415 |        0.020656 |

|   28 |       6 | Accept |   0.069319 |             1.806 |        0.006415 |         0.02077 |

|   29 |       6 | Accept |    0.59166 |            352.55 |        0.006415 |         0.02077 |
|      |         |        |            |                   |                 |                 |

|   30 |       6 | Accept |   0.043336 |            3.7865 |        0.006415 |        0.020133 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   31 |       6 | Accept |    0.10555 |            34.294 |        0.006415 |        0.020133 |

|   32 |       6 | Accept |   0.021276 |             4.582 |        0.006415 |        0.018661 |

|   33 |       5 | Accept |   0.030829 |            159.57 |        0.006415 |        0.018642 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|   34 |       5 | Accept |   0.014307 |            49.903 |        0.006415 |        0.018642 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |

|   35 |       5 | Accept |   0.050212 |             1.025 |        0.006415 |        0.018642 |
|      |         |        |            |                   |                 |                 |

|   36 |       6 | Accept |    0.74165 |            24.131 |        0.006415 |        0.018661 |
|      |         |        |            |                   |                 |                 |
```

```
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   37 |       6 | Accept |     0.4226 |           558.84 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|   38 |       6 | Accept |    0.57615 |           317.96 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   39 |       6 | Accept |    0.59166 |            26.64 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|   40 |       6 | Accept |   0.087087 |           30.271 |        0.006415 |        0.018661 |

|=========================================================================================================
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=========================================================================================================
|   41 |       6 | Accept |    0.73985 |           551.47 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|   42 |       4 | Accept |   0.025983 |           146.08 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   43 |       4 | Accept |    0.02566 |           146.73 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   44 |       4 | Accept |   0.024922 |           124.65 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   45 |       6 | Accept |   0.025891 |           23.638 |        0.006415 |        0.018661 |
|   46 |       5 | Accept |   0.025891 |           24.038 |        0.006415 |        0.018661 |
|   47 |       5 | Accept |   0.025891 |           23.646 |        0.006415 |        0.018661 |
|   48 |       6 | Accept |    0.03009 |           188.48 |        0.006415 |        0.018661 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   49 |       6 | Accept |   0.017214 |           6.0889 |        0.006415 |        0.017935 |
|   50 |       6 | Accept |    0.01726 |           5.6027 |        0.006415 |        0.017303 |

|=========================================================================================================
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=========================================================================================================
|   51 |       6 | Accept |   0.037244 |            158.2 |        0.006415 |        0.017303 |
|      |         |        |            |                  |                |                |
|   52 |       6 | Accept |   0.046474 |           190.01 |        0.006415 |        0.017303 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
```

```
|   53 |       6 | Accept |   0.032398 |           155.47 |        0.006415 |        0.017303 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|   54 |       6 | Accept |   0.054135 |           3.0156 |        0.006415 |        0.017093 |

|   55 |       6 | Accept |   0.049797 |           28.421 |        0.006415 |        0.017093 |

|   56 |       6 | Accept |   0.046566 |           27.524 |        0.006415 |        0.017093 |

|   57 |       6 | Accept |    0.36307 |           711.57 |        0.006415 |        0.017093 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|   58 |       6 | Accept |   0.022706 |           23.417 |        0.006415 |        0.017093 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|   59 |       6 | Accept |   0.028798 |           4.1232 |        0.006415 |         0.01733 |

|   60 |       6 | Accept |   0.041351 |           28.037 |        0.006415 |         0.01733 |

|======|=========|========|============|==================|=================|=================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======|=========|========|============|==================|=================|=================|
|   61 |       6 | Accept |   0.030044 |           26.265 |        0.006415 |         0.01733 |

|   62 |       6 | Accept |    0.11838 |           284.42 |        0.006415 |         0.01733 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|   63 |       6 | Accept |    0.47116 |           54.038 |        0.006415 |         0.01733 |
|      |         |        |            |                  |                 |                 |

|   64 |       6 | Accept |    0.26574 |           104.86 |        0.006415 |         0.01733 |
|      |         |        |            |                  |                 |                 |

|   65 |       6 | Accept |   0.050212 |          0.72243 |        0.006415 |         0.01733 |
|      |         |        |            |                  |                 |                 |

|   66 |       6 | Accept |    0.01726 |           5.6362 |        0.006415 |        0.016846 |

|   67 |       6 | Accept |   0.077072 |           268.17 |        0.006415 |        0.016846 |
|      |         |        |            |                  |                 |                 |

|   68 |       6 | Accept |   0.031613 |           25.909 |        0.006415 |        0.016846 |

|   69 |       6 | Accept |    0.02003 |           92.379 |        0.006415 |        0.016846 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|   70 |       6 | Accept |     0.1145 |           92.795 |        0.006415 |        0.016846 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |
|      |         |        |            |                  |                 |                 |

|======|=========|========|============|==================|=================|=================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
```

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
71	6	Accept	0.012968	53.299	0.006415	0.016846
72	6	Accept	0.011999	53.54	0.006415	0.016846
73	6	Accept	0.011999	50.887	0.006415	0.012076
74	6	Accept	0.039921	40.808	0.006415	0.012733
75	6	Accept	0.04232	98.684	0.006415	0.012754
76	6	Accept	0.094702	397.88	0.006415	0.012754
77	6	Accept	0.065165	44.801	0.006415	0.013179
78	6	Accept	0.22503	164.64	0.006415	0.013179
79	6	Accept	0.03069	44.42	0.006415	0.013027
80	6	Accept	0.11459	71.56	0.006415	0.012064
81	6	Accept	0.044582	41.462	0.006415	0.013176

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
82	6	Accept	0.014722	223.7	0.006415	0.013176
83	6	Accept	0.018322	23.34	0.006415	0.013176
84	6	Accept	0.016984	28.833	0.006415	0.013176
85	6	Accept	0.59166	2135.3	0.006415	0.013176
86	6	Accept	0.012461	39.005	0.006415	0.013176
87	6	Accept	0.016753	22.649	0.006415	0.013176
88	6	Accept	0.016845	28.292	0.006415	0.013176
89	6	Accept	0.016799	20.771	0.006415	0.013176
90	6	Accept	0.041259	175.27	0.006415	0.013176
91	6	Accept	0.017953	19.778	0.006415	0.013176
92	6	Accept	0.019568	19.245	0.006415	0.013176
93	6	Accept	0.016061	19.511	0.006415	0.013176
94	6	Accept	0.11847	149.5	0.006415	0.013176
95	6	Accept	0.017953	18.428	0.006415	0.013176

```
|   96 |       6 | Accept |   0.025106 |            25.204 |        0.006415 |        0.013176 |
|   97 |       6 | Accept |   0.011676 |            70.358 |        0.006415 |        0.012446 |
|   98 |       6 | Accept |   0.031983 |            37.179 |        0.006415 |        0.012446 |
|   99 |       6 | Accept |  0.0097379 |            76.639 |        0.006415 |        0.011402 |
|  100 |       6 | Accept |    0.22416 |            684.65 |        0.006415 |        0.011402 |
|======|=========|========|============|===================|=================|=================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======|=========|========|============|===================|=================|=================|
|  101 |       6 | Accept |    0.11478 |            87.263 |        0.006415 |        0.011913 |
|  102 |       6 | Accept |  0.0081687 |            75.177 |        0.006415 |         0.01045 |
|  103 |       6 | Accept |   0.010753 |             70.01 |        0.006415 |        0.010258 |
|  104 |       6 | Accept |    0.36076 |                77 |        0.006415 |        0.010258 |
|  105 |       6 | Accept |  0.0084456 |            62.748 |        0.006415 |       0.0092051 |
|  106 |       6 | Accept |   0.012738 |             32.34 |        0.006415 |       0.0092051 |
|  107 |       6 | Accept |   0.031521 |            41.482 |        0.006415 |       0.0092051 |
|  108 |       6 | Accept |   0.021368 |            38.604 |        0.006415 |       0.0092051 |
```

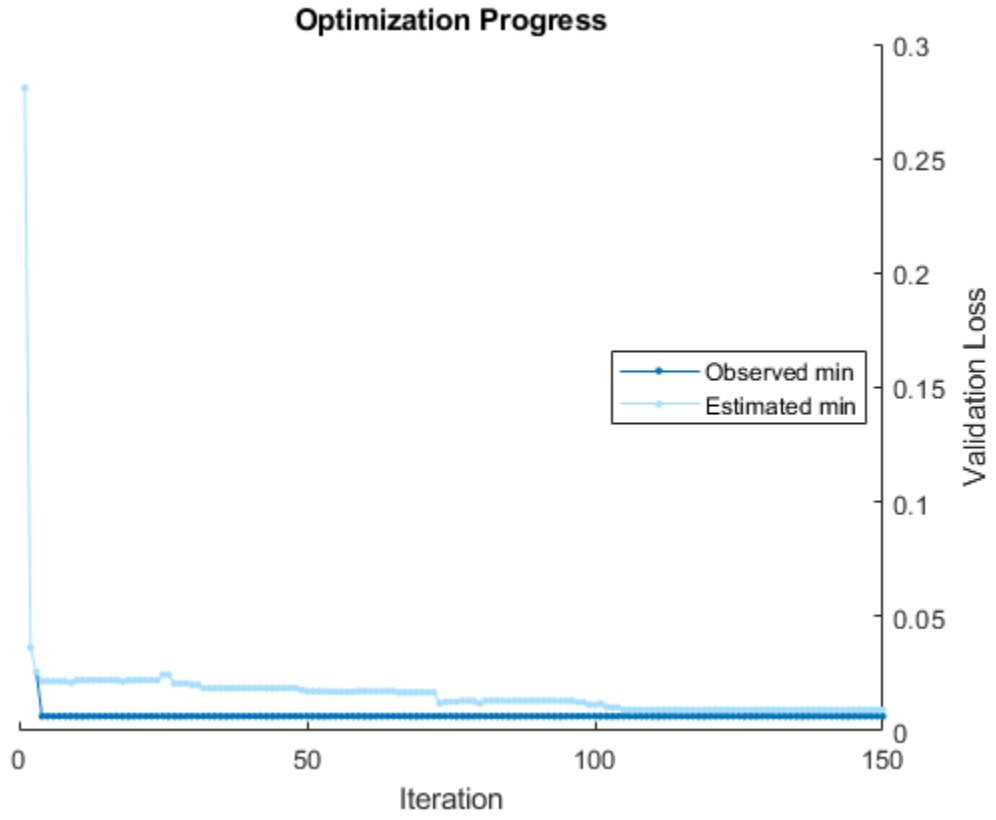
Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
109	6	Accept	0.050212	0.59557	0.006415	0.0092051
110	6	Accept	0.57883	127.15	0.006415	0.0092051

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
111	6	Accept	0.050212	0.685	0.006415	0.0092051
112	6	Accept	0.029583	33.504	0.006415	0.0092051
113	6	Accept	0.012138	41.159	0.006415	0.0092051
114	6	Accept	0.030921	38.482	0.006415	0.0092051
115	6	Accept	0.1295	52.128	0.006415	0.0092051
116	6	Accept	0.028475	33.553	0.006415	0.0092051
117	6	Accept	0.050212	0.62316	0.006415	0.0092051
118	6	Accept	0.030737	37	0.006415	0.0092051
119	6	Accept	0.021229	36.97	0.006415	0.0092051
120	6	Accept	0.015184	24.004	0.006415	0.0092051

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
121	6	Accept	0.06955	101.38	0.006415	0.0092051
122	6	Accept	0.028291	32.637	0.006415	0.0092051

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
123	6	Accept	0.041951	51.65	0.006415	0.0092051
124	6	Accept	0.039921	216.38	0.006415	0.0092051
125	6	Accept	0.048828	91.888	0.006415	0.0092051
126	6	Accept	0.030644	38.563	0.006415	0.0092051
127	6	Accept	0.029537	37.409	0.006415	0.0092051
128	6	Accept	0.02949	31.222	0.006415	0.0092051
129	6	Accept	0.029537	130.19	0.006415	0.0092051
130	6	Accept	0.094702	174.14	0.006415	0.0092051
131	6	Accept	0.037382	46.207	0.006415	0.0092051
132	6	Accept	0.022152	40.588	0.006415	0.0092051
133	6	Accept	0.032629	32.651	0.006415	0.0092051
134	6	Accept	0.02806	131.12	0.006415	0.0092051
135	6	Accept	0.59166	3480	0.006415	0.0092051
136	6	Accept	0.59166	3481.3	0.006415	0.0092051

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
137	6	Accept	0.032121	44.172	0.006415	0.0092051
138	6	Accept	0.30755	71.9	0.006415	0.0092051
139	6	Accept	0.040936	233.89	0.006415	0.0092051
140	5	Accept	0.040521	224.72	0.006415	0.0092051
141	5	Accept	0.22208	64.659	0.006415	0.0092051
142	6	Accept	0.029444	30.014	0.006415	0.0092051
143	6	Accept	0.046013	49.025	0.006415	0.0092051
144	6	Accept	0.013291	42.726	0.006415	0.0092051
145	6	Accept	0.59166	3493.5	0.006415	0.0092051
146	6	Accept	0.089533	189.71	0.006415	0.0092051
147	6	Accept	0.03369	49.586	0.006415	0.0092051
148	6	Accept	0.013384	42.876	0.006415	0.0092051
149	6	Accept	0.027368	117.6	0.006415	0.0092051
150	6	Accept	0.041674	234.61	0.006415	0.0092051




```
Optimization completed.
Total iterations: 150
Total elapsed time: 4534.2478 seconds
Total time for training and validation: 24883.8563 seconds

Best observed learner is an ensemble model with:
    Method:          AdaBoostM2
    NumLearningCycles:      214
    MinLeafSize:              5
    MaxNumSplits:            23
Observed validation loss: 0.006415
Time for training and validation: 60.9987 seconds

Best estimated learner (returned model) is an ensemble model with:
    Method:          AdaBoostM2
    NumLearningCycles:      214
    MinLeafSize:              3
    MaxNumSplits:            16
Estimated validation loss: 0.0092051
Estimated time for training and validation: 57.8146 seconds

Documentation for fitcauto display
```

The final model returned by `fitcauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`XTrain` and `YTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

Evaluate the final model performance on the test data set.

```
testAccuracy = 1 - loss(Mdl,XTest,YTest)
```

```
testAccuracy = 0.9917
```

The final model correctly classifies over 99% of the observations.

Combine Feature Selection and Automated Classifier Selection

Use `fitcauto` to automatically select a classification model with optimized hyperparameters, given predictor and response data stored in a table. Before passing data to `fitcauto`, perform feature selection to remove unimportant predictors from the data set.

Load and Partition Data

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable('CreditRating_Historical.dat');
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____     _____     _______    ________    _____    ________    _______

    62394     0.013     0.104     0.036       0.447      0.142        3       {'BB' }
    48608     0.232     0.335     0.062       1.969      0.281        8       {'A'  }
    42444     0.311     0.367     0.074       1.935      0.366        1       {'A'  }
    48631     0.194     0.263     0.062       1.017      0.228        4       {'BBB'}
    43768     0.121     0.413     0.057       3.647      0.466       12       {'AAA'}
    39255    -0.117    -0.799      0.01       0.179      0.082        4       {'CCC'}
    62236     0.087     0.158     0.049       0.816      0.324        2       {'BBB'}
    39354     0.005     0.181     0.034       2.597      0.388        7       {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,'ID');
creditrating.Industry = categorical(creditrating.Industry);
```

Partition the data into training and test sets. Use approximately 85% of the observations for the model selection and hyperparameter tuning process, and 15% of the observations to test the performance of the final model returned by `fitcauto` on new data. Use `cvpartition` to partition the data.

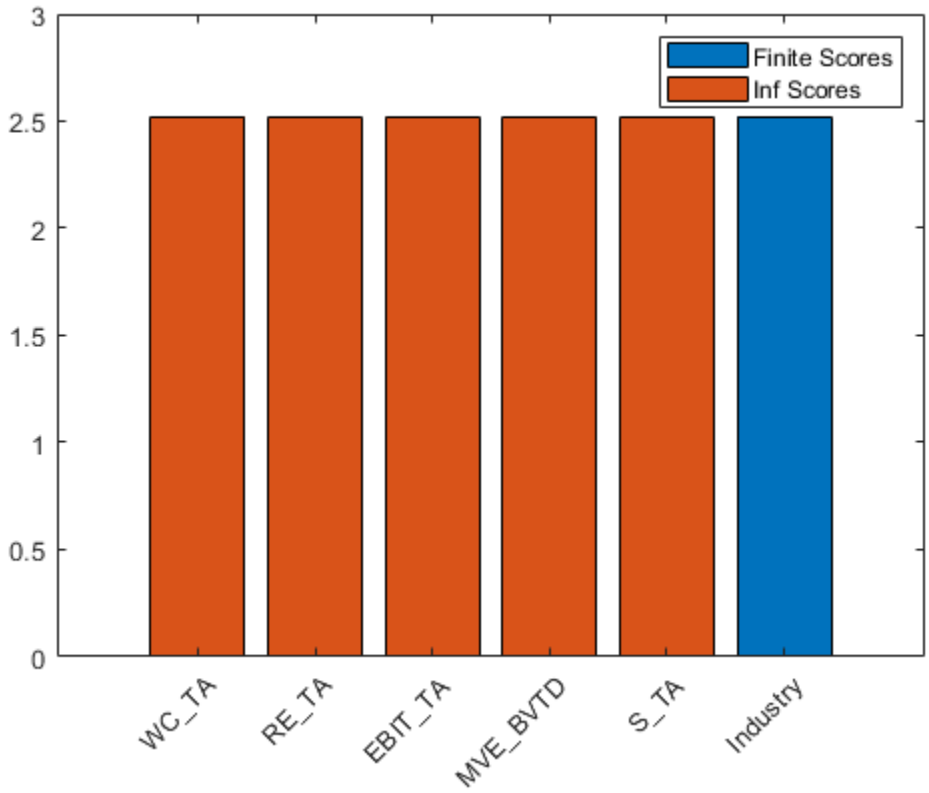
```
rng('default') % For reproducibility of the partition
c = cvpartition(creditrating.Rating,'Holdout',0.15);
trainingIndices = training(c); % Indices for the training set
```

```
testIndices = test(c); % Indices for the test set
creditTrain = creditrating(trainingIndices,:);
creditTest = creditrating(testIndices,:);
```

Perform Feature Selection

Before passing the training data to `fitcauto`, find the important predictors by using the `fscchi2` function. Visualize the predictor scores by using the `bar` function. Because some scores can be `Inf`, and `bar` discards `Inf` values, plot the finite scores first and then plot a finite representation of the `Inf` scores in a different color.

```
[idx,scores] = fscchi2(creditTrain,'Rating');
bar(scores(idx)) % Represents finite scores
hold on
veryImportant = isinf(scores);
finiteMax = max(scores(~veryImportant));
bar(finiteMax*veryImportant(idx)) % Represents Inf scores
hold off
xticklabels(strrep(creditTrain.Properties.VariableNames(idx),'_','\_'))
xtickangle(45)
legend({'Finite Scores','Inf Scores'})
```



Notice that the `Industry` predictor has a low score corresponding to a *p*-value that is greater than 0.05, which indicates that `Industry` might not be an important feature. Remove the `Industry` feature from the training and test data sets.

```
creditTrain = removevars(creditTrain,'Industry');
creditTest = removevars(creditTest,'Industry');
```

Run `fitcauto`

Pass the training data to `fitcauto`. The function uses Bayesian optimization to select models and their hyperparameter values, and returns a trained model `Mdl` with the best expected performance. Specify to try all available learner types and run the optimization in parallel (requires Parallel Computing Toolbox™). Return a second output `Results` that contains the details of the Bayesian optimization.

Expect this process to take some time. By default, `fitcauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-1568.

```
options = struct('UseParallel',true);
[Mdl,Results] = fitcauto(creditTrain,'Rating', ...
    'Learners','all','HyperparameterOptimizationOptions',options);
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the

Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

Copying objective function to workers...
Done copying objective function to workers.

Learner types to explore: discr, ensemble, kernel, knn, linear, nb, svm, tree
Total iterations (MaxObjectiveEvaluations): 240
Total time (MaxTime): Inf

|=====================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|=====================================================================================================|
|    1 |       6 | Best   |    0.42716 |            3.0379 |         0.42716 |         0.42716 |
|      |         |        |            |                   |                 |                 |
|    2 |       4 | Accept |    0.74185 |            4.7899 |         0.24948 |         0.29794 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    3 |       4 | Best   |    0.24948 |            5.0813 |         0.24948 |         0.29794 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    4 |       4 | Accept |    0.29794 |            3.7295 |         0.24948 |         0.29794 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    5 |       3 | Accept |    0.25097 |            9.2655 |         0.24948 |         0.25067 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    6 |       3 | Accept |    0.25067 |           0.81139 |         0.24948 |         0.25067 |
|      |         |        |            |                   |                 |                 |
|    7 |       6 | Accept |    0.52917 |            2.3362 |         0.24948 |         0.25067 |
|      |         |        |            |                   |                 |                 |
|      |         |        |            |                   |                 |                 |
|    8 |       3 | Accept |    0.55818 |           0.63908 |         0.24948 |         0.25067 |
|      |         |        |            |                   |                 |                 |
|    9 |       3 | Accept |     0.3781 |            1.6777 |         0.24948 |         0.25067 |
```

```
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   10 |       3 | Accept |    0.43225 |          0.80766 |        0.24948 |        0.25067 |
|      |         |        |            |                  |                |                |
|==========================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min   | Estimated min  |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|==========================================================================================|
|   11 |       3 | Accept |    0.47712 |           3.3756 |        0.24948 |        0.25067 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   12 |       6 | Accept |    0.25695 |           2.3709 |        0.24948 |        0.25067 |
|      |         |        |            |                  |                |                |
|   13 |       4 | Accept |    0.26413 |          0.49941 |        0.24379 |        0.25067 |
|   14 |       4 | Accept |    0.42327 |          0.85101 |        0.24379 |        0.25067 |
|      |         |        |            |                  |                |                |
|   15 |       4 | Best   |    0.24379 |           1.8084 |        0.24379 |        0.25067 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   16 |       3 | Accept |    0.81544 |           4.9586 |        0.24379 |        0.25067 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   17 |       3 | Accept |    0.45169 |          0.96723 |        0.24379 |        0.25067 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   18 |       6 | Accept |    0.33712 |          0.19972 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|   19 |       3 | Accept |     0.4834 |          0.38951 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|   20 |       3 | Accept |    0.46336 |          0.78881 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|==========================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min   | Estimated min  |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|==========================================================================================|
|   21 |       3 | Accept |    0.82082 |           2.9223 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   22 |       3 | Accept |    0.61292 |           1.8557 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   23 |       6 | Accept |    0.43255 |          0.68689 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|   24 |       4 | Accept |    0.28866 |           1.9017 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   25 |       4 | Accept |    0.74185 |           1.3546 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
|   26 |       4 | Accept |    0.42746 |          0.69002 |        0.24379 |        0.25695 |
|      |         |        |            |                  |                |                |
```

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
27	3	Accept	0.25606	12.143	0.24379	0.25498
28	3	Accept	0.25366	2.4732	0.24379	0.25498
29	6	Accept	0.66796	0.22938	0.24379	0.25498
30	4	Accept	0.69488	1.9352	0.242	0.25498
31	4	Best	0.242	1.9467	0.242	0.25498
32	4	Accept	0.32306	0.48104	0.242	0.25498
33	4	Accept	0.43225	0.10463	0.242	0.25498
34	4	Accept	0.32994	0.2065	0.242	0.25498
35	6	Accept	0.53814	2.8748	0.242	0.25498
36	4	Accept	0.24529	105.62	0.242	0.25498
37	4	Accept	0.53814	3.3272	0.242	0.25498
38	4	Accept	0.53814	3.8593	0.242	0.25498
39	4	Accept	0.25965	14.211	0.242	0.25498
40	5	Accept	0.42656	0.16731	0.242	0.25498

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
41	5	Accept	0.42656	0.11476	0.242	0.25498
42	4	Accept	0.2767	21.389	0.242	0.25498

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
43	4	Accept	0.29973	0.1848	0.242	0.25498
44	4	Accept	0.25935	20.084	0.242	0.25498
45	3	Accept	0.24499	6.5071	0.242	0.26328
46	3	Accept	0.28059	0.19378	0.242	0.26328
47	6	Accept	0.27281	0.13181	0.242	0.26328
48	3	Accept	0.2429	1.5474	0.242	0.26328
49	3	Accept	0.2423	1.5219	0.242	0.26328
50	3	Accept	0.67125	0.68464	0.242	0.26328

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
51	3	Accept	0.46485	0.8417	0.242	0.26328
52	6	Accept	0.63147	3.1926	0.242	0.26328
53	4	Accept	0.37571	0.11086	0.242	0.26706
54	4	Accept	0.29136	0.43908	0.242	0.26706
55	4	Accept	0.28059	0.50642	0.242	0.26706
56	3	Accept	0.36375	5.2617	0.242	0.26706
57	3	Accept	0.27251	0.13425	0.242	0.26706
58	6	Accept	0.43225	0.083255	0.242	0.26706
59	3	Accept	0.28059	0.10921	0.242	0.26106
60	3	Accept	0.42537	0.15885	0.242	0.26106

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
61	3	Accept	0.81484	2.9209	0.242	0.26106

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
62	3	Accept	0.24948	2.4493	0.242	0.26106
63	6	Accept	0.68531	0.40944	0.242	0.26106
64	3	Accept	0.32426	9.8071	0.242	0.26007
65	3	Accept	0.55369	0.69919	0.242	0.26007
66	3	Accept	0.24319	1.5742	0.242	0.26007
67	3	Accept	0.70894	1.9822	0.242	0.26007
68	6	Accept	0.46485	0.17082	0.242	0.26007
69	5	Accept	0.74185	2.1172	0.242	0.26007
70	5	Accept	0.83548	6.3585	0.242	0.26007
71	3	Accept	0.25905	16.096	0.242	0.26007
72	3	Accept	0.28448	13.826	0.242	0.26007
73	3	Accept	0.25695	0.24642	0.242	0.26007
74	6	Accept	0.32396	1.8799	0.242	0.26007
75	4	Accept	0.32456	0.20481	0.242	0.26007
76	4	Accept	0.32994	0.35799	0.242	0.26007
77	4	Accept	0.26054	4.1734	0.242	0.26007
78	4	Accept	0.43703	0.92964	0.242	0.24831
79	4	Accept	0.31588	17.116	0.242	0.24831

```
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   80 |       3 | Accept |    0.25277 |           7.8173 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min  |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   81 |       3 | Accept |    0.43015 |          0.10239 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|   82 |       6 | Accept |    0.42208 |         0.096953 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|   83 |       3 | Accept |    0.52617 |           2.3915 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   84 |       3 | Accept |    0.43344 |           1.2875 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|   85 |       3 | Accept |    0.30093 |           1.3003 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   86 |       3 | Accept |    0.42267 |          0.64506 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|   87 |       6 | Accept |    0.32905 |          0.26891 |          0.242 |        0.24831 |
|      |         |        |            |                  |                |                |
|   88 |       4 | Accept |    0.24349 |           1.9282 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   89 |       4 | Accept |    0.24499 |           1.6543 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   90 |       4 | Accept |    0.24469 |           2.1463 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min  |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|   91 |       4 | Accept |    0.28059 |          0.15481 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|   92 |       5 | Accept |    0.28059 |          0.38138 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|   93 |       5 | Accept |    0.28059 |          0.15804 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|   94 |       4 | Accept |    0.29674 |           11.234 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|      |         |        |            |                  |                |                |
|   95 |       4 | Accept |    0.28059 |          0.12279 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
|   96 |       4 | Accept |    0.29704 |           1.1016 |          0.242 |        0.24684 |
|      |         |        |            |                  |                |                |
```

97	4	Accept	0.28059	0.099833	0.242	0.24684
98	4	Best	0.2411	1.3634	0.2411	0.24471
99	4	Accept	0.74185	0.10441	0.2411	0.24471
100	4	Accept	0.30093	0.11043	0.2411	0.24471

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
101	4	Accept	0.46844	0.17182	0.2411	0.24471
102	4	Accept	0.25426	3.1634	0.2411	0.24471
103	4	Accept	0.24469	1.6558	0.2411	0.24558
104	4	Accept	0.43823	3.6572	0.2411	0.24558
105	4	Accept	0.31977	1.3423	0.2411	0.24556
106	6	Accept	0.24678	2.0849	0.2411	0.24529
107	5	Accept	0.24678	1.9627	0.2411	0.24441
108	5	Accept	0.24678	2.3413	0.2411	0.24441
109	4	Accept	0.29435	14.698	0.2411	0.24441
110	4	Accept	0.37152	0.098701	0.2411	0.24441

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss

```
|  111 |       4 | Accept |    0.25666 |           5.7386 |          0.2411 |         0.24441 |
|  112 |       4 | Accept |    0.28059 |          0.10645 |          0.2411 |         0.24441 |
|  113 |       4 | Accept |    0.28059 |          0.10515 |          0.2411 |         0.24441 |
|  114 |       6 | Accept |    0.74185 |           2.7447 |          0.2411 |         0.24441 |
|  115 |       4 | Accept |    0.78552 |           10.495 |          0.2411 |         0.24441 |
|  116 |       4 | Accept |    0.74185 |           2.7544 |          0.2411 |         0.24441 |
|  117 |       4 | Accept |    0.74185 |           2.6565 |          0.2411 |         0.24441 |
|  118 |       4 | Accept |    0.45797 |           11.614 |          0.2411 |         0.24441 |
|  119 |       4 | Accept |    0.65271 |           1.6037 |          0.2411 |         0.24441 |
|  120 |       4 | Accept |     0.3108 |           1.4953 |          0.2411 |         0.24441 |

|======================================================================================================|
| Iter | Active  | Eval   | Validation | Time for training | Observed min    | Estimated min   |
|      | workers | result | loss       | & validation (sec)| validation loss | validation loss |
|======================================================================================================|
|  121 |       4 | Accept |     0.2414 |           1.2954 |          0.2411 |         0.24441 |
|  122 |       4 | Accept |     0.5157 |           9.6657 |          0.2411 |         0.24441 |
|  123 |       4 | Accept |    0.53186 |           390.32 |          0.2402 |         0.24441 |
|  124 |       4 | Best   |     0.2402 |           1.4571 |          0.2402 |         0.24441 |
|  125 |       4 | Accept |     0.2402 |           1.3713 |          0.2402 |         0.24441 |
|  126 |       3 | Accept |    0.24529 |           39.503 |          0.2402 |         0.24441 |
|  127 |       3 | Accept |    0.25576 |           1.4031 |          0.2402 |         0.24441 |
```

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
128	6	Accept	0.2426	1.3555	0.2402	0.24229
129	3	Accept	0.24499	9.5236	0.2402	0.24419
130	3	Accept	0.28059	0.12828	0.2402	0.24419
131	3	Accept	0.6886	1.4213	0.2402	0.24419
132	3	Accept	0.4487	3.111	0.2402	0.24419
133	6	Accept	0.24469	8.3042	0.2402	0.24441
134	3	Accept	0.26204	1.7109	0.2402	0.24441
135	3	Accept	0.74185	4.0388	0.2402	0.24441
136	3	Accept	0.29734	9.9598	0.2402	0.24441
137	3	Accept	0.44391	2.0399	0.2402	0.24441
138	6	Accept	0.24559	3.2102	0.2402	0.24276
139	4	Accept	0.25067	1.4995	0.2402	0.24276
140	4	Accept	0.24559	1.6227	0.2402	0.24276

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
141	4	Accept	0.74634	7.0945	0.2402	0.24276

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
142	4	Accept	0.29076	1.3294	0.2402	0.24334
143	4	Accept	0.24379	1.4595	0.2402	0.24269
144	4	Accept	0.24499	4.9134	0.2402	0.24226
145	4	Accept	0.24469	12.208	0.2402	0.24242
146	4	Accept	0.38588	1.5037	0.2402	0.24388
147	4	Accept	0.24589	1.3026	0.2402	0.24228
148	4	Accept	0.2408	1.2582	0.2402	0.24165
149	4	Accept	0.24469	1.2764	0.2402	0.24197
150	5	Accept	0.24469	6.7111	0.2402	0.24222
151	5	Accept	0.30422	1.5471	0.2402	0.24244
152	6	Accept	0.24559	1.5701	0.2402	0.2414
153	5	Accept	0.24529	66.311	0.2402	0.24243
154	5	Accept	0.24589	1.604	0.2402	0.24243
155	5	Accept	0.29345	1.4823	0.2402	0.24262

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
156	6	Accept	0.25247	1.4792	0.2402	0.24289
157	6	Accept	0.2405	1.5732	0.2402	0.242
158	6	Accept	0.2426	1.5232	0.2402	0.24223
159	6	Accept	0.25456	1.9467	0.2402	0.24253
160	5	Accept	0.24559	133.37	0.2402	0.24253
161	5	Accept	0.57314	8.1723	0.2402	0.24253
162	5	Accept	0.24649	1.6878	0.2402	0.24325
163	5	Accept	0.2417	1.5824	0.2402	0.24297
164	4	Accept	0.45528	140.46	0.2402	0.24441
165	4	Accept	0.31409	1.5573	0.2402	0.24441
166	4	Accept	0.46575	2.2317	0.2402	0.24441
167	4	Accept	0.24529	3.3215	0.2402	0.24334
168	4	Accept	0.25396	1.3297	0.2402	0.24353
169	5	Accept	0.2402	1.5257	0.2402	0.24313
170	6	Accept	0.3464	1.8583	0.2402	0.24237

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
171	6	Accept	0.24499	12.298	0.2402	0.24203
172	6	Accept	0.30272	1.6031	0.2402	0.24186
173	6	Accept	0.24349	1.4247	0.2402	0.24134
174	6	Accept	0.29794	1.6502	0.2402	0.24299
175	5	Accept	0.24529	109.43	0.2402	0.24345
176	5	Accept	0.25007	1.6406	0.2402	0.24345
177	5	Accept	0.24619	2.171	0.2402	0.24259
178	5	Accept	0.32576	1.6841	0.2402	0.24185
179	5	Accept	0.8262	890.49	0.2402	0.24179
180	5	Accept	0.30212	1.5671	0.2402	0.24174

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
181	5	Accept	0.24499	14.697	0.2402	0.24192
182	5	Accept	0.24529	61.944	0.2402	0.2414
183	5	Accept	0.24499	17.644	0.2402	0.24186

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
184	5	Accept	0.29794	1.6007	0.2402	0.24189
185	4	Accept	0.24918	279.94	0.2402	0.24236
186	4	Accept	0.26593	1.4006	0.2402	0.24236
187	4	Accept	0.24768	1.438	0.2402	0.24161
188	4	Accept	0.24589	2.5142	0.2402	0.2416
189	4	Accept	0.2405	1.2662	0.2402	0.24215
190	4	Accept	0.24678	1.2604	0.2402	0.24174
191	4	Accept	0.25426	1.2846	0.2402	0.24186
192	5	Accept	0.24768	1.3295	0.2402	0.24188
193	6	Accept	0.2414	1.3979	0.2402	0.24154
194	6	Accept	0.24499	18.984	0.2402	0.24225
195	6	Accept	0.25037	1.5112	0.2402	0.24212
196	6	Accept	0.24499	6.007	0.2402	0.24191
197	6	Accept	0.24529	111.16	0.2402	0.24215
198	5	Accept	0.24499	18.688	0.2402	0.24186

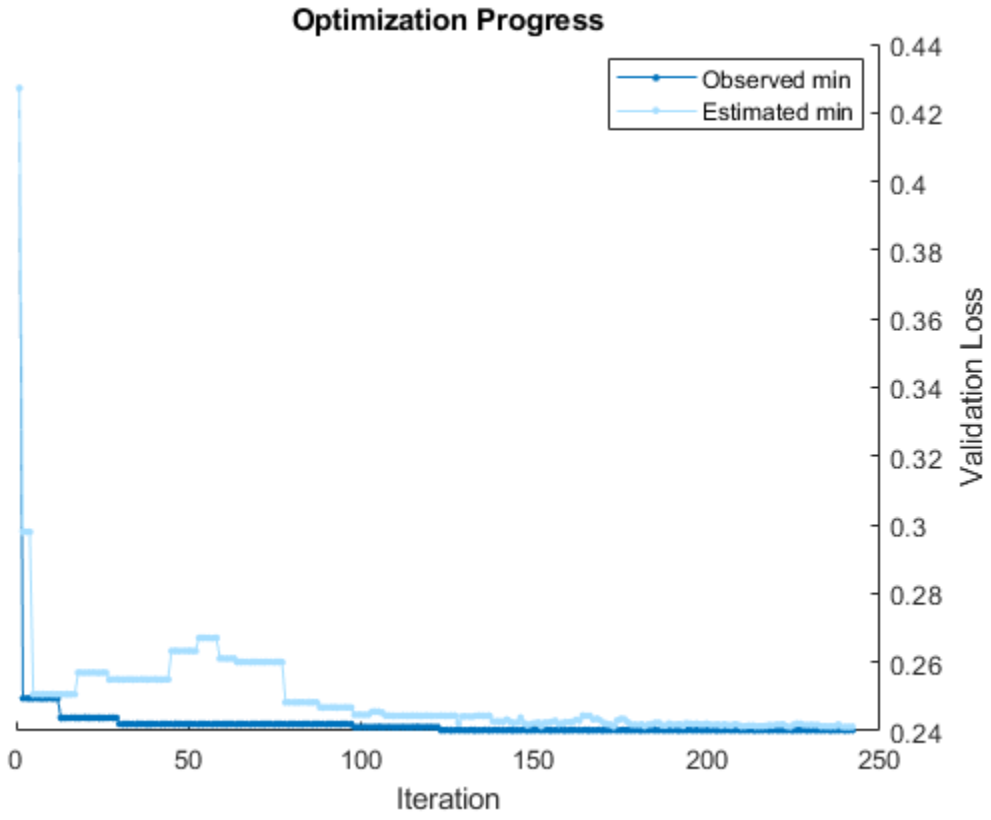
Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
199	5	Accept	0.2426	1.4789	0.2402	0.24186
200	6	Accept	0.2417	1.499	0.2402	0.24212
201	5	Accept	0.26473	277.53	0.2402	0.24176
202	5	Accept	0.30332	1.5632	0.2402	0.24176
203	5	Accept	0.3108	1.5519	0.2402	0.24188
204	5	Accept	0.2414	1.4413	0.2402	0.24147
205	5	Accept	0.3108	1.5576	0.2402	0.24195
206	5	Accept	0.2411	1.3939	0.2402	0.24163
207	5	Accept	0.29704	1.5568	0.2402	0.24155
208	5	Accept	0.2426	1.4114	0.2402	0.24195
209	5	Accept	0.24559	3.445	0.2402	0.2417
210	4	Accept	0.24678	266.64	0.2402	0.24136

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
211	4	Accept	0.24589	2.1278	0.2402	0.24136
212	3	Accept	0.95034	1042.9	0.2402	0.24153

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
213	3	Accept	0.24649	2.0006	0.2402	0.24153
214	6	Accept	0.26653	1.2521	0.2402	0.24136
215	4	Accept	0.2429	1.5969	0.2402	0.24136
216	4	Accept	0.33353	1.2604	0.2402	0.24136
217	4	Accept	0.28059	0.2661	0.2402	0.24136
218	4	Accept	0.24798	1.3024	0.2402	0.24173
219	4	Accept	0.24529	1.255	0.2402	0.24156
220	4	Accept	0.2408	1.2539	0.2402	0.24183
221	4	Accept	0.24589	1.8906	0.2402	0.24185
222	4	Accept	0.2414	1.2821	0.2402	0.24187
223	4	Accept	0.2423	1.2804	0.2402	0.24131
224	4	Accept	0.24559	3.1158	0.2402	0.24116
225	4	Accept	0.25067	1.3228	0.2402	0.24178
226	3	Accept	0.28358	328	0.2402	0.24207
227	3	Accept	0.24499	6.2167	0.2402	0.24207

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
228	6	Accept	0.2402	1.2501	0.2402	0.24173
229	3	Accept	0.24589	1.4248	0.2402	0.24173
230	3	Accept	0.43015	0.13765	0.2402	0.24173
231	3	Accept	0.47383	2.5354	0.2402	0.24173
232	3	Accept	0.67664	0.30377	0.2402	0.24173
233	6	Accept	0.2414	1.2369	0.2402	0.24133
234	3	Accept	0.27759	1.4452	0.2402	0.24133
235	3	Accept	0.28059	0.13623	0.2402	0.24133
236	3	Accept	0.52408	1.503	0.2402	0.24133
237	3	Accept	0.27789	5.8867	0.2402	0.24133
238	6	Accept	0.24499	1.2819	0.2402	0.24188
239	3	Accept	0.25097	1.4345	0.2402	0.24116
240	3	Accept	0.30212	0.12275	0.2402	0.24116

Iter	Active workers	Eval result	Validation loss	Time for training & validation (sec)	Observed min validation loss	Estimated min validation loss
241	3	Accept	0.25307	11.978	0.2402	0.24116
242	3	Accept	0.44182	11.435	0.2402	0.24116




```
_____________________________________________________________
Optimization completed.
Total iterations: 242
Total elapsed time: 1907.7403 seconds
Total time for training and validation: 4936.5339 seconds

Best observed learner is a multiclass svm model with:
    Coding (ECOC):     onevsone
    BoxConstraint:     0.033468
    KernelScale:       0.073489
Observed validation loss: 0.2402
Time for training and validation: 1.4571 seconds

Best estimated learner (returned model) is a multiclass svm model with:
    Coding (ECOC):     onevsone
    BoxConstraint:    0.0010378
    KernelScale:       0.012749
Estimated validation loss: 0.24116
Estimated time for training and validation: 1.5595 seconds

Documentation for fitcauto display
```

The final model returned by `fitcauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`creditTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

The model `Mdl` corresponds to the best point in the Bayesian optimization according to the `'min-visited-mean'` criterion. To gauge how the model will perform on new data, look at the observed cross-validation accuracy of the model (`cvAccuracy`) and its general estimated performance based on the Bayesian optimization (`estimatedAccuracy`).

```
[x,~,iteration] = bestPoint(Results,'Criterion','min-visited-mean');

cvError = Results.ObjectiveTrace(iteration);
cvAccuracy = 1 - cvError
```

```
cvAccuracy = 0.7598
```

```
estimatedError = predictObjective(Results,x);
estimatedAccuracy = 1 - estimatedError
```

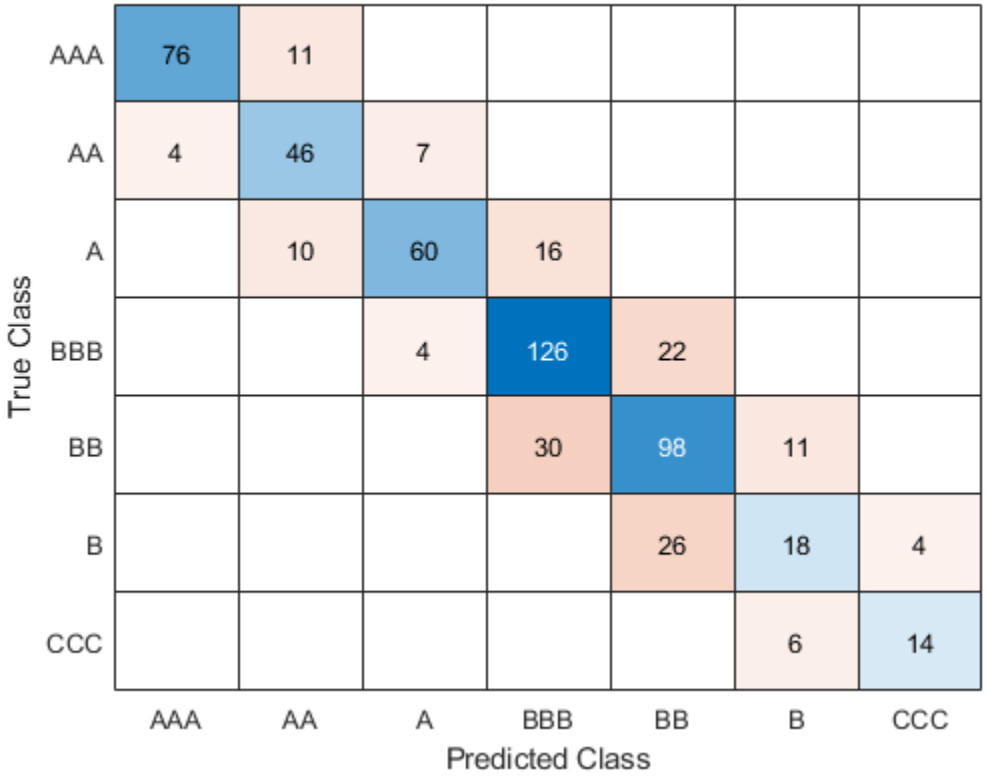
```
estimatedAccuracy = 0.7588
```

Evaluate the performance of the model on the test set. Create a confusion matrix from the results, and specify the order of the classes in the confusion matrix.

```
testAccuracy = 1 - loss(Mdl,creditTest,'Rating')
```

```
testAccuracy = 0.7438
```

```
cm = confusionchart(creditTest.Rating,predict(Mdl,creditTest));
sortClasses(cm,{'AAA','AA','A','BBB','BB','B','CCC'})
```



Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not accepted.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, specify a formula using `formula`.

If `Tbl` does not contain the response variable, specify a response variable using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class labels, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors.

- If `Y` is a character array, then each element of the class labels must correspond to one row of the array.
- The length of `Y` must be equal to the number of rows in `Tbl` or `X`.
- A good practice is to specify the class order by using the `ClassNames` name-value pair argument.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one predictor.

The length of `Y` and the number of rows in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data. The software removes rows of data corresponding to missing values in the response variable. However, the treatment of missing values in the predictor data `X` or `Tbl` varies among models (or learners).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',200,'Verbose',2)` specifies to run 200 iterations of the optimization process (that is, try 200 model hyperparameter combinations), and to display information in the Command Window about the next model hyperparameter combination to be evaluated.

Optimization Options

`Learners` — Types of classification models

`'auto'` (default) | `'all'` | `'all-linear'` | `'all-nonlinear'` | one or more learner names

Types of classification models to try during the optimization, specified as the comma-separated pair consisting of `'Learners'` and a value in the first table below or one or more learner names in the second table. Specify multiple learner names as a string or cell array.

Value	Description
`'auto'`	`fitcauto` automatically selects a subset of learners, suitable for the given predictor and response data. The learners can have model hyperparameter values that differ from the default. For more information, see “Automatic Selection of Learners” on page 33-1570.
`'all'`	`fitcauto` selects all possible learners.
`'all-linear'`	`fitcauto` selects linear learners: `'discr'` (with a linear discriminant type) and `'linear'`.
`'all-nonlinear'`	`fitcauto` selects all nonlinear learners: `'discr'` (with a quadratic discriminant type), `'ensemble'`, `'kernel'`, `'knn'`, `'nb'`, `'svm'` (with a Gaussian or polynomial kernel), and `'tree'`.

Note For greater efficiency, `fitcauto` does not select the following combinations of models when you specify one of the previous values.

- `'kernel'` and `'svm'` (with a Gaussian kernel) — `fitcauto` chooses the first when the predictor data has more than 11,000 observations, and the second otherwise.
- `'linear'` and `'svm'` (with a linear kernel) — `fitcauto` chooses the first.

Learner Name	Description
'discr'	Discriminant analysis classifier
'ensemble'	Ensemble classification model
'kernel'	Kernel classification model
'knn'	k-nearest neighbor model
'linear'	Linear classification model
'nb'	Naive Bayes classifier
'svm'	Support vector machine classifier
'tree'	Binary decision classification tree

Example: 'Learners','all'

Example: 'Learners','ensemble'

Example: 'Learners',{'svm','tree'}

Data Types: char | string | cell

OptimizeHyperparameters — Hyperparameters to optimize
'auto' (default) | 'all'

Hyperparameters to optimize, specified as the comma-separated pair consisting of 'OptimizeHyperparameters' and 'auto' or 'all'. The optimizable hyperparameters depend on the model (or learner), as described in this table.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'discr'	Delta, Gamma	DiscrimType	• When the Learners value is 'all-linear', the fitcauto function chooses among the DiscrimType values of 'linear', 'diaglinear', and 'pseudolinear', regardless of the OptimizeHyperparameters value. • When the Learners value is 'all-nonlinear', the fitcauto function chooses among the DiscrimType values of 'quadratic', 'diagquadratic', and 'pseudoquadratic', regardless of the OptimizeHyperparameters value. For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitcauto.
'ensemble'	Method, NumLearningCycles, LearnRate, MinLeafSize	MaxNumSplits, NumVariablesToSample, SplitCriterion	When the ensemble Method value is a boosting method, the ensemble NumBins value is 50. For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitcauto.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'kernel'	KernelScale, Lambda, Coding (for three or more classes only)	Learner, NumExpansionDimensions	For more information, including hyperparameter search ranges, see OptimizeHyperparameters and OptimizeHyperparameters (for three or more classes only). Note that you cannot change hyperparameter search ranges when you use fitcauto.
'knn'	Distance, NumNeighbors	DistanceWeight, Exponent, Standardize	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitcauto.
'linear'	Lambda, Learner, Coding (for three or more classes only)	Regularization	For more information, including hyperparameter search ranges, see OptimizeHyperparameters and OptimizeHyperparameters (for three or more classes only). Note that you cannot change hyperparameter search ranges when you use fitcauto.
'nb'	DistributionNames, Width	Kernel	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitcauto.

Learner Name	Hyperparameters for `'auto'`	Additional Hyperparameters for `'all'`	Notes
`'svm'`	`BoxConstraint`, `KernelScale`, `Coding` (for three or more classes only)	`KernelFunction`, `PolynomialOrder`, `Standardize`	• When the `Learners` value is `'all-linear'`, the `fitcauto` function does not optimize the `KernelFunction` or `PolynomialOrder` hyperparameters when the `OptimizeHyperparameters` value is `'all'`. • When the `Learners` value is `'all-nonlinear'`, the `fitcauto` function chooses among the `KernelFunction` values of `'gaussian'` and `'polynomial'`, regardless of the `OptimizeHyperparameters` value. • The `Standardize` value is `true` when the `OptimizeHyperparameters` value is `'auto'`. For more information, including hyperparameter search ranges, see `OptimizeHyperparameters` and `OptimizeHyperparameters` (for three or more classes only). Note that you cannot change hyperparameter search ranges when you use `fitcauto`.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'tree'	MinLeafSize	MaxNumSplits, SplitCriterion	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitcauto.

Note When 'Learners' is set to a value other than 'auto', the default values for the model hyperparameters not being optimized match the default fit function values, unless otherwise indicated in the table notes. When 'Learners' is set to 'auto', the optimized hyperparameter search ranges and nonoptimized hyperparameter values can vary, depending on the characteristics of the training data. For more information, see "Automatic Selection of Learners" on page 33-1570.

Example: 'OptimizeHyperparameters','all'

HyperparameterOptimizationOptions — Options for optimization
structure

Options for the optimization, specified as the comma-separated pair consisting of 'HyperparameterOptimizationOptions' and a structure. All fields in the structure are optional.

Field Name	Values	Default
MaxObjectiveEvaluations	Maximum number of iterations (objective function evaluations)	30*L, where L is the number of learners (see Learners)
MaxTime	Time limit, specified as a positive real number. The time limit is in seconds, as measured by tic and toc. Run time can exceed MaxTime because MaxTime does not interrupt function evaluations.	Inf
ShowPlots	Logical value indicating whether to show plots. If true, this field plots the best observed and estimated objective function values (so far) against the iteration number.	true
SaveIntermediateResults	Logical value indicating whether to save results. If true, this field overwrites a workspace variable named 'BayesoptResults' at each iteration. The variable is a BayesianOptimization object.	false

Field Name	Values	Default
`Verbose`	Display at the command line: • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with additional information about the next point to be evaluated	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Specify only one of the following three options.		
`CVPartition`	`cvpartition` object, created by `cvpartition`	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	Scalar in the range `(0,1)` representing the holdout fraction	
`Kfold`	Integer greater than 1	

Example: `'HyperparameterOptimizationOptions',struct('UseParallel',true)`

Data Types: `struct`

Classification Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcauto` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitcauto` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. However, learners that use decision trees assume that mathematically ordered categorical vectors are continuous variables. If the predictor data is a matrix (`X`), `fitcauto` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value pair argument.

For more information on how fitting functions treat categorical predictors, see “Automatic Creation of Dummy Variables” on page 2-49.

Note

- `fitcauto` does not support categorical predictors for discriminant analysis classifiers. That is, if you want `Learners` to include `'discr'` models, you cannot specify the `'CategoricalPredictors'` name-value pair argument or use a table of sample data (`Tbl`) containing categorical predictors.
- `fitcauto` does not support a mix of numeric and categorical predictors for k-nearest neighbor models. That is, if you want `Learners` to include `'knn'` models, you must specify the `'CategoricalPredictors'` value as `'all'` or `[]`.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure array.

- If you specify a square matrix `Cost` and the true class of an observation is `i`, then `Cost(i,j)` is the cost of classifying a point into class `j`. That is, rows correspond to the true classes and columns correspond to the predicted classes. To specify the class order for the corresponding rows and columns of `Cost`, also specify the `ClassNames` name-value pair argument.
- If you specify a structure `S`, then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default value for `Cost` is `ones(K) - eye(K)`, where `K` is the number of distinct classes.

Example: `'Cost',[0 1; 2 0]`

Data Types: `single` | `double` | `struct`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.

  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcauto` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to $1/K$, where K is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements to sum to `1`.
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements to sum to `1`.

Example: `'Prior',struct('ClassNames',{{'b','g'}},'ClassProbs',1:2)`

Data Types: `single` | `double` | `char` | `string` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
positive numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a positive numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`Mdl` — Trained classification model

classification model object

Trained classification model, returned as one of the classification model objects in this table.

Learner Name	Returned Model Object
`'discr'`	`CompactClassificationDiscriminant`
`'ensemble'`	`CompactClassificationEnsemble`
`'kernel'`	• `ClassificationKernel` for binary classification • `CompactClassificationECOC` for multiclass classification
`'knn'`	`ClassificationKNN`
`'linear'`	• `ClassificationLinear` for binary classification • `CompactClassificationECOC` for multiclass classification
`'nb'`	`CompactClassificationNaiveBayes`
`'svm'`	• `CompactClassificationSVM` for binary classification • `CompactClassificationECOC` for multiclass classification
`'tree'`	`CompactClassificationTree`

`OptimizationResults` — Optimization results

`BayesianOptimization` object

Optimization results, returned as a `BayesianOptimization` object. For more information on the Bayesian optimization process, see “Bayesian Optimization” on page 33-1571.

More About

Verbose Display

When you set the `Verbose` field of the `HyperparameterOptimizationOptions` name-value pair argument to `1` or `2`, the `fitcauto` function provides an iterative display of the optimization results.

The following table describes the columns in the display and their entries.

Column Name	Description
`Iter`	Iteration number — You can set a limit to the number of iterations by using the `MaxObjectiveEvaluations` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

Column Name	Description
`Active workers`	Number of active parallel workers — This column appears only when you run the optimization in parallel by setting the `UseParallel` field of the `'HyperparameterOptimizationOptions'` name-value pair argument to `true`.
`Eval result`	One of the following evaluation results: • `Best` — The learner and hyperparameter values at this iteration give the minimum observed validation loss computed so far. That is, the `Validation loss` value is the smallest computed so far. • `Accept` — The learner and hyperparameter values at this iteration give meaningful (for example, non-`NaN`) observed and estimated validation loss values. • `Error` — The learner and hyperparameter values at this iteration result in an error (for example, a `Validation loss` value of `NaN`).
`Validation loss`	Validation loss computed for the learner and hyperparameter values at this iteration — In particular, `fitcauto` computes the cross-validation classification error by default. You can change the validation scheme by using the `CVPartition`, `Holdout`, or `Kfold` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.
`Time for training & validation (sec)`	Time taken to train and compute the validation loss for the model with the learner and hyperparameter values at this iteration (in seconds) — In particular, this value excludes the time required to update the objective function model maintained by the Bayesian optimization process. For more details, see “Bayesian Optimization” on page 33-1571.
`Observed min validation loss`	Observed minimum validation loss computed so far — This value corresponds to the smallest `Validation loss` value computed so far in the optimization process. By default, `fitcauto` returns a plot of the optimization that displays dark blue points for the observed minimum validation loss values. This plot does not appear when the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument is set to `false`.

Column Name	Description
`Estimated min validation loss`	Estimated minimum validation loss — At each iteration, `fitcauto` updates an objective function model maintained by the Bayesian optimization process and uses this model to estimate the minimum validation loss. For more details, see “Bayesian Optimization” on page 33-2113. By default, `fitcauto` returns a plot of the optimization that displays light blue points for the estimated minimum validation loss values. This plot does not appear when the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument is set to `false`.
`Learner`	Model type evaluated at this iteration — Specify the learners used in the optimization by using the `'Learners'` name-value pair argument.
`Hyperparameter: Value`	Hyperparameter values at this iteration — Specify the hyperparameters used in the optimization by using the `'OptimizeHyperparameters'` name-value pair argument.

The display also includes a description of two models:

- `Best observed learner` — This model, with the listed learner type and hyperparameter values, yields the final observed minimum validation loss.
- `Best estimated learner` — This model, with the listed learner type and hyperparameter values, yields the final estimated minimum validation loss. `fitcauto` retrains the model on the entire training data set and returns it as the `Mdl` output.

Tips

- Depending on the size of your data and the number of learners you specify, `fitcauto` can take some time to run. If you have a Parallel Computing Toolbox license, you can speed up computations by running the optimization in parallel. To do so, specify `'HyperparameterOptimizationOptions',struct('UseParallel',true)`. You can include other fields in the structure to control other aspects of the optimization. See `HyperparameterOptimizationOptions`.

Algorithms

Automatic Selection of Learners

When you specify `'Learners','auto'`, the `fitcauto` function analyzes the predictor and response data in order to choose appropriate learners. The function considers whether the data set has any of these characteristics:

- Categorical predictors
- Missing values for more than 5% of the data
- Imbalanced data, where the ratio of the number of observations in the largest class to the number of observations in the smallest class is greater than 5
- More than 100 observations in the smallest class
- Wide data, where the number of predictors is greater than or equal to the number of observations
- High-dimensional data, where the number of predictors is greater than 100
- Large data, where the number of observations is greater than 50,000
- Binary response variable
- Ordinal response variable

The selected learners are always a subset of those listed in the `Learners` table. However, the associated models tried during the optimization process can have different default values for hyperparameters not being optimized, as well as different search ranges for hyperparameters being optimized.

Bayesian Optimization

The goal of Bayesian optimization, and optimization in general, is to find a point that minimizes an objective function. In the context of `fitcauto`, a point is a learner type together with a set of hyperparameter values for the learner (see `Learners` and `OptimizeHyperparameters`), and the objective function is the cross-validation classification error, by default. The Bayesian optimization implemented in `fitcauto` internally maintains a multi-`TreeBagger` model of the objective function. That is, the objective function model splits along the learner type and, for a given learner, the model is a `TreeBagger` ensemble for regression. (This underlying model differs from the Gaussian process model employed by other Statistics and Machine Learning Toolbox functions that use Bayesian optimization.) Bayesian optimization trains the underlying model by using objective function evaluations, and determines the next point to evaluate by using an acquisition function (`'expected-improvement'`). For more information, see “Expected Improvement” on page 10-4. The acquisition function balances between sampling at points with low modeled objective function values and exploring areas that are not well modeled yet. At the end of the optimization, `fitcauto` chooses the point with the minimum objective function model value, among the points evaluated during the optimization. For more information, see the `'Criterion','min-visited-mean'` name-value pair argument of `bestPoint`.

Alternative Functionality

- If you are unsure which models work best for your data set, you can alternatively use the **Classification Learner** app. Using the app, you can perform hyperparameter tuning for different models, and choose the optimized model that performs best. Although you must select a specific model before you can tune the model hyperparameters, **Classification Learner** provides greater flexibility for selecting optimizable hyperparameters and setting hyperparameter values. However, you cannot optimize in parallel, choose `'linear'` or `'kernel'` learners, specify observation weights, or specify prior probabilities in the app. For more information, see “Hyperparameter Optimization in Classification Learner App” on page 23-54.
- If you know which models might suit your data, you can alternatively use the corresponding model fit functions and specify the `'OptimizeHyperparameters'` name-value pair argument to tune hyperparameters. You can compare the results across the models to select the best classifier. For an example of this process, see “Moving Towards Automating Model Selection Using Bayesian Optimization” on page 18-197.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions',struct('UseParallel',true)` name-value pair argument in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`fitcdiscr` | `fitcecoc` | `fitcensemble` | `fitckernel` | `fitcknn` | `fitclinear` | `fitcnb` | `fitcsvm` | `fitctree`

Topics

“Automated Classifier Selection with Bayesian Optimization” on page 18-205
“Hyperparameter Optimization in Classification Learner App” on page 23-54

Introduced in R2020a

fitcdiscr

Fit discriminant analysis classifier

Syntax

```
Mdl = fitcdiscr(Tbl,ResponseVarName)
Mdl = fitcdiscr(Tbl,formula)
Mdl = fitcdiscr(Tbl,Y)

Mdl = fitcdiscr(X,Y)

Mdl = fitcdiscr(___,Name,Value)
```

Description

`Mdl = fitcdiscr(Tbl,ResponseVarName)` returns a fitted discriminant analysis model based on the input variables (also known as predictors, features, or attributes) contained in the table `Tbl` and output (response or labels) contained in `ResponseVarName`.

`Mdl = fitcdiscr(Tbl,formula)` returns a fitted discriminant analysis model based on the input variables contained in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitcdiscr(Tbl,Y)` returns a fitted discriminant analysis model based on the input variables contained in the table `Tbl` and response `Y`.

`Mdl = fitcdiscr(X,Y)` returns a discriminant analysis classifier based on the input variables `X` and response `Y`.

`Mdl = fitcdiscr(___,Name,Value)` fits a classifier with additional options specified by one or more name-value pair arguments, using any of the previous syntaxes. For example, you can optimize hyperparameters to minimize the model's cross-validation loss, or specify the cost of misclassification, the prior probabilities for each class, or the observation weights.

Examples

Train Discriminant Analysis Model

Load Fisher's iris data set.

```
load fisheriris
```

Train a discriminant analysis model using the entire data set.

```
Mdl = fitcdiscr(meas,species)
```

```
Mdl =
  ClassificationDiscriminant
             ResponseName: 'Y'
    CategoricalPredictors: []
```

```
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'
        NumObservations: 150
            DiscrimType: 'linear'
                     Mu: [3x4 double]
                 Coeffs: [3x3 struct]


  Properties, Methods
```

`Mdl` is a `ClassificationDiscriminant` model. To access its properties, use dot notation. For example, display the group means for each predictor.

```
Mdl.Mu
```

```
ans = 3×4

    5.0060    3.4280    1.4620    0.2460
    5.9360    2.7700    4.2600    1.3260
    6.5880    2.9740    5.5520    2.0260
```

To predict labels for new observations, pass `Mdl` and predictor data to `predict`.

Optimize Discriminant Analysis Model

This example shows how to optimize hyperparameters automatically using `fitcdiscr`. The example uses Fisher's iris data.

Load the data.

```
load fisheriris
```

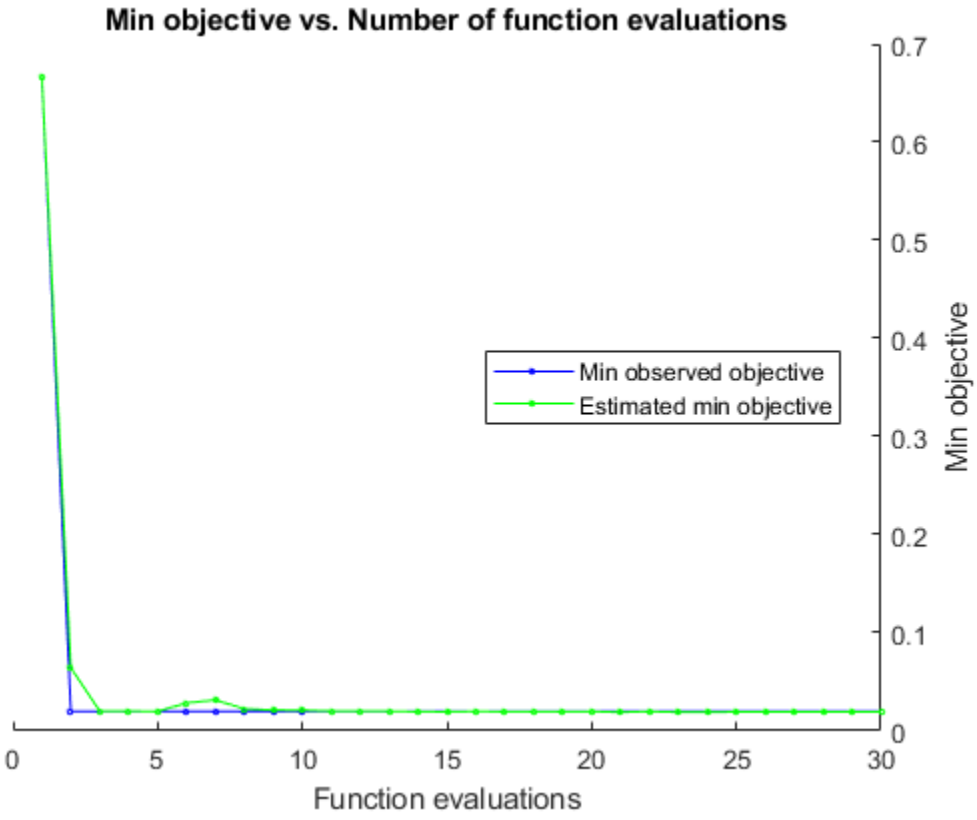
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

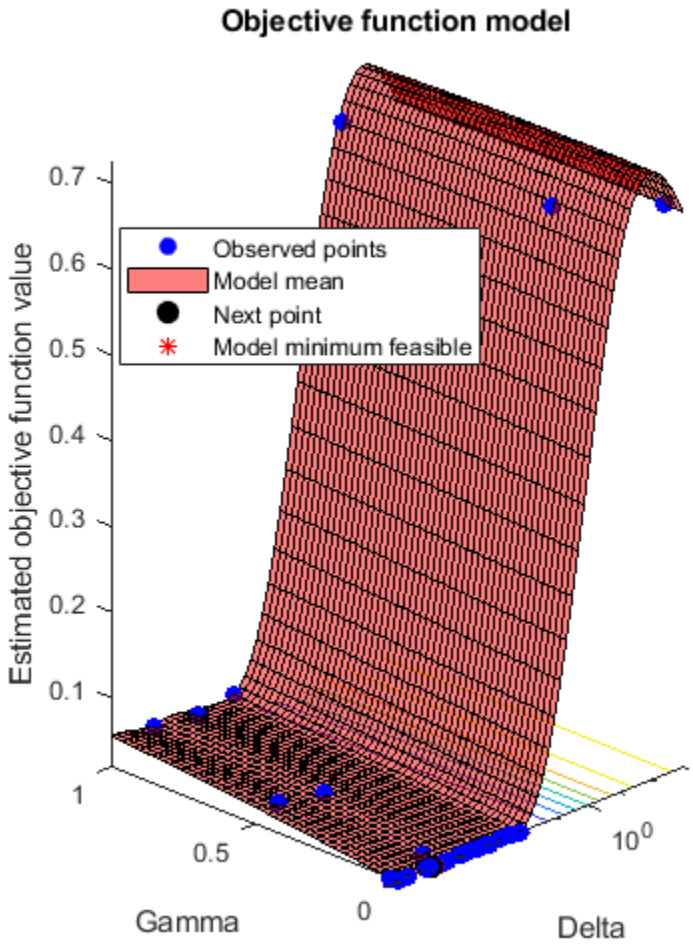
For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng(1)
Mdl = fitcdiscr(meas,species,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        Delta |         C
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.66667 |     0.86213 |     0.66667 |     0.66667 |       13.261 |      0.2
|    2 | Best   |        0.02 |     0.26312 |        0.02 |    0.064227 |   2.7404e-05 |     0.01
|    3 | Accept |        0.04 |     0.22833 |        0.02 |    0.020084 |   3.2455e-06 |      0.4
|    4 | Accept |     0.66667 |     0.22273 |        0.02 |    0.020118 |       14.879 |      0.9
|    5 | Accept |    0.046667 |     0.21553 |        0.02 |    0.019907 |   0.00031449 |      0.9
|    6 | Accept |        0.04 |     0.16288 |        0.02 |    0.028438 |   4.5092e-05 |      0.4
```

```
|    7 | Accept |    0.046667 |     0.16118 |        0.02 |    0.031424 |   2.0973e-05 |        0
|    8 | Accept |        0.02 |      0.1494 |        0.02 |    0.022424 |   1.0554e-06 |    0.002
|    9 | Accept |        0.02 |     0.16584 |        0.02 |    0.021105 |   1.1232e-06 |   0.0001
|   10 | Accept |        0.02 |     0.15767 |        0.02 |    0.020948 |   0.00011837 |    0.003
|   11 | Accept |        0.02 |     0.13495 |        0.02 |    0.020172 |   1.0292e-06 |     0.02
|   12 | Accept |        0.02 |     0.12469 |        0.02 |    0.020105 |   9.7792e-05 |    0.002
|   13 | Accept |        0.02 |     0.17518 |        0.02 |    0.020038 |   0.00036014 |    0.001
|   14 | Accept |        0.02 |     0.12286 |        0.02 |    0.019597 |   0.00021059 |    0.004
|   15 | Accept |        0.02 |     0.10558 |        0.02 |    0.019461 |   1.1911e-05 |    0.001
|   16 | Accept |        0.02 |     0.14187 |        0.02 |     0.01993 |    0.0017896 |   0.0007
|   17 | Accept |        0.02 |     0.12236 |        0.02 |    0.019551 |   0.00073745 |    0.006
|   18 | Accept |        0.02 |     0.12795 |        0.02 |    0.019776 |   0.00079304 |   0.0001
|   19 | Accept |        0.02 |     0.16053 |        0.02 |    0.019678 |     0.007292 |    0.000
|   20 | Accept |    0.046667 |     0.16801 |        0.02 |    0.019785 |    0.0074408 |      0.9
|======|========|=============|=============|=============|=============|==============|=========
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |        Delta |        C
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|======|========|=============|=============|=============|=============|==============|=========
|   21 | Accept |        0.02 |      0.1667 |        0.02 |    0.019043 |    0.0036004 |    0.002
|   22 | Accept |        0.02 |     0.15705 |        0.02 |    0.019755 |   2.5238e-05 |    0.001
|   23 | Accept |        0.02 |     0.22521 |        0.02 |      0.0191 |   1.5478e-05 |    0.002
|   24 | Accept |        0.02 |     0.18189 |        0.02 |    0.019081 |    0.0040557 |   0.0004
|   25 | Accept |        0.02 |     0.16974 |        0.02 |    0.019333 |    2.959e-05 |    0.001
|   26 | Accept |        0.02 |     0.18289 |        0.02 |    0.019369 |   2.3111e-06 |    0.002
|   27 | Accept |        0.02 |     0.15596 |        0.02 |    0.019455 |   3.8898e-05 |    0.001
|   28 | Accept |        0.02 |     0.17513 |        0.02 |    0.019449 |    0.0035925 |    0.002
|   29 | Accept |     0.66667 |     0.18111 |        0.02 |    0.019479 |       998.93 |     0.00
|   30 | Accept |        0.02 |      0.1495 |        0.02 |     0.01947 |   8.1557e-06 |    0.000
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25
30




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 51.8808 seconds
Total objective function evaluation time: 5.718

Best observed feasible point:
      Delta        Gamma
    __________    ________

    2.7404e-05    0.073264

Observed objective function value = 0.02
Estimated objective function value = 0.022693
Function evaluation time = 0.26312

Best estimated feasible point (according to models):
      Delta         Gamma
    __________    _________
```

```
    2.5238e-05    0.0015542

Estimated objective function value = 0.01947
Estimated function evaluation time = 0.15389

Mdl =
  ClassificationDiscriminant
                         ResponseName: 'Y'
                CategoricalPredictors: []
                           ClassNames: {'setosa'  'versicolor'  'virginica'}
                      ScoreTransform: 'none'
                     NumObservations: 150
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                         DiscrimType: 'linear'
                                  Mu: [3x4 double]
                              Coeffs: [3x3 struct]


  Properties, Methods
```

The fit achieves about 2% loss for the default 5-fold cross validation.

Optimize Discriminant Analysis Model on Tall Array

This example shows how to optimize hyperparameters of a discriminant analysis model automatically using a tall array. The sample data set `airlinesmall.csv` is a large data set that contains a tabular file of airline flight data. This example creates a tall table containing the data and uses it to run the optimization procedure.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Create a datastore that references the folder location with the data. Select a subset of the variables to work with, and treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Create a tall table that contains the data in the datastore.

```
ds = datastore('airlinesmall.csv');
ds.SelectedVariableNames = {'Month','DayofMonth','DayOfWeek',...
                            'DepTime','ArrDelay','Distance','DepDelay'};
ds.TreatAsMissing = 'NA';
tt  = tall(ds) % Tall table
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

tt =

  M×7 tall table

    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    Distance    DepDelay
    _____    __________    _________    _______    ________    ________    ________
```

```
    10          21           3          642           8          308          12
    10          26           1         1021           8          296           1
    10          23           5         2055          21          480          20
    10          23           5         1332          13          296          12
    10          22           4          629           4          373          -1
    10          28           3         1446          59          308          63
    10           8           4          928           3          447          -2
    10          10           6          859          11          954          -1
     :           :           :            :           :            :           :
     :           :           :            :           :            :           :
```

Determine the flights that are late by 10 minutes or more by defining a logical variable that is true for a late flight. This variable contains the class labels. A preview of this variable includes the first few rows.

```
Y = tt.DepDelay > 10 % Class labels
```

```
Y =

  M×1 tall logical array

   1
   0
   1
   1
   0
   1
   0
   0
   :
   :
```

Create a tall array for the predictor data.

```
X = tt{:,1:end-1} % Predictor data
```

```
X =

  M×6 tall double matrix

          10          21           3         642           8         308
          10          26           1        1021           8         296
          10          23           5        2055          21         480
          10          23           5        1332          13         296
          10          22           4         629           4         373
          10          28           3        1446          59         308
          10           8           4         928           3         447
          10          10           6         859          11         954
           :           :           :           :           :           :
           :           :           :           :           :           :
```

Remove rows in `X` and `Y` that contain missing data.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:end-1);
Y = R(:,end);
```

Standardize the predictor variables.

```
Z = zscore(X);
```

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value pair argument. Find the optimal `'DiscrimType'` value that minimizes holdout cross-validation loss. (Specifying `'auto'` uses `'DiscrimType'`.) For reproducibility, use the `'expected-improvement-plus'` acquisition function and set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

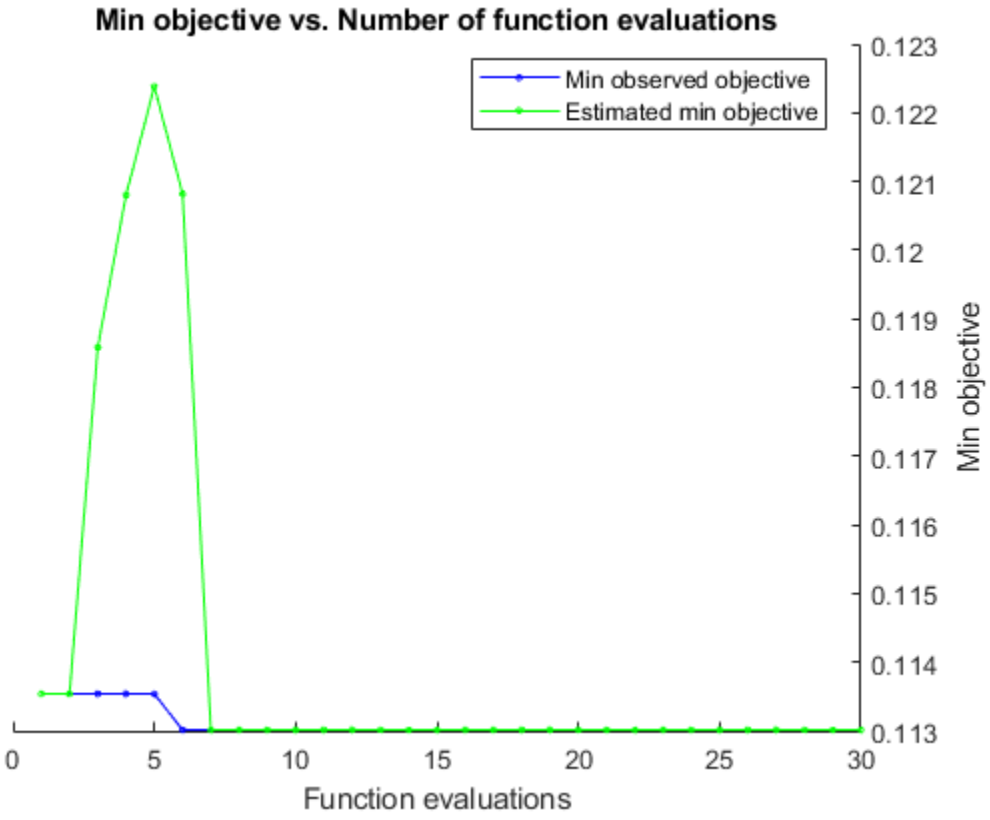
```
rng('default')
tallrng('default')
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitcdiscr(Z,Y,...
    'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('Holdout',0.3,...
    'AcquisitionFunctionName','expected-improvement-plus'))

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 5.7 sec
- Pass 2 of 2: Completed in 4.3 sec
Evaluation completed in 16 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 2.5 sec
Evaluation completed in 2.8 sec
|=====================================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  DiscrimType |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================|
|    1 | Best   |     0.11354 |      25.315 |     0.11354 |     0.11354 |    quadratic |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.5 sec
Evaluation completed in 2.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 1.6 sec
|    2 | Accept |     0.11354 |      7.9367 |     0.11354 |     0.11354 | pseudoQuadra |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.78 sec
Evaluation completed in 0.91 sec
|    3 | Accept |     0.12869 |      6.5057 |     0.11354 |     0.11859 | pseudoLinear |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.9 sec
Evaluation completed in 1.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec
|    4 | Accept |     0.12745 |      6.4167 |     0.11354 |      0.1208 |   diagLinear |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.85 sec
Evaluation completed in 1.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.8 sec
```

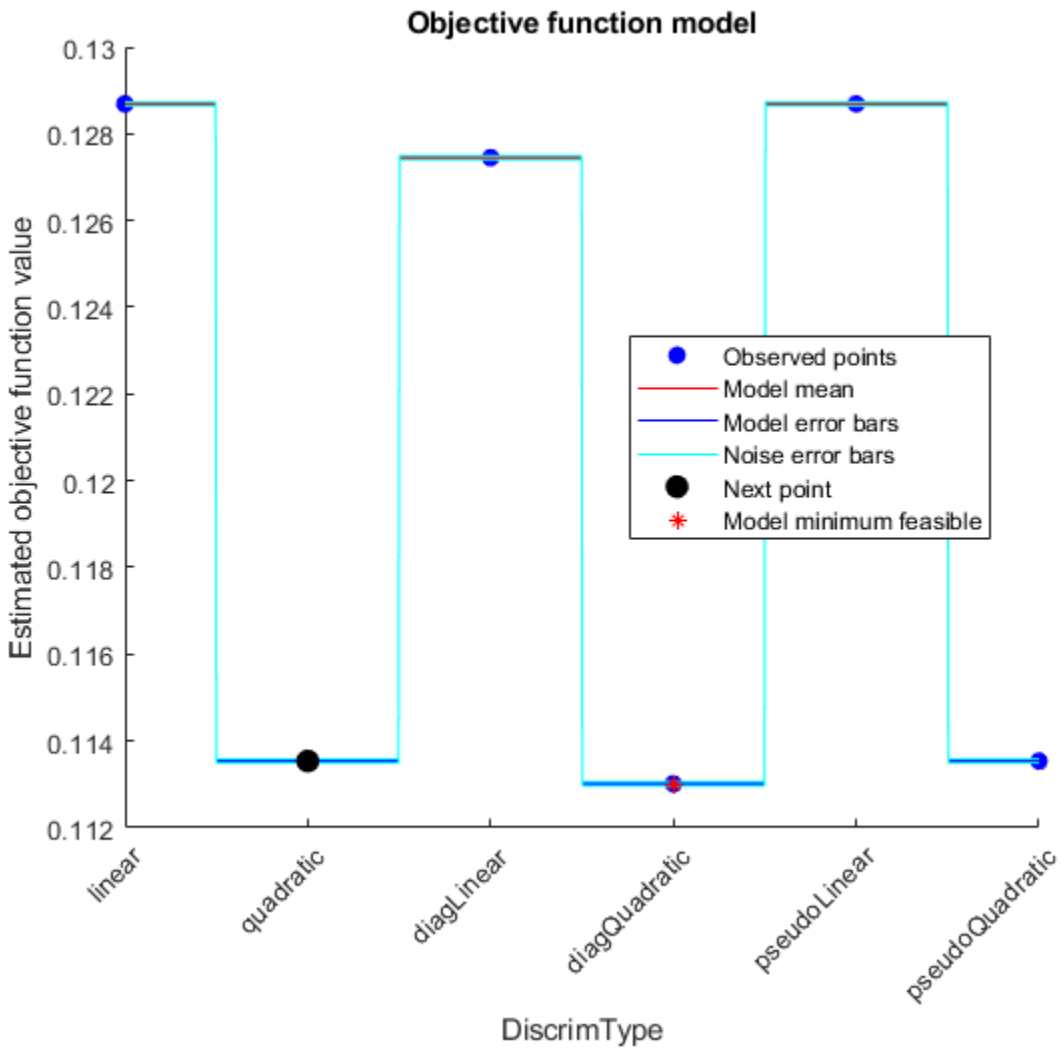
```
Evaluation completed in 0.93 sec
|    5 | Accept |     0.12869 |      6.1236 |     0.11354 |     0.12238 |       linear |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.85 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.9 sec
|    6 | Best   |     0.11301 |      5.4147 |     0.11301 |     0.12082 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.82 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.89 sec
|    7 | Accept |     0.11301 |       5.297 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.84 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.8 sec
Evaluation completed in 0.93 sec
|    8 | Accept |     0.11301 |      5.6152 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 2.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.88 sec
|    9 | Accept |     0.11301 |      5.9147 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec
|   10 | Accept |     0.11301 |      6.0504 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.82 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec
|   11 | Accept |     0.11301 |      5.9595 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.76 sec
Evaluation completed in 0.91 sec
|   12 | Accept |     0.11301 |      5.4266 |     0.11301 |     0.11301 | diagQuadrati |
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.87 sec
|   13 | Accept |     0.11301 |      5.3869 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.8 sec
Evaluation completed in 0.97 sec
|   14 | Accept |     0.11301 |      5.4876 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.85 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.73 sec
Evaluation completed in 0.85 sec
|   15 | Accept |     0.11301 |      5.4052 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.78 sec
Evaluation completed in 0.9 sec
|   16 | Accept |     0.11301 |      5.4434 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.8 sec
Evaluation completed in 0.93 sec
|   17 | Accept |     0.11301 |      5.5804 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.94 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.79 sec
Evaluation completed in 0.92 sec
|   18 | Accept |     0.11354 |       5.616 |     0.11301 |     0.11301 | pseudoQuadra |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.85 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.76 sec
Evaluation completed in 0.88 sec
|   19 | Accept |     0.11301 |      5.4031 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.76 sec
Evaluation completed in 1.4 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.88 sec
|   20 | Accept |     0.11301 |      5.1974 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 1.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.87 sec
|======================================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  DiscrimType |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|======================================================================================|
|   21 | Accept |     0.11301 |      5.1418 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.73 sec
Evaluation completed in 0.86 sec
|   22 | Accept |     0.11301 |      5.9864 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.78 sec
Evaluation completed in 0.91 sec
|   23 | Accept |     0.11354 |      5.5656 |     0.11301 |     0.11301 |    quadratic |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.82 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.9 sec
|   24 | Accept |     0.11354 |      5.3012 |     0.11301 |     0.11301 | pseudoQuadra |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 2.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.9 sec
|   25 | Accept |     0.11301 |      6.2276 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.89 sec
|   26 | Accept |     0.11301 |      5.5308 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.92 sec
```

```
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
|   27 | Accept |     0.11301 |      5.7396 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.78 sec
Evaluation completed in 0.9 sec
|   28 | Accept |     0.11354 |      5.4403 |     0.11301 |     0.11301 |    quadratic |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.81 sec
Evaluation completed in 0.93 sec
|   29 | Accept |     0.11301 |      5.3572 |     0.11301 |     0.11301 | diagQuadrati |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.74 sec
Evaluation completed in 0.85 sec
|   30 | Accept |     0.11354 |      5.2718 |     0.11301 |     0.11301 |    quadratic |
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
0.123
0.122
0.121
0.12
0.119
0.118
0.117
0.116
0.115
0.114
0.113
Min objective
0
5
10
15
20
25
30
Function evaluations




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 229.5689 seconds.
Total objective function evaluation time: 191.058

Best observed feasible point:
     DiscrimType
    _____________

    diagQuadratic

Observed objective function value = 0.11301
Estimated objective function value = 0.11301
Function evaluation time = 5.4147

Best estimated feasible point (according to models):
     DiscrimType
```

```
    _____________

    diagQuadratic

Estimated objective function value = 0.11301
Estimated function evaluation time = 5.784

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.76 sec
Evaluation completed in 1.4 sec

Mdl =
  CompactClassificationDiscriminant
           PredictorNames: {'x1'  'x2'  'x3'  'x4'  'x5'  'x6'}
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [0 1]
           ScoreTransform: 'none'
              DiscrimType: 'diagQuadratic'
                       Mu: [2×6 double]
                   Coeffs: [2×2 struct]


  Properties, Methods


FitInfo = struct with no fields.



HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                      ObjectiveFcn: @createObjFcn/tallObjFcn
              VariableDescriptions: [1×1 optimizableVariable]
                           Options: [1×1 struct]
                      MinObjective: 0.1130
                   XAtMinObjective: [1×1 table]
             MinEstimatedObjective: 0.1130
          XAtMinEstimatedObjective: [1×1 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 229.5689
                         NextPoint: [1×1 table]
                            XTrace: [30×1 table]
                    ObjectiveTrace: [30×1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30×1 cell}
      ObjectiveEvaluationTimeTrace: [30×1 double]
                IterationTimeTrace: [30×1 double]
                        ErrorTrace: [30×1 double]
                  FeasibilityTrace: [30×1 logical]
       FeasibilityProbabilityTrace: [30×1 double]
                IndexOfMinimumTrace: [30×1 double]
              ObjectiveMinimumTrace: [30×1 double]
    EstimatedObjectiveMinimumTrace: [30×1 double]
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X`.

The software considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. Consequently, the software does not train using observations with a missing response.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

X — Predictor data

numeric matrix

Predictor values, specified as a numeric matrix. Each column of `X` represents one variable, and each row represents one observation.

`fitcdiscr` considers `NaN` values in `X` as missing values. `fitcdiscr` does not use observations with missing values for `X` in the fit.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'DiscrimType','quadratic','SaveMemory','on'` specifies a quadratic discriminant classifier and does not store the covariance matrix in the output object.

Model Parameters

ClassNames — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.

- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Cost of misclassification
square matrix | structure

Cost of misclassification of a point, specified as the comma-separated pair consisting of `'Cost'` and one of the following:

- Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- Structure `S` having two fields: `S.ClassNames` containing the group names as a variable of the same type as `Y`, and `S.ClassificationCosts` containing the cost matrix.

The default is `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`.

Data Types: `single` | `double` | `struct`

`Delta` — Linear coefficient threshold
`0` (default) | nonnegative scalar value

Linear coefficient threshold, specified as the comma-separated pair consisting of `'Delta'` and a nonnegative scalar value. If a coefficient of `Mdl` has magnitude smaller than `Delta`, `Mdl` sets this coefficient to `0`, and you can eliminate the corresponding predictor from the model. Set `Delta` to a higher value to eliminate more predictors.

`Delta` must be `0` for quadratic discriminant models.

Data Types: `single` | `double`

`DiscrimType` — Discriminant type
`'linear'` (default) | `'quadratic'` | `'diaglinear'` | `'diagquadratic'` | `'pseudolinear'` | `'pseudoquadratic'`

Discriminant type, specified as the comma-separated pair consisting of `'DiscrimType'` and a character vector or string scalar in this table.

Value	Description	Predictor Covariance Treatment
`'linear'`	Regularized linear discriminant analysis (LDA)	• All classes have the same covariance matrix. • $\widehat{\Sigma}_\gamma = (1-\gamma)\widehat{\Sigma} + \gamma \mathrm{diag}\left(\widehat{\Sigma}\right)$. $\widehat{\Sigma}$ is the empirical, pooled covariance matrix and γ is the amount of regularization.
`'diaglinear'`	LDA	All classes have the same, diagonal covariance matrix.
`'pseudolinear'`	LDA	All classes have the same covariance matrix. The software inverts the covariance matrix using the pseudo inverse.
`'quadratic'`	Quadratic discriminant analysis (QDA)	The covariance matrices can vary among classes.
`'diagquadratic'`	QDA	The covariance matrices are diagonal and can vary among classes.
`'pseudoquadratic'`	QDA	The covariance matrices can vary among classes. The software inverts the covariance matrix using the pseudo inverse.

Note To use regularization, you must specify `'linear'`. To specify the amount of regularization, use the `Gamma` name-value pair argument.

Example: `'DiscrimType','quadratic'`

`FillCoeffs` — `Coeffs` property flag

`'on'` | `'off'`

`Coeffs` property flag, specified as the comma-separated pair consisting of `'FillCoeffs'` and `'on'` or `'off'`. Setting the flag to `'on'` populates the `Coeffs` property in the classifier object. This can be computationally intensive, especially when cross-validating. The default is `'on'`, unless you specify a cross-validation name-value pair, in which case the flag is set to `'off'` by default.

Example: `'FillCoeffs','off'`

`Gamma` — Amount of regularization

scalar value in the interval [0,1]

Amount of regularization to apply when estimating the covariance matrix of the predictors, specified as the comma-separated pair consisting of `'Gamma'` and a scalar value in the interval [0,1]. `Gamma` provides finer control over the covariance matrix structure than `DiscrimType`.

- If you specify `0`, then the software does not use regularization to adjust the covariance matrix. That is, the software estimates and uses the unrestricted, empirical covariance matrix.

  - For linear discriminant analysis, if the empirical covariance matrix is singular, then the software automatically applies the minimal regularization required to invert the covariance matrix. You can display the chosen regularization amount by entering `Mdl.Gamma` at the command line.
  - For quadratic discriminant analysis, if at least one class has an empirical covariance matrix that is singular, then the software throws an error.
- If you specify a value in the interval (0,1), then you must implement linear discriminant analysis, otherwise the software throws an error. Consequently, the software sets `DiscrimType` to `'linear'`.
- If you specify `1`, then the software uses maximum regularization for covariance matrix estimation. That is, the software restricts the covariance matrix to be diagonal. Alternatively, you can set `DiscrimType` to `'diagLinear'` or `'diagQuadratic'` for diagonal covariance matrices.

Example: `'Gamma',1`

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcdiscr` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

If you set values for both `Weights` and `Prior`, the weights are renormalized to add up to the value of the prior probability in the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`SaveMemory` — Flag to save covariance matrix
`'off'` (default) | `'on'`

Flag to save covariance matrix, specified as the comma-separated pair consisting of `'SaveMemory'` and either `'on'` or `'off'`. If you specify `'on'`, then `fitcdiscr` does not store the full covariance matrix, but instead stores enough information to compute the matrix. The `predict` method computes the full covariance matrix for prediction, and does not store the matrix. If you specify `'off'`, then `fitcdiscr` computes and stores the full covariance matrix in `Mdl`.

Specify `SaveMemory` as `'on'` when the input matrix contains thousands of predictors.

Example: `'SaveMemory','on'`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'CrossVal','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Delta','Gamma'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitcdiscr` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcdiscr` are:

- `Delta` — `fitcdiscr` searches among positive values, by default log-scaled in the range `[1e-6,1e3]`.

- `DiscrimType` — `fitcdiscr` searches among `'linear'`, `'quadratic'`, `'diagLinear'`, `'diagQuadratic'`, `'pseudoLinear'`, and `'pseudoQuadratic'`.
- `Gamma` — `fitcdiscr` searches among real values in the range `[0,1]`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitcdiscr',meas,species);
params(1).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Discriminant Analysis Model” on page 33-1574.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`

Field Name	Values	Default
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`

Field Name	Values	Default
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see "Parallel Bayesian Optimization" on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained discriminant analysis classification model

`ClassificationDiscriminant` model object | `ClassificationPartitionedModel` cross-validated model object

Trained discriminant analysis classification model, returned as a `ClassificationDiscriminant` model object or a `ClassificationPartitionedModel` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `ClassificationPartitionedModel` cross-validated model object. Otherwise, `Mdl` is a `ClassificationDiscriminant` model object.

To reference properties of `Mdl`, use dot notation. For example, to display the estimated component means at the Command Window, enter `Mdl.Mu`.

More About

Discriminant Classification

The model for discriminant analysis is:

- Each class (`Y`) generates data (`X`) using a multivariate normal distribution. That is, the model assumes `X` has a Gaussian mixture distribution (`gmdistribution`).
  - For linear discriminant analysis, the model has the same covariance matrix for each class, only the means vary.
  - For quadratic discriminant analysis, both means and covariances of each class vary.

`predict` classifies so as to minimize the expected classification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{k = 1}^{K} \widehat{P}(k|x) C(y|k),$$

where

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(k|x)$ is the posterior probability on page 20-6 of class k for observation x.
- $C(y|k)$ is the cost on page 20-7 of classifying an observation as y when its true class is k.

For details, see “Prediction Using Discriminant Analysis Models” on page 20-6.

Tips

After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Alternative Functionality

Functions

The `classify` function also performs discriminant analysis. `classify` is usually more awkward to use.

- `classify` requires you to fit the classifier every time you make a new prediction.
- `classify` does not perform cross-validation or hyperparameter optimization.
- `classify` requires you to fit the classifier when changing prior probabilities.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Supported syntaxes are:
  - `Mdl = fitcdiscr(Tbl,Y)`
  - `Mdl = fitcdiscr(X,Y)`
  - `Mdl = fitcdiscr(___,Name,Value)`
  - `[Mdl,FitInfo,HyperparameterOptimizationResults] = fitcdiscr(___,Name,Value)` — `fitcdiscr` returns the additional output arguments `FitInfo` and `HyperparameterOptimizationResults` when you specify the `'OptimizeHyperparameters'` name-value pair argument.
- The `FitInfo` output argument is an empty structure array currently reserved for possible future use.
- The `HyperparameterOptimizationResults` output argument is a `BayesianOptimization` object or a table of hyperparameters with associated values that describe the cross-validation optimization of hyperparameters.

  `'HyperparameterOptimizationResults'` is nonempty when the `'OptimizeHyperparameters'` name-value pair argument is nonempty at the time you create the model. The values in `'HyperparameterOptimizationResults'` depend on the value you specify for the `'HyperparameterOptimizationOptions'` name-value pair argument when you create the model.
  - If you specify `'bayesopt'` (default), then `HyperparameterOptimizationResults` is an object of class `BayesianOptimization`.
  - If you specify `'gridsearch'` or `'randomsearch'`, then `HyperparameterOptimizationResults` is a table of the hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst).
- Supported name-value pair arguments, and any differences, are:
  - `'ClassNames'`
  - `'Cost'`
  - `'DiscrimType'`
  - `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.
  - `'OptimizeHyperparameters'` — The only eligible parameter to optimize is `'DiscrimType'`. Specifying `'auto'` uses `'DiscrimType'`.
  - `'PredictorNames'`
  - `'Prior'`

- 'ResponseName'
- 'ScoreTransform'
- 'Weights'

- For tall arrays and tall tables, fitcdiscr returns a CompactClassificationDiscriminant object, which contains most of the same properties as a ClassificationDiscriminant object. The main difference is that the compact object is sensitive to memory requirements. The compact object does not include properties that include the data, or that include an array of the same size as the data. The compact object does not contain these ClassificationDiscriminant properties:

  - ModelParameters
  - NumObservations
  - HyperparameterOptimizationResults
  - RowsUsed
  - XCentered
  - W
  - X
  - Y

  Additionally, the compact object does not support these ClassificationDiscriminant methods:

  - compact
  - crossval
  - cvshrink
  - resubEdge
  - resubLoss
  - resubMargin
  - resubPredict

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the 'HyperparameterOptimizationOptions', struct('UseParallel',true) name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

ClassificationDiscriminant | ClassificationPartitionedModel | classify | crossval | predict

Topics

“Discriminant Analysis Classification” on page 20-2
“Improving Discriminant Analysis Models” on page 20-15
“Regularize Discriminant Analysis Classifier” on page 20-21

Introduced in R2014a

fitcecoc

Fit multiclass models for support vector machines or other classifiers

Syntax

```
Mdl = fitcecoc(Tbl,ResponseVarName)
Mdl = fitcecoc(Tbl,formula)
Mdl = fitcecoc(Tbl,Y)

Mdl = fitcecoc(X,Y)

Mdl = fitcecoc(___,Name,Value)
[Mdl,HyperparameterOptimizationResults] = fitcecoc(___,Name,Value)
```

Description

`Mdl = fitcecoc(Tbl,ResponseVarName)` returns a full, trained, multiclass, error-correcting output codes (ECOC) model on page 33-1643 using the predictors in table `Tbl` and the class labels in `Tbl.ResponseVarName`. `fitcecoc` uses $K(K – 1)/2$ binary support vector machine (SVM) models using the one-versus-one coding design on page 33-1640, where K is the number of unique class labels (levels). `Mdl` is a `ClassificationECOC` model.

`Mdl = fitcecoc(Tbl,formula)` returns an ECOC model using the predictors in table `Tbl` and the class labels. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used for training.

`Mdl = fitcecoc(Tbl,Y)` returns an ECOC model using the predictors in table `Tbl` and the class labels in vector `Y`.

`Mdl = fitcecoc(X,Y)` returns a trained ECOC model using the predictors `X` and the class labels `Y`.

`Mdl = fitcecoc(___,Name,Value)` returns an ECOC model with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

For example, specify different binary learners, a different coding design, or to cross-validate. It is good practice to cross-validate using the `Kfold` `Name,Value` pair argument. The cross-validation results determine how well the model generalizes.

`[Mdl,HyperparameterOptimizationResults] = fitcecoc(___,Name,Value)` also returns hyperparameter optimization details when you specify the `OptimizeHyperparameters` name-value pair argument and use linear or kernel binary learners. For other `Learners`, the `HyperparameterOptimizationResults` property of `Mdl` contains the results.

Examples

Train Multiclass Model Using SVM Learners

Train a multiclass error-correcting output codes (ECOC) model using support vector machine (SVM) binary learners.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
```

Train a multiclass ECOC model using the default options.

```
Mdl = fitcecoc(X,Y)
```

```
Mdl =
  ClassificationECOC
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
           BinaryLearners: {3x1 cell}
               CodingName: 'onevsone'


  Properties, Methods
```

`Mdl` is a `ClassificationECOC` model. By default, `fitcecoc` uses SVM binary learners and a one-versus-one coding design. You can access `Mdl` properties using dot notation.

Display the class names and the coding design matrix.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
CodingMat = Mdl.CodingMatrix
```

```
CodingMat = 3×3

     1     1     0
    -1     0     1
     0    -1    -1
```

A one-versus-one coding design for three classes yields three binary learners. The columns of `CodingMat` correspond to the learners, and the rows correspond to the classes. The class order is the same as the order in `Mdl.ClassNames`. For example, `CodingMat(:,1)` is `[1; –1; 0]` and indicates that the software trains the first SVM binary learner using all observations classified as `'setosa'` and `'versicolor'`. Because `'setosa'` corresponds to `1`, it is the positive class; `'versicolor'` corresponds to `–1`, so it is the negative class.

You can access each binary learner using cell indexing and dot notation.

```
Mdl.BinaryLearners{1}   % The first binary learner
```

```
ans =
  CompactClassificationSVM
             ResponseName: 'Y'
```

```
    CategoricalPredictors: []
               ClassNames: [-1 1]
          ScoreTransform: 'none'
                    Beta: [4x1 double]
                    Bias: 1.4505
         KernelParameters: [1x1 struct]


  Properties, Methods
```

Compute the resubstitution classification error.

```
error = resubLoss(Mdl)
```

```
error = 0.0067
```

The classification error on the training data is small, but the classifier might be an overfitted model. You can cross-validate the classifier using `crossval` and compute the cross-validation classification error instead.

Train Multiclass Linear Classification Model

Train an ECOC model composed of multiple binary, linear classification models.

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels. There are more than two classes in the data.

Create a default linear-classification-model template.

```
t = templateLinear();
```

To adjust the default values, see the “Name-Value Pair Arguments” on page 33-6131 on `templateLinear` page.

Train an ECOC model composed of multiple binary, linear classification models that can identify the product given the frequency distribution of words on a documentation web page. For faster training time, transpose the predictor data, and specify that observations correspond to columns.

```
X = X';
rng(1); % For reproducibility
Mdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns')
```

```
Mdl =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [1x13 categorical]
    ScoreTransform: 'none'
    BinaryLearners: {78x1 cell}
      CodingMatrix: [13x78 double]
```

```
  Properties, Methods
```

Alternatively, you can train an ECOC model composed of default linear classification models using `'Learners','Linear'`.

To conserve memory, `fitcecoc` returns trained ECOC models composed of linear classification learners in `CompactClassificationECOC` model objects.

Cross-Validate ECOC Classifier

Cross-validate an ECOC classifier with SVM binary learners, and estimate the generalized classification error.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
rng(1); % For reproducibility
```

Create an SVM template, and standardize the predictors.

```
t = templateSVM('Standardize',true)
```

```
t =
Fit template for classification SVM.

                     Alpha: [0x1 double]
             BoxConstraint: []
                 CacheSize: []
             CachingMethod: ''
                ClipAlphas: []
    DeltaGradientTolerance: []
                   Epsilon: []
              GapTolerance: []
              KKTTolerance: []
            IterationLimit: []
            KernelFunction: ''
               KernelScale: []
              KernelOffset: []
     KernelPolynomialOrder: []
                 NumPrint: []
                       Nu: []
          OutlierFraction: []
         RemoveDuplicates: []
          ShrinkagePeriod: []
                   Solver: ''
          StandardizeData: 1
       SaveSupportVectors: []
           VerbosityLevel: []
                  Version: 2
                   Method: 'SVM'
                     Type: 'classification'
```

`t` is an SVM template. Most of the template object properties are empty. When training the ECOC classifier, the software sets the applicable properties to their default values.

Train the ECOC classifier, and specify the class order.

```
Mdl = fitcecoc(X,Y,'Learners',t,...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a `ClassificationECOC` classifier. You can access its properties using dot notation.

Cross-validate `Mdl` using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

`CVMdl` is a `ClassificationPartitionedECOC` cross-validated ECOC classifier.

Estimate the generalized classification error.

```
genError = kfoldLoss(CVMdl)
```

```
genError = 0.0400
```

The generalized classification error is 4%, which indicates that the ECOC classifier generalizes fairly well.

Estimate Posterior Probabilities Using ECOC Classifier

Train an ECOC classifier using SVM binary learners. First predict the training-sample labels and class posterior probabilities. Then predict the maximum class posterior probability at each point in a grid. Visualize the results.

Load Fisher's iris data set. Specify the petal dimensions as the predictors and the species names as the response.

```
load fisheriris
X = meas(:,3:4);
Y = species;
rng(1); % For reproducibility
```

Create an SVM template. Standardize the predictors, and specify the Gaussian kernel.

```
t = templateSVM('Standardize',true,'KernelFunction','gaussian');
```

`t` is an SVM template. Most of its properties are empty. When the software trains the ECOC classifier, it sets the applicable properties to their default values.

Train the ECOC classifier using the SVM template. Transform classification scores to class posterior probabilities (which are returned by `predict` or `resubPredict`) using the `'FitPosterior'` name-value pair argument. Specify the class order using the `'ClassNames'` name-value pair argument. Display diagnostic messages during training by using the `'Verbose'` name-value pair argument.

```
Mdl = fitcecoc(X,Y,'Learners',t,'FitPosterior',true,...
    'ClassNames',{'setosa','versicolor','virginica'},...
    'Verbose',2);
```

```
Training binary learner 1 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 2
Positive class indices: 1

Fitting posterior probabilities for learner 1 (SVM).
Training binary learner 2 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 1

Fitting posterior probabilities for learner 2 (SVM).
Training binary learner 3 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 2

Fitting posterior probabilities for learner 3 (SVM).
```

`Mdl` is a `ClassificationECOC` model. The same SVM template applies to each binary learner, but you can adjust options for each binary learner by passing in a cell vector of templates.

Predict the training-sample labels and class posterior probabilities. Display diagnostic messages during the computation of labels and class posterior probabilities by using the `'Verbose'` name-value pair argument.

```
[label,~,~,Posterior] = resubPredict(Mdl,'Verbose',1);
```

```
Predictions from all learners have been computed.
Loss for all observations has been computed.
Computing posterior probabilities...
```

```
Mdl.BinaryLoss
```

```
ans =
'quadratic'
```

The software assigns an observation to the class that yields the smallest average binary loss. Because all binary learners are computing posterior probabilities, the binary loss function is `quadratic`.

Display a random set of results.

```
idx = randsample(size(X,1),10,1);
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
table(Y(idx),label(idx),Posterior(idx,:),...
    'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

```
ans=10×3 table
      TrueLabel         PredLabel                    Posterior
    ______________    ______________    ___________________________________

    {'virginica' }    {'virginica' }     0.0039319     0.0039866       0.99208
    {'virginica' }    {'virginica' }      0.017066      0.018262       0.96467
    {'virginica' }    {'virginica' }      0.014947      0.015855        0.9692
    {'versicolor'}    {'versicolor'}    2.2197e-14       0.87318       0.12682
```

```
    {'setosa'    }    {'setosa'    }         0.999    0.00025091    0.00074639
    {'versicolor'}    {'virginica' }    2.2195e-14      0.059427       0.94057
    {'versicolor'}    {'versicolor'}    2.2194e-14       0.97002      0.029984
    {'setosa'    }    {'setosa'    }         0.999     0.0002499    0.00074741
    {'versicolor'}    {'versicolor'}     0.0085638       0.98259     0.0088482
    {'setosa'    }    {'setosa'    }         0.999    0.00025013    0.00074718
```

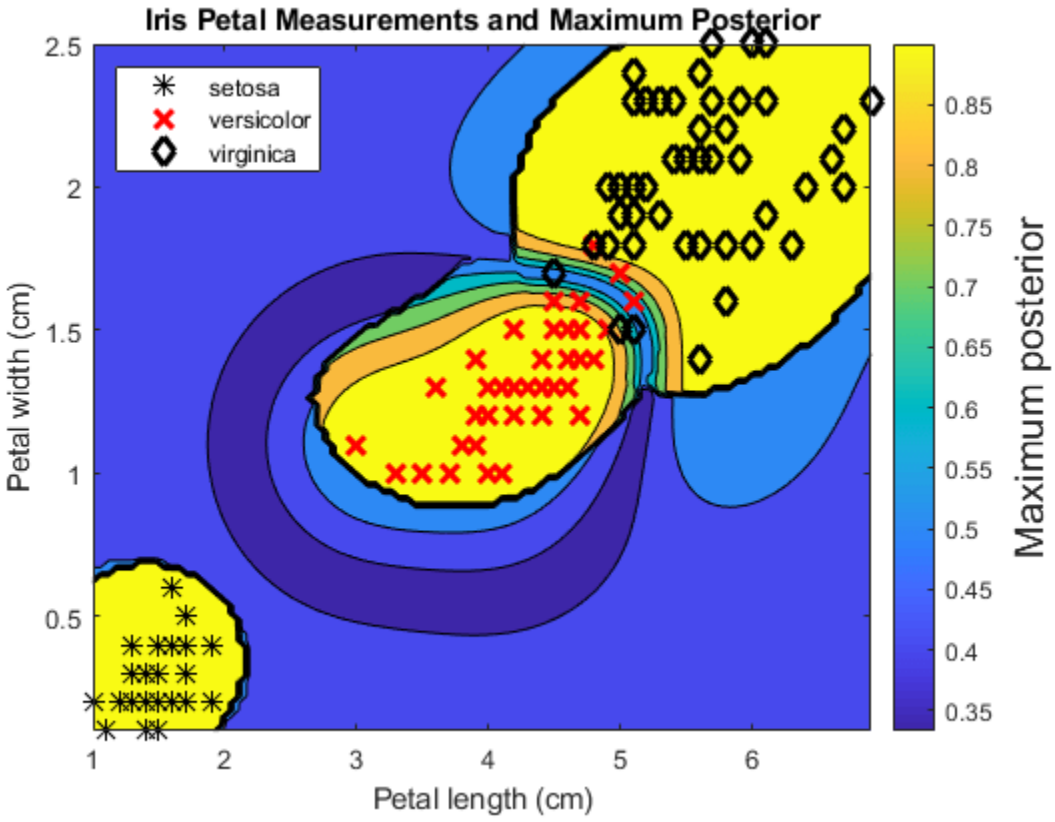
The columns of `Posterior` correspond to the class order of `Mdl.ClassNames`.

Define a grid of values in the observed predictor space. Predict the posterior probabilities for each instance in the grid.

```
xMax = max(X);
xMin = min(X);

x1Pts = linspace(xMin(1),xMax(1));
x2Pts = linspace(xMin(2),xMax(2));
[x1Grid,x2Grid] = meshgrid(x1Pts,x2Pts);

[~,~,~,PosteriorRegion] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
```

For each coordinate on the grid, plot the maximum class posterior probability among all classes.

```
contourf(x1Grid,x2Grid,...
        reshape(max(PosteriorRegion,[],2),size(x1Grid,1),size(x1Grid,2)));
h = colorbar;
h.YLabel.String = 'Maximum posterior';
h.YLabel.FontSize = 15;

hold on
gh = gscatter(X(:,1),X(:,2),Y,'krk','*xd',8);
gh(2).LineWidth = 2;
gh(3).LineWidth = 2;

title('Iris Petal Measurements and Maximum Posterior')
xlabel('Petal length (cm)')
ylabel('Petal width (cm)')
axis tight
legend(gh,'Location','NorthWest')
hold off
```



Speed Up Training ECOC Classifiers Using Binning and Parallel Computing

Train a one-versus-all ECOC classifier using a `GentleBoost` ensemble of decision trees with surrogate splits. To speed up training, bin numeric predictors and use parallel computing. Binning is valid only when `fitcecoc` uses a tree learner. After training, estimate the classification error using 10-fold cross-validation. Note that parallel computing requires Parallel Computing Toolbox™.

Load Sample Data

Load and inspect the `arrhythmia` data set.

```
load arrhythmia
[n,p] = size(X)
```

```
n = 452
```

```
p = 279
```

```
isLabels = unique(Y);
nLabels = numel(isLabels)
```

```
nLabels = 13
```

```
tabulate(categorical(Y))
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

The data set contains 279 predictors, and the sample size of 452 is relatively small. Of the 16 distinct labels, only 13 are represented in the response (Y). Each label describes various degrees of arrhythmia, and 54.20% of the observations are in class 1.

Train One-Versus-All ECOC Classifier

Create an ensemble template. You must specify at least three arguments: a method, a number of learners, and the type of learner. For this example, specify `'GentleBoost'` for the method, `100` for the number of learners, and a decision tree template that uses surrogate splits because there are missing observations.

```
tTree = templateTree('surrogate','on');
tEnsemble = templateEnsemble('GentleBoost',100,tTree);
```

`tEnsemble` is a template object. Most of its properties are empty, but the software fills them with their default values during training.

Train a one-versus-all ECOC classifier using the ensembles of decision trees as binary learners. To speed up training, use binning and parallel computing.

- Binning (`'NumBins',50`) — When you have a large training data set, you can speed up training (a potential decrease in accuracy) by using the `'NumBins'` name-value pair argument. This argument is valid only when `fitcecoc` uses a tree learner. If you specify the `'NumBins'` value, then the software bins every numeric predictor into a specified number of equiprobable bins, and then grows trees on the bin indices instead of the original data. You can try `'NumBins',50` first, and then change the `'NumBins'` value depending on the accuracy and training speed.
- Parallel computing (`'Options',statset('UseParallel',true)`) — With a Parallel Computing Toolbox license, you can speed up the computation by using parallel computing, which sends each binary learner to a worker in the pool. The number of workers depends on your system configuration. When you use decision trees for binary learners, `fitcecoc` parallelizes training using Intel® Threading Building Blocks (TBB) for dual-core systems and above. Therefore, specifying the `'UseParallel'` option is not helpful on a single computer. Use this option on a cluster.

Additionally, specify that the prior probabilities are $1/K$, where $K = 13$ is the number of distinct classes.

```
options = statset('UseParallel',true);
Mdl = fitcecoc(X,Y,'Coding','onevsall','Learners',tEnsemble,...
                'Prior','uniform','NumBins',50,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

`Mdl` is a `ClassificationECOC` model.

Cross-Validation

Cross-validate the ECOC classifier using 10-fold cross-validation.

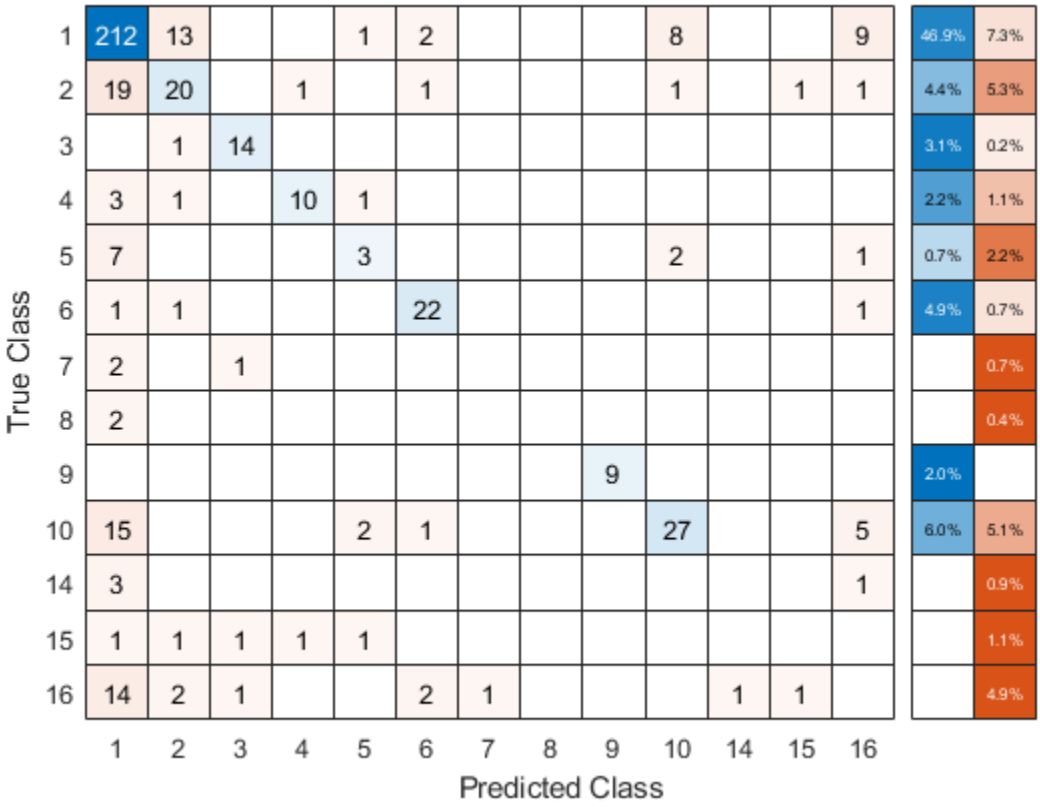
```
CVMdl = crossval(Mdl,'Options',options);
```

```
Warning: One or more folds do not contain points from all the groups.
```

`CVMdl` is a `ClassificationPartitionedECOC` model. The warning indicates that some classes are not represented while the software trains at least one fold. Therefore, those folds cannot predict labels for the missing classes. You can inspect the results of a fold using cell indexing and dot notation. For example, access the results of the first fold by entering `CVMdl.Trained{1}`.

Use the cross-validated ECOC classifier to predict validation-fold labels. You can compute the confusion matrix by using `confusionchart`. Move and resize the chart by changing the inner position property to ensure that the percentages appear in the row summary.

```
oofLabel = kfoldPredict(CVMdl,'Options',options);
ConfMat = confusionchart(Y,oofLabel,'RowSummary','total-normalized');
ConfMat.InnerPosition = [0.10 0.12 0.85 0.85];
```



Reproduce Binned Data

Reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

```
X = Mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = Mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are `0` for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Optimize ECOC Classifier

Optimize hyperparameters automatically using `fitcecoc`.

Load the `fisheriris` data set.

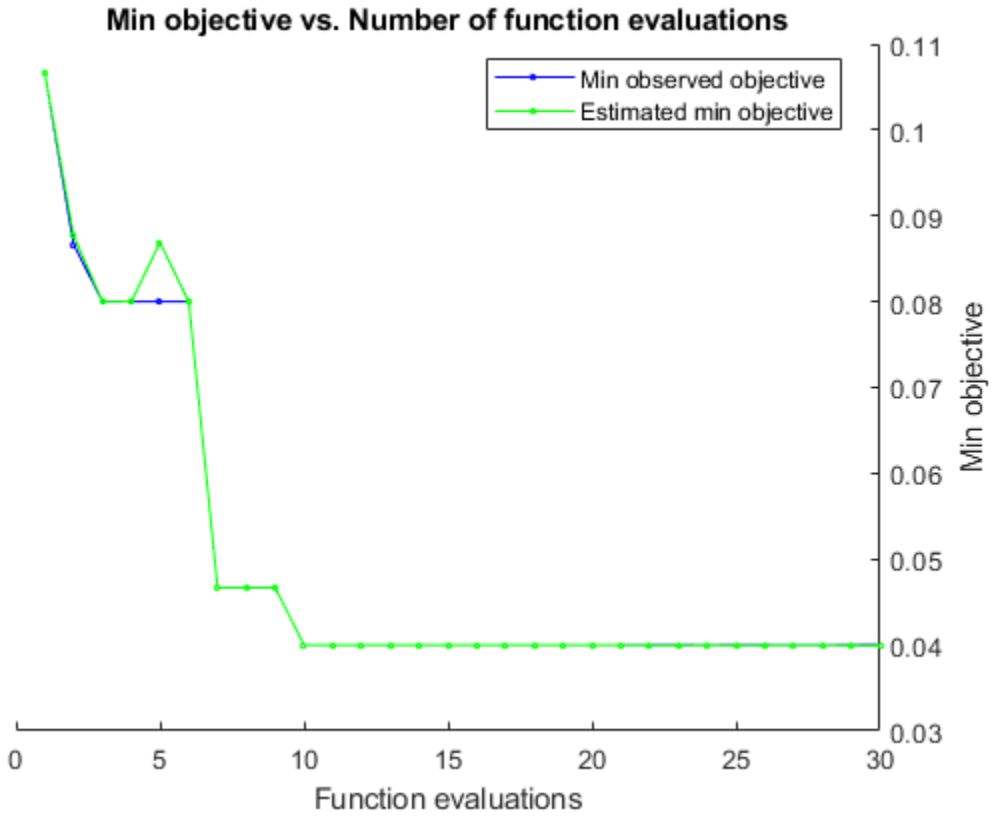
```
load fisheriris
X = meas;
Y = species;
```

Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization. For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
Mdl = fitcecoc(X,Y,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName',...
    'expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Coding | BoxCons
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.10667 |       1.376 |     0.10667 |     0.10667 |     onevsone |       5
|    2 | Best   |    0.086667 |      4.9839 |    0.086667 |    0.087701 |     onevsone |      94
|    3 | Best   |        0.08 |     0.39732 |        0.08 |    0.080045 |     onevsall |     0.0
|    4 | Accept |        0.08 |      0.4637 |        0.08 |    0.080001 |     onevsall |
|    5 | Accept |        0.38 |      21.284 |        0.08 |    0.086871 |     onevsall |      49
```

```
|    6 | Accept |        0.08 |     0.49032 |        0.08 |    0.079988 |     onevsall |     0.0
|    7 | Best   |    0.046667 |      6.3737 |    0.046667 |    0.046675 |     onevsall |    0.00
|    8 | Accept |    0.046667 |      3.1951 |    0.046667 |    0.046674 |     onevsone |      0.5
|    9 | Accept |    0.046667 |     0.45293 |    0.046667 |    0.046671 |     onevsone |     0.0
|   10 | Best   |        0.04 |     0.46167 |        0.04 |    0.040005 |     onevsone |    0.00
|   11 | Accept |     0.33333 |     0.39845 |        0.04 |    0.040007 |     onevsall |       5
|   12 | Accept |     0.10667 |     0.39703 |        0.04 |    0.040003 |     onevsone |       0
|   13 | Accept |     0.10667 |     0.40644 |        0.04 |    0.040015 |     onevsone |    0.00
|   14 | Accept |     0.33333 |     0.41831 |        0.04 |    0.040004 |     onevsall |    0.00
|   15 | Accept |     0.10667 |     0.43089 |        0.04 |    0.040004 |     onevsone |    0.00
|   16 | Accept |        0.06 |     0.47767 |        0.04 |    0.040005 |     onevsall |    0.00
|   17 | Accept |        0.04 |     0.43374 |        0.04 |    0.040007 |     onevsone |      28
|   18 | Accept |        0.04 |     0.42916 |        0.04 |    0.040001 |     onevsone |       1
|   19 | Accept |        0.04 |     0.65399 |        0.04 |    0.040002 |     onevsall |      99
|   20 | Accept |        0.04 |     0.46015 |        0.04 |     0.03999 |     onevsone |       9
|======|========|=============|=============|=============|=============|==============|=========
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Coding | BoxCons
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|======|========|=============|=============|=============|=============|==============|=========
|   21 | Accept |    0.046667 |     0.69592 |        0.04 |    0.039989 |     onevsone |      98
|   22 | Accept |        0.04 |     0.49686 |        0.04 |    0.039951 |     onevsone |     0.0
|   23 | Accept |        0.08 |     0.44333 |        0.04 |    0.039952 |     onevsall |      24
|   24 | Accept |        0.04 |     0.55151 |        0.04 |     0.03996 |     onevsone |    0.00
|   25 | Accept |        0.04 |     0.49853 |        0.04 |    0.039951 |     onevsone |      0.
|   26 | Accept |     0.10667 |     0.42697 |        0.04 |    0.039951 |     onevsone |      90
|   27 | Accept |    0.046667 |      4.6991 |        0.04 |     0.03998 |     onevsall |    0.00
|   28 | Accept |        0.04 |      0.3757 |        0.04 |    0.039975 |     onevsone |      11
|   29 | Accept |        0.04 |     0.37354 |        0.04 |    0.039935 |     onevsone |      21
|   30 | Accept |        0.04 |     0.39876 |        0.04 |    0.039909 |     onevsone |    0.00
```




```
_______________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 94.4407 seconds
Total objective function evaluation time: 52.9448

Best observed feasible point:
     Coding     BoxConstraint    KernelScale
    ________    _____________    ___________

    onevsone      0.0010187       0.0010007

Observed objective function value = 0.04
Estimated objective function value = 0.039964
Function evaluation time = 0.46167

Best estimated feasible point (according to models):
     Coding     BoxConstraint    KernelScale
    ________    _____________    ___________

    onevsone      0.022563        0.002403

Estimated objective function value = 0.039909
Estimated function evaluation time = 0.53445

Mdl =
  ClassificationECOC
```

```
                    ResponseName: 'Y'
           CategoricalPredictors: []
                      ClassNames: {'setosa'  'versicolor'  'virginica'}
                  ScoreTransform: 'none'
                  BinaryLearners: {3x1 cell}
                      CodingName: 'onevsone'
  HyperparameterOptimizationResults: [1x1 BayesianOptimization]


  Properties, Methods
```

Train Multiclass ECOC Model with SVMs and Tall Arrays

Create two multiclass ECOC models trained on tall data. Use linear binary learners for one of the models and kernel binary learners for the other. Compare the resubstitution classification error of the two models.

In general, you can perform multiclass classification of tall data by using `fitcecoc` with linear or kernel binary learners. When you use `fitcecoc` to train a model on tall arrays, you cannot use SVM binary learners directly. However, you can use either linear or kernel binary classification models that use SVMs.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Create a datastore that references the folder containing Fisher's iris data set. Specify `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Create tall versions of the predictor and response data.

```
ds = datastore('fisheriris.csv','TreatAsMissing','NA');
t = tall(ds);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
X = [t.SepalLength t.SepalWidth t.PetalLength t.PetalWidth];
Y = t.Species;
```

Standardize the predictor data.

```
Z = zscore(X);
```

Train a multiclass ECOC model that uses tall data and linear binary learners. By default, when you pass tall arrays to `fitcecoc`, the software trains linear binary learners that use SVMs. Because the response data contains only three unique classes, change the coding scheme from one-versus-all (which is the default when you use tall data) to one-versus-one (which is the default when you use in-memory data).

For reproducibility, set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
mdlLinear = fitcecoc(Z,Y,'Coding','onevsone')
```

```
Training binary learner 1 (Linear) out of 3.
Training binary learner 2 (Linear) out of 3.
Training binary learner 3 (Linear) out of 3.

mdlLinear =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: {'setosa'  'versicolor'  'virginica'}
    ScoreTransform: 'none'
    BinaryLearners: {3×1 cell}
      CodingMatrix: [3×3 double]


  Properties, Methods
```

`mdlLinear` is a `CompactClassificationECOC` model composed of three binary learners.

Train a multiclass ECOC model that uses tall data and kernel binary learners. First, create a `templateKernel` object to specify the properties of the kernel binary learners; in particular, increase the number of expansion dimensions to 2^{16}.

```
tKernel = templateKernel('NumExpansionDimensions',2^16)
```

```
tKernel =
Fit template for classification Kernel.

             BetaTolerance: []
                 BlockSize: []
             BoxConstraint: []
                   Epsilon: []
    NumExpansionDimensions: 65536
         GradientTolerance: []
        HessianHistorySize: []
            IterationLimit: []
               KernelScale: []
                    Lambda: []
                   Learner: 'svm'
              LossFunction: []
                    Stream: []
          VerbosityLevel: []
                   Version: 1
                    Method: 'Kernel'
                      Type: 'classification'
```

By default, the kernel binary learners use SVMs.

Pass the `templateKernel` object to `fitcecoc` and change the coding scheme to one-versus-one.

```
mdlKernel = fitcecoc(Z,Y,'Learners',tKernel,'Coding','onevsone')
```

```
Training binary learner 1 (Kernel) out of 3.
Training binary learner 2 (Kernel) out of 3.
Training binary learner 3 (Kernel) out of 3.
```

```
mdlKernel =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: {'setosa'  'versicolor'  'virginica'}
    ScoreTransform: 'none'
    BinaryLearners: {3×1 cell}
      CodingMatrix: [3×3 double]


  Properties, Methods
```

`mdlKernel` is also a `CompactClassificationECOC` model composed of three binary learners.

Compare the resubstitution classification error of the two models.

```
errorLinear = gather(loss(mdlLinear,Z,Y))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 1.6 sec

errorLinear = 0.0333
```

```
errorKernel = gather(loss(mdlKernel,Z,Y))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 15 sec
Evaluation completed in 16 sec

errorKernel = 0.0067
```

`mdlKernel` misclassifies a smaller percentage of the training data than `mdlLinear`.

Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not accepted.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, specify a formula using `formula`.

If `Tbl` does not contain the response variable, specify a response variable using `Y`. The length of response variable and the number of `Tbl` rows must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels to which the ECOC model is trained, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

If `Y` is a character array, then each element must correspond to one row of the array.

The length of `Y` and the number of rows of `Tbl` or `X` must be equal.

It is good practice to specify the class order using the `ClassNames` name-value pair argument.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
full matrix | sparse matrix

Predictor data, specified as a full or sparse matrix.

The length of `Y` and the number of observations in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Note

- For linear classification learners, if you orient `X` so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you can experience a significant reduction in optimization-execution time.
- For all other learners, orient `X` so that observations correspond to rows.
- `fitcecoc` supports sparse matrices for training linear classification models only.

Data Types: `double` | `single`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data. The software removes rows of `X` corresponding to missing values in `Y`. However, the treatment of missing values in `X` varies among binary learners. For details, see the training functions for your binary learners: `fitcdiscr`, `fitckernel`, `fitcknn`, `fitclinear`, `fitcnb`, `fitcsvm`, `fitctree`, or `fitcensemble`. Removing observations decreases the effective training or cross-validation sample size.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'Learners','tree','Coding','onevsone','CrossVal','on'` specifies to use decision trees for all binary learners, a one-versus-one coding design, and to implement 10-fold cross-validation.

ECOC Classifier Options

`Coding` — Coding design

`'onevsone'` (default) | `'allpairs'` | `'binarycomplete'` | `'denserandom'` | `'onevsall'` | `'ordinal'` | `'sparserandom'` | `'ternarycomplete'` | numeric matrix

Coding design name, specified as the comma-separated pair consisting of `'Coding'` and a numeric matrix or a value in this table.

Value	Number of Binary Learners	Description
`'allpairs'` and `'onevsone'`	$K(K-1)/2$	For each binary learner, one class is positive, another is negative, and the software ignores the rest. This design exhausts all combinations of class pair assignments.
`'binarycomplete'`	$2^{(K-1)}-1$	This design partitions the classes into all binary combinations, and does not ignore any classes. For each binary learner, all class assignments are `-1` and `1` with at least one positive and negative class in the assignment.
`'denserandom'`	Random, but approximately $10\log_2 K$	For each binary learner, the software randomly assigns classes into positive or negative classes, with at least one of each type. For more details, see “Random Coding Design Matrices” on page 33-1646.
`'onevsall'`	K	For each binary learner, one class is positive and the rest are negative. This design exhausts all combinations of positive class assignments.
`'ordinal'`	$K-1$	For the first binary learner, the first class is negative, and the rest positive. For the second binary learner, the first two classes are negative, the rest positive, and so on.
`'sparserandom'`	Random, but approximately $15\log_2 K$	For each binary learner, the software randomly assigns classes as positive or negative with probability 0.25 for each, and ignores classes with probability 0.5. For more details, see “Random Coding Design Matrices” on page 33-1646.

Value	Number of Binary Learners	Description
`'ternarycomplete'`	$\left(3^K - 2^{(K+1)} + 1\right)/2$	This design partitions the classes into all ternary combinations. All class assignments are `0`, `-1`, and `1` with at least one positive and one negative class in the assignment.

You can also specify a coding design using a custom coding matrix. The custom coding matrix is a *K*-by-*L* matrix. Each row corresponds to a class and each column corresponds to a binary learner. The class order (rows) corresponds to the order in `ClassNames`. Compose the matrix by following these guidelines:

- Every element of the custom coding matrix must be `-1`, `0`, or `1`, and the value must correspond to a dichotomous class assignment. This table describes the meaning of `Coding(i,j)`, that is, the class that learner `j` assigns to observations in class `i`.

Value	Dichotomous Class Assignment
−1	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

- Every column must contain at least one `-1` or `1`.
- For all column indices `i`,`j` such that `i` ≠ `j`, `Coding(:,i)` cannot equal `Coding(:,j)` and `Coding(:,i)` cannot equal `-Coding(:,j)`.
- All rows of the custom coding matrix must be different.

For more details on the form of custom coding design matrices, see “Custom Coding Design Matrices” on page 33-1644.

Example: `'Coding','ternarycomplete'`

Data Types: `char` | `string` | `double` | `single` | `int16` | `int32` | `int64` | `int8`

`FitPosterior` — Flag indicating whether to transform scores to posterior probabilities
`false` or `0` (default) | `true` or `1`

Flag indicating whether to transform scores to posterior probabilities, specified as the comma-separated pair consisting of `'FitPosterior'` and a `true` (`1`) or `false` (`0`).

If `FitPosterior` is `true`, then the software transforms binary-learner classification scores to posterior probabilities. You can obtain posterior probabilities by using `kfoldPredict`, `predict`, or `resubPredict`.

`fitcecoc` does not support fitting posterior probabilities if:

- The ensemble method is `AdaBoostM2`, `LPBoost`, `RUSBoost`, `RobustBoost`, or `TotalBoost`.

- The binary learners (`Learners`) are linear or kernel classification models that implement SVM. To obtain posterior probabilities for linear or kernel classification models, implement logistic regression instead.

Example: `'FitPosterior',true`

Data Types: `logical`

`Learners` — Binary learner templates
`'svm'` (default) | `'discriminant'` | `'kernel'` | `'knn'` | `'linear'` | `'naivebayes'` | `'tree'` | template object | cell vector of template objects

Binary learner templates, specified as the comma-separated pair consisting of `'Learners'` and a character vector, string scalar, template object, or cell vector of template objects. Specifically, you can specify binary classifiers such as SVM, and the ensembles that use `GentleBoost`, `LogitBoost`, and `RobustBoost`, to solve multiclass problems. However, `fitcecoc` also supports multiclass models as binary classifiers.

- If `Learners` is a character vector or string scalar, then the software trains each binary learner using the default values of the specified algorithm. This table summarizes the available algorithms.

Value	Description
`'discriminant'`	Discriminant analysis. For default options, see `templateDiscriminant`.
`'kernel'`	Kernel classification model. For default options, see `templateKernel`.
`'knn'`	*k*-nearest neighbors. For default options, see `templateKNN`.
`'linear'`	Linear classification model. For default options, see `templateLinear`.
`'naivebayes'`	Naive Bayes. For default options, see `templateNaiveBayes`.
`'svm'`	SVM. For default options, see `templateSVM`.
`'tree'`	Classification trees. For default options, see `templateTree`.

- If `Learners` is a template object, then each binary learner trains according to the stored options. You can create a template object using:
  - `templateDiscriminant`, for discriminant analysis.
  - `templateEnsemble`, for ensemble learning. You must at least specify the learning method (`Method`), the number of learners (`NLearn`), and the type of learner (`Learners`). You cannot use the `AdaBoostM2` ensemble method for binary learning.
  - `templateKernel`, for kernel classification.
  - `templateKNN`, for *k*-nearest neighbors.
  - `templateLinear`, for linear classification.
  - `templateNaiveBayes`, for naive Bayes.
  - `templateSVM`, for SVM.

  - `templateTree`, for classification trees.
- If `Learners` is a cell vector of template objects, then:
  - Cell *j* corresponds to binary learner *j* (in other words, column *j* of the coding design matrix), and the cell vector must have length *L*. *L* is the number of columns in the coding design matrix. For details, see `Coding`.
  - To use one of the built-in loss functions for prediction, then all binary learners must return a score in the same range. For example, you cannot include default SVM binary learners with default naive Bayes binary learners. The former returns a score in the range (-∞,∞), and the latter returns a posterior probability as a score. Otherwise, you must provide a custom loss as a function handle to functions such as `predict` and `loss`.
  - You cannot specify linear classification model learner templates with any other template.
  - Similarly, you cannot specify kernel classification model learner templates with any other template.

By default, the software trains learners using default SVM templates.

Example: `'Learners','tree'`

`NumBins` — Number of bins for numeric predictors
`[]` (empty) (default) | positive integer scalar

Number of bins for numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar. This argument is valid only when `fitcecoc` uses a tree learner, that is, `'Learners'` is either `'tree'` or a template object created by using `templateTree`, or a template object created by using `templateEnsemble` with tree weak learners.

- If the `'NumBins'` value is empty (default), then `fitcecoc` does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitcecoc` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
  - `fitcecoc` does not bin categorical predictors.

When you use a large training data set, this binning option speeds up training but might cause a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the value depending on the accuracy and training speed.

A trained model stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`NumConcurrent` — Number of binary learners concurrently trained
`1` (default) | positive integer scalar

Number of binary learners concurrently trained, specified as the comma-separated pair consisting of `'NumConcurrent'` and a positive integer scalar. The default value is `1`, which means `fitcecoc` trains the binary learners sequentially.

Note This option applies only when you use `fitcecoc` on tall arrays. See “Tall Arrays” on page 33-1647 for more information.

Data Types: `single` | `double`

`ObservationsIn` — Predictor data observation dimension

`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`.

Note

- For linear classification learners, if you orient `X` so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you can experience a significant reduction in optimization-execution time.
- For all other learners, orient `X` so that observations correspond to rows.

Example: `'ObservationsIn','columns'`

`Verbose` — Verbosity level

`0` (default) | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. `Verbose` controls the amount of diagnostic information per binary learner that the software displays in the Command Window.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software does not display diagnostic information.
`1`	The software displays diagnostic messages every time it trains a new binary learner.
`2`	The software displays extra diagnostic messages every time it trains a new binary learner.

Each binary learner has its own verbosity level that is independent of this name-value pair argument. To change the verbosity level of a binary learner, create a template object and specify the `'Verbose'` name-value pair argument. Then, pass the template object to `fitcecoc` by using the `'Learners'` name-value pair argument.

Example: `'Verbose',1`

Data Types: `double` | `single`

Cross-Validation Options

`CrossVal` — Flag to train cross-validated classifier

`'off'` (default) | `'on'`

Flag to train a cross-validated classifier, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated classifier with 10 folds.

You can override this cross-validation setting using one of the `CVPartition`, `Holdout`, `KFold`, or `Leaveout` name-value pair arguments. You can only use one cross-validation name-value pair argument at a time to create a cross-validated model.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. If you specify `'Leaveout','on'`, then, for each of the *n* observations, where *n* is `size(Mdl.X,1)`, the software:

1. Reserves the observation as validation data, and trains the model using the other *n* – 1 observations
2. Stores the *n* compact, trained models in the cells of a *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Note Leave-one-out is not recommended for cross-validating ECOC models composed of linear or kernel classification model learners.

Example: `'Leaveout','on'`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcecoc` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.

Value	Description
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

Specification of `'CategoricalPredictors'` is appropriate if:

- At least one predictor is categorical and all binary learners are classification trees, naive Bayes learners, SVMs, linear learners, kernel learners, or ensembles of classification trees.
- All predictors are categorical and at least one binary learner is *k*NN.

If you specify `'CategoricalPredictors'` for any other learner, then the software warns that it cannot train that binary learner. For example, the software cannot train discriminant analysis classifiers using categorical predictors.

Each learner identifies and treats categorical predictors in the same way as the fitting function corresponding to the learner. See `'CategoricalPredictors'` of `fitckernel` for kernel learners, `'CategoricalPredictors'` of `fitcknn` for *k*-nearest learners, `'CategoricalPredictors'` of `fitclinear` for linear learners, `'CategoricalPredictors'` of `fitcnb` for naive Bayes learners, `'CategoricalPredictors'` of `fitcsvm` for SVM learners, and `'CategoricalPredictors'` of `fitctree` for tree learners.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure. If you specify:

- The square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- The structure `S`, then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default is `ones(K) - eye(K)`, where K is the number of distinct classes.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `double` | `single` | `struct`

`Options` — Parallel computing options
`[]` (default) | structure array returned by `statset`

Parallel computing options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. These options require Parallel Computing Toolbox. `fitcecoc` uses `'Streams'`, `'UseParallel'`, and `'UseSubtreams'` fields.

This table summarizes the available options.

Option	Description
`'Streams'`	A `RandStream` object or cell array of such objects. If you do not specify `Streams`, the software uses the default stream or streams. If you specify `Streams`, use a single object except when the following are true: • You have an open parallel pool. • `UseParallel` is `true`. • `UseSubstreams` is `false`. In that case, use a cell array of the same size as the parallel pool. If a parallel pool is not open, then the software tries to open one (depending on your preferences), and `Streams` must supply a single random number stream.

Option	Description
'UseParallel'	If you have Parallel Computing Toolbox, then you can invoke a pool of workers by setting 'UseParallel',true. The fitcecoc function sends each binary learner to a worker in the pool. When you use decision trees for binary learners, fitcecoc parallelizes training using Intel Threading Building Blocks (TBB) for dual-core systems and above. Therefore, specifying the 'UseParallel' option is not helpful on a single computer. Use this option on a cluster. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.
'UseSubstreams'	Set to true to compute in parallel using the stream specified by 'Streams'. Default is false. For example, set Streams to a type allowing substreams, such as 'mlfg6331_64' or 'mrg32k3a'.

A best practice to ensure more predictable results is to use parpool and explicitly create a parallel pool before you invoke parallel computing using fitcecoc.

Example: 'Options',statset('UseParallel',true)

Data Types: struct

PredictorNames — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of PredictorNames depends on the way you supply the training data.

- If you supply X and Y, then you can use PredictorNames to assign names to the predictor variables in X.
  - The order of the names in PredictorNames must correspond to the column order of X. That is, PredictorNames{1} is the name of X(:,1), PredictorNames{2} is the name of X(:,2), and so on. Also, size(X,2) and numel(PredictorNames) must be equal.
  - By default, PredictorNames is {'x1','x2',...}.
- If you supply Tbl, then you can use PredictorNames to choose which predictor variables to use in training. That is, fitcecoc uses only the predictor variables in PredictorNames and the response variable during training.
  - PredictorNames must be a subset of Tbl.Properties.VariableNames and cannot include the name of the response variable.
  - By default, PredictorNames contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either 'PredictorNames' or formula, but not both.

Example: 'PredictorNames',
{'SepalLength','SepalWidth','PetalLength','PetalWidth'}

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

For more details on how the software incorporates class prior probabilities, see “Prior Probabilities and Cost” on page 33-1645.

Example: `struct('ClassNames',{{'setosa','versicolor','virginica'}},'ClassProbs',1:3)`

Data Types: `single` | `double` | `char` | `string` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize

`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Coding'}` along with the default parameters for the specified `Learners`:
  - `Learners = 'svm'` (default) — `{'BoxConstraint','KernelScale'}`
  - `Learners = 'discriminant'` — `{'Delta','Gamma'}`
  - `Learners = 'kernel'` — `{'KernelScale','Lambda'}`
  - `Learners = 'knn'` — `{'Distance','NumNeighbors'}`
  - `Learners = 'linear'` — `{'Lambda','Learner'}`
  - `Learners = 'naivebayes'` — `{'DistributionNames','Width'}`
  - `Learners = 'tree'` — `{'MinLeafSize'}`
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`

The optimization attempts to minimize the cross-validation loss (error) for `fitcecoc` by varying the parameters. For information about cross-validation loss in a different context, see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcecoc` are:

- `Coding` — `fitcecoc` searches among `'onevsall'` and `'onevsone'`.
- The eligible hyperparameters for the chosen `Learners`, as specified in this table.

Learners	Eligible Hyperparameters (Bold = Default)	Default Range
`'discriminant'`	**`Delta`**	Log-scaled in the range `[1e-6,1e3]`
	`DiscrimType`	`'linear'`, `'quadratic'`, `'diagLinear'`, `'diagQuadratic'`, `'pseudoLinear'`, and `'pseudoQuadratic'`
	`Gamma`	Real values in `[0,1]`
`'kernel'`	**`Lambda`**	Positive values log-scaled in the range `[1e-3/NumObservations,1e3/NumObservations]`
	`KernelScale`	Positive values log-scaled in the range `[1e-3,1e3]`
	`Learner`	`'svm'` and `'logistic'`
	`NumExpansionDimensions`	Integers log-scaled in the range `[100,10000]`

Learners	Eligible Hyperparameters (Bold = Default)	Default Range
'knn'	**Distance**	'cityblock', 'chebychev', 'correlation', 'cosine', 'euclidean', 'hamming', 'jaccard', 'mahalanobis', 'minkowski', 'seuclidean', and 'spearman'
	DistanceWeight	'equal', 'inverse', and 'squaredinverse'
	Exponent	Positive values in [0.5,3]
	NumNeighbors	Positive integer values log-scaled in the range [1, max(2,round(NumObservations/2))]
	Standardize	'true' and 'false'
'linear'	**Lambda**	Positive values log-scaled in the range [1e-5/NumObservations,1e5/NumObservations]
	Learner	'svm' and 'logistic'
	Regularization	'ridge' and 'lasso'
'naivebayes'	**DistributionNames**	'normal' and 'kernel'
	Width	Positive values log-scaled in the range [MinPredictorDiff/4,max(MaxPredictorRange,MinPredictorDiff)]
	Kernel	'normal', 'box', 'epanechnikov', and 'triangle'
'svm'	**BoxConstraint**	Positive values log-scaled in the range [1e-3,1e3]
	KernelScale	Positive values log-scaled in the range [1e-3,1e3]
	KernelFunction	'gaussian', 'linear', and 'polynomial'
	PolynomialOrder	Integers in the range [2,4]
	Standardize	'true' and 'false'
'tree'	MaxNumSplits	Integers log-scaled in the range [1,max(2,NumObservations-1)]
	MinLeafSize	Integers log-scaled in the range [1,max(2,floor(NumObservations/2))]
	NumVariablesToSample	Integers in the range [1,max(2,NumPredictors)]
	SplitCriterion	'gdi', 'deviance', and 'twoing'

Alternatively, use `hyperparameters` with your chosen `Learners`, such as

```
load fisheriris % hyperparameters requires data and learner
params = hyperparameters('fitcecoc',meas,species,'svm');
```

To see the eligible and default hyperparameters, examine `params`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitcecoc',meas,species,'svm');
params(2).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize ECOC Classifier” on page 33-1614.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`

Field Name	Values	Default
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`

Field Name	Values	Default
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see "Parallel Bayesian Optimization" on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained ECOC model

`ClassificationECOC` model object | `CompactClassificationECOC` model object | `ClassificationPartitionedECOC` cross-validated model object | `ClassificationPartitionedLinearECOC` cross-validated model object | `ClassificationPartitionedKernelECOC` cross-validated model object

Trained ECOC classifier, returned as a `ClassificationECOC` or `CompactClassificationECOC` model object, or a `ClassificationPartitionedECOC`, `ClassificationPartitionedLinearECOC`, or `ClassificationPartitionedKernelECOC` cross-validated model object.

This table shows how the types of model objects returned by `fitcecoc` depend on the type of binary learners you specify and whether you perform cross-validation.

Linear Classification Model Learners	Kernel Classification Model Learners	Cross-Validation	Returned Model Object
No	No	No	`ClassificationECOC`
No	No	Yes	`ClassificationPartitionedECOC`
Yes	No	No	`CompactClassificationECOC`
Yes	No	Yes	`ClassificationPartitionedLinearECOC`
No	Yes	No	`CompactClassificationECOC`
No	Yes	Yes	`ClassificationPartitionedKernelECOC`

`HyperparameterOptimizationResults` — Description of cross-validation optimization of hyperparameters

`BayesianOptimization` object | table of hyperparameters and associated values

Description of the cross-validation optimization of hyperparameters, returned as a `BayesianOptimization` object or a table of hyperparameters and associated values. `HyperparameterOptimizationResults` is nonempty when the `OptimizeHyperparameters` name-value pair argument is nonempty and the `Learners` name-value pair argument designates linear or kernel binary learners. The value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair argument:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observation from smallest (best) to highest (worst)

Data Types: `table`

Limitations

- `fitcecoc` supports sparse matrices for training linear classification models only. For all other models, supply a full matrix of predictor data instead.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).

- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Coding Design

A coding design is a matrix where elements direct which classes are trained by each binary learner, that is, how the multiclass problem is reduced to a series of binary problems.

Each row of the coding design corresponds to a distinct class, and each column corresponds to a binary learner. In a ternary coding design, for a particular column (or binary learner):

- A row containing 1 directs the binary learner to group all observations in the corresponding class into a positive class.
- A row containing –1 directs the binary learner to group all observations in the corresponding class into a negative class.
- A row containing 0 directs the binary learner to ignore all observations in the corresponding class.

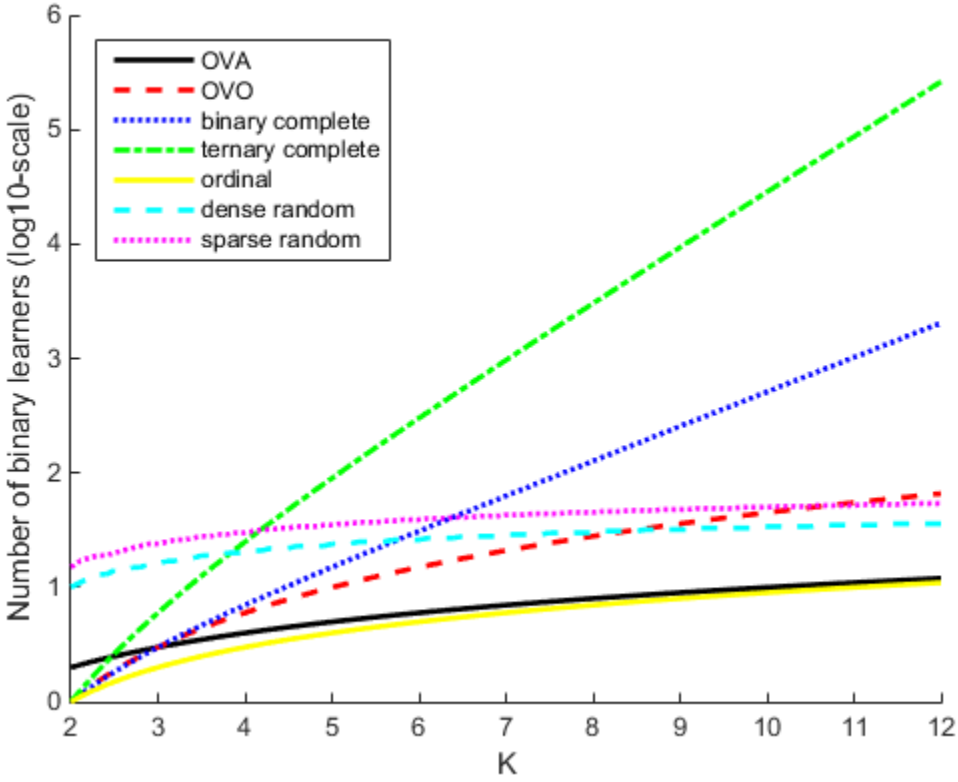
Coding design matrices with large, minimal, pairwise row distances based on the Hamming measure are optimal. For details on the pairwise row distance, see “Random Coding Design Matrices” on page 33-1646 and [4].

This table describes popular coding designs.

Coding Design	Description	Number of Learners	Minimal Pairwise Row Distance
one-versus-all (OVA)	For each binary learner, one class is positive and the rest are negative. This design exhausts all combinations of positive class assignments.	K	2
one-versus-one (OVO)	For each binary learner, one class is positive, another is negative, and the rest are ignored. This design exhausts all combinations of class pair assignments.	$K(K – 1)/2$	1
binary complete	This design partitions the classes into all binary combinations, and does not ignore any classes. That is, all class assignments are `−1` and `1` with at least one positive class and one negative class in the assignment for each binary learner.	$2^{K-1} – 1$	2^{K-2}
ternary complete	This design partitions the classes into all ternary combinations. That is, all class assignments are `0`, `−1`, and `1` with at least one positive class and one negative class in the assignment for each binary learner.	$(3^K – 2^{K+1} + 1)/2$	3^{K-2}

Coding Design	Description	Number of Learners	Minimal Pairwise Row Distance
ordinal	For the first binary learner, the first class is negative and the rest are positive. For the second binary learner, the first two classes are negative and the rest are positive, and so on.	K – 1	1
dense random	For each binary learner, the software randomly assigns classes into positive or negative classes, with at least one of each type. For more details, see “Random Coding Design Matrices” on page 33-1646.	Random, but approximately 10 $\log_2 K$	Variable
sparse random	For each binary learner, the software randomly assigns classes as positive or negative with probability 0.25 for each, and ignores classes with probability 0.5. For more details, see “Random Coding Design Matrices” on page 33-1646.	Random, but approximately 15 $\log_2 K$	Variable

This plot compares the number of binary learners for the coding designs with increasing K.



Error-Correcting Output Codes Model

An error-correcting output codes (ECOC) model reduces the problem of classification with three or more classes to a set of binary classification problems.

ECOC classification requires a coding design, which determines the classes that the binary learners train on, and a decoding scheme, which determines how the results (predictions) of the binary classifiers are aggregated.

Assume the following:

- The classification problem has three classes.
- The coding design is one-versus-one. For three classes, this coding design is

	Learner 1	Learner 2	Learner 3
Class 1	1	1	0
Class 2	−1	0	1
Class 3	0	−1	−1

- The decoding scheme uses loss g.
- The learners are SVMs.

To build this classification model, the ECOC algorithm follows these steps.

1 Learner 1 trains on observations in Class 1 or Class 2, and treats Class 1 as the positive class and Class 2 as the negative class. The other learners are trained similarly.

2 Let M be the coding design matrix with elements m_{kl}, and s_l be the predicted classification score for the positive class of learner l. The algorithm assigns a new observation to the class ($\widehat{k}$) that minimizes the aggregation of the losses for the L binary learners.

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{l=1}^{L} |m_{kl}| g(m_{kl}, s_l)}{\sum_{l=1}^{L} |m_{kl}|}.$$

ECOC models can improve classification accuracy, compared to other multiclass models [2].

Tips

- The number of binary learners grows with the number of classes. For a problem with many classes, the `binarycomplete` and `ternarycomplete` coding designs are not efficient. However:
  - If $K \leq 4$, then use `ternarycomplete` coding design rather than `sparserandom`.
  - If $K \leq 5$, then use `binarycomplete` coding design rather than `denserandom`.

  You can display the coding design matrix of a trained ECOC classifier by entering `Mdl.CodingMatrix` into the Command Window.
- You should form a coding matrix using intimate knowledge of the application, and taking into account computational constraints. If you have sufficient computational power and time, then try several coding matrices and choose the one with the best performance (e.g., check the confusion matrices for each model using `confusionchart`).
- Leave-one-out cross-validation (`Leaveout`) is inefficient for data sets with many observations. Instead, use k-fold cross-validation (`KFold`).
- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

Custom Coding Design Matrices

Custom coding matrices must have a certain form. The software validates custom coding matrices by ensuring:

- Every element is -1, 0, or 1.
- Every column contains as least one -1 and one 1.
- For all distinct column vectors u and v, $u \neq v$ and $u \neq -v$.
- All rows vectors are unique.
- The matrix can separate any two classes. That is, you can travel from any row to any other row following these rules:
  - You can move vertically from 1 to -1 or -1 to 1.
  - You can move horizontally from a nonzero element to another nonzero element.

- You can use a column of the matrix for a vertical move only once.

If it is not possible to move from row i to row j using these rules, then classes i and j cannot be separated by the design. For example, in the coding design

$$\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$$

classes 1 and 2 cannot be separated from classes 3 and 4 (that is, you cannot move horizontally from the -1 in row 2 to column 2 since there is a 0 in that position). Therefore, the software rejects this coding design.

Parallel Computing

If you use parallel computing (see `Options`), then `fitcecoc` trains binary learners in parallel.

Prior Probabilities and Cost

- Prior probabilities — The software normalizes the specified class prior probabilities (`Prior`) for each binary learner. Let M be the coding design matrix and $I(A,c)$ be an indicator matrix. The indicator matrix has the same dimensions as A. If the corresponding element of A is c, then the indicator matrix has elements equaling one, and zero otherwise. Let M_{+1} and M_{-1} be K-by-L matrices such that:
  - $M_{+1} = M \bigcirc I(M,1)$, where $\bigcirc$ is element-wise multiplication (that is, `Mplus = M.*(M == 1)`). Also, let $m_l^{(+1)}$ be column vector l of M_{+1}.
  - $M_{-1} = -M \bigcirc I(M,-1)$ (that is, `Mminus = -M.*(M == -1)`). Also, let $m_l^{(-1)}$ be column vector l of M_{-1}.

  Let $\pi_l^{+1} = m_l^{(+1)} \circ \pi$ and $\pi_l^{-1} = m_l^{(-1)} \circ \pi$, where π is the vector of specified, class prior probabilities (`Prior`).

  Then, the positive and negative, scalar class prior probabilities for binary learner l are

  $$\widehat{\pi}_l^{(j)} = \frac{\left\| \pi_l^{(j)} \right\|_1}{\left\| \pi_l^{(+1)} \right\|_1 + \left\| \pi_l^{(-1)} \right\|_1},$$

  where $j = \{-1,1\}$ and $\|a\|_1$ is the one-norm of a.
- Cost — The software normalizes the K-by-K cost matrix C (`Cost`) for each binary learner. For binary learner l, the cost of classifying a negative-class observation into the positive class is

  $$c_l^{-+} = \left(\pi_l^{(-1)} \right)^{\top} C \pi_l^{(+1)}.$$

  Similarly, the cost of classifying a positive-class observation into the negative class is

  $$c_l^{+-} = \left(\pi_l^{(+1)} \right)^{\top} C \pi_l^{(-1)}.$$

  The cost matrix for binary learner l is

$$C_l = \begin{bmatrix} 0 & c_l^{-+} \\ c_l^{+-} & 0 \end{bmatrix}.$$

ECOC models accommodate misclassification costs by incorporating them with class prior probabilities. If you specify `Prior` and `Cost`, then the software adjusts the class prior probabilities as follows:

$$\tilde{\pi}_l^{-1} = \frac{c_l^{-+}\hat{\pi}_l^{-1}}{c_l^{-+}\hat{\pi}_l^{-1} + c^{+-}\hat{\pi}_l^{+1}}$$

$$\tilde{\pi}_l^{+1} = \frac{c_l^{+-}\hat{\pi}_l^{+1}}{c_l^{-+}\hat{\pi}_l^{-1} + c^{+-}\hat{\pi}_l^{+1}}.$$

Random Coding Design Matrices

For a given number of classes K, the software generates random coding design matrices as follows.

1. The software generates one of these matrices:
   a. Dense random — The software assigns 1 or –1 with equal probability to each element of the K-by-L_d coding design matrix, where $L_d \approx \lceil 10\log_2 K \rceil$.
   b. Sparse random — The software assigns 1 to each element of the K-by-L_s coding design matrix with probability 0.25, –1 with probability 0.25, and 0 with probability 0.5, where $L_s \approx \lceil 15\log_2 K \rceil$.
2. If a column does not contain at least one 1 and at least one –1, then the software removes that column.
3. For distinct columns u and v, if $u = v$ or $u = -v$, then the software removes v from the coding design matrix.

The software randomly generates 10,000 matrices by default, and retains the matrix with the largest, minimal, pairwise row distance based on the Hamming measure ([4]) given by

$$\Delta(k_1, k_2) = 0.5\sum_{l=1}^{L} \left|m_{k_1 l}\right|\left|m_{k_2 l}\right|\left|m_{k_1 l} - m_{k_2 l}\right|,$$

where $m_{k_j l}$ is an element of coding design matrix j.

Support Vector Storage

By default and for efficiency, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties for all linear SVM binary learners. `fitcecoc` lists `Beta`, rather than `Alpha`, in the model display.

To store `Alpha`, `SupportVectorLabels`, and `SupportVectors`, pass a linear SVM template that specifies storing support vectors to `fitcecoc`. For example, enter:

```
t = templateSVM('SaveSupportVectors',true)
Mdl = fitcecoc(X,Y,'Learners',t);
```

You can remove the support vectors and related values by passing the resulting `ClassificationECOC` model to `discardSupportVectors`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Fürnkranz, Johannes, “Round Robin Classification.” *J. Mach. Learn. Res.*, Vol. 2, 2002, pp. 721–747.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recog. Lett.*, Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Supported syntaxes are:
  - `Mdl = fitcecoc(X,Y)`
  - `Mdl = fitcecoc(X,Y,Name,Value)`
  - `[Mdl,FitInfo,HyperparameterOptimizationResults] = fitcecoc(X,Y,Name,Value)` — `fitcecoc` returns the additional output arguments `FitInfo` and `HyperparameterOptimizationResults` when you specify the `'OptimizeHyperparameters'` name-value pair argument.
- The `FitInfo` output argument is an empty structure array currently reserved for possible future use.
- Options related to cross-validation are not supported. The supported name-value pair arguments are:
  - `'ClassNames'`
  - `'Cost'`
  - `'Coding'` — Default value is `'onevsall'`.
  - `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.
  - `'Learners'` — Default value is `'linear'`. You can specify `'linear'`,`'kernel'`, a `templateLinear` or `templateKernel` object, or a cell array of such objects.

- `'OptimizeHyperparameters'` — When you use linear binary learners, the value of the `'Regularization'` hyperparameter must be `'ridge'`.
- `'Prior'`
- `'Verbose'` — Default value is `1`.
- `'Weights'`

- This additional name-value pair argument is specific to tall arrays:
  - `'NumConcurrent'` — A positive integer scalar specifying the number of binary learners that are trained concurrently by combining file I/O operations. The default value for `'NumConcurrent'` is `1`, which means `fitcecoc` trains the binary learners sequentially. `'NumConcurrent'` is most beneficial when the input arrays cannot fit into the distributed cluster memory. Otherwise, the input arrays can be cached and speedup is negligible.

    If you run your code on Apache Spark™, `NumConcurrent` is upper bounded by the memory available for communications. Check the `'spark.executor.memory'` and `'spark.driver.memory'` properties in your Apache Spark configuration. See `parallel.cluster.Hadoop` (Parallel Computing Toolbox) for more details. For more information on Apache Spark and other execution environments that control where your code runs, see “Extend Tall Arrays with Other Products”.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` option to `true` in one of these ways:

- Set the `'UseParallel'` field of the options structure to `true` using `statset` and specify the `'Options'` name-value pair argument in the call to `fitceoc`.

  For example: `'Options',statset('UseParallel',true)`

  For more information, see the `'Options'` name-value pair argument.
- Perform parallel hyperparameter optimization by using the `'HyperparameterOptions',struct('UseParallel',true)` name-value pair argument in the call to `fitceoc`.

  For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `ClassificationPartitionedKernelECOC` | `ClassificationPartitionedLinearECOC` | `CompactClassificationECOC` | `designecoc` | `loss` | `predict` | `statset`

Topics
“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

fitcensemble

Fit ensemble of learners for classification

Syntax

```
Mdl = fitcensemble(Tbl,ResponseVarName)
Mdl = fitcensemble(Tbl,formula)
Mdl = fitcensemble(Tbl,Y)

Mdl = fitcensemble(X,Y)

Mdl = fitcensemble( ___ ,Name,Value)
```

Description

`Mdl = fitcensemble(Tbl,ResponseVarName)` returns the trained classification ensemble model object (`Mdl`) that contains the results of boosting 100 classification trees and the predictor and response data in the table `Tbl`. `ResponseVarName` is the name of the response variable in `Tbl`. By default, `fitcensemble` uses LogitBoost for binary classification and AdaBoostM2 for multiclass classification.

`Mdl = fitcensemble(Tbl,formula)` applies `formula` to fit the model to the predictor and response data in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`. For example, `'Y~X1+X2+X3'` fits the response variable `Tbl.Y` as a function of the predictor variables `Tbl.X1`, `Tbl.X2`, and `Tbl.X3`.

`Mdl = fitcensemble(Tbl,Y)` treats all variables in the table `Tbl` as predictor variables. `Y` is the array of class labels that is not in `Tbl`.

`Mdl = fitcensemble(X,Y)` uses the predictor data in the matrix `X` and the array of class labels in `Y`.

`Mdl = fitcensemble( ___ ,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments and any of the input arguments in the previous syntaxes. For example, you can specify the number of learning cycles, the ensemble aggregation method, or to implement 10-fold cross-validation.

Examples

Train Classification Ensemble

Create a predictive classification ensemble using all available predictor variables in the data. Then, train another ensemble using fewer predictors. Compare the in-sample predictive accuracies of the ensembles.

Load the `census1994` data set.

```
load census1994
```

Train an ensemble of classification models using the entire data set and default options.

```
Mdl1 = fitcensemble(adultdata,'salary')
```

```
Mdl1 =
  ClassificationEnsemble
             PredictorNames: {1x14 cell}
               ResponseName: 'salary'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
            ScoreTransform: 'none'
           NumObservations: 32561
                NumTrained: 100
                    Method: 'LogitBoost'
              LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}


  Properties, Methods
```

`Mdl` is a `ClassificationEnsemble` model. Some notable characteristics of `Mdl` are:

- Because two classes are represented in the data, LogitBoost is the ensemble aggregation algorithm.
- Because the ensemble aggregation method is a boosting algorithm, classification trees that allow a maximum of 10 splits compose the ensemble.
- One hundred trees compose the ensemble.

Use the classification ensemble to predict the labels of a random set of five observations from the data. Compare the predicted labels with their true values.

```
rng(1) % For reproducibility
[pX,pIdx] = datasample(adultdata,5);
label = predict(Mdl1,pX);
table(label,adultdata.salary(pIdx),'VariableNames',{'Predicted','Truth'})
```

```
ans=5×2 table
    Predicted    Truth
    _________    _____

      <=50K      <=50K
      <=50K      <=50K
      <=50K      <=50K
      <=50K      <=50K
      <=50K      <=50K
```

Train a new ensemble using `age` and `education` only.

```
Mdl2 = fitcensemble(adultdata,'salary ~ age + education');
```

Compare the resubstitution losses between `Mdl1` and `Mdl2`.

```
rsLoss1 = resubLoss(Mdl1)
```

```
rsLoss1 = 0.1058
```

```
rsLoss2 = resubLoss(Mdl2)
```

```
rsLoss2 = 0.2037
```

The in-sample misclassification rate for the ensemble that uses all predictors is lower.

Speed Up Training by Binning Numeric Predictor Values

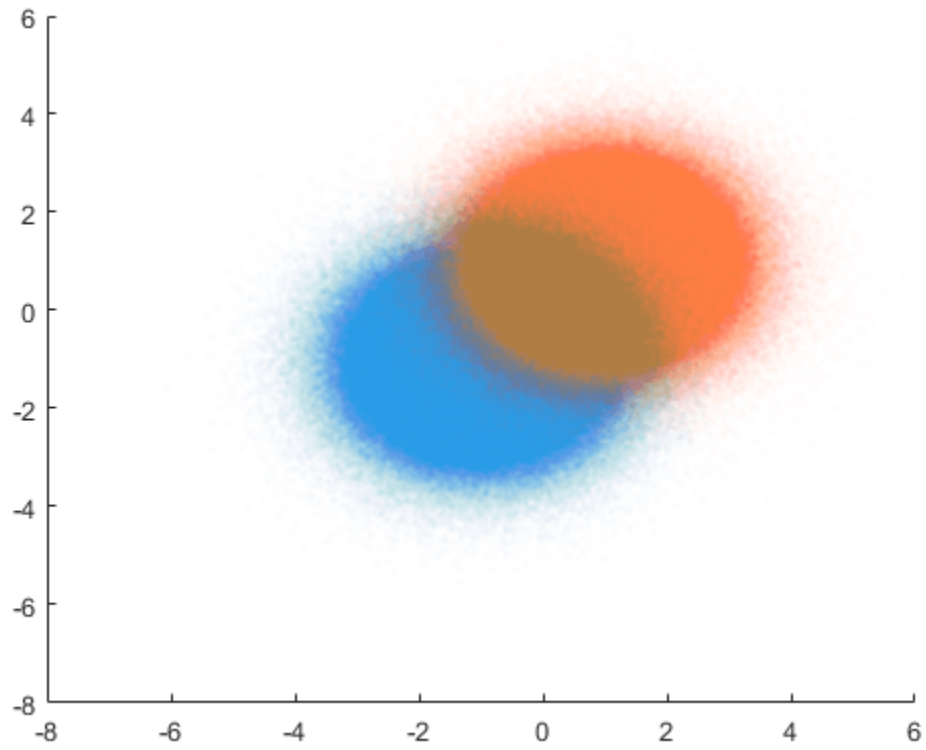
Train an ensemble of boosted classification trees by using `fitcensemble`. Reduce training time by specifying the `'NumBins'` name-value pair argument to bin numeric predictors. This argument is valid only when `fitcensemble` uses a tree learner. After training, you can reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

Generate a sample data set.

```
rng('default') % For reproducibility
N = 1e6;
X = [mvnrnd([-1 -1],eye(2),N); mvnrnd([1 1],eye(2),N)];
y = [zeros(N,1); ones(N,1)];
```

Visualize the data set.

```
figure
scatter(X(1:N,1),X(1:N,2),'Marker','.','MarkerEdgeAlpha',0.01)
hold on
scatter(X(N+1:2*N,1),X(N+1:2*N,2),'Marker','.','MarkerEdgeAlpha',0.01)
```



Train an ensemble of boosted classification trees using adaptive logistic regression (`LogitBoost`, the default for binary classification). Time the function for comparison purposes.

```
tic
Mdl1 = fitcensemble(X,y);
toc
```

```
Elapsed time is 478.988422 seconds.
```

Speed up training by using the `'NumBins'` name-value pair argument. If you specify the `'NumBins'` value as a positive integer scalar, then the software bins every numeric predictor into a specified number of equiprobable bins, and then grows trees on the bin indices instead of the original data. The software does not bin categorical predictors.

```
tic
Mdl2 = fitcensemble(X,y,'NumBins',50);
toc
```

```
Elapsed time is 165.598434 seconds.
```

The process is about three times faster when you use binned data instead of the original data. Note that the elapsed time can vary depending on your operating system.

Compare the classification errors by resubstitution.

```
rsLoss1 = resubLoss(Mdl1)
```

```
rsLoss1 = 0.0788
```

```
rsLoss2 = resubLoss(Mdl2)
```

```
rsLoss2 = 0.0788
```

In this example, binning predictor values reduces training time without loss of accuracy. In general, when you have a large data set like the one in this example, using the binning option speeds up training but causes a potential decrease in accuracy. If you want to reduce training time further, specify a smaller number of bins.

Reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

```
X = Mdl2.X; % Predictor data
Xbinned = zeros(size(X));
edges = Mdl2.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are `0` for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Estimate Generalization Error of Boosting Ensemble

Estimate the generalization error of ensemble of boosted classification trees.

Load the `ionosphere` data set.

```
load ionosphere
```

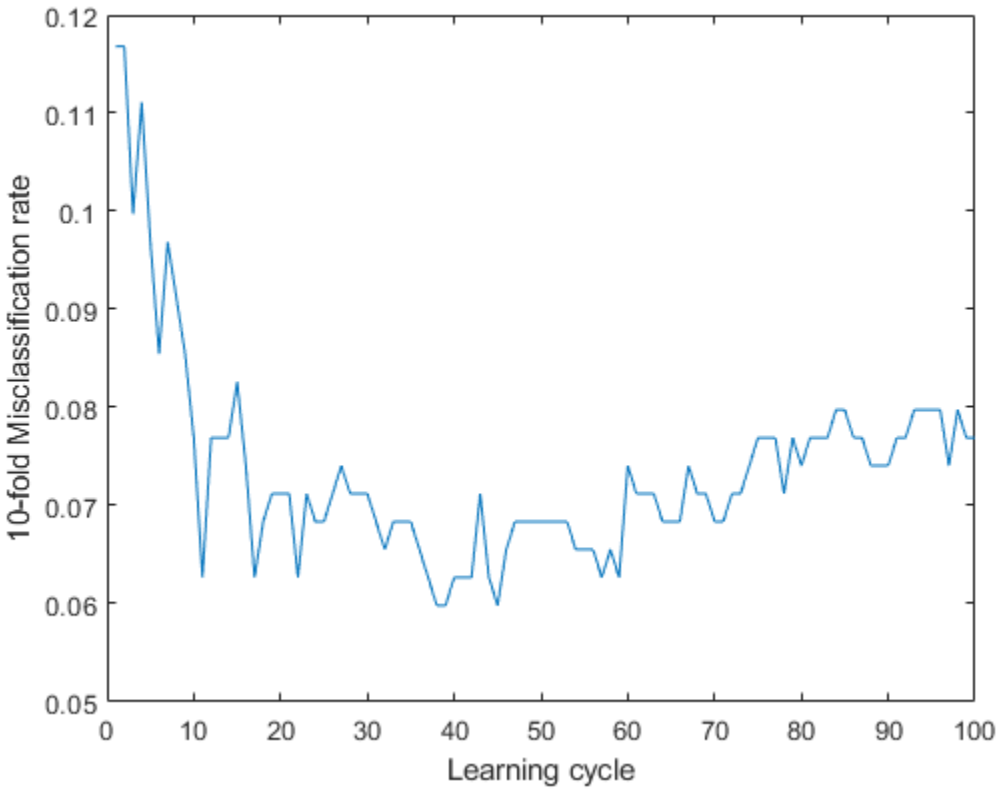
Cross-validate an ensemble of classification trees using AdaBoostM1 and 10-fold cross-validation. Specify that each tree should be split a maximum of five times using a decision tree template.

```
rng(5); % For reproducibility
t = templateTree('MaxNumSplits',5);
Mdl = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t,'CrossVal','on');
```

`Mdl` is a `ClassificationPartitionedEnsemble` model.

Plot the cumulative, 10-fold cross-validated, misclassification rate. Display the estimated generalization error of the ensemble.

```
kflc = kfoldLoss(Mdl,'Mode','cumulative');
figure;
plot(kflc);
ylabel('10-fold Misclassification rate');
xlabel('Learning cycle');
```




```
estGenError = kflc(end)
```

```
estGenError = 0.0769
```

`kfoldLoss` returns the generalization error by default. However, plotting the cumulative loss allows you to monitor how the loss changes as weak learners accumulate in the ensemble.

The ensemble achieves a misclassification rate of around 0.06 after accumulating about 50 weak learners. Then, the misclassification rate increase slightly as more weak learners enter the ensemble.

If you are satisfied with the generalization error of the ensemble, then, to create a predictive model, train the ensemble again using all of the settings except cross-validation. However, it is good practice to tune hyperparameters, such as the maximum number of decision splits per tree and the number of learning cycles.

Optimize Classification Ensemble

Optimize hyperparameters automatically using `fitcensemble`.

Load the `ionosphere` data set.

```
load ionosphere
```

You can find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

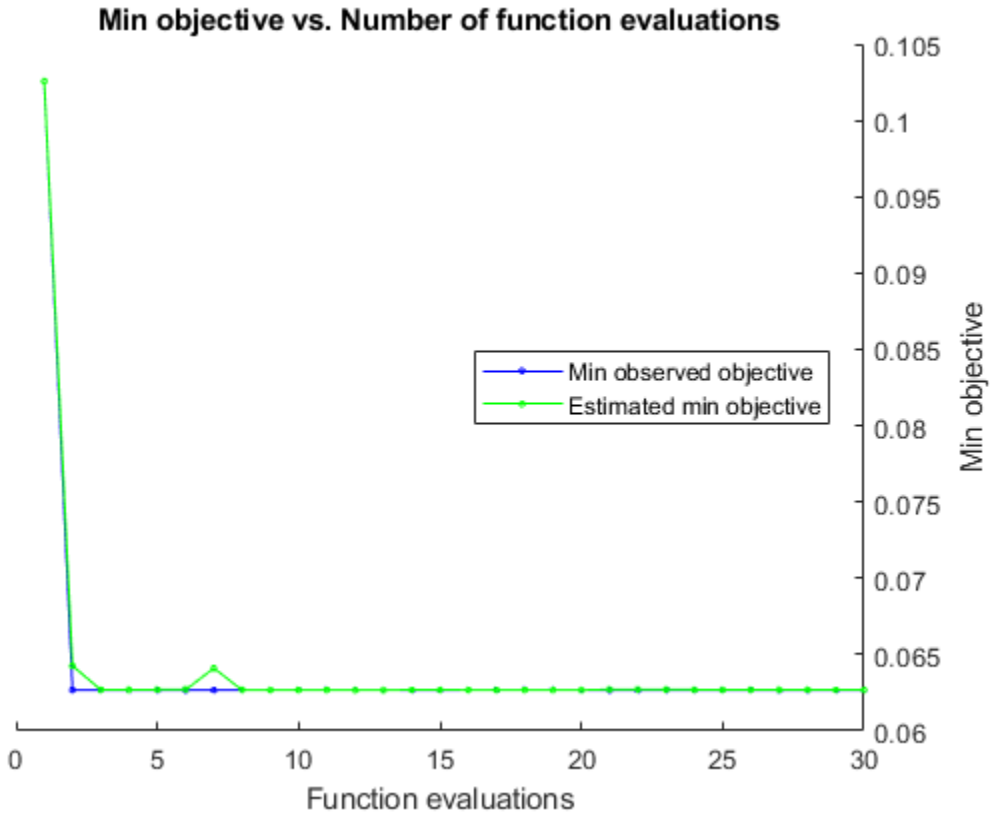
```
Mdl = fitcensemble(X,Y,'OptimizeHyperparameters','auto')
```

In this example, for reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function. Also, for reproducibility of random forest algorithm, specify the `'Reproducible'` name-value pair argument as `true` for tree learners.

```
rng('default')
t = templateTree('Reproducible',true);
Mdl = fitcensemble(X,Y,'OptimizeHyperparameters','auto','Learners',t, ...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName','expected-improvement-p
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result |             | runtime     | (observed)  | (estim.)    |              | ycles
|=====================================================================================================
|    1 | Best   |     0.10256 |      1.6978 |     0.10256 |     0.10256 |     RUSBoost |
|    2 | Best   |    0.062678 |      9.4354 |    0.062678 |    0.064264 |   LogitBoost |
|    3 | Accept |    0.099715 |       7.614 |    0.062678 |    0.062688 |   AdaBoostM1 |
|    4 | Accept |    0.068376 |      1.6045 |    0.062678 |    0.062681 |          Bag |
|    5 | Accept |    0.065527 |      20.359 |    0.062678 |    0.062699 |   LogitBoost |
|    6 | Accept |    0.074074 |      7.1054 |    0.062678 |      0.0627 |  GentleBoost |
|    7 | Accept |    0.082621 |      0.9688 |    0.062678 |    0.064102 |  GentleBoost |
|    8 | Accept |     0.17379 |     0.49564 |    0.062678 |     0.06268 |   LogitBoost |
|    9 | Accept |    0.076923 |      21.003 |    0.062678 |    0.062676 |  GentleBoost |
|   10 | Accept |    0.068376 |      13.575 |    0.062678 |    0.062676 |   AdaBoostM1 |
|   11 | Accept |     0.10541 |      3.6394 |    0.062678 |    0.062676 |     RUSBoost |
|   12 | Accept |    0.068376 |      3.3423 |    0.062678 |    0.062674 |   AdaBoostM1 |
|   13 | Accept |    0.096866 |      1.6005 |    0.062678 |    0.062672 |     RUSBoost |
|   14 | Accept |    0.071225 |       1.201 |    0.062678 |    0.062688 |   LogitBoost |
|   15 | Accept |    0.082621 |     0.87944 |    0.062678 |    0.062687 |   AdaBoostM1 |
|   16 | Accept |    0.079772 |      29.788 |    0.062678 |    0.062679 |   AdaBoostM1 |
|   17 | Accept |     0.35897 |      23.651 |    0.062678 |     0.06267 |          Bag |
|   18 | Accept |    0.074074 |       0.653 |    0.062678 |    0.062674 |          Bag |
|   19 | Accept |    0.088319 |      32.811 |    0.062678 |    0.062674 |     RUSBoost |
|   20 | Accept |    0.068376 |      6.1279 |    0.062678 |    0.062673 |  GentleBoost |
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Method | NumLearn
|      | result |             | runtime     | (observed)  | (estim.)    |              | ycles
```

```
|=====================================================================================================|
|   21 | Accept |      0.17379 |      22.601 |    0.062678 |     0.06271 |   LogitBoost |
|   22 | Accept |     0.071225 |      2.9727 |    0.062678 |    0.062713 |  GentleBoost |
|   23 | Accept |      0.64103 |      1.1288 |    0.062678 |    0.062706 |     RUSBoost |
|   24 | Accept |      0.11111 |       1.537 |    0.062678 |    0.062697 |     RUSBoost |
|   25 | Accept |      0.17379 |      5.5632 |    0.062678 |    0.062686 |   LogitBoost |
|   26 | Accept |      0.35897 |      8.0556 |    0.062678 |    0.062686 |   AdaBoostM1 |
|   27 | Accept |     0.065527 |      1.0791 |    0.062678 |    0.062686 |          Bag |
|   28 | Accept |      0.17379 |      1.7562 |    0.062678 |    0.062689 |   LogitBoost |
|   29 | Accept |     0.074074 |      4.3825 |    0.062678 |    0.062689 |  GentleBoost |
|   30 | Accept |     0.065527 |      1.4893 |    0.062678 |    0.062689 |   LogitBoost |
```




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 278.5509 seconds.
Total objective function evaluation time: 238.1176

Best observed feasible point:
```

```
      Method      NumLearningCycles    LearnRate    MinLeafSize
    ___________    _________________    _________    ___________

    LogitBoost           206            0.96537         33

Observed objective function value = 0.062678
Estimated objective function value = 0.062689
Function evaluation time = 9.4354

Best estimated feasible point (according to models):
      Method      NumLearningCycles    LearnRate    MinLeafSize
    ___________    _________________    _________    ___________

    LogitBoost           206            0.96537         33

Estimated objective function value = 0.062689
Estimated function evaluation time = 9.4324

Mdl =
  ClassificationEnsemble
                         ResponseName: 'Y'
                CategoricalPredictors: []
                           ClassNames: {'b'  'g'}
                       ScoreTransform: 'none'
                      NumObservations: 351
    HyperparameterOptimizationResults: [1×1 BayesianOptimization]
                           NumTrained: 206
                               Method: 'LogitBoost'
                         LearnerNames: {'Tree'}
                 ReasonForTermination: 'Terminated normally after completing the requested numbe
                              FitInfo: [206×1 double]
                   FitInfoDescription: {2×1 cell}


  Properties, Methods
```

The optimization searched over the ensemble aggregation methods for binary classification, over `NumLearningCycles`, over the `LearnRate` for applicable methods, and over the tree learner `MinLeafSize`. The output is the ensemble classifier with the minimum estimated cross-validation loss.

Optimize Classification Ensemble Using Cross-Validation

One way to create an ensemble of boosted classification trees that has satisfactory predictive performance is by tuning the decision tree complexity level using cross-validation. While searching for an optimal complexity level, tune the learning rate to minimize the number of learning cycles.

This example manually finds optimal parameters by using the cross-validation option (the `'KFold'` name-value pair argument) and the `kfoldLoss` function. Alternatively, you can use the `'OptimizeHyperparameters'` name-value pair argument to optimize hyperparameters automatically. See “Optimize Classification Ensemble” on page 33-1654.

Load the `ionosphere` data set.

```
load ionosphere
```

To search for the optimal tree-complexity level:

1. Cross-validate a set of ensembles. Exponentially increase the tree-complexity level for subsequent ensembles from decision stump (one split) to at most n - 1 splits. n is the sample size. Also, vary the learning rate for each ensemble between 0.1 to 1.
2. Estimate the cross-validated misclassification rate of each ensemble.
3. For tree-complexity level j, $j = 1 \ldots J$, compare the cumulative, cross-validated misclassification rate of the ensembles by plotting them against number of learning cycles. Plot separate curves for each learning rate on the same figure.
4. Choose the curve that achieves the minimal misclassification rate, and note the corresponding learning cycle and learning rate.

Cross-validate a deep classification tree and a stump. These classification trees serve as benchmarks.

```
rng(1) % For reproducibility
MdlDeep = fitctree(X,Y,'CrossVal','on','MergeLeaves','off', ...
    'MinParentSize',1);
MdlStump = fitctree(X,Y,'MaxNumSplits',1,'CrossVal','on');
```

Cross-validate an ensemble of 150 boosted classification trees using 5-fold cross-validation. Using a tree template, vary the maximum number of splits using the values in the sequence $\{3^0, 3^1, \ldots, 3^m\}$. m is such that 3^m is no greater than n - 1. For each variant, adjust the learning rate using each value in the set {0.1, 0.25, 0.5, 1};

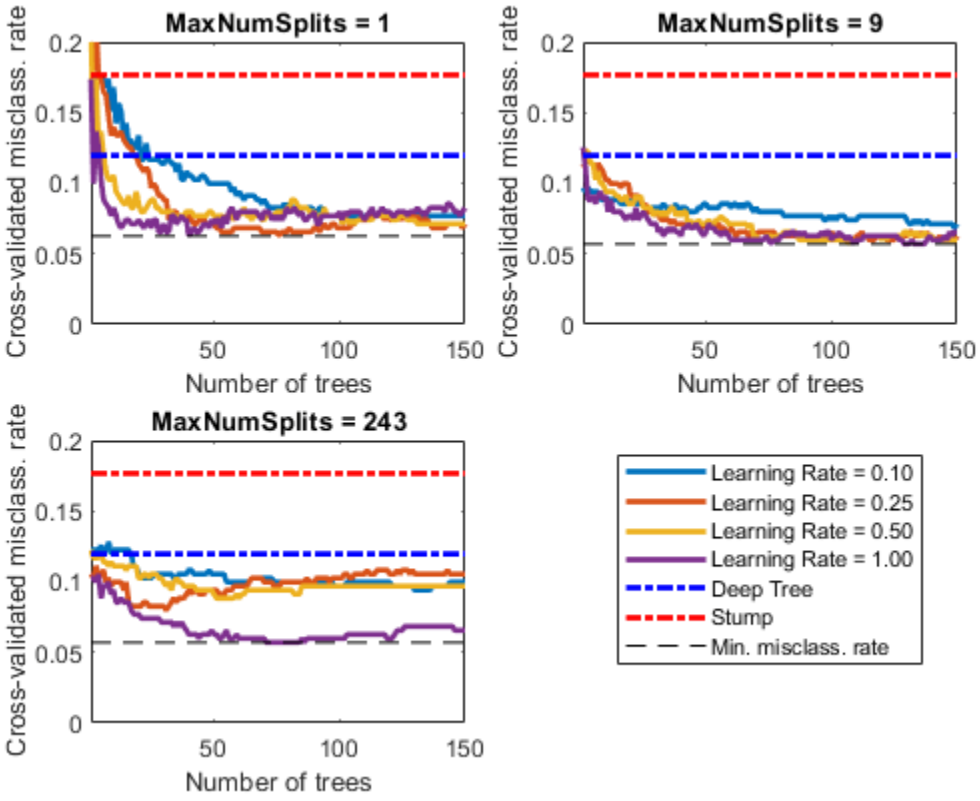
```
n = size(X,1);
m = floor(log(n - 1)/log(3));
learnRate = [0.1 0.25 0.5 1];
numLR = numel(learnRate);
maxNumSplits = 3.^(0:m);
numMNS = numel(maxNumSplits);
numTrees = 150;
Mdl = cell(numMNS,numLR);

for k = 1:numLR
    for j = 1:numMNS
        t = templateTree('MaxNumSplits',maxNumSplits(j));
        Mdl{j,k} = fitcensemble(X,Y,'NumLearningCycles',numTrees,...
            'Learners',t,'KFold',5,'LearnRate',learnRate(k));
    end
end
```

Estimate the cumulative, cross-validated misclassification rate for each ensemble and the classification trees serving as benchmarks.

```
kflAll = @(x)kfoldLoss(x,'Mode','cumulative');
errorCell = cellfun(kflAll,Mdl,'Uniform',false);
error = reshape(cell2mat(errorCell),[numTrees numel(maxNumSplits) numel(learnRate)]);
errorDeep = kfoldLoss(MdlDeep);
errorStump = kfoldLoss(MdlStump);
```

Plot how the cross-validated misclassification rate behaves as the number of trees in the ensemble increases. Plot the curves with respect to learning rate on the same plot, and plot separate plots for varying tree-complexity levels. Choose a subset of tree complexity levels to plot.

```
mnsPlot = [1 round(numel(maxNumSplits)/2) numel(maxNumSplits)];
figure
for k = 1:3
    subplot(2,2,k)
    plot(squeeze(error(:,mnsPlot(k),:)),'LineWidth',2)
    axis tight
    hold on
    h = gca;
    plot(h.XLim,[errorDeep errorDeep],'-.b','LineWidth',2)
    plot(h.XLim,[errorStump errorStump],'-.r','LineWidth',2)
    plot(h.XLim,min(min(error(:,mnsPlot(k),:))).*[1 1],'--k')
    h.YLim = [0 0.2];
    xlabel('Number of trees')
    ylabel('Cross-validated misclass. rate')
    title(sprintf('MaxNumSplits = %0.3g', maxNumSplits(mnsPlot(k))))
    hold off
end
hL = legend([cellstr(num2str(learnRate','Learning Rate = %0.2f')); ...
        'Deep Tree';'Stump';'Min. misclass. rate']);
hL.Position(1) = 0.6;
```



Each curve contains a minimum cross-validated misclassification rate occurring at the optimal number of trees in the ensemble.

Identify the maximum number of splits, number of trees, and learning rate that yields the lowest misclassification rate overall.

```
[minErr,minErrIdxLin] = min(error(:));
[idxNumTrees,idxMNS,idxLR] = ind2sub(size(error),minErrIdxLin);

fprintf('\nMin. misclass. rate = %0.5f',minErr)

Min. misclass. rate = 0.05413

fprintf('\nOptimal Parameter Values:\nNum. Trees = %d',idxNumTrees);

Optimal Parameter Values:
Num. Trees = 47

fprintf('\nMaxNumSplits = %d\nLearning Rate = %0.2f\n',...
    maxNumSplits(idxMNS),learnRate(idxLR))

MaxNumSplits = 3
Learning Rate = 0.25
```

Create a predictive ensemble based on the optimal hyperparameters and the entire training set.

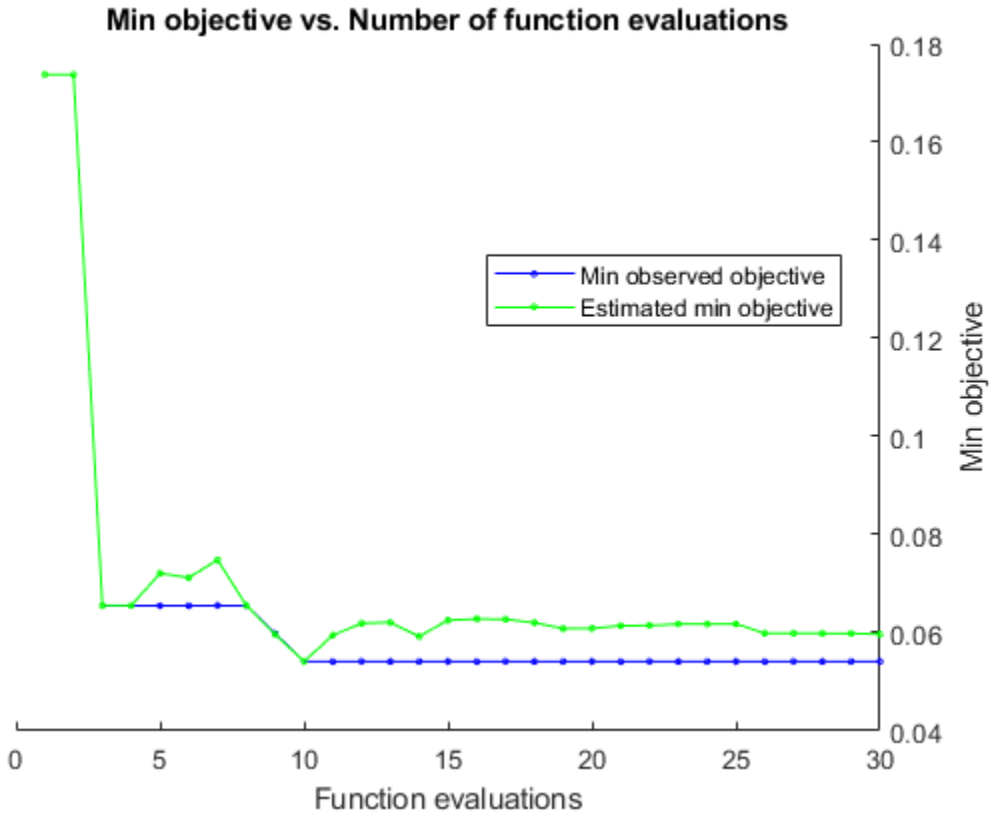
```
tFinal = templateTree('MaxNumSplits',maxNumSplits(idxMNS));
MdlFinal = fitcensemble(X,Y,'NumLearningCycles',idxNumTrees,...
    'Learners',tFinal,'LearnRate',learnRate(idxLR))

MdlFinal =
  ClassificationEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
               NumTrained: 47
                   Method: 'LogitBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [47×1 double]
       FitInfoDescription: {2×1 cell}


  Properties, Methods
```

`MdlFinal` is a `ClassificationEnsemble`. To predict whether a radar return is good given predictor data, you can pass the predictor data and `MdlFinal` to `predict`.

Instead of searching optimal values manually by using the cross-validation option (`'KFold'`) and the `kfoldLoss` function, you can use the `'OptimizeHyperparameters'` name-value pair argument. When you specify `'OptimizeHyperparameters'`, the software finds optimal parameters automatically using Bayesian optimization. The optimal values obtained by using `'OptimizeHyperparameters'` can be different from those obtained using manual search.

```
mdl = fitcensemble(X,Y,'OptimizeHyperparameters',{'NumLearningCycles','LearnRate','MaxNumSplits'

|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | NumLearningC-|    Learn
|      | result |             | runtime     | (observed)  | (estim.)    | ycles        |
|=====================================================================================================
|    1 | Best   |     0.17379 |      6.2592 |     0.17379 |     0.17379 |          137 |     0.00
```

```
|    2 | Accept |     0.17379 |     0.79961 |     0.17379 |     0.17379 |           15 |     0.0
|    3 | Best   |    0.065527 |      1.4585 |    0.065527 |    0.065538 |           31 |      0.4
|    4 | Accept |    0.074074 |      13.988 |    0.065527 |    0.065549 |          340 |      0.9
|    5 | Accept |    0.088319 |     0.92718 |    0.065527 |    0.072102 |           22 |       0
|    6 | Accept |    0.074074 |     0.44748 |    0.065527 |    0.071237 |           10 |       0
|    7 | Accept |     0.08547 |     0.52207 |    0.065527 |    0.074847 |           10 |      0.5
|    8 | Accept |    0.074074 |     0.59154 |    0.065527 |    0.065556 |           11 |      0.9
|    9 | Best   |    0.059829 |      1.6809 |    0.059829 |    0.059648 |           42 |      0.9
|   10 | Best   |    0.054131 |      2.2481 |    0.054131 |    0.054148 |           49 |      0.9
|   11 | Accept |    0.065527 |      2.1686 |    0.054131 |    0.059479 |           48 |       0
|   12 | Accept |    0.068376 |      2.5909 |    0.054131 |    0.061923 |           58 |      0.9
|   13 | Accept |     0.17379 |     0.48621 |    0.054131 |    0.062113 |           10 |    0.001
|   14 | Accept |     0.17379 |     0.55949 |    0.054131 |    0.059231 |           10 |     0.05
|   15 | Accept |    0.065527 |      1.9568 |    0.054131 |    0.062559 |           46 |      0.7
|   16 | Accept |    0.065527 |      2.5692 |    0.054131 |    0.062807 |           57 |      0.6
|   17 | Accept |     0.17379 |     0.55723 |    0.054131 |    0.062748 |           10 |    0.003
|   18 | Accept |     0.12821 |      1.9669 |    0.054131 |    0.062043 |           47 |     0.05
|   19 | Accept |     0.05698 |      1.2814 |    0.054131 |    0.060837 |           27 |      0.9
|   20 | Accept |    0.059829 |      1.1975 |    0.054131 |    0.060881 |           26 |      0.9
|======|========|=============|=============|=============|=============|==============|=========
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | NumLearningC-|    Learn
|      | result |             | runtime     | (observed)  | (estim.)    | ycles        |
|======|========|=============|=============|=============|=============|==============|=========
|   21 | Accept |    0.065527 |      1.2255 |    0.054131 |    0.061441 |           25 |      0.9
|   22 | Accept |     0.17379 |      1.3748 |    0.054131 |    0.061461 |           29 |    0.003
|   23 | Accept |    0.068376 |       3.055 |    0.054131 |    0.061768 |           67 |      0.1
|   24 | Accept |    0.059829 |      5.0035 |    0.054131 |    0.061785 |          119 |       0
|   25 | Accept |    0.059829 |      7.6141 |    0.054131 |    0.061793 |          176 |      0.2
|   26 | Accept |    0.059829 |      5.1133 |    0.054131 |     0.05988 |          115 |      0.1
|   27 | Accept |    0.059829 |      7.4027 |    0.054131 |    0.059895 |          178 |      0.2
|   28 | Accept |    0.059829 |      5.2506 |    0.054131 |    0.059872 |          118 |      0.1
|   29 | Accept |    0.062678 |      10.523 |    0.054131 |    0.059871 |          238 |      0.2
|   30 | Accept |     0.14815 |     0.57384 |    0.054131 |    0.059705 |           10 |      0.1
```




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 122.7983 seconds.
Total objective function evaluation time: 91.3933

Best observed feasible point:
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           49              0.97807          37

Observed objective function value = 0.054131
Estimated objective function value = 0.062545
Function evaluation time = 2.2481

Best estimated feasible point (according to models):
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           119              0.3125           1

Estimated objective function value = 0.059705
Estimated function evaluation time = 5.1842

mdl =
  ClassificationEnsemble
```

```
                      ResponseName: 'Y'
             CategoricalPredictors: []
                        ClassNames: {'b'  'g'}
                   ScoreTransform: 'none'
                  NumObservations: 351
HyperparameterOptimizationResults: [1×1 BayesianOptimization]
                       NumTrained: 119
                           Method: 'LogitBoost'
                     LearnerNames: {'Tree'}
            ReasonForTermination: 'Terminated normally after completing the requested numbe
                         FitInfo: [119×1 double]
               FitInfoDescription: {2×1 cell}


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable and you want to use all remaining variables as predictors, then specify the response variable using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use a subset of the remaining variables only as predictors, then specify a formula using `formula`.
- If `Tbl` does not contain the response variable, then specify the response data using `Y`. The length of response variable and the number of rows of `Tbl` must be equal.

Note To save memory and execution time, supply `X` and `Y` instead of `Tbl`.

Data Types: `table`

`ResponseVarName` — Response variable name

name of response variable in `Tbl`

Response variable name, specified as the name of the response variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl.Y` is the response variable, then specify `ResponseVarName` as `'Y'`. Otherwise, `fitcensemble` treats all columns of `Tbl` as predictor variables.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

For classification, you can specify the order of the classes using the `ClassNames` name-value pair argument. Otherwise, `fitcensemble` determines the class order, and stores it in the `Mdl.ClassNames`.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`X` — Predictor data
numeric matrix

Predictor data, specified as numeric matrix.

Each row corresponds to one observation, and each column corresponds to one predictor variable.

The length of `Y` and the number of rows of `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

`Y` — Response data
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Response data, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. Each entry in `Y` is the response to or label for the observation in the corresponding row of `X` or `Tbl`. The length of `Y` and the number of rows of `X` or `Tbl` must be equal. If the response variable is a character array, then each element must correspond to one row of the array.

You can specify the order of the classes using the `ClassNames` name-value pair argument. Otherwise, `fitcensemble` determines the class order, and stores it in the `Mdl.ClassNames`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'CrossVal','on','LearnRate',0.05` specifies to implement 10-fold cross-validation and to use `0.05` as the learning rate.

General Ensemble Options

`Method` — Ensemble aggregation method
`'Bag'` | `'Subspace'` | `'AdaBoostM1'` | `'AdaBoostM2'` | `'GentleBoost'` | `'LogitBoost'` | `'LPBoost'` | `'RobustBoost'` | `'RUSBoost'` | `'TotalBoost'`

Ensemble aggregation method, specified as the comma-separated pair consisting of `'Method'` and one of the following values.

Value	Method	Classification Problem Support	Related Name-Value Pair Arguments
`'Bag'`	Bootstrap aggregation (bagging, for example, random forest[2]) — If `'Method'` is `'Bag'`, then `fitcensemble` uses bagging with random predictor selections at each split (random forest) by default. To use bagging without the random selections, use tree learners whose `'NumVariablesToSample'` value is `'all'` or use discriminant analysis learners.	Binary and multiclass	N/A
`'Subspace'`	Random subspace	Binary and multiclass	`NPredToSample`
`'AdaBoostM1'`	Adaptive boosting	Binary only	`LearnRate`
`'AdaBoostM2'`	Adaptive boosting	Multiclass only	`LearnRate`
`'GentleBoost'`	Gentle adaptive boosting	Binary only	`LearnRate`
`'LogitBoost'`	Adaptive logistic regression	Binary only	`LearnRate`
`'LPBoost'`	Linear programming boosting — Requires Optimization Toolbox	Binary and multiclass	`MarginPrecision`

Value	Method	Classification Problem Support	Related Name-Value Pair Arguments
`'RobustBoost'`	Robust boosting — Requires Optimization Toolbox	Binary only	`RobustErrorGoal`, `RobustMarginSigma`, `RobustMaxMargin`
`'RUSBoost'`	Random undersampling boosting	Binary and multiclass	`LearnRate`, `RatioToSmallest`
`'TotalBoost'`	Totally corrective boosting — Requires Optimization Toolbox	Binary and multiclass	`MarginPrecision`

You can specify sampling options (`FResample`, `Replace`, `Resample`) for training data when you use bagging (`'Bag'`) or boosting (`'TotalBoost'`, `'RUSBoost'`, `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, `'LogitBoost'`, `'RobustBoost'`, or `'LPBoost'`).

The defaults are:

- `'LogitBoost'` for binary problems and `'AdaBoostM2'` for multiclass problems if `'Learners'` includes only tree learners
- `'AdaBoostM1'` for binary problems and `'AdaBoostM2'` for multiclass problems if `'Learners'` includes both tree and discriminant analysis learners
- `'Subspace'` if `'Learners'` does not include tree learners

For details about ensemble aggregation algorithms and examples, see “Algorithms” on page 33-1684, “Tips” on page 33-1683, “Ensemble Algorithms” on page 18-39, and “Choose an Applicable Ensemble Aggregation Method” on page 18-32.

Example: `'Method','Bag'`

`NumLearningCycles` — Number of ensemble learning cycles
`100` (default) | positive integer | `'AllPredictorCombinations'`

Number of ensemble learning cycles, specified as the comma-separated pair consisting of `'NumLearningCycles'` and a positive integer or `'AllPredictorCombinations'`.

- If you specify a positive integer, then, at every learning cycle, the software trains one weak learner for every template object in `Learners`. Consequently, the software trains `NumLearningCycles*numel(Learners)` learners.
- If you specify `'AllPredictorCombinations'`, then set `Method` to `'Subspace'` and specify one learner only for `Learners`. With these settings, the software trains learners for all possible combinations of predictors taken `NPredToSample` at a time. Consequently, the software trains `nchoosek(size(X,2),NPredToSample)` learners.

The software composes the ensemble using all trained learners and stores them in `Mdl.Trained`.

For more details, see “Tips” on page 33-1683.

Example: `'NumLearningCycles',500`

Data Types: `single` | `double` | `char` | `string`

`Learners` — Weak learners to use in ensemble
`'discriminant'` | `'knn'` | `'tree'` | weak-learner template object | cell vector of weak-learner template objects

Weak learners to use in the ensemble, specified as the comma-separated pair consisting of `'Learners'` and a weak-learner name, weak-learner template object, or cell vector of weak-learner template objects.

Weak Learner	Weak-Learner Name	Template Object Creation Function	Method Setting
Discriminant analysis	`'discriminant'`	`templateDiscriminant`	Recommended for `'Subspace'`
k-nearest neighbors	`'knn'`	`templateKNN`	For `'Subspace'` only
Decision tree	`'tree'`	`templateTree`	All methods except `'Subspace'`

- Weak-learner name (`'discriminant'`, `'knn'`, or `'tree'`) — `fitcensemble` uses weak learners created by a template object creation function with default settings. For example, specifying `'Learners','discriminant'` is the same as specifying `'Learners',templateDiscriminant()`. See the template object creation function pages for the default settings of a weak learner.
- Weak-learner template object — `fitcensemble` uses the weak learners created by a template object creation function. Use the name-value pair arguments of the template object creation function to specify the settings of the weak learners.
- Cell vector of *m* weak-learner template objects — `fitcensemble` grows *m* learners per learning cycle (see `NumLearningCycles`). For example, for an ensemble composed of two types of classification trees, supply `{t1 t2}`, where `t1` and `t2` are classification tree template objects returned by `templateTree`.

The default `'Learners'` value is `'knn'` if `'Method'` is `'Subspace'`.

The default `'Learners'` value is `'tree'` if `'Method'` is `'Bag'` or any boosting method. The default values of `templateTree()` depend on the value of `'Method'`.

- For bagged decision trees, the maximum number of decision splits (`'MaxNumSplits'`) is `n–1`, where `n` is the number of observations. The number of predictors to select at random for each split (`'NumVariablesToSample'`) is the square root of the number of predictors. Therefore, `fitcensemble` grows deep decision trees. You can grow shallower trees to reduce model complexity or computation time.
- For boosted decision trees, `'MaxNumSplits'` is 10 and `'NumVariablesToSample'` is `'all'`. Therefore, `fitcensemble` grows shallow decision trees. You can grow deeper trees for better accuracy.

See `templateTree` for the default settings of a weak learner. To obtain reproducible results, you must specify the `'Reproducible'` name-value pair argument of `templateTree` as `true` if `'NumVariablesToSample'` is not `'all'`.

For details on the number of learners to train, see `NumLearningCycles` and “Tips” on page 33-1683.

Example: `'Learners',templateTree('MaxNumSplits',5)`

`NPrint` — Printout frequency
`'off'` (default) | positive integer

Printout frequency, specified as the comma-separated pair consisting of `'NPrint'` and a positive integer or `'off'`.

To track the number of *weak learners* or *folds* that `fitcensemble` trained so far, specify a positive integer. That is, if you specify the positive integer *m*:

- Without also specifying any cross-validation option (for example, `CrossVal`), then `fitcensemble` displays a message to the command line every time it completes training *m* weak learners.
- And a cross-validation option, then `fitcensemble` displays a message to the command line every time it finishes training *m* folds.

If you specify `'off'`, then `fitcensemble` does not display a message when it completes training weak learners.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Example: `'NPrint',5`

Data Types: `single` | `double` | `char` | `string`

`NumBins` — Number of bins for numeric predictors

`[]` (empty) (default) | positive integer scalar

Number of bins for numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar. This argument is valid only when `fitcensemble` uses a tree learner, that is, `'Learners'` is either `'tree'` or a template object created by using `templateTree`.

- If the `'NumBins'` value is empty (default), then `fitcensemble` does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitcensemble` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
  - `fitcensemble` does not bin categorical predictors.

When you use a large training data set, this binning option speeds up training but might cause a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the value depending on the accuracy and training speed.

A trained model stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcensemble` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

Specification of `'CategoricalPredictors'` is appropriate if:

- `'Learners'` specifies tree learners.
- `'Learners'` specifies *k*-nearest learners where all predictors are categorical.

Each learner identifies and treats categorical predictors in the same way as the fitting function corresponding to the learner. See `'CategoricalPredictors'` of `fitcknn` for *k*-nearest learners and `'CategoricalPredictors'` of `fitctree` for tree learners.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcensemble` uses only the predictor variables in `PredictorNames` and the response variable during training.

- `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
- By default, `PredictorNames` contains the names of all predictor variables.
- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

Parallel Options

`Options` — Options for computing in parallel and setting random numbers
structure

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`

Field Name	Value	Default
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`. Also, use a tree template with the `'Reproducible'` name-value argument set to `true`. See “Reproducibility in Parallel Statistical Computations” on page 31-13.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `fitcensemble` uses the default stream or streams.

For an example using reproducible parallel training, see “Train Classification Ensemble in Parallel” on page 18-109.

For dual-core systems and above, `fitcensemble` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross-validate later by passing `Mdl` to `crossval` or `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other n – 1 observations.
2. Store the n compact, trained models in an n-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Other Classification Options

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost

square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure. If you specify:

- The square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, also specify the `ClassNames` name-value pair argument.
- The structure `S`, then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default is `ones(K) - eye(K)`, where K is the number of distinct classes.

Note `fitcensemble` uses `Cost` to adjust the prior class probabilities specified in `Prior`. Then, `fitcensemble` uses the adjusted prior probabilities for training and resets the cost matrix to its default.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `double` | `single` | `struct`

`Prior` — Prior probabilities

`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to $1/K$, where K is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.
structure array	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

`fitcensemble` normalizes the prior probabilities in `Prior` to sum to 1.

Example: `struct('ClassNames', {{'setosa','versicolor','virginica'}},'ClassProbs',1:3)`

Data Types: `char` | `string` | `double` | `single` | `struct`

`ScoreTransform` — Score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$

Value	Description
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Sampling Options for Boosting Methods and Bagging

`FResample` — Fraction of training set to resample

1 (default) | positive scalar in (0,1]

Fraction of the training set to resample for every weak learner, specified as the comma-separated pair consisting of `'FResample'` and a positive scalar in (0,1].

To use `'FResample'`, specify `'bag'` for `Method` or set `Resample` to `'on'`.

Example: 'FResample',0.75

Data Types: single | double

Replace — Flag indicating to sample with replacement
'on' (default) | 'off'

Flag indicating sampling with replacement, specified as the comma-separated pair consisting of 'Replace' and 'off' or 'on'.

- For 'on', the software samples the training observations with replacement.
- For 'off', the software samples the training observations without replacement. If you set Resample to 'on', then the software samples training observations assuming uniform weights. If you also specify a boosting method, then the software boosts by reweighting observations.

Unless you set Method to 'bag' or set Resample to 'on', Replace has no effect.

Example: 'Replace','off'

Resample — Flag indicating to resample
'off' | 'on'

Flag indicating to resample, specified as the comma-separated pair consisting of 'Resample' and 'off' or 'on'.

- If Method is a boosting method, then:
  - 'Resample','on' specifies to sample training observations using updated weights as the multinomial sampling probabilities.
  - 'Resample','off'(default) specifies to reweight observations at every learning iteration.
- If Method is 'bag', then 'Resample' must be 'on'. The software resamples a fraction of the training observations (see FResample) with or without replacement (see Replace).

If you specify to resample using Resample, then it is good practice to resample to entire data set. That is, use the default setting of 1 for FResample.

AdaBoostM1, AdaBoostM2, LogitBoost, and GentleBoost Method Options

LearnRate — Learning rate for shrinkage
1 (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of 'LearnRate' and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set LearnRate to a value less than 1, for example, 0.1 is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: 'LearnRate',0.1

Data Types: single | double

RUSBoost Method Options

LearnRate — Learning rate for shrinkage
1 (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

`RatioToSmallest` — Sampling proportion with respect to lowest-represented class

positive numeric scalar | numeric vector of positive values

Sampling proportion with respect to the lowest-represented class, specified as the comma-separated pair consisting of `'RatioToSmallest'` and a numeric scalar or numeric vector of positive values with length equal to the number of distinct classes in the training data.

Suppose that there are K classes in the training data and the lowest-represented class has m observations in the training data.

- If you specify the positive numeric scalar s, then `fitcensemble` samples $s*m$ observations from each class, that is, it uses the same sampling proportion for each class. For more details, see “Algorithms” on page 33-2275.
- If you specify the numeric vector $[s1,s2,...,sK]$, then `fitcensemble` samples $si*m$ observations from class i, $i = 1,...,K$. The elements of `RatioToSmallest` correspond to the order of the class names specified using `ClassNames` (see “Tips” on page 33-2274).

The default value is `ones(`K`,1)`, which specifies to sample m observations from each class.

Example: `'RatioToSmallest',[2,1]`

Data Types: `single` | `double`

LPBoost and TotalBoost Method Options

`MarginPrecision` — Margin precision to control convergence speed

`0.1` (default) | numeric scalar in [0,1]

Margin precision to control convergence speed, specified as the comma-separated pair consisting of `'MarginPrecision'` and a numeric scalar in the interval [0,1]. `MarginPrecision` affects the number of boosting iterations required for convergence.

Tip To train an ensemble using many learners, specify a small value for `MarginPrecision`. For training using a few learners, specify a large value.

Example: `'MarginPrecision',0.5`

Data Types: `single` | `double`

RobustBoost Method Options

`RobustErrorGoal` — Target classification error

`0.1` (default) | nonnegative numeric scalar

Target classification error, specified as the comma-separated pair consisting of `'RobustErrorGoal'` and a nonnegative numeric scalar. The upper bound on possible values depends on the values of `RobustMarginSigma` and `RobustMaxMargin`. However, the upper bound cannot exceed `1`.

Tip For a particular training set, usually there is an optimal range for `RobustErrorGoal`. If you set it too low or too high, then the software can produce a model with poor classification accuracy. Try cross-validating to search for the appropriate value.

Example: `'RobustErrorGoal',0.05`

Data Types: `single` | `double`

`RobustMarginSigma` — Classification margin distribution spread

`0.1` (default) | positive numeric scalar

Classification margin distribution spread over the training data, specified as the comma-separated pair consisting of `'RobustMarginSigma'` and a positive numeric scalar. Before specifying `RobustMarginSigma`, consult the literature on `RobustBoost`, for example, [19].

Example: `'RobustMarginSigma',0.5`

Data Types: `single` | `double`

`RobustMaxMargin` — Maximal classification margin

`0` (default) | nonnegative numeric scalar

Maximal classification margin in the training data, specified as the comma-separated pair consisting of `'RobustMaxMargin'` and a nonnegative numeric scalar. The software minimizes the number of observations in the training data having classification margins below `RobustMaxMargin`.

Example: `'RobustMaxMargin',1`

Data Types: `single` | `double`

Random Subspace Method Options

`NPredToSample` — Number of predictors to sample

`1` (default) | positive integer

Number of predictors to sample for each random subspace learner, specified as the comma-separated pair consisting of `'NPredToSample'` and a positive integer in the interval 1,...,*p*, where *p* is the number of predictor variables (`size(X,2)` or `size(Tbl,2)`).

Data Types: `single` | `double`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize

`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.

- `'auto'` — Use `{'Method','NumLearningCycles','LearnRate'}` along with the default parameters for the specified `Learners`:
  - `Learners = 'tree'` (default) — `{'MinLeafSize'}`
  - `Learners = 'discriminant'` — `{'Delta','Gamma'}`
  - `Learners = 'knn'` — `{'Distance','NumNeighbors'}`

  Note For hyperparameter optimization, `Learners` must be a single argument, not a string array or cell array.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`

The optimization attempts to minimize the cross-validation loss (error) for `fitcensemble` by varying the parameters. For information about cross-validation loss (albeit in a different context), see "Classification Loss" on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcensemble` are:

- `Method` — Depends on the number of classes.
  - Two classes — Eligible methods are `'Bag'`, `'GentleBoost'`, `'LogitBoost'`, `'AdaBoostM1'`, and `'RUSBoost'`.
  - Three or more classes — Eligible methods are `'Bag'`, `'AdaBoostM2'`, and `'RUSBoost'`.
- `NumLearningCycles` — `fitcensemble` searches among positive integers, by default log-scaled with range `[10,500]`.
- `LearnRate` — `fitcensemble` searches among positive reals, by default log-scaled with range `[1e-3,1]`.
- The eligible hyperparameters for the chosen `Learners`:

Learners	Eligible Hyperparameters Bold = Used By Default	Default Range
`'discriminant'`	**`Delta`**	Log-scaled in the range `[1e-6,1e3]`
	`DiscrimType`	`'linear'`, `'quadratic'`, `'diagLinear'`, `'diagQuadratic'`, `'pseudoLinear'`, and `'pseudoQuadratic'`
	`Gamma`	Real values in `[0,1]`

Learners	Eligible Hyperparameters Bold = Used By Default	Default Range
'knn'	**Distance**	'cityblock', 'chebychev', 'correlation', 'cosine', 'euclidean', 'hamming', 'jaccard', 'mahalanobis', 'minkowski', 'seuclidean', and 'spearman'
	DistanceWeight	'equal', 'inverse', and 'squaredinverse'
	Exponent	Positive values in [0.5,3]
	NumNeighbors	Positive integer values log-scaled in the range [1, max(2,round(NumObservations/2))]
	Standardize	'true' and 'false'
'tree'	MaxNumSplits	Integers log-scaled in the range [1,max(2,NumObservations-1)]
	MinLeafSize	Integers log-scaled in the range [1,max(2,floor(NumObservations/2))]
	NumVariablesToSample	Integers in the range [1,max(2,NumPredictors)]
	SplitCriterion	'gdi', 'deviance', and 'twoing'

Alternatively, use `hyperparameters` with your chosen `Learners`. Note that you must specify the predictor data and response when creating an `optimizableVariable` object.

```
load fisheriris
params = hyperparameters('fitcensemble',meas,species,'Tree');
```

To see the eligible and default hyperparameters, examine `params`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitcensemble',meas,species,'Tree');
params(4).Range = [1,30];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Classification Ensemble” on page 33-1654.

Example: `'OptimizeHyperparameters',{'Method','NumLearningCycles','LearnRate','MinLeafSize','MaxNumSplits'}`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`

Field Name	Values	Default
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	

Field Name	Values	Default
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained classification ensemble model

`ClassificationBaggedEnsemble` model object | `ClassificationEnsemble` model object | `ClassificationPartitionedEnsemble` cross-validated model object

Trained ensemble model, returned as one of the model objects in this table.

Model Object	Specify Any Cross-Validation Options?	`Method` Setting	`Resample` Setting
`ClassificationBaggedEnsemble`	No	`'Bag'`	`'on'`
`ClassificationEnsemble`	No	Any ensemble aggregation method for classification	`'off'`
`ClassificationPartitionedEnsemble`	Yes	Any ensemble aggregation method for classification	`'off'` or `'on'`

The name-value pair arguments that control cross-validation are `CrossVal`, `Holdout`, `KFold`, `Leaveout`, and `CVPartition`.

To reference properties of `Mdl`, use dot notation. For example, to access or display the cell vector of weak learner model objects for an ensemble that has not been cross-validated, enter `Mdl.Trained` at the command line.

Tips

- `NumLearningCycles` can vary from a few dozen to a few thousand. Usually, an ensemble with good predictive power requires from a few hundred to a few thousand weak learners. However, you do not have to train an ensemble for that many cycles at once. You can start by growing a few dozen learners, inspect the ensemble performance and then, if necessary, train more weak learners using `resume` for classification problems.
- Ensemble performance depends on the ensemble setting and the setting of the weak learners. That is, if you specify weak learners with default parameters, then the ensemble can perform poorly. Therefore, like ensemble settings, it is good practice to adjust the parameters of the weak learners using templates, and to choose values that minimize generalization error.
- If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of `1` for `FResample`.
- If the ensemble aggregation method (`Method`) is `'bag'` and:
  - The misclassification cost (`Cost`) is highly imbalanced, then, for in-bag samples, the software oversamples unique observations from the class that has a large penalty.

- The class prior probabilities (`Prior`) are highly skewed, the software oversamples unique observations from the class that has a large prior probability.

For smaller sample sizes, these combinations can result in a low relative frequency of out-of-bag observations from the class that has a large penalty or prior probability. Consequently, the estimated out-of-bag error is highly variable and it can be difficult to interpret. To avoid large estimated out-of-bag error variances, particularly for small sample sizes, set a more balanced misclassification cost matrix using `Cost` or a less skewed prior probability vector using `Prior`.

- Because the order of some input and output arguments correspond to the distinct classes in the training data, it is good practice to specify the class order using the `ClassNames` name-value pair argument.

  - To determine the class order quickly, remove all observations from the training data that are unclassified (that is, have a missing label), obtain and display an array of all the distinct classes, and then specify the array for `ClassNames`. For example, suppose the response variable (`Y`) is a cell array of labels. This code specifies the class order in the variable `classNames`.

    ```
    Ycat = categorical(Y);
    classNames = categories(Ycat)
    ```

    `categorical` assigns `<undefined>` to unclassified observations and `categories` excludes `<undefined>` from its output. Therefore, if you use this code for cell arrays of labels or similar code for categorical arrays, then you do not have to remove observations with missing labels to obtain a list of the distinct classes.

  - To specify that the class order from lowest-represented label to most-represented, then quickly determine the class order (as in the previous bullet), but arrange the classes in the list by frequency before passing the list to `ClassNames`. Following from the previous example, this code specifies the class order from lowest- to most-represented in `classNamesLH`.

    ```
    Ycat = categorical(Y);
    classNames = categories(Ycat);
    freq = countcats(Ycat);
    [~,idx] = sort(freq);
    classNamesLH = classNames(idx);
    ```

- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- For details of ensemble aggregation algorithms, see “Ensemble Algorithms” on page 18-39.
- If you set `Method` to be a boosting algorithm and `Learners` to be decision trees, then the software grows shallow decision trees by default. You can adjust tree depth by specifying the `MaxNumSplits`, `MinLeafSize`, and `MinParentSize` name-value pair arguments using `templateTree`.
- For bagging (`'Method','Bag'`), `fitcensemble` generates in-bag samples by oversampling classes with large misclassification costs and undersampling classes with small misclassification costs. Consequently, out-of-bag samples have fewer observations from classes with large misclassification costs and more observations from classes with small misclassification costs. If you train a classification ensemble using a small data set and a highly skewed cost matrix, then the number of out-of-bag observations per class can be low. Therefore, the estimated out-of-bag

error can have a large variance and can be difficult to interpret. The same phenomenon can occur for classes with large prior probabilities.

- For the RUSBoost ensemble aggregation method (`'Method','RUSBoost'`), the name-value pair argument `RatioToSmallest` specifies the sampling proportion for each class with respect to the lowest-represented class. For example, suppose that there are two classes in the training data: *A* and *B*. *A* has 100 observations and *B* has 10 observations. Suppose also that the lowest-represented class has *m* observations in the training data.
  - If you set `'RatioToSmallest',2`, then *s***m* = `2*10` = `20`. Consequently, `fitcensemble` trains every learner using 20 observations from class *A* and 20 observations from class *B*. If you set `'RatioToSmallest',[2 2]`, then you obtain the same result.
  - If you set `'RatioToSmallest',[2,1]`, then *s1***m* = `2*10` = `20` and *s2***m* = `1*10` = `10`. Consequently, `fitcensemble` trains every learner using 20 observations from class *A* and 10 observations from class *B*.
- For dual-core systems and above, `fitcensemble` parallelizes training using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L. “Bagging Predictors.” *Machine Learning*. Vol. 26, pp. 123–140, 1996.

[2] Breiman, L. “Random Forests.” *Machine Learning*. Vol. 45, pp. 5–32, 2001.

[3] Freund, Y. “A more robust boosting algorithm.” *arXiv:0905.2138v1*, 2009.

[4] Freund, Y. and R. E. Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *J. of Computer and System Sciences*, Vol. 55, pp. 119–139, 1997.

[5] Friedman, J. “Greedy function approximation: A gradient boosting machine.” *Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.

[6] Friedman, J., T. Hastie, and R. Tibshirani. “Additive logistic regression: A statistical view of boosting.” *Annals of Statistics*, Vol. 28, No. 2, pp. 337–407, 2000.

[7] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning* section edition, Springer, New York, 2008.

[8] Ho, T. K. “The random subspace method for constructing decision forests.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832–844, 1998.

[9] Schapire, R. E., Y. Freund, P. Bartlett, and W.S. Lee. “Boosting the margin: A new explanation for the effectiveness of voting methods.” *Annals of Statistics*, Vol. 26, No. 5, pp. 1651–1686, 1998.

[10] Seiffert, C., T. Khoshgoftaar, J. Hulse, and A. Napolitano. “RUSBoost: Improving classification performance when training data is skewed.” *19th International Conference on Pattern Recognition*, pp. 1–4, 2008.

[11] Warmuth, M., J. Liao, and G. Ratsch. “Totally corrective boosting algorithms that maximize the margin.” *Proc. 23rd Int’l. Conf. on Machine Learning, ACM*, New York, pp. 1001–1008, 2006.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`fitcensemble` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationBaggedEnsemble` | `ClassificationEnsemble` | `ClassificationPartitionedEnsemble` | `predict` | `templateDiscriminant` | `templateKNN` | `templateTree`

Topics

“Supervised Learning Workflow and Algorithms” on page 18-3
“Framework for Ensemble Learning” on page 18-31
“Ensemble Algorithms” on page 18-39

Introduced in R2016b

fitcgam

Fit generalized additive model (GAM) for binary classification

Syntax

```
Mdl = fitcgam(Tbl,ResponseVarName)
Mdl = fitcgam(Tbl,formula)
Mdl = fitcgam(Tbl,Y)
Mdl = fitcgam(X,Y)
Mdl = fitcgam(___,Name,Value)
```

Description

`Mdl = fitcgam(Tbl,ResponseVarName)` returns a generalized additive model on page 33-1707 `Mdl` trained using the sample data contained in the table `Tbl`. The input argument `ResponseVarName` is the name of the variable in `Tbl` that contains the class labels for binary classification.

`Mdl = fitcgam(Tbl,formula)` uses the model specification argument `formula` to specify the class labels and predictor variables in `Tbl`. You can specify a subset of predictor variables and interaction terms for predictor variables by using `formula`.

`Mdl = fitcgam(Tbl,Y)` uses the predictor variables in the table `Tbl` and the class labels in the vector `Y`.

`Mdl = fitcgam(X,Y)` uses the predictors in the matrix `X` and the class labels in the vector `Y`.

`Mdl = fitcgam(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in the previous syntaxes. For example, `'Interactions',5` specifies to include five interaction terms in the model. You can also specify a list of interaction terms using the `'Interactions'` name-value argument.

Examples

Train Generalized Additive Model

Train a univariate generalized additive model, which contains linear terms for predictors. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a univariate GAM that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

```
Mdl = fitcgam(X,Y)
```

```
Mdl =
  ClassificationGAM
```

```
           ResponseName: 'Y'
  CategoricalPredictors: []
             ClassNames: {'b'  'g'}
         ScoreTransform: 'logit'
              Intercept: 2.2715
        NumObservations: 351


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. The model display shows a partial list of the model properties. To view the full list of properties, double-click the variable name `Mdl` in the Workspace. The Variables editor opens for `Mdl`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the class order of `Mdl`.

```
classOrder = Mdl.ClassNames
```

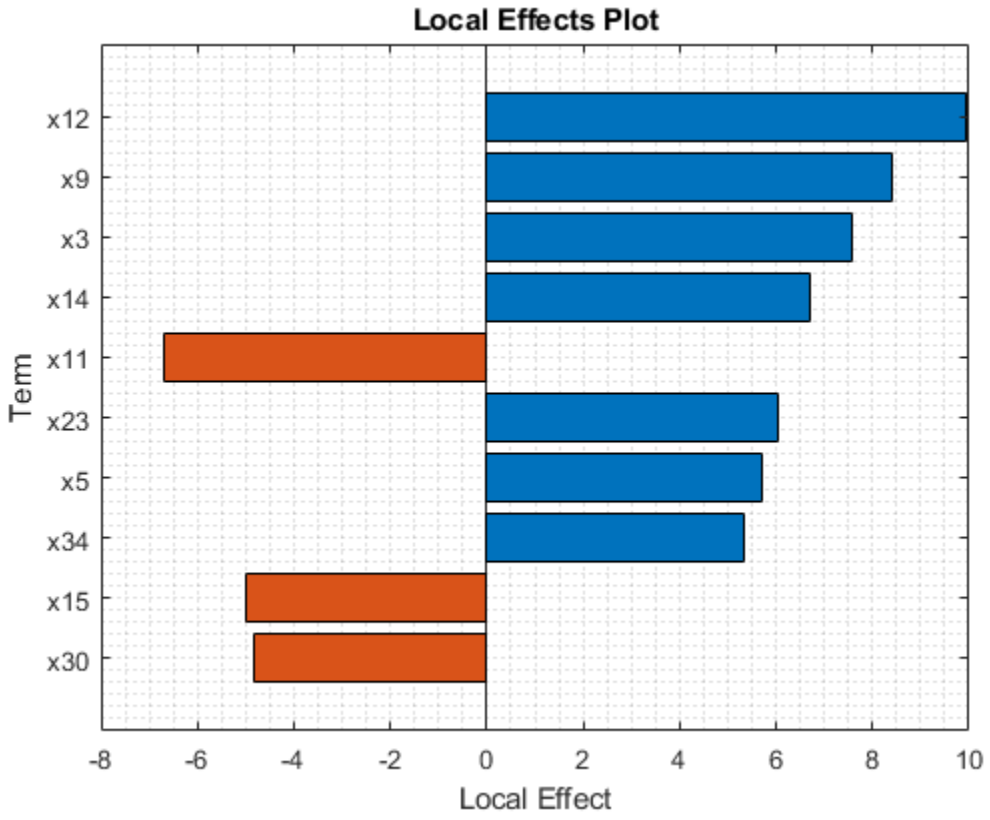
```
classOrder = 2x1 cell
    {'b'}
    {'g'}
```

Classify the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction.

```
label = predict(Mdl,X(1,:))
```

```
label = 1x1 cell array
    {'g'}
```

```
plotLocalEffects(Mdl,X(1,:))
```



The `predict` function classifies the first observation `X(1,:)` as `'g'`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the 10 most important terms on the prediction. Each local effect value shows the contribution of each term to the classification score for `'g'`, which is the logit of the posterior probability that the classification is `'g'` for the observation.

Train GAM with Interaction Terms

Train a generalized additive model that contains linear and interaction terms for predictors in three different ways:

- Specify the interaction terms using the `formula` input argument.
- Specify the `'Interactions'` name-value argument.
- Build a model with linear terms first and add interaction terms to the model by using the `addInteractions` function.

Load Fisher's iris data set. Create a table that contains observations for versicolor and virginica.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
tbl = array2table(meas(inds,:),'VariableNames',["x1","x2","x3","x4"]);
tbl.Y = species(inds,:);
```

Specify `formula`

Train a GAM that contains the four linear terms (`x1`, `x2`, `x3`, and `x4`) and two interaction terms (`x1*x2` and `x2*x3`). Specify the terms using a formula in the form `'Y ~ terms'`.

```
Mdl1 = fitcgam(tbl,'Y ~ x1 + x2 + x3 + x4 + x1:x2 + x2:x3');
```

The function adds interaction terms to the model in the order of importance. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitcgam` adds them to the model. Display the `Interactions` property.

```
Mdl1.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term.

Specify `'Interactions'`

Pass the training data (`tbl`) and the name of the response variable in `tbl` to `fitcgam`, so that the function includes the linear terms for all the other variables as predictors. Specify the `'Interactions'` name-value argument using a logical matrix to include the two interaction terms, `x1*x2` and `x2*x3`.

```
Mdl2 = fitcgam(tbl,'Y','Interactions',logical([1 1 0 0; 0 1 1 0]));
Mdl2.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

You can also specify `'Interactions'` as the number of interaction terms or as `'all'` to include all available interaction terms. Among the specified interaction terms, `fitcgam` identifies those whose *p*-values are not greater than the `'MaxPValue'` value and adds them to the model. The default `'MaxPValue'` is 1 so that the function adds all specified interaction terms to the model.

Specify `'Interactions','all'` and set the `'MaxPValue'` name-value argument to 0.01.

```
Mdl3 = fitcgam(tbl,'Y','Interactions','all','MaxPValue',0.01);
Mdl3.Interactions
```

```
ans = 5×2

     3     4
     2     4
     1     4
     2     3
     1     3
```

`Mdl3` includes five of the six available pairs of interaction terms.

Use `addInteractions` Function

Train a univariate GAM that contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function. Specify the second input argument of `addInteractions` in the same way you specify the `'Interactions'` name-value argument of `fitcgam`. You can specify the list of interaction terms using a logical matrix, the number of interaction terms, or `'all'`.

Specify the number of interaction terms as 5 to add the five most important interaction terms to the trained model.

```
Mdl4 = fitcgam(tbl,'Y');
UpdatedMdl4 = addInteractions(Mdl4,5);
UpdatedMdl4.Interactions
```

```
ans = 5×2

     3     4
     2     4
     1     4
     2     3
     1     3
```

`Mdl4` is a univariate GAM, and `UpdatedMdl4` is an updated GAM that contains all the terms in `Mdl4` and five additional interaction terms.

Create Cross-Validated GAM Using `fitcgam`

Train a cross-validated GAM with 10 folds, which is the default cross-validation option, by using `fitcgam`. Then, use `kfoldPredict` to predict class labels for validation-fold observations using a model trained on training-fold observations.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`.

```
rng('default') % For reproducibility
CVMdl = fitcgam(X,Y,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedGAM
    CrossValidatedModel: 'GAM'
         PredictorNames: {1x34 cell}
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
      NumTrainedPerFold: [1x1 struct]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'logit'
```

```
  Properties, Methods
```

The `fitcgam` function creates a `ClassificationPartitionedGAM` model object `CVMdl` with 10 folds. During cross-validation, the software completes these steps:

1. Randomly partition the data into 10 sets.
2. For each set, reserve the set as validation data, and train the model using the other 9 sets.
3. Store the 10 compact, trained models in a 10-by-1 cell vector in the `Trained` property of the cross-validated model object `ClassificationPartitionedGAM`.

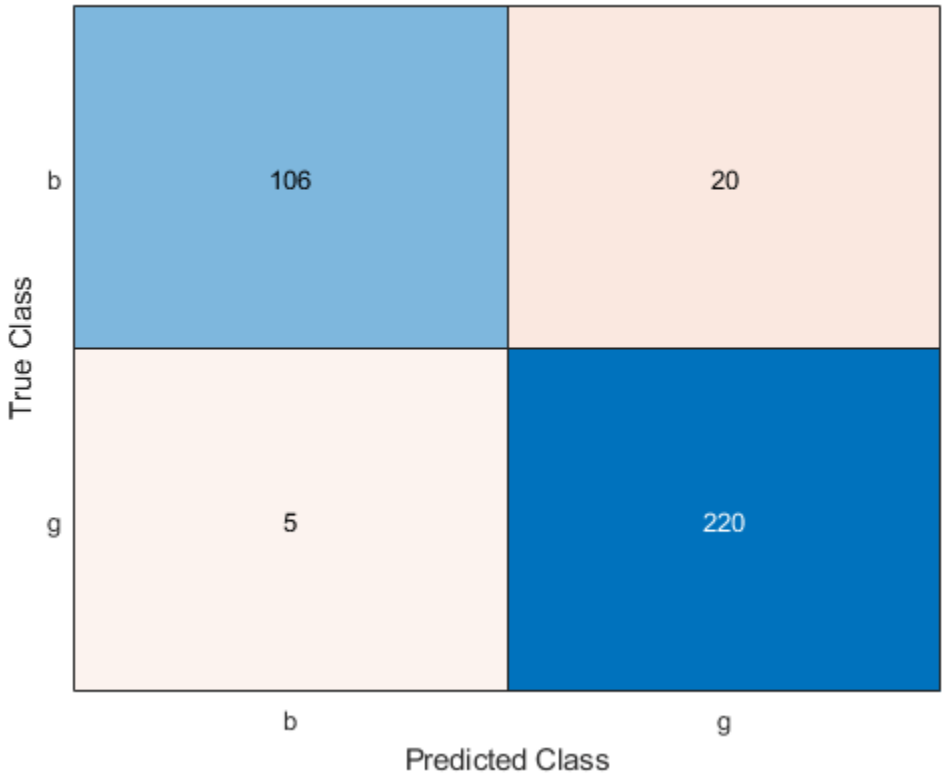
You can override the default cross-validation setting by using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument.

Classify the observations in `X` by using `kfoldPredict`. The function predicts class labels for every observation using the model trained without that observation.

```
label = kfoldPredict(CVMdl);
```

Create a confusion matrix to compare the true classes of the observations to their predicted labels.

```
C = confusionchart(Y,label);
```



Compute the classification error.

```
L = kfoldLoss(CVMdl)
```

```
L = 0.0712
```

The average misclassification rate over 10 folds is about 7%.

Optimize Cross-Validated GAM Using bayesopt

Optimize the parameters of a GAM with respect to cross-validation by using the `bayesopt` function.

Load the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

```
load census1994
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. To reduce the running time for this example, subsample 500 training observations from `adultdata` by using the `datasample` function.

```
rng('default')
NumSamples = 5e2;
adultdata = datasample(adultdata,NumSamples,'Replace',false);
```

Prepare `optimizableVariable` objects for the name-value arguments that you want to optimize using Bayesian optimization. This example finds optimal values for the `MaxNumSplitsPerPredictor` and `NumTreesPerPredictor` arguments of `fitcgam`.

```
maxNumSplits = optimizableVariable('maxNumSplits',[1,10],'Type','integer');
numTrees = optimizableVariable('numTrees',[1,500],'Type','integer');
```

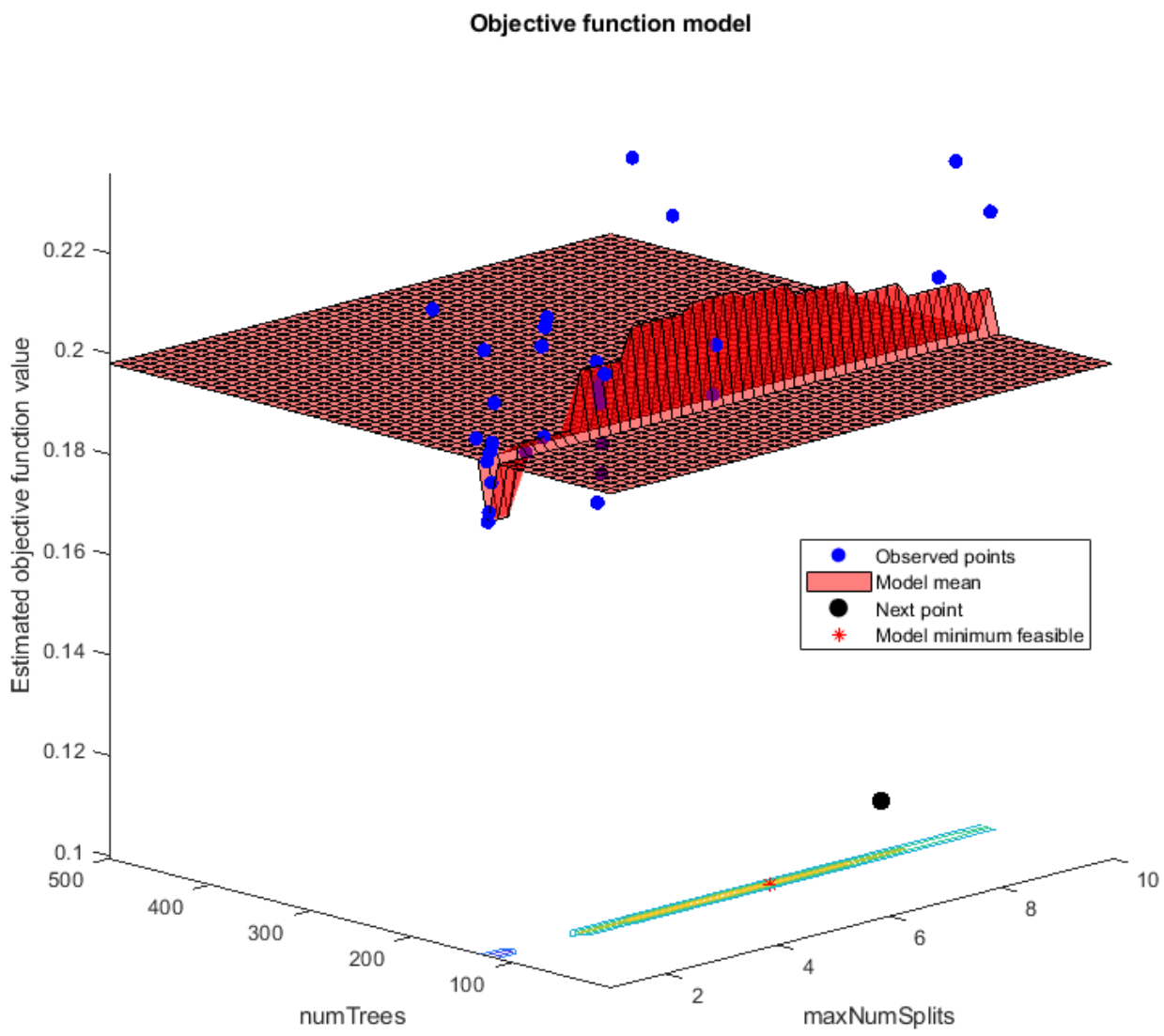
Create an objective function that takes an input `z = [maxNumSplits,numTrees]` and returns the cross-validated loss value of `z`.

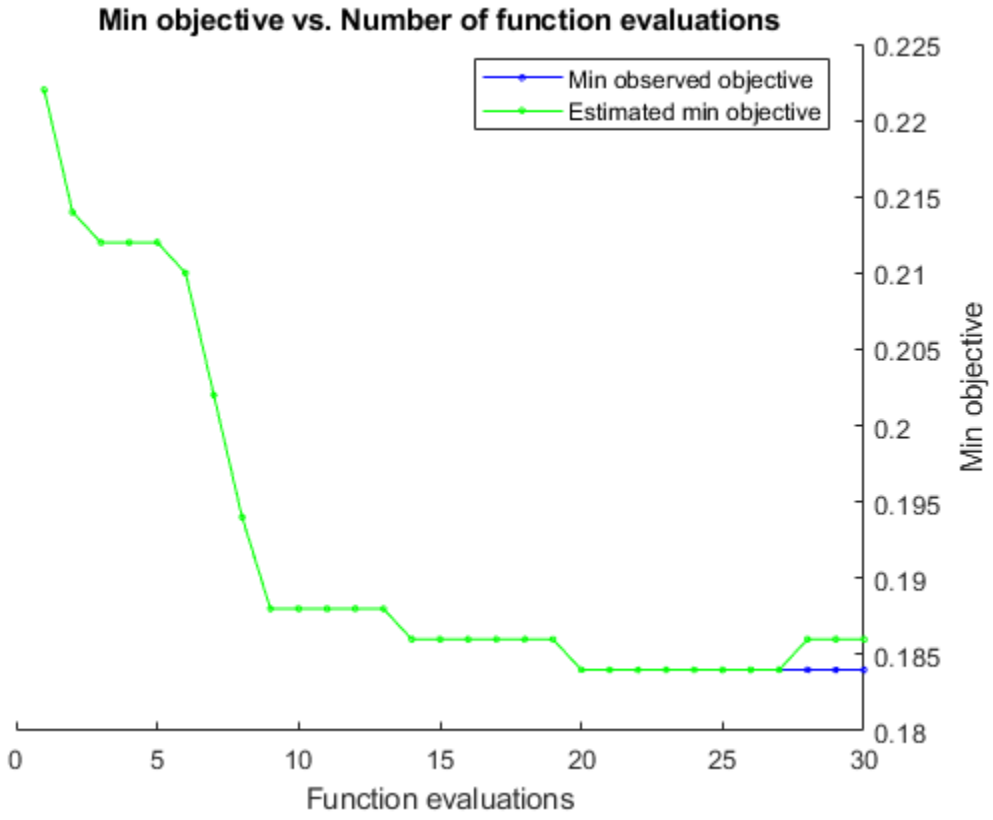
```
minfun = @(z)kfoldLoss(fitcgam(adultdata,'salary','CrossVal','on', ...
    'MaxNumSplitsPerPredictor',z.maxNumSplits, ...
    'NumTreesPerPredictor',z.numTrees));
```

If you specify the cross-validation option (`'CrossVal','on'`), then the `fitcgam` function returns a cross-validated model object `ClassificationPartitionedGAM`. The `kfoldLoss` function returns the classification loss obtained by the cross-validated model. Therefore, the function handle `minfun` computes the cross-validation loss at the parameters in `z`.

Search for the best parameters `[maxNumSplits,numTrees]` using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time and, therefore, can give varying results.

```
results = bayesopt(minfun,[maxNumSplits,numTrees],'Verbose',0, ...
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');
```


Objective function model
Estimated objective function value
0.22
0.2
0.18
0.16
0.14
0.12
0.1
500
400
300
200
100
numTrees
2
4
6
8
10
maxNumSplits
Observed points
Model mean
Next point
Model minimum feasible



Obtain the best point from `results`.

```
zbest = bestPoint(results)
```

```
zbest=1×2 table
    maxNumSplits    numTrees
    ____________    ________

          1             123
```

Train an optimized GAM using the `zbest` values.

```
Mdl = fitcgam(adultdata,'salary', ...
    'MaxNumSplitsPerPredictor',zbest.maxNumSplits, ...
    'NumTreesPerPredictor',zbest.numTrees);
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights.
  - The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. A good practice is to specify the order of the classes in the response variable by using the `'ClassNames'` name-value argument.
  - The column for the weights must be a numeric vector.

  You must specify the response variable in `Tbl` by using `ResponseVarName` or `formula` and specify the observation weights in `Tbl` by using `'Weights'`.
  - Specify the response variable by using `ResponseVarName` — `fitcgam` uses the remaining variables as predictors. To use a subset of the remaining variables in `Tbl` as predictors, specify predictor variables by using `'PredictorNames'`.
  - Define a model specification by using `formula` — `fitcgam` uses a subset of the variables in `Tbl` as predictor variables and the response variable, as specified in `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable `Y` and the number of rows in `Tbl` must be equal. To use a subset of the variables in `Tbl` as predictors, specify predictor variables by using `'PredictorNames'`.

`fitcgam` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Tbl` to be missing values.

- `fitcgam` does not use observations with all missing values in the fit.
- `fitcgam` does not use observations with missing response values in the fit.
- `fitcgam` uses observations with some missing values for predictors to find splits on variables for which these observations have valid values.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`formula` — Model specification
character vector | string scalar

Model specification, specified as a character vector or string scalar in the form `'Y ~ terms'`. The `formula` argument specifies a response variable and linear and interaction terms for predictor variables. Use `formula` to specify a subset of variables in `Tbl` as predictors for training the model. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

For example, specify `'Y~x1+x2+x3+x1:x2'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the linear terms for the predictor variables. `x1:x2` represents the interaction term for `x1` and `x2`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable

names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Alternatively, you can specify a response variable and linear terms for predictors using `formula`, and specify interaction terms for predictors using `'Interactions'`.

`fitcgam` builds a set of interaction trees using only the terms whose *p*-values are not greater than the `'MaxPValue'` value.

Example: `'Y~x1+x2+x3+x1:x2'`

Data Types: `char` | `string`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X` or `Tbl`.

A good practice is to specify the order of the classes by using the `'ClassNames'` name-value argument.

`fitcgam` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. `fitcgam` does not use observations with missing response values in the fit.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

`fitcgam` considers `NaN` values in `X` as missing values. The function does not use observations with all missing values in the fit. `fitcgam` uses observations with some missing values for `X` to find splits on variables for which these observations have valid values.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Interactions','all','MaxPValue',0.05` specifies to include all available interaction terms whose *p*-values are not greater than 0.05.

GAM Options

InitialLearnRateForInteractions — Initial learning rate of gradient boosting for interaction terms

1 (default) | numeric scalar in (0,1]

Initial learning rate of gradient boosting for interaction terms, specified as a numeric scalar in the interval (0,1].

For each boosting iteration for interaction trees, `fitcgam` starts fitting with the initial learning rate. The function halves the learning rate until it finds a rate that improves the model fit.

Training a model using a small learning rate requires more learning iterations, but often achieves better accuracy.

For more details about gradient boosting, see “Gradient Boosting Algorithm” on page 33-1708.

Example: `'InitialLearnRateForInteractions',0.1`

Data Types: `single` | `double`

`InitialLearnRateForPredictors` — Initial learning rate of gradient boosting for linear terms

`1` (default) | numeric scalar in (0,1]

Initial learning rate of gradient boosting for linear terms, specified as a numeric scalar in the interval (0,1].

For each boosting iteration for predictor trees, `fitcgam` starts fitting with the initial learning rate. The function halves the learning rate until it finds a rate that improves the model fit.

Training a model using a small learning rate requires more learning iterations, but often achieves better accuracy.

For more details about gradient boosting, see “Gradient Boosting Algorithm” on page 33-1708.

Example: `'InitialLearnRateForPredictors',0.1`

Data Types: `single` | `double`

`Interactions` — Number or list of interaction terms

`0` (default) | nonnegative integer scalar | logical matrix | `'all'`

Number or list of interaction terms to include in the candidate set *S*, specified as a nonnegative integer scalar, a logical matrix, or `'all'`.

- Number of interaction terms, specified as a nonnegative integer — *S* includes the specified number of important interaction terms, selected based on the *p*-values of the terms.
- List of interaction terms, specified as a logical matrix — *S* includes the terms specified by a `t`-by-`p` logical matrix, where `t` is the number of interaction terms, and `p` is the number of predictors used to train the model. For example, `logical([1 1 0; 0 1 1])` represents two pairs of interaction terms: a pair of the first and second predictors, and a pair of the second and third predictors.

  If `fitcgam` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. That is, the column indexes of the logical matrix do not count the response and observation weight variables. The indexes also do not count any variables not used by the function.
- `'all'` — *S* includes all possible pairs of interaction terms, which is `p*(p – 1)/2` number of terms in total.

Among the interaction terms in *S*, the `fitcgam` function identifies those whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees. Use the default value (`'MaxPValue',1`) to build interaction trees using all terms in *S*.

Example: `'Interactions','all'`

Data Types: `single` | `double` | `logical` | `char` | `string`

`MaxNumSplitsPerInteraction` — Maximum number of decision splits per interaction tree
4 (default) | positive integer scalar

Maximum number of decision splits (or branch nodes) for each interaction tree (boosted tree for an interaction term), specified as a positive integer scalar.

Example: `'MaxNumSplitsPerInteraction',5`

Data Types: `single` | `double`

`MaxNumSplitsPerPredictor` — Maximum number of decision splits per predictor tree
1 (default) | positive integer scalar

Maximum number of decision splits (or branch nodes) for each predictor tree (boosted tree for a linear term), specified as a positive integer scalar. By default, `fitcgam` uses a tree stump for a predictor tree.

Example: `'MaxNumSplitsPerPredictor',5`

Data Types: `single` | `double`

`MaxPValue` — Maximum *p*-value for detecting interaction terms
1 (default) | numeric scalar in [0,1]

Maximum *p*-value for detecting interaction terms, specified as a numeric scalar in the interval [0,1].

`fitcgam` first finds the candidate set *S* of interaction terms from `formula` or `'Interactions'`. Then the function identifies the interaction terms whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees.

The default value (`'MaxPValue',1`) builds interaction trees for all interaction terms in the candidate set *S*.

For more details about detecting interaction terms, see “Interaction Term Detection” on page 33-1709.

Example: `'MaxPValue',0.05`

Data Types: `single` | `double`

`NumBins` — Number of bins for numeric predictors
`256` (default) | positive integer scalar | `[]` (empty)

Number of bins for numeric predictors, specified as a positive integer scalar or `[]` (empty).

- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitcgam` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.

- `fitcgam` does not bin categorical predictors.

- If the `'NumBins'` value is empty (`[]`), then `fitcgam` does not bin any predictors.

When you use a large training data set, this binning option speeds up training but might cause a decrease in accuracy. You can first use the default value of `'NumBins'`, and then change the value depending on the accuracy and training speed.

The trained model `Mdl` stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`NumTreesPerInteraction` — Number of trees per interaction term
100 (default) | positive integer scalar

Number of trees per interaction term, specified as a positive integer scalar.

The `'NumTreesPerInteraction'` value is equivalent to the number of gradient boosting iterations for the interaction terms for predictors. For each iteration, `fitcgam` adds a set of interaction trees to the model, one tree for each interaction term. To learn about the gradient boosting algorithm, see “Gradient Boosting Algorithm” on page 33-1708.

You can determine whether the fitted model has the specified number of trees by viewing the diagnostic message displayed when `'Verbose'` is 1 or 2, or by checking the `ReasonForTermination` property value of the model `Mdl`.

Example: `'NumTreesPerInteraction',500`

Data Types: `single` | `double`

`NumTreesPerPredictor` — Number of trees per linear term
300 (default) | positive integer scalar

Number of trees per linear term, specified as a positive integer scalar.

The `'NumTreesPerPredictor'` value is equivalent to the number of gradient boosting iterations for the linear terms for predictors. For each iteration, `fitcgam` adds a set of predictor trees to the model, one tree for each predictor. To learn about the gradient boosting algorithm, see “Gradient Boosting Algorithm” on page 33-1708.

You can determine whether the fitted model has the specified number of trees by viewing the diagnostic message displayed when `'Verbose'` is 1 or 2, or by checking the `ReasonForTermination` property value of the model `Mdl`.

Example: `'NumTreesPerPredictor',500`

Data Types: `single` | `double`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcgam` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitcgam` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitcgam` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
`[0 1; 1 0]` (default) | 2-by-2 numeric matrix | structure

Misclassification cost of a point, specified as one of the following:

- 2-by-2 numeric matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (that is, the rows correspond to the true class and the columns correspond to the predicted class). To specify the class order for the corresponding rows and columns of `Cost`, set the `'ClassNames'` name-value argument.
- Structure `S` with two fields: `S.ClassNames`, which contains the group names as a variable of the same data type as the response variable in `Tbl` or `Y`; and `S.ClassificationCosts`, which contains the cost matrix.

Example: `'Cost',[0 2; 1 0]`

Data Types: `single` | `double` | `struct`

`NumPrint` — Number of iterations between diagnostic message printouts
`10` (default) | nonnegative integer scalar

Number of iterations between diagnostic message printouts, specified as a nonnegative integer scalar. This argument is valid only when you specify `'Verbose'` as 1.

If you specify `'Verbose',1` and `'NumPrint',numPrint`, then the software displays diagnostic messages every `numPrint` iterations in the Command Window.

Example: `'NumPrint',500`

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcgam` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.

- By default, `PredictorNames` contains the names of all predictor variables.
- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as one of the following:

- Character vector or string scalar.
  - `'empirical'` determines class probabilities from class frequencies in the response variable in `Y` or `Tbl`. If you pass observation weights, `fitcgam` uses the weights to compute the class probabilities.
  - `'uniform'` sets all class probabilities to be equal.
- Vector (one scalar value for each class). To specify the class order for the corresponding elements of `'Prior'`, set the `'ClassNames'` name-value argument.
- Structure `S` with two fields.
  - `S.ClassNames` contains the class names as a variable of the same type as the response variable in `Y` or `Tbl`.
  - `S.ClassProbs` contains a vector of corresponding probabilities.

`fitcgam` normalizes the weights in each class (`'Weights'`) to add up to the value of the prior probability of the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation
`'logit'` (default) | `'none'` | function handle | ...

Score transformation, specified as a built-in transformation function name or function handle.

This table summarizes the available score transformations. Specify one using its corresponding character vector or string scalar.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

This argument determines the output score computation for object functions such as `predict`, `margin`, and `edge`. Use `'logit'` (default) to compute posterior probabilities, and use `'none'` to compute the logit of posterior probabilities.

Example: `'ScoreTransform','none'`

Data Types: `char` | `string` | `function_handle`

`Verbose` — Verbosity level
`0` (default) | `1` | `2`

Verbosity level, specified as `0`, `1`, or `2`. The `Verbose` value controls the amount of information that the software displays in the Command Window.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software displays no information.
`1`	The software displays diagnostic messages every `numPrint` iterations, where `numPrint` is the `'NumPrint'` value.
`2`	The software displays diagnostic messages at every iteration.

Each line of the diagnostic messages shows the information about each boosting iteration and includes the following columns:

- `Type` — Type of trained trees, `1D` (predictor trees, or boosted trees for linear terms for predictors) or `2D` (interaction trees, or boosted trees for interaction terms for predictors)
- `NumTrees` — Number of trees per linear term or interaction term that `fitcgam` added to the model so far

- `Deviance` — "Deviance" on page 33-1708 of the model
- `RelTol` — Relative change of model predictions: $(\widehat{y}_k - \widehat{y}_{k-1})'(\widehat{y}_k - \widehat{y}_{k-1})/\widehat{y}_k'\widehat{y}_k$, where $\widehat{y}_k$ is a column vector of model predictions at iteration k
- `LearnRate` — Learning rate used for the current iteration

Example: `'Verbose',1`

Data Types: `single` | `double`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored in `Tbl.W`, then specify it as `'W'`.

`fitcgam` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Flag to train cross-validated model

`'off'` (default) | `'on'`

Flag to train a cross-validated model, specified as `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated model with 10 folds.

You can override this cross-validation setting using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument. You can use only one cross-validation name-value argument at a time to create a cross-validated model.

Alternatively, cross-validate after creating a model by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Output Arguments

`Mdl` — Trained generalized additive model

`ClassificationGAM` model object | `ClassificationPartitionedGAM` cross-validated model object

Trained generalized additive model, returned as one of the model objects in this table.

Model Object	Cross-Validation Options to Train Model Object	Ways to Classify Observations Using Model Object
`ClassificationGAM`	None	Use `predict` to classify new observations, and use `resubPredict` to classify training observations.
`ClassificationPartitionedGAM`	Specify `KFold`, `Holdout`, `Leaveout`, `CrossVal`, or `CVPartition`	Use `kfoldPredict` to classify observations that `fitcgam` holds out during training. `kfoldPredict` predicts a class label for every observation by using the model trained without that observation.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.Interactions` in the Command Window to display the interaction terms in `Mdl`.

More About

Generalized Additive Model (GAM) for Binary Classification

A generalized additive model (GAM) is an interpretable model that explains class scores (the logit of class probabilities) using a sum of univariate and bivariate shape functions of predictors.

`fitcgam` uses a boosted tree as a shape function for each predictor and, optionally, each pair of predictors; therefore, the function can capture a nonlinear relation between a predictor and the response variable. Because contributions of individual shape functions to the prediction (classification score) are well separated, the model is easy to interpret.

The standard GAM uses a univariate shape function for each predictor.

$$y \sim Binomial(n, \mu)$$

$$g(\mu) = \log\frac{\mu}{1-\mu} = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where y is a response variable that follows the binomial distribution with the probability of success (probability of positive class) μ in n observations. $g(\mu)$ is a logit link function, and c is an intercept (constant) term. $f_i(x_i)$ is a univariate shape function for the ith predictor, which is a boosted tree for a linear term for the predictor (predictor tree).

You can include interactions between predictors in a model by adding bivariate shape functions of important interaction terms to the model.

$$g(\mu) = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \sum_{i,\, j \in \{1, 2, \cdots, p\}} f_{ij}(x_i x_j),$$

where $f_{ij}(x_i x_j)$ is a bivariate shape function for the ith and jth predictors, which is a boosted tree for an interaction term for the predictors (interaction tree).

`fitcgam` finds important interaction terms based on the p-values of F-tests. For details, see "Interaction Term Detection" on page 33-1709.

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to the saturated model.

The deviance of a fitted model is twice the difference between the loglikelihoods of the model and the saturated model:

$$-2(\log L - \log L_s),$$

where L and L_s are the likelihoods of the fitted model and the saturated model, respectively. The saturated model is the model with the maximum number of parameters that you can estimate.

`fitcgam` uses the deviance to measure the goodness of model fit and finds a learning rate that reduces the deviance at each iteration. Specify `'Verbose'` as 1 or 2 to display the deviance and learning rate in the Command Window.

Algorithms

Gradient Boosting Algorithm

`fitcgam` fits a generalized additive model using a gradient boosting algorithm ("Adaptive Logistic Regression" on page 18-48).

`fitcgam` first builds sets of predictor trees (boosted trees for linear terms for predictors) and then builds sets of interaction trees (boosted trees for interaction terms for predictors). The boosting algorithm iterates for at most `'NumTreesPerPredictor'` times for predictor trees, and then iterates for at most `'NumTreesPerInteraction'` times for interaction trees.

For each boosting iteration, `fitcgam` builds a set of predictor trees with the initial learning rate `'InitialLearnRateForPredictors'`, or builds a set of interaction trees with the initial learning rate `'InitialLearnRateForInteractions'`.

- When building a set of trees, the function trains one tree at a time. It fits a tree to the residual that is the difference between the response and the aggregated prediction from all trees grown previously. To control the boosting learning speed, the function shrinks the tree by the learning rate and then adds the tree to the model and updates the residual.
  - Updated model = current model + (learning rate)·(new tree)
  - Updated residual = current residual – (learning rate)·(response explained by new tree)
- If adding the set of trees improves the model fit (that is, reduces the deviance of the fit), then `fitcgam` moves to the next iteration.
- Otherwise, `fitcgam` halves the learning rate and uses it to update the model and residual. The function continues to halve the learning rate until it finds a rate that improves the model fit.
  - If the function cannot find such a learning rate for predictor trees, then it stops boosting iterations for linear terms and starts boosting iterations for interaction terms.

- If the function cannot find such a learning rate for interaction trees, then it terminates the model fitting.

You can determine why training stopped by checking the `ReasonForTermination` property of the trained model.

Interaction Term Detection

For each pairwise interaction term $x_i x_j$ (specified by `formula` or `'Interactions'`), the software performs an *F*-test to examine whether the term is statistically significant.

To speed up the process, `fitcgam` bins numeric predictors into at most 8 equiprobable bins. The number of bins can be less than 8 if a predictor has fewer than 8 unique values. The *F*-test examines the null hypothesis that the bins created by x_i and x_j have equal responses versus the alternative that at least one bin has a different response value from the others. A small *p*-value indicates that differences are significant, which implies that the corresponding interaction term is significant and, therefore, including the term can improve the model fit.

`fitcgam` builds a set of interaction trees using the terms whose *p*-values are not greater than the `'MaxPValue'` value. You can use the default `'MaxPValue'` value `1` to build interaction trees using all terms specified by `formula` or `'Interactions'`.

`fitcgam` adds interaction terms to the model in the order of importance based on the *p*-values. Use the `Interactions` property of the returned model to check the order of the interaction terms added to the model.

References

[1] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible Models for Classification and Regression." *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12).* Beijing, China: ACM Press, 2012, pp. 150–158.

[2] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate Intelligible Models with Pairwise Interactions." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)* Chicago, Illinois, USA: ACM Press, 2013, pp. 623–631.

See Also

`ClassificationGAM` | `ClassificationPartitionedGAM` | `addInteractions` | `predict` | `resume`

Topics

"Train Generalized Additive Model for Binary Classification" on page 12-77

Introduced in R2021a

fitcknn

Fit *k*-nearest neighbor classifier

Syntax

```
Mdl = fitcknn(Tbl,ResponseVarName)
Mdl = fitcknn(Tbl,formula)
Mdl = fitcknn(Tbl,Y)

Mdl = fitcknn(X,Y)

Mdl = fitcknn(___,Name,Value)
```

Description

`Mdl = fitcknn(Tbl,ResponseVarName)` returns a *k*-nearest neighbor classification model based on the input variables (also known as predictors, features, or attributes) in the table `Tbl` and output (response) `Tbl.ResponseVarName`.

`Mdl = fitcknn(Tbl,formula)` returns a *k*-nearest neighbor classification model based on the input variables in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl`.

`Mdl = fitcknn(Tbl,Y)` returns a *k*-nearest neighbor classification model based on the predictor variables in the table `Tbl` and response array `Y`.

`Mdl = fitcknn(X,Y)` returns a *k*-nearest neighbor classification model based on the predictor data `X` and response `Y`.

`Mdl = fitcknn(___,Name,Value)` fits a model with additional options specified by one or more name-value pair arguments, using any of the previous syntaxes. For example, you can specify the tie-breaking algorithm, distance metric, or observation weights.

Examples

Train *k*-Nearest Neighbor Classifier

Train a *k*-nearest neighbor classifier for Fisher's iris data, where *k*, the number of nearest neighbors in the predictors, is 5.

Load Fisher's iris data.

```
load fisheriris
X = meas;
Y = species;
```

`X` is a numeric matrix that contains four petal measurements for 150 irises. `Y` is a cell array of character vectors that contains the corresponding iris species.

Train a 5-nearest neighbor classifier. Standardize the noncategorical predictor data.

```
Mdl = fitcknn(X,Y,'NumNeighbors',5,'Standardize',1)
```

```
Mdl =
  ClassificationKNN
              ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
                 Distance: 'euclidean'
             NumNeighbors: 5


  Properties, Methods
```

`Mdl` is a trained `ClassificationKNN` classifier, and some of its properties appear in the Command Window.

To access the properties of `Mdl`, use dot notation.

```
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl.Prior
```

```
ans = 1×3

    0.3333    0.3333    0.3333
```

`Mdl.Prior` contains the class prior probabilities, which you can specify using the `'Prior'` name-value pair argument in `fitcknn`. The order of the class prior probabilities corresponds to the order of the classes in `Mdl.ClassNames`. By default, the prior probabilities are the respective relative frequencies of the classes in the data.

You can also reset the prior probabilities after training. For example, set the prior probabilities to 0.5, 0.2, and 0.3, respectively.

```
Mdl.Prior = [0.5 0.2 0.3];
```

You can pass `Mdl` to `predict` to label new measurements or `crossval` to cross-validate the classifier.

Train a *k*-Nearest Neighbor Classifier Using the Minkowski Metric

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

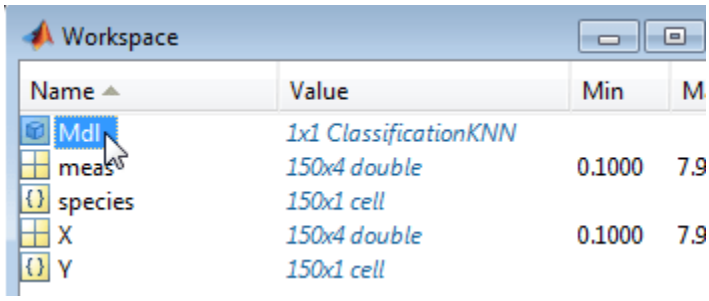
X is a numeric matrix that contains four petal measurements for 150 irises. Y is a cell array of character vectors that contains the corresponding iris species.

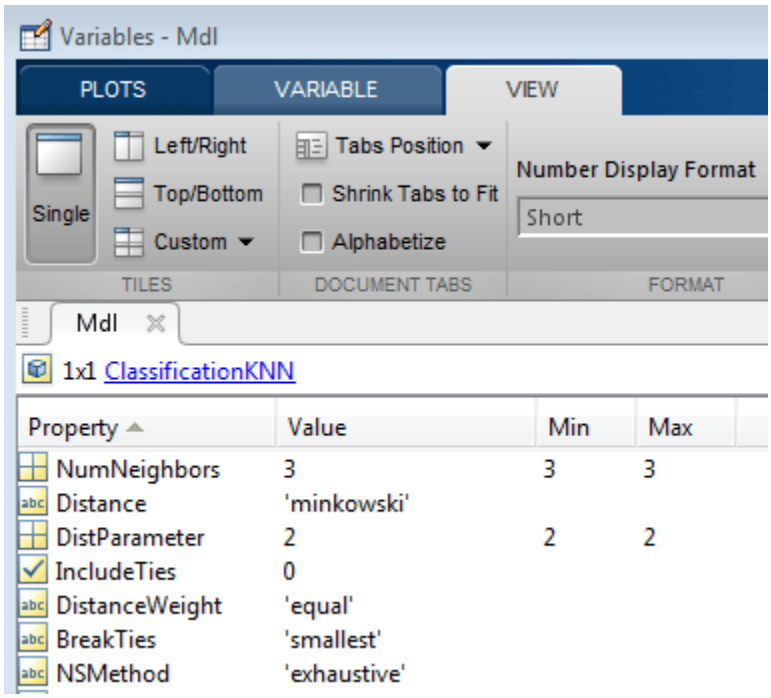
Train a 3-nearest neighbors classifier using the Minkowski metric. To use the Minkowski metric, you must use an exhaustive searcher. It is good practice to standardize noncategorical predictor data.

```
Mdl = fitcknn(X,Y,'NumNeighbors',3,...
    'NSMethod','exhaustive','Distance','minkowski',...
    'Standardize',1);
```

`Mdl` is a `ClassificationKNN` classifier.

You can examine the properties of `Mdl` by double-clicking `Mdl` in the Workspace window. This opens the Variable Editor.





Train a *k*-Nearest Neighbor Classifier Using a Custom Distance Metric

Train a *k*-nearest neighbor classifier using the chi-square distance.

Load Fisher's iris data set.

```
load fisheriris
X = meas;    % Predictors
Y = species; % Response
```

The chi-square distance between j-dimensional points x and z is

$$\chi(x,z) = \sqrt{\sum_{j=1}^{J} w_j (x_j - z_j)^2},$$

where w_j is a weight associated with dimension j.

Specify the chi-square distance function. The distance function must:

- Take one row of `X`, e.g., `x`, and the matrix `Z`.
- Compare `x` to each row of `Z`.
- Return a vector `D` of length n_z, where n_z is the number of rows of `Z`. Each element of `D` is the distance between the observation corresponding to `x` and the observations corresponding to each row of `Z`.

```
chiSqrDist = @(x,Z,wt)sqrt((bsxfun(@minus,x,Z).^2)*wt);
```

This example uses arbitrary weights for illustration.

Train a 3-nearest neighbor classifier. It is good practice to standardize noncategorical predictor data.

```
k = 3;
w = [0.3; 0.3; 0.2; 0.2];
KNNMdl = fitcknn(X,Y,'Distance',@(x,Z)chiSqrDist(x,Z,w),...
    'NumNeighbors',k,'Standardize',1);
```

`KNNMdl` is a `ClassificationKNN` classifier.

Cross validate the KNN classifier using the default 10-fold cross validation. Examine the classification error.

```
rng(1); % For reproducibility
CVKNNMdl = crossval(KNNMdl);
classError = kfoldLoss(CVKNNMdl)
```

```
classError = 0.0600
```

`CVKNNMdl` is a `ClassificationPartitionedModel` classifier. The 10-fold classification error is 4%.

Compare the classifier with one that uses a different weighting scheme.

```
w2 = [0.2; 0.2; 0.3; 0.3];
CVKNNMdl2 = fitcknn(X,Y,'Distance',@(x,Z)chiSqrDist(x,Z,w2),...
    'NumNeighbors',k,'KFold',10,'Standardize',1);
classError2 = kfoldLoss(CVKNNMdl2)
```

```
classError2 = 0.0400
```

The second weighting scheme yields a classifier that has better out-of-sample performance.

Optimize Fitted KNN Classifier

This example shows how to optimize hyperparameters automatically using `fitcknn`. The example uses the Fisher iris data.

Load the data.

```
load fisheriris
X = meas;
Y = species;
```

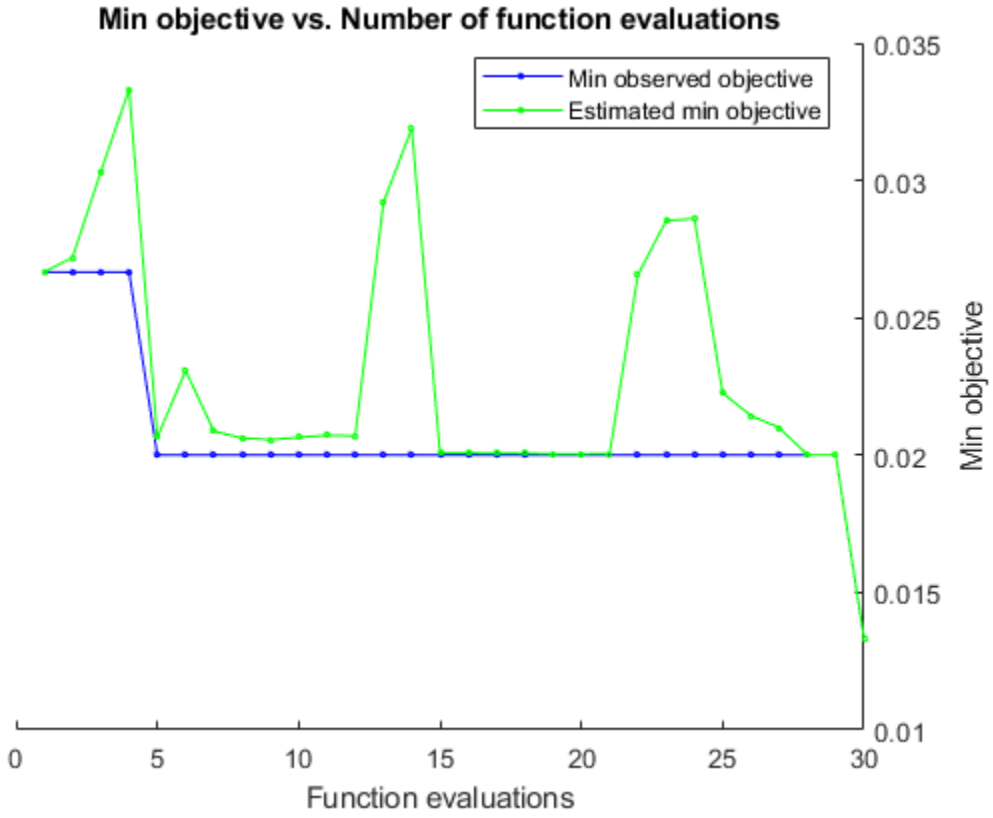
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

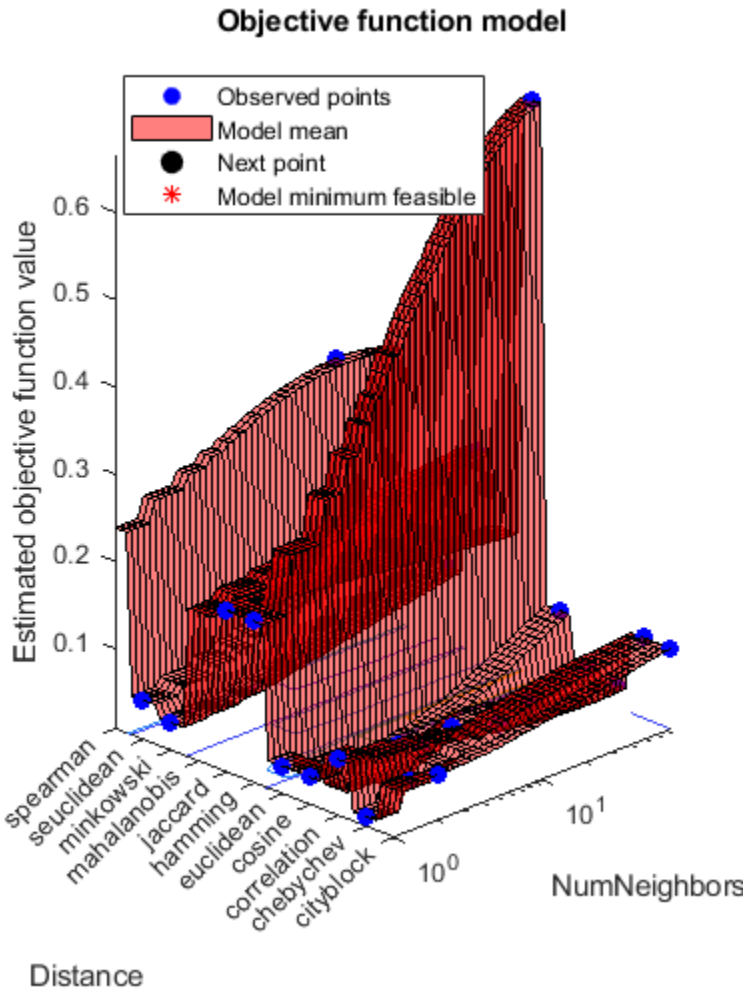
For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng(1)
Mdl = fitcknn(X,Y,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | NumNeighbors |     Dist
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |    0.026667 |     0.40383 |    0.026667 |    0.026667 |           30 |       co
|    2 | Accept |        0.04 |      0.1555 |    0.026667 |    0.027197 |            2 |    cheby
|    3 | Accept |     0.19333 |     0.15825 |    0.026667 |    0.030324 |            1 |      ham
|    4 | Accept |     0.33333 |     0.27148 |    0.026667 |    0.033313 |           31 |     spea
|    5 | Best   |        0.02 |     0.10956 |        0.02 |    0.020648 |            6 |       co
|    6 | Accept |    0.073333 |     0.17189 |        0.02 |    0.023082 |            1 |  correla
|    7 | Accept |        0.06 |     0.13061 |        0.02 |    0.020875 |            2 |    cityb
|    8 | Accept |        0.04 |     0.12303 |        0.02 |    0.020622 |            1 |    eucli
|    9 | Accept |        0.24 |     0.24423 |        0.02 |    0.020562 |           74 |  mahalan
|   10 | Accept |        0.04 |     0.16504 |        0.02 |    0.020649 |            1 |    minko
|   11 | Accept |    0.053333 |     0.24475 |        0.02 |    0.020722 |            1 |   seucli
|   12 | Accept |     0.19333 |     0.14521 |        0.02 |    0.020701 |            1 |      jac
|   13 | Accept |        0.04 |     0.11858 |        0.02 |    0.029203 |            1 |       co
|   14 | Accept |        0.04 |      0.1735 |        0.02 |    0.031888 |           75 |       co
|   15 | Accept |        0.04 |     0.15935 |        0.02 |    0.020076 |            1 |       co
|   16 | Accept |    0.093333 |     0.14057 |        0.02 |    0.020073 |           75 |    eucli
|   17 | Accept |    0.093333 |     0.11912 |        0.02 |     0.02007 |           75 |    minko
|   18 | Accept |         0.1 |     0.15151 |        0.02 |    0.020061 |           75 |    cheby
|   19 | Accept |     0.15333 |     0.12604 |        0.02 |    0.020044 |           75 |   seucli
|   20 | Accept |         0.1 |     0.14656 |        0.02 |    0.020044 |           75 |    cityb
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | NumNeighbors |     Dist
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|   21 | Accept |    0.033333 |     0.14888 |        0.02 |    0.020046 |           75 |  correla
|   22 | Accept |    0.033333 |     0.15147 |        0.02 |     0.02656 |            9 |       co
|   23 | Accept |    0.033333 |     0.13347 |        0.02 |     0.02854 |            9 |       co
|   24 | Accept |        0.02 |     0.11808 |        0.02 |    0.028607 |            1 |    cheby
|   25 | Accept |        0.02 |     0.17591 |        0.02 |    0.022264 |            1 |    cheby
```

```
|   26 | Accept |        0.02 |     0.14906 |        0.02 |    0.021439 |           1 |    cheby
|   27 | Accept |        0.02 |     0.12712 |        0.02 |    0.020999 |           1 |    cheby
|   28 | Accept |     0.66667 |     0.10841 |        0.02 |    0.020008 |          75 |      har
|   29 | Accept |        0.04 |     0.15102 |        0.02 |    0.020008 |          12 |  correla
|   30 | Best   |    0.013333 |     0.15388 |    0.013333 |    0.013351 |           6 |    eucli
```






```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 57.3265 seconds
Total objective function evaluation time: 4.8759

Best observed feasible point:
    NumNeighbors    Distance
    ____________    _________

          6          euclidean

Observed objective function value = 0.013333
Estimated objective function value = 0.013351
Function evaluation time = 0.15388

Best estimated feasible point (according to models):
    NumNeighbors    Distance
    ____________    _________
```

```
         6          euclidean

Estimated objective function value = 0.013351
Estimated function evaluation time = 0.15533

Mdl =
  ClassificationKNN
                           ResponseName: 'Y'
                  CategoricalPredictors: []
                             ClassNames: {'setosa'  'versicolor'  'virginica'}
                         ScoreTransform: 'none'
                        NumObservations: 150
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                               Distance: 'euclidean'
                           NumNeighbors: 6


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X`.

The software considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. Consequently, the software does not train using observations with a missing response.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as numeric matrix.

Each row corresponds to one observation (also known as an instance or example), and each column corresponds to one predictor variable (also known as a feature).

The length of `Y` and the number of rows of `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'NumNeighbors',3,'NSMethod','exhaustive','Distance','minkowski'` specifies a classifier for three-nearest neighbors using the nearest neighbor search method and the Minkowski metric.

Model Parameters

`BreakTies` — Tie-breaking algorithm

`'smallest'` (default) | `'nearest'` | `'random'`

Tie-breaking algorithm used by the `predict` method if multiple classes have the same smallest cost, specified as the comma-separated pair consisting of `'BreakTies'` and one of the following:

- `'smallest'` — Use the smallest index among tied groups.
- `'nearest'` — Use the class with the nearest neighbor among tied groups.
- `'random'` — Use a random tiebreaker among tied groups.

By default, ties occur when multiple classes have the same number of nearest points among the `K` nearest neighbors.

Example: `'BreakTies','nearest'`

`BucketSize` — Maximum data points in node

`50` (default) | positive integer value

Maximum number of data points in the leaf node of the *k*d-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer value. This argument is meaningful only when `NSMethod` is `'kdtree'`.

Example: `'BucketSize',40`

Data Types: `single` | `double`

`CategoricalPredictors` — Categorical predictor flag

`[]` | `'all'`

Categorical predictor flag, specified as the comma-separated pair consisting of `'CategoricalPredictors'` and one of the following:

- `'all'` — All predictors are categorical.
- `[]` — No predictors are categorical.

The predictor data for `fitcknn` must be either all continuous or all categorical.

- If the predictor data is in a table (`Tbl`), `fitcknn` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If `Tbl` includes both continuous and categorical values, then you must specify the value of `'CategoricalPredictors'` so that `fitcknn` can determine how to treat all predictors, as either continuous or categorical variables.
- If the predictor data is a matrix (`X`), `fitcknn` assumes that all predictors are continuous. To identify all predictors in `X` as categorical, specify `'CategoricalPredictors'` as `'all'`.

When you set `CategoricalPredictors` to `'all'`, the default `Distance` is `'hamming'`.

Example: `'CategoricalPredictors','all'`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Cost of misclassification

square matrix | structure

Cost of misclassification of a point, specified as the comma-separated pair consisting of `'Cost'` and one of the following:

- Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- Structure `S` having two fields: `S.ClassNames` containing the group names as a variable of the same type as `Y`, and `S.ClassificationCosts` containing the cost matrix.

The default is `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`.

Data Types: `single` | `double` | `struct`

`Cov` — Covariance matrix

`cov(X,'omitrows')` (default) | positive definite matrix of scalar values

Covariance matrix, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix of scalar values representing the covariance matrix when computing the Mahalanobis distance. This argument is only valid when `'Distance'` is `'mahalanobis'`.

You cannot simultaneously specify `'Standardize'` and either of `'Scale'` or `'Cov'`.

Data Types: `single` | `double`

`Distance` — Distance metric
`'cityblock'` | `'chebychev'` | `'correlation'` | `'cosine'` | `'euclidean'` | `'hamming'` | function handle | ...

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a valid distance metric name or function handle. The allowable distance metric names depend on your choice of a neighbor-searcher method (see `NSMethod`).

NSMethod	Distance Metric Names
`exhaustive`	Any distance metric of `ExhaustiveSearcher`
`kdtree`	`'cityblock'`, `'chebychev'`, `'euclidean'`, or `'minkowski'`

This table includes valid distance metrics of `ExhaustiveSearcher`.

Distance Metric Names	Description
`'cityblock'`	City block distance.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors).
`'euclidean'`	Euclidean distance.
`'hamming'`	Hamming distance, percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix `C`. The default value of `C` is the sample covariance matrix of `X`, as computed by `cov(X,'omitrows')`. To specify a different value for `C`, use the `'Cov'` name-value pair argument.
`'minkowski'`	Minkowski distance. The default exponent is `2`. To specify a different exponent, use the `'Exponent'` name-value pair argument.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between `X` and a query point is scaled, meaning divided by a scale value `S`. The default value of `S` is the standard deviation computed from `X`, `S = std(X,'omitnan')`. To specify another value for `S`, use the `Scale` name-value pair argument.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

Distance Metric Names	Description

`@distfun`	Distance function handle. *`distfun`* has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of  distance`
	`...`
	where
	• `ZI` is a 1-by-`N` vector containing one row of `X` or `Y`.
	• `ZJ` is an `M2`-by-`N` matrix containing multiple rows of `X` or `Y`.
	• `D2` is an `M2`-by-1 vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.

If you specify `CategoricalPredictors` as `'all'`, then the default distance metric is `'hamming'`. Otherwise, the default distance metric is `'euclidean'`.

For definitions, see “Distance Metrics” on page 18-12.

Example: `'Distance','minkowski'`

Data Types: `char` | `string` | `function_handle`

`DistanceWeight` — Distance weighting function
`'equal'` (default) | `'inverse'` | `'squaredinverse'` | function handle

Distance weighting function, specified as the comma-separated pair consisting of `'DistanceWeight'` and either a function handle or one of the values in this table.

Value	Description
`'equal'`	No weighting
`'inverse'`	Weight is 1/distance
`'squaredinverse'`	Weight is $1/\text{distance}^2$
`@fcn`	*`fcn`* is a function that accepts a matrix of nonnegative distances, and returns a matrix the same size containing nonnegative distance weights. For example, `'squaredinverse'` is equivalent to `@(d)d.^(-2)`.

Example: `'DistanceWeight','inverse'`

Data Types: `char` | `string` | `function_handle`

`Exponent` — Minkowski distance exponent
2 (default) | positive scalar value

Minkowski distance exponent, specified as the comma-separated pair consisting of `'Exponent'` and a positive scalar value. This argument is only valid when `'Distance'` is `'minkowski'`.

Example: `'Exponent',3`

Data Types: `single` | `double`

`IncludeTies` — Tie inclusion flag
`false` (default) | `true`

Tie inclusion flag, specified as the comma-separated pair consisting of `'IncludeTies'` and a logical value indicating whether `predict` includes all the neighbors whose distance values are equal to the `K`th smallest distance. If `IncludeTies` is `true`, `predict` includes all these neighbors. Otherwise, `predict` uses exactly `K` neighbors.

Example: `'IncludeTies',true`

Data Types: `logical`

`NSMethod` — Nearest neighbor search method
`'kdtree'` | `'exhaustive'`

Nearest neighbor search method, specified as the comma-separated pair consisting of `'NSMethod'` and `'kdtree'` or `'exhaustive'`.

- `'kdtree'` — Creates and uses a *k*d-tree to find nearest neighbors. `'kdtree'` is valid when the distance metric is one of the following:
  - `'euclidean'`
  - `'cityblock'`
  - `'minkowski'`
  - `'chebychev'`
- `'exhaustive'` — Uses the exhaustive search algorithm. When predicting the class of a new point `xnew`, the software computes the distance values from all points in `X` to `xnew` to find nearest neighbors.

The default is `'kdtree'` when `X` has `10` or fewer columns, `X` is not sparse or a `gpuArray`, and the distance metric is a `'kdtree'` type; otherwise, `'exhaustive'`.

Example: `'NSMethod','exhaustive'`

`NumNeighbors` — Number of nearest neighbors to find
`1` (default) | positive integer value

Number of nearest neighbors in `X` to find for classifying each point when predicting, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer value.

Example: `'NumNeighbors',3`

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.

- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcknn` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

If you set values for both `Weights` and `Prior`, the weights are renormalized to add up to the value of the prior probability in the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Scale` — Distance scale

`std(X,'omitnan')` (default) | vector of nonnegative scalar values

Distance scale, specified as the comma-separated pair consisting of `'Scale'` and a vector containing nonnegative scalar values with length equal to the number of columns in `X`. Each coordinate difference between `X` and a query point is scaled by the corresponding element of `Scale`. This argument is only valid when `'Distance'` is `'seuclidean'`.

You cannot simultaneously specify `'Standardize'` and either of `'Scale'` or `'Cov'`.

Data Types: `single` | `double`

`ScoreTransform` — Score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Standardize` — Flag to standardize predictors

`false` (default) | `true`

Flag to standardize the predictors, specified as the comma-separated pair consisting of `'Standardize'` and `true` (`1`) or `false` (`0`).

If you set `'Standardize',true`, then the software centers and scales each column of the predictor data (`X`) by the column mean and standard deviation, respectively.

The software does not standardize categorical predictors, and throws an error if all predictors are categorical.

You cannot simultaneously specify `'Standardize',1` and either of `'Scale'` or `'Cov'`.

It is good practice to standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical`

`Weights` — Observation weights
numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Cross Validation Options

`CrossVal` — Cross-validation flag
`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross validate `Mdl` later using the `crossval` method.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.

2 Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Distance','NumNeighbors'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitcknn` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcknn` are:

- `Distance` — `fitcknn` searches among `'cityblock'`, `'chebychev'`, `'correlation'`, `'cosine'`, `'euclidean'`, `'hamming'`, `'jaccard'`, `'mahalanobis'`, `'minkowski'`, `'seuclidean'`, and `'spearman'`.
- `DistanceWeight` — `fitcknn` searches among `'equal'`, `'inverse'`, and `'squaredinverse'`.
- `Exponent` — `fitcknn` searches among positive real values, by default in the range `[0.5,3]`.
- `NumNeighbors` — `fitcknn` searches among positive integer values, by default log-scaled in the range `[1, max(2,round(NumObservations/2))]`.
- `Standardize` — `fitcknn` searches among the values `'true'` and `'false'`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitcknn',meas,species);
params(1).Range = [1,20];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Fitted KNN Classifier” on page 33-1714.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`

Field Name	Values	Default
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`

Field Name	Values	Default
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained *k*-nearest neighbor classification model

`ClassificationKNN` model object | `ClassificationPartitionedModel` cross-validated model object

Trained *k*-nearest neighbor classification model, returned as a `ClassificationKNN` model object or a `ClassificationPartitionedModel` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `ClassificationPartitionedModel` cross-validated model object. Otherwise, `Mdl` is a `ClassificationKNN` model object.

To reference properties of `Mdl`, use dot notation. For example, to display the distance metric at the Command Window, enter `Mdl.Distance`.

More About

Prediction

`ClassificationKNN` predicts the classification of a point `xnew` using a procedure equivalent to this:

1. Find the `NumNeighbors` points in the training set `X` that are nearest to `xnew`.
2. Find the `NumNeighbors` response values `Y` to those nearest points.
3. Assign the classification label `ynew` that has the largest posterior probability among the values in `Y`.

For details, see “Posterior Probability” on page 33-4786 in the `predict` documentation.

Tips

After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- `NaN`s or `<undefined>`s indicate missing observations. The following describes the behavior of `fitcknn` when the data set or weights contain missing observations.
  - If any value of `Y` or any weight is missing, then `fitcknn` removes those values from `Y`, the weights, and the corresponding rows of `X` from the data. The software renormalizes the weights to sum to `1`.
  - If you specify to standardize predictors (`'Standardize',1`) or the standardized Euclidean distance (`'Distance','seuclidean'`) without a scale, then `fitcknn` removes missing observations from individual predictors before computing the mean and standard deviation. In other words, the software implements `mean` and `std` with the `'omitnan'` option on each predictor.
  - If you specify the Mahalanobis distance (`'Distance','mahalanbois'`) without its covariance matrix, then `fitcknn` removes rows of `X` that contain at least one missing value. In other words, the software implements `cov` with the `'omitrows'` option on the predictor matrix `X`.
- Suppose that you set `'Standardize',1`.
  - If you also specify `Prior` or `Weights`, then the software takes the observation weights into account. Specifically, the weighted mean of predictor j is

    $$\bar{x}_j = \sum_{B_j} w_k x_{jk}$$

    and the weighted standard deviation is

    $$s_j = \sum_{B_j} w_k (x_{jk} - \bar{x}_j),$$

    where B_j is the set of indices k for which x_{jk} and w_k are not missing.
  - If you also set `'Distance','mahalanobis'` or `'Distance','seuclidean'`, then you cannot specify `Scale` or `Cov`. Instead, the software:
    1. Computes the means and standard deviations of each predictor
    2. Standardizes the data using the results of step 1
    3. Computes the distance parameter values using their respective default.
- If you specify `Scale` and either of `Prior` or `Weights`, then the software scales observed distances by the weighted standard deviations.
- If you specify `Cov` and either of `Prior` or `Weights`, then the software applies the weighted covariance matrix to the distances. In other words,

$$Cov = \frac{\sum_B w_j}{\left(\sum_B w_j\right)^2 - \sum_B w_j^2} \sum_B w_j (x_j - \bar{x})'(x_j - \bar{x}),$$

where B is the set of indices j for which the observation x_j does not have any missing values and w_j is not missing.

Alternatives

Although `fitcknn` can train a multiclass KNN classifier, you can reduce a multiclass learning problem to a series of KNN binary learners using `fitcecoc`.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- By default, `fitcknn` uses the exhaustive nearest neighbor search algorithm for `gpuArray` input arguments.
- You cannot specify the name-value argument `'NSMethod'` as `'kdtree'`.
- You cannot specify the name-value argument `'Distance'` as a function handle.
- `fitcknn` returns a classification model fitted with GPU array input arguments in these cases only:
  - The input argument `X` is a `gpuArray`.
  - The input argument `Tbl` contains `gpuArray` elements.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationKNN` | `ClassificationPartitionedModel` | `fitcecoc` | `fitcensemble` | `predict` | `templateKNN`

Topics

“Construct KNN Classifier” on page 18-28
“Modify KNN Classifier” on page 18-29

“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2014a

fitclinear

Fit linear classification model to high-dimensional data

Syntax

```
Mdl = fitclinear(X,Y)

Mdl = fitclinear(Tbl,ResponseVarName)
Mdl = fitclinear(Tbl,formula)
Mdl = fitclinear(Tbl,Y)

Mdl = fitclinear(X,Y,Name,Value)
[Mdl,FitInfo] = fitclinear( ___ )
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitclinear( ___ )
```

Description

`fitclinear` trains linear classification models for two-class (binary) learning with high-dimensional, full or sparse predictor data. Available linear classification models include regularized support vector machines (SVM) and logistic regression models. `fitclinear` minimizes the objective function using techniques that reduce computing time (e.g., stochastic gradient descent).

For reduced computation time on a high-dimensional data set that includes many predictor variables, train a linear classification model by using `fitclinear`. For low- through medium-dimensional predictor data sets, see “Alternatives for Lower-Dimensional Data” on page 33-1767.

To train a linear classification model for multiclass learning by combining SVM or logistic regression binary classifiers using error-correcting output codes, see `fitcecoc`.

`Mdl = fitclinear(X,Y)` returns a trained linear classification model object `Mdl` that contains the results of fitting a binary support vector machine to the predictors `X` and class labels `Y`.

`Mdl = fitclinear(Tbl,ResponseVarName)` returns a linear classification model using the predictor variables in the table `Tbl` and the class labels in `Tbl.ResponseVarName`.

`Mdl = fitclinear(Tbl,formula)` returns a linear classification model using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitclinear(Tbl,Y)` returns a linear classification model using the predictor variables in the table `Tbl` and the class labels in vector `Y`.

`Mdl = fitclinear(X,Y,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that the columns of the predictor matrix correspond to observations, implement logistic regression, or specify to cross-validate. A good practice is to cross-validate using the `'Kfold'` name-value pair argument. The cross-validation results determine how well the model generalizes.

`[Mdl,FitInfo] = fitclinear( ___ )` also returns optimization details using any of the previous syntaxes. You cannot request `FitInfo` for cross-validated models.

`[Mdl,FitInfo,HyperparameterOptimizationResults] = fitclinear( ___ )` also returns hyperparameter optimization details when you pass an `OptimizeHyperparameters` name-value pair.

Examples

Train Linear Classification Model

Train a binary, linear classification model using support vector machines, dual SGD, and ridge regularization.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

Identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. Train the model using the entire data set. Determine how well the optimization algorithm fit the model to the data by extracting a fit summary.

```
rng(1); % For reproducibility
[Mdl,FitInfo] = fitclinear(X,Ystats)
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [34023x1 double]
              Bias: -1.0059
            Lambda: 3.1674e-05
           Learner: 'svm'


  Properties, Methods


FitInfo = struct with fields:
                     Lambda: 3.1674e-05
                  Objective: 5.3783e-04
                  PassLimit: 10
                  NumPasses: 10
                 BatchLimit: []
             NumIterations: 238561
              GradientNorm: NaN
         GradientTolerance: 0
      RelativeChangeInBeta: 0.0562
             BetaTolerance: 1.0000e-04
```

```
          DeltaGradient: 1.4582
    DeltaGradientTolerance: 1
           TerminationCode: 0
         TerminationStatus: {'Iteration limit exceeded.'}
                     Alpha: [31572x1 double]
                   History: []
                   FitTime: 0.2014
                    Solver: {'dual'}
```

`Mdl` is a `ClassificationLinear` model. You can pass `Mdl` and the training or new data to `loss` to inspect the in-sample classification error. Or, you can pass `Mdl` and new predictor data to `predict` to predict class labels for new observations.

`FitInfo` is a structure array containing, among other things, the termination status (`TerminationStatus`) and how long the solver took to fit the model to the data (`FitTime`). It is good practice to use `FitInfo` to determine whether optimization-termination measurements are satisfactory. Because training time is small, you can try to retrain the model, but increase the number of passes through the data. This can improve measures like `DeltaGradient`.

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, implement 5-fold cross-validation.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Estimate the coefficients using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
X = X';
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns','KFold',5,...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
```

```
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

```
numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `ClassificationPartitionedLinear` model. Because `fitclinear` implements 5-fold cross-validation, `CVMdl` contains 5 `ClassificationLinear` models that the software trains on each fold.

Display the first trained linear classification model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl1` is a `ClassificationLinear` model object. `fitclinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 11 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated classification error.

```
ce = kfoldLoss(CVMdl);
```

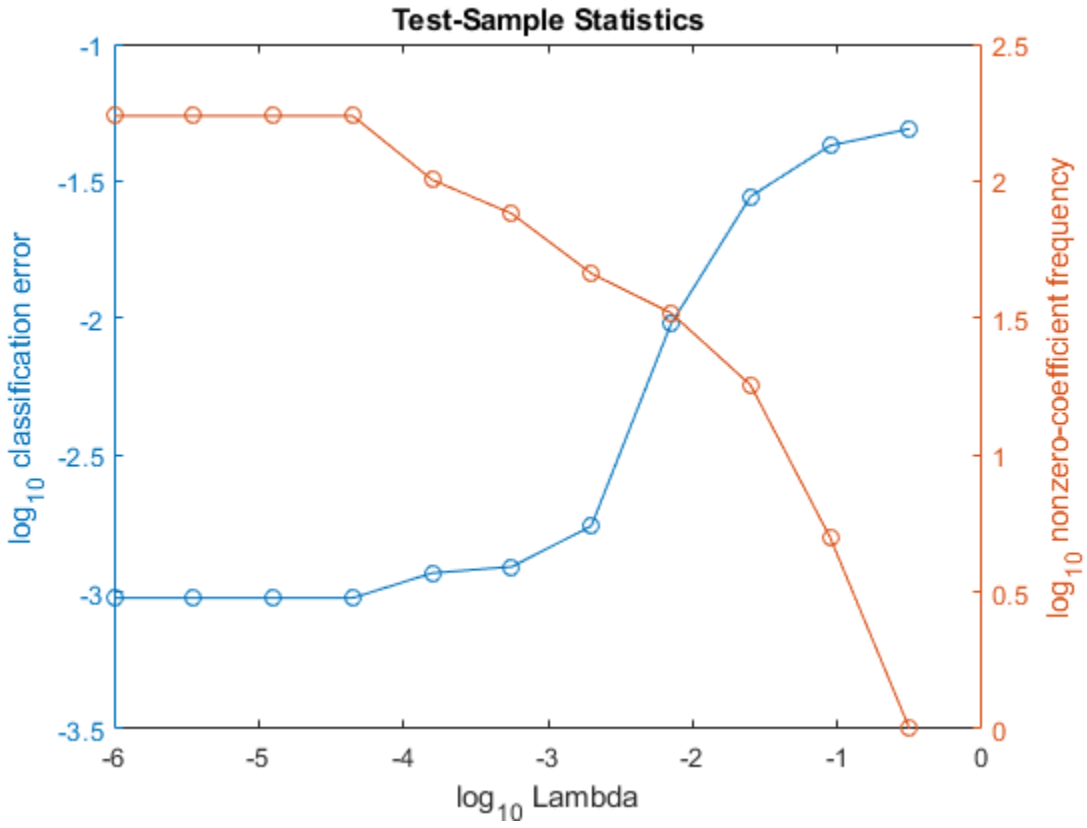
Because there are 11 regularization strengths, `ce` is a 1-by-11 vector of classification error rates.

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated, classification error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(ce),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} classification error')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
title('Test-Sample Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low classification error. In this case, a value between 10^{-4} to 10^{-1} should suffice.

```
idxFinal = 7;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

Optimize Linear Classifier

This example shows how to minimize the cross-validation error in a linear classifier using `fitclinear`. The example uses the NLP data set.

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. Identify the relevant labels.

```
X = X';
Ystats = Y == 'stats';
```

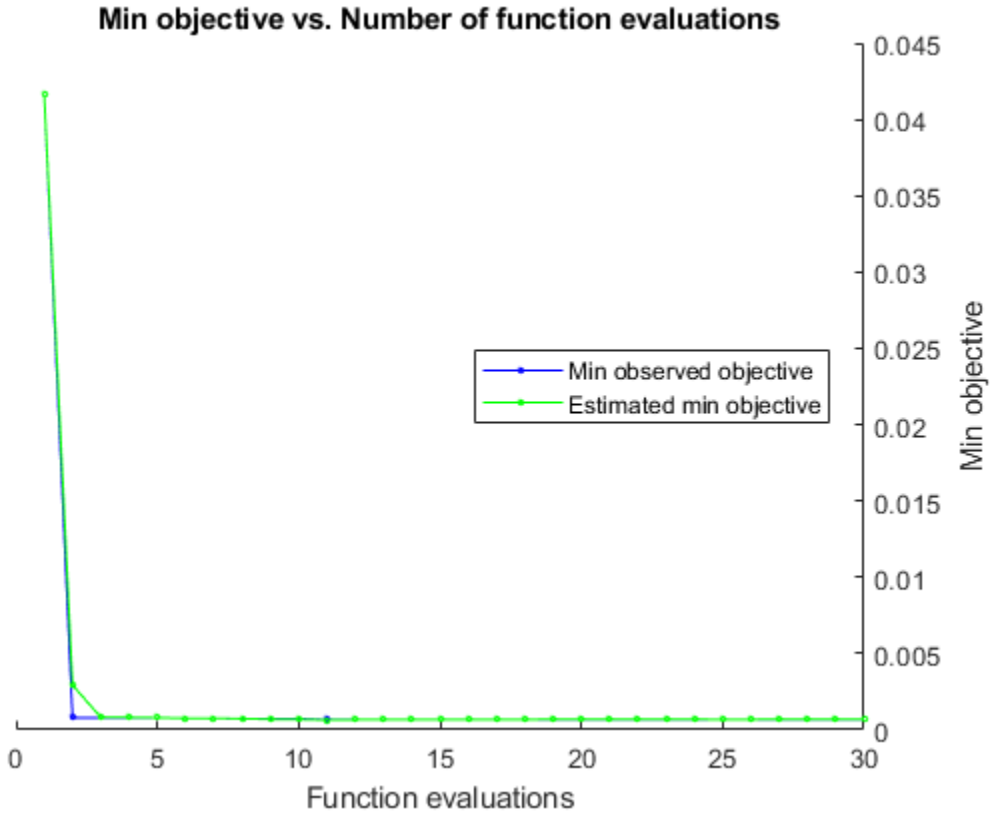
Optimize the classification using the `'auto'` parameters.

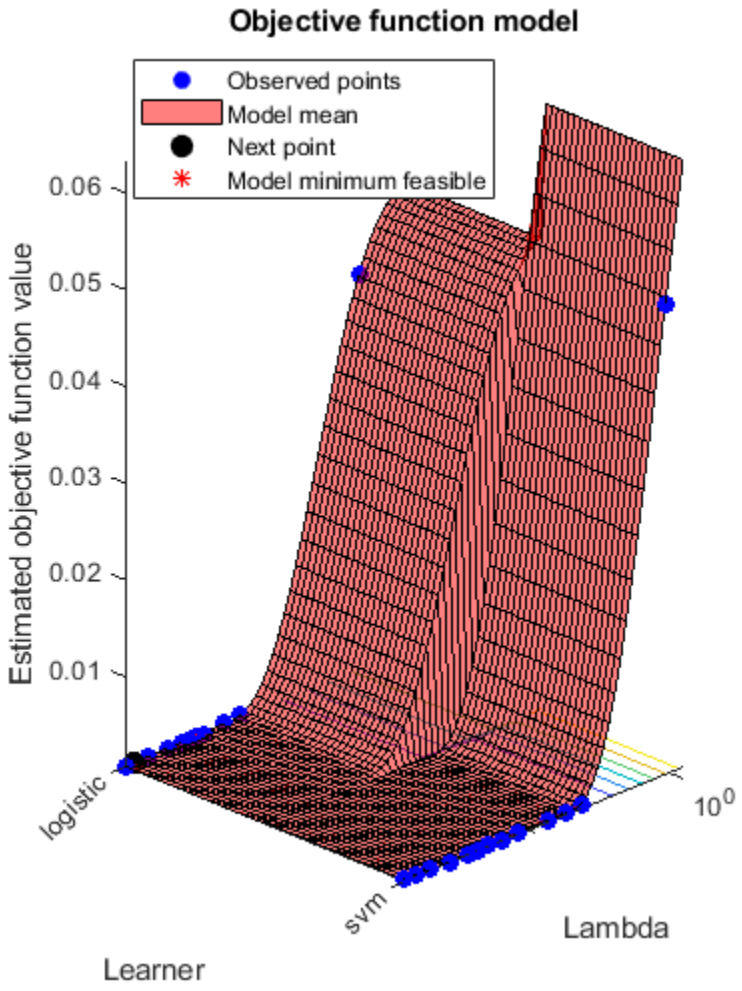
For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns','Solver','sparsa',...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |       Lea
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |    0.041619 |      6.4162 |    0.041619 |    0.041619 |     0.077903 |     logi
|    2 | Best   |  0.00079184 |      6.9651 |  0.00079184 |   0.0029367 |   2.1405e-09 |     logi
|    3 | Accept |    0.049221 |      8.0834 |  0.00079184 |  0.00082068 |      0.72101 |
|    4 | Accept |  0.00079184 |      7.7786 |  0.00079184 |    0.000815 |   3.4734e-07 |
|    5 | Accept |  0.00079184 |      7.4182 |  0.00079184 |  0.00079162 |   6.3377e-08 |
|    6 | Best   |  0.00076017 |      8.2979 |  0.00076017 |  0.00075995 |   3.1802e-10 |     logi
|    7 | Accept |  0.00088686 |      9.0694 |  0.00076017 |  0.00075831 |   3.1843e-10 |
|    8 | Best   |  0.00072849 |      11.275 |  0.00072849 |  0.00073699 |   1.9789e-07 |     logi
|    9 | Accept |  0.00072849 |      7.1112 |  0.00072849 |  0.00069632 |   3.3537e-08 |     logi
|   10 | Accept |  0.00085519 |      9.0838 |  0.00072849 |  0.00068786 |   2.6584e-09 |
|   11 | Best   |  0.00066515 |      10.303 |  0.00066515 |  0.00065095 |   7.3921e-08 |     logi
|   12 | Accept |  0.00069682 |      14.268 |  0.00066515 |  0.00066577 |   7.6302e-08 |     logi
|   13 | Accept |  0.00079184 |      13.226 |  0.00066515 |   0.0006663 |   1.3253e-07 |
|   14 | Accept |  0.00066515 |      12.937 |  0.00066515 |  0.00066622 |    7.944e-08 |     logi
|   15 | Accept |   0.0011719 |       17.91 |  0.00066515 |  0.00066815 |   0.00072837 |
|   16 | Accept |  0.00091854 |      14.372 |  0.00066515 |  0.00067008 |   4.6412e-05 |
|   17 | Accept |  0.00091854 |      8.7692 |  0.00066515 |  0.00067065 |   0.00021095 |
|   18 | Accept |  0.00082351 |      13.964 |  0.00066515 |  0.00067095 |   3.6744e-06 |
|   19 | Accept |   0.0010769 |      48.007 |  0.00066515 |   0.0006907 |   4.4922e-06 |     logi
|   20 | Accept |  0.00095021 |      54.565 |  0.00066515 |  0.00068144 |   1.0621e-06 |     logi
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |       Lambda |       Lea
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|   21 | Accept |  0.00076017 |      10.595 |  0.00066515 |  0.00068192 |   1.3685e-08 |
|   22 | Accept |  0.00076017 |      10.429 |  0.00066515 |  0.00068255 |    1.026e-08 |     logi
```

```
|   23 | Accept |  0.00072849 |      11.835 |  0.00066515 |  0.00068244 |   7.3386e-10 |     logi
|   24 | Accept |  0.00076017 |      13.756 |  0.00066515 |  0.00068257 |   1.0619e-06 |
|   25 | Accept |  0.00066515 |      11.287 |  0.00066515 |  0.00067854 |   5.3127e-08 |     logi
|   26 | Accept |    0.049221 |      1.0728 |  0.00066515 |  0.00067885 |       3.1627 |     logi
|   27 | Accept |  0.00076017 |      10.107 |  0.00066515 |  0.00068568 |    3.966e-08 |     logi
|   28 | Accept |  0.00076017 |       17.23 |  0.00066515 |  0.00069591 |   1.0856e-07 |     logi
|   29 | Accept |  0.00072849 |      11.337 |  0.00066515 |  0.00069581 |   3.1871e-10 |     logi
|   30 | Accept |  0.00085519 |      22.621 |  0.00066515 |  0.00069593 |   8.6432e-10 |
```






```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 457.5373 seconds
Total objective function evaluation time: 410.09

Best observed feasible point:
      Lambda      Learner
    __________    ________

    7.3921e-08    logistic

Observed objective function value = 0.00066515
Estimated objective function value = 0.00069593
Function evaluation time = 10.3034

Best estimated feasible point (according to models):
      Lambda      Learner
    __________    ________
```

```
    7.3921e-08    logistic

Estimated objective function value = 0.00069593
Estimated function evaluation time = 12.1805

Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x1 double]
              Bias: -10.1388
            Lambda: 7.3921e-08
           Learner: 'logistic'


  Properties, Methods
```

Input Arguments

X — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix.

The length of `Y` and the number of observations in `X` must be equal.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels to which the classification model is trained, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

- `fitclinear` only supports binary classification. Either `Y` must contain exactly two distinct classes, or you must specify two classes for training using the `'ClassNames'` name-value pair argument. For multiclass learning, see `fitcecoc`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.
- A good practice is to specify the class order using the `ClassNames` name-value pair argument.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response variable (for example, `Y` or `ValidationData{2}`)

- At least one missing value in a predictor observation (for example, row in `X` or `ValidationData{1}`)
- `NaN` value or `0` weight (for example, value in `Weights` or `ValidationData{3}`)

For memory-usage economy, it is best practice to remove observations containing missing values from your training data manually before training.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'ObservationsIn','columns','Learner','logistic','CrossVal','on'` specifies that the columns of the predictor matrix corresponds to observations, to implement logistic regression, to implement 10-fold cross-validation.

Linear Classification Options

`Lambda` — Regularization term strength
`'auto'` (default) | nonnegative scalar | vector of nonnegative values

Regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and `'auto'`, a nonnegative scalar, or a vector of nonnegative values.

- For `'auto'`, `Lambda` = 1/*n*.
  - If you specify a cross-validation, name-value pair argument (e.g., `CrossVal`), then *n* is the number of in-fold observations.
  - Otherwise, *n* is the training sample size.
- For a vector of nonnegative values, `fitclinear` sequentially optimizes the objective function for each distinct value in `Lambda` in ascending order.
  - If `Solver` is `'sgd'` or `'asgd'` and `Regularization` is `'lasso'`, `fitclinear` does not use the previous coefficient estimates as a warm start on page 33-1767 for the next optimization iteration. Otherwise, `fitclinear` uses warm starts.
  - If `Regularization` is `'lasso'`, then any coefficient estimate of 0 retains its value when `fitclinear` optimizes using subsequent values in `Lambda`.
  - `fitclinear` returns coefficient estimates for each specified regularization strength.

Example: `'Lambda',10.^(-(10:-2:2))`

Data Types: `char` | `string` | `double` | `single`

`Learner` — Linear classification model type
`'svm'` (default) | `'logistic'`

Linear classification model type, specified as the comma-separated pair consisting of `'Learner'` and `'svm'` or `'logistic'`.

In this table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.
- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Response Range	Loss Function
`'svm'`	Support vector machine	$y \in \{-1,1\}$; 1 for the positive class and –1 otherwise	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$
`'logistic'`	Logistic regression	Same as `'svm'`	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$

Example: `'Learner','logistic'`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

`Regularization` — Complexity penalty type
`'lasso'` | `'ridge'`

Complexity penalty type, specified as the comma-separated pair consisting of `'Regularization'` and `'lasso'` or `'ridge'`.

The software composes the objective function for minimization from the sum of the average loss function (see `Learner`) and the regularization term in this table.

Value	Description
`'lasso'`	Lasso (L1) penalty: $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$
`'ridge'`	Ridge (L2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

To specify the regularization term strength, which is λ in the expressions, use `Lambda`.

The software excludes the bias term (β_0) from the regularization penalty.

If `Solver` is `'sparsa'`, then the default value of `Regularization` is `'lasso'`. Otherwise, the default is `'ridge'`.

Tip

- For predictor variable selection, specify `'lasso'`. For more on variable selection, see “Introduction to Feature Selection” on page 15-49.
- For optimization accuracy, specify `'ridge'`.

Example: `'Regularization','lasso'`

`Solver` — Objective function minimization technique

`'sgd'` | `'asgd'` | `'dual'` | `'bfgs'` | `'lbfgs'` | `'sparsa'` | string array | cell array of character vectors

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a character vector or string scalar, a string array, or a cell array of character vectors with values from this table.

Value	Description	Restrictions
`'sgd'`	Stochastic gradient descent (SGD) [4][2]	
`'asgd'`	Average stochastic gradient descent (ASGD) [7]	
`'dual'`	Dual SGD for SVM [1][6]	`Regularization` must be `'ridge'` and `Learner` must be `'svm'`.
`'bfgs'`	Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (BFGS) [3]	Inefficient if `X` is very high-dimensional.
`'lbfgs'`	Limited-memory BFGS (LBFGS) [3]	`Regularization` must be `'ridge'`.
`'sparsa'`	Sparse Reconstruction by Separable Approximation (SpaRSA) [5]	`Regularization` must be `'lasso'`.

If you specify:

- A ridge penalty (see `Regularization`) and `X` contains 100 or fewer predictor variables, then the default solver is `'bfgs'`.
- An SVM model (see `Learner`), a ridge penalty, and `X` contains more than 100 predictor variables, then the default solver is `'dual'`.
- A lasso penalty and `X` contains 100 or fewer predictor variables, then the default solver is `'sparsa'`.

Otherwise, the default solver is `'sgd'`.

If you specify a string array or cell array of solver names, then the software uses all solvers in the specified order for each `Lambda`.

For more details on which solver to choose, see “Tips” on page 33-1768.

Example: `'Solver',{'sgd','lbfgs'}`

`Beta` — Initial linear coefficient estimates
`zeros(p,1)` (default) | numeric vector | numeric matrix

Initial linear coefficient estimates (β), specified as the comma-separated pair consisting of `'Beta'` and a p-dimensional numeric vector or a p-by-L numeric matrix. p is the number of predictor variables in `X` and L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a p-dimensional vector, then the software optimizes the objective function L times using this process.
  1. The software optimizes using `Beta` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The software optimizes again using the resulting estimate from the previous optimization as a warm start on page 33-1767, and the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.
- If you specify a p-by-L matrix, then the software optimizes the objective function L times. At iteration j, the software uses `Beta(:,j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.

If you set `'Solver','dual'`, then the software ignores `Beta`.

Data Types: `single` | `double`

`Bias` — Initial intercept estimate
numeric scalar | numeric vector

Initial intercept estimate (b), specified as the comma-separated pair consisting of `'Bias'` and a numeric scalar or an L-dimensional numeric vector. L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a scalar, then the software optimizes the objective function L times using this process.
  1. The software optimizes using `Bias` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The uses the resulting estimate as a warm start to the next optimization iteration, and uses the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.
- If you specify an L-dimensional vector, then the software optimizes the objective function L times. At iteration j, the software uses `Bias(j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.
- By default:
  - If `Learner` is `'logistic'`, then let g_j be 1 if `Y(j)` is the positive class, and -1 otherwise. `Bias` is the weighted average of the g for training or, for cross-validation, in-fold observations.

- If `Learner` is `'svm'`, then `Bias` is 0.

Data Types: `single` | `double`

`FitBias` — Linear model intercept inclusion flag

`true` (default) | `false`

Linear model intercept inclusion flag, specified as the comma-separated pair consisting of `'FitBias'` and `true` or `false`.

Value	Description
`true`	The software includes the bias term b in the linear model, and then estimates it.
`false`	The software sets $b = 0$ during estimation.

Example: `'FitBias',false`

Data Types: `logical`

`PostFitBias` — Flag to fit linear model intercept after optimization

`false` (default) | `true`

Flag to fit the linear model intercept after optimization, specified as the comma-separated pair consisting of `'PostFitBias'` and `true` or `false`.

Value	Description
`false`	The software estimates the bias term b and the coefficients β during optimization.
`true`	To estimate b, the software: **1** Estimates β and b using the model **2** Estimates classification scores **3** Refits b by placing the threshold on the classification scores that attains maximum accuracy

If you specify `true`, then `FitBias` must be true.

Example: `'PostFitBias',true`

Data Types: `logical`

`Verbose` — Verbosity level

`0` (default) | nonnegative integer

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and a nonnegative integer. `Verbose` controls the amount of diagnostic information `fitclinear` displays at the command line.

Value	Description
`0`	`fitclinear` does not display diagnostic information.

Value	Description
1	`fitclinear` periodically displays and stores the value of the objective function, gradient magnitude, and other diagnostic information. `FitInfo.History` contains the diagnostic information.
Any other positive integer	`fitclinear` displays and stores diagnostic information at each optimization iteration. `FitInfo.History` contains the diagnostic information.

Example: `'Verbose',1`

Data Types: `double` | `single`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size
positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration, the software estimates the subgradient using `BatchSize` observations from the training data.

- If `X` is a numeric matrix, then the default value is `10`.
- If `X` is a sparse matrix, then the default value is `max([10,ceil(sqrt(ff))])`, where `ff = numel(X)/nnz(X)` (the fullness factor of `X`).

Example: `'BatchSize',100`

Data Types: `single` | `double`

`LearnRate` — Learning rate
positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and a positive scalar. `LearnRate` controls the optimization step size by scaling the subgradient.

- If `Regularization` is `'ridge'`, then `LearnRate` specifies the initial learning rate γ_0. `fitclinear` determines the learning rate for iteration t, γ_t, using

  $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$

  - λ is the value of `Lambda`.
  - If `Solver` is `'sgd'`, then $c = 1$.
  - If `Solver` is `'asgd'`, then c is 0.75 [7].
- If `Regularization` is `'lasso'`, then, for all iterations, `LearnRate` is constant.

By default, `LearnRate` is `1/sqrt(1+max((sum(X.^2,obsDim))))`, where `obsDim` is `1` if the observations compose the columns of the predictor data `X`, and `2` otherwise.

Example: `'LearnRate',0.01`

Data Types: `single` | `double`

`OptimizeLearnRate` — Flag to decrease learning rate

`true` (default) | `false`

Flag to decrease the learning rate when the software detects divergence (that is, over-stepping the minimum), specified as the comma-separated pair consisting of `'OptimizeLearnRate'` and `true` or `false`.

If `OptimizeLearnRate` is `'true'`, then:

1. For the few optimization iterations, the software starts optimization using `LearnRate` as the learning rate.
2. If the value of the objective function increases, then the software restarts and uses half of the current value of the learning rate.
3. The software iterates step 2 until the objective function decreases.

Example: `'OptimizeLearnRate',true`

Data Types: `logical`

`TruncationPeriod` — Number of mini-batches between lasso truncation runs

`10` (default) | positive integer

Number of mini-batches between lasso truncation runs, specified as the comma-separated pair consisting of `'TruncationPeriod'` and a positive integer.

After a truncation run, the software applies a soft threshold to the linear coefficients. That is, after processing k = `TruncationPeriod` mini-batches, the software truncates the estimated coefficient j using

$$\widehat{\beta}_j^* = \begin{cases} \widehat{\beta}_j - u_t & \text{if } \widehat{\beta}_j > u_t, \\ 0 & \text{if } \left|\widehat{\beta}_j\right| \le u_t, \\ \widehat{\beta}_j + u_t & \text{if } \widehat{\beta}_j < -u_t. \end{cases}$$

- For SGD, $\widehat{\beta}_j$ is the estimate of coefficient j after processing k mini-batches. $u_t = k\gamma_t\lambda$. γ_t is the learning rate at iteration t. λ is the value of `Lambda`.
- For ASGD, $\widehat{\beta}_j$ is the averaged estimate coefficient j after processing k mini-batches, $u_t = k\lambda$.

If `Regularization` is `'ridge'`, then the software ignores `TruncationPeriod`.

Example: `'TruncationPeriod',100`

Data Types: `single` | `double`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table. The descriptions assume that the predictor data has observations in rows and predictors in columns.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitclinear` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitclinear` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitclinear` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitclinear` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitclinear` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitclinear` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.

- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure.

- If you specify the square matrix `cost` (`'Cost',cost`), then `cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class, and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `cost`, use the `ClassNames` name-value pair argument.
- If you specify the structure `S` (`'Cost',S`), then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default value for `Cost` is `ones(`K`) − eye(`K`)`, where K is the number of distinct classes.

`fitclinear` uses `Cost` to adjust the prior class probabilities specified in `Prior`. Then, `fitclinear` uses the adjusted prior probabilities for training and resets the cost matrix to its default.

Example: `'Cost',[0 2; 1 0]`

Data Types: `single` | `double` | `struct`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the predictor order in `X`. Assuming that `X` has the default orientation, with observations in rows and predictors in columns, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitclinear` uses only the predictor variables in `PredictorNames` and the response variable during training.

- `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
- By default, `PredictorNames` contains the names of all predictor variables.
- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities

`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and `'empirical'`, `'uniform'`, a numeric vector, or a structure array.

This table summarizes the available options for setting prior probabilities.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to their order in `Y`. If you specify the order using the `'ClassNames'` name-value pair argument, then order the elements accordingly.
structure array	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities.

`fitclinear` normalizes the prior probabilities in `Prior` to sum to 1.

Example: `'Prior',struct('ClassNames',{{'setosa','versicolor'}},'ClassProbs',1:2)`

Data Types: `char` | `string` | `double` | `single` | `struct`

`ResponseName` — Response variable name

`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, or `KFold`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` partition object as created by `cvpartition`. The partition object specifies the type of cross-validation, and also the indexing for training and validation sets.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of data used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range (0,1). If you specify `'Holdout',`*p*, then the software:

1. Randomly reserves *p*`*100`% of the data as validation data, and trains the model using the rest of the data
2. Stores the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated classifier, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1. If you specify, e.g., `'KFold',`*k*, then the software:

1. Randomly partitions the data into *k* sets
2. For each set, reserves the set as validation data, and trains the model using the other *k* – 1 sets
3. Stores the *k* compact, trained models in the cells of a *k*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

Example: `'KFold',8`

Data Types: `single` | `double`

SGD and ASGD Convergence Controls

`BatchLimit` — Maximal number of batches

positive integer

Maximal number of batches to process, specified as the comma-separated pair consisting of `'BatchLimit'` and a positive integer. When the software processes `BatchLimit` batches, it terminates optimization.

- By default:
  - The software passes through the data `PassLimit` times.
  - If you specify multiple solvers, and use (A)SGD to get an initial approximation for the next solver, then the default value is `ceil(1e6/BatchSize)`. `BatchSize` is the value of the `'BatchSize'` name-value pair argument.
- If you specify `'BatchLimit'` and `'PassLimit'`, then the software chooses the argument that results in processing the fewest observations.
- If you specify `'BatchLimit'` but not `'PassLimit'`, then the software processes enough batches to complete up to one entire pass through the data.

Example: `'BatchLimit',100`

Data Types: `single` | `double`

`BetaTolerance` — Relative tolerance on linear coefficients and bias term

`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`NumCheckConvergence` — Number of batches to process before next convergence check

positive integer

Number of batches to process before next convergence check, specified as the comma-separated pair consisting of `'NumCheckConvergence'` and a positive integer.

To specify the batch size, see `BatchSize`.

The software checks for convergence about 10 times per pass through the entire data set by default.

Example: `'NumCheckConvergence',100`

Data Types: `single` | `double`

`PassLimit` — Maximal number of passes
1 (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer.

`fitclinear` processes all observations when it completes one pass through the data.

When `fitclinear` passes through the data `PassLimit` times, it terminates optimization.

If you specify `'BatchLimit'` and `'PassLimit'`, then `fitclinear` chooses the argument that results in processing the fewest observations.

Example: `'PassLimit',5`

Data Types: `single` | `double`

`ValidationData` — Validation data for optimization convergence detection
cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of `'ValidationData'` and a cell array or table.

During optimization, the software periodically estimates the loss of `ValidationData`. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-1769. To optimize hyperparameters using cross-validation, see cross-validation options such as `CrossVal`.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of class labels, then it must have the same number of elements as the number of observations in `ValidationData{1}`. The set of all distinct labels of `ValidationData{2}` must be a subset of all distinct labels of `Y`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

Dual SGD Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `DeltaGradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`DeltaGradientTolerance` — Gradient-difference tolerance
`1` (default) | nonnegative scalar

Gradient-difference tolerance between upper and lower pool Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862 violators, specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar.

- If the magnitude of the KKT violators is less than `DeltaGradientTolerance`, then the software terminates optimization.
- If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'DeltaGapTolerance',1e-2`

Data Types: `double` | `single`

`NumCheckConvergence` — Number of passes through entire data set to process before next convergence check
`5` (default) | positive integer

Number of passes through entire data set to process before next convergence check, specified as the comma-separated pair consisting of `'NumCheckConvergence'` and a positive integer.

Example: `'NumCheckConvergence',100`

Data Types: `single` | `double`

`PassLimit` — Maximal number of passes
`10` (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer.

When the software completes one pass through the data, it has processed all observations.

When the software passes through the data `PassLimit` times, it terminates optimization.

Example: `'PassLimit',5`

Data Types: `single` | `double`

`ValidationData` — Validation data for optimization convergence detection
cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of `'ValidationData'` and a cell array or table.

During optimization, the software periodically estimates the loss of `ValidationData`. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-1769. To optimize hyperparameters using cross-validation, see cross-validation options such as `CrossVal`.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of class labels, then it must have the same number of elements as the number of observations in `ValidationData{1}`. The set of all distinct labels of `ValidationData{2}` must be a subset of all distinct labels of `Y`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

BFGS, LBFGS, and SpaRSA Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term

`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance

`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in the software, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`HessianHistorySize` — Size of history buffer for Hessian approximation

`15` (default) | positive integer

Size of history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. That is, at each iteration, the software composes the Hessian using statistics from the latest `HessianHistorySize` iterations.

The software does not support `'HessianHistorySize'` for SpaRSA.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`IterationLimit` — Maximal number of optimization iterations
`1000` (default) | positive integer

Maximal number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer. `IterationLimit` applies to these values of `Solver`: `'bfgs'`, `'lbfgs'`, and `'sparsa'`.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

`ValidationData` — Validation data for optimization convergence detection
cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of `'ValidationData'` and a cell array or table.

During optimization, the software periodically estimates the loss of `ValidationData`. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-1769. To optimize hyperparameters using cross-validation, see cross-validation options such as `CrossVal`.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of class labels, then it must have the same number of elements as the number of observations in `ValidationData{1}`. The set of all distinct labels of `ValidationData{2}` must be a subset of all distinct labels of `Y`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize

`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Lambda','Learner'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitclinear` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitclinear` are:

- `Lambda` — `fitclinear` searches among positive values, by default log-scaled in the range `[1e-5/NumObservations,1e5/NumObservations]`.
- `Learner` — `fitclinear` searches among `'svm'` and `'logistic'`.
- `Regularization` — `fitclinear` searches among `'ridge'` and `'lasso'`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitclinear',meas,species);
params(1).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Linear Classifier” on page 33-1739.

Example: `'OptimizeHyperparameters','auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`

Field Name	Values	Default
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	

Field Name	Values	Default
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained linear classification model

`ClassificationLinear` model object | `ClassificationPartitionedLinear` cross-validated model object

Trained linear classification model, returned as a `ClassificationLinear` model object or `ClassificationPartitionedLinear` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `ClassificationPartitionedLinear` cross-validated model object. Otherwise, `Mdl` is a `ClassificationLinear` model object.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.Beta` in the Command Window to display the vector or matrix of estimated coefficients.

Note Unlike other classification models, and for economical memory usage, `ClassificationLinear` and `ClassificationPartitionedLinear` model objects do not store the training data or training process details (for example, convergence history).

`FitInfo` — Optimization details

structure array

Optimization details, returned as a structure array.

Fields specify final values or name-value pair argument specifications, for example, `Objective` is the value of the objective function when optimization terminates. Rows of multidimensional fields correspond to values of `Lambda` and columns correspond to values of `Solver`.

This table describes some notable fields.

Field	Description
`TerminationStatus`	• Reason for optimization termination • Corresponds to a value in `TerminationCode`
`FitTime`	Elapsed, wall-clock time in seconds

<table>
<tr><th>Field</th><th>Description</th></tr>
<tr><td>History</td><td>A structure array of optimization information for each iteration. The field Solver stores solver types using integer coding.
<table>
<tr><th>Integer</th><th>Solver</th></tr>
<tr><td>1</td><td>SGD</td></tr>
<tr><td>2</td><td>ASGD</td></tr>
<tr><td>3</td><td>Dual SGD for SVM</td></tr>
<tr><td>4</td><td>LBFGS</td></tr>
<tr><td>5</td><td>BFGS</td></tr>
<tr><td>6</td><td>SpaRSA</td></tr>
</table></td></tr>
</table>

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.History.Objective`.

It is good practice to examine `FitInfo` to assess whether convergence is satisfactory.

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table of hyperparameters and associated values

Cross-validation optimization of hyperparameters, returned as a `BayesianOptimization` object or a table of hyperparameters and associated values. The output is nonempty when the value of `'OptimizeHyperparameters'` is not `'none'`. The output value depends on the `Optimizer` field value of the `'HyperparameterOptimizationOptions'` name-value pair argument:

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

More About

Warm Start

A warm start is initial estimates of the beta coefficients and bias term supplied to an optimization routine for quicker convergence.

Alternatives for Lower-Dimensional Data

High-dimensional linear classification and regression models minimize objective functions relatively quickly, but at the cost of some accuracy, the numeric-only predictor variables restriction, and the model must be linear with respect to the parameters. If your predictor data set is low- through medium-dimensional, or contains heterogeneous variables, then you should use the appropriate classification or regression fitting function. To help you decide which fitting function is appropriate for your low-dimensional data set, use this table.

Model to Fit	Function	Notable Algorithmic Differences
SVM	• Binary classification: `fitcsvm` • Multiclass classification: `fitcecoc` • Regression: `fitrsvm`	• Computes the Gram matrix of the predictor variables, which is convenient for nonlinear kernel transformations. • Solves dual problem using SMO, ISDA, or *L*1 minimization via quadratic programming using `quadprog`.
Linear regression	• Least-squares without regularization: `fitlm` • Regularized least-squares using a lasso penalty: `lasso` • Ridge regression: `ridge` or `lasso`	• `lasso` implements cyclic coordinate descent.
Logistic regression	• Logistic regression without regularization: `fitglm`. • Regularized logistic regression using a lasso penalty: `lassoglm`	• `fitglm` implements iteratively reweighted least squares. • `lassoglm` implements cyclic coordinate descent.

Tips

- It is a best practice to orient your predictor matrix so that observations correspond to columns and to specify `'ObservationsIn','columns'`. As a result, you can experience a significant reduction in optimization-execution time.
- For better optimization accuracy when you have high-dimensional predictor data and the `Regularization` value is `'ridge'`, set any of these options for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'dual'` if `Learner` is `'svm'`
  - `{'sgd','lbfgs'}`
  - `{'asgd','lbfgs'}`
  - `{'dual','lbfgs'}` if `Learner` is `'svm'`

  Other options can result in poor optimization accuracy.
- For better optimization accuracy when you have moderate- to low-dimensional predictor data and the `Regularization` value is `'ridge'`, set `Solver` to `'bfgs'`.
- If `Regularization` is `'lasso'`, set any of these options for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'sparsa'`
  - `{'sgd','sparsa'}`

  - `{'asgd','sparsa'}`
- When choosing between SGD and ASGD, consider that:
  - SGD takes less time per iteration, but requires more iterations to converge.
  - ASGD requires fewer iterations to converge, but takes more time per iteration.
- If your predictor data has few observations but many predictor variables, then:
  - Specify `'PostFitBias',true`.
  - For SGD or ASGD solvers, set `PassLimit` to a positive integer that is greater than 1, for example, 5 or 10. This setting often results in better accuracy.
- For SGD and ASGD solvers, `BatchSize` affects the rate of convergence.
  - If `BatchSize` is too small, then `fitclinear` achieves the minimum in many iterations, but computes the gradient per iteration quickly.
  - If `BatchSize` is too large, then `fitclinear` achieves the minimum in fewer iterations, but computes the gradient per iteration slowly.
- Large learning rates (see `LearnRate`) speed up convergence to the minimum, but can lead to divergence (that is, over-stepping the minimum). Small learning rates ensure convergence to the minimum, but can lead to slow termination.
- When using lasso penalties, experiment with various values of `TruncationPeriod`. For example, set `TruncationPeriod` to `1`, `10`, and then `100`.
- For efficiency, `fitclinear` does not standardize predictor data. To standardize `X`, enter

  ```
  X = bsxfun(@rdivide,bsxfun(@minus,X,mean(X,2)),std(X,0,2));
  ```

  The code requires that you orient the predictors and observations as the rows and columns of `X`, respectively. Also, for memory-usage economy, the code replaces the original predictor data the standardized data.
- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- If you specify `ValidationData`, then, during objective-function optimization:
  - `fitclinear` estimates the validation loss of `ValidationData` periodically using the current model, and tracks the minimal estimate.
  - When `fitclinear` estimates a validation loss, it compares the estimate to the minimal estimate.
  - When subsequent, validation loss estimates exceed the minimal estimate five times, `fitclinear` terminates optimization.
- If you specify `ValidationData` and to implement a cross-validation routine (`CrossVal`, `CVPartition`, `Holdout`, or `KFold`), then:

  1. `fitclinear` randomly partitions `X` and `Y` (or `Tbl`) according to the cross-validation routine that you choose.

2. `fitclinear` trains the model using the training-data partition. During objective-function optimization, `fitclinear` uses `ValidationData` as another possible way to terminate optimization (for details, see the previous bullet).
3. Once `fitclinear` satisfies a stopping criterion, it constructs a trained model based on the optimized linear coefficients and intercept.
   a. If you implement *k*-fold cross-validation, and `fitclinear` has not exhausted all training-set folds, then `fitclinear` returns to Step 2 to train using the next training-set fold.
   b. Otherwise, `fitclinear` terminates training, and then returns the cross-validated model.
4. You can determine the quality of the cross-validated model. For example:
   - To determine the validation loss using the holdout or out-of-fold data from step 1, pass the cross-validated model to `kfoldLoss`.
   - To predict observations on the holdout or out-of-fold data from step 1, pass the cross-validated model to `kfoldPredict`.

References

[1] Hsieh, C. J., K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. "A Dual Coordinate Descent Method for Large-Scale Linear SVM." *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, 2001, pp. 408–415.

[2] Langford, J., L. Li, and T. Zhang. "Sparse Online Learning Via Truncated Gradient." *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

[4] Shalev-Shwartz, S., Y. Singer, and N. Srebro. "Pegasos: Primal Estimated Sub-Gradient Solver for SVM." *Proceedings of the 24th International Conference on Machine Learning, ICML '07*, 2007, pp. 807–814.

[5] Wright, S. J., R. D. Nowak, and M. A. T. Figueiredo. "Sparse Reconstruction by Separable Approximation." *Trans. Sig. Proc.*, Vol. 57, No 7, 2009, pp. 2479–2493.

[6] Xiao, Lin. "Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization." *J. Mach. Learn. Res.*, Vol. 11, 2010, pp. 2543–2596.

[7] Xu, Wei. "Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent." *CoRR*, abs/1107.2490, 2011.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `fitclinear` does not support tall `table` data.
- Some name-value pair arguments have different defaults compared to the default values for the in-memory `fitclinear` function. Supported name-value pair arguments, and any differences, are:

- `'ObservationsIn'` — Supports only `'rows'`.
- `'Lambda'` — Can be `'auto'` (default) or a scalar.
- `'Learner'`
- `'Regularization'` — Supports only `'ridge'`.
- `'Solver'` — Supports only `'lbfgs'`.
- `'FitBias'` — Supports only `true`.
- `'Verbose'` — Default value is `1`.
- `'Beta'`
- `'Bias'`
- `'ClassNames'`
- `'Cost'`
- `'Prior'`
- `'Weights'` — Value must be a tall array.
- `'HessianHistorySize'`
- `'BetaTolerance'` — Default value is relaxed to `1e–3`.
- `'GradientTolerance'` — Default value is relaxed to `1e–3`.
- `'IterationLimit'` — Default value is relaxed to `20`.
- `'OptimizeHyperparameters'` — Value of `'Regularization'` parameter must be `'ridge'`.
- `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.

- For tall arrays, `fitclinear` implements LBFGS by distributing the calculation of the loss and gradient among different parts of the tall array at each iteration. Other solvers are not available for tall arrays.

  When initial values for `Beta` and `Bias` are not given, `fitclinear` refines the initial estimates of the parameters by fitting the model locally to parts of the data and combining the coefficients by averaging.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationLinear` | `ClassificationPartitionedLinear` | `fitcecoc` | `fitckernel` | `fitcsvm` | `fitglm` | `fitrlinear` | `kfoldLoss` | `kfoldPredict` | `lassoglm` | `predict` | `templateLinear` | `testcholdout`

Introduced in R2016a

fitcnb

Train multiclass naive Bayes model

Syntax

```
Mdl = fitcnb(Tbl,ResponseVarName)
Mdl = fitcnb(Tbl,formula)
Mdl = fitcnb(Tbl,Y)

Mdl = fitcnb(X,Y)

Mdl = fitcnb( ___ ,Name,Value)
```

Description

`Mdl = fitcnb(Tbl,ResponseVarName)` returns a multiclass naive Bayes model (`Mdl`), trained by the predictors in table `Tbl` and class labels in the variable `Tbl.ResponseVarName`.

`Mdl = fitcnb(Tbl,formula)` returns a multiclass naive Bayes model (`Mdl`), trained by the predictors in table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitcnb(Tbl,Y)` returns a multiclass naive Bayes model (`Mdl`), trained by the predictors in the table `Tbl` and class labels in the array `Y`.

`Mdl = fitcnb(X,Y)` returns a multiclass naive Bayes model (`Mdl`), trained by predictors `X` and class labels `Y`.

`Mdl = fitcnb( ___ ,Name,Value)` returns a naive Bayes classifier with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can specify a distribution to model the data, prior probabilities for the classes, or the kernel smoothing window bandwidth.

Examples

Train a Naive Bayes Classifier

Load Fisher's iris data set.

```
load fisheriris
X = meas(:,3:4);
Y = species;
tabulate(Y)
```

```
       Value    Count   Percent
      setosa       50     33.33%
  versicolor       50     33.33%
   virginica       50     33.33%
```

The software can classify data with more than two classes using naive Bayes methods.

Train a naive Bayes classifier. It is good practice to specify the class order.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
         DistributionNames: {'normal'  'normal'}
    DistributionParameters: {3x2 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

By default, the software models the predictor distribution within each class using a Gaussian distribution having some mean and standard deviation. Use dot notation to display the parameters of a particular Gaussian fit, e.g., display the fit for the first feature within `setosa`.

```
setosaIndex = strcmp(Mdl.ClassNames,'setosa');
estimates = Mdl.DistributionParameters{setosaIndex,1}
```

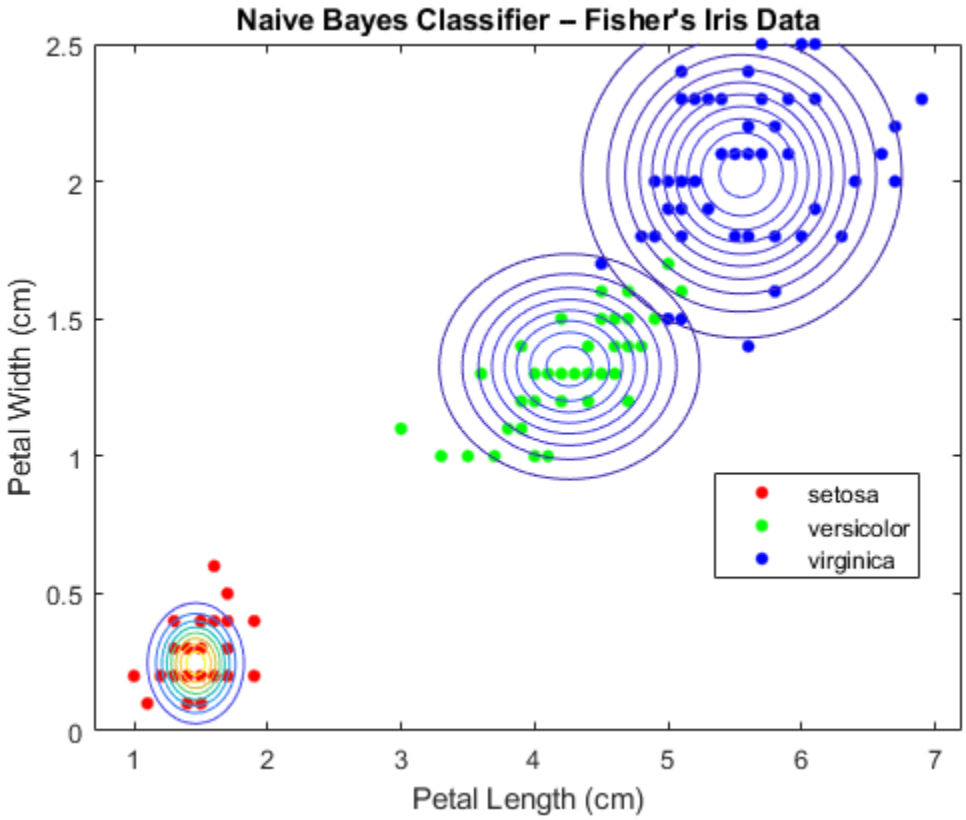
```
estimates = 2×1

    1.4620
    0.1737
```

The mean is `1.4620` and the standard deviation is `0.1737`.

Plot the Gaussian contours.

```
figure
gscatter(X(:,1),X(:,2),Y);
h = gca;
cxlim = h.XLim;
cylim = h.YLim;
hold on
Params = cell2mat(Mdl.DistributionParameters);
Mu = Params(2*(1:3)-1,1:2); % Extract the means
Sigma = zeros(2,2,3);
for j = 1:3
    Sigma(:,:,j) = diag(Params(2*j,:)).^2; % Create diagonal covariance matrix
    xlim = Mu(j,1) + 4*[-1 1]*sqrt(Sigma(1,1,j));
    ylim = Mu(j,2) + 4*[-1 1]*sqrt(Sigma(2,2,j));
    f = @(x,y) arrayfun(@(x0,y0) mvnpdf([x0 y0],Mu(j,:),Sigma(:,:,j)),x,y);
    fcontour(f,[xlim ylim]) % Draw contours for the multivariate normal distributions
end
h.XLim = cxlim;
h.YLim = cylim;
title('Naive Bayes Classifier -- Fisher''s Iris Data')
xlabel('Petal Length (cm)')
ylabel('Petal Width (cm)')
```

```
legend('setosa','versicolor','virginica')
hold off
```



You can change the default distribution using the name-value pair argument `'DistributionNames'`. For example, if some predictors are categorical, then you can specify that they are multivariate, multinomial random variables using `'DistributionNames','mvmn'`.

Specify Prior Probabilities When Training Naive Bayes Classifiers

Construct a naive Bayes classifier for Fisher's iris data set. Also, specify prior probabilities during training.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
classNames = {'setosa','versicolor','virginica'}; % Class order
```

X is a numeric matrix that contains four petal measurements for 150 irises. Y is a cell array of character vectors that contains the corresponding iris species.

By default, the prior class probability distribution is the relative frequency distribution of the classes in the data set. In this case the prior probability is 33% for each species. However, suppose you know

that in the population 50% of the irises are setosa, 20% are versicolor, and 30% are virginica. You can incorporate this information by specifying this distribution as a prior probability during training.

Train a naive Bayes classifier. Specify the class order and prior class probability distribution.

```
prior = [0.5 0.2 0.3];
Mdl = fitcnb(X,Y,'ClassNames',classNames,'Prior',prior)
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier, and some of its properties appear in the Command Window. The software treats the predictors as independent given a class, and, by default, fits them using normal distributions.

The naive Bayes algorithm does not use the prior class probabilities during training. Therefore, you can specify prior class probabilities after training using dot notation. For example, suppose that you want to see the difference in performance between a model that uses the default prior class probabilities and a model that uses different `prior`.

Create a new naive Bayes model based on `Mdl`, and specify that the prior class probability distribution is an empirical class distribution.

```
defaultPriorMdl = Mdl;
FreqDist = cell2table(tabulate(Y));
defaultPriorMdl.Prior = FreqDist{:,3};
```

The software normalizes the prior class probabilities to sum to `1`.

Estimate the cross-validation error for both models using 10-fold cross-validation.

```
rng(1); % For reproducibility
defaultCVMdl = crossval(defaultPriorMdl);
defaultLoss = kfoldLoss(defaultCVMdl)
```

```
defaultLoss = 0.0533
```

```
CVMdl = crossval(Mdl);
Loss = kfoldLoss(CVMdl)
```

```
Loss = 0.0340
```

`Mdl` performs better than `defaultPriorMdl`.

Specify Predictor Distributions for Naive Bayes Classifiers

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using every predictor. It is good practice to specify the class order.

```
Mdl1 = fitcnb(X,Y,...
    'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl1 =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

```
Mdl1.DistributionParameters
```

```
ans=3×4 cell array
    {2x1 double}    {2x1 double}    {2x1 double}    {2x1 double}
    {2x1 double}    {2x1 double}    {2x1 double}    {2x1 double}
    {2x1 double}    {2x1 double}    {2x1 double}    {2x1 double}
```

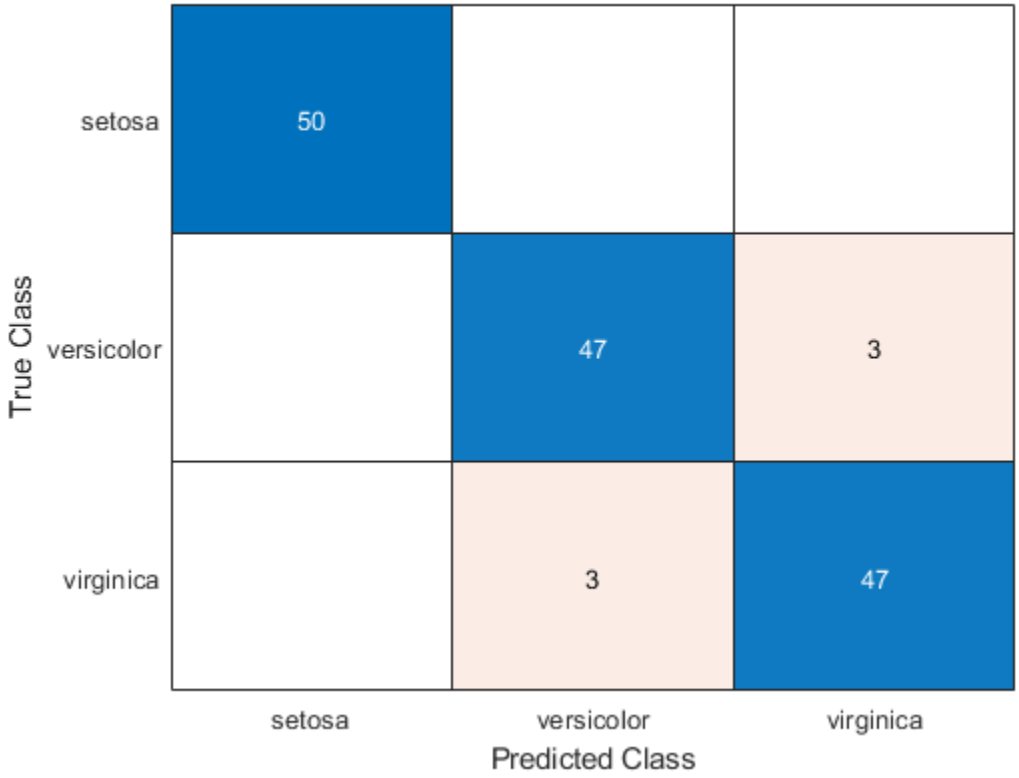
```
Mdl1.DistributionParameters{1,2}
```

```
ans = 2×1

    3.4280
    0.3791
```

By default, the software models the predictor distribution within each class as a Gaussian with some mean and standard deviation. There are four predictors and three class levels. Each cell in `Mdl1.DistributionParameters` corresponds to a numeric vector containing the mean and standard deviation of each distribution, e.g., the mean and standard deviation for setosa iris sepal widths are `3.4280` and `0.3791`, respectively.

Estimate the confusion matrix for `Mdl1`.

```
isLabels1 = resubPredict(Mdl1);
ConfusionMat1 = confusionchart(Y,isLabels1);
```



Element (*j*, *k*) of the confusion matrix chart represents the number of observations that the software classifies as *k*, but are truly in class *j* according to the data.

Retrain the classifier using the Gaussian distribution for predictors 1 and 2 (the sepal lengths and widths), and the default normal kernel density for predictors 3 and 4 (the petal lengths and widths).

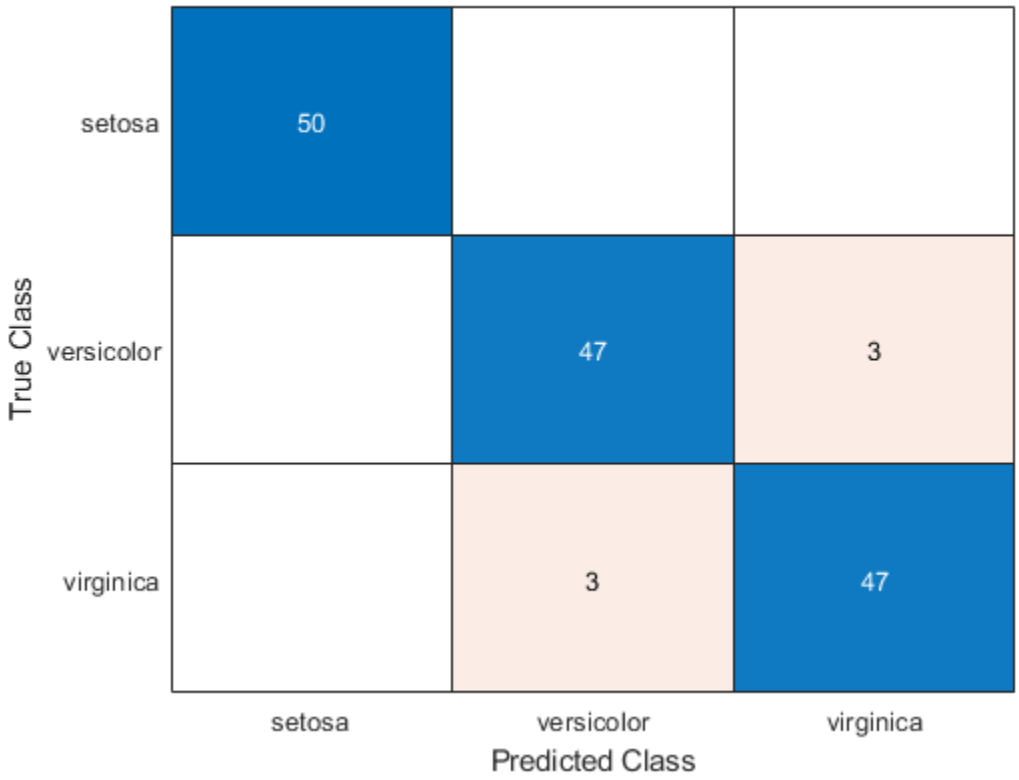
```
Mdl2 = fitcnb(X,Y,...
    'DistributionNames',{'normal','normal','kernel','kernel'},...
    'ClassNames',{'setosa','versicolor','virginica'});
Mdl2.DistributionParameters{1,2}
```

```
ans = 2×1

    3.4280
    0.3791
```

The software does not train parameters to the kernel density. Rather, the software chooses an optimal width. However, you can specify a width using the `'Width'` name-value pair argument.

Estimate the confusion matrix for `Mdl2`.

```
isLabels2 = resubPredict(Mdl2);
ConfusionMat2 = confusionchart(Y,isLabels2);
```



Based on the confusion matrices, the two classifiers perform similarly in the training sample.

Compare Classifiers Using Cross-Validation

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
rng(1); % For reproducibility
```

Train and cross-validate a naive Bayes classifier using the default options and *k*-fold cross-validation. It is good practice to specify the class order.

```
CVMdl1 = fitcnb(X,Y,...
    'ClassNames',{'setosa','versicolor','virginica'},...
    'CrossVal','on');
```

By default, the software models the predictor distribution within each class as a Gaussian with some mean and standard deviation. `CVMdl1` is a `ClassificationPartitionedModel` model.

Create a default naive Bayes binary classifier template, and train an error-correcting, output codes multiclass model.

```
t = templateNaiveBayes();
CVMdl2 = fitcecoc(X,Y,'CrossVal','on','Learners',t);
```

`CVMdl2` is a `ClassificationPartitionedECOC` model. You can specify options for the naive Bayes binary learners using the same name-value pair arguments as for `fitcnb`.

Compare the out-of-sample *k*-fold classification error (proportion of misclassified observations).

```
classErr1 = kfoldLoss(CVMdl1,'LossFun','ClassifErr')
```

```
classErr1 = 0.0533
```

```
classErr2 = kfoldLoss(CVMdl2,'LossFun','ClassifErr')
```

```
classErr2 = 0.0467
```

`Mdl2` has a lower generalization error.

Train Naive Bayes Classifiers Using Multinomial Predictors

Some spam filters classify an incoming email as spam based on how many times a word or punctuation (called tokens) occurs in an email. The predictors are the frequencies of particular words or punctuations in an email. Therefore, the predictors compose multinomial random variables.

This example illustrates classification using naive Bayes and multinomial predictors.

Create Training Data

Suppose you observed 1000 emails and classified them as spam or not spam. Do this by randomly assigning -1 or 1 to `y` for each email.

```
n = 1000;                        % Sample size
rng(1);                          % For reproducibility
Y = randsample([-1 1],n,true);  % Random labels
```

To build the predictor data, suppose that there are five tokens in the vocabulary, and 20 observed tokens per email. Generate predictor data from the five tokens by drawing random, multinomial deviates. The relative frequencies for tokens corresponding to spam emails should differ from emails that are not spam.

```
tokenProbs = [0.2 0.3 0.1 0.15 0.25;...
    0.4 0.1 0.3 0.05 0.15];             % Token relative frequencies
tokensPerEmail = 20;                    % Fixed for convenience
X = zeros(n,5);
X(Y == 1,:) = mnrnd(tokensPerEmail,tokenProbs(1,:),sum(Y == 1));
X(Y == -1,:) = mnrnd(tokensPerEmail,tokenProbs(2,:),sum(Y == -1));
```

Train the Classifier

Train a naive Bayes classifier. Specify that the predictors are multinomial.

```
Mdl = fitcnb(X,Y,'DistributionNames','mn');
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Assess the in-sample performance of `Mdl` by estimating the misclassification error.

```
isGenRate = resubLoss(Mdl,'LossFun','ClassifErr')
```

```
isGenRate = 0.0200
```

The in-sample misclassification rate is 2%.

Create New Data

Randomly generate deviates that represent a new batch of emails.

```
newN = 500;
newY = randsample([-1 1],newN,true);
newX = zeros(newN,5);
newX(newY == 1,:) = mnrnd(tokensPerEmail,tokenProbs(1,:),...
    sum(newY == 1));
newX(newY == -1,:) = mnrnd(tokensPerEmail,tokenProbs(2,:),...
    sum(newY == -1));
```

Assess Classifier Performance

Classify the new emails using the trained naive Bayes classifier `Mdl`, and determine whether the algorithm generalizes.

```
oosGenRate = loss(Mdl,newX,newY)
```

```
oosGenRate = 0.0261
```

The out-of-sample misclassification rate is 2.6% indicating that the classifier generalizes fairly well.

Optimize Naive Bayes Classifier

This example shows how to use the `OptimizeHyperparameters` name-value pair to minimize cross-validation loss in a naive Bayes classifier using `fitcnb`. The example uses Fisher's iris data.

Load Fisher's iris data.

```
load fisheriris
X = meas;
Y = species;
classNames = {'setosa','versicolor','virginica'};
```

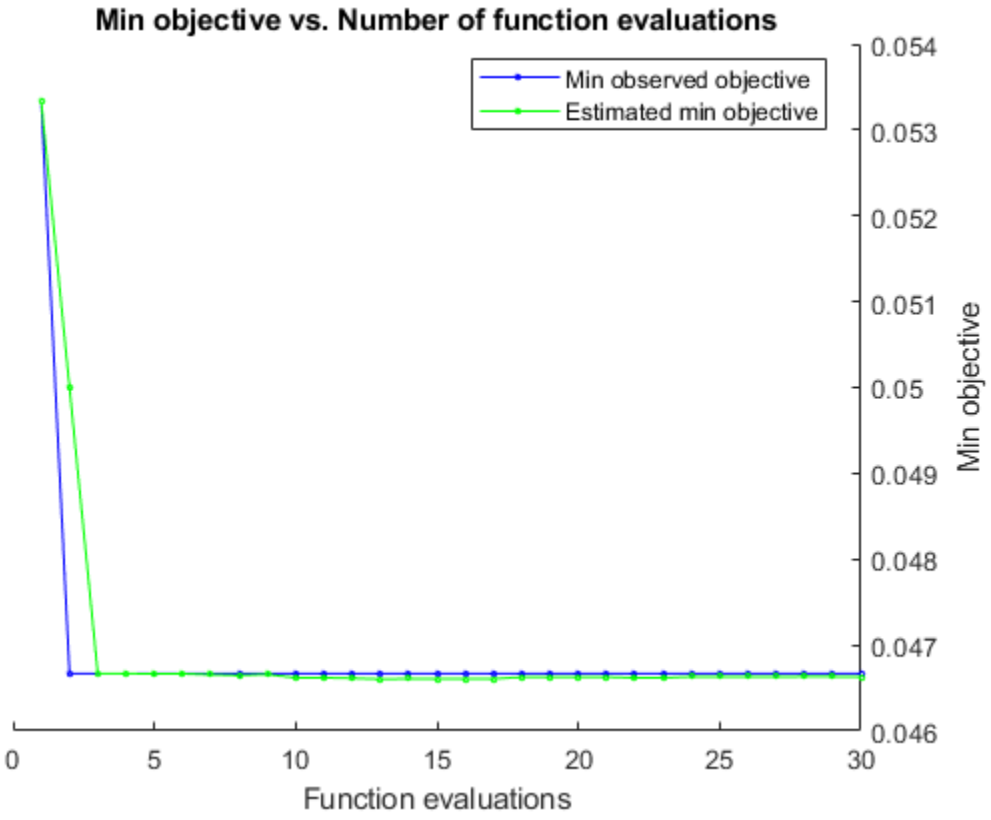
Optimize the classification using the 'auto' parameters.

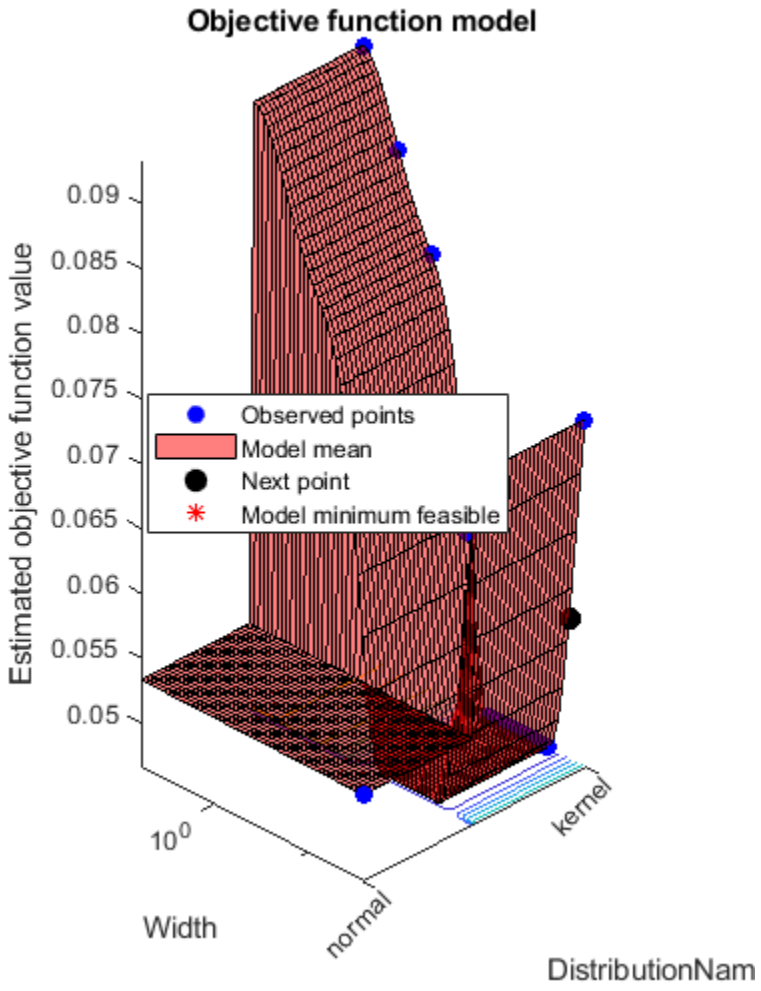
For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
Mdl = fitcnb(X,Y,'ClassNames',classNames,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName',...
    'expected-improvement-plus'))
```

```
Warning: It is recommended that you first standardize all numeric predictors when optimizing the

|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | Distribution-|        W
|      | result |             | runtime     | (observed)  | (estim.)    | Names        |
|=====================================================================================================
```

```
|    1 | Best   |    0.053333 |     0.88135 |    0.053333 |    0.053333 |       normal |
|    2 | Best   |    0.046667 |      1.0238 |    0.046667 |    0.049998 |       kernel |      0.1
|    3 | Accept |    0.053333 |      0.1809 |    0.046667 |    0.046667 |       normal |
|    4 | Accept |    0.086667 |     0.62091 |    0.046667 |    0.046668 |       kernel |        2
|    5 | Accept |    0.046667 |     0.63972 |    0.046667 |    0.046663 |       kernel |      0.1
|    6 | Accept |    0.073333 |     0.65658 |    0.046667 |    0.046665 |       kernel |     0.02
|    7 | Accept |    0.046667 |     0.69229 |    0.046667 |    0.046655 |       kernel |      0.2
|    8 | Accept |    0.046667 |      0.6157 |    0.046667 |    0.046647 |       kernel |        0
|    9 | Accept |        0.06 |     0.60976 |    0.046667 |    0.046658 |       kernel |      0.4
|   10 | Accept |    0.046667 |      0.6296 |    0.046667 |    0.046618 |       kernel |        0
|   11 | Accept |    0.046667 |     0.59977 |    0.046667 |    0.046619 |       kernel |     0.07
|   12 | Accept |    0.046667 |     0.61286 |    0.046667 |    0.046612 |       kernel |     0.08
|   13 | Accept |    0.046667 |     0.58707 |    0.046667 |    0.046603 |       kernel |      0.1
|   14 | Accept |    0.046667 |     0.56566 |    0.046667 |    0.046607 |       kernel |      0.2
|   15 | Accept |    0.046667 |     0.56852 |    0.046667 |    0.046606 |       kernel |      0.1
|   16 | Accept |    0.046667 |     0.59963 |    0.046667 |    0.046606 |       kernel |      0.1
|   17 | Accept |    0.046667 |     0.56891 |    0.046667 |    0.046606 |       kernel |     0.07
|   18 | Accept |    0.046667 |     0.58038 |    0.046667 |    0.046626 |       kernel |      0.2
|   19 | Accept |    0.046667 |      0.6118 |    0.046667 |    0.046626 |       kernel |     0.09
|   20 | Accept |    0.046667 |      0.5667 |    0.046667 |    0.046627 |       kernel |     0.06
|======|========|=============|=============|=============|=============|==============|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | Distribution-|
|      | result |             | runtime     | (observed)  | (estim.)    | Names        |
|======|========|=============|=============|=============|=============|==============|
|   21 | Accept |    0.046667 |     0.60543 |    0.046667 |    0.046627 |       kernel |     0.06
|   22 | Accept |    0.093333 |     0.56964 |    0.046667 |    0.046618 |       kernel |        5
|   23 | Accept |    0.046667 |     0.57433 |    0.046667 |    0.046619 |       kernel |     0.06
|   24 | Accept |    0.046667 |     0.58053 |    0.046667 |     0.04663 |       kernel |      0.2
|   25 | Accept |    0.046667 |     0.55826 |    0.046667 |     0.04663 |       kernel |        0
|   26 | Accept |    0.046667 |      0.5539 |    0.046667 |    0.046631 |       kernel |      0.1
|   27 | Accept |    0.046667 |     0.55845 |    0.046667 |    0.046631 |       kernel |      0.1
|   28 | Accept |    0.046667 |     0.57042 |    0.046667 |    0.046631 |       kernel |     0.06
|   29 | Accept |    0.046667 |     0.57446 |    0.046667 |    0.046631 |       kernel |        0
|   30 | Accept |        0.08 |     0.53443 |    0.046667 |    0.046628 |       kernel |        1
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
0.054
0.053
0.052
0.051
0.05
0.049
0.048
0.047
0.046
Min objective
0
5
10
15
20
25
30
Function evaluations




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 74.7772 seconds
Total objective function evaluation time: 18.0918

Best observed feasible point:
    DistributionNames     Width
    _________________    _______

          kernel          0.11903

Observed objective function value = 0.046667
Estimated objective function value = 0.046667
Function evaluation time = 1.0238

Best estimated feasible point (according to models):
    DistributionNames     Width
    _________________    _______
```

```
         kernel          0.25613

Estimated objective function value = 0.046628
Estimated function evaluation time = 0.58714

Mdl =
  ClassificationNaiveBayes
                          ResponseName: 'Y'
                 CategoricalPredictors: []
                            ClassNames: {'setosa'  'versicolor'  'virginica'}
                       ScoreTransform: 'none'
                      NumObservations: 150
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                     DistributionNames: {1x4 cell}
                DistributionParameters: {3x4 cell}
                               Kernel: {1x4 cell}
                              Support: {1x4 cell}
                                Width: [3x4 double]


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels to which the naive Bayes classifier is trained, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each element of `Y` defines the class membership of the corresponding row of `X`. `Y` supports K class levels.

If `Y` is a character array, then each row must correspond to one class label.

The length of `Y` and the number of rows of `X` must be equivalent.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature).

The length of `Y` and the number of rows of `X` must be equivalent.

Data Types: `double`

Note: The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data values.

- If `Y` contains missing values, then the software removes them and the corresponding rows of `X`.
- If `X` contains any rows composed entirely of missing values, then the software removes those rows and the corresponding elements of `Y`.
- If `X` contains missing values and you set `'DistributionNames','mn'`, then the software removes those rows of `X` and the corresponding elements of `Y`.

- If a predictor is not represented in a class, that is, if all of its values are `NaN` within a class, then the software returns an error.

Removing rows of `X` and corresponding elements of `Y` decreases the effective training or cross-validation sample size.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'DistributionNames','mn','Prior','uniform','KSWidth',0.5` specifies that the data distribution is multinomial, the prior probabilities for all classes are equal, and the kernel smoothing window bandwidth for all classes is `0.5` units.

Naive Bayes Options

`DistributionNames` — Data distributions
`'kernel'` | `'mn'` | `'mvmn'` | `'normal'` | string array | cell array of character vectors

Data distributions `fitcnb` uses to model the data, specified as the comma-separated pair consisting of `'DistributionNames'` and a character vector or string scalar, a string array, or a cell array of character vectors with values from this table.

Value	Description
`'kernel'`	Kernel smoothing density estimate.
`'mn'`	Multinomial distribution. If you specify `mn`, then all features are components of a multinomial distribution. Therefore, you cannot include `'mn'` as an element of a string array or a cell array of character vectors. For details, see “Algorithms” on page 33-1799.
`'mvmn'`	Multivariate multinomial distribution. For details, see “Algorithms” on page 33-1799.
`'normal'`	Normal (Gaussian) distribution.

If you specify a character vector or string scalar, then the software models all the features using that distribution. If you specify a 1-by-*P* string array or cell array of character vectors, then the software models feature *j* using the distribution in element *j* of the array.

By default, the software sets all predictors specified as categorical predictors (using the `CategoricalPredictors` name-value pair argument) to `'mvmn'`. Otherwise, the default distribution is `'normal'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'DistributionNames','mn'`

Example: `'DistributionNames',{'kernel','normal','kernel'}`

`Kernel` — Kernel smoother type
`'normal'` (default) | `'box'` | `'epanechnikov'` | `'triangle'` | string array | cell array of character vectors

Kernel smoother type, specified as the comma-separated pair consisting of `'Kernel'` and a character vector or string scalar, a string array, or a cell array of character vectors.

This table summarizes the available options for setting the kernel smoothing density region. Let $I\{u\}$ denote the indicator function.

Value	Kernel	Formula
`'box'`	Box (uniform)	$f(x) = 0.5I\{\|x\| \le 1\}$
`'epanechnikov'`	Epanechnikov	$f(x) = 0.75\left(1 - x^2\right)I\{\|x\| \le 1\}$
`'normal'`	Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-0.5x^2\right)$
`'triangle'`	Triangular	$f(x) = (1 - \|x\|)I\{\|x\| \le 1\}$

If you specify a 1-by-*P* string array or cell array, with each element of the array containing any value in the table, then the software trains the classifier using the kernel smoother type in element *j* for feature *j* in `X`. The software ignores elements of `Kernel` not corresponding to a predictor whose distribution is `'kernel'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'Kernel',{'epanechnikov','normal'}`

`Support` — Kernel smoothing density support
`'unbounded'` (default) | `'positive'` | string array | cell array | numeric row vector

Kernel smoothing density support, specified as the comma-separated pair consisting of `'Support'` and `'positive'`, `'unbounded'`, a string array, a cell array, or a numeric row vector. The software applies the kernel smoothing density to the specified region.

This table summarizes the available options for setting the kernel smoothing density region.

Value	Description
1-by-2 numeric row vector	For example, `[L,U]`, where `L` and `U` are the finite lower and upper bounds, respectively, for the density support.
`'positive'`	The density support is all positive real values.
`'unbounded'`	The density support is all real values.

If you specify a 1-by-*P* string array or cell array, with each element in the string array containing any text value in the table and each element in the cell array containing any value in the table, then the

software trains the classifier using the kernel support in element *j* for feature *j* in `X`. The software ignores elements of `Kernel` not corresponding to a predictor whose distribution is `'kernel'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'KSSupport',{[-10,20],'unbounded'}`

Data Types: `char` | `string` | `cell` | `double`

`Width` — Kernel smoothing window width

matrix of numeric values | numeric column vector | numeric row vector | scalar

Kernel smoothing window width, specified as the comma-separated pair consisting of `'Width'` and a matrix of numeric values, numeric column vector, numeric row vector, or scalar.

Suppose there are *K* class levels and *P* predictors. This table summarizes the available options for setting the kernel smoothing window width.

Value	Description
K-by-*P* matrix of numeric values	Element (*k*,*j*) specifies the width for predictor *j* in class *k*.
K-by-1 numeric column vector	Element *k* specifies the width for all predictors in class *k*.
1-by-*P* numeric row vector	Element *j* specifies the width in all class levels for predictor *j*.
scalar	Specifies the bandwidth for all features in all classes.

By default, the software selects a default width automatically for each combination of predictor and class by using a value that is optimal for a Gaussian distribution. If you specify `Width` and it contains `NaN`s, then the software selects widths for the elements containing `NaN`s.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'Width',[NaN NaN]`

Data Types: `double` | `struct`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'CrossVal','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other n – 1 observations.
2. Store the n compact, trained models in an n-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcnb` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitcnb` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitcnb` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitcnb` uses multivariate multinomial distributions. For details, see `DistributionNames` and “Algorithms” on page 33-1799.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Cost of misclassification
square matrix | structure

Cost of misclassification of a point, specified as the comma-separated pair consisting of `'Cost'` and one of the following:

- Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- Structure `S` having two fields: `S.ClassNames` containing the group names as a variable of the same type as `Y`, and `S.ClassificationCosts` containing the cost matrix.

The default is `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`.

Example: `'Cost',struct('ClassNames',{{'b','g'}},'ClassificationCosts',[0 0.5; 1 0])`

Data Types: `single` | `double` | `struct`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.

  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcnb` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities

`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to $1/K$, where K is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

If you set values for both `Weights` and `Prior`, the weights are renormalized to add up to the value of the prior probability in the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string

scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'DistributionNames','Width'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitcnb` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcnb` are:

- `DistributionNames` — `fitcnb` searches among `'normal'` and `'kernel'`.
- `Width` — `fitcnb` searches among real values, by default log-scaled in the range `[MinPredictorDiff/4,max(MaxPredictorRange,MinPredictorDiff)]`.
- `Kernel` — `fitcnb` searches among `'normal'`, `'box'`, `'epanechnikov'`, and `'triangle'`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitcnb',meas,species);
params(2).Range = [1e-2,1e2];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is

log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Naive Bayes Classifier” on page 33-1781.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`

Field Name	Values	Default
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		

Field Name	Values	Default
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained naive Bayes classification model

`ClassificationNaiveBayes` model object | `ClassificationPartitionedModel` cross-validated model object

Trained naive Bayes classification model, returned as a `ClassificationNaiveBayes` model object or a `ClassificationPartitionedModel` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `ClassificationPartitionedModel` cross-validated model object. Otherwise, `Mdl` is a `ClassificationNaiveBayes` model object.

To reference properties of `Mdl`, use dot notation. For example, to access the estimated distribution parameters, enter `Mdl.DistributionParameters`.

More About

Bag-of-Tokens Model

In the bag-of-tokens model, the value of predictor j is the nonnegative number of occurrences of token j in the observation. The number of categories (bins) in the multinomial model is the number of distinct tokens (number of predictors).

Naive Bayes

Naive Bayes is a classification algorithm that applies density estimation to the data.

The algorithm leverages Bayes theorem, and (naively) assumes that the predictors are conditionally independent, given the class. Although the assumption is usually violated in practice, naive Bayes classifiers tend to yield posterior distributions that are robust to biased class density estimates, particularly where the posterior is 0.5 (the decision boundary) [1].

Naive Bayes classifiers assign observations to the most probable class (in other words, the maximum a posteriori decision rule). Explicitly, the algorithm takes these steps:

1. Estimate the densities of the predictors within each class.
2. Model posterior probabilities according to Bayes rule. That is, for all $k = 1,...,K$,

$$\widehat{P}(Y = k \mid X_1, .., X_P) = \frac{\pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)},$$

where:

- Y is the random variable corresponding to the class index of an observation.
- $X_1,...,X_P$ are the random predictors of an observation.
- $\pi(Y = k)$ is the prior probability that a class index is k.

3 Classify an observation by estimating the posterior probability for each class, and then assign the observation to the class yielding the maximum posterior probability.

If the predictors compose a multinomial distribution, then the posterior probability $\widehat{P}(Y = k \mid X_1, .., X_P) \propto \pi(Y = k) P_{mn}(X_1, ..., X_P \mid Y = k)$, where $P_{mn}(X_1, ..., X_P \mid Y = k)$ is the probability mass function of a multinomial distribution.

Tips

- For classifying count-based data, such as the bag-of-tokens model on page 33-1798, use the multinomial distribution (e.g., set `'DistributionNames','mn'`).
- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- If you specify `'DistributionNames','mn'` when training `Mdl` using `fitcnb`, then the software fits a multinomial distribution using the bag-of-tokens model on page 33-1798. The software stores the probability that token j appears in class k in the property `DistributionParameters{k,j}`. Using additive smoothing [2], the estimated probability is

  $$P(\text{token } j \,|\, \text{class } k) = \frac{1 + c_{j|k}}{P + c_k},$$

  where:

  - $c_{j|k} = n_k \frac{\sum_{i: y_i \in \text{class } k} x_{ij} w_i}{\sum_{i: y_i \in \text{class } k} w_i}$, which is the weighted number of occurrences of token j in class k.
  - n_k is the number of observations in class k.
  - w_i is the weight for observation i. The software normalizes weights within a class such that they sum to the prior probability for that class.
  - $c_k = \sum_{j=1}^{P} c_{j|k}$, which is the total weighted number of occurrences of all tokens in class k.
- If you specify `'DistributionNames','mvmn'` when training `Mdl` using `fitcnb`, then:

1. For each predictor, the software collects a list of the unique levels, stores the sorted list in `CategoricalLevels`, and considers each level a bin. Each predictor/class combination is a separate, independent multinomial random variable.
2. For predictor j in class k, the software counts instances of each categorical level using the list stored in `CategoricalLevels{j}`.
3. The software stores the probability that predictor j, in class k, has level L in the property `DistributionParameters{k,j}`, for all levels in `CategoricalLevels{j}`. Using additive smoothing [2], the estimated probability is

$$P(\text{predictor } j = L \,|\, \text{class } k) = \frac{1 + m_{j|k}(L)}{m_j + m_k},$$

where:

- $m_{j|k}(L) = n_k \dfrac{\sum_{i:\, y_i \in \text{class } k} I\{x_{ij} = L\} w_i}{\sum_{i:\, y_i \in \text{class } k} w_i}$, which is the weighted number of observations for which predictor j equals L in class k.
- n_k is the number of observations in class k.
- $I\{x_{ij} = L\} = 1$ if $x_{ij} = L$, 0 otherwise.
- w_i is the weight for observation i. The software normalizes weights within a class such that they sum to the prior probability for that class.
- m_j is the number of distinct levels in predictor j.
- m_k is the weighted number of observations in class k.

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[2] Manning, C. D., P. Raghavan, and M. Schütze. *Introduction to Information Retrieval*, NY: Cambridge University Press, 2008.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays with the limitations:

- Supported syntaxes are:
  - `Mdl = fitcnb(Tbl,Y)`
  - `Mdl = fitcnb(X,Y)`
  - `Mdl = fitcnb(___,Name,Value)`
- Options related to kernel densities, cross-validation, and hyperparameter optimization are not supported. The supported name-value pair arguments are:

- `'DistributionNames'` — `'kernel'` value is not supported.
- `'CategoricalPredictors'`
- `'Cost'`
- `'PredictorNames'`
- `'Prior'`
- `'ResponseName'`
- `'ScoreTransform'`
- `'Weights'` — Value must be a tall array.

For more information, see “Tall Arrays for Out-of-Memory Data”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationNaiveBayes` | `ClassificationPartitionedModel` | `predict` | `templateNaiveBayes`

Topics
“Naive Bayes Classification” on page 21-2
“Grouping Variables” on page 2-45

Introduced in R2014b

fitcnet

Train neural network classification model

Syntax

```
Mdl = fitcnet(Tbl,ResponseVarName)
Mdl = fitcnet(Tbl,formula)
Mdl = fitcnet(Tbl,Y)

Mdl = fitcnet(X,Y)

Mdl = fitcnet( ___ ,Name,Value)
```

Description

Use `fitcnet` to train a feedforward, fully connected neural network for classification. The first fully connected layer of the neural network has a connection from the network input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer. The final fully connected layer and the subsequent softmax activation function produce the network's output, namely classification scores (posterior probabilities) and predicted labels. For more information, see “Neural Network Structure” on page 33-1823.

`Mdl = fitcnet(Tbl,ResponseVarName)` returns a neural network classification model `Mdl` trained using the predictors in the table `Tbl` and the class labels in the `ResponseVarName` table variable.

`Mdl = fitcnet(Tbl,formula)` returns a neural network classification model trained using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of the predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitcnet(Tbl,Y)` returns a neural network classification model using the predictor variables in the table `Tbl` and the class labels in vector `Y`.

`Mdl = fitcnet(X,Y)` returns a neural network classification model trained using the predictors in the matrix `X` and the class labels in vector `Y`.

`Mdl = fitcnet( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can adjust the number of outputs and the activation functions for the fully connected layers by specifying the `LayerSizes` and `Activations` name-value arguments.

Examples

Train Neural Network Classifier

Train a neural network classifier, and assess the performance of the classifier on a test set.

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response

variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable("CreditRating_Historical.dat");
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    ______    ______    _______    ________    _____    ________    _______

    62394     0.013     0.104     0.036      0.447      0.142        3       {'BB' }
    48608     0.232     0.335     0.062      1.969      0.281        8       {'A'  }
    42444     0.311     0.367     0.074      1.935      0.366        1       {'A'  }
    48631     0.194     0.263     0.062      1.017      0.228        4       {'BBB'}
    43768     0.121     0.413     0.057      3.647      0.466       12       {'AAA'}
    39255    -0.117    -0.799      0.01      0.179      0.082        4       {'CCC'}
    62236     0.087     0.158     0.049      0.816      0.324        2       {'BBB'}
    39354     0.005     0.181     0.034      2.597      0.388        7       {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,"ID");
creditrating.Industry = categorical(creditrating.Industry);
```

Convert the `Rating` response variable to an ordinal `categorical` variable.

```
creditrating.Rating = categorical(creditrating.Rating, ...
    ["AAA","AA","A","BBB","BB","B","CCC"],"Ordinal",true);
```

Partition the data into training and test sets. Use approximately 80% of the observations to train a neural network model, and 20% of the observations to test the performance of the trained model on new data. Use `cvpartition` to partition the data.

```
rng("default") % For reproducibility of the partition
c = cvpartition(creditrating.Rating,"Holdout",0.20);
trainingIndices = training(c); % Indices for the training set
testIndices = test(c); % Indices for the test set
creditTrain = creditrating(trainingIndices,:);
creditTest = creditrating(testIndices,:);
```

Train a neural network classifier by passing the training data `creditTrain` to the `fitcnet` function.

```
Mdl = fitcnet(creditTrain,"Rating")
```

```
Mdl =
  ClassificationNeuralNetwork
           PredictorNames: {'WC_TA'  'RE_TA'  'EBIT_TA'  'MVE_BVTD'  'S_TA'  'Industry'}
             ResponseName: 'Rating'
    CategoricalPredictors: 6
               ClassNames: [AAA    AA    A    BBB    BB    B    CCC]
           ScoreTransform: 'none'
          NumObservations: 3146
               LayerSizes: 10
              Activations: 'relu'
```

```
    OutputLayerActivation: 'softmax'
                   Solver: 'LBFGS'
         ConvergenceInfo: [1×1 struct]
         TrainingHistory: [1000×7 table]


  Properties, Methods
```

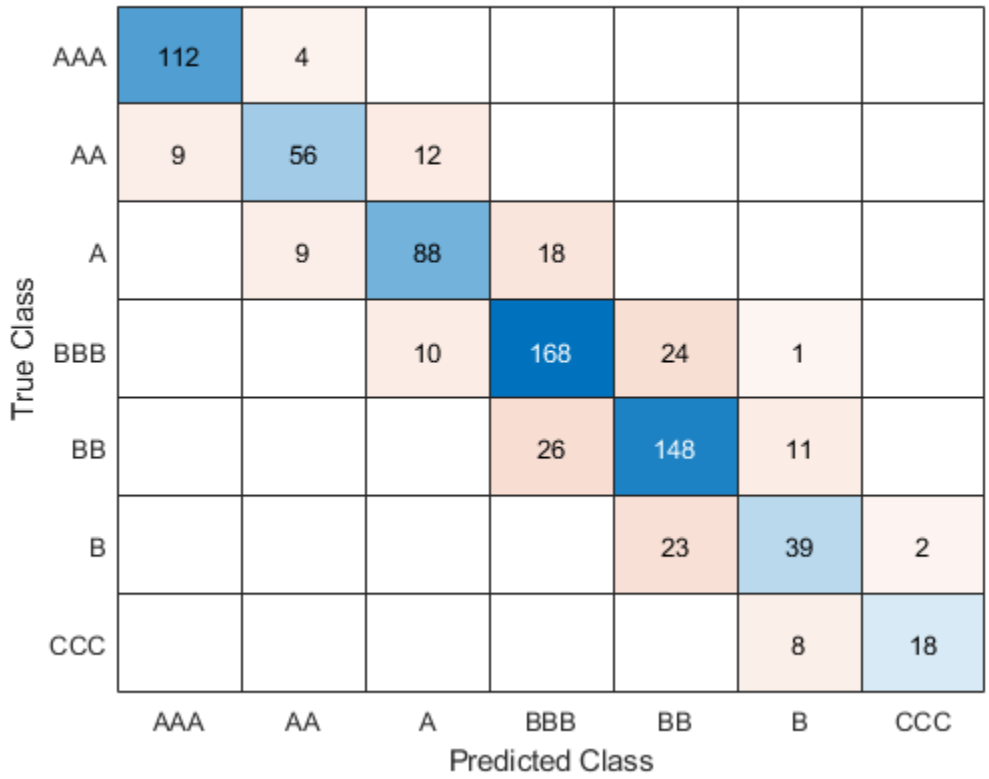
`Mdl` is a trained `ClassificationNeuralNetwork` classifier. You can use dot notation to access the properties of `Mdl`. For example, you can specify `Mdl.TrainingHistory` to get more information about the training history of the neural network model.

Evaluate the performance of the classifier on the test set by computing the test set classification error. Visualize the results by using a confusion matrix.

```
testAccuracy = 1 - loss(Mdl,creditTest,"Rating", ...
    "LossFun","classiferror")
```

```
testAccuracy = 0.8003
```

```
confusionchart(creditTest.Rating,predict(Mdl,creditTest))
```



Specify Neural Network Classifier Architecture

Specify the structure of a neural network classifier, including the size of the fully connected layers.

Load the `ionosphere` data set, which includes radar signal data. `X` contains the predictor data, and `Y` is the response variable, whose values represent either good ("g") or bad ("b") radar signals.

```
load ionosphere
```

Separate the data into training data (`XTrain` and `YTrain`) and test data (`XTest` and `YTest`) by using a stratified holdout partition. Reserve approximately 30% of the observations for testing, and use the rest of the observations for training.

```
rng("default") % For reproducibility of the partition
cvp = cvpartition(Y,"Holdout",0.3);
XTrain = X(training(cvp),:);
YTrain = Y(training(cvp));
XTest = X(test(cvp),:);
YTest = Y(test(cvp));
```

Train a neural network classifier. Specify to have 35 outputs in the first fully connected layer and 20 outputs in the second fully connected layer. By default, both layers use a rectified linear unit (ReLU) activation function. You can change the activation functions for the fully connected layers by using the `Activations` name-value argument.

```
Mdl = fitcnet(XTrain,YTrain, ...
    "LayerSizes",[35 20])
```

```
Mdl =
  ClassificationNeuralNetwork
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 246
               LayerSizes: [35 20]
              Activations: 'relu'
    OutputLayerActivation: 'softmax'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [47×7 table]


  Properties, Methods
```

Access the weights and biases for the fully connected layers of the trained classifier by using the `LayerWeights` and `LayerBiases` properties of `Mdl`. The first two elements of each property correspond to the values for the first two fully connected layers, and the third element corresponds to the values for the final fully connected layer with a softmax activation function for classification. For example, display the weights and biases for the second fully connected layer.

```
Mdl.LayerWeights{2}
```

```
ans = 20×35

    0.0481    0.2501   -0.1535   -0.0934    0.0760   -0.0579   -0.2465    1.0411    0.3712   -1.2
   -0.9489   -1.8343    0.5510   -0.5751   -0.8726    0.8815    0.0203   -1.6379    2.0315    1.7
```

```
   -0.1910    0.0246   -0.3511    0.0097    0.3160   -0.0693    0.2270   -0.0783   -0.1626   -0.3
   -0.0415   -0.0059   -0.0753   -0.1477   -0.1621   -0.1762    0.2164    0.1710   -0.0610   -0.1
    1.1848    1.6142   -0.1352    0.5774    0.5491    0.0103    0.0209    0.7219   -0.8643   -0.5
    0.2486   -0.2920   -0.0004    0.2806    0.2987   -0.2709    0.1473   -0.2580   -0.0499   -0.0
   -0.0516    0.0640    0.1824   -0.0675   -0.2065   -0.0052   -0.1682   -0.1520    0.0060    0.0
   -0.6192   -0.7804   -0.0506   -0.4205   -0.2584   -0.2020   -0.0008    0.0534    1.0185   -0.0
    0.5049   -0.1362   -0.2218    0.1637   -0.1282   -0.1008    0.1445    0.4527   -0.4887    0.0
    1.1109   -0.0466    0.4044    0.6366    0.1863    0.5660    0.2839    0.8793   -0.5497    0.0
      ⋮
```

```
Mdl.LayerBiases{2}
```

```
ans = 20×1

    0.6147
    0.1891
   -0.2767
   -0.2977
    1.3655
    0.0347
    0.1509
   -0.4839
   -0.3960
    0.9248
      ⋮
```

The final fully connected layer has two outputs, one for each class in the response variable. The number of layer outputs corresponds to the first dimension of the layer weights and layer biases.

```
size(Mdl.LayerWeights{end})
```

```
ans = 1×2

     2    20
```

```
size(Mdl.LayerBiases{end})
```

```
ans = 1×2

     2     1
```

To estimate the performance of the trained classifier, compute the test set classification error for `Mdl`.

```
testError = loss(Mdl,XTest,YTest, ...
    "LossFun","classiferror")
```

```
testError = 0.0774
```

```
accuracy = 1 - testError
```

```
accuracy = 0.9226
```

`Mdl` accurately classifies approximately 92% of the observations in the test set.

Stop Neural Network Training Early Using Validation Data

At each iteration of the training process, compute the validation loss of the neural network. Stop the training process early if the validation loss reaches a reasonable minimum.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Separate the data into a training set `tblTrain` and a validation set `tblValidation` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the validation data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(tbl.Smoker,"Holdout",0.30);
trainingIndices = training(c);
validationIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblValidation = tbl(validationIndices,:);
```

Train a neural network classifier by using the training set. Specify the `Smoker` column of `tblTrain` as the response variable. Evaluate the model at each iteration by using the validation set. Specify to display the training information at each iteration by using the `Verbose` name-value argument. By default, the training process ends early if the validation cross-entropy loss is greater than or equal to the minimum validation cross-entropy loss computed so far, six times in a row. To change the number of times the validation loss is allowed to be greater than or equal to the minimum, specify the `ValidationPatience` name-value argument.

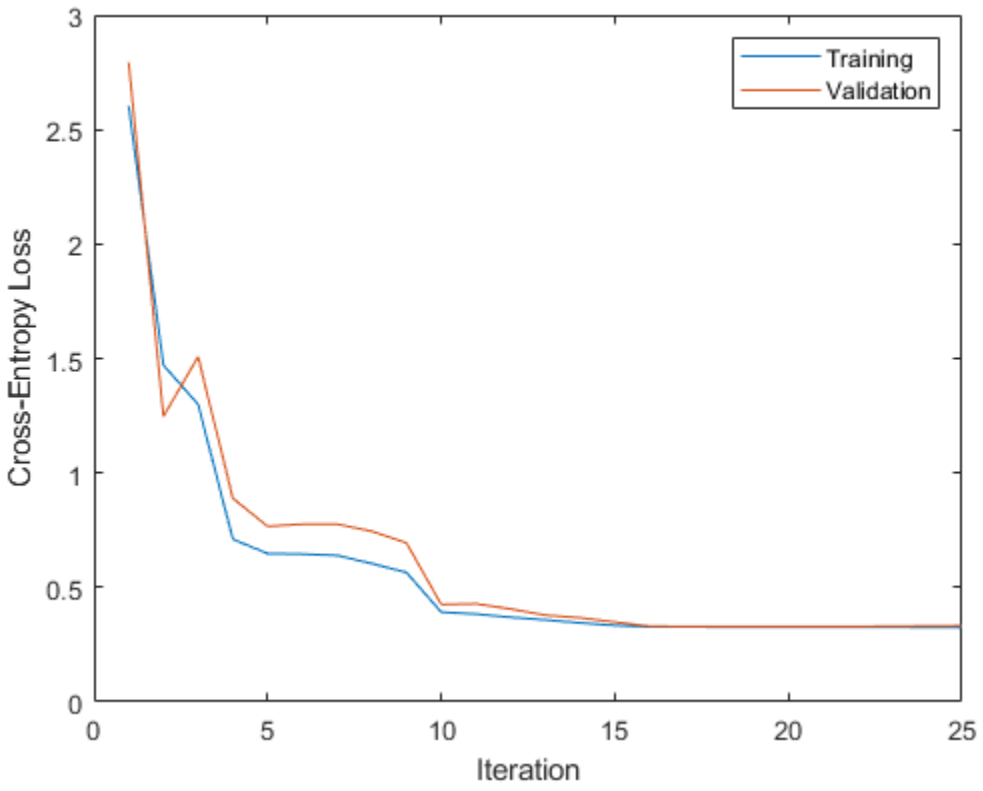
```
Mdl = fitcnet(tblTrain,"Smoker", ...
    "ValidationData",tblValidation, ...
    "Verbose",1);
```

```
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|           1|    2.602935|   26.866935|    0.262009|    0.001800|    2.793048|           0|
|           2|    1.470816|   42.594723|    0.058323|    0.001460|    1.247046|           0|
|           3|    1.299292|   25.854432|    0.034910|    0.000456|    1.507857|           1|
|           4|    0.710465|   11.629107|    0.013616|    0.000617|    0.889157|           0|
|           5|    0.647783|    2.561740|    0.005753|    0.000957|    0.766728|           0|
|           6|    0.645541|    0.681579|    0.001000|    0.000706|    0.776072|           1|
|           7|    0.639611|    1.544692|    0.007013|    0.005517|    0.776320|           2|
|           8|    0.604189|    5.045676|    0.064190|    0.000534|    0.744919|           0|
|           9|    0.565364|    5.851552|    0.068845|    0.000504|    0.694226|           0|
|          10|    0.391994|    8.377717|    0.560480|    0.000370|    0.425466|           0|
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|          11|    0.383843|    0.630246|    0.110270|    0.000749|    0.428487|           1|
|          12|    0.369289|    2.404750|    0.084395|    0.000531|    0.405728|           0|
|          13|    0.357839|    6.220679|    0.199197|    0.000353|    0.378480|           0|
|          14|    0.344974|    2.752717|    0.029013|    0.000330|    0.367279|           0|
```

```
|           15|    0.333747|    0.711398|    0.074513|    0.000328|    0.348499|           0|
|           16|    0.327763|    0.804818|    0.122178|    0.000348|    0.330237|           0|
|           17|    0.327702|    0.778169|    0.009810|    0.000365|    0.329095|           0|
|           18|    0.327277|    0.020615|    0.004377|    0.000380|    0.329141|           1|
|           19|    0.327273|    0.010018|    0.003313|    0.000432|    0.328773|           0|
|           20|    0.327268|    0.019497|    0.000805|    0.000776|    0.328831|           1|
|=============|============|============|============|============|============|============|
| Iteration   | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|             |            |            |            | Time (sec) | Loss       | Checks     |
|=============|============|============|============|============|============|============|
|           21|    0.327228|    0.113983|    0.005397|    0.000509|    0.329085|           2|
|           22|    0.327138|    0.240166|    0.012159|    0.000333|    0.329406|           3|
|           23|    0.326865|    0.428912|    0.036841|    0.000381|    0.329952|           4|
|           24|    0.325797|    0.255227|    0.139585|    0.000339|    0.331246|           5|
|           25|    0.325181|    0.758050|    0.135868|    0.000890|    0.332035|           6|
|=============|============|============|============|============|============|============|
```

Create a plot that compares the training cross-entropy loss and the validation cross-entropy loss at each iteration. By default, `fitcnet` stores the loss information inside the `TrainingHistory` property of the object `Mdl`. You can access this information by using dot notation.

```
iteration = Mdl.TrainingHistory.Iteration;
trainLosses = Mdl.TrainingHistory.TrainingLoss;
valLosses = Mdl.TrainingHistory.ValidationLoss;

plot(iteration,trainLosses,iteration,valLosses)
legend(["Training","Validation"])
xlabel("Iteration")
ylabel("Cross-Entropy Loss")
```



Check the iteration that corresponds to the minimum validation loss. The final returned model `Mdl` is the model trained at this iteration.

```
[~,minIdx] = min(valLosses);
iteration(minIdx)
```

```
ans = 19
```

Find Good Regularization Strength for Neural Network Using Cross-Validation

Assess the cross-validation loss of neural network models with different regularization strengths, and choose the regularization strength corresponding to the best performing model.

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable("CreditRating_Historical.dat");
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____     _____     _______    ________    _____    ________    ______
```

```
    62394     0.013     0.104     0.036     0.447     0.142         3       {'BB' }
    48608     0.232     0.335     0.062     1.969     0.281         8       {'A'  }
    42444     0.311     0.367     0.074     1.935     0.366         1       {'A'  }
    48631     0.194     0.263     0.062     1.017     0.228         4       {'BBB'}
    43768     0.121     0.413     0.057     3.647     0.466        12       {'AAA'}
    39255    -0.117    -0.799      0.01     0.179     0.082         4       {'CCC'}
    62236     0.087     0.158     0.049     0.816     0.324         2       {'BBB'}
    39354     0.005     0.181     0.034     2.597     0.388         7       {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,"ID");
creditrating.Industry = categorical(creditrating.Industry);
```

Convert the `Rating` response variable to an ordinal `categorical` variable.

```
creditrating.Rating = categorical(creditrating.Rating, ...
    ["AAA","AA","A","BBB","BB","B","CCC"],"Ordinal",true);
```

Create a `cvpartition` object for stratified 5-fold cross-validation. `cvp` partitions the data into five folds, where each fold has roughly the same proportions of different credit ratings. Set the random seed to the default value for reproducibility of the partition.

```
rng("default")
cvp = cvpartition(creditrating.Rating,"KFold",5);
```

Compute the cross-validation classification error for neural network classifiers with different regularization strengths. Try regularization strengths on the order of 1/*n*, where *n* is the number of observations. Specify to standardize the data before training the neural network models.

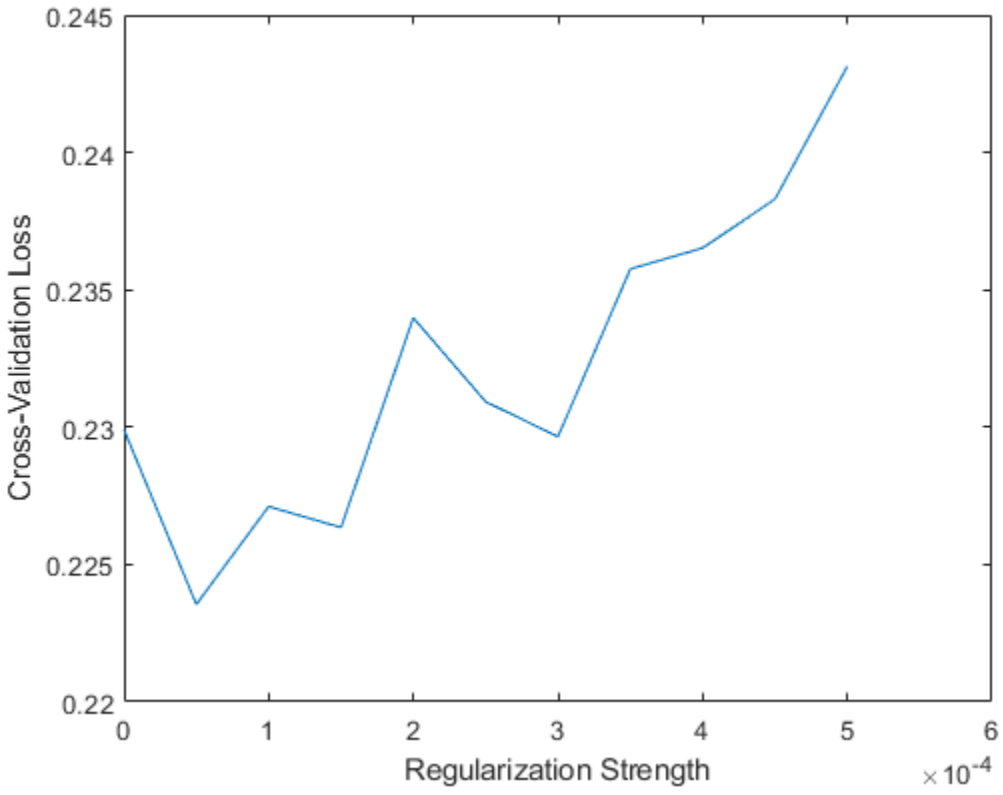
```
1/size(creditrating,1)
```

```
ans = 2.5432e-04
```

```
lambda = (0:0.5:5)*1e-4;
cvloss = zeros(length(lambda),1);

for i = 1:length(lambda)
    cvMdl = fitcnet(creditrating,"Rating","Lambda",lambda(i), ...
        "CVPartition",cvp,"Standardize",true);
    cvloss(i) = kfoldLoss(cvMdl,"LossFun","classiferror");
end
```

Plot the results. Find the regularization strength corresponding to the lowest cross-validation classification error.

```
plot(lambda,cvloss)
xlabel("Regularization Strength")
ylabel("Cross-Validation Loss")
```




```
[~,idx] = min(cvloss);
bestLambda = lambda(idx)
```

```
bestLambda = 5.0000e-05
```

Train a neural network classifier using the `bestLambda` regularization strength.

```
Mdl = fitcnet(creditrating,"Rating","Lambda",bestLambda, ...
    "Standardize",true)
```

```
Mdl =
  ClassificationNeuralNetwork
           PredictorNames: {'WC_TA'  'RE_TA'  'EBIT_TA'  'MVE_BVTD'  'S_TA'  'Industry'}
             ResponseName: 'Rating'
    CategoricalPredictors: 6
               ClassNames: [AAA    AA    A    BBB    BB    B    CCC]
          ScoreTransform: 'none'
         NumObservations: 3932
              LayerSizes: 10
             Activations: 'relu'
    OutputLayerActivation: 'softmax'
                  Solver: 'LBFGS'
        ConvergenceInfo: [1×1 struct]
        TrainingHistory: [1000×7 table]


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Y — Class labels

numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class labels used to train the model, specified as a numeric, categorical, or logical vector; a character or string array; or a cell array of character vectors.

- If `Y` is a character array, then each element of the class labels must correspond to one row of the array.
- The length of `Y` must be equal to the number of rows in `Tbl` or `X`.
- A good practice is to specify the class order by using the `ClassNames` name-value argument.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

X — Predictor data

numeric matrix

Predictor data used to train the model, specified as a numeric matrix.

By default, the software treats each row of `X` as one observation, and each column as one predictor.

The length of `Y` and the number of observations in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value argument.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response variable (for example, `Y` or `ValidationData{2}`)
- At least one missing value in a predictor observation (for example, row in `X` or `ValidationData{1}`)
- `NaN` value or `0` weight (for example, value in `Weights` or `ValidationData{3}`)

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitcnet(X,Y,'LayerSizes',[10 10],'Activations',["relu","tanh"])` specifies to create a neural network with two fully connected layers, each with 10 outputs. The first layer uses

a rectified linear unit (ReLU) activation function, and the second uses a hyperbolic tangent activation function.

Neural Network Options

`LayerSizes` — Sizes of fully connected layers

`10` (default) | positive integer vector

Sizes of the fully connected layers in the neural network model, specified as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer that uses a softmax activation function. For more information, see “Neural Network Structure” on page 33-1823.

Example: `'LayerSizes',[100 25 10]`

`Activations` — Activation functions for fully connected layers

`'relu'` (default) | `'tanh'` | `'sigmoid'` | `'none'` | string array | cell array of character vectors

Activation functions for the fully connected layers of the neural network model, specified as a character vector, string scalar, string array, or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If you specify one activation function only, then `Activations` is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer. The activation function for the final fully connected layer is always softmax (see “Neural Network Structure” on page 33-1823).
- If you specify an array of activation functions, then the *i*th element of `Activations` is the activation function for the *i*th layer of the neural network model.

Example: `'Activations','sigmoid'`

`LayerWeightsInitializer` — Function to initialize fully connected layer weights
`'glorot'` (default) | `'he'`

Function to initialize the fully connected layer weights, specified as `'glorot'` or `'he'`.

Value	Description
`'glorot'`	Initialize the weights with the Glorot initializer [1] (also known as the Xavier initializer). For each layer, the Glorot initalizer independently samples from a uniform distribution with zero mean and variable `2/(I+O)`, where `I` is the input size and `O` is the output size for the layer.
`'he'`	Initialize the weights with the He initializer [2]. For each layer, the He initializer samples from a normal distribution with zero mean and variance `2/I`, where `I` is the input size for the layer.

Example: `'LayerWeightsFunction','he'`

`LayerBiasesInitializer` — Type of initial fully connected layer biases
`'zeros'` (default) | `'ones'`

Type of initial fully connected layer biases, specified as `'zeros'` or `'ones'`.

- If you specify the value `'zeros'`, then each fully connected layer has an initial bias of 0.
- If you specify the value `'ones'`, then each fully connected layer has an initial bias of 1.

Example: `'LayerBiasesInitializer','ones'`

Data Types: `char` | `string`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

`Lambda` — Regularization term strength
`0` (default) | nonnegative scalar

Regularization term strength, specified as a nonnegative scalar. The software composes the objective function for minimization from the cross-entropy loss function and the ridge (L2) penalty term.

Example: `'Lambda',1e-4`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`false` or `0` (default) | `true` or `1`

Flag to standardize the predictor data, specified as a numeric or logical `0` (`false`) or `1` (`true`). If you set `Standardize` to `true`, then the software centers and scales each numeric predictor variable by the corresponding column mean and standard deviation. The software does not standardize the categorical predictors.

Example: `'Standardize',true`

Data Types: `single` | `double` | `logical`

Convergence Control Options

`Verbose` — Verbosity level

`0` (default) | `1`

Verbosity level, specified as `0` or `1`. The `'Verbose'` name-value argument controls the amount of diagnostic information that `fitcnet` displays at the command line.

Value	Description
`0`	`fitcnet` does not display diagnostic information.
`1`	`fitcnet` periodically displays diagnostic information.

By default, `StoreHistory` is set to `true` and `fitcnet` stores the diagnostic information inside of `Mdl`. Use `Mdl.TrainingHistory` to access the diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

`VerboseFrequency` — Frequency of verbose printing

`1` (default) | positive integer scalar

Frequency of verbose printing, which is the number of iterations between printing to the command window, specified as a positive integer scalar. A value of 1 indicates to print diagnostic information at every iteration.

Note To use this name-value argument, set `Verbose` to `1`.

Example: `'VerboseFrequency',5`

Data Types: `single` | `double`

`StoreHistory` — Flag to store training history

`true` or `1` (default) | `false` or `0`

Flag to store the training history, specified as a numeric or logical `0` (`false`) or `1` (`true`). If `StoreHistory` is set to `true`, then the software stores diagnostic information inside of `Mdl`, which you can access by using `Mdl.TrainingHistory`.

Example: `'StoreHistory',false`

Data Types: `single` | `double` | `logical`

`IterationLimit` — Maximum number of training iterations
`1e3` (default) | positive integer scalar

Maximum number of training iterations, specified as a positive integer scalar.

The software returns a trained model regardless of whether the training routine successfully converges. `Mdl.ConvergenceInfo` contains convergence information.

Example: `'IterationLimit',1e8`

Data Types: `single` | `double`

`GradientTolerance` — Relative gradient tolerance
`1e-6` (default) | nonnegative scalar

Relative gradient tolerance, specified as a nonnegative scalar.

Let $\mathscr{L}_t$ be the loss function at training iteration t, $\nabla\mathscr{L}_t$ be the gradient of the loss function with respect to the weights and biases at iteration t, and $\nabla\mathscr{L}_0$ be the gradient of the loss function at an initial point. If $\max|\nabla\mathscr{L}_t| \le a \cdot \text{GradientTolerance}$, where $a = \max(1, \min|\mathscr{L}_t|, \max|\nabla\mathscr{L}_0|)$, then the training process terminates.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`LossTolerance` — Loss tolerance
`1e-6` (default) | nonnegative scalar

Loss tolerance, specified as a nonnegative scalar.

If the function loss at some iteration is smaller than `LossTolerance`, then the training process terminates.

Example: `'LossTolerance',1e-8`

Data Types: `single` | `double`

`StepTolerance` — Step size tolerance
`1e-6` (default) | nonnegative scalar

Step size tolerance, specified as a nonnegative scalar.

If the step size at some iteration is smaller than `StepTolerance`, then the training process terminates.

Example: `'StepTolerance',1e-4`

Data Types: `single` | `double`

`ValidationData` — Validation data for training convergence detection
cell array | table

Validation data for training convergence detection, specified as a cell array or table.

During the training process, the software periodically estimates the validation loss by using `ValidationData`. If the validation loss increases more than `ValidationPatience` times in a row, then the software terminates the training.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of class labels, then it must have the same number of elements as the number of observations in `ValidationData{1}`. The set of all distinct labels of `ValidationData{2}` must be a subset of all distinct labels of `Y`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, set `Verbose` to `1`.

`ValidationFrequency` — Number of iterations between validation evaluations
1 (default) | positive integer scalar

Number of iterations between validation evaluations, specified as a positive integer scalar. A value of 1 indicates to evaluate validation metrics at every iteration.

Note To use this name-value argument, you must specify `ValidationData`.

Example: `'ValidationFrequency',5`

Data Types: `single` | `double`

`ValidationPatience` — Stopping condition for validation evaluations
6 (default) | nonnegative integer scalar

Stopping condition for validation evaluations, specified as a nonnegative integer scalar. The training process stops if the validation loss is greater than or equal to the minimum validation loss computed so far, `ValidationPatience` times in a row. You can check the `Mdl.TrainingHistory` table to see the running total of times that the validation loss is greater than or equal to the minimum (`Validation Checks`).

Example: `'ValidationPatience',10`

Data Types: `single` | `double`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table. The descriptions assume that the predictor data has observations in rows and predictors in columns.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcnet` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitcnet` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitcnet` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitcnet` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitcnet` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitcnet` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the predictor order in `X`. Assuming that `X` has the default orientation, with observations in rows and predictors in columns, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitcnet` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Flag to train cross-validated classifier

`'off'` (default) | `'on'`

Flag to train a cross-validated classifier, specified as `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated classifier with 10 folds.

You can override this cross-validation setting using the `CVPartition`, `Holdout`, `KFold`, or `Leaveout` name-value argument. You can use only one cross-validation name-value argument at a time to create a cross-validated model.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

Data Types: `char` | `string`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Output Arguments

`Mdl` — Trained neural network classifier
`ClassificationNeuralNetwork` object | `ClassificationPartitionedModel` object

Trained neural network classifier, returned as a `ClassificationNeuralNetwork` or `ClassificationPartitionedModel` object.

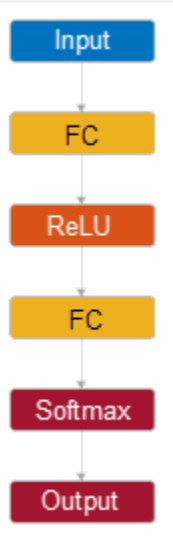
If you set any of the name-value arguments `CrossVal`, `CVPartition`, `Holdout`, `KFold`, or `Leaveout`, then `Mdl` is a `ClassificationPartitionedModel` object. Otherwise, `Mdl` is a `ClassificationNeuralNetwork` model.

To reference properties of `Mdl`, use dot notation.

More About

Neural Network Structure

The default neural network classifier has the following layer structure.

Structure	Description
	Input — This layer corresponds to the predictor data in `Tbl` or `X`.
	First fully connected layer — This layer has 10 outputs by default. • You can widen the layer or add more fully connected layers to the network by specifying the `LayerSizes` name-value argument. • You can find the weights and biases for this layer in the `Mdl.LayerWeights{1}` and `Mdl.LayerBiases{1}` properties of `Mdl`, respectively.
	ReLU activation function — `fitcnet` applies this activation function to the first fully connected layer. • You can change the activation function by specifying the `Activations` name-value argument.
	Final fully connected layer — This layer has K outputs, where K is the number of classes in the response variable. • You can find the weights and biases for this layer in the `Mdl.LayerWeights{end}` and `Mdl.LayerBiases{end}` properties of `Mdl`, respectively.
	Softmax function (for both binary and multiclass classification) — `fitcnet` applies this activation function to the final fully connected layer. The function takes each input x_i and returns the following, where K is the number of classes in the response variable: $$f(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{K} \exp(x_j)}.$$ The results correspond to the predicted classification scores (or posterior probabilities).
	Output — This layer corresponds to the predicted class labels.

For an example that shows how a neural network classifier with this layer structure returns predictions, see “Predict Using Layer Structure of Neural Network Classifier” on page 33-4850.

Tips

- Always try to standardize the numeric predictors (see `Standardize`). Standardization makes predictors insensitive to the scales on which they are measured.

Algorithms

Training Solver

`fitcnet` uses a limited-memory Broyden-Flecter-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) [3] as its loss function minimization technique, where the software minimizes the cross-entropy loss.

References

[1] Glorot, Xavier, and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks.” In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. 2010.

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.” In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034. 2015.

[3] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

See Also

`ClassificationNeuralNetwork` | `ClassificationPartitionedModel` | `CompactClassificationNeuralNetwork` | `edge` | `loss` | `margin` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

fitcox

Create Cox proportional hazards model

Syntax

```
coxMdl = fitcox(X,T)
coxMdl = fitcox(X,T,Name,Value)
```

Description

The `fitcox` function creates a Cox proportional hazards model for lifetime data. The basic Cox model includes a hazard function $h_0(t)$ and model coefficients b such that, for predictor `X`, the hazard rate at time t is

$$h(X_i,t) = h_0(t)\exp\left[\sum_{j=1}^{p} x_{ij}b_j\right],$$

where the b coefficients do not depend on time. `fitcox` infers both the model coefficients b and the hazard rate $h_0(t)$, and stores them as properties in the resulting `CoxModel` object.

The full Cox model includes extensions to the basic model, such as hazards with respect to different baselines or the inclusion of stratification variables. See “Extension of Cox Proportional Hazards Model” on page 14-27.

`coxMdl = fitcox(X,T)` returns a Cox proportional hazards model object `coxMdl` using the predictor values `X` and event times `T`.

`coxMdl = fitcox(X,T,Name,Value)` modifies the fit using one or more `Name,Value` arguments. For example, when the data includes censoring (values that are not observed), the `Censoring` argument specifies the censored data.

Examples

Estimate Cox Proportional Hazard Regression

Weibull random variables with the same shape parameter have proportional hazard rates; see “Weibull Distribution” on page B-170. The hazard rate with scale parameter a and shape parameter b at time t is

$$\frac{b}{a^b}t^{b-1}.$$

Generate pseudorandom samples from the Weibull distribution with scale parameters 1, 5, and 1/3, and with the same shape parameter `B`.

```
rng default % For reproducibility
B = 2;
A = ones(100,1);
data1 = wblrnd(A,B);
```

```
A2 = 5*A;
data2 = wblrnd(A2,B);
A3 = A/3;
data3 = wblrnd(A3,B);
```

Create a table of data. The predictors are the three variable types, 1, 2, or 3.

```
predictors = categorical([A;2*A;3*A]);
data = table(predictors,[data1;data2;data3],'VariableNames',["Predictors" "Times"]);
```

Fit a Cox regression to the data.

```
mdl = fitcox(data,"Times")
```

```
mdl =
Cox Proportional Hazards regression model:

                     Beta        SE        zStat       pValue
                   _______    _______    _______    __________

    Predictors_2    -3.5834    0.33187    -10.798    3.5299e-27
    Predictors_3     2.1668    0.20802     10.416    2.0899e-25
```

```
rates = exp(mdl.Coefficients.Beta)
```

```
rates = 2×1

    0.0278
    8.7301
```

Fit Cox Proportional Hazards Model to Lifetime Data

Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Load the `lightbulb` data set.

```
load lightbulb
```

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

```
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3))
```

```
coxMdl =
Cox Proportional Hazards regression model:

           Beta       SE      zStat       pValue
         _______    ______    ______    __________
```

```
    X1    4.7262    1.0372    4.5568    5.1936e-06
```

Find the hazard rate of incandescent bulbs compared to fluorescent bulbs by evaluating exp(Beta).

```
hr = exp(coxMdl.Coefficients.Beta)
```

```
hr = 112.8646
```

The estimate of the hazard ratio is e^{Beta} = 112.8646, which means that the estimated hazard for the incandescent bulbs is 112.86 times the hazard for the fluorescent bulbs. The small value of `coxMdl.Coefficients.pValue` indicates there is a negligible chance that the two types of light bulbs have identical hazard rates, which would mean `Beta` = 0.

Input Arguments

X — Predictor values

matrix | table

Predictor values, specified as a matrix or table.

- A matrix contains one column for each predictor and one row for each observation.
- A table contains one row for each observation. A table can also contain the time data as well as the predictors.

By default, if the predictor data is in a table, `fitcox` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix, `fitcox` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `CategoricalPredictors` name-value argument.

If `X`, `T`, the value of `'Frequency'`, or the value of `'Stratification'` contains `NaN` values, then `fitcox` removes rows with `NaN` values from all data when fitting a Cox model.

Data Types: `double` | `table` | `categorical`

T — Event times

real column vector | real matrix with two columns | name of column in table `X` | formula in Wilkinson notation for table `X`

Event times, specified as one of the following:

- Real column vector.
- Real matrix with two columns representing the start and stop times.
- Name of a column in the table `X`.
- Formula in Wilkinson notation for the table `X`. For example, to specify that the table columns `'x'` and `'y'` are in the model, use

  ```
  'T ~ x + y'
  ```

  See “Wilkinson Notation” on page 11-91.

For vector or matrix entries, the number of rows of `T` must be the same as the number of rows of `X`.

Use the two-column form of `T` to fit a model with time-varying coefficients. See “Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35.

Data Types: `single` | `double` | `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: To fit data with censored values `cens`, specify `'Censoring',cens`.

`Beta` — Coefficient initial values

`0.01/std(X)` (default) | numeric vector

Coefficient initial values, specified as a numeric vector of coefficient values. These values initiate the likelihood maximization iterations performed by `fitcox`.

Data Types: `double`

`Baseline` — `X` values at which to compute baseline hazard

`mean(X)`, the default for continuous predictors | `0`, the default for categorical predictors | real scalar | real row vector

`X` values at which to compute the baseline hazard, specified as a real scalar or row vector. If `Baseline` is a row vector, its length is the number of predictors, so there is one baseline for each predictor.

The default baseline for continuous predictors is `mean(X)`, so the default hazard rate at `X` for these predictors is `h(t)*exp((X – mean(X))*b)`. The default baseline for categorical predictors is `0`. Enter `0` to compute the baseline for all predictors relative to 0, so the hazard rate at `X` is `h(t)*exp(X*b)`. Changing the baseline changes the hazard ratio, but does not affect the coefficient estimates.

For the identified categorical predictors, `fitcox` creates dummy variables. `fitcox` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'Baseline',0`

Data Types: `double`

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data (`X`) that contains a categorical variable.
Logical vector	A `true` entry means that the corresponding column of predictor data (`X`) is a categorical variable.

Value	Description
Character matrix	Each row of the matrix is the name of a predictor variable in the table X. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable in the table X. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table, `fitcox` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix, `fitcox` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitcox` creates dummy variables. `fitcox` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`Censoring` — Indicator for censoring
array of 0s (default) | array of 0s and 1s | name of a column in table X

Indicator for censoring, specified as a Boolean vector with the same number of rows as X or the name of a column in the table X. Use 1 for observations that are right censored and 0 for observations that are fully observed. By default, all observations are fully observed. For an example, see “Cox Proportional Hazards Model for Censored Data” on page 14-31.

Example: `'Censoring',cens`

Data Types: `logical`

`Frequency` — Frequency or weights of observations
array of 1s (default) | vector of nonnegative scalar values

Frequency or weights of observations, specified as an array the same size as T containing nonnegative scalar values. The array can contain integer values corresponding to frequencies of observations or nonnegative values corresponding to observation weights.

The default is 1 per row of X and T.

If X, T, the value of `'Frequency'`, or the value of `'Stratification'` contains `NaN` values, then `fitcox` removes rows with `NaN` values from all data when fitting a Cox model.

Example: `'Frequency',w`

Data Types: `double`

`OptimizationOptions` — Algorithm control parameters
structure

Algorithm control parameters for the iterative algorithm `fitcox` uses to estimate the solution, specified as a structure. Create this structure using `statset`. For parameter names and default values, see the following table or enter `statset('fitcox')`.

In the table, "termination tolerance" means that if the internal iterations cause a change in the stated value less than the tolerance, the iterations stop.

Field in Structure	Description	Values
`Display`	Amount of information returned to the command line	• `'off'` — None (default) • `'final'` — Final output • `'iter'` — Output at each iteration
`MaxFunEvals`	Maximum number of function evaluations	Positive integer; default is `200`
`MaxIter`	Maximum number of iterations	Positive integer; default is `100`
`TolFun`	Termination tolerance on change in likelihood; see “Cox Proportional Hazards Model” on page 14-26	Positive scalar; default is `1e-8`
`TolX`	Termination tolerance for parameter (predictor estimate) change	Positive scalar; default is `1e-8`

Example: `'OptimizationOptions',statset('TolX',1e-6,'MaxIter',200)`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on how you supply the training data.

- If you supply `X` as a numeric array, then you can use `'PredictorNames'` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'X1','X2',...}`.
- If you supply `X` as a table, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitcox` uses only the predictor variables in `PredictorNames` and the time variable during training.
  - `PredictorNames` must be a subset of `X.Properties.VariableNames` and cannot include the name of the time variable `T`.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - Specify the predictors for training using either `'PredictorNames'` or a formula in Wilkinson notation, but not both.

Example: `'PredictorNames',{'Sex','Age','Weight','Smoker'}`

Data Types: `string` | `cell`

`Stratification` — Stratification variables
`[]` (default) | matrix of real values | name of column in table `X` | array of categorical variables

Stratification variables, specified as a matrix of real values, the name of a column in table `X`, or an array of categorical variables. The matrix must have the same number of rows as `T`, with each row corresponding to an observation.

The default `[]` is no stratification variable.

If `X`, `T`, the value of `'Frequency'`, or the value of `'Stratification'` contains `NaN` values, then `fitcox` removes rows with `NaN` values from all data when fitting a Cox model.

Example: `'Stratification',Gender`

Data Types: `single` | `double` | `char` | `string` | `categorical`

`TieBreakMethod` — Method to handle tied failure times
`'breslow'` (default) | `'efron'`

Method to handle tied failure times, specified as `'breslow'` (Breslow's method) or `'efron'` (Efron's method). See “Partial Likelihood Function for Tied Events” on page 14-28.

Example: `'TieBreakMethod','efron'`

Data Types: `char` | `string`

See Also

`CoxModel` | `coefci` | `coxphfit` | `hazardratio` | `linhyptest` | `plotSurvival` | `survival`

Topics
“Cox Proportional Hazards Model Object” on page 14-39
“What Is Survival Analysis?” on page 14-2
“Cox Proportional Hazards Model” on page 14-26
“Cox Proportional Hazards Model for Censored Data” on page 14-31
“Cox Proportional Hazards Model with Time-Dependent Covariates” on page 14-35
“Analyzing Survival or Reliability Data” on page 14-48

Introduced in R2021a

fitcsvm

Train support vector machine (SVM) classifier for one-class and binary classification

Syntax

```
Mdl = fitcsvm(Tbl,ResponseVarName)
Mdl = fitcsvm(Tbl,formula)
Mdl = fitcsvm(Tbl,Y)

Mdl = fitcsvm(X,Y)

Mdl = fitcsvm( ___ ,Name,Value)
```

Description

`fitcsvm` trains or cross-validates a support vector machine (SVM) model for one-class and two-class (binary) classification on a low-dimensional or moderate-dimensional predictor data set. `fitcsvm` supports mapping the predictor data using kernel functions, and supports sequential minimal optimization (SMO), iterative single data algorithm (ISDA), or $L1$ soft-margin minimization via quadratic programming for objective-function minimization.

To train a linear SVM model for binary classification on a high-dimensional data set, that is, a data set that includes many predictor variables, use `fitclinear` instead.

For multiclass learning with combined binary SVM models, use error-correcting output codes (ECOC). For more details, see `fitcecoc`.

To train an SVM regression model, see `fitrsvm` for low-dimensional and moderate-dimensional predictor data sets, or `fitrlinear` for high-dimensional data sets.

`Mdl = fitcsvm(Tbl,ResponseVarName)` returns a support vector machine (SVM) classifier on page 33-1863 `Mdl` trained using the sample data contained in the table `Tbl`. `ResponseVarName` is the name of the variable in `Tbl` that contains the class labels for one-class or two-class classification.

`Mdl = fitcsvm(Tbl,formula)` returns an SVM classifier trained using the sample data contained in the table `Tbl`. `formula` is an explanatory model of the response and a subset of the predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitcsvm(Tbl,Y)` returns an SVM classifier trained using the predictor variables in the table `Tbl` and the class labels in vector `Y`.

`Mdl = fitcsvm(X,Y)` returns an SVM classifier trained using the predictors in the matrix `X` and the class labels in vector `Y` for one-class or two-class classification.

`Mdl = fitcsvm( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input arguments in previous syntaxes. For example, you can specify the type of cross-validation, the cost for misclassification, and the type of score transformation function.

Examples

Train SVM Classifier

Load Fisher's iris data set. Remove the sepal lengths and widths and all observed setosa irises.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,3:4);
y = species(inds);
```

Train an SVM classifier using the processed data set.

```
SVMModel = fitcsvm(X,y)
```

```
SVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 100
                    Alpha: [24x1 double]
                     Bias: -14.4149
         KernelParameters: [1x1 struct]
           BoxConstraints: [100x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [100x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`SVMModel` is a trained `ClassificationSVM` classifier. Display the properties of `SVMModel`. For example, to determine the class order, use dot notation.

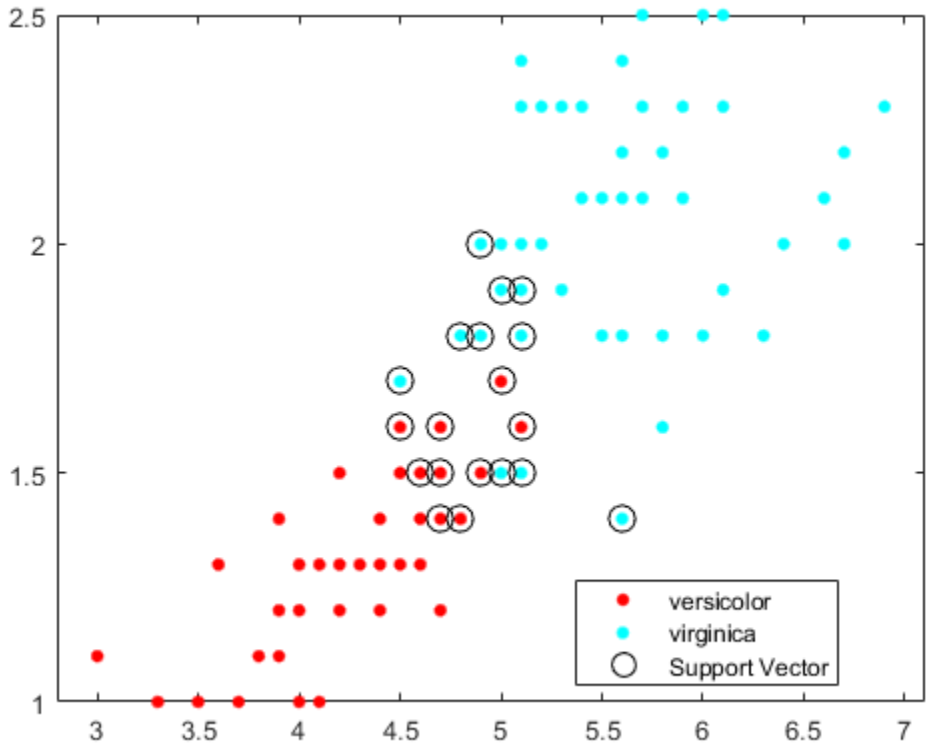
```
classOrder = SVMModel.ClassNames
```

```
classOrder = 2x1 cell
    {'versicolor'}
    {'virginica' }
```

The first class (`'versicolor'`) is the negative class, and the second (`'virginica'`) is the positive class. You can change the class order during training by using the `'ClassNames'` name-value pair argument.

Plot a scatter diagram of the data and circle the support vectors.

```
sv = SVMModel.SupportVectors;
figure
gscatter(X(:,1),X(:,2),y)
hold on
plot(sv(:,1),sv(:,2),'ko','MarkerSize',10)
legend('versicolor','virginica','Support Vector')
hold off
```



The support vectors are observations that occur on or beyond their estimated class boundaries.

You can adjust the boundaries (and, therefore, the number of support vectors) by setting a box constraint during training using the `'BoxConstraint'` name-value pair argument.

Train and Cross-Validate SVM Classifier

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier using the radial basis kernel. Let the software find a scale value for the kernel function. Standardize the predictors.

```
SVMModel = fitcsvm(X,Y,'Standardize',true,'KernelFunction','RBF',...
    'KernelScale','auto');
```

`SVMModel` is a trained `ClassificationSVM` classifier.

Cross-validate the SVM classifier. By default, the software uses 10-fold cross-validation.

```
CVSVMModel = crossval(SVMModel);
```

`CVSVMModel` is a `ClassificationPartitionedModel` cross-validated classifier.

Estimate the out-of-sample misclassification rate.

```
classLoss = kfoldLoss(CVSVMModel)
```

```
classLoss = 0.0484
```

The generalization rate is approximately 5%.

Detect Outliers Using SVM and One-Class Learning

Modify Fisher's iris data set by assigning all the irises to the same class. Detect outliers in the modified data set, and confirm the expected proportion of the observations that are outliers.

Load Fisher's iris data set. Remove the petal lengths and widths. Treat all irises as coming from the same class.

```
load fisheriris
X = meas(:,1:2);
y = ones(size(X,1),1);
```

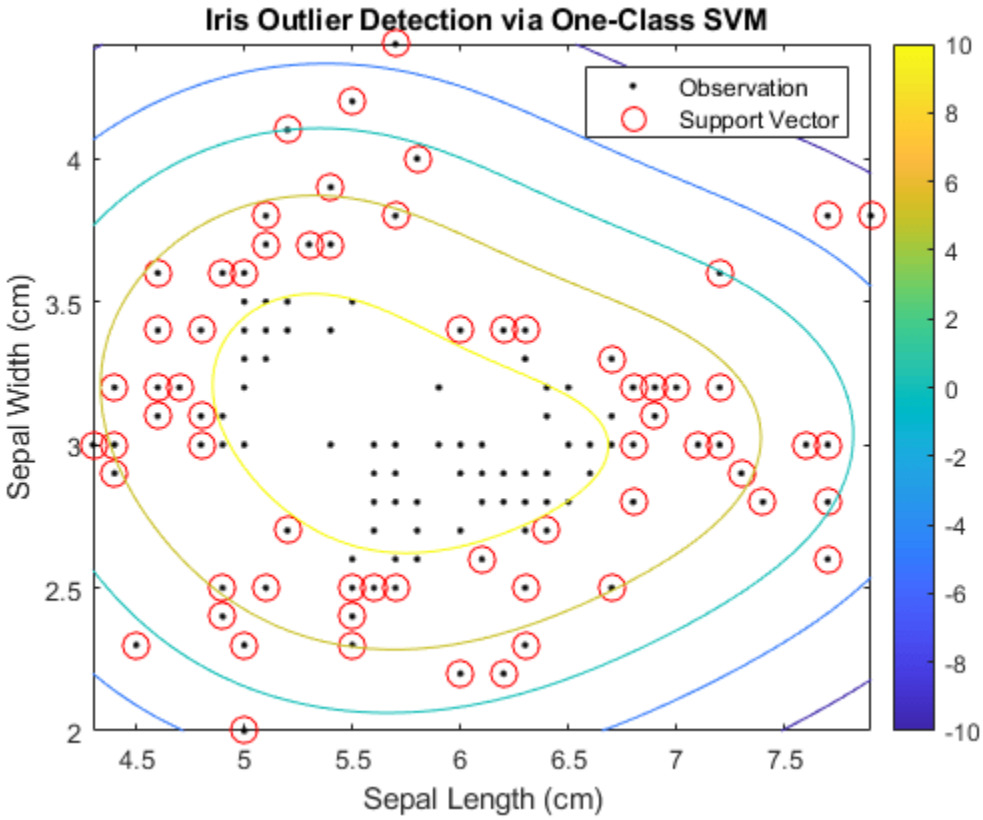
Train an SVM classifier using the modified data set. Assume that 5% of the observations are outliers. Standardize the predictors.

```
rng(1);
SVMModel = fitcsvm(X,y,'KernelScale','auto','Standardize',true,...
    'OutlierFraction',0.05);
```

`SVMModel` is a trained `ClassificationSVM` classifier. By default, the software uses the Gaussian kernel for one-class learning.

Plot the observations and the decision boundary. Flag the support vectors and potential outliers.

```
svInd = SVMModel.IsSupportVector;
h = 0.02; % Mesh grid step size
[X1,X2] = meshgrid(min(X(:,1)):h:max(X(:,1)),...
    min(X(:,2)):h:max(X(:,2)));
[~,score] = predict(SVMModel,[X1(:),X2(:)]);
scoreGrid = reshape(score,size(X1,1),size(X2,2));

figure
plot(X(:,1),X(:,2),'k.')
hold on
plot(X(svInd,1),X(svInd,2),'ro','MarkerSize',10)
contour(X1,X2,scoreGrid)
colorbar;
title('{\bf Iris Outlier Detection via One-Class SVM}')
xlabel('Sepal Length (cm)')
ylabel('Sepal Width (cm)')
legend('Observation','Support Vector')
hold off
```



The boundary separating the outliers from the rest of the data occurs where the contour value is `0`.

Verify that the fraction of observations with negative scores in the cross-validated data is close to 5%.

```
CVSVMModel = crossval(SVMModel);
[~,scorePred] = kfoldPredict(CVSVMModel);
outlierRate = mean(scorePred<0)
```

```
outlierRate = 0.0467
```

Find Multiple Class Boundaries Using Binary SVM

Create a scatter plot of the `fisheriris` data set. Treat coordinates of a grid within the plot as new observations from the distribution of the data set, and find class boundaries by assigning the coordinates to one of the three classes in the data set.

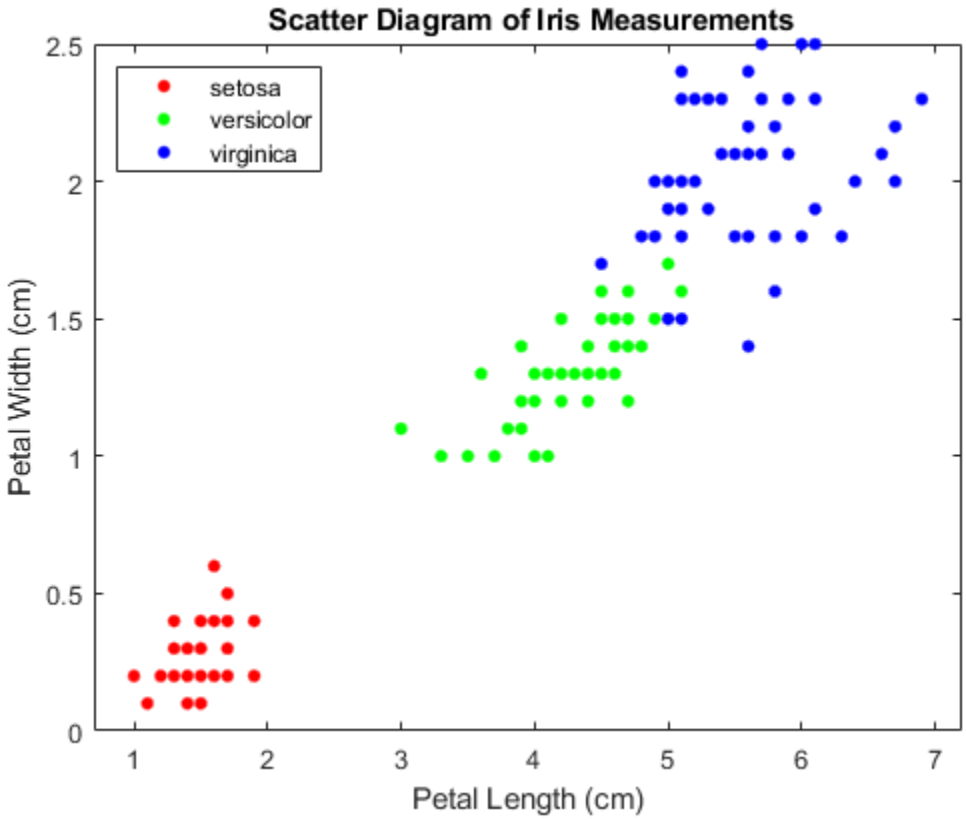
Load Fisher's iris data set. Use the petal lengths and widths as the predictors.

```
load fisheriris
X = meas(:,3:4);
Y = species;
```

Examine a scatter plot of the data.

```
figure
gscatter(X(:,1),X(:,2),Y);
```

```
h = gca;
lims = [h.XLim h.YLim]; % Extract the x and y axis limits
title('{\bf Scatter Diagram of Iris Measurements}');
xlabel('Petal Length (cm)');
ylabel('Petal Width (cm)');
legend('Location','Northwest');
```



The data contains three classes, one of which is linearly separable from the others.

For each class:

1. Create a logical vector (`indx`) indicating whether an observation is a member of the class.
2. Train an SVM classifier using the predictor data and `indx`.
3. Store the classifier in a cell of a cell array.

Define the class order.

```
SVMModels = cell(3,1);
classes = unique(Y);
rng(1); % For reproducibility

for j = 1:numel(classes)
    indx = strcmp(Y,classes(j)); % Create binary classes for each classifier
    SVMModels{j} = fitcsvm(X,indx,'ClassNames',[false true],'Standardize',true,...
        'KernelFunction','rbf','BoxConstraint',1);
end
```

`SVMModels` is a 3-by-1 cell array, with each cell containing a `ClassificationSVM` classifier. For each cell, the positive class is setosa, versicolor, and virginica, respectively.

Define a fine grid within the plot, and treat the coordinates as new observations from the distribution of the training data. Estimate the score of the new observations using each classifier.

```
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(X(:,1)):d:max(X(:,1)),...
    min(X(:,2)):d:max(X(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];
N = size(xGrid,1);
Scores = zeros(N,numel(classes));

for j = 1:numel(classes)
    [~,score] = predict(SVMModels{j},xGrid);
    Scores(:,j) = score(:,2); % Second column contains positive-class scores
end
```

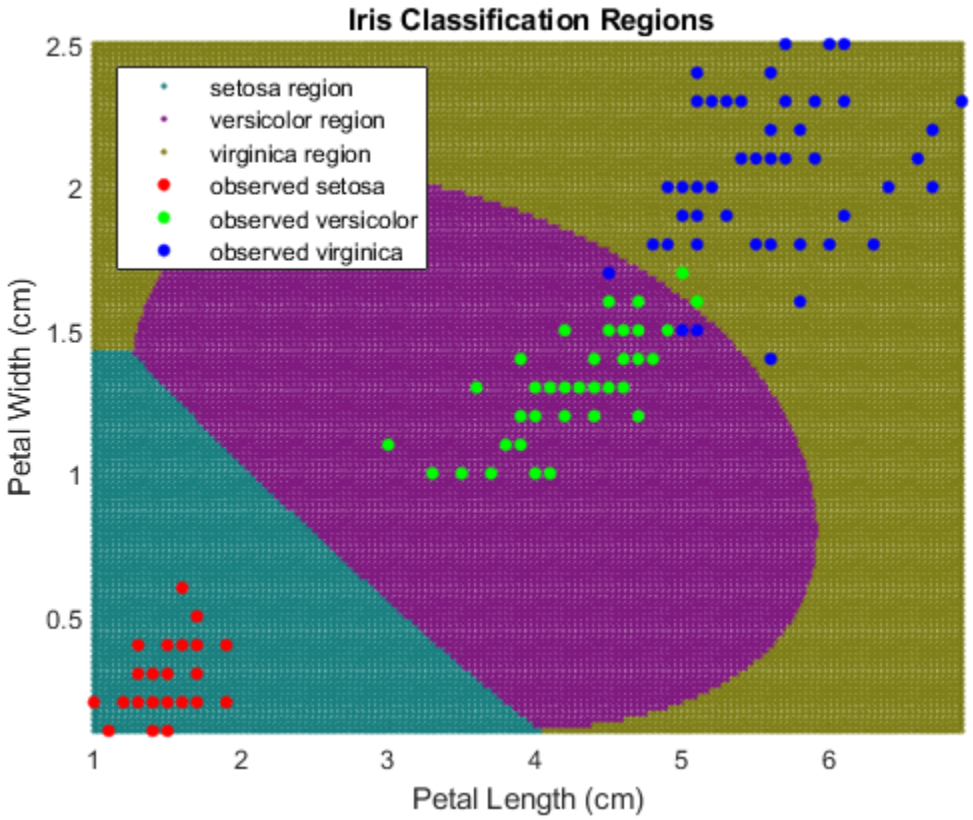
Each row of `Scores` contains three scores. The index of the element with the largest score is the index of the class to which the new class observation most likely belongs.

Associate each new observation with the classifier that gives it the maximum score.

```
[~,maxScore] = max(Scores,[],2);
```

Color in the regions of the plot based on the class to which the corresponding new observation belongs.

```
figure
h(1:3) = gscatter(xGrid(:,1),xGrid(:,2),maxScore,...
    [0.1 0.5 0.5; 0.5 0.1 0.5; 0.5 0.5 0.1]);
hold on
h(4:6) = gscatter(X(:,1),X(:,2),Y);
title('{\bf Iris Classification Regions}');
xlabel('Petal Length (cm)');
ylabel('Petal Width (cm)');
legend(h,{'setosa region','versicolor region','virginica region',...
    'observed setosa','observed versicolor','observed virginica'},...
    'Location','Northwest');
axis tight
hold off
```



Optimize SVM Classifier

Optimize hyperparameters automatically using `fitcsvm`.

Load the `ionosphere` data set.

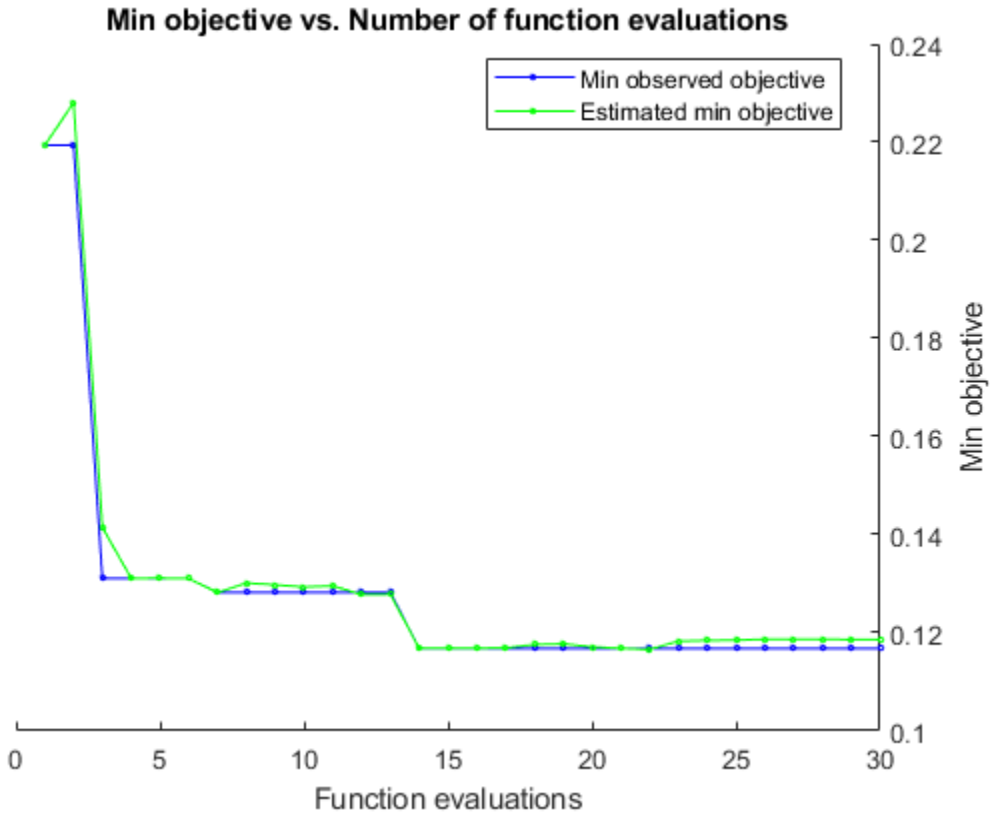
```
load ionosphere
```

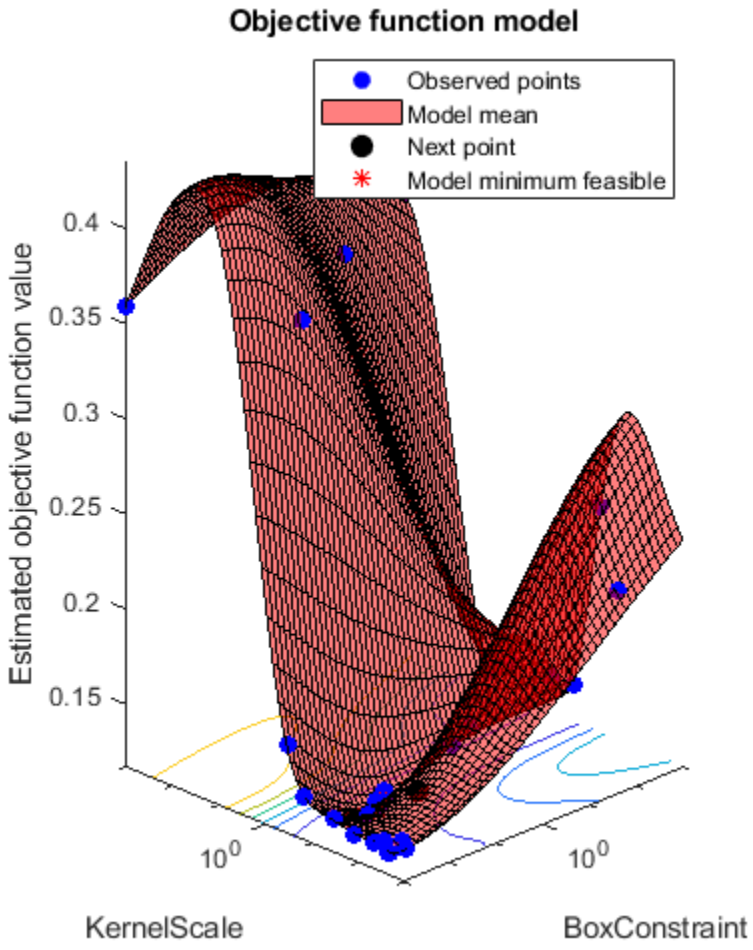
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization. For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
Mdl = fitcsvm(X,Y,'OptimizeHyperparameters','auto', ...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName', ...
    'expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.21937 |      22.967 |     0.21937 |     0.21937 |       64.836 |    0.001
|    2 | Accept |     0.35897 |     0.18996 |     0.21937 |     0.22807 |     0.036335 |        5
|    3 | Best   |     0.13105 |      9.0818 |     0.13105 |     0.14149 |    0.0022147 |    0.002
|    4 | Accept |     0.35897 |     0.21223 |     0.13105 |     0.13108 |       5.1259 |        9
```

```
|    5 | Accept |      0.1339 |      19.595 |     0.13105 |     0.13111 |    0.0011599 |    0.001
|    6 | Accept |     0.13105 |     0.93214 |     0.13105 |     0.13106 |    0.0010151 |    0.004
|    7 | Best   |     0.12821 |      11.314 |     0.12821 |     0.12819 |    0.0010563 |    0.002
|    8 | Accept |      0.1339 |      14.269 |     0.12821 |     0.13014 |    0.0010113 |    0.002
|    9 | Accept |     0.12821 |      6.9845 |     0.12821 |     0.12977 |    0.0010934 |    0.002
|   10 | Accept |     0.12821 |      3.8724 |     0.12821 |     0.12934 |    0.0010315 |    0.002
|   11 | Accept |     0.13675 |       16.06 |     0.12821 |     0.12954 |       994.04 |       0.2
|   12 | Accept |     0.23647 |      25.025 |     0.12821 |     0.12778 |       959.35 |      0.05
|   13 | Accept |     0.13105 |     0.35947 |     0.12821 |     0.12784 |    0.0010239 |      0.0
|   14 | Best   |     0.11681 |     0.21069 |     0.11681 |     0.11682 |      0.11417 |       0.2
|   15 | Accept |     0.13105 |      2.2964 |     0.11681 |     0.11683 |       11.334 |       0.4
|   16 | Accept |      0.1396 |     0.17818 |     0.11681 |     0.11684 |    0.0010029 |       0.1
|   17 | Accept |     0.16239 |     0.19468 |     0.11681 |     0.11689 |    0.0011338 |       0.3
|   18 | Accept |     0.13105 |      1.1004 |     0.11681 |     0.11758 |       2.2742 |       0.2
|   19 | Accept |     0.13675 |       18.78 |     0.11681 |     0.11772 |       970.79 |       0.6
|   20 | Accept |     0.11681 |     0.19186 |     0.11681 |     0.11711 |     0.047986 |       0.1
|======|========|=============|=============|=============|=============|==============|==========
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|======|========|=============|=============|=============|=============|==============|==========
|   21 | Accept |     0.12536 |     0.17522 |     0.11681 |     0.11681 |    0.0082154 |      0.05
|   22 | Accept |     0.11681 |      0.2207 |     0.11681 |     0.11659 |     0.063229 |       0.1
|   23 | Accept |     0.12251 |     0.21603 |     0.11681 |     0.11829 |     0.080255 |       0.2
|   24 | Accept |     0.12251 |      0.1714 |     0.11681 |     0.11842 |     0.031721 |       0.1
|   25 | Accept |      0.1339 |     0.16614 |     0.11681 |     0.11848 |    0.0010942 |      0.03
|   26 | Accept |     0.11966 |     0.19795 |     0.11681 |     0.11869 |      0.09066 |       0.2
|   27 | Accept |     0.35897 |     0.15896 |     0.11681 |     0.11869 |    0.0010252 |        99
|   28 | Accept |     0.35897 |     0.17082 |     0.11681 |     0.11869 |       990.01 |        99
|   29 | Accept |     0.12536 |     0.26753 |     0.11681 |     0.11862 |       995.77 |        13
|   30 | Accept |      0.1339 |     0.73655 |     0.11681 |     0.11862 |       991.43 |         5
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
Min objective
0
5
10
15
20
25
30
Function evaluations




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 203.4505 seconds
Total objective function evaluation time: 156.2959

Best observed feasible point:
    BoxConstraint    KernelScale
    _____________    ___________

       0.11417         0.26314

Observed objective function value = 0.11681
Estimated objective function value = 0.11926
Function evaluation time = 0.21069

Best estimated feasible point (according to models):
    BoxConstraint    KernelScale
    _____________    ___________
```

```
      0.080255         0.21171

Estimated objective function value = 0.11862
Estimated function evaluation time = 0.22213

Mdl =
  ClassificationSVM
                           ResponseName: 'Y'
                  CategoricalPredictors: []
                             ClassNames: {'b'  'g'}
                         ScoreTransform: 'none'
                        NumObservations: 351
      HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                                  Alpha: [98x1 double]
                                   Bias: -4.8344
                       KernelParameters: [1x1 struct]
                         BoxConstraints: [351x1 double]
                        ConvergenceInfo: [1x1 struct]
                        IsSupportVector: [351x1 logical]
                                 Solver: 'SMO'


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels to which the SVM model is trained, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

- `Y` must contain at most two distinct classes. For multiclass learning, see `fitcecoc`.
- If `Y` is a character array, then each element of the class labels must correspond to one row of the array.
- The length of `Y` and the number of rows in `Tbl` or `X` must be equal.
- It is a good practice to specify the class order by using the `ClassNames` name-value pair argument.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
matrix of numeric values

Predictor data to which the SVM classifier is trained, specified as a matrix of numeric values.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one predictor (also known as a feature).

The length of `Y` and the number of rows in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `'PredictorNames'` name-value pair argument.

Data Types: `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitcsvm(X,Y,'KFold',10,'Cost',[0 2;1 0],'ScoreTransform','sign')` performs 10-fold cross-validation, applies double the penalty to false positives compared to false negatives, and transforms scores using the sign function.

SVM Options

`BoxConstraint` — Box constraint
1 (default) | positive scalar

Box constraint on page 33-1862, specified as the comma-separated pair consisting of `'BoxConstraint'` and a positive scalar.

For one-class learning, the software always sets the box constraint to `1`.

For more details on the relationships and algorithmic behavior of `BoxConstraint`, `Cost`, `Prior`, `Standardize`, and `Weights`, see "Algorithms" on page 33-1865.

Example: `'BoxConstraint',100`

Data Types: `double` | `single`

`KernelFunction` — Kernel function
`'linear'` | `'gaussian'` | `'rbf'` | `'polynomial'` | function name

Kernel function used to compute the elements of the Gram matrix on page 33-1862, specified as the comma-separated pair consisting of `'KernelFunction'` and a kernel function name. Suppose $G(x_j,x_k)$ is element (j,k) of the Gram matrix, where x_j and x_k are p-dimensional vectors representing observations j and k in `X`. This table describes supported kernel function names and their functional forms.

Kernel Function Name	Description	Formula
`'gaussian'` or `'rbf'`	Gaussian or Radial Basis Function (RBF) kernel, default for one-class learning	$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right)$
`'linear'`	Linear kernel, default for two-class learning	$G(x_j, x_k) = x_j' x_k$
`'polynomial'`	Polynomial kernel. Use `'PolynomialOrder',q` to specify a polynomial kernel of order q.	$G(x_j, x_k) = (1 + x_j' x_k)^q$

You can set your own kernel function, for example, `kernel`, by setting `'KernelFunction','kernel'`. The value `kernel` must have this form.

```
function G = kernel(U,V)
```

where:

- `U` is an *m*-by-*p* matrix. Columns correspond to predictor variables, and rows correspond to observations.
- `V` is an *n*-by-*p* matrix. Columns correspond to predictor variables, and rows correspond to observations.
- `G` is an *m*-by-*n* Gram matrix on page 33-1862 of the rows of `U` and `V`.

`kernel.m` must be on the MATLAB path.

It is a good practice to avoid using generic names for kernel functions. For example, call a sigmoid kernel function `'mysigmoid'` rather than `'sigmoid'`.

Example: `'KernelFunction','gaussian'`

Data Types: `char` | `string`

`KernelScale` — Kernel scale parameter
1 (default) | `'auto'` | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software divides all elements of the predictor matrix `X` by the value of `KernelScale`. Then, the software applies the appropriate kernel norm to compute the Gram matrix.

- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed using `rng` before training.
- If you specify `KernelScale` and your own kernel function, for example, `'KernelFunction','kernel'`, then the software throws an error. You must apply scaling within `kernel`.

Example: `'KernelScale','auto'`

Data Types: `double` | `single` | `char` | `string`

`PolynomialOrder` — Polynomial kernel function order
3 (default) | positive integer

Polynomial kernel function order, specified as the comma-separated pair consisting of `'PolynomialOrder'` and a positive integer.

If you set `'PolynomialOrder'` and `KernelFunction` is not `'polynomial'`, then the software throws an error.

Example: `'PolynomialOrder',2`

Data Types: `double` | `single`

`KernelOffset` — Kernel offset parameter
nonnegative scalar

Kernel offset parameter, specified as the comma-separated pair consisting of `'KernelOffset'` and a nonnegative scalar.

The software adds `KernelOffset` to each element of the Gram matrix.

The defaults are:

- `0` if the solver is SMO (that is, you set `'Solver','SMO'`)
- `0.1` if the solver is ISDA (that is, you set `'Solver','ISDA'`)

Example: `'KernelOffset',0`

Data Types: `double` | `single`

`Standardize` — Flag to standardize predictor data
`false` (default) | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and `true` (`1`) or `false` (`0`).

If you set `'Standardize',true`:

- The software centers and scales each predictor variable (`X` or `Tbl`) by the corresponding weighted column mean and standard deviation. For details on weighted standardizing, see “Algorithms” on page 33-1865. MATLAB does not standardize the data contained in the dummy variable columns generated for categorical predictors.
- The software trains the classifier using the standardized predictors, but stores the unstandardized predictors as a matrix or table in the classifier property `X`.

Example: `'Standardize',true`

Data Types: `logical`

`Solver` — Optimization routine
`'ISDA'` | `'L1QP'` | `'SMO'`

Optimization routine, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description
`'ISDA'`	Iterative Single Data Algorithm (see [30])
`'L1QP'`	Uses `quadprog` to implement $L1$ soft-margin minimization by quadratic programming. This option requires an Optimization Toolbox license. For more details, see “Quadratic Programming Definition” (Optimization Toolbox).
`'SMO'`	Sequential Minimal Optimization (see [17])

The default value is `'ISDA'` if you set `'OutlierFraction'` to a positive value for two-class learning, and `'SMO'` otherwise.

Example: `'Solver','ISDA'`

`Alpha` — Initial estimates of alpha coefficients
numeric vector of nonnegative values

Initial estimates of alpha coefficients, specified as the comma-separated pair consisting of `'Alpha'` and a numeric vector of nonnegative values. The length of `Alpha` must be equal to the number of rows in `X`.

- Each element of `'Alpha'` corresponds to an observation in `X`.

- `'Alpha'` cannot contain any `NaN`s.
- If you specify `'Alpha'` and any one of the cross-validation name-value pair arguments (`'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`), then the software returns an error.

If `Y` contains any missing values, then remove all rows of `Y`, `X`, and `'Alpha'` that correspond to the missing values. That is, enter:

```
idx = ~isundefined(categorical(Y));
Y = Y(idx,:);
X = X(idx,:);
alpha = alpha(idx);
```

Then pass `Y`, `X`, and `alpha` as the response, predictors, and initial alpha estimates, respectively.

The default values are:

- `0.5*ones(size(X,1),1)` for one-class learning
- `zeros(size(X,1),1)` for two-class learning

Example: `'Alpha',0.1*ones(size(X,1),1)`

Data Types: `double` | `single`

`CacheSize` — Cache size
`1000` (default) | `'maximal'` | positive scalar

Cache size, specified as the comma-separated pair consisting of `'CacheSize'` and `'maximal'` or a positive scalar.

If `CacheSize` is `'maximal'`, then the software reserves enough memory to hold the entire *n*-by-*n* Gram matrix on page 33-1862.

If `CacheSize` is a positive scalar, then the software reserves `CacheSize` megabytes of memory for training the model.

Example: `'CacheSize','maximal'`

Data Types: `double` | `single` | `char` | `string`

`ClipAlphas` — Flag to clip alpha coefficients
`true` (default) | `false`

Flag to clip alpha coefficients, specified as the comma-separated pair consisting of `'ClipAlphas'` and either `true` or `false`.

Suppose that the alpha coefficient for observation j is α_j and the box constraint of observation j is C_j, $j = 1,...,n$, where n is the training sample size.

Value	Description
`true`	At each iteration, if α_j is near 0 or near C_j, then MATLAB sets α_j to 0 or to C_j, respectively.
`false`	MATLAB does not change the alpha coefficients during optimization.

MATLAB stores the final values of α in the `Alpha` property of the trained SVM model object.

`ClipAlphas` can affect SMO and ISDA convergence.

Example: `'ClipAlphas',false`

Data Types: `logical`

`Nu` — ν parameter for one-class learning
`0.5` (default) | positive scalar

ν parameter for "One-Class Learning" on page 33-1863, specified as the comma-separated pair consisting of `'Nu'` and a positive scalar. `Nu` must be greater than `0` and at most `1`.

Set `Nu` to control the tradeoff between ensuring that most training examples are in the positive class and minimizing the weights in the score function.

Example: `'Nu',0.25`

Data Types: `double` | `single`

`NumPrint` — Number of iterations between optimization diagnostic message output
`1000` (default) | nonnegative integer

Number of iterations between optimization diagnostic message output, specified as the comma-separated pair consisting of `'NumPrint'` and a nonnegative integer.

If you specify `'Verbose',1` and `'NumPrint',numprint`, then the software displays all optimization diagnostic messages from SMO and ISDA every `numprint` iterations in the Command Window.

Example: `'NumPrint',500`

Data Types: `double` | `single`

`OutlierFraction` — Expected proportion of outliers in training data
`0` (default) | numeric scalar in the interval [0,1)

Expected proportion of outliers in the training data, specified as the comma-separated pair consisting of `'OutlierFraction'` and a numeric scalar in the interval [0,1).

Suppose that you set `'OutlierFraction',outlierfraction`, where `outlierfraction` is a value greater than 0.

- For two-class learning, the software implements robust learning. In other words, the software attempts to remove 100*`outlierfraction`% of the observations when the optimization algorithm converges. The removed observations correspond to gradients that are large in magnitude.
- For one-class learning, the software finds an appropriate bias term such that `outlierfraction` of the observations in the training set have negative scores.

Example: `'OutlierFraction',0.01`

Data Types: `double` | `single`

`RemoveDuplicates` — Flag to replace duplicate observations with single observations
`false` (default) | `true`

Flag to replace duplicate observations with single observations in the training data, specified as the comma-separated pair consisting of `'RemoveDuplicates'` and `true` or `false`.

If `RemoveDuplicates` is `true`, then `fitcsvm` replaces duplicate observations in the training data with a single observation of the same value. The weight of the single observation is equal to the sum of the weights of the corresponding removed duplicates (see `Weights`).

Tip If your data set contains many duplicate observations, then specifying `'RemoveDuplicates',true` can decrease convergence time considerably.

Data Types: `logical`

`Verbose` — Verbosity level

`0` (default) | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. The value of `Verbose` controls the amount of optimization information that the software displays in the Command Window and saves the information as a structure to `Mdl.ConvergenceInfo.History`.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software does not display or save convergence information.
`1`	The software displays diagnostic messages and saves convergence criteria every `numprint` iterations, where `numprint` is the value of the name-value pair argument `'NumPrint'`.
`2`	The software displays diagnostic messages and saves convergence criteria at every iteration.

Example: `'Verbose',1`

Data Types: `double` | `single`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitcsvm` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.

Value	Description
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitcsvm` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitcsvm` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitcsvm` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitcsvm` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitcsvm` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost

square matrix | structure array

Misclassification cost specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure array.

- If you specify the square matrix `Cost` and the true class of an observation is `i`, then `Cost(i,j)` is the cost of classifying a point into class `j`. That is, rows correspond to the true classes and columns correspond to predicted classes. To specify the class order for the corresponding rows and columns of `Cost`, also specify the `ClassNames` name-value pair argument.
- If you specify the structure `S`, then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

For two-class learning, if you specify a cost matrix, then the software updates the prior probabilities by incorporating the penalties described in the cost matrix. Consequently, the cost matrix resets to the default. For more details on the relationships and algorithmic behavior of `BoxConstraint`, `Cost`, `Prior`, `Standardize`, and `Weights`, see “Algorithms” on page 33-1865.

The default values are:

- `Cost = 0` for one-class learning
- `Cost(i,j) = 1` if `i ~= j` and `Cost(i,j) = 0` if `i = j` for two-class learning

Example: `'Cost',[0,1;2,0]`

Data Types: `double` | `single` | `struct`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitcsvm` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: 'PredictorNames',
{'SepalLength','SepalWidth','PetalLength','PetalWidth'}

Data Types: string | cell

Prior — Prior probabilities
'empirical' (default) | 'uniform' | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of 'Prior' and a value in this table.

Value	Description
'empirical'	The class prior probabilities are the class relative frequencies in Y.
'uniform'	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element in the vector is a class prior probability. Order the elements according to Mdl.ClassNames or specify the order using the ClassNames name-value pair argument. The software normalizes the elements to sum to 1.
structure	A structure S with two fields: • S.ClassNames contains the class names as a variable of the same type as Y. • S.ClassProbs contains a vector of corresponding prior probabilities. The software normalizes the elements of the vector to sum to 1.

For two-class learning, if you specify a cost matrix, then the software updates the prior probabilities by incorporating the penalties described in the cost matrix. For more details on the relationships and algorithmic behavior of BoxConstraint, Cost, Prior, Standardize, and Weights, see “Algorithms” on page 33-1865.

Example: struct('ClassNames',
{{'setosa','versicolor','virginica'}},'ClassProbs',1:3)

Data Types: char | string | double | single | struct

ResponseName — Response variable name
'Y' (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply Y, then you can use 'ResponseName' to specify a name for the response variable.
- If you supply ResponseVarName or formula, then you cannot use 'ResponseName'.

Example: 'ResponseName','response'

Data Types: char | string

ScoreTransform — Score transformation
'none' (default) | 'doublelogit' | 'invlogit' | 'ismax' | 'logit' | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or the name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class. For more details on the relationships and algorithmic behavior of `BoxConstraint`, `Cost`, `Prior`, `Standardize`, and `Weights`, see “Algorithms” on page 33-1865.

Data Types: `double` | `single` | `char` | `string`

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Cross-Validation Options

`CrossVal` — Flag to train cross-validated classifier

`'off'` (default) | `'on'`

Flag to train a cross-validated classifier, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated classifier with 10 folds.

You can override this cross-validation setting using the `CVPartition`, `Holdout`, `KFold`, or `Leaveout` name-value pair argument. You can use only one cross-validation name-value pair argument at a time to create a cross-validated model.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.

2 For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.

3 Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag

`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1 Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.

2 Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Convergence Control Options

`DeltaGradientTolerance` — Tolerance for gradient difference

nonnegative scalar

Tolerance for the gradient difference between upper and lower violators obtained by Sequential Minimal Optimization (SMO) or Iterative Single Data Algorithm (ISDA), specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar.

If `DeltaGradientTolerance` is `0`, then the software does not use the tolerance for the gradient difference to check for optimization convergence.

The default values are:

- `1e-3` if the solver is SMO (for example, you set `'Solver','SMO'`)
- `0` if the solver is ISDA (for example, you set `'Solver','ISDA'`)

Example: `'DeltaGradientTolerance',1e-2`

Data Types: `double` | `single`

`GapTolerance` — Feasibility gap tolerance

`0` (default) | nonnegative scalar

Feasibility gap tolerance obtained by SMO or ISDA, specified as the comma-separated pair consisting of `'GapTolerance'` and a nonnegative scalar.

If `GapTolerance` is `0`, then the software does not use the feasibility gap tolerance to check for optimization convergence.

Example: `'GapTolerance',1e-2`

Data Types: `double` | `single`

`IterationLimit` — Maximal number of numerical optimization iterations
`1e6` (default) | positive integer

Maximal number of numerical optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The software returns a trained model regardless of whether the optimization routine successfully converges. `Mdl.ConvergenceInfo` contains convergence information.

Example: `'IterationLimit',1e8`

Data Types: `double` | `single`

`KKTTolerance` — Karush-Kuhn-Tucker complementarity conditions violation tolerance
nonnegative scalar

Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862 violation tolerance, specified as the comma-separated pair consisting of `'KKTTolerance'` and a nonnegative scalar.

If `KKTTolerance` is `0`, then the software does not use the KKT complementarity conditions violation tolerance to check for optimization convergence.

The default values are:

- `0` if the solver is SMO (for example, you set `'Solver','SMO'`)
- `1e-3` if the solver is ISDA (for example, you set `'Solver','ISDA'`)

Example: `'KKTTolerance',1e-2`

Data Types: `double` | `single`

`ShrinkagePeriod` — Number of iterations between reductions of active set
`0` (default) | nonnegative integer

Number of iterations between reductions of the active set, specified as the comma-separated pair consisting of `'ShrinkagePeriod'` and a nonnegative integer.

If you set `'ShrinkagePeriod',0`, then the software does not shrink the active set.

Example: `'ShrinkagePeriod',1000`

Data Types: `double` | `single`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of these values:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'BoxConstraint','KernelScale'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitcsvm` by varying the parameters. For information about cross-validation loss see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair argument.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitcsvm` are:

- `BoxConstraint` — `fitcsvm` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.
- `KernelScale` — `fitcsvm` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.
- `KernelFunction` — `fitcsvm` searches among `'gaussian'`, `'linear'`, and `'polynomial'`.
- `PolynomialOrder` — `fitcsvm` searches among integers in the range `[2,4]`.
- `Standardize` — `fitcsvm` searches among `'true'` and `'false'`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example:

```
load fisheriris
params = hyperparameters('fitcsvm',meas,species);
params(1).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize SVM Classifier” on page 33-1840.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained SVM classification model

`ClassificationSVM` model object | `ClassificationPartitionedModel` cross-validated model object

Trained SVM classification model, returned as a `ClassificationSVM` model object or `ClassificationPartitionedModel` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `Leaveout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `ClassificationPartitionedModel` cross-validated classifier. Otherwise, `Mdl` is a `ClassificationSVM` classifier.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.Alpha` in the Command Window to display the trained Lagrange multipliers.

Limitations

- `fitcsvm` trains SVM classifiers for one-class or two-class learning applications. To train SVM classifiers using data with more than two classes, use `fitcecoc`.
- `fitcsvm` supports low-dimensional and moderate-dimensional data sets. For high-dimensional data sets, use `fitclinear` instead.

More About

Box Constraint

A box constraint is a parameter that controls the maximum penalty imposed on margin-violating observations, which helps to prevent overfitting (regularization).

If you increase the box constraint, then the SVM classifier assigns fewer support vectors. However, increasing the box constraint can lead to longer training times.

Gram Matrix

The Gram matrix of a set of n vectors $\{x_1,..,x_n;\ x_j \in R^p\}$ is an n-by-n matrix with element (j,k) defined as $G(x_j,x_k) = <\phi(x_j),\phi(x_k)>$, an inner product of the transformed predictors using the kernel function ϕ.

For nonlinear SVM, the algorithm forms a Gram matrix using the rows of the predictor data X. The dual formalization replaces the inner product of the observations in X with corresponding elements of the resulting Gram matrix (called the “kernel trick”). Consequently, nonlinear SVM operates in the transformed predictor space to find a separating hyperplane.

Karush-Kuhn-Tucker (KKT) Complementarity Conditions

KKT complementarity conditions are optimization constraints required for optimal nonlinear programming solutions.

In SVM, the KKT complementarity conditions are

$$\begin{cases} \alpha_j\left[y_j f\left(x_j\right) - 1 + \xi_j\right] = 0 \\ \xi_j\left(C - \alpha_j\right) = 0 \end{cases}$$

for all $j = 1,...,n$, where $f(x_j) = \phi(x_j)'\beta + b$, ϕ is a kernel function (see Gram matrix on page 33-1862), and ξ_j is a slack variable. If the classes are perfectly separable, then $\xi_j = 0$ for all $j = 1,...,n$.

One-Class Learning

One-class learning, or unsupervised SVM, aims to separate data from the origin in the high-dimensional predictor space (not the original predictor space), and is an algorithm used for outlier detection.

The algorithm resembles that of SVM for binary classification on page 33-1863. The objective is to minimize the dual expression

$$0.5\sum_{jk}\alpha_j\alpha_k G(x_j, x_k)$$

with respect to $\alpha_1, ..., \alpha_n$, subject to

$$\sum \alpha_j = n\nu$$

and $0 \le \alpha_j \le 1$ for all $j = 1,...,n$. The value of $G(x_j,x_k)$ is in element (j,k) of the Gram matrix on page 33-1862.

A small value of ν leads to fewer support vectors and, therefore, a smooth, crude decision boundary. A large value of ν leads to more support vectors and, therefore, a curvy, flexible decision boundary. The optimal value of ν should be large enough to capture the data complexity and small enough to avoid overtraining. Also, $0 < \nu \le 1$.

For more details, see [5].

Support Vector

Support vectors are observations corresponding to strictly positive estimates of $\alpha_1,...,\alpha_n$.

SVM classifiers that yield fewer support vectors for a given training set are preferred.

Support Vector Machines for Binary Classification

The SVM binary classification algorithm searches for an optimal hyperplane that separates the data into two classes. For separable classes, the optimal hyperplane maximizes a margin (space that does not contain any observations) surrounding itself, which creates boundaries for the positive and negative classes. For inseparable classes, the objective is the same, but the algorithm imposes a penalty on the length of the margin for every observation that is on the wrong side of its class boundary.

The linear SVM score function is

$$f(x) = x'\beta + b,$$

where:

- x is an observation (corresponding to a row of `X`).
- The vector β contains the coefficients that define an orthogonal vector to the hyperplane (corresponding to `Mdl.Beta`). For separable data, the optimal margin length is $2/\|\beta\|$.
- b is the bias term (corresponding to `Mdl.Bias`).

The root of $f(x)$ for particular coefficients defines a hyperplane. For a particular hyperplane, $f(z)$ is the distance from point z to the hyperplane.

The algorithm searches for the maximum margin length, while keeping observations in the positive ($y = 1$) and negative ($y = –1$) classes separate.

- For separable classes, the objective is to minimize $\|\beta\|$ with respect to the β and b subject to $y_j f(x_j) \geq 1$, for all $j = 1,..,n$. This is the primal formalization for separable classes.
- For inseparable classes, the algorithm uses slack variables (ξ_j) to penalize the objective function for observations that cross the margin boundary for their class. $\xi_j = 0$ for observations that do not cross the margin boundary for their class, otherwise $\xi_j \geq 0$.

  The objective is to minimize $0.5\|\beta\|^2 + C\sum \xi_j$ with respect to the β, b, and ξ_j subject to $y_j f(x_j) \geq 1 - \xi_j$ and $\xi_j \geq 0$ for all $j = 1,..,n$, and for a positive scalar box constraint on page 33-1862 C. This is the primal formalization for inseparable classes.

The algorithm uses the Lagrange multipliers method to optimize the objective, which introduces n coefficients $\alpha_1,...,\alpha_n$ (corresponding to `Mdl.Alpha`). The dual formalizations for linear SVM are as follows:

- For separable classes, minimize

  $$0.5 \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k x_j' x_k - \sum_{j=1}^{n} \alpha_j$$

  with respect to $\alpha_1,...,\alpha_n$, subject to $\sum \alpha_j y_j = 0$, $\alpha_j \geq 0$ for all $j = 1,...,n$, and Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862.
- For inseparable classes, the objective is the same as for separable classes, except for the additional condition $0 \leq \alpha_j \leq C$ for all $j = 1,..,n$.

The resulting score function is

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\alpha}_j y_j x' x_j + \widehat{b} \ .$$

$\widehat{b}$ is the estimate of the bias and $\widehat{\alpha}_j$ is the jth estimate of the vector $\widehat{\alpha}$, $j = 1,...,n$. Written this way, the score function is free of the estimate of β as a result of the primal formalization.

The SVM algorithm classifies a new observation z using $\text{sign}\left(\widehat{f}(z)\right)$.

In some cases, a nonlinear boundary separates the classes. Nonlinear SVM works in a transformed predictor space to find an optimal, separating hyperplane.

The dual formalization for nonlinear SVM is

$$0.5 \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_j \alpha_k y_j y_k G(x_j, x_k) - \sum_{j=1}^{n} \alpha_j$$

with respect to $\alpha_1,...,\alpha_n$, subject to $\sum \alpha_j y_j = 0$, $0 \le \alpha_j \le C$ for all $j = 1,..,n$, and the KKT complementarity conditions. $G(x_k,x_j)$ are elements of the Gram matrix on page 33-1862. The resulting score function is

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\alpha}_j y_j G(x, x_j) + \widehat{b} .$$

For more details, see Understanding Support Vector Machines on page 18-150, [1], and [3].

Tips

- Unless your data set is large, always try to standardize the predictors (see `Standardize`). Standardization makes predictors insensitive to the scales on which they are measured.
- It is a good practice to cross-validate using the `KFold` name-value pair argument. The cross-validation results determine how well the SVM classifier generalizes.
- For one-class learning:
  - The default setting for the name-value pair argument `Alpha` can lead to long training times. To speed up training, set `Alpha` to a vector mostly composed of `0`s.
  - Set the name-value pair argument `Nu` to a value closer to `0` to yield fewer support vectors and, therefore, a smoother but crude decision boundary.
- Sparsity in support vectors is a desirable property of an SVM classifier. To decrease the number of support vectors, set `BoxConstraint` to a large value. This action increases the training time.
- For optimal training time, set `CacheSize` as high as the memory limit your computer allows.
- If you expect many fewer support vectors than observations in the training set, then you can significantly speed up convergence by shrinking the active set using the name-value pair argument `'ShrinkagePeriod'`. It is a good practice to specify `'ShrinkagePeriod',1000`.
- Duplicate observations that are far from the decision boundary do not affect convergence. However, just a few duplicate observations that occur near the decision boundary can slow down convergence considerably. To speed up convergence, specify `'RemoveDuplicates',true` if:
  - Your data set contains many duplicate observations.
  - You suspect that a few duplicate observations fall near the decision boundary.

  To maintain the original data set during training, `fitcsvm` must temporarily store separate data sets: the original and one without the duplicate observations. Therefore, if you specify `true` for data sets containing few duplicates, then `fitcsvm` consumes close to double the memory of the original data.
- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- For the mathematical formulation of the SVM binary classification algorithm, see “Support Vector Machines for Binary Classification” on page 33-1863 and “Understanding Support Vector Machines” on page 18-150.
- `NaN`, `<undefined>`, empty character vector (`''`), empty string (`""`), and `<missing>` values indicate missing values. `fitcsvm` removes entire rows of data corresponding to a missing

response. When computing total weights (see the next bullets), `fitcsvm` ignores any weight corresponding to an observation with at least one missing predictor. This action can lead to unbalanced prior probabilities in balanced-class problems. Consequently, observation box constraints might not equal `BoxConstraint`.

- `fitcsvm` removes observations that have zero weight or prior probability.
- For two-class learning, if you specify the cost matrix C (see `Cost`), then the software updates the class prior probabilities p (see `Prior`) to p_c by incorporating the penalties described in C.

  Specifically, `fitcsvm` completes these steps:

  1. Compute $p_c^* = p'\mathrm{C}$.
  2. Normalize p_c^* so that the updated prior probabilities sum to 1.

  $$p_c = \frac{1}{\sum_{j=1}^{K} p_{c,j}^*} p_c^* .$$

  K is the number of classes.
  3. Reset the cost matrix to the default

  $$\mathrm{C} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

  4. Remove observations from the training data corresponding to classes with zero prior probability.
- For two-class learning, `fitcsvm` normalizes all observation weights (see `Weights`) to sum to 1. The function then renormalizes the normalized weights to sum up to the updated prior probability of the class to which the observation belongs. That is, the total weight for observation j in class k is

  $$w_j^* = \frac{w_j}{\sum_{\forall j \in \text{Class } k} w_j} p_{c,k} .$$

  w_j is the normalized weight for observation j; $p_{c,k}$ is the updated prior probability of class k (see previous bullet).
- For two-class learning, `fitcsvm` assigns a box constraint to each observation in the training data. The formula for the box constraint of observation j is

  $$C_j = nC_0 w_j^* .$$

  n is the training sample size, C_0 is the initial box constraint (see the `'BoxConstraint'` name-value pair argument), and w_j^* is the total weight of observation j (see previous bullet).
- If you set `'Standardize',true` and the `'Cost'`, `'Prior'`, or `'Weights'` name-value pair argument, then `fitcsvm` standardizes the predictors using their corresponding weighted means and weighted standard deviations. That is, `fitcsvm` standardizes predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*} .$$

$$\mu_j^* = \frac{1}{\sum_k w_k^*} \sum_k w_k^* x_{jk} .$$

x_{jk} is observation k (row) of predictor j (column).

$$\left(\sigma_j^*\right)^2 = \frac{v_1}{v_1^2 - v_2} \sum_k w_k^* \left(x_{jk} - \mu_j^*\right)^2 .$$

$$v_1 = \sum_j w_j^* .$$

$$v_2 = \sum_j \left(w_j^*\right)^2 .$$

- Assume that `p` is the proportion of outliers that you expect in the training data, and that you set `'OutlierFraction',p`.
  - For one-class learning, the software trains the bias term such that 100`p`% of the observations in the training data have negative scores.
  - The software implements robust learning for two-class learning. In other words, the software attempts to remove 100`p`% of the observations when the optimization algorithm converges. The removed observations correspond to gradients that are large in magnitude.
- If your predictor data contains categorical variables, then the software generally uses full dummy encoding for these variables. The software creates one dummy variable for each level of each categorical variable.
  - The `PredictorNames` property stores one element for each of the original predictor variable names. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `PredictorNames` is a 1-by-3 cell array of character vectors containing the original names of the predictor variables.
  - The `ExpandedPredictorNames` property stores one element for each of the predictor variables, including the dummy variables. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `ExpandedPredictorNames` is a 1-by-5 cell array of character vectors containing the names of the predictor variables and the new dummy variables.
  - Similarly, the `Beta` property stores one beta coefficient for each predictor, including the dummy variables.
  - The `SupportVectors` property stores the predictor values for the support vectors, including the dummy variables. For example, assume that there are m support vectors and three predictors, one of which is a categorical variable with three levels. Then `SupportVectors` is an n-by-5 matrix.
  - The `X` property stores the training data as originally input and does not include the dummy variables. When the input is a table, `X` contains only the columns used as predictors.
- For predictors specified in a table, if any of the variables contain ordered (ordinal) categories, the software uses ordinal encoding for these variables.
  - For a variable with k ordered levels, the software creates k – 1 dummy variables. The jth dummy variable is –1 for levels up to j, and +1 for levels j + 1 through k.

- The names of the dummy variables stored in the `ExpandedPredictorNames` property indicate the first level with the value +1. The software stores k – 1 additional predictor names for the dummy variables, including the names of levels 2, 3, ..., k.

- All solvers implement $L1$ soft-margin minimization.
- For one-class learning, the software estimates the Lagrange multipliers, $\alpha_1,...,\alpha_n$, such that

$$\sum_{j=1}^{n} \alpha_j = n\nu .$$

References

[1] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[2] Fan, R.-E., P.-H. Chen, and C.-J. Lin. "Working set selection using second order information for training support vector machines." *Journal of Machine Learning Research*, Vol. 6, 2005, pp. 1889–1918.

[3] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[4] Kecman V., T. -M. Huang, and M. Vogt. "Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance." *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

[5] Scholkopf, B., J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. "Estimating the Support of a High-Dimensional Distribution." *Neural Comput.*, Vol. 13, Number 7, 2001, pp. 1443–1471.

[6] Scholkopf, B., and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press, 2002.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see "Parallel Bayesian Optimization" on page 10-7.

For general information about parallel computing, see "Run MATLAB Functions with Automatic Parallel Support" (Parallel Computing Toolbox).

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'KernelFunction'` option must not specify a custom kernel.
- The `'Solver'` option can be set to only `'SMO'`.
- The `'OutlierFraction'` option is not supported.
- The `'Alpha'` option must specify a feasible starting point.
- The `'OptimizeHyperparameters'` option is not supported.
- One-class classification is not supported. The labels must contain two different classes.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `ClassificationSVM` | `CompactClassificationSVM` | `fitSVMPosterior` | `fitcecoc` | `fitclinear` | `predict` | `quadprog` | `rng`

Topics

“Train SVM Classifiers Using a Gaussian Kernel” on page 18-156
“Train SVM Classifier Using Custom Kernel” on page 18-159
“Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45
“Optimize an SVM Classifier Fit Using Bayesian Optimization” on page 10-55
“Understanding Support Vector Machines” on page 18-150

Introduced in R2014a

fitctree

Fit binary decision tree for multiclass classification

Syntax

```
tree = fitctree(Tbl,ResponseVarName)
tree = fitctree(Tbl,formula)
tree = fitctree(Tbl,Y)

tree = fitctree(X,Y)

tree = fitctree(___,Name,Value)
```

Description

`tree = fitctree(Tbl,ResponseVarName)` returns a fitted binary classification decision tree based on the input variables (also known as predictors, features, or attributes) contained in the table `Tbl` and output (response or labels) contained in `Tbl.ResponseVarName`. The returned binary tree splits branching nodes based on the values of a column of `Tbl`.

`tree = fitctree(Tbl,formula)` returns a fitted binary classification decision tree based on the input variables contained in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `tree`.

`tree = fitctree(Tbl,Y)` returns a fitted binary classification decision tree based on the input variables contained in the table `Tbl` and output in vector `Y`.

`tree = fitctree(X,Y)` returns a fitted binary classification decision tree based on the input variables contained in matrix `X` and output `Y`. The returned binary tree splits branching nodes based on the values of a column of `X`.

`tree = fitctree(___,Name,Value)` fits a tree with additional options specified by one or more name-value pair arguments, using any of the previous syntaxes. For example, you can specify the algorithm used to find the best split on a categorical predictor, grow a cross-validated tree, or hold out a fraction of the input data for validation.

Examples

Grow a Classification Tree

Grow a classification tree using the `ionosphere` data set.

```
load ionosphere
tc = fitctree(X,Y)
```

```
tc =
  ClassificationTree
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
```

```
          ScoreTransform: 'none'
         NumObservations: 351


  Properties, Methods
```

Control Tree Depth

You can control the depth of the trees using the `MaxNumSplits`, `MinLeafSize`, or `MinParentSize` name-value pair parameters. `fitctree` grows deep decision trees by default. You can grow shallower trees to reduce model complexity or computation time.

Load the `ionosphere` data set.

```
load ionosphere
```

The default values of the tree depth controllers for growing classification trees are:

- `n - 1` for `MaxNumSplits`. `n` is the training sample size.
- `1` for `MinLeafSize`.
- `10` for `MinParentSize`.

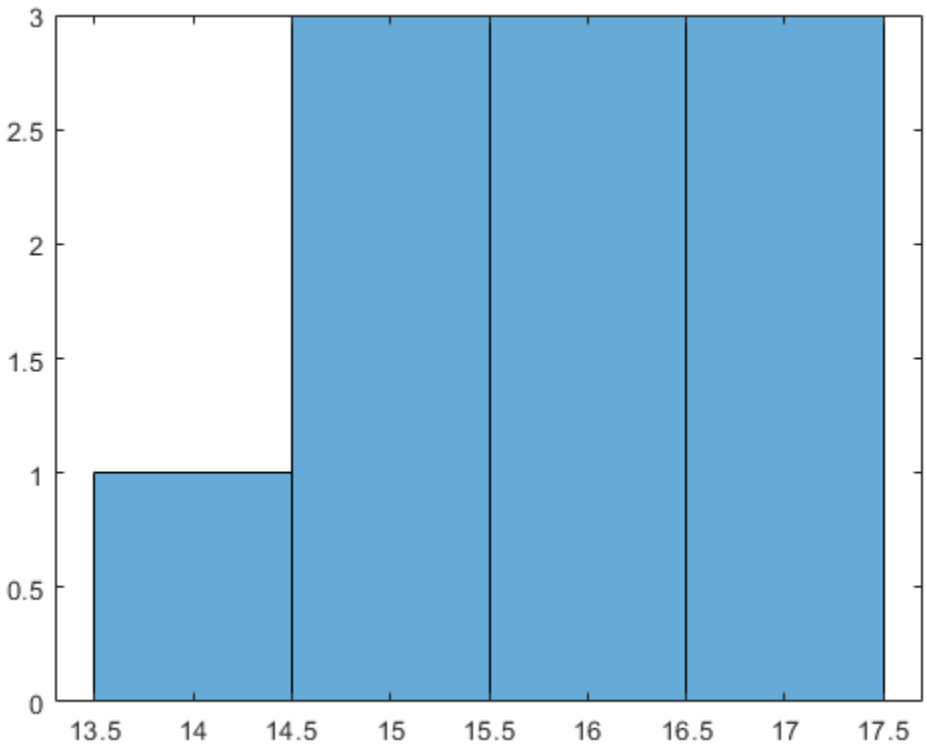
These default values tend to grow deep trees for large training sample sizes.

Train a classification tree using the default values for tree depth control. Cross-validate the model by using 10-fold cross-validation.

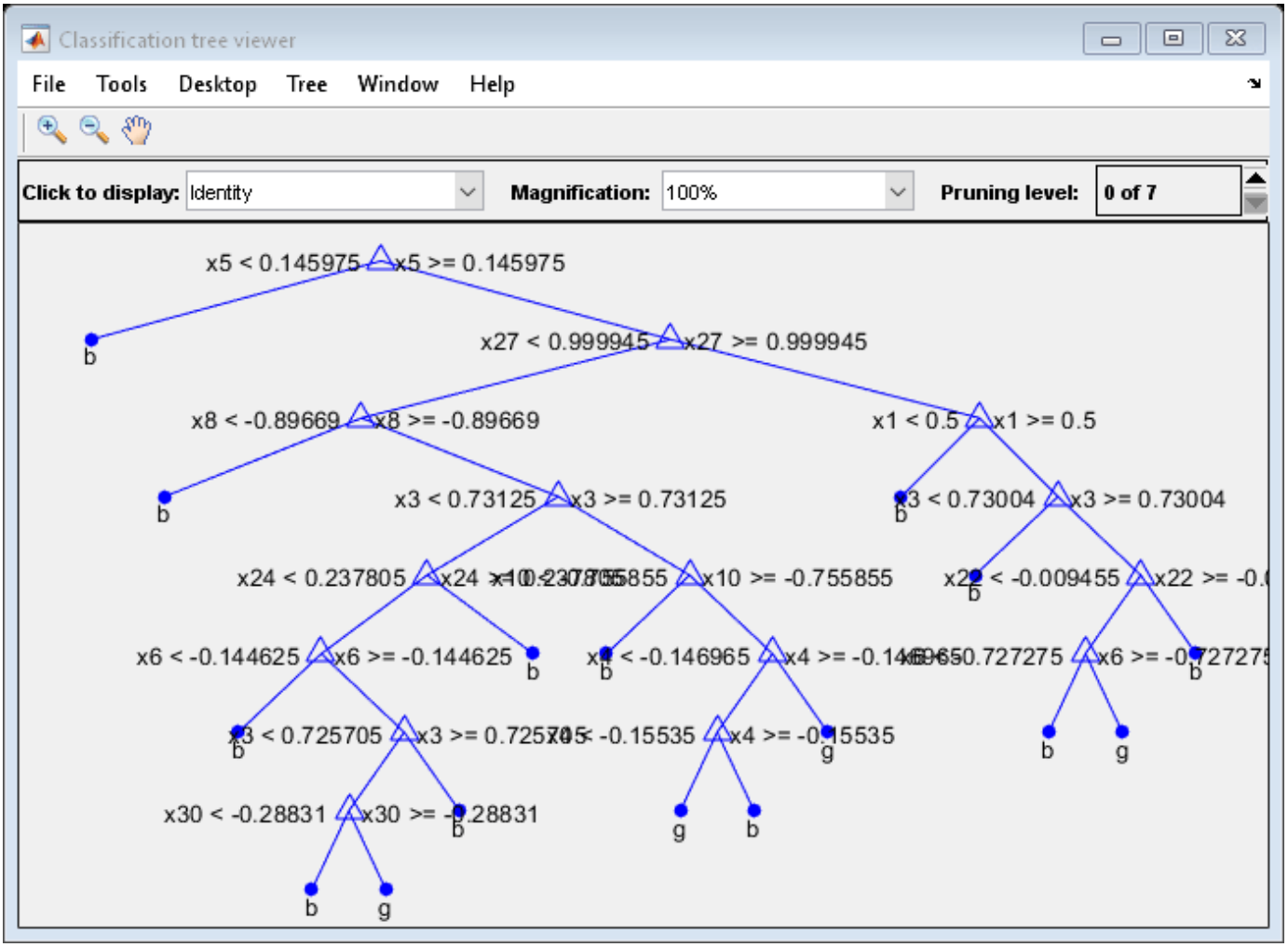
```
rng(1); % For reproducibility
MdlDefault = fitctree(X,Y,'CrossVal','on');
```

Draw a histogram of the number of imposed splits on the trees. Also, view one of the trees.

```
numBranches = @(x)sum(x.IsBranch);
mdlDefaultNumSplits = cellfun(numBranches, MdlDefault.Trained);

figure;
histogram(mdlDefaultNumSplits)
```



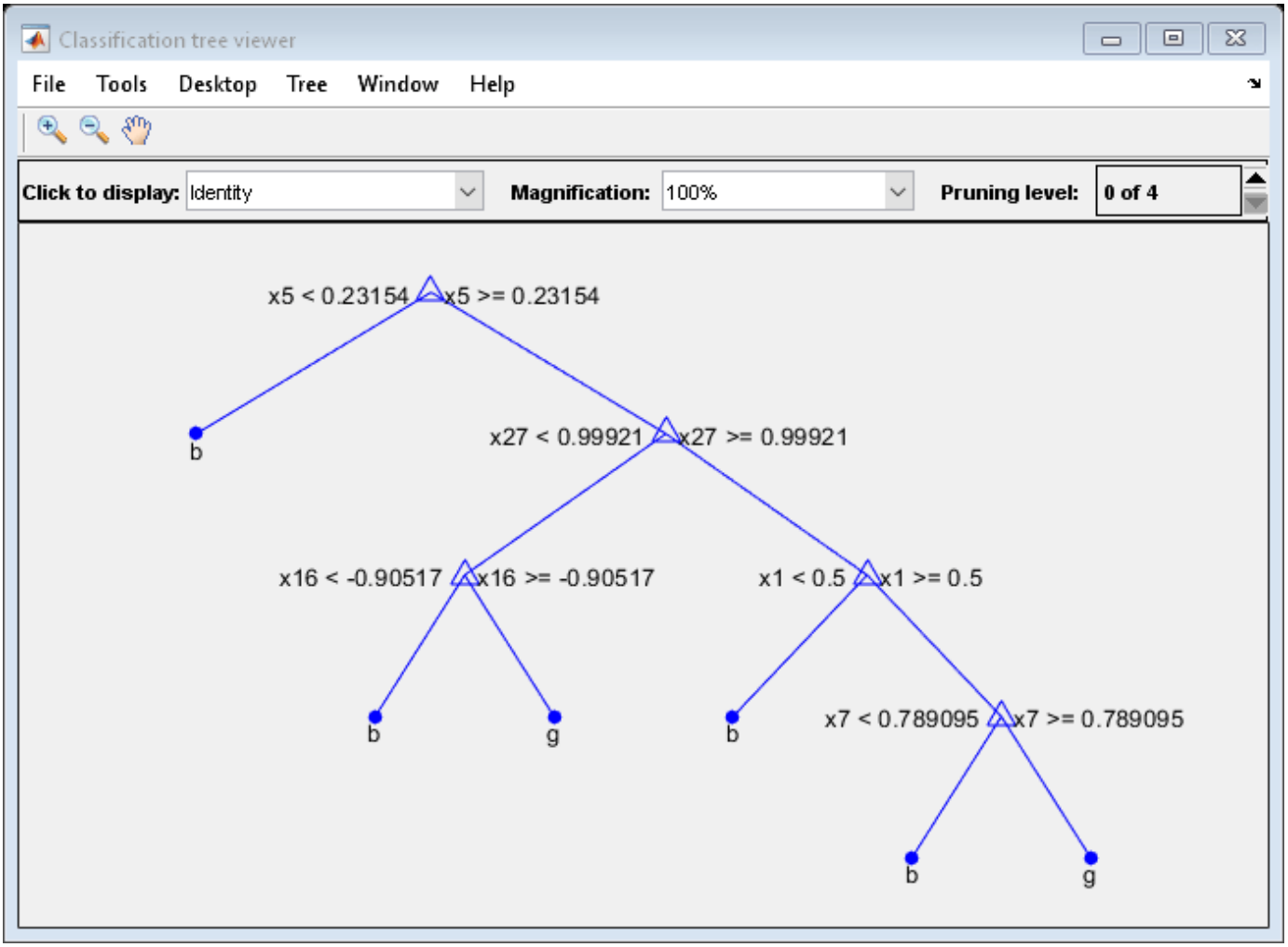

```
view(MdlDefault.Trained{1},'Mode','graph')
```



The average number of splits is around 15.

Suppose that you want a classification tree that is not as complex (deep) as the ones trained using the default number of splits. Train another classification tree, but set the maximum number of splits at 7, which is about half the mean number of splits from the default classification tree. Cross-validate the model by using 10-fold cross-validation.

```
Mdl7 = fitctree(X,Y,'MaxNumSplits',7,'CrossVal','on');
view(Mdl7.Trained{1},'Mode','graph')
```



Compare the cross-validation classification errors of the models.

```
classErrorDefault = kfoldLoss(MdlDefault)
```

```
classErrorDefault = 0.1168
```

```
classError7 = kfoldLoss(Mdl7)
```

```
classError7 = 0.1311
```

`Mdl7` is much less complex and performs only slightly worse than `MdlDefault`.

Optimize Classification Tree

This example shows how to optimize hyperparameters automatically using `fitctree`. The example uses Fisher's iris data.

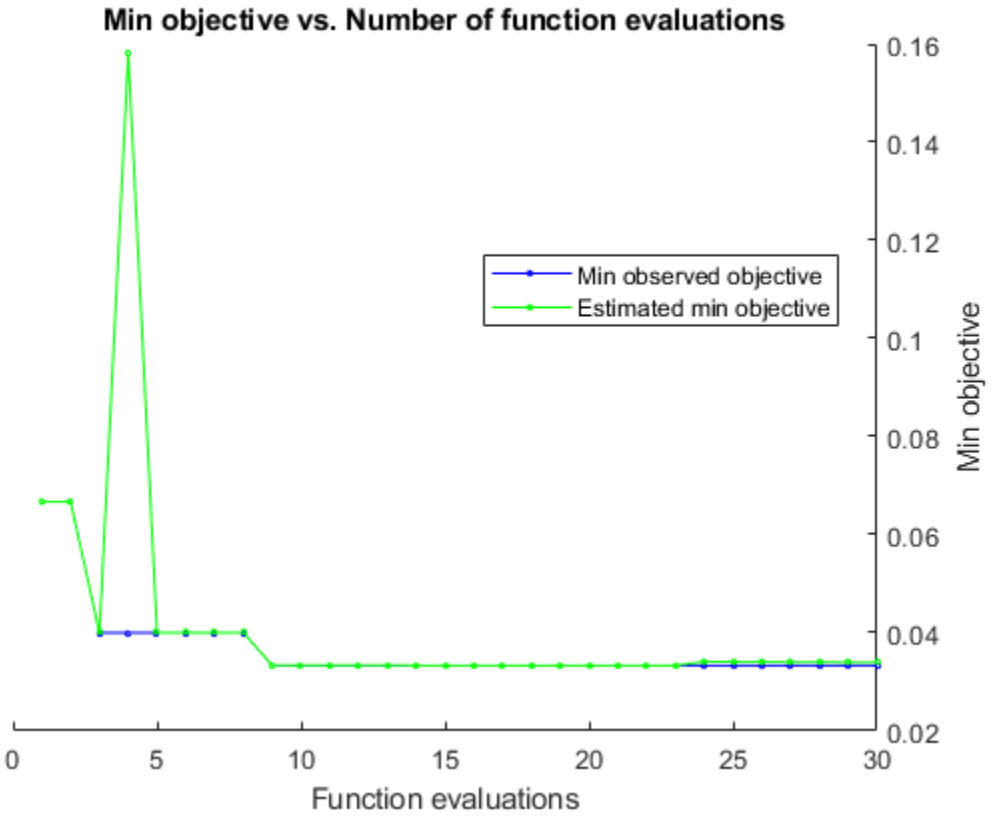
Load Fisher's iris data.

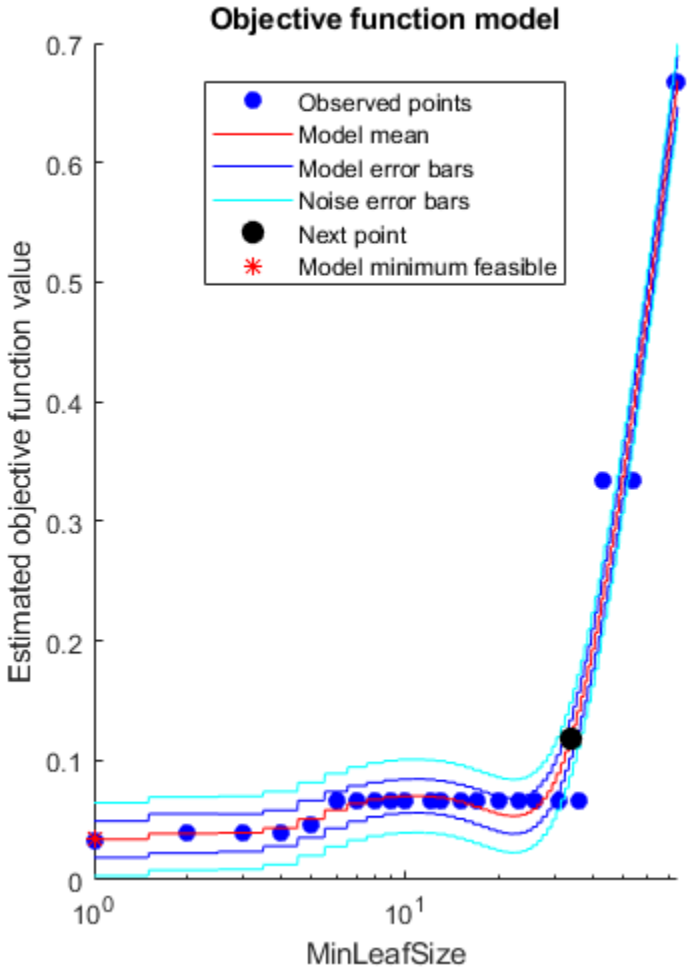
```
load fisheriris
```

Optimize the cross-validation loss of the classifier, using the data in `meas` to predict the response in `species`.

```
X = meas;
Y = species;
Mdl = fitctree(X,Y,'OptimizeHyperparameters','auto')
```

```
|=====================================================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  MinLeafSize |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|    1 | Best   |    0.066667 |     0.76731 |    0.066667 |    0.066667 |           31 |
|    2 | Accept |    0.066667 |     0.26782 |    0.066667 |    0.066667 |           12 |
|    3 | Best   |        0.04 |     0.22031 |        0.04 |    0.040003 |            2 |
|    4 | Accept |     0.66667 |     0.16866 |        0.04 |     0.15796 |           73 |
|    5 | Accept |        0.04 |     0.14894 |        0.04 |    0.040009 |            2 |
|    6 | Accept |     0.66667 |     0.16882 |        0.04 |    0.040012 |           74 |
|    7 | Accept |    0.066667 |     0.15502 |        0.04 |    0.040012 |           20 |
|    8 | Accept |        0.04 |     0.16831 |        0.04 |    0.040008 |            4 |
|    9 | Best   |    0.033333 |      0.1614 |    0.033333 |     0.03335 |            1 |
|   10 | Accept |    0.066667 |      0.1672 |    0.033333 |     0.03335 |            7 |
|   11 | Accept |        0.04 |     0.16183 |    0.033333 |    0.033348 |            3 |
|   12 | Accept |    0.066667 |     0.14057 |    0.033333 |    0.033348 |           26 |
|   13 | Accept |    0.046667 |     0.14334 |    0.033333 |    0.033347 |            5 |
|   14 | Accept |    0.033333 |     0.15167 |    0.033333 |     0.03334 |            1 |
|   15 | Accept |    0.066667 |     0.14523 |    0.033333 |    0.033339 |           15 |
|   16 | Accept |    0.033333 |     0.15999 |    0.033333 |    0.033337 |            1 |
|   17 | Accept |    0.033333 |     0.13856 |    0.033333 |    0.033336 |            1 |
|   18 | Accept |     0.33333 |     0.12745 |    0.033333 |    0.033336 |           43 |
|   19 | Accept |    0.066667 |     0.15121 |    0.033333 |    0.033336 |            9 |
|   20 | Accept |    0.066667 |     0.15772 |    0.033333 |    0.033336 |            6 |
|=====================================================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  MinLeafSize |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================|
|   21 | Accept |    0.066667 |     0.17974 |    0.033333 |    0.033336 |           17 |
|   22 | Accept |    0.066667 |     0.17015 |    0.033333 |    0.033336 |           10 |
|   23 | Accept |    0.066667 |     0.16476 |    0.033333 |    0.033336 |           36 |
|   24 | Accept |     0.33333 |     0.16325 |    0.033333 |    0.034075 |           54 |
|   25 | Accept |        0.04 |     0.17066 |    0.033333 |    0.034054 |            2 |
|   26 | Accept |        0.04 |     0.17939 |    0.033333 |    0.034022 |            3 |
|   27 | Accept |        0.04 |     0.17819 |    0.033333 |    0.033997 |            4 |
|   28 | Accept |    0.066667 |     0.16231 |    0.033333 |    0.033973 |           23 |
|   29 | Accept |    0.066667 |     0.17389 |    0.033333 |    0.033946 |            8 |
|   30 | Accept |    0.066667 |     0.17803 |    0.033333 |    0.033922 |           13 |
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
20
25
30




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 57.1119 seconds
Total objective function evaluation time: 5.5917

Best observed feasible point:
    MinLeafSize
    ___________

          1

Observed objective function value = 0.033333
Estimated objective function value = 0.033922
Function evaluation time = 0.1614

Best estimated feasible point (according to models):
    MinLeafSize
    ___________
```

```
         1

Estimated objective function value = 0.033922
Estimated function evaluation time = 0.17337

Mdl =
  ClassificationTree
                          ResponseName: 'Y'
                 CategoricalPredictors: []
                            ClassNames: {'setosa'  'versicolor'  'virginica'}
                       ScoreTransform: 'none'
                      NumObservations: 150
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]


  Properties, Methods
```

Unbiased Predictor Importance Estimates

Load the `census1994` data set. Consider a model that predicts a person's salary category given their age, working class, education level, martial status, race, sex, capital gain and loss, and number of working hours per week.

```
load census1994
X = adultdata(:,{'age','workClass','education_num','marital_status','race',...
    'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

Display the number of categories represented in the categorical variables using `summary`.

```
summary(X)
```

```
Variables:

    age: 32561x1 double

        Values:

            Min          17
            Median       37
            Max          90

    workClass: 32561x1 categorical

        Values:

            Federal-gov            960
            Local-gov             2093
            Never-worked             7
            Private              22696
            Self-emp-inc          1116
            Self-emp-not-inc      2541
            State-gov             1298
            Without-pay             14
            NumMissing            1836
```

```
education_num: 32561x1 double

    Values:

        Min           1
        Median       10
        Max          16

marital_status: 32561x1 categorical

    Values:

        Divorced                  4443
        Married-AF-spouse           23
        Married-civ-spouse       14976
        Married-spouse-absent      418
        Never-married            10683
        Separated                 1025
        Widowed                    993

race: 32561x1 categorical

    Values:

        Amer-Indian-Eskimo        311
        Asian-Pac-Islander       1039
        Black                    3124
        Other                     271
        White                   27816

sex: 32561x1 categorical

    Values:

        Female     10771
        Male       21790

capital_gain: 32561x1 double

    Values:

        Min            0
        Median         0
        Max        99999

capital_loss: 32561x1 double

    Values:

        Min            0
        Median         0
        Max         4356

hours_per_week: 32561x1 double

    Values:

        Min           1
```

```
            Median       40
            Max          99

    salary: 32561x1 categorical

        Values:

            <=50K     24720
            >50K       7841
```

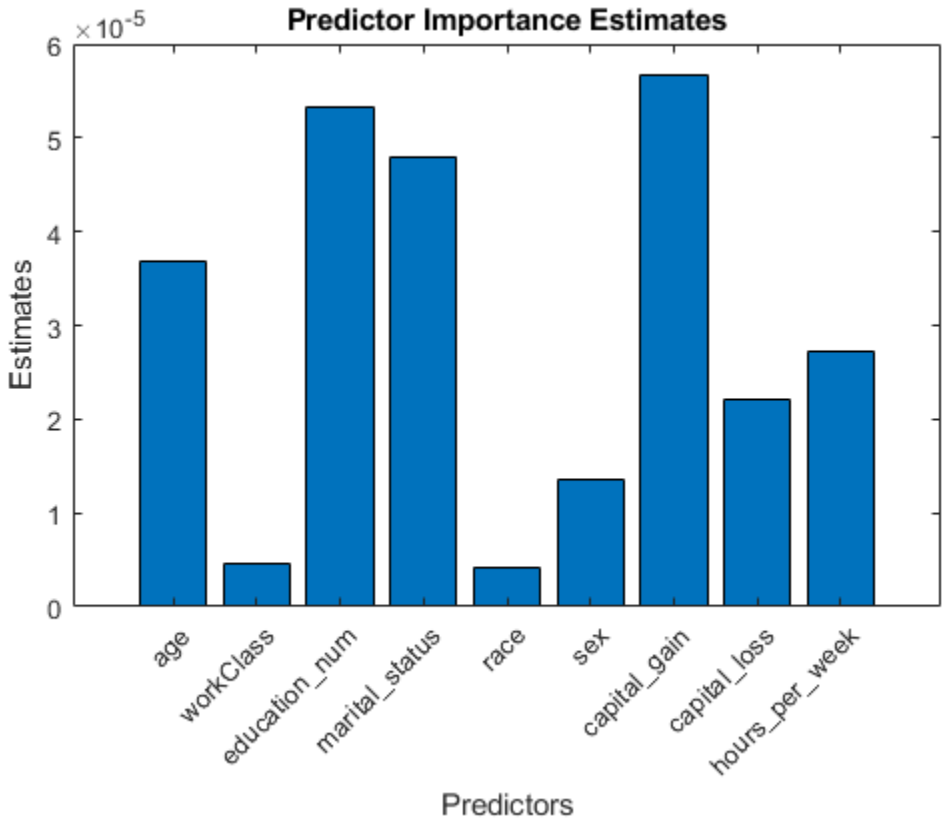
Because there are few categories represented in the categorical variables compared to levels in the continuous variables, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over the categorical variables.

Train a classification tree using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing observations in the data, specify usage of surrogate splits.

```
Mdl = fitctree(X,'salary','PredictorSelection','curvature',...
    'Surrogate','on');
```

Estimate predictor importance values by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. Compare the estimates using a bar graph.

```
imp = predictorImportance(Mdl);

figure;
bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `capital_gain` is the most important predictor, followed by `education_num`.

Optimize Classification Tree on Tall Array

This example shows how to optimize hyperparameters of a classification tree automatically using a tall array. The sample data set `airlinesmall.csv` is a large data set that contains a tabular file of airline flight data. This example creates a tall table containing the data and uses it to run the optimization procedure.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Create a datastore that references the folder location with the data. Select a subset of the variables to work with, and treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Create a tall table that contains the data in the datastore.

```
ds = datastore('airlinesmall.csv');
ds.SelectedVariableNames = {'Month','DayofMonth','DayOfWeek',...
                            'DepTime','ArrDelay','Distance','DepDelay'};
ds.TreatAsMissing = 'NA';
tt  = tall(ds) % Tall table
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

tt =

  M×7 tall table

    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    Distance    DepDelay
    _____    __________    _________    _______    ________    ________    ________

      10          21            3          642          8          308          12
      10          26            1         1021          8          296           1
      10          23            5         2055         21          480          20
      10          23            5         1332         13          296          12
      10          22            4          629          4          373          -1
      10          28            3         1446         59          308          63
      10           8            4          928          3          447          -2
      10          10            6          859         11          954          -1
       :           :            :            :          :            :           :
       :           :            :            :          :            :           :
```

Determine the flights that are late by 10 minutes or more by defining a logical variable that is true for a late flight. This variable contains the class labels. A preview of this variable includes the first few rows.

```
Y = tt.DepDelay > 10 % Class labels
```

```
Y =

  M×1 tall logical array

   1
   0
   1
   1
   0
   1
   0
   0
   :
   :
```

Create a tall array for the predictor data.

```
X = tt{:,1:end-1} % Predictor data
```

```
X =

  M×6 tall double matrix

         10          21           3         642           8         308
         10          26           1        1021           8         296
         10          23           5        2055          21         480
         10          23           5        1332          13         296
         10          22           4         629           4         373
         10          28           3        1446          59         308
         10           8           4         928           3         447
         10          10           6         859          11         954
```

```
          :          :          :          :          :          :
          :          :          :          :          :          :
```

Remove rows in X and Y that contain missing data.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:end-1);
Y = R(:,end);
```

Standardize the predictor variables.

```
Z = zscore(X);
```

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value pair argument. Find the optimal `'MinLeafSize'` value that minimizes holdout cross-validation loss. (Specifying `'auto'` uses `'MinLeafSize'`.) For reproducibility, use the `'expected-improvement-plus'` acquisition function and set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitctree(Z,Y,...
    'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('Holdout',0.3,...
    'AcquisitionFunctionName','expected-improvement-plus'))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 3: Completed in 5.6 sec
- Pass 2 of 3: Completed in 2.1 sec
- Pass 3 of 3: Completed in 3.4 sec
Evaluation completed in 13 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.73 sec
Evaluation completed in 0.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.2 sec
Evaluation completed in 1.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.64 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 0.72 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 5.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
```

```
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.87 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.57 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.68 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.65 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 4.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.63 sec
- Pass 2 of 4: Completed in 0.85 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 4.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
- Pass 2 of 4: Completed in 0.88 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 3 sec
Evaluation completed in 6.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.72 sec
- Pass 2 of 4: Completed in 0.96 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 4.2 sec
Evaluation completed in 7.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.77 sec
- Pass 2 of 4: Completed in 0.95 sec
- Pass 3 of 4: Completed in 0.65 sec
- Pass 4 of 4: Completed in 4.8 sec
Evaluation completed in 7.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.79 sec
- Pass 2 of 4: Completed in 1 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 5.1 sec
Evaluation completed in 8.4 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.89 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 5.8 sec
Evaluation completed in 9.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.63 sec
- Pass 4 of 4: Completed in 5.2 sec
Evaluation completed in 8.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.6 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.74 sec
- Pass 4 of 4: Completed in 4.8 sec
Evaluation completed in 9.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.68 sec
- Pass 4 of 4: Completed in 3.9 sec
Evaluation completed in 7.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.6 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.7 sec
- Pass 4 of 4: Completed in 3 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.66 sec
- Pass 4 of 4: Completed in 2.5 sec
Evaluation completed in 6.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.66 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 5.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.69 sec
- Pass 4 of 4: Completed in 1.9 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 0.67 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.3 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 0.65 sec
```

```
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.67 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.73 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.65 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 2.4 sec
Evaluation completed in 2.6 sec
|=====================================================================================|
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  MinLeafSize |
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================|
|    1 | Best   |     0.11572 |      197.12 |     0.11572 |     0.11572 |           10 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.4 sec
Evaluation completed in 0.56 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.93 sec
Evaluation completed in 1.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 0.84 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 1.6 sec
|    2 | Accept |     0.19635 |      10.496 |     0.11572 |     0.12008 |        48298 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.33 sec
Evaluation completed in 0.47 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 0.99 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.74 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
```

```
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.73 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.68 sec
- Pass 4 of 4: Completed in 0.77 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.5 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.75 sec
Evaluation completed in 0.87 sec
|    3 | Best   |      0.1048 |      44.614 |      0.1048 |     0.11431 |         3166 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.3 sec
Evaluation completed in 0.45 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 0.97 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.99 sec
```

```
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.73 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.82 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.6 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.5 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.63 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 4.1 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.66 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.62 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 4.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 1 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.8 sec
Evaluation completed in 0.94 sec
|    4 | Best   |     0.10094 |      91.723 |     0.10094 |     0.10574 |          180 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.3 sec
Evaluation completed in 0.42 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.93 sec
Evaluation completed in 1.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.66 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.1 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 0.88 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.5 sec
- Pass 4 of 4: Completed in 0.98 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.5 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.7 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.52 sec
```

```
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.64 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.57 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.97 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 0.85 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 0.82 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec
|    5 | Best   |     0.10087 |       82.84 |     0.10087 |     0.10085 |          219 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.32 sec
Evaluation completed in 0.45 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.66 sec
- Pass 3 of 4: Completed in 0.5 sec
```

```
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.85 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.6 sec
- Pass 4 of 4: Completed in 0.84 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.87 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 0.92 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.77 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.68 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
```

```
Evaluation completed in 0.93 sec
|    6 | Accept |     0.10155 |      61.043 |     0.10087 |     0.10089 |         1089 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.33 sec
Evaluation completed in 0.46 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.85 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.87 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.98 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.62 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
```

```
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 4.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 2.7 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.64 sec
- Pass 2 of 4: Completed in 0.87 sec
- Pass 3 of 4: Completed in 1.2 sec
- Pass 4 of 4: Completed in 3.7 sec
Evaluation completed in 7.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.73 sec
- Pass 2 of 4: Completed in 0.92 sec
- Pass 3 of 4: Completed in 0.6 sec
- Pass 4 of 4: Completed in 4.4 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.86 sec
- Pass 2 of 4: Completed in 1.5 sec
- Pass 3 of 4: Completed in 0.64 sec
- Pass 4 of 4: Completed in 4.8 sec
Evaluation completed in 8.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.9 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 0.65 sec
- Pass 4 of 4: Completed in 5.2 sec
Evaluation completed in 8.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.73 sec
- Pass 4 of 4: Completed in 5.6 sec
Evaluation completed in 9.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.5 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 0.75 sec
- Pass 4 of 4: Completed in 5.8 sec
Evaluation completed in 10 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.3 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 1.2 sec
- Pass 4 of 4: Completed in 5.1 sec
Evaluation completed in 9.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.4 sec
- Pass 2 of 4: Completed in 1.5 sec
- Pass 3 of 4: Completed in 0.7 sec
- Pass 4 of 4: Completed in 4.1 sec
Evaluation completed in 8.5 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.4 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 0.71 sec
- Pass 4 of 4: Completed in 3.6 sec
Evaluation completed in 7.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.5 sec
- Pass 2 of 4: Completed in 1.8 sec
- Pass 3 of 4: Completed in 0.74 sec
- Pass 4 of 4: Completed in 3.2 sec
Evaluation completed in 7.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.4 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 0.73 sec
- Pass 4 of 4: Completed in 2.8 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.5 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 0.82 sec
- Pass 4 of 4: Completed in 2.4 sec
Evaluation completed in 7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 2 sec
- Pass 2 of 4: Completed in 1.9 sec
- Pass 3 of 4: Completed in 0.79 sec
- Pass 4 of 4: Completed in 2.3 sec
Evaluation completed in 7.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.6 sec
- Pass 2 of 4: Completed in 1.8 sec
- Pass 3 of 4: Completed in 0.73 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 6.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.6 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 0.79 sec
- Pass 4 of 4: Completed in 2.3 sec
Evaluation completed in 7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.7 sec
- Pass 2 of 4: Completed in 1.9 sec
- Pass 3 of 4: Completed in 0.8 sec
- Pass 4 of 4: Completed in 1.8 sec
Evaluation completed in 6.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.7 sec
- Pass 2 of 4: Completed in 1.8 sec
- Pass 3 of 4: Completed in 0.77 sec
- Pass 4 of 4: Completed in 1.8 sec
Evaluation completed in 6.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.4 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 0.73 sec
```

```
- Pass 4 of 4: Completed in 1.8 sec
Evaluation completed in 6.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.5 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 1.3 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 6.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.5 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 0.73 sec
- Pass 4 of 4: Completed in 1.8 sec
Evaluation completed in 6.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.5 sec
|    7 | Accept |     0.13495 |      241.76 |     0.10087 |     0.10089 |            1 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.31 sec
Evaluation completed in 0.44 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 0.67 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.74 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.85 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.6 sec
```

```
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.62 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 4.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
- Pass 2 of 4: Completed in 1.5 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 1.9 sec
Evaluation completed in 6.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.69 sec
- Pass 2 of 4: Completed in 0.88 sec
- Pass 3 of 4: Completed in 0.75 sec
- Pass 4 of 4: Completed in 2.1 sec
Evaluation completed in 5.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 4.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.84 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 4.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.68 sec
- Pass 2 of 4: Completed in 0.85 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 4.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.68 sec
- Pass 2 of 4: Completed in 0.91 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 2.4 sec
Evaluation completed in 5.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.92 sec
- Pass 2 of 4: Completed in 0.86 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 4.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.69 sec
```

```
- Pass 2 of 4: Completed in 0.91 sec
- Pass 3 of 4: Completed in 0.63 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.67 sec
- Pass 2 of 4: Completed in 0.86 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.99 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
- Pass 2 of 4: Completed in 0.9 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.95 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.73 sec
- Pass 2 of 4: Completed in 0.91 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.91 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.76 sec
- Pass 2 of 4: Completed in 0.93 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.91 sec
Evaluation completed in 1.1 sec
|    8 | Accept |     0.10246 |      115.31 |     0.10087 |     0.10089 |           58 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.34 sec
Evaluation completed in 0.49 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.87 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
```

```
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.6 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.95 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.94 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.4 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.77 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.72 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.6 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.3 sec
Evaluation completed in 1.4 sec
|    9 | Accept |     0.10173 |      77.229 |     0.10087 |     0.10086 |          418 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.32 sec
Evaluation completed in 0.46 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.84 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.75 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.91 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.82 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 0.82 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.95 sec
Evaluation completed in 3.9 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.71 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 5.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.83 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.67 sec
- Pass 2 of 4: Completed in 0.87 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.82 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.84 sec
- Pass 3 of 4: Completed in 0.62 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.64 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.56 sec
```

```
- Pass 4 of 4: Completed in 0.88 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.62 sec
- Pass 2 of 4: Completed in 0.9 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.89 sec
|   10 | Accept |     0.10114 |      94.532 |     0.10087 |     0.10091 |          123 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.86 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 0.99 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.73 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.85 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.82 sec
- Pass 3 of 4: Completed in 0.64 sec
- Pass 4 of 4: Completed in 0.94 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.53 sec
```

```
- Pass 4 of 4: Completed in 0.97 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 4.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.84 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 4.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.63 sec
- Pass 2 of 4: Completed in 0.84 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.2 sec
```

```
- Pass 2 of 4: Completed in 0.83 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.81 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.77 sec
Evaluation completed in 0.89 sec
|   11 | Best   |      0.1008 |      90.637 |      0.1008 |     0.10088 |          178 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.38 sec
Evaluation completed in 0.52 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.59 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.93 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.57 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.91 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.85 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
```

```
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 1.1 sec
- Pass 2 of 4: Completed in 0.81 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 3.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.82 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.66 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.62 sec
- Pass 2 of 4: Completed in 0.85 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 0.86 sec
- Pass 3 of 4: Completed in 1.1 sec
- Pass 4 of 4: Completed in 0.96 sec
Evaluation completed in 4.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 0.8 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.5 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.69 sec
- Pass 2 of 4: Completed in 0.84 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.83 sec
Evaluation completed in 3.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.76 sec
Evaluation completed in 0.89 sec
|   12 | Accept |      0.1008 |      90.267 |      0.1008 |     0.10086 |          179 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.32 sec
Evaluation completed in 0.45 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.9 sec
Evaluation completed in 1.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.77 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.77 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.72 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.74 sec
- Pass 3 of 4: Completed in 0.51 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 3.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.74 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 0.97 sec
|   13 | Accept |     0.11126 |      32.134 |      0.1008 |     0.10084 |        10251 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.32 sec
```

```
Evaluation completed in 0.45 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.85 sec
Evaluation completed in 0.99 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.75 sec
- Pass 3 of 4: Completed in 0.55 sec
- Pass 4 of 4: Completed in 0.74 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.7 sec
- Pass 3 of 4: Completed in 0.57 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.68 sec
- Pass 3 of 4: Completed in 0.53 sec
- Pass 4 of 4: Completed in 0.79 sec
Evaluation completed in 3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.91 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.86 sec
Evaluation completed in 3.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 1 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.71 sec
- Pass 3 of 4: Completed in 0.64 sec
- Pass 4 of 4: Completed in 0.99 sec
Evaluation completed in 3.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 0.58 sec
- Pass 4 of 4: Completed in 0.94 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 0.77 sec
- Pass 3 of 4: Completed in 0.59 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.3 sec
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 0.78 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.9 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 0.72 sec
- Pass 3 of 4: Completed in 0.52 sec
- Pass 4 of 4: Completed in 0.89 sec
Evaluation completed in 3.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.58 sec
- Pass 2 of 4: Completed in 0.76 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.8 sec
Evaluation completed in 3.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 0.61 sec
- Pass 4 of 4: Completed in 0.76 sec
Evaluation completed in 3.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.83 sec
Evaluation completed in 0.97 sec
|   14 | Accept |     0.10154 |      66.262 |      0.1008 |     0.10085 |          736 |
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.36 sec
Evaluation completed in 0.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.89 sec
Evaluation completed in 1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.74 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 0.67 sec
- Pass 3 of 4: Completed in 0.56 sec
- Pass 4 of 4: Completed in 0.78 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 0.69 sec
- Pass 3 of 4: Completed in 0.54 sec
- Pass 4 of 4: Completed in 0.84 sec
Evaluation completed in 3.1 sec
Evaluating tall expression using the Parallel Pool 'local':
Evaluation 0% ...

Mdl =
  CompactClassificationTree
             ResponseName: 'Y'
```

```
    CategoricalPredictors: []
               ClassNames: [0 1]
          ScoreTransform: 'none'


  Properties, Methods


FitInfo = struct with no fields.



HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                      ObjectiveFcn: @createObjFcn/tallObjFcn
              VariableDescriptions: [4×1 optimizableVariable]
                           Options: [1×1 struct]
                      MinObjective: 0.1004
                   XAtMinObjective: [1×1 table]
             MinEstimatedObjective: 0.1008
          XAtMinEstimatedObjective: [1×1 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 3.0367e+03
                         NextPoint: [1×1 table]
                             XTrace: [30×1 table]
                    ObjectiveTrace: [30×1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30×1 cell}
       ObjectiveEvaluationTimeTrace: [30×1 double]
                 IterationTimeTrace: [30×1 double]
                         ErrorTrace: [30×1 double]
                  FeasibilityTrace: [30×1 logical]
        FeasibilityProbabilityTrace: [30×1 double]
                 IndexOfMinimumTrace: [30×1 double]
               ObjectiveMinimumTrace: [30×1 double]
    EstimatedObjectiveMinimumTrace: [30×1 double]
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: table

ResponseVarName — Response variable name
name of variable in Tbl

Response variable name, specified as the name of a variable in Tbl.

You must specify ResponseVarName as a character vector or string scalar. For example, if the response variable Y is stored as Tbl.Y, then specify it as 'Y'. Otherwise, the software treats all columns of Tbl, including Y, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If Y is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the ClassNames name-value argument.

Data Types: char | string

formula — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form 'Y~x1+x2+x3'. In this form, Y represents the response variable, and x1, x2, and x3 represent the predictor variables.

To specify a subset of variables in Tbl as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in Tbl that do not appear in formula.

The variable names in the formula must be both variable names in Tbl (Tbl.Properties.VariableNames) and valid MATLAB identifiers. You can verify the variable names in Tbl by using the isvarname function. If the variable names are not valid, then you can convert them by using the matlab.lang.makeValidName function.

Data Types: char | string

Y — Class labels
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class labels, specified as a numeric vector, categorical vector, logical vector, character array, string array, or cell array of character vectors. Each row of Y represents the classification of the corresponding row of X.

When fitting the tree, fitctree considers NaN, '' (empty character vector), "" (empty string), <missing>, and <undefined> values in Y to be missing values. fitctree does not use observations with missing values for Y in the fit.

For numeric Y, consider fitting a regression tree using fitrtree instead.

Data Types: single | double | categorical | logical | char | string | cell

X — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

`fitctree` considers `NaN` values in `X` as missing values. `fitctree` does not use observations with all missing values for `X` in the fit. `fitctree` uses observations with some missing values for `X` to find splits on variables for which these observations have valid values.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'CrossVal','on','MinLeafSize',40` specifies a cross-validated classification tree with a minimum of 40 observations per leaf.

Model Parameters

`AlgorithmForCategorical` — Algorithm for best categorical predictor split
`'Exact'` | `'PullLeft'` | `'PCA'` | `'OVAbyClass'`

Algorithm to find the best split on a categorical predictor with C categories for data and $K \geq 3$ classes, specified as the comma-separated pair consisting of `'AlgorithmForCategorical'` and one of the following values.

Value	Description
`'Exact'`	Consider all $2^{C-1} - 1$ combinations.
`'PullLeft'`	Start with all C categories on the right branch. Consider moving each category to the left branch as it achieves the minimum impurity for the K classes among the remaining categories. From this sequence, choose the split that has the lowest impurity.
`'PCA'`	Compute a score for each category using the inner product between the first principal component of a weighted covariance matrix (of the centered class probability matrix) and the vector of class probabilities for that category. Sort the scores in ascending order, and consider all $C - 1$ splits.

Value	Description
`'OVAbyClass'`	Start with all C categories on the right branch. For each class, order the categories based on their probability for that class. For the first class, consider moving each category to the left branch in order, recording the impurity criterion at each move. Repeat for the remaining classes. From this sequence, choose the split that has the minimum impurity.

`fitctree` automatically selects the optimal subset of algorithms for each split using the known number of classes and levels of a categorical predictor. For $K = 2$ classes, `fitctree` always performs the exact search. To specify a particular algorithm, use the `'AlgorithmForCategorical'` name-value pair argument.

For more details, see “Splitting Categorical Predictors in Classification Trees” on page 19-25.

Example: `'AlgorithmForCategorical','PCA'`

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitctree` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitctree` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitctree` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Cost of misclassification
square matrix | structure

Cost of misclassification of a point, specified as the comma-separated pair consisting of `'Cost'` and one of the following:

- Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). To specify the class order for the corresponding rows and columns of `Cost`, also specify the `ClassNames` name-value pair argument.
- Structure `S` having two fields: `S.ClassNames` containing the group names as a variable of the same data type as `Y`, and `S.ClassificationCosts` containing the cost matrix.

The default is `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`.

Data Types: `single` | `double` | `struct`

`MaxDepth` — Maximum tree depth
positive integer

Maximum tree depth, specified as the comma-separated pair consisting of `'MaxDepth'` and a positive integer. Specify a value for this argument to return a tree that has fewer levels and requires fewer passes through the tall array to compute. Generally, the algorithm of `fitctree` takes one pass through the data and an additional pass for each tree level. The function does not set a maximum tree depth, by default.

Note This option applies only when you use `fitctree` on tall arrays. See Tall Arrays on page 33-1931 for more information.

`MaxNumCategories` — Maximum category levels
`10` (default) | nonnegative scalar value

Maximum category levels, specified as the comma-separated pair consisting of `'MaxNumCategories'` and a nonnegative scalar value. `fitctree` splits a categorical predictor using the exact search algorithm if the predictor has at most `MaxNumCategories` levels in the split node. Otherwise, `fitctree` finds the best categorical split using one of the inexact algorithms.

Passing a small value can lead to loss of accuracy and passing a large value can increase computation time and memory overload.

Example: `'MaxNumCategories',8`

`MaxNumSplits` — Maximal number of decision splits
`size(X,1) - 1` (default) | positive integer

Maximal number of decision splits (or branch nodes), specified as the comma-separated pair consisting of `'MaxNumSplits'` and a positive integer. `fitctree` splits `MaxNumSplits` or fewer branch nodes. For more details on splitting behavior, see Algorithms on page 33-1928.

Example: `'MaxNumSplits',5`

Data Types: `single` | `double`

`MergeLeaves` — Leaf merge flag
`'on'` (default) | `'off'`

Leaf merge flag, specified as the comma-separated pair consisting of `'MergeLeaves'` and `'on'` or `'off'`.

If `MergeLeaves` is `'on'`, then `fitctree`:

- Merges leaves that originate from the same parent node, and that yields a sum of risk values greater or equal to the risk associated with the parent node
- Estimates the optimal sequence of pruned subtrees, but does not prune the classification tree

Otherwise, `fitctree` does not merge leaves.

Example: `'MergeLeaves','off'`

`MinLeafSize` — Minimum number of leaf node observations
`1` (default) | positive integer value

Minimum number of leaf node observations, specified as the comma-separated pair consisting of `'MinLeafSize'` and a positive integer value. Each leaf has at least `MinLeafSize` observations per tree leaf. If you supply both `MinParentSize` and `MinLeafSize`, `fitctree` uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

Example: `'MinLeafSize',3`

Data Types: `single` | `double`

`MinParentSize` — Minimum number of branch node observations
`10` (default) | positive integer value

Minimum number of branch node observations, specified as the comma-separated pair consisting of `'MinParentSize'` and a positive integer value. Each branch node in the tree has at least `MinParentSize` observations. If you supply both `MinParentSize` and `MinLeafSize`, `fitctree` uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

Example: `'MinParentSize',8`

Data Types: `single` | `double`

`NumBins` — Number of bins for numeric predictors
`[]` (empty) (default) | positive integer scalar

Number of bins for numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

- If the `'NumBins'` value is empty (default), then `fitctree` does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitctree` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
  - `fitctree` does not bin categorical predictors.

When you use a large training data set, this binning option speeds up training but might cause a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the value depending on the accuracy and training speed.

A trained model stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`NumVariablesToSample` — Number of predictors to select at random for each split
`'all'` (default) | positive integer value

Number of predictors to select at random for each split, specified as the comma-separated pair consisting of `'NumVariablesToSample'` and a positive integer value. Alternatively, you can specify `'all'` to use all available predictors.

If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` as `'all'`. Otherwise, the software might not select some predictors, underestimating their importance.

To reproduce the random selections, you must set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

Example: `'NumVariablesToSample',3`

Data Types: `char` | `string` | `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitctree` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`PredictorSelection` — Algorithm used to select the best split predictor
`'allsplits'` (default) | `'curvature'` | `'interaction-curvature'`

Algorithm used to select the best split predictor at each node, specified as the comma-separated pair consisting of `'PredictorSelection'` and a value in this table.

Value	Description
`'allsplits'`	Standard CART — Selects the split predictor that maximizes the split-criterion gain over all possible splits of all predictors [1].
`'curvature'`	Curvature test on page 33-1925 — Selects the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response [4]. Training speed is similar to standard CART.
`'interaction-curvature'`	Interaction test on page 33-1927 — Chooses the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response, and that minimizes the *p*-value of a chi-square test of independence between each pair of predictors and response [3]. Training speed can be slower than standard CART.

For `'curvature'` and `'interaction-curvature'`, if all tests yield *p*-values greater than 0.05, then `fitctree` stops splitting nodes.

Tip

- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [4]. Consider specifying the curvature or interaction test if any of the following are true:
  - If there are predictors that have relatively fewer distinct values than other predictors, for example, if the predictor data set is heterogeneous.
  - If an analysis of predictor importance is your goal. For more on predictor importance estimation, see `predictorImportance` and "Introduction to Feature Selection" on page 15-49.
- Trees grown using standard CART are not sensitive to predictor variable interactions. Also, such trees are less likely to identify important variables in the presence of many irrelevant predictors than the application of the interaction test. Therefore, to account for predictor interactions and identify importance variables in the presence of many irrelevant variables, specify the interaction test [3].
- Prediction speed is unaffected by the value of `'PredictorSelection'`.

For details on how `fitctree` selects split predictors, see "Node Splitting Rules" on page 33-1928 and "Choose Split Predictor Selection Technique" on page 19-14.

Example: `'PredictorSelection','curvature'`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as one of the following:

- Character vector or string scalar.
  - `'empirical'` determines class probabilities from class frequencies in the response variable in `Y` or `Tbl`. If you pass observation weights, `fitctree` uses the weights to compute the class probabilities.
  - `'uniform'` sets all class probabilities to be equal.
- Vector (one scalar value for each class). To specify the class order for the corresponding elements of `'Prior'`, set the `'ClassNames'` name-value argument.
- Structure `S` with two fields.
  - `S.ClassNames` contains the class names as a variable of the same type as the response variable in `Y` or `Tbl`.
  - `S.ClassProbs` contains a vector of corresponding probabilities.

`fitctree` normalizes the weights in each class (`'Weights'`) to add up to the value of the prior probability of the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`Prune` — Flag to estimate optimal sequence of pruned subtrees
`'on'` (default) | `'off'`

Flag to estimate the optimal sequence of pruned subtrees, specified as the comma-separated pair consisting of `'Prune'` and `'on'` or `'off'`.

If `Prune` is `'on'`, then `fitctree` grows the classification tree without pruning it, but estimates the optimal sequence of pruned subtrees. Otherwise, `fitctree` grows the classification tree without estimating the optimal sequence of pruned subtrees.

To prune a trained `ClassificationTree` model, pass it to `prune`.

Example: `'Prune','off'`

`PruneCriterion` — Pruning criterion
`'error'` (default) | `'impurity'`

Pruning criterion, specified as the comma-separated pair consisting of `'PruneCriterion'` and `'error'` or `'impurity'`.

If you specify `'impurity'`, then `fitctree` uses the impurity measure specified by the `'SplitCriterion'` name-value pair argument.

For details, see “Impurity and Node Error” on page 33-1926.

Example: `'PruneCriterion','impurity'`

`Reproducible` — Flag to enforce reproducibility
`false` (logical `0`) (default) | `true` (logical `1`)

Flag to enforce reproducibility over repeated runs of training a model, specified as the comma-separated pair consisting of `'Reproducible'` and either `false` or `true`.

If `'NumVariablesToSample'` is not `'all'`, then the software selects predictors at random for each split. To reproduce the random selections, you must specify `'Reproducible',true` and set the seed of the random number generator by using `rng`. Note that setting `'Reproducible'` to `true` can slow down training.

Example: `'Reproducible',true`

Data Types: `logical`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as the comma-separated pair consisting of `'ResponseName'` and a character vector or string scalar representing the name of the response variable.

This name-value pair is not valid when using the `ResponseVarName` or `formula` input arguments.

Example: `'ResponseName','IrisType'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation
`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$

Value	Description
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`SplitCriterion` — Split criterion
`'gdi'` (default) | `'twoing'` | `'deviance'`

Split criterion, specified as the comma-separated pair consisting of `'SplitCriterion'` and `'gdi'` (Gini's diversity index), `'twoing'` for the twoing rule, or `'deviance'` for maximum deviance reduction (also known as cross entropy).

For details, see “Impurity and Node Error” on page 33-1926.

Example: `'SplitCriterion','deviance'`

`Surrogate` — Surrogate decision splits flag
`'off'` (default) | `'on'` | `'all'` | positive integer value

Surrogate decision splits on page 33-1928 flag, specified as the comma-separated pair consisting of `'Surrogate'` and `'on'`, `'off'`, `'all'`, or a positive integer value.

- When set to `'on'`, `fitctree` finds at most 10 surrogate splits at each branch node.
- When set to `'all'`, `fitctree` finds all surrogate splits at each branch node. The `'all'` setting can use considerable time and memory.
- When set to a positive integer value, `fitctree` finds at most the specified number of surrogate splits at each branch node.

Use surrogate splits to improve the accuracy of predictions for data with missing values. The setting also lets you compute measures of predictive association between predictors. For more details, see “Node Splitting Rules” on page 33-1928.

Example: `'Surrogate','on'`

Data Types: `single` | `double` | `char` | `string`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors when training the model.

`fitctree` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Flag to grow cross-validated decision tree

`'off'` (default) | `'on'`

Flag to grow a cross-validated decision tree, specified as the comma-separated pair consisting of `'CrossVal'` and `'on'` or `'off'`.

If `'on'`, `fitctree` grows a cross-validated decision tree with 10 folds. You can override this cross-validation setting using one of the `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'` name-value pair arguments. You can only use one of these four arguments at a time when creating a cross-validated tree.

Alternatively, cross-validate `tree` later using the `crossval` method.

Example: `'CrossVal','on'`

`CVPartition` — Partition for cross-validated tree

`cvpartition` object

Partition to use in a cross-validated tree, specified as the comma-separated pair consisting of `'CVPartition'` and an object created using `cvpartition`.

If you use `'CVPartition'`, you cannot use any of the `'KFold'`, `'Holdout'`, or `'Leaveout'` name-value pair arguments.

`Holdout` — Fraction of data for holdout validation

`0` (default) | scalar value in the range `[0,1]`

Fraction of data used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range `[0,1]`. Holdout validation tests the specified fraction of the data, and uses the rest of the data for training.

If you use `'Holdout'`, you cannot use any of the `'CVPartition'`, `'KFold'`, or `'Leaveout'` name-value pair arguments.

Example: `'Holdout',0.1`

Data Types: `single` | `double`

KFold — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated classifier, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1. If you specify, e.g., `'KFold',k`, then the software:

1. Randomly partitions the data into k sets
2. For each set, reserves the set as validation data, and trains the model using the other k – 1 sets
3. Stores the k compact, trained models in the cells of a k-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',8`

Data Types: `single` | `double`

Leaveout — Leave-one-out cross-validation flag

`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. Specify `'on'` to use leave-one-out cross-validation.

If you use `'Leaveout'`, you cannot use any of the `'CVPartition'`, `'Holdout'`, or `'KFold'` name-value pair arguments.

Example: `'Leaveout','on'`

Hyperparameter Optimization Options

OptimizeHyperparameters — Parameters to optimize

`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'MinLeafSize'}`
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`

The optimization attempts to minimize the cross-validation loss (error) for `fitctree` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitctree` are:

- `MaxNumSplits` — `fitctree` searches among integers, by default log-scaled in the range `[1,max(2,NumObservations-1)]`.
- `MinLeafSize` — `fitctree` searches among integers, by default log-scaled in the range `[1,max(2,floor(NumObservations/2))]`.
- `SplitCriterion` — For two classes, `fitctree` searches among `'gdi'` and `'deviance'`. For three or more classes, `fitctree` also searches among `'twoing'`.
- `NumVariablesToSample` — `fitctree` does not optimize over this hyperparameter. If you pass `NumVariablesToSample` as a parameter name, `fitctree` simply uses the full number of predictors. However, `fitcensemble` does optimize over this hyperparameter.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitctree',meas,species);
params(1).Range = [1,30];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Classification Tree” on page 33-1874.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`tree` — Classification tree

classification tree object

Classification tree, returned as a classification tree object.

Using the `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'` options results in a tree of class `ClassificationPartitionedModel`. You cannot use a partitioned tree for prediction, so this kind of tree does not have a `predict` method. Instead, use `kfoldPredict` to predict responses for observations not used for training.

Otherwise, `tree` is of class `ClassificationTree`, and you can use the `predict` method to make predictions.

More About

Curvature Test

The curvature test is a statistical test assessing the null hypothesis that two variables are unassociated.

The curvature test between predictor variable x and y is conducted using this process.

1. If x is continuous, then partition it into its quartiles. Create a nominal variable that bins observations according to which section of the partition they occupy. If there are missing values, then create an extra bin for them.
2. For each level in the partitioned predictor $j = 1..,J$ and class in the response $k = 1,...,K$, compute the weighted proportion of observations in class k

   $$\widehat{\pi}_{jk} = \sum_{i=1}^{n} I\{y_i = k\} w_i .$$

   w_i is the weight of observation i, $\sum w_i = 1$, I is the indicator function, and n is the sample size. If all observations have the same weight, then $\widehat{\pi}_{jk} = \frac{n_{jk}}{n}$, where n_{jk} is the number of observations in level j of the predictor that are in class k.
3. Compute the test statistic

   $$t = n \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(\widehat{\pi}_{jk} - \widehat{\pi}_{j+}\widehat{\pi}_{+k}\right)^2}{\widehat{\pi}_{j+}\widehat{\pi}_{+k}}$$

   $\widehat{\pi}_{j+} = \sum_{k} \widehat{\pi}_{jk}$, that is, the marginal probability of observing the predictor at level j. $\widehat{\pi}_{+k} = \sum_{j} \widehat{\pi}_{jk}$, that is the marginal probability of observing class k. If n is large enough, then t is distributed as a χ^2 with $(K - 1)(J - 1)$ degrees of freedom.
4. If the p-value for the test is less than 0.05, then reject the null hypothesis that there is no association between x and y.

When determining the best split predictor at each node, the standard CART algorithm prefers to select continuous predictors that have many levels. Sometimes, such a selection can be spurious and can also mask more important predictors that have fewer levels, such as categorical predictors.

The curvature test can be applied instead of standard CART to determine the best split predictor at each node. In that case, the best split predictor variable is the one that minimizes the significant *p*-values (those less than 0.05) of curvature tests between each predictor and the response variable. Such a selection is robust to the number of levels in individual predictors.

Note If levels of a predictor are pure for a particular class, then `fitctree` merges those levels. Therefore, in step 3 of the algorithm, J can be less than the actual number of levels in the predictor. For example, if x has 4 levels, and all observations in bins 1 and 2 belong to class 1, then those levels are pure for class 1. Consequently, `fitctree` merges the observations in bins 1 and 2, and J reduces to 3.

For more details on how the curvature test applies to growing classification trees, see “Node Splitting Rules” on page 33-1928 and [4].

Impurity and Node Error

A decision tree splits nodes based on either impurity or node error.

Impurity means one of several things, depending on your choice of the `SplitCriterion` name-value pair argument:

- Gini's Diversity Index (`gdi`) — The Gini index of a node is

  $$1 - \sum_i p^2(i),$$

  where the sum is over the classes i at the node, and $p(i)$ is the observed fraction of classes with class i that reach the node. A node with just one class (a pure node) has Gini index `0`; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.
- Deviance (`'deviance'`) — With $p(i)$ defined the same as for the Gini index, the deviance of a node is

  $$-\sum_i p(i)\log_2 p(i)\,.$$

  A pure node has deviance `0`; otherwise, the deviance is positive.
- Twoing rule (`'twoing'`) — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class i in the left child node after a split, and $R(i)$ denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

  $$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2,$$

  where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.
- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $1 – p(j)$.

Interaction Test

The interaction test is a statistical test that assesses the null hypothesis that there is no interaction between a pair of predictor variables and the response variable.

The interaction test assessing the association between predictor variables x_1 and x_2 with respect to y is conducted using this process.

1. If x_1 or x_2 is continuous, then partition that variable into its quartiles. Create a nominal variable that bins observations according to which section of the partition they occupy. If there are missing values, then create an extra bin for them.
2. Create the nominal variable z with $J = J_1J_2$ levels that assigns an index to observation i according to which levels of x_1 and x_2 it belongs. Remove any levels of z that do not correspond to any observations.
3. Conduct a curvature test on page 33-1925 between z and y.

When growing decision trees, if there are important interactions between pairs of predictors, but there are also many other less important predictors in the data, then standard CART tends to miss the important interactions. However, conducting curvature and interaction tests for predictor selection instead can improve detection of important interactions, which can yield more accurate decision trees.

For more details on how the interaction test applies to growing decision trees, see “Curvature Test” on page 33-1925, “Node Splitting Rules” on page 33-1928 and [3].

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_jL_k} - P_{R_jR_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_jL_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in $(-\infty,1]$. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Surrogate Decision Splits

A surrogate decision split is an alternative to the optimal decision split at a given node in a decision tree. The optimal split is found by growing the tree; the surrogate split uses a similar or correlated predictor variable and split criterion.

When the value of the optimal split predictor for an observation is missing, the observation is sent to the left or right child node using the best surrogate predictor. When the value of the best surrogate split predictor for the observation is also missing, the observation is sent to the left or right child node using the second-best surrogate predictor, and so on. Candidate splits are sorted in descending order by their predictive measure of association on page 33-2412.

Tip

- By default, `Prune` is `'on'`. However, this specification does not prune the classification tree. To prune a trained classification tree, pass the classification tree to `prune`.
- After training a model, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

Node Splitting Rules

`fitctree` uses these processes to determine how to split node t.

- For standard CART (that is, if `PredictorSelection` is `'allpairs'`) and for all predictors x_i, $i = 1,\ldots,p$:

  1. `fitctree` computes the weighted impurity of node t, i_t. For supported impurity measures, see `SplitCriterion`.
  2. `fitctree` estimates the probability that an observation is in node t using

     $$P(T) = \sum_{j \in T} w_j .$$

     w_j is the weight of observation j, and T is the set of all observation indices in node t. If you do not specify `Prior` or `Weights`, then $w_j = 1/n$, where n is the sample size.
  3. `fitctree` sorts x_i in ascending order. Each element of the sorted predictor is a splitting candidate or cut point. `fitctree` stores any indices corresponding to missing values in the set T_U, which is the unsplit set.
  4. `fitctree` determines the best way to split node t using x_i by maximizing the impurity gain (ΔI) over all splitting candidates. That is, for all splitting candidates in x_i:

     a. `fitctree` splits the observations in node t into left and right child nodes (t_L and t_R, respectively).

     b. `fitctree` computes ΔI. Suppose that for a particular splitting candidate, t_L and t_R contain observation indices in the sets T_L and T_R, respectively.

        - If x_i does not contain any missing values, then the impurity gain for the current splitting candidate is

$\Delta I = P(T)i_t - P(T_L)i_{t_L} - P(T_R)i_{t_R}$.

- If x_i contains missing values then, assuming that the observations are missing at random, the impurity gain is

$\Delta I_U = P(T - T_U)i_t - P(T_L)i_{t_L} - P(T_R)i_{t_R}$.

$T - T_U$ is the set of all observation indices in node t that are not missing.

- If you use surrogate decision splits on page 33-1928, then:

  i `fitctree` computes the predictive measures of association on page 33-1927 between the decision split $x_j < u$ and all possible decision splits $x_k < v, j \neq k$.

  ii `fitctree` sorts the possible alternative decision splits in descending order by their predictive measure of association with the optimal split. The surrogate split is the decision split yielding the largest measure.

  iii `fitctree` decides the child node assignments for observations with a missing value for x_i using the surrogate split. If the surrogate predictor also contains a missing value, then `fitctree` uses the decision split with the second largest measure, and so on, until there are no other surrogates. It is possible for `fitctree` to split two different observations at node t using two different surrogate splits. For example, suppose the predictors x_1 and x_2 are the best and second best surrogates, respectively, for the predictor x_i, $i \notin \{1,2\}$, at node t. If observation m of predictor x_i is missing (i.e., x_{mi} is missing), but x_{m1} is not missing, then x_1 is the surrogate predictor for observation x_{mi}. If observations $x_{(m+1),i}$ and $x(m+1),1$ are missing, but $x_{(m+1),2}$ is not missing, then x_2 is the surrogate predictor for observation $m + 1$.

  iv `fitctree` uses the appropriate impurity gain formula. That is, if `fitctree` fails to assign all missing observations in node t to children nodes using surrogate splits, then the impurity gain is ΔI_U. Otherwise, `fitctree` uses ΔI for the impurity gain.

c `fitctree` chooses the candidate that yields the largest impurity gain.

`fitctree` splits the predictor variable at the cut point that maximizes the impurity gain.

- For the curvature test (that is, if `PredictorSelection` is `'curvature'`):

  1 `fitctree` conducts curvature tests on page 33-1925 between each predictor and the response for observations in node t.

  - If all p-values are at least 0.05, then `fitctree` does not split node t.
  - If there is a minimal p-value, then `fitctree` chooses the corresponding predictor to split node t.
  - If more than one p-value is zero due to underflow, then `fitctree` applies standard CART to the corresponding predictors to choose the split predictor.

  2 If `fitctree` chooses a split predictor, then it uses standard CART to choose the cut point (see step 4 in the standard CART process).

- For the interaction test (that is, if `PredictorSelection` is `'interaction-curvature'`):

  1 For observations in node t, `fitctree` conducts curvature tests on page 33-1925 between each predictor and the response and interaction tests on page 33-1927 between each pair of predictors and the response.

- If all *p*-values are at least 0.05, then `fitctree` does not split node *t*.
- If there is a minimal *p*-value and it is the result of a curvature test, then `fitctree` chooses the corresponding predictor to split node *t*.
- If there is a minimal *p*-value and it is the result of an interaction test, then `fitctree` chooses the split predictor using standard CART on the corresponding pair of predictors.
- If more than one *p*-value is zero due to underflow, then `fitctree` applies standard CART to the corresponding predictors to choose the split predictor.

2. If `fitctree` chooses a split predictor, then it uses standard CART to choose the cut point (see step 4 in the standard CART process).

Tree Depth Control

- If `MergeLeaves` is `'on'` and `PruneCriterion` is `'error'` (which are the default values for these name-value pair arguments), then the software applies pruning only to the leaves and by using classification error. This specification amounts to merging leaves that share the most popular class per leaf.
- To accommodate `MaxNumSplits`, `fitctree` splits all nodes in the current layer, and then counts the number of branch nodes. A layer is the set of nodes that are equidistant from the root node. If the number of branch nodes exceeds `MaxNumSplits`, `fitctree` follows this procedure:

  1. Determine how many branch nodes in the current layer must be unsplit so that there are at most `MaxNumSplits` branch nodes.
  2. Sort the branch nodes by their impurity gains.
  3. Unsplit the number of least successful branches.
  4. Return the decision tree grown so far.

  This procedure produces maximally balanced trees.
- The software splits branch nodes layer by layer until at least one of these events occurs:
  - There are `MaxNumSplits` branch nodes.
  - A proposed split causes the number of observations in at least one branch node to be fewer than `MinParentSize`.
  - A proposed split causes the number of observations in at least one leaf node to be fewer than `MinLeafSize`.
  - The algorithm cannot find a good split within a layer (i.e., the pruning criterion (see `PruneCriterion`), does not improve for all proposed splits in a layer). A special case is when all nodes are pure (i.e., all observations in the node have the same class).
  - For values `'curvature'` or `'interaction-curvature'` of `PredictorSelection`, all tests yield *p*-values greater than 0.05.

  `MaxNumSplits` and `MinLeafSize` do not affect splitting at their default values. Therefore, if you set `'MaxNumSplits'`, splitting might stop due to the value of `MinParentSize`, before `MaxNumSplits` splits occur.

Parallelization

For dual-core systems and above, `fitctree` parallelizes training decision trees using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[2] Coppersmith, D., S. J. Hong, and J. R. M. Hosking. “Partitioning Nominal Attributes in Decision Trees.” *Data Mining and Knowledge Discovery*, Vol. 3, 1999, pp. 197–217.

[3] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[4] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Supported syntaxes are:
  - `tree = fitctree(Tbl,Y)`
  - `tree = fitctree(X,Y)`
  - `tree = fitctree(___,Name,Value)`
  - `[tree,FitInfo,HyperparameterOptimizationResults] = fitctree(___,Name,Value)` — `fitctree` returns the additional output arguments `FitInfo` and `HyperparameterOptimizationResults` when you specify the `'OptimizeHyperparameters'` name-value pair argument.
- The `FitInfo` output argument is an empty structure array currently reserved for possible future use.
- The `HyperparameterOptimizationResults` output argument is a `BayesianOptimization` object or a table of hyperparameters with associated values that describe the cross-validation optimization of hyperparameters.

  `'HyperparameterOptimizationResults'` is nonempty when the `'OptimizeHyperparameters'` name-value pair argument is nonempty at the time you create the model. The values in `'HyperparameterOptimizationResults'` depend on the value you specify for the `'HyperparameterOptimizationOptions'` name-value pair argument when you create the model.
  - If you specify `'bayesopt'` (default), then `HyperparameterOptimizationResults` is an object of class `BayesianOptimization`.
  - If you specify `'gridsearch'` or `'randomsearch'`, then `HyperparameterOptimizationResults` is a table of the hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst).
- Supported name-value pair arguments, and any differences, are:
  - `'AlgorithmForCategorical'`

- 'CategoricalPredictors'
- 'ClassNames'
- 'Cost'
- 'HyperparameterOptimizationOptions' — For cross-validation, tall optimization supports only 'Holdout' validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify 'HyperparameterOptimizationOptions',struct('Holdout',0.3) to reserve 30% of the data as validation data.
- 'MaxNumCategories'
- 'MaxNumSplits'— for tall optimization, fitctree searches among integers, by default log-scaled in the range [1,max(2,min(10000,NumObservations-1))].
- 'MergeLeaves'
- 'MinLeafSize'
- 'MinParentSize'
- 'NumVariablesToSample'
- 'OptimizeHyperparameters'
- 'PredictorNames'
- 'Prior'
- 'ResponseName'
- 'ScoreTransform'
- 'SplitCriterion'
- 'Weights'

- This additional name-value pair argument is specific to tall arrays:
  - 'MaxDepth' — A positive integer specifying the maximum depth of the output tree. Specify a value for this argument to return a tree that has fewer levels and requires fewer passes through the tall array to compute. Generally, the algorithm of fitctree takes one pass through the data and an additional pass for each tree level. The function does not set a maximum tree depth, by default.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the 'HyperparameterOptimizationOptions', struct('UseParallel',true) name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `ClassificationTree` | `kfoldPredict` | `predict` | `prune`

Topics

“Splitting Categorical Predictors in Classification Trees” on page 19-25

Introduced in R2014a

fitglm

Create generalized linear regression model

Syntax

```
mdl = fitglm(tbl)
mdl = fitglm(X,y)
mdl = fitglm(___,modelspec)
mdl = fitglm(___,Name,Value)
```

Description

`mdl = fitglm(tbl)` returns a generalized linear model fit to variables in the table or dataset array `tbl`. By default, `fitglm` takes the last variable as the response variable.

`mdl = fitglm(X,y)` returns a generalized linear model of the responses `y`, fit to the data matrix `X`.

`mdl = fitglm(___,modelspec)` returns a generalized linear model of the type you specify in `modelspec`.

`mdl = fitglm(___,Name,Value)` returns a generalized linear model with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify which variables are categorical, the distribution of the response variable, and the link function to use.

Examples

Fit a Logistic Regression Model

Make a logistic binomial model of the probability of smoking as a function of age, weight, and sex, using a two-way interactions model.

Load the `hospital` dataset array.

```
load hospital
dsa = hospital;
```

Specify the model using a formula that allows up to two-way interactions between the variables age, weight, and sex. Smoker is the response variable.

```
modelspec = 'Smoker ~ Age*Weight*Sex - Age:Weight:Sex';
```

Fit a logistic binomial model.

```
mdl = fitglm(dsa,modelspec,'Distribution','binomial')
```

```
mdl =
Generalized linear regression model:
    logit(Smoker) ~ 1 + Sex*Age + Sex*Weight + Age*Weight
```

```
    Distribution = Binomial

Estimated Coefficients:
                        Estimate         SE         tStat      pValue
                       ___________    _________    ________    _______

    (Intercept)            -6.0492       19.749     -0.3063    0.75938
    Sex_Male               -2.2859       12.424    -0.18399    0.85402
    Age                    0.11691      0.50977     0.22934    0.81861
    Weight                0.031109      0.15208     0.20455    0.83792
    Sex_Male:Age          0.020734      0.20681     0.10025    0.92014
    Sex_Male:Weight        0.01216     0.053168     0.22871     0.8191
    Age:Weight         -0.00071959    0.0038964    -0.18468    0.85348


100 observations, 93 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 5.07, p-value = 0.535
```

All of the p-values (under `pValue`) are large. This means none of the coefficients are significant. The large *p*-value for the test of the model, 0.535, indicates that this model might not differ statistically from a constant model.

GLM for Poisson Response

Create sample data with 20 predictors, and Poisson response using just three of the predictors, plus a constant.

```
rng('default') % for reproducibility
X = randn(100,7);
mu = exp(X(:,[1 3 6])*[.4;.2;.3] + 1);
y = poissrnd(mu);
```

Fit a generalized linear model using the Poisson distribution.

```
mdl =  fitglm(X,y,'linear','Distribution','poisson')

mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7
    Distribution = Poisson

Estimated Coefficients:
                    Estimate        SE        tStat        pValue
                    _________    ________    ________    __________

    (Intercept)       0.88723    0.070969      12.502    7.3149e-36
    x1                0.44413    0.052337      8.4858    2.1416e-17
    x2              0.0083388    0.056527     0.14752       0.88272
    x3                0.21518    0.063416      3.3932    0.00069087
    x4               -0.058386    0.065503    -0.89135       0.37274
    x5               -0.060824    0.073441     -0.8282       0.40756
    x6                0.34267    0.056778      6.0352    1.5878e-09
    x7                0.04316     0.06146     0.70225       0.48252
```

```
100 observations, 92 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 119, p-value = 1.55e-22
```

The p-values of 2.14e-17, 0.00069, and 1.58e-09 indicate that the coefficients of the variables `x1`, `x3`, and `x6` are statistically significant.

Input Arguments

`tbl` — Input data

table | dataset array

Input data including predictor and response variables, specified as a table or dataset array. The predictor variables and response variable can be numeric, logical, categorical, character, or string. The response variable can have a data type other than numeric only if `'Distribution'` is `'binomial'`.

- By default, `fitglm` takes the last variable as the response variable and the others as the predictor variables.
- To set a different column as the response variable, use the `ResponseVar` name-value pair argument.
- To use a subset of the columns as predictors, use the `PredictorVars` name-value pair argument.
- To define a model specification, set the `modelspec` argument using a formula or terms matrix. The formula or terms matrix specifies which columns to use as the predictor or response variables.

The variable names in a table do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot specify `modelspec` using a formula.
- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiseglm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

You can verify the variable names in `tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

`X` — Predictor variables

matrix

Predictor variables, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

By default, there is a constant term in the model, unless you explicitly remove it, so do not include a column of 1s in `X`.

Data Types: `single` | `double`

y — Response variable

vector | matrix

Response variable, specified as a vector or matrix.

- If `'Distribution'` is not `'binomial'`, then `y` must be an *n*-by-1 vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`. The data type must be single or double.
- If `'Distribution'` is `'binomial'`, then `y` can be an *n*-by-1 vector or *n*-by-2 matrix with counts in column 1 and `BinomialSize` in column 2.

Data Types: `single` | `double` | `logical` | `categorical`

modelspec — Model specification

`'linear'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'y ~ terms'`

Model specification, specified as one of these values.

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and x_1*x_2^2 terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-1943, specifying terms in the model, where *t* is the number of terms and *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar “Formula” on page 33-1943 in the form `'y ~ terms'`,

where the `terms` are in “Wilkinson Notation” on page 33-1944. The variable names in the formula must be variable names in `tbl` or variable names specified by `Varnames`. Also, the variable names must be valid MATLAB identifiers.

The software determines the order of terms in a fitted model by using the order of terms in `tbl` or `X`. Therefore, the order of terms in the model can be different from the order of terms in the specified formula.

Example: `'quadratic'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distribution','normal','link','probit','Exclude',[23,59]` specifies that the distribution of the response is normal, and instructs `fitglm` to use the probit link function and exclude the 23rd and 59th observations from the fit.

`BinomialSize` — Number of trials for binomial distribution

1 (default) | numeric scalar | numeric vector | character vector | string scalar

Number of trials for binomial distribution, that is the sample size, specified as the comma-separated pair consisting of `'BinomialSize'` and the variable name in `tbl`, a numeric scalar, or a numeric vector of the same length as the response. This is the parameter `n` for the fitted binomial distribution. `BinomialSize` applies only when the `Distribution` parameter is `'binomial'`.

If `BinomialSize` is a scalar value, that means all observations have the same number of trials.

As an alternative to `BinomialSize`, you can specify the response as a two-column matrix with counts in column 1 and `BinomialSize` in column 2.

Data Types: `single` | `double` | `char` | `string`

`B0` — Initial values for coefficient estimates

numeric vector

Initial values for the coefficient estimates, specified as a numeric vector. The default values are initial fitted values derived from the input data.

Data Types: `single` | `double`

`CategoricalVars` — Categorical variable list

string array | cell array of character vectors | logical or numeric index vector

Categorical variable list, specified as the comma-separated pair consisting of `'CategoricalVars'` and either a string array or cell array of character vectors containing categorical variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are categorical.

- If data is in a table or dataset array `tbl`, then, by default, `fitglm` treats all categorical values, logical values, character arrays, string arrays, and cell arrays of character vectors as categorical variables.
- If data is in matrix `X`, then the default value of `'CategoricalVars'` is an empty matrix `[]`. That is, no variable is categorical unless you specify it as categorical.

For example, you can specify the second and third variables out of six as categorical using either of the following:

Example: `'CategoricalVars',[2,3]`

Example: `'CategoricalVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`DispersionFlag` — Indicator to compute dispersion parameter
`false` for `'binomial'` and `'poisson'` distributions (default) | `true`

Indicator to compute dispersion parameter for `'binomial'` and `'poisson'` distributions, specified as the comma-separated pair consisting of `'DispersionFlag'` and one of the following.

`true`	Estimate a dispersion parameter when computing standard errors. The estimated dispersion parameter value is the sum of squared Pearson residuals divided by the degrees of freedom for error (DFE).
`false`	Default. Use the theoretical value of 1 when computing standard errors.

The fitting function always estimates the dispersion for other distributions.

Example: `'DispersionFlag',true`

`Distribution` — Distribution of the response variable
`'normal'` (default) | `'binomial'` | `'poisson'` | `'gamma'` | `'inverse gaussian'`

Distribution of the response variable, specified as the comma-separated pair consisting of `'Distribution'` and one of the following.

`'normal'`	Normal distribution
`'binomial'`	Binomial distribution
`'poisson'`	Poisson distribution
`'gamma'`	Gamma distribution
`'inverse gaussian'`	Inverse Gaussian distribution

Example: `'Distribution','gamma'`

`Exclude` — Observations to exclude
logical or numeric index vector

Observations to exclude from the fit, specified as the comma-separated pair consisting of `'Exclude'` and a logical or numeric index vector indicating which observations to exclude from the fit.

For example, you can exclude observations 2 and 3 out of 6 using either of the following examples.

Example: `'Exclude',[2,3]`

Example: `'Exclude',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical`

`Intercept` — Indicator for constant term
`true` (default) | `false`

Indicator for the constant term (intercept) in the fit, specified as the comma-separated pair consisting of `'Intercept'` and either `true` to include or `false` to remove the constant term from the model.

Use `'Intercept'` only when specifying the model using a character vector or string scalar, not a formula or matrix.

Example: `'Intercept',false`

`Link` — Link function
canonical link function (default) | scalar value | structure

Link function to use in place of the canonical link function, specified as the comma-separated pair consisting of `'Link'` and one of the following.

Link Function Name	Link Function	Mean (Inverse) Function
`'identity'`	$f(\mu) = \mu$	$\mu = Xb$
`'log'`	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
`'logit'`	$f(\mu) = \log(\mu/(1–\mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'probit'`	$f(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function of the standard normal distribution.	$\mu = \Phi(Xb)$
`'comploglog'`	$f(\mu) = \log(–\log(1 – \mu))$	$\mu = 1 – \exp(–\exp(Xb))$
`'reciprocal'`	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
`p` (a number)	$f(\mu) = \mu^p$	$\mu = Xb^{1/p}$
`S` (a structure) with three fields. Each field holds a function handle that accepts a vector of inputs and returns a vector of the same size: • `S.Link` — The link function • `S.Inverse` — The inverse link function • `S.Derivative` — The derivative of the link function	$f(\mu) =$ `S.Link`(μ)	$\mu =$ `S.Inverse`(Xb)

The link function defines the relationship $f(\mu) = X*b$ between the mean response μ and the linear combination of predictors $X*b$.

For more information on the canonical link functions, see “Canonical Link Function” on page 33-1944.

Example: `'Link','probit'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`Options` — Optimization options
`statset('fitglm')` (default) | structure

Optimization options, specified as a structure. This argument determines the control parameters for the iterative algorithm that `fitglm` uses.

Create the `'Options'` value by using the function `statset` or by creating a structure array containing the fields and values described in this table.

Field Name	Value	Default Value
`Display`	Amount of information displayed by the algorithm • `'off'` — Displays no information • `'final'` — Displays the final output	`'off'`
`MaxIter`	Maximum number of iterations allowed, specified as a positive integer	`100`
`TolX`	Termination tolerance for the parameters, specified as a positive scalar	`1e-6`

You can also enter `statset('fitglm')` in the Command Window to see the names and default values of the fields that `fitglm` accepts in the `'Options'` name-value argument.

Example: `'Options',statset('Display','final','MaxIter',1000)` specifies to display the final information of the iterative algorithm results, and change the maximum number of iterations allowed to 1000.

Data Types: `struct`

`Offset` — Offset variable

[ ] (default) | numeric vector | character vector | string scalar

Offset variable in the fit, specified as the comma-separated pair consisting of `'Offset'` and the variable name in `tbl` or a numeric vector with the same length as the response.

`fitglm` uses `Offset` as an additional predictor with a coefficient value fixed at 1. In other words, the formula for fitting is

$f(\mu)$ = Offset + $X*b$,

where f is the link function, μ is the mean response, and $X*b$ is the linear combination of predictors X. The `Offset` predictor has coefficient `1`.

For example, consider a Poisson regression model. Suppose the number of counts is known for theoretical reasons to be proportional to a predictor `A`. By using the log link function and by specifying `log(A)` as an offset, you can force the model to satisfy this theoretical constraint.

Data Types: `single` | `double` | `char` | `string`

`PredictorVars` — Predictor variables

string array | cell array of character vectors | logical or numeric index vector

Predictor variables to use in the fit, specified as the comma-separated pair consisting of `'PredictorVars'` and either a string array or cell array of character vectors of the variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are predictor variables.

The string values or character vectors should be among the names in `tbl`, or the names you specify using the `'VarNames'` name-value pair argument.

The default is all variables in X, or all variables in `tbl` except for `ResponseVar`.

For example, you can specify the second and third variables as the predictor variables using either of the following examples.

Example: `'PredictorVars',[2,3]`

Example: `'PredictorVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`ResponseVar` — Response variable
last column in `tbl` (default) | character vector or string scalar containing variable name | logical or numeric index vector

Response variable to use in the fit, specified as the comma-separated pair consisting of `'ResponseVar'` and either a character vector or string scalar containing the variable name in the table or dataset array `tbl`, or a logical or numeric index vector indicating which column is the response variable. You typically need to use `'ResponseVar'` when fitting a table or dataset array `tbl`.

For example, you can specify the fourth variable, say `yield`, as the response out of six variables, in one of the following ways.

Example: `'ResponseVar','yield'`

Example: `'ResponseVar',[4]`

Example: `'ResponseVar',logical([0 0 0 1 0 0])`

Data Types: `single` | `double` | `logical` | `char` | `string`

`VarNames` — Names of variables
`{'x1','x2',...,'xn','y'}` (default) | string array | cell array of character vectors

Names of variables, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors including the names for the columns of X first, and the name for the response variable `y` last.

`'VarNames'` is not applicable to variables in a table or dataset array, because those variables already have names.

The variable names do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiseglm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

Before specifying `'VarNames',varNames`, you can verify the variable names in `varNames` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Example: `'VarNames',{'Horsepower','Acceleration','Model_Year','MPG'}`

Data Types: `string` | `cell`

`Weights` — Observation weights
`ones(n,1)` (default) | *n*-by-1 vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations.

Data Types: `single` | `double`

Output Arguments

`mdl` — Generalized linear regression model
`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'`*`y ~ terms`*`'`.

- *`y`* is the response name.
- *`terms`* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.

- `'y ~ x1 + x2 + x3 − 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include −1 in the formula.

A formula includes a constant term unless you explicitly remove the term using −1.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.
- − means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `( )` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
−x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 − x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Canonical Link Function

The default link function for a generalized linear model is the canonical link function.

Distribution	Canonical Link Function Name	Link Function	Mean (Inverse) Function
`'normal'`	`'identity'`	$f(\mu) = \mu$	$\mu = Xb$
`'binomial'`	`'logit'`	$f(\mu) = \log(\mu/(1 - \mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'poisson'`	`'log'`	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$

Distribution	Canonical Link Function Name	Link Function	Mean (Inverse) Function
'gamma'	-1	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
'inverse gaussian'	-2	$f(\mu) = 1/\mu^2$	$\mu = (Xb)^{-1/2}$

Tips

- The generalized linear model `mdl` is a standard linear model unless you specify otherwise with the `Distribution` name-value pair.
- For methods such as `plotResiduals` or `devianceTest`, or properties of the `GeneralizedLinearModel` object, see `GeneralizedLinearModel`.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- `fitglm` treats a categorical predictor as follows:
  - A model with a categorical predictor that has L levels (categories) includes L – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `fitglm` treats the group of L – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.
- `fitglm` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `tbl`, `X`, and `Y` to be missing values. `fitglm` does not use observations with missing values in the fit. The `ObservationInfo` property of a fitted model indicates whether or not `fitglm` uses each observation in the fit.

Alternative Functionality

- Use `stepwiseglm` to select a model specification automatically. Use `step`, `addTerms`, or `removeTerms` to adjust a fitted model.

References

[1] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

[2] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[3] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- If any input argument to `fitglm` is a tall array, then all of the other inputs must be tall arrays as well. This includes nonempty variables supplied with the `'Weights'`, `'Exclude'`, `'Offset'`, and `'BinomialSize'` name-value pairs.
- The default number of iterations is 5. You can change the number of iterations using the `'Options'` name-value pair to pass in an options structure. Create an options structure using `statset` to specify a different value for `MaxIter`.
- For tall data, `fitglm` returns a `CompactGeneralizedLinearModel` object that contains most of the same properties as a `GeneralizedLinearModel` object. The main difference is that the compact object is sensitive to memory requirements. The compact object does not include properties that include the data, or that include an array of the same size as the data. The compact object does not contain these `GeneralizedLinearModel` properties:

  - `Diagnostics`
  - `Fitted`
  - `Offset`
  - `ObservationInfo`
  - `ObservationNames`
  - `Residuals`
  - `Steps`
  - `Variables`

  You can compute the residuals directly from the compact object returned by `GLM = fitglm(X,Y)` using

```
RES = Y - predict(GLM,X);
S = sqrt(GLM.SSE/GLM.DFE);
histogram(RES,linspace(-3*S,3*S,51))
```

For more information, see “Tall Arrays for Out-of-Memory Data”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `glmfit` | `predict` | `stepwiseglm`

Topics

“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2013b

fitglme

Fit generalized linear mixed-effects model

Syntax

```
glme = fitglme(tbl,formula)
glme = fitglme(tbl,formula,Name,Value)
```

Description

`glme = fitglme(tbl,formula)` returns a generalized linear mixed-effects model, `glme`. The model is specified by `formula` and fitted to the predictor variables in the table or dataset array, `tbl`.

`glme = fitglme(tbl,formula,Name,Value)` returns a generalized linear mixed-effects model using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the distribution of the response, the link function, or the covariance pattern of the random-effects terms.

Examples

Fit a Generalized Linear Mixed-Effects Model

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this

model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}\left(\mu_{ij}\right).$$

This corresponds to the generalized linear mixed-effects model

$$\log\left(\mu_{ij}\right) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)', ..
    'Distribution','Poisson','Link','log','FitMethod','Laplace', ...
    'DummyVarCoding','effects');
```

Display the model.

```
disp(glme)
```

```
Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35

Fixed effects coefficients (95% CIs):
```

```
    Name                   Estimate     SE          tStat       DF    pValue
    {'(Intercept)'}           1.4689     0.15988      9.1875    94    9.8194e-15
    {'newprocess' }          -0.36766    0.17755     -2.0708    94      0.041122
    {'time_dev'   }          -0.094521   0.82849     -0.11409   94       0.90941
    {'temp_dev'   }          -0.28317     0.9617     -0.29444   94       0.76907
    {'supplier_C' }          -0.071868   0.078024     -0.9211   94       0.35936
    {'supplier_B' }           0.071072    0.07739     0.91836   94       0.36078


    Lower        Upper
       1.1515       1.7864
    -0.72019    -0.015134
     -1.7395       1.5505
     -2.1926       1.6263
    -0.22679     0.083051
    -0.082588     0.22473

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.31381

Group: Error
    Name                        Estimate
    {'sqrt(Dispersion)'}        1
```

The `Model information` table displays the total number of observations in the sample data (100), the number of fixed- and random-effects coefficients (6 and 20, respectively), and the number of covariance parameters (1). It also indicates that the response variable has a `Poisson` distribution, the link function is `Log`, and the fit method is `Laplace`.

`Formula` indicates the model specification using Wilkinson's notation.

The `Model fit statistics` table displays statistics used to assess the goodness of fit of the model. This includes the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`) values, log likelihood (`LogLikelihood`), and deviance (`Deviance`) values.

The `Fixed effects coefficients` table indicates that `fitglme` returned 95% confidence intervals. It contains one row for each fixed-effects predictor, and each column contains statistics corresponding to that predictor. Column 1 (`Name`) contains the name of each fixed-effects coefficient, column 2 (`Estimate`) contains its estimated value, and column 3 (`SE`) contains the standard error of the coefficient. Column 4 (`tStat`) contains the *t*-statistic for a hypothesis test that the coefficient is equal to 0. Column 5 (`DF`) and column 6 (`pValue`) contain the degrees of freedom and *p*-value that correspond to the *t*-statistic, respectively. The last two columns (`Lower` and `Upper`) display the lower and upper limits, respectively, of the 95% confidence interval for each fixed-effects coefficient.

`Random effects covariance parameters` displays a table for each grouping variable (here, only `factory`), including its total number of levels (20), and the type and estimate of the covariance parameter. Here, `std` indicates that `fitglme` returns the standard deviation of the random effect associated with the factory predictor, which has an estimated value of 0.31381. It also displays a table containing the error parameter type (here, the square root of the dispersion parameter), and its estimated value of 1.

The standard display generated by `fitglme` does not provide confidence intervals for the random-effects parameters. To compute and display these values, use `covarianceParameters`.

Input Arguments

`tbl` — Input data

table | dataset array

Input data, which includes the response variable, predictor variables, and grouping variables, specified as a table or dataset array. The predictor variables can be continuous or grouping variables (see “Grouping Variables” on page 2-45). You must specify the model for the variables using `formula`.

`formula` — Formula for model specification

character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`

Formula for model specification, specified as a character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`. The formula is case sensitive. For a full description, see “Formula” on page 33-1961.

Example: `'y ~ treatment + (1|block)'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects'` specifies the response variable distribution as Poisson, the link function as log, the fit method as Laplace, and dummy variable coding where the coefficients sum to 0.

`BinomialSize` — Number of trials for binomial distribution

1 (default) | scalar value | vector | variable name

Number of trials for binomial distribution, that is the sample size, specified as the comma-separated pair consisting of a scalar value, a vector of the same length as the response, or the name of a variable in the input table. If you specify the name of a variable, then the variable must be of the same length as the response. `BinomialSize` applies only when the `Distribution` parameter is `'binomial'`.

If `BinomialSize` is a scalar value, that means all observations have the same number of trials.

Data Types: `single` | `double`

`CheckHessian` — Indicator to check positive definiteness of Hessian

`false` (default) | `true`

Indicator to check the positive definiteness of the Hessian of the objective function with respect to unconstrained parameters at convergence, specified as the comma-separated pair consisting of `'CheckHessian'` and either `false` or `true`. Default is `false`.

Specify `'CheckHessian'` as `true` to verify optimality of the solution or to determine if the model is overparameterized in the number of covariance parameters.

If you specify 'FitMethod' as 'MPL' or 'REMPL', then the covariance of the fixed effects and the covariance parameters is based on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

Example: 'CheckHessian',true

CovarianceMethod — Method to compute covariance of estimated parameters
'conditional' (default) | 'JointHessian'

Method to compute covariance of estimated parameters, specified as the comma-separated pair consisting of 'CovarianceMethod' and either 'conditional' or 'JointHessian'. If you specify 'conditional', then fitglme computes a fast approximation to the covariance of fixed effects given the estimated covariance parameters. It does not compute the covariance of covariance parameters. If you specify 'JointHessian', then fitglme computes the joint covariance of fixed effects and covariance parameters via the observed information matrix using the Laplacian loglikelihood.

If you specify 'FitMethod' as 'MPL' or 'REMPL', then the covariance of the fixed effects and the covariance parameters is based on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

Example: 'CovarianceMethod','JointHessian'

CovariancePattern — Pattern of covariance matrix
'FullCholesky' | 'Isotropic' | 'Full' | 'Diagonal' | 'CompSymm' | square symmetric logical matrix | string array | cell array of character vectors or logical matrices

Pattern of the covariance matrix of the random effects, specified as the comma-separated pair consisting of 'CovariancePattern' and 'FullCholesky', 'Isotropic', 'Full', 'Diagonal', 'CompSymm', a square symmetric logical matrix, a string array, or a cell array containing character vectors or logical matrices.

If there are R random-effects terms, then the value of 'CovariancePattern' must be a string array or cell array of length R, where each element r of the array specifies the pattern of the covariance matrix of the random-effects vector associated with the rth random-effects term. The options for each element follow.

Value	Description
'FullCholesky'	Full covariance matrix using the Cholesky parameterization. fitglme estimates all elements of the covariance matrix.

Value	Description
`'Isotropic'`	Diagonal covariance matrix with equal variances. That is, off-diagonal elements of the covariance matrix are constrained to be 0, and the diagonal elements are constrained to be equal. For example, if there are three random-effects terms with an isotropic covariance structure, this covariance matrix looks like $\begin{pmatrix} \sigma_b^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_b^2 \end{pmatrix}$ where σ^2_1 is the common variance of the random-effects terms.
`'Full'`	Full covariance matrix, using the log-Cholesky parameterization. `fitlme` estimates all elements of the covariance matrix.
`'Diagonal'`	Diagonal covariance matrix. That is, off-diagonal elements of the covariance matrix are constrained to be 0. $\begin{pmatrix} \sigma_{b1}^2 & 0 & 0 \\ 0 & \sigma_{b2}^2 & 0 \\ 0 & 0 & \sigma_{b3}^2 \end{pmatrix}$
`'CompSymm'`	Compound symmetry structure. That is, common variance along diagonals and equal correlation between all random effects. For example, if there are three random-effects terms with a covariance matrix having a compound symmetry structure, this covariance matrix looks like $\begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1,b2} & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1}^2 & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1,b2} & \sigma_{b1}^2 \end{pmatrix}$ where σ^2_{b1} is the common variance of the random-effects terms and $\sigma_{b1,b2}$ is the common covariance between any two random-effects term .
`PAT`	Square symmetric logical matrix. If `'CovariancePattern'` is defined by the matrix `PAT`, and if `PAT(a,b) = false`, then the `(a,b)` element of the corresponding covariance matrix is constrained to be 0.

For scalar random-effects terms, the default is `'Isotropic'`. Otherwise, the default is `'FullCholesky'`.

Example: `'CovariancePattern','Diagonal'`

Example: `'CovariancePattern',{'Full','Diagonal'}`

Data Types: `char` | `string` | `logical` | `cell`

`DispersionFlag` — Indicator to compute dispersion parameter
`false` for `'binomial'` and `'poisson'` distributions (default) | `true`

Indicator to compute dispersion parameter for `'binomial'` and `'poisson'` distributions, specified as the comma-separated pair consisting of `'DispersionFlag'` and one of the following.

Value	Description
`true`	Estimate a dispersion parameter when computing standard errors
`false`	Use the theoretical value of `1.0` when computing standard errors

`'DispersionFlag'` only applies if `'FitMethod'` is `'MPL'` or `'REMPL'`.

The fitting function always estimates the dispersion for other distributions.

Example: `'DispersionFlag',true`

`Distribution` — Distribution of the response variable
`'Normal'` (default) | `'Binomial'` | `'Poisson'` | `'Gamma'` | `'InverseGaussian'`

Distribution of the response variable, specified as the comma-separated pair consisting of `'Distribution'` and one of the following.

Value	Description
`'Normal'`	Normal distribution
`'Binomial'`	Binomial distribution
`'Poisson'`	Poisson distribution
`'Gamma'`	Gamma distribution
`'InverseGaussian'`	Inverse Gaussian distribution

Example: `'Distribution','Binomial'`

`DummyVarCoding` — Coding to use for dummy variables
`'reference'` (default) | `'effects'` | `'full'`

Coding to use for dummy variables created from the categorical variables, specified as the comma-separated pair consisting of `'DummyVarCoding'` and one of the variables in this table.

Value	Description
`'reference'` (default)	`fitglme` creates dummy variables with a reference group. This scheme treats the first category as a reference group and creates one less dummy variables than the number of categories. You can check the category order of a categorical variable by using the `categories` function, and change the order by using the `reordercats` function.
`'effects'`	`fitglme` creates dummy variables using effects coding. This scheme uses –1 to represent the last category. This scheme creates one less dummy variables than the number of categories.
`'full'`	`fitglme` creates full dummy variables. This scheme creates one dummy variable for each category.

For more details about creating dummy variables, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'DummyVarCoding','effects'`

`EBMethod` — Method used to approximate empirical Bayes estimates of random effects
`'Auto'` (default) | `'LineSearchNewton'` | `'TrustRegion2D'` | `'fsolve'`

Method used to approximate empirical Bayes estimates of random effects, specified as the comma-separated pair consisting of `'EBMethod'` and one of the following.

- `'Auto'`
- `'LineSearchNewton'`
- `'TrustRegion2D'`
- `'fsolve'`

`'Auto'` is similar to `'LineSearchNewton'` but uses a different convergence criterion and does not display iterative progress. `'Auto'` and `'LineSearchNewton'` may fail for non-canonical link functions. For non-canonical link functions, `'TrustRegion2D'` or `'fsolve'` are recommended. You must have Optimization Toolbox to use `'fsolve'`.

Example: `'EBMethod','LineSearchNewton'`

`EBOptions` — Options for empirical Bayes optimization
structure

Options for empirical Bayes optimization, specified as the comma-separated pair consisting of `'EBOptions'` and a structure containing the following.

Value	Description
`'TolFun'`	Relative tolerance on the gradient norm. Default is 1e-6.
`'TolX'`	Absolute tolerance on the step size. Default is 1e-8.

Value	Description
'MaxIter'	Maximum number of iterations. Default is 100.
'Display'	'off', 'iter', or 'final'. Default is 'off'.

If EBMethod is 'Auto' and 'FitMethod' is 'Laplace', TolFun is the relative tolerance on the linear predictor of the model, and the 'Display' option does not apply.

If 'EBMethod' is 'fsolve', then 'EBOptions' must be specified as an object created by optimoptions('fsolve').

Data Types: struct

Exclude — Indices for rows to exclude
use all rows without NaNs (default) | vector of integer or logical values

Indices for rows to exclude from the generalized linear mixed-effects model in the data, specified as the comma-separated pair consisting of 'Exclude' and a vector of integer or logical values.

For example, you can exclude the 13th and 67th rows from the fit as follows.

Example: 'Exclude',[13,67]

Data Types: single | double | logical

FitMethod — Method for estimating model parameters
'MPL' (default) | 'REMPL' | 'Laplace' | 'ApproximateLaplace

Method for estimating model parameters, specified as the comma-separated pair consisting of 'FitMethod' and one of the following.

- 'MPL' — Maximum pseudo likelihood
- 'REMPL' — Restricted maximum pseudo likelihood
- 'Laplace' — Maximum likelihood using Laplace approximation
- 'ApproximateLaplace' — Maximum likelihood using approximate Laplace approximation with fixed effects profiled out

Example: 'FitMethod','REMPL'

InitPLIterations — Initial number of pseudo likelihood iterations
10 (default) | integer value in the range [1,∞)

Initial number of pseudo likelihood iterations used to initialize parameters for ApproximateLaplace and Laplace fit methods, specified as the comma-separated pair consisting of 'InitPLIterations' and an integer value greater than or equal to 1.

Data Types: single | double

Link — Link function
'identity' | 'log' | 'logit' | 'probit' | 'comploglog' | 'reciprocal' | scalar value | structure

Link function, specified as the comma-separated pair consisting of 'Link' and one of the following.

Value	Description
`'identity'`	`g(mu) = mu` This is the default for the normal distribution.
`'log'`	`g(mu) = log(mu)` This is the default for the Poisson distribution.
`'logit'`	`g(mu) = log(mu/(1-mu))` This is the default for the binomial distribution.
`'loglog'`	`g(mu) = log(-log(mu))`
`'probit'`	`g(mu) = norminv(mu)`
`'comploglog'`	`g(mu) = log(-log(1-mu))`
`'reciprocal'`	`g(mu) = mu.^(-1)`
Scalar value `P`	`g(mu) = mu.^P`
Structure `S`	A structure containing four fields whose values are function handles with the following names: • `S.Link` — Link function • `S.Derivative` — Derivative • `S.SecondDerivative` — Second derivative • `S.Inverse` — Inverse of link Specification of `S.SecondDerivative` can be omitted if `FitMethod` is `MPL` or `REMPL`, or if `S` is the canonical link for the specified distribution.

The default link function used by `fitglme` is the canonical link that depends on the distribution of the response.

Response Distribution	Canonical Link Function
`'Normal'`	`'identity'`
`'Binomial'`	`'logit'`
`'Poisson'`	`'log'`
`'Gamma'`	`-1`
`'InverseGaussian'`	`-2`

Example: `'Link','log'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`MuStart` — Starting value for conditional mean
scalar value

Starting value for conditional mean, specified as the comma-separated pair consisting of `'MuStart'` and a scalar value. Valid values are as follows.

Response Distribution	Valid Values
'Normal'	(-Inf,Inf)
'Binomial'	(0,1)
'Poisson'	(0,Inf)
'Gamma'	(0,Inf)
'InverseGaussian'	(0,Inf)

Data Types: single | double

Offset — Offset
zeros(n,1) (default) | *n*-by-1 vector of scalar values

Offset, specified as the comma-separated pair consisting of 'Offset' and an *n*-by-1 vector of scalar values, where *n* is the length of the response vector. You can also specify the variable name of an *n*-by-1 vector of scalar values. 'Offset' is used as an additional predictor that has a coefficient value fixed at 1.0.

Data Types: single | double

Optimizer — Optimization algorithm
'quasinewton' (default) | 'fminsearch' | 'fminunc'

Optimization algorithm, specified as the comma-separated pair consisting of 'Optimizer' and either of the following.

Value	Description
'quasinewton'	Uses a trust region based quasi-Newton optimizer. You can change the options of the algorithm using statset('fitglme'). If you do not specify the options, then fitglme uses the default options of statset('fitglme').
'fminsearch'	Uses a derivative-free Nelder-Mead method. You can change the options of the algorithm using optimset('fminsearch'). If you do not specify the options, then fitglme uses the default options of optimset('fminsearch').
'fminunc'	Uses a line search-based quasi-Newton method. You must have Optimization Toolbox to specify this option. You can change the options of the algorithm using optimoptions('fminunc'). If you do not specify the options, then fitglme uses the default options of optimoptions('fminunc') with 'Algorithm' set to 'quasi-newton'.

Example: 'Optimizer','fminsearch'

OptimizerOptions — Options for optimization algorithm
structure returned by statset | structure returned by optimset | object returned by optimoptions

Options for the optimization algorithm, specified as the comma-separated pair consisting of `'OptimizerOptions'` and a structure returned by `statset('fitglme')`, a structure created by `optimset('fminsearch')`, or an object returned by `optimoptions('fminunc')`.

- If `'Optimizer'` is `'fminsearch'`, then use `optimset('fminsearch')` to change the options of the algorithm. If `'Optimizer'` is `'fminsearch'` and you do not supply `'OptimizerOptions'`, then the defaults used in `fitglme` are the default options created by `optimset('fminsearch')`.
- If `'Optimizer'` is `'fminunc'`, then use `optimoptions('fminunc')` to change the options of the optimization algorithm. See `optimoptions` for the options `'fminunc'` uses. If `'Optimizer'` is `'fminunc'` and you do not supply `'OptimizerOptions'`, then the defaults used in `fitglme` are the default options created by `optimoptions('fminunc')` with `'Algorithm'` set to `'quasi-newton'`.
- If `'Optimizer'` is `'quasinewton'`, then use `statset('fitglme')` to change the optimization parameters. If `'Optimizer'` is `'quasinewton'` and you do not change the optimization parameters using `statset`, then `fitglme` uses the default options created by `statset('fitglme')`.

The `'quasinewton'` optimizer uses the following fields in the structure created by `statset('fitglme')`.

`TolFun` — Relative tolerance on gradient of objective function
`1e-6` (default) | positive scalar value

Relative tolerance on the gradient of the objective function, specified as a positive scalar value.

`TolX` — Absolute tolerance on step size
`1e-12` (default) | positive scalar value

Absolute tolerance on the step size, specified as a positive scalar value.

`MaxIter` — Maximum number of iterations allowed
`10000` (default) | positive scalar value

Maximum number of iterations allowed, specified as a positive scalar value.

`Display` — Level of display
`'off'` (default) | `'iter'` | `'final'`

Level of display, specified as one of `'off'`, `'iter'`, or `'final'`.

`PLIterations` — Maximum number of pseudo likelihood iterations
`100` (default) | positive integer value

Maximum number of pseudo likelihood (PL) iterations, specified as the comma-separated pair consisting of `'PLIterations'` and a positive integer value. PL is used for fitting the model if `'FitMethod'` is `'MPL'` or `'REMPL'`. For other `'FitMethod'` values, PL iterations are used to initialize parameters for subsequent optimization.

Example: `'PLIterations',200`

Data Types: `single` | `double`

`PLTolerance` — Relative tolerance factor for pseudo likelihood iterations
`1e–08` (default) | positive scalar value

Relative tolerance factor for pseudo likelihood iterations, specified as the comma-separated pair consisting of `'PLTolerance'` and a positive scalar value.

Example: `'PLTolerance',1e-06`

Data Types: `single` | `double`

`StartMethod` — Method to start iterative optimization
`'default'` (default) | `'random'`

Method to start iterative optimization, specified as the comma-separated pair consisting of `'StartMethod'` and either of the following.

Value	Description
`'default'`	An internally defined default value
`'random'`	A random initial value

Example: `'StartMethod','random'`

`UseSequentialFitting` — Initial fitting type
`false` (default) | `true`

, specified as the comma-separated pair consisting of `'UseSequentialFitting'` and either `false` or `true`. If `'UseSequentialFitting'` is `false`, all maximum likelihood methods are initialized using one or more pseudo likelihood iterations. If `'UseSequentialFitting'` is `true`, the initial values from pseudo likelihood iterations are refined using `'ApproximateLaplace'` for `'Laplace'` fitting.

Example: `'UseSequentialFitting',true`

`Verbose` — Indicator to display optimization process on screen
`0` (default) | `1` | `2`

Indicator to display the optimization process on screen, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. If `'Verbose'` is specified as `1` or `2`, then `fitglme` displays the progress of the iterative model-fitting process. Specifying `'Verbose'` as `2` displays iterative optimization information from the individual pseudo likelihood iterations. Specifying `'Verbose'` as `1` omits this display.

The setting for `'Verbose'` overrides the field `'Display'` in `'OptimizerOptions'`.

Example: `'Verbose',1`

`Weights` — Observation weights
vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations. If the response distribution is binomial or Poisson, then `'Weights'` must be a vector of positive integers.

Data Types: `single` | `double`

Output Arguments

`glme` — Generalized linear mixed-effects model
`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

More About

Formula

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For the generalized linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms.

Suppose a table `tbl` contains the following:

- A response variable, `y`
- Predictor variables, X_j, which can be continuous or grouping variables
- Grouping variables, g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form, `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random`$_1$ is a specification of the random-effects design matrix Z_1 corresponding to grouping variable g_1, and similarly `random`$_R$ is a specification of the random-effects design matrix Z_R corresponding to grouping variable g_R. You can express the `fixed` and `random` terms using Wilkinson notation.

Wilkinson notation describes the factors present in models. The notation relates to factors present in models, not to the multipliers (coefficients) of those factors.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	`X`, X^2, ..., X^k
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (elementwise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only
`- X2`	Do not include `X2`
`X1*X2 + X3`	`X1`, `X2`, `X3`, `X1*X2`
`X1 + X2 + X3 + X1:X2`	`X1`, `X2`, `X3`, `X1*X2`
`X1*X2*X3 - X1:X2:X3`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`, `X2*X3`
`X1*(X2 + X3)`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`. Here are some examples for generalized linear mixed-effects model specification.

Examples:

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, `X1` and `X2`. This is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept and fixed effects for `X1` and `X2`. The implicit intercept term is suppressed by including `-1`.
`'y ~ 1 + (1 | g1)'`	Fixed effects for the intercept plus random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

See Also

`GeneralizedLinearMixedModel`

Topics

“Fit a Generalized Linear Mixed-Effects Model” on page 12-57
“Generalized Linear Mixed-Effects Models” on page 12-48

Introduced in R2014b

fitgmdist

Fit Gaussian mixture model to data

Syntax

```
GMModel = fitgmdist(X,k)
GMModel = fitgmdist(X,k,Name,Value)
```

Description

`GMModel = fitgmdist(X,k)` returns a Gaussian mixture distribution model (`GMModel`) with `k` components fitted to data (`X`).

`GMModel = fitgmdist(X,k,Name,Value)` returns a Gaussian mixture distribution model with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify a regularization value or the covariance type.

Examples

Cluster Data Using a Gaussian Mixture Model

Generate data from a mixture of two bivariate Gaussian distributions.

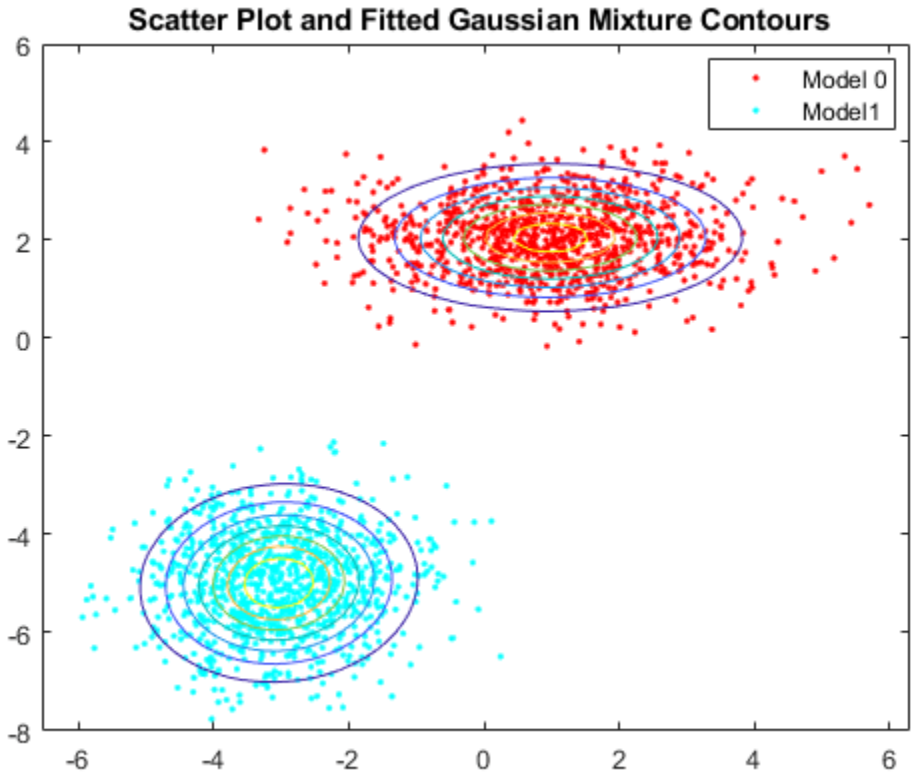
```
mu1 = [1 2];
Sigma1 = [2 0; 0 0.5];
mu2 = [-3 -5];
Sigma2 = [1 0;0 1];
rng(1); % For reproducibility
X = [mvnrnd(mu1,Sigma1,1000); mvnrnd(mu2,Sigma2,1000)];
```

Fit a Gaussian mixture model. Specify that there are two components.

```
GMModel = fitgmdist(X,2);
```

Plot the data over the fitted Gaussian mixture model contours.

```
figure
y = [zeros(1000,1);ones(1000,1)];
h = gscatter(X(:,1),X(:,2),y);
hold on
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(GMModel,[x0 y0]),x,y);
g = gca;
fcontour(gmPDF,[g.XLim g.YLim])
title('{\bf Scatter Plot and Fitted Gaussian Mixture Contours}')
legend(h,'Model 0','Model1')
hold off
```



Regularize Gaussian Mixture Model Estimation

Generate data from a mixture of two bivariate Gaussian distributions. Create a third predictor that is the sum of the first and second predictors.

```
mu1 = [1 2];
Sigma1 = [1 0; 0 1];
mu2 = [3 4];
Sigma2 = [0.5 0; 0 0.5];
rng(3); % For reproducibility
X1 = [mvnrnd(mu1,Sigma1,100);mvnrnd(mu2,Sigma2,100)];
X = [X1,X1(:,1)+X1(:,2)];
```

The columns of `X` are linearly dependent. This can cause ill-conditioned covariance estimates.

Fit a Gaussian mixture model to the data. You can use `try` / `catch` statements to help manage error messages.

```
rng(1); % Reset seed for common start values
try
    GMModel = fitgmdist(X,2)
catch exception
    disp('There was an error fitting the Gaussian mixture model')
    error = exception.message
end
```

```
There was an error fitting the Gaussian mixture model

error =
'Ill-conditioned covariance created at iteration 2.'
```

The covariance estimates are ill-conditioned. Consequently, optimization stops and an error appears.

Fit a Gaussian mixture model again, but use regularization.

```
rng(3); % Reset seed for common start values
GMModel = fitgmdist(X,2,'RegularizationValue',0.1)

GMModel =

Gaussian mixture distribution with 2 components in 3 dimensions
Component 1:
Mixing proportion: 0.536725
Mean:    2.8831    3.9506    6.8338

Component 2:
Mixing proportion: 0.463275
Mean:    0.8813    1.9758    2.8571
```

In this case, the algorithm converges to a solution due to regularization.

Select the Number of Gaussian Mixture Model Components Using PCA

Gaussian mixture models require that you specify a number of components before being fit to data. For many applications, it might be difficult to know the appropriate number of components. This example shows how to explore the data, and try to get an initial guess at the number of components using principal component analysis.

Load Fisher's iris data set.

```
load fisheriris
classes = unique(species)

classes = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

The data set contains three classes of iris species. The analysis proceeds as if this is unknown.

Use principal component analysis to reduce the dimension of the data to two dimensions for visualization.

```
[~,score] = pca(meas,'NumComponents',2);
```

Fit three Gaussian mixture models to the data by specifying 1, 2, and 3 components. Increase the number of optimization iterations to `1000`. Use dot notation to store the final parameter estimates. By default, the software fits full and different covariances for each component.

```
GMModels = cell(3,1); % Preallocation
options = statset('MaxIter',1000);
```

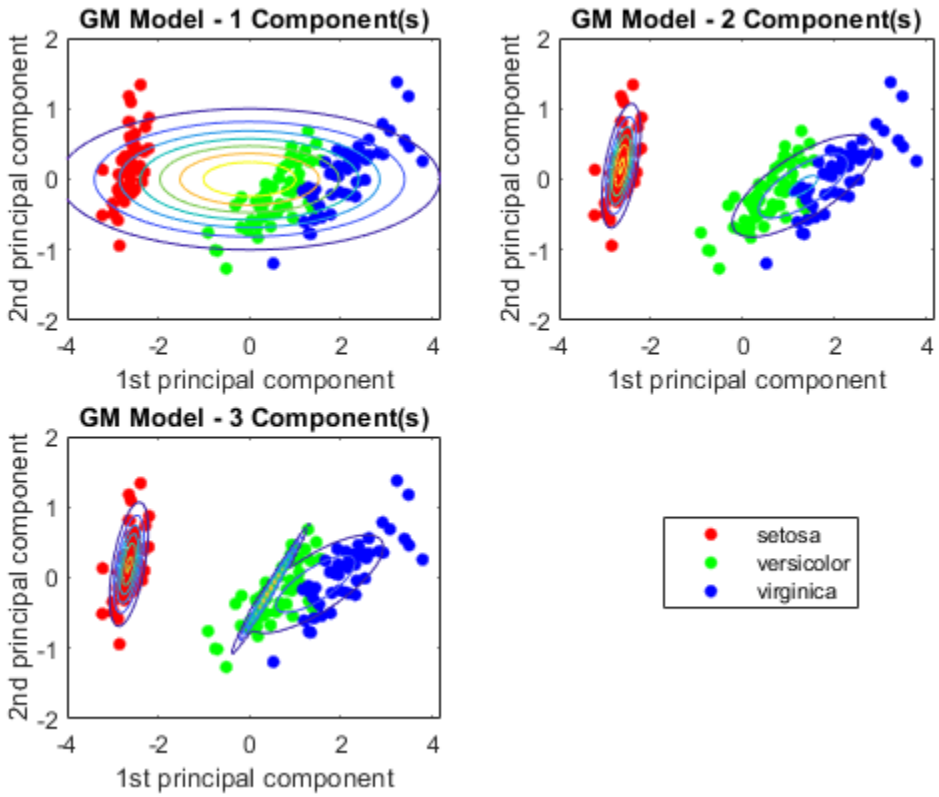
```
rng(1); % For reproducibility

for j = 1:3
    GMModels{j} = fitgmdist(score,j,'Options',options);
    fprintf('\n GM Mean for %i Component(s)\n',j)
    Mu = GMModels{j}.mu
end
```

```
 GM Mean for 1 Component(s)

Mu = 1×2
10^-14 ×

    0.1051   -0.0438


 GM Mean for 2 Component(s)

Mu = 2×2

    1.3212   -0.0954
   -2.6424    0.1909


 GM Mean for 3 Component(s)

Mu = 3×2

    0.4856   -0.1287
    1.4484   -0.0904
   -2.6424    0.1909
```

`GMModels` is a cell array containing three, fitted `gmdistribution` models. The means in the three component models are different, suggesting that the model distinguishes among the three iris species.

Plot the scores over the fitted Gaussian mixture model contours. Since the data set includes labels, use `gscatter` to distinguish between the true number of components.

```
figure
for j = 1:3
    subplot(2,2,j)
    h1 = gscatter(score(:,1),score(:,2),species);
    h = gca;
    hold on
    gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(GMModels{j},[x0 y0]),x,y);
    fcontour(gmPDF,[h.XLim h.YLim],'MeshDensity',100)
    title(sprintf('GM Model - %i Component(s)',j));
    xlabel('1st principal component');
    ylabel('2nd principal component');
    if(j ~= 3)
        legend off;
    end
    hold off
end
g = legend(h1);
g.Position = [0.7 0.25 0.1 0.1];
```



The three-component Gaussian mixture model, in conjunction with PCA, looks like it distinguishes between the three iris species.

There are other options you can use to help select the appropriate number of components for a Gaussian mixture model. For example,

- Compare multiple models with varying numbers of components using information criteria, e.g., AIC or BIC.
- Estimate the number of clusters using `evalclusters`, which supports, the Calinski-Harabasz criterion and the gap statistic, or other criteria.

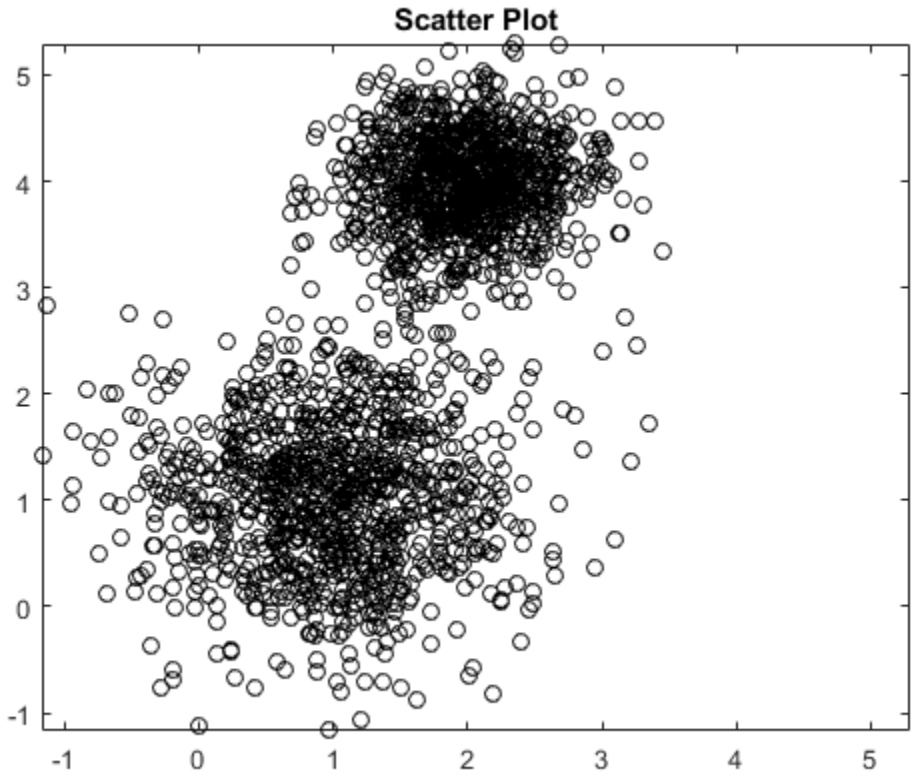
Determine the Best Gaussian Mixture Fit Using AIC

Gaussian mixture models require that you specify a number of components before being fit to data. For many applications, it might be difficult to know the appropriate number of components. This example uses the AIC fit statistic to help you choose the best fitting Gaussian mixture model over varying numbers of components.

Generate data from a mixture of two bivariate Gaussian distributions.

```
mu1 = [1 1];
Sigma1 = [0.5 0; 0 0.5];
mu2 = [2 4];
```

```
Sigma2 = [0.2 0; 0 0.2];
rng(1);
X = [mvnrnd(mu1,Sigma1,1000);mvnrnd(mu2,Sigma2,1000)];

plot(X(:,1),X(:,2),'ko')
title('Scatter Plot')
xlim([min(X(:)) max(X(:))]) % Make axes have the same scale
ylim([min(X(:)) max(X(:))])
```



Supposing that you do not know the underlying parameter values, the scatter plots suggests:

- There are two components.
- The variances between the clusters are different.
- The variance within the clusters is the same.
- There is no covariance within the clusters.

Fit a two-component Gaussian mixture model. Based on the scatter plot inspection, specify that the covariance matrices are diagonal. Print the final iteration and loglikelihood statistic to the Command Window by passing a `statset` structure as the value of the `Options` name-value pair argument.

```
options = statset('Display','final');
GMModel = fitgmdist(X,2,'CovarianceType','diagonal','Options',options);
```

```
11 iterations, log-likelihood = -4787.38
```

`GMModel` is a fitted `gmdistribution` model.

Examine the AIC over varying numbers of components.

```
AIC = zeros(1,4);
GMModels = cell(1,4);
options = statset('MaxIter',500);
for k = 1:4
    GMModels{k} = fitgmdist(X,k,'Options',options,'CovarianceType','diagonal');
    AIC(k)= GMModels{k}.AIC;
end

[minAIC,numComponents] = min(AIC);
numComponents
```

```
numComponents = 2
```

```
BestModel = GMModels{numComponents}
```

```
BestModel =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.501719
Mean:    1.9824    4.0013

Component 2:
Mixing proportion: 0.498281
Mean:    0.9880    1.0511
```

The smallest AIC occurs when the software fits the two-component Gaussian mixture model.

Set Initial Values When Fitting Gaussian Mixture Models

Gaussian mixture model parameter estimates might vary with different initial values. This example shows how to control initial values when you fit Gaussian mixture models using `fitgmdist`.

Load Fisher's iris data set. Use the petal lengths and widths as predictors.

```
load fisheriris
X = meas(:,3:4);
```

Fit a Gaussian mixture model to the data using default initial values. There are three iris species, so specify $k = 3$ components.

```
rng(10); % For reproducibility
GMModel1 = fitgmdist(X,3);
```

By default, the software:

1. Implements the “k-means++ Algorithm for Initialization” on page 33-1976 to choose $k = 3$ initial cluster centers.
2. Sets the initial covariance matrices as diagonal, where element `(j, j)` is the variance of `X(:,j)`.
3. Treats the initial mixing proportions as uniform.

Fit a Gaussian mixture model by connecting each observation to its label.

```
y = ones(size(X,1),1);
y(strcmp(species,'setosa')) = 2;
y(strcmp(species,'virginica')) = 3;

GMModel2 = fitgmdist(X,3,'Start',y);
```

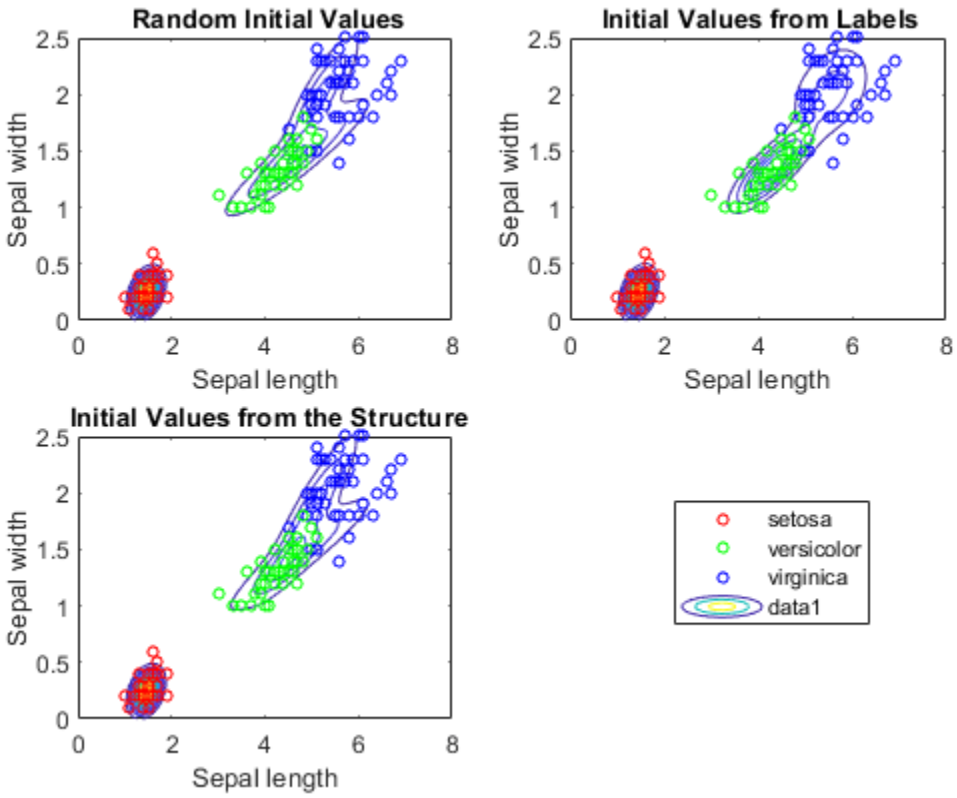
Fit a Gaussian mixture model by explicitly specifying the initial means, covariance matrices, and mixing proportions.

```
Mu = [1 1; 2 2; 3 3];
Sigma(:,:,1) = [1 1; 1 2];
Sigma(:,:,2) = 2*[1 1; 1 2];
Sigma(:,:,3) = 3*[1 1; 1 2];
PComponents = [1/2,1/4,1/4];
S = struct('mu',Mu,'Sigma',Sigma,'ComponentProportion',PComponents);

GMModel3 = fitgmdist(X,3,'Start',S);
```

Use `gscatter` to plot a scatter diagram that distinguishes between the iris species. For each model, plot the fitted Gaussian mixture model contours.

```
figure
subplot(2,2,1)
h = gscatter(X(:,1),X(:,2),species,[],'o',4);
haxis = gca;
xlim = haxis.XLim;
ylim = haxis.YLim;
d = (max([xlim ylim])-min([xlim ylim]))/1000;
[X1Grid,X2Grid] = meshgrid(xlim(1):d:xlim(2),ylim(1):d:ylim(2));
hold on
contour(X1Grid,X2Grid,reshape(pdf(GMModel1,[X1Grid(:) X2Grid(:)]),...
    size(X1Grid,1),size(X1Grid,2)),20)
uistack(h,'top')
title('{\bf Random Initial Values}');
xlabel('Sepal length');
ylabel('Sepal width');
legend off;
hold off
subplot(2,2,2)
h = gscatter(X(:,1),X(:,2),species,[],'o',4);
hold on
contour(X1Grid,X2Grid,reshape(pdf(GMModel2,[X1Grid(:) X2Grid(:)]),...
    size(X1Grid,1),size(X1Grid,2)),20)
uistack(h,'top')
title('{\bf Initial Values from Labels}');
xlabel('Sepal length');
ylabel('Sepal width');
legend off
hold off
subplot(2,2,3)
h = gscatter(X(:,1),X(:,2),species,[],'o',4);
hold on
contour(X1Grid,X2Grid,reshape(pdf(GMModel3,[X1Grid(:) X2Grid(:)]),...
    size(X1Grid,1),size(X1Grid,2)),20)
uistack(h,'top')
title('{\bf Initial Values from the Structure}');
xlabel('Sepal length');
ylabel('Sepal width');
```

```
legend('Location',[0.7,0.25,0.1,0.1]);
hold off
```



According to the contours, `GMModel2` seems to suggest a slight trimodality, while the others suggest bimodal distributions.

Display the estimated component means.

```
table(GMModel1.mu,GMModel2.mu,GMModel3.mu,'VariableNames',...
    {'Model1','Model2','Model3'})
```

```
ans=3×3 table
        Model1              Model2              Model3
    ________________    ________________    ________________

    5.2115     2.0119    4.2857    1.3339    1.4604    0.2429
     1.461    0.24423     1.462     0.246    4.7509    1.4629
    4.6829     1.4429    5.5507    2.0316    5.0158    1.8592
```

`GMModel2` seems to distinguish between the iris species the best.

Input Arguments

X — Data

numeric matrix

Data to which the Gaussian mixture model is fit, specified as a numeric matrix.

The rows of `X` correspond to observations, and the columns of `X` correspond to variables. The number of observations must be larger than each of the following: the number of variables and the number of components.

`NaN`s indicate missing values. The software removes rows of `X` containing at least one `NaN` before fitting, which decreases the effective sample size.

Data Types: `single` | `double`

`k` — Number of components
positive integer

Number of components to use when fitting Gaussian mixture model, specified as a positive integer. For example, if you specify `k = 3`, then the software fits a Gaussian mixture model with three distinct means, covariances matrices, and component proportions to the data (`X`).

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'RegularizationValue',0.1,'CovarianceType','diagonal'` specifies a regularization parameter value of 0.1 and to fit diagonal covariance matrices.

`CovarianceType` — Type of covariance matrix
`'full'` (default) | `'diagonal'`

Type of covariance matrix to fit to the data, specified as the comma-separated pair consisting of `'CovarianceType'` and either `'diagonal'` or `'full'`.

If you set `'diagonal'`, then the software fits diagonal covariance matrices. In this case, the software estimates `k*d` covariance parameters, where `d` is the number of columns in `X` (i.e., `d = size(X,2)`).

Otherwise, the software fits full covariance matrices. In this case, the software estimates `k*d*(d+1)/2` covariance parameters.

Example: `'CovarianceType','diagonal'`

`Options` — Iterative EM algorithm optimization options
`statset` options structure

Iterative EM algorithm optimization options, specified as the comma-separated pair consisting of `'Options'` and a `statset` options structure.

This table describes the available name-value pair arguments.

Name	Value
'Display'	'final': Display the final output. 'iter': Display iterative output to the Command Window for some functions; otherwise display the final output. 'off': Do not display optimization information.
'MaxIter'	Positive integer indicating the maximum number of iterations allowed. The default is 100
'TolFun'	Positive scalar indicating the termination tolerance for the loglikelihood function value. The default is 1e-6.

Example: 'Options',statset('Display','final','MaxIter',1500,'TolFun',1e-5)

ProbabilityTolerance — Tolerance for posterior probabilities
1e-8 (default) | nonnegative scalar value in range [0,1e-6]

Tolerance for posterior probabilities, specified as the comma-separated pair consisting of ProbabilityTolerance and a nonnegative scalar value in the range [0,1e-6].

In each iteration, after the estimation of posterior probabilities, fitgmdist sets any posterior probability that is not larger than the tolerance value to zero. Using a nonzero tolerance might speed up fitgmdist.

Example: 'ProbabilityTolerance',0.0000025

Data Types: single | double

RegularizationValue — Regularization parameter value
0 (default) | nonnegative scalar

Regularization parameter value, specified as the comma-separated pair consisting of 'RegularizationValue' and a nonnegative scalar.

Set RegularizationValue to a small positive scalar to ensure that the estimated covariance matrices are positive definite.

Example: 'RegularizationValue',0.01

Data Types: single | double

Replicates — Number of times to repeat EM algorithm
1 (default) | positive integer

Number of times to repeat the EM algorithm using a new set of initial values, specified as the comma-separated pair consisting of 'Replicates' and a positive integer.

If Replicates is greater than 1, then:

- The name-value pair argument Start must be plus (the default) or randSample.
- GMModel is the fit with the largest loglikelihood.

Example: 'Replicates',10

Data Types: `single` | `double`

`SharedCovariance` — Flag indicating whether all covariance matrices are identical
logical `false` (default) | logical `true`

Flag indicating whether all covariance matrices are identical (i.e., fit a pooled estimate), specified as the comma-separated pair consisting of `'SharedCovariance'` and either logical value `false` or `true`.

If `SharedCovariance` is `true`, then all `k` covariance matrices are equal, and the number of covariance parameters is scaled down by a factor of `k`.

`Start` — Initial value setting method
`'plus'` (default) | `'randSample'` | vector of integers | structure array

Initial value setting method, specified as the comma-separated pair consisting of `'Start'` and `'randSample'`, `'plus'`, a vector of integers, or a structure array.

The value of `Start` determines the initial values required by the optimization routine for each Gaussian component parameter — mean, covariance, and mixing proportion. This table summarizes the available options.

Value	Description
`'randSample'`	The software selects `k` observations from `X` at random as initial component means. The mixing proportions are uniform. The initial covariance matrices for all components are diagonal, where the element `j` on the diagonal is the variance of `X(:,j)`.
`'plus'`	The software selects `k` observations from `X` using the *k*-means++ algorithm on page 33-1976. The initial mixing proportions are uniform. The initial covariance matrices for all components are diagonal, where the element `j` on the diagonal is the variance of `X(:,j)`.
Vector of integers	A vector of length *n* (the number of observations) containing an initial guess of the component index for each point. That is, each element is an integer from 1 to `k`, which corresponds to a component. The software collects all observations corresponding to the same component, computes means, covariances, and mixing proportions for each, and sets the initial values to these statistics.

Value	Description

Structure array	Suppose that there are `d` variables (i.e., `d = size(X,2)`). The structure array, e.g., `S`, must have three fields:
	• `S.mu`: A `k`-by-`d` matrix specifying the initial mean of each component
	• `S.Sigma`: A numeric array specifying the covariance matrix of each component. `Sigma` is one of the following:
	  • A `d`-by-`d`-by-`k` array. `Sigma(:,:,j)` is the initial covariance matrix of component `j`.
	  • A `1`-by-`d`-by-`k` array. `diag(Sigma(:,:,j))` is the initial covariance matrix of component `j`.
	  • A `d`-by-`d` matrix. `Sigma` is the initial covariance matrix for all components.
	  • A `1`-by-`d` vector. `diag(Sigma)` is the initial covariance matrix for all components.
	• `S.ComponentProportion`: A `1`-by-`k` vector of scalars specifying the initial mixing proportions of each component. The default is uniform.

Example: `'Start',ones(n,1)`

Data Types: `single` | `double` | `char` | `string` | `struct`

Output Arguments

`GMModel` — Fitted Gaussian mixture model

`gmdistribution` model

Fitted Gaussian mixture model, returned as a `gmdistribution` model.

Access properties of `GMModel` using dot notation. For example, display the AIC by entering `GMModel.AIC`.

Tips

`fitgmdist` might:

- Converge to a solution where one or more of the components has an ill-conditioned or singular covariance matrix.

  The following issues might result in an ill-conditioned covariance matrix:

  - The number of dimensions of your data is relatively high and there are not enough observations.
  - Some of the predictors (variables) of your data are highly correlated.
  - Some or all the features are discrete.
  - You tried to fit the data to too many components.

  In general, you can avoid getting ill-conditioned covariance matrices by using one of the following precautions:

- Preprocess your data to remove correlated features.
- Set `'SharedCovariance'` to `true` to use an equal covariance matrix for every component.
- Set `'CovarianceType'` to `'diagonal'`.
- Use `'RegularizationValue'` to add a very small positive number to the diagonal of every covariance matrix.
- Try another set of initial values.

- Pass through an intermediate step where one or more of the components has an ill-conditioned covariance matrix. Try another set of initial values to avoid this issue without altering your data or model.

Algorithms

Gaussian Mixture Model Likelihood Optimization

The software optimizes the Gaussian mixture model likelihood using the iterative Expectation-Maximization (EM) algorithm.

`fitgmdist` fits GMMs to data using the iterative Expectation-Maximization (EM) algorithm. Using initial values for component means, covariance matrices, and mixing proportions, the EM algorithm proceeds using these steps.

1. For each observation, the algorithm computes posterior probabilities of component memberships. You can think of the result as an n-by-k matrix, where element (i,j) contains the posterior probability that observation i is from component j. This is the E-step of the EM algorithm.
2. Using the component-membership posterior probabilities as weights, the algorithm estimates the component means, covariance matrices, and mixing proportions by applying maximum likelihood. This is the M-step of the EM algorithm.

The algorithm iterates over these steps until convergence. The likelihood surface is complex, and the algorithm might converge to a local optimum. Also, the resulting local optimum might depend on the initial conditions. `fitgmdist` has several options for choosing initial conditions, including random component assignments for the observations and the k-means ++ algorithm.

k-means++ Algorithm for Initialization

The k-means++ algorithm uses an heuristic to find centroid seeds for k-means clustering. `fitgmdist` can apply the same principle to initialize the EM algorithm by using the k-means++ algorithm to select the initial parameter values for a fitted Gaussian mixture model.

The k-means++ algorithm assumes the number of clusters is k and chooses the initial parameter values as follows.

1. Select the component mixture probability to be the uniform probability $p_i = \frac{1}{k}$, where $i = 1, \ldots, k$.
2. Select the covariance matrices to be diagonal and identical, where $\sigma_i = diag(a_1, a_2, \ldots, a_k)$ and $a_j = \mathrm{var}(X_j)$.
3. Select the first initial component center μ_1 uniformly from all data points in X.
4. To choose center j:
   a. Compute the Mahalanobis distances from each observation to each centroid, and assign each observation to its closest centroid.

b For $m = 1,...,n$ and $p = 1,...,j - 1$, select centroid j at random from X with probability

$$\frac{d^2(x_m, \mu_p)}{\sum_{h;\, x_h \in M_p} d^2(x_h, \mu_p)}$$

where $d(x_m, \mu_p)$ is the distance between observation m and μ_p, and M_p is the set of all observations closest to centroid μ_p and x_m belongs to M_p.

That is, select each subsequent center with a probability proportional to the distance from itself to the closest center that you already chose.

5 Repeat step 4 until k centroids are chosen.

References

[1] McLachlan, G., and D. Peel. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons, Inc., 2000.

See Also

`cluster` | `gmdistribution`

Topics

"Tune Gaussian Mixture Models" on page 16-57
"Cluster Using Gaussian Mixture Model" on page 16-39

Introduced in R2014a

fitlm

Fit linear regression model

Syntax

```
mdl = fitlm(tbl)
mdl = fitlm(X,y)
mdl = fitlm(___,modelspec)
mdl = fitlm(___,Name,Value)
```

Description

`mdl = fitlm(tbl)` returns a linear regression model fit to variables in the table or dataset array `tbl`. By default, `fitlm` takes the last variable as the response variable.

`mdl = fitlm(X,y)` returns a linear regression model of the responses `y`, fit to the data matrix `X`.

`mdl = fitlm(___,modelspec)` defines the model specification using any of the input argument combinations in the previous syntaxes.

`mdl = fitlm(___,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify which variables are categorical, perform robust regression, or use observation weights.

Examples

Fit Linear Regression Using Data in Matrix

Fit a linear regression model using a matrix input data set.

Load the `carsmall` data set, a matrix input data set.

```
load carsmall
X = [Weight,Horsepower,Acceleration];
```

Fit a linear regression model by using `fitlm`.

```
mdl = fitlm(X,MPG)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                  __________    _________    _________    __________

    (Intercept)        47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
```

```
    x3              -0.011583      0.19333    -0.059913       0.95236


Number of observations: 93, Error degrees of freedom: 89
Root Mean Squared Error: 4.09
R-squared: 0.752,  Adjusted R-Squared: 0.744
F-statistic vs. constant model: 90, p-value = 7.38e-27
```

The model display includes the model formula, estimated coefficients, and model summary statistics.

The model formula in the display, `y ~ 1 + x1 + x2 + x3`, corresponds to $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

The model display also shows the estimated coefficient information, which is stored in the `Coefficients` property. Display the `Coefficients` property.

```
mdl.Coefficients
```

```
ans=4×4 table
                    Estimate        SE          tStat        pValue
                   __________    _________    _________    __________

    (Intercept)       47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
    x3              -0.011583      0.19333    -0.059913       0.95236
```

The `Coefficient` property includes these columns:

- `Estimate` — Coefficient estimates for each corresponding term in the model. For example, the estimate for the constant term (`intercept`) is 47.977.
- `SE` — Standard error of the coefficients.
- `tStat` — *t*-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. Note that `tStat = Estimate/SE`. For example, the *t*-statistic for the intercept is 47.977/3.8785 = 12.37.
- `pValue` — *p*-value for the *t*-statistic of the hypothesis test that the corresponding coefficient is equal to zero or not. For example, the *p*-value of the *t*-statistic for `x2` is greater than 0.05, so this term is not significant at the 5% significance level given the other terms in the model.

The summary statistics of the model are:

- `Number of observations` — Number of rows without any `NaN` values. For example, `Number of observations` is 93 because the `MPG` data vector has six `NaN` values and the `Horsepower` data vector has one `NaN` value for a different observation, where the number of rows in `X` and `MPG` is 100.
- `Error degrees of freedom` — $n - p$, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, the model has four predictors, so the `Error degrees of freedom` is 93 – 4 = 89.
- `Root mean squared error` — Square root of the mean squared error, which estimates the standard deviation of the error distribution.

- `R-squared` and `Adjusted R-squared` — Coefficient of determination and adjusted coefficient of determination, respectively. For example, the `R-squared` value suggests that the model explains approximately 75% of the variability in the response variable `MPG`.
- `F-statistic vs. constant model` — Test statistic for the *F*-test on the regression model, which tests whether the model fits significantly better than a degenerate model consisting of only a constant term.
- `p-value` — *p*-value for the *F*-test on the model. For example, the model is significant with a *p*-value of 7.3816e-27.

You can find these statistics in the model properties (`NumObservations`, `DFE`, `RMSE`, and `Rsquared`) and by using the `anova` function.

```
anova(mdl,'summary')
```

```
ans=3×5 table
                SumSq     DF    MeanSq      F         pValue
                ______    __    ______    ______    __________

    Total       6004.8    92    65.269
    Model         4516     3    1505.3    89.987    7.3816e-27
    Residual    1488.8    89    16.728
```

Fit Linear Regression Using Data in Table

Load the sample data.

```
load carsmall
```

Store the variables in a table.

```
tbl = table(Weight,Acceleration,MPG,'VariableNames',{'Weight','Acceleration','MPG'});
```

Display the first five rows of the table.

```
tbl(1:5,:)
```

```
ans=5×3 table
    Weight    Acceleration    MPG
    ______    ____________    ___

     3504           12         18
     3693         11.5         15
     3436           11         18
     3433           12         16
     3449         10.5         17
```

Fit a linear regression model for miles per gallon (MPG). Specify the model formula by using Wilkinson notation.

```
lm = fitlm(tbl,'MPG~Weight+Acceleration')
```

```
lm =
Linear regression model:
```

```
    MPG ~ 1 + Weight + Acceleration

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________

    (Intercept)         45.155        3.4659     13.028    1.6266e-22
    Weight          -0.0082475    0.00059836    -13.783    5.3165e-24
    Acceleration       0.19694       0.14743     1.3359       0.18493


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 4.12
R-squared: 0.743,  Adjusted R-Squared: 0.738
F-statistic vs. constant model: 132, p-value = 1.38e-27
```

The model `'MPG~Weight+Acceleration'` in this example is equivalent to set the model specification as `'linear'`. For example,

```
lm2 = fitlm(tbl,'linear');
```

If you use a character vector for model specification and you do not specify the response variable, then `fitlm` accepts the last variable in `tbl` as the response variable and the other variables as the predictor variables.

Fit Linear Regression Using Specified Model Formula

Fit a linear regression model using a model formula specified by Wilkinson notation.

Load the sample data.

```
load carsmall
```

Store the variables in a table.

```
tbl = table(Weight,Acceleration,Model_Year,MPG,'VariableNames',{'Weight','Acceleration','Model_Ye
```

Fit a linear regression model for miles per gallon (MPG) with weight and acceleration as the predictor variables.

```
lm = fitlm(tbl,'MPG~Weight+Acceleration')
```

```
lm =
Linear regression model:
    MPG ~ 1 + Weight + Acceleration

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________

    (Intercept)         45.155        3.4659     13.028    1.6266e-22
    Weight          -0.0082475    0.00059836    -13.783    5.3165e-24
    Acceleration       0.19694       0.14743     1.3359       0.18493
```

```
Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 4.12
R-squared: 0.743,  Adjusted R-Squared: 0.738
F-statistic vs. constant model: 132, p-value = 1.38e-27
```

The *p*-value of 0.18493 indicates that `Acceleration` does not have a significant impact on `MPG`.

Remove `Acceleration` from the model, and try improving the model by adding the predictor variable `Model_Year`. First define `Model_Year` as a categorical variable.

```
tbl.Model_Year = categorical(tbl.Model_Year);
lm = fitlm(tbl,'MPG~Weight+Model_Year')
```

```
lm =
Linear regression model:
    MPG ~ 1 + Weight + Model_Year

Estimated Coefficients:
                       Estimate         SE         tStat       pValue
                      __________    __________    _______    __________

    (Intercept)           40.11        1.5418     26.016    1.2024e-43
    Weight           -0.0066475    0.00042802    -15.531    3.3639e-27
    Model_Year_76        1.9291       0.74761     2.5804      0.011488
    Model_Year_82        7.9093       0.84975     9.3078    7.8681e-15


Number of observations: 94, Error degrees of freedom: 90
Root Mean Squared Error: 2.92
R-squared: 0.873,  Adjusted R-Squared: 0.868
F-statistic vs. constant model: 206, p-value = 3.83e-40
```

Specifying `modelspec` using Wilkinson notation enables you to update the model without having to change the design matrix. `fitlm` uses only the variables that are specified in the formula. It also creates the necessary two dummy indicator variables for the categorical variable `Model_Year`.

Fit Linear Regression Using Terms Matrix

Fit a linear regression model using a terms matrix.

Terms Matrix for Table Input

If the model variables are in a table, then a column of `0`s in a terms matrix represents the position of the response variable.

Load the `hospital` data set.

```
load hospital
```

Store the variables in a table.

```
t = table(hospital.Sex,hospital.BloodPressure(:,1),hospital.Age,hospital.Smoker, ...
    'VariableNames',{'Sex','BloodPressure','Age','Smoker'});
```

Represent the linear model `'BloodPressure ~ 1 + Sex + Age + Smoker'` using a terms matrix. The response variable is in the second column of the table, so the second column of the terms matrix must be a column of `0`s for the response variable.

```
T = [0 0 0 0;1 0 0 0;0 0 1 0;0 0 0 1]
```

```
T = 4×4

     0     0     0     0
     1     0     0     0
     0     0     1     0
     0     0     0     1
```

Fit a linear model.

```
mdl1 = fitlm(t,T)
```

```
mdl1 =
Linear regression model:
    BloodPressure ~ 1 + Sex + Age + Smoker

Estimated Coefficients:
                   Estimate       SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      116.14      2.6107      44.485    7.1287e-66
    Sex_Male       0.050106     0.98364    0.050939       0.95948
    Age            0.085276    0.066945      1.2738        0.2058
    Smoker_1           9.87      1.0346      9.5395    1.4516e-15


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 4.78
R-squared: 0.507,  Adjusted R-Squared: 0.492
F-statistic vs. constant model: 33, p-value = 9.91e-15
```

Terms Matrix for Matrix Input

If the predictor and response variables are in a matrix and column vector, then you must include `0` for the response variable at the end of each row in a terms matrix.

Load the `carsmall` data set and define the matrix of predictors.

```
load carsmall
X = [Acceleration,Weight];
```

Specify the model `'MPG ~ Acceleration + Weight + Acceleration:Weight + Weight^2'` using a terms matrix. This model includes the main effect and two-way interaction terms for the variables `Acceleration` and `Weight`, and a second-order term for the variable `Weight`.

```
T = [0 0 0;1 0 0;0 1 0;1 1 0;0 2 0]
```

```
T = 5×3

     0     0     0
     1     0     0
     0     1     0
```

```
     1     1     0
     0     2     0
```

Fit a linear model.

```
mdl2 = fitlm(X,MPG,T)
```

```
mdl2 =
Linear regression model:
    y ~ 1 + x1*x2 + x2^2

Estimated Coefficients:
                    Estimate          SE         tStat       pValue
                   ___________    __________    _______    __________

    (Intercept)         48.906        12.589     3.8847    0.00019665
    x1                 0.54418       0.57125    0.95261       0.34337
    x2                -0.012781     0.0060312    -2.1192      0.036857
    x1:x2          -0.00010892    0.00017925    -0.6076         0.545
    x2^2            9.7518e-07    7.5389e-07     1.2935       0.19917


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 4.1
R-squared: 0.751,  Adjusted R-Squared: 0.739
F-statistic vs. constant model: 67, p-value = 4.99e-26
```

Only the intercept and `x2` term, which corresponds to the `Weight` variable, are significant at the 5% significance level.

Linear Regression with Categorical Predictor

Fit a linear regression model that contains a categorical predictor. Reorder the categories of the categorical predictor to control the reference level in the model. Then, use `anova` to test the significance of the categorical variable.

Model with Categorical Predictor

Load the `carsmall` data set and create a linear regression model of `MPG` as a function of `Model_Year`. To treat the numeric vector `Model_Year` as a categorical variable, identify the predictor using the `'CategoricalVars'` name-value pair argument.

```
load carsmall
mdl = fitlm(Model_Year,MPG,'CategoricalVars',1,'VarNames',{'Model_Year','MPG'})
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                    Estimate      SE      tStat       pValue
                   _________    ______    ______    __________

    (Intercept)        17.69    1.0328    17.127    3.2371e-30
    Model_Year_76     3.8839    1.4059    2.7625     0.0069402
```

```
    Model_Year_82      14.02     1.4369    9.7571    8.2164e-16


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

The model formula in the display, `MPG ~ 1 + Model_Year`, corresponds to

$$\mathrm{MPG} = \beta_0 + \beta_1 I_{\mathrm{Year}=76} + \beta_2 I_{\mathrm{Year}=82} + \epsilon,$$

where $I_{\mathrm{Year}=76}$ and $I_{\mathrm{Year}=82}$ are indicator variables whose value is one if the value of `Model_Year` is 76 and 82, respectively. The `Model_Year` variable includes three distinct values, which you can check by using the `unique` function.

```
unique(Model_Year)
```

```
ans = 3×1

    70
    76
    82
```

`fitlm` chooses the smallest value in `Model_Year` as a reference level (`'70'`) and creates two indicator variables $I_{\mathrm{Year}=76}$ and $I_{\mathrm{Year}=82}$. The model includes only two indicator variables because the design matrix becomes rank deficient if the model includes three indicator variables (one for each level) and an intercept term.

Model with Full Indicator Variables

You can interpret the model formula of `mdl` as a model that has three indicator variables without an intercept term:

$$y = \beta_0 I_{x_1=70} + (\beta_0 + \beta_1) I_{x_1=76} + (\beta_0 + \beta_2) I_{x_2=82} + \epsilon.$$

Alternatively, you can create a model that has three indicator variables without an intercept term by manually creating indicator variables and specifying the model formula.

```
temp_Year = dummyvar(categorical(Model_Year));
Model_Year_70 = temp_Year(:,1);
Model_Year_76 = temp_Year(:,2);
Model_Year_82 = temp_Year(:,3);
tbl = table(Model_Year_70,Model_Year_76,Model_Year_82,MPG);
mdl = fitlm(tbl,'MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82 - 1')
```

```
mdl =
Linear regression model:
    MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82

Estimated Coefficients:
                    Estimate      SE       tStat       pValue
                    ________    _______    ______    __________

    Model_Year_70      17.69     1.0328    17.127    3.2371e-30
    Model_Year_76     21.574    0.95387    22.617    4.0156e-39
```

```
    Model_Year_82      31.71     0.99896    31.743    5.2234e-51


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
```

Choose Reference Level in Model

You can choose a reference level by modifying the order of categories in a categorical variable. First, create a categorical variable `Year`.

```
Year = categorical(Model_Year);
```

Check the order of categories by using the `categories` function.

```
categories(Year)
```

```
ans = 3x1 cell
    {'70'}
    {'76'}
    {'82'}
```

If you use `Year` as a predictor variable, then `fitlm` chooses the first category `'70'` as a reference level. Reorder `Year` by using the `reordercats` function.

```
Year_reordered = reordercats(Year,{'76','70','82'});
categories(Year_reordered)
```

```
ans = 3x1 cell
    {'76'}
    {'70'}
    {'82'}
```

The first category of `Year_reordered` is `'76'`. Create a linear regression model of `MPG` as a function of `Year_reordered`.

```
mdl2 = fitlm(Year_reordered,MPG,'VarNames',{'Model_Year','MPG'})
```

```
mdl2 =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                     Estimate      SE        tStat       pValue
                     ________    _______    _______    __________

    (Intercept)       21.574     0.95387     22.617    4.0156e-39
    Model_Year_70    -3.8839      1.4059    -2.7625     0.0069402
    Model_Year_82     10.136      1.3812     7.3385    8.7634e-11


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

`mdl2` uses `'76'` as a reference level and includes two indicator variables $I_{\text{Year = 70}}$ and $I_{\text{Year = 82}}$.

Evaluate Categorical Predictor

The model display of `mdl2` includes a *p*-value of each term to test whether or not the corresponding coefficient is equal to zero. Each *p*-value examines each indicator variable. To examine the categorical variable `Model_Year` as a group of indicator variables, use `anova`. Use the `'components'`(default) option to return a component ANOVA table that includes ANOVA statistics for each variable in the model except the constant term.

```
anova(mdl2,'components')
```

```
ans=2×5 table
                SumSq     DF    MeanSq      F        pValue
                ______    __    ______    _____    __________

    Model_Year    3190.1     2    1595.1    51.56    1.0694e-15
    Error         2815.2    91    30.936
```

The component ANOVA table includes the *p*-value of the `Model_Year` variable, which is smaller than the *p*-values of the indicator variables.

Specify Response and Predictor Variables for Linear Model

Fit a linear regression model to sample data. Specify the response and predictor variables, and include only pairwise interaction terms in the model.

Load sample data.

```
load hospital
```

Fit a linear model with interaction terms to the data. Specify weight as the response variable, and sex, age, and smoking status as the predictor variables. Also, specify that sex and smoking status are categorical variables.

```
mdl = fitlm(hospital,'interactions','ResponseVar','Weight',...
    'PredictorVars',{'Sex','Age','Smoker'},...
    'CategoricalVar',{'Sex','Smoker'})
```

```
mdl =
Linear regression model:
    Weight ~ 1 + Sex*Age + Sex*Smoker + Age*Smoker

Estimated Coefficients:
                        Estimate      SE        tStat        pValue
                        ________    _______    ________    __________

    (Intercept)            118.7     7.0718      16.785     6.821e-30
    Sex_Male               68.336     9.7153      7.0339    3.3386e-10
    Age                   0.31068    0.18531      1.6765      0.096991
    Smoker_1               3.0425     10.446     0.29127       0.77149
    Sex_Male:Age          -0.49094    0.24764     -1.9825      0.050377
    Sex_Male:Smoker_1       0.9509     3.8031     0.25003       0.80312
    Age:Smoker_1          -0.07288    0.26275    -0.27737       0.78211


Number of observations: 100, Error degrees of freedom: 93
```

```
Root Mean Squared Error: 8.75
R-squared: 0.898,  Adjusted R-Squared: 0.892
F-statistic vs. constant model: 137, p-value = 6.91e-44
```

The weight of the patients do not seem to differ significantly according to age, or the status of smoking, or interaction of these factors with patient sex at the 5% significance level.

Fit Robust Linear Regression Model

Load the `hald` data set, which measures the effect of cement composition on its hardening heat.

```
load hald
```

This data set includes the variables `ingredients` and `heat`. The matrix `ingredients` contains the percent composition of four chemicals present in the cement. The vector `heat` contains the values for the heat hardening after 180 days for each cement sample.

Fit a robust linear regression model to the data.

```
mdl = fitlm(ingredients,heat,'RobustOpts','on')
```

```
mdl =
Linear regression model (robust fit):
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    ________

    (Intercept)       60.09     75.818     0.79256      0.4509
    x1               1.5753    0.80585      1.9548    0.086346
    x2               0.5322    0.78315     0.67957     0.51596
    x3              0.13346     0.8166     0.16343     0.87424
    x4             -0.12052     0.7672    -0.15709     0.87906


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 2.65
R-squared: 0.979,  Adjusted R-Squared: 0.969
F-statistic vs. constant model: 94.6, p-value = 9.03e-07
```

For more details, see the topic “Reduce Outlier Effects Using Robust Regression” on page 11-104, which compares the results of a robust fit to a standard least-squares fit.

Compute Mean Absolute Error Using Cross-Validation

Compute the mean absolute error of a regression model by using 10-fold cross-validation.

Load the `carsmall` data set. Specify the `Acceleration` and `Displacement` variables as predictors and the `Weight` variable as the response.

```
load carsmall
X1 = Acceleration;
```

```
X2 = Displacement;
y = Weight;
```

Create the custom function `regf` (shown at the end of this example). This function fits a regression model to training data and then computes predicted car weights on a test set. The function compares the predicted car weight values to the true values, and then computes the mean absolute error (MAE) and the MAE adjusted to the range of the test set car weights.

Note: If you use the live script file for this example, the `regf` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

By default, `crossval` performs 10-fold cross-validation. For each of the 10 training and test set partitions of the data in `X1`, `X2`, and `y`, compute the MAE and adjusted MAE values using the `regf` function. Find the mean MAE and mean adjusted MAE.

```
rng('default') % For reproducibility
values = crossval(@regf,X1,X2,y)
```

```
values = 10×2

  319.2261    0.1132
  342.3722    0.1240
  214.3735    0.0902
  174.7247    0.1128
  189.4835    0.0832
  249.4359    0.1003
  194.4210    0.0845
  348.7437    0.1700
  283.1761    0.1187
  210.7444    0.1325
```

```
mean(values)
```

```
ans = 1×2

  252.6701    0.1129
```

This code creates the function `regf`.

```
function errors = regf(X1train,X2train,ytrain,X1test,X2test,ytest)
tbltrain = table(X1train,X2train,ytrain, ...
    'VariableNames',{'Acceleration','Displacement','Weight'});
tbltest = table(X1test,X2test,ytest, ...
    'VariableNames',{'Acceleration','Displacement','Weight'});
mdl = fitlm(tbltrain,'Weight ~ Acceleration + Displacement');
yfit = predict(mdl,tbltest);
MAE = mean(abs(yfit-tbltest.Weight));
adjMAE = MAE/range(tbltest.Weight);
errors = [MAE adjMAE];
end
```

Input Arguments

`tbl` — Input data

table | dataset array

Input data including predictor and response variables, specified as a table or dataset array. The predictor variables can be numeric, logical, categorical, character, or string. The response variable must be numeric or logical.

- By default, `fitlm` takes the last variable as the response variable and the others as the predictor variables.
- To set a different column as the response variable, use the `ResponseVar` name-value pair argument.
- To use a subset of the columns as predictors, use the `PredictorVars` name-value pair argument.
- To define a model specification, set the `modelspec` argument using a formula or terms matrix. The formula or terms matrix specifies which columns to use as the predictor or response variables.

The variable names in a table do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot specify `modelspec` using a formula.
- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiselm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

You can verify the variable names in `tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

`X` — Predictor variables

matrix

Predictor variables, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

By default, there is a constant term in the model, unless you explicitly remove it, so do not include a column of 1s in `X`.

Data Types: `single` | `double`

`y` — Response variable

vector

Response variable, specified as an *n*-by-1 vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`.

Data Types: `single` | `double` | `logical`

`modelspec` — Model specification

`'linear'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'y ~ terms'`

Model specification, specified as one of these values.

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and $x_1*x_2^2$ terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-1995, specifying terms in the model, where *t* is the number of terms and *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar “Formula” on page 33-1996 in the form
  `'y ~ terms'`,
  where the `terms` are in “Wilkinson Notation” on page 33-1996. The variable names in the formula must be variable names in `tbl` or variable names specified by `Varnames`. Also, the variable names must be valid MATLAB identifiers.

  The software determines the order of terms in a fitted model by using the order of terms in `tbl` or `X`. Therefore, the order of terms in the model can be different from the order of terms in the specified formula.

Example: `'quadratic'`

Example: `'y ~ x1 + x2^2 + x1:x2'`

Data Types: `single` | `double` | `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Intercept',false,'PredictorVars',[1,3],'ResponseVar',5,'RobustOpts','logistic'` specifies a robust regression model with no constant term, where the algorithm uses the logistic weighting function with the default tuning constant, first and third variables are the predictor variables, and fifth variable is the response variable.

`CategoricalVars` — Categorical variable list
string array | cell array of character vectors | logical or numeric index vector

Categorical variable list, specified as the comma-separated pair consisting of `'CategoricalVars'` and either a string array or cell array of character vectors containing categorical variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are categorical.

- If data is in a table or dataset array `tbl`, then, by default, `fitlm` treats all categorical values, logical values, character arrays, string arrays, and cell arrays of character vectors as categorical variables.
- If data is in matrix `X`, then the default value of `'CategoricalVars'` is an empty matrix `[]`. That is, no variable is categorical unless you specify it as categorical.

For example, you can specify the second and third variables out of six as categorical using either of the following:

Example: `'CategoricalVars',[2,3]`

Example: `'CategoricalVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`Exclude` — Observations to exclude
logical or numeric index vector

Observations to exclude from the fit, specified as the comma-separated pair consisting of `'Exclude'` and a logical or numeric index vector indicating which observations to exclude from the fit.

For example, you can exclude observations 2 and 3 out of 6 using either of the following examples.

Example: `'Exclude',[2,3]`

Example: `'Exclude',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical`

`Intercept` — Indicator for constant term
`true` (default) | `false`

Indicator for the constant term (intercept) in the fit, specified as the comma-separated pair consisting of `'Intercept'` and either `true` to include or `false` to remove the constant term from the model.

Use `'Intercept'` only when specifying the model using a character vector or string scalar, not a formula or matrix.

Example: `'Intercept',false`

`PredictorVars` — Predictor variables
string array | cell array of character vectors | logical or numeric index vector

Predictor variables to use in the fit, specified as the comma-separated pair consisting of `'PredictorVars'` and either a string array or cell array of character vectors of the variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are predictor variables.

The string values or character vectors should be among the names in `tbl`, or the names you specify using the `'VarNames'` name-value pair argument.

The default is all variables in `X`, or all variables in `tbl` except for `ResponseVar`.

For example, you can specify the second and third variables as the predictor variables using either of the following examples.

Example: `'PredictorVars',[2,3]`

Example: `'PredictorVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`ResponseVar` — Response variable
last column in `tbl` (default) | character vector or string scalar containing variable name | logical or numeric index vector

Response variable to use in the fit, specified as the comma-separated pair consisting of `'ResponseVar'` and either a character vector or string scalar containing the variable name in the table or dataset array `tbl`, or a logical or numeric index vector indicating which column is the response variable. You typically need to use `'ResponseVar'` when fitting a table or dataset array `tbl`.

For example, you can specify the fourth variable, say `yield`, as the response out of six variables, in one of the following ways.

Example: `'ResponseVar','yield'`

Example: `'ResponseVar',[4]`

Example: `'ResponseVar',logical([0 0 0 1 0 0])`

Data Types: `single` | `double` | `logical` | `char` | `string`

`RobustOpts` — Indicator of robust fitting type
`'off'` (default) | `'on'` | character vector | string scalar | structure

Indicator of the robust fitting type to use, specified as the comma-separated pair consisting of `'RobustOpts'` and one of these values.

- `'off'` — No robust fitting. `fitlm` uses ordinary least squares.
- `'on'` — Robust fitting using the `'bisquare'` weight function with the default tuning constant.
- Character vector or string scalar — Name of a robust fitting weight function from the following table. `fitlm` uses the corresponding default tuning constant specified in the table.
- Structure with the two fields `RobustWgtFun` and `Tune`.
  - The `RobustWgtFun` field contains the name of a robust fitting weight function from the following table or a function handle of a custom weight function.
  - The `Tune` field contains a tuning constant. If you do not set the `Tune` field, `fitlm` uses the corresponding default tuning constant.

Weight Function	Description	Default Tuning Constant
'andrews'	w = (abs(r)<pi) .* sin(r) ./ r	1.339
'bisquare'	w = (abs(r)<1) .* (1 - r.^2).^2 (also called biweight)	4.685
'cauchy'	w = 1 ./ (1 + r.^2)	2.385
'fair'	w = 1 ./ (1 + abs(r))	1.400
'huber'	w = 1 ./ max(1, abs(r))	1.345
'logistic'	w = tanh(r) ./ r	1.205
'ols'	Ordinary least squares (no weighting function)	None
'talwar'	w = 1 * (abs(r)<1)	2.795
'welsch'	w = exp(-(r.^2))	2.985
function handle	Custom weight function that accepts a vector r of scaled residuals, and returns a vector of weights the same size as r	1

- The default tuning constants of built-in weight functions give coefficient estimates that are approximately 95% as statistically efficient as the ordinary least-squares estimates, provided the response has a normal distribution with no outliers. Decreasing the tuning constant increases the downweight assigned to large residuals; increasing the tuning constant decreases the downweight assigned to large residuals.
- The value *r* in the weight functions is

  r = resid/(tune*s*sqrt(1—h)),

  where resid is the vector of residuals from the previous iteration, tune is the tuning constant, h is the vector of leverage values from a least-squares fit, and s is an estimate of the standard deviation of the error term given by

  s = MAD/0.6745.

  MAD is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution. If X has *p* columns, the software excludes the smallest *p* absolute deviations when computing the median.

For robust fitting, fitlm uses M-estimation to formulate estimating equations and solves them using the method of “Iteratively Reweighted Least Squares” on page 11-104 (IRLS).

Example: 'RobustOpts','andrews'

VarNames — Names of variables
{'x1','x2',...,'xn','y'} (default) | string array | cell array of character vectors

Names of variables, specified as the comma-separated pair consisting of 'VarNames' and a string array or cell array of character vectors including the names for the columns of X first, and the name for the response variable y last.

'VarNames' is not applicable to variables in a table or dataset array, because those variables already have names.

The variable names do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiselm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

Before specifying `'VarNames',varNames`, you can verify the variable names in `varNames` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Example: `'VarNames',{'Horsepower','Acceleration','Model_Year','MPG'}`

Data Types: `string` | `cell`

`Weights` — Observation weights
`ones(n,1)` (default) | *n*-by-1 vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations.

Data Types: `single` | `double`

Output Arguments

`mdl` — Linear model
`LinearModel` object

Linear model representing a least-squares fit of the response to the data, returned as a `LinearModel` object.

If the value of the `'RobustOpts'` name-value pair is not `[]` or `'ols'`, the model is not a least-squares fit, but uses the robust fitting function.

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'y ~ terms'`.

- *y* is the response name.
- *terms* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.
- `'y ~ x1 + x2 + x3 − 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include `−1` in the formula.

A formula includes a constant term unless you explicitly remove the term using `−1`.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- `+` means include the next variable.
- `−` means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- `*` defines an interaction and all lower-order terms.
- `^` raises the predictor to a power, exactly as in `*` repeated, so `^` includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
`1`	Constant (intercept) term
`x1^k`, where `k` is a positive integer	`x1`, $x1^2$, ..., $x1^k$
`x1 + x2`	`x1`, `x2`
`x1*x2`	`x1`, `x2`, `x1*x2`
`x1:x2`	`x1*x2` only
`−x2`	Do not include `x2`
`x1*x2 + x3`	`x1`, `x2`, `x3`, `x1*x2`

Wilkinson Notation	Terms in Standard Notation
`x1 + x2 + x3 + x1:x2`	x1, x2, x3, x1*x2
`x1*x2*x3 − x1:x2:x3`	x1, x2, x3, x1*x2, x1*x3, x2*x3
`x1*(x2 + x3)`	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Tips

- To access the model properties of the `LinearModel` object `mdl`, you can use dot notation. For example, `mdl.Residuals` returns a table of the raw, Pearson, Studentized, and standardized residual values for the model.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- The main fitting algorithm is QR decomposition. For robust fitting, `fitlm` uses M-estimation to formulate estimating equations and solves them using the method of “Iteratively Reweighted Least Squares” on page 11-104 (IRLS).
- `fitlm` treats a categorical predictor as follows:
  - A model with a categorical predictor that has L levels (categories) includes L – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `fitlm` treats the group of L – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.
- `fitlm` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `tbl`, `X`, and `Y` to be missing values. `fitlm` does not use observations with missing values in the fit. The `ObservationInfo` property of a fitted model indicates whether or not `fitlm` uses each observation in the fit.

Alternative Functionality

- For reduced computation time on high-dimensional data sets, fit a linear regression model using the `fitrlinear` function.
- To regularize a regression, use `fitrlinear`, `lasso`, `ridge`, or `plsregress`.
  - `fitrlinear` regularizes a regression for high-dimensional data sets using lasso or ridge regression.
  - `lasso` removes redundant predictors in linear regression using lasso or elastic net.
  - `ridge` regularizes a regression with correlated terms using ridge regression.
  - `plsregress` regularizes a regression with correlated terms using partial least squares.

References

[1] DuMouchel, W. H., and F. L. O'Brien. "Integrating a Robust Option into a Multiple Regression Computing Environment." *Computer Science and Statistics*: *Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[2] Holland, P. W., and R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares." *Communications in Statistics: Theory and Methods*, *A6*, 1977, pp. 813–827.

[3] Huber, P. J. *Robust Statistics*. Hoboken, NJ: John Wiley & Sons, Inc., 1981.

[4] Street, J. O., R. J. Carroll, and D. Ruppert. "A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares." *The American Statistician*. Vol. 42, 1988, pp. 152–154.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- If any input argument to `fitlm` is a tall array, then all of the other inputs must be tall arrays as well. This includes nonempty variables supplied with the `'Weights'` and `'Exclude'` name-value pairs.
- The `'RobustOpts'` name-value pair is not supported with tall arrays.
- For tall data, `fitlm` returns a `CompactLinearModel` object that contains most of the same properties as a `LinearModel` object. The main difference is that the compact object is sensitive to memory requirements. The compact object does not include properties that include the data, or that include an array of the same size as the data. The compact object does not contain these `LinearModel` properties:
  - `Diagnostics`
  - `Fitted`
  - `ObservationInfo`
  - `ObservationNames`
  - `Residuals`
  - `Steps`

- `Variables`

  You can compute the residuals directly from the compact object returned by `LM = fitlm(X,Y)` using

  ```
  RES = Y - predict(LM,X);
  S = LM.RMSE;
  histogram(RES,linspace(-3*S,3*S,51))
  ```

- If the `CompactLinearModel` object is missing lower order terms that include categorical factors:
  - The `plotEffects` and `plotInteraction` methods are not supported.
  - The `anova` method with the `'components'` option is not supported.

For more information, see “Tall Arrays for Out-of-Memory Data”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `fitrlinear` | `predict` | `stepwiselm`

Topics

“What Is a Linear Regression Model?” on page 11-6
“Linear Regression” on page 11-9
“Linear Regression Workflow” on page 11-35
“Train Linear Regression Model” on page 11-161
“Predict or Simulate Responses to New Data” on page 11-31
“Examine Quality and Adjust Fitted Model” on page 11-14
“Linear Regression with Categorical Covariates” on page 2-52
“Reduce Outlier Effects Using Robust Regression” on page 11-104
“Stepwise Regression” on page 11-99

Introduced in R2013b

fitlme

Fit linear mixed-effects model

Syntax

```
lme = fitlme(tbl,formula)
lme = fitlme(tbl,formula,Name,Value)
```

Description

`lme = fitlme(tbl,formula)` returns a linear mixed-effects model, specified by `formula`, fitted to the variables in the table or dataset array `tbl`.

`lme = fitlme(tbl,formula,Name,Value)` returns a linear mixed-effects model with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the covariance pattern of the random-effects terms, the method to use in estimating the parameters, or options for the optimization algorithm.

Examples

Fit Linear Mixed-Effects Model

Load the sample data.

```
load imports-85
```

Store the variables in a table.

```
tbl = table(X(:,12),X(:,14),X(:,24),'VariableNames',{'Horsepower','CityMPG','EngineType'});
```

Display the first five rows of the table.

```
tbl(1:5,:)
```

```
ans=5×3 table
    Horsepower    CityMPG    EngineType
    __________    _______    __________

       111          21           13
       111          21           13
       154          19           37
       102          24           35
       115          18           35
```

Fit a linear mixed-effects model for miles per gallon in the city, with fixed effects for horsepower, and uncorrelated random effect for intercept and horsepower grouped by the engine type.

```
lme = fitlme(tbl,'CityMPG~Horsepower+(1|EngineType)+(Horsepower-1|EngineType)');
```

In this model, `CityMPG` is the response variable, horsepower is the predictor variable, and engine type is the grouping variable. The fixed-effects portion of the model corresponds to `1 + Horsepower`, because the intercept is included by default.

Since the random-effect terms for intercept and horsepower are uncorrelated, these terms are specified separately. Because the second random-effect term is only for horsepower, you must include a `−1` to eliminate the intercept from the second random-effect term.

Display the model.

```
lme
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             203
    Fixed effects coefficients           2
    Random effects coefficients         14
    Covariance parameters                3

Formula:
    CityMPG ~ 1 + Horsepower + (1 | EngineType) + (Horsepower | EngineType)

Model fit statistics:
    AIC       BIC     LogLikelihood    Deviance
    1099.5    1116    -544.73          1089.5

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE         tStat     DF     pValue
    {'(Intercept)'}          37.276     2.8556    13.054    201    1.3147e-28
    {'Horsepower' }        -0.12631    0.02284     -5.53    201    9.8848e-08


    Lower       Upper
      31.645       42.906
    -0.17134    -0.081269

Random effects covariance parameters (95% CIs):
Group: EngineType (7 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        5.7338


    Lower     Upper
    2.3773    13.829

Group: EngineType (7 Levels)
    Name1                 Name2                 Type           Estimate
    {'Horsepower'}        {'Horsepower'}        {'std'}        0.050357


    Lower      Upper
    0.02307    0.10992

Group: Error
    Name               Estimate    Lower     Upper
```

```
    {'Res Std'}        3.226       2.9078    3.5789
```

Note that the random-effects covariance parameters for intercept and horsepower are separate in the display.

Now, fit a linear mixed-effects model for miles per gallon in the city, with the same fixed-effects term and potentially correlated random effect for intercept and horsepower grouped by the engine type.

```
lme2 = fitlme(tbl,'CityMPG~Horsepower+(Horsepower|EngineType)');
```

Because the random-effect term includes the intercept by default, you do not have to add `1`, the random effect term is equivalent to `(1 + Horsepower|EngineType)`.

Display the model.

```
lme2
```

```
lme2 =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             203
    Fixed effects coefficients           2
    Random effects coefficients         14
    Covariance parameters                4

Formula:
    CityMPG ~ 1 + Horsepower + (1 + Horsepower | EngineType)

Model fit statistics:
    AIC     BIC       LogLikelihood    Deviance
    1089    1108.9    -538.52          1077

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat      DF     pValue
    {'(Intercept)'}         33.824       4.0181     8.4178    201    7.1678e-15
    {'Horsepower' }        -0.1087     0.032912    -3.3029    201     0.0011328


    Lower      Upper
     25.901       41.747
    -0.1736    -0.043806

Random effects covariance parameters (95% CIs):
Group: EngineType (7 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std' }         9.4952
    {'Horsepower' }        {'(Intercept)'}        {'corr'}        -0.96843
    {'Horsepower' }        {'Horsepower' }        {'std' }        0.078874


    Lower       Upper
      4.7022      19.174
    -0.99568    -0.78738
    0.039917     0.15585

Group: Error
```

```
    Name               Estimate    Lower     Upper
    {'Res Std'}        3.1845      2.8774    3.5243
```

Note that the random effects covariance parameters for intercept and horsepower are together in the display, and it includes the correlation (`'corr'`) between the intercept and horsepower.

Fit Random Intercept LME Model

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the Centers for Disease Control and Prevention, CDC).

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the new response variable `FluRate`, the nominal variable `Region` that shows which region each estimate is from, the nationwide estimate `WtdILI`, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate', ...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Display the first six rows of `flu2`.

```
flu2(1:6,:)
```

```
ans =
    Date         WtdILI    Region       FluRate
    10/9/2005    1.182     NE            0.97
    10/9/2005    1.182     MidAtl       1.025
    10/9/2005    1.182     ENCentral    1.232
    10/9/2005    1.182     WNCentral    1.286
    10/9/2005    1.182     SAtl         1.082
    10/9/2005    1.182     ESCentral    1.457
```

Fit a linear mixed-effects model with a fixed-effects term for the nationwide estimate, `WtdILI`, and a random intercept that varies by `Date`. The model corresponds to

$$y_{im} = \beta_0 + \beta_1 \mathrm{WtdILI}_{im} + b_{0m} + \varepsilon_{im}, \quad i = 1, 2, \ldots, 468, \quad m = 1, 2, \ldots, 52,$$

where y_{im} is the observation i for level m of grouping variable `Date`, b_{0m} is the random effect for level m of the grouping variable `Date`, and ε_{im} is the observation error for observation i. The random effect has the prior distribution,

$$b_{0m} \sim N(0, \sigma_b^2),$$

and the error term has the distribution,

$\varepsilon_{im} \sim N(0, \sigma^2)$.

```
lme = fitlme(flu2,'FluRate ~ 1 + WtdILI + (1|Date)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             468
    Fixed effects coefficients           2
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    FluRate ~ 1 + WtdILI + (1 | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    286.24    302.83    -139.12          278.24

Fixed effects coefficients (95% CIs):
    Name                  Estimate    SE          tStat     DF     pValue
    {'(Intercept)'}       0.16385     0.057525    2.8484    466     0.0045885
    {'WtdILI'     }        0.7236     0.032219    22.459    466    3.0502e-76


    Lower       Upper
    0.050813    0.27689
     0.66028    0.78691

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       0.17146


    Lower      Upper
    0.13227    0.22226

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.30201     0.28217    0.32324
```

Estimated covariance parameters are displayed in the section titled "Random effects covariance parameters". The estimated value of σ_b is 0.17146 and its 95% confidence interval is [0.13227, 0.22226]. Since this interval does not include 0, the random-effects term is significant. You can formally test the significance of any random-effects term using a likelihood ratio test via the `compare` method.

The estimated response at an observation is the sum of the fixed effects and the random-effect value at the grouping variable level corresponding to that observation. For example, the estimated flu rate for observation 28 is

$$\begin{aligned}\hat{y}_{28} &= \hat{\beta}_0 + \hat{\beta}_1 \mathrm{WtdILI}_{28} + \hat{b}_{10/30/2005} \\ &= 0.1639 + 0.7236*(1.343) + 0.3318 \\ &= 1.46749,\end{aligned}$$

where $\hat{b}$ is the estimated best linear unbiased predictor (BLUP) of the random effects for the intercept. You can compute this value as follows.

```
beta = fixedEffects(lme);
[~,~,STATS] = randomEffects(lme); % Compute the random-effects statistics (STATS)
STATS.Level = nominal(STATS.Level);
y_hat = beta(1) + beta(2)*flu2.WtdILI(28) + STATS.Estimate(STATS.Level=='10/30/2005')
```

```
y_hat = 1.4674
```

You can display the fitted value using the `fitted` method.

```
F = fitted(lme);
F(28)
```

```
ans = 1.4674
```

LME Model for Randomized Block Design

Load the sample data.

```
load('shift.mat')
```

The data shows the absolute deviations from the target quality characteristic measured from the products each of five operators manufacture during three shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the absolute deviations of the quality characteristics from the target value. This is simulated data.

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if performance significantly differs according to the time of the shift. Use the restricted maximum likelihood method and `'effects'` contrasts.

`'effects'` contrasts mean that the coefficients sum to 0, and `fitlme` creates a matrix called a *fixed effects design matrix* to describe the effect of shift. This matrix has two columns, Shift_Evening and Shift_Morning, where

$$\text{Shift\_Evening} = \begin{cases} 0, & \text{if Morning} \\ 1, & \text{if Evening} \\ -1, & \text{if Night} \end{cases}$$

$$\text{Shift\_Morning} = \begin{cases} 1, & \text{if Morning} \\ 0, & \text{if Evening} \\ -1, & \text{if Night} \end{cases}$$

The model corresponds to

$$\text{Morning Shift:} \quad \text{QCDev}_{im} = \beta_0 + \beta_2 \text{Shift\_Morning}_i + b_{0m} + \epsilon_{im},$$
$$\text{Evening Shift:} \quad \text{QCDev}_{im} = \beta_0 + \beta_1 \text{Shift\_Evening}_i + b_{0m} + \epsilon_{im},$$
$$\text{Night Shift:} \quad \text{QCDev}_{im} = \beta_0 - \beta_1 \text{Shift\_Evening}_i - \beta_2 \text{Shift\_Morning}_i + b_{0m} + \epsilon_{im},$$

where i represents the observations, and m represents the operators, i = 1, 2, ..., 15, and m = 1, 2, ..., 5. The random effects and the observation error have the following distributions:

$$b_{0m} \sim N(0, \sigma_b^2)$$

and

$$\varepsilon_{im} \sim N(0, \sigma^2).$$

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)',...
'FitMethod','REML','DummyVarCoding','effects')
```

```
lme =
Linear mixed-effects model fit by REML

Model information:
    Number of observations             15
    Fixed effects coefficients          3
    Random effects coefficients         5
    Covariance parameters               2

Formula:
    QCDev ~ 1 + Shift + (1 | Operator)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    58.913    61.337    -24.456          48.913

Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE         tStat      DF    pValue
    {'(Intercept)'  }         3.6525    0.94109     3.8812    12    0.0021832
    {'Shift_Evening'}        -0.53293   0.31206    -1.7078    12      0.11339
    {'Shift_Morning'}        -0.91973   0.31206    -2.9473    12     0.012206


    Lower      Upper
     1.6021       5.703
    -1.2129     0.14699
    -1.5997    -0.23981

Random effects covariance parameters (95% CIs):
Group: Operator (5 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       2.0457


    Lower      Upper
    0.98207    4.2612

Group: Error
    Name              Estimate    Lower      Upper
```

```
    {'Res Std'}        0.85462     0.52357    1.395
```

Compute the best linear unbiased predictor (BLUP) estimates of random effects.

```
B = randomEffects(lme)
```

```
B = 5×1

    0.5775
    1.1757
   -2.1715
    2.3655
   -1.9472
```

The estimated absolute deviation from the target quality characteristics for the third operator working the evening shift is

$$\begin{aligned}\hat{y}_{\text{Evening, Operator3}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{Shift\_Evening} + \hat{b}_{03} \\ &= 3.6525 - 0.53293 - 2.1715 \\ &= 0.94807.\end{aligned}$$

You can also display this value as follows.

```
F = fitted(lme);
F(shift.Shift=='Evening' & shift.Operator=='3')
```

```
ans = 0.9481
```

Similarly, you can calculate the estimated absolute deviation from the target quality characteristics for the third operator working the morning shift as

$$\begin{aligned}\hat{y}_{\text{Morning, Operator3}} &= \hat{\beta}_0 + \hat{\beta}_2 \text{Shift\_Morning} + \hat{b}_{03} \\ &= 3.6525 - 0.91973 - 2.1715 \\ &= 0.56127.\end{aligned}$$

You can also display this value as follows.

```
F(shift.Shift=='Morning' & shift.Operator=='3')
```

```
ans = 0.5613
```

The operator tends to make a smaller magnitude of error during the morning shift.

LME Model for Split-Plot Experiment

Load the sample data.

```
load('fertilizer.mat')
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five types of

tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type) and the plots within blocks (tomato types within soil types) independently.

This model corresponds to

$$y_{imjk} = \beta_0 + \sum_{m=2}^{4} \beta_{1m} I[F]_{im} + \sum_{j=2}^{5} \beta_{2j} I[T]_{ij} + \sum_{j=2}^{5} \sum_{m=2}^{4} \beta_{3mj} I[F]_{im} I[T]_{ij} + b_{0k} S_k + b_{0jk} (S*T)_{jk} + \varepsilon_{imjk},$$

where $i = 1, 2, \ldots, 60$, index m corresponds to the fertilizer types, j corresponds to the tomato types, and $k = 1, 2, 3$ corresponds to the blocks (soil). S_k represents the k th soil type, and $(S*T)_{jk}$ represents the j th tomato type nested in the k th soil type. $I[F]_{im}$ is the dummy variable representing level m of the fertilizer. Similarly, $I[T]_{ij}$ is the dummy variable representing level j of the tomato type.

The random effects and observation error have these prior distributions: $b_{0k} \sim N(0, \sigma_S^2)$, $b_{0jk} \sim N(0, \sigma_{S*T}^2)$, and $\epsilon_{imjk} \sim N(0, \sigma^2)$.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             60
    Fixed effects coefficients         20
    Random effects coefficients        18
    Covariance parameters               3

Formula:
    Yield ~ 1 + Tomato*Fertilizer + (1 | Soil) + (1 | Soil:Tomato)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    522.57    570.74    -238.29          476.57

Fixed effects coefficients (95% CIs):
    Name                                  Estimate    SE        tStat       DF
    {'(Intercept)'                }            77     8.5836      8.9706    40
    {'Tomato_Grape'               }           -16     11.966     -1.3371    40
    {'Tomato_Heirloom'            }        -6.6667     11.966    -0.55714    40
    {'Tomato_Plum'                }         32.333     11.966      2.7022    40
```

```
    {'Tomato_Vine'                 }           -13     11.966     -1.0864    40
    {'Fertilizer_2'                }        34.667      8.572      4.0442    40
    {'Fertilizer_3'                }        33.667      8.572      3.9275    40
    {'Fertilizer_4'                }        47.667      8.572      5.5607    40
    {'Tomato_Grape:Fertilizer_2'   }       -2.6667     12.123    -0.21997    40
    {'Tomato_Heirloom:Fertilizer_2'}           -8      12.123    -0.65992    40
    {'Tomato_Plum:Fertilizer_2'    }           -15     12.123     -1.2374    40
    {'Tomato_Vine:Fertilizer_2'    }           -16     12.123     -1.3198    40
    {'Tomato_Grape:Fertilizer_3'   }        16.667     12.123      1.3748    40
    {'Tomato_Heirloom:Fertilizer_3'}        3.3333     12.123     0.27497    40
    {'Tomato_Plum:Fertilizer_3'    }        3.6667     12.123     0.30246    40
    {'Tomato_Vine:Fertilizer_3'    }             3     12.123     0.24747    40
    {'Tomato_Grape:Fertilizer_4'   }        13.333     12.123      1.0999    40
    {'Tomato_Heirloom:Fertilizer_4'}           -19     12.123     -1.5673    40
    {'Tomato_Plum:Fertilizer_4'    }       -2.6667     12.123    -0.21997    40
    {'Tomato_Vine:Fertilizer_4'    }        8.6667     12.123     0.71492    40


    pValue        Lower     Upper
    4.0206e-11     59.652    94.348
       0.18873    -40.184    8.1837
       0.58053     -30.85    17.517
      0.010059     8.1496    56.517
       0.28379    -37.184    11.184
    0.00023272     17.342    51.991
    0.00033057     16.342    50.991
    1.9567e-06     30.342    64.991
       0.82701    -27.167    21.834
       0.51309    -32.501    16.501
       0.22317    -39.501    9.5007
       0.19439    -40.501    8.5007
       0.17683    -7.8341    41.167
       0.78476    -21.167    27.834
       0.76387    -20.834    28.167
       0.80581    -21.501    27.501
       0.27796    -11.167    37.834
       0.12492    -43.501    5.5007
       0.82701    -27.167    21.834
       0.47881    -15.834    33.167

Random effects covariance parameters (95% CIs):
Group: Soil (3 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       2.5028


    Lower       Upper
    0.027711    226.05

Group: Soil:Tomato (15 Levels)
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       10.225


    Lower     Upper
    6.1497    17.001

Group: Error
```

```
    Name               Estimate    Lower     Upper
    {'Res Std'}        10.499      8.5389    12.908
```

The *p*-values corresponding to the last 12 rows in the fixed-effects coefficients display (0.82701 to 0.47881) indicate that interaction coefficients between the tomato and fertilizer types are not significant. To test for the overall interaction between tomato and fertilizer, use the `anova` method after refitting the model using `'effects'` contrasts.

The confidence interval for the standard deviations of the random-effects terms (σ_S^2), where the intercept is grouped by soil, is very large. This term does not appear significant.

Refit the model after removing the interaction term `Tomato:Fertilizer` and the random-effects term `(1 | Soil)`.

```
lme = fitlme(ds,'Yield ~ Fertilizer + Tomato + (1|Soil:Tomato)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             60
    Fixed effects coefficients          8
    Random effects coefficients        15
    Covariance parameters               2

Formula:
    Yield ~ 1 + Tomato + Fertilizer + (1 | Soil:Tomato)

Model fit statistics:
    AIC       BIC    LogLikelihood    Deviance
    511.06    532    -245.53          491.06

Fixed effects coefficients (95% CIs):
    Name                       Estimate    SE        tStat       DF
    {'(Intercept)'    }         77.733     7.3293      10.606    52
    {'Tomato_Grape'   }        -9.1667     9.6045    -0.95441    52
    {'Tomato_Heirloom'}        -12.583     9.6045     -1.3102    52
    {'Tomato_Plum'    }         28.833     9.6045      3.0021    52
    {'Tomato_Vine'    }        -14.083     9.6045     -1.4663    52
    {'Fertilizer_2'   }         26.333     4.5004      5.8514    52
    {'Fertilizer_3'   }             39     4.5004      8.6659    52
    {'Fertilizer_4'   }         47.733     4.5004      10.607    52


    pValue        Lower      Upper
    1.3108e-14     63.026    92.441
       0.34429    -28.439    10.106
        0.1959    -31.856    6.6895
     0.0041138     9.5605    48.106
       0.14858    -33.356    5.1895
    3.3024e-07     17.303    35.364
    1.1459e-11     29.969    48.031
     1.308e-14     38.703    56.764

Random effects covariance parameters (95% CIs):
Group: Soil:Tomato (15 Levels)
```

```
    Name1                 Name2                 Type          Estimate
    {'(Intercept)'}       {'(Intercept)'}       {'std'}       10.02


    Lower     Upper
    6.0812    16.509

Group: Error
    Name              Estimate    Lower     Upper
    {'Res Std'}       12.325      10.024    15.153
```

You can compare the two models using the `compare` method with the simulated likelihood ratio test since both a fixed-effect and a random-effect term are tested.

Longitudinal Study with a Covariate

Load the sample data.

```
load('weight.mat')
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs (A, B, C, D), and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

`fitlme` uses program A as a reference and creates the necessary dummy variables $I[.]$. Since the model already has an intercept, `fitlme` only creates dummy variables for programs B, C, and D. This is also known as the `'reference'` method of coding dummy variables.

This model corresponds to

$$\begin{aligned} y_{im} = {} & \beta_0 + \beta_1 \mathrm{IW}_i + \beta_2 \mathrm{Week}_i + \beta_3 I[\mathrm{PB}]_i + \beta_4 I[\mathrm{PC}]_i + \beta_5 I[\mathrm{PD}]_i \\ & + \beta_6(\mathrm{Week}_i * I[\mathrm{PB}]_i) + \beta_7(\mathrm{Week}_i * I[\mathrm{PC}]_i) + \beta_8(\mathrm{Week}_i * I[\mathrm{PD}]_i) \\ & + b_{0m} + b_{1m}\mathrm{Week}_{im} + \varepsilon_{im}, \end{aligned}$$

where $i = 1, 2, \ldots, 120$, and $m = 1, 2, \ldots, 20$. β_j are the fixed-effects coefficients, $j = 0, 1, \ldots, 8$, and b_{0m} and b_{1m} are random effects. IW stands for initial weight and $I[\cdot]$ is a dummy variable representing a type of program. For example, $I[\mathrm{PB}]_i$ is the dummy variable representing program type B. The random effects and observation error have the following prior distributions:

$$b_{0m} \sim N(0, \sigma_0^2)$$

$$b_{1m} \sim N(0, \sigma_1^2)$$

$\varepsilon_{im} \sim N(0, \sigma^2)$.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             120
    Fixed effects coefficients           9
    Random effects coefficients         40
    Covariance parameters                4

Formula:
    y ~ 1 + InitialWeight + Program*Week + (1 + Week | Subject)

Model fit statistics:
    AIC        BIC       LogLikelihood    Deviance
    -22.981    13.257    24.49            -48.981

Fixed effects coefficients (95% CIs):
    Name                      Estimate     SE           tStat       DF
    {'(Intercept)'   }          0.66105      0.25892      2.5531    111
    {'InitialWeight' }        0.0031879    0.0013814      2.3078    111
    {'Program_B'     }          0.36079      0.13139       2.746    111
    {'Program_C'     }        -0.033263      0.13117    -0.25358    111
    {'Program_D'     }          0.11317      0.13132     0.86175    111
    {'Week'          }           0.1732     0.067454      2.5677    111
    {'Program_B:Week'}         0.038771     0.095394     0.40644    111
    {'Program_C:Week'}         0.030543     0.095394     0.32018    111
    {'Program_D:Week'}         0.033114     0.095394     0.34713    111


    pValue       Lower         Upper
     0.012034       0.14798       1.1741
     0.022863    0.00045067    0.0059252
    0.0070394       0.10044      0.62113
      0.80029      -0.29319      0.22666
      0.39068      -0.14706       0.3734
     0.011567      0.039536      0.30686
      0.68521      -0.15026       0.2278
      0.74944      -0.15849      0.21957
      0.72915      -0.15592      0.22214

Random effects covariance parameters (95% CIs):
Group: Subject (20 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std' }       0.18407
    {'Week'       }        {'(Intercept)'}        {'corr'}       0.66841
    {'Week'       }        {'Week'       }        {'std' }       0.15033


    Lower      Upper
    0.12281    0.27587
    0.21076    0.88573
    0.11004    0.20537

Group: Error
```

```
    Name               Estimate    Lower       Upper
    {'Res Std'}        0.10261     0.087882    0.11981
```

The *p*-values 0.022863 and 0.011567 indicate significant effects of subject initial weights and time in the amount of weight lost. The weight loss of subjects who are in program B is significantly different relative to the weight loss of subjects who are in program A. The lower and upper limits of the covariance parameters for the random effects do not include 0, thus they are significant. You can also test the significance of the random effects using the `compare` method.

Input Arguments

`tbl` — Input data

table | `dataset` array

Input data, which includes the response variable, predictor variables, and grouping variables, specified as a table or `dataset` array. The predictor variables can be continuous or grouping variables (see “Grouping Variables” on page 2-45). You must specify the model for the variables using `formula`.

Data Types: `table`

`formula` — Formula for model specification

character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`

Formula for model specification, specified as a character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`. The formula is case sensitive. For a full description, see “Formula” on page 33-2018.

Example: `'y ~ treatment + (1|block)'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'CovariancePattern','Diagonal','Optimizer','fminunc','OptimizerOptions',opt` specifies a model, where the random-effects terms have a diagonal covariance matrix structure, and `fitlme` uses the `fminunc` optimization algorithm with the custom optimization parameters defined in variable `opt`.

`CovariancePattern` — Pattern of covariance matrix

`'FullCholesky'` (default) | character vector | string scalar | square symmetric logical matrix | string array | cell array of character vectors or logical matrices

Pattern of the covariance matrix of the random effects, specified as the comma-separated pair consisting of `'CovariancePattern'` and a character vector, a string scalar, a square symmetric logical matrix, a string array, or a cell array of character vectors or logical matrices.

If there are *R* random-effects terms, then the value of `'CovariancePattern'` must be a string array or cell array of length *R*, where each element *r* of the array specifies the pattern of the covariance

matrix of the random-effects vector associated with the *r*th random-effects term. The options for each element follow.

`'FullCholesky'`	Default. Full covariance matrix using the Cholesky parameterization. `fitlme` estimates all elements of the covariance matrix.
`'Full'`	Full covariance matrix, using the log-Cholesky parameterization. `fitlme` estimates all elements of the covariance matrix.
`'Diagonal'`	Diagonal covariance matrix. That is, off-diagonal elements of the covariance matrix are constrained to be 0. $\begin{pmatrix} \sigma_{b1}^2 & 0 & 0 \\ 0 & \sigma_{b2}^2 & 0 \\ 0 & 0 & \sigma_{b3}^2 \end{pmatrix}$
`'Isotropic'`	Diagonal covariance matrix with equal variances. That is, off-diagonal elements of the covariance matrix are constrained to be 0, and the diagonal elements are constrained to be equal. For example, if there are three random-effects terms with an isotropic covariance structure, this covariance matrix looks like $\begin{pmatrix} \sigma_b^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_b^2 \end{pmatrix}$ where σ^2_b is the common variance of the random-effects terms.
`'CompSymm'`	Compound symmetry structure. That is, common variance along diagonals and equal correlation between all random effects. For example, if there are three random-effects terms with a covariance matrix having a compound symmetry structure, this covariance matrix looks like $\begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1,b2} & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1}^2 & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1,b2} & \sigma_{b1}^2 \end{pmatrix}$ where σ^2_{b1} is the common variance of the random-effects terms and $\sigma_{b1,b2}$ is the common covariance between any two random-effects term .

`PAT`	Square symmetric logical matrix. If `'CovariancePattern'` is defined by the matrix `PAT`, and if `PAT(a,b) = false`, then the `(a,b)` element of the corresponding covariance matrix is constrained to be 0.

Example: `'CovariancePattern','Diagonal'`

Example: `'CovariancePattern',{'Full','Diagonal'}`

Data Types: `char` | `string` | `logical` | `cell`

`FitMethod` — Method for estimating parameters
`'ML'` (default) | `'REML'`

Method for estimating parameters of the linear mixed-effects model, specified as the comma-separated pair consisting of `'FitMethod'` and either of the following.

`'ML'`	Default. Maximum likelihood estimation
`'REML'`	Restricted maximum likelihood estimation

Example: `'FitMethod','REML'`

`Weights` — Observation weights
vector of scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of length n, where n is the number of observations.

Data Types: `single` | `double`

`Exclude` — Indices for rows to exclude
use all rows without `NaNs` (default) | vector of integer or logical values

Indices for rows to exclude from the linear mixed-effects model in the data, specified as the comma-separated pair consisting of `'Exclude'` and a vector of integer or logical values.

For example, you can exclude the 13th and 67th rows from the fit as follows.

Example: `'Exclude',[13,67]`

Data Types: `single` | `double` | `logical`

`DummyVarCoding` — Coding to use for dummy variables
`'reference'` (default) | `'effects'` | `'full'`

Coding to use for dummy variables created from the categorical variables, specified as the comma-separated pair consisting of `'DummyVarCoding'` and one of the variables in this table.

Value	Description
`'reference'` (default)	`fitlme` creates dummy variables with a reference group. This scheme treats the first category as a reference group and creates one less dummy variables than the number of categories. You can check the category order of a categorical variable by using the `categories` function, and change the order by using the `reordercats` function.
`'effects'`	`fitlme` creates dummy variables using effects coding. This scheme uses –1 to represent the last category. This scheme creates one less dummy variables than the number of categories.
`'full'`	`fitlme` creates full dummy variables. This scheme creates one dummy variable for each category.

For more details about creating dummy variables, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'DummyVarCoding','effects'`

`Optimizer` — Optimization algorithm
`'quasinewton'` (default) | `'fminunc'`

Optimization algorithm, specified as the comma-separated pair consisting of `'Optimizer'` and either of the following.

`'quasinewton'`	Default. Uses a trust region based quasi-Newton optimizer. Change the options of the algorithm using `statset('LinearMixedModel')`. If you don’t specify the options, then `LinearMixedModel` uses the default options of `statset('LinearMixedModel')`.
`'fminunc'`	You must have Optimization Toolbox to specify this option. Change the options of the algorithm using `optimoptions('fminunc')`. If you don’t specify the options, then `LinearMixedModel` uses the default options of `optimoptions('fminunc')` with `'Algorithm'` set to `'quasi-newton'`.

Example: `'Optimizer','fminunc'`

`OptimizerOptions` — Options for optimization algorithm
structure returned by `statset` | object returned by `optimoptions`

Options for the optimization algorithm, specified as the comma-separated pair consisting of `'OptimizerOptions'` and a structure returned by `statset('LinearMixedModel')` or an object returned by `optimoptions('fminunc')`.

- If `'Optimizer'` is `'fminunc'`, then use `optimoptions('fminunc')` to change the options of the optimization algorithm. See `optimoptions` for the options `'fminunc'` uses. If `'Optimizer'`

is `'fminunc'` and you do not supply `'OptimizerOptions'`, then the default for `LinearMixedModel` is the default options created by `optimoptions('fminunc')` with `'Algorithm'` set to `'quasi-newton'`.

- If `'Optimizer'` is `'quasinewton'`, then use `statset('LinearMixedModel')` to change the optimization parameters. If you don't change the optimization parameters, then `LinearMixedModel` uses the default options created by `statset('LinearMixedModel')`:

The `'quasinewton'` optimizer uses the following fields in the structure created by `statset('LinearMixedModel')`.

`TolFun` — Relative tolerance on gradient of objective function
`1e-6` (default) | positive scalar value

Relative tolerance on the gradient of the objective function, specified as a positive scalar value.

`TolX` — Absolute tolerance on step size
`1e-12` (default) | positive scalar value

Absolute tolerance on the step size, specified as a positive scalar value.

`MaxIter` — Maximum number of iterations allowed
`10000` (default) | positive scalar value

Maximum number of iterations allowed, specified as a positive scalar value.

`Display` — Level of display
`'off'` (default) | `'iter'` | `'final'`

Level of display, specified as one of `'off'`, `'iter'`, or `'final'`.

`StartMethod` — Method to start iterative optimization
`'default'` (default) | `'random'`

Method to start iterative optimization, specified as the comma-separated pair consisting of `'StartMethod'` and either of the following.

Value	Description
`'default'`	An internally defined default value
`'random'`	A random initial value

Example: `'StartMethod','random'`

`Verbose` — Indicator to display optimization process on screen
`false` (default) | `true`

Indicator to display the optimization process on screen, specified as the comma-separated pair consisting of `'Verbose'` and either `false` or `true`. Default is `false`.

The setting for `'Verbose'` overrides the field `'Display'` in `'OptimizerOptions'`.

Example: `'Verbose',true`

`CheckHessian` — Indicator to check positive definiteness of Hessian
`false` (default) | `true`

Indicator to check the positive definiteness of the Hessian of the objective function with respect to unconstrained parameters at convergence, specified as the comma-separated pair consisting of `'CheckHessian'` and either `false` or `true`. Default is `false`.

Specify `'CheckHessian'` as `true` to verify optimality of the solution or to determine if the model is overparameterized in the number of covariance parameters.

Example: `'CheckHessian',true`

Output Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, returned as a `LinearMixedModel` object.

More About

Formula

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For the linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms.

Suppose a table `tbl` contains the following:

- A response variable, `y`
- Predictor variables, X_j, which can be continuous or grouping variables
- Grouping variables, g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form, `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random`$_1$ is a specification of the random-effects design matrix Z_1 corresponding to grouping variable g_1, and similarly `random`$_R$ is a specification of the random-effects design matrix Z_R corresponding to grouping variable g_R. You can express the `fixed` and `random` terms using Wilkinson notation.

Wilkinson notation describes the factors present in models. The notation relates to factors present in models, not to the multipliers (coefficients) of those factors.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	`X`, X^2, ..., X^k
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (elementwise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only

Wilkinson Notation	Factors in Standard Notation
`- X2`	Do not include X2
`X1*X2 + X3`	X1, X2, X3, X1*X2
`X1 + X2 + X3 + X1:X2`	X1, X2, X3, X1*X2
`X1*X2*X3 - X1:X2:X3`	X1, X2, X3, X1*X2, X1*X3, X2*X3
`X1*(X2 + X3)`	X1, X2, X3, X1*X2, X1*X3

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`. Here are some examples for linear mixed-effects model specification.

Examples:

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, X1 and X2. This is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept and fixed effects for X1 and X2. The implicit intercept term is suppressed by including `-1`.
`'y ~ 1 + (1 | g1)'`	Fixed effects for the intercept plus random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

Cholesky Parameterization

One of the assumptions of linear mixed-effects models is that the random effects have the following prior distribution.

$$b \sim N\left(0, \sigma^2 D(\theta)\right),$$

where D is a q-by-q symmetric and positive semidefinite matrix, parameterized by a variance component vector θ, q is the number of variables in the random-effects term, and σ^2 is the observation error variance. Since the covariance matrix of the random effects, D, is symmetric, it has $q(q+1)/2$ free parameters. Suppose L is the lower triangular Cholesky factor of $D(\theta)$ such that

$$D(\theta) = L(\theta)L(\theta)^T,$$

then the $q*(q+1)/2$-by-1 unconstrained parameter vector θ is formed from elements in the lower triangular part of L.

For example, if

$$L = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix},$$

then

$$\theta = \begin{bmatrix} L_{11} \\ L_{21} \\ L_{31} \\ L_{22} \\ L_{32} \\ L_{33} \end{bmatrix}.$$

Log-Cholesky Parameterization

When the diagonal elements of L in Cholesky parameterization are constrained to be positive, then the solution for L is unique. Log-Cholesky parameterization is the same as Cholesky parameterization except that the logarithm of the diagonal elements of L are used to guarantee unique parameterization.

For example, for the 3-by-3 example in Cholesky parameterization, enforcing $L_{ii} \geq 0$,

$$\theta = \begin{bmatrix} \log(L_{11}) \\ L_{21} \\ L_{31} \\ \log(L_{22}) \\ L_{32} \\ \log(L_{33}) \end{bmatrix}.$$

Alternatives

If your model is not easily described using a formula, you can create matrices to define the fixed and random effects, and fit the model using `fitlmematrix(X,y,Z,G)`.

References

[1] Pinherio, J. C., and D. M. Bates. “Unconstrained Parametrizations for Variance-Covariance Matrices”. *Statistics and Computing*, Vol. 6, 1996, pp. 289–296.

See Also

`LinearMixedModel` | `fitlmematrix`

Topics

“Relationship Between Formula and Design Matrices” on page 11-138

Introduced in R2013b

fitlmematrix

Fit linear mixed-effects model

Syntax

```
lme = fitlmematrix(X,y,Z,[])
lme = fitlmematrix(X,y,Z,G)
lme = fitlmematrix( ___ ,Name,Value)
```

Description

`lme = fitlmematrix(X,y,Z,[])` creates a linear mixed-effects model of the responses `y` using the fixed-effects design matrix `X` and random-effects design matrix or matrices in `Z`.

`[]` implies that there is one group. That is, the grouping variable `G` is `ones(n,1)`, where *n* is the number of observations. Using `fitlmematrix(X,Y,Z,[])` without a specified covariance pattern most likely results in a nonidentifiable model. This syntax is recommended only if you build the grouping information into the random effects design `Z` and specify a covariance pattern for the random effects using the `'CovariancePattern'` name-value pair argument.

`lme = fitlmematrix(X,y,Z,G)` creates a linear mixed-effects model of the responses `y` using the fixed-effects design matrix `X` and random-effects design matrix `Z` or matrices in `Z`, and the grouping variable or variables in `G`.

`lme = fitlmematrix( ___ ,Name,Value)` also creates a linear mixed-effects model with additional options specified by one or more `Name,Value` pair arguments, using any of the previous input arguments.

For example, you can specify the names of the response, predictor, and grouping variables. You can also specify the covariance pattern, fitting method, or the optimization algorithm.

Examples

No Grouping Variable Specified

Load the sample data.

```
load carsmall
```

Fit a linear mixed-effects model, where miles per gallon (MPG) is the response, weight is the predictor variable, and the intercept varies by model year. First, define the design matrices. Then, fit the model using the specified design matrices.

```
y = MPG;
X = [ones(size(Weight)), Weight];
Z = ones(size(y));
lme = fitlmematrix(X,y,Z,Model_Year)
```

```
lme =
Linear mixed-effects model fit by ML
```

```
Model information:
    Number of observations             94
    Fixed effects coefficients          2
    Random effects coefficients         3
    Covariance parameters               2

Formula:
    y ~ x1 + x2 + (z11 | g1)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    486.09    496.26    -239.04          478.09

Fixed effects coefficients (95% CIs):
    Name          Estimate      SE           tStat      DF    pValue
    {'x1'}            43.575       2.3038     18.915    92    1.8371e-33
    {'x2'}        -0.0067097    0.0004242    -15.817    92    5.5373e-28


    Lower         Upper
            39        48.151
    -0.0075522    -0.0058672

Random effects covariance parameters (95% CIs):
Group: g1 (3 Levels)
    Name1          Name2          Type           Estimate    Lower     Upper
    {'z11'}        {'z11'}        {'std'}        3.301       1.4448    7.5421

Group: Error
    Name              Estimate    Lower     Upper
    {'Res Std'}       2.8997      2.5075    3.3532
```

Now, fit the same model by building the grouping into the Z matrix.

```
Z = double([Model_Year==70, Model_Year==76, Model_Year==82]);
lme = fitlmematrix(X,y,Z,[],'Covariancepattern','Isotropic')

lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             94
    Fixed effects coefficients          2
    Random effects coefficients         3
    Covariance parameters               2

Formula:
    y ~ x1 + x2 + (z11 + z12 + z13 | g1)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    486.09    496.26    -239.04          478.09

Fixed effects coefficients (95% CIs):
    Name          Estimate      SE           tStat      DF    pValue
    {'x1'}            43.575       2.3038     18.915    92    1.8371e-33
```

```
    {'x2'}        -0.0067097    0.0004242    -15.817    92    5.5373e-28


    Lower         Upper
            39        48.151
    -0.0075522    -0.0058672

Random effects covariance parameters (95% CIs):
Group: g1 (1 Levels)
    Name1          Name2          Type           Estimate    Lower     Upper
    {'z11'}        {'z11'}        {'std'}        3.301       1.4448    7.5421

Group: Error
    Name               Estimate    Lower     Upper
    {'Res Std'}        2.8997      2.5075    3.3532
```

Longitudinal Study with a Covariate

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned 4 exercise programs (A, B, C, D) and their weight loss is recorded over six 2-week time periods. This is simulated data.

Define `Subject` and `Program` as categorical variables. Create the design matrices for a linear mixed-effects model, with the initial weight, type of program, week, and the interaction between the week and type of program as the fixed effects. The intercept and coefficient of week vary by subject.

This model corresponds to

$$\begin{aligned} y_{im} = {} & \beta_0 + \beta_1 \mathrm{IW}_i + \beta_2 \mathrm{Week}_i + \beta_3 I[\mathrm{PB}]_i + \beta_4 I[\mathrm{PC}]_i + \beta_5 I[\mathrm{PD}]_i \\ & + \beta_6 (\mathrm{Week}_i * I[\mathrm{PB}]_i) + \beta_7 (\mathrm{Week}_i * I[\mathrm{PC}]_i) + \beta_8 (\mathrm{Week}_i * I[\mathrm{PD}]_i) \\ & + b_{0m} + b_{1m} \mathrm{Week}_{im} + \varepsilon_{im}, \end{aligned}$$

where $i = 1, 2, \ldots, 120$, and $m = 1, 2, \ldots, 20$. β_j are the fixed-effects coefficients, $j = 0, 1, \ldots, 8$, and b_{0m} and b_{1m} are random effects. IW stands for initial weight and $I[\cdot]$ is a dummy variable representing a type of program. For example, $I[\mathrm{PB}]_i$ is the dummy variable representing program type B. The random effects and observation error have the following prior distributions:

$$b_{0m} \sim N(0, \sigma_0^2)$$

$$b_{1m} \sim N(0, \sigma_1^2)$$

$$\varepsilon_{im} \sim N(0, \sigma^2).$$

```
Subject = nominal(Subject);
Program = nominal(Program);
D = dummyvar(Program); % Create dummy variables for Program
```

```
X = [ones(120,1), InitialWeight, D(:,2:4), Week,...
        D(:,2).*Week, D(:,3).*Week, D(:,4).*Week];
Z = [ones(120,1), Week];
G = Subject;
```

Since the model has an intercept, you only need the dummy variables for programs B, C, and D. This is also known as the `'reference'` method of coding dummy variables.

Fit the model using `fitlmematrix` with the defined design matrices and grouping variables.

```
lme = fitlmematrix(X,y,Z,G,'FixedEffectPredictors',...
{'Intercept','InitWeight','PrgB','PrgC','PrgD','Week','Week_PrgB','Week_PrgC','Week_PrgD'},...
'RandomEffectPredictors',{{'Intercept','Week'}},'RandomEffectGroups',{'Subject'})
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             120
    Fixed effects coefficients           9
    Random effects coefficients         40
    Covariance parameters                4

Formula:
    Linear Mixed Formula with 10 predictors.

Model fit statistics:
    AIC        BIC       LogLikelihood    Deviance
    -22.981    13.257    24.49            -48.981

Fixed effects coefficients (95% CIs):
    Name                  Estimate     SE           tStat       DF     pValue
    {'Intercept' }          0.66105      0.25892      2.5531    111     0.012034
    {'InitWeight'}        0.0031879    0.0013814      2.3078    111     0.022863
    {'PrgB'      }          0.36079      0.13139       2.746    111    0.0070394
    {'PrgC'      }        -0.033263      0.13117    -0.25358    111      0.80029
    {'PrgD'      }          0.11317      0.13132     0.86175    111      0.39068
    {'Week'      }           0.1732     0.067454      2.5677    111     0.011567
    {'Week_PrgB' }         0.038771     0.095394     0.40644    111      0.68521
    {'Week_PrgC' }         0.030543     0.095394     0.32018    111      0.74944
    {'Week_PrgD' }         0.033114     0.095394     0.34713    111      0.72915


    Lower         Upper
       0.14798       1.1741
    0.00045067    0.0059252
       0.10044      0.62113
      -0.29319      0.22666
      -0.14706       0.3734
      0.039536      0.30686
      -0.15026       0.2278
      -0.15849      0.21957
      -0.15592      0.22214

Random effects covariance parameters (95% CIs):
Group: Subject (20 Levels)
    Name1               Name2               Type            Estimate
    {'Intercept'}       {'Intercept'}       {'std' }        0.18407
```

```
    {'Week'     }        {'Intercept'}        {'corr'}        0.66841
    {'Week'     }        {'Week'     }        {'std' }        0.15033


    Lower      Upper
    0.12281    0.27587
    0.21076    0.88573
    0.11004    0.20537

Group: Error
    Name               Estimate    Lower       Upper
    {'Res Std'}        0.10261     0.087882    0.11981
```

Examine the fixed effects coefficients table. The row labeled `'InitWeight'` has a *p*-value of 0.0228, and the row labeled `'Week'` has a *p*-value of 0.0115. These *p*-values indicate significant effects of the initial weights of the subjects and the time factor in the amount of weight lost. The weight loss of subjects who are in program B is significantly different relative to the weight loss of subjects who are in program A. The lower and upper limits of the covariance parameters for the random effects do not include zero, thus they seem significant. You can also test the significance of the random-effects using the `compare` method.

Random Intercept Model

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables for estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the Centers for Disease Control and Prevention, CDC).

To fit a linear-mixed effects model, where the influenza rates are the responses, combine the nine columns corresponding to the regions into an array that has a single response variable, `FluRate`, and a nominal variable, `Region`, the nationwide estimate `WtdILI`, that shows which region each estimate is from, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate',...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Define the design matrices for a random-intercept linear mixed-effects model, where the intercept varies by `Date`. The corresponding model is

$$y_{im} = \beta_0 + \beta_1 \mathrm{WtdILI}_{im} + b_{0m} + \varepsilon_{im}, \quad i = 1, 2, \ldots, 468, \quad m = 1, 2, \ldots, 52,$$

where y_{im} is the observation i for level m of grouping variable `Date`, b_{0m} is the random effect for level m of the grouping variable `Date`, and ε_{im} is the observation error for observation i. The random effect has the prior distribution,

$$b_{0m} \sim N(0, \sigma_b^2),$$

and the error term has the distribution,

$\varepsilon_{im} \sim N(0, \sigma^2)$.

```
y = flu2.FluRate;
X = [ones(468,1) flu2.WtdILI];
Z = [ones(468,1)];
G = flu2.Date;
```

Fit the linear mixed-effects model.

```
lme = fitlmematrix(X,y,Z,G,'FixedEffectPredictors',{'Intercept','NationalRate'},...
'RandomEffectPredictors',{{'Intercept'}},'RandomEffectGroups',{'Date'})
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             468
    Fixed effects coefficients           2
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    y ~ Intercept + NationalRate + (Intercept | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    286.24    302.83    -139.12          278.24

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat     DF     pValue
    {'Intercept'   }        0.16385    0.057525    2.8484    466     0.0045885
    {'NationalRate'}         0.7236    0.032219    22.459    466    3.0502e-76


    Lower       Upper
    0.050813    0.27689
     0.66028    0.78691

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1               Name2               Type           Estimate    Lower
    {'Intercept'}       {'Intercept'}       {'std'}        0.17146     0.13227


    Upper
    0.22226

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.30201     0.28217    0.32324
```

The confidence limits of the standard deviation of the random-effects term σ_b, do not include zero (0.13227, 0.22226), which indicates that the random-effects term is significant. You can also test the significance of the random-effects using `compare` method.

The estimated value of an observation is the sum of the fixed-effects values and value of the random effect at the grouping variable level corresponding to that observation. For example, the estimated flu rate for observation 28

$$\begin{aligned}\hat{y}_{28} &= \hat{\beta}_0 + \hat{\beta}_1 \mathrm{WtdILI}_{28} + \hat{b}_{10/30/2005} \\ &= 0.1639 + 0.7236*(1.343) + 0.3318 \\ &= 1.46749,\end{aligned}$$

where $\hat{b}$ is the best linear unbiased predictor (BLUP) of the random effects for the intercept. You can compute this value as follows.

```
beta = fixedEffects(lme);
[~,~,STATS] = randomEffects(lme); % compute the random effects statistics STATS
STATS.Level = nominal(STATS.Level);
y_hat = beta(1) + beta(2)*flu2.WtdILI(28) + STATS.Estimate(STATS.Level=='10/30/2005')
```

```
y_hat = 1.4674
```

You can simply display the fitted value using the `fitted(lme)` method.

```
F = fitted(lme);
F(28)
```

```
ans = 1.4674
```

Randomized Block Design

Load the sample data.

```
load('shift.mat');
```

The data shows the deviations from the target quality characteristic measured from the products that five operators manufacture during three shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the deviations of the quality characteristics from the target value. This is simulated data.

Define the design matrices for a linear mixed-effects model with a random intercept grouped by operator, and shift as the fixed effects. Use the `'effects'` contrasts. `'effects'` contrasts mean that the coefficients sum to 0. You need to create two contrast coded variables in the fixed-effects design matrix, `X1` and `X2`, where

$$\text{Shift\_Evening} = \begin{cases} 0, & \text{if Morning} \\ 1, & \text{if Evening} \\ -1, & \text{if Night} \end{cases}$$

$$\text{Shift\_Morning} = \begin{cases} 1, & \text{if Morning} \\ 0, & \text{if Evening} \\ -1, & \text{if Night} \end{cases}$$

The model corresponds to

$$\text{Morning Shift:}\quad \text{QCDev}_{im} = \beta_0 + \beta_2 \text{Shift\_Morning}_i + b_{0m} + \epsilon_{im},$$
$$\text{Evening Shift:}\quad \text{QCDev}_{im} = \beta_0 + \beta_1 \text{Shift\_Evening}_i + b_{0m} + \epsilon_{im},$$
$$\text{Night Shift:}\quad \text{QCDev}_{im} = \beta_0 - \beta_1 \text{Shift\_Evening}_i - \beta_2 \text{Shift\_Morning}_i + b_{0m} + \epsilon_{im},$$

where i represents the observations, and m represents the operators, i = 1, 2, ..., 15, and m = 1, 2, ..., 5. The random effects and the observation error have the following distributions:

$$b_{0m} \sim N(0, \sigma_b^2)$$

and

$$\varepsilon_{im} \sim N(0, \sigma^2).$$

```
S = shift.Shift;
X1 = (S=='Morning') - (S=='Night');
X2 = (S=='Evening') - (S=='Night');
X = [ones(15,1), X1, X2];
y = shift.QCDev;
Z = ones(15,1);
G = shift.Operator;
```

Fit a linear mixed-effects model using the specified design matrices and restricted maximum likelihood method.

```
lme = fitlmematrix(X,y,Z,G,'FitMethod','REML','FixedEffectPredictors',....
{'Intercept','S_Morning','S_Evening'},'RandomEffectPredictors',{{'Intercept'}},...
'RandomEffectGroups',{'Operator'},'DummyVarCoding','effects')
```

```
lme =
Linear mixed-effects model fit by REML

Model information:
    Number of observations             15
    Fixed effects coefficients          3
    Random effects coefficients         5
    Covariance parameters               2

Formula:
    y ~ Intercept + S_Morning + S_Evening + (Intercept | Operator)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    58.913    61.337    -24.456          48.913

Fixed effects coefficients (95% CIs):
    Name                 Estimate    SE         tStat      DF    pValue
    {'Intercept'}          3.6525    0.94109     3.8812    12    0.0021832
    {'S_Morning'}        -0.91973    0.31206    -2.9473    12     0.012206
    {'S_Evening'}        -0.53293    0.31206    -1.7078    12      0.11339


    Lower      Upper
     1.6021       5.703
    -1.5997    -0.23981
    -1.2129     0.14699
```

```
Random effects covariance parameters (95% CIs):
Group: Operator (5 Levels)
    Name1               Name2               Type           Estimate    Lower
    {'Intercept'}       {'Intercept'}       {'std'}        2.0457      0.98207


    Upper
    4.2612

Group: Error
    Name              Estimate    Lower      Upper
    {'Res Std'}       0.85462     0.52357    1.395
```

Compute the best linear unbiased predictor (BLUP) estimates of random effects.

```
B = randomEffects(lme)
```

```
B = 5×1

    0.5775
    1.1757
   -2.1715
    2.3655
   -1.9472
```

The estimated deviation from the target quality characteristics for the third operator working the evening shift is

$$\begin{aligned}\hat{y}_{\text{Evening, Operator3}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{Shift\_Evening} + \hat{b}_{03} \\ &= 3.6525 - 0.53293 - 2.1715 \\ &= 0.94807.\end{aligned}$$

You can also display this value as follows.

```
F = fitted(lme);
F(shift.Shift=='Evening' & shift.Operator=='3')
```

```
ans = 0.9481
```

Correlated and Uncorrelated Random-Effects Terms

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration and horsepower, and uncorrelated random effect for intercept and acceleration grouped by the model year. This model corresponds to

$$\text{MPG}_{im} = \beta_0 + \beta_1 \text{Acc}_i + \beta_2 \text{HP} + b_{0m} + b_{1m} \text{Acc}_{im} + \varepsilon_{im}, \quad m = 1, 2, 3,$$

with the random-effects terms having the following prior distributions:

$$b_{0m} \sim N(0, \sigma_0^2),$$

$$b_{1m} \sim N(0, \sigma_1^2),$$

where m represents the model year.

First, prepare the design matrices for fitting the linear mixed-effects model.

```
X = [ones(406,1) Acceleration Horsepower];
Z = {ones(406,1),Acceleration};
G = {Model_Year,Model_Year};
Model_Year = nominal(Model_Year);
```

Now, fit the model using `fitlmematrix` with the defined design matrices and grouping variables.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept'},{'Acceleration'}},'RandomEffectGroups',{'Model_Year','Model_Year'})
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             392
    Fixed effects coefficients           3
    Random effects coefficients         26
    Covariance parameters                3

Formula:
    Linear Mixed Formula with 4 predictors.

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    2194.5    2218.3    -1091.3          2182.5

Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE           tStat      DF
    {'Intercept'   }          49.839       2.0518     24.291    389
    {'Acceleration'}        -0.58565      0.10846    -5.3995    389
    {'Horsepower'  }        -0.16534    0.0071227    -23.213    389


    pValue        Lower       Upper
    5.6168e-80      45.806      53.873
    1.1652e-07     -0.7989     -0.3724
    1.9755e-75    -0.17934    -0.15133

Random effects covariance parameters (95% CIs):
Group: Model_Year (13 Levels)
    Name1               Name2               Type           Estimate      Lower
    {'Intercept'}       {'Intercept'}       {'std'}        8.2415e-07    NaN


    Upper
    NaN

Group: Model_Year (13 Levels)
```

```
    Name1                   Name2                   Type           Estimate
    {'Acceleration'}        {'Acceleration'}        {'std'}        0.18783


    Lower      Upper
    0.12523    0.28172

Group: Error
    Name               Estimate    Lower     Upper
    {'Res Std'}        3.7258      3.4698    4.0007
```

Note that the random effects covariance parameters for intercept and acceleration are separate in the display. The standard deviation of the random effect for the intercept does not seem significant.

Refit the model with potentially correlated random effects for intercept and acceleration. In this case, the random-effects terms has this prior distribution

$$b_m = \begin{pmatrix} b_{0m} \\ b_{1m} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{pmatrix}\right),$$

where m represents the model year.

First, prepare the random-effects design matrix and grouping variable.

```
Z = [ones(406,1) Acceleration];
G = Model_Year;

lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'}},'RandomEffectGroups',{'Model_Year'})

lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             392
    Fixed effects coefficients           3
    Random effects coefficients         26
    Covariance parameters                4

Formula:
    Linear Mixed Formula with 4 predictors.

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    2193.5    2221.3    -1089.7          2179.5

Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE           tStat      DF
    {'Intercept'   }          50.133       2.2652     22.132    389
    {'Acceleration'}        -0.58327      0.13394    -4.3545    389
    {'Horsepower'  }        -0.16954    0.0072609     -23.35    389


    pValue        Lower       Upper
    7.7727e-71      45.679      54.586
```

```
    1.7075e-05    -0.84661    -0.31992
     5.188e-76    -0.18382    -0.15527

Random effects covariance parameters (95% CIs):
Group: Model_Year (13 Levels)
    Name1                  Name2                  Type            Estimate
    {'Intercept'   }       {'Intercept'   }       {'std' }          3.3475
    {'Acceleration'}       {'Intercept'   }       {'corr'}        -0.87971
    {'Acceleration'}       {'Acceleration'}       {'std' }         0.33789


    Lower       Upper
      1.2862      8.7119
    -0.98501    -0.29675
      0.1825     0.62558

Group: Error
    Name               Estimate    Lower     Upper
    {'Res Std'}        3.6874      3.4298    3.9644
```

Note that the random effects covariance parameters for intercept and acceleration are together in the display, with an addition of the correlation between the intercept and acceleration. The confidence intervals for the standard deviations and the correlation between the random effects for intercept and acceleration do not include 0s, hence they seem significant. You can compare these two models using the `compare` method.

Specify the Covariance Pattern

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Define `Subject` and `Program` as categorical variables.

```
Subject = nominal(Subject);
Program = nominal(Program);
```

Create the design matrices for a linear mixed-effects model, with the initial weight, type of program, and week as the fixed effects.

```
D = dummyvar(Program);
X = [ones(120,1), InitialWeight, D(:,2:4), Week];
Z = [ones(120,1) Week];
G = Subject;
```

This model corresponds to

$$\begin{aligned} y_{im} = {} & \beta_0 + \beta_1 \mathrm{IW}_i + \beta_2 \mathrm{Week}_i + \beta_3 I[\mathrm{PB}]_i + \beta_4 I[\mathrm{PC}]_i + \beta_5 I[\mathrm{PD}]_i \\ & + b_{0m} + b_{1m}\mathrm{Week2}_{im} + b_{2m}\mathrm{Week4}_{im} + b_{3m}\mathrm{Week6}_{im} + b_{4m}\mathrm{Week8}_{im} \\ & + b_{5m}\mathrm{Week10}_{im} + b_{6m}\mathrm{Week12}_{im} + \varepsilon_{im}, \end{aligned}$$

where i = 1, 2, ..., 120, and m = 1, 2, ..., 20.

β_j are the fixed-effects coefficients, j = 0, 1, ...,8, and b_{0m} and b_{1m} are random effects. IW stands for initial weight and $I[\,.\,]$ is a dummy variable representing a type of program. For example, $I[\mathrm{PB}]_i$ is the dummy variable representing program type B. The random effects and observation error have the following prior distributions:

$$b_{0m} \sim N(0, \sigma_0^2)$$

$$b_{1m} \sim N(0, \sigma_1^2)$$

$$\varepsilon_{im} \sim N(0, \sigma^2)\,.$$

Fit the model using `fitlmematrix` with the defined design matrices and grouping variables. Assume the repeated observations collected on a subject have common variance along diagonals.

```
lme = fitlmematrix(X,y,Z,G,'FixedEffectPredictors',...
{'Intercept','InitWeight','PrgB','PrgC','PrgD','Week'},...
'RandomEffectPredictors',{{'Intercept','Week'}},...
'RandomEffectGroups',{'Subject'},'CovariancePattern','Isotropic')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             120
    Fixed effects coefficients           6
    Random effects coefficients         40
    Covariance parameters                2

Formula:
    Linear Mixed Formula with 7 predictors.

Model fit statistics:
    AIC        BIC       LogLikelihood    Deviance
    -24.783    -2.483    20.391           -40.783

Fixed effects coefficients (95% CIs):
    Name                  Estimate     SE           tStat       DF
    {'Intercept' }           0.4208      0.28169      1.4938    114
    {'InitWeight'}        0.0045552    0.0015338      2.9699    114
    {'PrgB'      }          0.36993      0.12119      3.0525    114
    {'PrgC'      }        -0.034009       0.1209    -0.28129    114
    {'PrgD'      }            0.121      0.12111     0.99911    114
    {'Week'      }          0.19881     0.037134      5.3538    114


    pValue        Lower        Upper
       0.13799     -0.13723      0.97883
     0.0036324    0.0015168    0.0075935
     0.0028242      0.12986         0.61
       0.77899     -0.27351       0.2055
       0.31986     -0.11891      0.36091
    4.5191e-07      0.12525      0.27237

Random effects covariance parameters (95% CIs):
```

```
Group: Subject (20 Levels)
    Name1                Name2                Type           Estimate    Lower
    {'Intercept'}        {'Intercept'}        {'std'}        0.16561     0.12896


    Upper
    0.21269

Group: Error
    Name               Estimate    Lower       Upper
    {'Res Std'}        0.10272     0.088014    0.11987
```

Input Arguments

X — Fixed-effects design matrix

n-by-*p* matrix

Fixed-effects design matrix, specified as an *n*-by-*p* matrix, where *n* is the number of observations, and *p* is the number of fixed-effects predictor variables. Each row of X corresponds to one observation, and each column of X corresponds to one variable.

Data Types: `single` | `double`

y — Response values

n-by-1 vector

Response values, specified as an *n*-by-1 vector, where *n* is the number of observations.

Data Types: `single` | `double`

Z — Random-effects design

n-by-*q* matrix | cell array of *R* *n*-by-*q*(*r*) matrices, $r = 1, 2, ..., R$

Random-effects design, specified as either of the following.

- If there is one random-effects term in the model, then Z must be an *n*-by-*q* matrix, where *n* is the number of observations and *q* is the number of variables in the random-effects term.
- If there are *R* random-effects terms, then Z must be a cell array of length *R*. Each cell of Z contains an *n*-by-*q*(*r*) design matrix `Z{r}`, $r = 1, 2, ..., R$, corresponding to each random-effects term. Here, *q*(*r*) is the number of random effects term in the *r*th random effects design matrix, `Z{r}`.

Data Types: `single` | `double` | `cell`

G — Grouping variable or variables

n-by-1 vector | cell array of *R* *n*-by-1 vectors

Grouping variable or variables on page 2-45, specified as either of the following.

- If there is one random-effects term, then G must be an *n*-by-1 vector corresponding to a single grouping variable with *M* levels or groups.

  G can be a categorical vector, logical vector, numeric vector, character array, string array, or cell array of character vectors.

- If there are multiple random-effects terms, then `G` must be a cell array of length R. Each cell of `G` contains a grouping variable `G{r}`, $r = 1, 2, ..., R$, with $M(r)$ levels.

  `G{r}` can be a categorical vector, logical vector, numeric vector, character array, string array, or cell array of character vectors.

Data Types: `categorical` | `logical` | `single` | `double` | `char` | `string` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'CovariancePattern','Diagonal','DummyVarCoding','full','Optimizer','fminunc'` specifies a random-effects covariance pattern with zero off-diagonal elements, creates a dummy variable for each level of a categorical variable, and uses the `fminunc` optimization algorithm.

`FixedEffectPredictors` — Names of columns in fixed-effects design matrix
`{'x1','x2',...,'xP'}` (default) | string array or cell array of length p

Names of columns in the fixed-effects design matrix `X`, specified as the comma-separated pair consisting of `'FixedEffectPredictors'` and a string array or cell array of length p.

For example, if you have a constant term and two predictors, say `TimeSpent` and `Gender`, where `Female` is the reference level for `Gender`, as the fixed effects, then you can specify the names of your fixed effects in the following way. `Gender_Male` represents the dummy variable you must create for category `Male`. You can choose different names for these variables.

Example: `'FixedEffectPredictors',{'Intercept','TimeSpent','Gender_Male'},`

Data Types: `string` | `cell`

`RandomEffectPredictors` — Names of columns in random-effects design matrix or cell array
string array or cell array of length q | cell array of length R with elements of length $q(r)$, $r = 1, 2, ..., R$

Names of columns in the random-effects design matrix or cell array `Z`, specified as the comma-separated pair consisting of `'RandomEffectPredictors'` and either of the following:

- A string array or cell array of length q when `Z` is an n-by-q design matrix. In this case, the default is `{'z1','z2',...,'zQ'}`.
- A cell array of length R, when `Z` is a cell array of length R with each element `Z{r}` of length $q(r)$, $r = 1, 2, ..., R$. In this case, the default is `{'z11','z12',...,'z1Q(1)'},...,{'zr1','zr2',...,'zrQ(r)'}`.

For example, suppose you have correlated random effects for intercept and a variable named `Acceleration`. Then, you can specify the random-effects predictor names as follows.

Example: `'RandomEffectPredictors',{'Intercept','Acceleration'}`

If you have two random effects terms, one for the intercept and the variable `Acceleration` grouped by variable `g1`, and the second for the intercept, grouped by the variable `g2`, then you specify the random-effects predictor names as follows.

Example: `'RandomEffectPredictors',{{'Intercept','Acceleration'},{'Intercept'}}`

Data Types: `string` | `cell`

`ResponseVarName` — Name of response variable
`'y'` (default) | character vector | string scalar

Name of response variable, specified as the comma-separated pair consisting of `'ResponseVarName'` and a character vector or string scalar.

For example, if your response variable name is `score`, then you can specify it as follows.

Example: `'ResponseVarName','score'`

Data Types: `char` | `string`

`RandomEffectGroups` — Names of random effects grouping variables
`'g'` or `{'g1','g2',...,'gR'}` (default) | character vector | string scalar | string array | cell array of character vectors

Names of random effects grouping variables, specified as the comma-separated pair `'RandomEffectGroups'` and either of the following:

- Character vector or string scalar — If there is only one random-effects term, that is, if `G` is a vector, then the value of `'RandomEffectGroups'` is the name for the grouping variable `G`. The default is `'g'`.
- String array or cell array of character vectors — If there are multiple random-effects terms, that is, if `G` is a cell array of length *R*, then the value of `'RandomEffectGroups'` is a string array or cell array of length *R*, where each element is the name for the grouping variable `G{r}`. The default is `{'g1','g2',...,'gR'}`.

For example, if you have two random-effects terms, `z1` and `z2`, grouped by the grouping variables `sex` and `subject`, then you can specify the names of your grouping variables as follows.

Example: `'RandomEffectGroups',{'sex','subject'}`

Data Types: `char` | `string` | `cell`

`CovariancePattern` — Pattern of covariance matrix
`'FullCholesky'` (default) | character vector | string scalar | square symmetric logical matrix | string array | cell array of character vectors or logical matrices

Pattern of the covariance matrix of the random effects, specified as the comma-separated pair consisting of `'CovariancePattern'` and a character vector, a string scalar, a square symmetric logical matrix, a string array, or a cell array of character vectors or logical matrices.

If there are *R* random-effects terms, then the value of `'CovariancePattern'` must be a string array or cell array of length *R*, where each element *r* of the array specifies the pattern of the covariance matrix of the random-effects vector associated with the *r*th random-effects term. The options for each element follow.

`'FullCholesky'`	Default. Full covariance matrix using the Cholesky parameterization. `fitlme` estimates all elements of the covariance matrix.

`'Full'`	Full covariance matrix, using the log-Cholesky parameterization. `fitlme` estimates all elements of the covariance matrix.
`'Diagonal'`	Diagonal covariance matrix. That is, off-diagonal elements of the covariance matrix are constrained to be 0. $\begin{pmatrix} \sigma_{b1}^2 & 0 & 0 \\ 0 & \sigma_{b2}^2 & 0 \\ 0 & 0 & \sigma_{b3}^2 \end{pmatrix}$
`'Isotropic'`	Diagonal covariance matrix with equal variances. That is, off-diagonal elements of the covariance matrix are constrained to be 0, and the diagonal elements are constrained to be equal. For example, if there are three random-effects terms with an isotropic covariance structure, this covariance matrix looks like $\begin{pmatrix} \sigma_b^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_b^2 \end{pmatrix}$ where σ^2_b is the common variance of the random-effects terms.
`'CompSymm'`	Compound symmetry structure. That is, common variance along diagonals and equal correlation between all random effects. For example, if there are three random-effects terms with a covariance matrix having a compound symmetry structure, this covariance matrix looks like $\begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1,b2} & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1}^2 & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b1,b2} & \sigma_{b1}^2 \end{pmatrix}$ where σ^2_{b1} is the common variance of the random-effects terms and $\sigma_{b1,b2}$ is the common covariance between any two random-effects term .
`PAT`	Square symmetric logical matrix. If `'CovariancePattern'` is defined by the matrix `PAT`, and if `PAT(a,b) = false`, then the `(a,b)` element of the corresponding covariance matrix is constrained to be 0.

Example: `'CovariancePattern','Diagonal'`

Example: `'CovariancePattern',{'Full','Diagonal'}`

Data Types: `char` | `string` | `logical` | `cell`

`FitMethod` — Method for estimating parameters
`'ML'` (default) | `'REML'`

Method for estimating parameters of the linear mixed-effects model, specified as the comma-separated pair consisting of `'FitMethod'` and either of the following.

`'ML'`	Default. Maximum likelihood estimation
`'REML'`	Restricted maximum likelihood estimation

Example: `'FitMethod','REML'`

`Weights` — Observation weights
vector of scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of length *n*, where *n* is the number of observations.

Data Types: `single` | `double`

`Exclude` — Indices for rows to exclude
use all rows without `NaN`s (default) | vector of integer or logical values

Indices for rows to exclude from the linear mixed-effects model in the data, specified as the comma-separated pair consisting of `'Exclude'` and a vector of integer or logical values.

For example, you can exclude the 13th and 67th rows from the fit as follows.

Example: `'Exclude',[13,67]`

Data Types: `single` | `double` | `logical`

`DummyVarCoding` — Coding to use for dummy variables
`'reference'` (default) | `'effects'` | `'full'`

Coding to use for dummy variables created from the categorical variables, specified as the comma-separated pair consisting of `'DummyVarCoding'` and one of the variables in this table.

Value	Description
`'reference'` (default)	`fitlmematrix` creates dummy variables with a reference group. This scheme treats the first category as a reference group and creates one less dummy variables than the number of categories. You can check the category order of a categorical variable by using the `categories` function, and change the order by using the `reordercats` function.
`'effects'`	`fitlmematrix` creates dummy variables using effects coding. This scheme uses –1 to represent the last category. This scheme creates one less dummy variables than the number of categories.

Value	Description
`'full'`	`fitlmematrix` creates full dummy variables. This scheme creates one dummy variable for each category.

For more details about creating dummy variables, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'DummyVarCoding','effects'`

`Optimizer` — Optimization algorithm
`'quasinewton'` (default) | `'fminunc'`

Optimization algorithm, specified as the comma-separated pair consisting of `'Optimizer'` and either of the following.

`'quasinewton'`	Default. Uses a trust region based quasi-Newton optimizer. Change the options of the algorithm using `statset('LinearMixedModel')`. If you don’t specify the options, then `LinearMixedModel` uses the default options of `statset('LinearMixedModel')`.
`'fminunc'`	You must have Optimization Toolbox to specify this option. Change the options of the algorithm using `optimoptions('fminunc')`. If you don’t specify the options, then `LinearMixedModel` uses the default options of `optimoptions('fminunc')` with `'Algorithm'` set to `'quasi-newton'`.

Example: `'Optimizer','fminunc'`

`OptimizerOptions` — Options for optimization algorithm
structure returned by `statset` | object returned by `optimoptions`

Options for the optimization algorithm, specified as the comma-separated pair consisting of `'OptimizerOptions'` and a structure returned by `statset('LinearMixedModel')` or an object returned by `optimoptions('fminunc')`.

- If `'Optimizer'` is `'fminunc'`, then use `optimoptions('fminunc')` to change the options of the optimization algorithm. See `optimoptions` for the options `'fminunc'` uses. If `'Optimizer'` is `'fminunc'` and you do not supply `'OptimizerOptions'`, then the default for `LinearMixedModel` is the default options created by `optimoptions('fminunc')` with `'Algorithm'` set to `'quasi-newton'`.
- If `'Optimizer'` is `'quasinewton'`, then use `statset('LinearMixedModel')` to change the optimization parameters. If you don’t change the optimization parameters, then `LinearMixedModel` uses the default options created by `statset('LinearMixedModel')`:

The `'quasinewton'` optimizer uses the following fields in the structure created by `statset('LinearMixedModel')`.

`TolFun` — Relative tolerance on gradient of objective function
`1e-6` (default) | positive scalar value

Relative tolerance on the gradient of the objective function, specified as a positive scalar value.

`TolX` — Absolute tolerance on step size
`1e-12` (default) | positive scalar value

Absolute tolerance on the step size, specified as a positive scalar value.

`MaxIter` — Maximum number of iterations allowed
`10000` (default) | positive scalar value

Maximum number of iterations allowed, specified as a positive scalar value.

`Display` — Level of display
`'off'` (default) | `'iter'` | `'final'`

Level of display, specified as one of `'off'`, `'iter'`, or `'final'`.

`StartMethod` — Method to start iterative optimization
`'default'` (default) | `'random'`

Method to start iterative optimization, specified as the comma-separated pair consisting of `'StartMethod'` and either of the following.

Value	Description
`'default'`	An internally defined default value
`'random'`	A random initial value

Example: `'StartMethod','random'`

`Verbose` — Indicator to display optimization process on screen
`false` (default) | `true`

Indicator to display the optimization process on screen, specified as the comma-separated pair consisting of `'Verbose'` and either `false` or `true`. Default is `false`.

The setting for `'Verbose'` overrides the field `'Display'` in `'OptimizerOptions'`.

Example: `'Verbose',true`

`CheckHessian` — Indicator to check positive definiteness of Hessian
`false` (default) | `true`

Indicator to check the positive definiteness of the Hessian of the objective function with respect to unconstrained parameters at convergence, specified as the comma-separated pair consisting of `'CheckHessian'` and either `false` or `true`. Default is `false`.

Specify `'CheckHessian'` as `true` to verify optimality of the solution or to determine if the model is overparameterized in the number of covariance parameters.

Example: `'CheckHessian',true`

Output Arguments

`lme` — Linear mixed-effects model
`LinearMixedModel` object

Linear mixed-effects model, returned as a `LinearMixedModel` object.

More About

Cholesky Parameterization

One of the assumptions of linear mixed-effects models is that the random effects have the following prior distribution.

$$b \sim N\left(0, \sigma^2 D(\theta)\right),$$

where D is a q-by-q symmetric and positive semidefinite matrix, parameterized by a variance component vector θ, q is the number of variables in the random-effects term, and σ^2 is the observation error variance. Since the covariance matrix of the random effects, D, is symmetric, it has $q(q+1)/2$ free parameters. Suppose L is the lower triangular Cholesky factor of $D(\theta)$ such that

$$D(\theta) = L(\theta)L(\theta)^T,$$

then the $q*(q+1)/2$-by-1 unconstrained parameter vector θ is formed from elements in the lower triangular part of L.

For example, if

$$L = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix},$$

then

$$\theta = \begin{bmatrix} L_{11} \\ L_{21} \\ L_{31} \\ L_{22} \\ L_{32} \\ L_{33} \end{bmatrix}.$$

Log-Cholesky Parameterization

When the diagonal elements of L in Cholesky parameterization are constrained to be positive, then the solution for L is unique. Log-Cholesky parameterization is the same as Cholesky parameterization except that the logarithm of the diagonal elements of L are used to guarantee unique parameterization.

For example, for the 3-by-3 example in Cholesky parameterization, enforcing $L_{ii} \geq 0$,

$$\theta = \begin{bmatrix} \log(L_{11}) \\ L_{21} \\ L_{31} \\ \log(L_{22}) \\ L_{32} \\ \log(L_{33}) \end{bmatrix}.$$

Alternative Functionality

You can also fit a linear mixed-effects model using `fitlme(tbl,formula)`, where `tbl` is a table or dataset array containing the response `y`, the predictor variables `X`, and the grouping variables, and `formula` is of the form `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`.

See Also

`LinearMixedModel` | `compare` | `fitlme`

Introduced in R2013b

fitsemigraph

Label data using semi-supervised graph-based method

Syntax

```
Mdl = fitsemigraph(Tbl,ResponseVarName,UnlabeledTbl)
Mdl = fitsemigraph(Tbl,formula,UnlabeledTbl)
Mdl = fitsemigraph(Tbl,Y,UnlabeledTbl)

Mdl = fitsemigraph(X,Y,UnlabeledX)

Mdl = fitsemigraph( ___ ,Name,Value)
```

Description

`fitsemigraph` creates a semi-supervised graph-based model given labeled data, labels, and unlabeled data. The returned model contains the fitted labels for the unlabeled data and the corresponding scores. This model can also predict labels for unseen data using the `predict` object function. For more information on the different labeling algorithms, see “Algorithms” on page 33-2062.

`Mdl = fitsemigraph(Tbl,ResponseVarName,UnlabeledTbl)` uses the labeled data in `Tbl`, where `Tbl.ResponseVarName` contains the labels for the labeled data, and returns fitted labels for the unlabeled data in `UnlabeledTbl`. The function stores the fitted labels and the corresponding scores in the `FittedLabels` and `LabelScores` properties of the object `Mdl`, respectively.

`Mdl = fitsemigraph(Tbl,formula,UnlabeledTbl)` uses `formula` to specify the response variable (vector of labels) and the predictor variables to use among the variables in `Tbl`. The function uses these variables to label the data in `UnlabeledTbl`.

`Mdl = fitsemigraph(Tbl,Y,UnlabeledTbl)` uses the predictor data in `Tbl` and the labels in `Y` to label the data in `UnlabeledTbl`.

`Mdl = fitsemigraph(X,Y,UnlabeledX)` uses the predictor data in `X` and the labels in `Y` to label the data in `UnlabeledX`.

`Mdl = fitsemigraph( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify the labeling method, number of iterations, and score threshold to use in the labeling algorithm.

Examples

Fit Labels to Unlabeled Data

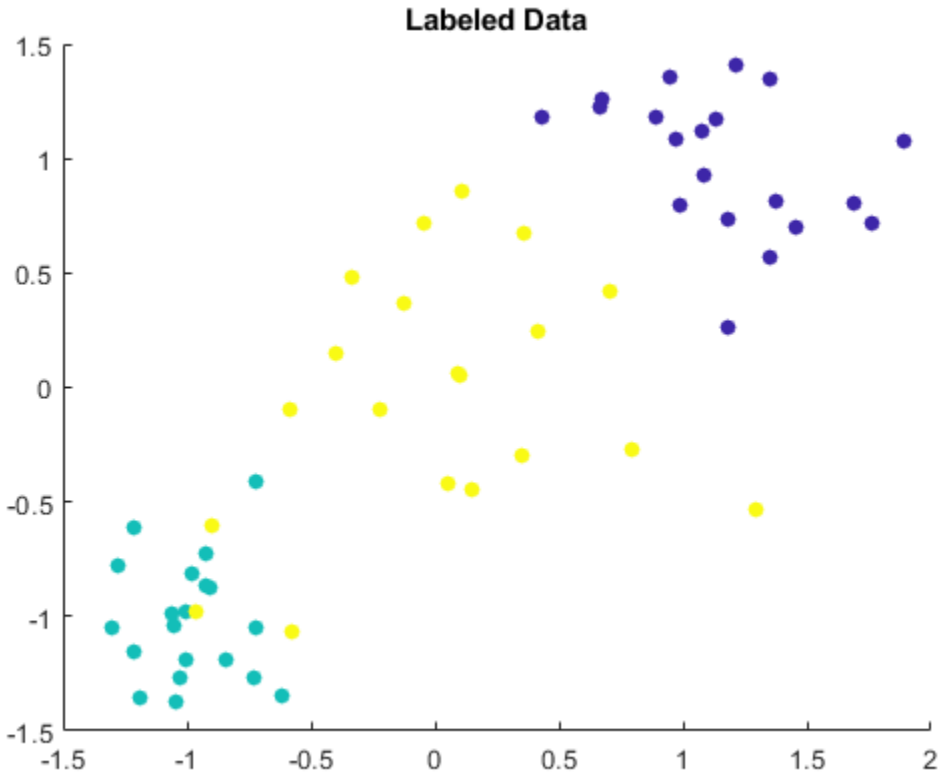
Fit labels to unlabeled data by using a semi-supervised graph-based method.

Randomly generate 60 observations of labeled data, with 20 observations in each of three classes.

```
rng('default') % For reproducibility

labeledX = [randn(20,2)*0.25 + ones(20,2);
            randn(20,2)*0.25 - ones(20,2);
            randn(20,2)*0.5];
Y = [ones(20,1); ones(20,1)*2; ones(20,1)*3];
```

Visualize the labeled data by using a scatter plot. Observations in the same class have the same color. Notice that the data is split into three clusters with very little overlap.

```
scatter(labeledX(:,1),labeledX(:,2),[],Y,'filled')
title('Labeled Data')
```



Randomly generate 300 additional observations of unlabeled data, with 100 observations per class. For the purposes of validation, keep track of the true labels for the unlabeled data.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
trueLabels = [ones(100,1); ones(100,1)*2; ones(100,1)*3];
```

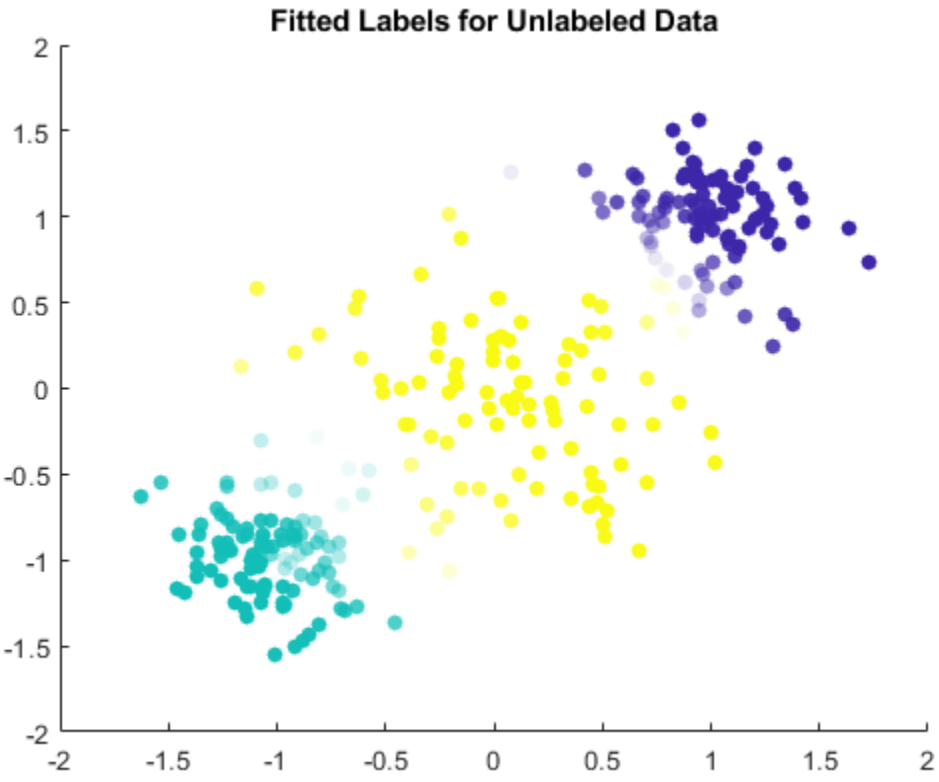
Fit labels to the unlabeled data by using a semi-supervised graph-based method. The function `fitsemigraph` returns a `SemiSupervisedGraphModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemigraph(labeledX,Y,unlabeledX)
```

```
Mdl =
  SemiSupervisedGraphModel with properties:

            FittedLabels: [300x1 double]
             LabelScores: [300x3 double]
              ClassNames: [1 2 3]
            ResponseName: 'Y'
    CategoricalPredictors: []
                  Method: 'labelpropagation'


  Properties, Methods
```

Visualize the fitted label results by using a scatter plot. Use the fitted labels to set the color of the observations, and use the maximum label scores to set the transparency of the observations. Observations with less transparency are labeled with greater confidence. Notice that observations that lie closer to the cluster boundaries are labeled with more uncertainty.

```
maxLabelScores = max(Mdl.LabelScores,[],2);
rescaledScores = rescale(maxLabelScores,0.05,0.95);
scatter(unlabeledX(:,1),unlabeledX(:,2),[],Mdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledScores);
title('Fitted Labels for Unlabeled Data')
```



Determine the accuracy of the labeling by using the true labels for the unlabeled data.

```
numWrongLabels = sum(trueLabels ~= Mdl.FittedLabels)
```

```
numWrongLabels = 10
```

Only 10 of the 300 observations in `unlabeledX` are mislabeled.

Specify Graph Type Used to Fit Labels

Fit labels to unlabeled data by using a semi-supervised graph-based method. Specify the type of nearest neighbor graph.

Load the `patients` data set. Create a table from the variables `Distolic`, `Gender`, and so on. For each observation, or row in the table, treat the `Smoker` value as the label for that observation.

```
load patients
Tbl = table(Diastolic,Gender,Height,Systolic,Weight,Smoker);
```

Suppose only 20% of the observations are labeled. To recreate this scenario, randomly sample 20 labeled observations and store them in the table `unlabeledTbl`. Remove the label from the rest of the observations and store them in the table `unlabeledTbl`. To verify the accuracy of the label fitting at the end of the example, retain the true labels for the unlabeled data in the variable `trueLabels`.

```
rng('default') % For reproducibility of the sampling
[labeledTbl,Idx] = datasample(Tbl,20,'Replace',false);

unlabeledTbl = Tbl;
unlabeledTbl(Idx,:) = [];
trueLabels = unlabeledTbl.Smoker;
unlabeledTbl.Smoker = [];
```

Fit labels to the unlabeled data by using a semi-supervised graph-based method. Use a mutual type of nearest neighbor graph, where two points are connected when they are nearest neighbors of each other. Specify to standardize the numeric predictors. The function `fitsemigraph` returns an object whose `FittedLabels` property contains the fitted labels for the unlabeled data.

```
Mdl = fitsemigraph(labeledTbl,'Smoker',unlabeledTbl,'KNNGraphType','mutual', ...
    'Standardize',true);
fittedLabels = Mdl.FittedLabels;
```

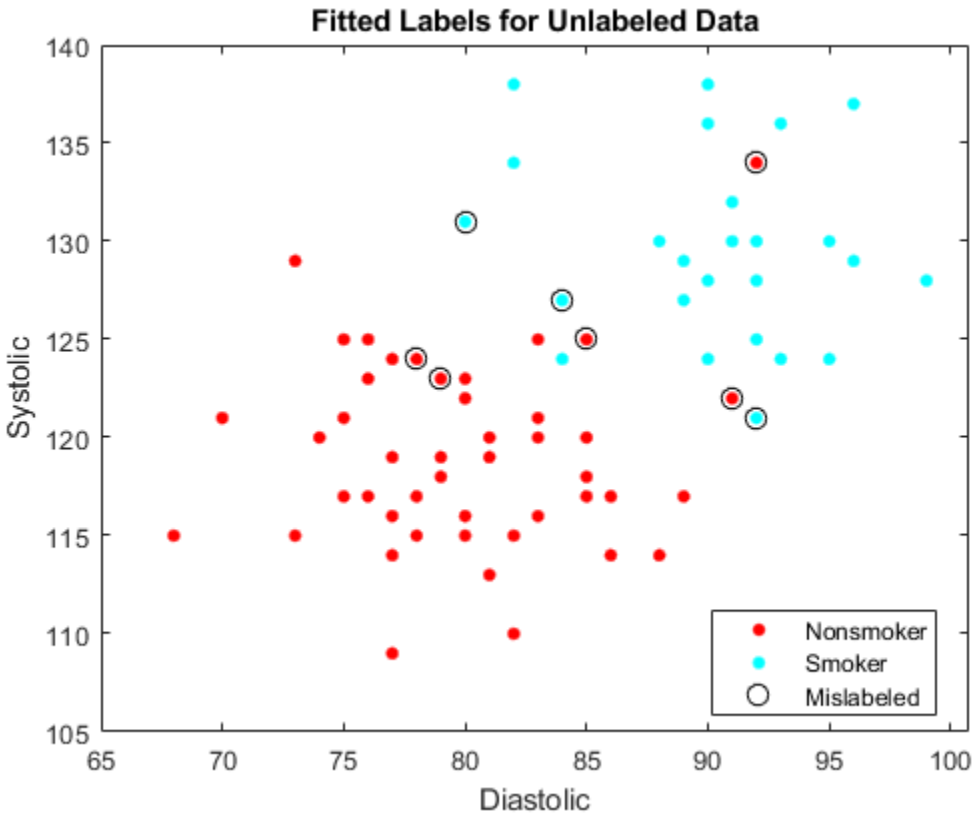
Identify the observations that are incorrectly labeled by comparing the stored true labels for the unlabeled data to the fitted labels returned by the semi-supervised graph-based method.

```
wrongIdx = (trueLabels ~= fittedLabels);
wrongTbl = unlabeledTbl(wrongIdx,:);
```

Visualize the fitted label results for the unlabeled data. Mislabeled observations are circled in the plot.

```
gscatter(unlabeledTbl.Diastolic,unlabeledTbl.Systolic, ...
    fittedLabels)
hold on
plot(wrongTbl.Diastolic,wrongTbl.Systolic, ...
    'ko','MarkerSize',8)
xlabel('Diastolic')
```

```
ylabel('Systolic')
legend('Nonsmoker','Smoker','Mislabeled')
title('Fitted Labels for Unlabeled Data')
```



Input Arguments

`Tbl` — Labeled sample data

table

Labeled sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor. Optionally, `Tbl` can contain one additional column for the response variable (vector of labels). Multicolumn variables, cell arrays other than cell arrays of character vectors, and ordinal categorical variables are not supported.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, specify a formula using `formula`.

If `Tbl` does not contain the response variable, specify a response variable using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`UnlabeledTbl` — Unlabeled sample data
table

Unlabeled sample data, specified as a table. Each row of `UnlabeledTbl` corresponds to one observation, and each column corresponds to one predictor. `UnlabeledTbl` must contain the same predictors as those contained in `Tbl`.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable contains the class labels for the sample data in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors.

The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value pair argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~X1+X2+X3'`. In this form, `Y` represents the response variable, and `X1`, `X2`, and `X3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class labels, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors.

- If `Y` is a character array, then each element of the class labels must correspond to one row of the array.
- The length of `Y` must be equal to the number of rows in `Tbl` or `X`.

- A good practice is to specify the class order by using the `ClassNames` name-value pair argument.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Labeled predictor data
numeric matrix

Labeled predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one predictor.

The length of `Y` and the number of rows in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

`UnlabeledX` — Unlabeled predictor data
numeric matrix

Unlabeled predictor data, specified as a numeric matrix. Each row of `UnlabeledX` corresponds to one observation, and each column corresponds to one predictor. `UnlabeledX` must have the same predictors as `X`, in the same order.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data. The software removes rows of the predictor data (observations) with missing values.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitsemigraph(Tbl,'Y',UnlabeledTbl,'Method','labelspreading','IterationLimit',2e3)` specifies to use a label spreading labeling technique and run a maximum of 2000 iterations.

Labeling Algorithm Options

`Method` — Labeling technique
`'labelpropagation'` (default) | `'labelpropagationexact'` | `'labelspreading'` | `'labelspreadingexact'`

Labeling technique, specified as the comma-separated pair consisting of `'Method'` and one of these values.

Value	Description	Method-Specific Name-Value Pair Arguments
`'labelpropagation'`	Iteratively propagate labels across the nodes in the similarity graph. For more information, see "Label Propagation" on page 33-2062.	`'IterationLimit'` — Maximum number of iterations `'Tolerance'` — Tolerance for the score absolute difference in subsequent iterations
`'labelpropagationexact'`	Use an exact formula to propagate labels. For more information, see "Label Propagation" on page 33-2062.	None
`'labelspreading'`	Iteratively spread labels across the nodes in the similarity graph. For more information, see "Label Spreading" on page 33-2063.	`'Alpha'` — Relative weight of neighboring labels to the initial label `'IterationLimit'` — Maximum number of iterations `'Tolerance'` — Tolerance for the score absolute difference in subsequent iterations
`'labelspreadingexact'`	Use an exact formula to spread labels. For more information, see "Label Spreading" on page 33-2063.	`'Alpha'` — Relative weight of neighboring labels to the initial label

Example: `'Method','labelspreading'`

Data Types: `char` | `string`

`Alpha` — Relative weight of neighboring labels to initial label
`0.01` (default) | scalar value in the range (0,1)

Relative weight of neighboring labels to the initial label for labeled observations in `X` or `Tbl`, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). A value close to 0 indicates that `fitsemigraph` treats labels of initially labeled observations almost like ground truth. A value close to 1 indicates that `fitsemigraph` treats labels of initially labeled observations almost like noise.

Note This argument is valid only when the `Method` value is `'labelspreading'` or `'labelspreadingexact'`.

Example: `'Alpha',0.05`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of iterations
`1e3` (default) | positive integer scalar

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer scalar. The `fitsemigraph` function returns `Mdl`, which

contains the fitted labels and scores, when this limit is reached, even if the algorithm does not converge.

Note This argument is valid only when the `Method` value is `'labelpropagation'` or `'labelspreading'`.

Example: `'IterationLimit',2e3`

Data Types: `single` | `double`

`Tolerance` — Tolerance for score absolute difference in subsequent iterations
`1e-3` (default) | nonnegative scalar

Tolerance for score absolute difference in subsequent iterations, specified as the comma-separated pair consisting of `'Tolerance'` and a nonnegative scalar.

Note This argument is valid only when the `Method` value is `'labelpropagation'` or `'labelspreading'`.

Example: `'Tolerance',1e-4`

Data Types: `single` | `double`

Classification Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitsemigraph` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.

Value	Description
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table, `fitsemigraph` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. Ordinal categorical variables are not supported. If the predictor data is a matrix, `fitsemigraph` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value pair argument.

`fitsemigraph` encodes categorical variables as numeric variables by assigning a positive integer value to each category. When you use categorical predictors, ensure that you use an appropriate distance metric (`Distance`).

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for labeling

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of the classes to use for labeling, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `ClassNames` must have the same data type as `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `'ClassNames'` to:

- Order the classes.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `'ClassNames'` to specify the column order of classification scores in `Mdl.LabelScores`.
- Select a subset of classes for labeling. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To use observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the predictor data.

- If you supply `X`, `Y`, and `UnlabeledX`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X` and `UnlabeledX`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl` and `UnlabeledTbl`, then you can use `'PredictorNames'` to choose which predictor variables to use. That is, `fitsemigraph` uses only the predictor variables in `PredictorNames` and the response variable to label the unlabeled data.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as the comma-separated pair consisting of `'ResponseName'` and a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Standardize` — Flag to standardize predictor data
`false` or `0` (default) | `true` or `1`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a numeric or logical `0` (`false`) or `1` (`true`). If you set `'Standardize',true`, the software combines the labeled and unlabeled predictor data, and then centers and scales each numeric predictor variable by the corresponding column mean and standard deviation. The software does not standardize the categorical predictors.

Example: `'Standardize',true`

Data Types: `double` | `logical`

Distance Metric Options

`Distance` — Distance metric
character vector | string scalar

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector or string scalar.

- If all the predictor variables are continuous (numeric) variables, then you can specify one of these distance metrics.

Value	Description
`'euclidean'`	Euclidean distance
`'seuclidean'`	Standardized Euclidean distance — Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation `S = std(PD,'omitnan')`, where `PD` is the predictor data, both labeled and unlabeled. To specify another scale parameter, use the `'Scale'` name-value pair argument.
`'mahalanobis'`	Mahalanobis distance — By default, the distance is computed using `C = cov(PD,'omitrows')`, the covariance of `PD`. To change the value of the covariance matrix, use the `'Cov'` name-value pair argument.
`'minkowski'`	Minkowski distance — The default exponent is 2. To specify a different exponent, use the `'P'` name-value pair argument.
`'chebychev'`	Chebychev distance (maximum coordinate difference)
`'cityblock'`	City block distance
`'correlation'`	One minus the sample correlation between observations (treated as sequences of values)
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors)
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

Note If you specify one of these distance metrics and the predictor data includes categorical predictors, then the software treats each categorical predictor as a numeric variable, with each category represented by a positive integer.

- If all the predictor variables are categorical variables, then you can specify one of these distance metrics.

Value	Description
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ

Note If you specify one of these distance metrics and the predictor data includes continuous (numeric) predictors, then the software treats each continuous predictor as a categorical variable.

- If the predictor variables are a mix of continuous (numeric) and categorical variables, then you can specify one of these distance metrics.

Value	Description
`'goodall3'`	Modified Goodall distance
`'ofd'`	Occurrence frequency distance

The default value is `'euclidean'` if all the predictor variables are continuous, and `'goodall3'` if any of the predictor variables are categorical. For more information on the various distance metrics, see “Distance Metrics” on page 33-2059.

Example: `'Distance','ofd'`

Data Types: `char` | `string`

`Scale` — Scale parameter for standardized Euclidean distance metric
nonnegative vector

Scale parameter for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative vector. The length of `Scale` is equal to the number of predictors. Each coordinate difference between two observations is scaled by the corresponding element of `Scale`.

The default scale parameter is `std(PD,'omitnan')`, where `PD` is the predictor data, both labeled and unlabeled.

Note This argument is valid only if `Distance` is `'seuclidean'`.

Example: `'Scale',iqr(X)`

Data Types: `single` | `double`

`Cov` — Covariance matrix for Mahalanobis distance metric
positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a *p*-by-*p* positive definite matrix, where *p* is the number of predictors.

The default covariance matrix is `cov(PD,'omitrows')`, where `PD` is the predictor data, both labeled and unlabeled.

Note This argument is valid only if `Distance` is `'mahalanobis'`.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

`P` — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar.

Note This argument is valid only if `Distance` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

Graph Options

`SimilarityGraph` — Type of similarity graph
`'knn'` (default) | `'epsilon'`

Type of similarity graph used in the labeling algorithm, specified as the comma-separated pair consisting of `'SimilarityGraph'` and one of these values.

Value	Description	Graph-Specific Name-Value Pair Arguments
`'knn'`	Construct the graph using nearest neighbors.	`'NumNeighbors'` — Number of nearest neighbors used to construct the similarity graph `'KNNGraphType'` — Type of nearest neighbor graph
`'epsilon'`	Construct the graph by using radius search. You must specify a value for `Radius` if you use this option.	`'Radius'` — Search radius for the nearest neighbors used to construct the similarity graph

For more information, see “Similarity Graph” on page 33-2061.

Example: `'SimilarityGraph','epsilon','Radius',2`

Data Types: `char` | `string`

`NumNeighbors` — Number of nearest neighbors
positive integer scalar

Number of nearest neighbors used to construct the similarity graph, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer scalar.

The default number of neighbors is `log(n)`, where `n` is the number of observations in the predictor data, both labeled and unlabeled.

Note This argument is valid only if `SimilarityGraph` is `'knn'`.

Example: `'NumNeighbors',10`

Data Types: `single` | `double`

`KNNGraphType` — Type of nearest neighbor graph
`'complete'` (default) | `'mutual'`

Type of nearest neighbor graph, specified as the comma-separated pair consisting of `'KNNGraphType'` and one of these values.

Value	Description
`'complete'`	Connects two points *i* and *j*, when either *i* is a nearest neighbor of *j* or *j* is a nearest neighbor of *i*. This option leads to a denser representation of the similarity matrix.
`'mutual'`	Connects two points *i* and *j*, when *i* is a nearest neighbor of *j* and *j* is a nearest neighbor of *i*. This option leads to a sparser representation of the similarity matrix.

Note This argument is valid only if `SimilarityGraph` is `'knn'`.

Example: `'KNNGraphType','mutual'`

Data Types: `char` | `string`

`Radius` — Search radius
nonnegative scalar

Search radius for the nearest neighbors used to construct the similarity graph, specified as the comma-separated pair consisting of `'Radius'` and a nonnegative scalar.

Note You must specify this argument if `SimilarityGraph` is `'epsilon'`.

Example: `'Radius',5`

Data Types: `single` | `double`

`KernelScale` — Scale factor
1 (default) | `'auto'` | positive scalar

Scale factor for the kernel, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software uses the scale factor to transform distances to similarity measures.

- The `'auto'` option is supported only for the `'euclidean'` and `'seuclidean'` distance metrics.
- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. To reproduce results, set a random number seed using `rng` before calling `fitsemigraph`.

Example: `'KernelScale','auto'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`Mdl` — Semi-supervised graph-based classifier

`SemiSupervisedGraphModel` object

Semi-supervised graph-based classifier, returned as a `SemiSupervisedGraphModel` object. Use dot notation to access the object properties. For example, to get the fitted labels for the unlabeled data and their corresponding scores, enter `Mdl.FittedLabels` and `Mdl.LabelScores`, respectively.

More About

Distance Metrics

A distance metric is a function that defines a distance between two observations. `fitsemigraph` supports various distance metrics for continuous (numeric) predictors, categorical predictors, and a mix of continuous and categorical predictors.

Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, x_2, ..., x_{mx}, and an my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, y_2, ...,y_{my}, the various distances between the vector x_s and y_t are defined as follows:

- Distance metrics for continuous (numeric) variables
  - Euclidean distance

    $$d_{st}^2 = (x_s - y_t)(x_s - y_t)'.$$

    The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.
  - Standardized Euclidean distance

    $$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)',$$

    where V is the n-by-n diagonal matrix whose jth diagonal element is $(S(j))^2$, where S is a vector of scaling factors for each dimension.
  - Mahalanobis distance

    $$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)',$$

    where C is the covariance matrix.
  - Minkowski distance

    $$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left|x_{sj} - y_{tj}\right|^p}.$$

    For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.
  - Chebychev distance

    $$d_{st} = \max_j\left\{\left|x_{sj} - y_{tj}\right|\right\}.$$

    The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- City block distance

$$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right|.$$

The city block distance is a special case of the Minkowski distance, where $p = 1$.

- Correlation distance

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}},$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

and

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}.$$

- Cosine distance

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right).$$

- Spearman distance

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$, as computed by `tiedrank`.
  - r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, that is, $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.
  - $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.
  - $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

- Distance metrics for categorical variables

  - Hamming distance

$$d_{st} = \left(\#\,(x_{sj} \neq y_{tj})/n \right).$$

  - Jaccard distance

$$d_{st} = \frac{\#\left[\left(x_{sj} \neq y_{tj}\right) \cap \left(\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right)\right]}{\#\left[\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right]}.$$

- Distance metrics for a mix of continuous (numeric) and categorical variables
  - Modified Goodall distance

    This distance is a variant of the Goodall distance, which assigns a small distance if the matching values are infrequent regardless of the frequencies of the other values. For mismatches, the distance contribution of the predictor is 1/(number of variables).
  - Occurrence frequency distance

    For a match, the occurrence frequency distance assigns zero distance. For a mismatch, the occurrence frequency distance assigns a higher distance on a less frequent value and a lower distance on a more frequent value.

Similarity Graph

A similarity graph models the local neighborhood relationships between observations in the predictor data, both labeled and unlabeled, as an undirected graph. The nodes in the graph represent observations, and the edges, which are directionless, represent the connections between the observations.

If the pairwise distance $Dist_{i,j}$ between any two nodes i and j is positive (or larger than a certain threshold), then the similarity graph connects the two nodes using an edge. The edge between the two nodes is weighted by the pairwise similarity $S_{i,j}$, where $S_{i,\,j} = \exp\left(-\left(\frac{Dist_{i,\,j}}{\sigma}\right)^2\right)$, for a specified kernel scale σ value.

`fitsemigraph` supports these two methods of constructing a similarity graph:

- *Nearest neighbor* method (if `SimilarityGraph` is `'knn'` (default)): `fitsemigraph` connects observations in the predictor data, both labeled and unlabeled, that are nearest neighbors.
  - Use the `'NumNeighbors'` name-value pair argument to specify the number of nearest neighbors.
  - Use the `'KNNGraphType'` name-value pair argument to specify whether to make a `'complete'` or `'mutual'` connection of points.
- *Radius search* method (if `SimilarityGraph` is `'epsilon'`): `fitsemigraph` connects observations whose pairwise distances are smaller than the specified search radius. You must specify the search radius for nearest neighbors used to construct the similarity graph by using the `'Radius'` name-value pair argument.

Similarity Matrix

A similarity matrix is a matrix representation of a similarity graph on page 33-2061. The n-by-n matrix $S = (S_{i,\,j})_{i,\,j\,=\,1,\,\ldots,\,n}$ contains pairwise similarity values between connected nodes in the similarity graph. The similarity matrix of a graph is also called an adjacency matrix.

The similarity matrix is symmetric because the edges of the similarity graph are directionless. A value of $S_{i,\,j} = 0$ means that nodes i and j of the similarity graph are not connected.

Algorithms

The software constructs a similarity graph (`SimilarityGraph`) with both labeled and unlabeled observations as nodes, and distributes label information from labeled observations to unlabeled observations by using either label propagation or label spreading.

Label Propagation

To propagate labels across the nodes in the similarity graph, the iterative label propagation algorithm (where `Method` is `'labelpropagation'`) follows these steps:

1. Initialize an n-by-K matrix $F(0)$, where n is the number of nodes (or observations) and K is the number of classes.
   - The first l rows correspond to labeled observations. Each row contains a 1 in the column corresponding to the true class label for that observation, and a 0 in every other column.
   - The last u rows correspond to the unlabeled observations, and contain a 0 in all columns.
2. At iteration t (starting with $t = 1$), update the F matrix by using the probabilistic transition matrix P, so that $F(t) = PF(t-1)$, where $P_{i,j} = \frac{S_{i,j}}{\sum_{k=1}^{n} S_{i,k}}$.
   - $P_{i,j}$ is the probability of transmitting label information from node i to node j.
   - $S_{i,j}$ is the weight of the edge between node i and node j. For the definition of $S_{i,j}$, see “Similarity Graph” on page 33-2061.
3. To complete iteration t, clamp the labels for the labeled observations. That is, keep the first l rows of $F(t)$ equal to their initial values in $F(0)$.
4. Repeat the second and third steps until the F values converge. You can use the `IterationLimit` and `Tolerance` values to control the convergence.

   The final F matrix corresponds to the scores for the labeled data and the unlabeled data (`LabelScores`). For each observation, or row in F, the column with the maximum score corresponds to the fitted class label (`FittedLabels`).

For more details, see [2].

Instead of using an iterative label propagation algorithm, you can use an exact method for label propagation (where `Method` is `'labelpropagationexact'`). In this case, the u-by-K matrix of label information for the unlabeled data is $F_U = (I - P_{UU})^{-1}P_{UL}F_L$ where:

- I is the identity matrix.
- P_{UU} and P_{UL} are the labeled (L) and unlabeled (U) submatrices of P such that $P = \begin{bmatrix} P_{LL} & P_{LU} \\ P_{UL} & P_{UU} \end{bmatrix}$.
- F_L is the l-by-K matrix of label information for the labeled data. Each row contains a 1 in the column corresponding to the true class label for that observation, and a 0 in every other column.

The F_U matrix corresponds to the scores for the unlabeled data (`LabelScores`). For each observation, or row in F_U, the column with the maximum score corresponds to the fitted class label (`FittedLabels`). For more details, see [3].

Label Spreading

To spread labels across the nodes in the similarity graph, the iterative spreading propagation algorithm (where `Method` is `'labelspreading'`) follows these steps:

1. Create an n-by-K matrix Y, where n is the number of nodes (or observations) and K is the number of classes.
   - The first l rows correspond to labeled observations. Each row contains a 1 in the column corresponding to the true class label for that observation, and a 0 in every other column.
   - The last u rows correspond to the unlabeled observations, and contain a 0 in all columns.
2. Create a matrix A that is a normalized version of the n-by-n similarity matrix on page 33-2061 S, with pairwise similarity values $S_{i,j}$ as defined in "Similarity Graph" on page 33-2061. Let $A = D^{-1/2}SD^{-1/2}$, where D is the n-by-n diagonal matrix $D = \begin{bmatrix} \sum_{j=1}^{n} S_{1,n} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=1}^{n} S_{n,n} \end{bmatrix}$.
3. At iteration t (starting with $t = 1$), update the F matrix by using the matrix A and the neighboring label weight parameter α (`Alpha`), so that $F(t) = \alpha AF(t - 1) + (1 - \alpha)Y$. Let $F(0)$ equal Y.
4. Repeat the third step until the F values converge. You can use the `IterationLimit` and `Tolerance` values to control the convergence.
5. Take the limit of the sequence $\{F(t)\}_{t=1,..,T}$. This final matrix corresponds to the scores for the labeled data and the unlabeled data (`LabelScores`). For each observation, or row in the matrix, the column with the maximum score corresponds to the fitted class label (`FittedLabels`).

Instead of using an iterative label spreading algorithm, you can use an exact method for label spreading (where `Method` is `'labelspreadingexact'`). In this case, the n-by-K matrix of label information for the labeled and unlabeled data is $F = (I - \alpha A)^{-1}Y$, where I is the identity matrix. The F matrix corresponds to the scores for the labeled data and the unlabeled data (`LabelScores`). For each observation, or row in F, the column with the maximum score corresponds to the fitted class label (`FittedLabels`).

For more details, see [1].

References

[1] Zhou, Dengyong, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. "Learning with Local and Global Consistency." *Advances in Neural Information Processing Systems 16 (NIPS)*. 2003.

[2] Zhu, Xiaojin, and Zoubin Ghahramani. "Learning from Labeled and Unlabeled Data with Label Propagation." *CMU CALD tech report CMU-CALD-02-107*. 2002.

[3] Zhu, Xiaojin, Zoubin Ghahramani, and John Lafferty. "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions." *The Twentieth International Conference on Machine Learning (ICML)*. 2003.

See Also

`SemiSupervisedGraphModel` | `fitsemiself` | `predict`

Topics
“Label Data Using Semi-Supervised Learning Techniques” on page 18-250

Introduced in R2020b

fitsemiself

Label data using semi-supervised self-training method

Syntax

```
Mdl = fitsemiself(Tbl,ResponseVarName,UnlabeledTbl)
Mdl = fitsemiself(Tbl,formula,UnlabeledTbl)
Mdl = fitsemiself(Tbl,Y,UnlabeledTbl)

Mdl = fitsemiself(X,Y,UnlabeledX)

Mdl = fitsemiself( ___ ,Name,Value)
```

Description

`fitsemiself` creates a semi-supervised self-training model given labeled data, labels, and unlabeled data. The returned model contains the fitted labels for the unlabeled data and the corresponding scores. This model can also predict labels for unseen data using the `predict` object function. For more information on the labeling algorithm, see “Algorithms” on page 33-2076.

`Mdl = fitsemiself(Tbl,ResponseVarName,UnlabeledTbl)` uses the labeled data in `Tbl`, where `Tbl.ResponseVarName` contains the labels for the labeled data, and returns fitted labels for the unlabeled data in `UnlabeledTbl`. The function stores the fitted labels and the corresponding scores in the `FittedLabels` and `LabelScores` properties of the object `Mdl`, respectively.

`Mdl = fitsemiself(Tbl,formula,UnlabeledTbl)` uses `formula` to specify the response variable (vector of labels) and the predictor variables to use among the variables in `Tbl`. The function uses these variables to label the data in `UnlabeledTbl`.

`Mdl = fitsemiself(Tbl,Y,UnlabeledTbl)` uses the predictor data in `Tbl` and the labels in `Y` to label the data in `UnlabeledTbl`.

`Mdl = fitsemiself(X,Y,UnlabeledX)` uses the predictor data in `X` and the labels in `Y` to label the data in `UnlabeledX`.

`Mdl = fitsemiself( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify the type of learner, number of iterations, and score threshold to use in the labeling algorithm.

Examples

Fit Labels to Unlabeled Data

Fit labels to unlabeled data by using a semi-supervised self-training method.

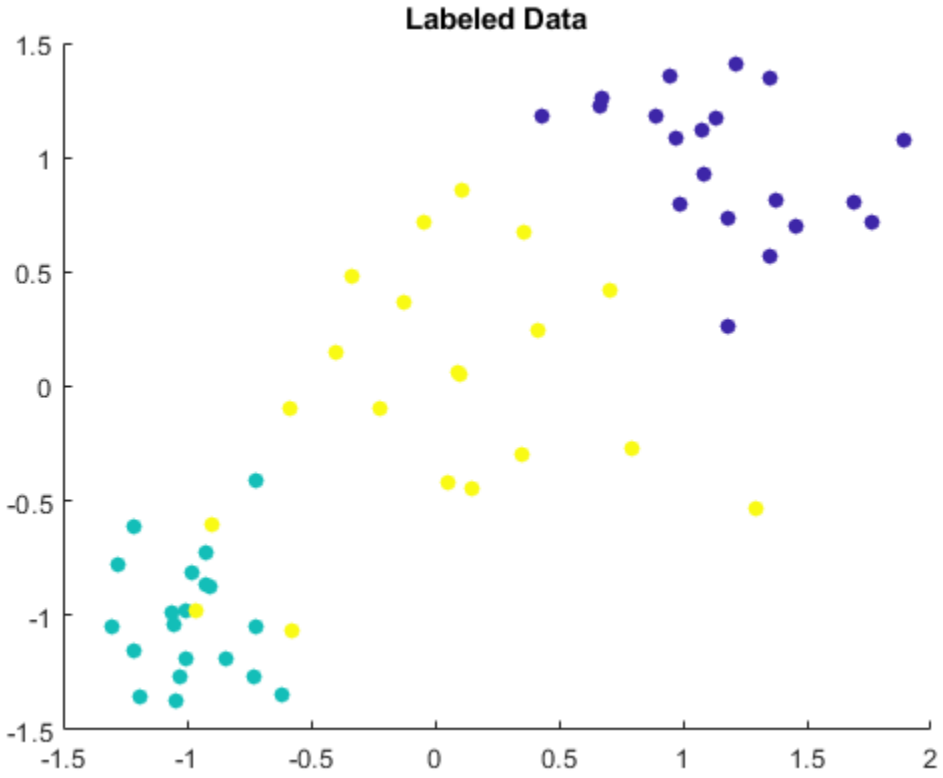
Randomly generate 60 observations of labeled data, with 20 observations in each of three classes.

```
rng('default') % For reproducibility
```

```
labeledX = [randn(20,2)*0.25 + ones(20,2);
            randn(20,2)*0.25 - ones(20,2);
            randn(20,2)*0.5];
Y = [ones(20,1); ones(20,1)*2; ones(20,1)*3];
```

Visualize the labeled data by using a scatter plot. Observations in the same class have the same color. Notice that the data is split into three clusters with very little overlap.

```
scatter(labeledX(:,1),labeledX(:,2),[],Y,'filled')
title('Labeled Data')
```



Randomly generate 300 additional observations of unlabeled data, with 100 observations per class. For the purposes of validation, keep track of the true labels for the unlabeled data.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
trueLabels = [ones(100,1); ones(100,1)*2; ones(100,1)*3];
```

Fit labels to the unlabeled data by using a semi-supervised self-training method. The function `fitsemiself` returns a `SemiSupervisedSelfTrainingModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

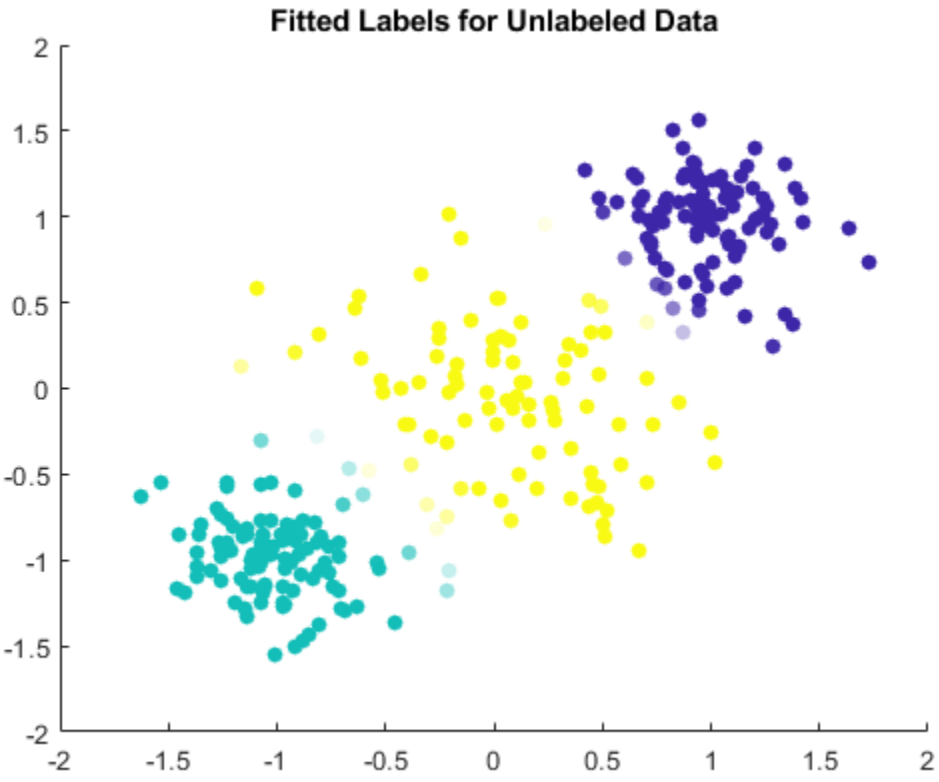
```
Mdl = fitsemiself(labeledX,Y,unlabeledX)
```

```
Mdl =
  SemiSupervisedSelfTrainingModel with properties:
```

```
            FittedLabels: [300x1 double]
             LabelScores: [300x3 double]
              ClassNames: [1 2 3]
            ResponseName: 'Y'
  CategoricalPredictors: []
                Learner: [1x1 classreg.learning.classif.CompactClassificationECOC]


  Properties, Methods
```

Visualize the fitted label results by using a scatter plot. Use the fitted labels to set the color of the observations, and use the maximum label scores to set the transparency of the observations. Observations with less transparency are labeled with greater confidence. Notice that observations that lie closer to the cluster boundaries are labeled with more uncertainty.

```
maxLabelScores = max(Mdl.LabelScores,[],2);
rescaledScores = rescale(maxLabelScores,0.05,0.95);
scatter(unlabeledX(:,1),unlabeledX(:,2),[],Mdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledScores);
title('Fitted Labels for Unlabeled Data')
```



Determine the accuracy of the labeling by using the true labels for the unlabeled data.

```
numWrongLabels = sum(trueLabels ~= Mdl.FittedLabels)
```

```
numWrongLabels = 8
```

Only 8 of the 300 observations in `unlabeledX` are mislabeled.

Specify Learner Used to Fit Labels

Fit labels to unlabeled data by using a semi-supervised self-training method. Specify the type of learner used to fit the labels.

Load the `carsmall` data set. Create a table from the variables `Acceleration`, `Displacement`, and so on. For each observation, or row in the table, treat the `Cylinders` value as the label for that observation.

```
load carsmall
Tbl = table(Acceleration,Displacement,Horsepower,Weight,Cylinders);
```

Suppose only 20% of the observations are labeled. To recreate this scenario, randomly sample 20 labeled observations and store them in the table `unlabeledTbl`. Remove the label from the rest of the observations and store them in the table `unlabeledTbl`. To verify the accuracy of the label fitting at the end of the example, retain the true labels for the unlabeled data in the variable `trueLabels`.

```
rng('default') % For reproducibility of the sampling
[labeledTbl,Idx] = datasample(Tbl,20,'Replace',false);

unlabeledTbl = Tbl;
unlabeledTbl(Idx,:) = [];
trueLabels = unlabeledTbl.Cylinders;
unlabeledTbl.Cylinders = [];
```

Fit labels to the unlabeled data by using a semi-supervised self-training method. Use a multiclass SVM (ECOC) model to iteratively label the unlabeled observations. Specify to standardize the numeric predictors and use a linear kernel function for the SVM binary learners. The function `fitsemiself` returns an object whose `FittedLabels` property contains the fitted labels for the unlabeled data.

```
Mdl = fitsemiself(labeledTbl,'Cylinders',unlabeledTbl, ...
    'Learner',templateECOC('Learner',templateSVM('Standardize',true, ...
    'KernelFunction','linear')));
fittedLabels = Mdl.FittedLabels;
```

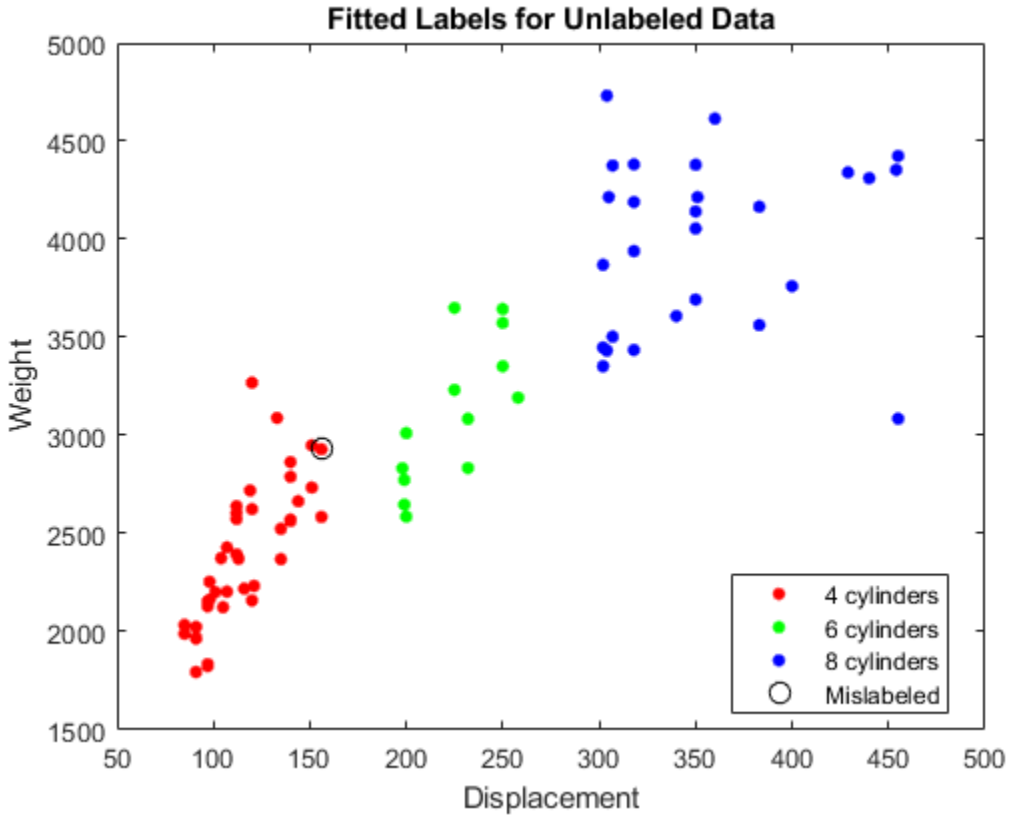
Identify the observations that are incorrectly labeled by comparing the stored true labels for the unlabeled data to the fitted labels returned by the semi-supervised self-training method.

```
wrongIdx = (trueLabels ~= fittedLabels);
wrongTbl = unlabeledTbl(wrongIdx,:);
```

Visualize the fitted label results for the unlabeled data. Mislabeled observations are circled in the plot.

```
gscatter(unlabeledTbl.Displacement,unlabeledTbl.Weight, ...
    fittedLabels)
hold on
plot(wrongTbl.Displacement,wrongTbl.Weight, ...
```

```
    'ko','MarkerSize',8)
xlabel('Displacement')
ylabel('Weight')
legend('4 cylinders','6 cylinders','8 cylinders','Mislabeled')
title('Fitted Labels for Unlabeled Data')
```



Input Arguments

`Tbl` — Labeled sample data

table

Labeled sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor. Optionally, `Tbl` can contain one additional column for the response variable (vector of labels). Multicolumn variables and cell arrays other than cell arrays of character vectors are not supported.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, specify a formula using `formula`.

If `Tbl` does not contain the response variable, specify a response variable using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`UnlabeledTbl` — Unlabeled sample data
table

Unlabeled sample data, specified as a table. Each row of `UnlabeledTbl` corresponds to one observation, and each column corresponds to one predictor. `UnlabeledTbl` must contain the same predictors as those contained in `Tbl`.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable contains the class labels for the sample data in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors.

The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value pair argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~X1+X2+X3'`. In this form, `Y` represents the response variable, and `X1`, `X2`, and `X3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Class labels
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class labels, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors.

- If `Y` is a character array, then each element of the class labels must correspond to one row of the array.
- The length of `Y` must be equal to the number of rows in `Tbl` or `X`.

- A good practice is to specify the class order by using the `ClassNames` name-value pair argument.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Labeled predictor data

numeric matrix

Labeled predictor data, specified as a numeric matrix.

By default, each row of `X` corresponds to one observation, and each column corresponds to one predictor.

The length of `Y` and the number of observations in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

`UnlabeledX` — Unlabeled predictor data

numeric matrix

Unlabeled predictor data, specified as a numeric matrix. By default, each row of `UnlabeledX` corresponds to one observation, and each column corresponds to one predictor. `UnlabeledX` must have the same predictors as `X`, in the same order.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data. Whether the software removes observations with missing values depends on the underlying classifier type (`Learner`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitsemiself(Tbl,'Y',UnlabeledTbl,'Learner',templateSVM('Standardize',true),'IterationLimit',2e3)` specifies to use a binary support vector machine (SVM) learner, standardize the numeric predictors, and run a maximum of 2000 iterations.

`Learner` — Underlying classifier type

`'svm'` | `'discriminant'` | `'kernel'` | `'knn'` | `'linear'` | `'naivebayes'` | `'tree'` | ...

Underlying classifier type, specified as the comma-separated pair consisting of `'Learner'` and one of the values in this table.

Value	Description
`'discriminant'` or `templateDiscriminant` object	Discriminant analysis classifier

Value	Description
`templateECOC` object	Multiclass error-correcting output codes (ECOC) model — `templateECOC('Learners',templateSVM('KernelFunction','gaussian'))` is the default for multiclass classification.
`templateEnsemble` object	Ensemble classification model
`'kernel'` or `templateKernel` object	Kernel classification model (for binary classification only)
`'knn'` or `templateKNN` object	k-nearest neighbor model
`'linear'` or `templateLinear` object	Linear classification model (for binary classification only)
`'svm'` or `templateSVM` object	Support vector machine (SVM) classifier (for binary classification only) — `templateSVM('KernelFunction','gaussian')` is the default for binary classification.
`'tree'` or `templateTree` object	Binary decision classification tree

Example: `'Learner','tree'`

Example: `'Learner',templateEnsemble('AdaBoostM1',100,'tree')`

`IterationLimit` — Maximum number of self-training iterations
`1e3` (default) | positive integer scalar

Maximum number of self-training iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer scalar. The `fitsemiself` function returns `Mdl`, which contains the fitted labels and scores, when this limit is reached, even if the algorithm does not converge.

Example: `'IterationLimit',2e3`

Data Types: `single` | `double`

`ScoreThreshold` — Score threshold for fitted labels
numeric scalar

Score threshold for fitted labels, specified as the comma-separated pair consisting of `'ScoreThreshold'` and a numeric scalar. At each iteration of the algorithm, the software makes label predictions for the unlabeled observations by using the specified `Learner`, and calculates scores for these predictions. Unlabeled observations with prediction scores greater than or equal to the score threshold are treated as labeled observations in the next iteration, where the label is the predicted label. By default, `ScoreThreshold` is `0.1` for binary classification and `–0.1` for multiclass classification.

Example: `'ScoreThreshold',0.2`

Data Types: `single` | `double`

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table. The descriptions assume that the predictor data has observations in rows and predictors in columns.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitsemiself` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table, `fitsemiself` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. However, learners that use decision trees assume that mathematically ordered categorical vectors are continuous variables. If the predictor data is a matrix, `fitsemiself` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value pair argument.

For more information on how different fitting functions and, therefore, different learners treat categorical predictors, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for labeling

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of the classes to use for labeling, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `ClassNames` must have the same data type as `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `'ClassNames'` to:

- Order the classes.

- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `'ClassNames'` to specify the column order of classification scores in `Mdl.LabelScores`.
- Select a subset of classes for labeling. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the underlying classifier `Learner` using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply predictor data.

- If you supply `X`, `Y`, and `UnlabeledX`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X` and `UnlabeledX`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. Assuming that `X` has the default orientation, with observations in rows and predictors in columns, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl` and `UnlabeledTbl`, then you can use `'PredictorNames'` to choose which predictor variables to use. That is, `fitsemiself` uses only the predictor variables in `PredictorNames` and the response variable to label the unlabeled data.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as the comma-separated pair consisting of `'ResponseName'` and a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`NumBins` — Number of bins for numeric predictors
`[]` (default) | positive integer scalar

Number of bins for the numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

- If the `'NumBins'` value is empty (default), then the software does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar, then the software bins every numeric predictor into a specified number of equiprobable bins, and then grows trees on the bin indices instead of the original data.

  - If the `'NumBins'` value exceeds the number (u) of unique values for a predictor, then `fitsemiself` bins the predictor into u bins.
  - `fitsemiself` does not bin categorical predictors.

  When you use a large data set, this binning option speeds up classifier training, but causes a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the `'NumBins'` value depending on the accuracy and training speed.

Note This argument is valid only when the `Learner` value is a `templateECOC` or `templateEnsemble` object that uses tree learners.

Example: `'NumBins',50`

Data Types: `single` | `double`

`ObservationsIn` — Observation dimension for predictor data `X` and `UnlabeledX`
`'rows'` (default) | `'columns'`

Observation dimension for the predictor data `X` and `UnlabeledX`, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'rows'` or `'columns'`. For linear classification models, if you orient `X` and `UnlabeledX` so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you can experience a reduction in execution time.

Note The `'columns'` value is valid only when the `Learner` value is a binary linear classification model (`'linear'` or `templateLinear`) or an ECOC model with linear binary learners (for example, `templateECOC('Learners','linear')`).

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

Output Arguments

`Mdl` — Semi-supervised self-training classifier
`SemiSupervisedSelfTrainingModel` object

Semi-supervised self-training classifier, returned as a `SemiSupervisedSelfTrainingModel` object. Use dot notation to access the object properties. For example, to get the fitted labels for the unlabeled data and their corresponding scores, enter `Mdl.FittedLabels` and `Mdl.LabelScores`, respectively.

Algorithms

The algorithm begins by training a user-specified classifier (`Learner`), first trained on the labeled data alone, and then uses that classifier to make label predictions for the unlabeled data. Next, the algorithm provides scores for the predictions, and then treats the predictions as true labels for the next training cycle of the classifier if the scores are above a threshold (`ScoreThreshold`). This process repeats until the label predictions converge or the iteration limit (`IterationLimit`) is reached.

References

[1] Abney, Steven. "Understanding the Yarowsky Algorithm." *Computational Linguistics* 30, no. 3 (September 2004): 365–95. https://doi.org/10.1162/0891201041850876.

[2] Yarowsky, David. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189–96. Cambridge, Massachusetts: Association for Computational Linguistics, 1995. https://doi.org/10.3115/981658.981684.

See Also

`SemiSupervisedSelfTrainingModel` | `fitsemigraph` | `predict` | `templateDiscriminant` | `templateECOC` | `templateEnsemble` | `templateKNN` | `templateKernel` | `templateLinear` | `templateSVM` | `templateTree`

Topics

"Label Data Using Semi-Supervised Learning Techniques" on page 18-250

Introduced in R2020b

fitrauto

Automatically select regression model with optimized hyperparameters

Syntax

```
Mdl = fitrauto(Tbl,ResponseVarName)
Mdl = fitrauto(Tbl,formula)
Mdl = fitrauto(Tbl,Y)

Mdl = fitrauto(X,Y)

Mdl = fitrauto( ___ ,Name,Value)
[Mdl,OptimizationResults] = fitrauto( ___ )
```

Description

Given predictor and response data, `fitrauto` automatically tries a selection of regression model types with different hyperparameter values. The function uses Bayesian optimization to select models and their hyperparameter values, and computes the following for each model: log(1 + *valLoss*), where *valLoss* is the cross-validation mean squared error (MSE). After the optimization is complete, `fitrauto` returns the model, trained on the entire data set, that is expected to best predict the responses for new data. You can use the `predict` and `loss` object functions of the returned model to predict on new data and compute the test set MSE, respectively.

Use `fitrauto` when you are uncertain which model types best suit your data. For information on alternative methods for tuning hyperparameters of regression models, see “Alternative Functionality” on page 33-2114.

`Mdl = fitrauto(Tbl,ResponseVarName)` returns a regression model `Mdl` with tuned hyperparameters. The table `Tbl` contains the predictor variables and the response variable, where `ResponseVarName` is the name of the response variable.

`Mdl = fitrauto(Tbl,formula)` uses `formula` to specify the response variable and the predictor variables to consider among the variables in `Tbl`.

`Mdl = fitrauto(Tbl,Y)` uses the predictor variables in table `Tbl` and the response values in vector `Y`.

`Mdl = fitrauto(X,Y)` uses the predictor variables in matrix `X` and the response values in vector `Y`.

`Mdl = fitrauto( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, use the `HyperparameterOptimizationOptions` name-value pair argument to specify how the Bayesian optimization is performed.

`[Mdl,OptimizationResults] = fitrauto( ___ )` additionally returns `OptimizationResults`, a `BayesianOptimization` object containing the results of the model selection and hyperparameter tuning process.

Examples

Automatically Select Regression Model Using Table Data

Use `fitrauto` to automatically select a regression model with optimized hyperparameters, given predictor and response data stored in a table.

Load Data

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `MPG`.

```
cars = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,Origin,Weight,MPG);
```

Partition Data

Partition the data into training and test sets. Use approximately 80% of the observations for the model selection and hyperparameter tuning process, and 20% of the observations to test the performance of the final model returned by `fitrauto`. Use `cvpartition` to partition the data.

```
rng('default') % For reproducibility of the data partition
c = cvpartition(length(MPG),'Holdout',0.2);
trainingIdx = training(c); % Training set indices
carsTrain = cars(trainingIdx,:);
testIdx = test(c); % Test set indices
carsTest = cars(testIdx,:);
```

Run `fitrauto`

Pass the training data to `fitrauto`. By default, `fitrauto` determines appropriate model types to try, uses Bayesian optimization to find good hyperparameter values, and returns a trained model `Mdl` with the best expected performance. Additionally, `fitrauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-2111.

Expect this process to take some time. To speed up the optimization process, consider running the optimization in parallel, if you have a Parallel Computing Toolbox™ license. To do so, pass `'HyperparameterOptimizationOptions',struct('UseParallel',true)` to `fitrauto` as a name-value pair argument.

```
Mdl = fitrauto(carsTrain,'MPG');
```

```
Learner types to explore: ensemble, svm, tree
Total iterations (MaxObjectiveEvaluations): 90
Total time (MaxTime): Inf

|=====================================================================================================
| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|=====================================================================================================
|    1 | Best   |           2.5161 |            1.4197 |           2.5161 |           2.5161 |
```

2	Accept	4.1439	0.68799	2.5161	2.5161	
3	Accept	4.144	5.5942	2.5161	2.5161	
4	Accept	3.1976	20.387	2.5161	2.5161	
5	Best	2.5041	0.13106	2.5041	2.5101	
6	Best	2.2096	7.0177	2.2096	2.5101	
7	Accept	2.7182	0.085053	2.2096	2.5057	
8	Accept	17.207	23.312	2.2096	2.5057	
9	Accept	4.1439	0.057226	2.2096	2.5057	
10	Best	2.1916	6.4368	2.1916	2.5057	

Iter	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)	Le
11	Accept	2.8889	0.1325	2.1916	2.5057	
12	Accept	4.1439	0.06362	2.1916	2.5057	
13	Accept	4.1439	0.048941	2.1916	2.5057	
14	Accept	2.2844	6.7538	2.1916	2.4806	
15	Accept	4.1439	0.054943	2.1916	2.4806	
16	Accept	4.1439	0.055455	2.1916	2.4806	

```
|   17 | Accept |           2.2075 |            6.7931 |           2.1916 |           2.1942 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   18 | Accept |           2.6056 |           0.08425 |           2.1916 |           2.1942 |

|   19 | Accept |           2.6056 |          0.087135 |           2.1916 |           2.1942 |

|   20 | Accept |           2.7182 |          0.062848 |           2.1916 |           2.1942 |

|======================================================================================================
| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|======================================================================================================
|   21 | Accept |           2.2402 |            6.8909 |           2.1916 |           2.2011 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   22 | Accept |           2.6056 |          0.059067 |           2.1916 |           2.2011 |

|   23 | Accept |           2.3016 |            4.8177 |           2.1916 |           2.1911 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   24 | Accept |           4.1439 |           0.05258 |           2.1916 |           2.1911 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   25 | Accept |            3.352 |          0.043391 |           2.1916 |           2.1911 |

|   26 | Accept |           4.1439 |          0.062799 |           2.1916 |           2.1911 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   27 | Accept |           2.3188 |            5.1633 |           2.1916 |           2.1884 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   28 | Accept |           2.4271 |            5.8444 |           2.1916 |           2.1908 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   29 | Accept |           2.6056 |           0.05574 |           2.1916 |           2.1908 |

|   30 | Accept |           4.1439 |          0.054692 |           2.1916 |           2.1908 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|======================================================================================================
| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|======================================================================================================
|   31 | Accept |           2.5241 |          0.051793 |           2.1916 |           2.1908 |

|   32 | Accept |           2.6443 |            5.2572 |           2.1916 |           2.1969 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   33 | Accept |           2.2537 |             5.641 |           2.1916 |           2.1931 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |
```

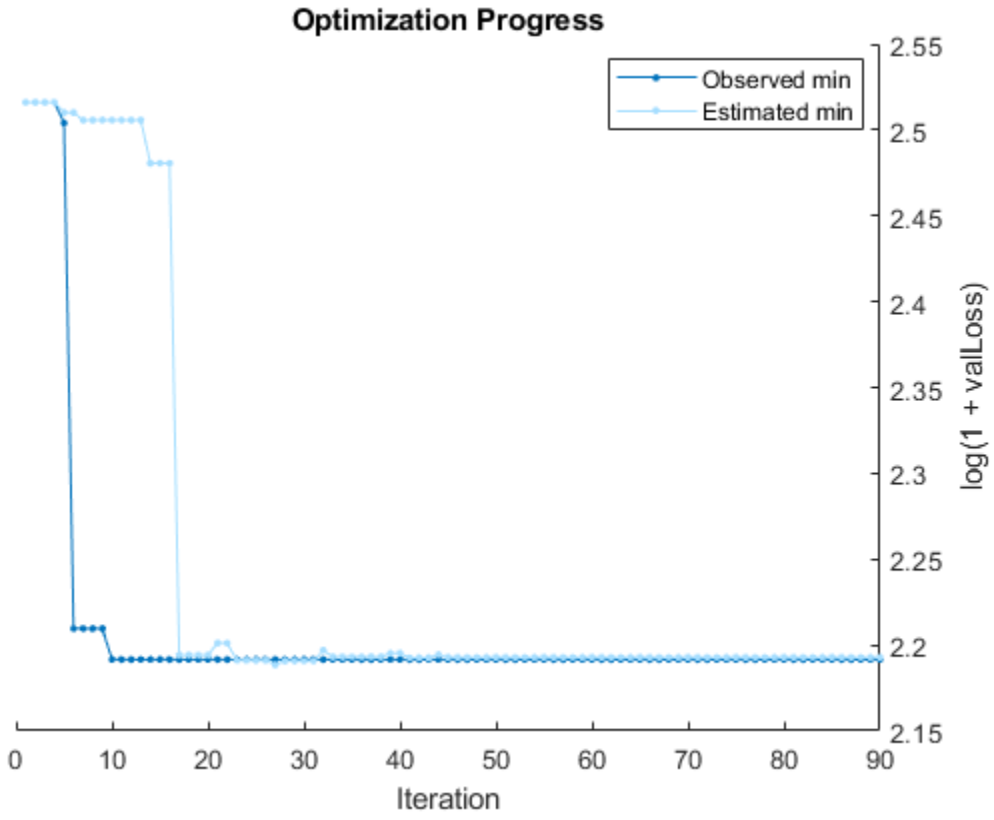
```
|   34 | Accept |           2.5448 |          0.053688 |           2.1916 |           2.1931 |
|   35 | Accept |           2.4438 |          0.045817 |           2.1916 |           2.1931 |
|   36 | Accept |           2.7182 |          0.062646 |           2.1916 |           2.1931 |
|   37 | Accept |           2.4749 |          0.048072 |           2.1916 |           2.1931 |
|   38 | Accept |           13.083 |            31.488 |           2.1916 |           2.1931 |
|   39 | Accept |           2.2446 |            5.2866 |           2.1916 |           2.1952 |
|   40 | Accept |           3.0919 |          0.040315 |           2.1916 |           2.1952 |

| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|   41 | Accept |           2.6335 |            6.4626 |           2.1916 |           2.1926 |
|   42 | Accept |           4.1439 |          0.048186 |           2.1916 |           2.1926 |
|   43 | Accept |           2.8766 |          0.075435 |           2.1916 |           2.1926 |
|   44 | Accept |           2.2402 |            5.6507 |           2.1916 |           2.1944 |
|   45 | Accept |           2.4576 |            5.4632 |           2.1916 |           2.1928 |
|   46 | Accept |            3.003 |          0.069043 |           2.1916 |           2.1928 |
|   47 | Accept |           4.1439 |           0.05544 |           2.1916 |           2.1928 |
|   48 | Accept |           11.118 |            31.888 |           2.1916 |           2.1928 |
|   49 | Accept |           3.0019 |          0.064077 |           2.1916 |           2.1928 |
|   50 | Accept |           4.1439 |          0.069826 |           2.1916 |           2.1928 |
```

Iter	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)	Le
51	Accept	4.1439	0.050005	2.1916	2.1928	
52	Accept	4.1439	0.054745	2.1916	2.1928	
53	Accept	2.8779	0.11876	2.1916	2.1928	
54	Accept	12.921	18.315	2.1916	2.1928	
55	Accept	2.9117	0.076534	2.1916	2.1928	
56	Accept	3.0276	0.064801	2.1916	2.1928	
57	Accept	7.1555	20.87	2.1916	2.1928	
58	Accept	2.9075	0.077548	2.1916	2.1928	
59	Accept	2.9168	0.068648	2.1916	2.1928	
60	Accept	4.1439	0.053635	2.1916	2.1928	

Iter	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)	Le
61	Accept	2.9456	0.073895	2.1916	2.1928	
62	Accept	4.1439	0.04979	2.1916	2.1928	
63	Accept	2.9095	0.065115	2.1916	2.1928	

```
|   64 | Accept |           4.1439 |          0.065248 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   65 | Accept |           4.1439 |          0.050181 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   66 | Accept |           2.6802 |            2.1221 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   67 | Accept |           4.1439 |          0.055315 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   68 | Accept |           4.1439 |          0.063152 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   69 | Accept |           4.1439 |          0.054083 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   70 | Accept |           4.1439 |           0.05377 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|======|========|==================|===================|==================|==================|
| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|======|========|==================|===================|==================|==================|
|   71 | Accept |           4.1439 |          0.053553 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   72 | Accept |           4.1439 |          0.050525 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   73 | Accept |           4.1439 |          0.068567 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   74 | Accept |           4.1439 |          0.055336 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   75 | Accept |           2.8707 |          0.095257 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   76 | Accept |           4.1439 |          0.059476 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   77 | Accept |           2.8682 |           0.10867 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |
```

```
|   78 | Accept |           2.7111 |           0.59286 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   79 | Accept |           2.7625 |           0.45668 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   80 | Accept |           3.2211 |          0.058812 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|======================================================================================================
| Iter | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated min    | Le
|      | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + valLoss) |
|======================================================================================================
|   81 | Accept |           4.1439 |           0.05754 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   82 | Accept |           4.1439 |          0.055356 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   83 | Accept |           2.7643 |           0.51644 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   84 | Accept |           4.1439 |           0.05615 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   85 | Accept |           4.1439 |          0.054645 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   86 | Accept |           2.8797 |          0.077707 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   87 | Accept |           4.1439 |          0.056472 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   88 | Accept |           4.1439 |          0.053269 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   89 | Accept |           4.1439 |          0.054778 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |

|   90 | Accept |           2.8676 |          0.097394 |           2.1916 |           2.1928 |
|      |        |                  |                   |                  |                  |
|      |        |                  |                   |                  |                  |
```




```
Optimization completed.
Total iterations: 90
Total elapsed time: 589.6892 seconds
Total time for training and validation: 245.2123 seconds

Best observed learner is an ensemble model with:
    Method:             LSBoost
    NumLearningCycles:      278
    MinLeafSize:             13
Observed log(1 + valLoss): 2.1916
Time for training and validation: 6.4368 seconds

Best estimated learner (returned model) is an ensemble model with:
    Method:             LSBoost
    NumLearningCycles:      278
    MinLeafSize:             13
Estimated log(1 + valLoss): 2.1928
Estimated time for training and validation: 5.8977 seconds

Documentation for fitrauto display
```

The final model returned by `fitrauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`carsTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

Evaluate the performance of the model on the test set. `testError` is based on the test set mean squared error (MSE). Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,carsTest,'MPG');
testError = log(1 + testMSE)
```

```
testError = 2.3194
```

Automatically Select Regression Model Using Matrix Data

Use `fitrauto` to automatically select a regression model with optimized hyperparameters, given predictor and response data stored in separate variables.

Load Data

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a matrix `X` containing the predictor variables `Acceleration`, `Cylinders`, and so on. Store the response variable `MPG` in the variable `Y`.

```
X = [Acceleration Cylinders Displacement Weight];
Y = MPG;
```

Create a variable indicating which predictors are categorical. `Cylinders` is the only categorical variable in `X`.

```
categoricalVars = [false true false false];
```

Partition Data

Partition the data into training and test sets. Use approximately 80% of the observations for the model selection and hyperparameter tuning process, and 20% of the observations to test the performance of the final model returned by `fitrauto`. Use `cvpartition` to partition the data.

```
rng('default') % For reproducibility of the partition
c = cvpartition(length(Y),'Holdout',0.20);
trainingIdx = training(c); % Indices for the training set
XTrain = X(trainingIdx,:);
YTrain = Y(trainingIdx);
testIdx = test(c); % Indices for the test set
XTest = X(testIdx,:);
YTest = Y(testIdx);
```

Run `fitrauto`

Pass the training data to `fitrauto`. By default, `fitrauto` determines appropriate model (or learner) types to try, uses Bayesian optimization to find good hyperparameter values for those models, and returns a trained model `Mdl` with the best expected performance. Specify the categorical predictors, and run the optimization in parallel (requires Parallel Computing Toolbox™). Return a second output `OptimizationResults` that contains the details of the Bayesian optimization.

Expect this process to take some time. By default, `fitrauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-2111.

```
options = struct('UseParallel',true);
[Mdl,OptimizationResults] = fitrauto(XTrain,YTrain, ...
    'CategoricalPredictors',categoricalVars, ...
    'HyperparameterOptimizationOptions',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
Copying objective function to workers...
Done copying objective function to workers.

Learner types to explore: ensemble, svm, tree
Total iterations (MaxObjectiveEvaluations): 90
Total time (MaxTime): Inf
```

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + va
1	5	Best	3.0205	1.7491	3.0205	3
2	5	Accept	3.0453	1.6641	3.0205	3
3	4	Accept	4.143	2.6491	3.0205	3
4	4	Accept	4.1434	7.5491	3.0205	3
5	3	Best	2.9188	8.6194	2.9188	3
6	3	Accept	4.143	0.83485	2.9188	3
7	6	Accept	2.9275	0.19147	2.9188	2
8	4	Accept	4.143	0.09353	2.9188	2
9	4	Accept	4.143	0.62222	2.9188	2
10	4	Accept	3.0205	0.21301	2.9188	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + va
11	6	Best	2.9167	7.3685	2.9167	2

```
|   12 |       4 | Accept |           3.6673 |            22.826 |           2.8696 |
|   13 |       4 | Best   |           2.8696 |            8.0371 |           2.8696 |
|   14 |       4 | Accept |           2.8696 |            7.6155 |           2.8696 |
|   15 |       3 | Accept |           2.8696 |            8.2039 |           2.8696 |
|   16 |       3 | Accept |           3.1971 |            0.1933 |           2.8696 |
|   17 |       6 | Accept |           3.0139 |           0.12079 |           2.8696 |
|   18 |       5 | Accept |            4.143 |           0.23444 |           2.8696 |
|   19 |       5 | Accept |           3.3225 |           0.19712 |           2.8696 |
|   20 |       3 | Accept |           2.8717 |            5.5602 |           2.8217 |
|======|=========|========|==================|===================|==================|
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + val
|======|=========|========|==================|===================|==================|
|   21 |       3 | Best   |           2.8217 |            4.2702 |           2.8217 |
|   22 |       3 | Accept |            4.143 |           0.11628 |           2.8217 |
|   23 |       6 | Accept |            4.143 |          0.067367 |           2.8217 |
|   24 |       5 | Accept |           3.1971 |          0.094668 |           2.8217 |
|   25 |       5 | Accept |            4.143 |           0.10279 |           2.8217 |
|   26 |       5 | Accept |            4.143 |           0.07812 |           2.8217 |
|   27 |       6 | Accept |            11.41 |            32.991 |           2.8217 |
|   28 |       6 | Accept |           2.8657 |            5.3498 |           2.8217 |
|   29 |       3 | Accept |           2.8689 |            6.3959 |           2.8217 |
```

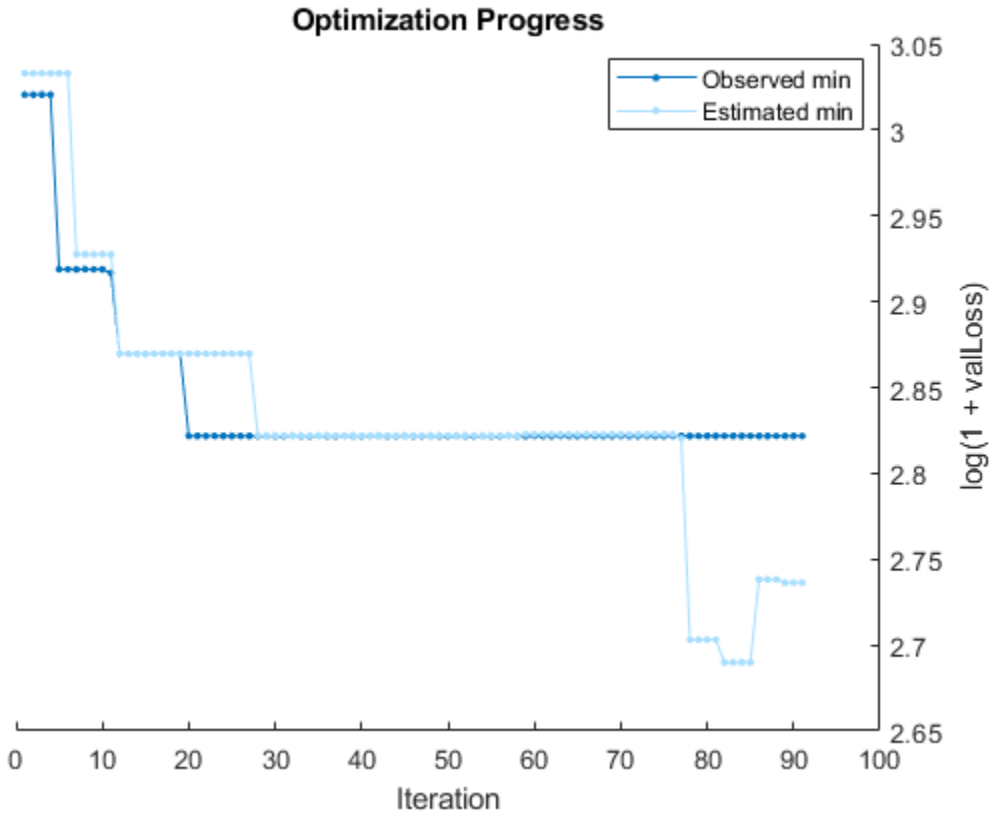
Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
30	3	Accept	2.8689	5.5649	2.8217	
31	3	Accept	2.9197	0.043492	2.8217	
32	3	Accept	2.9403	0.050907	2.8217	
33	6	Accept	4.143	0.1117	2.8217	
34	5	Accept	4.1924	16.905	2.8217	
35	5	Accept	2.9804	0.16413	2.8217	
36	4	Accept	2.8842	4.4972	2.8217	
37	4	Accept	3.112	0.40292	2.8217	
38	3	Accept	2.8686	5.942	2.8217	
39	3	Accept	4.143	0.048389	2.8217	
40	6	Accept	2.9487	0.053171	2.8217	

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
41	3	Accept	12.196	29.49	2.8217	
42	3	Accept	2.8717	4.5788	2.8217	
43	3	Accept	2.9093	0.24973	2.8217	
44	3	Accept	2.9804	0.18263	2.8217	
45	6	Accept	4.143	0.10875	2.8217	
46	5	Accept	2.8478	4.7428	2.8217	
47	5	Accept	4.143	1.2887	2.8217	
48	4	Accept	3.2214	16.267	2.8217	

```
|      |         |         |                |                |                |
|   49 |       4 | Accept  |          2.924 |         0.18896 |         2.8217 |
|      |         |         |                |                |                |
|   50 |       3 | Accept  |         3.1542 |           15.88 |         2.8217 |
|      |         |         |                |                |                |
|======|=========|=========|================|================|================|
| Iter | Active  | Eval    | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result  |                | & validation (sec)| log(1 + valLoss) | log(1 + val
|======|=========|=========|================|================|================|
|   51 |       3 | Accept  |         2.9143 |          3.0279 |         2.8217 |
|      |         |         |                |                |                |
|   52 |       6 | Accept  |          10.06 |          18.169 |         2.8217 |
|      |         |         |                |                |                |
|   53 |       2 | Accept  |         2.8989 |          7.2493 |         2.8217 |
|      |         |         |                |                |                |
|   54 |       2 | Accept  |         2.9049 |         0.55744 |         2.8217 |
|      |         |         |                |                |                |
|   55 |       2 | Accept  |         2.9804 |        0.066427 |         2.8217 |
|   56 |       2 | Accept  |         2.9032 |         0.17339 |         2.8217 |
|   57 |       2 | Accept  |          4.143 |        0.058535 |         2.8217 |
|      |         |         |                |                |                |
|   58 |       6 | Accept  |         2.9268 |         0.13315 |         2.8217 |
|      |         |         |                |                |                |
|   59 |       4 | Accept  |         2.9081 |          1.8508 |         2.8217 |
|      |         |         |                |                |                |
|   60 |       4 | Accept  |          4.143 |        0.059851 |         2.8217 |
|      |         |         |                |                |                |
|======|=========|=========|================|================|================|
| Iter | Active  | Eval    | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result  |                | & validation (sec)| log(1 + valLoss) | log(1 + val
|======|=========|=========|================|================|================|
|   61 |       4 | Accept  |         2.8335 |          4.6933 |         2.8217 |
|      |         |         |                |                |                |
|   62 |       5 | Accept  |         2.9049 |           1.902 |         2.8217 |
|      |         |         |                |                |                |
|   63 |       3 | Accept  |         2.8816 |          15.271 |         2.8217 |
|      |         |         |                |                |                |
|   64 |       3 | Accept  |         2.8773 |          9.1951 |         2.8217 |
|      |         |         |                |                |                |
```

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
65	3	Accept	2.9573	3.597	2.8217	2
66	6	Accept	2.9045	0.8479	2.8217	2
67	5	Accept	3.1507	0.089958	2.8217	2
68	5	Accept	3.1971	0.14914	2.8217	2
69	5	Accept	4.143	30.354	2.8217	2
70	5	Accept	2.9037	1.2103	2.8217	2
71	4	Accept	6.6283	17.347	2.8217	2
72	4	Accept	2.9655	15.866	2.8217	2
73	4	Accept	3.0653	0.064937	2.8217	2
74	2	Accept	5.9041	19.341	2.8217	2
75	2	Accept	2.874	14.077	2.8217	2
76	2	Accept	2.9116	0.1932	2.8217	2
77	6	Accept	2.9266	0.089378	2.8217	2
78	3	Accept	2.9134	0.2501	2.8217	2
79	3	Accept	4.143	0.086465	2.8217	2
80	3	Accept	2.9804	0.070321	2.8217	2
Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)

```
|   81 |       3 | Accept |            4.143 |          0.057551 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   82 |       6 | Accept |            2.905 |           0.35685 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   83 |       4 | Accept |            4.143 |          0.057428 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|   84 |       4 | Accept |           2.9968 |          0.053485 |           2.8217 |           2
|   85 |       4 | Accept |           2.9487 |          0.067551 |           2.8217 |           2

|   86 |       2 | Accept |           11.192 |             19.605 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|   87 |       2 | Accept |           2.8848 |            6.0635 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |
|   88 |       2 | Accept |           2.8919 |            2.8002 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   89 |       6 | Accept |           2.9712 |          0.073966 |           2.8217 |           2
|      |         |        |                  |                   |                  |
|      |         |        |                  |                   |                  |

|   90 |       5 | Accept |           3.0045 |            0.1149 |           2.8217 |           2
|======================================================================================================
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + val
|======================================================================================================
|   91 |       5 | Accept |            3.031 |          0.041383 |           2.8217 |           2
```




```
_____________________________________________________________
Optimization completed.
Total iterations: 91
Total elapsed time: 201.923 seconds
Total time for training and validation: 436.1019 seconds

Best observed learner is an ensemble model with:
    Method:                Bag
    NumLearningCycles:     202
    MinLeafSize:             9
Observed log(1 + valLoss): 2.8217
Time for training and validation: 4.2702 seconds

Best estimated learner (returned model) is an svm model with:
    BoxConstraint:    0.0013447
    KernelScale:         362.68
    Epsilon:           0.096474
Estimated log(1 + valLoss): 2.7363
Estimated time for training and validation: 0.29514 seconds

Documentation for fitrauto display
```

The final model returned by `fitrauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`XTrain` and `YTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Evaluate Test Set Performance

Evaluate the performance of the model on the test set. `testError` is based on the test set mean squared error (MSE). Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,XTest,YTest);
testError = log(1 + testMSE)
```

```
testError = 2.9873
```

Compare Optimized and Simple Linear Regression Model

Use `fitrauto` to automatically select a regression model with optimized hyperparameters, given predictor and response data stored in a table. Compare the performance of the resulting regression model to the performance of a simple linear regression model created with `fitlm`.

Load and Partition Data

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Convert the `Cylinders` variable to a `categorical` variable. Create a table containing the predictor variables `Acceleration`, `Cylinders`, `Displacement`, and so on, as well as the response variable `MPG`.

```
load carbig
Cylinders = categorical(Cylinders);
cars = table(Acceleration,Cylinders,Displacement, ...
    Horsepower,Model_Year,Origin,Weight,MPG);
```

Partition the data into training and test sets. Use approximately 80% of the observations for training, and 20% of the observations for testing. Use `cvpartition` to partition the data.

```
rng('default') % For reproducibility of the data partition
c = cvpartition(length(MPG),'Holdout',0.2);
trainingIdx = training(c); % Training set indices
carsTrain = cars(trainingIdx,:);
testIdx = test(c); % Test set indices
carsTest = cars(testIdx,:);
```

Run `fitrauto`

Pass the training data to `fitrauto`. By default, `fitrauto` determines appropriate model types to try, uses Bayesian optimization to find good hyperparameter values, and returns a trained model `autoMdl` with the best expected performance. Specify to optimize over all optimizable hyperparameters and run the optimization in parallel (requires Parallel Computing Toolbox™).

Expect this process to take some time. By default, `fitrauto` provides a plot of the optimization and an iterative display of the optimization results. For more information on how to interpret these results, see “Verbose Display” on page 33-2111.

```
options = struct('UseParallel',true);
autoMdl = fitrauto(carsTrain,'MPG','OptimizeHyperparameters','all', ...
    'HyperparameterOptimizationOptions',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
Copying objective function to workers...
Done copying objective function to workers.
Learner types to explore: ensemble, svm, tree
Total iterations (MaxObjectiveEvaluations): 90
Total time (MaxTime): Inf
```

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
1	4	Accept	3.5168	2.3522	2.6193	2
2	4	Accept	2.6633	2.3211	2.6193	2
3	4	Best	2.6193	2.3426	2.6193	2
4	4	Accept	4.1439	3.798	2.6193	2
5	2	Accept	2.3998	8.4115	2.2082	2
6	2	Best	2.2082	9.4573	2.2082	2
7	2	Accept	4.1439	0.37564	2.2082	2
8	6	Accept	4.1439	1.1014	2.2082	2
9	6	Accept	2.7394	1.8778	2.2082	2
10	3	Accept	2.5987	6.4235	2.2082	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
11	3	Accept	3.2336	6.93	2.2082	2
12	3	Accept	2.682	6.4489	2.2082	2
13	3	Accept	2.4795	0.40654	2.2082	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
14	5	Accept	4.1439	0.28349	2.2082	2
15	5	Accept	2.845	0.3422	2.2082	2
16	5	Accept	2.7106	5.6397	2.2082	2
17	3	Accept	6.0492	4.8396	2.2082	2
18	3	Accept	2.2932	5.4011	2.2082	2
19	3	Accept	2.6686	0.22924	2.2082	2
20	6	Accept	4.1439	1.5528	2.2082	2
21	4	Accept	4.143	0.10249	2.2082	2
22	4	Accept	4.143	0.31889	2.2082	2
23	4	Accept	3.5024	0.16777	2.2082	2
24	2	Accept	4.1439	33.887	2.2082	2
25	2	Accept	2.3583	5.6469	2.2082	2
26	2	Accept	4.1439	0.08045	2.2082	2
27	6	Accept	4.1439	0.11643	2.2082	2
28	6	Accept	2.6334	0.69581	2.2082	2
29	2	Accept	5.9862	5.4794	2.2082	2
30	2	Accept	2.2456	7.9201	2.2082	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
31	2	Accept	2.9745	7.6297	2.2082	2
32	2	Accept	2.6819	6.0671	2.2082	2
33	2	Accept	2.927	0.11096	2.2082	2
34	4	Accept	4.1439	0.11416	2.2082	2
35	4	Accept	2.9904	0.096707	2.2082	2
36	4	Accept	2.6513	0.10209	2.2082	2
37	4	Accept	2.2442	7.3117	2.2082	2
38	4	Accept	2.345	5.3353	2.2082	2
39	3	Accept	3.1314	4.9709	2.2082	2
40	3	Accept	4.1439	0.076184	2.2082	2
Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
41	5	Accept	2.611	15.327	2.2082	2
42	5	Accept	4.1439	0.11014	2.2082	2
43	3	Best	2.1839	5.4291	2.1839	2
44	3	Accept	2.9934	0.080707	2.1839	2

45	3	Accept	4.1439	0.090101	2.1839	2
46	6	Accept	2.2144	5.1987	2.1839	2
47	4	Accept	2.2877	5.9143	2.1839	2
48	4	Accept	3.0763	0.13281	2.1839	2
49	4	Accept	4.1439	1.3244	2.1839	2
50	3	Accept	4.1439	30.675	2.1839	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
51	3	Accept	2.2728	5.7106	2.1839	2
52	6	Accept	2.4156	4.7264	2.1839	2
53	4	Accept	4.1439	0.07497	2.1839	2
54	4	Accept	2.3527	5.0415	2.1839	2
55	4	Accept	2.5975	0.086879	2.1839	2
56	3	Accept	4.1439	30.921	2.1839	2
57	3	Accept	2.2488	6.5278	2.1839	2
58	5	Accept	2.2375	6.3071	2.1839	2
59	5	Accept	2.2145	5.1044	2.1839	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
60	2	Accept	2.2771	5.7956	2.1839	
61	2	Accept	2.4755	4.7265	2.1839	
62	2	Accept	2.6456	0.082657	2.1839	
63	2	Accept	2.5294	5.4057	2.1839	
64	6	Accept	2.3409	5.5648	2.1839	
65	2	Accept	2.3209	7.4037	2.1839	
66	2	Accept	2.5782	5.5342	2.1839	
67	2	Accept	2.2537	7.2619	2.1839	
68	2	Accept	4.1426	4.7455	2.1839	
69	2	Accept	2.9366	0.094426	2.1839	
70	6	Accept	2.2338	5.1862	2.1839	

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated min log(1 + valLoss)
71	3	Accept	4.143	4.4939	2.1839	
72	3	Accept	4.1439	0.076138	2.1839	
73	3	Accept	4.143	0.074774	2.1839	

74	3	Accept	2.3197	5.1527	2.1839	2
75	6	Accept	2.5309	4.7886	2.1839	2
76	3	Accept	2.2935	5.3774	2.1839	2
77	3	Accept	2.3649	4.9029	2.1839	2
78	3	Accept	3.3541	0.065977	2.1839	2
79	3	Accept	2.3592	5.0572	2.1839	2
80	6	Accept	2.4009	4.5836	2.1839	2

Iter	Active workers	Eval result	log(1 + valLoss)	Time for training & validation (sec)	Observed min log(1 + valLoss)	Estimated m log(1 + val
81	2	Accept	10.894	31.052	2.1839	2
82	2	Accept	2.216	5.8077	2.1839	2
83	2	Accept	2.2547	5.6201	2.1839	2
84	2	Accept	4.1439	0.11346	2.1839	2
85	2	Accept	4.1439	0.10597	2.1839	2
86	6	Accept	2.219	6.1221	2.1839	2
87	2	Accept	2.8447	0.17007	2.1839	2
88	2	Accept	4.1439	0.079667	2.1839	2

```
|      |         |        |                  |                   |                  |                
|      |         |        |                  |                   |                  |                
|   89 |       2 | Accept |           2.2928 |            4.9488 |           2.1839 |              2
|      |         |        |                  |                   |                  |                
|      |         |        |                  |                   |                  |                
|      |         |        |                  |                   |                  |                
|   90 |       2 | Accept |           2.4525 |           0.12704 |           2.1839 |              2
|      |         |        |                  |                   |                  |                
|      |         |        |                  |                   |                  |                
|=============================================================================================
| Iter | Active  | Eval   | log(1 + valLoss) | Time for training | Observed min     | Estimated m
|      | workers | result |                  | & validation (sec)| log(1 + valLoss) | log(1 + val
|=============================================================================================
|   91 |       2 | Accept |           4.1439 |          0.093014 |           2.1839 |              2
|      |         |        |                  |                   |                  |                
|      |         |        |                  |                   |                  |                

__________________________________________________________
Optimization completed.
Total iterations: 91
Total elapsed time: 235.4082 seconds
Total time for training and validation: 426.1606 seconds

Best observed learner is an ensemble model with:
    Method:             LSBoost
    LearnRate:         0.079786
    MinLeafSize:             16
    NumVariablesToSample:   NaN
Observed log(1 + valLoss): 2.1839
Time for training and validation: 5.4291 seconds

Best estimated learner (returned model) is an ensemble model with:
    Method:             LSBoost
    LearnRate:         0.079786
    MinLeafSize:             16
    NumVariablesToSample:   NaN
Estimated log(1 + valLoss): 2.1981
Estimated time for training and validation: 5.3328 seconds

Documentation for fitrauto display
```

The final model returned by `fitrauto` corresponds to the best estimated learner. Before returning the model, the function retrains it using the entire training data (`carsTrain`), the listed `Learner` (or model) type, and the displayed hyperparameter values.

Create Simple Model

Create a simple linear regression model `linearMdl` by using the `fitlm` function.

```
linearMdl = fitlm(carsTrain);
```

Although the `linearMdl` object does not have the exact same properties and methods as the `autoMdl` object, you can use both models to predict response values for new data by using the `predict` object function.

Compare Test Set Performance of Models

Compare the performance of the `linearMdl` and `autoMdl` models on the test data set. For each model, compute the test set mean squared error (MSE). Smaller MSE values indicate better performance.

```
ypred = predict(linearMdl,carsTest);
linearMSE = mean((carsTest.MPG-ypred).^2,'omitnan')
```

```
linearMSE = 11.0981
```

```
autoMSE = loss(autoMdl,carsTest,'MPG')
```

```
autoMSE = 8.8024
```

The `autoMdl` model seems to outperform the `linearMdl` model.

Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not accepted.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, specify a formula using `formula`.

If `Tbl` does not contain the response variable, specify a response variable using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training a model.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Y — Response data
numeric vector

Response data, specified as a numeric vector. The length of `Y` must be equal to the number of rows in `Tbl` or `X`.

To specify the response variable name, use the `ResponseName` name-value pair argument.

Data Types: `single` | `double`

X — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one predictor.

The length of `Y` and the number of rows in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing data. The software removes rows of data corresponding to missing values in the response variable. However, the treatment of missing values in the predictor data `X` or `Tbl` varies among models (or learners).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',200,'Verbose',2)` specifies to run 200 iterations of the optimization process (that is, try 200 model hyperparameter combinations), and to display information in the Command Window about the next model hyperparameter combination to be evaluated.

Optimizer Options

Learners — Types of regression models

'auto' (default) | 'all' | 'all-linear' | 'all-nonlinear' | one or more learner names

Types of regression models to try during the optimization, specified as the comma-separated pair consisting of 'Learners' and a value in the first table below or one or more learner names in the second table. Specify multiple learner names as a string or cell array.

Value	Description
'auto'	fitrauto automatically selects a subset of learners, suitable for the given predictor and response data. The learners can have model hyperparameter values that differ from the default. For more information, see "Automatic Selection of Learners" on page 33-2113.
'all'	fitrauto selects all possible learners.
'all-linear'	fitrauto selects linear ('linear') learners.
'all-nonlinear'	fitrauto selects all nonlinear learners: 'ensemble', 'gp', 'kernel', 'svm' (with a Gaussian or polynomial kernel), and 'tree'.

Note For greater efficiency, fitrauto does not select the following combinations of models when you specify one of the previous values.

- 'kernel' and 'svm' (with a Gaussian kernel) — fitrauto chooses the first when the predictor data has more than 11,000 observations, and the second otherwise.
- 'linear' and 'svm' (with a linear kernel) — fitrauto chooses the first.

Learner Name	Description
'ensemble'	Ensemble regression model
'gp'	Gaussian process regression model
'kernel'	Kernel regression model
'linear'	Linear regression model for high-dimensional data
'svm'	Support vector machine regression model
'tree'	Binary decision regression tree

Example: 'Learners','all'

Example: 'Learners','ensemble'

Example: 'Learners',{'gp','svm'}

OptimizeHyperparameters — Hyperparameters to optimize

'auto' (default) | 'all'

Hyperparameters to optimize, specified as the comma-separated pair consisting of 'OptimizeHyperparameters' and 'auto' or 'all'. The optimizable hyperparameters depend on the model (or learner), as described in this table.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'ensemble'	Method, NumLearningCycles, LearnRate, MinLeafSize	MaxNumSplits, NumVariablesToSample	When the ensemble Method value is a boosting method, the ensemble NumBins value is 50. For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.
'gp'	Sigma	BasisFunction, KernelFunction, KernelScale (KernelParameters), Standardize	The fitrauto function ignores all ARD kernel options and, therefore, chooses among the KernelFunction values of 'exponential', 'matern32', 'matern52', 'rationalquadratic', and 'squaredexponential' when the OptimizeHyperparameters value is 'all'. For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.
'kernel'	Epsilon, KernelScale, Lambda	Learner, NumExpansionDimensions	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'linear'	Lambda, Learner	Regularization	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.
'svm'	BoxConstraint, Epsilon, KernelScale	KernelFunction, PolynomialOrder, Standardize	• When the Learners value is 'all-linear', the fitrauto function does not optimize the KernelFunction or PolynomialOrder hyperparameters when the OptimizeHyperparameters value is 'all'. • When the Learners value is 'all-nonlinear', the fitrauto function chooses among the KernelFunction values of 'gaussian' and 'polynomial', regardless of the OptimizeHyperparameters value. • The Standardize value is true when the OptimizeHyperparameters value is 'auto'. For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.

Learner Name	Hyperparameters for 'auto'	Additional Hyperparameters for 'all'	Notes
'tree'	MinLeafSize	MaxNumSplits	For more information, including hyperparameter search ranges, see OptimizeHyperparameters. Note that you cannot change hyperparameter search ranges when you use fitrauto.

Note When 'Learners' is set to a value other than 'auto', the default values for the model hyperparameters not being optimized match the default fit function values, unless otherwise indicated in the table notes. When 'Learners' is set to 'auto', the optimized hyperparameter search ranges and nonoptimized hyperparameter values can vary, depending on the characteristics of the training data. For more information, see "Automatic Selection of Learners" on page 33-2113.

Example: 'OptimizeHyperparameters','all'

HyperparameterOptimizationOptions — Options for optimization
structure

Options for the optimization, specified as the comma-separated pair consisting of 'HyperparameterOptimizationOptions' and a structure. All fields in the structure are optional.

Field Name	Values	Default
MaxObjectiveEvaluations	Maximum number of iterations (objective function evaluations)	30*L, where L is the number of learners (see Learners)
MaxTime	Time limit, specified as a positive real number. The time limit is in seconds, as measured by tic and toc. Run time can exceed MaxTime because MaxTime does not interrupt function evaluations.	Inf
ShowPlots	Logical value indicating whether to show plots. If true, this field plots the best observed and estimated objective function values (so far) against the iteration number.	true
SaveIntermediateResults	Logical value indicating whether to save results. If true, this field overwrites a workspace variable named 'BayesoptResults' at each iteration. The variable is a BayesianOptimization object.	false

Field Name	Values	Default
`Verbose`	Display at the command line: • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with additional information about the next point to be evaluated	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Specify only one of the following three options.		
`CVPartition`	`cvpartition` object, created by `cvpartition`	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	Scalar in the range `(0,1)` representing the holdout fraction	
`Kfold`	Integer greater than 1	

Example: `'HyperparameterOptimizationOptions',struct('UseParallel',true)`

Regression Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrauto` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrauto` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. However, learners that use decision trees assume that mathematically ordered categorical vectors are continuous variables. If the predictor data is a matrix (`X`), `fitrauto` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value pair argument.

For more information on how fitting functions treat categorical predictors, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrauto` uses only the predictor variables in `PredictorNames` and the response variable during training.

- `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
- By default, `PredictorNames` contains the names of all predictor variables.
- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Weights` — Observation weights
positive numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a positive numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`Mdl` — Trained regression model
regression model object

Trained regression model, returned as one of the regression model objects in this table.

Learner Name	Returned Model Object
`'ensemble'`	`CompactRegressionEnsemble`
`'gp'`	`CompactRegressionGP`

Learner Name	Returned Model Object
'kernel'	RegressionKernel
'linear'	RegressionLinear
'svm'	CompactRegressionSVM
'tree'	CompactRegressionTree

OptimizationResults — Optimization results
BayesianOptimization object

Optimization results, returned as a BayesianOptimization object. For more information on the Bayesian optimization process, see “Bayesian Optimization” on page 33-2113.

More About

Verbose Display

When you set the Verbose field of the HyperparameterOptimizationOptions name-value pair argument to 1 or 2, the fitrauto function provides an iterative display of the optimization results.

The following table describes the columns in the display and their entries.

Column Name	Description
Iter	Iteration number — You can set a limit to the number of iterations by using the MaxObjectiveEvaluations field of the 'HyperparameterOptimizationOptions' name-value pair argument.
Active workers	Number of active parallel workers — This column appears only when you run the optimization in parallel by setting the UseParallel field of the 'HyperparameterOptimizationOptions' name-value pair argument to true.
Eval result	One of the following evaluation results: • Best — The learner and hyperparameter values at this iteration give the minimum observed validation loss computed so far. That is, the log(1 + valLoss) value is the smallest computed so far. • Accept — The learner and hyperparameter values at this iteration give meaningful (for example, non-NaN) observed and estimated validation loss values. • Error — The learner and hyperparameter values at this iteration result in an error (for example, a log(1 + valLoss) value of NaN).

Column Name	Description
`log(1 + valLoss)`	Log-transformed validation loss computed for the learner and hyperparameter values at this iteration — In particular, `fitrauto` computes log(1 + *valLoss*), where *valLoss* is the cross-validation mean squared error (MSE) by default. You can change the validation scheme by using the `CVPartition`, `Holdout`, or `Kfold` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.
`Time for training & validation (sec)`	Time taken to train and compute the validation loss for the model with the learner and hyperparameter values at this iteration (in seconds) — In particular, this value excludes the time required to update the objective function model maintained by the Bayesian optimization process. For more details, see “Bayesian Optimization” on page 33-2113.
`Observed min log(1 + valLoss)`	Observed minimum log-transformed validation loss computed so far — This value corresponds to the smallest `log(1 + valLoss)` value computed so far in the optimization process. By default, `fitrauto` returns a plot of the optimization that displays dark blue points for the observed minimum log-transformed validation loss values. This plot does not appear when the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument is set to `false`.
`Estimated min log(1 + valLoss)`	Estimated minimum log-transformed validation loss — At each iteration, `fitrauto` updates an objective function model maintained by the Bayesian optimization process and uses this model to estimate the minimum log-transformed validation loss. For more details, see “Bayesian Optimization” on page 33-2113. By default, `fitrauto` returns a plot of the optimization that displays light blue points for the estimated minimum log-transformed validation loss values. This plot does not appear when the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument is set to `false`.
`Learner`	Model type evaluated at this iteration — Specify the learners used in the optimization by using the `'Learners'` name-value pair argument.

Column Name	Description
`Hyperparameter: Value`	Hyperparameter values at this iteration — Specify the hyperparameters used in the optimization by using the `'OptimizeHyperparameters'` name-value pair argument.

The display also includes a description of two models:

- `Best observed learner` — This model, with the listed learner type and hyperparameter values, yields the final observed minimum validation loss (log-transformed).
- `Best estimated learner` — This model, with the listed learner type and hyperparameter values, yields the final estimated minimum validation loss (log-transformed). `fitrauto` retrains the model on the entire training data set and returns it as the `Mdl` output.

Tips

- Depending on the size of your data and the number of learners you specify, `fitrauto` can take some time to run. If you have a Parallel Computing Toolbox license, you can speed up computations by running the optimization in parallel. To do so, specify `'HyperparameterOptimizationOptions',struct('UseParallel',true)`. You can include other fields in the structure to control other aspects of the optimization. See `HyperparameterOptimizationOptions`.

Algorithms

Automatic Selection of Learners

When you specify `'Learners','auto'`, the `fitrauto` function analyzes the predictor and response data in order to choose appropriate learners. The function considers whether the data set has any of these characteristics:

- Categorical predictors
- Missing values for more than 5% of the data
- Wide data, where the number of predictors is greater than or equal to the number of observations
- High-dimensional data, where the number of predictors is greater than 100
- Large data, where the number of observations is greater than 50,000

The selected learners are always a subset of those listed in the `Learners` table. However, the associated models tried during the optimization process can have different default values for hyperparameters not being optimized, as well as different search ranges for hyperparameters being optimized.

Bayesian Optimization

The goal of Bayesian optimization, and optimization in general, is to find a point that minimizes an objective function. In the context of `fitrauto`, a point is a learner type together with a set of hyperparameter values for the learner (see `Learners` and `OptimizeHyperparameters`), and the objective function is log(1 + *valLoss*), where *valLoss* is the cross-validation mean squared error (MSE), by default. The Bayesian optimization implemented in `fitrauto` internally maintains a multi-

RegressionGP model of the objective function. That is, the objective function model splits along the learner type and, for a given learner, the model is a Gaussian process regression (GPR) model. (This underlying model differs from the single GPR model employed by other Statistics and Machine Learning Toolbox functions that use Bayesian optimization.) Bayesian optimization trains the underlying model by using objective function evaluations, and determines the next point to evaluate by using an acquisition function ('expected-improvement'). For more information, see "Expected Improvement" on page 10-4. The acquisition function balances between sampling at points with low modeled objective function values and exploring areas that are not well modeled yet. At the end of the optimization, fitrauto chooses the point with the minimum objective function model value, among the points evaluated during the optimization. For more information, see the 'Criterion','min-visited-mean' name-value pair argument of bestPoint.

Alternative Functionality

- If you are unsure which models work best for your data set, you can alternatively use the **Regression Learner** app. Using the app, you can perform hyperparameter tuning for different models, and choose the optimized model that performs best. Although you must select a specific model before you can tune the model hyperparameters, **Regression Learner** provides greater flexibility for selecting optimizable hyperparameters and setting hyperparameter values. The app also allows you to train a variety of linear regression models (see "Linear Regression Models" on page 24-14). However, you cannot optimize in parallel, choose 'linear' or 'kernel' learners, or specify observation weights in the app. For more information, see "Hyperparameter Optimization in Regression Learner App" on page 24-30.
- If you know which models might suit your data, you can alternatively use the corresponding model fit functions and specify the 'OptimizeHyperparameters' name-value pair argument to tune hyperparameters. You can compare the results across the models to select the best regression model. For an example of this process applied to classification models, see "Moving Towards Automating Model Selection Using Bayesian Optimization" on page 18-197.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the 'HyperparameterOptimizationOptions',struct('UseParallel',true) name-value pair argument in the call to this function.

For more general information about parallel computing, see "Run MATLAB Functions with Automatic Parallel Support" (Parallel Computing Toolbox).

See Also

fitrensemble | fitrgp | fitrkernel | fitrlinear | fitrsvm | fitrtree

Topics

"Automated Regression Model Selection with Bayesian Optimization" on page 18-214
"Hyperparameter Optimization in Regression Learner App" on page 24-30

Introduced in R2020b

fitrgam

Fit generalized additive model (GAM) for regression

Syntax

```
Mdl = fitrgam(Tbl,ResponseVarName)
Mdl = fitrgam(Tbl,formula)
Mdl = fitrgam(Tbl,Y)
Mdl = fitrgam(X,Y)
Mdl = fitrgam(___,Name,Value)
```

Description

`Mdl = fitrgam(Tbl,ResponseVarName)` returns a generalized additive model on page 33-2133 `Mdl` trained using the sample data contained in the table `Tbl`. The input argument `ResponseVarName` is the name of the variable in `Tbl` that contains the response values for regression.

`Mdl = fitrgam(Tbl,formula)` uses the model specification argument `formula` to specify the response variable and predictor variables in `Tbl`. You can specify a subset of predictor variables and interaction terms for predictor variables by using `formula`.

`Mdl = fitrgam(Tbl,Y)` uses the predictor variables in the table `Tbl` and the response values in the vector `Y`.

`Mdl = fitrgam(X,Y)` uses the predictors in the matrix `X` and the response values in the vector `Y`.

`Mdl = fitrgam(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in the previous syntaxes. For example, `'Interactions',5` specifies to include five interaction terms in the model. You can also specify a list of interaction terms using the `'Interactions'` name-value argument.

Examples

Train Generalized Additive Model

Train a univariate GAM, which contains linear terms for predictors. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 10 variables with information on the sales of properties in New York City in 2015. This example uses these variables to analyze the sale prices (`SALEPRICE`).

Preprocess the data set. Remove outliers, convert the `datetime` array (`SALEDATE`) to the month numbers, and move the response variable (`SALEPRICE`) to the last column.

```
idx = isoutlier(NYCHousing2015.SALEPRICE);
NYCHousing2015(idx,:) = [];
NYCHousing2015.SALEDATE = month(NYCHousing2015.SALEDATE);
NYCHousing2015 = movevars(NYCHousing2015,'SALEPRICE','After','SALEDATE');
```

Display the first three rows of the table.

```
head(NYCHousing2015,3)
```

```
ans=3×10 table
    BOROUGH    NEIGHBORHOOD       BUILDINGCLASSCATEGORY        RESIDENTIALUNITS    COMMERCIALUNI
    _______    ____________    ___________________________    ________________    _____________

       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}           1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}           1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}           1                   1
```

Train a univariate GAM for the sale prices. Specify the variables for `BOROUGH`, `NEIGHBORHOOD`, `BUILDINGCLASSCATEGORY`, and `SALEDATE` as categorical predictors.

```
Mdl = fitrgam(NYCHousing2015,'SALEPRICE','CategoricalPredictors',[1 2 3 9])
```

```
Mdl =
  RegressionGAM
           PredictorNames: {1x9 cell}
             ResponseName: 'SALEPRICE'
    CategoricalPredictors: [1 2 3 9]
        ResponseTransform: 'none'
                Intercept: 3.7518e+05
          NumObservations: 83517


  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object. The model display shows a partial list of the model properties. To view the full list of properties, double-click the variable name `Mdl` in the Workspace. The Variables editor opens for `Mdl`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the estimated intercept (constant) term of `Mdl`.

```
Mdl.Intercept
```

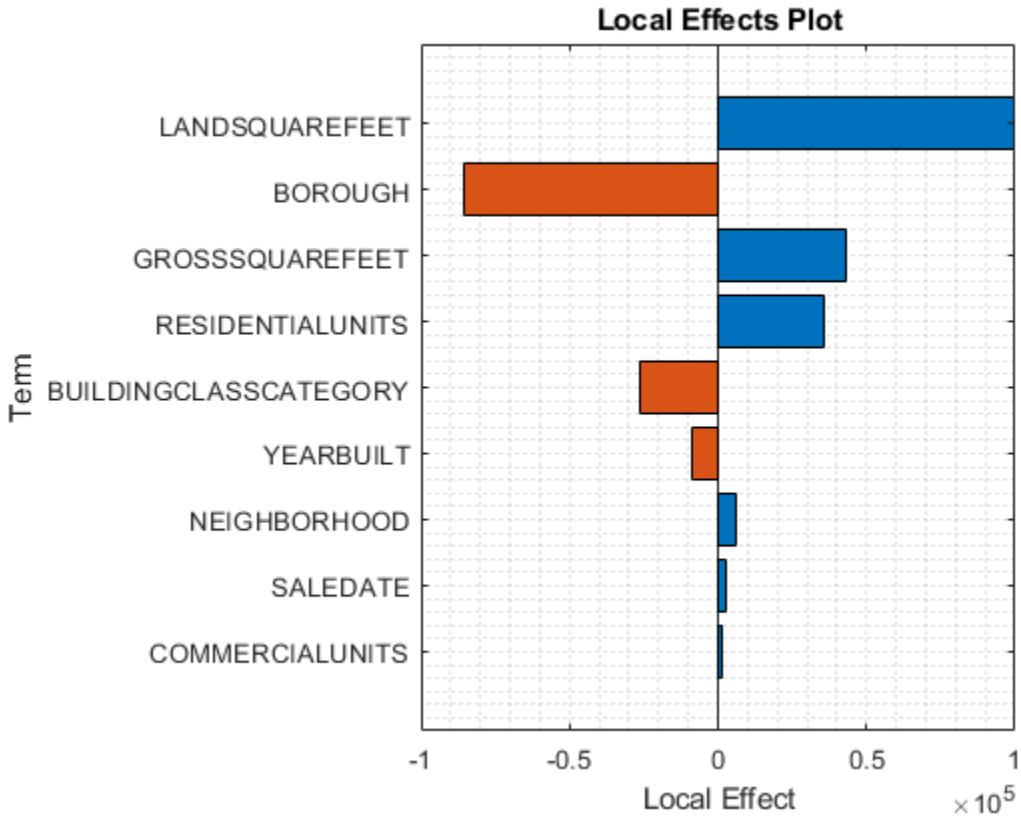
```
ans = 3.7518e+05
```

Predict the sale price for the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction.

```
yFit = predict(Mdl,NYCHousing2015(1,:))
```

```
yFit = 4.4421e+05
```

```
plotLocalEffects(Mdl,NYCHousing2015(1,:))
```



The `predict` function predicts the sale price for the first observation as `4.4421e5`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the terms in `Mdl` on the prediction. Each local effect value shows the contribution of each term to the predicted sale price.

Train GAM with Interaction Terms

Train a generalized additive model that contains linear and interaction terms for predictors in three different ways:

- Specify the interaction terms using the `formula` input argument.
- Specify the `'Interactions'` name-value argument.
- Build a model with linear terms first and add interaction terms to the model by using the `addInteractions` function.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table that contains the predictor variables (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and the response variable (`MPG`).

```
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Specify formula

Train a GAM that contains the four linear terms (Acceleration, Displacement, Horsepower, and Weight) and two interaction terms (Acceleration*Displacement and Displacement*Horsepower). Specify the terms using a formula in the form 'Y ~ terms'.

```
Mdl1 = fitrgam(tbl,'MPG ~ Acceleration + Displacement + Horsepower + Weight + Acceleration:Displa
```

The function adds interaction terms to the model in the order of importance. You can use the Interactions property to check the interaction terms in the model and the order in which fitrgam adds them to the model. Display the Interactions property.

```
Mdl1.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

Each row of Interactions represents one interaction term and contains the column indexes of the predictor variables for the interaction term.

Specify 'Interactions'

Pass the training data (tbl) and the name of the response variable in tbl to fitrgam, so that the function includes the linear terms for all the other variables as predictors. Specify the 'Interactions' name-value argument using a logical matrix to include the two interaction terms, x1*x2 and x2*x3.

```
Mdl2 = fitrgam(tbl,'MPG','Interactions',logical([1 1 0 0; 0 1 1 0]));
Mdl2.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

You can also specify 'Interactions' as the number of interaction terms or as 'all' to include all available interaction terms. Among the specified interaction terms, fitrgam identifies those whose *p*-values are not greater than the 'MaxPValue' value and adds them to the model. The default 'MaxPValue' is 1 so that the function adds all specified interaction terms to the model.

Specify 'Interactions','all' and set the 'MaxPValue' name-value argument to 0.05.

```
Mdl3 = fitrgam(tbl,'MPG','Interactions','all','MaxPValue',0.05);
```

```
Warning: Model does not include interaction terms because all interaction terms have p-values gre
```

```
Mdl3.Interactions
```

```
ans =

  0x2 empty double matrix
```

Mdl3 includes no interaction terms, which implies one of the following: all interaction terms have *p*-values greater than 0.05, or adding the interaction terms does not improve the model fit.

Use `addInteractions` Function

Train a univariate GAM that contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function. Specify the second input argument of `addInteractions` in the same way you specify the `'Interactions'` name-value argument of `fitrgam`. You can specify the list of interaction terms using a logical matrix, the number of interaction terms, or `'all'`.

Specify the number of interaction terms as 3 to add the three most important interaction terms to the trained model.

```
Mdl4 = fitrgam(tbl,'MPG');
UpdatedMdl4 = addInteractions(Mdl4,3);
UpdatedMdl4.Interactions
```

```
ans = 3×2

     2     3
     1     2
     3     4
```

`Mdl4` is a univariate GAM, and `UpdatedMdl4` is an updated GAM that contains all the terms in `Mdl4` and three additional interaction terms.

Create Cross-Validated GAM Using `fitrgam`

Train a cross-validated GAM with 10 folds, which is the default cross-validation option, by using `fitrgam`. Then, use `kfoldPredict` to predict responses for validation-fold observations using a model trained on training-fold observations.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table that contains the predictor variables (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and the response variable (`MPG`).

```
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`.

```
rng('default') % For reproducibility
CVMdl = fitrgam(tbl,'MPG','CrossVal','on')
```

```
CVMdl =
  RegressionPartitionedGAM
    CrossValidatedModel: 'GAM'
         PredictorNames: {1x4 cell}
           ResponseName: 'MPG'
        NumObservations: 398
                  KFold: 10
              Partition: [1x1 cvpartition]
      NumTrainedPerFold: [1x1 struct]
      ResponseTransform: 'none'
```

```
  Properties, Methods
```

The `fitrgam` function creates a `RegressionPartitionedGAM` model object `CVMdl` with 10 folds. During cross-validation, the software completes these steps:

1. Randomly partition the data into 10 sets.
2. For each set, reserve the set as validation data, and train the model using the other 9 sets.
3. Store the 10 compact, trained models a in a 10-by-1 cell vector in the `Trained` property of the cross-validated model object `RegressionPartitionedGAM`.

You can override the default cross-validation setting by using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument.

Predict responses for the observations in `tbl` by using `kfoldPredict`. The function predicts responses for every observation using the model trained without that observation.

```
yHat = kfoldPredict(CVMdl);
```

`yHat` is a numeric vector. Display the first five predicted responses.

```
yHat(1:5)
```

```
ans = 5×1

   19.4848
   15.7203
   15.5742
   15.3185
   17.8223
```

Compute the regression loss (mean squared error).

```
L = kfoldLoss(CVMdl)
```

```
L = 17.7248
```

`kfoldLoss` returns the average mean squared error over 10 folds.

Optimize Cross-Validated GAM Using bayesopt

Optimize the parameters of a GAM with respect to cross-validation by using the `bayesopt` function.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Prepare `optimizableVariable` objects for the name-value arguments that you want to optimize using Bayesian optimization. This example finds optimal values for the `MaxNumSplitsPerPredictor` and `NumTreesPerPredictor` arguments of `fitrgam`.

```
maxNumSplits = optimizableVariable('maxNumSplits',[1,10],'Type','integer');
numTrees = optimizableVariable('numTrees',[1,500],'Type','integer');
```

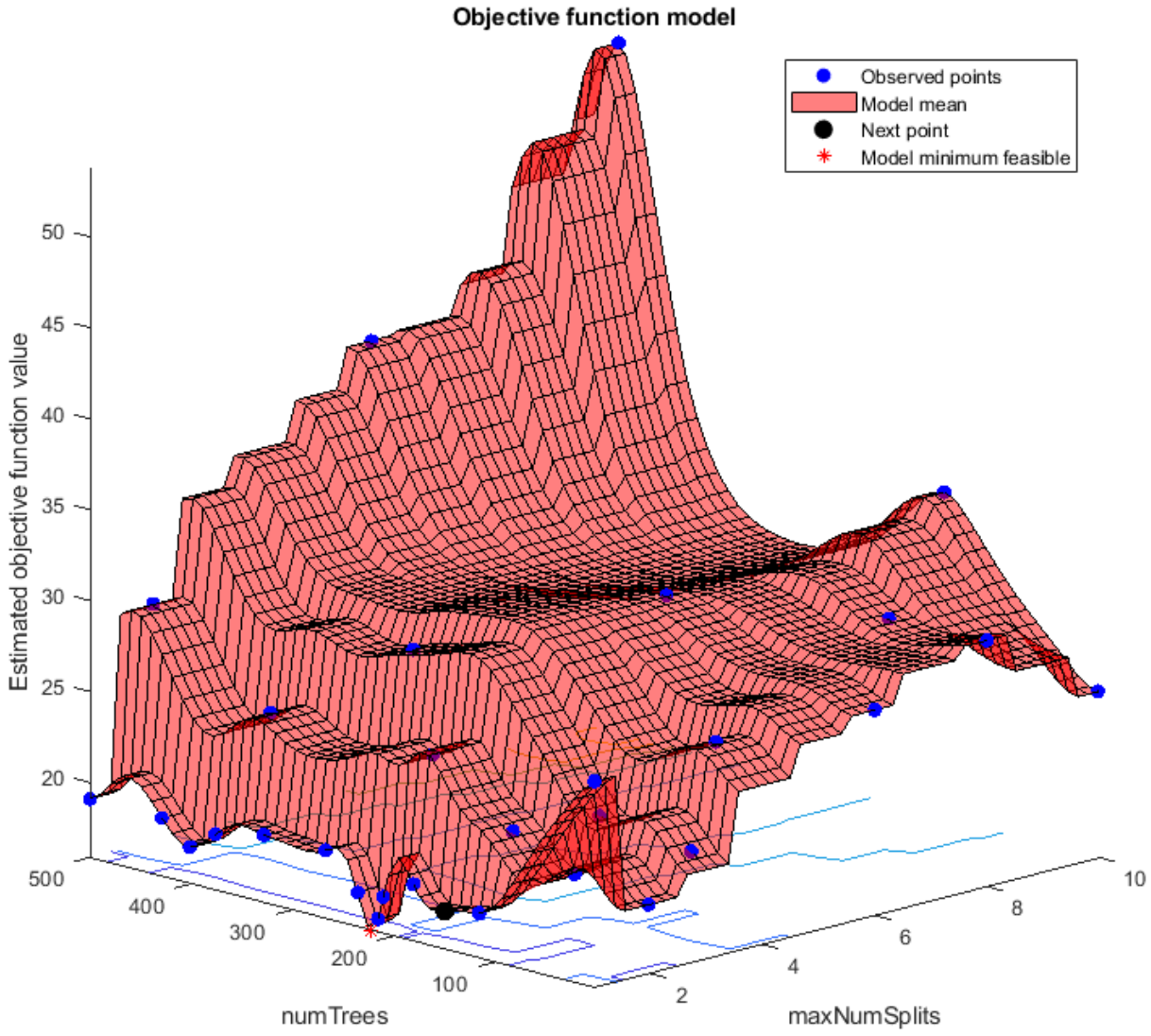
Create an objective function that takes an input `z = [maxNumSplits,numTrees]` and returns the cross-validated loss value of `z`.

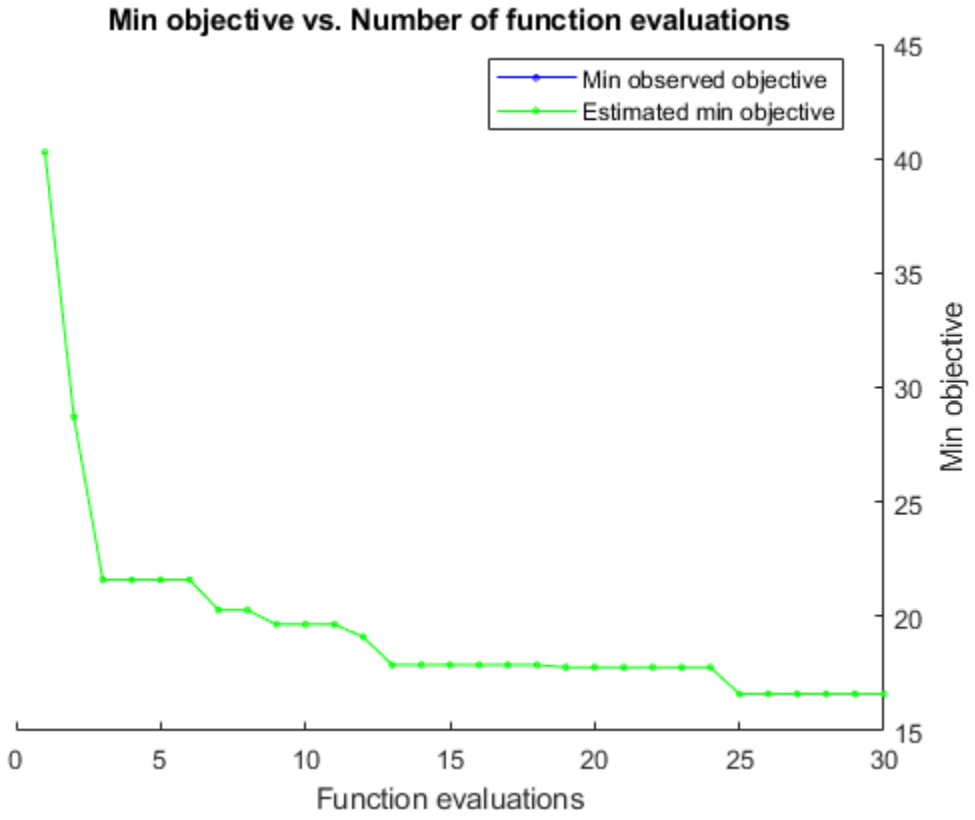
```
minfun = @(z)kfoldLoss(fitrgam(X,Y,'CrossVal','on', ...
    'MaxNumSplitsPerPredictor',z.maxNumSplits, ...
    'NumTreesPerPredictor',z.numTrees));
```

If you specify the cross-validation option (`'CrossVal','on'`), then the `fitrgam` function returns a cross-validated model object `RegressionPartitionedGAM`. The `kfoldLoss` function returns the regression loss (mean squared error) obtained by the cross-validated model. Therefore, the function handle `minfun` computes the cross-validation loss at the parameters in `z`.

Search for the best parameters `[maxNumSplits,numTrees]` using `bayesopt`. For reproducibility, choose the `'expected-improvement-plus'` acquisition function. The default acquisition function depends on run time and, therefore, can give varying results.

```
rng('default')
results = bayesopt(minfun,[maxNumSplits,numTrees],'Verbose',0, ...
    'IsObjectiveDeterministic',true, ...
    'AcquisitionFunctionName','expected-improvement-plus');
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
50
45
40
35
30
25
20
500
400
300
200
100
numTrees
2
4
6
8
10
maxNumSplits



Obtain the best point from `results`.

```
zbest = bestPoint(results)
```

```
zbest=1×2 table
    maxNumSplits    numTrees
    ____________    ________

          1             215
```

Train an optimized GAM using the `zbest` values.

```
Mdl = fitrgam(X,Y, ...
    'MaxNumSplitsPerPredictor',zbest.maxNumSplits, ...
    'NumTreesPerPredictor',zbest.numTrees);
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights. The response variable and the weight values must be numeric vectors.

  You must specify the response variable in `Tbl` by using `ResponseVarName` or `formula` and specify the observation weights in `Tbl` by using `'Weights'`.

  - Specify the response variable by using `ResponseVarName` — `fitrgam` uses the remaining variables as predictors. To use a subset of the remaining variables in `Tbl` as predictors, specify predictor variables by using `'PredictorNames'`.
  - Define a model specification by using `formula` — `fitrgam` uses a subset of the variables in `Tbl` as predictor variables and the response variable, as specified in `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable `Y` and the number of rows in `Tbl` must be equal. To use a subset of the variables in `Tbl` as predictors, specify predictor variables by using `'PredictorNames'`.

`fitrgam` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Tbl` to be missing values.

- `fitrgam` does not use observations with all missing values in the fit.
- `fitrgam` does not use observations with missing response values in the fit.
- `fitrgam` uses observations with some missing values for predictors to find splits on variables for which these observations have valid values.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`formula` — Model specification
character vector | string scalar

Model specification, specified as a character vector or string scalar in the form `'Y ~ terms'`. The `formula` argument specifies a response variable and linear and interaction terms for predictor variables. Use `formula` to specify a subset of variables in `Tbl` as predictors for training the model. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

For example, specify `'Y~x1+x2+x3+x1:x2'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the linear terms for the predictor variables. `x1:x2` represents the interaction term for `x1` and `x2`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Alternatively, you can specify a response variable and linear terms for predictors using `formula`, and specify interaction terms for predictors using `'Interactions'`.

`fitrgam` builds a set of interaction trees using only the terms whose *p*-values are not greater than the `'MaxPValue'` value.

Example: `'Y~x1+x2+x3+x1:x2'`

Data Types: `char` | `string`

Y — Response data

numeric column vector

Response data, specified as a numeric column vector. Each entry in `Y` is the response to the data in the corresponding row of `X` or `Tbl`.

The software considers `NaN` values in `Y` to be missing values. `fitrgam` does not use observations with missing response values in the fit.

Data Types: `single` | `double`

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

`fitrgam` considers `NaN` values in `X` as missing values. The function does not use observations with all missing values in the fit. `fitrgam` uses observations with some missing values for `X` to find splits on variables for which these observations have valid values.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Interactions','all','MaxPValue',0.05` specifies to include all available interaction terms whose *p*-values are not greater than 0.05.

GAM Options

InitialLearnRateForInteractions — Initial learning rate of gradient boosting for interaction terms

1 (default) | numeric scalar in (0,1]

Initial learning rate of gradient boosting for interaction terms, specified as a numeric scalar in the interval (0,1].

For each boosting iteration for interaction trees, `fitrgam` starts fitting with the initial learning rate. The function halves the learning rate until it finds a rate that improves the model fit.

Training a model using a small learning rate requires more learning iterations, but often achieves better accuracy.

For more details about gradient boosting, see “Gradient Boosting Algorithm” on page 33-2134.

Example: `'InitialLearnRateForInteractions',0.1`

Data Types: `single` | `double`

`InitialLearnRateForPredictors` — Initial learning rate of gradient boosting for linear terms
`1` (default) | numeric scalar in (0,1]

Initial learning rate of gradient boosting for linear terms, specified as a numeric scalar in the interval (0,1].

For each boosting iteration for predictor trees, `fitrgam` starts fitting with the initial learning rate. The function halves the learning rate until it finds a rate that improves the model fit.

Training a model using a small learning rate requires more learning iterations, but often achieves better accuracy.

For more details about gradient boosting, see “Gradient Boosting Algorithm” on page 33-2134.

Example: `'InitialLearnRateForPredictors',0.1`

Data Types: `single` | `double`

`Interactions` — Number or list of interaction terms
`0` (default) | nonnegative integer scalar | logical matrix | `'all'`

Number or list of interaction terms to include in the candidate set *S*, specified as a nonnegative integer scalar, a logical matrix, or `'all'`.

- Number of interaction terms, specified as a nonnegative integer — *S* includes the specified number of important interaction terms, selected based on the *p*-values of the terms.
- List of interaction terms, specified as a logical matrix — *S* includes the terms specified by a `t`-by-`p` logical matrix, where `t` is the number of interaction terms, and `p` is the number of predictors used to train the model. For example, `logical([1 1 0; 0 1 1])` represents two pairs of interaction terms: a pair of the first and second predictors, and a pair of the second and third predictors.

  If `fitrgam` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. That is, the column indexes of the logical matrix do not count the response and observation weight variables. The indexes also do not count any variables not used by the function.
- `'all'` — *S* includes all possible pairs of interaction terms, which is `p*(p − 1)/2` number of terms in total.

Among the interaction terms in *S*, the `fitrgam` function identifies those whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees. Use the default value (`'MaxPValue'`,1) to build interaction trees using all terms in *S*.

Example: `'Interactions','all'`

Data Types: `single` | `double` | `logical` | `char` | `string`

`MaxNumSplitsPerInteraction` — Maximum number of decision splits per interaction tree
4 (default) | positive integer scalar

Maximum number of decision splits (or branch nodes) for each interaction tree (boosted tree for an interaction term), specified as a positive integer scalar.

Example: `'MaxNumSplitsPerInteraction',5`

Data Types: `single` | `double`

`MaxNumSplitsPerPredictor` — Maximum number of decision splits per predictor tree
1 (default) | positive integer scalar

Maximum number of decision splits (or branch nodes) for each predictor tree (boosted tree for a linear term), specified as a positive integer scalar. By default, `fitrgam` uses a tree stump for a predictor tree.

Example: `'MaxNumSplitsPerPredictor',5`

Data Types: `single` | `double`

`MaxPValue` — Maximum *p*-value for detecting interaction terms
1 (default) | numeric scalar in [0,1]

Maximum *p*-value for detecting interaction terms, specified as a numeric scalar in the interval [0,1].

`fitrgam` first finds the candidate set *S* of interaction terms from `formula` or `'Interactions'`. Then the function identifies the interaction terms whose *p*-values are not greater than the `'MaxPValue'` value and uses them to build a set of interaction trees.

The default value (`'MaxPValue',1`) builds interaction trees for all interaction terms in the candidate set *S*.

For more details about detecting interaction terms, see “Interaction Term Detection” on page 33-2135.

Example: `'MaxPValue',0.05`

Data Types: `single` | `double`

`NumBins` — Number of bins for numeric predictors
`256` (default) | positive integer scalar | `[]` (empty)

Number of bins for numeric predictors, specified as a positive integer scalar or `[]` (empty).

- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitrgam` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
  - `fitrgam` does not bin categorical predictors.
- If the `'NumBins'` value is empty (`[]`), then `fitrgam` does not bin any predictors.

When you use a large training data set, this binning option speeds up training but might cause a decrease in accuracy. You can first use the default value of `'NumBins'`, and then change the value depending on the accuracy and training speed.

The trained model `Mdl` stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`NumTreesPerInteraction` — Number of trees per interaction term
100 (default) | positive integer scalar

Number of trees per interaction term, specified as a positive integer scalar.

The `'NumTreesPerInteraction'` value is equivalent to the number of gradient boosting iterations for the interaction terms for predictors. For each iteration, `fitrgam` adds a set of interaction trees to the model, one tree for each interaction term. To learn about the gradient boosting algorithm, see “Gradient Boosting Algorithm” on page 33-2134.

You can determine whether the fitted model has the specified number of trees by viewing the diagnostic message displayed when `'Verbose'` is 1 or 2, or by checking the `ReasonForTermination` property value of the model `Mdl`.

Example: `'NumTreesPerInteraction',500`

Data Types: `single` | `double`

`NumTreesPerPredictor` — Number of trees per linear term
300 (default) | positive integer scalar

Number of trees per linear term, specified as a positive integer scalar.

The `'NumTreesPerPredictor'` value is equivalent to the number of gradient boosting iterations for the linear terms for predictors. For each iteration, `fitrgam` adds a set of predictor trees to the model, one tree for each predictor. To learn about the gradient boosting algorithm, see “Gradient Boosting Algorithm” on page 33-2134.

You can determine whether the fitted model has the specified number of trees by viewing the diagnostic message displayed when `'Verbose'` is 1 or 2, or by checking the `ReasonForTermination` property value of the model `Mdl`.

Example: `'NumTreesPerPredictor',500`

Data Types: `single` | `double`

Other Regression Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrgam` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.

Value	Description
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrgam` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrgam` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`NumPrint` — Number of iterations between diagnostic message printouts
`10` (default) | nonnegative integer scalar

Number of iterations between diagnostic message printouts, specified as a nonnegative integer scalar. This argument is valid only when you specify `'Verbose'` as 1.

If you specify `'Verbose',1` and `'NumPrint',numPrint`, then the software displays diagnostic messages every `numPrint` iterations in the Command Window.

Example: `'NumPrint',500`

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrgam` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.

- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

`Verbose` — Verbosity level
`0` (default) | `1` | `2`

Verbosity level, specified as `0`, `1`, or `2`. The `Verbose` value controls the amount of information that the software displays in the Command Window.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software displays no information.
`1`	The software displays diagnostic messages every `numPrint` iterations, where `numPrint` is the `'NumPrint'` value.
`2`	The software displays diagnostic messages at every iteration.

Each line of the diagnostic messages shows the information about each boosting iteration and includes the following columns:

- `Type` — Type of trained trees, `1D` (predictor trees, or boosted trees for linear terms for predictors) or `2D` (interaction trees, or boosted trees for interaction terms for predictors)
- `NumTrees` — Number of trees per linear term or interaction term that `fitrgam` added to the model so far

- `Deviance` — "Deviance" on page 33-2134 of the model
- `RelTol` — Relative change of model predictions: $(\widehat{y}_k - \widehat{y}_{k-1})'(\widehat{y}_k - \widehat{y}_{k-1})/\widehat{y}_k'\widehat{y}_k$, where $\widehat{y}_k$ is a column vector of model predictions at iteration k
- `LearnRate` — Learning rate used for the current iteration

Example: `'Verbose',1`

Data Types: `single` | `double`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`.

`fitrgam` normalizes the values of `Weights` to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Flag to train cross-validated model

`'off'` (default) | `'on'`

Flag to train a cross-validated model, specified as `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated model with 10 folds.

You can override this cross-validation setting using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument. You can use only one cross-validation name-value argument at a time to create a cross-validated model.

Alternatively, cross-validate after creating a model by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Output Arguments

`Mdl` — Trained generalized additive model
`RegressionGAM` model object | `RegressionPartitionedGAM` cross-validated model object

Trained generalized additive model, returned as one of the model objects in this table.

Model Object	Cross-Validation Options to Train Model Object	Ways to Predict Responses Using Model Object
`RegressionGAM`	None	Use `predict` to predict responses for new observations, and use `resubPredict` to predict responses for training observations.
`RegressionPartitionedGAM`	Specify the name-value argument `KFold`, `Holdout`, `Leaveout`, `CrossVal`, or `CVPartition`	Use `kfoldPredict` to predict responses for observations that `fitrgam` holds out during training. `kfoldPredict` predicts a response for every observation by using the model trained without that observation.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.Interactions` in the Command Window to display the interaction terms in `Mdl`.

More About

Generalized Additive Model (GAM) for Regression

A generalized additive model (GAM) is an interpretable model that explains a response variable using a sum of univariate and bivariate shape functions of predictors.

`fitrgam` uses a boosted tree as a shape function for each predictor and, optionally, each pair of predictors; therefore, the function can capture a nonlinear relation between a predictor and the response variable. Because contributions of individual shape functions to the prediction (response value) are well separated, the model is easy to interpret.

The standard GAM uses a univariate shape function for each predictor.

$$y \sim N\left(\mu, \sigma^2\right)$$
$$g(\mu) = \mu = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where y is a response variable that follows the normal distribution with mean μ and standard deviation σ. $g(\mu)$ is an identity link function, and c is an intercept (constant) term. $f_i(x_i)$ is a univariate shape function for the ith predictor, which is a boosted tree for a linear term for the predictor (predictor tree).

You can include interactions between predictors in a model by adding bivariate shape functions of important interaction terms to the model.

$$\mu = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \sum_{i,\, j \in \{1, 2, \cdots, p\}} f_{ij}(x_i x_j),$$

where $f_{ij}(x_i x_j)$ is a bivariate shape function for the ith and jth predictors, which is a boosted tree for an interaction term for the predictors (interaction tree).

`fitrgam` finds important interaction terms based on the *p*-values of *F*-tests. For details, see “Interaction Term Detection” on page 33-2135.

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to the saturated model.

The deviance of a fitted model is twice the difference between the loglikelihoods of the model and the saturated model:

$$-2(\log L - \log L_s),$$

where L and L_s are the likelihoods of the fitted model and the saturated model, respectively. The saturated model is the model with the maximum number of parameters that you can estimate.

`fitrgam` uses the deviance to measure the goodness of model fit and finds a learning rate that reduces the deviance at each iteration. Specify `'Verbose'` as 1 or 2 to display the deviance and learning rate in the Command Window.

Algorithms

Gradient Boosting Algorithm

`fitrgam` fits a generalized additive model using a gradient boosting algorithm (“Least-Squares Boosting” on page 18-50).

`fitrgam` first builds sets of predictor trees (boosted trees for linear terms for predictors) and then builds sets of interaction trees (boosted trees for interaction terms for predictors). The boosting algorithm iterates for at most `'NumTreesPerPredictor'` times for predictor trees, and then iterates for at most `'NumTreesPerInteraction'` times for interaction trees.

For each boosting iteration, `fitrgam` builds a set of predictor trees with the initial learning rate `'InitialLearnRateForPredictors'`, or builds a set of interaction trees with the initial learning rate `'InitialLearnRateForInteractions'`.

- When building a set of trees, the function trains one tree at a time. It fits a tree to the residual that is the difference between the response and the aggregated prediction from all trees grown previously. To control the boosting learning speed, the function shrinks the tree by the learning rate and then adds the tree to the model and updates the residual.
  - Updated model = current model + (learning rate)·(new tree)
  - Updated residual = current residual – (learning rate)·(response explained by new tree)
- If adding the set of trees improves the model fit (that is, reduces the deviance of the fit), then `fitrgam` moves to the next iteration.
- Otherwise, `fitrgam` halves the learning rate and uses it to update the model and residual. The function continues to halve the learning rate until it finds a rate that improves the model fit.
  - If the function cannot find such a learning rate for predictor trees, then it stops boosting iterations for linear terms and starts boosting iterations for interaction terms.
  - If the function cannot find such a learning rate for interaction trees, then it terminates the model fitting.

  You can determine why training stopped by checking the `ReasonForTermination` property of the trained model.

Interaction Term Detection

For each pairwise interaction term $x_i x_j$ (specified by `formula` or `'Interactions'`), the software performs an *F*-test to examine whether the term is statistically significant.

To speed up the process, `fitrgam` bins numeric predictors into at most 8 equiprobable bins. The number of bins can be less than 8 if a predictor has fewer than 8 unique values. The *F*-test examines the null hypothesis that the bins created by x_i and x_j have equal responses versus the alternative that at least one bin has a different response value from the others. A small *p*-value indicates that differences are significant, which implies that the corresponding interaction term is significant and, therefore, including the term can improve the model fit.

`fitrgam` builds a set of interaction trees using the terms whose *p*-values are not greater than the `'MaxPValue'` value. You can use the default `'MaxPValue'` value `1` to build interaction trees using all terms specified by `formula` or `'Interactions'`.

`fitrgam` adds interaction terms to the model in the order of importance based on the *p*-values. Use the `Interactions` property of the returned model to check the order of the interaction terms added to the model.

References

[1] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible Models for Classification and Regression." *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12).* Beijing, China: ACM Press, 2012, pp. 150–158.

[2] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate Intelligible Models with Pairwise Interactions." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)* Chicago, Illinois, USA: ACM Press, 2013, pp. 623–631.

See Also

`RegressionGAM` | `RegressionPartitionedGAM` | `addInteractions` | `predict` | `resume`

Topics

"Train Generalized Additive Model for Regression" on page 12-91

Introduced in R2021a

fitrgp

Fit a Gaussian process regression (GPR) model

Syntax

```
gprMdl = fitrgp(Tbl,ResponseVarName)
gprMdl = fitrgp(Tbl,formula)
gprMdl = fitrgp(Tbl,y)

gprMdl = fitrgp(X,y)

gprMdl = fitrgp(___,Name,Value)
```

Description

`gprMdl = fitrgp(Tbl,ResponseVarName)` returns a Gaussian process regression (GPR) model trained using the sample data in `Tbl`, where `ResponseVarName` is the name of the response variable in `Tbl`.

`gprMdl = fitrgp(Tbl,formula)` returns a Gaussian process regression (GPR) model, trained using the sample data in `Tbl`, for the predictor variables and response variables identified by `formula`.

`gprMdl = fitrgp(Tbl,y)` returns a GPR model for the predictors in table `Tbl` and continuous response vector `y`.

`gprMdl = fitrgp(X,y)` returns a GPR model for predictors `X` and continuous response vector `y`.

`gprMdl = fitrgp(___,Name,Value)` returns a GPR model for any of the input arguments in the previous syntaxes, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the fitting method, the prediction method, the covariance function, or the active set selection method. You can also train a cross-validated model.

`gprMdl` is a `RegressionGP` object. For methods and properties of this class, see `RegressionGP` class page.

If you train a cross-validated model, then `gprMdl` is a `RegressionPartitionedModel` object. For further analysis on the cross-validated object, use the methods of `RegressionPartitionedModel` class. For the methods of this class, see the `RegressionPartitionedModel` class page.

Examples

Train GPR Model Using Data in Table

This example uses the abalone data [1], [2], from the UCI Machine Learning Repository [3] . Download the data and save it in your current folder with the name `abalone.data`.

Store the data into a table. Display the first seven rows.

```
tbl = readtable('abalone.data','Filetype','text',...
     'ReadVariableNames',false);
tbl.Properties.VariableNames = {'Sex','Length','Diameter','Height',...
     'WWeight','SWeight','VWeight','ShWeight','NoShellRings'};
tbl(1:7,:)
```

```
ans =

    Sex    Length    Diameter    Height    WWeight    SWeight    VWeight    ShWeight    NoShellRi
    ___    ______    ________    ______    _______    _______    _______    ________    _________

    'M'    0.455     0.365       0.095      0.514     0.2245      0.101      0.15       15
    'M'     0.35     0.265        0.09     0.2255     0.0995     0.0485      0.07        7
    'F'     0.53      0.42       0.135      0.677     0.2565     0.1415      0.21        9
    'M'     0.44     0.365       0.125      0.516     0.2155      0.114     0.155       10
    'I'     0.33     0.255        0.08      0.205     0.0895     0.0395     0.055        7
    'I'    0.425       0.3       0.095     0.3515      0.141     0.0775      0.12        8
    'F'     0.53     0.415        0.15     0.7775      0.237     0.1415      0.33       20
```

The dataset has 4177 observations. The goal is to predict the age of abalone from eight physical measurements. The last variable, number of shell rings shows the age of the abalone. The first predictor is a categorical variable. The last variable in the table is the response variable.

Fit a GPR model using the subset of regressors method for parameter estimation and fully independent conditional method for prediction. Standardize the predictors.

```
gprMdl = fitrgp(tbl,'NoShellRings','KernelFunction','ardsquaredexponential',...
      'FitMethod','sr','PredictMethod','fic','Standardize',1)
```

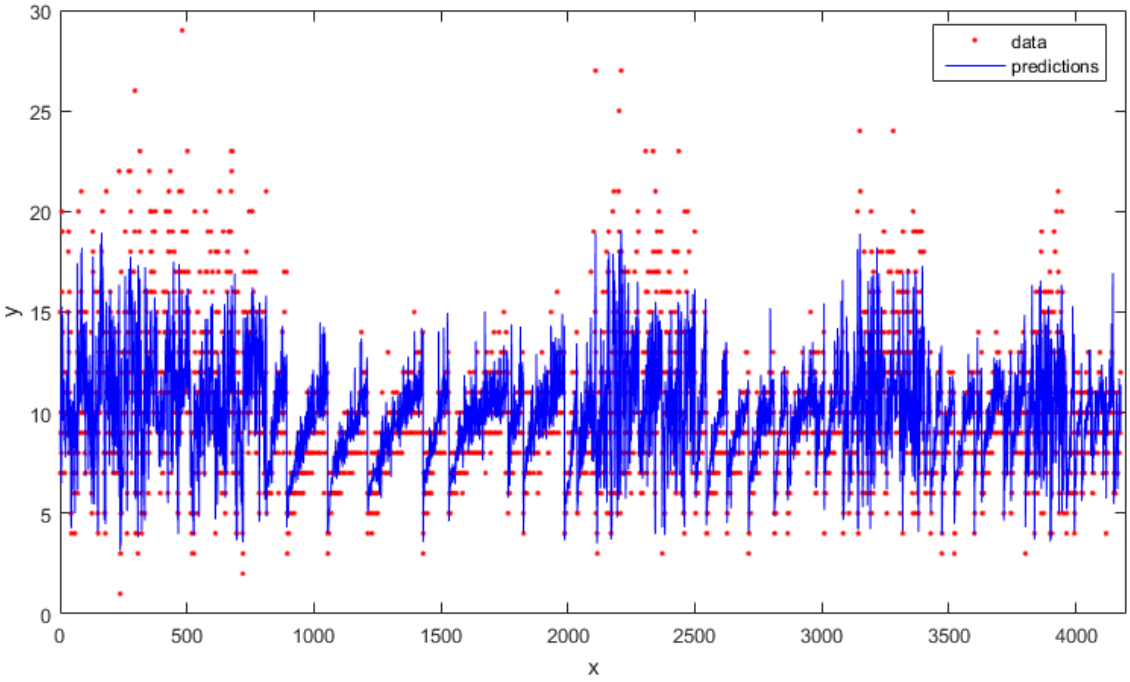
```
grMdl =

  RegressionGP
       PredictorNames: {1x8 cell}
         ResponseName: 'Var9'
    ResponseTransform: 'none'
      NumObservations: 4177
       KernelFunction: 'ARDSquaredExponential'
    KernelInformation: [1x1 struct]
        BasisFunction: 'Constant'
                 Beta: 10.9148
                Sigma: 2.0243
    PredictorLocation: [10x1 double]
       PredictorScale: [10x1 double]
                Alpha: [1000x1 double]
     ActiveSetVectors: [1000x10 double]
        PredictMethod: 'FIC'
        ActiveSetSize: 1000
            FitMethod: 'SR'
      ActiveSetMethod: 'Random'
    IsActiveSetVector: [4177x1 logical]
        LogLikelihood: -9.0013e+03
     ActiveSetHistory: [1x1 struct]
       BCDInformation: []
```

Predict the responses using the trained model.

```
ypred = resubPredict(gprMdl);
```

Plot the true response and the predicted responses.

```
figure();
plot(tbl.NoShellRings,'r.');
hold on
plot(ypred,'b');
xlabel('x');
ylabel('y');
legend({'data','predictions'},'Location','Best');
axis([0 4300 0 30]);
hold off;
```



Compute the regression loss on the training data (resubstitution loss) for the trained model.

```
L = resubLoss(gprMdl)
```

```
L =

    4.0064
```

Train GPR Model and Plot Predictions

Generate sample data.

```
rng(0,'twister'); % For reproducibility
n = 1000;
x = linspace(-10,10,n)';
y = 1 + x*5e-2 + sin(x)./x + 0.2*randn(n,1);
```

Fit a GPR model using a linear basis function and the exact fitting method to estimate the parameters. Also use the exact prediction method.

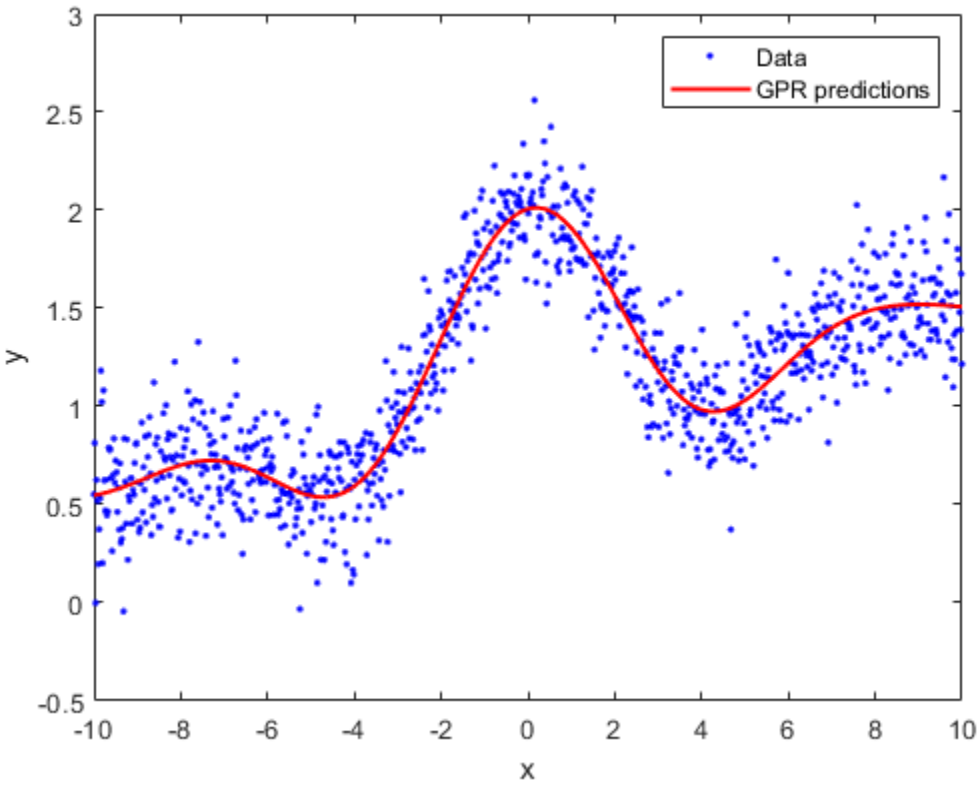
```
gprMdl = fitrgp(x,y,'Basis','linear',...
      'FitMethod','exact','PredictMethod','exact');
```

Predict the response corresponding to the rows of x (resubstitution predictions) using the trained model.

```
ypred = resubPredict(gprMdl);
```

Plot the true response with the predicted values.

```
plot(x,y,'b.');
hold on;
plot(x,ypred,'r','LineWidth',1.5);
xlabel('x');
ylabel('y');
legend('Data','GPR predictions');
hold off
```



Impact of Specifying Initial Kernel Parameter Values

Load the sample data.

```
load('gprdata2.mat')
```

The data has one predictor variable and continuous response. This is simulated data.

Fit a GPR model using the squared exponential kernel function with default kernel parameters.

```
gprMdl1 = fitrgp(x,y,'KernelFunction','squaredexponential');
```

Now, fit a second model, where you specify the initial values for the kernel parameters.

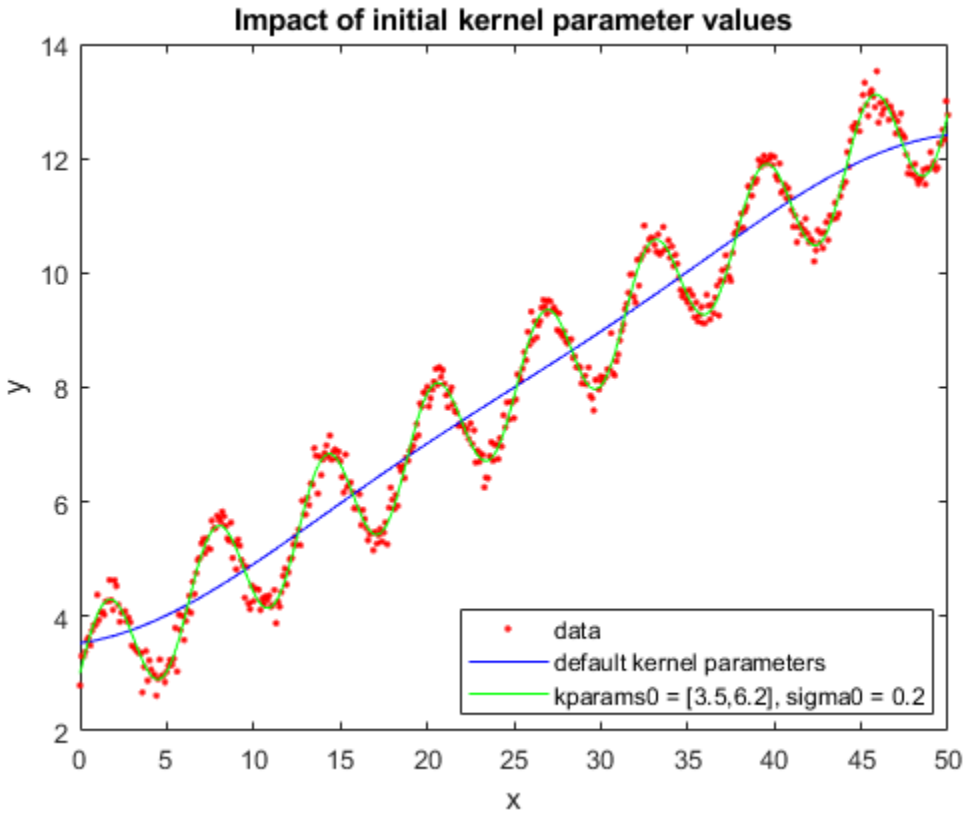
```
sigma0 = 0.2;
kparams0 = [3.5, 6.2];
gprMdl2 = fitrgp(x,y,'KernelFunction','squaredexponential',...
     'KernelParameters',kparams0,'Sigma',sigma0);
```

Compute the resubstitution predictions from both models.

```
ypred1 = resubPredict(gprMdl1);
ypred2 = resubPredict(gprMdl2);
```

Plot the response predictions from both models and the responses in training data.

```
figure();
plot(x,y,'r.');
hold on
plot(x,ypred1,'b');
plot(x,ypred2,'g');
xlabel('x');
ylabel('y');
legend({'data','default kernel parameters',...
'kparams0 = [3.5,6.2], sigma0 = 0.2'},...
'Location','Best');
title('Impact of initial kernel parameter values');
hold off
```



The marginal log likelihood that `fitrgp` maximizes to estimate GPR parameters has multiple local solutions; the solution that it converges to depends on the initial point. Each local solution corresponds to a particular interpretation of the data. In this example, the solution with the default initial kernel parameters corresponds to a low frequency signal with high noise whereas the second solution with custom initial kernel parameters corresponds to a high frequency signal with low noise.

Use Separate Length Scales for Predictors

Load the sample data.

```
load('gprdata.mat')
```

There are six continuous predictor variables. There are 500 observations in the training data set and 100 observations in the test data set. This is simulated data.

Fit a GPR model using the squared exponential kernel function with a separate length scale for each predictor. This covariance function is defined as:

$$k(x_i,x_j|\theta)=\sigma_f^2\exp\left[-\frac{1}{2}\sum_{m=1}^{d}\frac{(x_{im}-x_{jm})^2}{\sigma_m^2}\right].$$

where σ_m represents the length scale for predictor m, $m = 1, 2, \ldots, d$ and σ_f is the signal standard deviation. The unconstrained parametrization θ is

$\theta_m = \log\sigma_m, \quad \text{for } m = 1, 2, \ldots, d$

$\theta_{d+1} = \log\sigma_f .$

Initialize length scales of the kernel function at 10 and signal and noise standard deviations at the standard deviation of the response.

```
sigma0 = std(ytrain);
sigmaF0 = sigma0;
d = size(Xtrain,2);
sigmaM0 = 10*ones(d,1);
```

Fit the GPR model using the initial kernel parameter values. Standardize the predictors in the training data. Use the exact fitting and prediction methods.

```
gprMdl = fitrgp(Xtrain,ytrain,'Basis','constant','FitMethod','exact',...
'PredictMethod','exact','KernelFunction','ardsquaredexponential',...
'KernelParameters',[sigmaM0;sigmaF0],'Sigma',sigma0,'Standardize',1);
```

Compute the regression loss on the test data.

```
L = loss(gprMdl,Xtest,ytest)
```

```
L = 0.6919
```

Access the kernel information.

```
gprMdl.KernelInformation
```

```
ans = struct with fields:
                    Name: 'ARDSquaredExponential'
        KernelParameters: [7x1 double]
    KernelParameterNames: {7x1 cell}
```

Display the kernel parameter names.

```
gprMdl.KernelInformation.KernelParameterNames
```

```
ans = 7x1 cell
    {'LengthScale1'}
    {'LengthScale2'}
    {'LengthScale3'}
    {'LengthScale4'}
    {'LengthScale5'}
    {'LengthScale6'}
    {'SigmaF'      }
```

Display the kernel parameters.

```
sigmaM = gprMdl.KernelInformation.KernelParameters(1:end-1,1)
```

```
sigmaM = 6×1
10^4 ×

    0.0004
    0.0007
    0.0004
```

```
    5.2711
    0.1018
    0.0056
```

```
sigmaF = gprMdl.KernelInformation.KernelParameters(end)
```

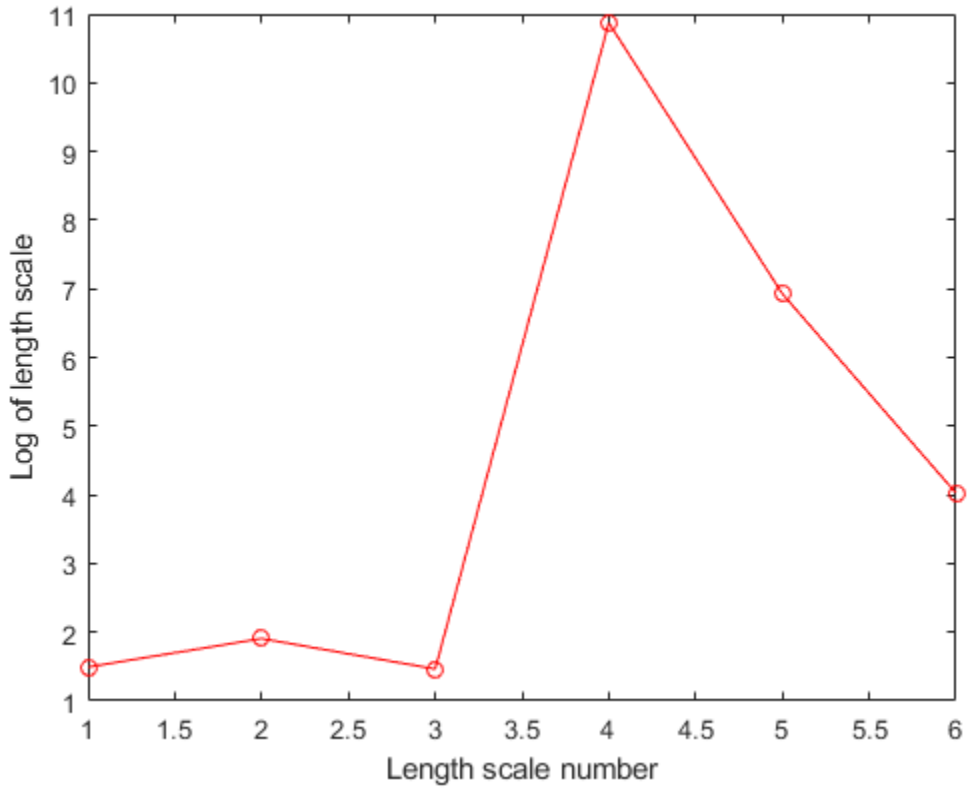
```
sigmaF = 28.1721
```

```
sigma  = gprMdl.Sigma
```

```
sigma = 0.8162
```

Plot the log of learned length scales.

```
figure()
plot((1:d)',log(sigmaM),'ro-');
xlabel('Length scale number');
ylabel('Log of length scale');
```



The log of length scale for the 4th and 5th predictor variables are high relative to the others. These predictor variables do not seem to be as influential on the response as the other predictor variables.

Fit the GPR model without using the 4th and 5th variables as the predictor variables.

```
X = [Xtrain(:,1:3) Xtrain(:,6)];
sigma0 = std(ytrain);
sigmaF0 = sigma0;
```

```
d = size(X,2);
sigmaM0 = 10*ones(d,1);

gprMdl = fitrgp(X,ytrain,'Basis','constant','FitMethod','exact',...
'PredictMethod','exact','KernelFunction','ardsquaredexponential',...
'KernelParameters',[sigmaM0;sigmaF0],'Sigma',sigma0,'Standardize',1);
```

Compute the regression error on the test data.

```
xtest = [Xtest(:,1:3) Xtest(:,6)];
L = loss(gprMdl,xtest,ytest)
```

```
L = 0.6928
```

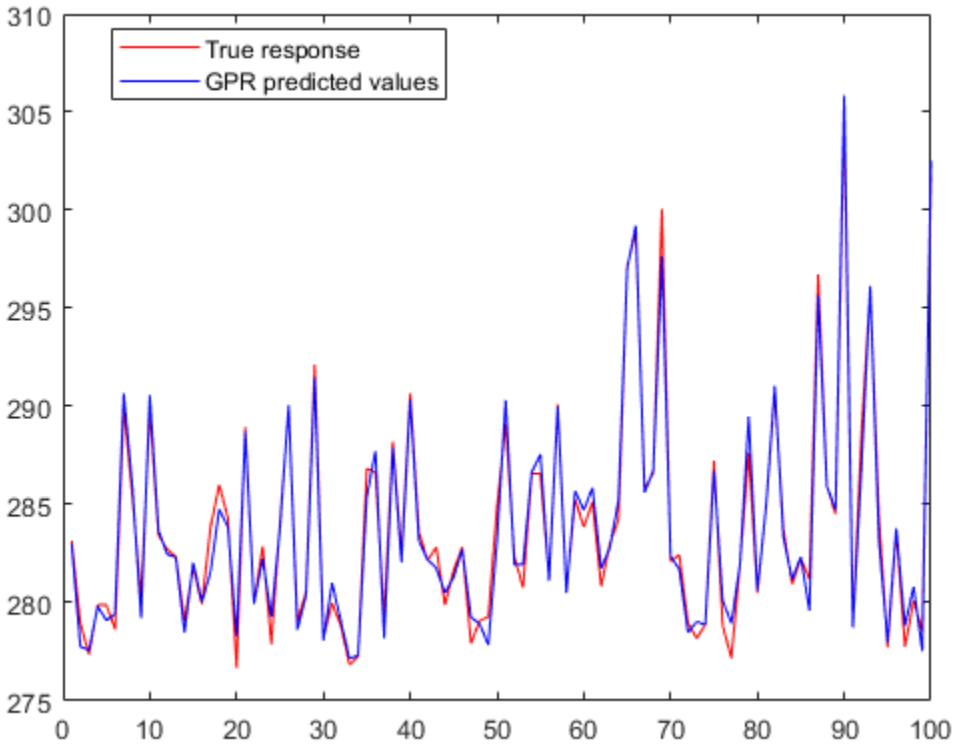
The loss is similar to the one when all variables are used as predictor variables.

Compute the predicted response for the test data.

```
 ypred = predict(gprMdl,xtest);
```

Plot the original response along with the fitted values.

```
figure;
plot(ytest,'r');
hold on;
plot(ypred,'b');
legend('True response','GPR predicted values','Location','Best');
hold off
```



Optimize GPR Regression

This example shows how to optimize hyperparameters automatically using `fitrgp`. The example uses the `gprdata2` data that ships with your software.

Load the data.

```
load('gprdata2.mat')
```

The data has one predictor variable and continuous response. This is simulated data.

Fit a GPR model using the squared exponential kernel function with default kernel parameters.

```
gprMdl1 = fitrgp(x,y,'KernelFunction','squaredexponential');
```

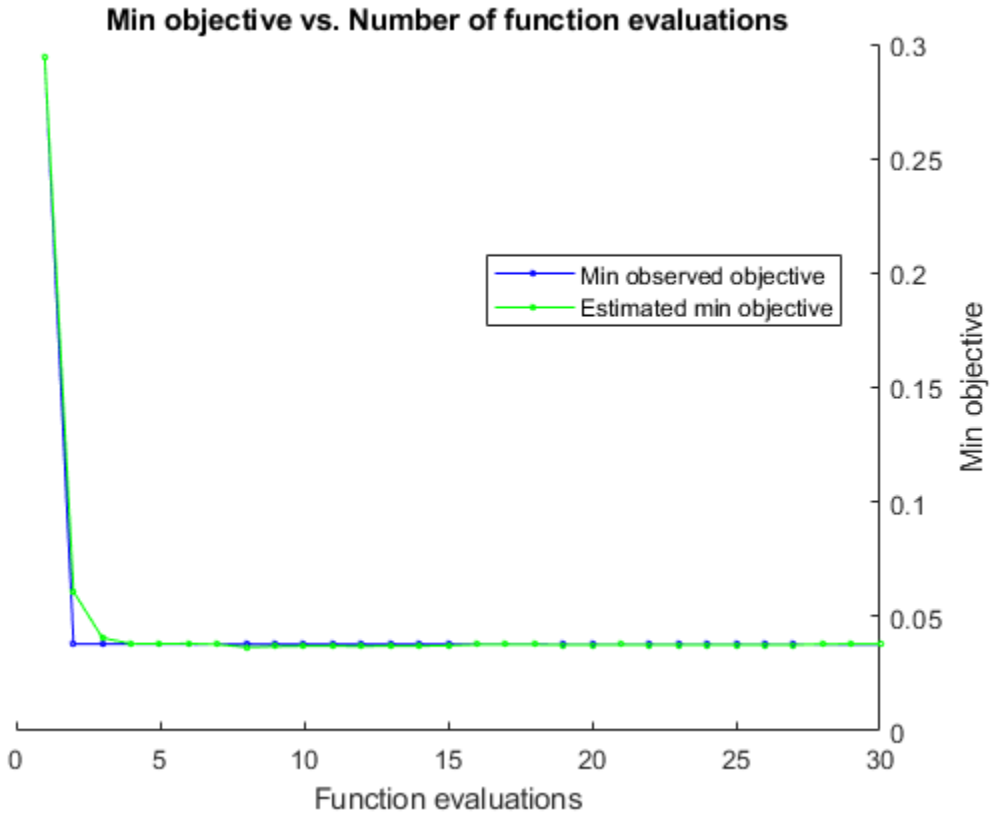
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

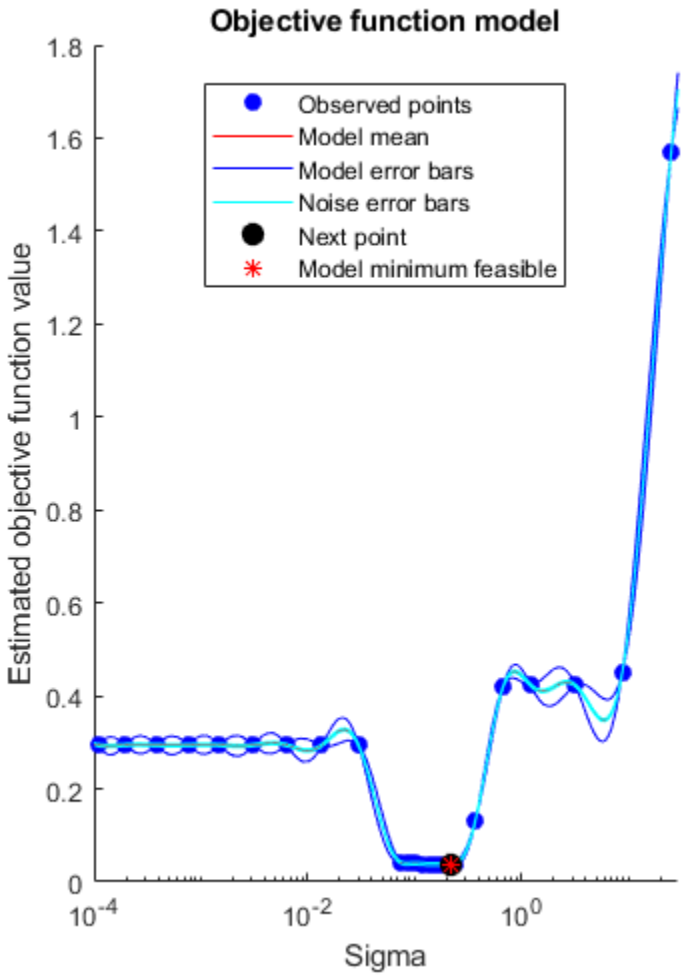
For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
gprMdl2 = fitrgp(x,y,'KernelFunction','squaredexponential',...
    'OptimizeHyperparameters','auto','HyperparameterOptimizationOptions',...
    struct('AcquisitionFunctionName','expected-improvement-plus'));
```

```
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |        Sigma |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              |
|=====================================================================================================|
|    1 | Best   |      0.29417 |       3.5408 |      0.29417 |      0.29417 |    0.0015045 |
|    2 | Best   |     0.037898 |       2.9479 |     0.037898 |     0.060792 |       0.14147 |
|    3 | Accept |       1.5693 |       1.5771 |     0.037898 |     0.040633 |       25.279 |
|    4 | Accept |      0.29417 |       2.9349 |     0.037898 |     0.037984 |    0.0001091 |
|    5 | Accept |      0.29393 |       3.7733 |     0.037898 |     0.038029 |     0.029932 |
|    6 | Accept |      0.13152 |       2.3301 |     0.037898 |     0.038127 |       0.37127 |
|    7 | Best   |     0.037785 |       3.6938 |     0.037785 |     0.037728 |       0.18116 |
|    8 | Accept |      0.03783 |       2.7556 |     0.037785 |     0.036524 |       0.16251 |
|    9 | Accept |     0.037833 |       3.6962 |     0.037785 |     0.036854 |       0.16159 |
|   10 | Accept |     0.037835 |       3.4367 |     0.037785 |     0.037052 |       0.16072 |
|   11 | Accept |      0.29417 |       3.3626 |     0.037785 |      0.03705 |   0.00038214 |
|   12 | Accept |      0.42256 |       1.9746 |     0.037785 |      0.03696 |       3.2067 |
|   13 | Accept |      0.03786 |       3.3984 |     0.037785 |     0.037087 |       0.15245 |
|   14 | Accept |      0.29417 |       3.0409 |     0.037785 |     0.037043 |    0.0063584 |
|   15 | Accept |      0.42302 |       1.9847 |     0.037785 |      0.03725 |       1.2221 |
|   16 | Accept |     0.039486 |       2.3734 |     0.037785 |     0.037672 |       0.10069 |
|   17 | Accept |     0.038591 |       2.5061 |     0.037785 |     0.037687 |       0.12077 |
|   18 | Accept |     0.038513 |       3.3339 |     0.037785 |     0.037696 |        0.1227 |
|   19 | Best   |     0.037757 |       3.6132 |     0.037757 |     0.037572 |       0.19621 |
|   20 | Accept |     0.037787 |       3.6766 |     0.037757 |     0.037601 |       0.18068 |
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |        Sigma |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              |
|=====================================================================================================|
|   21 | Accept |      0.44917 |       2.6148 |     0.037757 |      0.03766 |       8.7818 |
```

```
|   22 | Accept |    0.040201 |      2.3509 |    0.037757 |    0.037601 |     0.075414 |
|   23 | Accept |    0.040142 |      2.1471 |    0.037757 |    0.037607 |     0.087198 |
|   24 | Accept |     0.29417 |      2.8884 |    0.037757 |     0.03758 |    0.0031018 |
|   25 | Accept |     0.29417 |      2.9629 |    0.037757 |    0.037555 |   0.00019545 |
|   26 | Accept |     0.29417 |      3.2002 |    0.037757 |    0.037582 |     0.013608 |
|   27 | Accept |     0.29417 |      2.6996 |    0.037757 |    0.037556 |   0.00076147 |
|   28 | Accept |     0.42162 |      1.7615 |    0.037757 |    0.037854 |       0.6791 |
|   29 | Best   |    0.037704 |      2.7105 |    0.037704 |    0.037908 |       0.2367 |
|   30 | Accept |    0.037725 |       3.413 |    0.037704 |    0.037881 |      0.21743 |
```



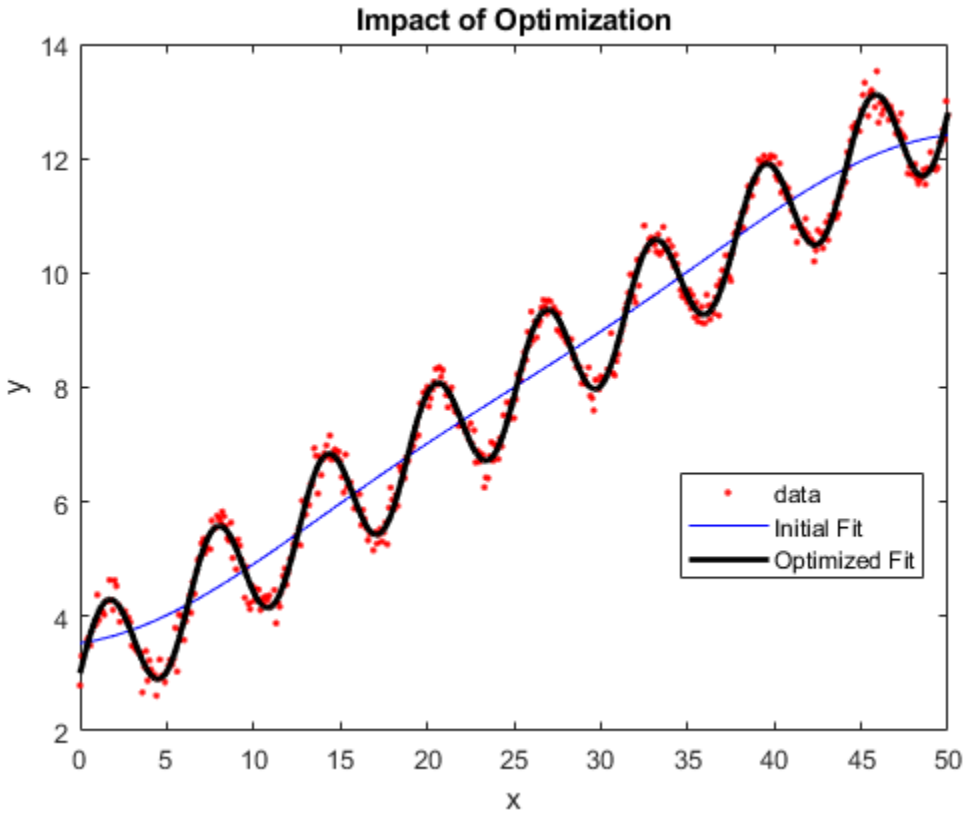



```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 121.0818 seconds
Total objective function evaluation time: 86.6996

Best observed feasible point:
    Sigma
    ______

    0.2367

Observed objective function value = 0.037704
Estimated objective function value = 0.038223
Function evaluation time = 2.7105

Best estimated feasible point (according to models):
     Sigma
    _______
```

```
    0.16159

Estimated objective function value = 0.037881
Estimated function evaluation time = 3.3496
```

Compare the pre- and post-optimization fits.

```
ypred1 = resubPredict(gprMdl1);
ypred2 = resubPredict(gprMdl2);

figure();
plot(x,y,'r.');
hold on
plot(x,ypred1,'b');
plot(x,ypred2,'k','LineWidth',2);
xlabel('x');
ylabel('y');
legend({'data','Initial Fit','Optimized Fit'},'Location','Best');
title('Impact of Optimization');
hold off
```



Train GPR Model Using Cross-Validation

This example uses the abalone data [1], [2], from the UCI Machine Learning Repository [3]. Download the data and save it in your current folder with the name `abalone.data`.

Store the data into a `table`. Display the first seven rows.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
tbl.Properties.VariableNames = {'Sex','Length','Diameter','Height','WWeight','SWeight','VWeight'
tbl(1:7,:)
```

```
ans =

    Sex    Length    Diameter    Height    WWeight    SWeight    VWeight    ShWeight    NoShellRi
    ___    ______    ________    ______    _______    _______    _______    ________    _________

    'M'    0.455     0.365       0.095      0.514     0.2245      0.101      0.15       15
    'M'     0.35     0.265        0.09     0.2255     0.0995     0.0485      0.07        7
    'F'     0.53      0.42       0.135      0.677     0.2565     0.1415      0.21        9
    'M'     0.44     0.365       0.125      0.516     0.2155      0.114     0.155       10
    'I'     0.33     0.255        0.08      0.205     0.0895     0.0395     0.055        7
    'I'    0.425       0.3       0.095     0.3515      0.141     0.0775      0.12        8
    'F'     0.53     0.415        0.15     0.7775      0.237     0.1415      0.33       20
```

The dataset has 4177 observations. The goal is to predict the age of abalone from eight physical measurements. The last variable, number of shell rings shows the age of the abalone. The first predictor is a categorical variable. The last variable in the table is the response variable.

Train a cross-validated GPR model using the 25% of the data for validation.

```
rng('default') % For reproducibility
cvgprMdl = fitrgp(tbl,'NoShellRings','Standardize',1,'Holdout',0.25);
```

Compute the average loss on folds using models trained on out-of-fold observations.

```
kfoldLoss(cvgprMdl)
```

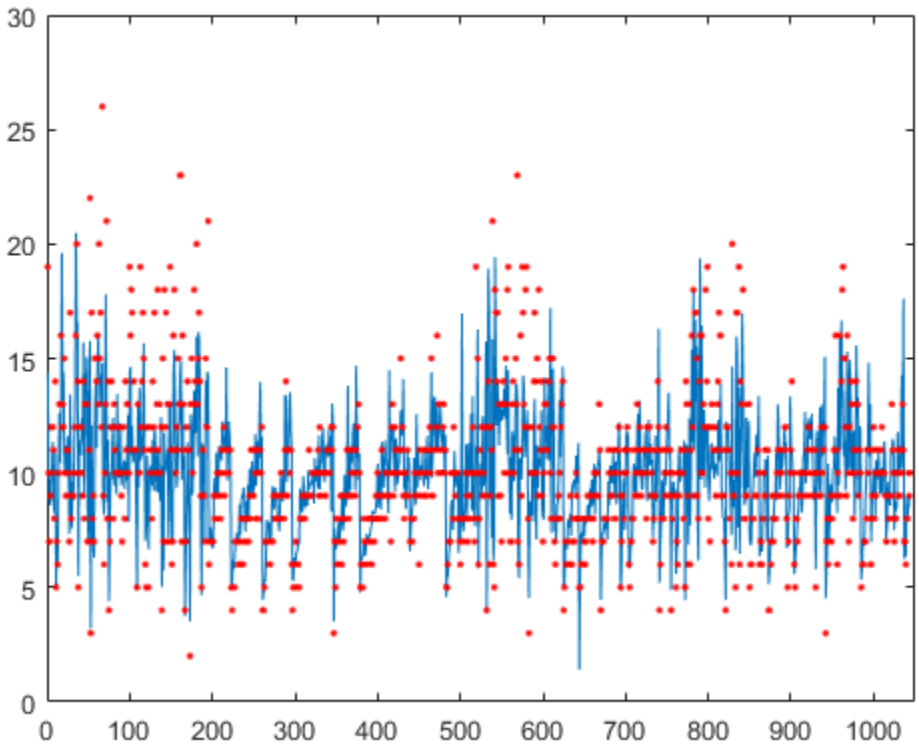
```
ans =
   4.6409
```

Predict the responses for out-of-fold data.

```
ypred = kfoldPredict(cvgprMdl);
```

Plot the true responses used for testing and the predictions.

```
figure();
plot(ypred(cvgprMdl.Partition.test));
hold on;
y = table2array(tbl(:,end));
plot(y(cvgprMdl.Partition.test),'r.');
axis([0 1050 0 30]);
xlabel('x')
ylabel('y')
hold off;
```



Fit GPR Model Using Custom Kernel Function

Generate the sample data.

```
rng(0,'twister'); % For reproducibility
n = 1000;
x = linspace(-10,10,n)';
y = 1 + x*5e-2 + sin(x)./x + 0.2*randn(n,1);
```

Define the squared exponential kernel function as a custom kernel function.

You can compute the squared exponential kernel function as

$$k(x_i,x_j|\theta) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{(x_i-x_j)^T(x_i-x_j)}{\sigma_l^2}\right),$$

where σ_f is the signal standard deviation, σ_l is the length scale. Both σ_f and σ_l must be greater than zero. This condition can be enforced by the unconstrained parametrization, $\sigma_l = \exp(\theta(1))$ and $\sigma_f = \exp(\theta(2))$, for some unconstrained parametrization vector θ.

Hence, you can define the squared exponential kernel function as a custom kernel function as follows:

```
kfcn = @(XN,XM,theta) (exp(theta(2))^2)*exp(-(pdist2(XN,XM).^2)/(2*exp(theta(1))^2));
```

Here `pdist2(XN,XM).^2` computes the distance matrix.

Fit a GPR model using the custom kernel function, `kfcn`. Specify the initial values of the kernel parameters (Because you use a custom kernel function, you must provide initial values for the unconstrained parametrization vector, `theta`).

```
theta0 = [1.5,0.2];
gprMdl = fitrgp(x,y,'KernelFunction',kfcn,'KernelParameters',theta0);
```

`fitrgp` uses analytical derivatives to estimate parameters when using a built-in kernel function, whereas when using a custom kernel function it uses numerical derivatives.

Compute the resubstitution loss for this model.

```
L = resubLoss(gprMdl)
```

```
L = 0.0391
```

Fit the GPR model using the built-in squared exponential kernel function option. Specify the initial values of the kernel parameters (Because you use the built-in custom kernel function and specifying initial parameter values, you must provide the initial values for the signal standard deviation and length scale(s) directly).

```
sigmaL0 = exp(1.5);
sigmaF0 = exp(0.2);
gprMdl2 = fitrgp(x,y,'KernelFunction','squaredexponential','KernelParameters',[sigmaL0,sigmaF0]);
```

Compute the resubstitution loss for this model.

```
L2 = resubLoss(gprMdl2)
```

```
L2 = 0.0391
```

The two loss values are the same as expected.

Specify Initial Step Size for LBFGS Optimization

Train a GPR model on generated data with many predictors. Specify the initial step size for the LBFGS optimizer.

Set the seed and type of the random number generator for reproducibility of the results.

```
rng(0,'twister'); % For reproducibility
```

Generate sample data with 300 observations and 3000 predictors, where the response variable depends on the 4th, 7th, and 13th predictors.

```
N = 300;
P = 3000;
X = rand(N,P);
y = cos(X(:,7)) + sin(X(:,4).*X(:,13)) + 0.1*randn(N,1);
```

Set initial values for the kernel parameters.

```
sigmaL0 = sqrt(P)*ones(P,1); % Length scale for predictors
sigmaF0 = 1; % Signal standard deviation
```

Set initial noise standard deviation to 1.

```
sigmaN0 = 1;
```

Specify `1e-2` as the termination tolerance for the relative gradient norm.

```
opts = statset('fitrgp');
opts.TolFun = 1e-2;
```

Fit a GPR model using the initial kernel parameter values, initial noise standard deviation, and an automatic relevance determination (ARD) squared exponential kernel function.

Specify the initial step size as 1 for determining the initial Hessian approximation for an LBFGS optimizer.

```
gpr = fitrgp(X,y,'KernelFunction','ardsquaredexponential','Verbose',1, ...
    'Optimizer','lbfgs','OptimizerOptions',opts, ...
    'KernelParameters',[sigmaL0;sigmaF0],'Sigma',sigmaN0,'InitialStepSize',1);
```

```
o Parameter estimation: FitMethod = Exact, Optimizer = lbfgs

 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=========================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=========================================================================================================
|        0 |  3.004966e+02 |   2.569e+02 |   0.000e+00 |        |   3.893e-03 |   0.000e+00 |
|        1 |  9.525779e+01 |   1.281e+02 |   1.003e+00 |     OK |   6.913e-03 |   1.000e+00 |
|        2 |  3.972026e+01 |   1.647e+01 |   7.639e-01 |     OK |   4.718e-03 |   5.000e-01 |
|        3 |  3.893873e+01 |   1.073e+01 |   1.057e-01 |     OK |   3.243e-03 |   1.000e+00 |
|        4 |  3.859904e+01 |   5.659e+00 |   3.282e-02 |     OK |   3.346e-03 |   1.000e+00 |
|        5 |  3.748912e+01 |   1.030e+01 |   1.395e-01 |     OK |   1.460e-03 |   1.000e+00 |
|        6 |  2.028104e+01 |   1.380e+02 |   2.010e+00 |     OK |   2.326e-03 |   1.000e+00 |
|        7 |  2.001849e+01 |   1.510e+01 |   9.685e-01 |     OK |   2.344e-03 |   1.000e+00 |
|        8 | -7.706109e+00 |   8.340e+01 |   1.125e+00 |     OK |   5.771e-04 |   1.000e+00 |
|        9 | -1.786074e+01 |   2.323e+02 |   2.647e+00 |     OK |   4.217e-03 |   1.250e-01 |
|       10 | -4.058422e+01 |   1.972e+02 |   6.796e-01 |     OK |   7.035e-03 |   1.000e+00 |
|       11 | -7.850209e+01 |   4.432e+01 |   8.335e-01 |     OK |   3.099e-03 |   1.000e+00 |
|       12 | -1.312162e+02 |   3.334e+01 |   1.277e+00 |     OK |   5.432e-02 |   1.000e+00 |
|       13 | -2.005064e+02 |   9.519e+01 |   2.828e+00 |     OK |   5.292e-03 |   1.000e+00 |
|       14 | -2.070150e+02 |   1.898e+01 |   1.641e+00 |     OK |   6.817e-03 |   1.000e+00 |
|       15 | -2.108086e+02 |   3.793e+01 |   7.685e-01 |     OK |   3.479e-03 |   1.000e+00 |
|       16 | -2.122920e+02 |   7.057e+00 |   1.591e-01 |     OK |   2.055e-03 |   1.000e+00 |
|       17 | -2.125610e+02 |   4.337e+00 |   4.818e-02 |     OK |   1.974e-03 |   1.000e+00 |
|       18 | -2.130162e+02 |   1.178e+01 |   8.891e-02 |     OK |   2.856e-03 |   1.000e+00 |
|       19 | -2.139378e+02 |   1.933e+01 |   2.371e-01 |     OK |   1.029e-02 |   1.000e+00 |

|=========================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=========================================================================================================
|       20 | -2.151111e+02 |   1.550e+01 |   3.015e-01 |     OK |   2.765e-02 |   1.000e+00 |
|       21 | -2.173046e+02 |   5.856e+00 |   6.537e-01 |     OK |   1.414e-02 |   1.000e+00 |
|       22 | -2.201781e+02 |   8.918e+00 |   8.484e-01 |     OK |   6.381e-03 |   1.000e+00 |
|       23 | -2.288858e+02 |   4.846e+01 |   2.311e+00 |     OK |   2.661e-03 |   1.000e+00 |
|       24 | -2.392171e+02 |   1.190e+02 |   6.283e+00 |     OK |   8.113e-03 |   1.000e+00 |
|       25 | -2.511145e+02 |   1.008e+02 |   1.198e+00 |     OK |   1.605e-02 |   1.000e+00 |
|       26 | -2.742547e+02 |   2.207e+01 |   1.231e+00 |     OK |   3.191e-03 |   1.000e+00 |
|       27 | -2.849931e+02 |   5.067e+01 |   3.660e+00 |     OK |   5.184e-03 |   1.000e+00 |
```

```
|       28 | -2.899797e+02 |   2.068e+01 |   1.162e+00 |    OK   |   6.270e-03 |   1.000e+00 |   Y
|       29 | -2.916723e+02 |   1.816e+01 |   3.213e-01 |    OK   |   1.415e-02 |   1.000e+00 |   Y
|       30 | -2.947674e+02 |   6.965e+00 |   1.126e+00 |    OK   |   6.339e-03 |   1.000e+00 |   Y
|       31 | -2.962491e+02 |   1.349e+01 |   2.352e-01 |    OK   |   8.999e-03 |   1.000e+00 |   Y
|       32 | -3.004921e+02 |   1.586e+01 |   9.880e-01 |    OK   |   3.940e-02 |   1.000e+00 |   Y
|       33 | -3.118906e+02 |   1.889e+01 |   3.318e+00 |    OK   |   1.213e-01 |   1.000e+00 |   Y
|       34 | -3.189215e+02 |   7.086e+01 |   3.070e+00 |    OK   |   8.095e-03 |   1.000e+00 |   Y
|       35 | -3.245557e+02 |   4.366e+00 |   1.397e+00 |    OK   |   2.718e-03 |   1.000e+00 |   Y
|       36 | -3.254613e+02 |   3.751e+00 |   6.546e-01 |    OK   |   1.004e-02 |   1.000e+00 |   Y
|       37 | -3.262823e+02 |   4.011e+00 |   2.026e-01 |    OK   |   2.441e-02 |   1.000e+00 |   Y
|       38 | -3.325606e+02 |   1.773e+01 |   2.427e+00 |    OK   |   5.234e-02 |   1.000e+00 |   Y
|       39 | -3.350374e+02 |   1.201e+01 |   1.603e+00 |    OK   |   2.674e-02 |   1.000e+00 |   Y

|=============================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=============================================================================================
|       40 | -3.379112e+02 |   5.280e+00 |   1.393e+00 |    OK   |   1.177e-02 |   1.000e+00 |   Y
|       41 | -3.389136e+02 |   3.061e+00 |   7.121e-01 |    OK   |   2.935e-02 |   1.000e+00 |   Y
|       42 | -3.401070e+02 |   4.094e+00 |   6.224e-01 |    OK   |   3.399e-02 |   1.000e+00 |   Y
|       43 | -3.436291e+02 |   8.833e+00 |   1.707e+00 |    OK   |   5.231e-02 |   1.000e+00 |   Y
|       44 | -3.456295e+02 |   5.891e+00 |   1.424e+00 |    OK   |   3.772e-02 |   1.000e+00 |   Y
|       45 | -3.460069e+02 |   1.126e+01 |   2.580e+00 |    OK   |   3.907e-02 |   1.000e+00 |   Y
|       46 | -3.481756e+02 |   1.546e+00 |   8.142e-01 |    OK   |   1.565e-02 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 1.546e+00
             Two norm of the final step     = 8.142e-01, TolX   = 1.000e-12
Relative infinity norm of the final gradient = 6.016e-03, TolFun = 1.000e-02
EXIT: Local minimum found.

o Alpha estimation: PredictMethod = Exact
```

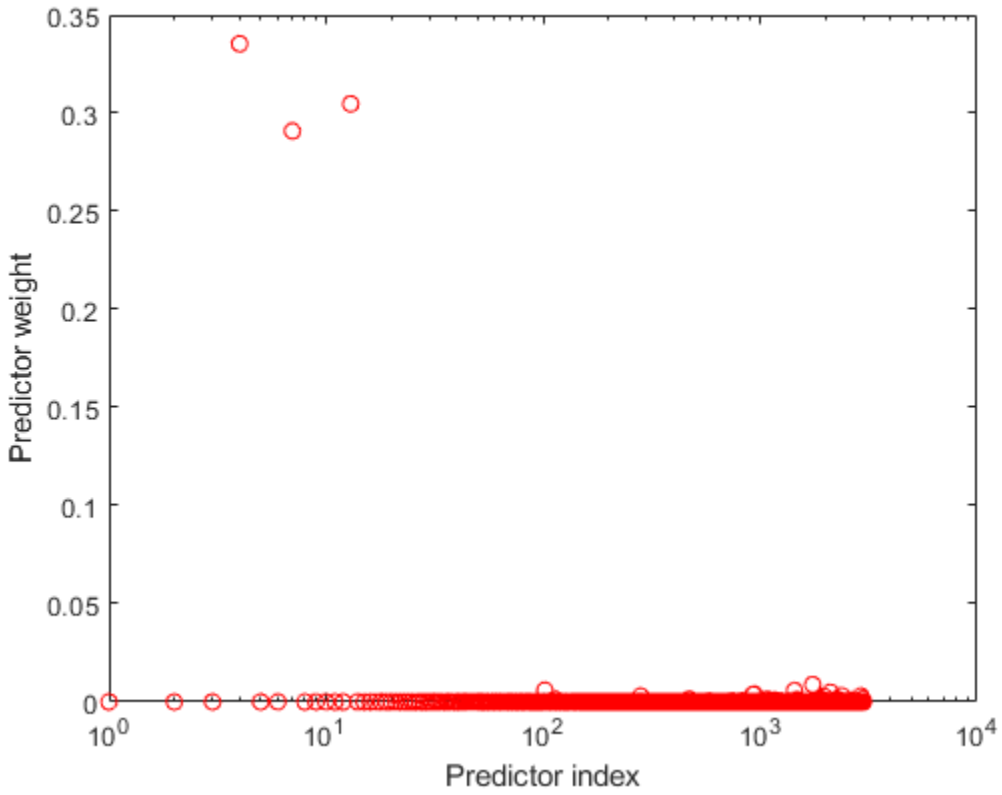
Because the GPR model uses an ARD kernel with many predictors, using an LBFGS approximation to the Hessian is more memory efficient than storing the full Hessian matrix. Also, using the initial step size to determine the initial Hessian approximation may help speed up optimization.

Find the predictor weights by taking the exponential of the negative learned length scales. Normalize the weights.

```
sigmaL = gpr.KernelInformation.KernelParameters(1:end-1); % Learned length scales
weights = exp(-sigmaL); % Predictor weights
weights = weights/sum(weights); % Normalized predictor weights
```

Plot the normalized predictor weights.

```
figure;
semilogx(weights,'ro');
xlabel('Predictor index');
ylabel('Predictor weight');
```



The trained GPR model assigns the largest weights to the 4th, 7th, and 13th predictors. The irrelevant predictors have weights close to zero.

Input Arguments

`Tbl` — Sample data

`table`

Sample data used to train the model, specified as a `table`. Each row of `Tbl` corresponds to one observation, and each column corresponds to one variable. `Tbl` contains the predictor variables, and optionally it can also contain one column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all the remaining variables as predictors, then specify the response variable using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the predictors in training the model, then specify the response variable and the predictor variables using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable using `y`. The length of the response variable and the number of rows in `Tbl` must be equal.

For more information on the `table` data type, see `table`.

If your predictor data contains categorical variables, then `fitrgp` creates dummy variables. For details, see `CategoricalPredictors`.

Data Types: `table`

`ResponseVarName` — Response variable name
name of a variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `y` is stored in `Tbl` (as `Tbl.y`), then specify it as `'y'`. Otherwise, the software treats all the columns of `Tbl`, including `y`, as predictors when training the model.

Data Types: `char` | `string`

`formula` — Response and predictor variables to use in model training
character vector or string scalar in the form of `'y~x1+x2+x3'`

Response and predictor variables to use in model training, specified as a character vector or string scalar in the form of `'y~x1+x2+x3'`. In this form, `y` represents the response variable; `x1`, `x2`, `x3` represent the predictor variables to use in training the model.

Use a formula if you want to specify a subset of variables in `Tbl` as predictors to use when training the model. If you specify a formula, then any variables that do not appear in `formula` are not used to train the model.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

The formula does not indicate the form of the `BasisFunction`.

Example: `'PetalLength~PetalWidth+Species'` identifies the variable `PetalLength` as the response variable, and `PetalWidth` and `Species` as the predictor variables.

Data Types: `char` | `string`

`X` — Predictor data for the GPR model
n-by-*d* matrix

Predictor data for the GPR model, specified as an *n*-by-*d* matrix. *n* is the number of observations (rows), and *d* is the number of predictors (columns).

The length of `y` and the number of rows of `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `double`

`y` — Response data for the GPR model
n-by-1 vector

Response data for the GPR model, specified as an *n*-by-1 vector. You can omit `y` if you provide the `Tbl` training data that also includes `y`. In that case, use `ResponseVarName` to identify the response variable or use `formula` to identify the response and predictor variables.

Data Types: `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'FitMethod','sr','BasisFunction','linear','ActiveSetMethod','sgma','PredictMethod','fic'` trains the GPR model using the subset of regressors approximation method for parameter estimation, uses a linear basis function, uses sparse greedy matrix approximation for active selection, and fully independent conditional approximation method for prediction.

Fitting

`FitMethod` — Method to estimate parameters of the GPR model

`'none'` | `'exact'` | `'sd'` | `'sr'` | `'fic'`

Method to estimate parameters of the GPR model, specified as the comma-separated pair consisting of `'FitMethod'` and one of the following.

Fit Method	Description
`'none'`	No estimation, use the initial parameter values as the known parameter values.
`'exact'`	Exact Gaussian process regression. Default if $n \le 2000$, where n is the number of observations.
`'sd'`	Subset of data points approximation. Default if $n > 2000$, where n is the number of observations.
`'sr'`	Subset of regressors approximation.
`'fic'`	Fully independent conditional approximation.

Example: `'FitMethod','fic'`

`BasisFunction` — Explicit basis in the GPR model

`'constant'` (default) | `'none'` | `'linear'` | `'pureQuadratic'` | function handle

Explicit basis in the GPR model, specified as the comma-separated pair consisting of `'BasisFunction'` and one of the following. If n is the number of observations, the basis function adds the term $\boldsymbol{H}*\boldsymbol{\beta}$ to the model, where $\boldsymbol{H}$ is the basis matrix and $\boldsymbol{\beta}$ is a p-by-1 vector of basis coefficients.

Explicit Basis	Basis Matrix
`'none'`	Empty matrix.

Explicit Basis	Basis Matrix
`'constant'`	$H = 1$ (n-by-1 vector of 1s, where n is the number of observations)
`'linear'`	$H = [1, X]$
`'pureQuadratic'`	$H = [1, X, X_2]$, where $X_2 = \begin{bmatrix} x_{11}^2 & x_{12}^2 & \cdots & x_{1d}^2 \\ x_{21}^2 & x_{22}^2 & \cdots & x_{2d}^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 & x_{n2}^2 & \cdots & x_{nd}^2 \end{bmatrix}$.
Function handle	Function handle, `hfcn`, that `fitrgp` calls as: $H = hfcn(X)$, where X is an n-by-d matrix of predictors and H is an n-by-p matrix of basis functions.

Example: `'BasisFunction','pureQuadratic'`

Data Types: `char` | `string` | `function_handle`

`Beta` — Initial value of the coefficients
p-by-1 vector

Initial value of the coefficients for the explicit basis, specified as the comma-separated pair consisting of `'Beta'` and *p*-by-1 vector, where *p* is the number of columns in the basis matrix $\boldsymbol{H}$.

The basis matrix depends on the choice of the explicit basis function as follows (also see `BasisFunction`).

`fitrgp` uses the coefficient initial values as the known coefficient values, only when `FitMethod` is `'none'`.

Data Types: `double`

`Sigma` — Initial value for the noise standard deviation of the Gaussian process model
`std(y)/sqrt(2)` (default) | positive scalar value

Initial value for the noise standard deviation of the Gaussian process model, specified as the comma-separated pair consisting of `'Sigma'` and a positive scalar value.

Example: `'Sigma',2`

Data Types: `double`

`ConstantSigma` — Constant value of `Sigma` for the noise standard deviation of the Gaussian process model
`false` (default) | `true`

Constant value of `Sigma` for the noise standard deviation of the Gaussian process model, specified as a logical scalar. When `ConstantSigma` is `true`, `fitrgp` does not optimize the value of `Sigma`, but instead takes the initial value as the value throughout its computations.

Example: `'ConstantSigma',true`

Data Types: `logical`

`SigmaLowerBound` — Lower bound on the noise standard deviation
`1e-2*std(y)` (default) | positive scalar value

Lower bound on the noise standard deviation, specified as the comma-separated pair consisting of `'SigmaLowerBound'` and a positive scalar value.

Example: `'SigmaLowerBound',0.02`

Data Types: `double`

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrgp` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrgp` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrgp` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitrgp` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered

categorical variable, `fitrgp` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitrgp` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`Standardize` — Indicator to standardize data
`0 (false)` (default) | logical value

Indicator to standardize data, specified as the comma-separated pair consisting of `'Standardize'` and a logical value.

If you set `'Standardize',1`, then the software centers and scales each column of the predictor data, by the column mean and standard deviation, respectively. The software does not standardize the data contained in the dummy variable columns that it generates for categorical predictors.

Example: `'Standardize',1`

Example: `'Standardize',true`

Data Types: `logical`

`Regularization` — Regularization standard deviation
`1e-2*std(y)` (default) | positive scalar value

Regularization standard deviation for sparse methods subset of regressors (`'sr'`) and fully independent conditional (`'fic'`), specified as the comma-separated pair consisting of `'Regularization'` and a positive scalar value.

Example: `'Regularization',0.2`

Data Types: `double`

`ComputationMethod` — Method for computing log likelihood and gradient
`'qr'` (default) | `'v'`

Method for computing the log likelihood and gradient for parameter estimation using subset of regressors (`'sr'`) and fully independent conditional (`'fic'`) approximation methods, specified as the comma-separated pair consisting of `'ComputationMethod'` and one of the following.

- `'qr'` — Use QR factorization based approach, this option provides better accuracy.
- `'v'` — Use V-method-based approach. This option provides faster computation of log likelihood gradients.

Example: `'ComputationMethod','v'`

Kernel (Covariance) Function

`KernelFunction` — Form of the covariance function
`'squaredexponential'` (default) | `'exponential'` | `'matern32'` | `'matern52'` | `'rationalquadratic'` | `'ardsquaredexponential'` | `'ardexponential'` | `'ardmatern32'` | `'ardmatern52'` | `'ardrationalquadratic'` | function handle

Form of the covariance function, specified as the comma-separated pair consisting of `'KernelFunction'` and one of the following.

Function	Description
`'exponential'`	Exponential kernel.
`'squaredexponential'`	Squared exponential kernel.
`'matern32'`	Matern kernel with parameter 3/2.
`'matern52'`	Matern kernel with parameter 5/2.
`'rationalquadratic'`	Rational quadratic kernel.
`'ardexponential'`	Exponential kernel with a separate length scale per predictor.
`'ardsquaredexponential'`	Squared exponential kernel with a separate length scale per predictor.
`'ardmatern32'`	Matern kernel with parameter 3/2 and a separate length scale per predictor.
`'ardmatern52'`	Matern kernel with parameter 5/2 and a separate length scale per predictor.
`'ardrationalquadratic'`	Rational quadratic kernel with a separate length scale per predictor.
Function handle	A function handle that can be called like this: `Kmn = kfcn(Xm,Xn,theta)` where `Xm` is an *m*-by-*d* matrix, `Xn` is an *n*-by-*d* matrix and `Kmn` is an *m*-by-*n* matrix of kernel products such that `Kmn`(*i*,*j*) is the kernel product between `Xm`(*i*,:) and `Xn`(*j*,:). `theta` is the *r*-by-1 unconstrained parameter vector for `kfcn`.

For more information on the kernel functions, see “Kernel (Covariance) Function Options” on page 6-6.

Example: `'KernelFunction','matern32'`

Data Types: `char` | `string` | `function_handle`

`KernelParameters` — Initial values for the kernel parameters
vector

Initial values for the kernel parameters, specified as the comma-separated pair consisting of `'KernelParameters'` and a vector. The size of the vector and the values depend on the form of the covariance function, specified by the `KernelFunction` name-value pair argument.

'KernelFunction'	'KernelParameters'
'exponential', 'squaredexponential', 'matern32', or 'matern52'	2-by-1 vector phi, where phi(1) contains the length scale and phi(2) contains the signal standard deviation. Default initial value of the length scale parameter is the mean of standard deviations of the predictors, and the signal standard deviation is the standard deviation of the responses divided by square root of 2. That is, phi = [mean(std(X));std(y)/sqrt(2)]
'rationalquadratic'	3-by-1 vector phi, where phi(1) contains the length scale, phi(2) contains the scale-mixture parameter, and phi(3) contains the signal standard deviation. Default initial value of the length scale parameter is the mean of standard deviations of the predictors, and the signal standard deviation is the standard deviation of the responses divided by square root of 2. Default initial value for the scale-mixture parameter is 1. That is, phi = [mean(std(X));1;std(y)/sqrt(2)]
'ardexponential', 'ardsquaredexponential', 'ardmatern32', or 'ardmatern52'	(d+1)-by-1 vector phi, where phi(i) contains the length scale for predictor i, and phi(d+1) contains the signal standard deviation. d is the number of predictor variables after dummy variables are created for categorical variables. For details about creating dummy variables, see CategoricalPredictors. Default initial value of the length scale parameters are the standard deviations of the predictors and the signal standard deviation is the standard deviation of the responses divided by square root of 2. That is, phi = [std(X)';std(y)/sqrt(2)]
'ardrationalquadratic'	(d+2)-by-1 vector phi, where phi(i) contains the length scale for predictor i, phi(d+1) contains the scale-mixture parameter, and phi(d+2) contains signal standard deviation. Default initial value of the length scale parameters are the standard deviations of the predictors and the signal standard deviation is the standard deviation of the responses divided by square root of 2. Default initial value for the scale-mixture parameter is 1. That is, phi = [std(X)';1;std(y)/sqrt(2)]

'KernelFunction'	'KernelParameters'
Function handle	*r*-by-1 vector as the initial value of the unconstrained parameter vector `phi` for the custom kernel function `kfcn`. When `KernelFunction` is a function handle, you must supply initial values for the kernel parameters.

For more information on the kernel functions, see “Kernel (Covariance) Function Options” on page 6-6.

Example: `'KernelParameters',theta`

Data Types: `double`

`DistanceMethod` — Method for computing inter-point distances
`'fast'` (default) | `'accurate'`

Method for computing inter-point distances to evaluate built-in kernel functions, specified as the comma-separated pair consisting of `'DistanceMethod'` and either `'fast'` or `'accurate'`.

`fitrgp` computes $(x-y)^2$ as $x^2+y^2-2*x*y$ when you choose the `fast` option and as $(x-y)^2$ when you choose the `accurate` option.

Example: `'DistanceMethod','accurate'`

Active Set Selection

`ActiveSet` — Observations in the active set
`[]` (default) | *m*-by-1 vector of integers ranging from 1 to *n* ($m \le n$) | logical vector of length *n*

Observations in the active set, specified as the comma-separated pair consisting of `'ActiveSet'` and an *m*-by-1 vector of integers ranging from 1 to *n* ($m \le n$) or a logical vector of length *n* with at least one `true` element. *n* is the total number of observations in the training data.

`fitrgp` uses the observations indicated by `ActiveSet` to train the GPR model. The active set cannot have duplicate elements.

If you supply `ActiveSet`, then:

- `fitrgp` does not use `ActiveSetSize` and `ActiveSetMethod`.
- You cannot perform cross-validation on this model.

Data Types: `double` | `logical`

`ActiveSetSize` — Size of the active set for sparse methods
an integer *m* ($1 \le m \le n$)

Size of the active set for sparse methods (`'sd'`, `'sr'`, `'fic'`), specified as the comma-separated pair consisting of `'ActiveSetSize'` and an integer *m*, $1 \le m \le n$, where *n* is the number of observations.

Default is min(1000,*n*) when `FitMethod` is `'sr'` or `'fic'`, and min(2000,*n*), otherwise.

Example: `'ActiveSetSize',100`

Data Types: `double`

`ActiveSetMethod` — Active set selection method
`'random'` (default) | `'sgma'` | `'entropy'` | `'likelihood'`

Active set selection method, specified as the comma-separated pair consisting of `'ActiveSetMethod'` and one of the following.

Method	Description
`'random'`	Random selection
`'sgma'`	Sparse greedy matrix approximation
`'entropy'`	Differential entropy-based selection
`'likelihood'`	Subset of regressors log likelihood-based selection

All active set selection methods (except `'random'`) require the storage of an *n*-by-*m* matrix, where *m* is the size of the active set and *n* is the number of observations.

Example: `'ActiveSetMethod','entropy'`

`RandomSearchSetSize` — Random search set size
59 (default) | integer value

Random search set size per greedy inclusion for active set selection, specified as the comma-separated pair consisting of `'RandomSearchSetSize'` and an integer value.

Example: `'RandomSearchSetSize',30`

Data Types: `double`

`ToleranceActiveSet` — Relative tolerance for terminating active set selection
1e-06 (default) | positive scalar

Relative tolerance for terminating active set selection, specified as the comma-separated pair consisting of `'ToleranceActiveset'` and a positive scalar value.

Example: `'ToleranceActiveset',0.0002`

Data Types: `double`

`NumActiveSetRepeats` — Number of repetitions
3 (default) | integer value

Number of repetitions for interleaved active set selection and parameter estimation on page 33-2172 when `ActiveSetMethod` is not `'random'`, specified as the comma-separated pair consisting of `'NumActiveSetRepeats'` and an integer value.

Example: `'NumActiveSetRepeats',5`

Data Types: `double`

Prediction

`PredictMethod` — Method used to make predictions
`'exact'` | `'bcd'` | `'sd'` | `'sr'` | `'fic'`

Method used to make predictions from a Gaussian process model given the parameters, specified as the comma-separated pair consisting of `'PredictMethod'` and one of the following.

Method	Description
'exact'	Exact Gaussian process regression method. Default, if $n \le 10000$.
'bcd'	Block coordinate descent. Default, if $n > 10000$.
'sd'	Subset of data points approximation.
'sr'	Subset of regressors approximation.
'fic'	Fully independent conditional approximation.

Example: 'PredictMethod','bcd'

BlockSizeBCD — Block size
minimum of 1000 or *n* (default) | integer in the range from 1 to *n*

Block size for block coordinate descent method ('bcd'), specified as the comma-separated pair consisting of 'BlockSizeBCD' and an integer in the range from 1 to *n*, where *n* is the number of observations.

Example: 'BlockSizeBCD',1500

Data Types: double

NumGreedyBCD — Number of greedy selections
minimum of 100 and BlockSizeBCD (default) | integer value in the range from 1 to BlockSizeBCD

Number of greedy selections for block coordinate descent method ('bcd'), specified as the comma-separated pair consisting of 'NumGreedyBCD' and an integer in the range from 1 to BlockSizeBCD.

Example: 'NumGreedyBCD',150

Data Types: double

ToleranceBCD — Relative tolerance on gradient norm
1e-3 (default) | positive scalar

Relative tolerance on gradient norm for terminating block coordinate descent method ('bcd') iterations, specified as the comma-separated pair consisting of 'ToleranceBCD' and a positive scalar.

Example: 'ToleranceBCD',0.002

Data Types: double

StepToleranceBCD — Absolute tolerance on step size
1e-3 (default) | positive scalar

Absolute tolerance on step size for terminating block coordinate descent method ('bcd') iterations, specified as the comma-separated pair consisting of 'StepToleranceBCD' and a positive scalar.

Example: 'StepToleranceBCD',0.002

Data Types: double

IterationLimitBCD — Maximum number of BCD iterations
1000000 (default) | integer value

Maximum number of block coordinate descent method (`'bcd'`) iterations, specified as the comma-separated pair consisting of `'IterationLimitBCD'` and an integer value.

Example: `'IterationLimitBCD',10000`

Data Types: `double`

Optimization

`Optimizer` — Optimizer to use for parameter estimation

`'quasinewton'` (default) | `'lbfgs'` | `'fminsearch'` | `'fminunc'` | `'fmincon'`

Optimizer to use for parameter estimation, specified as the comma-separated pair consisting of `'Optimizer'` and one of the values in this table.

Value	Description
`'quasinewton'`	Dense, symmetric rank-1-based, quasi-Newton approximation to the Hessian
`'lbfgs'`	LBFGS-based quasi-Newton approximation to the Hessian
`'fminsearch'`	Unconstrained nonlinear optimization using the simplex search method of Lagarias et al. [5]
`'fminunc'`	Unconstrained nonlinear optimization (requires an Optimization Toolbox license)
`'fmincon'`	Constrained nonlinear optimization (requires an Optimization Toolbox license)

For more information on the optimizers, see Algorithms on page 33-2173.

Example: `'Optimizer','fmincon'`

`OptimizerOptions` — Options for the optimizer

structure | object

Options for the optimizer you choose using the `Optimizer` name-value pair argument, specified as the comma-separated pair consisting of `'OptimizerOptions'` and a structure or object created by `optimset`, `statset('fitrgp')`, or `optimoptions`.

Optimizer	Create Optimizer Options Using
`'fminsearch'`	`optimset` (structure)
`'quasinewton'` or `'lbfgs'`	`statset('fitrgp')` (structure)
`'fminunc'` or `'fmincon'`	`optimoptions` (object)

The default options depend on the type of optimizer.

Example: `'OptimizerOptions',opt`

`InitialStepSize` — Initial step size

`[]` (default) | real positive scalar | `'auto'`

Initial step size, specified as the comma-separated pair consisting of `'InitialStepSize'` and a real positive scalar or `'auto'`.

'InitialStepSize' is the approximate maximum absolute value of the first optimization step when the optimizer is 'quasinewton' or 'lbfgs'. The initial step size can determine the initial Hessian approximation during optimization.

By default, fitrgp does not use the initial step size to determine the initial Hessian approximation. To use the initial step size, set a value for the 'InitialStepSize' name-value pair argument, or specify 'InitialStepSize','auto' to have fitrgp determine a value automatically. For more information on 'auto', see Algorithms on page 33-2173.

Example: 'InitialStepSize','auto'

Cross-Validation

CrossVal — Indicator for cross-validation

'off' (default) | 'on'

Indicator for cross-validation, specified as the comma-separated pair consisting of 'CrossVal' and either 'off' or 'on'. If it is 'on', then fitrgp returns a GPR model cross-validated with 10 folds.

You can use one of the KFold, Holdout, Leaveout or CVPartition name-value pair arguments to change the default cross-validation settings. You can use only one of these name-value pairs at a time.

As an alternative, you can use the crossval method for your model.

Example: 'CrossVal','on'

CVPartition — Random partition for a stratified *k*-fold cross-validation

cvpartition object

Random partition for a stratified *k*-fold cross-validation, specified as the comma-separated pair consisting of 'CVPartition' and a cvpartition object.

Example: 'CVPartition',cvp uses the random partition defined by cvp.

If you specify CVPartition, then you cannot specify Holdout, KFold, or Leaveout.

Holdout — Fraction of data to use for testing

scalar value in the range from 0 to 1

Fraction of the data to use for testing in holdout validation, specified as the comma-separated pair consisting of 'Holdout' and a scalar value in the range from 0 to 1. If you specify 'Holdout',p, then the software:
1. Randomly reserves around *p**100% of the data as validation data, and trains the model using the rest of the data
2. Stores the compact, trained model in cvgprMdl.Trained.

Example: 'Holdout', 0.3 uses 30% of the data for testing and 70% of the data for training.

If you specify Holdout, then you cannot specify CVPartition, KFold, or Leaveout.

Data Types: double

KFold — Number of folds

10 (default) | positive integer value

Number of folds to use in cross-validated GPR model, specified as the comma-separated pair consisting of 'KFold' and a positive integer value. KFold must be greater than 1. If you specify 'KFold',k then the software:

1. Randomly partitions the data into *k* sets.
2. For each set, reserves the set as test data, and trains the model using the other *k* – 1 sets.
3. Stores the *k* compact, trained models in the cells of a *k*-by-1 cell array in `cvgprMdl.Trained`.

Example: `'KFold',5` uses 5 folds in cross-validation. That is, for each fold, uses that fold as test data, and trains the model on the remaining 4 folds.

If you specify `KFold`, then you cannot specify `CVPartition`, `Holdout`, or `Leaveout`.

Data Types: `double`

`Leaveout` — Indicator for leave-one-out cross-validation
`'off'` (default) | `'on'`

Indicator for leave-one-out cross-validation, specified as the comma-separated pair consisting of `'Leaveout'` and either `'off'` or `'on'`.

If you specify `'Leaveout','on'`, then, for each of the *n* observations, the software:
1. Reserves the observation as test data, and trains the model using the other *n* – 1 observations.
2. Stores the compact, trained model in a cell in the *n*-by-1 cell array `cvgprMdl.Trained`.

Example: `'Leaveout','on'`

If you specify `Leaveout`, then you cannot specify `CVPartition`, `Holdout`, or `KFold`.

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Sigma'}`.
- `'all'` — Optimize all eligible parameters, equivalent to`{'BasisFunction','KernelFunction','KernelScale','Sigma','Standardize'}`.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitrgp` by varying the parameters. For information about cross-validation loss (albeit in a different context), see “Classification Loss” on page 33-3184. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitrgp` are:

- `BasisFunction` — `fitrgp` searches among `'constant'`, `'none'`, `'linear'`, and `'pureQuadratic'`.

- `KernelFunction` — `fitrgp` searches among `'ardexponential'`, `'ardmatern32'`, `'ardmatern52'`, `'ardrationalquadratic'`, `'ardsquaredexponential'`, `'exponential'`, `'matern32'`, `'matern52'`, `'rationalquadratic'`, and `'squaredexponential'`.
- `KernelScale` — `fitrgp` uses the `KernelParameters` argument to specify the value of the kernel scale parameter, which is held constant during fitting. In this case, all input dimensions are constrained to have the same `KernelScale` value. `fitrgp` searches among real value in the range `[1e-3*MaxPredictorRange,MaxPredictorRange]`, where

  `MaxPredictorRange = max(max(X) - min(X))`.

  `KernelScale` cannot be optimized for any of the ARD kernels.
- `Sigma` — `fitrgp` searches among real value in the range `[1e-4, max(1e-3,10*ResponseStd)]`, where

  `ResponseStd = std(y)`.

  Internally, `fitrgp` sets the `ConstantSigma` name-value pair to `true` so the value of `Sigma` is constant during the fitting.
- `Standardize` — `fitrgp` searches among `true` and `false`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load fisheriris
params = hyperparameters('fitrgp',meas,species);
params(1).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize GPR Regression” on page 33-2145.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Other

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array of unique names or a cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.

- If you supply `X` and `y`, then you can use `'PredictorNames'` to give the predictor variables in `X` names.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitrgp` uses the predictor variables in `PredictorNames` and the response only in training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - It good practice to specify the predictors for training using one of `'PredictorNames'` or `formula` only.

Example: `'PredictorNames',{'PedalLength','PedalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name

`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Verbose` — Verbosity level

0 (default) | 1

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and one of the following.

- 0 — `fitrgp` suppresses diagnostic messages related to active set selection and block coordinate descent but displays the messages related to parameter estimation, depending on the value of `'Display'` in `OptimizerOptions`.
- 1 — `fitrgp` displays the iterative diagnostic messages related to parameter estimation, active set selection, and block coordinate descent.

Example: `'Verbose',1`

`CacheSize` — Cache size in megabytes
1000 MB (default) | positive scalar

Cache size in megabytes (MB), specified as the comma-separated pair consisting of `'CacheSize'` and a positive scalar. Cache size is the extra memory that is available in addition to that required for fitting and active set selection. `fitrgp` uses `CacheSize` to:

- Decide whether interpoint distances should be cached when estimating parameters.
- Decide how matrix vector products should be computed for block coordinate descent method and for making predictions.

Example: `'CacheSize',2000`

Data Types: `double`

Output Arguments

`gprMdl` — Gaussian process regression model
`RegressionGP` object | `RegressionPartitionedModel` object

Gaussian process regression model, returned as a `RegressionGP` or a `RegressionPartitionedModel` object.

- If you cross-validate, that is, if you use one of the `'Crossval'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'` name-value pairs, then `gprMdl` is a `RegressionPartitionedModel` object. You cannot use a `RegressionPartitionedModel` object to make predictions using `predict`. For more information on the methods and properties of this object, see `RegressionPartitionedModel`.
- If you do not cross-validate, then `gprMdl` is a `RegressionGP` object. You can use this object for predictions using the `predict` method. For more information on the methods and properties of this object, see `RegressionGP`.

More About

Active Set Selection and Parameter Estimation

For subset of data, subset of regressors, or fully independent conditional approximation fitting methods (`FitMethod` equal to `'sd'`, `'sr'`, or `'fic'`), if you do not provide the active set, `fitrgp` selects the active set and computes the parameter estimates in a series of iterations.

In the first iteration, the software uses the initial parameter values in vector $\eta_0 = [\beta_0,\sigma_0,\theta_0]$ to select an active set A_1. It maximizes the GPR marginal log likelihood or its approximation using η_0 as the initial values and A_1 to compute the new parameter estimates η_1. Next, it computes the new log likelihood L_1 using η_1 and A_1.

In the second iteration, the software selects the active set A_2 using the parameter values in η_1. Then, using η_1 as the initial values and A_2, it maximizes the GPR marginal log likelihood or its approximation and estimates the new parameter values η_2. Then using η_2 and A_2, computes the new log likelihood value L_2.

The following table summarizes the iterations and what is computed at each iteration.

Iteration Number	Active Set	Parameter Vector	Log Likelihood
1	A_1	η_1	L_1
2	A_2	η_2	L_2
3	A_3	η_3	L_3
…	…	…	…

The software iterates similarly for a specified number of repetitions. You can specify the number of replications for active set selection using the `NumActiveSetRepeats` name-value pair argument.

Tips

- `fitrgp` accepts any combination of fitting, prediction, and active set selection methods. In some cases it might not be possible to compute the standard deviations of the predicted responses, hence the prediction intervals. See `predict`. And in some cases, using the exact method might be expensive due to the size of the training data.
- The `PredictorNames` property stores one element for each of the original predictor variable names. For example, if there are three predictors, one of which is a categorical variable with three levels, `PredictorNames` is a 1-by-3 cell array of character vectors.
- The `ExpandedPredictorNames` property stores one element for each of the predictor variables, including the dummy variables. For example, if there are three predictors, one of which is a categorical variable with three levels, then `ExpandedPredictorNames` is a 1-by-5 cell array of character vectors.
- Similarly, the `Beta` property stores one beta coefficient for each predictor, including the dummy variables.
- The `X` property stores the training data as originally input. It does not include the dummy variables.
- The default approach to initializing the Hessian approximation in `fitrgp` can be slow when you have a GPR model with many kernel parameters, such as when using an ARD kernel with many predictors. In this case, consider specifying `'auto'` or a value for the initial step size.

  You can set `'Verbose',1` for display of iterative diagnostic messages, and begin training a GPR model using an LBFGS or quasi-Newton optimizer with the default `fitrgp` optimization. If the iterative diagnostic messages are not displayed after a few seconds, it is possible that initialization of the Hessian approximation is taking too long. In this case, consider restarting training and using the initial step size to speed up optimization.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2..

Algorithms

- Fitting a GPR model involves estimating the following model parameters from the data:
  - Covariance function $k\left(x_i, x_j \middle| \theta\right)$ parameterized in terms of kernel parameters in vector θ (see “Kernel (Covariance) Function Options” on page 6-6)
  - Noise variance, σ^2
  - Coefficient vector of fixed basis functions, β

The value of the `'KernelParameters'` name-value pair argument is a vector that consists of initial values for the signal standard deviation σ_f and the characteristic length scales σ_l. The `fitrgp` function uses these values to determine the kernel parameters. Similarly, the `'Sigma'` name-value pair argument contains the initial value for the noise standard deviation σ.

- During optimization, `fitrgp` creates a vector of unconstrained initial parameter values η_0 by using the initial values for the noise standard deviation and the kernel parameters.
- `fitrgp` analytically determines the explicit basis coefficients β, specified by the `'Beta'` name-value pair argument, from estimated values of θ and σ^2. Therefore, β does not appear in the η_0 vector when `fitrgp` initializes numerical optimization.

Note If you specify no estimation of parameters for the GPR model, `fitrgp` uses the value of the `'Beta'` name-value pair argument and other initial parameter values as the known GPR parameter values (see `Beta`). In all other cases, the value of the `'Beta'` argument is optimized analytically from the objective function.

- The quasi-Newton optimizer uses a trust-region method with a dense, symmetric rank-1-based (SR1), quasi-Newton approximation to the Hessian, while the LBFGS optimizer uses a standard line-search method with a limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton approximation to the Hessian. See Nocedal and Wright [6].
- If you set the `'InitialStepSize'` name-value pair argument to `'auto'`, `fitrgp` determines the initial step size, $\|s_0\|_\infty$, by using $\|s_0\|_\infty = 0.5\|\eta_0\|_\infty + 0.1$.

  s_0 is the initial step vector, and η_0 is the vector of unconstrained initial parameter values.
- During optimization, `fitrgp` uses the initial step size, $\|s_0\|_\infty$, as follows:

  If you use `'Optimizer','quasinewton'` with the initial step size, then the initial Hessian approximation is $\frac{\|g_0\|_\infty}{\|s_0\|_\infty} I$.

  If you use `'Optimizer','lbfgs'` with the initial step size, then the initial inverse-Hessian approximation is $\frac{\|s_0\|_\infty}{\|g_0\|_\infty} I$.

  g_0 is the initial gradient vector, and I is the identity matrix.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

[4] Rasmussen, C. E. and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press. Cambridge, Massachusetts, 2006.

[5] Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright. "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions." *SIAM Journal of Optimization*. Vol. 9, Number 1, 1998, pp. 112–147.

[6] Nocedal, J. and S. J. Wright. *Numerical Optimization*, Second Edition. Springer Series in Operations Research, Springer Verlag, 2006.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionGP` | `compact` | `predict`

Topics

“Gaussian Process Regression Models” on page 6-2
“Kernel (Covariance) Function Options” on page 6-6
“Introduction to Feature Selection” on page 15-49

Introduced in R2015b

fitrlinear

Fit linear regression model to high-dimensional data

Syntax

```
Mdl = fitrlinear(X,Y)

Mdl = fitrlinear(Tbl,ResponseVarName)
Mdl = fitrlinear(Tbl,formula)
Mdl = fitrlinear(Tbl,Y)

Mdl = fitrlinear(X,Y,Name,Value)
[Mdl,FitInfo] = fitrlinear(___)
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrlinear(___)
```

Description

`fitrlinear` efficiently trains linear regression models with high-dimensional, full or sparse predictor data. Available linear regression models include regularized support vector machines (SVM) and least-squares regression methods. `fitrlinear` minimizes the objective function using techniques that reduce computing time (e.g., stochastic gradient descent).

For reduced computation time on a high-dimensional data set that includes many predictor variables, train a linear regression model by using `fitrlinear`. For low- through medium-dimensional predictor data sets, see “Alternatives for Lower-Dimensional Data” on page 33-2208.

`Mdl = fitrlinear(X,Y)` returns a trained regression model object `Mdl` that contains the results of fitting a support vector machine regression model to the predictors `X` and response `Y`.

`Mdl = fitrlinear(Tbl,ResponseVarName)` returns a linear regression model using the predictor variables in the table `Tbl` and the response values in `Tbl.ResponseVarName`.

`Mdl = fitrlinear(Tbl,formula)` returns a linear regression model using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitrlinear(Tbl,Y)` returns a linear regression model using the predictor variables in the table `Tbl` and the response values in vector `Y`.

`Mdl = fitrlinear(X,Y,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify to cross-validate, implement least-squares regression, or specify the type of regularization. A good practice is to cross-validate using the `'Kfold'` name-value pair argument. The cross-validation results determine how well the model generalizes.

`[Mdl,FitInfo] = fitrlinear(___)` also returns optimization details using any of the previous syntaxes. You cannot request `FitInfo` for cross-validated models.

`[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrlinear(___)` also returns hyperparameter optimization details when you pass an `OptimizeHyperparameters` name-value pair.

Examples

Train Linear Regression Model

Train a linear regression model using SVM, dual SGD, and ridge regularization.

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model. By default, `fitrlinear` uses support vector machines with a ridge penalty, and optimizes using dual SGD for SVM. Determine how well the optimization algorithm fit the model to the data by extracting a fit summary.

```
[Mdl,FitInfo] = fitrlinear(X,Y)
```

```
Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0056
               Lambda: 1.0000e-04
              Learner: 'svm'


  Properties, Methods


FitInfo = struct with fields:
                    Lambda: 1.0000e-04
                 Objective: 0.2725
                 PassLimit: 10
                 NumPasses: 10
                BatchLimit: []
             NumIterations: 100000
              GradientNorm: NaN
         GradientTolerance: 0
      RelativeChangeInBeta: 0.4907
             BetaTolerance: 1.0000e-04
             DeltaGradient: 1.5816
    DeltaGradientTolerance: 0.1000
           TerminationCode: 0
         TerminationStatus: {'Iteration limit exceeded.'}
                     Alpha: [10000x1 double]
                   History: []
```

```
                FitTime: 0.1318
                 Solver: {'dual'}
```

`Mdl` is a `RegressionLinear` model. You can pass `Mdl` and the training or new data to `loss` to inspect the in-sample mean-squared error. Or, you can pass `Mdl` and new predictor data to `predict` to predict responses for new observations.

`FitInfo` is a structure array containing, among other things, the termination status (`TerminationStatus`) and how long the solver took to fit the model to the data (`FitTime`). It is good practice to use `FitInfo` to determine whether optimization-termination measurements are satisfactory. In this case, `fitrlinear` reached the maximum number of iterations. Because training time is fast, you can retrain the model, but increase the number of passes through the data. Or, try another solver, such as LBFGS.

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for a linear regression model that uses least squares, implement 5-fold cross-validation.

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-5} through 10^{-1}.

```
Lambda = logspace(-5,-1,15);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Optimize the objective function using SpaRSA.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','KFold',5,'Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');

numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `RegressionPartitionedLinear` model. Because `fitrlinear` implements 5-fold cross-validation, `CVMdl` contains 5 `RegressionLinear` models that the software trains on each fold.

Display the first trained linear regression model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x15 double]
                 Bias: [1x15 double]
               Lambda: [1x15 double]
              Learner: 'leastsquares'


  Properties, Methods
```

`Mdl1` is a `RegressionLinear` model object. `fitrlinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 15 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated MSE.

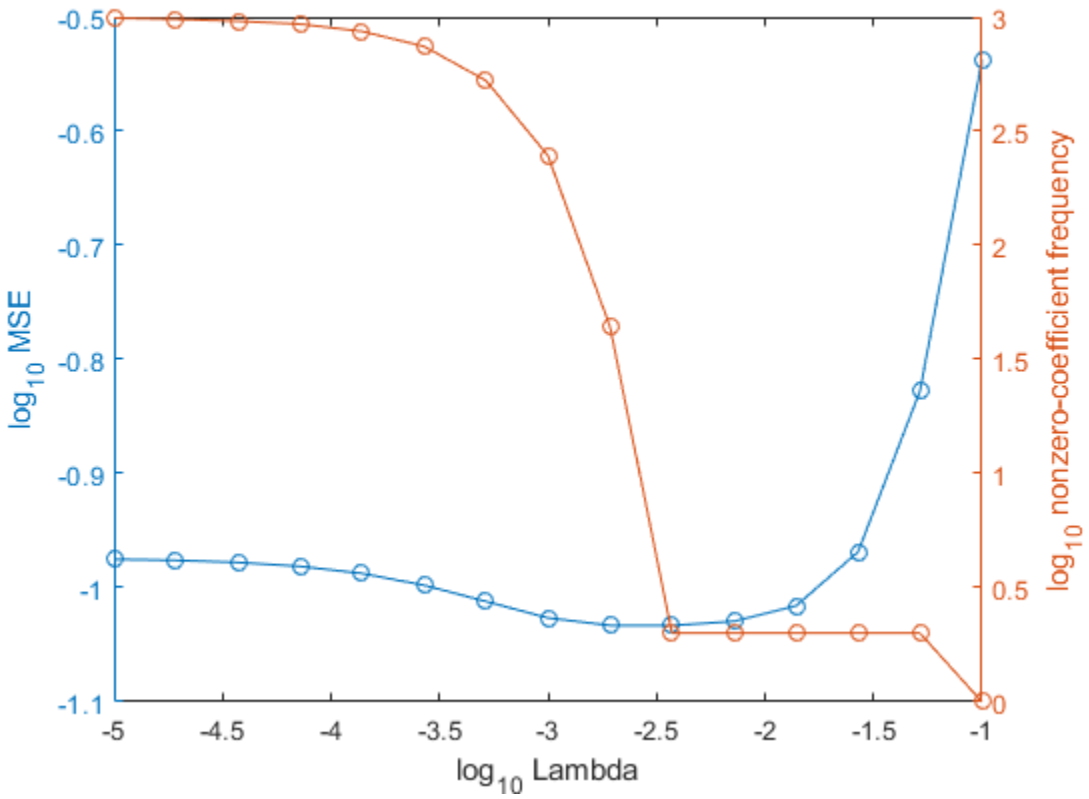
```
mse = kfoldLoss(CVMdl);
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. For each regularization strength, train a linear regression model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low MSE (for example, `Lambda(10)`).

```
idxFinal = 10;
```

Extract the model with corresponding to the minimal MSE.

```
MdlFinal = selectModels(Mdl,idxFinal)
```

```
MdlFinal =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0050
               Lambda: 0.0037
              Learner: 'leastsquares'


  Properties, Methods
```

```
idxNZCoeff = find(MdlFinal.Beta~=0)
```

```
idxNZCoeff = 2×1

   100
   200
```

```
EstCoeff = Mdl.Beta(idxNZCoeff)
```

```
EstCoeff = 2×1

    1.0051
    1.9965
```

`MdlFinal` is a `RegressionLinear` model with one regularization strength. The nonzero coefficients `EstCoeff` are close to the coefficients that simulated the data.

Optimize a Linear Regression

This example shows how to optimize hyperparameters automatically using `fitrlinear`. The example uses artificial (simulated) data for the model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

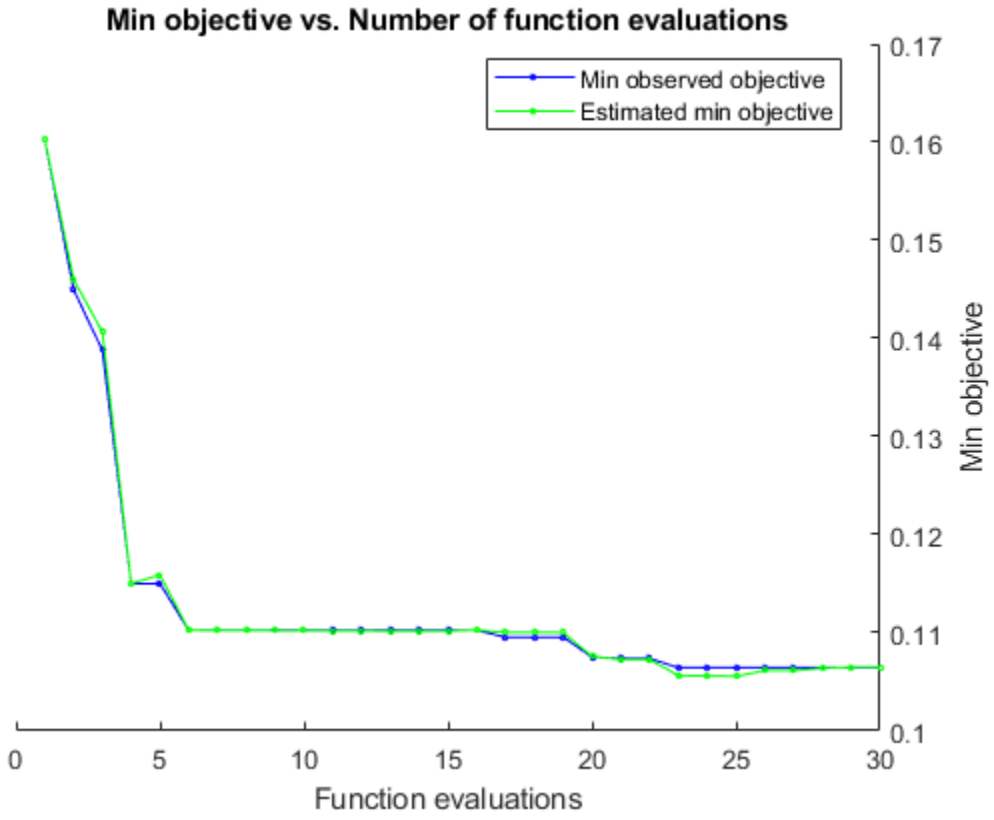
Find hyperparameters that minimize five-fold cross validation loss by using automatic hyperparameter optimization.

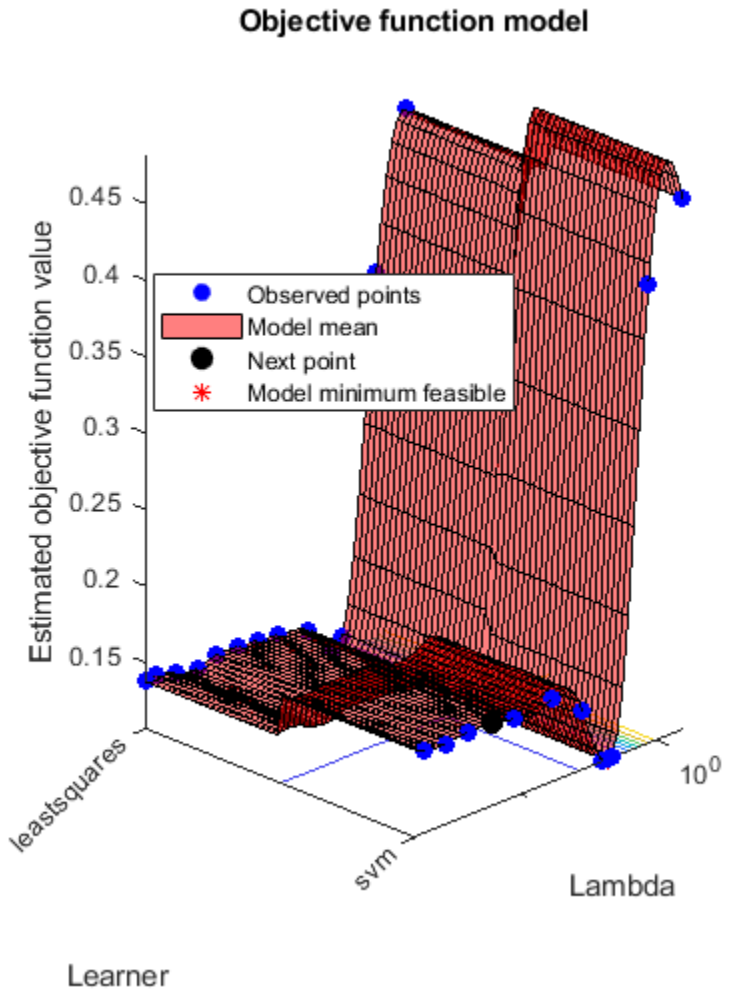
For reproducibility, use the `'expected-improvement-plus'` acquisition function.

```
hyperopts = struct('AcquisitionFunctionName','expected-improvement-plus');
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrlinear(X,Y,...
    'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',hyperopts)
```

```
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |       Lambda |      Lea
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.16029 |      1.1906 |     0.16029 |     0.16029 |   2.4206e-09 |
|    2 | Best   |     0.14496 |     0.82201 |     0.14496 |     0.14601 |     0.001807 |
|    3 | Best   |     0.13879 |       0.619 |     0.13879 |     0.14065 |   2.4681e-09 | leastsqu
|    4 | Best   |       0.115 |     0.64208 |       0.115 |     0.11501 |     0.021027 | leastsqu
|    5 | Accept |     0.44352 |     0.54092 |       0.115 |      0.1159 |       4.6795 | leastsqu
|    6 | Best   |     0.11025 |     0.45833 |     0.11025 |     0.11024 |     0.010671 | leastsqu
|    7 | Accept |     0.13222 |     0.43591 |     0.11025 |     0.11024 |   8.6067e-08 | leastsqu
|    8 | Accept |     0.13262 |     0.52531 |     0.11025 |     0.11023 |   8.5109e-05 | leastsqu
|    9 | Accept |     0.13543 |     0.66604 |     0.11025 |     0.11021 |   2.7562e-06 | leastsqu
|   10 | Accept |     0.15751 |     0.61219 |     0.11025 |     0.11022 |   5.0559e-06 |
|   11 | Accept |     0.40673 |     0.74857 |     0.11025 |      0.1102 |      0.52074 |
```

```
|   12 | Accept |     0.16057 |     0.81917 |     0.11025 |      0.1102 |   0.00014292 |
|   13 | Accept |     0.16105 |     0.72032 |     0.11025 |     0.11018 |   1.0079e-07 |
|   14 | Accept |     0.12763 |     0.52869 |     0.11025 |     0.11019 |    0.0012085 | leastsqu
|   15 | Accept |     0.13504 |     0.56049 |     0.11025 |     0.11019 |   1.3981e-08 | leastsqu
|   16 | Accept |     0.11041 |     0.54033 |     0.11025 |     0.11026 |    0.0093968 | leastsqu
|   17 | Best   |     0.10954 |     0.53996 |     0.10954 |     0.11003 |     0.010393 | leastsqu
|   18 | Accept |     0.10998 |     0.55792 |     0.10954 |     0.11002 |     0.010254 | leastsqu
|   19 | Accept |     0.45314 |     0.61113 |     0.10954 |     0.11001 |       9.9932 |
|   20 | Best   |     0.10753 |     0.92836 |     0.10753 |     0.10759 |     0.022576 |
|======|========|=============|=============|=============|=============|==============|=========
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |       Lambda |      Lea
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|======|========|=============|=============|=============|=============|==============|=========
|   21 | Best   |     0.10737 |     0.75253 |     0.10737 |     0.10728 |     0.010171 |
|   22 | Accept |     0.13448 |     0.53024 |     0.10737 |     0.10727 |   1.5344e-05 | leastsqu
|   23 | Best   |     0.10645 |     0.69952 |     0.10645 |     0.10565 |     0.015495 |
|   24 | Accept |     0.13598 |     0.50842 |     0.10645 |     0.10559 |   4.5984e-07 | leastsqu
|   25 | Accept |     0.15962 |     0.59813 |     0.10645 |     0.10556 |   1.4302e-08 |
|   26 | Accept |     0.10689 |     0.64169 |     0.10645 |     0.10616 |     0.015391 |
|   27 | Accept |     0.13748 |     0.49041 |     0.10645 |     0.10614 |    1.001e-09 | leastsqu
|   28 | Accept |     0.10692 |     0.61478 |     0.10645 |     0.10639 |     0.015761 |
|   29 | Accept |     0.10681 |     0.59589 |     0.10645 |     0.10649 |     0.015777 |
|   30 | Accept |     0.34314 |     0.53931 |     0.10645 |     0.10651 |      0.39671 | leastsqu
```






```
_____________________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 63.4088 seconds
Total objective function evaluation time: 19.0383

Best observed feasible point:
     Lambda     Learner
    ________    _______

    0.015495      svm

Observed objective function value = 0.10645
Estimated objective function value = 0.10651
Function evaluation time = 0.69952

Best estimated feasible point (according to models):
     Lambda     Learner
    ________    _______
```

```
    0.015777      svm

Estimated objective function value = 0.10651
Estimated function evaluation time = 0.71643

Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0018
               Lambda: 0.0158
              Learner: 'svm'


  Properties, Methods


FitInfo = struct with fields:
                    Lambda: 0.0158
                 Objective: 0.2309
                 PassLimit: 10
                 NumPasses: 10
                BatchLimit: []
             NumIterations: 99989
              GradientNorm: NaN
          GradientTolerance: 0
      RelativeChangeInBeta: 0.0641
             BetaTolerance: 1.0000e-04
             DeltaGradient: 1.1697
    DeltaGradientTolerance: 0.1000
           TerminationCode: 0
         TerminationStatus: {'Iteration limit exceeded.'}
                     Alpha: [10000x1 double]
                   History: []
                   FitTime: 0.0989
                    Solver: {'dual'}


HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                      ObjectiveFcn: @createObjFcn/inMemoryObjFcn
              VariableDescriptions: [3x1 optimizableVariable]
                           Options: [1x1 struct]
                      MinObjective: 0.1065
                   XAtMinObjective: [1x2 table]
             MinEstimatedObjective: 0.1065
          XAtMinEstimatedObjective: [1x2 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 63.4088
                         NextPoint: [1x2 table]
                            XTrace: [30x2 table]
                    ObjectiveTrace: [30x1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30x1 cell}
      ObjectiveEvaluationTimeTrace: [30x1 double]
                IterationTimeTrace: [30x1 double]
                        ErrorTrace: [30x1 double]
```

```
                FeasibilityTrace: [30x1 logical]
     FeasibilityProbabilityTrace: [30x1 double]
             IndexOfMinimumTrace: [30x1 double]
          ObjectiveMinimumTrace: [30x1 double]
EstimatedObjectiveMinimumTrace: [30x1 double]
```

This optimization technique is simpler than that shown in “Find Good Lasso Penalty Using Cross-Validation” on page 33-2178, but does not allow you to trade off model complexity and cross-validation loss.

Input Arguments

X — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix.

The length of `Y` and the number of observations in `X` must be equal.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time.

Data Types: `single` | `double`

Y — Response data

numeric vector

Response data, specified as an *n*-dimensional numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl` stores the response variable `Y` as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response (for example, `Y` or `ValidationData{2}`)
- At least one missing value in a predictor observation (for example, row in `X` or `ValidationData{1}`)
- `NaN` value or `0` weight (for example, value in `Weights` or `ValidationData{3}`)

For memory-usage economy, it is best practice to remove observations containing missing values from your training data manually before training.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `Mdl = fitrlinear(X,Y,'Learner','leastsquares','CrossVal','on','Regularization','lasso')` specifies to implement least-squares regression, implement 10-fold cross-validation, and specifies to include a lasso regularization term.

Linear Regression Options

Epsilon — Half the width of epsilon-insensitive band

`iqr(Y)/13.49` (default) | nonnegative scalar value

Half the width of the epsilon-insensitive band, specified as the comma-separated pair consisting of `'Epsilon'` and a nonnegative scalar value. `'Epsilon'` applies to SVM learners only.

The default `Epsilon` value is `iqr(Y)/13.49`, which is an estimate of standard deviation using the interquartile range of the response variable `Y`. If `iqr(Y)` is equal to zero, then the default `Epsilon` value is 0.1.

Example: `'Epsilon',0.3`

Data Types: `single` | `double`

Lambda — Regularization term strength

`'auto'` (default) | nonnegative scalar | vector of nonnegative values

Regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and `'auto'`, a nonnegative scalar, or a vector of nonnegative values.

- For `'auto'`, `Lambda` = $1/n$.
  - If you specify a cross-validation, name-value pair argument (e.g., `CrossVal`), then n is the number of in-fold observations.
  - Otherwise, n is the training sample size.
- For a vector of nonnegative values, `fitrlinear` sequentially optimizes the objective function for each distinct value in `Lambda` in ascending order.
  - If `Solver` is `'sgd'` or `'asgd'` and `Regularization` is `'lasso'`, `fitrlinear` does not use the previous coefficient estimates as a warm start on page 33-1767 for the next optimization iteration. Otherwise, `fitrlinear` uses warm starts.
  - If `Regularization` is `'lasso'`, then any coefficient estimate of 0 retains its value when `fitrlinear` optimizes using subsequent values in `Lambda`.
  - `fitrlinear` returns coefficient estimates for each specified regularization strength.

Example: `'Lambda',10.^(-(10:-2:2))`

Data Types: `char` | `string` | `double` | `single`

Learner — Linear regression model type

`'svm'` (default) | `'leastsquares'`

Linear regression model type, specified as the comma-separated pair consisting of `'Learner'` and `'svm'` or `'leastsquares'`.

In this table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.

- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Response range	Loss function
`'leastsquares'`	Linear regression via ordinary least squares	$y \in (-\infty,\infty)$	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$
`'svm'`	Support vector machine regression	Same as `'leastsquares'`	Epsilon-insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$

Example: `'Learner','leastsquares'`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

`Regularization` — Complexity penalty type
`'lasso'` | `'ridge'`

Complexity penalty type, specified as the comma-separated pair consisting of `'Regularization'` and `'lasso'` or `'ridge'`.

The software composes the objective function for minimization from the sum of the average loss function (see `Learner`) and the regularization term in this table.

Value	Description
`'lasso'`	Lasso (L1) penalty: $\lambda \sum_{j=1}^{p} \lvert\beta_j\rvert$
`'ridge'`	Ridge (L2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

To specify the regularization term strength, which is λ in the expressions, use `Lambda`.

The software excludes the bias term (β_0) from the regularization penalty.

If `Solver` is `'sparsa'`, then the default value of `Regularization` is `'lasso'`. Otherwise, the default is `'ridge'`.

Tip

- For predictor variable selection, specify `'lasso'`. For more on variable selection, see “Introduction to Feature Selection” on page 15-49.
- For optimization accuracy, specify `'ridge'`.

Example: `'Regularization','lasso'`

`Solver` — Objective function minimization technique
`'sgd'` | `'asgd'` | `'dual'` | `'bfgs'` | `'lbfgs'` | `'sparsa'` | string array | cell array of character vectors

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a character vector or string scalar, a string array, or a cell array of character vectors with values from this table.

Value	Description	Restrictions
`'sgd'`	Stochastic gradient descent (SGD) [5][3]	
`'asgd'`	Average stochastic gradient descent (ASGD) [8]	
`'dual'`	Dual SGD for SVM [2][7]	`Regularization` must be `'ridge'` and `Learner` must be `'svm'`.
`'bfgs'`	Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (BFGS) [4]	Inefficient if `X` is very high-dimensional.
`'lbfgs'`	Limited-memory BFGS (LBFGS) [4]	`Regularization` must be `'ridge'`.
`'sparsa'`	Sparse Reconstruction by Separable Approximation (SpaRSA) [6]	`Regularization` must be `'lasso'`.

If you specify:

- A ridge penalty (see `Regularization`) and `size(X,1) <= 100` (100 or fewer predictor variables), then the default solver is `'bfgs'`.
- An SVM regression model (see `Learner`), a ridge penalty, and `size(X,1) > 100` (more than 100 predictor variables), then the default solver is `'dual'`.
- A lasso penalty and `X` contains 100 or fewer predictor variables, then the default solver is `'sparsa'`.

Otherwise, the default solver is `'sgd'`.

If you specify a string array or cell array of solver names, then the software uses all solvers in the specified order for each `Lambda`.

For more details on which solver to choose, see “Tips” on page 33-2209.

Example: `'Solver',{'sgd','lbfgs'}`

Beta — Initial linear coefficient estimates
`zeros(p,1)` (default) | numeric vector | numeric matrix

Initial linear coefficient estimates (β), specified as the comma-separated pair consisting of `'Beta'` and a p-dimensional numeric vector or a p-by-L numeric matrix. p is the number of predictor variables in `X` and L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a p-dimensional vector, then the software optimizes the objective function L times using this process.
  1. The software optimizes using `Beta` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The software optimizes again using the resulting estimate from the previous optimization as a warm start on page 33-1767, and the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.
- If you specify a p-by-L matrix, then the software optimizes the objective function L times. At iteration j, the software uses `Beta(:,j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.

If you set `'Solver','dual'`, then the software ignores `Beta`.

Data Types: `single` | `double`

Bias — Initial intercept estimate
numeric scalar | numeric vector

Initial intercept estimate (b), specified as the comma-separated pair consisting of `'Bias'` and a numeric scalar or an L-dimensional numeric vector. L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a scalar, then the software optimizes the objective function L times using this process.
  1. The software optimizes using `Bias` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The uses the resulting estimate as a warm start on page 33-2208 to the next optimization iteration, and uses the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.
- If you specify an L-dimensional vector, then the software optimizes the objective function L times. At iteration j, the software uses `Bias(j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.
- By default:
  - If `Learner` is `'leastsquares'`, then `Bias` is the weighted average of `Y` for training or, for cross-validation, in-fold responses.
  - If `Learner` is `'svm'`, then `Bias` is the weighted median of `Y` for all training or, for cross-validation, in-fold observations that are greater than `Epsilon`.

Data Types: `single` | `double`

FitBias — Linear model intercept inclusion flag
`true` (default) | `false`

Linear model intercept inclusion flag, specified as the comma-separated pair consisting of `'FitBias'` and `true` or `false`.

Value	Description
`true`	The software includes the bias term b in the linear model, and then estimates it.
`false`	The software sets $b = 0$ during estimation.

Example: `'FitBias',false`

Data Types: `logical`

`PostFitBias` — Flag to fit linear model intercept after optimization

`false` (default) | `true`

Flag to fit the linear model intercept after optimization, specified as the comma-separated pair consisting of `'PostFitBias'` and `true` or `false`.

Value	Description
`false`	The software estimates the bias term b and the coefficients β during optimization.
`true`	To estimate b, the software: **1** Estimates β and b using the model. **2** Computes residuals. **3** Refits b. For least squares, b is the weighted average of the residuals. For SVM regression, b is the weighted median between all residuals with magnitude greater than `Epsilon`.

If you specify `true`, then `FitBias` must be true.

Example: `'PostFitBias',true`

Data Types: `logical`

`Verbose` — Verbosity level

`0` (default) | nonnegative integer

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and a nonnegative integer. `Verbose` controls the amount of diagnostic information `fitrlinear` displays at the command line.

Value	Description
`0`	`fitrlinear` does not display diagnostic information.

Value	Description
1	`fitrlinear` periodically displays and stores the value of the objective function, gradient magnitude, and other diagnostic information. `FitInfo.History` contains the diagnostic information.
Any other positive integer	`fitrlinear` displays and stores diagnostic information at each optimization iteration. `FitInfo.History` contains the diagnostic information.

Example: `'Verbose',1`

Data Types: `double` | `single`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size
positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration, the software estimates the subgradient using `BatchSize` observations from the training data.

- If `X` is a numeric matrix, then the default value is `10`.
- If `X` is a sparse matrix, then the default value is `max([10,ceil(sqrt(ff))])`, where `ff = numel(X)/nnz(X)` (the fullness factor of `X`).

Example: `'BatchSize',100`

Data Types: `single` | `double`

`LearnRate` — Learning rate
positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and a positive scalar. `LearnRate` specifies how many steps to take per iteration. At each iteration, the gradient specifies the direction and magnitude of each step.

- If `Regularization` is `'ridge'`, then `LearnRate` specifies the initial learning rate γ_0. The software determines the learning rate for iteration t, γ_t, using

  $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$

  - λ is the value of `Lambda`.
  - If `Solver` is `'sgd'`, $c = 1$.
  - If `Solver` is `'asgd'`:
    - $c = 2/3$ if `Learner` is `'leastsquares'`
    - $c = 3/4$ if `Learner` is `'svm'` [8]
- If `Regularization` is `'lasso'`, then, for all iterations, `LearnRate` is constant.

By default, `LearnRate` is `1/sqrt(1+max((sum(X.^2,obsDim))))`, where `obsDim` is `1` if the observations compose the columns of `X`, and `2` otherwise.

Example: `'LearnRate',0.01`

Data Types: `single` | `double`

`OptimizeLearnRate` — Flag to decrease learning rate
`true` (default) | `false`

Flag to decrease the learning rate when the software detects divergence (that is, over-stepping the minimum), specified as the comma-separated pair consisting of `'OptimizeLearnRate'` and `true` or `false`.

If `OptimizeLearnRate` is `'true'`, then:

1. For the few optimization iterations, the software starts optimization using `LearnRate` as the learning rate.
2. If the value of the objective function increases, then the software restarts and uses half of the current value of the learning rate.
3. The software iterates step 2 until the objective function decreases.

Example: `'OptimizeLearnRate',true`

Data Types: `logical`

`TruncationPeriod` — Number of mini-batches between lasso truncation runs
`10` (default) | positive integer

Number of mini-batches between lasso truncation runs, specified as the comma-separated pair consisting of `'TruncationPeriod'` and a positive integer.

After a truncation run, the software applies a soft threshold to the linear coefficients. That is, after processing k = `TruncationPeriod` mini-batches, the software truncates the estimated coefficient j using

$$\widehat{\beta}_j^* = \begin{cases} \widehat{\beta}_j - u_t & \text{if } \widehat{\beta}_j > u_t, \\ 0 & \text{if } \left|\widehat{\beta}_j\right| \le u_t, \\ \widehat{\beta}_j + u_t & \text{if } \widehat{\beta}_j < -u_t. \end{cases}$$

- For SGD, $\widehat{\beta}_j$ is the estimate of coefficient j after processing k mini-batches. $u_t = k\gamma_t\lambda$. γ_t is the learning rate at iteration t. λ is the value of `Lambda`.
- For ASGD, $\widehat{\beta}_j$ is the averaged estimate coefficient j after processing k mini-batches, $u_t = k\lambda$.

If `Regularization` is `'ridge'`, then the software ignores `TruncationPeriod`.

Example: `'TruncationPeriod',100`

Data Types: `single` | `double`

Other Regression Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table. The descriptions assume that the predictor data has observations in rows and predictors in columns.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrlinear` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrlinear` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrlinear` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitrlinear` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitrlinear` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitrlinear` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the predictor order in `X`. Assuming that `X` has the default orientation, with observations in rows and predictors in columns, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitrlinear` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
positive numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a positive numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl`

with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

`fitrlinear` normalizes the weights to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Cross-validation flag
`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, or `KFold`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` partition object as created by `cvpartition`. The partition object specifies the type of cross-validation, and also the indexing for training and validation sets.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of data used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range (0,1). If you specify `'Holdout',`*p*, then the software:

1. Randomly reserves *p*`*100`% of the data as validation data, and trains the model using the rest of the data
2. Stores the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

KFold — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated classifier, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1. If you specify, e.g., `'KFold',k`, then the software:

1. Randomly partitions the data into k sets
2. For each set, reserves the set as validation data, and trains the model using the other k – 1 sets
3. Stores the k compact, trained models in the cells of a k-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four options only: `'CVPartition'`, `'Holdout'`, or `'KFold'`.

Example: `'KFold',8`

Data Types: `single` | `double`

SGD and ASGD Convergence Controls

BatchLimit — Maximal number of batches

positive integer

Maximal number of batches to process, specified as the comma-separated pair consisting of `'BatchLimit'` and a positive integer. When the software processes `BatchLimit` batches, it terminates optimization.

- By default:
  - The software passes through the data `PassLimit` times.
  - If you specify multiple solvers, and use (A)SGD to get an initial approximation for the next solver, then the default value is `ceil(1e6/BatchSize)`. `BatchSize` is the value of the `'BatchSize'` name-value pair argument.
- If you specify `'BatchLimit'` and `'PassLimit'`, then the software chooses the argument that results in processing the fewest observations.
- If you specify `'BatchLimit'` but not `'PassLimit'`, then the software processes enough batches to complete up to one entire pass through the data.

Example: `'BatchLimit',100`

Data Types: `single` | `double`

BetaTolerance — Relative tolerance on linear coefficients and bias term

`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: 'BetaTolerance',1e-6

Data Types: single | double

NumCheckConvergence — Number of batches to process before next convergence check
positive integer

Number of batches to process before next convergence check, specified as the comma-separated pair consisting of 'NumCheckConvergence' and a positive integer.

To specify the batch size, see BatchSize.

The software checks for convergence about 10 times per pass through the entire data set by default.

Example: 'NumCheckConvergence',100

Data Types: single | double

PassLimit — Maximal number of passes
1 (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of 'PassLimit' and a positive integer.

fitrlinear processes all observations when it completes one pass through the data.

When fitrlinear passes through the data PassLimit times, it terminates optimization.

If you specify 'BatchLimit' and PassLimit, then fitrlinear chooses the argument that results in processing the fewest observations. For more details, see “Algorithms” on page 33-2210.

Example: 'PassLimit',5

Data Types: single | double

ValidationData — Validation data for optimization convergence detection
cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of 'ValidationData' and a cell array or table.

During optimization, the software periodically estimates the loss of ValidationData. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-2210. To optimize hyperparameters using cross-validation, see cross-validation options such as CrossVal.

You can specify ValidationData as a table if you use a table Tbl of predictor data that contains the response variable. In this case, ValidationData must contain the same predictors and response contained in Tbl. The software does not apply weights to observations, even if Tbl contains a vector of weights. To specify weights, you must specify ValidationData as a cell array.

If you specify ValidationData as a cell array, then it must have the following format:

- ValidationData{1} must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix X, then ValidationData{1} must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as X. The predictor variables in the training data X and ValidationData{1} must correspond. Similarly, if you use a predictor table

`Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.

- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of responses, then it must have the same number of elements as the number of observations in `ValidationData{1}`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

`GradientTolerance` — Absolute gradient tolerance

`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar. `GradientTolerance` applies to these values of `Solver`: `'bfgs'`, `'lbfgs'`, and `'sparsa'`.

Let $\nabla\mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla\mathscr{L}_t\|_\infty = \max|\nabla\mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when `fitrlinear` satisfies either stopping criterion.

If `fitrlinear` converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, `fitrlinear` uses the next solver specified in `Solver`.

Example: `'GradientTolerance',eps`

Data Types: `single` | `double`

`IterationLimit` — Maximal number of optimization iterations

`1000` (default) | positive integer

Maximal number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer. `IterationLimit` applies to these values of `Solver`: `'bfgs'`, `'lbfgs'`, and `'sparsa'`.

Example: `'IterationLimit',1e7`

Data Types: `single` | `double`

Dual SGD Optimization Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term

`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `DeltaGradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`DeltaGradientTolerance` — Gradient-difference tolerance

`0.1` (default) | nonnegative scalar

Gradient-difference tolerance between upper and lower pool Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862 violators, specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar. `DeltaGradientTolerance` applies to the `'dual'` value of `Solver` only.

- If the magnitude of the KKT violators is less than `DeltaGradientTolerance`, then `fitrlinear` terminates optimization.
- If `fitrlinear` converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, `fitrlinear` uses the next solver specified in `Solver`.

Example: `'DeltaGapTolerance',1e-2`

Data Types: `double` | `single`

`NumCheckConvergence` — Number of passes through entire data set to process before next convergence check

`5` (default) | positive integer

Number of passes through entire data set to process before next convergence check, specified as the comma-separated pair consisting of `'NumCheckConvergence'` and a positive integer.

Example: `'NumCheckConvergence',100`

Data Types: `single` | `double`

`PassLimit` — Maximal number of passes

`10` (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer.

When the software completes one pass through the data, it has processed all observations.

When the software passes through the data `PassLimit` times, it terminates optimization.

Example: `'PassLimit',5`

Data Types: `single` | `double`

`ValidationData` — Validation data for optimization convergence detection

cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of `'ValidationData'` and a cell array or table.

During optimization, the software periodically estimates the loss of `ValidationData`. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-2210. To optimize hyperparameters using cross-validation, see cross-validation options such as `CrossVal`.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of responses, then it must have the same number of elements as the number of observations in `ValidationData{1}`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

BFGS, LBFGS, and SpaRSA Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term

`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in the software, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. That is, at each iteration, the software composes the Hessian using statistics from the latest `HessianHistorySize` iterations.

The software does not support `'HessianHistorySize'` for SpaRSA.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`IterationLimit` — Maximal number of optimization iterations
`1000` (default) | positive integer

Maximal number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer. `IterationLimit` applies to these values of `Solver`: `'bfgs'`, `'lbfgs'`, and `'sparsa'`.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

`ValidationData` — Validation data for optimization convergence detection

cell array | table

Validation data for optimization convergence detection, specified as the comma-separated pair consisting of `'ValidationData'` and a cell array or table.

During optimization, the software periodically estimates the loss of `ValidationData`. If the validation-data loss increases, then the software terminates optimization. For more details, see “Algorithms” on page 33-2210. To optimize hyperparameters using cross-validation, see cross-validation options such as `CrossVal`.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* full or sparse matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X` and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.
- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of responses, then it must have the same number of elements as the number of observations in `ValidationData{1}`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, specify a value larger than 0 for `Verbose`.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

By default, the software does not detect convergence by monitoring validation-data loss.

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize

`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- 'none' — Do not optimize.
- 'auto' — Use {'Lambda','Learner'}.
- 'all' — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of optimizableVariable objects, typically the output of hyperparameters.

The optimization attempts to minimize the cross-validation loss (error) for fitrlinear by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the HyperparameterOptimizationOptions name-value pair.

Note 'OptimizeHyperparameters' values override any values you set using other name-value pair arguments. For example, setting 'OptimizeHyperparameters' to 'auto' causes the 'auto' values to apply.

The eligible parameters for fitrlinear are:

- Lambda — fitrlinear searches among positive values, by default log-scaled in the range [1e-5/NumObservations,1e5/NumObservations].
- Learner — fitrlinear searches among 'svm' and 'leastsquares'.
- Regularization — fitrlinear searches among 'ridge' and 'lasso'.

Set nondefault parameters by passing a vector of optimizableVariable objects that have nondefault values. For example,

```
load carsmall
params = hyperparameters('fitrlinear',[Horsepower,Weight],MPG);
params(1).Range = [1e-3,2e4];
```

Pass params as the value of OptimizeHyperparameters.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the Verbose field of the 'HyperparameterOptimizationOptions' name-value pair argument. To control the plots, set the ShowPlots field of the 'HyperparameterOptimizationOptions' name-value pair argument.

For an example, see “Optimize a Linear Regression” on page 33-2181.

Example: 'OptimizeHyperparameters','auto'

HyperparameterOptimizationOptions — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of 'HyperparameterOptimizationOptions' and a structure. This argument modifies the effect of the OptimizeHyperparameters name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained linear regression model

`RegressionLinear` model object | `RegressionPartitionedLinear` cross-validated model object

Trained linear regression model, returned as a `RegressionLinear` model object or `RegressionPartitionedLinear` cross-validated model object.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `RegressionPartitionedLinear` cross-validated model object. Otherwise, `Mdl` is a `RegressionLinear` model object.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.Beta` in the Command Window to display the vector or matrix of estimated coefficients.

Note Unlike other regression models, and for economical memory usage, `RegressionLinear` and `RegressionPartitionedLinear` model objects do not store the training data or optimization details (for example, convergence history).

`FitInfo` — Optimization details

structure array

Optimization details, returned as a structure array.

Fields specify final values or name-value pair argument specifications, for example, `Objective` is the value of the objective function when optimization terminates. Rows of multidimensional fields correspond to values of `Lambda` and columns correspond to values of `Solver`.

This table describes some notable fields.

Field	Description
`TerminationStatus`	• Reason for optimization termination • Corresponds to a value in `TerminationCode`
`FitTime`	Elapsed, wall-clock time in seconds
`History`	A structure array of optimization information for each iteration. The field `Solver` stores solver types using integer coding. **Integer** / **Solver** 1 / SGD 2 / ASGD 3 / Dual SGD for SVM 4 / LBFGS 5 / BFGS 6 / SpaRSA

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.History.Objective`.

It is good practice to examine `FitInfo` to assess whether convergence is satisfactory.

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table of hyperparameters and associated values

Cross-validation optimization of hyperparameters, returned as a `BayesianOptimization` object or a table of hyperparameters and associated values. The output is nonempty when the value of `'OptimizeHyperparameters'` is not `'none'`. The output value depends on the `Optimizer` field value of the `'HyperparameterOptimizationOptions'` name-value pair argument:

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Note If `Learner` is `'leastsquares'`, then the loss term in the objective function is half of the MSE. `loss` returns the MSE by default. Therefore, if you use `loss` to check the resubstitution, or training, error then there is a discrepancy between the MSE returned by `loss` and optimization results in `FitInfo` or returned to the command line by setting a positive verbosity level using `Verbose`.

More About

Warm Start

A warm start is initial estimates of the beta coefficients and bias term supplied to an optimization routine for quicker convergence.

Alternatives for Lower-Dimensional Data

High-dimensional linear classification and regression models minimize objective functions relatively quickly, but at the cost of some accuracy, the numeric-only predictor variables restriction, and the model must be linear with respect to the parameters. If your predictor data set is low- through medium-dimensional, or contains heterogeneous variables, then you should use the appropriate classification or regression fitting function. To help you decide which fitting function is appropriate for your low-dimensional data set, use this table.

Model to Fit	Function	Notable Algorithmic Differences
SVM	• Binary classification: `fitcsvm` • Multiclass classification: `fitcecoc` • Regression: `fitrsvm`	• Computes the Gram matrix of the predictor variables, which is convenient for nonlinear kernel transformations. • Solves dual problem using SMO, ISDA, or *L*1 minimization via quadratic programming using `quadprog`.

Model to Fit	Function	Notable Algorithmic Differences
Linear regression	• Least-squares without regularization: `fitlm` • Regularized least-squares using a lasso penalty: `lasso` • Ridge regression: `ridge` or `lasso`	• `lasso` implements cyclic coordinate descent.
Logistic regression	• Logistic regression without regularization: `fitglm`. • Regularized logistic regression using a lasso penalty: `lassoglm`	• `fitglm` implements iteratively reweighted least squares. • `lassoglm` implements cyclic coordinate descent.

Tips

- It is a best practice to orient your predictor matrix so that observations correspond to columns and to specify `'ObservationsIn','columns'`. As a result, you can experience a significant reduction in optimization-execution time.
- For better optimization accuracy when you have high-dimensional predictor data and the `Regularization` value is `'ridge'`, set any of these options for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'dual'` if `Learner` is `'svm'`
  - `{'sgd','lbfgs'}`
  - `{'asgd','lbfgs'}`
  - `{'dual','lbfgs'}` if `Learner` is `'svm'`

  Other options can result in poor optimization accuracy.
- For better optimization accuracy when you have moderate- to low-dimensional predictor data and the `Regularization` value is `'ridge'`, set `Solver` to `'bfgs'`.
- If `Regularization` is `'lasso'`, set any of these options for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'sparsa'`
  - `{'sgd','sparsa'}`
  - `{'asgd','sparsa'}`
- When choosing between SGD and ASGD, consider that:
  - SGD takes less time per iteration, but requires more iterations to converge.
  - ASGD requires fewer iterations to converge, but takes more time per iteration.
- If your predictor data has few observations but many predictor variables, then:

  - Specify `'PostFitBias',true`.
  - For SGD or ASGD solvers, set `PassLimit` to a positive integer that is greater than 1, for example, 5 or 10. This setting often results in better accuracy.
- For SGD and ASGD solvers, `BatchSize` affects the rate of convergence.
  - If `BatchSize` is too small, then `fitrlinear` achieves the minimum in many iterations, but computes the gradient per iteration quickly.
  - If `BatchSize` is too large, then `fitrlinear` achieves the minimum in fewer iterations, but computes the gradient per iteration slowly.
- Large learning rates (see `LearnRate`) speed up convergence to the minimum, but can lead to divergence (that is, over-stepping the minimum). Small learning rates ensure convergence to the minimum, but can lead to slow termination.
- When using lasso penalties, experiment with various values of `TruncationPeriod`. For example, set `TruncationPeriod` to `1`, `10`, and then `100`.
- For efficiency, `fitrlinear` does not standardize predictor data. To standardize `X`, enter

  ```
  X = bsxfun(@rdivide,bsxfun(@minus,X,mean(X,2)),std(X,0,2));
  ```

  The code requires that you orient the predictors and observations as the rows and columns of `X`, respectively. Also, for memory-usage economy, the code replaces the original predictor data the standardized data.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- If you specify `ValidationData`, then, during objective-function optimization:
  - `fitrlinear` estimates the validation loss of `ValidationData` periodically using the current model, and tracks the minimal estimate.
  - When `fitrlinear` estimates a validation loss, it compares the estimate to the minimal estimate.
  - When subsequent, validation loss estimates exceed the minimal estimate five times, `fitrlinear` terminates optimization.
- If you specify `ValidationData` and to implement a cross-validation routine (`CrossVal`, `CVPartition`, `Holdout`, or `KFold`), then:

  1. `fitrlinear` randomly partitions `X` and `Y` (or `Tbl`) according to the cross-validation routine that you choose.
  2. `fitrlinear` trains the model using the training-data partition. During objective-function optimization, `fitrlinear` uses `ValidationData` as another possible way to terminate optimization (for details, see the previous bullet).
  3. Once `fitrlinear` satisfies a stopping criterion, it constructs a trained model based on the optimized linear coefficients and intercept.
     a. If you implement *k*-fold cross-validation, and `fitrlinear` has not exhausted all training-set folds, then `fitrlinear` returns to Step 2 to train using the next training-set fold.

   b Otherwise, `fitrlinear` terminates training, and then returns the cross-validated model.

4 You can determine the quality of the cross-validated model. For example:

   - To determine the validation loss using the holdout or out-of-fold data from step 1, pass the cross-validated model to `kfoldLoss`.
   - To predict observations on the holdout or out-of-fold data from step 1, pass the cross-validated model to `kfoldPredict`.

References

[1] Ho, C. H. and C. J. Lin. “Large-Scale Linear Support Vector Regression.” *Journal of Machine Learning Research*, Vol. 13, 2012, pp. 3323–3348.

[2] Hsieh, C. J., K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. “A Dual Coordinate Descent Method for Large-Scale Linear SVM.” *Proceedings of the 25th International Conference on Machine Learning, ICML ’08*, 2001, pp. 408–415.

[3] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[4] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

[5] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[6] Wright, S. J., R. D. Nowak, and M. A. T. Figueiredo. “Sparse Reconstruction by Separable Approximation.” *Trans. Sig. Proc.*, Vol. 57, No 7, 2009, pp. 2479–2493.

[7] Xiao, Lin. “Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization.” *J. Mach. Learn. Res.*, Vol. 11, 2010, pp. 2543–2596.

[8] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `fitrlinear` does not support tall `table` data.
- Some name-value pair arguments have different defaults and values compared to the in-memory `fitrlinear` function. Supported name-value pair arguments, and any differences, are:
  - `'Epsilon'`
  - `'ObservationsIn'` — Supports only `'rows'`.
  - `'Lambda'` — Can be `'auto'` (default) or a scalar.
  - `'Learner'`
  - `'Regularization'` — Supports only `'ridge'`.

- `'Solver'` — Supports only `'lbfgs'`.
- `'Verbose'` — Default value is `1`
- `'Beta'`
- `'Bias'`
- `'FitBias'` — Supports only `true`.
- `'Weights'` — Value must be a tall array.
- `'HessianHistorySize'`
- `'BetaTolerance'` — Default value is relaxed to `1e-3`.
- `'GradientTolerance'` — Default value is relaxed to `1e-3`.
- `'IterationLimit'` — Default value is relaxed to `20`.
- `'OptimizeHyperparameters'` — Value of `'Regularization'` parameter must be `'ridge'`.
- `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.

- For tall arrays `fitrlinear` implements LBFGS by distributing the calculation of the loss and the gradient among different parts of the tall array at each iteration. Other solvers are not available for tall arrays.

  When initial values for `Beta` and `Bias` are not given, `fitrlinear` first refines the initial estimates of the parameters by fitting the model locally to parts of the data and combining the coefficients by averaging.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionLinear` | `RegressionPartitionedLinear` | `fitclinear` | `fitlm` | `fitrsvm` | `kfoldLoss` | `kfoldPredict` | `lasso` | `predict` | `ridge`

Introduced in R2016a

fitrm

Fit repeated measures model

Syntax

```
rm = fitrm(t,modelspec)
rm = fitrm(t,modelspec,Name,Value)
```

Description

`rm = fitrm(t,modelspec)` returns a repeated measures model, specified by `modelspec`, fitted to the variables in the table or dataset array `t`.

`rm = fitrm(t,modelspec,Name,Value)` returns a repeated measures model, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the hypothesis for the within-subject factors.

Examples

Fit a Repeated Measures Model

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = table([1 2 3 4]','VariableNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas)
```

```
rm =
  RepeatedMeasuresModel with properties:

   Between Subjects:
         BetweenDesign: [150x5 table]
         ResponseNames: {'meas1'  'meas2'  'meas3'  'meas4'}
    BetweenFactorNames: {'species'}
          BetweenModel: '1 + species'

   Within Subjects:
```

```
          WithinDesign: [4x1 table]
     WithinFactorNames: {'Measurements'}
           WithinModel: 'separatemeans'

   Estimates:
          Coefficients: [3x4 table]
            Covariance: [4x4 table]
```

Display the coefficients.

```
rm.Coefficients
```

```
ans=3×4 table
                          meas1       meas2      meas3      meas4
                        ________    ________    ______    ________

    (Intercept)           5.8433      3.0573     3.758      1.1993
    species_setosa       -0.83733     0.37067   -2.296     -0.95333
    species_versicolor    0.092667   -0.28733    0.502      0.12667
```

`fitrm` uses the `'effects'` contrasts which means that the coefficients sum to 0. The `rm.DesignMatrix` has one column of 1s for the intercept, and two other columns `species_setosa` and `species_versicolor`, which are as follows:

$$\text{species\_setosa} = \begin{cases} 1 & \text{if setosa} \\ 0 & \text{if versicolor} \\ -1 & \text{if virginica} \end{cases} \quad \text{and} \quad \text{species\_versicolor} = \begin{cases} 0 & \text{if setosa} \\ 1 & \text{if versicolor} \\ -1 & \text{if virginica} \end{cases}$$

Display the covariance matrix.

```
rm.Covariance
```

```
ans=4×4 table
              meas1       meas2       meas3       meas4
            ________    ________    ________    ________

    meas1     0.26501    0.092721     0.16751    0.038401
    meas2    0.092721     0.11539    0.055244     0.03271
    meas3     0.16751    0.055244     0.18519    0.042665
    meas4    0.038401     0.03271    0.042665    0.041882
```

Specify the Within-Subject Hypothesis

Load the sample data.

```
load('longitudinalData.mat');
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (time = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual, and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Define a variable that stores gender information.

```
Gender = ['F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']';
```

Store the data in a proper table array format to conduct repeated measures analysis.

```
t = table(Gender,Y(:,1),Y(:,2),Y(:,3),Y(:,4),Y(:,5),...
'VariableNames',{'Gender','t0','t2','t4','t6','t8'});
```

Define the within-subjects variable.

```
Time = [0 2 4 6 8]';
```

Fit a repeated measures model, where blood levels are the responses and gender is the predictor variable. Also define the hypothesis for within-subject factors.

```
rm = fitrm(t,'t0-t8 ~ Gender','WithinDesign',Time,'WithinModel','orthogonalcontrasts')
```

```
rm =
  RepeatedMeasuresModel with properties:

   Between Subjects:
         BetweenDesign: [16x6 table]
         ResponseNames: {'t0'  't2'  't4'  't6'  't8'}
    BetweenFactorNames: {'Gender'}
          BetweenModel: '1 + Gender'

   Within Subjects:
          WithinDesign: [5x1 table]
     WithinFactorNames: {'Time'}
           WithinModel: 'orthogonalcontrasts'

   Estimates:
          Coefficients: [2x5 table]
            Covariance: [5x5 table]
```

Fit a Model with Covariates

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the eight repeated measurements, `y1` through `y8`, as responses and the between-subject factors `Group`, `Gender`, `IQ`, and `Age`. `IQ` and `Age` as continuous variables. The table `within` includes the within-subject factors `w1` and `w2`.

Fit a repeated measures model, where age, IQ, group, and gender are the predictor variables, and the model includes the interaction effect of group and gender. Also define the within-subject factors.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender+Age+IQ','WithinDesign',within)
```

```
rm =
  RepeatedMeasuresModel with properties:

   Between Subjects:
```

```
        BetweenDesign: [30x12 table]
        ResponseNames: {'y1'  'y2'  'y3'  'y4'  'y5'  'y6'  'y7'  'y8'}
    BetweenFactorNames: {'Age'  'IQ'  'Group'  'Gender'}
         BetweenModel: '1 + Age + IQ + Group*Gender'

  Within Subjects:
         WithinDesign: [8x2 table]
     WithinFactorNames: {'w1'  'w2'}
          WithinModel: 'separatemeans'

  Estimates:
         Coefficients: [8x8 table]
           Covariance: [8x8 table]
```

Display the coefficients.

```
rm.Coefficients
```

```
ans=8×8 table
                             y1          y2          y3          y4          y5          y6
                          _______    ________    _______    ________    _________    _______

    (Intercept)            141.38      195.25      9.8663     -49.154      157.77     0.23762
    Age                   0.32042     -4.7672     -1.2748      0.6216     -1.0621     0.89927
    IQ                    -1.2671     -1.1653     0.05862      0.4288     -1.4518    -0.25501
    Group_A               -1.2195     -9.6186      22.532      15.303      12.602      12.886
    Group_B                2.5186       1.417     -2.2501     0.50181      8.0907      3.1957
    Gender_Female          5.3957     -3.9719      8.5225      9.3403      6.0909       1.642
    Group_A:Gender_Female  4.1046      10.064     -7.3053     -3.3085      4.6751      2.4907
    Group_B:Gender_Female -0.48486     -2.9202      1.1222     0.69715    -0.065945    0.079468
```

The display shows the coefficients for fitting the repeated measures as a function of the terms in the between-subjects model.

Input Arguments

`t` — Input data

table

Input data, which includes the values of the response variables and the between-subject factors to use as predictors in the repeated measures model, specified as a table.

The variable names in `t` must be valid MATLAB identifiers. You can verify the variable names by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `table`

`modelspec` — Formula for model specification

character vector or string scalar of the form `'y1-yk ~ terms'`

Formula for model specification, specified as a character vector or string scalar of the form `'y1-yk ~ terms'`. The responses and terms are specified using Wilkinson notation on page 33-2218. `fitrm`

treats the variables used in model terms as categorical if they are categorical (nominal or ordinal), logical, character arrays, string arrays, or cell arrays of character vectors.

For example, if you have four repeated measures as responses and the factors `x1`, `x2`, and `x3` as the predictor variables, then you can define a repeated measures model as follows.

Example: `'y1-y4 ~ x1 + x2 * x3'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'WithinDesign','W','WithinModel','w1+w2'` specifies the matrix `w` as the design matrix for within-subject factors, and the model for within-subject factors `w1` and `w2` is `'w1+w2'`.

`WithinDesign` — Design for within-subject factors

numeric vector of length *r* (default) | *r*-by-*k* numeric matrix | *r*-by-*k* table

Design for within-subject factors, specified as the comma-separated pair consisting of `'WithinDesign'` and one of the following:

- Numeric vector of length *r*, where *r* is the number of repeated measures.

  In this case, `fitrm` treats the values in the vector as continuous, and these are typically time values.
- *r*-by-*k* numeric matrix of the values of the *k* within-subject factors, w_1, w_2, ..., w_k.

  In this case,`fitrm` treats all *k* variables as continuous.
- *r*-by-*k* table that contains the values of the *k* within-subject factors.

  In this case, `fitrm` treats all numeric variables as continuous, and all categorical variables as categorical.

For example, if the table `weeks` contains the values of the within-subject factors, then you can define the design table as follows.

Example: `'WithinDesign',weeks`

Data Types: `single` | `double` | `table`

`WithinModel` — Model specifying within-subject hypothesis test

`'separatemeans'` (default) | `'orthogonalcontrasts'` | character vector or string scalar that defines a model

Model specifying the within-subject hypothesis test, specified as the comma-separated pair consisting of `'WithinModel'` and one of the following:

- `'separatemeans'` — Compute a separate mean for each group.
- `'orthogonalcontrasts'` — This is valid only when the within-subject model has a single numeric factor *T*. Responses are the average, the slope of centered *T*, and, in general, all orthogonal contrasts for a polynomial up to *T*^(*p* – 1), where *p* is the number if rows in the within-subject model.
- A character vector or string scalar that defines a model specification in the within-subject factors. You can define the model based on the rules for the `terms` in `modelspec`.

For example, if there are three within-subject factors `w1`, `w2`, and `w3`, then you can specify a model for the within-subject factors as follows.

Example: `'WithinModel','w1+w2+w2*w3'`

Data Types: `char` | `string`

Output Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

More About

Model Specification Using Wilkinson Notation

Wilkinson notation describes the factors present in models. It does not describe the multipliers (coefficients) of those factors.

The following rules specify the responses in `modelspec`.

Wilkinson Notation	Meaning
`Y1,Y2,Y3`	Specific list of variables
`Y1-Y5`	All table variables from `Y1` through `Y5`

The following rules specify terms in `modelspec`.

Wilkinson notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	X, X^2, ..., X^k
`X1 + X2`	X1, X2
`X1*X2`	X1, X2, X1*X2
`X1:X2`	X1*X2 only
`-X2`	Do not include X2
`X1*X2 + X3`	X1, X2, X3, X1*X2
`X1 + X2 + X3 + X1:X2`	X1, X2, X3, X1*X2
`X1*X2*X3 - X1:X2:X3`	X1, X2, X3, X1*X2, X1*X3, X2*X3
`X1*(X2 + X3)`	X1, X2, X3, X1*X2, X1*X3

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`.

See Also

`RepeatedMeasuresModel`

Introduced in R2014a

fitrnet

Train neural network regression model

Syntax

```
Mdl = fitrnet(Tbl,ResponseVarName)
Mdl = fitrnet(Tbl,formula)
Mdl = fitrnet(Tbl,Y)

Mdl = fitrnet(X,Y)

Mdl = fitrnet( ___ ,Name,Value)
```

Description

Use `fitrnet` to train a feedforward, fully connected neural network for regression. The first fully connected layer of the neural network has a connection from the network input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer, excluding the last. The final fully connected layer produces the network's output, namely predicted response values. For more information, see “Neural Network Structure” on page 33-2239.

`Mdl = fitrnet(Tbl,ResponseVarName)` returns a neural network regression model `Mdl` trained using the predictors in the table `Tbl` and the response values in the `ResponseVarName` table variable.

`Mdl = fitrnet(Tbl,formula)` returns a neural network regression model trained using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of the predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitrnet(Tbl,Y)` returns a neural network regression model using the predictor variables in the table `Tbl` and the response values in vector `Y`.

`Mdl = fitrnet(X,Y)` returns a neural network regression model trained using the predictors in the matrix `X` and the response values in vector `Y`.

`Mdl = fitrnet( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can adjust the number of outputs and the activation functions for the fully connected layers by specifying the `LayerSizes` and `Activations` name-value arguments.

Examples

Train Neural Network Regression Model

Train a neural network regression model, and assess the performance of the model on a test set.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `MPG`.

```
load carbig
cars = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,Origin,Weight,MPG);
```

Partition the data into training and test sets. Use approximately 80% of the observations to train a neural network model, and 20% of the observations to test the performance of the trained model on new data. Use `cvpartition` to partition the data.

```
rng("default") % For reproducibility of the data partition
c = cvpartition(length(MPG),"Holdout",0.20);
trainingIdx = training(c); % Training set indices
carsTrain = cars(trainingIdx,:);
testIdx = test(c); % Test set indices
carsTest = cars(testIdx,:);
```

Train a neural network regression model by passing the `carsTrain` training data to the `fitrnet` function. For better results, specify to standardize the predictor data.

```
Mdl = fitrnet(carsTrain,"MPG","Standardize",true)
```

```
Mdl =
  RegressionNeuralNetwork
           PredictorNames: {'Acceleration'  'Displacement'  'Horsepower'  'Model_Year'  'Origin'
             ResponseName: 'MPG'
    CategoricalPredictors: 5
        ResponseTransform: 'none'
          NumObservations: 314
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'linear'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [1000×7 table]


  Properties, Methods
```

`Mdl` is a trained `RegressionNeuralNetwork` model. You can use dot notation to access the properties of `Mdl`. For example, you can specify `Mdl.TrainingHistory` to get more information about the training history of the neural network model.

Evaluate the performance of the regression model on the test set by computing the test mean squared error (MSE). Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,carsTest,"MPG")
```

```
testMSE = 16.6154
```

Specify Neural Network Regression Model Architecture

Specify the structure of the neural network regression model, including the size of the fully connected layers.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a matrix `X` containing the predictor variables `Acceleration`, `Cylinders`, and so on. Store the response variable `MPG` in the variable `Y`.

```
load carbig
X = [Acceleration Cylinders Displacement Weight];
Y = MPG;
```

Partition the data into training data (`XTrain` and `YTrain`) and test data (`XTest` and `YTest`). Reserve approximately 20% of the observations for testing, and use the rest of the observations for training.

```
rng("default") % For reproducibility of the partition
c = cvpartition(length(Y),"Holdout",0.20);
trainingIdx = training(c); % Indices for the training set
XTrain = X(trainingIdx,:);
YTrain = Y(trainingIdx);
testIdx = test(c); % Indices for the test set
XTest = X(testIdx,:);
YTest = Y(testIdx);
```

Train a neural network regression model. Specify to standardize the predictor data, and to have 30 outputs in the first fully connected layer and 10 outputs in the second fully connected layer. By default, both layers use a rectified linear unit (ReLU) activation function. You can change the activation functions for the fully connected layers by using the `Activations` name-value argument.

```
Mdl = fitrnet(XTrain,YTrain,"Standardize",true, ...
    "LayerSizes",[30 10])
```

```
Mdl =
  RegressionNeuralNetwork
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 318
              LayerSizes: [30 10]
             Activations: 'relu'
    OutputLayerActivation: 'linear'
                  Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [1000×7 table]


  Properties, Methods
```

Access the weights and biases for the fully connected layers of the trained model by using the `LayerWeights` and `LayerBiases` properties of `Mdl`. The first two elements of each property correspond to the values for the first two fully connected layers, and the third element corresponds to the values for the final fully connected layer for regression. For example, display the weights and biases for the first fully connected layer.

```
Mdl.LayerWeights{1}
```

```
ans = 30×4

   -1.0617    0.1287    0.0797    0.4648
   -0.6497   -1.4565   -2.6026    2.6962
   -0.6420    0.2744   -0.0234   -0.0252
   -1.9727   -0.4665   -0.5833    0.9371
   -0.4373    0.1607    0.3930    0.7859
    0.5091   -0.0032   -0.6503   -1.6694
    0.0123   -0.2624   -2.2928   -1.0965
   -0.1386    1.2747    0.4085    0.5395
   -0.1755    1.5641   -3.1896   -1.1336
    0.4401    0.4942    1.8957   -1.1617
      ⋮
```

```
Mdl.LayerBiases{1}
```

```
ans = 30×1

   -1.3086
   -1.6205
   -0.7815
    1.5382
   -0.5256
    1.2394
   -2.3078
   -1.0709
   -1.8898
    1.9443
      ⋮
```

The final fully connected layer has one output. The number of layer outputs corresponds to the first dimension of the layer weights and layer biases.

```
size(Mdl.LayerWeights{end})
```

```
ans = 1×2

     1    10
```

```
size(Mdl.LayerBiases{end})
```

```
ans = 1×2

     1     1
```

To estimate the performance of the trained model, compute the test set mean squared error (MSE) for `Mdl`. Smaller MSE values indicate better performance.

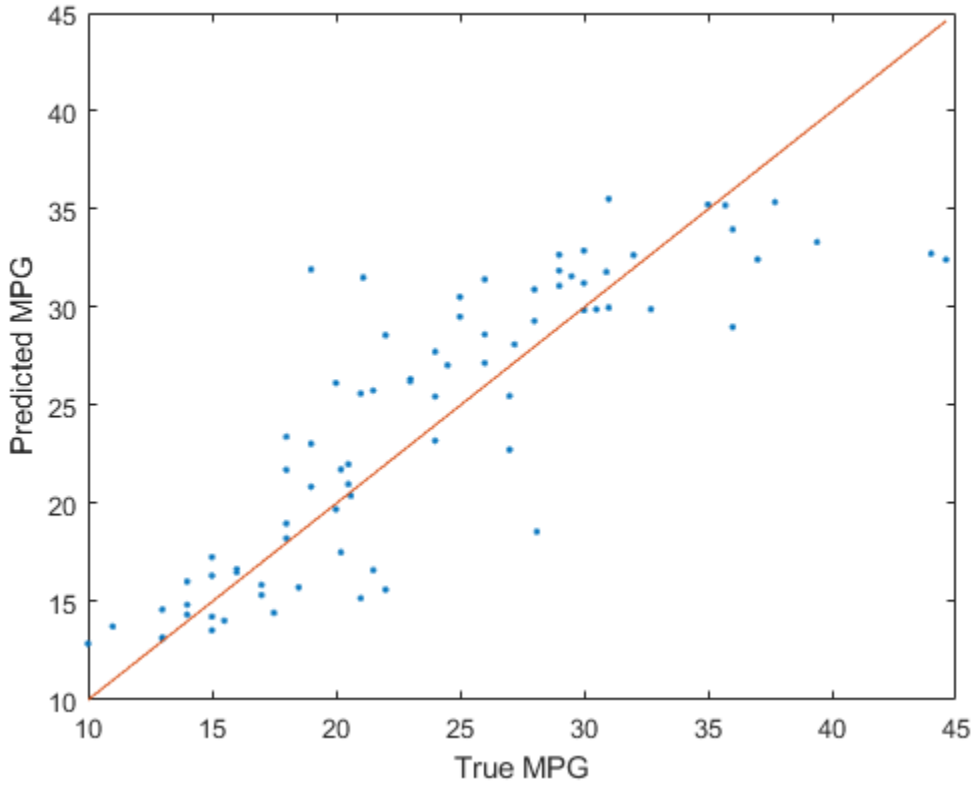
```
testMSE = loss(Mdl,XTest,YTest)
```

```
testMSE = 17.2022
```

Compare the predicted test set response values to the true response values. Plot the predicted miles per gallon (MPG) along the vertical axis and the true MPG along the horizontal axis. Points on the

reference line indicate correct predictions. A good model produces predictions that are scattered near the line.

```
testPredictions = predict(Mdl,XTest);
plot(YTest,testPredictions,".")
hold on
plot(YTest,YTest)
hold off
xlabel("True MPG")
ylabel("Predicted MPG")
```



Stop Neural Network Training Early Using Validation Data

At each iteration of the training process, compute the validation loss of the neural network. Stop the training process early if the validation loss reaches a reasonable minimum.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Systolic` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Age,Diastolic,Gender,Height,Smoker,Weight,Systolic);
```

Separate the data into a training set `tblTrain` and a validation set `tblValidation`. The software reserves approximately 30% of the observations for the validation data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(size(tbl,1),"Holdout",0.30);
trainingIndices = training(c);
validationIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblValidation = tbl(validationIndices,:);
```

Train a neural network regression model by using the training set. Specify the `Systolic` column of `tblTrain` as the response variable. Evaluate the model at each iteration by using the validation set. Specify to display the training information at each iteration by using the `Verbose` name-value argument. By default, the training process ends early if the validation loss is greater than or equal to the minimum validation loss computed so far, six times in a row. To change the number of times the validation loss is allowed to be greater than or equal to the minimum, specify the `ValidationPatience` name-value argument.

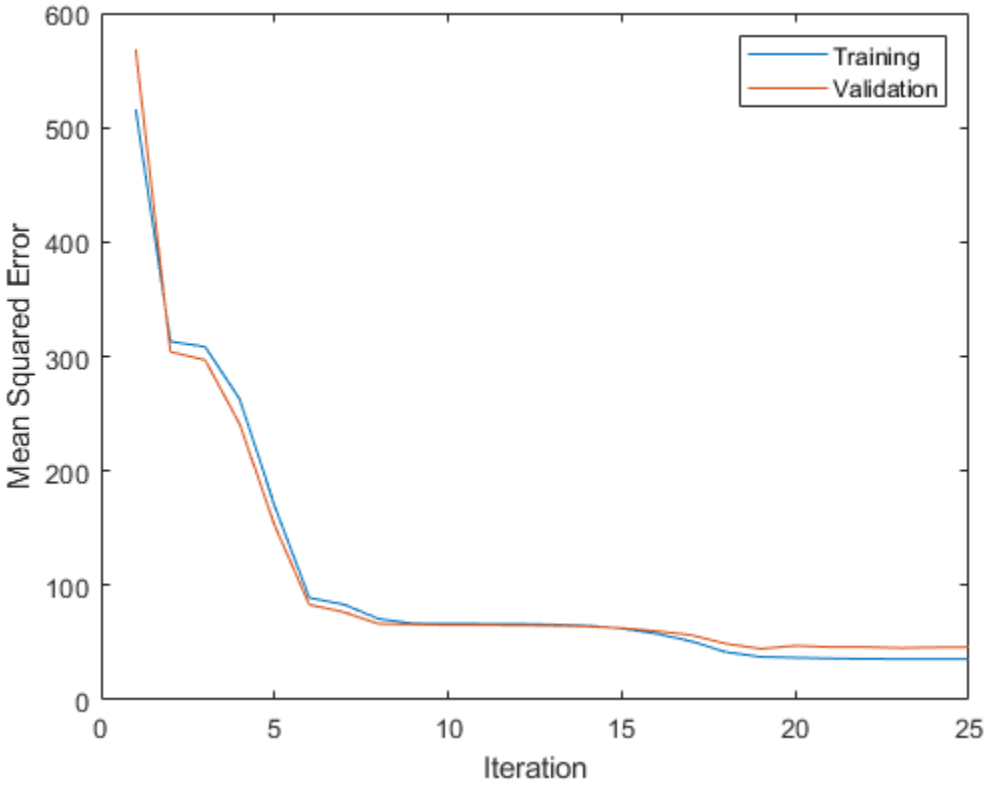
```
Mdl = fitrnet(tblTrain,"Systolic", ...
    "ValidationData",tblValidation, ...
    "Verbose",1);
```

```
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|           1|  516.021993| 3220.880047|    0.644473|    0.005193|  568.289202|           0|
|           2|  313.056754|  229.931405|    0.067026|    0.002658|  304.023695|           0|
|           3|  308.461807|  277.166516|    0.011122|    0.001363|  296.935608|           0|
|           4|  262.492770|  844.627934|    0.143022|    0.000531|  240.559640|           0|
|           5|  169.558740| 1131.714363|    0.336463|    0.000652|  152.531663|           0|
|           6|   89.134368|  362.084104|    0.382677|    0.001059|   83.147478|           0|
|           7|   83.309729|  994.830303|    0.199923|    0.000515|   76.634122|           0|
|           8|   70.731524|  327.637362|    0.041366|    0.000361|   66.421750|           0|
|           9|   66.650091|  124.369963|    0.125232|    0.000380|   65.914063|           0|
|          10|   66.404753|   36.699328|    0.016768|    0.000363|   65.357335|           0|
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|          11|   66.357143|   46.712988|    0.009405|    0.001130|   65.306106|           0|
|          12|   66.268225|   54.079264|    0.007953|    0.001023|   65.234391|           0|
|          13|   65.788550|   99.453225|    0.030942|    0.000436|   64.869708|           0|
|          14|   64.821095|  186.344649|    0.048078|    0.000295|   64.191533|           0|
|          15|   62.353896|  319.273873|    0.107160|    0.000290|   62.618374|           0|
|          16|   57.836593|  447.826470|    0.184985|    0.000287|   60.087065|           0|
|          17|   51.188884|  524.631067|    0.253062|    0.000287|   56.646294|           0|
|          18|   41.755601|  189.072516|    0.318515|    0.000286|   49.046823|           0|
|          19|   37.539854|   78.602559|    0.382284|    0.000290|   44.633562|           0|
|          20|   36.845322|  151.837884|    0.211286|    0.000286|   47.291367|           1|
|=========================================================================================|
| Iteration  | Train Loss | Gradient   | Step       | Iteration  | Validation | Validation |
|            |            |            |            | Time (sec) | Loss       | Checks     |
|=========================================================================================|
|          21|   36.218289|   62.826818|    0.142748|    0.000362|   46.139104|           2|
|          22|   35.776921|   53.606315|    0.215188|    0.000321|   46.170460|           3|
```

```
|          23|   35.729085|   24.400342|    0.060096|    0.001023|   45.318023|          4|
|          24|   35.622031|    9.602277|    0.121153|    0.000289|   45.791861|          5|
|          25|   35.573317|   10.735070|    0.126854|    0.000291|   46.062826|          6|
|=========================================================================================|
```

Create a plot that compares the training mean squared error (MSE) and the validation MSE at each iteration. By default, `fitrnet` stores the loss information inside the `TrainingHistory` property of the object `Mdl`. You can access this information by using dot notation.

```
iteration = Mdl.TrainingHistory.Iteration;
trainLosses = Mdl.TrainingHistory.TrainingLoss;
valLosses = Mdl.TrainingHistory.ValidationLoss;
plot(iteration,trainLosses,iteration,valLosses)
legend(["Training","Validation"])
xlabel("Iteration")
ylabel("Mean Squared Error")
```



Check the iteration that corresponds to the minimum validation MSE. The final returned model `Mdl` is the model trained at this iteration.

```
[~,minIdx] = min(valLosses);
iteration(minIdx)
```

```
ans = 19
```

Find Good Regularization Strength for Neural Network Using Cross-Validation

Assess the cross-validation loss of neural network models with different regularization strengths, and choose the regularization strength corresponding to the best performing model.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `MPG`. Remove observations from the table with missing values.

```
load carbig
tbl = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,Origin,Weight,MPG);
cars = rmmissing(tbl);
```

Create a `cvpartition` object for 5-fold cross-validation. `cvp` partitions the data into five folds, where each fold has roughly the same number of observations. Set the random seed to the default value for reproducibility of the partition.

```
rng("default")
n = size(cars,1);
cvp = cvpartition(n,"KFold",5);
```

Compute the cross-validation mean squared error (MSE) for neural network regression models with different regularization strengths. Try regularization strengths on the order of $1/n$, where n is the number of observations. Specify to standardize the data before training the neural network models.

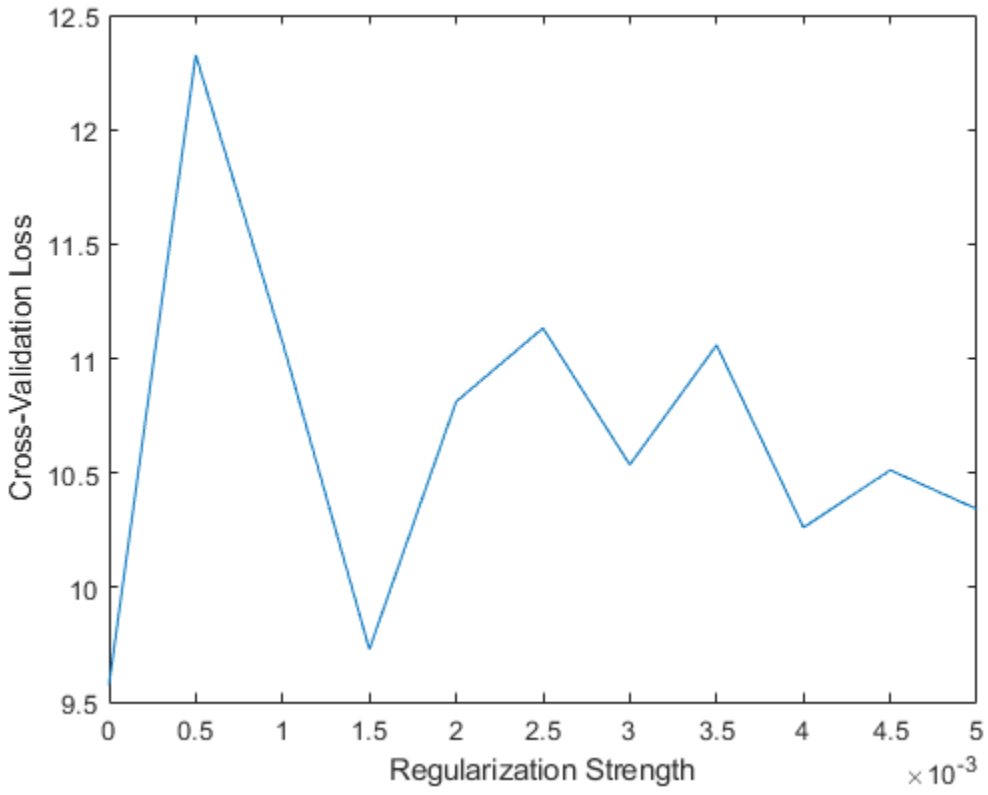
```
1/n
```

```
ans = 0.0026
```

```
lambda = (0:0.5:5)*1e-3;
cvloss = zeros(length(lambda),1);
for i = 1:length(lambda)
    cvMdl = fitrnet(cars,"MPG","Lambda",lambda(i), ...
        "CVPartition",cvp,"Standardize",true);
    cvloss(i) = kfoldLoss(cvMdl);
end
```

Plot the results. Find the regularization strength corresponding to the lowest cross-validation MSE.

```
plot(lambda,cvloss)
xlabel("Regularization Strength")
ylabel("Cross-Validation Loss")
```




```
[~,idx] = min(cvloss);
bestLambda = lambda(idx)
```

```
bestLambda = 0
```

Train a neural network regression model using the `bestLambda` regularization strength.

```
Mdl = fitrnet(cars,"MPG","Lambda",bestLambda, ...
    "Standardize",true)
```

```
Mdl =
  RegressionNeuralNetwork
           PredictorNames: {'Acceleration'  'Displacement'  'Horsepower'  'Model_Year'  'Origin'
             ResponseName: 'MPG'
    CategoricalPredictors: 5
        ResponseTransform: 'none'
          NumObservations: 392
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'linear'
                   Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [1000×7 table]


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl` stores the response variable `Y` as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Response data

numeric vector

Response data, specified as a numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

X — Predictor data

numeric matrix

Predictor data used to train the model, specified as a numeric matrix.

By default, the software treats each row of `X` as one observation, and each column as one predictor.

The length of `Y` and the number of observations in `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value argument.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

Data Types: `single` | `double`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response (for example, `Y` or `ValidationData{2}`)
- At least one missing value in a predictor observation (for example, row in `X` or `ValidationData{1}`)
- `NaN` value or `0` weight (for example, value in `Weights` or `ValidationData{3}`)

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitrnet(X,Y,'LayerSizes',[10 10],'Activations',["relu","tanh"])` specifies to create a neural network with two fully connected layers, each with 10 outputs. The first layer uses a rectified linear unit (ReLU) activation function, and the second uses a hyperbolic tangent activation function.

Neural Network Options

LayerSizes — Sizes of fully connected layers

`10` (default) | positive integer vector

Sizes of the fully connected layers in the neural network model, specified as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer. For more information, see “Neural Network Structure” on page 33-2239.

Example: `'LayerSizes',[100 25 10]`

`Activations` — Activation functions for fully connected layers
`'relu'` (default) | `'tanh'` | `'sigmoid'` | `'none'` | string array | cell array of character vectors

Activation functions for the fully connected layers of the neural network model, specified as a character vector, string scalar, string array, or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If you specify one activation function only, then `Activations` is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer (see “Neural Network Structure” on page 33-2239).
- If you specify an array of activation functions, then the *i*th element of `Activations` is the activation function for the *i*th layer of the neural network model.

Example: `'Activations','sigmoid'`

`LayerWeightsInitializer` — Function to initialize fully connected layer weights
`'glorot'` (default) | `'he'`

Function to initialize the fully connected layer weights, specified as `'glorot'` or `'he'`.

Value	Description
`'glorot'`	Initialize the weights with the Glorot initializer [1] (also known as the Xavier initializer). For each layer, the Glorot initalizer independently samples from a uniform distribution with zero mean and variable `2/(I+O)`, where `I` is the input size and `O` is the output size for the layer.

Value	Description
`'he'`	Initialize the weights with the He initializer [2]. For each layer, the He initializer samples from a normal distribution with zero mean and variance `2/I`, where `I` is the input size for the layer.

Example: `'LayerWeightsFunction','he'`

`LayerBiasesInitializer` — Type of initial fully connected layer biases
`'zeros'` (default) | `'ones'`

Type of initial fully connected layer biases, specified as `'zeros'` or `'ones'`.

- If you specify the value `'zeros'`, then each fully connected layer has an initial bias of 0.
- If you specify the value `'ones'`, then each fully connected layer has an initial bias of 1.

Example: `'LayerBiasesInitializer','ones'`

Data Types: `char` | `string`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

`Lambda` — Regularization term strength
`0` (default) | nonnegative scalar

Regularization term strength, specified as a nonnegative scalar. The software composes the objective function for minimization from the mean squared error (MSE) loss function and the ridge (L2) penalty term.

Example: `'Lambda',1e-4`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data
`false` or `0` (default) | `true` or `1`

Flag to standardize the predictor data, specified as a numeric or logical `0` (`false`) or `1` (`true`). If you set `Standardize` to `true`, then the software centers and scales each numeric predictor variable by the corresponding column mean and standard deviation. The software does not standardize the categorical predictors.

Example: `'Standardize',true`

Data Types: `single` | `double` | `logical`

Convergence Control Options

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as `0` or `1`. The `'Verbose'` name-value argument controls the amount of diagnostic information that `fitrnet` displays at the command line.

Value	Description
`0`	`fitrnet` does not display diagnostic information.
`1`	`fitrnet` periodically displays diagnostic information.

By default, `StoreHistory` is set to `true` and `fitrnet` stores the diagnostic information inside of `Mdl`. Use `Mdl.TrainingHistory` to access the diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

`VerboseFrequency` — Frequency of verbose printing
`1` (default) | positive integer scalar

Frequency of verbose printing, which is the number of iterations between printing to the command window, specified as a positive integer scalar. A value of 1 indicates to print diagnostic information at every iteration.

Note To use this name-value argument, set `Verbose` to `1`.

Example: `'VerboseFrequency',5`

Data Types: `single` | `double`

`StoreHistory` — Flag to store training history
`true` or `1` (default) | `false` or `0`

Flag to store the training history, specified as a numeric or logical `0` (`false`) or `1` (`true`). If `StoreHistory` is set to `true`, then the software stores diagnostic information inside of `Mdl`, which you can access by using `Mdl.TrainingHistory`.

Example: `'StoreHistory',false`

Data Types: `single` | `double` | `logical`

`IterationLimit` — Maximum number of training iterations
`1e3` (default) | positive integer scalar

Maximum number of training iterations, specified as a positive integer scalar.

The software returns a trained model regardless of whether the training routine successfully converges. `Mdl.ConvergenceInfo` contains convergence information.

Example: `'IterationLimit',1e8`

Data Types: `single` | `double`

`GradientTolerance` — Relative gradient tolerance
`1e-6` (default) | nonnegative scalar

Relative gradient tolerance, specified as a nonnegative scalar.

Let $\mathscr{L}_t$ be the loss function at training iteration t, $\nabla\mathscr{L}_t$ be the gradient of the loss function with respect to the weights and biases at iteration t, and $\nabla\mathscr{L}_0$ be the gradient of the loss function at an initial point. If $\max|\nabla\mathscr{L}_t| \le a \cdot \text{GradientTolerance}$, where $a = \max(1, \min|\mathscr{L}_t|, \max|\nabla\mathscr{L}_0|)$, then the training process terminates.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`LossTolerance` — Loss tolerance
`1e-6` (default) | nonnegative scalar

Loss tolerance, specified as a nonnegative scalar.

If the function loss at some iteration is smaller than `LossTolerance`, then the training process terminates.

Example: `'LossTolerance',1e-8`

Data Types: `single` | `double`

`StepTolerance` — Step size tolerance
`1e-6` (default) | nonnegative scalar

Step size tolerance, specified as a nonnegative scalar.

If the step size at some iteration is smaller than `StepTolerance`, then the training process terminates.

Example: `'StepTolerance',1e-4`

Data Types: `single` | `double`

`ValidationData` — Validation data for training convergence detection
cell array | table

Validation data for training convergence detection, specified as a cell array or table.

During the training process, the software periodically estimates the validation loss by using `ValidationData`. If the validation loss increases more than `ValidationPatience` times in a row, then the software terminates the training.

You can specify `ValidationData` as a table if you use a table `Tbl` of predictor data that contains the response variable. In this case, `ValidationData` must contain the same predictors and response contained in `Tbl`. The software does not apply weights to observations, even if `Tbl` contains a vector of weights. To specify weights, you must specify `ValidationData` as a cell array.

If you specify `ValidationData` as a cell array, then it must have the following format:

- `ValidationData{1}` must have the same data type and orientation as the predictor data. That is, if you use a predictor matrix `X`, then `ValidationData{1}` must be an *m*-by-*p* or *p*-by-*m* matrix of predictor data that has the same orientation as `X`. The predictor variables in the training data `X`

and `ValidationData{1}` must correspond. Similarly, if you use a predictor table `Tbl` of predictor data, then `ValidationData{1}` must be a table containing the same predictor variables contained in `Tbl`. The number of observations in `ValidationData{1}` and the predictor data can vary.

- `ValidationData{2}` must match the data type and format of the response variable, either `Y` or `ResponseVarName`. If `ValidationData{2}` is an array of responses, then it must have the same number of elements as the number of observations in `ValidationData{1}`. If `ValidationData{1}` is a table, then `ValidationData{2}` can be the name of the response variable in the table. If you want to use the same `ResponseVarName` or `formula`, you can specify `ValidationData{2}` as `[]`.
- Optionally, you can specify `ValidationData{3}` as an *m*-dimensional numeric vector of observation weights or the name of a variable in the table `ValidationData{1}` that contains observation weights. The software normalizes the weights with the validation data so that they sum to 1.

If you specify `ValidationData` and want to display the validation loss at the command line, set `Verbose` to `1`.

`ValidationFrequency` — Number of iterations between validation evaluations

`1` (default) | positive integer scalar

Number of iterations between validation evaluations, specified as a positive integer scalar. A value of 1 indicates to evaluate validation metrics at every iteration.

Note To use this name-value argument, you must specify `ValidationData`.

Example: `'ValidationFrequency',5`

Data Types: `single` | `double`

`ValidationPatience` — Stopping condition for validation evaluations

`6` (default) | nonnegative integer scalar

Stopping condition for validation evaluations, specified as a nonnegative integer scalar. Training stops if the validation loss is greater than or equal to the minimum validation loss computed so far, `ValidationPatience` times in a row. You can check the `Mdl.TrainingHistory` table to see the running total of times that the validation loss is greater than or equal to the minimum (`Validation Checks`).

Example: `'ValidationPatience',10`

Data Types: `single` | `double`

Other Regression Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table. The descriptions assume that the predictor data has observations in rows and predictors in columns.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrnet` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrnet` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrnet` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitrnet` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitrnet` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitrnet` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `'PredictorNames'` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the predictor order in `X`. Assuming that `X` has the default orientation, with observations in rows and predictors in columns, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.

  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `fitrnet` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Weights` — Observation weights
nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

`fitrnet` normalizes the weights to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Flag to train cross-validated model
`'off'` (default) | `'on'`

Flag to train a cross-validated model, specified as `'on'` or `'off'`.

If you specify `'on'`, then the software trains a cross-validated model with 10 folds.

You can override this cross-validation setting using the `CVPartition`, `Holdout`, `KFold`, or `Leaveout` name-value argument. You can use only one cross-validation name-value argument at a time to create a cross-validated model.

Alternatively, cross-validate later by passing `Mdl` to `crossval`.

Example: `'Crossval','on'`

Data Types: `char` | `string`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Output Arguments

`Mdl` — Trained neural network regression model
`RegressionNeuralNetwork` object | `RegressionPartitionedModel` object

Trained neural network regression model, returned as a `RegressionNeuralNetwork` or `RegressionPartitionedModel`object.

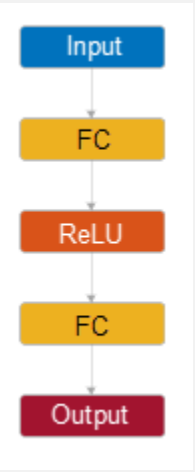
If you set any of the name-value arguments `CrossVal`, `CVPartition`, `Holdout`, `KFold`, or `Leaveout`, then `Mdl` is a `RegressionPartitionedModel` object. Otherwise, `Mdl` is a `RegressionNeuralNetwork` model.

To reference properties of `Mdl`, use dot notation.

More About

Neural Network Structure

The default neural network regression model has the following layer structure.

Structure	Description
	Input — This layer corresponds to the predictor data in `Tbl` or `X`.
	First fully connected layer — This layer has 10 outputs by default. • You can widen the layer or add more fully connected layers to the network by specifying the `LayerSizes` name-value argument. • You can find the weights and biases for this layer in the `Mdl.LayerWeights{1}` and `Mdl.LayerBiases{1}` properties of `Mdl`, respectively.
	ReLU activation function — `fitrnet` applies this activation function to the first fully connected layer. • You can change the activation function by specifying the `Activations` name-value argument.
	Final fully connected layer — This layer has one output. • You can find the weights and biases for this layer in the `Mdl.LayerWeights{end}` and `Mdl.LayerBiases{end}` properties of `Mdl`, respectively.
	Output — This layer corresponds to the predicted response values.

For an example that shows how a regression neural network model with this layer structure returns predictions, see “Predict Using Layer Structure of Regression Neural Network Model” on page 33-4889.

Tips

- Always try to standardize the numeric predictors (see `Standardize`). Standardization makes predictors insensitive to the scales on which they are measured.

Algorithms

Training Solver

`fitrnet` uses a limited-memory Broyden-Flecter-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) [3] as its loss function minimization technique, where the software minimizes the mean squared error (MSE).

References

[1] Glorot, Xavier, and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks.” In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. 2010.

[2] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.” In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034. 2015.

[3] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

See Also

`CompactRegressionNeuralNetwork` | `RegressionNeuralNetwork` | `RegressionPartitionedModel` | `loss` | `predict`

Topics

“Assess Regression Neural Network Performance” on page 18-184

Introduced in R2021a

fitdist

Fit probability distribution object to data

Syntax

```
pd = fitdist(x,distname)
pd = fitdist(x,distname,Name,Value)

[pdca,gn,gl] = fitdist(x,distname,'By',groupvar)
[pdca,gn,gl] = fitdist(x,distname,'By',groupvar,Name,Value)
```

Description

`pd = fitdist(x,distname)` creates a probability distribution object by fitting the distribution specified by `distname` to the data in column vector `x`.

`pd = fitdist(x,distname,Name,Value)` creates the probability distribution object with additional options specified by one or more name-value pair arguments. For example, you can indicate censored data or specify control parameters for the iterative fitting algorithm.

`[pdca,gn,gl] = fitdist(x,distname,'By',groupvar)` creates probability distribution objects by fitting the distribution specified by `distname` to the data in `x` based on the grouping variable `groupvar`. It returns a cell array of fitted probability distribution objects, `pdca`, a cell array of group labels, `gn`, and a cell array of grouping variable levels, `gl`.

`[pdca,gn,gl] = fitdist(x,distname,'By',groupvar,Name,Value)` returns the above output arguments using additional options specified by one or more name-value pair arguments. For example, you can indicate censored data or specify control parameters for the iterative fitting algorithm.

Examples

Fit a Normal Distribution to Data

Load the sample data. Create a vector containing the patients' weight data.

```
load hospital
x = hospital.Weight;
```

Create a normal distribution object by fitting it to the data.

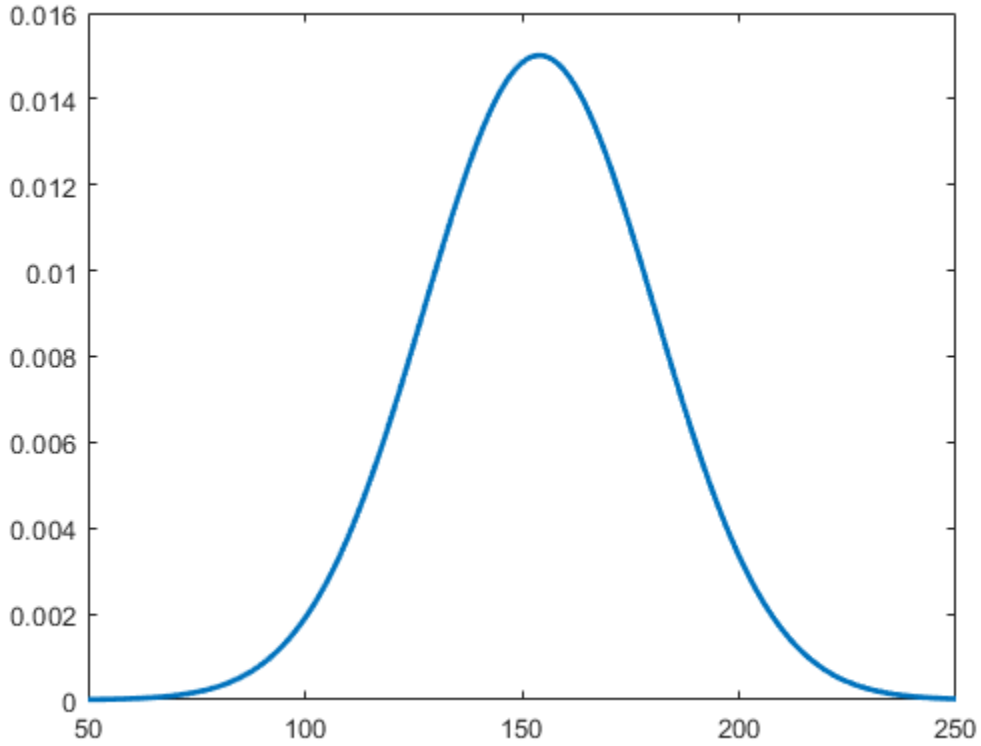
```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu =     154   [148.728, 159.272]
    sigma = 26.5714   [23.3299, 30.8674]
```

The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

Plot the pdf of the distribution.

```
x_values = 50:1:250;
y = pdf(pd,x_values);
plot(x_values,y,'LineWidth',2)
```



Fit a Kernel Distribution to Data

Load the sample data. Create a vector containing the patients' weight data.

```
load hospital
x = hospital.Weight;
```

Create a kernel distribution object by fitting it to the data. Use the Epanechnikov kernel function.

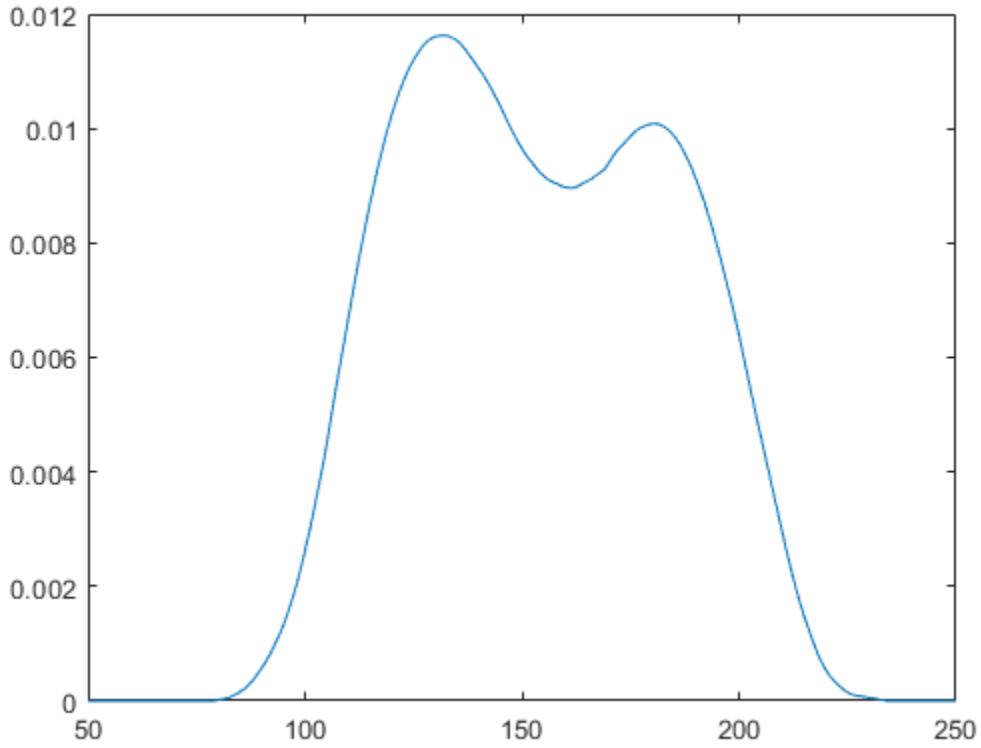
```
pd = fitdist(x,'Kernel','Kernel','epanechnikov')
```

```
pd =
  KernelDistribution

    Kernel = epanechnikov
    Bandwidth = 14.3792
```

```
    Support = unbounded
```

Plot the pdf of the distribution.

```
x_values = 50:1:250;
y = pdf(pd,x_values);
plot(x_values,y)
```



Fit Normal Distributions to Grouped Data

Load the sample data. Create a vector containing the patients' weight data.

```
load hospital
x = hospital.Weight;
```

Create normal distribution objects by fitting them to the data, grouped by patient gender.

```
gender = hospital.Sex;
[pdca,gn,gl] = fitdist(x,'Normal','By',gender)
```

```
pdca=1×2 cell array
    {1x1 prob.NormalDistribution}    {1x1 prob.NormalDistribution}
```

```
gn = 2x1 cell
    {'Female'}
    {'Male'  }


gl = 2x1 cell
    {'Female'}
    {'Male'  }
```

The cell array `pdca` contains two probability distribution objects, one for each gender group. The cell array `gn` contains two group labels. The cell array `gl` contains two group levels.

View each distribution in the cell array `pdca` to compare the mean, `mu`, and the standard deviation, `sigma`, grouped by patient gender.

```
female = pdca{1}  % Distribution for females
```

```
female =
  NormalDistribution

  Normal distribution
       mu = 130.472   [128.183, 132.76]
    sigma = 8.30339   [6.96947, 10.2736]
```

```
male = pdca{2}  % Distribution for males
```

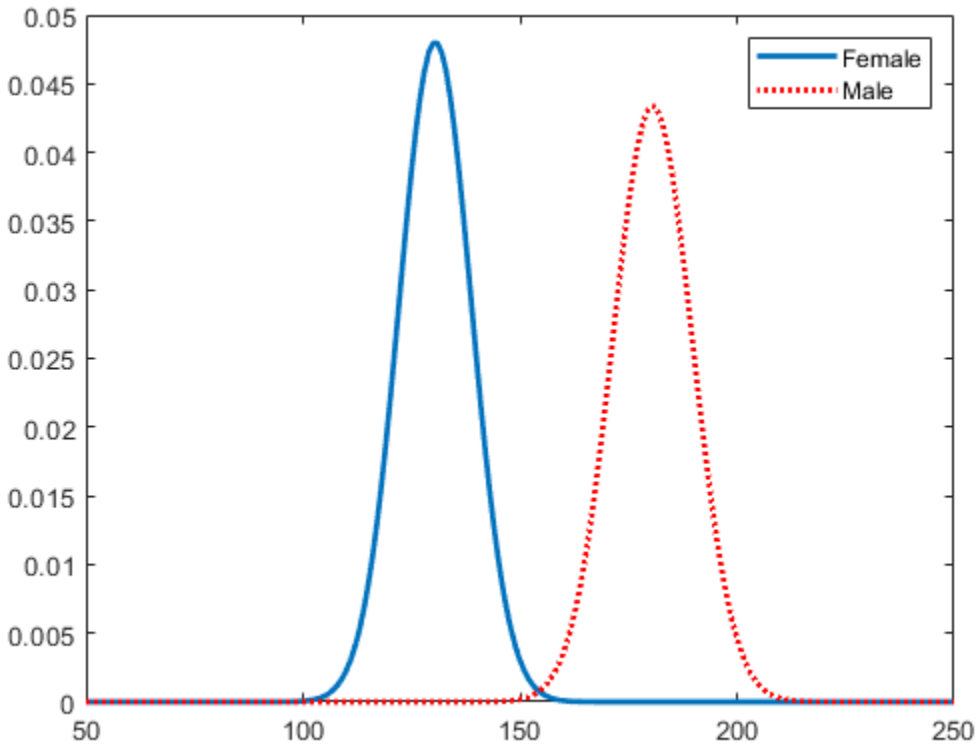
```
male =
  NormalDistribution

  Normal distribution
       mu = 180.532   [177.833, 183.231]
    sigma = 9.19322   [7.63933, 11.5466]
```

Compute the pdf of each distribution.

```
x_values = 50:1:250;
femalepdf = pdf(female,x_values);
malepdf = pdf(male,x_values);
```

Plot the pdfs for a visual comparison of weight distribution by gender.

```
figure
plot(x_values,femalepdf,'LineWidth',2)
hold on
plot(x_values,malepdf,'Color','r','LineStyle',':','LineWidth',2)
legend(gn,'Location','NorthEast')
hold off
```



Fit Kernel Distributions to Grouped Data

Load the sample data. Create a vector containing the patients' weight data.

```
load hospital
x = hospital.Weight;
```

Create kernel distribution objects by fitting them to the data, grouped by patient gender. Use a triangular kernel function.

```
gender = hospital.Sex;
[pdca,gn,gl] = fitdist(x,'Kernel','By',gender,'Kernel','triangle');
```

View each distribution in the cell array `pdca` to see the kernel distributions for each gender.

```
female = pdca{1}  % Distribution for females
```

```
female =
  KernelDistribution

    Kernel = triangle
    Bandwidth = 4.25894
    Support = unbounded
```

```
male = pdca{2}  % Distribution for males
```

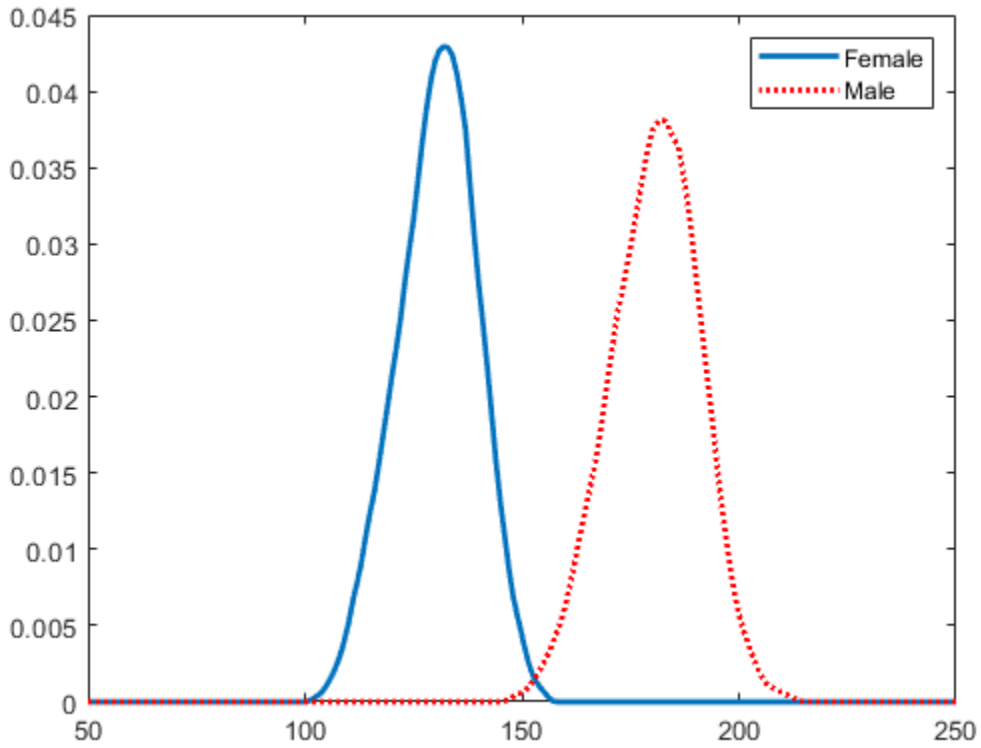
```
male =
  KernelDistribution

    Kernel = triangle
    Bandwidth = 5.08961
    Support = unbounded
```

Compute the pdf of each distribution.

```
x_values = 50:1:250;
femalepdf = pdf(female,x_values);
malepdf = pdf(male,x_values);
```

Plot the pdfs for a visual comparison of weight distribution by gender.

```
figure
plot(x_values,femalepdf,'LineWidth',2)
hold on
plot(x_values,malepdf,'Color','r','LineStyle',':','LineWidth',2)
legend(gn,'Location','NorthEast')
hold off
```



Input Arguments

x — Input data

column vector

Input data, specified as a column vector. `fitdist` ignores `NaN` values in `x`. Additionally, any `NaN` values in the censoring vector or frequency vector cause `fitdist` to ignore the corresponding values in `x`.

Data Types: `double`

`distname` — Distribution name
character vector | string scalar

Distribution name, specified as one of the following character vectors or string scalars. The distribution specified by `distname` determines the type of the returned probability distribution object.

Distribution Name	Description	Distribution Object
`'Beta'`	Beta distribution	`BetaDistribution`
`'Binomial'`	Binomial distribution	`BinomialDistribution`
`'BirnbaumSaunders'`	Birnbaum-Saunders distribution	`BirnbaumSaundersDistribution`
`'Burr'`	Burr distribution	`BurrDistribution`
`'Exponential'`	Exponential distribution	`ExponentialDistribution`
`'ExtremeValue'`	Extreme Value distribution	`ExtremeValueDistribution`
`'Gamma'`	Gamma distribution	`GammaDistribution`
`'GeneralizedExtremeValue'`	Generalized Extreme Value distribution	`GeneralizedExtremeValueDistribution`
`'GeneralizedPareto'`	Generalized Pareto distribution	`GeneralizedParetoDistribution`
`'HalfNormal'`	Half-normal distribution	`HalfNormalDistribution`
`'InverseGaussian'`	Inverse Gaussian distribution	`InverseGaussianDistribution`
`'Kernel'`	Kernel distribution	`KernelDistribution`
`'Logistic'`	Logistic distribution	`LogisticDistribution`
`'Loglogistic'`	Loglogistic distribution	`LoglogisticDistribution`
`'Lognormal'`	Lognormal distribution	`LognormalDistribution`
`'Nakagami'`	Nakagami distribution	`NakagamiDistribution`
`'NegativeBinomial'`	Negative Binomial distribution	`NegativeBinomialDistribution`
`'Normal'`	Normal distribution	`NormalDistribution`
`'Poisson'`	Poisson distribution	`PoissonDistribution`
`'Rayleigh'`	Rayleigh distribution	`RayleighDistribution`
`'Rician'`	Rician distribution	`RicianDistribution`
`'Stable'`	Stable distribution	`StableDistribution`
`'tLocationScale'`	*t* Location-Scale distribution	`tLocationScaleDistribution`

Distribution Name	Description	Distribution Object
`'Weibull'`	Weibull distribution	`WeibullDistribution`

`groupvar` — Grouping variable

categorical array | logical or numeric vector | character array | string array | cell array of character vectors

Grouping variable, specified as a categorical array, logical or numeric vector, character array, string array, or cell array of character vectors. Each unique value in a grouping variable defines a group.

For example, if `Gender` is a cell array of character vectors with values `'Male'` and `'Female'`, you can use `Gender` as a grouping variable to fit a distribution to your data by gender.

More than one grouping variable can be used by specifying a cell array of grouping variables. Observations are placed in the same group if they have common values of all specified grouping variables.

For example, if `Smoker` is a logical vector with values `0` for nonsmokers and `1` for smokers, then specifying the cell array `{Gender,Smoker}` divides observations into four groups: Male Smoker, Male Nonsmoker, Female Smoker, and Female Nonsmoker.

Example: `{Gender,Smoker}`

Data Types: `categorical` | `logical` | `single` | `double` | `char` | `string` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitdist(x,'Kernel','Kernel','triangle')` fits a kernel distribution object to the data in `x` using a triangular kernel function.

`Censoring` — Logical flag for censored data

`0` (default) | vector of logical values

Logical flag for censored data, specified as the comma-separated pair consisting of `'Censoring'` and a vector of logical values that is the same size as input vector `x`. The value is `1` when the corresponding element in `x` is a right-censored observation and `0` when the corresponding element is an exact observation. The default is a vector of `0`s, indicating that all observations are exact.

`fitdist` ignores any `NaN` values in this censoring vector. Additionally, any `NaN` values in `x` or the frequency vector cause `fitdist` to ignore the corresponding values in the censoring vector.

This argument is valid only if `distname` is `'BirnbaumSaunders'`, `'Burr'`, `'Exponential'`, `'ExtremeValue'`, `'Gamma'`, `'InverseGaussian'`, `'Kernel'`, `'Logistic'`, `'Loglogistic'`, `'Lognormal'`, `'Nakagami'`, `'Normal'`, `'Rician'`, `'tLocationScale'`, or `'Weibull'`.

Data Types: `logical`

`Frequency` — Observation frequency

`1` (default) | vector of nonnegative integer values

Observation frequency, specified as the comma-separated pair consisting of `'Frequency'` and a vector of nonnegative integer values that is the same size as input vector `x`. Each element of the

frequency vector specifies the frequencies for the corresponding elements in `x`. The default is a vector of `1`s, indicating that each value in `x` only appears once.

`fitdist` ignores any `NaN` values in this frequency vector. Additionally, any `NaN` values in `x` or the censoring vector cause `fitdist` to ignore the corresponding values in the frequency vector.

Data Types: `single` | `double`

`Options` — Control parameters
structure

Control parameters for the iterative fitting algorithm, specified as the comma-separated pair consisting of `'Options'` and a structure you create using `statset`.

Data Types: `struct`

`NTrials` — Number of trials
positive integer value

Number of trials for the binomial distribution, specified as the comma-separated pair consisting of `'NTrials'` and a positive integer value. You must specify `distname` as `'Binomial'` to use this option.

Data Types: `single` | `double`

`Theta` — Threshold parameter
`0` (default) | scalar value

Threshold parameter for the generalized Pareto distribution, specified as the comma-separated pair consisting of `'Theta'` and a scalar value. You must specify `distname` as `'GeneralizedPareto'` to use this option.

Data Types: `single` | `double`

`mu` — Location parameter
0 (default) | scalar value

Location parameter for the half-normal distribution, specified as the comma-separated pair consisting of `'mu'` and a scalar value. You must specify `distname` as `'HalfNormal'` to use this option.

Data Types: `single` | `double`

`Kernel` — Kernel smoother type
`'normal'` (default) | `'box'` | `'triangle'` | `'epanechnikov'`

Kernel smoother type, specified as the comma-separated pair consisting of `'Kernel'` and one of the following:

- `'normal'`
- `'box'`
- `'triangle'`
- `'epanechnikov'`

You must specify `distname` as `'Kernel'` to use this option.

`Support` — Kernel density support
`'unbounded'` (default) | `'positive'` | two-element vector

Kernel density support, specified as the comma-separated pair consisting of `'Support'` and `'unbounded'`, `'positive'`, or a two-element vector.

`'unbounded'`	Density can extend over the whole real line.
`'positive'`	Density is restricted to positive values.

Alternatively, you can specify a two-element vector giving finite lower and upper limits for the support of the density.

You must specify `distname` as `'Kernel'` to use this option.

Data Types: `single` | `double` | `char` | `string`

`Width` — Bandwidth of kernel smoothing window
scalar value

Bandwidth of the kernel smoothing window, specified as the comma-separated pair consisting of `'Width'` and a scalar value. The default value used by `fitdist` is optimal for estimating normal densities, but you might want to choose a smaller value to reveal features such as multiple modes. You must specify `distname` as `'Kernel'` to use this option.

Data Types: `single` | `double`

Output Arguments

`pd` — Probability distribution
probability distribution object

Probability distribution, returned as a probability distribution object. The distribution specified by `distname` determines the class type of the returned probability distribution object. For the list of `distname` values and corresponding probability distribution objects, see `distname`.

`pdca` — Probability distribution objects
cell array

Probability distribution objects of the type specified by `distname`, returned as a cell array. For the list of `distname` values and corresponding probability distribution objects, see `distname`.

`gn` — Group labels
cell array of character vectors

Group labels, returned as a cell array of character vectors.

`gl` — Grouping variable levels
cell array of character vectors

Grouping variable levels, returned as a cell array of character vectors containing one column for each grouping variable.

Algorithms

The `fitdist` function fits most distributions using maximum likelihood estimation. Two exceptions are the normal and lognormal distributions with uncensored data.

- For the uncensored normal distribution, the estimated value of the sigma parameter is the square root of the unbiased estimate of the variance.
- For the uncensored lognormal distribution, the estimated value of the sigma parameter is the square root of the unbiased estimate of the variance of the log of the data.

Alternative Functionality

App

The **Distribution Fitter** app opens a graphical user interface for you to import data from the workspace and interactively fit a probability distribution to that data. You can then save the distribution to the workspace as a probability distribution object. Open the Distribution Fitter app using `distributionFitter`, or click Distribution Fitter on the Apps tab.

References

[1] Johnson, N. L., S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Vol. 1, Hoboken, NJ: Wiley-Interscience, 1993.

[2] Johnson, N. L., S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Vol. 2, Hoboken, NJ: Wiley-Interscience, 1994.

[3] Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press, 1997.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Supported syntaxes are:

  ```
  pd = fitdist(x,distname)
  pd = fitdist(x,distname,Name,Value)
  ```

  Code generation does not support the syntaxes that include the grouping variable `'By',groupvar` and the related output arguments `pdca`, `gn`, and `gl`.
- `fitdist` supports code generation for beta, exponential, extreme value, lognormal, normal, and Weibull distributions.
  - The value of `distname` can be `'Beta'`, `'Exponential'`, `'ExtremeValue'`, `'Lognormal'`, `'Normal'` or `'Weibull'`.
  - The value of `distname` must be a compile-time constant.
- The values of `x`, `'Censoring'`, and `'Frequency'` must not contain `NaN` values.
- Code generation ignores the `'Frequency'` value for the beta distribution. Instead of specifying the `'Frequency'` value, manually add duplicated values to `x` so that the values in `x` have the frequency you want.
- Code generation does not support these input arguments: `groupvar`, `NTrials`, `Theta`, `mu`, `Kernel`, `Support`, and `Width`.

- Names in name-value pair arguments must be compile-time constants.
- These object functions of `pd` support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Probability Distribution Objects” on page 32-82.

See Also

`distributionFitter` | `histfit` | `makedist` | `mle` | `paramci`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2009a

fitensemble

Fit ensemble of learners for classification and regression

Syntax

```
Mdl = fitensemble(Tbl,ResponseVarName,Method,NLearn,Learners)
Mdl = fitensemble(Tbl,formula,Method,NLearn,Learners)
Mdl = fitensemble(Tbl,Y,Method,NLearn,Learners)

Mdl = fitensemble(X,Y,Method,NLearn,Learners)

Mdl = fitensemble( ___ ,Name,Value)
```

Description

`fitensemble` can boost or bag decision tree learners or discriminant analysis classifiers. The function can also train random subspace ensembles of KNN or discriminant analysis classifiers.

For simpler interfaces that fit classification and regression ensembles, instead use `fitcensemble` and `fitrensemble`, respectively. Also, `fitcensemble` and `fitrensemble` provide options for Bayesian optimization.

`Mdl = fitensemble(Tbl,ResponseVarName,Method,NLearn,Learners)` returns a trained ensemble model object that contains the results of fitting an ensemble of `NLearn` classification or regression learners (`Learners`) to all variables in the table `Tbl`. `ResponseVarName` is the name of the response variable in `Tbl`. `Method` is the ensemble-aggregation method.

`Mdl = fitensemble(Tbl,formula,Method,NLearn,Learners)` fits the model specified by `formula`.

`Mdl = fitensemble(Tbl,Y,Method,NLearn,Learners)` treats all variables in `Tbl` as predictor variables. `Y` is the response variable that is not in `Tbl`.

`Mdl = fitensemble(X,Y,Method,NLearn,Learners)` trains an ensemble using the predictor data in `X` and response data in `Y`.

`Mdl = fitensemble( ___ ,Name,Value)` trains an ensemble using additional options specified by one or more `Name,Value` pair arguments and any of the previous syntaxes. For example, you can specify the class order, to implement 10–fold cross-validation, or the learning rate.

Examples

Estimate the Resubstitution Loss of a Boosting Ensemble

Estimate the resubstitution loss of a trained, boosting classification ensemble of decision trees.

Load the `ionosphere` data set.

```
load ionosphere;
```

Train a decision tree ensemble using AdaBoost, 100 learning cycles, and the entire data set.

```
ClassTreeEns = fitensemble(X,Y,'AdaBoostM1',100,'Tree');
```

`ClassTreeEns` is a trained `ClassificationEnsemble` ensemble classifier.

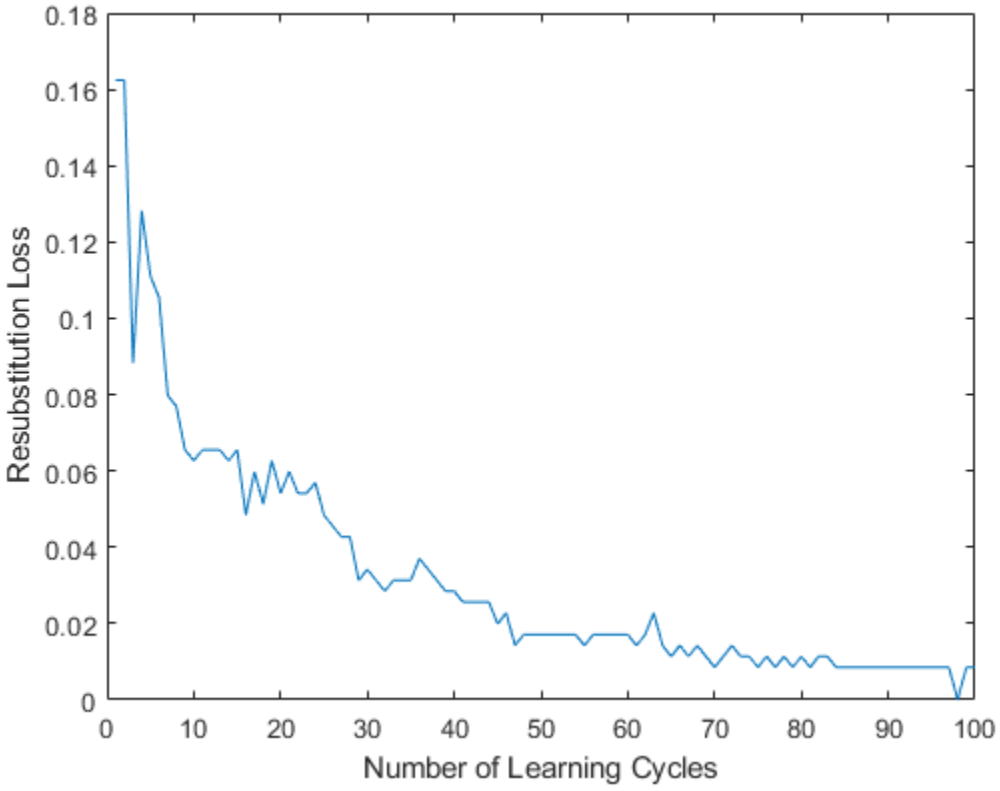
Determine the cumulative resubstitution losses (i.e., the cumulative misclassification error of the labels in the training data).

```
rsLoss = resubLoss(ClassTreeEns,'Mode','Cumulative');
```

`rsLoss` is a 100-by-1 vector, where element *k* contains the resubstitution loss after the first *k* learning cycles.

Plot the cumulative resubstitution loss over the number of learning cycles.

```
plot(rsLoss);
xlabel('Number of Learning Cycles');
ylabel('Resubstitution Loss');
```



In general, as the number of decision trees in the trained classification ensemble increases, the resubstitution loss decreases.

A decrease in resubstitution loss might indicate that the software trained the ensemble sensibly. However, you cannot infer the predictive power of the ensemble by this decrease. To measure the predictive power of an ensemble, estimate the generalization error by:

1. Randomly partitioning the data into training and cross-validation sets. Do this by specifying `'holdout',holdoutProportion` when you train the ensemble using `fitensemble`.
2. Passing the trained ensemble to `kfoldLoss`, which estimates the generalization error.

Train Regression Ensemble

Use a trained, boosted regression tree ensemble to predict the fuel economy of a car. Choose the number of cylinders, volume displaced by the cylinders, horsepower, and weight as predictors. Then, train an ensemble using fewer predictors and compare its in-sample predictive accuracy against the first ensemble.

Load the `carsmall` data set. Store the training data in a table.

```
load carsmall
Tbl = table(Cylinders,Displacement,Horsepower,Weight,MPG);
```

Specify a regression tree template that uses surrogate splits to improve predictive accuracy in the presence of `NaN` values.

```
t = templateTree('Surrogate','On');
```

Train the regression tree ensemble using LSBoost and 100 learning cycles.

```
Mdl1 = fitensemble(Tbl,'MPG','LSBoost',100,t);
```

`Mdl1` is a trained `RegressionEnsemble` regression ensemble. Because `MPG` is a variable in the MATLAB® Workspace, you can obtain the same result by entering

```
Mdl1 = fitensemble(Tbl,MPG,'LSBoost',100,t);
```

Use the trained regression ensemble to predict the fuel economy for a four-cylinder car with a 200-cubic inch displacement, 150 horsepower, and weighing 3000 lbs.

```
predMPG = predict(Mdl1,[4 200 150 3000])
```

```
predMPG = 22.8462
```

The average fuel economy of a car with these specifications is 21.78 mpg.

Train a new ensemble using all predictors in `Tbl` except `Displacement`.

```
formula = 'MPG ~ Cylinders + Horsepower + Weight';
Mdl2 = fitensemble(Tbl,formula,'LSBoost',100,t);
```

Compare the resubstitution MSEs between `Mdl1` and `Mdl2`.

```
mse1 = resubLoss(Mdl1)
```

```
mse1 = 6.4721
```

```
mse2 = resubLoss(Mdl2)
```

```
mse2 = 7.8599
```

The in-sample MSE for the ensemble that trains on all predictors is lower.

Estimate the Generalization Error of a Boosting Ensemble

Estimate the generalization error of a trained, boosting classification ensemble of decision trees.

Load the `ionosphere` data set.

```
load ionosphere;
```

Train a decision tree ensemble using AdaBoostM1, 100 learning cycles, and half of the data chosen randomly. The software validates the algorithm using the remaining half.

```
rng(2); % For reproducibility
ClassTreeEns = fitensemble(X,Y,'AdaBoostM1',100,'Tree',...
    'Holdout',0.5);
```

`ClassTreeEns` is a trained `ClassificationEnsemble` ensemble classifier.

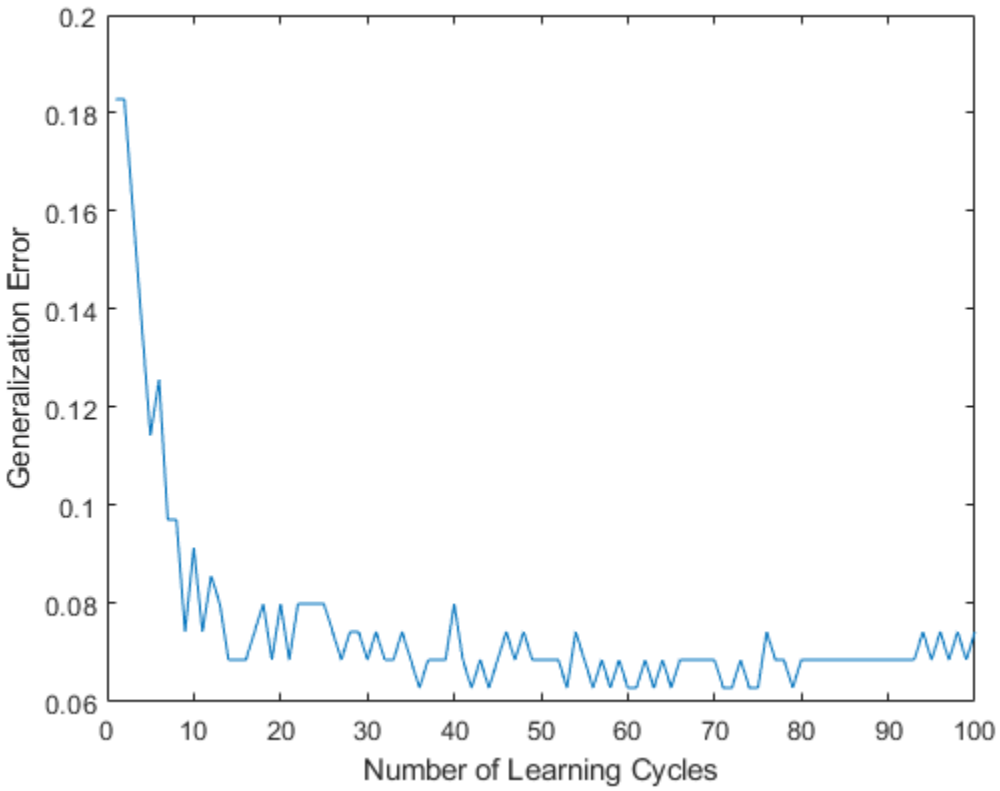
Determine the cumulative generalization error, i.e., the cumulative misclassification error of the labels in the validation data).

```
genError = kfoldLoss(ClassTreeEns,'Mode','Cumulative');
```

`genError` is a 100-by-1 vector, where element *k* contains the generalization error after the first *k* learning cycles.

Plot the generalization error over the number of learning cycles.

```
plot(genError);
xlabel('Number of Learning Cycles');
ylabel('Generalization Error');
```



The cumulative generalization error decreases to approximately 7% when 25 weak learners compose the ensemble classifier.

Find the Optimal Number of Splits and Trees for an Ensemble

You can control the depth of the trees in an ensemble of decision trees. You can also control the tree depth in an ECOC model containing decision tree binary learners using the `MaxNumSplits`, `MinLeafSize`, or `MinParentSize` name-value pair parameters.

- When bagging decision trees, `fitensemble` grows deep decision trees by default. You can grow shallower trees to reduce model complexity or computation time.
- When boosting decision trees, `fitensemble` grows stumps (a tree with one split) by default. You can grow deeper trees for better accuracy.

Load the `carsmall` data set. Specify the variables `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as predictors, and `MPG` as the response.

```
load carsmall
X = [Acceleration Displacement Horsepower Weight];
Y = MPG;
```

The default values of the tree depth controllers for boosting regression trees are:

- `1` for `MaxNumSplits`. This option grows stumps.
- `5` for `MinLeafSize`
- `10` for `MinParentSize`

To search for the optimal number of splits:

1. Train a set of ensembles. Exponentially increase the maximum number of splits for subsequent ensembles from stump to at most *n* - 1 splits, where *n* is the training sample size. Also, decrease the learning rate for each ensemble from 1 to 0.1.
2. Cross validate the ensembles.
3. Estimate the cross-validated mean-squared error (MSE) for each ensemble.
4. Compare the cross-validated MSEs. The ensemble with the lowest one performs the best, and indicates the optimal maximum number of splits, number of trees, and learning rate for the data set.

Grow and cross validate a deep regression tree and a stump. Specify to use surrogate splits because the data contains missing values. These serve as benchmarks.

```
MdlDeep = fitrtree(X,Y,'CrossVal','on','MergeLeaves','off',...
    'MinParentSize',1,'Surrogate','on');
MdlStump = fitrtree(X,Y,'MaxNumSplits',1,'CrossVal','on','Surrogate','on');
```

Train the boosting ensembles using 150 regression trees. Cross validate the ensemble using 5-fold cross validation. Vary the maximum number of splits using the values in the sequence $\{2^0, 2^1, \ldots, 2^m\}$, where m is such that 2^m is no greater than n - 1, where n is the training sample size. For each variant, adjust the learning rate to each value in the set {0.1, 0.25, 0.5, 1};

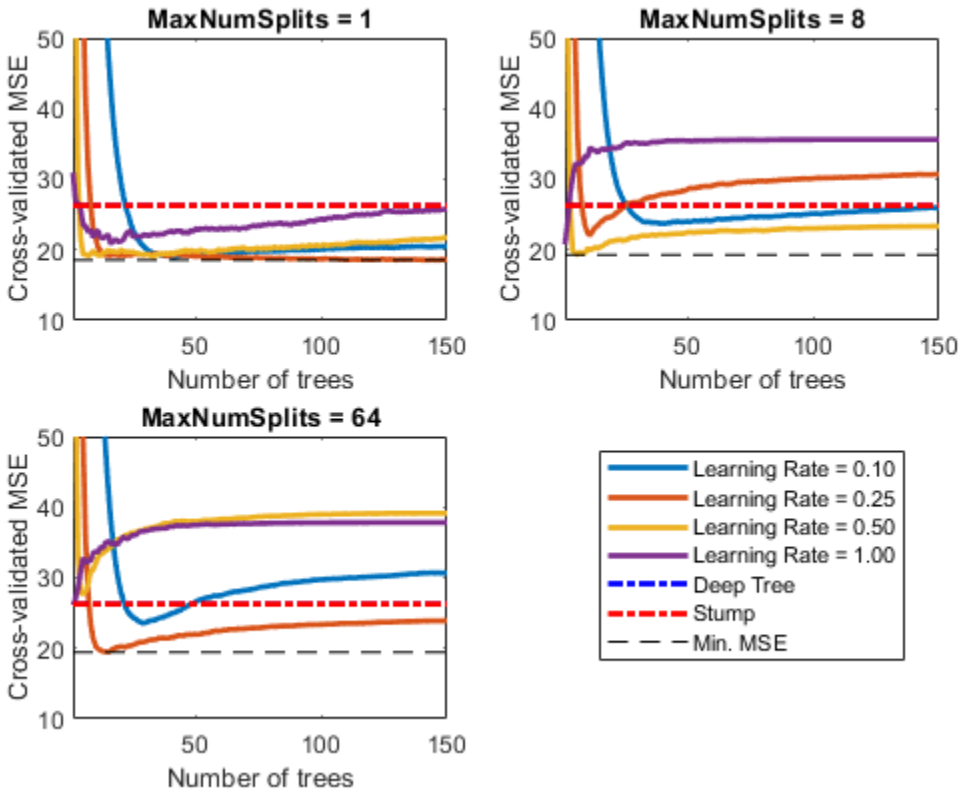
```
n = size(X,1);
m = floor(log2(n - 1));
lr = [0.1 0.25 0.5 1];
maxNumSplits = 2.^(0:m);
numTrees = 150;
Mdl = cell(numel(maxNumSplits),numel(lr));
rng(1); % For reproducibility
for k = 1:numel(lr);
    for j = 1:numel(maxNumSplits);
        t = templateTree('MaxNumSplits',maxNumSplits(j),'Surrogate','on');
        Mdl{j,k} = fitensemble(X,Y,'LSBoost',numTrees,t,...
            'Type','regression','KFold',5,'LearnRate',lr(k));
    end;
end;
```

Compute the cross-validated MSE for each ensemble.

```
kflAll = @(x)kfoldLoss(x,'Mode','cumulative');
errorCell = cellfun(kflAll,Mdl,'Uniform',false);
error = reshape(cell2mat(errorCell),[numTrees numel(maxNumSplits) numel(lr)]);
errorDeep = kfoldLoss(MdlDeep);
errorStump = kfoldLoss(MdlStump);
```

Plot how the cross-validated MSE behaves as the number of trees in the ensemble increases for a few of the ensembles, the deep tree, and the stump. Plot the curves with respect to learning rate in the same plot, and plot separate plots for varying tree complexities. Choose a subset of tree complexity levels.

```
mnsPlot = [1 round(numel(maxNumSplits)/2) numel(maxNumSplits)];
figure;
for k = 1:3;
    subplot(2,2,k);
    plot(squeeze(error(:,mnsPlot(k),:)),'LineWidth',2);
    axis tight;
    hold on;
    h = gca;
    plot(h.XLim,[errorDeep errorDeep],'-.b','LineWidth',2);
    plot(h.XLim,[errorStump errorStump],'-.r','LineWidth',2);
    plot(h.XLim,min(min(error(:,mnsPlot(k),:))).*[1 1],'--k');
    h.YLim = [10 50];
    xlabel 'Number of trees';
    ylabel 'Cross-validated MSE';
    title(sprintf('MaxNumSplits = %0.3g', maxNumSplits(mnsPlot(k))));
    hold off;
end;
hL = legend([cellstr(num2str(lr','Learning Rate = %0.2f'));...
        'Deep Tree';'Stump';'Min. MSE']);
hL.Position(1) = 0.6;
```



Each curve contains a minimum cross-validated MSE occurring at the optimal number of trees in the ensemble.

Identify the maximum number of splits, number of trees, and learning rate that yields the lowest MSE overall.

```
[minErr,minErrIdxLin] = min(error(:));
[idxNumTrees,idxMNS,idxLR] = ind2sub(size(error),minErrIdxLin);

fprintf('\nMin. MSE = %0.5f',minErr)

Min. MSE = 18.42979

fprintf('\nOptimal Parameter Values:\nNum. Trees = %d',idxNumTrees);

Optimal Parameter Values:
Num. Trees = 1

fprintf('\nMaxNumSplits = %d\nLearning Rate = %0.2f\n',...
    maxNumSplits(idxMNS),lr(idxLR))

MaxNumSplits = 4
Learning Rate = 1.00
```

For a different approach to optimizing this ensemble, see “Optimize a Boosted Regression Ensemble” on page 10-63.

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable and you want to use all remaining variables as predictors, then specify the response variable using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use a subset of the remaining variables only as predictors, then specify a formula using `formula`.
- If `Tbl` does not contain the response variable, then specify the response data using `Y`. The length of response variable and the number of rows of `Tbl` must be equal.

Note To save memory and execution time, supply `X` and `Y` instead of `Tbl`.

Data Types: `table`

`ResponseVarName` — Response variable name

name of response variable in `Tbl`

Response variable name, specified as the name of the response variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl.Y` is the response variable, then specify `ResponseVarName` as `'Y'`. Otherwise, `fitensemble` treats all columns of `Tbl` as predictor variables.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

For classification, you can specify the order of the classes using the `ClassNames` name-value pair argument. Otherwise, `fitensemble` determines the class order, and stores it in the `Mdl.ClassNames`.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`X` — Predictor data
numeric matrix

Predictor data, specified as numeric matrix.

Each row corresponds to one observation, and each column corresponds to one predictor variable.

The length of `Y` and the number of rows of `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

`Y` — Response data
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Response data, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each entry in `Y` is the response to or label for the observation in the corresponding row of `X` or `Tbl`. The length of `Y` and the number of rows of `X` or `Tbl` must be equal. If the response variable is a character array, then each element must correspond to one row of the array.

- For classification, `Y` can be any of the supported data types. You can specify the order of the classes using the `ClassNames` name-value pair argument. Otherwise, `fitensemble` determines the class order, and stores it in the `Mdl.ClassNames`.

- For regression, `Y` must be a numeric column vector.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Method` — Ensemble aggregation method
`'AdaBoostM1'` | `'LogitBoost'` | `'GentleBoost'` | `'RUSBoost'` | `'Subspace'` | `'Bag'` | `'AdaBoostM2'` | `'LSBoost'` | ...

Ensemble aggregation method, specified as one of the method names in this list.

- For classification with two classes:
  - `'AdaBoostM1'`
  - `'LogitBoost'`
  - `'GentleBoost'`
  - `'RobustBoost'` (requires Optimization Toolbox)
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For classification with three or more classes:
  - `'AdaBoostM2'`
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For regression:
  - `'LSBoost'`
  - `'Bag'`

If you specify `'Method','Bag'`, then specify the problem type using the `Type` name-value pair argument, because you can specify `'Bag'` for classification and regression problems.

For details about ensemble aggregation algorithms and examples, see “Ensemble Algorithms” on page 18-39 and “Choose an Applicable Ensemble Aggregation Method” on page 18-32.

`NLearn` — Number of ensemble learning cycles
positive integer | `'AllPredictorCombinations'`

Number of ensemble learning cycles, specified as a positive integer or `'AllPredictorCombinations'`.

- If you specify a positive integer, then, at every learning cycle, the software trains one weak learner for every template object in `Learners`. Consequently, the software trains `NLearn*numel(Learners)` learners.

- If you specify `'AllPredictorCombinations'`, then set `Method` to `'Subspace'` and specify one learner only in `Learners`. With these settings, the software trains learners for all possible combinations of predictors taken `NPredToSample` at a time. Consequently, the software trains `nchoosek(size(X,2),NPredToSample)` learners.

The software composes the ensemble using all trained learners and stores them in `Mdl.Trained`.

For more details, see “Tips” on page 33-2274.

Data Types: `single` | `double` | `char` | `string`

`Learners` — Weak learners to use in ensemble

weak-learner name | weak-learner template object | cell vector of weak-learner template objects

Weak learners to use in the ensemble, specified as a weak-learner name, weak-learner template object, or cell array of weak-learner template objects.

Weak Learner	Weak-Learner Name	Template Object Creation Function	Method Settings
Discriminant analysis	`'Discriminant'`	`templateDiscriminant`	Recommended for `'Subspace'`
k nearest neighbors	`'KNN'`	`templateKNN`	For `'Subspace'` only
Decision tree	`'Tree'`	`templateTree`	All methods except `'Subspace'`

For more details, see `NLearn` and “Tips” on page 33-2274.

Example: For an ensemble composed of two types of classification trees, supply `{t1 t2}`, where `t1` and `t2` are classification tree templates.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'CrossVal','on','LearnRate',0.05` specifies to implement 10-fold cross-validation and to use `0.05` as the learning rate.

General Ensemble Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitensemble` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

Specification of `'CategoricalPredictors'` is appropriate if:

- `'Learners'` specifies tree learners.
- `'Learners'` specifies *k*-nearest learners where all predictors are categorical.

Each learner identifies and treats categorical predictors in the same way as the fitting function corresponding to the learner. See `'CategoricalPredictors'` of `fitcknn` for *k*-nearest learners and `'CategoricalPredictors'` of `fitctree` for tree learners.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`NPrint` — Printout frequency
`'off'` (default) | positive integer

Printout frequency, specified as the comma-separated pair consisting of `'NPrint'` and a positive integer or `'off'`.

To track the number of *weak learners* or *folds* that `fitensemble` trained so far, specify a positive integer. That is, if you specify the positive integer *m*:

- Without also specifying any cross-validation option (for example, `CrossVal`), then `fitensemble` displays a message to the command line every time it completes training *m* weak learners.
- And a cross-validation option, then `fitensemble` displays a message to the command line every time it finishes training *m* folds.

If you specify `'off'`, then `fitensemble` does not display a message when it completes training weak learners.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Example: `'NPrint',5`

Data Types: `single` | `double` | `char` | `string`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitensemble` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`Type` — Supervised learning type
`'classification'` | `'regression'`

Supervised learning type, specified as the comma-separated pair consisting of `'Type'` and `'classification'` or `'regression'`.

- If `Method` is `'bag'`, then the supervised learning type is ambiguous. Therefore, specify `Type` when bagging.
- Otherwise, the value of `Method` determines the supervised learning type.

Example: `'Type','classification'`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross-validate later by passing `Mdl` to `crossval` or `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Other Classification or Regression Options

`ClassNames` — Names of classes to use for training
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.
- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure. If you specify:

- The square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, also specify the `ClassNames` name-value pair argument.
- The structure `S`, then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default is `ones(`*K*`) - eye(`*K*`)`, where *K* is the number of distinct classes.

Note `fitensemble` uses `Cost` to adjust the prior class probabilities specified in `Prior`. Then, `fitensemble` uses the adjusted prior probabilities for training and resets the cost matrix to its default.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `double` | `single` | `struct`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and a value in this table.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to `Mdl.ClassNames` or specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.

Value	Description
structure array	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

`fitensemble` normalizes the prior probabilities in `Prior` to sum to 1.

Example: `struct('ClassNames', {{'setosa','versicolor','virginica'}},'ClassProbs',1:3)`

Data Types: `char` | `string` | `double` | `single` | `struct`

`Weights` — Observation weights
numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

By default, `Weights` is `ones(`*n*`,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Sampling Options for Boosting Methods and Bagging

`FResample` — Fraction of training set to resample
`1` (default) | positive scalar in (0,1]

Fraction of the training set to resample for every weak learner, specified as the comma-separated pair consisting of `'FResample'` and a positive scalar in (0,1].

To use `'FResample'`, specify `'bag'` for `Method` or set `Resample` to `'on'`.

Example: `'FResample',0.75`

Data Types: `single` | `double`

`Replace` — Flag indicating to sample with replacement
`'on'` (default) | `'off'`

Flag indicating sampling with replacement, specified as the comma-separated pair consisting of `'Replace'` and `'off'` or `'on'`.

- For `'on'`, the software samples the training observations with replacement.
- For `'off'`, the software samples the training observations without replacement. If you set `Resample` to `'on'`, then the software samples training observations assuming uniform weights. If you also specify a boosting method, then the software boosts by reweighting observations.

Unless you set `Method` to `'bag'` or set `Resample` to `'on'`, `Replace` has no effect.

Example: `'Replace','off'`

`Resample` — Flag indicating to resample

`'off'` | `'on'`

Flag indicating to resample, specified as the comma-separated pair consisting of `'Resample'` and `'off'` or `'on'`.

- If `Method` is a boosting method, then:
  - `'Resample','on'` specifies to sample training observations using updated weights as the multinomial sampling probabilities.
  - `'Resample','off'`(default) specifies to reweight observations at every learning iteration.
- If `Method` is `'bag'`, then `'Resample'` must be `'on'`. The software resamples a fraction of the training observations (see `FResample`) with or without replacement (see `Replace`).

If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of 1 for `FResample`.

AdaBoostM1, AdaBoostM2, LogitBoost, GentleBoost, and LSBoost Method Options

`LearnRate` — Learning rate for shrinkage

`1` (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

RUSBoost Method Options

`LearnRate` — Learning rate for shrinkage

`1` (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

`RatioToSmallest` — Sampling proportion with respect to lowest-represented class
positive numeric scalar | numeric vector of positive values

Sampling proportion with respect to the lowest-represented class, specified as the comma-separated pair consisting of `'RatioToSmallest'` and a numeric scalar or numeric vector of positive values with length equal to the number of distinct classes in the training data.

Suppose that there are K classes in the training data and the lowest-represented class has m observations in the training data.

- If you specify the positive numeric scalar s, then `fitensemble` samples $s*m$ observations from each class, that is, it uses the same sampling proportion for each class. For more details, see "Algorithms" on page 33-2275.
- If you specify the numeric vector `[s1,s2,...,sK]`, then `fitensemble` samples $si*m$ observations from class i, $i = 1,...,K$. The elements of `RatioToSmallest` correspond to the order of the class names specified using `ClassNames` (see "Tips" on page 33-2274).

The default value is `ones(K,1)`, which specifies to sample m observations from each class.

Example: `'RatioToSmallest',[2,1]`

Data Types: `single` | `double`

LPBoost and TotalBoost Method Options

`MarginPrecision` — Margin precision to control convergence speed
`0.1` (default) | numeric scalar in [0,1]

Margin precision to control convergence speed, specified as the comma-separated pair consisting of `'MarginPrecision'` and a numeric scalar in the interval [0,1]. `MarginPrecision` affects the number of boosting iterations required for convergence.

Tip To train an ensemble using many learners, specify a small value for `MarginPrecision`. For training using a few learners, specify a large value.

Example: `'MarginPrecision',0.5`

Data Types: `single` | `double`

RobustBoost Method Options

`RobustErrorGoal` — Target classification error
`0.1` (default) | nonnegative numeric scalar

Target classification error, specified as the comma-separated pair consisting of `'RobustErrorGoal'` and a nonnegative numeric scalar. The upper bound on possible values depends on the values of `RobustMarginSigma` and `RobustMaxMargin`. However, the upper bound cannot exceed `1`.

Tip For a particular training set, usually there is an optimal range for `RobustErrorGoal`. If you set it too low or too high, then the software can produce a model with poor classification accuracy. Try cross-validating to search for the appropriate value.

Example: `'RobustErrorGoal',0.05`

Data Types: `single` | `double`

`RobustMarginSigma` — Classification margin distribution spread
`0.1` (default) | positive numeric scalar

Classification margin distribution spread over the training data, specified as the comma-separated pair consisting of `'RobustMarginSigma'` and a positive numeric scalar. Before specifying `RobustMarginSigma`, consult the literature on `RobustBoost`, for example, [19].

Example: `'RobustMarginSigma',0.5`

Data Types: `single` | `double`

`RobustMaxMargin` — Maximal classification margin
`0` (default) | nonnegative numeric scalar

Maximal classification margin in the training data, specified as the comma-separated pair consisting of `'RobustMaxMargin'` and a nonnegative numeric scalar. The software minimizes the number of observations in the training data having classification margins below `RobustMaxMargin`.

Example: `'RobustMaxMargin',1`

Data Types: `single` | `double`

Random Subspace Method Options

`NPredToSample` — Number of predictors to sample
`1` (default) | positive integer

Number of predictors to sample for each random subspace learner, specified as the comma-separated pair consisting of `'NPredToSample'` and a positive integer in the interval 1,...,*p*, where *p* is the number of predictor variables (`size(X,2)` or `size(Tbl,2)`).

Data Types: `single` | `double`

Output Arguments

`Mdl` — Trained ensemble model
`ClassificationBaggedEnsemble` model object | `ClassificationEnsemble` model object | `ClassificationPartitionedEnsemble` cross-validated model object | `RegressionBaggedEnsemble` model object | `RegressionEnsemble` model object | `RegressionPartitionedEnsemble` cross-validated model object

Trained ensemble model, returned as one of the model objects in this table.

Model Object	Type Setting	Specify Any Cross-Validation Options?	Method Setting	Resample Setting
`ClassificationBaggedEnsemble`	`'classification'`	No	`'Bag'`	`'on'`
`ClassificationEnsemble`	`'classification'`	No	Any ensemble-aggregation method for classification	`'off'`

Model Object	Type Setting	Specify Any Cross-Validation Options?	Method Setting	Resample Setting
ClassificationPartitionedEnsemble	'classification'	Yes	Any classification ensemble-aggregation method	'off' or 'on'
RegressionBaggedEnsemble	'regression'	No	'Bag'	'on'
RegressionEnsemble	'regression'	No	'LSBoost'	'off'
RegressionPartitionedEnsemble	'regression'	Yes	'LSBoost' or 'Bag'	'off' or 'on'

The name-value pair arguments that control cross-validation are `CrossVal`, `Holdout`, `KFold`, `Leaveout`, and `CVPartition`.

To reference properties of `Mdl`, use dot notation. For example, to access or display the cell vector of weak learner model objects for an ensemble that has not been cross-validated, enter `Mdl.Trained` at the command line.

Tips

- `NLearn` can vary from a few dozen to a few thousand. Usually, an ensemble with good predictive power requires from a few hundred to a few thousand weak learners. However, you do not have to train an ensemble for that many cycles at once. You can start by growing a few dozen learners, inspect the ensemble performance and then, if necessary, train more weak learners using `resume` for classification problems, or `resume` for regression problems.
- Ensemble performance depends on the ensemble setting and the setting of the weak learners. That is, if you specify weak learners with default parameters, then the ensemble can perform poorly. Therefore, like ensemble settings, it is good practice to adjust the parameters of the weak learners using templates, and to choose values that minimize generalization error.
- If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of `1` for `FResample`.
- In classification problems (that is, `Type` is `'classification'`):
  - If the ensemble-aggregation method (`Method`) is `'bag'` and:
    - The misclassification cost (`Cost`) is highly imbalanced, then, for in-bag samples, the software oversamples unique observations from the class that has a large penalty.
    - The class prior probabilities (`Prior`) are highly skewed, the software oversamples unique observations from the class that has a large prior probability.

    For smaller sample sizes, these combinations can result in a low relative frequency of out-of-bag observations from the class that has a large penalty or prior probability. Consequently, the estimated out-of-bag error is highly variable and it can be difficult to interpret. To avoid large estimated out-of-bag error variances, particularly for small sample sizes, set a more balanced misclassification cost matrix using `Cost` or a less skewed prior probability vector using `Prior`.
  - Because the order of some input and output arguments correspond to the distinct classes in the training data, it is good practice to specify the class order using the `ClassNames` name-value pair argument.

- To determine the class order quickly, remove all observations from the training data that are unclassified (that is, have a missing label), obtain and display an array of all the distinct classes, and then specify the array for `ClassNames`. For example, suppose the response variable (`Y`) is a cell array of labels. This code specifies the class order in the variable `classNames`.

  ```
  Ycat = categorical(Y);
  classNames = categories(Ycat)
  ```

  `categorical` assigns `<undefined>` to unclassified observations and `categories` excludes `<undefined>` from its output. Therefore, if you use this code for cell arrays of labels or similar code for categorical arrays, then you do not have to remove observations with missing labels to obtain a list of the distinct classes.
- To specify that the class order from lowest-represented label to most-represented, then quickly determine the class order (as in the previous bullet), but arrange the classes in the list by frequency before passing the list to `ClassNames`. Following from the previous example, this code specifies the class order from lowest- to most-represented in `classNamesLH`.

  ```
  Ycat = categorical(Y);
  classNames = categories(Ycat);
  freq = countcats(Ycat);
  [~,idx] = sort(freq);
  classNamesLH = classNames(idx);
  ```

Algorithms

- For details of ensemble-aggregation algorithms, see “Ensemble Algorithms” on page 18-39.
- If you specify `Method` to be a boosting algorithm and `Learners` to be decision trees, then the software grows stumps by default. A decision stump is one root node connected to two terminal, leaf nodes. You can adjust tree depth by specifying the `MaxNumSplits`, `MinLeafSize`, and `MinParentSize` name-value pair arguments using `templateTree`.
- `fitensemble` generates in-bag samples by oversampling classes with large misclassification costs and undersampling classes with small misclassification costs. Consequently, out-of-bag samples have fewer observations from classes with large misclassification costs and more observations from classes with small misclassification costs. If you train a classification ensemble using a small data set and a highly skewed cost matrix, then the number of out-of-bag observations per class can be low. Therefore, the estimated out-of-bag error can have a large variance and can be difficult to interpret. The same phenomenon can occur for classes with large prior probabilities.
- For the RUSBoost ensemble-aggregation method (`Method`), the name-value pair argument `RatioToSmallest` specifies the sampling proportion for each class with respect to the lowest-represented class. For example, suppose that there are two classes in the training data: A and B. A have 100 observations and B have 10 observations. Also, suppose that the lowest-represented class has m observations in the training data.

  - If you set `'RatioToSmallest',2`, then $s*m = 2*10 = 20$. Consequently, `fitensemble` trains every learner using 20 observations from class A and 20 observations from class B. If you set `'RatioToSmallest',[2 2]`, then you obtain the same result.
  - If you set `'RatioToSmallest',[2,1]`, then $s1*m = 2*10 = 20$ and $s2*m = 1*10 = 10$. Consequently, `fitensemble` trains every learner using 20 observations from class A and 10 observations from class B.

- For ensembles of decision trees, and for dual-core systems and above, `fitensemble` parallelizes training using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L. "Bagging Predictors." *Machine Learning*. Vol. 26, pp. 123–140, 1996.

[2] Breiman, L. "Random Forests." *Machine Learning*. Vol. 45, pp. 5–32, 2001.

[3] Freund, Y. "A more robust boosting algorithm." *arXiv:0905.2138v1*, 2009.

[4] Freund, Y. and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *J. of Computer and System Sciences*, Vol. 55, pp. 119–139, 1997.

[5] Friedman, J. "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.

[6] Friedman, J., T. Hastie, and R. Tibshirani. "Additive logistic regression: A statistical view of boosting." *Annals of Statistics*, Vol. 28, No. 2, pp. 337–407, 2000.

[7] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning* section edition, Springer, New York, 2008.

[8] Ho, T. K. "The random subspace method for constructing decision forests." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832–844, 1998.

[9] Schapire, R. E., Y. Freund, P. Bartlett, and W.S. Lee. "Boosting the margin: A new explanation for the effectiveness of voting methods." *Annals of Statistics*, Vol. 26, No. 5, pp. 1651–1686, 1998.

[10] Seiffert, C., T. Khoshgoftaar, J. Hulse, and A. Napolitano. "RUSBoost: Improving classification performance when training data is skewed." *19th International Conference on Pattern Recognition*, pp. 1–4, 2008.

[11] Warmuth, M., J. Liao, and G. Ratsch. "Totally corrective boosting algorithms that maximize the margin." *Proc. 23rd Int'l. Conf. on Machine Learning, ACM*, New York, pp. 1001–1008, 2006.

See Also

`ClassificationBaggedEnsemble` | `ClassificationEnsemble` | `ClassificationPartitionedEnsemble` | `RegressionBaggedEnsemble` | `RegressionEnsemble` | `RegressionPartitionedEnsemble` | `templateDiscriminant` | `templateKNN` | `templateTree`

Topics

"Supervised Learning Workflow and Algorithms" on page 18-3
"Framework for Ensemble Learning" on page 18-31

Introduced in R2011a

fitnlm

Fit nonlinear regression model

Syntax

```
mdl = fitnlm(tbl,modelfun,beta0)
mdl = fitnlm(X,y,modelfun,beta0)
mdl = fitnlm(___,modelfun,beta0,Name,Value)
```

Description

`mdl = fitnlm(tbl,modelfun,beta0)` fits the model specified by `modelfun` to variables in the table or dataset array `tbl`, and returns the nonlinear model `mdl`.

`fitnlm` estimates model coefficients using an iterative procedure starting from the initial values in `beta0`.

`mdl = fitnlm(X,y,modelfun,beta0)` fits a nonlinear regression model using the column vector `y` as a response variable and the columns of the matrix `X` as predictor variables.

`mdl = fitnlm(___,modelfun,beta0,Name,Value)` fits a nonlinear regression model with additional options specified by one or more `Name,Value` pair arguments.

Examples

Nonlinear Model from Table

Create a nonlinear model for auto mileage based on the `carbig` data.

Load the data and create a nonlinear model.

```
load carbig
tbl = table(Horsepower,Weight,MPG);
modelfun = @(b,x)b(1) + b(2)*x(:,1).^b(3) + ...
    b(4)*x(:,2).^b(5);
beta0 = [-50 500 -1 500 -1];
mdl = fitnlm(tbl,modelfun,beta0)
```

```
mdl =
Nonlinear regression model:
    MPG ~ b1 + b2*Horsepower^b3 + b4*Weight^b5

Estimated Coefficients:
          Estimate      SE        tStat       pValue
          ________    _______    ________    ________

    b1     -49.383     119.97    -0.41164     0.68083
    b2      376.43     567.05     0.66384     0.50719
    b3    -0.78193    0.47168     -1.6578    0.098177
    b4      422.37     776.02     0.54428     0.58656
```

```
    b5    -0.24127    0.48325    -0.49926     0.61788


Number of observations: 392, Error degrees of freedom: 387
Root Mean Squared Error: 3.96
R-Squared: 0.745,  Adjusted R-Squared 0.743
F-statistic vs. constant model: 283, p-value = 1.79e-113
```

Nonlinear Model from Matrix Data

Create a nonlinear model for auto mileage based on the `carbig` data.

Load the data and create a nonlinear model.

```
load carbig
X = [Horsepower,Weight];
y = MPG;
modelfun = @(b,x)b(1) + b(2)*x(:,1).^b(3) + ...
    b(4)*x(:,2).^b(5);
beta0 = [-50 500 -1 500 -1];
mdl = fitnlm(X,y,modelfun,beta0)
```

```
mdl =
Nonlinear regression model:
    y ~ b1 + b2*x1^b3 + b4*x2^b5

Estimated Coefficients:
          Estimate      SE        tStat       pValue
          ________    _______    ________    ________

    b1     -49.383     119.97    -0.41164     0.68083
    b2      376.43     567.05     0.66384     0.50719
    b3    -0.78193    0.47168     -1.6578    0.098177
    b4      422.37     776.02     0.54428     0.58656
    b5    -0.24127    0.48325    -0.49926     0.61788


Number of observations: 392, Error degrees of freedom: 387
Root Mean Squared Error: 3.96
R-Squared: 0.745,  Adjusted R-Squared 0.743
F-statistic vs. constant model: 283, p-value = 1.79e-113
```

Adjust Fitting Options in Nonlinear Model

Create a nonlinear model for auto mileage based on the `carbig` data. Strive for more accuracy by lowering the `TolFun` option, and observe the iterations by setting the `Display` option.

Load the data and create a nonlinear model.

```
load carbig
X = [Horsepower,Weight];
y = MPG;
modelfun = @(b,x)b(1) + b(2)*x(:,1).^b(3) + ...
```

```
    b(4)*x(:,2).^b(5);
beta0 = [-50 500 -1 500 -1];
```

Create options to lower `TolFun` and to report iterative display, and create a model using the options.

```
opts = statset('Display','iter','TolFun',1e-10);
mdl = fitnlm(X,y,modelfun,beta0,'Options',opts);
```

```
                                 Norm of         Norm of
   Iteration             SSE     Gradient          Step
  ------------------------------------------------------------
           0     1.82248e+06
           1          678600          788810         1691.07
           2          616716     6.12739e+06         45.4738
           3          249831      3.9532e+06         293.557
           4           17675          361544         369.284
           5         11746.6         69670.5         169.079
           6         7242.22          343738         394.822
           7         6250.32          159719         452.941
           8         6172.87         91622.9         268.674
           9            6077         6957.44         100.208
          10         6076.34          6370.4         88.1905
          11         6075.75         5199.08         77.9694
          12          6075.3         4646.61          69.764
          13         6074.91         4235.96         62.9114
          14         6074.55         3885.28         57.0647
          15         6074.23          3571.1         52.0036
          16         6073.93         3286.48         47.5795
          17         6073.66         3028.34         43.6844
          18          6073.4         2794.31         40.2352
          19         6073.17         2582.15         37.1663
          20         6072.95         2389.68         34.4243
          21         6072.74         2214.84          31.965
          22         6072.55         2055.78         29.7516
          23         6072.37         1910.83          27.753
          24         6072.21         1778.51         25.9428
          25         6072.05          1657.5         24.2986
          26          6071.9         1546.65         22.8011
          27         6071.76         1444.93         21.4338
          28         6071.63         1351.44         20.1822
          29         6071.51         1265.39         19.0339
          30         6071.39         1186.06          17.978
          31         6071.28         1112.83         17.0052
          32         6071.17         1045.13          16.107
          33         6071.07         982.465         15.2762
          34         6070.98         924.389         14.5063
          35         6070.89         870.498         13.7916
          36          6070.8         820.434          13.127
          37         6070.72         773.872         12.5081
          38         6070.64         730.521         11.9307
          39         6070.57         690.117         11.3914
          40          6070.5         652.422          10.887
          41         6070.43         617.219         10.4144
          42         6070.37         584.315         9.97114
          43         6070.31          553.53         9.55489
          44         6070.25         524.703          9.1635
          45         6070.19         497.686         8.79506
```

46	6070.14	472.345	8.44785
47	6070.08	448.557	8.12028
48	6070.03	426.21	7.81092
49	6069.99	405.201	7.51845
50	6069.94	385.435	7.2417
51	6069.9	366.825	6.97956
52	6069.85	349.293	6.73104
53	6069.81	332.764	6.49523
54	6069.77	317.171	6.27127
55	6069.74	302.452	6.0584
56	6069.7	288.55	5.85591
57	6069.66	275.411	5.66315
58	6069.63	262.986	5.47949
59	6069.6	251.23	5.3044
60	6069.57	240.1	5.13734
61	6069.54	229.558	4.97784
62	6069.51	219.567	4.82545
63	6069.48	210.094	4.67977
64	6069.45	201.108	4.5404
65	6069.43	192.578	4.407
66	6069.4	184.479	4.27923
67	6069.38	176.785	4.15678
68	6069.35	169.472	4.03935
69	6069.33	162.518	3.9267
70	6069.31	155.903	3.81855
71	6069.29	149.608	3.71468
72	6069.26	143.615	3.61486
73	6069.24	137.907	3.51889
74	6069.22	132.468	3.42658
75	6069.21	127.283	3.33774
76	6069.19	122.339	3.25221
77	6069.17	117.623	3.16981
78	6069.15	113.123	3.09041
79	6069.14	108.827	3.01386
80	6069.12	104.725	2.94002
81	6069.1	100.806	2.86877
82	6069.09	97.0611	2.8
83	6069.07	93.4814	2.73358
84	6069.06	90.0584	2.66942
85	6069.05	86.7842	2.60741
86	6069.03	83.6513	2.54745
87	6069.02	80.6528	2.48947
88	6069.01	77.7821	2.43338
89	6068.99	75.0327	2.37908
90	6068.98	72.399	2.32652
91	6068.97	69.8752	2.27561
92	6068.96	67.4561	2.22629
93	6068.95	65.1367	2.17849
94	6068.94	62.9123	2.13216
95	6068.93	60.7784	2.08723
96	6068.92	58.7308	2.04364
97	6068.91	56.7655	2.00135
98	6068.9	54.8787	1.9603
99	6068.89	4349.28	18.1917
100	6068.77	2416.27	14.4439
101	6068.71	1721.26	12.1305
102	6068.66	1228.78	10.289
103	6068.63	884.002	8.82019

```
          104          6068.6         639.615         7.62745
          105         6068.58          464.84         6.64627
          106         6068.56         338.878         5.82964
          107         6068.55         247.508         5.14297
          108         6068.54         180.879         4.56032
          109         6068.53         132.084         4.06194
          110         6068.52         96.2342         3.63255
          111         6068.51         69.8363          3.2602
          112         6068.51         50.3734         2.93541
          113          6068.5         36.0206         2.65062
          114          6068.5         25.4451         2.39969
          115         6068.49         17.6692         2.17764
          116         6068.49         1027.39         14.0164
          117         6068.48         544.039         5.31369
          118         6068.48         94.0567         2.86662
          119         6068.48         113.636         3.73504
          120         6068.48         0.51821         1.37054
          121         6068.48         4.59524        0.912906
          122         6068.48         1.56406         0.62935
          123         6068.48         1.13837         0.43259
          124         6068.48        0.296041        0.297545
Iterations terminated: relative change in SSE less than OPTIONS.TolFun
```

Specify Nonlinear Regression Using Model Name Syntax

Specify a nonlinear regression model for estimation using a function handle or model syntax.

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Use a function handle to specify the Hougen-Watson model for the rate data.

```
mdl = fitnlm(X,y,@hougen,beta0)
```

```
mdl =
Nonlinear regression model:
    y ~ hougen(b,X)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1      1.2526     0.86701    1.4447    0.18654
    b2    0.062776    0.043561    1.4411    0.18753
    b3    0.040048    0.030885    1.2967    0.23089
    b4     0.11242    0.075157    1.4957    0.17309
    b5      1.1914     0.83671    1.4239     0.1923


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 0.193
```

```
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 3.91e+03, p-value = 2.54e-13
```

Alternatively, you can use an expression to specify the Hougen-Watson model for the rate data.

```
myfun = 'y~(b1*x2-x3/b5)/(1+b2*x1+b3*x2+b4*x3)';
mdl2 = fitnlm(X,y,myfun,beta0)
```

```
mdl2 =
Nonlinear regression model:
    y ~ (b1*x2 - x3/b5)/(1 + b2*x1 + b3*x2 + b4*x3)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1      1.2526     0.86701    1.4447    0.18654
    b2    0.062776    0.043561    1.4411    0.18753
    b3    0.040048    0.030885    1.2967    0.23089
    b4     0.11242    0.075157    1.4957    0.17309
    b5      1.1914     0.83671    1.4239     0.1923


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 0.193
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 3.91e+03, p-value = 2.54e-13
```

Estimate Nonlinear Regression Using Robust Fitting Options

Generate sample data from the nonlinear regression model

$$y = b_1 + b_2\exp(-b_3x) + \varepsilon,$$

where b_1, b_2, and b_3 are coefficients, and the error term is normally distributed with mean 0 and standard deviation 0.5.

```
modelfun = @(b,x)(b(1)+b(2)*exp(-b(3)*x));

rng('default') % for reproducibility
b = [1;3;2];
x = exprnd(2,100,1);
y = modelfun(b,x) + normrnd(0,0.5,100,1);
```

Set robust fitting options.

```
opts = statset('nlinfit');
opts.RobustWgtFun = 'bisquare';
```

Fit the nonlinear model using the robust fitting options. Here, use an expression to specify the model.

```
b0 = [2;2;2];
modelstr = 'y ~ b1 + b2*exp(-b3*x)';

mdl = fitnlm(x,y,modelstr,b0,'Options',opts)
```

```
mdl =
Nonlinear regression model (robust fit):
    y ~ b1 + b2*exp( - b3*x)

Estimated Coefficients:
          Estimate      SE       tStat       pValue
          ________    _______    ______    __________

    b1     1.0218     0.07202    14.188    2.1344e-25
    b2     3.6619     0.25429    14.401     7.974e-26
    b3     2.9732     0.38496    7.7232    1.0346e-11


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 0.501
R-Squared: 0.807,  Adjusted R-Squared 0.803
F-statistic vs. constant model: 203, p-value = 2.34e-35
```

Fit Nonlinear Regression Model Using Weights Function Handle

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Specify a function handle for observation weights. The function accepts the model fitted values as input, and returns a vector of weights.

```
 a = 1; b = 1;
 weights = @(yhat) 1./((a + b*abs(yhat)).^2);
```

Fit the Hougen-Watson model to the rate data using the specified observation weights function.

```
mdl = fitnlm(X,y,@hougen,beta0,'Weights',weights)
```

```
mdl =
Nonlinear regression model:
    y ~ hougen(b,X)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1     0.83085     0.58224     1.427    0.19142
    b2     0.04095    0.029663    1.3805    0.20477
    b3    0.025063    0.019673     1.274    0.23842
    b4    0.080053    0.057812    1.3847    0.20353
    b5      1.8261       1.281    1.4256    0.19183


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 0.037
R-Squared: 0.998,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 1.14e+03, p-value = 3.49e-11
```

Nonlinear Regression Model Using Nonconstant Error Model

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the combined error variance model.

```
mdl = fitnlm(X,y,@hougen,beta0,'ErrorModel','combined')
```

```
mdl =
Nonlinear regression model:
    y ~ hougen(b,X)

Estimated Coefficients:
          Estimate       SE       tStat     pValue
          ________    ________    ______    _______

    b1      1.2526     0.86702    1.4447    0.18654
    b2    0.062776    0.043561    1.4411    0.18753
    b3    0.040048    0.030885    1.2967    0.23089
    b4     0.11242    0.075158    1.4957    0.17309
    b5      1.1914     0.83671    1.4239     0.1923


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 1.27
R-Squared: 0.999,  Adjusted R-Squared 0.998
F-statistic vs. zero model: 3.91e+03, p-value = 2.54e-13
```

Input Arguments

`tbl` — Input data

table | dataset array

Input data including predictor and response variables, specified as a table or dataset array. The predictor variables and response variable must be numeric.

- If you specify `modelfun` using a formula, the model specification in the formula specifies the predictor and response variables.
- If you specify `modelfun` using a function handle, the last variable is the response variable and the others are the predictor variables, by default. You can set a different column as the response variable by using the `ResponseVar` name-value pair argument. To select a subset of the columns as predictors, use the `PredictorVars` name-value pair argument.

The variable names in a table do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot specify `modelfun` using a formula.

You can verify the variable names in `tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `table`

`X` — Predictor variables

matrix

Predictor variables, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

Data Types: `single` | `double`

`y` — Response variable

vector

Response variable, specified as an *n*-by-1 vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`.

Data Types: `single` | `double`

`modelfun` — Functional form of the model

function handle | character vector or string scalar formula in the form `'y ~ f(b1,b2,...,bj,x1,x2,...,xk)'`

Functional form of the model, specified as either of the following.

- Function handle `@modelfun` or `@(b,x)modelfun`, where
  - `b` is a coefficient vector with the same number of elements as `beta0`.
  - `x` is a matrix with the same number of columns as `X` or the number of predictor variable columns of `tbl`.

  `modelfun(b,x)` returns a column vector that contains the same number of rows as `x`. Each row of the vector is the result of evaluating `modelfun` on the corresponding row of `x`. In other words, `modelfun` is a vectorized function, one that operates on all data rows and returns all evaluations in one function call. `modelfun` should return real numbers to obtain meaningful coefficients.
- Character vector or string scalar formula in the form `'y ~ f(b1,b2,...,bj,x1,x2,...,xk)'`, where `f` represents a scalar function of the scalar coefficient variables `b1`,...,`bj` and the scalar data variables `x1`,...,`xk`. The variable names in the formula must be valid MATLAB identifiers.

Data Types: `function_handle` | `char` | `string`

`beta0` — Coefficients

numeric vector

Coefficients for the nonlinear model, specified as a numeric vector. `NonLinearModel` starts its search for optimal coefficients from `beta0`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ErrorModel','combined','Exclude',2,'Options',opt` specifies the error model as the combined model, excludes the second observation from the fit, and uses the options defined in the structure `opt` to control the iterative fitting procedure.

`CoefficientNames` — Names of the model coefficients
`{'b1','b2',...,'bk'}` (default) | string array | cell array of character vectors

Names of the model coefficients, specified as a string array or cell array of character vectors.

Data Types: `string` | `cell`

`ErrorModel` — Form of the error variance model
`'constant'` (default) | `'proportional'` | `'combined'`

Form of the error variance model, specified as one of the following. Each model defines the error using a standard mean-zero and unit-variance variable e in combination with independent components: the function value f, and one or two parameters a and b

`'constant'` (default)	$y = f + ae$
`'proportional'`	$y = f + bfe$
`'combined'`	$y = f + (a + b\|f\|)e$

The only allowed error model when using `Weights` is `'constant'`.

Note `options.RobustWgtFun` must have value `[]` when using an error model other than `'constant'`.

Example: `'ErrorModel','proportional'`

`ErrorParameters` — Initial estimates of the error model parameters
numeric array

Initial estimates of the error model parameters for the chosen `ErrorModel`, specified as a numeric array.

Error Model	Parameters	Default Values
`'constant'`	a	1
`'proportional'`	b	1
`'combined'`	a, b	`[1,1]`

You can only use the `'constant'` error model when using `Weights`.

Note `options.RobustWgtFun` must have value `[]` when using an error model other than `'constant'`.

For example, if `'ErrorModel'` has the value `'combined'`, you can specify the starting value 1 for a and the starting value 2 for b as follows.

Example: `'ErrorParameters',[1,2]`

Data Types: `single` | `double`

`Exclude` — Observations to exclude
logical or numeric index vector

Observations to exclude from the fit, specified as the comma-separated pair consisting of `'Exclude'` and a logical or numeric index vector indicating which observations to exclude from the fit.

For example, you can exclude observations 2 and 3 out of 6 using either of the following examples.

Example: `'Exclude',[2,3]`

Example: `'Exclude',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical`

`Options` — Options for controlling the iterative fitting procedure
[ ] (default) | structure

Options for controlling the iterative fitting procedure, specified as a structure created by `statset`. The relevant fields are the nonempty fields in the structure returned by the call `statset('fitnlm')`.

Option	Meaning	Default
`DerivStep`	Relative difference used in finite difference derivative calculations. A positive scalar, or a vector of positive scalars the same size as the vector of parameters estimated by the Statistics and Machine Learning Toolbox function using the options structure.	`eps^(1/3)`
`Display`	Amount of information displayed by the fitting algorithm. • `'off'` — Displays no information. • `'final'` — Displays the final output. • `'iter'` — Displays iterative output to the Command Window.	`'off'`
`FunValCheck`	Character vector or string scalar indicating to check for invalid values, such as `NaN` or `Inf`, from the model function.	`'on'`
`MaxIter`	Maximum number of iterations allowed. Positive integer.	`200`
`RobustWgtFun`	Weight function for robust fitting. Can also be a function handle that accepts a normalized residual as input and returns the robust weights as output. If you use a function handle, give a `Tune` constant. See “Robust Options” on page 33-2289	`[]`
`Tune`	Tuning constant used in robust fitting to normalize the residuals before applying the weight function. A positive scalar. Required if the weight function is specified as a function handle.	See “Robust Options” on page 33-2289 for the default, which depends on `RobustWgtFun`.
`TolFun`	Termination tolerance for the objective function value. Positive scalar.	`1e-8`

Option	Meaning	Default
TolX	Termination tolerance for the parameters. Positive scalar.	1e-8

Data Types: struct

PredictorVars — Predictor variables
string array | cell array of character vectors | logical or numeric index vector

Predictor variables to use in the fit, specified as the comma-separated pair consisting of 'PredictorVars' and either a string array or cell array of character vectors of the variable names in the table or dataset array tbl, or a logical or numeric index vector indicating which columns are predictor variables.

The string values or character vectors should be among the names in tbl, or the names you specify using the 'VarNames' name-value pair argument.

The default is all variables in X, or all variables in tbl except for ResponseVar.

For example, you can specify the second and third variables as the predictor variables using either of the following examples.

Example: 'PredictorVars',[2,3]

Example: 'PredictorVars',logical([0 1 1 0 0 0])

Data Types: single | double | logical | string | cell

ResponseVar — Response variable
last column of tbl (default) | variable name | logical or numeric index vector

Response variable to use in the fit, specified as the comma-separated pair consisting of 'ResponseVar' and either a variable name in the table or dataset array tbl, or a logical or numeric index vector indicating which column is the response variable.

If you specify a model, it specifies the response variable. Otherwise, when fitting a table or dataset array, 'ResponseVar' indicates which variable fitnlm should use as the response.

For example, you can specify the fourth variable, say yield, as the response out of six variables, in one of the following ways.

Example: 'ResponseVar','yield'

Example: 'ResponseVar',[4]

Example: 'ResponseVar',logical([0 0 0 1 0 0])

Data Types: single | double | logical | char | string

VarNames — Names of variables
{'x1','x2',...,'xn','y'} (default) | string array | cell array of character vectors

Names of variables, specified as the comma-separated pair consisting of 'VarNames' and a string array or cell array of character vectors including the names for the columns of X first, and the name for the response variable y last.

'VarNames' is not applicable to variables in a table or dataset array, because those variables already have names.

Example: `'VarNames',{'Horsepower','Acceleration','Model_Year','MPG'}`

Data Types: `string` | `cell`

`Weights` — Observation weights
`ones(n,1)` (default) | vector of nonnegative scalar values | function handle

Observation weights, specified as a vector of nonnegative scalar values or function handle.

- If you specify a vector, then it must have *n* elements, where *n* is the number of rows in `tbl` or `y`.
- If you specify a function handle, then the function must accept a vector of predicted response values as input, and return a vector of real positive weights as output.

Given weights, `W`, `NonLinearModel` estimates the error variance at observation `i` by `MSE*(1/W(i))`, where MSE is the mean squared error.

Data Types: `single` | `double` | `function_handle`

Output Arguments

`mdl` — Nonlinear model
`NonLinearModel` object

Nonlinear model representing a least-squares fit of the response to the data, returned as a `NonLinearModel` object.

If the `Options` structure contains a nonempty `RobustWgtFun` field, the model is not a least-squares fit, but uses the `RobustWgtFun` robust fitting function.

For properties and methods of the nonlinear model object, `mdl`, see the `NonLinearModel` class page.

More About

Robust Options

Weight Function	Equation	Default Tuning Constant
`'andrews'`	`w = (abs(r)<pi) .* sin(r) ./ r`	1.339
`'bisquare'` (default)	`w = (abs(r)<1) .* (1 - r.^2).^2`	4.685
`'cauchy'`	`w = 1 ./ (1 + r.^2)`	2.385
`'fair'`	`w = 1 ./ (1 + abs(r))`	1.400
`'huber'`	`w = 1 ./ max(1, abs(r))`	1.345
`'logistic'`	`w = tanh(r) ./ r`	1.205
`'talwar'`	`w = 1 * (abs(r)<1)`	2.795
`'welsch'`	`w = exp(-(r.^2))`	2.985
`[]`	No robust fitting	—

Algorithms

- `fitnlm` uses the same fitting algorithm as `nlinfit`.
- `fitnlm` considers `NaN` values in `tbl`, `X`, and `y` to be missing values. When fitting a model, `fitnlm` does not use observations with missing values or observations at which `modelfun` returns `NaN` values. The `ObservationInfo` property of a fitted model contains information regarding whether or not `fitnlm` uses each observation in the fit.

References

[1] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

[2] DuMouchel, W. H., and F. L. O'Brien. "Integrating a Robust Option into a Multiple Regression Computing Environment." *Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[3] Holland, P. W., and R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares." *Communications in Statistics: Theory and Methods*, *A6*, 1977, pp. 813–827.

See Also

`NonLinearModel` | `nlinfit`

Topics

"Examine Quality and Adjust the Fitted Nonlinear Model" on page 13-6
"Predict or Simulate Responses Using a Nonlinear Model" on page 13-9
"Nonlinear Regression Workflow" on page 13-12
"Nonlinear Regression" on page 13-2

Introduced in R2013b

fitPosterior

Fit posterior probabilities for support vector machine (SVM) classifier

Syntax

```
ScoreSVMModel = fitPosterior(SVMModel)
[ScoreSVMModel,ScoreTransform] = fitPosterior(SVMModel)
[ScoreSVMModel,ScoreTransform] = fitPosterior(SVMModel,Name,Value)
```

Description

`ScoreSVMModel = fitPosterior(SVMModel)` returns a trained support vector machine (SVM) classifier `ScoreSVMModel` containing the optimal score-to-posterior-probability transformation function for two-class learning. For more details, see “Algorithms” on page 33-2299.

`[ScoreSVMModel,ScoreTransform] = fitPosterior(SVMModel)` additionally returns the optimal score-to-posterior-probability transformation function parameters.

`[ScoreSVMModel,ScoreTransform] = fitPosterior(SVMModel,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify the number of folds or the holdout sample proportion.

Examples

Estimate In-Sample Posterior Probabilities of SVM Classifier

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a support vector machine (SVM) classifier. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X,Y,'ClassNames',{'b','g'},'Standardize',true);
```

`SVMModel` is a `ClassificationSVM` classifier.

Fit the optimal score-to-posterior-probability transformation function.

```
rng(1); % For reproducibility
ScoreSVMModel = fitPosterior(SVMModel)
```

```
ScoreSVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: '@(S)sigmoid(S,-9.481840e-01,-1.218721e-01)'
         NumObservations: 351
```

```
                Alpha: [90x1 double]
                 Bias: -0.1343
     KernelParameters: [1x1 struct]
                   Mu: [1x34 double]
                Sigma: [1x34 double]
       BoxConstraints: [351x1 double]
      ConvergenceInfo: [1x1 struct]
      IsSupportVector: [351x1 logical]
               Solver: 'SMO'


  Properties, Methods
```

Because the classes are inseparable, the score transformation function (`ScoreSVMModel.ScoreTransform`) is the sigmoid function.

Estimate scores and positive class posterior probabilities for the training data. Display the results for the first 10 observations.

```
[label,scores] = resubPredict(SVMModel);
[~,postProbs] = resubPredict(ScoreSVMModel);
table(Y(1:10),label(1:10),scores(1:10,2),postProbs(1:10,2),'VariableNames',...
    {'TrueLabel','PredictedLabel','Score','PosteriorProbability'})
```

```
ans=10×4 table
    TrueLabel    PredictedLabel     Score     PosteriorProbability
    _________    ______________    _______    ____________________

      {'g'}          {'g'}          1.4861           0.82215
      {'b'}          {'b'}         -1.0002           0.30439
      {'g'}          {'g'}          1.8686           0.86917
      {'b'}          {'b'}         -2.6456          0.084197
      {'g'}          {'g'}          1.2806           0.79185
      {'b'}          {'b'}         -1.4617           0.22026
      {'g'}          {'g'}          2.1671           0.89814
      {'b'}          {'b'}         -5.7089         0.0050106
      {'g'}          {'g'}          2.4796           0.92223
      {'b'}          {'b'}         -2.7812          0.074801
```

Plot Posterior Probability Contours for Multiple Classes

Train a multiclass SVM classifier through the process of one-versus-all (OVA) classification, and then plot probability contours for each class. To implement OVA directly, see `fitcecoc`.

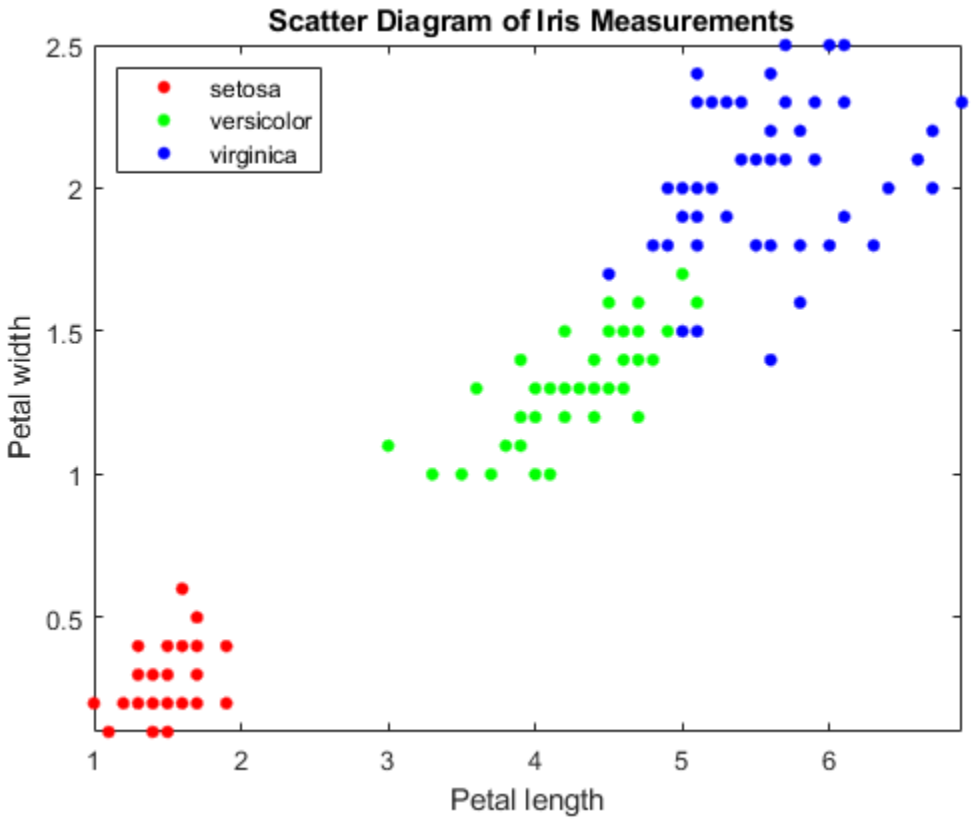
Load Fisher's iris data set. Use the petal lengths and widths as the predictor data.

```
load fisheriris
X = meas(:,3:4);
Y = species;
```

Examine a scatter plot of the data.

```
figure
gscatter(X(:,1),X(:,2),Y);
```

```
title('{\bf Scatter Diagram of Iris Measurements}');
xlabel('Petal length');
ylabel('Petal width');
legend('Location','Northwest');
axis tight
```



Train three binary SVM classifiers that separate each type of iris from the others. Assume that a radial basis function is an appropriate kernel for each, and allow the algorithm to choose a kernel scale. Define the class order.

```
classNames = {'setosa'; 'virginica'; 'versicolor'};
numClasses = size(classNames,1);
inds = cell(3,1); % Preallocation
SVMModel = cell(3,1);

rng(1); % For reproducibility
for j = 1:numClasses
    inds{j} = strcmp(Y,classNames{j});  % OVA classification
    SVMModel{j} = fitcsvm(X,inds{j},'ClassNames',[false true],...
        'Standardize',true,'KernelFunction','rbf','KernelScale','auto');
end
```

`fitcsvm` uses a heuristic procedure that involves subsampling to compute the value of the kernel scale.

Fit the optimal score-to-posterior-probability transformation function for each classifier.

```
for j = 1:numClasses
    SVMModel{j} = fitPosterior(SVMModel{j});
end
```

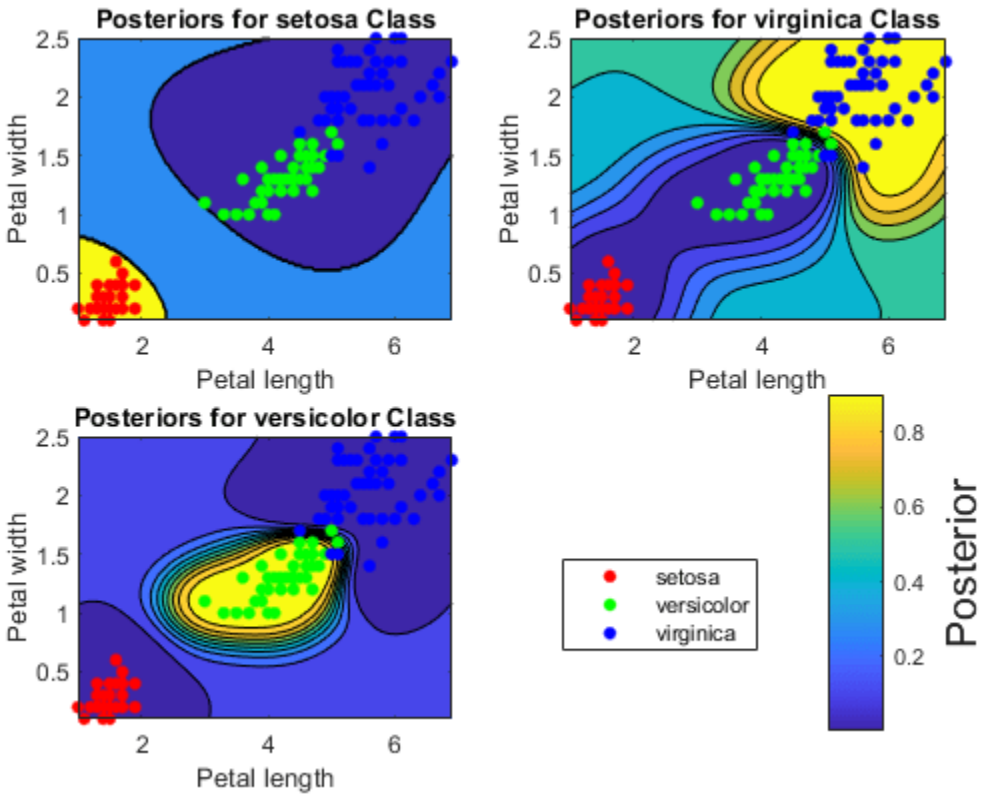
```
Warning: Classes are perfectly separated. The optimal score-to-posterior transformation is a step
```

Define a grid to plot the posterior probability contours. Estimate the posterior probabilities over the grid for each classifier.

```
d = 0.02;
[x1Grid,x2Grid] = meshgrid(min(X(:,1)):d:max(X(:,1)),...
    min(X(:,2)):d:max(X(:,2)));
xGrid = [x1Grid(:),x2Grid(:)];

posterior = cell(3,1);
for j = 1:numClasses
    [~,posterior{j}] = predict(SVMModel{j},xGrid);
end
```

For each SVM classifier, plot the posterior probability contour under the scatter plot of the data.

```
figure
h = zeros(numClasses + 1,1); % Preallocation for graphics handles
for j = 1:numClasses
subplot(2,2,j)
contourf(x1Grid,x2Grid,reshape(posterior{j}(:,2),size(x1Grid,1),size(x1Grid,2)));
hold on
h(1:numClasses) = gscatter(X(:,1),X(:,2),Y);
title(sprintf('Posteriors for %s Class',classNames{j}));
xlabel('Petal length');
ylabel('Petal width');
legend off
axis tight
hold off
end
h(numClasses + 1) = colorbar('Location','EastOutside',...
    'Position',[[0.8,0.1,0.05,0.4]]);
set(get(h(numClasses + 1),'YLabel'),'String','Posterior','FontSize',16);
legend(h(1:numClasses),'Location',[0.6,0.2,0.1,0.1]);
```



Fit Optimal Posterior Probability Function Using Holdout Cross-Validation

Estimate the score-to-posterior-probability transformation function after training an SVM classifier. Use cross-validation during the estimation to reduce bias, and compare the run times for 10-fold cross-validation and holdout cross-validation.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X,Y,'ClassNames',{'b','g'},'Standardize',true);
```

`SVMModel` is a `ClassificationSVM` classifier.

Fit the optimal score-to-posterior-probability transformation function. Compare the run times from using 10-fold cross-validation (the default) and a 10% holdout test sample.

```
rng(1); % For reproducibility
tic;    % Start the stopwatch
SVMModel_10FCV = fitPosterior(SVMModel);
toc     % Stop the stopwatch and display the run time
```

```
Elapsed time is 1.081044 seconds.
```

```
tic;
SVMModel_HO = fitPosterior(SVMModel,'Holdout',0.10);
toc
```

```
Elapsed time is 0.269525 seconds.
```

Although both run times are short because the data set is relatively small, `SVMModel_HO` fits the score transformation function much faster than `SVMModel_10FCV`. You can specify holdout cross-validation (instead of the default 10-fold cross validation) to reduce run time for larger data sets.

Input Arguments

`SVMModel` — Full, trained SVM classifier

`ClassificationSVM` classifier

Full, trained SVM classifier, specified as a `ClassificationSVM` model trained with `fitcsvm`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `fitPosterior(SVMModel,'KFold',5)` uses five folds in a cross-validated model.

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition

Cross-validation partition used to compute the transformation function, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` partition object as created by `cvpartition`. You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

The `crossval` name-value pair argument of `fitcsvm` splits the data into subsets using `cvpartition`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data for holdout validation used to compute the transformation function, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range (0,1). Holdout validation tests the specified fraction of the data and uses the remaining data for training.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use when computing the transformation function, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'KFold',8`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag indicating whether to use leave-one-out cross-validation to compute the transformation function, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. Use leave-one-out cross-validation by specifying `'Leaveout','on'`.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'Leaveout','on'`

Output Arguments

`ScoreSVMModel` — Trained SVM classifier
`ClassificationSVM` classifier

Trained SVM classifier, returned as a `ClassificationSVM` classifier. The trained classifier contains the estimated score-to-posterior-probability transformation function.

To estimate posterior probabilities for the training set observations, pass `ScoreSVMModel` to `resubPredict`.

To estimate posterior probabilities for new observations, pass the new observations and `ScoreSVMModel` to `predict`.

`ScoreTransform` — Optimal score-to-posterior-probability transformation function parameters
structure array

Optimal score-to-posterior-probability transformation function parameters, returned as a structure array.

- If the value of the `Type` field of `ScoreTransform` is `sigmoid`, then `ScoreTransform` also has these fields:
  - `Slope`: The value of A in the sigmoid function on page 33-2307
  - `Intercept`: The value of B in the sigmoid function
- If the value of the `Type` field of `ScoreTransform` is `step`, then `ScoreTransform` also has these fields:
  - `PositiveClassProbability`: The value of π in the step function on page 33-2307. This value represents the probability that an observation is in the positive class or the posterior

probability that an observation is in the positive class given that its score is in the interval (`LowerBound`,`UpperBound`).

- `LowerBound`: The value $\max_{y_n = -1} s_n$ in the step function. This value represents the lower bound of the score interval that assigns observations with scores in the interval the posterior probability of being in the positive class `PositiveClassProbability`. Any observation with a score less than `LowerBound` has the posterior probability of being in the positive class equal to `0`.
- `UpperBound`: The value $\min_{y_n = +1} s_n$ in the step function. This value represents the upper bound of the score interval that assigns observations with scores in the interval the posterior probability of being in the positive class `PositiveClassProbability`. Any observation with a score greater than `UpperBound` has the posterior probability of being in the positive class equal to `1`.

- If the value of the `Type` field of `ScoreTransform` is `constant`, then `ScoreTransform.PredictedClass` contains the name of the class prediction.

  This result is the same as `SVMModel.ClassNames`. The posterior probability of an observation being in `ScoreTransform.PredictedClass` is always `1`.

More About

Sigmoid Function

The sigmoid function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \frac{1}{1 + \exp(As_j + B)}.$$

If the value of the `Type` field of `ScoreTransform` is `sigmoid`, then parameters A and B correspond to the fields `Scale` and `Intercept` of `ScoreTransform`, respectively.

Step Function

The step function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \begin{cases} 0; & s < \max_{y_k = -1} s_k \\ \pi; & \max_{y_k = -1} s_k \le s_j \le \min_{y_k = +1} s_k, \\ 1; & s_j > \min_{y_k = +1} s_k \end{cases}$$

where:

- s_j is the score of observation j.
- +1 and –1 denote the positive and negative classes, respectively.
- π is the prior probability that an observation is in the positive class.

If the value of the `Type` field of `ScoreTransform` is `step`, then the quantities $\max_{y_k = -1} s_k$ and $\min_{y_k = +1} s_k$ correspond to the fields `LowerBound` and `UpperBound` of `ScoreTransform`, respectively.

Constant Function

The constant function maps all scores in a sample to posterior probabilities 1 or 0.

If all observations have posterior probability 1, then they are expected to come from the positive class.

If all observations have posterior probability 0, then they are not expected to come from the positive class.

Tips

- This process describes one way to predict positive class posterior probabilities.

  1. Train an SVM classifier by passing the data to `fitcsvm`. The result is a trained SVM classifier, such as `SVMModel`, that stores the data. The software sets the score transformation function property (`SVMModel.ScoreTransformation`) to `none`.
  2. Pass the trained SVM classifier `SVMModel` to `fitSVMPosterior` or `fitPosterior`. The result, such as, `ScoreSVMModel`, is the same trained SVM classifier as `SVMModel`, except the software sets `ScoreSVMModel.ScoreTransformation` to the optimal score transformation function.
  3. Pass the predictor data matrix and the trained SVM classifier containing the optimal score transformation function (`ScoreSVMModel`) to `predict`. The second column in the second output argument of `predict` stores the positive class posterior probabilities corresponding to each row of the predictor data matrix.

     If you skip step 2, then `predict` returns the positive class score rather than the positive class posterior probability.

- After fitting posterior probabilities, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

The software fits the appropriate score-to-posterior-probability transformation function by using the SVM classifier `SVMModel` and by conducting 10-fold cross-validation using the stored predictor data (`SVMModel.X`) and the class labels (`SVMModel.Y`), as outlined in [1]. The transformation function computes the posterior probability that an observation is classified into the positive class (`SVMModel.Classnames(2)`).

- If the classes are inseparable, then the transformation function is the sigmoid function on page 33-2307.
- If the classes are perfectly separable, then the transformation function is the step function on page 33-2307.
- In two-class learning, if one of the two classes has a relative frequency of 0, then the transformation function is the constant function on page 33-2428. The `fitPosterior` function is not appropriate for one-class learning.

- The software stores the optimal score-to-posterior-probability transformation function in `ScoreSVMModel.ScoreTransform`.

If you re-estimate the score-to-posterior-probability transformation function, that is, if you pass an SVM classifier to `fitPosterior` or `fitSVMPosterior` and its `ScoreTransform` property is not `none`, then the software:

- Displays a warning
- Resets the original transformation function to `'none'` before estimating the new one

Alternative Functionality

You can also fit the posterior probability function by using `fitSVMPosterior`. This function is similar to `fitPosterior`, except it is more broad because it accepts a wider range of SVM classifier types.

References

[1] Platt, J. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.” *Advances in Large Margin Classifiers*. Cambridge, MA: The MIT Press, 2000, pp. 61–74.

See Also

`ClassificationSVM` | `fitSVMPosterior` | `fitcsvm` | `predict`

Introduced in R2014a

fitPosterior

Package: `classreg.learning.classif`

Fit posterior probabilities for compact support vector machine (SVM) classifier

Syntax

```
ScoreSVMModel = fitPosterior(SVMModel,TBL,Y)
ScoreSVMModel = fitPosterior(SVMModel,X,Y)
[ScoreSVMModel,ScoreTransform] = fitPosterior( ___ )
```

Description

`ScoreSVMModel = fitPosterior(SVMModel,TBL,Y)` returns a trained support vector machine (SVM) classifier `ScoreSVMModel` containing the optimal score-to-posterior-probability transformation function for two-class learning. For more details, see “Algorithms” on page 33-2308. If you train `SVMModel` using a table, then you must use a table as input for `fitPosterior`.

`ScoreSVMModel = fitPosterior(SVMModel,X,Y)` returns a trained SVM classifier `ScoreSVMModel` containing the optimal score-to-posterior-probability transformation function for two-class learning. If you train `SVMModel` using a matrix, then you must use a matrix as input for `fitPosterior`.

`[ScoreSVMModel,ScoreTransform] = fitPosterior( ___ )` additionally returns the optimal score-to-posterior-probability transformation function parameters (`ScoreTransform`) for any of the input argument combinations in the previous syntaxes.

Examples

Estimate Posterior Probabilities for Data with Inseparable Classes

Load the `ionosphere` data set. Reserve 20 random observations of the data, and consider this set new data.

```
load ionosphere
n = size(X,1);
rng(1);  % For reproducibility

indx = ~ismember([1:n],randsample(n,20)); % Indices for the training data
```

The classes of this data set are inseparable.

Train an SVM classifier using the training data. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X(indx,:),Y(indx),'ClassNames',{'b','g'},...
    'Standardize',true);
```

`SVMModel` is a `ClassificationSVM` classifier.

Use the new data set to estimate the optimal score-to-posterior-probability transformation function for mapping scores to the posterior probability of an observation being classified as `g`. For efficiency, make a compact version of `SVMModel`, and pass it and the new data to `fitPosterior`.

```
CompactSVMModel = compact(SVMModel);
[ScoreCSVMModel,ScoreParameters] = fitPosterior(CompactSVMModel,...
    X(~indx,:),Y(~indx));

ScoreTransform = ScoreCSVMModel.ScoreTransform
```

```
ScoreTransform =
'@(S)sigmoid(S,-1.099032e+00,4.521358e-01)'
```

```
ScoreParameters
```

```
ScoreParameters = struct with fields:
         Type: 'sigmoid'
        Slope: -1.0990
    Intercept: 0.4521
```

`ScoreTransform` is the optimal score transformation function. `ScoreParameters` is a structure array with three fields: the score transformation function name (`Type`), the sigmoid slope (`Slope`), and the sigmoid intercept estimates (`Intercept`).

Alternatively, you can pass `SVMModel` and the new data to `fitSVMPosterior`, but this process is not as efficient.

Estimate the posterior probabilities that the observations in the new data are in class `g`.

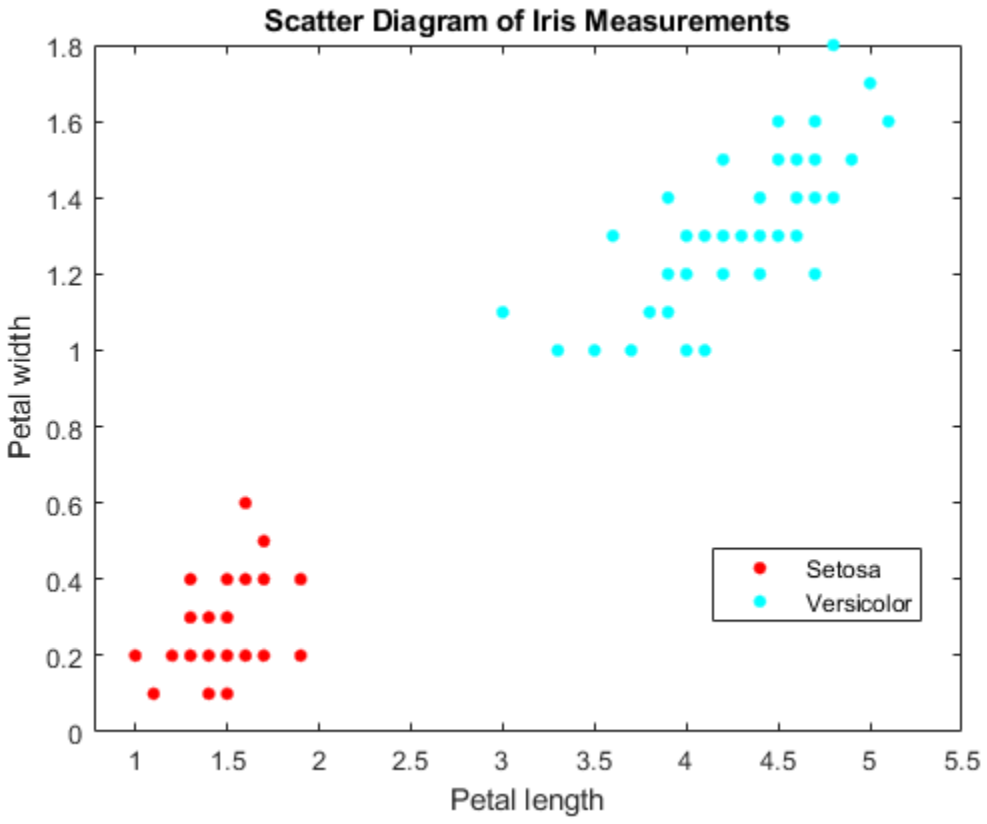
```
[labels,postProbs] = predict(ScoreCSVMModel,X(~indx,:));
table(Y(~indx),labels,postProbs(:,2),...
    'VariableNames',{'TrueLabel','PredictedLabel','PosteriorProbability'})
```

```
ans=20×3 table
    TrueLabel    PredictedLabel    PosteriorProbability
    _________    ______________    ____________________

      {'g'}          {'g'}                0.78441
      {'b'}          {'b'}               0.024573
      {'g'}          {'g'}                0.82404
      {'b'}          {'b'}              0.0061609
      {'b'}          {'b'}             3.6018e-06
      {'b'}          {'b'}                0.15688
      {'b'}          {'g'}                0.96219
      {'b'}          {'b'}             6.1253e-09
      {'b'}          {'b'}              0.0019635
      {'g'}          {'g'}                0.72509
      {'g'}          {'g'}                0.70264
      {'b'}          {'b'}               0.075291
      {'g'}          {'g'}                0.90693
      {'g'}          {'g'}                 0.8285
      {'b'}          {'b'}               0.051175
      {'g'}          {'g'}                0.95333
      ⋮
```

Estimate Posterior Probabilities for Data with Separable Classes

Load Fisher's iris data set. Use the petal lengths and widths as the predictor data, and remove the virginica species from the data. Reserve 10 random observations of the data, and consider this set new data.

```
load fisheriris
classKeep = ~strcmp(species,'virginica');
X = meas(classKeep,3:4);
Y = species(classKeep);

rng(1);  % For reproducibility
indx1 = 1:numel(species);
indx2 = indx1(classKeep);
indx = ~ismember(indx2,randsample(indx2,10)); % Indices for the training data

gscatter(X(indx,1),X(indx,2),Y(indx));
title('Scatter Diagram of Iris Measurements')
xlabel('Petal length')
ylabel('Petal width')
legend('Setosa','Versicolor')
```



The classes are perfectly separable. Therefore, the score-to-posterior-probability transformation function is a step function.

Train an SVM classifier. Standardize the data and specify that `versicolor` is the positive class.

```
SVMModel = fitcsvm(X(indx,:),Y(indx),...
    'ClassNames',{'setosa','versicolor'},'Standardize',true);
```

`SVMModel` is a `ClassificationSVM` classifier.

Use the new data set to estimate the optimal score-to-posterior-probability transformation function for mapping scores to the posterior probability of an observation being classified as `versicolor`. For efficiency, make a compact version `SVMModel`, and pass it and the new data to `fitPosterior`.

```
CompactSVMModel = compact(SVMModel);
[ScoreCSVMModel,ScoreParameters] = fitPosterior(CompactSVMModel,...
    X(~indx,:),Y(~indx));
```

```
Warning: Classes are perfectly separated. The optimal score-to-posterior transformation is a step
```

```
ScoreTransform = ScoreCSVMModel.ScoreTransform
```

```
ScoreTransform =
'@(S)step(S,-1.338450e+00,2.012495e+00,5.333333e-01)'
```

`fitPosterior` displays a warning whenever the classes are separable, and stores the step function in `ScoreSVMModel.ScoreTransform`.

Display the score function type and its estimated values.

```
ScoreParameters
```

```
ScoreParameters = struct with fields:
                      Type: 'step'
                LowerBound: -1.3385
                UpperBound: 2.0125
    PositiveClassProbability: 0.5333
```

`ScoreParameters` is a structure array with four fields:

- Score transformation function type (`Type`)
- Score corresponding to the negative class boundary (`LowerBound`)
- Score corresponding to the positive class boundary (`UpperBound`)
- Positive class probability (`PositiveClassProbability`)

Alternatively, you can pass `SVMModel` and the new data to `fitSVMPosterior`, but this process is not as efficient.

Estimate the posterior probabilities that the observations in the new data are versicolor irises.

```
[labels,postProbs] = predict(ScoreCSVMModel,X(~indx,:));
table(Y(~indx),labels,postProbs(:,2),...
    'VariableNames',{'TrueLabel','PredictedLabel','PosteriorProbability'})
```

```
ans=10×3 table
      TrueLabel       PredictedLabel    PosteriorProbability
    ______________    ______________    ____________________

    {'setosa'    }    {'setosa'    }             0
    {'setosa'    }    {'setosa'    }             0
```

```
    {'setosa'    }    {'setosa'    }            0
    {'setosa'    }    {'setosa'    }            0
    {'setosa'    }    {'setosa'    }            0
    {'setosa'    }    {'setosa'    }            0
    {'setosa'    }    {'setosa'    }            0
    {'setosa'    }    {'setosa'    }            0
    {'versicolor'}    {'versicolor'}            1
    {'versicolor'}    {'versicolor'}            1
```

Because the classes are separable, the step function transforms the positive-class score to:

- `0` if the score is less than `ScoreParameters.LowerBound`
- `1` if the score is greater than `ScoreParameters.UpperBound`
- `ScoreParameters.PositiveClassProbability` if the score is in the interval [`ScoreParameters.LowerBound` , `ScoreParameters.LowerBound`]

Input Arguments

`SVMModel` — Trained, compact SVM classifier

`CompactClassificationSVM` classifier

Trained, compact SVM classifier, specified as a `CompactClassificationSVM` model returned by `compact`.

`TBL` — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. `TBL` must contain all of the predictors used to train `SVMModel`. Optionally, `TBL` can contain an additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `Y`. If `TBL` does not include the response variable, then the length of `Y` must be equal to the number of rows in `TBL`.

If the sample data used to train `SVMModel` is a `table`, then you must specify the input data for `fitPosterior` as a table.

If you set `'Standardize',true` in `fitcsvm` when training `SVMModel`, then the software fits the transformation function parameter estimates using standardized data.

Data Types: `table`

`X` — Predictor data

matrix

Predictor data used to estimate the score-to-posterior-probability transformation function, specified as a matrix.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature).

The length of Y and the number of rows in X must be equal.

If you set 'Standardize',true in fitcsvm when training SVMModel, then the software fits the transformation function parameter estimates using standardized data.

Data Types: double | single

Y — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels used to estimate the score-to-posterior-probability transformation function, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors.

If Y is a character array, then each element must correspond to one class label.

The length of Y and the number of rows in X must be equal.

Data Types: categorical | char | string | logical | single | double | cell

Output Arguments

ScoreSVMModel — Trained, compact SVM classifier
CompactClassificationSVM classifier

Trained, compact SVM classifier containing the estimated score-to-posterior-probability transformation function, returned as a CompactClassificationSVM classifier.

To estimate posterior probabilities for new observations, pass ScoreSVMModel and the new observations to predict.

ScoreTransform — Optimal score-to-posterior-probability transformation function parameters
structure array

Optimal score-to-posterior-probability transformation function parameters, returned as a structure array.

- If the value of the Type field of ScoreTransform is sigmoid, then ScoreTransform also has these fields:
  - Slope: The value of A in the sigmoid function on page 33-2307
  - Intercept: The value of B in the sigmoid function
- If the value of the Type field of ScoreTransform is step, then ScoreTransform also has these fields:
  - PositiveClassProbability: The value of π in the step function on page 33-2307. This value represents the probability that an observation is in the positive class or the posterior probability that an observation is in the positive class given that its score is in the interval (LowerBound,UpperBound).
  - LowerBound: The value $\max_{y_n = -1} s_n$ in the step function. This value represents the lower bound of the score interval that assigns observations with scores in the interval the posterior probability of being in the positive class PositiveClassProbability. Any observation with

a score less than `LowerBound` has the posterior probability of being in the positive class equal to `0`.

- `UpperBound`: The value $\min_{y_n = +1} s_n$ in the step function. This value represents the upper bound of the score interval that assigns observations with scores in the interval the posterior probability of being in the positive class `PositiveClassProbability`. Any observation with a score greater than `UpperBound` has the posterior probability of being in the positive class equal to `1`.

- If the value of the `Type` field of `ScoreTransform` is `constant`, then `ScoreTransform.PredictedClass` contains the name of the class prediction.

  This result is the same as `SVMModel.ClassNames`. The posterior probability of an observation being in `ScoreTransform.PredictedClass` is always `1`.

More About

Sigmoid Function

The sigmoid function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \frac{1}{1 + \exp(As_j + B)}.$$

If the value of the `Type` field of `ScoreTransform` is `sigmoid`, then parameters A and B correspond to the fields `Scale` and `Intercept` of `ScoreTransform`, respectively.

Step Function

The step function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \begin{cases} 0; & s < \max_{y_k = -1} s_k \\ \pi; & \max_{y_k = -1} s_k \le s_j \le \min_{y_k = +1} s_k, \\ 1; & s_j > \min_{y_k = +1} s_k \end{cases}$$

where:

- s_j is the score of observation j.
- +1 and –1 denote the positive and negative classes, respectively.
- π is the prior probability that an observation is in the positive class.

If the value of the `Type` field of `ScoreTransform` is `step`, then the quantities $\max_{y_k = -1} s_k$ and $\min_{y_k = +1} s_k$ correspond to the fields `LowerBound` and `UpperBound` of `ScoreTransform`, respectively.

Constant Function

The constant function maps all scores in a sample to posterior probabilities 1 or 0.

If all observations have posterior probability 1, then they are expected to come from the positive class.

If all observations have posterior probability 0, then they are not expected to come from the positive class.

Tips

- This process describes one way to predict positive class posterior probabilities.

  1. Train an SVM classifier by passing the data to `fitcsvm`. The result is a trained SVM classifier, such as `SVMModel`, that stores the data. The software sets the score transformation function property (`SVMModel.ScoreTransformation`) to `none`.
  2. Pass the trained SVM classifier `SVMModel` to `fitSVMPosterior` or `fitPosterior`. The result, such as, `ScoreSVMModel`, is the same trained SVM classifier as `SVMModel`, except the software sets `ScoreSVMModel.ScoreTransformation` to the optimal score transformation function.
  3. Pass the predictor data matrix and the trained SVM classifier containing the optimal score transformation function (`ScoreSVMModel`) to `predict`. The second column in the second output argument of `predict` stores the positive class posterior probabilities corresponding to each row of the predictor data matrix.

     If you skip step 2, then `predict` returns the positive class score rather than the positive class posterior probability.

- After fitting posterior probabilities, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

The software fits the appropriate score-to-posterior-probability transformation function by using the SVM classifier `SVMModel` and by conducting 10-fold cross-validation using the stored predictor data (`SVMModel.X`) and the class labels (`SVMModel.Y`), as outlined in [1]. The transformation function computes the posterior probability that an observation is classified into the positive class (`SVMModel.Classnames(2)`).

- If the classes are inseparable, then the transformation function is the sigmoid function on page 33-2307.
- If the classes are perfectly separable, then the transformation function is the step function on page 33-2307.
- In two-class learning, if one of the two classes has a relative frequency of 0, then the transformation function is the constant function on page 33-2428. The `fitPosterior` function is not appropriate for one-class learning.
- The software stores the optimal score-to-posterior-probability transformation function in `ScoreSVMModel.ScoreTransform`.

If you re-estimate the score-to-posterior-probability transformation function, that is, if you pass an SVM classifier to `fitPosterior` or `fitSVMPosterior` and its `ScoreTransform` property is not `none`, then the software:

- Displays a warning
- Resets the original transformation function to `'none'` before estimating the new one

Alternative Functionality

You can also fit the optimal score-to-posterior-probability function by using `fitSVMPosterior`. This function is similar to `fitPosterior`, except it is more broad because it accepts a wider range of SVM classifier types.

References

[1] Platt, J. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.” *Advances in Large Margin Classifiers*. Cambridge, MA: The MIT Press, 2000, pp. 61–74.

See Also

`CompactClassificationSVM` | `fitSVMPosterior` | `fitcsvm` | `predict`

Introduced in R2014a

fitrensemble

Fit ensemble of learners for regression

Syntax

```
Mdl = fitrensemble(Tbl,ResponseVarName)
Mdl = fitrensemble(Tbl,formula)
Mdl = fitrensemble(Tbl,Y)

Mdl = fitrensemble(X,Y)

Mdl = fitrensemble( ___ ,Name,Value)
```

Description

`Mdl = fitrensemble(Tbl,ResponseVarName)` returns the trained regression ensemble model object (`Mdl`) that contains the results of boosting 100 regression trees using LSBoost and the predictor and response data in the table `Tbl`. `ResponseVarName` is the name of the response variable in `Tbl`.

`Mdl = fitrensemble(Tbl,formula)` applies `formula` to fit the model to the predictor and response data in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`. For example, `'Y~X1+X2+X3'` fits the response variable `Tbl.Y` as a function of the predictor variables `Tbl.X1`, `Tbl.X2`, and `Tbl.X3`.

`Mdl = fitrensemble(Tbl,Y)` treats all variables in the table `Tbl` as predictor variables. `Y` is the vector of responses that is not in `Tbl`.

`Mdl = fitrensemble(X,Y)` uses the predictor data in the matrix `X` and response data in the vector `Y`.

`Mdl = fitrensemble( ___ ,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments and any of the input arguments in the previous syntaxes. For example, you can specify the number of learning cycles, the ensemble aggregation method, or to implement 10-fold cross-validation.

Examples

Train Regression Ensemble

Create a regression ensemble that predicts the fuel economy of a car given the number of cylinders, volume displaced by the cylinders, horsepower, and weight. Then, train another ensemble using fewer predictors. Compare the in-sample predictive accuracies of the ensembles.

Load the `carsmall` data set. Store the variables to be used in training in a table.

```
load carsmall
Tbl = table(Cylinders,Displacement,Horsepower,Weight,MPG);
```

Train a regression ensemble.

```
Mdl1 = fitrensemble(Tbl,'MPG');
```

`Mdl1` is a `RegressionEnsemble` model. Some notable characteristics of `Mdl1` are:

- The ensemble aggregation algorithm is `'LSBoost'`.
- Because the ensemble aggregation method is a boosting algorithm, regression trees that allow a maximum of 10 splits compose the ensemble.
- One hundred trees compose the ensemble.

Because `MPG` is a variable in the MATLAB® Workspace, you can obtain the same result by entering

```
Mdl1 = fitrensemble(Tbl,MPG);
```

Use the trained regression ensemble to predict the fuel economy for a four-cylinder car with a 200-cubic inch displacement, 150 horsepower, and weighing 3000 lbs.

```
pMPG = predict(Mdl1,[4 200 150 3000])
```

```
pMPG = 25.6467
```

Train a new ensemble using all predictors in `Tbl` except `Displacement`.

```
formula = 'MPG ~ Cylinders + Horsepower + Weight';
Mdl2 = fitrensemble(Tbl,formula);
```

Compare the resubstitution MSEs between `Mdl1` and `Mdl2`.

```
mse1 = resubLoss(Mdl1)
```

```
mse1 = 0.3096
```

```
mse2 = resubLoss(Mdl2)
```

```
mse2 = 0.5861
```

The in-sample MSE for the ensemble that trains on all predictors is lower.

Speed Up Training by Binning Numeric Predictor Values

Train an ensemble of boosted regression trees by using `fitrensemble`. Reduce training time by specifying the `'NumBins'` name-value pair argument to bin numeric predictors. After training, you can reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

Generate a sample data set.

```
rng('default') % For reproducibility
N = 1e6;
X1 = randi([-1,5],[N,1]);
X2 = randi([5,10],[N,1]);
X3 = randi([0,5],[N,1]);
X4 = randi([1,10],[N,1]);
X = [X1 X2 X3 X4];
y = X1 + X2 + X3 + X4 + normrnd(0,1,[N,1]);
```

Train an ensemble of boosted regression trees using least-squares boosting (`LSBoost`, the default value). Time the function for comparison purposes.

```
tic
Mdl1 = fitrensemble(X,y);
toc
```

```
Elapsed time is 78.662954 seconds.
```

Speed up training by using the `'NumBins'` name-value pair argument. If you specify the `'NumBins'` value as a positive integer scalar, then the software bins every numeric predictor into a specified number of equiprobable bins, and then grows trees on the bin indices instead of the original data. The software does not bin categorical predictors.

```
tic
Mdl2 = fitrensemble(X,y,'NumBins',50);
toc
```

```
Elapsed time is 43.353208 seconds.
```

The process is about two times faster when you use binned data instead of the original data. Note that the elapsed time can vary depending on your operating system.

Compare the regression errors by resubstitution.

```
rsLoss = resubLoss(Mdl1)
```

```
rsLoss = 1.0134
```

```
rsLoss2 = resubLoss(Mdl2)
```

```
rsLoss2 = 1.0133
```

In this example, binning predictor values reduces training time without a significant loss of accuracy. In general, when you have a large data set like the one in this example, using the binning option speeds up training but causes a potential decrease in accuracy. If you want to reduce training time further, specify a smaller number of bins.

Reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

```
X = Mdl2.X; % Predictor data
Xbinned = zeros(size(X));
edges = Mdl2.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are `0` for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Estimate Generalization Error of Boosting Ensemble

Estimate the generalization error of an ensemble of boosted regression trees.

Load the `carsmall` data set. Choose the number of cylinders, volume displaced by the cylinders, horsepower, and weight as predictors of fuel economy.

```
load carsmall
X = [Cylinders Displacement Horsepower Weight];
```

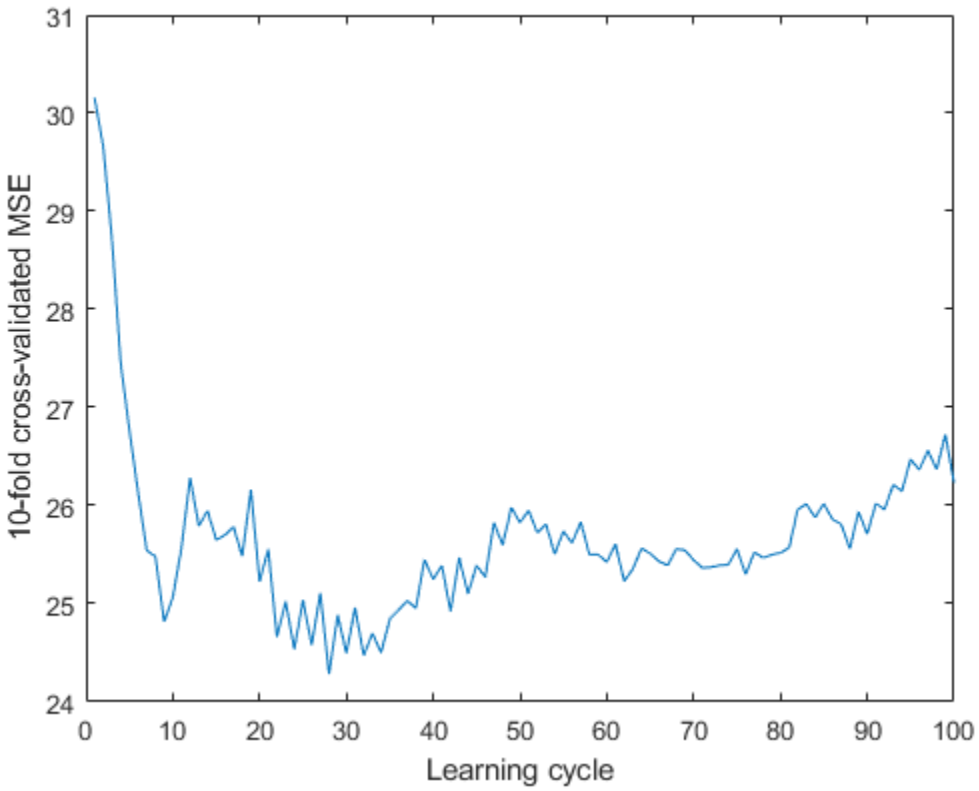
Cross-validate an ensemble of regression trees using 10-fold cross-validation. Using a decision tree template, specify that each tree should be a split once only.

```
rng(1); % For reproducibility
t = templateTree('MaxNumSplits',1);
Mdl = fitrensemble(X,MPG,'Learners',t,'CrossVal','on');
```

`Mdl` is a `RegressionPartitionedEnsemble` model.

Plot the cumulative, 10-fold cross-validated, mean-squared error (MSE). Display the estimated generalization error of the ensemble.

```
kflc = kfoldLoss(Mdl,'Mode','cumulative');
figure;
plot(kflc);
ylabel('10-fold cross-validated MSE');
xlabel('Learning cycle');
```




```
estGenError = kflc(end)
```

```
estGenError = 26.2356
```

`kfoldLoss` returns the generalization error by default. However, plotting the cumulative loss allows you to monitor how the loss changes as weak learners accumulate in the ensemble.

The ensemble achieves an MSE of around 23.5 after accumulating about 30 weak learners.

If you are satisfied with the generalization error of the ensemble, then, to create a predictive model, train the ensemble again using all of the settings except cross-validation. However, it is good practice to tune hyperparameters such as the maximum number of decision splits per tree and the number of learning cycles..

Optimize Regression Ensemble

This example shows how to optimize hyperparameters automatically using `fitrensemble`. The example uses the `carsmall` data.

Load the data.

```
load carsmall
```

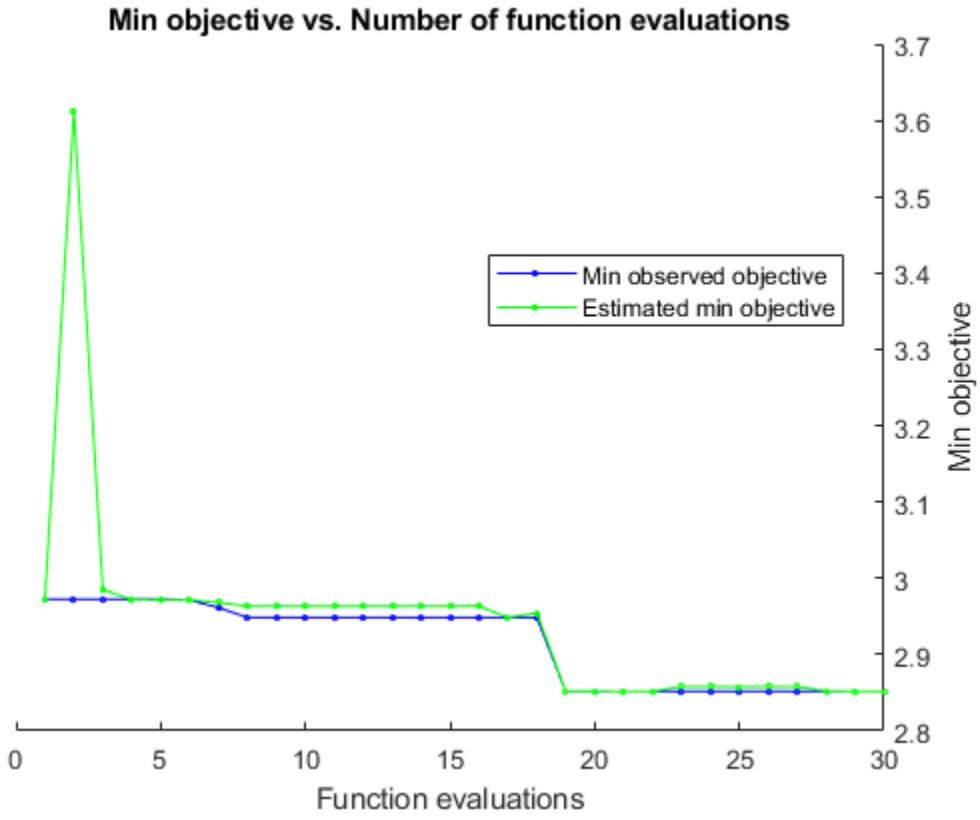
You can find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

```
Mdl = fitrensemble([Horsepower,Weight],MPG,'OptimizeHyperparameters','auto')
```

In this example, for reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function. Also, for reproducibility of random forest algorithm, specify the `'Reproducible'` name-value pair argument as `true` for tree learners.

```
rng('default')
t = templateTree('Reproducible',true);
Mdl = fitrensemble([Horsepower,Weight],MPG,'OptimizeHyperparameters','auto','Learners',t, ...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName','expected-improvement-p
```

```
|=====================================================================================================
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |       Method | NumLearn
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              | ycles
|=====================================================================================================
|    1 | Best   |       2.9726 |       18.409 |       2.9726 |       2.9726 |          Bag |
|    2 | Accept |       6.2619 |       2.4114 |       2.9726 |       3.6133 |      LSBoost |
|    3 | Accept |       2.9975 |       1.1166 |       2.9726 |       2.9852 |          Bag |
|    4 | Accept |       4.1897 |       2.1627 |       2.9726 |        2.972 |          Bag |
|    5 | Accept |       6.3321 |       2.8207 |       2.9726 |       2.9715 |      LSBoost |
|    6 | Best   |       2.9714 |       1.3552 |       2.9714 |       2.9715 |          Bag |
|    7 | Best   |       2.9615 |       2.0096 |       2.9615 |       2.9681 |          Bag |
|    8 | Best   |       2.9487 |       4.1304 |       2.9487 |       2.9633 |          Bag |
|    9 | Accept |       4.1881 |       1.5529 |       2.9487 |       2.9634 |      LSBoost |
|   10 | Accept |       3.6848 |       8.3732 |       2.9487 |       2.9634 |      LSBoost |
|   11 | Accept |       3.4935 |       10.675 |       2.9487 |       2.9634 |      LSBoost |
|   12 | Accept |       4.1881 |       8.7065 |       2.9487 |       2.9634 |      LSBoost |
|   13 | Accept |       3.2462 |       11.662 |       2.9487 |       2.9634 |      LSBoost |
|   14 | Accept |       3.4681 |       15.903 |       2.9487 |       2.9634 |      LSBoost |
|   15 | Accept |       3.4935 |        1.726 |       2.9487 |       2.9634 |      LSBoost |
|   16 | Accept |       5.5968 |       1.2402 |       2.9487 |       2.9636 |      LSBoost |
|   17 | Accept |       6.3387 |       2.8586 |       2.9487 |       2.9487 |      LSBoost |
|   18 | Accept |       3.0275 |      0.43687 |       2.9487 |       2.9545 |          Bag |
|   19 | Best   |       2.8517 |       21.359 |       2.8517 |       2.8521 |          Bag |
|   20 | Accept |        2.904 |      0.49595 |       2.8517 |       2.8521 |          Bag |
|=====================================================================================================
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |       Method | NumLearn
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              | ycles
|=====================================================================================================
|   21 | Accept |       4.1873 |        26.26 |       2.8517 |       2.8519 |          Bag |
|   22 | Accept |       3.4552 |       24.178 |       2.8517 |       2.8519 |      LSBoost |
|   23 | Accept |       2.8921 |       2.7981 |       2.8517 |       2.8578 |          Bag |
|   24 | Accept |       3.0779 |       1.3836 |       2.8517 |       2.8578 |      LSBoost |
|   25 | Accept |       2.9363 |       2.8504 |       2.8517 |       2.8577 |      LSBoost |
|   26 | Accept |       3.4972 |       18.743 |       2.8517 |       2.8578 |      LSBoost |
|   27 | Accept |       2.8988 |       1.5132 |       2.8517 |       2.8579 |      LSBoost |
|   28 | Accept |       3.5779 |      0.71655 |       2.8517 |       2.8523 |      LSBoost |
|   29 | Accept |       4.1881 |       21.054 |       2.8517 |       2.8518 |      LSBoost |
|   30 | Accept |       3.4165 |      0.52229 |       2.8517 |       2.8518 |      LSBoost |
```




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 295.891 seconds
Total objective function evaluation time: 219.4244

Best observed feasible point:
    Method    NumLearningCycles    LearnRate    MinLeafSize
    ______    _________________    _________    ___________

     Bag             496               NaN           5

Observed objective function value = 2.8517
Estimated objective function value = 2.8518
Function evaluation time = 21.3586

Best estimated feasible point (according to models):
    Method    NumLearningCycles    LearnRate    MinLeafSize
    ______    _________________    _________    ___________

     Bag             496               NaN           5

Estimated objective function value = 2.8518
Estimated function evaluation time = 22.9034

Mdl =
  RegressionBaggedEnsemble
```

```
                      ResponseName: 'Y'
             CategoricalPredictors: []
                 ResponseTransform: 'none'
                   NumObservations: 94
HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                        NumTrained: 496
                            Method: 'Bag'
                      LearnerNames: {'Tree'}
            ReasonForTermination: 'Terminated normally after completing the requested numbe
                           FitInfo: []
                FitInfoDescription: 'None'
                    Regularization: []
                         FResample: 1
                           Replace: 1
                UseObsForLearner: [94x496 logical]


  Properties, Methods
```

The optimization searched over the methods for regression (`Bag` and `LSBoost`), over `NumLearningCycles`, over the `LearnRate` for `LSBoost`, and over the tree learner `MinLeafSize`. The output is the ensemble regression with the minimum estimated cross-validation loss.

Optimize Regression Ensemble Using Cross-Validation

One way to create an ensemble of boosted regression trees that has satisfactory predictive performance is to tune the decision tree complexity level using cross-validation. While searching for an optimal complexity level, tune the learning rate to minimize the number of learning cycles as well.

This example manually finds optimal parameters by using the cross-validation option (the `'KFold'` name-value pair argument) and the `kfoldLoss` function. Alternatively, you can use the `'OptimizeHyperparameters'` name-value pair argument to optimize hyperparameters automatically. See “Optimize Regression Ensemble” on page 33-2314.

Load the `carsmall` data set. Choose the number of cylinders, volume displaced by the cylinders, horsepower, and weight as predictors of fuel economy.

```
load carsmall
Tbl = table(Cylinders,Displacement,Horsepower,Weight,MPG);
```

The default values of the tree depth controllers for boosting regression trees are:

- `10` for `MaxNumSplits`.
- `5` for `MinLeafSize`
- `10` for `MinParentSize`

To search for the optimal tree-complexity level:

1. Cross-validate a set of ensembles. Exponentially increase the tree-complexity level for subsequent ensembles from decision stump (one split) to at most n - 1 splits. n is the sample size. Also, vary the learning rate for each ensemble between 0.1 to 1.

2. Estimate the cross-validated mean-squared error (MSE) for each ensemble.
3. For tree-complexity level j, $j = 1 \ldots J$, compare the cumulative, cross-validated MSE of the ensembles by plotting them against number of learning cycles. Plot separate curves for each learning rate on the same figure.
4. Choose the curve that achieves the minimal MSE, and note the corresponding learning cycle and learning rate.

Cross-validate a deep regression tree and a stump. Because the data contain missing values, use surrogate splits. These regression trees serve as benchmarks.

```
rng(1) % For reproducibility
MdlDeep = fitrtree(Tbl,'MPG','CrossVal','on','MergeLeaves','off', ...
    'MinParentSize',1,'Surrogate','on');
MdlStump = fitrtree(Tbl,'MPG','MaxNumSplits',1,'CrossVal','on', ...
    'Surrogate','on');
```

Cross-validate an ensemble of 150 boosted regression trees using 5-fold cross-validation. Using a tree template:

- Vary the maximum number of splits using the values in the sequence $\{2^0, 2^1, \ldots, 2^m\}$. m is such that 2^m is no greater than n - 1.
- Turn on surrogate splits.

For each variant, adjust the learning rate using each value in the set {0.1, 0.25, 0.5, 1}.

```
n = size(Tbl,1);
m = floor(log2(n - 1));
learnRate = [0.1 0.25 0.5 1];
numLR = numel(learnRate);
maxNumSplits = 2.^(0:m);
numMNS = numel(maxNumSplits);
numTrees = 150;
Mdl = cell(numMNS,numLR);

for k = 1:numLR
    for j = 1:numMNS
        t = templateTree('MaxNumSplits',maxNumSplits(j),'Surrogate','on');
        Mdl{j,k} = fitrensemble(Tbl,'MPG','NumLearningCycles',numTrees, ...
            'Learners',t,'KFold',5,'LearnRate',learnRate(k));
    end
end
```

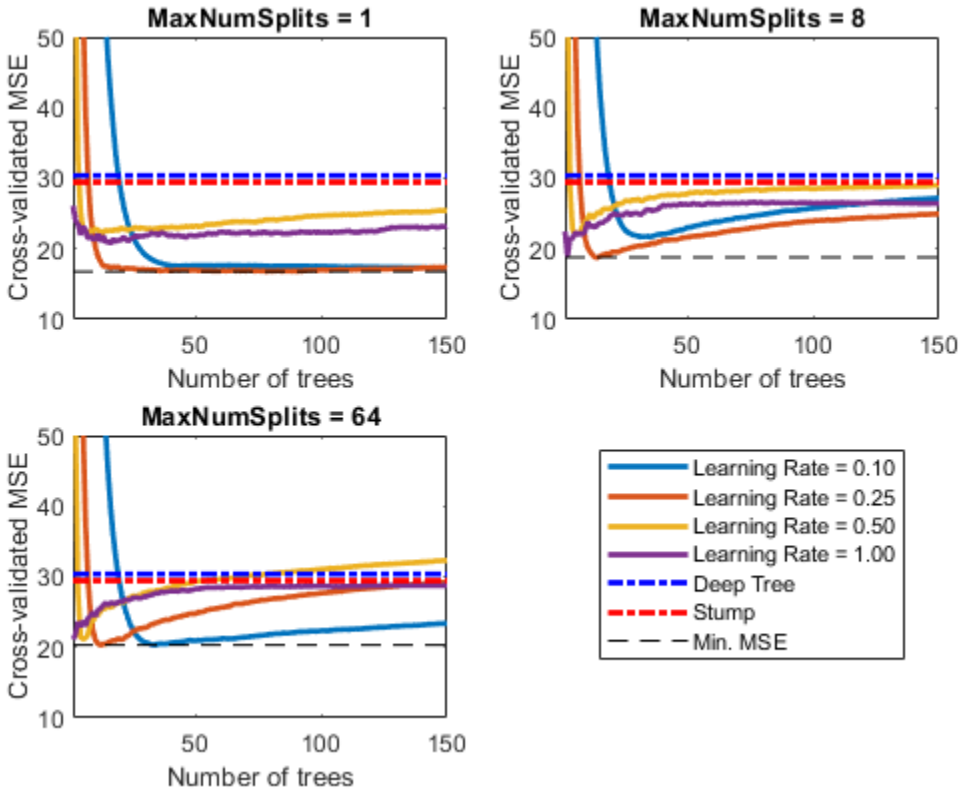
Estimate the cumulative, cross-validated MSE of each ensemble.

```
kflAll = @(x)kfoldLoss(x,'Mode','cumulative');
errorCell = cellfun(kflAll,Mdl,'Uniform',false);
error = reshape(cell2mat(errorCell),[numTrees numel(maxNumSplits) numel(learnRate)]);
errorDeep = kfoldLoss(MdlDeep);
errorStump = kfoldLoss(MdlStump);
```

Plot how the cross-validated MSE behaves as the number of trees in the ensemble increases. Plot the curves with respect to learning rate on the same plot, and plot separate plots for varying tree-complexity levels. Choose a subset of tree complexity levels to plot.

```
mnsPlot = [1 round(numel(maxNumSplits)/2) numel(maxNumSplits)];
figure;
```

```
for k = 1:3
    subplot(2,2,k)
    plot(squeeze(error(:,mnsPlot(k),:)),'LineWidth',2)
    axis tight
    hold on
    h = gca;
    plot(h.XLim,[errorDeep errorDeep],'-.b','LineWidth',2)
    plot(h.XLim,[errorStump errorStump],'-.r','LineWidth',2)
    plot(h.XLim,min(min(error(:,mnsPlot(k),:))).*[1 1],'--k')
    h.YLim = [10 50];
    xlabel('Number of trees')
    ylabel('Cross-validated MSE')
    title(sprintf('MaxNumSplits = %0.3g', maxNumSplits(mnsPlot(k))))
    hold off
end
hL = legend([cellstr(num2str(learnRate','Learning Rate = %0.2f')); ...
        'Deep Tree';'Stump';'Min. MSE']);
hL.Position(1) = 0.6;
```



Each curve contains a minimum cross-validated MSE occurring at the optimal number of trees in the ensemble.

Identify the maximum number of splits, number of trees, and learning rate that yields the lowest MSE overall.

```
[minErr,minErrIdxLin] = min(error(:));
[idxNumTrees,idxMNS,idxLR] = ind2sub(size(error),minErrIdxLin);
fprintf('\nMin. MSE = %0.5f',minErr)
```

```
Min. MSE = 16.77593

fprintf('\nOptimal Parameter Values:\nNum. Trees = %d',idxNumTrees);

Optimal Parameter Values:
Num. Trees = 78

fprintf('\nMaxNumSplits = %d\nLearning Rate = %0.2f\n',...
    maxNumSplits(idxMNS),learnRate(idxLR))

MaxNumSplits = 1
Learning Rate = 0.25
```

Create a predictive ensemble based on the optimal hyperparameters and the entire training set.

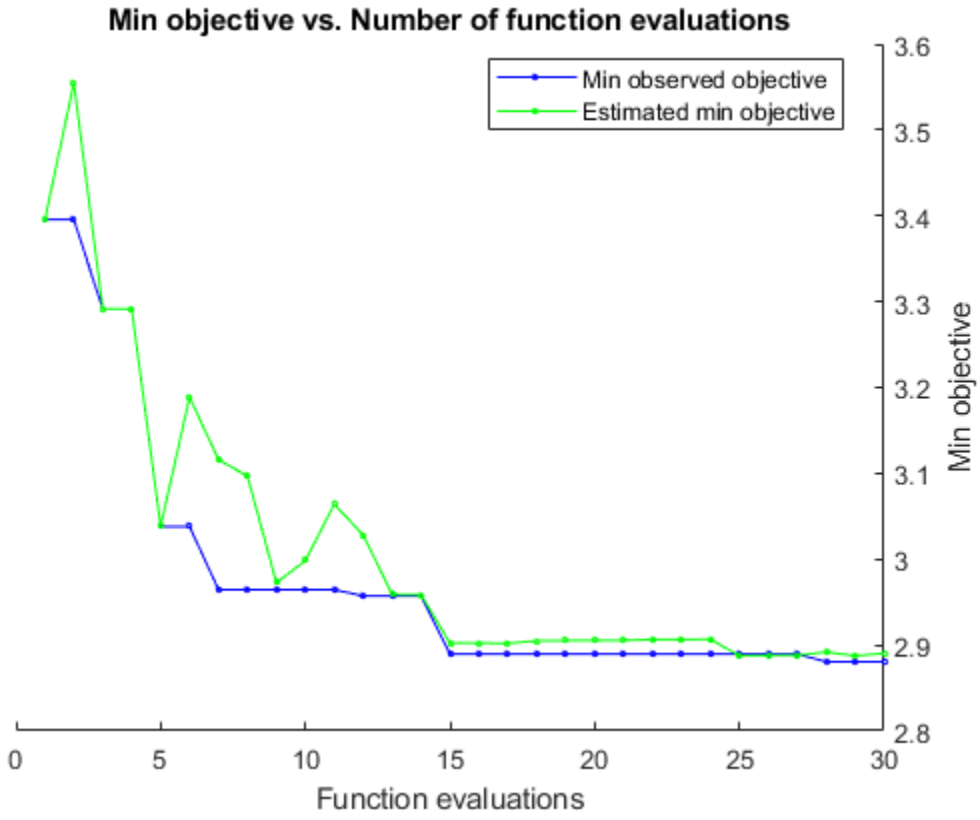
```
tFinal = templateTree('MaxNumSplits',maxNumSplits(idxMNS),'Surrogate','on');
MdlFinal = fitrensemble(Tbl,'MPG','NumLearningCycles',idxNumTrees, ...
    'Learners',tFinal,'LearnRate',learnRate(idxLR))

MdlFinal =
  RegressionEnsemble
           PredictorNames: {1x4 cell}
             ResponseName: 'MPG'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94
               NumTrained: 78
                   Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [78x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

`MdlFinal` is a `RegressionEnsemble`. To predict the fuel economy of a car given its number of cylinders, volume displaced by the cylinders, horsepower, and weight, you can pass the predictor data and `MdlFinal` to `predict`.

Instead of searching optimal values manually by using the cross-validation option (`'KFold'`) and the `kfoldLoss` function, you can use the `'OptimizeHyperparameters'` name-value pair argument. When you specify `'OptimizeHyperparameters'`, the software finds optimal parameters automatically using Bayesian optimization. The optimal values obtained by using `'OptimizeHyperparameters'` can be different from those obtained using manual search.

```
t = templateTree('Surrogate','on');
mdl = fitrensemble(Tbl,'MPG','Learners',t, ...
    'OptimizeHyperparameters',{'NumLearningCycles','LearnRate','MaxNumSplits'})

|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   | NumLearningC-|    Learn
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    | ycles        |
|=====================================================================================================
|    1 | Best   |      3.3955 |       1.274 |      3.3955 |      3.3955 |           26 |     0.01
|    2 | Accept |      6.0976 |      6.5794 |      3.3955 |      3.5549 |          170 |    0.001
|    3 | Best   |      3.2914 |      10.972 |      3.2914 |      3.2917 |          273 |      0.0
```

```
|    4 | Accept |      6.1839 |      3.1871 |      3.2914 |      3.2915 |           80 |    0.001
|    5 | Best   |      3.0379 |     0.89567 |      3.0379 |      3.0384 |           18 |      0.2
|    6 | Accept |      3.3628 |     0.49197 |      3.0379 |      3.1888 |           10 |      0.1
|    7 | Best   |      2.9646 |     0.52503 |      2.9646 |      3.1158 |           10 |      0.2
|    8 | Accept |      3.0528 |     0.50685 |      2.9646 |      3.0968 |           10 |      0.2
|    9 | Accept |      2.9789 |     0.47815 |      2.9646 |      2.9727 |           10 |      0.2
|   10 | Accept |      3.0605 |     0.46351 |      2.9646 |      2.9984 |           10 |      0.8
|   11 | Accept |       3.161 |     0.88622 |      2.9646 |      3.0639 |           21 |      0.4
|   12 | Best   |      2.9571 |     0.46911 |      2.9571 |      3.0282 |           10 |      0.4
|   13 | Accept |      6.1344 |     0.47308 |      2.9571 |      2.9588 |           10 |     0.01
|   14 | Accept |      2.9729 |       0.466 |      2.9571 |      2.9586 |           10 |      0.5
|   15 | Best   |      2.8895 |     0.46415 |      2.8895 |      2.9022 |           10 |      0.3
|   16 | Accept |      2.9254 |      2.5157 |      2.8895 |      2.9023 |           69 |      0.1
|   17 | Accept |      2.9254 |      1.3488 |      2.8895 |       2.902 |           35 |        0
|   18 | Accept |      2.9271 |     0.62763 |      2.8895 |      2.9048 |           14 |      0.3
|   19 | Accept |      2.9254 |      4.0751 |      2.8895 |      2.9051 |          116 |     0.06
|   20 | Accept |      2.8966 |       7.672 |      2.8895 |      2.9053 |          223 |      0.1
|======|========|=============|=============|=============|=============|==============|=========
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   | NumLearningC-|    Learn
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    | ycles        |
|======|========|=============|=============|=============|=============|==============|=========
|   21 | Accept |      2.9346 |      1.9984 |      2.8895 |      2.9054 |           49 |      0.2
|   22 | Accept |      3.0867 |      6.7252 |      2.8895 |      2.9063 |          123 |      0.0
|   23 | Accept |      2.8982 |      18.831 |      2.8895 |      2.9064 |          460 |     0.05
|   24 | Accept |       2.915 |       19.45 |      2.8895 |      2.9066 |          498 |        0
|   25 | Best   |      2.8894 |      23.436 |      2.8894 |      2.8875 |          475 |     0.08
|   26 | Accept |      2.9043 |      15.603 |      2.8894 |      2.8879 |          267 |     0.00
|   27 | Accept |      2.9468 |      1.1468 |      2.8894 |      2.8879 |           22 |      0.2
|   28 | Best   |      2.8805 |      7.1523 |      2.8805 |      2.8918 |          146 |      0.1
|   29 | Accept |       3.338 |     0.76699 |      2.8805 |      2.8874 |           11 |      0.1
|   30 | Accept |      2.9067 |      4.6689 |      2.8805 |      2.8903 |           97 |      0.1
```




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 192.1184 seconds
Total objective function evaluation time: 144.1501

Best observed feasible point:
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           146             0.19004          1

Observed objective function value = 2.8805
Estimated objective function value = 2.8903
Function evaluation time = 7.1523

Best estimated feasible point (according to models):
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           146             0.19004          1

Estimated objective function value = 2.8903
Estimated function evaluation time = 6.1119

mdl =
  RegressionEnsemble
```

```
                    PredictorNames: {1x4 cell}
                      ResponseName: 'MPG'
             CategoricalPredictors: []
                 ResponseTransform: 'none'
                   NumObservations: 94
HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                       NumTrained: 146
                           Method: 'LSBoost'
                     LearnerNames: {'Tree'}
              ReasonForTermination: 'Terminated normally after completing the requested numbe
                          FitInfo: [146x1 double]
                FitInfoDescription: {2x1 cell}
                    Regularization: []


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable and you want to use all remaining variables as predictors, then specify the response variable using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use a subset of the remaining variables only as predictors, then specify a formula using `formula`.
- If `Tbl` does not contain the response variable, then specify the response data using `Y`. The length of response variable and the number of rows of `Tbl` must be equal.

Note To save memory and execution time, supply `X` and `Y` instead of `Tbl`.

Data Types: `table`

`ResponseVarName` — Response variable name

name of response variable in `Tbl`

Response variable name, specified as the name of the response variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl.Y` is the response variable, then specify `ResponseVarName` as `'Y'`. Otherwise, `fitrensemble` treats all columns of `Tbl` as predictor variables.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

X — Predictor data
numeric matrix

Predictor data, specified as numeric matrix.

Each row corresponds to one observation, and each column corresponds to one predictor variable.

The length of `Y` and the number of rows of `X` must be equal.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

Y — Response
numeric vector

Response, specified as a numeric vector. Each element in `Y` is the response to the observation in the corresponding row of `X` or `Tbl`. The length of `Y` and the number of rows of `X` or `Tbl` must be equal.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'NumLearningCycles',500,'Method','Bag','Learners',templateTree(),'CrossVal','on'` cross-validates an ensemble of 500 bagged regression trees using 10-fold cross-validation.

General Ensemble Options

`Method` — Ensemble aggregation method

`'LSBoost'` (default) | `'Bag'`

Ensemble aggregation method, specified as the comma-separated pair consisting of `'Method'` and `'LSBoost'` or `'Bag'`.

Value	Method	Notes
`'LSBoost'`	Least-squares boosting (LSBoost)	You can specify the learning rate for shrinkage by using the `'LearnRate'` name-value pair argument.
`'Bag'`	Bootstrap aggregation (bagging, for example, random forest[2])	`fitrensemble` uses bagging with random predictor selections at each split (random forest) by default. To use bagging without the random selections, use tree learners whose `'NumVariablesToSample'` value is `'all'`.

For details about ensemble aggregation algorithms and examples, see “Algorithms” on page 33-2337, “Ensemble Algorithms” on page 18-39, and “Choose an Applicable Ensemble Aggregation Method” on page 18-32.

Example: `'Method','Bag'`

`NumLearningCycles` — Number of ensemble learning cycles

`100` (default) | positive integer

Number of ensemble learning cycles, specified as the comma-separated pair consisting of `'NumLearningCycles'` and a positive integer. At every learning cycle, the software trains one weak learner for every template object in `Learners`. Consequently, the software trains `NumLearningCycles*numel(Learners)` learners.

The software composes the ensemble using all trained learners and stores them in `Mdl.Trained`.

For more details, see “Tips” on page 33-2337.

Example: `'NumLearningCycles',500`

Data Types: `single` | `double`

`Learners` — Weak learners to use in ensemble

`'tree'` (default) | tree template object | cell vector of tree template objects

Weak learners to use in the ensemble, specified as the comma-separated pair consisting of `'Learners'` and `'tree'`, a tree template object, or a cell vector of tree template objects.

- `'tree'` (default) — `fitrensemble` uses default regression tree learners, which is the same as using `templateTree()`. The default values of `templateTree()` depend on the value of `'Method'`.

- For bagged decision trees, the maximum number of decision splits (`'MaxNumSplits'`) is `n–1`, where `n` is the number of observations. The number of predictors to select at random for each split (`'NumVariablesToSample'`) is one third of the number of predictors. Therefore, `fitrensemble` grows deep decision trees. You can grow shallower trees to reduce model complexity or computation time.
- For boosted decision trees, `'MaxNumSplits'` is 10 and `'NumVariablesToSample'` is `'all'`. Therefore, `fitrensemble` grows shallow decision trees. You can grow deeper trees for better accuracy.

See `templateTree` for the default settings of a weak learner.

- Tree template object — `fitrensemble` uses the tree template object created by `templateTree`. Use the name-value pair arguments of `templateTree` to specify settings of the tree learners.
- Cell vector of *m* tree template objects — `fitrensemble` grows *m* regression trees per learning cycle (see `NumLearningCycles`). For example, for an ensemble composed of two types of regression trees, supply `{t1 t2}`, where `t1` and `t2` are regression tree template objects returned by `templateTree`.

To obtain reproducible results, you must specify the `'Reproducible'` name-value pair argument of `templateTree` as `true` if `'NumVariablesToSample'` is not `'all'`.

For details on the number of learners to train, see `NumLearningCycles` and “Tips” on page 33-2337.

Example: `'Learners',templateTree('MaxNumSplits',5)`

`NPrint` — Printout frequency
`'off'` (default) | positive integer

Printout frequency, specified as the comma-separated pair consisting of `'NPrint'` and a positive integer or `'off'`.

To track the number of *weak learners* or *folds* that `fitrensemble` trained so far, specify a positive integer. That is, if you specify the positive integer *m*:

- Without also specifying any cross-validation option (for example, `CrossVal`), then `fitrensemble` displays a message to the command line every time it completes training *m* weak learners.
- And a cross-validation option, then `fitrensemble` displays a message to the command line every time it finishes training *m* folds.

If you specify `'off'`, then `fitrensemble` does not display a message when it completes training weak learners.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Example: `'NPrint',5`

Data Types: `single` | `double` | `char` | `string`

`NumBins` — Number of bins for numeric predictors
`[]` (empty) (default) | positive integer scalar

Number of bins for numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

- If the `'NumBins'` value is empty (default), then `fitrensemble` does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitrensemble` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.
  - The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
  - `fitrensemble` does not bin categorical predictors.

When you use a large training data set, this binning option speeds up training but might cause a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the value depending on the accuracy and training speed.

A trained model stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrensemble` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrensemble` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrensemble` assumes that all

predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrensemble` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

Parallel Options

`Options` — Options for computing in parallel and setting random numbers

structure

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`. Also, use a tree template with the `'Reproducible'` name-value argument set to `true`. See “Reproducibility in Parallel Statistical Computations” on page 31-13.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `fitrensemble` uses the default stream or streams.

For an example using reproducible parallel training, see “Train Classification Ensemble in Parallel” on page 18-109.

For dual-core systems and above, `fitrensemble` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Cross-Validation Options

`CrossVal` — Cross-validation flag

`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, cross-validate later by passing `Mdl` to `crossval` or `crossval`.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition

`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation

scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds

`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag

`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Other Regression Options

`Weights` — Observation weights

numeric vector of positive values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or name of a variable in `Tbl`. The software weighs the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows of `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response when training the model.

The software normalizes the values of `Weights` to sum to 1.

By default, `Weights` is `ones(n,1)`, where *n* is the number of observations in `X` or `Tbl`.

Data Types: `double` | `single` | `char` | `string`

Sampling Options

`FResample` — Fraction of training set to resample
1 (default) | positive scalar in (0,1]

Fraction of the training set to resample for every weak learner, specified as the comma-separated pair consisting of `'FResample'` and a positive scalar in (0,1].

To use `'FResample'`, specify `'bag'` for `Method` or set `Resample` to `'on'`.

Example: `'FResample',0.75`

Data Types: `single` | `double`

`Replace` — Flag indicating to sample with replacement
`'on'` (default) | `'off'`

Flag indicating sampling with replacement, specified as the comma-separated pair consisting of `'Replace'` and `'off'` or `'on'`.

- For `'on'`, the software samples the training observations with replacement.
- For `'off'`, the software samples the training observations without replacement. If you set `Resample` to `'on'`, then the software samples training observations assuming uniform weights. If you also specify a boosting method, then the software boosts by reweighting observations.

Unless you set `Method` to `'bag'` or set `Resample` to `'on'`, `Replace` has no effect.

Example: `'Replace','off'`

`Resample` — Flag indicating to resample
`'off'` | `'on'`

Flag indicating to resample, specified as the comma-separated pair consisting of `'Resample'` and `'off'` or `'on'`.

- If `Method` is a boosting method, then:
  - `'Resample','on'` specifies to sample training observations using updated weights as the multinomial sampling probabilities.
  - `'Resample','off'`(default) specifies to reweight observations at every learning iteration.
- If `Method` is `'bag'`, then `'Resample'` must be `'on'`. The software resamples a fraction of the training observations (see `FResample`) with or without replacement (see `Replace`).

If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of 1 for `FResample`.

LSBoost Method Options

`LearnRate` — Learning rate for shrinkage
1 (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'Method','NumLearningCycles','LearnRate'}` along with the default parameters for the specified `Learners`:
  - `Learners = 'tree'` (default) — `{'MinLeafSize'}`

  Note For hyperparameter optimization, `Learners` must be a single argument, not a string array or cell array.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`

The optimization attempts to minimize the cross-validation loss (error) for `fitrensemble` by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitrensemble` are:

- `Method` — Eligible methods are `'Bag'` or `'LSBoost'`.
- `NumLearningCycles` — `fitrensemble` searches among positive integers, by default log-scaled with range `[10,500]`.
- `LearnRate` — `fitrensemble` searches among positive reals, by default log-scaled with range `[1e-3,1]`.
- `MinLeafSize` — `fitrensemble` searches among integers log-scaled in the range `[1,max(2,floor(NumObservations/2))]`.
- `MaxNumSplits` — `fitrensemble` searches among integers log-scaled in the range `[1,max(2,NumObservations-1)]`.
- `NumVariablesToSample` — `fitrensemble` searches among integers in the range `[1,max(2,NumPredictors)]`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load carsmall
params = hyperparameters('fitrensemble',[Horsepower,Weight],MPG,'Tree');
params(4).Range = [1,20];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Regression Ensemble” on page 33-2314.

Example: `'OptimizeHyperparameters',{'Method','NumLearningCycles','LearnRate','MinLeafSize','MaxNumSplits'}`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`

Field Name	Values	Default
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`

Field Name	Values	Default
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see "Parallel Bayesian Optimization" on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained regression ensemble model

`RegressionBaggedEnsemble` model object | `RegressionEnsemble` model object | `RegressionPartitionedEnsemble` cross-validated model object

Trained ensemble model, returned as one of the model objects in this table.

Model Object	Specify Any Cross-Validation Options?	`Method` Setting	`Resample` Setting
`RegressionBaggedEnsemble`	No	`'Bag'`	`'on'`

Model Object	Specify Any Cross-Validation Options?	Method Setting	Resample Setting
`RegressionEnsemble`	No	`'LSBoost'`	`'off'`
`RegressionPartitionedEnsemble`	Yes	`'LSBoost'` or `'Bag'`	`'off'` or `'on'`

The name-value pair arguments that control cross-validation are `CrossVal`, `Holdout`, `KFold`, `Leaveout`, and `CVPartition`.

To reference properties of `Mdl`, use dot notation. For example, to access or display the cell vector of weak learner model objects for an ensemble that has not been cross-validated, enter `Mdl.Trained` at the command line.

Tips

- `NumLearningCycles` can vary from a few dozen to a few thousand. Usually, an ensemble with good predictive power requires from a few hundred to a few thousand weak learners. However, you do not have to train an ensemble for that many cycles at once. You can start by growing a few dozen learners, inspect the ensemble performance and then, if necessary, train more weak learners using `resume`.
- Ensemble performance depends on the ensemble setting and the setting of the weak learners. That is, if you specify weak learners with default parameters, then the ensemble can perform poorly. Therefore, like ensemble settings, it is good practice to adjust the parameters of the weak learners using templates, and to choose values that minimize generalization error.
- If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of `1` for `FResample`.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- For details of ensemble aggregation algorithms, see “Ensemble Algorithms” on page 18-39.
- If you specify `'Method','LSBoost'`, then the software grows shallow decision trees by default. You can adjust tree depth by specifying the `MaxNumSplits`, `MinLeafSize`, and `MinParentSize` name-value pair arguments using `templateTree`.
- For dual-core systems and above, `fitrensemble` parallelizes training using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L. “Bagging Predictors.” *Machine Learning*. Vol. 26, pp. 123–140, 1996.

[2] Breiman, L. “Random Forests.” *Machine Learning*. Vol. 45, pp. 5–32, 2001.

[3] Freund, Y. and R. E. Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting.” *J. of Computer and System Sciences*, Vol. 55, pp. 119–139, 1997.

[4] Friedman, J. “Greedy function approximation: A gradient boosting machine.” *Annals of Statistics*, Vol. 29, No. 5, pp. 1189–1232, 2001.

[5] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning* section edition, Springer, New York, 2008.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`fitrensemble` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionBaggedEnsemble` | `RegressionEnsemble` | `RegressionPartitionedEnsemble` | `predict` | `templateTree`

Topics

“Supervised Learning Workflow and Algorithms” on page 18-3
“Framework for Ensemble Learning” on page 18-31
“Ensemble Algorithms” on page 18-39

Introduced in R2016b

fitrsvm

Fit a support vector machine regression model

Syntax

```
Mdl = fitrsvm(Tbl,ResponseVarName)
Mdl = fitrsvm(Tbl,formula)
Mdl = fitrsvm(Tbl,Y)

Mdl = fitrsvm(X,Y)

Mdl = fitrsvm( ___ ,Name,Value)
```

Description

`fitrsvm` trains or cross-validates a support vector machine (SVM) regression model on a low- through moderate-dimensional predictor data set. `fitrsvm` supports mapping the predictor data using kernel functions, and supports SMO, ISDA, or $L1$ soft-margin minimization via quadratic programming for objective-function minimization.

To train a linear SVM regression model on a high-dimensional data set, that is, data sets that include many predictor variables, use `fitrlinear` instead.

To train an SVM model for binary classification, see `fitcsvm` for low- through moderate-dimensional predictor data sets, or `fitclinear` for high-dimensional data sets.

`Mdl = fitrsvm(Tbl,ResponseVarName)` returns a full, trained support vector machine (SVM) regression model `Mdl` trained using the predictors values in the table `Tbl` and the response values in `Tbl.ResponseVarName`.

`Mdl = fitrsvm(Tbl,formula)` returns a full SVM regression model trained using the predictors values in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitrsvm(Tbl,Y)` returns a full, trained SVM regression model trained using the predictors values in the table `Tbl` and the response values in the vector `Y`.

`Mdl = fitrsvm(X,Y)` returns a full, trained SVM regression model trained using the predictors values in the matrix `X` and the response values in the vector `Y`.

`Mdl = fitrsvm( ___ ,Name,Value)` returns an SVM regression model with additional options specified by one or more name-value pair arguments, using any of the previous syntaxes. For example, you can specify the kernel function or train a cross-validated model.

Examples

Train Linear Support Vector Machine Regression Model

Train a support vector machine (SVM) regression model using sample data stored in matrices.

Load the `carsmall` data set.

```
load carsmall
rng 'default'  % For reproducibility
```

Specify `Horsepower` and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Horsepower,Weight];
Y = MPG;
```

Train a default SVM regression model.

```
Mdl = fitrsvm(X,Y)
```

```
Mdl =
  RegressionSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                    Alpha: [75x1 double]
                     Bias: 57.3800
          KernelParameters: [1x1 struct]
           NumObservations: 93
            BoxConstraints: [93x1 double]
           ConvergenceInfo: [1x1 struct]
           IsSupportVector: [93x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`Mdl` is a trained `RegressionSVM` model.

Check the model for convergence.

```
Mdl.ConvergenceInfo.Converged
```

```
ans = logical
   0
```

`0` indicates that the model did not converge.

Retrain the model using standardized data.

```
MdlStd = fitrsvm(X,Y,'Standardize',true)
```

```
MdlStd =
  RegressionSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                    Alpha: [77x1 double]
                     Bias: 22.9131
          KernelParameters: [1x1 struct]
                       Mu: [109.3441 2.9625e+03]
                    Sigma: [45.3545 805.9668]
           NumObservations: 93
```

```
          BoxConstraints: [93x1 double]
         ConvergenceInfo: [1x1 struct]
         IsSupportVector: [93x1 logical]
                  Solver: 'SMO'


  Properties, Methods
```

Check the model for convergence.

```
MdlStd.ConvergenceInfo.Converged
```

```
ans = logical
   1
```

1 indicates that the model did converge.

Compute the resubstitution (in-sample) mean-squared error for the new model.

```
lStd = resubLoss(MdlStd)
```

```
lStd = 17.0256
```

Train Support Vector Machine Regression Model

Train a support vector machine regression model using the abalone data from the UCI Machine Learning Repository.

Download the data and save it in your current folder with the name `'abalone.csv'`.

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data';
websave('abalone.csv',url);
```

Read the data into a table. Specify the variable names.

```
varnames = {'Sex'; 'Length'; 'Diameter'; 'Height'; 'Whole_weight';...
    'Shucked_weight'; 'Viscera_weight'; 'Shell_weight'; 'Rings'};
Tbl = readtable('abalone.csv','Filetype','text','ReadVariableNames',false);
Tbl.Properties.VariableNames = varnames;
```

The sample data contains 4177 observations. All the predictor variables are continuous except for `Sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings (stored in `Rings`) on the abalone and determine its age using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with an automatic kernel scale. Standardize the data.

```
rng default  % For reproducibility
Mdl = fitrsvm(Tbl,'Rings','KernelFunction','gaussian','KernelScale','auto',...
    'Standardize',true)
```

```
Mdl =
```

```
  RegressionSVM
           PredictorNames: {1×8 cell}
             ResponseName: 'Rings'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635×1 double]
                     Bias: 10.8144
          KernelParameters: [1×1 struct]
                       Mu: [1×10 double]
                    Sigma: [1×10 double]
           NumObservations: 4177
            BoxConstraints: [4177×1 double]
           ConvergenceInfo: [1×1 struct]
           IsSupportVector: [4177×1 logical]
                    Solver: 'SMO'
```

The Command Window shows that `Mdl` is a trained `RegressionSVM` model and displays a property list.

Display the properties of `Mdl` using dot notation. For example, check to confirm whether the model converged and how many iterations it completed.

```
conv = Mdl.ConvergenceInfo.Converged
iter = Mdl.NumIterations
```

```
conv =

  logical

   1


iter =

        2759
```

The returned results indicate that the model converged after 2759 iterations.

Cross-Validate SVM Regression Model

Load the `carsmall` data set.

```
load carsmall
rng 'default'  % For reproducibility
```

Specify `Horsepower` and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Horsepower Weight];
Y = MPG;
```

Cross-validate two SVM regression models using 5-fold cross-validation. For both models, specify to standardize the predictors. For one of the models, specify to train using the default linear kernel, and the Gaussian kernel for the other model.

```
MdlLin = fitrsvm(X,Y,'Standardize',true,'KFold',5)
```

```
MdlLin =
  RegressionPartitionedSVM
    CrossValidatedModel: 'SVM'
         PredictorNames: {'x1'  'x2'}
           ResponseName: 'Y'
        NumObservations: 94
                  KFold: 5
              Partition: [1x1 cvpartition]
      ResponseTransform: 'none'


  Properties, Methods
```

```
MdlGau = fitrsvm(X,Y,'Standardize',true,'KFold',5,'KernelFunction','gaussian')
```

```
MdlGau =
  RegressionPartitionedSVM
    CrossValidatedModel: 'SVM'
         PredictorNames: {'x1'  'x2'}
           ResponseName: 'Y'
        NumObservations: 94
                  KFold: 5
              Partition: [1x1 cvpartition]
      ResponseTransform: 'none'


  Properties, Methods
```

```
MdlLin.Trained
```

```
ans=5×1 cell array
    {1x1 classreg.learning.regr.CompactRegressionSVM}
    {1x1 classreg.learning.regr.CompactRegressionSVM}
    {1x1 classreg.learning.regr.CompactRegressionSVM}
    {1x1 classreg.learning.regr.CompactRegressionSVM}
    {1x1 classreg.learning.regr.CompactRegressionSVM}
```

`MdlLin` and `MdlGau` are `RegressionPartitionedSVM` cross-validated models. The `Trained` property of each model is a 5-by-1 cell array of `CompactRegressionSVM` models. The models in the cell store the results of training on 4 folds of observations, and leaving one fold of observations out.

Compare the generalization error of the models. In this case, the generalization error is the out-of-sample mean-squared error.

```
mseLin = kfoldLoss(MdlLin)
```

```
mseLin = 17.4417
```

```
mseGau = kfoldLoss(MdlGau)
```

```
mseGau = 16.7333
```

The SVM regression model using the Gaussian kernel performs better than the one using the linear kernel.

Create a model suitable for making predictions by passing the entire data set to `fitrsvm`, and specify all name-value pair arguments that yielded the better-performing model. However, do not specify any cross-validation options.

```
MdlGau = fitrsvm(X,Y,'Standardize',true,'KernelFunction','gaussian');
```

To predict the MPG of a set of cars, pass `Mdl` and a table containing the horsepower and weight measurements of the cars to `predict`.

Optimize SVM Regression

This example shows how to optimize hyperparameters automatically using `fitrsvm`. The example uses the `carsmall` data.

Load the `carsmall` data set.

```
load carsmall
```

Specify `Horsepower` and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Horsepower Weight];
Y = MPG;
```

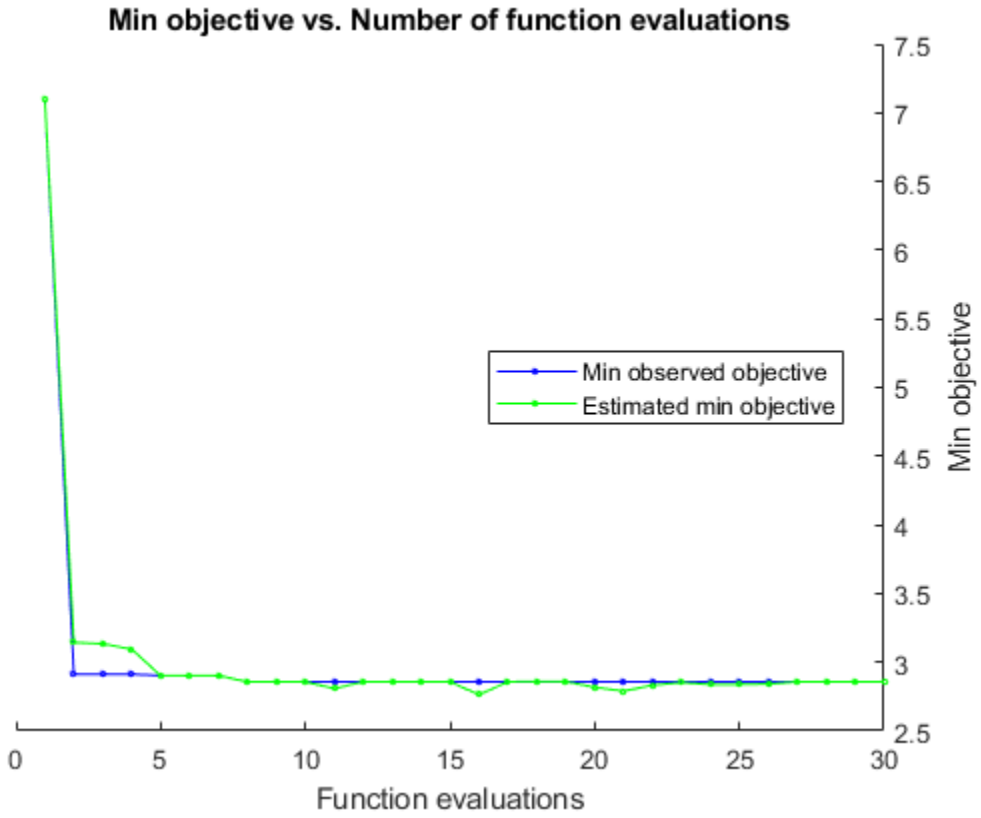
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
Mdl = fitrsvm(X,Y,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName',...
    'expected-improvement-plus'))
```

```
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |      7.0963 |      17.231 |      7.0963 |      7.0963 |      0.35664 |     0.04
|    2 | Best   |      2.9121 |     0.14958 |      2.9121 |       3.143 |        70.67 |       71
|    3 | Accept |      4.1884 |     0.13475 |      2.9121 |      3.1342 |       14.367 |    0.005
|    4 | Accept |       4.159 |     0.11746 |      2.9121 |       3.094 |    0.0030879 |       71
|    5 | Best   |      2.9027 |     0.29517 |      2.9027 |       2.902 |       969.07 |        1
|    6 | Accept |      4.1884 |     0.10564 |      2.9027 |      2.9023 |       993.93 |       91
|    7 | Accept |      2.9321 |     0.15536 |      2.9027 |      2.9024 |       221.61 |       59
|    8 | Best   |      2.8565 |     0.54052 |      2.8565 |      2.8567 |       940.66 |       35
|    9 | Accept |      4.1884 |     0.11782 |      2.8565 |      2.8567 |       514.19 |       51
|   10 | Accept |      2.9223 |     0.18092 |      2.8565 |       2.857 |       59.312 |       29
|   11 | Accept |      2.9033 |     0.18672 |      2.8565 |      2.8076 |       134.08 |       48
|   12 | Accept |      2.8628 |      12.579 |      2.8565 |      2.8571 |       999.43 |       74
```

```
|   13 | Accept |      2.9711 |      13.884 |      2.8565 |      2.8571 |       852.78 |        6
|   14 | Accept |      3.0558 |     0.11595 |      2.8565 |      2.8571 |     0.015133 |       15
|   15 | Accept |      2.8967 |      2.3951 |      2.8565 |      2.8574 |       989.37 |        2
|   16 | Accept |      4.1884 |     0.14236 |      2.8565 |      2.7627 |       964.51 |       33
|   17 | Accept |      2.9391 |      5.1201 |      2.8565 |      2.8572 |       993.07 |       10
|   18 | Accept |      2.9392 |     0.15187 |      2.8565 |      2.8599 |       912.86 |       98
|   19 | Accept |      2.9355 |     0.14043 |      2.8565 |      2.8583 |       1.1277 |       38
|   20 | Accept |      2.9084 |     0.14044 |      2.8565 |      2.8167 |       3.9293 |       93
|======================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   | BoxConstraint|  KernelS
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|======================================================================================================
|   21 | Accept |      2.9307 |     0.11397 |      2.8565 |      2.7894 |    0.0010707 |       22
|   22 | Accept |      4.5677 |      13.282 |      2.8565 |      2.8319 |       971.48 |       11
|   23 | Accept |      3.0589 |     0.12041 |      2.8565 |      2.8517 |    0.0012016 |       47
|   24 | Accept |      2.9353 |     0.14985 |      2.8565 |      2.8377 |       19.318 |       13
|   25 | Accept |      2.9386 |    0.079296 |      2.8565 |      2.8378 |    0.0010304 |       28
|   26 | Accept |      4.1884 |     0.11381 |      2.8565 |      2.8411 |    0.0018407 |    0.001
|   27 | Accept |      2.9259 |     0.36396 |      2.8565 |      2.8572 |       992.81 |       46
|   28 | Accept |      2.9209 |       0.119 |      2.8565 |      2.8573 |       252.51 |       15
|   29 | Accept |      3.6378 |    0.098302 |      2.8565 |      2.8573 |     0.073707 |       98
|   30 | Accept |      2.9167 |     0.11261 |      2.8565 |      2.8573 |    0.0010818 |       17
```




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
```

```
Total elapsed time: 134.5666 seconds
Total objective function evaluation time: 68.4375

Best observed feasible point:
    BoxConstraint    KernelScale    Epsilon
    _____________    ___________    _______

       940.66          350.46       5.2391

Observed objective function value = 2.8565
Estimated objective function value = 2.8573
Function evaluation time = 0.54052

Best estimated feasible point (according to models):
    BoxConstraint    KernelScale    Epsilon
    _____________    ___________    _______

       940.66          350.46       5.2391

Estimated objective function value = 2.8573
Estimated function evaluation time = 0.5406

Mdl =
  RegressionSVM
                          ResponseName: 'Y'
                 CategoricalPredictors: []
                     ResponseTransform: 'none'
                                 Alpha: [19x1 double]
                                  Bias: 49.4606
                      KernelParameters: [1x1 struct]
                       NumObservations: 93
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                        BoxConstraints: [93x1 double]
                       ConvergenceInfo: [1x1 struct]
                       IsSupportVector: [93x1 logical]
                                Solver: 'SMO'


  Properties, Methods
```

The optimization searched over `BoxConstraint`, `KernelScale`, and `Epsilon`. The output is the regression with the minimum estimated cross-validation loss.

Input Arguments

`Tbl` — Predictor data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula using `formula`.

If `Tbl` does not contain the response variable, then specify a response variable using `Y`. The length of response variable and the number of rows of `Tbl` must be equal.

If a row of `Tbl` or an element of `Y` contains at least one `NaN`, then `fitrsvm` removes those rows and elements from both arguments when training the model.

To specify the names of the predictors in the order of their appearance in `Tbl`, use the `PredictorNames` name-value pair argument.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl` stores the response variable `Y` as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Response data
numeric vector

Response data, specified as an *n*-by-1 numeric vector. The length of `Y` and the number of rows of `Tbl` or `X` must be equal.

If a row of `Tbl` or `X`, or an element of `Y`, contains at least one `NaN`, then `fitrsvm` removes those rows and elements from both arguments when training the model.

To specify the response variable name, use the `ResponseName` name-value pair argument.

Data Types: `single` | `double`

X — Predictor data

numeric matrix

Predictor data to which the SVM regression model is fit, specified as an *n*-by-*p* numeric matrix. *n* is the number of observations and *p* is the number of predictor variables.

The length of `Y` and the number of rows of `X` must be equal.

If a row of `X` or an element of `Y` contains at least one `NaN`, then `fitrsvm` removes those rows and elements from both arguments.

To specify the names of the predictors in the order of their appearance in `X`, use the `PredictorNames` name-value pair argument.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'KernelFunction','gaussian','Standardize',true,'CrossVal','on'` trains a 10-fold cross-validated SVM regression model using a Gaussian kernel and standardized training data.

Support Vector Machine Options

BoxConstraint — Box constraint

positive scalar value

Box constraint for the alpha coefficients, specified as the comma-separated pair consisting of `'BoxConstraint'` and a positive scalar value.

The absolute value of the `Alpha` coefficients cannot exceed the value of `BoxConstraint`.

The default `BoxConstraint` value for the `'gaussian'` or `'rbf'` kernel function is `iqr(Y)/1.349`, where `iqr(Y)` is the interquartile range of response variable `Y`. For all other kernels, the default `BoxConstraint` value is 1.

Example: `BoxConstraint,10`

Data Types: `single` | `double`

KernelFunction — Kernel function

`'linear'` (default) | `'gaussian'` | `'rbf'` | `'polynomial'` | function name

Kernel function used to compute the Gram matrix on page 33-1862, specified as the comma-separated pair consisting of `'KernelFunction'` and a value in this table.

Value	Description	Formula
`'gaussian'` or `'rbf'`	Gaussian or Radial Basis Function (RBF) kernel	$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right)$
`'linear'`	Linear kernel	$G(x_j, x_k) = x_j' x_k$
`'polynomial'`	Polynomial kernel. Use `'PolynomialOrder',q` to specify a polynomial kernel of order q.	$G(x_j, x_k) = (1 + x_j' x_k)^q$

You can set your own kernel function, for example, `kernel`, by setting `'KernelFunction','kernel'`. `kernel` must have the following form:

```
function G = kernel(U,V)
```

where:

- `U` is an m-by-p matrix.
- `V` is an n-by-p matrix.
- `G` is an m-by-n Gram matrix of the rows of `U` and `V`.

And `kernel.m` must be on the MATLAB path.

It is good practice to avoid using generic names for kernel functions. For example, call a sigmoid kernel function `'mysigmoid'` rather than `'sigmoid'`.

Example: `'KernelFunction','gaussian'`

Data Types: `char` | `string`

`KernelScale` — Kernel scale parameter

1 (default) | `'auto'` | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software divides all elements of the predictor matrix `X` by the value of `KernelScale`. Then, the software applies the appropriate kernel norm to compute the Gram matrix.

- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed using `rng` before training.
- If you specify `KernelScale` and your own kernel function, for example, `'KernelFunction','kernel'`, then the software throws an error. You must apply scaling within `kernel`.

Example: `'KernelScale','auto'`

Data Types: `double` | `single` | `char` | `string`

`PolynomialOrder` — Polynomial kernel function order

3 (default) | positive integer

Polynomial kernel function order, specified as the comma-separated pair consisting of `'PolynomialOrder'` and a positive integer.

If you set 'PolynomialOrder' and KernelFunction is not 'polynomial', then the software throws an error.

Example: 'PolynomialOrder',2

Data Types: double | single

KernelOffset — Kernel offset parameter
nonnegative scalar

Kernel offset parameter, specified as the comma-separated pair consisting of 'KernelOffset' and a nonnegative scalar.

The software adds KernelOffset to each element of the Gram matrix.

The defaults are:

- 0 if the solver is SMO (that is, you set 'Solver','SMO')
- 0.1 if the solver is ISDA (that is, you set 'Solver','ISDA')

Example: 'KernelOffset',0

Data Types: double | single

Epsilon — Half the width of epsilon-insensitive band
iqr(Y)/13.49 (default) | nonnegative scalar value

Half the width of the epsilon-insensitive band, specified as the comma-separated pair consisting of 'Epsilon' and a nonnegative scalar value.

The default Epsilon value is iqr(Y)/13.49, which is an estimate of a tenth of the standard deviation using the interquartile range of the response variable Y. If iqr(Y) is equal to zero, then the default Epsilon value is 0.1.

Example: 'Epsilon',0.3

Data Types: single | double

Standardize — Flag to standardize predictor data
false (default) | true

Flag to standardize the predictor data, specified as the comma-separated pair consisting of 'Standardize' and true (1) or false (0).

If you set 'Standardize',true:

- The software centers and scales each column of the predictor data (X) by the weighted column mean and standard deviation, respectively (for details on weighted standardizing, see “Algorithms” on page 33-1865). MATLAB does not standardize the data contained in the dummy variable columns generated for categorical predictors.
- The software trains the model using the standardized predictor matrix, but stores the unstandardized data in the model property X.

Example: 'Standardize',true

Data Types: logical

`Solver` — Optimization routine

`'ISDA'` | `'L1QP'` | `'SMO'`

Optimization routine, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description
`'ISDA'`	Iterative Single Data Algorithm (see [30])
`'L1QP'`	Uses `quadprog` to implement $L1$ soft-margin minimization by quadratic programming. This option requires an Optimization Toolbox license. For more details, see “Quadratic Programming Definition” (Optimization Toolbox).
`'SMO'`	Sequential Minimal Optimization (see [17])

The defaults are:

- `'ISDA'` if you set `'OutlierFraction'` to a positive value
- `'SMO'` otherwise

Example: `'Solver','ISDA'`

`Alpha` — Initial estimates of alpha coefficients

numeric vector

Initial estimates of alpha coefficients, specified as the comma-separated pair consisting of `'Alpha'` and a numeric vector. The length of `Alpha` must be equal to the number of rows of `X`.

- Each element of `Alpha` corresponds to an observation in `X`.
- `Alpha` cannot contain any `NaN`s.
- If you specify `Alpha` and any one of the cross-validation name-value pair arguments (`'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`), then the software returns an error.

If `Y` contains any missing values, then remove all rows of `Y`, `X`, and `Alpha` that correspond to the missing values. That is, enter:

```
idx = ~isnan(Y);
Y = Y(idx);
X = X(idx,:);
alpha = alpha(idx);
```

Then, pass `Y`, `X`, and `alpha` as the response, predictors, and initial alpha estimates, respectively.

The default is `zeros(size(Y,1))`.

Example: `'Alpha',0.1*ones(size(X,1),1)`

Data Types: `single` | `double`

`CacheSize` — Cache size

`1000` (default) | `'maximal'` | positive scalar

Cache size, specified as the comma-separated pair consisting of `'CacheSize'` and `'maximal'` or a positive scalar.

If `CacheSize` is `'maximal'`, then the software reserves enough memory to hold the entire *n*-by-*n* Gram matrix on page 33-1862.

If `CacheSize` is a positive scalar, then the software reserves `CacheSize` megabytes of memory for training the model.

Example: `'CacheSize','maximal'`

Data Types: `double` | `single` | `char` | `string`

`ClipAlphas` — Flag to clip alpha coefficients
`true` (default) | `false`

Flag to clip alpha coefficients, specified as the comma-separated pair consisting of `'ClipAlphas'` and either `true` or `false`.

Suppose that the alpha coefficient for observation j is α_j and the box constraint of observation j is C_j, $j = 1,...,n$, where n is the training sample size.

Value	Description
`true`	At each iteration, if α_j is near 0 or near C_j, then MATLAB sets α_j to 0 or to C_j, respectively.
`false`	MATLAB does not change the alpha coefficients during optimization.

MATLAB stores the final values of α in the `Alpha` property of the trained SVM model object.

`ClipAlphas` can affect SMO and ISDA convergence.

Example: `'ClipAlphas',false`

Data Types: `logical`

`NumPrint` — Number of iterations between optimization diagnostic message output
`1000` (default) | nonnegative integer

Number of iterations between optimization diagnostic message output, specified as the comma-separated pair consisting of `'NumPrint'` and a nonnegative integer.

If you specify `'Verbose',1` and `'NumPrint',numprint`, then the software displays all optimization diagnostic messages from SMO and ISDA every `numprint` iterations in the Command Window.

Example: `'NumPrint',500`

Data Types: `double` | `single`

`OutlierFraction` — Expected proportion of outliers in training data
0 (default) | numeric scalar in the interval [0,1)

Expected proportion of outliers in training data, specified as the comma-separated pair consisting of `'OutlierFraction'` and a numeric scalar in the interval [0,1). `fitrsvm` removes observations with large gradients, ensuring that `fitrsvm` removes the fraction of observations specified by `OutlierFraction` by the time convergence is reached. This name-value pair is only valid when `'Solver'` is `'ISDA'`.

Example: `'OutlierFraction',0.1`

Data Types: `single` | `double`

`RemoveDuplicates` — Flag to replace duplicate observations with single observations
`false` (default) | `true`

Flag to replace duplicate observations with single observations in the training data, specified as the comma-separated pair consisting of `'RemoveDuplicates'` and `true` or `false`.

If `RemoveDuplicates` is `true`, then `fitrsvm` replaces duplicate observations in the training data with a single observation of the same value. The weight of the single observation is equal to the sum of the weights of the corresponding removed duplicates (see `Weights`).

Tip If your data set contains many duplicate observations, then specifying `'RemoveDuplicates',true` can decrease convergence time considerably.

Data Types: `logical`

`Verbose` — Verbosity level
`0` (default) | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. The value of `Verbose` controls the amount of optimization information that the software displays in the Command Window and saves the information as a structure to `Mdl.ConvergenceInfo.History`.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software does not display or save convergence information.
`1`	The software displays diagnostic messages and saves convergence criteria every `numprint` iterations, where `numprint` is the value of the name-value pair argument `'NumPrint'`.
`2`	The software displays diagnostic messages and saves convergence criteria at every iteration.

Example: `'Verbose',1`

Data Types: `double` | `single`

Other Regression Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrsvm` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrsvm` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrsvm` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitrsvm` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitrsvm` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitrsvm` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names

string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.

- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrsvm` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of numeric values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of numeric values. The size of `Weights` must equal the number of rows in `X`. `fitrsvm` normalizes the values of `Weights` to sum to 1.

Data Types: `single` | `double`

Cross-Validation Options

`CrossVal` — Cross-validation flag
`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'CrossVal'` and either `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

To override this cross-validation setting, use one of these name-value pair arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`. To create a cross-validated model, you can use one cross-validation name-value pair argument at a time only.

Alternatively, you can cross-validate the model later using the `crossval` method.

Example: `'CrossVal','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag

`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as `'on'` or `'off'`. If you specify `'Leaveout','on'`, then for each of the *n* observations (where *n* is the number of observations, excluding missing observations, specified in the `NumObservations` property of the model), the software completes these steps:

1. Reserve the one observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Convergence Controls

`DeltaGradientTolerance` — Tolerance for gradient difference

0 (default) | nonnegative scalar

Tolerance for gradient difference between upper and lower violators obtained by SMO or ISDA, specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar.

Example: `'DeltaGradientTolerance',1e-4`

Data Types: `single` | `double`

`GapTolerance` — Feasibility gap tolerance

`1e-3` (default) | nonnegative scalar

Feasibility gap tolerance obtained by SMO or ISDA, specified as the comma-separated pair consisting of `'GapTolerance'` and a nonnegative scalar.

If `GapTolerance` is `0`, then `fitrsvm` does not use this parameter to check convergence.

Example: `'GapTolerance',1e-4`

Data Types: `single` | `double`

`IterationLimit` — Maximal number of numerical optimization iterations

`1e6` (default) | positive integer

Maximal number of numerical optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The software returns a trained model regardless of whether the optimization routine successfully converges. `Mdl.ConvergenceInfo` contains convergence information.

Example: `'IterationLimit',1e8`

Data Types: `double` | `single`

KKTTolerance — Tolerance for KKT violation
0 | nonnegative scalar value

Tolerance for Karush-Kuhn-Tucker (KKT) violation, specified as the comma-separated pair consisting of `'KKTTolerance'` and a nonnegative scalar value.

This name-value pair applies only if `'Solver'` is `'SMO'` or `'ISDA'`.

If `KKTTolerance` is `0`, then `fitrsvm` does not use this parameter to check convergence.

Example: `'KKTTolerance',1e-4`

Data Types: `single` | `double`

ShrinkagePeriod — Number of iterations between reductions of active set
`0` (default) | nonnegative integer

Number of iterations between reductions of the active set, specified as the comma-separated pair consisting of `'ShrinkagePeriod'` and a nonnegative integer.

If you set `'ShrinkagePeriod',0`, then the software does not shrink the active set.

Example: `'ShrinkagePeriod',1000`

Data Types: `double` | `single`

Hyperparameter Optimization

OptimizeHyperparameters — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'BoxConstraint','KernelScale','Epsilon'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitrsvm` by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitrsvm` are:

- `BoxConstraint` — `fitrsvm` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.

- `KernelScale` — `fitrsvm` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.
- `Epsilon` — `fitrsvm` searches among positive values, by default log-scaled in the range `[1e-3,1e2]*iqr(Y)/1.349`.
- `KernelFunction` — `fitrsvm` searches among `'gaussian'`, `'linear'`, and `'polynomial'`.
- `PolynomialOrder` — `fitrsvm` searches among integers in the range `[2,4]`.
- `Standardize` — `fitrsvm` searches among `'true'` and `'false'`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load carsmall
params = hyperparameters('fitrsvm',[Horsepower,Weight],MPG);
params(1).Range = [1e-4,1e6];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize SVM Regression” on page 33-2344.

Example: `'OptimizeHyperparameters','auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`

Field Name	Values	Default
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`

Field Name	Values	Default
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see "Parallel Bayesian Optimization" on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained SVM regression model

`RegressionSVM` model | `RegressionPartitionedSVM` cross-validated model

Trained SVM regression model, returned as a `RegressionSVM` model or `RegressionPartitionedSVM` cross-validated model.

If you set any of the name-value pair arguments `KFold`, `Holdout`, `Leaveout`, `CrossVal`, or `CVPartition`, then `Mdl` is a `RegressionPartitionedSVM` cross-validated model. Otherwise, `Mdl` is a `RegressionSVM` model.

Limitations

`fitrsvm` supports low- through moderate-dimensional data sets. For high-dimensional data set, use `fitrlinear` instead.

Tips

- Unless your data set is large, always try to standardize the predictors (see `Standardize`). Standardization makes predictors insensitive to the scales on which they are measured.
- It is good practice to cross-validate using the `KFold` name-value pair argument. The cross-validation results determine how well the SVM model generalizes.
- Sparsity in support vectors is a desirable property of an SVM model. To decrease the number of support vectors, set the `BoxConstraint` name-value pair argument to a large value. This action also increases the training time.
- For optimal training time, set `CacheSize` as high as the memory limit on your computer allows.
- If you expect many fewer support vectors than observations in the training set, then you can significantly speed up convergence by shrinking the active-set using the name-value pair argument `'ShrinkagePeriod'`. It is good practice to use `'ShrinkagePeriod',1000`.
- Duplicate observations that are far from the regression line do not affect convergence. However, just a few duplicate observations that occur near the regression line can slow down convergence considerably. To speed up convergence, specify `'RemoveDuplicates',true` if:

  - Your data set contains many duplicate observations.
  - You suspect that a few duplicate observations can fall near the regression line.

  However, to maintain the original data set during training, `fitrsvm` must temporarily store separate data sets: the original and one without the duplicate observations. Therefore, if you specify `true` for data sets containing few duplicates, then `fitrsvm` consumes close to double the memory of the original data.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- For the mathematical formulation of linear and nonlinear SVM regression problems and the solver algorithms, see “Understanding Support Vector Machine Regression” on page 25-2.
- `NaN`, `<undefined>`, empty character vector (`''`), empty string (`""`), and `<missing>` values indicate missing data values. `fitrsvm` removes entire rows of data corresponding to a missing response. When normalizing weights, `fitrsvm` ignores any weight corresponding to an observation with at least one missing predictor. Consequently, observation box constraints might not equal `BoxConstraint`.
- `fitrsvm` removes observations that have zero weight.
- If you set `'Standardize',true` and `'Weights'`, then `fitrsvm` standardizes the predictors using their corresponding weighted means and weighted standard deviations. That is, `fitrsvm` standardizes predictor j (x_j) using

$$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

- $\mu_j^* = \frac{1}{\sum_k w_k} \sum_k w_k x_{jk}$.
- x_{jk} is observation k (row) of predictor j (column).
- $\left(\sigma_j^*\right)^2 = \frac{v_1}{v_1^2 - v_2} \sum_k w_k \left(x_{jk} - \mu_j^*\right)^2$.
- $v_1 = \sum_j w_j$.
- $v_2 = \sum_j \left(w_j\right)^2$.

- If your predictor data contains categorical variables, then the software generally uses full dummy encoding for these variables. The software creates one dummy variable for each level of each categorical variable.
  - The `PredictorNames` property stores one element for each of the original predictor variable names. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `PredictorNames` is a 1-by-3 cell array of character vectors containing the original names of the predictor variables.
  - The `ExpandedPredictorNames` property stores one element for each of the predictor variables, including the dummy variables. For example, assume that there are three predictors, one of which is a categorical variable with three levels. Then `ExpandedPredictorNames` is a 1-by-5 cell array of character vectors containing the names of the predictor variables and the new dummy variables.
  - Similarly, the `Beta` property stores one beta coefficient for each predictor, including the dummy variables.
  - The `SupportVectors` property stores the predictor values for the support vectors, including the dummy variables. For example, assume that there are m support vectors and three predictors, one of which is a categorical variable with three levels. Then `SupportVectors` is an m-by-5 matrix.
  - The `X` property stores the training data as originally input. It does not include the dummy variables. When the input is a table, `X` contains only the columns used as predictors.
- For predictors specified in a table, if any of the variables contain ordered (ordinal) categories, the software uses ordinal encoding for these variables.
  - For a variable having k ordered levels, the software creates k – 1 dummy variables. The jth dummy variable is -1 for levels up to j, and +1 for levels j + 1 through k.
  - The names of the dummy variables stored in the `ExpandedPredictorNames` property indicate the first level with the value +1. The software stores k – 1 additional predictor names for the dummy variables, including the names of levels 2, 3, ..., k.
- All solvers implement $L1$ soft-margin minimization.
- Let `p` be the proportion of outliers that you expect in the training data. If you set `'OutlierFraction',p`, then the software implements robust learning. In other words, the software attempts to remove 100`p`% of the observations when the optimization algorithm converges. The removed observations correspond to gradients that are large in magnitude.

References

[1] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[2] Fan, R.-E., P.-H. Chen, and C.-J. Lin. “Working set selection using second order information for training support vector machines.” *Journal of Machine Learning Research*, Vol 6, 2005, pp. 1889–1918.

[3] Kecman V., T. -M. Huang, and M. Vogt. “Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance.” In *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[6] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`CompactRegressionSVM` | `RegressionPartitionedSVM` | `RegressionSVM` | `predict`

Topics

“Understanding Support Vector Machine Regression” on page 25-2

Introduced in R2015b

fitrtree

Fit binary decision tree for regression

Syntax

```
tree = fitrtree(Tbl,ResponseVarName)
tree = fitrtree(Tbl,formula)
tree = fitrtree(Tbl,Y)

tree = fitrtree(X,Y)

tree = fitrtree(___,Name,Value)
```

Description

`tree = fitrtree(Tbl,ResponseVarName)` returns a regression tree based on the input variables (also known as predictors, features, or attributes) in the table `Tbl` and the output (response) contained in `Tbl.ResponseVarName`. The returned `tree` is a binary tree where each branching node is split based on the values of a column of `Tbl`.

`tree = fitrtree(Tbl,formula)` returns a regression tree based on the input variables contained in the table `Tbl`. The input `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `tree`.

`tree = fitrtree(Tbl,Y)` returns a regression tree based on the input variables contained in the table `Tbl` and the output in vector `Y`.

`tree = fitrtree(X,Y)` returns a regression tree based on the input variables `X` and the output `Y`. The returned `tree` is a binary tree where each branching node is split based on the values of a column of `X`.

`tree = fitrtree(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify observation weights or train a cross-validated model.

Examples

Construct Regression Tree

Load the sample data.

```
load carsmall
```

Construct a regression tree using the sample data. The response variable is miles per gallon, MPG.

```
tree = fitrtree([Weight, Cylinders],MPG,...
                'CategoricalPredictors',2,'MinParentSize',20,...
                'PredictorNames',{'W','C'})
```

```
tree =
  RegressionTree
```

```
          PredictorNames: {'W'  'C'}
            ResponseName: 'Y'
    CategoricalPredictors: 2
        ResponseTransform: 'none'
          NumObservations: 94


  Properties, Methods
```

Predict the mileage of 4,000-pound cars with 4, 6, and 8 cylinders.

```
MPG4Kpred = predict(tree,[4000 4; 4000 6; 4000 8])
```

```
MPG4Kpred = 3×1

   19.2778
   19.2778
   14.3889
```

Control Regression Tree Depth

`fitrtree` grows deep decision trees by default. You can grow shallower trees to reduce model complexity or computation time. To control the depth of trees, use the `'MaxNumSplits'`, `'MinLeafSize'`, or `'MinParentSize'` name-value pair arguments.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

The default values of the tree-depth controllers for growing regression trees are:

- `n - 1` for `MaxNumSplits`. `n` is the training sample size.
- `1` for `MinLeafSize`.
- `10` for `MinParentSize`.

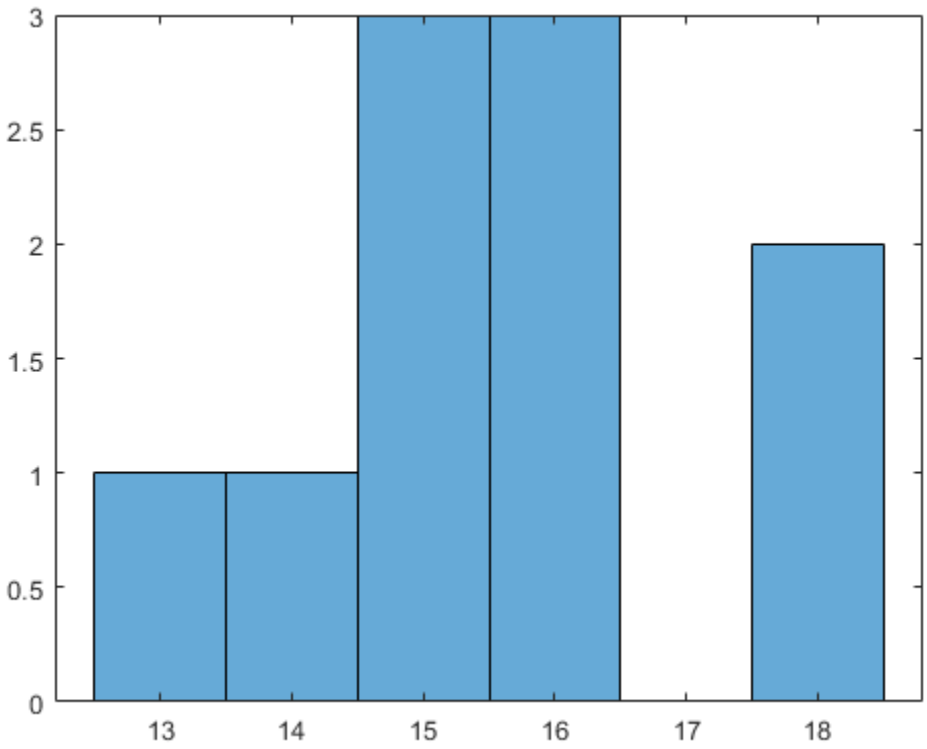
These default values tend to grow deep trees for large training sample sizes.

Train a regression tree using the default values for tree-depth control. Cross-validate the model using 10-fold cross-validation.

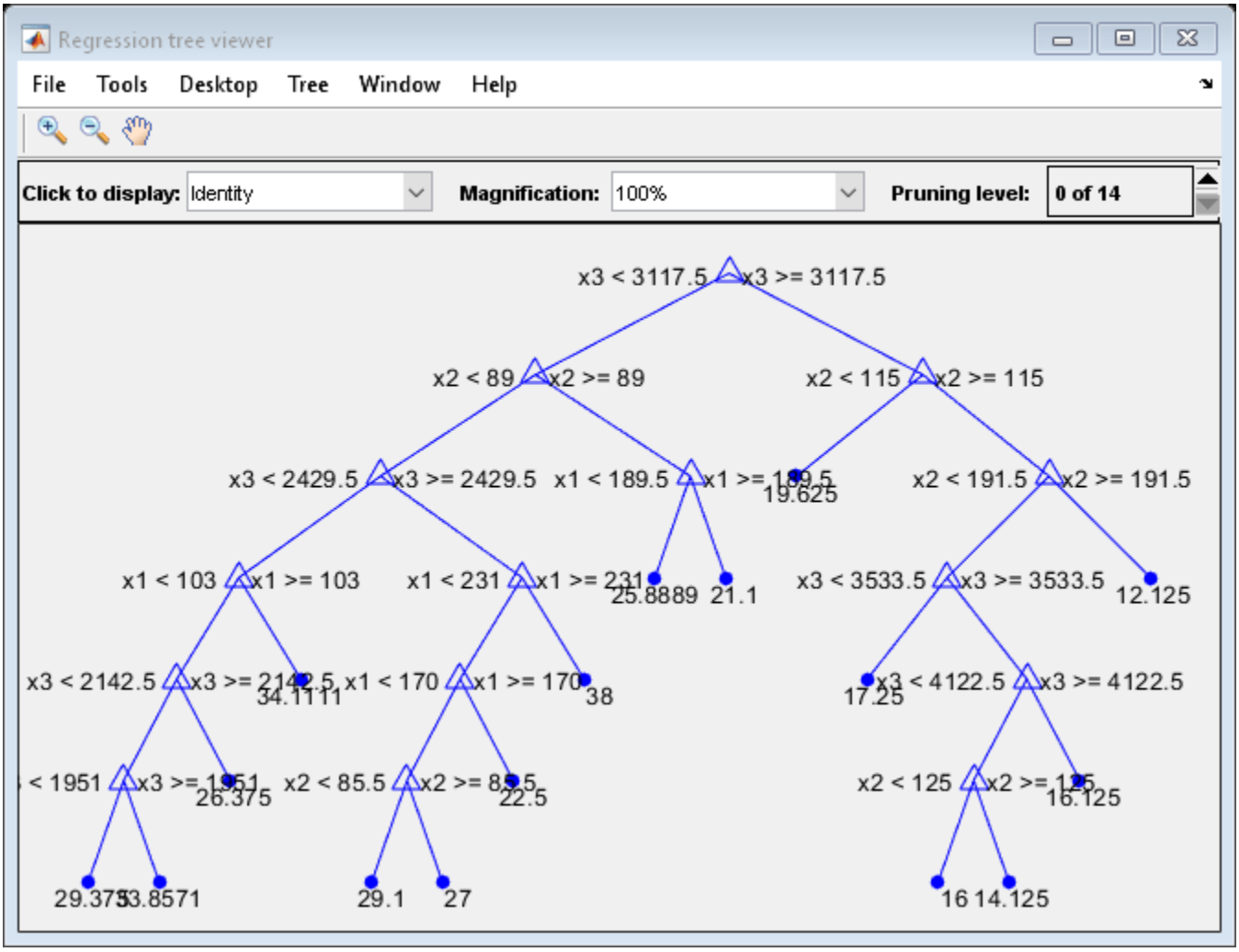
```
rng(1); % For reproducibility
MdlDefault = fitrtree(X,MPG,'CrossVal','on');
```

Draw a histogram of the number of imposed splits on the trees. The number of imposed splits is one less than the number of leaves. Also, view one of the trees.

```
numBranches = @(x)sum(x.IsBranch);
mdlDefaultNumSplits = cellfun(numBranches, MdlDefault.Trained);

figure;
histogram(mdlDefaultNumSplits)
```



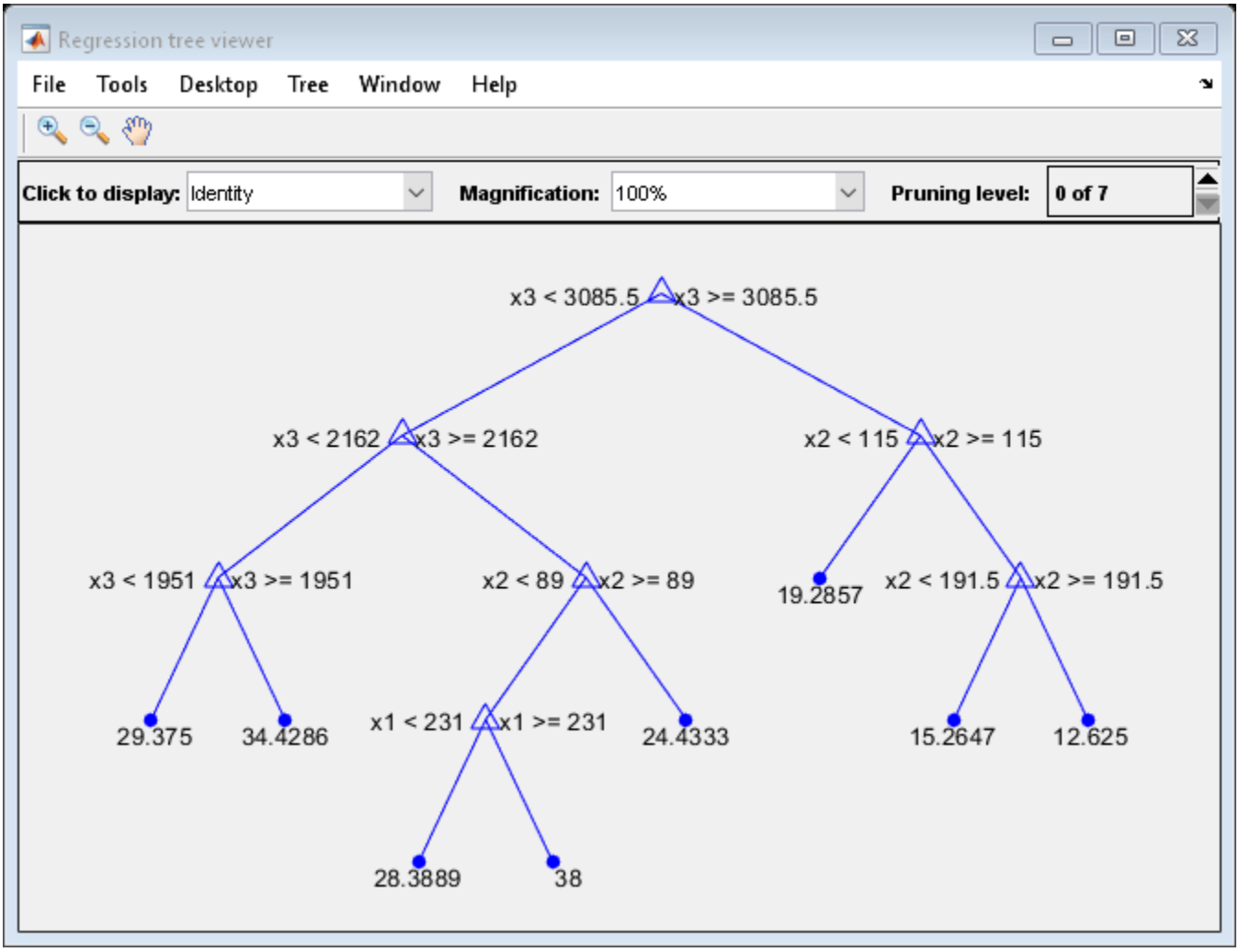

```
view(MdlDefault.Trained{1},'Mode','graph')
```



The average number of splits is between 14 and 15.

Suppose that you want a regression tree that is not as complex (deep) as the ones trained using the default number of splits. Train another regression tree, but set the maximum number of splits at 7, which is about half the mean number of splits from the default regression tree. Cross-validate the model using 10-fold cross-validation.

```
Mdl7 = fitrtree(X,MPG,'MaxNumSplits',7,'CrossVal','on');
view(Mdl7.Trained{1},'Mode','graph')
```



Compare the cross-validation mean squared errors (MSEs) of the models.

```
mseDefault = kfoldLoss(MdlDefault)
```

```
mseDefault = 25.7383
```

```
mse7 = kfoldLoss(Mdl7)
```

```
mse7 = 26.5748
```

`Mdl7` is much less complex and performs only slightly worse than `MdlDefault`.

Optimize Regression Tree

Optimize hyperparameters automatically using `fitrtree`.

Load the `carsmall` data set.

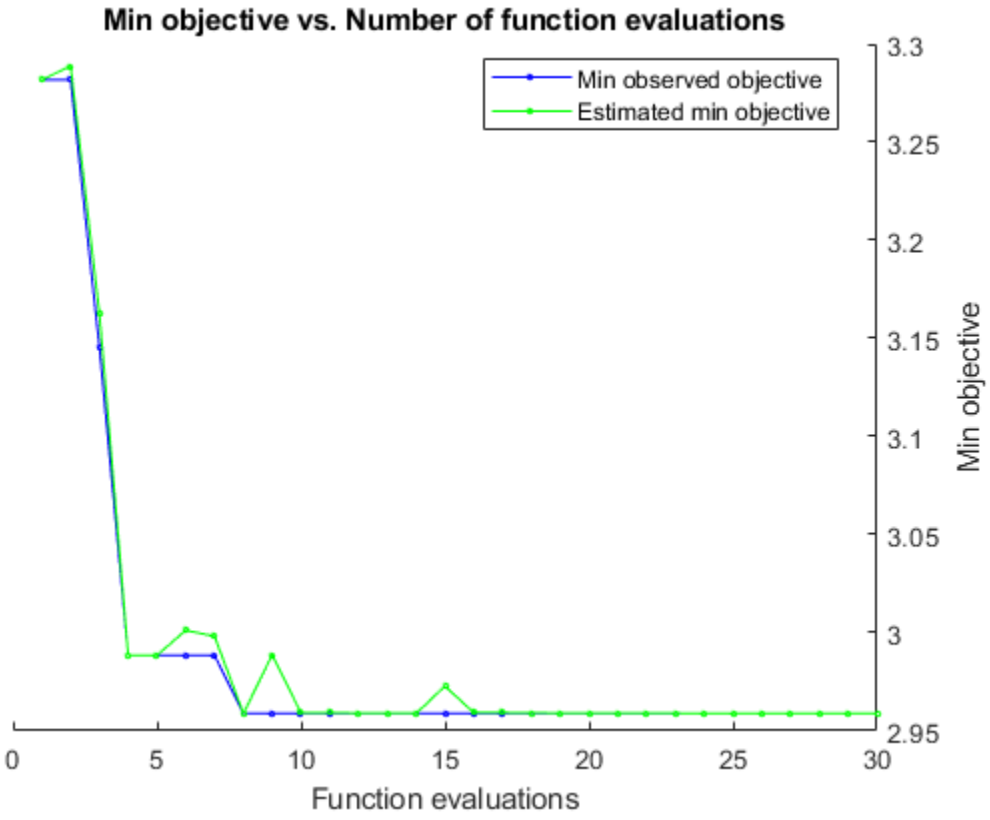
```
load carsmall
```

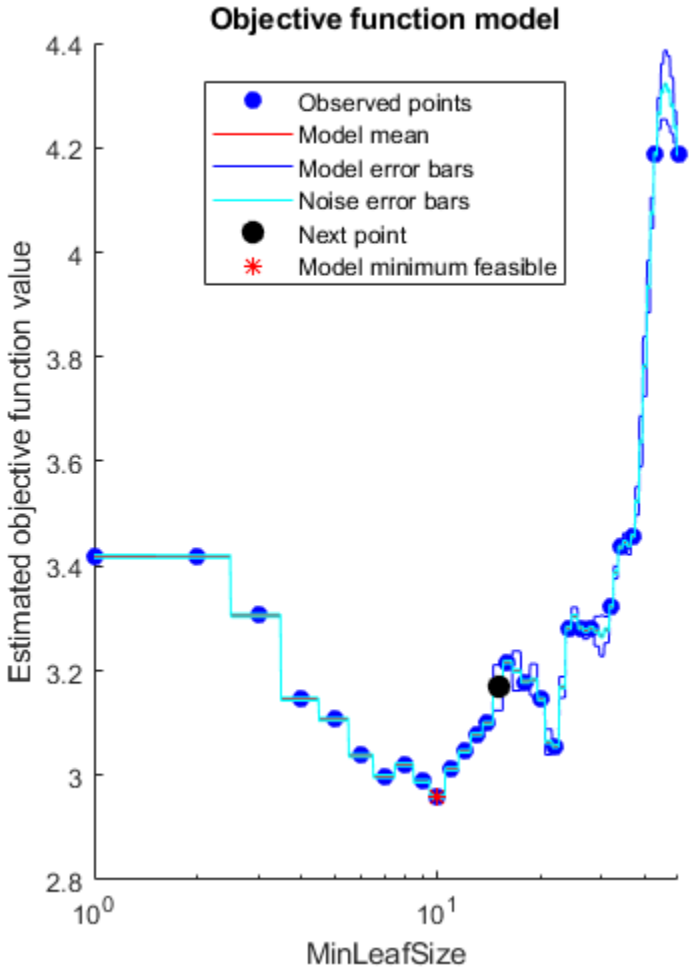
Use `Weight` and `Horsepower` as predictors for `MPG`. Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization.

For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
X = [Weight,Horsepower];
Y = MPG;
rng default
Mdl = fitrtree(X,Y,'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName',...
    'expected-improvement-plus'))
```

```
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |  MinLeafSize |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              |
|=====================================================================================================|
|    1 | Best   |       3.2818 |      0.12606 |       3.2818 |       3.2818 |           28 |
|    2 | Accept |       3.4183 |     0.080602 |       3.2818 |       3.2888 |            1 |
|    3 | Best   |       3.1457 |     0.075198 |       3.1457 |       3.1628 |            4 |
|    4 | Best   |       2.9885 |     0.068214 |       2.9885 |       2.9885 |            9 |
|    5 | Accept |       2.9978 |     0.056384 |       2.9885 |       2.9885 |            7 |
|    6 | Accept |       3.0203 |     0.062221 |       2.9885 |       3.0013 |            8 |
|    7 | Accept |       2.9885 |     0.057122 |       2.9885 |       2.9981 |            9 |
|    8 | Best   |       2.9589 |     0.069876 |       2.9589 |       2.9589 |           10 |
|    9 | Accept |        3.078 |     0.056591 |       2.9589 |       2.9888 |           13 |
|   10 | Accept |       4.1881 |     0.082283 |       2.9589 |       2.9592 |           50 |
|   11 | Accept |       3.4182 |     0.085051 |       2.9589 |       2.9592 |            2 |
|   12 | Accept |       3.0376 |     0.072658 |       2.9589 |       2.9591 |            6 |
|   13 | Accept |       3.1453 |      0.06178 |       2.9589 |       2.9591 |           20 |
|   14 | Accept |       2.9589 |     0.079448 |       2.9589 |        2.959 |           10 |
|   15 | Accept |       3.0123 |     0.084837 |       2.9589 |       2.9728 |           11 |
|   16 | Accept |       2.9589 |     0.064132 |       2.9589 |       2.9593 |           10 |
|   17 | Accept |       3.3055 |      0.09149 |       2.9589 |       2.9593 |            3 |
|   18 | Accept |       2.9589 |     0.068534 |       2.9589 |       2.9592 |           10 |
|   19 | Accept |       3.4577 |     0.066947 |       2.9589 |       2.9591 |           37 |
|   20 | Accept |       3.2166 |     0.075825 |       2.9589 |        2.959 |           16 |
|=====================================================================================================|
| Iter | Eval   | Objective:   | Objective    | BestSoFar    | BestSoFar    |  MinLeafSize |
|      | result | log(1+loss)  | runtime      | (observed)   | (estim.)     |              |
|=====================================================================================================|
|   21 | Accept |        3.107 |     0.067325 |       2.9589 |       2.9591 |            5 |
|   22 | Accept |       3.2818 |     0.062028 |       2.9589 |        2.959 |           24 |
|   23 | Accept |       3.3226 |     0.081007 |       2.9589 |        2.959 |           32 |
|   24 | Accept |       4.1881 |     0.067852 |       2.9589 |       2.9589 |           43 |
|   25 | Accept |       3.1789 |     0.082533 |       2.9589 |       2.9589 |           18 |
|   26 | Accept |       3.0992 |     0.074624 |       2.9589 |       2.9589 |           14 |
|   27 | Accept |       3.0556 |     0.077162 |       2.9589 |       2.9589 |           22 |
|   28 | Accept |       3.0459 |     0.096438 |       2.9589 |       2.9589 |           12 |
|   29 | Accept |       3.2818 |     0.078972 |       2.9589 |       2.9589 |           26 |
|   30 | Accept |       3.4361 |       0.0649 |       2.9589 |       2.9589 |           34 |
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
3.3
3.25
3.2
3.15
3.1
3.05
3
2.95
Min objective
0
5
10
15
20
25
30
Function evaluations




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 43.5879 seconds
Total objective function evaluation time: 2.2381

Best observed feasible point:
    MinLeafSize
    ___________

        10

Observed objective function value = 2.9589
Estimated objective function value = 2.9589
Function evaluation time = 0.069876

Best estimated feasible point (according to models):
    MinLeafSize
    ___________
```

```
        10

Estimated objective function value = 2.9589
Estimated function evaluation time = 0.073471

Mdl =
  RegressionTree
                           ResponseName: 'Y'
                  CategoricalPredictors: []
                      ResponseTransform: 'none'
                        NumObservations: 94
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]


  Properties, Methods
```

Unbiased Predictor Importance Estimates

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

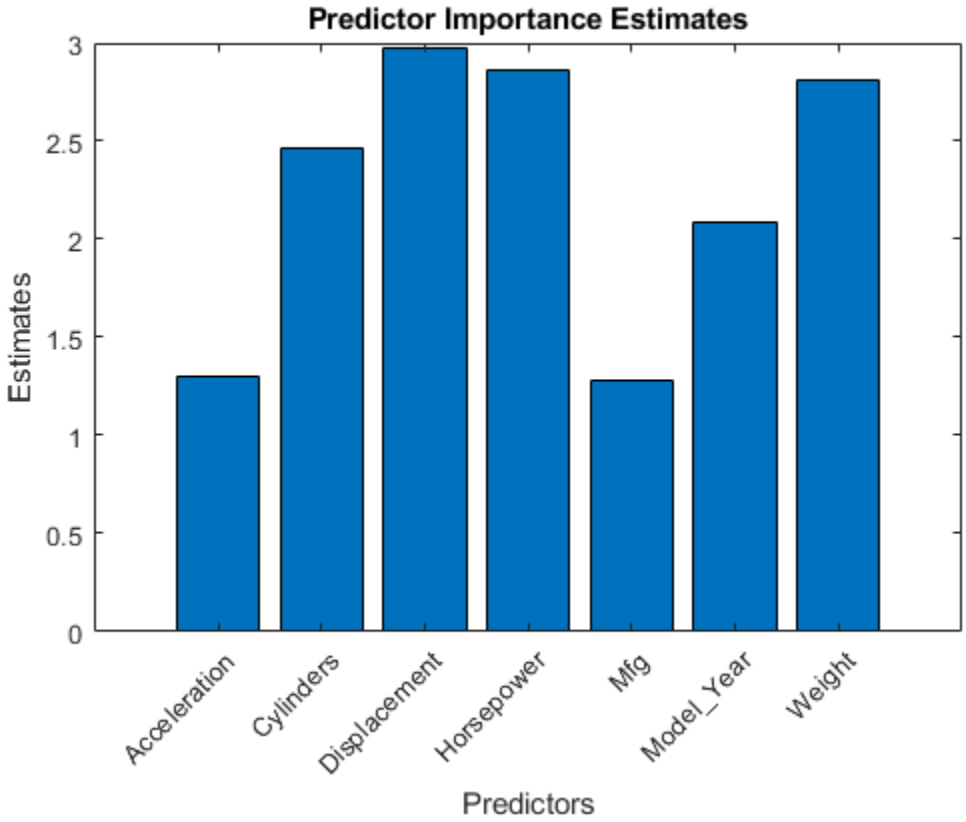
Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a regression tree using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits.

```
Mdl = fitrtree(X,'MPG','PredictorSelection','curvature','Surrogate','on');
```

Estimate predictor importance values by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. Compare the estimates using a bar graph.

```
imp = predictorImportance(Mdl);

figure;
bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Displacement` is the most important predictor, followed by `Horsepower`.

Control Maximum Tree Depth on Tall Array

`fitrtree` grows deep decision trees by default. Build a shallower tree that requires fewer passes through a tall array. Use the `'MaxDepth'` name-value pair argument to control the maximum tree depth.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Convert the in-memory arrays `X` and `MPG` to tall arrays.

```
tx = tall(X);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
ty = tall(MPG);
```

Grow a regression tree using all observations. Allow the tree to grow to the maximum possible depth.

For reproducibility, set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

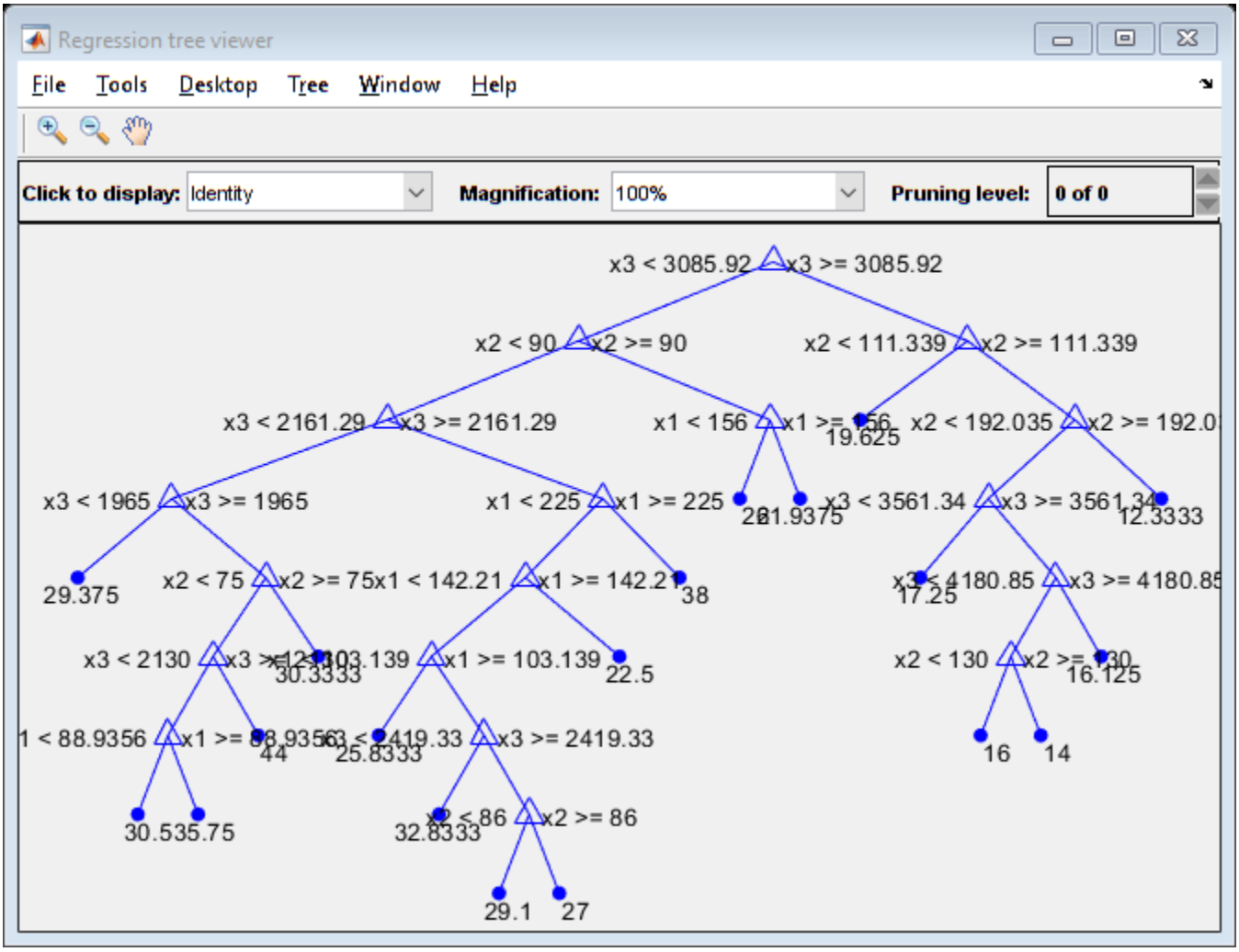
```
rng('default')
tallrng('default')
Mdl = fitrtree(tx,ty);
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 4.1 sec
- Pass 2 of 2: Completed in 0.71 sec
Evaluation completed in 6.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 1.4 sec
- Pass 2 of 7: Completed in 0.29 sec
- Pass 3 of 7: Completed in 1.5 sec
- Pass 4 of 7: Completed in 3.3 sec
- Pass 5 of 7: Completed in 0.63 sec
- Pass 6 of 7: Completed in 1.2 sec
- Pass 7 of 7: Completed in 2.6 sec
Evaluation completed in 12 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.36 sec
- Pass 2 of 7: Completed in 0.27 sec
- Pass 3 of 7: Completed in 0.85 sec
- Pass 4 of 7: Completed in 2 sec
- Pass 5 of 7: Completed in 0.55 sec
- Pass 6 of 7: Completed in 0.92 sec
- Pass 7 of 7: Completed in 1.6 sec
Evaluation completed in 7.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.32 sec
- Pass 2 of 7: Completed in 0.29 sec
- Pass 3 of 7: Completed in 0.89 sec
- Pass 4 of 7: Completed in 1.9 sec
- Pass 5 of 7: Completed in 0.83 sec
- Pass 6 of 7: Completed in 1.2 sec
- Pass 7 of 7: Completed in 2.4 sec
Evaluation completed in 9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.33 sec
```

```
- Pass 2 of 7: Completed in 0.28 sec
- Pass 3 of 7: Completed in 0.89 sec
- Pass 4 of 7: Completed in 2.4 sec
- Pass 5 of 7: Completed in 0.76 sec
- Pass 6 of 7: Completed in 1 sec
- Pass 7 of 7: Completed in 1.7 sec
Evaluation completed in 8.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.34 sec
- Pass 2 of 7: Completed in 0.26 sec
- Pass 3 of 7: Completed in 0.81 sec
- Pass 4 of 7: Completed in 1.7 sec
- Pass 5 of 7: Completed in 0.56 sec
- Pass 6 of 7: Completed in 1 sec
- Pass 7 of 7: Completed in 1.9 sec
Evaluation completed in 7.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.35 sec
- Pass 2 of 7: Completed in 0.28 sec
- Pass 3 of 7: Completed in 0.81 sec
- Pass 4 of 7: Completed in 1.8 sec
- Pass 5 of 7: Completed in 0.76 sec
- Pass 6 of 7: Completed in 0.96 sec
- Pass 7 of 7: Completed in 2.2 sec
Evaluation completed in 8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.35 sec
- Pass 2 of 7: Completed in 0.32 sec
- Pass 3 of 7: Completed in 0.92 sec
- Pass 4 of 7: Completed in 1.9 sec
- Pass 5 of 7: Completed in 1 sec
- Pass 6 of 7: Completed in 1.5 sec
- Pass 7 of 7: Completed in 2.1 sec
Evaluation completed in 9.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.33 sec
- Pass 2 of 7: Completed in 0.28 sec
- Pass 3 of 7: Completed in 0.82 sec
- Pass 4 of 7: Completed in 1.4 sec
- Pass 5 of 7: Completed in 0.61 sec
- Pass 6 of 7: Completed in 0.93 sec
- Pass 7 of 7: Completed in 1.5 sec
Evaluation completed in 6.6 sec
```

View the trained tree `Mdl`.

```
view(Mdl,'Mode','graph')
```



`Mdl` is a tree of depth 8.

Estimate the in-sample mean squared error.

```
MSE_Mdl = gather(loss(Mdl,tx,ty))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.6 sec
Evaluation completed in 1.9 sec

MSE_Mdl = 4.9078
```

Grow a regression tree using all observations. Limit the tree depth by specifying a maximum tree depth of 4.

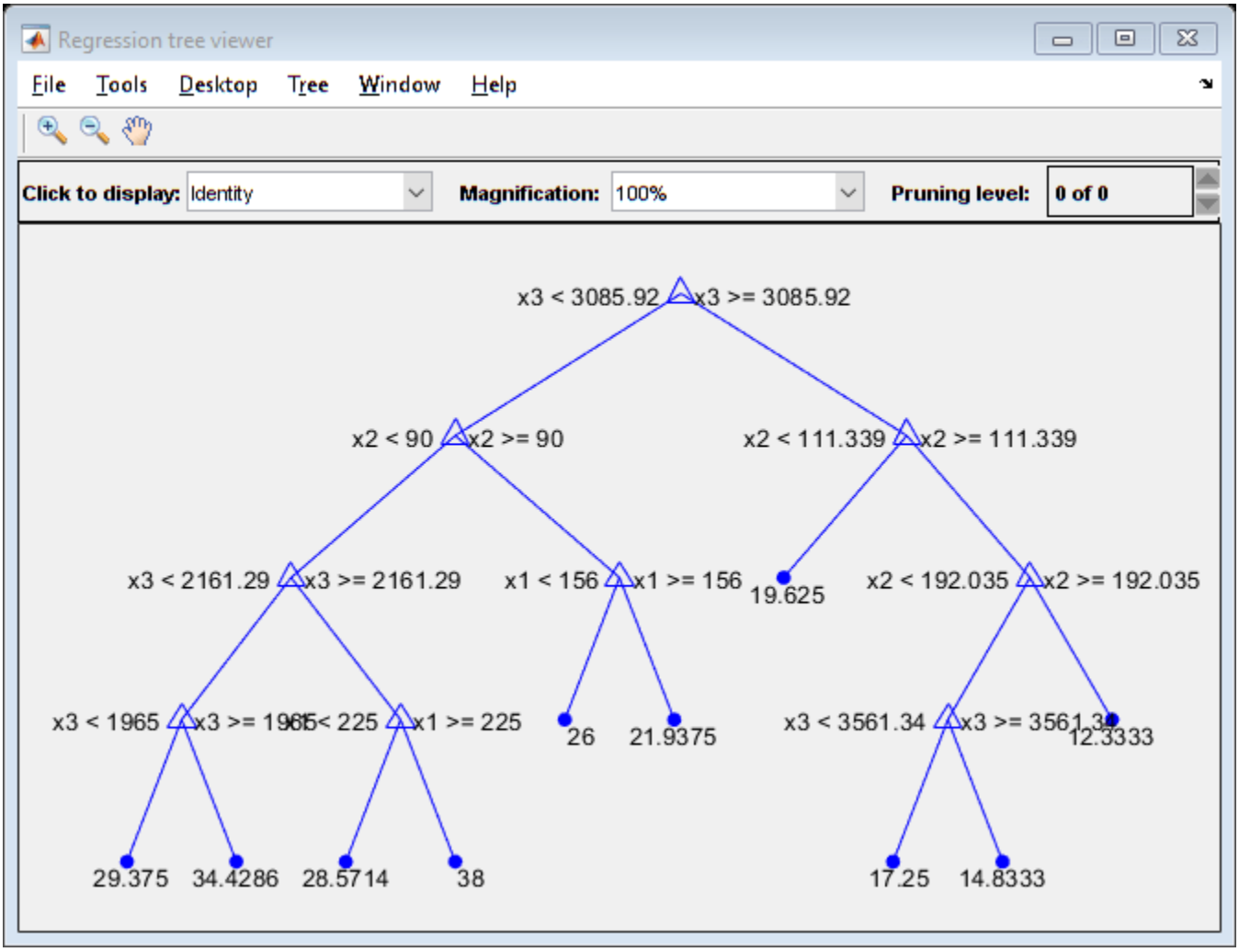
```
Mdl2 = fitrtree(tx,ty,'MaxDepth',4);
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 0.27 sec
- Pass 2 of 2: Completed in 0.28 sec
Evaluation completed in 0.84 sec
Evaluating tall expression using the Parallel Pool 'local':
```

```
- Pass 1 of 7: Completed in 0.36 sec
- Pass 2 of 7: Completed in 0.3 sec
- Pass 3 of 7: Completed in 0.95 sec
- Pass 4 of 7: Completed in 1.6 sec
- Pass 5 of 7: Completed in 0.55 sec
- Pass 6 of 7: Completed in 0.93 sec
- Pass 7 of 7: Completed in 1.5 sec
Evaluation completed in 7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.34 sec
- Pass 2 of 7: Completed in 0.3 sec
- Pass 3 of 7: Completed in 0.95 sec
- Pass 4 of 7: Completed in 1.7 sec
- Pass 5 of 7: Completed in 0.57 sec
- Pass 6 of 7: Completed in 0.94 sec
- Pass 7 of 7: Completed in 1.8 sec
Evaluation completed in 7.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.34 sec
- Pass 2 of 7: Completed in 0.3 sec
- Pass 3 of 7: Completed in 0.87 sec
- Pass 4 of 7: Completed in 1.5 sec
- Pass 5 of 7: Completed in 0.57 sec
- Pass 6 of 7: Completed in 0.81 sec
- Pass 7 of 7: Completed in 1.7 sec
Evaluation completed in 6.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.32 sec
- Pass 2 of 7: Completed in 0.27 sec
- Pass 3 of 7: Completed in 0.85 sec
- Pass 4 of 7: Completed in 1.6 sec
- Pass 5 of 7: Completed in 0.63 sec
- Pass 6 of 7: Completed in 0.9 sec
- Pass 7 of 7: Completed in 1.6 sec
Evaluation completed in 7 sec
```

View the trained tree `Mdl2`.

```
view(Mdl2,'Mode','graph')
```



Estimate the in-sample mean squared error.

```
MSE_Mdl2 = gather(loss(Mdl2,tx,ty))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.73 sec
Evaluation completed in 1 sec

MSE_Mdl2 = 9.3903
```

`Mdl2` is a less complex tree with a depth of 4 and an in-sample mean squared error that is higher than the mean squared error of `Mdl`.

Optimize Regression Tree on Tall Array

Optimize hyperparameters of a regression tree automatically using a tall array. The sample data set is the `carsmall` data set. This example converts the data set to a tall array and uses it to run the optimization procedure.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. If you want to run the example using the local MATLAB session when you have Parallel Computing Toolbox, you can change the global execution environment by using the `mapreducer` function.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Convert the in-memory arrays `X` and `MPG` to tall arrays.

```
tx = tall(X);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
ty = tall(MPG);
```

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value pair argument. Find the optimal `'MinLeafSize'` value that minimizes holdout cross-validation loss. (Specifying `'auto'` uses `'MinLeafSize'`.) For reproducibility, use the `'expected-improvement-plus'` acquisition function and set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrtree(tx,ty,...
    'OptimizeHyperparameters','auto',...
    'HyperparameterOptimizationOptions',struct('Holdout',0.3,...
    'AcquisitionFunctionName','expected-improvement-plus'))
```

```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 4.4 sec
Evaluation completed in 6.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.97 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 3.6 sec
- Pass 4 of 4: Completed in 2.4 sec
Evaluation completed in 9.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 2.7 sec
- Pass 4 of 4: Completed in 1.9 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.52 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 3 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 8.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.55 sec
- Pass 2 of 4: Completed in 1.4 sec
```

```
- Pass 3 of 4: Completed in 2.6 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 7.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.61 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2.1 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 6.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2.4 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 6.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 1.4 sec
Evaluation completed in 1.7 sec
|=====================================================================================|
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |  MinLeafSize |
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================|
|    1 | Best   |      3.2007 |      69.013 |      3.2007 |      3.2007 |            2 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.52 sec
Evaluation completed in 0.83 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.65 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 3 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 8.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.79 sec
Evaluation completed in 1 sec
|    2 | Error  |         NaN |      13.772 |         NaN |      3.2007 |           46 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.52 sec
Evaluation completed in 0.81 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.57 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 2.2 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 6.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2.7 sec
- Pass 4 of 4: Completed in 1.7 sec
Evaluation completed in 6.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2.1 sec
- Pass 4 of 4: Completed in 1.9 sec
```

```
Evaluation completed in 6.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.72 sec
Evaluation completed in 0.99 sec
|    3 | Best   |      3.1876 |      29.091 |      3.1876 |      3.1884 |           18 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.48 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.54 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.64 sec
Evaluation completed in 0.92 sec
|    4 | Best   |      2.9048 |      33.465 |      2.9048 |      2.9537 |            6 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.71 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
```

```
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.66 sec
Evaluation completed in 0.92 sec
|    5 | Accept |      3.2895 |      25.902 |      2.9048 |      2.9048 |           15 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.54 sec
Evaluation completed in 0.82 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2.1 sec
- Pass 4 of 4: Completed in 1.9 sec
Evaluation completed in 6.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 2 sec
Evaluation completed in 6.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.68 sec
Evaluation completed in 0.99 sec
|    6 | Accept |      3.1641 |      35.522 |      2.9048 |      3.1493 |            5 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.51 sec
Evaluation completed in 0.79 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.67 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 6.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
```

```
Evaluation completed in 5.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.63 sec
Evaluation completed in 0.89 sec
|    7 | Accept |      2.9048 |      33.755 |      2.9048 |      2.9048 |            6 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.45 sec
Evaluation completed in 0.75 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2.2 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 6.1 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.68 sec
Evaluation completed in 0.97 sec
|    8 | Accept |      2.9522 |      33.362 |      2.9048 |      2.9048 |            7 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.42 sec
Evaluation completed in 0.71 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
```

```
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.49 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.64 sec
Evaluation completed in 0.9 sec
|    9 | Accept |      2.9985 |      32.674 |      2.9048 |      2.9048 |            8 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.43 sec
Evaluation completed in 0.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.56 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 5.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.88 sec
Evaluation completed in 1.2 sec
|   10 | Accept |      3.0185 |      33.922 |      2.9048 |      2.9048 |           10 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.74 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
```

```
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 6.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.73 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 6.2 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.63 sec
Evaluation completed in 0.88 sec
|   11 | Accept |      3.2895 |      26.625 |      2.9048 |      2.9048 |           14 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.48 sec
Evaluation completed in 0.78 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.51 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.65 sec
Evaluation completed in 0.9 sec
|   12 | Accept |      3.4798 |      18.111 |      2.9048 |      2.9049 |           31 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.71 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
```

```
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.64 sec
Evaluation completed in 0.91 sec
|   13 | Accept |      3.2248 |      47.436 |      2.9048 |      2.9048 |            1 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.46 sec
Evaluation completed in 0.74 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.6 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.57 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2.6 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 6.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.62 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.6 sec
Evaluation completed in 6.1 sec
Evaluating tall expression using the Parallel Pool 'local':
```

```
- Pass 1 of 1: Completed in 0.61 sec
Evaluation completed in 0.88 sec
|   14 | Accept |      3.1498 |      42.062 |      2.9048 |      2.9048 |            3 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.46 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.67 sec
- Pass 2 of 4: Completed in 1.3 sec
- Pass 3 of 4: Completed in 2.3 sec
- Pass 4 of 4: Completed in 2.2 sec
Evaluation completed in 7.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.6 sec
Evaluation completed in 0.86 sec
|   15 | Accept |      2.9048 |        34.3 |      2.9048 |      2.9048 |            6 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.48 sec
Evaluation completed in 0.78 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
```

```
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 2 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.88 sec
|   16 | Accept |      2.9048 |       32.97 |      2.9048 |      2.9048 |            6 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.43 sec
Evaluation completed in 0.73 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.9 sec
|   17 | Accept |      3.1847 |       17.47 |      2.9048 |      2.9048 |           23 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.43 sec
Evaluation completed in 0.72 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.68 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 6.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
```

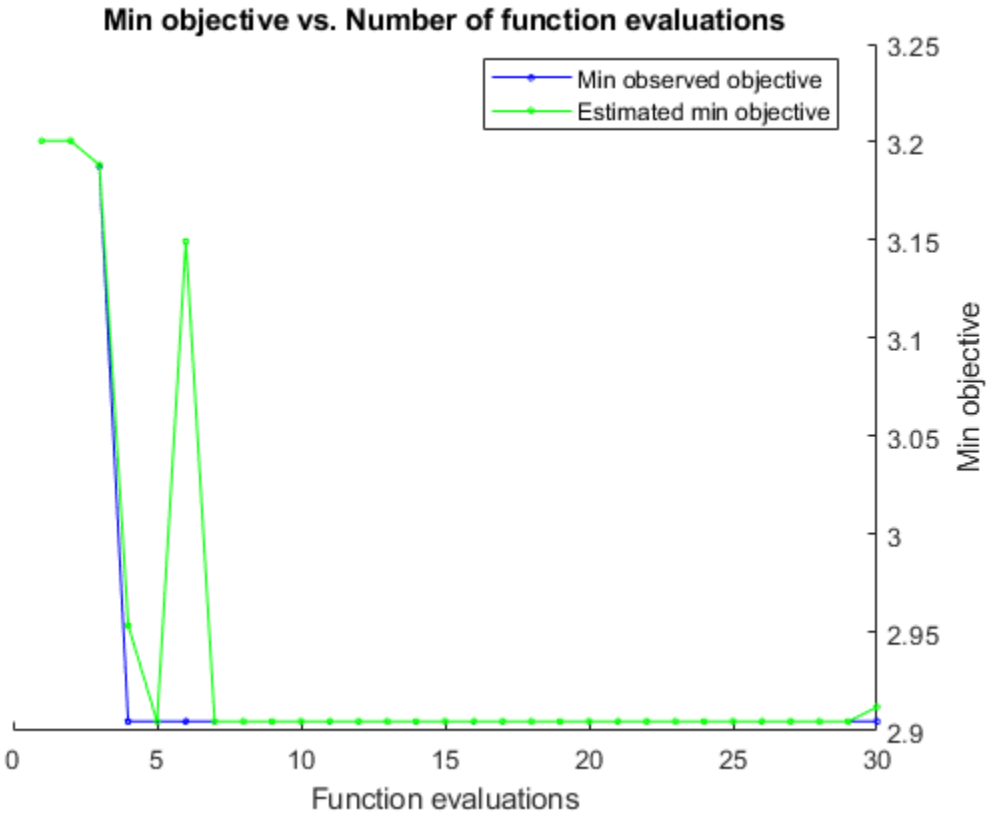
```
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.93 sec
|   18 | Accept |      3.1817 |      33.346 |      2.9048 |      2.9048 |            4 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.43 sec
Evaluation completed in 0.72 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.86 sec
|   19 | Error  |         NaN |      10.235 |      2.9048 |      2.9048 |           38 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.47 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.63 sec
Evaluation completed in 0.89 sec
|   20 | Accept |      3.0628 |      32.459 |      2.9048 |      2.9048 |           12 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.46 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.48 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
```

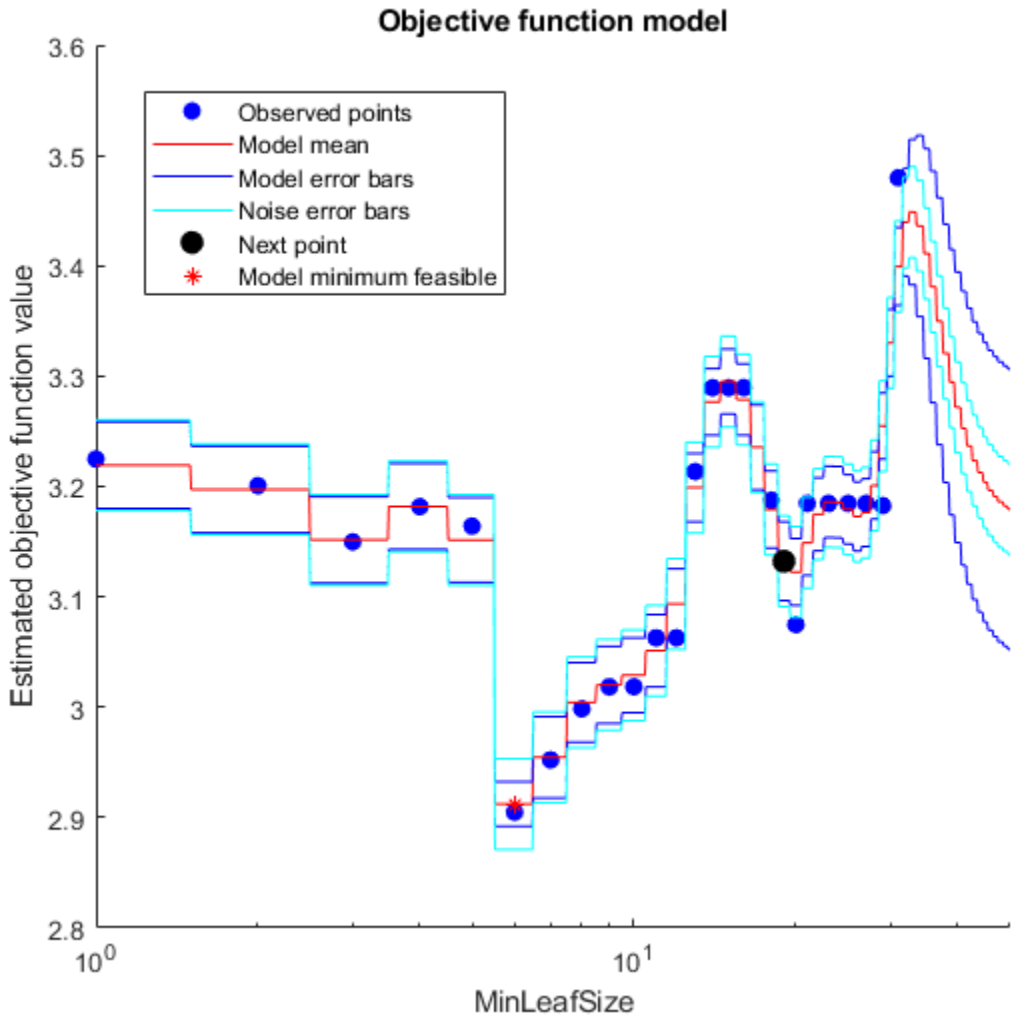
```
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.68 sec
- Pass 2 of 4: Completed in 1.7 sec
- Pass 3 of 4: Completed in 2.1 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 6.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.64 sec
Evaluation completed in 0.9 sec
|=====================================================================================|
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |  MinLeafSize |
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================|
|   21 | Accept |      3.1847 |       19.02 |      2.9048 |      2.9048 |           27 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.45 sec
Evaluation completed in 0.75 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.47 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.5 sec
- Pass 2 of 4: Completed in 1.6 sec
- Pass 3 of 4: Completed in 2.4 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 6.8 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.5 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.63 sec
Evaluation completed in 0.89 sec
|   22 | Accept |      3.0185 |      33.933 |      2.9048 |      2.9048 |            9 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.46 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
```

```
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.64 sec
Evaluation completed in 0.89 sec
|   23 | Accept |      3.0749 |      25.147 |      2.9048 |      2.9048 |           20 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.73 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.42 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.53 sec
- Pass 2 of 4: Completed in 1.4 sec
- Pass 3 of 4: Completed in 1.9 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.9 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.88 sec
|   24 | Accept |      3.0628 |      32.764 |      2.9048 |      2.9048 |           11 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.73 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.2 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.61 sec
```

```
Evaluation completed in 0.87 sec
|   25 | Error  |         NaN |      10.294 |      2.9048 |      2.9048 |           34 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.73 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.87 sec
|   26 | Accept |      3.1847 |      17.587 |      2.9048 |      2.9048 |           25 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.45 sec
Evaluation completed in 0.73 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.3 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.66 sec
Evaluation completed in 0.96 sec
|   27 | Accept |      3.2895 |      24.867 |      2.9048 |      2.9048 |           16 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.44 sec
Evaluation completed in 0.74 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
```

```
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.4 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.6 sec
Evaluation completed in 0.88 sec
|   28 | Accept |      3.2135 |      24.928 |      2.9048 |      2.9048 |           13 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.47 sec
Evaluation completed in 0.76 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.45 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.46 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.62 sec
Evaluation completed in 0.87 sec
|   29 | Accept |      3.1847 |      17.582 |      2.9048 |      2.9048 |           21 |

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.53 sec
Evaluation completed in 0.81 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.44 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 4: Completed in 0.43 sec
- Pass 2 of 4: Completed in 1.1 sec
- Pass 3 of 4: Completed in 1.8 sec
- Pass 4 of 4: Completed in 1.3 sec
Evaluation completed in 5.4 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 1: Completed in 0.63 sec
Evaluation completed in 0.88 sec
|   30 | Accept |      3.1827 |      17.597 |      2.9048 |      2.9122 |           29 |
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
3.25
3.2
3.15
3.1
3.05
3
2.95
2.9
Min objective
0
5
10
15
20
25
30
Function evaluations




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 882.5668 seconds.
Total objective function evaluation time: 859.2122

Best observed feasible point:
    MinLeafSize
    ___________

         6

Observed objective function value = 2.9048
Estimated objective function value = 2.9122
Function evaluation time = 33.4655

Best estimated feasible point (according to models):
    MinLeafSize
```

```
    ____________

          6

Estimated objective function value = 2.9122
Estimated function evaluation time = 33.6594

Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 2: Completed in 0.26 sec
- Pass 2 of 2: Completed in 0.26 sec
Evaluation completed in 0.84 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.31 sec
- Pass 2 of 7: Completed in 0.25 sec
- Pass 3 of 7: Completed in 0.75 sec
- Pass 4 of 7: Completed in 1.2 sec
- Pass 5 of 7: Completed in 0.45 sec
- Pass 6 of 7: Completed in 0.69 sec
- Pass 7 of 7: Completed in 1.2 sec
Evaluation completed in 5.7 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.28 sec
- Pass 2 of 7: Completed in 0.24 sec
- Pass 3 of 7: Completed in 0.75 sec
- Pass 4 of 7: Completed in 1.2 sec
- Pass 5 of 7: Completed in 0.46 sec
- Pass 6 of 7: Completed in 0.67 sec
- Pass 7 of 7: Completed in 1.2 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.32 sec
- Pass 2 of 7: Completed in 0.25 sec
- Pass 3 of 7: Completed in 0.71 sec
- Pass 4 of 7: Completed in 1.2 sec
- Pass 5 of 7: Completed in 0.47 sec
- Pass 6 of 7: Completed in 0.66 sec
- Pass 7 of 7: Completed in 1.2 sec
Evaluation completed in 5.6 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.29 sec
- Pass 2 of 7: Completed in 0.25 sec
- Pass 3 of 7: Completed in 0.73 sec
- Pass 4 of 7: Completed in 1.2 sec
- Pass 5 of 7: Completed in 0.46 sec
- Pass 6 of 7: Completed in 0.68 sec
- Pass 7 of 7: Completed in 1.2 sec
Evaluation completed in 5.5 sec
Evaluating tall expression using the Parallel Pool 'local':
- Pass 1 of 7: Completed in 0.27 sec
- Pass 2 of 7: Completed in 0.25 sec
- Pass 3 of 7: Completed in 0.75 sec
- Pass 4 of 7: Completed in 1.2 sec
- Pass 5 of 7: Completed in 0.47 sec
- Pass 6 of 7: Completed in 0.69 sec
- Pass 7 of 7: Completed in 1.2 sec
Evaluation completed in 5.6 sec
```

```
Mdl =
  CompactRegressionTree
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'


  Properties, Methods


FitInfo = struct with no fields.



HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                      ObjectiveFcn: @createObjFcn/tallObjFcn
              VariableDescriptions: [3×1 optimizableVariable]
                           Options: [1×1 struct]
                      MinObjective: 2.9048
                   XAtMinObjective: [1×1 table]
             MinEstimatedObjective: 2.9122
          XAtMinEstimatedObjective: [1×1 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 882.5668
                         NextPoint: [1×1 table]
                            XTrace: [30×1 table]
                    ObjectiveTrace: [30×1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30×1 cell}
      ObjectiveEvaluationTimeTrace: [30×1 double]
                IterationTimeTrace: [30×1 double]
                        ErrorTrace: [30×1 double]
                  FeasibilityTrace: [30×1 logical]
       FeasibilityProbabilityTrace: [30×1 double]
              IndexOfMinimumTrace: [30×1 double]
            ObjectiveMinimumTrace: [30×1 double]
    EstimatedObjectiveMinimumTrace: [30×1 double]
```

Input Arguments

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.

- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl` stores the response variable `Y` as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Response data
numeric column vector

Response data, specified as a numeric column vector with the same number of rows as `X`. Each entry in `Y` is the response to the data in the corresponding row of `X`.

The software considers `NaN` values in `Y` to be missing values. `fitrtree` does not use observations with missing values for `Y` in the fit.

Data Types: `single` | `double`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each column of `X` represents one variable, and each row represents one observation.

`fitrtree` considers `NaN` values in `X` as missing values. `fitrtree` does not use observations with all missing values for `X` in the fit. `fitrtree` uses observations with some missing values for `X` to find splits on variables for which these observations have valid values.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `'CrossVal','on','MinParentSize',30` specifies a cross-validated regression tree with a minimum of 30 observations per branch node.

Model Parameters

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrtree` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrtree` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrtree` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`MaxDepth` — Maximum tree depth
positive integer

Maximum tree depth, specified as the comma-separated pair consisting of `'MaxDepth'` and a positive integer. Specify a value for this argument to return a tree that has fewer levels and requires fewer passes through the tall array to compute. Generally, the algorithm of `fitrtree` takes one pass through the data and an additional pass for each tree level. The function does not set a maximum tree depth, by default.

Note This option applies only when you use `fitrtree` on tall arrays. See Tall Arrays on page 33-2416 for more information.

`MergeLeaves` — Leaf merge flag
`'on'` (default) | `'off'`

Leaf merge flag, specified as the comma-separated pair consisting of `'MergeLeaves'` and `'on'` or `'off'`.

If `MergeLeaves` is `'on'`, then `fitrtree`:

- Merges leaves that originate from the same parent node and yield a sum of risk values greater than or equal to the risk associated with the parent node
- Estimates the optimal sequence of pruned subtrees, but does not prune the regression tree

Otherwise, `fitrtree` does not merge leaves.

Example: `'MergeLeaves','off'`

`MinParentSize` — Minimum number of branch node observations
`10` (default) | positive integer value

Minimum number of branch node observations, specified as the comma-separated pair consisting of `'MinParentSize'` and a positive integer value. Each branch node in the tree has at least `MinParentSize` observations. If you supply both `MinParentSize` and `MinLeafSize`, `fitrtree` uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

Example: `'MinParentSize',8`

Data Types: `single` | `double`

`NumBins` — Number of bins for numeric predictors
`[]`(empty) (default) | positive integer scalar

Number of bins for numeric predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

- If the `'NumBins'` value is empty (default), then `fitrtree` does not bin any predictors.
- If you specify the `'NumBins'` value as a positive integer scalar (`numBins`), then `fitrtree` bins every numeric predictor into at most `numBins` equiprobable bins, and then grows trees on the bin indices instead of the original data.

- The number of bins can be less than `numBins` if a predictor has fewer than `numBins` unique values.
- `fitrtree` does not bin categorical predictors.

When you use a large training data set, this binning option speeds up training but might cause a potential decrease in accuracy. You can try `'NumBins',50` first, and then change the value depending on the accuracy and training speed.

A trained model stores the bin edges in the `BinEdges` property.

Example: `'NumBins',50`

Data Types: `single` | `double`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrtree` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`PredictorSelection` — Algorithm used to select the best split predictor
`'allsplits'` (default) | `'curvature'` | `'interaction-curvature'`

Algorithm used to select the best split predictor at each node, specified as the comma-separated pair consisting of `'PredictorSelection'` and a value in this table.

Value	Description
`'allsplits'`	Standard CART — Selects the split predictor that maximizes the split-criterion gain over all possible splits of all predictors [1].

Value	Description
`'curvature'`	Curvature test on page 33-2411 — Selects the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response [2]. Training speed is similar to standard CART.
`'interaction-curvature'`	Interaction test on page 33-2412 — Chooses the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response (that is, conducts curvature tests), and that minimizes the *p*-value of a chi-square test of independence between each pair of predictors and response [2]. Training speed can be slower than standard CART.

For `'curvature'` and `'interaction-curvature'`, if all tests yield *p*-values greater than 0.05, then `fitrtree` stops splitting nodes.

Tip

- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [3]. Consider specifying the curvature or interaction test if any of the following are true:
  - If there are predictors that have relatively fewer distinct values than other predictors, for example, if the predictor data set is heterogeneous.
  - If an analysis of predictor importance is your goal. For more on predictor importance estimation, see `predictorImportance` and “Introduction to Feature Selection” on page 15-49.
- Trees grown using standard CART are not sensitive to predictor variable interactions. Also, such trees are less likely to identify important variables in the presence of many irrelevant predictors than the application of the interaction test. Therefore, to account for predictor interactions and identify importance variables in the presence of many irrelevant variables, specify the interaction test .
- Prediction speed is unaffected by the value of `'PredictorSelection'`.

For details on how `fitrtree` selects split predictors, see “Node Splitting Rules” on page 33-2413 and “Choose Split Predictor Selection Technique” on page 19-14.

Example: `'PredictorSelection','curvature'`

`Prune` — Flag to estimate optimal sequence of pruned subtrees
`'on'` (default) | `'off'`

Flag to estimate the optimal sequence of pruned subtrees, specified as the comma-separated pair consisting of `'Prune'` and `'on'` or `'off'`.

If `Prune` is `'on'`, then `fitrtree` grows the regression tree and estimates the optimal sequence of pruned subtrees, but does not prune the regression tree. Otherwise, `fitrtree` grows the regression tree without estimating the optimal sequence of pruned subtrees.

To prune a trained regression tree, pass the regression tree to `prune`.

Example: `'Prune','off'`

`PruneCriterion` — Pruning criterion
`'mse'` (default)

Pruning criterion, specified as the comma-separated pair consisting of `'PruneCriterion'` and `'mse'`.

`QuadraticErrorTolerance` — Quadratic error tolerance
`1e-6` (default) | positive scalar value

Quadratic error tolerance per node, specified as the comma-separated pair consisting of `'QuadraticErrorTolerance'` and a positive scalar value. The function stops splitting nodes when the weighted mean squared error per node drops below `QuadraticErrorTolerance*ε`, where `ε` is the weighted mean squared error of all n responses computed before growing the decision tree.

$$\varepsilon = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2 .$$

w_i is the weight of observation i, given that the weights of all the observations sum to one ($\sum_{i=1}^{n} w_i = 1$), and

$$\bar{y} = \sum_{i=1}^{n} w_i y_i$$

is the weighted average of all the responses.

For more details on node splitting, see Node Splitting Rules on page 33-2413.

Example: `'QuadraticErrorTolerance',1e-4`

`Reproducible` — Flag to enforce reproducibility
`false` (logical `0`) (default) | `true` (logical `1`)

Flag to enforce reproducibility over repeated runs of training a model, specified as the comma-separated pair consisting of `'Reproducible'` and either `false` or `true`.

If `'NumVariablesToSample'` is not `'all'`, then the software selects predictors at random for each split. To reproduce the random selections, you must specify `'Reproducible',true` and set the seed of the random number generator by using `rng`. Note that setting `'Reproducible'` to `true` can slow down training.

Example: `'Reproducible',true`

Data Types: `logical`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

`SplitCriterion` — Split criterion
`'MSE'` (default)

Split criterion, specified as the comma-separated pair consisting of `'SplitCriterion'` and `'MSE'`, meaning mean squared error.

Example: `'SplitCriterion','MSE'`

`Surrogate` — Surrogate decision splits flag
`'off'` (default) | `'on'` | `'all'` | positive integer

Surrogate decision splits flag, specified as the comma-separated pair consisting of `'Surrogate'` and `'on'`, `'off'`, `'all'`, or a positive integer.

- When `'on'`, `fitrtree` finds at most 10 surrogate splits at each branch node.
- When set to a positive integer, `fitrtree` finds at most the specified number of surrogate splits at each branch node.
- When set to `'all'`, `fitrtree` finds all surrogate splits at each branch node. The `'all'` setting can use much time and memory.

Use surrogate splits to improve the accuracy of predictions for data with missing values. The setting also enables you to compute measures of predictive association between predictors.

Example: `'Surrogate','on'`

Data Types: `single` | `double` | `char` | `string`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors when training the model.

`fitrtree` normalizes the values of `Weights` to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation

`CrossVal` — Cross-validation flag
`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'CrossVal'` and either `'on'` or `'off'`.

If `'on'`, `fitrtree` grows a cross-validated decision tree with 10 folds. You can override this cross-validation setting using one of the `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'` name-value pair arguments. You can only use one of these four options (`'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`) at a time when creating a cross-validated tree.

Alternatively, cross-validate `tree` later using the `crossval` method.

Example: `'CrossVal','on'`

`CVPartition` — Partition for cross-validation tree
`cvpartition` object

Partition for cross-validated tree, specified as the comma-separated pair consisting of `'CVPartition'` and an object created using `cvpartition`.

If you use `'CVPartition'`, you cannot use any of the `'KFold'`, `'Holdout'`, or `'Leaveout'` name-value pair arguments.

`Holdout` — Fraction of data for holdout validation
`0` (default) | scalar value in the range `[0,1]`

Fraction of data used for holdout validation, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range `[0,1]`. Holdout validation tests the specified fraction of the data, and uses the rest of the data for training.

If you use `'Holdout'`, you cannot use any of the `'CVPartition'`, `'KFold'`, or `'Leaveout'` name-value pair arguments.

Example: `'Holdout',0.1`

Data Types: `single` | `double`

`KFold` — Number of folds
`10` (default) | positive integer greater than 1

Number of folds to use in a cross-validated tree, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1.

If you use `'KFold'`, you cannot use any of the `'CVPartition'`, `'Holdout'`, or `'Leaveout'` name-value pair arguments.

Example: `'KFold',8`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and either `'on'` or `'off'`. Use leave-one-out cross-validation by setting to `'on'`.

If you use `'Leaveout'`, you cannot use any of the `'CVPartition'`, `'Holdout'`, or `'KFold'` name-value pair arguments.

Example: `'Leaveout','on'`

Hyperparameters

`MaxNumSplits` — Maximal number of decision splits
`size(X,1) - 1` (default) | positive integer

Maximal number of decision splits (or branch nodes), specified as the comma-separated pair consisting of `'MaxNumSplits'` and a positive integer. `fitrtree` splits `MaxNumSplits` or fewer branch nodes. For more details on splitting behavior, see “Tree Depth Control” on page 33-2415.

Example: `'MaxNumSplits',5`

Data Types: `single` | `double`

`MinLeafSize` — Minimum number of leaf node observations
`1` (default) | positive integer value

Minimum number of leaf node observations, specified as the comma-separated pair consisting of `'MinLeafSize'` and a positive integer value. Each leaf has at least `MinLeafSize` observations per tree leaf. If you supply both `MinParentSize` and `MinLeafSize`, `fitrtree` uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

Example: `'MinLeafSize',3`

Data Types: `single` | `double`

`NumVariablesToSample` — Number of predictors to select at random for each split
`'all'` (default) | positive integer value

Number of predictors to select at random for each split, specified as the comma-separated pair consisting of `'NumVariablesToSample'` and a positive integer value. Alternatively, you can specify `'all'` to use all available predictors.

If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` as `'all'`. Otherwise, the software might not select some predictors, underestimating their importance.

To reproduce the random selections, you must set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

Example: `'NumVariablesToSample',3`

Data Types: `char` | `string` | `single` | `double`

Hyperparameter Optimization

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of the following:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'MinLeafSize'}`.
- `'all'` — Optimize all eligible parameters.
- String array or cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitrtree` by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitrtree` are:

- `MaxNumSplits` — `fitrtree` searches among integers, by default log-scaled in the range `[1,max(2,NumObservations-1)]`.
- `MinLeafSize` — `fitrtree` searches among integers, by default log-scaled in the range `[1,max(2,floor(NumObservations/2))]`.
- `NumVariablesToSample` — `fitrtree` does not optimize over this hyperparameter. If you pass `NumVariablesToSample` as a parameter name, `fitrtree` simply uses the full number of predictors. However, `fitrensemble` does optimize over this hyperparameter.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example,

```
load carsmall
params = hyperparameters('fitrtree',[Horsepower,Weight],MPG);
params(1).Range = [1,30];
```

Pass `params` as the value of `OptimizeHyperparameters`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Regression Tree” on page 33-2369.

Example: `'auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`tree` — Regression tree

regression tree object

Regression tree, returned as a regression tree object. Using the `'Crossval'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'` options results in a tree of class `RegressionPartitionedModel`. You cannot use a partitioned tree for prediction, so this kind of tree does not have a `predict` method.

Otherwise, `tree` is of class `RegressionTree`, and you can use the `predict` method to make predictions.

More About

Curvature Test

The curvature test is a statistical test assessing the null hypothesis that two variables are unassociated.

The curvature test between predictor variable x and y is conducted using this process.

1. If x is continuous, then partition it into its quartiles. Create a nominal variable that bins observations according to which section of the partition they occupy. If there are missing values, then create an extra bin for them.
2. For each level in the partitioned predictor $j = 1...J$ and class in the response $k = 1,...,K$, compute the weighted proportion of observations in class k

   $$\widehat{\pi}_{jk} = \sum_{i=1}^{n} I\{y_i = k\} w_i .$$

   w_i is the weight of observation i, $\sum w_i = 1$, I is the indicator function, and n is the sample size. If all observations have the same weight, then $\widehat{\pi}_{jk} = \frac{n_{jk}}{n}$, where n_{jk} is the number of observations in level j of the predictor that are in class k.
3. Compute the test statistic

   $$t = n \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(\widehat{\pi}_{jk} - \widehat{\pi}_{j+}\widehat{\pi}_{+k}\right)^2}{\widehat{\pi}_{j+}\widehat{\pi}_{+k}}$$

   $\widehat{\pi}_{j+} = \sum_{k} \widehat{\pi}_{jk}$, that is, the marginal probability of observing the predictor at level j. $\widehat{\pi}_{+k} = \sum_{j} \widehat{\pi}_{jk}$, that is the marginal probability of observing class k. If n is large enough, then t is distributed as a χ^2 with $(K - 1)(J - 1)$ degrees of freedom.
4. If the p-value for the test is less than 0.05, then reject the null hypothesis that there is no association between x and y.

When determining the best split predictor at each node, the standard CART algorithm prefers to select continuous predictors that have many levels. Sometimes, such a selection can be spurious and can also mask more important predictors that have fewer levels, such as categorical predictors.

The curvature test can be applied instead of standard CART to determine the best split predictor at each node. In that case, the best split predictor variable is the one that minimizes the significant *p*-values (those less than 0.05) of curvature tests between each predictor and the response variable. Such a selection is robust to the number of levels in individual predictors.

For more details on how the curvature test applies to growing regression trees, see "Node Splitting Rules" on page 33-2413 and [3].

Interaction Test

The interaction test is a statistical test that assesses the null hypothesis that there is no interaction between a pair of predictor variables and the response variable.

The interaction test assessing the association between predictor variables x_1 and x_2 with respect to y is conducted using this process.

1. If x_1 or x_2 is continuous, then partition that variable into its quartiles. Create a nominal variable that bins observations according to which section of the partition they occupy. If there are missing values, then create an extra bin for them.
2. Create the nominal variable z with $J = J_1J_2$ levels that assigns an index to observation i according to which levels of x_1 and x_2 it belongs. Remove any levels of z that do not correspond to any observations.
3. Conduct a curvature test on page 33-2411 between z and y.

When growing decision trees, if there are important interactions between pairs of predictors, but there are also many other less important predictors in the data, then standard CART tends to miss the important interactions. However, conducting curvature and interaction tests for predictor selection instead can improve detection of important interactions, which can yield more accurate decision trees.

For more details on how the interaction test applies to growing decision trees, see "Curvature Test" on page 33-2411, "Node Splitting Rules" on page 33-2413 and [2].

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_jL_k} - P_{R_jR_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_jL_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.

- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in (–∞,1]. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Surrogate Decision Splits

A surrogate decision split is an alternative to the optimal decision split at a given node in a decision tree. The optimal split is found by growing the tree; the surrogate split uses a similar or correlated predictor variable and split criterion.

When the value of the optimal split predictor for an observation is missing, the observation is sent to the left or right child node using the best surrogate predictor. When the value of the best surrogate split predictor for the observation is also missing, the observation is sent to the left or right child node using the second-best surrogate predictor, and so on. Candidate splits are sorted in descending order by their predictive measure of association on page 33-2412.

Tip

- By default, `Prune` is `'on'`. However, this specification does not prune the regression tree. To prune a trained regression tree, pass the regression tree to `prune`.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

Node Splitting Rules

`fitrtree` uses these processes to determine how to split node t.

- For standard CART (that is, if `PredictorSelection` is `'allpairs'`) and for all predictors x_i, $i = 1,\ldots,p$:

  1. `fitrtree` computes the weighted mean squared error (MSE) of the responses in node t using

     $$\varepsilon_t = \sum_{j \in T} w_j \left(y_j - \bar{y}_t\right)^2 .$$

     w_j is the weight of observation j, and T is the set of all observation indices in node t. If you do not specify `Weights`, then $w_j = 1/n$, where n is the sample size.

  2. `fitrtree` estimates the probability that an observation is in node t using

     $$P(T) = \sum_{j \in T} w_j .$$

  3. `fitrtree` sorts x_i in ascending order. Each element of the sorted predictor is a splitting candidate or cut point. `fitrtree` records any indices corresponding to missing values in the set T_U, which is the unsplit set.
  4. `fitrtree` determines the best way to split node t using x_i by maximizing the reduction in MSE (ΔI) over all splitting candidates. That is, for all splitting candidates in x_i:

a `fitrtree` splits the observations in node t into left and right child nodes (t_L and t_R, respectively).

b `fitrtree` computes ΔI. Suppose that for a particular splitting candidate, t_L and t_R contain observation indices in the sets T_L and T_R, respectively.

- If x_i does not contain any missing values, then the reduction in MSE for the current splitting candidate is

$$\Delta I = P(T)\varepsilon_t - P(T_L)\varepsilon_{t_L} - P(T_R)\varepsilon_{t_R}.$$

- If x_i contains missing values, then, assuming that the observations are missing at random, the reduction in MSE is

$$\Delta I_U = P(T - T_U)\varepsilon_t - P(T_L)\varepsilon_{t_L} - P(T_R)\varepsilon_{t_R}.$$

  $T - T_U$ is the set of all observation indices in node t that are not missing.

- If you use surrogate decision splits on page 33-2413, then:

  i `fitrtree` computes the predictive measures of association on page 33-2412 between the decision split $x_j < u$ and all possible decision splits $x_k < v, j \neq k$.

  ii `fitrtree` sorts the possible alternative decision splits in descending order by their predictive measure of association with the optimal split. The surrogate split is the decision split yielding the largest measure.

  iii `fitrtree` decides the child node assignments for observations with a missing value for x_i using the surrogate split. If the surrogate predictor also contains a missing value, then `fitrtree` uses the decision split with the second largest measure, and so on, until there are no other surrogates. It is possible for `fitrtree` to split two different observations at node t using two different surrogate splits. For example, suppose the predictors x_1 and x_2 are the best and second best surrogates, respectively, for the predictor x_i, $i \notin \{1,2\}$, at node t. If observation m of predictor x_i is missing (i.e., x_{mi} is missing), but x_{m1} is not missing, then x_1 is the surrogate predictor for observation x_{mi}. If observations $x_{(m+1),i}$ and $x(m+1),1$ are missing, but $x_{(m+1),2}$ is not missing, then x_2 is the surrogate predictor for observation $m+1$.

  iv `fitrtree` uses the appropriate MSE reduction formula. That is, if `fitrtree` fails to assign all missing observations in node t to children nodes using surrogate splits, then the MSE reduction is ΔI_U. Otherwise, `fitrtree` uses ΔI for the MSE reduction.

c `fitrtree` chooses the candidate that yields the largest MSE reduction.

`fitrtree` splits the predictor variable at the cut point that maximizes the MSE reduction.

- For the curvature test (that is, if `PredictorSelection` is `'curvature'`):

1 `fitrtree` computes the residuals $r_{ti} = y_{ti} - \bar{y}_t$ for all observations in node t.

$\bar{y}_t = \frac{1}{\sum_i w_i}\sum_i w_i y_{ti}$, which is the weighted average of the responses in node t. The weights are the observation weights in `Weights`.

2 `fitrtree` assigns observations to one of two bins according to the sign of the corresponding residuals. Let z_t be a nominal variable that contains the bin assignments for the observations in node t.

3. `fitrtree` conducts curvature tests on page 33-2411 between each predictor and z_t. For regression trees, $K = 2$.
   - If all p-values are at least 0.05, then `fitrtree` does not split node t.
   - If there is a minimal p-value, then `fitrtree` chooses the corresponding predictor to split node t.
   - If more than one p-value is zero due to underflow, then `fitrtree` applies standard CART to the corresponding predictors to choose the split predictor.
4. If `fitrtree` chooses a split predictor, then it uses standard CART to choose the cut point (see step 4 in the standard CART process).

- For the interaction test (that is, if `PredictorSelection` is `'interaction-curvature'`):
  1. For observations in node t, `fitrtree` conducts curvature tests on page 33-2411 between each predictor and the response and interaction tests on page 33-2412 between each pair of predictors and the response.
     - If all p-values are at least 0.05, then `fitrtree` does not split node t.
     - If there is a minimal p-value and it is the result of a curvature test, then `fitrtree` chooses the corresponding predictor to split node t.
     - If there is a minimal p-value and it is the result of an interaction test, then `fitrtree` chooses the split predictor using standard CART on the corresponding pair of predictors.
     - If more than one p-value is zero due to underflow, then `fitrtree` applies standard CART to the corresponding predictors to choose the split predictor.
  2. If `fitrtree` chooses a split predictor, then it uses standard CART to choose the cut point (see step 4 in the standard CART process).

Tree Depth Control

- If `MergeLeaves` is `'on'` and `PruneCriterion` is `'mse'` (which are the default values for these name-value pair arguments), then the software applies pruning only to the leaves and by using MSE. This specification amounts to merging leaves coming from the same parent node whose MSE is at most the sum of the MSE of its two leaves.
- To accommodate `MaxNumSplits`, `fitrtree` splits all nodes in the current layer, and then counts the number of branch nodes. A layer is the set of nodes that are equidistant from the root node. If the number of branch nodes exceeds `MaxNumSplits`, `fitrtree` follows this procedure:
  1. Determine how many branch nodes in the current layer must be unsplit so that there are at most `MaxNumSplits` branch nodes.
  2. Sort the branch nodes by their impurity gains.
  3. Unsplit the number of least successful branches.
  4. Return the decision tree grown so far.

  This procedure produces maximally balanced trees.
- The software splits branch nodes layer by layer until at least one of these events occurs:
  - There are `MaxNumSplits` branch nodes.
  - A proposed split causes the number of observations in at least one branch node to be fewer than `MinParentSize`.
  - A proposed split causes the number of observations in at least one leaf node to be fewer than `MinLeafSize`.

- The algorithm cannot find a good split within a layer (i.e., the pruning criterion (see `PruneCriterion`), does not improve for all proposed splits in a layer). A special case is when all nodes are pure (i.e., all observations in the node have the same class).
- For values `'curvature'` or `'interaction-curvature'` of `PredictorSelection`, all tests yield *p*-values greater than 0.05.

`MaxNumSplits` and `MinLeafSize` do not affect splitting at their default values. Therefore, if you set `'MaxNumSplits'`, splitting might stop due to the value of `MinParentSize`, before `MaxNumSplits` splits occur.

Parallelization

For dual-core systems and above, `fitrtree` parallelizes training decision trees using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[2] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[3] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Supported syntaxes are:
  - `tree = fitrtree(Tbl,Y)`
  - `tree = fitrtree(X,Y)`
  - `tree = fitrtree(___,Name,Value)`
  - `[tree,FitInfo,HyperparameterOptimizationResults] = fitrtree(___,Name,Value)` — `fitrtree` returns the additional output arguments `FitInfo` and `HyperparameterOptimizationResults` when you specify the `'OptimizeHyperparameters'` name-value pair argument.

  `tree` is a `CompactRegressionTree` object; therefore, it does not include the data used in training the regression tree.
- The `FitInfo` output argument is an empty structure array currently reserved for possible future use.
- The `HyperparameterOptimizationResults` output argument is a `BayesianOptimization` object or a table of hyperparameters with associated values that describe the cross-validation optimization of hyperparameters.

`'HyperparameterOptimizationResults'` is nonempty when the `'OptimizeHyperparameters'` name-value pair argument is nonempty at the time you create the model. The values in `'HyperparameterOptimizationResults'` depend on the value you specify for the `'HyperparameterOptimizationOptions'` name-value pair argument when you create the model.

- If you specify `'bayesopt'` (default), then `HyperparameterOptimizationResults` is an object of class `BayesianOptimization`.
- If you specify `'gridsearch'` or `'randomsearch'`, then `HyperparameterOptimizationResults` is a table of the hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst).

- Supported name-value pair arguments are:
  - `'CategoricalPredictors'`
  - `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.
  - `'MaxNumSplits'` — For tall optimization, `fitrtree` searches among integers, log-scaled (by default) in the range `[1,max(2,min(10000,NumObservations−1))]`.
  - `'MergeLeaves'`
  - `'MinLeafSize'` — For tall optimization, `fitrtree` searches among integers, log-scaled (by default) in the range `[1,max(2,floor(NumObservations/2))]`.
  - `'MinParentSize'`
  - `'NumVariablesToSample'` — For tall optimization, `fitrtree` searches among integers in the range `[1,max(2,NumPredictors)]`.
  - `'OptimizeHyperparameters'`
  - `'PredictorNames'`
  - `'QuadraticErrorTolerance'`
  - `'ResponseName'`
  - `'ResponseTransform'`
  - `'SplitCriterion'`
  - `'Weights'`
- This additional name-value pair argument is specific to tall arrays:
  - `'MaxDepth'` — A positive integer specifying the maximum depth of the output tree. Specify a value for this argument to return a tree that has fewer levels and requires fewer passes through the tall array to compute. Generally, the algorithm of `fitrtree` takes one pass through the data and an additional pass for each tree level. The function does not set a maximum tree depth, by default.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionPartitionedModel` | `RegressionTree` | `predict` | `prune` | `surrogateAssociation`

Topics

“Splitting Categorical Predictors in Classification Trees” on page 19-25

Introduced in R2014a

fitSVMPosterior

Fit posterior probabilities

Syntax

```
ScoreSVMModel = fitSVMPosterior(SVMModel)

ScoreSVMModel = fitSVMPosterior(SVMModel,TBL,ResponseVarName)
ScoreSVMModel = fitSVMPosterior(SVMModel,TBL,Y)

ScoreSVMModel = fitSVMPosterior(SVMModel,X,Y)

ScoreSVMModel = fitSVMPosterior(___,Name,Value)

[ScoreSVMModel,ScoreTransform] = fitSVMPosterior(___)
```

Description

`ScoreSVMModel = fitSVMPosterior(SVMModel)` returns `ScoreSVMModel`, which is a trained, support vector machine (SVM) classifier containing the optimal score-to-posterior-probability transformation function for two-class learning.

The software fits the appropriate score-to-posterior-probability transformation function using the SVM classifier `SVMModel`, and by cross validation using the stored predictor data (`SVMModel.X`) and the class labels (`SVMModel.Y`). The transformation function computes the posterior probability that an observation is classified into the positive class (`SVMModel.Classnames(2)`).

- If the classes are inseparable, then the transformation function is the sigmoid function on page 33-2427.
- If the classes are perfectly separable, the transformation function is the step function on page 33-2427.
- In two-class learning, if one of the two classes has a relative frequency of 0, then the transformation function is the constant function on page 33-2428. `fitSVMPosterior` is not appropriate for one-class learning.
- If `SVMModel` is a `ClassificationSVM` classifier, then the software estimates the optimal transformation function by 10-fold cross validation as outlined in [1]. Otherwise, `SVMModel` must be a `ClassificationPartitionedModel` classifier. `SVMModel` specifies the cross-validation method.
- The software stores the optimal transformation function in `ScoreSVMModel.ScoreTransform`.

`ScoreSVMModel = fitSVMPosterior(SVMModel,TBL,ResponseVarName)` returns a trained support vector classifier containing the transformation function from the trained, compact SVM classifier `SVMModel`. The software estimates the score transformation function using predictor data in the table `TBL` and class labels `TBL.ResponseVarName`.

`ScoreSVMModel = fitSVMPosterior(SVMModel,TBL,Y)` returns a trained support vector classifier containing the transformation function from the trained, compact SVM classifier `SVMModel`. The software estimates the score transformation function using predictor data in the table `TBL` and class labels `Y`.

`ScoreSVMModel = fitSVMPosterior(SVMModel,X,Y)` returns a trained support vector classifier containing the transformation function from the trained, compact SVM classifier `SVMModel`. The software estimates the score transformation function using predictor data `X` and class labels `Y`.

`ScoreSVMModel = fitSVMPosterior(___,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments provided `SVMModel` is a `ClassificationSVM` classifier. For example, you can specify the number of folds to use in *k*-fold cross validation.

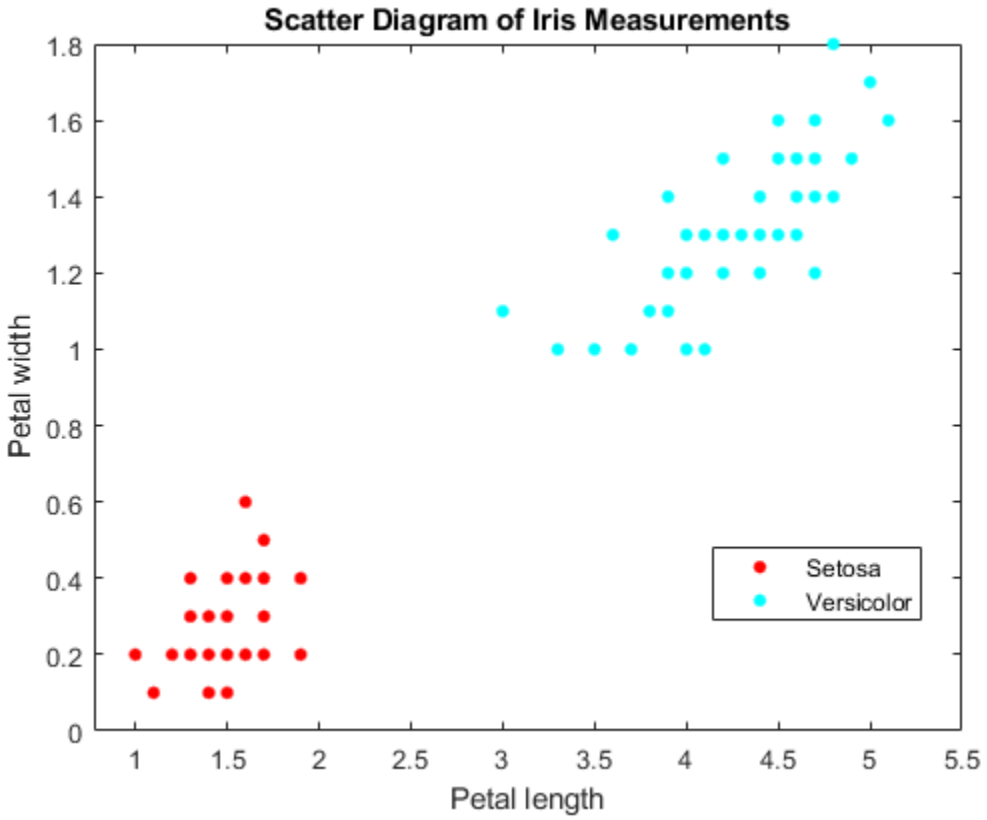
`[ScoreSVMModel,ScoreTransform] = fitSVMPosterior(___)` additionally returns the transformation function parameters (`ScoreTransform`) using any of the input arguments in the previous syntaxes.

Examples

Fit the Score-to-Posterior Probability Function for Separable Classes

Load Fisher's iris data set. Train the classifier using the petal lengths and widths, and remove the virginica species from the data.

```
load fisheriris
classKeep = ~strcmp(species,'virginica');
X = meas(classKeep,3:4);
y = species(classKeep);

gscatter(X(:,1),X(:,2),y);
title('Scatter Diagram of Iris Measurements')
xlabel('Petal length')
ylabel('Petal width')
legend('Setosa','Versicolor')
```



The classes are perfectly separable. Therefore, the score transformation function is a step function.

Train an SVM classifier using the data. Cross validate the classifier using 10-fold cross validation (the default).

```
rng(1);
CVSVMModel = fitcsvm(X,y,'CrossVal','on');
```

`CVSVMModel` is a trained `ClassificationPartitionedModel` SVM classifier.

Estimate the step function that transforms scores to posterior probabilities.

```
[ScoreCVSVMModel,ScoreParameters] = fitSVMPosterior(CVSVMModel);
```

```
Warning: Classes are perfectly separated. The optimal score-to-posterior transformation is a step
```

`fitSVMPosterior` does the following:

- Uses the data that the software stored in `CVSVMModel` to fit the transformation function
- Warns whenever the classes are separable
- Stores the step function in `ScoreCSVMModel.ScoreTransform`

Display the score function type and its parameter values.

```
ScoreParameters
```

```
ScoreParameters = struct with fields:
                     Type: 'step'
```

```
                LowerBound: -0.8431
                UpperBound: 0.6897
    PositiveClassProbability: 0.5000
```

`ScoreParameters` is a structure array with four fields:

- The score transformation function type (`Type`)
- The score corresponding to the negative class boundary (`LowerBound`)
- The score corresponding to the positive class boundary (`UpperBound`)
- The positive class probability (`PositiveClassProbability`)

Since the classes are separable, the step function transforms the score to either `0` or `1`, which is the posterior probability that an observation is a versicolor iris.

Fit the Score-to-Posterior Probability Function for Inseparable Classes

Load the `ionosphere` data set.

```
load ionosphere
```

The classes of this data set are not separable.

Train an SVM classifier. Cross validate using 10-fold cross validation (the default). It is good practice to standardize the predictors and specify the class order.

```
rng(1) % For reproducibility
CVSVMModel = fitcsvm(X,Y,'ClassNames',{'b','g'},'Standardize',true,...
    'CrossVal','on');
ScoreTransform = CVSVMModel.ScoreTransform
```

```
ScoreTransform =
'none'
```

`CVSVMModel` is a trained `ClassificationPartitionedModel` SVM classifier. The positive class is `'g'`. The `ScoreTransform` property is `none`.

Estimate the optimal score function for mapping observation scores to posterior probabilities of an observation being classified as `'g'`.

```
[ScoreCVSVMModel,ScoreParameters] = fitSVMPosterior(CVSVMModel);
ScoreTransform = ScoreCVSVMModel.ScoreTransform
```

```
ScoreTransform =
'@(S)sigmoid(S,-9.481989e-01,-1.218252e-01)'
```

```
ScoreParameters
```

```
ScoreParameters = struct with fields:
         Type: 'sigmoid'
        Slope: -0.9482
    Intercept: -0.1218
```

`ScoreTransform` is the optimal score transform function. `ScoreParameters` contains the score transformation function, slope estimate, and the intercept estimate.

You can estimate test-sample, posterior probabilities by passing `ScoreCVSVMModel` to `kfoldPredict`.

Estimate Posterior Probabilities for Test Samples

Estimate positive class posterior probabilities for the test set of an SVM algorithm.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Specify a 20% holdout sample. It is good practice to standardize the predictors and specify the class order.

```
rng(1) % For reproducibility
CVSVMModel = fitcsvm(X,Y,'Holdout',0.2,'Standardize',true,...
    'ClassNames',{'b','g'});
```

`CVSVMModel` is a trained `ClassificationPartitionedModel` cross-validated classifier.

Estimate the optimal score function for mapping observation scores to posterior probabilities of an observation being classified as `'g'`.

```
ScoreCVSVMModel = fitSVMPosterior(CVSVMModel);
```

`ScoreSVMModel` is a trained `ClassificationPartitionedModel` cross-validated classifier containing the optimal score transformation function estimated from the training data.

Estimate the out-of-sample positive class posterior probabilities. Display the results for the first 10 out-of-sample observations.

```
[~,OOSPostProbs] = kfoldPredict(ScoreCVSVMModel);
indx = ~isnan(OOSPostProbs(:,2));
hoObs = find(indx); % Holdout observation numbers
OOSPostProbs = [hoObs, OOSPostProbs(indx,2)];
table(OOSPostProbs(1:10,1),OOSPostProbs(1:10,2),...
    'VariableNames',{'ObservationIndex','PosteriorProbability'})
```

```
ans=10×2 table
    ObservationIndex    PosteriorProbability
    ________________    ____________________

           6                    0.17381
           7                    0.89639
           8                  0.0076613
           9                    0.91602
          16                   0.026722
          22                  4.6114e-06
          23                      0.9024
          24                  2.4137e-06
          38                  0.00042705
```

```
    41                  0.86427
```

Input Arguments

SVMModel — Trained SVM classifier

`ClassificationSVM` classifier | `CompactClassificationSVM` classifier | `ClassificationPartitionedModel` classifier

Trained SVM classifier, specified as a `ClassificationSVM`, `CompactClassificationSVM`, or `ClassificationPartitionedModel` classifier.

If `SVMModel` is a `ClassificationSVM` classifier, then you can set optional name-value pair arguments.

If `SVMModel` is a `CompactClassificationSVM` classifier, then you must input predictor data `X` and class labels `Y`.

TBL — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all of the predictors used to train `SVMModel`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName` or `Y`.

If you trained `SVMModel` using sample data contained in a table, then the input data for `fitSVMPosterior` must also be in a table.

If you set `'Standardize',true` in `fitcsvm` when training `SVMModel`, then the software standardizes the columns of the predictor data using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `table`

X — Predictor data

matrix

Predictor data used to estimate the score-to-posterior-probability transformation function, specified as a matrix.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature).

The length of `Y` and the number of rows in `X` must be equal.

If you set `'Standardize',true` in `fitcsvm` when training `SVMModel`, then the software fits the transformation function parameter estimates using standardized data.

Data Types: `double` | `single`

`ResponseVarName` — Response variable name
name of variable in `TBL`

Response variable name, specified as the name of a variable in `TBL`. If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `TBL.Response`, then specify `ResponseVarName` as `'Response'`. Otherwise, the software treats all columns of `TBL`, including `TBL.Response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels used to estimate the score-to-posterior-probability transformation function, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors.

If `Y` is a character array, then each element must correspond to one class label.

The length of `Y` and the number of rows in `X` must be equal.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KFold',8` performs 8-fold cross validation when `SVMModel` is a `ClassificationSVM` classifier.

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition

Cross-validation partition used to compute the transformation function, specified as the comma-separated pair consisting of `'CVPartition'` and a `cvpartition` partition object as created by `cvpartition`. You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

The `crossval` name-value pair argument of `fitcsvm` splits the data into subsets using `cvpartition`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data for holdout validation used to compute the transformation function, specified as the comma-separated pair consisting of `'Holdout'` and a scalar value in the range (0,1). Holdout validation tests the specified fraction of the data and uses the remaining data for training.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use when computing the transformation function, specified as the comma-separated pair consisting of `'KFold'` and a positive integer value greater than 1.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'KFold',8`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag indicating whether to use leave-one-out cross-validation to compute the transformation function, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. Use leave-one-out cross-validation by specifying `'Leaveout','on'`.

You can use only one of these four options at a time for creating a cross-validated model: `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`.

Example: `'Leaveout','on'`

Output Arguments

`ScoreSVMModel` — Trained SVM classifier
`ClassificationSVM` classifier | `CompactClassificationSVM` classifier | `ClassificationPartitionedModel` classifier

Trained SVM classifier containing the estimated score transformation function, returned as a `ClassificationSVM`, `CompactClassificationSVM`, or `ClassificationPartitionedModel` classifier.

The `ScoreSVMModel` classifier type is the same as the `SVMModel` classifier type.

To estimate posterior probabilities, pass `ScoreSVMModel` and predictor data to `predict`. If you set `'Standardize',true` in `fitcsvm` to train `SVMModel`, then `predict` standardizes the columns of `X` using the corresponding means in `SVMModel.Mu` and standard deviations in `SVMModel.Sigma`.

`ScoreTransform` — Optimal score-to-posterior-probability transformation function parameters
structure array

Optimal score-to-posterior-probability transformation function parameters, specified as a structure array. If field `Type` is:

- `sigmoid`, then `ScoreTransform` has these fields:
  - `Slope` — The value of A in the sigmoid function on page 33-2427
  - `Intercept` — The value of `B` in the sigmoid function
- `step`, then `ScoreTransform` has these fields:
  - `PositiveClassProbability`: the value of π in the step function on page 33-2427. π represents:
    - The probability that an observation is in the positive class.
    - The posterior probability that a score is in the interval (`LowerBound`,`UpperBound`).
  - `LowerBound`: the value $\max_{y_n = -1} s_n$ in the step function. It represents the lower bound of the interval that assigns the posterior probability of being in the positive class `PositiveClassProbability` to scores. Any observation with a score less than `LowerBound` has posterior probability of being the positive class `0`.
  - `UpperBound`: the value $\min_{y_n = +1} s_n$ in the step function. It represents the upper bound of the interval that assigns the posterior probability of being in the positive class `PositiveClassProbability`. Any observation with a score greater than `UpperBound` has posterior probability of being the positive class `1`.
- `constant`, then `ScoreTransform.PredictedClass` contains the name of the class prediction.

  This result is the same as `SVMModel.ClassNames`. The posterior probability of an observation being in `ScoreTransform.PredictedClass` is always `1`.

More About

Sigmoid Function

The sigmoid function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \frac{1}{1 + \exp(As_j + B)}.$$

If the value of the `Type` field of `ScoreTransform` is `sigmoid`, then parameters A and B correspond to the fields `Scale` and `Intercept` of `ScoreTransform`, respectively.

Step Function

The step function that maps score s_j corresponding to observation j to the positive class posterior probability is

$$P(s_j) = \begin{cases} 0; & s < \max_{y_k = -1} s_k \\ \pi; & \max_{y_k = -1} s_k \le s_j \le \min_{y_k = +1} s_k, \\ 1; & s_j > \min_{y_k = +1} s_k \end{cases}$$

where:

- s_j is the score of observation j.
- +1 and –1 denote the positive and negative classes, respectively.
- π is the prior probability that an observation is in the positive class.

If the value of the `Type` field of `ScoreTransform` is `step`, then the quantities $\max_{y_k = -1} s_k$ and $\min_{y_k = +1} s_k$ correspond to the fields `LowerBound` and `UpperBound` of `ScoreTransform`, respectively.

Constant Function

The constant function maps all scores in a sample to posterior probabilities 1 or 0.

If all observations have posterior probability 1, then they are expected to come from the positive class.

If all observations have posterior probability 0, then they are not expected to come from the positive class.

Tips

- This process describes one way to predict positive class posterior probabilities.

  1. Train an SVM classifier by passing the data to `fitcsvm`. The result is a trained SVM classifier, such as `SVMModel`, that stores the data. The software sets the score transformation function property (`SVMModel.ScoreTransformation`) to `none`.
  2. Pass the trained SVM classifier `SVMModel` to `fitSVMPosterior` or `fitPosterior`. The result, such as, `ScoreSVMModel`, is the same trained SVM classifier as `SVMModel`, except the software sets `ScoreSVMModel.ScoreTransformation` to the optimal score transformation function.
  3. Pass the predictor data matrix and the trained SVM classifier containing the optimal score transformation function (`ScoreSVMModel`) to `predict`. The second column in the second output argument of `predict` stores the positive class posterior probabilities corresponding to each row of the predictor data matrix.

     If you skip step 2, then `predict` returns the positive class score rather than the positive class posterior probability.

- After fitting posterior probabilities, you can generate C/C++ code that predicts labels for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

If you re-estimate the score-to-posterior-probability transformation function, that is, if you pass an SVM classifier to `fitPosterior` or `fitSVMPosterior` and its `ScoreTransform` property is not `none`, then the software:

- Displays a warning
- Resets the original transformation function to `'none'` before estimating the new one

References

[1] Platt, J. “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods”. In: *Advances in Large Margin Classifiers*. Cambridge, MA: The MIT Press, 2000, pp. 61–74.

See Also

`ClassificationPartitionedModel` | `ClassificationSVM` | `CompactClassificationSVM` | `fitPosterior` | `fitPosterior` | `fitcsvm` | `kfoldPredict` | `predict`

Introduced in R2014a

fitted

Class: `GeneralizedLinearMixedModel`

Fitted responses from generalized linear mixed-effects model

Syntax

```
mufit = fitted(glme)
mufit = fitted(glme,Name,Value)
```

Description

`mufit = fitted(glme)` returns the fitted conditional response of the generalized linear mixed-effects model `glme`.

`mufit = fitted(glme,Name,Value)` returns the fitted response with additional options specified by one or more name-value pair arguments. For example, you can specify to compute the marginal fitted response.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Conditional` — Indicator for conditional response

`true` (default) | `false`

Indicator for conditional response, specified as the comma-separated pair consisting of `'Conditional'` and one of the following.

Value	Description
`true`	Contributions from both fixed effects and random effects (conditional)
`false`	Contribution from only fixed effects (marginal)

To obtain fitted marginal response values, `fitted` computes the conditional mean of the response with the empirical Bayes predictor vector of random effects *b* set equal to 0. For more information, see “Conditional and Marginal Response” on page 33-2433

Example: `'Conditional',false`

Output Arguments

`mufit` — Fitted response values

n-by-1 vector

Fitted response values, returned as an *n*-by-1 vector, where *n* is the number of observations.

Examples

Plot Observed Versus Fitted Values

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

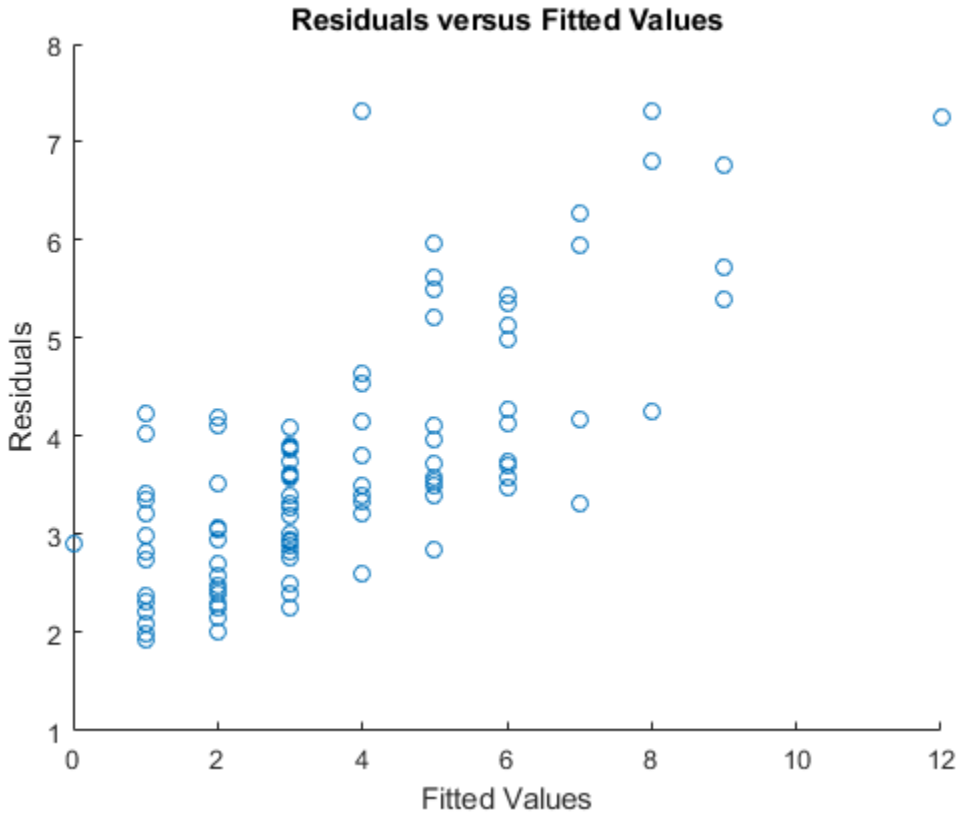
```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)', ..
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Generate the fitted conditional mean values for the model.

```
mufit = fitted(glme);
```

Create a scatterplot of the observed values versus fitted values.

```
figure
scatter(mfr.defects,mufit)
title('Residuals versus Fitted Values')
xlabel('Fitted Values')
ylabel('Residuals')
```



More About

Conditional and Marginal Response

A *conditional response* includes contributions from both fixed- and random-effects predictors. A *marginal response* includes contribution from only fixed effects.

Suppose the generalized linear mixed-effects model `glme` has an *n*-by-*p* fixed-effects design matrix `X` and an *n*-by-*q* random-effects design matrix `Z`. Also, suppose the estimated *p*-by-1 fixed-effects vector is $\widehat{\beta}$, and the *q*-by-1 empirical Bayes predictor vector of random effects is $\widehat{b}$.

The fitted conditional response corresponds to the `'Conditional',true` name-value pair argument, and is defined as

$$\widehat{\mu}_{cond} = g^{-1}(\widehat{\eta}_{ME}),$$

where $\widehat{\eta}_{ME}$ is the linear predictor including the fixed- and random-effects of the generalized linear mixed-effects model

$$\widehat{\eta}_{ME} = X\widehat{\beta} + Z\widehat{b} + \delta .$$

The fitted marginal response corresponds to the `'Conditional',false` name-value pair argument, and is defined as

$$\widehat{\mu}_{mar} = g^{-1}(\widehat{\eta}_{FE}),$$

where $\widehat{\eta}_{FE}$ is the linear predictor including only the fixed-effects portion of the generalized linear mixed-effects model

$$\widehat{\eta}_{FE} = X\widehat{\beta} + \delta \ .$$

See Also

`GeneralizedLinearMixedModel` | `designMatrix` | `fitglme` | `residuals` | `response`

fitted

Class: `LinearMixedModel`

Fitted responses from a linear mixed-effects model

Syntax

```
yfit = fitted(lme)
yfit = fitted(lme,Name,Value)
```

Description

`yfit = fitted(lme)` returns the fitted conditional response on page 33-2440 from the linear mixed-effects model `lme`.

`yfit = fitted(lme,Name,Value)` returns the fitted response from the linear mixed-effects model `lme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify if you want to compute the fitted marginal response on page 33-2440.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Conditional` — Indicator for conditional response

`true` (default) | false

Indicator for conditional response, specified as the comma-separated pair consisting of `'Conditional'` and either of the following.

`true`	Contribution from both fixed effects and random effects (conditional)
`false`	Contribution from only fixed effects (marginal)

Example: `'Conditional',false`

Data Types: `logical`

Output Arguments

yfit — Fitted response values

n-by-1 vector

Fitted response values, returned as an *n*-by-1 vector, where *n* is the number of observations.

Examples

Compute Fitted Conditional and Marginal Responses

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the Center for Disease Control and Prevention, CDC).

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses and region as the predictor variable, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the response variable, `FluRate`, the nominal variable, `Region`, that shows which region each estimate is from, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate','IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Fit a linear mixed-effects model with fixed effects for region and a random intercept that varies by `Date`.

Region is a categorical variable. You can specify the contrasts for categorical variables using the `DummyVarCoding` name-value pair argument when fitting the model. When you do not specify the contrasts, `fitlme` uses the `'reference'` contrast by default. Because the model has an intercept, `fitlme` takes the first region, `NE`, as the reference and creates eight dummy variables representing the other eight regions. For example, $I[\text{MidAtl}]$ is the dummy variable representing the region `MidAtl`. For details, see “Dummy Variables” on page 2-48.

The corresponding model is

$$y_{im} = \beta_0 + \beta_1 I[\text{MidAtl}]_i + \beta_2 I[\text{ENCentral}]_i + \beta_3 I[\text{WNCentral}]_i + \beta_4 I[\text{SAtl}]_i + \beta_5 I[\text{ESCentral}]_i + \beta_6 I[\text{WSCentral}]_i + \beta_7 I[\text{Mtn}]_i + \beta_8 I[\text{Pac}]_i + b_{0m} + \varepsilon_{im}, \quad m = 1, 2, \ldots, 52,$$

where y_{im} is the observation i for level m of grouping variable `Date`, β_j, $j = 0, 1, \ldots, 8$, are the fixed-effects coefficients, with β_0 being the coefficient for region `NE`. b_{0m} is the random effect for level m of the grouping variable `Date`, and ε_{im} is the observation error for observation i. The random effect has the prior distribution, $b_{0m} \sim N(0, \sigma_b^2)$ and the error term has the distribution, $\varepsilon_{im} \sim N(0, \sigma^2)$.

```
lme = fitlme(flu2,'FluRate ~ 1 + Region + (1|Date)')
```

```
lme =
Linear mixed-effects model fit by ML
```

```
Model information:
    Number of observations             468
    Fixed effects coefficients           9
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    FluRate ~ 1 + Region + (1 | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    318.71    364.35    -148.36          296.71

Fixed effects coefficients (95% CIs):
    Name                        Estimate    SE          tStat      DF
    {'(Intercept)'     }          1.2233    0.096678     12.654    459
    {'Region_MidAtl'   }        0.010192    0.052221    0.19518    459
    {'Region_ENCentral'}        0.051923    0.052221     0.9943    459
    {'Region_WNCentral'}         0.23687    0.052221     4.5359    459
    {'Region_SAtl'     }        0.075481    0.052221     1.4454    459
    {'Region_ESCentral'}         0.33917    0.052221      6.495    459
    {'Region_WSCentral'}           0.069    0.052221     1.3213    459
    {'Region_Mtn'      }        0.046673    0.052221    0.89377    459
    {'Region_Pac'      }        -0.16013    0.052221    -3.0665    459


    pValue        Lower        Upper
     1.085e-31       1.0334       1.4133
       0.84534    -0.092429      0.11281
        0.3206    -0.050698      0.15454
    7.3324e-06      0.13424      0.33949
       0.14902     -0.02714       0.1781
    2.1623e-10      0.23655      0.44179
       0.18705    -0.033621      0.17162
       0.37191    -0.055948      0.14929
     0.0022936     -0.26276    -0.057514

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.6443


    Lower     Upper
    0.5297    0.78368

Group: Error
    Name               Estimate    Lower      Upper
    {'Res Std'}        0.26627     0.24878    0.285
```

The *p*-values 7.3324e-06 and 2.1623e-10 respectively show that the fixed effects of the flu rates in regions `WNCentral` and `ESCentral` are significantly different relative to the flu rates in region `NE`.

The confidence limits for the standard deviation of the random-effects term, σ_b, do not include 0 (0.5297, 0.78368), which indicates that the random-effects term is significant. You can also test the significance of the random-effects terms using the `compare` method.

The conditional fitted response from the model at a given observation includes contributions from fixed and random effects. For example, the estimated best linear unbiased predictor (BLUP) of the flu rate for region `WNCentral` in week 10/9/2005 is

$$\begin{aligned}\hat{y}_{\text{WNCentral, 10/9/2005}} &= \hat{\beta}_0 + \hat{\beta}_3 I[\text{WNCentral}] + \hat{b}_{10/9/2005} \\ &= 1.2233 + 0.23687 - 0.1718 \\ &= 1.28837.\end{aligned}$$

This is the fitted conditional response, since it includes contributions to the estimate from both the fixed and random effects. You can compute this value as follows.

```
beta = fixedEffects(lme);
[~,~,STATS] = randomEffects(lme); % Compute the random-effects statistics (STATS)
STATS.Level = nominal(STATS.Level);
y_hat = beta(1) + beta(4) + STATS.Estimate(STATS.Level=='10/9/2005')
```

```
y_hat = 1.2884
```

In the previous calculation, `beta(1)` corresponds to the estimate for β_0 and `beta(4)` corresponds to the estimate for β_3. You can simply display the fitted value using the `fitted` method.

```
F = fitted(lme);
F(flu2.Date == '10/9/2005' & flu2.Region == 'WNCentral')
```

```
ans = 1.2884
```

The estimated marginal response for region `WNCentral` in week 10/9/2005 is

$$\begin{aligned}\hat{y}^{(\text{marginal})}_{\text{WNCentral, 10/9/2005}} &= \hat{\beta}_0 + \hat{\beta}_3 I[\text{WNCentral}] \\ &= 1.2233 + 0.23687 \\ &= 1.46017.\end{aligned}$$

Compute the fitted marginal response.

```
F = fitted(lme,'Conditional',false);
F(flu2.Date == '10/9/2005' & flu2.Region == 'WNCentral')
```

```
ans = 1.4602
```

Plot Residuals vs. Fitted Values

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

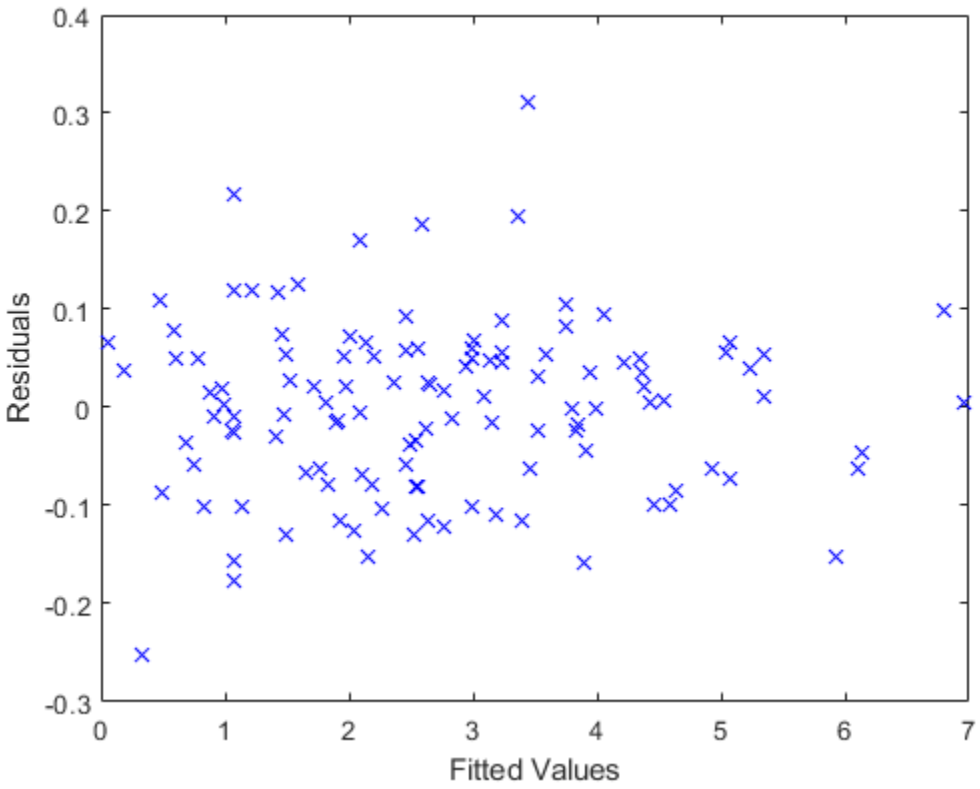
```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

Compute the fitted values and raw residuals.

```
F = fitted(lme);
R = residuals(lme);
```

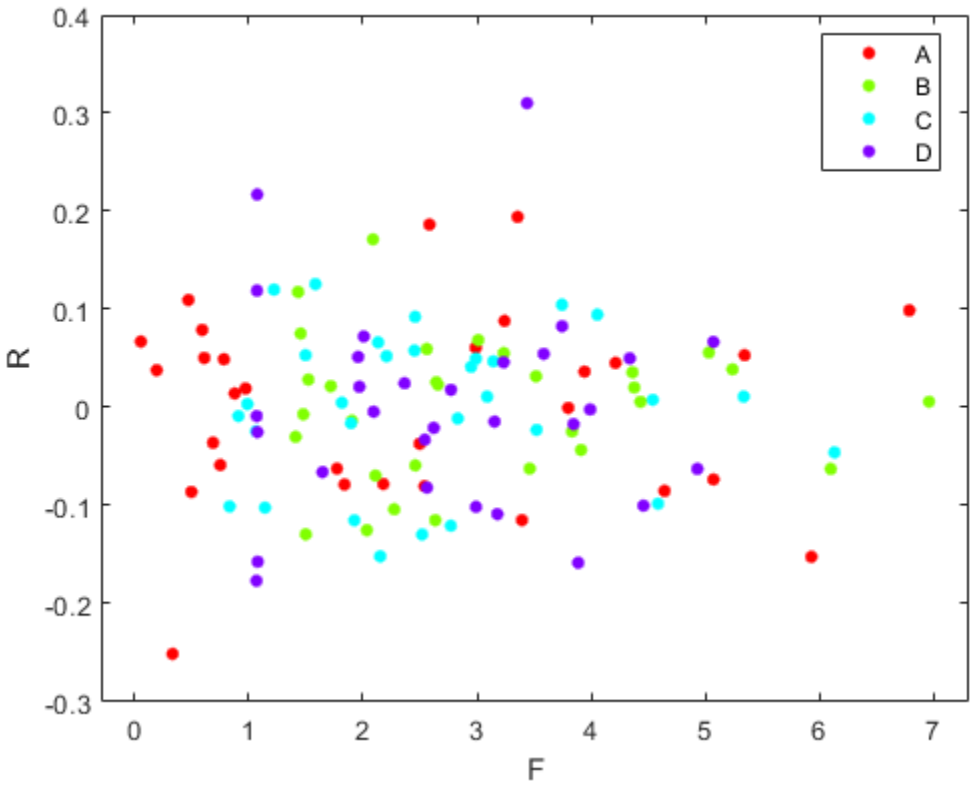
Plot the residuals versus the fitted values.

```
plot(F,R,'bx')
xlabel('Fitted Values')
ylabel('Residuals')
```



Now, plot the residuals versus the fitted values, grouped by program.

```
figure()
gscatter(F,R,Program)
```



More About

Fitted Conditional and Marginal Response

A conditional response includes contributions from both fixed and random effects, whereas a marginal response includes contribution from only fixed effects.

Suppose the linear mixed-effects model, `lme`, has an n-by-p fixed-effects design matrix X and an n-by-q random-effects design matrix Z. Also, suppose the p-by-1 estimated fixed-effects vector is $\widehat{\beta}$, and the q-by-1 estimated best linear unbiased predictor (BLUP) vector of random effects is $\widehat{b}$. The fitted conditional response is

$$\widehat{y}_{Cond} = X\widehat{\beta} + Z\widehat{b},$$

and the fitted marginal response is

$$\widehat{y}_{Mar} = X\widehat{\beta},$$

See Also

`LinearMixedModel` | `residuals` | `response`

fixedEffects

Class: `GeneralizedLinearMixedModel`

Estimates of fixed effects and related statistics

Syntax

```
beta = fixedEffects(glme)
[beta,betanames] = fixedEffects(glme)
[beta,betanames,stats] = fixedEffects(glme)
[___] = fixedEffects(glme,Name,Value)
```

Description

`beta = fixedEffects(glme)` returns the estimated fixed-effects coefficients, `beta`, of the generalized linear mixed-effects model `glme`.

`[beta,betanames] = fixedEffects(glme)` also returns the names of estimated fixed-effects coefficients in `betanames`. Each name corresponds to a fixed-effects coefficient in `beta`.

`[beta,betanames,stats] = fixedEffects(glme)` also returns a table of statistics, `stats`, related to the estimated fixed-effects coefficients of `glme`.

`[___] = fixedEffects(glme,Name,Value)` returns any of the output arguments in previous syntaxes using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence level, or the method for computing the approximate degrees of freedom for the *t*-statistic.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range [0,1]

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range [0,1]. For a value α, the confidence level is 100 × (1 – α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`DFMethod` — Method for computing approximate degrees of freedom
`'residual'` (default) | `'none'`

Method for computing approximate degrees of freedom, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom value is assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	The degrees of freedom is set to infinity.

Example: `'DFMethod','none'`

Output Arguments

`beta` — Estimated fixed-effects coefficients
vector

Estimated fixed-effects coefficients of the fitted generalized linear mixed-effects model `glme`, returned as a vector.

`betanames` — Names of fixed-effects coefficients
table

Names of fixed-effects coefficients in `beta`, returned as a table.

`stats` — Fixed-effects estimates and related statistics
dataset array

Fixed-effects estimates and related statistics, returned as a dataset array that has one row for each of the fixed effects and one column for each of the following statistics.

Column Name	Description
`Name`	Name of the fixed-effects coefficient
`Estimate`	Estimated coefficient value
`SE`	Standard error of the estimate
`tStat`	*t*-statistic for a test that the coefficient is 0
`DF`	Estimated degrees of freedom for the *t*-statistic
`pValue`	*p*-value for the *t*-statistic
`Lower`	Lower limit of a 95% confidence interval for the fixed-effects coefficient
`Upper`	Upper limit of a 95% confidence interval for the fixed-effects coefficient

When fitting a model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`), if you specify the `'CovarianceMethod'` name-value pair argument as

`'conditional'`, then `SE` does not account for the uncertainty in estimating the covariance parameters. To account for this uncertainty, specify `'CovarianceMethod'` as `'JointHessian'`.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `fixedEffects` bases the fixed effects estimates and related statistics on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

Examples

Estimate Fixed-Effects Coefficients

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)', ..
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Compute and display the estimated fixed-effects coefficient values and related statistics.

```
[beta,betanames,stats] = fixedEffects(glme);
stats
```

```
stats =
    Fixed effect coefficients: DFMethod = 'residual', Alpha = 0.05

    Name                    Estimate     SE          tStat       DF    pValue
    {'(Intercept)'}            1.4689     0.15988      9.1875    94    9.8194e-15
    {'newprocess' }          -0.36766     0.17755     -2.0708    94      0.041122
    {'time_dev'   }          -0.094521    0.82849    -0.11409    94       0.90941
    {'temp_dev'   }          -0.28317      0.9617    -0.29444    94       0.76907
    {'supplier_C' }          -0.071868    0.078024     -0.9211    94       0.35936
    {'supplier_B' }           0.071072     0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
    -0.72019    -0.015134
     -1.7395       1.5505
     -2.1926       1.6263
    -0.22679     0.083051
    -0.082588      0.22473
```

The returned results indicate, for example, that the estimated coefficient for `temp_dev` is –0.28317. Its large *p*-value, 0.76907, indicates that it is not a statistically significant predictor at the 5% significance level. Additionally, the confidence interval boundaries `Lower` and `Upper` indicate that the 95% confidence interval for the coefficient for `temp_dev` is [-2.1926 , 1.6263]. This interval contains 0, which supports the conclusion that `temp_dev` is not statistically significant at the 5% significance level.

See Also

`GeneralizedLinearMixedModel` | `coefCI` | `coefTest` | `fitglme` | `randomEffects`

fixedEffects

Class: `LinearMixedModel`

Estimates of fixed effects and related statistics

Syntax

```
beta = fixedEffects(lme)
[beta,betanames] = fixedEffects(lme)
[beta,betanames,stats] = fixedEffects(lme)
[beta,betanames,stats] = fixedEffects(lme,Name,Value)
```

Description

`beta = fixedEffects(lme)` returns the estimated fixed-effects coefficients, `beta`, of the linear mixed-effects model `lme`.

`[beta,betanames] = fixedEffects(lme)` also returns the names of estimated fixed-effects coefficients in `betanames`. Each name corresponds to a fixed-effects coefficient in `beta`.

`[beta,betanames,stats] = fixedEffects(lme)` also returns the estimated fixed-effects coefficients of the linear mixed-effects model `lme` and related statistics in `stats`.

`[beta,betanames,stats] = fixedEffects(lme,Name,Value)` also returns the estimated fixed-effects coefficients of the linear mixed-effects model `lme` and related statistics with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`DFMethod` — Method for computing approximate degrees of freedom
`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing approximate degrees of freedom for the *t*-statistic that tests the fixed-effects coefficients against 0, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

Output Arguments

`beta` — Fixed-effects coefficients estimates
vector

Fixed-effects coefficients estimates of the fitted linear mixed-effects model `lme`, returned as a vector.

`betanames` — Names of fixed-effects coefficients
table

Names of fixed-effects coefficients in `beta`, returned as a table.

`stats` — Fixed-effects estimates and related statistics
dataset array

Fixed-effects estimates and related statistics, returned as a dataset array that has one row for each of the fixed effects and one column for each of the following statistics.

`Name`	Name of the fixed effect coefficient
`Estimate`	Estimated coefficient value
`SE`	Standard error of the estimate
`tStat`	*t*-statistic for a test that the coefficient is zero
`DF`	Estimated degrees of freedom for the *t*-statistic
`pValue`	*p*-value for the *t*-statistic
`Lower`	Lower limit of a 95% confidence interval for the fixed-effect coefficient
`Upper`	Upper limit of a 95% confidence interval for the fixed-effect coefficient

Examples

Display Fixed-Effects Coefficient Estimates and Names

Load the sample data.

```
load('weight.mat');
```

The data set `weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between week and program are the fixed effects. The intercept and week vary by subject.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

Display the fixed-effects coefficient estimates and corresponding fixed-effects names.

```
[beta,betanames] = fixedEffects(lme)
```

```
beta = 9×1

    0.6610
    0.0032
    0.3608
   -0.0333
    0.1132
    0.1732
    0.0388
    0.0305
    0.0331


betanames=9×1 table
           Name
    ________________

    {'(Intercept)'   }
    {'InitialWeight' }
    {'Program_B'     }
    {'Program_C'     }
    {'Program_D'     }
    {'Week'          }
    {'Program_B:Week'}
    {'Program_C:Week'}
    {'Program_D:Week'}
```

Compute Coefficient Estimates and Related Statistics

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration and horsepower, and potentially correlated random effects for intercept and acceleration grouped by model year. First, store the data in a table.

```
tbl = table(Acceleration,Horsepower,Model_Year,MPG);
```

Fit the model.

```
lme = fitlme(tbl, 'MPG ~ Acceleration + Horsepower + (Acceleration|Model_Year)');
```

Compute the fixed-effects coefficients estimates and related statistics.

```
[~,~,stats] = fixedEffects(lme)
```

```
stats =
    Fixed effect coefficients: DFMethod = 'Residual', Alpha = 0.05

    Name                    Estimate    SE           tStat      DF
    {'(Intercept)' }          50.133       2.2652     22.132    389
    {'Acceleration'}        -0.58327      0.13394    -4.3545    389
    {'Horsepower'  }        -0.16954    0.0072609     -23.35    389


    pValue        Lower       Upper
    7.7727e-71      45.679      54.586
    1.7075e-05    -0.84661    -0.31992
     5.188e-76    -0.18382    -0.15527
```

The small *p*-values (under `pValue`) indicate that all fixed-effects coefficients are significant.

Compute Confidence Intervals with Specified Options

Load the sample data.

```
load('shift.mat');
```

The data shows the deviations from the target quality characteristic measured from the products that five operators manufacture during three shifts: morning, evening, and night. This is a randomized block design, where the operators are the blocks. The experiment is designed to study the impact of the time of shift on the performance. The performance measure is the deviation of the quality characteristics from the target value. This is simulated data.

`Shift` and `Operator` are nominal variables.

```
shift.Shift = nominal(shift.Shift);
shift.Operator = nominal(shift.Operator);
```

Fit a linear mixed-effects model with a random intercept grouped by operator to assess if performance significantly differs according to the time of the shift.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)');
```

Compute the 99% confidence intervals for fixed-effects coefficients, using the residual method to compute the degrees of freedom. This is the default method.

```
[~,~,stats] = fixedEffects(lme,'alpha',0.01)
```

```
stats =
    Fixed effect coefficients: DFMethod = 'Residual', Alpha = 0.01

    Name                    Estimate    SE         tStat       DF    pValue
    {'(Intercept)'  }        3.1196     0.88681      3.5178    12    0.0042407
    {'Shift_Morning'}       -0.3868     0.48344    -0.80009    12      0.43921
    {'Shift_Night'  }        1.9856     0.48344      4.1072    12    0.0014535


    Lower      Upper
    0.41081    5.8284
    -1.8635    1.0899
     0.5089    3.4623
```

Compute the 99% confidence intervals for fixed-effects coefficients, using the Satterthwaite approximation to compute the degrees of freedom.

```
[~,~,stats] = fixedEffects(lme,'DFMethod','satterthwaite','alpha',0.01)
```

```
stats =
    Fixed effect coefficients: DFMethod = 'Satterthwaite', Alpha = 0.01

    Name                    Estimate    SE         tStat       DF       pValue
    {'(Intercept)'  }        3.1196     0.88681      3.5178    6.123    0.01214
    {'Shift_Morning'}       -0.3868     0.48344    -0.80009       10    0.44225
    {'Shift_Night'  }        1.9856     0.48344      4.1072       10    0.00212


    Lower       Upper
    -0.14122    6.3804
      -1.919    1.1454
     0.45343    3.5178
```

The Satterthwaite approximation usually produces smaller `DF` values than the residual method. That is why it produces larger *p*-values (`pValue`) and larger confidence intervals (see `Lower` and `Upper`).

See Also

`LinearMixedModel` | `coefCI` | `coefTest` | `fitlme` | `randomEffects`

fpdf

F probability density function

Syntax

```
Y = fpdf(X,V1,V2)
```

Description

`Y = fpdf(X,V1,V2)` computes the *F* pdf at each of the values in `X` using the corresponding numerator degrees of freedom `V1` and denominator degrees of freedom `V2`. `X`, `V1`, and `V2` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. `V1` and `V2` parameters must contain real positive values, and the values in `X` must lie on the interval `[0 Inf]`.

The probability density function for the *F* distribution is

$$y = f(x \mid \nu_1, \nu_2) = \frac{\Gamma\left[\frac{(\nu_1 + \nu_2)}{2}\right]}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{x^{\frac{\nu_1 - 2}{2}}}{\left[1 + \left(\frac{\nu_1}{\nu_2}\right)x\right]^{\frac{\nu_1 + \nu_2}{2}}}$$

Examples

```
y = fpdf(1:6,2,2)
y =
  0.2500  0.1111  0.0625  0.0400  0.0278  0.0204

z = fpdf(3,5:10,5:10)
z =
  0.0689  0.0659  0.0620  0.0577  0.0532  0.0487
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

See Also

`fcdf` | `finv` | `frnd` | `fstat` | `pdf`

Topics

“F Distribution” on page B-45

Introduced before R2006a

fracfact

Fractional factorial design

Syntax

```
X = fracfact(gen)
[X,conf] = fracfact(gen)
[X,conf] = fracfact(gen,Name,Value)
```

Description

`X = fracfact(gen)` creates the two-level fractional factorial design defined by the generator `gen`.

`[X,conf] = fracfact(gen)` returns a cell array of character vectors containing the confounding pattern for the design.

`[X,conf] = fracfact(gen,Name,Value)` creates a fractional factorial designs with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`gen`

Either a string array or cell array of character vectors where each element contains one “word,” or a character array or string scalar consisting of “words” separated by spaces. “Words” consist of case-sensitive letters or groups of letters, where `'a'` represents value `1`, `'b'` represents value `2`, ..., `'A'` represents value `27`, ..., `'Z'` represents value `52`.

Each word defines how the corresponding factor’s levels are defined as products of generators from a `2^K` full-factorial design. `K` is the number of letters of the alphabet in `gen`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`FactorNames`

String array or cell array specifying the name for each factor.

Default: `{'X1','X2',...}`

`MaxInt`

Positive integer setting the maximum level of interaction to include in the confounding output.

Default: 2

Output Arguments

X

The two-level fractional factorial design. X is a matrix of size N-by-P, where

- N = 2^K, where K is the number of letters of the alphabet in gen.
- P is the number of words in gen.

Because X is a two-level design, the components of X are ±1. For the meaning of X, see “Fractional Factorial Designs” on page 28-5.

conf

Cell array of character vectors containing the confounding pattern for the design.

Examples

Generate a fractional factorial design for four variables, where the fourth variable is the product of the first three:

```
x = fracfact('a b c abc')

x =
    -1    -1    -1    -1
    -1    -1     1     1
    -1     1    -1     1
    -1     1     1    -1
     1    -1    -1     1
     1    -1     1    -1
     1     1    -1    -1
     1     1     1     1
```

Find generators for a six-factor design that uses four factors and achieves resolution IV using fracfactgen. Use the result to specify the design:

```
generators = fracfactgen('a b c d e f',4, ... % 4 factors
    4) % resolution 4

generators =
    'a'
    'b'
    'c'
    'd'
    'bcd'
    'acd'

x = fracfact(generators)

x =
    -1    -1    -1    -1    -1    -1
    -1    -1    -1     1     1     1
    -1    -1     1    -1     1     1
    -1    -1     1     1    -1    -1
    -1     1    -1    -1     1    -1
    -1     1    -1     1    -1     1
```

```
    -1     1     1    -1    -1     1
    -1     1     1     1     1    -1
     1    -1    -1    -1    -1     1
     1    -1    -1     1     1    -1
     1    -1     1    -1     1    -1
     1    -1     1     1    -1     1
     1     1    -1    -1     1     1
     1     1    -1     1    -1    -1
     1     1     1    -1    -1    -1
     1     1     1     1     1     1
```

References

[1] Box, G. E. P., W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. Hoboken, NJ: Wiley-Interscience, 1978.

See Also

`ff2n` | `fracfactgen` | `fullfact` | `hadamard`

Topics

“Fractional Factorial Designs” on page 28-5

Introduced before R2006a

fracfactgen

Fractional factorial design generators

Syntax

```
generators = fracfactgen(terms)
generators = fracfactgen(terms,k)
generators = fracfactgen(terms,k,R)
generators = fracfactgen(terms,k,R,basic)
```

Description

`generators = fracfactgen(terms)` uses the Franklin-Bailey algorithm to find generators for the smallest two-level fractional-factorial design for estimating linear model terms specified by `terms`. `terms` is a character vector or string scalar consisting of words formed from the `52` case-sensitive letters `a`-`Z`, separated by spaces. Use `'a'-'z'` for the first 26 factors, and, if necessary, `'A'-'Z'` for the remaining factors. For example, `terms = 'a b c ab ac'`. Single-letter words indicate main effects to be estimated; multiple-letter words indicate interactions. Alternatively, `terms` is an *m*-by-*n* matrix of `0`s and `1`s where *m* is the number of model terms to be estimated and *n* is the number of factors. For example, if `terms` contains rows `[0 1 0 0]` and `[1 0 0 1]`, then the factor `b` and the interaction between factors `a` and `d` are included in the model. `generators` is a cell array of character vectors with one generator per cell. Pass `generators` to `fracfact` to produce the fractional-factorial design and corresponding confounding pattern.

`generators = fracfactgen(terms,k)` returns generators for a two-level fractional-factorial design with 2^k-runs, if possible. If `k` is `[]`, `fracfactgen` finds the smallest design.

`generators = fracfactgen(terms,k,R)` finds a design with resolution `R`, if possible. The default resolution is `3`.

A design of resolution *R* is one in which no *n*-factor interaction is confounded with any other effect containing less than *R* – *n* factors. Thus a resolution III design does not confound main effects with one another but may confound them with two-way interactions, while a resolution IV design does not confound either main effects or two-way interactions but may confound two-way interactions with each other.

If `fracfactgen` is unable to find a design at the requested resolution, it tries to find a lower-resolution design sufficient to calibrate the model. If it is successful, it returns the generators for the lower-resolution design along with a warning. If it fails, it returns an error.

`generators = fracfactgen(terms,k,R,basic)` also accepts a vector `basic` specifying the indices of factors that are to be treated as basic. These factors receive full-factorial treatments in the design. The default includes factors that are part of the highest-order interaction in `terms`.

Examples

Suppose you wish to determine the effects of four two-level factors, for which there may be two-way interactions. A full-factorial design would require $2^4 = 16$ runs. The `fracfactgen` function finds generators for a resolution IV (separating main effects) fractional-factorial design that requires only $2^3 = 8$ runs:

```
generators = fracfactgen('a b c d',3,4)
generators =
    'a'
    'b'
    'c'
    'abc'
```

The more economical design and the corresponding confounding pattern are returned by `fracfact`:

```
[dfF,confounding] = fracfact(generators)
dfF =
    -1    -1    -1    -1
    -1    -1     1     1
    -1     1    -1     1
    -1     1     1    -1
     1    -1    -1     1
     1    -1     1    -1
     1     1    -1    -1
     1     1     1     1
confounding =
    'Term'     'Generator'    'Confounding'
    'X1'       'a'            'X1'
    'X2'       'b'            'X2'
    'X3'       'c'            'X3'
    'X4'       'abc'          'X4'
    'X1*X2'    'ab'           'X1*X2 + X3*X4'
    'X1*X3'    'ac'           'X1*X3 + X2*X4'
    'X1*X4'    'bc'           'X1*X4 + X2*X3'
    'X2*X3'    'bc'           'X1*X4 + X2*X3'
    'X2*X4'    'ac'           'X1*X3 + X2*X4'
    'X3*X4'    'ab'           'X1*X2 + X3*X4'
```

The confounding pattern shows, for example, that the two-way interaction between `X1` and `X2` is confounded by the two-way interaction between `X3` and `X4`.

References

[1] Box, G. E. P., W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. Hoboken, NJ: Wiley-Interscience, 1978.

See Also

`fracfact` | `hadamard`

Topics

“Fractional Factorial Designs” on page 28-5

Introduced in R2006a

friedman

Friedman’s test

Syntax

```
p = friedman(x,reps)
p = friedman(x,reps,displayopt)
[p,tbl] = friedman( ___ )
[p,tbl,stats] = friedman( ___ )
```

Description

`p = friedman(x,reps)` returns the *p*-value for the nonparametric Friedman's test to compare column effects in a two-way layout. `friedman` tests the null hypothesis that the column effects are all the same against the alternative that they are not all the same.

`p = friedman(x,reps,displayopt)` enables the ANOVA table display when `displayopt` is `'on'` (default) and suppresses the display when `displayopt` is `'off'`.

`[p,tbl] = friedman( ___ )` returns the ANOVA table (including column and row labels) in cell array `tbl`.

`[p,tbl,stats] = friedman( ___ )` also returns a structure `stats` that you can use to perform a follow-up multiple comparison test.

Examples

Test For Column Effects Using Friedman's Test

This example shows how to test for column effects in a two-way layout using Friedman's test.

Load the sample data.

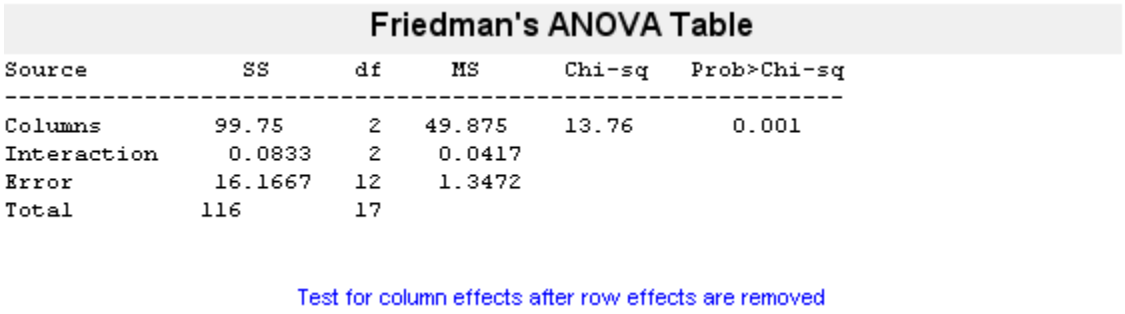
```
load popcorn
popcorn
```

```
popcorn = 6×3

    5.5000    4.5000    3.5000
    5.5000    4.5000    4.0000
    6.0000    4.0000    3.0000
    6.5000    5.0000    4.0000
    7.0000    5.5000    5.0000
    7.0000    5.0000    4.5000
```

This data comes from a study of popcorn brands and popper type (Hogg 1987). The columns of the matrix `popcorn` are brands (Gourmet, National, and Generic). The rows are popper type (Oil and Air). The study popped a batch of each brand three times with each popper. The values are the yield in cups of popped popcorn.

Use Friedman's test to determine whether the popcorn brand affects the yield of popcorn.

```
p = friedman(popcorn,3)
```




```
p = 0.0010
```

The small value of `p = 0.001` indicates the popcorn brand affects the yield of popcorn.

Input Arguments

x — Sample data

matrix

Sample data for the hypothesis test, specified as a matrix. The columns of `x` represent changes in a factor A. The rows represent changes in a blocking factor B. If there is more than one observation for each combination of factors, input reps indicates the number of replicates in each “cell,” which must be constant.

Data Types: `single` | `double`

reps — Number of replicates

1 (default) | positive integer value

Number of replicates for each combination of groups, specified as a positive integer value. For example, the following data has two replicates (`reps = 2`) for each group combination of row factor A and column factor B.

$$\begin{matrix} B=1 & B=2 \end{matrix}$$
$$\begin{bmatrix} x_{111} & x_{121} \\ x_{112} & x_{122} \\ x_{211} & x_{221} \\ x_{212} & x_{222} \\ x_{311} & x_{321} \\ x_{312} & x_{322} \end{bmatrix} \begin{matrix} \} A=1 \\ \} A=2 \\ \} A=3 \end{matrix}$$

Data Types: `single` | `double`

displayopt — ANOVA table display option

`'off'` (default) | `'on'`

ANOVA table display option, specified as `'off'` or `'on'`.

If `displayopt` is `'on'`, then `friedman` displays a figure showing an ANOVA table, which divides the variability of the ranks into two or three parts:

- The variability due to the differences among the column effects
- The variability due to the interaction between rows and columns (if reps is greater than its default value of 1)
- The remaining variability not explained by any systematic source

The ANOVA table has six columns:

- The first shows the source of the variability.
- The second shows the Sum of Squares (SS) due to each source.
- The third shows the degrees of freedom (df) associated with each source.
- The fourth shows the Mean Squares (MS), which is the ratio SS/df.
- The fifth shows Friedman's chi-square statistic.
- The sixth shows the *p* value for the chi-square statistic.

You can copy a text version of the ANOVA table to the clipboard by selecting `Copy Text` from the **Edit** menu.

Output Arguments

`p` — *p*-value

scalar value in the range `[0,1]`

p-value of the test, returned as a scalar value in the range `[0,1]`. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`tbl` — ANOVA table

cell array

ANOVA table, including column and row labels, returned as a cell array. The ANOVA table has six columns:

- The first shows the source of the variability.
- The second shows the Sum of Squares (SS) due to each source.
- The third shows the degrees of freedom (df) associated with each source.
- The fourth shows the Mean Squares (MS), which is the ratio SS/df.
- The fifth shows Friedman's chi-square statistic.
- The sixth shows the *p* value for the chi-square statistic.

You can copy a text version of the ANOVA table to the clipboard by selecting `Copy Text` from the **Edit** menu.

`stats` — Test data

structure

Test data, returned as a structure. `friedman` evaluates the hypothesis that the column effects are all the same against the alternative that they are not all the same. However, sometimes it is preferable to

perform a test to determine which pairs of column effects are significantly different, and which are not. You can use the `multcompare` function to perform such tests by supplying `stats` as the input value.

More About

Friedman's Test

Friedman's test is similar to classical balanced two-way ANOVA, but it tests only for column effects after adjusting for possible row effects. It does not test for row effects or interaction effects. Friedman's test is appropriate when columns represent treatments that are under study, and rows represent nuisance effects (blocks) that need to be taken into account but are not of any interest.

The different columns of `X` represent changes in a factor A. The different rows represent changes in a blocking factor B. If there is more than one observation for each combination of factors, input `reps` indicates the number of replicates in each "cell," which must be constant.

The matrix below illustrates the format for a set-up where column factor A has three levels, row factor B has two levels, and there are two replicates (`reps=2`). The subscripts indicate row, column, and replicate, respectively.

$$\begin{bmatrix} x_{111} & x_{121} & x_{131} \\ x_{112} & x_{122} & x_{132} \\ x_{211} & x_{221} & x_{231} \\ x_{212} & x_{222} & x_{232} \end{bmatrix}$$

Friedman's test assumes a model of the form

$$x_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where μ is an overall location parameter, α_i represents the column effect, β_j represents the row effect, and ε_{ijk} represents the error. This test ranks the data within each level of B, and tests for a difference across levels of A. The `p` that `friedman` returns is the p value for the null hypothesis that $\alpha_i = 0$. If the p value is near zero, this casts doubt on the null hypothesis. A sufficiently small p value suggests that at least one column-sample median is significantly different than the others; i.e., there is a main effect due to factor A. The choice of a critical p value to determine whether a result is "statistically significant" is left to the researcher. It is common to declare a result significant if the p value is less than 0.05 or 0.01.

Friedman's test makes the following assumptions about the data in `X`:

- All data come from populations having the same continuous distribution, apart from possibly different locations due to column and row effects.
- All observations are mutually independent.

The classical two-way ANOVA replaces the first assumption with the stronger assumption that data come from normal distributions.

References

[1] Hogg, R. V., and J. Ledolter. *Engineering Statistics*. New York: MacMillan, 1987.

[2] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

See Also

`anova2` | `kruskalwallis` | `multcompare`

Introduced before R2006a

frnd

F random numbers

Syntax

```
R = frnd(V1,V2)
R = frnd(V1,V2,m,n,...)
R = frnd(V1,V2,[m,n,...])
```

Description

`R = frnd(V1,V2)` generates random numbers from the *F* distribution with numerator degrees of freedom `V1` and denominator degrees of freedom `V2`. `V1` and `V2` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input for `V1` or `V2` is expanded to a constant array with the same dimensions as the other input. `V1` and `V2` parameters must contain real positive values.

`R = frnd(V1,V2,m,n,...)` or `R = frnd(V1,V2,[m,n,...])` generates an `m`-by-`n`-by-... array containing random numbers from the *F* distribution with parameters `V1` and `V2`. `V1` and `V2` can each be scalars or arrays of the same size as `R`.

Examples

```
n1 = frnd(1:6,1:6)
n1 =
  0.0022  0.3121  3.0528  0.3189  0.2715  0.9539

n2 = frnd(2,2,[2 3])
n2 =
  0.3186  0.9727  3.0268
  0.2052 148.5816  0.2191

n3 = frnd([1 2 3;4 5 6],1,2,3)
n3 =
  0.6233  0.2322  31.5458
  2.5848  0.2121  4.4955
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`fcdf` | `finv` | `fpdf` | `fstat` | `random`

Topics

“F Distribution” on page B-45

Introduced before R2006a

fscchi2

Univariate feature ranking for classification using chi-square tests

Syntax

```
idx = fscchi2(Tbl,ResponseVarName)
idx = fscchi2(Tbl,formula)
idx = fscchi2(Tbl,Y)

idx = fscchi2(X,Y)

idx = fscchi2(___,Name,Value)
[idx,scores] = fscchi2(___)
```

Description

`idx = fscchi2(Tbl,ResponseVarName)` ranks features (predictors) using chi-square tests on page 33-2471. The table `Tbl` contains predictor variables and a response variable, and `ResponseVarName` is the name of the response variable in `Tbl`. The function returns `idx`, which contains the indices of predictors ordered by predictor importance, meaning `idx(1)` is the index of the most important predictor. You can use `idx` to select important predictors for classification problems.

`idx = fscchi2(Tbl,formula)` specifies a response variable and predictor variables to consider among the variables in `Tbl` by using `formula`.

`idx = fscchi2(Tbl,Y)` ranks predictors in `Tbl` using the response variable `Y`.

`idx = fscchi2(X,Y)` ranks predictors in `X` using the response variable `Y`.

`idx = fscchi2(___,Name,Value)` specifies additional options using one or more name-value pair arguments in addition to any of the input argument combinations in the previous syntaxes. For example, you can specify prior probabilities and observation weights.

`[idx,scores] = fscchi2(___)` also returns the predictor scores `scores`. A large score value indicates that the corresponding predictor is important.

Examples

Rank Predictors in Matrix

Rank predictors in a numeric matrix and create a bar plot of predictor importance scores.

Load the sample data.

```
load ionosphere
```

`ionosphere` contains predictor variables (`X`) and a response variable (`Y`).

Rank the predictors using chi-square tests.

```
[idx,scores] = fscchi2(X,Y);
```

The values in `scores` are the negative logs of the *p*-values. If a *p*-value is smaller than `eps(0)`, then the corresponding score value is `Inf`. Before creating a bar plot, determine whether `scores` includes `Inf` values.

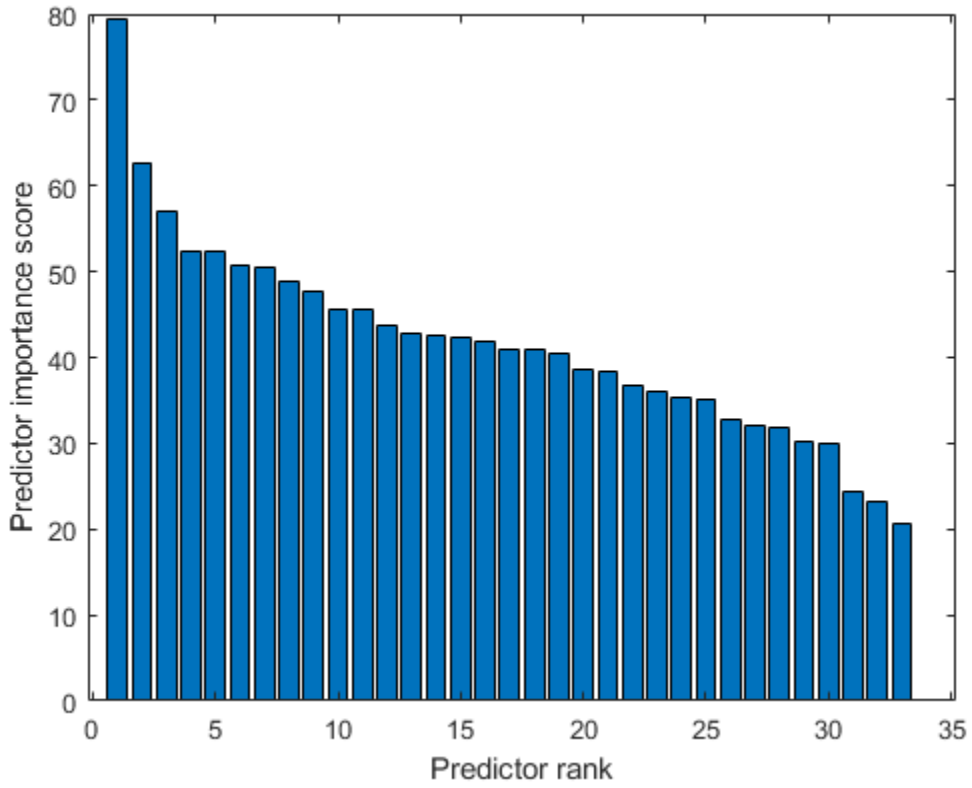
```
find(isinf(scores))
```

```
ans =

  1x0 empty double row vector
```

`scores` does not include `Inf` values. If `scores` includes `Inf` values, you can replace `Inf` by a large numeric number before creating a bar plot for visualization purposes. For details, see “Rank Predictors in Table” on page 33-2465.

Create a bar plot of the predictor importance scores.

```
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
```



Select the top five most important predictors. Find the columns of these predictors in X.

```
idx(1:5)
```

```
ans = 1×5
```

```
     5     7     3     8     6
```

The fifth column of X is the most important predictor of Y.

Rank Predictors in Table

Rank predictors in a table and create a bar plot of predictor importance scores.

If your data is in a table and `fscchi2` ranks a subset of the variables in the table, then the function indexes the variables using only the subset. Therefore, a good practice is to move the predictors that you do not want to rank to the end of the table. Move the response variable and observation weight vector as well. Then, the indexes of the output arguments are consistent with the indexes of the table.

Load the census1994 data set.

```
load census1994
```

The table `adultdata` in `census1994` contains demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. Display the first three rows of the table.

```
head(adultdata,3)
```

```
ans=3×15 table
    age       workClass          fnlwgt      education    education_num      marital_status
    ___    _________________    __________    _________    _____________    ____________________

    39     State-gov                77516    Bachelors          13         Never-married
    50     Self-emp-not-inc         83311    Bachelors          13         Married-civ-spouse
    38     Private             2.1565e+05    HS-grad             9         Divorced
```

In the table `adultdata`, the third column `fnlwgt` is the weight of the samples, and the last column `salary` is the response variable. Move `fnlwgt` to the left of `salary` by using the `movevars` function.

```
adultdata = movevars(adultdata,'fnlwgt','before','salary');
head(adultdata,3)
```

```
ans=3×15 table
    age       workClass          education    education_num      marital_status          occupation
    ___    _________________    _________    _____________    ____________________    _____________

    39     State-gov            Bachelors          13         Never-married          Adm-clerical
    50     Self-emp-not-inc     Bachelors          13         Married-civ-spouse     Exec-managerial
    38     Private              HS-grad             9         Divorced               Handlers-clean
```

Rank the predictors in `adultdata`. Specify the column `salary` as a response variable, and specify the column `fnlwgt` as observation weights.

```
[idx,scores] = fscchi2(adultdata,'salary','Weights','fnlwgt');
```

The values in `scores` are the negative logs of the *p*-values. If a *p*-value is smaller than `eps(0)`, then the corresponding score value is `Inf`. Before creating a bar plot, determine whether `scores` includes `Inf` values.

```
idxInf = find(isinf(scores))
```

```
idxInf = 1×8

     1     3     4     5     6     7    10    12
```

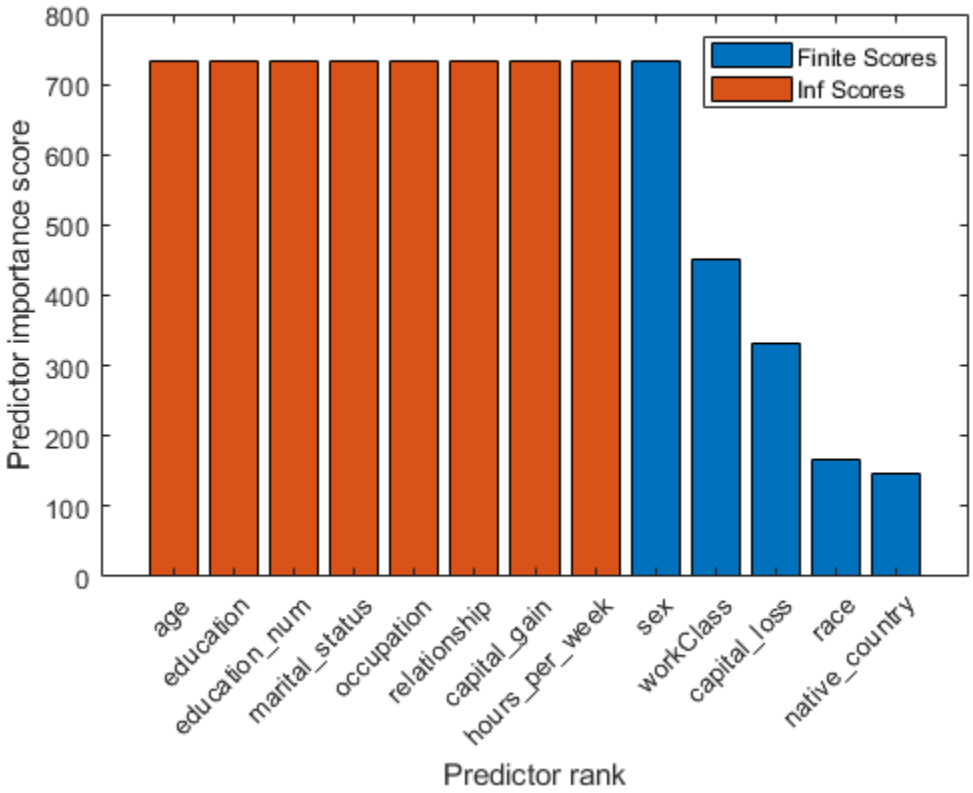
`scores` includes eight `Inf` values.

Create a bar plot of predictor importance scores. Use the predictor names for the *x*-axis tick labels.

```
figure
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
xticklabels(strrep(adultdata.Properties.VariableNames(idx),'_','\_'))
xtickangle(45)
```

The `bar` function does not plot any bars for the `Inf` values. For the `Inf` values, plot bars that have the same length as the largest finite score.

```
hold on
bar(scores(idx(length(idxInf)+1))*ones(length(idxInf),1))
legend('Finite Scores','Inf Scores')
hold off
```



The bar graph displays finite scores and Inf scores using different colors.

Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for a response variable and observation weights.

A response variable can be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element of the response variable must correspond to one row of the array.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`. If `Tbl` also contains the observation weights, then you can specify the weights by using `Weights`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify the subset of variables by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The response variable and `Tbl` must have the same number of rows.

If `fscchi2` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. The values in the `'CategoricalPredictors'` name-value pair argument and the output argument `idx` do not count the predictors that the function does not rank.

`fscchi2` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Tbl` for a response variable to be missing values. `fscchi2` does not use observations with missing values for a response variable.

Data Types: `table`

`ResponseVarName` — Response variable name

character vector or string scalar containing name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of a variable in `Tbl`.

For example, if a response variable is the column `Y` of `Tbl` (`Tbl.Y`), then specify `ResponseVarName` as `'Y'`.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y ~ x1 + x2 + x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors, use a formula. If you specify a formula, then `fscchi2` does not rank any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Y — Response variable
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Response variable, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors. Each row of `Y` represents the labels of the corresponding row of `X`.

`fscchi2` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. `fscchi2` does not use observations with missing values for `Y`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

X — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumBins',20,'UseMissing',true` sets the number of bins as 20 and specifies to use missing values in predictors for ranking.

`CategoricalPredictors` — List of categorical predictors
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

List of categorical predictors, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fscchi2` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.

Value	Description
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the names in `Tbl`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the names in `Tbl`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fscchi2` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fscchi2` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for ranking

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of the classes to use for ranking, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `ClassNames` must have the same data type as `Y` or the response variable in `Tbl`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `'ClassNames'` to:

- Specify the order of the `Prior` dimensions that corresponds to the class order.
- Select a subset of classes for ranking. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To rank predictors using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `'ClassNames'` is the set of all distinct class names in `Y` or the response variable in `Tbl`. The default `'ClassNames'` value has mathematical ordering if the response variable is ordinal. Otherwise, the default value has alphabetical ordering.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`NumBins` — Number of bins for binning continuous predictors

10 (default) | positive integer scalar

Number of bins for binning continuous predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

Example: `'NumBins',50`

Data Types: `single` | `double`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as one of the following:

- Character vector or string scalar.
  - `'empirical'` determines class probabilities from class frequencies in the response variable in `Y` or `Tbl`. If you pass observation weights, `fscchi2` uses the weights to compute the class probabilities.
  - `'uniform'` sets all class probabilities to be equal.
- Vector (one scalar value for each class). To specify the class order for the corresponding elements of `'Prior'`, set the `'ClassNames'` name-value argument.
- Structure `S` with two fields.
  - `S.ClassNames` contains the class names as a variable of the same type as the response variable in `Y` or `Tbl`.
  - `S.ClassProbs` contains a vector of corresponding probabilities.

`fscchi2` normalizes the weights in each class (`'Weights'`) to add up to the value of the prior probability of the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`UseMissing` — Indicator for whether to use or discard missing values in predictors
`false` (default) | `true`

Indicator for whether to use or discard missing values in predictors, specified as the comma-separated pair consisting of `'UseMissing'` and either `true` to use or `false` to discard missing values in predictors for ranking.

`fscchi2` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values to be missing values.

If you specify `'UseMissing',true`, then `fscchi2` uses missing values for ranking. For a categorical variable, `fscchi2` treats missing values as an extra category. For a continuous variable, `fscchi2` places `NaN` values in a separate bin for binning.

If you specify `'UseMissing',false`, then `fscchi2` does not use missing values for ranking. Because `fscchi2` computes importance scores individually for each predictor, the function does not discard an entire row when values in the row are partially missing. For each variable, `fscchi2` uses all values that are not missing.

Example: `'UseMissing',true`

Data Types: `logical`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The function weights the observations in each row of `X`

or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weight vector is the column `W` of `Tbl` (`Tbl.W`), then specify `'Weights','W'`.

`fscchi2` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`idx` — Indices of predictors ordered by predictor importance

numeric vector

Indices of predictors in `X` or `Tbl` ordered by predictor importance, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

If `fscchi2` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. For example, suppose `Tbl` includes 10 columns and you specify the last five columns of `Tbl` as the predictor variables by using `formula`. If `idx(3)` is `5`, then the third most important predictor is the 10th column in `Tbl`, which is the fifth predictor in the subset.

`scores` — Predictor scores

numeric vector

Predictor scores, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

A large score value indicates that the corresponding predictor is important.

- If you use `X` to specify the predictors or use all the variables in `Tbl` as predictors, then the values in `scores` have the same order as the predictors in `X` or `Tbl`.
- If you specify a subset of variables in `Tbl` as predictors, then the values in `scores` have the same order as the subset.

For example, suppose `Tbl` includes 10 columns and you specify the last five columns of `Tbl` as the predictor variables by using `formula`. Then, `score(3)` contains the score value of the 8th column in `Tbl`, which is the third predictor in the subset.

Algorithms

Univariate Feature Ranking Using Chi-Square Tests

- `fscchi2` examines whether each predictor variable is independent of a response variable by using individual chi-square tests. A small *p*-value of the test statistic indicates that the corresponding predictor variable is dependent on the response variable, and, therefore is an important feature.
- The output `scores` is –log(*p*). Therefore, a large score value indicates that the corresponding predictor is important. If a *p*-value is smaller than `eps(0)`, then the output is `Inf`.

- `fscchi2` examines a continuous variable after binning, or discretizing, the variable. You can specify the number of bins using the `'NumBins'` name-value pair argument.

See Also

`fscmrmr` | `fscnca` | `relieff` | `sequentialfs`

Topics

“Introduction to Feature Selection” on page 15-49

Introduced in R2020a

fscmrmr

Rank features for classification using minimum redundancy maximum relevance (MRMR) algorithm

Syntax

```
idx = fscmrmr(Tbl,ResponseVarName)
idx = fscmrmr(Tbl,formula)
idx = fscmrmr(Tbl,Y)

idx = fscmrmr(X,Y)

idx = fscmrmr( ___ ,Name,Value)
[idx,scores] = fscmrmr( ___ )
```

Description

`idx = fscmrmr(Tbl,ResponseVarName)` ranks features (predictors) using the MRMR algorithm on page 33-2482. The table `Tbl` contains predictor variables and a response variable, and `ResponseVarName` is the name of the response variable in `Tbl`. The function returns `idx`, which contains the indices of predictors ordered by predictor importance. You can use `idx` to select important predictors for classification problems.

`idx = fscmrmr(Tbl,formula)` specifies a response variable and predictor variables to consider among the variables in `Tbl` by using `formula`.

`idx = fscmrmr(Tbl,Y)` ranks predictors in `Tbl` using the response variable `Y`.

`idx = fscmrmr(X,Y)` ranks predictors in `X` using the response variable `Y`.

`idx = fscmrmr( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments in addition to any of the input argument combinations in the previous syntaxes. For example, you can specify prior probabilities and observation weights.

`[idx,scores] = fscmrmr( ___ )` also returns the predictor scores `scores`. A large score value indicates that the corresponding predictor is important.

Examples

Rank Predictors by Importance

Load the sample data.

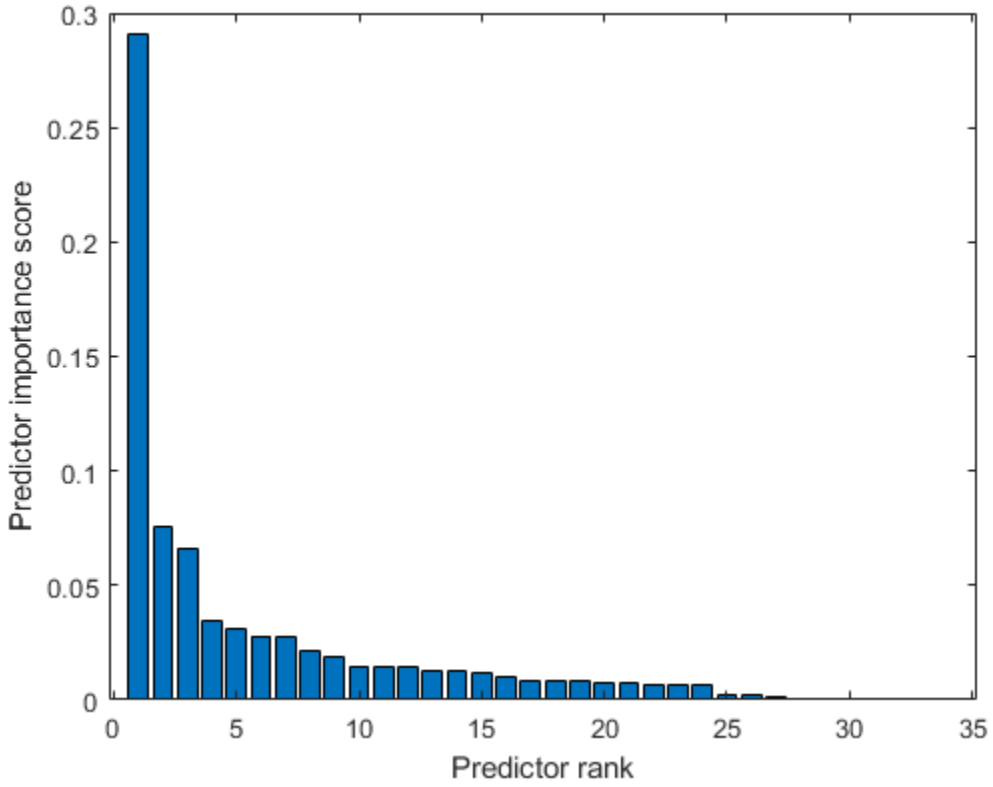
```
load ionosphere
```

Rank the predictors based on importance.

```
[idx,scores] = fscmrmr(X,Y);
```

Create a bar plot of the predictor importance scores.

```
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
```



The drop in score between the first and second most important predictors is large, while the drops after the sixth predictor are relatively small. A drop in the importance score represents the confidence of feature selection. Therefore, the large drop implies that the software is confident of selecting the most important predictor. The small drops indicate that the difference in predictor importance are not significant.

Select the top five most important predictors. Find the columns of these predictors in X.

```
idx(1:5)
```

```
ans = 1×5

     5     4     1     7    24
```

The fifth column of X is the most important predictor of Y.

Select Features and Compare Accuracies of Two Classification Models

Find important predictors by using `fscmrmr`. Then compare the accuracies of the full classification model (which uses all the predictors) and a reduced model that uses the five most important predictors by using `testckfold`.

Load the census1994 data set.

```
load census1994
```

The table `adultdata` in `census1994` contains demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. Display the first three rows of the table.

```
head(adultdata,3)
```

```
ans=3×15 table
    age       workClass          fnlwgt      education    education_num      marital_status
    ___    _________________    __________    _________    _____________    ____________________

    39     State-gov                77516    Bachelors         13          Never-married
    50     Self-emp-not-inc         83311    Bachelors         13          Married-civ-spouse
    38     Private              2.1565e+05    HS-grad            9          Divorced
```

The output arguments of `fscmrmr` include only the variables ranked by the function. Before passing a table to the function, move the variables that you do not want to rank, including the response variable and weight, to the end of the table so that the order of the output arguments is consistent with the order of the table.

In the table `adultdata`, the third column `fnlwgt` is the weight of the samples, and the last column `salary` is the response variable. Move `fnlwgt` to the left of `salary` by using the `movevars` function.

```
adultdata = movevars(adultdata,'fnlwgt','before','salary');
head(adultdata,3)
```

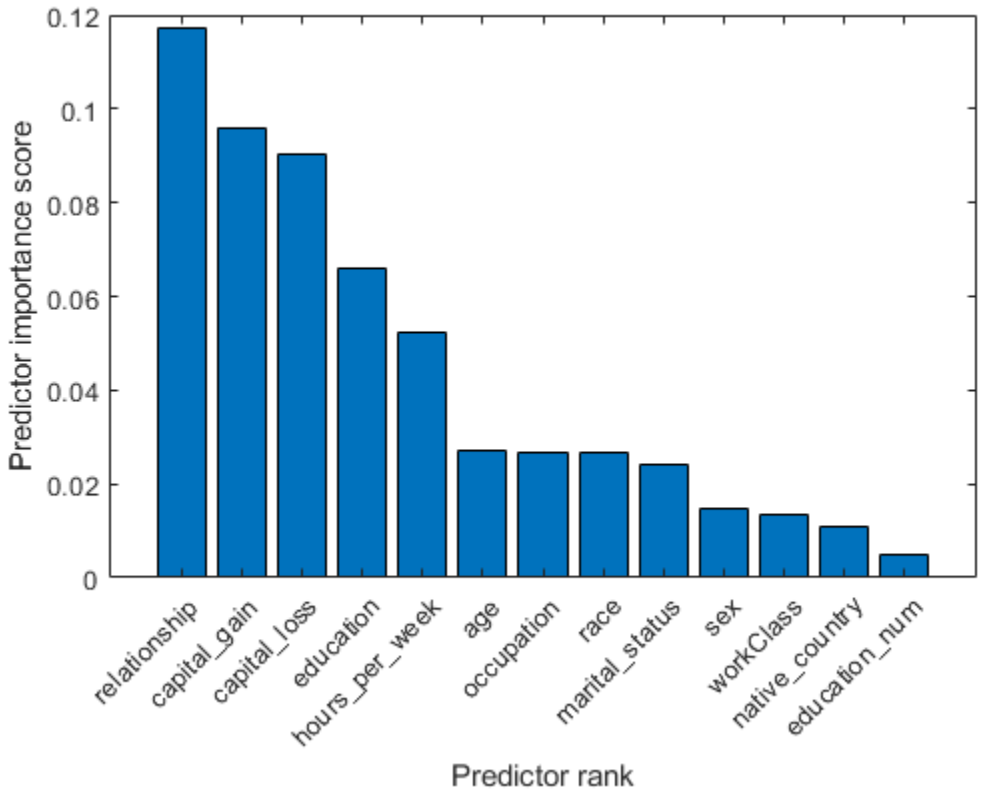
```
ans=3×15 table
    age       workClass          education    education_num      marital_status          occupation
    ___    _________________    _________    _____________    ____________________    _____________

    39     State-gov            Bachelors         13          Never-married          Adm-clerical
    50     Self-emp-not-inc     Bachelors         13          Married-civ-spouse     Exec-managerial
    38     Private              HS-grad            9          Divorced               Handlers-cleaners
```

Rank the predictors in `adultdata`. Specify the column `salary` as the response variable.

```
[idx,scores] = fscmrmr(adultdata,'salary','Weights','fnlwgt');
```

Create a bar plot of predictor importance scores. Use the predictor names for the *x*-axis tick labels.

```
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
xticklabels(strrep(adultdata.Properties.VariableNames(idx),'_','\_'))
xtickangle(45)
```



The five most important predictors are `relationship`, `capital_loss`, `capital_gain`, `education`, and `hours_per_week`.

Compare the accuracy of a classification tree trained with all predictors to the accuracy of one trained with the five most important predictors.

Create a classification tree template using the default options.

```
C = templateTree;
```

Define the table `tbl1` to contain all predictors and the table `tbl2` to contain the five most important predictors.

```
tbl1 = adultdata(:,adultdata.Properties.VariableNames(idx(1:13)));
tbl2 = adultdata(:,adultdata.Properties.VariableNames(idx(1:5)));
```

Pass the classification tree template and the two tables to the `testckfold` function. The function compares the accuracies of the two models by repeated cross-validation. Specify `'Alternative','greater'` to test the null hypothesis that the model with all predictors is, at most, as accurate as the model with the five predictors. The `'greater'` option is available when `'Test'` is `'5x2t'` (5-by-2 paired *t* test) or `'10x10t'` (10-by-10 repeated cross-validation *t* test).

```
[h,p] = testckfold(C,C,tbl1,tbl2,adultdata.salary,'Weights',adultdata.fnlwgt,'Alternative','grea
```

```
h = logical
   0
```

```
p = 0.9969
```

`h` equals 0 and the *p*-value is almost 1, indicating failure to reject the null hypothesis. Using the model with the five predictors does not result in loss of accuracy compared to the model with all the predictors.

Now train a classification tree using the selected predictors.

```
mdl = fitctree(adultdata,'salary ~ relationship + capital_loss + capital_gain + education +  hou
    'Weights',adultdata.fnlwgt)
```

```
mdl =
  ClassificationTree
           PredictorNames: {1x5 cell}
             ResponseName: 'salary'
    CategoricalPredictors: [1 2]
               ClassNames: [<=50K    >50K]
           ScoreTransform: 'none'
          NumObservations: 32561


  Properties, Methods
```

Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for a response variable and observation weights.

A response variable can be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element of the response variable must correspond to one row of the array.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`. If `Tbl` also contains the observation weights, then you can specify the weights by using `Weights`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify the subset of variables by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The response variable and `Tbl` must have the same number of rows.

If `fscmrmr` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. The values in the `'CategoricalPredictors'` name-value pair argument and the output argument `idx` do not count the predictors that the function does not rank.

`fscmrmr` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Tbl` for a response variable to be missing values. `fscmrmr` does not use observations with missing values for a response variable.

Data Types: `table`

`ResponseVarName` — Response variable name
character vector or string scalar containing name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of a variable in `Tbl`.

For example, if a response variable is the column `Y` of `Tbl` (`Tbl.Y`), then specify `ResponseVarName` as `'Y'`.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y ~ x1 + x2 + x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors, use a formula. If you specify a formula, then `fscmrmr` does not rank any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Response variable
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Response variable, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors. Each row of `Y` represents the labels of the corresponding row of `X`.

`fscmrmr` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. `fscmrmr` does not use observations with missing values for `Y`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'CategoricalPredictors',[1 2],'Verbose',2` specifies the first two predictor variables as categorical variables and specifies the verbosity level as 2.

`CategoricalPredictors` — List of categorical predictors

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

List of categorical predictors, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fscmrmr` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the names in `Tbl`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the names in `Tbl`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fscmrmr` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fscmrmr` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for ranking

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of the classes to use for ranking, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, a logical or numeric vector, or a cell

array of character vectors. `ClassNames` must have the same data type as `Y` or the response variable in `Tbl`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `'ClassNames'` to:

- Specify the order of the `Prior` dimensions that corresponds to the class order.
- Select a subset of classes for ranking. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To rank predictors using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `'ClassNames'` is the set of all distinct class names in `Y` or the response variable in `Tbl`. The default `'ClassNames'` value has mathematical ordering if the response variable is ordinal. Otherwise, the default value has alphabetical ordering.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Prior` — Prior probabilities
`'empirical'` (default) | `'uniform'` | vector of scalar values | structure

Prior probabilities for each class, specified as one of the following:

- Character vector or string scalar.
  - `'empirical'` determines class probabilities from class frequencies in the response variable in `Y` or `Tbl`. If you pass observation weights, `fscmrmr` uses the weights to compute the class probabilities.
  - `'uniform'` sets all class probabilities to be equal.
- Vector (one scalar value for each class). To specify the class order for the corresponding elements of `'Prior'`, set the `'ClassNames'` name-value argument.
- Structure `S` with two fields.
  - `S.ClassNames` contains the class names as a variable of the same type as the response variable in `Y` or `Tbl`.
  - `S.ClassProbs` contains a vector of corresponding probabilities.

`fscmrmr` normalizes the weights in each class (`'Weights'`) to add up to the value of the prior probability of the respective class.

Example: `'Prior','uniform'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`UseMissing` — Indicator for whether to use or discard missing values in predictors
`false` (default) | `true`

Indicator for whether to use or discard missing values in predictors, specified as the comma-separated pair consisting of `'UseMissing'` and either `true` to use or `false` to discard missing values in predictors for ranking.

`fscmrmr` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values to be missing values.

If you specify `'UseMissing',true`, then `fscmrmr` uses missing values for ranking. For a categorical variable, `fscmrmr` treats missing values as an extra category. For a continuous variable, `fscmrmr` places `NaN` values in a separate bin for binning.

If you specify `'UseMissing',false`, then `fscmrmr` does not use missing values for ranking. Because `fscmrmr` computes mutual information for each pair of variables, the function does not discard an entire row when values in the row are partially missing. `fscmrmr` uses all pair values that do not include missing values.

Example: `'UseMissing',true`

Data Types: `logical`

`Verbose` — Verbosity level
`0` (default) | nonnegative integer

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and a nonnegative integer. The value of `Verbose` controls the amount of diagnostic information that the software displays in the Command Window.

- 0 — `fscmrmr` does not display any diagnostic information.
- 1 — `fscmrmr` displays the elapsed times for computing “Mutual Information” on page 33-2482 and ranking predictors.
- ≥ 2 — `fscmrmr` displays the elapsed times and more messages related to computing mutual information. The amount of information increases as you increase the `'Verbose'` value.

Example: `'Verbose',1`

Data Types: `single` | `double`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The function weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weight vector is the column `W` of `Tbl` (`Tbl.W`), then specify `'Weights','W'`.

`fscmrmr` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`idx` — Indices of predictors ordered by predictor importance
numeric vector

Indices of predictors in `X` or `Tbl` ordered by predictor importance, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

If `fscmrmr` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. For example, suppose `Tbl` includes 10 columns and you specify the last five columns of `Tbl` as the predictor variables by using `formula`. If `idx(3)` is `5`, then the third most important predictor is the 10th column in `Tbl`, which is the fifth predictor in the subset.

`scores` — Predictor scores
numeric vector

Predictor scores, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

A large score value indicates that the corresponding predictor is important. Also, a drop in the feature importance score represents the confidence of feature selection. For example, if the software is confident of selecting a feature *x*, then the score value of the next most important feature is much smaller than the score value of *x*.

- If you use `X` to specify the predictors or use all the variables in `Tbl` as predictors, then the values in `scores` have the same order as the predictors in `X` or `Tbl`.
- If you specify a subset of variables in `Tbl` as predictors, then the values in `scores` have the same order as the subset.

For example, suppose `Tbl` includes 10 columns and you specify the last five columns of `Tbl` as the predictor variables by using `formula`. Then, `score(3)` contains the score value of the 8th column in `Tbl`, which is the third predictor in the subset.

More About

Mutual Information

The mutual information between two variables measures how much uncertainty of one variable can be reduced by knowing the other variable.

The mutual information I of the discrete random variables X and Z is defined as

$$I(X,Z) = \sum_{i,j} P\left(X = x_i, Z = z_j\right) \log \frac{P\left(X = x_i, Z = z_j\right)}{P(X = x_i) P\left(Z = z_j\right)}.$$

If X and Z are independent, then I equals 0. If X and Z are the same random variable, then I equals the entropy of X.

The `fscmrmr` function uses this definition to compute the mutual information values for both categorical (discrete) and continuous variables. `fscmrmr` discretizes a continuous variable into 256 bins or the number of unique values in the variable if it is less than 256. The function finds optimal bivariate bins for each pair of variables using the adaptive algorithm [2].

Algorithms

Minimum Redundancy Maximum Relevance (MRMR) Algorithm

The MRMR algorithm[1] finds an optimal set of features that is mutually and maximally dissimilar and can represent the response variable effectively. The algorithm minimizes the redundancy of a feature set and maximizing the relevance of a feature set to the response variable. The algorithm quantifies the redundancy and relevance using the mutual information of variables—pairwise mutual

information of features and mutual information of a feature and the response. You can use this algorithm for classification problems.

The goal of the MRMR algorithm is to find an optimal set S of features that maximizes V_S, the relevance of S with respect to a response variable y, and minimizes W_S, the redundancy of S, where V_S and W_S are defined with mutual information on page 33-2482 I:

$$V_S = \frac{1}{|S|}\sum_{x \in S} I(x, y),$$

$$W_S = \frac{1}{|S|^2}\sum_{x, z \in S} I(x, z)\,.$$

$|S|$ is the number of features in S.

Finding an optimal set S requires considering all $2^{|\Omega|}$ combinations, where Ω is the entire feature set. Instead, the MRMR algorithm ranks features through the forward addition scheme, which requires $O(|\Omega|\cdot|S|)$ computations, by using the mutual information quotient (MIQ) value.

$$\mathrm{MIQ}_x = \frac{V_x}{W_x},$$

where V_x and W_x are the relevance and redundancy of a feature, respectively:

$$V_x = I(x, y),$$

$$W_x = \frac{1}{|S|}\sum_{z \in S} I(x, z)\,.$$

The `fscmrmr` function ranks all features in Ω and returns `idx` (the indices of features ordered by feature importance) using the MRMR algorithm. Therefore, the computation cost becomes $O(|\Omega|^2)$. The function quantifies the importance of a feature using a heuristic algorithm and returns `score`. A large score value indicates that the corresponding predictor is important. Also, a drop in the feature importance score represents the confidence of feature selection. For example, if the software is confident of selecting a feature x, then the score value of the next most important feature is much smaller than the score value of x. You can use the outputs to find an optimal set S for a given number of features.

`fscmrmr` ranks features as follows:

1. Select the feature with the largest relevance, $\max_{x \in \Omega} V_x$. Add the selected feature to an empty set S.
2. Find the features with nonzero relevance and zero redundancy in the complement of S, S^c.
   - If S^c does not include a feature with nonzero relevance and zero redundancy, go to step 4.
   - Otherwise, select the feature with the largest relevance, $\max_{x \in S^C,\, W_X = 0} V_x$. Add the selected feature to the set S.
3. Repeat Step 2 until the redundancy is not zero for all features in S^c.
4. Select the feature that has the largest MIQ value with nonzero relevance and nonzero redundancy in S^c, and add the selected feature to the set S.

$$\max_{x \in S^c} \mathrm{MIQ}_x = \max_{x \in S^c} \frac{I(x, y)}{\frac{1}{|S|}\sum_{z \in S} I(x, z)}\,.$$

5 Repeat Step 4 until the relevance is zero for all features in S^c.

6 Add the features with zero relevance to S in random order.

The software can skip any step if it cannot find a feature that satisfies the conditions described in the step.

Compatibility Considerations

Specify `'UseMissing',true` to use missing values in predictors for ranking

Behavior changed in R2020a

Starting in R2020a, you can specify whether to use or discard missing values in predictors for ranking by using the `'UseMissing'` name-value pair argument. The default value of `'UseMissing'` is `false` because most classification training functions in Statistics and Machine Learning Toolbox do not use missing values for training.

In R2019b, `fscmrmr` used missing values in predictors by default. To update your code, specify `'UseMissing',true`.

References

[1] Ding, C., and H. Peng. "Minimum redundancy feature selection from microarray gene expression data." *Journal of Bioinformatics and Computational Biology*. Vol. 3, Number 2, 2005, pp. 185–205.

[2] Darbellay, G. A., and I. Vajda. "Estimation of the information by an adaptive partitioning of the observation space." *IEEE Transactions on Information Theory*. Vol. 45, Number 4, 1999, pp. 1315–1321.

See Also

`fscnca` | `fsulaplacian` | `relieff` | `sequentialfs`

Topics

“Introduction to Feature Selection” on page 15-49
“Sequential Feature Selection” on page 15-61

Introduced in R2019b

fscnca

Feature selection using neighborhood component analysis for classification

Syntax

```
mdl = fscnca(X,Y)
mdl = fscnca(X,Y,Name,Value)
```

Description

`mdl = fscnca(X,Y)` performs feature selection for classification using the predictors in `X` and responses in `Y`.

`fscnca` learns the feature weights by using a diagonal adaptation of neighborhood component analysis (NCA) with regularization.

`mdl = fscnca(X,Y,Name,Value)` performs feature selection for classification with additional options specified by one or more name-value pair arguments.

Examples

Detect Relevant Features in Data Using NCA for Classification

Generate toy data where the response variable depends on the 3rd, 9th, and 15th predictors.

```
rng(0,'twister'); % For reproducibility
N = 100;
X = rand(N,20);
y = -ones(N,1);
y(X(:,3).*X(:,9)./X(:,15) < 0.4) = 1;
```

Fit the neighborhood component analysis model for classification.

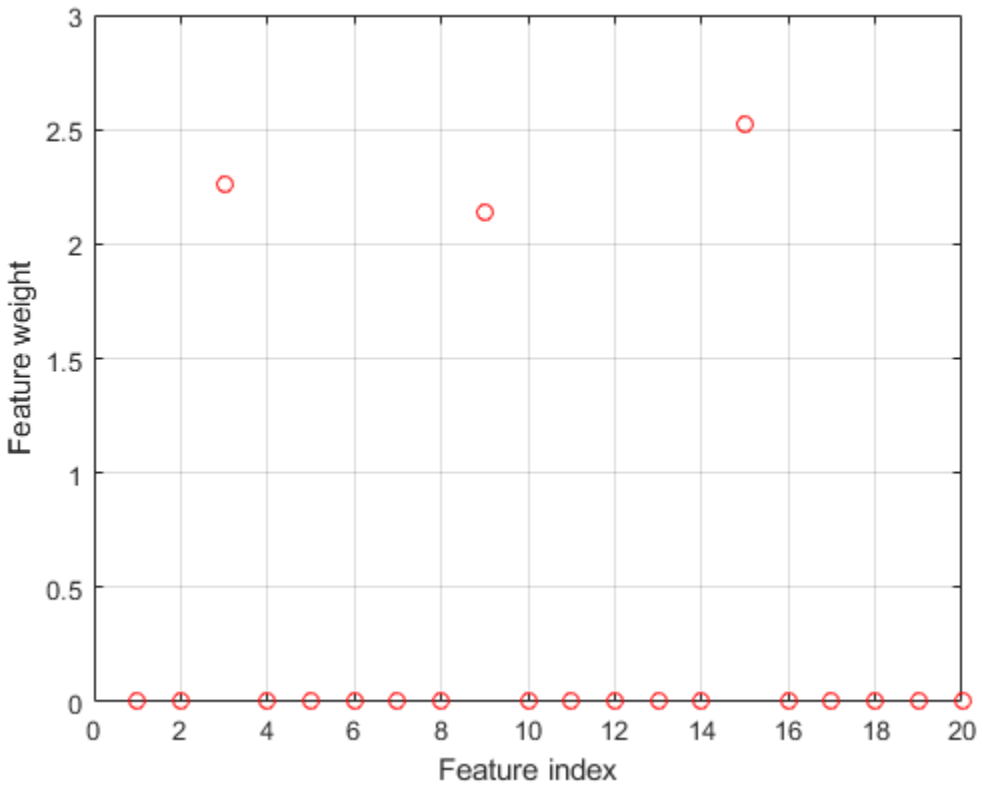
```
mdl = fscnca(X,y,'Solver','sgd','Verbose',1);
```

```
 o Tuning initial learning rate: NumTuningIterations = 20, TuningSubsetSize = 100

|=============================================|
|    TUNING    | TUNING SUBSET |    LEARNING    |
|     ITER     |   FUN VALUE   |      RATE      |
|=============================================|
|            1 | -3.755936e-01 |   2.000000e-01 |
|            2 | -3.950971e-01 |   4.000000e-01 |
|            3 | -4.311848e-01 |   8.000000e-01 |
|            4 | -4.903195e-01 |   1.600000e+00 |
|            5 | -5.630190e-01 |   3.200000e+00 |
|            6 | -6.166993e-01 |   6.400000e+00 |
|            7 | -6.255669e-01 |   1.280000e+01 |
|            8 | -6.255669e-01 |   1.280000e+01 |
|            9 | -6.255669e-01 |   1.280000e+01 |
|           10 | -6.255669e-01 |   1.280000e+01 |
```

```
|           11 | -6.255669e-01 |   1.280000e+01 |
|           12 | -6.255669e-01 |   1.280000e+01 |
|           13 | -6.255669e-01 |   1.280000e+01 |
|           14 | -6.279210e-01 |   2.560000e+01 |
|           15 | -6.279210e-01 |   2.560000e+01 |
|           16 | -6.279210e-01 |   2.560000e+01 |
|           17 | -6.279210e-01 |   2.560000e+01 |
|           18 | -6.279210e-01 |   2.560000e+01 |
|           19 | -6.279210e-01 |   2.560000e+01 |
|           20 | -6.279210e-01 |   2.560000e+01 |

 o Solver = SGD, MiniBatchSize = 10, PassLimit = 5

|=============================================================================================|
|   PASS   |     ITER     | AVG MINIBATCH | AVG MINIBATCH |   NORM STEP   |    LEARNING    |
|          |              |   FUN VALUE   |   NORM GRAD   |               |      RATE      |
|=============================================================================================|
|        0 |            9 | -5.658450e-01 |  4.492407e-02 |  9.290605e-01 |   2.560000e+01 |
|        1 |           19 | -6.131382e-01 |  4.923625e-02 |  7.421541e-01 |   1.280000e+01 |
|        2 |           29 | -6.225056e-01 |  3.738784e-02 |  3.277588e-01 |   8.533333e+00 |
|        3 |           39 | -6.233366e-01 |  4.947901e-02 |  5.431133e-01 |   6.400000e+00 |
|        4 |           49 | -6.238576e-01 |  3.445763e-02 |  2.946188e-01 |   5.120000e+00 |

         Two norm of the final step = 2.946e-01
Relative two norm of the final step = 6.588e-02, TolX = 1.000e-06
EXIT: Iteration or pass limit reached.
```

Plot the selected features. The weights of the irrelevant features should be close to zero.

```
figure()
plot(mdl.FeatureWeights,'ro')
grid on
xlabel('Feature index')
ylabel('Feature weight')
```



`fscnca` correctly detects the relevant features.

Identify Relevant Features for Classification

Load sample data

```
load ovariancancer;
whos
```

```
  Name        Size                Bytes  Class     Attributes

  grp       216x1                 25056  cell
  obs       216x4000            3456000  single
```

This example uses the high-resolution ovarian cancer data set that was generated using the WCX2 protein array. After some preprocessing steps, the data set has two variables: `obs` and `grp`. The `obs` variable consists 216 observations with 4000 features. Each element in `grp` defines the group to which the corresponding row of `obs` belongs.

Divide data into training and test sets

Use `cvpartition` to divide data into a training set of size 160 and a test set of size 56. Both the training set and the test set have roughly the same group proportions as in `grp`.

```
rng(1); % For reproducibility
cvp = cvpartition(grp,'holdout',56)

cvp =
Hold-out cross validation partition
   NumObservations: 216
       NumTestSets: 1
         TrainSize: 160
          TestSize: 56

Xtrain = obs(cvp.training,:);
ytrain = grp(cvp.training,:);
Xtest  = obs(cvp.test,:);
ytest  = grp(cvp.test,:);
```

Determine if feature selection is necessary

Compute generalization error without fitting.

```
nca = fscnca(Xtrain,ytrain,'FitMethod','none');
L = loss(nca,Xtest,ytest)

L = 0.0893
```

This option computes the generalization error of the neighborhood component analysis (NCA) feature selection model using the initial feature weights (in this case the default feature weights) provided in `fscnca`.

Fit NCA without regularization parameter (Lambda = 0)

```
nca = fscnca(Xtrain,ytrain,'FitMethod','exact','Lambda',0,...
      'Solver','sgd','Standardize',true);
L = loss(nca,Xtest,ytest)

L = 0.0714
```

The improvement on the loss value suggests that feature selection is a good idea. Tuning the λ value usually improves the results.

Tune the regularization parameter for NCA using five-fold cross-validation

Tuning λ means finding the λ value that produces the minimum classification loss. To tune λ using cross-validation:

1. Partition the training data into five folds and extract the number of validation (test) sets. For each fold, `cvpartition` assigns four-fifths of the data as a training set, and one-fifth of the data as a test set.

```
cvp = cvpartition(ytrain,'kfold',5);
numvalidsets = cvp.NumTestSets;
```

Assign λ values and create an array to store the loss function values.

```
n = length(ytrain);
lambdavals = linspace(0,20,20)/n;
lossvals = zeros(length(lambdavals),numvalidsets);
```

2. Train the NCA model for each λ value, using the training set in each fold.

3. Compute the classification loss for the corresponding test set in the fold using the NCA model. Record the loss value.

4. Repeat this process for all λ values.

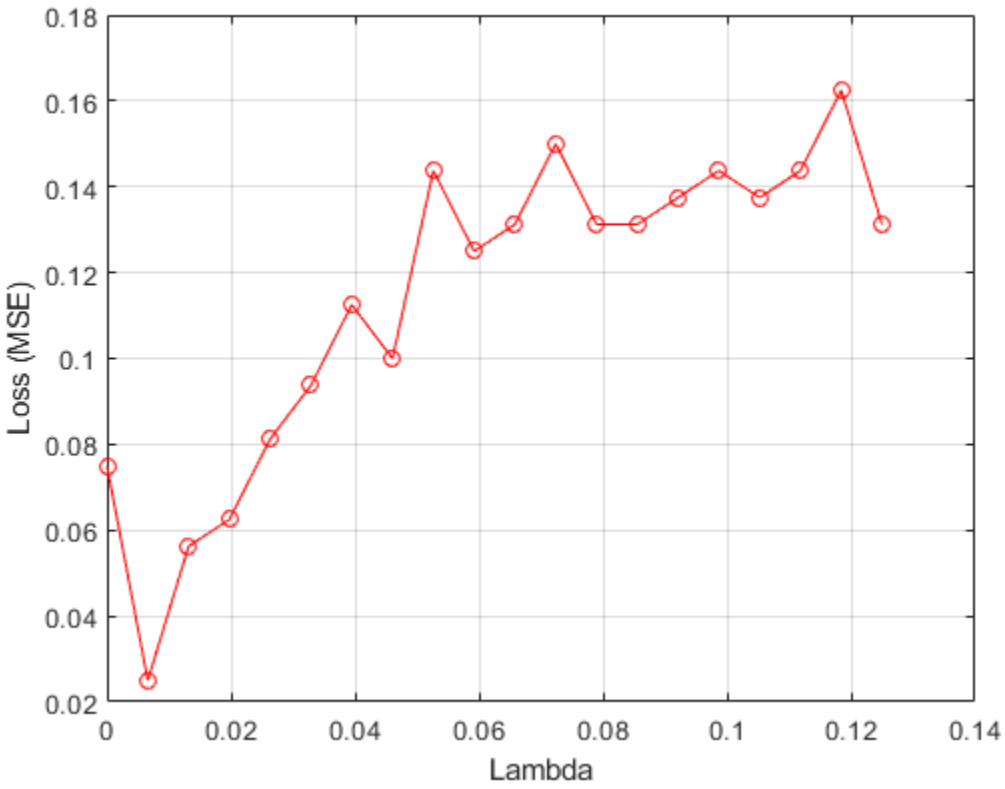
```
for i = 1:length(lambdavals)
    for k = 1:numvalidsets
        X = Xtrain(cvp.training(k),:);
        y = ytrain(cvp.training(k),:);
        Xvalid = Xtrain(cvp.test(k),:);
        yvalid = ytrain(cvp.test(k),:);

        nca = fscnca(X,y,'FitMethod','exact', ...
             'Solver','sgd','Lambda',lambdavals(i), ...
             'IterationLimit',30,'GradientTolerance',1e-4, ...
             'Standardize',true);

        lossvals(i,k) = loss(nca,Xvalid,yvalid,'LossFunction','classiferror');
    end
end
```

Compute the average loss obtained from the folds for each λ value.

```
meanloss = mean(lossvals,2);
```

Plot the average loss values versus the λ values.

```
figure()
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the best lambda value that corresponds to the minimum average loss.

```
[~,idx] = min(meanloss) % Find the index
```

```
idx = 2
```

```
bestlambda = lambdavals(idx) % Find the best lambda value
```

```
bestlambda = 0.0066
```

```
bestloss = meanloss(idx)
```

```
bestloss = 0.0250
```

Fit the nca model on all data using best λ and plot the feature weights

Use the solver lbfgs and standardize the predictor values.

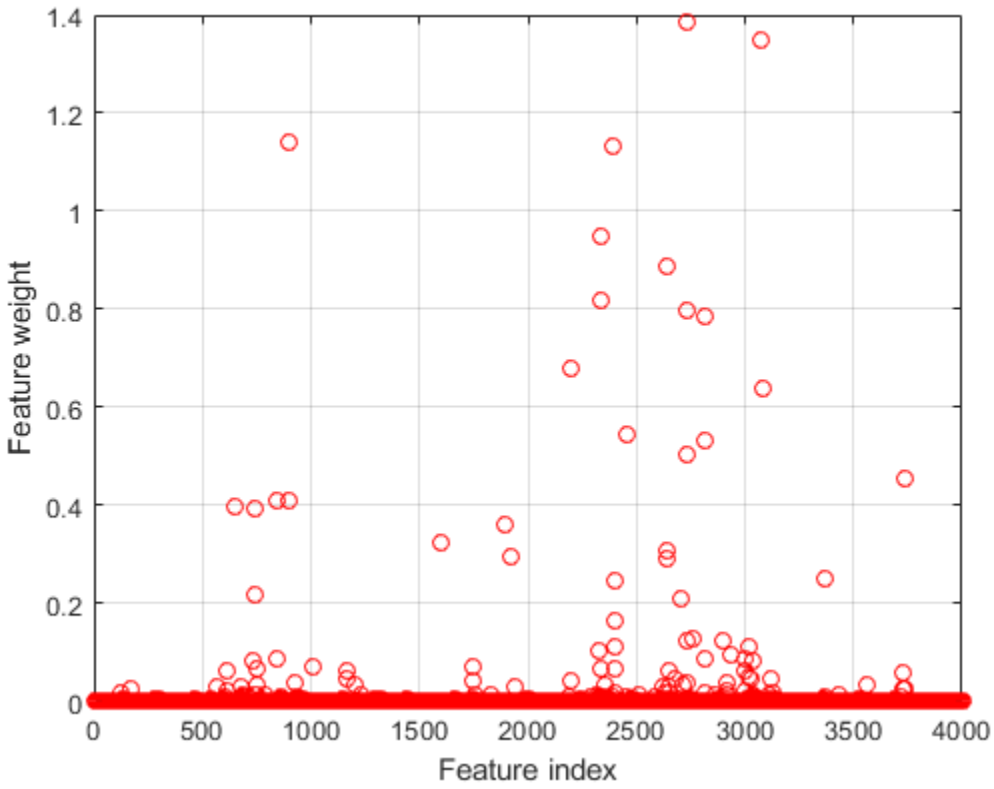
```
nca = fscnca(Xtrain,ytrain,'FitMethod','exact','Solver','sgd',...
    'Lambda',bestlambda,'Standardize',true,'Verbose',1);
```

```
 o Tuning initial learning rate: NumTuningIterations = 20, TuningSubsetSize = 100

|=============================================|
|    TUNING    | TUNING SUBSET |    LEARNING    |
|     ITER     |   FUN VALUE   |      RATE      |
|=============================================|
|            1 |  2.403497e+01 |   2.000000e-01 |
|            2 |  2.275050e+01 |   4.000000e-01 |
```

```
|            3 |  2.036845e+01 |   8.000000e-01 |
|            4 |  1.627647e+01 |   1.600000e+00 |
|            5 |  1.023512e+01 |   3.200000e+00 |
|            6 |  3.864283e+00 |   6.400000e+00 |
|            7 |  4.743816e-01 |   1.280000e+01 |
|            8 | -7.260138e-01 |   2.560000e+01 |
|            9 | -7.260138e-01 |   2.560000e+01 |
|           10 | -7.260138e-01 |   2.560000e+01 |
|           11 | -7.260138e-01 |   2.560000e+01 |
|           12 | -7.260138e-01 |   2.560000e+01 |
|           13 | -7.260138e-01 |   2.560000e+01 |
|           14 | -7.260138e-01 |   2.560000e+01 |
|           15 | -7.260138e-01 |   2.560000e+01 |
|           16 | -7.260138e-01 |   2.560000e+01 |
|           17 | -7.260138e-01 |   2.560000e+01 |
|           18 | -7.260138e-01 |   2.560000e+01 |
|           19 | -7.260138e-01 |   2.560000e+01 |
|           20 | -7.260138e-01 |   2.560000e+01 |

 o Solver = SGD, MiniBatchSize = 10, PassLimit = 5

|=========================================================================================|
|   PASS   |     ITER     | AVG MINIBATCH | AVG MINIBATCH |   NORM STEP   |    LEARNING    |
|          |              |   FUN VALUE   |   NORM GRAD   |               |      RATE      |
|=========================================================================================|
|        0 |            9 |  4.016078e+00 |  2.835465e-02 |  5.395984e+00 |   2.560000e+01 |
|        1 |           19 | -6.726156e-01 |  6.111354e-02 |  5.021138e-01 |   1.280000e+01 |
|        1 |           29 | -8.316555e-01 |  4.024185e-02 |  1.196030e+00 |   1.280000e+01 |
|        2 |           39 | -8.838656e-01 |  2.333418e-02 |  1.225839e-01 |   8.533333e+00 |
|        3 |           49 | -8.669035e-01 |  3.413150e-02 |  3.421881e-01 |   6.400000e+00 |
|        3 |           59 | -8.906935e-01 |  1.946293e-02 |  2.232510e-01 |   6.400000e+00 |
|        4 |           69 | -8.778630e-01 |  3.561283e-02 |  3.290643e-01 |   5.120000e+00 |
|        4 |           79 | -8.857136e-01 |  2.516633e-02 |  3.902977e-01 |   5.120000e+00 |

        Two norm of the final step = 3.903e-01
Relative two norm of the final step = 6.171e-03, TolX = 1.000e-06
EXIT: Iteration or pass limit reached.
```

Plot the feature weights.

```
figure()
plot(nca.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



Select features using the feature weights and a relative threshold.

```
tol    = 0.02;
selidx = find(nca.FeatureWeights > tol*max(1,max(nca.FeatureWeights)))
```

```
selidx = 72×1

   565
   611
   654
   681
   737
   743
   744
   750
   754
   839
    ⋮
```

Compute the classification loss using the test set.

```
L = loss(nca,Xtest,ytest)
```

```
L = 0.0179
```

Classify observations using the selected features

Extract the features with feature weights greater than 0 from the training data.

```
features = Xtrain(:,selidx);
```

Apply a support vector machine classifier using the selected features to the reduced training set.

```
svmMdl = fitcsvm(features,ytrain);
```

Evaluate the accuracy of the trained classifier on the test data which has not been used for selecting features.

```
L = loss(svmMdl,Xtest(:,selidx),ytest)
```

```
L = single
    0
```

Input Arguments

X — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

Y — Class labels

categorical vector | logical vector | numeric vector | string array | cell array of character vectors of length *n* | character matrix with *n* rows

Class labels, specified as a categorical vector, logical vector, numeric vector, string array, cell array of character vectors of length *n*, or character matrix with *n* rows, where *n* is the number of observations. Element *i* or row *i* of `Y` is the class label corresponding to row *i* of `X` (observation *i*).

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','sgd','Weights',W,'Lambda',0.0003` specifies the solver as the stochastic gradient descent, the observation weights as the values in the vector `W`, and sets the regularization parameter at 0.0003.

Fitting Options

`FitMethod` — Method for fitting the model

`'exact'` (default) | `'none'` | `'average'`

Method for fitting the model, specified as the comma-separated pair consisting of `'FitMethod'` and one of the following:

- `'exact'` — Performs fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to fscnca.

- `'average'` — Divides the data into partitions (subsets), fits each partition using the `exact` method, and returns the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

Example: `'FitMethod','none'`

`NumPartitions` — Number of partitions
`max(2,min(10,n))` (default) | integer between 2 and *n*

Number of partitions to split the data for using with `'FitMethod','average'` option, specified as the comma-separated pair consisting of `'NumPartitions'` and an integer value between 2 and *n*, where *n* is the number of observations.

Example: `'NumPartitions',15`

Data Types: `double` | `single`

`Lambda` — Regularization parameter
1/*n* (default) | nonnegative scalar

Regularization parameter to prevent overfitting, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

As the number of observations *n* increases, the chance of overfitting decreases and the required amount of regularization also decreases. See “Identify Relevant Features for Classification” on page 33-2487 and “Tune Regularization Parameter to Detect Features Using NCA for Classification” on page 15-210 to learn how to tune the regularization parameter.

Example: `'Lambda',0.002`

Data Types: `double` | `single`

`LengthScale` — Width of the kernel
1 (default) | positive real scalar

Width of the kernel, specified as the comma-separated pair consisting of `'LengthScale'` and a positive real scalar.

A length scale value of 1 is sensible when all predictors are on the same scale. If the predictors in `X` are of very different magnitudes, then consider standardizing the predictor values using `'Standardize',true` and setting `'LengthScale',1`.

Example: `'LengthScale',1.5`

Data Types: `double` | `single`

`InitialFeatureWeights` — Initial feature weights
`ones(p,1)` (default) | *p*-by-1 vector of real positive scalars

Initial feature weights, specified as the comma-separated pair consisting of `'InitialFeatureWeights'` and a *p*-by-1 vector of real positive scalars, where *p* is the number of predictors in the training data.

The regularized objective function for optimizing feature weights is nonconvex. As a result, using different initial feature weights can give different results. Setting all initial feature weights to 1 generally works well, but in some cases, random initialization using `rand(p,1)` can give better quality solutions.

Data Types: `double` | `single`

`Weights` — Observation weights
n-by-1 vector of 1s (default) | *n*-by-1 vector of real positive scalars

Observation weights, specified as the comma-separated pair consisting of `'ObservationWeights'` and an *n*-by-1 vector of real positive scalars. Use observation weights to specify higher importance of some observations compared to others. The default weights assign equal importance to all observations.

Data Types: `double` | `single`

`Prior` — Prior probabilities for each class
`'empirical'` (default) | `'uniform'` | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and one of the following:

- `'empirical'` — `fscnca` obtains the prior class probabilities from class frequencies.
- `'uniform'` — `fscnca` sets all class probabilities equal.
- Structure with two fields:
  - `ClassProbs` — Vector of class probabilities. If these are numeric values with a total greater than 1, `fsnca` normalizes them to add up to 1.
  - `ClassNames` — Class names corresponding to the class probabilities in `ClassProbs`.

Example: `'Prior','uniform'`

`Standardize` — Indicator for standardizing predictor data
`false` (default) | `true`

Indicator for standardizing the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and either `false` or `true`. For more information, see “Impact of Standardization” on page 15-102.

Example: `'Standardize',true`

Data Types: `logical`

`Verbose` — Verbosity level indicator
0 (default) | 1 | >1

Verbosity level indicator for the convergence summary display, specified as the comma-separated pair consisting of `'Verbose'` and one of the following:

- 0 — No convergence summary
- 1 — Convergence summary, including norm of gradient and objective function values
- > 1 — More convergence information, depending on the fitting algorithm

  When using `'minibatch-lbfgs'` solver and verbosity level > 1, the convergence information includes iteration the log from intermediate mini-batch LBFGS fits.

Example: `'Verbose',1`

Data Types: `double` | `single`

`Solver` — Solver type
`'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Solver type for estimating feature weights, specified as the comma-separated pair consisting of `'Solver'` and one of the following:

- `'lbfgs'` — Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm
- `'sgd'` — Stochastic gradient descent (SGD) algorithm
- `'minibatch-lbfgs'` — Stochastic gradient descent with LBFGS algorithm applied to mini-batches

Default is `'lbfgs'` for $n \le 1000$, and `'sgd'` for $n > 1000$.

Example: `'solver','minibatch-lbfgs'`

`LossFunction` — Loss function
`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFunction'` and one of the following.

- `'classiferror'` — Misclassification error

  $$l(y_i, y_j) = \begin{cases} 1 & \text{if } y_i \neq y_j, \\ 0 & \text{otherwise.} \end{cases}$$

- `@lossfun` — Custom loss function handle. A loss function has this form.

  ```
  function L = lossfun(Yu,Yv)
  % calculation of loss
  ...
  ```

  `Yu` is a u-by-1 vector and `Yv` is a v-by-1 vector. `L` is a u-by-v matrix of loss values such that `L(i,j)` is the loss value for `Yu(i)` and `Yv(j)`.

The objective function for minimization includes the loss function $l(y_i,y_j)$ as follows:

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} p_{ij} l(y_i, y_j) + \lambda \sum_{r=1}^{p} w_r^2,$$

where w is the feature weight vector, n is the number of observations, and p is the number of predictor variables. p_{ij} is the probability that x_j is the reference point for x_i. For details, see “NCA Feature Selection for Classification” on page 15-99.

Example: `'LossFunction',@lossfun`

`CacheSize` — Memory size
`1000MB` (default) | integer

Memory size, in MB, to use for objective function and gradient computation, specified as the comma-separated pair consisting of `'CacheSize'` and an integer.

Example: `'CacheSize',1500MB`

Data Types: `double` | `single`

LBFGS Options

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of history buffer for Hessian approximation for the `'lbfgs'` solver, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. At each iteration the function uses the most recent `HessianHistorySize` iterations to build an approximation to the inverse Hessian.

Example: `'HessianHistorySize',20`

Data Types: `double` | `single`

`InitialStepSize` — Initial step size
`'auto'` (default) | positive real scalar

Initial step size for the `'lbfgs'` solver, specified as the comma-separated pair consisting of `'InitialStepSize'` and a positive real scalar. By default, the function determines the initial step size automatically.

Data Types: `double` | `single`

`LineSearchMethod` — Line search method
`'weakwolfe'` (default) | `'strongwolfe'` | `'backtracking'`

Line search method, specified as the comma-separated pair consisting of `'LineSearchMethod'` and one of the following:

- `'weakwolfe'` — Weak Wolfe line search
- `'strongwolfe'` — Strong Wolfe line search
- `'backtracking'` — Backtracking line search

Example: `'LineSearchMethod','backtracking'`

`MaxLineSearchIterations` — Maximum number of line search iterations
`20` (default) | positive integer

Maximum number of line search iterations, specified as the comma-separated pair consisting of `'MaxLineSearchIterations'` and a positive integer.

Example: `'MaxLineSearchIterations',25`

Data Types: `double` | `single`

`GradientTolerance` — Relative convergence tolerance
`1e-6` (default) | positive real scalar

Relative convergence tolerance on the gradient norm for solver `lbfgs`, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive real scalar.

Example: `'GradientTolerance',0.000002`

Data Types: `double` | `single`

SGD Options

`InitialLearningRate` — Initial learning rate for `'sgd'` solver
`'auto'` (default) | positive real scalar

Initial learning rate for the `'sgd'` solver, specified as the comma-separated pair consisting of `'InitialLearningRate'` and a positive real scalar.

When using solver type `'sgd'`, the learning rate decays over iterations starting with the value specified for `'InitialLearningRate'`.

The default `'auto'` means that the initial learning rate is determined using experiments on small subsets of data. Use the `NumTuningIterations` name-value pair argument to specify the number of iterations for automatically tuning the initial learning rate. Use the `TuningSubsetSize` name-value pair argument to specify the number of observations to use for automatically tuning the initial learning rate.

For solver type `'minibatch-lbfgs'`, you can set `'InitialLearningRate'` to a very high value. In this case, the function applies LBFGS to each mini-batch separately with initial feature weights from the previous mini-batch.

To make sure the chosen initial learning rate decreases the objective value with each iteration, plot the `Iteration` versus the `Objective` values saved in the `mdl.FitInfo` property.

You can use the `refit` method with `'InitialFeatureWeights'` equal to `mdl.FeatureWeights` to start from the current solution and run additional iterations

Example: `'InitialLearningRate',0.9`

Data Types: `double` | `single`

`MiniBatchSize` — Number of observations to use in each batch for the `'sgd'` solver
min(10,*n*) (default) | positive integer value from 1 to *n*

Number of observations to use in each batch for the `'sgd'` solver, specified as the comma-separated pair consisting of `'MiniBatchSize'` and a positive integer from 1 to *n*.

Example: `'MiniBatchSize',25`

Data Types: `double` | `single`

`PassLimit` — Maximum number of passes for solver `'sgd'`
5 (default) | positive integer

Maximum number of passes through all *n* observations for solver `'sgd'`, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer. Each pass through all of the data is called an epoch.

Example: `'PassLimit',10`

Data Types: `double` | `single`

`NumPrint` — Frequency of batches for displaying convergence summary
10 (default) | positive integer value

Frequency of batches for displaying convergence summary for the `'sgd'` solver , specified as the comma-separated pair consisting of `'NumPrint'` and a positive integer. This argument applies when

the `'Verbose'` value is greater than 0. `NumPrint` mini-batches are processed for each line of the convergence summary that is displayed on the command line.

Example: `'NumPrint',5`

Data Types: `double` | `single`

`NumTuningIterations` — Number of tuning iterations
20 (default) | positive integer

Number of tuning iterations for the `'sgd'` solver, specified as the comma-separated pair consisting of `'NumTuningIterations'` and a positive integer. This option is valid only for `'InitialLearningRate','auto'`.

Example: `'NumTuningIterations',15`

Data Types: `double` | `single`

`TuningSubsetSize` — Number of observations to use for tuning initial learning rate
min(100,*n*) (default) | positive integer value from 1 to *n*

Number of observations to use for tuning the initial learning rate, specified as the comma-separated pair consisting of `'TuningSubsetSize'` and a positive integer value from 1 to *n*. This option is valid only for `'InitialLearningRate','auto'`.

Example: `'TuningSubsetSize',25`

Data Types: `double` | `single`

SGD or LBFGS Options

`IterationLimit` — Maximum number of iterations
positive integer

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer. The default is 10000 for SGD and 1000 for LBFGS and mini-batch LBFGS.

Each pass through a batch is an iteration. Each pass through all of the data is an epoch. If the data is divided into *k* mini-batches, then every epoch is equivalent to *k* iterations.

Example: `'IterationLimit',250`

Data Types: `double` | `single`

`StepTolerance` — Convergence tolerance on the step size
1e-6 (default) | positive real scalar

Convergence tolerance on the step size, specified as the comma-separated pair consisting of `'StepTolerance'` and a positive real scalar. The `'lbfgs'` solver uses an absolute step tolerance, and the `'sgd'` solver uses a relative step tolerance.

Example: `'StepTolerance',0.000005`

Data Types: `double` | `single`

Mini-Batch LBFGS Options

`MiniBatchLBFGSIterations` — Maximum number of iterations per mini-batch LBFGS step
10 (default) | positive integer

Maximum number of iterations per mini-batch LBFGS step, specified as the comma-separated pair consisting of `'MiniBatchLBFGSIterations'` and a positive integer.

Example: `'MiniBatchLBFGSIterations',15`

Mini-batch LBFGS algorithm is a combination of SGD and LBFGS methods. Therefore, all of the name-value pair arguments that apply to SGD and LBFGS solvers also apply to the mini-batch LBFGS algorithm.

Data Types: `double` | `single`

Output Arguments

`mdl` — Neighborhood component analysis model for classification

`FeatureSelectionNCAClassification` object

Neighborhood component analysis model for classification, returned as a `FeatureSelectionNCAClassification` object.

See Also

`FeatureSelectionNCAClassification` | `loss` | `predict` | `refit`

Topics

“Tune Regularization Parameter to Detect Features Using NCA for Classification” on page 15-210
“Neighborhood Component Analysis (NCA) Feature Selection” on page 15-99
“Introduction to Feature Selection” on page 15-49

Introduced in R2016b

fsrnca

Feature selection using neighborhood component analysis for regression

Syntax

```
mdl = fsrnca(X,Y)
mdl = fsrnca(X,Y,Name,Value)
```

Description

`mdl = fsrnca(X,Y)` performs feature selection for regression using the predictors in `X` and responses in `Y`.

`fsrnca` learns the feature weights by a diagonal adaptation of neighborhood component analysis (NCA) with regularization.

`mdl = fsrnca(X,Y,Name,Value)` performs feature selection for regression with additional options specified by one or more name-value pair arguments.

Examples

Detect Relevant Features in Data Using NCA for Regression

Generate toy data where the response variable depends on the 3rd, 9th, and 15th predictors.

```
rng(0,'twister'); % For reproducibility
N = 100;
X = rand(N,20);
y = 1 + X(:,3)*5 + sin(X(:,9)./X(:,15) + 0.25*randn(N,1));
```

Fit the neighborhood component analysis model for regression.

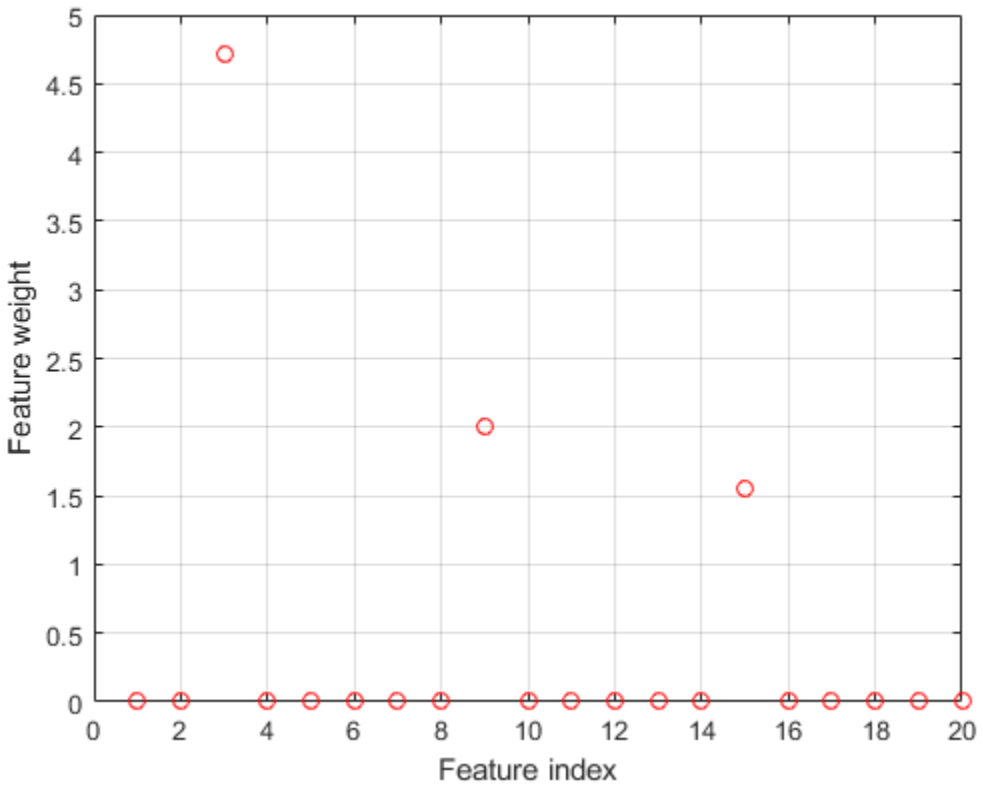
```
mdl = fsrnca(X,y,'Verbose',1,'Lambda',0.5/N);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 |  1.636932e+00 |   3.688e-01 |   0.000e+00 |        |   1.627e+00 |   0.000e+00 |
|        1 |  8.304833e-01 |   1.083e-01 |   2.449e+00 |    OK  |   9.194e+00 |   4.000e+00 |
|        2 |  7.548105e-01 |   1.341e-02 |   1.164e+00 |    OK  |   1.095e+01 |   1.000e+00 |
|        3 |  7.346997e-01 |   9.752e-03 |   6.383e-01 |    OK  |   2.979e+01 |   1.000e+00 |
|        4 |  7.053407e-01 |   1.605e-02 |   1.712e+00 |    OK  |   5.809e+01 |   1.000e+00 |
|        5 |  6.970502e-01 |   9.106e-03 |   8.818e-01 |    OK  |   6.223e+01 |   1.000e+00 |
|        6 |  6.952347e-01 |   5.522e-03 |   6.382e-01 |    OK  |   3.280e+01 |   1.000e+00 |
|        7 |  6.946302e-01 |   9.102e-04 |   1.952e-01 |    OK  |   3.380e+01 |   1.000e+00 |
|        8 |  6.945037e-01 |   6.557e-04 |   9.942e-02 |    OK  |   8.490e+01 |   1.000e+00 |
|        9 |  6.943908e-01 |   1.997e-04 |   1.756e-01 |    OK  |   1.124e+02 |   1.000e+00 |
|       10 |  6.943785e-01 |   3.478e-04 |   7.755e-02 |    OK  |   7.621e+01 |   1.000e+00 |
|       11 |  6.943728e-01 |   1.428e-04 |   3.416e-02 |    OK  |   3.649e+01 |   1.000e+00 |
```

```
|       12 |  6.943711e-01 |   1.128e-04 |   1.231e-02 |    OK  |   6.092e+01 |   1.000e+00 |   Y
|       13 |  6.943688e-01 |   1.066e-04 |   2.326e-02 |    OK  |   9.319e+01 |   1.000e+00 |   Y
|       14 |  6.943655e-01 |   9.324e-05 |   4.399e-02 |    OK  |   1.810e+02 |   1.000e+00 |   Y
|       15 |  6.943603e-01 |   1.206e-04 |   8.823e-02 |    OK  |   4.609e+02 |   1.000e+00 |   Y
|       16 |  6.943582e-01 |   1.701e-04 |   6.669e-02 |    OK  |   8.425e+01 |   5.000e-01 |   Y
|       17 |  6.943552e-01 |   5.160e-05 |   6.473e-02 |    OK  |   8.832e+01 |   1.000e+00 |   Y
|       18 |  6.943546e-01 |   2.477e-05 |   1.215e-02 |    OK  |   7.925e+01 |   1.000e+00 |   Y
|       19 |  6.943546e-01 |   1.077e-05 |   6.086e-03 |    OK  |   1.378e+02 |   1.000e+00 |   Y

|=============================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=============================================================================================================
|       20 |  6.943545e-01 |   2.260e-05 |   4.071e-03 |    OK  |   5.856e+01 |   1.000e+00 |   Y
|       21 |  6.943545e-01 |   4.250e-06 |   1.109e-03 |    OK  |   2.964e+01 |   1.000e+00 |   Y
|       22 |  6.943545e-01 |   1.916e-06 |   8.356e-04 |    OK  |   8.649e+01 |   1.000e+00 |   Y
|       23 |  6.943545e-01 |   1.083e-06 |   5.270e-04 |    OK  |   1.168e+02 |   1.000e+00 |   Y
|       24 |  6.943545e-01 |   1.791e-06 |   2.673e-04 |    OK  |   4.016e+01 |   1.000e+00 |   Y
|       25 |  6.943545e-01 |   2.596e-07 |   1.111e-04 |    OK  |   3.154e+01 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 2.596e-07
             Two norm of the final step     = 1.111e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 2.596e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the selected features. The weights of the irrelevant features should be close to zero.

```
figure()
plot(mdl.FeatureWeights,'ro')
grid on
xlabel('Feature index')
ylabel('Feature weight')
```



`fsrnca` correctly detects the relevant predictors for this response.

Tune Regularization Parameter in NCA for Regression

Load the sample data.

```
load robotarm.mat
```

The `robotarm` (pumadyn32nm) dataset is created using a robot arm simulator with 7168 training observations and 1024 test observations with 32 features [1][2]. This is a preprocessed version of the original data set. The data are preprocessed by subtracting off a linear regression fit, followed by normalization of all features to unit variance.

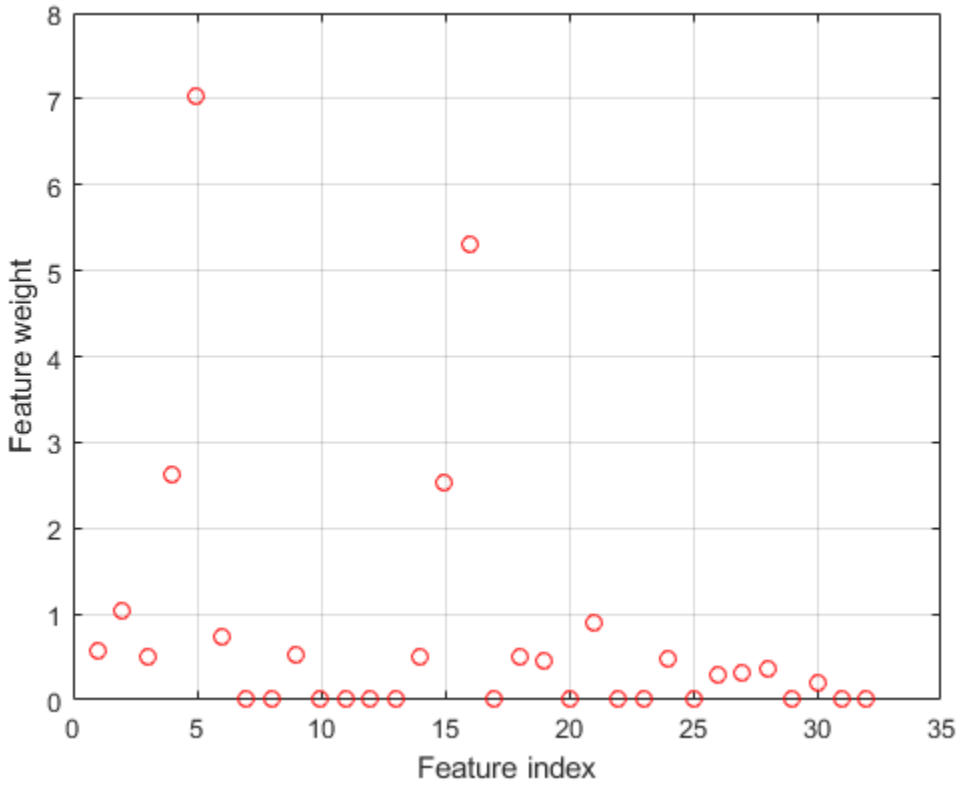
Perform neighborhood component analysis (NCA) feature selection for regression with the default λ (regularization parameter) value.

```
nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
    'Solver','lbfgs');
```

Plot the selected values.

```
figure
plot(nca.FeatureWeights,'ro')
xlabel('Feature index')
```

```
ylabel('Feature weight')
grid on
```



More than half of the feature weights are nonzero. Compute the loss using the test set as a measure of the performance by using the selected features.

```
L = loss(nca,Xtest,ytest)
```

```
L = 0.0837
```

Try improving the performance. Tune the regularization parameter λ for feature selection using five-fold cross-validation. Tuning λ means finding the λ value that produces the minimum regression loss. To tune λ using cross-validation:

1. Partition the data into five folds. For each fold, `cvpartition` assigns 4/5th of the data as a training set, and 1/5th of the data as a test set.

```
rng(1) % For reproducibility
n = length(ytrain);
cvp = cvpartition(length(ytrain),'kfold',5);
numvalidsets = cvp.NumTestSets;
```

Assign the λ values for the search. Multiplying response values by a constant increases the loss function term by a factor of the constant. Therefore, including the `std(ytrain)` factor in the λ values balances the default loss function (`'mad'`, mean absolute deviation) term and the regularization term in the objective function. In this example, the `std(ytrain)` factor is one because the loaded sample data is a preprocessed version of the original data set.

```
lambdavals = linspace(0,50,20)*std(ytrain)/n;
```

Create an array to store the loss values.

```
lossvals = zeros(length(lambdavals),numvalidsets);
```

2. Train the NCA model for each λ value, using the training set in each fold.

3. Compute the regression loss for the corresponding test set in the fold using the NCA model. Record the loss value.

4. Repeat this for each λ value and each fold.

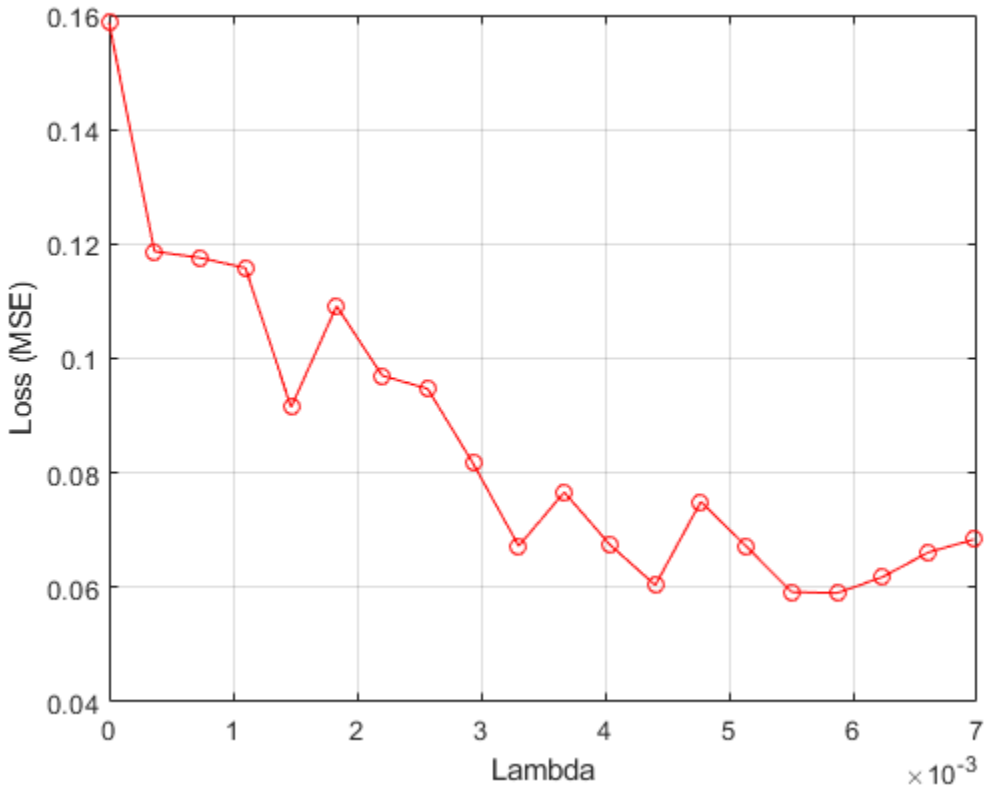
```
for i = 1:length(lambdavals)
    for k = 1:numvalidsets
        X = Xtrain(cvp.training(k),:);
        y = ytrain(cvp.training(k),:);
        Xvalid = Xtrain(cvp.test(k),:);
        yvalid = ytrain(cvp.test(k),:);

        nca = fsrnca(X,y,'FitMethod','exact', ...
             'Solver','minibatch-lbfgs','Lambda',lambdavals(i), ...
             'GradientTolerance',1e-4,'IterationLimit',30);

        lossvals(i,k) = loss(nca,Xvalid,yvalid,'LossFunction','mse');
    end
end
```

Compute the average loss obtained from the folds for each λ value.

```
meanloss = mean(lossvals,2);
```

Plot the mean loss versus the λ values.

```
figure
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the λ value that gives the minimum loss value.

```
[~,idx] = min(meanloss)
```

```
idx = 17
```

```
bestlambda = lambdavals(idx)
```

```
bestlambda = 0.0059
```

```
bestloss = meanloss(idx)
```

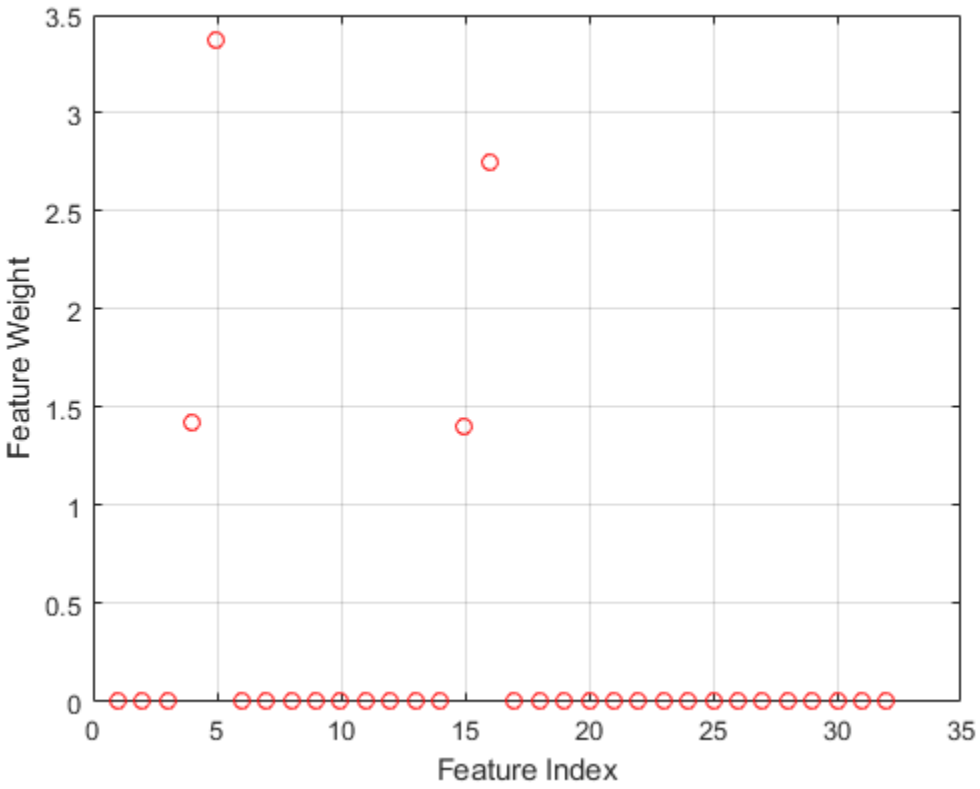
```
bestloss = 0.0590
```

Fit the NCA feature selection model for regression using the best λ value.

```
nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
    'Solver','lbfgs','Lambda',bestlambda);
```

Plot the selected features.

```
figure
plot(nca.FeatureWeights,'ro')
xlabel('Feature Index')
ylabel('Feature Weight')
grid on
```



Most of the feature weights are zero. `fsrnca` identifies the four most relevant features.

Compute the loss for the test set.

```
L = loss(nca,Xtest,ytest)
```

```
L = 0.0571
```

Tuning the regularization parameter, λ, eliminated more of the irrelevant features and improved the performance.

Compare NCA and ARD Feature Selection

This example uses the Abalone data [3][4] from the UCI Machine Learning Repository [5]. Download the data and save it in your current folder with the name `'abalone.data'`.

Store the data into a table. Display the first seven rows.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
tbl.Properties.VariableNames = {'Sex','Length','Diameter','Height', ...
    'WWeight','SWeight','VWeight','ShWeight','NoShellRings'};
tbl(1:7,:)
```

```
ans=7×9 table
     Sex      Length     Diameter     Height     WWeight     SWeight     VWeight     ShWeight     NoShel
```

```
_____    ______    ________    ______    _______    _______    _______    ________    _____

{'M'}    0.455      0.365      0.095      0.514      0.2245      0.101       0.15        15
{'M'}     0.35      0.265       0.09     0.2255      0.0995     0.0485       0.07         7
{'F'}     0.53       0.42      0.135      0.677      0.2565     0.1415       0.21         9
{'M'}     0.44      0.365      0.125      0.516      0.2155      0.114      0.155        10
{'I'}     0.33      0.255       0.08      0.205      0.0895     0.0395      0.055         7
{'I'}    0.425        0.3      0.095     0.3515       0.141     0.0775       0.12         8
{'F'}     0.53      0.415       0.15     0.7775       0.237     0.1415       0.33        20
```

The dataset has 4177 observations. The goal is to predict the age of abalone from eight physical measurements. The last variable, the number of shell rings, shows the age of the abalone. The first predictor is a categorical variable. The last variable in the table is the response variable.

Prepare the predictor and response variables for `fsrnca`. The last column of `tbl` contains the number of shell rings, which is the response variable. The first predictor variable, sex, is categorical. You must create dummy variables.

```
y = table2array(tbl(:,end));
X(:,1:3) = dummyvar(categorical(tbl.Sex));
X = [X,table2array(tbl(:,2:end-1))];
```

Use four-fold cross-validation to tune the regularization parameter in the NCA model. First partition the data into four folds.

```
rng('default') % For reproducibility
n = length(y);
cvp = cvpartition(n,'kfold',4);
numtestsets = cvp.NumTestSets;
```

`cvpartition` divides the data into four partitions (folds). In each fold, about three-fourths of the data is assigned as a training set and one-fourth is assigned as a test set.

Generate a variety of λ (regularization parameter) values for fitting the model to determine the best λ value. Create a vector to collect the loss values from each fit.

```
lambdavals = linspace(0,25,20)*std(y)/n;
lossvals = zeros(length(lambdavals),numtestsets);
```

The rows of `lossvals` corresponds to the λ values and the columns correspond to the folds.

Fit the NCA model for regression using `fsrnca` to the data from each fold using each λ value. Compute the loss for each model using the test data from each fold.

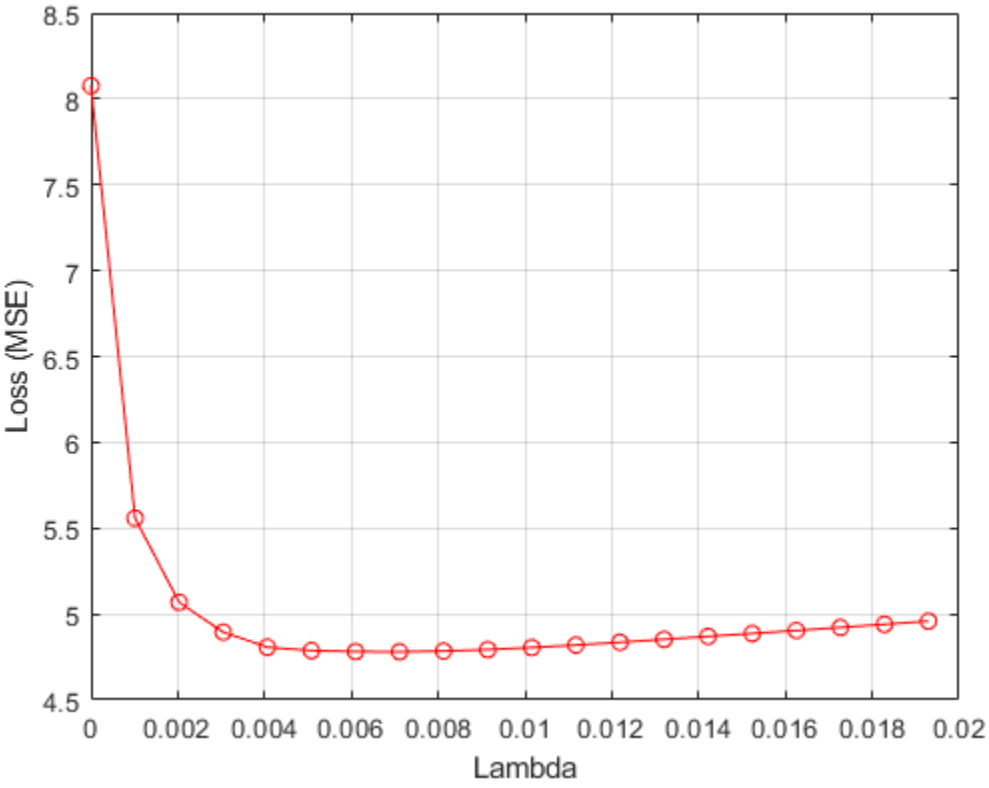
```
for i = 1:length(lambdavals)
   for k = 1:numtestsets
       Xtrain = X(cvp.training(k),:);
       ytrain = y(cvp.training(k),:);
       Xtest = X(cvp.test(k),:);
       ytest = y(cvp.test(k),:);

       nca = fsrnca(Xtrain,ytrain,'FitMethod','exact', ...
                 'Solver','lbfgs','Lambda',lambdavals(i),'Standardize',true);

       lossvals(i,k) = loss(nca,Xtest,ytest,'LossFunction','mse');
    end
end
```

Compute the average loss for the folds, that is, compute the mean in the second dimension of `lossvals`.

```
meanloss = mean(lossvals,2);
```

Plot the λ values versus the mean loss from the four folds.

```
figure
plot(lambdavals,meanloss,'ro-')
xlabel('Lambda')
ylabel('Loss (MSE)')
grid on
```



Find the λ value that minimizes the mean loss.

```
[~,idx] = min(meanloss);
bestlambda = lambdavals(idx)
```

```
bestlambda = 0.0071
```

Compute the best loss value.

```
bestloss = meanloss(idx)
```

```
bestloss = 4.7799
```

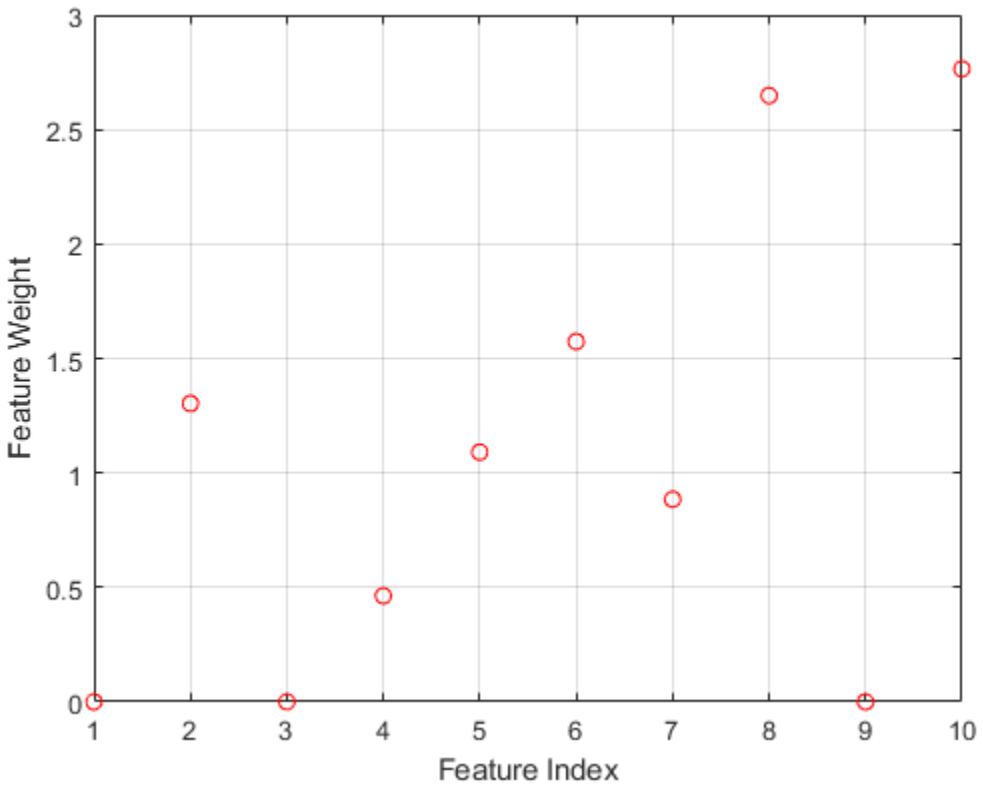
Fit the NCA model on all of the data using the best λ value.

```
nca = fsrnca(X,y,'FitMethod','exact','Solver','lbfgs', ...
    'Verbose',1,'Lambda',bestlambda,'Standardize',true);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=====================================================================================================
|        0 |  2.469168e+00 |   1.266e-01 |   0.000e+00 |        |   4.741e+00 |   0.000e+00 |
|        1 |  2.375166e+00 |   8.265e-02 |   7.268e-01 |    OK  |   1.054e+01 |   1.000e+00 |
|        2 |  2.293528e+00 |   2.067e-02 |   2.034e+00 |    OK  |   1.569e+01 |   1.000e+00 |
|        3 |  2.286703e+00 |   1.031e-02 |   3.158e-01 |    OK  |   2.213e+01 |   1.000e+00 |
|        4 |  2.279928e+00 |   2.023e-02 |   9.374e-01 |    OK  |   1.953e+01 |   1.000e+00 |
|        5 |  2.276258e+00 |   6.884e-03 |   2.497e-01 |    OK  |   1.439e+01 |   1.000e+00 |
|        6 |  2.274358e+00 |   1.792e-03 |   4.010e-01 |    OK  |   3.109e+01 |   1.000e+00 |
|        7 |  2.274105e+00 |   2.412e-03 |   2.399e-01 |    OK  |   3.557e+01 |   1.000e+00 |
|        8 |  2.274073e+00 |   1.459e-03 |   7.684e-02 |    OK  |   1.356e+01 |   1.000e+00 |
|        9 |  2.274050e+00 |   3.733e-04 |   3.797e-02 |    OK  |   1.725e+01 |   1.000e+00 |
|       10 |  2.274043e+00 |   2.750e-04 |   1.379e-02 |    OK  |   2.445e+01 |   1.000e+00 |
|       11 |  2.274027e+00 |   2.682e-04 |   5.701e-02 |    OK  |   7.386e+01 |   1.000e+00 |
|       12 |  2.274020e+00 |   1.712e-04 |   4.107e-02 |    OK  |   9.461e+01 |   1.000e+00 |
|       13 |  2.274014e+00 |   2.633e-04 |   6.720e-02 |    OK  |   7.469e+01 |   1.000e+00 |
|       14 |  2.274012e+00 |   9.818e-05 |   2.263e-02 |    OK  |   3.275e+01 |   1.000e+00 |
|       15 |  2.274012e+00 |   4.220e-05 |   6.188e-03 |    OK  |   2.799e+01 |   1.000e+00 |
|       16 |  2.274012e+00 |   2.859e-05 |   4.979e-03 |    OK  |   6.628e+01 |   1.000e+00 |
|       17 |  2.274011e+00 |   1.582e-05 |   6.767e-03 |    OK  |   1.439e+02 |   1.000e+00 |
|       18 |  2.274011e+00 |   7.623e-06 |   4.311e-03 |    OK  |   1.211e+02 |   1.000e+00 |
|       19 |  2.274011e+00 |   3.038e-06 |   2.528e-04 |    OK  |   1.798e+01 |   5.000e-01 |

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=====================================================================================================
|       20 |  2.274011e+00 |   6.710e-07 |   2.325e-04 |    OK  |   2.721e+01 |   1.000e+00 |

         Infinity norm of the final gradient = 6.710e-07
             Two norm of the final step     = 2.325e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 6.710e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the selected features.

```
figure
plot(nca.FeatureWeights,'ro')
xlabel('Feature Index')
ylabel('Feature Weight')
grid on
```



The irrelevant features have zero weights. According to this figure, the features 1, 3, and 9 are not selected.

Fit a Gaussian process regression (GPR) model using the subset of regressors method for parameter estimation and the fully independent conditional method for prediction. Use the ARD squared exponential kernel function, which assigns an individual weight to each predictor. Standardize the predictors.

```
gprMdl = fitrgp(tbl,'NoShellRings','KernelFunction','ardsquaredexponential', ...
      'FitMethod','sr','PredictMethod','fic','Standardize',true)
```

```
gprMdl =
  RegressionGP
           PredictorNames: {'Sex'  'Length'  'Diameter'  'Height'  'WWeight'  'SWeight'  'VWeigh
             ResponseName: 'NoShellRings'
    CategoricalPredictors: 1
         ResponseTransform: 'none'
           NumObservations: 4177
            KernelFunction: 'ARDSquaredExponential'
         KernelInformation: [1×1 struct]
             BasisFunction: 'Constant'
                      Beta: 11.4959
                     Sigma: 2.0282
         PredictorLocation: [10×1 double]
            PredictorScale: [10×1 double]
                     Alpha: [1000×1 double]
          ActiveSetVectors: [1000×10 double]
```

```
            PredictMethod: 'FIC'
            ActiveSetSize: 1000
                FitMethod: 'SR'
          ActiveSetMethod: 'Random'
        IsActiveSetVector: [4177×1 logical]
            LogLikelihood: -9.0019e+03
         ActiveSetHistory: [1×1 struct]
           BCDInformation: []


  Properties, Methods
```

Compute the regression loss on the training data (resubstitution loss) for the trained model.

```
L = resubLoss(gprMdl)
```

```
L = 4.0306
```

The smallest cross-validated loss using `fsrnca` is comparable to the loss obtained using a GPR model with an ARD kernel.

Input Arguments

X — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

Y — Response values

numeric real vector of length *n*

Response values, specified as a numeric real vector of length *n*, where *n* is the number of observations.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','sgd','Weights',W,'Lambda',0.0003` specifies the solver as the stochastic gradient descent, the observation weights as the values in the vector `W`, and sets the regularization parameter at 0.0003.

Fitting Options

FitMethod — Method for fitting the model

`'exact'` (default) | `'none'` | `'average'`

Method for fitting the model, specified as the comma-separated pair consisting of `'FitMethod'` and one of the following:

- `'exact'` — Performs fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to fsrnca.
- `'average'` — Divides the data into partitions (subsets), fits each partition using the `exact` method, and returns the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

Example: `'FitMethod','none'`

`NumPartitions` — Number of partitions
`max(2,min(10,n))` (default) | integer between 2 and *n*

Number of partitions to split the data for using with `'FitMethod','average'` option, specified as the comma-separated pair consisting of `'NumPartitions'` and an integer value between 2 and *n*, where *n* is the number of observations.

Example: `'NumPartitions',15`

Data Types: `double` | `single`

`Lambda` — Regularization parameter
1/*n* (default) | nonnegative scalar

Regularization parameter to prevent overfitting, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

As the number of observations *n* increases, the chance of overfitting decreases and the required amount of regularization also decreases. See “Tune Regularization Parameter in NCA for Regression” on page 33-2503 to learn how to tune the regularization parameter.

Example: `'Lambda',0.002`

Data Types: `double` | `single`

`LengthScale` — Width of the kernel
1 (default) | positive real scalar

Width of the kernel, specified as the comma-separated pair consisting of `'LengthScale'` and a positive real scalar.

A length scale value of 1 is sensible when all predictors are on the same scale. If the predictors in `X` are of very different magnitudes, then consider standardizing the predictor values using `'Standardize',true` and setting `'LengthScale',1`.

Example: `'LengthScale',1.5`

Data Types: `double` | `single`

`InitialFeatureWeights` — Initial feature weights
`ones(p,1)` (default) | *p*-by-1 vector of real positive scalars

Initial feature weights, specified as the comma-separated pair consisting of `'InitialFeatureWeights'` and a *p*-by-1 vector of real positive scalars, where *p* is the number of predictors in the training data.

The regularized objective function for optimizing feature weights is nonconvex. As a result, using different initial feature weights can give different results. Setting all initial feature weights to 1

generally works well, but in some cases, random initialization using `rand(p,1)` can give better quality solutions.

Data Types: `double` | `single`

`Weights` — Observation weights
n-by-1 vector of 1s (default) | *n*-by-1 vector of real positive scalars

Observation weights, specified as the comma-separated pair consisting of `'ObservationWeights'` and an *n*-by-1 vector of real positive scalars. Use observation weights to specify higher importance of some observations compared to others. The default weights assign equal importance to all observations.

Data Types: `double` | `single`

`Standardize` — Indicator for standardizing predictor data
`false` (default) | `true`

Indicator for standardizing the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and either `false` or `true`. For more information, see “Impact of Standardization” on page 15-102.

Example: `'Standardize',true`

Data Types: `logical`

`Verbose` — Verbosity level indicator
0 (default) | 1 | >1

Verbosity level indicator for the convergence summary display, specified as the comma-separated pair consisting of `'Verbose'` and one of the following:

- 0 — No convergence summary
- 1 — Convergence summary, including norm of gradient and objective function values
- > 1 — More convergence information, depending on the fitting algorithm

  When using `'minibatch-lbfgs'` solver and verbosity level > 1, the convergence information includes iteration the log from intermediate mini-batch LBFGS fits.

Example: `'Verbose',1`

Data Types: `double` | `single`

`Solver` — Solver type
`'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Solver type for estimating feature weights, specified as the comma-separated pair consisting of `'Solver'` and one of the following:

- `'lbfgs'` — Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm
- `'sgd'` — Stochastic gradient descent (SGD) algorithm
- `'minibatch-lbfgs'` — Stochastic gradient descent with LBFGS algorithm applied to mini-batches

Default is `'lbfgs'` for $n \le 1000$, and `'sgd'` for $n > 1000$.

Example: `'solver','minibatch-lbfgs'`

`LossFunction` — Loss function

`'mad'` (default) | `'mse'` | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFunction'` and one of the following:

- `'mad'` — Mean absolute deviation

  $$l(y_i, y_j) = |y_i - y_j|.$$

- `'mse'` — Mean squared error

  $$l(y_i, y_j) = (y_i - y_j)^2.$$

- `'epsiloninsensitive'` — ε-insensitive loss function

  $$l(y_i, y_j) = \max(0, |y_i - y_j| - \epsilon).$$

  This loss function is more robust to outliers than mean squared error or mean absolute deviation.

- `@lossfun` — Custom loss function handle. A loss function has this form.

  ```
  function L = lossfun(Yu,Yv)
  % calculation of loss
  ...
  ```

  `Yu` is a *u*-by-1 vector and `Yv` is a *v*-by-1 vector. `L` is a *u*-by-*v* matrix of loss values such that `L(i,j)` is the loss value for `Yu(i)` and `Yv(j)`.

The objective function for minimization includes the loss function $l(y_i,y_j)$ as follows:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} p_{ij} l(y_i, y_j) + \lambda \sum_{r=1}^{p} w_r^2,$$

where w is the feature weight vector, n is the number of observations, and p is the number of predictor variables. p_{ij} is the probability that x_j is the reference point for x_i. For details, see “NCA Feature Selection for Regression” on page 15-101.

Example: `'LossFunction',@lossfun`

`Epsilon` — Epsilon value

`iqr(Y)/13.49` (default) | nonnegative real scalar

Epsilon value for the `'LossFunction','epsiloninsensitive'` option, specified as the comma-separated pair consisting of `'LossFunction'` and a nonnegative real scalar. The default value is an estimate of the sample standard deviation using the interquartile range of the response variable.

Example: `'Epsilon',0.1`

Data Types: `double` | `single`

`CacheSize` — Memory size

`1000MB` (default) | integer

Memory size, in MB, to use for objective function and gradient computation, specified as the comma-separated pair consisting of `'CacheSize'` and an integer.

Example: `'CacheSize',1500MB`

Data Types: `double` | `single`

LBFGS Options

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of history buffer for Hessian approximation for the `'lbfgs'` solver, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. At each iteration the function uses the most recent `HessianHistorySize` iterations to build an approximation to the inverse Hessian.

Example: `'HessianHistorySize',20`

Data Types: `double` | `single`

`InitialStepSize` — Initial step size
`'auto'` (default) | positive real scalar

Initial step size for the `'lbfgs'` solver, specified as the comma-separated pair consisting of `'InitialStepSize'` and a positive real scalar. By default, the function determines the initial step size automatically.

Data Types: `double` | `single`

`LineSearchMethod` — Line search method
`'weakwolfe'` (default) | `'strongwolfe'` | `'backtracking'`

Line search method, specified as the comma-separated pair consisting of `'LineSearchMethod'` and one of the following:

- `'weakwolfe'` — Weak Wolfe line search
- `'strongwolfe'` — Strong Wolfe line search
- `'backtracking'` — Backtracking line search

Example: `'LineSearchMethod','backtracking'`

`MaxLineSearchIterations` — Maximum number of line search iterations
`20` (default) | positive integer

Maximum number of line search iterations, specified as the comma-separated pair consisting of `'MaxLineSearchIterations'` and a positive integer.

Example: `'MaxLineSearchIterations',25`

Data Types: `double` | `single`

`GradientTolerance` — Relative convergence tolerance
`1e-6` (default) | positive real scalar

Relative convergence tolerance on the gradient norm for solver `lbfgs`, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive real scalar.

Example: `'GradientTolerance',0.000002`

Data Types: `double` | `single`

SGD Options

`InitialLearningRate` — Initial learning rate for `'sgd'` solver
`'auto'` (default) | positive real scalar

Initial learning rate for the `'sgd'` solver, specified as the comma-separated pair consisting of `'InitialLearningRate'` and a positive real scalar.

When using solver type `'sgd'`, the learning rate decays over iterations starting with the value specified for `'InitialLearningRate'`.

The default `'auto'` means that the initial learning rate is determined using experiments on small subsets of data. Use the `NumTuningIterations` name-value pair argument to specify the number of iterations for automatically tuning the initial learning rate. Use the `TuningSubsetSize` name-value pair argument to specify the number of observations to use for automatically tuning the initial learning rate.

For solver type `'minibatch-lbfgs'`, you can set `'InitialLearningRate'` to a very high value. In this case, the function applies LBFGS to each mini-batch separately with initial feature weights from the previous mini-batch.

To make sure the chosen initial learning rate decreases the objective value with each iteration, plot the `Iteration` versus the `Objective` values saved in the `mdl.FitInfo` property.

You can use the `refit` method with `'InitialFeatureWeights'` equal to `mdl.FeatureWeights` to start from the current solution and run additional iterations

Example: `'InitialLearningRate',0.9`

Data Types: `double` | `single`

`MiniBatchSize` — Number of observations to use in each batch for the `'sgd'` solver
min(10,*n*) (default) | positive integer value from 1 to *n*

Number of observations to use in each batch for the `'sgd'` solver, specified as the comma-separated pair consisting of `'MiniBatchSize'` and a positive integer from 1 to *n*.

Example: `'MiniBatchSize',25`

Data Types: `double` | `single`

`PassLimit` — Maximum number of passes for solver `'sgd'`
5 (default) | positive integer

Maximum number of passes through all *n* observations for solver `'sgd'`, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer. Each pass through all of the data is called an epoch.

Example: `'PassLimit',10`

Data Types: `double` | `single`

`NumPrint` — Frequency of batches for displaying convergence summary
10 (default) | positive integer value

Frequency of batches for displaying convergence summary for the `'sgd'` solver , specified as the comma-separated pair consisting of `'NumPrint'` and a positive integer. This argument applies when

the 'Verbose' value is greater than 0. NumPrint mini-batches are processed for each line of the convergence summary that is displayed on the command line.

Example: 'NumPrint',5

Data Types: double | single

NumTuningIterations — Number of tuning iterations
20 (default) | positive integer

Number of tuning iterations for the 'sgd' solver, specified as the comma-separated pair consisting of 'NumTuningIterations' and a positive integer. This option is valid only for 'InitialLearningRate','auto'.

Example: 'NumTuningIterations',15

Data Types: double | single

TuningSubsetSize — Number of observations to use for tuning initial learning rate
min(100,*n*) (default) | positive integer value from 1 to *n*

Number of observations to use for tuning the initial learning rate, specified as the comma-separated pair consisting of 'TuningSubsetSize' and a positive integer value from 1 to *n*. This option is valid only for 'InitialLearningRate','auto'.

Example: 'TuningSubsetSize',25

Data Types: double | single

SGD or LBFGS Options

IterationLimit — Maximum number of iterations
positive integer

Maximum number of iterations, specified as the comma-separated pair consisting of 'IterationLimit' and a positive integer. The default is 10000 for SGD and 1000 for LBFGS and mini-batch LBFGS.

Each pass through a batch is an iteration. Each pass through all of the data is an epoch. If the data is divided into *k* mini-batches, then every epoch is equivalent to *k* iterations.

Example: 'IterationLimit',250

Data Types: double | single

StepTolerance — Convergence tolerance on the step size
1e-6 (default) | positive real scalar

Convergence tolerance on the step size, specified as the comma-separated pair consisting of 'StepTolerance' and a positive real scalar. The 'lbfgs' solver uses an absolute step tolerance, and the 'sgd' solver uses a relative step tolerance.

Example: 'StepTolerance',0.000005

Data Types: double | single

Mini-batch LBFGS Options

MiniBatchLBFGSIterations — Maximum number of iterations per mini-batch LBFGS step
10 (default) | positive integer

Maximum number of iterations per mini-batch LBFGS step, specified as the comma-separated pair consisting of `'MiniBatchLBFGSIterations'` and a positive integer.

Example: `'MiniBatchLBFGSIterations',15`

Mini-batch LBFGS algorithm is a combination of SGD and LBFGS methods. Therefore, all of the name-value pair arguments that apply to SGD and LBFGS solvers also apply to the mini-batch LBFGS algorithm.

Data Types: `double` | `single`

Output Arguments

`mdl` — Neighborhood component analysis model for regression

`FeatureSelectionNCARegression` object

Neighborhood component analysis model for regression, returned as a `FeatureSelectionNCARegression` object.

References

[1] Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Campand, M. Revow, Z. Ghahramani, R. Kustra, R. Tibshirani. The DELVE Manual, 1996, http://mlg.eng.cam.ac.uk/pub/pdf/RasNeaHinetal96.pdf.

[2] University of Toronto, Computer Science Department. Delve Datasets. http://www.cs.toronto.edu/~delve/data/datasets.html.

[3] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[4] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[5] Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

See Also

`FeatureSelectionNCARegression` | `loss` | `predict` | `refit`

Topics

“Robust Feature Selection Using NCA for Regression” on page 15-85
“Neighborhood Component Analysis (NCA) Feature Selection” on page 15-99
“Introduction to Feature Selection” on page 15-49

Introduced in R2016b

fsrftest

Univariate feature ranking for regression using *F*-tests

Syntax

```
idx = fsrftest(Tbl,ResponseVarName)
idx = fsrftest(Tbl,formula)
idx = fsrftest(Tbl,Y)

idx = fsrftest(X,Y)

idx = fsrftest(___,Name,Value)
[idx,scores] = fsrftest(___)
```

Description

`idx = fsrftest(Tbl,ResponseVarName)` ranks features (predictors) using *F*-tests on page 33-2528. The table `Tbl` contains predictor variables and a response variable, and `ResponseVarName` is the name of the response variable in `Tbl`. The function returns `idx`, which contains the indices of predictors ordered by predictor importance, meaning `idx(1)` is the index of the most important predictor. You can use `idx` to select important predictors for regression problems.

`idx = fsrftest(Tbl,formula)` specifies a response variable and predictor variables to consider among the variables in `Tbl` by using `formula`.

`idx = fsrftest(Tbl,Y)` ranks predictors in `Tbl` using the response variable `Y`.

`idx = fsrftest(X,Y)` ranks predictors in `X` using the response variable `Y`.

`idx = fsrftest(___,Name,Value)` specifies additional options using one or more name-value pair arguments in addition to any of the input argument combinations in the previous syntaxes. For example, you can specify categorical predictors and observation weights.

`[idx,scores] = fsrftest(___)` also returns the predictor scores `scores`. A large score value indicates that the corresponding predictor is important.

Examples

Rank Predictors in Matrix

Rank predictors in a numeric matrix and create a bar plot of predictor importance scores.

Load the sample data.

```
load robotarm.mat
```

The `robotarm` data set contains 7168 training observations (`Xtrain` and `ytrain`) and 1024 test observations (`Xtest` and `ytest`) with 32 features [1][2].

Rank the predictors using the training observations.

```
[idx,scores] = fsrftest(Xtrain,ytrain);
```

The values in `scores` are the negative logs of the *p*-values. If a *p*-value is smaller than `eps(0)`, then the corresponding score value is `Inf`. Before creating a bar plot, determine whether `scores` includes `Inf` values.

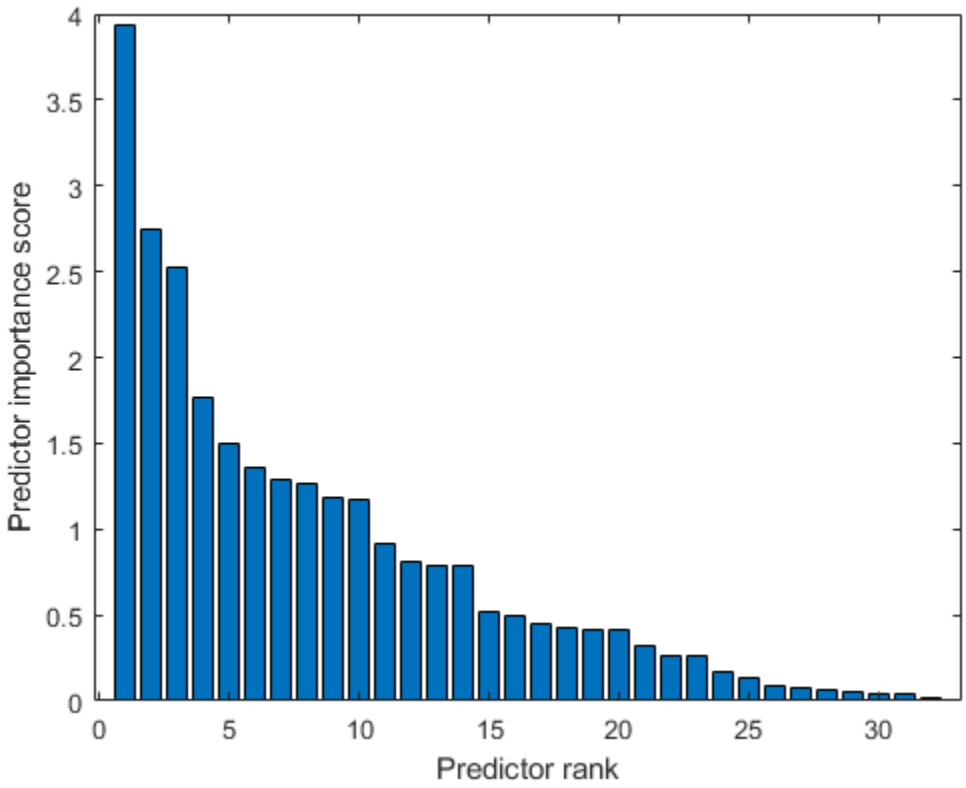
```
find(isinf(scores))
```

```
ans =

  1x0 empty double row vector
```

`scores` does not include `Inf` values. If `scores` includes `Inf` values, you can replace `Inf` by a large numeric number before creating a bar plot for visualization purposes. For details, see “Rank Predictors in Table” on page 33-2522.

Create a bar plot of the predictor importance scores.

```
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
```



Select the top five most important predictors. Find the columns of these predictors in `Xtrain`.

```
idx(1:5)
```

```
ans = 1×5
```

```
    30    24    10     4     5
```

The 30th column of `Xtrain` is the most important predictor of `ytrain`.

Rank Predictors in Table

Rank predictors in a table and create a bar plot of predictor importance scores.

If your data is in a table and `fsrftest` ranks a subset of the variables in the table, then the function indexes the variables using only the subset. Therefore, a good practice is to move the predictors that you do not want to rank to the end of the table. Move the response variable and observation weight vector as well. Then, the indexes of the output arguments are consistent with the indexes of the table. You can move variables in a table using the `movevars` function.

This example uses the Abalone data [3][4] from the UCI Machine Learning Repository [5]. Download the data and save it in your current folder with the name `'abalone.data'`.

Store the data in a table.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
tbl.Properties.VariableNames = {'Sex','Length','Diameter','Height', ...
    'WWeight','SWeight','VWeight','ShWeight','NoShellRings'};
```

Preview the first few rows of the table.

```
head(tbl)
```

```
ans=8×9 table
     Sex     Length    Diameter    Height    WWeight    SWeight    VWeight    ShWeight    NoShell
    _____    ______    ________    ______    _______    _______    _______    ________    _______

    {'M'}    0.455      0.365      0.095      0.514     0.2245      0.101       0.15          15
    {'M'}     0.35      0.265       0.09     0.2255     0.0995     0.0485       0.07           7
    {'F'}     0.53       0.42      0.135      0.677     0.2565     0.1415       0.21           9
    {'M'}     0.44      0.365      0.125      0.516     0.2155      0.114      0.155          10
    {'I'}     0.33      0.255       0.08      0.205     0.0895     0.0395      0.055           7
    {'I'}    0.425        0.3      0.095     0.3515      0.141     0.0775       0.12           8
    {'F'}     0.53      0.415       0.15     0.7775      0.237     0.1415       0.33          20
    {'F'}    0.545      0.425      0.125      0.768      0.294     0.1495       0.26          16
```

The last variable in the table is a response variable.

Rank the predictors in `tbl`. Specify the last column `NoShellRings` as a response variable.

```
[idx,scores] = fsrftest(tbl,'NoShellRings')
```

```
idx = 1×8

     3     4     5     7     8     2     6     1


scores = 1×8
```

```
  447.6891  736.9619       Inf       Inf       Inf  604.6692       Inf       Inf
```

The values in `scores` are the negative logs of the *p*-values. If a *p*-value is smaller than `eps(0)`, then the corresponding score value is `Inf`. Before creating a bar plot, determine whether `scores` includes `Inf` values.

```
idxInf = find(isinf(scores))
```

```
idxInf = 1×5

     3     4     5     7     8
```

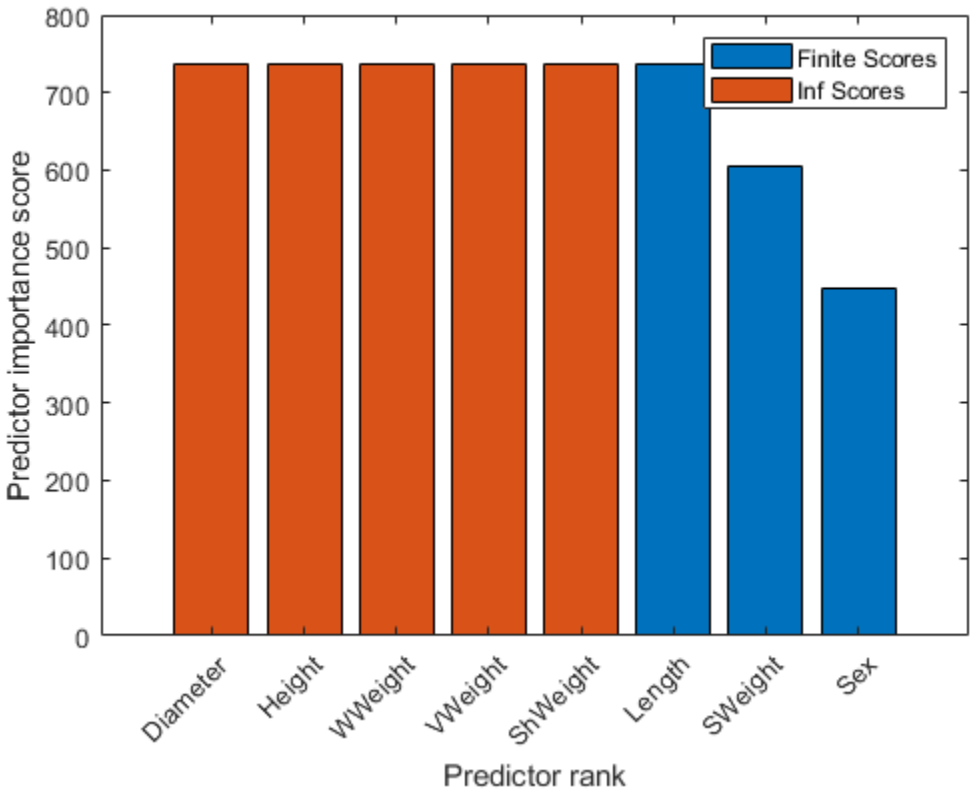
`scores` includes five `Inf` values.

Create a bar plot of predictor importance scores. Use the predictor names for the *x*-axis tick labels.

```
bar(scores(idx))
xlabel('Predictor rank')
ylabel('Predictor importance score')
xticklabels(strrep(tbl.Properties.VariableNames(idx),'_','\_'))
xtickangle(45)
```

The `bar` function does not plot any bars for the `Inf` values. For the `Inf` values, plot bars that have the same length as the largest finite score.

```
hold on
bar(scores(idx(length(idxInf)+1))*ones(length(idxInf),1))
legend('Finite Scores','Inf Scores')
hold off
```



The bar graph displays finite scores and Inf scores using different colors.

Input Arguments

`Tbl` — Sample data

table

Sample data, specified as a table. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for a response variable and observation weights.

A response variable can be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element of the response variable must correspond to one row of the array.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`. If `Tbl` also contains the observation weights, then you can specify the weights by using `Weights`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify the subset of variables by using `formula`.

- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The response variable and `Tbl` must have the same number of rows.

If `fsrftest` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. The values in the `'CategoricalPredictors'` name-value pair argument and the output argument `idx` do not count the predictors that the function does not rank.

`fsrftest` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Tbl` for a response variable to be missing values. `fsrftest` does not use observations with missing values for a response variable.

Data Types: `table`

`ResponseVarName` — Response variable name
character vector or string scalar containing name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of a variable in `Tbl`.

For example, if a response variable is the column `Y` of `Tbl` (`Tbl.Y`), then specify `ResponseVarName` as `'Y'`.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y ~ x1 + x2 + x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors, use a formula. If you specify a formula, then `fsrftest` does not rank any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

`Y` — Response variable
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Response variable, specified as a numeric, categorical, or logical vector, a character or string array, or a cell array of character vectors. Each row of `Y` represents the labels of the corresponding row of `X`.

`fsrftest` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `Y` to be missing values. `fsrftest` does not use observations with missing values for `Y`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumBins',20,'UseMissing',true` sets the number of bins as 20 and specifies to use missing values in predictors for ranking.

`CategoricalPredictors` — List of categorical predictors

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

List of categorical predictors, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fsrftest` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the names in `Tbl`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the names in `Tbl`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fsrftest` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fsrftest` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`NumBins` — Number of bins for binning continuous predictors

10 (default) | positive integer scalar

Number of bins for binning continuous predictors, specified as the comma-separated pair consisting of `'NumBins'` and a positive integer scalar.

Example: `'NumBins',50`

Data Types: `single` | `double`

`UseMissing` — Indicator for whether to use or discard missing values in predictors
`false` (default) | `true`

Indicator for whether to use or discard missing values in predictors, specified as the comma-separated pair consisting of `'UseMissing'` and either `true` to use or `false` to discard missing values in predictors for ranking.

`fsrftest` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values to be missing values.

If you specify `'UseMissing',true`, then `fsrftest` uses missing values for ranking. For a categorical variable, `fsrftest` treats missing values as an extra category. For a continuous variable, `fsrftest` places `NaN` values in a separate bin for binning.

If you specify `'UseMissing',false`, then `fsrftest` does not use missing values for ranking. Because `fsrftest` computes importance scores individually for each predictor, the function does not discard an entire row when values in the row are partially missing. For each variable, `fsrftest` uses all values that are not missing.

Example: `'UseMissing',true`

Data Types: `logical`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The function weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weight vector is the column `W` of `Tbl` (`Tbl.W`), then specify `'Weights','W'`.

`fsrftest` normalizes the weights to add up to one.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`idx` — Indices of predictors ordered by predictor importance
numeric vector

Indices of predictors in `X` or `Tbl` ordered by predictor importance, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

If `fsrftest` uses a subset of variables in `Tbl` as predictors, then the function indexes the predictors using only the subset. For example, suppose `Tbl` includes 10 columns and you specify the last five

columns of `Tbl` as the predictor variables by using `formula`. If `idx(3)` is `5`, then the third most important predictor is the 10th column in `Tbl`, which is the fifth predictor in the subset.

`scores` — Predictor scores
numeric vector

Predictor scores, returned as a 1-by-*r* numeric vector, where *r* is the number of ranked predictors.

A large score value indicates that the corresponding predictor is important.

- If you use `X` to specify the predictors or use all the variables in `Tbl` as predictors, then the values in `scores` have the same order as the predictors in `X` or `Tbl`.
- If you specify a subset of variables in `Tbl` as predictors, then the values in `scores` have the same order as the subset.

For example, suppose `Tbl` includes 10 columns and you specify the last five columns of `Tbl` as the predictor variables by using `formula`. Then, `score(3)` contains the score value of the 8th column in `Tbl`, which is the third predictor in the subset.

Algorithms

Univariate Feature Ranking Using *F*-Tests

- `fsrftest` examines the importance of each predictor individually using an *F*-test. Each *F*-test tests the hypothesis that the response values grouped by predictor variable values are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. A small *p*-value of the test statistic indicates that the corresponding predictor is important.
- The output `scores` is –log(*p*). Therefore, a large score value indicates that the corresponding predictor is important. If a *p*-value is smaller than `eps(0)`, then the output is `Inf`.
- `fsrftest` examines a continuous variable after binning, or discretizing, the variable. You can specify the number of bins using the `'NumBins'` name-value pair argument.

References

[1] Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Camp, M. Revow, Z. Ghahramani, R. Kustra, and R. Tibshirani. The DELVE Manual, 1996.

[2] University of Toronto, Computer Science Department. Delve Datasets.

[3] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[4] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[5] Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

See Also

`fsrnca` | `relieff` | `sequentialfs`

Topics

“Introduction to Feature Selection” on page 15-49

Introduced in R2020a

fstat

F mean and variance

Syntax

```
[M,V] = fstat(V1,V2)
```

Description

`[M,V] = fstat(V1,V2)` returns the mean of and variance for the *F* distribution with numerator degrees of freedom `V1` and denominator degrees of freedom `V2`. `V1` and `V2` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `M` and `V`. A scalar input for `V1` or `V2` is expanded to a constant arrays with the same dimensions as the other input. `V1` and `V2` parameters must contain real positive values.

The mean of the *F* distribution for values of ν_2 greater than 2 is

$$\frac{\nu_2}{\nu_2 - 2}$$

The variance of the *F* distribution for values of ν_2 greater than 4 is

$$\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

The mean of the *F* distribution is undefined if ν_2 is less than 3. The variance is undefined for ν_2 less than 5.

Examples

`fstat` returns `NaN` when the mean and variance are undefined.

```
[m,v] = fstat(1:5,1:5)
m =
    NaN  NaN  3.0000  2.0000  1.6667
v =
    NaN  NaN  NaN  NaN  8.8889
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`fcdf` | `finv` | `fpdf` | `frnd`

Topics

“F Distribution” on page B-45

Introduced before R2006a

fsulaplacian

Rank features for unsupervised learning using Laplacian scores

Syntax

```
idx = fsulaplacian(X)
idx = fsulaplacian(X,Name,Value)
[idx,scores] = fsulaplacian( ___ )
```

Description

`idx = fsulaplacian(X)` ranks features (variables) in `X` using the Laplacian scores on page 33-2538. The function returns `idx`, which contains the indices of features ordered by feature importance. You can use `idx` to select important features for unsupervised learning.

`idx = fsulaplacian(X,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify `'NumNeighbors',10` to create a similarity graph on page 33-2537 using 10 nearest neighbors.

`[idx,scores] = fsulaplacian( ___ )` also returns the feature scores `scores`, using any of the input argument combinations in the previous syntaxes. A large score value indicates that the corresponding feature is important.

Examples

Rank Features by Importance

Load the sample data.

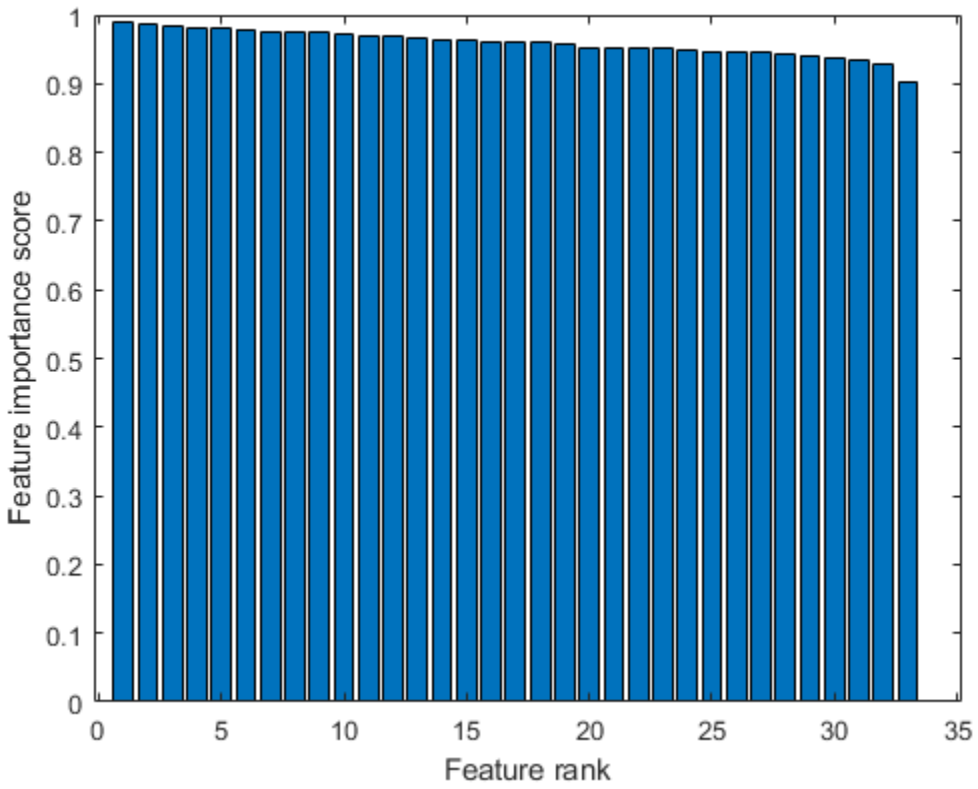
```
load ionosphere
```

Rank the features based on importance.

```
[idx,scores] = fsulaplacian(X);
```

Create a bar plot of the feature importance scores.

```
bar(scores(idx))
xlabel('Feature rank')
ylabel('Feature importance score')
```



Select the top five most important features. Find the columns of these features in X.

```
idx(1:5)
```

```
ans = 1×5

    15    13    17    21    19
```

The 15th column of X is the most important feature.

Rank Features Using Specified Similarity Matrix

Compute a similarity matrix from Fisher's iris data set and rank the features using the similarity matrix.

Load Fisher's iris data set.

```
load fisheriris
```

Find the distance between each pair of observations in `meas` by using the `pdist` and `squareform` functions with the default Euclidean distance metric.

```
D = pdist(meas);
Z = squareform(D);
```

Construct the similarity matrix and confirm that it is symmetric.

```
S = exp(-Z.^2);
issymmetric(S)
```

```
ans = logical
   1
```

Rank the features.

```
idx = fsulaplacian(meas,'Similarity',S)
```

```
idx = 1×4

     3     4     1     2
```

Ranking using the similarity matrix `S` is the same as ranking by specifying `'NumNeighbors'` as `size(meas,1)`.

```
idx2 = fsulaplacian(meas,'NumNeighbors',size(meas,1))
```

```
idx2 = 1×4

     3     4     1     2
```

Input Arguments

X — Input data

numeric matrix

Input data, specified as an *n*-by-*p* numeric matrix. The rows of `X` correspond to observations (or points), and the columns correspond to features.

The software treats `NaN`s in `X` as missing data and ignores any row of `X` containing at least one `NaN`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumNeighbors',10,'KernelScale','auto'` specifies the number of nearest neighbors as 10 and the kernel scale factor as `'auto'`.

Similarity — Similarity matrix

`[]` (empty matrix) (default) | symmetric matrix

Similarity matrix, specified as the comma-separated pair consisting of `'Similarity'` and an *n*-by-*n* symmetric matrix, where *n* is the number of observations. The similarity matrix (or adjacency matrix) represents the input data by modeling local neighborhood relationships among the data points. The values in a similarity matrix represent the edges (or connections) between nodes (data points) that

are connected in a similarity graph on page 33-2537. For more information, see “Similarity Matrix” on page 33-2538.

If you specify the `'Similarity'` value, then you cannot specify any other name-value pair argument. If you do not specify the `'Similarity'` value, then the software computes a similarity matrix using the options specified by the other name-value pair arguments.

Data Types: `single` | `double`

`Distance` — Distance metric
character vector | string scalar | function handle

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle, as described in this table.

Value	Description
`'euclidean'`	Euclidean distance (default)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation computed from `X`. Use the `Scale` name-value pair argument to specify a different scaling factor.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use the `Cov` name-value pair argument to specify a different covariance matrix.
`'cityblock'`	City block distance
`'minkowski'`	Minkowski distance. The default exponent is 2. Use the `P` name-value pair argument to specify a different exponent, where `P` is a positive scalar value.
`'chebychev'`	Chebychev distance (maximum coordinate difference)
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors)
`'correlation'`	One minus the sample correlation between observations (treated as sequences of values)
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

Value	Description

`@distfun`	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For more information, see “Distance Metrics” on page 18-12.

When you use the `'seuclidean'`, `'minkowski'`, or `'mahalanobis'` distance metric, you can specify the additional name-value pair argument `'Scale'`, `'P'`, or `'Cov'`, respectively, to control the distance metrics.

Example: `'Distance','minkowski','P',3` specifies to use the Minkowski distance metric with an exponent of 3.

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar.

This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

Cov — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix.

This argument is valid only if `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(4)`

Data Types: `single` | `double`

Scale — Scaling factors for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | numeric vector of nonnegative values

Scaling factors for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a numeric vector of nonnegative values.

`Scale` has length *p* (the number of columns in `X`), because each dimension (column) of `X` has a corresponding value in `Scale`. For each dimension of `X`, `fsulaplacian` uses the corresponding value in `Scale` to standardize the difference between observations.

This argument is valid only if `'Distance'` is `'seuclidean'`.

Data Types: `single` | `double`

`NumNeighbors` — Number of nearest neighbors
`log(size(X,1))` (default) | positive integer

Number of nearest neighbors used to construct the similarity graph, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer.

Example: `'NumNeighbors',10`

Data Types: `single` | `double`

`KernelScale` — Scale factor
`1` (default) | `'auto'` | positive scalar

Scale factor for the kernel, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software uses the scale factor to transform distances to similarity measures. For more information, see “Similarity Graph” on page 33-2537.

- The `'auto'` option is supported only for the `'euclidean'` and `'seuclidean'` distance metrics.
- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. To reproduce results, set a random number seed using `rng` before calling `fsulaplacian`.

Example: `'KernelScale','auto'`

Output Arguments

`idx` — Indices of features ordered by feature importance
numeric vector

Indices of the features in `X` ordered by feature importance, returned as a numeric vector. For example, if `idx(3)` is `5`, then the third most important feature is the fifth column in `X`.

`scores` — Feature scores
numeric vector

Feature scores, returned as a numeric vector. A large score value in `scores` indicates that the corresponding feature is important. The values in `scores` have the same order as the features in `X`.

More About

Similarity Graph

A similarity graph models the local neighborhood relationships between data points in `X` as an undirected graph. The nodes in the graph represent data points, and the edges, which are directionless, represent the connections between the data points.

If the pairwise distance $Dist_{i,j}$ between any two nodes i and j is positive (or larger than a certain threshold), then the similarity graph connects the two nodes using an edge [2]. The edge between the two nodes is weighted by the pairwise similarity $S_{i,j}$, where $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$, for a specified kernel scale σ value.

`fsulaplacian` constructs a similarity graph using the nearest neighbor method. The function connects points in `X` that are nearest neighbors. Use `'NumNeighbors'` to specify the number of nearest neighbors.

Similarity Matrix

A similarity matrix is a matrix representation of a similarity graph on page 33-2537. The n-by-n matrix $S = (S_{i,j})_{i,j = 1,\ldots,n}$ contains pairwise similarity values between connected nodes in the similarity graph. The similarity matrix of a graph is also called an adjacency matrix.

The similarity matrix is symmetric because the edges of the similarity graph are directionless. A value of $S_{i,j} = 0$ means that nodes i and j of the similarity graph are not connected.

Degree Matrix

A degree matrix D_g is an n-by-n diagonal matrix obtained by summing the rows of the similarity matrix on page 33-2538 S. That is, the ith diagonal element of D_g is $D_g(i,i) = \sum_{j=1}^{n} S_{i,j}$.

Laplacian Matrix

A Laplacian matrix, which is one way of representing a similarity graph on page 33-2537, is defined as the difference between the degree matrix on page 33-2538 D_g and the similarity matrix on page 33-2538 S.

$$L = D_g - S.$$

Algorithms

Laplacian Score

The `fsulaplacian` function ranks features using Laplacian scores[1] obtained from a nearest neighbor similarity graph on page 33-2537.

`fsulaplacian` computes the values in `scores` as follows:

1. For each data point in `X`, define a local neighborhood using the nearest neighbor method, and find pairwise distances $Dist_{i,j}$ for all points i and j in the neighborhood.
2. Convert the distances to the similarity matrix on page 33-2538 S using the kernel transformation $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$, where σ is the scale factor for the kernel as specified by the `'KernelScale'` name-value pair argument.
3. Center each feature by removing its mean.

$$\tilde{x}_r = x_r - \frac{x_r^T D_g \mathbf{1}}{\mathbf{1}^T D_g \mathbf{1}}\mathbf{1},$$

where x_r is the rth feature, D_g is the degree matrix on page 33-2538, and $\mathbf{1}^T = [1, \cdots, 1]^T$.

4 Compute the score s_r for each feature.

$$s_r = \frac{\tilde{x}_r^T S \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r}.$$

Note that [1] defines the Laplacian score as

$$L_r = \frac{\tilde{x}_r^T L \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r} = 1 - \frac{\tilde{x}_r^T S \tilde{x}_r}{\tilde{x}_r^T D_g \tilde{x}_r},$$

where L is the Laplacian matrix on page 33-2538, defined as the difference between D_g and S. The `fsulaplacian` function uses only the second term of this equation for the score value of `scores` so that a large score value indicates an important feature.

Selecting features using the Laplacian score is consistent with minimizing the value

$$\frac{\sum_{i,j} (x_{ir} - x_{jr})^2 S_{i,j}}{Var(x_r)},$$

where x_{ir} represents the ith observation of the rth feature. Minimizing this value implies that the algorithm prefers features with large variance. Also, the algorithm assumes that two data points of an important feature are close if and only if the similarity graph has an edge between the two data points.

References

[1] He, X., D. Cai, and P. Niyogi. "Laplacian Score for Feature Selection." *NIPS Proceedings*. 2005.

[2] Von Luxburg, U. “A Tutorial on Spectral Clustering.” *Statistics and Computing Journal*. Vol.17, Number 4, 2007, pp. 395–416.

See Also

`fscmrmr` | `relieff` | `sequentialfs`

Topics

“Introduction to Feature Selection” on page 15-49

Introduced in R2019b

fsurfht

Interactive contour plot

Syntax

```
fsurfht(fun,xlims,ylims)
fsurfht(fun,xlims,ylims,p1,p2,p3,p4,p5)
```

Description

`fsurfht(fun,xlims,ylims)` is an interactive contour plot of the function specified by the text variable `fun`. The *x*-axis limits are specified by `xlims` in the form `[xmin xmax]`, and the *y*-axis limits are specified by `ylims` in the form `[ymin ymax]`.

`fsurfht(fun,xlims,ylims,p1,p2,p3,p4,p5)` allows for five optional parameters that you can supply to the function `fun`.

The intersection of the vertical and horizontal reference lines on the plot defines the current *x* value and *y* value. You can drag these reference lines and watch the calculated *z*-values (at the top of the plot) update simultaneously. Alternatively, you can type the *x* value and *y* value into editable text fields on the *x*-axis and *y*-axis.

Examples

Plot the Gaussian likelihood function for the `gas.mat` data.

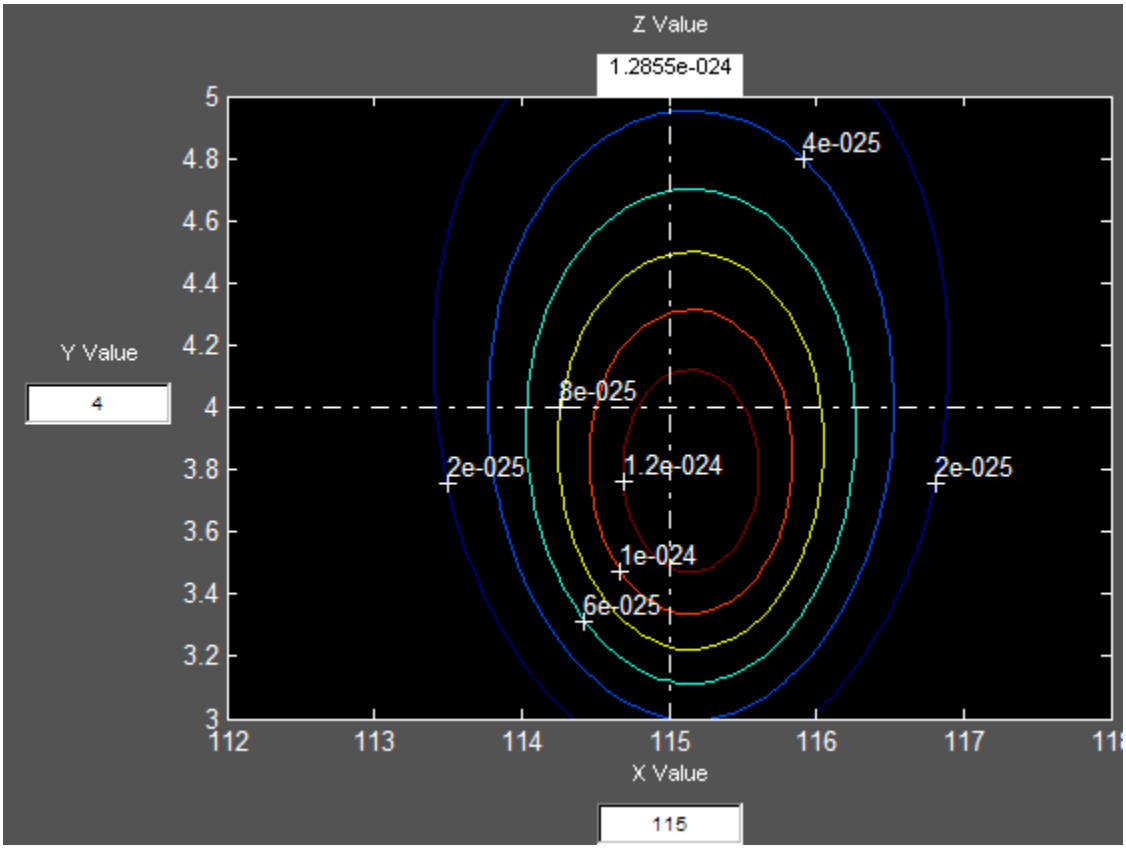
```
load gas
```

Create a function containing the following commands, and name it `gauslike.m`.

```
function z = gauslike(mu,sigma,p1)
n = length(p1);
z = ones(size(mu));
for i = 1:n
z = z .* (normpdf(p1(i),mu,sigma));
end
```

The `gauslike` function calls `normpdf`, treating the data sample as fixed and the parameters μ and σ as variables. Assume that the gas prices are normally distributed, and plot the likelihood surface of the sample.

```
fsurfht('gauslike',[112 118],[3 5],price1)
```



The sample mean is the *x* value at the maximum, but the sample standard deviation is *not* the *y* value at the maximum.

```
mumax = mean(price1)
mumax =
 115.1500
sigmamax = std(price1)*sqrt(19/20)
sigmamax =
  3.7719
```

Introduced before R2006a

fullfact

Full factorial design

Syntax

```
dFF = fullfact(levels)
```

Description

`dFF = fullfact(levels)` gives factor settings `dFF` for a full factorial design with *n* factors, where the number of levels for each factor is given by the vector `levels` of length *n*. `dFF` is *m*-by-*n*, where *m* is the number of treatments in the full-factorial design. Each row of `dFF` corresponds to a single treatment. Each column contains the settings for a single factor, with integer values from one to the number of levels.

Examples

The following generates an eight-run full-factorial design with two levels in the first factor and four levels in the second factor:

```
dFF = fullfact([2 4])
dFF =
   1   1
   2   1
   1   2
   2   2
   1   3
   2   3
   1   4
   2   4
```

See Also

`ff2n`

Introduced before R2006a

gagerr

Gage repeatability and reproducibility study

Syntax

```
gagerr(y,{part,operator})
gagerr(y,GROUP)
gagerr(y,part)
gagerr(...,param1,val1,param2,val2,...)
[TABLE, stats] = gagerr(...)
```

Description

`gagerr(y,{part,operator})` performs a gage repeatability and reproducibility study on measurements in `y` collected by `operator` on `part`. `y` is a column vector containing the measurements on different parts. `part` and `operator` are categorical variables, numeric vectors, character matrices, string arrays, or cell arrays of character vectors. The number of elements in `part` and `operator` should be the same as in `y`.

`gagerr` prints a table in the command window in which the decomposition of variance, standard deviation, study var (5.15 x standard deviation) are listed with respective percentages for different sources. Summary statistics are printed below the table giving the number of distinct categories (NDC) and the percentage of Gage R&R of total variations (PRR).

`gagerr` also plots a bar graph showing the percentage of different components of variations. Gage R&R, repeatability, reproducibility, and part-to-part variations are plotted as four vertical bars. Variance and study var are plotted as two groups.

To determine the capability of a measurement system using NDC, use the following guidelines:

- If NDC > 5, the measurement system is capable.
- If NDC < 2, the measurement system is not capable.
- Otherwise, the measurement system may be acceptable.

To determine the capability of a measurement system using PRR, use the following guidelines:

- If PRR < 10%, the measurement system is capable.
- If PRR > 30%, the measurement system is not capable.
- Otherwise, the measurement system may be acceptable.

`gagerr(y,GROUP)` performs a gage R&R study on measurements in `y` with `part` and `operator` represented in `GROUP`. `GROUP` is a numeric matrix whose first and second columns specify different parts and operators, respectively. The number of rows in `GROUP` should be the same as the number of elements in `y`.

`gagerr(y,part)` performs a gage R&R study on measurements in `y` without operator information. The assumption is that all variability is contributed by `part`.

`gagerr(...,param1,val1,param2,val2,...)` performs a gage R&R study using one or more of the following parameter name/value pairs:

- `'spec'` — A two-element vector that defines the lower and upper limit of the process, respectively. In this case, summary statistics printed in the command window include Precision-to-Tolerance Ratio (PTR). Also, the bar graph includes an additional group, the percentage of tolerance.

  To determine the capability of a measurement system using PTR, use the following guidelines:

  - If PTR < 0.1, the measurement system is capable.
  - If PTR > 0.3, the measurement system is not capable.
  - Otherwise, the measurement system may be acceptable.
- `'printtable'` — A value `'on'` or `'off'` that indicates whether the tabular output should be printed in the command window or not. The default value is `'on'`.
- `'printgraph'` — A value `'on'` or `'off'` that indicates whether the bar graph should be plotted or not. The default value is `'on'`.
- `'randomoperator'` — A logical value, `true` or `false`, that indicates whether the effect of `operator` is random or not. The default value is `true`.
- `'model'` — The model to use, specified by one of:

  - `'linear'` — Main effects only (default)
  - `'interaction'` — Main effects plus two-factor interactions
  - `'nested'` — Nest `operator` in `part`

  The default value is `'linear'`.

`[TABLE, stats] = gagerr(...)` returns a 6-by-5 matrix `TABLE` and a structure `stats`. The columns of `TABLE`, from left to right, represent variance, percentage of variance, standard deviations, study var, and percentage of study var. The rows of `TABLE`, from top to bottom, represent different sources of variations: gage R&R, repeatability, reproducibility, operator, operator and part interactions, and part. `stats` is a structure containing summary statistics for the performance of the measurement system. The fields of `stats` are:

- `ndc` — Number of distinct categories
- `prr` — Percentage of gage R&R of total variations
- `ptr` — Precision-to-tolerance ratio. The value is `NaN` if the parameter `'spec'` is not given.

Examples

Gage R&R Study

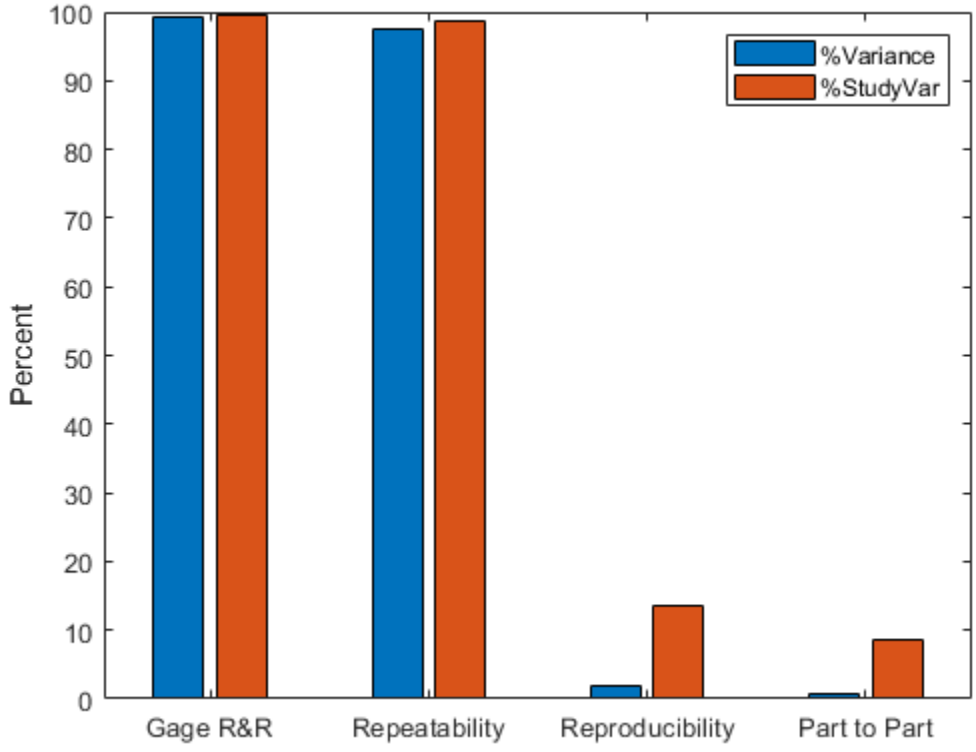
Simulate a measurement system by randomly generating the operators, parts, and the measurements, `y` , operators do on the parts.

```
rng(1234,'twister')              % for reproducibility
y = randn(100,1);                % measurements
part = ceil(3*rand(100,1));      % parts
operator = ceil(4*rand(100,1));  % operators
```

Conduct a gage R&R study for this system using a mixed ANOVA model without interactions.

```
gagerr(y,{part, operator},'randomoperator',true)
```

```
  Columns 1 through 4

    {'Source'           }    {'Variance'}    {'% Variance'}    {'sigma' }
    {'Gage R&R'         }    {[  0.9715]}    {[   99.2653]}    {[0.9857]}
    {'  Repeatability'  }    {[  0.9535]}    {[   97.4201]}    {[0.9765]}
    {'  Reproducibility'}    {[  0.0181]}    {[    1.8452]}    {[0.1344]}
    {'   Operator'      }    {[  0.0181]}    {[    1.8452]}    {[0.1344]}
    {'Part'             }    {[  0.0072]}    {[    0.7347]}    {[0.0848]}
    {'Total'            }    {[  0.9787]}    {[       100]}    {[0.9893]}

  Columns 5 through 6

    {'5.15*sigma'}    {'% 5.15*sigma'}
    {[    5.0762]}    {[     99.6320]}
    {[    5.0288]}    {[     98.7016]}
    {[    0.6921]}    {[     13.5838]}
    {[    0.6921]}    {[     13.5838]}
    {[    0.4367]}    {[      8.5716]}
    {[    5.0949]}    {0x0 char      }

Number of distinct categories (NDC):0
% of Gage R&R of total variations (PRR): 99.63
Note: The last column of the above table does not have to sum to 100%
```



References

[1] Burdick, Richard K., Connie M. Borror, and Douglas C. Montgomery. *Design and Analysis of Gauge R&R Studies: Making Decisions with Confidence Intervals in Random and Mixed ANOVA Models*. Society for Industrial Applied Mathematics: American Statistical Association, 2005.

See Also

Topics

“Grouping Variables” on page 2-45

Introduced in R2006b

gamcdf

Gamma cumulative distribution function

Syntax

```
p = gamcdf(x,a)
p = gamcdf(x,a,b)

[p,pLo,pUp] = gamcdf(x,a,b,pCov)
[p,pLo,pUp] = gamcdf(x,a,b,pCov,alpha)

___ = gamcdf( ___ ,'upper')
```

Description

`p = gamcdf(x,a)` returns the cumulative distribution function (cdf) of the standard gamma distribution with the shape parameters in `a`, evaluated at the values in `x`.

`p = gamcdf(x,a,b)` returns the cdf of the gamma distribution with the shape parameters in `a` and scale parameters in `b`, evaluated at the values in `x`.

`[p,pLo,pUp] = gamcdf(x,a,b,pCov)` also returns the 95% confidence interval [`pLo`,`pUp`] of `p` when `a` and `b` are estimates. `pCov` is the covariance matrix of the estimated parameters.

`[p,pLo,pUp] = gamcdf(x,a,b,pCov,alpha)` specifies the confidence level for the confidence interval [`pLo` `pUp`] to be `100(1–alpha)`%.

`___ = gamcdf( ___ ,'upper')` returns the complement of the cdf, evaluated at the values in `x`, using an algorithm that more accurately computes the extreme upper-tail probabilities than subtracting the lower tail value from 1. `'upper'` can follow any of the input argument combinations in the previous syntaxes.

Examples

Compute Gamma Distribution cdf

Compute the cdf of the mean of the gamma distribution, which is equal to the product of the parameters *ab*.

```
a = 1:6;
b = 5:10;
prob = gamcdf(a.*b,a,b)
```

```
prob = 1×6

    0.6321    0.5940    0.5768    0.5665    0.5595    0.5543
```

As `ab` increases, the distribution becomes more symmetric, and the mean approaches the median.

Confidence Interval of Gamma cdf Value

Find a confidence interval estimating the probability that an observation is in the interval `[0 10]` using gamma distributed data.

Generate a sample of `1000` gamma distributed random numbers with shape `2` and scale `5`.

```
x = gamrnd(2,5,1000,1);
```

Compute estimates for the parameters.

```
[params,~] = gamfit(x)
```

```
params = 1×2

    2.1089    4.8147
```

Store the parameters as `ahat` and `bhat`.

```
ahat = params(1);
bhat = params(2);
```

Find the covariance of the parameter estimates.

```
[~,nCov] = gamlike(params,x)
```

```
nCov = 2×2

    0.0077   -0.0176
   -0.0176    0.0512
```

Create a confidence interval estimating the probability that an observation is in the interval `[0 10]`.

```
[prob,pLo,pUp] = gamcdf(10,ahat,bhat,nCov)
```

```
prob = 0.5830

pLo = 0.5587

pUp = 0.6069
```

Complementary cdf (Tail Distribution)

Determine the probability that an observation from the gamma distribution with shape parameter 2 and scale parameter 3 will is in the interval `[150 Inf]`.

```
p1 = 1 - gamcdf(150,2,3)
```

```
p1 = 0
```

`gamcdf(150, 2, 3)` is nearly `1`, so `p1` becomes `0`. Specify `'upper'` so that `gamcdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = gamcdf(150,2,3,'upper')
```

```
p2 = 9.8366e-21
```

Input Arguments

x — Values at which to evaluate cdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the cdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `x` must be a scalar value.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gamcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

a — Shape parameter

positive scalar value | array of positive scalar values

Shape of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gamcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

b — Scale parameter

1 (default) | positive scalar value | array of positive scalar values

Scale of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gamcdf` expands each scalar input into a constant array of the same size as the array

inputs. Each element in p is the cdf value of the distribution specified by the corresponding elements in a and b, evaluated at the corresponding element in x.

Example: [1 1 2 2]

Data Types: single | double

pCov — Covariance of estimates
2-by-2 numeric matrix

Covariance of the estimates a and b, specified as a 2-by-2 matrix.

If you specify pCov to compute the confidence interval [pLo,pUp], then x, a, and b must be scalar values.

You can estimate a and b by using gamfit or mle, and estimate the covariance of a and b by using gamlike. For an example, see “Confidence Interval of Gamma cdf Value” on page 33-2548.

Data Types: single | double

alpha — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is 100(1–alpha)%, where alpha is the probability that the confidence interval does not contain the true value.

Example: 0.01

Data Types: single | double

Output Arguments

p — cdf values
scalar value | array of scalar values

cdf values evaluated at the values in x, returned as a scalar value or an array of scalar values. p is the same size as x, a, and b after any necessary scalar expansion. Each element in p is the cdf value of the distribution specified by the corresponding elements in a and b, evaluated at the corresponding element in x.

pLo — Lower confidence bound for p
scalar value | array of scalar values

Lower confidence bound for p, returned as a scalar value or an array of scalar values. pLo has the same size as p.

pUp — Upper confidence bound for p
scalar value | array of scalar values

Upper confidence bound for p, returned as a scalar value or an array of scalar values. pUp has the same size as p.

More About

Gamma cdf

The gamma distribution is a two-parameter family of curves. The parameters a and b are shape and scale, respectively.

The gamma cdf is

$$p = F(x \mid a, b) = \frac{1}{b^a \Gamma(a)} \int_0^x t^{a-1} e^{\frac{-t}{b}} dt .$$

The result p is the probability that a single observation from a gamma distribution with parameters a and b falls in the interval $[0,x]$.

The gamma cdf is related to the incomplete gamma function `gammainc` by

$$f(x \mid a, b) = \text{gammainc}\left(\frac{x}{b}, a\right).$$

The standard gamma distribution occurs when $b = 1$, which coincides with the incomplete gamma function precisely.

For more information, see “Gamma Distribution” on page B-47.

Alternative Functionality

- `gamcdf` is a function specific to the gamma distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, create a `GammaDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `gamcdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GammaDistribution` | `cdf` | `gamfit` | `gaminv` | `gamlike` | `gamma` | `gampdf` | `gamrnd` | `gamstat`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gamfit

Gamma parameter estimates

Syntax

```
phat = gamfit(data)
[phat,pci] = gamfit(data)
[phat,pci] = gamfit(data,alpha)
[...] = gamfit(data,alpha,censoring,freq,options)
```

Description

`phat = gamfit(data)` returns the maximum likelihood estimates (MLEs) for the parameters of the gamma distribution given the data in vector `data`.

`[phat,pci] = gamfit(data)` returns MLEs and 95% percent confidence intervals. The first row of `pci` is the lower bound of the confidence intervals; the last row is the upper bound.

`[phat,pci] = gamfit(data,alpha)` returns `100(1 - alpha)`% confidence intervals. For example, `alpha = 0.01` yields 99% confidence intervals.

`[...] = gamfit(data,alpha,censoring)` accepts a Boolean vector of the same size as `data` that is 1 for observations that are right-censored and 0 for observations that are observed exactly.

`[...] = gamfit(data,alpha,censoring,freq)` accepts a frequency vector of the same size as `data`. `freq` typically contains integer frequencies for the corresponding elements in `data`, but may contain any nonnegative values.

`[...] = gamfit(data,alpha,censoring,freq,options)` accepts a structure, `options`, that specifies control parameters for the iterative algorithm the function uses to compute maximum likelihood estimates. The gamma fit function accepts an `options` structure which can be created using the function `statset`. Enter `statset('gamfit')` to see the names and default values of the parameters that `gamfit` accepts in the `options` structure.

Examples

Fit a gamma distribution to random data generated from a specified gamma distribution:

```
a = 2; b = 4;
data = gamrnd(a,b,100,1);

[p,ci] = gamfit(data)
p =
  2.1990  3.7426
ci =
  1.6840  2.8298
  2.7141  4.6554
```

References

[1] Hahn, Gerald J., and S. S. Shapiro. *Statistical Models in Engineering*. Hoboken, NJ: John Wiley & Sons, Inc., 1994, p. 88.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gamcdf` | `gaminv` | `gamlike` | `gampdf` | `gamrnd` | `gamstat` | `mle`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gaminv

Gamma inverse cumulative distribution function

Syntax

```
x = gaminv(p,a)
x = gaminv(p,a,b)

[x,xLo,xUp] = gaminv(p,a,b,pCov)
[x,xLo,xUp] = gaminv(p,a,b,pCov,alpha)
```

Description

`x = gaminv(p,a)` returns the inverse cumulative distribution function (icdf) of the standard gamma distribution with the shape parameter `a`, evaluated at the values in `p`.

`x = gaminv(p,a,b)` returns the icdf of the gamma distribution with shape parameter `a` and the scale parameter `b`, evaluated at the values in `p`.

`[x,xLo,xUp] = gaminv(p,a,b,pCov)` also returns the 95% confidence interval `[xLo,xUp]` of `x` when `a` and `b` are estimates. `pCov` is the covariance matrix of the estimated parameters.

`[x,xLo,xUp] = gaminv(p,a,b,pCov,alpha)` specifies the confidence level for the confidence interval `[xLo,xUp]` to be `100(1–alpha)`%.

Examples

Compute Gamma icdf

Find the median of the gamma distribution with shape parameter `3` and scale parameter `5`.

```
x = gaminv(0.5,3,5)
```

```
x = 13.3703
```

Confidence Interval of Gamma icdf Value

Find a confidence interval estimating the median using gamma distributed data.

Generate a sample of `500` gamma distributed random numbers with shape `2` and scale `5`.

```
x = gamrnd(2,5,500,1);
```

Compute estimates for the parameters.

```
params = gamfit(x)
```

```
params = 1×2
```

```
    1.9820    5.0601
```

Store the estimates for the parameters as `ahat` and `bhat`.

```
ahat = params(1);
bhat = params(2);
```

Compute the covariance of the parameter estimates.

```
[~,nCov] = gamlike(params,x)
```

```
nCov = 2×2

    0.0135   -0.0346
   -0.0346    0.1141
```

Create a confidence interval estimating x.

```
[x,xLo,xUp] = gaminv(0.50,ahat,bhat,nCov)
```

```
x = 8.4021

xLo = 7.8669

xUp = 8.9737
```

Input Arguments

p — Probability values at which to evaluate icdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the inverse cdf (icdf), specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `p` must be a scalar value (not an array).

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `p`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gaminv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `p`.

Example: `[0.1,0.5,0.9]`

Data Types: `single` | `double`

a — Shape Parameter

positive scalar value | array of positive scalar values

Shape parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `p`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gaminv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `p`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

b — Scale Parameter

1 (default) | positive scalar value | array of positive scalar values

Scale parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `p`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gaminv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `p`.

Example: `[1 1 2 2]`

Data Types: `single` | `double`

pCov — Covariance of estimates

2-by-2 numeric matrix

Covariance of the estimates `a` and `b`, specified as a 2-by-2 matrix.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `p`, `a`, and `b` must be scalar values.

You can estimate `a` and `b` by using `gamfit` or `mle`, and estimate the covariance of `a` and `b` by using `gamlike`. For an example, see “Confidence Interval of Gamma icdf Value” on page 33-2555.

Data Types: `single` | `double`

alpha — Significance level

0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

x — icdf values

scalar value | array of scalar values

icdf values evaluated at the probability values in `p`, returned as a scalar value or an array of scalar values. `x` is the same size as `p`, `a`, and `b`, after any necessary scalar expansion. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `p`.

xLo — Lower confidence bound for x

scalar value | array of scalar values

Lower confidence bound for `x`, returned as a scalar value or an array of scalar values. `xLo` has the same size as `x`.

xUp — Upper confidence bound for x

scalar value | array of scalar values

Upper confidence bound for `x`, returned as a scalar value or an array of scalar values. `xUp` has the same size as `x`.

More About

Gamma icdf

The gamma distribution is a two-parameter family of curves. The parameters *a* and *b* are shape and scale, respectively.

The gamma inverse function in terms of the gamma cdf is

$$x = F^{-1}(p\,|\,a,b) = \left\{x : F(x\,|\,a,b) = p\right\},$$

where

$$p = F(x\,|\,a,b) = \frac{1}{b^a \Gamma(a)} \int_0^x t^{a-1} e^{\frac{-t}{b}} dt\,.$$

The result *x* is the value such that an observation from the gamma distribution with parameters *a* and *b* falls in [0,*x*] with probability *p*.

For more information, see “Gamma Distribution” on page B-47.

Algorithms

No known analytical solution exists for the integral equation shown in “Gamma icdf” on page 33-2558. `gaminv` uses an iterative approach (Newton's method) to converge on the solution.

Alternative Functionality

- `gaminv` is a function specific to the gamma distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use

`icdf`, create a `GammaDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `gaminv` is faster than the generic function `icdf`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GammaDistribution` | `gamcdf` | `gamfit` | `gamlike` | `gampdf` | `gamrnd` | `gamstat` | `icdf`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gamlike

Gamma negative log-likelihood

Syntax

```
nlogL = gamlike(params,data)
[nlogL,AVAR] = gamlike(params,data)
```

Description

`nlogL = gamlike(params,data)` returns the negative of the gamma log-likelihood of the parameters, `params`, given `data`. `params(1)=A`, shape parameters, and `params(2)=B`, scale parameters. The parameters in `params` must all be positive

`[nlogL,AVAR] = gamlike(params,data)` also returns `AVAR`, which is the asymptotic variance-covariance matrix of the parameter estimates when the values in `params` are the maximum likelihood estimates. `AVAR` is the inverse of Fisher's information matrix. The diagonal elements of `AVAR` are the asymptotic variances of their respective parameters.

`[...] = gamlike(params,data,censoring)` accepts a Boolean vector of the same size as `data` that is 1 for observations that are right-censored and 0 for observations that are observed exactly.

`[...] = gamfit(params,data,censoring,freq)` accepts a frequency vector of the same size as `data`. `freq` typically contains integer frequencies for the corresponding elements in `data`, but may contain any non-negative values.

`gamlike` is a utility function for maximum likelihood estimation of the gamma distribution. Since `gamlike` returns the negative gamma log-likelihood function, minimizing `gamlike` using `fminsearch` is the same as maximizing the likelihood.

Examples

Compute the negative log-likelihood of parameter estimates computed by the `gamfit` function:

```
a = 2; b = 3;
r = gamrnd(a,b,100,1);

[nlogL,AVAR] = gamlike(gamfit(r),r)
nlogL =
  267.5648
AVAR =
  0.0788  -0.1104
 -0.1104  0.1955
```

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gamcdf` | `gamfit` | `gaminv` | `gampdf` | `gamrnd` | `gamstat`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gampdf

Gamma probability density function

Syntax

```
y = gampdf(x,a)
y = gampdf(x,a,b)
```

Description

`y = gampdf(x,a)` returns the probability density function (pdf) of the standard gamma distribution with the shape parameter `a`, evaluated at the values in `x`.

`y = gampdf(x,a,b)` returns the pdf of the gamma distribution with the shape parameter `a` and the scale parameter `b`, evaluated at the values in `x`.

Examples

Compute Gamma pdf

Compute the density of the observed value `5` in the standard gamma distribution with shape parameter `2`.

```
y1 = gampdf(5,2)
```

```
y1 = 0.0337
```

Compute the density of the observed value `5` in the gamma distributions with shape parameter `2` and scale parameters `1` through `5`.

```
y2 = gampdf(5,2,1:5)
```

```
y2 = 1×5

    0.0337    0.1026    0.1049    0.0895    0.0736
```

Input Arguments

x — Values at which to evaluate pdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the pdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gampdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

`a` — Shape parameter
positive scalar value | array of positive scalar values

Shape parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gampdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

`b` — Scale parameter
1 (default) | positive scalar value | array of positive scalar values

Scale parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `gampdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 1 2 2]`

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x`, `a`, and `b` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

More About

Gamma pdf

The gamma distribution is a two-parameter family of curves. The parameters a and b are shape and scale, respectively.

The gamma pdf is

$$y = f(x|a,b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{\frac{-x}{b}},$$

where $\Gamma(\cdot)$ is the Gamma function.

The standard gamma distribution occurs when $b = 1$.

For more information, see “Gamma Distribution” on page B-47.

Alternative Functionality

- `gampdf` is a function specific to the gamma distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create a `GammaDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `gampdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GammaDistribution` | `gamcdf` | `gamfit` | `gaminv` | `gamlike` | `gamrnd` | `gamstat` | `pdf`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gamrnd

Gamma random numbers

Syntax

```
r = gamrnd(a,b)
r = gamrnd(a,b,sz1,...,szN)
r = gamrnd(a,b,sz)
```

Description

`r = gamrnd(a,b)` generates a random number from the gamma distribution with the shape parameter `a` and the scale parameter `b`.

`r = gamrnd(a,b,sz1,...,szN)` generates an array of random numbers from the gamma distribution, where `sz1,...,szN` indicates the size of each dimension.

`r = gamrnd(a,b,sz)` generates an array of random numbers from the gamma distribution, where vector `sz` specifies `size(r)`.

Examples

Generate Gamma Random Number

Generate a single random number from the gamma distribution with shape 5 and scale 7.

```
r = gamrnd(5,7)
```

```
r = 68.9857
```

Generate Array of Gamma Random Numbers

Generate five random numbers from the gamma distributions with shape parameter values 1 through 5 and scale parameter 2.

```
a1 = 1:5;
b1 = 2;
r1 = gamrnd(a1,b1)
```

```
r1 = 1×5

    7.1297    6.0918    2.1010    8.7253   29.5447
```

By default, `gamrnd` generates an array that is the same size as `a` and `b` after any necessary scalar expansion so that all scalars are expanded to match the dimensions of the other inputs.

If you specify array dimensions `sz1,...,szN` or `sz`, they must match the dimensions of `a` and `b` after any necessary scalar expansion.

Generate a 2-by-3 array of random numbers from the gamma distribution with shape parameter `3` and scale parameter `7`.

```
sz = [2 3];
r2 = gamrnd(3,7,sz)
```

```
r2 = 2×3

   17.9551   41.3983    7.9865
   16.4204   40.0048   44.1909
```

Generate six random numbers from the gamma distributions with shape parameter values `1` through `6` and scale parameter values `5` through `10` respectively.

```
a3 = 1:6;
b3 = 5:10;
r3 = gamrnd(a3,b3,1,6)
```

```
r3 = 1×6

    9.5930    7.8289   11.0360   15.0367   28.1456   98.2664
```

Input Arguments

a — Shape parameter

positive scalar value | array of positive scalar values

Shape parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If both `a` and `b` are arrays, then the array sizes must be the same. If either `a` or `b` is a scalar, then `gamrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

b — Scale parameter

positive scalar value | array of positive scalar values

Scale parameter of the gamma distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If both `a` and `b` are arrays, then the array sizes must be the same. If either `a` or `b` is a scalar, then `gamrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[1 1 2 2]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as separate arguments of integers.

If either `a` or `b` is an array, then the specified dimensions `sz1,...,szN` must match the common dimensions of `a` and `b` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `gamrnd` ignores trailing dimensions with a size of 1. For example, `gamrnd(2,5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the gamma distribution with shape 2 and scale 5.

Example: `2,4`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If either `a` or `b` is an array, then the specified dimensions `sz` must match the common dimensions of `a` and `b` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `gamrnd` ignores trailing dimensions with a size of 1. For example, `gamrnd(2,5,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the gamma distribution with shape 2 and scale 5.

Example: `[2 4]`

Data Types: `single` | `double`

Output Arguments

`r` — Gamma random numbers
nonnegative scalar value | array of nonnegative scalar values

Gamma random numbers, returned as a nonnegative scalar value or an array of nonnegative scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Alternative Functionality

- `gamrnd` is a function specific to the gamma distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use

`random`, create a `GammaDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `gamrnd` is faster than the generic function `random`.

- Use `randg` to generate random numbers from the standard gamma distribution (unit scale).
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

References

[1] Marsaglia, George, and Wai Wan Tsang. "A Simple Method for Generating Gamma Variables." *ACM Transactions on Mathematical Software* 26, no. 3 (September 1, 2000): 363–72. https://doi.org/10.1145/358407.358414.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`GammaDistribution` | `gamcdf` | `gamfit` | `gaminv` | `gamlike` | `gampdf` | `gamstat` | `randg` | `random`

Topics

"Gamma Distribution" on page B-47

Introduced before R2006a

gamstat

Gamma mean and variance

Syntax

```
[M,V] = gamstat(A,B)
```

Description

`[M,V] = gamstat(A,B)` returns the mean of and variance for the gamma distribution with shape parameters in `A` and scale parameters in `B`. `A` and `B` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `M` and `V`. A scalar input for `A` or `B` is expanded to a constant array with the same dimensions as the other input. The parameters in `A` and `B` must be positive.

The mean of the gamma distribution with parameters A and B is AB. The variance is AB^2.

Examples

```
[m,v] = gamstat(1:5,1:5)
m =
   1   4   9  16  25
v =
   1   8  27  64  125

[m,v] = gamstat(1:5,1./(1:5))
m =
   1   1   1   1   1
v =
  1.0000  0.5000  0.3333  0.2500  0.2000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gamcdf` | `gamfit` | `gaminv` | `gamlike` | `gampdf` | `gamrnd`

Topics

“Gamma Distribution” on page B-47

Introduced before R2006a

gather

Gather properties of machine learning model from GPU

Syntax

```
gatheredMdl = gather(mdl)
[gatheredMdl1,gatheredMdl2,...,gatheredMdln] = gather(mdl1,mdl2,...,mdln)
```

Description

`gatheredMdl = gather(mdl)` gathers all properties of the input regression or classification model `mdl` and returns the gathered model `gatheredMdl`. All properties of the output model are stored in the local workspace.

Use `gather` to create a machine learning model with properties stored in the local workspace from a model fitted using data stored as a GPU array. For more details on GPU arrays, see `gpuArray`.

`[gatheredMdl1,gatheredMdl2,...,gatheredMdln] = gather(mdl1,mdl2,...,mdln)` gathers the properties of multiple models `mdl1,mdl2,...,mdln` and returns the corresponding gathered models `gatheredMdl1,gatheredMdl2,...,gatheredMdln`. The number of input arguments and output arguments must match.

Examples

Gather Properties of Linear Regression Model

Gather the properties of a linear regression model fitted with GPU array data.

Load the `carsmall` data set. Create `X` as a numeric matrix that contains three car performance metrics. Create `Y` as a numeric vector that contains the corresponding miles per gallon.

```
load carsmall
X = [Weight,Horsepower,Acceleration];
Y = MPG;
```

Convert the predictor `X` and response `Y` to `gpuArray` (Parallel Computing Toolbox) objects.

```
X = gpuArray(X);
Y = gpuArray(Y);
```

Fit a linear regression model `mdl` by using `fitlm`.

```
mdl = fitlm(X,Y);
```

Display the coefficients of `mdl` and determine whether the estimated coefficient values are GPU arrays.

```
mdl.Coefficients
```

```
ans=4×4 table
                    Estimate        SE          tStat        pValue
```

```
                    ____________    _________    _________    _____________

    (Intercept)         47.977        3.8785        12.37     4.8957e-21
    x1                -0.0065416    0.0011274      -5.8023     9.8742e-08
    x2                 -0.042943     0.024313      -1.7663        0.08078
    x3                 -0.011583      0.19333     -0.059913        0.95236
```

```
isgpuarray(mdl.Coefficients.Estimate)
```

```
ans = logical
   1
```

Gather the properties of the linear regression model.

```
gatheredMdl = gather(mdl);
```

Display the coefficients of `gatheredMdl` and determine whether the estimated coefficient values are GPU arrays.

```
gatheredMdl.Coefficients
```

```
ans=4×4 table
                    Estimate        SE          tStat        pValue
                    ____________    _________    _________    _____________

    (Intercept)         47.977        3.8785        12.37     4.8957e-21
    x1                -0.0065416    0.0011274      -5.8023     9.8742e-08
    x2                 -0.042943     0.024313      -1.7663        0.08078
    x3                 -0.011583      0.19333     -0.059913        0.95236
```

```
isgpuarray(gatheredMdl.Coefficients.Estimate)
```

```
ans = logical
   0
```

Gather Properties of Multiple Models

Gather the properties of a linear regression model and a *k*-nearest neighbor classifier. Both models are fitted using GPU array data.

Load the `carsmall` data set. Create `X` as a numeric matrix that contains three car performance metrics, and convert the predictor `X` to a `gpuArray` object.

```
load carsmall
X = [Weight,Horsepower,Acceleration];
X = gpuArray(X);
```

Fit a linear regression model of `MPG` (miles per gallon) as a function of the predictor `X`.

```
mdlLinear = fitlm(X,MPG);
```

Train a 3-nearest neighbor classifier using the predictor `X` and the classes `Cylinders`. Standardize the noncategorical predictor data.

```
mdlKNN = fitcknn(X,Cylinders,'NumNeighbors',3,'Standardize',1);
```

Gather the properties of the `mdLinear` and `mdlKNN` models.

```
[gMdlLinear,gMdlKNN] = gather(mdlLinear,mdlKNN);
```

Determine whether the *p*-value of the Durbin-Watson test for the regression model `mdlLinear` is a GPU array.

```
isgpuarray(dwtest(mdlLinear))
```

```
ans = logical
   1
```

Determine whether the *p*-value of the Durbin-Watson test for the gathered regression model `gMdlLinear` is a GPU array.

```
isgpuarray(dwtest(gMdlLinear))
```

```
ans = logical
   0
```

Determine whether the resubstitution loss of the classifier `mdlKNN` is a GPU array.

```
isgpuarray(resubLoss(mdlKNN))
```

```
ans = logical
   1
```

Determine whether the resubstitution loss of the gathered classifier `gMdlKNN` is a GPU array.

```
isgpuarray(resubLoss(gMdlKNN))
```

```
ans = logical
   1
```

Input Arguments

`mdl` — Machine learning model fitted with GPU arrays

regression model object | classification model object

Machine learning model fitted with GPU arrays, specified as a regression or classification model object, as given in the following table of supported models.

Model Object Name	Model Description	Model Creation Function
`LinearModel`	Full linear regression model	`fitlm`
`CompactLinearModel`	Compact linear regression model	`LinearModel` object function `compact`
`GeneralizedLinearModel`	Full generalized linear regression model	`fitglm`

Model Object Name	Model Description	Model Creation Function
`CompactGeneralizedLinearModel`	Compact generalized linear regression model	`GeneralizedLinearModel` object function `compact`
`ClassificationKNN`	*k*-nearest neighbor classification model	`fitcknn`

If you want to create a compact model fitted with GPU arrays, the input argument `mdl` of `compact` must be a full model object fitted with GPU array input arguments.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationKNN` | `CompactGeneralizedLinearModel` | `CompactLinearModel` | `GeneralizedLinearModel` | `LinearModel` | `gather` | `gpuArray`

Topics

“Linear Regression” on page 11-9
“Generalized Linear Models” on page 12-9
“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2020b

ge

Class: `qrandstream`

Greater than or equal relation for handles

Syntax

```
h1 >= h2
```

Description

`h1 >= h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise `>=` result.

If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = ge(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`eq` | `gt` | `le` | `lt` | `ne` | `qrandstream`

GeneralizedLinearMixedModel class

Generalized linear mixed-effects model class

Description

A `GeneralizedLinearMixedModel` object represents a regression model of a response variable that contains both fixed and random effects. The object comprises data, a model description, fitted coefficients, covariance parameters, design matrices, residuals, residual plots, and other diagnostic information for a generalized linear mixed-effects (GLME) model. You can predict model responses with the `predict` function and generate random data at new design points using the `random` function.

Construction

You can fit a generalized linear mixed-effects (GLME) model to sample data using `fitglme(tbl, formula)`. For more information, see `fitglme`.

Input Arguments

`tbl` — Input data

table | dataset array

Input data, which includes the response variable, predictor variables, and grouping variables, specified as a table or dataset array. The predictor variables can be continuous or grouping variables (see “Grouping Variables” on page 2-45). You must specify the model for the variables using `formula`.

Data Types: `table`

`formula` — Formula for model specification

character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`

Formula for model specification, specified as a character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`. For a full description, see Formula on page 33-2587.

Example: `'y ~ treatment +(1|block)'`

Properties

`Coefficients` — Estimates of fixed-effects coefficients

dataset array

Estimates of fixed-effects coefficients and related statistics, stored as a dataset array that has one row for each coefficient and the following columns:

- `Name` — Name of the coefficient
- `Estimate` — Estimated coefficient value

- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is equal to 0
- `DF` — Degrees of freedom associated with the *t* statistic
- `pValue` — *p*-value for the *t*-statistic
- `Lower` — Lower confidence limit
- `Upper` — Upper confidence limit

To obtain any of these columns as a vector, index into the property using dot notation.

Use the `coefTest` method to perform other tests on the coefficients.

`CoefficientCovariance` — Covariance of estimated fixed-effects vector
matrix

Covariance of estimated fixed-effects vector, stored as a matrix.

Data Types: `single` | `double`

`CoefficientNames` — Names of fixed-effects coefficients
cell array of character vectors

Names of fixed-effects coefficients, stored as a cell array of character vectors. The label for the coefficient of the constant term is `(Intercept)`. The labels for other coefficients indicate the terms that they multiply. When the term includes a categorical predictor, the label also indicates the level of that predictor.

Data Types: `cell`

`DFE` — Degrees of freedom for error
positive integer value

Degrees of freedom for error, stored as a positive integer value. `DFE` is the number of observations minus the number of estimated coefficients.

`DFE` contains the degrees of freedom corresponding to the `'Residual'` method of calculating denominator degrees of freedom for hypothesis tests on fixed-effects coefficients. If n is the number of observations and p is the number of fixed-effects coefficients, then `DFE` is equal to $n - p$.

Data Types: `double`

`Dispersion` — Model dispersion parameter
scalar value

Model dispersion parameter, stored as a scalar value. The dispersion parameter defines the conditional variance of the response.

For observation i, the conditional variance of the response y_i, given the conditional mean μ_i and the dispersion parameter σ^2, in a generalized linear mixed-effects model is

$$\mathrm{var}\left(y_i \middle| \mu_i, \sigma^2\right) = \frac{\sigma^2}{w_i} v(\mu_i),$$

where w_i is the ith observation weight and v is the variance function for the specified conditional distribution of the response. The `Dispersion` property contains an estimate of σ^2 for the specified

GLME model. The value of `Dispersion` depends on the specified conditional distribution of the response. For binomial and Poisson distributions, the theoretical value of `Dispersion` is equal to $\sigma^2 = 1.0$.

- If `FitMethod` is `MPL` or `REMPL` and the `'DispersionFlag'` name-value pair argument in `fitglme` is `true`, then a dispersion parameter is estimated from data for all distributions, including binomial and Poisson distributions.
- If `FitMethod` is `ApproximateLaplace` or `Laplace`, then the `'DispersionFlag'` name-value pair argument in `fitglme` does not apply, and the dispersion parameter is fixed at 1.0 for binomial and Poisson distributions. For all other distributions, `Dispersion` is estimated from data.

Data Types: `double`

`DispersionEstimated` — Flag indicating if dispersion parameter was estimated
`true` | `false`

Flag indicating estimated dispersion parameter, stored as a logical value.

- If `FitMethod` is `ApproximateLaplace` or `Laplace`, then the dispersion parameter is fixed at its theoretical value of 1.0 for binomial and Poisson distributions, and `DispersionEstimated` is `false`. For other distributions, the dispersion parameter is estimated from the data, and `DispersionEstimated` is `true`.
- If `FitMethod` is `MPL` or `REMPL`, and the `'DispersionFlag'` name-value pair argument in `fitglme` is specified as `true`, then the dispersion parameter is estimated for all distributions, including binomial and Poisson distributions, and `DispersionEstimated` is `true`.
- If `FitMethod` is `MPL` or `REMPL`, and the `'DispersionFlag'` name-value pair argument in `fitglme` is specified as `false`, then the dispersion parameter is fixed at its theoretical value for binomial and Poisson distributions, and `DispersionEstimated` is `false`. For distributions other than binomial and Poisson, the dispersion parameter is estimated from the data, and `DispersionEstimated` is `true`.

Data Types: `logical`

`Distribution` — Response distribution name
`'Normal'` | `'Binomial'` | `'Poisson'` | `'Gamma'` | `'InverseGaussian'`

Response distribution name, stored as one of the following:

- `'Normal'` — Normal distribution
- `'Binomial'` — Binomial distribution
- `'Poisson'` — Poisson distribution
- `'Gamma'` — Gamma distribution
- `'InverseGaussian'` — Inverse Gaussian distribution

`FitMethod` — Method used to fit the model
`'MPL'` | `'REMPL'` | `'ApproximateLaplace'` | `'Laplace'`

Method used to fit the model, stored as one of the following.

- `'MPL'` — Maximum pseudo likelihood
- `'REMPL'` — Restricted maximum pseudo likelihood

- `'ApproximateLaplace'` — Maximum likelihood using the approximate Laplace method, with fixed effects profiled out
- `'Laplace'` — Maximum likelihood using the Laplace method

`Formula` — Model specification formula

object

Model specification formula, stored as an object. The model specification formula uses Wilkinson's notation to describe the relationship between the fixed-effects terms, random-effects terms, and grouping variables in the GLME model. For more information see Formula on page 33-2587.

`Link` — Link function characteristics

structure

Link function characteristics, stored as a structure containing the following fields. The link is a function `G` that links the distribution parameter `MU` to the linear predictor `ETA` as follows: `G(MU) = ETA`.

Field	Description
`Name`	Name of the link function
`Link`	Function that defines `G`
`Derivative`	Derivative of `G`
`SecondDerivative`	Second derivative of `G`
`Inverse`	Inverse of `G`

Data Types: `struct`

`LogLikelihood` — Log of likelihood function

scalar value

Log of likelihood function evaluated at the estimated coefficient values, stored as a scalar value. `LogLikelihood` depends on the method used to fit the model.

- If you use `'Laplace'` or `'ApproximateLaplace'`, then `LogLikelihood` is the maximized log likelihood.
- If you use `'MPL'`, then `LogLikelihood` is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration.
- If you use `'REMPL'`, then `LogLikelihood` is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.

Data Types: `double`

`ModelCriterion` — Model criterion

table

Model criterion to compare fitted generalized linear mixed-effects models, stored as a table with the following fields.

Field	Description
`AIC`	Akaike information criterion

Field	Description
BIC	Bayesian information criterion
LogLikelihood	• For a model fit using 'Laplace' or 'ApproximateLaplace', LogLikelihood is the maximized log likelihood. • For a model fit using 'MPL', LogLikelihood is the maximized log likelihood of the pseudo data from the final pseudo likelihood iteration. • For a model fit using 'REMPL', LogLikelihood is the maximized restricted log likelihood of the pseudo data from the final pseudo likelihood iteration.
Deviance	–2 times LogLikelihood

NumCoefficients — Number of fixed-effects coefficients
positive integer value

Number of fixed-effects coefficients in the fitted generalized linear mixed-effects model, stored as a positive integer value.

Data Types: double

NumEstimatedCoefficients — Number of estimated fixed-effects coefficients
positive integer value

Number of estimated fixed-effects coefficients in the fitted generalized linear mixed-effects model, stored as a positive integer value.

Data Types: double

NumObservations — Number of observations
positive integer value

Number of observations used in the fit, stored as a positive integer value. NumObservations is the number of rows in the table or dataset array tbl, minus rows excluded using the 'Exclude' name-value pair of fitglme or rows containing NaN values.

Data Types: double

NumPredictors — Number of predictors
positive integer value

Number of variables used as predictors in the generalized linear mixed-effects model, stored as a positive integer value.

Data Types: double

NumVariables — Total number of variables
positive integer value

Total number of variables, including the response and predictors, stored as a positive integer value. If the sample data is in a table or dataset array tbl, then NumVariables is the total number of variables in tbl, including the response variable. NumVariables includes variables, if any, that are not used as predictors or as the response.

Data Types: `double`

`ObservationInfo` — Information about the observations

table

Information about the observations used in the fit, stored as a table.

`ObservationInfo` has one row for each observation and the following columns.

Name	Description
`Weights`	The weight value for the observation. The default value is 1.
`Excluded`	If the observation was excluded from the fit using the `'Exclude'` name-value pair argument in `fitglme`, then `Excluded` is `true`, or `1`. Otherwise, `Excluded` is `false`, or `0`.
`Missing`	If the observation was excluded from the fit because any response or predictor value is missing, then `Missing` is `true`. Otherwise, `Missing` is `false`. Missing values include `NaN` for numeric variables, empty cells for cell arrays, blank rows for character arrays, and the `<undefined>` value for categorical arrays.
`Subset`	If the observation was used in the fit, then `Subset` is `true`. If the observation was not used in the fit because it is missing or excluded, then `Subset` is `false`.
`BinomSize`	Binomial size for each observation. This column only applies when fitting a binomial distribution.

Data Types: `table`

`ObservationNames` — Names of observations

cell array of character vectors

Names of observations used in the fit, stored as a cell array of character vectors.

- If the data is in a table or dataset array `tbl` that contains observation names, then `ObservationNames` uses those names.
- If the data is provided in matrices, or in a table or dataset array without observation names, then `ObservationNames` is an empty cell array.

Data Types: `cell`

`PredictorNames` — Names of predictors

cell array of character vectors

Names of the variables used as predictors in the fit, stored as a cell array of character vectors that has the same length as `NumPredictors`.

Data Types: `cell`

ResponseName — Name of response variable
character vector

Name of the variable used as the response variable in the fit, stored as a character vector.

Data Types: `char`

Rsquared — Proportion of variability in the response explained by the fitted model
structure

Proportion of variability in the response explained by the fitted model, stored as a structure. `Rsquared` contains the *R*-squared value of the fitted model, also known as the multiple correlation coefficient. `Rsquared` contains the following fields.

Field	Description
`Ordinary`	R-squared value, stored as a scalar value in a structure. `Rsquared.Ordinary = 1 − SSE./SST`
`Adjusted`	R-squared value adjusted for the number of fixed-effects coefficients, stored as a scalar value in a structure. `Rsquared.Adjusted = 1 − (SSE./SST)*(DFT./DFE)`, where `DFE = n − p`, `DFT = n − 1`, `n` is the total number of observations, and `p` is the number of fixed-effects coefficients.

Data Types: `struct`

SSE — Error sum of squares
positive scalar value

Error sum of squares, stored as a positive scalar value. `SSE` is the weighted sum of the squared conditional residuals, and is calculated as

$$SSE = \sum_{i=1}^{n} w_i^{eff}(y_i - f_i)^2,$$

where n is the number of observations, w_i^{eff} is the ith effective weight, y_i is the ith response, and f_i is the ith fitted value.

The ith effective weight is calculated as

$$w_i^{eff} = \left\{ \frac{w_i}{v_i\left(\mu_i\left(\widehat{\beta}, \widehat{b}\right)\right)} \right\},$$

where v_i is the variance term for the ith observation, $\widehat{\beta}$ and $\widehat{b}$ are estimated values of β and b, respectively.

The ith fitted value is calculated as

$$f_i = g^{-1}\left(x_i^T\widehat{\beta} + z_i^T\widehat{b} + \delta_i\right),$$

where x_i^T is the ith row of the fixed-effects design matrix X, and z_i^T is the ith row of the random-effects design matrix Z. δ_i is the ith offset value.

Data Types: `double`

SSR — Regression sum of squares
positive scalar value

Regression sum of squares, stored as a positive scalar value. SSR is the sum of squares explained by the generalized linear mixed-effects regression, or equivalently the weighted sum of the squared deviations of the conditional fitted values from their weighted mean. SSR is calculated as

$$SSR = \sum_{i=1}^{N} w_i^{eff}\left(f_i - \bar{f}\right)^2,$$

where n is the number of observations, w_i^{eff} is the ith effective weight, f_i is the ith fitted value, and $\bar{f}$ is a weighted average of the fitted values.

The ith effective weight is calculated as

$$w_i^{eff} = \left\{\frac{w_i}{v_i\left(\mu_i\left(\widehat{\beta}, \widehat{b}\right)\right)}\right\},$$

where $\widehat{\beta}$ and $\widehat{b}$ are estimated values of β and b, respectively.

The ith fitted value is calculated as

$$f_i = g^{-1}\left(x_i^T\widehat{\beta} + z_i^T\widehat{b} + \delta_i\right),$$

where x_i^T is the ith row of the fixed-effects design matrix X, and z_i^T is the ith row of the random-effects design matrix Z. δ_i is the ith offset value.

The weighted average of fitted values is calculated as

$$\bar{f} = \frac{\left[\sum_{i=1}^{n} w_i^{eff} f_i\right]}{\sum_{i=1}^{n} w_i^{eff}}.$$

Data Types: `double`

SST — Total sum of squares
positive scalar value

Total sum of squares, stored as a positive scalar value. For a GLME model, SST is defined as SST = SSE + SSR.

Data Types: `double`

`VariableInfo` — Information about the variables
table

Information about the variables used in the fit, stored as a table. `VariableInfo` has one row for each variable and contains the following columns.

Column Name	Description
`Class`	Class of the variable (`'double'`, `'cell'`, `'nominal'`, and so on).
`Range`	Value range of the variable. • For a numerical variable, `Range` is a two-element vector of the form `[min,max]`. • For a cell or categorical variable, `Range` is a cell or categorical array containing all unique values of the variable.
`InModel`	If the variable is a predictor in the fitted model, `InModel` is `true`. If the variable is not in the fitted model, `InModel` is `false`.
`IsCategorical`	If the variable type is treated as a categorical predictor (such as cell, logical, or categorical), then `IsCategorical` is `true`. If the variable is a continuous predictor, then `IsCategorical` is `false`.

Data Types: `table`

`VariableNames` — Names of the variables
cell array of character vectors

Names of all the variables contained in the table or dataset array `tbl`, stored as a cell array of character vectors.

Data Types: `cell`

`Variables` — Variables
table

Variables, stored as a table. If the fit is based on a table or dataset array `tbl`, then `Variables` is identical to `tbl`.

Data Types: `table`

Object Functions

anova	Analysis of variance for generalized linear mixed-effects model
coefCI	Confidence intervals for coefficients of generalized linear mixed-effects model
coefTest	Hypothesis test on fixed and random effects of generalized linear mixed-effects model
compare	Compare generalized linear mixed-effects models
covarianceParameters	Extract covariance parameters of generalized linear mixed-effects model
designMatrix	Fixed- and random-effects design matrices

fitted	Fitted responses from generalized linear mixed-effects model
fixedEffects	Estimates of fixed effects and related statistics
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotResiduals	Plot residuals of generalized linear mixed-effects model
predict	Predict response of generalized linear mixed-effects model
random	Generate random responses from fitted generalized linear mixed-effects model
randomEffects	Estimates of random effects and related statistics
refit	Refit generalized linear mixed-effects model
residuals	Residuals of fitted generalized linear mixed-effects model
response	Response vector of generalized linear mixed-effects model

Examples

Fit a Generalized Linear Mixed-Effects Model

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)', ..
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Display the model.

```
disp(glme)
```

```
Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    416.35    434.58    -201.17          402.35

Fixed effects coefficients (95% CIs):
    Name                   Estimate     SE          tStat       DF    pValue
    {'(Intercept)'}           1.4689     0.15988      9.1875    94    9.8194e-15
    {'newprocess' }         -0.36766     0.17755     -2.0708    94      0.041122
    {'time_dev'   }        -0.094521     0.82849    -0.11409    94       0.90941
    {'temp_dev'   }         -0.28317      0.9617    -0.29444    94       0.76907
    {'supplier_C' }        -0.071868    0.078024     -0.9211    94       0.35936
    {'supplier_B' }         0.071072     0.07739     0.91836    94       0.36078


    Lower        Upper
       1.1515       1.7864
```

```
     -0.72019    -0.015134
      -1.7395       1.5505
      -2.1926       1.6263
     -0.22679     0.083051
    -0.082588      0.22473

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1                Name2                Type           Estimate
    {'(Intercept)'}      {'(Intercept)'}      {'std'}        0.31381

Group: Error
    Name                       Estimate
    {'sqrt(Dispersion)'}       1
```

The `Model information` table displays the total number of observations in the sample data (100), the number of fixed- and random-effects coefficients (6 and 20, respectively), and the number of covariance parameters (1). It also indicates that the response variable has a `Poisson` distribution, the link function is `Log`, and the fit method is `Laplace`.

`Formula` indicates the model specification using Wilkinson's notation.

The `Model fit statistics` table displays statistics used to assess the goodness of fit of the model. This includes the Akaike information criterion (`AIC`), Bayesian information criterion (`BIC`) values, log likelihood (`LogLikelihood`), and deviance (`Deviance`) values.

The `Fixed effects coefficients` table indicates that `fitglme` returned 95% confidence intervals. It contains one row for each fixed-effects predictor, and each column contains statistics corresponding to that predictor. Column 1 (`Name`) contains the name of each fixed-effects coefficient, column 2 (`Estimate`) contains its estimated value, and column 3 (`SE`) contains the standard error of the coefficient. Column 4 (`tStat`) contains the *t*-statistic for a hypothesis test that the coefficient is equal to 0. Column 5 (`DF`) and column 6 (`pValue`) contain the degrees of freedom and *p*-value that correspond to the *t*-statistic, respectively. The last two columns (`Lower` and `Upper`) display the lower and upper limits, respectively, of the 95% confidence interval for each fixed-effects coefficient.

`Random effects covariance parameters` displays a table for each grouping variable (here, only `factory`), including its total number of levels (20), and the type and estimate of the covariance parameter. Here, `std` indicates that `fitglme` returns the standard deviation of the random effect associated with the factory predictor, which has an estimated value of 0.31381. It also displays a table containing the error parameter type (here, the square root of the dispersion parameter), and its estimated value of 1.

The standard display generated by `fitglme` does not provide confidence intervals for the random-effects parameters. To compute and display these values, use `covarianceParameters`.

More About

Formula

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For generalized linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms, respectively, and *R* is the number of grouping variables in the model.

Suppose a table `tbl` contains the following:

- A response variable, `y`
- Predictor variables, X_j, which can be continuous or grouping variables
- Grouping variables, g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form, `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random`$_1$ is a specification of the random-effects design matrix Z_1 corresponding to grouping variable g_1, and similarly `random`$_R$ is a specification of the random-effects design matrix Z_R corresponding to grouping variable g_R. You can express the `fixed` and `random` terms using Wilkinson notation.

Wilkinson notation describes the factors present in models. The notation relates to factors present in models, not to the multipliers (coefficients) of those factors.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	`X`, X^2, ..., X^k
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (elementwise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only
`- X2`	Do not include `X2`
`X1*X2 + X3`	`X1`, `X2`, `X3`, `X1*X2`
`X1 + X2 + X3 + X1:X2`	`X1`, `X2`, `X3`, `X1*X2`
`X1*X2*X3 - X1:X2:X3`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`, `X2*X3`
`X1*(X2 + X3)`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`. Here are some examples for linear mixed-effects model specification.

Examples:

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, `X1` and `X2`. This is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept and fixed effects for `X1` and `X2`. The implicit intercept term is suppressed by including `-1`.
`'y ~ 1 + (1 | g1)'`	Fixed effects for the intercept plus random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.

Formula	Description
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

See Also

`fitglme`

Topics

“Fit a Generalized Linear Mixed-Effects Model” on page 12-57
“Generalized Linear Mixed-Effects Models” on page 12-48

GeneralizedLinearModel

Generalized linear regression model class

Description

`GeneralizedLinearModel` is a fitted generalized linear regression model. A generalized linear regression model is a special class of nonlinear models that describe a nonlinear relationship between a response and predictors. A generalized linear regression model has generalized characteristics of a linear regression model. The response variable follows a normal, binomial, Poisson, gamma, or inverse Gaussian distribution with parameters including the mean response μ. A link function f defines the relationship between μ and the linear combination of predictors.

Use the properties of a `GeneralizedLinearModel` object to investigate a fitted generalized linear regression model. The object properties include information about coefficient estimates, summary statistics, fitting method, and input data. Use the object functions to predict responses and to modify, evaluate, and visualize the model.

Creation

Create a `GeneralizedLinearModel` object by using `fitglm` or `stepwiseglm`.

`fitglm` fits a generalized linear regression model to data using a fixed model specification. Use `addTerms`, `removeTerms`, or `step` to add or remove terms from the model. Alternatively, use `stepwiseglm` to fit a model using stepwise generalized linear regression.

Properties

Coefficient Estimates

`CoefficientCovariance` — Covariance matrix of coefficient estimates
numeric matrix

This property is read-only.

Covariance matrix of coefficient estimates, specified as a p-by-p matrix of numeric values. p is the number of coefficients in the fitted model.

For details, see “Coefficient Standard Errors and Confidence Intervals” on page 11-58.

Data Types: `single` | `double`

`CoefficientNames` — Coefficient names
cell array of character vectors

This property is read-only.

Coefficient names, specified as a cell array of character vectors, each containing the name of the corresponding term.

Data Types: `cell`

`Coefficients` — Coefficient values

table

This property is read-only.

Coefficient values, specified as a table. `Coefficients` contains one row for each coefficient and these columns:

- `Estimate` — Estimated coefficient value
- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is zero
- `pValue` — *p*-value for the *t*-statistic

Use `coefTest` to perform linear hypothesis tests on the coefficients. Use `coefCI` to find the confidence intervals of the coefficient estimates.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the estimated coefficient vector in the model `mdl`:

```
beta = mdl.Coefficients.Estimate
```

Data Types: `table`

`NumCoefficients` — Number of model coefficients

positive integer

This property is read-only.

Number of model coefficients, specified as a positive integer. `NumCoefficients` includes coefficients that are set to zero when the model terms are rank deficient.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated coefficients

positive integer

This property is read-only.

Number of estimated coefficients in the model, specified as a positive integer. `NumEstimatedCoefficients` does not include coefficients that are set to zero when the model terms are rank deficient. `NumEstimatedCoefficients` is the degrees of freedom for regression.

Data Types: `double`

Summary Statistics

`Deviance` — Deviance of fit

numeric value

This property is read-only.

Deviance of the fit, specified as a numeric value. The deviance is useful for comparing two models when one model is a special case of the other model. The difference between the deviance of the two models has a chi-square distribution with degrees of freedom equal to the difference in the number of

estimated parameters between the two models. For more information, see “Deviance” on page 33-2606.

Data Types: `single` | `double`

`DFE` — Degrees of freedom for error
positive integer

This property is read-only.

Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients, specified as a positive integer.

Data Types: `double`

`Diagnostics` — Observation diagnostics
table

This property is read-only.

Observation diagnostics, specified as a table that contains one row for each observation and the columns described in this table.

Column	Meaning	Description
`Leverage`	Diagonal elements of `HatMatrix`	`Leverage` for each observation indicates to what extent the fit is determined by the observed predictor values. A value close to `1` indicates that the fit is largely determined by that observation, with little contribution from the other observations. A value close to `0` indicates that the fit is largely determined by the other observations. For a model with `P` coefficients and `N` observations, the average value of `Leverage` is `P/N`. A `Leverage` value greater than `2*P/N` indicates high leverage.
`CooksDistance`	Cook's distance of scaled change in fitted values	`CooksDistance` is a measure of scaled change in fitted values. An observation with `CooksDistance` greater than three times the mean Cook's distance can be an outlier.
`HatMatrix`	Projection matrix to compute fitted from observed responses	`HatMatrix` is an `N`-by-`N` matrix such that `Fitted = HatMatrix*Y`, where `Y` is the response vector and `Fitted` is the vector of fitted response values.

The software computes these values on the scale of the linear combination of the predictors, stored in the `LinearPredictor` field of the `Fitted` and `Residuals` properties. For example, the software computes the diagnostic values by using the fitted response and adjusted response values from the model `mdl`.

```
Yfit = mdl.Fitted.LinearPredictor
Yadjusted = mdl.Fitted.LinearPredictor + mdl.Residuals.LinearPredictor
```

`Diagnostics` contains information that is helpful in finding outliers and influential observations. For more details, see “Leverage” on page 33-2605, “Cook’s Distance” on page 33-2604, and “Hat Matrix” on page 33-2605.

Use `plotDiagnostics` to plot observation diagnostics.

Rows not used in the fit because of missing values (in `ObservationInfo.Missing`) or excluded values (in `ObservationInfo.Excluded`) contain `NaN` values in the `CooksDistance` column and zeros in the `Leverage` and `HatMatrix` columns.

To obtain any of these columns as an array, index into the property using dot notation. For example, obtain the hat matrix in the model `mdl`:

```
HatMatrix = mdl.Diagnostics.HatMatrix;
```

Data Types: `table`

`Dispersion` — Scale factor of variance of response
numeric scalar

This property is read-only.

Scale factor of the variance of the response, specified as a numeric scalar.

If the `'DispersionFlag'` name-value pair argument of `fitglm` or `stepwiseglm` is `true`, then the function estimates the `Dispersion` scale factor in computing the variance of the response. The variance of the response equals the theoretical variance multiplied by the scale factor.

For example, the variance function for the binomial distribution is $p(1–p)/n$, where p is the probability parameter and n is the sample size parameter. If `Dispersion` is near `1`, the variance of the data appears to agree with the theoretical variance of the binomial distribution. If `Dispersion` is larger than `1`, the data set is “overdispersed” relative to the binomial distribution.

Data Types: `double`

`DispersionEstimated` — Flag to indicate use of dispersion scale factor
logical value

This property is read-only.

Flag to indicate whether `fitglm` used the `Dispersion` scale factor to compute standard errors for the coefficients in `Coefficients.SE`, specified as a logical value. If `DispersionEstimated` is `false`, `fitglm` used the theoretical value of the variance.

- `DispersionEstimated` can be `false` only for the binomial and Poisson distributions.
- Set `DispersionEstimated` by setting the `'DispersionFlag'` name-value pair argument of `fitglm` or `stepwiseglm`.

Data Types: `logical`

`Fitted` — Fitted response values based on input data
table

This property is read-only.

Fitted (predicted) values based on the input data, specified as a table that contains one row for each observation and the columns described in this table.

Column	Description
`Response`	Predicted values on the scale of the response

Column	Description
`LinearPredictor`	Predicted values on the scale of the linear combination of the predictors (same as the link function applied to the `Response` fitted values)
`Probability`	Fitted probabilities (included only with the binomial distribution)

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the vector `f` of fitted values on the response scale in the model `mdl`:

```
f = mdl.Fitted.Response
```

Use `predict` to compute predictions for other predictor values, or to compute confidence bounds on `Fitted`.

Data Types: `table`

`LogLikelihood` — Loglikelihood
numeric value

This property is read-only.

Loglikelihood of the model distribution at the response values, specified as a numeric value. The mean is fitted from the model, and other parameters are estimated as part of the model fit.

Data Types: `single` | `double`

`ModelCriterion` — Criterion for model comparison
structure

This property is read-only.

Criterion for model comparison, specified as a structure with these fields:

- `AIC` — Akaike information criterion. `AIC = –2*logL + 2*m`, where `logL` is the loglikelihood and `m` is the number of estimated parameters.
- `AICc` — Akaike information criterion corrected for the sample size. `AICc = AIC + (2*m*(m + 1))/(n – m – 1)`, where `n` is the number of observations.
- `BIC` — Bayesian information criterion. `BIC = –2*logL + m*log(n)`.
- `CAIC` — Consistent Akaike information criterion. `CAIC = –2*logL + m*(log(n) + 1)`.

Information criteria are model selection tools that you can use to compare multiple models fit to the same data. These criteria are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). Different information criteria are distinguished by the form of the penalty.

When you compare multiple models, the model with the lowest information criterion value is the best-fitting model. The best-fitting model can vary depending on the criterion used for model comparison.

To obtain any of the criterion values as a scalar, index into the property using dot notation. For example, obtain the AIC value `aic` in the model `mdl`:

```
aic = mdl.ModelCriterion.AIC
```

Data Types: `struct`

`Residuals` — Residuals for fitted model

table

This property is read-only.

Residuals for the fitted model, specified as a table that contains one row for each observation and the columns described in this table.

Column	Description
`Raw`	Observed minus fitted values
`LinearPredictor`	Residuals on the linear predictor scale, equal to the adjusted response value minus the fitted linear combination of the predictors
`Pearson`	Raw residuals divided by the estimated standard deviation of the response
`Anscombe`	Residuals defined on transformed data with the transformation selected to remove skewness
`Deviance`	Residuals based on the contribution of each observation to the deviance

Rows not used in the fit because of missing values (in `ObservationInfo.Missing`) contain `NaN` values.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the ordinary raw residual vector `r` in the model `mdl`:

```
r = mdl.Residuals.Raw
```

Data Types: `table`

`Rsquared` — R-squared value for model

structure

This property is read-only.

R-squared value for the model, specified as a structure with five fields.

Field	Description	Equation
`Ordinary`	Ordinary (unadjusted) R-squared	$R^2_{\text{Ordinary}} = 1 - \frac{\text{SSE}}{\text{SST}}$ `SSE` is the sum of squared errors, and `SST` is the total sum of squared deviations of the response vector from the mean of the response vector.
`Adjusted`	R-squared adjusted for the number of coefficients	$R^2_{\text{Adjusted}} = 1 - \frac{\text{SSE}}{\text{SST}} \cdot \frac{N-1}{\text{DFE}}$ N is the number of observations (`NumObservations`), and `DFE` is the degrees of freedom for the error (residuals).

Field	Description	Equation
`LLR`	Loglikelihood ratio	$R^2_{\text{LLR}} = 1 - \frac{L}{L_0}$ L is the loglikelihood of the fitted model (`LogLikelihood`), and L_0 is the loglikelihood of a model that includes only a constant term. R^2_{LLR} is the McFadden pseudo R-squared value [1] for logistic regression models.
`Deviance`	Deviance R-squared	$R^2_{\text{Deviance}} = 1 - \frac{D}{D_0}$ D is the deviance of the fitted model (`Deviance`), and D_0 is the deviance of a model that includes only a constant term.
`AdjGeneralized`	Adjusted generalized R-squared	$R^2_{\text{AdjGeneralized}} = \frac{1 - \exp\left(\frac{2(L_0 - L)}{N}\right)}{1 - \exp\left(\frac{2L_0}{N}\right)}$ $R^2_{\text{AdjGeneralized}}$ is the Nagelkerke adjustment [2] to a formula proposed by Maddala [3], Cox and Snell [4], and Magee [5] for logistic regression models.

To obtain any of these values as a scalar, index into the property using dot notation. For example, to obtain the adjusted R-squared value in the model `mdl`, enter:

```
r2 = mdl.Rsquared.Adjusted
```

Data Types: `struct`

SSE — Sum of squared errors
numeric value

This property is read-only.

Sum of squared errors (residuals), specified as a numeric value.

Data Types: `single` | `double`

SSR — Regression sum of squares
numeric value

This property is read-only.

Regression sum of squares, specified as a numeric value. The regression sum of squares is equal to the sum of squared deviations of the fitted values from their mean.

Data Types: `single` | `double`

SST — Total sum of squares
numeric value

This property is read-only.

Total sum of squares, specified as a numeric value. The total sum of squares is equal to the sum of squared deviations of the response vector `y` from the `mean(y)`.

Data Types: `single` | `double`

Fitting Information

`Steps` — Stepwise fitting information

structure

This property is read-only.

Stepwise fitting information, specified as a structure with the fields described in this table.

Field	Description
`Start`	Formula representing the starting model
`Lower`	Formula representing the lower bound model. The terms in `Lower` must remain in the model.
`Upper`	Formula representing the upper bound model. The model cannot contain more terms than `Upper`.
`Criterion`	Criterion used for the stepwise algorithm, such as `'sse'`
`PEnter`	Threshold for `Criterion` to add a term
`PRemove`	Threshold for `Criterion` to remove a term
`History`	Table representing the steps taken in the fit

The `History` table contains one row for each step, including the initial fit, and the columns described in this table.

Column	Description
`Action`	Action taken during the step: • `'Start'` — First step • `'Add'` — A term is added • `'Remove'` — A term is removed
`TermName`	• If `Action` is `'Start'`, `TermName` specifies the starting model specification. • If `Action` is `'Add'` or `'Remove'`, `TermName` specifies the term added or removed in the step.
`Terms`	Model specification in a “Terms Matrix” on page 33-2606
`DF`	Regression degrees of freedom after the step
`delDF`	Change in regression degrees of freedom from the previous step (negative for steps that remove a term)
`Deviance`	Deviance (residual sum of squares) at the step (only for a generalized linear regression model)
`FStat`	*F*-statistic that leads to the step
`PValue`	*p*-value of the *F*-statistic

The structure is empty unless you fit the model using stepwise regression.

Data Types: `struct`

Input Data

`Distribution` — Generalized distribution information

structure

This property is read-only.

Generalized distribution information, specified as a structure with the fields described in this table.

Field	Description
`Name`	Name of the distribution: `'normal'`, `'binomial'`, `'poisson'`, `'gamma'`, or `'inverse gaussian'`
`DevianceFunction`	Function that computes the components of the deviance as a function of the fitted parameter values and the response values
`VarianceFunction`	Function that computes the theoretical variance for the distribution as a function of the fitted parameter values. When `DispersionEstimated` is `true`, the software multiplies the variance function by `Dispersion` in the computation of the coefficient standard errors.

Data Types: `struct`

`Formula` — Model information

`LinearFormula` object

This property is read-only.

Model information, specified as a `LinearFormula` object.

Display the formula of the fitted model `mdl` using dot notation:

```
mdl.Formula
```

`Link` — Link function

structure

This property is read-only.

Link function, specified as a structure with the fields described in this table.

Field	Description
`Name`	Name of the link function, specified as a character vector. If you specify the link function using a function handle, then `Name` is `''`.
`Link`	Function f that defines the link function, specified as a function handle
`Derivative`	Derivative of f, specified as a function handle
`Inverse`	Inverse of f, specified as a function handle

The link function is a function f that links the distribution parameter μ to the fitted linear combination Xb of the predictors:

$$f(\mu) = Xb.$$

Data Types: `struct`

`NumObservations` — Number of observations

positive integer

This property is read-only.

Number of observations the fitting function used in fitting, specified as a positive integer. `NumObservations` is the number of observations supplied in the original table, dataset, or matrix, minus any excluded rows (set with the `'Exclude'` name-value pair argument) or rows with missing values.

Data Types: `double`

`NumPredictors` — Number of predictor variables

positive integer

This property is read-only.

Number of predictor variables used to fit the model, specified as a positive integer.

Data Types: `double`

`NumVariables` — Number of variables

positive integer

This property is read-only.

Number of variables in the input data, specified as a positive integer. `NumVariables` is the number of variables in the original table or dataset, or the total number of columns in the predictor matrix and response vector.

`NumVariables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `double`

`ObservationInfo` — Observation information

table

This property is read-only.

Observation information, specified as an *n*-by-4 table, where *n* is equal to the number of rows of input data. `ObservationInfo` contains the columns described in this table.

Column	Description
`Weights`	Observation weights, specified as a numeric value. The default value is `1`.
`Excluded`	Indicator of excluded observations, specified as a logical value. The value is `true` if you exclude the observation from the fit by using the `'Exclude'` name-value pair argument.
`Missing`	Indicator of missing observations, specified as a logical value. The value is `true` if the observation is missing.
`Subset`	Indicator of whether or not the fitting function uses the observation, specified as a logical value. The value is `true` if the observation is not excluded or missing, meaning the fitting function uses the observation.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the weight vector `w` of the model `mdl`:

```
w = mdl.ObservationInfo.Weights
```

Data Types: `table`

`ObservationNames` — Observation names
cell array of character vectors

This property is read-only.

Observation names, specified as a cell array of character vectors containing the names of the observations used in the fit.

- If the fit is based on a table or dataset containing observation names, `ObservationNames` uses those names.
- Otherwise, `ObservationNames` is an empty cell array.

Data Types: `cell`

`Offset` — Offset variable
numeric vector

This property is read-only.

Offset variable, specified as a numeric vector with the same length as the number of rows in the data. `Offset` is passed from `fitglm` or `stepwiseglm` in the `'Offset'` name-value pair argument. The fitting functions use `Offset` as an additional predictor variable with a coefficient value fixed at `1`. In other words, the formula for fitting is

$f(\mu)$ ~ `Offset + (terms involving real predictors)`

where f is the link function. The `Offset` predictor has coefficient `1`.

For example, consider a Poisson regression model. Suppose the number of counts is known for theoretical reasons to be proportional to a predictor `A`. By using the log link function and by specifying `log(A)` as an offset, you can force the model to satisfy this theoretical constraint.

Data Types: `double`

`PredictorNames` — Names of predictors used to fit model
cell array of character vectors

This property is read-only.

Names of predictors used to fit the model, specified as a cell array of character vectors.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`VariableInfo` — Information about variables

table

This property is read-only.

Information about variables contained in `Variables`, specified as a table with one row for each variable and the columns described in this table.

Column	Description
`Class`	Variable class, specified as a cell array of character vectors, such as `'double'` and `'categorical'`
`Range`	Variable range, specified as a cell array of vectors • Continuous variable — Two-element vector `[min,max]`, the minimum and maximum values • Categorical variable — Vector of distinct variable values
`InModel`	Indicator of which variables are in the fitted model, specified as a logical vector. The value is `true` if the model includes the variable.
`IsCategorical`	Indicator of categorical variables, specified as a logical vector. The value is `true` if the variable is categorical.

`VariableInfo` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

`VariableNames` — Names of variables

cell array of character vectors

This property is read-only.

Names of variables, specified as a cell array of character vectors.

- If the fit is based on a table or dataset, this property provides the names of the variables in the table or dataset.
- If the fit is based on a predictor matrix and response vector, `VariableNames` contains the values specified by the `'VarNames'` name-value pair argument of the fitting method. The default value of `'VarNames'` is `{'x1','x2',...,'xn','y'}`.

`VariableNames` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `cell`

`Variables` — Input data

table

This property is read-only.

Input data, specified as a table. `Variables` contains both predictor and response values. If the fit is based on a table or dataset array, `Variables` contains all the data from the table or dataset array. Otherwise, `Variables` is a table created from the input data matrix `X` and the response vector `y`.

`Variables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

Object Functions

Create `CompactGeneralizedLinearModel`

compact	Compact generalized linear regression model

Add or Remove Terms from Generalized Linear Model

addTerms	Add terms to generalized linear regression model
removeTerms	Remove terms from generalized linear regression model
step	Improve generalized linear regression model by adding or removing terms

Predict Responses

feval	Predict responses of generalized linear regression model using one input for each predictor
predict	Predict responses of generalized linear regression model
random	Simulate responses with random noise for generalized linear regression model

Evaluate Generalized Linear Model

coefCI	Confidence intervals of coefficient estimates of generalized linear regression model
coefTest	Linear hypothesis test on generalized linear regression model coefficients
devianceTest	Analysis of deviance for generalized linear regression model
partialDependence	Compute partial dependence

Visualize Generalized Linear Model and Summary Statistics

plotDiagnostics	Plot observation diagnostics of generalized linear regression model
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotResiduals	Plot residuals of generalized linear regression model
plotSlice	Plot of slices through fitted generalized linear regression surface

Gather Properties of Generalized Linear Model

gather	Gather properties of machine learning model from GPU

Examples

Create Generalized Linear Regression Model

Fit a logistic regression model of the probability of smoking as a function of age, weight, and sex, using a two-way interaction model.

Load the `hospital` data set.

```
load hospital
```

Convert the dataset array to a table.

```
tbl = dataset2table(hospital);
```

Specify the model using a formula that includes two-way interactions and lower-order terms.

```
modelspec = 'Smoker ~ Age*Weight*Sex - Age:Weight:Sex';
```

Create the generalized linear model.

```
mdl = fitglm(tbl,modelspec,'Distribution','binomial')
```

```
mdl =
Generalized linear regression model:
    logit(Smoker) ~ 1 + Sex*Age + Sex*Weight + Age*Weight
    Distribution = Binomial

Estimated Coefficients:
                        Estimate         SE         tStat      pValue
                       ___________    _________    ________    _______

    (Intercept)            -6.0492       19.749     -0.3063    0.75938
    Sex_Male               -2.2859       12.424    -0.18399    0.85402
    Age                    0.11691      0.50977     0.22934    0.81861
    Weight                0.031109      0.15208     0.20455    0.83792
    Sex_Male:Age          0.020734      0.20681     0.10025    0.92014
    Sex_Male:Weight        0.01216     0.053168     0.22871     0.8191
    Age:Weight         -0.00071959    0.0038964    -0.18468    0.85348


100 observations, 93 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 5.07, p-value = 0.535
```

The large *p*-value indicates that the model might not differ statistically from a constant.

Create Generalized Linear Regression Model Using Stepwise Regression

Create response data using three of 20 predictor variables, and create a generalized linear model using stepwise regression from a constant model to see if `stepwiseglm` finds the correct predictors.

Generate sample data that has 20 predictor variables. Use three of the predictors to generate the Poisson response variable.

```
rng default % for reproducibility
X = randn(100,20);
mu = exp(X(:,[5 10 15])*[.4;.2;.3] + 1);
y = poissrnd(mu);
```

Fit a generalized linear regression model using the Poisson distribution. Specify the starting model as a model that contains only a constant (intercept) term. Also, specify a model with an intercept and linear term for each predictor as the largest model to consider as the fit by using the `'Upper'` name-value pair argument.

```
mdl =  stepwiseglm(X,y,'constant','Upper','linear','Distribution','poisson')
```

```
1. Adding x5, Deviance = 134.439, Chi2Stat = 52.24814, PValue = 4.891229e-13
2. Adding x15, Deviance = 106.285, Chi2Stat = 28.15393, PValue = 1.1204e-07
3. Adding x10, Deviance = 95.0207, Chi2Stat = 11.2644, PValue = 0.000790094

mdl =
Generalized linear regression model:
    log(y) ~ 1 + x5 + x10 + x15
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.0115     0.064275    15.737    8.4217e-56
    x5             0.39508     0.066665    5.9263    3.0977e-09
    x10            0.18863      0.05534    3.4085     0.0006532
    x15            0.29295     0.053269    5.4995    3.8089e-08


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 91.7, p-value = 9.61e-20
```

`stepwiseglm` finds the three correct predictors: `x5`, `x10`, and `x15`.

More About

Canonical Link Function

The default link function for a generalized linear model is the canonical link function. You can specify a link function when you fit a model with `fitglm` or `stepwiseglm` by using the `'Link'` name-value pair argument.

Distribution	Canonical Link Function Name	Link Function	Mean (Inverse) Function
`'normal'`	`'identity'`	$f(\mu) = \mu$	$\mu = Xb$
`'binomial'`	`'logit'`	$f(\mu) = \log(\mu/(1 - \mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'poisson'`	`'log'`	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
`'gamma'`	`-1`	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
`'inverse gaussian'`	`-2`	$f(\mu) = 1/\mu^2$	$\mu = (Xb)^{-1/2}$

Cook's Distance

Cook's distance is the scaled change in fitted values, which is useful for identifying outliers in the observations for predictor variables. Cook's distance shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier.

The Cook's distance D_i of observation i is

$$D_i = w_i \frac{e_i^2}{p\widehat{\varphi}} \frac{h_{ii}}{(1 - h_{ii})^2},$$

where

- $\widehat{\varphi}$ is the dispersion parameter (estimated or theoretical).
- e_i is the linear predictor residual, $g(y_i) - x_i\widehat{\beta}$, where
  - g is the link function.
  - y_i is the observed response.
  - x_i is the observation.
  - $\widehat{\beta}$ is the estimated coefficient vector.
- p is the number of coefficients in the regression model.
- h_{ii} is the ith diagonal element of the Hat Matrix on page 33-2605 H.

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

Hat Matrix

The hat matrix is a projection matrix that projects the vector of response observations onto the vector of predictions.

The hat matrix H is defined in terms of the data matrix X and a diagonal weight matrix W:

$H = X(X^TWX)^{-1}X^TW^T$.

W has diagonal elements w_i:

$$w_i = \frac{g'(\mu_i)}{\sqrt{V(\mu_i)}},$$

where

- g is the link function mapping y_i to x_ib.
- g' is the derivative of the link function g.
- V is the variance function.
- μ_i is the ith mean.

The diagonal elements H_{ii} satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where n is the number of observations (rows of X), and p is the number of coefficients in the regression model.

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to a saturated model.

Deviance of a model M_1 is twice the difference between the loglikelihood of the model M_1 and the saturated model M_s. A saturated model is a model with the maximum number of parameters that you can estimate.

For example, if you have n observations (y_i, $i = 1, 2, ..., n$) with potentially different values for $X_i^T\beta$, then you can define a saturated model with n parameters. Let $L(b,y)$ denote the maximum value of the likelihood function for a model with the parameters b. Then the deviance of the model M_1 is

$$-2(\log L(b_1, y) - \log L(b_S, y)),$$

where b_1 and b_s contain the estimated parameters for the model M_1 and the saturated model, respectively. The deviance has a chi-square distribution with n – p degrees of freedom, where n is the number of parameters in the saturated model and p is the number of parameters in the model M_1.

Assume you have two different generalized linear regression models M_1 and M_2, and M_1 has a subset of the terms in M_2. You can assess the fit of the models by comparing the deviances D_1 and D_2 of the two models. The difference of the deviances is

$$\begin{aligned} D = D_2 - D_1 &= -2(\log L(b_2, y) - \log L(b_S, y)) + 2(\log L(b_1, y) - \log L(b_S, y)) \\ &= -2(\log L(b_2, y) - \log L(b_1, y)). \end{aligned}$$

Asymptotically, the difference D has a chi-square distribution with degrees of freedom v equal to the difference in the number of parameters estimated in M_1 and M_2. You can obtain the p-value for this test by using `1 - chi2cdf(D,v)`.

Typically, you examine D using a model M_2 with a constant term and no predictors. Therefore, D has a chi-square distribution with p – 1 degrees of freedom. If the dispersion is estimated, the difference divided by the estimated dispersion has an F distribution with p – 1 numerator degrees of freedom and n – p denominator degrees of freedom.

Terms Matrix

A terms matrix `T` is a t-by-(p + 1) matrix specifying terms in a model, where t is the number of terms, p is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and

response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

References

[1] McFadden, Daniel. "Conditional logit analysis of qualitative choice behavior." in *Frontiers in Econometrics*, edited by P. Zarembka,105–42. New York: Academic Press, 1974.

[2] Nagelkerke, N. J. D. "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78, no. 3 (1991): 691–92.

[3] Maddala, Gangadharrao S. *Limited-Dependent and Qualitative Variables in Econometrics.* Econometric Society Monographs. New York, NY: Cambridge University Press, 1983.

[4] Cox, D. R., and E. J. Snell. *Analysis of Binary Data.* 2nd ed. Monographs on Statistics and Applied Probability 32. London; New York: Chapman and Hall, 1989.

[5] Magee, Lonnie. "R 2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests." *The American Statistician* 44, no. 3 (August 1990): 250–53.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `random` functions support code generation.
- When you fit a model by using `fitglm` or `stepwiseglm`, you cannot specify `Link`, `Derivative`, and `Inverse` fields of the `'Link'` name-value pair argument as anonymous functions. That is, you cannot generate code using a generalized linear model that was created using anonymous functions for links. Instead, define functions for link components.

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions fully support GPU arrays:
  - `feval`
  - `predict`
  - `random`
  - `partialDependence`
  - `plotPartialDependence`
- The following object functions support model objects fitted with GPU array input arguments:
  - `compact`
  - `addTerms`

- `removeTerms`
- `step`
- `coefCI`
- `coefTest`
- `devianceTest`
- `plotDiagnostics`
- `plotResiduals`
- `plotSlice`
- `gather`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `LinearModel` | `NonLinearModel` | `fitglm` | `stepwiseglm`

Topics

“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2012a

generateCode

Package: `classreg.learning.coder.config.svm`

Generate C/C++ code using coder configurer

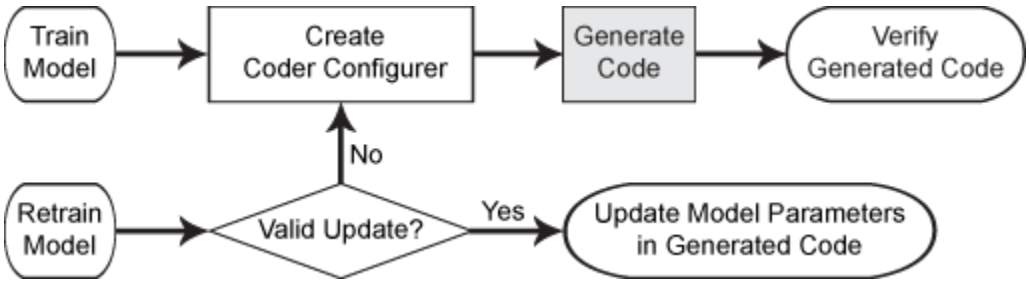
Syntax

```
generateCode(configurer)
generateCode(configurer,cfg)
generateCode( ___ ,'OutputPath',outputPath)
```

Description

After training a machine learning model, create a coder configurer for the model by using `learnerCoderConfigurer`. Modify the properties of the configurer to specify code generation options. Then use `generateCode` to generate C/C++ code for the `predict` and `update` functions of the machine learning model. Generating C/C++ code requires MATLAB Coder.

This flow chart shows the code generation workflow using a coder configurer. Use `generateCode` for the highlighted step.



`generateCode(configurer)` generates a MEX (MATLAB Executable) function for the `predict` and `update` functions of a machine learning model by using `configurer`. The generated MEX function is named *`outputFileName`*, which is the file name stored in the `OutputFileName` property of `configurer`.

To generate a MEX function, `generateCode` first generates the following MATLAB files required to generate code and stores them in the current folder:

- `predict.m`, `update.m`, and `initialize.m` — `predict.m` and `update.m` are the entry-point functions for the `predict` and `update` functions of the machine learning model, respectively, and these two functions call `initialize.m`.
- A MAT-file that includes machine learning model information — `generateCode` uses the `saveLearnerForCoder` function to save machine learning model information in a MAT-file whose file name is stored in the `OutputFileName` property of a coder configurer. `initialize.m` loads the saved MAT-file by using the `loadLearnerForCoder` function.

After generating the necessary MATLAB files, `generateCode` creates the MEX function and the code for the MEX function in the `codegen\mex\`*`outputFileName`* folder and copies the MEX function to the current folder.

`generateCode(configurer,cfg)` generates C/C++ code using the build type specified by `cfg`.

`generateCode( ___ ,'OutputPath',outputPath)` specifies the folder path for the output files in addition to any of the input arguments in previous syntaxes. `generateCode` generates the MATLAB files in the folder specified by `outputPath` and generates C/C++ code in the folder `outputPath \codegen\`*`type`*`\`*`outputFileName`* where *`type`* is the build type specified by `cfg`.

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `carsmall` data set and train a support vector machine (SVM) regression model.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X,Y);
```

`Mdl` is a `RegressionSVM` object.

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  RegressionSVMCoderConfigurer with properties:

   Update Inputs:
             Alpha: [1x1 LearnerCoderInput]
    SupportVectors: [1x1 LearnerCoderInput]
             Scale: [1x1 LearnerCoderInput]
              Bias: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'RegressionSVMModel'


  Properties, Methods
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionSVMModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionSVMModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('RegressionSVMModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Alpha
        %                       SupportVectors
        %                       Scale
        %                       Bias
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
```

```
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Specify Build Type

Train a machine learning model and generate code by using the coder configurer of the trained model. When generating code, specify the build type and other configuration options using a code generation configuration object.

Load the `ionosphere` data set and train a binary support vector machine (SVM) classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` object.

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Create a code generation configuration object by using `coder.config` (MATLAB Coder). Specify `'dll'` to generate a dynamic library and specify the `GenerateReport` property as `true` to enable the code generation report.

```
cfg = coder.config('dll');
cfg.GenerateReport = true;
```

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` and the configuration object `cfg` to generate code. Also, specify the output folder path.

```
generateCode(configurer,cfg,'OutputPath','testPath')
```

```
Specified folder does not exist. Folder has been created.
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful: To view the report, open('codegen\dll\ClassificationSVMModel\html\rep
```

`generateCode` creates the specified folder. The function also generates the MATLAB files required to generate code and stores them in the folder. Then `generateCode` generates C code in the `testPath\codegen\dll\ClassificationSVMModel` folder.

Update Parameters of ECOC Classification Model in Generated Code

Train an error-correcting output codes (ECOC) model using SVM binary learners and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the ECOC model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using different settings, and update parameters in the generated code without regenerating the code.

Train Model

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create an SVM binary learner template to use a Gaussian kernel function and to standardize predictor data.

```
t = templateSVM('KernelFunction','gaussian','Standardize',true);
```

Train a multiclass ECOC model using the template `t`.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

`Mdl` is a `ClassificationECOC` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationECOC` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns the first two outputs of the `predict` function, which are the predicted labels and negated average binary losses.

```
configurer = learnerCoderConfigurer(Mdl,X,'NumOutputs',2)
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

`configurer` is a `ClassificationECOCCoderConfigurer` object, which is a coder configurer of a `ClassificationECOC` object. The display shows the tunable input arguments of `predict` and `update`: `X`, `BinaryLearners`, `Prior`, and `Cost`.

Specify Coder Attributes of Parameters

Specify the coder attributes of `predict` arguments (predictor data and the name-value pair arguments `'Decoding'` and `'BinaryLoss'`) and `update` arguments (support vectors of the SVM learners) so that you can use these arguments as the input arguments of `predict` and `update` in the generated code.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 4];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 4 predictors, so the second value of the `SizeVector` attribute must be 4 and the second value of the `VariableDimensions` attribute must be `false`.

Next, modify the coder attributes of `BinaryLoss` and `Decoding` to use the `'BinaryLoss'` and `'Decoding'` name-value pair arguments in the generated code. Display the coder attributes of `BinaryLoss`.

```
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

            Value: 'hinge'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 0
```

To use a nondefault value in the generated code, you must specify the value before generating the code. Specify the `Value` attribute of `BinaryLoss` as `'exponential'`.

```
configurer.BinaryLoss.Value = 'exponential';
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

            Value: 'exponential'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
```

```
        Tunability: 1
```

If you modify attribute values when `Tunability` is `false` (logical 0), the software sets the `Tunability` to `true` (logical 1).

Display the coder attributes of `Decoding`.

```
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: 'lossweighted'
    SelectedOption: 'Built-in'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 1
        Tunability: 0
```

Specify the `IsConstant` attribute of `Decoding` as `false` so that you can use all available values in `BuiltInOptions` in the generated code.

```
configurer.Decoding.IsConstant = false;
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: [1x1 LearnerCoderInput]
    SelectedOption: 'NonConstant'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 0
        Tunability: 1
```

The software changes the `Value` attribute of `Decoding` to a `LearnerCoderInput` object so that you can use both `'lossweighted'` and `'lossbased'` as the value of `'Decoding'`. Also, the software sets the `SelectedOption` to `'NonConstant'` and the `Tunability` to `true`.

Finally, modify the coder attributes of `SupportVectors` in `BinaryLearners`. Display the coder attributes of `SupportVectors`.

```
configurer.BinaryLearners.SupportVectors
```

```
ans =
  LearnerCoderInput with properties:

            SizeVector: [54 4]
    VariableDimensions: [1 0]
              DataType: 'double'
            Tunability: 1
```

The default value of `VariableDimensions` is `[true false]` because each learner has a different number of support vectors. If you retrain the ECOC model using new data or different settings, the number of support vectors in the SVM learners can vary. Therefore, increase the upper bound of the number of support vectors.

```
configurer.BinaryLearners.SupportVectors.SizeVector = [150 4];

SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Display the coder configurer.

```
configurer
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
        BinaryLoss: [1x1 EnumeratedInput]
          Decoding: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

The display now includes `BinaryLoss` and `Decoding` as well.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the ECOC classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationECOCModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationECOCModel` for the two entry-point functions.

- Create the code for the MEX function in the `codegen\mex\ClassificationECOCModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument. Because you specified `'Decoding'` as a tunable input argument by changing the `IsConstant` attribute before generating the code, you also need to specify it in the call to the MEX function, even though `'lossweighted'` is the default value of `'Decoding'`.

```
[label,NegLoss] = predict(Mdl,X,'BinaryLoss','exponential');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
```

Compare `label` to `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`NegLoss_mex` might include round-off differences compared to `NegLoss`. In this case, compare `NegLoss_mex` to `NegLoss`, allowing a small tolerance.

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `NegLoss` and `NegLoss_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using a different setting. Specify `'KernelScale'` as `'auto'` so that the software selects an appropriate scale factor using a heuristic procedure.

```
t_new = templateSVM('KernelFunction','gaussian','Standardize',true,'KernelScale','auto');
retrainedMdl = fitcecoc(X,Y,'Learners',t_new);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationECOCModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` to the outputs from the `predict` function in the updated MEX function.

```
[label,NegLoss] = predict(retrainedMdl,X,'BinaryLoss','exponential','Decoding','lossbased');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
isequal(label,label_mex)

ans = logical
   1


find(abs(NegLoss-NegLoss_mex) > 1e-8)

ans =

  0x1 empty double column vector
```

The comparison confirms that `label` and `label_mex` are equal, and `NegLoss` and `NegLoss_mex` are equal within the tolerance.

Input Arguments

`configurer` — Coder configurer

coder configurer object

Coder configurer of a machine learning model, specified as a coder configurer object created by using `learnerCoderConfigurer`.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

`cfg` — Build type

`'mex'` (default) | `'dll'` | `'lib'` | code generation configuration object

Build type, specified as `'mex'`, `'dll'`, `'lib'`, or a code generation configuration object created by `coder.config`.

`generateCode` generates C/C++ code using one of the following build types.

- `'mex'` — Generates a MEX function that has a platform-dependent extension. A MEX function is a C/C++ program that is executable from the Command Window. Before generating a C/C++ library for deployment, generate a MEX function to verify that the generated code provides the correct functionality.

- `'dll'` — Generate a dynamic C/C++ library.
- `'lib'` — Generate a static C/C++ library.
- Code generation configuration object created by `coder.config` — Generate C/C++ code using the code generation configuration object to customize code generation options. You can specify the build type and other configuration options using the object. For example, modify the `GenerateReport` parameter to enable the code generation report, and modify the `TargetLang` parameter to generate C++ code. The default value of the `TargetLang` parameter is `'C'`, generating C code.

  ```
  cfg = coder.config('mex');
  cfg.GenerateReport = true;
  cfg.TargetLang = 'C++';
  ```

  For details, see the `-config` option of `codegen`, `coder.config`, and “Configure Build Settings” (MATLAB Coder).

`generateCode` generates C/C++ code in the folder `outputPath\codegen\`*`type`*`\`*`outputFileName`*, where *`type`* is the build type specified by the `cfg` argument and *`outputFileName`* is the file name stored in the `OutputFileName` property of `configurer`.

`outputPath` — Folder path for output files
current folder (default) | character vector | string scalar

Folder path for the output files of `generateCode`, specified as a character vector or string array.

The specified folder path can be an absolute path or a relative path to the current folder path.

- The path must not contain spaces because they can lead to code generation failures in certain operating system configurations.
- The path also cannot contain non-7-bit ASCII characters, such as Japanese characters.

If the specified folder does not exist, then `generateCode` creates the folder.

`generateCode` searches the specified folder for the four MATLAB files: `predict.m`, `update.m`, `initialize.m`, and a MAT-file that includes machine learning model information. If the four files do not exist in the folder, then `generateCode` generates the files. Otherwise, `generateCode` does not generate any MATLAB files.

`generateCode` generates C/C++ code in the folder `outputPath\codegen\`*`type`*`\`*`outputFileName`*, where *`type`* is the build type specified by the `cfg` argument and *`outputFileName`* is the file name stored in the `OutputFileName` property of `configurer`.

Example: `'C:\myfiles'`

Data Types: `char` | `string`

Limitations

- The `generateCode` function uses the `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` functions, so the code generation limitations of these functions also apply to the `generateCode` function. For details, see the function reference pages `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
- For the code generation usage notes and limitations of a machine learning model and its object functions, see the Code Generation sections of the corresponding reference pages.

Model	Model Object	predict Function
Binary decision tree for multiclass classification	`CompactClassificationTree`	`predict`
SVM for one-class and binary classification	`CompactClassificationSVM`	`predict`
Linear model for binary classification	`ClassificationLinear`	`predict`
Multiclass model for SVMs and linear models	`CompactClassificationECOC`	`predict`
Binary decision tree for regression	`CompactRegressionTree`	`predict`
SVM regression	`CompactRegressionSVM`	`predict`
Linear regression	`RegressionLinear`	`predict`

Alternative Functionality

- If you want to modify the MATLAB files (`predict.m`, `update.m`, and `initialize.m`) according to your code generation workflow, then use `generateFiles` to generate these files and use `codegen` to generate code.

See Also

`generateFiles` | `learnerCoderConfigurer` | `update` | `validatedUpdateInputs`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

generateFiles

Package: `classreg.learning.coder.config.svm`

Generate MATLAB files for code generation using coder configurer

Syntax

```
generateFiles(configurer)
generateFiles(configurer,'OutputPath',outputPath)
```

Description

`generateFiles(configurer)` generates the MATLAB files required to generate C/C++ code by using the coder configurer `configurer`, and saves the generated files in the current folder.

To customize the code generation workflow, use `generateFiles` and `codegen`. If you do not need to customize your workflow, use `generateCode`.

`generateFiles` generates the following MATLAB files:

- `predict.m`, `update.m`, and `initialize.m` — `predict.m` and `update.m` are the entry-point functions for the `predict` and `update` functions of the machine learning model, respectively, and these two functions call `initialize.m`. You can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project.
- A MAT-file that includes machine learning model information — `generateFiles` uses the `saveLearnerForCoder` function to save machine learning model information in a MAT-file whose file name is stored in the `OutputFileName` property of a coder configurer. `initialize.m` loads the saved MAT-file by using the `loadLearnerForCoder` function.

After you generate these files, generate C/C++ code by using `codegen` and the prepared `codegen` argument stored in the `CodeGenerationArguments` property of a coder configurer.

If the folder already includes all four MATLAB files, then `generateFiles` does not generate any files.

`generateFiles(configurer,'OutputPath',outputPath)` generates the MATLAB files in the folder specified by `outputPath`.

Examples

Generate MATLAB® Files for Code Generation

Train a machine learning model and then generate the MATLAB files required to generate C/C++ code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `ionosphere` data set and train a binary support vector machine (SVM) classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` object.

Create a coder configurer for the `ClassificationSVM` object.

```
configurer = learnerCoderConfigurer(Mdl,X);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Use `generateFiles` to generate the MATLAB files required to generate C/C++ code for the `predict` and `update` functions of the model.

```
generateFiles(configurer)
```

`generateFiles` generates `predict.m`, `update.m`, `initialize.m`, and `ClassificationSVMModel.mat` (a MAT-file that includes machine learning model information).

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files.

```
type predict.m % Display contents of predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:03:16
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m % Display contents of update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:03:16
initialize('update',varargin{:});
end
```

```
type initialize.m % Display contents of initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:03:16
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('ClassificationSVMModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Alpha
        %                       SupportVectors
        %                       SupportVectorLabels
        %                       Scale
        %                       Bias
        %                       Prior
        %                       Cost
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
```

```
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Generate C/C++ code by using `codegen` (MATLAB Coder) and the prepared `codegen` argument stored in the `CodeGenerationArguments` property of `configurer`.

```
cfArgs = configurer.CodeGenerationArguments;
codegen(cfArgs{:})
```

```
Code generation successful.
```

Input Arguments

`configurer` — Coder configurer

coder configurer object

Coder configurer of a machine learning model, specified as a coder configurer object created by using `learnerCoderConfigurer`.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

`outputPath` — Folder path for output files

current folder (default) | character vector | string scalar

Folder path for the output files of `generateFiles`, specified as a character vector or string array.

The specified folder path can be an absolute path or a relative path to the current folder path.

- The path must not contain spaces because they can lead to code generation failures in certain operating system configurations.
- The path also cannot contain non-7-bit ASCII characters, such as Japanese characters.

If the specified folder does not exist, then `generateFiles` creates the folder.

`generateFiles` searches the specified folder for the four MATLAB files: `predict.m`, `update.m`, `initialize.m`, and a MAT-file that includes machine learning model information. If the four files do

not exist in the folder, then `generateFiles` generates the files. Otherwise, `generateFiles` does not generate any MATLAB files.

Example: `'C:\myfiles'`

Data Types: `char` | `string`

Alternative Functionality

- To customize the code generation workflow, use `generateFiles` and `codegen`. If you do not need to customize your workflow, use `generateCode`. In addition to generating the four MATLAB files generated by `generateFiles`, the `generateCode` function also generates the C/C++ code.

See Also

`generateCode` | `learnerCoderConfigurer` | `update` | `validatedUpdateInputs`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

generateLearnerDataTypeFcn

Generate function that defines data types for fixed-point code generation

Syntax

```
generateLearnerDataTypeFcn(filename,X)
generateLearnerDataTypeFcn(filename,X,Name,Value)
```

Description

To generate fixed-point C/C++ code for the `predict` function of a machine learning model, use `generateLearnerDataTypeFcn`, `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.

- After training a machine learning model, save the model using `saveLearnerForCoder`.
- Create a structure that defines fixed-point data types by using the function generated from `generateLearnerDataTypeFcn`.
- Define an entry-point function that loads the model by using both `loadLearnerForCoder` and the structure, and then calls the `predict` function.
- Generate code using `codegen`, and then verify the generated code.

The `generateLearnerDataTypeFcn` function requires Fixed-Point Designer, and generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.

This flow chart shows the fixed-point code generation workflow for the `predict` function of a machine learning model. Use `generateLearnerDataTypeFcn` for the highlighted step.

`generateLearnerDataTypeFcn(filename,X)` generates a data type function on page 33-2631 that defines fixed-point data types for the variables required to generate fixed-point C/C++ code for prediction of a machine learning model. `filename` stores the machine learning model, and `X` contains the predictor data for the `predict` function of the model.

Use the generated function to create a structure that defines fixed-point data types. Then, use the structure as the input argument `T` of `loadLearnerForCoder`.

`generateLearnerDataTypeFcn(filename,X,Name,Value)` specifies additional options by using one or more name-value pair arguments. For example, you can specify `'WordLength',32` to use 32-bit word length for the fixed-point data types.

Examples

Generate Fixed-Point C/C++ Code for Prediction

After training a machine learning model, save the model using `saveLearnerForCoder`. For fixed-point code generation, specify the fixed-point data types of the variables required for prediction by

using the data type function generated by `generateLearnerDataTypeFcn`. Then, define an entry-point function that loads the model by using both `loadLearnerForCoder` and the specified fixed-point data types, and calls the `predict` function of the model. Use `codegen` (MATLAB Coder) to generate fixed-point C/C++ code for the entry-point function, and then verify the generated code.

Before generating code using `codegen`, you can use `buildInstrumentedMex` (Fixed-Point Designer) and `showInstrumentationResults` (Fixed-Point Designer) to optimize the fixed-point data types to improve the performance of the fixed-point code. Record minimum and maximum values of named and internal variables for prediction by using `buildInstrumentedMex`. View the instrumentation results using `showInstrumentationResults`; then, based on the results, tune the fixed-point data type properties of the variables. For details regarding this optional step, see “Fixed-Point Code Generation for Prediction of SVM” on page 32-87.

Train Model

Load the `ionosphere` data set and train a binary SVM classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y,'KernelFunction','gaussian');
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `myMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'myMdl');
```

Define Fixed-Point Data Types

Use `generateLearnerDataTypeFcn` to generate a function that defines the fixed-point data types of the variables required for prediction of the SVM model.

```
generateLearnerDataTypeFcn('myMdl',X)
```

`generateLearnerDataTypeFcn` generates the `myMdl_datatype` function.

Create a structure `T` that defines the fixed-point data types by using `myMdl_datatype`.

```
T = myMdl_datatype('Fixed')
```

```
T = struct with fields:
               XDataType: [0x0 embedded.fi]
           ScoreDataType: [0x0 embedded.fi]
    InnerProductDataType: [0x0 embedded.fi]
```

The structure `T` includes the fields for the named and internal variables required to run the `predict` function. Each field contains a fixed-point object, returned by `fi` (Fixed-Point Designer). The fixed-point object specifies fixed-point data type properties, such as word length and fraction length. For example, display the fixed-point data type properties of the predictor data.

```
T.XDataType
```

```
ans =

[]
```

```
          DataTypeMode: Fixed-point: binary point scaling
            Signedness: Signed
            WordLength: 16
        FractionLength: 14

        RoundingMethod: Floor
        OverflowAction: Wrap
           ProductMode: FullPrecision
  MaxProductWordLength: 128
               SumMode: FullPrecision
      MaxSumWordLength: 128
```

Define Entry-Point Function

Define an entry-point function named `myFixedPointPredict` that does the following:

- Accept the predictor data `X` and the fixed-point data type structure `T`.
- Load a fixed-point version of a trained SVM classification model by using both `loadLearnerForCoder` and the structure `T`.
- Predict labels and scores using the loaded model.

```
type myFixedPointPredict.m % Display contents of myFixedPointPredict.m file
```

```
function [label,score] = myFixedPointPredict(X,T) %#codegen
Mdl = loadLearnerForCoder('myMdl','DataType',T);
[label,score] = predict(Mdl,X);
end
```

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function file.

Generate Code

The `XDataType` field of the structure `T` specifies the fixed-point data type of the predictor data. Convert `X` to the type specified in `T.XDataType` by using the `cast` (Fixed-Point Designer) function.

```
X_fx = cast(X,'like',T.XDataType);
```

Generate code for the entry-point function using `codegen`. Specify `X_fx` and constant folded `T` as input arguments of the entry-point function.

```
codegen myFixedPointPredict -args {X_fx,coder.Constant(T)}
```

```
Code generation successful.
```

`codegen` generates the MEX function `myFixedPointPredict_mex` with a platform-dependent extension.

Verify Generated Code

Pass predictor data to `predict` and `myFixedPointPredict_mex` to compare the outputs.

```
[labels,scores] = predict(Mdl,X);
[labels_fx,scores_fx] = myFixedPointPredict_mex(X_fx,T);
```

Compare the outputs from `predict` and `myFixedPointPredict_mex`.

```
verify_labels = isequal(labels,labels_fx)
```

```
verify_labels = logical
   1
```

`isequal` returns logical 1 (true), which means `labels` and `labels_fx` are equal. If the labels are not equal, you can compute the percentage of incorrectly classified labels as follows.

```
sum(strcmp(labels_fx,labels)==0)/numel(labels_fx)*100
```

```
ans = 0
```

Find the maximum of the relative differences between the score outputs.

```
relDiff_scores = max(abs((scores_fx.double(:,1)-scores(:,1))./scores(:,1)))
```

```
relDiff_scores = 0.0055
```

If you are not satisfied with the comparison results and want to improve the precision of the generated code, you can tune the fixed-point data types and regenerate the code. For details, see "Tips" on page 33-2632 in `generateLearnerDataTypeFcn`, "Data Type Function" on page 33-2631, and "Fixed-Point Code Generation for Prediction of SVM" on page 32-87.

Input Arguments

`filename` — Name of MAT-file that contains structure array representing model object

character vector | string scalar

Name of the MATLAB formatted binary file (MAT-file) that contains the structure array representing a model object, specified as a character vector or string scalar.

You must create the `filename` file using `saveLearnerForCoder`, and the model in `filename` can be one of the following:

- Classification model
  - Decision tree (`CompactClassificationTree`)
  - Ensemble of decision trees (`CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`)
  - SVM (support vector machine) (`CompactClassificationSVM`)
- Regression model
  - Decision tree (`CompactRegressionTree`)
  - Ensemble of decision trees (`CompactRegressionEnsemble`, `RegressionBaggedEnsemble`)
  - SVM (`CompactRegressionSVM`)

The extension of the `filename` file must be `.mat`. If `filename` has no extension, then `generateLearnerDataTypeFcn` appends `.mat`.

If `filename` does not include a full path, then `generateLearnerDataTypeFcn` loads the file from the current folder.

Example: `'myMdl'`

Data Types: `char` | `string`

`X` — Predictor data
numeric matrix

Predictor data for the `predict` function of the model stored in `filename`, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `generateLearnerDataTypeFcn(filename,X,'OutputFunctionName','myDataTypeFcn','WordLength',32)` generates a data type function named `myDataTypeFcn` that uses 32 bits for the word length when defining the fixed-point data type for each variable.

`OutputFunctionName` — Name of generated function
`filename` plus `_datatype` (default) | character vector | string scalar

Name of the generated function, specified as the comma-separated pair consisting of `'OutputFunctionName'` and a character vector or string scalar. The `'OutputFunctionName'` value must be a valid MATLAB function name.

The default function name is the file name in `filename` followed by `_datatype`. For example, if `filename` is `myMdl`, then the default function name is `myMdl_datatype`.

Example: `'OutputFunctionName','myDataTypeFcn'`

Data Types: `char` | `string`

`WordLength` — Word length in bits
16 (default) | numeric scalar

Word length in bits, specified as the comma-separated pair consisting of `'WordLength'` and a numeric scalar.

The generated data type function on page 33-2631 defines a fixed-point object for each variable using the specified `'WordLength'` value. If a variable requires a longer word length than the specified value, the software doubles the word length for the variable.

The optimal word length depends on your target hardware properties. When the specified word length is longer than the longest word size of your target hardware, the generated code contains multiword operations.

For details, see “Fixed-Point Data Types” (Fixed-Point Designer).

Example: `'WordLength',32`

Data Types: `single` | `double`

`OutputRange` — Range of `predict` output
range simulated using `X` (default) | numeric vector of two elements

Range of the output argument of the `predict` function, specified as the comma-separated pair consisting of `'OutputRange'` and a numeric vector of two elements (minimum and maximum values of the output).

The `'OutputRange'` value specifies the range of predicted class scores for a classification model and the range of predicted responses for a regression model. The following tables list the output arguments for which you can specify the range by using the `'OutputRange'` name-value pair argument.

Classification Model

Model	`predict` Function of Model	Output Argument
Decision tree	`predict`	`score`
Ensemble of decision trees	`predict`	`score`
SVM	`predict`	`score`

Regression Model

Model	`predict` Function of Model	Output Argument
Decision tree	`predict`	`Yfit`
Ensemble of decision trees	`predict`	`Yfit`
SVM	`predict`	`yfit`

When `X` contains a large number of observations and the range for the output argument is known, specify the `'OutputRange'` value to reduce the amount of computation.

If you do not specify the `'OutputRange'` value, then the software simulates the output range using the predictor data `X` and the `predict` function.

The software determines the span of numbers that the fixed-point data can represent by using the `'OutputRange'` value and the `'PercentSafetyMargin'` value.

Example: `'OutputRange',[0,1]`

Data Types: `single` | `double`

`PercentSafetyMargin` — Safety margin percentage
10 (default) | numeric scalar

Safety margin percentage, specified as the comma-separated pair consisting of `'PercentSafetyMargin'` and a numeric scalar.

For each variable, the software simulates the range of the variable and adds the specified safety margin to determine the span of numbers that the fixed-point data can represent. Then, the software proposes the maximum fraction length that does not cause overflows.

Use caution when you specify the `'PercentSafetyMargin'` value. If a variable range is large, then increasing the safety margin can cause underflow, because the software decreases fraction length to represent a larger range using a given word length.

Example: `'PercentSafetyMargin',15`

Data Types: `single` | `double`

More About

Data Type Function

Use the data type function generated by `generateLearnerDataTypeFcn` to create a structure that defines fixed-point data types for the variables required to generate fixed-point C/C++ code for prediction of a machine learning model. Use the output structure of the data type function as the input argument `T` of `loadLearnerForCoder`.

If `filename` is `'myMdl'`, then `generateLearnerDataTypeFcn` generates a data type function named `myMdl_datatype`. The `myMdl_datatype` function supports this syntax:

```
T = myMdl_datatype(dt)
```

`T = myMdl_datatype(dt)` returns a data type structure that defines data types for the variables required to generate fixed-point C/C++ code for prediction of a machine learning model.

Each field of `T` contains a fixed-point object returned by `fi`. The input argument `dt` specifies the `DataType` property of the fixed-point object.

- Specify `dt` as `'Fixed'`(default) for fixed-point code generation.
- Specify `dt` as `'Double'` to simulate floating-point behavior of the fixed-point code.

Use the output structure `T` as the second input argument of `loadLearnerForCoder`.

The structure `T` contains the fields in the following table. These fields define the data types for the variables that directly influence the precision of the model. These variables, along with other named and internal variables, are required to run the `predict` function of the model.

Description	Fields

Common fields for classification	• `XDataType` (input)
	• `ScoreDataType` (output or internal variable) and `TransformedScoreDataType` (output)
	  • If you train a model using the default `'ScoreTransform'` value of `'none'` or `'identity'` (that is, you do not transform predicted scores), then the `ScoreDataType` field influences the precision of the output scores.
	  • If you train a model using a value of `'ScoreTransform'` other than `'none'` or `'identity'` (that is, you do transform predicted scores), then the `ScoreDataType` field influences the precision of the internal untransformed scores. The `TransformedScoreDataType` field influences the precision of the transformed output scores.

Common fields for regression	• `XDataType` (input) • `YFitDataType` (output)

Description	Fields
Additional fields for an ensemble of decision trees	• `WeakLearnerOutputDataType` (internal variable) — Data type for outputs from weak learners. • `AggregatedLearnerWeightsDataType` (internal variable) — Data type for a weighted aggregate of the outputs from weak learners, applicable only if you train a model using bagging (`'Method','bag'`). The software computes predicted scores (`ScoreDataType`) by dividing the aggregate by the sum of learner weights.
Additional fields for SVM	• `XnormDataType` (internal variable), applicable only if you train a model using `'Standardize'` or `'KernelScale'` • `InnerProductDataType` (internal variable)

The software proposes the maximum fraction length that does not cause overflows, based on the default word length (16) and safety margin (10%) for each variable.

The following code shows the data type function `myMdl_datatype`, generated by `generateLearnerDataTypeFcn` when `filename` is `'myMdl'` and the model in the `filename` file is an SVM classifier.

```
function T = myMdl_datatype(dt)

if nargin < 1
    dt = 'Fixed';
end

% Set fixed-point math settings
fm = fimath('RoundingMethod','Floor', ...
    'OverflowAction','Wrap', ...
    'ProductMode','FullPrecision', ...
    'MaxProductWordLength',128, ...
    'SumMode','FullPrecision', ...
    'MaxSumWordLength',128);

% Data type for predictor data
T.XDataType = fi([],true,16,14,fm,'DataType',dt);

% Data type for output score
T.ScoreDataType = fi([],true,16,14,fm,'DataType',dt);

% Internal variables
% Data type of the squared distance dist = (x-sv)^2 for the Gaussian kernel G(x,sv) = exp(-dist)
% where x is the predictor data for an observation and sv is a support vector
T.InnerProductDataType = fi([],true,16,6,fm,'DataType',dt);

end
```

Tips

- To improve the precision of the generated fixed-point code, you can tune the fixed-point data types. Modify the fixed-point data types by updating the data type function on page 33-2631 (`myMdl_datatype`) and creating a new structure, and then regenerate the code using the new structure. You can update the `myMdl_datatype` function in one of two ways:

- Regenerate the `myMdl_datatype` function by using `generateLearnerDataTypeFcn` and its name-value pair arguments.
  - Increase the word length by using the `'WordLength'` name-value pair argument.
  - Decrease the safety margin by using the `'PercentSafetyMargin'` name-value pair argument.

  If you increase the word length or decrease the safety margin, the software can propose a longer fraction length, and therefore, improve the precision of the generated code based on the given data set.
- Manually modify the fixed-point data types in the function file (`myMdl_datatype.m`). For each variable, you can tune the word length and fraction length and specify fixed-point math settings using a `fimath` object.

- In the generated fixed-point code, a large number of operations or a large variable range can result in loss of precision, compared to the precision of the corresponding floating-point code. When training an SVM model, keep the following tips in mind to avoid loss of precision in the generated fixed-point code:
  - Data standardization (`'Standardize'`) — To avoid overflows in the model property values of support vectors in an SVM model, you can standardize the predictor data. Instead of using the `'Standardize'` name-value pair argument when training the model, standardize the predictor data before passing the data to the fitting function and the `predict` function so that the fixed-point code does not include the operations for the standardization.
  - Kernel function (`'KernelFunction'`) — Using the Gaussian kernel or linear kernel is preferable to using a polynomial kernel. A polynomial kernel requires higher computational complexity than the other kernels, and the output of a polynomial kernel function is unbounded.
  - Kernel scale (`'KernelScale'`) — Using a kernel scale requires additional operations if the value of `'KernelScale'` is not 1.
  - The prediction of a one-class classification problem might have loss of precision if the predicted class score values have a large range.

Compatibility Considerations

Specify precision of transformed scores before and after their transformation
Behavior changed in R2020a

In R2019b, you could train an SVM classifier for fixed-point code generation with some nondefault score transforms (`'ismax'`, `'sign'`, `'symmetric'`, or `'symmetricismax'`). The `generateLearnerDataTypeFcn` function generated a data type function whose output structure contained only one field related to the scores, `ScoreDataType`. This field influenced the precision of the output scores both before and after their transformation.

Starting in R2020a, when you train an SVM classifier with a score transform other than `'none'` or `'identity'`, the `generateLearnerDataTypeFcn` function generates a data type function whose output structure `T` contains two fields related to the scores: `ScoreDataType` and `TransformedScoreDataType`. Use these fields to influence the precision of the output scores before and after their transformation, respectively. For more details, see “Data Type Function” on page 33-2631.

To update your code, rerun the `generateLearnerDataTypeFcn` function and then regenerate the output structure `T`.

See Also

`buildInstrumentedMex` | `codegen` | `fi` | `loadLearnerForCoder` | `saveLearnerForCoder` | `showInstrumentationResults`

Topics

“Fixed-Point Code Generation for Prediction of SVM” on page 32-87

Introduced in R2019b

gencfeatures

Perform automated feature engineering for classification

Syntax

```
[Transformer,NewTbl] = gencfeatures(Tbl,ResponseVarName,q)
[Transformer,NewTbl] = gencfeatures(Tbl,Y,q)
[Transformer,NewTbl] = gencfeatures(Tbl,formula,q)
[Transformer,NewTbl] = gencfeatures(___,Name,Value)
```

Description

The `gencfeatures` function enables you to automate the feature engineering process in the context of a machine learning workflow. Before passing tabular training data to a classifier, you can create new features from the predictors in the data by using `gencfeatures`. Use the returned data to train the classifier.

To better understand the generated features, use the `describe` function of the returned `FeatureTransformer` object. To apply the same training set feature transformations to a test set, use the `transform` function of the `FeatureTransformer` object.

`[Transformer,NewTbl] = gencfeatures(Tbl,ResponseVarName,q)` uses automated feature engineering to create `q` features from the predictors in `Tbl`. The software assumes that the `ResponseVarName` variable in `Tbl` is the response and does not create new features from this variable. `gencfeatures` returns a `FeatureTransformer` object (`Transformer`) and a new table (`NewTbl`) that contains the transformed features.

By default, `gencfeatures` assumes that generated features are used to train an interpretable linear model with a binary response variable. If you have a multiclass response variable and you want to generate features to improve the accuracy of a bagged ensemble, specify `'TargetLearner','bag'`.

`[Transformer,NewTbl] = gencfeatures(Tbl,Y,q)` assumes that the vector `Y` is the response variable and creates new features from the variables in `Tbl`.

`[Transformer,NewTbl] = gencfeatures(Tbl,formula,q)` uses the explanatory model `formula` to determine the response variable in `Tbl` and the subset of `Tbl` predictors from which to create new features.

`[Transformer,NewTbl] = gencfeatures(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can change the expected learner type, the method for selecting new features, and the standardization method for transformed data.

Examples

Interpret Linear Model with Generated Features

Use automated feature engineering to generate new features. Train a linear classifier using the generated features. Interpret the relationship between the generated features and the trained model.

Load the `patients` data set. Create a table from a subset of the variables.

```
load patients
Tbl = table(Age,Diastolic,Gender,Height,SelfAssessedHealthStatus, ...
    Systolic,Weight,Smoker);
```

Generate 10 new features from the variables in `Tbl`. Specify the `Smoker` variable as the response. By default, `gencfeatures` assumes that the new features will be used to train a binary linear classifier.

```
rng("default") % For reproducibility
[T,NewTbl] = gencfeatures(Tbl,"Smoker",10)
```

```
T =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'linear'
    NumEngineeredFeatures: 10
      NumOriginalFeatures: 0
         TotalNumFeatures: 10


NewTbl=100×11 table
    zsc(Systolic.^2)    eb8(Diastolic)    q8(Systolic)    eb8(Systolic)    q8(Diastolic)    zsc(
    ________________    ______________    ____________    _____________    _____________    ____

         0.15379              8                6               4               8           -1.
         -1.9421              2                1               1               2          -0.2
         0.30311              4                6               5               5           0.5
        -0.85785              2                2               2               2           0.8
        -0.14125              3                5               4               4             1
        -0.28697              1                4               3               1           0.6
          1.0677              6                8               6               6          -0.4
         -1.1361              4                2               2               5          -0.7
         -1.1361              3                2               2               3          -0.8
        -0.71693              5                3               3               6           0.3
         -1.2734              2                1               1               2           1.2
         -1.1361              1                2               2               1             1
         0.60534              1                6               5               1          -0.9
          1.0677              8                8               6               8          -0.2
         -1.2734              3                1               1               4           0.9
          1.0677              7                8               6               8          -0.9
      ⋮
```

`T` is a `FeatureTransformer` object that can be used to transform new data, and `newTbl` contains the new features generated from the `Tbl` data.

To better understand the generated features, use the `describe` object function of the `FeatureTransformer` object. For example, inspect the first two generated features.

```
describe(T,1:2)
```

```
                          Type        IsOriginal    InputVariables                          Tr
                      ____________    __________    ______________    ____________________________

    zsc(Systolic.^2)  Numeric           false         Systolic        power(  ,2)
                                                                      Standardization with z-sc
    eb8(Diastolic)    Categorical       false         Diastolic       Equal-width binning (numbe
```

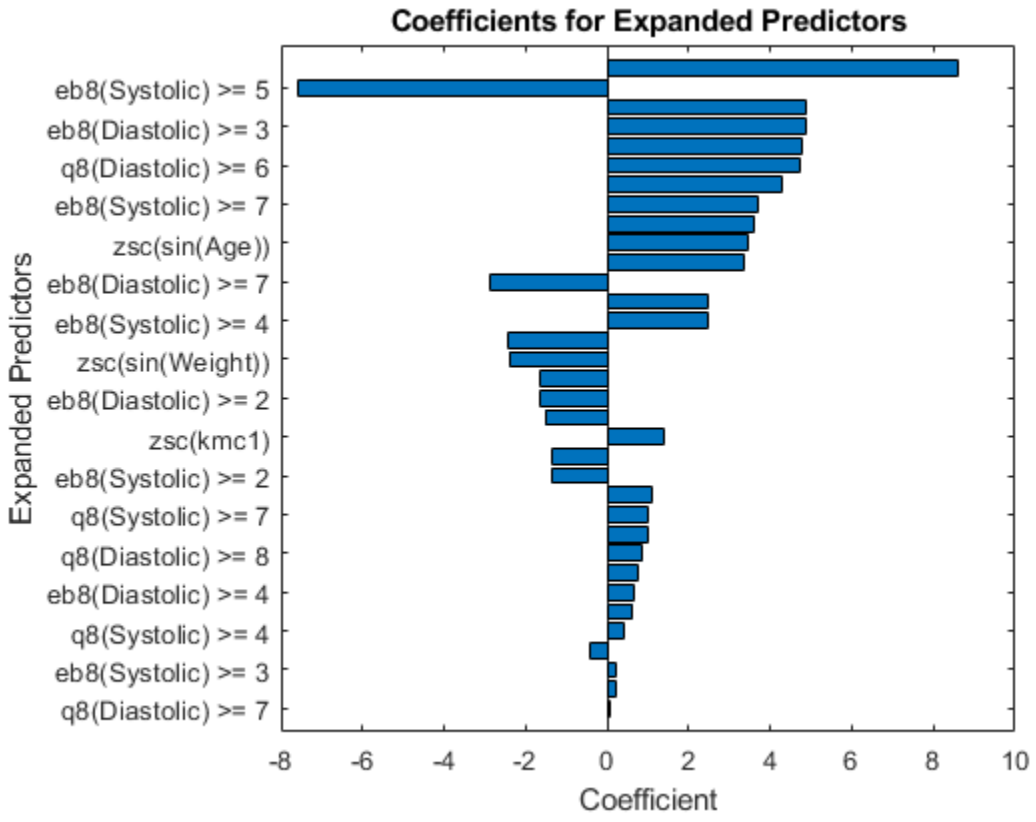
The first feature in `newTbl` is a numeric variable, created by first squaring the values of the `Systolic` variable and then converting the results to z-scores. The second feature in `newTbl` is a categorical variable, created by binning the values of the `Systolic` variable into 50 equiprobable bins.

Use the generated features to fit a linear classifier without any regularization.

```
Mdl = fitclinear(NewTbl,"Smoker","Lambda",0);
```

Plot the coefficients of the predictors used to train `Mdl`. Note that `fitclinear` expands categorical predictors before fitting a model.

```
p = length(Mdl.Beta);
[sortedCoefs,expandedIndex] = sort(Mdl.Beta,"ComparisonMethod","abs");
sortedExpandedPreds = Mdl.ExpandedPredictorNames(expandedIndex);
bar(sortedCoefs,"Horizontal","on")
yticks(1:2:p)
yticklabels(sortedExpandedPreds(1:2:end))
xlabel("Coefficient")
ylabel("Expanded Predictors")
title("Coefficients for Expanded Predictors")
```



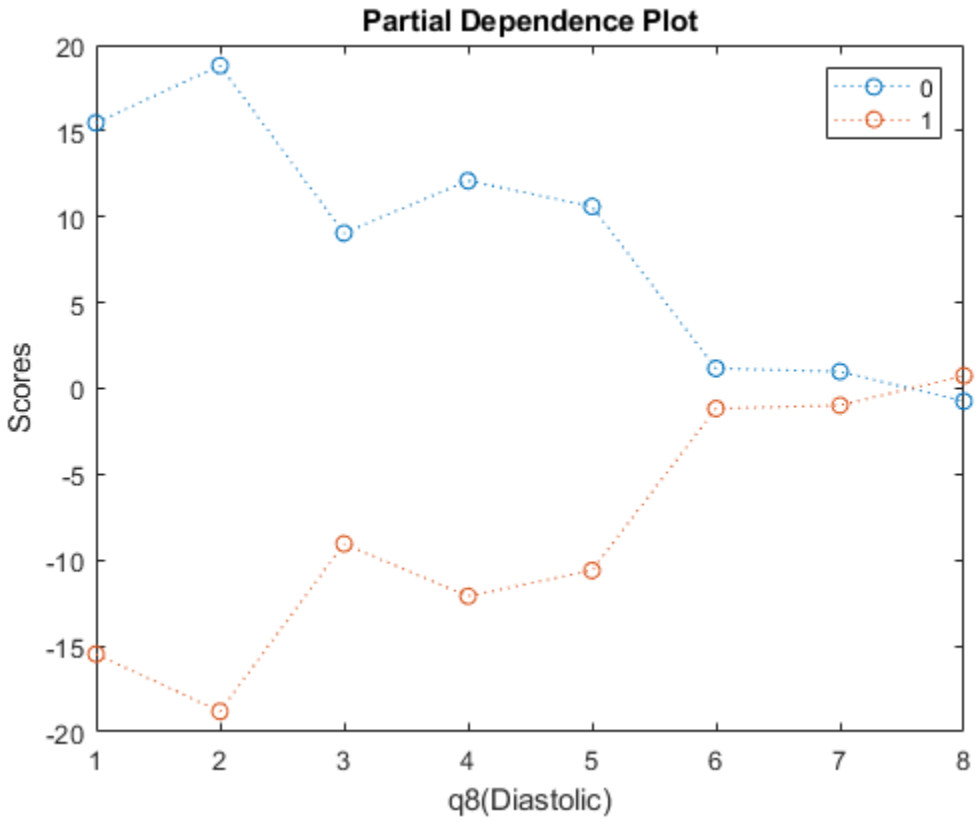
Identify the predictors whose coefficients have larger absolute values.

```
bigCoefs = abs(sortedCoefs) >= 4;
flip(sortedExpandedPreds(bigCoefs))
```

```
ans = 1×7 cell
    {'zsc(Systolic.^2)'}    {'eb8(Systolic) >= 5'}    {'q8(Diastolic) >= 3'}    {'eb8(Diastolic)
```

You can use partial dependence plots to analyze the categorical features whose levels have large coefficients in terms of absolute value. For example, inspect the partial dependence plot for the `q8(Diastolic)` variable, whose levels `q8(Diastolic) >= 3` and `q8(Diastolic) >= 6` have coefficients with large absolute values. These two levels correspond to noticeable changes in the predicted scores.

```
plotPartialDependence(Mdl,"q8(Diastolic)",Mdl.ClassNames,NewTbl);
```



Improve Accuracy for Interpretable Linear Model

Generate new features to improve the model accuracy for an interpretable linear model. Compare the test set accuracy of a linear model trained on the original data to the test set accuracy of a linear model trained on the transformed features.

Load the `ionosphere` data set. Convert the matrix of predictors `X` to a table.

```
load ionosphere
tbl = array2table(X);
```

Partition the data into training and test sets. Use approximately 70% of the observations as training data, and 30% of the observations as test data. Partition the data using `cvpartition`.

```
rng("default") % For reproducibility of the partition
cvp = cvpartition(Y,"Holdout",0.3);

trainIdx = training(cvp);
trainTbl = tbl(training(cvp),:);
trainY = Y(trainIdx);

testIdx = test(cvp);
testTbl = tbl(testIdx,:);
testY = Y(testIdx);
```

Use the training data to generate 45 new features. Inspect the returned `FeatureTransformer` object.

```
[T,newTrainTbl] = gencfeatures(trainTbl,trainY,45);
T
```

```
T =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'linear'
    NumEngineeredFeatures: 45
      NumOriginalFeatures: 0
         TotalNumFeatures: 45
```

All the generated features are engineered features rather than original features in `trainTbl`.

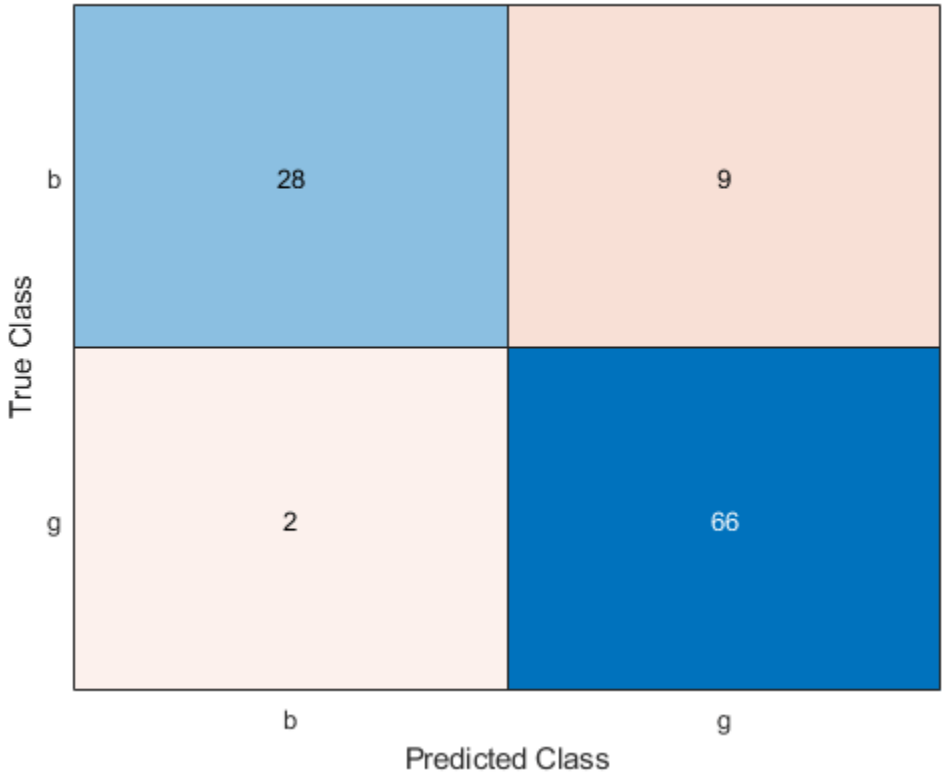
Apply the transformations stored in the object `T` to the test data.

```
newTestTbl = transform(T,testTbl);
```

Compare the test set performances of a linear classifier trained on the original features and a linear classifier trained on the new features.

Fit a linear model without transforming the data. Check the test set performance of the model using a confusion matrix.

```
originalMdl = fitclinear(trainTbl,trainY);
originalPredictedLabels = predict(originalMdl,testTbl);
cm = confusionchart(testY,originalPredictedLabels);
```

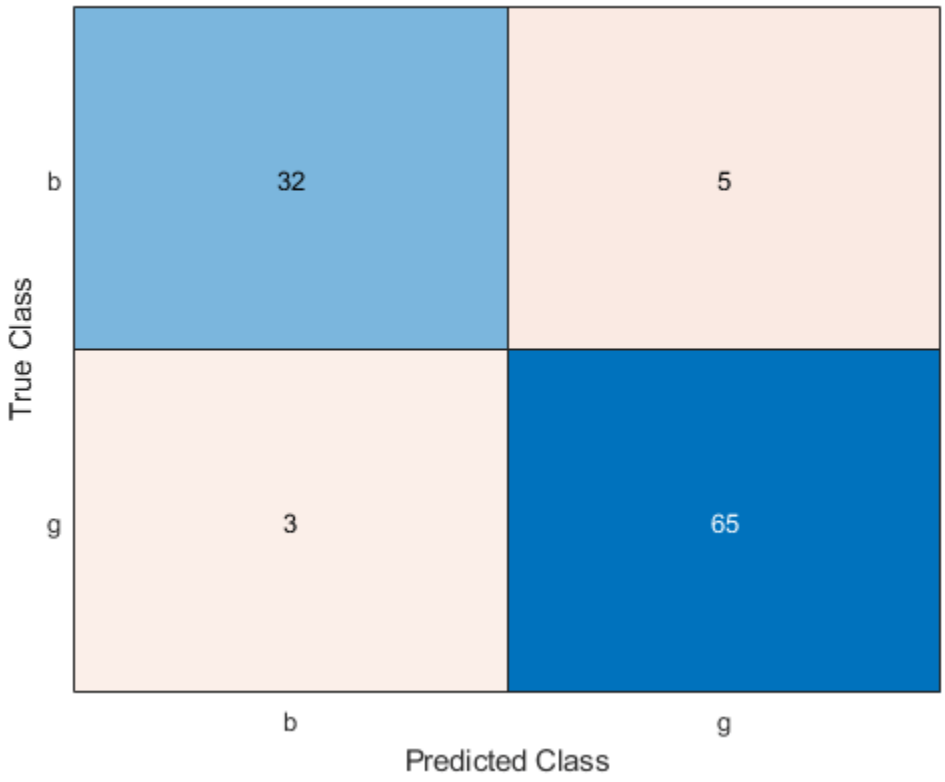



```
confusionMatrix = cm.NormalizedValues;
originalTestAccuracy = sum(diag(confusionMatrix))/sum(confusionMatrix,"all")
```

```
originalTestAccuracy = 0.8952
```

Fit a linear model with the transformed data. Check the test set performance of the model using a confusion matrix.

```
newMdl = fitclinear(newTrainTbl,trainY);
newPredictedLabels = predict(newMdl,newTestTbl);
newcm = confusionchart(testY,newPredictedLabels);
```




```
newConfusionMatrix = newcm.NormalizedValues;
newTestAccuracy = sum(diag(newConfusionMatrix))/sum(newConfusionMatrix,"all")
```

```
newTestAccuracy = 0.9238
```

The linear classifier trained on the transformed data seems to outperform the linear classifier trained on the original data.

Generate New Features to Improve Bagged Ensemble Accuracy

Use `gencfeatures` to engineer new features before training a bagged ensemble classifier. Before making predictions on new data, apply the same feature transformations to the new data set. Compare the test set performance of the ensemble that uses the engineered features to the test set performance of the ensemble that uses the original features.

Read the sample file `CreditRating_Historical.dat` into a table. The predictor data consists of financial ratios and industry sector information for a list of corporate customers. The response variable consists of credit ratings assigned by a rating agency. Preview the first few rows of the data set.

```
creditrating = readtable("CreditRating_Historical.dat");
head(creditrating)
```

```
ans=8×8 table
     ID      WC_TA     RE_TA     EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
```

```
    _____    ______    ______    _______    ________    _____    ________    _______

    62394     0.013     0.104     0.036      0.447      0.142        3       {'BB' }
    48608     0.232     0.335     0.062      1.969      0.281        8       {'A'  }
    42444     0.311     0.367     0.074      1.935      0.366        1       {'A'  }
    48631     0.194     0.263     0.062      1.017      0.228        4       {'BBB'}
    43768     0.121     0.413     0.057      3.647      0.466       12       {'AAA'}
    39255    -0.117    -0.799      0.01      0.179      0.082        4       {'CCC'}
    62236     0.087     0.158     0.049      0.816      0.324        2       {'BBB'}
    39354     0.005     0.181     0.034      2.597      0.388        7       {'AA' }
```

Because each value in the `ID` variable is a unique customer ID, that is, `length(unique(creditrating.ID))` is equal to the number of observations in `creditrating`, the `ID` variable is a poor predictor. Remove the `ID` variable from the table, and convert the `Industry` variable to a `categorical` variable.

```
creditrating = removevars(creditrating,"ID");
creditrating.Industry = categorical(creditrating.Industry);
```

Convert the `Rating` response variable to an ordinal `categorical` variable.

```
creditrating.Rating = categorical(creditrating.Rating, ...
    ["AAA","AA","A","BBB","BB","B","CCC"],"Ordinal",true);
```

Partition the data into training and test sets. Use approximately 75% of the observations as training data, and 25% of the observations as test data. Partition the data using `cvpartition`.

```
rng("default") % For reproducibility of the partition
c = cvpartition(creditrating.Rating,"Holdout",0.25);
trainingIndices = training(c); % Indices for the training set
testIndices = test(c); % Indices for the test set
creditTrain = creditrating(trainingIndices,:);
creditTest = creditrating(testIndices,:);
```

Use the training data to generate 40 new features to fit a bagged ensemble. By default, the 40 features can include original features if the software considers them to be important variables.

```
[T,newCreditTrain] = gencfeatures(creditTrain,"Rating",40, ...
    "TargetLearner","bag");
T
```

```
T =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'bag'
    NumEngineeredFeatures: 34
      NumOriginalFeatures: 6
         TotalNumFeatures: 40
```

Because `T.NumOriginalFeatures` is `6`, the function keeps all the original predictors.

Create `newCreditTest` by applying the transformations stored in the object `T` to the test data.

```
newCreditTest = transform(T,creditTest);
```

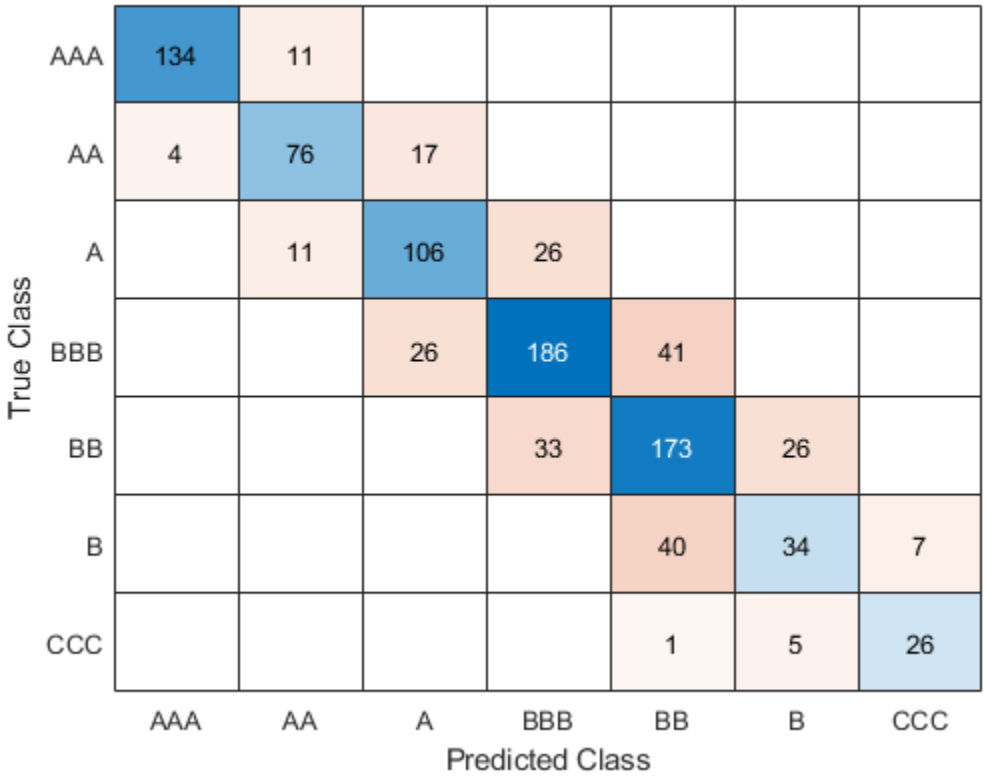
Compare the test set performances of a bagged ensemble trained on the original features and a bagged ensemble trained on the new features.

Train a bagged ensemble using the original training set `creditTrain`. Compute the accuracy of the model on the original test set `creditTest`. Visualize the results using a confusion matrix.

```
originalMdl = fitcensemble(creditTrain,"Rating","Method","Bag");
originalTestAccuracy = 1 - loss(originalMdl,creditTest, ...
    "Rating","LossFun","classiferror")
```

```
originalTestAccuracy = 0.7481
```

```
predictedTestLabels = predict(originalMdl,creditTest);
confusionchart(creditTest.Rating,predictedTestLabels);
```

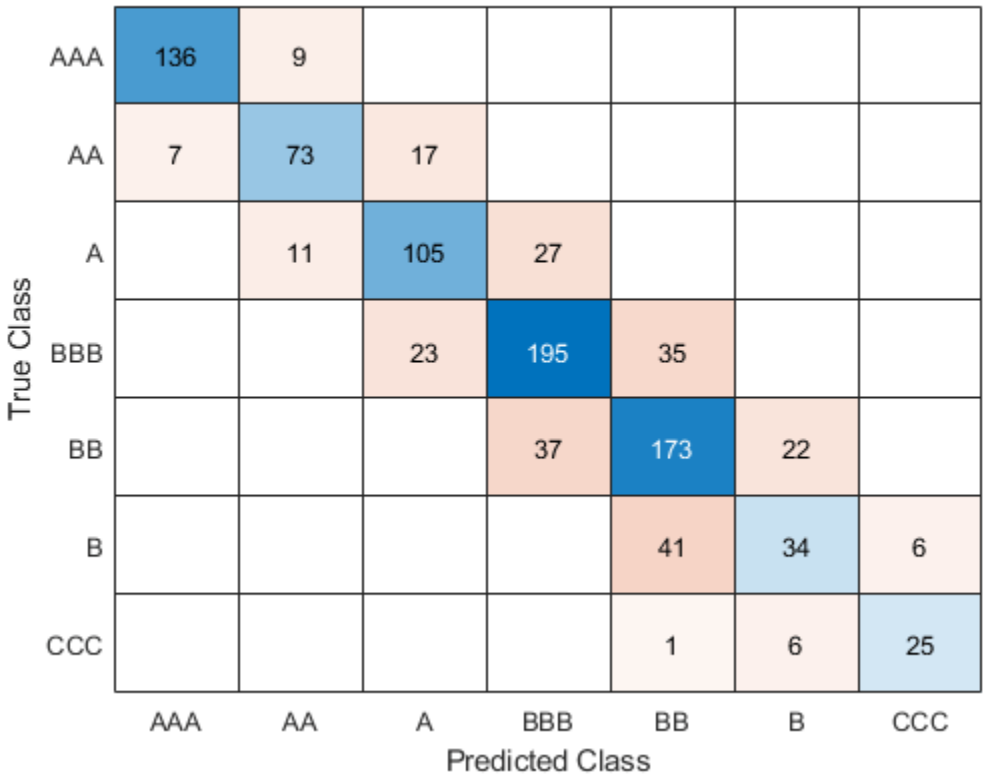


Train a bagged ensemble using the transformed training set `newCreditTrain`. Compute the accuracy of the model on the transformed test set `newCreditTest`. Visualize the results using a confusion matrix.

```
newMdl = fitcensemble(newCreditTrain,"Rating","Method","Bag");
newTestAccuracy = 1 - loss(newMdl,newCreditTest, ...
    "Rating","LossFun","classiferror")
```

```
newTestAccuracy = 0.7543
```

```
newPredictedTestLabels = predict(newMdl,newCreditTest);
confusionchart(newCreditTest.Rating,newPredictedTestLabels)
```



The bagged ensemble trained on the transformed data seems to outperform the bagged ensemble trained on the original data.

Compute Cross-Validation Loss Using Generated Features

Generate features to train a linear classifier. Compute the cross-validation classification error of the model by using the `crossval` function.

Load the `ionosphere` data set, and create a table containing the predictor data.

```
load ionosphere
Tbl = array2table(X);
```

Create a random partition for stratified 5-fold cross-validation.

```
rng("default") % For reproducibility of the partition
cvp = cvpartition(Y,"KFold",5);
```

Compute the cross-validation classification loss for a linear model trained on the original features in `Tbl`.

```
CVMdl = fitclinear(Tbl,Y,"CVPartition",cvp);
cvloss = kfoldLoss(CVMdl)
```

```
cvloss = 0.1339
```

Create the custom function `myloss` (shown at the end of this example). This function generates 20 features from the training data, and then applies the same training set transformations to the test data. The function then fits a linear classifier to the training data and computes the test set loss.

Note: If you use the live script file for this example, the `myloss` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Compute the cross-validation classification loss for a linear model trained on features generated from the predictors in `Tbl`.

```
newcvloss = mean(crossval(@myloss,Tbl,Y,"Partition",cvp))
```

```
newcvloss = 0.0770
```

```
function testloss = myloss(TrainTbl,trainY,TestTbl,testY)
[Transformer,NewTrainTbl] = gencfeatures(TrainTbl,trainY,20);
NewTestTbl = transform(Transformer,TestTbl);
Mdl = fitclinear(NewTrainTbl,trainY);
testloss = loss(Mdl,NewTestTbl,testY, ...
    "LossFun","classiferror");
end
```

Input Arguments

`Tbl` — Original features

table

Original features, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable, and you want to create new features from any of the remaining variables in `Tbl`, then specify the response variable by using `ResponseVarName`.

If `Tbl` contains the response variable, and you want to create new features from only a subset of the remaining variables in `Tbl`, then specify a formula by using `formula`.

If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl` as predictors, and might create new features from `Y`.

Data Types: `char` | `string`

q — Number of features

positive integer scalar

Number of features, specified as a positive integer scalar. For example, you can set `q` to approximately `1.5*size(Tbl,2)`, which is about 1.5 times the number of original features.

Data Types: `single` | `double`

Y — Response variable

numeric vector | categorical vector

Response variable, specified as a numeric or categorical vector. The length of `Y` must be equal to the number of rows in `Tbl`.

formula — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~X1+X2+X3'`. In this form, `Y` represents the response variable, and `X1`, `X2`, and `X3` represent the predictor variables.

To create new features from only a subset of the predictor variables in `Tbl`, use a formula. If you specify a formula, then the software does not create new features from any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `gencfeatures(Tbl,'Y',10,'TargetLearner','bag','FeatureSelection','oob')` specifies that the expected learner type is a bagged ensemble classifier and the method for selecting features is an out-of-bag, predictor importance technique.

TargetLearner — Expected learner type

`'linear'` (default) | `'bag'`

Expected learner type, specified as `'linear'` or `'bag'`. The software creates and selects new features assuming that they will be used to train this type of model.

Value	Expected Model
`'linear'`	`ClassificationLinear` — Appropriate for binary classification only

Value	Expected Model
`'bag'`	`ClassificationBaggedEnsemble` — Appropriate for binary and multiclass classification

By default, `TargetLearner` is `'linear'`, which supports binary response variables only. If you have a multiclass response variable and you want to generate new features, you must set `TargetLearner` to `'bag'`.

Example: `'TargetLearner','bag'`

`IncludeInputVariables` — Method for including original features in `Tbl`
`'auto'` (default) | `'include'` | `'select'` | `'omit'`

Method for including the original features in `Tbl` in the new table `NewTbl`, specified as one of the values in this table.

Value	Description
`'auto'`	This value is equivalent to: • `'select'` when `TargetLearner` is `'linear'` • `'include'` when `TargetLearner` is `'bag'`
`'include'`	The software includes all original features that can be used as predictors by the target learner, and excludes unsupported features, such as `datetime` and `duration` variables.
`'select'`	The software includes original features that are supported by the target learner and considered to be important by the specified feature selection method (`FeatureSelectionMethod`).
`'omit'`	The software omits the original features.

Example: `'IncludeInputVariables','include'`

Data Types: `logical`

`FeatureSelectionMethod` — Method for selecting new features
`'auto'` (default) | `'oob'` | `'mrmr'` | `'lasso'`

Method for selecting new features, specified as one of the values in this table. The software generates many features and uses this method to select the important features to include in `NewTbl`.

Value	Description
`'auto'`	This value is equivalent to: • `'lasso'` when `TargetLearner` is `'linear'` • `'oob'` when `TargetLearner` is `'bag'`

Value	Description
`'oob'`	Out-of-bag, predictor importance estimates by permutation — Available when `TargetLearner` is `'bag'`
`'mrmr'`	Minimum redundancy maximum relevance (MRMR) — Available when `TargetLearner` is `'linear'` or `'bag'`
`'lasso'`	Lasso regularization — Available when `TargetLearner` is `'linear'`

Example: `'FeatureSelection','mrmr'`

`TransformedDataStandardization` — Standardization method for transformed data
`'auto'` (default) | `'none'` | `'zscore'` | `'mad'` | `'range'`

Standardization method for the transformed data, specified as one of the values in this table.

Value	Description
`'auto'`	This value is equivalent to: • `'zscore'` when `TargetLearner` is `'linear'` • `'none'` when `TargetLearner` is `'bag'`
`'none'`	Use raw data
`'zscore'`	Center and scale to have mean 0 and standard deviation 1
`'mad'`	Center and scale to have median 0 and median absolute deviation 1
`'range'`	Scale range of data to [0,1]

Example: `'TransformedDataStandardization','range'`

`CategoricalEncodingLimit` — Maximum number of categories allowed in categorical predictor
nonnegative integer scalar | `Inf`

Maximum number of categories allowed in a categorical predictor, specified as a nonnegative integer scalar. If a categorical predictor has more than the specified number of categories, than `gencfeatures` does not create new features from the predictor. The default value is `50` when `TargetLearner` is `'linear'` and `Inf` when `TargetLearner` is `'ensemble'`.

Example: `'CategoricalEncodingLimit',20`

Data Types: `single` | `double`

Output Arguments

`Transformer` — Engineered feature transformer
`FeatureTransformer` object

Engineered feature transformer, returned as a `FeatureTransformer` object. To better understand the engineered features, use the `describe` object function of `Transformer`. To apply the same feature transformations on a new data set, use the `transform` object function of `Transformer`.

`NewTbl` — Generated features
table

Generated features, returned as a table. Each row corresponds to an observation, and each column corresponds to a generated feature. If the response variable is included in `Tbl`, then `NewTbl` also includes the response variable. Use this table to train a classification model of type `TargetLearner`.

`NewTbl` contains generated features in the following order: original features, engineered features as ranked by the feature selection method, and the response variable.

Tips

- By default, when `TargetLearner` is `'linear'`, the software generates new features from numeric predictors by using z-scores (see `TransformedDataStandardization`). You can change the type of standardization for the transformed features; however, using some method of standardization, thereby avoiding the `'none'` specification, is strongly recommended. Linear model fitting works best with standardized data.

See Also

`FeatureTransformer` | `describe` | `fitcensemble` | `fitclinear` | `plotPartialDependence` | `transform`

Topics

“Automated Feature Engineering for Classification” on page 18-190

Introduced in R2021a

geocdf

Geometric cumulative distribution function

Syntax

```
y = geocdf(x,p)
y = geocdf(x,p,'upper')
```

Description

`y = geocdf(x,p)` returns the cumulative distribution function (cdf) of the geometric distribution at each value in `x` using the corresponding probabilities in `p`. `x` and `p` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other input. The parameters in `p` must lie on the interval `[0,1]`.

`y = geocdf(x,p,'upper')` returns the complement of the geometric distribution cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

Examples

Compute Geometric Distribution cdf

Suppose you toss a fair coin repeatedly, and a "success" occurs when the coin lands with heads facing up. What is the probability of observing three or fewer tails ("failures") before tossing a heads?

To solve, determine the value of the cumulative distribution function (cdf) for the geometric distribution at x equal to 3. The probability of success (tossing a heads) p in any given trial is 0.5.

```
x = 3;
p = 0.5;
y = geocdf(x,p)
```

```
y = 0.9375
```

The returned value of y indicates that the probability of observing three or fewer tails before tossing a heads is 0.9375.

More About

Geometric Distribution cdf

The cumulative distribution function (cdf) of the geometric distribution is

$$y = F(x \mid p) = 1 - (1 - p)^{x + 1};\ x = 0, 1, 2, \ldots,$$

where p is the probability of success, and x is the number of failures before the first success. The result y is the probability of observing up to x trials before a success, when the probability of success in any given trial is p.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `geoinv` | `geopdf` | `geornd` | `geostat` | `mle`

Topics

“Geometric Distribution” on page B-63

Introduced before R2006a

geoinv

Geometric inverse cumulative distribution function

Syntax

```
x = geoinv(y,p)
```

Description

`x = geoinv(y,p)` returns the inverse cumulative distribution function (icdf) of the geometric distribution at each value in `y` using the corresponding probabilities in `p`.

`geoinv` returns the smallest positive integer `x` such that the geometric cdf evaluated at `x` is equal to or exceeds `y`. You can think of `y` as the probability of observing `x` successes in a row in independent trials, where `p` is the probability of success in each trial.

`y` and `p` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input for `p` or `y` is expanded to a constant array with the same dimensions as the other input. The values in `p` and `y` must lie on the interval `[0,1]`.

Examples

Compute Geometric Distribution icdf

Suppose the probability of a five-year-old car battery not starting in cold weather is 0.03. If we want no more than a ten percent chance that the car does not start, what is the maximum number of days in a row that we should try to start the car?

To solve, compute the inverse cdf of the geometric distribution. In this example, a "success" means the car does not start, while a "failure" means the car does start. The probability of success for each trial *p* equals 0.03, while the probability of observing *x* failures in a row before observing a success *y* equals 0.1.

```
y = 0.1;
p = 0.03;
x = geoinv(y,p)
```

```
x = 3
```

The returned result indicates that if we start the car three times, there is at least a ten percent chance that it will not start on one of those tries. Therefore, if we want no greater than a ten percent chance that the car will not start, we should only attempt to start it for a maximum of two days in a row.

We can confirm this result by evaluating the cdf at values of *x* equal to 2 and 3, given the probability of success for each trial *p* equal to 0.03.

```
y2 = geocdf(2,p)  % cdf for x = 2
```

```
y2 = 0.0873
```

```
y3 = geocdf(3,p)  % cdf for x = 3
```

```
y3 = 0.1147
```

The returned results indicate an 8.7% chance of the car not starting if we try two days in a row, and an 11.5% chance of not starting if we try three days in a row.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`geocdf` | `geopdf` | `geornd` | `geostat` | `icdf`

Topics

“Geometric Distribution” on page B-63

Introduced before R2006a

geomean

Geometric mean

Syntax

```
m = geomean(X)
m = geomean(X,'all')
m = geomean(X,dim)
m = geomean(X,vecdim)
m = geomean( ___ ,nanflag)
```

Description

`m = geomean(X)` returns the geometric mean on page 33-2659 of `X`.

- If `X` is a vector, then `geomean(X)` is the geometric mean of the elements in `X`.
- If `X` is a matrix, then `geomean(X)` is a row vector containing the geometric mean of each column of `X`.
- If `X` is a multidimensional array, then `geomean` operates along the first nonsingleton dimension of `X`.

`m = geomean(X,'all')` returns the geometric mean of all the elements in `X`.

`m = geomean(X,dim)` returns the geometric mean along the operating dimension `dim` of `X`.

`m = geomean(X,vecdim)` returns the geometric mean over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `geomean(X,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the geometric mean of the elements on the corresponding page of `X`.

`m = geomean( ___ ,nanflag)` specifies whether to exclude `NaN` values from the calculation, using any of the input argument combinations in previous syntaxes. By default, `geomean` includes `NaN` values in the calculation (`nanflag` has the value `'includenan'`). To exclude `NaN` values, set the value of `nanflag` to `'omitnan'`.

Examples

Compare Geometric and Arithmetic Mean

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a matrix of exponential random numbers with 5 rows and 4 columns.

```
X = exprnd(1,5,4)
```

```
X = 5×4
```

```
    0.2049    2.3275    1.8476    1.9527
    0.0989    1.2783    0.0298    0.8633
    2.0637    0.6035    0.0438    0.0880
    0.0906    0.0434    0.7228    0.2329
    0.4583    0.0357    0.2228    0.0414
```

Compute the geometric and arithmetic means of the columns of X.

```
geometric = geomean(X)
```

```
geometric = 1×4

    0.2805    0.3083    0.2079    0.2698
```

```
arithmetic = mean(X)
```

```
arithmetic = 1×4

    0.5833    0.8577    0.5734    0.6357
```

The arithmetic mean is greater than the geometric mean for all the columns of X.

Geometric Mean of All Elements

Find the geometric mean over multiple dimensions by using the `'all'` input argument.

Create a 2-by-5-by-4 array X.

```
X = reshape(1:40,[2 5 4])
```

```
X =
X(:,:,1) =

     1     3     5     7     9
     2     4     6     8    10


X(:,:,2) =

    11    13    15    17    19
    12    14    16    18    20


X(:,:,3) =

    21    23    25    27    29
    22    24    26    28    30


X(:,:,4) =

    31    33    35    37    39
    32    34    36    38    40
```

Find the geometric mean of all the elements of X.

```
m = geomean(X,'all')
```

```
m = 15.7685
```

m is the geometric mean of the entire array X.

Geometric Mean Along Specified Dimensions

Find the geometric mean along different operating dimensions and vectors of dimensions for a multidimensional array.

Create a 3-by-5-by-2 array X.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the geometric mean of X along the default dimension.

```
gmean1 = geomean(X)
```

```
gmean1 =
gmean1(:,:,1) =

    1.8171    4.9324    7.9581   10.9696   13.9761


gmean1(:,:,2) =

   16.9804   19.9833   22.9855   25.9872   28.9885
```

By default, geomean operates along the first dimension of X whose size does not equal 1. In this case, this dimension is the first dimension of X. Therefore, gmean1 is a 1-by-5-by-2 array.

Find the geometric mean of X along the second dimension.

```
gmean2 = geomean(X,2)
```

```
gmean2 =
gmean2(:,:,1) =
```

```
    5.1549
    6.5784
    7.8155


gmean2(:,:,2) =

   21.5814
   22.6004
   23.6177
```

`gmean2` is a 3-by-1-by-2 array.

Find the geometric mean of `X` along the third dimension.

```
gmean3 = geomean(X,3)

gmean3 = 3×5

    4.0000    8.7178   12.4097   15.8114   19.0788
    5.8310   10.0000   13.5647   16.9115   20.1494
    7.3485   11.2250   14.6969   18.0000   21.2132
```

`gmean3` is a 3-by-5 array.

Find the geometric mean of each page of `X` by specifying the first and second dimensions using the `vecdim` input argument.

```
mpage = geomean(X,[1 2])

mpage =
mpage(:,:,1) =

    6.4234


mpage(:,:,2) =

   22.5845
```

For example, `mpage(1,1,2)` is the geometric mean of the elements in `X(:,:,2)`.

Find the geometric mean of the elements in each `X(i,:,:)` slice by specifying the second and third dimensions.

```
mrow = geomean(X,[2 3])

mrow = 3×1

   10.5475
   12.1932
   13.5862
```

For example, `mrow(3)` is the geometric mean of the elements in `X(3,:,:)`, and is equivalent to specifying `geomean(X(3,:,:),'all')`.

Geometric Mean Excluding NaN

Create a vector and compute its `geomean`, excluding `NaN` values.

```
x = 1:10;
x(3) = nan; % Replace the third element of x with a NaN value
n = geomean(x,'omitnan')
```

```
n = 4.7408
```

If you do not specify `'omitnan'`, then `geomean(x)` returns `NaN`.

Input Arguments

X — Input data

nonnegative vector | nonnegative matrix | nonnegative multidimensional array

Input data that represents a sample from a population, specified as a nonnegative vector, matrix, or multidimensional array.

- If `X` is a vector, then `geomean(X)` is the geometric mean of the elements in `X`.
- If `X` is a matrix, then `geomean(X)` is a row vector containing the geometric mean of each column of `X`.
- If `X` is a multidimensional array, then `geomean` operates along the first nonsingleton dimension of `X`.

To specify the operating dimension when `X` is a matrix or an array, use the `dim` input argument.

Data Types: `single` | `double`

dim — Dimension

positive integer scalar

Dimension along which to operate, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension of `X` whose size does not equal 1.

Consider a two-dimensional array `X`:

- If `dim` is equal to 1, then `geomean(X,1)` returns a row vector containing the geometric mean for each column in `X`.
- If `dim` is equal to 2, then `geomean(X,2)` returns a column vector containing the geometric mean for each row in `X`.

If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `geomean` returns `X`.

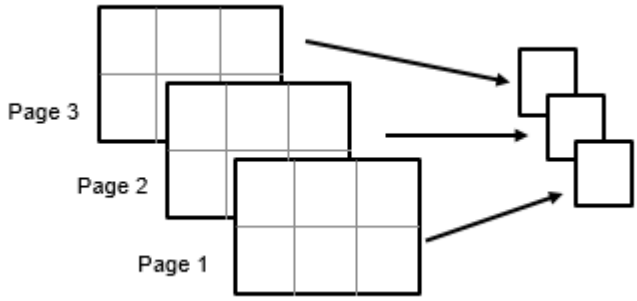
Data Types: `single` | `double`

vecdim — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `m` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `m`.

For example, if `X` is a 2-by-3-by-3 array, then `geomean(X,[1 2])` returns a 1-by-1-by-3 array. Each element of the output is the geometric mean of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

`nanflag` — `NaN` condition
`'includenan'` (default) | `'omitnan'`

`NaN` condition, specified as one of these values:

- `'includenan'` — Include `NaN` values when computing the `geomean`. This returns `NaN`.
- `'omitnan'` — Ignore `NaN` values in the input.

Data Types: `char` | `string`

Output Arguments

`m` — Geometric mean
scalar | vector | matrix | multidimensional array

Geometric mean, returned as a scalar, vector, matrix, or multidimensional array.

More About

Geometric Mean

The geometric mean of a sample `X` is

$$m = \left[\prod_{i=1}^{n} x_i \right]^{\frac{1}{n}}$$

where n is the number of values in `X`.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- These input arguments are not supported: `'all'`, `vecdim`, and `nanflag`.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`harmmean` | `mean` | `median` | `trimmean`

Topics

“Geometric Distribution” on page B-63

Introduced before R2006a

geopdf

Geometric probability density function

Syntax

```
y = geopdf(x,p)
```

Description

`y = geopdf(x,p)` returns the probability density function (pdf) of the geometric distribution at each value in `x` using the corresponding probabilities in `p`. `x` and `p` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other input. The parameters in `p` must lie on the interval `[0,1]`.

Examples

Compute Geometric Distribution pdf

Suppose you toss a fair coin repeatedly, and a "success" occurs when the coin lands with heads facing up. What is the probability of observing exactly three tails ("failures") before tossing a heads?

To solve, determine the value of the probability density function (pdf) for the geometric distribution at `x` equal to `3`. The probability of success (tossing a heads) `p` in any given trial is `0.5`.

```
x = 3;
p = 0.5;
y = geopdf(x,p)
```

```
y = 0.0625
```

The returned value of `y` indicates that the probability of observing exactly three tails before tossing a heads is `0.0625`.

More About

Geometric Distribution pdf

The probability density function (pdf) of the geometric distribution is

$$y = f(x \mid p) = p(1-p)^x \quad ; \quad x = 0, 1, 2, \ldots,$$

where p is the probability of success, and x is the number of failures before the first success. The result y is the probability of observing exactly x trials before a success, when the probability of success in any given trial is p. For discrete distributions, the pdf is also known as the probability mass function (pmf).

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`geocdf` | `geoinv` | `geornd` | `geostat` | `mle` | `pdf`

Topics

“Geometric Distribution” on page B-63

Introduced before R2006a

geornd

Geometric random numbers

Syntax

```
r = geornd(p)
r = geornd(p,m,n,...)
r = geornd(p,[m,n,...])
```

Description

`r = geornd(p)` generates random numbers from a geometric distribution with probability parameter `p`. `p` can be a vector, a matrix, or a multidimensional array. The size of `r` is equal to the size of `p`. The parameters in `p` must lie in the interval `[0,1]`.

`r = geornd(p,m,n,...)` or `r = geornd(p,[m,n,...])` generates a multidimensional `m`-by-`n`-by-`...` array containing random numbers from the geometric distribution with probability parameter `p`. `p` can be a scalar or an array of the same size as `r`.

The geometric distribution is useful to model the number of failures before one success in a series of independent trials, where each trial results in either success or failure, and the probability of success in any individual trial is the constant `p`.

Examples

Generate Random Numbers from Geometric Distribution

Generate a single random number from a geometric distribution with probability parameter *p* equal to 0.01.

```
rng default  % For reproducibility
p = 0.01;
r1 = geornd(0.01)
```

```
r1 = 20
```

The returned random number represents a single experiment in which 20 failures were observed before a success, where each independent trial has a probability of success *p* equal to 0.01.

Generate a 1-by-5 array of random numbers from a geometric distribution with probability parameter *p* equal to 0.01.

```
r2 = geornd(p,1,5)
```

```
r2 = 1×5

     9   205     9    45   231
```

Each random number in the returned array represents the result of an experiment to determine the number of failures observed before a success, where each independent trial has a probability of success *p* equal to 0.01.

Generate a 1-by-3 array containing one random number from each of the three geometric distributions corresponding to the parameters in the 1-by-3 array of probabilities *p*.

```
p = [0.01 0.1 0.5];
r3 = geornd(p,[1 3])
```

```
r3 = 1×3

   127     5     0
```

Each element of the returned 1-by-3 array `r3` contains one random number generated from the geometric distribution described by the corresponding parameter in `P`. For example, the first element in `r3` represents an experiment in which 127 failures were observed before a success, where each independent trial has a probability of success *p* equal to 0.01. The second element in `r3` represents an experiment in which 5 failures were observed before a success, where each independent trial has a probability of success *p* equal to 0.1. The third element in `r3` represents an experiment in which zero failures were observed before a success - in other words, the first attempt was a success - where each independent trial has a probability of success *p* equal to 0.5.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`geocdf` | `geoinv` | `geopdf` | `geostat` | `random`

Topics

“Geometric Distribution” on page B-63

Introduced before R2006a

geostat

Geometric mean and variance

Syntax

```
[m,v] = geostat(p)
```

Description

`[m,v] = geostat(p)` returns the mean `m` and variance `v` of a geometric distribution with corresponding probability parameters in `p`. `p` can be a vector, a matrix, or a multidimensional array. The parameters in `p` must lie in the interval `[0,1]`.

Examples

Compute Mean and Variance of Geometric Distribution

Define a probability vector that contains six different parameter values.

```
p = 1./(1:6)
```

```
p = 1×6

    1.0000    0.5000    0.3333    0.2500    0.2000    0.1667
```

Compute the mean and variance of the geometric distribution that corresponds to each value contained in probability vector.

```
[m,v] = geostat(1./(1:6))
```

```
m = 1×6

         0    1.0000    2.0000    3.0000    4.0000    5.0000


v = 1×6

         0    2.0000    6.0000   12.0000   20.0000   30.0000
```

The returned values indicate that, for example, the mean of a geometric distribution with probability parameter *p* equal to 1/3 is 2, and its variance is 6.

More About

Geometric Distribution Mean and Variance

The mean of the geometric distribution is mean = $\frac{1-p}{p}$, and the variance of the geometric distribution is var = $\frac{1-p}{p^2}$, where p is the probability of success.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`geocdf` | `geoinv` | `geopdf` | `geornd`

Topics
“Geometric Distribution” on page B-63

Introduced before R2006a

GapEvaluation

Package: `clustering.evaluation`
Superclasses: `ClusterCriterion`

Gap criterion clustering evaluation object

Description

`GapEvaluation` is an object consisting of sample data, clustering data, and gap criterion values used to evaluate the optimal number of clusters. Create a gap criterion clustering evaluation object using `evalclusters`.

Construction

`eva = evalclusters(x,clust,'Gap')` creates a gap criterion clustering evaluation object.

`eva = evalclusters(x,clust,'Gap',Name,Value)` creates a gap criterion clustering evaluation object using additional options specified by one or more name-value pair arguments.

Input Arguments

x — Input data

matrix

Input data, specified as an *N*-by-*P* matrix. *N* is the number of observations, and *P* is the number of variables.

Data Types: `single` | `double`

clust — Clustering algorithm

`'kmeans'` | `'linkage'` | `'gmdistribution'` | matrix of clustering solutions | function handle

Clustering algorithm, specified as one of the following.

`'kmeans'`	Cluster the data in `x` using the `kmeans` clustering algorithm, with `'EmptyAction'` set to `'singleton'` and `'Replicates'` set to `5`.
`'linkage'`	Cluster the data in `x` using the `clusterdata` agglomerative clustering algorithm, with `'Linkage'` set to `'ward'`.
`'gmdistribution'`	Cluster the data in `x` using the `gmdistribution` Gaussian mixture distribution algorithm, with `'SharedCov'` set to `true` and `'Replicates'` set to `5`.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can specify a clustering algorithm using a function handle. The function must be of the form `C = clustfun(DATA,K)`, where `DATA` is the data to be clustered, and `K` is the number of clusters. The output of `clustfun` must be one of the following:

- A vector of integers representing the cluster index for each observation in `DATA`. There must be `K` unique values in this vector.

- A numeric *n*-by-*K* matrix of score for *n* observations and *K* classes. In this case, the cluster index for each observation is determined by taking the largest score value in each row.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can also specify `clust` as a *n*-by-*K* matrix containing the proposed clustering solutions. *n* is the number of observations in the sample data, and *K* is the number of proposed clustering solutions. Column *j* contains the cluster indices for each of the *N* points in the *j*th clustering solution.

Data Types: `single` | `double` | `char` | `string` | `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KList',[1:5],'Distance','cityblock'` specifies to test 1, 2, 3, 4, and 5 clusters using the city block distance metric.

B — Number of reference data sets

`100` (default) | positive integer value

Number of reference data sets generated from the reference distribution `ReferenceDistribution`, specified as the comma-separated pair consisting of `'B'` and a positive integer value.

Example: `'B',150`

Data Types: `single` | `double`

Distance — Distance metric

`'sqEuclidean'` (default) | `'Euclidean'` | `'cityblock'` | function | ...

Distance metric used for computing the criterion values, specified as the comma-separated pair consisting of `'Distance'` and one of the following.

`'sqEuclidean'`	Squared Euclidean distance
`'Euclidean'`	Euclidean distance
`'cityblock'`	Sum of absolute differences
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)

For detailed information about each distance metric, see `pdist`.

You can also specify a function for the distance metric by using a function handle. The distance function must be of the form

```
d2 = distfun(XI,XJ),
```

where `XI` is a 1-by-*n* vector corresponding to a single row of the input matrix `X`, and `XJ` is an m_2-by-*n* matrix corresponding to multiple rows of `X`. `distfun` must return an m_2-by-1 vector of distances `d2`, whose *k*th element is the distance between `XI` and `XJ(k,:)`.

`Distance` only accepts a function handle if the clustering algorithm `clust` accepts a function handle as the distance metric. For example, the `kmeans` clustering algorithm does not accept a function

handle as the distance metric. Therefore, if you use the `kmeans` algorithm and then specify a function handle for `Distance`, the software errors.

- When `clust` is `'kmeans'` or `'gmdistribution'`, `evalclusters` uses the distance metric specified for `Distance` to cluster the data.
- If `clust` is `'linkage'`, and `Distance` is either `'sqEuclidean'` or `'Euclidean'`, then the clustering algorithm uses Euclidean distance and Ward linkage.
- If `clust` is `'linkage'` and `Distance` is any other metric, then the clustering algorithm uses the specified distance metric and average linkage.
- In all other cases, the distance metric specified for `Distance` must match the distance metric used in the clustering algorithm to obtain meaningful results.

Example: `'Distance','Euclidean'`

Data Types: `single` | `double` | `char` | `string` | `function_handle`

`KList` — List of number of clusters to evaluate
vector

List of number of clusters to evaluate, specified as the comma-separated pair consisting of `'KList'` and a vector of positive integer values. You must specify `KList` when `clust` is a clustering algorithm name or a function handle. When `criterion` is `'gap'`, `clust` must be a character vector, a string scalar, or a function handle, and you must specify `KList`.

Example: `'KList',[1:6]`

Data Types: `single` | `double`

`ReferenceDistribution` — Reference data generation method
`'PCA'` (default) | `'uniform'`

Reference data generation method, specified as the comma-separated pair consisting of `'ReferenceDistributions'` and one of the following.

`'PCA'`	Generate reference data from a uniform distribution over a box aligned with the principal components of the data matrix `x`.
`'uniform'`	Generate reference data uniformly over the range of each feature in the data matrix `x`.

Example: `'ReferenceDistribution','uniform'`

`SearchMethod` — Method for selecting optimal number of clusters
`'globalMaxSE'` (default) | `'firstMaxSE'`

Method for selecting the optimal number of clusters, specified as the comma-separated pair consisting of `'SearchMethod'` and one of the following.

`'globalMaxSE'`	Evaluate each proposed number of clusters in `KList` and select the smallest number of clusters satisfying $$\text{Gap}(K) \geq GAPMAX - \text{SE}(GAPMAX),$$ where K is the number of clusters, $\text{Gap}(K)$ is the gap value for the clustering solution with K clusters, $GAPMAX$ is the largest gap value, and $\text{SE}(GAPMAX)$ is the standard error corresponding to the largest gap value.
`'firstMaxSE'`	Evaluate each proposed number of clusters in `KList` and select the smallest number of clusters satisfying $$\text{Gap}(K) \geq \text{Gap}(K+1) - \text{SE}(K+1),$$ where K is the number of clusters, $\text{Gap}(K)$ is the gap value for the clustering solution with K clusters, and $\text{SE}(K+1)$ is the standard error of the clustering solution with $K+1$ clusters.

Example: `'SearchMethod','globalMaxSE'`

Properties

`B`

Number of data sets generated from the reference distribution, stored as a positive integer value.

`ClusteringFunction`

Clustering algorithm used to cluster the input data, stored as a valid clustering algorithm name or function handle. If the clustering solutions are provided in the input, `ClusteringFunction` is empty.

`CriterionName`

Name of the criterion used for clustering evaluation, stored as a valid criterion name.

`CriterionValues`

Criterion values corresponding to each proposed number of clusters in `InspectedK`, stored as a vector of numerical values.

`Distance`

Distance metric used for clustering data, stored as a valid distance metric name.

`ExpectedLogW`

Expectation of the natural logarithm of W based on the generated reference data, stored as a vector of scalar values. W is the within-cluster dispersion computed using the distance metric `Distance`.

`InspectedK`

List of the number of proposed clusters for which to compute criterion values, stored as a vector of positive integer values.

LogW

Natural logarithm of *W* based on the input data, stored as a vector of scalar values. *W* is the within-cluster dispersion computed using the distance metric `Distance`.

Missing

Logical flag for excluded data, stored as a column vector of logical values. If `Missing` equals `true`, then the corresponding value in the data matrix `x` is not used in the clustering solution.

NumObservations

Number of observations in the data matrix `X`, minus the number of missing (`NaN`) values in `X`, stored as a positive integer value.

OptimalK

Optimal number of clusters, stored as a positive integer value.

OptimalY

Optimal clustering solution corresponding to `OptimalK`, stored as a column vector of positive integer values. If the clustering solutions are provided in the input, `OptimalY` is empty.

ReferenceDistribution

Reference data generation method, stored as a valid reference distribution name.

SE

Standard error of the natural logarithm of *W* with respect to the reference data for each number of clusters in `InspectedK`, stored as a vector of scalar values. *W* is the within-cluster dispersion computed using the distance metric `Distance`.

SearchMethod

Method for determining the optimal number of clusters, stored as a valid search method name.

StdLogW

Standard deviation of the natural logarithm of *W* with respect to the reference data for each number of clusters in `InspectedK`. *W* is the within-cluster dispersion computed using the distance metric `Distance`.

X

Data used for clustering, stored as a matrix of numerical values.

Methods

increaseB	Increase reference data sets

Inherited Methods

addK	Evaluate additional numbers of clusters
plot	Plot clustering evaluation object criterion values
compact	Compact clustering evaluation object

Examples

Evaluate Clustering Solution Using Gap Criterion

Evaluate the optimal number of clusters using the gap clustering evaluation criterion.

Load the sample data.

```
load fisheriris
```

The data contains sepal and petal measurements from three species of iris flowers.

Evaluate the number of clusters based on the gap criterion values. Cluster the data using `kmeans`.

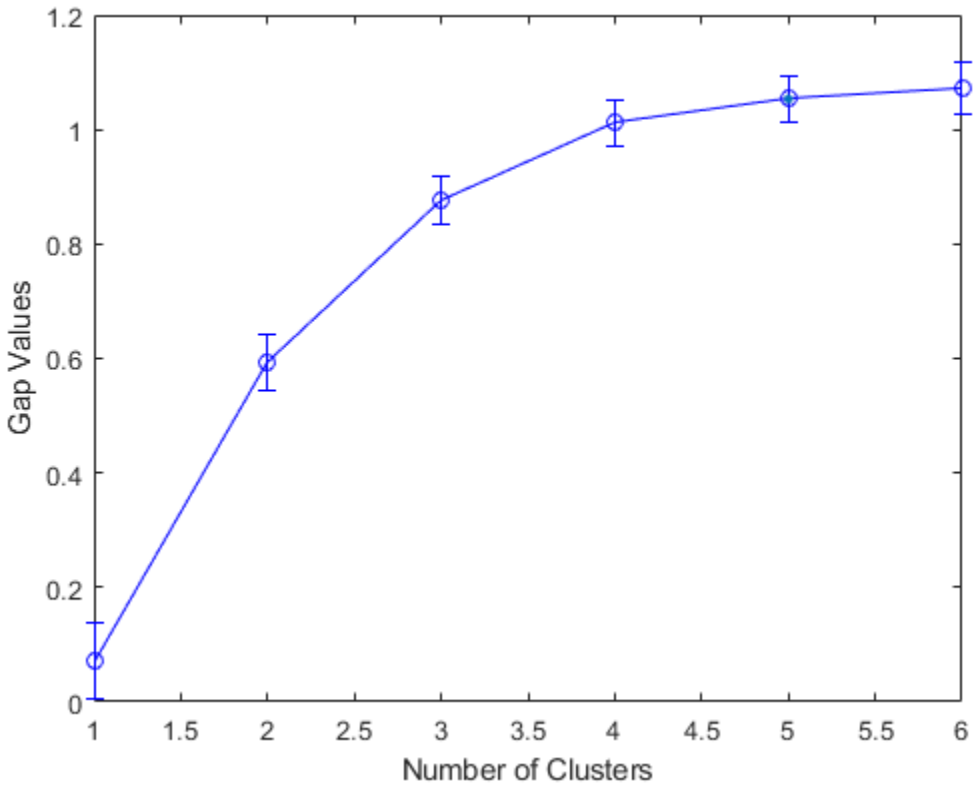
```
rng('default');  % For reproducibility
eva = evalclusters(meas,'kmeans','gap','KList',[1:6])
```

```
eva =
  GapEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [0.0720 0.5928 0.8762 1.0114 1.0534 1.0720]
           OptimalK: 5
```

The `OptimalK` value indicates that, based on the gap criterion, the optimal number of clusters is five.

Plot the gap criterion values for each number of clusters tested.

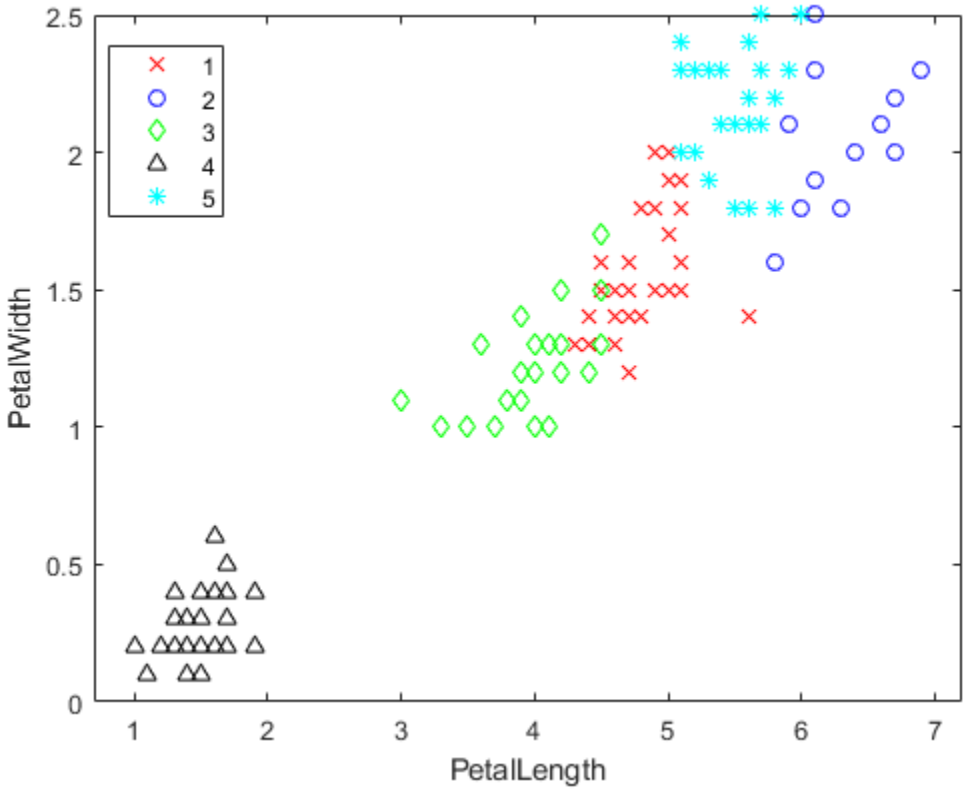
```
plot(eva)
```



Based on the plot, the maximum value of the gap criterion occurs at six clusters. However, the value at five clusters is within one standard error of the maximum, so the suggested optimal number of clusters is five.

Create a grouped scatter plot to examine the relationship between petal length and width. Group the data by suggested clusters.

```
figure
PetalLength = meas(:,3);
PetalWidth = meas(:,4);
ClusterGroup = eva.OptimalY;
gscatter(PetalLength,PetalWidth,ClusterGroup,'rbgkc','xod^*');
```



The plot shows cluster 4 in the lower-left corner, completely separated from the other four clusters. Cluster 4 contains flowers with the smallest petal widths and lengths. Cluster 2 is in the upper-right corner and contains flowers with the largest petal widths and lengths. Cluster 5 is next to cluster 2 and contains flowers with similar petal widths as the flowers in cluster 2, but smaller petal lengths than the flowers in cluster 2. Clusters 1 and 3 are near the center of the plot and contain flowers with measurements between the extremes.

More About

Gap Value

A common graphical approach to cluster evaluation involves plotting an error measurement versus several proposed numbers of clusters, and locating the “elbow” of this plot. The “elbow” occurs at the most dramatic decrease in error measurement. The gap criterion formalizes this approach by estimating the “elbow” location as the number of clusters with the largest gap value. Therefore, under the gap criterion, the optimal number of clusters occurs at the solution with the largest local or global gap value within a tolerance range.

The gap value is defined as

$$Gap_n(k) = E_n^*\{\log(W_k)\} - \log(W_k),$$

where n is the sample size, k is the number of clusters being evaluated, and W_k is the pooled within-cluster dispersion measurement

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r,$$

where n_r is the number of data points in cluster r, and D_r is the sum of the pairwise distances for all points in cluster r.

The expected value $E_n^*\{\log(W_k)\}$ is determined by Monte Carlo sampling from a reference distribution, and `log(`W_k`)` is computed from the sample data.

The gap value is defined even for clustering solutions that contain only one cluster, and can be used with any distance metric. However, the gap criterion is more computationally expensive than other cluster evaluation criteria, because the clustering algorithm must be applied to the reference data for each proposed clustering solution.

References

[1] Tibshirani, R., G. Walther, and T. Hastie. “Estimating the number of clusters in a data set via the gap statistic.” *Journal of the Royal Statistical Society: Series B*. Vol. 63, Part 2, 2001, pp. 411–423.

See Also

`CalinskiHarabaszEvaluation` | `DaviesBouldinEvaluation` | `SilhouetteEvaluation` | `evalclusters`

Topics
Class Attributes
Property Attributes

get

Class: `dataset`

(Not Recommended) Access dataset array properties

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
get(A)
s = get(A)
p = get(A,PropertyName)
p = get(A,{PropertyName1,PropertyName2,...})
```

Description

`get(A)` displays a list of property/value pairs for the dataset array `A`.

`s = get(A)` returns the values in a scalar structure `s` with field names given by the properties.

`p = get(A,PropertyName)` returns the value of the property specified by *`PropertyName`*.

`p = get(A,{PropertyName1,PropertyName2,...})` allows multiple property names to be specified and returns their values in a cell array.

Examples

Create a dataset array from Fisher's iris data and access the information.

```
load fisheriris
NumObs = size(meas,1);
NameObs = strcat({'Obs'},num2str((1:NumObs)','%-d'));
iris = dataset({nominal(species),'species'},...
               {meas,'SL','SW','PL','PW'},...
               'ObsNames',NameObs);

get(iris)
   Description: ''
   Units: {}
   DimNames: {'Observations' 'Variables'}
   UserData: []
   ObsNames: {150x1 cell}
   VarNames: {'species' 'SL' 'SW' 'PL' 'PW'}

ON = get(iris,'ObsNames');
ON(1:3)
ans =
    'Obs1'
    'Obs2'
    'Obs3'
```

See Also

`set` | `summary`

getlabels

(Not Recommended) Access nominal or ordinal array labels

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
labels = getlabels(A)
```

Description

`labels = getlabels(A)` returns the labels of the levels in the nominal or ordinal array `A` as a cell array of character vectors, `labels`. If `A` is an ordinal array, `getlabels` returns the labels in the order of the levels.

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

See Also

`getlevels` | `nominal` | `ordinal`

Topics

“Change Category Labels” on page 2-7

Introduced in R2007a

getlevels

(Not Recommended) Access nominal or ordinal array levels

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
L = getlevels(A)
```

Description

`L = getlevels(A)` returns the levels in the nominal or ordinal array `A`. `L` is a vector of the same type as `A`.

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

See Also

`getlabels` | `nominal` | `ordinal`

Topics

“Add and Drop Category Levels” on page 2-18
“Merge Category Levels” on page 2-16
“Reorder Category Levels” on page 2-9

Introduced in R2009a

gevcdf

Generalized extreme value cumulative distribution function

Syntax

```
p = gevcdf(x,k,sigma,mu)
p = gevcdf(x,k,sigma,mu,'upper')
```

Description

`p = gevcdf(x,k,sigma,mu)` returns the cdf of the generalized extreme value (GEV) distribution with shape parameter `k`, scale parameter `sigma`, and location parameter, `mu`, evaluated at the values in `x`. The size of `p` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

`p = gevcdf(x,k,sigma,mu,'upper')` returns the complement of the cdf of the GEV distribution, using an algorithm that more accurately computes the extreme upper tail probabilities.

Default values for `k`, `sigma`, and `mu` are 0, 1, and 0, respectively.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblcdf` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evcdf` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `gevfit` | `gevinv` | `gevlike` | `gevpdf` | `gevrnd` | `gevstat`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevfit

Generalized extreme value parameter estimates

Syntax

```
parmhat = gevfit(X)
[parmhat,parmci] = gevfit(X)
[parmhat,parmci] = gevfit(X,alpha)
[...] = gevfit(X,alpha,options)
```

Description

`parmhat = gevfit(X)` returns maximum likelihood estimates of the parameters for the generalized extreme value (GEV) distribution given the data in X. `parmhat(1)` is the shape parameter, `k`, `parmhat(2)` is the scale parameter, `sigma`, and `parmhat(3)` is the location parameter, `mu`.

`[parmhat,parmci] = gevfit(X)` returns 95% confidence intervals for the parameter estimates.

`[parmhat,parmci] = gevfit(X,alpha)` returns `100(1-alpha)`% confidence intervals for the parameter estimates.

`[...] = gevfit(X,alpha,options)` specifies control parameters for the iterative algorithm used to compute ML estimates. This argument can be created by a call to `statset`. See `statset('gevfit')` for parameter names and default values. Pass in `[]` for `alpha` to use the default values.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblfit` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evfit` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution is defined for `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

See Also

`gevcdf` | `gevinv` | `gevlike` | `gevpdf` | `gevrnd` | `gevstat` | `mle`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevinv

Generalized extreme value inverse cumulative distribution function

Syntax

```
X = gevinv(P,k,sigma,mu)
```

Description

`X = gevinv(P,k,sigma,mu)` returns the inverse cdf of the generalized extreme value (GEV) distribution with shape parameter `k`, scale parameter `sigma`, and location parameter `mu`, evaluated at the values in `P`. The size of `X` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

Default values for `k`, `sigma`, and `mu` are 0, 1, and 0, respectively.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblinv` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evinv` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that $k*(X-mu)/sigma > -1$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gevcdf` | `gevfit` | `gevlike` | `gevpdf` | `gevrnd` | `gevstat` | `icdf`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevlike

Generalized extreme value negative log-likelihood

Syntax

```
nlogL = gevlike(params,data)
[nlogL,ACOV] = gevlike(params,data)
```

Description

`nlogL = gevlike(params,data)` returns the negative of the log-likelihood `nlogL` for the generalized extreme value (GEV) distribution, evaluated at parameters `params`. `params(1)` is the shape parameter, `k`, `params(2)` is the scale parameter, `sigma`, and `params(3)` is the location parameter, `mu`.

`[nlogL,ACOV] = gevlike(params,data)` returns the inverse of Fisher's information matrix, `ACOV`. If the input parameter values in `params` are the maximum likelihood estimates, the diagonal elements of `ACOV` are their asymptotic variances. `ACOV` is based on the observed Fisher's information, not the expected information.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wbllike` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evlike` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah.*Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

See Also

`gevcdf` | `gevfit` | `gevinv` | `gevpdf` | `gevrnd` | `gevstat`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevpdf

Generalized extreme value probability density function

Syntax

```
Y = gevpdf(X,k,sigma,mu)
```

Description

`Y = gevpdf(X,k,sigma,mu)` returns the pdf of the generalized extreme value (GEV) distribution with shape parameter `k`, scale parameter `sigma`, and location parameter, `mu`, evaluated at the values in `X`. The size of `Y` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

Default values for `k`, `sigma`, and `mu` are 0, 1, and 0, respectively.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblpdf` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evcdf` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gevcdf` | `gevfit` | `gevinv` | `gevlike` | `gevrnd` | `gevstat` | `pdf`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevrnd

Generalized extreme value random numbers

Syntax

```
R = gevrnd(k,sigma,mu)
R = gevrnd(k,sigma,mu,m,n,...)
R = gevrnd(k,sigma,mu,[m,n,...])
```

Description

`R = gevrnd(k,sigma,mu)` returns an array of random numbers chosen from the generalized extreme value (GEV) distribution with shape parameter `k`, scale parameter `sigma`, and location parameter, `mu`. The size of `R` is the common size of the input arguments if all are arrays. If any parameter is a scalar, the size of `R` is the size of the other parameters.

`R = gevrnd(k,sigma,mu,m,n,...)` or `R = gevrnd(k,sigma,mu,[m,n,...])` generates an `m`-by-`n`-by-... array containing random numbers from the GEV distribution with parameters `k`, `sigma`, and `mu`. The `k`, `sigma`, `mu` parameters can each be scalars or arrays of the same size as `R`.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblrnd` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evrnd` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gevcdf` | `gevfit` | `gevinv` | `gevlike` | `gevpdf` | `gevstat` | `random`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gevstat

Generalized extreme value mean and variance

Syntax

```
[M,V] = gevstat(k,sigma,mu)
```

Description

`[M,V] = gevstat(k,sigma,mu)` returns the mean of and variance for the generalized extreme value (GEV) distribution with shape parameter `k`, scale parameter `sigma`, and location parameter, `mu`. The sizes of `M` and `V` are the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

When `k < 0`, the GEV is the type III extreme value distribution. When `k > 0`, the GEV distribution is the type II, or Frechet, extreme value distribution. If `w` has a Weibull distribution as computed by the `wblstat` function, then `-w` has a type III extreme value distribution and `1/w` has a type II extreme value distribution. In the limit as `k` approaches 0, the GEV is the mirror image of the type I extreme value distribution as computed by the `evstat` function.

The mean of the GEV distribution is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. The GEV distribution has positive density only for values of `X` such that `k*(X-mu)/sigma > -1`.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gevcdf` | `gevfit` | `gevinv` | `gevlike` | `gevpdf` | `gevrnd`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced before R2006a

gline

Interactively add line to plot

Syntax

```
gline(h)
gline
hline = gline(...)
```

Description

`gline(h)` allows you to draw a line segment in the figure with handle `h` by clicking the pointer at the two endpoints. A rubber-band line tracks the pointer movement.

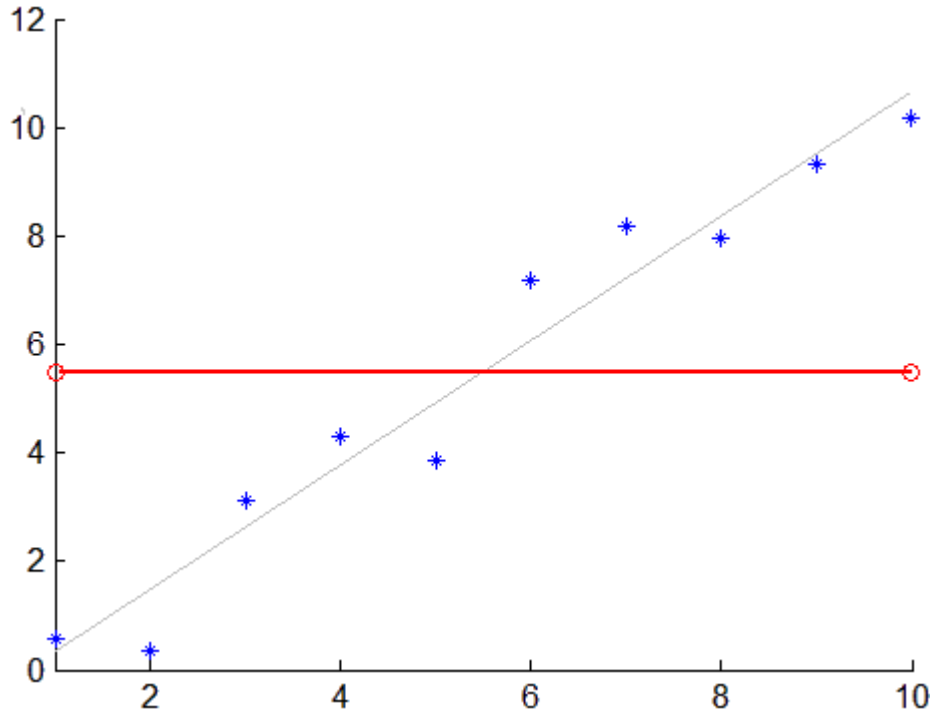
`gline` with no input arguments defaults to `h = gcf` and draws in the current figure.

`hline = gline(...)` returns the handle `hline` to the line.

Examples

Use `gline` to connect two points in a plot:

```
x = 1:10;

y = x + randn(1,10);
scatter(x,y,25,'b','*')

lsline

mu = mean(y);
hold on
plot([1 10],[mu mu],'ro')

hline = gline; % Connect circles
set(hline,'Color','r')
```



See Also

`lsline` | `refcurve` | `refline`

Introduced before R2006a

glmfit

Fit generalized linear regression model

Syntax

```
b = glmfit(X,y,distr)
b = glmfit(X,y,distr,Name,Value)
[b,dev] = glmfit(___)
[b,dev,stats] = glmfit(___)
```

Description

`b = glmfit(X,y,distr)` returns a vector `b` of coefficient estimates for a generalized linear regression model of the responses in `y` on the predictors in `X`, using the distribution `distr`.

`b = glmfit(X,y,distr,Name,Value)` specifies additional options using one or more name-value arguments. For example, you can specify `'Constant','off'` to omit the constant term from the model.

`[b,dev] = glmfit(___)` also returns the value `dev`, the deviance on page 33-2701 of the fit.

`[b,dev,stats] = glmfit(___)` also returns the model statistics `stats`.

Examples

Fit Generalized Linear Model with Probit Link

Fit a generalized linear regression model, and compute predicted (estimated) values for the predictor data using the fitted model.

Create a sample data set.

```
x = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
n = [48 42 31 34 31 21 23 23 21 16 17 21]';
y = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

`x` contains the predictor variable values. Each `y` value is the number of successes in the corresponding number of trials in `n`.

Fit a probit regression model for `y` on `x`.

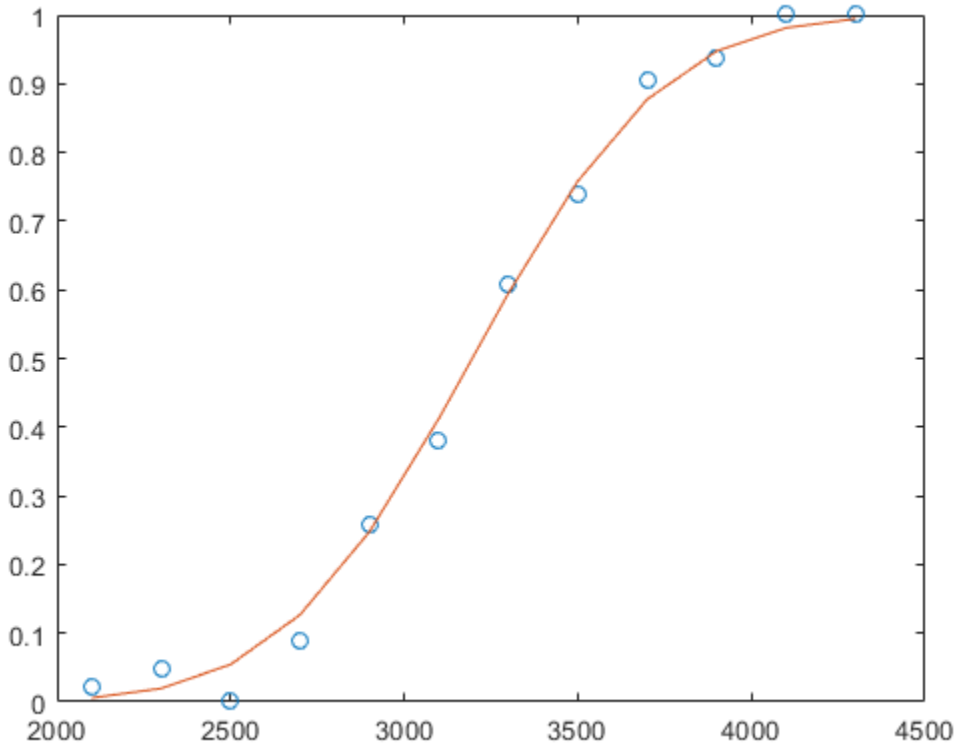
```
b = glmfit(x,[y n],'binomial','Link','probit');
```

Compute the estimated number of successes.

```
yfit = glmval(b,x,'probit','Size',n);
```

Plot the observed success percent and estimated success percent versus the `x` values.

```
plot(x,y./n,'o',x,yfit./n,'-')
```



Fit Generalized Linear Model Using Custom Link Function

Define a custom link function and use it to fit a generalized linear regression model.

Load the sample data.

```
load fisheriris
```

The column vector `species` contains iris flowers of three different species: setosa, versicolor, and virginica. The matrix `meas` contains four types of measurements for the flowers, the length and width of sepals and petals in centimeters.

Define the predictor variables and response variable.

```
X = meas(51:end,:);
y = strcmp('versicolor',species(51:end));
```

Define a custom link function for a logit link function. Create three function handles that define the link function, the derivative of the link function, and the inverse link function. Store them in a cell array.

```
link = @(mu) log(mu./(1-mu));
derlink = @(mu) 1./(mu.*(1-mu));
invlink = @(resp) 1./(1+exp(-resp));
F = {link,derlink,invlink};
```

Fit a logistic regression model using `glmfit` with the custom link function.

```
b = glmfit(X,y,'binomial','link',F)
```

```
b = 5×1

   42.6378
    2.4652
    6.6809
   -9.4294
  -18.2861
```

Fit a generalized linear model by using the built-in `logit` link function, and compare the results.

```
b = glmfit(X,y,'binomial','link','logit')
```

```
b = 5×1

   42.6378
    2.4652
    6.6809
   -9.4294
  -18.2861
```

Perform Deviance Test

Fit a generalized linear regression model that contains an intercept and linear term for each predictor. Perform a deviance test that determines whether the model fits significantly better than a constant model.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Fit a generalized linear regression model that contains an intercept and linear term for each predictor.

```
[b,dev] = glmfit(X,y,'poisson');
```

The second output argument `dev` is a “Deviance” on page 33-2701 of the fit.

Fit a generalized linear regression model that contains only an intercept. Specify the predictor variable as a column of 1s, and specify `'Constant'` as `'off'` so that `glmfit` does not include a constant term in the model.

```
[~,dev_noconstant] = glmfit(ones(100,1),y,'poisson','Constant','off');
```

Compute the difference between `dev_constant` and `dev`.

```
D = dev_noconstant - dev

D = 2.9533e+05
```

`D` has a chi-square distribution with 2 degrees of freedom. The degrees of freedom equal the difference in the number of estimated parameters in the model corresponding to `dev` and the number of estimated parameters in the constant model. Find the *p*-value for a deviance test.

```
p = 1 - chi2cdf(D,2)

p = 0
```

The small *p*-value indicates that the model differs significantly from a constant.

Alternatively, you can create a generalized linear regression model of Poisson data by using the `fitglm` function. The model display includes the statistic (`Chi^2-statistic vs. constant model`) and *p*-value.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')

mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

You can also use the `devianceTest` function with the fitted model object.

```
devianceTest(mdl)

ans=2×4 table
                            Deviance     DFE     chi2Stat     pValue
                           __________    ___    __________    ______

    log(y) ~ 1             2.9544e+05    99
    log(y) ~ 1 + x1 + x2         107.4    97     2.9533e+05       0
```

Input Arguments

X — Predictor variables

numeric matrix

Predictor variables, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

By default, `glmfit` includes a constant term in the model. Do not add a column of 1s directly to `X`. You can change the default behavior of `glmfit` by specifying the `'Constant'` name-value argument.

Data Types: `single` | `double`

`y` — Response variable
vector | matrix

Response variable, specified as a vector or matrix.

- If `distr` is not `'binomial'`, then `y` must be an *n*-by-1 vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`. The data type must be single or double.
- If `distr` is `'binomial'`, then `y` is an *n*-by-1 vector indicating success or failure at each observation, or an *n*-by-2 matrix whose first column indicates the number of successes for each observation and second column indicates the number of trials for each observation.

Data Types: `single` | `double` | `logical` | `categorical`

`distr` — Distribution of response variable
`'normal'` (default) | `'binomial'` | `'poisson'` | `'gamma'` | `'inverse gaussian'`

Distribution of the response variable, specified as one of the values in this table.

Value	Description
`'normal'`	Normal distribution (default)
`'binomial'`	Binomial distribution
`'poisson'`	Poisson distribution
`'gamma'`	Gamma distribution
`'inverse gaussian'`	Inverse Gaussian distribution

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `b = glmfit(X,y,'normal','link','probit')` specifies that the distribution of the response is normal and instructs `glmfit` to use the probit link function.

`B0` — Initial values for coefficient estimates
numeric vector

Initial values for the coefficient estimates, specified as a numeric vector. The default values are initial fitted values derived from the input data.

Data Types: `single` | `double`

`Constant` — Indicator for constant term
`'on'` (default) | `'off'`

Indicator for the constant term (intercept) in the fit, specified as either `'on'` to include the constant term or `'off'` to remove it from the model.

- `'on'` (default) — `glmfit` includes a constant term in the model and returns a (p + 1)-by-1 vector of coefficient estimates `b`, where p is the number of predictors in `X`. The coefficient of the constant term is the first element of `b`.
- `'off'` — `glmfit` omits the constant term and returns a p-by-1 vector of coefficient estimates `b`.

Example: `'Constant','off'`

`EstDisp` — Indicator to compute dispersion parameter
`'off'` for `'binomial'` and `'poisson'` distributions (default) | `'on'`

Indicator to compute a dispersion parameter for `'binomial'` and `'poisson'` distributions, specified as `'on'` or `'off'`.

Value	Description
`'on'`	Estimate a dispersion parameter when computing standard errors. The estimated dispersion parameter value is the sum of squared Pearson residuals divided by the degrees of freedom for error (DFE).
`'off'`	Use the theoretical value of 1 when computing standard errors (default).

The fitting function always estimates the dispersion for other distributions.

Example: `'EstDisp','on'`

`Link` — Link function
canonical link function (default) | scalar value | structure or cell array of custom link function

Link function to use in place of the canonical link function, specified as one of the built-in link functions in the following table or a custom link function.

Link Function Name	Link Function	Mean (Inverse) Function
`'identity'` (default for `'normal'` distribution)	$f(\mu) = \mu$	$\mu = Xb$
`'log'` (default for `'poisson'` distribution)	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
`'logit'` (default for `'binomial'` distribution)	$f(\mu) = \log(\mu/(1 - \mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'probit'`	$f(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function of the standard normal distribution	$\mu = \Phi(Xb)$
`'loglog'`	$f(\mu) = \log(-\log(\mu))$	$\mu = \exp(-\exp(Xb))$
`'comploglog'`	$f(\mu) = \log(-\log(1 - \mu))$	$\mu = 1 - \exp(-\exp(Xb))$
`'reciprocal'` (default for `'gamma'` distribution)	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$

Link Function Name	Link Function	Mean (Inverse) Function
p (a number, default for the 'inverse gaussian' distribution with $p = -2$)	$f(\mu) = \mu^p$	$\mu = Xb^{1/p}$

The default `'Link'` value is the canonical link function, which depends on the distribution of the response variable specified by the `distr` argument.

You can specify a custom link function using a structure or cell array.

- Structure with three fields. Each field of the structure (for example, `S`) holds a function handle that accepts a vector of inputs and returns a vector of the same size:
  - `S.Link` — Link function, $f(\mu)$ = `S.Link`(μ)
  - `S.Derivative` — Derivative of the link function
  - `S.Inverse` — Inverse link function, μ = `S.Inverse`(Xb)
- Cell array of the form `{FL FD FI}` that defines the link function (`FL(mu)`), the derivative of the link function (`FD = dFL(mu)/dmu`), and the inverse link function (`FI = FL^(−1)`). Each entry holds a function handle that accepts a vector of inputs and returns a vector of the same size.

The link function defines the relationship $f(\mu) = X*b$ between the mean response μ and the linear combination of predictors $X*b$.

Example: `'Link','probit'`

Data Types: `single` | `double` | `char` | `string` | `struct` | `cell`

`Offset` — Offset variable

[ ] (default) | numeric vector

Offset variable in the fit, specified as a numeric vector with the same length as the response `y`.

`glmfit` uses `Offset` as an additional predictor with a coefficient value fixed at 1. In other words, the formula for fitting is

$f(\mu)$ = `Offset` + $X*b$,

where f is the link function, μ is the mean response, and $X*b$ is the linear combination of predictors X. The `Offset` predictor has coefficient `1`.

For example, consider a Poisson regression model. Suppose, for theoretical reasons, the number of counts is to be proportional to a predictor `A`. By using the log link function and specifying `log(A)` as an offset, you can force the model to satisfy this theoretical constraint.

Data Types: `single` | `double`

`Options` — Optimization options

`statset('glmfit')` (default) | structure

Optimization options, specified as a structure. This argument determines the control parameters for the iterative algorithm that `glmfit` uses.

Create the `'Options'` value by using the function `statset` or by creating a structure array containing the fields and values described in this table.

Field Name	Value	Default Value
`Display`	Amount of information displayed by the algorithm • `'off'` — Displays no information • `'final'` — Displays the final output	`'off'`
`MaxIter`	Maximum number of iterations allowed, specified as a positive integer	`100`
`TolX`	Termination tolerance for the parameters, specified as a positive scalar	`1e-6`

You can also enter `statset('glmfit')` in the Command Window to see the names and default values of the fields that `glmfit` accepts in the `'Options'` name-value argument.

Example: `'Options',statset('Display','final','MaxIter',1000)` specifies to display the final information of the iterative algorithm results, and change the maximum number of iterations allowed to 1000.

Data Types: `struct`

`Weights` — Observation weights
`ones(n,1)` (default) | *n*-by-1 vector of nonnegative scalar values

Observation weights, specified as an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations.

Data Types: `single` | `double`

Output Arguments

`b` — Coefficient estimates
numeric vector

Coefficient estimates, returned as a numeric vector.

- If `'Constant'` is `'on'` (default), then `glmfit` includes a constant term in the model and returns a (p + 1)-by-1 vector of coefficient estimates `b`, where p is the number of predictors in `X`. The coefficient of the constant term is the first element of `b`.
- If `'Constant'` is `'off'`, then `glmfit` omits the constant term and returns a p-by-1 vector of coefficient estimates `b`.

`dev` — Deviance of fit
numeric value

Deviance of the fit, returned as a numeric value. The deviance is useful for comparing two models when one model is a special case of the other model. The difference between the deviance of the two models has a chi-square distribution with degrees of freedom equal to the difference in the number of estimated parameters between the two models.

For more information, see “Deviance” on page 33-2701.

`stats` — Model statistics
structure

Model statistics, returned as a structure with the following fields:

- `beta` — Coefficient estimates `b`
- `dfe` — Degrees of freedom for error
- `sfit` — Estimated dispersion parameter
- `s` — Theoretical or estimated dispersion parameter
- `estdisp` — 0 when `'EstDisp'` is `'off'` and 1 when `'EstDisp'` is `'on'`
- `covb` — Estimated covariance matrix for `b`
- `se` — Vector of standard errors of the coefficient estimates `b`
- `coeffcorr` — Correlation matrix for `b`
- `t` — *t* statistics for `b`
- `p` — *p*-values for `b`
- `resid` — Vector of residuals
- `residp` — Vector of Pearson residuals
- `residd` — Vector of deviance residuals
- `resida` — Vector of Anscombe residuals

If you estimate a dispersion parameter for the binomial or Poisson distribution, then `stats.s` is equal to `stats.sfit`. Also, the elements of `stats.se` differ by the factor `stats.s` from their theoretical values.

More About

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to a saturated model.

Deviance of a model M_1 is twice the difference between the loglikelihood of the model M_1 and the saturated model M_s. A saturated model is a model with the maximum number of parameters that you can estimate.

For example, if you have n observations (y_i, $i = 1, 2, \ldots, n$) with potentially different values for $X_i^T\beta$, then you can define a saturated model with n parameters. Let $L(b,y)$ denote the maximum value of the likelihood function for a model with the parameters b. Then the deviance of the model M_1 is

$$-2(\log L(b_1, y) - \log L(b_S, y)),$$

where b_1 and b_s contain the estimated parameters for the model M_1 and the saturated model, respectively. The deviance has a chi-square distribution with n – p degrees of freedom, where n is the number of parameters in the saturated model and p is the number of parameters in the model M_1.

Assume you have two different generalized linear regression models M_1 and M_2, and M_1 has a subset of the terms in M_2. You can assess the fit of the models by comparing the deviances D_1 and D_2 of the two models. The difference of the deviances is

$$\begin{aligned} D = D_2 - D_1 &= -2(\log L(b_2, y) - \log L(b_S, y)) + 2(\log L(b_1, y) - \log L(b_S, y)) \\ &= -2(\log L(b_2, y) - \log L(b_1, y)). \end{aligned}$$

Asymptotically, the difference D has a chi-square distribution with degrees of freedom v equal to the difference in the number of parameters estimated in M_1 and M_2. You can obtain the p-value for this test by using `1 - chi2cdf(D,v)`.

Typically, you examine D using a model M_2 with a constant term and no predictors. Therefore, D has a chi-square distribution with p – 1 degrees of freedom. If the dispersion is estimated, the difference divided by the estimated dispersion has an F distribution with p – 1 numerator degrees of freedom and n – p denominator degrees of freedom.

Tips

- `glmfit` treats `NaN`s in `X` or `y` as missing values and ignores them.

Alternative Functionality

`glmfit` is useful when you simply need the output arguments of the function or when you want to repeat fitting a model multiple times in a loop. If you need to investigate a fitted model further, create a generalized linear regression model object `GeneralizedLinearModel` by using `fitglm` or `stepwiseglm`. A `GeneralizedLinearModel` object provides more features than `glmfit`.

- Use the properties of `GeneralizedLinearModel` to investigate a fitted model. The object properties include information about the coefficient estimates, summary statistics, fitting method, and input data.
- Use the object functions of `GeneralizedLinearModel` to predict responses and to modify, evaluate, and visualize the generalized linear regression model.
- You can find the information in the output of `glmfit` using the properties and object functions of `GeneralizedLinearModel`.

Output of `glmfit`	**Equivalent Values in `GeneralizedLinearModel`**
`b`	See the `Estimate` column of the `Coefficients` property.
`dev`	See the `Deviance` property.
`stats`	See the model display in the Command Window. You can find the statistics in the model properties (`CoefficientCovariance`, `Coefficients`, `Dispersion`, `DispersionEstimated`, and `Residuals`). The dispersion parameter in `stats.s` of `glmfit` is the scale factor for the standard errors of coefficients, whereas the dispersion parameter in the `Dispersion` property of a generalized linear model is the scale factor for the variance of the response. Therefore, `stats.s` is the square root of the `Dispersion` value.

References

[1] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[3] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `fitglm` | `glmval` | `regress` | `regstats` | `stepwiseglm`

Topics

“Fitting Data with Generalized Linear Models” on page 12-65
“Generalized Linear Models” on page 12-9

Introduced before R2006a

glmval

Generalized linear model values

Syntax

```
yhat = glmval(b,X,link)
[yhat,dylo,dyhi] = glmval(b,X,link,stats)
[...] = glmval(...,param1,val1,param2,val2,...)
```

Description

`yhat = glmval(b,X,link)` computes predicted values for the generalized linear model with link function `link` and predictors `X`. Distinct predictor variables should appear in different columns of `X`. `b` is a vector of coefficient estimates as returned by the `glmfit` function. `link` can be any of the character vectors, string scalars, or custom-defined link functions used as values for the `'link'` name-value pair argument in the `glmfit` function.

Note By default, `glmval` adds a first column of 1s to `X`, corresponding to a constant term in the model. Do not enter a column of 1s directly into `X`. You can change the default behavior of `glmval` using the `'constant'` parameter.

`[yhat,dylo,dyhi] = glmval(b,X,link,stats)` also computes 95% confidence bounds for the predicted values. When the `stats` structure output of the `glmfit` function is specified, `dylo` and `dyhi` are also returned. `dylo` and `dyhi` define a lower confidence bound of `yhat-dylo`, and an upper confidence bound of `yhat+dyhi`. Confidence bounds are nonsimultaneous, and apply to the fitted curve, not to a new observation.

`[...] = glmval(...,param1,val1,param2,val2,...)` specifies optional parameter name/value pairs to control the predicted values. Acceptable parameters are listed in this table:

Parameter	Value
`'confidence'` — the confidence level for the confidence bounds	A scalar between 0 and 1
`'size'` — the size parameter (N) for a binomial model	A scalar, or a vector with one value for each row of X
`'offset'` — used as an additional predictor variable, but with a coefficient value fixed at 1.0	A vector
`'constant'`	• `'on'` — Includes a constant term in the model. The coefficient of the constant term is the first element of `b`. • `'off'` — Omit the constant term
`'simultaneous'` — Compute simultaneous confidence intervals (`true`), or compute non-simultaneous confidence intervals (default `false`)	`true` or `false`

Examples

Fit Generalized Linear Model with Probit Link

Fit a generalized linear regression model, and compute predicted (estimated) values for the predictor data using the fitted model.

Create a sample data set.

```
x = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
n = [48 42 31 34 31 21 23 23 21 16 17 21]';
y = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

`x` contains the predictor variable values. Each `y` value is the number of successes in the corresponding number of trials in `n`.

Fit a probit regression model for `y` on `x`.

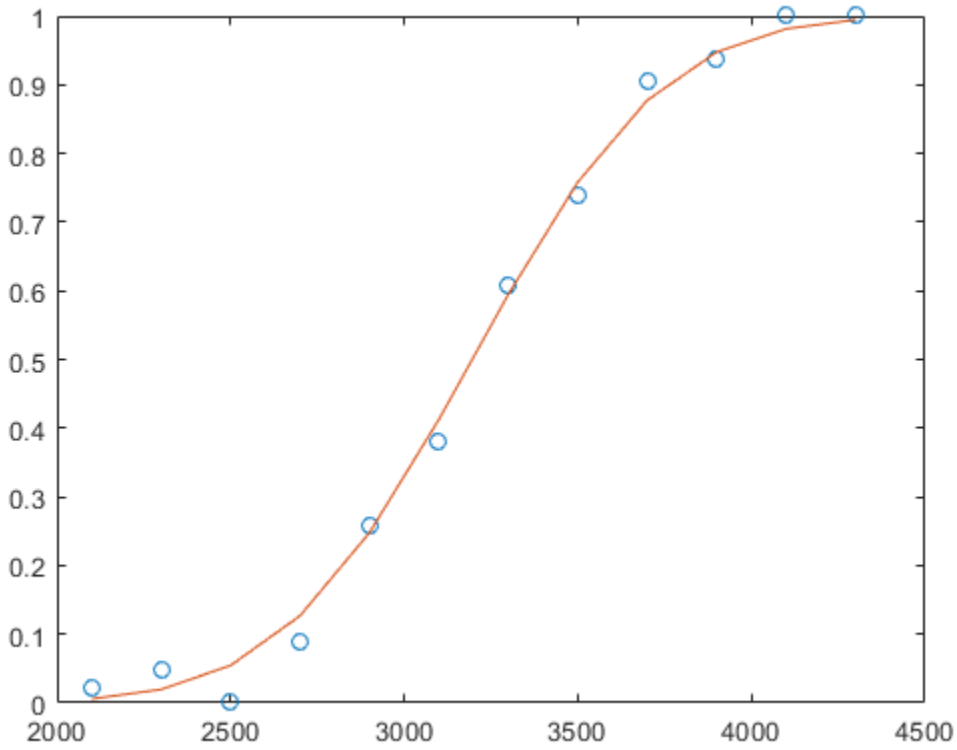
```
b = glmfit(x,[y n],'binomial','Link','probit');
```

Compute the estimated number of successes.

```
yfit = glmval(b,x,'probit','Size',n);
```

Plot the observed success percent and estimated success percent versus the `x` values.

```
plot(x,y./n,'o',x,yfit./n,'-')
```



Use Custom-Defined Link Function

Enter sample data.

```
x = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
n = [48 42 31 34 31 21 23 23 21 16 17 21]';
y = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

Each `y` value is the number of successes in corresponding number of trials in `n`, and `x` contains the predictor variable values.

Define three function handles, created by using `@`, that define the link, the derivative of the link, and the inverse link for a probit link function,. Store the handles in a cell array.

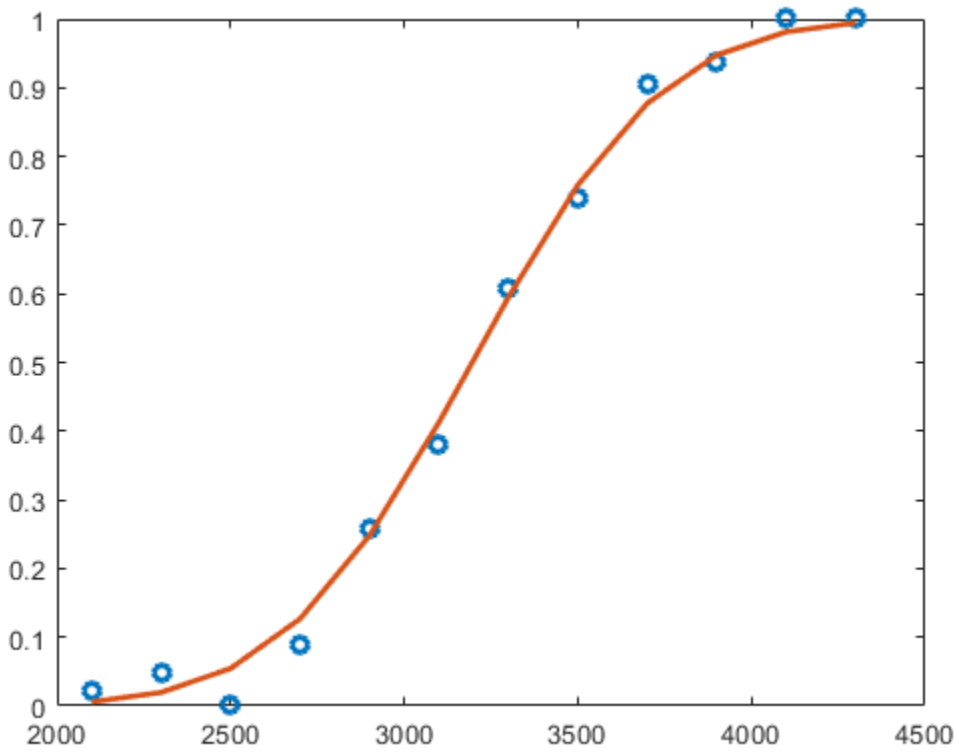
```
link = @(mu) norminv(mu);
derlink = @(mu) 1 ./ normpdf(norminv(mu));
invlink = @(resp) normcdf(resp);
F = {link, derlink, invlink};
```

Fit a generalized linear model for `y` on `x` by using the link function that you defined.

```
b = glmfit(x,[y n],'binomial','link',F);
```

Compute the estimated number of successes. Plot the observed and estimated percent success versus the `x` values.

```
yfit = glmval(b,x,F,'size',n);
plot(x, y./n,'o',x,yfit./n,'-','LineWidth',2)
```



Generate Code from Function That Predicts Responses Given New Data

Train a generalized linear model, and then generate code from a function that classifies new observations based on the model. This example is based on the “Use Custom-Defined Link Function” on page 33-2706 example.

Enter the sample data.

```
x = [2100 2300 2500 2700 2900 3100 ...
     3300 3500 3700 3900 4100 4300]';
n = [48 42 31 34 31 21 23 23 21 16 17 21]';
y = [1 2 0 3 8 8 14 17 19 15 17 21]';
```

Suppose that the inverse normal pdf is an appropriate link function for the problem.

Define a function named `myInvNorm.m` that accepts values of $X\beta$ and returns corresponding values of the inverse of the standard normal cdf.

```
function in = myInvNorm(mu) %#codegen
%myInvNorm Inverse of standard normal cdf for code generation
%   myInvNorm is a GLM link function that accepts a numeric vector mu, and
%   returns in, which is a numeric vector of corresponding values of the
%   inverse of the standard normal cdf.
```

```
%
in = norminv(mu);
end
```

Define another function named `myDInvNorm.m` that accepts values of $X\beta$ and returns corresponding values of the derivative of the link function.

```
function din = myDInvNorm(mu) %#codegen
%myDInvNorm Derivative of inverse of standard normal cdf for code
%generation
%   myDInvNorm corresponds to the derivative of the GLM link function
%   myInvNorm. myDInvNorm accepts a numeric vector mu, and returns din,
%   which is a numeric vector of corresponding derivatives of the inverse
%   of the standard normal cdf.
%
din = 1./normpdf(norminv(mu));
end
```

Define another function named `myInvInvNorm.m` that accepts values of $X\beta$ and returns corresponding values of the inverse of the link function.

```
function iin = myInvInvNorm(mu) %#codegen
%myInvInvNorm Standard normal cdf for code generation
%   myInvInvNorm is the inverse of the GLM link function myInvNorm.
%   myInvInvNorm accepts a numeric vector mu, and returns iin, which is a
%   numeric vector of corresponding values of the standard normal cdf.
%
iin = normcdf(mu);
end
```

Create a structure array that specifies each of the link functions. Specifically, the structure array contains fields named `'Link'`, `'Derivative'`, and `'Inverse'`. The corresponding values are the names of the functions.

```
link = struct('Link','myInvNorm','Derivative','myDInvNorm',...
    'Inverse','myInvInvNorm')
```

```
link =

  struct with fields:

         Link: 'myInvNorm'
   Derivative: 'myDInvNorm'
      Inverse: 'myInvInvNorm'
```

Fit a GLM for `y` on `x` using the link function `link`. Return the structure array of statistics.

```
[b,~,stats] = glmfit(x,[y n],'binomial','link',link);
```

`b` is a 2-by-1 vector of regression coefficients.

In your current working folder, define a function called `classifyGLM.m` that:

- Accepts measurements with columns corresponding to those in `x`, regression coefficients whose dimensions correspond to `b`, a link function, the structure of GLM statistics, and any valid `glmval` name-value pair argument
- Returns predictions and confidence interval margins of error

```
function [yhat,lo,hi] = classifyGLM(b,x,link,varargin) %#codegen
%CLASSIFYGLM Classify measurements using GLM model
%   CLASSIFYGLM classifies the n observations in the n-by-1 vector x using
%   the GLM model with regression coefficients b and link function link,
%   and then returns the n-by-1 vector of predicted values in yhat.
%   CLASSIFYGLM also returns margins of error for the predictions using
%   additional information in the GLM statistics structure stats.
narginchk(3,Inf);
if(isstruct(varargin{1}))
    stats = varargin{1};
    [yhat,lo,hi] = glmval(b,x,link,stats,varargin{2:end});
else
    yhat = glmval(b,x,link,varargin{:});
end
end
```

Generate a MEX function from `classifyGLM.m`. Because C uses static typing, `codegen` must determine the properties of all variables in MATLAB® files at compile time. To ensure that the MEX function can use the same inputs, use the `-args` argument to specify the following in the order given:

- Regression coefficients `b` as a compile-time constant
- In-sample observations `x`
- Link function as a compile-time constant
- Resulting GLM statistics as a compile-time constant
- Name `'Confidence'` as a compile-time constant
- Confidence level 0.9

To designate arguments as compile-time constants, use `coder.Constant`.

```
codegen -config:mex classifyGLM -args {coder.Constant(b),x,coder.Constant(link),coder.Constant(s
```

```
Code generation successful.
```

`codegen` generates the MEX file `classifyGLM_mex.mexw64` in your current folder. The file extension depends on your system platform.

Compare predictions by using `glmval` and `classifyGLM_mex`. Specify name-value pair arguments in the same order as in the `-args` argument in the call to `codegen`.

```
[yhat1,melo1,mehi1] = glmval(b,x,link,stats,'Confidence',0.9);
[yhat2,melo2,mehi2] = classifyGLM_mex(b,x,link,stats,'Confidence',0.9);

comp1 = (yhat1 - yhat2)'*(yhat1 - yhat2);
```

```
agree1 = comp1 < eps
comp2 = (melo1 - melo2)'*(melo1 - melo2);
agree2 = comp2 < eps
comp3 = (mehi1 - mehi2)'*(mehi1 - mehi2);
agree3 = comp3 < eps
```

```
agree1 =

  logical

   1


agree2 =

  logical

   1


agree3 =

  logical

   1
```

The generated MEX function produces the same predictions as `predict`.

References

[1] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[3] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

Argument	Notes and Limitations
`b`	`b` must be a compile-time constant (the value of `b` cannot change while `codegen` generates the C/C++ code). You can designate variables as compile-time constants using `coder.Constant`.
`X`	Must be a double- or single-precision numeric matrix
`link`	• Does not support function handles, that is, functions created using `@`. • Must be a compile-time constant.

Argument	Notes and Limitations
`stats`	Must be a compile-time constant
`'Constant'` name-value pair argument	Must be a compile-time constant
Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to specify a confidence level of 0.9, include `{coder.Constant('Confidence'),coder.Constant(0.9)}` in the `-args` value of `codegen`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `fitglm` | `glmfit` | `stepwiseglm`

Topics

“Introduction to Code Generation” on page 32-2

Introduced before R2006a

glyphplot

Glyph plot

Syntax

```
glyphplot(X)
glyphplot(X,'glyph','face')
glyphplot(X,'glyph','face','features',f)
glyphplot(X,...,'grid',[rows,cols])
glyphplot(X,...,'grid',[rows,cols],'page',p)
glyphplot(X,...,'centers',C)
glyphplot(X,...,'centers',C,'radius',r)
glyphplot(X,...,'obslabels',labels)
glyphplot(X,...,'standardize',method)
glyphplot(X,...,prop1,val1,...)
h = glyphplot(X,...)
```

Description

`glyphplot(X)` creates a star plot from the multivariate data in the *n*-by-*p* matrix `X`. Rows of `X` correspond to observations, columns to variables. A star plot represents each observation as a "star" whose *i*th spoke is proportional in length to the *i*th coordinate of that observation. `glyphplot` standardizes `X` by shifting and scaling each column separately onto the interval `[0,1]` before making the plot, and centers the glyphs on a rectangular grid that is as close to square as possible. `glyphplot` treats `NaN`s in `X` as missing values, and does not plot the corresponding rows of `X`. `glyphplot(X,'glyph','star')` is a synonym for `glyphplot(X)`.

`glyphplot(X,'glyph','face')` creates a face plot from `X`. A face plot represents each observation as a "face," whose *i*th facial feature is drawn with a characteristic proportional to the *i*th coordinate of that observation. The features are described in "Face Features" on page 33-2713.

`glyphplot(X,'glyph','face','features',f)` creates a face plot where the *i*th element of the index vector `f` defines which facial feature will represent the *i*th column of `X`. `f` must contain integers from `0` to `17`, where `0` indicate that the corresponding column of `X` should not be plotted. See "Face Features" on page 33-2713 for more information.

`glyphplot(X,...,'grid',[rows,cols])` organizes the glyphs into a `rows`-by-`cols` grid.

`glyphplot(X,...,'grid',[rows,cols],'page',p)` organizes the glyph into one or more pages of a `rows`-by-`cols` grid, and displays the page `p`. If `p` is a vector, `glyphplot` displays multiple pages in succession. If `p` is `'all'`, `glyphplot` displays all pages. If `p` is `'scroll'`, `glyphplot` displays a single plot with a scrollbar.

`glyphplot(X,...,'centers',C)` creates a plot with each glyph centered at the locations in the *n*-by-2 matrix `C`.

`glyphplot(X,...,'centers',C,'radius',r)` creates a plot with glyphs positioned using `C`, and scale the glyphs so the largest has radius `r`.

`glyphplot(X,...,'obslabels',labels)` labels each glyph with the text in `labels`. By default, the glyphs are labelled `1:N`. Use `''` for blank labels.

`glyphplot(X,...,'standardize',method)` standardizes `X` before making the plot. Choices for *`method`* are

- `'column'` — Maps each column of `X` separately onto the interval [0,1]. This is the default.
- `'matrix'` — Maps the entire matrix `X` onto the interval `[0,1]`.
- `'PCA'` — Transforms `X` to its principal component scores, in order of decreasing eigenvalue, and maps each one onto the interval `[0,1]`.
- `'off'` — No standardization. Negative values in `X` may make a star plot uninterpretable.

`glyphplot(X,...,prop1,val1,...)` sets properties to the specified property values for all line graphics objects created by `glyphplot`.

`h = glyphplot(X,...)` returns a matrix of handles to the graphics objects created by `glyphplot`. For a star plot, `h(:,1)` and `h(:,2)` contain handles to the line objects for each star's perimeter and spokes, respectively. For a face plot, `h(:,1)` and `h(:,2)` contain object handles to the lines making up each face and to the pupils, respectively. `h(:,3)` contains handles to the text objects for the labels, if present.

Face Features

The following table describes the correspondence between the columns of the vector `f`, the value of the `'Features'` input parameter, and the facial features of the glyph plot. If `X` has fewer than 17 columns, unused features are displayed at their default value.

Column	Facial Feature
1	Size of face
2	Forehead/jaw relative arc length
3	Shape of forehead
4	Shape of jaw
5	Width between eyes
6	Vertical position of eyes
7	Height of eyes
8	Width of eyes (this also affects eyebrow width)
9	Angle of eyes (this also affects eyebrow angle)
10	Vertical position of eyebrows
11	Width of eyebrows (relative to eyes)
12	Angle of eyebrows (relative to eyes)
13	Direction of pupils
14	Length of nose
15	Vertical position of mouth
16	Shape of mouth
17	Mouth arc length

Examples

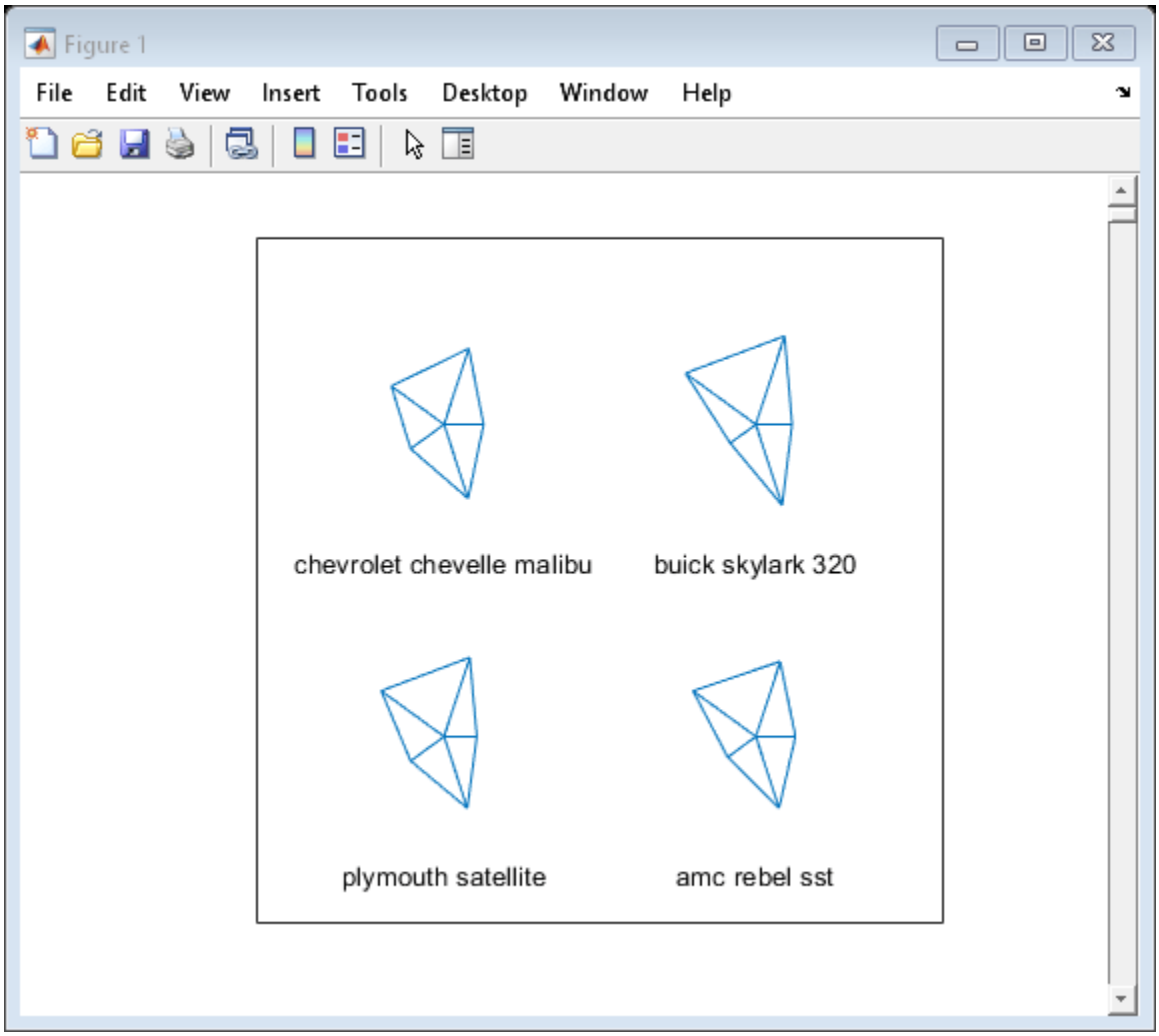
Star and Face Plots of Multivariate Data

Load the sample data.

```
load carsmall
X = [Acceleration Displacement Horsepower MPG Weight];
```

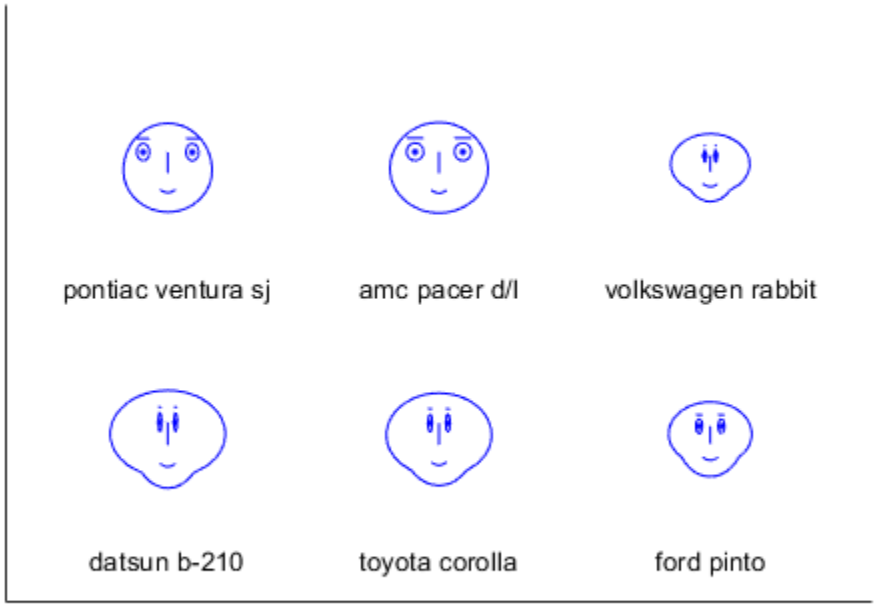
Create a star plot of the data in X . Standardize the data before plotting.

```
glyphplot(X,'standardize','column','obslabels',Model,'grid',[2 2],...
            'page','scroll');
```



Create a face plot of the data in X .

```
glyphplot(X,'glyph','face','obslabels',Model,'grid',[2 3],'page',9);
```



See Also

`andrewsplot` | `parallelcoords`

Introduced before R2006a

gmdistribution

Create Gaussian mixture model

Description

A `gmdistribution` object stores a Gaussian mixture distribution, also called a Gaussian mixture model (GMM), which is a multivariate distribution that consists of multivariate Gaussian distribution components. Each component is defined by its mean and covariance. The mixture is defined by a vector of mixing proportions, where each mixing proportion represents the fraction of the population described by a corresponding component.

Creation

You can create a `gmdistribution` model object in two ways.

- Use the `gmdistribution` function (described here) to create a `gmdistribution` model object by specifying the distribution parameters.
- Use the `fitgmdist` function to fit a `gmdistribution` model object to data given a fixed number of components.

Syntax

```
gm = gmdistribution(mu,sigma)
gm = gmdistribution(mu,sigma,p)
```

Description

`gm = gmdistribution(mu,sigma)` creates a `gmdistribution` model object using the specified means `mu` and covariances `sigma` with equal mixing proportions.

`gm = gmdistribution(mu,sigma,p)` specifies the mixing proportions of multivariate Gaussian distribution components.

Input Arguments

`mu` — Means
k-by-*m* numeric matrix

Means of multivariate Gaussian distribution components, specified as a *k*-by-*m* numeric matrix, where *k* is the number of components and *m* is the number of variables in each component. `mu(i,:)` is the mean of component `i`.

Data Types: `single` | `double`

`sigma` — Covariances
numeric vector | numeric matrix | numeric array

Covariances of multivariate Gaussian distribution components, specified as a numeric vector, matrix, or array.

Given that *k* is the number of components and *m* is the number of variables in each component, `sigma` is one of the values in this table.

Value	Description
m-by-*m*-by-*k* array	`sigma(:,:,i)` is the covariance matrix of component `i`.
1-by-*m*-by-*k* array	Covariance matrices are diagonal. `sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.
m-by-*m* matrix	Covariance matrices are the same across components.
1-by-*m* vector	Covariance matrices are diagonal and the same across components.

Data Types: `single` | `double`

`p` — Mixing proportions of mixture components
numeric vector of length *k*

Mixing proportions of mixture components, specified as a numeric vector of length *k*, where *k* is the number of components. The default is a row vector of (1/*k*)s, which sets equal proportions. If `p` does not sum to `1`, `gmdistribution` normalizes it.

Data Types: `single` | `double`

Properties

Distribution Parameters

`mu` — Means
k-by-*m* numeric matrix

This property is read-only.

Means of multivariate Gaussian distribution components, specified as a *k*-by-*m* numeric matrix, where *k* is the number of components and *m* is the number of variables in each component. `mu(i,:)` is the mean of component `i`.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the `mu` input argument of `gmdistribution` sets this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then `fitgmdist` estimates this property.

Data Types: `single` | `double`

`Sigma` — Covariances
numeric vector | numeric matrix | numeric array

This property is read-only.

Covariances of multivariate Gaussian distribution components, specified as a numeric vector, matrix, or array.

Given that *k* is the number of components and *m* is the number of variables in each component, `Sigma` is one of the values in this table.

Value	Description
m-by-*m*-by-*k* array	`Sigma(:,:,i)` is the covariance matrix of component `i`.
1-by-*m*-by-*k* array	Covariance matrices are diagonal. `Sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.
m-by-*m* matrix	Covariance matrices are the same across components.
1-by-*m* vector	Covariance matrices are diagonal and the same across components.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the `sigma` input argument of `gmdistribution` sets this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then `fitgmdist` estimates this property.

Data Types: `single` | `double`

`ComponentProportion` — Mixing proportions of mixture components
1-by-*k* numeric vector

This property is read-only.

Mixing proportions of mixture components, specified as a 1-by-*k* numeric vector.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the `p` input argument of `gmdistribution` sets this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then `fitgmdist` estimates this property.

Data Types: `single` | `double`

Distribution Characteristics

`CovarianceType` — Type of covariance matrices
`'diagonal'` | `'full'`

This property is read-only.

Type of covariance matrices, specified as either `'diagonal'` or `'full'`.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the type of covariance matrices in the `sigma` input argument of `gmdistribution` sets this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then the `'CovarianceType'` name-value pair argument of `fitgmdist` sets this property.

`DistributionName` — Distribution name
`'gaussian mixture distribution'` (default)

This property is read-only.

Distribution name, specified as `'gaussian mixture distribution'`.

`NumComponents` — Number of mixture components
positive integer

This property is read-only.

Number of mixture components, k, specified as a positive integer.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the input arguments `mu`, `sigma`, and `p` of `gmdistribution` set this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then the `k` input argument of `fitgmdist` sets this property.

Data Types: `single` | `double`

`NumVariables` — Number of variables
positive integer

This property is read-only.

Number of variables in the multivariate Gaussian distribution components, m, specified as a positive integer.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the input arguments `mu`, `sigma`, and `p` of `gmdistribution` set this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then the input data `X` of `fitgmdist` sets this property.

Data Types: `double`

`SharedCovariance` — Flag indicating shared covariance
`true` | `false`

This property is read-only.

Flag indicating whether a covariance matrix is shared across mixture components, specified as `true` or `false`.

- If you create a `gmdistribution` object by using the `gmdistribution` function, then the type of covariance matrices in the `sigma` input argument of `gmdistribution` sets this property.
- If you fit a `gmdistribution` object to data by using the `fitgmdist` function, then the `'SharedCovariance'` name-value pair argument of `fitgmdist` sets this property.

Data Types: `logical`

Properties for Fitted Object

The following properties apply only to a fitted object you create by using `fitgmdist`. The values of these properties are empty if you create a `gmdistribution` object by using the `gmdistribution` function.

`AIC` — Akaike Information Criterion
scalar

This property is read-only.

Akaike information criterion (AIC), specified as a scalar. `AIC = 2*NlogL + 2*p`, where `NlogL` is the negative loglikelihood (the `NegativeLogLikelihood` property) and `p` is the number of estimated parameters.

AIC is a model selection tool you can use to compare multiple models fit to the same data. AIC is a likelihood-based measure of model fit that includes a penalty for complexity, specifically, the number of parameters. When you compare multiple models, a model with a smaller value of AIC is better.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `single` | `double`

`BIC` — Bayes Information Criterion
scalar

This property is read-only.

Bayes information criterion (BIC), specified as a scalar. `BIC = 2*NlogL + p*log(n)`, where `NlogL` is the negative loglikelihood (the `NegativeLogLikelihood` property), `n` is the number of observations, and `p` is the number of estimated parameters.

BIC is a model selection tool you can use to compare multiple models fit to the same data. BIC is a likelihood-based measure of model fit that includes a penalty for complexity, specifically, the number of parameters. When you compare multiple models, a model with the lowest BIC value is the best fitting model.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `single` | `double`

`Converged` — Flag indicating convergence
`true` | `false`

This property is read-only.

Flag indicating whether the Expectation-Maximization (EM) algorithm is converged when fitting a Gaussian mixture model, specified as `true` or `false`.

You can change the optimization options by using the `'Options'` name-value pair argument of `fitgmdist`.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `logical`

`NegativeLogLikelihood` — Negative loglikelihood
scalar

This property is read-only.

Negative loglikelihood of the fitted Gaussian mixture model given the input data `X` of `fitgmdist`, specified as a scalar.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `single` | `double`

`NumIterations` — Number of iterations
positive integer

This property is read-only.

Number of iterations in the Expectation-Maximization (EM) algorithm, specified as a positive integer.

You can change the optimization options, including the maximum number of iterations allowed, by using the `'Options'` name-value pair argument of `fitgmdist`.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `double`

`ProbabilityTolerance` — Tolerance for posterior probabilities
nonnegative scalar value in range `[0,1e-6]`

This property is read-only.

Tolerance for posterior probabilities, specified as a nonnegative scalar value in the range `[0,1e-6]`.

The `'ProbabilityTolerance'` name-value pair argument of `fitgmdist` sets this property.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `single` | `double`

`RegularizationValue` — Regularization parameter value
nonnegative scalar

This property is read-only.

Regularization parameter value, specified as a nonnegative scalar.

The `'RegularizationValue'` name-value pair argument of `fitgmdist` sets this property.

This property is empty if you create a `gmdistribution` object by using the `gmdistribution` function.

Data Types: `single` | `double`

Object Functions

cdf	Cumulative distribution function for Gaussian mixture distribution
cluster	Construct clusters from Gaussian mixture distribution
mahal	Mahalanobis distance to Gaussian mixture component
pdf	Probability density function for Gaussian mixture distribution
posterior	Posterior probability of Gaussian mixture component
random	Random variate from Gaussian mixture distribution

Examples

Create Gaussian Mixture Distribution Using gmdistribution

Create a two-component bivariate Gaussian mixture distribution by using the `gmdistribution` function.

Define the distribution parameters (means and covariances) of two bivariate Gaussian mixture components.

```
mu = [1 2;-3 -5];
sigma = cat(3,[2 .5],[1 1]) % 1-by-2-by-2 array
```

```
sigma =
sigma(:,:,1) =

    2.0000    0.5000


sigma(:,:,2) =

     1     1
```

The `cat` function concatenates the covariances along the third array dimension. The defined covariance matrices are diagonal matrices. `sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.

Create a `gmdistribution` object. By default, the `gmdistribution` function creates an equal proportion mixture.

```
gm = gmdistribution(mu,sigma)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:     1     2

Component 2:
Mixing proportion: 0.500000
Mean:    -3    -5
```

List the properties of the `gm` object.

```
properties(gm)
```

```
Properties for class gmdistribution:

    NumVariables
    DistributionName
    NumComponents
    ComponentProportion
    SharedCovariance
    NumIterations
    RegularizationValue
    NegativeLogLikelihood
    CovarianceType
    mu
```

```
    Sigma
    AIC
    BIC
    Converged
    ProbabilityTolerance
```

You can access these properties by using dot notation. For example, access the `ComponentProportion` property, which represents the mixing proportions of mixture components.

```
gm.ComponentProportion
```

```
ans = 1×2

    0.5000    0.5000
```

A `gmdistribution` object has properties that apply only to a fitted object. The fitted object properties are `AIC`, `BIC`, `Converged`, `NegativeLogLikelihood`, `NumIterations`, `ProbabilityTolerance`, and `RegularizationValue`. The values of the fitted object properties are empty if you create an object by using the `gmdistribution` function and specifying distribution parameters. For example, access the `NegativeLogLikelihood` property by using dot notation.

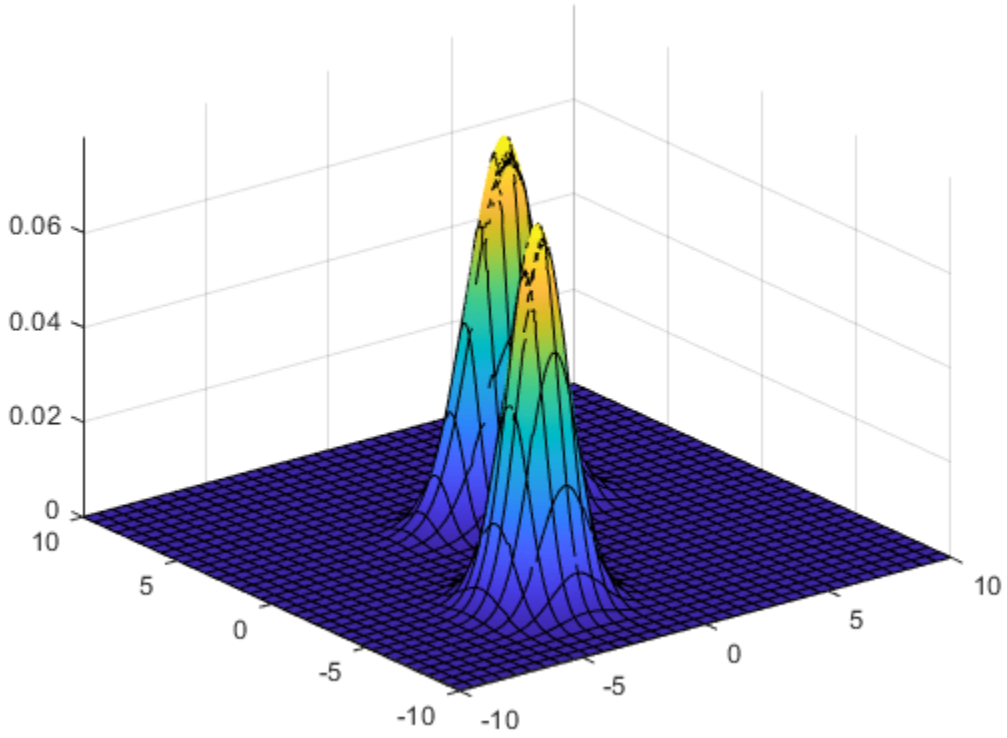
```
gm.NegativeLogLikelihood
```

```
ans =

     []
```

After you create a `gmdistribution` object, you can use the object functions. Use `cdf` and `pdf` to compute the values of the cumulative distribution function (cdf) and the probability density function (pdf). Use `random` to generate random vectors. Use `cluster`, `mahal`, and `posterior` for cluster analysis.

Visualize the object by using `pdf` and `fsurf`.

```
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fsurf(gmPDF,[-10 10])
```



Fit Gaussian Mixture Model to Data Using `fitgmdist`

Generate random variates that follow a mixture of two bivariate Gaussian distributions by using the `mvnrnd` function. Fit a Gaussian mixture model (GMM) to the generated data by using the `fitgmdist` function.

Define the distribution parameters (means and covariances) of two bivariate Gaussian mixture components.

```
mu1 = [1 2];           % Mean of the 1st component
sigma1 = [2 0; 0 .5]; % Covariance of the 1st component
mu2 = [-3 -5];         % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Covariance of the 2nd component
```

Generate an equal number of random variates from each component, and combine the two sets of random variates.

```
rng('default') % For reproducibility
r1 = mvnrnd(mu1,sigma1,1000);
r2 = mvnrnd(mu2,sigma2,1000);
X = [r1; r2];
```

The combined data set X contains random variates following a mixture of two bivariate Gaussian distributions.

Fit a two-component GMM to X.

```
gm = fitgmdist(X,2)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:   -2.9617   -4.9727

Component 2:
Mixing proportion: 0.500000
Mean:    0.9539    2.0261
```

List the properties of the gm object.

```
properties(gm)
```

```
Properties for class gmdistribution:

    NumVariables
    DistributionName
    NumComponents
    ComponentProportion
    SharedCovariance
    NumIterations
    RegularizationValue
    NegativeLogLikelihood
    CovarianceType
    mu
    Sigma
    AIC
    BIC
    Converged
    ProbabilityTolerance
```

You can access these properties by using dot notation. For example, access the NegativeLogLikelihood property, which represents the negative loglikelihood of the data X given the fitted model.

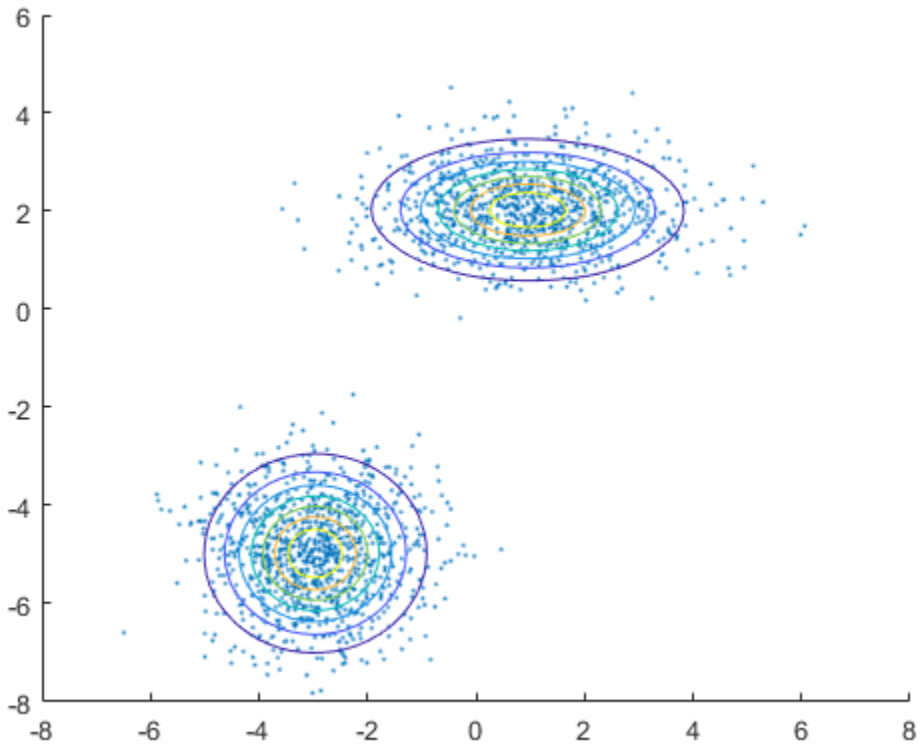
```
gm.NegativeLogLikelihood
```

```
ans = 7.0584e+03
```

After you create a gmdistribution object, you can use the object functions. Use cdf and pdf to compute the values of the cumulative distribution function (cdf) and the probability density function (pdf). Use random to generate random variates. Use cluster, mahal, and posterior for cluster analysis.

Plot X by using scatter. Visualize the fitted model gm by using pdf and fcontour.

```
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fcontour(gmPDF,[-8 6])
```



References

[1] McLachlan, G., and D. Peel. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons, Inc., 2000.

See Also

`fitgmdist`

Topics

“Simulate Data from Gaussian Mixture Model” on page 5-119
“Cluster Using Gaussian Mixture Model” on page 16-39

Introduced in R2007b

gname

Add case names to plot

Syntax

```
gname(cases)
gname
h = gname(cases,line_handle)
```

Description

`gname(cases)` displays a figure window and waits for you to press a mouse button or a keyboard key. The input argument `cases` is a character array, string array, or cell array of character vectors, in which each row of the character array or each element of the string array or cell array contains the case name of a point. Moving the mouse over the graph displays a pair of cross-hairs. If you position the cross-hairs near a point with the mouse and click once, the graph displays the label corresponding to that point. Alternatively, you can click and drag the mouse to create a rectangle around several points. When you release the mouse button, the graph displays the labels for all points in the rectangle. Right-click a point to remove its label. When you are done labelling points, press the **Enter** or **Escape** key to stop labeling.

`gname` with no arguments labels each case with its case number.

`cases` typically contains unique case names for each point, and is a string array, cell array of character vectors, or character matrix with each row representing a name. `cases` can also be any grouping variable, which `gname` converts to labels.

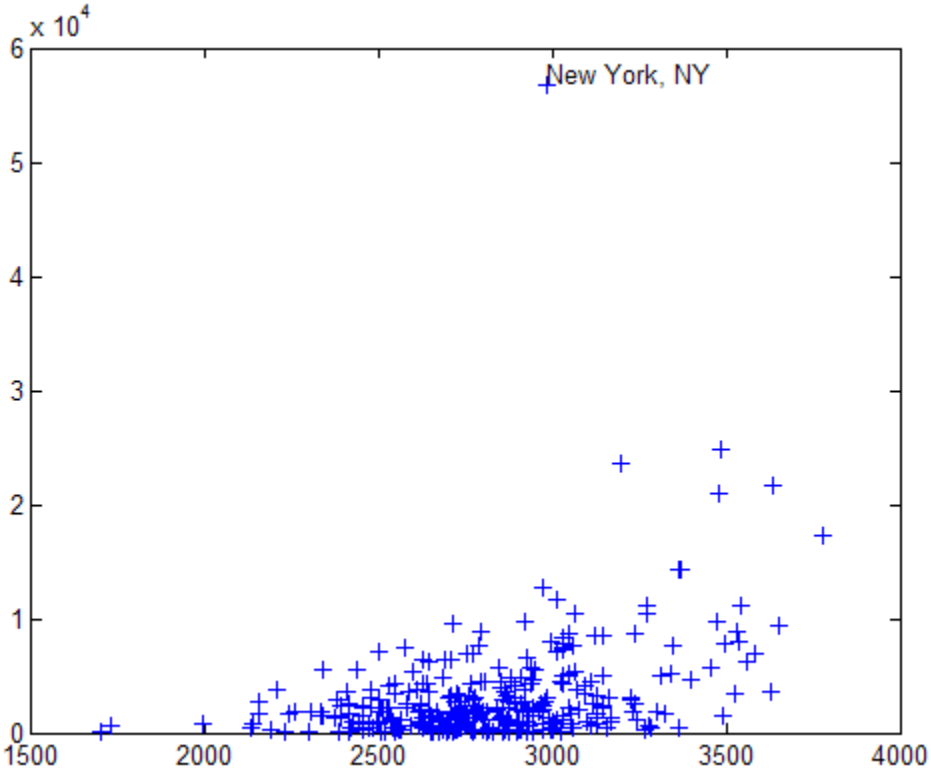
`h = gname(cases,line_handle)` returns a vector of handles to the text objects on the plot. Use the scalar `line_handle` to identify the correct line if there is more than one line object on the plot.

You can use `gname` to label plots created by the `plot`, `scatter`, `gscatter`, `plotmatrix`, and `gplotmatrix` functions.

Examples

This example uses the city ratings data sets to find out which cities are the best and worst for education and the arts.

```
load cities
education = ratings(:,6);
arts = ratings(:,7);
plot(education,arts,'+')
gname(names)
```



Click the point at the top of the graph to display its label, “New York.”

See Also

`gplotmatrix` | `gscatter` | `gtext`

Introduced before R2006a

gpcdf

Generalized Pareto cumulative distribution function

Syntax

```
p = gpcdf(x,k,sigma,theta)
p = gpcdf(x,k,sigma,theta,'upper')
```

Description

`p = gpcdf(x,k,sigma,theta)` returns the cdf of the generalized Pareto (GP) distribution with the tail index (shape) parameter `k`, scale parameter `sigma`, and threshold (location) parameter, `theta`, evaluated at the values in `x`. The size of `p` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

`p = gpcdf(x,k,sigma,theta,'upper')` returns the complement of the cdf of the generalized Pareto (GP) distribution, using an algorithm that more accurately computes the extreme upper tail probabilities.

Default values for `k`, `sigma`, and `theta` are 0, 1, and 0, respectively.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. When $k \geq 0$, the GP has positive density for

`x > theta`, or, when

$k < 0$, $0 \leq \frac{x-\theta}{\sigma} \leq -\frac{1}{k}$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `gpfit` | `gpinv` | `gplike` | `gppdf` | `gprnd` | `gpstat`

Topics

“Generalized Pareto Distribution” on page B-59

“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gpfit

Generalized Pareto parameter estimates

Syntax

```
parmhat = gpfit(x)
[parmhat,parmci] = gpfit(x)
[parmhat,parmci] = gpfit(x,alpha)
[...] = gpfit(x,alpha,options)
```

Description

`parmhat = gpfit(x)` returns maximum likelihood estimates of the parameters for the two-parameter generalized Pareto (GP) distribution given the data in `x`. `parmhat(1)` is the tail index (shape) parameter, `k` and `parmhat(2)` is the scale parameter, `sigma`. `gpfit` does not fit a threshold (location) parameter.

`[parmhat,parmci] = gpfit(x)` returns 95% confidence intervals for the parameter estimates.

`[parmhat,parmci] = gpfit(x,alpha)` returns `100(1-alpha)`% confidence intervals for the parameter estimates.

`[...] = gpfit(x,alpha,options)` specifies control parameters for the iterative algorithm used to compute ML estimates. This argument can be created by a call to `statset`. See `statset('gpfit')` for parameter names and default values.

Other functions for the generalized Pareto, such as `gpcdf` allow a threshold parameter, `theta`. However, `gpfit` does not estimate theta. It is assumed to be known, and subtracted from `x` before calling `gpfit`.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. When $k \geq 0$, the GP has positive density for

`k > theta`, or, when $k < 0$, for

$$0 \leq \frac{x-\theta}{\sigma} \leq -\frac{1}{k}$$

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

See Also

`gpcdf` | `gpinv` | `gplike` | `gppdf` | `gprnd` | `gpstat` | `mle`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gpinv

Generalized Pareto inverse cumulative distribution function

Syntax

```
x = gpinv(p,k,sigma,theta)
```

Description

`x = gpinv(p,k,sigma,theta)` returns the inverse cdf for a generalized Pareto (GP) distribution with tail index (shape) parameter `k`, scale parameter `sigma`, and threshold (location) parameter `theta`, evaluated at the values in `p`. The size of `x` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

Default values for `k`, `sigma`, and `theta` are 0, 1, and 0, respectively.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. When $k \geq 0$, the GP has positive density for

`x > theta`, or, when

$k < 0$, $0 \leq \frac{x-\theta}{\sigma} \leq -\frac{1}{k}$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gpcdf` | `gpfit` | `gplike` | `gppdf` | `gprnd` | `gpstat` | `icdf`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gplike

Generalized Pareto negative loglikelihood

Syntax

```
nlogL = gplike(params,data)
[nlogL,acov] = gplike(params,data)
```

Description

`nlogL = gplike(params,data)` returns the negative of the loglikelihood `nlogL` for the two-parameter generalized Pareto (GP) distribution, evaluated at parameters `params`. `params(1)` is the tail index (shape) parameter, `k`, and `params(2)` is the scale parameter. `gplike` does not allow a threshold (location) parameter.

`[nlogL,acov] = gplike(params,data)` returns the inverse of Fisher's information matrix, `acov`. If the input parameter values in `params` are the maximum likelihood estimates, the diagonal elements of `acov` are their asymptotic variances. `acov` is based on the observed Fisher's information, not the expected information.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. When $k \geq 0$, the GP has positive density for

`x > theta`, or, when

$k < 0$, $0 \leq \frac{x-\theta}{\sigma} \leq -\frac{1}{k}$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

See Also

`gpcdf` | `gpfit` | `gpinv` | `gppdf` | `gprnd` | `gpstat`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gppdf

Generalized Pareto probability density function

Syntax

```
p = gppdf(x,k,sigma,theta)
```

Description

`p = gppdf(x,k,sigma,theta)` returns the pdf of the generalized Pareto (GP) distribution with the tail index (shape) parameter `k`, scale parameter `sigma`, and threshold (location) parameter, `theta`, evaluated at the values in `x`. The size of `p` is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

Default values for `k`, `sigma`, and `theta` are 0, 1, and 0, respectively.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when `k ≥ 1`, and the variance is not finite when `k ≥ 1/2`. When `k ≥ 0`, the GP has positive density for

`x > theta`, or, when

`k < 0`, $0 \le \frac{x-\theta}{\sigma} \le -\frac{1}{k}$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gpcdf` | `gpfit` | `gpinv` | `gplike` | `gprnd` | `gpstat` | `pdf`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gplotmatrix

Matrix of scatter plots by group

Syntax

```
gplotmatrix(X,[],group)
gplotmatrix(X,Y,group)
gplotmatrix(X,Y,group,clr,sym,siz)
gplotmatrix(X,Y,group,clr,sym,siz,doleg)
gplotmatrix(X,[],group,clr,sym,siz,doleg,dispopt)
gplotmatrix(X,[],group,clr,sym,siz,doleg,dispopt,xnam)
gplotmatrix(X,Y,group,clr,sym,siz,doleg,[],xnam,ynam)

gplotmatrix(parent,___)

[h,ax,bigax] = gplotmatrix(___)
```

Description

`gplotmatrix(X,[],group)` creates a matrix of scatter plots and histograms of the data in `X`, grouped by the grouping variable in `group`. Each off-diagonal plot in the resulting figure is a scatter plot of a column of `X` against another column of `X`. The software also plots the outlines of the grouped histograms in the diagonal plots of the plot matrix. `X` and `group` must have the same number of rows.

`gplotmatrix(X,Y,group)` creates a matrix of scatter plots. Each plot in the resulting figure is a scatter plot of a column of `X` against a column of `Y`. For example, if `X` has p columns and `Y` has q columns, then the figure contains a q-by-p matrix of scatter plots. All plots are grouped by the grouping variable `group`. The input arguments `X`, `Y`, and `group` must all have the same number of rows.

`gplotmatrix(X,Y,group,clr,sym,siz)` specifies the marker color `clr`, symbol `sym`, and size `siz` for each group.

`gplotmatrix(X,Y,group,clr,sym,siz,doleg)` controls whether a legend is displayed in the figure. `gplotmatrix` creates a legend by default.

`gplotmatrix(X,[],group,clr,sym,siz,doleg,dispopt)` controls the display options for the diagonal plots in the plot matrix of `X`.

`gplotmatrix(X,[],group,clr,sym,siz,doleg,dispopt,xnam)` labels the x-axes and y-axes of the scatter plots using the column names specified in `xnam`. The input argument `xnam` must contain one name for each column of `X`. Set `dispopt` to `'variable'` to display the variable names along the diagonal of the scatter plot matrix.

`gplotmatrix(X,Y,group,clr,sym,siz,doleg,[],xnam,ynam)` labels the x-axes and y-axes of the scatter plots using the column names specified in `xnam` and `ynam`. The input arguments `xnam` and `ynam` must contain one name for each column of `X` and `Y`, respectively.

`gplotmatrix(parent,___)` creates the scatter plot matrix in the figure or panel specified by `parent`. Specify `parent` as the first input argument followed by any of the input argument combinations in the previous syntaxes.

`[h,ax,bigax] = gplotmatrix( ___ )` returns graphics handles to the individual plots and the entire scatter plot matrix.

You can pass in `[]` for `clr`, `sym`, `siz`, `doleg`, and `dispopt` to use their default values.

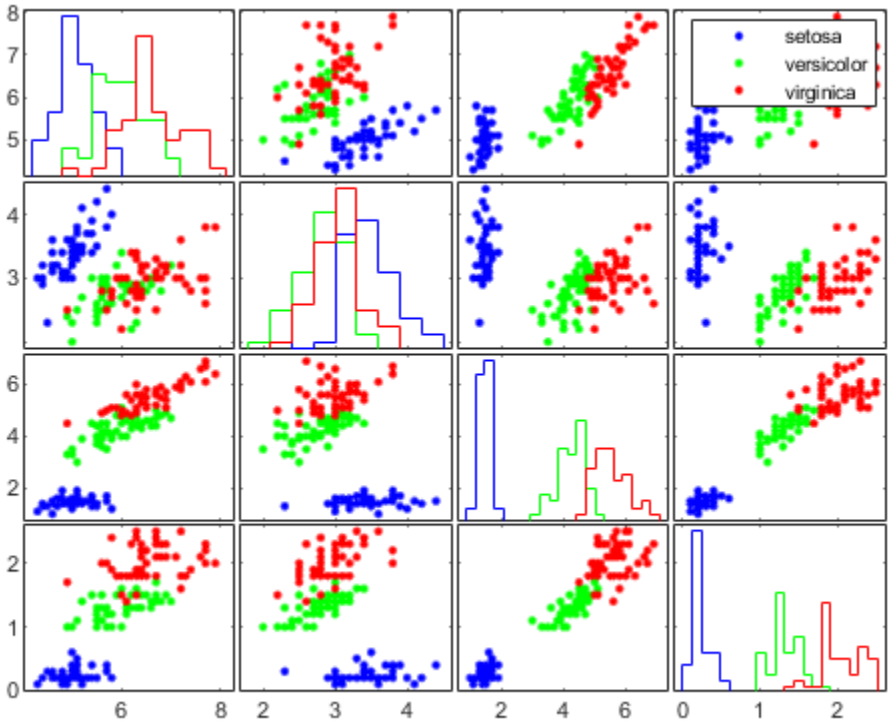
Examples

Scatter Plots with Grouped Data

Create a matrix of scatter plots for each combination of variables in a data set. Group the data according to a separate variable.

Load the `fisheriris` data set, which contains flower data. The four columns of `meas` are the sepal length, sepal width, petal length, and petal width of the flowers. `species` contains the flower species names: setosa, versicolor, and virginica. Visually compare the flower measurements across flower species.

```
load fisheriris
gplotmatrix(meas,[],species)
```



In the matrix of scatter plots, the x-axis of the leftmost column of scatter plots corresponds to sepal length, the first column in `meas`. Similarly, the y-axis of the bottom row of scatter plots corresponds to petal width, the last column in `meas`. Therefore, the scatter plot in the bottom left of the matrix compares sepal length values (along the x-axis) to petal width values (along the y-axis). The color of each point depends on the species of the flower.

The diagonal plots are histograms rather than scatter plots. For example, the plot in the top left of the matrix shows the distribution of sepal length values for each species of flower.

Create Scatter Plot Matrix with Subset of Variables

Create scatter plots comparing a subset of the variables in a data set to another subset of variables. Group the data according to a separate variable.

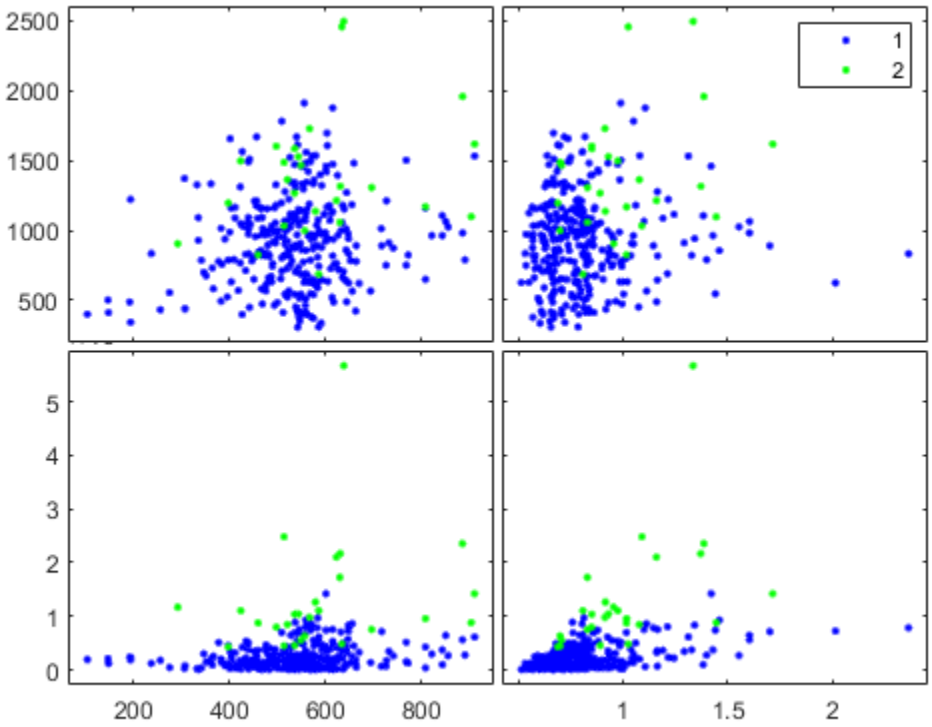
Load the `discrim` data set.

```
load discrim
```

The `ratings` array contains rating values of 329 US cities for the nine categories listed in the `categories` array. The `group` array contains a city size code that is equal to `2` for the 26 largest cities, and `1` otherwise.

Create a matrix of scatter plots to compare the first two categories, `climate` and `housing`, with the fourth and seventh categories, `crime` and `arts`. Specify `group` as the grouping variable to visually distinguish the data for large and small cities.

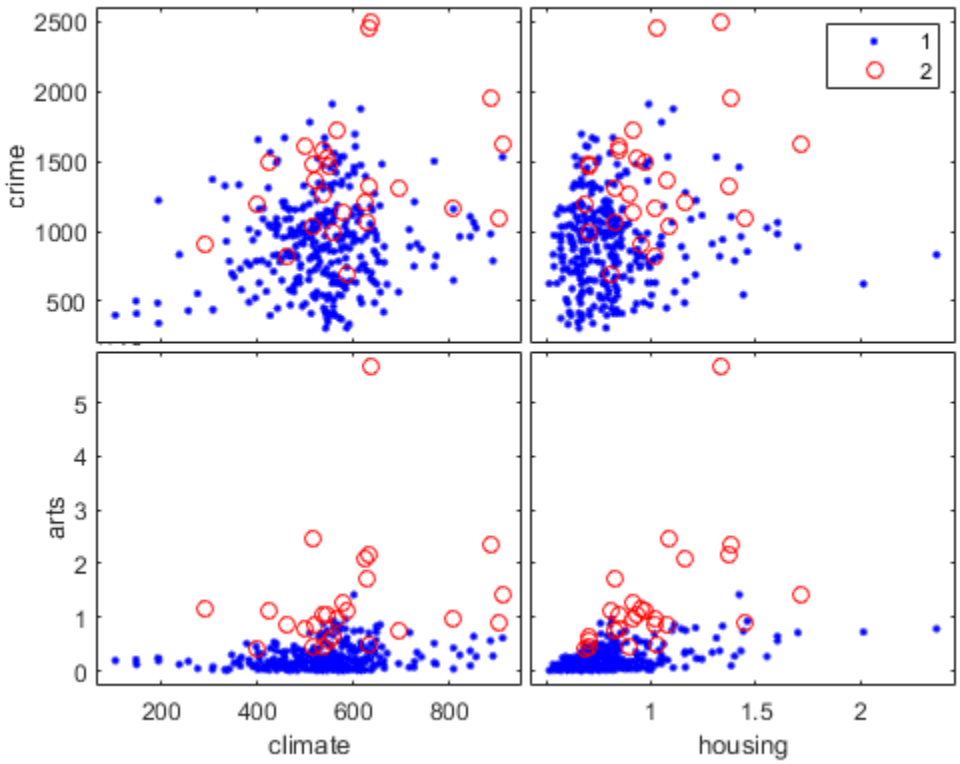
```
X = ratings(:,1:2);
Y = ratings(:,[4 7]);
gplotmatrix(X,Y,group)
```



The matrix of scatter plots shows the specified comparisons, with each city size group represented by a different color.

Adjust the appearance of the plots by specifying marker colors and symbols, and labeling the axes with the rating categories.

```
xnames = categories(1:2,:);
ynames = categories([4 7],:);
gplotmatrix(X,Y,group,'br','.o',[],'on',[],xnames,ynames)
```



Scatter Plot Matrix with Multiple Grouping Variables

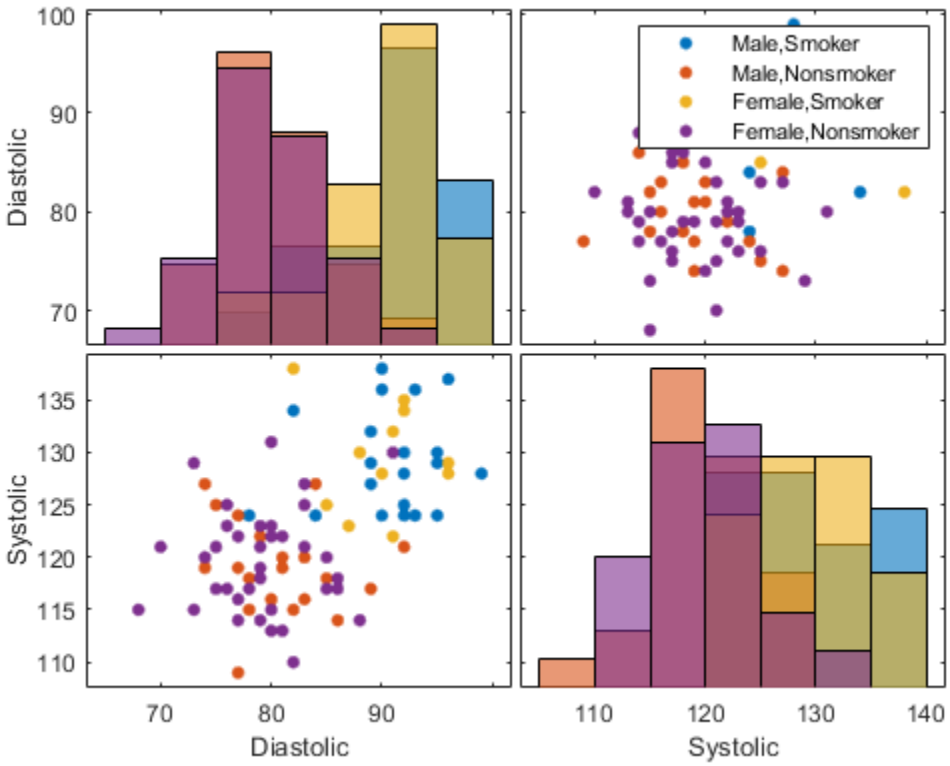
Create a matrix of scatter plots comparing data variables by using two grouping variables.

Load the `patients` data set. Compare patient diastolic and systolic blood pressure values. Group the patients according to their gender and smoker status. Convert `Smoker` to a categorical variable to have more descriptive labels in the legend. Display grouped histograms along the diagonal of the plot matrix by using the `'grpbars'` display option, and label the axes.

```
load patients
X = [Diastolic Systolic];
labeledSmoker = categorical(Smoker,[true false],{'Smoker','Nonsmoker'});
group = {Gender,labeledSmoker};
color = lines(4)
```

```
color = 4×3

         0    0.4470    0.7410
    0.8500    0.3250    0.0980
    0.9290    0.6940    0.1250
    0.4940    0.1840    0.5560
```

```
xnames = {'Diastolic','Systolic'};
gplotmatrix(X,[],group,color,[],[],[],'grpbars',xnames)
```



For example, the scatter plot in the bottom left of the matrix shows that smokers (blue and yellow markers) tend to have higher diastolic and systolic blood pressure values, regardless of gender.

Modify Scatter Plot Matrix Appearance

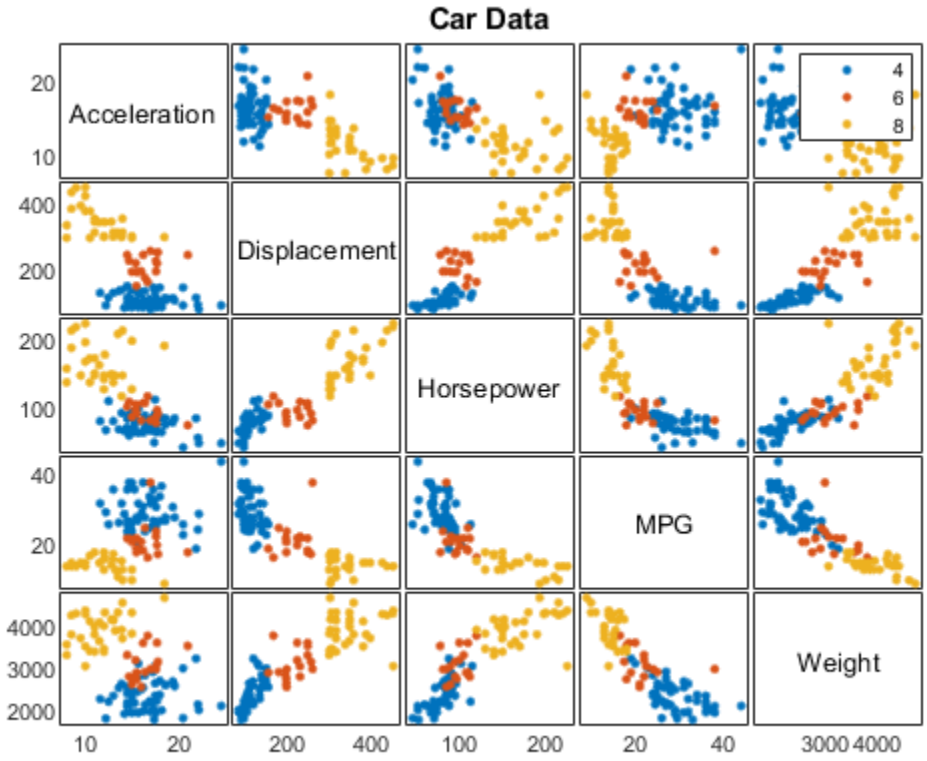
Create a matrix of scatter plots that display grouped data. Modify the appearance of one of the scatter plots.

Load the `carsmall` data set. Create a scatter plot matrix using different car measurements. Group the cars by the number of cylinders. Specify the group colors, and display the car variable names along the diagonal of the plot matrix. Add a title to the plot matrix.

```
load carsmall
X = [Acceleration Displacement Horsepower MPG Weight];
color = lines(3)
```

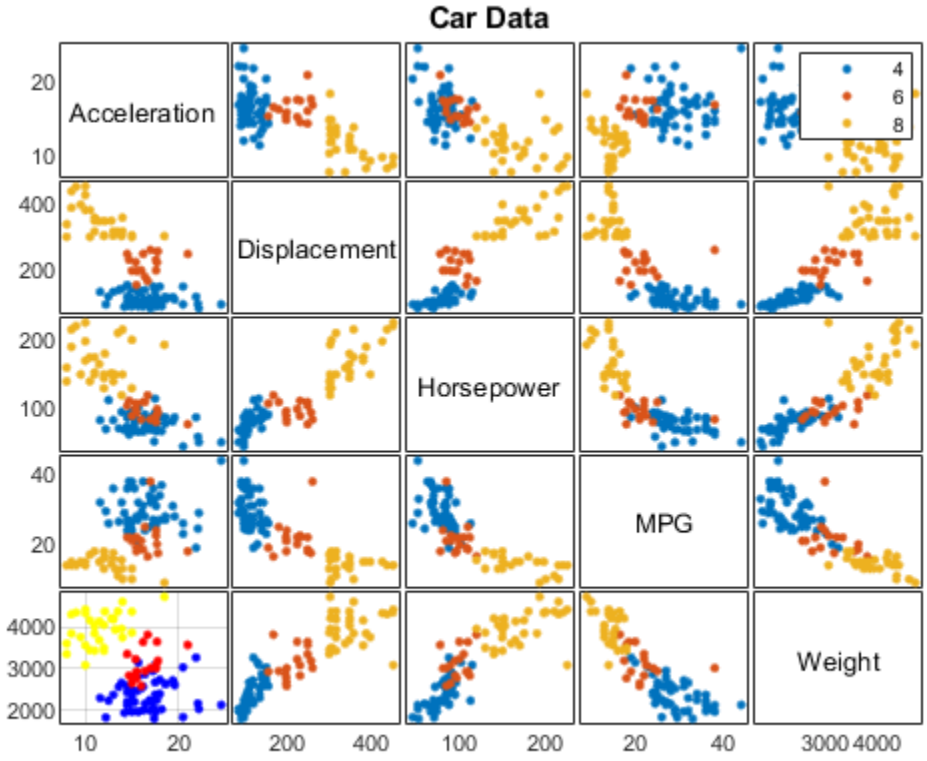
```
color = 3×3

         0    0.4470    0.7410
    0.8500    0.3250    0.0980
    0.9290    0.6940    0.1250
```

```
xnames = {'Acceleration','Displacement','Horsepower','MPG','Weight'};
[h,ax] = gplotmatrix(X,[],Cylinders,color,[],[],[],'variable',xnames);
title('Car Data')
```



Change the appearance of the scatter plot in the bottom left of the matrix by using `h` and `ax`. First, change the colors of the data points in the scatter plot. Then, add grid lines to the scatter plot.

```
bottomleftPlot = h(5,1,:);
bottomleftPlot(1).Color = 'blue';
bottomleftPlot(2).Color = 'red';
bottomleftPlot(3).Color = 'yellow';

bottomleftAxes = ax(5,1);
bottomleftAxes.XGrid = 'on';
bottomleftAxes.YGrid = 'on';
```



Input Arguments

X — Input data

numeric matrix | datetime array | duration array

Input data, specified as an *n*-by-*p* numeric matrix, datetime array, or duration array. `gplotmatrix` creates a matrix of plots using the columns of `X`. If you do not specify an additional input matrix `Y`, then `gplotmatrix` creates a *p*-by-*p* matrix of plots. The off-diagonal plots are scatter plots, and the diagonal plots depend on the value of `dispopt`. In each scatter plot, `gplotmatrix` plots one column of `X` against another column of `X`. The points in the scatter plots are grouped according to `group`.

If you specify `Y`, then `gplotmatrix` creates a *q*-by-*p* matrix of scatter plots using the *p* columns of `X` and the *q* columns of `Y`.

Data Types: `single` | `double` | `datetime` | `duration`

Y — Input data

numeric matrix | datetime array | duration array

Input data, specified as an *n*-by-*q* numeric matrix, datetime array, or duration array. `gplotmatrix` creates a *q*-by-*p* matrix of scatter plots using the *p* columns of `X` and the *q* columns of `Y`. For each column of the plot matrix, the *x*-axis values of the scatter plots are the same as the values in the corresponding column of `X`. Similarly, for each row of the plot matrix, the *y*-axis values of the scatter plots are the same as the values in the corresponding column of `Y`. The points in the scatter plots are grouped according to `group`.

X and Y must have the same number of rows.

Data Types: `single` | `double` | `datetime` | `duration`

`group` — Grouping variable

categorical vector | numeric vector | logical vector | character array | string array | cell array

Grouping variable, specified as a categorical vector, numeric vector, logical vector, character array, string array, or cell array of character vectors. Alternatively, `group` can be a cell array containing several grouping variables (such as `{g1 g2 g3}`), in which case observations are in the same group if they have common values of all grouping variables. In any case, `group` must have the same number of rows as X. Points in the same group appear on the graph with the same marker color, symbol, and size.

Example:
```
categorical({'blue','red','yellow','blue','yellow','red','red','yellow','blue
','red'})
```

Example: `{Smoker,Gender}` where `Smoker` and `Gender` are grouping variables

Data Types: `categorical` | `single` | `double` | `logical` | `char` | `string` | `cell`

`clr` — Marker colors

character vector | string scalar | string array | cell array of character vectors | matrix of RGB values

Marker colors, specified as one of the following:

- Character vector or string scalar of color short names.
- String array or cell array of character vectors designating color names or short names.
- Three-column matrix of RGB values in the range [0,1]. The three columns represent the R (red) value, G (green) value, and B (blue) value.

You can choose among these predefined colors and their equivalent RGB triplets.

Color Name	Short Name	RGB Triplet	Appearance
`'red'`	`'r'`	`[1 0 0]`	
`'green'`	`'g'`	`[0 1 0]`	
`'blue'`	`'b'`	`[0 0 1]`	
`'cyan'`	`'c'`	`[0 1 1]`	
`'magenta'`	`'m'`	`[1 0 1]`	
`'yellow'`	`'y'`	`[1 1 0]`	
`'black'`	`'k'`	`[0 0 0]`	
`'white'`	`'w'`	`[1 1 1]`	

When the total number of groups exceeds the number of specified colors, `gplotmatrix` cycles through the specified colors.

Example: `{'blue','black','green'}`

Example: `[0 0 1; 0 0.5 0.5; 0.5 0.5 0.5]`

Data Types: `char` | `string` | `cell` | `single` | `double`

`sym` — Marker symbols

`'.'` (default) | character vector | string scalar

Marker symbols, specified as a character vector or string scalar.

You can choose among these marker options.

Value	Description
`'o'`	Circle
`'+'`	Plus sign
`'*'`	Asterisk
`'.'`	Point
`'x'`	Cross
`'s'`	Square
`'d'`	Diamond
`'^'`	Upward-pointing triangle
`'v'`	Downward-pointing triangle
`'>'`	Right-pointing triangle
`'<'`	Left-pointing triangle
`'p'`	Five-pointed star (pentagram)
`'h'`	Six-pointed star (hexagram)
`'none'`	No markers

By default, `gplotmatrix` assigns `'.'` as the marker symbol for each group. When the total number of groups exceeds the number of specified symbols, `gplotmatrix` cycles through the specified symbols.

Example: `'x'`

Example: `'xo+'`

Data Types: `char` | `string`

`siz` — Marker sizes

positive numeric vector

Marker sizes, specified as a positive numeric vector. The default value is determined by the number of observations. When the total number of groups exceeds the number of specified sizes, `gplotmatrix` cycles through the specified sizes.

Example: `[6 12]`

Data Types: `single` | `double`

`doleg` — Option to include legend

`'on'` (default) | `'off'`

Option to include a legend, specified as either `'on'` or `'off'`. By default, the legend is displayed in the figure.

dispopt — Display options for diagonal plots
'stairs' (default) | 'hist' | 'grpbars' | 'none' | 'variable'

Display options for the diagonal plots in the plot matrix, specified as 'stairs', 'hist','grpbars', 'none', or 'variable'. This table describes the different display options.

Value	Description
'stairs'	Plot the outlines of grouped histograms.
'hist'	Plot histograms.
'grpbars'	Plot grouped histograms.
'none'	Display blank plots.
'variable'	Display variable names. To use this display option, you must specify xnam.

The default is 'stairs' when group contains more than one group. Otherwise, gplotmatrix displays a single histogram in each diagonal plot.

To generate the histograms, gplotmatrix uses the 'pdf' type of normalization for numeric data and the 'count' type of normalization for datetime and duration data. (See the 'Normalization' name-value pair argument of histogram.) Note that the y-axis tick mark labels do not apply to the histograms. Use data tips to see the correct histogram values.

xnam — X column names
character array | string array | cell array of character vectors

X column names, specified as a character array, string array, or cell array of character vectors. xnam must contain one name for each column of X.

Example: {'Sepal Length','Sepal Width','Petal Length','Petal Width'}

Data Types: char | string | cell

ynam — Y column names
character array | string array | cell array of character vectors

Y column names, specified as a character array, string array, or cell array of character vectors. ynam must contain one name for each column of Y.

Example: {'Diastolic','Systolic'}

Data Types: char | string | cell

parent — Parent container
Figure object | Panel object

Parent container, specified as a Figure or Panel object.

Output Arguments

h — Line handles to individual plots
array of Line and Histogram objects

Line handles to individual plots, returned as a one of these arrays:

- *p*-by-*p*-by-*k* array of `Line` and `Histogram` objects if you do not specify `Y`
- *q*-by-*p*-by-*k* array of `Line` objects if you specify both `X` and `Y`

p is the number of columns in `X`, *q* is the number of columns in `Y`, and *k* is the number of unique groups in `group`.

Each scatter plot has *k* corresponding `Line` objects in `h`, and each histogram has *k* corresponding `Histogram` objects in `h`.

ax — Axes handles to individual plots
matrix of `Axes` objects

Axes handles to individual plots, returned as a matrix of `Axes` objects. If `dispopt` is `'hist'`, `'stairs'`, or `'grpbars'`, then `ax` contains one extra row of handles to invisible axes where the histograms are plotted.

bigax — Axes handle to entire plot matrix
`Axes` object

Axes handle to the entire plot matrix, returned as an `Axes` object. `bigax` points to the current axes, so a subsequent `title`, `xlabel`, or `ylabel` command produces labels that are centered with respect to the entire plot matrix.

See Also

`grpstats` | `gscatter` | `plotmatrix`

Topics
“Create Scatter Plots Using Grouped Data” on page 4-2
“MANOVA” on page 9-49
“Grouping Variables” on page 2-45

Introduced before R2006a

gprnd

Generalized Pareto random numbers

Syntax

```
r = gprnd(k,sigma,theta)
r = gprnd(k,sigma,theta,m,n,...)
R = gprnd(K,sigma,theta,[m,n,...])
```

Description

`r = gprnd(k,sigma,theta)` returns an array of random numbers chosen from the generalized Pareto (GP) distribution with tail index (shape) parameter `k`, scale parameter `sigma`, and threshold (location) parameter, `theta`. The size of `r` is the common size of the input arguments if all are arrays. If any parameter is a scalar, the size of `r` is the size of the other parameters.

`r = gprnd(k,sigma,theta,m,n,...)` or `R = gprnd(K,sigma,theta,[m,n,...])` generates an `m`-by-`n`-by-... array. The `k`, `sigma`, `theta` parameters can each be scalars or arrays of the same size as `r`.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when `k ≥ 1`, and the variance is not finite when `k ≥ 1/2`. When `k ≥ 0`, the GP has positive density for

`x > theta`, or, when

$$0 \le \frac{x-\theta}{\sigma} \le -\frac{1}{k}$$

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gpcdf` | `gpfit` | `gpinv` | `gplike` | `gppdf` | `gpstat` | `random`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

gpstat

Generalized Pareto mean and variance

Syntax

```
[m,v] = gpstat(k,sigma,theta)
```

Description

`[m,v] = gpstat(k,sigma,theta)` returns the mean of and variance for the generalized Pareto (GP) distribution with the tail index (shape) parameter `k`, scale parameter `sigma`, and threshold (location) parameter, `theta`.

The default value for `theta` is 0.

When `k = 0` and `theta = 0`, the GP is equivalent to the exponential distribution. When `k > 0` and `theta = sigma/k`, the GP is equivalent to a Pareto distribution with a scale parameter equal to `sigma/k` and a shape parameter equal to `1/k`. The mean of the GP is not finite when $k \geq 1$, and the variance is not finite when $k \geq 1/2$. When $k \geq 0$, the GP has positive density for `x > theta`, or when

$k < 0$, $0 \leq \frac{x-\theta}{\sigma} \leq -\frac{1}{k}$.

References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`gpcdf` | `gpfit` | `gpinv` | `gplike` | `gppdf` | `gprnd`

Topics

“Generalized Pareto Distribution” on page B-59
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Supported Distributions” on page 5-14

Introduced before R2006a

growTrees

Class: `TreeBagger`

Train additional trees and add to ensemble

Syntax

```
B = growTrees(B,ntrees)
B = growTrees(B,ntrees,'param1',val1,'param2',val2,...)
```

Description

`B = growTrees(B,ntrees)` grows `ntrees` new trees and appends them to those trees already stored in the ensemble `B`.

`B = growTrees(B,ntrees,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'NumPrint'`	Specifies that a diagnostic message showing training progress should display after every `value` training cycles (grown trees). Default is no diagnostic messages.

`'Options'`	A `struct` that specifies options that govern computation when growing the ensemble of decision trees. One option requests that the computation of decision trees on multiple bootstrap replicates uses multiple processors, if the Parallel Computing Toolbox is available. Two options specify the random number streams to use in selecting bootstrap replicates. You can create this argument with a call to `statset`. You can retrieve values of the individual fields with a call to `statget`. Applicable `statset` parameters are:
	• `'UseParallel'` — If `true` and Parallel Computing Toolbox is installed, then the software uses an existing parallel pool for parallel trees, or, depending on parallel preferences, the software opens and uses a new pool if none is currently open. Otherwise, the software computes in serial. Default is `false`, meaning serial computation.
	For dual-core systems and above, `TreeBagger` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, using the `'UseParallel'` option on a single computer may not speed up computation much and may consume more memory than in serial. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.
	• `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. Default is `false`. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
	• `Streams` — A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `growTrees` uses the default stream or streams. If you choose to specify `Streams`, use a single object except in the case
	  • `UseParallel` is `true`
	  • `UseSubstreams` is `false`
	In that case, use a cell array the same size as the Parallel pool.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`TreeBagger` | `TreeBagger` | `fitctree` | `fitrtree` | `statget` | `statset`

grp2idx

Create index vector from grouping variable

Syntax

```
[g,gN] = grp2idx(s)
[g,gN,gL] = grp2idx(s)
```

Description

`[g,gN] = grp2idx(s)` creates an index vector `g` from the grouping variable `s`. The output `g` is a vector of integer values from 1 up to the number *K* of distinct groups. `gN` is a cell array of character vectors representing the list of group names.

`[g,gN,gL] = grp2idx(s)` also returns a column vector `gL` representing the list of the group levels with the same data type as `s`.

Examples

Create Index Vector from Categorical Vector

Create a categorical vector by using `discretize` and convert it to an index vector by using `grp2idx`.

Load the `hospital` data set and convert the ages in `hospital.Ages` to categorical values representing the ages by decade.

```
load hospital
edges = 0:10:100; % Bin edges
labels = strcat(num2str((0:10:90)','%d'),{'s'}); % Labels for the bins
s = discretize(hospital.Age,edges,'Categorical',labels);
```

Display the ages and the groups of ages for the first five samples.

```
ages = hospital.Age(1:5)
```

```
ages = 5×1

    38
    43
    38
    40
    49
```

```
groups = s(1:5)
```

```
groups = 5x1 categorical
     30s
     40s
     30s
```

```
     40s
     40s
```

Create an index vector from the categorical vector `s`.

```
[g,gN,gL] = grp2idx(s);
```

Display the index values corresponding to the first five samples.

```
g(1:5)
```

```
ans = 5×1

     4
     5
     4
     5
     5
```

Reproduce the input argument `s` using the output `gL`.

```
gL(g(1:5))
```

```
ans = 5x1 categorical
     30s
     40s
     30s
     40s
     40s
```

Use `gN(g)` to reproduce the input argument `s` as a cell array of character vectors.

```
gN(g(1:5))
```

```
ans = 5x1 cell
    {'30s'}
    {'40s'}
    {'30s'}
    {'40s'}
    {'40s'}
```

Input Arguments

`s` — Grouping variable

categorical vector | numeric vector | logical vector | `datetime` vector | duration vector | string array | cell array of character vectors | character array

Grouping variable, specified as a categorical, numeric, logical, `datetime`, or duration vector, a string array, a cell array of character vectors, or a character array with each row representing a group label.

`grp2idx` treats `NaN`s (numeric, duration, or logical), `''` (empty character arrays or cell arrays of character vectors), `""` (empty strings), `<missing>` values (string), `<undefined>` values

(categorical), and NaTs (`datetime`) in `s` as missing values and returns `NaN`s in the corresponding rows of `g`. The outputs `gN` and `gL` do not include entries for missing values.

Data Types: `single` | `double` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `logical` | `char` | `string` | `cell` | `categorical` | `datetime` | `duration`

Output Arguments

`g` — Group index

positive integer vector

Group index, returned as a positive integer vector with values from 1 up to the number K of distinct groups in `s`.

`gN` — List of group names

cell array of character vectors

List of group names, returned as a cell array of character vectors.

The order of `gN` depends on the data type of the grouping variable `s`.

- For numeric and logical vectors, the order is the sorted order of `s`.
- For categorical vectors, the order is the order of `categories(s)`.
- For other data types, the order is the order of first appearance in `s`.

`gN(g)` reproduces the contents of `s` in a cell array.

`gL` — List of group levels

categorical vector | numeric vector | logical vector | `datetime` vector | duration vector | cell array of character vectors | character array

List of group levels, returned as the same data type as `s`: a categorical, numeric, logical, `datetime`, or duration vector, a cell array of character vectors, or a character array with each row representing a group label. (The software treats string arrays as cell arrays of character vectors.)

The set of groups and their order in `gL` are the same as those in `gN`, but `gL` has the same data type as `s`.

If `s` is a character matrix, then `gL(g,:)` reproduces `s`; otherwise, `gL(g)` reproduces `s`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `s` can be a numeric, logical, or character vector or a cell array of character vectors. Code generation does not support a categorical, `datetime`, or duration vector or a string array for the input argument.
- In the generated code, the second and third outputs, `gN` and `gL`, are identical. `gN` and `gL` have the same data type as the input argument `s`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`categories` | `crosstab` | `findgroups` | `grpstats` | `gscatter`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

grpstats

Summary statistics organized by group

Syntax

```
statarray = grpstats(tbl,groupvar)
statarray = grpstats(tbl,groupvar,whichstats)
statarray = grpstats(tbl,groupvar,whichstats,Name,Value)

means = grpstats(X,group)
[stats1,...,statsN] = grpstats(X,group,whichstats)
[stats1,...,statsN] = grpstats(X,group,whichstats,'Alpha',alpha)

grpstats(X,group,alpha)
```

Description

`statarray = grpstats(tbl,groupvar)` returns a table or dataset array with the means for the data groups specified in `tbl` determined by the values of the grouping variable or variables specified in `groupvar`.

- If there is a single grouping variable, then there is a row in `statarray` for each value of the grouping variable. `grpstats` sorts the groups by order of appearance (if the grouping variable is a character vector or string scalar), in ascending numeric order (if the grouping variable is numeric), or in order of the levels (if the grouping variable is categorical).
- If `groupvar` is a string array or cell array of character vectors containing multiple grouping variable names, or a vector of column numbers, then there is a row in `statarray` for each observed unique combination of values of the grouping variables. `grpstats` sorts the groups by the values of the first grouping variable, then the second grouping variable, and so on.
- If any variables in `tbl` (other than those specified in `groupvar`) are not numeric or logical arrays, then you must specify the names or column numbers of the numeric and logical variables for which you want to calculate means using the name-value pair argument, `DataVars`.

`statarray = grpstats(tbl,groupvar,whichstats)` returns the group values for the summary statistics types specified in `whichstats`.

`statarray = grpstats(tbl,groupvar,whichstats,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments.

`means = grpstats(X,group)` returns a column vector or matrix with the means of the groups of the data in the matrix or vector `X` determined by the values of the grouping variable or variables, `group`. The rows of `means` correspond to the grouping variable values.

- If there is a single grouping variable, then there is a row in `means` for each value of the grouping variable. `grpstats` sorts the groups by order of appearance (if the grouping variable is a character vector or string scalar), in ascending numeric order (if the grouping variable is numeric), or in order of the levels (if the grouping variable is categorical).
- If `group` is a string array or cell array of grouping variables, then there is a row in `means` for each observed unique combination of values of the grouping variables. `grpstats` sorts the groups by the values of the first grouping variable, then the second grouping variable, and so on.

- If X is a matrix, then means is a matrix with the same number of columns as X. Each column of means has the group means for the corresponding column of X.

[stats1,...,statsN] = grpstats(X,group,whichstats) returns column vectors or arrays with group values for the summary statistic types specified in whichstats.

[stats1,...,statsN] = grpstats(X,group,whichstats,'Alpha',alpha) specifies the significance level for confidence and prediction intervals.

grpstats(X,group,alpha) plots the means of the groups of data in the vector or matrix X determined by the values of the grouping variable, group. The grouping variable values are on the horizontal plot axis. Each group mean has 100×(1 – alpha)% confidence intervals.

- If X is a matrix, then grpstats plots the means and confidence intervals for each column of X.
- If group is a cell array of grouping variables, then grpstats plots the means and confidence intervals for the groups of data in X determined by the unique combinations of values of the grouping variables. For example, if there are two grouping variables, each with two values, there are four possible combinations of grouping variable values. The plot includes only the combinations of values that exist in the input grouping variables (not all possible combinations).

Examples

Dataset Array Summary Statistics Organized by Group

Load the sample data.

```
load('hospital')
```

The dataset array hospital has 100 observations and 7 variables.

Create a dataset array with only the variables Sex, Age, Weight, and Smoker.

```
dsa = hospital(:,{'Sex','Age','Weight','Smoker'});
```

Sex is a nominal array, with levels Male and Female. The variables Age and Weight have numeric values, and Smoker has logical values.

Compute the mean for the numeric and logical arrays, Age, Weight, and Smoker, grouped by the levels in Sex.

```
statarray = grpstats(dsa,'Sex')
```

```
statarray =
              Sex       GroupCount    mean_Age    mean_Weight    mean_Smoker
    Female    Female    53            37.717      130.47         0.24528
    Male      Male      47            38.915      180.53         0.44681
```

statarray is a dataset array with two rows, corresponding to the levels in Sex. GroupCount is the number of observations in each group. The means of Age, Weight, and Smoker, grouped by Sex, are given in mean_Age, mean_Weight, and mean_Smoker.

Compute the mean for Age and Weight, grouped by the values in Smoker.

```
statarray = grpstats(dsa,'Smoker','mean','DataVars',{'Age','Weight'})
```

```
statarray =
         Smoker    GroupCount    mean_Age    mean_Weight
    0    false     66             37.97      149.91
    1    true      34            38.882      161.94
```

In this case, not all variables in `dsa` (excluding the grouping variable, `Smoker`) are numeric or logical arrays; the variable `Sex` is a nominal array. When not all variables in the input dataset array are numeric or logical arrays, you must specify the variables for which you want to calculate summary statistics using `DataVars`.

Compute the minimum and maximum weight, grouped by the combinations of values in `Sex` and `Smoker`.

```
statarray = grpstats(dsa,{'Sex','Smoker'},{'min','max'},...
                     'DataVars','Weight')
```

```
statarray =
                Sex       Smoker    GroupCount    min_Weight    max_Weight
    Female_0    Female    false     40            111           147
    Female_1    Female    true      13            115           146
    Male_0      Male      false     26            158           194
    Male_1      Male      true      21            164           202
```

There are two unique values in `Smoker` and two levels in `Sex`, for a total of four possible combinations of values: Female Nonsmoker (`Female_0`), Female Smoker (`Female_1`), Male Nonsmoker (`Male_0`), and Male Smoker (`Male_1`).

Specify the names for the columns in the output.

```
statarray = grpstats(dsa,{'Sex','Smoker'},{'min','max'},...
          'DataVars','Weight','VarNames',{'Gender','Smoker',...
                    'GroupCount','LowestWeight','HighestWeight'})
```

```
statarray =
                Gender    Smoker    GroupCount    LowestWeight    HighestWeight
    Female_0    Female    false     40            111             147
    Female_1    Female    true      13            115             146
    Male_0      Male      false     26            158             194
    Male_1      Male      true      21            164             202
```

Summary Statistics for a Dataset Array Without Grouping

Load the sample data.

```
load('hospital')
```

The dataset array `hospital` has 100 observations and 7 variables.

Create a dataset array with only the variables `Age`, `Weight`, and `Smoker`.

```
dsa = hospital(:,{'Age','Weight','Smoker'});
```

The variables `Age` and `Weight` have numeric values, and `Smoker` has logical values.

Compute the mean, minimum, and maximum for the numeric and logical arrays, `Age`, `Weight`, and `Smoker`, with no grouping.

```
statarray = grpstats(dsa,[],{'mean','min','max'})
```

```
statarray =
           GroupCount    mean_Age    min_Age    max_Age    mean_Weight
    All    100           38.28       25         50         154


           min_Weight    max_Weight    mean_Smoker    min_Smoker    max_Smoker
    All    111           202           0.34           false         true
```

The observation name `All` indicates that all observations in `dsa` were used to compute the summary statistics.

Group Means for a Matrix Using One or More Grouping Variables

Load the sample data.

```
load('carsmall')
```

All variables are measured for 100 cars. `Origin` is the country of origin for each car (France, Germany, Italy, Japan, Sweden, or USA). `Cylinders` has three unique values, `4`, `6`, and `8`, indicating the number of cylinders in each car.

Calculate the mean acceleration, grouped by country of origin.

```
means = grpstats(Acceleration,Origin)
```

```
means = 6×1

   14.4377
   18.0500
   15.8867
   16.3778
   16.6000
   15.5000
```

`means` is a 6-by-1 vector of mean accelerations, where each value corresponds to a country of origin.

Calculate the mean acceleration, grouped by both country of origin and number of cylinders.

```
means = grpstats(Acceleration,{Origin,Cylinders})
```

```
means = 10×1

   17.0818
   16.5267
   11.6406
   18.0500
   15.9143
   15.5000
   16.3375
```

```
   16.7000
   16.6000
   15.5000
```

There are 18 possible combinations of grouping variable values because `Origin` has 6 unique values and `Cylinders` has 3 unique values. Only 10 of the possible combinations appear in the data, so `means` is a 10-by-1 vector of group means corresponding to the observed combinations of values.

Return the group names along with the mean acceleration for each group.

```
[means,grps] = grpstats(Acceleration,{Origin,Cylinders},{'mean','gname'})
```

```
means = 10×1

   17.0818
   16.5267
   11.6406
   18.0500
   15.9143
   15.5000
   16.3375
   16.7000
   16.6000
   15.5000


grps = 10x2 cell
    {'USA'    }    {'4'}
    {'USA'    }    {'6'}
    {'USA'    }    {'8'}
    {'France' }    {'4'}
    {'Japan'  }    {'4'}
    {'Japan'  }    {'6'}
    {'Germany'}    {'4'}
    {'Germany'}    {'6'}
    {'Sweden' }    {'4'}
    {'Italy'  }    {'4'}
```

The output `grps` shows the 10 observed combinations of grouping variable values. For example, the mean acceleration of 4-cylinder cars made in France is 18.05.

Multiple Summary Statistics for a Matrix Organized by Group

Load the sample data.

```
load carsmall
```

The variable `Acceleration` was measured for 100 cars. The variable `Origin` is the country of origin for each car (France, Germany, Italy, Japan, Sweden, or USA).

Return the minimum and maximum acceleration grouped by country of origin.

```
[grpMin,grpMax,grp] = grpstats(Acceleration,Origin,{'min','max','gname'})
```

```
grpMin = 6×1

    8.0000
   15.3000
   13.9000
   12.2000
   15.7000
   15.5000


grpMax = 6×1

   22.2000
   21.9000
   18.2000
   24.6000
   17.5000
   15.5000


grp = 6x1 cell
    {'USA'    }
    {'France' }
    {'Japan'  }
    {'Germany'}
    {'Sweden' }
    {'Italy'  }
```

The sample car with the lowest acceleration is made in the USA, and the sample car with the highest acceleration is made in Germany.

Plot Prediction Intervals for a New Observation in Each Group

Load the sample data.

```
load('carsmall')
```

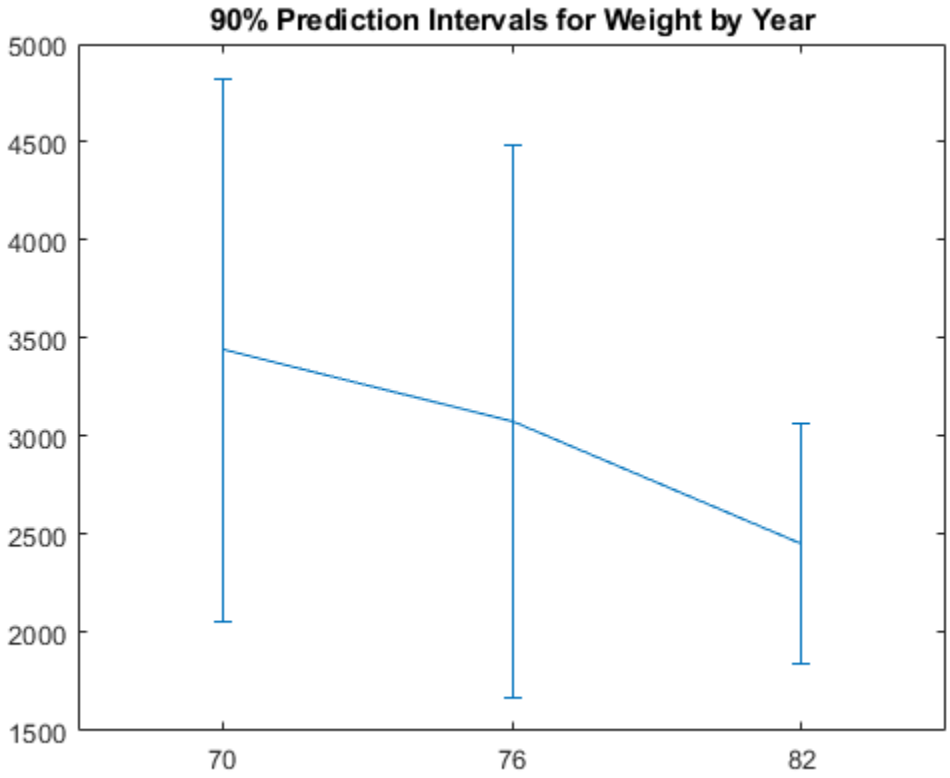
The variable `Weight` was measured for 100 cars. The variable `Model_Year` has three unique values, `70`, `76`, and `82`, which correspond to model years 1970, 1976, and 1982.

Calculate the mean weight and 90% prediction intervals for each model year.

```
[means,pred,grp] = grpstats(Weight,Model_Year,...
                      {'mean','predci','gname'},'Alpha',0.1);
```

Plot error bars showing the mean weight and 90% prediction intervals, grouped by model year. Label the horizontal axis with the group names.

```
ngrps = length(grp); % Number of groups
errorbar((1:ngrps)',means,pred(:,2)-means)
xlim([0.5 3.5])
set(gca,'xtick',1:ngrps,'xticklabel',grp)
title('90% Prediction Intervals for Weight by Year')
```



Plot Group Means and Confidence Intervals

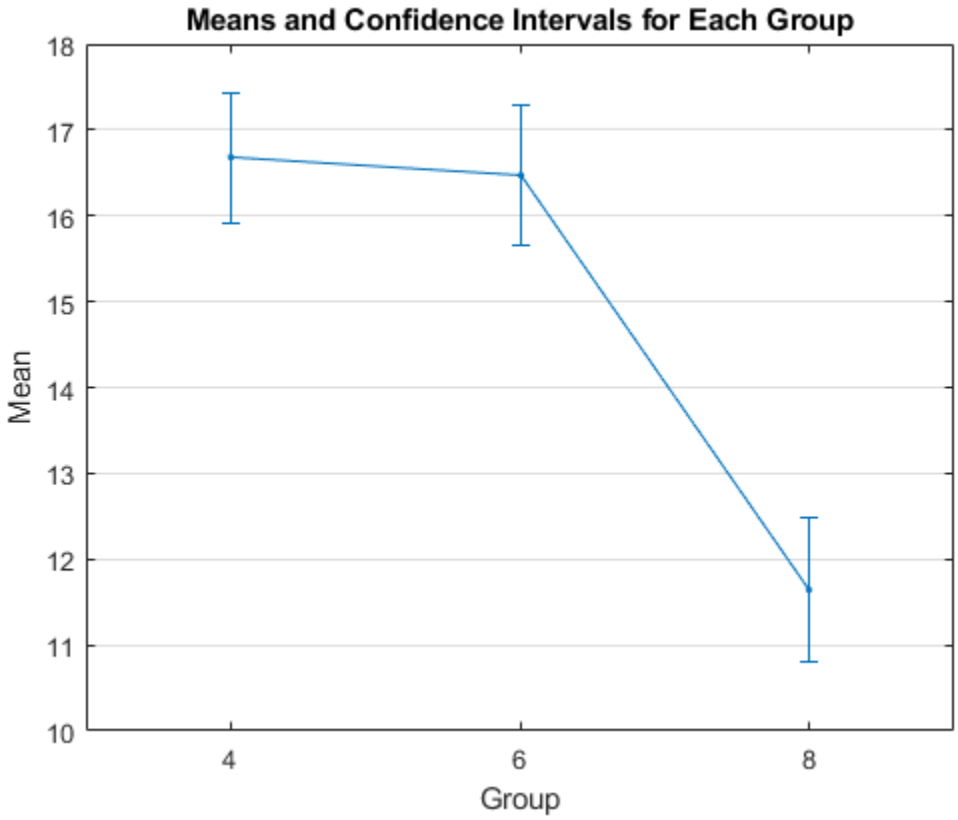
Load the sample data.

```
load('carsmall')
```

The variables `Acceleration` and `Weight` are the acceleration and weight values measured for 100 cars. The variable `Cylinders` is the number of cylinders in each car. The variable `Model_Year` has three unique values, `70`, `76`, and `82`, which correspond to model years 1970, 1976, and 1982.

Plot mean acceleration, grouped by `Cylinders`, with 95% confidence intervals.

```
grpstats(Acceleration,Cylinders,0.05)
```

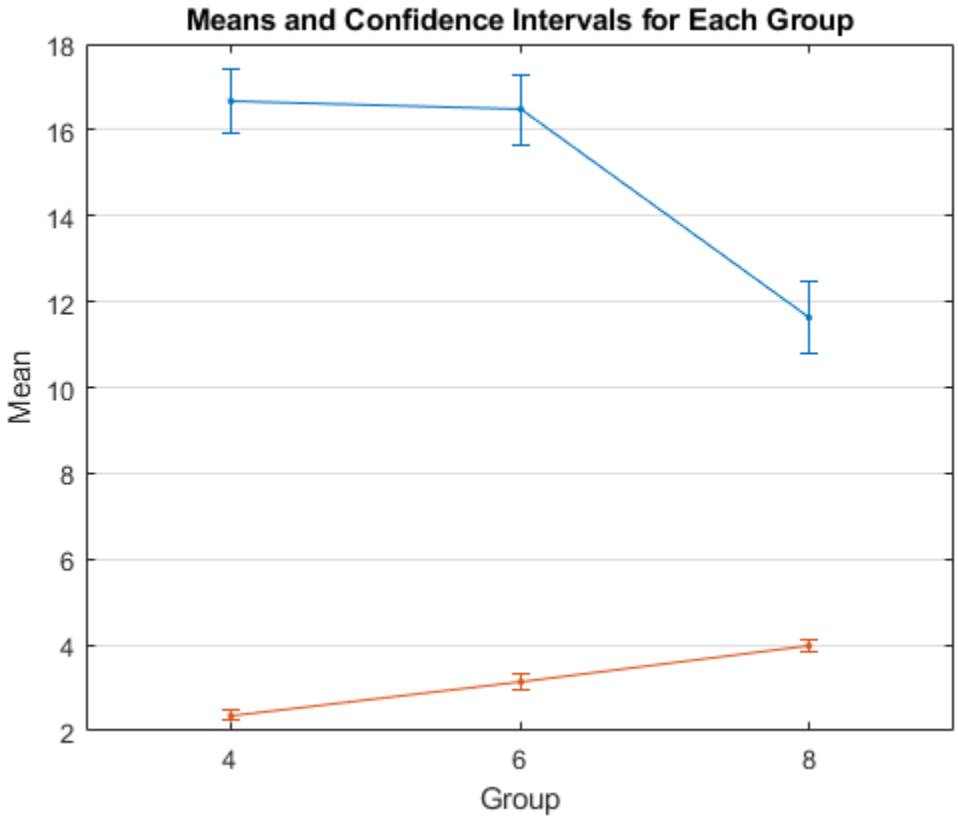



```
ans = 3×1

   16.6706
   16.4765
   11.6406
```

The mean acceleration for cars with 8 cylinders is significantly lower than for cars with 4 or 6 cylinders.

Plot mean acceleration and weight, grouped by `Cylinders`, and 95% confidence intervals. Scale the `Weight` values by 1000 so the means of `Weight` and `Acceleration` are the same order of magnitude.

```
grpstats([Acceleration,Weight/1000],Cylinders,0.05)
```

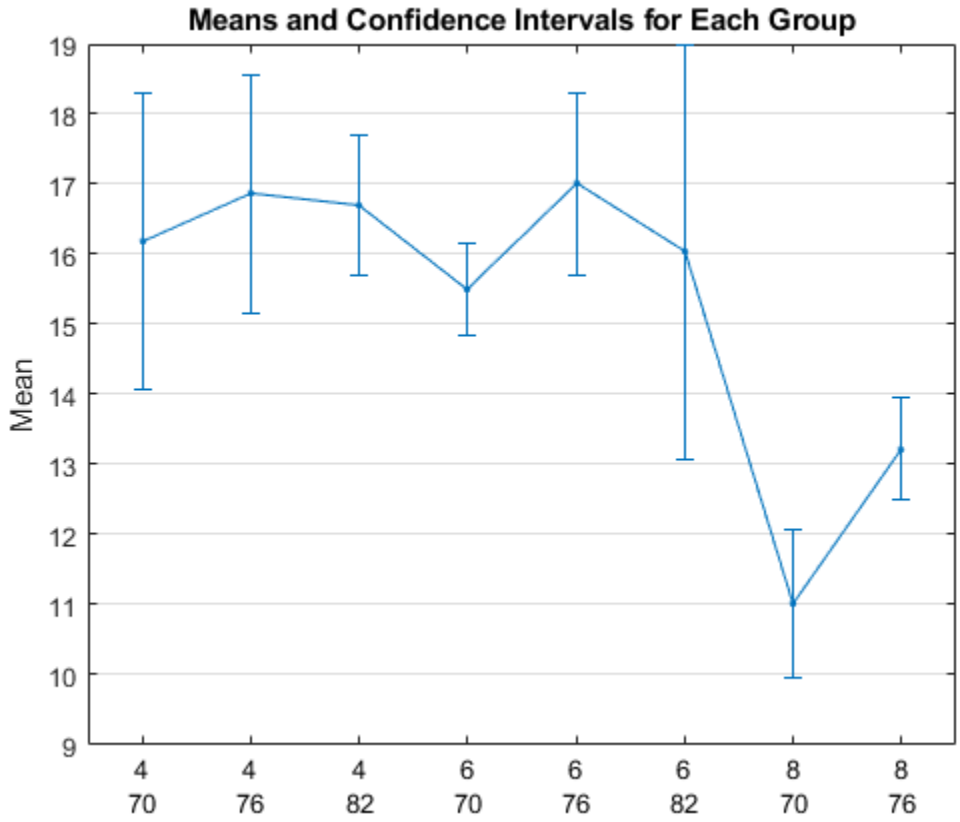



```
ans = 3×2

   16.6706    2.3726
   16.4765    3.1255
   11.6406    3.9703
```

The average weight of cars increases with the number of cylinders, and the average acceleration decreases with the number of cylinders.

Plot mean acceleration, grouped by both `Cylinders` and `Model_Year`. Specify 95% confidence intervals.

```
grpstats(Acceleration,{Cylinders,Model_Year},0.05)
```




```
ans = 8×1

   16.1875
   16.8667
   16.7036
   15.5000
   17.0000
   16.0333
   11.0217
   13.2222
```

There are nine possible combinations of grouping variable values because there are three unique values in `Cylinders` and three unique values in `Model_Year`. The plot does not show 8-cylinder cars with model year 1982 because the data did not include this combination.

The mean acceleration of 8-cylinder cars made in 1976 is significantly larger than the mean acceleration of 8-cylinder cars made in 1970.

Input Arguments

`tbl` — Input data

table | dataset array

Input data, specified as a table or dataset array. `tbl` must include at least one variable that is a grouping variable.

Summary statistics can only be calculated for variables that have a numeric or logical data type. If any variables in `tbl` (other than the grouping variables) are not numeric or logical arrays, then use the name-value pair argument `DataVars` to specify the names or column numbers of the numeric and logical variables for which to calculate summary statistics.

`groupvar` — Identifiers for the grouping variables

character vector | string array | cell array of character vectors | vector of positive integers | logical vector | `[]`

Identifiers for the grouping variables in the input data, `tbl`, specified as one of the following:

Character vector, string array, or cell array of character vectors	Names of the grouping variables
Positive integer or vector of positive integers	Variable numbers of the grouping variables
Vector of logical values with number of elements equal to the number of variables in `tbl`	Logical indicator with value `true` for grouping variables and `false` otherwise
`[]`	No groups (returns summary statistics for all data)

Any variable that is identified by `groupvar` as a grouping variable must have a valid grouping variable data type: categorical array, logical or numeric vector, datetime or duration vector, string array, or cell array of character vectors.

For example, consider an input table, `tbl`, with six variables. The fourth variable is named `Gender`. To be a valid grouping variable, the data type of `Gender` might be a string array, a cell array of character vectors, or a nominal array, with the unique values `Male` and `Female`. To specify the variable `Gender` as the grouping variable, you can use any of these syntaxes:

- `statarray = grpstats(tbl,'Gender')`
- `statarray = grpstats(tbl,4)`
- `statarray = grpstats(tbl,logical([0 0 0 1 0 0]))`

Data Types: `double` | `logical` | `char` | `string` | `cell`

`whichstats` — Type of summary statistics

character vector | string scalar | function handle | string array | cell array of character vectors or function handles

Type of summary statistics to compute, specified as one of the following values.

- Character vector or string scalar specifying the type of summary statistics, as described in this table.

Type	Description
`'mean'`	Mean
`'sem'`	Standard error of the mean
`'numel'`	Count, or number, of non-`NaN` elements
`'gname'`	Group name
`'std'`	Standard deviation

Type	Description
'var'	Variance
'min'	Minimum
'max'	Maximum
'range'	Range
'meanci'	95% confidence interval for the mean. You can specify different significance levels using the Alpha name-value pair argument.
'predci'	95% prediction interval for a new observation. You can specify different significance levels using the Alpha name-value pair argument.

- Function handle to specify any other type of summary statistics. You can use the handle to any function that accepts a column or matrix of data, and returns the same size output each time grpstats calls the function handle (even if the output for some groups is empty).
  - If the function accepts a column of data, then the function can return either a scalar value or an *nvals*-by-1 column vector for descriptive statistics of length *nvals* (for example, a confidence interval has length two). If the function accepts a matrix, the function must return either a 1-by-*ncols* row vector or an *nvals*-by-*ncols* matrix, where *ncols* is the number of columns in the input data matrix.
  - For functions that do not compute column-wise statistics, specify the computation direction while specifying the function. For example, to use the sum function, specify the function handle as @(x)sum(x,1) because sum computes column-wise statistics for matrices with two or more rows, but not for single-row matrices.
- String array or a cell array of character vectors or function handles to specify multiple types of summary statistics.

Example: stat1 = grpstats(X,group,'sem')

Example: stat1 = grpstats(X,group,@(x)sum(x,1))

Example: [stat1,stat2,stat3] = grpstats(X,group,{'mean','std',@skewness})

alpha — Significance level
scalar value in the range (0,1)

Significance level, specified as a scalar value in the range (0,1).

- When you specify 'meanci' or 'predci' in whichstats, you can use alpha to specify the significance level for the confidence or prediction intervals. If you specify alpha, then grpstats returns 100×(1 – alpha)% confidence or prediction intervals. If you do not specify alpha, then grpstats returns 95% intervals (alpha = 0.05).
- Use alpha with the syntax on page 33-2759 to plot group means and corresponding 100×(1 – alpha)% confidence intervals.

Data Types: double

X — Input data
vector | matrix

Input data, specified as a vector or a matrix. If `X` is a matrix, then `grpstats` returns summary statistics for each column of `X`.

Data Types: `double` | `single`

`group` — Grouping variable

categorical array | logical or numeric vector | datetime or duration vector | string array | cell array of character vectors | `[]`

Grouping variable, specified as a categorical array, logical or numeric vector, datetime or duration vector, string array, or cell array of character vectors. Each unique value in a grouping variable defines a group. `grpstats` groups data for summary statistics using the grouping variable values.

There must be a grouping variable value for each row of the input data `X`. Observations (rows) with the same value of the grouping variable are in the same group. Use `[]` to compute summary statistics for all data, without using groups.

For example, if `Gender` is a string array or cell array of character vectors with values `'Male'` and `'Female'`, you can use `Gender` as a grouping variable to summarize your data by gender.

You can also use more than one grouping variable to group data for summary statistics. In this case, specify a cell array of grouping variables.

For example, if `Smoker` is a logical vector with values `0` for nonsmokers and `1` for smokers, then specifying the cell array `{Gender,Smoker}` divides observations into four groups: Male Smoker, Male Nonsmoker, Female Smoker, and Female Nonsmoker. `grpstats` returns summary statistics only for the combinations of values that exist in the input grouping variables (not all possible combinations).

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical` | `datetime` | `duration`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'DataVars',[1,3,4],'Alpha',0.01` specifies that summary statistics be calculated for the 1st, 3rd, and 4th variables in a dataset array, with 99% confidence intervals.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level for confidence and prediction intervals, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

When you include `'meanci'` or `'predci'` in `whichstats`, you can use `Alpha` to specify the significance level for confidence or prediction intervals. If you specify the value α, then `grpstats` returns $100\times(1 - \alpha)$% confidence or prediction intervals.

If you do not specify a value for `Alpha`, then `grpstats` returns 95% intervals ($\alpha = 0.05$).

Example: `'Alpha',0.1`

Data Types: `double`

`DataVars` — Variable names or columns
string array | cell array of character vectors | vector of positive integers | logical vector

Variable names or columns indicating which variables in the input data `tbl` you want to compute summary statistics for, specified as the comma-separated pair consisting of `'DataVars'` and a string array, cell array of character vectors, vector of positive integers, or logical vector. Use a character vector or string scalar to specify a variable name, a positive integer to specify a variable column number, or logical values to indicate which variables to include (`true` if you want to compute summary statistics, `false` otherwise).

You must specify `DataVars` if there are any variables in `tbl` (other than the grouping variables specified in `groupvar`) that are not numeric or logical arrays. Summary statistics can only be calculated for variables that have a numeric or logical data type.

Example: `'DataVars',{'Height','Weight'}`

Data Types: `double` | `string` | `cell` | `char`

`VarNames` — Variable names for output
string array | cell array of character vectors

Variable names for the output `statarray`, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors. By default, `grpstats` constructs output variable names by appending a prefix to the variable names from the input data `tbl`. This prefix corresponds to the summary statistic name.

Example: `'VarNames',{'Gender','GroupCount','MaleMean','FemaleMean'}`

Data Types: `string` | `cell`

Output Arguments

`statarray` — Group summary statistics
table | dataset array

Group summary statistics, returned as a table or a dataset array. If `tbl` is a table, `grpstats` returns `statarray` as a table. If `tbl` is a dataset array, `grpstats` returns `statarray` as a dataset array.

`statarray` contains summary statistic values for the groups of data in `tbl` determined by the levels of the grouping variables specified by `groupvar`. There is a row in `statarray` for each observed value or combination of values in the variables specified by `groupvar`. The output `statarray` contains:

- All grouping variables specified by `groupvar`.
- The variable `GroupCount`, containing the number of observations in each group.
- Group summary statistic values for all variables in `tbl` (other than those specified by `groupvar`), or for only the variables specified using `DataVars`.

The total number of variables in `statarray` is *ngroupvars* + 1 + *ndatavars*×*nstats*, where *ngroupvars* is the number of variables in `groupvar`, *ndatavars* is the number of variables for which summary statistics are computed, and *nstats* is the number of summary statistic types specified in `whichstats`.

`grpstats` assigns default names to the variables in `statarray`, unless you specify variable names using the name-value pair argument `VarNames`.

means — Group means
column vector | array

Group means for the groups of data in the vector or matrix X determined by the levels of group, returned as an *ngroups*-by-*ncols* array. Here, *ngroups* is the number of unique values in the grouping variable, and *ncols* is the number of columns in X. If X is a vector, then means is a column vector.

stats1,...,statsN — Group summary statistics
column vectors | arrays

Group summary statistics for the groups of data in the vector or matrix X determined by the levels of group, returned as *ngroups*-by-*ncols* arrays. Here, *ngroups* is the number of unique values in the grouping variable, and *ncols* is the number of columns in X. You must specify an output argument for each type of summary statistic specified in whichstats.

If a summary statistic type in whichstats returns a value of length *nvals* (for example, a confidence interval is a descriptive statistic of length two), then the corresponding output argument is an *ngroups*-by-*ncols*-by-*nvals* array.

Algorithms

- grpstats treats NaNs as missing values, and removes them from the input data before calculating summary statistics.
- grpstats ignores empty group names.

Alternative Functionality

MATLAB includes the function groupsummary, which also returns group summaries and is recommended when you are working with a table.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- If the input data is a tall array, then all grouping variables must also be tall and have the same number of rows as the data.
- The whichstats option cannot be specified as a function handle. In addition to the current built-in options, whichstats can also be:
  - 'Count' — Number of non-NaNs.
  - 'NNZ' — Number of nonzeros and non-NaNs.
  - 'Kurtosis' — Compute kurtosis.
  - 'Skewness' — Compute skewness.
  - 'all-stats' — Compute all summary statistics.
- Group order is not guaranteed to be the same as the in-memory grpstats computation. Specify 'gname' as the whichstats option to return the order of rows of the summary statistics. For

example `[means,grpname] = grpstats(x,bins,{'mean','gname'})` returns the means of groups in `x` in the same order that the groups appear in `grpname`.

- Summary statistics for *nonnumeric* variables return NaNs.
- `grpstats` always operates on the first dimension.
- If the input is a tall table, then the output is also a tall table. However, rather than including row names, the output tall table contains an extra variable `GroupLabel` that contains the same information.

For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`dataset` | `groupsummary` | `table`

Topics

“Summary Statistics Grouped by Category” on page 2-32
“Test Differences Between Category Means” on page 2-25
“Plot Data Grouped by Category” on page 2-21
“Calculations on Dataset Arrays” on page 2-92
“Dataset Arrays” on page 2-112
“Grouping Variables” on page 2-45
“Nominal and Ordinal Arrays” on page 2-36

Introduced before R2006a

grpstats

Class: `RepeatedMeasuresModel`

Compute descriptive statistics of repeated measures data by group

Syntax

```
statstbl = grpstats(rm,g)
statstbl = grpstats(rm,g,stats)
```

Description

`statstbl = grpstats(rm,g)` returns the count, mean, and variance for the data used to fit the repeated measures model `rm`, grouped by the factors, `g`.

`statstbl = grpstats(rm,g,stats)` returns the statistics specified by `stats` for the data used to fit the repeated measures model `rm`, grouped by the factors, `g`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`g` — Name of grouping factor or factors

character vector | string array | cell array of character vectors

Name of grouping factor or factors, specified as a character vector, string array, or cell array of character vectors.

Example: `'Drug'`

Example: `{'Drug','Sex'}`

Data Types: `char` | `string` | `cell`

`stats` — Statistics to compute

character vector | string scalar | function handle | string array | cell array of multiple character vectors and function handles

Statistics to compute, specified as one of the following:

- Character vector or string scalar specifying the name of the statistics to compute. Names can be one of the following.

Name	Description
`'mean'`	Mean

Name	Description
'sem'	Standard error of the mean
'numel'	Count or number of elements
'gname'	Group name
'std'	Standard deviation
'var'	Variance
'min'	Minimum
'max'	Maximum
'range'	Maximum minus minimum
'meanci'	95% confidence interval for the mean
'predci'	95% prediction interval for a new observation

- Function handle — The function you specify must accept a vector of response values for a single group, and compute descriptive statistics for it. A function should typically return a value that has one row. A function must return the same size output each time `grpstats` calls it, even if the input for some groups is empty.
- A string array or cell array of character vectors and function handles.

Example: `@median`

Example: `@skewness`

Example: `'gname'`

Example: `{'gname','range','predci'}`

Output Arguments

`statstbl` — Statistics values for each group

table

Statistics values for each group, returned as a table.

Examples

Compute Group Statistics

Load the sample data.

```
load fisheriris
```

The column vector, `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Compute group counts, mean, and standard deviation with respect to species.

```
grpstats(rm,'species')
```

```
ans=3×4 table
       species        GroupCount     mean      std
    ______________    __________    ______    ______

    {'setosa'    }       200        2.5355    1.8483
    {'versicolor'}       200         3.573    1.7624
    {'virginica' }       200         4.285    1.9154
```

Now, compute the range of data and 95% confidence intervals for the group means for the factor species. Also display the group name.

```
grpstats(rm,'species',{'gname','range','predci'})
```

```
ans=3×5 table
       species            gname          GroupCount    range           predci
    ______________    ______________    __________    _____    ____________________

    {'setosa'    }    {'setosa'    }       200         5.7      -1.1185      6.1895
    {'versicolor'}    {'versicolor'}       200           6     0.088976       7.057
    {'virginica' }    {'virginica' }       200         6.5       0.4985      8.0715
```

Statistics for Data Grouped by Two Factors

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Compute group counts, mean, standard deviation, skewness, and kurtosis of data grouped by the factors `Group` and `Gender`.

```
GS = grpstats(rm,{'Group','Gender'},{'mean','std',@skewness,@kurtosis})
```

```
GS=6×7 table
    Group    Gender    GroupCount     mean       std      skewness    kurtosis
    _____    ______    __________    _______    ______    ________    ________
```

```
A      Female        40         16.554    21.498     0.35324     3.7807
A      Male          40         9.8335    20.602    -0.38722     2.7834
B      Female        40         11.261    25.779    -0.49177     4.1484
B      Male          40         3.6078    24.646     0.55447     2.7966
C      Female        40        -11.335    27.186      1.7499     6.1429
C      Male          40        -14.028    31.984      1.7362      5.141
```

Tips

- `grpstats` computes results separately for each group. The results do not depend on the fitted repeated measures model. It computes the results on all available data, without omitting entire rows that contain `NaN`s.

See Also

`fitrm` | `plot`

gscatter

Scatter plot by group

Syntax

```
gscatter(x,y,g)
gscatter(x,y,g,clr,sym,siz)
gscatter(x,y,g,clr,sym,siz,doleg)
gscatter(x,y,g,clr,sym,siz,doleg,xnam,ynam)

gscatter(ax,___)

h = gscatter(___)
```

Description

`gscatter(x,y,g)` creates a scatter plot of `x` and `y`, grouped by `g`. The inputs `x` and `y` are vectors of the same size.

`gscatter(x,y,g,clr,sym,siz)` specifies the marker color `clr`, symbol `sym`, and size `siz` for each group.

`gscatter(x,y,g,clr,sym,siz,doleg)` controls whether a legend is displayed on the graph. `gscatter` creates a legend by default.

`gscatter(x,y,g,clr,sym,siz,doleg,xnam,ynam)` specifies the names to use for the x-axis and y-axis labels. If you do not provide `xnam` and `ynam`, and the `x` and `y` inputs are variables with names, then `gscatter` labels the axes with the variable names.

`gscatter(ax,___)` uses the plot axes specified by the axes object `ax`. Specify `ax` as the first input argument followed by any of the input argument combinations in the previous syntaxes.

`h = gscatter(___)` returns graphics handles corresponding to the groups in `g`.

You can pass in `[]` for `clr`, `sym`, `siz`, and `doleg` to use their default values.

Examples

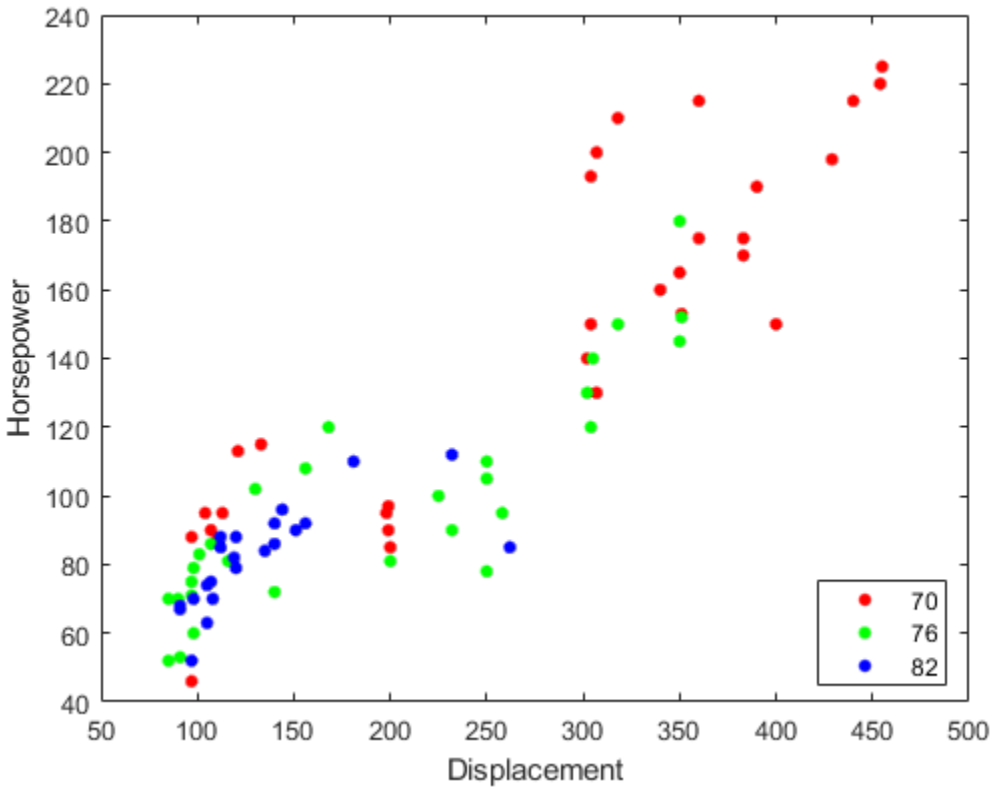
Scatter Plot with Default Settings

Load the `carsmall` data set.

```
load carsmall
```

Plot the `Displacement` values on the x-axis and the `Horsepower` values on the y-axis. `gscatter` uses the variable names as the default labels for the axes. Group the data points by `Model_Year`.

```
gscatter(Displacement,Horsepower,Model_Year)
```



Scatter Plot with One Grouping Variable

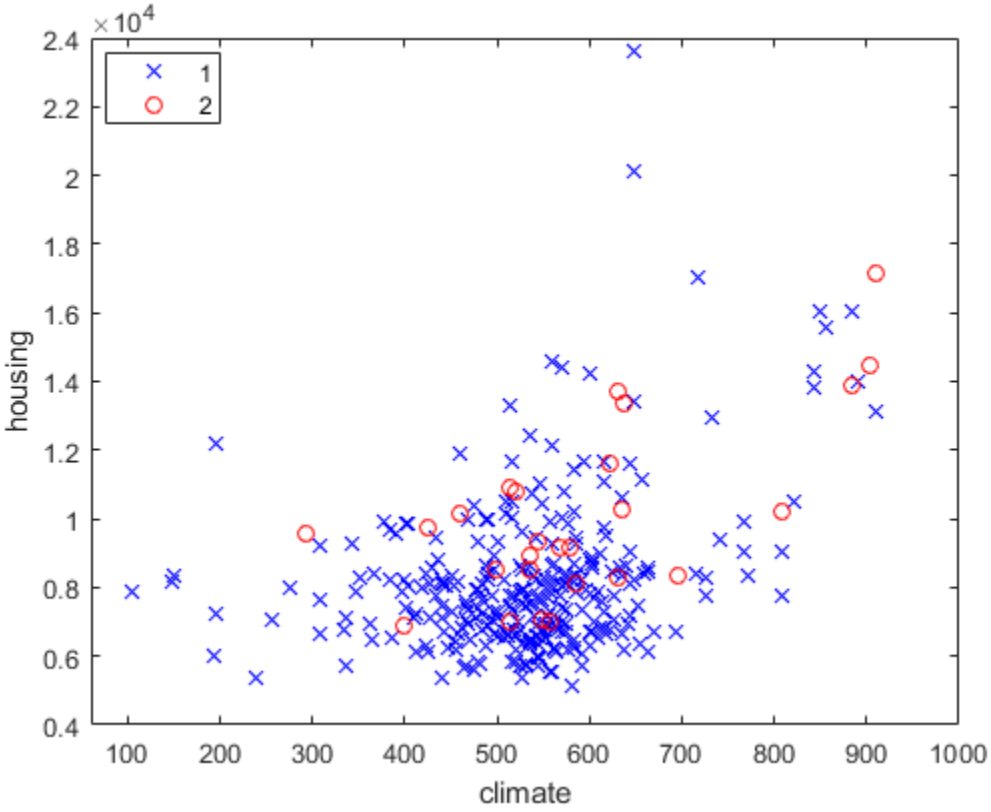
Load the `discrim` data set.

```
load discrim
```

The data set contains ratings of cities according to nine factors such as climate, housing, education, and health. The matrix `ratings` contains the ratings information.

Plot the relationship between the ratings for climate (first column) and housing (second column) grouped by city size in the matrix `group`. Choose different colors and plotting symbols for each group.

```
gscatter(ratings(:,1),ratings(:,2),group,'br','xo')
xlabel('climate')
ylabel('housing')
```



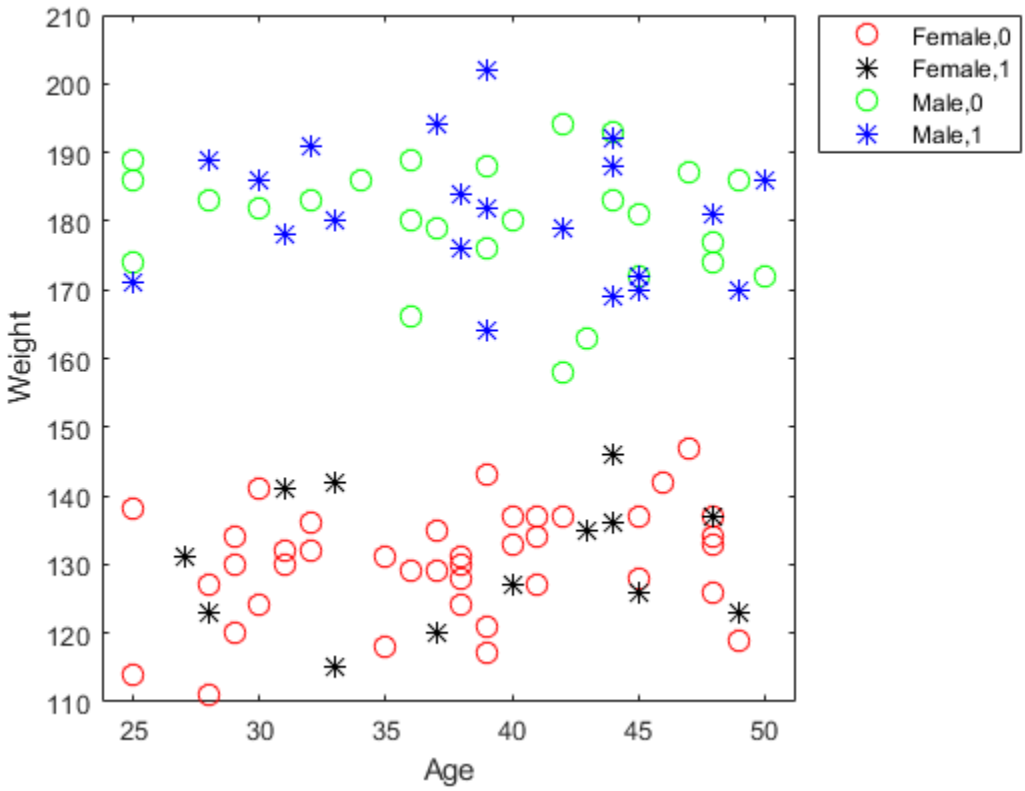
Scatter Plot with Multiple Grouping Variables

Load the `hospital` data set.

```
load hospital
```

Plot the ages and weights of the hospital patients. Group the patients according to their gender and smoker status. Use the `o` symbol to represent nonsmokers and the * symbol to represent smokers.

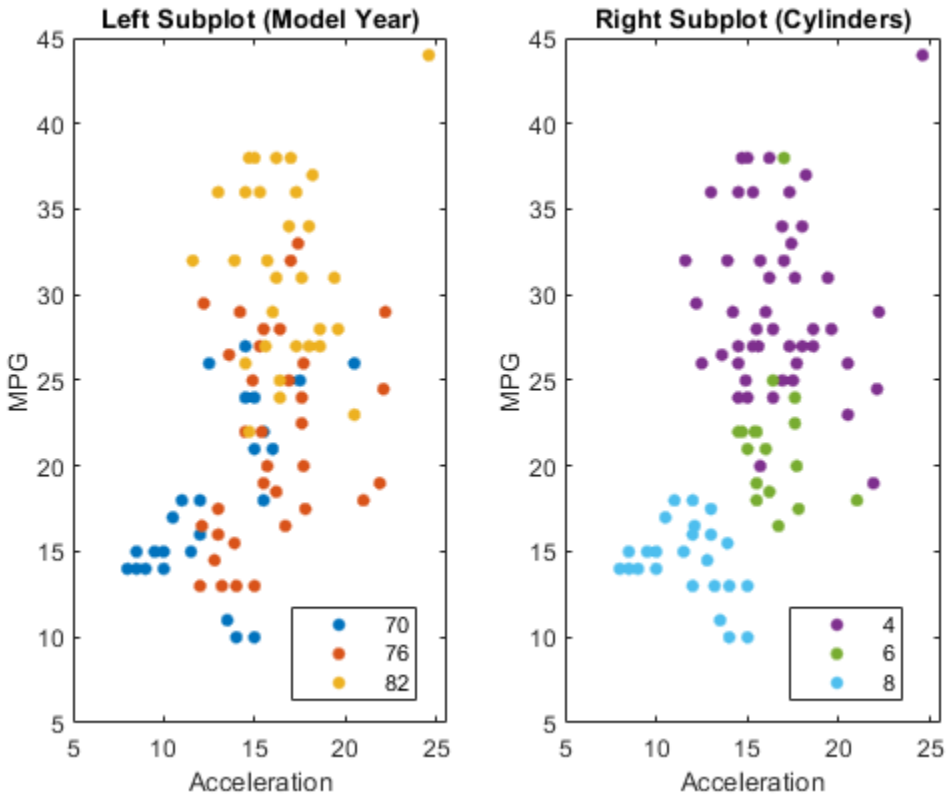
```
x = hospital.Age;
y = hospital.Weight;
g = {hospital.Sex,hospital.Smoker};
gscatter(x,y,g,'rkgb','o*',8,'on','Age','Weight')
legend('Location','northeastoutside')
```



Specify Axes for Scatter Plot

Load the `carsmall` data set. Create a figure with two subplots and return the `axes` objects as `ax1` and `ax2`. Create a scatter plot in each set of axes by referring to the corresponding `Axes` object. In the left subplot, group the data using the `Model_Year` variable. In the right subplot, group the data using the `Cylinders` variable. Add a title to each plot by passing the corresponding `Axes` object to the `title` function.

```
load carsmall
color = lines(6); % Generate color values

ax1 = subplot(1,2,1); % Left subplot
gscatter(ax1,Acceleration,MPG,Model_Year,color(1:3,:))
title(ax1,'Left Subplot (Model Year)')

ax2 = subplot(1,2,2); % Right subplot
gscatter(ax2,Acceleration,MPG,Cylinders,color(4:6,:))
title(ax2,'Right Subplot (Cylinders)')
```



Create and Modify Scatter Plot

Load the `carbig` data set.

```
load carbig
```

Create a scatter plot comparing `Acceleration` to `MPG`. Group data points based on `Origin`.

```
h = gscatter(Acceleration,MPG,Origin)
```

```
h =
  7x1 Line array:

  Line    (USA)
  Line    (France)
  Line    (Japan)
  Line    (Germany)
  Line    (Sweden)
  Line    (Italy)
  Line    (England)
```

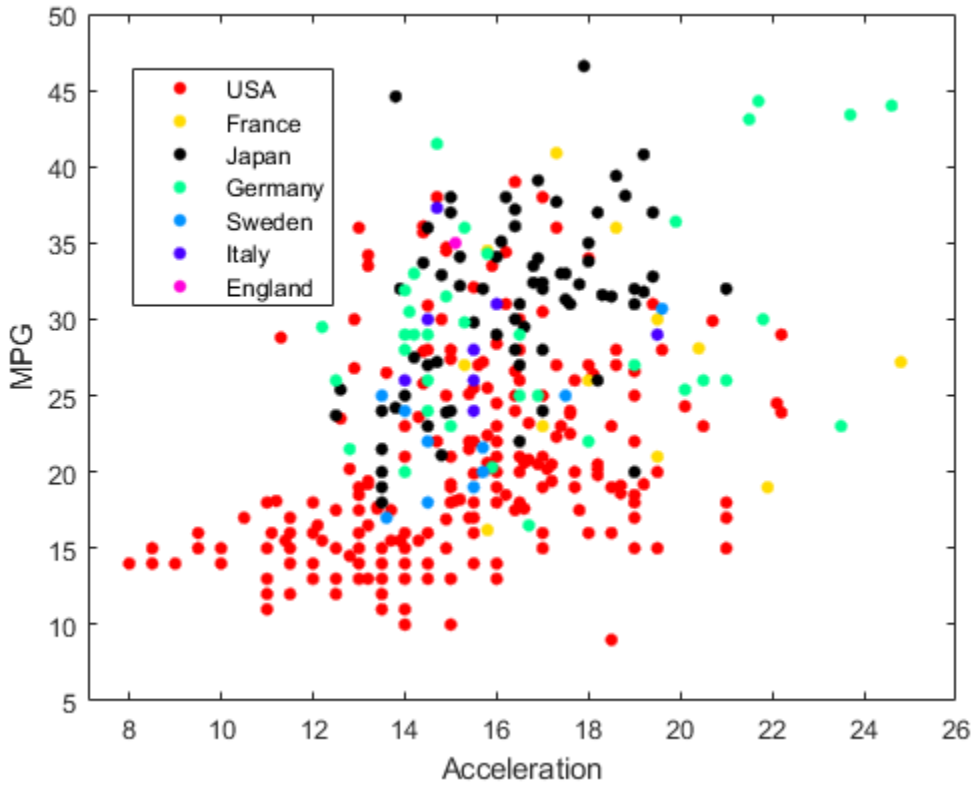
Display the `Line` object corresponding to the group labeled `(Japan)`.

```
jgroup = h(3)
```

```
jgroup =
  Line (Japan) with properties:

              Color: [0.2857 1 0]
          LineStyle: 'none'
          LineWidth: 0.5000
             Marker: '.'
         MarkerSize: 15
    MarkerFaceColor: 'none'
              XData: [1x79 double]
              YData: [1x79 double]
              ZData: [1x0 double]

  Show all properties
```

Change the marker color for the `Japan` group to black.

```
jgroup.Color = 'k';
```



Input Arguments

x — x-axis values

numeric vector

x-axis values, specified as a numeric vector. `x` must have the same size as `y`.

Data Types: `single` | `double`

y — y-axis values
numeric vector

y-axis values, specified as a numeric vector. `y` must have the same size as `x`.

Data Types: `single` | `double`

g — Grouping variable
categorical vector | logical vector | numeric vector | character array | string array | cell array of character vectors | cell array

Grouping variable, specified as a categorical vector, logical vector, numeric vector, character array, string array, or cell array of character vectors. Alternatively, `g` can be a cell array containing several grouping variables (such as `{g1 g2 g3}`), in which case observations are in the same group if they have common values of all grouping variables. Points in the same group appear on the scatter plot with the same marker color, symbol, and size.

The number of rows in `g` must be equal to the length of `x`.

Example: `species`

Example: `{Cylinders,Origin}`

Data Types: `categorical` | `logical` | `single` | `double` | `char` | `string` | `cell`

`clr` — Marker colors
character vector or string scalar of colors | matrix of RGB triplet values

Marker colors, specified as either a character vector or string scalar of colors recognized by the `plot` function or a matrix of RGB triplet values. Each RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color, respectively. Each intensity must be in the range [0,1].

This table lists the available color characters and their equivalent RGB triplet values.

Long Name	Short Name	RGB Triplet
Yellow	`'y'`	`[1 1 0]`
Magenta	`'m'`	`[1 0 1]`
Cyan	`'c'`	`[0 1 1]`
Red	`'r'`	`[1 0 0]`
Green	`'g'`	`[0 1 0]`
Blue	`'b'`	`[0 0 1]`
White	`'w'`	`[1 1 1]`
Black	`'k'`	`[0 0 0]`

If you do not specify enough values for all groups, then `gscatter` cycles through the specified values as needed.

Example: `'rgb'`

Example: `[0 0 1; 0 0 0]`

Data Types: `char` | `string` | `single` | `double`

`sym` — Marker symbols
`'.'` (default) | character vector or string scalar of symbols

Marker symbols, specified as a character vector or string scalar of symbols recognized by the `plot` function. This table lists the available marker symbols.

Value	Description
`'o'`	Circle
`'+'`	Plus sign
`'*'`	Asterisk
`'.'`	Point
`'x'`	Cross
`'s'`	Square
`'d'`	Diamond
`'^'`	Upward-pointing triangle
`'v'`	Downward-pointing triangle
`'>'`	Right-pointing triangle
`'<'`	Left-pointing triangle
`'p'`	Five-pointed star (pentagram)
`'h'`	Six-pointed star (hexagram)
`'none'`	No markers

If you do not specify enough values for all groups, then `gscatter` cycles through the specified values as needed.

Example: `'o+*v'`

Data Types: `char` | `string`

`siz` — Marker sizes
positive numeric vector

Marker sizes, specified as a positive numeric vector in points. The default value is determined by the number of observations. If you do not specify enough values for all groups, then `gscatter` cycles through the specified values as needed.

Example: `[6 12]`

Data Types: `single` | `double`

`doleg` — Option to include legend
`'on'` (default) | `'off'`

Option to include a legend, specified as either `'on'` or `'off'`. By default, the legend is displayed on the graph.

`xnam` — x-axis label
`x` variable name (default) | character vector | string scalar

x-axis label, specified as a character vector or string scalar.

Data Types: `char` | `string`

`ynam` — y-axis label
y variable name (default) | character vector | string scalar

y-axis label, specified as a character vector or string scalar.

Data Types: `char` | `string`

`ax` — Axes for plot
`Axes` object | `UIAxes` object

Axes for the plot, specified as an `Axes` or `UIAxes` object. If you do not specify `ax`, then `gscatter` creates the plot using the current axes. For more information on creating an axes object, see `axes` and `uiaxes`.

Output Arguments

`h` — Graphics handles
array of `Line` objects

Graphics handles, returned as an array of `Line` objects. Each `Line` object corresponds to one of the groups in `g`. You can use dot notation to query and set properties of the line objects. For a list of `Line` object properties, see Chart Line.

See Also

`gplotmatrix` | `grpstats` | `scatter`

Topics
“Create Scatter Plots Using Grouped Data” on page 4-2
“MANOVA” on page 9-49
“Grouping Variables” on page 2-45

Introduced before R2006a

gt

Class: `qrandstream`

Greater than relation for handles

Syntax

```
h1 > h2
```

Description

`h1 > h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise `>` result.

If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = gt(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`eq` | `ge` | `le` | `lt` | `ne` | `qrandstream`

haltonset

Halton quasirandom point set

Description

`haltonset` is a quasirandom point set object that produces points from the Halton sequence. The Halton sequence uses different prime bases in each dimension to fill space in a highly uniform manner.

Creation

Syntax

```
p = haltonset(d)
p = haltonset(d,Name,Value)
```

Description

`p = haltonset(d)` constructs a `d`-dimensional point set `p`, which is a `haltonset` object with default property settings. The input argument `d` corresponds to the `Dimensions` property of `p`.

`p = haltonset(d,Name,Value)` sets properties on page 33-2789 of `p` using one or more name-value pair arguments. Enclose each property name in quotes. For example, `haltonset(5,'Leap',2)` creates a five-dimensional point set from the first point, fourth point, seventh point, tenth point, and so on.

The returned object `p` encapsulates properties of a Halton quasirandom sequence. The point set is finite, with a length determined by the `Skip` and `Leap` properties and by limits on the size of the point set indices (maximum value of 2^{53}). Values of the point set are generated whenever you access `p` using `net` or parenthesis indexing. Values are not stored within `p`.

Properties

`Dimensions` — Number of dimensions

positive integer scalar

This property is read-only.

Number of dimensions of the points in the point set, specified as a positive integer scalar. For example, each point in the point set `p` with `p.Dimensions = 5` has five values.

Use the `d` input argument to specify the number of dimensions when you create a point set using the `haltonset` function.

Leap — Interval between points

`0` (default) | positive integer scalar

Interval between points in the sequence, specified as a positive integer scalar. In other words, the `Leap` property of a point set specifies the number of points in the sequence to leap over and omit for every point taken. The default `Leap` value is 0, which corresponds to taking every point from the sequence.

Leaping is a technique used to improve the quality of a point set. However, you must choose the `Leap` values with care. Many `Leap` values create sequences that fail to touch on large sub-hyper-rectangles of the unit hypercube and, therefore, fail to be a uniform quasirandom point set. For more information, see [1].

One rule for choosing `Leap` values for Halton sets is to set the value to `(n–1)`, where `n` is a prime number that has not been used to generate one of the dimensions. For example, for a `d`-dimensional point set, specify the `(d+1)`th or greater prime number for `n`.

Example: `p = haltonset(2,'Leap',4);` (where `d = 2` and `n = 5`)

Example: `p.Leap = 100;`

`ScrambleMethod` — Settings that control scrambling
`0x0` structure (default) | structure with `Type` and `Options` fields

Settings that control the scrambling of the sequence, specified as a structure with these fields:

- `Type` — A character vector containing the name of the scramble
- `Options` — A cell array of parameter values for the scramble

Use the `scramble` object function to set scrambles. For a list of valid scramble types, see the `type` input argument of `scramble`. An error occurs if you set an invalid scramble type for a given point set.

The `ScrambleMethod` property also accepts an empty matrix as a value. The software then clears all scrambling and sets the property to contain a `0x0` structure.

`Skip` — Number of initial points in sequence to omit
`0` (default) | positive integer scalar

Number of initial points in the sequence to omit from the point set, specified as a positive integer scalar.

Initial points of a sequence sometimes exhibit undesirable properties. For example, the first point is often `(0,0,0,...)`, which can cause the sequence to be unbalanced because the counterpart of the point, `(1,1,1,...)`, never appears. Also, initial points often exhibit correlations among different dimensions, and these correlations disappear later in the sequence.

Example: `p = haltonset(__,'Skip',2e3);`

Example: `p.Skip = 1e3;`

`Type` — Sequence type
`'Halton'` (default)

This property is read-only.

Sequence type on which the quasirandom point set `p` is based, specified as `'Halton'`.

Object Functions

net	Generate quasirandom point set
scramble	Scramble quasirandom point set

You can also use the following MATLAB functions with a `haltonset` object. The software treats the point set object like a matrix of multidimensional points.

length	Length of largest array dimension
size	Array size

Examples

Create Halton Point Set

Generate a three-dimensional Halton point set, skip the first 1000 values, and then retain every 101st point.

```
p = haltonset(3,'Skip',1e3,'Leap',1e2)
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
               Skip : 1000
               Leap : 100
    ScrambleMethod : none
```

Apply reverse-radix scrambling by using `scramble`.

```
p = scramble(p,'RR2')
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
               Skip : 1000
               Leap : 100
    ScrambleMethod : RR2
```

Generate the first four points by using `net`.

```
X0 = net(p,4)
```

```
X0 = 4×3

    0.0928    0.6950    0.0029
    0.6958    0.2958    0.8269
    0.3013    0.6497    0.4141
    0.9087    0.7883    0.2166
```

Generate every third point, up to the eleventh point, by using parenthesis indexing.

```
X = p(1:3:11,:)
```

```
X = 4×3

    0.0928    0.6950    0.0029
    0.9087    0.7883    0.2166
    0.3843    0.9840    0.9878
    0.6831    0.7357    0.7923
```

Tips

- The `Skip` and `Leap` properties are useful for parallel applications. For example, if you have a Parallel Computing Toolbox license, you can partition a sequence of points across N different workers by using the function `labindex`. On each nth worker, set the `Skip` property of the point set to n – 1 and the `Leap` property to N – 1. The following code shows how to partition a sequence across three workers.

```
Nworkers = 3;
p = haltonset(10,'Leap',Nworkers-1);
spmd(Nworkers)
    p.Skip = labindex - 1;

    % Compute something using points 1,4,7...
    % or points 2,5,8... or points 3,6,9...
end
```

Algorithms

Halton Sequence Generation

Consider a default `haltonset` object `p` that contains `d`-dimensional points. Each `p(i,:)` is a point in a Halton sequence. The jth coordinate of the point, `p(i,j)`, is equal to

$$\sum_k a_{ij}(k) b_j^{-k-1}.$$

- b_j is the jth prime.
- The $a_{ij}(k)$ coefficients are nonnegative integers less than b_j such that

$$i - 1 = \sum_{k=0} a_{ij}(k) b_j^k.$$

  In other words, the $a_{ij}(k)$ values are the base b_j digits of the integer i – 1.

For more information, see [1].

References

[1] Kocis, L., and W. J. Whiten. “Computational Investigations of Low-Discrepancy Sequences.” *ACM Transactions on Mathematical Software*. Vol. 23, No. 2, 1997, pp. 266–294.

See Also

`net` | `scramble` | `sobolset`

Topics
“Generating Quasi-Random Numbers” on page 7-12

Introduced in R2008a

harmmean

Harmonic mean

Syntax

```
m = harmmean(X)
m = harmmean(X,'all')
m = harmmean(X,dim)
m = harmmean(X,vecdim)
m = harmmean( ___ ,nanflag)
```

Description

`m = harmmean(X)` calculates the harmonic mean on page 33-2798 of a sample. For vectors, `harmmean(X)` is the harmonic mean of the elements in `X`. For matrices, `harmmean(X)` is a row vector containing the harmonic means of each column. For *N*-dimensional arrays, `harmmean` operates along the first nonsingleton dimension of `X`.

`m = harmmean(X,'all')` returns the harmonic mean of all elements of `X`.

`m = harmmean(X,dim)` takes the harmonic mean along the operating dimension `dim` of `X`.

`m = harmmean(X,vecdim)` returns the harmonic mean over the dimensions specified in the vector `vecdim`. Each element of `vecdim` represents a dimension of the input array `X`. The output `m` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `m`. For example, if `X` is a 2-by-3-by-4 array, then `harmmean(X,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the harmonic mean of the elements on the corresponding page of `X`.

`m = harmmean( ___ ,nanflag)` specifies whether to exclude `NaN` values from the calculation, using any of the input argument combinations in previous syntaxes. By default, `harmmean` includes `NaN` values in the calculation (`nanflag` has the value `'includenan'`). To exclude `NaN` values, set the value of `nanflag` to `'omitnan'`.

Examples

Compare Harmonic and Arithmetic Mean

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a matrix of exponential random numbers with 5 rows and 4 columns.

```
X = exprnd(1,5,4)
```

```
X = 5×4

    0.2049    2.3275    1.8476    1.9527
    0.0989    1.2783    0.0298    0.8633
    2.0637    0.6035    0.0438    0.0880
```

```
    0.0906    0.0434    0.7228    0.2329
    0.4583    0.0357    0.2228    0.0414
```

Compute the harmonic and arithmetic means of the columns of X.

```
harmonic = harmmean(X)
```

```
harmonic = 1×4

    0.1743    0.0928    0.0797    0.1205
```

```
arithmetic = mean(X)
```

```
arithmetic = 1×4

    0.5833    0.8577    0.5734    0.6357
```

The arithmetic mean is greater than the harmonic mean for all the columns of X.

Harmonic Mean of All Values

Find the harmonic mean of all the values in an array.

Create a 3-by-5-by-2 array X.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the harmonic mean of the elements of X.

```
m = harmmean(X,'all')
```

```
m = 7.5094
```

Harmonic Mean Along Specified Dimensions

Find the harmonic mean along different operating dimensions and vectors of dimensions for a multidimensional array.

Create a 3-by-5-by-2 array X.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the harmonic mean of X along the default dimension.

```
hmean1 = harmmean(X)
```

```
hmean1 =
hmean1(:,:,1) =

    1.6364    4.8649    7.9162   10.9392   13.9523


hmean1(:,:,2) =

   16.9607   19.9666   22.9710   25.9743   28.9770
```

By default, harmmean operates along the first dimension of X whose size does not equal 1. In this case, this dimension is the first dimension of X. Therefore, hmean1 is a 1-by-5-by-2 array.

Find the harmonic mean of X along the second dimension.

```
hmean2 = harmmean(X,2)
```

```
hmean2 =
hmean2(:,:,1) =

    3.1852
    5.0641
    6.5693


hmean2(:,:,2) =

   21.1595
   22.1979
```

```
   23.2329
```

`hmean2` is a 3-by-1-by-2 array.

Find the harmonic mean of `X` along the third dimension.

```
hmean3 = harmmean(X,3)
```

```
hmean3 = 3×5

    1.8824    6.6087   10.6207   14.2857   17.7561
    3.5789    8.0000   11.8710   15.4595   18.8837
    5.1429    9.3333   13.0909   16.6154   20.0000
```

`hmean3` is a 3-by-5 array.

Find the harmonic mean of each page of `X` by specifying the first and second dimensions using the `vecdim` input argument.

```
mpage = harmmean(X,[1 2])
```

```
mpage =
mpage(:,:,1) =

    4.5205


mpage(:,:,2) =

   22.1645
```

For example, `mpage(1,1,2)` is the harmonic mean of the elements in `X(:,:,2)`.

Find the harmonic mean of the elements in each `X(i,:,:)` slice by specifying the second and third dimensions.

```
mrow = harmmean(X,[2 3])
```

```
mrow = 3×1

    5.5369
    8.2469
   10.2425
```

For example, `mrow(3)` is the harmonic mean of the elements in `X(3,:,:)` and is equivalent to specifying `harmmean(X(3,:,:),'all')`.

Harmonic Mean Excluding NaN

Create a vector and compute its `harmmean`, excluding `NaN` values.

```
x = 1:10;
x(3) = nan; % Replace the third element of x with a NaN value
n = harmmean(x,'omitnan')
```

```
n = 3.4674
```

If you do not specify `'omitnan'`, then `harmmean(x)` returns `NaN`.

More About

Harmonic Mean

The harmonic mean of a sample `X` is

$$m = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

where n is the number of values in `X`.

Tips

- When `harmmean` computes the harmonic mean of an array containing `0`, the returned value is `0`.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- These input arguments are not supported: `'all'`, `vecdim`, and `nanflag`.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

geomean | mean | median | trimmean

Introduced before R2006a

hazardratio

Estimate Cox model hazard relative to baseline

Syntax

```
hazard = hazardratio(coxMdl,X)
hazard = hazardratio(coxMdl,X,Stratification)
hazard = hazardratio( ___ ,'Baseline',baseline)
```

Description

`hazard = hazardratio(coxMdl,X)` returns the estimated hazard relative to the baseline for a fitted Cox proportional hazards model `coxMdl` using the predictors `X`.

`hazard = hazardratio(coxMdl,X,Stratification)` returns the estimated hazard relative to the baseline using the predictors `X` and stratification levels `Stratification`. The number of rows in `X` and `Stratification` must be the same.

Note When you train `coxMdl` using stratification variables and pass predictor variables `X`, `hazardratio` also requires you to pass stratification variables.

`hazard = hazardratio( ___ ,'Baseline',baseline)` estimates the hazard relative to the supplied baseline using any of the input argument combinations in the previous syntaxes.

Examples

Compute Relative Hazard

Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

```
load lightbulb
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3));
```

View the default baseline for the fitted model.

```
defaultBaseline = coxMdl.Baseline
```

```
defaultBaseline = 0.5000
```

Compute the hazard ratio of an incandescent bulb (1) relative to this baseline.

```
defaultHazard = hazardratio(coxMdl,1)
```

```
defaultHazard = 10.6238
```

Compute the hazard ratio of an incandescent bulb relative to a fluorescent bulb (0).

```
relHazard = hazardratio(coxMdl,1,'Baseline',0)
```

```
relHazard = 112.8646
```

The hazard rate of an incandescent bulb is estimated to be over 100 times the hazard rate of a fluorescent bulb.

Compute Hazard Relative to Different Baseline Values

Create a Cox model from the `readmissiontimes` data. In this data, `0` indicates a male patient, and `1` indicates a female patient.

```
load readmissiontimes
coxMdl = fitcox([Age,Sex,Weight],ReadmissionTime,'Censoring',Censored);
```

Calculate the relative hazard of a 40-year-old man weighing 200 lbs. relative to the baseline hazard.

```
hazard = hazardratio(coxMdl,[40 0 200])
```

```
hazard = 4.3112
```

Calculate the hazard of this same man relative to a 50-year-old woman weighing 150 lbs.

```
hazard2 = hazardratio(coxMdl,[40 0 200],'Baseline',[50 1 150])
```

```
hazard2 = 5.2053
```

Hazard Ratios for Stratified Model

Load the `coxModel` data. (This simulated data is generated in the example “Cox Proportional Hazards Model Object” on page 14-39.) The model named `coxMdl` has three stratification levels (1, 2, and 3) and a predictor `X` with three categorical values (1, 1/20, and 1/100).

```
load coxModel
```

Find the hazard ratio of the predictor value `categorical(1)` and stratification level 3 with respect to the baseline.

```
X = categorical(1);
stratification = 3;
hazard = hazardratio(coxMdl,X,stratification)
```

```
hazard = 12.7096
```

Calculate the ratio with respect to a baseline of 0.

```
hazard = hazardratio(coxMdl,X,stratification,'Baseline',0)
```

```
hazard = 95.5127
```

Calculate the ratio of a `categorical(1/100)` predictor with respect to a baseline of 0.

```
X = categorical(1/100);
hazard = hazardratio(coxMdl,X,stratification,'Baseline',0)
```

```
hazard = 1
```

Input Arguments

`coxMdl` — Fitted Cox proportional hazards model

`CoxModel` object

Fitted Cox proportional hazards model, specified as a `CoxModel` object. Create `coxMdl` using `fitcox`.

`X` — Data for estimating hazard

matrix | table

Data for estimating the hazard, specified as a matrix or table. The data must be the same type as the data used to train `coxMdl`.

Data Types: `double` | `table` | `categorical`

`Stratification` — Stratification level

variable or variables of type used for training

Stratification level, specified as a variable or variables of the same type used for training `coxMdl`. Specify the same number of rows in `Stratification` as in `X`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `table` | `cell` | `categorical`

`baseline` — Baseline hazard

inferred from `coxMdl` (default) | real scalar | real row vector

Baseline hazard, specified as a real scalar or row vector.

- A scalar value applies to all predictors.
- A row vector value must have the same number of entries as the number of predictors.

The returned hazard ratio is relative to the baseline.

Example: `[1 20 100]`

Data Types: `single` | `double`

Output Arguments

`hazard` — Hazard ratio relative to baseline

nonnegative vector

Hazard ratio relative to the baseline, returned as a nonnegative vector. `hazard` gives the factor by which to multiply the baseline hazard, so you can obtain the relative hazard of an individual with predictor values `X` and, if applicable, stratification level `Stratification`.

See Also

`CoxModel` | `fitcox` | `plotSurvival` | `survival`

Topics

“Cox Proportional Hazards Model Object” on page 14-39

Introduced in R2021a

hist3

Bivariate histogram plot

Syntax

```
hist3(X)
hist3(X,'Nbins',nbins)
hist3(X,'Ctrs',ctrs)
hist3(X,'Edges',edges)
hist3(___,Name,Value)
hist3(ax,___)

N = hist3(___)
[N,c] = hist3(___)
```

Description

`hist3(X)` creates a bivariate histogram plot of `X(:,1)` and `X(:,2)` using 10-by-10 equally spaced bins. The `hist3` function displays the bins as 3-D rectangular bars, and the height of each bar indicates the number of elements in the bin.

`hist3(X,'Nbins',nbins)` specifies the number of bins in each dimension of the histogram. This syntax is equivalent to `hist3(X,nbins)`.

`hist3(X,'Ctrs',ctrs)` specifies the centers of the bins in each dimension of the histogram. This syntax is equivalent to `hist3(X,ctrs)`.

`hist3(X,'Edges',edges)` specifies the edges of the bins in each dimension.

`hist3(___,Name,Value)` specifies graphical properties using one or more name-value pair arguments in addition to the input arguments in the previous syntaxes. For example, `'FaceAlpha',0.5` creates a semitransparent histogram. For a list of properties, see Surface Properties.

`hist3(ax,___)` plots into the axes specified by `ax` instead of the current axes (`gca`). The option `ax` can precede any of the input argument combinations in the previous syntaxes.

`N = hist3(___)` returns the number of elements in `X` that fall in each bin. This syntax does not create a histogram.

`[N,c] = hist3(___)` also returns the bin centers. This syntax does not create a histogram.

Examples

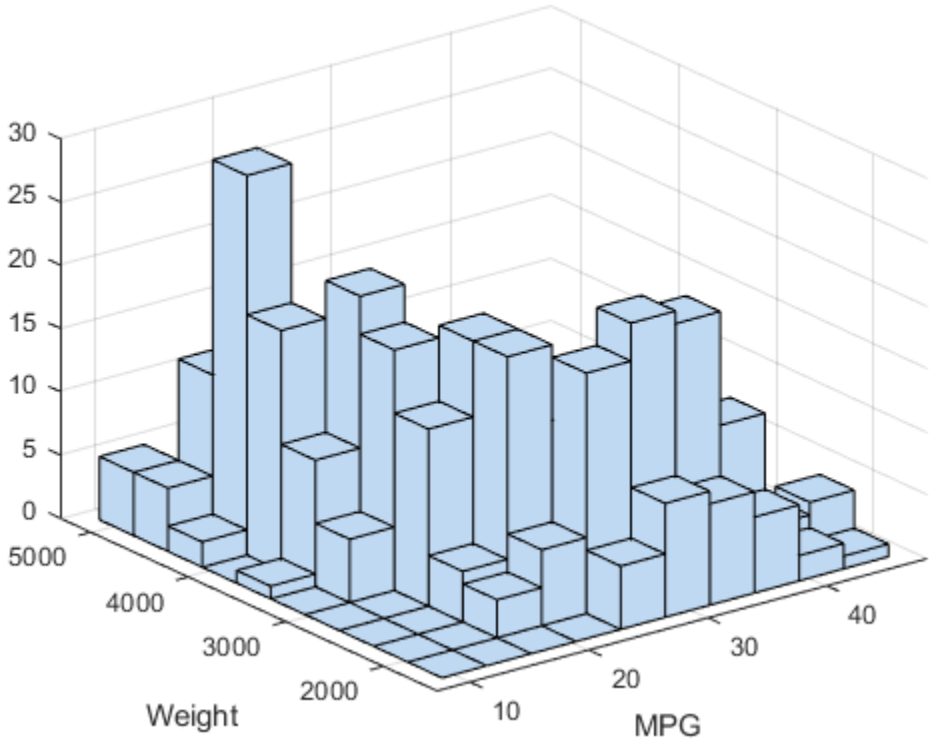
Histogram of Vectors

Load the sample data.

```
load carbig
```

Create a bivariate histogram with the default settings.

```
X = [MPG,Weight];
hist3(X)
xlabel('MPG')
ylabel('Weight')
```



Specify Centers of Histogram Bins

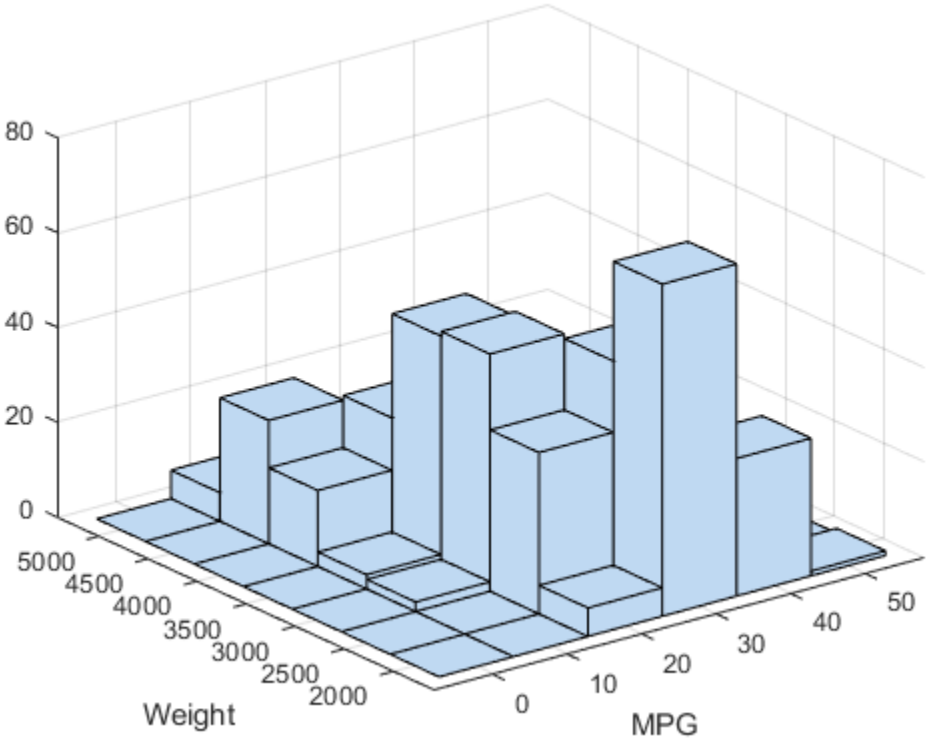
Create a bivariate histogram on the bins specified by the bin centers, and count the number of elements in each bin.

Load the sample data.

```
load carbig
```

Create a bivariate histogram. Specify the centers of the histogram bins using a two-element cell array.

```
X = [MPG,Weight];
hist3(X,'Ctrs',{0:10:50 2000:500:5000})
xlabel('MPG')
ylabel('Weight')
```



Count the number of elements in each bin.

```
N = hist3(X,'Ctrs',{0:10:50 2000:500:5000})
```

```
N = 6×7

     0     0     0     0     0     0     0
     0     0     2     3    16    26     6
     6    34    50    49    27    10     0
    70    49    11     3     0     0     0
    29     4     2     0     0     0     0
     1     0     0     0     0     0     0
```

Color Histogram Bars by Height

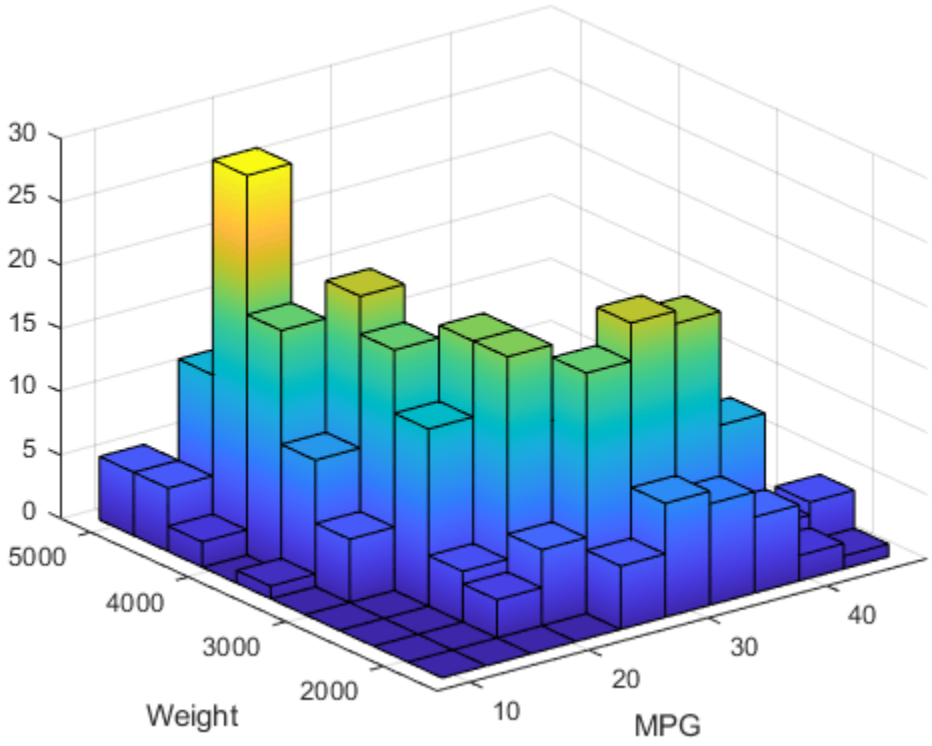
Load the sample data.

```
load carbig
```

Create a bivariate histogram. Specify graphical properties to color the histogram bars by height representing the frequency of the observations.

```
X = [MPG,Weight];
hist3(X,'CDataMode','auto','FaceColor','interp')
```

```
xlabel('MPG')
ylabel('Weight')
```



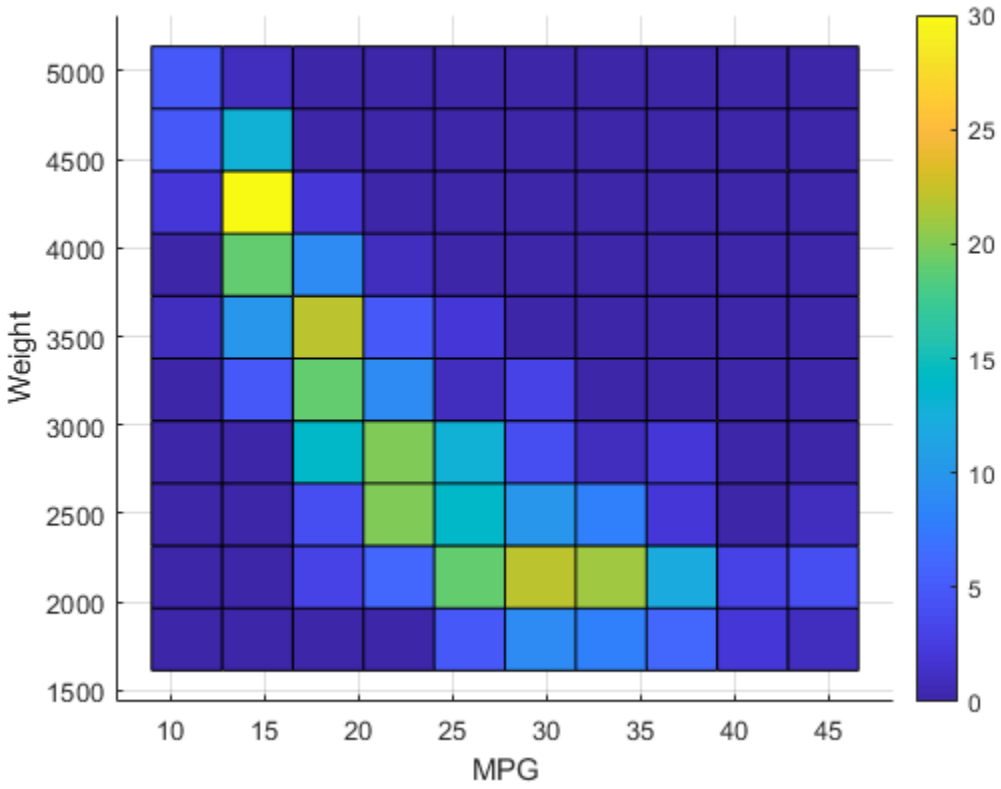
Tiled Histogram View

Load the sample data.

```
load carbig
```

Create a bivariate tiled histogram. Specify graphical properties to color the top surface of the histogram bars by the frequency of the observations. Change the view to two-dimensional.

```
X = [MPG,Weight];
hist3(X,'CdataMode','auto')
xlabel('MPG')
ylabel('Weight')
colorbar
view(2)
```



Adjust Graphical Properties

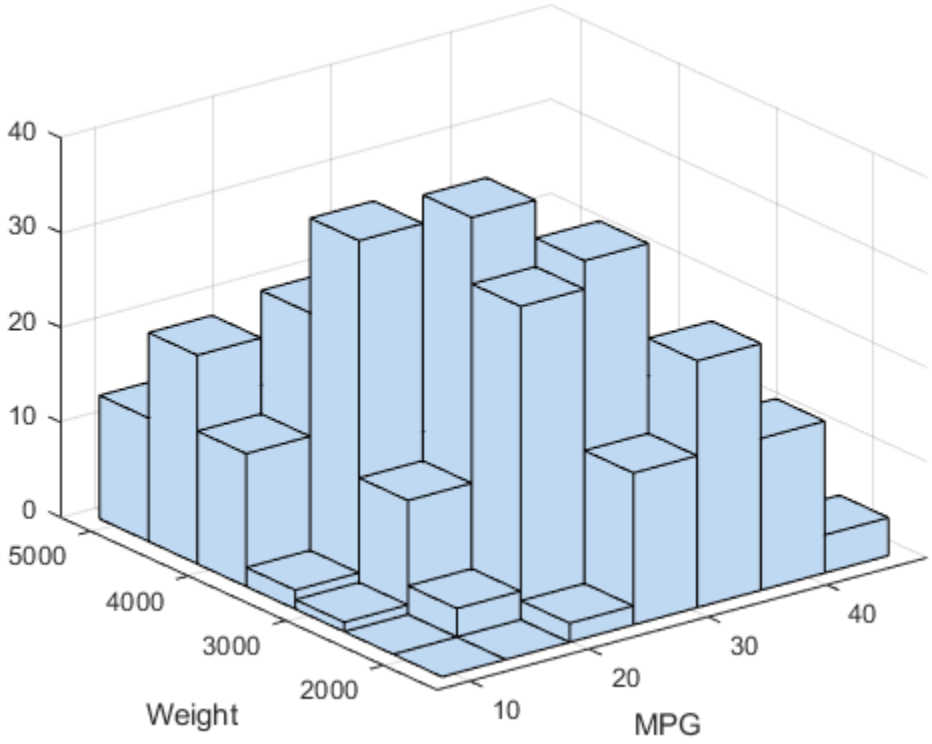
Create a bivariate histogram and adjust its graphical properties by using the handle of the histogram surface object.

Load the sample data.

```
load carbig
```

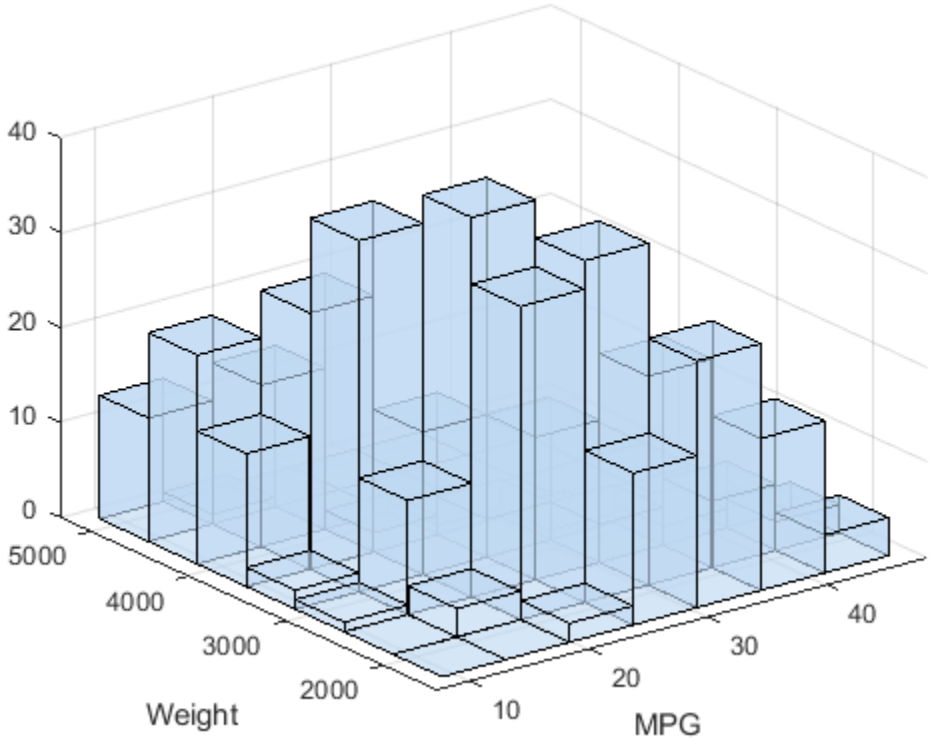
Create a bivariate histogram with 7 bins in each dimension.

```
X = [MPG,Weight];
hist3(X,'Nbins',[7 7])
xlabel('MPG')
ylabel('Weight')
```



The `hist3` function creates a bivariate histogram, which is a type of surface plot. Find the handle of the surface object and adjust the face transparency.

```
s = findobj(gca,'Type','Surface');
s.FaceAlpha = 0.65;
```



Plot Histogram with Intensity Map

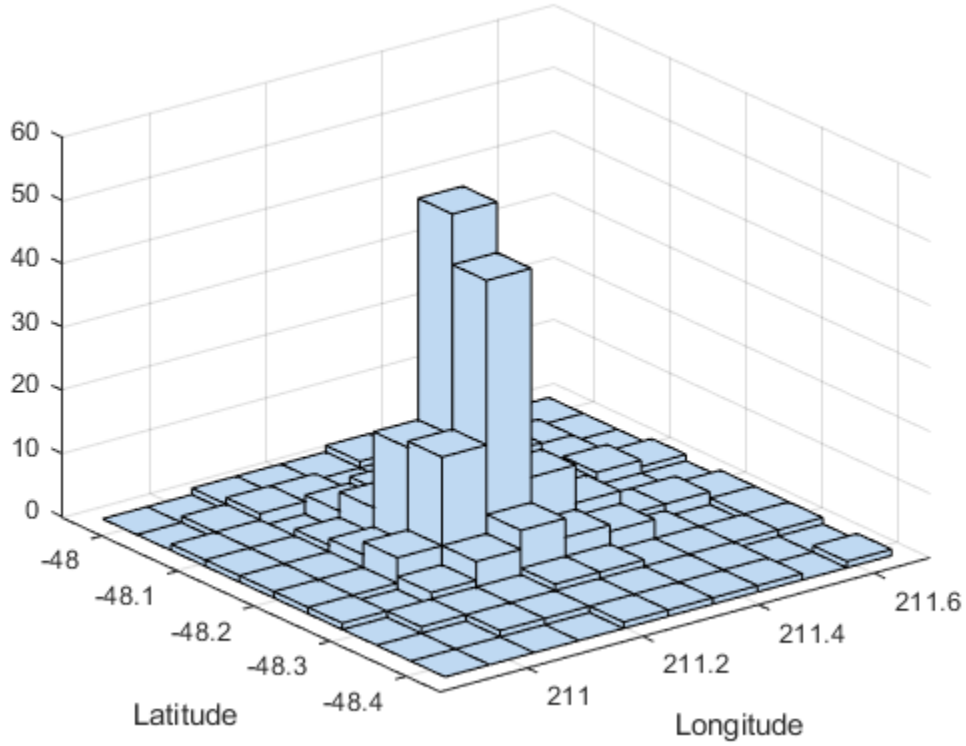
Create a bivariate histogram and add the 2-D projected view of intensities to the histogram.

Load the `seamount` data set (a *seamount* is an underwater mountain). The data set consists of a set of longitude (`x`) and latitude (`y`) locations, and the corresponding `seamount` elevations (`z`) measured at those coordinates. This example uses `x` and `y` to draw a bivariate histogram.

```
load seamount
```

Draw a bivariate histogram.

```
hist3([x,y])
xlabel('Longitude')
ylabel('Latitude')
hold on
```



Count the number of elements in each bin.

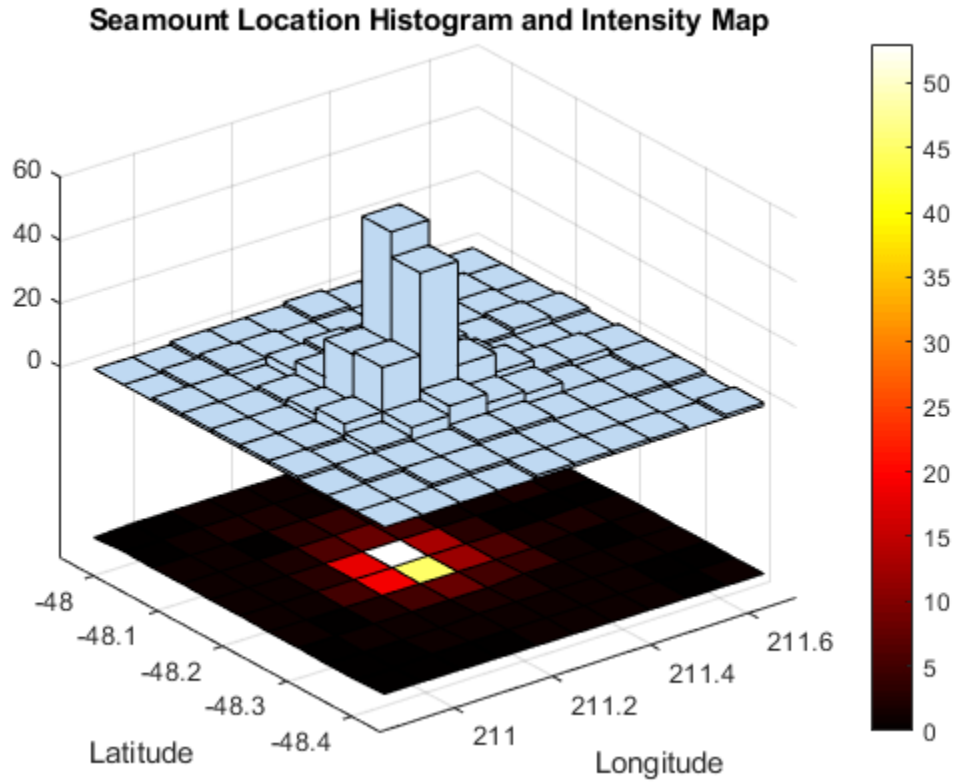
```
N = hist3([x,y]);
```

Generate a grid to draw the 2-D projected view of intensities by using `pcolor`.

```
N_pcolor = N';
N_pcolor(size(N_pcolor,1)+1,size(N_pcolor,2)+1) = 0;
xl = linspace(min(x),max(x),size(N_pcolor,2)); % Columns of N_pcolor
yl = linspace(min(y),max(y),size(N_pcolor,1)); % Rows of N_pcolor
```

Draw the intensity map by using `pcolor`. Set the *z*-level of the intensity map to view the histogram and the intensity map together.

```
h = pcolor(xl,yl,N_pcolor);
colormap('hot') % Change color scheme
colorbar % Display colorbar
h.ZData = -max(N_pcolor(:))*ones(size(N_pcolor));
ax = gca;
ax.ZTick(ax.ZTick < 0) = [];
title('Seamount Location Histogram and Intensity Map');
```



Input Arguments

X — Data to distribute among bins

m-by-2 numeric matrix

Data to distribute among the bins, specified as an *m*-by-2 numeric matrix, where *m* is the number of data points. Corresponding elements in `X(:,1)` and `X(:,2)` specify the *x* and *y* coordinates of 2-D data points.

`hist3` ignores all `NaN` values. Similarly, `hist3` ignores `Inf` and `–Inf` values unless you explicitly specify `Inf` or `–Inf` as a bin edge by using the `edges` input argument.

Data Types: `single` | `double`

nbins — Number of bins

`[10 10]` (default) | two-element vector of positive integers

Number of bins in each dimension, specified as a two-element vector of positive integers. `nbins(1)` specifies the number of bins in the first dimension, and `nbins(2)` specifies the number of bins in the second dimension.

Example: `[10 20]`

Data Types: `single` | `double`

`ctrs` — Bin centers

two-element cell array of numeric vectors

Bin centers in each dimension, specified as a two-element cell array of numeric vectors with monotonically nondecreasing values. `ctrs{1}` and `ctrs{2}` are the positions of the bin centers in the first and second dimensions, respectively.

`hist3` assigns rows of `X` falling outside the range of the grid to the bins along the outer edges of the grid.

Example: `{0:10:100 0:50:500}`

Data Types: `cell`

`edges` — Bin edges

two-element cell array of numeric vectors

Bin edges in each dimension, specified as a two-element cell array of numeric vectors with monotonically nondecreasing values. `edges{1}` and `edges{2}` are the positions of the bin edges in the first and second dimensions, respectively.

The value `X(k,:)` is in the `(i,j)`th bin if `edges{1}(i) ≤ X(k,1) < edges{1}(i+1)` and `edges{2}(j) ≤ X(k,2) < edges{2}(j+1)`.

The last bins in each dimension also include the last (outer) edge. For example, `X(k,:)` falls into the `(I,j)`th bin if `edges{1}(I–1) ≤ X(k,1) ≤ edges{1}(I)` and `edges{2}(j) ≤ X(k,2) < edges{2}(j+1)`, where `I` is the length of `edges{1}`. Also, `X(k,:)` falls into the `(i,J)`th bin if `edges{1}(i) ≤ X(k,1) < edges{1}(i+1)` and `edges{2}(J–1) ≤ X(k,2) ≤ edges{2}(J)`, where `J` is the length of `edges{2}`.

`hist3` does not count rows of `X` falling outside the range of the grid. Use `–Inf` and `Inf` in `edges` to include all non-`NaN` values.

Example: `{0:10:100 0:50:500}`

Data Types: `cell`

`ax` — Target axes

current axes (`gca`) (default) | `Axes` object

Target axes, specified as an axes object. If you do not specify an `Axes` object, then the `hist3` function uses the current axes (`gca`). For details, see Axes Properties.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `hist3(X,'FaceColor','interp','CDataMode','auto')` colors the histogram bars according to the height of the bars.

The graphical properties listed here are only a subset. For a full list, see Surface Properties.

`CDataMode` — Selection mode for vertex colors

`'manual'` (default) | `'auto'`

Selection mode for `CData` (vertex colors), specified as the comma-separated pair consisting of `'CDataMode'` and one of these values:

- `'manual'` — Use manually specified values in the `CData` property. The default color in `CData` is light steel blue corresponding to an RGB triple value of `[0.75 0.85 0.95]`.
- `'auto'` — Use the `ZData` values to set the colors. `ZData` contains the *z*-coordinate data for the eight corners of each bar.

Example: `'CDataMode','auto'`

`EdgeColor` — Edge line color
`[0 0 0]` (default) | `'none'` | `'flat'` | `'interp'` | RGB triplet | hexadecimal color code | color name | short name

Edge line color, specified as the comma-separated pair consisting of `'EdgeColor'` and one of these values:

- `'none'` — Do not draw the edges.
- `'flat'` — Use a different color for each edge based on the values in the `CData` property.
- `'interp'` — Use interpolated coloring for each edge based on the values in the `CData` property.
- RGB triplet, hexadecimal color code, color name, or short name — Use the specified color for all the edges. These values do not use the color values in the `CData` property.

The default color of `[0 0 0]` corresponds to black edges.

RGB triplets and hexadecimal color codes are useful for specifying custom colors.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
[0 0.4470 0.7410]	'#0072BD'	
[0.8500 0.3250 0.0980]	'#D95319'	
[0.9290 0.6940 0.1250]	'#EDB120'	
[0.4940 0.1840 0.5560]	'#7E2F8E'	
[0.4660 0.6740 0.1880]	'#77AC30'	
[0.3010 0.7450 0.9330]	'#4DBEEE'	
[0.6350 0.0780 0.1840]	'#A2142F'	

Example: `'EdgeColor','blue'`

`FaceAlpha` — Face transparency
1 (default) | scalar in the range `[0,1]` | `'flat'` | `'interp'` | `'texturemap'`

Face transparency, specified as the comma-separated pair consisting of `'FaceAlpha'` and one of these values:

- Scalar in the range `[0,1]` — Use uniform transparency across all the faces. A value of `1` is fully opaque and `0` is completely transparent. Values between `0` and `1` are semitransparent. This option does not use the transparency values in the `AlphaData` property.
- `'flat'` — Use a different transparency for each face based on the values in the `AlphaData` property. The transparency value at the first vertex determines the transparency for the entire face. This value applies only when you specify the `AlphaData` property and set the `FaceColor` property to `'flat'`.
- `'interp'` — Use interpolated transparency for each face based on the values in the `AlphaData` property. The transparency varies across each face by interpolating the values at the vertices. This value applies only when you specify the `AlphaData` property and set the `FaceColor` property to `'interp'`.
- `'texturemap'` — Transform the data in `AlphaData` so that it conforms to the surface.

Example: `'FaceAlpha',0.5`

`FaceColor` — Face color
`'flat'` (default) | `'interp'` | `'none'` | `'texturemap'` | RGB triplet | hexadecimal color code | color name | short name

Face color, specified as the comma-separated pair consisting of `'FaceColor'` and one of these values:

- `'flat'` — Use a different color for each face based on the values in the `CData` property.
- `'interp'` — Use interpolated coloring for each face based on the values in the `CData` property.
- `'none'` — Do not draw the faces.
- `'texturemap'` — Transform the color data in `CData` so that it conforms to the surface.
- RGB triplet, hexadecimal color code, color name, or short name — Use the specified color for all the faces. These values do not use the color values in the `CData` property.

RGB triplets and hexadecimal color codes are useful for specifying custom colors.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

Example: `'FaceColor','interp'`

`LineStyle` — Line style

`'-'` (default) | `'--'` | `':'` | `'-.'` | `'none'`

Line style, specified as the comma-separated pair consisting of `'LineStyle'` and one of the options in this table.

Line Style	Description	Resulting Line
`'-'`	Solid line	
`'--'`	Dashed line	

Line Style	Description	Resulting Line
':'	Dotted line	.....
'-.'	Dash-dotted line	_._._._.
'none'	No line	No line

Example: 'LineStyle',':'

LineWidth — Line width
0.5 (default) | positive value

Line width, specified as the comma-separated pair consisting of 'LineWidth' and a positive value in points.

Example: 'LineWidth',0.75

Data Types: single | double

Output Arguments

N — Number of elements in each bin
numeric matrix

Number of elements in X that fall in each bin, returned as a numeric matrix.

c — Bin centers
two-element cell array of numeric vectors

Bin centers in each dimension, returned as a two-element cell array of numeric vectors. c{1} and c{2} are the positions of the bin centers in the first and second dimensions, respectively.

Tips

The hist3 function creates a bivariate histogram, which is a type of surface plot. You can specify surface properties using one or more name-value pair arguments. Also, you can change the appearance of the histogram by changing the surface property values after you create a histogram. Get the handle of the surface object by using s = findobj(gca,'Type','Surface'), and then use s to modify the surface properties. For an example, see “Adjust Graphical Properties” on page 33-2808. For a list of properties, see Surface Properties.

Alternative Functionality

The histogram2 function enables you to create a bivariate histogram using a Histogram2 object. You can use the name-value pair arguments of histogram2 to use normalization ('Normalization'), adjust the width of the bins in each dimension ('BinWidth'), and display the histogram as a rectangular array of tiles instead of 3-D bars ('DisplayStyle').

See Also

accumarray | bar3 | binScatterPlot | histcounts2 | histogram2

Introduced before R2006a

histfit

Histogram with a distribution fit

Syntax

```
histfit(data)
histfit(data,nbins)
histfit(data,nbins,dist)

histfit(ax, ___ )

h = histfit( ___ )
```

Description

`histfit(data)` plots a histogram of values in `data` using the number of bins equal to the square root of the number of elements in `data` and fits a normal density function.

`histfit(data,nbins)` plots a histogram using `nbins` bins and fits a normal density function.

`histfit(data,nbins,dist)` plots a histogram with `nbins` bins and fits a density function from the distribution specified by `dist`.

`histfit(ax, ___ )` uses the plot axes specified by the `Axes` object `ax`. Specify `ax` as the first input argument followed by any of the input argument combinations in the previous syntaxes.

`h = histfit( ___ )` returns a vector of handles `h`, where `h(1)` is the handle to the histogram and `h(2)` is the handle to the density curve.

Examples

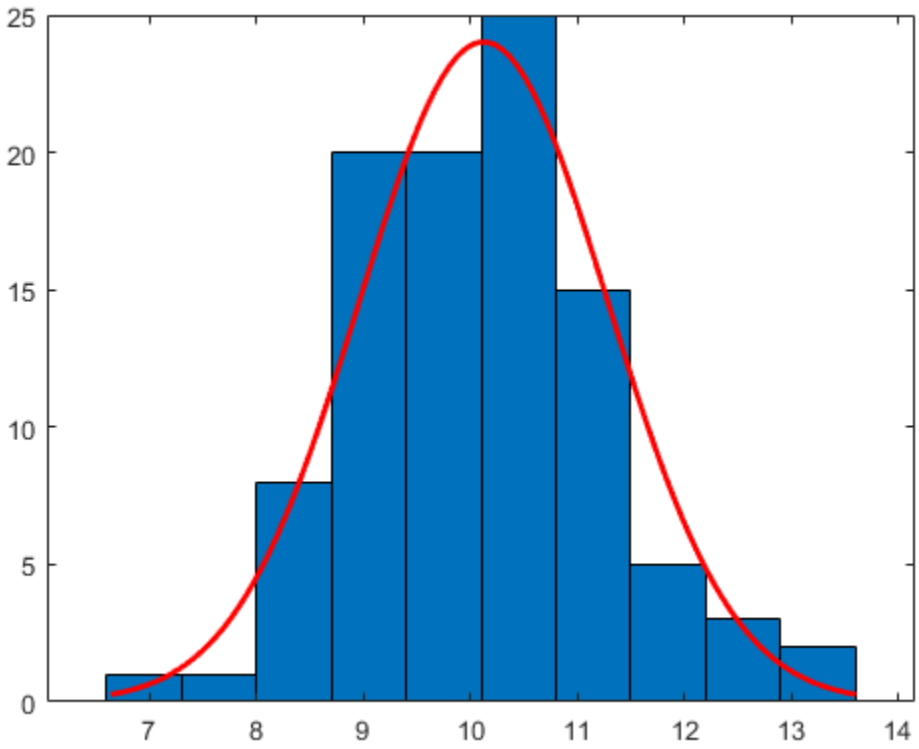
Histogram with a Normal Distribution Fit

Generate a sample of size 100 from a normal distribution with mean 10 and variance 1.

```
rng default; % For reproducibility
r = normrnd(10,1,100,1);
```

Construct a histogram with a normal distribution fit.

```
histfit(r)
```



`histfit` uses `fitdist` to fit a distribution to data. Use `fitdist` to obtain parameters used in fitting.

```
pd = fitdist(r,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 10.1231   [9.89244, 10.3537]
    sigma =  1.1624   [1.02059, 1.35033]
```

The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

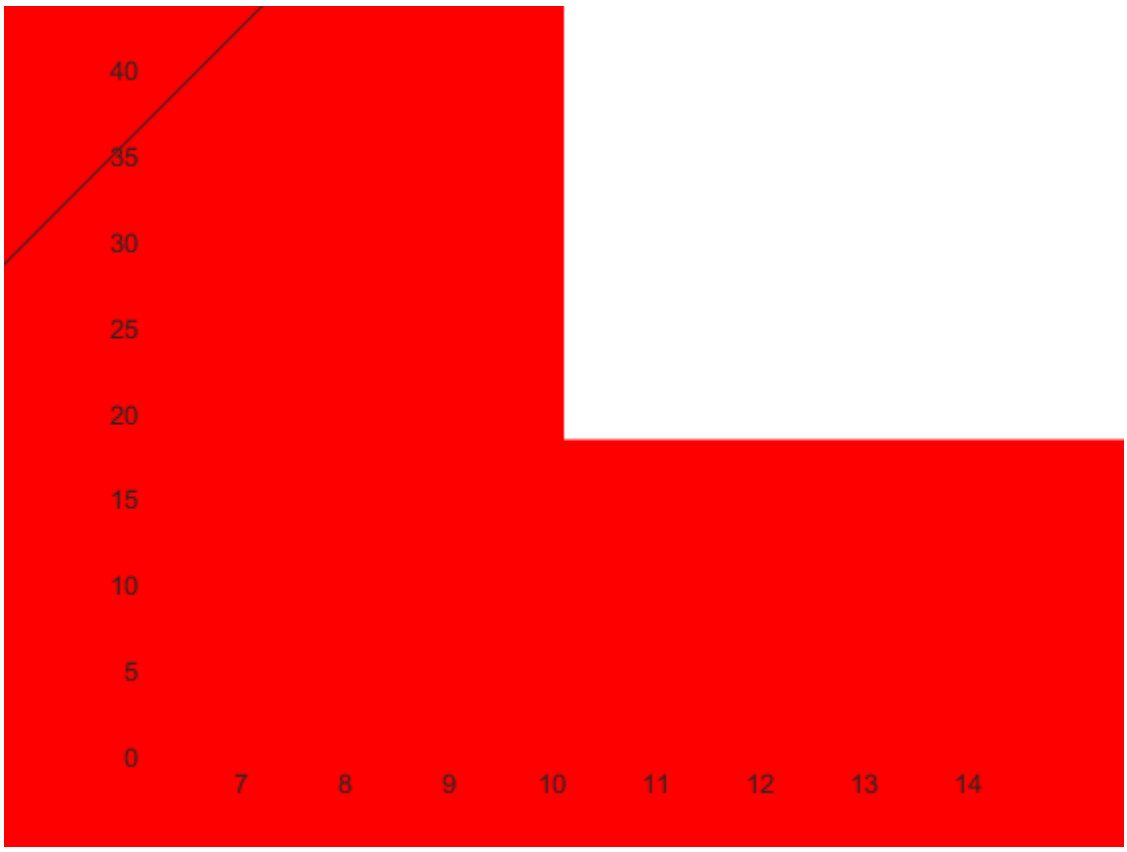
Histogram for a Given Number of Bins

Generate a sample of size 100 from a normal distribution with mean 10 and variance 1.

```
rng default; % For reproducibility
r = normrnd(10,1,100,1);
```

Construct a histogram using six bins with a normal distribution fit.

```
histfit(r,6)
```



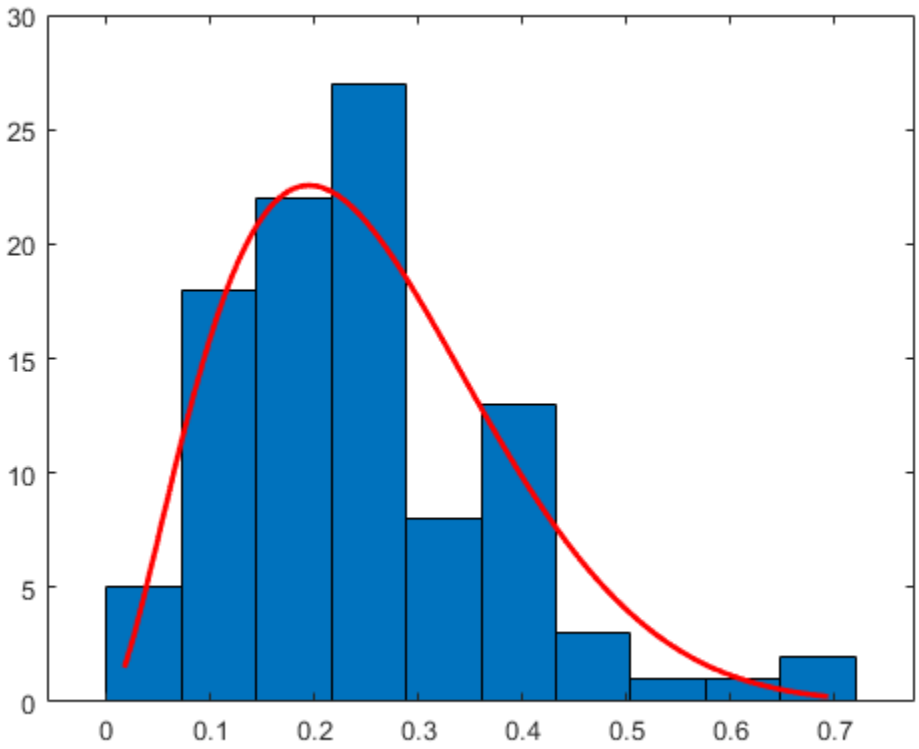
Histogram with a Specified Distribution Fit

Generate a sample of size 100 from a beta distribution with parameters (3,10).

```
rng default;  % For reproducibility
b = betarnd(3,10,100,1);
```

Construct a histogram using 10 bins with a beta distribution fit.

```
histfit(b,10,'beta')
```



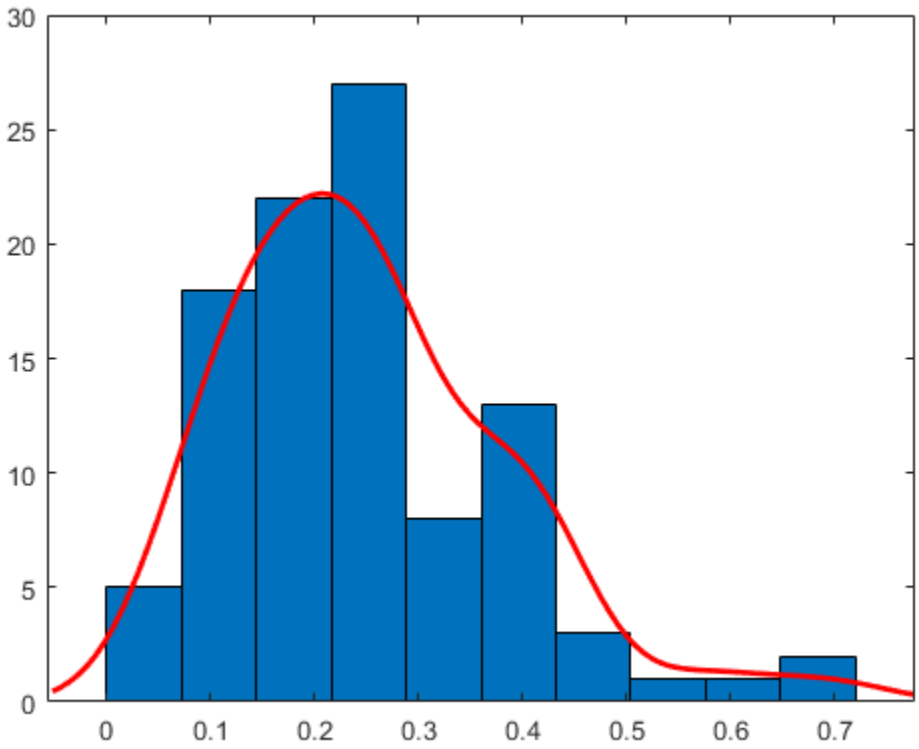
Histogram with a Kernel Smoothing Function Fit

Generate a sample of size 100 from a beta distribution with parameters (3,10).

```
rng default;  % For reproducibility
b = betarnd(3,10,[100,1]);
```

Construct a histogram using 10 bins with a smoothing function fit.

```
histfit(b,10,'kernel')
```



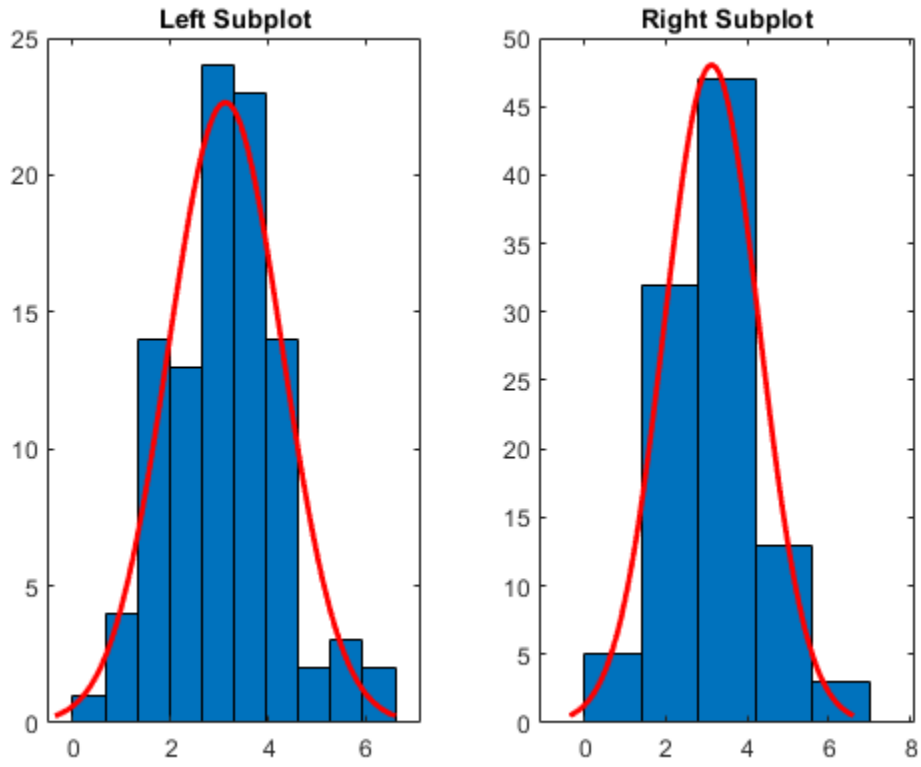
Specify Axes for Histogram with Distribution Fit

Generate a sample of size `100` from a normal distribution with mean `3` and variance `1`.

```
rng('default') % For reproducibility
r = normrnd(3,1,100,1);
```

Create a figure with two subplots and return the `Axes` objects as `ax1` and `ax2`. Create a histogram with a normal distribution fit in each set of axes by referring to the corresponding `Axes` object. In the left subplot, plot a histogram with 10 bins. In the right subplot, plot a histogram with 5 bins. Add a title to each plot by passing the corresponding `Axes` object to the `title` function.

```
ax1 = subplot(1,2,1); % Left subplot
histfit(ax1,r,10,'normal')
title(ax1,'Left Subplot')

ax2 = subplot(1,2,2); % Right subplot
histfit(ax2,r,5,'normal')
title(ax2,'Right Subplot')
```



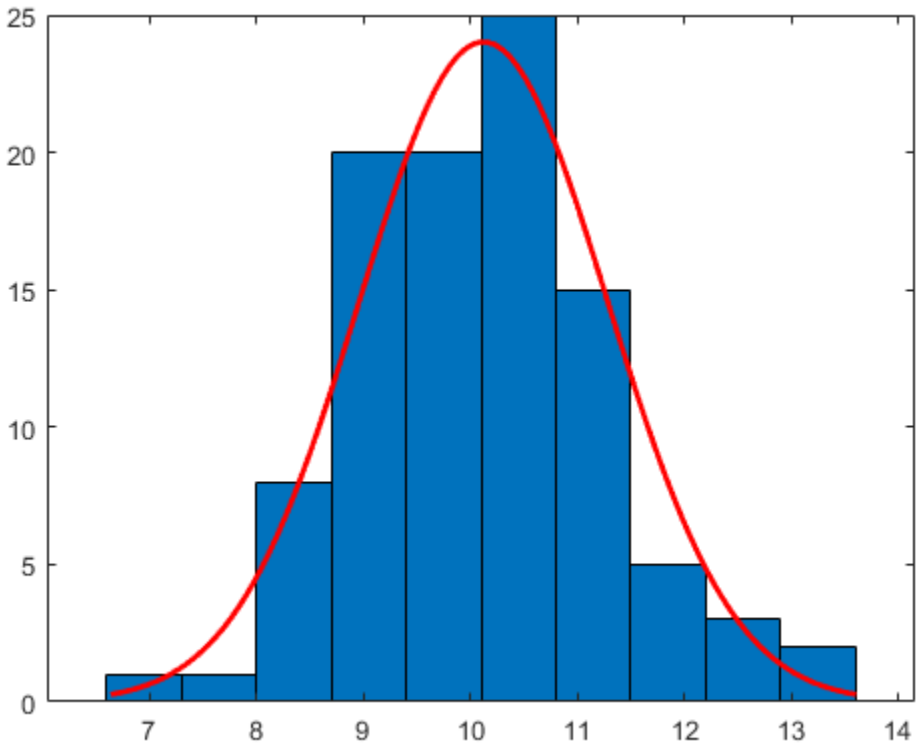
Handle for a Histogram with a Distribution Fit

Generate a sample of size 100 from a normal distribution with mean 10 and variance 1.

```
rng default % for reproducibility
r = normrnd(10,1,100,1);
```

Construct a histogram with a normal distribution fit.

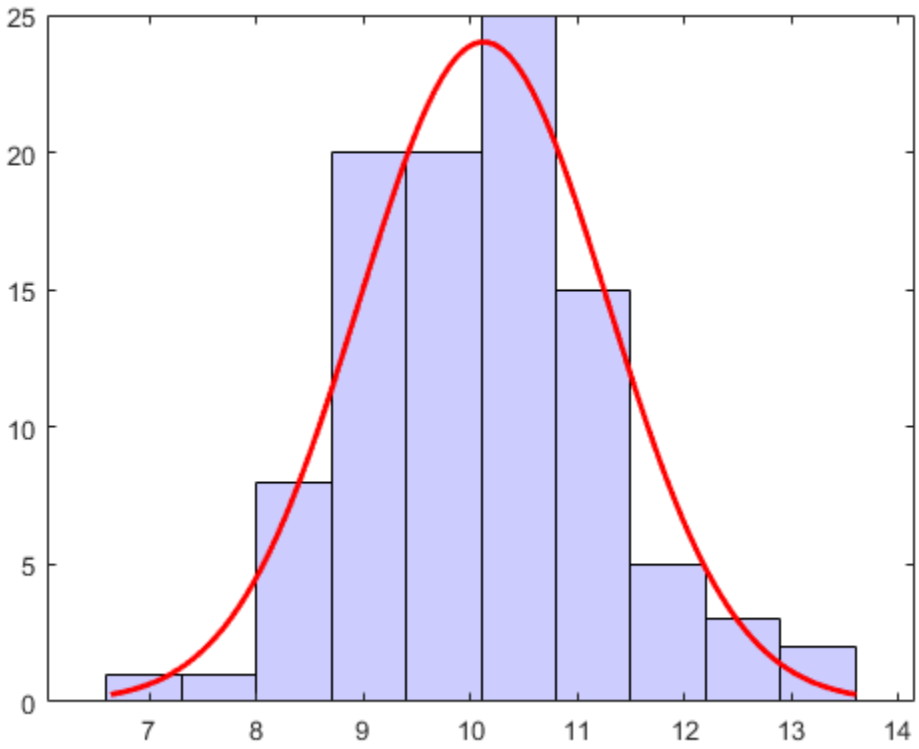
```
h = histfit(r,10,'normal')
```




```
h =
  2x1 graphics array:

  Bar
  Line
```

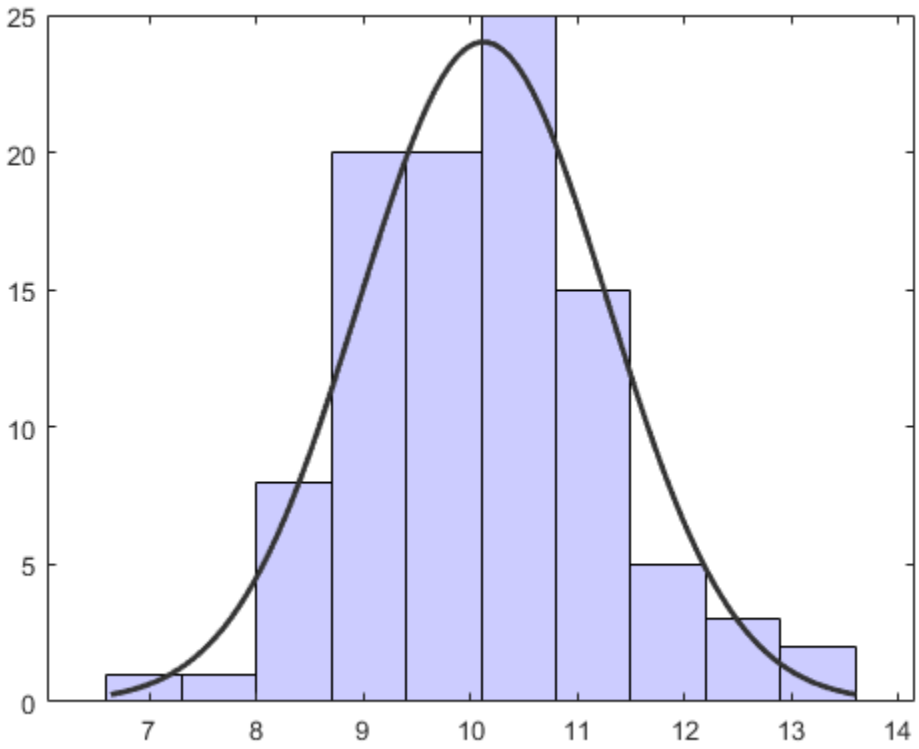
Change the bar colors of the histogram.

```
h(1).FaceColor = [.8 .8 1];
```



Change the color of the density curve.

```
h(2).Color = [.2 .2 .2];
```



Input Arguments

`data` — Input data

vector

Input data, specified as a vector.

Example: `data = [1.5 2.5 4.6 1.2 3.4]`

Example: `data = [1.5 2.5 4.6 1.2 3.4]'`

Data Types: `double` | `single`

`nbins` — Number of bins

positive integer | [ ]

Number of bins for the histogram, specified as a positive integer. Default value is the square root of the number of elements in `data`, rounded up. Use [ ] for the default number of bins when fitting a distribution.

Example: `y = histfit(x,8)`

Example: `y = histfit(x,10,'gamma')`

Example: `y = histfit(x,[ ],'weibull')`

Data Types: `double` | `single`

dist — Distribution to fit
'normal' (default) | character vector | string scalar

Distribution to fit to the histogram, specified as a character vector or string scalar. The following table shows the supported distributions.

dist	Description
'beta'	Beta
'birnbaumsaunders'	Birnbaum-Saunders
'burr'	Burr Type XII
'exponential'	Exponential
'extreme value' or 'ev'	Extreme value
'gamma'	Gamma
'generalized extreme value' or 'gev'	Generalized extreme value
'generalized pareto' or 'gp'	Generalized Pareto (threshold 0)
'inversegaussian'	Inverse Gaussian
'logistic'	Logistic
'loglogistic'	Loglogistic
'lognormal'	Lognormal
'nakagami'	Nakagami
'negative binomial' or 'nbin'	Negative binomial
'normal'	Normal
'poisson'	Poisson
'rayleigh'	Rayleigh
'rician'	Rician
'tlocationscale'	t location-scale
'weibull' or 'wbl'	Weibull
'kernel'	Nonparametric kernel-smoothing distribution. The density is evaluated at 100 equally spaced points that cover the range of the data in data. It works best with continuously distributed samples.

ax — Axes for plot
Axes object

Axes for the plot, specified as an Axes object. If you do not specify ax, then histfit creates the plot using the current axes. For more information on creating an Axes object, see axes.

Output Arguments

h — Handles for the plot
plot handle

Handles for the plot, returned as a vector, where `h(1)` is the handle to the histogram, and `h(2)` is the handle to the density curve. `histfit` normalizes the density to match the total area under the curve with that of the histogram.

Algorithms

`histfit` uses `fitdist` to fit a distribution to data. Use `fitdist` to obtain parameters used in fitting.

See Also

`distributionFitter` | `fitdist` | `histogram` | `normfit` | `paramci`

Introduced before R2006a

hmmdecode

Hidden Markov model posterior state probabilities

Syntax

```
PSTATES = hmmdecode(seq,TRANS,EMIS)
[PSTATES,logpseq] = hmmdecode(...)
[PSTATES,logpseq,FORWARD,BACKWARD,S] = hmmdecode(...)
hmmdecode(...,'Symbols',SYMBOLS)
```

Description

`PSTATES = hmmdecode(seq,TRANS,EMIS)` calculates the posterior state probabilities, `PSTATES`, of the sequence `seq`, from a hidden Markov model. The posterior state probabilities are the conditional probabilities of being at state *k* at step *i*, given the observed sequence of symbols, `sym`. You specify the model by a transition probability matrix, `TRANS`, and an emissions probability matrix, `EMIS`. `TRANS(i,j)` is the probability of transition from state `i` to state `j`. `EMIS(k,seq)` is the probability that symbol `seq` is emitted from state `k`.

`PSTATES` is an array with the same length as `seq` and one row for each state in the model. The (*i*, *j*)th element of `PSTATES` gives the probability that the model is in state *i* at the *j*th step, given the sequence `seq`.

Note The function `hmmdecode` begins with the model in state 1 at step 0, prior to the first emission. `hmmdecode` computes the probabilities in `PSTATES` based on the fact that the model begins in state 1.

`[PSTATES,logpseq] = hmmdecode(...)` returns `logpseq`, the logarithm of the probability of sequence `seq`, given transition matrix `TRANS` and emission matrix `EMIS`.

`[PSTATES,logpseq,FORWARD,BACKWARD,S] = hmmdecode(...)` returns the forward and backward probabilities of the sequence scaled by `S`.

`hmmdecode(...,'Symbols',SYMBOLS)` specifies the symbols that are emitted. `SYMBOLS` can be a numeric array, a string array, or a cell array of the names of the symbols. The default symbols are integers `1` through `N`, where `N` is the number of possible emissions.

Examples

```
trans = [0.95,0.05;
         0.10,0.90];
emis = [1/6 1/6 1/6 1/6 1/6 1/6;
   1/10 1/10 1/10 1/10 1/10 1/2];

[seq,states] = hmmgenerate(100,trans,emis);
pStates = hmmdecode(seq,trans,emis);
[seq,states] = hmmgenerate(100,trans,emis,...
   'Symbols',{'one','two','three','four','five','six'})
pStates = hmmdecode(seq,trans,emis,...
   'Symbols',{'one','two','three','four','five','six'});
```

References

[1] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

See Also

`hmmestimate` | `hmmgenerate` | `hmmtrain` | `hmmviterbi`

Introduced before R2006a

hmmestimate

Hidden Markov model parameter estimates from emissions and states

Syntax

```
[TRANS,EMIS] = hmmestimate(seq,states)
hmmestimate(...,'Symbols',SYMBOLS)
hmmestimate(...,'Statenames',STATENAMES)
hmmestimate(...,'Pseudoemissions',PSEUDOE)
hmmestimate(...,'Pseudotransitions',PSEUDOTR)
```

Description

`[TRANS,EMIS] = hmmestimate(seq,states)` calculates the maximum likelihood estimate of the transition, `TRANS`, and emission, `EMIS`, probabilities of a hidden Markov model for sequence, `seq`, with known states, `states`.

`hmmestimate(...,'Symbols',SYMBOLS)` specifies the symbols that are emitted. `SYMBOLS` can be a numeric array, a string array or a cell array of the names of the symbols. The default symbols are integers 1 through N, where N is the number of possible emissions.

`hmmestimate(...,'Statenames',STATENAMES)` specifies the names of the states. `STATENAMES` can be a numeric array, a string array, or a cell array of the names of the states. The default state names are 1 through `M`, where `M` is the number of states.

`hmmestimate(...,'Pseudoemissions',PSEUDOE)` specifies pseudocount emission values in the matrix `PSEUDOE`. Use this argument to avoid zero probability estimates for emissions with very low probability that might not be represented in the sample sequence. `PSEUDOE` should be a matrix of size *m*-by-*n*, where *m* is the number of states in the hidden Markov model and *n* is the number of possible emissions. If the $i \to k$ emission does not occur in `seq`, you can set `PSEUDOE(i,k)` to be a positive number representing an estimate of the expected number of such emissions in the sequence `seq`.

`hmmestimate(...,'Pseudotransitions',PSEUDOTR)` specifies pseudocount transition values. You can use this argument to avoid zero probability estimates for transitions with very low probability that might not be represented in the sample sequence. `PSEUDOTR` should be a matrix of size *m*-by-*m*, where *m* is the number of states in the hidden Markov model. If the $i \to j$ transition does not occur in `states`, you can set `PSEUDOTR(i,j)` to be a positive number representing an estimate of the expected number of such transitions in the sequence `states`.

Pseudotransitions and Pseudoemissions

If the probability of a specific transition or emission is very low, the transition might never occur in the sequence `states`, or the emission might never occur in the sequence `seq`. In either case, the algorithm returns a probability of 0 for the given transition or emission in `TRANS` or `EMIS`. You can compensate for the absence of transition with the `'Pseudotransitions'` and `'Pseudoemissions'` arguments. The simplest way to do this is to set the corresponding entry of `PSEUDOE` or `PSEUDOTR` to `1`. For example, if the transition $i \to j$ does not occur in `states`, set `PSEUDOTR(i,j) = 1`. This forces `TRANS(i,j)` to be positive. If you have an estimate for the expected number of transitions $i \to j$ in a sequence of the same length as `states`, and the actual

number of transitions $i \rightarrow j$ that occur in `seq` is substantially less than what you expect, you can set `PSEUDOTR(i,j)` to the expected number. This increases the value of `TRANS(i,j)`. For transitions that do occur in states with the frequency you expect, set the corresponding entry of `PSEUDOTR` to `0`, which does not increase the corresponding entry of `TRANS`.

If you do not know the sequence of states, use `hmmtrain` to estimate the model parameters.

Examples

```
trans = [0.95,0.05; 0.10,0.90];
emis = [1/6 1/6 1/6 1/6 1/6 1/6;
   1/10 1/10 1/10 1/10 1/10 1/2];

[seq,states] = hmmgenerate(1000,trans,emis);
[estimateTR,estimateE] = hmmestimate(seq,states);
```

References

[1] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

See Also

`hmmdecode` | `hmmgenerate` | `hmmtrain` | `hmmviterbi`

Introduced before R2006a

hmmgenerate

Hidden Markov model states and emissions

Syntax

```
[seq,states] = hmmgenerate(len,TRANS,EMIS)
hmmgenerate(...,'Symbols',SYMBOLS)
hmmgenerate(...,'Statenames',STATENAMES)
```

Description

`[seq,states] = hmmgenerate(len,TRANS,EMIS)` takes a known Markov model, specified by transition probability matrix `TRANS` and emission probability matrix `EMIS`, and uses it to generate

- A random sequence `seq` of emission symbols
- A random sequence `states` of states

The length of both `seq` and `states` is `len`. `TRANS(i,j)` is the probability of transition from state `i` to state `j`. `EMIS(k,l)` is the probability that symbol `l` is emitted from state `k`.

Note The function `hmmgenerate` begins with the model in state 1 at step 0, prior to the first emission. The model then makes a transition to state i_1, with probability T_{1i_1}, and generates an emission a_{k_l} with probability $E_{i_1k_{1_l}}$. `hmmgenerate` returns i_1 as the first entry of `states`, and a_{k_l} as the first entry of `seq`.

`hmmgenerate(...,'Symbols',SYMBOLS)` specifies the symbols that are emitted. `SYMBOLS` can be specified as a numeric array, a string array, or a cell array of character vectors. The default symbols are integers `1` through `N`, where `N` is the number of possible emissions.

`hmmgenerate(...,'Statenames',STATENAMES)` specifies the names of the states. `STATENAMES` can be specified as a numeric array, a string array, or a cell array of character vectors. The default state names are `1` through `M`, where `M` is the number of states.

Since the model always begins at state 1, whose transition probabilities are in the first row of `TRANS`, in the following example, the first entry of the output `states` is be 1 with probability 0.95 and 2 with probability 0.05.

Examples

```
 trans = [0.95,0.05;
          0.10,0.90];
 emis = [1/6 1/6 1/6 1/6 1/6 1/6;
    1/10 1/10 1/10 1/10 1/10 1/2];

[seq,states] = hmmgenerate(100,trans,emis)
[seq,states] = hmmgenerate(100,trans,emis,...
    'Symbols',{'one','two','three','four','five','six'},...
    'Statenames',{'fair';'loaded'})
```

See Also

hmmdecode | hmmestimate | hmmtrain | hmmviterbi

Introduced before R2006a

hmmtrain

Hidden Markov model parameter estimates from emissions

Syntax

```
[ESTTR,ESTEMIT] = hmmtrain(seq,TRGUESS,EMITGUESS)
hmmtrain(...,'Algorithm',algorithm)
hmmtrain(...,'Symbols',SYMBOLS)
hmmtrain(...,'Tolerance',tol)
hmmtrain(...,'Maxiterations',maxiter)
hmmtrain(...,'Verbose',true)
hmmtrain(...,'Pseudoemissions',PSEUDOE)
hmmtrain(...,'Pseudotransitions',PSEUDOTR)
```

Description

`[ESTTR,ESTEMIT] = hmmtrain(seq,TRGUESS,EMITGUESS)` estimates the transition and emission probabilities for a hidden Markov model using the Baum-Welch algorithm. `seq` can be a row vector containing a single sequence, a matrix with one row per sequence, or a cell array with each cell containing a sequence. `TRGUESS` and `EMITGUESS` are initial estimates of the transition and emission probability matrices. `TRGUESS(i,j)` is the estimated probability of transition from state `i` to state `j`. `EMITGUESS(i,k)` is the estimated probability that symbol `k` is emitted from state `i`.

`hmmtrain(...,'Algorithm',algorithm)` specifies the training algorithm. *`algorithm`* can be either `'BaumWelch'` or `'Viterbi'`. The default algorithm is `'BaumWelch'`.

`hmmtrain(...,'Symbols',SYMBOLS)` specifies the symbols that are emitted. `SYMBOLS` can be a numeric array, a string array, or a cell array of the names of the symbols. The default symbols are integers `1` through `N`, where `N` is the number of possible emissions.

`hmmtrain(...,'Tolerance',tol)` specifies the tolerance used for testing convergence of the iterative estimation process. The default tolerance is `1e-4`.

`hmmtrain(...,'Maxiterations',maxiter)` specifies the maximum number of iterations for the estimation process. The default maximum is `100`.

`hmmtrain(...,'Verbose',true)` returns the status of the algorithm at each iteration.

`hmmtrain(...,'Pseudoemissions',PSEUDOE)` specifies pseudocount emission values for the Viterbi training algorithm. Use this argument to avoid zero probability estimates for emissions with very low probability that might not be represented in the sample sequence. `PSEUDOE` should be a matrix of size *m*-by-*n*, where *m* is the number of states in the hidden Markov model and *n* is the number of possible emissions. If the *i*→*k* emission does not occur in `seq`, you can set `PSEUDOE(i,k)` to be a positive number representing an estimate of the expected number of such emissions in the sequence `seq`.

`hmmtrain(...,'Pseudotransitions',PSEUDOTR)` specifies pseudocount transition values for the Viterbi training algorithm. Use this argument to avoid zero probability estimates for transitions with very low probability that might not be represented in the sample sequence. `PSEUDOTR` should be a matrix of size *m*-by-*m*, where *m* is the number of states in the hidden Markov model. If the *i*→*j*

transition does not occur in `states`, you can set `PSEUDOTR(i,j)` to be a positive number representing an estimate of the expected number of such transitions in the sequence `states`.

If you know the states corresponding to the sequences, use `hmmestimate` to estimate the model parameters.

Tolerance

The input argument `'tolerance'` controls how many steps the `hmmtrain` algorithm executes before the function returns an answer. The algorithm terminates when all of the following three quantities are less than the value that you specify for `tolerance`:

- The log likelihood that the input sequence `seq` is generated by the currently estimated values of the transition and emission matrices
- The change in the norm of the transition matrix, normalized by the size of the matrix
- The change in the norm of the emission matrix, normalized by the size of the matrix

The default value of `'tolerance'` is `1e-6`. Increasing the tolerance decreases the number of steps the `hmmtrain` algorithm executes before it terminates.

`maxiterations`

The maximum number of iterations, `'maxiterations'`, controls the maximum number of steps the algorithm executes before it terminates. If the algorithm executes `maxiter` iterations before reaching the specified tolerance, the algorithm terminates and the function returns a warning. If this occurs, you can increase the value of `'maxiterations'` to make the algorithm reach the desired tolerance before terminating.

Examples

```
trans = [0.95,0.05;
      0.10,0.90];
emis = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6;
   1/10, 1/10, 1/10, 1/10, 1/10, 1/2];

seq1 = hmmgenerate(100,trans,emis);
seq2 = hmmgenerate(200,trans,emis);
seqs = {seq1,seq2};
[estTR,estE] = hmmtrain(seqs,trans,emis);
```

References

[1] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

See Also

`hmmdecode` | `hmmestimate` | `hmmgenerate` | `hmmviterbi`

Introduced before R2006a

hmmviterbi

Hidden Markov model most probable state path

Syntax

```
STATES = hmmviterbi(seq,TRANS,EMIS)
hmmviterbi(...,'Symbols',SYMBOLS)
hmmviterbi(...,'Statenames',STATENAMES)
```

Description

`STATES = hmmviterbi(seq,TRANS,EMIS)` given a sequence, `seq`, calculates the most likely path through the hidden Markov model specified by transition probability matrix, `TRANS`, and emission probability matrix `EMIS`. `TRANS(i,j)` is the probability of transition from state `i` to state `j`. `EMIS(i,k)` is the probability that symbol `k` is emitted from state `i`.

Note The function `hmmviterbi` begins with the model in state 1 at step 0, prior to the first emission. `hmmviterbi` computes the most likely path based on the fact that the model begins in state 1.

`hmmviterbi(...,'Symbols',SYMBOLS)` specifies the symbols that are emitted. `SYMBOLS` can be a numeric array, a string array, or a cell array of the names of the symbols. The default symbols are integers `1` through `N`, where `N` is the number of possible emissions.

`hmmviterbi(...,'Statenames',STATENAMES)` specifies the names of the states. `STATENAMES` can be a numeric array, a string array, or a cell array of the names of the states. The default state names are 1 through `M`, where `M` is the number of states.

Examples

```
trans = [0.95,0.05;
         0.10,0.90];
emis = [1/6 1/6 1/6 1/6 1/6 1/6;
   1/10 1/10 1/10 1/10 1/10 1/2];

[seq,states] = hmmgenerate(100,trans,emis);
estimatedStates = hmmviterbi(seq,trans,emis);

[seq,states] = ...
   hmmgenerate(100,trans,emis,...
               'Statenames',{'fair';'loaded'});
estimatesStates = ...
   hmmviterbi(seq,trans,emis,...
              'Statenames',{'fair';'loaded'});
```

References

[1] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

See Also

hmmdecode | hmmestimate | hmmgenerate | hmmtrain

Introduced before R2006a

horzcat

Class: `dataset`

(Not Recommended) Horizontal concatenation for dataset arrays

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = horzcat(ds1, ds2, ...)
```

Description

`ds = horzcat(ds1, ds2, ...)` horizontally concatenates the dataset arrays `ds1, ds2, ...` . You may concatenate dataset arrays that have duplicate variable names, however, the variables must contain identical data, and `horzcat` includes only one copy of the variable in the output dataset.

Observation names for all dataset arrays that have them must be identical except for order. `horzcat` concatenates by matching observation names when present, or by position for datasets that do not have observation names.

See Also

`cat` | `vertcat`

hougen

Hougen-Watson model

Syntax

```
yhat = hougen(beta,x)
```

Description

`yhat = hougen(beta,x)` returns the predicted values of the reaction rate, `yhat`, as a function of the vector of parameters, `beta`, and the matrix of data, `X`. `beta` must have 5 elements and `X` must have three columns.

`hougen` is a utility function for `rsmdemo`.

The model form is:

$$\widehat{y} = \frac{\beta_1 x_2 - x_3/\beta_5}{1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3}$$

References

[1] Bates, D. M., and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. Hoboken, NJ: John Wiley & Sons, Inc., 1988.

See Also

`rsmdemo`

Introduced before R2006a

hygecdf

Hypergeometric cumulative distribution function

Syntax

```
hygecdf(x,M,K,N)
hygecdf(x,M,K,N,'upper')
```

Description

`hygecdf(x,M,K,N)` computes the hypergeometric cdf at each of the values in `x` using the corresponding size of the population, `M`, number of items with the desired characteristic in the population, `K`, and number of samples drawn, `N`. Vector or matrix inputs for `x`, `M`, `K`, and `N` must all have the same size. A scalar input is expanded to a constant matrix with the same dimensions as the other inputs.

`hygecdf(x,M,K,N,'upper')` returns the complement of the hypergeometric cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The hypergeometric cdf is

$$p = F(x \mid M,K,N) = \sum_{i=0}^{x} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$$

The result, p, is the probability of drawing up to x of a possible K items in N drawings without replacement from a group of M objects.

Examples

Compute Hypergeometric Distribution CDF

Suppose you have a lot of 100 floppy disks and you know that 20 of them are defective. What is the probability of drawing zero to two defective floppies if you select 10 at random?

```
p = hygecdf(2,100,20,10)
```

```
p = 0.6812
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `hygeinv` | `hygepdf` | `hygernd` | `hygestat`

Introduced before R2006a

hygeinv

Hypergeometric inverse cumulative distribution function

Syntax

```
hygeinv(P,M,K,N)
```

Description

`hygeinv(P,M,K,N)` returns the smallest integer `X` such that the hypergeometric cdf evaluated at `X` equals or exceeds `P`. You can think of `P` as the probability of observing `X` defective items in `N` drawings without replacement from a group of `M` items where `K` are defective.

Examples

Suppose you are the Quality Assurance manager for a floppy disk manufacturer. The production line turns out floppy disks in batches of 1,000. You want to sample 50 disks from each batch to see if they have defects. You want to accept 99% of the batches if there are no more than 10 defective disks in the batch. What is the maximum number of defective disks should you allow in your sample of 50?

```
x = hygeinv(0.99,1000,10,50)
x =
   3
```

What is the median number of defective floppy disks in samples of 50 disks from batches with 10 defective disks?

```
x = hygeinv(0.50,1000,10,50)
x =
   0
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`hygecdf` | `hygepdf` | `hygernd` | `hygestat` | `icdf`

Introduced before R2006a

hygepdf

Hypergeometric probability density function

Syntax

```
Y = hygepdf(X,M,K,N)
```

Description

`Y = hygepdf(X,M,K,N)` computes the hypergeometric pdf at each of the values in `X` using the corresponding size of the population, `M`, number of items with the desired characteristic in the population, `K`, and number of samples drawn, `N`. `X`, `M`, `K`, and `N` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs.

The parameters in `M`, `K`, and `N` must all be positive integers, with `N` ≤ `M`. The values in `X` must be less than or equal to all the parameter values.

The hypergeometric pdf is

$$y = f(x \mid M, K, N) = \frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

The result, y, is the probability of drawing exactly x of a possible K items in n drawings without replacement from a group of M objects.

Examples

Suppose you have a lot of 100 floppy disks and you know that 20 of them are defective. What is the probability of drawing 0 through 5 defective floppy disks if you select 10 at random?

```
p = hygepdf(0:5,100,20,10)
p =
  0.0951  0.2679  0.3182  0.2092  0.0841  0.0215
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`hygecdf` | `hygeinv` | `hygernd` | `hygestat` | `pdf`

Introduced before R2006a

hygernd

Hypergeometric random numbers

Syntax

```
R = hygernd(M,K,N)
R = hygernd(M,K,N,m,n,...)
R = hygernd(M,K,N,[m,n,...])
```

Description

`R = hygernd(M,K,N)` generates random numbers from the hypergeometric distribution with corresponding size of the population, `M`, number of items with the desired characteristic in the population, `K`, and number of samples drawn, `N`. `M`, `K`, and `N` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `R`. A scalar input for `M`, `K`, or `N` is expanded to a constant array with the same dimensions as the other inputs.

`R = hygernd(M,K,N,m,n,...)` or `R = hygernd(M,K,N,[m,n,...])` generates an `m`-by-`n`-by-... array. The `M`, `K`, `N` parameters can each be scalars or arrays of the same size as `R`.

Examples

```
numbers = hygernd(1000,40,50)
numbers =
   1
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`hygecdf` | `hygeinv` | `hygepdf` | `hygestat` | `random`

Introduced before R2006a

hygestat

Hypergeometric mean and variance

Syntax

```
[MN,V] = hygestat(M,K,N)
```

Description

`[MN,V] = hygestat(M,K,N)` returns the mean of and variance for the hypergeometric distribution with corresponding size of the population, `M`, number of items with the desired characteristic in the population, `K`, and number of samples drawn, `N`. Vector or matrix inputs for `M`, `K`, and `N` must have the same size, which is also the size of `MN` and `V`. A scalar input for `M`, `K`, or `N` is expanded to a constant matrix with the same dimensions as the other inputs.

The mean of the hypergeometric distribution with parameters `M`, `K`, and `N` is `NK/M`, and the variance is `NK(M-K)(M-N)/[M^2(M-1)]`.

Examples

The hypergeometric distribution approaches the binomial distribution, where `p = K/M`, as `M` goes to infinity.

```
[m,v] = hygestat(10.^(1:4),10.^(0:3),9)
m =
  0.9000  0.9000  0.9000  0.9000
v =
  0.0900  0.7445  0.8035  0.8094

[m,v] = binostat(9,0.1)
m =
  0.9000
v =
  0.8100
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`hygecdf` | `hygeinv` | `hygepdf` | `hygernd`

Introduced before R2006a

hyperparameters

Variable descriptions for optimizing a fit function

Syntax

```
VariableDescriptions = hyperparameters(FitFcnName,predictors,response)
VariableDescriptions = hyperparameters(FitFcnName,predictors,response,
LearnerType)
```

Description

`VariableDescriptions = hyperparameters(FitFcnName,predictors,response)` returns the default variables for the given fit function. These are the variables that apply when you set the `OptimizeHyperparameters` name-value pair to `'auto'`.

`VariableDescriptions = hyperparameters(FitFcnName,predictors,response, LearnerType)` returns the variables for an ensemble fit with specified learner type. This syntax applies when `FitFcnName` is `'fitcecoc'`, `'fitcensemble'`, or `'fitrensemble'`.

Examples

Obtain Default Hyperparameters

Obtain the default hyperparameters for the `fitcsvm` classifier.

Load the `ionosphere` data.

```
load ionosphere
```

Obtain the hyperparameters.

```
VariableDescriptions = hyperparameters('fitcsvm',X,Y);
```

Examine all the hyperparameters.

```
for ii = 1:length(VariableDescriptions)
    disp(ii),disp(VariableDescriptions(ii))
end
```

```
     1

  optimizableVariable with properties:

         Name: 'BoxConstraint'
        Range: [1.0000e-03 1000]
         Type: 'real'
    Transform: 'log'
     Optimize: 1

     2
```

```
  optimizableVariable with properties:

        Name: 'KernelScale'
       Range: [1.0000e-03 1000]
        Type: 'real'
   Transform: 'log'
    Optimize: 1

    3

  optimizableVariable with properties:

        Name: 'KernelFunction'
       Range: {'gaussian'  'linear'  'polynomial'}
        Type: 'categorical'
   Transform: 'none'
    Optimize: 0

    4

  optimizableVariable with properties:

        Name: 'PolynomialOrder'
       Range: [2 4]
        Type: 'integer'
   Transform: 'none'
    Optimize: 0

    5

  optimizableVariable with properties:

        Name: 'Standardize'
       Range: {'true'  'false'}
        Type: 'categorical'
   Transform: 'none'
    Optimize: 0
```

Change the `PolynomialOrder` hyperparameter to have a wider range and to be used in an optimization.

```
VariableDescriptions(4).Range = [2,5];
VariableDescriptions(4).Optimize = true;
disp(VariableDescriptions(4))
```

```
  optimizableVariable with properties:

        Name: 'PolynomialOrder'
       Range: [2 5]
        Type: 'integer'
   Transform: 'none'
    Optimize: 1
```

Obtain Ensemble Hyperparameters

Obtain the default hyperparameters for the `fitrensemble` ensemble regression function.

Load the `carsmall` data.

```
load carsmall
```

Use `Horsepower` and `Weight` as predictor variables, and `MPG` as the response variable.

```
X = [Horsepower Weight];
Y = MPG;
```

Obtain the default hyperparameters for a `Tree` learner.

```
VariableDescriptions = hyperparameters('fitrensemble',X,Y,'Tree');
```

Examine all the hyperparameters.

```
for ii = 1:length(VariableDescriptions)
    disp(ii),disp(VariableDescriptions(ii))
end
```

```
     1

  optimizableVariable with properties:

         Name: 'Method'
        Range: {'Bag'  'LSBoost'}
         Type: 'categorical'
    Transform: 'none'
     Optimize: 1

     2

  optimizableVariable with properties:

         Name: 'NumLearningCycles'
        Range: [10 500]
         Type: 'integer'
    Transform: 'log'
     Optimize: 1

     3

  optimizableVariable with properties:

         Name: 'LearnRate'
        Range: [1.0000e-03 1]
         Type: 'real'
    Transform: 'log'
     Optimize: 1

     4

  optimizableVariable with properties:

         Name: 'MinLeafSize'
        Range: [1 50]
         Type: 'integer'
    Transform: 'log'
     Optimize: 1
```

```
    5

  optimizableVariable with properties:

        Name: 'MaxNumSplits'
       Range: [1 99]
        Type: 'integer'
   Transform: 'log'
    Optimize: 0

    6

  optimizableVariable with properties:

        Name: 'NumVariablesToSample'
       Range: [1 2]
        Type: 'integer'
   Transform: 'none'
    Optimize: 0
```

Change the `MaxNumSplits` hyperparameter to have a wider range and to be used in an optimization.

```
VariableDescriptions(5).Range = [1,200];
VariableDescriptions(5).Optimize = true;
disp(VariableDescriptions(5))
```

```
  optimizableVariable with properties:

        Name: 'MaxNumSplits'
       Range: [1 200]
        Type: 'integer'
   Transform: 'log'
    Optimize: 1
```

Input Arguments

`FitFcnName` — Name of fitting function

`'fitcdiscr'` | `'fitcecoc'` | `'fitcensemble'` | `'fitckernel'` | `'fitcknn'` | `'fitclinear'` | `'fitcnb'` | `'fitcsvm'` | `'fitctree'` | `'fitrensemble'` | `'fitrgp'` | `'fitrkernel'` | `'fitrlinear'` | `'fitrsvm'` | `'fitrtree'`

Name of fitting function, specified as one of the listed classification or regression fit function names.

If `FitFcnName` is `'fitcecoc'`, `'fitcensemble'`, or `'fitrensemble'`, then also specify the learner type in the `LearnerType` argument.

Example: `'fitctree'`

`predictors` — Predictor data

matrix with D predictor columns | table with D predictor columns

Predictor data, specified as a matrix with D predictor columns or as a table with D predictor columns, where D is the number of predictors.

Example: `X`

Data Types: `double` | `logical` | `char` | `string` | `table` | `cell` | `categorical` | `datetime`

response — Class labels or numeric response
grouping variable | scalar

Class labels or numeric response, specified as a grouping variable (see “Grouping Variables” on page 2-45) or as a scalar.

Example: `Y`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

LearnerType — Learner type for ensemble fit
`'Discriminant'` | `'Kernel'` | `'KNN'` | `'Linear'` | `'SVM'` | `'Tree'` | template of a listed learner

Learner type for ensemble fit, specified as `'Discriminant'`, `'Kernel'`, `'KNN'`, `'Linear'`, `'SVM'`, `'Tree'`, or a template of a listed learner. Use this argument when `FitFcnName` is `'fitcecoc'`, `'fitcensemble'`, or `'fitrensemble'`.

For `'fitcensemble'` you can use only `'Discriminant'`, `'KNN'`, `'Tree'`, or an associated template.

For `'fitrensemble'`, you can use only `'Tree'` or `templateTree`.

Example: `'Tree'`

Output Arguments

VariableDescriptions — Variable descriptions
vector of `optimizableVariable` objects

Variable descriptions, returned as a vector of `optimizableVariable` objects. The variables have their default parameters set, such as range and variable type. All eligible variables exist in the descriptions, but the variables unused in the `'auto'` setting have their `Optimize` property set to `false`. You can update the variables by using dot notation, as shown in “Examples” on page 33-0 .

See Also
`fitcdiscr` | `fitcecoc` | `fitcensemble` | `fitckernel` | `fitcknn` | `fitclinear` | `fitcnb` | `fitcsvm` | `fitctree` | `fitrensemble` | `fitrgp` | `fitrkernel` | `fitrlinear` | `fitrsvm` | `fitrtree`

Introduced in R2016b

icdf

Package: `prob`

Inverse cumulative distribution function

Syntax

```
x = icdf('name',p,A)
x = icdf('name',p,A,B)
x = icdf('name',p,A,B,C)
x = icdf('name',p,A,B,C,D)

x = icdf(pd,p)
```

Description

`x = icdf('name',p,A)` returns the inverse cumulative distribution function (icdf) for the one-parameter distribution family specified by `'name'` and the distribution parameter `A`, evaluated at the probability values in `p`.

`x = icdf('name',p,A,B)` returns the icdf for the two-parameter distribution family specified by `'name'` and the distribution parameters `A` and `B`, evaluated at the probability values in `p`.

`x = icdf('name',p,A,B,C)` returns the icdf for the three-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, and `C`, evaluated at the probability values in `p`.

`x = icdf('name',p,A,B,C,D)` returns the icdf for the four-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, `C`, and `D`, evaluated at the probability values in `p`.

`x = icdf(pd,p)` returns the icdf function of the probability distribution object `pd`, evaluated at the probability values in `p`.

Examples

Compute the Normal Distribution icdf

Create a standard normal distribution object with the mean, μ, equal to 0 and the standard deviation, σ, equal to 1.

```
mu = 0;
sigma = 1;
pd = makedist('Normal','mu',mu,'sigma',sigma);
```

Define the input vector p to contain the probability values at which to calculate the icdf.

```
p = [0.1,0.25,0.5,0.75,0.9];
```

Compute the icdf values for the standard normal distribution at the values in p.

```
x = icdf(pd,p)
```

```
x = 1×5

   -1.2816   -0.6745         0    0.6745    1.2816
```

Each value in *x* corresponds to a value in the input vector *p*. For example, at the value *p* equal to 0.9, the corresponding icdf value *x* is equal to 1.2816.

Alternatively, you can compute the same icdf values without creating a probability distribution object. Use the `icdf` function and specify a standard normal distribution using the same parameter values for μ and σ.

```
x2 = icdf('Normal',p,mu,sigma)
```

```
x2 = 1×5

   -1.2816   -0.6745         0    0.6745    1.2816
```

The icdf values are the same as those computed using the probability distribution object.

Compute the Poisson Distribution icdf

Create a Poisson distribution object with the rate parameter, λ, equal to 2.

```
lambda = 2;
pd = makedist('Poisson','lambda',lambda);
```

Define the input vector *p* to contain the probability values at which to calculate the icdf.

```
p = [0.1,0.25,0.5,0.75,0.9];
```

Compute the icdf values for the Poisson distribution at the values in *p*.

```
x = icdf(pd,p)
```

```
x = 1×5

     0     1     2     3     4
```

Each value in *x* corresponds to a value in the input vector *p*. For example, at the value *p* equal to 0.9, the corresponding icdf value *x* is equal to 4.

Alternatively, you can compute the same icdf values without creating a probability distribution object. Use the `icdf` function and specify a Poisson distribution using the same value for the rate parameter λ.

```
x2 = icdf('Poisson',p,lambda)
```

```
x2 = 1×5

     0     1     2     3     4
```

The icdf values are the same as those computed using the probability distribution object.

Compute Standard Normal Critical Values

Create a standard normal distribution object.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

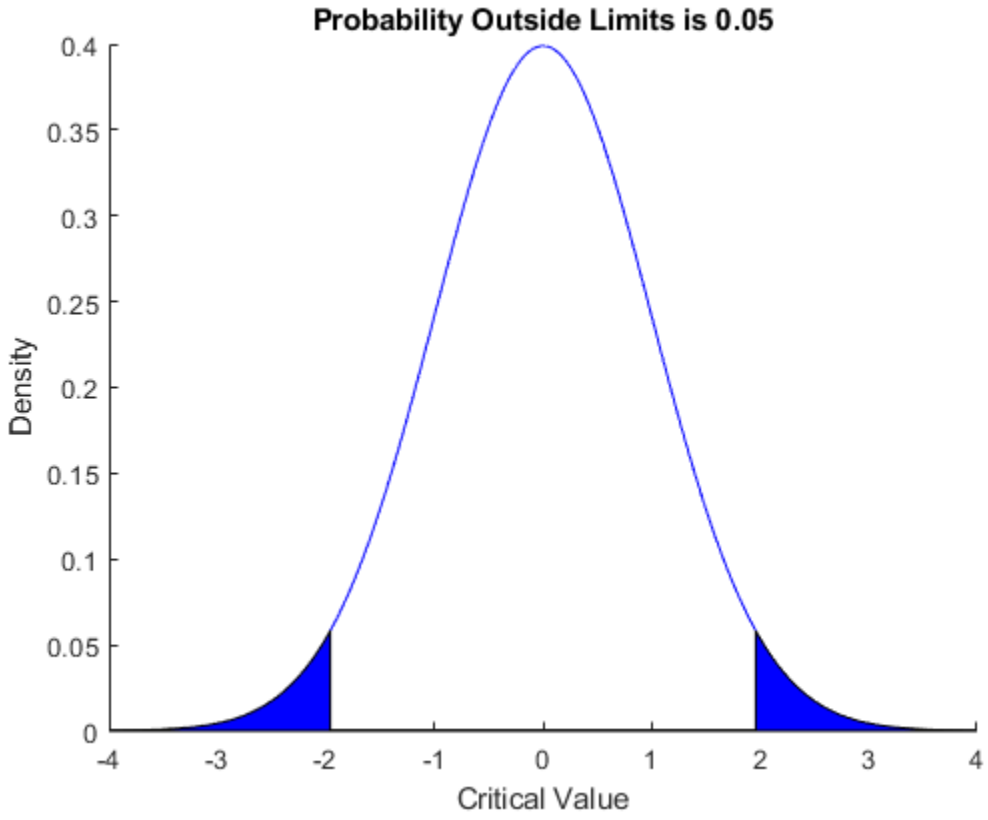
Determine the critical values at the 5% significance level for a test statistic with a standard normal distribution, by computing the upper and lower 2.5% values.

```
x = icdf(pd,[.025,.975])
```

```
x = 1×2

   -1.9600    1.9600
```

Plot the cdf and shade the critical regions.

```
p = normspec(x,0,1,'outside')
```




```
p = 0.0500
```

Input Arguments

`'name'` — Probability distribution name

character vector or string scalar of probability distribution name

Probability distribution name, specified as one of the probability distribution names in this table.

`'name'`	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
`'Beta'`	“Beta Distribution” on page B-6	a first shape parameter	b second shape parameter	—	—
`'Binomial'`	“Binomial Distribution” on page B-10	n number of trials	p probability of success for each trial	—	—
`'BirnbaumSaunders'`	“Birnbaum-Saunders Distribution” on page B-18	β scale parameter	γ shape parameter	—	—
`'Burr'`	“Burr Type XII Distribution” on page B-19	α scale parameter	c first shape parameter	k second shape parameter	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Chisquare'	“Chi-Square Distribution” on page B-28	ν degrees of freedom	—	—	—
'Exponential'	“Exponential Distribution” on page B-33	μ mean	—	—	—
'Extreme Value'	“Extreme Value Distribution” on page B-40	μ location parameter	σ scale parameter	—	—
'F'	“F Distribution” on page B-45	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	—	—
'Gamma'	“Gamma Distribution” on page B-47	a shape parameter	b scale parameter	—	—
'Generalized Extreme Value'	“Generalized Extreme Value Distribution” on page B-55	k shape parameter	σ scale parameter	μ location parameter	—
'Generalized Pareto'	“Generalized Pareto Distribution” on page B-59	k tail index (shape) parameter	σ scale parameter	μ threshold (location) parameter	—
'Geometric'	“Geometric Distribution” on page B-63	p probability parameter	—	—	—
'HalfNormal'	“Half-Normal Distribution” on page B-68	μ location parameter	σ scale parameter	—	—
'Hypergeometric'	“Hypergeometric Distribution” on page B-73	m size of the population	k number of items with the desired characteristic in the population	n number of samples drawn	—
'InverseGaussian'	“Inverse Gaussian Distribution” on page B-75	μ scale parameter	λ shape parameter	—	—
'Logistic'	“Logistic Distribution” on page B-85	μ mean	σ scale parameter	—	—
'LogLogistic'	“Loglogistic Distribution” on page B-86	μ mean of logarithmic values	σ scale parameter of logarithmic values	—	—
'Lognormal'	“Lognormal Distribution” on page B-88	μ mean of logarithmic values	σ standard deviation of logarithmic values	—	—
'Nakagami'	“Nakagami Distribution” on page B-108	μ shape parameter	ω scale parameter	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Negative Binomial'	"Negative Binomial Distribution" on page B-109	r number of successes	p probability of success in a single trial	—	—
'Noncentral F'	"Noncentral F Distribution" on page B-115	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	δ noncentrality parameter	—
'Noncentral t'	"Noncentral t Distribution" on page B-117	ν degrees of freedom	δ noncentrality parameter	—	—
'Noncentral Chi-square'	"Noncentral Chi-Square Distribution" on page B-113	ν degrees of freedom	δ noncentrality parameter	—	—
'Normal'	"Normal Distribution" on page B-119	μ mean	σ standard deviation	—	—
'Poisson'	"Poisson Distribution" on page B-131	λ mean	—	—	—
'Rayleigh'	"Rayleigh Distribution" on page B-137	b scale parameter	—	—	—
'Rician'	"Rician Distribution" on page B-139	s noncentrality parameter	σ scale parameter	—	—
'Stable'	"Stable Distribution" on page B-140	α first shape parameter	β second shape parameter	γ scale parameter	δ location parameter
'T'	"Student's t Distribution" on page B-149	ν degrees of freedom	—	—	—
'tLocationScale'	"t Location-Scale Distribution" on page B-156	μ location parameter	σ scale parameter	ν shape parameter	—
'Uniform'	"Uniform Distribution (Continuous)" on page B-163	a lower endpoint (minimum)	b upper endpoint (maximum)	—	—
'Discrete Uniform'	"Uniform Distribution (Discrete)" on page B-168	n maximum observable value	—	—	—
'Weibull'	"Weibull Distribution" on page B-170	a scale parameter	b shape parameter	—	—

Example: 'Normal'

p — Probability values at which to evaluate icdf
scalar value | array of scalar values

Probability values at which to evaluate the icdf, specified as a scalar value, or an array of scalar values in the range [0,1].

If one or more of the input arguments `p`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `icdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Example: `[0.1,0.25,0.5,0.75,0.9]`

Data Types: `single` | `double`

`A` — First probability distribution parameter
scalar value | array of scalar values

First probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `p`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `icdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`B` — Second probability distribution parameter
scalar value | array of scalar values

Second probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `p`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `icdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`C` — Third probability distribution parameter
scalar value | array of scalar values

Third probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `p`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `icdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`D` — Fourth probability distribution parameter
scalar value | array of scalar values

Fourth probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `p`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `icdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`pd` — Probability distribution
probability distribution object

Probability distribution, specified as a probability distribution object created with a function or app in this table.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.
`paretotails`	Create a piecewise distribution object that has generalized Pareto distributions in the tails.

Output Arguments

x — icdf values

scalar value | array of scalar values

icdf values, returned as a scalar value or an array of scalar values. `x` is the same size as `p` after any necessary scalar expansion. Each element in `x` is the icdf value of the distribution, specified by the corresponding elements in the distribution parameters (`A`, `B`, `C`, and `D`) or specified by the probability distribution object (`pd`), evaluated at the corresponding element in `p`.

Alternative Functionality

`icdf` is a generic function that accepts either a distribution by its name `'name'` or a probability distribution object `pd`. It is faster to use a distribution-specific function, such as `norminv` for the normal distribution and `binoinv` for the binomial distribution. For a list of distribution-specific functions, see “Supported Distributions” on page 5-14.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `'name'` must be a compile-time constant. For example, to use the normal distribution, include `coder.Constant('Normal')` in the `-args` value of `codegen`.
- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `cdf` | `fitdist` | `makedist` | `mle` | `paretotails` | `pdf` | `random`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

inconsistent

Inconsistency coefficient

Syntax

```
Y = inconsistent(Z)
Y = inconsistent(Z,d)
```

Description

`Y = inconsistent(Z)` returns the inconsistency coefficient for each link of the hierarchical cluster tree `Z` generated by the `linkage` function. `inconsistent` calculates the inconsistency coefficient for each link by comparing its height with the average height of other links at the same level of the hierarchy. The larger the coefficient, the greater the difference between the objects connected by the link. For more information, see “Algorithms” on page 33-2867.

`Y = inconsistent(Z,d)` returns the inconsistency coefficient for each link in the tree `Z` by searching to a depth `d` below each link.

Examples

Inconsistency Coefficient Calculation

Examine an inconsistency coefficient calculation for a hierarchical cluster tree.

Load the `examgrades` data set.

```
load examgrades
```

Create a hierarchical cluster tree.

```
Z = linkage(grades);
```

Create a matrix of inconsistency coefficient information using `inconsistent`. Examine the information for the 84th link.

```
Y = inconsistent(Z);
Y(84,:)
```

```
ans = 1×4

    7.2741    0.3624    3.0000    0.5774
```

The fourth column of `Y` contains the inconsistency coefficient, which is computed using the mean in the first column of `Y` and the standard deviation in the second column of `Y`.

Because the rows of `Y` correspond to the rows of `Z`, examine the 84th link in `Z`.

```
Z(84,:)
```

```
ans = 1×3

  190.0000  203.0000    7.4833
```

The 84th link connects the 190th and 203rd clusters in the tree and has a height of `7.4833`. The 190th cluster corresponds to the link of index 190 − 120 = 70, where 120 is the number of observations. The 203rd cluster corresponds to the 83rd link.

By default, `inconsistent` uses two levels of the tree to compute `Y`. Therefore, it uses only the 70th, 83rd, and 84th links to compute the inconsistency coefficient for the 84th link. Compare the values in `Y(84,:)` with the corresponding computations by using the link heights in `Z`.

```
mean84 = mean([Z(70,3) Z(83,3) Z(84,3)])
```

```
mean84 = 7.2741
```

```
std84 = std([Z(70,3) Z(83,3) Z(84,3)])
```

```
std84 = 0.3624
```

```
inconsistent84 = (Z(84,3)-mean84)/std84
```

```
inconsistent84 = 0.5774
```

Compute Inconsistency Coefficient

Create the sample data.

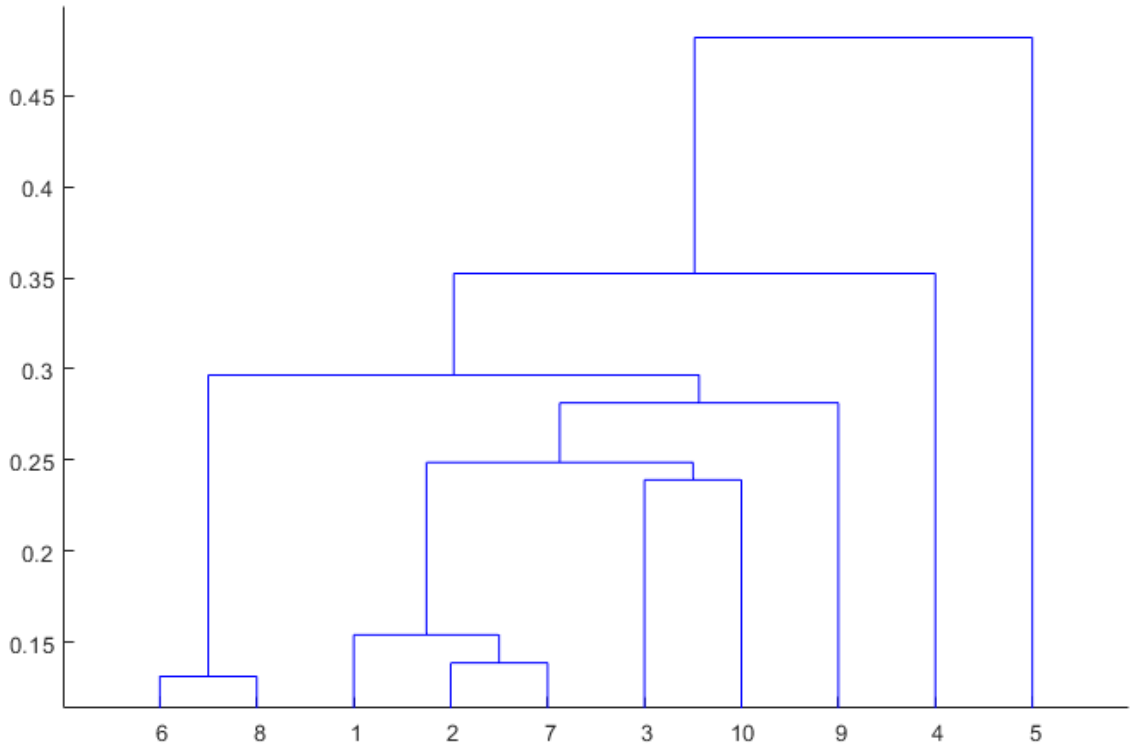
```
X = gallery('uniformdata',[10 2],12);
Y = pdist(X);
```

Generate the hierarchical cluster tree.

```
Z = linkage(Y,'single');
```

Generate a dendrogram plot of the hierarchical cluster tree.

```
dendrogram(Z)
```



Compute the inconsistency coefficient for each link in the cluster tree Z to depth 3.

```
W = inconsistent(Z,3)
```

```
W = 9×4

    0.1313         0    1.0000         0
    0.1386         0    1.0000         0
    0.1463    0.0109    2.0000    0.7071
    0.2391         0    1.0000         0
    0.1951    0.0568    4.0000    0.9425
    0.2308    0.0543    4.0000    0.9320
    0.2395    0.0748    4.0000    0.7636
    0.2654    0.0945    4.0000    0.9203
    0.3769    0.0950    3.0000    1.1040
```

Input Arguments

Z — Agglomerative hierarchical cluster tree

numeric matrix

Agglomerative hierarchical cluster tree, specified as a numeric matrix returned by `linkage`. Z is an (m – 1)-by-3 matrix, where m is the number of observations. Columns 1 and 2 of Z contain cluster

indices linked in pairs to form a binary tree. `Z(I,3)` contains the linkage distances between the two clusters merged in row `Z(I,:)`.

Data Types: `single` | `double`

d — Depth
2 (default) | positive integer scalar

Depth, specified as a positive integer scalar. For each link k, `inconsistent` calculates the corresponding inconsistency coefficient using all the links in the tree within `d` levels below k.

Data Types: `single` | `double`

Output Arguments

Y — Inconsistency coefficient information
numeric matrix

Inconsistency coefficient information, returned as an (m – 1)-by-4 matrix, where the (m – 1) rows correspond to the rows of `Z`. This table describes the columns of `Y`.

Column	Description
1	Mean of the heights of all the links included in the calculation
2	Standard deviation of the heights of all the links included in the calculation
3	Number of links included in the calculation
4	Inconsistency coefficient

Data Types: `double`

Algorithms

For each link k, the inconsistency coefficient is calculated as

$$Y(k,4) = (Z(k,3) - Y(k,1))/Y(k,2),$$

where `Y` is the inconsistency coefficient information for links in the hierarchical cluster tree `Z`.

For links that have no further links below them, the inconsistency coefficient is set to 0.

References

[1] Jain, A., and R. Dubes. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

[2] Zahn, C. T. “Graph-theoretical methods for detecting and describing Gestalt clusters.” *IEEE Transactions on Computers*. Vol. C-20, Issue 1, 1971, pp. 68–86.

See Also

`cluster` | `clusterdata` | `cophenet` | `dendrogram` | `linkage` | `pdist` | `squareform`

Topics
“Hierarchical Clustering” on page 16-6

Introduced before R2006a

increaseB

Class: `clustering.evaluation.GapEvaluation`
Package: `clustering.evaluation`

Increase reference data sets

Syntax

```
eva_out = increaseB(eva,nref)
```

Description

`eva_out = increaseB(eva,nref)` returns a gap criterion clustering evaluation object `eva_out` that uses the same evaluation criteria as the input object `eva` and an additional number of reference data sets as specified by `nref`.

Input Arguments

eva — Clustering evaluation data

clustering evaluation object

Clustering evaluation data, specified as a clustering evaluation object. Create a clustering evaluation object using `evalclusters`.

nref — Number of additional reference data sets

positive integer value

Number of additional reference data sets, specified as a positive integer value.

Output Arguments

eva_out — Updated clustering evaluation data

clustering evaluation object

Updated clustering evaluation data, returned as a gap criterion clustering evaluation object. `eva_out` contains evaluation data obtained using the reference data sets from the input object `eva` plus a number of additional reference data sets as specified in `nref`.

`increaseB` updates the `B` property of the input object `eva` to reflect the increase in the number of reference data sets used to compute the gap criterion values. `increaseB` also updates the `CriterionValues` property with gap criterion values computed using the total number of reference data sets. `increaseB` might also update the `OptimalK` and `OptimalY` properties to reflect the optimal number of clusters and optimal clustering solution as determined using the total number of reference data sets. Additionally, `increaseB` might also update the `LogW`, `ExpectedLogW`, `StdLogW`, and `SE` properties.

Examples

Evaluate Clustering Solutions Using Additional Reference Data

Create a gap clustering evaluation object using `evalclusters`, then use `increaseB` to increase the number of reference data sets used to compute the gap criterion values.

Load the sample data.

```
load fisheriris
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Cluster the flower measurement data using `kmeans`, and use the gap criterion to evaluate proposed solutions of one through five clusters. Use 50 reference data sets.

```
rng('default') % For reproducibility
eva = evalclusters(meas,'kmeans','gap','klist',1:5,'B',50)
```

```
eva =
  GapEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5]
    CriterionValues: [0.0870 0.5822 0.8766 1.0007 1.0465]
           OptimalK: 4
```

The clustering evaluation object `eva` contains data on each proposed clustering solution. The returned results indicate that the optimal number of clusters is four.

The value of the `B` property of `eva` shows 50 reference data sets.

```
eva.B
```

```
ans = 50
```

Increase the number of reference data sets by 100, for a total of 150 sets.

```
eva = increaseB(eva,100)
```

```
eva =
  GapEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5]
    CriterionValues: [0.0794 0.5850 0.8738 1.0034 1.0508]
           OptimalK: 5
```

The returned results now indicate that the optimal number of clusters is five.

The value of the `B` property of `eva` now shows 150 reference data sets.

```
eva.B
```

```
ans = 150
```

See Also

`evalclusters`

interactionplot

Interaction plot for grouped data

Syntax

```
interactionplot(Y,GROUP)
interactionplot(Y,GROUP,'varnames',VARNAMES)
[h,AX,bigax] = interactionplot(...)
```

Description

`interactionplot(Y,GROUP)` displays the two-factor interaction plot for the group means of matrix `Y` with groups defined by entries in `GROUP`, which can be a cell array or a matrix. `Y` is a numeric matrix or vector. If `Y` is a matrix, the rows represent different observations and the columns represent replications of each observation. If `Y` is a vector, the rows give the means of each entry in `GROUP`. If `GROUP` is a cell array, then each cell of `GROUP` must contain a grouping variable that is a categorical variable, numeric vector, character matrix, string array, or single-column cell array of character vectors. If `GROUP` is a matrix, then its columns represent different grouping variables. Each grouping variable must have the same number of rows as `Y`. The number of grouping variables must be greater than 1.

The interaction plot is a matrix plot, with the number of rows and columns both equal to the number of grouping variables. The grouping variable names are printed on the diagonal of the plot matrix. The plot at off-diagonal position (*i*,*j*) is the interaction of the two variables whose names are given at row diagonal (*i*,*i*) and column diagonal (*j*,*j*), respectively.

`interactionplot(Y,GROUP,'varnames',VARNAMES)` displays the interaction plot with user-specified grouping variable names `VARNAMES`. `VARNAMES` is a character matrix, a string array, or a cell array of character vectors, one per grouping variable. Default names are `'X1'`, `'X2'`, ...

`[h,AX,bigax] = interactionplot(...)` returns a handle `h` to the figure window, a matrix `AX` of handles to the subplot axes, and a handle `bigax` to the big (invisible) axes framing the subplots.

Examples

Display Interaction Plots

Randomly generate data for a response variable `y` .

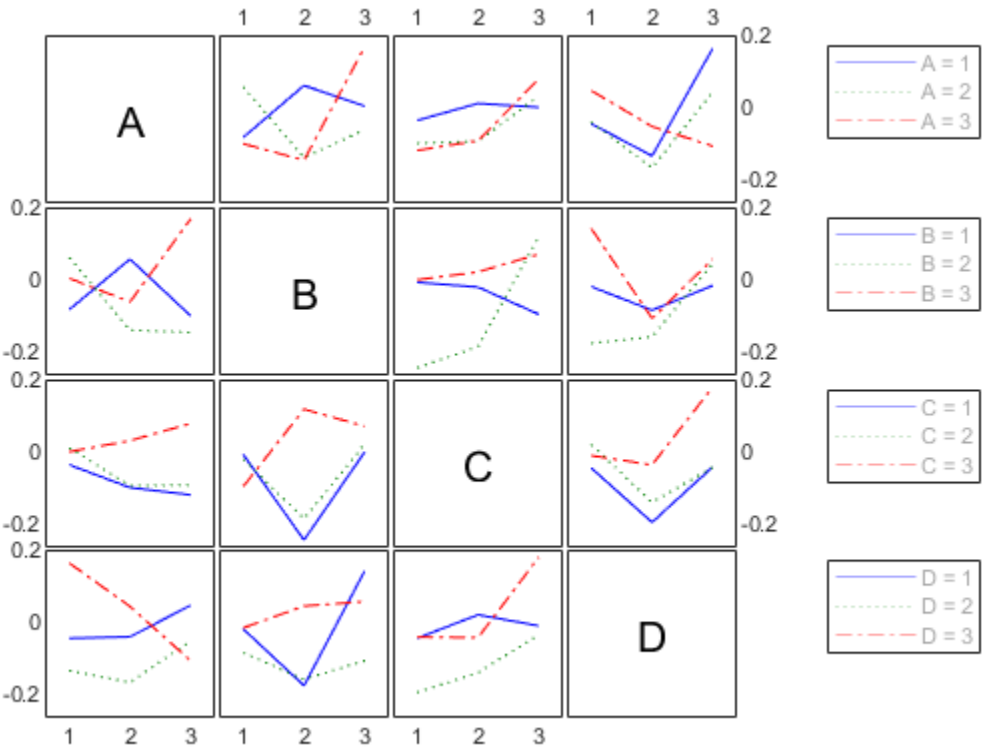
```
rng default;     % For reproducibility
y = randn(1000,1);
```

Randomly generate data for four three-level factors.

```
group = ceil(3*rand(1000,4));
```

Display the interaction plots for the factors and name the factors 'A', 'B', 'C', 'D'.

```
interactionplot(y,group,'varnames',{'A','B','C','D'})
```



See Also

`maineffectsplot` | `multivarichart`

Introduced in R2006b

intersect

Class: `dataset`

(Not Recommended) Set intersection for dataset array observations

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = intersect(A,B)
C = intersect(A,B,vars)
C = intersect(A,B,vars,setOrder)
[C,iA,iB] = intersect( ___ )
```

Description

`C = intersect(A,B)` for `dataset` arrays `A` and `B` returns the common set of observations from the two arrays, with repetitions removed. The observations in the dataset array `C` are in sorted order.

`C = intersect(A,B,vars)` returns the set of common observations from the two arrays, considering only the variables specified in `vars`, with repetitions removed. The observations in the dataset array `C` are sorted by those variables.

The values for variables not specified in `vars` for each observation in `C` are taken from the corresponding observations in `A`. If there are multiple observations in `A` that correspond to an observation in `C`, then those values are taken from the first occurrence.

`C = intersect(A,B,vars,setOrder)` returns the observations in `C` in the order specified by `setOrder`.

`[C,iA,iB] = intersect( ___ )` also returns index vectors `iA` and `iB` such that `C = A(iA,:)` and `C = B(iB,:)`. If there are repeated observations in `A` or `B`, then `intersect` returns the index of the first occurrence. You can use any of the previous input arguments.

Input Arguments

`A,B`

Input dataset arrays.

`vars`

String array or cell array of character vectors containing variable names, or a vector of integers containing variable column numbers. `vars` indicates the variables in `A` and `B` that `intersect` considers.

Specify `vars` as `[]` to use its default value of all variables.

`setOrder`

Flag indicating the sorting order for the observations in `C`. The possible values of `setOrder` are:

`'sorted'`	Observations in `C` are in sorted order (default).
`'stable'`	Observations in `C` are in the same order that they appear in `A`.

Output Arguments

`C`

Dataset array with the common set of observations in `A` and `B`, with repetitions removed. `C` is in sorted order (by default), or the order specified by `setOrder`.

`iA`

Index vector, indicating the observations in `A` that are common to `B`. The vector `iA` contains the index to the first occurrence of any repeated observations in `A`.

`iB`

Index vector, indicating the observations in `B` that are common to `A`. The vector `iB` contains the index to the first occurrence of any repeated observations in `B`.

Examples

Intersection of Two Dataset Arrays

Load sample data.

```
A = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'));
B = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),'Sh
```

Return the intersection and index vectors.

```
[C,iA,iB] = intersect(A,B);
```

```
C =

    id              name          sex        age    wgt    smoke
    'TRW-072'       'WHITE'       'm'        39     202    1
```

There is one observation in common between `A` and `B`.

Find the observation in the original dataset arrays.

```
A(iA,:)
```

```
ans =

    id              name          sex        age    wgt    smoke
    'TRW-072'       'WHITE'       'm'        39     202    1
```

```
B(iB,:)
```

```
ans =

    id              name          sex        age    wgt    smoke
    'TRW-072'       'WHITE'       'm'        39     202    1
```

See Also

dataset | ismember | setdiff | setxor | sortrows | union | unique

Topics

“Merge Dataset Arrays” on page 2-85
“Dataset Arrays” on page 2-112

invpred

Inverse prediction

Syntax

```
X0 = invpred(X,Y,Y0)
[X0,DXLO,DXUP] = invpred(X,Y,Y0)
[X0,DXLO,DXUP] = invpred(X,Y,Y0,name1,val1,name2,val2,...)
```

Description

`X0 = invpred(X,Y,Y0)` accepts vectors `X` and `Y` of the same length, fits a simple regression, and returns the estimated value `X0` for which the height of the line is equal to `Y0`. The output, `X0`, has the same size as `Y0`, and `Y0` can be an array of any size.

`[X0,DXLO,DXUP] = invpred(X,Y,Y0)` also computes 95% inverse prediction intervals. `DXLO` and `DXUP` define intervals with lower bound `X0–DXLO` and upper bound `X0+DXUP`. Both `DXLO` and `DXUP` have the same size as `Y0`.

The intervals are not simultaneous and are not necessarily finite. Some intervals may extend from a finite value to `-Inf` or `+Inf`, and some may extend over the entire real line.

`[X0,DXLO,DXUP] = invpred(X,Y,Y0,name1,val1,name2,val2,...)` specifies optional argument name/value pairs chosen from the following list. Argument names are case insensitive and partial matches are allowed.

Name	Value
`'alpha'`	A value between 0 and 1 specifying a confidence level of `100*(1-alpha)`%. Default is `alpha=0.05` for 95% confidence.
`'predopt'`	Either `'observation'`, the default value to compute the intervals for `X0` at which a new observation could equal `Y0`, or `'curve'` to compute intervals for the `X0` value at which the curve is equal to `Y0`.

Examples

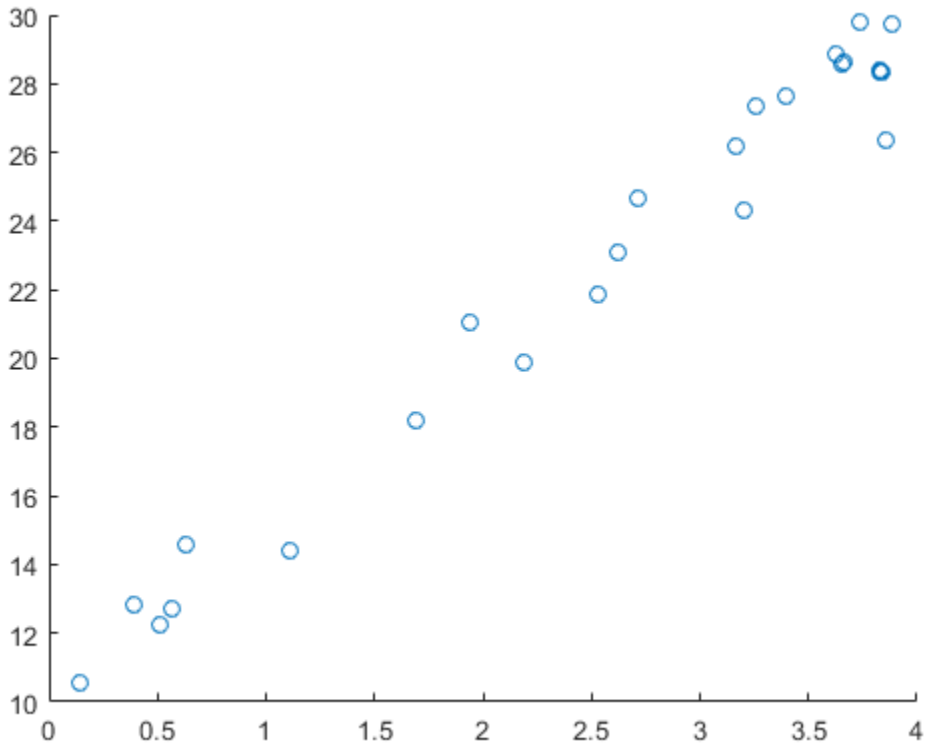
Inverse Prediction

Generate sample data.

```
x = 4*rand(25,1);
y = 10 + 5*x + randn(size(x));
```

Make a scatterplot of the data.

```
scatter(x,y)
```



Predict the *x* value for a given *y* value of 20.

```
x0 = invpred(x,y,20)
```

```
x0 = 1.9967
```

See Also

`polyconf` | `polyfit` | `polytool` | `polyval`

Introduced before R2006a

iqr

Package: `prob`

Interquartile range

Syntax

```
r = iqr(x)
r = iqr(x,'all')
r = iqr(x,dim)
r = iqr(x,vecdim)

r = iqr(pd)
```

Description

`r = iqr(x)` returns the interquartile range of the values in `x`.

- If `x` is a vector, then `r` is the difference between the 75th and the 25th percentiles of the data contained in `x`.
- If `x` is a matrix, then `r` is a row vector containing the difference between the 75th and the 25th percentiles of the sample data in each column of `x`.
- If `x` is a multidimensional array, then `iqr` operates along the first nonsingleton dimension of `x`. The size of this dimension becomes 1 while the sizes of all other dimensions remain the same.

`r = iqr(x,'all')` returns the interquartile range of all the values in `x`.

`r = iqr(x,dim)` returns the interquartile range along the dimension of `x` specified by `dim`.

`r = iqr(x,vecdim)` returns the interquartile range over the dimensions specified by `vecdim`. For example, if `x` is a matrix, then `iqr(x,[1 2])` is the interquartile range of all the elements of `x` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

`r = iqr(pd)` returns the interquartile range of the probability distribution `pd`.

Examples

Compute the Interquartile Range

Generate a 4-by-4 matrix of random data from a normal distribution with parameter values μ equal to 10 and σ equal to 1.

```
rng default  % For reproducibility
x = normrnd(10,1,4)
```

```
x = 4×4

   10.5377   10.3188   13.5784   10.7254
   11.8339    8.6923   12.7694    9.9369
```

```
    7.7412    9.5664    8.6501   10.7147
   10.8622   10.3426   13.0349    9.7950
```

Compute the interquartile range for each column of data.

```
r = iqr(x)
```

```
r = 1×4

    2.2086    1.2013    2.5969    0.8541
```

Compute the interquartile range for each row of data.

```
r2 = iqr(x,2)
```

```
r2 = 4×1

    1.7237
    2.9870
    1.9449
    1.8797
```

Compute Interquartile Range of Multidimensional Array

Compute the interquartile range of a multidimensional array over multiple dimensions by specifying the `'all'` and `vecdim` input arguments.

Create a 3-by-4-by-2 array `X`.

```
X = reshape(1:24,[3 4 2])
```

```
X =
X(:,:,1) =

     1     4     7    10
     2     5     8    11
     3     6     9    12


X(:,:,2) =

    13    16    19    22
    14    17    20    23
    15    18    21    24
```

Compute the interquartile range of all the values in `X`.

```
rall = iqr(X,'all')
```

```
rall = 12
```

Compute the interquartile range of each page of `X`. Specify the first and second dimensions as the operating dimensions along which the interquartile range is calculated.

```
rpage = iqr(X,[1 2])
```

```
rpage =
rpage(:,:,1) =

     6


rpage(:,:,2) =

     6
```

For example, `rpage(1,1,1)` is the interquartile range of all the elements in `X(:,:,1)`.

Compute the interquartile range of the elements in each `X(i,:,:)` slice by specifying the second and third dimensions as the operating dimensions.

```
rrow = iqr(X,[2 3])
```

```
rrow = 3×1

    12
    12
    12
```

For example, `rrow(3)` is the interquartile range of all the elements in `X(3,:,:)`.

Compute the Normal Distribution Interquartile Range

Create a standard normal distribution object with the mean, μ, equal to 0 and the standard deviation, σ, equal to 1.

```
pd = makedist('Normal','mu',0,'sigma',1);
```

Compute the interquartile range of the standard normal distribution.

```
r = iqr(pd)
```

```
r = 1.3490
```

The returned value is the difference between the 75th and the 25th percentile values for the distribution. This is equivalent to computing the difference between the inverse cumulative distribution function (icdf) values at the probabilities y equal to 0.75 and 0.25.

```
r2 = icdf(pd,0.75) - icdf(pd,0.25)
```

```
r2 = 1.3490
```

Interquartile Range of a Fitted Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades;
x = grades(:,1);
```

Create a normal distribution object by fitting it to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the interquartile range of the fitted distribution.

```
r = iqr(pd)
```

```
r = 11.7634
```

The returned result indicates that the difference between the 75th and 25th percentile of the students' grades is 11.7634.

Use `icdf` to determine the 75th and 25th percentiles of the students' grades.

```
y = icdf(pd,[0.25,0.75])
```

```
y = 1×2

   69.1266   80.8900
```

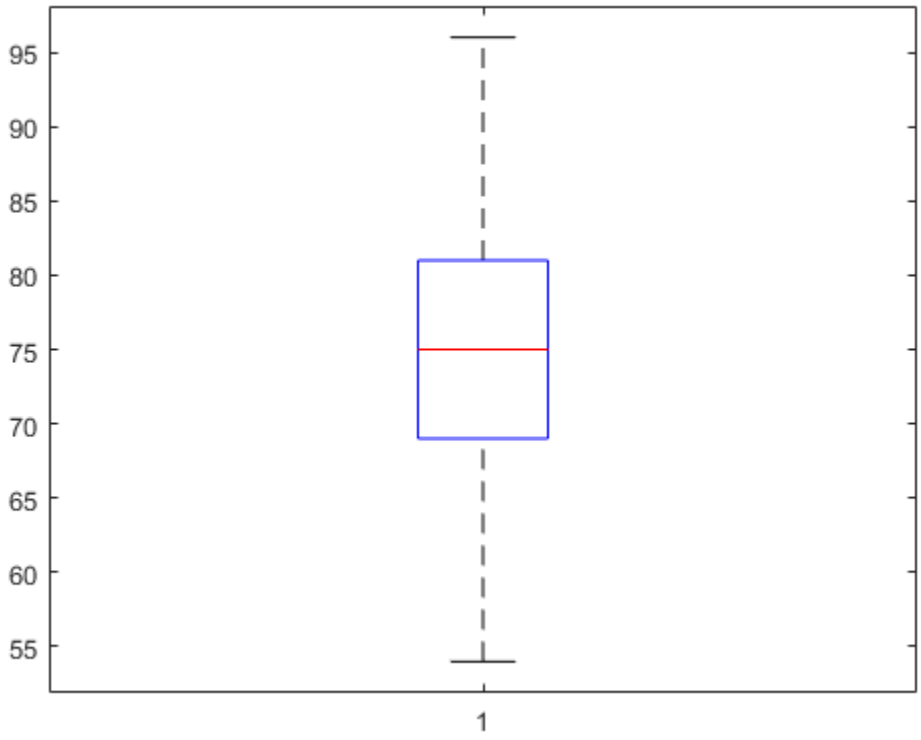
Calculate the difference between the 75th and 25th percentiles. This yields the same result as `iqr`.

```
y(2)-y(1)
```

```
ans = 11.7634
```

Use `boxplot` to visualize the interquartile range.

```
boxplot(x)
```



The top line of the box shows the 75th percentile, and the bottom line shows the 25th percentile. The center line shows the median, which is the 50th percentile.

Input Arguments

`x` — Input array

vector | matrix | multidimensional array

Input array, specified as a vector, matrix, or multidimensional array.

Data Types: `single` | `double`

`dim` — Dimension

positive integer value

Dimension along which the interquartile range is calculated, specified as a positive integer. For example, for a matrix `x`, when `dim` is equal to 1, `iqr` returns the interquartile range for the columns of `x`. When `dim` is equal to 2, `iqr` returns the interquartile range for the rows of `x`. For *n*-dimensional arrays, `iqr` operates along the first nonsingleton dimension of `x`.

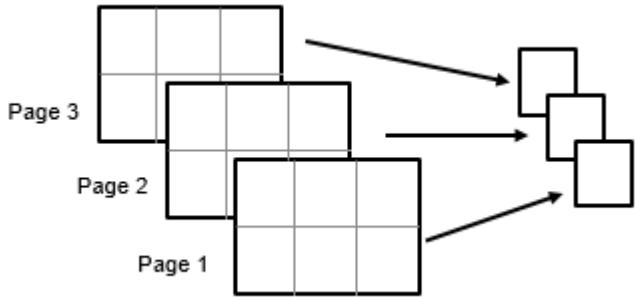
Data Types: `single` | `double`

`vecdim` — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `x`. The output `r` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `x` and `r`.

For example, if `x` is a 2-by-3-by-3 array, then `iqr(x,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the interquartile range of the elements on the corresponding page of `x`.



Data Types: `single` | `double`

pd — Probability distribution
probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

r — Interquartile range values
scalar | vector | matrix | multidimensional array

Interquartile range values, returned as a scalar, vector, matrix, or multidimensional array.

- If you input an array `x`, then the dimensions of `r` depend on whether the `'all'`, `dim`, or `vecdim` input arguments are specified. Each interquartile range value in `r` is the difference between the 75th and the 25th percentiles of the specified data contained in `x`.
- If you input a probability distribution `pd`, then the scalar value of `r` is the difference between the values of the 75th and 25th percentiles of the probability distribution.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- These input arguments are not supported: `'all'` and `vecdim`.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).
- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see "Code Generation for Probability Distribution Objects" on page 32-82.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- These input arguments are not supported: `'all'`, `vecdim`, and `pd`.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

Distribution Fitter | `boxplot` | `fitdist` | `icdf` | `mad` | `makedist` | `range` | `std`

Topics

"Working with Probability Distributions" on page 5-3
"Supported Distributions" on page 5-14

Introduced before R2006a

incrementalClassificationLinear

Binary classification linear model for incremental learning

Description

`incrementalClassificationLinear` creates an `incrementalClassificationLinear` model object, which represents a binary classification linear model for incremental learning. Supported learners include support vector machine (SVM) and logistic regression.

Unlike other Statistics and Machine Learning Toolbox model objects, `incrementalClassificationLinear` can be called directly. Also, you can specify learning options such as performance metrics configurations, parameter values, and the objective solver, before fitting the model to data. After you create an `incrementalClassificationLinear` object, it is prepared for incremental learning on page 33-2906.

`incrementalClassificationLinear` is best suited for incremental learning. For a traditional approach to training an SVM or linear model for binary classification (such as creating a model by fitting it to data, performing cross-validation, tuning hyperparameters, and so on), see `fitcsvm` or `fitclinear`. For multiclass incremental learning using the naive Bayes algorithm, see `incrementalClassificationNaiveBayes`.

Creation

You can create an `incrementalClassificationLinear` model object in several ways:

- **Call the function directly** — Configure incremental learning options, or specify initial values for linear model parameters and hyperparameters, by calling `incrementalClassificationLinear` directly. This approach is best when you do not have data yet or you want to start incremental learning immediately.
- **Convert a traditionally trained model** — To initialize a binary classification linear model for incremental learning using the model coefficients and hyperparameters of a trained SVM or binary classification linear model object, you can convert the traditionally trained model to an `incrementalClassificationLinear` model object by passing it to the `incrementalLearner` function. This table contains links to the appropriate reference pages.

Convertible Model Object	Conversion Function
`ClassificationSVM` or `CompactClassificationSVM`	`incrementalLearner`
`ClassificationLinear`	`incrementalLearner`

- **Call an incremental learning function** — `fit`, `updateMetrics`, and `updateMetricsAndFit` accept a configured `incrementalClassificationLinear` model object and data as input, and return an `incrementalClassificationLinear` model object updated with information learned from the input model and data.

Syntax

```
Mdl = incrementalClassificationLinear()
Mdl = incrementalClassificationLinear(Name,Value)
```

Description

`Mdl = incrementalClassificationLinear()` returns a default binary classification linear model object for incremental learning, `Mdl`. Properties of a default model contain placeholders for unknown model parameters. You must train a default model before you can track its performance or generate predictions from it.

`Mdl = incrementalClassificationLinear(Name,Value)` sets properties on page 33-2889 and additional options using name-value pair arguments. Enclose each name in quotes. For example, `incrementalClassificationLinear('Beta',[0.1 0.3],'Bias',1,'MetricsWarmupPeriod',100)` sets the vector of linear model coefficients β to `[0.1 0.3]`, the bias β_0 to `1`, and the metrics warm-up period to `100`.

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Standardize',true` standardizes the predictor data using the predictor means and standard deviations estimated during the estimation period.

`Metrics` — Model performance metrics to track during incremental learning
`"classiferror"` (default) | string vector | function handle | cell vector | structure array | `"binodeviance"` | `"exponential"` | `"hinge"` | `"logit"` | `"quadratic"` | ...

Model performance metrics to track during incremental learning, specified as a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

When `Mdl` is warm (see IsWarm), `updateMetrics` and `updateMetricsAndFit` track performance metrics in the Metrics property of `Mdl`.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"binodeviance"`	Binomial deviance
`"classiferror"`	Classification error
`"exponential"`	Exponential
`"hinge"`	Hinge
`"logit"`	Logistic
`"quadratic"`	Quadratic

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["classiferror" "hinge"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(C,S)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You select the function name (`customMetric`).
- `C` is an *n*-by-2 logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in the `ClassNames` property. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-2 numeric matrix of predicted classification scores. `S` is similar to the `Score` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in the `ClassNames` property. `S(p,q)` is the classification score of observation *p* being classified in class *q*.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'logit' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the `Metrics` property. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"classiferror"` is `"ClassificationError"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(C,S)customMetric(C,S)...` is `CustomMetric_1`

For more details on performance metrics options, see “Performance Metrics” on page 33-2909.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`Standardize` — Flag to standardize predictor data
`'auto'` (default) | `false` | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`'auto'`	`incrementalClassificationLinear` determines whether the predictor variables need to be standardized. See “Standardize Data” on page 33-2908.
`true`	The software standardizes the predictor data. For more details, see “Standardize Data” on page 33-2908.
`false`	The software does not standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical` | `char` | `string`

Properties

You can set most properties by using name-value pair argument syntax only when you call `incrementalClassificationLinear` directly. You can set some properties when you call `incrementalLearner` to convert a traditionally trained model. You cannot set the properties `FittedLoss`, `NumTrainingObservations`, `Mu`, `Sigma`, `SolverOptions`, and `IsWarm`.

Classification Model Parameters

`Beta` — Linear model coefficients β

numeric vector

This property is read-only.

Linear model coefficients β, specified as a `NumPredictors`-by-1 numeric vector.

If you convert a traditionally trained model to create `Mdl`, `Beta` is specified by the value of the `Beta` property of the traditionally trained model. Otherwise, by default, `Beta` is `zeros(NumPredictors,1)`.

Data Types: `single` | `double`

`Bias` — Model intercept β_0

numeric scalar

This property is read-only.

Model intercept β_0, or bias term, specified as a numeric scalar.

If you convert a traditionally trained model to create `Mdl`, `Bias` is specified by the value of the `Bias` property of the traditionally trained model. Otherwise, by default, `Bias` is `0`.

Data Types: `single` | `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used in training the model, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `incrementalClassificationLinear` stores a specified string vector as a cell array of character vectors. `ClassNames` and the response data must have the same data type.

- If you convert a traditionally trained model to create `Mdl`, `ClassNames` is the `ClassNames` property of the traditionally trained model.
- Otherwise, incremental fitting functions infer `ClassNames` during training.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`FittedLoss` — Loss function used to fit linear model
`'hinge'` | `'logit'`

This property is read-only.

Loss function used to fit the linear model, specified as `'hinge'` or `'logit'`.

Value	Algorithm	Loss Function	`Learner` Value
`'hinge'`	Support vector machine	Hinge: $\ell[y, f(x)] = \max [0, 1 - yf(x)]$	`'svm'`
`'logit'`	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log \{1 + \exp[-yf(x)]\}$	`'logistic'`

`Learner` — Linear classification model type
`'logistic'` | `'svm'`

This property is read-only.

Linear classification model type, specified as `'logistic'` or `'svm'`.

In the following table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.
- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Loss Function	`FittedLoss` Value
`'logistic'`	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log \{1 + \exp[-yf(x)]\}$	`'logit'`
`'svm'`	Support vector machine	Hinge: $\ell[y, f(x)] = \max [0, 1 - yf(x)]$	`'hinge'`

If you convert a traditionally trained model to create `Mdl`, `Learner` is the learner of the traditionally trained model.

- If the traditionally trained model is `ClassificationSVM` or `CompactClassificationSVM`, `Learner` is `'svm'`.

- If the traditionally trained model is `ClassificationLinear`, `Learner` is the value of the `Learner` property of the traditionally trained model.

`NumPredictors` — Number of predictor variables
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of predictor variables, specified as a nonnegative numeric scalar.

If you convert a traditionally trained model to create `Mdl`, `NumPredictors` is specified by the congruent property of the traditionally trained model. Otherwise, incremental fitting functions infer `NumPredictors` from the predictor data during training.

Data Types: `double`

`NumTrainingObservations` — Number of observations fit to incremental model
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of observations fit to the incremental model `Mdl`, specified as a nonnegative numeric scalar. `NumTrainingObservations` increases when you pass `Mdl` and training data to `fit` or `updateMetricsAndFit`.

Note If you convert a traditionally trained model to create `Mdl`, `incrementalClassificationLinear` does not add the number of observations fit to the traditionally trained model to `NumTrainingObservations`.

Data Types: `double`

`Prior` — Prior class probabilities
numeric vector | `'empirical'` | `'uniform'`

This property is read-only.

Prior class probabilities, specified as a value in this table. You can set this property using name-value pair argument syntax, but `incrementalClassificationLinear` always stores a numeric vector.

Value	Description
`'empirical'`	Incremental learning functions infer prior class probabilities from the observed class relative frequencies in the response data during incremental training (after the estimation period `EstimationPeriod`).
`'uniform'`	For each class, the prior probability is $1/K$, where K is the number of classes.
numeric vector	Custom, normalized prior probabilities. The order of the elements of `Prior` corresponds to the elements of the `ClassNames` property.

- If you convert a traditionally trained model to create `Mdl`, `incrementalClassificationLinear` uses the `Prior` property of the traditionally trained model.
- Otherwise, `Prior` is `'empirical'`.

Data Types: `single` | `double`

`ScoreTransform` — Score transformation function
character vector | string scalar | function handle

This property is read-only.

Score transformation function describing how incremental learning functions transform raw response values, specified as a character vector, string scalar, or function handle. `incrementalClassificationLinear` stores the specified value as a character vector or function handle.

This table describes the available built-in functions for score transformation.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function that you define, enter its function handle; for example, `'ScoreTransform',@`*`function`*, where:

- *`function`* accepts an n-by-K matrix (the original scores) and returns a matrix of the same size (the transformed scores).
- n is the number of observations, and row j of the matrix contains the class scores of observation j.
- K is the number of classes `numel(ClassNames)`, and column k is class `ClassNames(`*`k`*`)`.

By default:

- If you convert a traditionally trained model to create `Mdl`, `ScoreTransform` is specified by the congruent property of the traditionally trained model. For example, if the `ScoreTransform` property of the traditionally trained model is a score-to-posterior-probability transformation function, as computed by `fitPosterior` or `fitSVMPosterior`, `Mdl.ScoreTransform` contains an anonymous function.

- `ScoreTransform` is `'none'` when `Learner` is `'svm'`.
- `ScoreTransform` is `'logit'` when `Learner` is `'logistic'`.

Data Types: `char` | `function_handle`

Training Parameters

`EstimationPeriod` — Number of observations processed to estimate hyperparameters

nonnegative integer

This property is read-only.

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalClassificationLinear` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalClassificationLinear` sets `'EstimationPeriod'` to `1000`.

For more details, see “Estimation Period” on page 33-2907.

Data Types: `single` | `double`

`FitBias` — Linear model intercept inclusion flag

`true` | `false`

This property is read-only.

Linear model intercept inclusion flag, specified as `true` or `false`.

Value	Description
`true`	`incrementalClassificationLinear` includes the bias term β_0 in the linear model, which incremental fitting functions fit to data.
`false`	`incrementalClassificationLinear` sets $\beta_0 = 0$.

If `Bias` ≠ 0, `FitBias` must be `true`. In other words, `incrementalClassificationLinear` does not support an equality constraint on β_0.

If you convert a traditionally trained linear classification model (`ClassificationLinear`) to create `Mdl`, `FitBias` is specified by the value of the `ModelParameters.FitBias` property of the traditionally trained model.

Data Types: `logical`

`Mu` — Predictor means

vector of numeric values | `[]`

This property is read-only.

Predictor means, specified as a numeric vector.

If `Mu` is an empty array `[]` and you specify `'Standardize',true`, incremental fitting functions set `Mu` to the predictor variable means estimated during the estimation period specified by `EstimationPeriod`.

You cannot specify `Mu` directly.

Data Types: `single` | `double`

`Sigma` — Predictor standard deviations
vector of numeric values | `[]`

This property is read-only.

Predictor standard deviations, specified as a numeric vector.

If `Sigma` is an empty array `[]` and you specify `'Standardize',true`, incremental fitting functions set `Sigma` to the predictor variable standard deviations estimated during the estimation period specified by `EstimationPeriod`.

You cannot specify `Sigma` directly.

Data Types: `single` | `double`

`Solver` — Objective function minimization technique
`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

This property is read-only.

Objective function minimization technique, specified as a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-2907 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Parameters” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Parameters” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

If you convert a traditionally trained linear model for binary classification (`ClassificationLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `Solver` is specified by the `ModelParameters.Solver` property of the traditionally trained model.

Data Types: `char` | `string`

`SolverOptions` — Objective solver configurations

structure array

This property is read-only.

Objective solver configurations, specified as a structure array. The fields of `SolverOptions` are properties specific to the specified solver `Solver`.

Data Types: `struct`

SGD and ASGD Solver Parameters

`BatchSize` — Mini-batch size

positive integer

This property is read-only.

Mini-batch size, specified as a positive integer. At each iteration during training, `incrementalClassificationLinear` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

If you convert a traditionally trained linear model for binary classification (`ClassificationLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `BatchSize` is specified by the `ModelParameters.BatchSize` property of the traditionally trained model. Otherwise, the default is `10`.

Data Types: `single` | `double`

`Lambda` — Ridge (*L*2) regularization term strength

nonnegative scalar

This property is read-only.

Ridge (*L*2) regularization term strength, specified as a nonnegative scalar.

If you convert a traditionally trained linear model for binary classification with a ridge penalty (`ClassificationLinear` object with property `Regularization` equal to `'ridge (L2)'`) to create `Mdl`, `Lambda` is specified by the value of the `Lambda` property of the traditionally trained model. Otherwise, the default is `1e-5`.

Data Types: `double` | `single`

`LearnRate` — Learning rate

`'auto'` | positive scalar

This property is read-only.

Learning rate, specified as `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

When you specify `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either.

If you convert a traditionally trained linear model for binary classification (`ClassificationLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `LearnRate` is specified by the `ModelParameters.LearnRate` property of the traditionally trained model.

The `LearnRateSchedule` property determines the learning rate for subsequent learning cycles.

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule
`'decaying'` (default) | `'constant'`

This property is read-only.

Learning rate schedule, specified as a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.
`'decaying'`	The learning rate at learning cycle t is $$\gamma_t = \frac{\gamma_0}{(1+\lambda\gamma_0 t)^c}.$$ • λ is the value of `Lambda`. • If `Solver` is `'sgd'`, then $c = 1$. • If `Solver` is `'asgd'`, then c is 0.75 [4].

If you convert a traditionally trained linear model for binary classification (`ClassificationLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `LearnRate` is `'decaying'`.

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch
`true` (default) | `false`

This property is read-only.

Flag for shuffling the observations in the batch at each learning cycle, specified as a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Data Types: `logical`

Performance Metrics Parameters

`IsWarm` — Flag indicating whether model tracks performance metrics

`false` | `true`

This property is read-only.

Flag indicating whether the incremental model tracks performance metrics, specified as `false` or `true`. The incremental model `Mdl` is warm (`IsWarm` becomes `true`) after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (that is, `EstimationPeriod` + `MetricsWarmupPeriod` observations).

Value	Description
`true`	The incremental model `Mdl` is warm. Consequently, `updateMetrics` and `updateMetricsAndFit` track performance metrics in the `Metrics` property of `Mdl`.
`false`	`updateMetrics` and `updateMetricsAndFit` do not track performance metrics.

Data Types: `logical`

`Metrics` — Model performance metrics

table

This property is read-only.

Model performance metrics updated during incremental learning by `updateMetrics` and `updateMetricsAndFit`, specified as a table with two columns and *m* rows, where *m* is the number of metrics specified by the `'Metrics'` name-value pair argument.

The columns of `Metrics` are labeled `Cumulative` and `Window`.

- **`Cumulative`**: Element *j* is the model performance, as measured by metric *j*, from the time the model became warm (IsWarm is 1).
- **`Window`**: Element *j* is the model performance, as measured by metric *j*, evaluated over all observations within the window specified by the `MetricsWindowSize` property. The software updates `Window` after it processes `MetricsWindowSize` observations.

Rows are labeled by the specified metrics. For details, see `'Metrics'`.

Data Types: `table`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`1000` (default) | nonnegative integer

This property is read-only.

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as a nonnegative integer.

For more details, see “Performance Metrics” on page 33-2909.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer

This property is read-only.

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2909.

Data Types: `single` | `double`

Object Functions

fit	Train linear model for incremental learning
updateMetricsAndFit	Update performance metrics in linear model for incremental learning given new data and train model
updateMetrics	Update performance metrics in linear model for incremental learning given new data
loss	Loss of linear model for incremental learning on batch of data
predict	Predict responses for new observations from linear model for incremental learning

Examples

Create Incremental Learner Without Any Prior Information

Create a default incremental linear SVM model for binary classification.

```
Mdl = incrementalClassificationLinear()
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
              Beta: [0x1 double]
              Bias: 0
           Learner: 'svm'
```

```
  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Fit the incremental model to the training data by using the `updateMetricsAndfit` function. Simulate a data stream by processing chunks of 50 observations at a time. At each iteration:

- Process 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

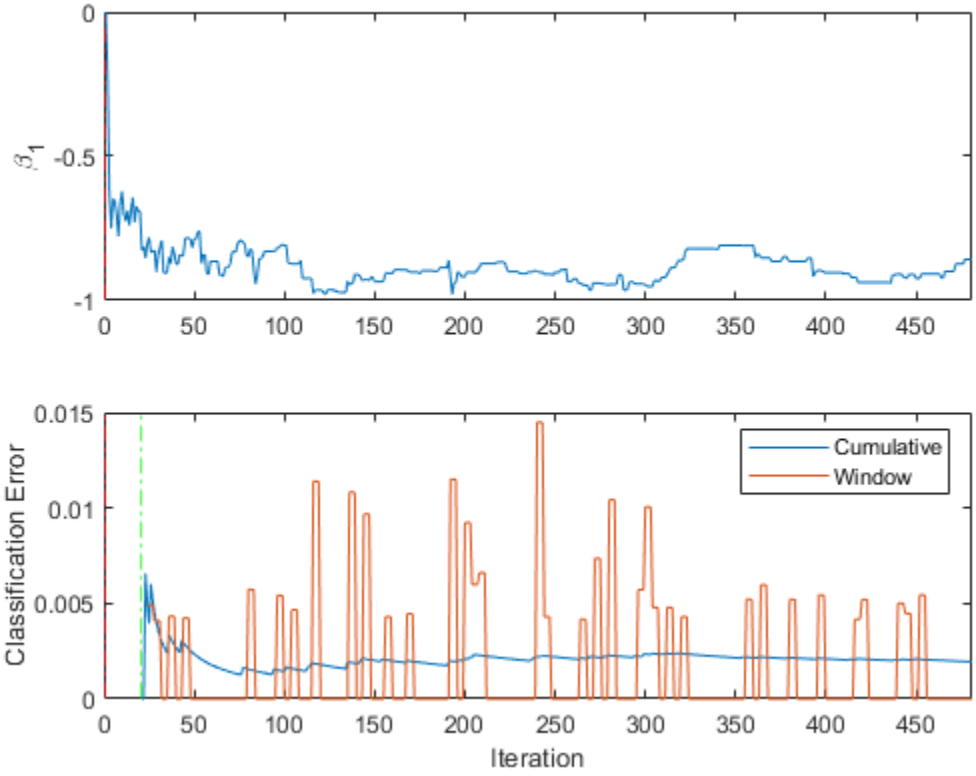
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    beta1(j + 1) = Mdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
```

```
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline((Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod)/numObsPerChunk,'g-.');
legend(h,ce.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_1 during all incremental learning iterations
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Configure Incremental Learning Options

Prepare an incremental binary SVM learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500

observations. Train the model by using SGD, and adjust the SGD batch size, learning rate, and regularization parameter.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Waling, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Create an incremental linear model for binary classification. Configure the model as follows:

- Specify that the incremental fitting functions process the raw (unstandardized) predictor data.
- Specify the SGD solver.
- Assume that a ridge regularization parameter value of 0.001, SGD batch size of 20, and learning rate of 0.002 work well for the problem.
- Specify a metrics warm-up period of 5000 observations.
- Specify a metrics window size of 500 observations.
- Track the classification and hinge error metrics to measure the performance of the model.

```
Mdl = incrementalClassificationLinear('Standardize',false,...
    'Solver','sgd','Lambda',0.001,'BatchSize',20,'LearnRate',0.002,...
    'MetricsWarmupPeriod',5000,'MetricsWindowSize',500,...
    'Metrics',{'classiferror' 'hinge'})
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [2x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
              Beta: [0x1 double]
              Bias: 0
           Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model object configured for incremental learning.

Fit the incremental model to the rest of the data by using the `updateMetricsAndfit` function. At each iteration:

- Simulate a data stream by processing a chunk of 50 observations. Note that chunk size is different from SGD batch size.

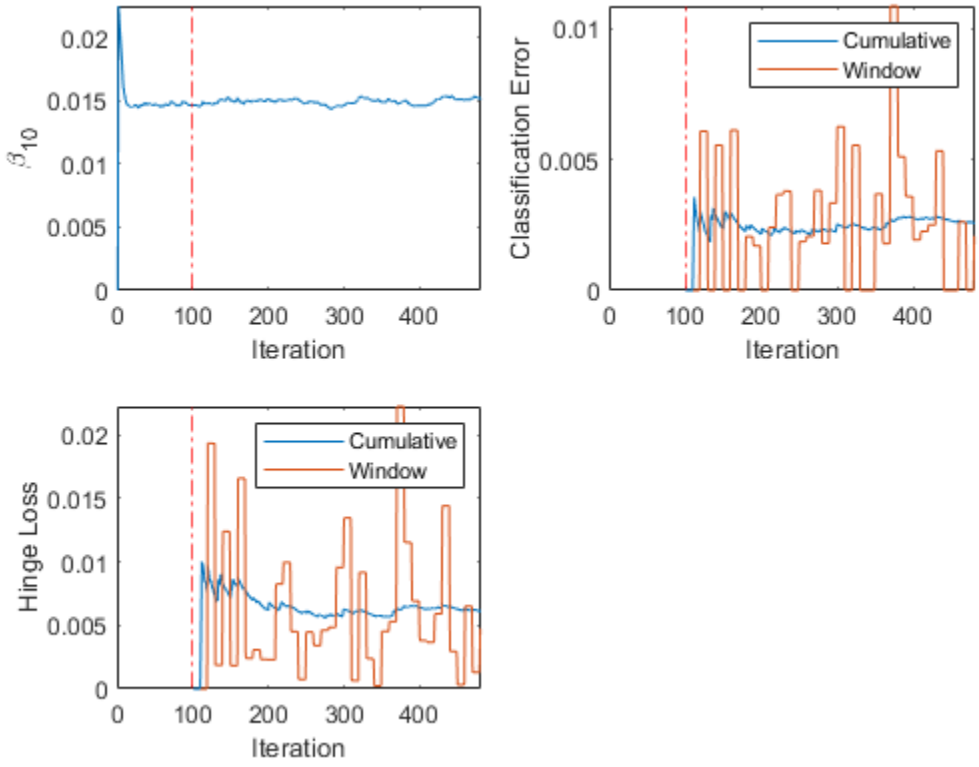
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the estimated coefficient β_{10}, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
hinge = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta10 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    hinge{j,:} = Mdl.Metrics{"HingeLoss",:};
    beta10(j + 1) = Mdl.Beta(10);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_{10} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(beta10)
ylabel('\beta_{10}')
xlim([0 nchunk]);
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
subplot(2,2,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ce.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,3)
h = plot(hinge.Variables);
xlim([0 nchunk]);
ylabel('Hinge Loss')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,hinge.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_{10} during all incremental learning iterations
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (10 iterations).

Convert Traditionally Trained Model to Incremental Learner

Train a linear model for binary classification by using `fitclinear`, convert it to an incremental learner, track its performance, and fit it to streaming data. Carry over training options from traditional to incremental learning.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data. Orient the observations of the predictor data in columns.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
```

```
X = feat(idx,:)';
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Suppose that the data collected when the subject was idle (`Y = false`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from an idle subject, and 1 to a moving subject.

```
W = ones(n,1) + ~Y;
```

Train Linear Model for Binary Classification

Fit a linear model for binary classification to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitclinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns',...
    'Weights',W(idxtt))
```

```
TTMdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.1107
            Lambda: 8.2967e-05
           Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear model for binary classification.

Convert Trained Model

Convert the traditionally trained classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

             IsWarm: 1
            Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.1107
           Learner: 'svm'
```

```
  Properties, Methods
```

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. Simulate a data stream by processing 50 observations at a time. At each iteration:

1. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify that the observations are oriented in columns, and specify the observation weights.
2. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify that the observations are oriented in columns, and specify the observation weights.
3. Store the classification error and first estimated coefficient β_1.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];
Xil = X(:,idxil);
Yil = Y(idxil);
Wil = W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(:,idx),Yil(idx),...
        'ObservationsIn','columns','Weights',Wil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(:,idx),Yil(idx),'ObservationsIn','columns',...
        'Weights',Wil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(end);
end
```

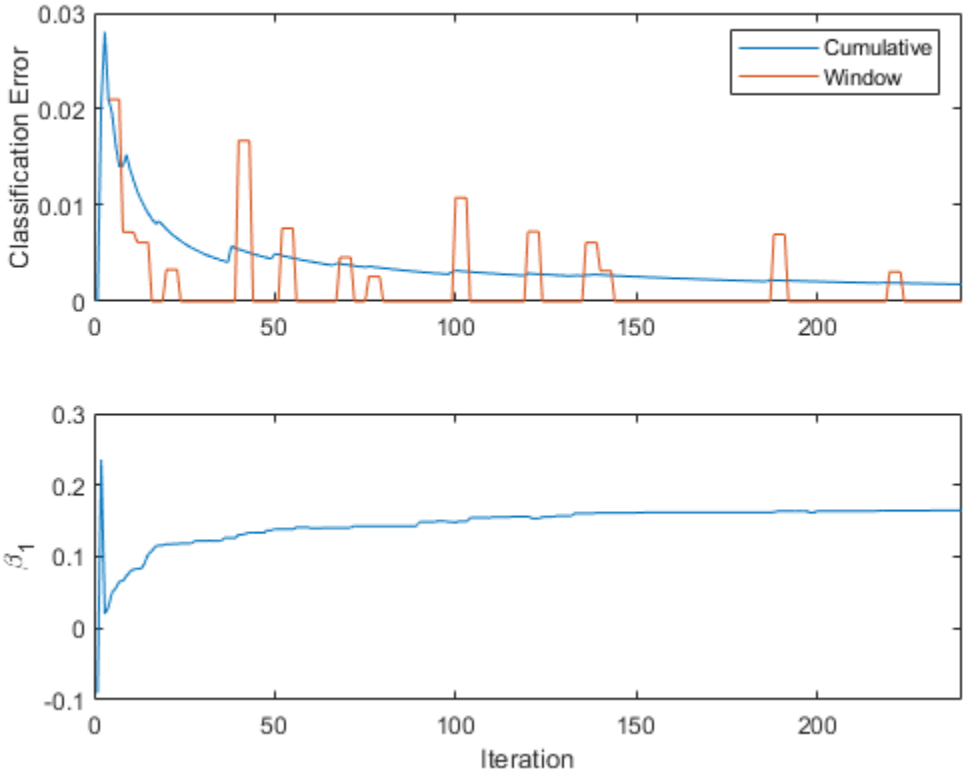
`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Plot a trace plot of the performance metrics and estimated coefficient β_1.

```
figure;
subplot(2,1,1)
h = plot(ce.Variables);
xlim([0 nchunk]);
```

```
ylabel('Classification Error')
legend(h,ce.Properties.VariableNames)
subplot(2,1,2)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable and decreases gradually, whereas the window loss jumps.

β_1 changes abruptly at first, then gradually levels off as `fit` processes more chunks.

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for classification `fit` and `updateMetricsAndFit` use the more aggressive ScInOL2 version of the algorithm.

Tips

- After creating a model, you can generate C/C++ code that performs incremental learning on a data stream. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned. Estimation occurs only when `EstimationPeriod` is positive.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When both these conditions apply: • Incremental fitting functions are configured to standardize predictor data (see “Standardize Data” on page 33-2908). • `Mdl.Mu` and `Mdl.Sigma` are empty arrays `[]`.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Learning rate	LearnRate	Adjust solver step size	When both these conditions apply: • The solver is SGD or ASGD (see Solver). • You do not set the `'LearnRate'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `Mdl`.

- When you set `'Standardize',true` and a positive estimation period (see EstimationPeriod), and `Mdl.Mu` and `Mdl.Sigma` are empty, incremental fitting functions estimate means and standard deviations using the estimation period observations.
- When you set `'Standardize','auto'` (the default), the following conditions apply:
  - If you create `incrementalClassificationLinear` by converting a traditionally trained binary linear SVM model (`ClassificationSVM` or `CompactClassificationSVM`), and the `Mu` and `Sigma` properties of the traditionally trained model are empty arrays `[]`, incremental learning functions do not standardize predictor variables. If the `Mu` and `Sigma` properties of the traditionally trained model are nonempty, incremental learning functions standardize the predictor variables using the specified means and standard deviations. Incremental fitting functions do not estimate new means and standard deviations, regardless of the length of the estimation period.
  - If you create `incrementalClassificationLinear` by converting a linear classification model (`ClassificationLinear`), incremental learning functions do not standardize the data, regardless of the length of the estimation period.
  - If you do not convert a traditionally trained model, incremental learning functions standardize the predictor data only when you specify an SGD solver (see `Solver`) and a positive estimation period (see EstimationPeriod).
- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

$$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k^*} \sum_k w_k^* x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k^*} \sum_k w_k^* \left(x_{jk} - \mu_j^*\right)^2$.

- $w_j^* = \frac{w_j}{\sum_{\forall j \in \text{Class } k} w_j} p_k$,

  - p_k is the prior probability of class k (`Prior` property of the incremental model).
  - w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (IsWarm property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to MetricsWarmupPeriod observations, which is the metrics warm-up period.

  If EstimationPeriod > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:

  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the MetricsWindowSize name-value pair argument. `MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- All object functions on page 33-2898 of an `incrementalClassificationLinear` model object support code generation.
- If you configure `Mdl` to shuffle data (see Solver and Shuffle), the `fit` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB.
- When you generate code that loads or creates an `incrementalClassificationLinear` model object, the following restrictions apply.
  - `Mdl` cannot represent a converted SVM model configured to return posterior probabilities as scores.
  - The `ClassNames` property must contain all expected class names.
  - The `NumPredictors` property must reflect the number of predictor variables.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Functions

`fit` | `incrementalLearner` | `incrementalLearner` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Objects

`ClassificationLinear` | `ClassificationSVM` | `CompactClassificationSVM`

Topics

“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19

“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26
“Initialize Incremental Learning Model from Logistic Regression Model Trained in Classification Learner” on page 26-36

Introduced in R2020b

incrementalClassificationNaiveBayes

Naive Bayes classification model for incremental learning

Description

`incrementalClassificationNaiveBayes` creates an `incrementalClassificationNaiveBayes` model object, which represents a naive Bayes multiclass classification model for incremental learning. `incrementalClassificationNaiveBayes` supports normally distributed predictor variables.

Unlike other Statistics and Machine Learning Toolbox model objects, `incrementalClassificationNaiveBayes` can be called directly. Also, you can specify learning options such as performance metrics configurations and prior class probabilities before fitting the model to data. After you create an `incrementalClassificationNaiveBayes` object, it is prepared for incremental learning on page 33-2932.

`incrementalClassificationNaiveBayes` is best suited for incremental learning. For a traditional approach to training a naive Bayes model for multiclass classification (such as creating a model by fitting it to data, performing cross-validation, tuning hyperparameters, and so on), see `fitcnb`.

Creation

You can create an `incrementalClassificationNaiveBayes` model object in several ways:

- **Call the function directly** — Configure incremental learning options, or specify learner-specific options, by calling `incrementalClassificationNaiveBayes` directly. This approach is best when you do not have data yet or you want to start incremental learning immediately. You must specify the maximum number of classes or all class names expected in the response data during incremental learning.
- **Convert a traditionally trained model** — To initialize a naive Bayes classification model for incremental learning using the model parameters of a trained naive Bayes model object, you can convert the traditionally trained model to an `incrementalClassificationNaiveBayes` model object by passing it to the `incrementalLearner` function.
- **Call an incremental learning function** — `fit`, `updateMetrics`, and `updateMetricsAndFit` accept a configured `incrementalClassificationNaiveBayes` model object and data as input, and return an `incrementalClassificationNaiveBayes` model object updated with information learned from the input model and data.

Syntax

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',MaxNumClasses)
Mdl = incrementalClassificationNaiveBayes('ClassNames',ClassNames)
Mdl = incrementalClassificationNaiveBayes(___,Name,Value)
```

Description

`Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',MaxNumClasses)` returns a default naive Bayes classification model object for incremental learning, `Mdl`, where `MaxNumClasses` is the maximum number of classes expected in the response data during incremental learning. Properties of a default model contain placeholders for unknown model parameters. You must train a default model before you can track its performance or generate predictions from it.

`Mdl = incrementalClassificationNaiveBayes('ClassNames',ClassNames)` specifies all class names `ClassNames` expected in the response data during incremental learning, and sets the `ClassNames` property.

`Mdl = incrementalClassificationNaiveBayes( ___ ,Name,Value)` uses any of the input-argument combinations in the previous syntaxes to set properties on page 33-2916 and additional options using name-value pair arguments . Enclose each name in quotes. For example, `incrementalClassificationNaiveBayes('MaxNumClasses',5,'MetricsWarmupPeriod',100)` sets the maximum number of classes expected in the response data to `5`, and sets the metrics warm-up period to `100`.

Input Arguments

`MaxNumClasses` — Maximum number of classes

positive integer

Maximum number of classes expected in the response data during incremental learning, specified as a positive integer.

If you do not specify `MaxNumClasses`, you must specify the `ClassNames` argument. In that case, `MaxNumClasses` is the number of class names in `ClassNames`.

Example: `'MaxNumClasses',5`

Data Types: `single` | `double`

`ClassNames` — All unique class labels

categorical array | character array | string vector | logical vector | numeric vector | cell array of character vectors

All unique class labels expected in the response data during incremental learning, specified as a categorical or character array; logical, numeric, or string vector, or cell array of character vectors. `ClassNames` and the response data must have the same data type. `ClassNames` sets the `ClassNames` property.

`ClassNames` specifies the order of any input or output argument dimension that corresponds to the class order. For example, set `'ClassNames'` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`

If you do not specify `ClassNames`, you must specify the `MaxNumClasses` argument. In that case, the software infers the `MaxNumClasses` `ClassNames` from the data during incremental learning.

Example: `'ClassNames',["virginica" "setosa" "versicolor"}`

Data Types: `single` | `double` | `logical` | `string` | `char` | `cell` | `categorical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumPredictors',4,'Prior',[0.3 0.3 0.4]` specifies 4 variables in the predictor data and a prior class probability distribution of `[0.3 0.3 0.4]`.

`Cost` — Cost of misclassifying an observation
square matrix | structure array

Cost of misclassifying an observation, specified as a value in this table, where are `MaxNumClasses` is the number of classes in the `ClassNames` property:

Value	Description
`MaxNumClasses`-by-`MaxNumClasses` numeric matrix	`Cost(i,j)` is the cost of classifying an observation into class `j` when its true class is `i`. In other words, the rows correspond to the true class and the columns correspond to the predicted class. For example, `Cost = [0 2;1 0]` applies double the penalty for misclassifying `ClassNames(1)` than for misclassifying `ClassNames(2)`.
Structure array	A structure array having two fields: • `ClassNames` containing the class names, the same value as `ClassNames` • `ClassificationCosts` containing the cost matrix, as previously described.

If you specify `Cost`, you must also specify the `ClassNames` argument. `Cost` sets the `Cost` property.

The default is a `MaxNumClasses`-by-`MaxNumClasses` matrix, where `Cost(`i`,`j`) = 1` for all $i \neq j$, and `Cost(`i`,`j`) = 0` for all $i = j$.

Example: `'Cost',struct('ClassNames',{'b','g'},'ClassificationCosts',[0 2; 1 0])`

Data Types: `single` | `double` | `struct`

`Metrics` — Model performance metrics to track during incremental learning
`"mincost"` (default) | `"classiferror"` | string vector | function handle | cell vector | structure array | `"binodeviance"` | `"exponential"` | `"hinge"` | `"logit"` | `"quadratic"` | ...

Model performance metrics to track during incremental learning, in addition to minimal expected misclassification cost, specified as a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

When `Mdl` is warm (see IsWarm), `updateMetrics` and `updateMetricsAndFit` track performance metrics in the Metrics property of `Mdl`.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"binodeviance"`	Binomial deviance
`"classiferror"`	Misclassification error rate
`"exponential"`	Exponential
`"hinge"`	Hinge
`"logit"`	Logistic
`'"mincost"`	Minimal expected misclassification cost (for classification scores that are posterior probabilities). `incrementalClassificationNaiveBayes` always tracks this metric.
`"quadratic"`	Quadratic

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["classiferror" "logit"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(C,S,Cost)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You select the function name (`customMetric`).
- `C` is an *n*-by-*K* logical matrix with rows indicating the class to which the corresponding observation belongs, where *K* is the number of classes. The column order corresponds to the class order in the `ClassNames` property. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-*K* numeric matrix of predicted classification scores. `S` is similar to the `Posterior` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in the `ClassNames` property. `S(p,q)` is the classification score of observation *p* being classified in class *q*.
- `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. See the `'Cost'` name-value argument.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'logit' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the `Metrics` property. The data type of `Metrics` determines the row names of the table.

'Metrics' Value Data Type	Description of Metrics Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for "classiferror" is "ClassificationError"
Structure array	Field name	Row name for struct('Metric1',@customMetric1) is "Metric1"
Function handle to function stored in a program file	Name of function	Row name for @customMetric is "customMetric"
Anonymous function	CustomMetric_*j*, where *j* is metric *j* in Metrics	Row name for @(C,S,Cost)customMetric(C,S,Cost)... is CustomMetric_1

For more details on performance metrics options, see “Performance Metrics” on page 33-2932.

Data Types: char | string | struct | cell | function_handle

Properties

You can set most properties by using name-value pair argument syntax only when you call incrementalClassificationNaiveBayes directly. You can set some properties when you call incrementalLearner to convert a traditionally trained model. You cannot set the properties DistributionParameters, IsWarm, and NumTrainingObservations.

Classification Model Parameters

Cost — Cost of misclassifying an observation

square numeric matrix

This property is read-only.

Cost of misclassifying an observation, specified as a MaxNumClasses-by-MaxNumClasses numeric matrix.

If you specify the 'Cost' name-value argument, its value sets Cost. If you specify a structure array, Cost is the value of the ClassificationCosts field.

If you convert a traditionally trained model to create Mdl, Cost is the Cost property of the traditionally trained model.

Data Types: single | double

ClassNames — All unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

All unique class labels expected in the response data during incremental learning, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors.

- If you specify the MaxNumClasses argument, the software infers ClassNames during incremental learning.

- If you specify the `ClassNames` argument, `incrementalClassificationNaiveBayes` stores your specification in `ClassNames`. If you specify a string vector, `incrementalClassificationNaiveBayes` stores it as a cell array of character vectors instead.
- If you convert a traditionally trained model to create `Mdl`, `ClassNames` is the `ClassNames` property of the traditionally trained model.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`NumPredictors` — Number of predictor variables
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of predictor variables, specified as a nonnegative numeric scalar.

If you convert a traditionally trained model to create `Mdl`, `NumPredictors` is specified by the congruent property of the traditionally trained model. Otherwise, incremental fitting functions infer `NumPredictors` from the predictor data during training.

Data Types: `double`

`NumTrainingObservations` — Number of observations fit to incremental model
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of observations fit to the incremental model `Mdl`, specified as a nonnegative numeric scalar. `NumTrainingObservations` increases when you pass `Mdl` and training data to `fit` or `updateMetricsAndFit`.

Note If you convert a traditionally trained model to create `Mdl`, `incrementalClassificationNaiveBayes` does not add the number of observations fit to the traditionally trained model to `NumTrainingObservations`.

Data Types: `double`

`Prior` — Prior class probabilities
numeric vector | `'empirical'` | `'uniform'`

This property is read-only.

Prior class probabilities, specified as a value in this table. You can set this property using name-value pair argument syntax, but `incrementalClassificationNaiveBayes` always stores a numeric vector.

Value	Description
`'empirical'`	Incremental learning functions infer prior class probabilities from the observed class relative frequencies in the response data during incremental training.

Value	Description
`'uniform'`	For each class, the prior probability is 1/*K*, where *K* is the number of classes.
numeric vector	Custom, normalized prior probabilities. The order of the elements of `Prior` corresponds to the elements of the `ClassNames` property.

- If you convert a traditionally trained model to create `Mdl`, `incrementalClassificationNaiveBayes` uses the `Prior` property of the traditionally trained model.
- Otherwise, `Prior` is `'empirical'`.

Data Types: `single` | `double`

`ScoreTransform` — Score transformation function
`'none'` (default) | string scalar | character vector | function handle

This property is read-only.

Score transformation function describing how incremental learning functions transform raw response values, specified as a character vector, string scalar, or function handle. `incrementalClassificationNaiveBayes` stores the specified value as a character vector or function handle.

This table describes the available built-in functions for score transformation.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function that you define, enter its function handle; for example, `'ScoreTransform',@`*`function`*, where:

- *`function`* accepts an *n*-by-*K* matrix (the original scores) and returns a matrix of the same size (the transformed scores).
- *n* is the number of observations, and row *j* of the matrix contains the class scores of observation *j*.
- *K* is the number of classes, and column *k* is class `ClassNames(`*`k`*`)`.

If you convert a traditionally trained model to create `Mdl`, `ScoreTransform` is specified by the congruent property of the traditionally trained model.

The default `'none'` specifies returning posterior class probabilities.

Data Types: `char` | `function_handle`

Training Parameters

`DistributionNames` — Predictor distributions

`'normal'` (default) | cell array of character vectors

This property is read-only.

Predictor distributions, specified as `'normal'` or a 1-by-`NumPredictors` cell array with all cells containing `'normal'`. The conditional distribution $P(x_j|c_k)$ is normal (Gaussian), for $j = 1, \ldots,$`NumPredictors` and each $k \in$ `ClassNames`.

Data Types: `char` | `string` | `cell`

`DistributionParameters` — Distribution parameter estimates

cell array

This property is read-only.

Distribution parameter estimates, specified as a cell array. `DistributionParameters` is a K-by-`NumPredictors` cell array, where K is the number of classes and cell (k,j) contains the distribution parameter estimates for instances of predictor j in class k. The order of the rows corresponds to the order of the classes in the property `ClassNames`, and the order of the columns corresponds to the order of the predictors in the predictor data.

If class k has no observations for predictor j, then `DistributionParameters{`k`,`j`}` is empty (`[]`).

Because all predictor distributions, specified by the `DistributionNames` property, are `'normal'`, each cell of `DistributionParameters` is a 2-by-1 numeric vector, where the first element is the sample mean and the second element is the sample standard deviation.

Data Types: `cell`

Performance Metrics Parameters

`IsWarm` — Flag indicating whether model tracks performance metrics

`false` | `true`

Flag indicating whether the incremental model tracks performance metrics, specified as `false` or `true`.

The incremental model `Mdl` is warm (`IsWarm` becomes `true`) when incremental fitting functions perform both of the following actions:

- Fit the incremental model to `MetricsWarmupPeriod` observations.
- Process `MaxNumClasses` classes or all class names specified by the `ClassNames` name-value argument.

Value	Description
`true`	The incremental model `Mdl` is warm. Consequently, `updateMetrics` and `updateMetricsAndFit` track performance metrics in the `Metrics` property of `Mdl`.
`false`	`updateMetrics` and `updateMetricsAndFit` do not track performance metrics.

Data Types: `logical`

`Metrics` — Model performance metrics
table

This property is read-only.

Model performance metrics updated during incremental learning by `updateMetrics` and `updateMetricsAndFit`, specified as a table with two columns and *m* rows, where *m* is the number of metrics specified by the `'Metrics'` name-value pair argument.

The columns of `Metrics` are labeled `Cumulative` and `Window`.

- **`Cumulative`**: Element *j* is the model performance, as measured by metric *j*, from the time the model became warm (IsWarm is `1`).
- **`Window`**: Element *j* is the model performance, as measured by metric *j*, evaluated over all observations within the window specified by the `MetricsWindowSize` property. The software updates `Window` after it processes `MetricsWindowSize` observations.

Rows are labeled by the specified metrics. For details, see `'Metrics'`.

Data Types: `table`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`1000` (default) | nonnegative integer

This property is read-only.

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as a nonnegative integer.

For more details, see “Performance Metrics” on page 33-2932.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer

This property is read-only.

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2932.

Data Types: `single` | `double`

Object Functions

fit	Train naive Bayes classification model for incremental learning
updateMetricsAndFit	Update performance metrics in naive Bayes classification model for incremental learning given new data and train model
updateMetrics	Update performance metrics in naive Bayes classification model for incremental learning given new data
logp	Log unconditional probability density of naive Bayes classification model for incremental learning
loss	Loss of naive Bayes classification model for incremental learning on batch of data
predict	Predict responses for new observations from naive Bayes classification model for incremental learning

Examples

Create Incremental Learner With Little Prior Information

To create a naive Bayes classification model for incremental learning, the least amount of information you must specify is the maximum number of classes that you expect the model to experience (`'MaxNumClasses'` name-value argument). As you fit the model to incoming batches of data by using an incremental fitting function, the model collects newly experienced classes in its `ClassNames` property. If the specified expected maximum number of classes is inaccurate, one of the following alternatives occurs:

- Before an incremental fitting function experiences the expected maximum number of classes, the model is cold. Consequently, the `updateMetrics` and `updateMetricsAndFit` functions do not measure performance metrics.
- If the number of classes exceeds the maximum expected, the incremental fitting function issues an error.

This example shows how to create a naive Bayes classification model for incremental learning when you know only the expected maximum number of classes in the data. Also, the example illustrates the consequences when incremental fitting functions experience all expected classes early and late in the sample.

For this example, consider training a device to predict whether a subject is sitting, standing, walking, running, or dancing based on biometric data measured on the subject. Therefore, the device has a maximum of 5 classes from which to choose.

Experience Expected Maximum Number of Classes Early in Sample

Create an incremental naive Bayes learner for multiclass learning. Specify a maximum of 5 classes in the data.

```
MdlEarly = incrementalClassificationNaiveBayes('MaxNumClasses',5)
```

```
MdlEarly =
  incrementalClassificationNaiveBayes

                  IsWarm: 0
                 Metrics: [1×2 table]
              ClassNames: [1×0 double]
          ScoreTransform: 'none'
```

```
         DistributionNames: 'normal'
    DistributionParameters: {}


  Properties, Methods
```

`MdlEarly` is an `incrementalClassificationNaiveBayes` model object. All its properties are read-only.

`MdlEarly` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Fit the incremental model to the training data by using the `updateMetricsAndfit` function. Simulate a data stream by processing chunks of 50 observations at a time. At each iteration:

- Process 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store μ_{11} (the mean of the first predictor variable in the first class), the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

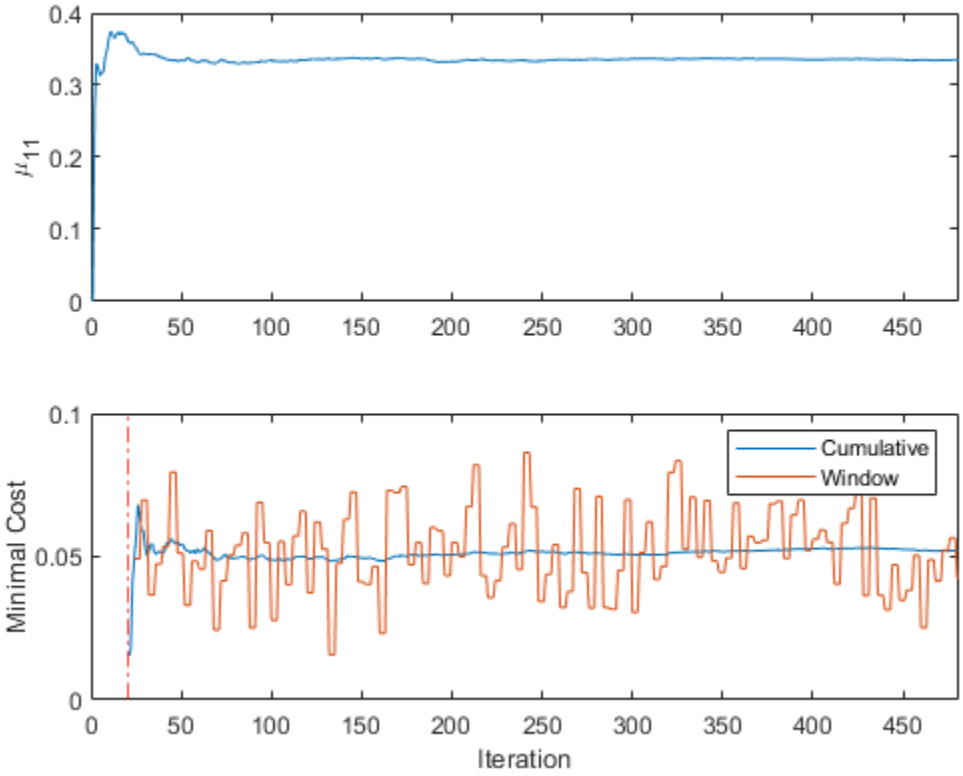
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu1 = zeros(nchunk,1);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    MdlEarly = updateMetricsAndFit(MdlEarly,X(idx,:),Y(idx));
    mc{j,:} = MdlEarly.Metrics{"MinimalCost",:};
    mu1(j + 1) = MdlEarly.DistributionParameters{1,1}(1);
end
```

`MdlEarly` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and μ_{11} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
```

```
plot(mu1)
ylabel('\mu_{11}')
xlim([0 nchunk]);
subplot(2,1,2)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
xline(MdlEarly.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mc.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit μ_{11} during all incremental learning iterations
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Experience Expected Maximum Number of Classes Late in Sample

Create a different naive Bayes model for incremental learning for the objective.

```
MdlLate = incrementalClassificationNaiveBayes('MaxNumClasses',5)
```

```
MdlLate =
  incrementalClassificationNaiveBayes
```

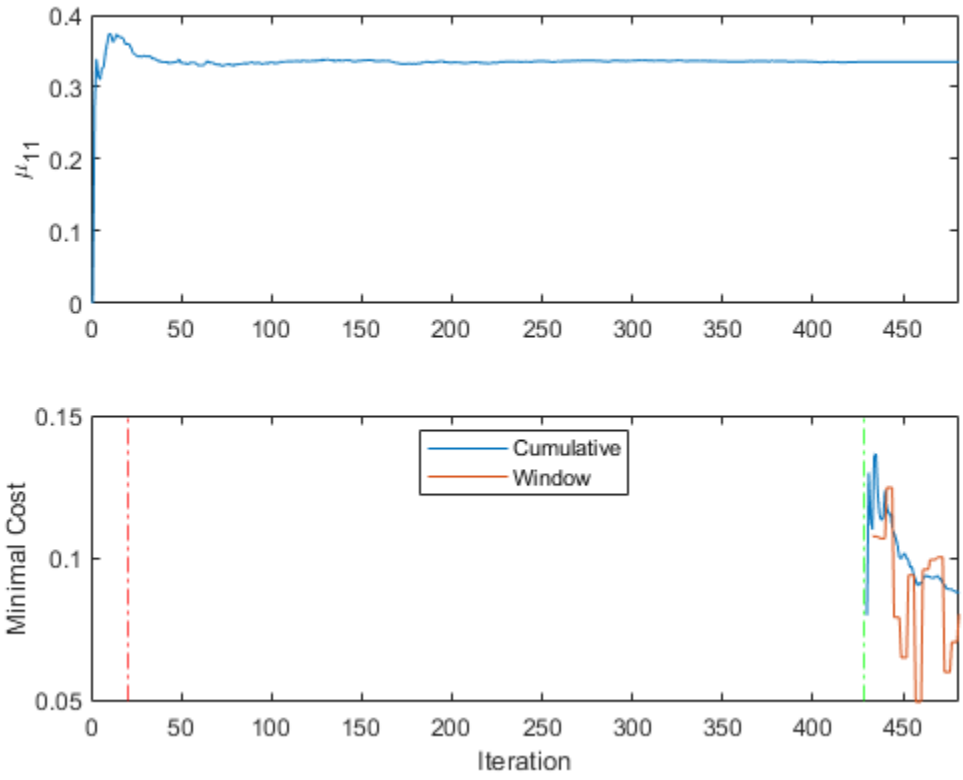
```
                  IsWarm: 0
                 Metrics: [1×2 table]
              ClassNames: [1×0 double]
          ScoreTransform: 'none'
       DistributionNames: 'normal'
  DistributionParameters: {}


  Properties, Methods
```

Move all observations labeled with class 5 to the end of the sample.

```
idx5 = Y == 5;
Xnew = [X(~idx5,:); X(idx5,:)];
Ynew = [Y(~idx5) ;Y(idx5)];
```

Fit the incremental model and plot results, as performed previously.

```
mcnew = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu1new = zeros(nchunk,1);

for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    MdlLate = updateMetricsAndFit(MdlLate,Xnew(idx,:),Ynew(idx));
    mcnew{j,:} = MdlLate.Metrics{"MinimalCost",:};
    mu1new(j + 1) = MdlLate.DistributionParameters{1,1}(1);
end

figure;
subplot(2,1,1)
plot(mu1new)
ylabel('\mu_{11}')
xlim([0 nchunk]);
subplot(2,1,2)
h = plot(mcnew.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
xline(MdlLate.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xline(sum(~idx5)/numObsPerChunk,'g-.');
legend(h,mcnew.Properties.VariableNames,'Location','best')
xlabel('Iteration')
```



The `updateMetricsAndFit` function trains throughout incremental learning, but the function starts tracking performance metrics after the model has been fit to all expected number of classes (the green vertical line in the bottom subplot).

Specify All Class Names

This example shows how to create an incremental naive Bayes learner when you know all the class names in the data.

Consider training a device to predict whether a subject is sitting, standing, walking, running, or dancing based on biometric data measured on the subject, and you know the class names map 1 through 5 to an activity.

Create an incremental naive Bayes learner for multiclass learning. Specify the class names.

```
classnames = 1:5;
Mdl = incrementalClassificationNaiveBayes('ClassNames',classnames)
```

```
Mdl =
  incrementalClassificationNaiveBayes

                  IsWarm: 0
                 Metrics: [1x2 table]
              ClassNames: [1 2 3 4 5]
          ScoreTransform: 'none'
```

```
        DistributionNames: 'normal'
    DistributionParameters: {5x0 cell}


  Properties, Methods
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Fit the incremental model to the training data by using the `updateMetricsAndfit` function. Simulate a data stream by processing chunks of 50 observations at a time. At each iteration:

- Process 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
end
```

Configure Incremental Learning Options

In addition to specifying the maximum number of class names, prepare an incremental naive Bayes learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Create an incremental linear model for binary classification. Configure the model as follows:

- Specify a metrics warm-up period of 5000 observations.
- Specify a metrics window size of 500 observations.
- Track the classification error and minimal cost to measure the performance of the model.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5,'MetricsWarmupPeriod',5000,...
    'MetricsWindowSize',500,'Metrics',["classiferror" "mincost"])
```

```
Mdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 0
                  Metrics: [2x2 table]
               ClassNames: [1x0 double]
           ScoreTransform: 'none'
        DistributionNames: 'normal'
    DistributionParameters: {}


  Properties, Methods
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object configured for incremental learning.

Fit the incremental model to the rest of the data by using the `updateMetricsAndfit` function. At each iteration:

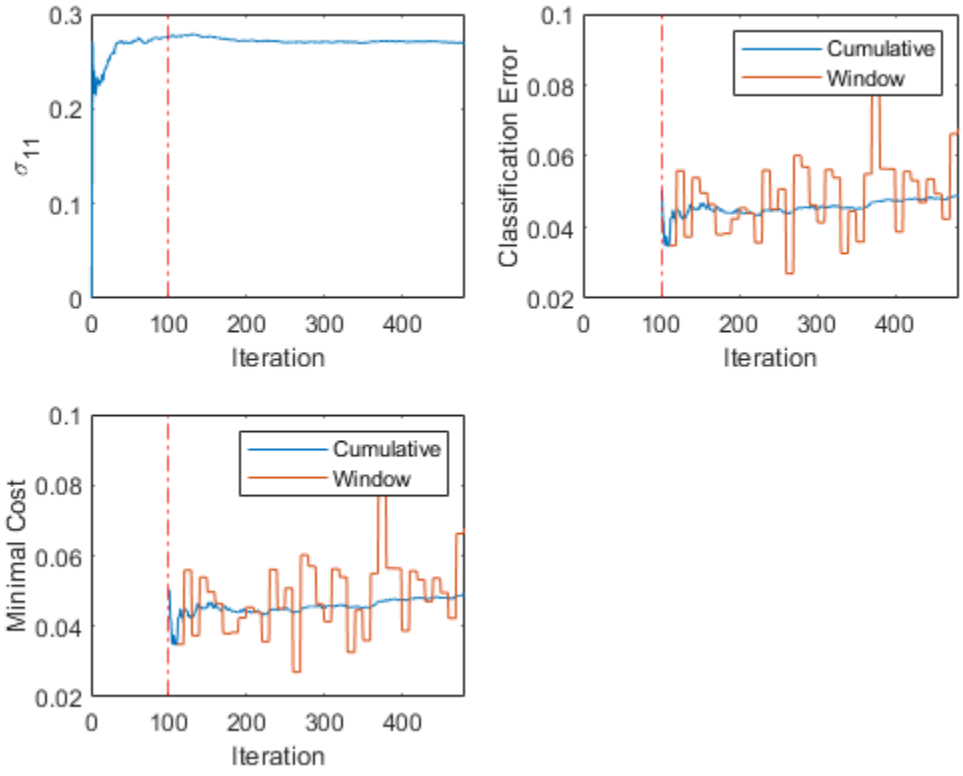
- Simulate a data stream by processing a chunk of 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store σ_{11} (the standard deviation of the first predictor variable in the first class), the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
sigma11 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    mc{j,:} = Mdl.Metrics{"MinimalCost",:};
    sigma11(j + 1) = Mdl.DistributionParameters{1,1}(2);
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and σ_{11} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(sigma11)
ylabel('\sigma_{11}')
xlim([0 nchunk]);
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
subplot(2,2,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ce.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,3)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mc.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit σ_{11} during all incremental learning iterations

- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (10 iterations).

Convert Traditionally Trained Model to Incremental Learner

Train a naive Bayes model for multiclass classification by using `fitcnb`, convert it to an incremental learner, track its performance, and fit it to streaming data. Carry over training options from traditional to incremental learning.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Suppose that the data collected when the subject was idle (`Y > 2`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from an idle subject, and 1 to a moving subject.

```
W = ones(n,1) + (Y < 3);
```

Train Naive Bayes Model

Fit a naive Bayes model for multiclass classification to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitcnb(X(idxtt,:),Y(idxtt),'Weights',W(idxtt))
```

```
TTMdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 12053
         DistributionNames: {1x60 cell}
    DistributionParameters: {5x60 cell}


  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained naive Bayes model.

Convert Trained Model

Convert the traditionally trained naive Bayes model to a naive Bayes classification model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 1
                  Metrics: [1x2 table]
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
        DistributionNames: {1x60 cell}
   DistributionParameters: {5x60 cell}


  Properties, Methods
```

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. Simulate a data stream by processing 50 observations at a time. At each iteration:

1. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify the observation weights.
2. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify the observation weights.
3. Store the minimal cost and mean of the first predictor variable of the first class μ_{11}.

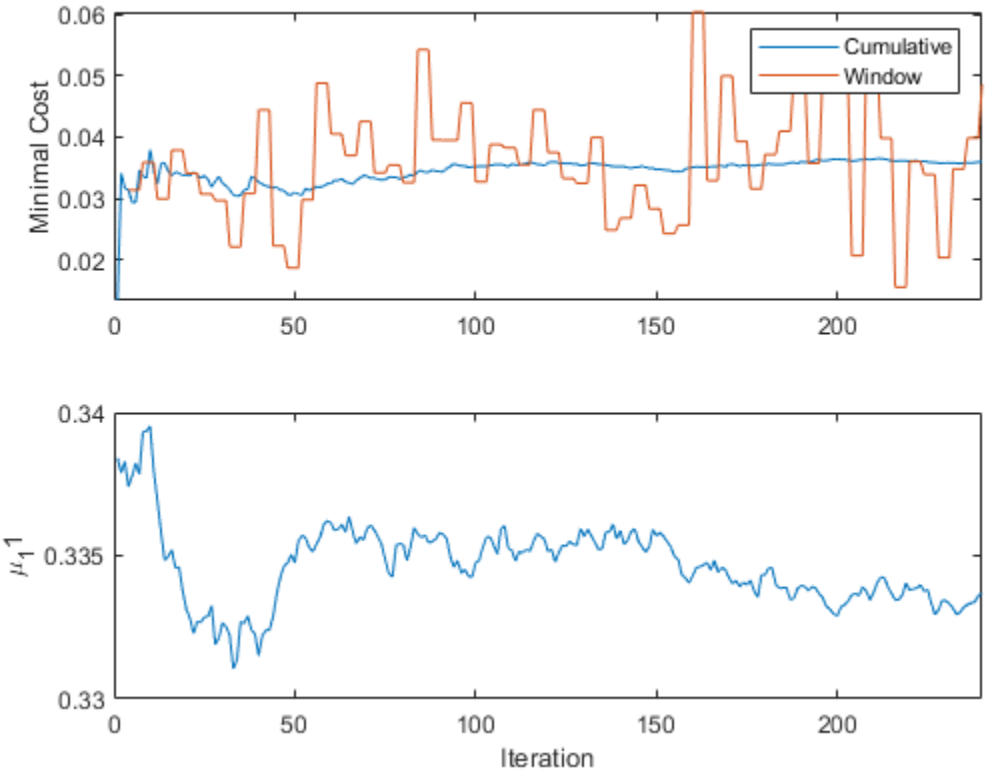
```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu11 = [IncrementalMdl.DistributionParameters{1,1}(1); zeros(nchunk,1)];
Xil = X(idxil,:);
Yil = Y(idxil);
Wil = W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(idx,:),Yil(idx),...
        'Weights',Wil(idx));
    mc{j,:} = IncrementalMdl.Metrics{"MinimalCost",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx),'Weights',Wil(idx));
    mu11(j + 1) = IncrementalMdl.DistributionParameters{1,1}(1);
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Plot a trace plot of the performance metrics and estimated coefficient μ_{11}.

```
figure;
subplot(2,1,1)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
legend(h,mc.Properties.VariableNames)
subplot(2,1,2)
plot(mu11)
ylabel('\mu_11')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable and decreases gradually, whereas the window loss jumps.

μ_{11} changes abruptly at first, then gradually levels off as `fit` processes more chunks.

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Algorithms

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm when `fit` or `updateMetricsAndFit` perform both of the following actions:
  - Fit the incremental model to MetricsWarmupPeriod observations, which is the metrics warm-up period.
  - Process `MaxNumClasses` classes or all class names specified by the `ClassNames` name-value argument.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:
  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the MetricsWindowSize name-value pair argument. `MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observation weights in the weights buffer.
3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

See Also

Functions

`fit` | `fitcnb` | `incrementalLearner` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics

“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2021a

incrementalLearner

Convert linear model for binary classification to incremental learner

Syntax

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl = incrementalLearner(Mdl,Name,Value)
```

Description

`IncrementalMdl = incrementalLearner(Mdl)` returns a binary classification linear model for incremental learning on page 33-2948, `IncrementalMdl`, using the hyperparameters and coefficients of the traditionally trained linear model for binary classification, `Mdl`. Because its property values reflect the knowledge gained from `Mdl`, `IncrementalMdl` can predict labels given new observations, and it is warm, meaning that its predictive performance is tracked.

`IncrementalMdl = incrementalLearner(Mdl,Name,Value)` uses additional options specified by one or more name-value pair arguments. Some options require you to train `IncrementalMdl` before its predictive performance is tracked. For example, `'MetricsWarmupPeriod',50,'MetricsWindowSize',100` specifies a preliminary incremental training period of 50 observations before performance metrics are tracked, and specifies processing 100 observations before updating the performance metrics.

Examples

Convert Traditionally Trained Model to Incremental Learner

Train a linear classification model for binary learning by using `fitclinear`, and then convert it to an incremental learner.

Load and Preprocess Data

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: `Sitting`, `Standing`, `Walking`, `Running`, or `Dancing`. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Train Linear Classification Model

Fit a linear classification model to the entire data set.

```
TTMdl = fitclinear(feat,Y)
```

```
TTMdl =
  ClassificationLinear
```

```
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2005
            Lambda: 4.1537e-05
           Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear classification model.

Convert Trained Model

Convert the traditionally trained linear classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

             IsWarm: 1
            Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2005
           Learner: 'svm'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object prepared for incremental learning using SVM.

- The `incrementalLearner` function Initializes the incremental learner by passing learned coefficients to it, along with other information `TTMdl` extracted from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.
- The `incrementalLearner` function trains the model using the adaptive scale-invariant solver, whereas `fitclinear` trained `TTMdl` using the BFGS solver.

Predict Responses

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict classification scores for all observations using both models.

```
[~,ttscores] = predict(TTMdl,feat);
[~,ilscores] = predict(IncrementalMdl,feat);
compareScores = norm(ttscores(:,1) - ilscores(:,1))
```

```
compareScores = 0
```

The difference between the scores generated by the models is 0.

Specify SGD Solver and Standardize Predictor Data

If you train a linear classification model using the SGD or ASGD solver, `incrementalLearner` preserves the solver, linear model type, and associated hyperparameter values when it converts the linear classification model.

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Waling, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Randomly split the data in half: the first half for training a model traditionally, and the second half for incremental learning.

```
n = numel(Y);
rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.5);
idxtt = training(cvp);
idxil = test(cvp);

% First half of data
Xtt = feat(idxtt,:);
Ytt = Y(idxtt);

% Second half of data
Xil = feat(idxil,:);
Yil = Y(idxil);
```

Create a set of 11 logarithmically spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Because the variables are on different scales, use implicit expansion to standardize the predictor data.

```
Xtt = (Xtt - mean(Xtt))./std(Xtt);
```

Tune the L2 regularization parameter by applying 5-fold cross-validation. Specify the standard SGD solver.

```
TTCVMdl = fitclinear(Xtt,Ytt,'KFold',5,'Learner','logistic',...
    'Solver','sgd','Lambda',Lambda);
```

`TTCVMdl` is a `ClassificationPartitionedLinear` model representing the five models created during cross-validation (see `TTCVMdl.Trained`). The cross-validation procedure includes training with each specified regularization value.

Compute the cross-validated classification error for each model and regularization.

```
cvloss = kfoldLoss(TTCVMdl)
```

```
cvloss = 1×11

    0.0054    0.0039    0.0034    0.0033    0.0030    0.0027    0.0027    0.0031    0.0036    0.0
```

`cvloss` contains the test-sample classification loss for each regularization value in `Lamba`.

Select the regularization value that minimizes the classification error. Train the model again using the selected regularization value.

```
[~,idxmin] = min(cvloss);
TTMdl = fitclinear(Xtt,Ytt,'Learner','logistic','Solver','sgd',...
    'Lambda',Lambda(idxmin));
```

`TTMdl` is a `ClassificationLinear` model.

Convert the traditionally trained linear classification model to a binary classification linear model for incremental learning. Specify the standard SGD solver. Prepare incremental learning functions to standardize the predictors. This action requires an initial period for estimating the predictor means and standard deviations. Specify an estimation period of `2000` observations (the default is `1000` when predictor moments are required).

```
IncrementalMdl = incrementalLearner(TTMdl,'Standardize',true,'EstimationPeriod',2000);
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object. `incrementalLearner` passes the solver and regularization strength, among other information learned from training `TTMdl`, to `IncrementalMdl`.

Fit the incremental model to the second half of the data by using the `fit` function. At each iteration:

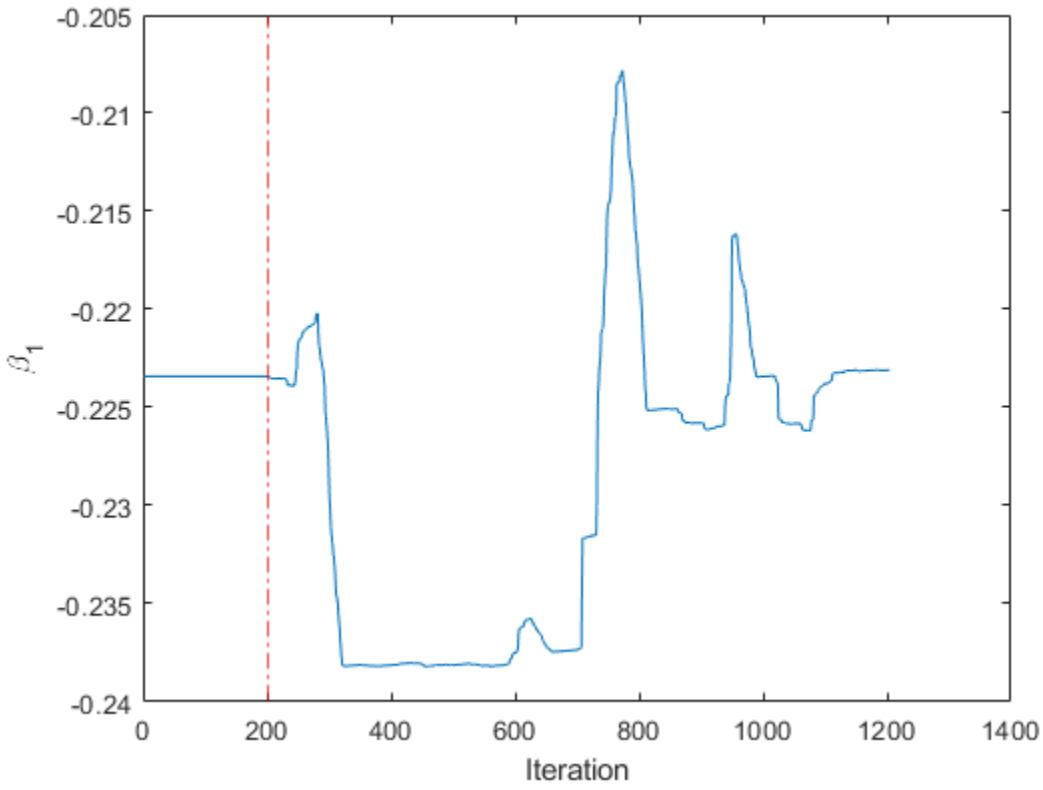
- Simulate a data stream by processing 10 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1 to see how it evolves during training.

```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 10;
nchunk = floor(nil/numObsPerChunk);
learnrate = [IncrementalMdl.LearnRate; zeros(nchunk,1)];
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Plot β_1 to see how it evolved.

```
plot(beta1)
ylabel('\beta_1')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
```



Because `fit` does not fit the model to the streaming data during the estimation period, β_1 is constant for the first 200 iterations (2000 observations). Then, β_1 changes abruptly during incremental fitting.

Configure Performance Metric Options

Use a trained linear classification model to initialize an incremental learner. Prepare the incremental learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, and Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Because the data set is grouped by activity, shuffle it for simplicity. Then, randomly split the data in half: the first half for training a model traditionally, and the second half for incremental learning.

```
n = numel(Y);

rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.5);
idxtt = training(cvp);
idxil = test(cvp);
shuffidx = randperm(n);
X = feat(shuffidx,:);
Y = Y(shuffidx);

% First half of data
Xtt = X(idxtt,:);
Ytt = Y(idxtt);

% Second half of data
Xil = X(idxil,:);
Yil = Y(idxil);
```

Fit a linear classification model to the first half of the data.

```
TTMdl = fitclinear(Xtt,Ytt);
```

Convert the traditionally trained linear classification model to a binary classification linear model for incremental learning. Specify the following:

- A performance metrics warm-up period of 2000 observations
- A metrics window size of 500 observations
- Use of classification error and hinge loss to measure the performance of the model

```
IncrementalMdl = incrementalLearner(TTMdl,'MetricsWarmupPeriod',2000,'MetricsWindowSize',500,...
    'Metrics',["classiferror" "hinge"]);
```

Fit the incremental model to the second half of the data by using the `updateMetricsAndFit` function. At each iteration:

- Simulate a data stream that processing a chunk of 20 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

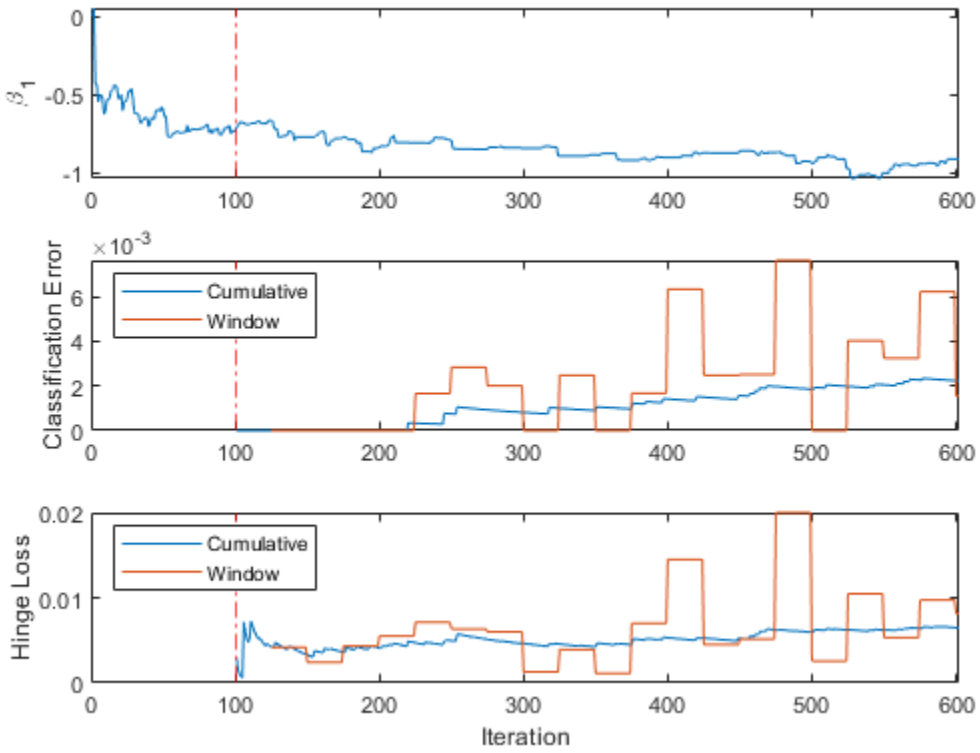
```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 20;
nchunk = ceil(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
hinge = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
```

```
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    hinge{j,:} = IncrementalMdl.Metrics{"HingeLoss",:};
    beta1(j + 1) = IncrementalMdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(3,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
subplot(3,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ce.Properties.VariableNames,'Location','northwest')
subplot(3,1,3)
h = plot(hinge.Variables);
xlim([0 nchunk]);
ylabel('Hinge Loss')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,hinge.Properties.VariableNames,'Location','northwest')
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_1 during all incremental learning iterations.
- Compute the performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (25 iterations).

Input Arguments

`Mdl` — Traditionally trained linear model for binary classification

`ClassificationLinear` model object

Traditionally trained linear model for binary classification, specified as a `ClassificationLinear` model object returned by `fitclinear`.

Note

- If `Mdl.Lambda` is a numeric vector, you must select the model corresponding to one regularization strength in the regularization path by using `selectModels`.
- Incremental learning functions support only numeric input predictor data. If `Mdl` was fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of

dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','scale-invariant','MetricsWindowSize',100` specifies the adaptive scale-invariant solver for objective optimization, and specifies processing 100 observations before updating the performance metrics.

General Options

`Solver` — Objective function minimization technique

`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-2948 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

Example: `'Solver','sgd'`

Data Types: `char` | `string`

`EstimationPeriod` — Number of observations processed to estimate hyperparameters

nonnegative integer

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as the comma-separated pair consisting of `'EstimationPeriod'` and a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalLearner` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalLearner` sets `'EstimationPeriod'` to `1000`.

For more details, see “Estimation Period” on page 33-2948.

Example: `'EstimationPeriod',100`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`false` (default) | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`true`	The software standardizes the predictor data. For more details, see “Standardize Data” on page 33-2949.
`false`	The software does not standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size

positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration during training, `incrementalLearner` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'BatchSize'`. Instead, `incrementalLearner` sets `'BatchSize'` to `Mdl.ModelParameters.BatchSize`.
- Otherwise, `BatchSize` is `10`.

Example: `'BatchSize',1`

Data Types: `single` | `double`

`Lambda` — Ridge (*L*2) regularization term strength

nonnegative scalar

Ridge (*L*2) regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

When `Mdl.Regularization` is `'ridge (L2)'`:

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'Lambda'`. Instead, `incrementalLearner` sets `'Lambda'` to `Mdl.Lambda`.
- Otherwise, `Lambda` is `1e-5`.

Note `incrementalLearner` does not support lasso regularization. If `Mdl.Regularization` is `'lasso (L1)'`, `incrementalLearner` uses ridge regularization instead, and sets the `'Solver'` name-value pair argument to `'scale-invariant'` by default.

Example: `'Lambda',0.01`

Data Types: `single` | `double`

`LearnRate` — Learning rate
`'auto'` | positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'LearnRate'`. Instead, `incrementalLearner` sets `'LearnRate'` to `Mdl.ModelParameters.LearnRate`.
- Otherwise, `LearnRate` is `'auto'`.

For `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either function.

The name-value pair argument `'LearnRateSchedule'` determines the learning rate for subsequent learning cycles.

Example: `'LearnRate',0.001`

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule
`'decaying'` (default) | `'constant'`

Learning rate schedule, specified as the comma-separated pair consisting of `'LearnRateSchedule'` and a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.

Value	Description
`'decaying'`	The learning rate at learning cycle t is $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$ • λ is the value of `Lambda`. • If `Solver` is `'sgd'`, then $c = 1$. • If `Solver` is `'asgd'`, then c is 0.75 [4].

If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'LearnRateSchedule'`.

Example: `'LearnRateSchedule','constant'`

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch

`true` (default) | `false`

Flag for shuffling the observations in the batch at each iteration, specified as the comma-separated pair consisting of `'Shuffle'` and a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Example: `'Shuffle',false`

Data Types: `logical`

Performance Metrics Options

`Metrics` — Model performance metrics to track during incremental learning

`"classiferror"` (default) | string vector | function handle | cell vector | structure array | `"binodeviance"` | `"exponential"` | `"hinge"` | `"logit"` | `"quadratic"` | ...

Model performance metrics to track during incremental learning with the `updateMetrics` or `updateMetricsAndFit` function, specified as a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"binodeviance"`	Binomial deviance

Name	Description
"classiferror"	Classification error
"exponential"	Exponential
"hinge"	Hinge
"logit"	Logistic
"quadratic"	Quadratic

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["classiferror" "hinge"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(C,S)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data by processing the incremental learning functions during a learning cycle.
- You specify the function name (`customMetric`).
- `C` is an *n*-by-2 logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-2 numeric matrix of predicted classification scores. `S` is similar to the `score` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in `Mdl.ClassNames`. `S(p,q)` is the classification score of observation *p* being classified in class *q*.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'logit' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `IncrementalMdl.Metrics`. The data type of `Metrics` determines the row names of the table.

'Metrics' Value Data Type	Description of Metrics Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"classiferror"` is `"ClassificationError"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`

'Metrics' Value Data Type	Description of Metrics Property Row Name	Example
Anonymous function	`CustomMetric_j`, where `j` is metric `j` in `Metrics`	Row name for `@(C,S)customMetric(C,S)...` is `CustomMetric_1`

For more details on performance metrics options, see “Performance Metrics” on page 33-2950.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`0` (default) | nonnegative integer | ...

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as the comma-separated pair consisting of `'MetricsWarmupPeriod'` and a nonnegative integer. The incremental model is warm after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (`EstimationPeriod` + `MetricsWarmupPeriod` observations).

For more details on performance metrics options, see “Performance Metrics” on page 33-2950.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer | ...

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2950.

Data Types: `single` | `double`

Output Arguments

`IncrementalMdl` — Binary classification linear model for incremental learning
`incrementalClassificationLinear` model object

Binary classification linear model for incremental learning, returned as an `incrementalClassificationLinear` model object. `IncrementalMdl` is also configured to generate predictions given new data (see `predict`).

To initialize `IncrementalMdl` for incremental learning, `incrementalLearner` passes the values of the `Mdl` properties in this table to congruent properties of `IncrementalMdl`.

Property	Description
`Beta`	Linear model coefficients, a numeric vector
`Bias`	Model intercept, a numeric scalar
`ClassNames`	Class labels for binary classification, a two-element list
`ModelParameters.FitBias`	Linear model intercept inclusion flag

Property	Description
`Learner`	Linear classification model type
`Prior`	Prior class label distribution, a numeric vector

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for classification `fit` and `updateMetricsAndFit` use the more aggressive ScInOL2 version of the algorithm.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When you set `'Standardize',true` (see "Standardize Data" on page 33-2949)
Learning rate	`LearnRate`	Adjust solver step size	When both of these conditions apply: • You change the solver of `Mdl` to SGD or ASGD (see `Solver`). • You do not set the `'LearnRate'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `IncrementalMdl`.

- When you set `'Standardize',true`, and `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, the following conditions apply:
  - If the estimation period is positive (see the `EstimationPeriod` property of `IncrementalMdl`), incremental fitting functions estimate means and standard deviations using the estimation period observations.
  - If the estimation period is 0, `incrementalLearner` forces the estimation period to `1000`. Consequently, incremental fitting functions estimate new predictor variable means and standard deviations during the forced estimation period.
- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  where
  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k^*} \sum_k w_k^* x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k^*} \sum_k w_k^* \left(x_{jk} - \mu_j^*\right)^2$.
  - $w_j^* = \frac{w_j}{\sum_{\forall j \in \text{Class } k} w_j} p_k$, where

- p_k is the prior probability of class k (`Prior` property of the incremental model).
- w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions are incremental learning functions that track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to `'MetricsWarmupPeriod'` observations, which is the metrics warm-up period.

  If `'EstimationPeriod'` > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:

  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the `'MetricsWindowSize'` name-value pair argument. `'MetricsWindowSize'` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `IncrementalMdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

See Also

Objects
`ClassificationLinear` | `incrementalClassificationLinear`

Functions
`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2020b

incrementalLearner

Convert naive Bayes classification model to incremental learner

Syntax

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl = incrementalLearner(Mdl,Name,Value)
```

Description

`IncrementalMdl = incrementalLearner(Mdl)` returns a naive Bayes classification model for incremental learning on page 33-2959, `IncrementalMdl`, using the hyperparameters of the traditionally trained naive Bayes classification model, `Mdl`. Because its property values reflect the knowledge gained from `Mdl`, `IncrementalMdl` can predict labels given new observations, and it is warm, meaning that its predictive performance is tracked.

`IncrementalMdl = incrementalLearner(Mdl,Name,Value)` uses additional options specified by one or more name-value pair arguments. Some options require you to train `IncrementalMdl` before its predictive performance is tracked. For example, `'MetricsWarmupPeriod',50,'MetricsWindowSize',100` specifies a preliminary incremental training period of 50 observations before performance metrics are tracked, and specifies processing 100 observations before updating the performance metrics.

Examples

Convert Traditionally Trained Model to Incremental Learner

Train a naive Bayes model by using `fitcnb`, and then convert it to an incremental learner.

Load and Preprocess Data

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Train Naive Bayes Model

Fit a naive Bayes classification model to the entire data set.

```
TTMdl = fitcnb(feat,actid);
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained naive Bayes classification model.

Convert Trained Model

Convert the traditionally trained naive Bayes classification model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 1
                  Metrics: [1x2 table]
              ClassNames: [1 2 3 4 5]
          ScoreTransform: 'none'
        DistributionNames: {1x60 cell}
   DistributionParameters: {5x60 cell}


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object prepared for incremental learning using naive Bayes classification.

- The `incrementalLearner` function Initializes the incremental learner by passing learned conditional predictor distribution parameters to it, along with other information `TTMdl` extracted from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can track performance metrics and make predictions.

Predict Responses

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict classification scores (class posterior probabilities) for all observations using both models.

```
[~,ttscores] = predict(TTMdl,feat);
[~,ilcores] = predict(IncrementalMdl,feat);
compareScores = norm(ttscores - ilcores)
```

```
compareScores = 0
```

The difference between the scores generated by the models is 0.

Configure Performance Metric Options

Use a trained naive Bayes model to initialize an incremental learner. Prepare the incremental learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line

Randomly split the data in half: the first half for training a model traditionally, and the second half for incremental learning.

```
n = numel(actid);
```

```
rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.5);
idxtt = training(cvp);
idxil = test(cvp);

% First half of data
Xtt = feat(idxtt,:);
Ytt = actid(idxtt);

% Second half of data
Xil = feat(idxil,:);
Yil = actid(idxil);
```

Fit a naive Bayes model to the first half of the data.

```
TTMdl = fitcnb(Xtt,Ytt);
```

Convert the traditionally trained naive Bayes model to a naive Bayes classification model for incremental learning. Specify the following:

- A performance metrics warm-up period of 2000 observations
- A metrics window size of 500 observations
- Use of classification error and minimal cost to measure the performance of the model

```
IncrementalMdl = incrementalLearner(TTMdl,'MetricsWarmupPeriod',2000,'MetricsWindowSize',500,...
    'Metrics',["classiferror" "hinge"]);
```

Fit the incremental model to the second half of the data by using the `updateMetricsAndfit` function. At each iteration:

- Simulate a data stream by processing 20 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the mean of the second predictor within the first class μ_{12}, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

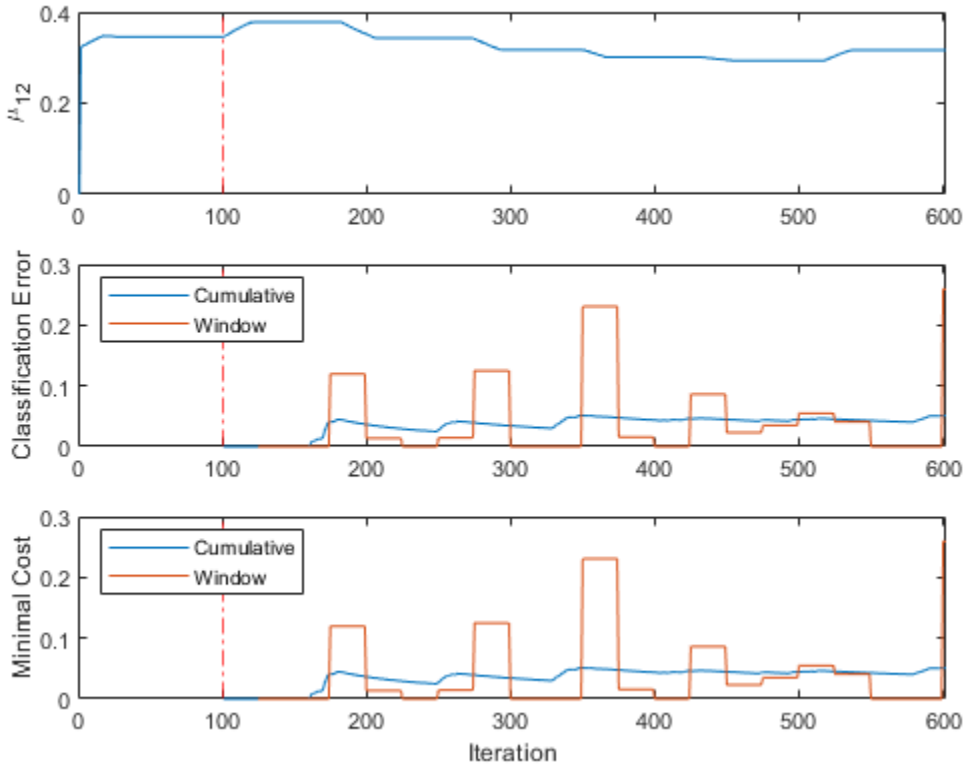
```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 20;
nchunk = ceil(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu12 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    mc{j,:} = IncrementalMdl.Metrics{"MinimalCost",:};
    mu12(j + 1) = IncrementalMdl.DistributionParameters{1,2}(1);
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up,

`updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and μ_{12} evolved during training, plot them on separate subplots.

```
figure;
subplot(3,1,1)
plot(mu12)
ylabel('\mu_{12}')
xlim([0 nchunk]);
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
subplot(3,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ce.Properties.VariableNames,'Location','northwest')
subplot(3,1,3)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mc.Properties.VariableNames,'Location','northwest')
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit μ_{12} during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (25 iterations).

Because the data is ordered by activity, the mean and performance metrics periodically change abruptly.

Input Arguments

`Mdl` — Traditionally trained naive Bayes model for multiclass classification

`ClassificationNaiveBayes` model object

Traditionally trained naive Bayes model for multiclass classification, specified as a `ClassificationNaiveBayes` model object returned by `fitcnb`. The conditional distribution of each predictor variable, as stored in `Mdl.DistributionNames`, must be normal.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Metrics',["classiferror" "mincost"],'MetricsWindowSize',100` specifies tracking the misclassification rate and minimal cost, and specifies processing 100 observations before updating the performance metrics.

`Metrics` — Model performance metrics to track during incremental learning

`"mincost"` (default) | `"classiferror"` | string vector | function handle | cell vector | structure array | `"binodeviance"` | `"exponential"` | `"hinge"` | `"logit"` | `"quadratic"` | ...

Model performance metrics to track during incremental learning with the `updateMetrics` or `updateMetricsAndFit` function, specified as a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"binodeviance"`	Binomial deviance
`"classiferror"`	Classification error
`"exponential"`	Exponential
`"hinge"`	Hinge
`"logit"`	Logistic
`'"mincost"`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)

Name	Description
`"quadratic"`	Quadratic

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["classiferror" "mincost"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(C,S,Cost)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You select the function name (`customMetric`).
- `C` is an *n*-by-*K* logical matrix with rows indicating the class to which the corresponding observation belongs, where *K* is the number of classes. The column order corresponds to the class order in the `ClassNames` property. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-*K* numeric matrix of predicted classification scores. `S` is similar to the `Score` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in the `ClassNames` property. `S(p,q)` is the classification score of observation *p* being classified in class *q*.
- `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. See the `'Cost'` name-value argument.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'logit' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `IncrementalMdl.Metrics`. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"classiferror"` is `"ClassificationError"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(C,S,Cost)customMetric(C,S,Cost)...` is `CustomMetric_1`

For more details on performance metrics options, see “Performance Metrics” on page 33-2959.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`1000` (default) | nonnegative integer

This property is read-only.

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as a nonnegative integer.

For more details, see “Performance Metrics” on page 33-2959.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer

This property is read-only.

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2959.

Data Types: `single` | `double`

Output Arguments

`IncrementalMdl` — Naive Bayes classification for incremental learning
`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning, returned as an `incrementalClassificationNaiveBayes` model object. `IncrementalMdl` is also configured to generate predictions given new data (see `predict`).

To initialize `IncrementalMdl` for incremental learning, `incrementalLearner` passes the values of the properties of `Mdl` in this table to congruent properties of `IncrementalMdl`.

Property	Description
`ClassNames`	Class labels for binary classification, a list of names
`Cost`	Misclassification costs, a numeric matrix
`DistributionNames`	Names of the conditional distributions of the predictor variables, a cell array in which each cell contains `'normal`
`DistributionParameters`	Parameter values of the conditional distributions of the predictor variables, a cell array of length 2 numeric vectors (for details, see `DistributionParameters`)
`Prior`	Prior class label distribution, a numeric vector
`ScoreTransform`	Score transformation function, a name or function handle.
`Y`	Class labels supplied to `fitcnb` to compute per class weights for incremental learning, stored as an array of labels (see `Y`)

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Algorithms

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm when `fit` or `updateMetricsAndFit` perform both of the following actions:
  - Fit the incremental model to MetricsWarmupPeriod observations, which is the metrics warm-up period.
  - Process `MaxNumClasses` classes or all class names specified by the `ClassNames` name-value argument.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:
  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the MetricsWindowSize name-value pair argument. `MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.

2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observation weights in the weights buffer.
3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8

Introduced in R2021a

incrementalLearner

Convert binary classification support vector machine (SVM) model to incremental learner

Syntax

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl = incrementalLearner(Mdl,Name,Value)
```

Description

`IncrementalMdl = incrementalLearner(Mdl)` returns a binary classification linear model for incremental learning on page 33-2974, `IncrementalMdl`, using the hyperparameters and coefficients of the traditionally trained linear SVM model for binary classification, `Mdl`. Because its property values reflect the knowledge gained from `Mdl`, `IncrementalMdl` can predict labels given new observations, and it is warm, meaning that its predictive performance is tracked.

`IncrementalMdl = incrementalLearner(Mdl,Name,Value)` uses additional options specified by one or more name-value pair arguments. Some options require you to train `IncrementalMdl` before its predictive performance is tracked. For example, `'MetricsWarmupPeriod',50,'MetricsWindowSize',100` specifies a preliminary incremental training period of 50 observations before performance metrics are tracked, and specifies processing 100 observations before updating the performance metrics.

Examples

Convert Traditionally Trained Model to Incremental Learner

Train an SVM model by using `fitcsvm`, and then convert it to an incremental learner.

Load and Preprocess Data

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: `Sitting`, `Standing`, `Walking`, `Running`, or `Dancing`. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Train SVM Model

Fit an SVM model to the entire data set. Discard the support vectors (`Alpha`) from the model so that the software uses the linear coefficients (`Beta`) for prediction.

```
TTMdl = fitcsvm(feat,Y);
TTMdl = discardSupportVectors(TTMdl)
```

```
TTMdl =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [0 1]
           ScoreTransform: 'none'
          NumObservations: 24075
                     Beta: [60x1 double]
                     Bias: -6.4221
         KernelParameters: [1x1 struct]
           BoxConstraints: [24075x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [24075x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`TTMdl` is a `ClassificationSVM` model object representing a traditionally trained SVM model.

Convert Trained Model

Convert the traditionally trained SVM model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

            IsWarm: 1
           Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -6.4221
           Learner: 'svm'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object prepared for incremental learning using SVM.

- The `incrementalLearner` function Initializes the incremental learner by passing learned coefficients to it, along with other information `TTMdl` extracted from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.
- The `incrementalLearner` function specifies to train the model using the adaptive scale-invariant solver, whereas `fitcsvm` trained `TTMdl` using the `SMO` solver.

Predict Responses

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict classification scores for all observations using both models.

```
[~,ttscores] = predict(TTMdl,feat);
[~,ilcores] = predict(IncrementalMdl,feat);
compareScores = norm(ttscores(:,1) - ilcores(:,1))
```

```
compareScores = 0
```

The difference between the scores generated by the models is 0.

Specify SGD Solver and Standardize Predictor Data

The default solver is the adaptive scale-invariant solver. If you specify this solver, you do not need to tune any parameters for training. However, if you specify either the standard SGD or ASGD solver instead, you can also specify an estimation period, during which the incremental fitting functions tune the learning rate.

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: `Sitting`, `Standing`, `Walking`, `Running`, and `Dancing`. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Randomly split the data in half: the first half have for training a model traditionally, and the second half for incremental learning.

```
n = numel(Y);

rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.5);
idxtt = training(cvp);
idxil = test(cvp);

% First half of data
Xtt = feat(idxtt,:);
Ytt = Y(idxtt);

% Second half of data
Xil = feat(idxil,:);
Yil = Y(idxil);
```

Fit an SVM model to the first half of the data. Standardize the predictor data by setting `'Standardize',true`.

```
TTMdl = fitcsvm(Xtt,Ytt,'Standardize',true);
```

The `Mu` and `Sigma` properties of `TTMdl` contain the predictor data sample means and standard deviations, respectively.

Suppose that the distribution of the predictors is not expected to change in the future. Convert the traditionally trained SVM model to a binary classification linear model for incremental learning.

Specify the standard SGD solver and an estimation period of `2000` observations (the default is `1000` when a learning rate is required).

```
IncrementalMdl = incrementalLearner(TTMdl,'Solver','sgd','EstimationPeriod',2000);
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object. Because the predictor data of `TTMdl` is standardized (`TTMdl.Mu` and `TTMdl.Sigma` are nonempty), `incrementalLearner` prepares incremental learning functions to standardize supplied predictor data by using the previously learned moments (stored in `IncrementalMdl.Mu` and `IncrementalMdl.Sigma`).

Fit the incremental model to the second half of the data by using the `fit` function. At each iteration:

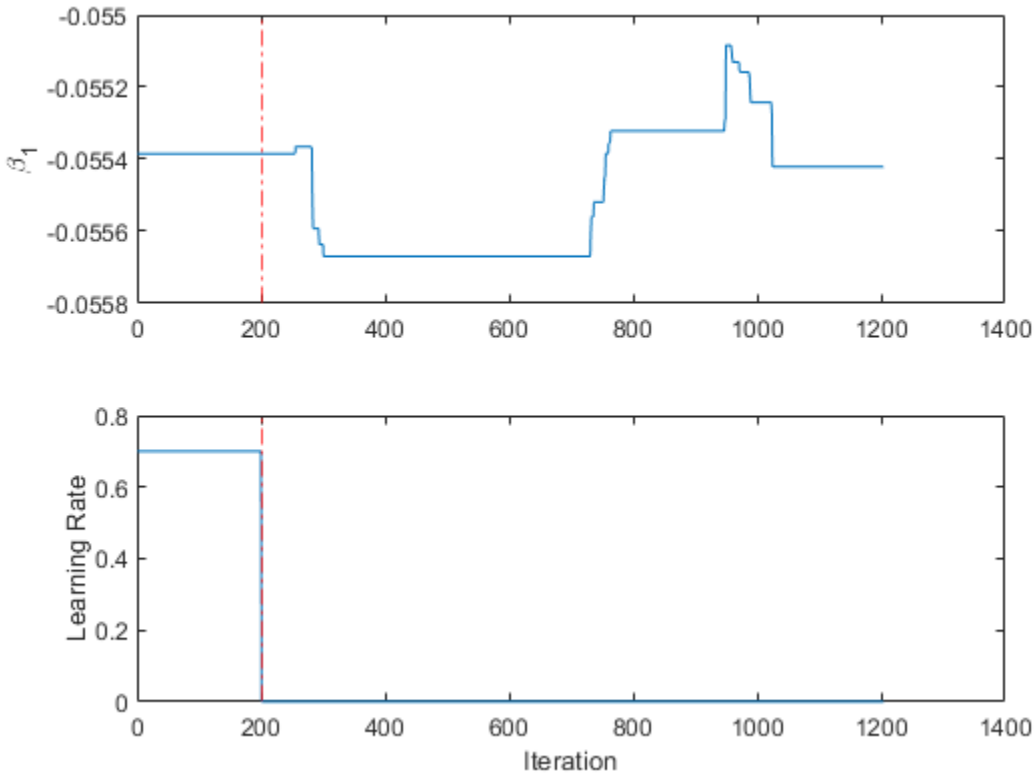
- Simulate a data stream by processing 10 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the learning rate and β_1 to see how the coefficients and learning rate evolve during training.

```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 10;
nchunk = floor(nil/numObsPerChunk);
learnrate = [IncrementalMdl.LearnRate; zeros(nchunk,1)];
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(1);
    learnrate(j + 1) = IncrementalMdl.LearnRate;
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

To see how the learning rate and β_1 evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
plot(learnrate)
ylabel('Learning Rate')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
```



The learning rate jumps to its auto-tuned value after the estimation period.

Because `fit` does not fit the model to the streaming data during the estimation period, β_1 is constant for the first 200 iterations (2000 observations). Then, β_1 changes during incremental fitting.

Configure Performance Metric Options

Use a trained SVM model to initialize an incremental learner. Prepare the incremental learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line

Responses can be one of five classes: `Sitting`, `Standing`, `Walking`, `Running`, and `Dancing`. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Because the data set is grouped by activity, shuffle it to reduce bias. Then, randomly split the data in half: the first half for training a model traditionally, and the second half for incremental learning.

```
n = numel(Y);

rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.5);
idxtt = training(cvp);
idxil = test(cvp);
shuffidx = randperm(n);
X = feat(shuffidx,:);
Y = Y(shuffidx);

% First half of data
Xtt = X(idxtt,:);
Ytt = Y(idxtt);

% Second half of data
Xil = X(idxil,:);
Yil = Y(idxil);
```

Fit an SVM model to the first half of the data.

```
TTMdl = fitcsvm(Xtt,Ytt);
```

Convert the traditionally trained SVM model to a binary classification linear model for incremental learning. Specify the following:

- A performance metrics warm-up period of 2000 observations
- A metrics window size of 500 observations
- Use of classification error and hinge loss to measure the performance of the model

```
IncrementalMdl = incrementalLearner(TTMdl,'MetricsWarmupPeriod',2000,'MetricsWindowSize',500,...
    'Metrics',["classiferror" "hinge"]);
```

Fit the incremental model to the second half of the data by using the `updateMetricsAndfit` function. At each iteration:

- Simulate a data stream by processing 20 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

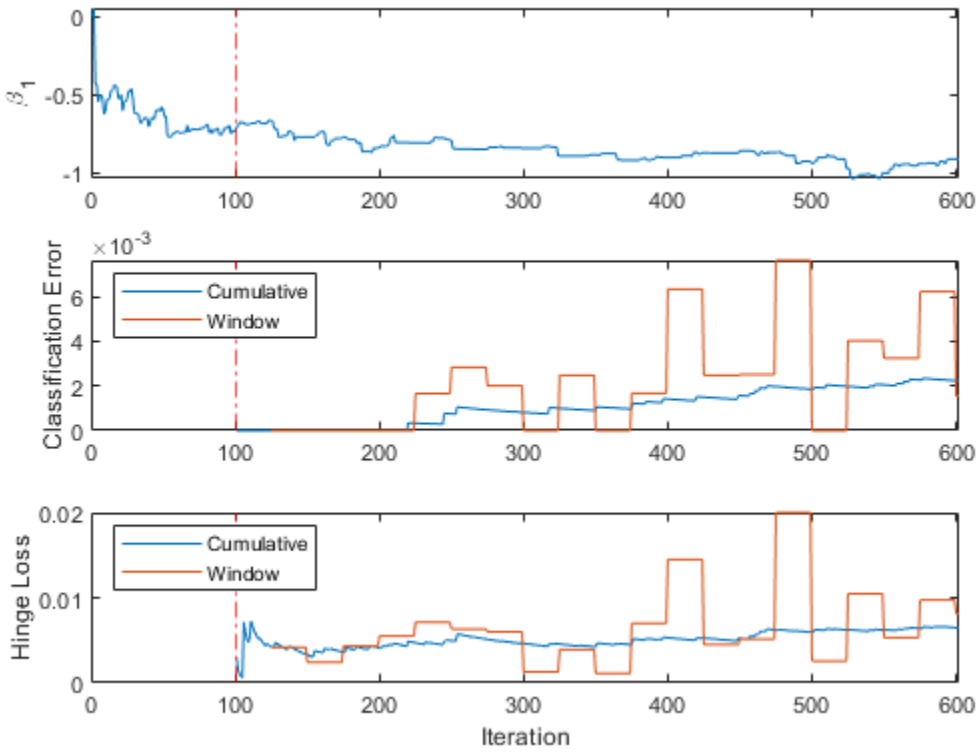
```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 20;
nchunk = ceil(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
hinge = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    hinge{j,:} = IncrementalMdl.Metrics{"HingeLoss",:};
```

```
    beta1(j + 1) = IncrementalMdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(3,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
subplot(3,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ce.Properties.VariableNames,'Location','northwest')
subplot(3,1,3)
h = plot(hinge.Variables);
xlim([0 nchunk]);
ylabel('Hinge Loss')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,hinge.Properties.VariableNames,'Location','northwest')
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (25 iterations).

Input Arguments

`Mdl` — Traditionally trained linear SVM model for binary classification

`ClassificationSVM` model object | `CompactClassificationSVM` model object

Traditionally trained linear SVM model for binary classification, specified as a model object returned by its training or processing function.

Model Object	Training or Processing Function
`ClassificationSVM`	`fitcsvm`
`CompactClassificationSVM`	`fitcsvm` or `compact`

Note

Incremental learning functions support only numeric input predictor data. If `Mdl` was fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','scale-invariant','MetricsWindowSize',100` specifies the adaptive scale-invariant solver for objective optimization, and specifies processing 100 observations before updating the performance metrics.

General Options

`Solver` — Objective function minimization technique
`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-2974 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

Example: `'Solver','sgd'`

Data Types: `char` | `string`

`EstimationPeriod` — Number of observations processed to estimate hyperparameters
nonnegative integer

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as the comma-separated pair consisting of `'EstimationPeriod'` and a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalLearner` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalLearner` sets `'EstimationPeriod'` to `1000`.

For more details, see "Estimation Period" on page 33-2975.

Example: `'EstimationPeriod',100`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`'auto'` (default) | `false` | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`'auto'`	`incrementalLearner` determines whether the predictor variables need to be standardized. See "Standardize Data" on page 33-2975.
`true`	The software standardizes the predictor data.
`false`	The software does not standardize the predictor data.

Under some conditions, `incrementalLearner` can override your specification. For more details, see "Standardize Data" on page 33-2975.

Example: `'Standardize',true`

Data Types: `logical` | `char` | `string`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size

`10` (default) | positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration during training, `incrementalLearner` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

Example: `'BatchSize',1`

Data Types: `single` | `double`

`Lambda` — Ridge (L^2) regularization term strength

`1e-5` (default) | nonnegative scalar

Ridge (L^2) regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

Example: `'Lambda',0.01`

Data Types: `single` | `double`

`LearnRate` — Learning rate

`'auto'` (default) | positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

For `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either function.

The name-value pair argument `'LearnRateSchedule'` determines the learning rate for subsequent learning cycles.

Example: `'LearnRate',0.001`

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule

`'decaying'` (default) | `'constant'`

Learning rate schedule, specified as the comma-separated pair consisting of `'LearnRateSchedule'` and a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.
`'decaying'`	The learning rate at learning cycle t is $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$ • λ is the value of `Lambda`. • If `Solver` is `'sgd'`, then $c = 1$. • If `Solver` is `'asgd'`, then c is 0.75 [4].

Example: `'LearnRateSchedule','constant'`

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch

`true` (default) | `false`

Flag for shuffling the observations in the batch at each iteration, specified as the comma-separated pair consisting of `'Shuffle'` and a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Example: `'Shuffle',false`

Data Types: `logical`

Performance Metrics Options

`Metrics` — Model performance metrics to track during incremental learning
`"classiferror"` (default) | string vector | function handle | cell vector | structure array | `"binodeviance"` | `"exponential"` | `"hinge"` | `"logit"` | `"quadratic"` | ...

Model performance metrics to track during incremental learning with the `updateMetrics` or `updateMetricsAndFit` function, specified as a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"binodeviance"`	Binomial deviance
`"classiferror"`	Classification error
`"exponential"`	Exponential
`"hinge"`	Hinge
`"logit"`	Logistic
`"quadratic"`	Quadratic

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["classiferror" "hinge"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(C,S)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data by processing the incremental learning functions during a learning cycle.
- You specify the function name (`customMetric`).
- `C` is an *n*-by-2 logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(`*p*`,`*q*`) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.

- `S` is an *n*-by-2 numeric matrix of predicted classification scores. `S` is similar to the `score` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in `Mdl.ClassNames`. `S(p,q)` is the classification score of observation *p* being classified in class *q*.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'logit' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `IncrementalMdl.Metrics`. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"classiferror"` is `"ClassificationError"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(C,S)customMetric(C,S)...` is `CustomMetric_1`

For more details on performance metrics options, see “Performance Metrics” on page 33-2976.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`0` (default) | nonnegative integer | ...

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as the comma-separated pair consisting of `'MetricsWarmupPeriod'` and a nonnegative integer. The incremental model is warm after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (`EstimationPeriod` + `MetricsWarmupPeriod` observations).

For more details on performance metrics options, see “Performance Metrics” on page 33-2976.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer | ...

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2976.

Data Types: `single` | `double`

Output Arguments

`IncrementalMdl` — Binary classification linear model for incremental learning
`incrementalClassificationLinear` model object

Binary classification linear model for incremental learning, returned as an `incrementalClassificationLinear` model object. `IncrementalMdl` is also configured to generate predictions given new data (see `predict`).

To initialize `IncrementalMdl` for incremental learning, `incrementalLearner` passes the values of the `Mdl` properties in this table to congruent properties of `IncrementalMdl`.

Property	Description
`Beta`	Linear model coefficients, a numeric vector
`Bias`	Model intercept, a numeric scalar
`ClassNames`	Class labels for binary classification, two-element list
`Mu`	Predictor variable means, a numeric vector
`Prior`	Prior class label distribution, a numeric vector
`Sigma`	Predictor variable standard deviations, a numeric vector

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for classification `fit` and `updateMetricsAndFit` use the more aggressive ScInOL2 version of the algorithm.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When both these conditions apply: • When you set `'Standardize',true` (see "Standardize Data" on page 33-2975) • `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty arrays `[]`.
Learning rate	`LearnRate`	Adjust solver step size	When both of these conditions apply: • You change the solver of `Mdl` to SGD or ASGD (see `Solver`). • You do not set the `'LearnRate'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `IncrementalMdl`.

- If you standardized the predictor data when you trained the input model `Mdl` by using `fitcsvm`, the following conditions apply:
  - `incrementalLearner` passes the means in `Mdl.Mu` and standard deviations in `Mdl.Sigma` to the congruent incremental learning model properties.

  - Incremental learning functions always standardize the predictor data, regardless of the value of the `'Standardize'` name-value pair argument.
- When you set `'Standardize',true`, and `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, the following conditions apply:
  - If the estimation period is positive (see the `EstimationPeriod` property of `IncrementalMdl`), incremental fitting functions estimate means and standard deviations using the estimation period observations.
  - If the estimation period is 0, `incrementalLearner` forces the estimation period to `1000`. Consequently, incremental fitting functions estimate new predictor variable means and standard deviations during the forced estimation period.
- When you set `'Standardize','auto'` (the default), the following conditions apply.
  - If `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, incremental learning functions do not standardize predictor variables.
  - Otherwise, incremental learning functions standardize the predictor variables using their means and standard deviations in `IncrementalMdl.Mu` and `IncrementalMdl.Sigma`, respectively. Incremental fitting functions do not estimate new means and standard deviations regardless of the length of the estimation period.
- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  where
  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k^*} \sum_k w_k^* x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k^*} \sum_k w_k^* \left(x_{jk} - \mu_j^*\right)^2$.
  - $w_j^* = \frac{w_j}{\sum_{\forall j \in \text{Class } k} w_j} p_k$, where
    - p_k is the prior probability of class k (`Prior` property of the incremental model).
    - w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions are incremental learning functions that track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to `'MetricsWarmupPeriod'` observations, which is the metrics warm-up period.

If `'EstimationPeriod'` > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.

- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:

  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the `'MetricsWindowSize'` name-value pair argument. `'MetricsWindowSize'` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `IncrementalMdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

See Also

Objects
`ClassificationSVM` | `CompactClassificationSVM` | `incrementalClassificationLinear`

Functions
`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2020b

incrementalLearner

Convert support vector machine (SVM) regression model to incremental learner

Syntax

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl = incrementalLearner(Mdl,Name,Value)
```

Description

`IncrementalMdl = incrementalLearner(Mdl)` returns a linear regression model for incremental learning on page 33-2992, `IncrementalMdl`, using the hyperparameters and coefficients of the traditionally trained linear SVM model for regression, `Mdl`. Because its property values reflect the knowledge gained from `Mdl`, `IncrementalMdl` can predict labels given new observations, and it is warm, meaning that its predictive performance is tracked.

`IncrementalMdl = incrementalLearner(Mdl,Name,Value)` uses additional options specified by one or more name-value pair arguments. Some options require you to train `IncrementalMdl` before its predictive performance is tracked. For example, `'MetricsWarmupPeriod',50,'MetricsWindowSize',100` specifies a preliminary incremental training period of 50 observations before performance metrics are tracked, and specifies processing 100 observations before updating the performance metrics.

Examples

Convert Traditionally Trained Model to Incremental Learner

Train an SVM regression model by using `fitrsvm`, and then convert it to an incremental learner.

Load and Preprocess Data

Load the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}];
```

Train SVM Regression Model

Fit an SVM regression model to 5000 randomly drawn observations from the data set. Discard the support vectors (`Alpha`) from the model so that the software uses linear coefficients (`Beta`) for prediction.

```
N = numel(Y);
n = 5000;
rng(1); % For reproducibility
idx = randsample(N,n);

TTMdl = fitrsvm(X(idx,:),Y(idx));
TTMdl = discardSupportVectors(TTMdl)
```

```
TTMdl =
  RegressionSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                     Beta: [312x1 double]
                     Bias: 64.5811
         KernelParameters: [1x1 struct]
          NumObservations: 5000
           BoxConstraints: [5000x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [5000x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

`TTMdl` is a `RegressionSVM` model object representing a traditionally trained SVM regression model.

Convert Trained Model

Convert the traditionally trained SVM regression model to a linear regression model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalRegressionLinear

               IsWarm: 1
              Metrics: [1x2 table]
    ResponseTransform: 'none'
                 Beta: [312x1 double]
                 Bias: 64.5811
              Learner: 'svm'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object prepared for incremental learning using SVM.

- The `incrementalLearner` function Initializes the incremental learner by passing learned coefficients to it, along with other information `TTMdl` extracted from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.
- The `incrementalLearner` function trains the model using the adaptive scale-invariant solver, whereas `fitrsvm` trained `TTMdl` using the SMO solver.

Predict Responses

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict sales prices for all observations using both models.

```
ttyfit = predict(TTMdl,X);
ilyfit = predict(IncrementalMdl,X);
compareyfit = norm(ttyfit - ilyfit)
```

```
compareyfit = 0
```

The difference between the fitted values generated by the models is 0.

Specify SGD Solver

The default solver is the adaptive scale-invariant solver. If you specify this solver, you do not need to tune any parameters for training. However, if you specify either the standard SGD or ASGD solver instead, you can also specify an estimation period, during which the incremental fitting functions tune the learning rate.

Load and shuffle the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015

rng(1) % For reproducibility
n = size(NYCHousing2015,1);
shuffidx = randsample(n,n);
NYCHousing2015 = NYCHousing2015(shuffidx,:);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}];
```

Randomly partition the data into 5% and 95% sets: the first set for training a model traditionally, and the second set for incremental learning.

```
cvp = cvpartition(n,'Holdout',0.95);
idxtt = training(cvp);
idxil = test(cvp);

% 5% set for traditional training
Xtt = X(idxtt,:);
Ytt = Y(idxtt);

% 95% set for incremental learning
Xil = X(idxil,:);
Yil = Y(idxil);
```

Fit an SVM regression model to 5% of the data.

```
TTMdl = fitrsvm(Xtt,Ytt);
```

Convert the traditionally trained SVM regression model to a linear regression model for incremental learning. Specify the standard SGD solver and an estimation period of `2e4` observations (the default is `1000` when a learning rate is required).

```
IncrementalMdl = incrementalLearner(TTMdl,'Solver','sgd','EstimationPeriod',2e4);
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object.

Fit the incremental model to the rest of the data by using the `fit` function. At each iteration:

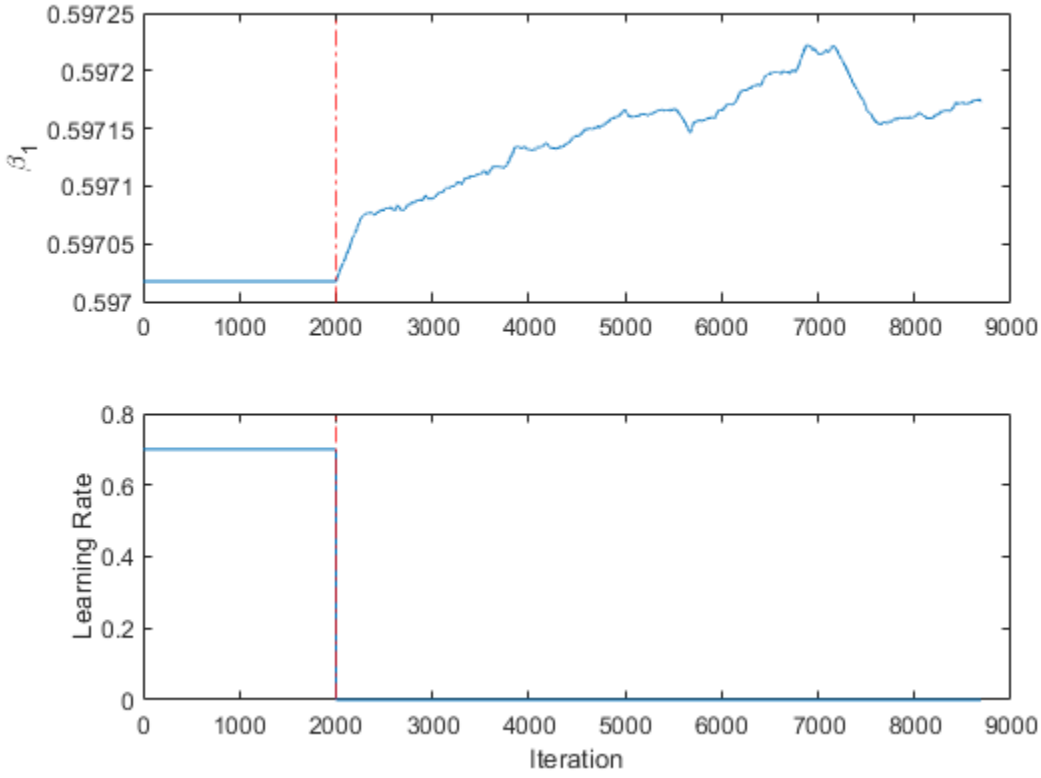
- Simulate a data stream by processing 10 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the learning rate and β_1 to see how the coefficients and learning rate evolve during training.

```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 10;
nchunk = floor(nil/numObsPerChunk);
learnrate = [IncrementalMdl.LearnRate; zeros(nchunk,1)];
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(1);
    learnrate(j + 1) = IncrementalMdl.LearnRate;
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

To see how the learning rate and β_1 evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(beta1)
hold on
ylabel('\beta_1')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
plot(learnrate)
ylabel('Learning Rate')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
```



The learning rate jumps to its auto-tuned value after the estimation period.

Because `fit` does not fit the model to the streaming data during the estimation period, β_1 is constant for the first 2000 iterations (20,000 observations). Then, β_1 changes slightly as `fit` fits the model to each new chunk of 10 observations.

Configure Performance Metric Options

Use a trained SVM regression model to initialize an incremental learner. Prepare the incremental learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the robot arm data set.

```
load robotarm
```

For details on the data set, enter `Description` at the command line.

Randomly partition the data into 5% and 95% sets: the first set for training a model traditionally, and the second set for incremental learning.

```
n = numel(ytrain);

rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.95);
idxtt = training(cvp);
idxil = test(cvp);

% 5% set for traditional training
Xtt = Xtrain(idxtt,:);
Ytt = ytrain(idxtt);

% 95% set for incremental learning
Xil = Xtrain(idxil,:);
Yil = ytrain(idxil);
```

Fit an SVM regression model to the first set.

```
TTMdl = fitrsvm(Xtt,Ytt);
```

Convert the traditionally trained SVM regression model to a linear regression model for incremental learning. Specify the following:

- A performance metrics warm-up period of 2000 observations.
- A metrics window size of 500 observations.
- Use of epsilon insensitive loss, MSE, and mean absolute error (MAE) to measure the performance of the model. The software supports epsilon insensitive loss and MSE. Create an anonymous function that measures the absolute error of each new observation. Create a structure array containing the name `MeanAbsoluteError` and its corresponding function.

```
maefcn = @(z,zfit)abs(z - zfit);
maemetric = struct("MeanAbsoluteError",maefcn);
IncrementalMdl = incrementalLearner(TTMdl,'MetricsWarmupPeriod',2000,'MetricsWindowSize',500,...
    'Metrics',{'epsiloninsensitive' 'mse' maemetric});
```

Fit the incremental model to the rest of the data by using the `updateMetricsAndfit` function. At each iteration:

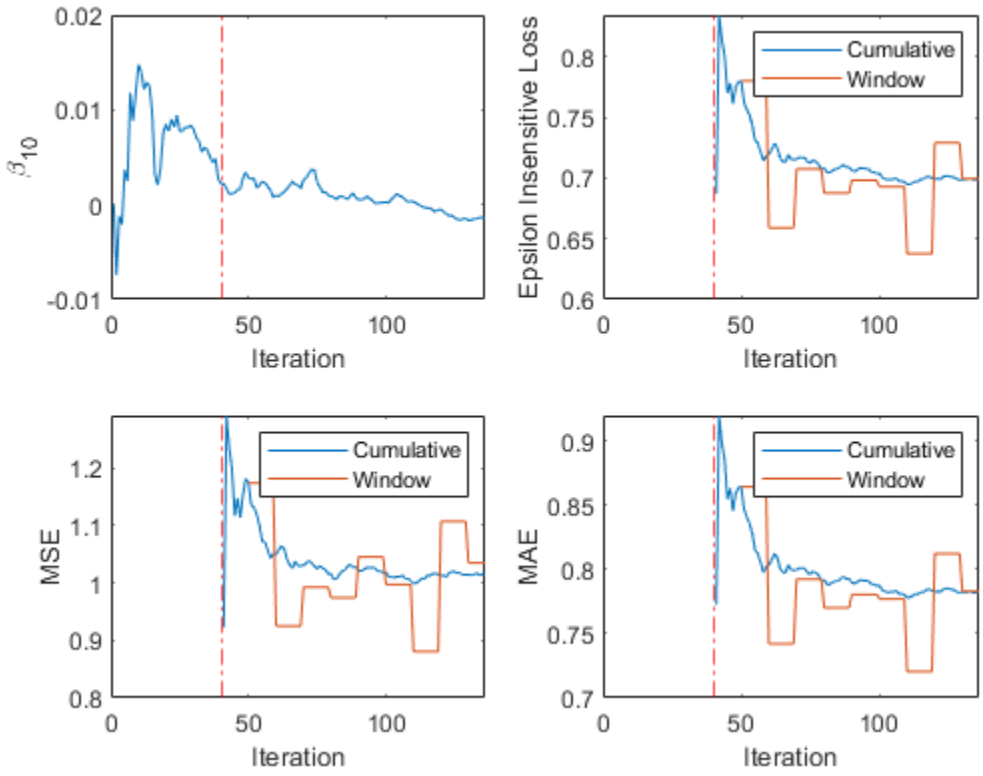
- Simulate a data stream by processing 50 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the estimated coefficient β_{10}, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mse = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mae = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    mse{j,:} = IncrementalMdl.Metrics{"MeanSquaredError",:};
    mae{j,:} = IncrementalMdl.Metrics{"MeanAbsoluteError",:};
    beta1(j + 1) = IncrementalMdl.Beta(10);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_{10} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(beta1)
ylabel('\beta_{10}')
xlim([0 nchunk]);
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
subplot(2,2,2)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ei.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,3)
h = plot(mse.Variables);
xlim([0 nchunk]);
ylabel('MSE')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mse.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,4)
h = plot(mae.Variables);
xlim([0 nchunk]);
ylabel('MAE')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mae.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_{10} during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Input Arguments

`Mdl` — Traditionally trained linear SVM model for regression

`RegressionSVM` model object | `CompactRegressionSVM` model object

Traditionally trained linear SVM model for regression, specified as a model object returned by its training or processing function.

Model Object	Training or Processing Function
`RegressionSVM`	`fitrsvm`
`CompactRegressionSVM`	`fitrsvm` or `compact`

Note

Incremental learning functions support only numeric input predictor data. If `Mdl` was fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','scale-invariant','MetricsWindowSize',100` specifies the adaptive scale-invariant solver for objective optimization, and specifies processing 100 observations before updating the performance metrics.

General Options

`Solver` — Objective function minimization technique
`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-2992 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

Example: `'Solver','sgd'`

Data Types: `char` | `string`

`EstimationPeriod` — Number of observations processed to estimate hyperparameters
nonnegative integer

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as the comma-separated pair consisting of `'EstimationPeriod'` and a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalLearner` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalLearner` sets `'EstimationPeriod'` to `1000`.

For more details, see “Estimation Period” on page 33-2993.

Example: `'EstimationPeriod',100`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`'auto'` (default) | `false` | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`'auto'`	`incrementalLearner` determines whether the predictor variables need to be standardized. See “Standardize Data” on page 33-2993.
`true`	The software standardizes the predictor data.
`false`	The software does not standardize the predictor data.

Under some conditions, `incrementalLearner` can override your specification. For more details, see “Standardize Data” on page 33-2993.

Example: `'Standardize',true`

Data Types: `logical` | `char` | `string`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size

`10` (default) | positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration during training, `incrementalLearner` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

Example: `'BatchSize',1`

Data Types: `single` | `double`

`Lambda` — Ridge (L^2) regularization term strength

`1e-5` (default) | nonnegative scalar

Ridge (L^2) regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

Example: `'Lambda',0.01`

Data Types: `single` | `double`

`LearnRate` — Learning rate
`'auto'` (default) | positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

For `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either function.

The name-value pair argument `'LearnRateSchedule'` determines the learning rate for subsequent learning cycles.

Example: `'LearnRate',0.001`

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule
`'decaying'` (default) | `'constant'`

Learning rate schedule, specified as the comma-separated pair consisting of `'LearnRateSchedule'` and a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.
`'decaying'`	The learning rate at learning cycle t is $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$ • λ is the value of `Lambda`. • If `Solver` is `'sgd'`, then $c = 1$. • If `Solver` is `'asgd'`, then c is 0.75 [4].

Example: `'LearnRateSchedule','constant'`

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch
`true` (default) | `false`

Flag for shuffling the observations in the batch at each iteration, specified as the comma-separated pair consisting of `'Shuffle'` and a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Example: `'Shuffle',false`

Data Types: `logical`

Performance Metrics Options

`Metrics` — Model performance metrics to track during incremental learning
`"epsiloninsensitive"` (default) | string vector | function handle | cell vector | structure array | `"mse"` | ...

Model performance metrics to track during incremental learning with `updateMetrics` and `updateMetricsAndFit`, specified as the comma-separated pair consisting of `'Metrics'` and a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
`"epsiloninsensitive"`	Epsilon insensitive loss
`"mse"`	Weighted mean squared error

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["epsiloninsensitive" "mse"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(Y,YFit)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You select the function name (`customMetric`).
- `Y` is a length *n* numeric vector of observed responses, where *n* is the sample size.
- `YFit` is a length *n* numeric vector of corresponding predicted responses.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'mse' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `IncrementalMdl.Metrics`. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"epsiloninsensitive"` is `"EpsilonInsensitiveLoss"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(Y,YFit)customMetric(Y,YFit)...` is `CustomMetric_1`

For more details on performance metrics options, see “Performance Metrics” on page 33-2994.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`0` (default) | nonnegative integer | ...

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as the comma-separated pair consisting of `'MetricsWarmupPeriod'` and a nonnegative integer. The incremental model is warm after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (`EstimationPeriod` + `MetricsWarmupPeriod` observations).

For more details on performance metrics options, see “Performance Metrics” on page 33-2994.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer | ...

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-2994.

Data Types: `single` | `double`

Output Arguments

`IncrementalMdl` — Linear regression model for incremental learning
`incrementalRegressionLinear` model object

Linear regression model for incremental learning, returned as an `incrementalRegressionLinear` model object. `IncrementalMdl` is also configured to generate predictions given new data (see `predict`).

To initialize `IncrementalMdl` for incremental learning, `incrementalLearner` passes the values of the `Mdl` properties in this table to congruent properties of `IncrementalMdl`.

Property	Description
`Beta`	Linear model coefficients, a numeric vector
`Bias`	Model intercept, a numeric scalar
`Epsilon`	Half the width of the epsilon insensitive band, a nonnegative scalar
`Mu`	Predictor variable means, a numeric vector
`Sigma`	Predictor variable standard deviations, a numeric vector

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for regression `fit` and `updateMetricsAndFit` use the more conservative ScInOL1 version of the algorithm.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When both these conditions apply: • When you set `'Standardize',true` (see "Standardize Data" on page 33-2993) • `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty arrays `[]`.
Learning rate	`LearnRate`	Adjust solver step size	When both of these conditions apply: • You change the solver of `Mdl` to SGD or ASGD (see `Solver`). • You do not set the `'LearnRate'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `IncrementalMdl`.

- If you standardized the predictor data when you trained the input model `Mdl` by using `fitrsvm`, the following conditions apply:
  - `incrementalLearner` passes the means in `Mdl.Mu` and standard deviations in `Mdl.Sigma` to the congruent incremental learning model properties.
  - Incremental learning functions always standardize the predictor data, regardless of the value of the `'Standardize'` name-value pair argument.
- When you set `'Standardize',true`, and `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, the following conditions apply:
  - If the estimation period is positive (see the `EstimationPeriod` property of `IncrementalMdl`), incremental fitting functions estimate means and standard deviations using the estimation period observations.

- If the estimation period is 0, `incrementalLearner` forces the estimation period to `1000`. Consequently, incremental fitting functions estimate new predictor variable means and standard deviations during the forced estimation period.

- If you set `'Standardize','auto'` (the default), the following conditions apply.
  - If `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, incremental learning functions do not standardize predictor variables.
  - Otherwise, incremental learning functions standardize the predictor variables using their means and standard deviations in `IncrementalMdl.Mu` and `IncrementalMdl.Sigma`, respectively. Incremental fitting functions do not estimate new means and standard deviations regardless of the length of the estimation period.
- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

$$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k}\sum_k w_k x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k}\sum_k w_k\left(x_{jk} - \mu_j^*\right)^2$.
  - w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions are incremental learning functions that track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to `'MetricsWarmupPeriod'` observations, which is the metrics warm-up period.

  If `'EstimationPeriod'` > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:
  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the `'MetricsWindowSize'` name-value pair argument. `'MetricsWindowSize'` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on

the last 20 observations in the supplied data (`X((end – 20 + 1):end,:)` and `Y((end – 20 + 1):end)`).

Incremental functions that track performance metrics within a window use the following process:

1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
3. When the buffer is filled, overwrite `IncrementalMdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

See Also

Objects
`CompactRegressionSVM` | `RegressionSVM` | `incrementalRegressionLinear`

Functions
`fit` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Linear Regression Using Flexible Workflow” on page 26-22

Introduced in R2020b

incrementalLearner

Convert linear regression model to incremental learner

Syntax

```
IncrementalMdl = incrementalLearner(Mdl)
IncrementalMdl = incrementalLearner(Mdl,Name,Value)
```

Description

`IncrementalMdl = incrementalLearner(Mdl)` returns a linear regression model for incremental learning on page 33-3009, `IncrementalMdl`, using the hyperparameters and coefficients of the traditionally trained linear regression model `Mdl`. Because its property values reflect the knowledge gained from `Mdl`, `IncrementalMdl` can predict labels given new observations, and it is warm, meaning that its predictive performance is tracked.

`IncrementalMdl = incrementalLearner(Mdl,Name,Value)` uses additional options specified by one or more name-value pair arguments. Some options require you to train `IncrementalMdl` before its predictive performance is tracked. For example, `'MetricsWarmupPeriod',50,'MetricsWindowSize',100` specifies a preliminary incremental training period of 50 observations before performance metrics are tracked, and specifies processing 100 observations before updating the performance metrics.

Examples

Convert Traditionally Trained Model to Incremental Learner

Train a linear regression model by using `fitrlinear`, and then convert it to an incremental learner.

Load and Preprocess Data

Load the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}];
```

Train Linear Regression Model

Fit an linear regression model to the entire data set.

```
TTMdl = fitrlinear(X,Y)
```

```
TTMdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [312x1 double]
                 Bias: 0.0956
               Lambda: 1.0935e-05
              Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `RegressionLinear` model object representing a traditionally trained linear regression model.

Convert Trained Model

Convert the traditionally trained linear regression model to a linear regression model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalRegressionLinear

               IsWarm: 1
              Metrics: [1x2 table]
    ResponseTransform: 'none'
                 Beta: [312x1 double]
                 Bias: 0.0956
              Learner: 'svm'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object prepared for incremental learning using SVM.

- The `incrementalLearner` function Initializes the incremental learner by passing learned coefficients to it, along with other information `TTMdl` extracted from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.
- The `incrementalLearner` function trains the model using the adaptive scale-invariant solver, whereas `fitrlinear` trained `TTMdl` using the dual SGD solver.

Predict Responses

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict sales prices for all observations using both models.

```
ttyfit = predict(TTMdl,X);
ilyfit = predict(IncrementalMdl,X);
compareyfit = norm(ttyfit - ilyfit)
```

```
compareyfit = 0
```

The difference between the fitted values generated by the models is 0.

Specify SGD Solver

The default solver is the adaptive scale-invariant solver. If you specify this solver, you do not need to tune any parameters for training. However, if you specify either the standard SGD or ASGD solver instead, you can also specify an estimation period, during which the incremental fitting functions tune the learning rate.

Load and shuffle the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015

rng(1) % For reproducibility
n = size(NYCHousing2015,1);
shuffidx = randsample(n,n);
NYCHousing2015 = NYCHousing2015(shuffidx,:);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}];
```

Randomly partition the data into 5% and 95% sets: the first set for training a model traditionally, and the second set for incremental learning.

```
cvp = cvpartition(n,'Holdout',0.95);
idxtt = training(cvp);
```

```
idxil = test(cvp);

% 5% set for traditional training
Xtt = X(idxtt,:);
Ytt = Y(idxtt);

% 95% set for incremental learning
Xil = X(idxil,:);
Yil = Y(idxil);
```

Fit a linear regression model to 5% of the data.

```
TTMdl = fitrlinear(Xtt,Ytt);
```

Convert the traditionally trained linear regression model to a linear regression model for incremental learning. Specify the standard SGD solver and an estimation period of `2e4` observations (the default is `1000` when a learning rate is required).

```
IncrementalMdl = incrementalLearner(TTMdl,'Solver','sgd','EstimationPeriod',2e4);
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object.

Fit the incremental model to the rest of the data by using the `fit` function. At each iteration:

- Simulate a data stream by processing 10 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the learning rate and β_1 to see how the coefficients and learning rate evolve during training.

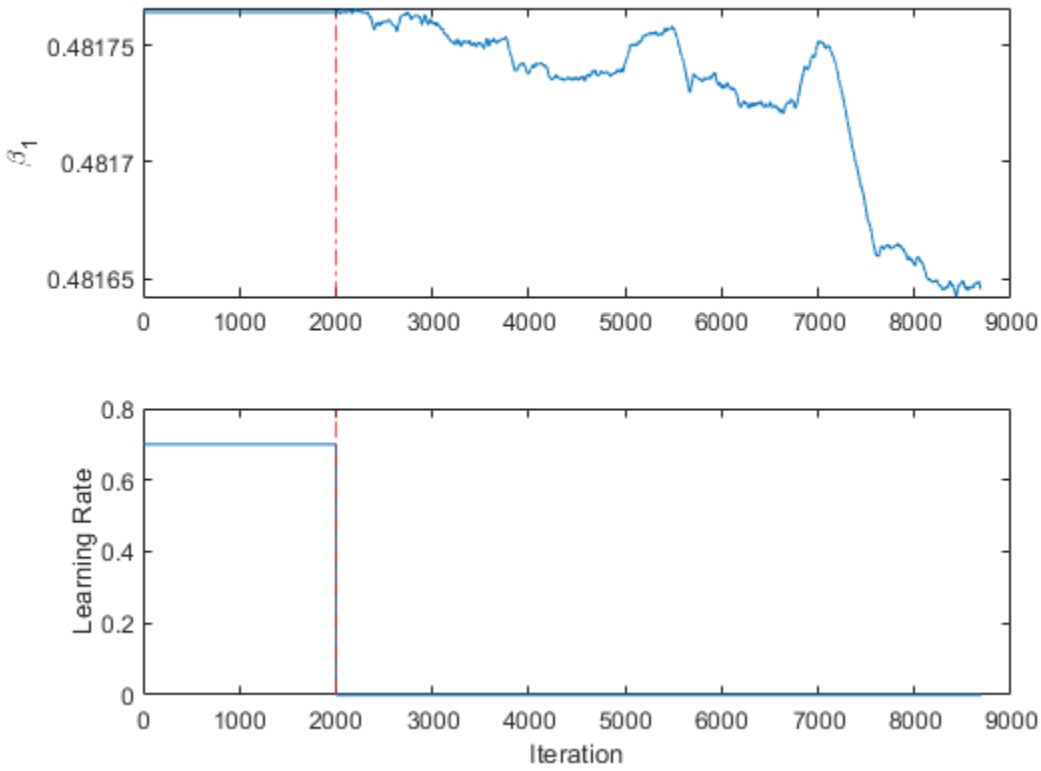
```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 10;
nchunk = floor(nil/numObsPerChunk);
learnrate = [IncrementalMdl.LearnRate; zeros(nchunk,1)];
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = fit(IncrementalMdl,Xil(idx,:),Yil(idx));
    beta1(j + 1) = IncrementalMdl.Beta(1);
    learnrate(j + 1) = IncrementalMdl.LearnRate;
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

To see how the learning rate and β_1 evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(beta1)
hold on
ylabel('\beta_1')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
plot(learnrate)
```

```
ylabel('Learning Rate')
xline(IncrementalMdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
```



The learning rate jumps to its auto-tuned value after the estimation period.

Because `fit` does not fit the model to the streaming data during the estimation period, β_1 is constant for the first 2000 iterations (20,000 observations). Then, β_1 changes slightly as `fit` fits the model to each new chunk of 10 observations.

Configure Performance Metric Options

Use a trained linear regression model to initialize an incremental learner. Prepare the incremental learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations.

Load the robot arm data set.

```
load robotarm
```

For details on the data set, enter `Description` at the command line.

Randomly partition the data into 5% and 95% sets: the first set for training a model traditionally, and the second set for incremental learning.

```
n = numel(ytrain);

rng(1) % For reproducibility
cvp = cvpartition(n,'Holdout',0.95);
idxtt = training(cvp);
idxil = test(cvp);

% 5% set for traditional training
Xtt = Xtrain(idxtt,:);
Ytt = ytrain(idxtt);

% 95% set for incremental learning
Xil = Xtrain(idxil,:);
Yil = ytrain(idxil);
```

Fit a linear regression model to the first set.

```
TTMdl = fitrlinear(Xtt,Ytt);
```

Convert the traditionally trained linear regression model to a linear regression model for incremental learning. Specify all of the following:

- A performance metrics warm-up period of 2000 observations.
- A metrics window size of 500 observations.
- Use of epsilon insensitive loss, MSE, and mean absolute error (MAE) to measure the performance of the model. The software supports epsilon insensitive loss and MSE. Create an anonymous function that measures the absolute error of each new observation. Create a structure array containing the name `MeanAbsoluteError` and its corresponding function.

```
maefcn = @(z,zfit)abs(z - zfit);
maemetric = struct("MeanAbsoluteError",maefcn);
IncrementalMdl = incrementalLearner(TTMdl,'MetricsWarmupPeriod',2000,'MetricsWindowSize',500,...
    'Metrics',{'epsiloninsensitive' 'mse' maemetric});
```

Fit the incremental model to the rest of the data by using the `updateMetricsAndFit` function. At each iteration:

- Simulate a data stream by processing 50 observations at a time.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

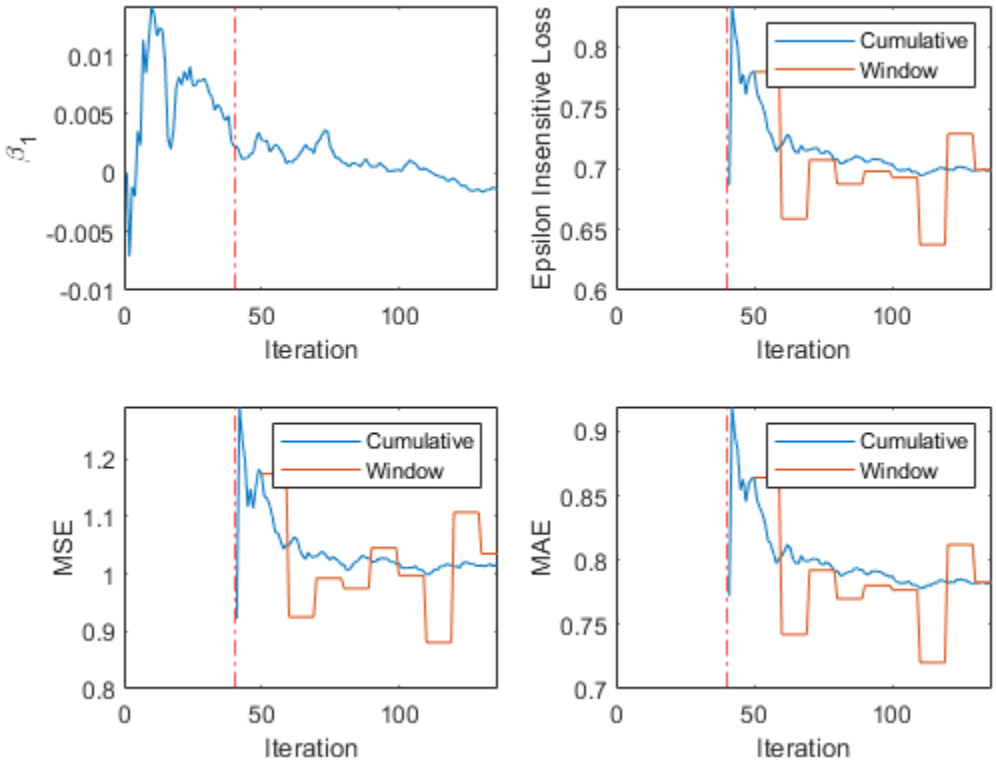
```
% Preallocation
nil = numel(Yil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mse = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mae = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
```

```
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx));
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    mse{j,:} = IncrementalMdl.Metrics{"MeanSquaredError",:};
    mae{j,:} = IncrementalMdl.Metrics{"MeanAbsoluteError",:};
    beta1(j + 1) = IncrementalMdl.Beta(10);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
subplot(2,2,2)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ei.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,3)
h = plot(mse.Variables);
xlim([0 nchunk]);
ylabel('MSE')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mse.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,4)
h = plot(mae.Variables);
xlim([0 nchunk]);
ylabel('MAE')
xline(IncrementalMdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mae.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Input Arguments

`Mdl` — Traditionally trained linear regression model

`RegressionLinear` model object

Traditionally trained linear regression model, specified as a `RegressionLinear` model object returned by `fitrlinear`.

Note

- If `Mdl.Lambda` is a numeric vector, you must select the model corresponding to one regularization strength in the regularization path by using `selectModels`.
- Incremental learning functions support only numeric input predictor data. If `Mdl` was fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of

dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Solver','scale-invariant','MetricsWindowSize',100` specifies the adaptive scale-invariant solver for objective optimization, and specifies processing 100 observations before updating the performance metrics.

General Options

`Solver` — Objective function minimization technique
`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-3010 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in “SGD and ASGD Solver Options” on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

Example: `'Solver','sgd'`

Data Types: `char` | `string`

`EstimationPeriod` — Number of observations processed to estimate hyperparameters
nonnegative integer

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as the comma-separated pair consisting of `'EstimationPeriod'` and a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalLearner` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalLearner` sets `'EstimationPeriod'` to `1000`.

For more details, see “Estimation Period” on page 33-3010.

Example: `'EstimationPeriod',100`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`false` (default) | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`true`	The software standardizes the predictor data. For more details, see “Standardize Data” on page 33-3011.
`false`	The software does not standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size

positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration during training, `incrementalLearner` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'BatchSize'`. Instead, `incrementalLearner` sets `'BatchSize'` to `Mdl.ModelParameters.BatchSize`.
- Otherwise, `BatchSize` is `10`.

Example: `'BatchSize',1`

Data Types: `single` | `double`

`Lambda` — Ridge (*L*2) regularization term strength

nonnegative scalar

Ridge (*L*2) regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and a nonnegative scalar.

When `Mdl.Regularization` is `'ridge (L2)'`:

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'Lambda'`. Instead, `incrementalLearner` sets `'Lambda'` to `Mdl.Lambda`.
- Otherwise, `Lambda` is `1e-5`.

Note `incrementalLearner` does not support lasso regularization. If `Mdl.Regularization` is `'lasso (L1)'`, `incrementalLearner` uses ridge regularization instead, and sets the `'Solver'` name-value pair argument to `'scale-invariant'` by default.

Example: `'Lambda',0.01`

Data Types: `single` | `double`

`LearnRate` — Learning rate
`'auto'` | positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

- If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'LearnRate'`. Instead, `incrementalLearner` sets `'LearnRate'` to `Mdl.ModelParameters.LearnRate`.
- Otherwise, `LearnRate` is `'auto'`.

For `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either function.

The name-value pair argument `'LearnRateSchedule'` determines the learning rate for subsequent learning cycles.

Example: `'LearnRate',0.001`

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule
`'decaying'` (default) | `'constant'`

Learning rate schedule, specified as the comma-separated pair consisting of `'LearnRateSchedule'` and a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.

Value	Description
`'decaying'`	The learning rate at learning cycle t is $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$ • λ is the value of `Lambda`. • If `Solver` is `'sgd'`, then $c = 1$. • If `Solver` is `'asgd'`, then c is 0.75 [4].

If `Mdl.ModelParameters.Solver` is `'sgd'` or `'asgd'`, you cannot set `'LearnRateSchedule'`.

Example: `'LearnRateSchedule','constant'`

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch

`true` (default) | `false`

Flag for shuffling the observations in the batch at each iteration, specified as the comma-separated pair consisting of `'Shuffle'` and a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Example: `'Shuffle',false`

Data Types: `logical`

Performance Metrics Options

`Metrics` — Model performance metrics to track during incremental learning

`"epsiloninsensitive"` | `"mse"` | string vector | function handle | cell vector | structure array | ...

Model performance metrics to track during incremental learning with `updateMetrics` and `updateMetricsAndFit`, specified as the comma-separated pair consisting of `'Metrics'` and a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

The following table lists the built-in loss function names and which learners, specified in `Mdl.Learner`, support them. You can specify more than one loss function by using a string vector.

Name	Description	Learners Supporting Metric
`"epsiloninsensitive"`	Epsilon insensitive loss	`'svm'`
`"mse"`	Weighted mean squared error	`'svm'` and `'leastsquares'`

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["epsiloninsensitive" "mse"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(Y,YFit)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You specify the function name (`customMetric`).
- `Y` is a length *n* numeric vector of observed responses, where *n* is the sample size.
- `YFit` is a length *n* numeric vector of corresponding predicted responses.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'mse' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `IncrementalMdl.Metrics`. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"epsiloninsensitive"` is `"EpsilonInsensitiveLoss"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(Y,YFit)customMetric(Y,YFit)...` is `CustomMetric_1`

By default:

- `Metrics` is `"epsiloninsensitive"` if `Mdl.Learner` is `'svm'`.
- `Metrics` is `"mse"` if `Mdl.Learner` is `'leastsquares'`.

For more details on performance metrics options, see “Performance Metrics” on page 33-3011.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`0` (default) | nonnegative integer | ...

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as the comma-separated pair consisting of `'MetricsWarmupPeriod'` and a nonnegative integer. The incremental model is warm after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (`EstimationPeriod` + `MetricsWarmupPeriod` observations).

For more details on performance metrics options, see “Performance Metrics” on page 33-3011.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer | ...

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-3011.

Data Types: `single` | `double`

Output Arguments

`IncrementalMdl` — Linear regression model for incremental learning
`incrementalRegressionLinear` model object

Linear regression model for incremental learning, returned as an `incrementalRegressionLinear` model object. `IncrementalMdl` is also configured to generate predictions given new data (see `predict`).

To initialize `IncrementalMdl` for incremental learning, `incrementalLearner` passes the values of the `Mdl` properties in this table to congruent properties of `IncrementalMdl`.

Property	Description
`Beta`	Linear model coefficients, a numeric vector
`Bias`	Model intercept, a numeric scalar
`Epsilon`	Half the width of the epsilon insensitive band, a nonnegative scalar
`Learner`	Linear regression model type
`ModelParameters.FitBias`	Linear model intercept inclusion flag
`Mu`	Predictor variable means, a numeric vector
`Sigma`	Predictor variable standard deviations, a numeric vector

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional

machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for regression `fit` and `updateMetricsAndFit` use the more conservative ScInOL1 version of the algorithm.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When you set `'Standardize',true` (see “Standardize Data” on page 33-3011)
Learning rate	`LearnRate`	Adjust solver step size	When both of these conditions apply: • You change the solver of `Mdl` to SGD or ASGD (see `Solver`). • You do not set the `'LearnRate'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `IncrementalMdl`.

- When you set `'Standardize',true`, and `IncrementalMdl.Mu` and `IncrementalMdl.Sigma` are empty, the following conditions apply:
  - If the estimation period is positive (see the `EstimationPeriod` property of `IncrementalMdl`), incremental fitting functions estimate means and standard deviations using the estimation period observations.
  - If the estimation period is 0, `incrementalLearner` forces the estimation period to `1000`. Consequently, incremental fitting functions estimate new predictor variable means and standard deviations during the forced estimation period.
- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  where
  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k}\sum_k w_k x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k}\sum_k w_k\left(x_{jk} - \mu_j^*\right)^2$.
  - w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions are incremental learning functions that track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (`IsWarm` property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to `'MetricsWarmupPeriod'` observations, which is the metrics warm-up period.

  If `'EstimationPeriod'` > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:

- `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
- `Window` — The functions compute metrics based on all observations within a window determined by the `'MetricsWindowSize'` name-value pair argument. `'MetricsWindowSize'` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

  Incremental functions that track performance metrics within a window use the following process:

  1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
  2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
  3. When the buffer is filled, overwrite `IncrementalMdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

See Also

Objects
`RegressionLinear` | `incrementalRegressionLinear`

Functions
`fit` | `updateMetrics`

Topics

Introduced in R2020b

incrementalRegressionLinear

Linear regression model for incremental learning

Description

`incrementalRegressionLinear` creates an `incrementalRegressionLinear` model object, which represents an incremental linear model for regression problems. Supported learners include support vector machine (SVM) and least squares.

Unlike other Statistics and Machine Learning Toolbox model objects, `incrementalRegressionLinear` can be called directly. Also, you can specify learning options, such as performance metrics configurations, parameter values, and the objective solver, before fitting the model to data. After you create an `incrementalRegressionLinear` object, it is prepared for incremental learning on page 33-3034.

`incrementalRegressionLinear` is best suited for incremental learning. For a traditional approach to training an SVM or linear regression model (such as creating a model by fitting it to data, performing cross-validation, tuning hyperparameters, and so on), see `fitrsvm` or `fitrlinear`.

Creation

You can create an `incrementalRegressionLinear` model object in several ways:

- **Call the function directly** — Configure incremental learning options, or specify initial values for linear model parameters and hyperparameters, by calling `incrementalRegressionLinear` directly. This approach is best when you do not have data yet or you want to start incremental learning immediately.
- **Convert a traditionally trained model** — To initialize an linear regression model for incremental learning using the model coefficients and hyperparameters of a trained SVM or linear regression model object, you can convert the traditionally trained model to an `incrementalRegressionLinear` model object by passing it to the `incrementalLearner` function. This table contains links to the appropriate reference pages.

Convertible Model Object	Conversion Function
`RegressionSVM` or `CompactRegressionSVM`	`incrementalLearner`
`RegressionLinear`	`incrementalLearner`

- **Call an incremental learning function** — `fit`, `updateMetrics`, and `updateMetricsAndFit` accept a configured `incrementalRegressionLinear` model object and data as input, and return an `incrementalRegressionLinear` model object updated with information learned from the input model and data.

Syntax

```
Mdl = incrementalRegressionLinear()
Mdl = incrementalRegressionLinear(Name,Value)
```

Description

`Mdl = incrementalRegressionLinear()` returns a default incremental linear model object for regression `Mdl`. Properties of a default model contain placeholders for unknown model parameters. You must train a default model before you can track its performance or generate predictions from it.

`Mdl = incrementalRegressionLinear(Name,Value)` sets properties on page 33-3017 and additional options using name-value pair arguments. Enclose each name in quotes. For example, `incrementalRegressionLinear('Beta',[0.1 0.3],'Bias',1,'MetricsWarmupPeriod',100)` sets the vector of linear model coefficients β to `[0.1 0.3]`, the bias β_0 to `1`, and the metrics warm-up period to `100`.

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Standardize',true` standardizes the predictor data using the predictor means and standard deviations estimated during the estimation period.

`Metrics` — Model performance metrics to track during incremental learning
`"epsiloninsensitive"` | `"mse"` | string vector | function handle | cell vector | structure array | ...

Model performance metrics to track during incremental learning, specified as the comma-separated pair consisting of `'Metrics'` and a built-in loss function name, string vector of names, function handle (`@metricName`), structure array of function handles, or cell vector of names, function handles, or structure arrays.

When `Mdl` is warm (see IsWarm), `updateMetrics` and `updateMetricsAndFit` track performance metrics in the Metrics property of `Mdl`.

The following table lists the built-in loss-function names and which learners, specified in `Mdl.Learner`, support them. You can specify more than one loss function by using a string vector.

Name	Description	Learners Supporting Metric
`"epsiloninsensitive"`	Epsilon insensitive loss	`'svm'`
`"mse"`	Weighted mean squared error	`'svm'` and `'leastsquares'`

For more details on the built-in loss functions, see `loss`.

Example: `'Metrics',["epsiloninsensitive" "mse"]`

To specify a custom function that returns a performance metric, use function handle notation. The function must have this form:

```
metric = customMetric(Y,YFit)
```

- The output argument `metric` is an *n*-by-1 numeric vector, where each element is the loss of the corresponding observation in the data processed by the incremental learning functions during a learning cycle.
- You select the function name (`customMetric`).

- `Y` is a length *n* numeric vector of observed responses, where *n* is the sample size.
- `YFit` is a length *n* numeric vector of corresponding predicted responses.

To specify multiple custom metrics and assign a custom name to each, use a structure array. To specify a combination of built-in and custom metrics, use a cell vector.

Example: `'Metrics',struct('Metric1',@customMetric1,'Metric2',@customMetric2)`

Example: `'Metrics',{@customMetric1 @customeMetric2 'mse' struct('Metric3',@customMetric3)}`

`updateMetrics` and `updateMetricsAndFit` store specified metrics in a table in the property `Metrics`. The data type of `Metrics` determines the row names of the table.

`'Metrics'` Value Data Type	Description of `Metrics` Property Row Name	Example
String or character vector	Name of corresponding built-in metric	Row name for `"epsiloninsensitive"` is `"EpsilonInsensitiveLoss"`
Structure array	Field name	Row name for `struct('Metric1',@customMetric1)` is `"Metric1"`
Function handle to function stored in a program file	Name of function	Row name for `@customMetric` is `"customMetric"`
Anonymous function	`CustomMetric_`*j*, where *j* is metric *j* in `Metrics`	Row name for `@(Y,YFit)customMetric(Y,YFit)...` is `CustomMetric_1`

By default:

- `Metrics` is `"epsiloninsensitive"` if `Mdl.Learner` is `'svm'`.
- `Metrics` is `"mse"` if `Mdl.Learner` is `'leastsquares'`.

For more details on performance metrics options, see “Performance Metrics” on page 33-3037.

Data Types: `char` | `string` | `struct` | `cell` | `function_handle`

`Standardize` — Flag to standardize predictor data
`'auto'` (default) | `false` | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and a value in this table.

Value	Description
`'auto'`	`incrementalRegressionLinear` determines whether the predictor variables need to be standardized. See “Standardize Data” on page 33-3036.

Value	Description
`true`	The software standardizes the predictor data. For more details, see “Standardize Data” on page 33-3036.
`false`	The software does not standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical` | `char` | `string`

Properties

You can set most properties by using name-value pair argument syntax only when you call `incrementalRegressionLinear`. You can set some properties when you call `incrementalLearner` to convert a traditionally trained model. You cannot set the properties `FittedLoss`, `NumTrainingObservations`, `Mu`, `Sigma`, `SolverOptions` , and `IsWarm`.

Regression Model Parameters

`Beta` — Linear model coefficients β

numeric vector

This property is read-only.

Linear model coefficients β, specified as a `NumPredictors`-by-1 numeric vector.

If you convert a traditionally trained model to create `Mdl`, `Beta` is specified by the value of the `Beta` property of the traditionally trained model. Otherwise, by default, `Beta` is `zeros(NumPredictors,1)`.

Data Types: `single` | `double`

`Bias` — Model intercept β_0

numeric scalar

This property is read-only.

Model intercept β_0, or bias term, specified as a numeric scalar.

If you convert a traditionally trained model to create `Mdl`, `Bias` is specified by the value of the `Bias` property of the traditionally trained model. Otherwise, by default, `Bias` is `0`.

Data Types: `single` | `double`

`Epsilon` — Half of the width of epsilon insensitive band

`'auto'` | nonnegative scalar

This property is read-only.

Half of the width of the epsilon insensitive-band, specified as `'auto'` or a nonnegative scalar.

If you specify `'auto'` when you call `incrementalRegressionLinear`, incremental fitting functions estimate `Epsilon` during the estimation period, specified by EstimationPeriod, using this procedure:

- If `iqr(Y)` ≠ 0, `Epsilon` is `iqr(Y)/13.49`, where `Y` is the estimation period response data.
- If `iqr(Y)` = 0 or before you fit `Mdl` to data, `Epsilon` is `0.1`.

If you convert a traditionally trained SVM regression model to create `Mdl` (`Learner` is `'svm'`), `Epsilon` is specified by the value of the `Epsilon` property of the traditionally trained model.

If `Learner` is `'leastsquares'`, you cannot set `Epsilon` and its value is `NaN`.

Data Types: `single` | `double`

`FittedLoss` — Loss function used to fit linear model
`'epsiloninsensitive'` | `'mse'`

This property is read-only.

Loss function used to fit the linear model, specified as `'epsiloninsensitive'` or `'mse'`.

Value	Algorithm	Loss function	`Learner` Value
`'epsiloninsensitive'`	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	`'svm'`
`'mse'`	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	`'leastsquares'`

`Learner` — Linear regression model type
`'leastsquares'` | `'svm'`

This property is read-only.

Linear regression model type, specified as `'leastsquares'` or `'svm'`.

In the following table, $f(x) = x\beta + b$.

- β is `Beta`.
- x is an observation from p predictor variables.
- β_0 is `Bias`.

Value	Algorithm	Loss function	`FittedLoss` Value
`'leastsquares'`	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	`'mse'`
`'svm'`	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	`'epsiloninsensitive'`

If you convert a traditionally trained model to create `Mdl`, `Learner` is the learner of the traditionally trained model.

- If the traditionally trained model is `CompactRegressionSVM` or `RegressionSVM`, `Learner` is `'svm'`.

- If the traditionally trained model is `RegressionLinear`, `Learner` is the value of the `Learner` property of the traditionally trained model.

`NumPredictors` — Number of predictor variables
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of predictor variables, specified as a nonnegative numeric scalar.

If you convert a traditionally trained model to create `Mdl`, `NumPredictors` is specified by the congruent property of the traditionally trained model. Otherwise, incremental fitting functions infer `NumPredictors` from the predictor data during training.

Data Types: `double`

`NumTrainingObservations` — Number of observations fit to incremental model
`0` (default) | nonnegative numeric scalar

This property is read-only.

Number of observations fit to the incremental model `Mdl`, specified as a nonnegative numeric scalar. `NumTrainingObservations` increases when you pass `Mdl` and training data to `fit` or `updateMetricsAndFit`.

Note If you convert a traditionally trained model to create `Mdl`, `incrementalRegressionLinear` does not add the number of observations fit to the traditionally trained model to `NumTrainingObservations`.

Data Types: `double`

`ResponseTransform` — Response transformation function
`'none'` | function handle

This property is read-only.

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how incremental learning functions transform raw response values.

For a MATLAB function or a function that you define, enter its function handle; for example, `'ResponseTransform',@`*`function`*, where *`function`* accepts an *n*-by-1 vector (the original responses) and returns a vector of the same length (the transformed responses).

- If you convert a traditionally trained model to create `Mdl`, `ResponseTransform` is specified by the congruent property of the traditionally trained model.
- Otherwise, `ResponseTransform` is `'none'`.

Data Types: `char` | `function_handle`

Training Parameters

`EstimationPeriod` — Number of observations processed to estimate hyperparameters
nonnegative integer

This property is read-only.

Number of observations processed by the incremental model to estimate hyperparameters before training or tracking performance metrics, specified as a nonnegative integer.

Note

- If `Mdl` is prepared for incremental learning (all hyperparameters required for training are specified), `incrementalRegressionLinear` forces `'EstimationPeriod'` to `0`.
- If `Mdl` is not prepared for incremental learning, `incrementalRegressionLinear` sets `'EstimationPeriod'` to `1000`.

For more details, see “Estimation Period” on page 33-3035.

Data Types: `single` | `double`

FitBias — Linear model intercept inclusion flag

`true` | `false`

This property is read-only.

Linear model intercept inclusion flag, specified as `true` or `false`.

Value	Description
`true`	`incrementalRegressionLinear` includes the bias term β_0 in the linear model, which incremental fitting functions fit to data.
`false`	`incrementalRegressionLinear` sets $\beta_0 = 0$.

If `Bias` $\neq 0$, `FitBias` must be `true`. In other words, `incrementalRegressionLinear` does not support an equality constraint on β_0.

If you convert a traditionally trained linear regression model (`RegressionLinear`) to create `Mdl`, `FitBias` is specified by the value of the `ModelParameters.FitBias` property of the traditionally trained model.

Data Types: `logical`

Mu — Predictor means

vector of numeric values | `[]`

This property is read-only.

Predictor means, specified as a numeric vector.

If `Mu` is an empty array `[]` and you specify `'Standardize',true`, incremental fitting functions set `Mu` to the predictor variable means estimated during the estimation period specified by `EstimationPeriod`.

You cannot specify `Mu` directly.

Data Types: `single` | `double`

`Sigma` — Predictor standard deviations

vector of numeric values | `[]`

This property is read-only.

Predictor standard deviations, specified as a numeric vector.

If `Sigma` is an empty array `[]` and you specify `'Standardize',true`, incremental fitting functions set `Sigma` to the predictor variable standard deviations estimated during the estimation period specified by `EstimationPeriod`.

You cannot specify `Sigma` directly.

Data Types: `single` | `double`

`Shuffle` — Flag for shuffling observations in batch

`true` (default) | `false`

This property is read-only.

Flag for shuffling the observations in the batch at each learning cycle, specified as a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Data Types: `logical`

`Solver` — Objective function minimization technique

`'scale-invariant'` (default) | `'sgd'` | `'asgd'`

This property is read-only.

Objective function minimization technique, specified as a value in this table.

Value	Description	Notes
`'scale-invariant'`	Adaptive scale-invariant solver for incremental learning on page 33-3035 [1]	• This algorithm is parameter free and can adapt to differences in predictor scales. Try this algorithm before using SGD or ASGD. • To shuffle incoming batches before the `fit` function fits the model, set `Shuffle` to `true`.

Value	Description	Notes
`'sgd'`	Stochastic gradient descent (SGD) [3][2]	• To train effectively with SGD, standardize the data and specify adequate values for hyperparameters using options listed in "SGD and ASGD Solver Parameters" on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.
`'asgd'`	Average stochastic gradient descent (ASGD) [4]	• To train effectively with ASGD, standardize the data and specify adequate values for hyperparameters using options listed in "SGD and ASGD Solver Parameters" on page 33-0 . • The `fit` function always shuffles an incoming batch of data before fitting the model.

If you convert a traditionally trained linear regression model (`RegressionLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `Solver` is specified by the `ModelParameters.Solver` property of the traditionally trained model.

Data Types: `char` | `string`

`SolverOptions` — Objective solver configurations
structure array

This property is read-only.

Objective solver configurations, specified as a structure array. The fields of `SolverOptions` are properties specific to the specified solver `Solver`.

Data Types: `struct`

SGD and ASGD Solver Parameters

`BatchSize` — Mini-batch size
positive integer

This property is read-only.

Mini-batch size, specified as a positive integer. At each iteration during training, `incrementalRegressionLinear` uses `min(BatchSize,numObs)` observations to compute the subgradient, where `numObs` is the number of observations in the training data passed to `fit` or `updateMetricsAndFit`.

If you convert a traditionally trained linear regression model (`RegressionLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `BatchSize` is specified by the `ModelParameters.BatchSize` property of the traditionally trained model. Otherwise, the default is `10`.

Data Types: `single` | `double`

`Lambda` — Ridge (*L*2) regularization term strength
nonnegative scalar

This property is read-only.

Ridge ($L2$) regularization term strength, specified as a nonnegative scalar.

If you convert a traditionally trained linear model for ridge regression (`RegressionLinear` object with the `Regularization` property equal to `'ridge (L2)'`) to create `Mdl`, `Lambda` is specified by the value of the `Lambda` property of the traditionally trained model. Otherwise, the default is `1e-5`.

Data Types: `double` | `single`

`LearnRate` — Learning rate
`'auto'` | positive scalar

This property is read-only.

Learning rate, specified as `'auto'` or a positive scalar. `LearnRate` controls the optimization step size by scaling the objective subgradient.

When you specify `'auto'`:

- If `EstimationPeriod` is `0`, the initial learning rate is `0.7`.
- If `EstimationPeriod` > `0`, the initial learning rate is `1/sqrt(1+max(sum(X.^2,obsDim)))`, where `obsDim` is `1` if the observations compose the columns of the predictor data, and `2` otherwise. `fit` and `updateMetricsAndFit` set the value when you pass the model and training data to either.

If you convert a traditionally trained linear regression model (`RegressionLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `LearnRate` is specified by the `ModelParameters.LearnRate` property of the traditionally trained model.

The `LearnRateSchedule` property determines the learning rate for subsequent learning cycles.

Example: `'LearnRate',0.001`

Data Types: `single` | `double` | `char` | `string`

`LearnRateSchedule` — Learning rate schedule
`'decaying'` (default) | `'constant'`

This property is read-only.

Learning rate schedule, specified as a value in this table, where `LearnRate` specifies the initial learning rate γ_0.

Value	Description
`'constant'`	The learning rate is γ_0 for all learning cycles.

Value	Description

`'decaying'`	The learning rate at learning cycle t is
	$$\gamma_t = \frac{\gamma_0}{(1+\lambda\gamma_0 t)^c}.$$
	• λ is the value of `Lambda`.
	• If `Solver` is `'sgd'`, $c = 1$.
	• If `Solver` is `'asgd'`:
	  • $c = 2/3$ if `Learner` is `'leastsquares'`
	  • $c = 3/4$ if `Learner` is `'svm'` [4]

If you convert a traditionally trained linear regression model (`RegressionLinear`) to create `Mdl`, whose `ModelParameters.Solver` property is `'sgd'` or `'asgd'`, `LearnRate` is `'decaying'`.

Data Types: `char` | `string`

Adaptive Scale-Invariant Solver Options

`Shuffle` — Flag for shuffling observations in batch

`true` (default) | `false`

This property is read-only.

Flag for shuffling the observations in the batch at each learning cycle, specified as a value in this table.

Value	Description
`true`	The software shuffles observations in each incoming batch of data before processing the set. This action reduces bias induced by the sampling scheme.
`false`	The software processes the data in the order received.

Data Types: `logical`

Performance Metrics Parameters

`IsWarm` — Flag indicating whether model tracks performance metrics

`false` | `true`

This property is read-only.

Flag indicating whether the incremental model tracks performance metrics, specified as `false` or `true`. The incremental model `Mdl` is warm (`IsWarm` becomes `true`) after incremental fitting functions fit `MetricsWarmupPeriod` observations to the incremental model (that is, `EstimationPeriod` + `MetricsWarmupPeriod` observations).

Value	Description
`true`	The incremental model `Mdl` is warm. Consequently, `updateMetrics` and `updateMetricsAndFit` track performance metrics in the `Metrics` property of `Mdl`.
`false`	`updateMetrics` and `updateMetricsAndFit` do not track performance metrics.

Data Types: `logical`

`Metrics` — Model performance metrics
table

This property is read-only.

Model performance metrics updated during incremental learning by `updateMetrics` and `updateMetricsAndFit`, specified as a table with two columns and *m* rows, where *m* is the number of metrics specified by the `'Metrics'` name-value pair argument.

The columns of `Metrics` are labeled `Cumulative` and `Window`.

- **`Cumulative`**: Element *j* is the model performance, as measured by metric *j*, from the time the model became warm (IsWarm is `1`).
- **`Window`**: Element *j* is the model performance, as measured by metric *j*, evaluated over all observations within the window specified by the `MetricsWindowSize` property. The software updates `Window` after it processes `MetricsWindowSize` observations.

Rows are labeled by the specified metrics. For details, see `'Metrics'`.

Data Types: `table`

`MetricsWarmupPeriod` — Number of observations fit before tracking performance metrics
`1000` (default) | nonnegative integer

This property is read-only.

Number of observations the incremental model must be fit to before it tracks performance metrics in its `Metrics` property, specified as a nonnegative integer.

For more details, see “Performance Metrics” on page 33-3037.

Data Types: `single` | `double`

`MetricsWindowSize` — Number of observations to use to compute window performance metrics
`200` (default) | positive integer

This property is read-only.

Number of observations to use to compute window performance metrics, specified as a positive integer.

For more details on performance metrics options, see “Performance Metrics” on page 33-3037.

Data Types: `single` | `double`

Object Functions

fit	Train linear model for incremental learning
updateMetricsAndFit	Update performance metrics in linear model for incremental learning given new data and train model
updateMetrics	Update performance metrics in linear model for incremental learning given new data
loss	Loss of linear model for incremental learning on batch of data
predict	Predict responses for new observations from linear model for incremental learning

Examples

Create Incremental Learner Without Any Prior Information

Create a default incremental linear model for regression.

```
Mdl = incrementalRegressionLinear()
```

```
Mdl =
  incrementalRegressionLinear

               IsWarm: 0
              Metrics: [1x2 table]
    ResponseTransform: 'none'
                 Beta: [0x1 double]
                 Bias: 0
              Learner: 'svm'


  Properties, Methods
```

```
Mdl.EstimationPeriod
```

```
ans = 1000
```

`Mdl` is an `incrementalRegressionLinear` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations. The software sets the estimation period to 1000 because half the width of the epsilon insensitive band `Epsilon` is unknown. You can set `Epsilon` to a positive floating point scalar by using the `'Epsilon'` name-value pair argument. This action results in a default estimation period of 0.

Load the robot arm data set.

```
load robotarm
```

For details on the data set, enter `Description` at the command line.

Fit the incremental model to the training data by using the `updateMetricsAndfit` function. To simulate a data stream fit the model in chunks of 50 observations at a time. At each iteration:

- Process 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.

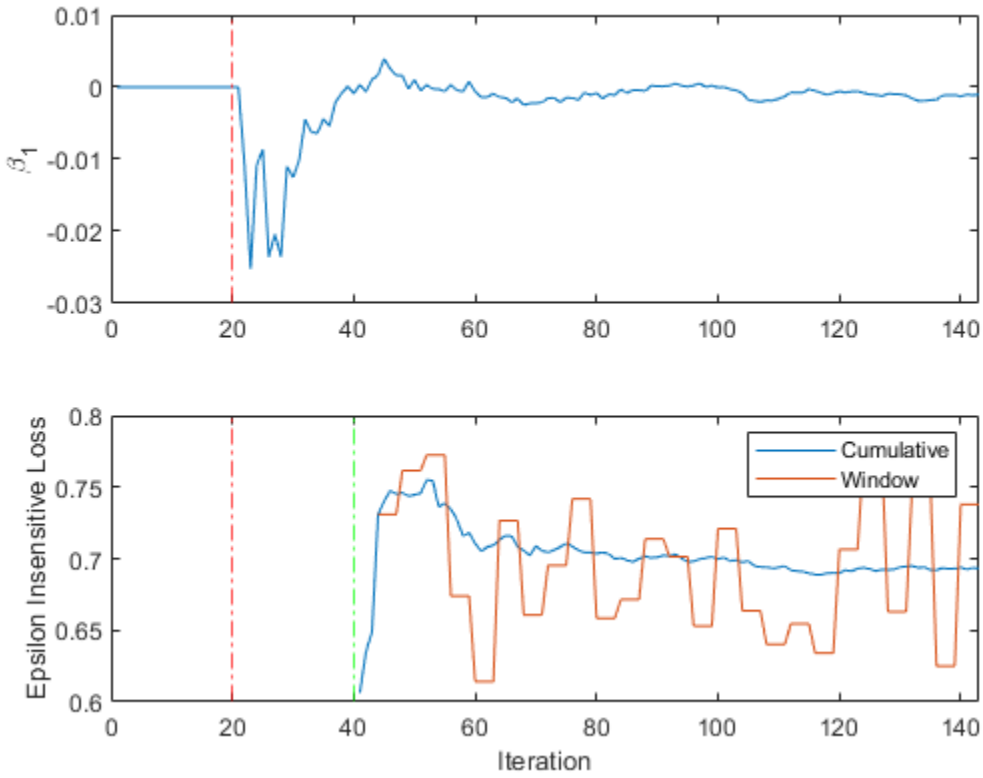
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

```
% Preallocation
n = numel(ytrain);
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,Xtrain(idx,:),ytrain(idx));
    ei{j,:} = Mdl.Metrics{"EpsilonInsensitiveLoss",:};
    beta1(j + 1) = Mdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. While `updateMetricsAndFit` processes the first 1000 observations, it stores a buffer to estimate `Epsilon`; the function does not fit the coefficients until after this estimation period. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline((Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod)/numObsPerChunk,'g-.');
legend(h,ei.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- After the estimation period (first 20 iterations), fit β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (4 iterations).

Configure Incremental Learning Options

Prepare an incremental regression learner by specifying a metrics warm-up period, during which the `updateMetricsAndFit` function only fits the model. Specify a metrics window size of 500 observations. Train the model by using SGD, and adjust the SGD batch size, learning rate, and regularization parameter.

Load the robot arm data set.

```
load robotarm
n = numel(ytrain);
```

For details on the data set, enter `Description` at the command line.

Create an incremental linear model for regression. Configure the model as follows:

- Specify the SGD solver.
- Assume that a ridge regularization parameter value of 0.001, SGD batch size of 20, learning rate of 0.002, and half the width of the epsilon insensitive band for SVM of 0.05 work well for the problem.
- Specify that the incremental fitting functions process the raw (unstandardized) predictor data.
- Specify a metrics warm-up period of 1000 observations.
- Specify a metrics window size of 500 observations.
- Track the epsilon insensitive loss, MSE, and mean absolute error (MAE) to measure the performance of the model. The software supports epsilon insensitive loss and MSE. Create an anonymous function that measures the absolute error of each new observation. Create a structure array containing the name `MeanAbsoluteError` and its corresponding function.

```
maefcn = @(z,zfit)abs(z - zfit);
maemetric = struct("MeanAbsoluteError",maefcn);

Mdl = incrementalRegressionLinear('Epsilon',0.05,...
    'Solver','sgd','Lambda',0.001,'BatchSize',20,'LearnRate',0.002,...
    'Standardize',false,...
    'MetricsWarmupPeriod',1000,'MetricsWindowSize',500,...
    'Metrics',{'epsiloninsensitive' 'mse' maemetric})
```

```
Mdl =
  incrementalRegressionLinear

               IsWarm: 0
              Metrics: [3x2 table]
    ResponseTransform: 'none'
                 Beta: [0x1 double]
                 Bias: 0
              Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalRegressionLinear` model object configured for incremental learning without an estimation period.

Fit the incremental model to the data by using `updateMetricsAndfit` function. At each iteration:

- Simulate a data stream by processing a chunk of 50 observations. Note that chunk size is different from SGD batch size.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the estimated coefficient β_{10}, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

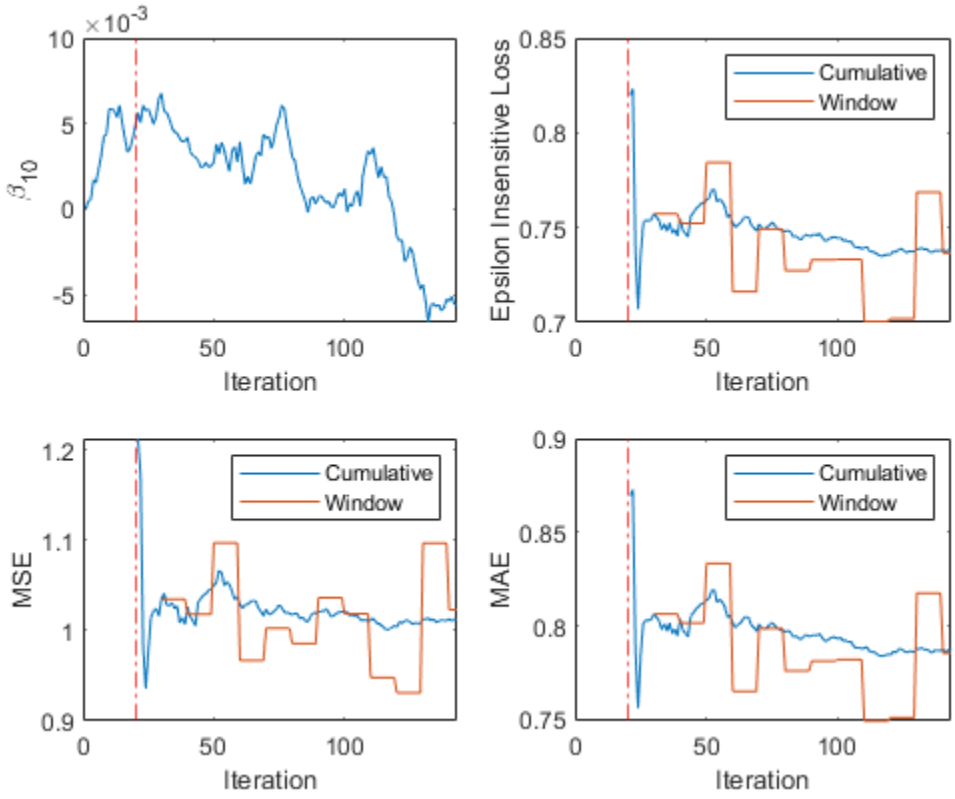
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mse = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mae = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta10 = zeros(nchunk,1);
```

```
% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,Xtrain(idx,:),ytrain(idx));
    ei{j,:} = Mdl.Metrics{"EpsilonInsensitiveLoss",:};
    mse{j,:} = Mdl.Metrics{"MeanSquaredError",:};
    mae{j,:} = Mdl.Metrics{"MeanAbsoluteError",:};
    beta10(j + 1) = Mdl.Beta(10);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_{10} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,2,1)
plot(beta10)
ylabel('\beta_{10}')
xlim([0 nchunk]);
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
subplot(2,2,2)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,ei.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,3)
h = plot(mse.Variables);
xlim([0 nchunk]);
ylabel('MSE')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mse.Properties.VariableNames)
xlabel('Iteration')
subplot(2,2,4)
h = plot(mae.Variables);
xlim([0 nchunk]);
ylabel('MAE')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mae.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fit β_{10} during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations (10 iterations).

Convert Traditionally Trained Model to Incremental Learner

Train a linear regression model by using `fitrlinear`, convert it to an incremental learner, track its performance, and fit it to streaming data. Carry over training options from traditional to incremental learning.

Load and Preprocess Data

Load the 2015 NYC housing data set, and shuffle the data. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
rng(1); % For reproducibility
n = size(NYCHousing2015,1);
idxshuff = randsample(n,n);
NYCHousing2015 = NYCHousing2015(idxshuff,:);
```

Suppose that the data collected from Manhattan (`BOROUGH = 1`) was collected using a new method that doubles its quality. Create a weight variable that attributes `2` to observations collected from Manhattan, and `1` to all other observations.

```
NYCHousing2015.W = ones(n,1) + (NYCHousing2015.BOROUGH == 1);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data. Transpose the resulting predictor matrix.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}]';
```

Train Linear Regression Model

Fit a linear regression model to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitrlinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns',...
    'Weights',NYCHousing2015.W(idxtt))
```

```
TTMdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [313x1 double]
                 Bias: 0.1116
               Lambda: 2.1977e-05
              Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `RegressionLinear` model object representing a traditionally trained linear regression model.

Convert Trained Model

Convert the traditionally trained linear regression model to a linear regression model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalRegressionLinear
```

```
              IsWarm: 1
             Metrics: [1x2 table]
    ResponseTransform: 'none'
                Beta: [313x1 double]
                Bias: 0.1116
             Learner: 'svm'


  Properties, Methods
```

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. Simulate a data stream by processing 500 observations at a time. At each iteration:

1. Call `updateMetrics` to update the cumulative and window epsilon insensitive loss of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model. Specify that the observations are oriented in columns, and specify the observation weights.
2. Call `fit` to fit the model to the incoming chunk of observations. Overwrite the previous incremental model to update the model parameters. Specify that the observations are oriented in columns, and specify the observation weights.
3. Store the losses and last estimated coefficient β_{313}.

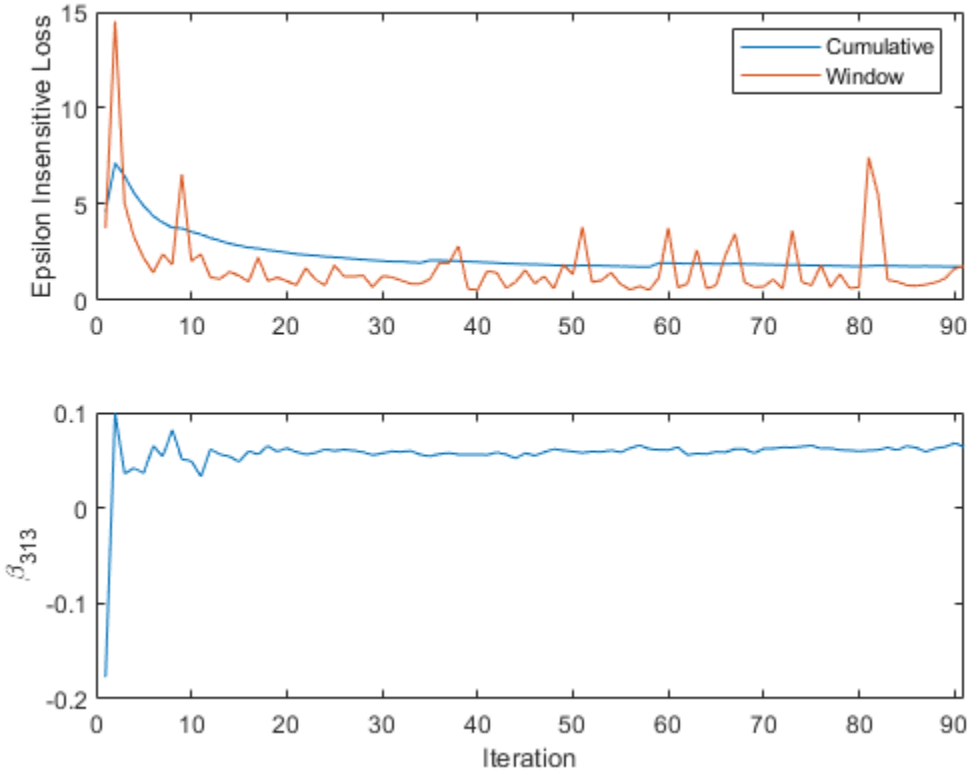
```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 500;
nchunk = floor(nil/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta313 = [IncrementalMdl.Beta(end); zeros(nchunk,1)];
Xil = X(:,idxil);
Yil = Y(idxil);
Wil = NYCHousing2015.W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(:,idx),Yil(idx),...
        'ObservationsIn','columns','Weights',Wil(idx));
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    IncrementalMdl = fit(IncrementalMdl,Xil(:,idx),Yil(idx),'ObservationsIn','columns',...
        'Weights',Wil(idx));
    beta313(j + 1) = IncrementalMdl.Beta(end);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

Alternatively, you can use `updateMetricsAndFit` to update performance metrics of the model given a new chunk of data, and then fit the model to the data.

Plot a trace plot of the performance metrics and estimated coefficient β_{313}.

```
figure;
subplot(2,1,1)
h = plot(ei.Variables);
xlim([0 nchunk]);
ylabel('Epsilon Insensitive Loss')
legend(h,ei.Properties.VariableNames)
subplot(2,1,2)
plot(beta313)
ylabel('\beta_{313}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss gradually changes with each iteration (chunk of 500 observations), whereas the window loss jumps. Because the metrics window is 200 by default, `updateMetrics` measures the performance based on the latest 200 observations in each 500 observation chunk.

β_{313} changes abruptly, then levels off as `fit` processes chunks of observations.

More About

Incremental Learning

Incremental learning, or online learning, is a branch of machine learning concerned with processing incoming data from a data stream, possibly given little to no knowledge of the distribution of the

predictor variables, aspects of the prediction or objective function (including tuning parameter values), or whether the observations are labeled. Incremental learning differs from traditional machine learning, where enough labeled data is available to fit to a model, perform cross-validation to tune hyperparameters, and infer the predictor distribution.

Given incoming observations, an incremental learning model processes data in any of the following ways, but usually in this order:

- Predict labels.
- Measure the predictive performance.
- Check for structural breaks or drift in the model.
- Fit the model to the incoming observations.

Adaptive Scale-Invariant Solver for Incremental Learning

The adaptive scale-invariant solver for incremental learning, introduced in [1], is a gradient-descent-based objective solver for training linear predictive models. The solver is hyperparameter free, insensitive to differences in predictor variable scales, and does not require prior knowledge of the distribution of the predictor variables. These characteristics make it well suited to incremental learning.

The standard SGD and ASGD solvers are sensitive to differing scales among the predictor variables, resulting in models that can perform poorly. To achieve better accuracy using SGD and ASGD, you can standardize the predictor data, and tune the regularization and learning rate parameters can require tuning. For traditional machine learning, enough data is available to enable hyperparameter tuning by cross-validation and predictor standardization. However, for incremental learning, enough data might not be available (for example, observations might be available only one at a time) and the distribution of the predictors might be unknown. These characteristics make parameter tuning and predictor standardization difficult or impossible to do during incremental learning.

The incremental fitting functions for regression `fit` and `updateMetricsAndFit` use the more conservative ScInOL1 version of the algorithm.

Tips

- After creating a model, you can generate C/C++ code that performs incremental learning on a data stream. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

Estimation Period

During the estimation period, incremental fitting functions `fit` and `updateMetricsAndFit` use the first incoming `EstimationPeriod` observations to estimate (tune) hyperparameters required for incremental training. This table describes the hyperparameters and when they are estimated or tuned. Estimation occurs only when `EstimationPeriod` is positive.

Hyperparameter	Model Property	Use	Hyperparameters Estimated
Predictor means and standard deviations	`Mu` and `Sigma`	Standardize predictor data	When both these conditions apply: • Incremental fitting functions are configured to standardize predictor data (see "Standardize Data" on page 33-3036). • `Mdl.Mu` and `Mdl.Sigma` are empty arrays `[]`.
Learning rate	LearnRate	Adjust solver step size	When both these conditions apply: • The solver is SGD or ASGD (see Solver). • You do not set the `'LearnRate'` name-value pair argument.
Half the width of the epsilon insensitive band	Epsilon	Control number of support vectors	When both these conditions apply: • The learner is SVM (see Learner). • You do not set the `'Epsilon'` name-value pair argument.

The functions fit only the last estimation period observation to the incremental model, and they do not use any of the observations to track the performance of the model. At the end of the estimation period, the functions update the properties that store the hyperparameters.

Standardize Data

If incremental learning functions are configured to standardize predictor variables, they do so using the means and standard deviations stored in the `Mu` and `Sigma` properties of the incremental learning model `Mdl`.

- When you set `'Standardize',true` and a positive estimation period (see EstimationPeriod), and `Mdl.Mu` and `Mdl.Sigma` are empty, incremental fitting functions estimate means and standard deviations using the estimation period observations.
- When you set `'Standardize','auto'` (the default), the following conditions apply.
  - If you create `incrementalRegressionLinear` by converting a traditionally trained SVM regression model (`CompactRegressionSVM` or `RegressionSVM`), and the `Mu` and `Sigma` properties of the model being converted are empty arrays `[]`, incremental learning functions do not standardize predictor variables. If the `Mu` and `Sigma` properties of the model being converted are nonempty, incremental learning functions standardize the predictor variables using the specified means and standard deviations. Incremental fitting functions do not estimate new means and standard deviations regardless of the length of the estimation period.
  - If you create `incrementalRegressionLinear` by converting a linear regression model (`RegressionLinear`), incremental learning functions does not standardize the data regardless of the length of the estimation period.
  - If you do not convert a traditionally trained model, incremental learning functions standardize the predictor data only when you specify an SGD solver (see `Solver`) and a positive estimation period (see EstimationPeriod).

- When incremental fitting functions estimate predictor means and standard deviations, the functions compute weighted means and weighted standard deviations using the estimation period observations. Specifically, the functions standardize predictor j (x_j) using

  $$x_j^* = \frac{x_j - \mu_j^*}{\sigma_j^*}.$$

  - x_j is predictor j, and x_{jk} is observation k of predictor j in the estimation period.
  - $\mu_j^* = \frac{1}{\sum_k w_k} \sum_k w_k x_{jk}$.
  - $\left(\sigma_j^*\right)^2 = \frac{1}{\sum_k w_k} \sum_k w_k \left(x_{jk} - \mu_j^*\right)^2$.
  - w_j is observation weight j.

Performance Metrics

- The `updateMetrics` and `updateMetricsAndFit` functions track model performance metrics (`'Metrics'`) from new data when the incremental model is warm (IsWarm property). An incremental model is warm after `fit` or `updateMetricsAndFit` fit the incremental model to MetricsWarmupPeriod observations, which is the metrics warm-up period.

  If EstimationPeriod > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` and `updateMetricsAndFit` update the metrics at the following frequencies:

  - `Cumulative` — The functions compute cumulative metrics since the start of model performance tracking. The functions update metrics every time you call the functions and base the calculation on the entire supplied data set.
  - `Window` — The functions compute metrics based on all observations within a window determined by the MetricsWindowSize name-value pair argument. `MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `MetricsWindowSize` is 20, the functions compute metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `MetricsWindowSize` observations enter the

buffer, and the earliest observations are removed from the buffer. For example, suppose `MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the functions use the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

References

[1] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[4] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- All object functions on page 33-3026 of an `incrementalRegressionLinear` model object support code generation.
- If you configure `Mdl` to shuffle data (see Solver and Shuffle), the `fit` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB.
- When you generate code that loads or creates an `incrementalRegressionLinear` model object, the `NumPredictors` property must reflect the number of predictor variables.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Functions

`fit` | `incrementalLearner` | `incrementalLearner` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Objects

`CompactRegressionSVM` | `RegressionLinear` | `RegressionSVM`

Topics

“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Linear Regression Using Succinct Workflow” on page 26-16

“Implement Incremental Learning for Linear Regression Using Flexible Workflow” on page 26-22
“Initialize Incremental Learning Model from SVM Regression Model Trained in Regression Learner” on page 26-30

Introduced in R2020b

isempty

Class: `dataset`

(Not Recommended) True for empty dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
tf = isempty(A)
```

Description

`tf = isempty(A)` returns true (`1`) if `A` is an empty dataset and false (`0`) otherwise. An empty array has no elements, that is `prod(size(A))==0`.

See Also

`size`

islevel

(Not Recommended) Determine if levels are in nominal or ordinal array

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
tf = islevel(levels,A)
```

Description

`tf = islevel(levels,A)` returns a logical array indicating which of the levels in `levels` correspond to a level in the nominal or ordinal array `A`.

Input Arguments

`A` — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

`levels` — Levels to test

character vector | string array | cell array of character vectors | 2-D character matrix

Levels to test, specified as a character vector, string array, cell array of character vectors, or 2-D character matrix.

Data Types: `char` | `string` | `cell`

Output Arguments

`tf` — Level indicator

logical array

Level indicator, returned as a logical array of the same size as `levels`. `tf` has the value `1` (true) when the corresponding element of `levels` is the label of a level in the nominal or ordinal array `A`, even if the level contains no elements. Otherwise, `tf` has the value `0` (false).

See Also

`isequal` | `ismember` | `nominal` | `ordinal`

Introduced in R2007a

ismember

Class: `dataset`

(Not Recommended) Dataset array elements that are members of set

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
LiA = ismember(A,B)
LiA = ismember(A,B,vars)
[LiA,LocB] = ismember( ___ )
```

Description

`LiA = ismember(A,B)` for `dataset` arrays `A` and `B` returns a vector of logical values the same length as `A`. The output vector, `LiA`, has value `1` (true) in the elements that correspond to observations in `A` that are also present in `B`, and `0` (false) otherwise.

`LiA = ismember(A,B,vars)` returns a vector of logical values the same length as `A`. The output vector, `LiA`, has value `1` (true) in the elements that correspond to observations in `A` that are also present in `B` for the variables specified in `vars` only, and `0` (false) otherwise.

`[LiA,LocB] = ismember( ___ )` also returns a vector the same length as `A` containing the index to the first observation in `B` that corresponds to each observation in `A`, or `0` if there is no such observation. You can use any of the previous input arguments.

Input Arguments

`A`

Query dataset array, containing the observations to be found in `B`.

`B`

Set dataset array. When an observation in `A` is found in `B`, for all variables or only those variables specified in `vars`, the corresponding element of `LiA` is `1`.

`vars`

String array or cell array of character vectors containing variable names, or a vector of integers containing variable column numbers. `vars` indicates which variables to match observations on in `A` and `B`.

Output Arguments

`LiA`

Vector of logical values the same length as `A`. `LiA` has value `1` (true) when the corresponding observation in `A` is present in `B`. Otherwise, `LiA` has value `0` (false).

If you specify `vars`, `LiA` has value `1` when the corresponding observation in `A` is present in `B` for the variables in `vars` only.

`LocB`

Vector the same length as `A` containing the index to the first observation in `B` that corresponds to each observation in `A`, for all variables or only those variables specified in `vars`.

Examples

Find Observations That Are Members of a Dataset Array

Load sample data.

```
load('hospital')
B = hospital(1:50,1:5);
```

This set dataset array, `B`, has 50 observations on 5 variables.

Specify a query dataset array.

```
rng('default')
rIx = randsample(100,10);
A = hospital(rIx,1:5)
```

```
A =
                LastName             Sex       Age    Weight    Smoker
    YLN-495    {'COLEMAN'  }        Male      39     188       false
    LQW-768    {'TAYLOR'   }        Female    31     132       false
    DGC-290    {'BUTLER'   }        Male      38     184       true
    DAU-529    {'REED'     }        Male      50     186       true
    REV-997    {'ALEXANDER'}        Male      25     171       true
    QEQ-082    {'COX'      }        Female    28     111       false
    AGR-528    {'SIMMONS'  }        Male      45     181       false
    PUE-347    {'YOUNG'    }        Female    25     114       false
    HVR-372    {'RUSSELL'  }        Male      44     188       true
    XUE-826    {'JACKSON'  }        Male      25     174       false
```

Check which observations in `A` are present in `B`.

```
LiA = ismember(A,B)
```

```
LiA = 10x1 logical array

   0
   1
   0
   0
```

```
   0
   0
   0
   1
   0
   1
```

Display the observations in A that are present in B.

```
A(LiA,:)
```

```
ans =
                LastName           Sex       Age    Weight    Smoker
    LQW-768    {'TAYLOR' }        Female    31     132       false
    PUE-347    {'YOUNG'  }        Female    25     114       false
    XUE-826    {'JACKSON'}        Male      25     174       false
```

Find the location of the observations in B.

```
[~,LocB] = ismember(A,B)
```

```
LocB = 10×1

     0
    10
     0
     0
     0
     0
     0
    28
     0
    13
```

Display the observations in B that match observations in A.

```
B(LocB(LocB>0),:)
```

```
ans =
                LastName           Sex       Age    Weight    Smoker
    LQW-768    {'TAYLOR' }        Female    31     132       false
    PUE-347    {'YOUNG'  }        Female    25     114       false
    XUE-826    {'JACKSON'}        Male      25     174       false
```

See Also

dataset | intersect | setdiff | setxor | sortrows | union | unique

Topics

“Dataset Arrays” on page 2-112

ismissing

Class: `dataset`

(Not Recommended) Find dataset array elements with missing values

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
I = ismissing(ds)
I = ismissing(ds,Name,Value)
```

Description

`I = ismissing(ds)` returns a logical array that indicates which elements in the dataset array, `ds`, contain a missing value. By default, `ismissing` recognizes `NaN` as a missing value in numeric variables, `''` as a missing value in character variables, and `<undefined>` as a missing value in categorical arrays.

- `ds2 = ds(~any(I,2),:)` creates a new dataset array containing only the complete observations in `ds`.
- `ds2 = ds(:,~any(I,1))` creates a new dataset array containing only the variables from `ds` with no missing values.

`I = ismissing(ds,Name,Value)` returns missing value indices with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ds`

`dataset` array

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NumericTreatAsMissing`

Vector of numeric values to treat as missing value indicators in floating point `ds` variables. `ismissing` always treats a `NaN` value as a missing value.

Default:

`StringTreatAsMissing`

Character vector, string array, or cell array of character vectors to treat as missing value indicators in character `ds` variables. `ismissing` always treats `''` as a missing value.

Output Arguments

`I`

Logical array indicating which elements in `ds` contain a missing value. `I` is the same size as `ds`, with value `1` for elements that contain a missing value.

See Also

`dataset` | `isempty` | `isnan` | `isundefined` | `replaceWithMissing`

Topics

“Clean Messy and Missing Data” on page 2-97
“Dataset Arrays” on page 2-112

isvalid

Class: `qrandstream`

Test handle validity

Syntax

```
tf = isvalid(h)
```

Description

`tf = isvalid(h)` performs an element-wise check for validity on the handle elements of `h`. The result is a logical array of the same dimensions as `h`, where each element is the element-wise validity result.

A handle is invalid if it has been deleted or if it is an element of a handle array and has not yet been initialized.

See Also

`delete` | `qrandstream`

iwishrnd

Inverse Wishart random numbers

Syntax

```
W = iwishrnd(Tau,df)
W = iwishrnd(Tau,df,DI)
[W,DI] = iwishrnd(Tau,df)
```

Description

`W = iwishrnd(Tau,df)` generates a random matrix `W` from the inverse Wishart distribution with parameters `Tau` and `df`. The inverse of `W` has the Wishart distribution with covariance matrix `Sigma = inv(Tau)` and with `df` degrees of freedom. `Tau` is a symmetric and positive definite matrix.

`W = iwishrnd(Tau,df,DI)` expects `DI` to be the transpose of the inverse of the Cholesky factor of `Tau`, so that `DI'*DI = inv(Tau)`, where `inv` is the MATLAB inverse function. `DI` is lower-triangular and the same size as `Tau`. If you call `iwishrnd` multiple times using the same value of `Tau`, it is more efficient to supply `DI` instead of computing it each time.

`[W,DI] = iwishrnd(Tau,df)` returns `DI` so you can use it as an input in future calls to `iwishrnd`.

Note that different sources use different parametrizations for the inverse Wishart distribution. This function defines the parameter `tau` so that the mean of the output matrix is `Tau/(df-d-1)` where `d` is the dimension of `Tau`.

See Also

`wishrnd`

Topics

“Inverse Wishart Distribution” on page B-76

Introduced before R2006a

jackknife

Jackknife sampling

Syntax

```
jackstat = jackknife(jackfun,X)
jackstat = jackknife(jackfun,X,Y,...)
jackstat = jackknife(jackfun,...,'Options',option)
```

Description

`jackstat = jackknife(jackfun,X)` draws jackknife data samples from the `n`-by-`p` data array `X`, computes statistics on each sample using the function `jackfun`, and returns the results in the matrix `jackstat`. `jackknife` regards each row of `X` as one data sample, so there are `n` data samples. Each of the `n` rows of `jackstat` contains the results of applying `jackfun` to one jackknife sample. `jackfun` is a function handle specified with `@`. Row `i` of `jackstat` contains the results for the sample consisting of `X` with the `i`th row omitted:

```
s = x;
s(i,:) = [];
jackstat(i,:) = jackfun(s);
```

If `jackfun` returns a matrix or array, then this output is converted to a row vector for storage in `jackstat`. If `X` is a row vector, it is converted to a column vector.

`jackstat = jackknife(jackfun,X,Y,...)` accepts additional arguments to be supplied as inputs to `jackfun`. They may be scalars, column vectors, or matrices. `jackknife` creates each jackknife sample by sampling with replacement from the rows of the non-scalar data arguments (these must have the same number of rows). Scalar data are passed to `jackfun` unchanged. Non-scalar arguments must have the same number of rows, and each jackknife sample omits the same row from these arguments.

`jackstat = jackknife(jackfun,...,'Options',option)` provides an option to perform jackknife iterations in parallel, if the Parallel Computing Toolbox is available. Set `'Options'` as a structure you create with `statset`. `jackknife` uses the following field in the structure:

`'UseParallel'`	If `true`, use multiple processors to compute jackknife iterations. If the Parallel Computing Toolbox is not installed, then computation occurs in serial mode. Default is `false`, meaning serial computation.

Examples

Estimate the bias of the MLE variance estimator of random samples taken from the vector `y` using `jackknife`. The bias has a known formula in this problem, so you can compare the `jackknife` value to this formula.

```
sigma = 5;
y = normrnd(0,sigma,100,1);
m = jackknife(@var,y,1);
n = length(y);
```

```
bias = -sigma^2/n % known bias formula
jbias = (n-1)*(mean(m)-var(y,1)) % jackknife bias estimate

bias =
   -0.2500

jbias =
   -0.3378
```

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`bootstrp` | `histogram` | `ksdensity` | `random` | `randsample`

Topics

“Jackknife Resampling” on page 3-16

Introduced in R2006a

jbtest

Jarque-Bera test

Syntax

```
h = jbtest(x)
h = jbtest(x,alpha)
h = jbtest(x,alpha,mctol)
[h,p] = jbtest(___)
[h,p,jbstat,critval] = jbtest(___)
```

Description

`h = jbtest(x)` returns a test decision for the null hypothesis that the data in vector `x` comes from a normal distribution with an unknown mean and variance, using the Jarque-Bera test on page 33-3054. The alternative hypothesis is that it does not come from such a distribution. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = jbtest(x,alpha)` returns a test decision for the null hypothesis at the significance level specified by `alpha`.

`h = jbtest(x,alpha,mctol)` returns a test decision based on a *p*-value computed using a Monte Carlo simulation with a maximum Monte Carlo standard error on page 33-3054 less than or equal to `mctol`.

`[h,p] = jbtest(___)` also returns the *p*-value `p` of the hypothesis test, using any of the input arguments from the previous syntaxes.

`[h,p,jbstat,critval] = jbtest(___)` also returns the test statistic `jbstat` and the critical value `critval` for the test.

Examples

Test for a Normal Distribution

Load the data set.

```
load carbig
```

Test the null hypothesis that car mileage, in miles per gallon (`MPG`), follows a normal distribution across different makes of cars.

```
h = jbtest(MPG)
```

```
h = 1
```

The returned value of `h = 1` indicates that `jbtest` rejects the null hypothesis at the default 5% significance level.

Test the Hypothesis at a Different Significance Level

Load the data set.

```
load carbig
```

Test the null hypothesis that car mileage in miles per gallon (MPG) follows a normal distribution across different makes of cars at the 1% significance level.

```
[h,p] = jbtest(MPG,0.01)
```

```
h = 1
```

```
p = 0.0022
```

The returned value of `h = 1`, and the returned *p*-value less than `α = 0.01` indicate that `jbtest` rejects the null hypothesis.

Test for a Normal Distribution Using Monte Carlo Simulation

Load the data set.

```
load carbig
```

Test the null hypothesis that car mileage, in miles per gallon (MPG), follows a normal distribution across different makes of cars. Use a Monte Carlo simulation to obtain an exact *p*-value.

```
[h,p,jbstat,critval] = jbtest(MPG,[],0.0001)
```

```
h = 1
```

```
p = 0.0022
```

```
jbstat = 18.2275
```

```
critval = 5.8461
```

The returned value of `h = 1` indicates that `jbtest` rejects the null hypothesis at the default 5% significance level. Additionally, the test statistic, `jbstat`, is larger than the critical value, `critval`, which indicates rejection of the null hypothesis.

Input Arguments

x — Sample data

vector

Sample data for the hypothesis test, specified as a vector. `jbtest` treats `NaN` values in `x` as missing values and ignores them.

Data Types: `single` | `double`

alpha — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as a scalar value in the range (0,1). If `alpha` is in the range [0.001,0.50], and if the sample size is less than or equal to 2000, `jbtest` looks up the critical value for the test in a table of precomputed values. To conduct the test at a significance level outside of these specifications, use `mctol`.

Example: `0.01`

Data Types: `single` | `double`

`mctol` — Maximum Monte Carlo standard error
nonnegative scalar value

Maximum Monte Carlo standard error on page 33-3054 for the *p*-value, `p`, specified as a nonnegative scalar value. If you specify a value for `mctol`, `jbtest` computes a Monte Carlo approximation for `p` directly, rather than interpolating into a table of precomputed values. `jbtest` chooses the number of Monte Carlo replications large enough to make the Monte Carlo standard error for `p` less than `mctol`.

If you specify a value for `mctol`, you must also specify a value for `alpha`. You can specify `alpha` as `[]` to use the default value of 0.05.

Example: `0.0001`

Data Types: `single` | `double`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `alpha` significance level.

`p` — *p*-value
scalar value in the range (0,1)

p-value of the test, returned as a scalar value in the range (0,1). `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`jbtest` warns when `p` is not found within the tabulated range of [0.001,0.50], and returns either the smallest or largest tabulated value. In this case, you can use `mctol` to compute a more accurate *p*-value.

`jbstat` — Test statistic
nonnegative scalar value

Test statistic for the Jarque-Bera test, returned as a nonnegative scalar value.

`critval` — Critical value
nonnegative scalar value

Critical value for the Jarque-Bera test at the `alpha` significance level, returned as a nonnegative scalar value. If `alpha` is in the range [0.001,0.50], and if the sample size is less than or equal to 2000,

`jbtest` looks up the critical value for the test in a table of precomputed values. If you use `mctol`, `jbtest` determines the critical value of the test using a Monte Carlo simulation. The null hypothesis is rejected when `jbstat > critval`.

More About

Jarque-Bera Test

The Jarque-Bera test is a two-sided goodness-of-fit test suitable when a fully specified null distribution is unknown and its parameters must be estimated.

The test is specifically designed for alternatives in the Pearson system of distributions. The test statistic is

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right),$$

where n is the sample size, s is the sample skewness, and k is the sample kurtosis. For large sample sizes, the test statistic has a chi-square distribution with two degrees of freedom.

Monte Carlo Standard Error

The Monte Carlo standard error is the error due to simulating the p-value.

The Monte Carlo standard error is calculated as

$$SE = \sqrt{\frac{(\widehat{p})(1-\widehat{p})}{\text{mcreps}}},$$

where $\widehat{p}$ is the estimated p-value of the hypothesis test, and `mcreps` is the number of Monte Carlo replications performed. `jbtest` chooses the number of Monte Carlo replications, `mcreps`, large enough to make the Monte Carlo standard error for $\widehat{p}$ less than the value specified for `mctol`.

Algorithms

Jarque-Bera tests often use the chi-square distribution to estimate critical values for large samples, deferring to the Lilliefors test (see `lillietest`) for small samples. `jbtest`, by contrast, uses a table of critical values computed using Monte Carlo simulation for sample sizes less than 2000 and significance levels from 0.001 to 0.50. Critical values for a test are computed by interpolating into the table, using the analytic chi-square approximation only when extrapolating for larger sample sizes.

References

[1] Jarque, C. M., and A. K. Bera. “A Test for Normality of Observations and Regression Residuals.” *International Statistical Review*. Vol. 55, No. 2, 1987, pp. 163–172.

[2] Deb, P., and M. Sefton. “The Distribution of a Lagrange Multiplier Test of Normality.” *Economics Letters*. Vol. 51, 1996, pp. 123–130. This paper proposed a Monte Carlo simulation for determining the distribution of the test statistic. The results of this function are based on an independent Monte Carlo simulation, not the results in this paper.

See Also

adtest | kstest | lillietest

Introduced before R2006a

johnsrnd

Johnson system random numbers

Syntax

```
r = johnsrnd(quantiles,m,n)
r = johnsrnd(quantiles)
[r,type] = johnsrnd(...)
[r,type,coefs] = johnsrnd(...)
```

Description

`r = johnsrnd(quantiles,m,n)` returns an `m`-by-`n` matrix of random numbers drawn from the distribution in the Johnson system that satisfies the quantile specification given by `quantiles`. `quantiles` is a four-element vector of quantiles for the desired distribution that correspond to the standard normal quantiles [–1.5 –0.5 0.5 1.5]. In other words, you specify a distribution from which to draw random values by designating quantiles that correspond to the cumulative probabilities [0.067 0.309 0.691 0.933]. `quantiles` may also be a 2-by-4 matrix whose first row contains four standard normal quantiles, and whose second row contains the corresponding quantiles of the desired distribution. The standard normal quantiles must be spaced evenly.

Note Because `r` is a random sample, its sample quantiles typically differ somewhat from the specified distribution quantiles.

`r = johnsrnd(quantiles)` returns a scalar value.

`r = johnsrnd(quantiles,m,n,...)` or `r = johnsrnd(quantiles,[m,n,...])` returns an `m`-by-`n`-by-... array.

`[r,type] = johnsrnd(...)` returns the type of the specified distribution within the Johnson system. `type` is `'SN'`, `'SL'`, `'SB'`, or `'SU'`. Set `m` and `n` to zero to identify the distribution type without generating any random values.

The four distribution types in the Johnson system correspond to the following transformations of a normal random variate:

- `'SN'` — Identity transformation (normal distribution on page B-119)
- `'SL'` — Exponential transformation (lognormal distribution on page B-88)
- `'SB'` — Logistic transformation (bounded)
- `'SU'` — Hyperbolic sine transformation (unbounded)

`[r,type,coefs] = johnsrnd(...)` returns coefficients `coefs` of the transformation that defines the distribution. `coefs` is `[gamma, eta, epsilon, lambda]`. If `z` is a standard normal random variable and `h` is one of the transformations defined above, `r = lambda*h((z-gamma)/eta) +epsilon` is a random variate from the distribution type corresponding to `h`.

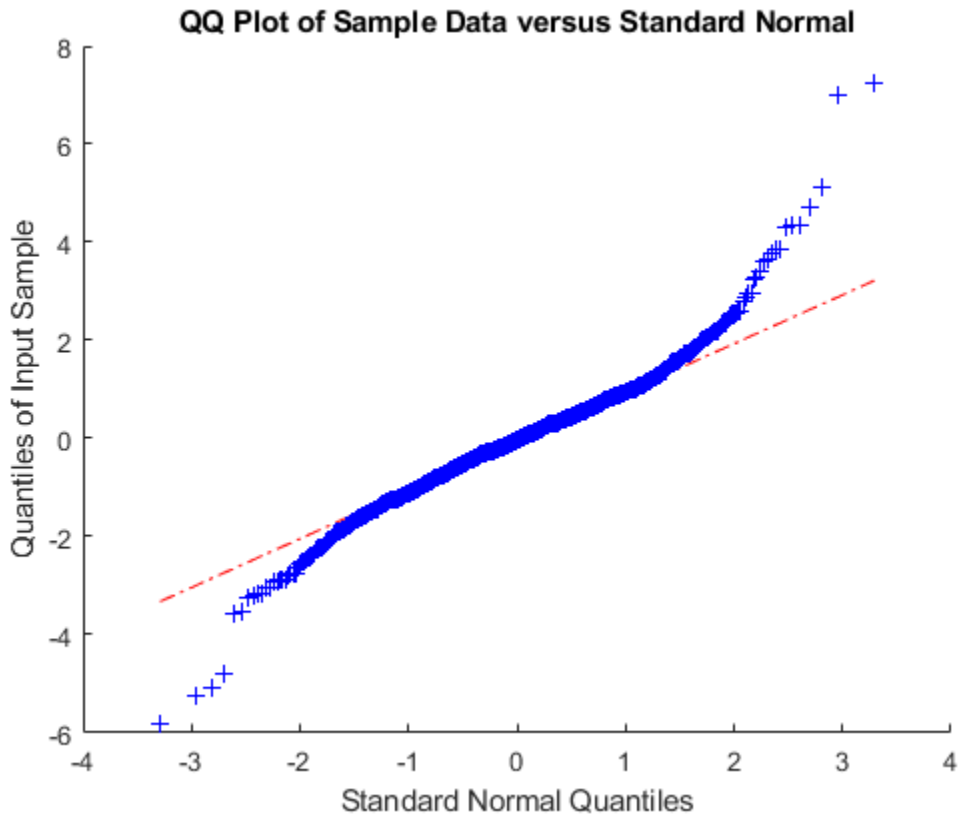
Examples

Generate Random Samples Using the Johnson System

This example shows several different approaches to using the Johnson system of flexible distribution families to generate random numbers and fit a distribution to sample data.

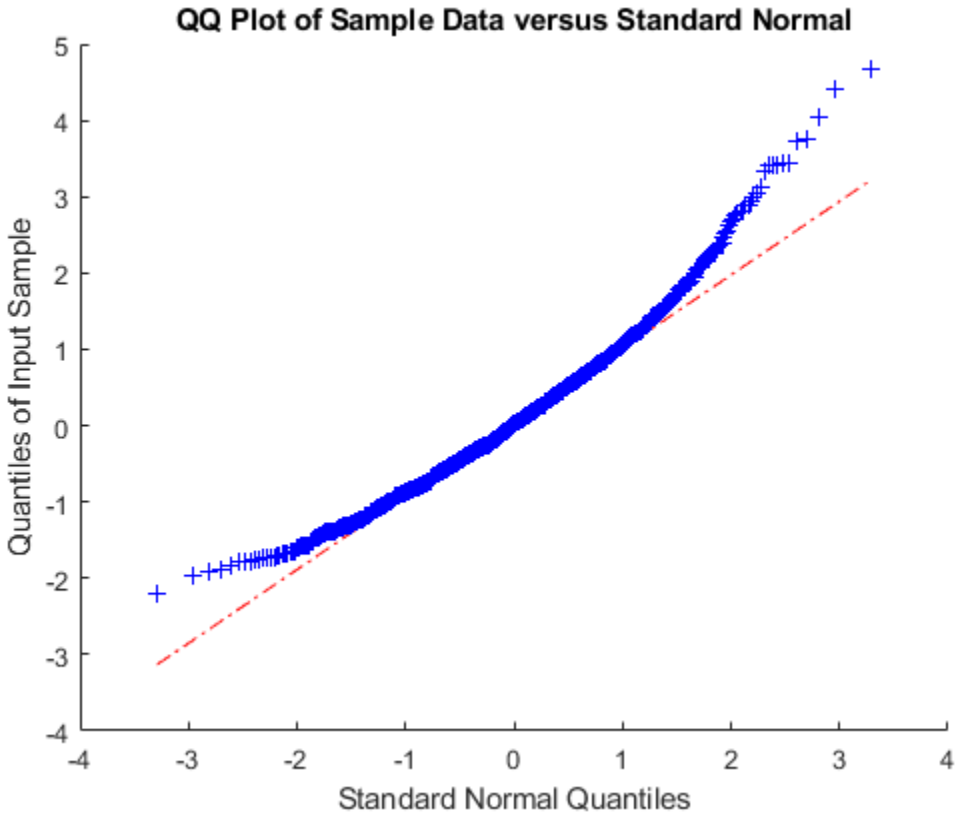
Generate random values with longer tails than a standard normal.

```
rng default;  % For reproducibility
r = johnsrnd([-1.7 -.5 .5 1.7],1000,1);
figure;
qqplot(r);
```



Generate random values skewed to the right.

```
r = johnsrnd([-1.3 -.5 .5 1.7],1000,1);
figure;
qqplot(r);
```



Generate random values that match some sample data well in the right-hand tail.

```
load carbig;
qnorm = [.5 1 1.5 2];
q = quantile(Acceleration, normcdf(qnorm));
r = johnsrnd([qnorm;q],1000,1);
[q;quantile(r,normcdf(qnorm))]
```

```
ans = 2×4

   16.7000   18.2086   19.5376   21.7263
   16.6986   18.2220   19.9078   22.0918
```

Determine the distribution type and the coefficients.

```
[r,type,coefs] = johnsrnd([qnorm;q],0)
```

```
r =

     []

type =
'SU'

coefs = 1×4

    1.0920    0.5829   18.4382    1.4494
```

See Also

`pearsrnd` | `random`

Topics

“Generating Data Using Flexible Families of Distributions” on page 7-20

Introduced in R2006a

join

Class: `dataset`

(Not Recommended) Merge dataset array observations

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = join(A,B)
C = join(A,B,keys)
C = join(A,B,param1,val1,param2,val2,...)
[C,IB] = join(...)
C = join(A,B,'Type',TYPE,...)
C = join(A,B,'Type',TYPE,'MergeKeys',true,...)
[C,IA,IB] = join(A,B,'Type',TYPE,...)
```

Description

`C = join(A,B)` creates a dataset array `C` by merging observations from the two dataset arrays `A` and `B`. `join` performs the merge by first finding *key variables*, that is, pairs of dataset variables, one in `A` and one in `B`, that share the same name. Each observation in `B` must contain a unique combination of values in the key variables, and must contain all combinations of values that are present in the keys from `A`. `join` then uses these key variables to define a many-to-one correspondence between observations in `A` and those in `B`. `join` uses this correspondence to replicate the observations in `B` and combine them with the observations in `A` to create `C`.

`C = join(A,B,keys)` performs the merge using the variables specified by `keys` as the key variables in both `A` and `B`. `keys` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

`C` contains one observation for each observation in `A`. Variables in `C` include all of the variables from `A`, as well as one variable corresponding to each variable in `B` (except for the keys from `B`). If `A` and `B` contain variables with identical names, `join` adds the suffix `'_left'` and `'_right'` to the corresponding variables in `C`.

`C = join(A,B,param1,val1,param2,val2,...)` specifies optional parameter name/value pairs to control how the dataset variables in `A` and `B` are used in the merge. Parameters are:

- `'Keys'` — Specifies the variables to use as keys in both `A` and `B`.
- `'LeftKeys'` — Specifies the variables to use as keys in `A`.
- `'RightKeys'` — Specifies the variables to use as keys in `B`.

You may provide either the `'Keys'` parameter, or both the `'LeftKeys'` and `'RightKeys'` parameters. The value for these parameters is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. `'LeftKeys'` or `'RightKeys'` must both specify the same number of key variables, and `join` pairs the left and right keys in the order specified.

- `'LeftVars'` — Specifies which variables from `A` to include in `C`. By default, `join` includes all variables from `A`.
- `'RightVars'` — Specifies which variables from `B` to include in `C`. By default, `join` includes all variables from `B` except the key variables.

You can use `'LeftVars'` or `'RightVars'` to include or exclude key variables as well as data variables. The value for these parameters is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

`[C,IB] = join(...)` returns an index vector `IB`, where `join` constructs `C` by horizontally concatenating `A(:,LeftVars)` and `B(IB,RightVars)`. `join` can also perform more complicated inner and outer join operations that allow a many-to-many correspondence between `A` and `B`, and allow unmatched observations in either `A` or `B`.

`C = join(A,B,'Type',TYPE,...)` performs the join operation specified by `TYPE`. `TYPE` is one of `'inner'`, `'leftouter'`, `'rightouter'`, `'fullouter'`, or `'outer'` (which is a synonym for `'fullouter'`). For an inner join, `C` only contains observations corresponding to a combination of key values that occurred in both `A` and `B`. For a left (or right) outer join, `C` also contains observations corresponding to keys in `A` (or `B`) that did not match any in `B` (or `A`). Variables in `C` taken from `A` (or `B`) contain null values in those observations. A full outer join is equivalent to a left and right outer join. `C` contains variables corresponding to the key variables from both `A` and `B`, and `join` sorts the observations in `C` by the key values.

For inner and outer joins, `C` contains variables corresponding to the key variables from both `A` and `B` by default, as well as all the remaining variables. `join` sorts the observations in the result `C` by the key values.

`C = join(A,B,'Type',TYPE,'MergeKeys',true,...)` includes a single variable in `C` for each key variable pair from `A` and `B`, rather than including two separate variables. For outer joins, `join` creates the single variable by merging the key values from `A` and `B`, taking values from `A` where a corresponding observation exists in `A`, and from `B` otherwise. Setting the `'MergeKeys'` parameter to `true` overrides inclusion or exclusion of any key variables specified via the `'LeftVars'` or `'RightVars'` parameter. Setting the `'MergeKeys'` parameter to `false` is equivalent to not passing in the `'MergeKeys'` parameter.

`[C,IA,IB] = join(A,B,'Type',TYPE,...)` returns index vectors `IA` and `IB` indicating the correspondence between observations in `C` and those in `A` and `B`. For an inner join, `join` constructs `C` by horizontally concatenating `A(IA,LeftVars)` and `B(IB,RightVars)`. For an outer join, `IA` or `IB` may also contain zeros, indicating the observations in `C` that do not correspond to observations in `A` or `B`, respectively.

Examples

Create a dataset array from Fisher's iris data:

```
load fisheriris
NumObs = size(meas,1);
NameObs = strcat({'Obs'},num2str((1:NumObs)','%-d'));
iris = dataset({nominal(species),'species'},...
               {meas,'SL','SW','PL','PW'},...
               'ObsNames',NameObs);
```

Create a separate dataset array with the diploid chromosome counts for each species of iris:

```
snames = nominal({'setosa';'versicolor';'virginica'});
CC = dataset({snames,'species'},{[38;108;70],'cc'})
CC =
    species       cc
    setosa         38
    versicolor    108
    virginica      70
```

Broadcast the data in `CC` to the rows of `iris` using the key variable `species` in each dataset:

```
iris2 = join(iris,CC);
iris2([1 2 51 52 101 102],:)
ans =
           species       SL     SW     PL     PW     cc
 Obs1      setosa        5.1    3.5    1.4    0.2     38
 Obs2      setosa        4.9      3    1.4    0.2     38
 Obs51     versicolor      7    3.2    4.7    1.4    108
 Obs52     versicolor    6.4    3.2    4.5    1.5    108
 Obs101    virginica     6.3    3.3      6    2.5     70
 Obs102    virginica     5.8    2.7    5.1    1.9     70
```

Create two datasets and join them using the `'MergeKeys'` flag:

```
% Create two data sets that both contain the key variable
% 'Key1'. The two arrays contain observations with common
% values of Key1, but each array also contains observations
% with values of Key1 not present in the other.
a = dataset({'a' 'b' 'c' 'e' 'h'}',[1 2 3 11 17]',...
   'VarNames',{'Key1' 'Var1'})
b = dataset({'a' 'b' 'd' 'e'}',[4 5 6 7]',...
   'VarNames',{'Key1' 'Var2'})

% Combine a and b with an outer join, which matches up
% observations with common key values, but also retains
% observations whose key values don't have a match.
% Keep the key values as separate variables in the result.
couter = join(a,b,'key','Key1','Type','outer')

% Join a and b, merging the key values as a single variable
% in the result.
coutermerge = join(a,b,'key','Key1','Type','outer',...
   'MergeKeys',true)

% Join a and b, retaining only observations whose key
% values match.
cinner = join(a,b,'key','Key1','Type','inner',...
   'MergeKeys',true)

a =

    Key1       Var1
    'a'         1
    'b'         2
    'c'         3
    'e'        11
    'h'        17
```

```
b =

    Key1       Var2
    'a'        4
    'b'        5
    'd'        6
    'e'        7


couter =

    Key1_left    Var1    Key1_right    Var2
    'a'            1     'a'             4
    'b'            2     'b'             5
    'c'            3     ''            NaN
    ''           NaN     'd'             6
    'e'           11     'e'             7
    'h'           17     ''            NaN


coutermerge =

    Key1       Var1    Var2
    'a'          1       4
    'b'          2       5
    'c'          3     NaN
    'd'        NaN       6
    'e'         11       7
    'h'         17     NaN


cinner =

    Key1       Var1    Var2
    'a'         1      4
    'b'         2      5
    'e'        11      7
```

See Also

sortrows

KDTreeSearcher

Create *K*d-tree nearest neighbor searcher

Description

`KDTreeSearcher` model objects store the results of a nearest neighbor search that uses the *K*d-tree algorithm. Results include the training data, distance metric and its parameters, and maximum number of data points in each leaf node (that is, the bucket size). The *K*d-tree algorithm partitions an *n*-by-*K* data set by recursively splitting *n* points in *K*-dimensional space into a binary tree.

Once you create a `KDTreeSearcher` model object, you can search the stored tree to find all neighboring points to the query data by performing a nearest neighbor search using `knnsearch` or a radius search using `rangesearch`. The *K*d-tree algorithm is more efficient than the exhaustive search algorithm when *K* is small (that is, $K \leq 10$), the training and query sets are not sparse, and the training and query sets have many observations.

Creation

Use either the `createns` function or the `KDTreeSearcher` function (described here) to create a `KDTreeSearcher` model object. Both functions use the same syntax except that the `createns` function has the `'NSMethod'` name-value pair argument, which you use to choose the nearest neighbor search method. The `createns` function also creates an `ExhaustiveSearcher` object. Specify `'NSMethod','kdtree'` to create a `KDTreeSearcher` object. The default is `'kdtree'` if *K* ≤ 10, the training data is not sparse, and the distance metric is Euclidean, city block, Chebychev, or Minkowski.

Syntax

```
Mdl = KDTreeSearcher(X)
Mdl = KDTreeSearcher(X,Name,Value)
```

Description

`Mdl = KDTreeSearcher(X)` grows a default *K*d-tree (`Mdl`) using the *n*-by-*K* numeric matrix of training data (`X`).

`Mdl = KDTreeSearcher(X,Name,Value)` specifies additional options using one or more name-value pair arguments. You can specify the maximum number of data points in each leaf node (that is, the bucket size) and the distance metric, and set the distance metric parameter (`DistParameter`) property. For example, `KDTreeSearcher(X,'Distance','minkowski','BucketSize',10)` specifies to use the Minkowski distance when searching for nearest neighbors and to use `10` for the bucket size. To specify `DistParameter`, use the `P` name-value pair argument.

Input Arguments

`X` — Training data
numeric matrix

Training data that grows the *K*d-tree, specified as a numeric matrix. `X` has *n* rows, each corresponding to an observation (that is, an instance or example), and *K* columns, each corresponding to a predictor (that is, a feature).

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distance','minkowski','P',3,'BucketSize',10` specifies to use the following when searching for nearest neighbors: the Minkowski distance, `3` for the Minkowski distance metric exponent, and `10` for the bucket size.

`Distance` — Distance metric

`'euclidean'` (default) | `'chebychev'` | `'cityblock'` | `'minkowski'`

Distance metric used when you call `knnsearch` or `rangesearch` to find nearest neighbors for future query points, specified as the comma-separated pair consisting of `'Distance'` and one of these values.

Value	Description
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cityblock'`	City block distance.
`'euclidean'`	Euclidean distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value pair argument.

For more details, see “Distance Metrics” on page 18-12.

The software does not use the distance metric for creating a `KDTreeSearcher` model object, so you can alter the distance metric by using dot notation after creating the object.

Example: `'Distance','minkowski'`

`P` — Exponent for Minkowski distance metric

`2` (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

`BucketSize` — Maximum number of data points in each leaf node

`50` (default) | positive integer

Maximum number of data points in each leaf node of the *K*d-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer.

Example: `'BucketSize',10`

Data Types: `single` | `double`

Properties

X — Training data
numeric matrix

This property is read-only.

Training data that grows the *K*d-tree, specified as a numeric matrix. `X` has *n* rows, each corresponding to an observation (that is, an instance or example), and *K* columns, each corresponding to a predictor (that is, a feature).

The input argument `X` of `createns` or `KDTreeSearcher` sets this property.

Data Types: `single` | `double`

`Distance` — Distance metric
`'chebychev'` | `'cityblock'` | `'minkowski'`

Distance metric used when you call `knnsearch` or `rangesearch` to find nearest neighbors for future query points, specified as `'chebychev'`, `'cityblock'`, `'euclidean'`, or `'minkowski'`.

The `'Distance'` name-value pair argument of `createns` or `KDTreeSearcher` sets this property.

The software does not use the distance metric for creating a `KDTreeSearcher` model object, so you can alter it by using dot notation.

`DistParameter` — Distance metric parameter values
`[]` | positive scalar

Distance metric parameter values, specified as empty (`[]`) or a positive scalar.

If `Distance` is `'minkowski'`, then `DistParameter` is the exponent in the Minkowski distance formula. Otherwise, `DistParameter` is `[]`, indicating that the specified distance metric formula has no parameters.

The `'P'` name-value pair argument of `createns` or `KDTreeSearcher` sets this property.

You can alter `DistParameter` by using dot notation, for example, `Mdl.DistParameter = PNew`, where `PNew` is a positive scalar.

Data Types: `single` | `double`

`BucketSize` — Maximum number of data points in each leaf node
positive integer

This property is read-only.

Maximum number of data points in each leaf node of the *K*d-tree, specified as a positive integer.

The `'BucketSize'` name-value pair argument of `createns` or `KDTreeSearcher` sets this property.

Data Types: `single` | `double`

Object Functions

knnsearch Find k-nearest neighbors using searcher object

rangesearch Find all neighbors within specified distance using searcher object

Examples

Grow Default *K*d-Tree

Grow a four-dimensional *K*d-tree that uses the Euclidean distance.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
[n,k] = size(X)
```

```
n = 150
```

```
k = 4
```

`X` has 150 observations and 4 predictors.

Grow a four-dimensional *K*d-tree using the entire data set as training data.

```
Mdl1 = KDTreeSearcher(X)
```

```
Mdl1 =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl1` is a `KDTreeSearcher` model object, and its properties appear in the Command Window. The object contains information about the grown four-dimensional *K*d-tree, such as the distance metric. You can alter property values using dot notation.

Alternatively, you can grow a *K*d-tree by using `createns`.

```
Mdl2 = createns(X)
```

```
Mdl2 =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

`Mdl2` is also a `KDTreeSearcher` model object, and it is equivalent to `Mdl1`. Because `X` has four columns and the default distance metric is Euclidean, `createns` creates a `KDTreeSearcher` model by default.

To find the nearest neighbors in `X` to a batch of query data, pass the `KDTreeSearcher` model object and the query data to `knnsearch` or `rangesearch`.

Specify the Minkowski Distance for Nearest Neighbor Search

Load Fisher's iris data. Focus on the petal dimensions.

```
load fisheriris
X = meas(:,[3 4]); % Predictors
```

Grow a two-dimensional *K*d-tree using `createns` and the training data. Specify the Minkowski distance metric.

```
Mdl = createns(X,'Distance','Minkowski')
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'minkowski'
    DistParameter: 2
                X: [150x2 double]
```

Because `X` has two columns and the distance metric is Minkowski, `createns` creates a `KDTreeSearcher` model object by default.

Access properties of `Mdl` by using dot notation. For example, use `Mdl.DistParameter` to access the Minkowski distance exponent.

```
Mdl.DistParameter
```

```
ans = 2
```

You can pass query data and `Mdl` to:

- `knnsearch` to find indices and distances of nearest neighbors
- `rangesearch` to find indices of all nearest neighbors within a distance that you specify

Alter Properties of `KDTreeSearcher` Model

Create a `KDTreeSearcher` model object and alter the `Distance` property by using dot notation.

Load Fisher's iris data set.

```
load fisheriris
X = meas;
```

Grow a default four-dimensional *K*d-tree using the entire data set as training data.

```
Mdl = KDTreeSearcher(X)
```

```
Mdl =
  KDTreeSearcher with properties:
```

```
       BucketSize: 50
         Distance: 'euclidean'
    DistParameter: []
                X: [150x4 double]
```

Specify that the neighbor searcher use the Minkowski metric to compute the distances between the training and query data.

```
Mdl.Distance = 'minkowski'
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'minkowski'
    DistParameter: 2
                X: [150x4 double]
```

You can pass `Mdl` and the query data to either `knnsearch` or `rangesearch` to find the nearest neighbors to the points in the query data based on the Minkowski distance.

Search for Nearest Neighbors of Query Data Using Minkowski Distance

Grow a *K*d-tree nearest neighbor searcher object by using the `createns` function. Pass the object and query data to the `knnsearch` function to find *k*-nearest neighbors.

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng(1);                      % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
tIdx = ~ismember(1:n,qIdx); % Indices of training data
Q = meas(qIdx,:);
X = meas(tIdx,:);
```

Grow a four-dimensional *K*d-tree using the training data. Specify the Minkowski distance for finding nearest neighbors.

```
Mdl = createns(X,'Distance','minkowski')
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'minkowski'
    DistParameter: 2
                X: [145x4 double]
```

Because `X` has four columns and the distance metric is Minkowski, `createns` creates a `KDTreeSearcher` model object by default. The Minkowski distance exponent is `2` by default.

Find the indices of the training data (`Mdl.X`) that are the two nearest neighbors of each point in the query data (`Q`).

```
IdxNN = knnsearch(Mdl,Q,'K',2)
```

```
IdxNN = 5×2

    17     4
     6     2
     1    12
    89    66
   124   100
```

Each row of `IdxNN` corresponds to a query data observation, and the column order corresponds to the order of the nearest neighbors, with respect to ascending distance. For example, based on the Minkowski distance, the second nearest neighbor of `Q(3,:)` is `X(12,:)`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations: The `knnsearch` and `rangesearch` functions support code generation.

For more information, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Nearest Neighbor Searcher” on page 32-19.

See Also

`ExhaustiveSearcher` | `createns`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2010a

kfoldEdge

Package: `classreg.learning.partition`

Classification edge for cross-validated ECOC model

Syntax

```
edge = kfoldEdge(CVMdl)
edge = kfoldEdge(CVMdl,Name,Value)
```

Description

`edge = kfoldEdge(CVMdl)` returns the classification edge on page 33-3076 obtained by the cross-validated ECOC model (`ClassificationPartitionedECOC`) `CVMdl`. For every fold, `kfoldEdge` computes the classification edge for validation-fold observations using an ECOC model trained on training-fold observations. `CVMdl.X` contains both sets of observations.

`edge = kfoldEdge(CVMdl,Name,Value)` returns the classification edge with additional options specified by one or more name-value pair arguments. For example, specify the number of folds, decoding scheme, or verbosity level.

Examples

Estimate *k*-Fold Cross-Validation Edge

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

Train and cross-validate an ECOC model using support vector machine (SVM) binary classifiers. Standardize the predictor data using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

Estimate the average of the edges.

```
edge = kfoldEdge(CVMdl)
```

```
edge = 0.4825
```

Alternatively, you can obtain the per-fold edges by specifying the name-value pair `'Mode','individual'` in `kfoldEdge`.

Display Individual Edges for Each Cross-Validation Fold

The classification edge is a relative measure of classifier quality. To determine which folds perform poorly, display the edges for each fold.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Use 8-fold cross-validation, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'KFold',8,'Learners',t,'ClassNames',classOrder);
```

Estimate the classification edge for each fold.

```
edges = kfoldEdge(CVMdl,'Mode','individual')
```

```
edges = 8×1

    0.4792
    0.4872
    0.4259
    0.5302
    0.5064
    0.4575
    0.4860
    0.4687
```

The edges have similar magnitudes across folds. Folds that perform poorly have small edges relative to the other folds.

To return the average classification edge across the folds that perform well, specify the `'Folds'` name-value pair argument.

Select ECOC Model Features by Comparing Cross-Validation Edges

The classifier edge measures the average of the classifier margins. One way to perform feature selection is to compare cross-validation edges from multiple models. Based solely on this criterion, the classifier with the greatest edge is the best classifier.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Define the following two data sets.

- `fullX` contains all the predictors.
- `partX` contains the petal dimensions.

```
fullX = X;
partX = X(:,3:4);
```

For each predictor set, train and cross-validate an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(fullX,Y,'CrossVal','on','Learners',t,...
    'ClassNames',classOrder);
PCVMdl = fitcecoc(partX,Y,'CrossVal','on','Learners',t,...
    'ClassNames',classOrder);
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedECOC` models. By default, the software implements 10-fold cross-validation.

Estimate the edge for each classifier.

```
fullEdge = kfoldEdge(CVMdl)
```

```
fullEdge = 0.4825
```

```
partEdge = kfoldEdge(PCVMdl)
```

```
partEdge = 0.4951
```

The two models have comparable edges.

Input Arguments

`CVMdl` — Cross-validated ECOC model

`ClassificationPartitionedECOC` model

Cross-validated ECOC model, specified as a `ClassificationPartitionedECOC` model. You can create a `ClassificationPartitionedECOC` model in two ways:

- Pass a trained ECOC model (`ClassificationECOC`) to `crossval`.
- Train an ECOC model using `fitcecoc` and specify any one of these cross-validation name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldEdge(CVMdl,'BinaryLoss','hinge')` specifies `'hinge'` as the binary learner loss function.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices for prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'`

Aggregation level for the output, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

This table describes the values.

Value	Description
`'average'`	The output is a scalar average over all folds.

Value	Description
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

edge — Classification edge
numeric scalar | numeric column vector

Classification edge on page 33-3076, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `edge` is the average classification edge over all folds. Otherwise, `edge` is a *k*-by-1 numeric column vector containing the classification edge for each fold, where *k* is the number of folds.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same

scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `edge` | `fitcecoc` | `kfoldMargin` | `kfoldPredict` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

kfoldEdge

Classification edge for observations not used for training

Syntax

```
E = kfoldEdge(obj)
E = kfoldEdge(obj,Name,Value)
```

Description

`E = kfoldEdge(obj)` returns classification edge (average classification margin) obtained by cross-validated classification ensemble `obj`. For every fold, this method computes classification edge for in-fold observations using an ensemble trained on out-of-fold observations.

`E = kfoldEdge(obj,Name,Value)` calculates edge with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

Object of class `ClassificationPartitionedEnsemble`. Create `ens` with `fitcensemble` along with one of the cross-validation options: `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`. Alternatively, create `ens` from a classification ensemble with `crossval`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`folds`

Indices of folds ranging from `1` to `ens.KFold`. Use only these folds for predictions.

Default: `1:ens.KFold`

`mode`

Character vector or string scalar representing the meaning of the output `edge`:

- `'average'` — `edge` is a scalar value, the average over all folds.
- `'individual'` — `edge` is a vector of length `ens.KFold` with one element per fold.
- `'cumulative'` — `edge` is a vector of length `min(ens.NTrainedPerFold)` in which element `J` is obtained by averaging values across all folds for weak learners `1:J` in each fold.

Default: `'average'`

Output Arguments

E

The average classification margin. `E` is a scalar or vector, depending on the setting of the `mode` name-value pair.

Examples

Compute K-Fold Edge of Held-Out Observations

Compute the k-fold edge for an ensemble trained on the Fisher iris data.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Learners',t);
```

Create a cross-validated ensemble from `ens` and find the classification edge.

```
rng(10,'twister') % For reproducibility
cvens = crossval(ens);
E = kfoldEdge(cvens)
```

```
E = 3.2033
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `obj.Prior`.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `obj.X`.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

See Also

`crossval` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict` | `kfoldfun`

kfoldEdge

Package: `classreg.learning.partition`

Classification edge for cross-validated kernel classification model

Syntax

```
edge = kfoldEdge(CVMdl)
edge = kfoldEdge(CVMdl,Name,Value)
```

Description

`edge = kfoldEdge(CVMdl)` returns the classification edge on page 33-3085 obtained by the cross-validated, binary kernel model (`ClassificationPartitionedKernel`) `CVMdl`. For every fold, `kfoldEdge` computes the classification edge for validation-fold observations using a model trained on training-fold observations.

`edge = kfoldEdge(CVMdl,Name,Value)` returns the classification edge with additional options specified by one or more name-value pair arguments. For example, specify the number of folds or the aggregation level.

Examples

Estimate *k*-Fold Cross-Validation Edge

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data.

```
CVMdl = fitckernel(X,Y,'Crossval','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the cross-validated classification edge.

```
edge = kfoldEdge(CVMdl)
```

```
edge = 1.5585
```

Alternatively, you can obtain the per-fold edges by specifying the name-value pair `'Mode','individual'` in `kfoldEdge`.

Feature Selection Using *k*-Fold Edges

Perform feature selection by comparing *k*-fold edges from multiple models. Based solely on this criterion, the classifier with the greatest edge is the best classifier.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Randomly choose half of the predictor variables.

```
rng(1); % For reproducibility
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.5*p));
```

Cross-validate two binary kernel classification models: one that uses all of the predictors, and one that uses half of the predictors.

```
CVMdl = fitckernel(X,Y,'CrossVal','on');
PCVMdl = fitckernel(X(:,idxPart),Y,'CrossVal','on');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedKernel` models. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the *k*-fold edge for each classifier.

```
fullEdge = kfoldEdge(CVMdl)
```

```
fullEdge = 1.5142
```

```
partEdge = kfoldEdge(PCVMdl)
```

```
partEdge = 1.8910
```

Based on the *k*-fold edges, the classifier that uses half of the predictors is the better model.

Input Arguments

`CVMdl` — Cross-validated, binary kernel classification model

`ClassificationPartitionedKernel` model object

Cross-validated, binary kernel classification model, specified as a `ClassificationPartitionedKernel` model object. You can create a

`ClassificationPartitionedKernel` model by using `fitckernel` and specifying any one of the cross-validation name-value pair arguments.

To obtain estimates, `kfoldEdge` applies the same data used to cross-validate the kernel classification model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldEdge(CVMdl,'Mode','individual')` returns the classification edge for each fold.

`Folds` — Fold indices for prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`Mode` — Aggregation level for output

`'average'` (default) | `'individual'`

Aggregation level for the output, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

This table describes the values.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Example: `'Mode','individual'`

Output Arguments

edge — Classification edge

numeric scalar | numeric column vector

Classification edge on page 33-3085, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `edge` is the average classification edge over all folds. Otherwise, `edge` is a *k*-by-1 numeric column vector containing the classification edge for each fold, where *k* is the number of folds.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationPartitionedKernel` | `fitckernel`

Introduced in R2018b

kfoldEdge

Package: `classreg.learning.partition`

Classification edge for cross-validated kernel ECOC model

Syntax

```
edge = kfoldEdge(CVMdl)
edge = kfoldEdge(CVMdl,Name,Value)
```

Description

`edge = kfoldEdge(CVMdl)` returns the classification edge on page 33-3090 obtained by the cross-validated kernel ECOC model (`ClassificationPartitionedKernelECOC`) `CVMdl`. For every fold, `kfoldEdge` computes the classification edge for validation-fold observations using a model trained on training-fold observations.

`edge = kfoldEdge(CVMdl,Name,Value)` returns the classification edge with additional options specified by one or more name-value pair arguments. For example, specify the number of folds, decoding scheme, or verbosity level.

Examples

Estimate *k*-Fold Cross-Validation Edge

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model composed of kernel binary learners.

```
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the cross-validated classification edges.

```
edge = kfoldEdge(CVMdl)
```

```
edge = 0.4145
```

Alternatively, you can obtain the per-fold edges by specifying the name-value pair `'Mode','individual'` in `kfoldEdge`.

Feature Selection Using *k*-Fold Edges

Perform feature selection by comparing *k*-fold edges from multiple models. Based solely on this criterion, the classifier with the greatest edge is the best classifier.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Randomly choose half of the predictor variables.

```
rng(1); % For reproducibility
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.5*p));
```

Cross-validate two ECOC models composed of kernel classification models: one that uses all of the predictors, and one that uses half of the predictors.

```
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on');
PCVMdl = fitcecoc(X(:,idxPart),Y,'Learners','kernel','CrossVal','on');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedKernelECOC` models. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the *k*-fold edge for each classifier.

```
fullEdge = kfoldEdge(CVMdl)
```

```
fullEdge = 0.4092
```

```
partEdge = kfoldEdge(PCVMdl)
```

```
partEdge = 0.4161
```

Based on the *k*-fold edges, the two classifiers are comparable.

Input Arguments

`CVMdl` — Cross-validated kernel ECOC model

`ClassificationPartitionedKernelECOC` model

Cross-validated kernel ECOC model, specified as a `ClassificationPartitionedKernelECOC` model. You can create a `ClassificationPartitionedKernelECOC` model by training an ECOC model using `fitcecoc` and specifying these name-value pair arguments:

- `'Learners'`– Set the value to `'kernel'`, a template object returned by `templateKernel`, or a cell array of such template objects.
- One of the arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldEdge(CVMdl,'BinaryLoss','hinge')` specifies `'hinge'` as the binary learner loss function.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example, `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

where:

- `M` is the *K*-by-*L* coding matrix stored in `Mdl.CodingMatrix`.
- `s` is the 1-by-*L* row vector of classification scores.
- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

By default, if all binary learners are kernel classification models using SVM, then `BinaryLoss` is `'hinge'`. If all binary learners are kernel classification models using logistic regression, then `BinaryLoss` is `'quadratic'`.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3090.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices for prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'`

Aggregation level for the output, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

This table describes the values.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

edge — Classification edge
numeric scalar | numeric column vector

Classification edge on page 33-3090, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `edge` is the average classification edge over all folds. Otherwise, `edge` is a *k*-by-1 numeric column vector containing the classification edge for each fold, where *k* is the number of folds.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_js_j)/2$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

See Also

`ClassificationPartitionedKernelECOC` | `fitcecoc`

Introduced in R2018b

kfoldEdge

Classification edge for observations not used for training

Syntax

```
e = kfoldEdge(CVMdl)
e = kfoldEdge(CVMdl,Name,Value)
```

Description

`e = kfoldEdge(CVMdl)` returns the cross-validated classification edges on page 33-3098 obtained by the cross-validated, binary, linear classification model `CVMdl`. That is, for every fold, `kfoldEdge` estimates the classification edge for observations that it holds out when it trains using all other observations.

`e` contains a classification edge for each regularization strength in the linear classification models that comprise `CVMdl`.

`e = kfoldEdge(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, indicate which folds to use for the edge calculation.

Input Arguments

`CVMdl` — Cross-validated, binary, linear classification model

`ClassificationPartitionedLinear` model object

Cross-validated, binary, linear classification model, specified as a `ClassificationPartitionedLinear` model object. You can create a `ClassificationPartitionedLinear` model using `fitclinear` and specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldEdge applies the same data used to cross-validate the linear classification model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Folds` — Fold indices to use for classification-score prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for classification-score prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

Mode — Edge aggregation level
`'average'` (default) | `'individual'`

Edge aggregation level, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

Value	Description
`'average'`	Returns classification edges averaged over all folds
`'individual'`	Returns classification edges for each fold

Example: `'Mode','individual'`

Output Arguments

e — Cross-validated classification edges
numeric scalar | numeric vector | numeric matrix

Cross-validated classification edges on page 33-3098, returned as a numeric scalar, vector, or matrix.

Let L be the number of regularization strengths in the cross-validated models (that is, L is `numel(CVMdl.Trained{1}.Lambda)`) and F be the number of folds (stored in `CVMdl.KFold`).

- If `Mode` is `'average'`, then `e` is a 1-by-L vector. `e(`j`)` is the average classification edge over all folds of the cross-validated model that uses regularization strength j.
- Otherwise, `e` is an F-by-L matrix. `e(`i`,`j`)` is the classification edge for fold i of the cross-validated model that uses regularization strength j.

To estimate `e`, `kfoldEdge` uses the data that created `CVMdl` (see `X` and `Y`).

Examples

Estimate *k*-Fold Cross-Validation Edge

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Cross-validate a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinear` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the average of the out-of-fold edges.

```
e = kfoldEdge(CVMdl)
```

```
e = 8.1243
```

Alternatively, you can obtain the per-fold edges by specifying the name-value pair `'Mode','individual'` in `kfoldEdge`.

Feature Selection Using *k*-fold Edges

One way to perform feature selection is to compare *k*-fold edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Edge” on page 33-3094.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create these two data sets:

- `fullX` contains all predictors.
- `partX` contains 1/2 of the predictors chosen at random.

```
rng(1); % For reproducibility
p = size(X,1); % Number of predictors
halfPredIdx = randsample(p,ceil(0.5*p));
fullX = X;
partX = X(halfPredIdx,:);
```

Cross-validate two binary, linear classification models: one that uses the all of the predictors and one that uses half of the predictors. Optimize the objective function using SpaRSA, and indicate that observations correspond to columns.

```
CVMdl = fitclinear(fullX,Ystats,'CrossVal','on','Solver','sparsa',...
    'ObservationsIn','columns');
PCVMdl = fitclinear(partX,Ystats,'CrossVal','on','Solver','sparsa',...
    'ObservationsIn','columns');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinear` models.

Estimate the *k*-fold edge for each classifier.

```
fullEdge = kfoldEdge(CVMdl)
```

```
fullEdge = 16.5629
```

```
partEdge = kfoldEdge(PCVMdl)
```

```
partEdge = 13.9030
```

Based on the *k*-fold edges, the classifier that uses all of the predictors is the better model.

Find Good Lasso Penalty Using *k*-fold Edge

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare k-fold edges.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Edge” on page 33-3094.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,11);
```

Cross-validate a binary, linear classification model using 5-fold cross-validation and that uses each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns','KFold',5,...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinear` model. Because `fitclinear` implements 5-fold cross-validation, `CVMdl` contains 5 `ClassificationLinear` models that the software trains on each fold.

Estimate the edges for each fold and regularization strength.

```
eFolds = kfoldEdge(CVMdl,'Mode','individual')
```

```
eFolds = 5×11

    0.9958    0.9958    0.9958    0.9958    0.9958    0.9924    0.9769    0.9204    0.8424    0.8
    0.9991    0.9991    0.9991    0.9991    0.9991    0.9938    0.9777    0.9177    0.8263    0.8
    0.9992    0.9992    0.9992    0.9992    0.9992    0.9942    0.9779    0.9210    0.8255    0.8
    0.9974    0.9974    0.9974    0.9974    0.9974    0.9931    0.9773    0.9131    0.8422    0.8
```

```
    0.9976    0.9976    0.9976    0.9976    0.9976    0.9942    0.9783    0.9197    0.8390    0.8
```

`eFolds` is a 5-by-11 matrix of edges. Rows correspond to folds and columns correspond to regularization strengths in `Lambda`. You can use `eFolds` to identify ill-performing folds, that is, unusually low edges.

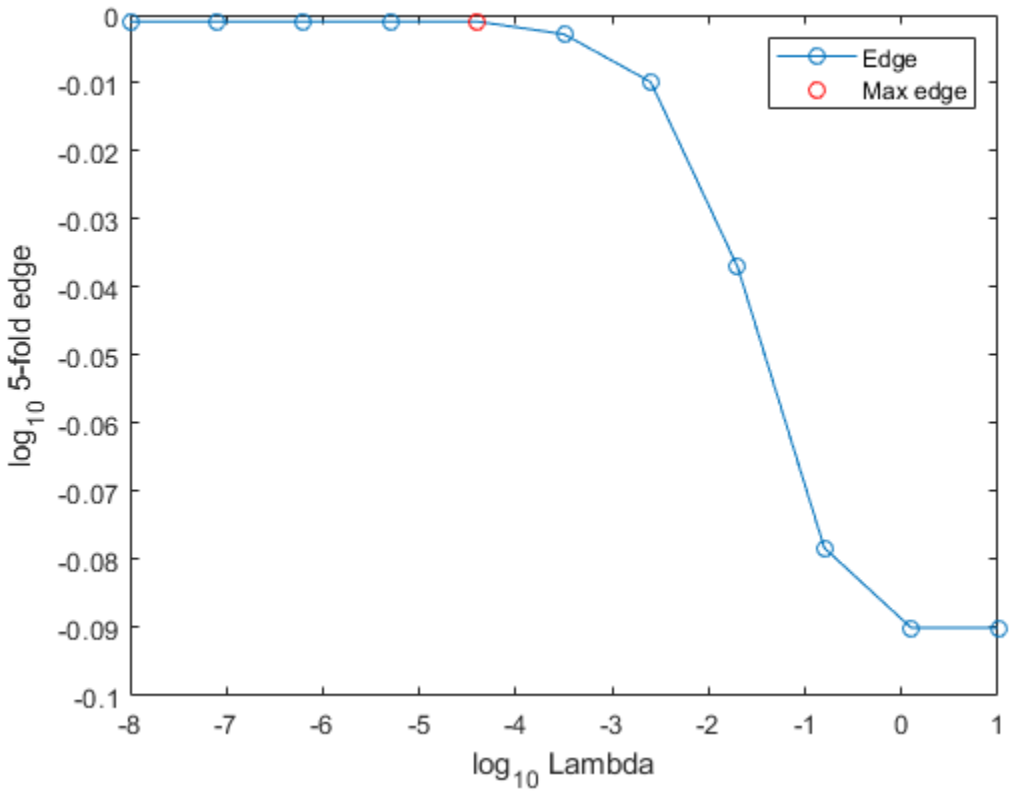
Estimate the average edge over all folds for each regularization strength.

```
e = kfoldEdge(CVMdl)
```

```
e = 1×11

    0.9978    0.9978    0.9978    0.9978    0.9978    0.9936    0.9776    0.9184    0.8351    0.8
```

Determine how well the models generalize by plotting the averages of the 5-fold edge for each regularization strength. Identify the regularization strength that maximizes the 5-fold edge over the grid.

```
figure;
plot(log10(Lambda),log10(e),'-o')
[~, maxEIdx] = max(e);
maxLambda = Lambda(maxEIdx);
hold on
plot(log10(maxLambda),log10(e(maxEIdx)),'ro');
ylabel('log_{10} 5-fold edge')
xlabel('log_{10} Lambda')
legend('Edge','Max edge')
hold off
```



Several values of `Lambda` yield similarly high edges. Higher values of lambda lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the edge starts decreasing.

```
LambdaFinal = Lambda(5);
```

Train a linear classification model using the entire data set and specify the regularization strength `LambdaFinal`.

```
MdlFinal = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',LambdaFinal);
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationLinear` | `ClassificationPartitionedLinear` | `edge` | `kfoldMargin` | `kfoldPredict`

Introduced in R2016a

kfoldEdge

Classification edge for observations not used for training

Syntax

```
e = kfoldEdge(CVMdl)
e = kfoldEdge(CVMdl,Name,Value)
```

Description

`e = kfoldEdge(CVMdl)` returns the cross-validated classification edges on page 33-3108 obtained by the cross-validated, error-correcting output codes (ECOC) model composed of linear classification models `CVMdl`. That is, for every fold, `kfoldEdge` estimates the classification edge for observations that it holds out when it trains using all other observations.

`e` contains a classification edge for each regularization strength in the linear classification models that comprise `CVMdl`.

`e = kfoldEdge(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify a decoding scheme, which folds to use for the edge calculation, or verbosity level.

Input Arguments

`CVMdl` — Cross-validated, ECOC model composed of linear classification models

`ClassificationPartitionedLinearECOC` model object

Cross-validated, ECOC model composed of linear classification models, specified as a `ClassificationPartitionedLinearECOC` model object. You can create a `ClassificationPartitionedLinearECOC` model using `fitcecoc` and by:

1. Specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`
2. Setting the name-value pair argument `Learners` to `'linear'` or a linear classification model template returned by `templateLinear`

To obtain estimates, kfoldEdge applies the same data used to cross-validate the ECOC model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in, loss-function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {-1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes the binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, e.g., `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` should have this form

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

By default, if all binary learners are linear classification models using:

- SVM, then `BinaryLoss` is `'hinge'`
- Logistic regression, then `BinaryLoss` is `'quadratic'`

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme

`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices to use for classification-score prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for classification-score prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`Mode` — Edge aggregation level
`'average'` (default) | `'individual'`

Edge aggregation level, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

Value	Description
`'average'`	Returns classification edges averaged over all folds
`'individual'`	Returns classification edges for each fold

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

e — Cross-validated classification edges

numeric scalar | numeric vector | numeric matrix

Cross-validated classification edges on page 33-3108, returned as a numeric scalar, vector, or matrix.

Let L be the number of regularization strengths in the cross-validated models (that is, L is `numel(CVMdl.Trained{1}.BinaryLearners{1}.Lambda)`) and F be the number of folds (stored in `CVMdl.KFold`).

- If `Mode` is `'average'`, then `e` is a 1-by-L vector. `e(j)` is the average classification edge over all folds of the cross-validated model that uses regularization strength j.
- Otherwise, `e` is a F-by-L matrix. `e(i,j)` is the classification edge for fold i of the cross-validated model that uses regularization strength j.

Examples

Estimate *k*-Fold Cross-Validation Edge

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

For simplicity, use the label 'others' for all observations in `Y` that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Cross-validate a multiclass, linear classification model.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learner','linear','CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the average of the out-of-fold edges.

```
e = kfoldEdge(CVMdl)
```

```
e = 0.7232
```

Alternatively, you can obtain the per-fold edges by specifying the name-value pair `'Mode','individual'` in `kfoldEdge`.

Feature Selection Using *k*-fold Edges

One way to perform feature selection is to compare *k*-fold edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Edge” on page 33-3103, and orient the predictor data so that observations correspond to columns.

```
load nlpdata
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
X = X';
```

Create these two data sets:

- `fullX` contains all predictors.
- `partX` contains a 1/2 of the predictors chosen at random.

```
rng(1); % For reproducibility
p = size(X,1); % Number of predictors
halfPredIdx = randsample(p,ceil(0.5*p));
fullX = X;
partX = X(halfPredIdx,:);
```

Create a linear classification model template that specifies to optimize the objective function using SpaRSA.

```
t = templateLinear('Solver','sparsa');
```

Cross-validate two ECOC models composed of binary, linear classification models: one that uses the all of the predictors and one that uses half of the predictors. Indicate that observations correspond to columns.

```
CVMdl = fitcecoc(fullX,Y,'Learners',t,'CrossVal','on',...
    'ObservationsIn','columns');
PCVMdl = fitcecoc(partX,Y,'Learners',t,'CrossVal','on',...
    'ObservationsIn','columns');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinearECOC` models.

Estimate the *k*-fold edge for each classifier.

```
fullEdge = kfoldEdge(CVMdl)
```

```
fullEdge = 0.3090
```

```
partEdge = kfoldEdge(PCVMdl)
```

```
partEdge = 0.2617
```

Based on the *k*-fold edges, the classifier that uses all of the predictors is the better model.

Find Good Lasso Penalty Using *k*-fold Edge

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare k-fold edges.

Load the NLP data set. Preprocess the data as in “Feature Selection Using k-fold Edges” on page 33-3103.

```
load nlpdata
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
X = X';
```

Create a set of 8 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,8);
```

Create a linear classification model template that specifies to use logistic regression with a lasso penalty, use each of the regularization strengths, optimize the objective function using SpaRSA, and reduce the tolerance on the gradient of the objective function to `1e-8`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8);
```

Cross-validate an ECOC model composed of binary, linear classification models using 5-fold cross-validation and that

```
rng(10) % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns','KFold',5)
```

```
CVMdl =
  ClassificationPartitionedLinearECOC
    CrossValidatedModel: 'LinearECOC'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [comm    dsp    simulink    others]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model.

Estimate the edges for each fold and regularization strength.

```
eFolds = kfoldEdge(CVMdl,'Mode','individual')
```

```
eFolds = 5×8

    0.5533    0.5553    0.5554    0.5541    0.4948    0.2930    0.1044    0.0855
    0.5249    0.5261    0.5265    0.5266    0.4796    0.2948    0.1063    0.0867
    0.5316    0.5323    0.5326    0.5322    0.4792    0.2888    0.1039    0.0868
    0.5375    0.5392    0.5402    0.5364    0.4835    0.2918    0.1018    0.0853
    0.5502    0.5560    0.5585    0.5575    0.4947    0.2929    0.1026    0.0849
```

`eFolds` is a 5-by-8 matrix of edges. Rows correspond to folds and columns correspond to regularization strengths in `Lambda`. You can use `eFolds` to identify ill-performing folds, that is, unusually low edges.

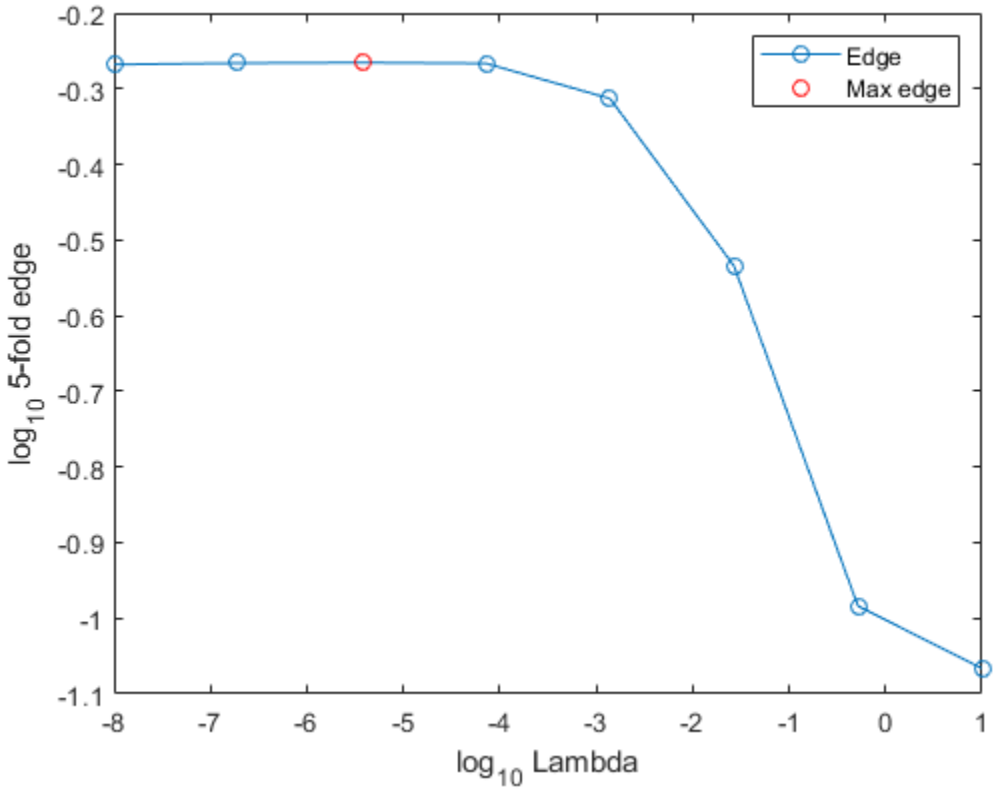
Estimate the average edge over all folds for each regularization strength.

```
e = kfoldEdge(CVMdl)
```

```
e = 1×8

    0.5395    0.5418    0.5426    0.5414    0.4863    0.2923    0.1038    0.0859
```

Determine how well the models generalize by plotting the averages of the 5-fold edge for each regularization strength. Identify the regularization strength that maximizes the 5-fold edge over the grid.

```
figure
plot(log10(Lambda),log10(e),'-o')
[~, maxEIdx] = max(e);
maxLambda = Lambda(maxEIdx);
hold on
plot(log10(maxLambda),log10(e(maxEIdx)),'ro')
ylabel('log_{10} 5-fold edge')
xlabel('log_{10} Lambda')
legend('Edge','Max edge')
hold off
```



Several values of `Lambda` yield similarly high edges. Greater regularization strength values lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the edge starts decreasing.

```
LambdaFinal = Lambda(4);
```

Train an ECOC model composed of linear classification model using the entire data set and specify the regularization strength `LambdaFinal`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',LambdaFinal,'GradientTolerance',1e-8);
MdlFinal = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns');
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$

Value	Description	Score Domain	$g(y_j,s_j)$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_j s_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_j s_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_j s_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationLinear` | `ClassificationPartitionedLinearECOC` | `edge` | `fitcecoc` | `kfoldMargin` | `kfoldPredict` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2016a

kfoldEdge

Package: `classreg.learning.partition`

Classification edge for cross-validated classification model

Syntax

```
E = kfoldEdge(CVMdl)
E = kfoldEdge(CVMdl,Name,Value)
```

Description

`E = kfoldEdge(CVMdl)` returns the classification edge on page 33-3114 obtained by the cross-validated classification model `CVMdl`. For every fold, `kfoldEdge` computes the classification edge for validation-fold observations using a classifier trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`E = kfoldEdge(CVMdl,Name,Value)` returns the classification edge with additional options specified by one or more name-value arguments. For example, specify the folds to use or specify to compute the classification edge for each individual fold.

Examples

Estimate *k*-fold Edge of Classifier

Compute the *k*-fold edge for a model trained on Fisher's iris data.

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification tree classifier.

```
tree = fitctree(meas,species);
```

Cross-validate the classifier using 10-fold cross-validation.

```
cvtree = crossval(tree);
```

Compute the *k*-fold edge.

```
edge = kfoldEdge(cvtree)
```

```
edge = 0.8578
```

Compute K-Fold Edge of Held-Out Observations

Compute the k-fold edge for an ensemble trained on the Fisher iris data.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Learners',t);
```

Create a cross-validated ensemble from `ens` and find the classification edge.

```
rng(10,'twister') % For reproducibility
cvens = crossval(ens);
E = kfoldEdge(cvens)
```

```
E = 3.2033
```

Input Arguments

`CVMdl` — Cross-validated partitioned classifier

`ClassificationPartitionedModel` object | `ClassificationPartitionedEnsemble` object | `ClassificationPartitionedGAM` object

Cross-validated partitioned classifier, specified as a `ClassificationPartitionedModel`, `ClassificationPartitionedEnsemble`, or `ClassificationPartitionedGAM` object. You can create the object in two ways:

- Pass a trained classification model listed in the following table to its `crossval` object function.
- Train a classification model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Classification Model	Function
`ClassificationDiscriminant`	`fitcdiscr`
`ClassificationEnsemble`	`fitcensemble`
`ClassificationGAM`	`fitcgam`
`ClassificationKNN`	`fitcknn`
`ClassificationNaiveBayes`	`fitcnb`
`ClassificationNeuralNetwork`	`fitcnet`
`ClassificationSVM`	`fitcsvm`
`ClassificationTree`	`fitctree`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldEdge(CVMdl,'Folds',[1 2 3 5])` specifies to use the first, second, third, and fifth folds to compute the classification edge, but to exclude the fourth fold.

`Folds` — Fold indices to use

`1:CVMdl.KFold` (default) | positive integer vector

Fold indices to use, specified as a positive integer vector. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `ClassificationPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'` | `'cumulative'`

Aggregation level for the output, specified as `'average'`, `'individual'`, or `'cumulative'`.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length k containing one value per fold, where k is the number of folds.

Value	Description

`'cumulative'`	**Note** If you want to specify this value, `CVMdl` must be a `ClassificationPartitionedEnsemble` object or `ClassificationPartitionedGAM` object.
	
	• If `CVMdl` is `ClassificationPartitionedEnsemble`, then the output is a vector of length `min(CVMdl.NumTrainedPerFold)`. Each element `j` is an average over all folds that the function obtains by using ensembles trained with weak learners `1:j`.
	• If `CVMdl` is `ClassificationPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.
	  • If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained only the intercept (constant) term. The `(j + 1)`th element of `L` is an average obtained using the intercept term and the first `j` predictor trees per linear term.
	  • If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is an average obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

Example: `'Mode','individual'`

Output Arguments

E — Classification edge

numeric scalar | numeric column vector

Classification edge on page 33-3114, returned as a numeric scalar or numeric column vector.

- If `Mode` is `'average'`, then `E` is the average classification edge over all folds.
- If `Mode` is `'individual'`, then `E` is a *k*-by-1 numeric column vector containing the classification edge for each fold, where *k* is the number of folds.
- If `Mode` is `'cumulative'` and `CVMdl` is `ClassificationPartitionedEnsemble`, then `E` is a `min(CVMdl.NumTrainedPerFold)`-by-1 numeric column vector. Each element `j` is the average classification edge over all folds that the function obtains by using ensembles trained with weak learners `1:j`.
- If `Mode` is `'cumulative'` and `CVMdl` is `ClassificationPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.
  - If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first

element of `L` is the average classification edge over all folds that is obtained using only the intercept (constant) term. The `(j + 1)`th element of `L` is the average edge obtained using the intercept term and the first `j` predictor trees per linear term.

- If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is the average classification edge over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is the average edge obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Algorithms

`kfoldEdge` computes the classification edge as described in the corresponding `edge` object function. For a model-specific description, see the appropriate `edge` function reference page in the following table.

Model Type	edge Function
Discriminant analysis classifier	`edge`
Ensemble classifier	`edge`
Generalized additive model classifier	`edge`
k-nearest neighbor classifier	`edge`
Naive Bayes classifier	`edge`
Neural network classifier	`edge`
Support vector machine classifier	`edge`
Binary decision tree for multiclass classification	`edge`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports *k*-nearest neighbor and SVM model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict` | `kfoldfun`

Introduced in R2011a

kfoldfun

Package: `classreg.learning.partition`

Cross-validate function using cross-validated ECOC model

Syntax

```
vals = kfoldfun(CVMdl,fun)
```

Description

`vals = kfoldfun(CVMdl,fun)` cross-validates the function `fun` by applying `fun` to the data stored in the cross-validated ECOC model `CVMdl`. You must pass `fun` as a function handle.

Examples

Estimate Classification Error Using Custom Loss Function

Train a multiclass ECOC classifier, and then cross-validate the model using a custom *k*-fold loss function.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Train and cross-validate an ECOC model using support vector machine (SVM) binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,...
    'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation.

Compute the classification error (proportion of misclassified observations) for the validation-fold observations.

```
L = kfoldLoss(CVMdl)
```

```
L = 0.0400
```

Examine the result when the cost of misclassifying a flower as `versicolor` is `10` and the cost of any other error is `1`. Write a function named `noversicolor` that assigns a cost of `1` for general misclassification and a cost of `10` for misclassifying a flower as `versicolor`.

If you use the live script file for this example, the `noversicolor` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB path.

Compute the mean misclassification error with the `noversicolor` cost.

```
foldLoss = kfoldfun(CVMdl,@noversicolor);
mean(foldLoss)
```

```
ans = 0.0667
```

This code creates the function `noversicolor`.

```
function averageCost = noversicolor(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)
% noversicolor: Example custom cross-validation function that assigns a cost of
%   10 for misclassifying versicolor irises and a cost of 1 for misclassifying
%   the other irises. This example function requires the fisheriris data
%   set.
Ypredict = predict(CMP,Xtest);
misclassified = not(strcmp(Ypredict,Ytest)); % Different result
classifiedAsVersicolor = strcmp(Ypredict,'versicolor'); % Index of bad decisions
cost = sum(misclassified) + ...
    9*sum(misclassified & classifiedAsVersicolor); % Total differences
averageCost = single(cost/numel(Ytest)); % Average error
end
```

Input Arguments

`CVMdl` — Cross-validated ECOC model

`ClassificationPartitionedECOC` model

Cross-validated ECOC model, specified as a `ClassificationPartitionedECOC` model.

`fun` — Cross-validated function

function handle

Cross-validated function, specified as a function handle. `fun` has this syntax:

```
testvals = fun(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)
```

- `CMP` is a compact model stored in one element of the `CVMdl.Trained` property.
- `Xtrain` is the training matrix of predictor values.
- `Ytrain` is the training array of response values.
- `Wtrain` is the set of training weights for observations.
- `Xtest` and `Ytest` are the validation data, with associated weights `Wtest`.
- The returned value `testvals` must have the same size across all folds.

Data Types: `function_handle`

Output Arguments

`vals` — Cross-validation results

numeric matrix

Cross-validation results, returned as a numeric matrix. `vals` corresponds to the arrays of the `testvals` output, concatenated vertically over all the folds. For example, if `testvals` from every fold is a numeric vector of length *n*, `kfoldfun` returns a `KFold`-by-*n* numeric matrix with one row per fold.

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `crossval` | `fitcecoc` | `kfoldEdge` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict`

Introduced in R2014b

kfoldfun

Package: `classreg.learning.partition`

Cross-validate function for classification

Syntax

```
vals = kfoldfun(CVMdl,fun)
```

Description

`vals = kfoldfun(CVMdl,fun)` cross-validates the function `fun` by applying `fun` to the data stored in the cross-validated model `CVMdl`. You must pass `fun` as a function handle.

Examples

Estimate Classification Loss Using Custom Loss Function

Train a classification tree classifier, and then cross-validate it using a custom *k*-fold loss function.

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification tree classifier.

```
Mdl = fitctree(meas,species);
```

`Mdl` is a `ClassificationTree` model.

Cross-validate `Mdl` using the default 10-fold cross-validation. Compute the classification error (proportion of misclassified observations) for the validation-fold observations.

```
rng(1); % For reproducibility
CVMdl = crossval(Mdl);
L = kfoldLoss(CVMdl,'LossFun','classiferror')
```

```
L = 0.0467
```

Examine the result when the cost of misclassifying a flower as `versicolor` is `10`, and the cost of any other misclassification is `1`. Create the custom function `noversicolor` (shown at the end of this example). This function attributes a cost of `10` for misclassifying a flower as `versicolor`, and a cost of `1` for any other misclassification.

Compute the mean misclassification error with the `noversicolor` cost.

```
mean(kfoldfun(CVMdl,@noversicolor))
```

```
ans = 0.2267
```

This code creates the function `noversicolor`.

```
function averageCost = noversicolor(CMP,~,~,~,Xtest,Ytest,~)
% noversicolor Example custom cross-validation function
%    Attributes a cost of 10 for misclassifying versicolor irises, and 1 for
%    the other irises.  This example function requires the fisheriris data
%    set.
Ypredict = predict(CMP,Xtest);
misclassified = not(strcmp(Ypredict,Ytest)); % Different result
classifiedAsVersicolor = strcmp(Ypredict,'versicolor'); % Index of bad decisions
cost = sum(misclassified) + ...
    9*sum(misclassified & classifiedAsVersicolor); % Total differences
averageCost = cost/numel(Ytest); % Average error
end
```

Input Arguments

`CVMdl` — Cross-validated model

`ClassificationPartitionedModel` object | `ClassificationPartitionedEnsemble` object | `ClassificationPartitionedGAM` object

Cross-validated model, specified as a `ClassificationPartitionedModel` object, `ClassificationPartitionedEnsemble` object, or `ClassificationPartitionedGAM` object.

`fun` — Cross-validated function

function handle

Cross-validated function, specified as a function handle. `fun` has the syntax:

```
testvals = fun(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)
```

- `CMP` is a compact model stored in one element of the `CVMdl.Trained` property.
- `Xtrain` is the training matrix of predictor values.
- `Ytrain` is the training array of response values.
- `Wtrain` are the training weights for observations.
- `Xtest` and `Ytest` are the test data, with associated weights `Wtest`.
- The returned value `testvals` must have the same size across all folds.

Data Types: `function_handle`

Output Arguments

`vals` — Cross-validation results

numeric matrix

Cross-validation results, returned as a numeric matrix. `vals` contains the arrays of `testvals` output, concatenated vertically over all folds. For example, if `testvals` from every fold is a numeric vector of length `N`, `kfoldfun` returns a `KFold`-by-`N` numeric matrix with one row per fold.

Data Types: `double`

See Also

`ClassificationPartitionedModel` | `kfoldEdge` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict`

Introduced in R2011a

kfoldfun

Package: `classreg.learning.partition`

Cross-validate function for regression

Syntax

```
vals = kfoldfun(CVMdl,fun)
```

Description

`vals = kfoldfun(CVMdl,fun)` cross-validates the function `fun` by applying `fun` to the data stored in the cross-validated model `CVMdl`. You must pass `fun` as a function handle.

Examples

Estimate Regression Loss Using Custom Loss Function

Train a regression tree model, and then cross-validate it using a custom *k*-fold loss function.

Load the `imports-85` data set. Train a regression tree using a subset of the data.

```
load imports-85
Mdl = fitrtree(X(:,[4 5]),X(:,16),...
    'PredictorNames',{'Length','Width'},...
    'ResponseName','Price');
```

Cross-validate the regression tree, and obtain the mean squared error.

```
CVMdl = crossval(Mdl);
L = kfoldLoss(CVMdl)
```

```
L = 1.9167e+07
```

Examine the error when you use a simple averaging of training responses instead of predictions in the calculation.

```
f = @(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)...
    mean((Ytest-mean(Ytrain)).^2)
```

```
f = function_handle with value:
    @(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)mean((Ytest-mean(Ytrain)).^2)
```

```
mean(kfoldfun(CVMdl,f))
```

```
ans = 6.3586e+07
```

Input Arguments

`CVMdl` — Cross-validated model

`RegressionPartitionedModel` object | `RegressionPartitionedEnsemble` object | `RegressionPartitionedGAM` object | `RegressionPartitionedSVM` object

Cross-validated model, specified as a `RegressionPartitionedModel` object, `RegressionPartitionedEnsemble` object, `RegressionPartitionedGAM` object, or `RegressionPartitionedSVM` object.

`fun` — Cross-validated function

function handle

Cross-validated function, specified as a function handle. `fun` has the syntax:

```
testvals = fun(CMP,Xtrain,Ytrain,Wtrain,Xtest,Ytest,Wtest)
```

- `CMP` is a compact model stored in one element of the `CVMdl.Trained` property.
- `Xtrain` is the training matrix of predictor values.
- `Ytrain` is the training array of response values.
- `Wtrain` are the training weights for observations.
- `Xtest` and `Ytest` are the test data, with associated weights `Wtest`.
- The returned value `testvals` must have the same size across all folds.

Data Types: `function_handle`

Output Arguments

`vals` — Cross-validation results

numeric matrix

Cross-validation results, returned as a numeric matrix. `vals` contains the arrays of `testvals` output, concatenated vertically over all folds. For example, if `testvals` from every fold is a numeric vector of length `N`, `kfoldfun` returns a `KFold`-by-`N` numeric matrix with one row per fold.

Data Types: `double`

See Also

`RegressionPartitionedEnsemble` | `RegressionPartitionedGAM` | `RegressionPartitionedModel` | `RegressionPartitionedSVM` | `kfoldLoss` | `kfoldPredict`

Introduced in R2011a

kfoldLoss

Package: `classreg.learning.partition`

Classification loss for cross-validated ECOC model

Syntax

```
loss = kfoldLoss(CVMdl)
loss = kfoldLoss(CVMdl,Name,Value)
```

Description

`loss = kfoldLoss(CVMdl)` returns the classification loss obtained by the cross-validated ECOC model (`ClassificationPartitionedECOC`) `CVMdl`. For every fold, `kfoldLoss` computes the classification loss for validation-fold observations using a model trained on training-fold observations. `CVMdl.X` contains both sets of observations.

`loss = kfoldLoss(CVMdl,Name,Value)` returns the classification loss with additional options specified by one or more name-value pair arguments. For example, specify the number of folds, decoding scheme, or verbosity level.

Examples

Determine *k*-Fold Cross-Validation Loss

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Train and cross-validate an ECOC model using support vector machine (SVM) binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

Estimate the average classification error.

```
L = kfoldLoss(CVMdl)
```

```
L = 0.0400
```

The average classification error for the folds is 4%.

Alternatively, you can obtain the per-fold losses by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Display Individual Losses for Each Cross-Validation Fold

The classification loss is a measure of classifier quality. To determine which folds perform poorly, display the losses for each fold.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Use 8-fold cross-validation, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'KFold',8,'Learners',t,'ClassNames',classOrder);
```

Estimate the average classification loss across all folds and the losses for each fold.

```
loss = kfoldLoss(CVMdl)
```

```
loss = 0.0333
```

```
losses = kfoldLoss(CVMdl,'Mode','individual')
```

```
losses = 8×1

    0.0556
    0.0526
    0.1579
         0
         0
         0
         0
         0
```

The third fold misclassifies a much higher percentage of observations than any other fold.

Return the average classification loss for the folds that perform well by specifying the `'Folds'` name-value pair argument.

```
newloss = kfoldLoss(CVMdl,'Folds',[1:2 4:8])
```

```
newloss = 0.0153
```

The total classification loss decreases by approximately half its original size.

Consider adjusting parameters of the binary classifiers or the coding design to see if performance for all folds improves.

Determine ECOC Model Quality Using Custom Cross-Validation Loss

In addition to knowing whether a model generally classifies observations correctly, you can determine how well the model classifies an observation into its predicted class. One way to determine this type of model quality is to pass a custom loss function to `kfoldLoss`.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y)  % Class order
```

```
classOrder = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
rng(1) % For reproducibility
```

Train and cross-validate an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

Create a custom function that takes the minimal loss for each observation, then averages the minimal losses for all observations. `S` corresponds to the `NegLoss` output of `kfoldPredict`.

```
lossfun = @(~,S,~,~)mean(min(-S,[],2));
```

Compute the cross-validated custom loss.

```
kfoldLoss(CVMdl,'LossFun',lossfun)
```

```
ans = 0.0101
```

The average minimal binary loss for the validation-fold observations is `0.0101`.

Input Arguments

`CVMdl` — Cross-validated ECOC model

`ClassificationPartitionedECOC` model

Cross-validated ECOC model, specified as a `ClassificationPartitionedECOC` model. You can create a `ClassificationPartitionedECOC` model in two ways:

- Pass a trained ECOC model (`ClassificationECOC`) to `crossval`.
- Train an ECOC model using `fitcecoc` and specify any one of these cross-validation name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldLoss(CVMdl,'Folds',[1 3 5])` specifies to use only the first, third, and fifth folds to calculate the classification loss.

`BinaryLoss` — Binary learner loss function
`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.

- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices for prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`LossFun` — Loss function
`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'classiferror'` or a function handle.

- Specify the built-in function `'classiferror'`. In this case, the loss function is the classification error on page 33-3130.
- Or, specify your own function using function handle notation.

  Assume that *n* is the number of observations in the training data (`CVMdl.NumObservations`) and *K* is the number of classes (`numel(CVMdl.ClassNames)`). Your function needs the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an *n*-by-*K* logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `CVMdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set every element of row `p` to `0`.
  - `S` is an *n*-by-*K* numeric matrix of negated loss values for the classes. Each row corresponds to an observation. The column order corresponds to the class order in `CVMdl.ClassNames`. The input `S` resembles the output argument `NegLoss` of `kfoldPredict`.
  - `W` is an *n*-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes its elements to sum to `1`.
  - `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of 0 for correct classification and 1 for misclassification.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'`

Aggregation level for the output, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

This table describes the values.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.

- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`loss` — Classification loss
numeric scalar | numeric column vector

Classification loss, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `loss` is the average classification loss over all folds. Otherwise, `loss` is a *k*-by-1 numeric column vector containing the classification loss for each fold, where *k* is the number of folds.

More About

Classification Error

The classification error is a binary classification error measure that has the form

$$L = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j},$$

where:

- w_j is the weight for observation j. The software renormalizes the weights to sum to 1.
- $e_j = 1$ if the predicted class of observation j differs from its true class, and 0 otherwise.

In other words, the classification error is the proportion of observations misclassified by the classifier.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).

- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `fitcecoc` | `kfoldPredict` | `loss` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

kfoldLoss

Classification loss for observations not used for training

Syntax

```
L = kfoldLoss(ens)
L = kfoldLoss(ens,Name,Value)
```

Description

`L = kfoldLoss(ens)` returns loss obtained by cross-validated classification model `ens`. For every fold, this method computes classification loss for in-fold observations using a model trained on out-of-fold observations.

`L = kfoldLoss(ens,Name,Value)` calculates loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

Object of class `ClassificationPartitionedEnsemble`. Create `ens` with `fitcensemble` along with one of the cross-validation options: `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`. Alternatively, create `ens` from a classification ensemble with `crossval`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`folds`

Indices of folds ranging from `1` to `ens.KFold`. Use only these folds for predictions.

Default: `1:ens.KFold`

`lossfun`

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal

Value	Description
'exponential'	Exponential loss
'hinge'	Hinge loss
'logit'	Logistic loss
'mincost'	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
'quadratic'	Quadratic loss

'mincost' is appropriate for classification scores that are posterior probabilities.

- Bagged and subspace ensembles return posterior probabilities by default (ens.Method is 'Bag' or 'Subspace').
- If the ensemble method is 'AdaBoostM1', 'AdaBoostM2', GentleBoost, or 'LogitBoost', then, to use posterior probabilities as classification scores, you must specify the double-logit score transform by entering

  ```
  ens.ScoreTransform = 'doublelogit';
  ```

- For all other ensemble methods, the software does not support posterior probabilities as classification scores.

- Specify your own function using function handle notation.

  Suppose that n be the number of observations in X and K be the number of distinct classes (numel(ens.ClassNames), ens is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument lossvalue is a scalar.
  - You choose the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in ens.ClassNames.

    Construct C by setting C(p,q) = 1 if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in ens.ClassNames. S is a matrix of classification scores, similar to the output of predict.
  - W is an n-by-1 numeric vector of observation weights. If you pass W, the software normalizes them to sum to 1.
  - Cost is a *K*-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) - eye(K) specifies a cost of 0 for correct classification, and 1 for misclassification.

  Specify your function using 'LossFun',@*lossfun*.

For more details on loss functions, see “Classification Loss” on page 33-3135.

Default: 'classiferror'

mode

A character vector or string scalar for determining the output of `kfoldLoss`:

- `'average'` — `L` is a scalar, the loss averaged over all folds.
- `'individual'` — `L` is a vector of length `ens.KFold`, where each entry is the loss for a fold.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'average'`

Output Arguments

L

Loss, by default the fraction of misclassified data. `L` can be a vector, and can mean different things, depending on the name-value pair settings.

Examples

Estimate Cross-Validated Classification Error

Load the `ionosphere` data set.

```
load ionosphere
```

Train a classification ensemble of 100 decision trees using AdaBoostM1. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
```

Cross-validate the ensemble using 10-fold cross-validation.

```
cvens = crossval(ens);
```

Estimate the cross-validated classification error.

```
L = kfoldLoss(cvens)
```

```
L = 0.0655
```

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.

- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

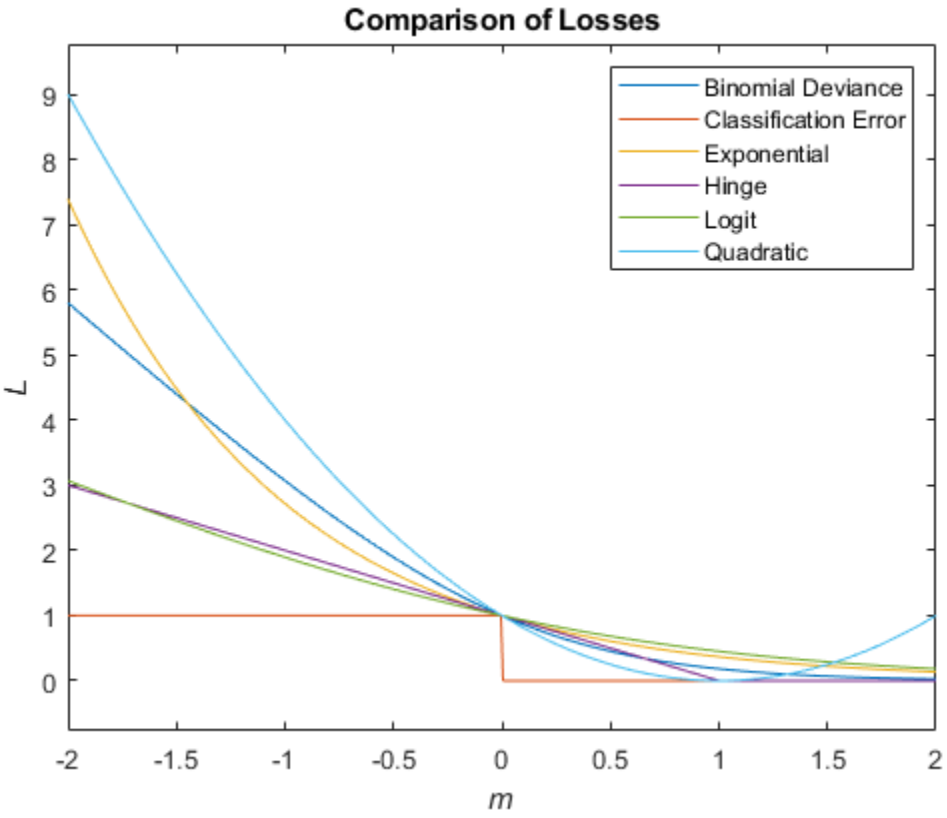
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.

Loss Function	Value of LossFun	Equation
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $$L = -\sum_{j=1}^{n}\frac{\tilde{w}_j \log(m_j)}{Kn},$$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`crossval` | `kfoldEdge` | `kfoldMargin` | `kfoldPredict` | `kfoldfun`

kfoldLoss

Package: `classreg.learning.partition`

Classification loss for cross-validated kernel classification model

Syntax

```
loss = kfoldLoss(CVMdl)
loss = kfoldLoss(CVMdl,Name,Value)
```

Description

`loss = kfoldLoss(CVMdl)` returns the classification loss on page 33-3144 obtained by the cross-validated, binary kernel model (`ClassificationPartitionedKernel`) `CVMdl`. For every fold, `kfoldLoss` computes the classification loss for validation-fold observations using a model trained on training-fold observations.

By default, `kfoldLoss` returns the classification error.

`loss = kfoldLoss(CVMdl,Name,Value)` returns the classification loss with additional options specified by one or more name-value pair arguments. For example, specify the classification loss function, number of folds, or aggregation level.

Examples

Estimate *k*-Fold Cross-Validation Classification Error

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data.

```
CVMdl = fitckernel(X,Y,'Crossval','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the cross-validated classification loss. By default, the software computes the classification error.

```
loss = kfoldLoss(CVMdl)
```

```
loss = 0.0940
```

Alternatively, you can obtain the per-fold classification errors by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Specify Custom Classification Loss

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data.

```
CVMdl = fitckernel(X,Y,'Crossval','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Create an anonymous function that measures linear loss, that is,

$$L = \frac{\sum_j - w_j y_j f_j}{\sum_j w_j}.$$

w_j is the weight for observation j, y_j is the response j (–1 for the negative class and 1 otherwise), and f_j is the raw classification score of observation j.

```
linearloss = @(C,S,W,Cost)sum(-W.*sum(S.*C,2))/sum(W);
```

Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `'LossFun'` name-value pair argument.

Estimate the cross-validated classification loss using the linear loss function.

```
loss = kfoldLoss(CVMdl,'LossFun',linearloss)
```

```
loss = -0.7792
```

Input Arguments

`CVMdl` — Cross-validated, binary kernel classification model

`ClassificationPartitionedKernel` model object

Cross-validated, binary kernel classification model, specified as a `ClassificationPartitionedKernel` model object. You can create a `ClassificationPartitionedKernel` model by using `fitckernel` and specifying any one of the cross-validation name-value pair arguments.

To obtain estimates, `kfoldLoss` applies the same data used to cross-validate the kernel classification model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldLoss(CVMdl,'Folds',[1 3 5])` specifies to use only the first, third, and fifth folds to calculate the classification loss.

`Folds` — Fold indices for prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`LossFun` — Loss function

`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding value.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal

Value	Description
'exponential'	Exponential loss
'hinge'	Hinge loss
'logit'	Logistic loss
'mincost'	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
'quadratic'	Quadratic loss

'mincost' is appropriate for classification scores that are posterior probabilities. For kernel classification models, logistic regression learners return posterior probabilities as classification scores by default, but SVM learners do not (see kfoldPredict).

- Specify your own function by using function handle notation.

  Assume that n is the number of observations in X, and K is the number of distinct classes (numel(CVMdl.ClassNames), where CVMdl is the input model). Your function must have this signature:

  lossvalue = *lossfun*(C,S,W,Cost)

  - The output argument lossvalue is a scalar.
  - You specify the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in CVMdl.ClassNames.

    Construct C by setting C(p,q) = 1, if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in CVMdl.ClassNames. S is a matrix of classification scores, similar to the output of kfoldPredict.
  - W is an n-by-1 numeric vector of observation weights. If you pass W, the software normalizes the weights to sum to 1.
  - Cost is a K-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) – eye(K) specifies a cost of 0 for correct classification, and 1 for misclassification.

Example: 'LossFun',@*lossfun*

Data Types: char | string | function_handle

Mode — Aggregation level for output
'average' (default) | 'individual'

Aggregation level for the output, specified as the comma-separated pair consisting of 'Mode' and 'average' or 'individual'.

This table describes the values.

Value	Description
'average'	The output is a scalar average over all folds.

Value	Description
`'individual'`	The output is a vector of length k containing one value per fold, where k is the number of folds.

Example: `'Mode','individual'`

Output Arguments

`loss` — Classification loss

numeric scalar | numeric column vector

Classification loss on page 33-3144, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `loss` is the average classification loss over all folds. Otherwise, `loss` is a k-by-1 numeric column vector containing the classification loss for each fold, where k is the number of folds.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Suppose the following:

- L is the weighted average classification loss.
- n is the sample size.
- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so that they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

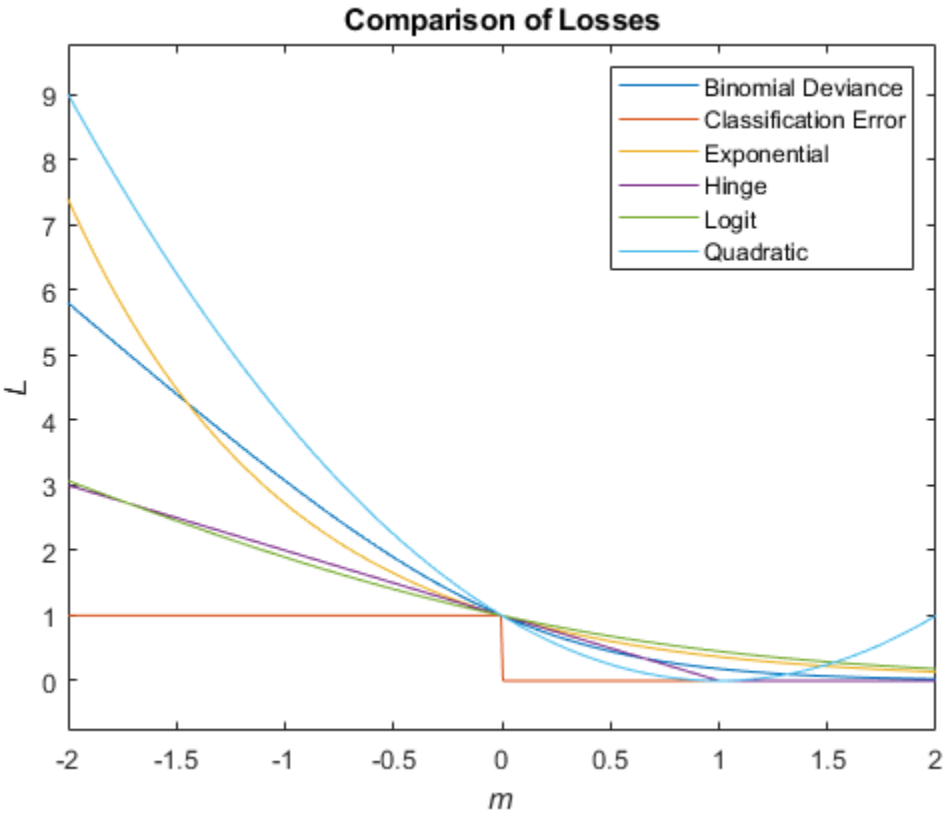
This table describes the supported loss functions that you can specify by using the `'LossFun'` name-value argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$

Loss Function	Value of LossFun	Equation
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`ClassificationPartitionedKernel` | `fitckernel`

Introduced in R2018b

kfoldLoss

Package: `classreg.learning.partition`

Classification loss for cross-validated kernel ECOC model

Syntax

```
loss = kfoldLoss(CVMdl)
loss = kfoldLoss(CVMdl,Name,Value)
```

Description

`loss = kfoldLoss(CVMdl)` returns the classification loss obtained by the cross-validated kernel ECOC model (`ClassificationPartitionedKernelECOC`) `CVMdl`. For every fold, `kfoldLoss` computes the classification loss for validation-fold observations using a model trained on training-fold observations. `kfoldLoss` applies the same data used to create `CVMdl` (see `fitcecoc`).

By default, `kfoldLoss` returns the classification error on page 33-3153.

`loss = kfoldLoss(CVMdl,Name,Value)` returns the classification loss with additional options specified by one or more name-value pair arguments. For example, specify the classification loss function, number of folds, decoding scheme, or verbosity level.

Examples

Estimate *k*-Fold Cross-Validation Classification Error

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model composed of kernel binary learners.

```
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the cross-validated classification loss. By default, the software computes the classification error.

```
loss = kfoldLoss(CVMdl)
```

```
loss = 0.0333
```

Alternatively, you can obtain the per-fold classification errors by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Determine Model Quality Using Custom Cross-Validation Loss

In addition to knowing whether a model generally classifies observations correctly, you can determine how well the model classifies an observation into its predicted class. One way to determine this type of model quality is to pass a custom loss function to `kfoldLoss`.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model composed of kernel binary learners.

```
rng(1) % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Create a custom function that takes the minimal loss for each observation, then averages the minimal losses for all observations. `S` corresponds to the `NegLoss` output of `kfoldPredict`.

```
lossfun = @(~,S,~,~)mean(min(-S,[],2));
```

Compute the cross-validated custom loss.

```
kfoldLoss(CVMdl,'LossFun',lossfun)
```

```
ans = 0.0199
```

The average minimal binary loss for the validation-fold observations is about 0.02.

Input Arguments

`CVMdl` — Cross-validated kernel ECOC model

`ClassificationPartitionedKernelECOC` model

Cross-validated kernel ECOC model, specified as a `ClassificationPartitionedKernelECOC` model. You can create a `ClassificationPartitionedKernelECOC` model by training an ECOC model using `fitcecoc` and specifying these name-value pair arguments:

- `'Learners'` – Set the value to `'kernel'`, a template object returned by `templateKernel`, or a cell array of such template objects.
- One of the arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldLoss(CVMdl,'Folds',[1 3 5])` specifies to use only the first, third, and fifth folds to calculate the classification loss.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$

Value	Description	Score Domain	$g(y_j,s_j)$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example, `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

By default, if all binary learners are kernel classification models using SVM, then `BinaryLoss` is `'hinge'`. If all binary learners are kernel classification models using logistic regression, then `BinaryLoss` is `'quadratic'`.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3153.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices for prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices for prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds` for prediction.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`LossFun` — Loss function
`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'classiferror'` or a function handle.

- Specify the built-in function `'classiferror'`. In this case, the loss function is the classification error on page 33-3153.
- Or, specify your own function using function handle notation.

  Assume that *n* is the number of observations in the training data (`CVMdl.NumObservations`) and *K* is the number of classes (`numel(CVMdl.ClassNames)`). Your function needs the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an *n*-by-*K* logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `CVMdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set every element of row `p` to `0`.
  - `S` is an *n*-by-*K* numeric matrix of negated loss values for the classes. Each row corresponds to an observation. The column order corresponds to the class order in `CVMdl.ClassNames`. The input `S` resembles the output argument `NegLoss` of `kfoldPredict`.
  - `W` is an *n*-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes its elements to sum to `1`.
  - `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of 0 for correct classification and 1 for misclassification.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'`

Aggregation level for the output, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

This table describes the values.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`loss` — Classification loss
numeric scalar | numeric column vector

Classification loss, returned as a numeric scalar or numeric column vector.

If `Mode` is `'average'`, then `loss` is the average classification loss over all folds. Otherwise, `loss` is a *k*-by-1 numeric column vector containing the classification loss for each fold, where *k* is the number of folds.

More About

Classification Error

The classification error is a binary classification error measure that has the form

$$L = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j},$$

where:

- w_j is the weight for observation j. The software renormalizes the weights to sum to 1.
- $e_j = 1$ if the predicted class of observation j differs from its true class, and 0 otherwise.

In other words, the classification error is the proportion of observations misclassified by the classifier.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).

- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

See Also

`ClassificationPartitionedKernelECOC` | `fitcecoc`

Introduced in R2018b

kfoldLoss

Classification loss for observations not used in training

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

`L = kfoldLoss(CVMdl)` returns the cross-validated classification losses on page 33-3162 obtained by the cross-validated, binary, linear classification model `CVMdl`. That is, for every fold, `kfoldLoss` estimates the classification loss for observations that it holds out when it trains using all other observations.

`L` contains a classification loss for each regularization strength in the linear classification models that compose `CVMdl`.

`L = kfoldLoss(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, indicate which folds to use for the loss calculation or specify the classification-loss function.

Input Arguments

`CVMdl` — Cross-validated, binary, linear classification model

`ClassificationPartitionedLinear` model object

Cross-validated, binary, linear classification model, specified as a `ClassificationPartitionedLinear` model object. You can create a `ClassificationPartitionedLinear` model using `fitclinear` and specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldLoss applies the same data used to cross-validate the linear classification model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Folds` — Fold indices to use for classification-score prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for classification-score prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

LossFun — Loss function
`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. For linear classification models, logistic regression learners return posterior probabilities as classification scores by default, but SVM learners do not (see `predict`).
- Specify your own function using function handle notation.

  Let `n` be the number of observations in `X` and `K` be the number of distinct classes (`numel(Mdl.ClassNames)`, `Mdl` is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.
  - `Cost` is a *K*-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Mode` — Loss aggregation level
`'average'` (default) | `'individual'`

Loss aggregation level, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

Value	Description
`'average'`	Returns losses averaged over all folds
`'individual'`	Returns losses for each fold

Example: `'Mode','individual'`

Output Arguments

`L` — Cross-validated classification losses
numeric scalar | numeric vector | numeric matrix

Cross-validated classification losses on page 33-3162, returned as a numeric scalar, vector, or matrix. The interpretation of `L` depends on `LossFun`.

Let *R* be the number of regularizations strengths is the cross-validated models (stored in `numel(CVMdl.Trained{1}.Lambda)`) and *F* be the number of folds (stored in `CVMdl.KFold`).

- If `Mode` is `'average'`, then `L` is a 1-by-*R* vector. `L(j)` is the average classification loss over all folds of the cross-validated model that uses regularization strength *j*.
- Otherwise, `L` is an *F*-by-*R* matrix. `L(i,j)` is the classification loss for fold *i* of the cross-validated model that uses regularization strength *j*.

To estimate `L`, `kfoldLoss` uses the data that created `CVMdl` (see `X` and `Y`).

Examples

Estimate *k*-Fold Cross-Validation Classification Error

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Cross-validate a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinear` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the average of the out-of-fold, classification error rates.

```
ce = kfoldLoss(CVMdl)
```

```
ce = 7.6017e-04
```

Alternatively, you can obtain the per-fold classification error rates by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Specify Custom Classification Loss

Load the NLP data set. Preprocess the data as in "Estimate k-Fold Cross-Validation Classification Error" on page 33-3158, and transpose the predictor data.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Cross-validate a binary, linear classification model using 5-fold cross-validation. Optimize the objective function using SpaRSA. Specify that the predictor observations correspond to columns.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','KFold',5,...
    'ObservationsIn','columns');
CMdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 5-by-1 cell array holding a `ClassificationLinear` models that the software trained using the training set of each fold.

Create an anonymous function that measures linear loss, that is,

$$L = \frac{\sum_j - w_j y_j f_j}{\sum_j w_j}.$$

w_j is the weight for observation j, y_j is response j (-1 for the negative class, and 1 otherwise), and f_j is the raw classification score of observation j. Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `LossFun` name-value pair argument. Because the function does not use classification cost, use ~ to have `kfoldLoss` ignore its position.

```
linearloss = @(C,S,W,~)sum(-W.*sum(S.*C,2))/sum(W);
```

Estimate the average cross-validated classification loss using the linear loss function. Also, obtain the loss for each fold.

```
ce = kfoldLoss(CVMdl,'LossFun',linearloss)
```

```
ce = -8.0982
```

```
ceFold = kfoldLoss(CVMdl,'LossFun',linearloss,'Mode','individual')
```

```
ceFold = 5×1

   -8.3165
   -8.7633
   -7.4342
   -8.0423
   -7.9347
```

Find Good Lasso Penalty Using *k*-fold Classification Loss

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare test-sample classification error rates.

Load the NLP data set. Preprocess the data as in “Specify Custom Classification Loss” on page 33-3159.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Cross-validate binary, linear classification models using 5-fold cross-validation, and that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'KFold',5,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract a trained linear classification model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  ClassificationLinear
```

```
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl1` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated classification error.

```
ce = kfoldLoss(CVMdl);
```

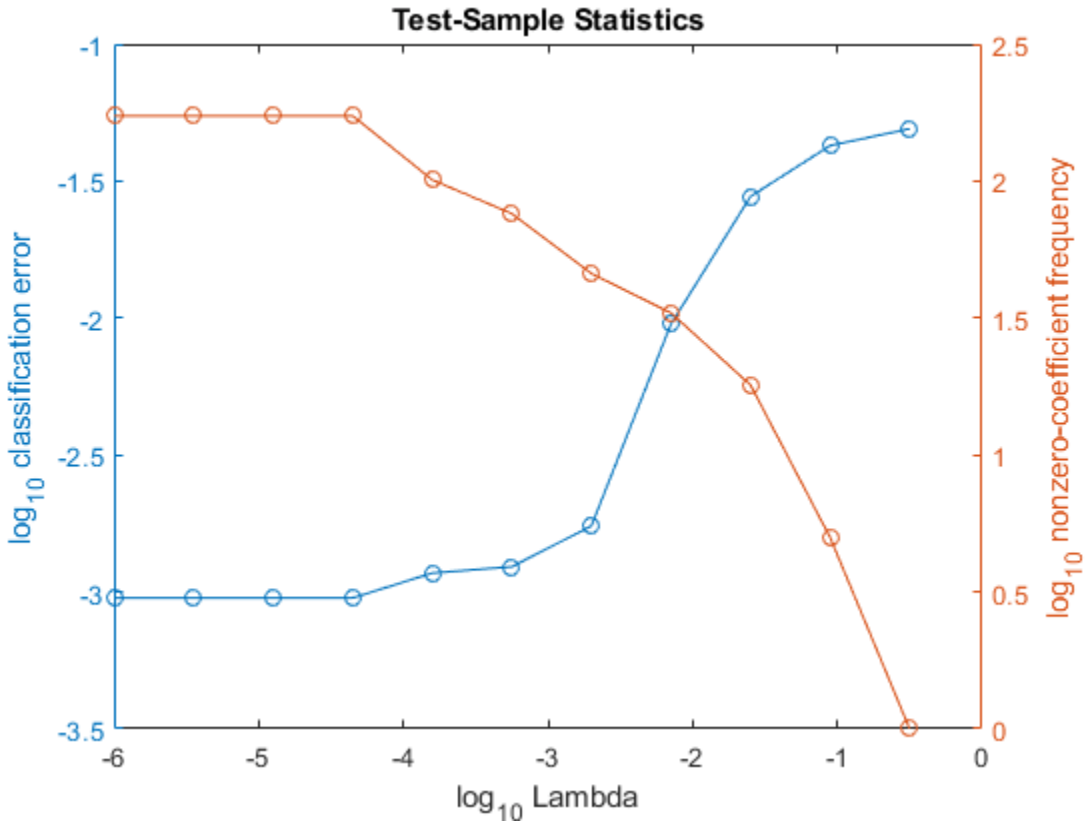
Because there are 11 regularization strengths, `ce` is a 1-by-11 vector of classification error rates.

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated, classification error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(ce),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} classification error')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
title('Test-Sample Statistics')
hold off
```



Choose the indexes of the regularization strength that balances predictor variable sparsity and low classification error. In this case, a value between 10^{-4} to 10^{-1} should suffice.

```
idxFinal = 7;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:

- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

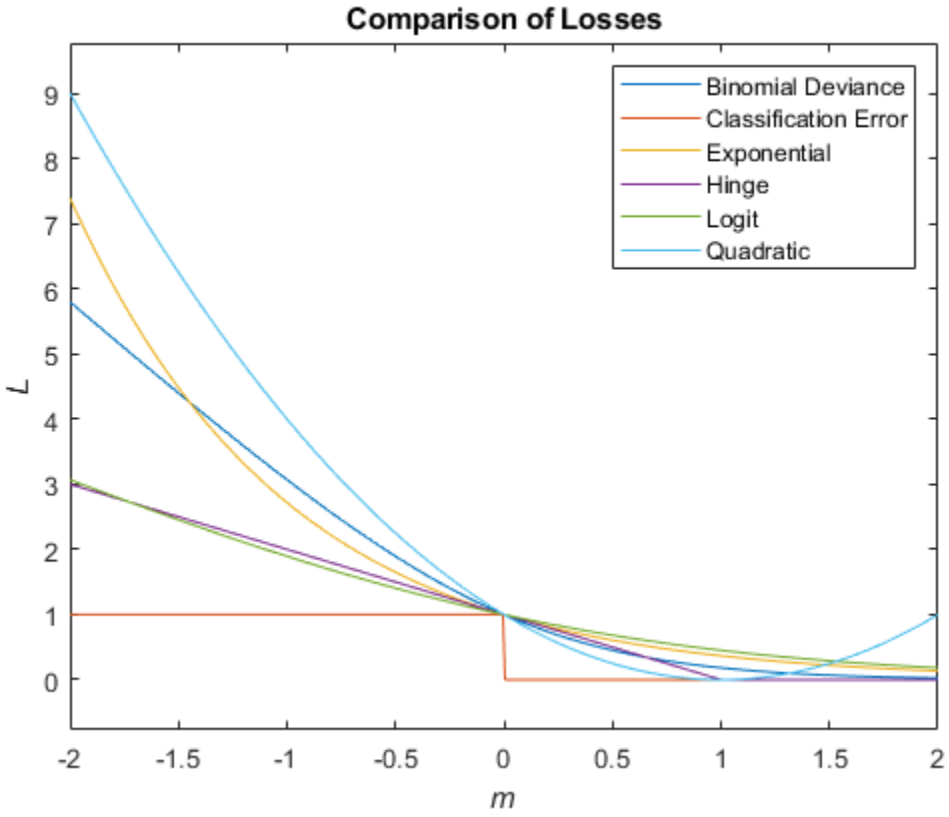
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.

Loss Function	Value of LossFun	Equation
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = (f(X_j)'C)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\hat{y}_j = \underset{k=1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`ClassificationLinear` | `ClassificationPartitionedLinear` | `kfoldPredict` | `loss`

Introduced in R2016a

kfoldLoss

Classification loss for observations not used in training

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

`L = kfoldLoss(CVMdl)` returns the cross-validated classification error on page 33-3175 rates estimated by the cross-validated, error-correcting output codes (ECOC) model composed of linear classification models `CVMdl`. That is, for every fold, `kfoldLoss` estimates the classification error rate for observations that it holds out when it trains using all other observations. `kfoldLoss` applies the same data used create `CVMdl` (see `fitcecoc`).

`L` contains a classification loss for each regularization strength in the linear classification models that compose `CVMdl`.

`L = kfoldLoss(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify a decoding scheme, which folds to use for the loss calculation, or verbosity level.

Input Arguments

`CVMdl` — Cross-validated, ECOC model composed of linear classification models

`ClassificationPartitionedLinearECOC` model object

Cross-validated, ECOC model composed of linear classification models, specified as a `ClassificationPartitionedLinearECOC` model object. You can create a `ClassificationPartitionedLinearECOC` model using `fitcecoc` and by:

1. Specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`
2. Setting the name-value pair argument `Learners` to `'linear'` or a linear classification model template returned by `templateLinear`

To obtain estimates, kfoldLoss applies the same data used to cross-validate the ECOC model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in, loss-function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {-1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes the binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, e.g., `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` should have this form

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

By default, if all binary learners are linear classification models using:

- SVM, then `BinaryLoss` is `'hinge'`
- Logistic regression, then `BinaryLoss` is `'quadratic'`

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme

`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Folds` — Fold indices to use for classification-score prediction
`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for classification-score prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`LossFun` — Loss function
`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a function handle or `'classiferror'`.

You can:

- Specify the built-in function `'classiferror'`, then the loss function is the classification error on page 33-3175.
- Specify your own function using function handle notation.

  For what follows, `n` is the number of observations in the training data (`CVMdl.NumObservations`) and `K` is the number of classes (`numel(CVMdl.ClassNames)`). Your function needs the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `CVMdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set every element of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of negated loss values for classes. Each row corresponds to an observation. The column order corresponds to the class order in `CVMdl.ClassNames`. `S` resembles the output argument `NegLoss` of `kfoldPredict`.
  - `W` is an `n`-by-`1` numeric vector of observation weights. If you pass `W`, the software normalizes its elements to sum to `1`.
  - `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of 0 for correct classification, and 1 for misclassification.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `function_handle` | `char` | `string`

`Mode` — Loss aggregation level
`'average'` (default) | `'individual'`

Loss aggregation level, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

Value	Description
`'average'`	Returns losses averaged over all folds
`'individual'`	Returns losses for each fold

Example: `'Mode','individual'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`L` — Cross-validated classification losses
numeric scalar | numeric vector | numeric matrix

Cross-validated classification losses on page 33-3175, returned as a numeric scalar, vector, or matrix. The interpretation of `L` depends on `LossFun`.

Let R be the number of regularizations strengths is the cross-validated models (`CVMdl.Trained{1}.BinaryLearners{1}.Lambda`) and F be the number of folds (stored in `CVMdl.KFold`).

- If `Mode` is `'average'`, then `L` is a 1-by-R vector. `L(`j`)` is the average classification loss over all folds of the cross-validated model that uses regularization strength j.
- Otherwise, `L` is a F-by-R matrix. `L(`i`,`j`)` is the classification loss for fold i of the cross-validated model that uses regularization strength j.

Examples

Estimate *k*-Fold Cross-Validation Classification Error

Load the NLP data set.

```
load nlpdata
```

X is a sparse matrix of predictor data, and Y is a categorical vector of class labels.

Cross-validate an ECOC model of linear classification models.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learner','linear','CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. By default, the software implements 10-fold cross validation.

Estimate the average of the out-of-fold classification error rates.

```
ce = kfoldLoss(CVMdl)
```

```
ce = 0.0958
```

Alternatively, you can obtain the per-fold classification error rates by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Specify Custom Classification Loss

Load the NLP data set. Transpose the predictor data.

```
load nlpdata
X = X';
```

For simplicity, use the label 'others' for all observations in Y that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Create a linear classification model template that specifies optimizing the objective function using SpaRSA.

```
t = templateLinear('Solver','sparsa');
```

Cross-validate an ECOC model of linear classification models using 5-fold cross-validation. Optimize the objective function using SpaRSA. Specify that the predictor observations correspond to columns.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'KFold',5,'ObservationsIn','columns');
CMdl1 = CVMdl.Trained{1}
```

```
CMdl1 =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [comm    dsp    simulink    others]
    ScoreTransform: 'none'
    BinaryLearners: {6x1 cell}
      CodingMatrix: [4x6 double]
```

```
  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. It contains the property `Trained`, which is a 5-by-1 cell array holding a `CompactClassificationECOC` models that the software trained using the training set of each fold.

Create a function that takes the minimal loss for each observation, and then averages the minimal losses across all observations. Because the function does not use the class-identifier matrix (`C`), observation weights (`W`), and classification cost (`Cost`), use `~` to have `kfoldLoss` ignore its their positions.

```
lossfun = @(~,S,~,~)mean(min(-S,[],2));
```

Estimate the average cross-validated classification loss using the minimal loss per observation function. Also, obtain the loss for each fold.

```
ce = kfoldLoss(CVMdl,'LossFun',lossfun)
```

```
ce = 0.0243
```

```
ceFold = kfoldLoss(CVMdl,'LossFun',lossfun,'Mode','individual')
```

```
ceFold = 5×1

    0.0244
    0.0255
    0.0248
    0.0240
    0.0226
```

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for an ECOC model composed of linear classification models that use logistic regression learners, implement 5-fold cross-validation.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

For simplicity, use the label 'others' for all observations in `Y` that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-7} through 10^{-2}.

```
Lambda = logspace(-7,-2,11);
```

Create a linear classification model template that specifies to use logistic regression learners, use lasso penalties with strengths in `Lambda`, train using SpaRSA, and lower the tolerance on the gradient of the objective function to `1e-8`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns.

```
X = X';
rng(10); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns','KFold',5);
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model.

Dissect `CVMdl`, and each model within it.

```
numECOCModels = numel(CVMdl.Trained)
```

```
numECOCModels = 5
```

```
ECOCMdl1 = CVMdl.Trained{1}
```

```
ECOCMdl1 =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [comm    dsp    simulink    others]
    ScoreTransform: 'none'
    BinaryLearners: {6×1 cell}
      CodingMatrix: [4×6 double]


  Properties, Methods
```

```
numCLModels = numel(ECOCMdl1.BinaryLearners)
```

```
numCLModels = 6
```

```
CLMdl1 = ECOCMdl1.BinaryLearners{1}
```

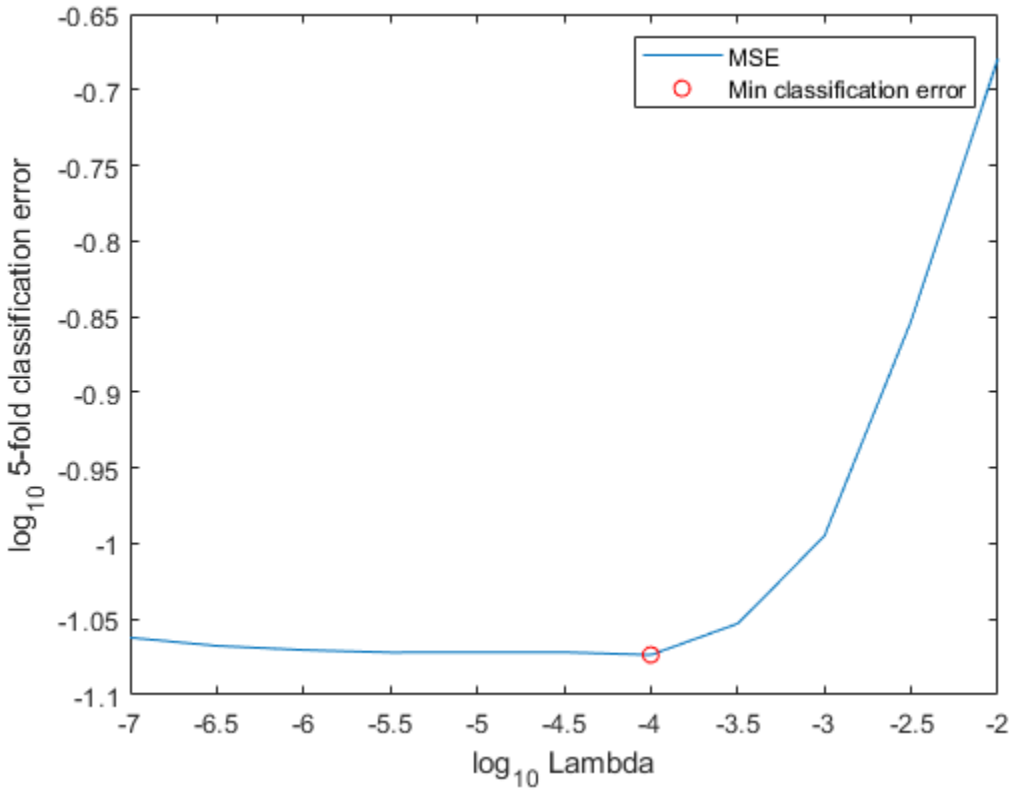
```
CLMdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [-1 1]
    ScoreTransform: 'logit'
              Beta: [34023×11 double]
              Bias: [-0.3169 -0.3169 -0.3168 -0.3168 -0.3168 -0.3167 -0.1725 -0.0805 -0.1762 -0.1
            Lambda: [1.0000e-07 3.1623e-07 1.0000e-06 3.1623e-06 1.0000e-05 3.1623e-05 1.0000e-04
           Learner: 'logistic'


  Properties, Methods
```

Because `fitcecoc` implements 5-fold cross-validation, `CVMdl` contains a 5-by-1 cell array of `CompactClassificationECOC` models that the software trains on each fold. The `BinaryLearners` property of each `CompactClassificationECOC` model contains the `ClassificationLinear`

models. The number of `ClassificationLinear` models within each compact ECOC model depends on the number of distinct labels and coding design. Because `Lambda` is a sequence of regularization strengths, you can think of `CLMdl1` as 11 models, one for each regularization strength in `Lambda`.

Determine how well the models generalize by plotting the averages of the 5-fold classification error for each regularization strength. Identify the regularization strength that minimizes the generalization error over the grid.

```
ce = kfoldLoss(CVMdl);
figure;
plot(log10(Lambda),log10(ce))
[~,minCEIdx] = min(ce);
minLambda = Lambda(minCEIdx);
hold on
plot(log10(minLambda),log10(ce(minCEIdx)),'ro');
ylabel('log_{10} 5-fold classification error')
xlabel('log_{10} Lambda')
legend('MSE','Min classification error')
hold off
```



Train an ECOC model composed of linear classification model using the entire data set, and specify the minimal regularization strength.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',minLambda,'GradientTolerance',1e-8);
MdlFinal = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns');
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Classification Error

The classification error is a binary classification error measure that has the form

$$L = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j},$$

where:

- w_j is the weight for observation j. The software renormalizes the weights to sum to 1.
- $e_j = 1$ if the predicted class of observation j differs from its true class, and 0 otherwise.

In other words, the classification error is the proportion of observations misclassified by the classifier.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationLinear` | `ClassificationPartitionedLinearECOC` | `fitcecoc` | `kfoldPredict` | `loss` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13

“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2016a

kfoldLoss

Package: `classreg.learning.partition`

Classification loss for cross-validated classification model

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

`L = kfoldLoss(CVMdl)` returns the classification loss obtained by the cross-validated classification model `CVMdl`. For every fold, `kfoldLoss` computes the classification loss for validation-fold observations using a classifier trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`L = kfoldLoss(CVMdl,Name,Value)` returns the classification loss with additional options specified by one or more name-value arguments. For example, you can specify a custom loss function.

Examples

Estimate Cross-Validated Classification Error

Load the `ionosphere` data set.

```
load ionosphere
```

Grow a classification tree.

```
tree = fitctree(X,Y);
```

Cross-validate the classification tree using 10-fold cross-validation.

```
cvtree = crossval(tree);
```

Estimate the cross-validated classification error.

```
L = kfoldLoss(cvtree)
```

```
L = 0.1083
```

Estimate Cross-Validated Classification Error

Load the `ionosphere` data set.

```
load ionosphere
```

Train a classification ensemble of 100 decision trees using AdaBoostM1. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
```

Cross-validate the ensemble using 10-fold cross-validation.

```
cvens = crossval(ens);
```

Estimate the cross-validated classification error.

```
L = kfoldLoss(cvens)
```

```
L = 0.0655
```

Find Optimal Number of Trees for GAM Using `kfoldLoss`

Train a cross-validated generalized additive model (GAM) with 10 folds. Then, use `kfoldLoss` to compute cumulative cross-validation classification errors (misclassification rate in decimal). Use the errors to determine the optimal number of trees per predictor (linear term for predictor) and the optimal number of trees per interaction term.

Alternatively, you can find optimal values of `fitcgam` name-value arguments by using the `bayesopt` function. For an example, see “Optimize Cross-Validated GAM Using bayesopt” on page 33-1693.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
rng('default') % For reproducibility
CVMdl = fitcgam(X,Y,'CrossVal','on','Interactions','all','MaxPValue',0.05);
```

If you specify `'Mode'` as `'cumulative'` for `kfoldLoss`, then the function returns cumulative errors, which are the average errors across all folds obtained using the same number of trees for each fold. Display the number of trees for each fold.

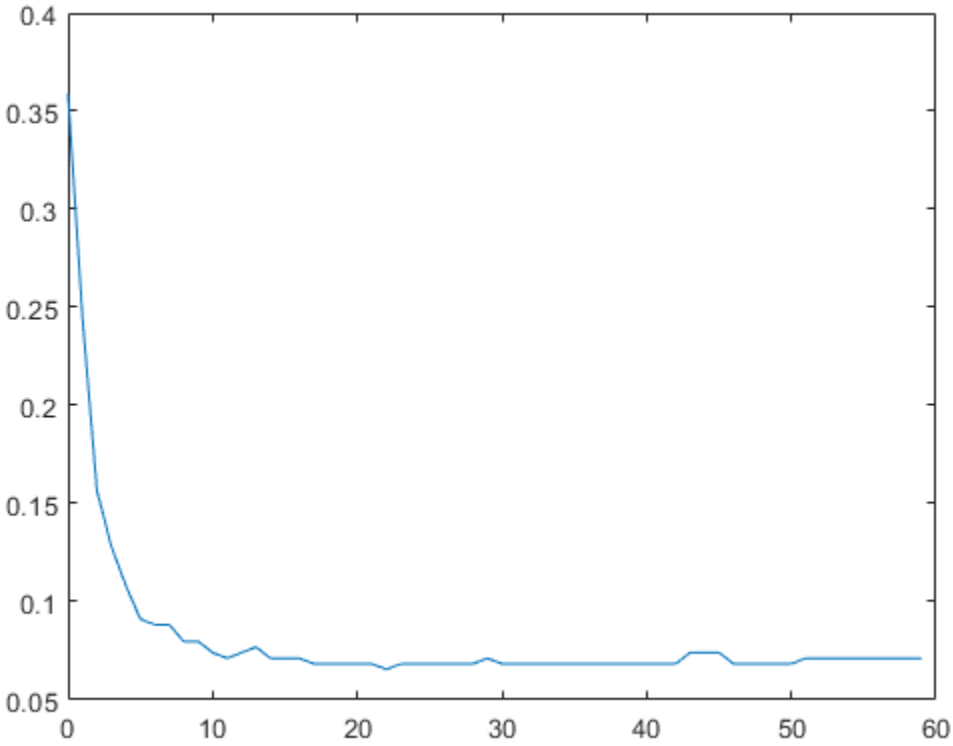
```
CVMdl.NumTrainedPerFold
```

```
ans = struct with fields:
      PredictorTrees: [65 64 59 61 60 66 65 62 64 61]
    InteractionTrees: [1 2 2 2 2 1 2 2 2 2]
```

`kfoldLoss` can compute cumulative errors using up to 59 predictor trees and one interaction tree.

Plot the cumulative, 10-fold cross-validated, classification error (misclassification rate in decimal). Specify `'IncludeInteractions'` as `false` to exclude interaction terms from the computation.

```
L_noInteractions = kfoldLoss(CVMdl,'Mode','cumulative','IncludeInteractions',false);
figure
plot(0:min(CVMdl.NumTrainedPerFold.PredictorTrees),L_noInteractions)
```



The first element of `L_noInteractions` is the average error over all folds obtained using only the intercept (constant) term. The (`J`+1)th element of `L_noInteractions` is the average error obtained using the intercept term and the first `J` predictor trees per linear term. Plotting the cumulative loss allows you to monitor how the error changes as the number of predictor trees in GAM increases.

Find the minimum error and the number of predictor trees used to achieve the minimum error.

```
[M,I] = min(L_noInteractions)
```

```
M = 0.0655
```

```
I = 23
```

The GAM achieves the minimum error when it includes 22 predictor trees.

Compute the cumulative classification error using both linear terms and interaction terms.

```
L = kfoldLoss(CVMdl,'Mode','cumulative')
```

```
L = 2×1

    0.0712
    0.0712
```

The first element of L is the average error over all folds obtained using the intercept (constant) term and all predictor trees per linear term. The second element of L is the average error obtained using the intercept term, all predictor trees per linear term, and one interaction tree per interaction term. The error does not decrease when interaction terms are added.

If you are satisfied with the error when the number of predictor trees is 22, you can create a predictive model by training the univariate GAM again and specifying 'NumTreesPerPredictor',22 without cross-validation.

Input Arguments

CVMdl — Cross-validated partitioned classifier

ClassificationPartitionedModel object | ClassificationPartitionedEnsemble object | ClassificationPartitionedGAM object

Cross-validated partitioned classifier, specified as a ClassificationPartitionedModel, ClassificationPartitionedEnsemble, or ClassificationPartitionedGAM object. You can create the object in two ways:

- Pass a trained classification model listed in the following table to its crossval object function.
- Train a classification model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Classification Model	Function
ClassificationDiscriminant	fitcdiscr
ClassificationEnsemble	fitcensemble
ClassificationGAM	fitcgam
ClassificationKNN	fitcknn
ClassificationNaiveBayes	fitcnb
ClassificationNeuralNetwork	fitcnet
ClassificationSVM	fitcsvm
ClassificationTree	fitctree

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example: kfoldLoss(CVMdl,'Folds',[1 2 3 5]) specifies to use the first, second, third, and fifth folds to compute the classification loss, but to exclude the fourth fold.

Folds — Fold indices to use

1:CVMdl.KFold (default) | positive integer vector

Fold indices to use, specified as a positive integer vector. The elements of Folds must be within the range from 1 to CVMdl.KFold.

The software uses only the folds specified in Folds.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `ClassificationPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Data Types: `logical`

`LossFun` — Loss function
`'classiferror'` | `'binodeviance'` | `'crossentropy'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as a built-in loss function name or a function handle. The default loss function depends on the model type of `CVMdl`.

- The default value is `'classiferror'` if the model type is an ensemble, generalized additive model, neural network, or support vector machine classifier.
- The default value is `'mincost'` if the model type is a discriminant analysis, *k*-nearest neighbor, naive Bayes, or tree classifier.

`'classiferror'` and `'mincost'` are equivalent when you use the default cost matrix. See “Algorithms” on page 33-3187 for more information.

- This table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'crossentropy'`	Cross-entropy loss (for neural networks only)
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. The `predict` and `kfoldPredict` functions of discriminant analysis, generalized additive model, *k*-nearest neighbor, naive Bayes, neural network, and tree classifiers return such scores by default.

  - For ensemble models that use `'Bag'` or `'Subspace'` methods, classification scores are posterior probabilities by default. For ensemble models that use `'AdaBoostM1'`,

'AdaBoostM2', GentleBoost, or 'LogitBoost' methods, you can use posterior probabilities as classification scores by specifying the double-logit score transform. For example, enter:

```
CVMdl.ScoreTransform = 'doublelogit';
```

For all other ensemble methods, the software does not support posterior probabilities as classification scores.

- For SVM models, you can specify to use posterior probabilities as classification scores by setting 'FitPosterior',true when you cross-validate the model using fitcsvm.

- Specify your own function using function handle notation.

  Suppose that n is the number of observations in the training data (CVMdl.NumObservations) and K is the number of classes (numel(CVMdl.ClassNames)). Your function must have the signature lossvalue = *lossfun*(C,S,W,Cost), where:

  - The output argument lossvalue is a scalar.
  - You specify the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in CVMdl.ClassNames.

    Construct C by setting C(p,q) = 1 if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in CVMdl.ClassNames. The input S resembles the output argument score of kfoldPredict.
  - W is an n-by-1 numeric vector of observation weights. If you pass W, the software normalizes its elements to sum to 1.
  - Cost is a K-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) – eye(K) specifies a cost of 0 for correct classification, and 1 for misclassification.

  Specify your function using 'LossFun',@*lossfun*.

For more details on loss functions, see “Classification Loss” on page 33-3184.

Example: 'LossFun','hinge'

Data Types: char | string | function_handle

Mode — Aggregation level for output

'average' (default) | 'individual' | 'cumulative'

Aggregation level for the output, specified as 'average', 'individual', or 'cumulative'.

Value	Description
'average'	The output is a scalar average over all folds.
'individual'	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Value	Description

`'cumulative'`	**Note** If you want to specify this value, `CVMdl` must be a `ClassificationPartitionedEnsemble` object or `ClassificationPartitionedGAM` object.
	
	• If `CVMdl` is `ClassificationPartitionedEnsemble`, then the output is a vector of length `min(CVMdl.NumTrainedPerFold)`. Each element `j` is an average over all folds that the function obtains by using ensembles trained with weak learners `1:j`.
	• If `CVMdl` is `ClassificationPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.
	  • If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained only the intercept (constant) term. The `(j + 1)`th element of `L` is an average obtained using the intercept term and the first `j` predictor trees per linear term.
	  • If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is an average obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

Example: `'Mode','individual'`

Output Arguments

L — Classification loss

numeric scalar | numeric column vector

Classification loss, returned as a numeric scalar or numeric column vector.

- If `Mode` is `'average'`, then `L` is the average classification loss over all folds.
- If `Mode` is `'individual'`, then `L` is a *k*-by-1 numeric column vector containing the classification loss for each fold, where *k* is the number of folds.
- If `Mode` is `'cumulative'` and `CVMdl` is `ClassificationPartitionedEnsemble`, then `L` is a `min(CVMdl.NumTrainedPerFold)`-by-1 numeric column vector. Each element `j` is the average classification loss over all folds that the function obtains by using ensembles trained with weak learners `1:j`.
- If `Mode` is `'cumulative'` and `CVMdl` is `ClassificationPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.
  - If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first

element of `L` is the average classification loss over all folds that is obtained using only the intercept (constant) term. The `(j + 1)`th element of `L` is the average loss obtained using the intercept term and the first `j` predictor trees per linear term.

- If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is the average classification loss over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is the average loss obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

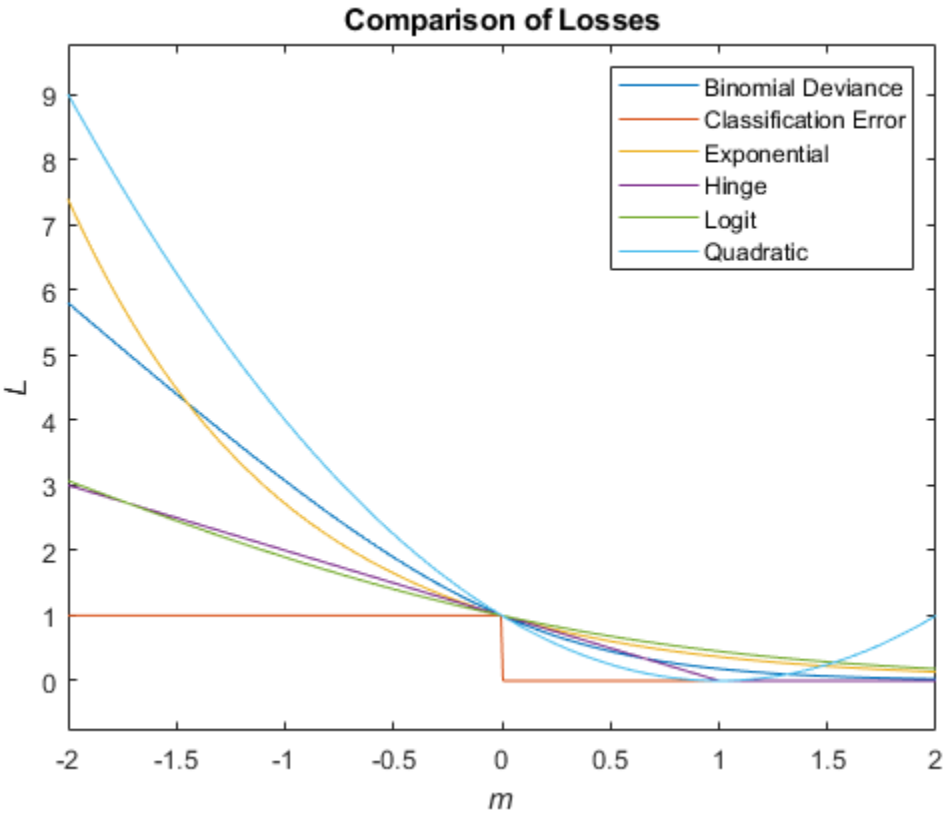
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Algorithms

`kfoldLoss` computes the classification loss as described in the corresponding `loss` object function. For a model-specific description, see the appropriate `loss` function reference page in the following table.

Model Type	`loss` Function
Discriminant analysis classifier	`loss`
Ensemble classifier	`loss`
Generalized additive model classifier	`loss`
k-nearest neighbor classifier	`loss`
Naive Bayes classifier	`loss`
Neural network classifier	`loss`
Support vector machine classifier	`loss`
Binary decision tree for multiclass classification	`loss`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports *k*-nearest neighbor and SVM model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `kfoldEdge` | `kfoldMargin` | `kfoldPredict` | `kfoldfun`

Topics

“Examine Quality of KNN Classifier” on page 18-28

Introduced in R2011a

kfoldLoss

Cross-validation loss of partitioned regression ensemble

Syntax

```
L = kfoldLoss(cvens)
L = kfoldLoss(cvens,Name,Value)
```

Description

`L = kfoldLoss(cvens)` returns the cross-validation loss of `cvens`.

`L = kfoldLoss(cvens,Name,Value)` returns cross-validation loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`cvens`

Object of class `RegressionPartitionedEnsemble`. Create `obj` with `fitrensemble` along with one of the cross-validation options: `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`. Alternatively, create `obj` from a regression ensemble with `crossval`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`folds`

Indices of folds ranging from `1` to `cvens.KFold`. Use only these folds for predictions.

Default: `1:cvens.KFold`

`lossfun`

Function handle for loss function, or `'mse'`, meaning mean squared error. If you pass a function handle `fun`, `loss` calls it as

```
fun(Y,Yfit,W)
```

where `Y`, `Yfit`, and `W` are numeric vectors of the same length.

- `Y` is the observed response.
- `Yfit` is the predicted response.
- `W` is the observation weights.

The returned value `fun(Y,Yfit,W)` should be a scalar.

Default: `'mse'`

`mode`

Method for computing cross-validation loss.

- `'average'` — `L` is a scalar value, the average loss over all folds.
- `'individual'` — `L` is a vector with one element per fold.
- `'cumulative'` — `L` is a vector with a length of minimum number of observations used for training in each fold. Each element in the vector `L` is obtained by taking the average of loss across all folds.

Default: `'average'`

Output Arguments

`L`

The loss (mean squared error) between the observations in a fold when compared against predictions made with an ensemble trained on the out-of-fold data. `L` can be a vector, and can mean different things, depending on the name-value pair settings.

Examples

Find Cross-Validation Loss for Regression Ensemble

Find the cross-validation loss for a regression ensemble of the `carsmall` data.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train an ensemble of regression trees.

```
rens = fitrensemble(X,MPG);
```

Create a cross-validated ensemble from `rens` and find the k-fold cross-validation loss.

```
rng(10,'twister') % For reproducibility
cvrens = crossval(rens);
L = kfoldLoss(cvrens)
```

```
L = 28.7114
```

See Also

`RegressionPartitionedEnsemble` | `kfoldPredict` | `loss`

kfoldLoss

Regression loss for observations not used in training

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

Description

`L = kfoldLoss(CVMdl)` returns the cross-validated mean squared error (MSE) obtained by the cross-validated, linear regression model `CVMdl`. That is, for every fold, `kfoldLoss` estimates the regression loss for observations that it holds out when it trains using all other observations.

`L` contains a regression loss for each regularization strength in the linear regression models that compose `CVMdl`.

`L = kfoldLoss(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, indicate which folds to use for the loss calculation or specify the regression-loss function.

Input Arguments

`CVMdl` — Cross-validated, linear regression model

`RegressionPartitionedLinear` model object

Cross-validated, linear regression model, specified as a `RegressionPartitionedLinear` model object. You can create a `RegressionPartitionedLinear` model using `fitrlinear` and specifying any of the one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldLoss applies the same data used to cross-validate the linear regression model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Folds` — Fold indices to use for response prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for response prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

LossFun — Loss function
'mse' (default) | 'epsiloninsensitive' | function handle

Loss function, specified as the comma-separated pair consisting of 'LossFun' and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar. Also, in the table, $f(x) = x\beta + b$.
  - β is a vector of p coefficients.
  - x is an observation from p predictor variables.
  - b is the scalar bias.

Value	Description
'epsiloninsensitive'	Epsilon-insensitive loss: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$
'mse'	MSE: $\ell[y, f(x)] = [y - f(x)]^2$

  'epsiloninsensitive' is appropriate for SVM learners only.
- Specify your own function using function handle notation.

  Let n be the number of observations in X. Your function must have this signature

  `lossvalue = lossfun(Y,Yhat,W)`

  where:

  - The output argument lossvalue is a scalar.
  - You choose the function name (*lossfun*).
  - Y is an n-dimensional vector of observed responses. kfoldLoss passes the input argument Y in for Y.
  - Yhat is an n-dimensional vector of predicted responses, which is similar to the output of predict.
  - W is an n-by-1 numeric vector of observation weights.

  Specify your function using 'LossFun',@*lossfun*.

Data Types: char | string | function_handle

Mode — Loss aggregation level
'average' (default) | 'individual'

Loss aggregation level, specified as the comma-separated pair consisting of 'Mode' and 'average' or 'individual'.

Value	Description
'average'	Returns losses averaged over all folds
'individual'	Returns losses for each fold

Example: 'Mode','individual'

Output Arguments

L — Cross-validated regression losses

numeric scalar | numeric vector | numeric matrix

Cross-validated regression losses, returned as a numeric scalar, vector, or matrix. The interpretation of `L` depends on `LossFun`.

Let R be the number of regularizations strengths is the cross-validated models (stored in `numel(CVMdl.Trained{1}.Lambda)`) and F be the number of folds (stored in `CVMdl.KFold`).

- If `Mode` is `'average'`, then `L` is a 1-by-R vector. `L(`j`)` is the average regression loss over all folds of the cross-validated model that uses regularization strength j.
- Otherwise, `L` is an F-by-R matrix. `L(`i`,`j`)` is the regression loss for fold i of the cross-validated model that uses regularization strength j.

To estimate `L`, `kfoldLoss` uses the data that created `CVMdl` (see `X` and `Y`).

Examples

Estimate *k*-Fold Mean Squared Error

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Cross-validate a linear regression model using SVM learners.

```
rng(1); % For reproducibility
CVMdl = fitrlinear(X,Y,'CrossVal','on');
```

`CVMdl` is a `RegressionPartitionedLinear` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the average of the test-sample MSEs.

```
mse = kfoldLoss(CVMdl)
```

```
mse = 0.1735
```

Alternatively, you can obtain the per-fold MSEs by specifying the name-value pair `'Mode','individual'` in `kfoldLoss`.

Specify Custom Regression Loss

Simulate data as in “Estimate k-Fold Mean Squared Error” on page 33-3193.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
X = X'; % Put observations in columns for faster training
```

Cross-validate a linear regression model using 10-fold cross-validation. Optimize the objective function using SpaRSA.

```
CVMdl = fitrlinear(X,Y,'CrossVal','on','ObservationsIn','columns',...
    'Solver','sparsa');
```

`CVMdl` is a `RegressionPartitionedLinear` model. It contains the property `Trained`, which is a 10-by-1 cell array holding `RegressionLinear` models that the software trained using the training set.

Create an anonymous function that measures Huber loss ($\delta = 1$), that is,

$$L = \frac{1}{\sum w_j} \sum_{j=1}^{n} w_j \ell_j ,$$

where

$$\ell_j = \begin{cases} 0.5\widehat{e}_j^2; & |\widehat{e}_j| \le 1 \\ |\widehat{e}_j| - 0.5; & |\widehat{e}_j| > 1 \end{cases}.$$

$\widehat{e}_j$ is the residual for observation j. Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `'LossFun'` name-value pair argument.

```
huberloss = @(Y,Yhat,W)sum(W.*((0.5*(abs(Y-Yhat)<=1).*(Y-Yhat).^2) + ...
    ((abs(Y-Yhat)>1).*abs(Y-Yhat)-0.5)))/sum(W);
```

Estimate the average Huber loss over the folds. Also, obtain the Huber loss for each fold.

```
mseAve = kfoldLoss(CVMdl,'LossFun',huberloss)
```

```
mseAve = -0.4447
```

```
mseFold = kfoldLoss(CVMdl,'LossFun',huberloss,'Mode','individual')
```

```
mseFold = 10×1

   -0.4454
   -0.4473
   -0.4452
   -0.4469
   -0.4434
```

```
   -0.4427
   -0.4471
   -0.4430
   -0.4438
   -0.4426
```

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for a linear regression model that uses least squares, implement 5-fold cross-validation.

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-5} through 10^{-1}.

```
Lambda = logspace(-5,-1,15);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Optimize the objective function using SpaRSA.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','KFold',5,'Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');

numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `RegressionPartitionedLinear` model. Because `fitrlinear` implements 5-fold cross-validation, `CVMdl` contains 5 `RegressionLinear` models that the software trains on each fold.

Display the first trained linear regression model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 = 
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x15 double]
```

```
                Bias: [1x15 double]
              Lambda: [1x15 double]
             Learner: 'leastsquares'


  Properties, Methods
```

`Mdl1` is a `RegressionLinear` model object. `fitrlinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 15 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated MSE.

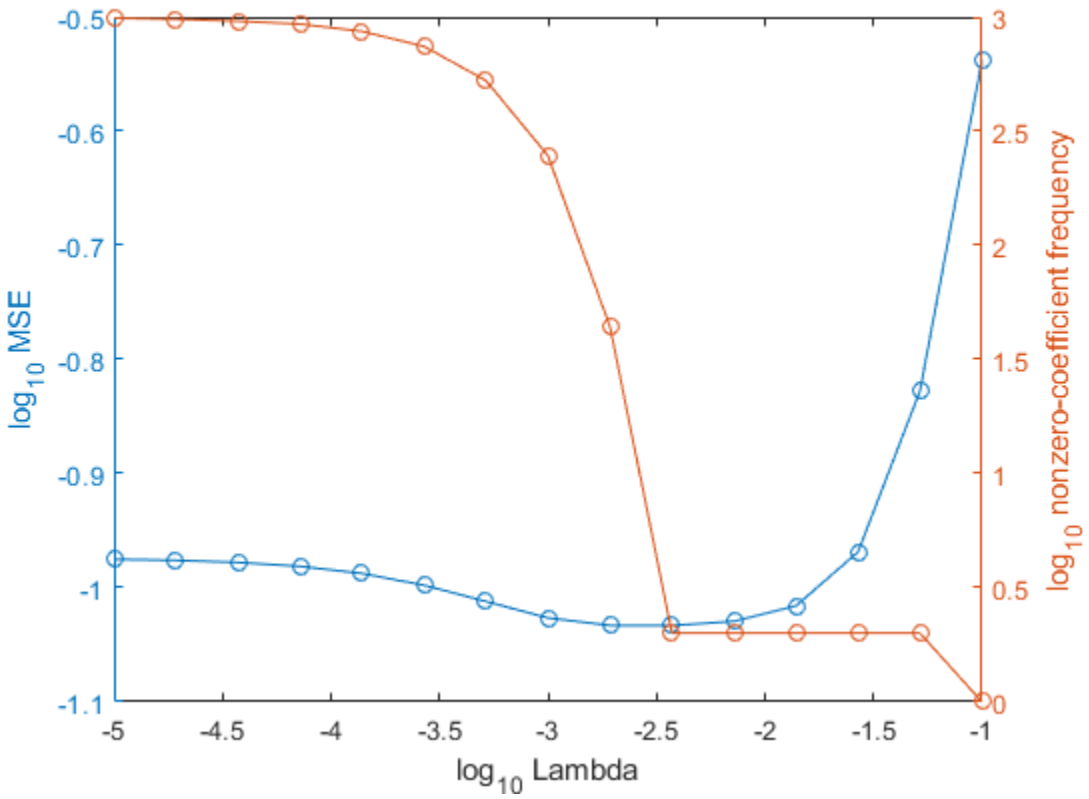
```
mse = kfoldLoss(CVMdl);
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. For each regularization strength, train a linear regression model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low MSE (for example, `Lambda(10)`).

```
idxFinal = 10;
```

Extract the model with corresponding to the minimal MSE.

```
MdlFinal = selectModels(Mdl,idxFinal)
```

```
MdlFinal =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0050
               Lambda: 0.0037
              Learner: 'leastsquares'


  Properties, Methods
```

```
idxNZCoeff = find(MdlFinal.Beta~=0)
```

```
idxNZCoeff = 2×1

   100
   200
```

```
EstCoeff = Mdl.Beta(idxNZCoeff)
```

```
EstCoeff = 2×1

    1.0051
    1.9965
```

`MdlFinal` is a `RegressionLinear` model with one regularization strength. The nonzero coefficients `EstCoeff` are close to the coefficients that simulated the data.

See Also

`RegressionLinear` | `RegressionPartitionedLinear` | `kfoldPredict` | `loss`

Introduced in R2016a

kfoldLoss

Package: `classreg.learning.partition`

Loss for cross-validated partitioned regression model

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

`L = kfoldLoss(CVMdl)` returns the loss (mean squared error) obtained by the cross-validated regression model `CVMdl`. For every fold, `kfoldLoss` computes the loss for validation-fold observations using a model trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`L = kfoldLoss(CVMdl,Name,Value)` returns the loss with additional options specified by one or more name-value arguments. For example, you can specify a custom loss function.

Examples

Find Cross-Validation Loss for Regression Ensemble

Find the cross-validation loss for a regression ensemble of the `carsmall` data.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train an ensemble of regression trees.

```
rens = fitrensemble(X,MPG);
```

Create a cross-validated ensemble from `rens` and find the k-fold cross-validation loss.

```
rng(10,'twister') % For reproducibility
cvrens = crossval(rens);
L = kfoldLoss(cvrens)
```

```
L = 28.7114
```

Display Individual Losses for Each Cross-Validation Fold

The mean squared error (MSE) is a measure of model quality. Examine the MSE for each fold of a cross-validated regression model.

Load the `carsmall` data set. Specify the predictor `X` and the response data `Y`.

```
load carsmall
X = [Cylinders Displacement Horsepower Weight];
Y = MPG;
```

Train a cross-validated regression tree model. By default, the software implements 10-fold cross-validation.

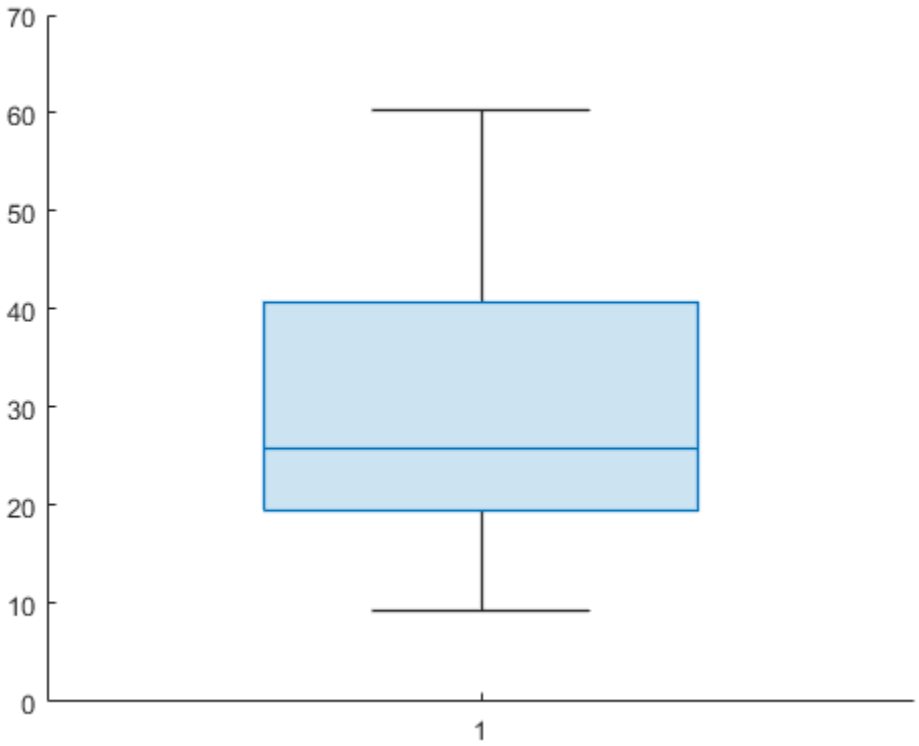
```
rng('default') % For reproducibility
CVMdl = fitrtree(X,Y,'CrossVal','on');
```

Compute the MSE for each fold. Visualize the distribution of the loss values by using a box plot. Notice that none of the values is an outlier.

```
losses = kfoldLoss(CVMdl,'Mode','individual')
```

```
losses = 10×1

   42.5072
   20.3995
   22.3737
   34.4255
   40.8005
   60.2755
   19.5562
    9.2060
   29.0788
   16.3386
```

```
boxchart(losses)
```



Find Optimal Number of Trees for GAM Using `kfoldLoss`

Train a cross-validated generalized additive model (GAM) with 10 folds. Then, use `kfoldLoss` to compute the cumulative cross-validation regression loss (mean squared errors). Use the errors to determine the optimal number of trees per predictor (linear term for predictor) and the optimal number of trees per interaction term.

Alternatively, you can find optimal values of `fitrgam` name-value arguments by using the `bayesopt` function. For an example, see “Optimize Cross-Validated GAM Using bayesopt” on page 33-2120.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, and `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`. Also, specify to include 5 interaction terms.

```
rng('default') % For reproducibility
CVMdl = fitrgam(tbl,'Systolic','CrossVal','on','Interactions',5);
```

If you specify `'Mode'` as `'cumulative'` for `kfoldLoss`, then the function returns cumulative errors, which are the average errors across all folds obtained using the same number of trees for each fold. Display the number of trees for each fold.

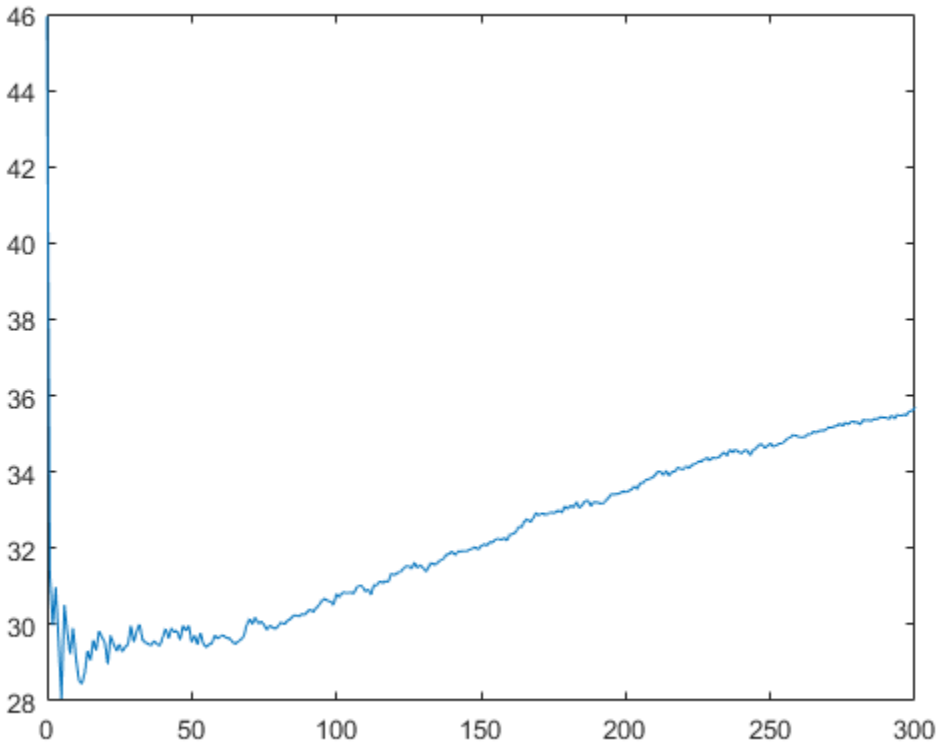
```
CVMdl.NumTrainedPerFold
```

```
ans = struct with fields:
      PredictorTrees: [300 300 300 300 300 300 300 300 300 300]
    InteractionTrees: [76 100 100 100 100 42 100 100 59 100]
```

`kfoldLoss` can compute cumulative errors using up to 300 predictor trees and 42 interaction trees.

Plot the cumulative, 10-fold cross-validated, mean squared errors. Specify `'IncludeInteractions'` as `false` to exclude interaction terms from the computation.

```
L_noInteractions = kfoldLoss(CVMdl,'Mode','cumulative','IncludeInteractions',false);
figure
plot(0:min(CVMdl.NumTrainedPerFold.PredictorTrees),L_noInteractions)
```



The first element of `L_noInteractions` is the average error over all folds obtained using only the intercept (constant) term. The (`J`+1)th element of `L_noInteractions` is the average error obtained using the intercept term and the first `J` predictor trees per linear term. Plotting the cumulative loss allows you to monitor how the error changes as the number of predictor trees in the GAM increases.

Find the minimum error and the number of predictor trees used to achieve the minimum error.

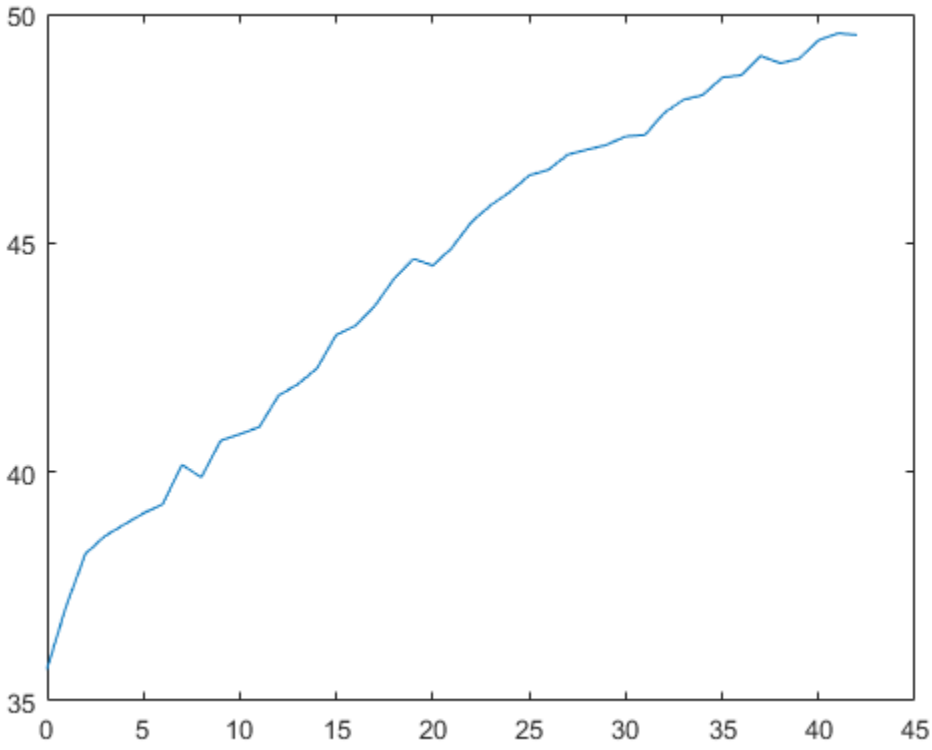
```
[M,I] = min(L_noInteractions)
```

```
M = 28.0506

I = 6
```

The GAM achieves the minimum error when it includes 5 predictor trees.

Compute the cumulative mean squared error using both linear terms and interaction terms.

```
L = kfoldLoss(CVMdl,'Mode','cumulative');
figure
plot(0:min(CVMdl.NumTrainedPerFold.InteractionTrees),L)
```



The first element of `L` is the average error over all folds obtained using the intercept (constant) term and all predictor trees per linear term. The (`J`+1)th element of `L` is the average error obtained using the intercept term, all predictor trees per linear term, and the first `J` interaction trees per interaction term. The plot shows that the error increases when interaction terms are added.

If you are satisfied with the error when the number of predictor trees is 5, you can create a predictive model by training the univariate GAM again and specifying `'NumTreesPerPredictor',5` without cross-validation.

Input Arguments

`CVMdl` — Cross-validated partitioned regression model

`RegressionPartitionedModel` object | `RegressionPartitionedEnsemble` object | `RegressionPartitionedGAM` object | `RegressionPartitionedSVM` object

Cross-validated partitioned regression model, specified as a `RegressionPartitionedModel`, `RegressionPartitionedEnsemble`, `RegressionPartitionedGAM`, or `RegressionPartitionedSVM` object. You can create the in two ways:

- Pass a trained regression model listed in the following table to its `crossval` object function.
- Train a regression model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Regression Model	Function
`RegressionEnsemble`	`fitrensemble`
`RegressionGAM`	`fitrgam`
`RegressionGP`	`fitrgp`
`RegressionNeuralNetwork`	`fitrnet`
`RegressionSVM`	`fitrsvm`
`RegressionTree`	`fitrtree`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldLoss(CVMdl,'Folds',[1 2 3 5])` specifies to use the first, second, third, and fifth folds to compute the mean squared error, but to exclude the fourth fold.

`Folds` — Fold indices to use
`1:CVMdl.KFold` (default) | positive integer vector

Fold indices to use, specified as a positive integer vector. The elements of `Folds` must be within the range from `1` to `CVMdl.KFold`.

The software uses only the folds specified in `Folds`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `RegressionPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`LossFun` — Loss function
`'mse'` (default) | function handle

Loss function, specified as `'mse'` or a function handle.

- Specify the built-in function `'mse'`. In this case, the loss function is the mean squared error.
- Specify your own function using function handle notation.

  Assume that *n* is the number of observations in the training data (`CVMdl.NumObservations`). Your function must have the signature `lossvalue = lossfun(Y,Yfit,W)`, where:

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `Y` is an *n*-by-1 numeric vector of observed responses.
  - `Yfit` is an *n*-by-1 numeric vector of predicted responses.
  - `W` is an *n*-by-1 numeric vector of observation weights.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Mode` — Aggregation level for output
`'average'` (default) | `'individual'` | `'cumulative'`

Aggregation level for the output, specified as `'average'`, `'individual'`, or `'cumulative'`.

Value	Description
`'average'`	The output is a scalar average over all folds.
`'individual'`	The output is a vector of length *k* containing one value per fold, where *k* is the number of folds.

Value	Description

`'cumulative'`	**Note** If you want to specify this value, `CVMdl` must be a `RegressionPartitionedEnsemble` object or `RegressionPartitionedGAM` object.
	
	• If `CVMdl` is `RegressionPartitionedEnsemble`, then the output is a vector of length `min(CVMdl.NumTrainedPerFold)`. Each element `j` is an average over all folds that the function obtains by using ensembles trained with weak learners `1:j`.
	• If `CVMdl` is `RegressionPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.
	  • If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained using only the intercept (constant) term. The `(j + 1)`th element of `L` is an average obtained using the intercept term and the first `j` predictor trees per linear term.
	  • If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is an average over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is an average obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

Example: `'Mode','individual'`

Output Arguments

L — Loss

numeric scalar | numeric column vector

Loss, returned as a numeric scalar or numeric column vector.

By default, the loss is the mean squared error between the validation-fold observations and the predictions made with a regression model trained on the training-fold observations.

- If `Mode` is `'average'`, then `L` is the average loss over all folds.
- If `Mode` is `'individual'`, then `L` is a *k*-by-1 numeric column vector containing the loss for each fold, where *k* is the number of folds.
- If `Mode` is `'cumulative'` and `CVMdl` is `RegressionPartitionedEnsemble`, then `L` is a `min(CVMdl.NumTrainedPerFold)`-by-1 numeric column vector. Each element `j` is the average loss over all folds that the function obtains using ensembles trained with weak learners `1:j`.
- If `Mode` is `'cumulative'` and `CVMdl` is `RegressionPartitionedGAM`, then the output value depends on the `IncludeInteractions` value.

- If `IncludeInteractions` is `false`, then `L` is a `(1 + min(NumTrainedPerFold.PredictorTrees))`-by-1 numeric column vector. The first element of `L` is the average loss over all folds that is obtained using only the intercept (constant) term. The `(j + 1)`th element of `L` is the average loss obtained using the intercept term and the first `j` predictor trees per linear term.
- If `IncludeInteractions` is `true`, then `L` is a `(1 + min(NumTrainedPerFold.InteractionTrees))`-by-1 numeric column vector. The first element of `L` is the average loss over all folds that is obtained using the intercept (constant) term and all predictor trees per linear term. The `(j + 1)`th element of `L` is the average loss obtained using the intercept term, all predictor trees per linear term, and the first `j` interaction trees per interaction term.

Alternative Functionality

If you want to compute the cross-validated loss of a tree model, you can avoid constructing a `RegressionPartitionedModel` object by calling `cvloss`. Creating a cross-validated tree object can save you time if you plan to examine it more than once.

See Also

`RegressionPartitionedEnsemble` | `RegressionPartitionedGAM` | `RegressionPartitionedModel` | `RegressionPartitionedSVM` | `kfoldPredict`

Introduced in R2011a

kfoldMargin

Package: `classreg.learning.partition`

Classification margins for cross-validated ECOC model

Syntax

```
margin = kfoldMargin(CVMdl)
margin = kfoldMargin(CVMdl,Name,Value)
```

Description

`margin = kfoldMargin(CVMdl)` returns classification margins on page 33-3213 obtained by the cross-validated ECOC model (`ClassificationPartitionedECOC`) `CVMdl`. For every fold, `kfoldMargin` computes classification margins for validation-fold observations using an ECOC model trained on training-fold observations. `CVMdl.X` contains both sets of observations.

`margin = kfoldMargin(CVMdl,Name,Value)` returns classification margins with additional options specified by one or more name-value pair arguments. For example, specify the binary learner loss function, decoding scheme, or verbosity level.

Examples

Estimate *k*-Fold Cross-Validation Margins

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

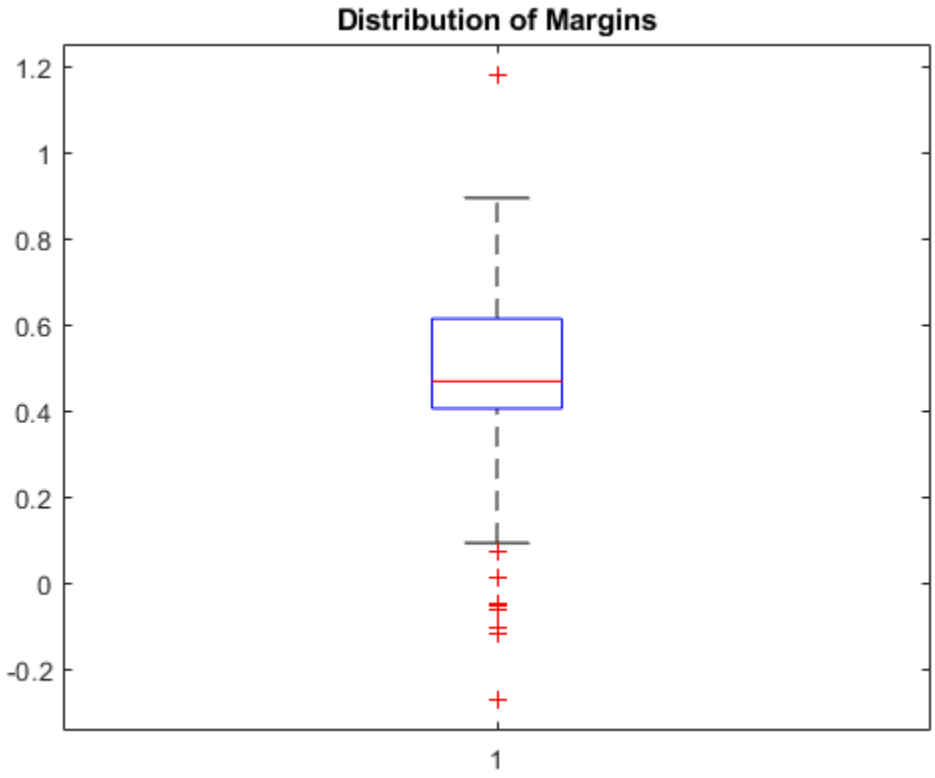
Train and cross-validate an ECOC model using support vector machine (SVM) binary classifiers. Standardize the predictor data using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

Estimate the margins for validation-fold observations. Display the distribution of the margins using a boxplot.

```
margin = kfoldMargin(CVMdl);

boxplot(margin)
title('Distribution of Margins')
```



Select ECOC Model Features by Comparing Cross-Validation Margins

One way to perform feature selection is to compare cross-validation margins from multiple models. Based solely on this criterion, the classifier with the greatest margins is the best classifier.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Define the following two data sets.

- `fullX` contains all the predictors.
- `partX` contains the petal dimensions.

```
fullX = X;
partX = X(:,3:4);
```

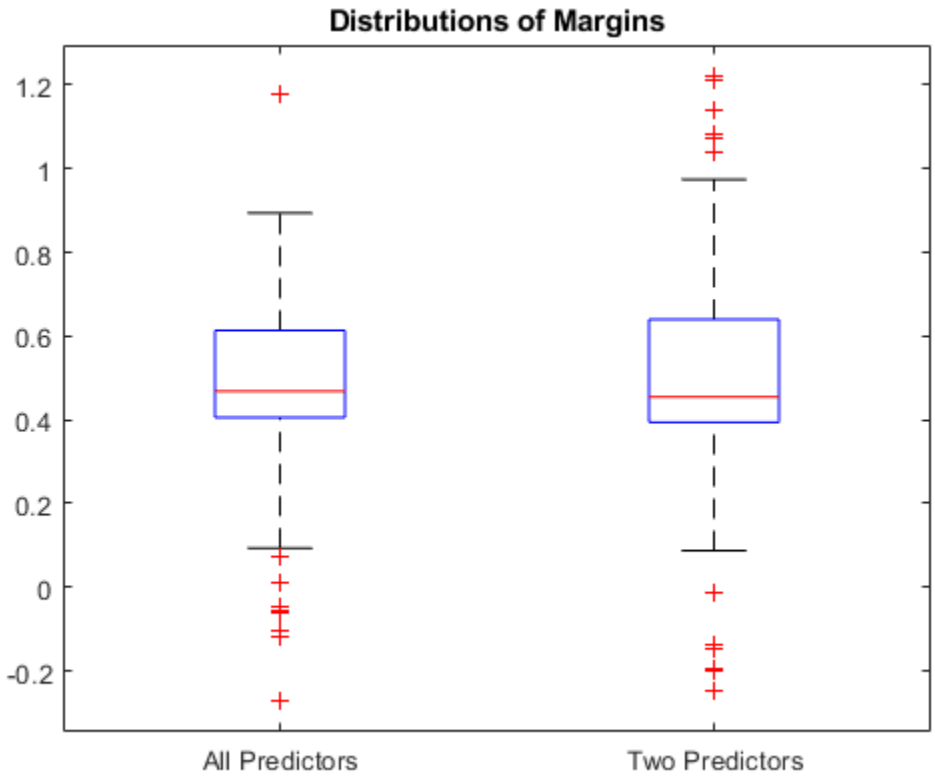
For each predictor set, train and cross-validate an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(fullX,Y,'CrossVal','on','Learners',t,...
    'ClassNames',classOrder);
PCVMdl = fitcecoc(partX,Y,'CrossVal','on','Learners',t,...
    'ClassNames',classOrder);
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedECOC` models. By default, the software implements 10-fold cross-validation.

Estimate the margins for each classifier. Use loss-based decoding for aggregating the binary learner results. For each model, display the distribution of the margins using a boxplot.

```
fullMargins = kfoldMargin(CVMdl,'Decoding','lossbased');
partMargins = kfoldMargin(PCVMdl,'Decoding','lossbased');

boxplot([fullMargins partMargins],'Labels',{'All Predictors','Two Predictors'})
title('Distributions of Margins')
```



The margin distributions are approximately the same.

Input Arguments

`CVMdl` — Cross-validated ECOC model

`ClassificationPartitionedECOC` model

Cross-validated ECOC model, specified as a `ClassificationPartitionedECOC` model. You can create a `ClassificationPartitionedECOC` model in two ways:

- Pass a trained ECOC model (`ClassificationECOC`) to `crossval`.
- Train an ECOC model using `fitcecoc` and specify any one of these cross-validation name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldMargin(CVMdl,'Verbose',1)` specifies to display diagnostic messages in the Command Window.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.

- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`margin` — Classification margins

numeric vector

Classification margins on page 33-3213, returned as a numeric vector. `margin` is an *n*-by-1 vector, where each row is the margin of the corresponding observation and *n* is the number of observations (`size(CVMdl.X,1)`).

More About

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} \left|m_{kj}\right| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|}.$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `fitcecoc` | `kfoldEdge` | `kfoldPredict` | `margin` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

kfoldMargin

Package: `classreg.learning.partition`

Classification margins for cross-validated kernel classification model

Syntax

```
margin = kfoldMargin(CVMdl)
```

Description

`margin = kfoldMargin(CVMdl)` returns the classification margins on page 33-3219 obtained by the cross-validated, binary kernel model (`ClassificationPartitionedKernel`) `CVMdl`. For every fold, `kfoldMargin` computes the classification margins for validation-fold observations using a model trained on training-fold observations.

Examples

Estimate *k*-Fold Cross-Validation Margins

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled as either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data.

```
CVMdl = fitckernel(X,Y,'Crossval','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the classification margins for validation-fold observations.

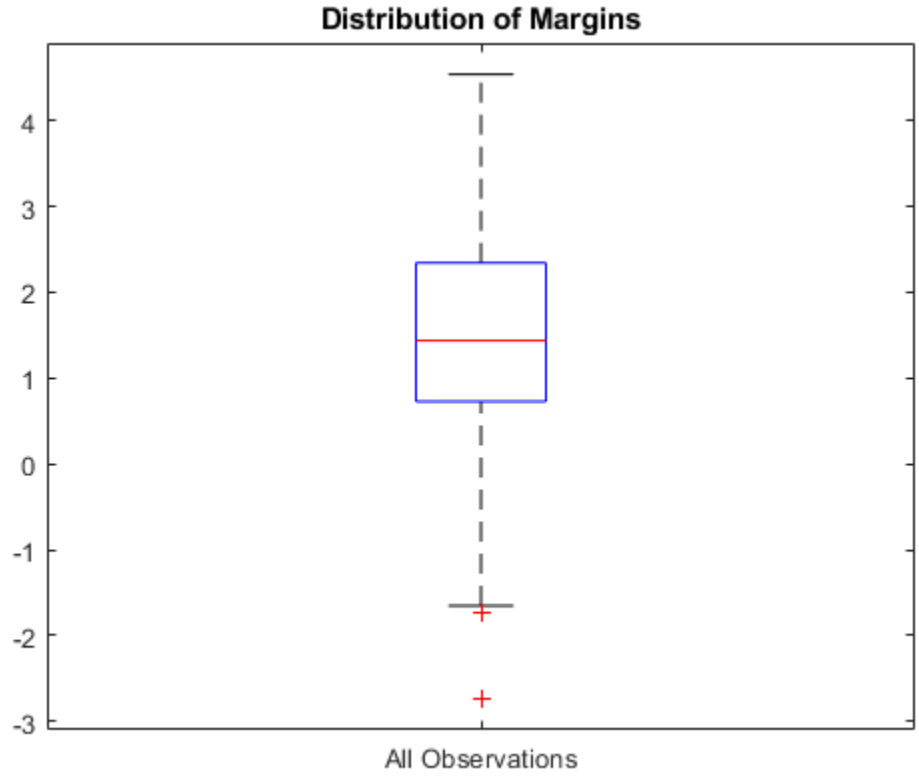
```
m = kfoldMargin(CVMdl);
size(m)
```

```
ans = 1×2

   351     1
```

`m` is a 351-by-1 vector. `m(j)` is the classification margin for observation `j`.

Plot the k-fold margins using a boxplot.

```
boxplot(m,'Labels','All Observations')
title('Distribution of Margins')
```



Feature Selection Using *k*-Fold Margins

Perform feature selection by comparing *k*-fold margins from multiple models. Based solely on this criterion, the classifier with the greatest margins is the best classifier.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Randomly choose 10% of the predictor variables.

```
rng(1); % For reproducibility
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.1*p));
```

Cross-validate two binary kernel classification models: one that uses all of the predictors, and one that uses 10% of the predictors.

```
CVMdl = fitckernel(X,Y,'CrossVal','on');
PCVMdl = fitckernel(X(:,idxPart),Y,'CrossVal','on');
```

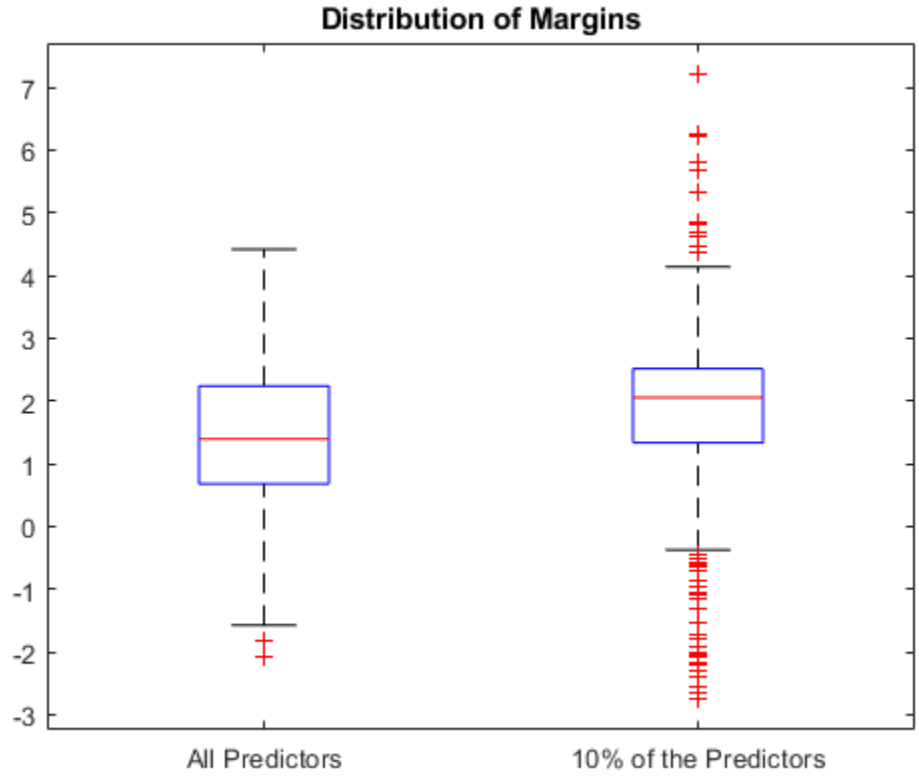
`CVMdl` and `PCVMdl` are `ClassificationPartitionedKernel` models. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the *k*-fold margins for each classifier.

```
fullMargins = kfoldMargin(CVMdl);
partMargins = kfoldMargin(PCVMdl);
```

Plot the distribution of the margin sets using box plots.

```
boxplot([fullMargins partMargins], ...
    'Labels',{'All Predictors','10% of the Predictors'});
title('Distribution of Margins')
```



The quartiles of the `PCVMdl` margin distribution are situated higher than the quartiles of the `CVMdl` margin distribution, indicating that the `PCVMdl` model is the better classifier.

Input Arguments

`CVMdl` — Cross-validated, binary kernel classification model

`ClassificationPartitionedKernel` model object

Cross-validated, binary kernel classification model, specified as a `ClassificationPartitionedKernel` model object. You can create a `ClassificationPartitionedKernel` model by using `fitckernel` and specifying any one of the cross-validation name-value pair arguments.

To obtain estimates, `kfoldMargin` applies the same data used to cross-validate the kernel classification model (`X` and `Y`).

Output Arguments

`margin` — Classification margins

numeric vector

Classification margins on page 33-3219, returned as a numeric vector. `margin` is an *n*-by-1 vector, where each row is the margin of the corresponding observation and *n* is the number of observations (`size(CVMdl.Y,1)`).

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationPartitionedKernel` | `fitckernel`

Introduced in R2018b

kfoldMargin

Package: `classreg.learning.partition`

Classification margins for cross-validated kernel ECOC model

Syntax

```
margin = kfoldMargin(CVMdl)
margin = kfoldMargin(CVMdl,Name,Value)
```

Description

`margin = kfoldMargin(CVMdl)` returns the classification margins on page 33-3226 obtained by the cross-validated kernel ECOC model (`ClassificationPartitionedKernelECOC`) `CVMdl`. For every fold, `kfoldMargin` computes the classification margins for validation-fold observations using a model trained on training-fold observations.

`margin = kfoldMargin(CVMdl,Name,Value)` returns classification margins with additional options specified by one or more name-value pair arguments. For example, specify the binary learner loss function, decoding scheme, or verbosity level.

Examples

Estimate *k*-Fold Cross-Validation Margins

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model composed of kernel binary learners.

```
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the classification margins for validation-fold observations.

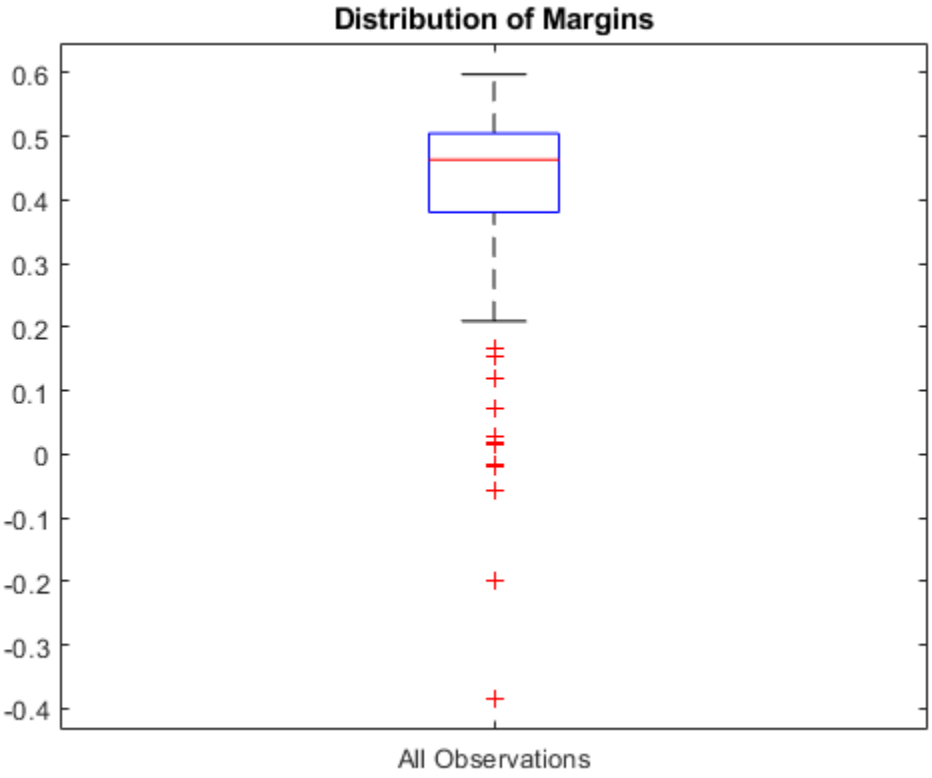
```
m = kfoldMargin(CVMdl);
size(m)
```

```
ans = 1×2

   150     1
```

`m` is a 150-by-1 vector. `m(j)` is the classification margin for observation `j`.

Plot the k-fold margins using a boxplot.

```
boxplot(m,'Labels','All Observations')
title('Distribution of Margins')
```



Feature Selection Using *k*-Fold Margins

Perform feature selection by comparing *k*-fold margins from multiple models. Based solely on this criterion, the classifier with the greatest margins is the best classifier.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Randomly choose half of the predictor variables.

```
rng(1); % For reproducibility
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.5*p));
```

Cross-validate two ECOC models composed of kernel classification models: one that uses all of the predictors, and one that uses half of the predictors.

```
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on');
PCVMdl = fitcecoc(X(:,idxPart),Y,'Learners','kernel','CrossVal','on');
```

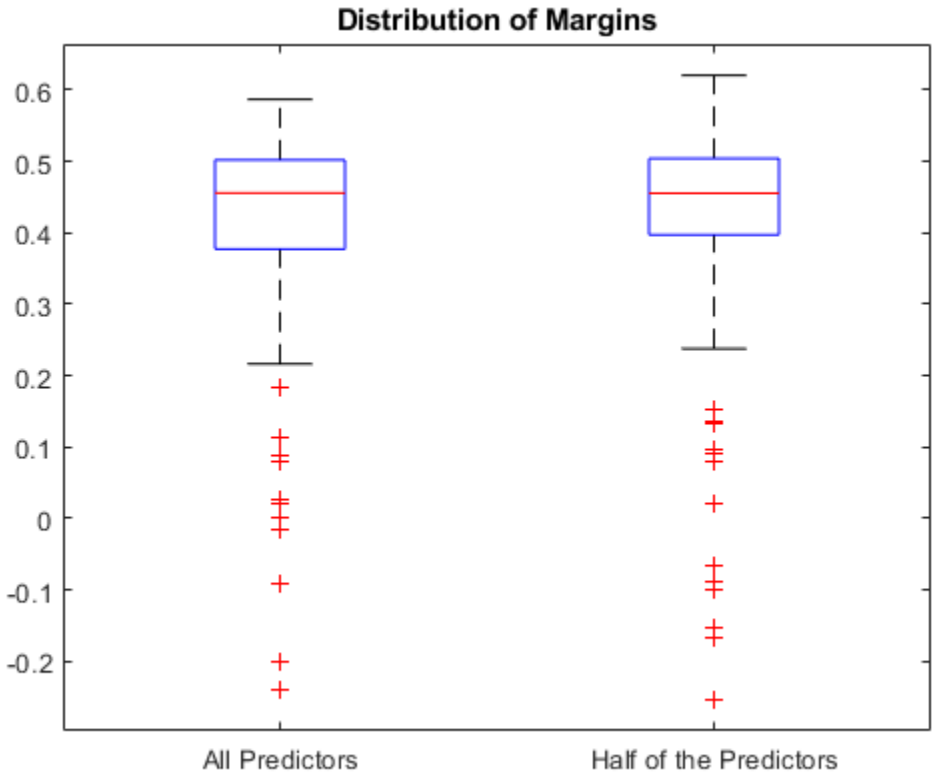
`CVMdl` and `PCVMdl` are `ClassificationPartitionedKernelECOC` models. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Estimate the *k*-fold margins for each classifier.

```
fullMargins = kfoldMargin(CVMdl);
partMargins = kfoldMargin(PCVMdl);
```

Plot the distribution of the margin sets using box plots.

```
boxplot([fullMargins partMargins], ...
    'Labels',{'All Predictors','Half of the Predictors'});
title('Distribution of Margins')
```



The `PCVMdl` margin distribution is similar to the `CVMdl` margin distribution.

Input Arguments

`CVMdl` — Cross-validated kernel ECOC model

`ClassificationPartitionedKernelECOC` model

Cross-validated kernel ECOC model, specified as a `ClassificationPartitionedKernelECOC` model. You can create a `ClassificationPartitionedKernelECOC` model by training an ECOC model using `fitcecoc` and specifying these name-value pair arguments:

- `'Learners'`– Set the value to `'kernel'`, a template object returned by `templateKernel`, or a cell array of such template objects.
- One of the arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldMargin(CVMdl,'Verbose',1)` specifies to display diagnostic messages in the Command Window.

`BinaryLoss` — Binary learner loss function
`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.
- For a custom binary loss function, for example, `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

By default, if all binary learners are kernel classification models using SVM, then `BinaryLoss` is `'hinge'`. If all binary learners are kernel classification models using logistic regression, then `BinaryLoss` is `'quadratic'`.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3226.

Example: `'Decoding','lossbased'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`margin` — Classification margins
numeric vector

Classification margins on page 33-3226, returned as a numeric vector. `margin` is an *n*-by-1 vector, where each row is the margin of the corresponding observation and *n* is the number of observations (`size(CVMdl.Y,1)`).

More About

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

See Also

`ClassificationPartitionedKernelECOC` | `fitcecoc`

Introduced in R2018b

kfoldMargin

Classification margins for observations not used in training

Syntax

```
m = kfoldMargin(CVMdl)
```

Description

`m = kfoldMargin(CVMdl)` returns the cross-validated classification margins on page 33-3235 obtained by the cross-validated, binary, linear classification model `CVMdl`. That is, for every fold, `kfoldMargin` estimates the classification margins for observations that it holds out when it trains using all other observations.

`m` contains classification margins for each regularization strength in the linear classification models that comprise `CVMdl`.

Input Arguments

`CVMdl` — Cross-validated, binary, linear classification model

`ClassificationPartitionedLinear` model object

Cross-validated, binary, linear classification model, specified as a `ClassificationPartitionedLinear` model object. You can create a `ClassificationPartitionedLinear` model using `fitclinear` and specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldMargin applies the same data used to cross-validate the linear classification model (`X` and `Y`).

Output Arguments

`m` — Cross-validated classification margins

numeric vector | numeric matrix

Cross-validated classification margins on page 33-3235, returned as a numeric vector or matrix.

`m` is *n*-by-*L*, where *n* is the number of observations in the data that created `CVMdl` (see `X` and `Y`) and *L* is the number of regularization strengths in `CVMdl` (that is, `numel(CVMdl.Trained{1}.Lambda)`).

`m(i,j)` is the cross-validated classification margin of observation *i* using the linear classification model that has regularization strength `CVMdl.Trained{1}.Lambda(j)`.

Data Types: `single` | `double`

Examples

Estimate *k*-Fold Cross-Validation Margins

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Cross-validate a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinear` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the cross-validated margins.

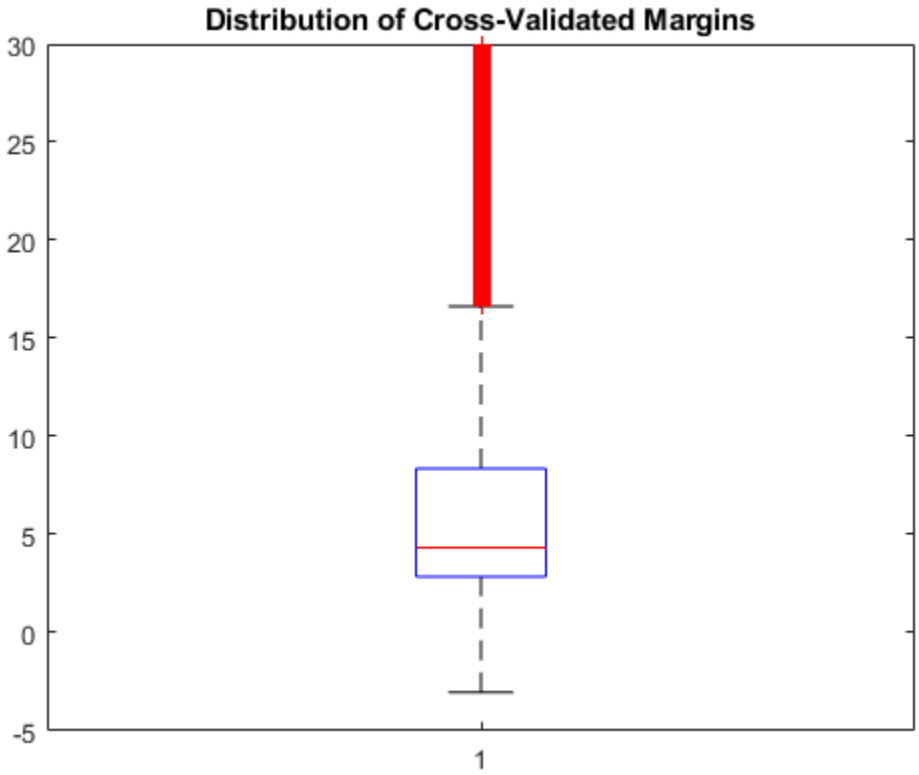
```
m = kfoldMargin(CVMdl);
size(m)
```

```
ans = 1×2

       31572           1
```

`m` is a 31572-by-1 vector. `m(j)` is the average of the out-of-fold margins for observation `j`.

Plot the *k*-fold margins using box plots.

```
figure;
boxplot(m);
h = gca;
h.YLim = [-5 30];
title('Distribution of Cross-Validated Margins')
```



Feature Selection Using *k*-fold Margins

One way to perform feature selection is to compare *k*-fold margins from multiple models. Based solely on this criterion, the classifier with the larger margins is the better classifier.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Margins” on page 33-3229.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create these two data sets:

- `fullX` contains all predictors.
- `partX` contains 1/2 of the predictors chosen at random.

```
rng(1); % For reproducibility
p = size(X,1); % Number of predictors
halfPredIdx = randsample(p,ceil(0.5*p));
fullX = X;
partX = X(halfPredIdx,:);
```

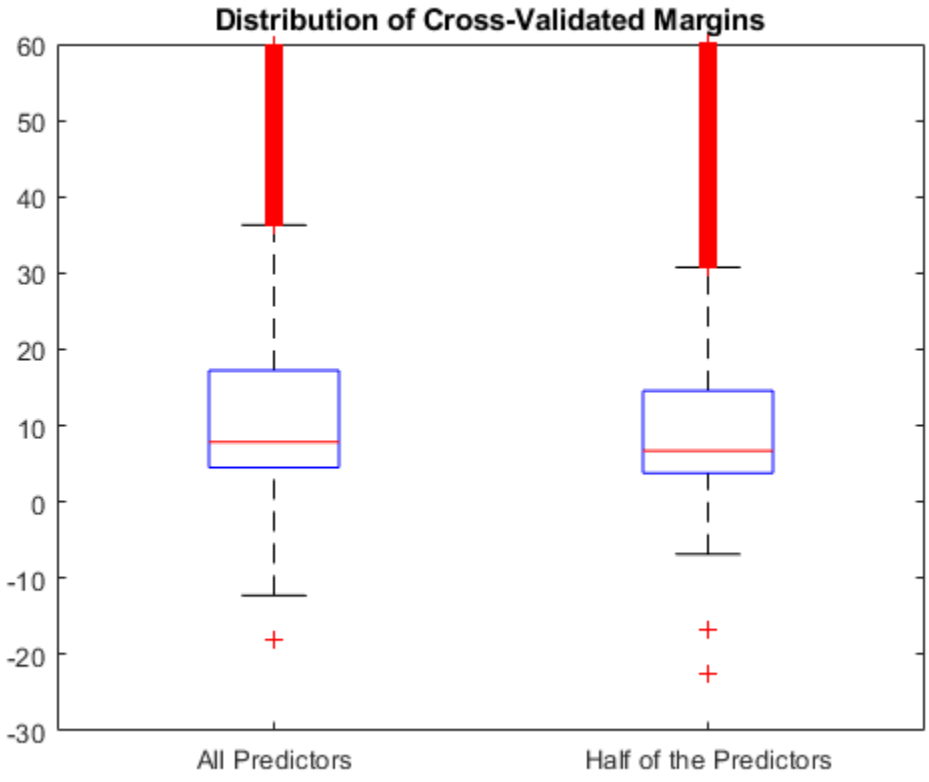
Cross-validate two binary, linear classification models: one that uses the all of the predictors and one that uses half of the predictors. Optimize the objective function using SpaRSA, and indicate that observations correspond to columns.

```
CVMdl = fitclinear(fullX,Ystats,'CrossVal','on','Solver','sparsa',...
    'ObservationsIn','columns');
PCVMdl = fitclinear(partX,Ystats,'CrossVal','on','Solver','sparsa',...
    'ObservationsIn','columns');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinear` models.

Estimate the *k*-fold margins for each classifier. Plot the distribution of the *k*-fold margins sets using box plots.

```
fullMargins = kfoldMargin(CVMdl);
partMargins = kfoldMargin(PCVMdl);

figure;
boxplot([fullMargins partMargins],'Labels',...
    {'All Predictors','Half of the Predictors'});
h = gca;
h.YLim = [-30 60];
title('Distribution of Cross-Validated Margins')
```



The distributions of the margins of the two classifiers are similar.

Find Good Lasso Penalty Using *k*-fold Margins

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare distributions of *k*-fold margins.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Margins” on page 33-3229.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,11);
```

Cross-validate a binary, linear classification model using 5-fold cross-validation and that uses each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns','KFold',5, ...
    'Learner','logistic','Solver','sparsa','Regularization','lasso', ...
    'Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinear` model. Because `fitclinear` implements 5-fold cross-validation, `CVMdl` contains 5 `ClassificationLinear` models that the software trains on each fold.

Estimate the *k*-fold margins for each regularization strength.

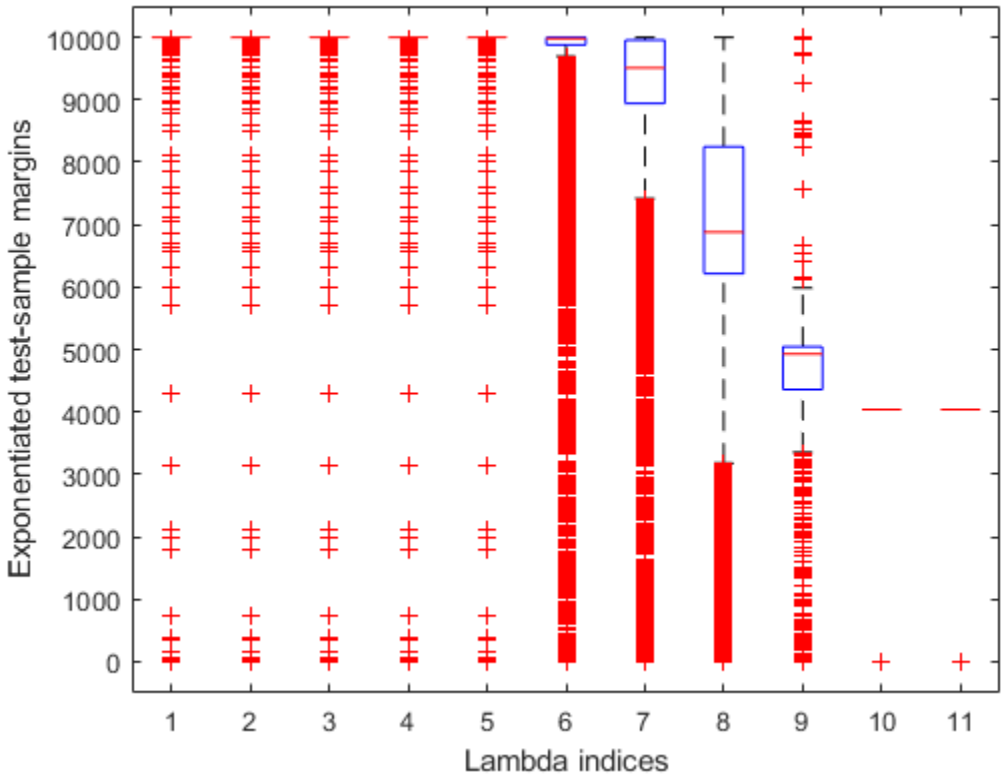
```
m = kfoldMargin(CVMdl);
size(m)
```

```
ans = 1×2

       31572          11
```

`m` is a 31572-by-11 matrix of cross-validated margins for each observation. The columns correspond to the regularization strengths.

Plot the *k*-fold margins for each regularization strength. Because logistic regression scores are in [0,1], margins are in [-1,1]. Rescale the margins to help identify the regularization strength that maximizes the margins over the grid.

```
figure
boxplot(10000.^m)
ylabel('Exponentiated test-sample margins')
xlabel('Lambda indices')
```



Several values of `Lambda` yield *k*-fold margin distributions that are compacted near 10000. Higher values of lambda lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the centers of the *k*-fold margin distributions start decreasing.

```
LambdaFinal = Lambda(5);
```

Train a linear classification model using the entire data set and specify the desired regularization strength.

```
MdlFinal = fitclinear(X,Ystats,'ObservationsIn','columns', ...
    'Learner','logistic','Solver','sparsa','Regularization','lasso', ...
    'Lambda',LambdaFinal);
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationLinear` | `ClassificationPartitionedLinear` | `kfoldEdge` | `kfoldPredict` | `margin`

Introduced in R2016a

kfoldMargin

Classification margins for observations not used in training

Syntax

```
m = kfoldMargin(CVMdl)
m = kfoldMargin(CVMdl,Name,Value)
```

Description

`m = kfoldMargin(CVMdl)` returns the cross-validated classification margins on page 33-3245 obtained by `CVMdl`, which is a cross-validated, error-correcting output codes (ECOC) model composed of linear classification models. That is, for every fold, `kfoldMargin` estimates the classification margins for observations that it holds out when it trains using all other observations.

`m` contains classification margins for each regularization strength in the linear classification models that comprise `CVMdl`.

`m = kfoldMargin(CVMdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify a decoding scheme or verbosity level.

Input Arguments

`CVMdl` — Cross-validated, ECOC model composed of linear classification models

`ClassificationPartitionedLinearECOC` model object

Cross-validated, ECOC model composed of linear classification models, specified as a `ClassificationPartitionedLinearECOC` model object. You can create a `ClassificationPartitionedLinearECOC` model using `fitcecoc` and by:

1. Specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`
2. Setting the name-value pair argument `Learners` to `'linear'` or a linear classification model template returned by `templateLinear`

To obtain estimates, kfoldMargin applies the same data used to cross-validate the ECOC model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in, loss-function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {-1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes the binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, e.g., `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` should have this form

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

By default, if all binary learners are linear classification models using:

- SVM, then `BinaryLoss` is `'hinge'`
- Logistic regression, then `BinaryLoss` is `'quadratic'`

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme

`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`m` — Cross-validated classification margins
numeric vector | numeric matrix

Cross-validated classification margins on page 33-3245, returned as a numeric vector or matrix.

`m` is *n*-by-*L*, where *n* is the number of observations in `X` and *L* is the number of regularization strengths in `Mdl` (that is, `numel(Mdl.Lambda)`).

`m(i,j)` is the cross-validated classification margin of observation *i* using the ECOC model, composed of linear classification models, that has regularization strength `Mdl.Lambda(j)`.

Examples

Estimate *k*-Fold Cross-Validation Margins

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

For simplicity, use the label 'others' for all observations in `Y` that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Cross-validate a multiclass, linear classification model.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learner','linear','CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Estimate the *k*-fold margins.

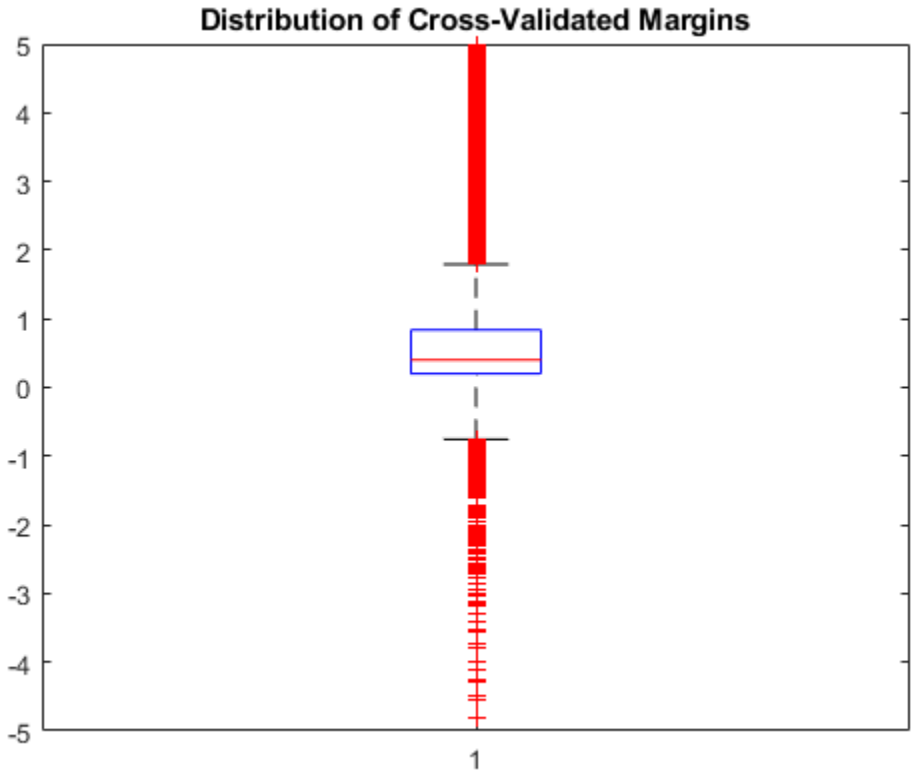
```
m = kfoldMargin(CVMdl);
size(m)
```

```
ans = 1×2

       31572           1
```

`m` is a 31572-by-1 vector. `m(j)` is the average of the out-of-fold margins for observation `j`.

Plot the *k*-fold margins using box plots.

```
figure;
boxplot(m);
h = gca;
h.YLim = [-5 5];
title('Distribution of Cross-Validated Margins')
```



Feature Selection Using *k*-fold Margins

One way to perform feature selection is to compare *k*-fold margins from multiple models. Based solely on this criterion, the classifier with the larger margins is the better classifier.

Load the NLP data set. Preprocess the data as in “Estimate k-Fold Cross-Validation Margins” on page 33-3238, and orient the predictor data so that observations correspond to columns.

```
load nlpdata
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
X = X';
```

Create these two data sets:

- `fullX` contains all predictors.
- `partX` contains 1/2 of the predictors chosen at random.

```
rng(1); % For reproducibility
p = size(X,1); % Number of predictors
halfPredIdx = randsample(p,ceil(0.5*p));
fullX = X;
partX = X(halfPredIdx,:);
```

Create a linear classification model template that specifies optimizing the objective function using SpaRSA.

```
t = templateLinear('Solver','sparsa');
```

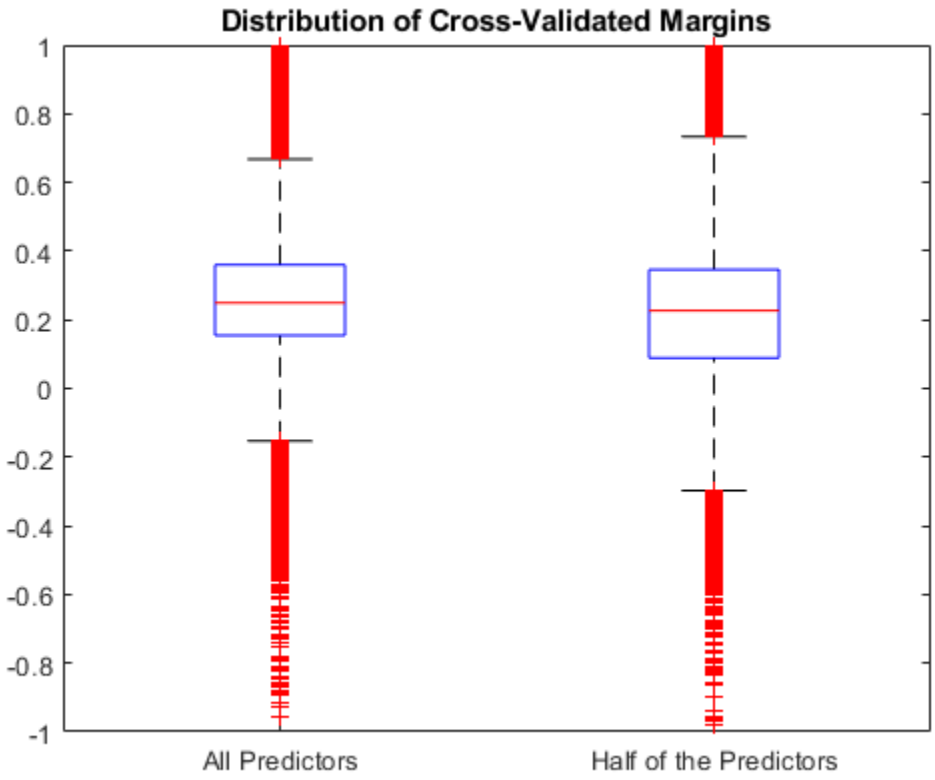
Cross-validate two ECOC models composed of binary, linear classification models: one that uses the all of the predictors and one that uses half of the predictors. Indicate that observations correspond to columns.

```
CVMdl = fitcecoc(fullX,Y,'Learners',t,'CrossVal','on',...
    'ObservationsIn','columns');
PCVMdl = fitcecoc(partX,Y,'Learners',t,'CrossVal','on',...
    'ObservationsIn','columns');
```

`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinearECOC` models.

Estimate the *k*-fold margins for each classifier. Plot the distribution of the *k*-fold margins sets using box plots.

```
fullMargins = kfoldMargin(CVMdl);
partMargins = kfoldMargin(PCVMdl);

figure;
boxplot([fullMargins partMargins],'Labels',...
    {'All Predictors','Half of the Predictors'});
h = gca;
h.YLim = [-1 1];
title('Distribution of Cross-Validated Margins')
```



The distributions of the *k*-fold margins of the two classifiers are similar.

Find Good Lasso Penalty Using *k*-fold Margins

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare distributions of *k*-fold margins.

Load the NLP data set. Preprocess the data as in “Feature Selection Using k-fold Margins” on page 33-3240.

```
load nlpdata
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
X = X';
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,11);
```

Create a linear classification model template that specifies using logistic regression with a lasso penalty, using each of the regularization strengths, optimizing the objective function using SpaRSA, and reducing the tolerance on the gradient of the objective function to `1e-8`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8);
```

Cross-validate an ECOC model composed of binary, linear classification models using 5-fold cross-validation and that

```
rng(10); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns','KFold',5)
```

```
CVMdl =
  ClassificationPartitionedLinearECOC
    CrossValidatedModel: 'LinearECOC'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [comm    dsp    simulink    others]
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model.

Estimate the *k*-fold margins for each regularization strength. The scores for logistic regression are in [0,1]. Apply the quadratic binary loss.

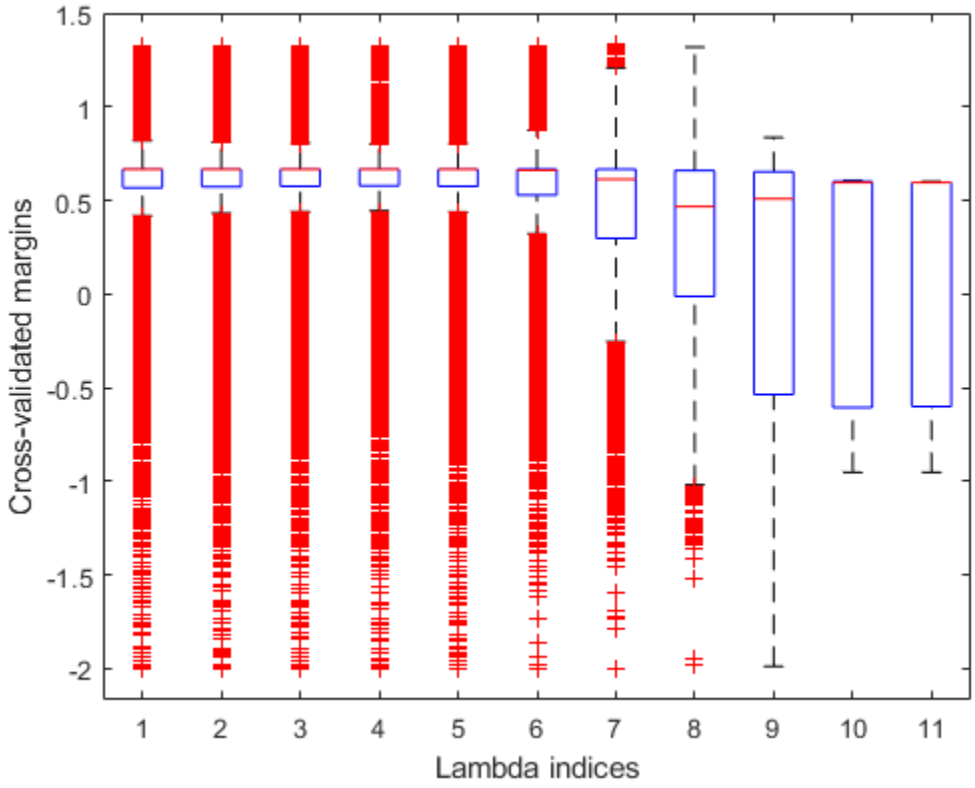
```
m = kfoldMargin(CVMdl,'BinaryLoss','quadratic');
size(m)
```

```
ans = 1×2

       31572          11
```

`m` is a 31572-by-11 matrix of cross-validated margins for each observation. The columns correspond to the regularization strengths.

Plot the *k*-fold margins for each regularization strength.

```
figure;
boxplot(m)
ylabel('Cross-validated margins')
xlabel('Lambda indices')
```



Several values of `Lambda` yield similarly high margin distribution centers with low spreads. Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the margin distribution center starts decreasing and spread starts increasing.

```
LambdaFinal = Lambda(5);
```

Train an ECOC model composed of linear classification model using the entire data set and specify the regularization strength `LambdaFinal`.

```
t = templateLinear('Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda(5),'GradientTolerance',1e-8);
MdlFinal = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns');
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_js_j)/2$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationLinear` | `ClassificationPartitionedLinearECOC` | `kfoldEdge` | `kfoldPredict` | `margin`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2016a

kfoldMargin

Package: `classreg.learning.partition`

Classification margins for cross-validated classification model

Syntax

```
M = kfoldMargin(CVMdl)
M = kfoldMargin(CVMdl,'IncludeInteractions',includeInteractions)
```

Description

`M = kfoldMargin(CVMdl)` returns classification margins on page 33-3248 obtained by the cross-validated classification model `CVMdl`. For every fold, `kfoldMargin` computes classification margins for validation-fold observations using a classifier trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`M = kfoldMargin(CVMdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

Examples

Estimate *k*-fold Margins of Classifier

Find the *k*-fold margins for an ensemble that classifies the `ionosphere` data.

Load the `ionosphere` data set.

```
load ionosphere
```

Create a template tree stump.

```
t = templateTree('MaxNumSplits',1);
```

Train a classification ensemble of decision trees. Specify `t` as the weak learner.

```
Mdl = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
```

Cross-validate the classifier using 10-fold cross-validation.

```
cvens = crossval(Mdl);
```

Compute the k-fold margins. Display summary statistics for the margins.

```
m = kfoldMargin(cvens);
marginStats = table(min(m),mean(m),max(m),...
    'VariableNames',{'Min','Mean','Max'})
```

```
marginStats=1×3 table
      Min       Mean      Max
```

```
    _______    _______    _______

    -11.312    7.3236     23.517
```

Input Arguments

`CVMdl` — Cross-validated partitioned classifier

`ClassificationPartitionedModel` object | `ClassificationPartitionedEnsemble` object | `ClassificationPartitionedGAM` object

Cross-validated partitioned classifier, specified as a `ClassificationPartitionedModel`, `ClassificationPartitionedEnsemble`, or `ClassificationPartitionedGAM` object. You can create the object in two ways:

- Pass a trained classification model listed in the following table to its `crossval` object function.
- Train a classification model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Classification Model	Function
`ClassificationDiscriminant`	`fitcdiscr`
`ClassificationEnsemble`	`fitcensemble`
`ClassificationGAM`	`fitcgam`
`ClassificationKNN`	`fitcknn`
`ClassificationNaiveBayes`	`fitcnb`
`ClassificationNeuralNetwork`	`fitcnet`
`ClassificationSVM`	`fitcsvm`
`ClassificationTree`	`fitctree`

`includeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `ClassificationPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Data Types: `logical`

Output Arguments

`M` — Classification margins

numeric vector

Classification margins on page 33-3248, returned as a numeric vector. `M` is an *n*-by-1 vector, where each row is the margin of the corresponding observation and *n* is the number of observations. (*n* is `size(CVMdl.X,1)` when observations are in rows.)

If you use a holdout validation technique to create `CVMdl` (that is, if `CVMdl.KFold` is `1`), then `M` has `NaN` values for training-fold observations.

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Algorithms

`kfoldMargin` computes classification margins as described in the corresponding `margin` object function. For a model-specific description, see the appropriate `margin` function reference page in the following table.

Model Type	`margin` Function
Discriminant analysis classifier	`margin`
Ensemble classifier	`margin`
Generalized additive model classifier	`margin`
k-nearest neighbor classifier	`margin`
Naive Bayes classifier	`margin`
Neural network classifier	`margin`
Support vector machine classifier	`margin`
Binary decision tree for multiclass classification	`margin`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports *k*-nearest neighbor and SVM model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `kfoldEdge` | `kfoldLoss` | `kfoldPredict` | `kfoldfun`

Introduced in R2011a

kfoldPredict

Package: `classreg.learning.partition`

Classify observations in cross-validated ECOC model

Syntax

```
label = kfoldPredict(CVMdl)
label = kfoldPredict(CVMdl,Name,Value)
[label,NegLoss,PBScore] = kfoldPredict(___)
[label,NegLoss,PBScore,Posterior] = kfoldPredict(___)
```

Description

`label = kfoldPredict(CVMdl)` returns class labels predicted by the cross-validated ECOC model (`ClassificationPartitionedECOC`) `CVMdl`. For every fold, `kfoldPredict` predicts class labels for observations that it holds out during training. `CVMdl.X` contains both sets of observations.

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`label = kfoldPredict(CVMdl,Name,Value)` returns predicted class labels with additional options specified by one or more name-value pair arguments. For example, specify the posterior probability estimation method, decoding scheme, or verbosity level.

`[label,NegLoss,PBScore] = kfoldPredict(___)` additionally returns negated values of the average binary loss per class (`NegLoss`) for validation-fold observations and positive-class scores (`PBScore`) for validation-fold observations classified by each binary learner, using any of the input argument combinations in the previous syntaxes.

If the coding matrix varies across folds (that is, the coding scheme is `sparserandom` or `denserandom`), then `PBScore` is empty (`[]`).

`[label,NegLoss,PBScore,Posterior] = kfoldPredict(___)` additionally returns posterior class probability estimates for validation-fold observations (`Posterior`).

To obtain posterior class probabilities, you must set `'FitPosterior',1` when training the cross-validated ECOC model using `fitcecoc`. Otherwise, `kfoldPredict` throws an error.

Examples

Predict *k*-Fold Cross-Validation Labels

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
```

```
classOrder = unique(Y);
rng(1); % For reproducibility
```

Train and cross-validate an ECOC model using support vector machine (SVM) binary classifiers. Standardize the predictor data using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

Predict the validation-fold labels. Print a random subset of true and predicted labels.

```
labels = kfoldPredict(CVMdl);
idx = randsample(numel(labels),10);
table(Y(idx),labels(idx),...
    'VariableNames',{'TrueLabels','PredictedLabels'})
```

```
ans=10×2 table
    TrueLabels    PredictedLabels
    __________    _______________

    setosa          setosa
    versicolor      versicolor
    setosa          setosa
    virginica       virginica
    versicolor      versicolor
    setosa          setosa
    virginica       virginica
    virginica       virginica
    setosa          setosa
    setosa          setosa
```

`CVMdl` correctly labels the validation-fold observations with indices `idx`.

Predict Cross-Validation Labels Using Custom Binary Loss Function

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
K = numel(classOrder);  % Number of classes
rng(1); % For reproducibility
```

Train and cross-validate an ECOC model using SVM binary classifiers. Standardize the predictor data using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',1);
CVMdl = fitcecoc(X,Y,'CrossVal','on','Learners',t,'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

SVM scores are signed distances from the observation to the decision boundary. Therefore, the domain is $(-\infty, \infty)$. Create a custom binary loss function that:

- Maps the coding design matrix (M) and positive-class classification scores (s) for each learner to the binary loss for each observation
- Uses linear loss
- Aggregates the binary learner loss using the median

You can create a separate function for the binary loss function, and then save it on the MATLAB® path. Alternatively, you can specify an anonymous binary loss function. In this case, create a function handle (`customBL`) to an anonymous binary loss function.

```
customBL = @(M,s)nanmedian(1 - bsxfun(@times,M,s),2)/2;
```

Predict cross-validation labels and estimate the median binary loss per class. Print the median negative binary losses per class for a random set of 10 validation-fold observations.

```
[label,NegLoss] = kfoldPredict(CVMdl,'BinaryLoss',customBL);

idx = randsample(numel(label),10);
classOrder
```

```
classOrder = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
table(Y(idx),label(idx),NegLoss(idx,:),'VariableNames',...
    {'TrueLabel','PredictedLabel','NegLoss'})
```

```
ans=10×3 table
    TrueLabel     PredictedLabel                NegLoss
    __________    ______________    ____________________________________

    setosa          versicolor      0.37141       2.1296       -4.001
    versicolor      versicolor      -1.2166      0.36678     -0.65021
    setosa          versicolor      0.23932       2.0793      -3.8186
    virginica       virginica       -1.9151     -0.19953      0.61467
    versicolor      versicolor      -1.3745      0.45532     -0.58077
    setosa          versicolor      0.20061       2.2774       -3.978
    virginica       versicolor      -1.4926     0.090706    -0.098127
    virginica       virginica       -1.7667     -0.13466      0.40134
    setosa          versicolor      0.20011       1.9111      -3.6112
    setosa          versicolor      0.16118       1.9679      -3.6291
```

The order of the columns corresponds to the elements of `classOrder`. The software predicts the label based on the maximum negated loss. The results indicate that the median of the linear losses might not perform as well as other losses.

Estimate Cross-Validation Posterior Probabilities

Load Fisher's iris data set. Use the petal dimensions as the predictor data X. Specify the response data Y and the order of the classes in Y.

```
load fisheriris
X = meas(:,3:4);
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

Create an SVM template. Standardize the predictors, and specify the Gaussian kernel.

```
t = templateSVM('Standardize',1,'KernelFunction','gaussian');
```

`t` is an SVM template. Most of its properties are empty. When training the ECOC classifier, the software sets the applicable properties to their default values.

Train and cross-validate an ECOC classifier using the SVM template. Transform classification scores to class posterior probabilities (returned by `kfoldPredict`) using the `'FitPosterior'` name-value pair argument. Specify the class order.

```
CVMdl = fitcecoc(X,Y,'Learners',t,'CrossVal','on','FitPosterior',true,...
    'ClassNames',classOrder);
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software uses 10-fold cross-validation.

Predict the validation-fold class posterior probabilities. Use 10 random initial values for the Kullback-Leibler algorithm.

```
[label,~,~,Posterior] = kfoldPredict(CVMdl,'NumKLInitializations',10);
```

The software assigns an observation to the class that yields the smallest average binary loss. Because all the binary learners compute posterior probabilities, the binary loss function is `quadratic`.

Display a random set of results.

```
idx = randsample(size(X,1),10);
CVMdl.ClassNames
```

```
ans = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
table(Y(idx),label(idx),Posterior(idx,:),...
    'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

```
ans=10×3 table
    TrueLabel     PredLabel                  Posterior
    __________    __________    ___________________________________

    versicolor    versicolor     0.0086404       0.98243     0.0089302
    versicolor    virginica     2.2197e-14       0.12437       0.87563
    setosa        setosa             0.999    0.00022837    0.00076884
    versicolor    versicolor    2.2194e-14       0.98916      0.010845
```

```
    virginica     virginica       0.012316      0.012923       0.97476
    virginica     virginica      0.0015569     0.0015636       0.99688
    virginica     virginica      0.0042886     0.0043547       0.99136
    setosa        setosa             0.999    0.00028329    0.00071382
    virginica     virginica      0.0094736     0.0098247        0.9807
    setosa        setosa             0.999    0.00013558    0.00086196
```

The columns of `Posterior` correspond to the class order of `CVMdl.ClassNames`.

Estimate Cross-Validation Posterior Probabilities Using Parallel Computing

Train a multiclass ECOC model and estimate the posterior probabilities using parallel computing.

Load the `arrhythmia` data set. Examine the response data `Y`.

```
load arrhythmia
Y = categorical(Y);
tabulate(Y)
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

```
n = numel(Y);
K = numel(unique(Y));
```

Several classes are not represented in the data, and many of the other classes have low relative frequencies.

Specify an ensemble learning template that uses the GentleBoost method and 50 weak classification tree learners.

```
t = templateEnsemble('GentleBoost',50,'Tree');
```

`t` is a template object. Most of the options are empty (`[]`). The software uses default values for all empty options during training.

Because the response variable contains many classes, specify a sparse random coding design.

```
rng(1); % For reproducibility
Coding = designecoc(K,'sparserandom');
```

Train and cross-validate an ECOC model using parallel computing. Fit posterior probabilities (returned by `kfoldPredict`).

```
pool = parpool;                       % Invokes workers

Starting parallel pool (parpool) using the 'local' profile ...
connected to 6 workers.

options = statset('UseParallel',1);
CVMdl = fitcecoc(X,Y,'Learner',t,'Options',options,'Coding',Coding,...
    'FitPosterior',1,'CrossVal','on');

Warning: One or more folds do not contain points from all the groups.
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software implements 10-fold cross-validation. You can specify a different number of folds using the `'KFold'` name-value pair argument.

The pool invokes six workers, although the number of workers might vary among systems. Because some classes have low relative frequency, one or more folds most likely do not contain observations from all classes.

Estimate posterior probabilities, and display the posterior probability of being classified as not having arrhythmia (class 1) given the data for a random set of validation-fold observations.

```
[~,~,~,posterior] = kfoldPredict(CVMdl,'Options',options);
idx = randsample(n,10);
table(idx,Y(idx),posterior(idx,1),...
    'VariableNames',{'OOFSampleIndex','TrueLabel','PosteriorNoArrhythmia'})

ans=10×3 table
    OOFSampleIndex    TrueLabel    PosteriorNoArrhythmia
    ______________    _________    _____________________

         171              1               0.33654
         221              1               0.85135
          72             16                0.9174
           3             10              0.025649
         202              1                0.8438
         243              1                0.9435
          18              1               0.81198
          49              6              0.090154
         234              1                0.61625
         315              1                0.97187
```

Input Arguments

`CVMdl` — Cross-validated ECOC model

`ClassificationPartitionedECOC` model

Cross-validated ECOC model, specified as a `ClassificationPartitionedECOC` model. You can create a `ClassificationPartitionedECOC` model in two ways:

- Pass a trained ECOC model (`ClassificationECOC`) to `crossval`.
- Train an ECOC model using `fitcecoc` and specify any one of these cross-validation name-value pair arguments: `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldPredict(CVMdl,'PosteriorMethod','qp')` specifies to estimate multiclass posterior probabilities by solving a least-squares problem using quadratic programming.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`NumKLInitializations` — Number of random initial values
`0` (default) | nonnegative integer scalar

Number of random initial values for fitting posterior probabilities by Kullback-Leibler divergence minimization, specified as the comma-separated pair consisting of `'NumKLInitializations'` and a nonnegative integer scalar.

If you do not request the fourth output argument (`Posterior`) and set `'PosteriorMethod','kl'` (the default), then the software ignores the value of `NumKLInitializations`.

For more details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3260.

Example: `'NumKLInitializations',5`

Data Types: `single` | `double`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.

- Specify `'Options',statset('UseParallel',true)`.

`PosteriorMethod` — Posterior probability estimation method
`'kl'` (default) | `'qp'`

Posterior probability estimation method, specified as the comma-separated pair consisting of `'PosteriorMethod'` and `'kl'` or `'qp'`.

- If `PosteriorMethod` is `'kl'`, then the software estimates multiclass posterior probabilities by minimizing the Kullback-Leibler divergence between the predicted and expected posterior probabilities returned by binary learners. For details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3260.
- If `PosteriorMethod` is `'qp'`, then the software estimates multiclass posterior probabilities by solving a least-squares problem using quadratic programming. You need an Optimization Toolbox license to use this option. For details, see “Posterior Estimation Using Quadratic Programming” on page 33-3261.
- If you do not request the fourth output argument (`Posterior`), then the software ignores the value of `PosteriorMethod`.

Example: `'PosteriorMethod','qp'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label` has the same data type and number of rows as `CVMdl.Y`.

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`NegLoss` — Negated average binary losses
numeric matrix

Negated average binary losses, returned as a numeric matrix. `NegLoss` is an n-by-K matrix, where n is the number of observations (`size(CVMdl.X,1)`) and K is the number of unique classes (`size(CVMdl.ClassNames,1)`).

`PBScore` — Positive-class scores
numeric matrix

Positive-class scores for each binary learner, returned as a numeric matrix. `PBScore` is an *n*-by-*L* matrix, where *n* is the number of observations (`size(CVMdl.X,1)`) and *L* is the number of binary learners (`size(CVMdl.CodingMatrix,2)`).

If the coding matrix varies across folds (that is, the coding scheme is `sparserandom` or `denserandom`), then `PBScore` is empty (`[]`).

`Posterior` — Posterior class probabilities
numeric matrix

Posterior class probabilities, returned as a numeric matrix. `Posterior` is an *n*-by-*K* matrix, where *n* is the number of observations (`size(CVMdl.X,1)`) and *K* is the number of unique classes (`size(CVMdl.ClassNames,1)`).

You must set `'FitPosterior',1` when training the cross-validated ECOC model using `fitcecoc` in order to request `Posterior`. Otherwise, the software throws an error.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} \left|m_{kj}\right| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} \left|m_{kj}\right| g(m_{kj}, s_j)}{\sum_{j=1}^{L} \left|m_{kj}\right|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Algorithms

The software can estimate class posterior probabilities by minimizing the Kullback-Leibler divergence or by using quadratic programming. For the following descriptions of the posterior estimation algorithms, assume that:

- m_{kj} is the element (k,j) of the coding design matrix M.
- I is the indicator function.
- $\widehat{p}_k$ is the class posterior probability estimate for class k of an observation, $k = 1,...,K$.
- r_j is the positive-class posterior probability for binary learner j. That is, r_j is the probability that binary learner j classifies an observation into the positive class, given the training data.

Posterior Estimation Using Kullback-Leibler Divergence

By default, the software minimizes the Kullback-Leibler divergence to estimate class posterior probabilities. The Kullback-Leibler divergence between the expected and observed positive-class posterior probabilities is

$$\Delta(r,\widehat{r}) = \sum_{j=1}^{L} w_j\left[r_j\log\frac{r_j}{\widehat{r}_j} + (1-r_j)\log\frac{1-r_j}{1-\widehat{r}_j}\right],$$

where $w_j = \sum_{S_j} w_i^*$ is the weight for binary learner j.

- S_j is the set of observation indices on which binary learner j is trained.
- w_i^* is the weight of observation i.

The software minimizes the divergence iteratively. The first step is to choose initial values $\widehat{p}_k^{(0)};\ k = 1, ..., K$ for the class posterior probabilities.

- If you do not specify `'NumKLIterations'`, then the software tries both sets of deterministic initial values described next, and selects the set that minimizes Δ.
  - $\widehat{p}_k^{(0)} = 1/K;\ k = 1, ..., K$.
  - $\widehat{p}_k^{(0)};\ k = 1, ..., K$ is the solution of the system

    $$M_{01}\widehat{p}^{(0)} = r,$$

    where M_{01} is M with all $m_{kj} = -1$ replaced with 0, and r is a vector of positive-class posterior probabilities returned by the L binary learners [Dietterich et al.] on page 18-279. The software uses `lsqnonneg` to solve the system.
- If you specify `'NumKLIterations',c`, where `c` is a natural number, then the software does the following to choose the set $\widehat{p}_k^{(0)};\ k = 1, ..., K$, and selects the set that minimizes Δ.
  - The software tries both sets of deterministic initial values as described previously.
  - The software randomly generates `c` vectors of length K using `rand`, and then normalizes each vector to sum to 1.

At iteration t, the software completes these steps:

1 Compute

$$\widehat{r}_j^{(t)} = \frac{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1)}{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1 \cup m_{kj} = -1)}.$$

2 Estimate the next class posterior probability using

$$\widehat{p}_k^{(t+1)} = \widehat{p}_k^{(t)} \frac{\sum_{j=1}^{L} w_j \left[r_j I\left(m_{kj} = +1\right) + \left(1 - r_j\right) I\left(m_{kj} = -1\right) \right]}{\sum_{j=1}^{L} w_j \left[\widehat{r}_j^{(t)} I\left(m_{kj} = +1\right) + \left(1 - \widehat{r}_j^{(t)}\right) I\left(m_{kj} = -1\right) \right]}.$$

3 Normalize $\widehat{p}_k^{(t+1)};\ k = 1, ..., K$ so that they sum to 1.

4 Check for convergence.

For more details, see [Hastie et al.] on page 18-280 and [Zadrozny] on page 18-281.

Posterior Estimation Using Quadratic Programming

Posterior probability estimation using quadratic programming requires an Optimization Toolbox license. To estimate posterior probabilities for an observation using this method, the software completes these steps:

1 Estimate the positive-class posterior probabilities, r_j, for binary learners $j = 1,...,L$.

2 Using the relationship between r_j and $\widehat{p}_k$ [Wu et al.] on page 18-281, minimize

$$\sum_{j=1}^{L}\left[-r_j\sum_{k=1}^{K}\widehat{p}_k I\left(m_{kj}=-1\right)+\left(1-r_j\right)\sum_{k=1}^{K}\widehat{p}_k I\left(m_{kj}=+1\right)\right]^2$$

with respect to $\widehat{p}_k$ and the restrictions

$$0\le\widehat{p}_k\le 1$$

$$\sum_k\widehat{p}_k=1.$$

The software performs minimization using `quadprog`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

[5] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[6] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[7] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationPartitionedECOC` | `edge` | `fitcecoc` | `predict` | `quadprog` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

kfoldPredict

Package: `classreg.learning.partition`

Classify observations in cross-validated kernel classification model

Syntax

```
label = kfoldPredict(CVMdl)
[label,score] = kfoldPredict(CVMdl)
```

Description

`label = kfoldPredict(CVMdl)` returns class labels predicted by the cross-validated, binary kernel model (`ClassificationPartitionedKernel`) `CVMdl`. For every fold, `kfoldPredict` predicts class labels for validation-fold observations using a model trained on training-fold observations.

`[label,score] = kfoldPredict(CVMdl)` also returns classification scores on page 33-3268 for both classes.

Examples

Classify Observations Using Cross-Validation

Classify observations using a cross-validated, binary kernel classifier, and display the confusion matrix for the resulting classification.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data.

```
rng(1); % For reproducibility
CVMdl = fitckernel(X,Y,'Crossval','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

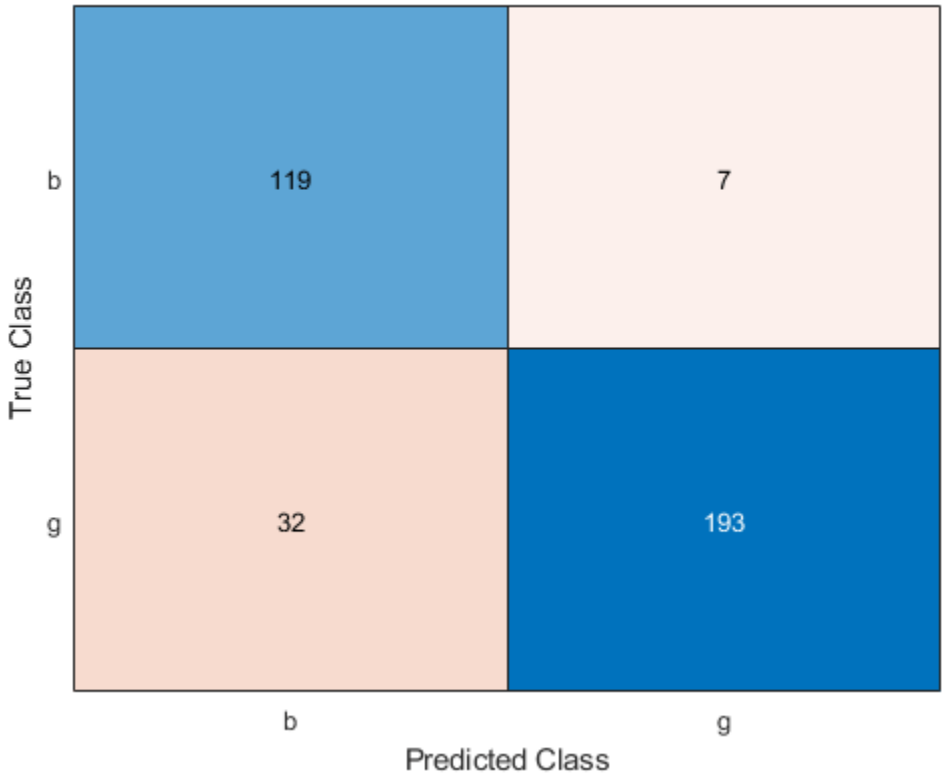
`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Classify the observations that `fitckernel` does not use in training the folds.

```
label = kfoldPredict(CVMdl);
```

Construct a confusion matrix to compare the true classes of the observations to their predicted labels.

```
C = confusionchart(Y,label);
```



The `CVMdl` model misclassifies 32 good (`'g'`) radar returns as being bad (`'b'`) and misclassifies 7 bad radar returns as being good.

Estimate *k*-Fold Cross-Validation Posterior Class Probabilities

Estimate posterior class probabilities using a cross-validated, binary kernel classifier, and determine the quality of the model by plotting a receiver operating characteristic (ROC) curve. Cross-validated kernel classification models return posterior probabilities for logistic regression learners only.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, which are labeled either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Cross-validate a binary kernel classification model using the data. Specify the class order, and fit logistic regression learners.

```
rng(1); % For reproducibility
CVMdl = fitckernel(X,Y,'Crossval','on', ...
    'ClassNames',{'b','g'},'Learner','logistic')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernel` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Predict the posterior class probabilities for the observations that `fitckernel` does not use in training the folds.

```
[~,posterior] = kfoldPredict(CVMdl);
```

The output `posterior` is a matrix with two columns and `n` rows, where `n` is the number of observations. Column `i` contains posterior probabilities of `CVMdl.ClassNames(i)` given a particular observation.

Obtain false and true positive rates, and estimate the area under the curve (AUC). Specify that the second class is the positive class.

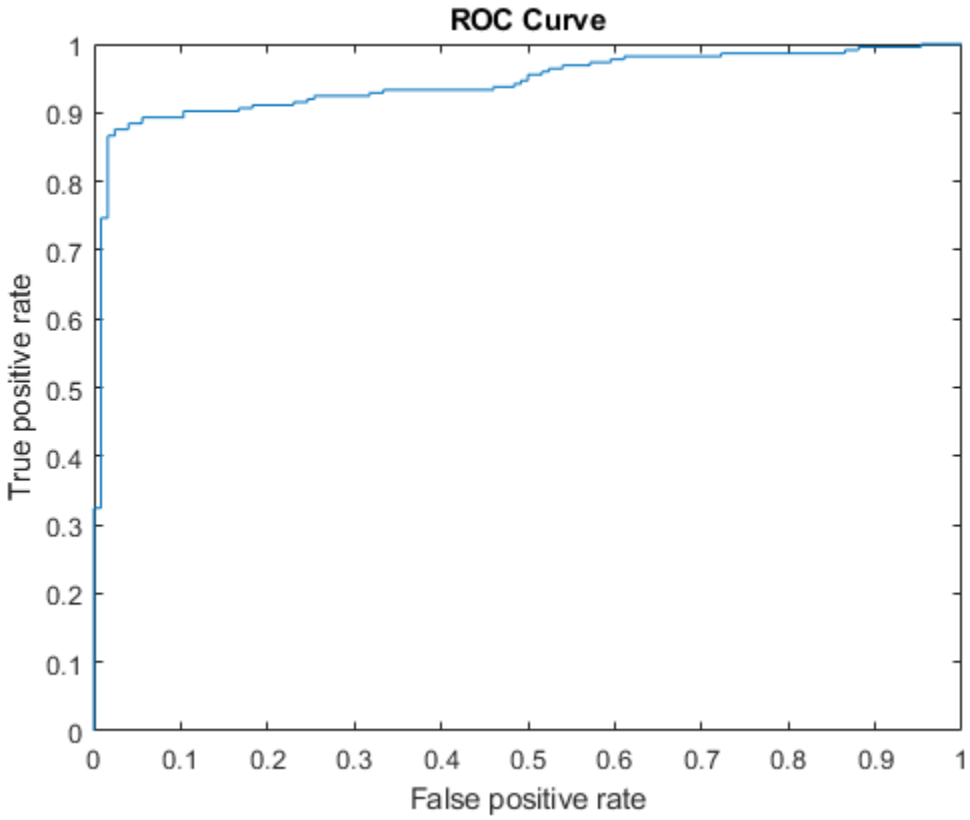
```
[fpr,tpr,~,auc] = perfcurve(Y,posterior(:,2),CVMdl.ClassNames(2));
auc
```

```
auc = 0.9441
```

The AUC is close to `1`, which indicates that the model predicts labels well.

Plot an ROC curve.

```
plot(fpr,tpr)
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve')
```



Input Arguments

`CVMdl` — Cross-validated, binary kernel classification model

`ClassificationPartitionedKernel` model object

Cross-validated, binary kernel classification model, specified as a `ClassificationPartitionedKernel` model object. You can create a `ClassificationPartitionedKernel` model by using `fitckernel` and specifying any one of the cross-validation name-value pair arguments.

To obtain estimates, `kfoldPredict` applies the same data used to cross-validate the kernel classification model (`X` and `Y`).

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical matrix | numeric matrix | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric matrix, or cell array of character vectors.

`label` has *n* rows, where *n* is the number of observations in `X`, and has the same data type as the observed class labels (`Y`) used to train `CVMdl`. (The software treats string arrays as cell arrays of character vectors.)

`kfoldPredict` classifies observations into the class yielding the highest score.

`score` — Classification scores
numeric array

Classification scores on page 33-3268, returned as an *n*-by-2 numeric array, where *n* is the number of observations in `X`. `score(i,j)` is the score for classifying observation `i` into class `j`. The order of the classes is stored in `CVMdl.ClassNames`.

If `CVMdl.Trained{1}.Learner` is `'logistic'`, then classification scores are posterior probabilities.

More About

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b\,.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationPartitionedKernel` | `fitckernel`

Introduced in R2018b

kfoldPredict

Package: `classreg.learning.partition`

Classify observations in cross-validated kernel ECOC model

Syntax

```
label = kfoldPredict(CVMdl)
label = kfoldPredict(CVMdl,Name,Value)
[label,NegLoss,PBScore] = kfoldPredict( ___ )
[label,NegLoss,PBScore,Posterior] = kfoldPredict( ___ )
```

Description

`label = kfoldPredict(CVMdl)` returns class labels predicted by the cross-validated kernel ECOC model (`ClassificationPartitionedKernelECOC`) `CVMdl`. For every fold, `kfoldPredict` predicts class labels for validation-fold observations using a model trained on training-fold observations. `kfoldPredict` applies the same data used to create `CVMdl` (see `fitcecoc`).

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`label = kfoldPredict(CVMdl,Name,Value)` returns predicted class labels with additional options specified by one or more name-value pair arguments. For example, specify the posterior probability estimation method, decoding scheme, or verbosity level.

`[label,NegLoss,PBScore] = kfoldPredict( ___ )` additionally returns negated values of the average binary loss per class (`NegLoss`) for validation-fold observations and positive-class scores (`PBScore`) for validation-fold observations classified by each binary learner, using any of the input argument combinations in the previous syntaxes.

If the coding matrix varies across folds (that is, the coding scheme is `sparserandom` or `denserandom`), then `PBScore` is empty (`[]`).

`[label,NegLoss,PBScore,Posterior] = kfoldPredict( ___ )` additionally returns posterior class probability estimates for validation-fold observations (`Posterior`).

To obtain posterior class probabilities, the kernel classification binary learners must be logistic regression models. Otherwise, `kfoldPredict` throws an error.

Examples

Classify Observations Using Cross-Validation

Classify observations using a cross-validated, multiclass kernel ECOC classifier, and display the confusion matrix for the resulting classification.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model composed of kernel binary learners.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners','kernel','CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'


  Properties, Methods
```

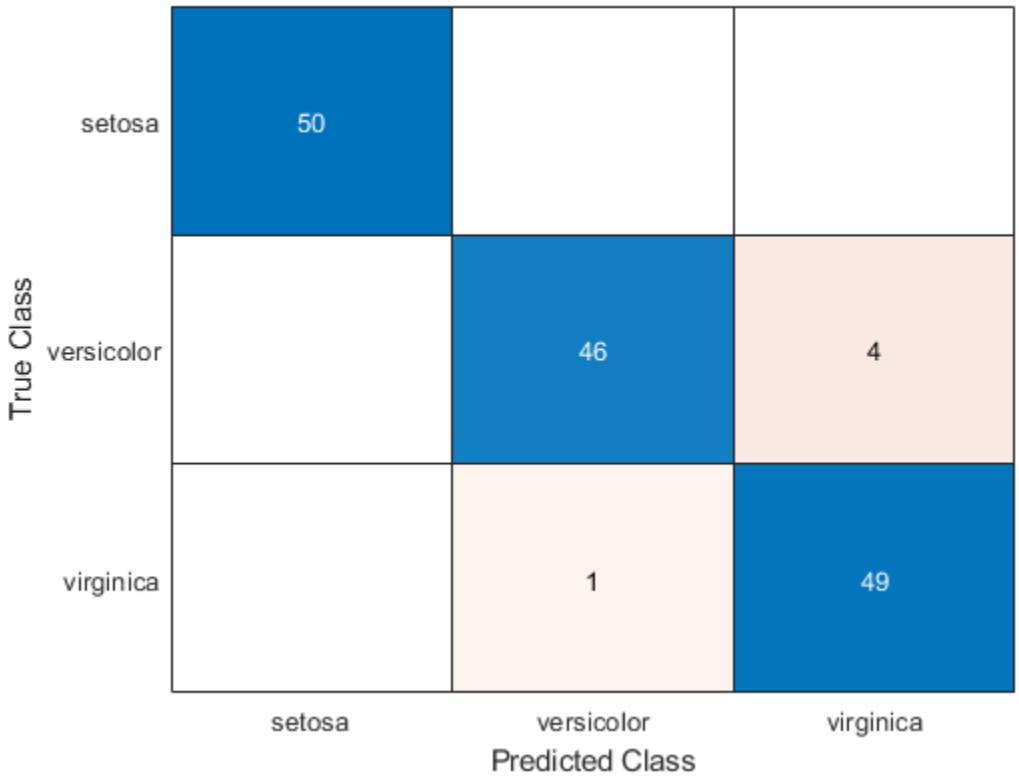
`CVMdl` is a `ClassificationPartitionedKernelECOC` model. By default, the software implements 10-fold cross-validation. To specify a different number of folds, use the `'KFold'` name-value pair argument instead of `'Crossval'`.

Classify the observations that `fitcecoc` does not use in training the folds.

```
label = kfoldPredict(CVMdl);
```

Construct a confusion matrix to compare the true classes of the observations to their predicted labels.

```
C = confusionchart(Y,label);
```



The `CVMdl` model misclassifies four `'versicolor'` irises as `'virginica'` irises and misclassifies one `'virginica'` iris as a `'versicolor'` iris.

Predict Cross-Validation Labels Using Custom Binary Loss

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Cross-validate an ECOC model of kernel classification models using 5-fold cross-validation.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners','kernel','KFold',5)
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
```

```
          ScoreTransform: 'none'

  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedKernelECOC` model. It contains the property `Trained`, which is a 5-by-1 cell array of `CompactClassificationECOC` models.

By default, the kernel classification models that compose the `CompactClassificationECOC` models use SVMs. SVM scores are signed distances from the observation to the decision boundary. Therefore, the domain is $(-\infty, \infty)$. Create a custom binary loss function that:

- Maps the coding design matrix (M) and positive-class classification scores (s) for each learner to the binary loss for each observation
- Uses linear loss
- Aggregates the binary learner loss using the median

You can create a separate function for the binary loss function, and then save it on the MATLAB® path. Or, you can specify an anonymous binary loss function. In this case, create a function handle (`customBL`) to an anonymous binary loss function.

```
customBL = @(M,s)nanmedian(1 - bsxfun(@times,M,s),2)/2;
```

Predict cross-validation labels and estimate the median binary loss per class. Print the median negative binary losses per class for a random set of 10 observations.

```
[label,NegLoss] = kfoldPredict(CVMdl,'BinaryLoss',customBL);

idx = randsample(numel(label),10);
table(Y(idx),label(idx),NegLoss(idx,1),NegLoss(idx,2),NegLoss(idx,3),...
    'VariableNames',[{'True'};{'Predicted'};...
    unique(CVMdl.ClassNames)])
```

```
ans=10×5 table
         True           Predicted        setosa     versicolor    virginica
    ________________    ____________    ________    __________    _________

    {'setosa'    }    {'setosa'    }     0.20926     -0.84572     -0.86354
    {'setosa'    }    {'setosa'    }     0.16144     -0.90572     -0.75572
    {'virginica' }    {'versicolor'}    -0.83532     -0.12157     -0.54311
    {'virginica' }    {'virginica' }    -0.97235     -0.69759      0.16994
    {'virginica' }    {'virginica' }    -0.89441     -0.69937     0.093778
    {'virginica' }    {'virginica' }    -0.86774     -0.47297     -0.15929
    {'setosa'    }    {'setosa'    }     -0.1026     -0.69671     -0.70069
    {'setosa'    }    {'setosa'    }      0.1001     -0.89163     -0.70848
    {'virginica' }    {'virginica' }     -1.0106     -0.52919     0.039829
    {'versicolor'}    {'versicolor'}     -1.0298     0.027354     -0.49757
```

The cross-validated model correctly predicts the labels for 9 of the 10 random observations.

Estimate *k*-Fold Cross-Validation Posterior Class Probabilities

Estimate posterior class probabilities using a cross-validated, multiclass kernel ECOC classification model. Kernel classification models return posterior probabilities for logistic regression learners only.

Load Fisher's iris data set. `X` contains flower measurements, and `Y` contains the names of flower species.

```
load fisheriris
X = meas;
Y = species;
```

Create a kernel template for the binary kernel classification models. Specify to fit logistic regression learners.

```
t = templateKernel('Learner','logistic')
```

```
t =
Fit template for classification Kernel.

             BetaTolerance: []
                 BlockSize: []
             BoxConstraint: []
                   Epsilon: []
    NumExpansionDimensions: []
         GradientTolerance: []
        HessianHistorySize: []
            IterationLimit: []
               KernelScale: []
                    Lambda: []
                   Learner: 'logistic'
              LossFunction: []
                    Stream: []
            VerbosityLevel: []
                   Version: 1
                    Method: 'Kernel'
                      Type: 'classification'
```

`t` is a kernel template. Most of its properties are empty. When training an ECOC classifier using the template, the software sets the applicable properties to their default values.

Cross-validate an ECOC model using the kernel template.

```
rng('default'); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernelECOC
    CrossValidatedModel: 'KernelECOC'
           ResponseName: 'Y'
        NumObservations: 150
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'setosa'  'versicolor'  'virginica'}
         ScoreTransform: 'none'
```

```
  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedECOC` model. By default, the software uses 10-fold cross-validation.

Predict the validation-fold class posterior probabilities.

```
[label,~,~,Posterior] = kfoldPredict(CVMdl);
```

The software assigns an observation to the class that yields the smallest average binary loss. Because all binary learners are computing posterior probabilities, the binary loss function is `quadratic`.

Display the posterior probabilities for 10 randomly selected observations.

```
idx = randsample(size(X,1),10);
CVMdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
table(Y(idx),label(idx),Posterior(idx,:),...
    'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

```
ans=10×3 table
      TrueLabel         PredLabel                  Posterior
    ______________    ______________    ___________________________________

    {'setosa'    }    {'setosa'    }     0.68216     0.18546     0.13238
    {'virginica' }    {'virginica' }      0.1581     0.14405     0.69785
    {'virginica' }    {'virginica' }    0.071807    0.093291      0.8349
    {'setosa'    }    {'setosa'    }     0.74918     0.11434     0.13648
    {'versicolor'}    {'versicolor'}     0.09375     0.67149     0.23476
    {'versicolor'}    {'versicolor'}    0.036202     0.85544     0.10836
    {'versicolor'}    {'versicolor'}      0.2252     0.50473     0.27007
    {'virginica' }    {'virginica' }    0.061562     0.11086     0.82758
    {'setosa'    }    {'setosa'    }     0.42448     0.21181     0.36371
    {'virginica' }    {'virginica' }    0.082705      0.1428      0.7745
```

The columns of `Posterior` correspond to the class order of `CVMdl.ClassNames`.

Input Arguments

`CVMdl` — Cross-validated kernel ECOC model

`ClassificationPartitionedKernelECOC` model

Cross-validated kernel ECOC model, specified as a `ClassificationPartitionedKernelECOC` model. You can create a `ClassificationPartitionedKernelECOC` model by training an ECOC model using `fitcecoc` and specifying these name-value pair arguments:

- `'Learners'` – Set the value to `'kernel'`, a template object returned by `templateKernel`, or a cell array of such template objects.

- One of the arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `kfoldPredict(CVMdl,'PosteriorMethod','qp')` specifies to estimate multiclass posterior probabilities by solving a least-squares problem using quadratic programming.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example, `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.

- L is the number of binary learners.

By default, if all binary learners are kernel classification models using SVM, then `BinaryLoss` is `'hinge'`. If all binary learners are kernel classification models using logistic regression, then `BinaryLoss` is `'quadratic'`.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3278.

Example: `'Decoding','lossbased'`

`NumKLInitializations` — Number of random initial values
`0` (default) | nonnegative integer scalar

Number of random initial values for fitting posterior probabilities by Kullback-Leibler divergence minimization, specified as the comma-separated pair consisting of `'NumKLInitializations'` and a nonnegative integer scalar.

If you do not request the fourth output argument (`Posterior`) and set `'PosteriorMethod','kl'` (the default), then the software ignores the value of `NumKLInitializations`.

For more details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3279.

Example: `'NumKLInitializations',5`

Data Types: `single` | `double`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`PosteriorMethod` — Posterior probability estimation method
`'kl'` (default) | `'qp'`

Posterior probability estimation method, specified as the comma-separated pair consisting of `'PosteriorMethod'` and `'kl'` or `'qp'`.

- If `PosteriorMethod` is `'kl'`, then the software estimates multiclass posterior probabilities by minimizing the Kullback-Leibler divergence between the predicted and expected posterior probabilities returned by binary learners. For details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3279.

- If `PosteriorMethod` is `'qp'`, then the software estimates multiclass posterior probabilities by solving a least-squares problem using quadratic programming. You need an Optimization Toolbox license to use this option. For details, see “Posterior Estimation Using Quadratic Programming” on page 33-3280.
- If you do not request the fourth output argument (`Posterior`), then the software ignores the value of `PosteriorMethod`.

Example: `'PosteriorMethod','qp'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label` has the same data type and number of rows as `CVMdl.Y`.

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`NegLoss` — Negated average binary losses
numeric matrix

Negated average binary losses, returned as a numeric matrix. `NegLoss` is an *n*-by-*K* matrix, where *n* is the number of observations (`size(CVMdl.Y,1)`) and *K* is the number of unique classes (`size(CVMdl.ClassNames,1)`).

`PBScore` — Positive-class scores
numeric matrix

Positive-class scores for each binary learner, returned as a numeric matrix. `PBScore` is an *n*-by-*L* matrix, where *n* is the number of observations (`size(CVMdl.Y,1)`) and *L* is the number of binary learners (`size(CVMdl.CodingMatrix,2)`).

If the coding matrix varies across folds (that is, the coding scheme is `sparserandom` or `denserandom`), then `PBScore` is empty (`[]`).

`Posterior` — Posterior class probabilities
numeric matrix

Posterior class probabilities, returned as a numeric matrix. `Posterior` is an n-by-K matrix, where n is the number of observations (`size(CVMdl.Y,1)`) and K is the number of unique classes (`size(CVMdl.ClassNames,1)`).

To return posterior probabilities, each kernel classification binary learner must have its `Learner` property set to `'logistic'`. Otherwise, the software throws an error.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$

Value	Description	Score Domain	$g(y_j,s_j)$
`'linear'`	Linear	(–∞,∞)	$(1 - y_j s_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_j s_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Algorithms

The software can estimate class posterior probabilities by minimizing the Kullback-Leibler divergence or by using quadratic programming. For the following descriptions of the posterior estimation algorithms, assume that:

- m_{kj} is the element (k,j) of the coding design matrix M.
- I is the indicator function.
- $\widehat{p}_k$ is the class posterior probability estimate for class k of an observation, $k = 1,...,K$.
- r_j is the positive-class posterior probability for binary learner j. That is, r_j is the probability that binary learner j classifies an observation into the positive class, given the training data.

Posterior Estimation Using Kullback-Leibler Divergence

By default, the software minimizes the Kullback-Leibler divergence to estimate class posterior probabilities. The Kullback-Leibler divergence between the expected and observed positive-class posterior probabilities is

$$\Delta(r,\widehat{r}) = \sum_{j=1}^{L} w_j \left[r_j \log\frac{r_j}{\widehat{r}_j} + (1 - r_j)\log\frac{1 - r_j}{1 - \widehat{r}_j} \right],$$

where $w_j = \sum_{S_j} w_i^*$ is the weight for binary learner j.

- S_j is the set of observation indices on which binary learner j is trained.
- w_i^* is the weight of observation i.

The software minimizes the divergence iteratively. The first step is to choose initial values $\widehat{p}_k^{(0)};\ k = 1,...,K$ for the class posterior probabilities.

- If you do not specify `'NumKLIterations'`, then the software tries both sets of deterministic initial values described next, and selects the set that minimizes Δ.
  - $\widehat{p}_k^{(0)} = 1/K;\ k = 1,...,K$.
  - $\widehat{p}_k^{(0)};\ k = 1,...,K$ is the solution of the system

$$M_{01}\hat{p}^{(0)} = r,$$

where M_{01} is M with all $m_{kj} = -1$ replaced with 0, and r is a vector of positive-class posterior probabilities returned by the L binary learners [Dietterich et al.] on page 18-279. The software uses `lsqnonneg` to solve the system.

- If you specify `'NumKLIterations',c`, where `c` is a natural number, then the software does the following to choose the set $\hat{p}_k^{(0)};\ k = 1,\ldots,K$, and selects the set that minimizes Δ.
  - The software tries both sets of deterministic initial values as described previously.
  - The software randomly generates `c` vectors of length K using `rand`, and then normalizes each vector to sum to 1.

At iteration t, the software completes these steps:

1 Compute

$$\hat{r}_j^{(t)} = \frac{\sum_{k=1}^{K} \hat{p}_k^{(t)} I(m_{kj} = +1)}{\sum_{k=1}^{K} \hat{p}_k^{(t)} I(m_{kj} = +1 \cup m_{kj} = -1)}.$$

2 Estimate the next class posterior probability using

$$\hat{p}_k^{(t+1)} = \hat{p}_k^{(t)} \frac{\sum_{j=1}^{L} w_j \left[r_j I(m_{kj} = +1) + (1 - r_j) I(m_{kj} = -1) \right]}{\sum_{j=1}^{L} w_j \left[\hat{r}_j^{(t)} I(m_{kj} = +1) + \left(1 - \hat{r}_j^{(t)}\right) I(m_{kj} = -1) \right]}.$$

3 Normalize $\hat{p}_k^{(t+1)};\ k = 1,\ldots,K$ so that they sum to 1.

4 Check for convergence.

For more details, see [Hastie et al.] on page 18-280 and [Zadrozny] on page 18-281.

Posterior Estimation Using Quadratic Programming

Posterior probability estimation using quadratic programming requires an Optimization Toolbox license. To estimate posterior probabilities for an observation using this method, the software completes these steps:

1 Estimate the positive-class posterior probabilities, r_j, for binary learners $j = 1,\ldots,L$.

2 Using the relationship between r_j and $\hat{p}_k$ [Wu et al.] on page 18-281, minimize

$$\sum_{j=1}^{L} \left[-r_j \sum_{k=1}^{K} \hat{p}_k I(m_{kj} = -1) + (1 - r_j) \sum_{k=1}^{K} \hat{p}_k I(m_{kj} = +1) \right]^2$$

with respect to $\hat{p}_k$ and the restrictions

$$0 \le \hat{p}_k \le 1$$

$$\sum_k \hat{p}_k = 1.$$

The software performs minimization using `quadprog`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

[5] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[6] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[7] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

See Also

`ClassificationPartitionedKernelECOC` | `fitcecoc`

Introduced in R2018b

kfoldPredict

Predict labels for observations not used for training

Syntax

```
Label = kfoldPredict(CVMdl)
[Label,Score] = kfoldPredict(CVMdl)
```

Description

`Label = kfoldPredict(CVMdl)` returns cross-validated class labels predicted by the cross-validated, binary, linear classification model `CVMdl`. That is, for every fold, `kfoldPredict` predicts class labels for observations that it holds out when it trains using all other observations.

`Label` contains predicted class labels for each regularization strength in the linear classification models that compose `CVMdl`.

`[Label,Score] = kfoldPredict(CVMdl)` also returns cross-validated classification scores on page 33-3289 for both classes. `Score` contains classification scores for each regularization strength in `CVMdl`.

Input Arguments

`CVMdl` — Cross-validated, binary, linear classification model

`ClassificationPartitionedLinear` model object

Cross-validated, binary, linear classification model, specified as a `ClassificationPartitionedLinear` model object. You can create a `ClassificationPartitionedLinear` model using `fitclinear` and specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldPredict applies the same data used to cross-validate the linear classification model (`X` and `Y`).

Output Arguments

`Label` — Cross-validated, predicted class labels

categorical array | character array | logical matrix | numeric matrix | cell array of character vectors

Cross-validated, predicted class labels, returned as a categorical or character array, logical or numeric matrix, or cell array of character vectors.

In most cases, `Label` is an n-by-L array of the same data type as the observed class labels (see `Y`) used to create `CVMdl`. (The software treats string arrays as cell arrays of character vectors.) n is the number of observations in the predictor data (see `X`) and L is the number of regularization strengths in `CVMdl.Trained{1}.Lambda`. That is, `Label(`i`,`j`)` is the predicted class label for observation i using the linear classification model that has regularization strength `CVMdl.Trained{1}.Lambda(`j`)`.

If `Y` is a character array and $L > 1$, then `Label` is a cell array of class labels.

`Score` — Cross-validated classification scores
numeric array

Cross-validated classification scores on page 33-3289, returned as an n-by-2-by-L numeric array. n is the number of observations in the predictor data that created `CVMdl` (see `X`) and L is the number of regularization strengths in `CVMdl.Trained{1}.Lambda`. `Score(`i`,`k`,`j`)` is the score for classifying observation i into class k using the linear classification model that has regularization strength `CVMdl.Trained{1}.Lambda(`j`)`. `CVMdl.ClassNames` stores the order of the classes.

If `CVMdl.Trained{1}.Learner` is `'logistic'`, then classification scores are posterior probabilities.

Examples

Predict *k*-fold Cross-Validation Labels

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Cross-validate a binary, linear classification model using the entire data set, which can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'CrossVal','on');
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [34023x1 double]
              Bias: -1.0008
            Lambda: 3.5193e-05
           Learner: 'svm'


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinear` model. By default, the software implements 10-fold cross validation. You can alter the number of folds using the `'KFold'` name-value pair argument.

Predict labels for the observations that `fitclinear` did not use in training the folds.

```
label = kfoldPredict(CVMdl);
```

Because there is one regularization strength in `Mdl1`, `label` is a column vector of predictions containing as many rows as observations in `X`.

Construct a confusion matrix.

```
ConfusionTrain = confusionchart(Ystats,label);
```

The model misclassifies 15 `'stats'` documentation pages as being outside of the Statistics and Machine Learning Toolbox documentation, and misclassifies nine pages as `'stats'` pages.

Estimate *k*-fold Cross-Validation Posterior Class Probabilities

Linear classification models return posterior probabilities for logistic regression learners only.

Load the NLP data set and preprocess it as in “Predict k-fold Cross-Validation Labels” on page 33-3283. Transpose the predictor data matrix.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Cross-validate binary, linear classification models using 5-fold cross-validation. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'KFold',5,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','GradientTolerance',1e-8);
```

Predict the posterior class probabilities for observations not used to train each fold.

```
[~,posterior] = kfoldPredict(CVMdl);
CVMdl.ClassNames
```

```
ans = 2x1 logical array

   0
   1
```

Because there is one regularization strength in `CVMdl`, `posterior` is a matrix with 2 columns and rows equal to the number of observations. Column *i* contains posterior probabilities of `Mdl.ClassNames(i)` given a particular observation.

Obtain false and true positive rates, and estimate the AUC. Specify that the second class is the positive class.

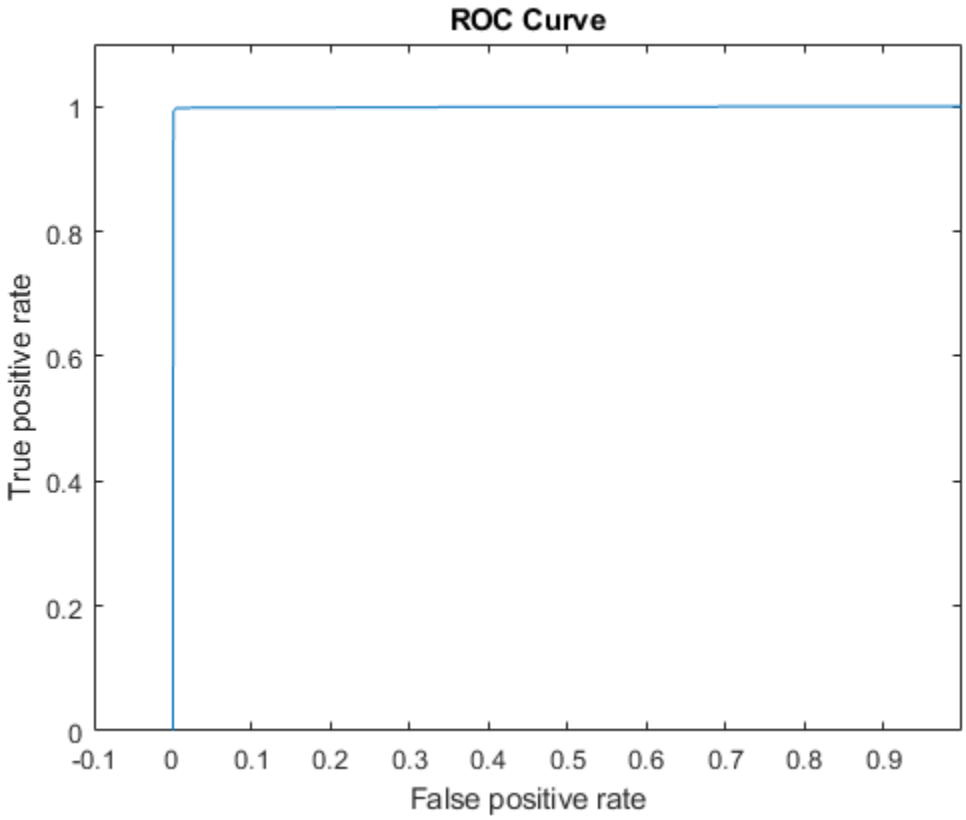
```
[fpr,tpr,~,auc] = perfcurve(Ystats,posterior(:,2),CVMdl.ClassNames(2));
auc
```

```
auc = 0.9990
```

The AUC is `0.9990`, which indicates a model that predicts well.

Plot an ROC curve.

```
figure;
plot(fpr,tpr)
h = gca;
h.XLim(1) = -0.1;
h.YLim(2) = 1.1;
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve')
```



The ROC curve indicates that the model classifies almost perfectly.

Find Good Lasso Penalty Using Cross-Validated AUC

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare cross-validated AUC values.

Load the NLP data set. Preprocess the data as in “Estimate k-fold Cross-Validation Posterior Class Probabilities” on page 33-3284.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

There are 9471 observations in the test sample.

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Cross-validate a binary, linear classification models that use each of the regularization strengths and 5-fold cross-validation. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10) % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns', ...
    'KFold',5,'Learner','logistic','Solver','sparsa', ...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)

CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 5
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods


Mdl1 = CVMdl.Trained{1}

Mdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl1` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 11 models, one for each regularization strength in `Lambda`.

Predict the cross-validated labels and posterior class probabilities.

```
[label,posterior] = kfoldPredict(CVMdl);
CVMdl.ClassNames;
[n,K,L] = size(posterior)

n = 31572

K = 2

L = 11

posterior(3,1,5)

ans = 1.0000
```

`label` is a 31572-by-11 matrix of predicted labels. Each column corresponds to the predicted labels of the model trained using the corresponding regularization strength. `posterior` is a 31572-by-2-by-11 matrix of posterior class probabilities. Columns correspond to classes and pages correspond to regularization strengths. For example, `posterior(3,1,5)` indicates that the posterior probability

that the first class (label `0`) is assigned to observation 3 by the model that uses `Lambda(5)` as a regularization strength is 1.0000.

For each model, compute the AUC. Designate the second class as the positive class.

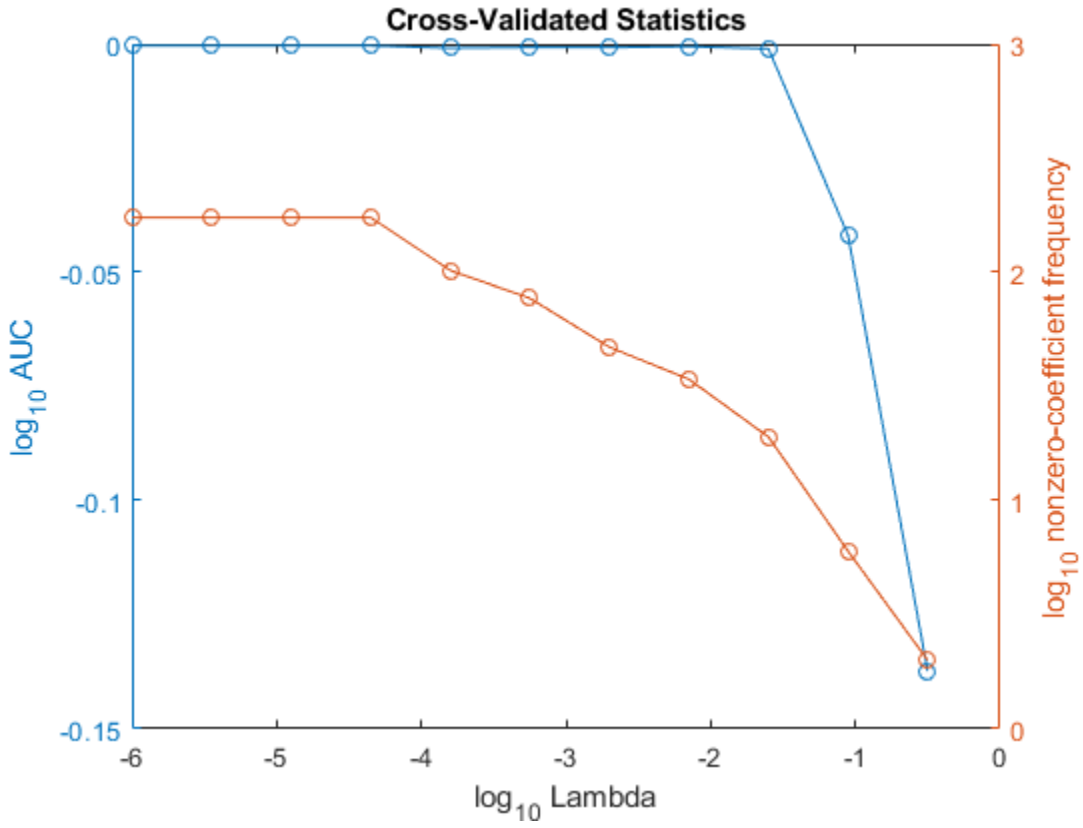
```
auc = 1:numel(Lambda);  % Preallocation
for j = 1:numel(Lambda)
    [~,~,~,auc(j)] = perfcurve(Ystats,posterior(:,2,j),CVMdl.ClassNames(2));
end
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you trained the model. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns', ...
    'Learner','logistic','Solver','sparsa','Regularization','lasso', ...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the test-sample error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure
[h,hL1,hL2] = plotyy(log10(Lambda),log10(auc), ...
    log10(Lambda),log10(numNZCoeff + 1));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} AUC')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
title('Cross-Validated Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and high AUC. In this case, a value between 10^{-3} to 10^{-1} should suffice.

```
idxFinal = 9;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

See Also

`ClassificationLinear` | `ClassificationPartitionedLinear` | `confusionchart` | `perfcurve` | `predict` | `testcholdout`

Introduced in R2016a

kfoldPredict

Predict labels for observations not used for training

Syntax

```
Label = kfoldPredict(CVMdl)
Label = kfoldPredict(CVMdl,Name,Value)
[Label,NegLoss,PBScore] = kfoldPredict( ___ )
[Label,NegLoss,PBScore,Posterior] = kfoldPredict( ___ )
```

Description

`Label = kfoldPredict(CVMdl)` returns class labels predicted by the cross-validated ECOC model composed of linear classification models `CVMdl`. That is, for every fold, `kfoldPredict` predicts class labels for observations that it holds out when it trains using all other observations. `kfoldPredict` applies the same data used create `CVMdl` (see `fitcecoc`).

Also, `Label` contains class labels for each regularization strength in the linear classification models that compose `CVMdl`.

`Label = kfoldPredict(CVMdl,Name,Value)` returns predicted class labels with additional options specified by one or more `Name,Value` pair arguments. For example, specify the posterior probability estimation method, decoding scheme, or verbosity level.

`[Label,NegLoss,PBScore] = kfoldPredict( ___ )` additionally returns, for held-out observations and each regularization strength:

- Negated values of the average binary loss per class (`NegLoss`).
- Positive-class scores (`PBScore`) for each binary learner.

`[Label,NegLoss,PBScore,Posterior] = kfoldPredict( ___ )` additionally returns posterior class probability estimates for held-out observations and for each regularization strength. To return posterior probabilities, the linear classification model learners must be logistic regression models.

Input Arguments

`CVMdl` — Cross-validated, ECOC model composed of linear classification models

`ClassificationPartitionedLinearECOC` model object

Cross-validated, ECOC model composed of linear classification models, specified as a `ClassificationPartitionedLinearECOC` model object. You can create a `ClassificationPartitionedLinearECOC` model using `fitcecoc` and by:

1. Specifying any one of the cross-validation, name-value pair arguments, for example, `CrossVal`
2. Setting the name-value pair argument `Learners` to `'linear'` or a linear classification model template returned by `templateLinear`

To obtain estimates, kfoldPredict applies the same data used to cross-validate the ECOC model (`X` and `Y`).

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`BinaryLoss` — Binary learner loss function
`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in, loss-function name or function handle.

- This table contains names and descriptions of the built-in functions, where y_j is a class label for a particular binary learner (in the set {-1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_js_j)/2$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes the binary losses such that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, e.g., `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` should have this form

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

By default, if all binary learners are linear classification models using:

- SVM, then `BinaryLoss` is `'hinge'`
- Logistic regression, then `BinaryLoss` is `'quadratic'`

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3259.

Example: `'Decoding','lossbased'`

`NumKLInitializations` — Number of random initial values
`0` (default) | nonnegative integer

Number of random initial values for fitting posterior probabilities by Kullback-Leibler divergence minimization, specified as the comma-separated pair consisting of `'NumKLInitializations'` and a nonnegative integer.

To use this option, you must:

- Return the fourth output argument (`Posterior`).
- The linear classification models that compose the ECOC models must use logistic regression learners (that is, `CVMdl.Trained{1}.BinaryLearners{1}.Learner` must be `'logistic'`).
- `PosteriorMethod` must be `'kl'`.

For more details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3301.

Example: `'NumKLInitializations',5`

Data Types: `single` | `double`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`PosteriorMethod` — Posterior probability estimation method
`'kl'` (default) | `'qp'`

Posterior probability estimation method, specified as the comma-separated pair consisting of `'PosteriorMethod'` and `'kl'` or `'qp'`.

- To use this option, you must return the fourth output argument (`Posterior`) and the linear classification models that compose the ECOC models must use logistic regression learners (that is, `CVMdl.Trained{1}.BinaryLearners{1}.Learner` must be `'logistic'`).

- If `PosteriorMethod` is `'kl'`, then the software estimates multiclass posterior probabilities by minimizing the Kullback-Leibler divergence between the predicted and expected posterior probabilities returned by binary learners. For details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-3301.
- If `PosteriorMethod` is `'qp'`, then the software estimates multiclass posterior probabilities by solving a least-squares problem using quadratic programming. You need an Optimization Toolbox license to use this option. For details, see “Posterior Estimation Using Quadratic Programming” on page 33-3302.

Example: `'PosteriorMethod','qp'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`Label` — Cross-validated, predicted class labels
categorical array | character array | logical matrix | numeric matrix | cell array of character vectors

Cross-validated, predicted class labels, returned as a categorical or character array, logical or numeric matrix, or cell array of character vectors.

In most cases, `Label` is an *n*-by-*L* array of the same data type as the observed class labels (`Y`) used to create `CVMdl`. (The software treats string arrays as cell arrays of character vectors.) *n* is the number of observations in the predictor data (`X`) and *L* is the number of regularization strengths in the linear classification models that compose the cross-validated ECOC model. That is, `Label(i,j)` is the predicted class label for observation `i` using the ECOC model of linear classification models that has regularization strength `CVMdl.Trained{1}.BinaryLearners{1}.Lambda(j)`.

If `Y` is a character array and $L > 1$, then `Label` is a cell array of class labels.

The software assigns the predicted label corresponding to the class with the largest, negated, average binary loss (`NegLoss`), or, equivalently, the smallest average binary loss.

`NegLoss` — Cross-validated, negated, average binary losses
numeric array

Cross-validated, negated, average binary losses, returned as an *n*-by-*K*-by-*L* numeric matrix or array. *K* is the number of distinct classes in the training data and columns correspond to the classes in `CVMdl.ClassNames`. For *n* and *L*, see `Label`. `NegLoss(i,k,j)` is the negated, average binary loss for classifying observation `i` into class `k` using the linear classification model that has regularization strength `CVMdl.Trained{1}.BinaryLoss{1}.Lambda(j)`.

`PBScore` — Cross-validated, positive-class scores
numeric array

Cross-validated, positive-class scores, returned as an *n*-by-*B*-by-*L* numeric array. *B* is the number of binary learners in the cross-validated ECOC model and columns correspond to the binary learners in `CVMdl.Trained{1}.BinaryLearners`. For *n* and *L*, see `Label`. `PBScore(i,b,j)` is the positive-class score of binary learner *b* for classifying observation `i` into its positive class, using the linear classification model that has regularization strength `CVMdl.Trained{1}.BinaryLearners{1}.Lambda(j)`.

If the coding matrix varies across folds (that is, if the coding scheme is `sparserandom` or `denserandom`), then `PBScore` is empty (`[]`).

`Posterior` — Cross-validated posterior class probabilities
numeric array

Cross-validated posterior class probabilities, returned as an *n*-by-*K*-by-*L* numeric array. For dimension definitions, see `NegLoss`. `Posterior(i,k,j)` is the posterior probability for classifying observation `i` into class *k* using the linear classification model that has regularization strength `CVMdl.Trained{1}.BinaryLearners{1}.Lambda(j)`.

To return posterior probabilities, `CVMdl.Trained{1}.BinaryLearner{1}.Learner` must be `'logistic'`.

Examples

Predict *k*-fold Cross-Validation Labels

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

Cross-validate an ECOC model of linear classification models.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learner','linear','CrossVal','on');
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. By default, the software implements 10-fold cross validation.

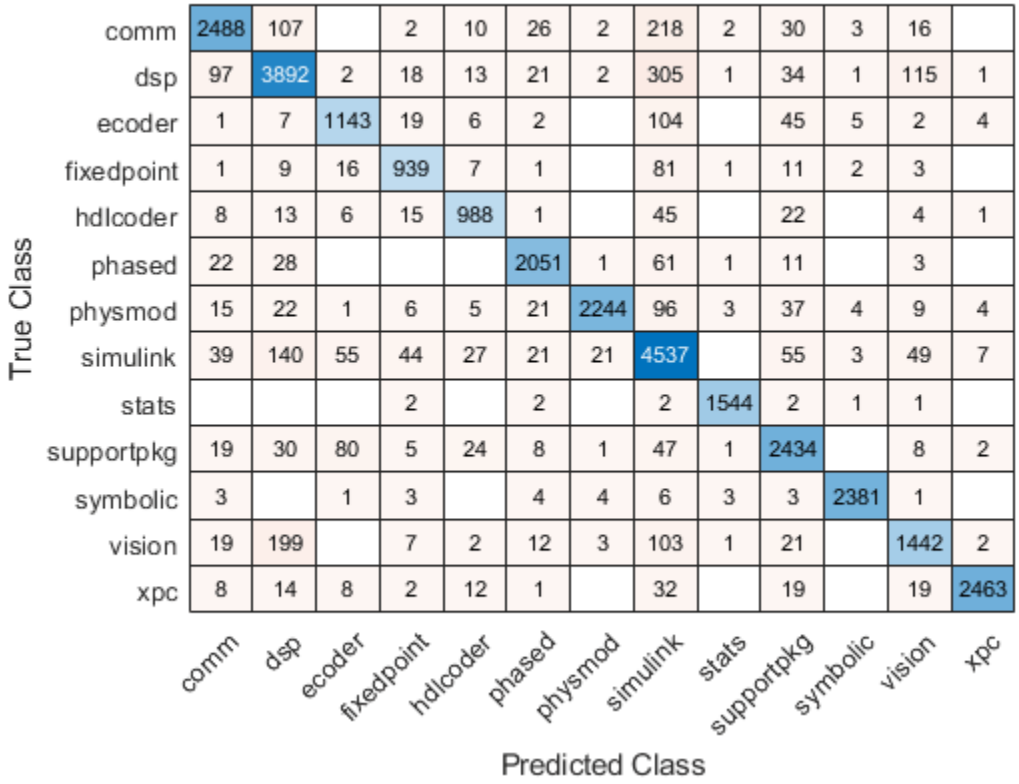
Predict labels for the observations that `fitcecoc` did not use in training the folds.

```
label = kfoldPredict(CVMdl);
```

Because there is one regularization strength in `CVMdl`, `label` is a column vector of predictions containing as many rows as observations in `X`.

Construct a confusion matrix.

```
cm = confusionchart(Y,label);
```



Specify Custom Binary Loss

Load the NLP data set. Transpose the predictor data.

```
load nlpdata
X = X';
```

For simplicity, use the label 'others' for all observations in Y that are not `'simulink'`, `'dsp'`, or `'comm'`.

```
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Create a linear classification model template that specifies optimizing the objective function using SpaRSA.

```
t = templateLinear('Solver','sparsa');
```

Cross-validate an ECOC model of linear classification models using 5-fold cross-validation. Specify that the predictor observations correspond to columns.

```
rng(1); % For reproducibility
CVMdl = fitcecoc(X,Y,'Learners',t,'KFold',5,'ObservationsIn','columns');
CMdl1 = CVMdl.Trained{1}
```

```
CMdl1 =
  CompactClassificationECOC
```

```
      ResponseName: 'Y'
        ClassNames: [comm    dsp    simulink    others]
    ScoreTransform: 'none'
    BinaryLearners: {6x1 cell}
      CodingMatrix: [4x6 double]


  Properties, Methods
```

`CVMdl` is a `ClassificationPartitionedLinearECOC` model. It contains the property `Trained`, which is a 5-by-1 cell array holding a `CompactClassificationECOC` models that the software trained using the training set of each fold.

By default, the linear classification models that compose the ECOC models use SVMs. SVM scores are signed distances from the observation to the decision boundary. Therefore, the domain is $(-\infty, \infty)$. Create a custom binary loss function that:

- Maps the coding design matrix (M) and positive-class classification scores (s) for each learner to the binary loss for each observation
- Uses linear loss
- Aggregates the binary learner loss using the median.

You can create a separate function for the binary loss function, and then save it on the MATLAB® path. Or, you can specify an anonymous binary loss function.

```
customBL = @(M,s)nanmedian(1 - bsxfun(@times,M,s),2)/2;
```

Predict cross-validation labels and estimate the median binary loss per class. Print the median negative binary losses per class for a random set of 10 out-of-fold observations.

```
[label,NegLoss] = kfoldPredict(CVMdl,'BinaryLoss',customBL);

idx = randsample(numel(label),10);
table(Y(idx),label(idx),NegLoss(idx,1),NegLoss(idx,2),NegLoss(idx,3),...
    NegLoss(idx,4),'VariableNames',[{'True'};{'Predicted'};...
    categories(CVMdl.ClassNames)])
```

```
ans=10×6 table
      True      Predicted      comm         dsp       simulink    others
    ________    _________    _________    ________    ________    _______

    others      others          -1.2319     -1.0488    0.048758     1.6175
    simulink    simulink        -16.407     -12.218      21.531     11.218
    dsp         dsp             -0.7387    -0.11534    -0.88466    -0.2613
    others      others          -0.1251     -0.8749    -0.99766    0.14517
    dsp         dsp              2.5867      6.4187     -3.5867    -4.4165
    others      others        -0.025358     -1.2287    -0.97464    0.19747
    others      others          -2.6725    -0.56708    -0.51092     2.7453
    others      others          -1.1605    -0.88321    -0.11679    0.43504
    others      others          -1.9511     -1.3175     0.24735    0.95111
    simulink    others           -7.848     -5.8203      4.8203     6.8457
```

The software predicts the label based on the maximum negated loss.

Estimate Posterior Class Probabilities

ECOC models composed of linear classification models return posterior probabilities for logistic regression learners only. This example requires the Parallel Computing Toolbox™ and the Optimization Toolbox™

Load the NLP data set and preprocess the data as in “Specify Custom Binary Loss” on page 33-3296.

```
load nlpdata
X = X';
Y(~(ismember(Y,{'simulink','dsp','comm'}))) = 'others';
```

Create a set of 5 logarithmically-spaced regularization strengths from 10^{-5} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,5);
```

Create a linear classification model template that specifies optimizing the objective function using SpaRSA and to use logistic regression learners.

```
t = templateLinear('Solver','sparsa','Learner','logistic','Lambda',Lambda);
```

Cross-validate an ECOC model of linear classification models using 5-fold cross-validation. Specify that the predictor observations correspond to columns, and to use parallel computing.

```
rng(1); % For reproducibility
Options = statset('UseParallel',true);
CVMdl = fitcecoc(X,Y,'Learners',t,'KFold',5,'ObservationsIn','columns',...
    'Options',Options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Predict the cross-validated posterior class probabilities. Specify to use parallel computing and to estimate posterior probabilities using quadratic programming.

```
[label,~,~,Posterior] = kfoldPredict(CVMdl,'Options',Options,...
    'PosteriorMethod','qp');
size(label)
label(3,4)
size(Posterior)
Posterior(3,:,4)
```

```
ans =

       31572           5


ans =

  categorical

     others


ans =
```

```
       31572           4           5

ans =

    0.0285    0.0373    0.1714    0.7627
```

Because there are five regularization strengths:

- `label` is a 31572-by-5 categorical array. `label(3,4)` is the predicted, cross-validated label for observation 3 using the model trained with regularization strength `Lambda(4)`.
- `Posterior` is a 31572-by-4-by-5 matrix. `Posterior(3,:,4)` is the vector of all estimated, posterior class probabilities for observation 3 using the model trained with regularization strength `Lambda(4)`. The order of the second dimension corresponds to `CVMdl.ClassNames`. Display a random set of 10 posterior class probabilities.

Display a random sample of cross-validated labels and posterior probabilities for the model trained using `Lambda(4)`.

```
idx = randsample(size(label,1),10);
table(Y(idx),label(idx,4),Posterior(idx,1,4),Posterior(idx,2,4),...
    Posterior(idx,3,4),Posterior(idx,4,4),...
    'VariableNames',[{'True'};{'Predicted'};categories(CVMdl.ClassNames)])
```

```
ans =

  10×6 table

      True       Predicted       comm          dsp         simulink      others
    _________    _________    __________    __________    _________    _________

    others       others          0.030275      0.022142      0.10416      0.84342
    simulink     simulink     3.4954e-05    4.2982e-05      0.99832    0.0016016
    dsp          others            0.15787       0.25718      0.18848      0.39647
    others       others          0.094177      0.062712      0.12921      0.71391
    dsp          dsp             0.0057979       0.89703    0.015098     0.082072
    others       others          0.086084      0.054836      0.086165     0.77292
    others       others          0.0062338     0.0060492     0.023816      0.9639
    others       others            0.06543      0.075097      0.17136      0.68812
    others       others          0.051843      0.025566      0.13299       0.7896
    simulink     simulink     0.00044059    0.00049753      0.70958      0.28948
```

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).

- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Algorithms

The software can estimate class posterior probabilities by minimizing the Kullback-Leibler divergence or by using quadratic programming. For the following descriptions of the posterior estimation algorithms, assume that:

- m_{kj} is the element (k,j) of the coding design matrix M.
- I is the indicator function.
- $\widehat{p}_k$ is the class posterior probability estimate for class k of an observation, $k = 1,...,K$.
- r_j is the positive-class posterior probability for binary learner j. That is, r_j is the probability that binary learner j classifies an observation into the positive class, given the training data.

Posterior Estimation Using Kullback-Leibler Divergence

By default, the software minimizes the Kullback-Leibler divergence to estimate class posterior probabilities. The Kullback-Leibler divergence between the expected and observed positive-class posterior probabilities is

$$\Delta(r,\widehat{r}) = \sum_{j=1}^{L} w_j \left[r_j \log\frac{r_j}{\widehat{r}_j} + (1-r_j)\log\frac{1-r_j}{1-\widehat{r}_j} \right],$$

where $w_j = \sum_{S_j} w_i^*$ is the weight for binary learner j.

- S_j is the set of observation indices on which binary learner j is trained.
- w_i^* is the weight of observation i.

The software minimizes the divergence iteratively. The first step is to choose initial values $\widehat{p}_k^{(0)};\ k = 1,...,K$ for the class posterior probabilities.

- If you do not specify `'NumKLIterations'`, then the software tries both sets of deterministic initial values described next, and selects the set that minimizes Δ.
  - $\widehat{p}_k^{(0)} = 1/K;\ k = 1,...,K$.
  - $\widehat{p}_k^{(0)};\ k = 1,...,K$ is the solution of the system

    $$M_{01}\widehat{p}^{(0)} = r,$$

    where M_{01} is M with all $m_{kj} = -1$ replaced with 0, and r is a vector of positive-class posterior probabilities returned by the L binary learners [Dietterich et al.] on page 18-279. The software uses `lsqnonneg` to solve the system.
- If you specify `'NumKLIterations',c`, where `c` is a natural number, then the software does the following to choose the set $\widehat{p}_k^{(0)};\ k = 1,...,K$, and selects the set that minimizes Δ.
  - The software tries both sets of deterministic initial values as described previously.
  - The software randomly generates `c` vectors of length K using `rand`, and then normalizes each vector to sum to 1.

At iteration t, the software completes these steps:

1 Compute

$$\widehat{r}_j^{(t)} = \frac{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1)}{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1 \cup m_{kj} = -1)}.$$

2 Estimate the next class posterior probability using

$$\widehat{p}_k^{(t+1)} = \widehat{p}_k^{(t)} \frac{\sum_{j=1}^{L} w_j \left[r_j I\left(m_{kj} = +1\right) + \left(1 - r_j\right) I\left(m_{kj} = -1\right) \right]}{\sum_{j=1}^{L} w_j \left[\widehat{r}_j^{(t)} I\left(m_{kj} = +1\right) + \left(1 - \widehat{r}_j^{(t)}\right) I\left(m_{kj} = -1\right) \right]}.$$

3 Normalize $\widehat{p}_k^{(t+1)};\ k = 1, \ldots, K$ so that they sum to 1.

4 Check for convergence.

For more details, see [Hastie et al.] on page 18-280 and [Zadrozny] on page 18-281.

Posterior Estimation Using Quadratic Programming

Posterior probability estimation using quadratic programming requires an Optimization Toolbox license. To estimate posterior probabilities for an observation using this method, the software completes these steps:

1 Estimate the positive-class posterior probabilities, r_j, for binary learners $j = 1,\ldots,L$.

2 Using the relationship between r_j and $\widehat{p}_k$ [Wu et al.] on page 18-281, minimize

$$\sum_{j=1}^{L} \left[-r_j \sum_{k=1}^{K} \widehat{p}_k I\left(m_{kj} = -1\right) + \left(1 - r_j\right) \sum_{k=1}^{K} \widehat{p}_k I\left(m_{kj} = +1\right) \right]^2$$

with respect to $\widehat{p}_k$ and the restrictions

$$0 \le \widehat{p}_k \le 1$$

$$\sum_k \widehat{p}_k = 1.$$

The software performs minimization using `quadprog`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

[5] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[6] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[7] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `ClassificationLinear` | `ClassificationPartitionedLinearECOC` | `confusionchart` | `fitcecoc` | `perfcurve` | `predict` | `statset` | `testcholdout`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2016a

kfoldPredict

Package: `classreg.learning.partition`

Classify observations in cross-validated classification model

Syntax

```
label = kfoldPredict(CVMdl)
label = kfoldPredict(CVMdl,'IncludeInteractions',includeInteractions)
[label,Score] = kfoldPredict( ___ )
[label,Score,Cost] = kfoldPredict(CVMdl)
```

Description

`label = kfoldPredict(CVMdl)` returns class labels predicted by the cross-validated classifier `CVMdl`. For every fold, `kfoldPredict` predicts class labels for validation-fold observations using a classifier trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`label = kfoldPredict(CVMdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

`[label,Score] = kfoldPredict( ___ )` additionally returns the predicted classification scores for validation-fold observations using a classifier trained on training-fold observations, with any of the input argument in the previous syntaxes.

`[label,Score,Cost] = kfoldPredict(CVMdl)` additionally returns the expected misclassification costs for discriminant analysis, *k*-nearest neighbor, naive Bayes, and tree classifiers.

Examples

Create Confusion Matrix Using Cross-Validation Predictions

Create a confusion matrix using the 10-fold cross-validation predictions of a discriminant analysis model.

Load the `fisheriris` data set. `X` contains flower measurements for 150 different flowers, and `y` lists the species, or class, for each flower. Create a variable `order` that specifies the order of the classes.

```
load fisheriris
X = meas;
y = species;
order = unique(y)
```

```
order = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

Create a 10-fold cross-validated discriminant analysis model by using the `fitcdiscr` function. By default, `fitcdiscr` ensures that training and test sets have roughly the same proportions of flower species. Specify the order of the flower classes.

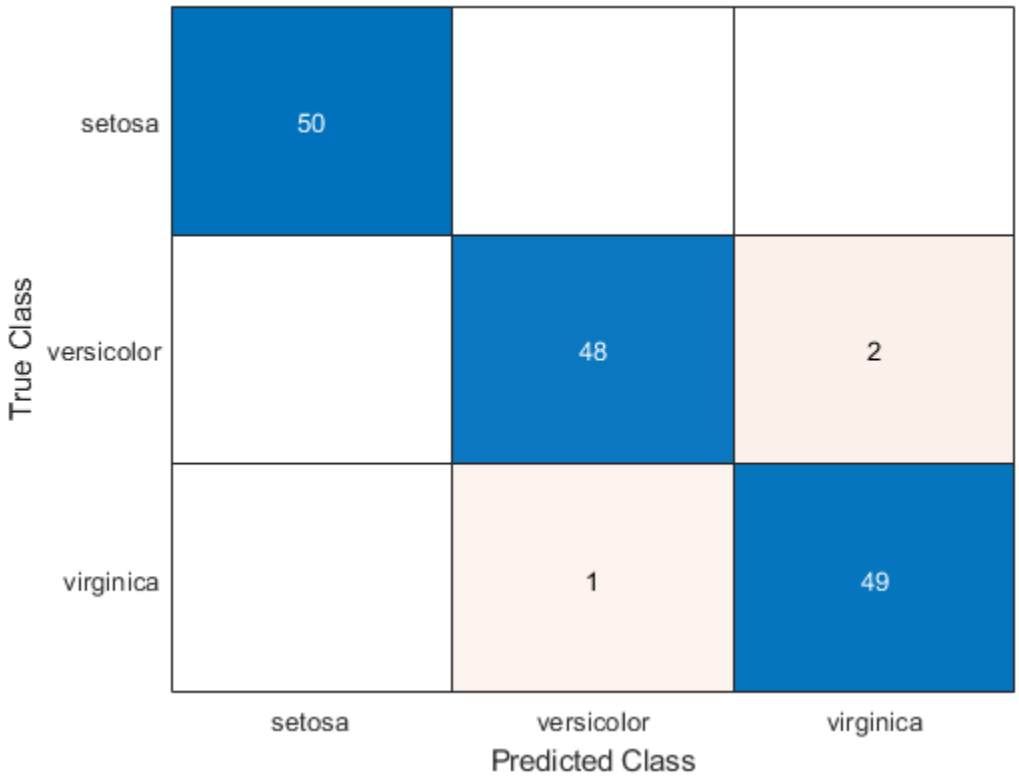
```
cvmdl = fitcdiscr(X,y,'KFold',10,'ClassNames',order);
```

Predict the species of the test set flowers.

```
predictedSpecies = kfoldPredict(cvmdl);
```

Create a confusion matrix that compares the true class values to the predicted class values.

```
confusionchart(y,predictedSpecies)
```



Estimate Cross-Validation Predictions from Ensemble

Find the cross-validation predictions for a model based on Fisher's iris data.

Load Fisher's iris data set.

```
load fisheriris
```

Train an ensemble of classification trees using AdaBoostM2. Specify tree stumps as the weak learners.

```
rng(1); % For reproducibility
t = templateTree('MaxNumSplits',1);
Mdl = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Cross-validate the trained ensemble using 10-fold cross-validation.

```
CVMdl = crossval(Mdl);
```

Estimate cross-validation predicted labels and scores.

```
[elabel,escore] = kfoldPredict(CVMdl);
```

Display the maximum and minimum scores of each class.

```
max(escore)
```

```
ans = 1×3

    9.3862    8.9871   10.1866
```

```
min(escore)
```

```
ans = 1×3

    0.0018    3.8359    0.9573
```

Input Arguments

`CVMdl` — Cross-validated partitioned classifier

`ClassificationPartitionedModel` object | `ClassificationPartitionedEnsemble` object | `ClassificationPartitionedGAM` object

Cross-validated partitioned classifier, specified as a `ClassificationPartitionedModel`, `ClassificationPartitionedEnsemble`, or `ClassificationPartitionedGAM` object. You can create the object in two ways:

- Pass a trained classification model listed in the following table to its `crossval` object function.
- Train a classification model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Classification Model	Function
`ClassificationDiscriminant`	`fitcdiscr`
`ClassificationEnsemble`	`fitcensemble`
`ClassificationGAM`	`fitcgam`
`ClassificationKNN`	`fitcknn`
`ClassificationNaiveBayes`	`fitcnb`
`ClassificationNeuralNetwork`	`fitcnet`
`ClassificationSVM`	`fitcsvm`
`ClassificationTree`	`fitctree`

`includeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `ClassificationPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Data Types: `logical`

Output Arguments

`label` — Predicted class labels
categorical vector | logical vector | numeric vector | character array | cell array of character vectors

Predicted class labels, returned as a categorical vector, logical vector, numeric vector, character array, or cell array of character vectors. `label` has the same data type and number of rows as `CVMdl.Y`. Each entry of `label` corresponds to the predicted class label for the corresponding observation in `CVMdl.X`.

If you use a holdout validation technique to create `CVMdl` (that is, if `CVMdl.KFold` is `1`), then ignore the `label` values for training-fold observations. These values match the class with the highest frequency.

`Score` — Classification scores
numeric matrix

Classification scores, returned as an *n*-by-*K* matrix, where *n* is the number of observations (`size(CVMdl.X,1)` when observations are in rows) and *K* is the number of unique classes (`size(CVMdl.ClassNames,1)`). The classification score `Score(i,j)` represents the confidence that the `i`th observation belongs to class `j`.

If you use a holdout validation technique to create `CVMdl` (that is, if `CVMdl.KFold` is `1`), then `Score` has `NaN` values for training-fold observations.

`Cost` — Expected misclassification costs
numeric matrix

Expected misclassification costs, returned as an *n*-by-*K* matrix, where *n* is the number of observations (`size(CVMdl.X,1)` when observations are in rows) and *K* is the number of unique classes (`size(CVMdl.ClassNames,1)`). The value `Cost(i,j)` is the average misclassification cost of predicting that the `i`th observation belongs to class `j`.

Note If you want to return this output argument, `CVMdl` must be a discriminant analysis, *k*-nearest neighbor, naive Bayes, or tree classifier.

If you use a holdout validation technique to create `CVMdl` (that is, if `CVMdl.KFold` is `1`), then `Cost` has `NaN` values for training-fold observations.

Algorithms

`kfoldPredict` computes predictions as described in the corresponding `predict` object function. For a model-specific description, see the appropriate `predict` function reference page in the following table.

Model Type	`predict` Function
Discriminant analysis classifier	`predict`
Ensemble classifier	`predict`
Generalized additive model classifier	`predict`
k-nearest neighbor classifier	`predict`
Naive Bayes classifier	`predict`
Neural network classifier	`predict`
Support vector machine classifier	`predict`
Binary decision tree for multiclass classification	`predict`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports *k*-nearest neighbor and SVM model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationPartitionedModel` | `kfoldEdge` | `kfoldLoss` | `kfoldMargin` | `kfoldfun`

Introduced in R2011a

kfoldPredict

Predict responses for observations not used for training

Syntax

```
YHat = kfoldPredict(CVMdl)
```

Description

`YHat = kfoldPredict(CVMdl)` returns cross-validated predicted responses by the cross-validated linear regression model `CVMdl`. That is, for every fold, `kfoldPredict` predicts responses for observations that it holds out when it trains using all other observations.

`YHat` contains predicted responses for each regularization strength in the linear regression models that compose `CVMdl`.

Input Arguments

`CVMdl` — Cross-validated, linear regression model

`RegressionPartitionedLinear` model object

Cross-validated, linear regression model, specified as a `RegressionPartitionedLinear` model object. You can create a `RegressionPartitionedLinear` model using `fitrlinear` and specifying any of the one of the cross-validation, name-value pair arguments, for example, `CrossVal`.

To obtain estimates, kfoldPredict applies the same data used to cross-validate the linear regression model (`X` and `Y`).

Output Arguments

`YHat` — Cross-validated predicted responses

numeric array

Cross-validated predicted responses, returned as an n-by-L numeric array. n is the number of observations in the predictor data that created `CVMdl` (see `X`) and L is the number of regularization strengths in `CVMdl.Trained{1}.Lambda`. `YHat(`i`,`j`)` is the predicted response for observation i using the linear regression model that has regularization strength `CVMdl.Trained{1}.Lambda(`j`)`.

The predicted response using the model with regularization strength j is $\widehat{y}_j = x\beta_j + b_j$.

- x is an observation from the predictor data matrix `X`, and is row vector.
- β_j is the estimated column vector of coefficients. The software stores this vector in `Mdl.Beta(:,`j`)`.
- b_j is the estimated, scalar bias, which the software stores in `Mdl.Bias(`j`)`.

Examples

Predict Cross-Validated Responses

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Cross-validate a linear regression model.

```
CVMdl = fitrlinear(X,Y,'CrossVal','on')
```

```
CVMdl =
  RegressionPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 10000
                  KFold: 10
              Partition: [1x1 cvpartition]
      ResponseTransform: 'none'


  Properties, Methods
```

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: 0.0107
               Lambda: 1.1111e-04
              Learner: 'svm'


  Properties, Methods
```

By default, `fitrlinear` implements 10-fold cross-validation. `CVMdl` is a `RegressionPartitionedLinear` model. It contains the property `Trained`, which is a 10-by-1 cell array holding 10 `RegressionLinear` models that the software trained using the training set.

Predict responses for observations that `fitrlinear` did not use in training the folds.

```
yHat = kfoldPredict(CVMdl);
```

Because there is one regularization strength in `Mdl`, `yHat` is a numeric vector.

Predict for Models Containing Several Regularization Strengths

Simulate 10000 observations as in “Predict Cross-Validated Responses” on page 33-3309.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-5} through 10^{-1}.

```
Lambda = logspace(-5,-1,15);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Specify using least squares with a lasso penalty and optimizing the objective function using SpaRSA.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','KFold',5,'Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');
```

`CVMdl` is a `RegressionPartitionedLinear` model. Its `Trained` property contains a 5-by-1 cell array of trained `RegressionLinear` models, each one holds out a different fold during training. Because `fitrlinear` trained using 15 regularization strengths, you can think of each `RegressionLinear` model as 15 models.

Predict cross-validated responses.

```
YHat = kfoldPredict(CVMdl);
size(YHat)
```

```
ans = 1×2

       10000          15
```

```
YHat(2,:)
```

```
ans = 1×15

   -1.7338   -1.7332   -1.7319   -1.7299   -1.7266   -1.7239   -1.7135   -1.7210   -1.7324   -1.7
```

`YHat` is a 10000-by-15 matrix. `YHat(2,:)` is the cross-validated response for observation 2 using the model regularized with all 15 regularization values.

See Also

`RegressionLinear` | `RegressionPartitionedLinear` | `fitrlinear` | `predict`

Introduced in R2016a

kfoldPredict

Package: `classreg.learning.partition`

Predict responses for observations in cross-validated regression model

Syntax

```
yfit = kfoldPredict(CVMdl)
yfit = kfoldPredict(CVMdl,'IncludeInteractions',includeInteractions)
```

Description

`yfit = kfoldPredict(CVMdl)` returns responses predicted by the cross-validated regression model `CVMdl`. For every fold, `kfoldPredict` predicts the responses for validation-fold observations using a model trained on training-fold observations. `CVMdl.X` and `CVMdl.Y` contain both sets of observations.

`yfit = kfoldPredict(CVMdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

Examples

Compute Cross-Validation Loss Manually

When you create a cross-validated regression model, you can compute the mean squared error (MSE) by using the `kfoldLoss` object function. Alternatively, you can predict responses for validation-fold observations using `kfoldPredict` and compute the MSE manually.

Load the `carsmall` data set. Specify the predictor data `X` and the response data `Y`.

```
load carsmall
X = [Cylinders Displacement Horsepower Weight];
Y = MPG;
```

Train a cross-validated regression tree model. By default, the software implements 10-fold cross-validation.

```
rng('default') % For reproducibility
CVMdl = fitrtree(X,Y,'CrossVal','on');
```

Compute the 10-fold cross-validation MSE by using `kfoldLoss`.

```
L = kfoldLoss(CVMdl)
```

```
L = 29.4963
```

Predict the responses `yfit` by using the cross-validated regression model. Compute the mean squared error between `yfit` and the true responses `CVMdl.Y`. The computed MSE matches the loss value returned by `kfoldLoss`.

```
yfit = kfoldPredict(CVMdl);
mse = mean((yfit - CVMdl.Y).^2)
```

```
mse = 29.4963
```

Input Arguments

`CVMdl` — Cross-validated partitioned regression model

`RegressionPartitionedModel` object | `RegressionPartitionedEnsemble` object | `RegressionPartitionedGAM` object | `RegressionPartitionedSVM` object

Cross-validated partitioned regression model, specified as a `RegressionPartitionedModel`, `RegressionPartitionedEnsemble`, `RegressionPartitionedGAM`, or `RegressionPartitionedSVM` object. You can create the in two ways:

- Pass a trained regression model listed in the following table to its `crossval` object function.
- Train a regression model using a function listed in the following table and specify one of the cross-validation name-value arguments for the function.

Regression Model	Function
`RegressionEnsemble`	`fitrensemble`
`RegressionGAM`	`fitrgam`
`RegressionGP`	`fitrgp`
`RegressionNeuralNetwork`	`fitrnet`
`RegressionSVM`	`fitrsvm`
`RegressionTree`	`fitrtree`

`includeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `CVMdl` is `RegressionPartitionedGAM`.

The default value is `true` if the models in `CVMdl` (`CVMdl.Trained`) contain interaction terms. The value must be `false` if the models do not contain interaction terms.

Data Types: `logical`

Output Arguments

`yfit` — Predicted responses

numeric vector

Predicted responses, returned as an *n*-by-1 numeric vector, where *n* is the number of observations. (*n* is `size(CVMdl.X,1)` when observations are in rows.) Each entry of `yfit` corresponds to the predicted response for the corresponding row of `CVMdl.X`.

If you use a holdout validation technique to create `CVMdl` (that is, if `CVMdl.KFold` is `1`), then `yfit` has `NaN` values for training-fold observations.

See Also

`RegressionPartitionedEnsemble` | `RegressionPartitionedGAM` | `RegressionPartitionedModel` | `RegressionPartitionedSVM` | `kfoldLoss`

Introduced in R2011a

kmeans

k-means clustering

Syntax

```
idx = kmeans(X,k)
idx = kmeans(X,k,Name,Value)
[idx,C] = kmeans( ___ )
[idx,C,sumd] = kmeans( ___ )
[idx,C,sumd,D] = kmeans( ___ )
```

Description

`idx = kmeans(X,k)` performs *k*-means clustering on page 33-3329 to partition the observations of the *n*-by-*p* data matrix `X` into `k` clusters, and returns an *n*-by-1 vector (`idx`) containing cluster indices of each observation. Rows of `X` correspond to points and columns correspond to variables.

By default, `kmeans` uses the squared Euclidean distance metric and the *k*-means++ algorithm on page 33-3329 for cluster center initialization.

`idx = kmeans(X,k,Name,Value)` returns the cluster indices with additional options specified by one or more `Name,Value` pair arguments.

For example, specify the cosine distance, the number of times to repeat the clustering using new initial values, or to use parallel computing.

`[idx,C] = kmeans( ___ )` returns the `k` cluster centroid locations in the `k`-by-*p* matrix `C`.

`[idx,C,sumd] = kmeans( ___ )` returns the within-cluster sums of point-to-centroid distances in the `k`-by-1 vector `sumd`.

`[idx,C,sumd,D] = kmeans( ___ )` returns distances from each point to every centroid in the *n*-by-`k` matrix `D`.

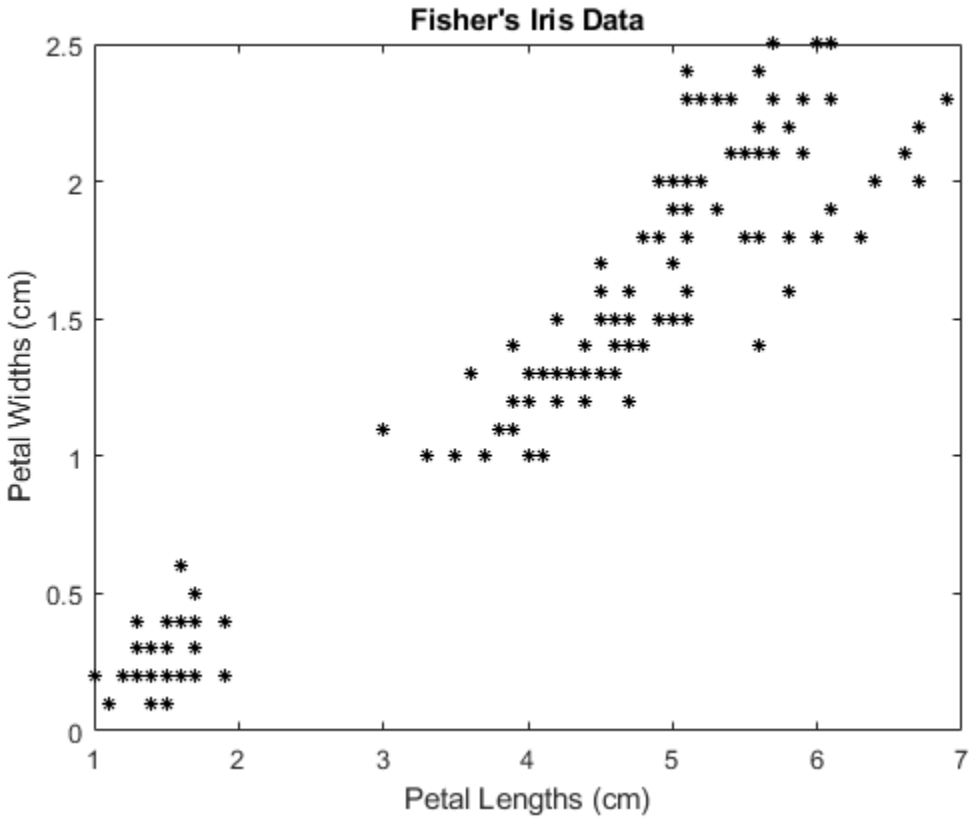
Examples

Train a *k*-Means Clustering Algorithm

Cluster data using *k*-means clustering, then plot the cluster regions.

Load Fisher's iris data set. Use the petal lengths and widths as predictors.

```
load fisheriris
X = meas(:,3:4);

figure;
plot(X(:,1),X(:,2),'k*','MarkerSize',5);
title 'Fisher''s Iris Data';
xlabel 'Petal Lengths (cm)';
ylabel 'Petal Widths (cm)';
```



The larger cluster seems to be split into a lower variance region and a higher variance region. This might indicate that the larger cluster is two, overlapping clusters.

Cluster the data. Specify $k = 3$ clusters.

```
rng(1); % For reproducibility
[idx,C] = kmeans(X,3);
```

`idx` is a vector of predicted cluster indices corresponding to the observations in `X`. `C` is a 3-by-2 matrix containing the final centroid locations.

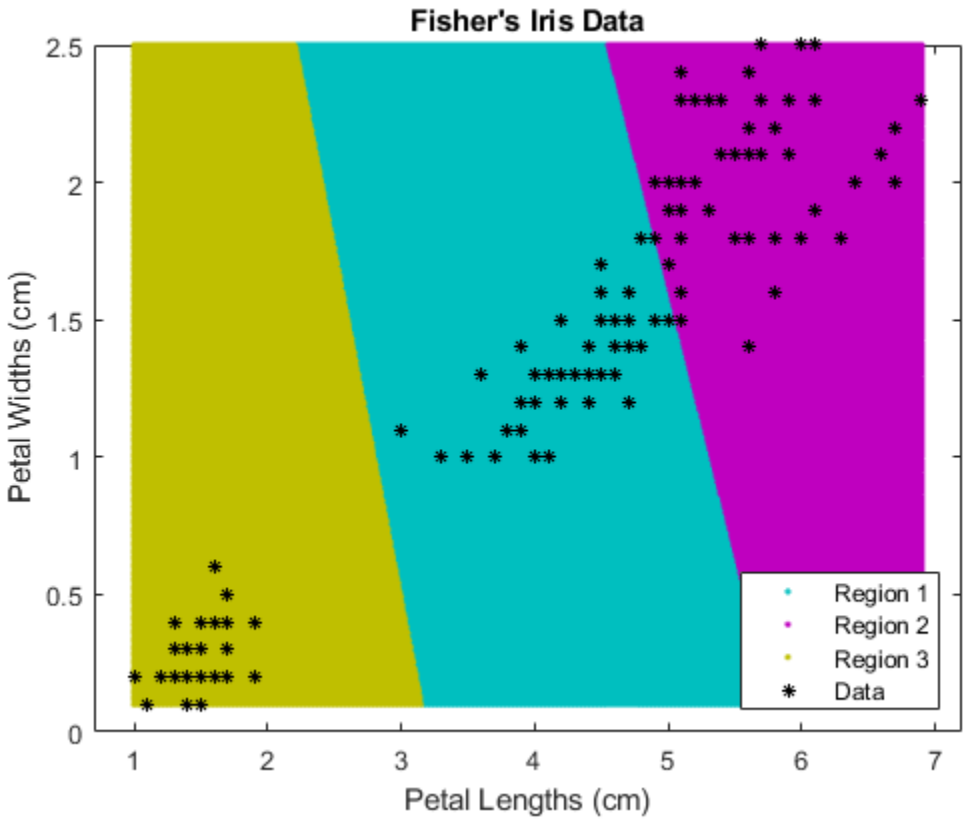
Use `kmeans` to compute the distance from each centroid to points on a grid. To do this, pass the centroids (`C`) and points on a grid to `kmeans`, and implement one iteration of the algorithm.

```
x1 = min(X(:,1)):0.01:max(X(:,1));
x2 = min(X(:,2)):0.01:max(X(:,2));
[x1G,x2G] = meshgrid(x1,x2);
XGrid = [x1G(:),x2G(:)]; % Defines a fine grid on the plot

idx2Region = kmeans(XGrid,3,'MaxIter',1,'Start',C);

Warning: Failed to converge in 1 iterations.

    % Assigns each node in the grid to the closest centroid
```

`kmeans` displays a warning stating that the algorithm did not converge, which you should expect since the software only implemented one iteration.

Plot the cluster regions.

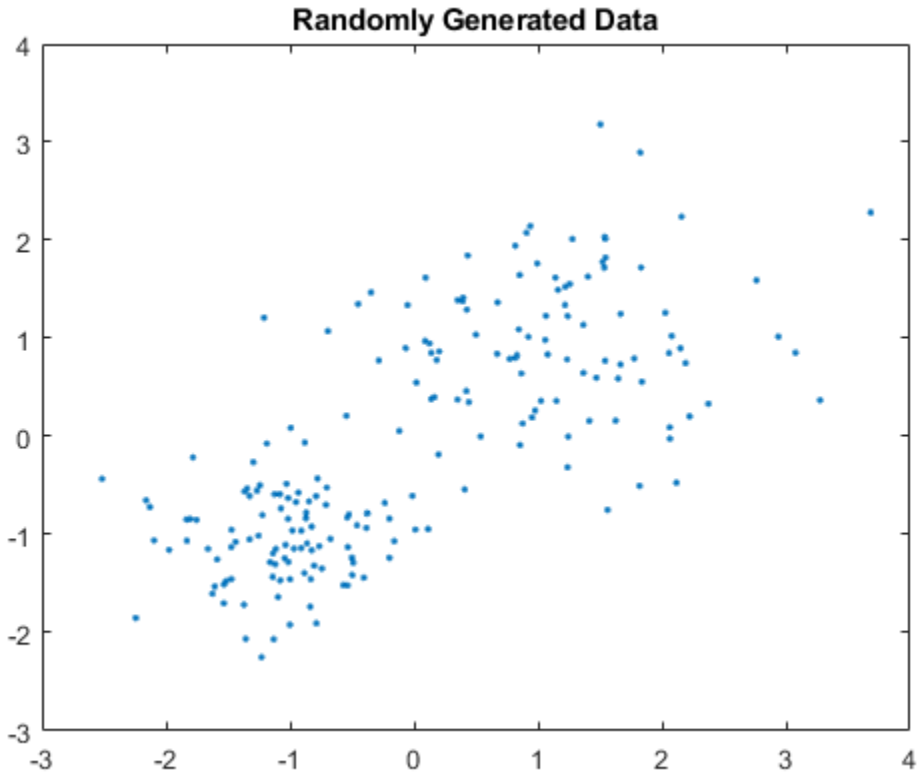
```
figure;
gscatter(XGrid(:,1),XGrid(:,2),idx2Region,...
    [0,0.75,0.75;0.75,0,0.75;0.75,0.75,0],'..');
hold on;
plot(X(:,1),X(:,2),'k*','MarkerSize',5);
title 'Fisher''s Iris Data';
xlabel 'Petal Lengths (cm)';
ylabel 'Petal Widths (cm)';
legend('Region 1','Region 2','Region 3','Data','Location','SouthEast');
hold off;
```



Partition Data into Two Clusters

Randomly generate the sample data.

```
rng default; % For reproducibility
X = [randn(100,2)*0.75+ones(100,2);
    randn(100,2)*0.5-ones(100,2)];

figure;
plot(X(:,1),X(:,2),'.');
title 'Randomly Generated Data';
```



There appears to be two clusters in the data.

Partition the data into two clusters, and choose the best arrangement out of five initializations. Display the final output.

```
opts = statset('Display','final');
[idx,C] = kmeans(X,2,'Distance','cityblock',...
    'Replicates',5,'Options',opts);
```

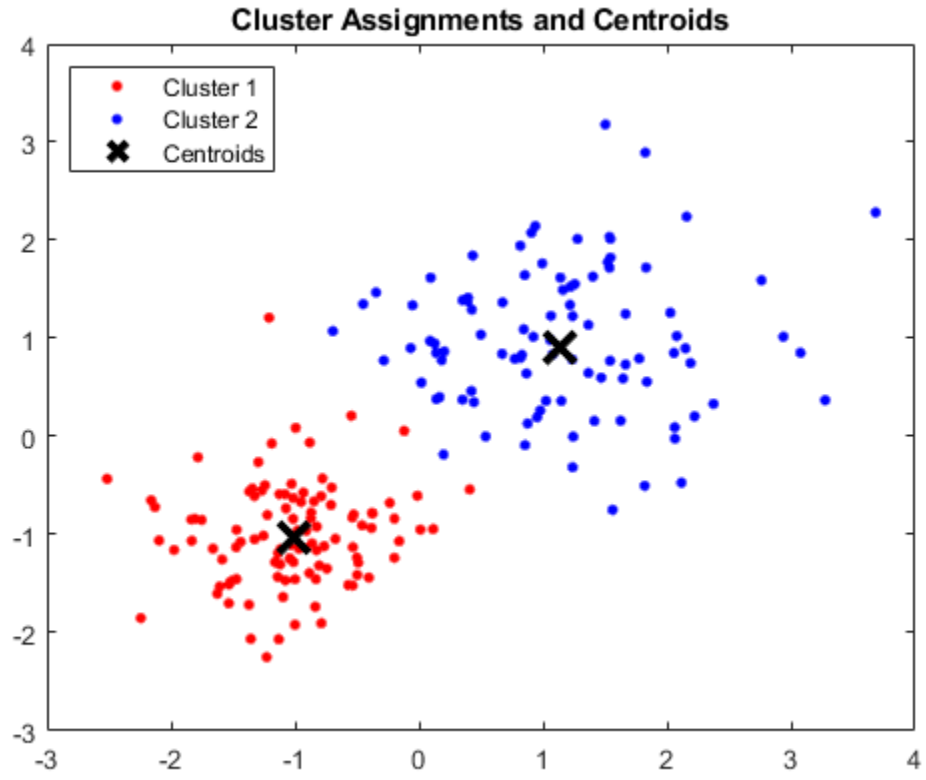
```
Replicate 1, 3 iterations, total sum of distances = 201.533.
Replicate 2, 5 iterations, total sum of distances = 201.533.
Replicate 3, 3 iterations, total sum of distances = 201.533.
Replicate 4, 3 iterations, total sum of distances = 201.533.
Replicate 5, 2 iterations, total sum of distances = 201.533.
Best total sum of distances = 201.533
```

By default, the software initializes the replicates separately using *k*-means++.

Plot the clusters and the cluster centroids.

```
figure;
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',12)
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',12)
plot(C(:,1),C(:,2),'kx',...
     'MarkerSize',15,'LineWidth',3)
legend('Cluster 1','Cluster 2','Centroids',...
       'Location','NW')
```

```
title 'Cluster Assignments and Centroids'
hold off
```



You can determine how well separated the clusters are by passing `idx` to `silhouette`.

Cluster Data Using Parallel Computing

Clustering large data sets might take time, particularly if you use online updates (set by default). If you have a Parallel Computing Toolbox ™ license and you set the options for parallel computing, then `kmeans` runs each clustering task (or replicate) in parallel. And, if `Replicates`>1, then parallel computing decreases time to convergence.

Randomly generate a large data set from a Gaussian mixture model.

```
Mu = bsxfun(@times,ones(20,30),(1:20)'); % Gaussian mixture mean
rn30 = randn(30,30);
Sigma = rn30'*rn30; % Symmetric and positive-definite covariance
Mdl = gmdistribution(Mu,Sigma); % Define the Gaussian mixture distribution

rng(1); % For reproducibility
X = random(Mdl,10000);
```

`Mdl` is a 30-dimensional `gmdistribution` model with 20 components. `X` is a 10000-by-30 matrix of data generated from `Mdl`.

Specify the options for parallel computing.

```
stream = RandStream('mlfg6331_64');  % Random number stream
options = statset('UseParallel',1,'UseSubstreams',1,...
    'Streams',stream);
```

The input argument `'mlfg6331_64'` of `RandStream` specifies to use the multiplicative lagged Fibonacci generator algorithm. `options` is a structure array with fields that specify options for controlling estimation.

Cluster the data using *k*-means clustering. Specify that there are $k = 20$ clusters in the data and increase the number of iterations. Typically, the objective function contains local minima. Specify 10 replicates to help find a lower, local minimum.

```
tic; % Start stopwatch timer
[idx,C,sumd,D] = kmeans(X,20,'Options',options,'MaxIter',10000,...
    'Display','final','Replicates',10);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
connected to 6 workers.
Replicate 5, 72 iterations, total sum of distances = 7.73161e+06.
Replicate 1, 64 iterations, total sum of distances = 7.72988e+06.
Replicate 3, 68 iterations, total sum of distances = 7.72576e+06.
Replicate 4, 84 iterations, total sum of distances = 7.72696e+06.
Replicate 6, 82 iterations, total sum of distances = 7.73006e+06.
Replicate 7, 40 iterations, total sum of distances = 7.73451e+06.
Replicate 2, 194 iterations, total sum of distances = 7.72953e+06.
Replicate 9, 105 iterations, total sum of distances = 7.72064e+06.
Replicate 10, 125 iterations, total sum of distances = 7.72816e+06.
Replicate 8, 70 iterations, total sum of distances = 7.73188e+06.
Best total sum of distances = 7.72064e+06
```

```
toc % Terminate stopwatch timer
```

```
Elapsed time is 61.915955 seconds.
```

The Command Window indicates that six workers are available. The number of workers might vary on your system. The Command Window displays the number of iterations and the terminal objective function value for each replicate. The output arguments contain the results of replicate 9 because it has the lowest total sum of distances.

Assign New Data to Existing Clusters and Generate C/C++ Code

`kmeans` performs *k*-means clustering to partition data into *k* clusters. When you have a new data set to cluster, you can create new clusters that include the existing data and the new data by using `kmeans`. The `kmeans` function supports C/C++ code generation, so you can generate code that accepts training data and returns clustering results, and then deploy the code to a device. In this workflow, you must pass training data, which can be of considerable size. To save memory on the device, you can separate training and prediction by using `kmeans` and `pdist2`, respectively.

Use `kmeans` to create clusters in MATLAB® and use `pdist2` in the generated code to assign new data to existing clusters. For code generation, define an entry-point function that accepts the cluster centroid positions and the new data set, and returns the index of the nearest cluster. Then, generate code for the entry-point function.

Generating C/C++ code requires MATLAB® Coder™.

Perform *k*-Means Clustering

Generate a training data set using three distributions.

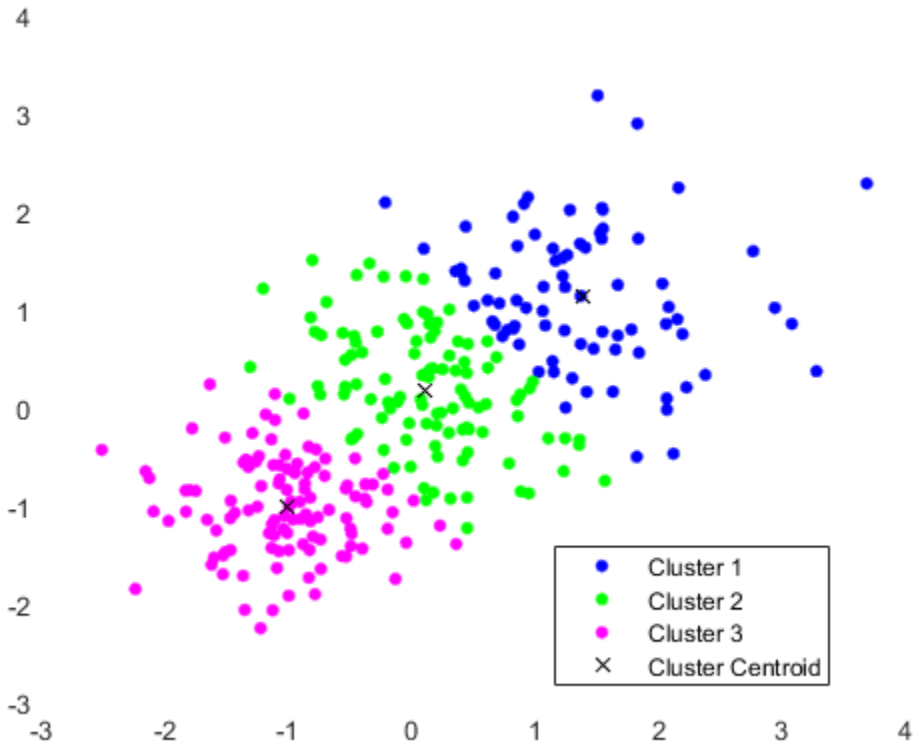
```
rng('default') % For reproducibility
X = [randn(100,2)*0.75+ones(100,2);
     randn(100,2)*0.5-ones(100,2);
     randn(100,2)*0.75];
```

Partition the training data into three clusters by using `kmeans`.

```
[idx,C] = kmeans(X,3);
```

Plot the clusters and the cluster centroids.

```
figure
gscatter(X(:,1),X(:,2),idx,'bgm')
hold on
plot(C(:,1),C(:,2),'kx')
legend('Cluster 1','Cluster 2','Cluster 3','Cluster Centroid')
```



Assign New Data to Existing Clusters

Generate a test data set.

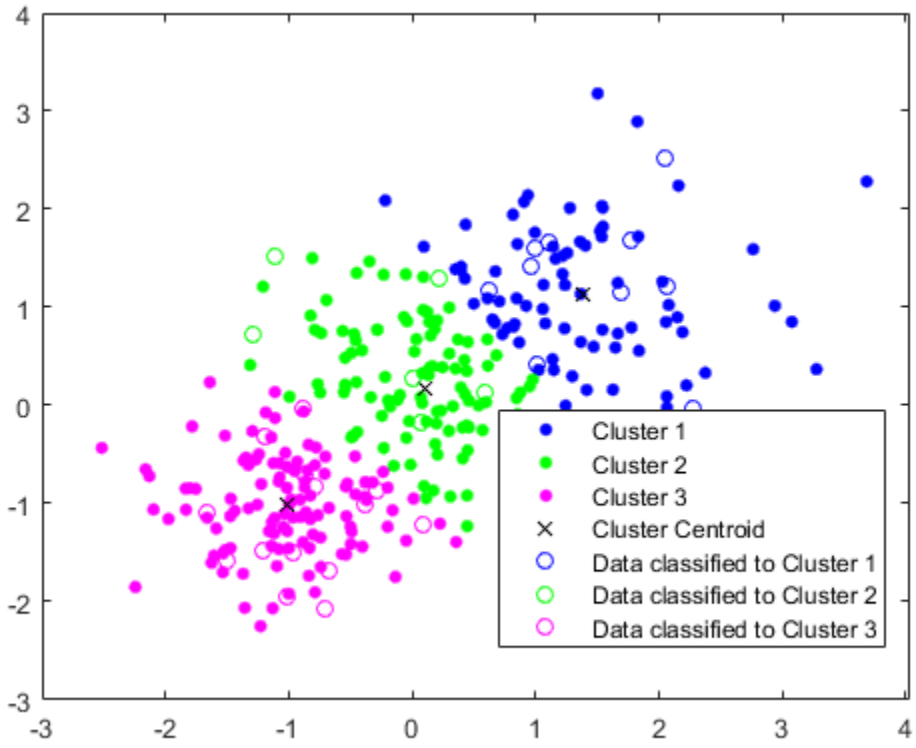
```
Xtest = [randn(10,2)*0.75+ones(10,2);
    randn(10,2)*0.5-ones(10,2);
    randn(10,2)*0.75];
```

Classify the test data set using the existing clusters. Find the nearest centroid from each test data point by using `pdist2`.

```
[~,idx_test] = pdist2(C,Xtest,'euclidean','Smallest',1);
```

Plot the test data and label the test data using `idx_test` by using `gscatter`.

```
gscatter(Xtest(:,1),Xtest(:,2),idx_test,'bgm','ooo')
legend('Cluster 1','Cluster 2','Cluster 3','Cluster Centroid', ...
    'Data classified to Cluster 1','Data classified to Cluster 2', ...
    'Data classified to Cluster 3')
```



Generate Code

Generate C code that assigns new data to the existing clusters. Note that generating C/C++ code requires MATLAB® Coder™.

Define an entry-point function named `findNearestCentroid` that accepts centroid positions and new data, and then find the nearest cluster by using `pdist2`.

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would cause errors during code generation.

```
type findNearestCentroid % Display contents of findNearestCentroid.m

function idx = findNearestCentroid(C,X) %#codegen
[~,idx] = pdist2(C,X,'euclidean','Smallest',1); % Find the nearest centroid
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate code by using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and array size of the inputs of `findNearestCentroid`, pass a MATLAB expression that represents the set of values with a certain data type and array size by using the `-args` option. For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45.

```
codegen findNearestCentroid -args {C,Xtest}

Code generation successful.
```

`codegen` generates the MEX function `findNearestCentroid_mex` with a platform-dependent extension.

Verify the generated code.

```
myIndx = findNearestCentroid(C,Xtest);
myIndex_mex = findNearestCentroid_mex(C,Xtest);
verifyMEX = isequal(idx_test,myIndx,myIndex_mex)

verifyMEX = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal. The comparison confirms that the `pdist2` function, the `findNearestCentroid` function, and the MEX function return the same index.

You can also generate optimized CUDA® code using GPU Coder™.

```
cfg = coder.gpuConfig('mex');
codegen -config cfg findNearestCentroid -args {C,Xtest}
```

For more information on code generation, see “General Code Generation Workflow” on page 32-5. For more information on GPU coder, see “Get Started with GPU Coder” (GPU Coder) and “Supported Functions” (GPU Coder).

Input Arguments

X — Data

numeric matrix

Data, specified as a numeric matrix. The rows of `X` correspond to observations, and the columns correspond to variables.

If `X` is a numeric vector, then `kmeans` treats it as an *n*-by-1 data matrix, regardless of its orientation.

The software treats NaNs in X as missing data and removes any row of X that contains at least one NaN. Removing rows of X reduces the sample size. The kmeans function returns NaN for the corresponding value in the output argument idx.

Data Types: single | double

k — Number of clusters
positive integer

Number of clusters in the data, specified as a positive integer.

Data Types: single | double

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example: 'Distance','cosine','Replicates',10,'Options',statset('UseParallel',1) specifies the cosine distance, 10 replicate clusters at different starting values, and to use parallel computing.

Display — Level of output to display
'off' (default) | 'final' | 'iter'

Level of output to display in the Command Window, specified as the comma-separated pair consisting of 'Display' and one of the following options:

- 'final' — Displays results of the final iteration
- 'iter' — Displays results of each iteration
- 'off' — Displays nothing

Example: 'Display','final'

Distance — Distance metric
'sqeuclidean' (default) | 'cityblock' | 'cosine' | 'correlation' | 'hamming'

Distance metric, in p-dimensional space, used for minimization, specified as the comma-separated pair consisting of 'Distance' and 'sqeuclidean', 'cityblock', 'cosine', 'correlation', or 'hamming'.

kmeans computes centroid clusters differently for the supported distance metrics. This table summarizes the available distance metrics. In the formulae, x is an observation (that is, a row of X) and c is a centroid (a row vector).

Distance Metric	Description	Formula
'sqeuclidean'	Squared Euclidean distance (default). Each centroid is the mean of the points in that cluster.	$d(x,c) = (x-c)(x-c)'$

Distance Metric	Description	Formula
`'cityblock'`	Sum of absolute differences, i.e., the $L1$ distance. Each centroid is the component-wise median of the points in that cluster.	$d(x,c) = \sum_{j=1}^{p} \left\lvert x_j - c_j \right\rvert$
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors). Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length.	$d(x,c) = 1 - \frac{xc'}{\sqrt{(xx')(cc')}}$
`'correlation'`	One minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to zero mean and unit standard deviation.	$d(x,c) = 1 - \frac{\left(x-\overrightarrow{\bar{x}}\right)\left(c-\overrightarrow{\bar{c}}\right)'}{\sqrt{\left(x-\overrightarrow{\bar{x}}\right)\left(x-\overrightarrow{\bar{x}}\right)'}\sqrt{\left(c-\overrightarrow{\bar{c}}\right)\left(c-\overrightarrow{\bar{c}}\right)'}},$ where • $\overrightarrow{\bar{x}} = \frac{1}{p}\left(\sum_{j=1}^{p} x_j\right)\overrightarrow{1}_p$ • $\overrightarrow{\bar{c}} = \frac{1}{p}\left(\sum_{j=1}^{p} c_j\right)\overrightarrow{1}_p$ • $\overrightarrow{1}_p$ is a row vector of p ones.
`'hamming'`	This metric is only suitable for binary data. It is the proportion of bits that differ. Each centroid is the component-wise median of points in that cluster.	$d(x,y) = \frac{1}{p}\sum_{j=1}^{p} I\{x_j \neq y_j\},$ where I is the indicator function.

Example: `'Distance','cityblock'`

`EmptyAction` — Action to take if cluster loses all member observations
`'singleton'` (default) | `'error'` | `'drop'`

Action to take if a cluster loses all its member observations, specified as the comma-separated pair consisting of `'EmptyAction'` and one of the following options.

Value	Description
`'error'`	Treat an empty cluster as an error.

Value	Description
'drop'	Remove any clusters that become empty. kmeans sets the corresponding return values in C and D to NaN.
'singleton'	Create a new cluster consisting of the one point furthest from its centroid (default).

Example: 'EmptyAction','error'

MaxIter — Maximum number of iterations
100 (default) | positive integer

Maximum number of iterations, specified as the comma-separated pair consisting of 'MaxIter' and a positive integer.

Example: 'MaxIter',1000

Data Types: double | single

OnlinePhase — Online update flag
'off' (default) | 'on'

Online update flag, specified as the comma-separated pair consisting of 'OnlinePhase' and 'off' or 'on'.

If OnlinePhase is on, then kmeans performs an online update phase in addition to a batch update phase. The online phase can be time consuming for large data sets, but guarantees a solution that is a local minimum of the distance criterion. In other words, the software finds a partition of the data in which moving any single point to a different cluster increases the total sum of distances.

Example: 'OnlinePhase','on'

Options — Options for controlling iterative algorithm for minimizing fitting criteria
[] (default) | structure array returned by statset

Options for controlling the iterative algorithm for minimizing the fitting criteria, specified as the comma-separated pair consisting of 'Options' and a structure array returned by statset. Supported fields of the structure array specify options for controlling the iterative algorithm.

This table summarizes the supported fields. Note that the supported fields require Parallel Computing Toolbox.

Field	Description
`'Streams'`	A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `kmeans` uses the default stream or streams. If you specify `Streams`, use a single object except when all of the following conditions exist: • You have an open parallel pool. • `UseParallel` is `true`. • `UseSubstreams` is `false`. In this case, use a cell array the same size as the parallel pool. If a parallel pool is not open, then `Streams` must supply a single random number stream.
`'UseParallel'`	• If `true` and `Replicates` > 1, then `kmeans` implements the *k*-means algorithm on each replicate in parallel. • If Parallel Computing Toolbox is not installed, then computation occurs in serial mode. The default is `false`, indicating serial computation.
`'UseSubstreams'`	Set to `true` to compute in parallel in a reproducible fashion. The default is `false`. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.

To ensure more predictable results, use `parpool` and explicitly create a parallel pool before invoking `kmeans` and setting `'Options',statset('UseParallel',1)`.

Example: `'Options',statset('UseParallel',1)`

Data Types: `struct`

`Replicates` — Number of times to repeat clustering using new initial cluster centroid positions
1 (default) | positive integer

Number of times to repeat clustering using new initial cluster centroid positions, specified as the comma-separated pair consisting of `'Replicates'` and an integer. `kmeans` returns the solution with the lowest `sumd`.

You can set `'Replicates'` implicitly by supplying a 3-D array as the value for the `'Start'` name-value pair argument.

Example: `'Replicates',5`

Data Types: `double` | `single`

`Start` — Method for choosing initial cluster centroid positions
`'plus'` (default) | `'cluster'` | `'sample'` | `'uniform'` | numeric matrix | numeric array

Method for choosing initial cluster centroid positions (or seeds), specified as the comma-separated pair consisting of `'Start'` and `'cluster'`, `'plus'`, `'sample'`, `'uniform'`, a numeric matrix, or a numeric array. This table summarizes the available options for choosing seeds.

Value	Description
`'cluster'`	Perform a preliminary clustering phase on a random 10% subsample of `X` when the number of observations in the subsample is greater than `k`. This preliminary phase is itself initialized using `'sample'`. If the number of observations in the random 10% subsample is less than `k`, then the software selects `k` observations from `X` at random.
`'plus'` (default)	Select `k` seeds by implementing the *k*-means++ algorithm on page 33-3329 for cluster center initialization.
`'sample'`	Select `k` observations from `X` at random.
`'uniform'`	Select `k` points uniformly at random from the range of `X`. Not valid with the Hamming distance.
numeric matrix	`k`-by-*p* matrix of centroid starting locations. The rows of `Start` correspond to seeds. The software infers `k` from the first dimension of `Start`, so you can pass in `[]` for `k`.
numeric array	`k`-by-*p*-by-*r* array of centroid starting locations. The rows of each page correspond to seeds. The third dimension invokes replication of the clustering routine. Page *j* contains the set of seeds for replicate *j*. The software infers the number of replicates (specified by the `'Replicates'` name-value pair argument) from the size of the third dimension.

Example: `'Start','sample'`

Data Types: `char` | `string` | `double` | `single`

Output Arguments

`idx` — Cluster indices

numeric column vector

Cluster indices, returned as a numeric column vector. `idx` has as many rows as `X`, and each row indicates the cluster assignment of the corresponding observation.

`C` — Cluster centroid locations

numeric matrix

Cluster centroid locations, returned as a numeric matrix. `C` is a `k`-by-*p* matrix, where row *j* is the centroid of cluster *j*.

`sumd` — Within-cluster sums of point-to-centroid distances
numeric column vector

Within-cluster sums of point-to-centroid distances, returned as a numeric column vector. `sumd` is a `k`-by-1 vector, where element *j* is the sum of point-to-centroid distances within cluster *j*. By default, `kmeans` uses the squared Euclidean distance (see `'Distance'` metrics).

`D` — Distances from each point to every centroid
numeric matrix

Distances from each point to every centroid, returned as a numeric matrix. `D` is an *n*-by-`k` matrix, where element (*j*,*m*) is the distance from observation *j* to centroid *m*. By default, `kmeans` uses the squared Euclidean distance (see `'Distance'` metrics).

More About

k-Means Clustering

k-means clustering, or Lloyd's algorithm [2], is an iterative, data-partitioning algorithm that assigns *n* observations to exactly one of *k* clusters defined by centroids, where *k* is chosen before the algorithm starts.

The algorithm proceeds as follows:

1. Choose *k* initial cluster centers (centroid). For example, choose *k* observations at random (by using `'Start','sample'`) or use the *k*-means ++ algorithm on page 33-3329 for cluster center initialization (the default).
2. Compute point-to-cluster-centroid distances of all observations to each centroid.
3. There are two ways to proceed (specified by `OnlinePhase`):
   - Batch update — Assign each observation to the cluster with the closest centroid.
   - Online update — Individually assign observations to a different centroid if the reassignment decreases the sum of the within-cluster, sum-of-squares point-to-cluster-centroid distances.

   For more details, see "Algorithms" on page 33-3330.
4. Compute the average of the observations in each cluster to obtain *k* new centroid locations.
5. Repeat steps 2 through 4 until cluster assignments do not change, or the maximum number of iterations is reached.

k-means++ Algorithm

The *k*-means++ algorithm uses an heuristic to find centroid seeds for *k*-means clustering. According to Arthur and Vassilvitskii [1], *k*-means++ improves the running time of Lloyd's algorithm, and the quality of the final solution.

The *k*-means++ algorithm chooses seeds as follows, assuming the number of clusters is *k*.

1. Select an observation uniformly at random from the data set, *X*. The chosen observation is the first centroid, and is denoted c_1.
2. Compute distances from each observation to c_1. Denote the distance between c_j and the observation m as $d(x_m, c_j)$.

3 Select the next centroid, c_2 at random from X with probability

$$\frac{d^2(x_m, c_1)}{\sum_{j=1}^{n} d^2(x_j, c_1)}.$$

4 To choose center j:

   a Compute the distances from each observation to each centroid, and assign each observation to its closest centroid.

   b For $m = 1,...,n$ and $p = 1,...,j - 1$, select centroid j at random from X with probability

$$\frac{d^2(x_m, c_p)}{\sum_{\{h;\, x_h \in C_p\}} d^2(x_h, c_p)},$$

   where C_p is the set of all observations closest to centroid c_p and x_m belongs to C_p.

   That is, select each subsequent center with a probability proportional to the distance from itself to the closest center that you already chose.

5 Repeat step 4 until k centroids are chosen.

Arthur and Vassilvitskii [1] demonstrate, using a simulation study for several cluster orientations, that k-means++ achieves faster convergence to a lower sum of within-cluster, sum-of-squares point-to-cluster-centroid distances than Lloyd's algorithm.

Algorithms

- `kmeans` uses a two-phase iterative algorithm to minimize the sum of point-to-centroid distances, summed over all `k` clusters.

  1 This first phase uses batch updates, where each iteration consists of reassigning points to their nearest cluster centroid, all at once, followed by recalculation of cluster centroids. This phase occasionally does not converge to solution that is a local minimum. That is, a partition of the data where moving any single point to a different cluster increases the total sum of distances. This is more likely for small data sets. The batch phase is fast, but potentially only approximates a solution as a starting point for the second phase.

  2 This second phase uses online updates, where points are individually reassigned if doing so reduces the sum of distances, and cluster centroids are recomputed after each reassignment. Each iteration during this phase consists of one pass though all the points. This phase converges to a local minimum, although there might be other local minima with lower total sum of distances. In general, finding the global minimum is solved by an exhaustive choice of starting points, but using several replicates with random starting points typically results in a solution that is a global minimum.

- If `Replicates` = $r > 1$ and `Start` is `plus` (the default), then the software selects r possibly different sets of seeds according to the k-means++ algorithm on page 33-3329.
- If you enable the `UseParallel` option in `Options` and `Replicates` > 1, then each worker selects seeds and clusters in parallel.

References

[1] Arthur, David, and Sergi Vassilvitskii. “K-means++: The Advantages of Careful Seeding.” *SODA ‘07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. 2007, pp. 1027–1035.

[2] Lloyd, Stuart P. “Least Squares Quantization in PCM.” *IEEE Transactions on Information Theory*. Vol. 28, 1982, pp. 129–137.

[3] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[4] Spath, H. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*. Translated by J. Goldschmidt. New York: Halsted Press, 1985.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Supported syntaxes are:
  - `idx = kmeans(X,k)`
  - `[idx,C] = kmeans(X,k)`
  - `[idx,C,sumd] = kmeans(X,k)`
  - `[___] = kmeans(___,Name,Value)`
- Supported name-value pair arguments, and any differences, are:
  - `'Display'` — Default value is `'iter'`.
  - `'MaxIter'`
  - `'Options'` — Supports only the `'TolFun'` field of the structure array created by `statset`. The default value of `'TolFun'` is `1e-4`. The `kmeans` function uses the value of `'TolFun'` as the termination tolerance for the within-cluster sums of point-to-centroid distances. For example, you can specify `'Options',statset('TolFun',1e-8)`.
  - `'Replicates'`
  - `'Start'` — Supports only `'plus'`, `'sample'`, and a numeric array.

For more information, see “Tall Arrays for Out-of-Memory Data”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- If the `Start` method uses random selections, the initial centroid cluster positions might not match MATLAB.
- If the number of rows in `X` is fixed, code generation does not remove rows of `X` that contain a `NaN`.
- The cluster centroid locations in `C` can have a different order than in MATLAB. In this case, the cluster indices in `idx` have corresponding differences.

- If you provide `Display`, its value must be `'off'`.
- If you provide `Streams`, it must be empty and `UseSubstreams` must be `false`.
- When you set the `UseParallel` option to `true`:
  - Some computations can execute in parallel even when `Replicates` is `1`. For large data sets, when `Replicates` is `1`, consider setting the `UseParallel` option to `true`.
  - `kmeans` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms. Loops that run in parallel can be faster than loops that run on a single thread. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`.
- To save memory on the device to which you deploy generated code, you can separate training and prediction by using `kmeans` and `pdist2`, respectively. Use `kmeans` to create clusters in MATLAB and use `pdist2` in the generated code to assign new data to existing clusters. For code generation, define an entry-point function that accepts the cluster centroid positions and the new data set, and returns the index of the nearest cluster. Then, generate code for the entry-point function. For an example, see “Assign New Data to Existing Clusters and Generate C/C++ Code” on page 33-3320.
- Starting in R2020a, `kmeans` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for stricter single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`clusterdata` | `gmdistribution` | `kmedoids` | `linkage` | `parpool` | `silhouette` | `statset`

Topics
“Compare k-Means Clustering Solutions” on page 16-33
“Introduction to k-Means Clustering” on page 16-33

Introduced before R2006a

kmedoids

k-medoids clustering

Syntax

```
idx = kmedoids(X,k)
idx = kmedoids(X,k,Name,Value)
[idx,C] = kmedoids( ___ )
[idx,C,sumd] = kmedoids( ___ )
[idx,C,sumd,D] = kmedoids( ___ )
[idx,C,sumd,D,midx] = kmedoids( ___ )
[idx,C,sumd,D,midx,info] = kmedoids( ___ )
```

Description

`idx = kmedoids(X,k)` performs “k-medoids Clustering” on page 33-3345 to partition the observations of the *n*-by-*p* matrix `X` into `k` clusters, and returns an *n*-by-1 vector `idx` containing cluster indices of each observation. Rows of `X` correspond to points and columns correspond to variables. By default, `kmedoids` uses squared Euclidean distance metric and the *k*-means++ algorithm on page 33-3329 for choosing initial cluster medoid positions.

`idx = kmedoids(X,k,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments.

`[idx,C] = kmedoids( ___ )` returns the `k` cluster medoid locations in the *k*-by-*p* matrix `C`.

`[idx,C,sumd] = kmedoids( ___ )` returns the within-cluster sums of point-to-medoid distances in the *k*-by-1 vector `sumd`.

`[idx,C,sumd,D] = kmedoids( ___ )` returns distances from each point to every medoid in the *n*-by-*k* matrix `D`.

`[idx,C,sumd,D,midx] = kmedoids( ___ )` returns the indices `midx` such that `C = X(midx,:)`. `midx` is a *k*-by-1 vector.

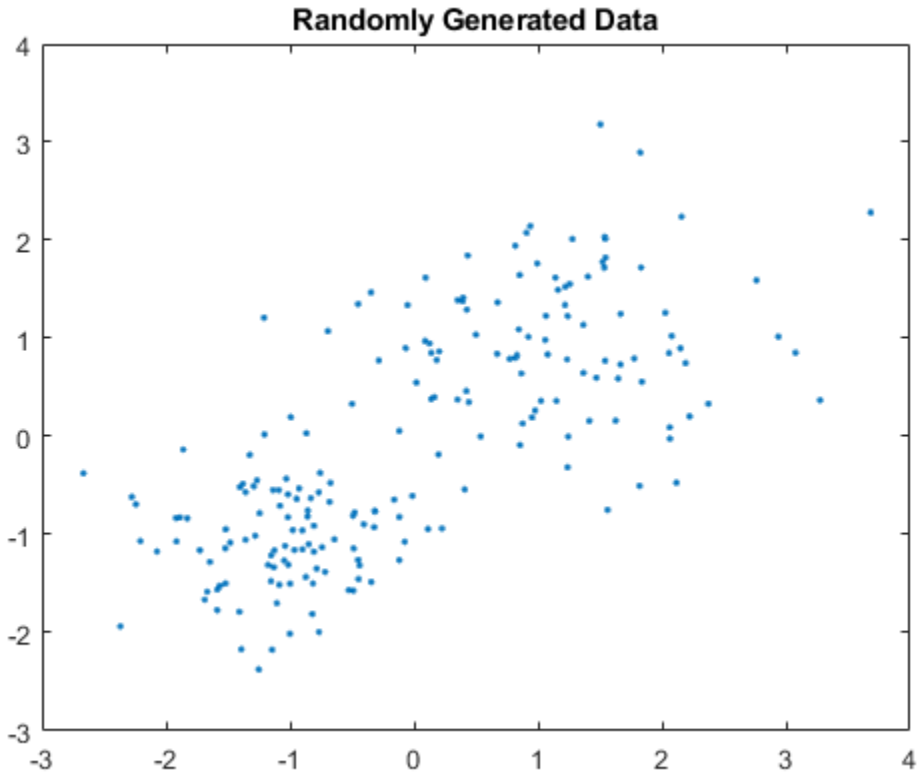
`[idx,C,sumd,D,midx,info] = kmedoids( ___ )` returns a structure `info` with information about the options used by the algorithm when executed.

Examples

Group Data into Two Clusters

Randomly generate data.

```
rng('default'); % For reproducibility
X = [randn(100,2)*0.75+ones(100,2);
    randn(100,2)*0.55-ones(100,2)];
figure;
plot(X(:,1),X(:,2),'.');
title('Randomly Generated Data');
```



Group data into two clusters using `kmedoids`. Use the `cityblock` distance metric.

```
opts = statset('Display','iter');
[idx,C,sumd,d,midx,info] = kmedoids(X,2,'Distance','cityblock','Options',opts);
```

```
   rep        iter             sum
     1           1         209.856
     1           2         209.856
Best total sum of distances = 209.856
```

`info` is a struct that contains information about how the algorithm was executed. For example, `bestReplicate` field indicates the replicate that was used to produce the final solution. In this example, the replicate number 1 was used since the default number of replicates is 1 for the default algorithm, which is `pam` in this case.

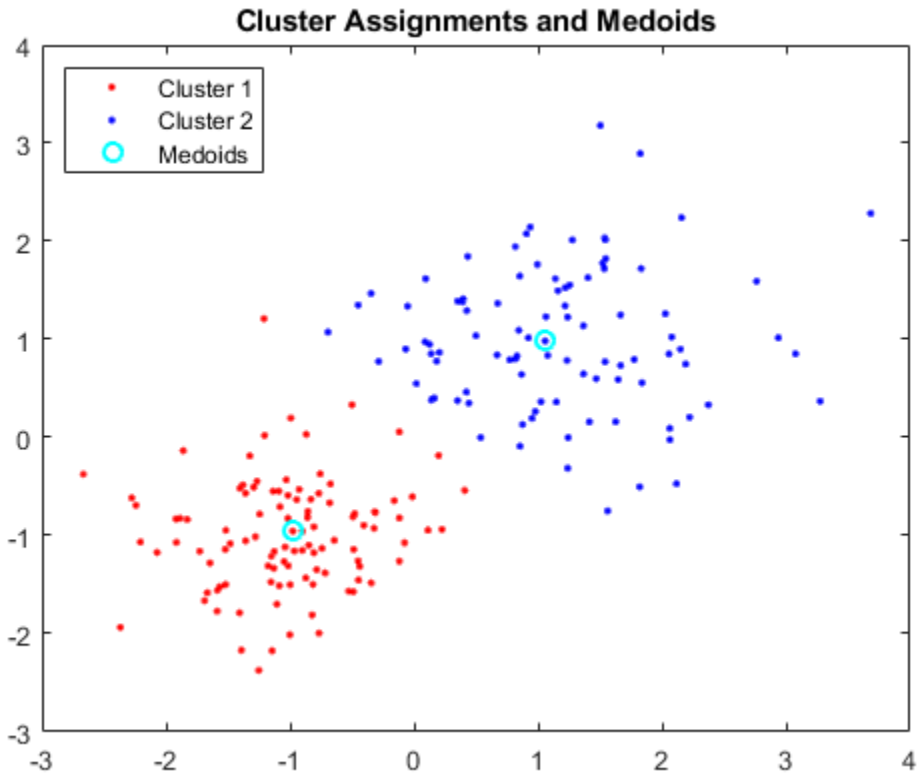
```
info
```

```
info = struct with fields:
        algorithm: 'pam'
            start: 'plus'
         distance: 'cityblock'
       iterations: 2
    bestReplicate: 1
```

Plot the clusters and the cluster medoids.

```
figure;
plot(X(idx==1,1),X(idx==1,2),'r.','MarkerSize',7)
```

```
hold on
plot(X(idx==2,1),X(idx==2,2),'b.','MarkerSize',7)
plot(C(:,1),C(:,2),'co',...
     'MarkerSize',7,'LineWidth',1.5)
legend('Cluster 1','Cluster 2','Medoids',...
       'Location','NW');
title('Cluster Assignments and Medoids');
hold off
```



Cluster Categorical Data Using k-Medoids

This example uses "Mushroom" data set [3][4][5] [6][7] from the UCI machine learning archive [7], described in http://archive.ics.uci.edu/ml/datasets/Mushroom. The data set includes 22 predictors for 8,124 observations of various mushrooms. The predictors are categorical data types. For example, cap shape is categorized with features of `'b'` for bell-shaped cap and `'c'` for conical. Mushroom color is also categorized with features of `'n'` for brown, and `'p'` for pink. The data set also includes a classification for each mushroom of either edible or poisonous.

Since the features of the mushroom data set are categorical, it is not possible to define the mean of several data points, and therefore the widely-used *k*-means clustering algorithm cannot be meaningfully applied to this data set. *k*-medoids is a related algorithm that partitions data into *k* distinct clusters, by finding medoids that minimize the sum of dissimilarities between points in the data and their nearest medoid.

The medoid of a set is a member of that set whose average dissimilarity with the other members of the set is the smallest. Similarity can be defined for many types of data that do not allow a mean to be calculated, allowing *k*-medoids to be used for a broader range of problems than *k*-means.

Using *k*-medoids, this example clusters the mushrooms into two groups, based on the predictors provided. It then explores the relationship between those clusters and the classifications of the mushrooms as either edible or poisonous.

This example assumes that you have downloaded the "Mushroom" data set [3][4][5] [6][7] from the UCI database (http://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/) and saved it in your current directory as a text file named agaricus-lepiota.txt. There is no column headers in the data, so `readtable` uses the default variable names.

```
clear all
data = readtable('agaricus-lepiota.txt','ReadVariableNames',false);
```

Display the first 5 mushrooms with their first few features.

```
data(1:5,1:10)
```

```
ans =

    Var1    Var2    Var3    Var4    Var5    Var6    Var7    Var8    Var9    Var10
    ____    ____    ____    ____    ____    ____    ____    ____    ____    _____

    'p'     'x'     's'     'n'     't'     'p'     'f'     'c'     'n'     'k'
    'e'     'x'     's'     'y'     't'     'a'     'f'     'c'     'b'     'k'
    'e'     'b'     's'     'w'     't'     'l'     'f'     'c'     'b'     'n'
    'p'     'x'     'y'     'w'     't'     'p'     'f'     'c'     'n'     'n'
    'e'     'x'     's'     'g'     'f'     'n'     'f'     'w'     'b'     'k'
```

Extract the first column, labeled data for edible and poisonous groups. Then delete the column.

```
labels = data(:,1);
labels = categorical(labels{:,:});
data(:,1) = [];
```

Store the names of predictors (features), which are described in http://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota.names.

```
VarNames = {'cap_shape' 'cap_surface' 'cap_color' 'bruises' 'odor' ...
    'gill_attachment' 'gill_spacing' 'gill_size' 'gill_color' ...
    'stalk_shape' 'stalk_root' 'stalk_surface_above_ring' ...
    'stalk_surface_below_ring' 'stalk_color_above_ring' ...
    'stalk_color_below_ring' 'veil_type' 'veil_color' 'ring_number' ....
    'ring_type' 'spore_print_color' 'population' 'habitat'};
```

Set the variable names.

```
data.Properties.VariableNames = VarNames;
```

There are a total of 2480 missing values denoted as `'?'`.

```
sum(char(data{:,:}) == '?')
```

```
ans =

        2480
```

Based on the inspection of the data set and its description, the missing values belong only to the 11th variable (`stalk_root`). Remove the column from the table.

```
data(:,11) = [];
```

`kmedoids` only accepts numeric data. You need to cast the categories you have into numeric type. The distance function you will use to define the dissimilarity of the data will be based on the double representation of the categorical data.

```
cats = categorical(data{:,:});
data = double(cats);
```

`kmedoids` can use any distance metric supported by pdist2 to cluster. For this example you will cluster the data using the Hamming distance because this is an appropriate distance metric for categorical data as illustrated below. The Hamming distance between two vectors is the percentage of the vector components that differ. For instance, consider these two vectors.

```
v1 = [1 0 2 1];
```

```
v2 = [1 1 2 1];
```

They are equal in the 1st, 3rd and 4th coordinate. Since 1 of the 4 coordinates differ, the Hamming distance between these two vectors is .25.

You can use the function `pdist2` to measure the Hamming distance between the first and second row of data, the numerical representation of the categorical mushroom data. The value .2857 means that 6 of the 21 features of the mushroom differ.

```
pdist2(data(1,:),data(2,:),'hamming')
```

```
ans =

    0.2857
```

In this example, you're clustering the mushroom data into two clusters based on features to see if the clustering corresponds to edibility. The `kmedoids` function is guaranteed to converge to a local minima of the clustering criterion; however, this may not be a global minimum for the problem. It is a good idea to cluster the problem a few times using the `'replicates'` parameter. When `'replicates'` is set to a value, *n*, greater than 1, the k-medoids algorithm is run *n* times, and the best result is returned.

To run `kmedoids` to cluster data into 2 clusters, based on the Hamming distance and to return the best result of 3 replicates, you run the following.

```
rng('default'); % For reproducibility
[IDX, C, SUMD, D, MIDX, INFO] = kmedoids(data,2,'distance','hamming','replicates',3);
```

Let's assume that mushrooms in the predicted group 1 are poisonous and group 2 are all edible. To determine the performance of clustering results, calculate how many mushrooms in group 1 are indeed poisonous and group 2 are edible based on the known labels. In other words, calculate the number of false positives, false negatives, as well as true positives and true negatives.

Construct a confusion matrix (or matching matrix), where the diagonal elements represent the number of true positives and true negatives, respectively. The off-diagonal elements represent false negatives and false positives, respectively. For convenience, use the `confusionmat` function, which calculates a confusion matrix given known labels and predicted labels. Get the predicted label

information from the IDX variable. IDX contains values of 1 and 2 for each data point, representing poisonous and edible groups, respectively.

```
predLabels = labels; % Initialize a vector for predicted labels.
predLabels(IDX==1) = categorical({'p'}); % Assign group 1 to be poisonous.
predLabels(IDX==2) = categorical({'e'}); % Assign group 2 to be edible.
confMatrix = confusionmat(labels,predLabels)
```

```
confMatrix =

        4176          32
         816        3100
```

Out of 4208 edible mushrooms, 4176 were correctly predicted to be in group 2 (edible group), and 32 were incorrectly predicted to be in group 1 (poisonous group). Similarly, out of 3916 poisonous mushrooms, 3100 were correctly predicted to be in group 1 (poisonous group), and 816 were incorrectly predicted to be in group 2 (edible group).

Given this confusion matrix, calculate the accuracy, which is the proportion of true results (both true positives and true negatives) against the overall data, and precision, which is the proportion of the true positives against all the positive results (true positives and false positives).

```
accuracy = (confMatrix(1,1)+confMatrix(2,2))/(sum(sum(confMatrix)))
```

```
accuracy =

    0.8956
```

```
precision = confMatrix(1,1) / (confMatrix(1,1)+confMatrix(2,1))
```

```
precision =

    0.8365
```

The results indicated that applying the k-medoids algorithm to the categorical features of mushrooms resulted in clusters that were associated with edibility.

Input Arguments

X — Data

numeric matrix

Data, specified as a numeric matrix. The rows of X correspond to observations, and the columns correspond to variables.

k — Number of medoids

positive integer

Number of medoids in the data, specified as a positive integer.

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example:
`'Distance','euclidean','Replicates',3,'Options',statset('UseParallel',1)`
specifies Euclidean distance, three replicate medoids at different starting values, and to use parallel computing.

`Algorithm` — Algorithm to find medoids
`'pam'` | `'small'` | `'clara'` | `'large'`

Algorithm to find medoids, specified as the comma-separated pair consisting of `'Algorithm'` and `'pam'`, `'small'`, `'clara'`, or `'large'`. The default algorithm depends on the number of rows of `X`.

- If the number of rows of `X` is less than 3000, `'pam'` is the default algorithm.
- If the number of rows is between 3000 and 10000, `'small'` is the default algorithm.
- For all other cases, `'large'` is the default algorithm.

You can override the default choice by explicitly stating the algorithm. This table summarizes the available algorithms.

Algorithm	Description
`'pam'`	Partitioning Around Medoids (PAM) is the classical algorithm for solving the *k*-medoids problem described in [1]. After applying the initialization function to select initial medoid positions, the program performs the swap-step of the PAM algorithm, that is, it searches over all possible swaps between medoids and non-medoids to see if the sum of medoid to cluster member distances goes down. You can specify which initialization function to use via the `'Start'` name-value pair argument. The algorithm proceeds as follows. **1** Build-step: Each of *k* clusters is associated with a potential medoid. This assignment is performed using a technique specified by the `'Start'` name-value pair argument. **2** Swap-step: Within each cluster, each point is tested as a potential medoid by checking if the sum of within-cluster distances gets smaller using that point as the medoid. If so, the point is defined as a new medoid. Every point is then assigned to the cluster with the closest medoid. The algorithm iterates the build- and swap-steps until the medoids do not change, or other termination criteria are met. The algorithm can produce better solutions than the other algorithms in some situations, but it can be prohibitively long running.

Algorithm	Description
`'small'`	Use an algorithm similar to the k-means algorithm to find `k` medoids. This option employs a variant of the Lloyd's iterations based on [2]. The algorithm proceeds as follows. **1** For each point in each cluster, calculate the sum of distances from the point to every other point in the cluster. Choose the point that minimizes the sum as the new medoid. **2** Update the cluster membership for each data point to reflect the new medoid. The algorithm repeats these steps until no further updates occur or other termination criteria are met. The algorithm has an optional PAM-like online update phase (specified by the `'OnlinePhase'` name-value pair argument) that improves cluster quality. It tends to return higher quality solutions than the `clara` or `large` algorithms, but it may not be the best choice for very large data.
`'clara'`	Clustering LARge Applications (CLARA) [1] repeatedly performs the PAM algorithm on random subsets of the data. It aims to overcome scaling challenges posed by the PAM algorithm through sampling. The algorithm proceeds as follows. **1** Select a subset of the data and apply the PAM algorithm to the subset. **2** Assign points of the full data set to clusters by picking the closest medoid. The algorithm repeats these steps until the medoids do not change, or other termination criteria are met. For the best performance, it is recommended that you perform multiple replicates. By default, the program performs five replicates. Each replicate samples `s` rows from X (specified by `'NumSamples'` name-value pair argument) to perform clustering on. By default, `40+2*k` samples are selected.
`'large'`	This is similar to the `small` scale algorithm and repeatedly performs searches using a k-means like update. However, the algorithm examines only a random sample of cluster members during each iteration. The user-adjustable parameter, `'PercentNeighbors'`, controls the number of neighbors to examine. If there is no improvement after the neighbors are examined, the algorithm terminates the local search. The algorithm performs a total of *r* replicates (specified by `'Replicates'` name-value pair argument) and returns the best clustering result. The algorithm has an optional PAM-like online phase (specified by the `'OnlinePhase'` name-value pair argument) that improves cluster quality.

Example: `'Algorithm','pam'`

`OnlinePhase` — Flag to perform PAM-like online update phase
`'on'` (default) | `'off'`

A flag to perform PAM-like online update phase, specified as a comma-separated pair consisting of `'OnlinePhase'` and `'on'` or `'off'`.

If it is `on`, then `kmedoids` performs a PAM-like update to the medoids after the Lloyd iterations in the `small` and `large` algorithms. During this online update phase, the algorithm chooses a small subset of data points in each cluster that are the furthest from and nearest to medoid. For each chosen point, it reassigns the clustering of the entire data set and check if this creates a smaller sum of distances than the best known.

In other words, the swap considerations are limited to the points near the medoids and far from the medoids. The near points are considered in order to refine the clustering. The far points are considered in order to escape local minima. Turning on this feature tends to improve the quality of solutions generated by both algorithms. Total run time tends to increase as well, but the increase typically is less than one iteration of PAM.

Example: `OnlinePhase,'off'`

`Distance` — Distance metric

`'sqEuclidean'` (default) | `'euclidean'` | character vector | string scalar | function handle | ...

Distance metric, specified as the name of a distance metric described in the following table, or a function handle. `kmedoids` minimizes the sum of medoid to cluster member distances.

Value	Description
`'sqEuclidean'`	Squared Euclidean distance (default)
`'euclidean'`	Euclidean distance
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`.
`'cityblock'`	City block distance
`'minkowski'`	Minkowski distance. The exponent is 2.
`'chebychev'`	Chebychev distance (maximum coordinate difference)
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ

Value	Description

`@distfun`	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For the definition of each distance metric, see “Distance Metrics” on page 33-4495.

Example: `'Distance','hamming'`

`Options` — Options to control iterative algorithm to minimize fitting criteria
`[]` (default) | structure array returned by `statset`

Options to control the iterative algorithm to minimize fitting criteria, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. Supported fields of the structure array specify options for controlling the iterative algorithm. This table summarizes the supported fields.

Field	Description
`Display`	Level of display output. Choices are `'off'` (default) and `'iter'`.
`MaxIter`	Maximum number of iterations allowed. The default is `100`.
`UseParallel`	If `true`, compute in parallel. If Parallel Computing Toolbox is not available, then computation occurs in serial mode. The default is `false`, meaning serial computation.
`UseSubstreams`	Set to `true` to compute in parallel in a reproducible fashion. The default is `false`. To compute reproducibly, you must also set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
`Streams`	A `RandStream` object or cell array of such objects. For details about these options and parallel computing in Statistics and Machine Learning Toolbox, see “Speed Up Statistical Computations” or enter `help parallelstats` at the command line.

Example: `'Options',statset('Display','off')`

`Replicates` — Number of times to repeat clustering using new initial cluster medoid positions
positive integer

Number of times to repeat clustering using new initial cluster medoid positions, specified as a positive integer. The default value depends on the choice of algorithm. For `pam` and `small`, the default is 1. For `clara`, the default is 5. For `large`, the default is 3.

Example: `'Replicates',4`

`NumSamples` — Number of samples to take from data when executing `clara` algorithm
`40+2*k` (default) | positive integer

Number of samples to take from the data when executing the `clara` algorithm, specified as a positive integer. The default number of samples is calculated as `40+2*k`.

Example: `'NumSamples',160`

`PercentNeighbors` — Percent of data set to examine using large algorithm
0.001 (default) | scalar value between 0 and 1

Percent of the data set to examine using the `large` algorithm, specified as a positive number.

The program examines `percentneighbors*size(X,1)` number of neighbors for the medoids. If there is no improvement in the within-cluster sum of distances, then the algorithm terminates.

The value of this parameter between 0 and 1, where a value closer to 1 tends to give higher quality solutions, but the algorithm takes longer to run, and a value closer to 0 tends to give lower quality solutions, but finishes faster.

Example: `'PercentNeighbors',0.01`

`Start` — Method for choosing initial cluster medoid positions
`'plus'` (default) | `'sample'` | `'cluster'` | matrix

Method for choosing initial cluster medoid positions, specified as the comma-separated pair consisting of `'Start'` and `'plus'`, `'sample'`, `'cluster'`, or a matrix. This table summarizes the available methods.

Method	Description
`'plus'` (default)	Select `k` observations from `X` according to the *k*-means++ algorithm on page 33-3329 for cluster center initialization.
`'sample'`	Select `k` observations from `X` at random.
`'cluster'`	Perform preliminary clustering phase on a random subsample (10%) of `X`. This preliminary phase is itself initialized using `sample`, that is, the observations are selected at random.
matrix	A custom `k`-by-*p* matrix of starting locations. In this case, you can pass in `[]` for the `k` input argument, and `kmedoids` infers `k` from the first dimension of the matrix. You can also supply a 3-D array, implying a value for `'Replicates'` from the array's third dimension.

Example: `'Start','sample'`

Data Types: `char` | `string` | `single` | `double`

Output Arguments

`idx` — Medoid indices

numeric column vector

Medoid indices, returned as a numeric column vector. `idx` has as many rows as `X`, and each row indicates the medoid assignment of the corresponding observation.

`C` — Cluster medoid locations

numeric matrix

Cluster medoid locations, returned as a numeric matrix. `C` is a *k*-by-*p* matrix, where row *j* is the medoid of cluster *j*

`sumd` — Within-cluster sums of point-to-medoid distances

numeric column vector

Within-cluster sums of point-to-medoid distances, returned as a numeric column vector. `sumd` is a `k`-by1 vector, where element *j* is the sum of point-to-medoid distances within cluster *j*.

`D` — Distances from each point to every medoid

numeric matrix

Distances from each point to every medoid, returned as a numeric matrix. `D` is an *n*-by-`k` matrix, where element (*j*,*m*) is the distance from observation *j* to medoid *m*.

`midx` — Index to `X`

column vector

Index to `X`, returned as a column vector of indices. `midx` is a `k`-by-1 vector and the indices satisfy `C = X(midx,:)`.

`info` — Algorithm information

struct

Algorithm information, returned as a struct. `info` contains options used by the function when executed such as *k*-medoid clustering algorithm (`algorithm`), method used to choose initial cluster medoid positions (`start`), distance metric (`distance`), number of iterations taken in the best replicate (`iterations`) and the replicate number of the returned results (`bestReplicate`).

More About

k-medoids Clustering

k-medoids clustering is a partitioning method commonly used in domains that require robustness to outlier data, arbitrary distance metrics, or ones for which the mean or median does not have a clear definition.

It is similar to *k*-means, and the goal of both methods is to divide a set of measurements or observations into *k* subsets or clusters so that the subsets minimize the sum of distances between a measurement and a center of the measurement's cluster. In the *k*-means algorithm, the center of the subset is the mean of measurements in the subset, often called a centroid. In the *k*-medoids algorithm, the center of the subset is a member of the subset, called a medoid.

The *k*-medoids algorithm returns medoids which are the actual data points in the data set. This allows you to use the algorithm in situations where the mean of the data does not exist within the data set. This is the main difference between *k*-medoids and *k*-means where the centroids returned by *k*-means may not be within the data set. Hence *k*-medoids is useful for clustering categorical data where a mean is impossible to define or interpret.

The function `kmedoids` provides several iterative algorithms that minimize the sum of distances from each object to its cluster medoid, over all clusters. One of the algorithms is called partitioning around medoids (PAM) [1] which proceeds in two steps.

1. Build-step: Each of *k* clusters is associated with a potential medoid. This assignment is performed using a technique specified by the `'Start'` name-value pair argument.
2. Swap-step: Within each cluster, each point is tested as a potential medoid by checking if the sum of within-cluster distances gets smaller using that point as the medoid. If so, the point is defined as a new medoid. Every point is then assigned to the cluster with the closest medoid.

The algorithm iterates the build- and swap-steps until the medoids do not change, or other termination criteria are met.

You can control the details of the minimization using several optional input parameters to `kmedoids`, including ones for the initial values of the cluster medoids, and for the maximum number of iterations. By default, `kmedoids` uses the *k*-means++ algorithm on page 33-3329 for cluster medoid initialization and the squared Euclidean metric to determine distances.

References

[1] Kaufman, L., and Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.

[2] Park, H-S, and Jun, C-H. (2009). A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. 36, 3336-3341.

[3] Schlimmer,J.S. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine.

[4] Iba,W., Wogulis,J., and Langley,P. (1988). Trading off Simplicity and Coverage in Incremental Concept Learning. In Proceedings of the 5th International Conference on Machine Learning, 73-79. Ann Arbor, Michigan: Morgan Kaufmann.

[5] Duch W, A.R., and Grabczewski, K. (1996) Extraction of logical rules from training data using backpropagation networks. Proc. of the 1st Online Workshop on Soft Computing, 19-30, pp. 25-30.

[6] Duch, W., Adamczak, R., Grabczewski, K., Ishikawa, M., and Ueda, H. (1997). Extraction of crisp logical rules using constrained backpropagation networks - comparison of two new approaches. Proc. of the European Symposium on Artificial Neural Networks (ESANN'97), Bruge, Belgium 16-18.

[7] Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`clusterdata` | `evalclusters` | `kmeans` | `linkage` | `linkage` | `pdist` | `silhouette`

Introduced in R2014b

knnsearch

Find *k*-nearest neighbors using searcher object

Syntax

```
Idx = knnsearch(Mdl,Y)
Idx = knnsearch(Mdl,Y,Name,Value)
[Idx,D] = knnsearch( ___ )
```

Description

`Idx = knnsearch(Mdl,Y)` searches for the nearest neighbor (i.e., the closest point, row, or observation) in `Mdl.X` to each point (i.e., row or observation) in the query data `Y` using an exhaustive search or a *K*d-tree. `knnsearch` returns `Idx`, which is a column vector of the indices in `Mdl.X` representing the nearest neighbors.

`Idx = knnsearch(Mdl,Y,Name,Value)` returns the indices of the closest points in `Mdl.X` to `Y` with additional options specified by one or more `Name,Value` pair arguments. For example, specify the number of nearest neighbors to search for, distance metric different from the one stored in `Mdl.Distance`. You can also specify which action to take if the closest distances are tied.

`[Idx,D] = knnsearch( ___ )` additionally returns the matrix `D` using any of the input arguments in the previous syntaxes. `D` contains the distances between each observation in `Y` that correspond to the closest observations in `Mdl.X`. By default, the function arranges the columns of `D` in ascending order by closeness, with respect to the distance metric.

Examples

Search for Nearest Neighbors Using *K*d-tree and Exhaustive Search

`knnsearch` accepts `ExhaustiveSearcher` or `KDTreeSearcher` model objects to search the training data for the nearest neighbors to the query data. An `ExhaustiveSearcher` model invokes the exhaustive searcher algorithm, and a `KDTreeSearcher` model defines a *K*d-tree, which `knnsearch` uses to search for nearest neighbors.

Load Fisher's iris data set. Randomly reserve five observations from the data for query data.

```
load fisheriris
rng(1); % For reproducibility
n = size(meas,1);
idx = randsample(n,5);
X = meas(~ismember(1:n,idx),:); % Training data
Y = meas(idx,:);                % Query data
```

The variable `meas` contains 4 predictors.

Grow a default four-dimensional *K*d-tree.

```
MdlKDT = KDTreeSearcher(X)
```

```
MdlKDT =
  KDTreeSearcher with properties:

      BucketSize: 50
        Distance: 'euclidean'
    DistParameter: []
                X: [145x4 double]
```

`MdlKDT` is a `KDTreeSearcher` model object. You can alter its writable properties using dot notation.

Prepare an exhaustive nearest neighbor searcher.

```
MdlES = ExhaustiveSearcher(X)
```

```
MdlES =
  ExhaustiveSearcher with properties:

        Distance: 'euclidean'
    DistParameter: []
                X: [145x4 double]
```

`MdlKDT` is an `ExhaustiveSearcher` model object. It contains the options, such as the distance metric, to use to find nearest neighbors.

Alternatively, you can grow a *K*d-tree or prepare an exhaustive nearest neighbor searcher using `createns`.

Search the training data for the nearest neighbors indices that correspond to each query observation. Conduct both types of searches using the default settings. By default, the number of neighbors to search for per query observation is `1`.

```
IdxKDT = knnsearch(MdlKDT,Y);
IdxES = knnsearch(MdlES,Y);
[IdxKDT IdxES]
```

```
ans = 5×2

    17    17
     6     6
     1     1
    89    89
   124   124
```

In this case, the results of the search are the same.

Search for Nearest Neighbors of Query Data Using Minkowski Distance

Grow a *K*d-tree nearest neighbor searcher object by using the `createns` function. Pass the object and query data to the `knnsearch` function to find *k*-nearest neighbors.

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng(1);                      % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
tIdx = ~ismember(1:n,qIdx); % Indices of training data
Q = meas(qIdx,:);
X = meas(tIdx,:);
```

Grow a four-dimensional *K*d-tree using the training data. Specify the Minkowski distance for finding nearest neighbors.

```
Mdl = createns(X,'Distance','minkowski')
```

```
Mdl =
  KDTreeSearcher with properties:

       BucketSize: 50
         Distance: 'minkowski'
    DistParameter: 2
                X: [145x4 double]
```

Because `X` has four columns and the distance metric is Minkowski, `createns` creates a `KDTreeSearcher` model object by default. The Minkowski distance exponent is `2` by default.

Find the indices of the training data (`Mdl.X`) that are the two nearest neighbors of each point in the query data (`Q`).

```
IdxNN = knnsearch(Mdl,Q,'K',2)
```

```
IdxNN = 5×2

    17     4
     6     2
     1    12
    89    66
   124   100
```

Each row of `IdxNN` corresponds to a query data observation, and the column order corresponds to the order of the nearest neighbors, with respect to ascending distance. For example, based on the Minkowski distance, the second nearest neighbor of `Q(3,:)` is `X(12,:)`.

Include Ties in Nearest Neighbor Search

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng(4);                      % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
```

```
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Grow a four-dimensional *K*d-tree using the training data. Specify the Minkowski distance for finding nearest neighbors.

```
Mdl = KDTreeSearcher(X);
```

`Mdl` is a `KDTreeSearcher` model object. By default, the distance metric for finding nearest neighbors is the Euclidean metric.

Find the indices of the training data (`X`) that are the seven nearest neighbors of each point in the query data (`Y`).

```
[Idx,D] = knnsearch(Mdl,Y,'K',7,'IncludeTies',true);
```

`Idx` and `D` are five-element cell arrays of vectors, with each vector having at least seven elements.

Display the lengths of the vectors in `Idx`.

```
cellfun('length',Idx)
```

```
ans = 5×1

     8
     7
     7
     7
     7
```

Because cell `1` contains a vector with length greater than $k = 7$, query observation 1 (`Y(1,:)`) is equally close to at least two observations in `X`.

Display the indices of the nearest neighbors to `Y(1,:)` and their distances.

```
nn5 = Idx{1}
```

```
nn5 = 1×8

    91    98    67    69    71    93    88    95
```

```
nn5d = D{1}
```

```
nn5d = 1×8

    0.1414    0.2646    0.2828    0.3000    0.3464    0.3742    0.3873    0.3873
```

Training observations `88` and `95` are 0.3873 cm away from query observation `1`.

Compare *k*-Nearest Neighbors Using Different Distance Metrics

Train two `KDTreeSearcher` models using different distance metrics, and compare *k*-nearest neighbors of query data for the two models.

Load Fisher's iris data set. Consider the petal measurements as predictors.

```
load fisheriris
X = meas(:,3:4); % Predictors
Y = species;     % Response
```

Train a `KDTreeSearcher` model object by using the predictors. Specify the Minkowski distance with exponent 5.

```
KDTreeMdl = KDTreeSearcher(X,'Distance','minkowski','P',5)
```

```
KDTreeMdl =
  KDTreeSearcher with properties:

      BucketSize: 50
        Distance: 'minkowski'
    DistParameter: 5
                X: [150x2 double]
```

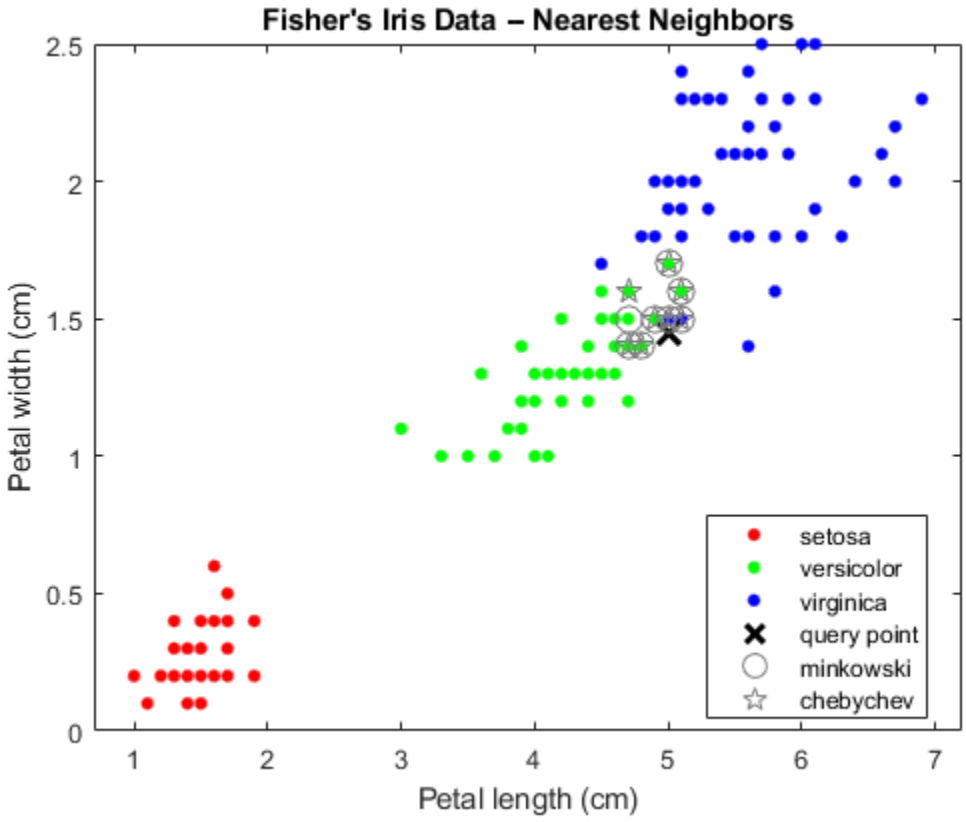
Find the 10 nearest neighbors from `X` to a query point (`newpoint`), first using Minkowski then Chebychev distance metrics. The query point must have the same column dimension as the data used to train the model.

```
newpoint = [5 1.45];
[IdxMk,DMk] = knnsearch(KDTreeMdl,newpoint,'k',10);
[IdxCb,DCb] = knnsearch(KDTreeMdl,newpoint,'k',10,'Distance','chebychev');
```

`IdxMk` and `IdxCb` are 1-by-10 matrices containing the row indices of `X` corresponding to the nearest neighbors to `newpoint` using Minkowski and Chebychev distances, respectively. Element (1,1) is the nearest, element (1,2) is the next nearest, and so on.

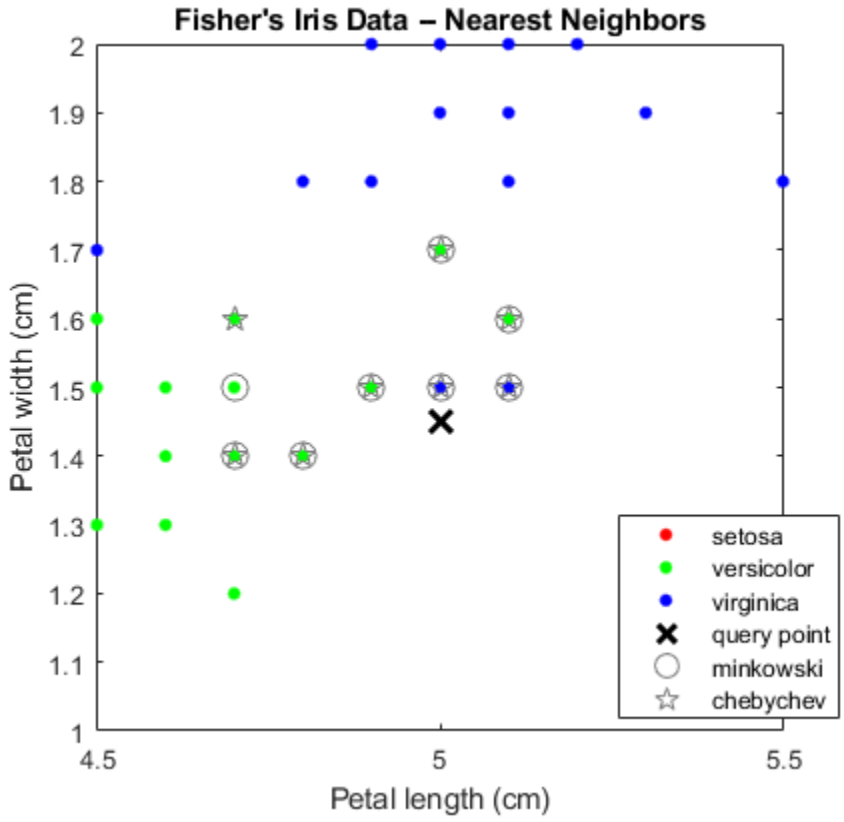
Plot the training data, query point, and nearest neighbors.

```
figure;
gscatter(X(:,1),X(:,2),Y);
title('Fisher''s Iris Data -- Nearest Neighbors');
xlabel('Petal length (cm)');
ylabel('Petal width (cm)');
hold on
plot(newpoint(1),newpoint(2),'kx','MarkerSize',10,'LineWidth',2);   % Query point
plot(X(IdxMk,1),X(IdxMk,2),'o','Color',[.5 .5 .5],'MarkerSize',10); % Minkowski nearest neighbors
plot(X(IdxCb,1),X(IdxCb,2),'p','Color',[.5 .5 .5],'MarkerSize',10); % Chebychev nearest neighbors
legend('setosa','versicolor','virginica','query point',...
   'minkowski','chebychev','Location','Best');
```



Zoom in on the points of interest.

```
h = gca; % Get current axis handle.
h.XLim = [4.5 5.5];
h.YLim = [1 2];
axis square;
```



Several observations are equal, which is why only eight nearest neighbors are identified in the plot.

Input Arguments

`Mdl` — Nearest neighbor searcher

`ExhaustiveSearcher` model object | `KDTreeSearcher` model object

Nearest neighbor searcher, specified as an `ExhaustiveSearcher` or `KDTreeSearcher` model object, respectively.

If `Mdl` is an `ExhaustiveSearcher` model, then `knnsearch` searches for nearest neighbors using an exhaustive search. Otherwise, `knnsearch` uses the grown *K*d-tree to search for nearest neighbors.

`Y` — Query data

numeric matrix

Query data, specified as a numeric matrix.

`Y` is an *m*-by-*K* matrix. Rows of `Y` correspond to observations (i.e., examples), and columns correspond to predictors (i.e., variables or features). `Y` must have the same number of columns as the training data stored in `Mdl.X`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'K',2,'Distance','minkowski'` specifies to find the two nearest neighbors of `Mdl.X` to each point in `Y` and to use the Minkowski distance metric.

For Both Nearest Neighbor Searchers

`Distance` — Distance metric
`Mdl.Distance` (default) | `'cityblock'` | `'euclidean'` | `'mahalanobis'` | `'minkowski'` | `'seuclidean'` | function handle | ...

Distance metric used to find neighbors of the training data to the query observations, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle.

For both types of nearest neighbor searchers, `knnsearch` supports these distance metrics.

Value	Description
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cityblock'`	City block distance.
`'euclidean'`	Euclidean distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value pair argument.

If `Mdl` is an `ExhaustiveSearcher` model object, then `knnsearch` also supports these distance metrics.

Value	Description
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as row vectors).
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix. To change the value of the covariance matrix, use the `'Cov'` name-value pair argument.

Value	Description
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in Mdl.X and the query matrix is scaled by dividing by the corresponding element of the standard deviation computed from Mdl.X. To specify another scaling, use the 'Scale' name-value pair argument.
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

If Mdl is an ExhaustiveSearcher model object, then you can also specify a function handle for a custom distance metric by using @ (for example, @distfun). The custom distance function must:

- Have the form function D2 = distfun(ZI,ZJ).
- Take as arguments:
  - A 1-by-*K* vector ZI containing a single row from Mdl.X or Y, where *K* is the number of columns of Mdl.X.
  - An *m*-by-*K* matrix ZJ containing multiple rows of Mdl.X or Y, where *m* is a positive integer.
- Return an *m*-by-1 vector of distances D2, where D2(*j*) is the distance between the observations ZI and ZJ(*j*,:).

For more details, see “Distance Metrics” on page 18-12.

Example: 'Distance','minkowski'

IncludeTies — Flag to include all nearest neighbors
false (0) (default) | true (1)

Flag to include nearest neighbors that have the same distance from query observations, specified as the comma-separated pair consisting of 'IncludeTies' and false (0) or true (1).

If IncludeTies is true, then:

- knnsearch includes all nearest neighbors whose distances are equal to the *k*th smallest distance in the output arguments, where *k* is the number of searched nearest neighbors specified by the 'K' name-value pair argument.
- Idx and D are *m*-by-1 cell arrays such that each cell contains a vector of at least *k* indices and distances, respectively. Each vector in D contains arranged distances in ascending order. Each row in Idx contains the indices of the nearest neighbors corresponding to the distances in D.

If IncludeTies is false, then knnsearch chooses the observation with the smallest index among the observations that have the same distance from a query point.

Example: 'IncludeTies',true

K — Number of nearest neighbors
1 (default) | positive integer

Number of nearest neighbors to search for in the training data per query observation, specified as the comma-separated pair consisting of 'K' and a positive integer.

Example: `'K',2`

Data Types: `single` | `double`

P — Exponent for Minkowski distance metric

2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

For Kd-Tree Nearest Neighbor Searchers

SortIndices — Flag to sort returned indices according to distance

`true` (`1`) (default) | `false` (`0`)

Flag to sort returned indices according to distance, specified as the comma-separated pair consisting of `'SortIndices'` and either `true` (`1`) or `false` (`0`).

For faster performance, you can set `SortIndices` to `false` when the following are true:

- `Y` contains many observations that have many nearest neighbors in `X`.
- `Mdl` is a `KDTreeSearcher` model object.
- `IncludeTies` is `false`.

In this case, `knnsearch` returns the indices of the nearest neighbors in no particular order. When `SortIndices` is `true`, the function arranges the nearest-neighbor indices in ascending order by distance.

`SortIndices` is `true` by default. When `Mdl` is an `ExhaustiveSearcher` model object or `IncludeTies` is `true`, the function always sorts the indices.

Example: `'SortIndices',false`

Data Types: `logical`

For Exhaustive Nearest Neighbor Searchers

Cov — Covariance matrix for Mahalanobis distance metric

`cov(Mdl.X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix. `Cov` is a K-by-K matrix, where K is the number of columns of `Mdl.X`. If you specify `Cov` and do not specify `'Distance','mahalanobis'`, then `knnsearch` returns an error message.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

Scale — Scale parameter value for standardized Euclidean distance metric

`std(Mdl.X,'omitnan')` (default) | nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector. `Scale` has length *K*, where *K* is the number of columns of `Mdl.X`.

The software scales each difference between the training and query data using the corresponding element of `Scale`. If you specify `Scale` and do not specify `'Distance','seuclidean'`, then `knnsearch` returns an error message.

Example: `'Scale',quantile(Mdl.X,0.75) - quantile(Mdl.X,0.25)`

Data Types: `single` | `double`

Note If you specify `'Distance'`, `'Cov'`, `'P'`, or `'Scale'`, then `Mdl.Distance` and `Mdl.DistParameter` do not change value.

Output Arguments

`Idx` — Training data indices of nearest neighbors

numeric matrix | cell array of numeric vectors

Training data indices of nearest neighbors, returned as a numeric matrix or cell array of numeric vectors.

- If you do not specify `IncludeTies` (`false` by default), then `Idx` is an *m*-by-*k* numeric matrix, where *m* is the number of rows in `Y` and *k* is the number of searched nearest neighbors specified by the `'K'` name-value pair argument. `Idx(j,i)` indicates that `Mdl.X(Idx(j,i),:)` is one of the *k* closest observations in `Mdl.X` to the query observation `Y(j,:)`.
- If you specify `'IncludeTies',true`, then `Idx` is an *m*-by-`1` cell array such that cell `j` (`Idx{j}`) contains a vector of at least *k* indices of the closest observations in `Mdl.X` to the query observation `Y(j,:)`.

If `SortIndices` is `true`, then `knnsearch` arranges the indices in ascending order by distance.

`D` — Distances of nearest neighbors

numeric matrix | cell array of numeric vectors

Distances of the nearest neighbors to the query data, returned as a numeric matrix or cell array of numeric vectors.

- If you do not specify `IncludeTies` (`false` by default), then `D` is an *m*-by-*k* numeric matrix, where *m* is the number of rows in `Y` and *k* is the number of searched nearest neighbors specified by the `'K'` name-value pair argument. `D(j,i)` is the distance between `Mdl.X(Idx(j,i),:)` and the query observation `Y(j,:)` with respect to the distance metric.
- If you specify `'IncludeTies',true`, then `D` is an *m*-by-`1` cell array such that cell `j` (`D{j}`) contains a vector of at least *k* distances of the closest observations in `Mdl.X` to the query observation `Y(j,:)`.

If `SortIndices` is `true`, then `knnsearch` arranges the distances in ascending order.

Tips

`knnsearch` finds the *k* (positive integer) points in `Mdl.X` that are *k*-nearest for each `Y` point. In contrast, `rangesearch` finds all the points in `Mdl.X` that are within distance `r` (positive scalar) of each `Y` point.

Alternative Functionality

- `knnsearch` is an object function that requires an `ExhaustiveSearcher` or a `KDTreeSearcher` model object and query data. Under equivalent conditions, the `knnsearch` object function returns the same results as the `knnsearch` function when you specify the name-value pair argument `'NSMethod','exhaustive'` or `'NSMethod','kdtree'`, respectively.
- For *k*-nearest neighbors classification, see `fitcknn` and `ClassificationKNN`.

References

[1] Friedman, J. H., Bentely, J., and Finkel, R. A. (1977). "An Algorithm for Finding Best Matches in Logarithmic Expected Time." *ACM Transactions on Mathematical Software* Vol. 3, Issue 3, Sept. 1977, pp. 209–226.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- This table contains notes about the arguments of `knnsearch`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	There are two ways to use `Mdl` in code generation. For an example, see “Code Generation for Nearest Neighbor Searcher” on page 32-19. • Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `knnsearch` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `knnsearch` function. Then use `codegen` to generate code for the entry-point function. • Include `coder.Constant(Mdl)` in the `-args` value of `codegen`. If `Mdl` is a `KDTreeSearcher` object, and the code generation build type is a MEX function, then `codegen` generates a MEX function using Intel Threading Building Blocks (TBB) for parallel computation. Otherwise, `codegen` generates code using `parfor`. • MEX function for the *k*d-tree search algorithm — `codegen` generates an optimized MEX function using Intel TBB for parallel computation on multicore platforms. You can use the MEX function to accelerate MATLAB algorithms. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb. If you generate the MEX function to test the generated code of the `parfor` version, you can disable the usage of Intel TBB. Set the `ExtrinsicCalls` property of the MEX configuration object to `false`. For details, see `coder.MexCodeConfig`. • MEX function for the exhaustive search algorithm and standalone C/C++ code for both algorithms — The generated code of `knnsearch` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.
`'Distance'`	• Cannot be a custom distance function. • Must be a compile-time constant; its value cannot change in the generated code.
`'IncludeTies'`	Must be a compile-time constant; its value cannot change in the generated code.
`'SortIndices'`	Not supported. The output arguments are always sorted.
Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined exponent for the Minkowski distance in the generated code, include `{coder.Constant('Distance'),coder.Constant('Minkowski'),coder.Constant('P'),0}` in the `-args` value of `codegen`.

Argument	Notes and Limitations
`Idx`	• When you specify `'IncludeTies'` as `true`, the sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision. • Starting in R2020a, `knnsearch` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for strict single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Nearest Neighbor Searcher” on page 32-19.

See Also

`ClassificationKNN` | `ExhaustiveSearcher` | `KDTreeSearcher` | `createns` | `fitcknn` | `knnsearch` | `rangesearch`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2010a

knnsearch

Find *k*-nearest neighbors using input data

Syntax

```
Idx = knnsearch(X,Y)
Idx = knnsearch(X,Y,Name,Value)
[Idx,D] = knnsearch( ___ )
```

Description

`Idx = knnsearch(X,Y)` finds the nearest neighbor in `X` for each query point in `Y` and returns the indices of the nearest neighbors in `Idx`, a column vector. `Idx` has the same number of rows as `Y`.

`Idx = knnsearch(X,Y,Name,Value)` returns `Idx` with additional options specified using one or more name-value pair arguments. For example, you can specify the number of nearest neighbors to search for and the distance metric used in the search.

`[Idx,D] = knnsearch( ___ )` additionally returns the matrix `D`, using any of the input arguments in the previous syntaxes. `D` contains the distances between each observation in `Y` and the corresponding closest observations in `X`.

Examples

Find Nearest Neighbors

Find the patients in the `hospital` data set that most closely resemble the patients in `Y`, according to age and weight.

Load the `hospital` data set.

```
load hospital;
X = [hospital.Age hospital.Weight];
Y = [20 162; 30 169; 40 168; 50 170; 60 171];   % New patients
```

Perform a `knnsearch` between `X` and `Y` to find indices of nearest neighbors.

```
Idx = knnsearch(X,Y);
```

Find the patients in `X` closest in age and weight to those in `Y`.

```
X(Idx,:)
```

```
ans = 5×2

    25   171
    25   171
    39   164
    49   170
    50   172
```

Find *k*-Nearest Neighbors Using Different Distance Metrics

Find the 10 nearest neighbors in X to each point in Y, first using the Minkowski distance metric and then using the Chebychev distance metric.

Load Fisher's iris data set.

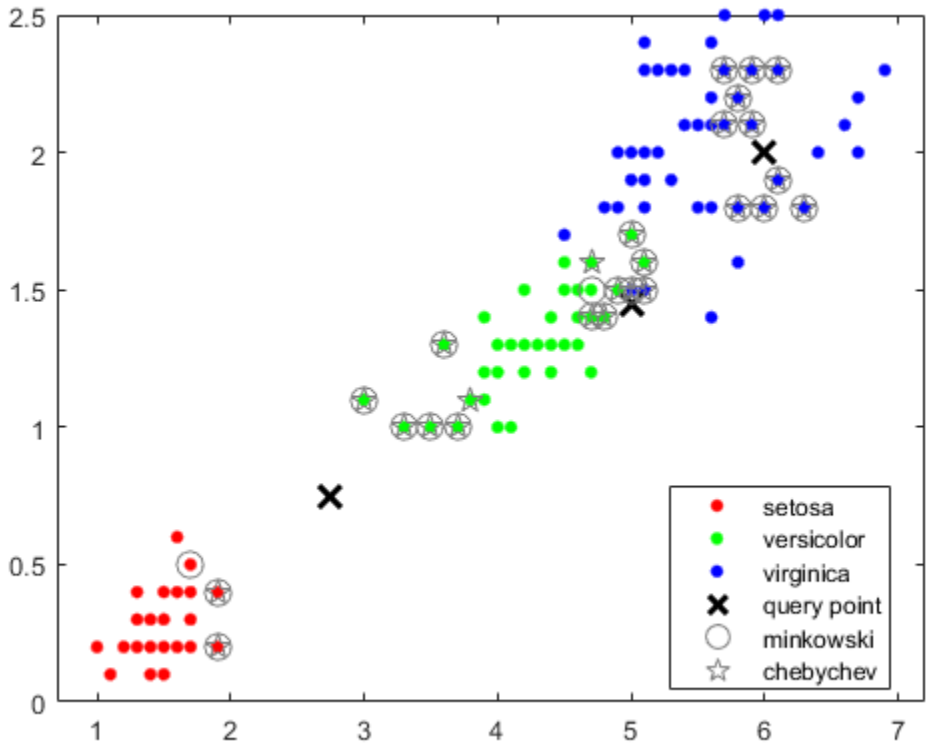
```
load fisheriris
X = meas(:,3:4);    % Measurements of original flowers
Y = [5 1.45;6 2;2.75 .75];  % New flower data
```

Perform a `knnsearch` between X and the query points Y using Minkowski and Chebychev distance metrics.

```
[mIdx,mD] = knnsearch(X,Y,'K',10,'Distance','minkowski','P',5);
[cIdx,cD] = knnsearch(X,Y,'K',10,'Distance','chebychev');
```

Visualize the results of the two nearest neighbor searches. Plot the training data. Plot the query points with the marker X. Use circles to denote the Minkowski nearest neighbors. Use pentagrams to denote the Chebychev nearest neighbors.

```
gscatter(X(:,1),X(:,2),species)
line(Y(:,1),Y(:,2),'Marker','x','Color','k',...
   'Markersize',10,'Linewidth',2,'Linestyle','none')
line(X(mIdx,1),X(mIdx,2),'Color',[.5 .5 .5],'Marker','o',...
   'Linestyle','none','Markersize',10)
line(X(cIdx,1),X(cIdx,2),'Color',[.5 .5 .5],'Marker','p',...
   'Linestyle','none','Markersize',10)
legend('setosa','versicolor','virginica','query point',...
'minkowski','chebychev','Location','best')
```



Input Arguments

X — Input data

numeric matrix

Input data, specified as a numeric matrix. Rows of X correspond to observations, and columns correspond to variables.

Data Types: `single` | `double`

Y — Query points

numeric matrix

Query points, specified as a numeric matrix. Rows of Y correspond to observations, and columns correspond to variables. Y must have the same number of columns as X.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `knnsearch(X,Y,'K',10,'IncludeTies',true,'Distance','cityblock')` searches for 10 nearest neighbors, including ties and using the city block distance.

K — Number of nearest neighbors

1 (default) | positive integer

Number of nearest neighbors to find in X for each point in Y, specified as the comma-separated pair consisting of 'K' and a positive integer.

Example: 'K',10

Data Types: single | double

IncludeTies — Flag to include all nearest neighbors

false (0) (default) | true (1)

Flag to include all nearest neighbors that have the same distance from query points, specified as the comma-separated pair consisting of 'IncludeTies' and false (0) or true (1).

If 'IncludeTies' is false, then knnsearch chooses the observation with the smallest index among the observations that have the same distance from a query point.

If 'IncludeTies' is true, then:

- knnsearch includes all nearest neighbors whose distances are equal to the *k*th smallest distance in the output arguments. To specify *k*, use the 'K' name-value pair argument.
- Idx and D are *m*-by-1 cell arrays such that each cell contains a vector of at least *k* indices and distances, respectively. Each vector in D contains distances arranged in ascending order. Each row in Idx contains the indices of the nearest neighbors corresponding to the distances in D.

Example: 'IncludeTies',true

NSMethod — Nearest neighbor search method

'kdtree' | 'exhaustive'

Nearest neighbor search method, specified as the comma-separated pair consisting of 'NSMethod' and one of these values.

- 'kdtree' — Creates and uses a *K*d-tree to find nearest neighbors. 'kdtree' is the default value when the number of columns in X is less than or equal to 10, X is not sparse, and the distance metric is 'euclidean', 'cityblock', 'chebychev', or 'minkowski'. Otherwise, the default value is 'exhaustive'.

  The value 'kdtree' is valid only when the distance metric is one of the four metrics noted above.
- 'exhaustive' — Uses the exhaustive search algorithm by computing the distance values from all the points in X to each point in Y.

Example: 'NSMethod','exhaustive'

Distance — Distance metric

'euclidean' (default) | 'seuclidean' | 'cityblock' | 'chebychev' | 'minkowski' | 'mahalanobis' | function handle | ...

Distance metric knnsearch uses, specified as the comma-separated pair consisting of 'Distance' and one of the values in this table or a function handle.

Value	Description
'euclidean'	Euclidean distance.
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows and the query matrix Y is scaled by dividing by the corresponding element standard deviation computed from X. To specify another scaling, use the 'S name-value pair argument.
'cityblock'	City block distance.
'chebychev'	Chebychev distance (maximum coordinate difference).
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different expon use the 'P' name-value pair argument.
'mahalanobis'	Mahalanobis distance, computed using a positive definite covariance matri change the value of the covariance matrix, use the 'Cov' name-value pair argument.
'cosine'	One minus the cosine of the included angle between observations (treated vectors).
'correlation'	One minus the sample linear correlation between observations (treated as sequences of values).
'spearman'	One minus the sample Spearman's rank correlation between observations (as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coor that differ.

You can also specify a function handle for a custom distance metric by using @ (for example, @distfun). A custom distance function must:

- Have the form `function D2 = distfun(ZI,ZJ)`.
- Take as arguments:
  - A 1-by-*n* vector ZI containing a single row from X or from the query points Y.
  - An m_2-by-*n* matrix ZJ containing multiple rows of X or Y.
- Return an m_2-by-1 vector of distances D2, whose jth element is the distance between the observations ZI and ZJ(j,:).

For more information, see “Distance Metrics” on page 18-12.

Example: 'Distance','chebychev'

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of 'P' and a positive scalar.

This argument is valid only if 'Distance' is 'minkowski'.

Example: 'P',3

Data Types: single | double

`Cov` — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix.

This argument is valid only if `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(4)`

Data Types: `single` | `double`

`Scale` — Scale parameter value for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector. `'Scale'` has length equal to the number of columns in `X`. When `knnsearch` computes the standardized Euclidean distance, each coordinate of `X` is scaled by the corresponding element of `'Scale'`, as is each query point. This argument is valid only when `'Distance'` is `'seuclidean'`.

Example: `'Scale',quantile(X,0.75) - quantile(X,0.25)`

Data Types: `single` | `double`

`BucketSize` — Maximum number of data points in leaf node of *K*d-tree
`50` (default) | positive integer

Maximum number of data points in the leaf node of the *K*d-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer. This argument is valid only when `NSMethod` is `'kdtree'`.

Example: `'BucketSize',20`

Data Types: `single` | `double`

`SortIndices` — Flag to sort returned indices according to distance
`true` (`1`) (default) | `false` (`0`)

Flag to sort returned indices according to distance, specified as the comma-separated pair consisting of `'SortIndices'` and either `true` (`1`) or `false` (`0`).

For faster performance, you can set `SortIndices` to `false` when the following are true:

- `Y` contains many observations that have many nearest neighbors in `X`.
- `NSMethod` is `'kdtree'`.
- `IncludeTies` is `false`.

In this case, `knnsearch` returns the indices of the nearest neighbors in no particular order. When `SortIndices` is `true`, the function arranges the nearest-neighbor indices in ascending order by distance.

`SortIndices` is `true` by default. When `NSMethod` is `'exhaustive'` or `IncludeTies` is `true`, the function always sorts the indices.

Example: `'SortIndices',false`

Data Types: `logical`

Output Arguments

`Idx` — Input data indices of nearest neighbors

numeric matrix | cell array of numeric vectors

Input data indices of the nearest neighbors, returned as a numeric matrix or cell array of numeric vectors.

- If you do not specify `IncludeTies` (`false` by default), then `Idx` is an *m*-by-*k* numeric matrix, where *m* is the number of rows in `Y` and *k* is the number of searched nearest neighbors. `Idx(j,i)` indicates that `X(Idx(j,i),:)` is one of the *k* closest observations in `X` to the query point `Y(j,:)`.
- If you specify `'IncludeTies',true`, then `Idx` is an *m*-by-`1` cell array such that cell `j` (`Idx{j}`) contains a vector of at least *k* indices of the closest observations in `X` to the query point `Y(j,:)`.

If `SortIndices` is `true`, then `knnsearch` arranges the indices in ascending order by distance.

`D` — Distances of nearest neighbors

numeric matrix | cell array of numeric vectors

Distances of the nearest neighbors to the query points, returned as a numeric matrix or cell array of numeric vectors.

- If you do not specify `IncludeTies` (`false` by default), then `D` is an *m*-by-*k* numeric matrix, where *m* is the number of rows in `Y` and *k* is the number of searched nearest neighbors. `D(j,i)` is the distance between `X(Idx(j,i),:)` and `Y(j,:)` with respect to the distance metric.
- If you specify `'IncludeTies',true`, then `D` is an *m*-by-`1` cell array such that cell `j` (`D{j}`) contains a vector of at least *k* distances of the closest observations in `X` to the query point `Y(j,:)`.

If `SortIndices` is `true`, then `knnsearch` arranges the distances in ascending order.

Tips

- For a fixed positive integer *k*, `knnsearch` finds the *k* points in `X` that are the nearest to each point in `Y`. To find all points in `X` within a fixed distance of each point in `Y`, use `rangesearch`.
- `knnsearch` does not save a search object. To create a search object, use `createns`.

Algorithms

For information on a specific search algorithm, see “k-Nearest Neighbor Search and Radius Search” on page 18-14.

Alternative Functionality

If you set the `knnsearch` function's `'NSMethod'` name-value pair argument to the appropriate value (`'exhaustive'` for an exhaustive search algorithm or `'kdtree'` for a *K*d-tree algorithm), then the search results are equivalent to the results obtained by conducting a distance search using the `knnsearch` object function. Unlike the `knnsearch` function, the `knnsearch` object function requires an `ExhaustiveSearcher` or a `KDTreeSearcher` model object.

References

[1] Friedman, J. H., J. Bentely, and R. A. Finkel. “An Algorithm for Finding Best Matches in Logarithmic Expected Time.” *ACM Transactions on Mathematical Software* 3, no. 3 (1977): 209–226.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- If `X` is a tall array, then `Y` cannot be a tall array. Similarly, if `Y` is a tall array, then `X` cannot be a tall array.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- For code generation, the default value of the `'NSMethod'` name-value pair argument is `'exhaustive'` when the number of columns in `X` is greater than 7.
- The value of the `'Distance'` name-value pair argument must be a compile-time constant and cannot be a custom distance function.
- The value of the `'IncludeTies'` name-value pair argument must be a compile-time constant.
- The `'SortIndices'` name-value pair argument is not supported. The output arguments are always sorted.
- Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined exponent for the Minkowski distance in the generated code, include `{coder.Constant('Distance'),coder.Constant('Minkowski'),coder.Constant('P'),0}` in the `-args` value of `codegen`.
- When you specify `'IncludeTies'` as `true`, the sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision.
- When `knnsearch` uses the *k*d-tree search algorithm, and the code generation build type is a MEX function, `codegen` generates a MEX function using Intel Threading Building Blocks (TBB) for parallel computation. Otherwise, `codegen` generates code using `parfor`.
  - MEX function for the *k*d-tree search algorithm — `codegen` generates an optimized MEX function using Intel TBB for parallel computation on multicore platforms. You can use the MEX function to accelerate MATLAB algorithms. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

    If you generate the MEX function to test the generated code of the `parfor` version, you can disable the usage of Intel TBB. Set the `ExtrinsicCalls` property of the MEX configuration object to `false`. For details, see `coder.MexCodeConfig`.
  - MEX function for the exhaustive search algorithm and standalone C/C++ code for both algorithms — The generated code of `knnsearch` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your

compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.

- Starting in R2020a, `knnsearch` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for strict single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'IncludeTies'`, `'NSMethod'`, and `'SortIndices'` name-value pair arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ExhaustiveSearcher` | `KDTreeSearcher` | `createns` | `knnsearch`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2010a

kruskalwallis

Kruskal-Wallis test

Syntax

```
p = kruskalwallis(x)
p = kruskalwallis(x,group)
p = kruskalwallis(x,group,displayopt)
[p,tbl,stats] = kruskalwallis( ___ )
```

Description

`p = kruskalwallis(x)` returns the *p*-value for the null hypothesis that the data in each column of the matrix `x` comes from the same distribution, using a Kruskal-Wallis test on page 33-3376. The alternative hypothesis is that not all samples come from the same distribution. `kruskalwallis` also returns an ANOVA table and a box plot.

`p = kruskalwallis(x,group)` returns the *p*-value for a test of the null hypothesis that the data in each categorical group, as specified by the grouping variable `group` comes from the same distribution. The alternative hypothesis is that not all groups come from the same distribution.

`p = kruskalwallis(x,group,displayopt)` returns the *p*-value of the test and lets you display or suppress the ANOVA table and box plot.

`[p,tbl,stats] = kruskalwallis( ___ )` also returns the ANOVA table as the cell array `tbl` and the structure `stats` containing information about the test statistics.

Examples

Test Data Samples for the Same Distribution

Create two different normal probability distribution objects. The first distribution has `mu = 0` and `sigma = 1`, and the second distribution has `mu = 2` and `sigma = 1`.

```
pd1 = makedist('Normal');
pd2 = makedist('Normal','mu',2,'sigma',1);
```

Create a matrix of sample data by generating random numbers from these two distributions.

```
rng('default'); % for reproducibility
x = [random(pd1,20,2),random(pd2,20,1)];
```

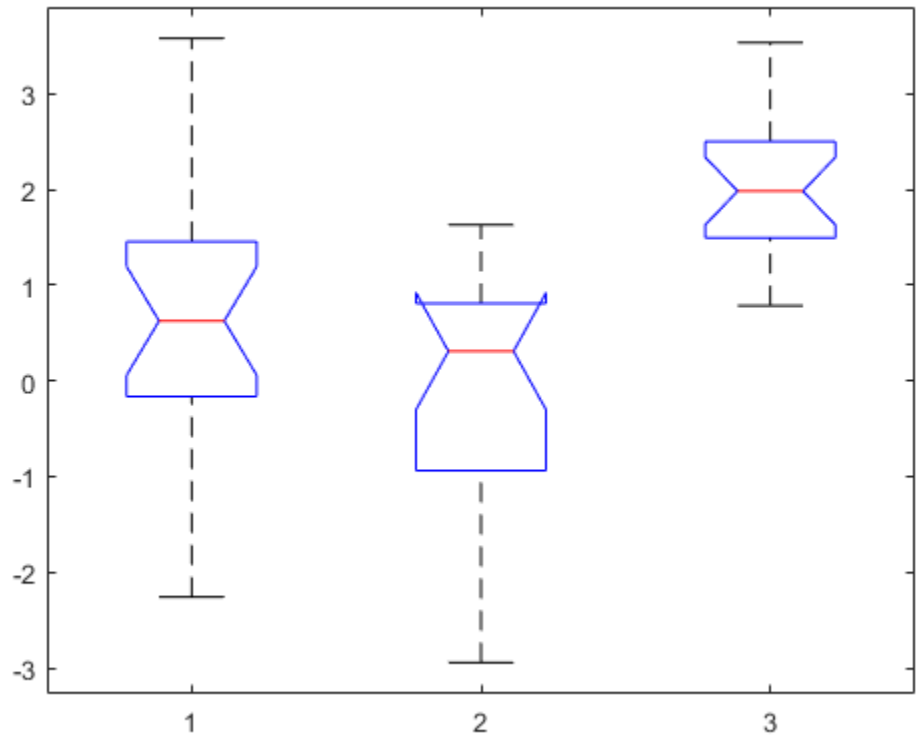
The first two columns of `x` contain data generated from the first distribution, while the third column contains data generated from the second distribution.

Test the null hypothesis that the sample data from each column in `x` comes from the same distribution.

```
p = kruskalwallis(x)
```

Kruskal-Wallis ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	7631.1	2	3815.55	25.02	3.68957e-06
Error	10363.9	57	181.82		
Total	17995	59			




```
p = 3.6896e-06
```

The returned value of `p` indicates that `kruskalwallis` rejects the null hypothesis that all three data samples come from the same distribution at a 1% significance level. The ANOVA table provides additional test results, and the box plot visually presents the summary statistics for each column in `x`.

Conduct Followup Tests for Unequal Medians

Create two different normal probability distribution objects. The first distribution has `mu = 0` and `sigma = 1`. The second distribution has `mu = 2` and `sigma = 1`.

```
pd1 = makedist('Normal');
pd2 = makedist('Normal','mu',2,'sigma',1);
```

Create a matrix of sample data by generating random numbers from these two distributions.

```
rng('default'); % for reproducibility
x = [random(pd1,20,2),random(pd2,20,1)];
```

The first two columns of `x` contain data generated from the first distribution, while the third column contains data generated from the second distribution.

Test the null hypothesis that the sample data from each column in `x` comes from the same distribution. Suppress the output displays, and generate the structure `stats` to use in further testing.

```
[p,tbl,stats] = kruskalwallis(x,[],'off')
```

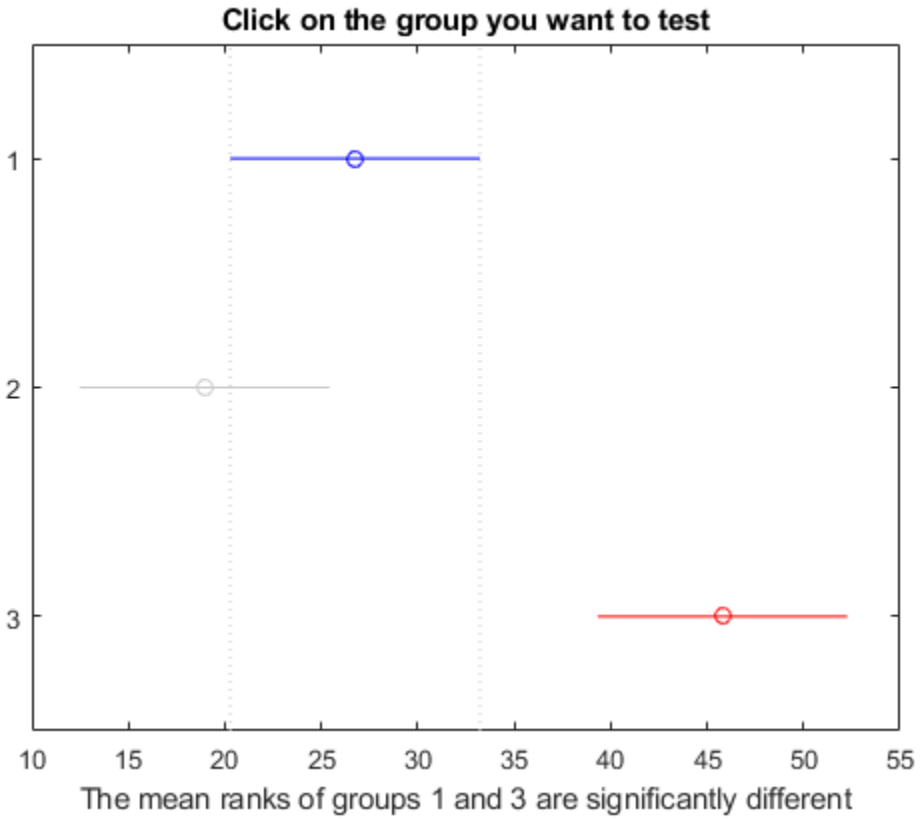
```
p = 3.6896e-06

tbl=4×6 cell array
  Columns 1 through 5

    {'Source' }    {'SS'        }    {'df'}    {'MS'        }    {'Chi-sq'  }
    {'Columns'}    {[7.6311e+03]}    {[ 2]}    {[3.8155e+03]}    {[ 25.0200]}
    {'Error'  }    {[1.0364e+04]}    {[57]}    {[  181.8228]}    {0x0 double}
    {'Total'  }    {[     17995]}    {[59]}    {0x0 double  }    {0x0 double}

  Column 6

    {'Prob>Chi-sq'}
    {[ 3.6896e-06]}
    {0x0 double   }
    {0x0 double   }


stats = struct with fields:
       gnames: [3x1 char]
            n: [20 20 20]
       source: 'kruskalwallis'
    meanranks: [26.7500 18.9500 45.8000]
         sumt: 0
```

The returned value of `p` indicates that the test rejects the null hypothesis at the 1% significance level. You can use the structure `stats` to perform additional followup testing. The cell array `tbl` contains the same data as the graphical ANOVA table, including column and row labels.

Conduct a followup test to identify which data sample comes from a different distribution.

```
c = multcompare(stats)
```

```
Note: Intervals can be used for testing but are not simultaneous confidence intervals.
```




```
c = 3×6

    1.0000    2.0000   -5.1435    7.8000   20.7435    0.3345
    1.0000    3.0000  -31.9935  -19.0500   -6.1065    0.0016
    2.0000    3.0000  -39.7935  -26.8500  -13.9065    0.0000
```

The results indicate that there is a significant difference between groups 1 and 3, so the test rejects the null hypothesis that the data in these two groups comes from the same distribution. The same is true for groups 2 and 3. However, there is not a significant difference between groups 1 and 2, so the test does not reject the null hypothesis that these two groups come from the same distribution. Therefore, these results suggest that the data in groups 1 and 2 come from the same distribution, and the data in group 3 comes from a different distribution.

Test for the Same Distribution Across Groups

Create a vector, `strength`, containing measurements of the strength of metal beams. Create a second vector, `alloy`, indicating the type of metal alloy from which the corresponding beam is made.

```
strength = [82 86 79 83 84 85 86 87 74 82 ...
            78 75 76 77 79 79 77 78 82 79];

alloy = {'st','st','st','st','st','st','st','st',...
         'al1','al1','al1','al1','al1','al1',...
         'al2','al2','al2','al2','al2','al2'};
```

Test the null hypothesis that the beam strength measurements have the same distribution across all three alloys.

```
p = kruskalwallis(strength,alloy,'off')
```

```
p = 0.0018
```

The returned value of p indicates that the test rejects the null hypothesis at the 1% significance level.

Input Arguments

x — Sample data

vector | matrix

Sample data for the hypothesis test, specified as a vector or an *m*-by-*n* matrix. If x is an *m*-by-*n* matrix, each of the *n* columns represents an independent sample containing *m* mutually independent observations.

Data Types: `single` | `double`

group — Grouping variable

numeric vector | logical vector | character array | string array | cell array of character vectors

Grouping variable, specified as a numeric or logical vector, a character or string array, or a cell array of character vectors.

- If x is a vector, then each element in `group` identifies the group to which the corresponding element in x belongs, and `group` must be a vector of the same length as x. If a row of `group` contains an empty value, that row and the corresponding observation in x are disregarded. `NaN` values in either x or `group` are similarly ignored.
- If x is a matrix, then each column in x represents a different group, and you can use `group` to specify labels for these columns. The number of elements in `group` and the number of columns in x must be equal.

The labels contained in `group` also annotate the box plot.

Example: `{'red','blue','green','blue','red','blue','green','green','red'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

displayopt — Display option

`'on'` (default) | `'off'`

Display option, specified as `'on'` or `'off'`. If `displayopt` is `'on'`, `kruskalwallis` displays the following figures:

- An ANOVA table containing the sums of squares, degrees of freedom, and other quantities calculated based on the ranks of the data in x.
- A box plot of the data in each column of the data matrix x. The box plots are based on the actual data values, rather than on the ranks.

If `displayopt` is `'off'`, `kruskalwallis` does not display these figures.

If you specify a value for `displayopt`, you must also specify a value for `group`. If you do not have a grouping variable, specify `group` as `[]`.

Example: `'off'`

Output Arguments

p — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`tbl` — ANOVA table
cell array

ANOVA table of test results, returned as a cell array. `tbl` includes the sums of squares, degrees of freedom, and other quantities calculated based on the ranks of the data in `x`, as well as column and row labels.

`stats` — Test data
structure

Test data, returned as a structure. You can perform followup multiple comparison tests on pairs of sample medians by using `multcompare`, with `stats` as the input value.

More About

Kruskal-Wallis Test

The Kruskal-Wallis test is a nonparametric version of classical one-way ANOVA, and an extension of the Wilcoxon rank sum test to more than two groups. The Kruskal-Wallis test is valid for data that has two or more groups. It compares the medians of the groups of data in `x` to determine if the samples come from the same population (or, equivalently, from different populations with the same distribution).

The Kruskal-Wallis test uses ranks of the data, rather than numeric values, to compute the test statistics. It finds ranks by ordering the data from smallest to largest across all groups, and taking the numeric index of this ordering. The rank for a tied observation is equal to the average rank of all observations tied with it. The *F*-statistic used in classical one-way ANOVA is replaced by a chi-square statistic, and the *p*-value measures the significance of the chi-square statistic.

The Kruskal-Wallis test assumes that all samples come from populations having the same continuous distribution, apart from possibly different locations due to group effects, and that all observations are mutually independent. By contrast, classical one-way ANOVA replaces the first assumption with the stronger assumption that the populations have normal distributions.

See Also

`anova1` | `boxplot` | `friedman` | `multcompare` | `ranksum`

Topics
“Grouping Variables” on page 2-45

Introduced before R2006a

ksdensity

Kernel smoothing function estimate for univariate and bivariate data

Syntax

```
[f,xi] = ksdensity(x)
[f,xi] = ksdensity(x,pts)
[f,xi] = ksdensity( ___ ,Name,Value)
[f,xi,bw] = ksdensity( ___ )

ksdensity( ___ )
ksdensity(ax, ___ )
```

Description

`[f,xi] = ksdensity(x)` returns a probability density estimate, `f`, for the sample data in the vector or two-column matrix `x`. The estimate is based on a normal kernel function, and is evaluated at equally-spaced points, `xi`, that cover the range of the data in `x`. `ksdensity` estimates the density at 100 points for univariate data, or 900 points for bivariate data.

`ksdensity` works best with continuously distributed samples.

`[f,xi] = ksdensity(x,pts)` specifies points (`pts`) to evaluate `f`. Here, `xi` and `pts` contain identical values.

`[f,xi] = ksdensity( ___ ,Name,Value)` uses additional options specified by one or more name-value pair arguments in addition to any of the input arguments in the previous syntaxes. For example, you can define the function type `ksdensity` evaluates, such as probability density, cumulative probability, survivor function, and so on. Or you can specify the bandwidth of the smoothing window.

`[f,xi,bw] = ksdensity( ___ )` also returns the bandwidth of the kernel smoothing window, `bw`. The default bandwidth is the optimal for normal densities.

`ksdensity( ___ )` plots the kernel smoothing function estimate.

`ksdensity(ax, ___ )` plots the results using axes with the handle, `ax`, instead of the current axes returned by `gca`.

Examples

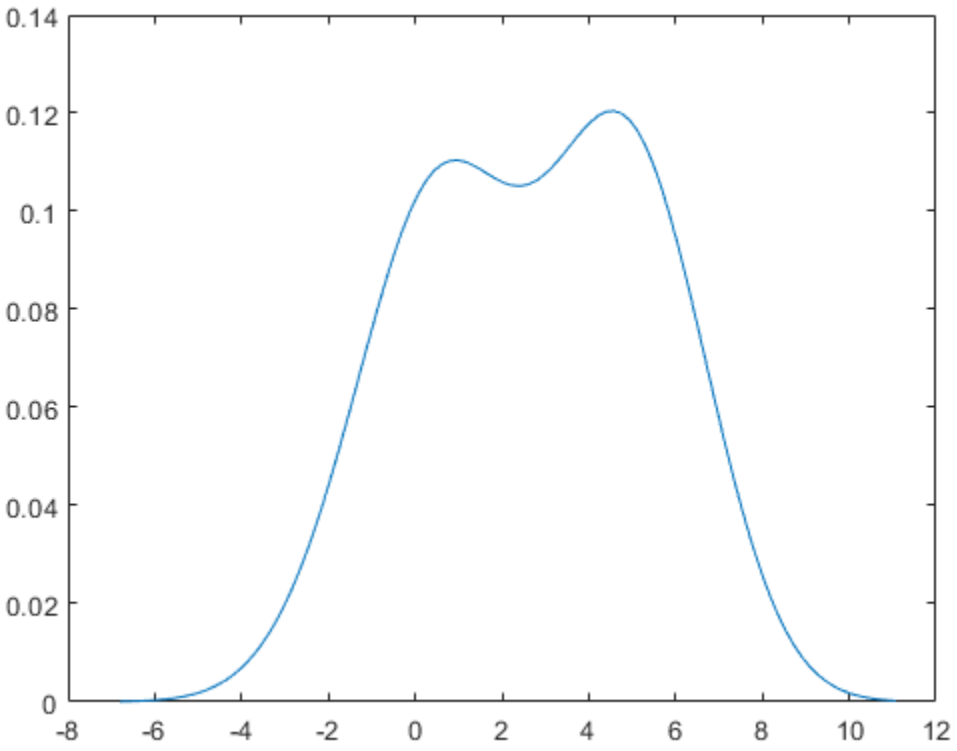
Estimate Density

Generate a sample data set from a mixture of two normal distributions.

```
rng('default')  % For reproducibility
x = [randn(30,1); 5+randn(30,1)];
```

Plot the estimated density.

```
[f,xi] = ksdensity(x);
figure
plot(xi,f);
```



The density estimate shows the bimodality of the sample.

Estimate Density with Boundary Correction

Generate a nonnegative sample data set from the half-normal distribution.

```
rng('default') % For reproducibility
pd = makedist('HalfNormal','mu',0,'sigma',1);
x = random(pd,100,1);
```

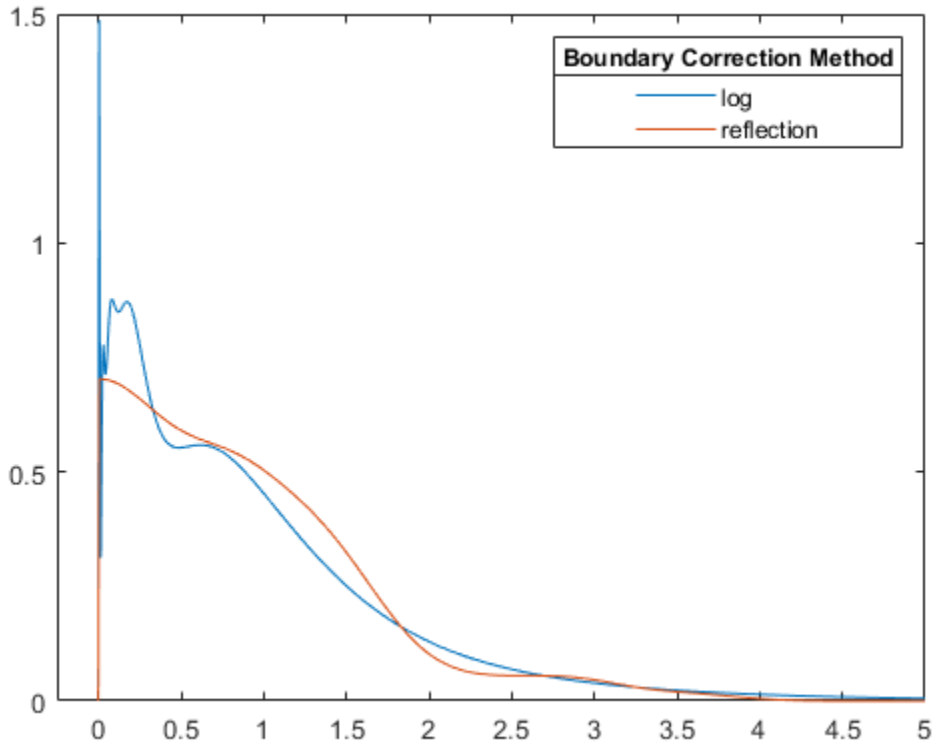
Estimate pdfs with two different boundary correction methods, log transformation and reflection, by using the `'BoundaryCorrection'` name-value pair argument.

```
pts = linspace(0,5,1000); % points to evaluate the estimator
[f1,xi1] = ksdensity(x,pts,'Support','positive');
[f2,xi2] = ksdensity(x,pts,'Support','positive','BoundaryCorrection','reflection');
```

Plot the two estimated pdfs.

```
plot(xi1,f1,xi2,f2)
lgd = legend('log','reflection');
title(lgd, 'Boundary Correction Method')
```

```
xl = xlim;
xlim([xl(1)-0.25 xl(2)])
```



`ksdensity` uses a boundary correction method when you specify either positive or bounded support. The default boundary correction method is log transformation. When `ksdensity` transforms the support back, it introduces the `1/x` term in the kernel density estimator. Therefore, the estimate has a peak near `x = 0`. On the other hand, the reflection method does not cause undesirable peaks near the boundary.

Estimate Cumulative Distribution Function at Specified Values

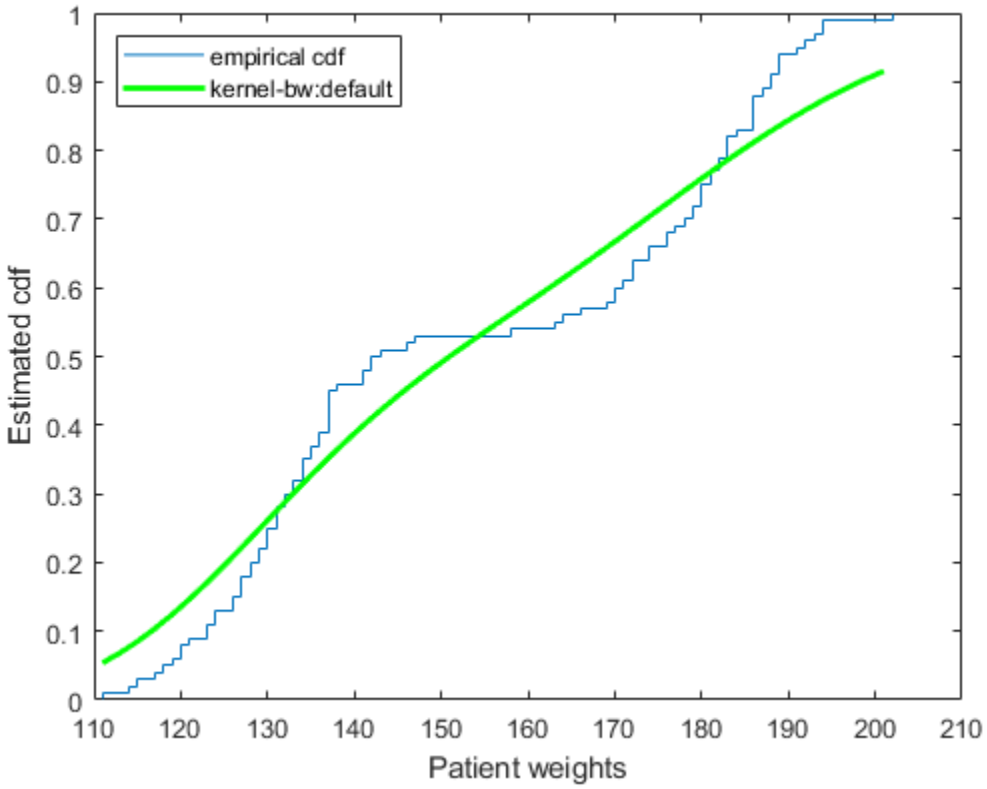
Load the sample data.

```
load hospital
```

Compute and plot the estimated cdf evaluated at a specified set of values.

```
pts = (min(hospital.Weight):2:max(hospital.Weight));
figure()
ecdf(hospital.Weight)
hold on
[f,xi,bw] = ksdensity(hospital.Weight,pts,'Support','positive',...
    'Function','cdf');
plot(xi,f,'-g','LineWidth',2)
legend('empirical cdf','kernel-bw:default','Location','northwest')
```

```
xlabel('Patient weights')
ylabel('Estimated cdf')
```



`ksdensity` seems to smooth the cumulative distribution function estimate too much. An estimate with a smaller bandwidth might produce a closer estimate to the empirical cumulative distribution function.

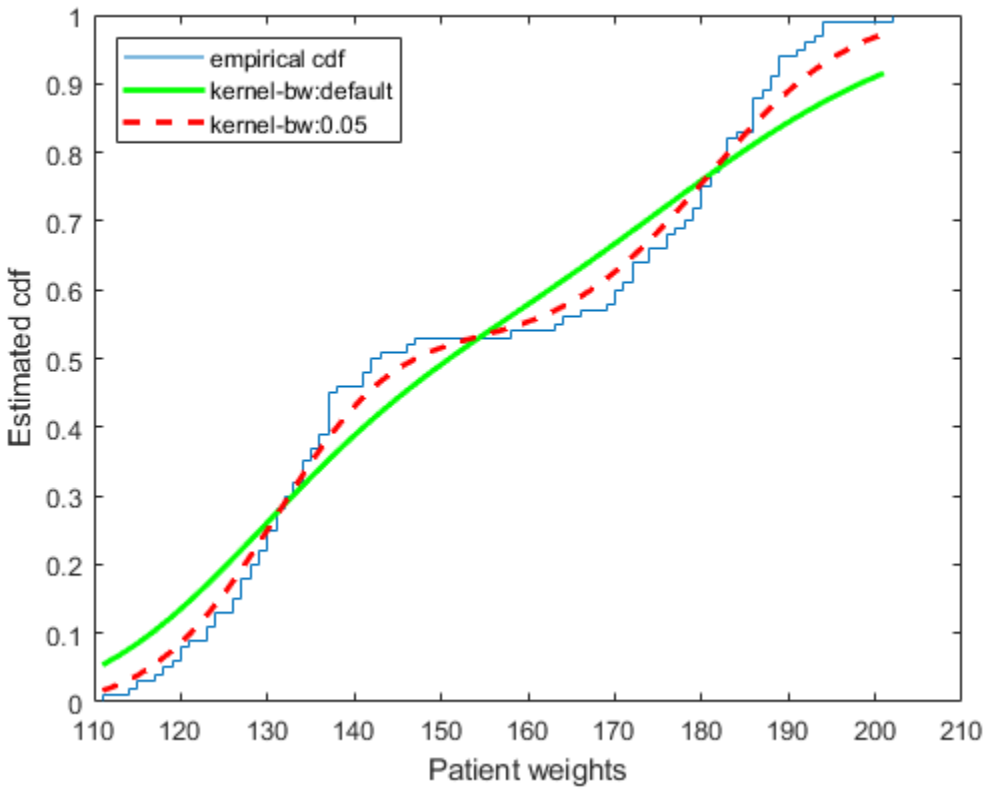
Return the bandwidth of the smoothing window.

```
bw
```

```
bw = 0.1070
```

Plot the cumulative distribution function estimate using a smaller bandwidth.

```
[f,xi] = ksdensity(hospital.Weight,pts,'Support','positive',...
    'Function','cdf','Bandwidth',0.05);
plot(xi,f,'--r','LineWidth',2)
legend('empirical cdf','kernel-bw:default','kernel-bw:0.05',...
    'Location','northwest')
hold off
```



The `ksdensity` estimate with a smaller bandwidth matches the empirical cumulative distribution function better.

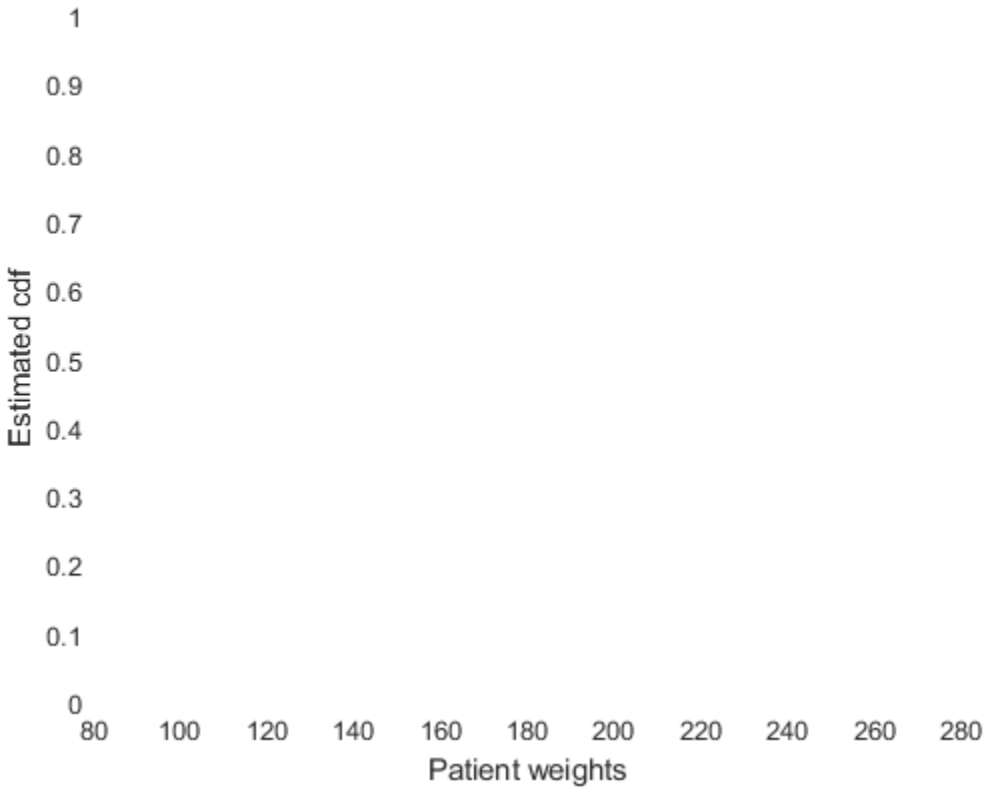
Plot Estimated Cumulative Density Function for Given Number of Points

Load the sample data.

```
load hospital
```

Plot the estimated cdf evaluated at 50 equally spaced points.

```
figure()
ksdensity(hospital.Weight,'Support','positive','Function','cdf',...
'NumPoints',50)
xlabel('Patient weights')
ylabel('Estimated cdf')
```



Estimate Survivor and Cumulative Hazard for Censored Failure Data

Generate sample data from an exponential distribution with mean 3.

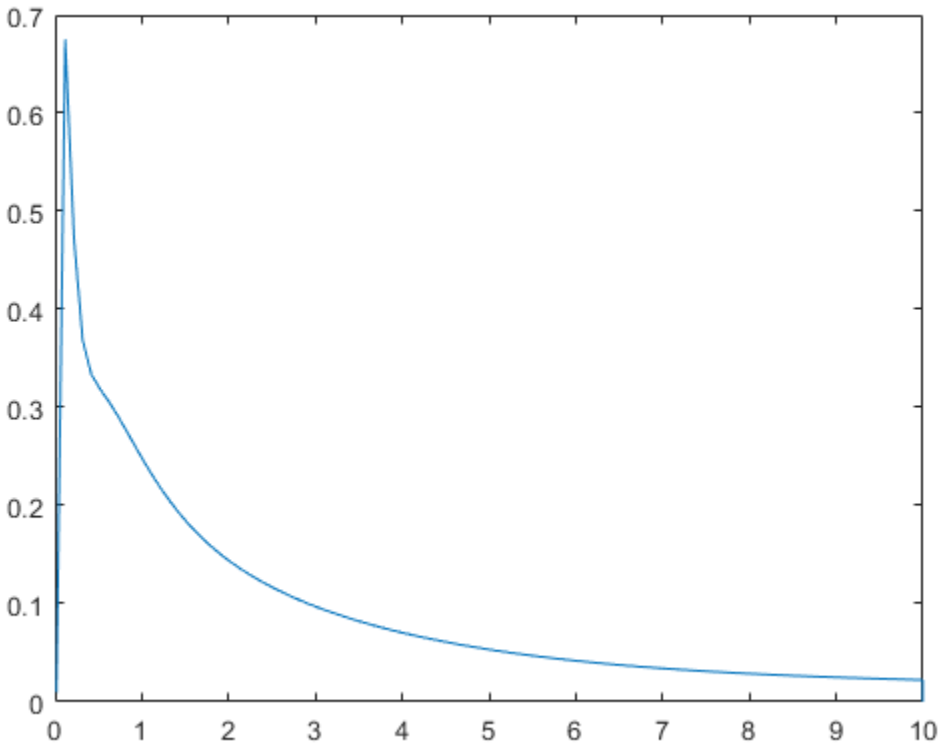
```
rng('default')  % For reproducibility
x = random('exp',3,100,1);
```

Create a logical vector that indicates censoring. Here, observations with lifetimes longer than 10 are censored.

```
T = 10;
cens = (x>T);
```

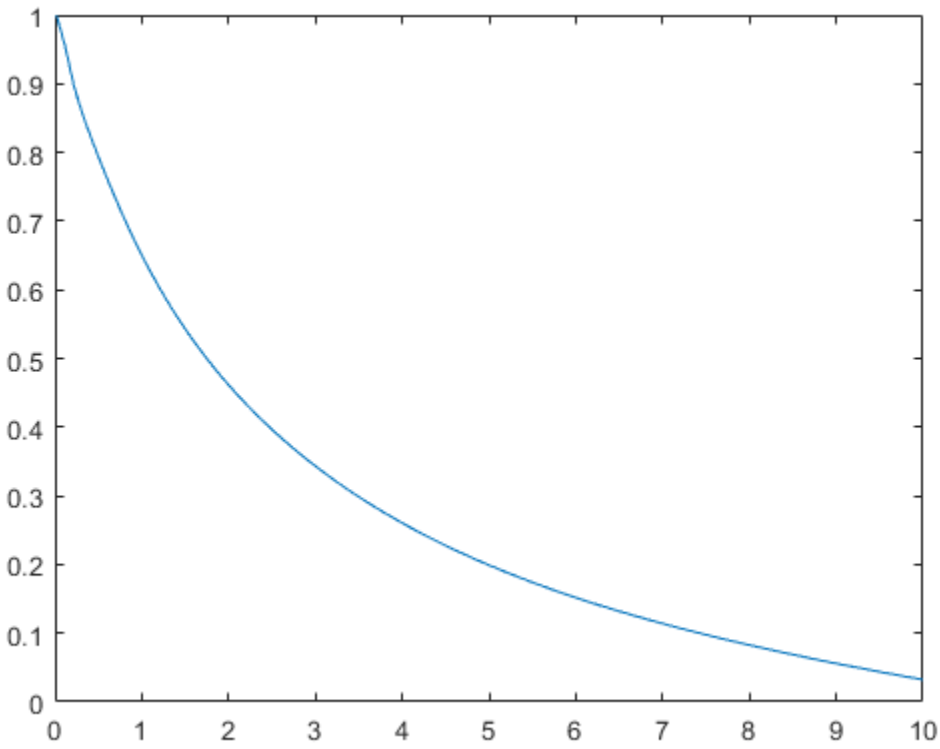
Compute and plot the estimated density function.

```
figure
ksdensity(x,'Support','positive','Censoring',cens);
```



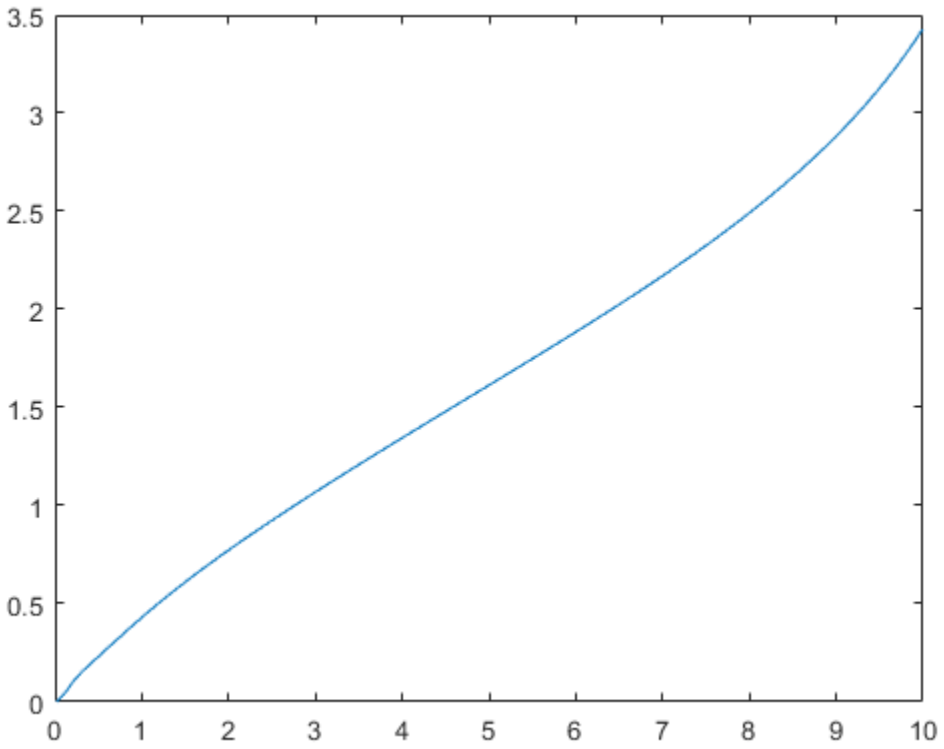
Compute and plot the survivor function.

```
figure
ksdensity(x,'Support','positive','Censoring',cens,...
'Function','survivor');
```



Compute and plot the cumulative hazard function.

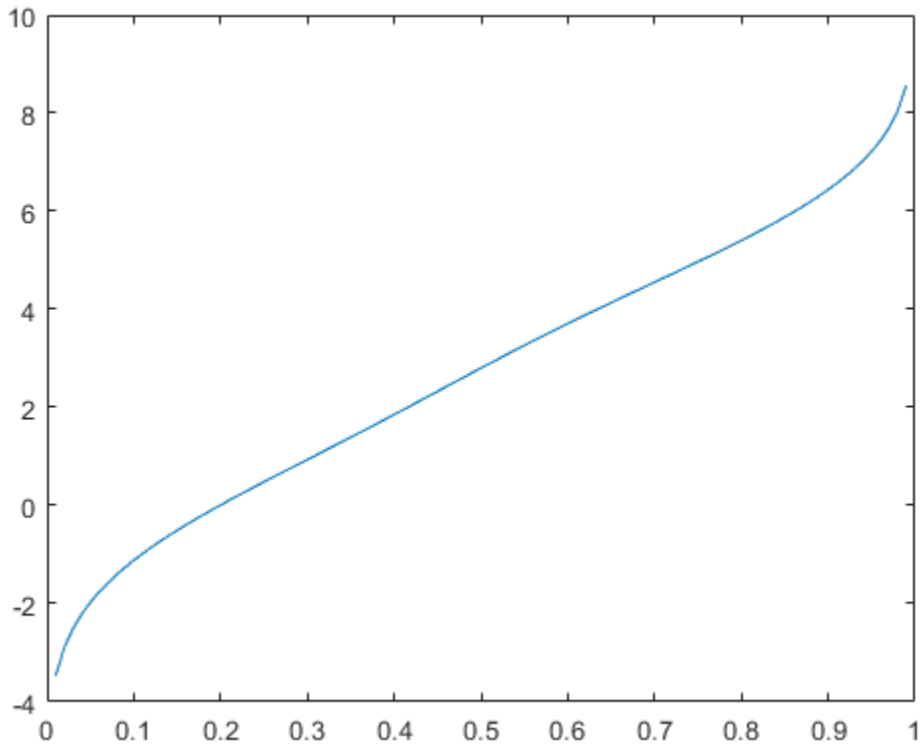
```
figure
ksdensity(x,'Support','positive','Censoring',cens,...
'Function','cumhazard');
```



Estimate Inverse Cumulative Distribution Function for Specified Probability Values

Generate a mixture of two normal distributions, and plot the estimated inverse cumulative distribution function at a specified set of probability values.

```
rng('default')  % For reproducibility
x = [randn(30,1); 5+randn(30,1)];
pi = linspace(.01,.99,99);
figure
ksdensity(x,pi,'Function','icdf');
```



Return Bandwidth of Smoothing Window

Generate a mixture of two normal distributions.

```
rng('default')  % For reproducibility
x = [randn(30,1); 5+randn(30,1)];
```

Return the bandwidth of the smoothing window for the probability density estimate.

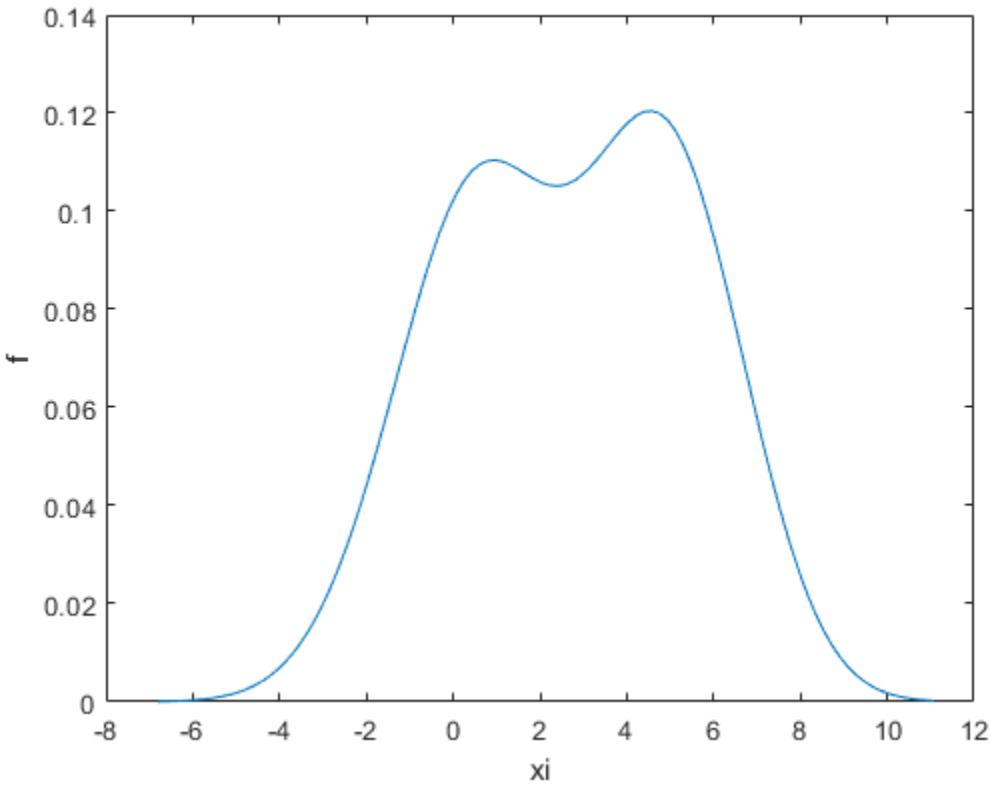
```
[f,xi,bw] = ksdensity(x);
bw
```

```
bw = 1.5141
```

The default bandwidth is optimal for normal densities.

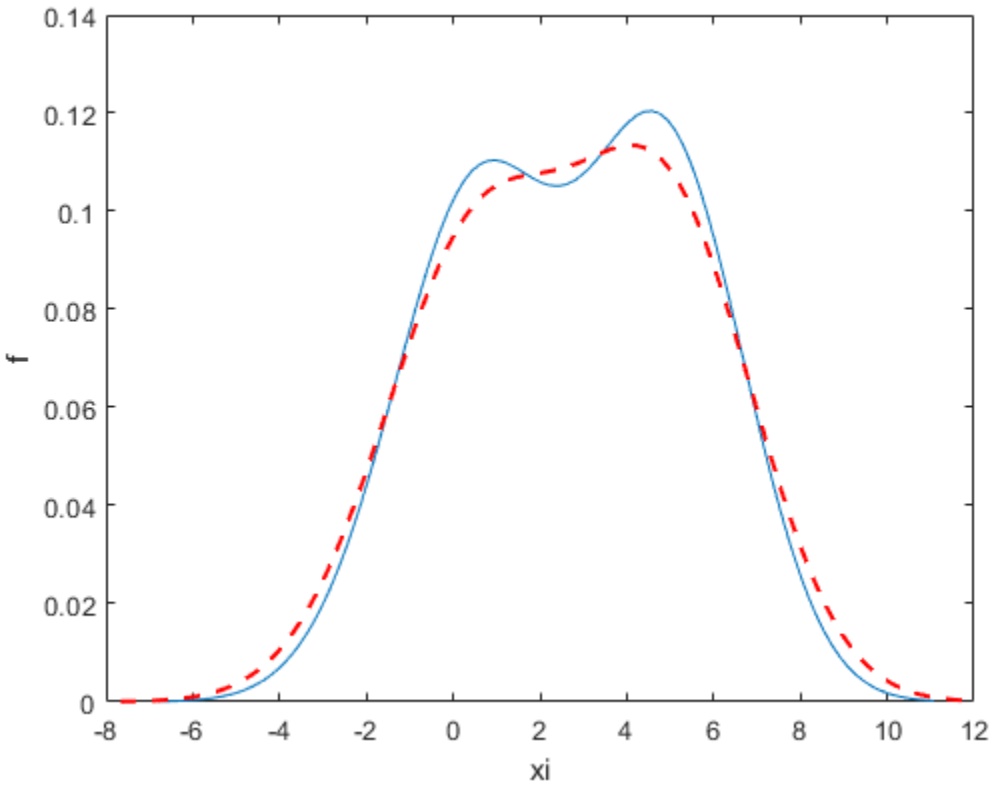
Plot the estimated density.

```
figure
plot(xi,f);
xlabel('xi')
ylabel('f')
hold on
```



Plot the density using an increased bandwidth value.

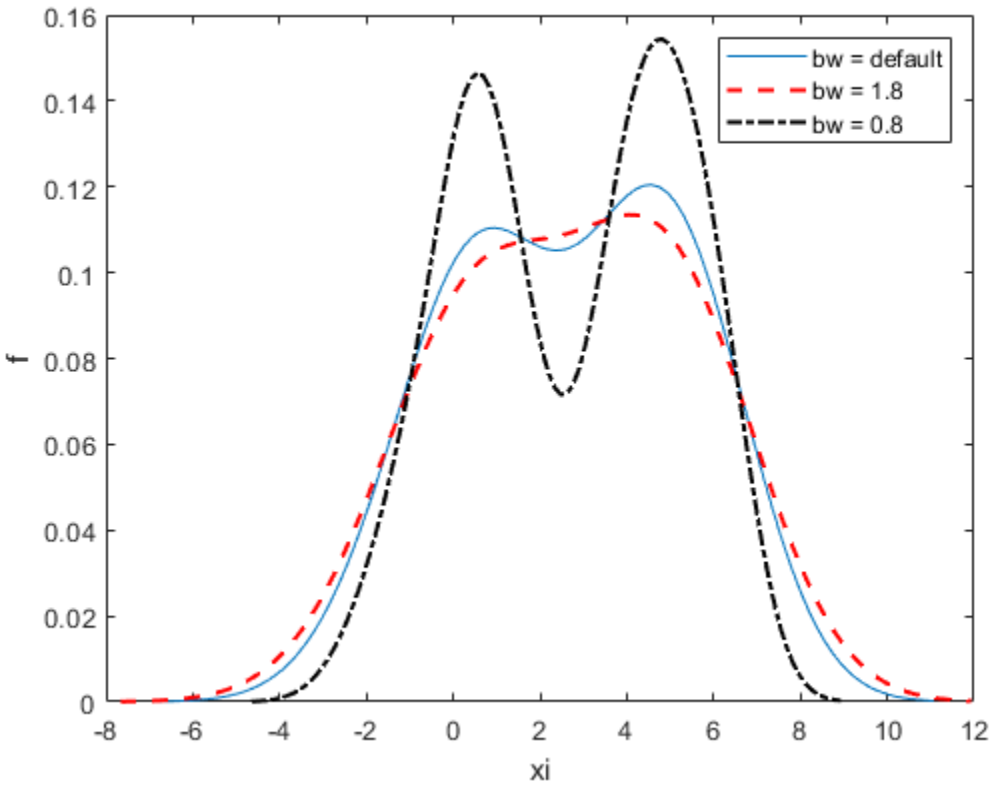
```
[f,xi] = ksdensity(x,'Bandwidth',1.8);
plot(xi,f,'--r','LineWidth',1.5)
```



A higher bandwidth further smooths the density estimate, which might mask some characteristics of the distribution.

Now, plot the density using a decreased bandwidth value.

```
[f,xi] = ksdensity(x,'Bandwidth',0.8);
plot(xi,f,'-.k','LineWidth',1.5)
legend('bw = default','bw = 1.8','bw = 0.8')
hold off
```



A smaller bandwidth smooths the density estimate less, which exaggerates some characteristics of the sample.

Plot Kernel Density Estimate of Bivariate Data

Create a two-column vector of points at which to evaluate the density.

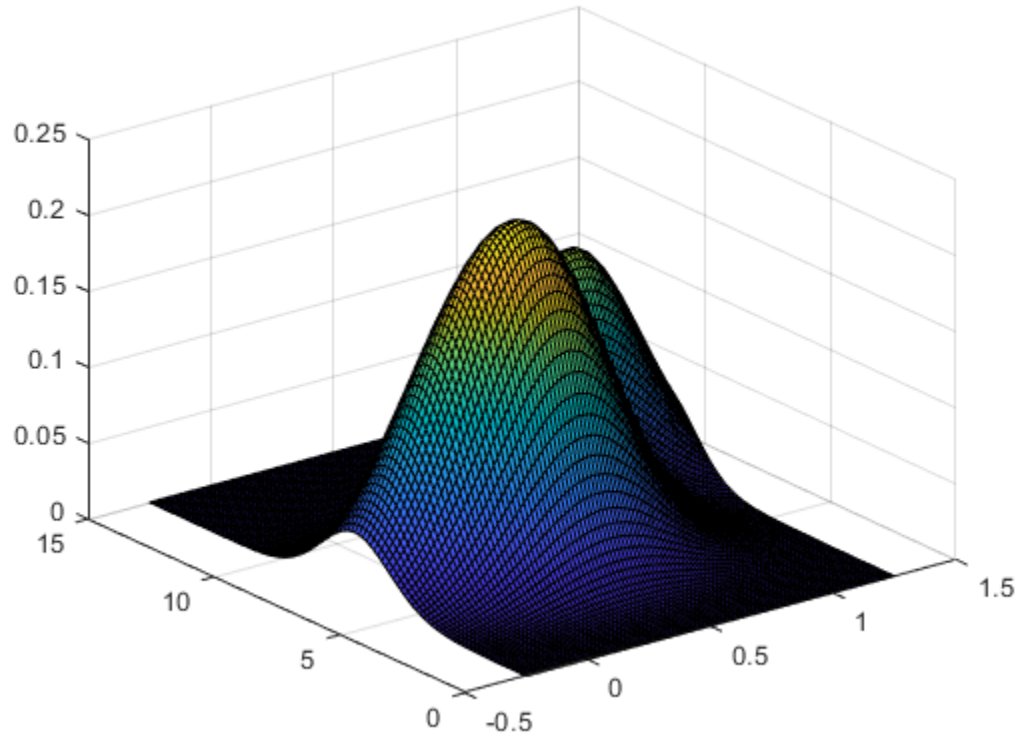
```
gridx1 = -0.25:.05:1.25;
gridx2 = 0:.1:15;
[x1,x2] = meshgrid(gridx1, gridx2);
x1 = x1(:);
x2 = x2(:);
xi = [x1 x2];
```

Generate a 30-by-2 matrix containing random numbers from a mixture of bivariate normal distributions.

```
rng('default')  % For reproducibility
x = [0+.5*rand(20,1) 5+2.5*rand(20,1);
            .75+.25*rand(10,1) 8.75+1.25*rand(10,1)];
```

Plot the estimated density of the sample data.

```
figure
ksdensity(x,xi);
```



Input Arguments

x — Sample data

column vector | two-column matrix

Sample data for which `ksdensity` returns `f` values, specified as a column vector or two-column matrix. Use a column vector for univariate data, and a two-column matrix for bivariate data.

Example: `[f,xi] = ksdensity(x)`

Data Types: `single` | `double`

pts — Points at which to evaluate f

vector | two-column matrix

Points at which to evaluate `f`, specified as a vector or two-column matrix. For univariate data, `pts` can be a row or column vector. The length of the returned output `f` is equal to the number of points in `pts`.

Example: `pts = (0:1:25); ksdensity(x,pts);`

Data Types: `single` | `double`

ax — Axes handle

handle

Axes handle for the figure `ksdensity` plots to, specified as a handle.

For example, if h is a handle for a figure, then ksdensity can plot to that figure as follows.

Example: ksdensity(h,x)

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example: 'Censoring',cens,'Kernel','triangle','NumPoints',20,'Function','cdf' specifies that ksdensity estimates the cdf by evaluating at 20 equally spaced points that covers the range of data, using the triangle kernel smoothing function and accounting for the censored data information in vector cens.

Bandwidth — Bandwidth of the kernel smoothing window

optimal value for normal densities (default) | scalar value | two-element vector

The bandwidth of the kernel-smoothing window, which is a function of the number of points in x, specified as the comma-separated pair consisting of 'Bandwidth' and a scalar value. If the sample data is bivariate, Bandwidth can also be a two-element vector. The default is optimal for estimating normal densities [1], but you might want to choose a larger or smaller value to smooth more or less.

If you specify 'BoundaryCorrection' as 'log'(default) and 'Support' as either 'positive' or a vector [L U], ksdensity converts bounded data to be unbounded by using log transformation. The value of 'Bandwidth' is on the scale of the transformed values.

Example: 'Bandwidth',0.8

Data Types: single | double

BoundaryCorrection — Boundary correction method

'log' (default) | 'reflection'

Boundary correction method, specified as the comma-separated pair consisting of 'BoundaryCorrection' and 'log' or 'reflection'.

Value	Description
'log'	ksdensity converts bounded data x to be unbounded by one of the following transformations. Then, it transforms back to the original bounded scale after density estimation. • For univariate data, if you specify 'Support','positive', then ksdensity applies log(x). • For univariate data, if you specify 'Support',[L U], where L and U are numeric scalars and L < U, then ksdensity applies log((x-L)/(U–x)). • For bivariate data, ksdensity transforms each column of x in the same way with the univariate data. The value of 'Bandwidth' and the bw output are on the scale of the transformed values.
'reflection'	ksdensity augments bounded data by adding reflected data near the boundaries, then it returns estimates corresponding to the original support. For details, see “Reflection Method” on page 33-3396.

`ksdensity` applies boundary correction only when you specify `'Support'` as a value other than `'unbounded'`.

Example: `'BoundaryCorrection','reflection'`

`Censoring` — Logical vector
vector of 0s (default) | vector of 0s and 1s

Logical vector indicating which entries are censored, specified as the comma-separated pair consisting of `'Censoring'` and a vector of binary values. A value of 0 indicates there is no censoring, 1 indicates that observation is censored. Default is there is no censoring. This name-value pair is only valid for univariate data.

Example: `'Censoring',censdata`

Data Types: `logical`

`Function` — Function to estimate
`'pdf'` (default) | `'cdf'` | `'icdf'` | `'survivor'` | `'cumhazard'`

Function to estimate, specified as the comma-separated pair consisting of `'Function'` and one of the following.

Value	Description
`'pdf'`	Probability density function.
`'cdf'`	Cumulative distribution function.
`'icdf'`	Inverse cumulative distribution function. `ksdensity` computes the estimated inverse cdf of the values in `x`, and evaluates it at the probability values specified in `pi`. This value is valid only for univariate data.
`'survivor'`	Survivor function.
`'cumhazard'`	Cumulative hazard function. This value is valid only for univariate data.

Example: `'Function','icdf'`

`Kernel` — Type of kernel smoother
`'normal'` (default) | `'box'` | `'triangle'` | `'epanechnikov'` | function handle | character vector | string scalar

Type of kernel smoother, specified as the comma-separated pair consisting of `'Kernel'` and one of the following.

- `'normal'` (default)
- `'box'`
- `'triangle'`
- `'epanechnikov'`
- A kernel function that is a custom or built-in function. Specify the function as a function handle (for example, `@myfunction` or `@normpdf`) or as a character vector or string scalar (for example, `'myfunction'` or `'normpdf'`). The software calls the specified function with one argument that

is an array of distances between data values and locations where the density is evaluated. The function must return an array of the same size containing corresponding values of the kernel function.

When `'Function'` is `'pdf'`, the kernel function returns density values. Otherwise, it returns cumulative probability values.

Specifying a custom kernel when `'Function'` is `'icdf'` returns an error.

For bivariate data, `ksdensity` applies the same kernel to each dimension.

Example: `'Kernel','box'`

`NumPoints` — Number of equally spaced points
100 (default) | scalar value

Number of equally spaced points in `xi`, specified as the comma-separated pair consisting of `'NumPoints'` and a scalar value. This name-value pair is only valid for univariate data.

For example, for a kernel smooth estimate of a specified function at 80 equally spaced points within the range of sample data, input:

Example: `'NumPoints',80`

Data Types: `single` | `double`

`Support` — Support for the density
`'unbounded'` (default) | `'positive'` | two-element vector, `[L U]` | two-by-two matrix, `[L1 L2; U1 U2]`

Support for the density, specified as the comma-separated pair consisting of `'support'` and one of the following.

Value	Description
`'unbounded'`	Default. Allow the density to extend over the whole real line.
`'positive'`	Restrict the density to positive values.
Two-element vector, `[L U]`	Give the finite lower and upper bounds for the support of the density. This option is only valid for univariate sample data.
Two-by-two matrix, `[L1 L2; U1 U2]`	Give the finite lower and upper bounds for the support of the density. The first row contains the lower limits and the second row contains the upper limits. This option is only valid for bivariate sample data.

For bivariate data, `'Support'` can be a combination of positive, unbounded, or bounded variables specified as `[0 -Inf; Inf Inf]` or `[0 L; Inf U]`.

Example: `'Support','positive'`

Example: `'Support',[0 10]`

Data Types: `single` | `double` | `char` | `string`

`PlotFcn` — Function used to create kernel density plot
`'surf'` (default) | `'contour'` | `'plot3'` | `'surfc'`

Function used to create kernel density plot, specified as the comma-separated pair consisting of `'PlotFcn'` and one of the following.

Value	Description
`'surf'`	3-D shaded surface plot, created using `surf`
`'contour'`	Contour plot, created using `contour`
`'plot3'`	3-D line plot, created using `plot3`
`'surfc'`	Contour plot under a 3-D shaded surface plot, created using `surfc`

This name-value pair is only valid for bivariate sample data.

Example: `'PlotFcn','contour'`

`Weights` — Weights for sample data
vector

Weights for sample data, specified as the comma-separated pair consisting of `'Weights'` and a vector of length `size(x,1)`, where `x` is the sample data.

Example: `'Weights',xw`

Data Types: `single` | `double`

Output Arguments

`f` — Estimated function values
vector

Estimated function values, returned as a vector whose length is equal to the number of points in `xi` or `pts`.

`xi` — Evaluation points
100 equally spaced points | 900 equally spaced points | vector | two-column matrix

Evaluation points at which `ksdensity` calculates `f`, returned as a vector or a two-column matrix. For univariate data, the default is 100 equally-spaced points that cover the range of data in `x`. For bivariate data, the default is 900 equally-spaced points created using `meshgrid` from 30 equally-spaced points in each dimension.

`bw` — Bandwidth of smoothing window
scalar value

Bandwidth of smoothing window, returned as a scalar value.

If you specify `'BoundaryCorrection'` as `'log'`(default) and `'Support'` as either `'positive'` or a vector `[L U]`, `ksdensity` converts bounded data to be unbounded by using log transformation. The value of `bw` is on the scale of the transformed values.

More About

Kernel Distribution

A kernel distribution is a nonparametric representation of the probability density function (pdf) of a random variable. You can use a kernel distribution when a parametric distribution cannot properly describe the data, or when you want to avoid making assumptions about the distribution of the data.

A kernel distribution is defined by a smoothing function and a bandwidth value, which control the smoothness of the resulting density curve.

The kernel density estimator is the estimated pdf of a random variable. For any real values of x, the kernel density estimator's formula is given by

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where x_1, x_2, ..., x_n are random samples from an unknown distribution, n is the sample size, $K(\cdot)$ is the kernel smoothing function, and h is the bandwidth.

The kernel estimator for the cumulative distribution function (cdf), for any real values of x, is given by

$$\widehat{F}_h(x) = \int_{-\infty}^{x} \widehat{f}_h(t)dt = \frac{1}{n}\sum_{i=1}^{n} G\left(\frac{x - x_i}{h}\right),$$

where $G(x) = \int_{-\infty}^{x} K(t)dt$.

For more details, see “Kernel Distribution” on page B-78.

Reflection Method

The reflection method is a boundary correction method that accurately finds kernel density estimators when a random variable has bounded support. If you specify `'BoundaryCorrection','reflection'`, `ksdensity` uses the reflection method. This method augments bounded data by adding reflected data near the boundaries, and estimates the pdf. Then, `ksdensity` returns the estimated pdf corresponding to the original support with proper normalization, so that the estimated pdf's integral over the original support is equal to one.

If you additionally specify `'Support',[L U]`, then `ksdensity` finds the kernel estimator as follows.

- If `'Function'` is `'pdf'`, then the kernel density estimator is

  $$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x - x_i^-}{h}\right) + K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x - x_i^+}{h}\right)\right] \text{ for } L \le x \le U,$$

  where $x_i^- = 2L - x_i$, $x_i^+ = 2U - x_i$, and x_i is the `i`th sample data.
- If `'Function'` is `'cdf'`, then the kernel estimator for cdf is

  $$\widehat{F}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\left[G\left(\frac{x - x_i^-}{h}\right) + G\left(\frac{x - x_i}{h}\right) + G\left(\frac{x - x_i^+}{h}\right)\right] - \frac{1}{n}\sum_{i=1}^{n}\left[G\left(\frac{L - x_i^-}{h}\right) + G\left(\frac{L - x_i}{h}\right) + G\left(\frac{L - x_i^+}{h}\right)\right]$$

  for $L \le x \le U$.
- To obtain a kernel estimator for an inverse cdf, a survivor function, or a cumulative hazard function (when `'Function'` is `'icdf'`, `'survivor'`, or `'cumhazrd'`), `ksdensity` uses both $\widehat{f}_h(x)$ and $\widehat{F}_h(x)$.

If you additionally specify `'Support'` as `'positive'` or `[0 inf]`, then `ksdensity` finds the kernel estimator by replacing `[L U]` with `[0 inf]` in the above equations.

References

[1] Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press Inc., 1997.

[2] Hill, P. D. “Kernel estimation of a distribution function.” *Communications in Statistics - Theory and Methods*. Vol 14, Issue. 3, 1985, pp. 605-620.

[3] Jones, M. C. “Simple boundary correction for kernel density estimation.” *Statistics and Computing*. Vol. 3, Issue 3, 1993, pp. 135-146.

[4] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- Some options that require extra passes or sorting of the input data are not supported:
  - `'BoundaryCorrection'`
  - `'Censoring'`
  - `'Support'` (support is always unbounded).
- Uses standard deviation (instead of median absolute deviation) to compute the bandwidth.

For more information, see “Tall Arrays for Out-of-Memory Data”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Plotting is not supported.
- Names in name-value pair arguments must be compile-time constants.
- Values in the following name-value pair arguments must also be compile-time constants: `'BoundaryCorrection'`, `'Function'`, and `'Kernel'`. For example, to use the `'Function','cdf'` name-value pair argument in the generated code, include `{coder.Constant('Function'),coder.Constant('cdf')}` in the `-args` value of `codegen`.
- The value of the `'Kernel'` name-value pair argument cannot be a custom function handle. To specify a custom kernel function, use a character vector or string scalar.
- For the value of the `'Support'` name-value pair argument, the compile-time data type must match the runtime data type.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`histogram` | `mvksdensity`

Topics

“Fit Kernel Distribution Using ksdensity” on page 5-39
“Fit Distributions to Grouped Data Using ksdensity” on page 5-41
“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Supported Distributions” on page 5-14

Introduced before R2006a

kstest

One-sample Kolmogorov-Smirnov test

Syntax

```
h = kstest(x)
h = kstest(x,Name,Value)
[h,p] = kstest( ___ )
[h,p,ksstat,cv] = kstest( ___ )
```

Description

`h = kstest(x)` returns a test decision for the null hypothesis that the data in vector `x` comes from a standard normal distribution, against the alternative that it does not come from such a distribution, using the one-sample Kolmogorov-Smirnov test on page 33-3405. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`h = kstest(x,Name,Value)` returns a test decision for the one-sample Kolmogorov-Smirnov test with additional options specified by one or more name-value pair arguments. For example, you can test for a distribution other than standard normal, change the significance level, or conduct a one-sided test.

`[h,p] = kstest( ___ )` also returns the *p*-value `p` of the hypothesis test, using any of the input arguments from the previous syntaxes.

`[h,p,ksstat,cv] = kstest( ___ )` also returns the value of the test statistic `ksstat` and the approximate critical value `cv` of the test.

Examples

Test for Standard Normal Distribution

Perform the one-sample Kolmogorov-Smirnov test by using `kstest`. Confirm the test decision by visually comparing the empirical cumulative distribution function (cdf) to the standard normal cdf.

Load the `examgrades` data set. Create a vector containing the first column of the exam grade data.

```
load examgrades
test1 = grades(:,1);
```

Test the null hypothesis that the data comes from a normal distribution with a mean of 75 and a standard deviation of 10. Use these parameters to center and scale each element of the data vector, because `kstest` tests for a standard normal distribution by default.

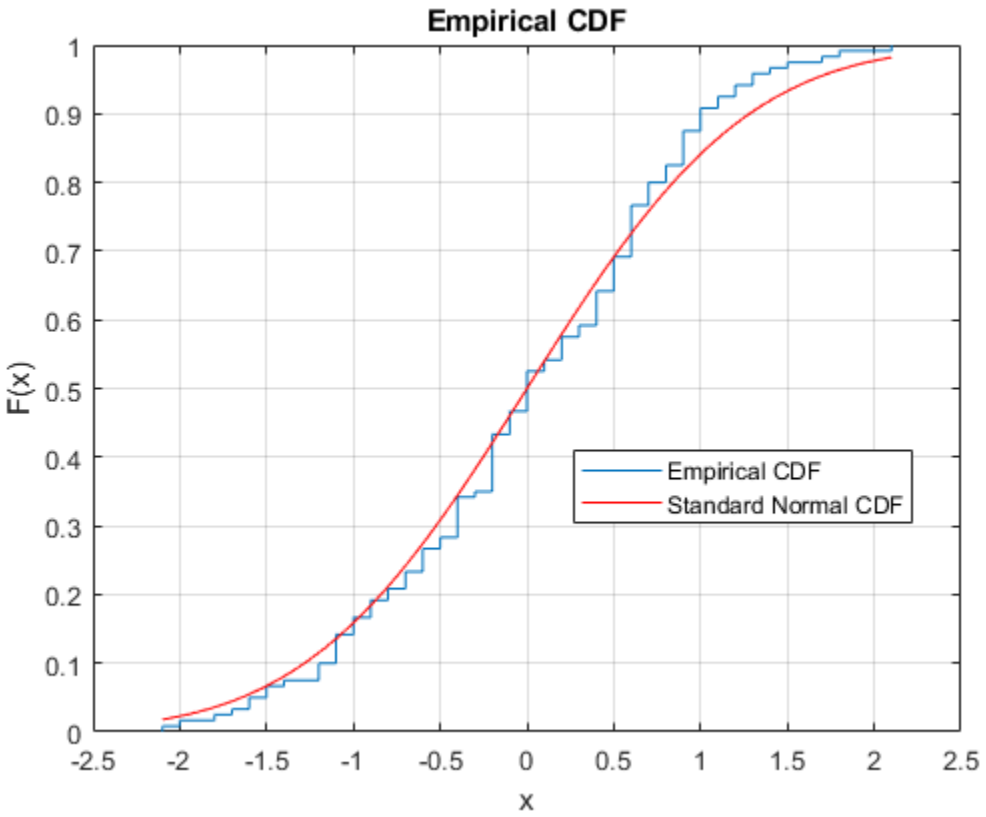
```
x = (test1-75)/10;
h = kstest(x)
```

```
h = logical
   0
```

The returned value of `h = 0` indicates that `kstest` fails to reject the null hypothesis at the default 5% significance level.

Plot the empirical cdf and the standard normal cdf for a visual comparison.

```
cdfplot(x)
hold on
x_values = linspace(min(x),max(x));
plot(x_values,normcdf(x_values,0,1),'r-')
legend('Empirical CDF','Standard Normal CDF','Location','best')
```



The figure shows the similarity between the empirical cdf of the centered and scaled data vector and the cdf of the standard normal distribution.

Specify the Hypothesized Distribution Using a Two-Column Matrix

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades;
x = grades(:,1);
```

Specify the hypothesized distribution as a two-column matrix. Column 1 contains the data vector `x`. Column 2 contains cdf values evaluated at each value in `x` for a hypothesized Student's *t* distribution with a location parameter of 75, a scale parameter of 10, and one degree of freedom.

```
test_cdf = [x,cdf('tlocationscale',x,75,10,1)];
```

Test if the data are from the hypothesized distribution.

```
h = kstest(x,'CDF',test_cdf)
```

```
h = logical
   1
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis at the default 5% significance level.

Specify the Hypothesized Distribution Using a Probability Distribution Object

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades;
x = grades(:,1);
```

Create a probability distribution object to test if the data comes from a Student's *t* distribution with a location parameter of 75, a scale parameter of 10, and one degree of freedom.

```
test_cdf = makedist('tlocationscale','mu',75,'sigma',10,'nu',1);
```

Test the null hypothesis that the data comes from the hypothesized distribution.

```
h = kstest(x,'CDF',test_cdf)
```

```
h = logical
   1
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis at the default 5% significance level.

Test the Hypothesis at Different Significance Levels

Load the sample data. Create a vector containing the first column of the students' exam grades.

```
load examgrades;
x = grades(:,1);
```

Create a probability distribution object to test if the data comes from a Student's *t* distribution with a location parameter of 75, a scale parameter of 10, and one degree of freedom.

```
test_cdf = makedist('tlocationscale','mu',75,'sigma',10,'nu',1);
```

Test the null hypothesis that data comes from the hypothesized distribution at the 1% significance level.

```
[h,p] = kstest(x,'CDF',test_cdf,'Alpha',0.01)
```

```
h = logical
   1


p = 0.0021
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis at the 1% significance level.

Conduct a One-Sided Hypothesis Test

Load the sample data. Create a vector containing the third column of the stock return data matrix.

```
load stockreturns;
x = stocks(:,3);
```

Test the null hypothesis that the data comes from a standard normal distribution, against the alternative hypothesis that the population cdf of the data is larger than the standard normal cdf.

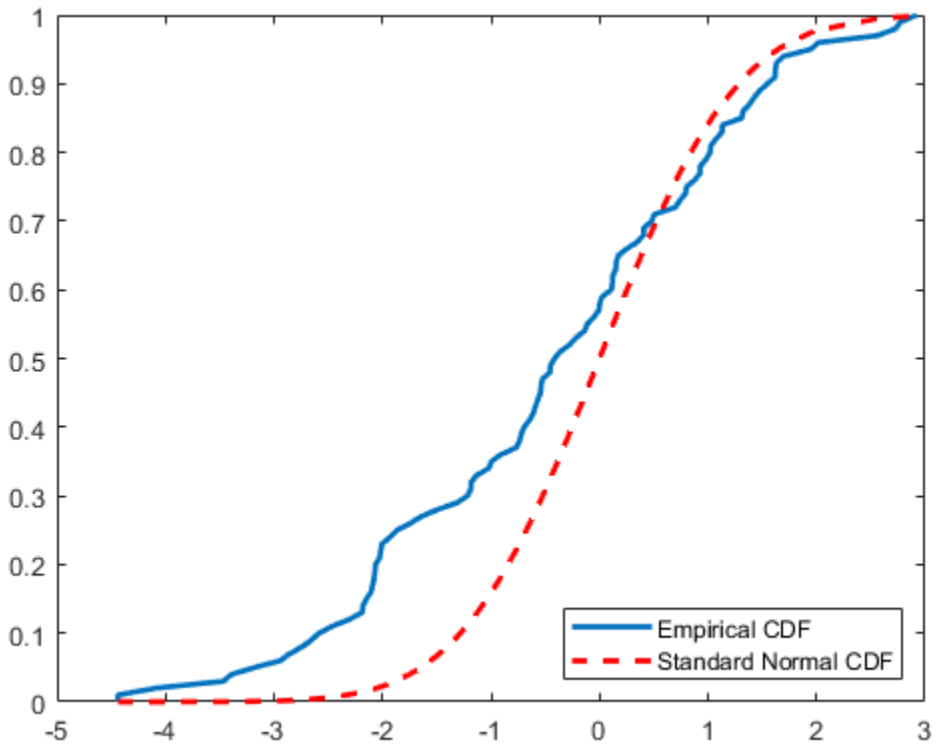
```
[h,p,k,c] = kstest(x,'Tail','larger')
```

```
h = logical
   1


p = 5.0854e-05

k = 0.2197

c = 0.1207
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis in favor of the alternative hypothesis at the default 5% significance level.

Plot the empirical cdf and the standard normal cdf for a visual comparison.

```
[f,x_values] = ecdf(x);
J = plot(x_values,f);
hold on;
K = plot(x_values,normcdf(x_values),'r--');
set(J,'LineWidth',2);
set(K,'LineWidth',2);
legend([J K],'Empirical CDF','Standard Normal CDF','Location','SE');
```



The plot shows the difference between the empirical cdf of the data vector `x` and the cdf of the standard normal distribution.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','larger','Alpha',0.01` specifies a test using the alternative hypothesis that the cdf of the population from which the sample data is drawn is greater than the cdf of the hypothesized distribution, conducted at the 1% significance level.

Alpha — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`CDF` — cdf of hypothesized continuous distribution
matrix | probability distribution object

cdf of hypothesized continuous distribution, specified the comma-separated pair consisting of `'CDF'` and either a two-column matrix or a continuous probability distribution object. When `CDF` is a matrix, column 1 contains a set of possible *x* values, and column 2 contains the corresponding hypothesized cumulative distribution function values $G(x)$. The calculation is most efficient if `CDF` is specified such that column 1 contains the values in the data vector `x`. If there are values in `x` not found in column 1 of `CDF`, `kstest` approximates $G(x)$ by interpolation. All values in `x` must lie in the interval between the smallest and largest values in the first column of `CDF`. By default, `kstest` tests for a standard normal distribution.

The one-sample Kolmogorov-Smirnov test on page 33-3405 is only valid for continuous cumulative distribution functions, and requires `CDF` to be predetermined. The result is not accurate if `CDF` is estimated from the data. To test `x` against the normal, lognormal, extreme value, Weibull, or exponential distribution without specifying distribution parameters, use `lillietest` instead.

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis
`'unequal'` (default) | `'larger'` | `'smaller'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'unequal'`	Test the alternative hypothesis that the cdf of the population from which `x` is drawn is not equal to the cdf of the hypothesized distribution.
`'larger'`	Test the alternative hypothesis that the cdf of the population from which `x` is drawn is greater than the cdf of the hypothesized distribution.
`'smaller'`	Test the alternative hypothesis that the cdf of the population from which `x` is drawn is less than the cdf of the hypothesized distribution.

If the values in the data vector `x` tend to be larger than expected from the hypothesized distribution, the empirical distribution function of `x` tends to be smaller, and vice versa.

Example: `'Tail','larger'`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as a logical value.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

p — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

ksstat — Test statistic
nonnegative scalar value

Test statistic of the hypothesis test, returned as a nonnegative scalar value.

cv — Critical value
nonnegative scalar value

Critical value, returned as a nonnegative scalar value.

More About

One-Sample Kolmogorov-Smirnov Test

The one-sample Kolmogorov-Smirnov test is a nonparametric test of the null hypothesis that the population cdf of the data is equal to the hypothesized cdf.

The two-sided test for "unequal" cdf functions tests the null hypothesis against the alternative that the population cdf of the data is not equal to the hypothesized cdf. The test statistic is the maximum absolute difference between the empirical cdf calculated from x and the hypothesized cdf:

$$D^* = \max_x\left(\left|\widehat{F}(x) - G(x)\right|\right),$$

where $\widehat{F}(x)$ is the empirical cdf and $G(x)$ is the cdf of the hypothesized distribution.

The one-sided test for a "larger" cdf function tests the null hypothesis against the alternative that the population cdf of the data is greater than the hypothesized cdf. The test statistic is the maximum amount by which the empirical cdf calculated from x exceeds the hypothesized cdf:

$$D^* = \max_x\left(\widehat{F}(x) - G(x)\right).$$

The one-sided test for a "smaller" cdf function tests the null hypothesis against the alternative that the population cdf of the data is less than the hypothesized cdf. The test statistic is the maximum amount by which the hypothesized cdf exceeds the empirical cdf calculated from x:

$$D^* = \max_x\left(G(x) - \widehat{F}(x)\right).$$

`kstest` computes the critical value `cv` using an approximate formula or by interpolation in a table. The formula and table cover the range `0.01 ≤ alpha ≤ 0.2` for two-sided tests and `0.005 ≤ alpha ≤ 0.1` for one-sided tests. `cv` is returned as `NaN` if `alpha` is outside this range.

Algorithms

`kstest` decides to reject the null hypothesis by comparing the *p*-value `p` with the significance level `Alpha`, not by comparing the test statistic `ksstat` with the critical value `cv`. Since `cv` is

approximate, comparing `ksstat` with `cv` occasionally leads to a different conclusion than comparing `p` with `Alpha`.

References

[1] Massey, F. J. “The Kolmogorov-Smirnov Test for Goodness of Fit.” *Journal of the American Statistical Association*. Vol. 46, No. 253, 1951, pp. 68–78.

[2] Miller, L. H. “Table of Percentage Points of Kolmogorov Statistics.” *Journal of the American Statistical Association*. Vol. 51, No. 273, 1956, pp. 111–121.

[3] Marsaglia, G., W. Tsang, and J. Wang. “Evaluating Kolmogorov’s Distribution.” *Journal of Statistical Software*. Vol. 8, Issue 18, 2003.

See Also

`adtest` | `kstest2` | `lillietest`

Introduced before R2006a

kstest2

Two-sample Kolmogorov-Smirnov test

Syntax

```
h = kstest2(x1,x2)
h = kstest2(x1,x2,Name,Value)
[h,p] = kstest2(___)
[h,p,ks2stat] = kstest2(___)
```

Description

`h = kstest2(x1,x2)` returns a test decision for the null hypothesis that the data in vectors `x1` and `x2` are from the same continuous distribution, using the two-sample Kolmogorov-Smirnov test on page 33-3410. The alternative hypothesis is that `x1` and `x2` are from different continuous distributions. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = kstest2(x1,x2,Name,Value)` returns a test decision for a two-sample Kolmogorov-Smirnov test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = kstest2(___)` also returns the asymptotic *p*-value `p`, using any of the input arguments from the previous syntaxes.

`[h,p,ks2stat] = kstest2(___)` also returns the test statistic `ks2stat`.

Examples

Test Two Samples for the Same Distribution

Generate sample data from two different Weibull distributions.

```
rng(1);     % For reproducibility
x1 = wblrnd(1,1,1,50);
x2 = wblrnd(1.2,2,1,50);
```

Test the null hypothesis that data in vectors `x1` and `x2` comes from populations with the same distribution.

```
h = kstest2(x1,x2)
```

```
h = logical
   1
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis at the default 5% significance level.

Test the Hypothesis at Different Significance Levels

Generate sample data from two different Weibull distributions.

```
rng(1);     % For reproducibility
x1 = wblrnd(1,1,1,50);
x2 = wblrnd(1.2,2,1,50);
```

Test the null hypothesis that data vectors `x1` and `x2` are from populations with the same distribution at the 1% significance level.

```
[h,p] = kstest2(x1,x2,'Alpha',0.01)
```

```
h = logical
   0


p = 0.0317
```

The returned value of `h = 0` indicates that `kstest` does not reject the null hypothesis at the 1% significance level.

One-Sided Hypothesis Test

Generate sample data from two different Weibull distributions.

```
rng(1);     % For reproducibility
x1 = wblrnd(1,1,1,50);
x2 = wblrnd(1.2,2,1,50);
```

Test the null hypothesis that data in vectors `x1` and `x2` comes from populations with the same distribution, against the alternative hypothesis that the cdf of the distribution of `x1` is larger than the cdf of the distribution of `x2`.

```
[h,p,k] = kstest2(x1,x2,'Tail','larger')
```

```
h = logical
   1


p = 0.0158

k = 0.2800
```

The returned value of `h = 1` indicates that `kstest` rejects the null hypothesis, in favor of the alternative hypothesis that the cdf of the distribution of `x1` is larger than the cdf of the distribution of `x2`, at the default 5% significance level. The returned value of `k` is the test statistic for the two-sample Kolmogorov-Smirnov test.

Input Arguments

x1 — Sample data

vector

Sample data from the first sample, specified as a vector. Data vectors `x1` and `x2` do not need to be the same size.

Data Types: `single` | `double`

x2 — Sample data

vector

Sample data from the second sample, specified as a vector. Data vectors `x1` and `x2` do not need to be the same size.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','larger','Alpha',0.01` specifies a test using the alternative hypothesis that the empirical cdf of `x1` is larger than the empirical cdf of `x2`, conducted at the 1% significance level.

Alpha — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

Tail — Type of alternative hypothesis

`'unequal'` (default) | `'larger'` | `'smaller'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'unequal'`	Test the alternative hypothesis that the empirical cdf of `x1` is unequal to the empirical cdf of `x2`.
`'larger'`	Test the alternative hypothesis that the empirical cdf of `x1` is larger than the empirical cdf of `x2`.
`'smaller'`	Test the alternative hypothesis that the empirical cdf of `x1` is smaller than the empirical cdf of `x2`.

If the data values in `x1` tend to be larger than those in `x2`, the empirical distribution function of `x1` tends to be smaller than that of `x2`, and vice versa.

Example: `'Tail','larger'`

Output Arguments

h — Hypothesis test result

`1` | `0`

Hypothesis test result, returned as a logical value.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — Asymptotic *p*-value
scalar value in the range (0,1)

Asymptotic *p*-value of the test, returned as a scalar value in the range (0,1). `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. The asymptotic *p*-value becomes very accurate for large sample sizes, and is believed to be reasonably accurate for sample sizes `n1` and `n2`, such that `(n1*n2)/(n1 + n2)` ≥ 4.

`ks2stat` — Test statistic
nonnegative scalar value

Test statistic, returned as a nonnegative scalar value.

More About

Two-Sample Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov test is a nonparametric hypothesis test that evaluates the difference between the cdfs of the distributions of the two sample data vectors over the range of *x* in each data set.

The two-sided test uses the maximum absolute difference between the cdfs of the distributions of the two data vectors. The test statistic is

$$D^* = \max_x\left(\left|\widehat{F}_1(x) - \widehat{F}_2(x)\right|\right),$$

where $\widehat{F}_1(x)$ is the proportion of `x1` values less than or equal to *x* and $\widehat{F}_2(x)$ is the proportion of `x2` values less than or equal to *x*.

The one-sided test uses the actual value of the difference between the cdfs of the distributions of the two data vectors rather than the absolute value. The test statistic is

$$D^* = \max_x\left(\widehat{F}_1(x) - \widehat{F}_2(x)\right).$$

Algorithms

In `kstest2`, the decision to reject the null hypothesis is based on comparing the *p*-value `p` with the significance level `Alpha`, not by comparing the test statistic `ks2stat` with a critical value.

References

[1] Massey, F. J. "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*. Vol. 46, No. 253, 1951, pp. 68–78.

[2] Miller, L. H. "Table of Percentage Points of Kolmogorov Statistics." *Journal of the American Statistical Association*. Vol. 51, No. 273, 1956, pp. 111–121.

[3] Marsaglia, G., W. Tsang, and J. Wang. “Evaluating Kolmogorov’s Distribution.” *Journal of Statistical Software*. Vol. 8, Issue 18, 2003.

See Also

`adtest` | `kstest` | `lillietest`

Introduced before R2006a

kurtosis

Kurtosis

Syntax

```
k = kurtosis(X)
k = kurtosis(X,flag)
k = kurtosis(X,flag,'all')
k = kurtosis(X,flag,dim)
k = kurtosis(X,flag,vecdim)
```

Description

`k = kurtosis(X)` returns the sample kurtosis of `X`.

- If `X` is a vector, then `kurtosis(X)` returns a scalar value that is the kurtosis of the elements in `X`.
- If `X` is a matrix, then `kurtosis(X)` returns a row vector that contains the sample kurtosis of each column in `X`.
- If `X` is a multidimensional array, then `kurtosis(X)` operates along the first nonsingleton dimension of `X`.

`k = kurtosis(X,flag)` specifies whether to correct for bias (`flag` is `0`) or not (`flag` is `1`, the default). When `X` represents a sample from a population, the kurtosis of `X` is biased, meaning it tends to differ from the population kurtosis by a systematic amount based on the sample size. You can set `flag` to `0` to correct for this systematic bias.

`k = kurtosis(X,flag,'all')` returns the kurtosis of all elements of `X`.

`k = kurtosis(X,flag,dim)` returns the kurtosis along the operating dimension `dim` of `X`.

`k = kurtosis(X,flag,vecdim)` returns the kurtosis over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `kurtosis(X,1,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the biased kurtosis of the elements on the corresponding page of `X`.

Examples

Find Kurtosis of Matrix

Set the random seed for reproducibility of the results.

```
rng('default')
```

Generate a matrix with 5 rows and 4 columns.

```
X = randn(5,4)
```

```
X = 5×4
```

```
    0.5377   -1.3077   -1.3499   -0.2050
    1.8339   -0.4336    3.0349   -0.1241
   -2.2588    0.3426    0.7254    1.4897
    0.8622    3.5784   -0.0631    1.4090
    0.3188    2.7694    0.7147    1.4172
```

Find the sample kurtosis of `X`.

```
k = kurtosis(X)
```

```
k = 1×4

    2.7067    1.4069    2.3783    1.1759
```

`k` is a row vector containing the sample kurtosis of each column in `X`.

Correct for Bias in Sample Kurtosis

For an input vector, correct for bias in the calculation of kurtosis by specifying the `flag` input argument.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Generate a vector of length 10.

```
x = randn(10,1)
```

```
x = 10×1

    0.5377
    1.8339
   -2.2588
    0.8622
    0.3188
   -1.3077
   -0.4336
    0.3426
    3.5784
    2.7694
```

Find the biased kurtosis of `x`. By default, `kurtosis` sets the value of `flag` to `1` for computing the biased kurtosis.

```
k1 = kurtosis(x) % flag is 1 by default
```

```
k1 = 2.3121
```

Find the bias-corrected kurtosis of `x` by setting the value of `flag` to `0`.

```
k2 = kurtosis(x,0)
```

```
k2 = 2.7483
```

Find Kurtosis Along Given Dimension

Find the kurtosis along different dimensions for a multidimensional array.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4,3,2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the kurtosis of `X` along the default dimension.

```
k1 = kurtosis(X)
```

```
k1 =
k1(:,:,1) =

    2.1350    1.7060    2.2789


k1(:,:,2) =

    1.0542    2.3278    2.0996
```

By default, `kurtosis` operates along the first dimension of `X` whose size does not equal 1. In this case, this dimension is the first dimension of `X`. Therefore, `k1` is a 1-by-3-by-2 array.

Find the biased kurtosis of `X` along the second dimension.

```
k2 = kurtosis(X,1,2)
```

```
k2 =
k2(:,:,1) =

    1.5000
    1.5000
```

```
    1.5000
    1.5000


k2(:,:,2) =

    1.5000
    1.5000
    1.5000
    1.5000
```

`k2` is a 4-by-1-by-2 array.

Find the biased kurtosis of `X` along the third dimension.

```
k3 = kurtosis(X,1,3)
```

```
k3 = 4×3

    1.0000    1.0000    1.0000
    1.0000    1.0000    1.0000
    1.0000    1.0000    1.0000
    1.0000    1.0000    1.0000
```

`k3` is a 4-by-3 matrix.

Find Kurtosis Along Vector of Dimensions

Find the kurtosis over multiple dimensions by using the `'all'` and `vecdim` input arguments.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4 3 2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the biased kurtosis of X.

```
kall = kurtosis(X,1,'all')
```

```
kall = 2.8029
```

`kall` is the biased kurtosis of the entire input data set X.

Find the biased kurtosis of each page of X by specifying the first and second dimensions.

```
kpage = kurtosis(X,1,[1 2])
```

```
kpage =
kpage(:,:,1) =

    1.9345


kpage(:,:,2) =

    2.5877
```

For example, `kpage(1,1,2)` is the biased kurtosis of the elements in `X(:,:,2)`.

Find the biased kurtosis of the elements in each `X(i,:,:)` slice by specifying the second and third dimensions.

```
krow = kurtosis(X,1,[2 3])
```

```
krow = 4×1

    3.8457
    1.4306
    1.7094
    2.3378
```

For example, `krow(3)` is the biased kurtosis of the elements in `X(3,:,:)`.

Input Arguments

X — Input data

vector | matrix | multidimensional array

Input data that represents a sample from a population, specified as a vector, matrix, or multidimensional array.

- If X is a vector, then `kurtosis(X)` returns a scalar value that is the kurtosis of the elements in X.
- If X is a matrix, then `kurtosis(X)` returns a row vector that contains the sample kurtosis of each column in X.
- If X is a multidimensional array, then `kurtosis(X)` operates along the first nonsingleton dimension of X.

To specify the operating dimension when X is a matrix or an array, use the `dim` input argument.

`kurtosis` treats `NaN` values in `X` as missing values and removes them.

Data Types: `single` | `double`

`flag` — Indicator for bias

1 (default) | 0

Indicator for the bias, specified as `0` or `1`.

- If `flag` is `1` (default), then the kurtosis of `X` is biased, meaning it tends to differ from the population kurtosis by a systematic amount based on the sample size.
- If `flag` is `0`, then `kurtosis` corrects for the systematic bias.

Data Types: `single` | `double` | `logical`

`dim` — Dimension

positive integer

Dimension along which to operate, specified as a positive integer. If you do not specify a value for `dim`, then the default is the first dimension of `X` whose size does not equal 1.

Consider the kurtosis of a matrix `X`:

- If `dim` is equal to 1, then `kurtosis` returns a row vector that contains the sample kurtosis of each column in `X`.
- If `dim` is equal to 2, then `kurtosis` returns a column vector that contains the sample kurtosis of each row in `X`.

If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `kurtosis` returns an array of `NaN`s the same size as `X`.

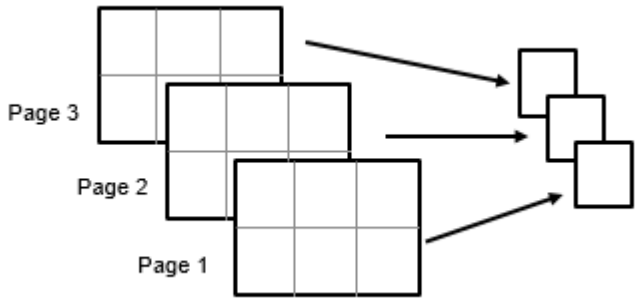
Data Types: `single` | `double`

`vecdim` — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `k` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `k`.

For example, if `X` is a 2-by-3-by-3 array, then `kurtosis(X,1,[1 2])` returns a 1-by-1-by-3 array. Each element of the output is the biased kurtosis of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

k — Kurtosis

scalar | vector | matrix | multidimensional array

Kurtosis, returned as a scalar, vector, matrix, or multidimensional array.

Algorithms

Kurtosis is a measure of how outlier-prone a distribution is. The kurtosis of the normal distribution on page B-119 is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3. Some definitions of kurtosis subtract 3 from the computed value, so that the normal distribution has kurtosis of 0. The `kurtosis` function does not use this convention.

The kurtosis of a distribution is defined as

$$k = \frac{E(x-\mu)^4}{\sigma^4},$$

where μ is the mean of x, σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t. The `kurtosis` function computes a sample version of this population value.

When you set `flag` to `1`, the kurtosis is biased, and the following equation applies:

$$k_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}.$$

When you set `flag` to `0`, `kurtosis` corrects for the systematic bias, and the following equation applies:

$$k_0 = \frac{n-1}{(n-2)(n-3)}((n+1)k_1 - 3(n-1)) + 3.$$

This bias-corrected equation requires that `X` contain at least four elements.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`mean` | `moment` | `skewness` | `std` | `var`

Topics

"Normal Distribution" on page B-119

Introduced before R2006a

lasso

Lasso or elastic net regularization for linear models

Syntax

```
B = lasso(X,y)
B = lasso(X,y,Name,Value)
[B,FitInfo] = lasso(___)
```

Description

`B = lasso(X,y)` returns fitted least-squares regression coefficients for linear models of the predictor data `X` and the response `y`. Each column of `B` corresponds to a particular regularization coefficient in `Lambda`. By default, `lasso` performs lasso regularization using a geometric sequence of `Lambda` values.

`B = lasso(X,y,Name,Value)` fits regularized regressions with additional options specified by one or more name-value pair arguments. For example, `'Alpha',0.5` sets elastic net as the regularization method, with the parameter `Alpha` equal to 0.5.

`[B,FitInfo] = lasso(___)` also returns the structure `FitInfo`, which contains information about the fit of the models, using any of the input arguments in the previous syntaxes.

Examples

Remove Redundant Predictors Using Lasso Regularization

Construct a data set with redundant predictors and identify those predictors by using `lasso`.

Create a matrix `X` of 100 five-dimensional normal variables. Create a response vector `y` from just two components of `X`, and add a small amount of noise.

```
rng default % For reproducibility
X = randn(100,5);
weights = [0;2;0;-3;0]; % Only two nonzero coefficients
y = X*weights + randn(100,1)*0.1; % Small added noise
```

Construct the default lasso fit.

```
B = lasso(X,y);
```

Find the coefficient vector for the 25th `Lambda` value in `B`.

```
B(:,25)
```

```
ans = 5×1

         0
    1.6093
         0
```

```
   -2.5865
         0
```

`lasso` identifies and removes the redundant predictors.

Create Linear Model Without Intercept Term Using Lasso Regularization

Create sample data with predictor variable X and response variable $y = 0 + 2X + \varepsilon$.

```
rng('default') % For reproducibility
X = rand(100,1);
y = 2*X + randn(100,1)/10;
```

Specify a regularization value, and find the coefficient of the regression model without an intercept term.

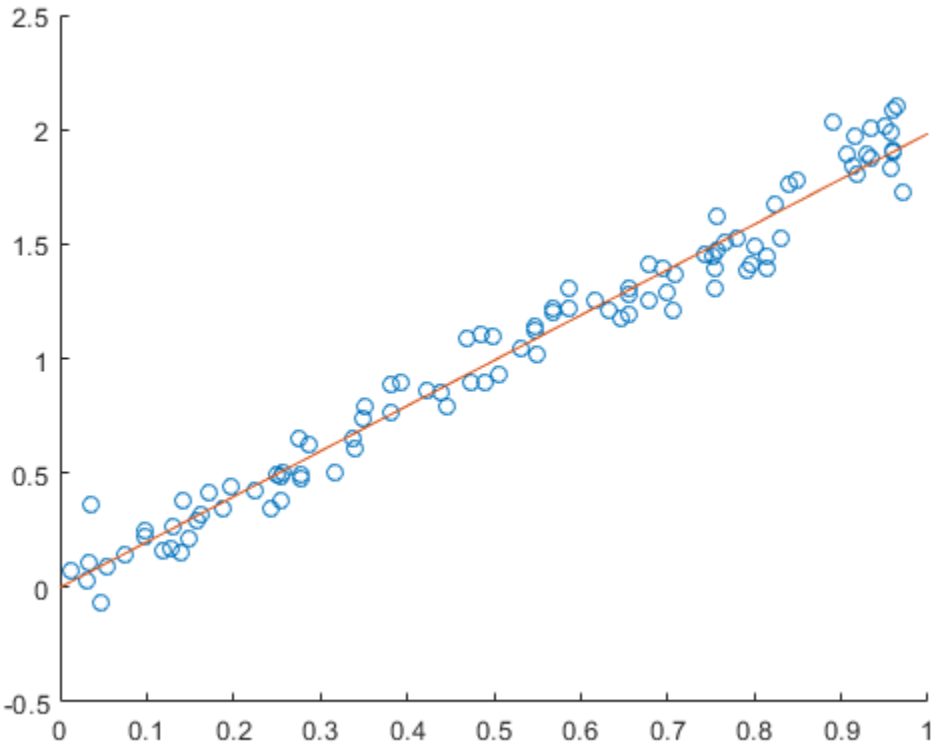
```
lambda = 1e-03;
B = lasso(X,y,'Lambda',lambda,'Intercept',false)
```

```
Warning: When the 'Intercept' value is false, the 'Standardize' value is set to false.
```

```
B = 1.9825
```

Plot the real values (points) against the predicted values (line).

```
scatter(X,y)
hold on
x = 0:0.1:1;
plot(x,x*B)
hold off
```



Remove Redundant Predictors by Using Cross-Validated Fits

Construct a data set with redundant predictors and identify those predictors by using cross-validated `lasso`.

Create a matrix `X` of 100 five-dimensional normal variables. Create a response vector `y` from two components of `X`, and add a small amount of noise.

```
rng default % For reproducibility
X = randn(100,5);
weights = [0;2;0;-3;0]; % Only two nonzero coefficients
y = X*weights + randn(100,1)*0.1; % Small added noise
```

Construct the lasso fit by using 10-fold cross-validation with labeled predictor variables.

```
[B,FitInfo] = lasso(X,y,'CV',10,'PredictorNames',{'x1','x2','x3','x4','x5'});
```

Display the variables in the model that corresponds to the minimum cross-validated mean squared error (MSE).

```
idxLambdaMinMSE = FitInfo.IndexMinMSE;
minMSEModelPredictors = FitInfo.PredictorNames(B(:,idxLambdaMinMSE)~=0)
```

```
minMSEModelPredictors = 1x2 cell
    {'x2'}    {'x4'}
```

Display the variables in the sparsest model within one standard error of the minimum MSE.

```
idxLambda1SE = FitInfo.Index1SE;
sparseModelPredictors = FitInfo.PredictorNames(B(:,idxLambda1SE)~=0)
```

```
sparseModelPredictors = 1x2 cell
    {'x2'}    {'x4'}
```

In this example, `lasso` identifies the same predictors for the two models and removes the redundant predictors.

Lasso Plot with Cross-Validated Fits

Visually examine the cross-validated error of various levels of regularization.

Load the sample data.

```
load acetylene
```

Create a design matrix with interactions and no constant term.

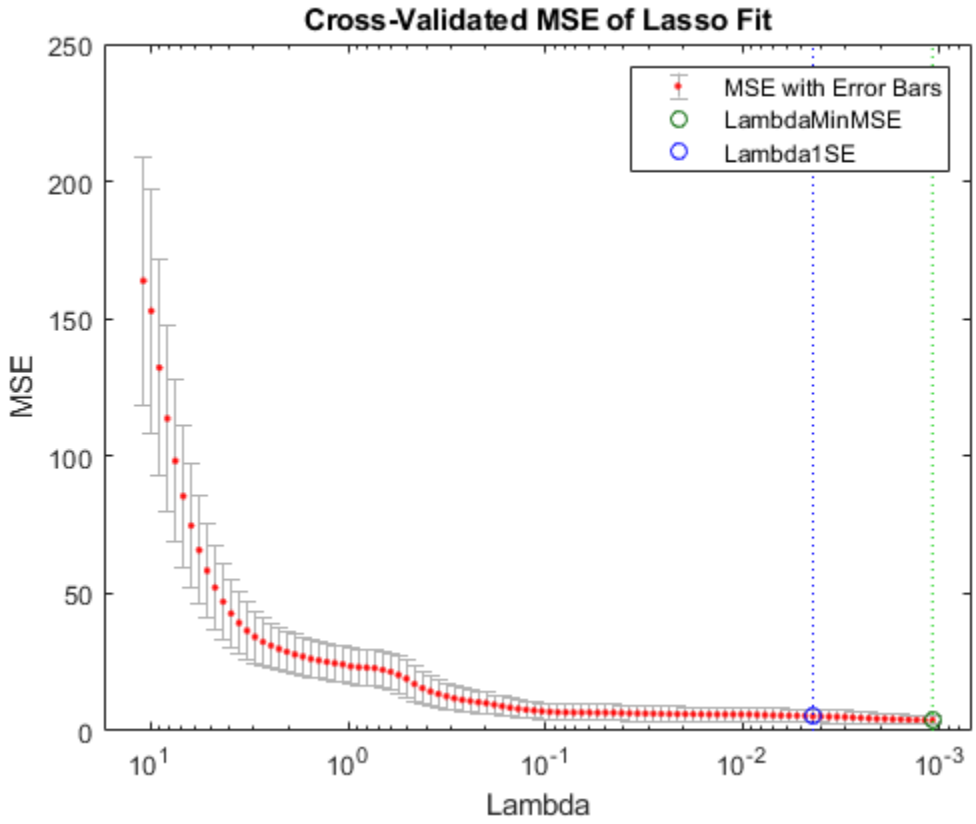
```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
```

Construct the lasso fit using 10-fold cross-validation. Include the `FitInfo` output so you can plot the result.

```
rng default % For reproducibility
[B,FitInfo] = lasso(D,y,'CV',10);
```

Plot the cross-validated fits.

```
lassoPlot(B,FitInfo,'PlotType','CV');
legend('show') % Show legend
```



The green circle and dotted line locate the `Lambda` with minimum cross-validation error. The blue circle and dotted line locate the point with minimum cross-validation error plus one standard deviation.

Predict Values Using Elastic Net Regularization

Predict students' exam scores using `lasso` and the elastic net method.

Load the `examgrades` data set.

```
load examgrades
X = grades(:,1:4);
y = grades(:,5);
```

Split the data into training and test sets.

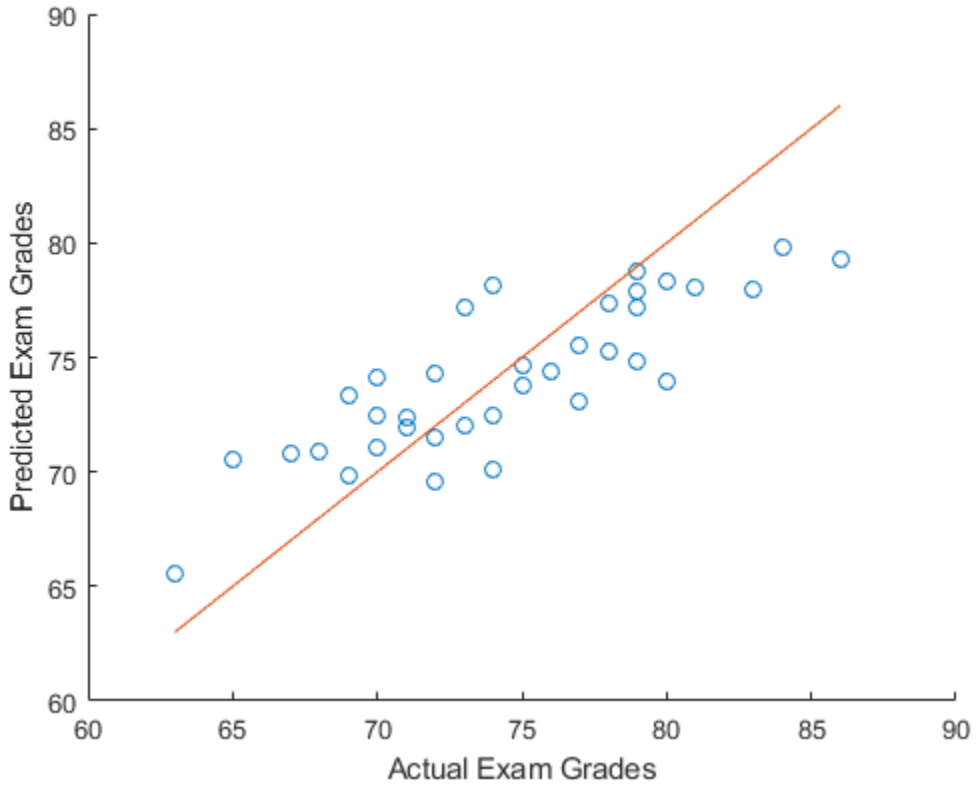
```
n = length(y);
c = cvpartition(n,'HoldOut',0.3);
idxTrain = training(c,1);
idxTest = ~idxTrain;
XTrain = X(idxTrain,:);
yTrain = y(idxTrain);
XTest = X(idxTest,:);
yTest = y(idxTest);
```

Find the coefficients of a regularized linear regression model using 10-fold cross-validation and the elastic net method with `Alpha` = 0.75. Use the largest `Lambda` value such that the mean squared error (MSE) is within one standard error of the minimum MSE.

```
[B,FitInfo] = lasso(XTrain,yTrain,'Alpha',0.75,'CV',10);
idxLambda1SE = FitInfo.Index1SE;
coef = B(:,idxLambda1SE);
coef0 = FitInfo.Intercept(idxLambda1SE);
```

Predict exam scores for the test data. Compare the predicted values to the actual exam grades using a reference line.

```
yhat = XTest*coef + coef0;
hold on
scatter(yTest,yhat)
plot(yTest,yTest)
xlabel('Actual Exam Grades')
ylabel('Predicted Exam Grades')
hold off
```



Input Arguments

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row represents one observation, and each column represents one predictor variable.

Data Types: `single` | `double`

y — Response data
numeric vector

Response data, specified as a numeric vector. `y` has length *n*, where *n* is the number of rows of `X`. The response `y(i)` corresponds to the *i*th row of `X`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `lasso(X,y,'Alpha',0.75,'CV',10)` performs elastic net regularization with 10-fold cross-validation. The `'Alpha',0.75` name-value pair argument sets the parameter used in the elastic net optimization.

`AbsTol` — Absolute error tolerance
`1e−4` (default) | positive scalar

Absolute error tolerance used to determine the convergence of the "ADMM Algorithm" on page 33-3432, specified as the comma-separated pair consisting of `'AbsTol'` and a positive scalar. The algorithm converges when successive estimates of the coefficient vector differ by an amount less than `AbsTol`.

Note This option applies only when you use `lasso` on tall arrays. See "Extended Capabilities" on page 33-0 for more information.

Example: `'AbsTol',1e−3`

Data Types: `single` | `double`

`Alpha` — Weight of lasso versus ridge optimization
`1` (default) | positive scalar

Weight of lasso (L^1) versus ridge (L^2) optimization, specified as the comma-separated pair consisting of `'Alpha'` and a positive scalar value in the interval `(0,1]`. The value `Alpha = 1` represents lasso regression, `Alpha` close to `0` approaches ridge regression on page 11-109, and other values represent elastic net optimization. See "Elastic Net" on page 33-3431.

Example: `'Alpha',0.5`

Data Types: `single` | `double`

`B0` — Initial values for *x*-coefficients in ADMM Algorithm
vector of zeros (default) | numeric vector

Initial values for *x*-coefficients in "ADMM Algorithm" on page 33-3432, specified as the comma-separated pair consisting of `'B0'` and a numeric vector.

Note This option applies only when you use `lasso` on tall arrays. See "Extended Capabilities" on page 33-0 for more information.

Data Types: `single` | `double`

`CV` — Cross-validation specification for estimating mean squared error

`'resubstitution'` (default) | positive integer scalar | `cvpartition` object

Cross-validation specification for estimating the mean squared error (MSE), specified as the comma-separated pair consisting of `'CV'` and one of the following:

- `'resubstitution'` — `lasso` uses `X` and `y` to fit the model and to estimate the MSE without cross-validation.
- Positive scalar integer `K` — `lasso` uses `K`-fold cross-validation.
- `cvpartition` object `cvp` — `lasso` uses the cross-validation method expressed in `cvp`. You cannot use a `'leaveout'` partition with `lasso`.

Example: `'CV',3`

`DFmax` — Maximum number of nonzero coefficients

`Inf` (default) | positive integer scalar

Maximum number of nonzero coefficients in the model, specified as the comma-separated pair consisting of `'DFmax'` and a positive integer scalar. `lasso` returns results only for `Lambda` values that satisfy this criterion.

Example: `'DFmax',5`

Data Types: `single` | `double`

`Intercept` — Flag for fitting the model with intercept term

`true` (default) | `false`

Flag for fitting the model with the intercept term, specified as the comma-separated pair consisting of `'Intercept'` and either `true` or `false`. The default value is `true`, which indicates to include the intercept term in the model. If `Intercept` is `false`, then the returned intercept value is 0.

Example: `'Intercept',false`

Data Types: `logical`

`Lambda` — Regularization coefficients

nonnegative vector

Regularization coefficients, specified as the comma-separated pair consisting of `'Lambda'` and a vector of nonnegative values. See "Lasso" on page 33-3431.

- If you do not supply `Lambda`, then `lasso` calculates the largest value of `Lambda` that gives a nonnull model. In this case, `LambdaRatio` gives the ratio of the smallest to the largest value of the sequence, and `NumLambda` gives the length of the vector.
- If you supply `Lambda`, then `lasso` ignores `LambdaRatio` and `NumLambda`.
- If `Standardize` is `true`, then `Lambda` is the set of values used to fit the models with the `X` data standardized to have zero mean and a variance of one.

The default is a geometric sequence of `NumLambda` values, with only the largest value able to produce `B = 0`.

Example: `'Lambda',linspace(0,1)`

Data Types: `single` | `double`

`LambdaRatio` — Ratio of smallest to largest `Lambda` values
`1e–4` (default) | positive scalar

Ratio of the smallest to the largest `Lambda` values when you do not supply `Lambda`, specified as the comma-separated pair consisting of `'LambdaRatio'` and a positive scalar.

If you set `LambdaRatio = 0`, then `lasso` generates a default sequence of `Lambda` values and replaces the smallest one with `0`.

Example: `'LambdaRatio',1e–2`

Data Types: `single` | `double`

`MaxIter` — Maximum number of iterations allowed
positive integer scalar

Maximum number of iterations allowed, specified as the comma-separated pair consisting of `'MaxIter'` and a positive integer scalar.

If the algorithm executes `MaxIter` iterations before reaching the convergence tolerance `RelTol`, then the function stops iterating and returns a warning message.

The function can return more than one warning when `NumLambda` is greater than `1`.

Default values are `1e5` for standard data and `1e4` for tall arrays.

Example: `'MaxIter',1e3`

Data Types: `single` | `double`

`MCReps` — Number of Monte Carlo repetitions for cross-validation
`1` (default) | positive integer scalar

Number of Monte Carlo repetitions for cross-validation, specified as the comma-separated pair consisting of `'MCReps'` and a positive integer scalar.

- If `CV` is `'resubstitution'` or a `cvpartition` of type `'resubstitution'`, then `MCReps` must be `1`.
- If `CV` is a `cvpartition` of type `'holdout'`, then `MCReps` must be greater than `1`.

Example: `'MCReps',5`

Data Types: `single` | `double`

`NumLambda` — Number of `Lambda` values
`100` (default) | positive integer scalar

Number of `Lambda` values `lasso` uses when you do not supply `Lambda`, specified as the comma-separated pair consisting of `'NumLambda'` and a positive integer scalar. `lasso` can return fewer than `NumLambda` fits if the residual error of the fits drops below a threshold fraction of the variance of `y`.

Example: `'NumLambda',50`

Data Types: `single` | `double`

`Options` — Option to cross-validate in parallel and specify random streams
structure

Option to cross-validate in parallel and specify the random streams, specified as the comma-separated pair consisting of `'Options'` and a structure. This option requires Parallel Computing Toolbox.

Create the `Options` structure with `statset`. The option fields are:

- `UseParallel` — Set to `true` to compute in parallel. The default is `false`.
- `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. For reproducibility, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`. The default is `false`.
- `Streams` — A `RandStream` object or cell array consisting of one such object. If you do not specify `Streams`, then `lasso` uses the default stream.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

`PredictorNames` — Names of predictor variables
`{}` (default) | string array | cell array of character vectors

Names of the predictor variables, in the order in which they appear in `X`, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array or cell array of character vectors.

Example: `'PredictorNames',{'x1','x2','x3','x4'}`

Data Types: `string` | `cell`

`RelTol` — Convergence threshold for coordinate descent algorithm
`1e–4` (default) | positive scalar

Convergence threshold for the coordinate descent algorithm [3], specified as the comma-separated pair consisting of `'RelTol'` and a positive scalar. The algorithm terminates when successive estimates of the coefficient vector differ in the L^2 norm by a relative amount less than `RelTol`.

Example: `'RelTol',5e–3`

Data Types: `single` | `double`

`Rho` — Augmented Lagrangian parameter
positive scalar

Augmented Lagrangian parameter ρ for the "ADMM Algorithm" on page 33-3432, specified as the comma-separated pair consisting of `'Rho'` and a positive scalar. The default is automatic selection.

Note This option applies only when you use `lasso` on tall arrays. See "Extended Capabilities" on page 33-0 for more information.

Example: `'Rho',2`

Data Types: `single` | `double`

`Standardize` — Flag for standardizing predictor data before fitting models
`true` (default) | `false`

Flag for standardizing the predictor data `X` before fitting the models, specified as the comma-separated pair consisting of `'Standardize'` and either `true` or `false`. If `Standardize` is `true`, then the `X` data is scaled to have zero mean and a variance of one. `Standardize` affects whether the regularization is applied to the coefficients on the standardized scale or the original scale. The results are always presented on the original data scale.

If `Intercept` is `false`, then the software sets `Standardize` to `false`, regardless of the `Standardize` value you specify.

`X` and `y` are always centered when `Intercept` is `true`.

Example: `'Standardize',false`

Data Types: `logical`

`U0` — Initial value of scaled dual variable
vector of zeros (default) | numeric vector

Initial value of the scaled dual variable *u* in the "ADMM Algorithm" on page 33-3432, specified as the comma-separated pair consisting of `'U0'` and a numeric vector.

Note This option applies only when you use `lasso` on tall arrays. See "Extended Capabilities" on page 33-0 for more information.

Data Types: `single` | `double`

`Weights` — Observation weights
`1/n*ones(n,1)` (default) | nonnegative vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a nonnegative vector. `Weights` has length *n*, where *n* is the number of rows of `X`. The `lasso` function scales `Weights` to sum to `1`.

Data Types: `single` | `double`

Output Arguments

`B` — Fitted coefficients
numeric matrix

Fitted coefficients, returned as a numeric matrix. `B` is a *p*-by-*L* matrix, where *p* is the number of predictors (columns) in `X`, and *L* is the number of `Lambda` values. You can specify the number of `Lambda` values using the `NumLambda` name-value pair argument.

The coefficient corresponding to the intercept term is a field in `FitInfo`.

Data Types: `single` | `double`

`FitInfo` — Fit information of models
structure

Fit information of the linear models, returned as a structure with the fields described in this table.

Field in `FitInfo`	Description
`Intercept`	Intercept term β_0 for each linear model, a 1-by-L vector
`Lambda`	Lambda parameters in ascending order, a 1-by-L vector
`Alpha`	Value of the `Alpha` parameter, a scalar
`DF`	Number of nonzero coefficients in `B` for each value of `Lambda`, a 1-by-L vector
`MSE`	Mean squared error (MSE), a 1-by-L vector
`PredictorNames`	Value of the `PredictorNames` parameter, stored as a cell array of character vectors

If you set the `CV` name-value pair argument to cross-validate, the `FitInfo` structure contains these additional fields.

Field in `FitInfo`	Description
`SE`	Standard error of MSE for each `Lambda`, as calculated during cross-validation, a 1-by-L vector
`LambdaMinMSE`	`Lambda` value with the minimum MSE, a scalar
`Lambda1SE`	Largest `Lambda` value such that MSE is within one standard error of the minimum MSE, a scalar
`IndexMinMSE`	Index of `Lambda` with the value `LambdaMinMSE`, a scalar
`Index1SE`	Index of `Lambda` with the value `Lambda1SE`, a scalar

More About

Lasso

For a given value of λ, a nonnegative parameter, `lasso` solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right).$$

- N is the number of observations.
- y_i is the response at observation i.
- x_i is data, a vector of length p at observation i.
- λ is a nonnegative regularization parameter corresponding to one value of `Lambda`.
- The parameters β_0 and β are a scalar and a vector of length p, respectively.

As λ increases, the number of nonzero components of β decreases.

The lasso problem involves the L^1 norm of β, as contrasted with the elastic net algorithm.

Elastic Net

For α strictly between 0 and 1, and nonnegative λ, elastic net solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda P_\alpha(\beta)\right),$$

where

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha\left|\beta_j\right|\right).$$

Elastic net is the same as lasso when $\alpha = 1$. For other values of α, the penalty term $P_\alpha(\beta)$ interpolates between the L^1 norm of β and the squared L^2 norm of β. As α shrinks toward 0, elastic net approaches `ridge` regression.

Algorithms

ADMM Algorithm

When operating on tall arrays, `lasso` uses an algorithm based on the Alternating Direction Method of Multipliers (ADMM) [5]. The notation used here is the same as in the reference paper. This method solves problems of the form

Minimize $l(x) + g(z)$

Subject to $Ax + Bz = c$

Using this notation, the lasso regression problem is

Minimize $l(x)+g(z)=\frac{1}{2}\|Ax-b\|_2^2+\lambda\|z\|_1$

Subject to $x - z = 0$

Because the loss function $l(x)=\frac{1}{2}\|Ax-b\|_2^2$ is quadratic, the iterative updates performed by the algorithm amount to solving a linear system of equations with a single coefficient matrix but several right-hand sides. The updates performed by the algorithm during each iteration are

$$x^{k+1}=\left(A^TA+\rho I\right)^{-1}\left(A^Tb+\rho\left(z^k-u^k\right)\right)$$
$$z^{k+1}=S_{\lambda/\rho}\left(x^{k+1}+u^k\right)$$
$$u^{k+1}=u^k+x^{k+1}-z^{k+1}$$

A is the dataset (a tall array), x contains the coefficients, ρ is the penalty parameter (augmented Lagrangian parameter), b is the response (a tall array), and S is the soft thresholding operator.

$$S_\kappa(a)=\begin{cases}a-\kappa, & a>\kappa\\ 0, & |a|\le\kappa\\ a+\kappa, & a<\kappa\end{cases}.$$

`lasso` solves the linear system using Cholesky factorization because the coefficient matrix $A^TA+\rho I$ is symmetric and positive definite. Because ρ does not change between iterations, the Cholesky factorization is cached between iterations.

Even though A and b are tall arrays, they appear only in the terms A^TA and A^Tb. The results of these two matrix multiplications are small enough to fit in memory, so they are precomputed and the iterative updates between iterations are performed entirely within memory.

References

[1] Tibshirani, R. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society.* Series B, Vol. 58, No. 1, 1996, pp. 267–288.

[2] Zou, H., and T. Hastie. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society.* Series B, Vol. 67, No. 2, 2005, pp. 301–320.

[3] Friedman, J., R. Tibshirani, and T. Hastie. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software.* Vol. 33, No. 1, 2010. `https://www.jstatsoft.org/v33/i01`

[4] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* 2nd edition. New York: Springer, 2008.

[5] Boyd, S. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers.” *Foundations and Trends in Machine Learning.* Vol. 3, No. 1, 2010, pp. 1–122.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- With tall arrays, `lasso` uses an algorithm based on ADMM (Alternating Direction Method of Multipliers).
- No elastic net support. The `'Alpha'` parameter is always 1.
- No cross-validation (`'CV'` parameter) support, which includes the related parameter `'MCReps'`.
- The output `FitInfo` does not contain the additional fields `'SE'`, `'LambdaMinMSE'`, `'Lambda1SE'`, `'IndexMinMSE'`, and `'Index1SE'`.
- The `'Options'` parameter is not supported because it does not contain options that apply to the ADMM algorithm. You can tune the ADMM algorithm using name-value pair arguments.
- Supported name-value pair arguments are:
  - `'Lambda'`
  - `'LambdaRatio'`
  - `'NumLambda'`
  - `'Standardize'`
  - `'PredictorNames'`
  - `'RelTol'`
  - `'Weights'`
- Additional name-value pair arguments to control the ADMM algorithm are:

- `'Rho'` — Augmented Lagrangian parameter, ρ. The default value is automatic selection.
- `'AbsTol'` — Absolute tolerance used to determine convergence. The default value is `1e–4`.
- `'MaxIter'` — Maximum number of iterations. The default value is `1e4`.
- `'B0'` — Initial values for the coefficients x. The default value is a vector of zeros.
- `'U0'` — Initial values of the scaled dual variable u. The default value is a vector of zeros.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also
`fitlm` | `fitrlinear` | `lassoPlot` | `lassoglm` | `ridge`

Topics
“Lasso Regularization” on page 11-120
“Lasso and Elastic Net with Cross Validation” on page 11-123
“Wide Data via Lasso and Parallel Computing” on page 11-115
“Lasso and Elastic Net” on page 11-112
“Introduction to Feature Selection” on page 15-49

Introduced in R2011b

lassoglm

Lasso or elastic net regularization for generalized linear models

Syntax

```
B = lassoglm(X,y)
B = lassoglm(X,y,distr)
B = lassoglm(X,y,distr,Name,Value)
[B,FitInfo] = lassoglm( ___ )
```

Description

`B = lassoglm(X,y)` returns penalized, maximum-likelihood fitted coefficients for generalized linear models of the predictor data `X` and the response `y`, where the values in `y` are assumed to have a normal probability distribution. Each column of `B` corresponds to a particular regularization coefficient in `Lambda`. By default, `lassoglm` performs lasso regularization using a geometric sequence of `Lambda` values.

`B = lassoglm(X,y,distr)` performs lasso regularization to fit the models using the probability distribution `distr` for `y`.

`B = lassoglm(X,y,distr,Name,Value)` fits regularized generalized linear regressions with additional options specified by one or more name-value pair arguments. For example, `'Alpha',0.5` sets elastic net as the regularization method, with the parameter `Alpha` equal to 0.5.

`[B,FitInfo] = lassoglm( ___ )` also returns the structure `FitInfo`, which contains information about the fit of the models, using any of the input arguments in the previous syntaxes.

Examples

Remove Redundant Predictors Using Lasso Regularization

Construct a data set with redundant predictors and identify those predictors by using `lassoglm`.

Create a random matrix `X` with 100 observations and 10 predictors. Create the normally distributed response `y` using only four of the predictors and a small amount of noise.

```
rng default
X = randn(100,10);
weights = [0.6;0.5;0.7;0.4];
y = X(:,[2 4 5 7])*weights + randn(100,1)*0.1; % Small added noise
```

Perform lasso regularization.

```
B = lassoglm(X,y);
```

Find the coefficient vector for the 75th `Lambda` value in `B`.

```
B(:,75)
```

```
ans = 10×1

         0
    0.5431
         0
    0.3944
    0.6173
         0
    0.3473
         0
         0
         0
```

`lassoglm` identifies and removes the redundant predictors.

Cross-Validated Lasso Regularization of Generalized Linear Model

Construct data from a Poisson model, and identify the important predictors by using `lassoglm`.

Create data with 20 predictors. Create a Poisson response variable using only three of the predictors plus a constant.

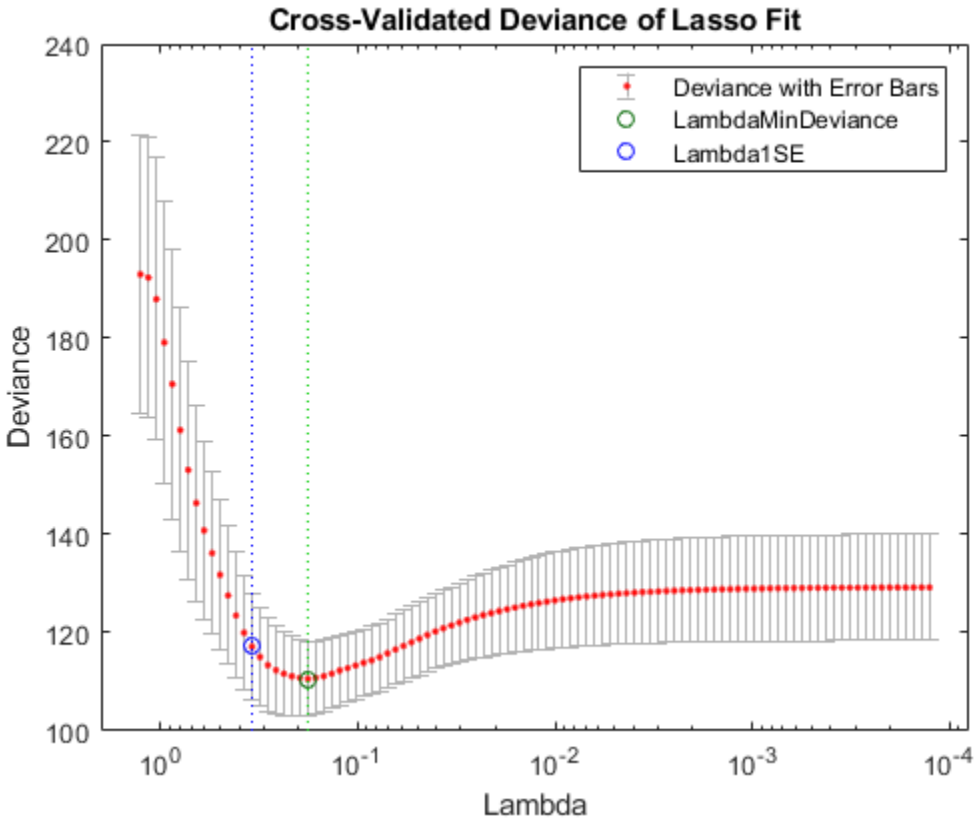
```
rng default % For reproducibility
X = randn(100,20);
weights = [.4;.2;.3];
mu = exp(X(:,[5 10 15])*weights + 1);
y = poissrnd(mu);
```

Construct a cross-validated lasso regularization of a Poisson regression model of the data.

```
[B,FitInfo] = lassoglm(X,y,'poisson','CV',10);
```

Examine the cross-validation plot to see the effect of the `Lambda` regularization parameter.

```
lassoPlot(B,FitInfo,'plottype','CV');
legend('show') % Show legend
```



The green circle and dotted line locate the `Lambda` with minimum cross-validation error. The blue circle and dotted line locate the point with minimum cross-validation error plus one standard deviation.

Find the nonzero model coefficients corresponding to the two identified points.

```
idxLambdaMinDeviance = FitInfo.IndexMinDeviance;
mincoefs = find(B(:,idxLambdaMinDeviance))
```

```
mincoefs = 7×1

     3
     5
     6
    10
    11
    15
    16
```

```
idxLambda1SE = FitInfo.Index1SE;
min1coefs = find(B(:,idxLambda1SE))
```

```
min1coefs = 3×1

     5
    10
```

```
    15
```

The coefficients from the minimum-plus-one standard error point are exactly those coefficients used to create the data.

Predict Values Using Lasso Regularization

Predict whether students got a B or above on their last exam by using `lassoglm`.

Load the `examgrades` data set. Convert the last exam grades to a logical vector, where `1` represents a grade of 80 or above and `0` represents a grade below 80.

```
load examgrades
X = grades(:,1:4);
y = grades(:,5);
yBinom = (y>=80);
```

Partition the data into training and test sets.

```
rng default    % Set the seed for reproducibility
c = cvpartition(yBinom,'HoldOut',0.3);
idxTrain = training(c,1);
idxTest = ~idxTrain;
XTrain = X(idxTrain,:);
yTrain = yBinom(idxTrain);
XTest = X(idxTest,:);
yTest = yBinom(idxTest);
```

Perform lasso regularization for generalized linear model regression with 3-fold cross-validation on the training data. Assume the values in `y` are binomially distributed. Choose model coefficients corresponding to the `Lambda` with minimum expected deviance.

```
[B,FitInfo] = lassoglm(XTrain,yTrain,'binomial','CV',3);
idxLambdaMinDeviance = FitInfo.IndexMinDeviance;
B0 = FitInfo.Intercept(idxLambdaMinDeviance);
coef = [B0; B(:,idxLambdaMinDeviance)]
```

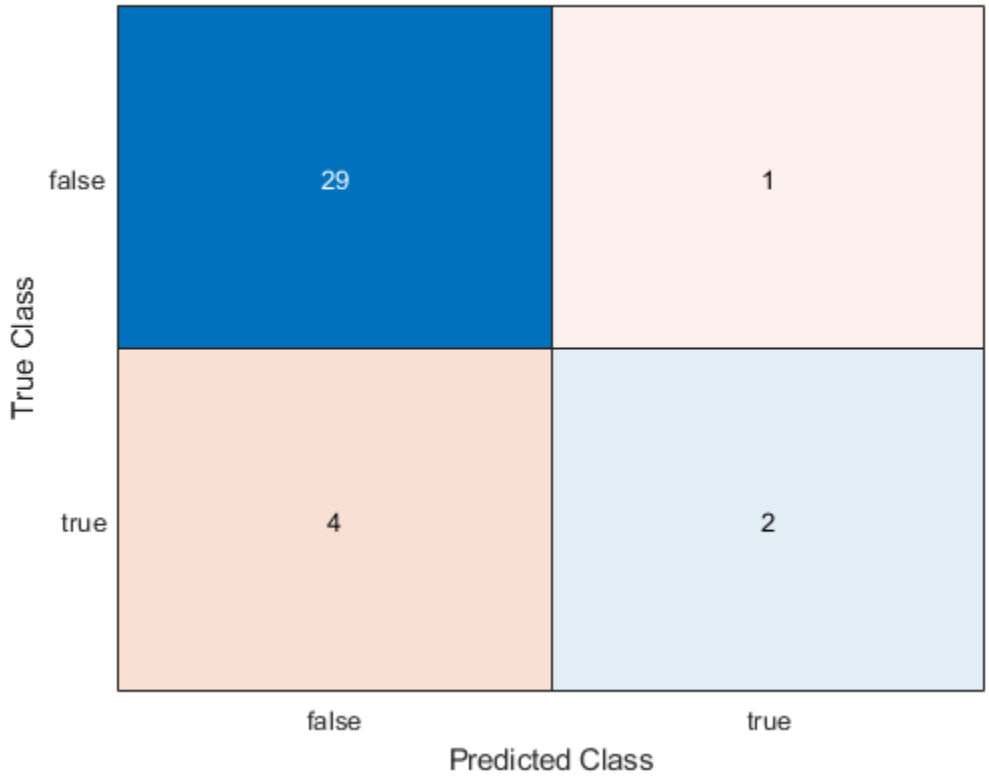
```
coef = 5×1

  -21.1911
    0.0235
    0.0670
    0.0693
    0.0949
```

Predict exam grades for the test data using the model coefficients found in the previous step. Specify the link function for a binomial response using `'logit'`. Convert the prediction values to a logical vector.

```
yhat = glmval(coef,XTest,'logit');
yhatBinom = (yhat>=0.5);
```

Determine the accuracy of the predictions using a confusion matrix.

```
c = confusionchart(yTest,yhatBinom);
```



The function correctly predicts 31 exam grades. However, the function incorrectly predicts that 1 student receives a B or above and 4 students receive a grade below a B.

Input Arguments

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row represents one observation, and each column represents one predictor variable.

Data Types: `single` | `double`

y — Response data

numeric vector | logical vector | categorical array | numeric matrix

Response data, specified as a numeric vector, logical vector, categorical array, or two-column numeric matrix.

- When `distr` is not `'binomial'`, `y` is a numeric vector or categorical array of length *n*, where *n* is the number of rows in `X`. The response `y(i)` corresponds to row *i* in `X`.
- When `distr` is `'binomial'`, `y` is one of the following:

- Numeric vector of length *n*, where each entry represents success (`1`) or failure (`0`)
- Logical vector of length *n*, where each entry represents success or failure
- Categorical array of length *n*, where each entry represents success or failure
- Two-column numeric matrix, where the first column contains the number of successes for each observation and the second column contains the total number of trials

Data Types: `single` | `double` | `logical` | `categorical`

`distr` — Distribution of response data
`'normal'` | `'binomial'` | `'poisson'` | `'gamma'` | `'inverse gaussian'`

Distribution of response data, specified as one of the following:

- `'normal'`
- `'binomial'`
- `'poisson'`
- `'gamma'`
- `'inverse gaussian'`

`lassoglm` uses the default link function on page 33-3445 corresponding to `distr`. Specify another link function using the `Link` name-value pair argument.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `lassoglm(X,y,'poisson','Alpha',0.5)` performs elastic net regularization assuming that the response values are Poisson distributed. The `'Alpha',0.5` name-value pair argument sets the parameter used in the elastic net optimization.

`Alpha` — Weight of lasso versus ridge optimization
`1` (default) | positive scalar

Weight of lasso (L^1) versus ridge (L^2) optimization, specified as the comma-separated pair consisting of `'Alpha'` and a positive scalar value in the interval `(0,1]`. The value `Alpha = 1` represents lasso regression, `Alpha` close to `0` approaches ridge regression, and other values represent elastic net optimization. See “Elastic Net” on page 33-3446.

Example: `'Alpha',0.75`

Data Types: `single` | `double`

`CV` — Cross-validation specification for estimating deviance
`'resubstitution'` (default) | positive integer scalar | `cvpartition` object

Cross-validation specification for estimating the deviance, specified as the comma-separated pair consisting of `'CV'` and one of the following:

- `'resubstitution'` — `lassoglm` uses `X` and `y` to fit the model and to estimate the deviance without cross-validation.
- Positive scalar integer `K` — `lassoglm` uses `K`-fold cross-validation.

- `cvpartition` object `cvp` — `lassoglm` uses the cross-validation method expressed in `cvp`. You cannot use a `'leaveout'` partition with `lassoglm`.

Example: `'CV',10`

`DFmax` — Maximum number of nonzero coefficients
`Inf` (default) | positive integer scalar

Maximum number of nonzero coefficients in the model, specified as the comma-separated pair consisting of `'DFmax'` and a positive integer scalar. `lassoglm` returns results only for `Lambda` values that satisfy this criterion.

Example: `'DFmax',25`

Data Types: `single` | `double`

`Lambda` — Regularization coefficients
nonnegative vector

Regularization coefficients, specified as the comma-separated pair consisting of `'Lambda'` and a vector of nonnegative values. See “Lasso” on page 33-3446.

- If you do not supply `Lambda`, then `lassoglm` estimates the largest value of `Lambda` that gives a nonnull model. In this case, `LambdaRatio` gives the ratio of the smallest to the largest value of the sequence, and `NumLambda` gives the length of the vector.
- If you supply `Lambda`, then `lassoglm` ignores `LambdaRatio` and `NumLambda`.
- If `Standardize` is `true`, then `Lambda` is the set of values used to fit the models with the `X` data standardized to have zero mean and a variance of one.

The default is a geometric sequence of `NumLambda` values, with only the largest value able to produce `B = 0`.

Data Types: `single` | `double`

`LambdaRatio` — Ratio of smallest to largest `Lambda` values
`1e–4` (default) | positive scalar

Ratio of the smallest to the largest `Lambda` values when you do not supply `Lambda`, specified as the comma-separated pair consisting of `'LambdaRatio'` and a positive scalar.

If you set `LambdaRatio = 0`, then `lassoglm` generates a default sequence of `Lambda` values and replaces the smallest one with `0`.

Example: `'LambdaRatio',1e–2`

Data Types: `single` | `double`

`Link` — Mapping between mean of response and linear predictor
`'comploglog'` | `'identity'` | `'log'` | `'logit'` | `'loglog'` | ...

Mapping between the mean μ of the response and the linear predictor Xb, specified as the comma-separated pair consisting of `'Link'` and one of the values in this table.

Value	Description
`'comploglog'`	$\log(-\log((1 - \mu))) = Xb$

Value	Description
`'identity'`, default for the distribution `'normal'`	$\mu = Xb$
`'log'`, default for the distribution `'poisson'`	$\log(\mu) = Xb$
`'logit'`, default for the distribution `'binomial'`	$\log(\mu/(1 - \mu)) = Xb$
`'loglog'`	$\log(-\log(\mu)) = Xb$
`'probit'`	$\Phi^{-1}(\mu) = Xb$, where Φ is the normal (Gaussian) cumulative distribution function
`'reciprocal'`, default for the distribution `'gamma'`	$\mu^{-1} = Xb$
`p` (a number), default for the distribution `'inverse gaussian'` (with $p = -2$)	$\mu^p = Xb$
A cell array of the form `{FL FD FI}`, containing three function handles created using `@`, which define the link (`FL`), the derivative of the link (`FD`), and the inverse link (`FI`). Or, a structure of function handles with the field `Link` containing `FL`, the field `Derivative` containing `FD`, and the field `Inverse` containing `FI`.	User-specified link function (see "Custom Link Function" on page 12-12)

Example: `'Link','probit'`

Data Types: `char` | `string` | `single` | `double` | `cell`

`MaxIter` — Maximum number of iterations allowed
`1e4` (default) | positive integer scalar

Maximum number of iterations allowed, specified as the comma-separated pair consisting of `'MaxIter'` and a positive integer scalar.

If the algorithm executes `MaxIter` iterations before reaching the convergence tolerance `RelTol`, then the function stops iterating and returns a warning message.

The function can return more than one warning when `NumLambda` is greater than `1`.

Example: `'MaxIter',1e3`

Data Types: `single` | `double`

`MCReps` — Number of Monte Carlo repetitions for cross-validation
`1` (default) | positive integer scalar

Number of Monte Carlo repetitions for cross-validation, specified as the comma-separated pair consisting of `'MCReps'` and a positive integer scalar.

- If `CV` is `'resubstitution'` or a `cvpartition` of type `'resubstitution'`, then `MCReps` must be `1`.
- If `CV` is a `cvpartition` of type `'holdout'`, then `MCReps` must be greater than `1`.

Example: `'MCReps',2`

Data Types: `single` | `double`

`NumLambda` — Number of Lambda values
`100` (default) | positive integer scalar

Number of `Lambda` values `lassoglm` uses when you do not supply `Lambda`, specified as the comma-separated pair consisting of `'NumLambda'` and a positive integer scalar. `lassoglm` can return fewer than `NumLambda` fits if the deviance of the fits drops below a threshold fraction of the null deviance (deviance of the fit without any predictors `X`).

Example: `'NumLambda',150`

Data Types: `single` | `double`

`Offset` — Additional predictor variable
numeric vector

Additional predictor variable, specified as the comma-separated pair consisting of `'Offset'` and a numeric vector with the same number of rows as `X`. The `lassoglm` function keeps the coefficient value of `Offset` fixed at `1.0`.

Data Types: `single` | `double`

`Options` — Option to cross-validate in parallel and specify random streams
structure

Option to cross-validate in parallel and specify the random streams, specified as the comma-separated pair consisting of `'Options'` and a structure. This option requires Parallel Computing Toolbox.

Create the `Options` structure with `statset`. The option fields are:

- `UseParallel` — Set to `true` to compute in parallel. The default is `false`.
- `UseSubstreams` — Set to `true` to compute in parallel in a reproducible fashion. For reproducibility, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`. The default is `false`.
- `Streams` — A `RandStream` object or cell array consisting of one such object. If you do not specify `Streams`, then `lassoglm` uses the default stream.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

`PredictorNames` — Names of predictor variables
`{}` (default) | string array | cell array of character vectors

Names of the predictor variables, in the order in which they appear in `X`, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array or cell array of character vectors.

Example: `'PredictorNames',{'Height','Weight','Age'}`

Data Types: `string` | `cell`

`RelTol` — Convergence threshold for coordinate descent algorithm
`1e–4` (default) | positive scalar

Convergence threshold for the coordinate descent algorithm [3], specified as the comma-separated pair consisting of `'RelTol'` and a positive scalar. The algorithm terminates when successive estimates of the coefficient vector differ in the L^2 norm by a relative amount less than `RelTol`.

Example: `'RelTol',2e–3`

Data Types: `single` | `double`

`Standardize` — Flag for standardizing predictor data before fitting models
`true` (default) | `false`

Flag for standardizing the predictor data `X` before fitting the models, specified as the comma-separated pair consisting of `'Standardize'` and either `true` or `false`. If `Standardize` is `true`, then the `X` data is scaled to have zero mean and a variance of one. `Standardize` affects whether the regularization is applied to the coefficients on the standardized scale or the original scale. The results are always presented on the original data scale.

Example: `'Standardize',false`

Data Types: `logical`

`Weights` — Observation weights
`1/n*ones(n,1)` (default) | nonnegative vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a nonnegative vector. `Weights` has length *n*, where *n* is the number of rows of `X`. At least two values must be positive.

Data Types: `single` | `double`

Output Arguments

`B` — Fitted coefficients
numeric matrix

Fitted coefficients, returned as a numeric matrix. `B` is a *p*-by-*L* matrix, where *p* is the number of predictors (columns) in `X`, and *L* is the number of `Lambda` values. You can specify the number of `Lambda` values using the `NumLambda` name-value pair argument.

The coefficient corresponding to the intercept term is a field in `FitInfo`.

Data Types: `single` | `double`

`FitInfo` — Fit information of models
structure

Fit information of the generalized linear models, returned as a structure with the fields described in this table.

Field in `FitInfo`	Description
`Intercept`	Intercept term β_0 for each linear model, a `1`-by-*L* vector
`Lambda`	Lambda parameters in ascending order, a `1`-by-*L* vector
`Alpha`	Value of the `Alpha` parameter, a scalar

Field in FitInfo	Description
DF	Number of nonzero coefficients in B for each value of Lambda, a 1-by-*L* vector
Deviance	Deviance of the fitted model for each value of Lambda, a 1-by-*L* vector If the model is cross-validated, then the values for Deviance represent the estimated expected deviance of the model applied to new data, as calculated by cross-validation. Otherwise, Deviance is the deviance of the fitted model applied to the data used to perform the fit.
PredictorNames	Value of the PredictorNames parameter, stored as a cell array of character vectors

If you set the CV name-value pair argument to cross-validate, the FitInfo structure contains these additional fields.

Field in FitInfo	Description
SE	Standard error of Deviance for each Lambda, as calculated during cross-validation, a 1-by-*L* vector
LambdaMinDeviance	Lambda value with minimum expected deviance, as calculated by cross-validation, a scalar
Lambda1SE	Largest Lambda value such that Deviance is within one standard error of the minimum, a scalar
IndexMinDeviance	Index of Lambda with the value LambdaMinDeviance, a scalar
Index1SE	Index of Lambda with the value Lambda1SE, a scalar

More About

Link Function

A link function $f(\mu)$ maps a distribution with mean μ to a linear model with data X and coefficient vector b using the formula

$f(\mu) = Xb$.

You can find the formulas for the link functions in the Link name-value pair argument description. This table lists the link functions that are typically used for each distribution.

Distribution Family	Default Link Function	Other Typical Link Functions
'normal'	'identity'	
'binomial'	'logit'	'comploglog', 'loglog', 'probit'
'poisson'	'log'	
'gamma'	'reciprocal'	
'inverse gaussian'	–2	

Lasso

For a nonnegative value of λ, `lassoglm` solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\text{Deviance}(\beta_0,\beta)+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right).$$

- The function Deviance in this equation is the deviance of the model fit to the responses using the intercept β_0 and the predictor coefficients β. The formula for Deviance depends on the `distr` parameter you supply to `lassoglm`. Minimizing the λ-penalized deviance is equivalent to maximizing the λ-penalized loglikelihood.
- N is the number of observations.
- λ is a nonnegative regularization parameter corresponding to one value of `Lambda`.
- The parameters β_0 and β are a scalar and a vector of length p, respectively.

As λ increases, the number of nonzero components of β decreases.

The lasso problem involves the L^1 norm of β, as contrasted with the elastic net algorithm.

Elastic Net

For α strictly between 0 and 1, and nonnegative λ, elastic net solves the problem

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\text{Deviance}(\beta_0,\beta)+\lambda P_\alpha(\beta)\right),$$

where

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha\left|\beta_j\right|\right).$$

Elastic net is the same as lasso when $\alpha = 1$. For other values of α, the penalty term $P_\alpha(\beta)$ interpolates between the L^1 norm of β and the squared L^2 norm of β. As α shrinks toward 0, elastic net approaches `ridge` regression.

References

[1] Tibshirani, R. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society.* Series B, Vol. 58, No. 1, 1996, pp. 267–288.

[2] Zou, H., and T. Hastie. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society.* Series B, Vol. 67, No. 2, 2005, pp. 301–320.

[3] Friedman, J., R. Tibshirani, and T. Hastie. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software.* Vol. 33, No. 1, 2010. `https://www.jstatsoft.org/v33/i01`

[4] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* 2nd edition. New York: Springer, 2008.

[5] Dobson, A. J. *An Introduction to Generalized Linear Models.* 2nd edition. New York: Chapman & Hall/CRC Press, 2002.

[6] McCullagh, P., and J. A. Nelder. *Generalized Linear Models.* 2nd edition. New York: Chapman & Hall/CRC Press, 1989.

[7] Collett, D. *Modelling Binary Data.* 2nd edition. New York: Chapman & Hall/CRC Press, 2003.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`fitglm` | `glmval` | `lasso` | `lassoPlot` | `ridge`

Topics

“Lasso Regularization of Generalized Linear Models” on page 12-32
“Regularize Poisson Regression” on page 12-34
“Regularize Logistic Regression” on page 12-36
“Regularize Wide Data in Parallel” on page 12-43
“Introduction to Feature Selection” on page 15-49

Introduced in R2012a

lassoPlot

Trace plot of lasso fit

Syntax

```
lassoPlot(B)
lassoPlot(B,FitInfo)
lassoPlot(B,FitInfo,Name,Value)
[ax,figh] = lassoPlot( ___ )
```

Description

`lassoPlot(B)` creates a trace plot of the values in `B` against the L^1 norm of `B`.

`lassoPlot(B,FitInfo)` creates a plot with type depending on the data type of `FitInfo` and the value, if any, of the `PlotType` name-value pair.

`lassoPlot(B,FitInfo,Name,Value)` creates a plot with additional options specified by one or more `Name,Value` pair arguments.

`[ax,figh] = lassoPlot( ___ )`, for any previous input syntax, returns a handle `ax` to the plot axis, and a handle `figh` to the figure window.

Input Arguments

`B`

Coefficients of a sequence of regression fits, as returned from the `lasso` or `lassoglm` functions. `B` is a `p`-by-`NLambda` matrix, where `p` is the number of predictors, and each column of `B` is a set of coefficients `lasso` calculates using one `Lambda` penalty value.

`FitInfo`

Information controlling the plot:

- `FitInfo` is a structure, especially as returned from `lasso` or `lassoglm` — `lassoPlot` creates a plot based on the `PlotType` name-value pair.
- `FitInfo` is a vector — `lassoPlot` forms the *x*-axis of the plot from the values in `FitInfo`. The length of `FitInfo` must equal the number of columns of `B`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Parent`

Axis in which to draw the plot.

Default: New plot

`PlotType`

Plot type when you specify a `FitInfo` vector or structure:

`PlotType`	Plot
`'L1'`	`lassoPlot` creates the *x*-axis from the L^1 norm of the coefficients in `B`. The *x*-axis at the top of the plot contains the degrees of freedom (`df`), meaning the number of nonzero coefficients of `B`.
`'Lambda'` When you choose this value, `FitInfo` must be a structure.	`lassoPlot` creates the *x*-axis from the `Lambda` field of `FitInfo`. The *x*-axis at the top of the plot contains the degrees of freedom (`df`), meaning the number of nonzero coefficients of `B`.
`'CV'` When you choose this value, `FitInfo` must be a cross-validated structure.	• For each `Lambda`, `lassoPlot` plots an estimate of the mean squared prediction error on new data for the model fitted by `lasso` with that value of `Lambda`. • `lassoPlot` plots error bars for the estimates.

If you include a cross-validated `FitInfo` structure, `lassoPlot` also indicates two specific `Lambda` values with green and blue dashed lines.

- A green, dashed line indicates the value of `Lambda` with a minimum cross-validated mean squared error (MSE).
- A blue, dashed line indicates the greatest `Lambda` that is within one standard error of the minimum MSE. This Lambda value makes the sparsest model with relatively low MSE.

To display the label for each plot in the legend of the figure, type `legend('show')` in the Command Window.

Default: `'L1'`

`PredictorNames`

String array or cell array of character vectors to label each coefficient of `B`. If the length of `PredictorNames` is less than the number of rows of `B`, the remaining labels are padded with default values.

`lassoPlot` uses `PredictorNames` in `FitInfo` only if:

- You created `FitInfo` with a call to `lasso` that included a `PredictorNames` name-value pair.
- You call `lassoPlot` *without* a `PredictorNames` name-value pair.
- You include `FitInfo` in your `lassoPlot` call.

For an example, see “Lasso Plot with Default Plot Type” on page 33-3450.

Default: `{'B1','B2',...}`

`XScale`

- `'linear'` for linear x-axis
- `'log'` for logarithmic scaled x-axis

Default: `'linear'`, except `'log'` for the `'CV'` plot type

Output Arguments

`ax`

Handle to the axis of the plot (see “Axes Appearance”).

`figh`

Handle to the figure window (see “Special Object Identifiers”).

Examples

Lasso Plot with Default Plot Type

Load the sample data

```
load acetylene
```

Prepare the design matrix for lasso fit with interactions.

```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
```

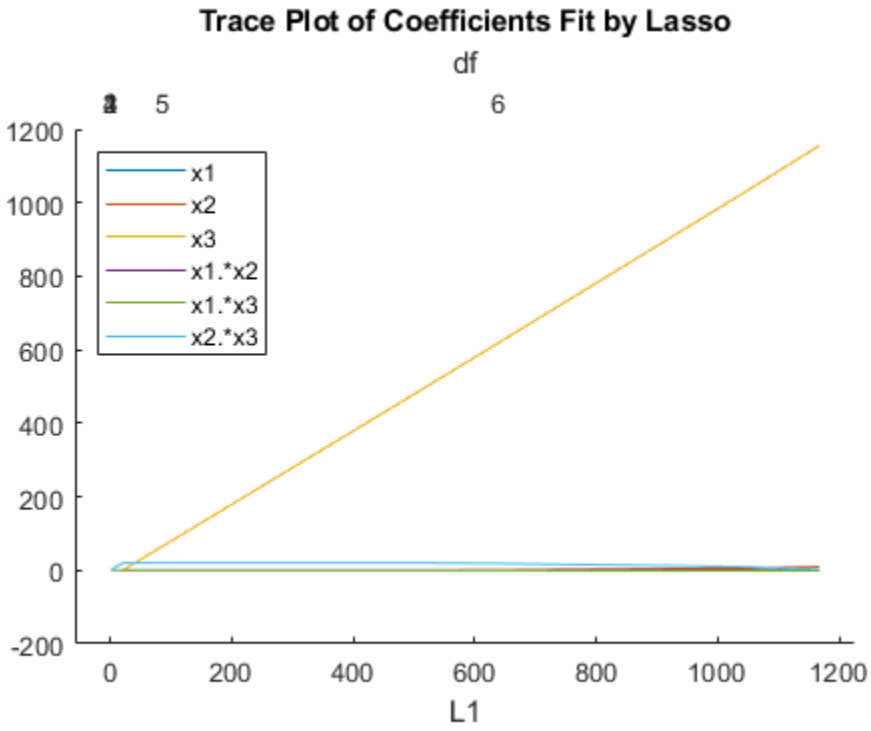
The `x2fx` function returns the quadratic model in the order of a constant term, linear terms and interaction terms: constant term, `x1`, `x2`, `x3`, `x1.*x2`, `x1.*x3`, and `x2.*x3`

Fit a regularized model of the data using `lasso`.

```
B = lasso(D,y);
```

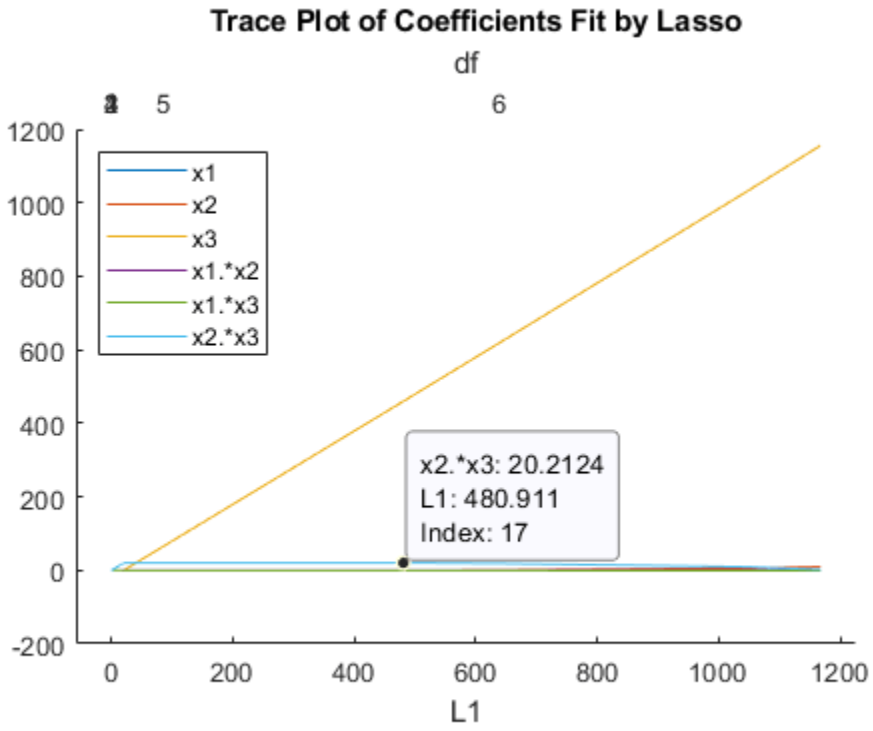
Plot the lasso fits with labeled coefficients by using the `PredictorNames` name-value pair.

```
lassoPlot(B,'PredictorNames',{'x1','x2','x3','x1.*x2','x1.*x3','x2.*x3'});
legend('show','Location','NorthWest') % Show legend
```



Each line represents a trace of the values in B for a single predictor variable: `x1`, `x2`, `x3`, `x1.*x2`, `x1.*x3`, and `x2.*x3`.

Display a data tip for the trace plot. A data tip appears when you hover over a data tip.



A data tip displays these lines of information: the name of the selected coefficient with a fitted value, the L1 norm of a set of coefficients including the selected coefficient, and the index of the corresponding Lambda.

Lasso Plot with Lambda Plot Type

Load the sample data.

```
load acetylene
```

Prepare the data for lasso fit with interactions.

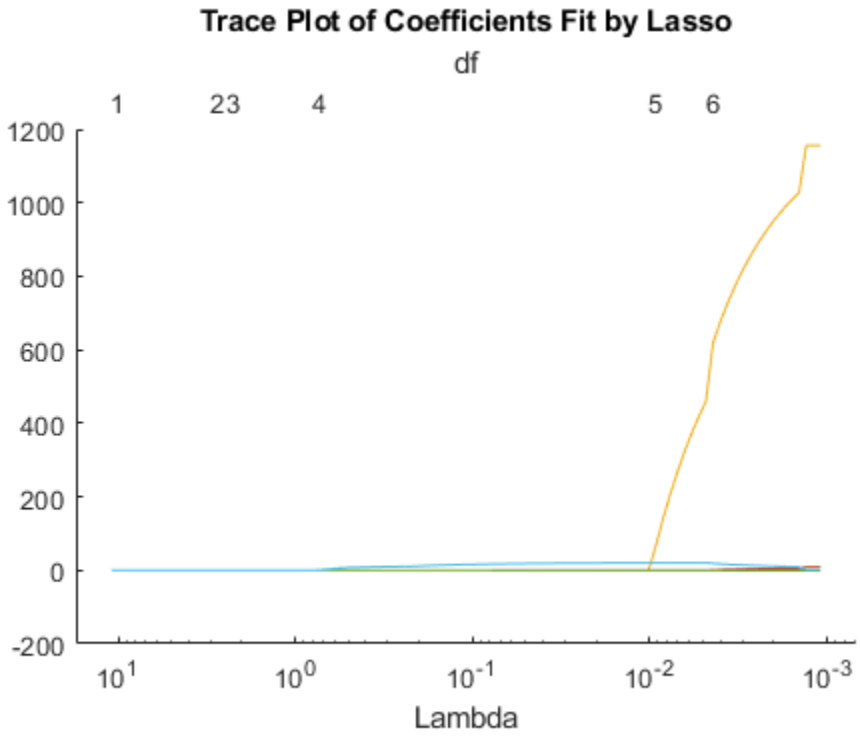
```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
```

Fit a regularized model of the data with `lasso`.

```
[B,FitInfo] = lasso(D,y);
```

Plot the fits with the `Lambda` plot type and logarithmic scaling.

```
lassoPlot(B,FitInfo,'PlotType','Lambda','XScale','log');
```



Lasso Plot with Cross-Validated Fits

Visually examine the cross-validated error of various levels of regularization.

Load the sample data.

```
load acetylene
```

Create a design matrix with interactions and no constant term.

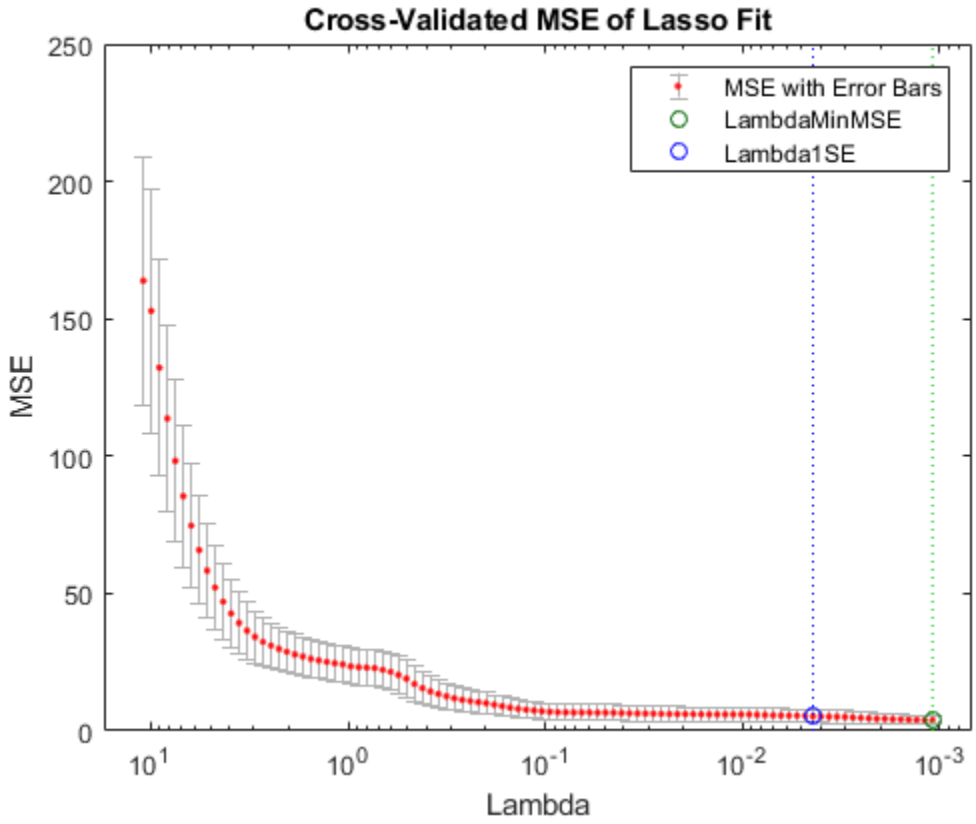
```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
```

Construct the lasso fit using 10-fold cross-validation. Include the `FitInfo` output so you can plot the result.

```
rng default % For reproducibility
[B,FitInfo] = lasso(D,y,'CV',10);
```

Plot the cross-validated fits.

```
lassoPlot(B,FitInfo,'PlotType','CV');
legend('show') % Show legend
```



The green circle and dotted line locate the `Lambda` with minimum cross-validation error. The blue circle and dotted line locate the point with minimum cross-validation error plus one standard deviation.

See Also

`lasso` | `lassoglm`

Topics

“Lasso and Elastic Net” on page 11-112

Introduced in R2011b

le

Class: `qrandstream`

Less than or equal relation for handles

Syntax

```
h1 <= h2
```

Description

Handles are equal if they are handles for the same object. All comparisons use a number associated with each handle object. Nothing can be assumed about the result of a handle comparison except that the repeated comparison of two handles in the same MATLAB session will yield the same result. The order of handle values is purely arbitrary and has no connection to the state of the handle objects being compared.

`h1 <= h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise `<=` result.

If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = le(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`eq` | `ge` | `gt` | `lt` | `ne` | `qrandstream`

learnerCoderConfigurer

Create coder configurer of machine learning model

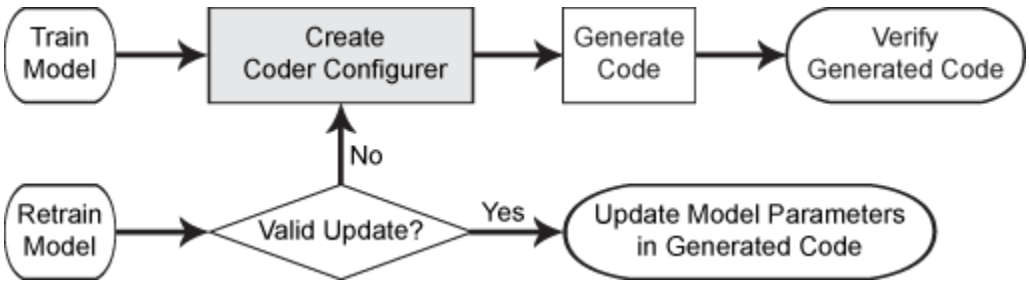
Syntax

```
configurer = learnerCoderConfigurer(Mdl,X)
configurer = learnerCoderConfigurer(Mdl,X,Name,Value)
```

Description

After training a machine learning model, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the object functions and properties of the configurer to specify code generation options and to generate C/C++ code for the `predict` and `update` functions of the machine learning model. Generating C/C++ code requires MATLAB Coder.

This flow chart shows the code generation workflow using a coder configurer. Use `learnerCoderConfigurer` for the highlighted step.



`configurer = learnerCoderConfigurer(Mdl,X)` returns the coder configurer `configurer` for the machine learning model `Mdl`. Specify the predictor data `X` for the `predict` function of `Mdl`.

`configurer = learnerCoderConfigurer(Mdl,X,Name,Value)` returns a coder configurer with additional options specified by one or more name-value pair arguments. For example, you can specify the number of output arguments in the `predict` function, the file name of generated C/C++ code, and the verbosity level of the coder configurer.

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `carsmall` data set and train a support vector machine (SVM) regression model.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X,Y);
```

`Mdl` is a `RegressionSVM` object.

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  RegressionSVMCoderConfigurer with properties:

   Update Inputs:
             Alpha: [1x1 LearnerCoderInput]
    SupportVectors: [1x1 LearnerCoderInput]
             Scale: [1x1 LearnerCoderInput]
              Bias: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'RegressionSVMModel'


  Properties, Methods
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionSVMModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionSVMModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('RegressionSVMModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Alpha
        %                       SupportVectors
        %                       Scale
        %                       Bias
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of SVM Classification Model in Generated Code

Train a SVM model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`). Train a binary SVM classification model using the first 50 observations.

```
load ionosphere
Mdl = fitcsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `ClassificationSVM` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted labels and scores.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:),'NumOutputs',2);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM classification model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 34];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 34 predictors, so the value of the `SizeVector` attribute must be 34 and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 34];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
VariableDimensions attribute for SupportVectorLabels has been modified to satisfy configuration c
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the

modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM classification model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `ClassificationSVMModel` for the two entry-point functions in the `codegen\mex\ClassificationSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score] = predict(Mdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
```

Compare `label` and `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`score_mex` might include round-off differences compared with `score`. In this case, compare `score_mex` and `score`, allowing a small tolerance.

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `score` and `score_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitcsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score] = predict(retrainedMdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `labels` and `labels_mex` are equal, and the score values are equal within the tolerance.

Input Arguments

`Mdl` — Machine learning model

full model object | compact model object

Machine learning model, specified as a full or compact model object, as given in this table of supported models.

Model	Full/Compact Model Object	Training Function
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`	`fitctree`
SVM for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	`fitcsvm`
Linear model for binary classification	`ClassificationLinear`	`fitclinear`

Model	Full/Compact Model Object	Training Function
Multiclass model for SVMs and linear models	`ClassificationECOC`, `CompactClassificationECOC`	`fitcecoc`
Binary decision tree for regression	`RegressionTree`, `CompactRegressionTree`	`fitrtree`
Support vector machine (SVM) regression	`RegressionSVM`, `CompactRegressionSVM`	`fitrsvm`
Linear regression	`RegressionLinear`	`fitrlinear`

For the code generation usage notes and limitations of a machine learning model, see the Code Generation section of the model object page.

`X` — Predictor data

numeric matrix

Predictor data for the `predict` function of `Mdl`, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictor variables. To instead specify `X` as a *p*-by-*n* matrix, where the observations correspond to columns, you must set the `'ObservationsIn'` name-value pair argument to `'columns'`. This option is available only for linear models and ECOC models with linear binary learners.

The `predict` function of a machine learning model predicts labels for classification and responses for regression for given predictor data. After creating the coder configurer `configurer`, you can use the `generateCode` function to generate C/C++ code for the `predict` function of `Mdl`. The generated code accepts predictor data that has the same size and data type of `X`. You can specify whether each dimension has a variable size or fixed size after creating `configurer`.

For example, if you want to generate C/C++ code that predicts labels using 100 observations with three predictor variables, then specify `X` as `zeros(100,3)`. The `learnerCoderConfigurer` function uses only the size and data type of `X`, not its values. Therefore, `X` can be predictor data or a MATLAB expression that represents the set of values with a certain data type. The output `configurer` stores the size and data type of `X` in the `X` property of `configurer`. You can modify the size and data type of `X` after creating `configurer`. For example, change the number of observations to 200 and the data type to `single`.

```
configurer.X.SizeVector = [200 3];
configurer.X.DataType = 'single';
```

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify `X` as `zeros(100,3)` and change the `VariableDimensions` property.

```
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

```
configurer.X.SizeVector = [Inf 3];
```

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `configurer = learnerCoderConfigurer(Mdl,X,'NumOutputs',2,'OutputFileName','myModel')` sets the number of outputs in `predict` to 2 and specifies the file name `'myModel'` for the generated C/C++ code.

`NumOutputs` — Number of outputs in `predict`

1 (default) | positive integer

Number of output arguments in the `predict` function of the machine learning model `Mdl`, specified as the comma-separated pair consisting of `'NumOutputs'` and a positive integer `n`.

This table lists the outputs for the `predict` function of different models. `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function in the order given in the Outputs column.

Model	`predict` Function of Model	Outputs
Binary decision tree for multiclass classification	`predict`	`label` (predicted class labels), `score` (posterior probabilities), `node` (node numbers for predicted classes), `cnum` (class numbers of predicted labels)
SVM for one-class and binary classification	`predict`	`label` (predicted class labels), `score` (scores or posterior probabilities)
Linear model for binary classification	`predict`	`Label` (predicted class labels), `Score` (classification scores)
Multiclass model for SVMs and linear models	`predict`	`label` (predicted class labels), `NegLoss` (negated average binary losses), `PBScore` (positive-class scores)
Binary decision tree for regression	`predict`	`Yfit` (predicted responses), `node` (node numbers for predictions)
SVM regression	`predict`	`yfit` (predicted responses)
Linear regression	`predict`	`YHat` (predicted responses)

For example, if you specify `'NumOutputs',1` for an SVM classification model, then `predict` returns predicted class labels in the generated C/C++ code.

After creating the coder configurer `configurer`, you can modify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `'NumOutputs'` name-value pair argument is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` of a coder configurer generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively—and generates C/C++ code for the two entry-point functions. The specified value for `'NumOutputs'` corresponds to the number of output arguments in `predict.m`.

Example: `'NumOutputs',2`

Data Types: `single` | `double`

`OutputFileName` — File name of generated C/C++ code
`Mdl` object name plus `'Model'` (default) | character vector | string scalar

File name of the generated C/C++ code, specified as the comma-separated pair consisting of `'OutputFileName'` and a character vector or string scalar.

The object function `generateCode` of a coder configurer generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

The default file name is the object name of `Mdl` followed by `'Model'`. For example, if `Mdl` is a `CompactClassificationSVM` or `ClassificationSVM` object, then the default name is `'ClassificationSVMModel'`.

After creating the coder configurer `configurer`, you can modify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Example: `'OutputFileName','myModel'`

Data Types: `char` | `string`

`Verbose` — Verbosity level
`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line for the coder configurer `configurer`.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Example: `'Verbose',false`

Data Types: `logical`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and either `'rows'` or `'columns'`. If you set `'ObservationsIn'` to `'columns'`, then the predictor data `X` must be oriented so that the observations correspond to columns.

Note The `'columns'` option is available only for linear models and ECOC models with linear binary learners.

Example: `'ObservationsIn','columns'`

Output Arguments

`configurer` — Coder configurer
coder configurer object

Coder configurer of a machine learning model, returned as one of the coder configurer objects in this table.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

Use the object functions and properties of a coder configurer object to configure code generation options and to generate C/C++ code for the `predict` and `update` functions of the machine learning model.

See Also
`generateCode` | `update` | `validatedUpdateInputs`

Topics
“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

length

Class: `dataset`

(Not Recommended) Length of dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
n = length(A)
```

Description

`n = length(A)` returns the number of observations in the dataset `A`. `length` is equivalent to `size(A,1)`.

See Also

`size`

levelcounts

(Not Recommended) Element counts by level of a nominal or ordinal array

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the "Categorical Arrays" data type instead.

Syntax

```
C = levelcounts(A)
C = levelcounts(A,dim)
```

Description

`C = levelcounts(A)` returns counts of the number of elements in the nominal or ordinal array `A` equal to each possible level in `A`.

- If `A` is a vector, then `C` is a vector containing as many elements as the number of levels in `A`.
- If `A` is a matrix, then `C` is a matrix of column counts.
- If `A` is an *N*-dimensional array, then `levelcounts` operates along the first nonsingleton dimension.

`C = levelcounts(A,dim)` operates along the dimension `dim`.

Examples

Count Observations in Each Level

Create a nominal array from data in a cell array.

```
colors = nominal({'r','b','g';'g','r','b';'b','r','g'},...
                 {'blue','green','red'})
```

```
colors = 3x3 nominal
     red        blue      green
     green      red       blue
     blue       red       green
```

Count the number of observations of each level in each column.

```
levelcounts(colors)
```

```
ans = 3×3

     1     1     1
     1     0     2
     1     2     0
```

Count the number of observations of each level in each row.

```
levelcounts(colors,2)
```

```
ans = 3×3

     1     1     1
     1     1     1
     1     1     1
```

Alternatively, you can use `summary` to display the counts with their labels. The default is to count elements in each column.

```
summary(colors)
```

```
     blue       1      1      1
     green      1      0      2
     red        1      2      0
```

You can also count elements in each row.

```
summary(colors,2)
```

```
     blue      green      red
     1         1          1
     1         1          1
     1         1          1
```

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

dim — Dimension along which to count

positive integer value

Dimension along which to count the number of elements in each level, specified as a positive integer value. For example, if the dimension is 1, then `levelcounts` counts along each column, while if the dimension is 2, then `levelcounts` counts along each row.

Data Types: `double` | `single`

See Also

`nominal` | `ordinal` | `summary`

Introduced in R2007a

leverage

Leverage

Syntax

```
h = leverage(data)
h = leverage(data,model)
```

Description

`h = leverage(data)` finds the leverage of each row (point) in the matrix `data` for a linear additive regression model.

`h = leverage(data,model)` finds the leverage on a regression, using a specified model type, where `model` can be one of the following:

- `'linear'` - includes constant and linear terms
- `'interaction'` - includes constant, linear, and cross product terms
- `'quadratic'` - includes interactions and squared terms
- `'purequadratic'` - includes constant, linear, and squared terms

Leverage is a measure of the influence of a given observation on a regression due to its location in the space of the inputs.

Examples

One rule of thumb is to compare the leverage to *2p/n* where *n* is the number of observations and *p* is the number of parameters in the model. For the Hald data set this value is 0.7692.

```
load hald
h = max(leverage(ingredients,'linear'))
h =
  0.7004
```

Since 0.7004 < 0.7692, there are no high leverage points using this rule.

Algorithms

```
[Q,R] = qr(x2fx(data,'model'),0);

leverage = (sum(Q'.*Q'))'
```

References

[1] Goodall, C. R. “Computation Using the QR Decomposition.” *Handbook in Statistics.* Vol. 9, Amsterdam: Elsevier/North-Holland, 1993.

See Also

Topics
`regstats`

Introduced before R2006a

lhsdesign

Latin hypercube sample

Syntax

```
X = lhsdesign(n,p)
X = lhsdesign(n,p,Name,Value)
```

Description

`X = lhsdesign(n,p)` returns a Latin hypercube sample matrix of size `n`-by-`p`. For each column of `X`, the `n` values are randomly distributed with one from each interval `(0,1/n)`, `(1/n,2/n)`, ..., `(1 - 1/n,1)`, and randomly permuted.

`X = lhsdesign(n,p,Name,Value)` modifies the resulting design using one or more name-value pair arguments. For example, you can obtain a discrete design by specifying `'Smooth','off'`.

Examples

Create Latin Hypercube Sample

Create a Latin hypercube sample of 10 rows and 4 columns.

```
rng default % For reproducibility
X = lhsdesign(10,4)
```

```
X = 10×4

    0.1893    0.2569    0.0147    0.5583
    0.8038    0.1089    0.9378    0.1950
    0.5995    0.6818    0.3649    0.3097
    0.3225    0.8736    0.4487    0.8055
    0.9183    0.9854    0.1598    0.2509
    0.0131    0.3864    0.5924    0.7511
    0.7916    0.7131    0.2760    0.6662
    0.6600    0.5420    0.6877    0.9100
    0.2740    0.0450    0.7816    0.0631
    0.4200    0.4855    0.8760    0.4889
```

Each column of `X` contains one random number in each interval `[0,0.1]`, `[0.1,0.2]`, `[0.2,0.3]`, `[0.3,0.4]`, `[0.4,0.5]`, `[0.5,0.6]`, `[0.6,0.7]`, `[0.7,0.8]`, `[0.8,0.9]`, and `[0.9,1]`.

Latin Hypercube Design with Nondefault Options

Determine the effects of various name-value pair arguments in `lhsdesign`. Start with a default design for 10 rows and four columns.

```
rng default % For reproducibility
X = lhsdesign(10,4)
```

```
X = 10×4

    0.1893    0.2569    0.0147    0.5583
    0.8038    0.1089    0.9378    0.1950
    0.5995    0.6818    0.3649    0.3097
    0.3225    0.8736    0.4487    0.8055
    0.9183    0.9854    0.1598    0.2509
    0.0131    0.3864    0.5924    0.7511
    0.7916    0.7131    0.2760    0.6662
    0.6600    0.5420    0.6877    0.9100
    0.2740    0.0450    0.7816    0.0631
    0.4200    0.4855    0.8760    0.4889
```

To obtain a discrete design. as opposed to a continuous design, set the `'Smooth'` name-value pair argument to `'off'`.

```
rng default % For reproducibility
X = lhsdesign(10,4,'Smooth','off')
```

```
X = 10×4

    0.2500    0.3500    0.7500    0.8500
    0.1500    0.8500    0.2500    0.3500
    0.8500    0.7500    0.4500    0.7500
    0.9500    0.1500    0.6500    0.1500
    0.0500    0.0500    0.8500    0.9500
    0.4500    0.5500    0.9500    0.4500
    0.3500    0.9500    0.5500    0.0500
    0.5500    0.4500    0.0500    0.2500
    0.6500    0.6500    0.1500    0.6500
    0.7500    0.2500    0.3500    0.5500
```

The resulting design is discrete.

Calculate the sum of squares of the between-column correlations of the returned design.

```
y = corr(X);
(sum(y(:).^2) - 4)/2 % Subtract 4 to remove the diagonal terms of corr(X)
```

```
ans = 0.4874
```

Observe the effect of changing the `'Criterion'` name-value pair argument to `'correlation'`, which minimizes the sum of between-column squared correlations. The `'correlation'` criterion always gives a discrete design, as if `'Smooth'` is set to `'off'`.

```
rng default % For reproducibility
X = lhsdesign(10,4,'Criterion','correlation')
```

```
X = 10×4

    0.6500    0.0500    0.4500    0.7500
    0.2500    0.3500    0.0500    0.1500
    0.1500    0.9500    0.8500    0.4500
    0.8500    0.5500    0.9500    0.0500
```

```
    0.5500    0.2500    0.5500    0.3500
    0.3500    0.4500    0.7500    0.8500
    0.4500    0.1500    0.6500    0.6500
    0.0500    0.6500    0.2500    0.5500
    0.9500    0.8500    0.3500    0.9500
    0.7500    0.7500    0.1500    0.2500
```

```
y = corr(X);
(sum(y(:).^2) - 4)/2
```

```
ans = 0.0102
```

Minimizing the correlations results in a design with much lower sum of squared correlations.

Specify fewer iterations to improve the criterion.

```
rng default % For reproducibility
X = lhsdesign(10,4,'Criterion','correlation','Iterations',2)
```

```
X = 10×4
```

```
    0.6500    0.0500    0.4500    0.7500
    0.3500    0.3500    0.0500    0.1500
    0.1500    0.9500    0.8500    0.4500
    0.9500    0.5500    0.9500    0.0500
    0.5500    0.2500    0.5500    0.3500
    0.2500    0.4500    0.7500    0.8500
    0.4500    0.1500    0.6500    0.6500
    0.0500    0.6500    0.2500    0.5500
    0.8500    0.8500    0.3500    0.9500
    0.7500    0.7500    0.1500    0.2500
```

```
y = corr(X);
(sum(y(:).^2) - 4)/2
```

```
ans = 0.0328
```

Lowering the number of iterations results in a worse design (higher sum of squared correlations).

Input Arguments

n — Number of returned samples

positive integer

Number of returned samples, specified as a positive integer.

Example: 24

Data Types: `single` | `double`

p — Number of returned variables

positive integer

Number of returned variables, specified as a positive integer.

Example: 4

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `X = lhsdesign(n,p,'Smooth','off')` returns a discrete Latin hypercube design

`Smooth` — Indication for continuous samples

`'on'` (default) | `'off'`

Indication for continuous samples, specified as the comma-separated pair consisting of `'Smooth'` and `'on'` (continuous samples) or `'off'` (discrete samples). When this option is `'off'`, the returned values in each column of `X` are a random permutation of the values `0.5/n`, `1.5/n`, ..., `1 − 0.5/n`.

Example: `'Smooth','off'`

Data Types: `char` | `string`

`Criterion` — Criterion for iterative sample generation

`'maximin'` (default) | `'none'` | `'correlation'`

Criterion for iterative sample generation, specified as the comma-separated pair consisting of `'Criterion'` and `'maximin'`, `'none'`, or `'correlation'`. The algorithm uses up to `Iterations` tries to improve the criterion.

Note The `'correlation'` criterion gives discrete samples, as if `Smooth` is set to `'off'`.

Criterion	Description
`'maximin'`	Maximize the minimum distance between points.
`'correlation'`	Minimize the sum of between-column squared correlations.
`'none'`	No iteration

Example: `'Criterion','correlation'`

Data Types: `char` | `string`

`Iterations` — Maximum number of iterations to improve criterion

5 (default) | positive integer

Maximum number of iterations to improve `Criterion`, specified as the comma-separated pair consisting of `'Iterations'` and a positive integer. The algorithm uses up to `Iterations` tries to improve the criterion.

Example: `'Iterations',10`

Data Types: `single` | `double`

See Also

`haltonset` | `lhsnorm` | `sobolset` | `unifrnd`

Introduced before R2006a

lhsnorm

Latin hypercube sample from normal distribution

Syntax

```
X = lhsnorm(mu,sigma,n)
X = lhsnorm(mu,sigma,n,flag)
[X,Z] = lhsnorm(...)
```

Description

`X = lhsnorm(mu,sigma,n)` returns an *n*-by-*p* matrix, `X`, containing a Latin hypercube sample of size `n` from a *p*-dimensional multivariate normal distribution with mean vector, `mu`, and covariance matrix, `sigma`.

`X` is similar to a random sample from the multivariate normal distribution, but the marginal distribution of each column is adjusted so that its sample marginal distribution is close to its theoretical normal distribution.

`X = lhsnorm(mu,sigma,n,flag)` controls the amount of smoothing in the sample. If *`flag`* is `'off'`, each column has points equally spaced on the probability scale. In other words, each column is a permutation of the values `G(0.5/n), G(1.5/n), ..., G(1-0.5/n)`, where `G` is the inverse normal cumulative distribution for that column's marginal distribution. If *`flag`* is `'on'` (the default), each column has points uniformly distributed on the probability scale. For example, in place of `0.5/n` you use a value having a uniform distribution on the interval `(0/n,1/n)`.

`[X,Z] = lhsnorm(...)` also returns `Z`, the original multivariate normal sample before the marginals are adjusted to obtain `X`.

References

[1] Stein, M. “Large sample properties of simulations using latin hypercube sampling.” *Technometrics*. Vol. 29, No. 2, 1987, pp. 143–151. Correction, Vol. 32, p. 367.

See Also

`lhsdesign` | `mvnrnd`

Introduced before R2006a

lillietest

Lilliefors test

Syntax

```
h = lillietest(x)
h = lillietest(x,Name,Value)
[h,p] = lillietest( ___ )
[h,p,kstat,critval] = lillietest( ___ )
```

Description

`h = lillietest(x)` returns a test decision for the null hypothesis that the data in vector `x` comes from a distribution in the normal family, against the alternative that it does not come from such a distribution, using a Lilliefors test. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = lillietest(x,Name,Value)` returns a test decision with additional options specified by one or more name-value pair arguments. For example, you can test the data against a different distribution family, change the significance level, or calculate the *p*-value using a Monte Carlo approximation.

`[h,p] = lillietest( ___ )` also returns the *p*-value `p`, using any of the input arguments from the previous syntaxes.

`[h,p,kstat,critval] = lillietest( ___ )` also returns the test statistic `kstat` and the critical value `critval` for the test.

Examples

Test for Normal Distribution

Load the sample data. Test the null hypothesis that car mileage, in miles per gallon (`MPG`), follows a normal distribution across different makes of cars.

```
load carbig
[h,p,k,c] = lillietest(MPG)
```

```
Warning: P is less than the smallest tabulated value, returning 0.001.
```

```
h = 1
```

```
p = 1.0000e-03
```

```
k = 0.0789
```

```
c = 0.0451
```

The test statistic `k` is greater than the critical value `c`, so `lillietest` returns a result of `h = 1` to indicate rejection of the null hypothesis at the default 5% significance level. The warning indicates

that the returned *p*-value is less than the smallest value in the table of precomputed values. To find a more accurate *p*-value, use `MCTol` to run a Monte Carlo approximation. See “Determine the p-value Using Monte Carlo Approximation” on page 33-3480.

Test the Hypothesis at Different Significance Levels

Load the sample data. Create a vector containing the first column of the students’ exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the sample data comes from a normal distribution at the 1% significance level.

```
[h,p] = lillietest(x,'Alpha',0.01)
```

```
h = 0
```

```
p = 0.0348
```

The returned value of `h = 0` indicates that `lillietest` does not reject the null hypothesis at the 1% significance level.

Test for Exponential Distribution

Load the sample data. Test the null hypothesis that car mileage, in miles per gallon (`MPG`), follows an exponential distribution across different makes of cars.

```
load carbig
h = lillietest(MPG,'Distribution','exponential')
```

```
h = 1
```

The returned value of `h = 1` indicates that `lillietest` rejects the null hypothesis at the default 5% significance level.

Test for Weibull Distribution

Generate two sample data sets, one from a Weibull distribution and another from a lognormal distribution. Perform the Lilliefors test to assess whether each data set is from a Weibull distribution. Confirm the test decision by performing a visual comparison using a Weibull probability plot (`wblplot`).

Generate samples from a Weibull distribution.

```
rng('default')
data1 = wblrnd(0.5,2,[500,1]);
```

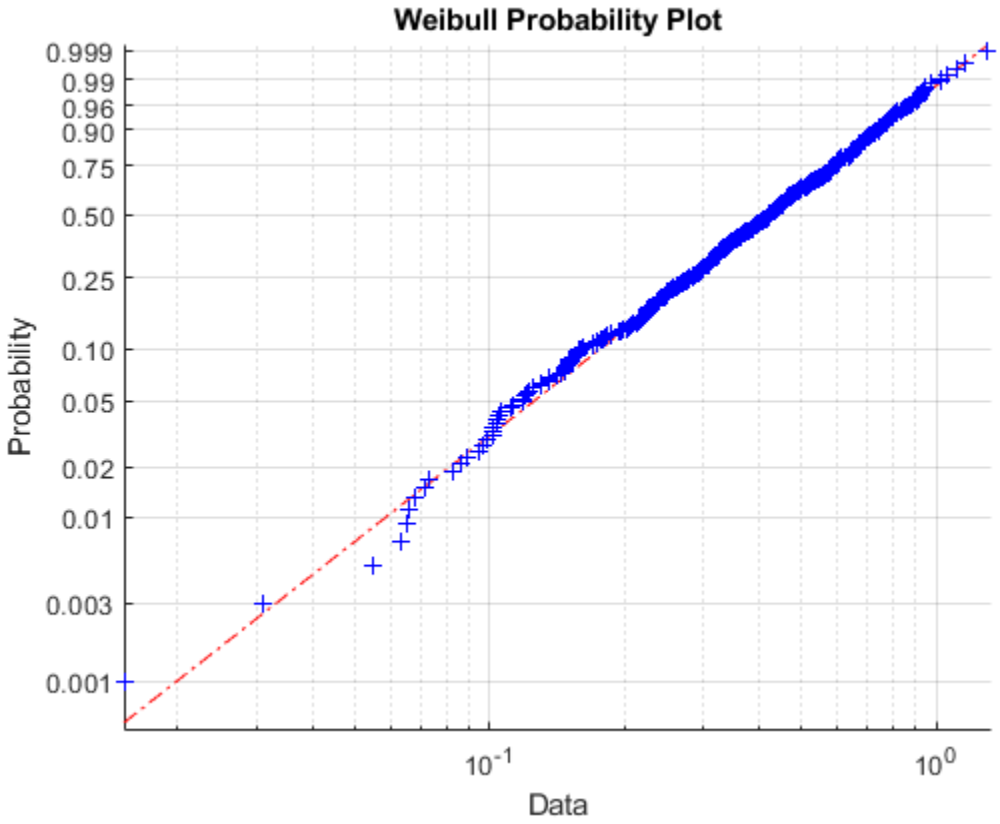
Perform the Lilliefors test by using the `lillietest`. To test data for a Weibull distribution, test if the logarithm of the data has an extreme value distribution.

```
h1 = lillietest(log(data1),'Distribution','extreme value')
```

```
h1 = 0
```

The returned value of `h1 = 0` indicates that `lillietest` fails to reject the null hypothesis at the default 5% significance level. Confirm the test decision using a Weibull probability plot.

```
wblplot(data1)
```



The plot indicates that the data follows a Weibull distribution.

Generate samples from a lognormal distribution.

```
data2 =lognrnd(5,2,[500,1]);
```

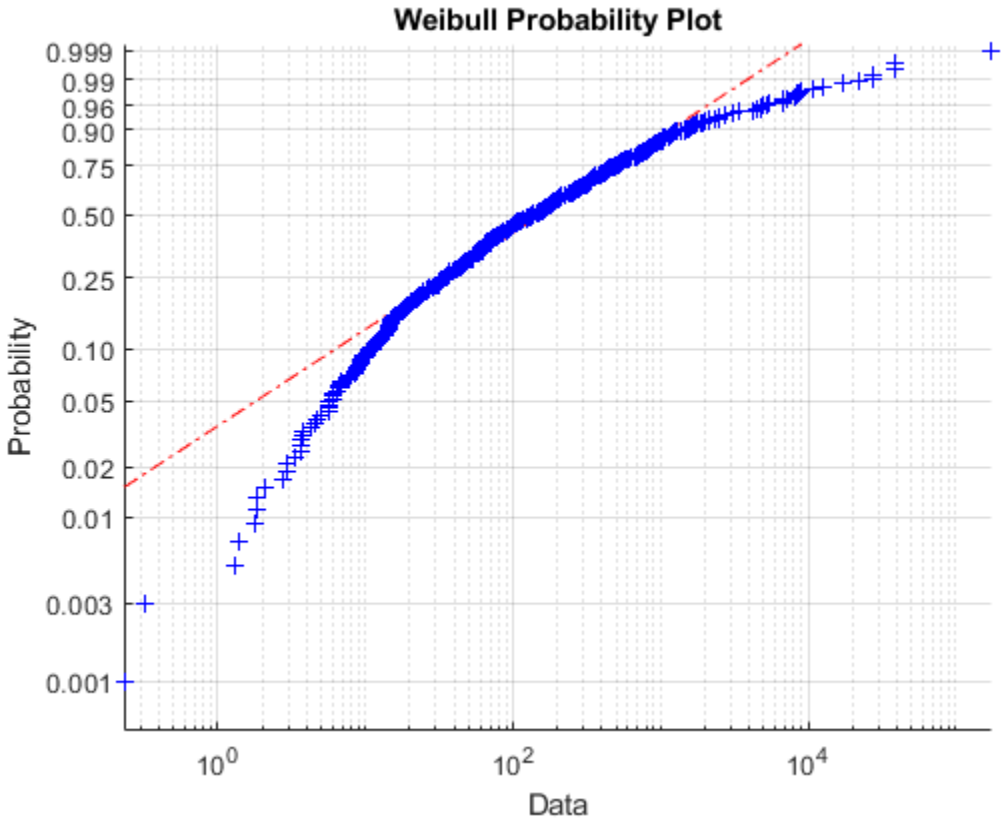
Perform the Lilliefors test.

```
h2 = lillietest(log(data2),'Distribution','extreme value')
```

```
h2 = 1
```

The returned value of `h2 = 1` indicates that `lillietest` rejects the null hypothesis at the default 5% significance level. Confirm the test decision using a Weibull probability plot.

```
wblplot(data2)
```



The plot indicates that the data does not follow a Weibull distribution.

Determine the p-value Using Monte Carlo Approximation

Load the sample data. Test the null hypothesis that car mileage, in miles per gallon (`MPG`), follows a normal distribution across different makes of cars. Determine the *p*-value using a Monte Carlo approximation with a maximum Monte Carlo standard error of `1e-4`.

```
load carbig
[h,p] = lillietest(MPG,'MCTol',1e-4)
```

```
h = 1
```

```
p = 8.3333e-06
```

The returned value of `h = 1` indicates that `lillietest` rejects the null hypothesis that the data comes from a normal distribution at the 5% significance level.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distribution','exponential','Alpha',0.01` tests the null hypothesis that the population distribution belongs to the exponential distribution family at the 1% significance level.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

- If `MCTol` is not used, `Alpha` must be in the range [0.001,0.50].
- If `MCTol` is used, `Alpha` must be in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Distribution` — Distribution family

`'normal'` (default) | `'exponential'` | `'extreme value'`

Distribution family for the hypothesis test, specified as the comma-separated pair consisting of `'Distr'` and one of the following.

`'normal'`	Normal distribution
`'exponential'`	Exponential distribution
`'extreme value'`	Extreme value distribution

- To test `x` for a lognormal distribution, test if `log(x)` has a normal distribution.
- To test `x` for a Weibull distribution, test if `log(x)` has an extreme value distribution.

Example: `'Distribution','exponential'`

`MCTol` — Maximum Monte Carlo standard error

scalar value in the range (0,1)

Maximum Monte Carlo standard error on page 33-3483 for `p`, the *p*-value of the test, specified as the comma-separated pair consisting of `'MCTol'` and a scalar value in the range (0,1).

Example: `'MCTol',0.001`

Data Types: `single` | `double`

Output Arguments

h — Hypothesis test result

1 | 0

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

p — *p*-value

scalar value in the range (0,1)

p-value of the test, returned as a scalar value in the range (0,1). `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

- If `MCTol` is not used, `p` is computed using inverse interpolation into the table of critical values, and is returned as a scalar value in the range [0.001,0.50]. `lillietest` warns when `p` is not found within the tabulated range and returns either the smallest or largest tabulated value.
- If `MCTol` is used, `lillietest` conducts a Monte Carlo simulation to compute a more accurate *p*-value, and `p` is returned as a scalar value in the range (0,1).

kstat — Test statistic

nonnegative scalar value

Test statistic, returned as a nonnegative scalar value.

critval — Critical value

nonnegative scalar value

Critical value for the hypothesis test, returned as a nonnegative scalar value.

More About

Lilliefors Test

The Lilliefors test is a two-sided goodness-of-fit test suitable when the parameters of the null distribution are unknown and must be estimated. This is in contrast to the one-sample Kolmogorov-Smirnov test, which requires the null distribution to be completely specified.

The Lilliefors test statistic is:

$$D^* = \max_x \left| \widehat{F}(x) - G(x) \right|,$$

where $\widehat{F}(x)$ is the empirical cdf of the sample data and $G(x)$ is the cdf of the hypothesized distribution with estimated parameters equal to the sample parameters.

`lillietest` can be used to test whether the data vector `x` has a lognormal or Weibull distribution by applying a transformation to the data vector and running the appropriate Lilliefors test:

- To test `x` for a lognormal distribution, test if `log(x)` has a normal distribution.

- To test `x` for a Weibull distribution, test if `log(x)` has an extreme value distribution.

The Lilliefors test cannot be used when the null hypothesis is not a location-scale family of distributions.

Monte Carlo Standard Error

The Monte Carlo standard error is the error due to simulating the *p*-value.

The Monte Carlo standard error is calculated as:

$$SE = \sqrt{\frac{(\widehat{p})(1-\widehat{p})}{\text{mcreps}}},$$

where $\widehat{p}$ is the estimated *p*-value of the hypothesis test, and `mcreps` is the number of Monte Carlo replications performed.

The number of Monte Carlo replications, `mcreps`, is determined such that the Monte Carlo standard error for $\widehat{p}$ less than the value specified for `MCTol`.

Algorithms

To compute the critical value for the hypothesis test, `lillietest` interpolates into a table of critical values pre-computed using Monte Carlo simulation for sample sizes less than 1000 and significance levels between 0.001 and 0.50. The table used by `lillietest` is larger and more accurate than the table originally introduced by Lilliefors. If a more accurate *p*-value is desired, or if the desired significance level is less than 0.001 or greater than 0.50, the `MCTol` input argument can be used to run a Monte Carlo simulation to calculate the *p*-value more exactly.

When the computed value of the test statistic is greater than the critical value, `lillietest` rejects the null hypothesis at significance level `Alpha`.

`lillietest` treats `NaN` values in `x` as missing values and ignores them.

References

[1] Conover, W. J. *Practical Nonparametric Statistics*. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

[2] Lilliefors, H. W. “On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown.” *Journal of the American Statistical Association*. Vol. 64, 1969, pp. 387–389.

[3] Lilliefors, H. W. “On the Kolmogorov-Smirnov test for normality with mean and variance unknown.” *Journal of the American Statistical Association*. Vol. 62, 1967, pp. 399–402.

See Also

`adtest` | `cdfplot` | `jbtest` | `kstest` | `kstest2`

Introduced before R2006a

lime

Local interpretable model-agnostic explanations (LIME)

Description

"LIME" on page 33-3505 explains a prediction of a machine learning model (classification or regression) for a query point by finding important predictors and fitting a simple interpretable model.

You can create a `lime` object for a machine learning model with a specified query point (`queryPoint`) and a specified number of important predictors (`numImportantPredictors`). The software generates a synthetic data set, and fits a simple interpretable model of important predictors that effectively explains the predictions for the synthetic data around the query point. The simple model can be a linear model (default) or decision tree model.

Use the fitted simple model to explain a prediction of the machine learning model locally, at the specified query point. Use the `plot` function to visualize the LIME results. Based on the local explanations, you can decide whether or not to trust the machine learning model.

Fit a new simple model for another query point by using the `fit` function.

Creation

Syntax

```
results = lime(blackbox)
results = lime(blackbox,X)
results = lime(blackbox,'CustomSyntheticData',customSyntheticData)

results = lime( ___ ,'QueryPoint',queryPoint,'NumImportantPredictors',
numImportantPredictors)

results = lime( ___ ,Name,Value)
```

Description

`results = lime(blackbox)` creates a `lime` object using a machine learning model object `blackbox` that contains predictor data. The `lime` function generates samples of a synthetic predictor data set and computes the predictions for the samples. To fit a simple model, use the `fit` function with `results`.

`results = lime(blackbox,X)` creates a `lime` object using the predictor data in `X`.

`results = lime(blackbox,'CustomSyntheticData',customSyntheticData)` creates a `lime` object using the pregenerated, custom synthetic predictor data set `customSyntheticData`. The `lime` function computes the predictions for the samples in `customSyntheticData`.

`results = lime( ___ ,'QueryPoint',queryPoint,'NumImportantPredictors', numImportantPredictors)` also finds the specified number of important predictors and fits a

linear simple model for the query point `queryPoint`. You can specify `queryPoint` and `numImportantPredictors` in addition to any of the input argument combinations in the previous syntaxes.

`results = lime( ___ ,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'SimpleModelType','tree'` specifies the type of simple model as a decision tree model.

Input Arguments

`blackbox` — Machine learning model to be interpreted

regression model object | classification model object | function handle

Machine learning model to be interpreted, specified as a full or compact regression or classification model object or a function handle.

- Full or compact model object — You can specify a full or compact regression or classification model object, which has a `predict` object function. The software uses the `predict` function to compute the predictions for the query point and the synthetic predictor data set.
  - If you specify a model object that does not contain predictor data (for example, a compact model), then you must provide predictor data using `X` or `customSyntheticData`.
  - `lime` does not support a model object trained with a sparse matrix. When you train a model, use a full numeric matrix or table for the predictor data where rows correspond to individual observations.

  Regression Model Object

Supported Model	Full or Compact Regression Model Object
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`
Generalized additive model	`RegressionGAM`, `CompactRegressionGAM`
Linear regression for high-dimensional data	`RegressionLinear`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`
Regression tree	`RegressionTree`, `CompactRegressionTree`
Support vector machine regression	`RegressionSVM`, `CompactRegressionSVM`

Classification Model Object

Supported Model	Full or Compact Classification Model Object
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`
Generalized additive model	`ClassificationGAM`, `CompactClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Linear classification model	`ClassificationLinear`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`
Naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`
Support vector machine for binary classification	`ClassificationSVM`, `CompactClassificationSVM`

- Function handle — You can specify a function handle that accepts predictor data and returns a column vector containing a prediction for each observation in the predictor data. The prediction is a predicted response for regression or a classified label for classification. You must provide the predictor data using `X` or `customSyntheticData` and specify the `'Type'` name-value argument.

`X` — Predictor data
numeric matrix | table

Predictor data, specified as a numeric matrix or table. Each row of `X` corresponds to one observation, and each column corresponds to one variable.

`X` must be consistent with the predictor data that trained `blackbox`, stored in `blackbox.X`. The specified value must not contain a response variable.

- `X` must have the same data types as the predictor variables (for example, `trainX`) that trained `blackbox`. The variables that make up the columns of `X` must have the same number and order as in `trainX`.
  - If you train `blackbox` using a numeric matrix, then `X` must be a numeric matrix.
  - If you train `blackbox` using a table, then `X` must be a table. All predictor variables in `X` must have the same variable names and data types as in `trainX`.
- `lime` does not support a sparse matrix.

If `blackbox` is a model object that does not contain predictor data or a function handle, you must provide `X` or `customSyntheticData`. If `blackbox` is a full machine learning model object and you specify this argument, then `lime` does not use the predictor data in `blackbox`. It uses the specified predictor data only.

Data Types: `single` | `double` | `table`

`customSyntheticData` — Pregenerated, custom synthetic predictor data set
`[]` (default) | numeric matrix | table

Pregenerated, custom synthetic predictor data set, specified as a numeric matrix or table.

If you provide a pregenerated data set, then `lime` uses the provided data set instead of generating a new synthetic predictor data set.

`customSyntheticData` must be consistent with the predictor data that trained `blackbox`, stored in `blackbox.X`. The specified value must not contain a response variable.

- `customSyntheticData` must have the same data types as the predictor variables (for example, `trainX`) that trained `blackbox`. The variables that make up the columns of `customSyntheticData` must have the same number and order as in `trainX`
  - If you train `blackbox` using a numeric matrix, then `customSyntheticData` must be a numeric matrix.
  - If you train `blackbox` using a table, then `customSyntheticData` must be a table. All predictor variables in `customSyntheticData` must have the same variable names and data types as in `trainX`.
- `lime` does not support a sparse matrix.

If `blackbox` is a model object that does not contain predictor data or a function handle, you must provide `X` or `customSyntheticData`. If `blackbox` is a full machine learning model object and you specify this argument, then `lime` does not use the predictor data in `blackbox`; it uses the specified predictor data only.

Data Types: `single` | `double` | `table`

`queryPoint` — Query point
row vector of numeric values | single-row table

Query point at which `lime` explains a prediction, specified as a row vector of numeric values or a single-row table. `queryPoint` must have the same data type and number of columns as `X`, `customSyntheticData`, or the predictor data in `blackbox`.

If you specify `numImportantPredictors` and `queryPoint`, then the `lime` function fits a simple model when creating a `lime` object.

`queryPoint` must not contain missing values.

Example: `blackbox.X(1,:)` specifies the query point as the first observation of the predictor data in the full machine learning model `blackbox`.

Data Types: `single` | `double` | `table`

`numImportantPredictors` — Number of important predictors to use in simple model
positive integer scalar value

Number of important predictors to use in the simple model, specified as a positive integer scalar value.

- If `'SimpleModelType'` is `'linear'`, then the software selects the specified number of important predictors and fits a linear model of the selected predictors.
- If `'SimpleModelType'` is `'tree'`, then the software specifies the maximum number of decision splits (or branch nodes) as the number of important predictors so that the fitted decision tree uses at most the specified number of predictors.

If you specify `numImportantPredictors` and `queryPoint`, then the `lime` function fits a simple model when creating a `lime` object.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `lime(blackbox,'QueryPoint',q,'NumImportantPredictors',n,'SimpleModelType','tree')` specifies the query point as `q`, the number of important predictors to use for the simple model as `n`, and the type of simple model as a decision tree model. `lime` generates samples of a synthetic predictor data set, computes the predictions for the samples, and fits a decision tree model for the query point using at most the specified number of predictors.

Options for Synthetic Predictor Data

`DataLocality` — Locality of synthetic data for data generation
`'global'` (default) | `'local'`

Locality of the synthetic data for data generation, specified as the comma-separated pair consisting of `'DataLocality'` and `'global'` or `'local'`.

- `'global'` — The software estimates distribution parameters using the whole predictor data set (`X` or the predictor data in `blackbox`). The software generates a synthetic predictor data set with the estimated parameters and uses the data set for simple model fitting of any query point.
- `'local'` — The software estimates the distribution parameters using the *k*-nearest neighbors of a query point, where *k* is the `'NumNeighbors'` value. The software generates a new synthetic predictor data set each time it fits a simple model for the specified query point.

For more details, see “LIME” on page 33-3505.

Example: `'DataLocality','local'`

Data Types: `char` | `string`

`NumNeighbors` — Number of neighbors of query point
1500 (default) | positive integer scalar value

Number of neighbors of the query point, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer scalar value. This argument is valid only when `'DataLocality'` is `'local'`.

If you specify a value larger than the number of observations in the predictor data set (`X` or the predictor data in `blackbox`), then `lime` uses all observations.

Example: `'NumNeighbors',2000`

Data Types: `single` | `double`

`NumSyntheticData` — Number of samples to generate for synthetic data set

5000 (default) | positive integer scalar value

Number of samples to generate for the synthetic data set, specified as the comma-separated pair consisting of `'NumSyntheticData'` and a positive integer scalar value. This argument is valid only when `'DataLocality'` is `'local'`.

Example: `'NumSyntheticData',2500`

Data Types: `single` | `double`

Options for Simple Model

`KernelWidth` — Kernel width

0.75 (default) | numeric scalar value

Kernel width of the squared exponential (or Gaussian) kernel function, specified as the comma-separated pair consisting of `'KernelWidth'` and a numeric scalar value.

The `lime` function computes distances between the query point and the samples in the synthetic predictor data set, and then converts the distances to weights by using the squared exponential kernel function. If you lower the `'KernelWidth'` value, then `lime` uses weights that are more focused on the samples near the query point. For details, see “LIME” on page 33-3505.

Example: `'KernelWidth',0.5`

Data Types: `single` | `double`

`SimpleModelType` — Type of simple model

`'linear'` (default) | `'tree'`

Type of the simple model, specified as the comma-separated pair consisting of `'SimpleModelType'` and `'linear'` or `'tree'`.

- `'linear'` — The software fits a linear model by using `fitrlinear` for regression or `fitclinear` for classification.
- `'tree'` — The software fits a decision tree model by using `fitrtree` for regression or `fitctree` for classification.

Example: `'SimpleModelType','tree'`

Data Types: `char` | `string`

Options for Machine Learning Model

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as the comma-separated pair consisting of `'CategoricalPredictors'` and one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `blackbox` uses a subset of input variables as predictors, then the software indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the variable names of the predictor data in the form of a table. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the variable names of the predictor data in the form of a table.
`'all'`	All predictors are categorical.

- If you specify `blackbox` as a function handle, then `lime` identifies categorical predictors from the predictor data `X` or `customSyntheticData`. If the predictor data is in a table, `lime` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix, `lime` assumes that all predictors are continuous.
- If you specify `blackbox` as a regression or classification model object, then `lime` identifies categorical predictors by using the `CategoricalPredictors` property of the model object.

`lime` does not support an ordered categorical predictor.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

Type — Type of machine learning model
`'regression'` | `'classification'`

Type of the machine learning model, specified as the comma-separated pair consisting of `'Type'` and `'regression` or `'classification'`.

You must specify this argument when you specify `blackbox` as a function handle. If you specify `blackbox` as a regression or classification model object, then `lime` determines the `'Type'` value depending on the model type.

Example: `'Type','classification'`

Data Types: `char` | `string`

Options for Computing Distances

`Distance` — Distance metric
character vector | string scalar | function handle

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle.

- If the predictor data includes only continuous variables, then `lime` supports these distance metrics.

Value	Description
`'euclidean'`	Euclidean distance.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(PD,'omitnan')`, where `PD` is the predictor data or synthetic predictor data. To specify different scaling, use the `'Scale'` name-value argument.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `PD`, `C = cov(PD,'omitrows')`. To change the value of the covariance matrix, use the `'Cov'` name-value argument.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. To specify a different exponent, use the `'P'` name-value argument.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`t` vector containing a single observation.
	• `ZJ` is an `s`-by-`t` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `s`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance metric instead of a function handle.

- If the predictor data includes both continuous and categorical variables, then `lime` supports these distance metrics.

Value	Description
`'goodall3'`	Modified Goodall distance

Value	Description
'ofd'	Occurrence frequency distance

For definitions, see “Distance Metrics” on page 33-3503.

The default value is 'euclidean' if the predictor data includes only continuous variables, or 'goodall3' if the predictor data includes both continuous and categorical variables.

Example: 'Distance','ofd'

Data Types: char | string | function_handle

Cov — Covariance matrix for Mahalanobis distance metric
positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of 'Cov' and a *K*-by-*K* positive definite matrix, where *K* is the number of predictors.

This argument is valid only if 'Distance' is 'mahalanobis'.

The default 'Cov' value is cov(PD,'omitrows'), where PD is the predictor data or synthetic predictor data. If you do not specify the 'Cov' value, then the software uses different covariance matrices when computing the distances for both the predictor data and the synthetic predictor data.

Example: 'Cov',eye(3)

Data Types: single | double

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of 'P' and a positive scalar.

This argument is valid only if 'Distance' is 'minkowski'.

Example: 'P',3

Data Types: single | double

Scale — Scale parameter value for standardized Euclidean distance metric
nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of 'Scale' and a nonnegative numeric vector of length *K*, where *K* is the number of predictors.

This argument is valid only if 'Distance' is 'seuclidean'.

The default 'Scale' value is std(PD,'omitnan'), where PD is the predictor data or synthetic predictor data. If you do not specify the 'Scale' value, then the software uses different scale parameters when computing the distances for both the predictor data and the synthetic predictor data.

Example: 'Scale',quantile(X,0.75) - quantile(X,0.25)

Data Types: single | double

Properties

Specified Properties

You can specify the following properties when creating a `lime` object.

`BlackboxModel` — Machine learning model to be interpreted
regression model object | classification model object | function handle

This property is read-only.

Machine learning model to be interpreted, specified as a regression or classification model object or a function handle.

The `blackbox` argument sets this property.

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

- If you specify `blackbox` using a function handle, then `lime` identifies categorical predictors from the predictor data `X` or `customSyntheticData`. If you specify the `'CategoricalPredictors'` name-value argument, then the argument sets this property.
- If you specify `blackbox` as a regression or classification model object, then `lime` determines this property by using the `CategoricalPredictors` property of the model object.

`lime` does not support an ordered categorical predictor.

If `'SimpleModelType'` is `'linear'`(default), then `lime` creates dummy variables for each identified categorical predictor. `lime` treats the category of the specified query point as a reference group and creates one less dummy variable than the number of categories. For more details, see “Dummy Variables with Reference Group” on page 2-48.

Data Types: `single` | `double`

`DataLocality` — Locality of synthetic data for data generation
`'global'` | `'local'`

This property is read-only.

Locality of the synthetic data for data generation, specified as `'global'` or `'local'`.

The `'DataLocality'` name-value argument sets this property.

`NumImportantPredictors` — Number of important predictors to use in simple model
positive integer scalar value

This property is read-only.

Number of important predictors to use in the simple model (`SimpleModel`), specified as a positive integer scalar value.

The `numImportantPredictors` argument of `lime` or the `numImportantPredictors` argument of `fit` sets this property.

Data Types: `single` | `double`

`NumSyntheticData` — Number of samples in synthetic data set
positive integer scalar value

This property is read-only.

Number of samples in the synthetic data set, specified as a positive integer scalar value.

- If you specify `customSyntheticData`, then the number of samples in the custom synthetic data set sets this property.
- Otherwise, the `'NumSyntheticData'` name-value argument of `lime` or the `'NumSyntheticData'` name-value argument of `fit` sets this property.

Data Types: `single` | `double`

`QueryPoint` — Query point
row vector of numeric values | single-row table

This property is read-only.

Query point at which `lime` explains a prediction using the simple model (`SimpleModel`), specified as a row vector of numeric values or single-row table.

The `queryPoint` argument of `lime` or the `queryPoint` argument of `fit` sets this property.

Data Types: `single` | `double` | `table`

`Type` — Type of machine learning model
`'regression'` | `'classification'`

This property is read-only.

Type of the machine learning model (`BlackboxModel`), specified as `'regression` or `'classification'`.

- If you specify `blackbox` as a regression or classification model object, then `lime` determines this property depending on the model type.
- If you specify `blackbox` using a function handle, then the `'Type'` name-value argument sets this property.

`X` — Predictor data
numeric matrix | table

This property is read-only.

Predictor data, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- If you specify the `X` argument, then the argument sets this property.
- If you specify the `customSyntheticData` argument, then this property is empty.

- If you specify `blackbox` as a full machine learning model object and do not specify `X` or `customSyntheticData`, then this property value is the predictor data used to train `blackbox`.

`lime` does not use rows that contain missing values and does not store the rows in `X`.

Data Types: `single` | `double` | `table`

Computed Properties

The software computes the following properties.

`BlackboxFitted` — Prediction for query point computed by machine learning model
scalar

This property is read-only.

Prediction for the query point computed by the machine learning model (`BlackboxModel`), specified as a scalar. The prediction is a predicted response for regression or a classified label for classification.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`Fitted` — Predictions for synthetic predictor data computed by machine learning model
vector

This property is read-only.

Predictions for synthetic predictor data computed by the machine learning model (`BlackboxModel`), specified as a vector.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`ImportantPredictors` — Important predictor indices
vector of positive integers

This property is read-only.

Important predictor indices, specified as a vector of positive integers. `ImportantPredictors` contains the index values corresponding to the columns of the predictors used in the simple model (`SimpleModel`).

Data Types: `single` | `double`

`SimpleModel` — Simple model
`RegressionLinear` model object | `RegressionTree` model object | `ClassificationLinear` model object | `ClassificationTree` model object

This property is read-only.

Simple model, specified as a `RegressionLinear`, `RegressionTree`, `ClassificationLinear`, or `ClassificationTree` model object. `lime` determines the type of simple model object depending on the type of the machine learning model (`Type`) and the type of the simple model (`'SimpleModelType'`).

`SimpleModelFitted` — Prediction for query point computed by simple model
scalar

This property is read-only.

Prediction for the query point computed by the simple model (`SimpleModel`), specified as a scalar.

If `SimpleModel` is `ClassificationLinear`, then the `SimpleModelFitted` value is 1 or –1.

- The `SimpleModelFitted` value is 1 if the prediction from the simple model is the same as `BlackboxFitted` (prediction from the machine learning model).
- The `SimpleModelFitted` value is –1 if the prediction from the simple model is different from `BlackboxFitted`. If the `BlackboxFitted` value is `A`, then the `plot` function displays the `SimpleModelFitted` value as `Not A`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`SyntheticData` — Synthetic predictor data
numeric matrix | table

This property is read-only.

Synthetic predictor data, specified as a numeric matrix or a table.

- If you specify the `customSyntheticData` input argument, then the argument sets this property.
- Otherwise, `lime` estimates distribution parameters from the predictor data `X` and generates a synthetic predictor data set.

Data Types: `single` | `double` | `table`

Object Functions

fit Fit simple model of local interpretable model-agnostic explanations (LIME)
plot Plot results of local interpretable model-agnostic explanations (LIME)

Examples

Explain Prediction with Decision Tree Simple Model

Train a classification model and create a `lime` object that uses a decision tree simple model. When you create a `lime` object, specify a query point and the number of important predictors so that the software generates samples of a synthetic data set and fits a simple model for the query point with important predictors. Then display the estimated predictor importance in the simple model by using the object function `plot`.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    _________    ______
```

```
    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

Create a table of predictor variables by removing the columns of customer IDs and ratings from `tbl`.

```
tblX = removevars(tbl,["ID","Rating"]);
```

Train a blackbox model of credit ratings by using the `fitcecoc` function.

```
blackbox = fitcecoc(tblX,tbl.Rating,'CategoricalPredictors','Industry');
```

Create a `lime` object that explains the prediction for the last observation using a decision tree simple model. Specify `'NumImportantPredictors'` as six to find at most 6 important predictors. If you specify the `'QueryPoint'` and `'NumImportantPredictors'` values when you create a `lime` object, then the software generates samples of a synthetic data set and fits a simple interpretable model to the synthetic data set.

```
queryPoint = tblX(end,:)
```

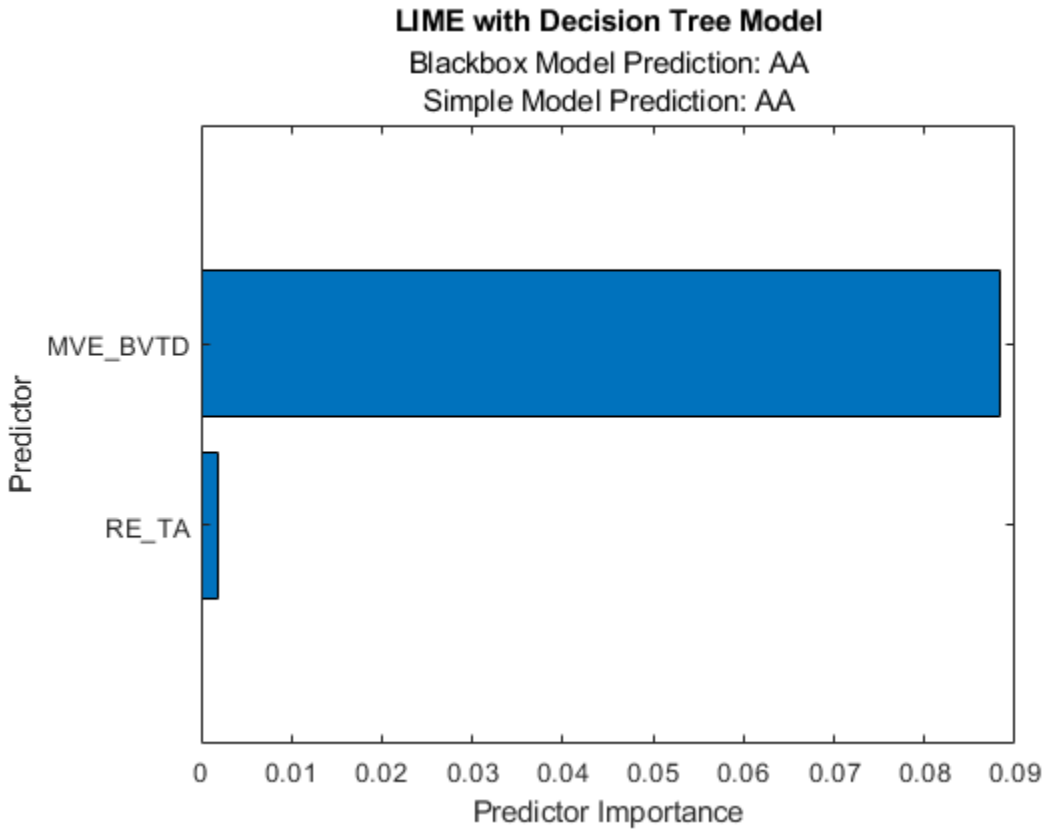
```
queryPoint=1×6 table
    WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA    Industry
    _____    _____    _______    ________    ____    ________

    0.239    0.463     0.065      2.924      0.34       2
```

```
rng('default') % For reproducibility
results = lime(blackbox,'QueryPoint',queryPoint,'NumImportantPredictors',6, ...
    'SimpleModelType','tree')
```

```
results =
  lime with properties:

             BlackboxModel: [1x1 ClassificationECOC]
              DataLocality: 'global'
     CategoricalPredictors: 6
                      Type: 'classification'
                         X: [3932x6 table]
                QueryPoint: [1x6 table]
    NumImportantPredictors: 6
          NumSyntheticData: 5000
             SyntheticData: [5000x6 table]
                    Fitted: {5000x1 cell}
               SimpleModel: [1x1 ClassificationTree]
       ImportantPredictors: [2x1 double]
            BlackboxFitted: {'AA'}
         SimpleModelFitted: {'AA'}
```

Plot the `lime` object `results` by using the object function `plot`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```

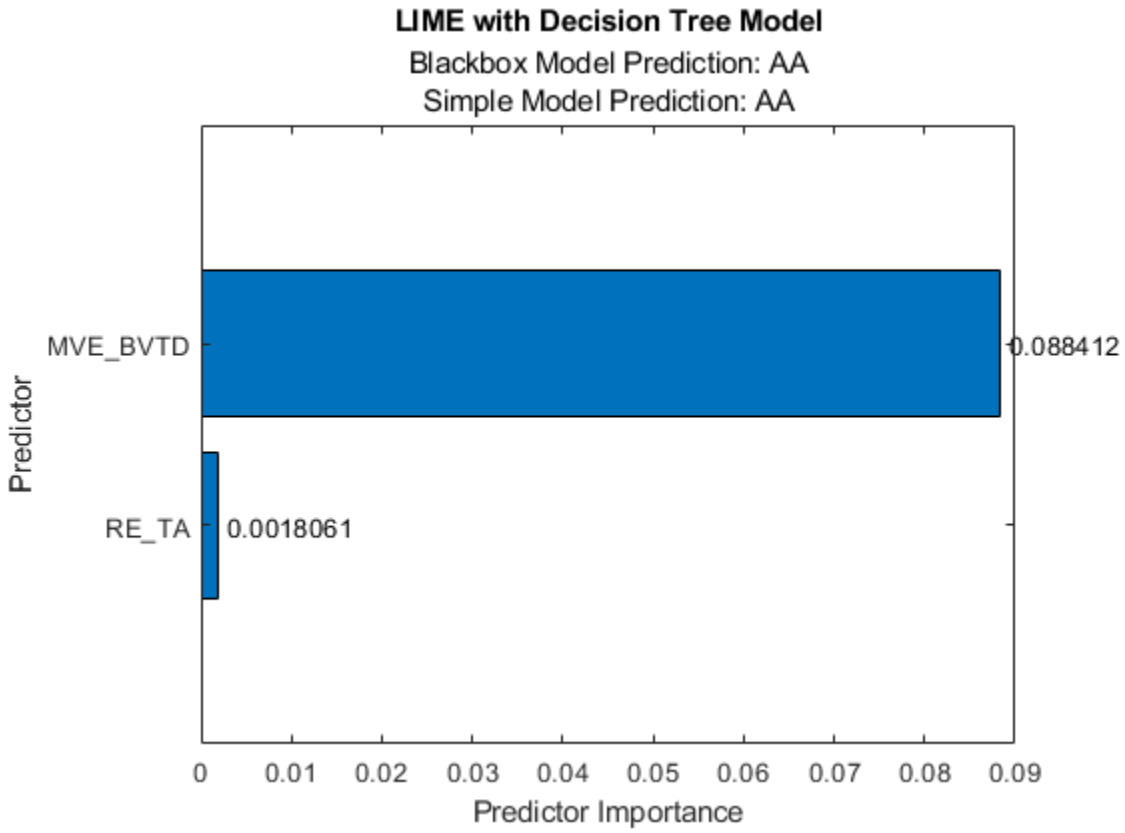


The plot displays two predictions for the query point, which correspond to the “BlackboxFitted” on page 33-0 property and the “SimpleModelFitted” on page 33-0 property of `results`.

The horizontal bar graph shows the sorted predictor importance values. `lime` finds the financial ratio variables `EBIT_TA` and `WC_TA` as important predictors for the query point.

You can read the bar lengths by using data tips or Bar Properties. For example, you can find `Bar` objects by using the `findobj` function and add labels to the ends of the bars by using the `text` function.

```
b = findobj(f,'Type','bar');
text(b.YEndPoints+0.001,b.XEndPoints,string(b.YData))
```



Alternatively, you can display the coefficient values in a table with the predictor variable names.

```
imp = b.YData;
flipud(array2table(imp', ...
    'RowNames',f.CurrentAxes.YTickLabel,'VariableNames',{'Predictor Importance'}))
```

```
ans=2×1 table
               Predictor Importance
               ____________________

    MVE_BVTD          0.088412
    RE_TA            0.0018061
```

Explain Prediction with Linear Simple Model

Train a regression model and create a `lime` object that uses a linear simple model. When you create a `lime` object, if you do not specify a query point and the number of important predictors, then the software generates samples of a synthetic data set but does not fit a simple model. Use the object function `fit` to fit a simple model for a query point. Then display the coefficients of the fitted linear simple model by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Create a table of predictor variables by removing the response variable from `tbl`.

```
tblX = removevars(tbl,'MPG');
```

Train a blackbox model of `MPG` by using the `fitrkernel` function.

```
rng('default') % For reproducibility
mdl = fitrkernel(tblX,tbl.MPG,'CategoricalPredictors',[2 5]);
```

Create a `lime` object. Specify a predictor data set because `mdl` does not contain predictor data.

```
results = lime(mdl,tblX)
```

```
results =
  lime with properties:

             BlackboxModel: [1x1 RegressionKernel]
              DataLocality: 'global'
     CategoricalPredictors: [2 5]
                      Type: 'regression'
                         X: [392x6 table]
                QueryPoint: []
    NumImportantPredictors: []
          NumSyntheticData: 5000
             SyntheticData: [5000x6 table]
                    Fitted: [5000x1 double]
               SimpleModel: []
       ImportantPredictors: []
            BlackboxFitted: []
          SimpleModelFitted: []
```

`results` contains the generated synthetic data set. The `SimpleModel` property is empty (`[]`).

Fit a linear simple model for the first observation in `tblX`. Specify the number of important predictors to find as 3.

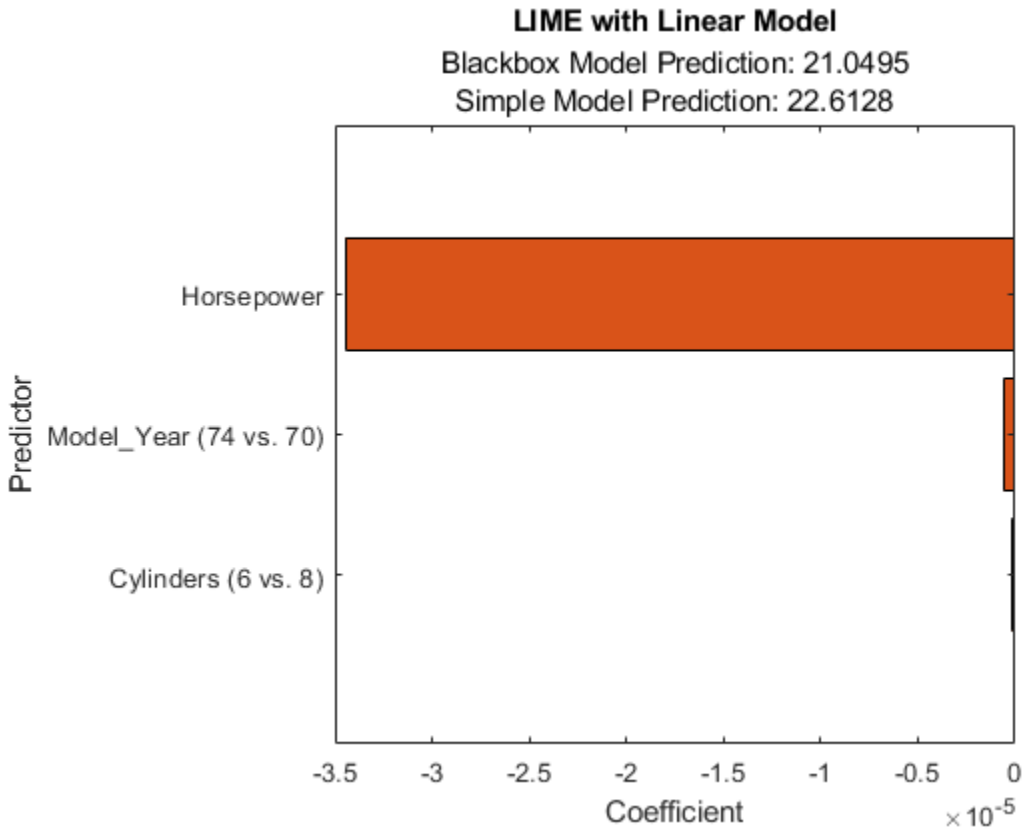
```
queryPoint = tblX(1,:)
```

```
queryPoint=1×6 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight
    ____________    _________    ____________    __________    __________    ______

         12             8            307            130            70         3504
```

```
results = fit(results,queryPoint,3);
```

Plot the `lime` object `results` by using the object function `plot`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The plot displays two predictions for the query point, which correspond to the “BlackboxFitted” on page 33-0 property and the “SimpleModelFitted” on page 33-0 property of `results`.

The horizontal bar graph shows the coefficient values of the simple model, sorted by their absolute values. LIME finds `Horsepower`, `Model_Year`, and `Cylinders` as important predictors for the query point.

`Model_Year` and `Cylinders` are categorical predictors that have multiple categories. For a linear simple model, the software creates one less dummy variable than the number of categories for each categorical predictor. The bar graph displays only the most important dummy variable. You can check the coefficients of the other dummy variables using the `SimpleModel` property of `results`. Display the sorted coefficient values, including all categorical dummy variables.

```
[~,I] = sort(abs(results.SimpleModel.Beta),'descend');
table(results.SimpleModel.ExpandedPredictorNames(I)',results.SimpleModel.Beta(I), ...
    'VariableNames',{'Exteded Predictor Name','Coefficient'})
```

```
ans=17×2 table
      Exteded Predictor Name      Coefficient
    ___________________________    ___________
```

```
    {'Horsepower'            }    -3.4485e-05
    {'Model_Year (74 vs. 70)'}    -6.1279e-07
    {'Model_Year (80 vs. 70)'}     -4.015e-07
    {'Model_Year (81 vs. 70)'}     3.4176e-07
    {'Model_Year (82 vs. 70)'}    -2.2483e-07
    {'Cylinders (6 vs. 8)'   }    -1.9024e-07
    {'Model_Year (76 vs. 70)'}     1.8136e-07
    {'Cylinders (5 vs. 8)'   }     1.7461e-07
    {'Model_Year (71 vs. 70)'}      1.558e-07
    {'Model_Year (75 vs. 70)'}     1.5456e-07
    {'Model_Year (77 vs. 70)'}      1.521e-07
    {'Model_Year (78 vs. 70)'}     1.4272e-07
    {'Model_Year (72 vs. 70)'}     6.7001e-08
    {'Model_Year (73 vs. 70)'}     4.7214e-08
    {'Cylinders (4 vs. 8)'   }     4.5118e-08
    {'Model_Year (79 vs. 70)'}    -2.2598e-08
      ⋮
```

Specify Blackbox Model as Function Handle

Train a regression model and create a `lime` object using a function handle to the `predict` function of the model. Use the object function `fit` to fit a simple model for the specified query point. Then display the coefficients of the fitted linear simple model by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight);
```

Train a blackbox model of `MPG` by using the `TreeBagger` function.

```
rng('default') % For reproducibility
Mdl = TreeBagger(100,tbl,MPG,'Method','regression','CategoricalPredictors',[2 5]);
```

`lime` does not support a `TreeBagger` object directly, so you cannot specify the first input argument (blackbox model) of `lime` as a `TreeBagger` object. Instead, you can use a function handle to the `predict` function. You can also specify options of the `predict` function using name-value arguments of the function.

Create the function handle to the `predict` function of the `TreeBagger` object `Mdl`. Specify the array of tree indices to use as `1:50`.

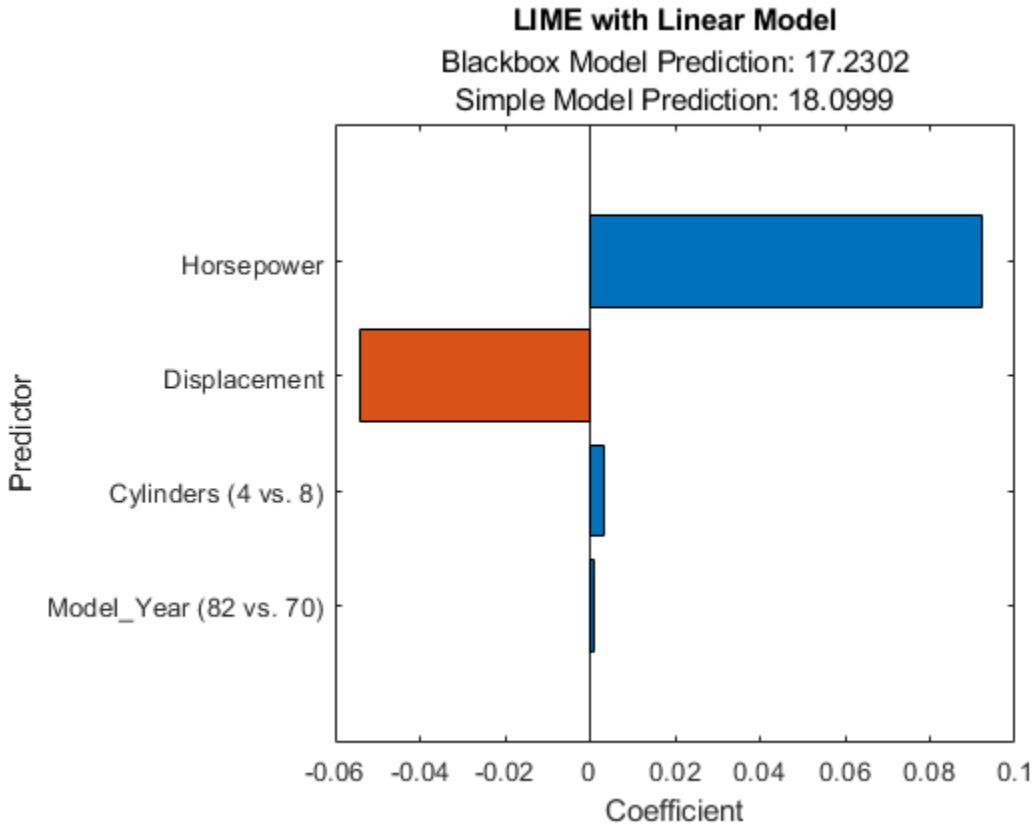
```
myPredict = @(tbl) predict(Mdl,tbl,'Trees',1:50);
```

Create a `lime` object using the function handle `myPredict`. When you specify a blackbox model as a function handle, you must provide the predictor data and specify the `'Type'` name-value argument. `tbl` includes categorical predictors (`Cylinder` and `Model_Year`) with the `double` data type. By default, `lime` does not treat variables with the `double` data type as categorical predictors. Specify the second (`Cylinder`) and fifth (`Model_Year`) variables as categorical predictors.

```
results = lime(myPredict,tbl,'Type','regression','CategoricalPredictors',[2 5]);
```

Fit a linear simple model for the first observation in `tbl`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
results = fit(results,tbl(1,:),4);
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```



`lime` finds `Horsepower`, `Displacement`, `Cylinders`, and `Model_Year` as important predictors.

More About

Distance Metrics

A distance metric is a function that defines a distance between two observations. `lime` supports various distance metrics for continuous variables and a mix of continuous and categorical variables.

- Distance metrics for continuous variables

  Given an mx-by-n data matrix X, which is treated as mx (1-by-n) row vectors x_1, x_2, ..., x_{mx}, and an my-by-n data matrix Y, which is treated as my (1-by-n) row vectors y_1, y_2, ...,y_{my}, the various distances between the vector x_s and y_t are defined as follows:

  - Euclidean distance

    $d_{st}^2 = (x_s - y_t)(x_s - y_t)'$.

The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.

- Standardized Euclidean distance

$$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)',$$

where V is the n-by-n diagonal matrix whose jth diagonal element is $(S(j))^2$, where S is a vector of scaling factors for each dimension.

- Mahalanobis distance

$$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)',$$

where C is the covariance matrix.

- City block distance

$$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right|.$$

The city block distance is a special case of the Minkowski distance, where $p = 1$.

- Minkowski distance

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right|^p}.$$

For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.

- Chebychev distance

$$d_{st} = \max_j \{ \left| x_{sj} - y_{tj} \right| \}.$$

The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- Cosine distance

$$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}} \right).$$

- Correlation distance

$$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}},$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

and

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}.$$

- Spearman distance

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

  where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$, as computed by `tiedrank`.
  - r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, that is, $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.
  - $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.
  - $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

- Distance metrics for a mix of continuous and categorical variables

  - Modified Goodall distance

    This distance is a variant of the Goodall distance, which assigns a small distance if the matching values are infrequent regardless of the frequencies of the other values. For mismatches, the distance contribution of the predictor is 1/(number of variables).

  - Occurrence frequency distance

    For a match, the occurrence frequency distance assigns zero distance. For a mismatch, the occurrence frequency distance assigns a higher distance on a less frequent value and a lower distance on a more frequent value.

Algorithms

LIME

To explain a prediction of a machine learning model using LIME [1], the software generates a synthetic data set and fits a simple interpretable model to the synthetic data set by using `lime` and `fit`, as described in steps 1–5.

- If you specify the `queryPoint` and `numImportantPredictors` values of `lime`, then the `lime` function performs all steps.
- If you do not specify `queryPoint` and `numImportantPredictors` and specify `'DataLocality'` as `'global'` (default), then the `lime` function generates a synthetic data set (steps 1–2), and the `fit` function fits a simple model (steps 3–5).
- If you do not specify `queryPoint` and `numImportantPredictors` and specify `'DataLocality'` as `'local'`, then the `fit` function performs all steps.

The `lime` and `fit` functions perform these steps:

1 Generate a synthetic predictor data set X_s using a multivariate normal distribution for continuous variables and a multinomial distribution for each categorical variable. You can specify the number of samples to generate by using the `'NumSyntheticData'` name-value argument.

- If `'DataLocality'` is `'global'` (default), then the software estimates the distribution parameters from the whole predictor data set (`X` or predictor data in `blackbox`).
- If `'DataLocality'` is `'local'`, then the software estimates the distribution parameters using the k-nearest neighbors of the query point, where k is the `'NumNeighbors'` value. You can specify a distance metric to find the nearest neighbors by using the `'Distance'` name-value argument.

The software ignores missing values in the predictor data set when estimating the distribution parameters.

Alternatively, you can provide a pregenerated, custom synthetic predictor data set by using the `customSyntheticData` input argument of `lime`.

2. Compute the predictions Y_s for the synthetic data set X_s. The predictions are predicted responses for regression or classified labels for classification. The software uses the `predict` function of the `blackbox` model to compute the predictions. If you specify `blackbox` as a function handle, then the software computes the predictions by using the function handle.
3. Compute the distances d between the query point and the samples in the synthetic predictor data set using the distance metric specified by `'Distance'`.
4. Compute the weight values w_q of the samples in the synthetic predictor data set with respect to the query point q using the squared exponential (or Gaussian) kernel function

$$w_q(x_s) = \exp\left(-\frac{1}{2}\left(\frac{d(x_s, q)}{\sqrt{p}\sigma}\right)^2\right).$$

- x_s is a sample in the synthetic predictor data set X_s.
- $d(x_s,q)$ is the distance between the sample x_s and the query point q.
- p is the number of predictors in X_s.
- σ is the kernel width, which you can specify by using the `'KernelWidth'` name-value argument. The default `'KernelWidth'` value is 0.75.

The weight value at the query point is 1, and then it converges to zero as the distance value increases. The `'KernelWidth'` value controls how fast the weight value converges to zero. The lower the `'KernelWidth'` value, the faster the weight value converges to zero. Therefore, the algorithm gives more weight to samples near the query point. Because this algorithm uses such weight values, the selected important predictors and fitted simple model effectively explain the predictions for the synthetic data locally, around the query point.

5. Fit a simple model.

- If `'SimpleModelType'` is `'linear'` (default), then the software selects important predictors and fits a linear model of the selected important predictors.
  - Select n important predictors ($\tilde{X}_s$) by using the group orthogonal matching pursuit (OMP) algorithm [2][3], where n is the `numImportantPredictors` value. This algorithm uses the synthetic predictor data set (X_s), predictions (Y_s), and weight values (w_q).
  - Fit a linear model of the selected important predictors ($\tilde{X}_s$) to the predictions (Y_s) using the weight values (w_q). The software uses `fitrlinear` for regression or `fitclinear` for classification. For a multiclass model, the software uses the one-versus-all scheme to construct a binary classification problem. The positive class is the predicted class for the query point from the `blackbox` model, and the negative class refers to the other classes.

- If `'SimpleModelType'` is `'tree'`, then the software fits a decision tree model by using `fitrtree` for regression or `fitctree` for classification. The software specifies the maximum number of decision splits (or branch nodes) as the number of important predictors so that the fitted decision tree uses at most the specified number of predictors.

References

[1] Ribeiro, Marco Tulio, S. Singh, and C. Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. San Francisco, California: ACM, 2016.

[2] Świrszcz, Grzegorz, Naoki Abe, and Aurélie C. Lozano. "Grouped Orthogonal Matching Pursuit for Variable Selection and Prediction." *Advances in Neural Information Processing Systems* (2009): 1150–58.

[3] Lozano, Aurélie C., Grzegorz Świrszcz, and Naoki Abe. "Group Orthogonal Matching Pursuit for Logistic Regression." *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (2011): 452–60.

See Also

`plotPartialDependence` | `shapley`

Topics

“Interpret Deep Network Predictions on Tabular Data Using LIME” (Deep Learning Toolbox)
“Interpret Machine Learning Models” on page 18-256

Introduced in R2020b

LinearModel

Linear regression model

Description

`LinearModel` is a fitted linear regression model object. A regression model describes the relationship between a response and predictors. The linearity in a linear regression model refers to the linearity of the predictor coefficients.

Use the properties of a `LinearModel` object to investigate a fitted linear regression model. The object properties include information about coefficient estimates, summary statistics, fitting method, and input data. Use the object functions to predict responses and to modify, evaluate, and visualize the linear regression model.

Creation

Create a `LinearModel` object by using `fitlm` or `stepwiselm`.

`fitlm` fits a linear regression model to data using a fixed model specification. Use `addTerms`, `removeTerms`, or `step` to add or remove terms from the model. Alternatively, use `stepwiselm` to fit a model using stepwise linear regression.

Properties

Coefficient Estimates

`CoefficientCovariance` — Covariance matrix of coefficient estimates
numeric matrix

This property is read-only.

Covariance matrix of coefficient estimates, specified as a *p*-by-*p* matrix of numeric values. *p* is the number of coefficients in the fitted model.

For details, see “Coefficient Standard Errors and Confidence Intervals” on page 11-58.

Data Types: `single` | `double`

`CoefficientNames` — Coefficient names
cell array of character vectors

This property is read-only.

Coefficient names, specified as a cell array of character vectors, each containing the name of the corresponding term.

Data Types: `cell`

`Coefficients` — Coefficient values
table

This property is read-only.

Coefficient values, specified as a table. `Coefficients` contains one row for each coefficient and these columns:

- `Estimate` — Estimated coefficient value
- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is zero
- `pValue` — *p*-value for the *t*-statistic

Use `anova` (only for a linear regression model) or `coefTest` to perform other tests on the coefficients. Use `coefCI` to find the confidence intervals of the coefficient estimates.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the estimated coefficient vector in the model `mdl`:

```
beta = mdl.Coefficients.Estimate
```

Data Types: `table`

`NumCoefficients` — Number of model coefficients
positive integer

This property is read-only.

Number of model coefficients, specified as a positive integer. `NumCoefficients` includes coefficients that are set to zero when the model terms are rank deficient.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated coefficients
positive integer

This property is read-only.

Number of estimated coefficients in the model, specified as a positive integer. `NumEstimatedCoefficients` does not include coefficients that are set to zero when the model terms are rank deficient. `NumEstimatedCoefficients` is the degrees of freedom for regression.

Data Types: `double`

Summary Statistics

`DFE` — Degrees of freedom for error
positive integer

This property is read-only.

Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients, specified as a positive integer.

Data Types: `double`

`Diagnostics` — Observation diagnostics
table

This property is read-only.

Observation diagnostics, specified as a table that contains one row for each observation and the columns described in this table.

Column	Meaning	Description
`Leverage`	Diagonal elements of `HatMatrix`	`Leverage` for each observation indicates to what extent the fit is determined by the observed predictor values. A value close to `1` indicates that the fit is largely determined by that observation, with little contribution from the other observations. A value close to `0` indicates that the fit is largely determined by the other observations. For a model with `P` coefficients and `N` observations, the average value of `Leverage` is `P/N`. A `Leverage` value greater than `2*P/N` indicates high leverage.
`CooksDistance`	Cook's distance	`CooksDistance` is a measure of scaled change in fitted values. An observation with `CooksDistance` greater than three times the mean Cook's distance can be an outlier.
`Dffits`	Delete-1 scaled differences in fitted values	`Dffits` is the scaled change in the fitted values for each observation that results from excluding that observation from the fit. Values greater than `2*sqrt(P/N)` in absolute value can be considered influential.
`S2_i`	Delete-1 variance	`S2_i` is a set of residual variance estimates obtained by deleting each observation in turn. These estimates can be compared with the mean squared error (MSE) value, stored in the `MSE` property.
`CovRatio`	Delete-1 ratio of determinant of covariance	`CovRatio` is the ratio of the determinant of the coefficient covariance matrix, with each observation deleted in turn, to the determinant of the covariance matrix for the full model. Values greater than `1 + 3*P/N` or less than `1 – 3*P/N` indicate influential points.
`Dfbetas`	Delete-1 scaled differences in coefficient estimates	`Dfbetas` is an `N`-by-`P` matrix of the scaled change in the coefficient estimates that results from excluding each observation in turn. Values greater than `3/sqrt(N)` in absolute value indicate that the observation has a significant influence on the corresponding coefficient.
`HatMatrix`	Projection matrix to compute `fitted` from observed responses	`HatMatrix` is an `N`-by-`N` matrix such that `Fitted = HatMatrix*Y`, where `Y` is the response vector and `Fitted` is the vector of fitted response values.

`Diagnostics` contains information that is helpful in finding outliers and influential observations. Delete-1 diagnostics capture the changes that result from excluding each observation in turn from the fit. For more details, see “Hat Matrix and Leverage” on page 11-77, “Cook’s Distance” on page 11-55, and “Delete-1 Statistics” on page 11-63.

Use `plotDiagnostics` to plot observation diagnostics.

Rows not used in the fit because of missing values (in `ObservationInfo.Missing`) or excluded values (in `ObservationInfo.Excluded`) contain `NaN` values in the `CooksDistance`, `Dffits`, `S2_i`, and `CovRatio` columns and zeros in the `Leverage`, `Dfbetas`, and `HatMatrix` columns.

To obtain any of these columns as an array, index into the property using dot notation. For example, obtain the delete-1 variance vector in the model `mdl`:

```
S2i = mdl.Diagnostics.S2_i;
```

Data Types: `table`

`Fitted` — Fitted response values based on input data

numeric vector

This property is read-only.

Fitted (predicted) response values based on input data, specified as an *n*-by-1 numeric vector. *n* is the number of observations in the input data. Use `predict` to compute predictions for other predictor values, or to compute confidence bounds on `Fitted`.

Data Types: `single` | `double`

`LogLikelihood` — Loglikelihood

numeric value

This property is read-only.

Loglikelihood of response values, specified as a numeric value, based on the assumption that each response value follows a normal distribution. The mean of the normal distribution is the fitted (predicted) response value, and the variance is the `MSE`.

Data Types: `single` | `double`

`ModelCriterion` — Criterion for model comparison

structure

This property is read-only.

Criterion for model comparison, specified as a structure with these fields:

- `AIC` — Akaike information criterion. `AIC = –2*logL + 2*m`, where `logL` is the loglikelihood and `m` is the number of estimated parameters.
- `AICc` — Akaike information criterion corrected for the sample size. `AICc = AIC + (2*m*(m + 1))/(n – m – 1)`, where `n` is the number of observations.
- `BIC` — Bayesian information criterion. `BIC = –2*logL + m*log(n)`.
- `CAIC` — Consistent Akaike information criterion. `CAIC = –2*logL + m*(log(n) + 1)`.

Information criteria are model selection tools that you can use to compare multiple models fit to the same data. These criteria are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). Different information criteria are distinguished by the form of the penalty.

When you compare multiple models, the model with the lowest information criterion value is the best-fitting model. The best-fitting model can vary depending on the criterion used for model comparison.

To obtain any of the criterion values as a scalar, index into the property using dot notation. For example, obtain the AIC value `aic` in the model `mdl`:

```
aic = mdl.ModelCriterion.AIC
```

Data Types: `struct`

`MSE` — Mean squared error
numeric value

This property is read-only.

Mean squared error (residuals), specified as a numeric value.

$$MSE = SSE / DFE,$$

where *MSE* is the mean squared error, *SSE* is the sum of squared errors, and *DFE* is the degrees of freedom.

Data Types: `single` | `double`

`Residuals` — Residuals for fitted model
table

This property is read-only.

Residuals for the fitted model, specified as a table that contains one row for each observation and the columns described in this table.

Column	Description
`Raw`	Observed minus fitted values
`Pearson`	Raw residuals divided by the root mean squared error (RMSE)
`Standardized`	Raw residuals divided by their estimated standard deviation
`Studentized`	Raw residual divided by an independent estimate of the residual standard deviation. The residual for observation *i* is divided by an estimate of the error standard deviation based on all observations except observation *i*.

Use `plotResiduals` to create a plot of the residuals. For details, see “Residuals” on page 11-80.

Rows not used in the fit because of missing values (in `ObservationInfo.Missing`) or excluded values (in `ObservationInfo.Excluded`) contain `NaN` values.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the raw residual vector `r` in the model `mdl`:

```
r = mdl.Residuals.Raw
```

Data Types: `table`

`RMSE` — Root mean squared error
numeric value

This property is read-only.

Root mean squared error (residuals), specified as a numeric value.

$$RMSE = \text{sqrt}(MSE),$$

where *RMSE* is the root mean squared error and *MSE* is the mean squared error.

Data Types: `single` | `double`

Rsquared — R-squared value for model

structure

This property is read-only.

R-squared value for the model, specified as a structure with two fields:

- `Ordinary` — Ordinary (unadjusted) R-squared
- `Adjusted` — R-squared adjusted for the number of coefficients

The R-squared value is the proportion of the total sum of squares explained by the model. The ordinary R-squared value relates to the SSR and SST properties:

$$Rsquared = SSR/SST,$$

where SST is the total sum of squares, and SSR is the regression sum of squares.

For details, see “Coefficient of Determination (R-Squared)” on page 11-61.

To obtain either of these values as a scalar, index into the property using dot notation. For example, obtain the adjusted R-squared value in the model `mdl`:

```
r2 = mdl.Rsquared.Adjusted
```

Data Types: `struct`

SSE — Sum of squared errors

numeric value

This property is read-only.

Sum of squared errors (residuals), specified as a numeric value.

The Pythagorean theorem implies

$$SST = SSE + SSR,$$

where SST is the total sum of squares, SSE is the sum of squared errors, and SSR is the regression sum of squares.

Data Types: `single` | `double`

SSR — Regression sum of squares

numeric value

This property is read-only.

Regression sum of squares, specified as a numeric value. The regression sum of squares is equal to the sum of squared deviations of the fitted values from their mean.

The Pythagorean theorem implies

$$SST = SSE + SSR,$$

where SST is the total sum of squares, SSE is the sum of squared errors, and SSR is the regression sum of squares.

Data Types: `single` | `double`

SST — Total sum of squares
numeric value

This property is read-only.

Total sum of squares, specified as a numeric value. The total sum of squares is equal to the sum of squared deviations of the response vector `y` from the `mean(y)`.

The Pythagorean theorem implies
`SST = SSE + SSR`,
where `SST` is the total sum of squares, `SSE` is the sum of squared errors, and `SSR` is the regression sum of squares.

Data Types: `single` | `double`

Fitting Method

`Robust` — Robust fit information
structure

This property is read-only.

Robust fit information, specified as a structure with the fields described in this table.

Field	Description
`WgtFun`	Robust weighting function, such as `'bisquare'` (see `'RobustOpts'`)
`Tune`	Tuning constant. This field is empty (`[]`) if `WgtFun` is `'ols'` or if `WgtFun` is a function handle for a custom weight function with the default tuning constant 1.
`Weights`	Vector of weights used in the final iteration of robust fit. This field is empty for a `CompactLinearModel` object.

This structure is empty unless you fit the model using robust regression.

Data Types: `struct`

`Steps` — Stepwise fitting information
structure

This property is read-only.

Stepwise fitting information, specified as a structure with the fields described in this table.

Field	Description
`Start`	Formula representing the starting model
`Lower`	Formula representing the lower bound model. The terms in `Lower` must remain in the model.
`Upper`	Formula representing the upper bound model. The model cannot contain more terms than `Upper`.
`Criterion`	Criterion used for the stepwise algorithm, such as `'sse'`

Field	Description
`PEnter`	Threshold for `Criterion` to add a term
`PRemove`	Threshold for `Criterion` to remove a term
`History`	Table representing the steps taken in the fit

The `History` table contains one row for each step, including the initial fit, and the columns described in this table.

Column	Description
`Action`	Action taken during the step: • `'Start'` — First step • `'Add'` — A term is added • `'Remove'` — A term is removed
`TermName`	• If `Action` is `'Start'`, `TermName` specifies the starting model specification. • If `Action` is `'Add'` or `'Remove'`, `TermName` specifies the term added or removed in the step.
`Terms`	Model specification in a “Terms Matrix” on page 33-3525
`DF`	Regression degrees of freedom after the step
`delDF`	Change in regression degrees of freedom from the previous step (negative for steps that remove a term)
`Deviance`	Deviance (residual sum of squares) at the step (only for a generalized linear regression model)
`FStat`	*F*-statistic that leads to the step
`PValue`	*p*-value of the *F*-statistic

The structure is empty unless you fit the model using stepwise regression.

Data Types: `struct`

Input Data

`Formula` — Model information

`LinearFormula` object

This property is read-only.

Model information, specified as a `LinearFormula` object.

Display the formula of the fitted model `mdl` using dot notation:

```
mdl.Formula
```

`NumObservations` — Number of observations

positive integer

This property is read-only.

Number of observations the fitting function used in fitting, specified as a positive integer. `NumObservations` is the number of observations supplied in the original table, dataset, or matrix, minus any excluded rows (set with the `'Exclude'` name-value pair argument) or rows with missing values.

Data Types: `double`

`NumPredictors` — Number of predictor variables
positive integer

This property is read-only.

Number of predictor variables used to fit the model, specified as a positive integer.

Data Types: `double`

`NumVariables` — Number of variables
positive integer

This property is read-only.

Number of variables in the input data, specified as a positive integer. `NumVariables` is the number of variables in the original table or dataset, or the total number of columns in the predictor matrix and response vector.

`NumVariables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `double`

`ObservationInfo` — Observation information
table

This property is read-only.

Observation information, specified as an *n*-by-4 table, where *n* is equal to the number of rows of input data. `ObservationInfo` contains the columns described in this table.

Column	Description
`Weights`	Observation weights, specified as a numeric value. The default value is `1`.
`Excluded`	Indicator of excluded observations, specified as a logical value. The value is `true` if you exclude the observation from the fit by using the `'Exclude'` name-value pair argument.
`Missing`	Indicator of missing observations, specified as a logical value. The value is `true` if the observation is missing.
`Subset`	Indicator of whether or not the fitting function uses the observation, specified as a logical value. The value is `true` if the observation is not excluded or missing, meaning the fitting function uses the observation.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the weight vector `w` of the model `mdl`:

```
w = mdl.ObservationInfo.Weights
```

Data Types: `table`

`ObservationNames` — Observation names

cell array of character vectors

This property is read-only.

Observation names, specified as a cell array of character vectors containing the names of the observations used in the fit.

- If the fit is based on a table or dataset containing observation names, `ObservationNames` uses those names.
- Otherwise, `ObservationNames` is an empty cell array.

Data Types: `cell`

`PredictorNames` — Names of predictors used to fit model

cell array of character vectors

This property is read-only.

Names of predictors used to fit the model, specified as a cell array of character vectors.

Data Types: `cell`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`VariableInfo` — Information about variables

table

This property is read-only.

Information about variables contained in `Variables`, specified as a table with one row for each variable and the columns described in this table.

Column	Description
`Class`	Variable class, specified as a cell array of character vectors, such as `'double'` and `'categorical'`
`Range`	Variable range, specified as a cell array of vectors • Continuous variable — Two-element vector `[`*`min`*`,`*`max`*`]`, the minimum and maximum values • Categorical variable — Vector of distinct variable values
`InModel`	Indicator of which variables are in the fitted model, specified as a logical vector. The value is `true` if the model includes the variable.
`IsCategorical`	Indicator of categorical variables, specified as a logical vector. The value is `true` if the variable is categorical.

`VariableInfo` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

`VariableNames` — Names of variables
cell array of character vectors

This property is read-only.

Names of variables, specified as a cell array of character vectors.

- If the fit is based on a table or dataset, this property provides the names of the variables in the table or dataset.
- If the fit is based on a predictor matrix and response vector, `VariableNames` contains the values specified by the `'VarNames'` name-value pair argument of the fitting method. The default value of `'VarNames'` is `{'x1','x2',...,'xn','y'}`.

`VariableNames` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `cell`

`Variables` — Input data
table

This property is read-only.

Input data, specified as a table. `Variables` contains both predictor and response values. If the fit is based on a table or dataset array, `Variables` contains all the data from the table or dataset array. Otherwise, `Variables` is a table created from the input data matrix `X` and the response vector `y`.

`Variables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

Object Functions

Create `CompactLinearModel`

compact	Compact linear regression model

Add or Remove Terms from Linear Model

addTerms	Add terms to linear regression model
removeTerms	Remove terms from linear regression model
step	Improve linear regression model by adding or removing terms

Predict Responses

feval	Predict responses of linear regression model using one input for each predictor
predict	Predict responses of linear regression model
random	Simulate responses with random noise for linear regression model

Evaluate Linear Model

anova	Analysis of variance for linear regression model
coefCI	Confidence intervals of coefficient estimates of linear regression model
coefTest	Linear hypothesis test on linear regression model coefficients
dwtest	Durbin-Watson test with linear regression model object
partialDependence	Compute partial dependence

Visualize Linear Model and Summary Statistics

plot	Scatter plot or added variable plot of linear regression model
plotAdded	Added variable plot of linear regression model
plotAdjustedResponse	Adjusted response plot of linear regression model
plotDiagnostics	Plot observation diagnostics of linear regression model
plotEffects	Plot main effects of predictors in linear regression model
plotInteraction	Plot interaction effects of two predictors in linear regression model
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotResiduals	Plot residuals of linear regression model
plotSlice	Plot of slices through fitted linear regression surface

Gather Properties of Linear Model

gather	Gather properties of machine learning model from GPU

Examples

Fit Linear Regression Using Data in Matrix

Fit a linear regression model using a matrix input data set.

Load the `carsmall` data set, a matrix input data set.

```
load carsmall
X = [Weight,Horsepower,Acceleration];
```

Fit a linear regression model by using `fitlm`.

```
mdl = fitlm(X,MPG)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate        SE          tStat        pValue
                  __________    _________    _________    __________

    (Intercept)        47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
    x3              -0.011583      0.19333    -0.059913       0.95236


Number of observations: 93, Error degrees of freedom: 89
```

```
Root Mean Squared Error: 4.09
R-squared: 0.752,  Adjusted R-Squared: 0.744
F-statistic vs. constant model: 90, p-value = 7.38e-27
```

The model display includes the model formula, estimated coefficients, and model summary statistics.

The model formula in the display, `y ~ 1 + x1 + x2 + x3`, corresponds to $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

The model display also shows the estimated coefficient information, which is stored in the `Coefficients` property. Display the `Coefficients` property.

```
mdl.Coefficients
```

```
ans=4×4 table
                    Estimate        SE          tStat        pValue
                   __________    _________    _________    __________

    (Intercept)       47.977       3.8785        12.37    4.8957e-21
    x1             -0.0065416    0.0011274      -5.8023    9.8742e-08
    x2              -0.042943     0.024313      -1.7663       0.08078
    x3              -0.011583      0.19333    -0.059913       0.95236
```

The `Coefficient` property includes these columns:

- `Estimate` — Coefficient estimates for each corresponding term in the model. For example, the estimate for the constant term (`intercept`) is 47.977.
- `SE` — Standard error of the coefficients.
- `tStat` — *t*-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. Note that `tStat = Estimate/SE`. For example, the *t*-statistic for the intercept is 47.977/3.8785 = 12.37.
- `pValue` — *p*-value for the *t*-statistic of the hypothesis test that the corresponding coefficient is equal to zero or not. For example, the *p*-value of the *t*-statistic for `x2` is greater than 0.05, so this term is not significant at the 5% significance level given the other terms in the model.

The summary statistics of the model are:

- `Number of observations` — Number of rows without any `NaN` values. For example, `Number of observations` is 93 because the `MPG` data vector has six `NaN` values and the `Horsepower` data vector has one `NaN` value for a different observation, where the number of rows in `X` and `MPG` is 100.
- `Error degrees of freedom` — $n - p$, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, the model has four predictors, so the `Error degrees of freedom` is 93 – 4 = 89.
- `Root mean squared error` — Square root of the mean squared error, which estimates the standard deviation of the error distribution.
- `R-squared` and `Adjusted R-squared` — Coefficient of determination and adjusted coefficient of determination, respectively. For example, the `R-squared` value suggests that the model explains approximately 75% of the variability in the response variable `MPG`.

- `F-statistic vs. constant model` — Test statistic for the *F*-test on the regression model, which tests whether the model fits significantly better than a degenerate model consisting of only a constant term.
- `p-value` — *p*-value for the *F*-test on the model. For example, the model is significant with a *p*-value of 7.3816e-27.

You can find these statistics in the model properties (`NumObservations`, `DFE`, `RMSE`, and `Rsquared`) and by using the `anova` function.

```
anova(mdl,'summary')
```

```
ans=3×5 table
                SumSq     DF    MeanSq      F         pValue
                ______    __    ______    ______    __________

    Total       6004.8    92    65.269
    Model         4516     3    1505.3    89.987    7.3816e-27
    Residual    1488.8    89    16.728
```

Linear Regression with Categorical Predictor

Fit a linear regression model that contains a categorical predictor. Reorder the categories of the categorical predictor to control the reference level in the model. Then, use `anova` to test the significance of the categorical variable.

Model with Categorical Predictor

Load the `carsmall` data set and create a linear regression model of `MPG` as a function of `Model_Year`. To treat the numeric vector `Model_Year` as a categorical variable, identify the predictor using the `'CategoricalVars'` name-value pair argument.

```
load carsmall
mdl = fitlm(Model_Year,MPG,'CategoricalVars',1,'VarNames',{'Model_Year','MPG'})
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                    Estimate      SE      tStat       pValue
                    ________    ______    ______    __________

    (Intercept)        17.69    1.0328    17.127    3.2371e-30
    Model_Year_76     3.8839    1.4059    2.7625     0.0069402
    Model_Year_82      14.02    1.4369    9.7571    8.2164e-16


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

The model formula in the display, `MPG ~ 1 + Model_Year`, corresponds to

$$\text{MPG} = \beta_0 + \beta_1 I_{\text{Year} = 76} + \beta_2 I_{\text{Year} = 82} + \epsilon,$$

where $I_{\text{Year} = 76}$ and $I_{\text{Year} = 82}$ are indicator variables whose value is one if the value of `Model_Year` is 76 and 82, respectively. The `Model_Year` variable includes three distinct values, which you can check by using the `unique` function.

```
unique(Model_Year)
```

```
ans = 3×1

    70
    76
    82
```

`fitlm` chooses the smallest value in `Model_Year` as a reference level (`'70'`) and creates two indicator variables $I_{\text{Year} = 76}$ and $I_{\text{Year} = 82}$. The model includes only two indicator variables because the design matrix becomes rank deficient if the model includes three indicator variables (one for each level) and an intercept term.

Model with Full Indicator Variables

You can interpret the model formula of `mdl` as a model that has three indicator variables without an intercept term:

$$y = \beta_0 I_{x_1 = 70} + (\beta_0 + \beta_1) I_{x_1 = 76} + (\beta_0 + \beta_2) I_{x_2 = 82} + \epsilon.$$

Alternatively, you can create a model that has three indicator variables without an intercept term by manually creating indicator variables and specifying the model formula.

```
temp_Year = dummyvar(categorical(Model_Year));
Model_Year_70 = temp_Year(:,1);
Model_Year_76 = temp_Year(:,2);
Model_Year_82 = temp_Year(:,3);
tbl = table(Model_Year_70,Model_Year_76,Model_Year_82,MPG);
mdl = fitlm(tbl,'MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82 - 1')
```

```
mdl =
Linear regression model:
    MPG ~ Model_Year_70 + Model_Year_76 + Model_Year_82

Estimated Coefficients:
                     Estimate      SE       tStat       pValue
                     ________    _______    ______    __________

    Model_Year_70      17.69      1.0328    17.127    3.2371e-30
    Model_Year_76     21.574     0.95387    22.617    4.0156e-39
    Model_Year_82      31.71     0.99896    31.743    5.2234e-51


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
```

Choose Reference Level in Model

You can choose a reference level by modifying the order of categories in a categorical variable. First, create a categorical variable `Year`.

```
Year = categorical(Model_Year);
```

Check the order of categories by using the `categories` function.

```
categories(Year)
```

```
ans = 3x1 cell
    {'70'}
    {'76'}
    {'82'}
```

If you use `Year` as a predictor variable, then `fitlm` chooses the first category `'70'` as a reference level. Reorder `Year` by using the `reordercats` function.

```
Year_reordered = reordercats(Year,{'76','70','82'});
categories(Year_reordered)
```

```
ans = 3x1 cell
    {'76'}
    {'70'}
    {'82'}
```

The first category of `Year_reordered` is `'76'`. Create a linear regression model of `MPG` as a function of `Year_reordered`.

```
mdl2 = fitlm(Year_reordered,MPG,'VarNames',{'Model_Year','MPG'})
```

```
mdl2 =
Linear regression model:
    MPG ~ 1 + Model_Year

Estimated Coefficients:
                    Estimate      SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)       21.574     0.95387     22.617    4.0156e-39
    Model_Year_70    -3.8839      1.4059    -2.7625     0.0069402
    Model_Year_82     10.136      1.3812     7.3385    8.7634e-11


Number of observations: 94, Error degrees of freedom: 91
Root Mean Squared Error: 5.56
R-squared: 0.531,  Adjusted R-Squared: 0.521
F-statistic vs. constant model: 51.6, p-value = 1.07e-15
```

`mdl2` uses `'76'` as a reference level and includes two indicator variables $I_{\text{Year = 70}}$ and $I_{\text{Year = 82}}$.

Evaluate Categorical Predictor

The model display of `mdl2` includes a *p*-value of each term to test whether or not the corresponding coefficient is equal to zero. Each *p*-value examines each indicator variable. To examine the categorical variable `Model_Year` as a group of indicator variables, use `anova`. Use the `'components'`(default) option to return a component ANOVA table that includes ANOVA statistics for each variable in the model except the constant term.

```
anova(mdl2,'components')
```

```
ans=2×5 table
                  SumSq     DF    MeanSq      F         pValue
                  ______    __    ______    _____    __________

    Model_Year    3190.1     2    1595.1    51.56    1.0694e-15
    Error         2815.2    91    30.936
```

The component ANOVA table includes the *p*-value of the `Model_Year` variable, which is smaller than the *p*-values of the indicator variables.

Fit Robust Linear Regression Model

Load the `hald` data set, which measures the effect of cement composition on its hardening heat.

```
load hald
```

This data set includes the variables `ingredients` and `heat`. The matrix `ingredients` contains the percent composition of four chemicals present in the cement. The vector `heat` contains the values for the heat hardening after 180 days for each cement sample.

Fit a robust linear regression model to the data.

```
mdl = fitlm(ingredients,heat,'RobustOpts','on')
```

```
mdl =
Linear regression model (robust fit):
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    ________

    (Intercept)       60.09     75.818     0.79256      0.4509
    x1               1.5753    0.80585      1.9548    0.086346
    x2               0.5322    0.78315     0.67957     0.51596
    x3              0.13346     0.8166     0.16343     0.87424
    x4             -0.12052     0.7672    -0.15709     0.87906


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 2.65
R-squared: 0.979,  Adjusted R-Squared: 0.969
F-statistic vs. constant model: 94.6, p-value = 9.03e-07
```

For more details, see the topic “Reduce Outlier Effects Using Robust Regression” on page 11-104, which compares the results of a robust fit to a standard least-squares fit.

Fit Linear Model Using Stepwise Regression

Load the `hald` data set, which measures the effect of cement composition on its hardening heat.

```
load hald
```

This data set includes the variables `ingredients` and `heat`. The matrix `ingredients` contains the percent composition of four chemicals present in the cement. The vector `heat` contains the values for the heat hardening after 180 days for each cement sample.

Fit a stepwise linear regression model to the data. Specify 0.06 as the threshold for the criterion to add a term to the model.

```
mdl = stepwiselm(ingredients,heat,'PEnter',0.06)
```

```
1. Adding x4, FStat = 22.7985, pValue = 0.000576232
2. Adding x1, FStat = 108.2239, pValue = 1.105281e-06
3. Adding x2, FStat = 5.0259, pValue = 0.051687
4. Removing x4, FStat = 1.8633, pValue = 0.2054

mdl =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     52.577       2.2862    22.998    5.4566e-10
    x1              1.4683       0.1213    12.105    2.6922e-07
    x2             0.66225     0.045855    14.442     5.029e-08


Number of observations: 13, Error degrees of freedom: 10
Root Mean Squared Error: 2.41
R-squared: 0.979,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 230, p-value = 4.41e-09
```

By default, the starting model is a constant model. `stepwiselm` performs forward selection and adds the `x4`, `x1`, and `x2` terms (in that order), because the corresponding *p*-values are less than the `PEnter` value of 0.06. `stepwiselm` then uses backward elimination and removes `x4` from the model because, once `x2` is in the model, the *p*-value of `x4` is greater than the default value of `PRemove`, 0.1.

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Alternative Functionality

- For reduced computation time on high-dimensional data sets, fit a linear regression model using the `fitrlinear` function.
- To regularize a regression, use `fitrlinear`, `lasso`, `ridge`, or `plsregress`.
  - `fitrlinear` regularizes a regression for high-dimensional data sets using lasso or ridge regression.
  - `lasso` removes redundant predictors in linear regression using lasso or elastic net.
  - `ridge` regularizes a regression with correlated terms using ridge regression.
  - `plsregress` regularizes a regression with correlated terms using partial least squares.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `random` functions support code generation.

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The following object functions fully support GPU arrays:
  - `feval`
  - `predict`
  - `random`
  - `partialDependence`
  - `plotPartialDependence`
- The following object functions support model objects fitted with GPU array input arguments:
  - `compact`
  - `addTerms`
  - `removeTerms`
  - `step`
  - `anova`
  - `coefCI`

- `coefTest`
- `dwtest`
- `plot`
- `plotAdded`
- `plotAdjustedResponse`
- `plotDiagnostics`
- `plotEffects`
- `plotInteraction`
- `plotResiduals`
- `plotSlice`
- `gather`

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `fitlm` | `stepwiselm`

Topics

“Linear Regression” on page 11-9
“What Is a Linear Regression Model?” on page 11-6

Introduced in R2012a

LinearMixedModel class

Linear mixed-effects model class

Description

A `LinearMixedModel` object represents a model of a response variable with fixed and random effects. It comprises data, a model description, fitted coefficients, covariance parameters, design matrices, residuals, residual plots, and other diagnostic information for a linear mixed-effects model. You can predict model responses with the `predict` function and generate random data at new design points using the `random` function.

Construction

You can fit a linear mixed-effects model using `fitlme(tbl,formula)` if your data is in a table or dataset array. Alternatively, if your model is not easily described using a formula, you can create matrices to define the fixed and random effects, and fit the model using `fitlmematrix(X,y,Z,G)`.

Input Arguments

`tbl` — Input data

table | `dataset` array

Input data, which includes the response variable, predictor variables, and grouping variables, specified as a table or `dataset` array. The predictor variables can be continuous or grouping variables (see “Grouping Variables” on page 2-45). You must specify the model for the variables using `formula`.

Data Types: `table`

`formula` — Formula for model specification

character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`

Formula for model specification, specified as a character vector or string scalar of the form `'y ~ fixed + (random1|grouping1) + ... + (randomR|groupingR)'`. For a full description, see “Formula” on page 33-3540.

Example: `'y ~ treatment +(1|block)'`

`X` — Fixed-effects design matrix

n-by-*p* matrix

Fixed-effects design matrix, specified as an *n*-by-*p* matrix, where *n* is the number of observations, and *p* is the number of fixed-effects predictor variables. Each row of `X` corresponds to one observation, and each column of `X` corresponds to one variable.

Data Types: `single` | `double`

`y` — Response values

n-by-1 vector

Response values, specified as an *n*-by-1 vector, where *n* is the number of observations.

Data Types: `single` | `double`

Z — Random-effects design

n-by-*q* matrix | cell array of *R* *n*-by-*q*(*r*) matrices, *r* = 1, 2, ..., *R*

Random-effects design, specified as either of the following.

- If there is one random-effects term in the model, then `Z` must be an *n*-by-*q* matrix, where *n* is the number of observations and *q* is the number of variables in the random-effects term.
- If there are *R* random-effects terms, then `Z` must be a cell array of length *R*. Each cell of `Z` contains an *n*-by-*q*(*r*) design matrix `Z{r}`, *r* = 1, 2, ..., *R*, corresponding to each random-effects term. Here, *q*(*r*) is the number of random effects term in the *r*th random effects design matrix, `Z{r}`.

Data Types: `single` | `double` | `cell`

G — Grouping variable or variables

n-by-1 vector | cell array of *R* *n*-by-1 vectors

Grouping variable or variables on page 2-45, specified as either of the following.

- If there is one random-effects term, then `G` must be an *n*-by-1 vector corresponding to a single grouping variable with *M* levels or groups.

  `G` can be a categorical vector, logical vector, numeric vector, character array, string array, or cell array of character vectors.
- If there are multiple random-effects terms, then `G` must be a cell array of length *R*. Each cell of `G` contains a grouping variable `G{r}`, *r* = 1, 2, ..., *R*, with *M*(*r*) levels.

  `G{r}` can be a categorical vector, logical vector, numeric vector, character array, string array, or cell array of character vectors.

Data Types: `categorical` | `logical` | `single` | `double` | `char` | `string` | `cell`

Properties

Coefficients — Fixed-effects coefficient estimates

dataset array

Fixed-effects coefficient estimates and related statistics, stored as a dataset array containing the following fields.

`Name`	Name of the term.
`Estimate`	Estimated value of the coefficient.
`SE`	Standard error of the coefficient.
`tStat`	*t*-statistics for testing the null hypothesis that the coefficient is equal to zero.

`DF`	Degrees of freedom for the *t*-test. Method to compute `DF` is specified by the `'DFMethod'` name-value pair argument. `Coefficients` always uses the `'Residual'` method for `'DFMethod'`.
`pValue`	*p*-value for the *t*-test.
`Lower`	Lower limit of the confidence interval for coefficient. `Coefficients` always uses the 95% confidence level, i.e.`'alpha'` is 0.05.
`Upper`	Upper limit of confidence interval for coefficient. `Coefficients` always uses the 95% confidence level, i.e.`'alpha'` is 0.05.

You can change `'DFMethod'` and `'alpha'` while computing confidence intervals for or testing hypotheses involving fixed- and random-effects, using the `coefCI` and `coefTest` methods.

`CoefficientCovariance` — Covariance of the estimated fixed-effects coefficients
p-by-*p* matrix

Covariance of the estimated fixed-effects coefficients of the linear mixed-effects model, stored as a *p*-by-*p* matrix, where *p* is the number of fixed-effects coefficients.

You can display the covariance parameters associated with the random effects using the `covarianceParameters` method.

Data Types: `double`

`CoefficientNames` — Names of the fixed-effects coefficients
1-by-*p* cell array of character vectors

Names of the fixed-effects coefficients of a linear mixed-effects model, stored as a 1-by-*p* cell array of character vectors.

Data Types: `cell`

`DFE` — Residual degrees of freedom
positive integer value

Residual degrees of freedom, stored as a positive integer value. $DFE = n - p$, where n is the number of observations, and p is the number of fixed-effects coefficients.

This corresponds to the `'Residual'` method of calculating degrees of freedom in the `fixedEffects` and `randomEffects` methods.

Data Types: `double`

`FitMethod` — Method used to fit the linear mixed-effects model
`ML` | `REML`

Method used to fit the linear mixed-effects model, stored as either of the following.

- `ML`, if the fitting method is maximum likelihood
- `REML`, if the fitting method is restricted maximum likelihood

Data Types: `char`

`Formula` — Specification of the fixed- and random-effects terms, and grouping variables

object

Specification of the fixed-effects terms, random-effects terms, and grouping variables that define the linear mixed-effects model, stored as an object.

For more information on how to specify the model to fit using a formula, see “Formula” on page 33-3540.

`LogLikelihood` — Maximized log or restricted log likelihood

scalar value

Maximized log likelihood or maximized restricted log likelihood of the fitted linear mixed-effects model depending on the fitting method you choose, stored as a scalar value.

Data Types: `double`

`ModelCriterion` — Model criterion

dataset array

Model criterion to compare fitted linear mixed-effects models, stored as a dataset array with the following columns.

`AIC`	Akaike Information Criterion
`BIC`	Bayesian Information Criterion
`Loglikelihood`	Log likelihood value of the model
`Deviance`	–2 times the log likelihood of the model

If n is the number of observations used in fitting the model, and p is the number of fixed-effects coefficients, then for calculating AIC and BIC,

- The total number of parameters is $nc + p + 1$, where nc is the total number of parameters in the random-effects covariance excluding the residual variance
- The effective number of observations is
  - n, when the fitting method is maximum likelihood (ML)
  - $n – p$, when the fitting method is restricted maximum likelihood (REML)

`MSE` — ML or REML estimate

positive scalar value

ML or REML estimate, based on the fitting method used for estimating σ^2, stored as a positive scalar value. σ^2 is the residual variance or variance of the observation error term of the linear mixed-effects model.

Data Types: `double`

`NumCoefficients` — Number of fixed-effects coefficients

positive integer value

Number of fixed-effects coefficients in the fitted linear mixed-effects model, stored as a positive integer value.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated fixed-effects coefficients
positive integer value

Number of estimated fixed-effects coefficients in the fitted linear mixed-effects model, stored as a positive integer value.

Data Types: `double`

`NumObservations` — Number of observations
positive integer value

Number of observations used in the fit, stored as a positive integer value. This is the number of rows in the table or dataset array, or the design matrices minus the excluded rows or rows with `NaN` values.

Data Types: `double`

`NumPredictors` — Number of predictors
positive integer value

Number of variables used as predictors in the linear mixed-effects model, stored as a positive integer value.

Data Types: `double`

`NumVariables` — Total number of variables
positive integer value

Total number of variables including the response and predictors, stored as a positive integer value.

- If the sample data is in a table or dataset array `tbl`, `NumVariables` is the total number of variables in `tbl` including the response variable.
- If the fit is based on matrix input, `NumVariables` is the total number of columns in the predictor matrix or matrices, and response vector.

`NumVariables` includes variables, if there are any, that are not used as predictors or as the response.

Data Types: `double`

`ObservationInfo` — Information about the observations
table

Information about the observations used in the fit, stored as a table.

`ObservationInfo` has one row for each observation and the following four columns.

`Weights`	The value of the weighted variable for that observation. Default value is 1.
`Excluded`	`true`, if the observation was excluded from the fit using the `'Exclude'` name-value pair argument, `false`, otherwise. 1 stands for `true` and 0 stands for `false`.

`Missing`	`true`, if the observation was excluded from the fit because any response or predictor value is missing, `false`, otherwise. Missing values include `NaN` for numeric variables, empty cells for cell arrays, blank rows for character arrays, and the `<undefined>` value for categorical arrays.
`Subset`	`true`, if the observation was used in the fit, `false`, if it was not used because it is missing or excluded.

Data Types: `table`

`ObservationNames` — Names of observations
cell array of character vectors

Names of observations used in the fit, stored as a cell array of character vectors.

- If the data is in a table or dataset array, `tbl`, containing observation names, `ObservationNames` has those names.
- If the data is provided in matrices, or a table or dataset array without observation names, then `ObservationNames` is an empty cell array.

Data Types: `cell`

`PredictorNames` — Names of predictors
cell array of character vectors

Names of the variables that you use as predictors in the fit, stored as a cell array of character vectors that has the same length as `NumPredictors`.

Data Types: `cell`

`ResponseName` — Names of response variable
character vector

Name of the variable used as the response variable in the fit, stored as a character vector.

Data Types: `char`

`Rsquared` — Proportion of variability in the response explained by the fitted model
structure

Proportion of variability in the response explained by the fitted model, stored as a structure. It is the multiple correlation coefficient or R-squared. `Rsquared` has two fields.

`Ordinary`	R-squared value, stored as a scalar value in a structure. `Rsquared.Ordinary = 1 – SSE./SST`

`Adjusted`	R-squared value adjusted for the number of fixed-effects coefficients, stored as a scalar value in a structure. `Rsquared.Adjusted = 1 – (SSE./SST)*(DFT./DFE)`, where `DFE = n – p`, `DFT = n – 1`, and `n` is the total number of observations, `p` is the number of fixed-effects coefficients.

Data Types: `struct`

SSE — Error sum of squares
positive scalar value

Error sum of squares, that is, sum of the squared conditional residuals, stored as a positive scalar value.

`SSE = sum((y – F).^2)`, where `y` is the response vector, and `F` is the fitted conditional response of the linear mixed-effects model. The conditional model has contributions from both fixed and random effects.

Data Types: `double`

SSR — Regression sum of squares
positive scalar value

Regression sum of squares, that is, the sum of squares explained by the linear mixed-effects regression, stored as a positive scalar value. It is the sum of squared deviations of the conditional fitted values from their mean.

`SSR = sum((F – mean(F)).^2)`, where `F` is the fitted conditional response of the linear mixed-effects model. The conditional model has contributions from both fixed and random effects.

Data Types: `double`

SST — Total sum of squares
positive scalar value

Total sum of squares, that is, the sum of the squared deviations of the observed response values from their mean, stored as a positive scalar value.

`SST = sum((y – mean(y)).^2) = SSR + SSE`, where `y` is the response vector.

Data Types: `double`

`Variables` — Variables
table

Variables, stored as a table.

- If the fit is based on a table or dataset array `tbl`, then `Variables` is identical to `tbl`.
- If the fit is based on matrix input, then `Variables` is a table containing all the variables in the predictor matrix or matrices, and response variable.

Data Types: `table`

`VariableInfo` — Information about the variables

table

Information about the variables used in the fit, stored as a table.

`VariableInfo` has one row for each variable and contains the following four columns.

`Class`	Class of the variable (`'double'`, `'cell'`, `'nominal'`, and so on).
`Range`	Value range of the variable. • For a numerical variable, it is a two-element vector of the form `[min,max]`. • For a cell or categorical variable, it is a cell or categorical array containing all unique values of the variable.
`InModel`	`true`, if the variable is a predictor in the fitted model. `false`, if the variable is not in the fitted model.
`IsCategorical`	`true`, if the variable has a type that is treated as a categorical predictor, such as cell, logical, or categorical, or if it is specified as categorical by the `'Categorical'` name-value pair argument of the `fit` method. `false`, if it is a continuous predictor.

Data Types: `table`

`VariableNames` — Names of the variables

cell array of character vectors

Names of the variables used in the fit, stored as a cell array of character vectors.

- If sample data is in a table or dataset array `tbl`, `VariableNames` contains the names of the variables in `tbl`.
- If sample data is in matrix format, then `VariableInfo` includes variable names you supply while fitting the model. If you do not supply the variable names, then `VariableInfo` contains the default names.

Data Types: `cell`

Object Functions

anova	Analysis of variance for linear mixed-effects model
coefCI	Confidence intervals for coefficients of linear mixed-effects model
coefTest	Hypothesis test on fixed and random effects of linear mixed-effects model
compare	Compare linear mixed-effects models
covarianceParameters	Extract covariance parameters of linear mixed-effects model
designMatrix	Fixed- and random-effects design matrices

fitted	Fitted responses from a linear mixed-effects model
fixedEffects	Estimates of fixed effects and related statistics
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotResiduals	Plot residuals of linear mixed-effects model
predict	Predict response of linear mixed-effects model
random	Generate random responses from fitted linear mixed-effects model
randomEffects	Estimates of random effects and related statistics
residuals	Residuals of fitted linear mixed-effects model
response	Response vector of the linear mixed-effects model

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Random Intercept Model with Categorical Predictor

Load the sample data.

```
load flu
```

The `flu` dataset array has a `Date` variable, and 10 variables containing estimated influenza rates (in 9 different regions, estimated from Google® searches, plus a nationwide estimate from the Center for Disease Control and Prevention, CDC).

To fit a linear-mixed effects model, your data must be in a properly formatted dataset array. To fit a linear mixed-effects model with the influenza rates as the responses and region as the predictor variable, combine the nine columns corresponding to the regions into an array. The new dataset array, `flu2`, must have the response variable, `FluRate`, the nominal variable, `Region`, that shows which region each estimate is from, and the grouping variable `Date`.

```
flu2 = stack(flu,2:10,'NewDataVarName','FluRate',...
    'IndVarName','Region');
flu2.Date = nominal(flu2.Date);
```

Fit a linear mixed-effects model with fixed effects for region and a random intercept that varies by `Date`.

Because region is a nominal variable, `fitlme` takes the first region, `NE`, as the reference and creates eight dummy variables representing the other eight regions. For example, I[MidAtl] is the dummy variable representing the region `MidAtl`. For details, see “Dummy Variables” on page 2-48.

The corresponding model is

$$y_{im} = \beta_0 + \beta_1 I[\text{MidAtl}]_i + \beta_2 I[\text{ENCentral}]_i + \beta_3 I[\text{WNCentral}]_i + \beta_4 I[\text{SAtl}]_i + \beta_5 I[\text{ESCentral}]_i + \beta_6 I[\text{WSCentral}]_i + \beta_7 I[\text{Mtn}]_i + \beta_8 I[\text{Pac}]_i + b_{0m} + \varepsilon_{im}, \quad m = 1, 2, \ldots, 52,$$

where y_{im} is the observation i for level m of grouping variable `Date`, β_j, $j = 0, 1, \ldots, 8$, are the fixed-effects coefficients, b_{0m} is the random effect for level m of the grouping variable `Date`, and ε_{im} is the

observation error for observation i. The random effect has the prior distribution, $b_{0m} \sim N(0, \sigma_b^2)$ and the error term has the distribution, $\varepsilon_{im} \sim N(0, \sigma^2)$.

```
lme = fitlme(flu2,'FluRate ~ 1 + Region + (1|Date)')
```

```
lme =
Linear mixed-effects model fit by ML

Model information:
    Number of observations             468
    Fixed effects coefficients           9
    Random effects coefficients         52
    Covariance parameters                2

Formula:
    FluRate ~ 1 + Region + (1 | Date)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    318.71    364.35    -148.36          296.71

Fixed effects coefficients (95% CIs):
    Name                        Estimate    SE          tStat      DF
    {'(Intercept)'     }          1.2233    0.096678     12.654    459
    {'Region_MidAtl'   }        0.010192    0.052221    0.19518    459
    {'Region_ENCentral'}        0.051923    0.052221     0.9943    459
    {'Region_WNCentral'}         0.23687    0.052221     4.5359    459
    {'Region_SAtl'     }        0.075481    0.052221     1.4454    459
    {'Region_ESCentral'}         0.33917    0.052221      6.495    459
    {'Region_WSCentral'}           0.069    0.052221     1.3213    459
    {'Region_Mtn'      }        0.046673    0.052221    0.89377    459
    {'Region_Pac'      }        -0.16013    0.052221    -3.0665    459


    pValue        Lower        Upper
     1.085e-31       1.0334       1.4133
       0.84534    -0.092429      0.11281
        0.3206    -0.050698      0.15454
    7.3324e-06      0.13424      0.33949
       0.14902     -0.02714       0.1781
    2.1623e-10      0.23655      0.44179
       0.18705    -0.033621      0.17162
       0.37191    -0.055948      0.14929
     0.0022936     -0.26276    -0.057514

Random effects covariance parameters (95% CIs):
Group: Date (52 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.6443


    Lower     Upper
    0.5297    0.78368

Group: Error
    Name               Estimate    Lower      Upper
```

```
    {'Res Std'}        0.26627     0.24878    0.285
```

The *p*-values 7.3324e-06 and 2.1623e-10 respectively show that the fixed effects of the flu rates in regions `WNCentral` and `ESCentral` are significantly different relative to the flu rates in region `NE`.

The confidence limits for the standard deviation of the random-effects term, σ_b, do not include 0 (0.5297, 0.78368), which indicates that the random-effects term is significant. You can also test the significance of the random-effects terms using the `compare` method.

The estimated value of an observation is the sum of the fixed effects and the random-effect value at the grouping variable level corresponding to that observation. For example, the estimated best linear unbiased predictor (BLUP) of the flu rate for region `WNCentral` in week 10/9/2005 is

$$\begin{aligned}\hat{y}_{\text{WNCentral, 10/9/2005}} &= \hat{\beta}_0 + \hat{\beta}_3 I[\text{WNCentral}] + \hat{b}_{10/9/2005} \\ &= 1.2233 + 0.23687 - 0.1718 \\ &= 1.28837.\end{aligned}$$

This is the fitted conditional response, since it includes contribution to the estimate from both the fixed and random effects. You can compute this value as follows.

```
beta = fixedEffects(lme);
[~,~,STATS] = randomEffects(lme); % Compute the random-effects statistics (STATS)
STATS.Level = nominal(STATS.Level);
y_hat = beta(1) + beta(4) + STATS.Estimate(STATS.Level=='10/9/2005')
```

```
y_hat = 1.2884
```

You can simply display the fitted value using the `fitted` method.

```
F = fitted(lme);
F(flu2.Date == '10/9/2005' & flu2.Region == 'WNCentral')
```

```
ans = 1.2884
```

Compute the fitted marginal response for region `WNCentral` in week 10/9/2005.

```
F = fitted(lme,'Conditional',false);
F(flu2.Date == '10/9/2005' & flu2.Region == 'WNCentral')
```

```
ans = 1.4602
```

Linear Mixed-Effects Model with a Random Slope

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower and the cylinders, and uncorrelated random-effect for intercept and acceleration grouped by the model year. This model corresponds to

$$\text{MPG}_{im} = \beta_0 + \beta_1 \text{Acc}_i + \beta_2 \text{HP} + b_{0m} + b_{1m}\text{Acc}_{im} + \varepsilon_{im}, \quad m = 1, 2, 3,$$

with the random-effects terms having the following prior distributions:

$$b_m = \begin{pmatrix} b_{0m} \\ b_{1m} \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{pmatrix}\right),$$

where m represents the model year.

First, prepare the design matrices for fitting the linear mixed-effects model.

```
X = [ones(406,1) Acceleration Horsepower];
Z = [ones(406,1) Acceleration];
Model_Year = nominal(Model_Year);
G = Model_Year;
```

Now, fit the model using `fitlmematrix` with the defined design matrices and grouping variables. Use the `'fminunc'` optimization algorithm.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'}},'RandomEffectGroups',{'Model_Year'},...
'FitMethod','REML')
```

```
lme =
Linear mixed-effects model fit by REML

Model information:
    Number of observations             392
    Fixed effects coefficients           3
    Random effects coefficients         26
    Covariance parameters                4

Formula:
    Linear Mixed Formula with 4 predictors.

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    2202.9    2230.7    -1094.5          2188.9

Fixed effects coefficients (95% CIs):
    Name                    Estimate    SE           tStat      DF
    {'Intercept'   }          50.064       2.3176     21.602    389
    {'Acceleration'}        -0.57897      0.13843    -4.1825    389
    {'Horsepower'  }        -0.16958    0.0073242    -23.153    389


    pValue        Lower       Upper
    1.4185e-68      45.507       54.62
    3.5654e-05    -0.85112    -0.30681
    3.5289e-75    -0.18398    -0.15518

Random effects covariance parameters (95% CIs):
Group: Model_Year (13 Levels)
    Name1                   Name2                   Type            Estimate
    {'Intercept'   }        {'Intercept'   }        {'std' }           3.72
    {'Acceleration'}        {'Intercept'   }        {'corr'}        -0.8769
    {'Acceleration'}        {'Acceleration'}        {'std' }         0.3593
```

```
    Lower       Upper
      1.5215      9.0954
    -0.98275    -0.33845
     0.19418     0.66483

Group: Error
    Name               Estimate    Lower     Upper
    {'Res Std'}        3.6913      3.4331    3.9688
```

The fixed effects coefficients display includes the estimate, standard errors (`SE`), and the 95% confidence interval limits (`Lower` and `Upper`). The *p*-values for (`pValue`) indicate that all three fixed-effects coefficients are significant.

The confidence intervals for the standard deviations and the correlation between the random effects for intercept and acceleration do not include zeros, hence they seem significant. Use the `compare` method to test for the random effects.

Display the covariance matrix of the estimated fixed-effects coefficients.

```
lme.CoefficientCovariance
```

```
ans = 3×3

    5.3711   -0.2809   -0.0126
   -0.2809    0.0192    0.0005
   -0.0126    0.0005    0.0001
```

The diagonal elements show the variances of the fixed-effects coefficient estimates. For example, the variance of the estimate of the intercept is 5.3711. Note that the standard errors of the estimates are the square roots of the variances. For example, the standard error of the intercept is 2.3176, which is `sqrt(5.3711)`.

The off-diagonal elements show the correlation between the fixed-effects coefficient estimates. For example, the correlation between the intercept and acceleration is –0.2809 and the correlation between acceleration and horsepower is 0.0005.

Display the coefficient of determination for the model.

```
lme.Rsquared
```

```
ans = struct with fields:
    Ordinary: 0.7866
    Adjusted: 0.7855
```

The adjusted value is the R-squared value adjusted for the number of predictors in the model.

More About

Formula

In general, a formula for model specification is a character vector or string scalar of the form `'y ~ terms'`. For the linear mixed-effects models, this formula is in the form `'y ~ fixed + (random1|`

`grouping1) + ... + (randomR|groupingR)'`, where `fixed` and `random` contain the fixed-effects and the random-effects terms.

Suppose a table `tbl` contains the following:

- A response variable, `y`
- Predictor variables, X_j, which can be continuous or grouping variables
- Grouping variables, g_1, g_2, ..., g_R,

where the grouping variables in X_j and g_r can be categorical, logical, character arrays, string arrays, or cell arrays of character vectors.

Then, in a formula of the form, `'y ~ fixed + (random`$_1$`|g`$_1$`) + ... + (random`$_R$`|g`$_R$`)'`, the term `fixed` corresponds to a specification of the fixed-effects design matrix `X`, `random`$_1$ is a specification of the random-effects design matrix Z_1 corresponding to grouping variable g_1, and similarly `random`$_R$ is a specification of the random-effects design matrix Z_R corresponding to grouping variable g_R. You can express the `fixed` and `random` terms using Wilkinson notation.

Wilkinson notation describes the factors present in models. The notation relates to factors present in models, not to the multipliers (coefficients) of those factors.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	`X`, X^2, ..., X^k
`X1 + X2`	`X1`, `X2`
`X1*X2`	`X1`, `X2`, `X1.*X2 (elementwise multiplication of X1 and X2)`
`X1:X2`	`X1.*X2` only
`- X2`	Do not include `X2`
`X1*X2 + X3`	`X1`, `X2`, `X3`, `X1*X2`
`X1 + X2 + X3 + X1:X2`	`X1`, `X2`, `X3`, `X1*X2`
`X1*X2*X3 - X1:X2:X3`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`, `X2*X3`
`X1*(X2 + X3)`	`X1`, `X2`, `X3`, `X1*X2`, `X1*X3`

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`. Here are some examples for linear mixed-effects model specification.

Examples:

Formula	Description
`'y ~ X1 + X2'`	Fixed effects for the intercept, `X1` and `X2`. This is equivalent to `'y ~ 1 + X1 + X2'`.
`'y ~ -1 + X1 + X2'`	No intercept and fixed effects for `X1` and `X2`. The implicit intercept term is suppressed by including `-1`.

Formula	Description
`'y ~ 1 + (1 | g1)'`	Fixed effects for the intercept plus random effect for the intercept for each level of the grouping variable `g1`.
`'y ~ X1 + (1 | g1)'`	Random intercept model with a fixed slope.
`'y ~ X1 + (X1 | g1)'`	Random intercept and slope, with possible correlation between them. This is equivalent to `'y ~ 1 + X1 + (1 + X1|g1)'`.
`'y ~ X1 + (1 | g1) + (-1 + X1 | g1)'`	Independent random effects terms for intercept and slope.
`'y ~ 1 + (1 | g1) + (1 | g2) + (1 | g1:g2)'`	Random intercept model with independent main effects for `g1` and `g2`, plus an independent interaction effect.

See Also

`fitlme` | `fitlmematrix`

linhyptest

Linear hypothesis test

Syntax

```
p = linhyptest(beta,COVB,c,H,dfe)
[p,t,r] = linhyptest(...)
```

Description

`p = linhyptest(beta,COVB,c,H,dfe)` returns the *p* value `p` of a hypothesis test on a vector of parameters. `beta` is a vector of *k* parameter estimates. `COVB` is the *k*-by-*k* estimated covariance matrix of the parameter estimates. `c` and `H` specify the null hypothesis in the form `H*b = c`, where `b` is the vector of unknown parameters estimated by `beta`. `dfe` is the degrees of freedom for the `COVB` estimate, or `Inf` if `COVB` is known rather than estimated.

`beta` is required. The remaining arguments have default values:

- `COVB = eye(k)`
- `c = zeros(k,1)`
- `H = eye(K)`
- `dfe = Inf`

If `H` is omitted, `c` must have *k* elements and it specifies the null hypothesis values for the entire parameter vector.

Note The following functions return outputs suitable for use as the `COVB` input argument to `linhyptest`: `nlinfit`, `coxphfit`, `glmfit`, `mnrfit`, `regstats`, `robustfit`. `nlinfit` returns `COVB` directly; the other functions return `COVB` in `stats.covb`.

`[p,t,r] = linhyptest(...)` also returns the test statistic `t` and the rank `r` of the hypothesis matrix `H`. If `dfe` is `Inf` or is not given, `t*r` is a chi-square statistic with `r` degrees of freedom . If `dfe` is specified as a finite value, `t` is an *F* statistic with `r` and `dfe` degrees of freedom.

`linhyptest` performs a test based on an asymptotic normal distribution for the parameter estimates. It can be used after any estimation procedure for which the parameter covariances are available, such as `regstats` or `glmfit`. For linear regression, the *p*-values are exact. For other procedures, the *p*-values are approximate, and may be less accurate than other procedures such as those based on a likelihood ratio.

Examples

Fit a multiple linear model to the data in `hald.mat`:

```
load hald
stats = regstats(heat,ingredients,'linear');
beta = stats.beta
```

```
beta =
   62.4054
    1.5511
    0.5102
    0.1019
   -0.1441
```

Perform an *F*-test that the last two coefficients are both 0:

```
SIGMA = stats.covb;
dfe = stats.fstat.dfe;
H = [0 0 0 1 0;0 0 0 0 1];
c = [0;0];
[p,F] = linhyptest(beta,SIGMA,c,H,dfe)
p =
    0.4668
F =
    0.8391
```

See Also

`coxphfit` | `glmfit` | `mnrfit` | `nlinfit` | `regstats` | `robustfit`

Introduced in R2007a

linhyptest

Linear hypothesis tests on Cox model coefficients

Syntax

```
testTable = linhyptest(coxMdl)
```

Description

`testTable = linhyptest(coxMdl)` returns an ANOVA-style table with *p*-values for tests that determine if sequential combinations of Cox model coefficient estimates are zero. `linhyptest` tests successive null hypotheses, starting with the hypothesis that all the coefficients are 0. The function then tests to determine if all but the first coefficient are 0, all but the first two coefficients are zero, and so on, up to the number of coefficients minus one. A significant *p*-value indicates that you can reject the null hypothesis, meaning the assumption that all coefficients in a particular combination of coefficients are 0.

Examples

Perform Linear Hypothesis Test

Examine the result of `linhyptest` on the `readmissiontimes` data set.

```
load readmissiontimes
coxMdl = fitcox([Age,Sex,Weight],ReadmissionTime,...
    'Censoring',Censored);
testTable = linhyptest(coxMdl)
```

```
testTable=3×2 table
       Predictor          pValue
    ________________    __________

    {'Empty Model'}    2.5612e-07
    {'X1'         }    7.9753e-08
    {'X1, X2'     }      0.095973
```

- The first row of the returned table indicates that you can reject the hypothesis that all model coefficients are 0 at the `.05` or `.01` significance levels.
- The second row indicates that you can reject the hypothesis that only the `Sex` and `Weight` coefficients are 0 at the `.05` or `.01` significance levels.
- The third row indicates that you cannot reject the hypothesis that only the `Weight` coefficient is 0 at the `.05` or `.01` significance levels.

Input Arguments

`coxMdl` — Fitted Cox proportional hazards model

`CoxModel` object

Fitted Cox proportional hazards model, specified as a `CoxModel` object. Create `coxMdl` using `fitcox`.

Output Arguments

`testTable` — Significance levels of cumulative hypothesis tests

table

Significance levels of cumulative hypothesis tests, returned as a table. The tested coefficients are in the `Coefficients` property of the model. The table returns tests of successive null hypotheses, starting with the hypothesis that all the coefficients are 0. The second row tests whether all but the first coefficient is 0. The third row tests whether all but the first two coefficients are zero, and so on. The last row tests whether all coefficients but the last are zero. A significant *p*-value indicates that you can reject the null hypothesis, meaning the assumption that all coefficients in a particular combination of coefficients are 0. A significant *p*-value is one that is smaller than a specified significance level.

See Also

`CoxModel` | `coefci` | `fitcox`

Introduced in R2021a

linkage

Agglomerative hierarchical cluster tree

Syntax

```
Z = linkage(X)
Z = linkage(X,method)
Z = linkage(X,method,metric)
Z = linkage(X,method,metric,'savememory',value)

Z = linkage(X,method,pdist_inputs)

Z = linkage(y)
Z = linkage(y,method)
```

Description

`Z = linkage(X)` returns a matrix `Z` that encodes a tree containing hierarchical clusters of the rows of the input data matrix `X`.

`Z = linkage(X,method)` creates the tree using the specified `method`, which describes how to measure the distance between clusters. For more information, see “Linkages” on page 33-3554.

`Z = linkage(X,method,metric)` performs clustering by passing `metric` to the `pdist` function, which computes the distance between the rows of `X`.

`Z = linkage(X,method,metric,'savememory',value)` uses a memory-saving algorithm when `value` is `'on'`, and uses the standard algorithm when `value` is `'off'`.

`Z = linkage(X,method,pdist_inputs)` passes `pdist_inputs` to the `pdist` function, which computes the distance between the rows of `X`. The `pdist_inputs` argument consists of the `'seuclidean'`, `'minkowski'`, or `'mahalanobis'` metric and an additional distance metric option.

`Z = linkage(y)` uses a vector representation `y` of a distance matrix. `y` is either computed by `pdist` or is a more general dissimilarity matrix conforming to the output format of `pdist`.

`Z = linkage(y,method)` creates the tree using the specified `method`, which describes how to measure the distance between clusters.

Examples

Cluster Data and Plot Result

Randomly generate sample data with 20,000 observations.

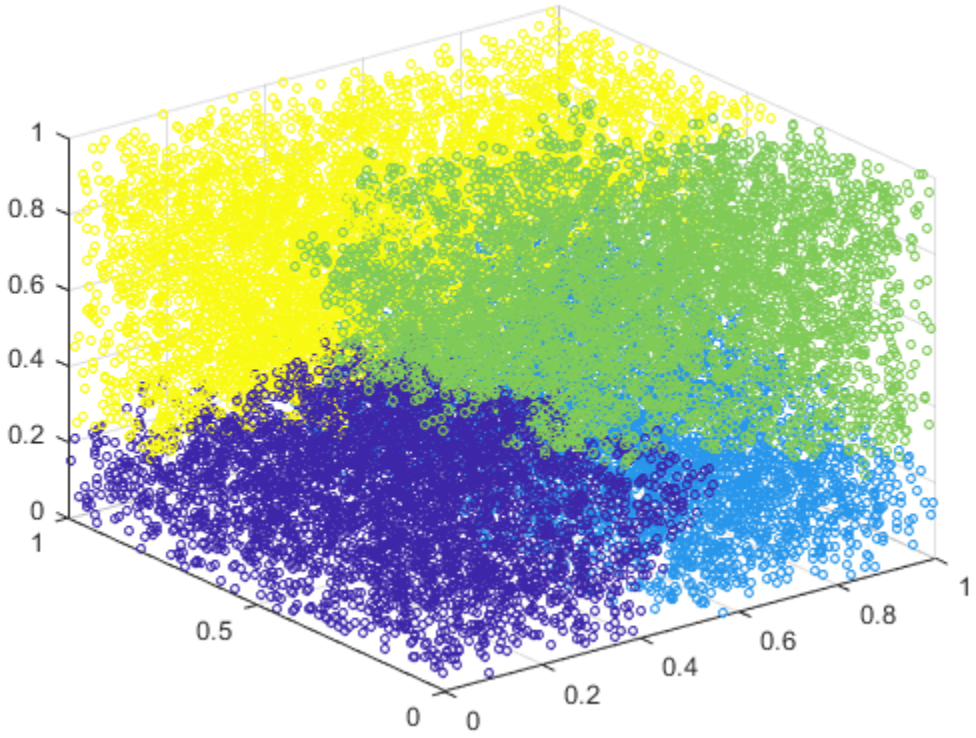
```
rng('default') % For reproducibility
X = rand(20000,3);
```

Create a hierarchical cluster tree using the `ward` linkage method. In this case, the `'SaveMemory'` option of the `clusterdata` function is set to `'on'` by default. In general, specify the best value for `'SaveMemory'` based on the dimensions of `X` and the available memory.

```
Z = linkage(X,'ward');
```

Cluster the data into a maximum of four groups and plot the result.

```
c = cluster(Z,'Maxclust',4);
scatter3(X(:,1),X(:,2),X(:,3),10,c)
```



`cluster` identifies four groups in the data.

Compare Cluster Assignments to Classes

Find a maximum of three clusters in the `fisheriris` data set and compare cluster assignments of the flowers to their known classification.

Load the sample data.

```
load fisheriris
```

Create a hierarchical cluster tree using the `'average'` method and the `'chebychev'` metric.

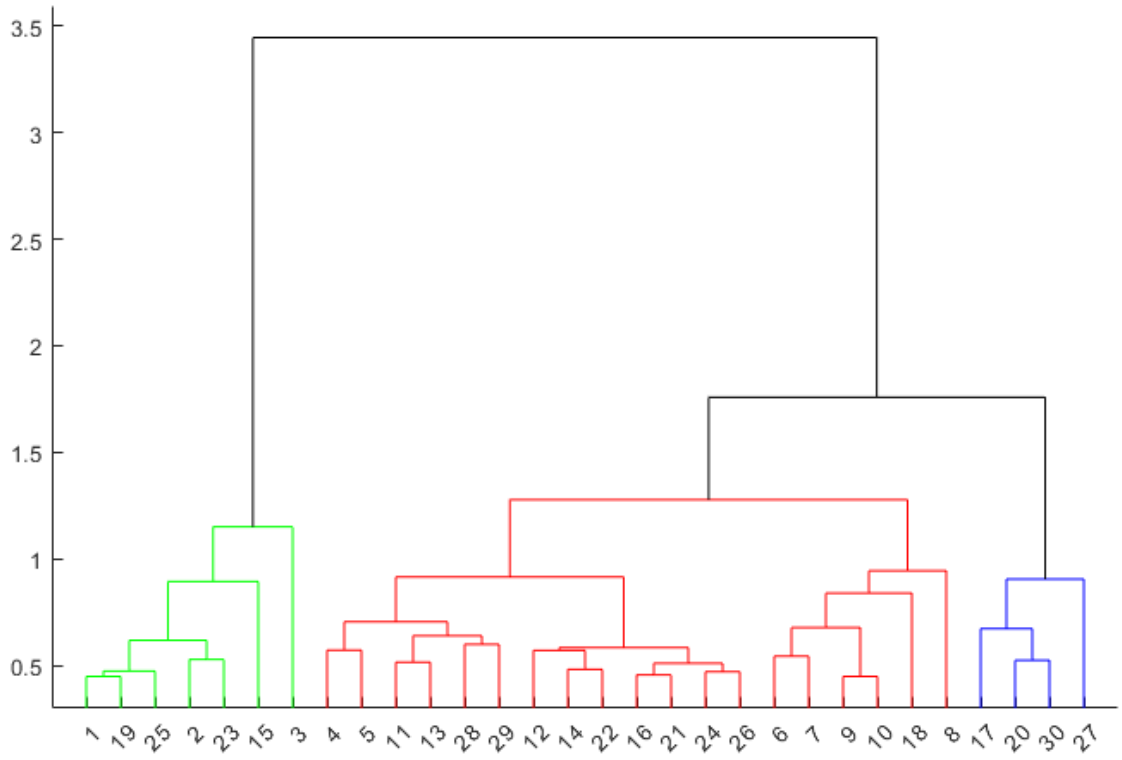
```
Z = linkage(meas,'average','chebychev');
```

Find a maximum of three clusters in the data.

```
T = cluster(Z,'maxclust',3);
```

Create a dendrogram plot of `Z`. To see the three clusters, use `'ColorThreshold'` with a cutoff halfway between the third-from-last and second-from-last linkages.

```
cutoff = median([Z(end-2,3) Z(end-1,3)]);
dendrogram(Z,'ColorThreshold',cutoff)
```



Display the last two rows of `Z` to see how the three clusters are combined into one. `linkage` combines the 293rd (blue) cluster with the 297th (red) cluster to form the 298th cluster with a linkage of `1.7583`. `linkage` then combines the 296th (green) cluster with the 298th cluster.

```
lastTwo = Z(end-1:end,:)
```

```
lastTwo = 2×3

  293.0000  297.0000    1.7583
  296.0000  298.0000    3.4445
```

See how the cluster assignments correspond to the three species. For example, one of the clusters contains `50` flowers of the second species and `40` flowers of the third species.

```
crosstab(T,species)
```

```
ans = 3×3

     0     0    10
     0    50    40
    50     0     0
```

Observe Clustering Step in Hierarchical Tree

Load the `examgrades` data set.

```
load examgrades
```

Create a hierarchical tree using `linkage`. Use the `'single'` method and the Minkowski metric with an exponent of 3.

```
Z = linkage(grades,'single',{'minkowski',3});
```

Observe the 25th clustering step.

```
Z(25,:)
```

```
ans = 1×3

   86.0000  137.0000    4.5307
```

`linkage` combines the 86th observation and the 137th cluster to form a cluster of index 120 + 25 = 145, where 120 is the total number of observations in `grades` and 25 is the row number in `Z`. The shortest distance between the 86th observation and any of the points in the 137th cluster is `4.5307`.

Cluster Data Using Dissimilarity Matrix

Create an agglomerative hierarchical cluster tree using a dissimilarity matrix.

Take a dissimilarity matrix `X` and convert it to a vector form that `linkage` accepts by using `squareform`.

```
X = [0 1 2 3; 1 0 4 5; 2 4 0 6; 3 5 6 0];
y = squareform(X);
```

Create a cluster tree using `linkage` with the `'complete'` method of calculating the distance between clusters. The first two columns of `Z` show how `linkage` combines clusters. The third column of `Z` gives the distance between clusters.

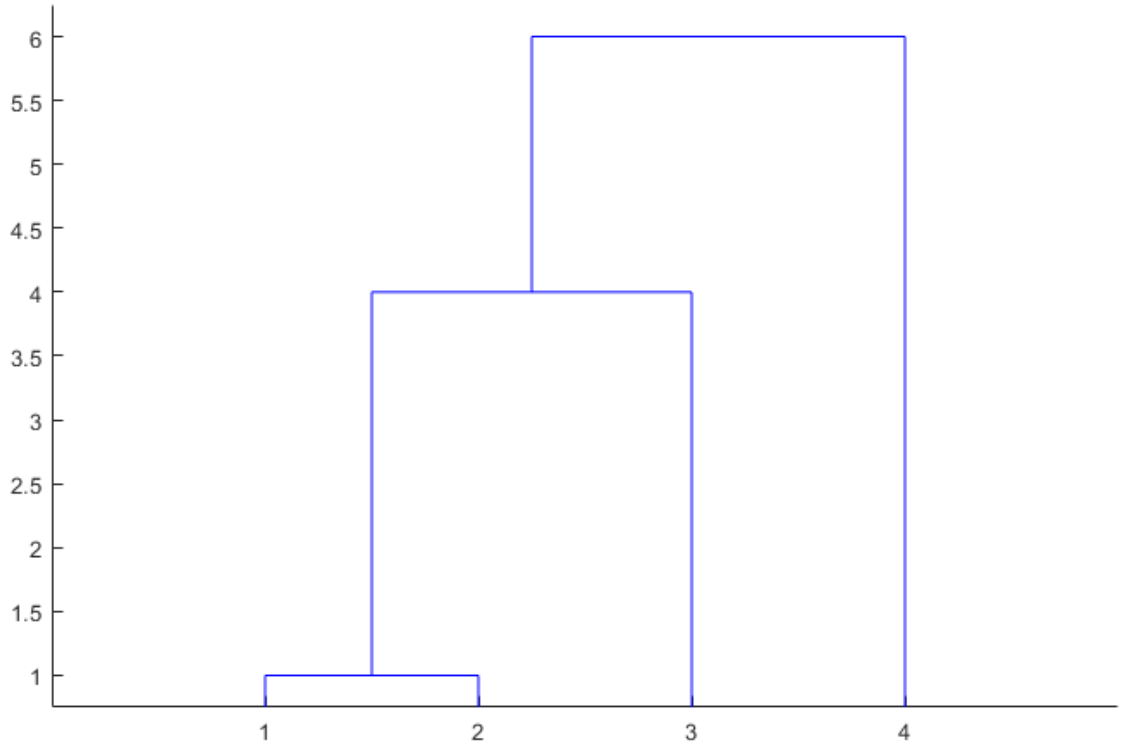
```
Z = linkage(y,'complete')
```

```
Z = 3×3

     1     2     1
     3     5     4
```

```
    4     6     6
```

Create a dendrogram plot of Z. The x-axis corresponds to the leaf nodes of the tree, and the y-axis corresponds to the linkage distances between clusters.

```
dendrogram(Z)
```



Input Arguments

X — Input data

numeric matrix

Input data, specified as a numeric matrix with two or more rows. The rows represent observations, and the columns represent categories or dimensions.

Data Types: `single` | `double`

`method` — Algorithm for computing distance between clusters

`'single'` (default) | `'average'` | `'centroid'` | `'complete'` | ...

Algorithm for computing the distance between clusters, specified as one of the values in this table.

Method	Description
'average'	Unweighted average distance (UPGMA)
'centroid'	Centroid distance (UPGMC), appropriate for Euclidean distances only
'complete'	Farthest distance
'median'	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
'single'	Shortest distance
'ward'	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
'weighted'	Weighted average distance (WPGMA)

For more information on these methods, see “Linkages” on page 33-3554.

`metric` — Distance metric

`'euclidean'` (default) | `'squaredeuclidean'` | `'seuclidean'` | `'mahalanobis'` | function handle | ...

Distance metric, specified as any metric accepted by the `pdist` function. These metrics are described in the following table.

Value	Description
'euclidean'	Euclidean distance (default).
'squaredeuclidean'	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`. Use `DistParameter` to specify another value for `S`.
'mahalanobis'	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use `DistParameter` to specify another value for `C`, where the matrix `C` is symmetric and positive definite.
'cityblock'	City block distance.
'minkowski'	Minkowski distance. The default exponent is 2. Use `DistParameter` to specify a different exponent `P`, where `P` is a positive scalar value of the exponent.
'chebychev'	Chebychev distance (maximum coordinate difference).
'cosine'	One minus the cosine of the included angle between points (treated as vectors).
'correlation'	One minus the sample correlation between points (treated as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

Value	Description

`@distfun`	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a 1-by-n vector containing a single observation.
	• `ZJ` is an m2-by-n matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an m2-by-1 vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For more information, see “Distance Metrics” on page 33-4495.

Use `pdist_inputs` instead of `metric` to specify the additional input argument `DistParameter` of `pdist` for `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

Data Types: `char` | `string` | `function_handle`

`pdist_inputs` — Distance metric and distance metric option
cell array

Distance metric and distance metric option, specified as a cell array of the comma-separated pair consisting of the two input arguments `Distance` and `DistParameter` of the function `pdist`. This argument is valid only for specifying `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

Example: `{'minkowski',5}`

Data Types: `cell`

`value` — Flag for `'savememory'` option
`'on'` | `'off'`

Flag for the `'savememory'` option, specified as either `'on'` or `'off'`. The `'on'` setting causes `linkage` to construct clusters without computing the distance matrix. The `'on'` setting is available only when `method` is `'centroid'`, `'median'`, or `'ward'` and `metric` is `'euclidean'`.

When `value` is `'on'`, the `linkage` run time is proportional to the number of dimensions (number of columns of `X`). When `value` is `'off'`, the `linkage` memory requirement is proportional to N^2, where N is the number of observations. The best (least-time) setting to use for `value` depends on the problem dimensions, number of observations, and available memory. The default `value` setting is a rough approximation of an optimal setting.

The default is `'on'` when `X` has 20 columns or fewer, or the computer does not have enough memory to store the distance matrix. Otherwise, the default is `'off'`.

Example: `'savememory','on'`

`y` — Distances
numeric vector

Distances, specified as a numeric vector with the same format as the output of the `pdist` function:

- A row vector of length $m(m – 1)/2$, corresponding to pairs of observations in a matrix with m rows
- Distances arranged in the order (2,1), (3,1), ..., (m,1), (3,2), ..., (m,2), ..., (m,m – 1))

`y` can be a more general dissimilarity matrix conforming to the output format of `pdist`.

Data Types: `single` | `double`

Output Arguments

Z — Agglomerative hierarchical cluster tree

numeric matrix

Agglomerative hierarchical cluster tree, returned as a numeric matrix. `Z` is an (m – 1)-by-3 matrix, where m is the number of observations in the original data. Columns 1 and 2 of `Z` contain cluster indices linked in pairs to form a binary tree. The leaf nodes are numbered from 1 to m. Leaf nodes are the singleton clusters from which all higher clusters are built. Each newly formed cluster, corresponding to row `Z(I,:)`, is assigned the index m + `I`. The entries `Z(I,1)` and `Z(I,2)` contain the indices of the two component clusters that form cluster m + `I`. The m – 1 higher clusters correspond to the interior nodes of the clustering tree. `Z(I,3)` contains the linkage distance between the two clusters merged in row `Z(I,:)`.

For example, consider building a tree with 30 initial nodes. Suppose that cluster 5 and cluster 7 are combined at step 12, and that the distance between them at that step is 1.5. Then `Z(12,:)` is `[5 7 1.5]`. The newly formed cluster has index 12 + 30 = 42. If cluster 42 appears in a later row, then the function is combining the cluster created at step 12 into a larger cluster.

Data Types: `single` | `double`

More About

Linkages

A *linkage* is the distance between two clusters.

The following notation describes the linkages used by the various methods:

- Cluster r is formed from clusters p and q.
- n_r is the number of objects in cluster r.
- x_{ri} is the ith object in cluster r.
- *Single linkage*, also called *nearest neighbor*, uses the smallest distance between objects in the two clusters.

  $$d(r,s) = \min(dist(x_{ri}, x_{sj})), i \in (i, ..., n_r), j \in (1, ..., n_s)$$

- *Complete linkage*, also called *farthest neighbor*, uses the largest distance between objects in the two clusters.

  $$d(r,s) = \max(dist(x_{ri}, x_{sj})), i \in (1, ..., n_r), j \in (1, ..., n_s)$$

- *Average linkage* uses the average distance between all pairs of objects in any two clusters.

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj})$$

- *Centroid linkage* uses the Euclidean distance between the centroids of the two clusters.

  $$d(r,s) = \|\bar{x}_r - \bar{x}_s\|_2,$$

  where

  $$\bar{x}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri}$$

- *Median linkage* uses the Euclidean distance between weighted centroids of the two clusters.

  $$d(r,s) = \|\tilde{x}_r - \tilde{x}_s\|_2,$$

  where $\tilde{x}_r$ and $\tilde{x}_s$ are weighted centroids for the clusters r and s. If cluster r was created by combining clusters p and q, $\tilde{x}_r$ is defined recursively as

  $$\tilde{x}_r = \frac{1}{2}(\tilde{x}_p + \tilde{x}_q)$$

- *Ward's linkage* uses the incremental sum of squares, that is, the increase in the total within-cluster sum of squares as a result of joining two clusters. The within-cluster sum of squares is defined as the sum of the squares of the distances between all objects in the cluster and the centroid of the cluster. The sum of squares metric is equivalent to the following distance metric $d(r,s)$, which is the formula `linkage` uses.

  $$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)}} \|\bar{x}_r - \bar{x}_s\|_2,$$

  where

  - $\|\|_2$ is the Euclidean distance.
  - $\bar{x}_r$ and $\bar{x}_s$ are the centroids of clusters r and s.
  - n_r and n_s are the number of elements in clusters r and s.

  In some references, Ward's linkage does not use the factor of 2 multiplying $n_r n_s$. The `linkage` function uses this factor so that the distance between two singleton clusters is the same as the Euclidean distance.

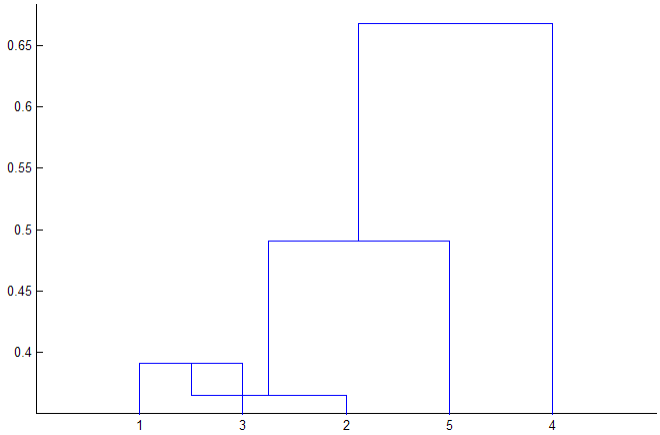
- *Weighted average linkage* uses a recursive definition for the distance between two clusters. If cluster r was created by combining clusters p and q, the distance between r and another cluster s is defined as the average of the distance between p and s and the distance between q and s.

  $$d(r,s) = \frac{(d(p,s) + d(q,s))}{2}$$

Tips

- Computing `linkage(y)` can be slow when `y` is a vector representation of the distance matrix. For the `'centroid'`, `'median'`, and `'ward'` methods, `linkage` checks whether `y` is a Euclidean distance. Avoid this time-consuming check by passing in `X` instead of `y`.

- The `'centroid'` and `'median'` methods can produce a cluster tree that is not monotonic. This result occurs when the distance from the union of two clusters, *r* and *s*, to a third cluster is less than the distance between *r* and *s*. In this case, in a dendrogram drawn with the default orientation, the path from a leaf to the root node takes some downward steps. To avoid this result, use another method. This figure shows a nonmonotonic cluster tree.



  In this case, cluster 1 and cluster 3 are joined into a new cluster, and the distance between this new cluster and cluster 2 is less than the distance between cluster 1 and cluster 3. The result is a nonmonotonic tree.

- You can provide the output `Z` to other functions including `dendrogram` to display the tree, `cluster` to assign points to clusters, `inconsistent` to compute inconsistent measures, and `cophenet` to compute the cophenetic correlation coefficient.

See Also

`cluster` | `clusterdata` | `cophenet` | `dendrogram` | `inconsistent` | `kmeans` | `pdist` | `silhouette` | `squareform`

Topics

“Hierarchical Clustering” on page 16-6

Introduced before R2006a

loadCompactModel

(To be removed) Reconstruct model object from saved model for code generation

Note `loadCompactModel` will be removed in a future release. Use `loadLearnerForCoder` instead. To update your code, simply replace instances of `loadCompactModel` with `loadLearnerForCoder`.

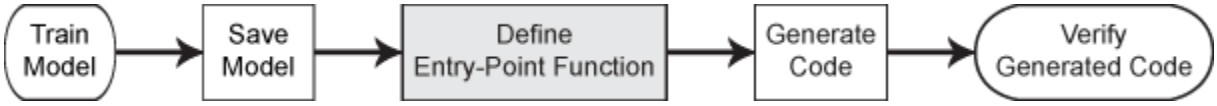
Syntax

```
Mdl = loadCompactModel(filename)
```

Description

To generate C/C++ code for the object functions (`predict`, `random`, `knnsearch`, or `rangesearch`) of machine learning models, use `saveCompactModel`, `loadCompactModel`, and `codegen`. After training a machine learning model, save the model by using `saveCompactModel`. Define an entry-point function that loads the model by using `loadCompactModel` and calls an object function. Then use `codegen` or the MATLAB Coder app to generate C/C++ code. Generating C/C++ code requires MATLAB Coder.

This flow chart shows the code generation workflow for the object functions of machine learning models. Use `loadCompactModel` for the highlighted step.



`Mdl = loadCompactModel(filename)` reconstructs a classification model, regression model, or nearest neighbor searcher (`Mdl`) from the model stored in the MATLAB formatted binary file (MAT-file) named `filename`. You must create the `filename` file by using `saveCompactModel`.

Examples

Generate C/C++ Code for Prediction

After training a machine learning model, save the model by using `saveCompactModel`. Define an entry-point function that loads the model by using `loadCompactModel` and calls the `predict` function of the trained model. Then use `codegen` (MATLAB Coder) to generate C/C++ code.

This example briefly explains the code generation workflow for the prediction of machine learning models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details. To learn about the code generation for finding nearest neighbors using a nearest neighbor searcher model, see “Code Generation for Nearest Neighbor Searcher” on page 32-19.

Train Model

Load Fisher's iris data set. Remove all observed setosa irises data so that `X` and `Y` contain data for two classes only.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,:);
Y = species(inds);
```

Train a support vector machine (SVM) classification model using the processed data set.

```
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `SVMIris.mat` by using `saveCompactModel`.

```
saveCompactModel(Mdl,'SVMIris');
```

```
Warning: saveCompactModel will be removed in a future release. Use <a href="matlab:doc saveLearne
```

Define Entry-Point Function

Define an entry-point function named `classifyIrises` that does the following:

- Accept iris flower measurements with columns corresponding to `meas`, and return predicted labels.
- Load a trained SVM classification model.
- Predict labels using the loaded classification model for the iris flower measurements.

```
type classifyIrises.m % Display contents of classifyIrises.m file
```

```
function label = classifyIrises(X) %#codegen
%CLASSIFYIRISES Classify iris species using SVM Model
%   CLASSIFYIRISES classifies the iris flower measurements in X using the
%   compact SVM model in the file SVMIris.mat, and then returns class
%   labels in label.
CompactMdl = loadCompactModel('SVMIris');
label = predict(CompactMdl,X);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point

function at compile time. Pass `X` as the value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see "Specify Variable-Size Arguments for Code Generation" on page 32-45 and "Specify Properties of Entry-Point Function Inputs" (MATLAB Coder).

```
codegen classifyIrises -args {X}
```

```
Code generation successful.
```

`codegen` generates the MEX function `classifyIrises_mex` with a platform-dependent extension.

Verify Generated Code

Compare the labels classified using `predict`, `classifyIrises`, and `classifyIrises_mex`.

```
label1 = predict(Mdl,X);
label2 = classifyIrises(X);
```

```
Warning: loadCompactModel will be removed in a future release. Use <a href="matlab:doc loadLearne
```

```
label3 = classifyIrises_mex(X);
```

```
Warning: loadCompactModel will be removed in a future release. Use <a href="matlab:doc loadLearne
```

```
verify_label = isequal(label1,label2,label3)
```

```
verify_label = logical
   1
```

`isequal` returns logical 1 (true), which means all the inputs are equal. The labels classified all three ways are the same.

Input Arguments

`filename` — Name of MAT-file that contains structure array representing a model object
character vector | string scalar

Name of the MAT-file that contains the structure array representing a model object, specified as a character vector or string scalar. You must create the `filename` file using `saveCompactModel`.

`loadCompactModel` reconstructs the model stored in the `filename` file at compile time. For supported models, see the `Mdl` input argument of `saveCompactModel`.

The extension of the `filename` file must be `.mat`. If `filename` has no extension, then `loadCompactModel` appends `.mat`.

If `filename` does not include a full path, then `loadCompactModel` loads the file from the current folder.

Example: `'Mdl'`

Data Types: `char` | `string`

Output Arguments

`Mdl` — Machine learning model
model object

Machine learning model, returned as one of these model objects:

- Classification model object
  - `ClassificationKNN`
  - `ClassificationLinear`
  - `CompactClassificationDiscriminant`
  - `CompactClassificationECOC`
  - `CompactClassificationEnsemble`
  - `CompactClassificationNaiveBayes`
  - `CompactClassificationSVM` — If you use `saveCompactModel` to save an SVM model that is equipped to predict posterior probabilities, and use `loadCompactModel` to load the model, then `loadCompactModel` cannot restore the ScoreTransform property into the MATLAB Workspace. However, `loadCompactModel` can load the model, including the ScoreTransform property, at compile time for code generation, within an entry-point function.
  - `CompactClassificationTree`
- Regression model object
  - `CompactGeneralizedLinearModel`
  - `CompactLinearModel` — Suppose you train a linear model by using `fitlm` and specifying `'RobustOpts'` as a structure with an anonymous function handle for the `RobustWgtFun` field, use `saveCompactModel` to save the model, and then use `loadCompactModel` to load the model. In this case, `loadCompactModel` cannot restore the Robust property into the MATLAB Workspace. However, `loadCompactModel` can load the model at compile time within an entry-point function for code generation.
  - `CompactRegressionEnsemble`
  - `CompactRegressionGP`
  - `CompactRegressionSVM`
  - `CompactRegressionTree`
  - `RegressionLinear`
- Nearest neighbor searcher object
  - `ExhaustiveSearcher`
  - `KDTreeSearcher`

Algorithms

`saveCompactModel` prepares a machine learning model (`Mdl`) for code generation. The function removes some properties that are not required for prediction.

- For a model that has a corresponding compact model, the `saveCompactModel` function applies the appropriate `compact` function to the model before saving it.

- For a model that does not have a corresponding compact model, such as `ClassificationKNN`, `ClassificationLinear`, `RegressionLinear`, `ExhaustiveSearcher`, and `KDTreeSearcher`, the `saveCompactModel` function removes properties such as hyperparameter optimization properties, training solver information, and others.

`loadCompactModel` loads the model saved by `saveCompactModel`.

Alternative Functionality

- Use a coder configurer created by `learnerCoderConfigurer` for the models listed in this table.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

  After training a machine learning model, create a coder configurer of the model. Use the object functions and properties of the configurer to configure code generation options and to generate code for the `predict` and `update` functions of the model. If you generate code using a coder configurer, you can update model parameters in the generated code without having to regenerate the code. For details, see “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80.

Compatibility Considerations

`loadCompactModel` will be removed

Warns starting in R2020b

`loadCompactModel` will be removed in a future release. Use `loadLearnerForCoder` instead.

`saveLearnerForCoder` and `loadLearnerForCoder` provide broader functionality, including fixed-point code generation for supported models.

This table shows how to update your code to use `loadLearnerForCoder`.

Not Recommended	Recommended
`Mdl = loadCompactModel('MyModel');`	`Mdl = loadLearnerForCoder('MyModel');`

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`codegen` | `loadLearnerForCoder` | `saveCompactModel`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
“Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
“Code Generation for Nearest Neighbor Searcher” on page 32-19
“Specify Variable-Size Arguments for Code Generation” on page 32-45

Introduced in R2016b

loadLearnerForCoder

Reconstruct model object from saved model for code generation

Syntax

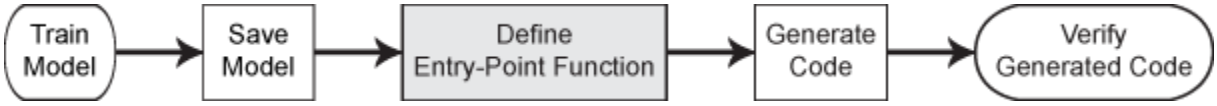
```
Mdl = loadLearnerForCoder(filename)
Mdl = loadLearnerForCoder(filename,'DataType','single')
Mdl = loadLearnerForCoder(filename,'DataType',T)
```

Description

To generate C/C++ code for the object functions of machine learning models (including `predict`, `random`, `knnsearch`, `rangesearch`, and incremental learning functions) , use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`. After training a machine learning model, save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls an object function. Then use `codegen` or the MATLAB Coder app to generate C/C++ code. Generating C/C++ code requires MATLAB Coder.

For functions that support single-precision C/C++ code generation, use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`; specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.

This flow chart shows the code generation workflow for the object functions of machine learning models. Use `loadLearnerForCoder` for the highlighted step.



Fixed-point C/C++ code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.

This flow chart shows the fixed-point code generation workflow for the `predict` function of a machine learning model. Use `loadLearnerForCoder` for the highlighted step.

`Mdl = loadLearnerForCoder(filename)` reconstructs a classification model, regression model, or nearest neighbor searcher (`Mdl`) from the model stored in the MATLAB formatted binary file (MAT-file) named `filename`. You must create the `filename` file by using `saveLearnerForCoder`.

`Mdl = loadLearnerForCoder(filename,'DataType','single')` reconstructs a single-precision classification or regression model (`Mdl`) from the model stored in the MATLAB formatted binary file (MAT-file) named `filename`.

`Mdl = loadLearnerForCoder(filename,'DataType',T)` returns a fixed-point version of the model stored in `filename`. The structure `T` contains the fields that specify the fixed-point data types

for the variables required to use the `predict` function of the model. Create `T` using the function generated by `generateLearnerDataTypeFcn`.

Use this syntax in an entry-point function, and use `codegen` to generate fixed-point code for the entry-point function. You can use this syntax only when generating code.

Examples

Generate C/C++ Code for Prediction

After training a machine learning model, save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the trained model. Then use `codegen` (MATLAB Coder) to generate C/C++ code.

This example briefly explains the code generation workflow for the prediction of machine learning models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details. To learn about the code generation for finding nearest neighbors using a nearest neighbor searcher model, see “Code Generation for Nearest Neighbor Searcher” on page 32-19.

Train Model

Load Fisher's iris data set. Remove all observed setosa irises data so that `X` and `Y` contain data for two classes only.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,:);
Y = species(inds);
```

Train a support vector machine (SVM) classification model using the processed data set.

```
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `SVMIris.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'SVMIris');
```

Define Entry-Point Function

Define an entry-point function named `classifyIris` that does the following:

- Accept iris flower measurements with columns corresponding to `meas`, and return predicted labels.
- Load a trained SVM classification model.
- Predict labels using the loaded classification model for the iris flower measurements.

```
type classifyIris.m % Display contents of classifyIris.m file
```

```
function label = classifyIris(X) %#codegen
%CLASSIFYIRIS Classify iris species using SVM Model
%   CLASSIFYIRIS classifies the iris flower measurements in X using the SVM
%   model in the file SVMIris.mat, and then returns class labels in label.
Mdl = loadLearnerForCoder('SVMIris');
label = predict(Mdl,X);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Pass `X` as the value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

```
codegen classifyIris -args {X}
```

```
Code generation successful.
```

`codegen` generates the MEX function `classifyIris_mex` with a platform-dependent extension.

Verify Generated Code

Compare the labels classified using `predict`, `classifyIris`, and `classifyIris_mex`.

```
label1 = predict(Mdl,X);
label2 = classifyIris(X);
label3 = classifyIris_mex(X);
verify_label = isequal(label1,label2,label3)
```

```
verify_label = logical
   1
```

`isequal` returns logical 1 (true), which means all the inputs are equal. The labels classified all three ways are the same.

Generate Single-Precision C/C++ Code for Prediction

After training a machine learning model, save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the trained model. Then use `codegen` (MATLAB Coder) to generate C/C++ code.

This example briefly explains the single-precision code generation workflow for the prediction of machine learning models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details.

Train Model

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using predictors `X` and class labels `Y`.

```
Mdl = fitcnb(X,Y);
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Save Model

Save the naive Bayes classification model to the file `naiveBayesIris.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'naiveBayesIris');
```

Define Entry-Point Function

Define an entry-point function named `classifyIrisSingle` that does the following:

- Accept iris flower measurements with columns corresponding to petal measurements, and return predicted labels.
- Load a trained naive Bayes classification model.
- Predict labels using the single-precision loaded classification model for the iris flower measurements.

```
type classifyIrisSingle.m
```

```
function label = classifyIrisSingle(X) %#codegen
% CLASSIFYIRISSINGLE Classify iris species using single-precision naive
% Bayes model
% CLASSIFYIRISSINGLE classifies the iris flower measurements in X using the
% single-precision naive Bayes model in the file naiveBayesIris.mat, and
% then returns the predicted labels in label.
Mdl = loadLearnerForCoder('naiveBayesIris','DataType','single');
label = predict(Mdl,X);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open this example in MATLAB, then MATLAB opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Pass `X` as the value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

```
Xpred = single(X);
codegen classifyIrisSingle -args Xpred
```

```
Code generation successful.
```

`codegen` generates the MEX function `classifyIrisSingle_mex` with a platform-dependent extension.

Verify Generated Code

Compare the labels classified using `predict`, `classifyIrisSingle`, and `classifyIrisSingle_mex`.

```
label1 = predict(Mdl,X);
label2 = classifyIrisSingle(X);
label3 = classifyIrisSingle_mex(Xpred);
verify_label = isequal(label1,label2,label3)
```

```
verify_label = logical
   1
```

`isequal` returns logical 1 (true), which means all the inputs are equal. The labels classified all three ways are the same. If the generated MEX function `classifyIrisSingle_mex` and the function `predict` do not produce the same classification results, you can compute the percentage of incorrectly classified labels.

```
sum(strcmp(label3,label1)==0)/numel(label1)*100
```

```
ans = 0
```

Generate Fixed-Point C/C++ Code for Prediction

After training a machine learning model, save the model using `saveLearnerForCoder`. For fixed-point code generation, specify the fixed-point data types of the variables required for prediction by using the data type function generated by `generateLearnerDataTypeFcn`. Then, define an entry-point function that loads the model by using both `loadLearnerForCoder` and the specified fixed-point data types, and calls the `predict` function of the model. Use `codegen` (MATLAB Coder) to generate fixed-point C/C++ code for the entry-point function, and then verify the generated code.

Before generating code using `codegen`, you can use `buildInstrumentedMex` (Fixed-Point Designer) and `showInstrumentationResults` (Fixed-Point Designer) to optimize the fixed-point

data types to improve the performance of the fixed-point code. Record minimum and maximum values of named and internal variables for prediction by using `buildInstrumentedMex`. View the instrumentation results using `showInstrumentationResults`; then, based on the results, tune the fixed-point data type properties of the variables. For details regarding this optional step, see “Fixed-Point Code Generation for Prediction of SVM” on page 32-87.

Train Model

Load the `ionosphere` data set and train a binary SVM classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y,'KernelFunction','gaussian');
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `myMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'myMdl');
```

Define Fixed-Point Data Types

Use `generateLearnerDataTypeFcn` to generate a function that defines the fixed-point data types of the variables required for prediction of the SVM model.

```
generateLearnerDataTypeFcn('myMdl',X)
```

`generateLearnerDataTypeFcn` generates the `myMdl_datatype` function.

Create a structure `T` that defines the fixed-point data types by using `myMdl_datatype`.

```
T = myMdl_datatype('Fixed')
```

```
T = struct with fields:
               XDataType: [0x0 embedded.fi]
           ScoreDataType: [0x0 embedded.fi]
    InnerProductDataType: [0x0 embedded.fi]
```

The structure `T` includes the fields for the named and internal variables required to run the `predict` function. Each field contains a fixed-point object, returned by `fi` (Fixed-Point Designer). The fixed-point object specifies fixed-point data type properties, such as word length and fraction length. For example, display the fixed-point data type properties of the predictor data.

```
T.XDataType
```

```
ans =

[]

          DataTypeMode: Fixed-point: binary point scaling
            Signedness: Signed
            WordLength: 16
        FractionLength: 14

        RoundingMethod: Floor
        OverflowAction: Wrap
```

```
          ProductMode: FullPrecision
  MaxProductWordLength: 128
              SumMode: FullPrecision
      MaxSumWordLength: 128
```

Define Entry-Point Function

Define an entry-point function named `myFixedPointPredict` that does the following:

- Accept the predictor data `X` and the fixed-point data type structure `T`.
- Load a fixed-point version of a trained SVM classification model by using both `loadLearnerForCoder` and the structure `T`.
- Predict labels and scores using the loaded model.

```
type myFixedPointPredict.m % Display contents of myFixedPointPredict.m file
```

```
function [label,score] = myFixedPointPredict(X,T) %#codegen
Mdl = loadLearnerForCoder('myMdl','DataType',T);
[label,score] = predict(Mdl,X);
end
```

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function file.

Generate Code

The `XDataType` field of the structure `T` specifies the fixed-point data type of the predictor data. Convert `X` to the type specified in `T.XDataType` by using the `cast` (Fixed-Point Designer) function.

```
X_fx = cast(X,'like',T.XDataType);
```

Generate code for the entry-point function using `codegen`. Specify `X_fx` and constant folded `T` as input arguments of the entry-point function.

```
codegen myFixedPointPredict -args {X_fx,coder.Constant(T)}
```

```
Code generation successful.
```

`codegen` generates the MEX function `myFixedPointPredict_mex` with a platform-dependent extension.

Verify Generated Code

Pass predictor data to `predict` and `myFixedPointPredict_mex` to compare the outputs.

```
[labels,scores] = predict(Mdl,X);
[labels_fx,scores_fx] = myFixedPointPredict_mex(X_fx,T);
```

Compare the outputs from `predict` and `myFixedPointPredict_mex`.

```
verify_labels = isequal(labels,labels_fx)
```

```
verify_labels = logical
   1
```

`isequal` returns logical 1 (true), which means `labels` and `labels_fx` are equal. If the labels are not equal, you can compute the percentage of incorrectly classified labels as follows.

```
sum(strcmp(labels_fx,labels)==0)/numel(labels_fx)*100
```

```
ans = 0
```

Find the maximum of the relative differences between the score outputs.

```
relDiff_scores = max(abs((scores_fx.double(:,1)-scores(:,1))./scores(:,1)))
```

```
relDiff_scores = 0.0055
```

If you are not satisfied with the comparison results and want to improve the precision of the generated code, you can tune the fixed-point data types and regenerate the code. For details, see “Tips” on page 33-2632 in `generateLearnerDataTypeFcn`, “Data Type Function” on page 33-2631, and “Fixed-Point Code Generation for Prediction of SVM” on page 32-87.

Input Arguments

`filename` — Name of MAT-file that contains structure array representing a model object

character vector | string scalar

Name of the MAT-file that contains the structure array representing a model object, specified as a character vector or string scalar. You must create the `filename` file using `saveLearnerForCoder`. `loadLearnerForCoder` reconstructs the model stored in the `filename` file at compile time.

The extension of the `filename` file must be `.mat`. If `filename` has no extension, then `loadLearnerForCoder` appends `.mat`.

If `filename` does not include a full path, then `loadLearnerForCoder` loads the file from the current folder.

These tables show the models you can save using `saveLearnerForCoder` and whether each model supports fixed-point and single-precision code generation.

- **Classification Model Object**

Model	Full/Compact Model Objects	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Binary decision tree for classification	`ClassificationTree`, `CompactClassificationTree`	Yes	Yes
Discriminant analysis classification	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`	No	Yes
Ensemble classifier	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`	Yes (Only for ensembles of decision trees)	Yes
Binary classification linear model for incremental learning	`incrementalClassificationLinear`	No	Yes
k-nearest neighbor classification	`ClassificationKNN`	No	Yes
Linear model for binary classification of high-dimensional data	`ClassificationLinear`	No	Yes
Multiclass model for support vector machines (SVMs) or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`	No	Yes
Naive Bayes classifier	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`	No	Yes
SVM for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	Yes	Yes

- **Regression Model Object**

Model	Full/Compact Model Object	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Ensemble regression	`RegressionEnsemble`, `CompactRegressionEnsemble`, `RegressionBaggedEnsemble`	Yes	Yes
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`	No	Yes (see "Tips" on page 33-3574)
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`	No	No
Linear regression model for incremental learning	`incrementalRegressionLinear`	No	Yes
Linear regression model	`LinearModel`, `CompactLinearModel`	No	Yes
Linear regression for high-dimensional data	`RegressionLinear`	No	No
Regression tree	`RegressionTree`, `CompactRegressionTree`	Yes	Yes
SVM regression	`RegressionSVM`, `CompactRegressionSVM`	Yes	Yes

- **Nearest Neighbor Searcher Object**

Model	Model Object	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Exhaustive nearest neighbor searcher	`ExhaustiveSearcher`	No	No
Nearest neighbor searcher using *K*d-tree	`KDTreeSearcher`	No	No

Example: `'Mdl'`

Data Types: `char` | `string`

T — Fixed-point data types
structure

Fixed-point data types, specified as a structure. This argument is for fixed-point C/C++ code generation.

Create `T` using a function generated by `generateLearnerDataTypeFcn`. For details about the generated function and the structure `T`, see `generateLearnerDataTypeFcn` and "Data Type Function" on page 33-2631.

You can use this argument when the model in the `filename` file is an SVM model, a decision tree model, and an ensemble of decision trees.

Data Types: `struct`

Output Arguments

`Mdl` — Machine learning model

model object

Machine learning model, returned as one of these model objects:

- Classification model object
  - `ClassificationKNN`
  - `ClassificationLinear`
  - `CompactClassificationDiscriminant`
  - `CompactClassificationECOC`
  - `CompactClassificationEnsemble`
  - `CompactClassificationNaiveBayes`
  - `CompactClassificationSVM` — If you use `saveLearnerForCoder` to save an SVM model that is equipped to predict posterior probabilities, and use `loadLearnerForCoder` to load the model, then `loadLearnerForCoder` cannot restore the ScoreTransform property into the MATLAB Workspace. However, `loadLearnerForCoder` can load the model, including the ScoreTransform property, at compile time for code generation, within an entry-point function.
  - `CompactClassificationTree`
  - `incrementalClassificationLinear`
- Regression model object
  - `CompactGeneralizedLinearModel`
  - `CompactLinearModel` — Suppose you train a linear model by using `fitlm` and specifying `'RobustOpts'` as a structure with an anonymous function handle for the `RobustWgtFun` field, use `saveLearnerForCoder` to save the model, and then use `loadLearnerForCoder` to load the model. In this case, `loadLearnerForCoder` cannot restore the Robust property into the MATLAB Workspace. However, `loadLearnerForCoder` can load the model at compile time within an entry-point function for code generation.
  - `CompactRegressionEnsemble`
  - `CompactRegressionGP`
  - `CompactRegressionSVM`
  - `CompactRegressionTree`
  - `incrementalRegressionLinear`

  - `RegressionLinear`
- Nearest neighbor searcher object
  - `ExhaustiveSearcher`
  - `KDTreeSearcher`

Tips

- For single-precision code generation for a Gaussian process regression (GPR) model, create the model by using `fitrgp(X,Y,'Standardize',1)`.

Algorithms

`saveLearnerForCoder` prepares a machine learning model (`Mdl`) for code generation. The function removes some unnecessary properties.

- For a model that has a corresponding compact model, the `saveLearnerForCoder` function applies the appropriate `compact` function to the model before saving it.
- For a model that does not have a corresponding compact model, such as `ClassificationKNN`, `ClassificationLinear`, `RegressionLinear`, `ExhaustiveSearcher`, and `KDTreeSearcher`, the `saveLearnerForCoder` function removes properties such as hyperparameter optimization properties, training solver information, and others.

`loadLearnerForCoder` loads the model saved by `saveLearnerForCoder`.

Alternative Functionality

- Use a coder configurer created by `learnerCoderConfigurer` for the models listed in this table.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

  After training a machine learning model, create a coder configurer of the model. Use the object functions and properties of the configurer to configure code generation options and to generate code for the `predict` and `update` functions of the model. If you generate code using a coder configurer, you can update model parameters in the generated code without having to regenerate the code. For details, see “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- For fixed-point C/C++ code generation, the input argument T must be a compile-time constant. For an example, see “Generate Fixed-Point C/C++ Code for Prediction” on page 33-3567.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`codegen` | `generateLearnerDataTypeFcn` | `saveLearnerForCoder`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
“Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
“Code Generation for Nearest Neighbor Searcher” on page 32-19
“Fixed-Point Code Generation for Prediction of SVM” on page 32-87
“Specify Variable-Size Arguments for Code Generation” on page 32-45

Introduced in R2019b

logncdf

Lognormal cumulative distribution function

Syntax

```
p = logncdf(x)
p = logncdf(x,mu)
p = logncdf(x,mu,sigma)

[p,pLo,pUp] = logncdf(x,mu,sigma,pCov)
[p,pLo,pUp] = logncdf(x,mu,sigma,pCov,alpha)

___ = logncdf( ___ ,'upper')
```

Description

`p = logncdf(x)` returns the cumulative distribution function (cdf) of the standard lognormal distribution, evaluated at the values in `x`. In the standard lognormal distribution, the mean and standard deviation of logarithmic values are 0 and 1, respectively.

`p = logncdf(x,mu)` returns the cdf of the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and 1 (standard deviation of logarithmic values), evaluated at the values in `x`.

`p = logncdf(x,mu,sigma)` returns the cdf of the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and `sigma` (standard deviation of logarithmic values), evaluated at the values in `x`.

`[p,pLo,pUp] = logncdf(x,mu,sigma,pCov)` also returns the 95% confidence bounds `[pLo,pUp]` of `p` using the estimated parameters (`mu` and `sigma`) and their covariance matrix `pCov`.

`[p,pLo,pUp] = logncdf(x,mu,sigma,pCov,alpha)` specifies the confidence level for the confidence interval `[pLo,pUp]` to be `100(1–alpha)`%.

`___ = logncdf( ___ ,'upper')` returns the complement of the cdf, evaluated at the values in `x`, using an algorithm that more accurately computes the extreme upper-tail probabilities. `'upper'` can follow any of the input argument combinations in the previous syntaxes.

Examples

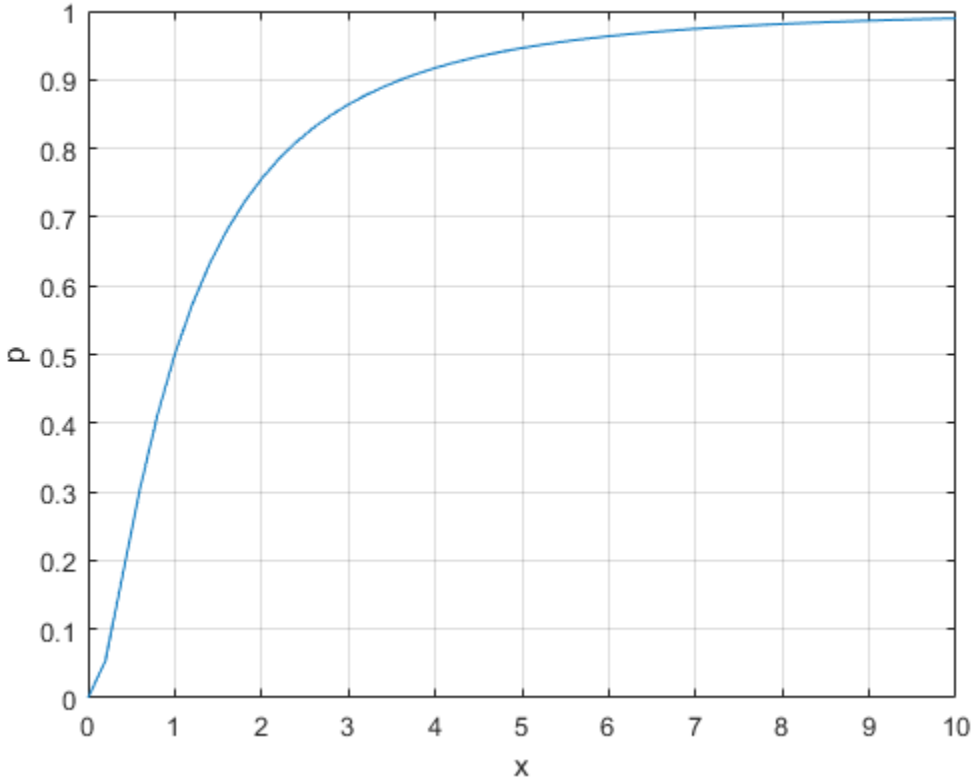
Compute Lognormal Distribution cdf

Compute the cdf values evaluated at the values in `x` for the lognormal distribution with mean `mu` and standard deviation `sigma`.

```
x = 0:0.2:10;
mu = 0;
sigma = 1;
p = logncdf(x,mu,sigma);
```

Plot the cdf.

```
plot(x,p)
grid on
xlabel('x')
ylabel('p')
```



Confidence Interval of Lognormal cdf Value

Find the maximum likelihood estimates (MLEs) of the lognormal distribution parameters, and then find the confidence interval of the corresponding cdf value.

Generate 1000 random numbers from the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,n,1);
```

Find the MLEs for the distribution parameters (mean and standard deviation of logarithmic values) by using `mle`.

```
phat = mle(x,'distribution','LogNormal')
```

```
phat = 1×2
```

```
    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `lognlike`. The function `lognlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = lognlike(phat,x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the cdf value at 0.5 and its 95% confidence interval.

```
[p,pLo,pUp] = logncdf(0.5,muHat,sigmaHat,pCov)
```

```
p = 0.0024
```

```
pLo = 0.0016
```

```
pUp = 0.0037
```

`p` is the cdf value of the lognormal distribution with the parameters `muHat` and `sigmaHat`. The interval `[pLo,pUp]` is the 95% confidence interval of the cdf evaluated at 0.5, considering the uncertainty of `muHat` and `sigmaHat` using `pCov`. The 95% confidence interval means the probability that `[pLo,pUp]` contains the true cdf value is 0.95.

Complementary cdf (Tail Distribution)

Determine the probability that an observation from a standard lognormal distribution will fall on the interval `[exp(10),Inf]`.

```
p1 = 1 - logncdf(exp(10))
```

```
p1 = 0
```

`logncdf(exp(10))` is nearly 1, so `p1` becomes 0. Specify `'upper'` so that `logncdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = logncdf(exp(10),'upper')
```

```
p2 = 7.6199e-24
```

You can also use `'upper'` to compute a right-tailed *p*-value.

Input Arguments

x — Values at which to evaluate cdf

positive scalar value | array of positive scalar values

Values at which to evaluate the cdf, specified as a positive scalar value or an array of positive scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `x` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logncdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[-1,0,3,4]`

Data Types: `single` | `double`

`mu` — Mean of logarithmic values
`0` (default) | scalar value | array of scalar values

Mean of logarithmic values for the lognormal distribution, specified as a scalar value or an array of scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `mu` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logncdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values
`1` (default) | positive scalar value | array of positive scalar values

Standard deviation of logarithmic values for the lognormal distribution, specified as a positive scalar value or an array of positive scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `sigma` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logncdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

`pCov` — Covariance of estimates
2-by-2 numeric matrix

Covariance of the estimates `mu` and `sigma`, specified as a 2-by-2 matrix.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `x`, `mu`, and `sigma` must be scalar values.

You can estimate the maximum likelihood estimates of `mu` and `sigma` by using `mle`, and estimate the covariance of `mu` and `sigma` by using `lognlike`. For an example, see “Confidence Interval of Lognormal cdf Value” on page 33-3577.

Data Types: `single` | `double`

`alpha` — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

`p` — cdf values
scalar value | array of scalar values

cdf values, evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `p` is the same size as `x`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

`pLo` — Lower confidence bound for `p`
scalar value | array of scalar values

Lower confidence bound for `p`, returned as a scalar value or an array of scalar values. `pLo` has the same size as `p`.

`pUp` — Upper confidence bound for `p`
scalar value | array of scalar values

Upper confidence bound for `p`, returned as a scalar value or an array of scalar values. `pUp` has the same size as `p`.

More About

Lognormal Distribution

The lognormal distribution is a probability distribution whose logarithm has a normal distribution.

The cumulative distribution function (cdf) of the lognormal distribution is

$$p = F(x \mid \mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^x \frac{1}{t}\exp\left\{\frac{-(\log t-\mu)^2}{2\sigma^2}\right\}dt, \quad \text{for } x > 0.$$

Algorithms

- The `logncdf` function uses the complementary error function `erfc`. The relationship between `logncdf` and `erfc` is

  $$\text{logncdf}(x, 0, 1) = \frac{1}{2}\text{erfc}\left(-\frac{\log x}{\sqrt{2}}\right).$$

  The complementary error function `erfc(x)` is defined as

  $$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt.$$

- The `logncdf` function computes confidence bounds for `p` by using the delta method. The normal distribution cdf value of `log(x)` with the parameters `mu` and `sigma` is equivalent to the cdf value of `(log(x)–mu)/sigma` with the parameters 0 and 1. Therefore, the `logncdf` function estimates the variance of `(log(x)–mu)/sigma` using the covariance matrix of `mu` and `sigma` by the delta method, and finds the confidence bounds of `(log(x)–mu)/sigma` using the estimates of this variance. Then, the function transforms the bounds to the scale of `p`. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pCov` from large samples.

Alternative Functionality

- `logncdf` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, create a `LognormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `logncdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

References

[1] Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1964.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `cdf` | `erfc` | `lognfit` | `logninv` | `lognlike` | `lognpdf` | `lognrnd` | `lognstat`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

lognfit

Lognormal parameter estimates

Syntax

```
pHat = lognfit(x)

[pHat,pCI] = lognfit(x)
[pHat,pCI] = lognfit(x,alpha)

[ ___ ] = lognfit(x,alpha,censoring)
[ ___ ] = lognfit(x,alpha,censoring,freq)
[ ___ ] = lognfit(x,alpha,censoring,freq,options)
```

Description

`pHat = lognfit(x)` returns unbiased estimates of lognormal distribution parameters, given the sample data in `x`. `pHat(1)` and `pHat(2)` are the mean and standard deviation of logarithmic values, respectively.

`[pHat,pCI] = lognfit(x)` also returns 95% confidence intervals for the parameter estimates.

`[pHat,pCI] = lognfit(x,alpha)` specifies the confidence level for the confidence intervals to be `100(1–alpha)`%.

`[ ___ ] = lognfit(x,alpha,censoring)` specifies whether each value in `x` is right-censored or not. Use the logical vector `censoring` in which 1 indicates observations that are right-censored and 0 indicates observations that are fully observed. With censoring, the `phat` values are the maximum likelihood estimates (MLEs).

`[ ___ ] = lognfit(x,alpha,censoring,freq)` specifies the frequency or weights of observations.

`[ ___ ] = lognfit(x,alpha,censoring,freq,options)` specifies optimization options for the iterative algorithm `lognfit` to use to compute MLEs with censoring. Create `options` by using the function `statset`.

You can pass in `[]` for `alpha`, `censoring`, and `freq` to use their default values.

Examples

Estimate Parameters and Confidence Intervals

Generate 1000 random numbers from the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,n,1);
```

Find the parameter estimates and the 99% confidence intervals.

```
[pHat,pCI] = lognfit(x,0.01)
```

```
pHat = 1×2

    4.9347    1.9979


pCI = 2×2

    4.7717    1.8887
    5.0978    2.1196
```

`pHat(1)` and `pHat(2)` are the mean and standard deviation of logarithmic values, respectively. `pCI` contains the 99% confidence intervals of the mean and standard deviation parameters. The values in the first row are the lower bounds, and the values in the second row are the upper bounds.

Change Algorithm Options

Find the MLEs of a data set with censoring by using `lognfit`. Use `statset` to specify the iterative algorithm options that `lognfit` uses to compute MLEs for censored data, and then find the MLEs again.

Generate the true times `x` that follow the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,n,1);
```

Generate the censoring times. Note that the censoring times must be independent of the true times `x`.

```
censtime = normrnd(150,20,size(x));
```

Specify the indicator for the censoring times and the observed times.

```
censoring = x>censtime;
y = min(x,censtime);
```

Find the MLEs of the lognormal distribution parameters. The second input argument of `lognfit` specifies the confidence level. Pass in `[]` to use its default value 0.05. The third input argument specifies the censorship information.

```
pHat = lognfit(y,[],censoring)
```

```
pHat = 1×2

    4.9535    1.9996
```

Display the default algorithm parameters that `lognfit` uses to estimate the lognormal distribution parameters.

```
statset('lognfit')
```

```
ans = struct with fields:
          Display: 'off'
```

```
      MaxFunEvals: 200
          MaxIter: 100
           TolBnd: 1.0000e-06
           TolFun: 1.0000e-08
        TolTypeFun: []
              TolX: 1.0000e-08
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

Save the options using a different name. Change how the results are displayed (`Display`) and the termination tolerance for the objective function (`TolFun`).

```
options = statset('lognfit');
options.Display = 'final';
options.TolFun = 1e-10;
```

Alternatively, you can specify algorithm parameters by using the name-value pair arguments of the function `statset`.

```
options = statset('Display','final','TolFun',1e-10);
```

Find the MLEs with the new algorithm parameters.

```
pHat = lognfit(y,[],censoring,[],options)
```

```
Successful convergence: Norm of gradient less than OPTIONS.TolFun

pHat = 1×2

    4.9535    1.9996
```

`lognfit` displays a report on the final iteration.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

alpha — Significance level

0.05 (default) | scalar in the range (0,1)

Significance level for the confidence intervals, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence intervals do not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

`censoring` — Indicator for censoring

array of 0s (default) | logical vector

Indicator for the censoring of each value in `x`, specified as a logical vector of the same size as `x`. Use 1 for observations that are right-censored and 0 for observations that are fully observed.

The default is an array of 0s, meaning that all observations are fully observed.

Data Types: `logical`

`freq` — Frequency or weights of observations

array of 1s (default) | nonnegative vector

Frequency or weights of observations, specified as a nonnegative vector that is the same size as `x`. The `freq` input argument typically contains nonnegative integer counts for the corresponding elements in `x`, but can contain any nonnegative values.

To obtain the weighted MLEs for a data set with censoring, specify weights of observations, normalized to the number of observations in `x`.

The default is an array of 1s, meaning one observation per element of `x`.

Data Types: `single` | `double`

`options` — Optimization options

`statset('lognfit')` (default) | structure

Optimization options, specified as a structure. `options` determines the control parameters for the iterative algorithm that `lognfit` uses to compute MLEs for censored data.

Create `options` by using the function `statset` or by creating a structure array containing the fields and values described in this table.

Field Name	Value	Default Value
`Display`	Amount of information displayed by the algorithm. • `'off'` — Displays no information. • `'final'` — Displays the final output.	`'off'`
`MaxFunEvals`	Maximum number of objective function evaluations allowed, specified as a positive integer.	`200`
`MaxIter`	Maximum number of iterations allowed, specified as a positive integer.	`100`

Field Name	Value	Default Value
`TolBnd`	Lower bound of the standard deviation parameter estimate, specified as a positive scalar. The bounds for the mean and standard deviation parameter estimates are `[–Inf,Inf]` and `[TolBnd,Inf]`, respectively.	`1e-6`
`TolFun`	Termination tolerance for the objective function value, specified as a positive scalar.	`1e-8`
`TolX`	Termination tolerance for the parameters, specified as a positive scalar.	`1e-8`

You can also enter `statset('lognfit')` in the Command Window to see the names and default values of the fields that `lognfit` accepts in the `options` structure.

Example: `statset('Display','final','MaxIter',1000)` specifies to display the final information of the iterative algorithm results, and change the maximum number of iterations allowed to 1000.

Data Types: `struct`

Output Arguments

`pHat` — Estimates of lognormal distribution parameters

1-by-2 vector

Estimates of lognormal distribution parameters, returned as a 1-by-2 vector. `pHat(1)` and `pHat(2)` are the mean and standard deviation of logarithmic values, respectively.

- With no censoring, the `pHat` values are unbiased estimates. To compute the MLEs with no censoring, use the `mle` function.
- With censoring, the `pHat` values are the MLEs. To compute the weighted MLEs, specify the weights of observations by using `freq`.

`pCI` — Confidence intervals for parameter estimates

2-by-2 matrix

Confidence intervals for parameter estimates of the lognormal distribution, returned as a 2-by-2 matrix containing the lower and upper bounds of the `100(1–alpha)`% confidence intervals.

The first and second rows correspond to the lower and upper bounds of the confidence intervals, respectively.

Algorithms

To compute the confidence intervals, `lognfit` uses the exact method for uncensored data and the Wald method for censored data. The exact method provides exact coverage for uncensored samples based on *t* and chi-square distributions.

Alternative Functionality

`lognfit` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic functions `mle`, `fitdist`, and `paramci` and the **Distribution Fitter** app, which support various probability distributions.

- `mle` returns MLEs and the confidence intervals of MLEs for the parameters of various probability distributions. You can specify the probability distribution name or a custom probability density function.
- Create a `LognormalDistribution` probability distribution object by fitting the distribution to data using the `fitdist` function or the **Distribution Fitter** app. The object properties `mu` and `sigma` store the parameter estimates. To obtain the confidence intervals for the parameter estimates, pass the object to `paramci`.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[3] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `fitdist` | `logncdf` | `logninv` | `lognlike` | `lognpdf` | `lognrnd` | `lognstat` | `mle` | `paramci` | `statset`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

logninv

Lognormal inverse cumulative distribution function

Syntax

```
x = logninv(p)
x = logninv(p,mu)
x = logninv(p,mu,sigma)

[x,xLo,xUp] = logninv(p,mu,sigma,pCov)
[x,xLo,xUp] = logninv(p,mu,sigma,pCov,alpha)
```

Description

`x = logninv(p)` returns the inverse of the standard lognormal cumulative distribution function (cdf), evaluated at the probability values in `p`. In the standard lognormal distribution, the mean and standard deviation of logarithmic values are 0 and 1, respectively.

`x = logninv(p,mu)` returns the inverse of the lognormal cdf with the distribution parameters `mu` (mean of logarithmic values) and 1 (standard deviation of logarithmic values), evaluated at the probability values in `p`.

`x = logninv(p,mu,sigma)` returns the inverse of the lognormal cdf with the distribution parameters `mu` (mean of logarithmic values) and `sigma` (standard deviation of logarithmic values), evaluated at the probability values in `p`.

`[x,xLo,xUp] = logninv(p,mu,sigma,pCov)` also returns the 95% confidence bounds [`xLo`,`xUp`] of `x` using the estimated parameters (`mu` and `sigma`) and their covariance matrix `pCov`.

`[x,xLo,xUp] = logninv(p,mu,sigma,pCov,alpha)` specifies the confidence level for the confidence interval [`xLo`,`xUp`] to be `100(1–alpha)`%.

Examples

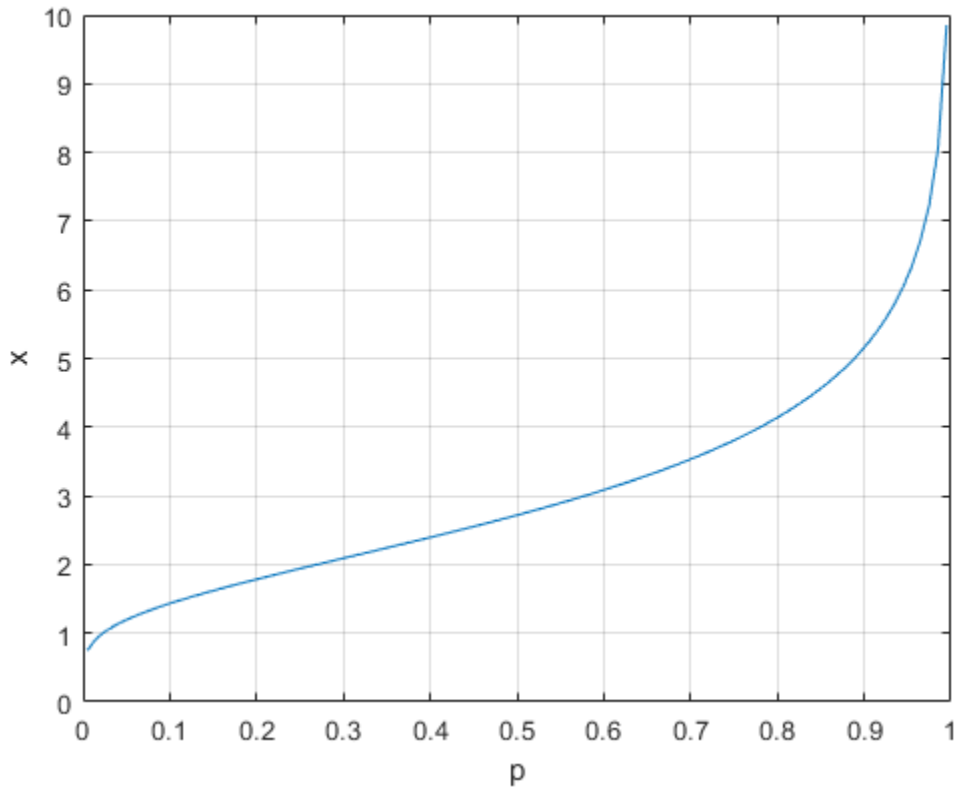
Inverse of Lognormal Distribution cdf

Compute the inverse of cdf values evaluated at the probability values in `p` for the lognormal distribution with mean `mu` and standard deviation `sigma`.

```
p = 0.005:0.01:0.995;
mu = 1;
sigma = 0.5;
x = logninv(p,mu,sigma);
```

Plot the inverse cdf.

```
plot(p,x)
grid on
xlabel('p');
ylabel('x');
```



Confidence Interval of Inverse Lognormal cdf Value

Find the maximum likelihood estimates (MLEs) of the lognormal distribution parameters, and then find the confidence interval of the corresponding inverse cdf value.

Generate 1000 random numbers from the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,[n,1]);
```

Find the MLEs for the distribution parameters (mean and standard deviation of logarithmic values) by using `mle`.

```
phat = mle(x,'distribution','LogNormal')
```

```
phat = 1×2

    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `lognlike`. The function `lognlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = lognlike(phat,x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the inverse cdf value at 0.5 and its 99% confidence interval.

```
[x,xLo,xUp] = logninv(0.5,muHat,sigmaHat,pCov,0.01)
```

```
x = 139.0364

xLo = 118.1643

xUp = 163.5953
```

`x` is the inverse cdf value using the lognormal distribution with the parameters `muHat` and `sigmaHat`. The interval `[xLo,xUp]` is the 99% confidence interval of the inverse cdf value evaluated at 0.5, considering the uncertainty of `muHat` and `sigmaHat` using `pCov`. The 99% confidence interval means the probability that `[xLo,xUp]` contains the true inverse cdf value is 0.99.

Input Arguments

`p` — Probability values at which to evaluate inverse of cdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the inverse of the cdf (icdf), specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `p` must be a scalar value.

To evaluate the icdf at multiple values, specify `p` using an array. To evaluate the icdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `p`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logninv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

Example: `[0.1,0.5,0.9]`

Data Types: `single` | `double`

`mu` — Mean of logarithmic values

`0` (default) | scalar value | array of scalar values

Mean of logarithmic values for the lognormal distribution, specified as a scalar value or an array of scalar values.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `mu` must be a scalar value.

To evaluate the icdf at multiple values, specify `p` using an array. To evaluate the icdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `p`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logninv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values

1 (default) | positive scalar value | array of positive scalar values

Standard deviation of logarithmic values for the lognormal distribution, specified as a positive scalar value or an array of positive scalar values.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `sigma` must be a scalar value.

To evaluate the icdf at multiple values, specify `p` using an array. To evaluate the icdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `p`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `logninv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

`pCov` — Covariance of estimates

2-by-2 numeric matrix

Covariance of the estimates `mu` and `sigma`, specified as a 2-by-2 matrix.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `p`, `mu`, and `sigma` must be scalar values.

You can estimate the maximum likelihood estimates of `mu` and `sigma` by using `mle`, and estimate the covariance of `mu` and `sigma` by using `lognlike`. For an example, see “Confidence Interval of Inverse Lognormal cdf Value” on page 33-3590.

Data Types: `single` | `double`

`alpha` — Significance level

0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

x — icdf values

scalar value | array of scalar values

icdf values, evaluated at the probability values in `p`, returned as a scalar value or an array of scalar values. `x` is the same size as `p`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

xLo — Lower confidence bound for x

scalar value | array of scalar values

Lower confidence bound for `x`, returned as a scalar value or an array of scalar values. `xLo` has the same size as `x`.

xUp — Upper confidence bound for x

scalar value | array of scalar values

Upper confidence bound for `x`, returned as a scalar value or an array of scalar values. `xUp` has the same size as `x`.

More About

Lognormal Distribution

The lognormal distribution is a probability distribution whose logarithm has a normal distribution.

The lognormal inverse function is defined in terms of the lognormal cdf as

$$x = F^{-1}(p\left|\mu,\sigma\right.) = \left\{x : F(x\left|\mu,\sigma\right.) = p\right\}$$

where

$$p = F(x\left|\mu,\sigma\right.) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^x \frac{1}{t}\exp\left\{\frac{-(\log t-\mu)^2}{2\sigma^2}\right\}dt, \quad \text{for } x > 0.$$

Algorithms

- The function `logninv` uses the inverse complementary error function `erfcinv`. The relationship between `logninv` and `erfcinv` is

  $$\text{logninv}(p,0,1) = \exp(-\sqrt{2}\,\text{erfcinv}(2p)).$$

  The inverse complementary error function `erfcinv(x)` is defined as `erfcinv(erfc(x))=x`, and the complementary error function `erfc(x)` is defined as

  $$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt.$$

- The `logninv` function computes confidence bounds for `x` by using the delta method. `log(logninv(p,mu,sigma))` is equivalent to `mu + sigma*log(logninv(p,0,1))`. Therefore, the `logninv` function estimates the variance of `mu + sigma*log(logninv(p,0,1))`

using the covariance matrix of `mu` and `sigma` by the delta method, and finds the confidence bounds using the estimates of this variance. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pCov` from large samples.

Alternative Functionality

- `logninv` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use `icdf`, create a `LognormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `logninv` is faster than the generic function `icdf`.

References

[1] Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1964.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. Hoboken, NJ: Wiley-Interscience, 2000. pp. 102–105.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `erfcinv` | `icdf` | `logncdf` | `lognfit` | `lognlike` | `lognpdf` | `lognrnd` | `lognstat`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

lognlike

Lognormal negative loglikelihood

Syntax

```
nlogL = lognlike(params,x)
nlogL = lognlike(params,x,censoring)
nlogL = lognlike(params,x,censoring,freq)

[nlogL,aVar] = lognlike( ___ )
```

Description

`nlogL = lognlike(params,x)` returns the lognormal negative loglikelihood of the distribution parameters (`params`) given the sample data (`x`). `params(1)` and `params(2)` are the mean and standard deviation of logarithmic values, respectively.

`nlogL = lognlike(params,x,censoring)` specifies whether each value in `x` is right-censored or not. Use the logical vector `censoring` in which 1 indicates observations that are right-censored and 0 indicates observations that are fully observed.

`nlogL = lognlike(params,x,censoring,freq)` specifies the frequency or weights of observations. To specify `freq` without specifying `censoring`, you can pass `[]` for `censoring`.

`[nlogL,aVar] = lognlike( ___ )` also returns the inverse of the Fisher information matrix `aVar`, using any of the input argument combinations in the previous syntaxes. If values in `params` are the maximum likelihood estimates (MLEs) of the parameters, `aVar` is an approximation to the asymptotic covariance matrix.

Examples

Negative Loglikelihood of MLEs

Find the MLEs of a data set with censoring by using `mle`, and then find the negative loglikelihood of the MLEs by using `lognlike`.

Generate 1000 random numbers from the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,[n,1]);
```

Find the MLEs for the distribution parameters (mean and standard deviation of logarithmic values) by using `mle`.

```
phat = mle(x,'distribution','LogNormal')
```

```
phat = 1×2
```

```
    4.9347    1.9969
```

Find the negative loglikelihood of the MLEs.

```
nlogL = lognlike(phat,x)
```

```
nlogL = 7.0453e+03
```

Confidence Interval of Lognormal cdf Value

Find the maximum likelihood estimates (MLEs) of the lognormal distribution parameters, and then find the confidence interval of the corresponding cdf value.

Generate 1000 random numbers from the lognormal distribution with the parameters 5 and 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = lognrnd(5,2,n,1);
```

Find the MLEs for the distribution parameters (mean and standard deviation of logarithmic values) by using `mle`.

```
phat = mle(x,'distribution','LogNormal')
```

```
phat = 1×2

    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `lognlike`. The function `lognlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = lognlike(phat,x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the cdf value at 0.5 and its 95% confidence interval.

```
[p,pLo,pUp] = logncdf(0.5,muHat,sigmaHat,pCov)
```

```
p = 0.0024
```

```
pLo = 0.0016
```

```
pUp = 0.0037
```

`p` is the cdf value of the lognormal distribution with the parameters `muHat` and `sigmaHat`. The interval `[pLo,pUp]` is the 95% confidence interval of the cdf evaluated at 0.5, considering the

uncertainty of `muHat` and `sigmaHat` using `pCov`. The 95% confidence interval means the probability that `[pLo,pUp]` contains the true cdf value is 0.95.

Input Arguments

`params` — Lognormal distribution parameters

vector of two numeric values

Lognormal distribution parameters, specified as a vector of two numeric values. `params(1)` and `params(2)` are the mean and standard deviation of logarithmic values, respectively. `params(2)` must be positive.

Example: `[0,1]`

Data Types: `single` | `double`

`x` — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

`censoring` — Indicator for censoring

array of 0s (default) | logical vector

Indicator for the censoring of each value in `x`, specified as a logical vector of the same size as `x`. Use 1 for observations that are right-censored and 0 for observations that are fully observed.

The default is an array of 0s, meaning that all observations are fully observed.

Data Types: `logical`

`freq` — Frequency or weights of observations

array of 1s (default) | nonnegative vector

Frequency or weights of observations, specified as a nonnegative vector that is the same size as `x`. The `freq` input argument typically contains nonnegative integer counts for the corresponding elements in `x`, but can contain any nonnegative values.

To obtain the weighted negative loglikelihood for a data set with censoring, specify weights of observations, normalized to the number of observations in `x`.

The default is an array of 1s, meaning one observation per element of `x`.

Data Types: `single` | `double`

Output Arguments

`nlogL` — Negative loglikelihood

numeric scalar

Negative loglikelihood value of the distribution parameters (`params`) given the sample data (`x`), returned as a numeric scalar.

aVar — Inverse of Fisher information matrix
numeric matrix

Inverse of the Fisher information matrix, returned as a 2-by-2 numeric matrix. `aVar` is based on the observed Fisher information given the observed data (`x`), not the expected information.

If values in `params` are the MLEs of the parameters, `aVar` is an approximation to the asymptotic variance-covariance matrix (also known as the asymptotic covariance matrix). To find the MLEs, use `mle`.

Alternative Functionality

`lognlike` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic functions `mlecov`, `fitdist`, `negloglik`, and `proflik` and the **Distribution Fitter** app, which support various probability distributions.

- `mlecov` returns the asymptotic covariance matrix of the MLEs of the parameters for a distribution specified by a custom probability density function. For example, `mlecov(params,x,'pdf',@lognpdf)` returns the asymptotic covariance matrix of the MLEs for the lognormal distribution.
- Create a `LognormalDistribution` probability distribution object by fitting the distribution to data using the `fitdist` function or the **Distribution Fitter** app. The object property `ParameterCovariance` stores the covariance matrix of the parameter estimates. To obtain the negative loglikelihood of the parameter estimates and the profile of the likelihood function, pass the object to `negloglik` and `proflik`, respectively.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[3] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `logncdf` | `lognfit` | `logninv` | `lognpdf` | `lognrnd` | `lognstat` | `mle` | `mlecov` | `negloglik` | `proflik`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

lognpdf

Lognormal probability density function

Syntax

```
y = lognpdf(x)
y = lognpdf(x,mu)
y = lognpdf(x,mu,sigma)
```

Description

`y = lognpdf(x)` returns the probability density function (pdf) of the standard lognormal distribution, evaluated at the values in `x`. In the standard lognormal distribution, the mean and standard deviation of logarithmic values are 0 and 1, respectively.

`y = lognpdf(x,mu)` returns the pdf of the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and 1 (standard deviation of logarithmic values), evaluated at the values in `x`.

`y = lognpdf(x,mu,sigma)` returns the pdf of the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and `sigma` (standard deviation of logarithmic values), evaluated at the values in `x`.

Examples

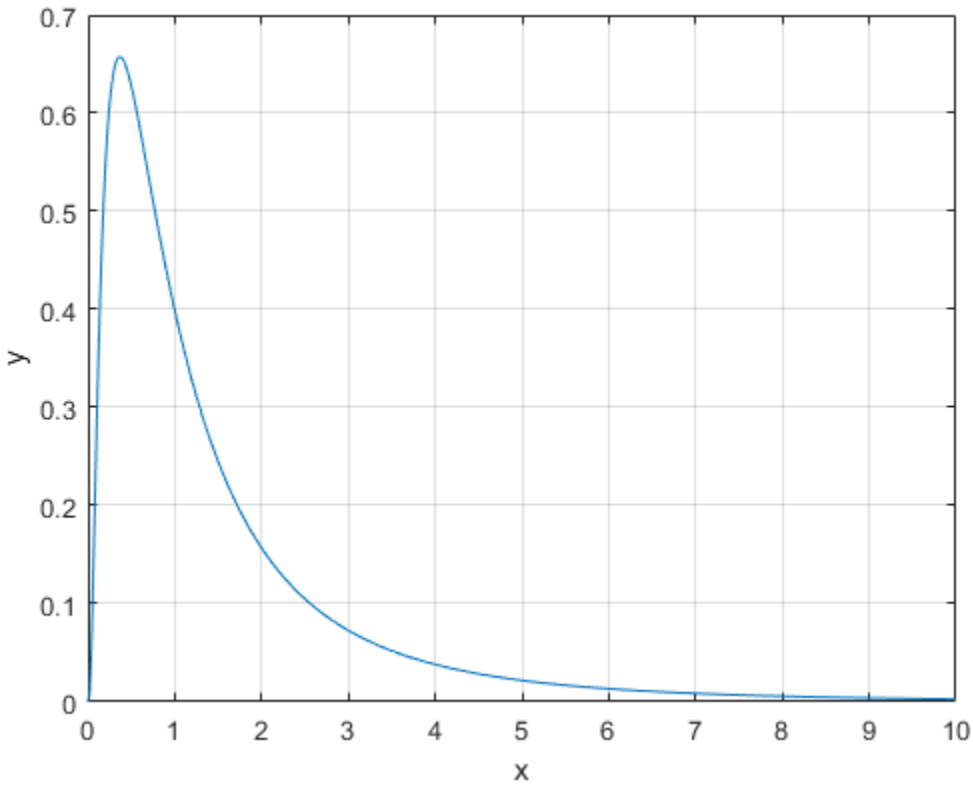
Compute Lognormal Distribution pdf

Compute the pdf values evaluated at the values in `x` for the lognormal distribution with mean `mu` and standard deviation `sigma`.

```
x = 0:0.02:10;
mu = 0;
sigma = 1;
y = lognpdf(x,mu,sigma);
```

Plot the pdf.

```
plot(x,y)
grid on
xlabel('x')
ylabel('y')
```



Input Arguments

x — Values at which to evaluate pdf

positive scalar value | array of positive scalar values

Values at which to evaluate the pdf, specified as a positive scalar value or an array of positive scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `lognpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[-1,0,3,4]`

Data Types: `single` | `double`

mu — Mean of logarithmic values

`0` (default) | scalar value | array of scalar values

Mean of logarithmic values for the lognormal distribution, specified as a scalar value or an array of scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `lognpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values
`1` (default) | positive scalar value | array of positive scalar values

Standard deviation of logarithmic values for the lognormal distribution, specified as a positive scalar value or an array of positive scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `lognpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values, evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

More About

Lognormal Distribution

The lognormal distribution is a probability distribution whose logarithm has a normal distribution.

The probability density function (pdf) of the lognormal distribution is

$$y = f(x\left|\mu,\sigma\right.) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{\frac{-(\log x-\mu)^2}{2\sigma^2}\right\},\quad \text{for } x > 0.$$

Alternative Functionality

- `lognpdf` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create a `LognormalDistribution` probability distribution object and pass the object as an input

argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `lognpdf` is faster than the generic function `pdf`.

- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

References

[1] Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. 3rd ed., New York: McGraw-Hill, 1974. pp. 540–541.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `logncdf` | `lognfit` | `logninv` | `lognlike` | `lognrnd` | `lognstat` | `pdf`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

lognrnd

Lognormal random numbers

Syntax

```
r = lognrnd(mu,sigma)
r = lognrnd(mu,sigma,sz1,...,szN)
r = lognrnd(mu,sigma,sz)
```

Description

`r = lognrnd(mu,sigma)` generates a random number from the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and `sigma` (standard deviation of logarithmic values).

`r = lognrnd(mu,sigma,sz1,...,szN)` generates an array of lognormal random numbers, where `sz1,...,szN` indicates the size of each dimension.

`r = lognrnd(mu,sigma,sz)` generates an array of lognormal random numbers, where vector `sz` specifies `size(r)`.

Examples

Generate Lognormal Random Number

Find the distribution parameters from the mean and variance of a lognormal distribution and generate a lognormal random value from the distribution.

Find the distribution parameters `mu` and `sigma` from the mean and variance.

```
m = 1; % mean
v = 2; % variance
mu = log((m^2)/sqrt(v+m^2))
```

```
mu = -0.5493
```

```
sigma = sqrt(log(v/(m^2)+1))
```

```
sigma = 1.0481
```

Generate a lognormal random value.

```
rng('default') % For reproducibility
r = lognrnd(mu,sigma)
```

```
r = 1.0144
```

Reset Random Number Generator

Save the current state of the random number generator. Then create a 1-by-5 vector of lognormal random numbers from the lognormal distribution with the parameters 3 and 10.

```
s = rng;
r = lognrnd(3,10,[1,5])
```

```
r = 1×5
10^9 ×

    0.0000    1.8507    0.0000    0.0001    0.0000
```

Restore the state of the random number generator to `s`, and then create a new 1-by-5 vector of random numbers. The values are the same as before.

```
rng(s);
r1 = lognrnd(3,10,[1,5])
```

```
r1 = 1×5
10^9 ×

    0.0000    1.8507    0.0000    0.0001    0.0000
```

Clone Size from Existing Array

Create a matrix of lognormally distributed random numbers with the same size as an existing array.

```
A = [3 2; -2 1];
sz = size(A);
R = lognrnd(0,1,sz)
```

```
R = 2×2

    1.7120    0.1045
    6.2582    2.3683
```

You can combine the previous two lines of code into a single line.

```
R = lognrnd(1,0,size(A));
```

Input Arguments

`mu` — Mean of logarithmic values

scalar value | array of scalar values

Mean of logarithmic values for the lognormal distribution, specified as a scalar value or an array of scalar values.

To generate random numbers from multiple distributions, specify `mu` and `sigma` using arrays. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar,

then `lognrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values
nonnegative scalar value | array of nonnegative scalar values

Standard deviation of logarithmic values for the lognormal distribution, specified as a nonnegative scalar value or an array of nonnegative scalar values.

If `sigma` is zero, then the output `r` is always equal to `exp(mu)`.

To generate random numbers from multiple distributions, specify `mu` and `sigma` using arrays. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `lognrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as separate arguments of integers. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the lognormal probability distribution.

If either `mu` or `sigma` is an array, then the specified dimensions `sz1,...,szN` must match the common dimensions of `mu` and `sigma` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `lognrnd` ignores trailing dimensions with a size of 1. For example, `lognrnd(mu,sigma,3,1,1,1)` produces a 3-by-1 vector of random numbers.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5 3 2]` generates a 5-by-3-by-2 array of random numbers from the lognormal probability distribution.

If either `mu` or `sigma` is an array, then the specified dimensions `sz` must match the common dimensions of `mu` and `sigma` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value `[sz1]`, then `r` is a square matrix of size `sz1`-by-`sz1`.

- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `lognrnd` ignores trailing dimensions with a size of 1. For example, `lognrnd(mu,sigma,[3,1,1,1])` produces a 3-by-1 vector of random numbers.

Example: `[5 3 2]`

Data Types: `single` | `double`

Output Arguments

`r` — Lognormal random numbers

scalar value | array of scalar values

Lognormal random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

More About

Lognormal Distribution

The lognormal distribution is a probability distribution whose logarithm has a normal distribution.

The mean m and variance v of a lognormal random variable are functions of the lognormal distribution parameters μ and σ:

$$m = \exp\left(\mu + \sigma^2/2\right)$$

$$v = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right)$$

Also, you can compute the lognormal distribution parameters μ and σ from the mean m and variance v:

$$\mu = \log\left(m^2/\sqrt{v + m^2}\right)$$

$$\sigma = \sqrt{\log\left(v/m^2 + 1\right)}$$

Alternative Functionality

- `lognrnd` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, create a `LognormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `lognrnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

References

[1] Marsaglia, G., and W. W. Tsang. “A Fast, Easily Implemented Method for Sampling from Decreasing or Symmetric Unimodal Density Functions.” *SIAM Journal on Scientific and Statistical Computing*. Vol. 5, Number 2, 1984, pp. 349–359.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `logncdf` | `lognfit` | `logninv` | `lognlike` | `lognpdf` | `lognstat` | `normrnd` | `random`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

lognstat

Lognormal mean and variance

Syntax

```
[m,v] = lognstat(mu,sigma)
```

Description

`[m,v] = lognstat(mu,sigma)` returns the mean and variance of the lognormal distribution with the distribution parameters `mu` (mean of logarithmic values) and `sigma` (standard deviation of logarithmic values).

Examples

Compute Mean and Variance

Compute the mean and variance of the lognormal distribution with parameters `mu` and `sigma`.

```
mu = 0;
sigma = 1;
[m,v] = lognstat(mu,sigma)
```

```
m = 1.6487
```

```
v = 4.6708
```

Input Arguments

`mu` — Mean of logarithmic values

scalar value | array of scalar values

Mean of logarithmic values for the lognormal distribution, specified as a scalar value or an array of scalar values.

To compute the means and variances of multiple distributions, specify distribution parameters using an array of scalar values. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `lognstat` expands the scalar argument into a constant array of the same size as the other argument. Each element in `m` and `v` is the mean and variance of the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values

positive scalar value | array of positive scalar values

Standard deviation of logarithmic values for the lognormal distribution, specified as a positive scalar value or an array of positive scalar values.

To compute the means and variances of multiple distributions, specify distribution parameters using an array of scalar values. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `lognstat` expands the scalar argument into a constant array of the same size as the other argument. Each element in `m` and `v` is the mean and variance of the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

Output Arguments

`m` — Mean

scalar value | array of scalar values

Mean of the lognormal distribution, returned as a scalar value or an array of scalar values. `m` is the same size as `mu` and `sigma` after any necessary scalar expansion. Each element in `m` is the mean of the lognormal distribution specified by the corresponding elements in `mu` and `sigma`.

`v` — Variance

scalar value | array of scalar values

Variance of the lognormal distribution, returned as a scalar value or an array of scalar values. `v` is the same size as `mu` and `sigma` after any necessary scalar expansion. Each element in `v` is the variance of the lognormal distribution specified by the corresponding elements in `mu` and `sigma`.

More About

Lognormal Distribution

The lognormal distribution is a probability distribution whose logarithm has a normal distribution.

The mean m and variance v of a lognormal random variable are functions of the lognormal distribution parameters μ and σ:

$$m = \exp\left(\mu + \sigma^2/2\right)$$

$$v = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right)$$

Also, you can compute the lognormal distribution parameters μ and σ from the mean m and variance v:

$$\mu = \log\left(m^2/\sqrt{v + m^2}\right)$$

$$\sigma = \sqrt{\log\left(v/m^2 + 1\right)}$$

Alternative Functionality

- `lognstat` is a function specific to lognormal distribution. Statistics and Machine Learning Toolbox also offers generic functions to compute summary statistics, including mean (`mean`), median (`median`), interquartile range (`iqr`), variance (`var`), and standard deviation (`std`). These generic functions support various probability distributions. To use these functions, create a `LognormalDistribution` probability distribution object and pass the object as an input argument.

References

[1] Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. 3rd ed., New York: McGraw-Hill, 1974. pp. 540–541.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LognormalDistribution` | `logncdf` | `lognfit` | `logninv` | `lognlike` | `lognpdf` | `lognrnd` | `mean` | `std` | `var`

Topics

“Lognormal Distribution” on page B-88

Introduced before R2006a

logp

Log unconditional probability density for discriminant analysis classifier

Syntax

```
lp = logp(obj,Xnew)
```

Description

`lp = logp(obj,Xnew)` returns the log of the unconditional probability density of each row of `Xnew`, computed using the discriminant analysis model `obj`.

Input Arguments

obj

Discriminant analysis classifier, produced using `fitcdiscr`.

Xnew

Matrix where each row represents an observation, and each column represents a predictor. The number of columns in `Xnew` must equal the number of predictors in `obj`.

Output Arguments

lp

Column vector with the same number of rows as `Xnew`. Each entry is the logarithm of the unconditional probability density of the corresponding row of `Xnew`.

Examples

Compute Log Unconditional Probability Density of an Observation

Construct a discriminant analysis classifier for Fisher's iris data, and examine its prediction for an average measurement.

Load Fisher's iris data and construct a default discriminant analysis classifier.

```
load fisheriris
Mdl = fitcdiscr(meas,species);
```

Find the log probability of the discriminant model applied to an average iris.

```
logpAverage = logp(Mdl,mean(meas))
```

```
logpAverage = -1.7254
```

More About

Unconditional Probability Density

The unconditional probability density of a point x of a discriminant analysis model is

$$P(x) = \sum_{k=1}^{K} P(x,k),$$

where $P(x,k)$ is the conditional density of the model at x for class k, when the total number of classes is K.

The conditional density $P(x,k)$ is

$P(x,k) = P(k)P(x|k)$,

where $P(k)$ is the prior probability of class k, and $P(x|k)$ is the conditional density of x given class k. The conditional density function of the multivariate normal with 1-by-d mean μ_k and d-by-d covariance Σ_k at a 1-by-d point x is

$$P(x|k) = \frac{1}{\left((2\pi)^d|\Sigma_k|\right)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^T\right),$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix.

See Also

`CompactClassificationDiscriminant` | `fitcdiscr` | `mahal`

Topics

“Discriminant Analysis Classification” on page 20-2

logp

Log unconditional probability density for naive Bayes classifier

Syntax

```
lp = logp(Mdl,tbl)
lp = logp(Mdl,X)
```

Description

`lp = logp(Mdl,tbl)` returns the log "Unconditional Probability Density" on page 33-3617 (`lp`) of the observations (rows) in `tbl` using the naive Bayes model `Mdl`. You can use `lp` to identify outliers in the training data.

`lp = logp(Mdl,X)` returns the log unconditional probability density of the observations (rows) in `X` using the naive Bayes model `Mdl`.

Examples

Compute Unconditional Probability Densities of Observations

Compute the unconditional probability densities of the in-sample observations of a naive Bayes classifier model.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Compute the unconditional probability densities of the in-sample observations.

```
lp = logp(Mdl,X);
```

Identify indices of observations that have very small or very large log unconditional probabilities (`ind`). Display lower (`L`) and upper (`U`) thresholds used by the outlier detection method.

```
[TF,L,U] = isoutlier(lp);
L
```

```
L = -6.9222
```

```
U
```

```
U = 3.0323
```

```
ind = find(TF)
```

```
ind = 4×1

    61
   118
   119
   132
```

Display the values of the outlier unconditional probability densities.

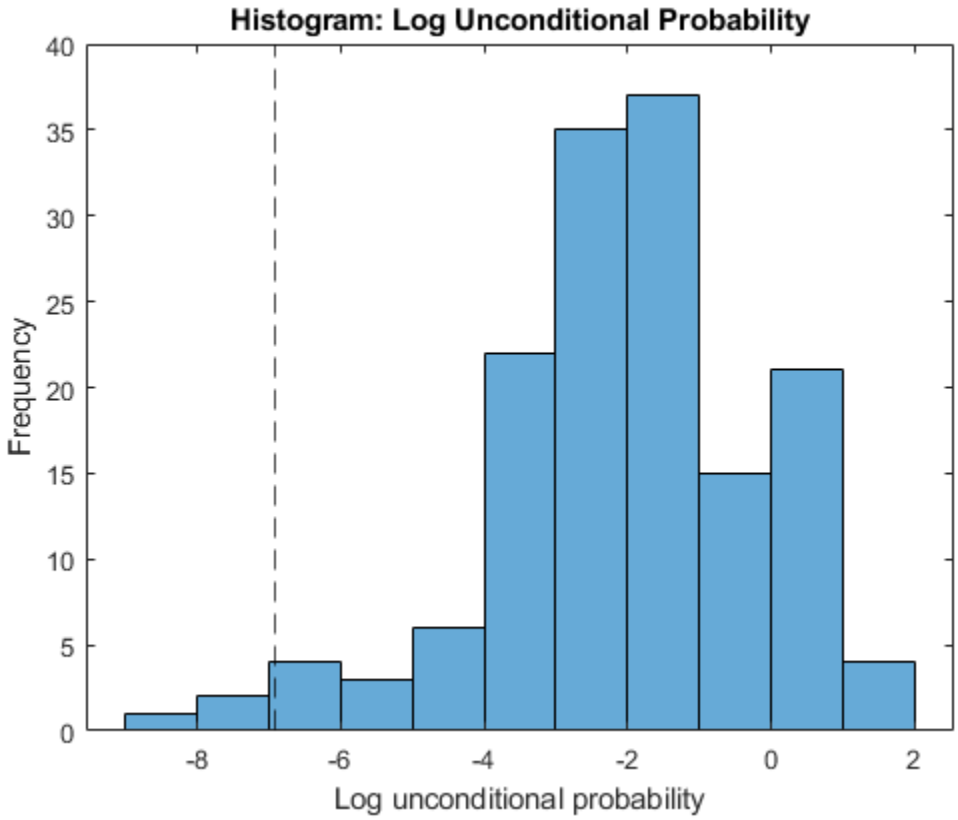
```
lp(ind)
```

```
ans = 4×1

   -7.8995
   -8.4765
   -6.9854
   -7.8969
```

All the outliers are smaller than the lower outlier detection threshold.

Plot the unconditional probability densities.

```
histogram(lp)
hold on
xline(L,'k--')
hold off
xlabel('Log unconditional probability')
ylabel('Frequency')
title('Histogram: Log Unconditional Probability')
```



Input Arguments

`Mdl` — Naive Bayes classification model

`ClassificationNaiveBayes` model object | `CompactClassificationNaiveBayes` model object

Naive Bayes classification model, specified as a `ClassificationNaiveBayes` model object or `CompactClassificationNaiveBayes` model object returned by `fitcnb` or `compact`, respectively.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed. Optionally, `tbl` can contain additional columns for the response variable and observation weights.

If you train `Mdl` using sample data contained in a table, then the input data for `logp` must also be in a table.

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an *instance* or *example*), and each column corresponds to one variable (also known as a *feature*). The variables in the columns of `X` must be the same as the variables that trained the `Mdl` classifier.

The length of `Y` and the number of rows of `X` must be equal.

Data Types: `double` | `single`

More About

Unconditional Probability Density

The unconditional probability density of the predictors is the density's distribution marginalized over the classes.

In other words, the unconditional probability density is

$$P(X_1,..,X_P) = \sum_{k=1}^{K} P(X_1,..,X_P, Y = k) = \sum_{k=1}^{K} P(X_1,..,X_P \Big| y = k)\pi(Y = k),$$

where $\pi(Y = k)$ is the class prior probability. The conditional distribution of the data given the class ($P(X_1,..,X_P|y = k)$) and the class prior probability distributions are training options (that is, you specify them when training the classifier).

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

See Also

`ClassificationNaiveBayes` | `CompactClassificationNaiveBayes` | `fitcnb` | `predict`

Topics

“Naive Bayes Classification” on page 21-2

Introduced in R2014b

logp

Log unconditional probability density of naive Bayes classification model for incremental learning

Syntax

```
lp = logp(Mdl,X)
```

Description

`lp = logp(Mdl,X)` returns the log unconditional probability densities on page 33-3621 `lp` of the observations in the predictor data `X` using the naive Bayes classification model for incremental learning `Mdl`. You can use `lp` to identify outliers in the training data.

Examples

Detect Outliers In Streaming Data

Train a naive Bayes classification model by using `fitcnb`, convert it to an incremental learner, and then use the incremental model to detect outliers in streaming data.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Train Naive Bayes Classification Model

Fit a naive Bayes classification model to a random sample of about 25% of the data.

```
idxtt = randsample([true false false false],n,true);
TTMdl = fitcnb(X(idxtt,:),Y(idxtt))
```

```
TTMdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 6167
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}
```

```
  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained model.

Convert Trained Model

Convert the traditionally trained model to a naive Bayes classification model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 1
                  Metrics: [1×2 table]
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
        DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` object. `IncrementalMdl` represents a naive Bayes classification model for incremental learning; the parameter values are the same as the parameters in `TTMdl`.

Detect Outliers

Determine unconditional density thresholds for outliers by using the traditionally trained model and training data. Observations in the streaming data yielding densities beyond the thresholds are considered outliers.

```
ttlp = logp(TTMdl,X(idxtt,:));
[~,lower,upper] = isoutlier(ttlp)
```

```
lower = -336.0424

upper = 399.9853
```

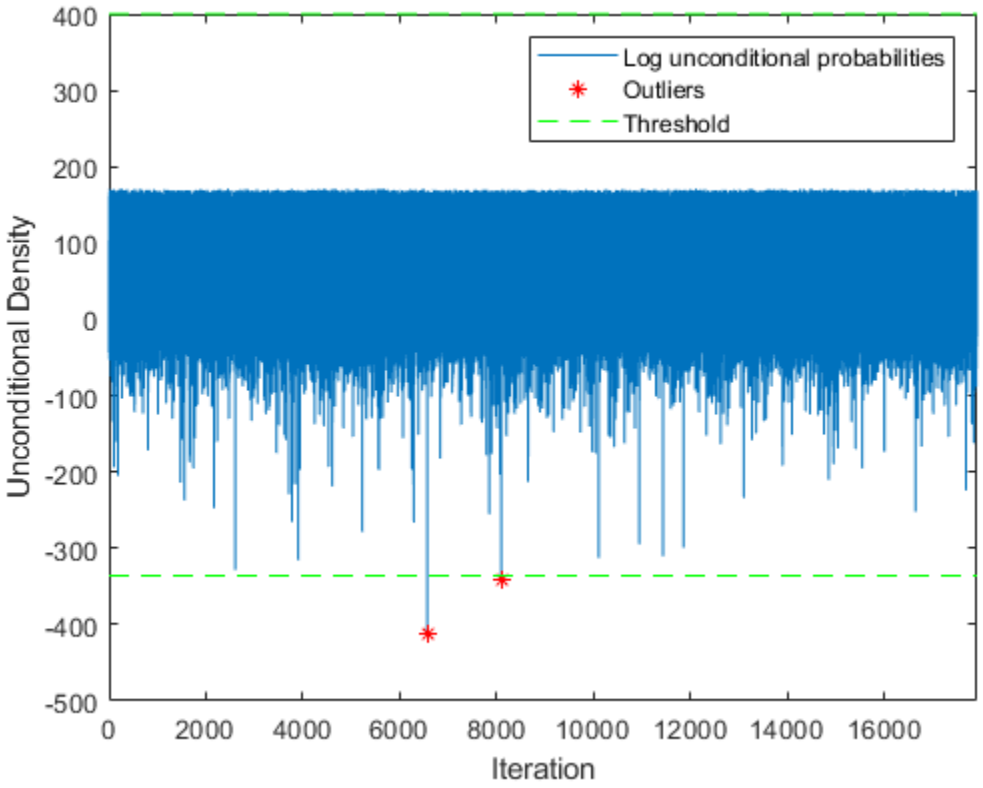
Detect outliers in the rest of the data, relative to what was learned to create `TTMdl`. Simulate a data stream by processing 1 observation at a time. At each iteration, call `logp` to compute the log unconditional probability density of the observation and store each value.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 1;
nchunk = floor(nil/numObsPerChunk);
lp = zeros(nchunk,1);
iso = false(nchunk,1);
Xil = X(idxil,:);
Yil = Y(idxil);

% Incremental fitting
for j = 1:nchunk
```

```
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    lp(j) = logp(IncrementalMdl,Xil(idx,:));
    iso(j) = ((lp(j) < lower) + (lp(j) > upper)) >= 1;
end
```

Plot the log unconditional probability densities of the streaming data. Identify the outliers.

```
figure;
h1 = plot(lp);
hold on
x = 1:nchunk;
h2 = plot(x(iso),lp(iso),'r*');
h3 = line(xlim,[lower lower],'Color','g','LineStyle','--');
line(xlim,[upper upper],'Color','g','LineStyle','--')
xlim([0 nchunk]);
ylabel('Unconditional Density')
xlabel('Iteration')
legend([h1 h2 h3],["Log unconditional probabilities" "Outliers" "Threshold"])
hold off
```



Input Arguments

`Mdl` — Naive Bayes classification model for incremental learning
`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

You must configure `Mdl` to compute the log conditional probability densities on a batch of observations.

- If `Mdl` is a converted, traditionally trained model, you can compute the log conditional probabilities without any modifications.
- Otherwise, `Mdl.DistributionParameters` must be a cell matrix with `Mdl.NumPredictors` > 0 columns and at least one row, where each row corresponds to each class name in `Mdl.ClassNames`.

`X` — Batch of predictor data
floating-point matrix

Batch of predictor data with which to compute the log conditional probability densities, specified as an *n*-by-`Mdl.NumPredictors` floating-point matrix.

Note

- `logp` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.
- For each $j = 1$ through n, if `X(j,:)` contains at least one `NaN`, `lp(j)` is `NaN`.

Data Types: `single` | `double`

Output Arguments

`lp` — Log conditional probability densities
floating-point vector

Log unconditional probability densities on page 33-3621, returned as an *n*-by-1 floating-point vector. `lp(j)` is the log unconditional probability density of the predictors evaluated at `X(j,:)`.

Data Types: `single` | `double`

More About

Unconditional Probability Density

The unconditional probability density of the predictors is the density's distribution marginalized over the classes.

In other words, the unconditional probability density is

$$P(X_1,..,X_P) = \sum_{k=1}^{K} P(X_1,..,X_P, Y = k) = \sum_{k=1}^{K} P(X_1,..,X_P \Big| y = k)\pi(Y = k),$$

where $\pi(Y = k)$ is the class prior probability. The conditional distribution of the data given the class ($P(X_1,..,X_P|y = k)$) and the class prior probability distributions are training options (that is, you specify them when training the classifier).

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8

Introduced in R2021a

loss

Package:

Classification loss for generalized additive model (GAM)

Syntax

```
L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)
L = loss(Mdl,X,Y)
L = loss(___,Name,Value)
```

Description

`L = loss(Mdl,Tbl,ResponseVarName)` returns the "Classification Loss" on page 33-3628 (`L`), a scalar representing how well the generalized additive model `Mdl` classifies the predictor data in `Tbl` compared to the true class labels in `Tbl.ResponseVarName`.

The interpretation of `L` depends on the loss function (`'LossFun'`) and weighting scheme (`'Weights'`). In general, better classifiers yield smaller classification loss values. The default `'LossFun'` value is `'classiferror'` (misclassification rate in decimal).

`L = loss(Mdl,Tbl,Y)` uses the predictor data in table `Tbl` and the true class labels in `Y`.

`L = loss(Mdl,X,Y)` uses the predictor data in matrix `X` and the true class labels in `Y`.

`L = loss(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, `'LossFun','mincost'` sets the loss function to the minimal expected misclassification cost function.

Examples

Determine Test Sample Classification Loss

Determine the test sample classification error (loss) of a generalized additive model. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(XTrain,YTrain,'ClassNames',{'b','g'});
```

`Mdl` is a `ClassificationGAM` model object.

Determine how well the algorithm generalizes by estimating the test sample classification error. By default, the `loss` function of `ClassificationGAM` estimates classification error by using the `'classiferror'` loss (misclassification rate in decimal).

```
L = loss(Mdl,XTest,YTest)
```

```
L = 0.1052
```

The trained classifier misclassifies approximately 11% of the test sample.

Compare GAMs by Examining Classification Loss

Train a generalized additive model (GAM) that contains both linear and interaction terms for predictors, and estimate the classification loss with and without interaction terms. Specify whether to include interaction terms when estimating the classification loss for training and test data.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 50 observations for the new test data set.

```
rng('default') % For reproducibility
n = size(X,1);
newInds = randsample(n,50);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X(inds,:),Y(inds),'ClassNames',{'b','g'},'Interactions',10)
```

```
Mdl =
  ClassificationGAM
```

```
          ResponseName: 'Y'
CategoricalPredictors: []
           ClassNames: {'b'  'g'}
       ScoreTransform: 'logit'
            Intercept: 2.0026
         Interactions: [10x2 double]
      NumObservations: 301


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object.

Compute the resubstitution classification loss both with and without interaction terms in `Mdl`. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
resubl = resubLoss(Mdl)
```

```
resubl = 0
```

```
resubl_nointeraction = resubLoss(Mdl,'IncludeInteractions',false)
```

```
resubl_nointeraction = 0
```

Estimate the classification loss both with and without interaction terms in `Mdl`.

```
l = loss(Mdl,XNew,YNew)
```

```
l = 0.0615
```

```
l_nointeraction = loss(Mdl,XNew,YNew,'IncludeInteractions',false)
```

```
l_nointeraction = 0.0615
```

Including interaction terms does not change the classification loss for `Mdl`. The trained model classifies all training samples correctly and misclassifies approximately 6% of the test samples.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `CompactClassificationGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `CompactClassificationGAM` model object.

- If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table (`Tbl`).
- If you trained `Mdl` using sample data contained in a matrix, then the input data for `loss` must also be in a matrix (`X`).

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`Tbl` must contain all the predictors used to train `Mdl`. Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights.

- The response variable must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.) If the response variable in `Tbl` has the same name as the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.
- The weight values must be a numeric vector. You must specify the observation weights in `Tbl` by using `'Weights'`.

If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X` or `Tbl`.

`Y` must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

If you trained `Mdl` using sample data contained in a matrix, then the input data for `loss` must also be in a matrix.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'IncludeInteractions',false,'Weights',w` specifies to exclude interaction terms from the model and to use the observation weights `w`.

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `'IncludeInteractions'` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`LossFun` — Loss function
`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  For more details on loss functions, see “Classification Loss” on page 33-3628.

- To specify a custom loss function, use function handle notation. The function must have this form:

  `lossvalue = lossfun(C,S,W,Cost)`

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. `n` is the number of observations in `Tbl` or `X`, and `K` is the number of distinct classes (`numel(Mdl.ClassNames)`. The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(p,q) = 1`, if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights.
  - `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

Example: `'LossFun','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored in `Tbl.W`, then specify it as `'W'`.

`loss` normalizes the weights in each class to add up to the value of the prior probability of the respective class.

Data Types: `single` | `double` | `char` | `string`

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Suppose the following:

- L is the weighted average classification loss.
- n is the sample size.
- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so that they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

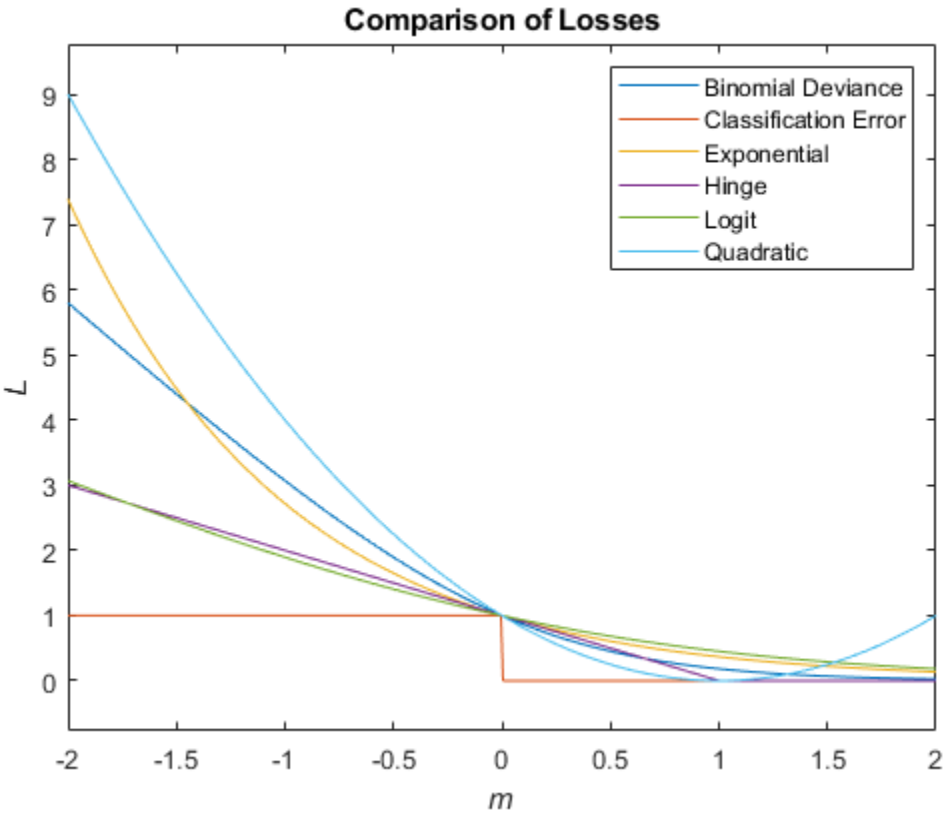
This table describes the supported loss functions that you can specify by using the `'LossFun'` name-value argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.

Loss Function	Value of LossFun	Equation
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`edge` | `margin` | `predict` | `resubLoss`

Topics

"Train Generalized Additive Model for Binary Classification" on page 12-77

Introduced in R2021a

loss

Loss of *k*-nearest neighbor classifier

Syntax

```
L = loss(mdl,tbl,ResponseVarName)
L = loss(mdl,tbl,Y)
L = loss(mdl,X,Y)
L = loss(___,Name,Value)
```

Description

`L = loss(mdl,tbl,ResponseVarName)` returns a scalar representing how well `mdl` classifies the data in `tbl` when `tbl.ResponseVarName` contains the true classifications. If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName`.

When computing the loss, the `loss` function normalizes the class probabilities in `tbl.ResponseVarName` to the class probabilities used for training, which are stored in the `Prior` property of `mdl`.

The meaning of the classification loss (`L`) depends on the loss function and weighting scheme, but, in general, better classifiers yield smaller classification loss values. For more details, see “Classification Loss” on page 33-3635.

`L = loss(mdl,tbl,Y)` returns a scalar representing how well `mdl` classifies the data in `tbl` when `Y` contains the true classifications.

When computing the loss, the `loss` function normalizes the class probabilities in `Y` to the class probabilities used for training, which are stored in the `Prior` property of `mdl`.

`L = loss(mdl,X,Y)` returns a scalar representing how well `mdl` classifies the data in `X` when `Y` contains the true classifications.

When computing the loss, the `loss` function normalizes the class probabilities in `Y` to the class probabilities used for training, which are stored in the `Prior` property of `mdl`.

`L = loss(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input arguments in previous syntaxes. For example, you can specify the loss function and the classification weights.

Examples

Loss Calculation

Create a *k*-nearest neighbor classifier for the Fisher iris data, where $k = 5$.

Load the Fisher iris data set.

```
load fisheriris
```

Create a classifier for five nearest neighbors.

```
mdl = fitcknn(meas,species,'NumNeighbors',5);
```

Examine the loss of the classifier for a mean observation classified as `'versicolor'`.

```
X = mean(meas);
Y = {'versicolor'};
L = loss(mdl,X,Y)
```

```
L = 0
```

All five nearest neighbors classify as `'versicolor'`.

Input Arguments

`mdl` — *k*-nearest neighbor classifier model

`ClassificationKNN` object

k-nearest neighbor classifier model, specified as a `ClassificationKNN` object.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `mdl` using sample data contained in a `table`, then the input data for `loss` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name

name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable is stored as `tbl.response`, then specify it as `'response'`. Otherwise, the software treats all columns of `tbl`, including `tbl.response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of X represents one observation, and each column represents one variable.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. Each row of Y represents the classification of the corresponding row of X.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(mdl,tbl,'response','LossFun','exponential','Weights','w')` returns the weighted exponential loss of `mdl` classifying the data in `tbl`. Here, `tbl.response` is the response variable, and `tbl.w` is the weight variable.

LossFun — Loss function

`'mincost'` (default) | `'binodeviance'` | `'classiferror'` | `'exponential'` | `'hinge'` | `'logit'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or a function handle.

- The following table lists the available loss functions.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. By default, *k*-nearest neighbor models return posterior probabilities as classification scores (see `predict`).

- You can specify a function handle for a custom loss function using `@` (for example, `@lossfun`). Let *n* be the number of observations in X and *K* be the number of distinct classes (`numel(mdl.ClassNames)`). Your custom loss function must have this form:

  ```
  function lossvalue = lossfun(C,S,W,Cost)
  ```

- `C` is an *n*-by-*K* logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `mdl.ClassNames`. Construct `C` by setting `C(p,q) = 1`, if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
- `S` is an *n*-by-*K* numeric matrix of classification scores. The column order corresponds to the class order in `mdl.ClassNames`. The argument `S` is a matrix of classification scores, similar to the output of `predict`.
- `W` is an *n*-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes the weights to sum to `1`.
- `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.
- The output argument `lossvalue` is a scalar.

For more details on loss functions, see “Classification Loss” on page 33-3635.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of a variable in `tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `tbl`.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `tbl`.

If you specify `Weights` as the name of a variable in `tbl`, the name must be a character vector or string scalar. For example, if the weights are stored as `tbl.w`, then specify `Weights` as `'w'`. Otherwise, the software treats all columns of `tbl`, including `tbl.w`, as predictors.

`loss` normalizes the weights so that observation weights in each class sum to the prior probability of that class. When you supply `Weights`, `loss` computes the weighted classification loss.

Example: `'Weights','w'`

Data Types: `single` | `double` | `char` | `string`

Algorithms

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.

- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

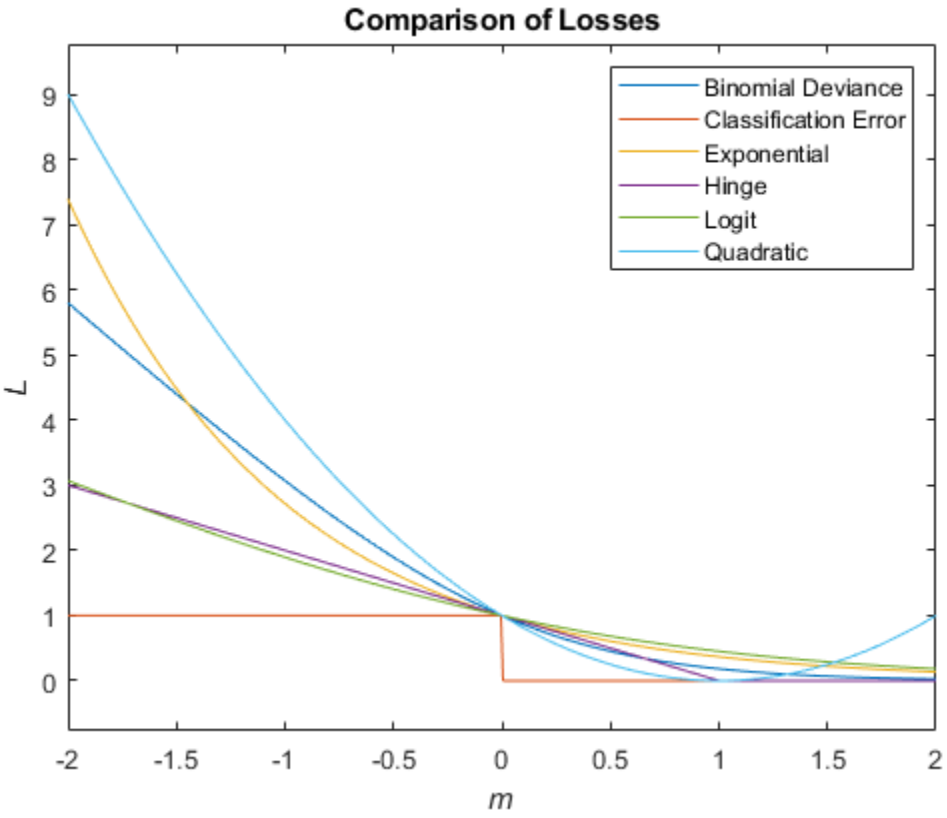
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.

Loss Function	Value of LossFun	Equation
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = (f(X_j)'C)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\hat{y}_j = \underset{k=1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



True Misclassification Cost

Two costs are associated with KNN classification: the true misclassification cost per class and the expected misclassification cost per observation.

You can set the true misclassification cost per class by using the `'Cost'` name-value pair argument when you run `fitcknn`. The value `Cost(i,j)` is the cost of classifying an observation into class `j` if its true class is `i`. By default, `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Expected Cost

Two costs are associated with KNN classification: the true misclassification cost per class and the expected misclassification cost per observation. The third output of `predict` is the expected misclassification cost per observation.

Suppose you have `Nobs` observations that you want to classify with a trained classifier `mdl`, and you have `K` classes. You place the observations into a matrix `Xnew` with one observation per row. The command

```
[label,score,cost] = predict(mdl,Xnew)
```

returns a matrix `cost` of size `Nobs`-by-`K`, among other outputs. Each row of the `cost` matrix contains the expected (average) cost of classifying the observation into each of the `K` classes. `cost(n,j)` is

$$\sum_{i=1}^{K} \widehat{P}(i|Xnew(n))C(j|i),$$

where

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability on page 33-4786 of class i for observation $Xnew(n)$.
- $C(j|i)$ is the true misclassification cost of classifying an observation as j when its true class is i.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.
- `loss` executes on a GPU in these cases only:
  - The input argument `X` is a `gpuArray`.
  - The input argument `tbl` contains `gpuArray` elements.
  - The input argument `mdl` was fitted with GPU array input arguments.

See Also

`ClassificationKNN` | `edge` | `fitcknn` | `margin`

Topics

“Examine Quality of KNN Classifier” on page 18-28
“Predict Classification Using KNN Classifier” on page 18-29
“Modify KNN Classifier” on page 18-29
“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2012a

loss

Class: `ClassificationLinear`

Classification loss for linear classification models

Syntax

```
L = loss(Mdl,X,Y)

L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(Mdl,X,Y)` returns the classification losses on page 33-3647 for the binary, linear classification model `Mdl` using predictor data in `X` and corresponding class labels in `Y`. `L` contains classification error rates for each regularization strength in `Mdl`.

`L = loss(Mdl,Tbl,ResponseVarName)` returns the classification losses for the predictor data in `Tbl` and the true class labels in `Tbl.ResponseVarName`.

`L = loss(Mdl,Tbl,Y)` returns the classification losses for the predictor data in table `Tbl` and the true class labels in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that columns in the predictor data correspond to observations or specify the classification loss function.

Input Arguments

`Mdl` — Binary, linear classification model

`ClassificationLinear` model object

Binary, linear classification model, specified as a `ClassificationLinear` model object. You can create a `ClassificationLinear` model object using `fitclinear`.

`X` — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix. This orientation of `X` indicates that rows correspond to individual observations, and columns correspond to individual predictor variables.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

ResponseVarName — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

LossFun — Loss function
'classiferror' (default) | 'binodeviance' | 'exponential' | 'hinge' | 'logit' | 'mincost' | 'quadratic' | function handle

Loss function, specified as the comma-separated pair consisting of 'LossFun' and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
'binodeviance'	Binomial deviance
'classiferror'	Misclassified rate in decimal
'exponential'	Exponential loss
'hinge'	Hinge loss
'logit'	Logistic loss
'mincost'	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
'quadratic'	Quadratic loss

  'mincost' is appropriate for classification scores that are posterior probabilities. For linear classification models, logistic regression learners return posterior probabilities as classification scores by default, but SVM learners do not (see predict).
- To specify a custom loss function, use function handle notation. The function must have this form:

  lossvalue = *lossfun*(C,S,W,Cost)

  - The output argument lossvalue is a scalar.
  - You specify the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating the class to which the corresponding observation belongs. n is the number of observations in Tbl or X, and K is the number of distinct classes (numel(Mdl.ClassNames). The column order corresponds to the class order in Mdl.ClassNames. Create C by setting C(p,q) = 1, if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in Mdl.ClassNames. S is a matrix of classification scores, similar to the output of predict.
  - W is an n-by-1 numeric vector of observation weights.
  - Cost is a K-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) – eye(K) specifies a cost of 0 for correct classification and 1 for misclassification.

Example: 'LossFun',@*lossfun*

Data Types: char | string | function_handle

ObservationsIn — Predictor data observation dimension
'rows' (default) | 'columns'

Predictor data observation dimension, specified as 'rows' or 'columns'.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of observations in `X` or `Tbl`.
- If you specify `Weights` as the name of a variable in `Tbl`, then the name must be a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, then for each regularization strength, `loss` computes the weighted classification loss on page 33-3647 and normalizes weights to sum up to the value of the prior probability in the respective class.

Data Types: `double` | `single`

Output Arguments

`L` — Classification losses
numeric scalar | numeric row vector

Classification losses, returned as a numeric scalar or row vector. The interpretation of `L` depends on `Weights` and `LossFun`.

`L` is the same size as `Mdl.Lambda`. `L(`*j*`)` is the classification loss of the linear classification model trained using the regularization strength `Mdl.Lambda(`*j*`)`.

Examples

Estimate Test-Sample Classification Loss

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. Specify to hold out 30% of the observations. Optimize the objective function using SpaRSA.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','Holdout',0.30);
CMdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `ClassificationLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Estimate the training- and test-sample classification error.

```
ceTrain = loss(CMdl,X(trainIdx,:),Ystats(trainIdx))
```

```
ceTrain = 1.3572e-04
```

```
ceTest = loss(CMdl,X(testIdx,:),Ystats(testIdx))
```

```
ceTest = 5.2804e-04
```

Because there is one regularization strength in `CMdl`, `ceTrain` and `ceTest` are numeric scalars.

Specify Custom Classification Loss

Load the NLP data set. Preprocess the data as in “Estimate Test-Sample Classification Loss” on page 33-3643, and transpose the predictor data.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Train a binary, linear classification model. Specify to hold out 30% of the observations. Optimize the objective function using SpaRSA. Specify that the predictor observations correspond to columns.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','Holdout',0.30,...
    'ObservationsIn','columns');
CMdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `ClassificationLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Create an anonymous function that measures linear loss, that is,

$$L = \frac{\sum_j - w_j y_j f_j}{\sum_j w_j}.$$

w_j is the weight for observation j, y_j is response j (-1 for the negative class, and 1 otherwise), and f_j is the raw classification score of observation j. Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `LossFun` name-value pair argument.

```
linearloss = @(C,S,W,Cost)sum(-W.*sum(S.*C,2))/sum(W);
```

Estimate the training- and test-sample classification loss using the linear loss function.

```
ceTrain = loss(CMdl,X(:,trainIdx),Ystats(trainIdx),'LossFun',linearloss,...
    'ObservationsIn','columns')
```

```
ceTrain = -7.8330
```

```
ceTest = loss(CMdl,X(:,testIdx),Ystats(testIdx),'LossFun',linearloss,...
    'ObservationsIn','columns')
```

```
ceTest = -7.7383
```

Find Good Lasso Penalty Using Classification Loss

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare test-sample classification error rates.

Load the NLP data set. Preprocess the data as in “Specify Custom Classification Loss” on page 33-3644.

```
load nlpdata
Ystats = Y == 'stats';
X = X';

rng(10); % For reproducibility
Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition);
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Train binary, linear classification models that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'CVPartition',Partition,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
```

```
        NumObservations: 31572
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract the trained linear classification model.

```
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the test-sample classification error.

```
ce = loss(Mdl,X(:,testIdx),Ystats(testIdx),'ObservationsIn','columns');
```

Because there are 11 regularization strengths, `ce` is a 1-by-11 vector of classification error rates.

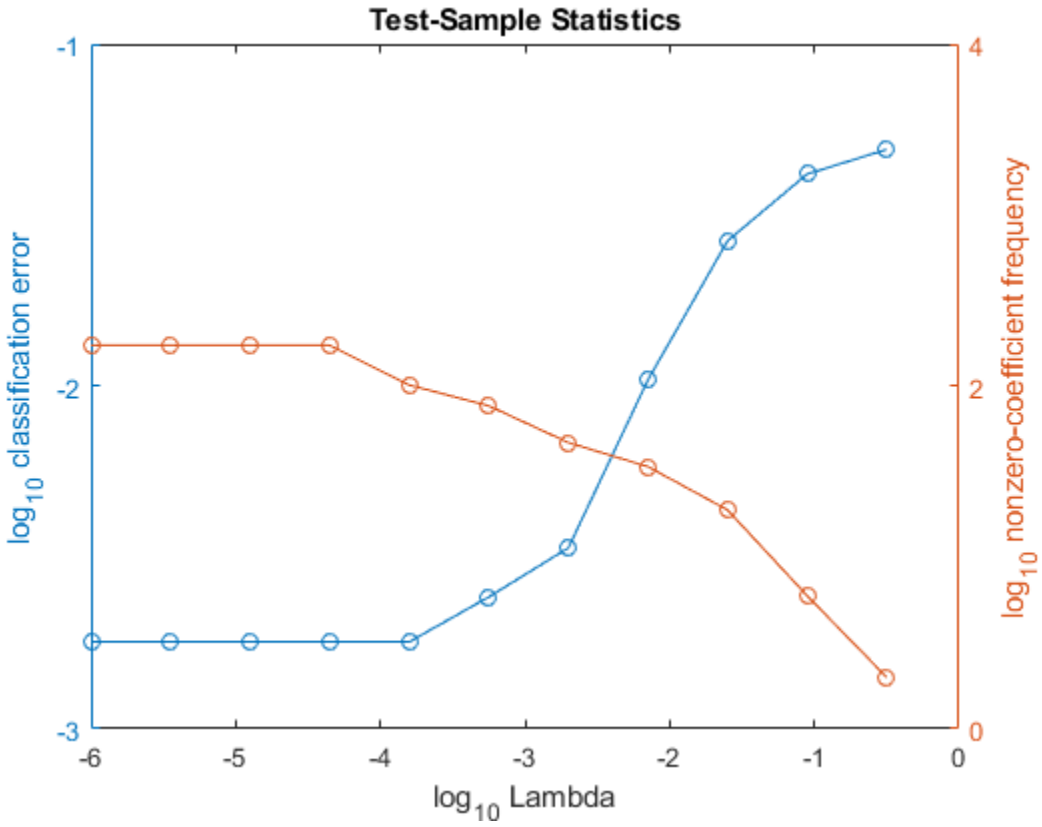
Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the test-sample error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(ce),...
    log10(Lambda),log10(numNZCoeff + 1));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} classification error')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
```

```
title('Test-Sample Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low classification error. In this case, a value between 10^{-4} to 10^{-1} should suffice.

```
idxFinal = 7;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.

- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

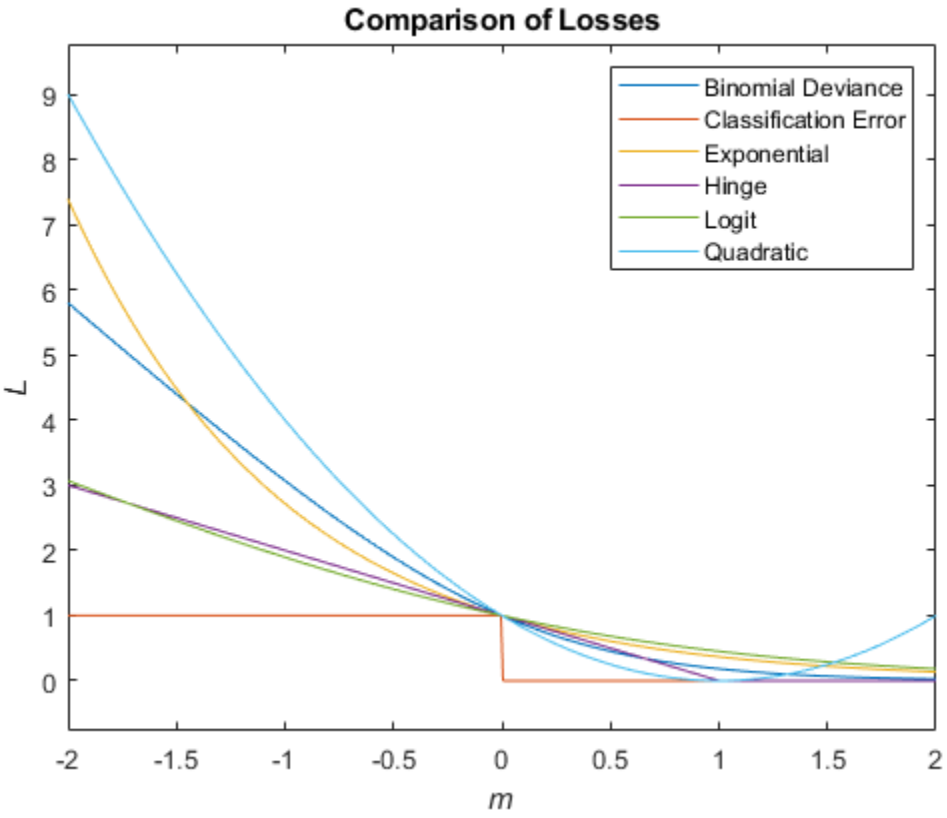
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.

Loss Function	Value of LossFun	Equation
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $$L = -\sum_{j=1}^{n}\frac{\tilde{w}_j \log(m_j)}{Kn},$$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Algorithms

By default, observation weights are prior class probabilities. If you supply weights using `Weights`, then the software normalizes them to sum to the prior probabilities in the respective classes. The software uses the renormalized weights to estimate the weighted classification loss.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `loss` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationLinear` | `fitclinear` | `predict`

Introduced in R2016a

loss

Classification error

Syntax

```
L = loss(obj,X,Y)
L = loss(obj,X,Y,Name,Value)
```

Description

`L = loss(obj,X,Y)` returns the classification loss on page 33-3654, which is a scalar representing how well `obj` classifies the data in `X`, when `Y` contains the true classifications.

When computing the loss, `loss` normalizes the class probabilities in `Y` to the class probabilities used for training, stored in the `Prior` property of `obj`.

`L = loss(obj,X,Y,Name,Value)` returns the loss with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier of class `ClassificationDiscriminant` or `CompactClassificationDiscriminant`, typically constructed with `fitcdiscr`.

`X`

Matrix where each row represents an observation, and each column represents a predictor. The number of columns in `X` must equal the number of predictors in `obj`.

`Y`

Class labels, with the same data type as exists in `obj`. The number of elements of `Y` must equal the number of rows of `X`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun`

Built-in, loss-function name (character vector or string scalar in the table) or function handle.

- The following table lists the available loss functions. Specify one using the corresponding value.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

`'mincost'` is appropriate for classification scores that are posterior probabilities. Discriminant analysis models return posterior probabilities as classification scores by default (see `predict`).

- Specify your own function using function handle notation.

  Suppose that *n* be the number of observations in `X` and *K* be the number of distinct classes (`numel(Mdl.ClassNames)`). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an *n*-by-*K* logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an *n*-by-*K* numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an *n*-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.
  - `Cost` is a *K*-by-*K* numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see “Classification Loss” on page 33-3654.

Default: `'mincost'`

`Weights`

Numeric vector of length `N`, where `N` is the number of rows of `X`. `weights` are nonnegative. `loss` normalizes the weights so that observation weights in each class sum to the prior probability of that class. When you supply `weights`, `loss` computes weighted classification loss.

Default: `ones(N,1)`

Output Arguments

L

Classification loss on page 33-3654, a scalar. The interpretation of `L` depends on the values in `weights` and `lossfun`.

Examples

Estimate Classification Error

Load Fisher's iris data set.

```
load fisheriris
```

Train a discriminant analysis model using all observations in the data.

```
Mdl = fitcdiscr(meas,species);
```

Estimate the classification error of the model using the training observations.

```
L = loss(Mdl,meas,species)
```

```
L = 0.0200
```

Alternatively, if `Mdl` is not compact, then you can estimate the training-sample classification error by passing `Mdl` to `resubLoss`.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* =$

[0 0 1 0]′. The order of the classes corresponds to the order in the `ClassNames` property of the input model.

- $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
- $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.

- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

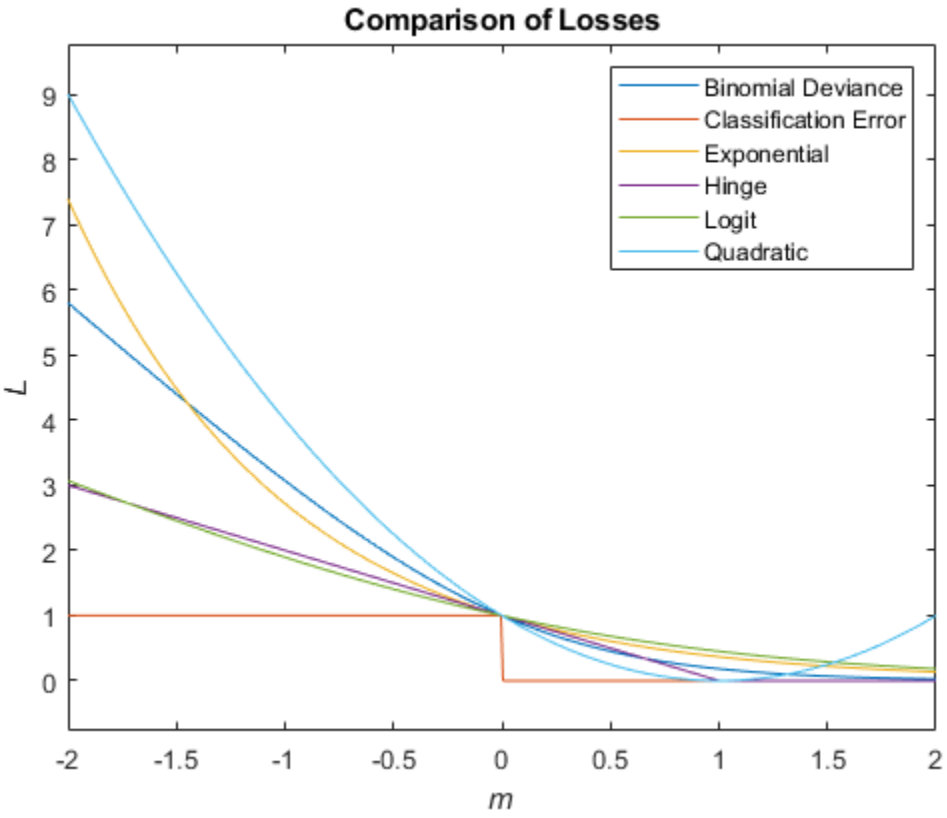
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp\left(-m_j\right).$
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\left\{0, 1 - m_j\right\}.$
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log\left(1 + \exp\left(-m_j\right)\right).$

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Posterior Probability

The posterior probability that a point x belongs to class k is the product of the prior probability and the multivariate normal density. The density function of the multivariate normal with 1-by-d mean μ_k and d-by-d covariance Σ_k at a 1-by-d point x is

$$P(x|k) = \frac{1}{\left((2\pi)^d|\Sigma_k|\right)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^T\right),$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix.

Let $P(k)$ represent the prior probability of class k. Then the posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)},$$

where $P(x)$ is a normalization constant, the sum over k of $P(x|k)P(k)$.

Prior Probability

The prior probability is one of three choices:

- `'uniform'` — The prior probability of class `k` is one over the total number of classes.
- `'empirical'` — The prior probability of class `k` is the number of training samples of class `k` divided by the total number of training samples.

- Custom — The prior probability of class `k` is the `k`th element of the `prior` vector. See `fitcdiscr`.

After creating a classification model (`Mdl`) you can set the prior using dot notation:

```
Mdl.Prior = v;
```

where `v` is a vector of positive elements representing the frequency with which each element occurs. You do not need to retrain the classifier when you set a new prior.

Cost

The matrix of expected costs per observation is defined in “Cost” on page 20-7.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

See Also
`ClassificationDiscriminant` | `edge` | `fitcdiscr` | `margin` | `predict`

Topics
“Discriminant Analysis Classification” on page 20-2

loss

Package:

Classification loss for multiclass error-correcting output codes (ECOC) model

Syntax

```
L = loss(Mdl,tbl,ResponseVarName)
L = loss(Mdl,tbl,Y)

L = loss(Mdl,X,Y)

L = loss(___,Name,Value)
```

Description

`L = loss(Mdl,tbl,ResponseVarName)` returns the classification loss (`L`), a scalar representing how well the trained multiclass error-correcting output codes (ECOC) model `Mdl` classifies the predictor data in `tbl` compared to the true class labels in `tbl.ResponseVarName`. By default, `loss` uses the classification error on page 33-3665 to compute `L`.

`L = loss(Mdl,tbl,Y)` returns the classification loss for the predictor data in table `tbl` and the true class labels in `Y`.

`L = loss(Mdl,X,Y)` returns the classification loss for the predictor data in matrix `X` and the true class labels in `Y`.

`L = loss(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify a decoding scheme, classification loss function, and verbosity level.

Examples

Determine Test-Sample Loss of ECOC Model

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Specify a 15% holdout sample, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.15,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

Estimate the test-sample classification error, which is the default classification loss.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
L = loss(Mdl,XTest,YTest)
```

```
L = 0
```

The ECOC model correctly classifies all irises in the test sample.

Determine ECOC Model Quality Using Custom Loss

Determine the quality of an ECOC model by using a custom loss function that considers the minimal binary loss for each observation.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);  % Class order
rng(1) % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Specify a 15% holdout sample, standardize the predictors using an SVM template, and define the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.15,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

Create a function that takes the minimal loss for each observation, then averages the minimal losses for all observations. `S` corresponds to the `NegLoss` output of `predict`.

```
lossfun = @(~,S,~,~)mean(min(-S,[],2));
```

Compute the test-sample custom loss.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
loss(Mdl,XTest,YTest,'LossFun',lossfun)
```

```
ans = 0.0033
```

The average minimal binary loss for the test-sample observations is `0.0033`.

Input Arguments

`Mdl` — Full or compact multiclass ECOC model

`ClassificationECOC` model object | `CompactClassificationECOC` model object

Full or compact multiclass ECOC model, specified as a `ClassificationECOC` or `CompactClassificationECOC` model object.

To create a full or compact ECOC model, see `ClassificationECOC` or `CompactClassificationECOC`.

`tbl` — Sample data

table

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain additional columns for the response variable and observation weights. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you train `Mdl` using sample data contained in a `table`, then the input data for `loss` must also be in a table.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `tbl.y`, then specify `ResponseVarName` as `'y'`. Otherwise, the software treats all columns of `tbl`, including `tbl.y`, as predictors.

The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one variable. The variables in the columns of `X` must be the same as the variables that trained the classifier `Mdl`.

The number of rows in `X` must equal the number of rows in `Y`.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `double` | `single`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `Y` must have the same data type as `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The number of rows in `Y` must equal the number of rows in `tbl` or `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(Mdl,X,Y,'BinaryLoss','hinge','LossFun',@lossfun)` specifies `'hinge'` as the binary learner loss function and the custom function handle `@lossfun` as the overall loss function.

BinaryLoss — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the *K*-by-*L* coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-*L* row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - *K* is the number of classes.
  - *L* is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme

`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3666.

Example: `'Decoding','lossbased'`

`LossFun` — Loss function

`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'classiferror'` or a function handle.

- Specify the built-in function `'classiferror'`. In this case, the loss function is the classification error on page 33-3665, which is the proportion of misclassified observations.
- Or, specify your own function using function handle notation.

  Assume that `n = size(X,1)` is the sample size and `K` is the number of classes. Your function must have the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of negated loss values for the classes. Each row corresponds to an observation. The column order corresponds to the class order in `Mdl.ClassNames`. The input `S` resembles the output argument `NegLoss` of `predict`.
  - `W` is an `n`-by-`1` numeric vector of observation weights. If you pass `W`, the software normalizes its elements to sum to `1`.
  - `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of 0 for correct classification and 1 for misclassification.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`. `Mdl.BinaryLearners` must contain `ClassificationLinear` models.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, you can experience a significant reduction in execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level

`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | numeric vector | name of variable in `tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `tbl`. If you supply weights, then `loss` computes the weighted loss.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `tbl`.

If you specify `Weights` as the name of a variable in `tbl`, you must do so as a character vector or string scalar. For example, if the weights are stored as `tbl.w`, then specify `Weights` as `'w'`. Otherwise, the software treats all columns of `tbl`, including `tbl.w`, as predictors.

If you do not specify your own loss function (using `LossFun`), then the software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`L` — Classification loss

numeric scalar | numeric row vector

Classification loss, returned as a numeric scalar or row vector. `L` is a generalization or resubstitution quality measure. Its interpretation depends on the loss function and weighting scheme, but in general, better classifiers yield smaller classification loss values.

If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `L` is a 1-by-ℓ vector, where ℓ is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`). The value `L(j)` is the loss for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.

Otherwise, `L` is a scalar value.

More About

Classification Error

The classification error is a binary classification error measure that has the form

$$L = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j},$$

where:

- w_j is the weight for observation j. The software renormalizes the weights to sum to 1.
- $e_j = 1$ if the predicted class of observation j differs from its true class, and 0 otherwise.

In other words, the classification error is the proportion of observations misclassified by the classifier.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j).$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|}.$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_j s_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_j s_j)/2$

Value	Description	Score Domain	$g(y_j,s_j)$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_j s_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_j s_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_j s_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_j s_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `loss` does not support tall `table` data when `Mdl` contains kernel or linear binary learners.

For more information, see “Tall Arrays”.

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `fitcecoc` | `predict` | `resubLoss`

Topics

Introduced in R2014b

loss

Classification error

Syntax

```
L = loss(ens,tbl,ResponseVarName)
L = loss(ens,tbl,Y)
L = loss(ens,X,Y)
L = loss( ___ ,Name,Value)
```

Description

`L = loss(ens,tbl,ResponseVarName)` returns the classification error for ensemble `ens` computed using table of predictors `tbl` and true class labels `tbl.ResponseVarName`.

`L = loss(ens,tbl,Y)` returns the classification error for ensemble `ens` computed using table of predictors `tbl` and true class labels `Y`.

`L = loss(ens,X,Y)` returns the classification error for ensemble `ens` computed using matrix of predictors `X` and true class labels `Y`.

`L = loss( ___ ,Name,Value)` computes classification error with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

When computing the loss, `loss` normalizes the class probabilities in `ResponseVarName` or `Y` to the class probabilities used for training, stored in the `Prior` property of `ens`.

Input Arguments

`ens`

Classification ensemble created with `fitcensemble`, or a compact classification ensemble created with `compact`.

`tbl`

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all of the predictors used to train the model. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `ens` using sample data contained in a `table`, then the input data for this method must also be in a table.

`ResponseVarName`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `tbl`, including `Y`, as predictors when training the model.

X

Matrix of data to classify. Each row of X represents one observation, and each column represents one predictor. X must have the same number of columns as the data used to train `ens`. X should have the same number of rows as the number of elements in Y.

If you trained `ens` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

Y

Class labels of observations in `tbl` or X. Y should be of the same type as the classification used to train `ens`, and its number of elements should equal the number of rows of `tbl` or X.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `loss` uses only these learners for calculating loss.

Default: `1:NumTrained`

`Lossfun`

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities.

  - Bagged and subspace ensembles return posterior probabilities by default (`ens.Method` is `'Bag'` or `'Subspace'`).

- If the ensemble method is `'AdaBoostM1'`, `'AdaBoostM2'`, `GentleBoost`, or `'LogitBoost'`, then, to use posterior probabilities as classification scores, you must specify the double-logit score transform by entering

  ```
  ens.ScoreTransform = 'doublelogit';
  ```

- For all other ensemble methods, the software does not support posterior probabilities as classification scores.

- Specify your own function using function handle notation.

  Suppose that `n` be the number of observations in `X` and `K` be the number of distinct classes (`numel(ens.ClassNames)`, `ens` is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `ens.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `ens.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.
  - `Cost` is a *K*-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see “Classification Loss” on page 33-3672.

Default: `'classiferror'`

`mode`

Meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

`UseObsForLearner`

A logical matrix of size `N`-by-`T`, where:

- N is the number of rows of X.
- T is the number of weak learners in ens.

When UseObsForLearner(i,j) is true, learner j is used in predicting the class of row i of X.

Default: true(N,T)

weights

Vector of observation weights, with nonnegative entries. The length of weights must equal the number of rows in X. When you specify weights, loss normalizes the weights so that observation weights in each class sum to the prior probability of that class.

Default: ones(size(X,1),1)

Output Arguments

L

Classification loss on page 33-3672, by default the fraction of misclassified data. L can be a vector, and can mean different things, depending on the name-value pair settings.

Examples

Estimate Classification Error

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification ensemble of 100 decision trees using AdaBoostM2. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Estimate the classification error of the model using the training observations.

```
L = loss(ens,meas,species)
```

```
L = 0.0333
```

Alternatively, if ens is not compact, then you can estimate the training-sample classification error by passing ens to resubLoss.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K - 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

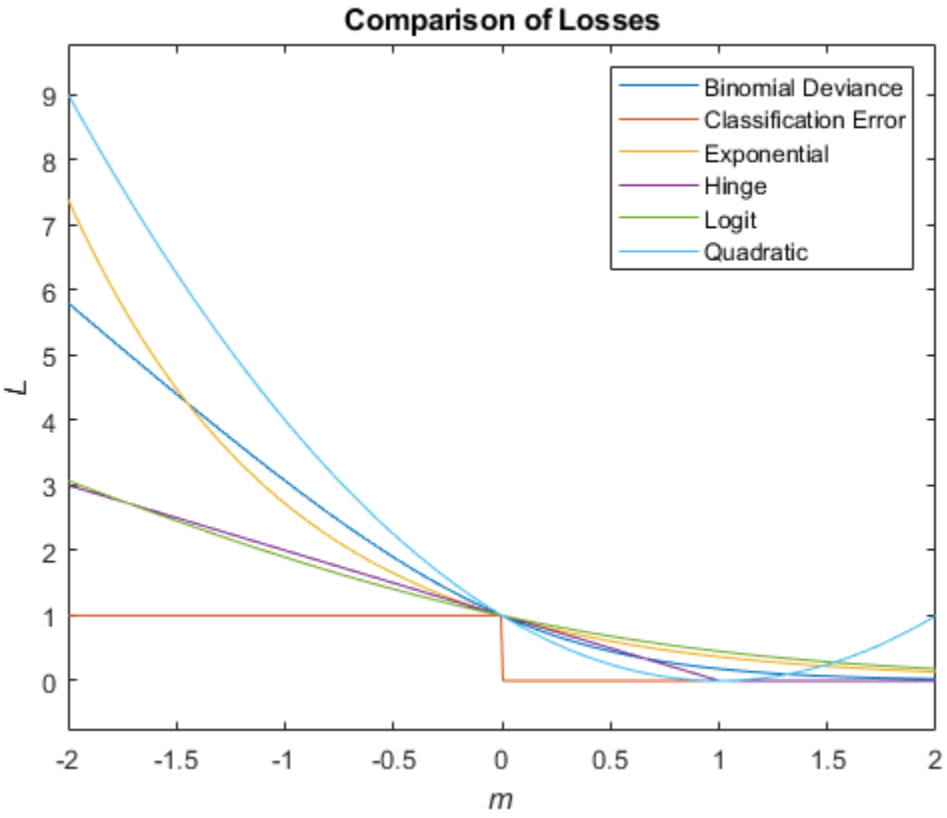
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.

Loss Function	Value of LossFun	Equation
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $$L = -\sum_{j=1}^{n}\frac{\tilde{w}_j \log(m_j)}{Kn},$$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`edge` | `loss` | `margin` | `predict`

loss

Classification loss for naive Bayes classifier

Syntax

```
L = loss(Mdl,tbl,ResponseVarName)
L = loss(Mdl,tbl,Y)

L = loss(Mdl,X,Y)

L = loss(___,Name,Value)
```

Description

`L = loss(Mdl,tbl,ResponseVarName)` returns the "Classification Loss" on page 33-3682, a scalar representing how well the trained naive Bayes classifier `Mdl` classifies the predictor data in table `tbl` compared to the true class labels in `tbl.ResponseVarName`.

`loss` normalizes the class probabilities in `tbl.ResponseVarName` to the prior class probabilities used by `fitcnb` for training, which are stored in the `Prior` property of `Mdl`.

`L = loss(Mdl,tbl,Y)` returns the classification loss for the predictor data in table `tbl` and the true class labels in `Y`.

`L = loss(Mdl,X,Y)` returns the classification loss based on the predictor data in matrix `X` compared to the true class labels in `Y`.

`L = loss(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify the loss function and the classification weights.

Examples

Determine Test Sample Classification Loss of Naive Bayes Classifier

Determine the test sample classification error (loss) of a naive Bayes classifier. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 105
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Determine how well the algorithm generalizes by estimating the test sample classification error.

```
L = loss(Mdl,XTest,YTest)
```

```
L = 0.0444
```

The naive Bayes classifier misclassifies approximately 4% of the test sample.

You might decrease the classification error by specifying better predictor distributions when you train the classifier with `fitcnb`.

Determine Test Sample Logit Loss of Naive Bayes Classifier

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
```

```
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Determine how well the algorithm generalizes by estimating the test sample logit loss.

```
L = loss(Mdl,XTest,YTest,'LossFun','logit')
```

```
L = 0.3359
```

The logit loss is approximately 0.34.

Input Arguments

`Mdl` — Naive Bayes classification model

`ClassificationNaiveBayes` model object | `CompactClassificationNaiveBayes` model object

Naive Bayes classification model, specified as a `ClassificationNaiveBayes` model object or `CompactClassificationNaiveBayes` model object returned by `fitcnb` or `compact`, respectively.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed. Optionally, `tbl` can contain additional columns for the response variable and observation weights.

If you train `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

ResponseVarName — Response variable name
name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `y` is stored as `tbl.y`, then specify it as `'y'`. Otherwise, the software treats all columns of `tbl`, including `y`, as predictors.

If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

X — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an *instance* or *example*), and each column corresponds to one variable (also known as a *feature*). The variables in the columns of `X` must be the same as the variables that trained the `Mdl` classifier.

The length of `Y` and the number of rows of `X` must be equal.

Data Types: `double` | `single`

Y — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. `Y` must have the same data type as `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The length of `Y` must be equal to the number of rows of `tbl` or `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(Mdl,tbl,Y,'Weights',W)` weighs the observations in each row of `tbl` using the corresponding weight in each row of the variable `W`.

LossFun — Loss function
`'mincost'` (default) | `'binodeviance'` | `'classiferror'` | `'exponential'` | `'hinge'` | `'logit'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. Naive Bayes models return posterior probabilities as classification scores by default (see `predict`).

- Specify your own function using function handle notation.

  Suppose that `n` is the number of observations in `X` and `K` is the number of distinct classes (`numel(Mdl.ClassNames)`, where `Mdl` is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`.

    Create `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes the weights to sum to `1`.
  - `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

  Specify your function using `'LossFun',@`*`lossfun`*.

For more details on loss functions, see “Classification Loss” on page 33-3682.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | numeric vector | name of a variable in `tbl`

Observation weights, specified as a numeric vector or the name of a variable in `tbl`. The software weighs the observations in each row of `X` or `tbl` with the corresponding weights in `Weights`.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows of `X` or `tbl`.

If you specify `Weights` as the name of a variable in `tbl`, then the name must be a character vector or string scalar. For example, if the weights are stored as `tbl.w`, then specify `Weights` as `'w'`. Otherwise, the software treats all columns of `tbl`, including `tbl.w`, as predictors.

If you do not specify a loss function, then the software normalizes `Weights` to add up to `1`.

Data Types: `double` | `char` | `string`

Output Arguments

`L` — Classification loss

scalar

Classification loss, returned as a scalar. `L` is a generalization or resubstitution quality measure. Its interpretation depends on the loss function and weighting scheme; in general, better classifiers yield smaller loss values.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.

- $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
- $m_j = y_j^* f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.

- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

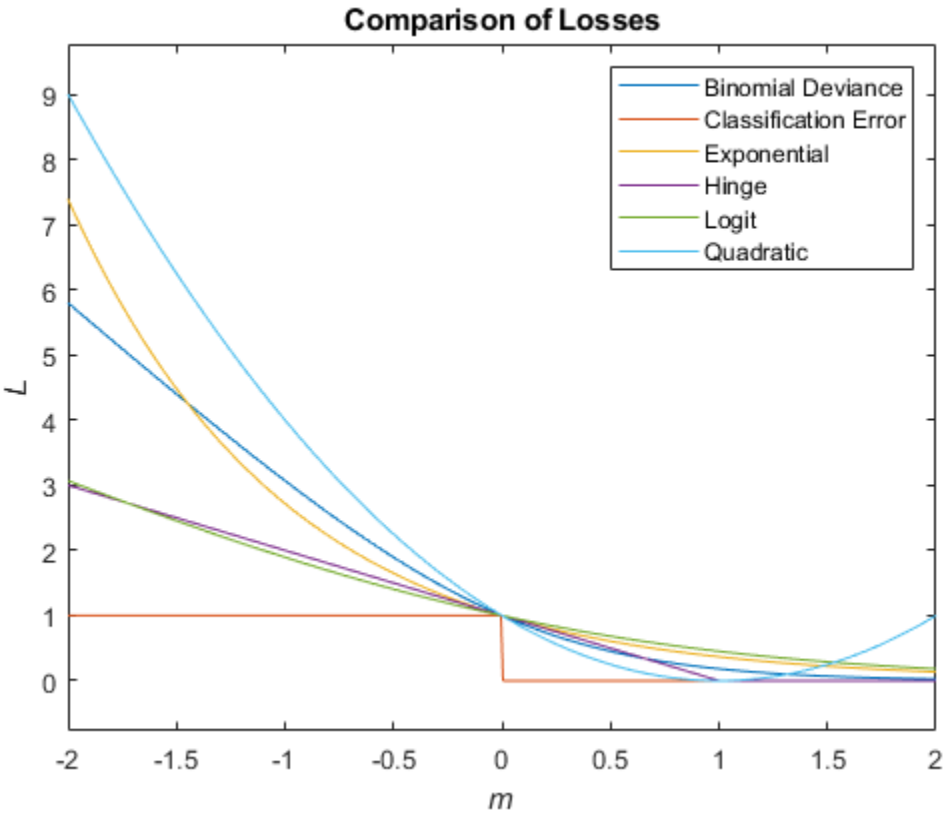
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\widehat{y}_j \neq y_j\right\}.$ $\widehat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\widetilde{w}_j \log(m_j)}{Kn},$ where the weights $\widetilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp\left(-m_j\right).$
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\left\{0, 1 - m_j\right\}.$
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log\left(1 + \exp\left(-m_j\right)\right).$

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Misclassification Cost

A misclassification cost is the relative severity of a classifier labeling an observation into the wrong class.

There are two types of misclassification costs: true and expected. Let K be the number of classes.

- True misclassification cost — A K-by-K matrix, where element (i,j) indicates the misclassification cost of predicting an observation into class j if its true class is i. The software stores the misclassification cost in the property `Mdl.Cost`, and uses it in computations. By default, `Mdl.Cost(i,j) = 1` if `i ≠ j`, and `Mdl.Cost(i,j) = 0` if `i = j`. In other words, the cost is `0` for correct classification and `1` for any incorrect classification.
- Expected misclassification cost — A K-dimensional vector, where element k is the weighted average misclassification cost of classifying an observation into class k, weighted by the class posterior probabilities.

  $$c_k = \sum_{j=1}^{K} \widehat{P}(Y = j | x_1, ..., x_P) Cost_{jk}.$$

  In other words, the software classifies observations to the class corresponding with the lowest expected misclassification cost.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For naive Bayes, the posterior probability that a classification is k for a given observation $(x_1,...,x_P)$ is

$$\widehat{P}(Y = k \,|\, x_1, .., x_P) = \frac{P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)}{P(X_1, ..., X_P)},$$

where:

- $P(X_1, ..., X_P \,|\, y = k)$ is the conditional joint density of the predictors given they are in class k. `Mdl.DistributionNames` stores the distribution names of the predictors.
- $\pi(Y = k)$ is the class prior probability distribution. `Mdl.Prior` stores the prior distribution.
- $P(X_1, .., X_P)$ is the joint density of the predictors. The classes are discrete, so $P(X_1, ..., X_P) = \sum_{k=1}^{K} P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)$.

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

See Also

`ClassificationNaiveBayes` | `CompactClassificationNaiveBayes` | `fitcnb` | `predict` | `resubLoss`

Topics

“Naive Bayes Classification” on page 21-2

Introduced in R2014b

loss

Package:

Classification loss for neural network classifier

Syntax

```
L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss(Mdl,X,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(Mdl,Tbl,ResponseVarName)` returns the classification loss on page 33-3694 for the trained neural network classifier `Mdl` using the predictor data in table `Tbl` and the class labels in the `ResponseVarName` table variable.

`L` is returned as a scalar value that represents the classification error by default.

`L = loss(Mdl,Tbl,Y)` returns the classification loss for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

`L = loss(Mdl,X,Y)` returns the classification loss for the trained neural network classifier `Mdl` using the predictor data `X` and the corresponding class labels in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that columns in the predictor data correspond to observations, specify the loss function, or supply observation weights.

Examples

Test Set Classification Error of Neural Network

Calculate the test set classification error of a neural network classifier.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(tbl.Smoker,"Holdout",0.30);
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a neural network classifier using the training set. Specify the `Smoker` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitcnet(tblTrain,"Smoker", ...
    "Standardize",true);
```

Calculate the test set classification error. Classification error is the default loss type for neural network classifiers.

```
testError = loss(Mdl,tblTest,"Smoker")
```

```
testError = 0.0671
```

```
testAccuracy = 1 - testError
```

```
testAccuracy = 0.9329
```

The neural network model correctly classifies approximately 93% of the test set observations.

Select Features to Include in Neural Network Classifier

Perform feature selection by comparing test set classification margins, edges, errors, and predictions. Compare the test set metrics for a model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(fishertable.Species,"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

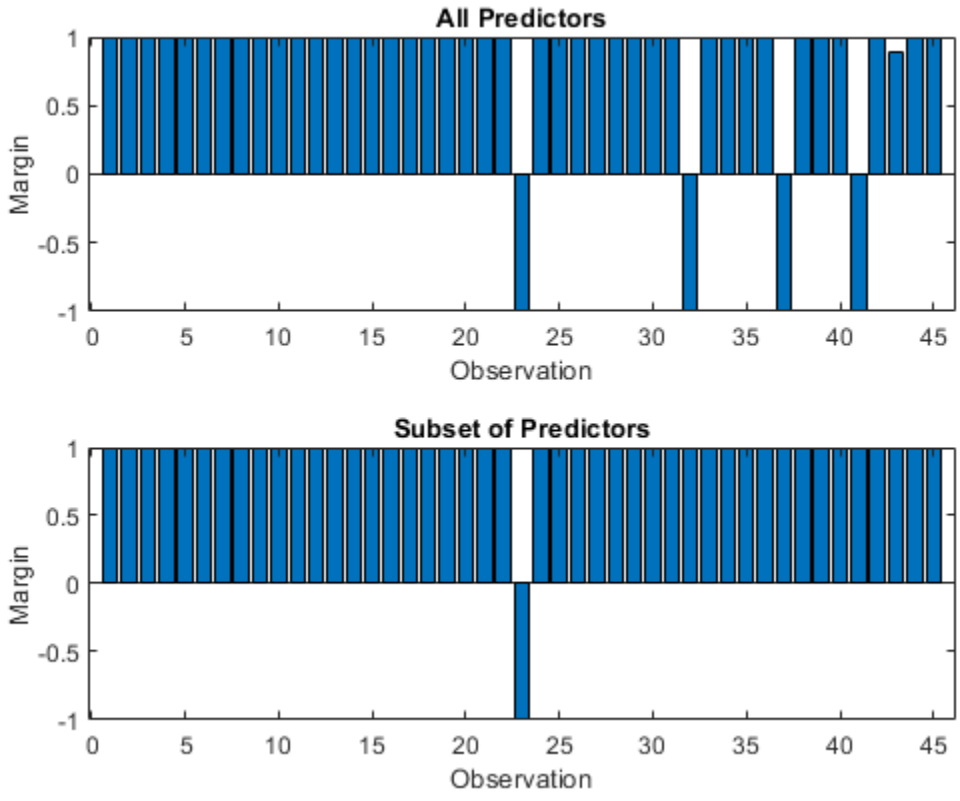
Train one neural network classifier using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `Species` as the response variable, and standardize the predictors.

```
allMdl = fitcnet(trainTbl,"Species","Standardize",true);
subsetMdl = fitcnet(trainTbl,"Species ~ SepalLength + SepalWidth + PetalLength", ...
    "Standardize",true);
```

Calculate the test set classification margins for the two models. Because the test set includes only 45 observations, display the margins using bar graphs.

For each observation, the classification margin is the difference between the classification score for the true class and the maximal score for the false classes. Because neural network classifiers return classification scores that are posterior probabilities, margin values close to 1 indicate confident classifications and negative margin values indicate misclassifications.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allMargins = margin(allMdl,testTbl);
bar(ax1,allMargins)
xlabel(ax1,"Observation")
ylabel(ax1,"Margin")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
subsetMargins = margin(subsetMdl,testTbl);
bar(ax2,subsetMargins)
xlabel(ax2,"Observation")
ylabel(ax2,"Margin")
title(ax2,"Subset of Predictors")
```



Compare the test set classification edge, or mean of the classification margins, of the two models.

```
allEdge = edge(allMdl,testTbl)
```

```
allEdge = 0.8198
```

```
subsetEdge = edge(subsetMdl,testTbl)
```

```
subsetEdge = 0.9556
```

Based on the test set classification margins and edges, the model trained on a subset of the predictors seems to outperform the model trained on all the predictors.

Compare the test set classification error of the two models.

```
allError = loss(allMdl,testTbl);
allAccuracy = 1-allError
```

```
allAccuracy = 0.9111
```

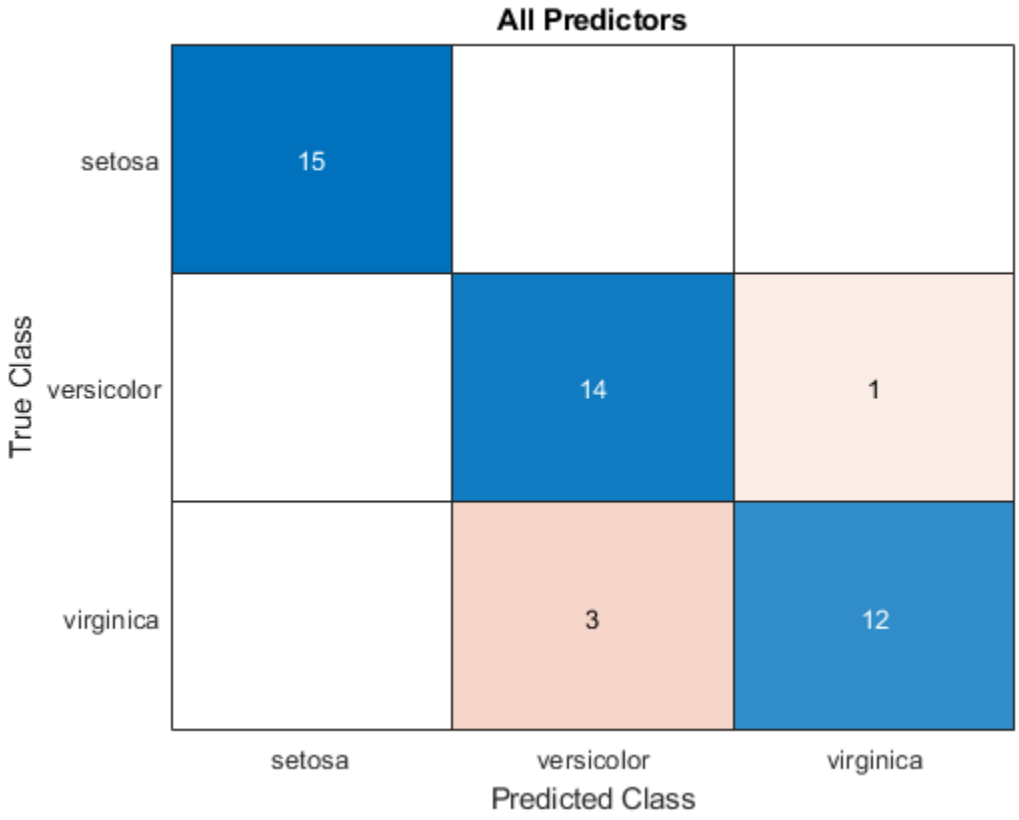
```
subsetError = loss(subsetMdl,testTbl);
subsetAccuracy = 1-subsetError
```

```
subsetAccuracy = 0.9778
```

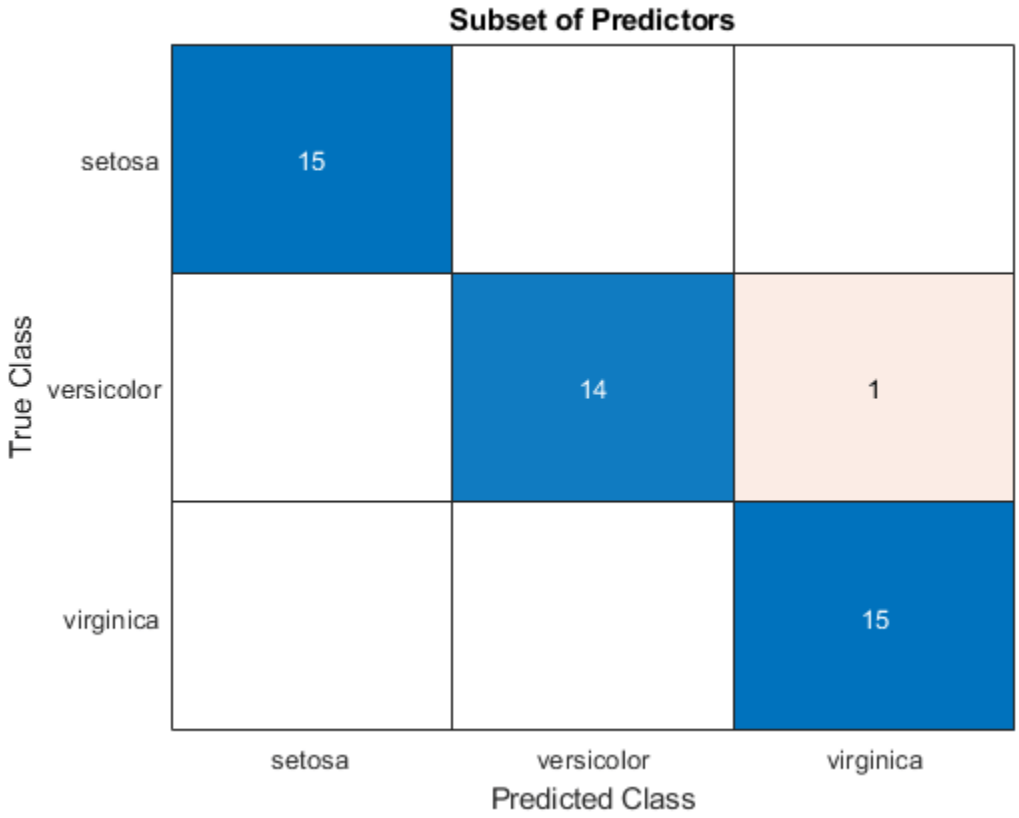
Again, the model trained using only a subset of the predictors seems to perform better than the model trained using all the predictors.

Visualize the test set classification results using confusion matrices.

```
allLabels = predict(allMdl,testTbl);
figure
confusionchart(testTbl.Species,allLabels)
title("All Predictors")
```



```
subsetLabels = predict(subsetMdl,testTbl);
figure
confusionchart(testTbl.Species,subsetLabels)
title("Subset of Predictors")
```



The model trained using all the predictors misclassifies four of the test set observations. The model trained using a subset of the predictors misclassifies only one of the test set observations.

Given the test set performance of the two models, consider using the model trained using all the predictors except `PetalWidth`.

Input Arguments

`Mdl` — Trained neural network classifier

`ClassificationNeuralNetwork` model object | `CompactClassificationNeuralNetwork` model object

Trained neural network classifier, specified as a `ClassificationNeuralNetwork` model object or `CompactClassificationNeuralNetwork` model object returned by `fitcnet` or `compact`, respectively.

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain an additional column for the

response variable. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.
- If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.
- If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. By default, `loss` assumes that each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(Mdl,Tbl,"Response","LossFun","crossentropy")` specifies to compute the cross-entropy loss for the model `Mdl`.

LossFun — Loss function

`'classiferror'` (default) | `'binodeviance'` | `'crossentropy'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'crossentropy'`	Cross-entropy loss (for neural networks only)
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  For more details on loss functions, see “Classification Loss” on page 33-3694.

- To specify a custom loss function, use function handle notation. The function must have this form:

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. `n` is the number of observations in `Tbl` or `X`, and `K` is the number of distinct classes (`numel(Mdl.ClassNames)`. The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(p,q) = 1`, if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.

- `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
- `W` is an `n`-by-1 numeric vector of observation weights.
- `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

Example: `'LossFun','crossentropy'`

Data Types: `char` | `string` | `function_handle`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights
nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

If you supply weights, then `loss` computes the weighted classification loss and normalizes weights to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:

- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

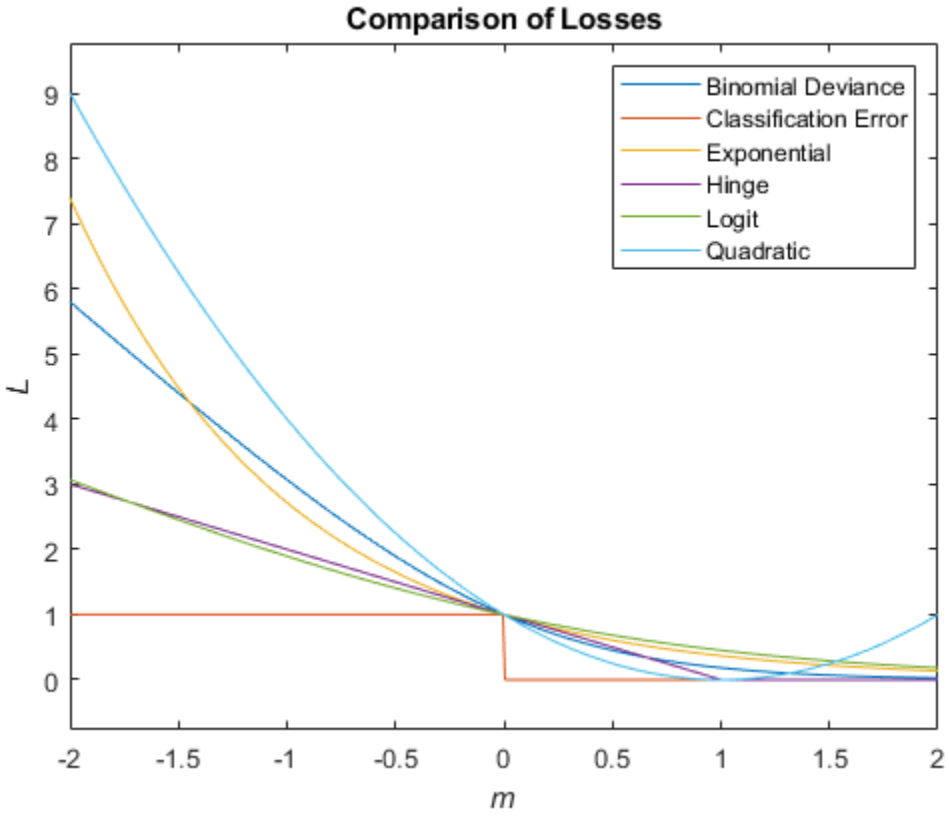
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.

Loss Function	Value of LossFun	Equation
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = (f(X_j)'C)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\hat{y}_j = \underset{k=1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`ClassificationNeuralNetwork` | `CompactClassificationNeuralNetwork` | `edge` | `fitcnet` | `margin` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

loss

Package: `classreg.learning.classif`

Find classification error for support vector machine (SVM) classifier

Syntax

```
L = loss(SVMModel,TBL,ResponseVarName)
L = loss(SVMModel,TBL,Y)

L = loss(SVMModel,X,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(SVMModel,TBL,ResponseVarName)` returns the classification error (see "Classification Loss" on page 33-3703), a scalar representing how well the trained support vector machine (SVM) classifier (`SVMModel`) classifies the predictor data in table `TBL` compared to the true class labels in `TBL.ResponseVarName`.

`loss` normalizes the class probabilities in `TBL.ResponseVarName` to the prior class probabilities that `fitcsvm` used for training, stored in the `Prior` property of `SVMModel`.

The classification loss (`L`) is a generalization or resubstitution quality measure. Its interpretation depends on the loss function and weighting scheme, but, in general, better classifiers yield smaller classification loss values.

`L = loss(SVMModel,TBL,Y)` returns the classification error for the predictor data in table `TBL` and the true class labels in `Y`.

`loss` normalizes the class probabilities in `Y` to the prior class probabilities that `fitcsvm` used for training, stored in the `Prior` property of `SVMModel`.

`L = loss(SVMModel,X,Y)` returns the classification error based on the predictor data in matrix `X` compared to the true class labels in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input arguments in previous syntaxes. For example, you can specify the loss function and the classification weights.

Examples

Determine Test Sample Classification Error of SVM Classifiers

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify a 15% holdout sample for testing, standardize the data, and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Holdout',0.15,'ClassNames',{'b','g'},...
    'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1}; % Extract the trained, compact classifier
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`CVSVMModel` is a `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Determine how well the algorithm generalizes by estimating the test sample classification error.

```
L = loss(CompactSVMModel,XTest,YTest)
```

```
L = 0.0787
```

The SVM classifier misclassifies approximately 8% of the test sample.

Determine Test Sample Hinge Loss of SVM Classifiers

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify a 15% holdout sample for testing, standardize the data, and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Holdout',0.15,'ClassNames',{'b','g'},...
    'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1}; % Extract the trained, compact classifier
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`CVSVMModel` is a `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Determine how well the algorithm generalizes by estimating the test sample hinge loss.

```
L = loss(CompactSVMModel,XTest,YTest,'LossFun','hinge')
```

```
L = 0.2998
```

The hinge loss is approximately 0.3. Classifiers with hinge losses close to 0 are preferred.

Input Arguments

SVMModel — SVM classification model

ClassificationSVM model object | CompactClassificationSVM model object

SVM classification model, specified as a ClassificationSVM model object or CompactClassificationSVM model object returned by fitcsvm or compact, respectively.

TBL — Sample data

table

Sample data, specified as a table. Each row of TBL corresponds to one observation, and each column corresponds to one predictor variable. Optionally, TBL can contain additional columns for the response variable and observation weights. TBL must contain all of the predictors used to train SVMModel. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If TBL contains the response variable used to train SVMModel, then you do not need to specify ResponseVarName or Y.

If you trained SVMModel using sample data contained in a table, then the input data for loss must also be in a table.

If you set 'Standardize',true in fitcsvm when training SVMModel, then the software standardizes the columns of the predictor data using the corresponding means in SVMModel.Mu and the standard deviations in SVMModel.Sigma.

Data Types: table

ResponseVarName — Response variable name

name of variable in TBL

Response variable name, specified as the name of a variable in TBL.

You must specify ResponseVarName as a character vector or string scalar. For example, if the response variable Y is stored as TBL.Y, then specify ResponseVarName as 'Y'. Otherwise, the software treats all columns of TBL, including Y, as predictors when training the model.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: char | string

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of X corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature). The variables in the columns of X must be the same as the variables that trained the SVMModel classifier.

The length of Y and the number of rows in X must be equal.

If you set `'Standardize',true` in `fitcsvm` to train `SVMModel`, then the software standardizes the columns of `X` using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `double` | `single`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. `Y` must be the same as the data type of `SVMModel.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The length of `Y` must equal the number of rows in `TBL` or the number of rows in `X`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(SVMModel,TBL,Y,'Weights',W)` weighs the observations in each row of `TBL` using the corresponding weight in each row of the variable `W` in `TBL`.

LossFun — Loss function

`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. You can specify to use posterior probabilities as classification scores for SVM models by setting `'FitPosterior',true` when you cross-validate the model using `fitcsvm`.

- Specify your own function by using function handle notation.

Suppose that `n` is the number of observations in `X`, and `K` is the number of distinct classes (`numel(SVMModel.ClassNames)`) used to create the input model (`SVMModel`). Your function must have this signature

```
lossvalue = lossfun(C,S,W,Cost)
```

where:

- The output argument `lossvalue` is a scalar.
- You choose the function name (*`lossfun`*).
- `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `SVMModel.ClassNames`.

  Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
- `S` is an `n`-by-`K` numeric matrix of classification scores, similar to the output of `predict`. The column order corresponds to the class order in `SVMModel.ClassNames`.
- `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes the weights to sum to `1`.
- `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see “Classification Loss” on page 33-3703.

Example: `'LossFun','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | numeric vector | name of variable in `TBL`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `TBL`. The software weighs the observations in each row of `X` or `TBL` with the corresponding weight in `Weights`.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `TBL`.

If you specify `Weights` as the name of a variable in `TBL`, you must do so as a character vector or string scalar. For example, if the weights are stored as `TBL.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `TBL`, including `TBL.W`, as predictors.

If you do not specify your own loss function, then the software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

Example: `'Weights','W'`

Data Types: `single` | `double` | `char` | `string`

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

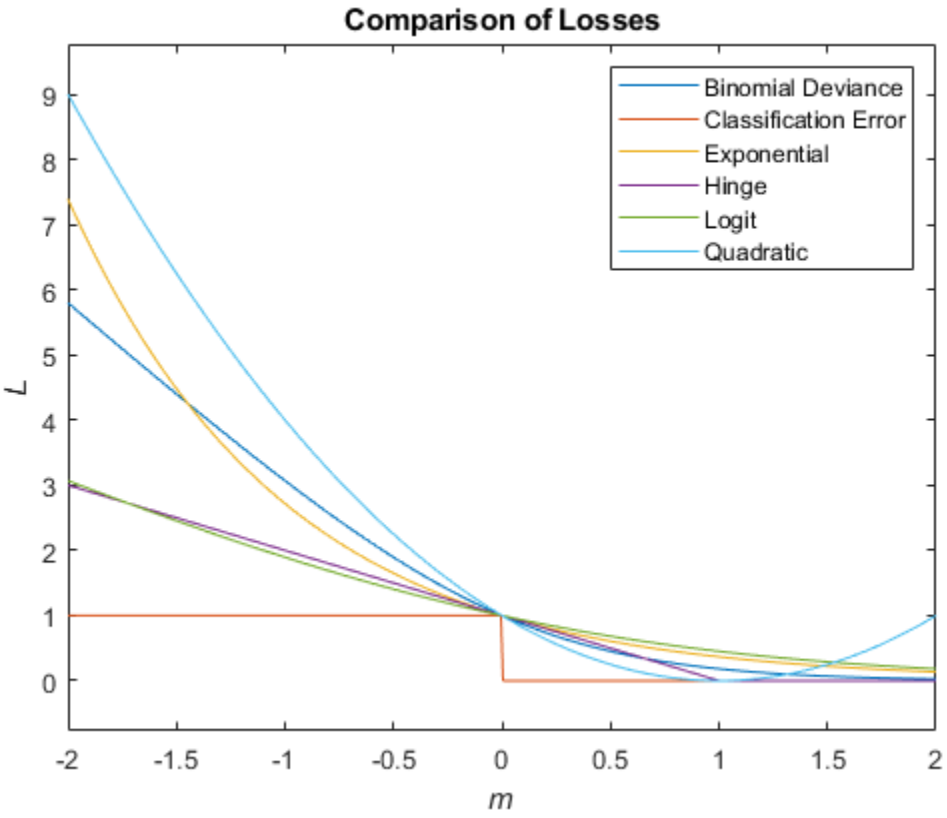
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of `LossFun`	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$

Loss Function	Value of LossFun	Equation
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Classification Score

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

The positive class classification score $f(x)$ is the trained SVM classification function. $f(x)$ is also the numerical predicted response for x, or the score for predicting x into the positive class.

$$f(x) = \sum_{j=1}^{n} \alpha_j y_j G(x_j, x) + b,$$

where $(\alpha_1, ..., \alpha_n, b)$ are the estimated SVM parameters, $G(x_j, x)$ is the dot product in the predictor space between x and the support vectors, and the sum includes the training set observations. The negative class classification score for x, or the score for predicting x into the negative class, is $-f(x)$.

If $G(x_j,x) = x_j'x$ (the linear kernel), then the score function reduces to

$$f(x) = (x/s)'\beta + b\,.$$

s is the kernel scale and β is the vector of fitted linear coefficients.

For more details, see “Understanding Support Vector Machines” on page 18-150.

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, second edition. Springer, New York, 2008.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationSVM` | `CompactClassificationSVM` | `fitcsvm` | `predict`

Introduced in R2014a

loss

Classification error

Syntax

```
L = loss(tree,TBL,ResponseVarName)
L = loss(tree,TBL,Y)
L = loss(tree,X,Y)

L = loss( ___ ,Name,Value)

[L,se,NLeaf,bestlevel] = loss( ___ )
```

Description

`L = loss(tree,TBL,ResponseVarName)` returns a scalar representing how well `tree` classifies the data in `TBL`, when `TBL.ResponseVarName` contains the true classifications.

When computing the loss, `loss` normalizes the class probabilities in `Y` to the class probabilities used for training, stored in the `Prior` property of `tree`.

`L = loss(tree,TBL,Y)` returns a scalar representing how well `tree` classifies the data in `TBL`, when `Y` contains the true classifications.

`L = loss(tree,X,Y)` returns a scalar representing how well `tree` classifies the data in `X`, when `Y` contains the true classifications.

`L = loss( ___ ,Name,Value)` returns the loss with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can specify the loss function or observation weights.

`[L,se,NLeaf,bestlevel] = loss( ___ )` also returns the vector of standard errors of the classification errors (`se`), the vector of numbers of leaf nodes in the trees of the pruning sequence (`NLeaf`), and the best pruning level as defined in the `TreeSize` name-value pair (`bestlevel`).

Input Arguments

`tree` — Trained classification tree

`ClassificationTree` model object | `CompactClassificationTree` model object

Trained classification tree, specified as a `ClassificationTree` or `CompactClassificationTree` model object. That is, `tree` is a trained classification model returned by `fitctree` or `compact`.

`TBL` — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all the predictors used to train `tree`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If TBL contains the response variable used to train tree, then you do not need to specify ResponseVarName or Y.

If you train tree using sample data contained in a table, then the input data for this method must also be in a table.

Data Types: table

X — Data to classify
numeric matrix

Data to classify, specified as a numeric matrix. Each row of X represents one observation, and each column represents one predictor. X must have the same number of columns as the data used to train tree. X must have the same number of rows as the number of elements in Y.

Data Types: single | double

ResponseVarName — Response variable name
name of a variable in TBL

Response variable name, specified as the name of a variable in TBL. If TBL contains the response variable used to train tree, then you do not need to specify ResponseVarName.

If you specify ResponseVarName, then you must do so as a character vector or string scalar. For example, if the response variable is stored as TBL.Response, then specify it as 'Response'. Otherwise, the software treats all columns of TBL, including TBL.ResponseVarName, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: char | string

Y — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Y must be of the same type as the classification used to train tree, and its number of elements must equal the number of rows of X.

Data Types: categorical | char | string | logical | single | double | cell

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

LossFun — Loss function
'mincost' (default) | 'binodeviance' | 'classiferror' | 'exponential' | 'hinge' | 'logit' | 'quadratic' | function handle

Loss function, specified as the comma-separated pair consisting of 'LossFun' and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. Classification trees return posterior probabilities as classification scores by default (see `predict`).

- Specify your own function using function handle notation.

  Suppose that `n` be the number of observations in `X` and `K` be the number of distinct classes (`numel(tree.ClassNames)`). Your function must have this signature

  `lossvalue = lossfun(C,S,W,Cost)`

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `tree.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `tree.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.
  - `Cost` is a *K*-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see “Classification Loss” on page 33-3715.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | name of a variable in `TBL` | numeric vector of positive values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values or the name of a variable in `TBL`.

If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `TBL`.

If you specify `Weights` as the name of a variable in `TBL`, you must do so as a character vector or string scalar. For example, if the weights are stored as `TBL.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `TBL`, including `TBL.W`, as predictors.

`loss` normalizes the weights so that observation weights in each class sum to the prior probability of that class. When you supply `Weights`, `loss` computes weighted classification loss.

Data Types: `single` | `double` | `char` | `string`

`Name,Value` arguments associated with pruning subtrees:

`Subtrees` — Pruning level
0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `loss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`loss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as the comma-separated pair consisting of `'TreeSize'` and one of the following values:

- `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L+se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Output Arguments

`L` — Classification loss
vector of scalar values

Classification loss on page 33-3715, returned as a vector the length of `Subtrees`. The meaning of the error depends on the values in `Weights` and `LossFun`.

se — Standard error of loss
vector of scalar values

Standard error of loss, returned as a vector the length of `Subtrees`.

NLeaf — Number of leaf nodes
vector of integer values

Number of leaves (terminal nodes) in the pruned subtrees, returned as a vector the length of `Subtrees`.

bestlevel — Best pruning level
scalar value

Best pruning level as defined in the `TreeSize` name-value pair, returned as a scalar whose value depends on `TreeSize`:

- `TreeSize` = `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L+se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `TreeSize` = `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

By default, `bestlevel` is the pruning level that gives loss within one standard deviation of minimal loss.

Examples

Compute the In-sample Classification Error

Compute the resubstituted classification error for the `ionosphere` data set.

```
load ionosphere
tree = fitctree(X,Y);
L = loss(tree,X,Y)
```

```
L = 0.0114
```

Examine the Classification Error for Each Subtree

Unpruned decision trees tend to overfit. One way to balance model complexity and out-of-sample performance is to prune a tree (or restrict its growth) so that in-sample and out-of-sample performance are satisfactory.

Load Fisher's iris data set. Partition the data into training (50%) and validation (50%) sets.

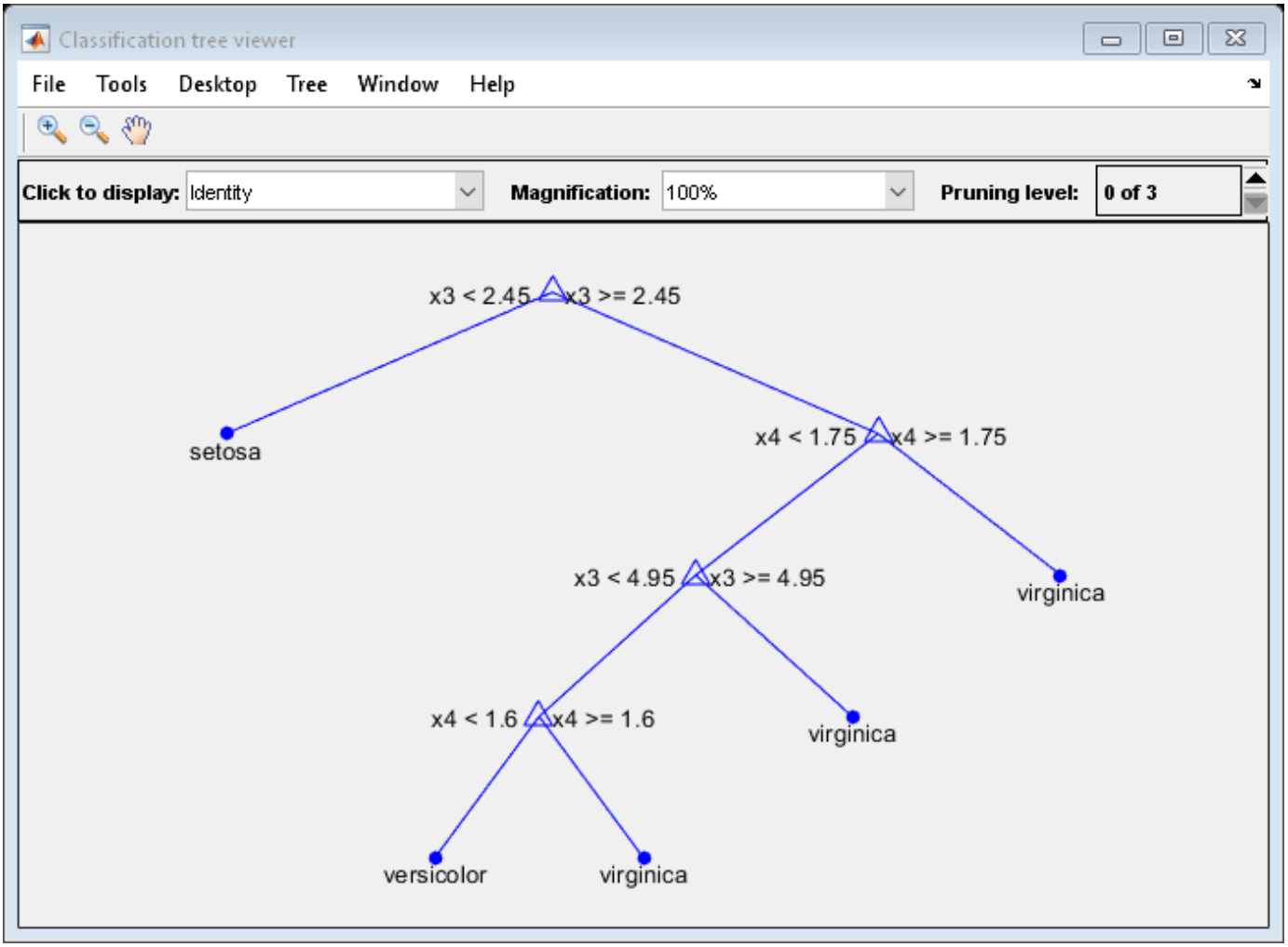
```
load fisheriris
n = size(meas,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a classification tree using the training set.

```
Mdl = fitctree(meas(idxTrn,:),species(idxTrn));
```

View the classification tree.

```
view(Mdl,'Mode','graph');
```



The classification tree has four pruning levels. Level 0 is the full, unpruned tree (as displayed). Level 3 is just the root node (i.e., no splits).

Examine the training sample classification error for each subtree (or pruning level) excluding the highest level.

```
m = max(Mdl.PruneList) - 1;
trnLoss = resubLoss(Mdl,'SubTrees',0:m)
```

```
trnLoss = 3×1

    0.0267
    0.0533
```

```
    0.3067
```

- The full, unpruned tree misclassifies about 2.7% of the training observations.
- The tree pruned to level 1 misclassifies about 5.3% of the training observations.
- The tree pruned to level 2 (i.e., a stump) misclassifies about 30.6% of the training observations.

Examine the validation sample classification error at each level excluding the highest level.

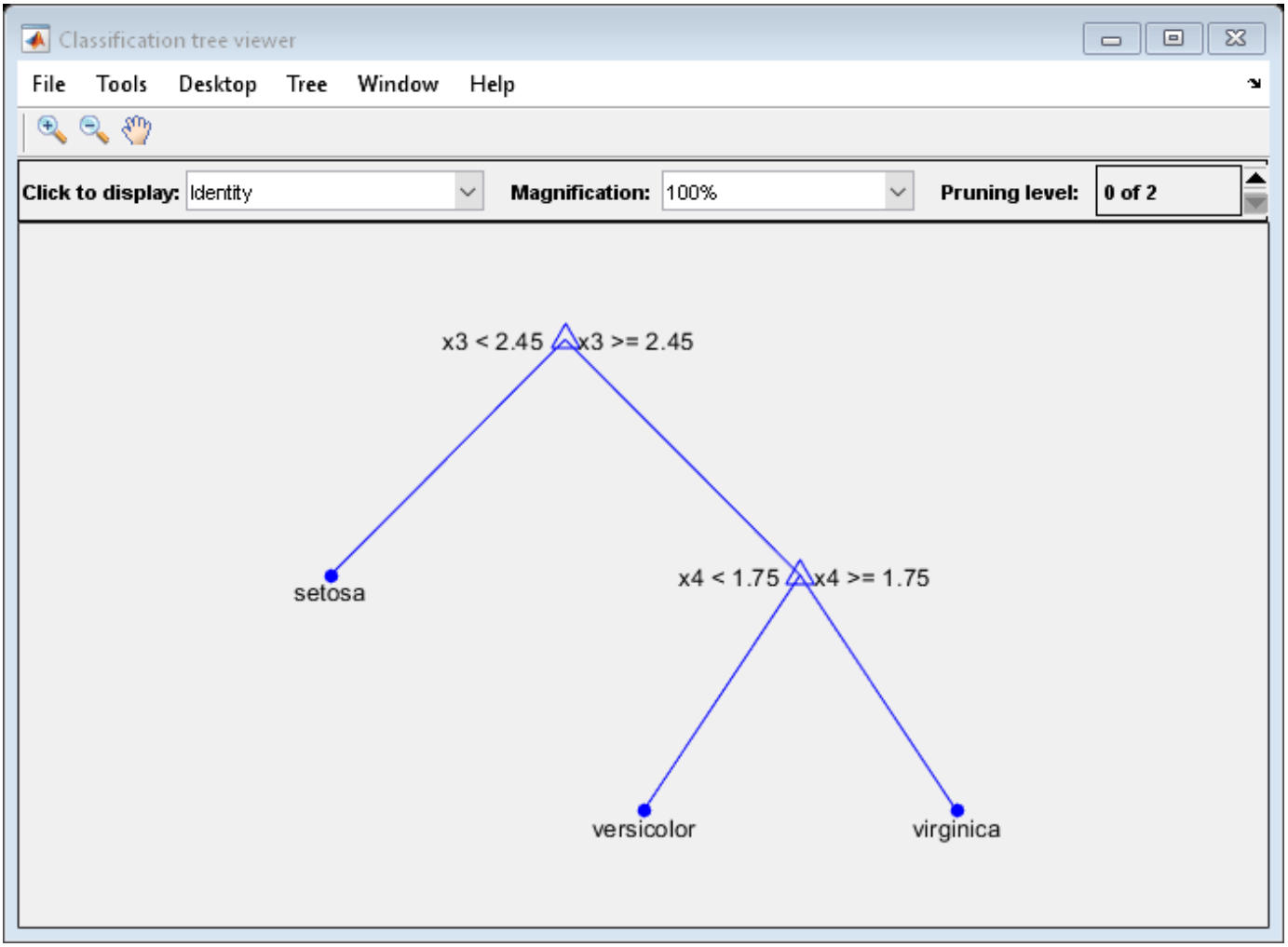
```
valLoss = loss(Mdl,meas(idxVal,:),species(idxVal),'SubTrees',0:m)
```

```
valLoss = 3×1

    0.0369
    0.0237
    0.3067
```

- The full, unpruned tree misclassifies about 3.7% of the validation observations.
- The tree pruned to level 1 misclassifies about 2.4% of the validation observations.
- The tree pruned to level 2 (i.e., a stump) misclassifies about 30.7% of the validation observations.

To balance model complexity and out-of-sample performance, consider pruning `Mdl` to level 1.

```
pruneMdl = prune(Mdl,'Level',1);
view(pruneMdl,'Mode','graph')
```



More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the

average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

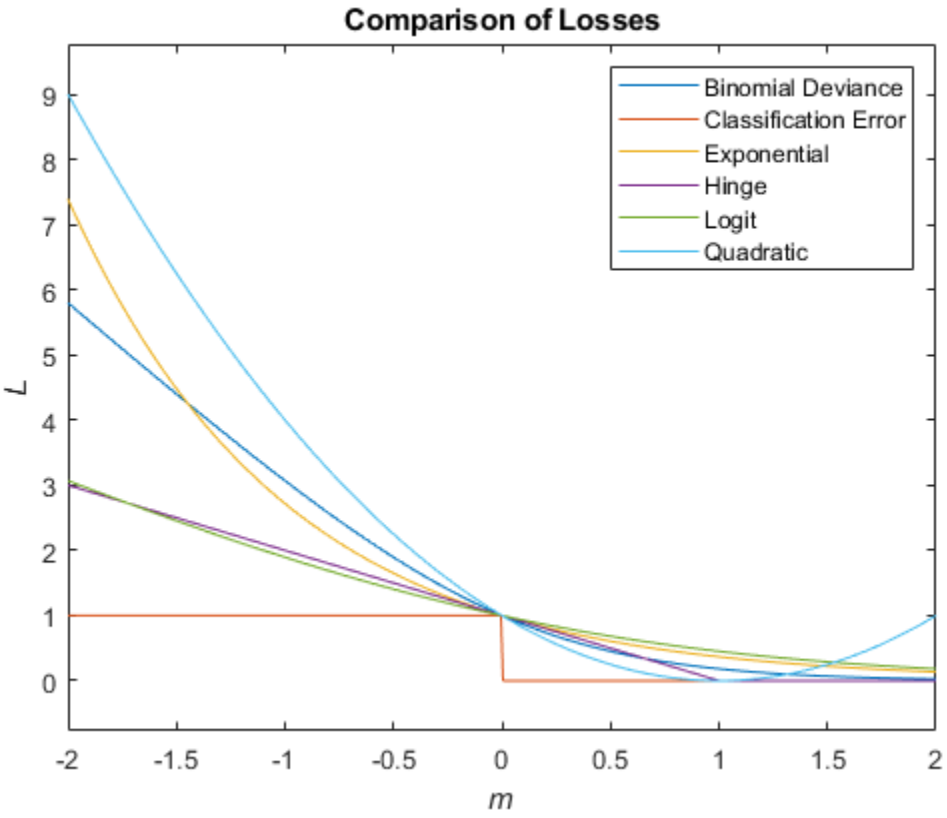
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp\left(-m_j\right).$

Loss Function	Value of LossFun	Equation
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = (f(X_j)'C)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\hat{y}_j = \underset{k=1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



True Misclassification Cost

The true misclassification cost is the cost of classifying an observation into an incorrect class.

You can set the true misclassification cost per class by using the `'Cost'` name-value argument when you create the classifier. `Cost(i,j)` is the cost of classifying an observation into class `j` when its true class is `i`. By default, `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Expected Misclassification Cost

The expected misclassification cost per observation is an averaged cost of classifying the observation into each class.

Suppose you have `Nobs` observations that you want to classify with a trained classifier, and you have `K` classes. You place the observations into a matrix `X` with one observation per row.

The expected cost matrix `CE` has size `Nobs`-by-`K`. Each row of `CE` contains the expected (average) cost of classifying the observation into each of the `K` classes. `CE(n,k)` is

$$\sum_{i=1}^{K} \widehat{P}(i|X(n))C(k|i),$$

where:

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability of class i for observation $X(n)$.
- $C(k|i)$ is the true misclassification cost of classifying an observation as k when its true class is i.

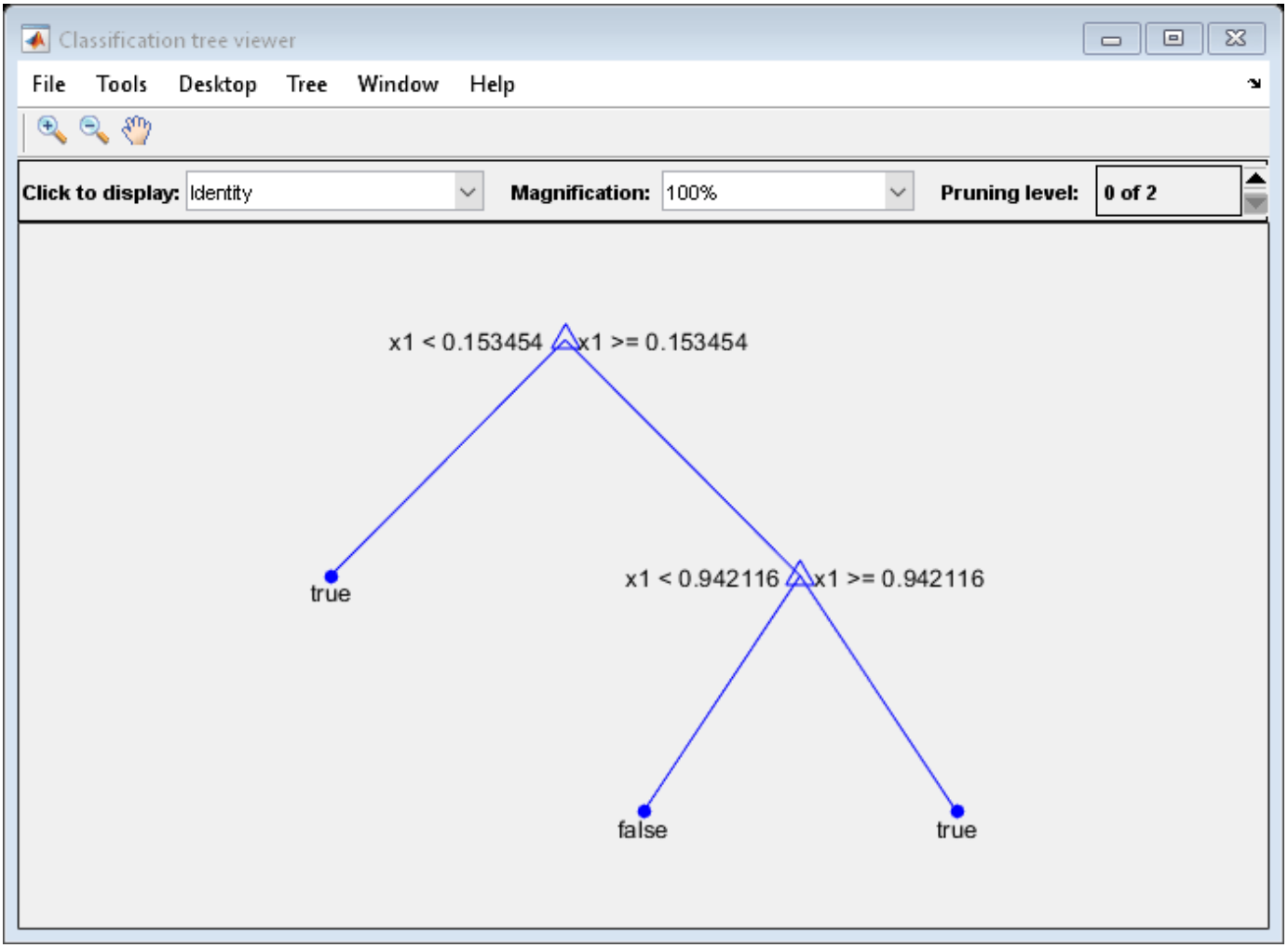
Score (tree)

For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor `X` as `true` when `X < 0.15` or `X > 0.95`, and `X` is false otherwise.

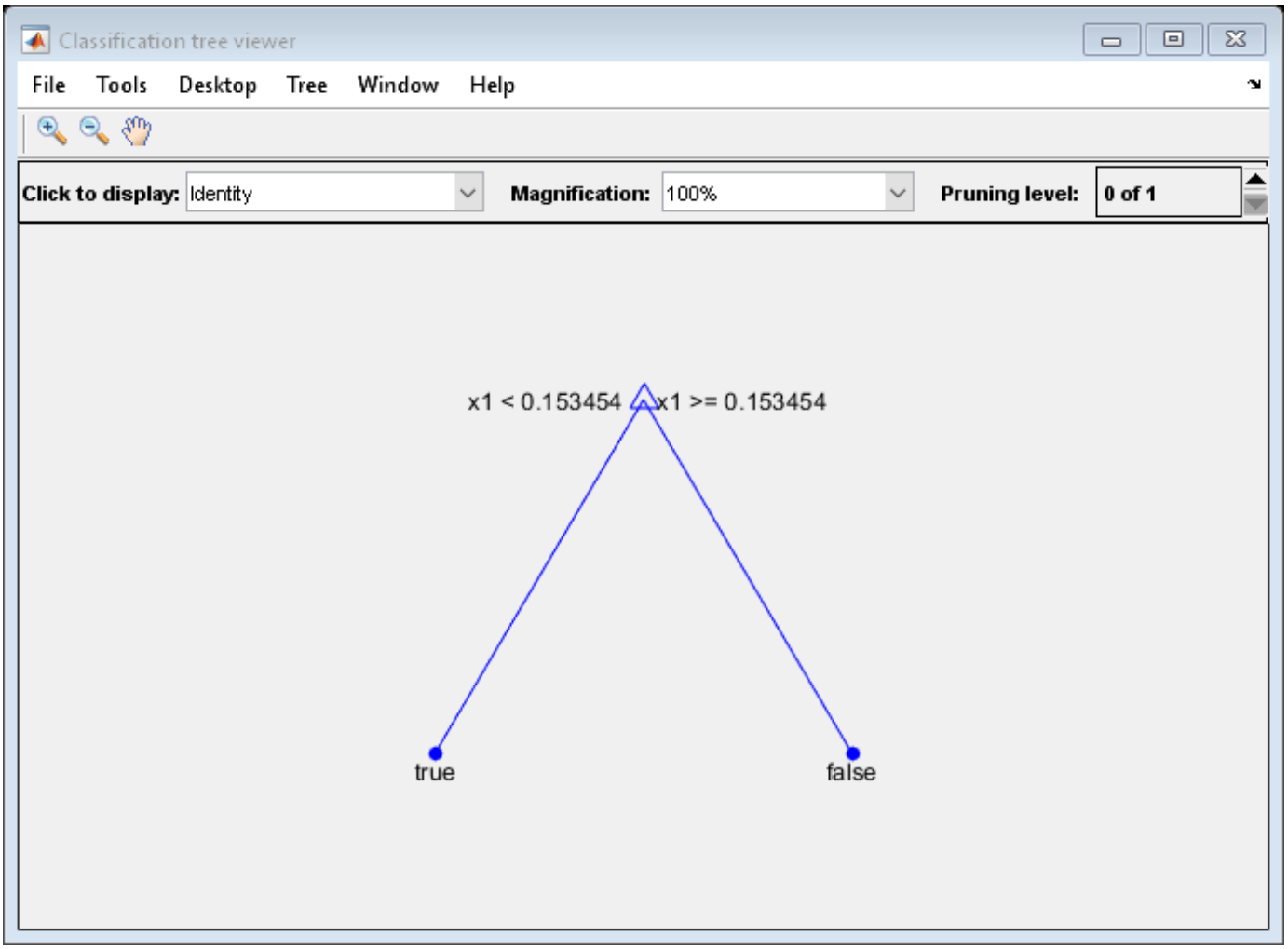
Generate 100 random points and classify them:

```
rng(0,'twister') % for reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
```

```
     0    1.0000    0.0975
0.9059    0.0941    0.2785
0.9059    0.0941    0.5469
0.9059    0.0941    0.9575
0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Only one output is supported.
- You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

See Also

`edge` | `fitctree` | `margin` | `predict`

loss

Regression error

Syntax

```
L = loss(ens,tbl,ResponseVarName)
L = loss(ens,tbl,Y)
L = loss(ens,X,Y)
L = loss( ___ ,Name,Value)
```

Description

`L = loss(ens,tbl,ResponseVarName)` returns the mean squared error between the predictions of `ens` to the data in `tbl`, compared to the true responses `tbl.ResponseVarName`.

`L = loss(ens,tbl,Y)` returns the mean squared error between the predictions of `ens` to the data in `tbl`, compared to the true responses `Y`.

`L = loss(ens,X,Y)` returns the mean squared error between the predictions of `ens` to the data in `X`, compared to the true responses `Y`.

`L = loss( ___ ,Name,Value)` computes the error in prediction with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`, or the `compact` method.

`tbl`

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all of the predictors used to train the model. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `ens` using sample data contained in a `table`, then the input data for this method must also be in a table.

`ResponseVarName`

Response variable name, specified as the name of a variable in `tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `tbl`, including `Y`, as predictors when training the model.

`X`

A matrix of predictor values. Each column of `X` represents one variable, and each row represents one observation.

`NaN` values in `X` are taken to be missing values. Observations with all missing values for `X` are not used in the calculation of loss.

If you trained `ens` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

`Y`

A numeric column vector with the same number of rows as `tbl` or `X`. Each entry in `Y` is the response to the data in the corresponding row of `tbl` or `X`.

`NaN` values in `Y` are taken to be missing values. Observations with missing values for `Y` are not used in the calculation of loss.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `oobEdge` uses only these learners for calculating loss.

Default: `1:NumTrained`

`lossfun`

Function handle for loss function, or `'mse'`, meaning mean squared error. If you pass a function handle `fun`, `loss` calls it as

```
fun(Y,Yfit,W)
```

where `Y`, `Yfit`, and `W` are numeric vectors of the same length.

- `Y` is the observed response.
- `Yfit` is the predicted response.
- `W` is the observation weights.

The returned value `fun(Y,Yfit,W)` should be a scalar.

Default: `'mse'`

`mode`

Meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.

- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

`UseObsForLearner`

A logical matrix of size `N`-by-`NumTrained`, where `N` is the number of observations in `ens.X`, and `NumTrained` is the number of weak learners. When `UseObsForLearner(I,J)` is `true`, `predict` uses learner `J` in predicting observation `I`.

Default: `true(N,NumTrained)`

`weights`

Numeric vector of observation weights with the same number of elements as `Y`. The formula for `loss` with `weights` is in “Weighted Mean Squared Error” on page 33-3724.

Default: `ones(size(Y))`

Output Arguments

`L`

Weighted mean squared error of predictions. The formula for `loss` is in “Weighted Mean Squared Error” on page 33-3724.

Examples

Find Mean-Squared Error of Ensemble Predictions

Find the loss of an ensemble predictor using the `carsmall` data set.

Load the `carsmall` data set and select engine displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train an ensemble of regression trees and find the regression error for predicting `MPG`.

```
ens = fitrensemble(X,MPG);
L = loss(ens,X,MPG)
```

```
L = 0.3463
```

More About

Weighted Mean Squared Error

Let n be the number of rows of data, x_j be the jth row of data, y_j be the true response to x_j, and let $f(x_j)$ be the response prediction of `ens` to x_j. Let w be the vector of weights (all one by default).

First the weights are divided by their sum so they add to one: $w \to w/\Sigma w$. The mean squared error L is

$$L = \sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j \right)^2 .$$

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`fitrensemble` | `predict`

loss

Regression error for Gaussian process regression model

Syntax

```
L = loss(gprMdl,Xnew,Ynew)
L = loss(gprMdl,Xnew,Ynew,Name,Value)
```

Description

`L = loss(gprMdl,Xnew,Ynew)` returns the mean squared error for the Gaussian process regression (GPR) model `gpr`, using the predictors in `Xnew` and observed response in `Ynew`.

`L = loss(gprMdl,Xnew,Ynew,Name,Value)` returns the mean squared error for the GPR model, `gpr`, with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify a custom loss function or the observation weights.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object | `CompactRegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` (full) or `CompactRegressionGP` (compact) object.

`Xnew` — New observed data

`table` | *m*-by-*d* matrix

New data, specified as a `table` or an *n*-by-*d* matrix, where *m* is the number of observations, and *d* is the number of predictor variables in the training data.

If you trained `gprMdl` on a `table`, then `Xnew` must be a `table` that contains all the predictor variables used to train `gprMdl`.

If `Xnew` is a table, then it can also contain `Ynew`. And if it does, then you do not have to specify `Ynew`.

If you trained `gprMdl` on a matrix, then `Xnew` must be a numeric matrix with *d* columns, and can only contain values for the predictor variables.

Data Types: `single` | `double` | `table`

`Ynew` — New response values

n-by-1 vector

New observed response values, that correspond to the predictor values in `Xnew`, specified as an *n*-by-1 vector. *n* is the number of rows in `Xnew`. Each entry in `Ynew` is the observed response based on the predictor data in the corresponding row of `Xnew`.

If `Xnew` is a table containing new response values, you do not have to specify `Ynew`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`lossfun` — Loss function
`'mse'` (default) | function handle

Loss function, specified as `'mse'` (mean squared error) or a function handle.

If you pass a function handle, say `fun`, `loss` calls it as shown below: `fun(Y,Ypred,W)`, where `Y`, `Ypred` and `W` are numeric vectors of length *n*, and *n* is the number of rows in `Xnew`. `Y` is the observed response, `Ypred` is the predicted response, and `W` is the observation weights.

Example: `'lossfun',Fct` calls the loss function `Fct`.

Data Types: `char` | `string` | `function_handle`

`weights` — Observation weights
vector of 1s (default) | *n*-by-1 vector

Observation weights, specified as *n*-by-1 vector, where *n* is the number of rows in `Xnew`. By default, the weight of each observation is 1.

Example: `'weights',W` uses the observation weights in vector `W`.

Data Types: `double` | `single`

Output Arguments

`L` — Regression error
scalar value

Regression error for the trained Gaussian process regression model, `gprMdl`, returned as a scalar value.

Examples

Compute Regression Loss for Test Data

Load the sample data.

```
load('gprdata.mat')
```

The data has 8 predictor variables and contains 500 observations in training data and 100 observations in test data. This is simulated data.

Fit a GPR model using the squared exponential kernel function with separate length scales for each predictor. Standardize the predictor values in the training data. Use the exact method for fitting and prediction.

```
gprMdl = fitrgp(Xtrain,ytrain,'FitMethod','exact',...
'PredictMethod','exact','KernelFunction','ardsquaredexponential',...
'Standardize',1);
```

Compute the regression error for the test data.

```
L = loss(gprMdl,Xtest,ytest)
```

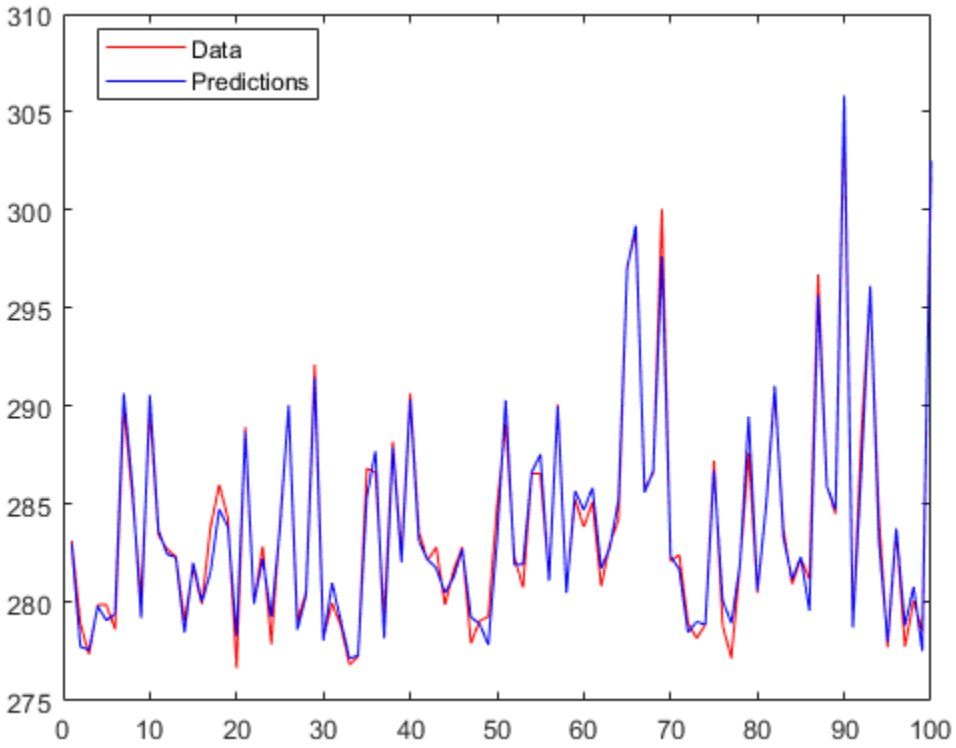
```
L = 0.6928
```

Predict the responses for test data.

```
ypredtest = predict(gprMdl,Xtest);
```

Plot the test response along with the predictions.

```
figure;
plot(ytest,'r');
hold on;
plot(ypredtest,'b');
legend('Data','Predictions','Location','Best');
```



Manually compute the regression loss.

```
L = (ytest - ypredtest)'*(ytest - ypredtest)/length(ytest)
```

```
L = 0.6928
```

Specify Custom Loss Function

Load the sample data and store in a `table`.

```
load fisheriris
tbl = table(meas(:,1),meas(:,2),meas(:,3),meas(:,4),species,...
'VariableNames',{'meas1','meas2','meas3','meas4','species'});
```

Fit a GPR model using the first measurement as the response and the other variables as the predictors.

```
mdl = fitrgp(tbl,'meas1');
```

Predict the responses using the trained model.

```
ypred = predict(mdl,tbl);
```

Compute the mean absolute error.

```
n = height(tbl);
y = tbl.meas1;
fun = @(y,ypred,w) sum(abs(y-ypred))/n;
L = loss(mdl,tbl,'lossfun',fun)
```

```
L = 0.2345
```

Alternatives

You can use `resubLoss` to compute the regression error for the trained GPR model at the observations in the training data.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`CompactRegressionGP` | `RegressionGP` | `compact` | `fitrgp` | `predict` | `resubLoss`

Introduced in R2015b

loss

Package:

Loss for regression neural network

Syntax

```
L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss(Mdl,X,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(Mdl,Tbl,ResponseVarName)` returns the regression loss for the trained regression neural network `Mdl` using the predictor data in table `Tbl` and the response values in the `ResponseVarName` table variable.

`L` is returned as a scalar value that represents the mean squared error (MSE) by default.

`L = loss(Mdl,Tbl,Y)` returns the regression loss for the model `Mdl` using the predictor data in table `Tbl` and the response values in vector `Y`.

`L = loss(Mdl,X,Y)` returns the regression loss for the trained regression neural network `Mdl` using the predictor data `X` and the corresponding response values in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify that columns in the predictor data correspond to observations, specify the loss function, or supply observation weights.

Examples

Test Set Mean Squared Error of Neural Network

Calculate the test set mean squared error (MSE) of a regression neural network model.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Systolic` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Age,Diastolic,Gender,Height,Smoker,Weight,Systolic);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a nonstratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(size(tbl,1),"Holdout",0.30);
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a regression neural network model using the training set. Specify the `Systolic` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitrnet(tblTrain,"Systolic", ...
    "Standardize",true);
```

Calculate the test set MSE. Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,tblTest,"Systolic")
```

```
testMSE = 49.9595
```

Select Features to Include in Regression Neural Network

Perform feature selection by comparing test set losses and predictions. Compare the test set metrics for a regression neural network model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a nonstratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(size(fishertable,1),"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

Train one regression neural network model using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `PetalLength` as the response variable, and standardize the predictors.

```
allMdl = fitrnet(trainTbl,"PetalLength","Standardize",true);
subsetMdl = fitrnet(trainTbl,"PetalLength ~ SepalLength + SepalWidth + Species", ...
    "Standardize",true);
```

Compare the test set mean squared error (MSE) of the two models. Smaller MSE values indicate better performance.

```
allMSE = loss(allMdl,testTbl)
```

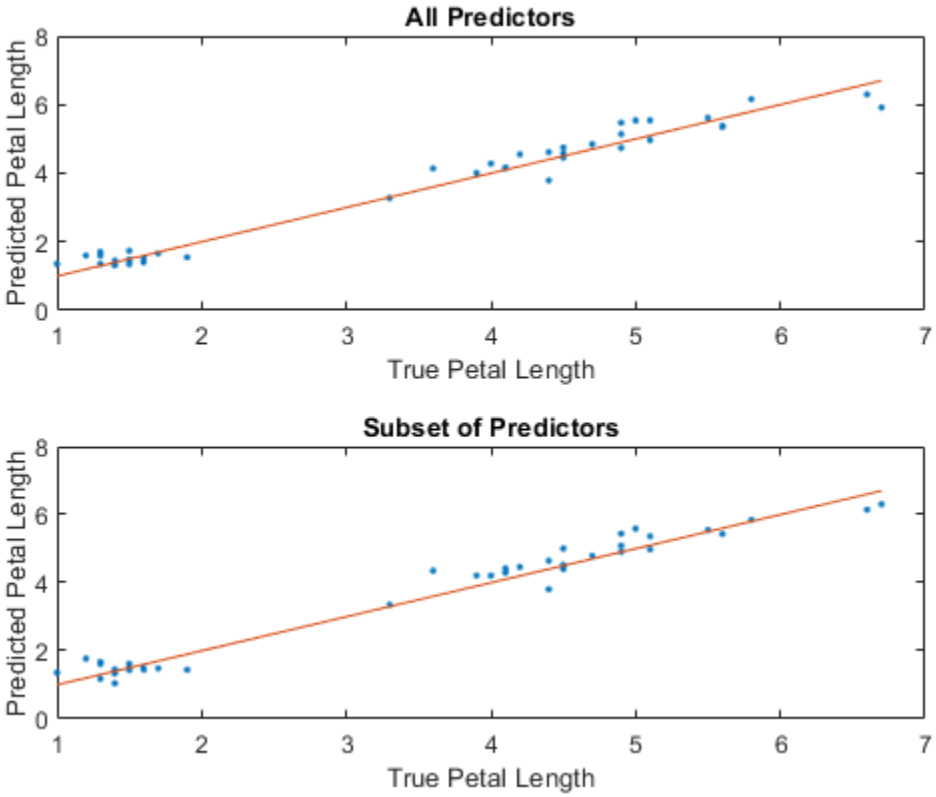
```
allMSE = 0.0834
```

```
subsetMSE = loss(subsetMdl,testTbl)
```

```
subsetMSE = 0.0887
```

For each model, compare the test set predicted petal lengths to the true petal lengths. Plot the predicted petal lengths along the vertical axis and the true petal lengths along the horizontal axis. Points on the reference line indicate correct predictions.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allPredictedY = predict(allMdl,testTbl);
plot(ax1,testTbl.PetalLength,allPredictedY,".")
hold on
plot(ax1,testTbl.PetalLength,testTbl.PetalLength)
hold off
xlabel(ax1,"True Petal Length")
ylabel(ax1,"Predicted Petal Length")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
subsetPredictedY = predict(subsetMdl,testTbl);
plot(ax2,testTbl.PetalLength,subsetPredictedY,".")
hold on
plot(ax2,testTbl.PetalLength,testTbl.PetalLength)
hold off
xlabel(ax2,"True Petal Length")
ylabel(ax2,"Predicted Petal Length")
title(ax2,"Subset of Predictors")
```



Because both models seems to perform well, with predictions scattered near the reference line, consider using the model trained using all predictors except `PetalWidth`.

Input Arguments

`Mdl` — Trained regression neural network

`RegressionNeuralNetwork` model object | `CompactRegressionNeuralNetwork` model object

Trained regression neural network, specified as a `RegressionNeuralNetwork` model object or `CompactRegressionNeuralNetwork` model object returned by `fitrnet` or `compact`, respectively.

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain an additional column for the response variable. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.
- If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

- If you set `'Standardize',true` in `fitrnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

Data Types: `char` | `string`

`Y` — Response data
numeric vector

Response data, specified as a numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. By default, `loss` assumes that each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

If you set `'Standardize',true` in `fitrnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `loss(Mdl,Tbl,"Response","Weights","W")` specifies to use the `Response` and `W` variables in the table `Tbl` as the response values and observation weights, respectively.

`LossFun` — Loss function
`'mse'` (default) | function handle

Loss function, specified as `'mse'` or a function handle.

- `'mse'` — Weighted mean squared error.
- Function handle — To specify a custom loss function, use a function handle. The function must have this form:

  ```
  lossval = lossfun(Y,YFit,W)
  ```

  - The output argument `lossval` is a floating-point scalar.
  - You specify the function name (*`lossfun`*).
  - `Y` is a length *n* numeric vector of observed responses, where *n* is the number of observations in `Tbl` or `X`.
  - `YFit` is a length *n* numeric vector of corresponding predicted responses.
  - `W` is an *n*-by-1 numeric vector of observation weights.

Example: `'LossFun',@lossfun`

Data Types: `char` | `string` | `function_handle`

`ObservationsIn` — Predictor data observation dimension

`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights

nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

If you supply weights, then `loss` computes the weighted regression loss and normalizes weights to sum to 1.

Data Types: `single` | `double` | `char` | `string`

See Also

`CompactRegressionNeuralNetwork` | `RegressionNeuralNetwork` | `fitrnet` | `predict`

Topics
“Assess Regression Neural Network Performance” on page 18-184

Introduced in R2021a

loss

Regression error for support vector machine regression model

Syntax

```
L = loss(mdl,tbl,ResponseVarName)
L = loss(mdl,tbl,Y)
L = loss(mdl,X,Y)
L = loss(___,Name,Value)
```

Description

`L = loss(mdl,tbl,ResponseVarName)` returns the loss for the predictions of the support vector machine (SVM) regression model, `mdl`, based on the predictor data in the table `tbl` and the true response values in `tbl.ResponseVarName`.

`L = loss(mdl,tbl,Y)` returns the loss for the predictions of the support vector machine (SVM) regression model, `mdl`, based on the predictor data in the table `X` and the true response values in the vector `Y`.

`L = loss(mdl,X,Y)` returns the loss for the predictions of the support vector machine (SVM) regression model, `mdl`, based on the predictor data in `X` and the true responses in `Y`.

`L = loss(___,Name,Value)` returns the loss with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can specify the loss function or observation weights.

Input Arguments

`mdl` — SVM regression model

`RegressionSVM` model | `CompactRegressionSVM` model

SVM regression model, specified as a `RegressionSVM` model or `CompactRegressionSVM` model returned by `fitrsvm` or `compact`, respectively.

`tbl` — Sample data

table

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain additional columns for the response variable and observation weights. `tbl` must contain all of the predictors used to train `mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `mdl` using sample data contained in a `table`, then the input data for this method must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name

name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

X — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix or table. Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature).

If you trained `mdl` using a matrix of predictor values, then `X` must be a numeric matrix with p columns. p is the number of predictors used to train `mdl`.

The length of `Y` and the number of rows of `X` must be equal.

Data Types: `single` | `double`

Y — Observed response values
vector of numeric values

Observed response values, specified as a vector of length n containing numeric values. Each entry in `Y` is the observed response based on the predictor data in the corresponding row of `X`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function
`'mse'` (default) | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'mse'`, `'epsiloninsensitive'`, or a function handle.

- The following table lists the available loss functions.

Value	Loss Function
`'mse'`	“Weighted Mean Squared Error” on page 33-3740
`'epsiloninsensitive'`	“Epsilon-Insensitive Loss Function” on page 33-3740

- Specify your own function using function handle notation.

  Your function must have the signature `lossvalue = lossfun(Y,Yfit,W)`, where:

- The output argument `lossvalue` is a scalar value.
- You choose the function name (*lossfun*).
- `Y` is an *n*-by-1 numeric vector of observed response values.
- `Yfit` is an *n*-by-1 numeric vector of predicted response values, calculated using the corresponding predictor values in `X` (similar to the output of `predict`).
- `W` is an *n*-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to 1.

Specify your function using `'LossFun',@`*`lossfun`*.

Example: `'LossFun','epsiloninsensitive'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector. `Weights` must be the same length as the number of rows in `X`. The software weighs the observations in each row of `X` using the corresponding weight value in `Weights`.

Weights are normalized to sum to 1.

Data Types: `single` | `double`

Output Arguments

`L` — Regression loss
scalar value

Regression loss, returned as a scalar value.

Examples

Calculate Test Sample Loss for SVM Regression Model

This example shows how to train an SVM regression model, then calculate the resubstitution mean square error and epsilon-insensitive error.

Load the `carsmall` sample data.

```
load carsmall
rng default  % for reproducibility
```

Specify `Horsepower` and `Weight` as the predictor variables (`X`), and `MPG` as the response variable (`Y`).

```
X = [Horsepower,Weight];
Y = MPG;
```

Train a linear SVM regression model. Standardize the data.

```
mdl = fitrsvm(X,Y,'Standardize',true);
```

`mdl` is a `RegressionSVM` model.

Determine how well the trained model generalizes to new predictor values by estimating the test sample mean square error and epsilon-insensitive error.

```
lossMSE = loss(mdl,X,Y)
lossEI = loss(mdl,X,Y,'LossFun','epsiloninsensitive')
```

```
lossMSE =

   17.0256


lossEI =

    2.2506
```

More About

Weighted Mean Squared Error

The weighted mean squared error is calculated as follows:

$$\text{mse} = \frac{\sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j\right)^2}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of rows of data
- x_j is the jth row of data
- y_j is the true response to x_j
- $f(x_j)$ is the response prediction of the SVM regression model `mdl` to x_j
- w is the vector of weights.

The weights in w are all equal to one by default. You can specify different values for weights using the `'Weights'` name-value pair argument. If you specify weights, each value is divided by the sum of all weights, such that the normalized weights add to one.

Epsilon-Insensitive Loss Function

The epsilon-insensitive loss function ignores errors that are within the distance epsilon (ε) of the function value. It is formally described as:

$$Loss_{\varepsilon} = \begin{cases} 0\,,\ if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon\,,\ otherwise\,. \end{cases}$$

The mean epsilon-insensitive loss is calculated as follows:

$$Loss = \frac{\sum_{j=1}^{n} w_j \max\left(0, \left|y_j - f\left(x_j\right)\right| - \varepsilon\right)}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of rows of data
- x_j is the jth row of data
- y_j is the true response to x_j
- $f(x_j)$ is the response prediction of the SVM regression model `mdl` to x_j
- w is the vector of weights.

The weights in w are all equal to one by default. You can specify different values for weights using the `'Weights'` name-value pair argument. If you specify weights, each value is divided by the sum of all weights, such that the normalized weights add to one.

Tips

- If `mdl` is a cross-validated `RegressionPartitionedSVM` model, use `kfoldLoss` instead of `loss` to calculate the regression error.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`CompactRegressionSVM` | `RegressionSVM` | `fitrsvm` | `kfoldLoss`

Introduced in R2015b

loss

Regression error

Syntax

```
L = loss(tree,tbl,ResponseVarName)
L = loss(tree,x,y)
L = loss( ___ ,Name,Value)
[L,se,NLeaf,bestlevel] = loss( ___ )
```

Description

`L = loss(tree,tbl,ResponseVarName)` returns the mean squared error between the predictions of `tree` to the data in `tbl`, compared to the true responses `tbl.ResponseVarName`.

`L = loss(tree,x,y)` returns the mean squared error between the predictions of `tree` to the data in `x`, compared to the true responses `y`.

`L = loss( ___ ,Name,Value)` computes the error in prediction with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

`[L,se,NLeaf,bestlevel] = loss( ___ )` also returns the standard error of the loss (`se`), the number of leaves (terminal nodes) in the tree (`NLeaf`), and the optimal pruning level for `tree` (`bestlevel`).

Input Arguments

`tree` — Trained regression tree
`RegressionTree` object | `CompactRegressionTree` object

Trained regression tree, specified as a `RegressionTree` object constructed by `fitrtree` or a `CompactRegressionTree` object constructed by `compact`.

`x` — Predictor values
matrix of floating-point values

Predictor values, specified as matrix of floating-point values. Each column of `x` represents one variable, and each row represents one observation.

Data Types: `single` | `double`

`ResponseVarName` — Response variable name
name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `y` is stored as `tbl.y`, then specify `ResponseVarName` as `'y'`. Otherwise, the software treats all columns of `tbl`, including `y`, as predictors when training the model.

Data Types: `char` | `string`

y — Response data

numeric column vector

Response data, specified as a numeric column vector with the same number of rows as `x`. Each entry in `y` is the response to the data in the corresponding row of `x`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

LossFun — Loss function

`'mse'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a function handle for loss, or `'mse'` representing mean-squared error. If you pass a function handle `fun`, `loss` calls `fun` as:

```
fun(Y,Yfit,W)
```

- `Y` is the vector of true responses.
- `Yfit` is the vector of predicted responses.
- `W` is the observation weights. If you pass `W`, the elements are normalized to sum to `1`.

All the vectors have the same number of rows as `Y`.

Example: `'LossFun','mse'`

Data Types: `function_handle` | `char` | `string`

Subtrees — Pruning level

0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `loss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`loss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as the comma-separated pair consisting of `'TreeSize'` and one of the following:

- `'se'` — `loss` returns `bestlevel` that corresponds to the smallest tree whose mean squared error (MSE) is within one standard error of the minimum MSE.
- `'min'` — `loss` returns `bestlevel` that corresponds to the minimal MSE tree.

Example: `'TreeSize','min'`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of a variable in `tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values. The software weights the observations in each row of `X` or `tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `tbl`.

If you specify the input data as a table `tbl`, then `Weights` can be the name of a variable in `tbl` that contains a numeric vector. In this case, you must specify `Weights` as a variable name. For example, if weights vector `W` is stored as `tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `tbl`, including `W`, as predictors when training the model.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`L` — Classification error
vector of scalar values

Classification error, returned as a vector the length of `Subtrees`. The error for each tree is the mean squared error, weighted with `Weights`. If you include `LossFun`, `L` reflects the loss calculated with `LossFun`.

`se` — Standard error of loss
vector of scalar values

Standard error of loss, returned as a vector the length of `Subtrees`.

`NLeaf` — Number of leaf nodes
vector of integer values

Number of leaves (terminal nodes) in the pruned subtrees, returned as a vector the length of `Subtrees`.

`bestlevel` — Best pruning level
scalar value

Best pruning level as defined in the `TreeSize` name-value pair, returned as a scalar whose value depends on `TreeSize`:

- `TreeSize` = `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L+se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `TreeSize` = `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Examples

Compute the In-Sample MSE

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using all observations.

```
tree = fitrtree(X,MPG);
```

Estimate the in-sample MSE.

```
L = loss(tree,X,MPG)
```

```
L = 4.8952
```

Find the Pruning Level Yielding the Optimal In-sample Loss

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

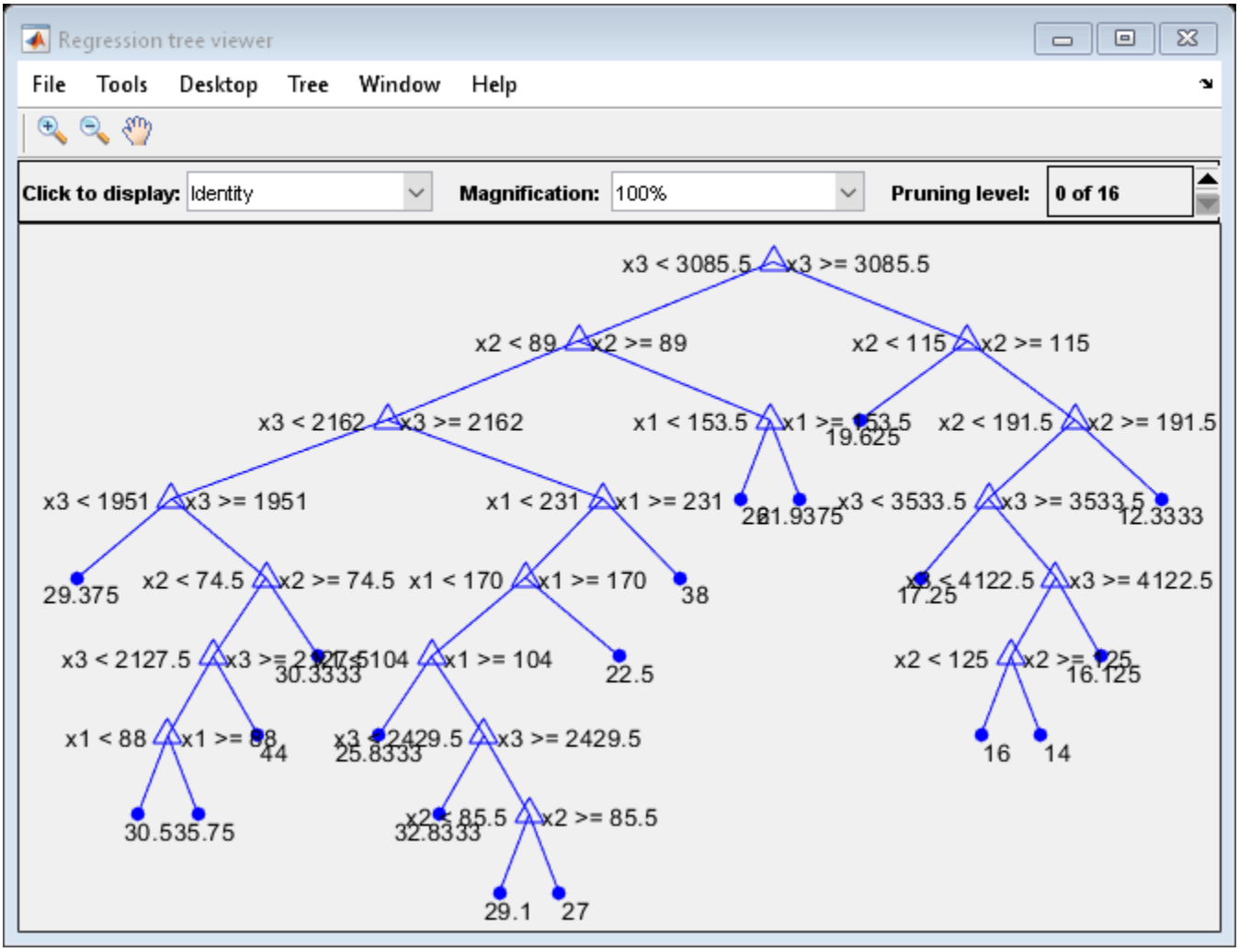
```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using all observations.

```
Mdl = fitrtree(X,MPG);
```

View the regression tree.

```
view(Mdl,'Mode','graph');
```



Find the best pruning level that yields the optimal in-sample loss.

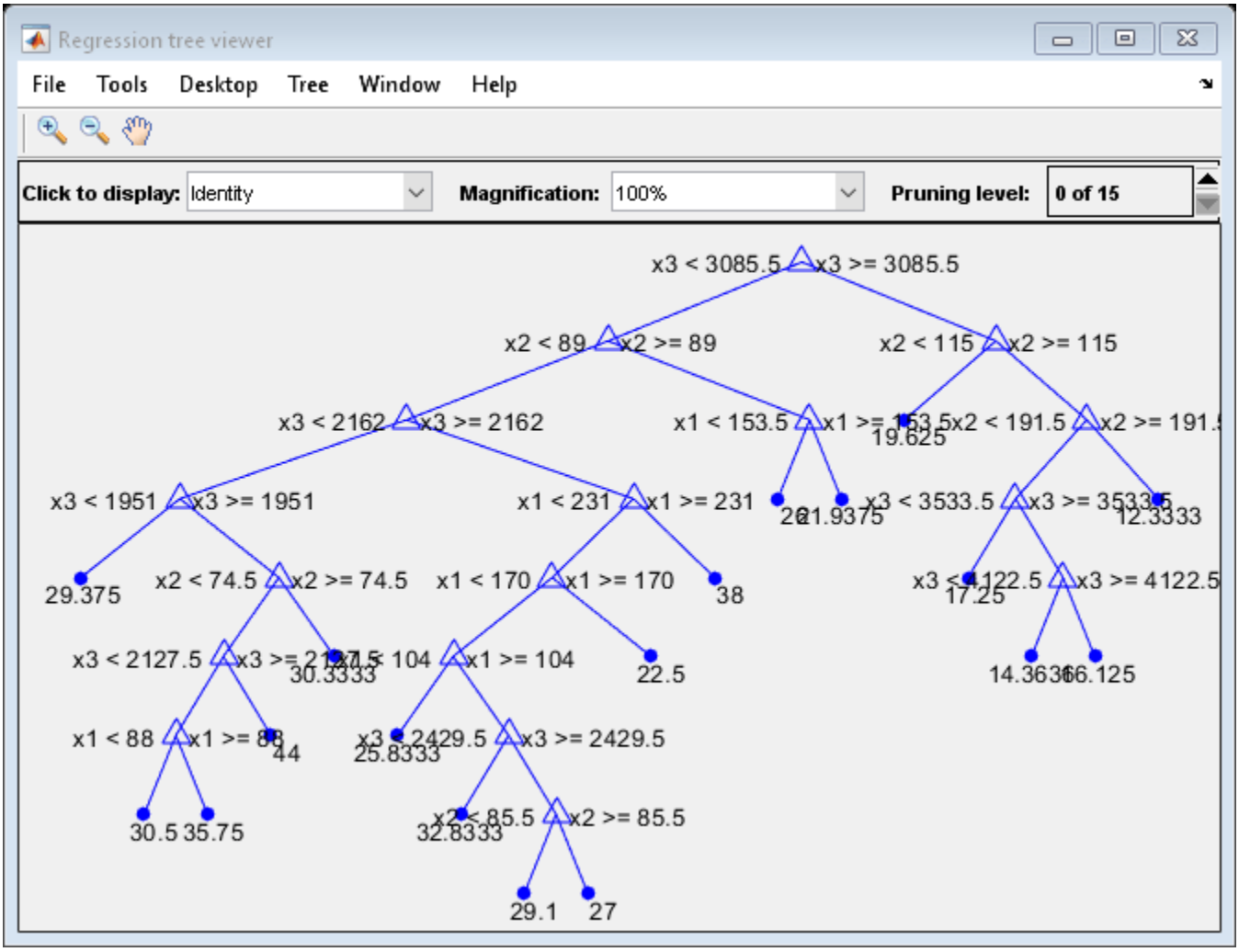
```
[L,se,NLeaf,bestLevel] = loss(Mdl,X,MPG,'Subtrees','all');
bestLevel
```

```
bestLevel = 1
```

The best pruning level is level 1.

Prune the tree to level 1.

```
pruneMdl = prune(Mdl,'Level',bestLevel);
view(pruneMdl,'Mode','graph');
```



Examine the MSE for Each Subtree

Unpruned decision trees tend to overfit. One way to balance model complexity and out-of-sample performance is to prune a tree (or restrict its growth) so that in-sample and out-of-sample performance are satisfactory.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
Y = MPG;
```

Partition the data into training (50%) and validation (50%) sets.

```
n = size(X,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
```

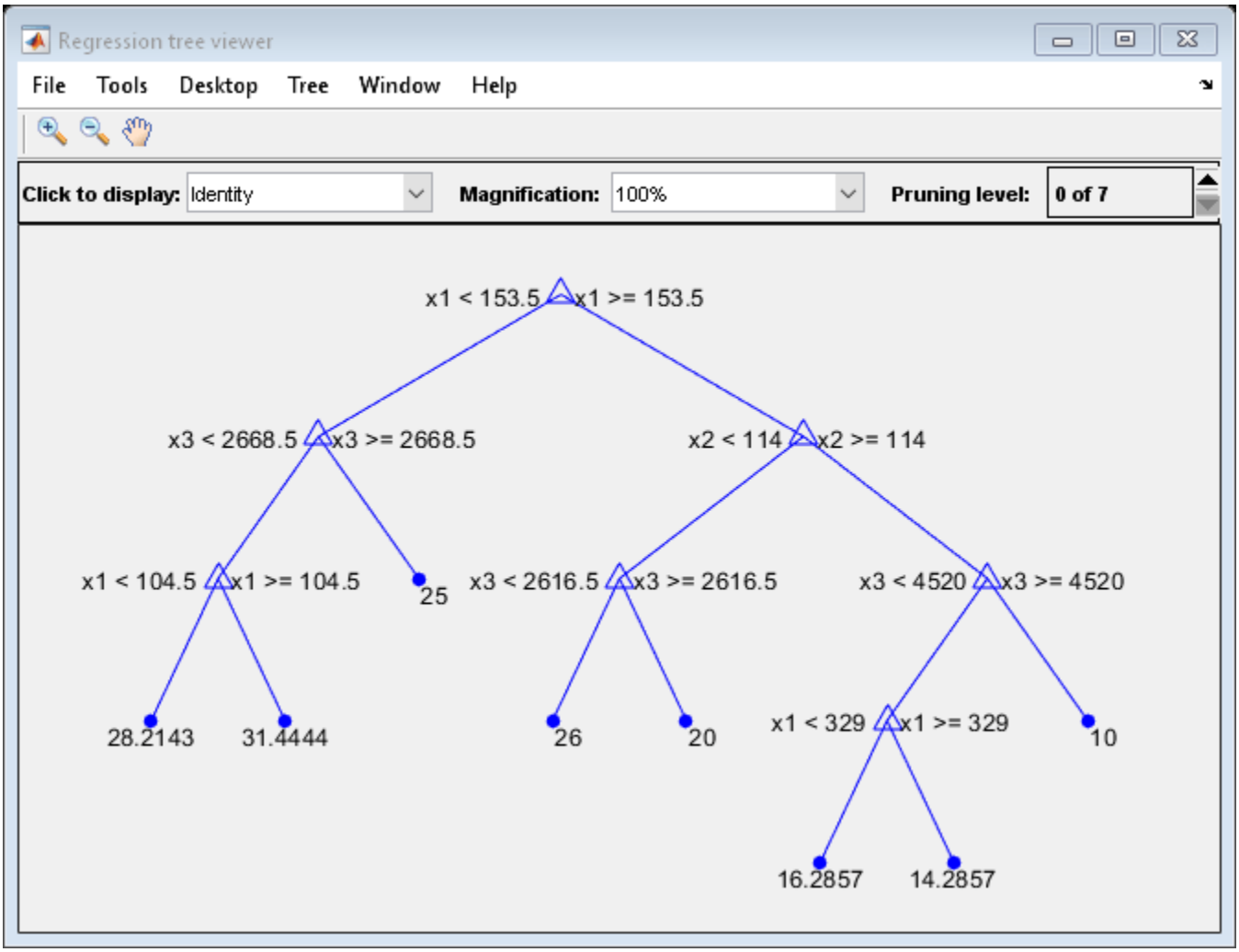
```
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a regression tree using the training set.

```
Mdl = fitrtree(X(idxTrn,:),Y(idxTrn));
```

View the regression tree.

```
view(Mdl,'Mode','graph');
```



The regression tree has seven pruning levels. Level 0 is the full, unpruned tree (as displayed). Level 7 is just the root node (i.e., no splits).

Examine the training sample MSE for each subtree (or pruning level) excluding the highest level.

```
m = max(Mdl.PruneList) - 1;
trnLoss = resubLoss(Mdl,'SubTrees',0:m)
```

```
trnLoss = 7×1

    5.9789
```

```
    6.2768
    6.8316
    7.5209
    8.3951
   10.7452
   14.8445
```

- The MSE for the full, unpruned tree is about 6 units.
- The MSE for the tree pruned to level 1 is about 6.3 units.
- The MSE for the tree pruned to level 6 (i.e., a stump) is about 14.8 units.

Examine the validation sample MSE at each level excluding the highest level.

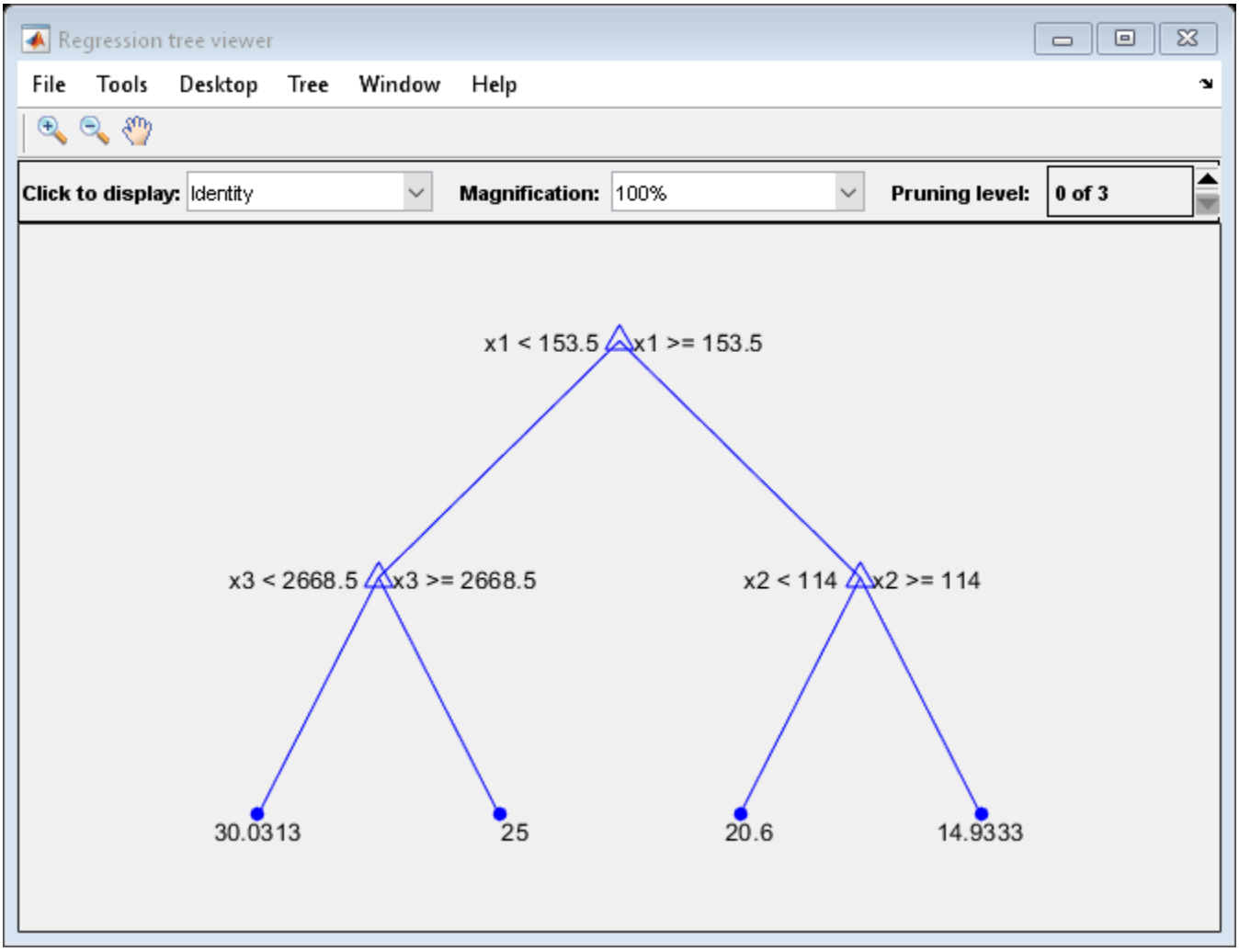
```
valLoss = loss(Mdl,X(idxVal,:),Y(idxVal),'SubTrees',0:m)
```

```
valLoss = 7×1

   32.1205
   31.5035
   32.0541
   30.8183
   26.3535
   30.0137
   38.4695
```

- The MSE for the full, unpruned tree (level 0) is about 32.1 units.
- The MSE for the tree pruned to level 4 is about 26.4 units.
- The MSE for the tree pruned to level 5 is about 30.0 units.
- The MSE for the tree pruned to level 6 (i.e., a stump) is about 38.5 units.

To balance model complexity and out-of-sample performance, consider pruning `Mdl` to level 4.

```
pruneMdl = prune(Mdl,'Level',4);
view(pruneMdl,'Mode','graph')
```



More About

Mean Squared Error

The mean squared error m of the predictions $f(X_n)$ with weight vector w is

$$m = \frac{\sum w_n (f(X_n) - Y_n)^2}{\sum w_n}.$$

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- Only one output is supported.

- You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

See Also

`fitrtree` | `predict`

loss

Loss of linear model for incremental learning on batch of data

Syntax

```
L = loss(Mdl,X,Y)
L = loss(Mdl,X,Y,Name,Value)
```

Description

`loss` returns the regression or classification loss of a configured incremental learning model for linear regression (`incrementalRegressionLinear` object) or linear binary classification (`incrementalClassificationLinear` object).

To measure model performance on a data stream and store the results in the output model, call `updateMetrics` or `updateMetricsAndFit`.

`L = loss(Mdl,X,Y)` returns the loss for the incremental learning model `Mdl` using the batch of predictor data `X` and corresponding responses `Y`.

`L = loss(Mdl,X,Y,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify that the columns of the predictor data matrix correspond to observations, or specify the classification loss function .

Examples

Measure Model Performance During Incremental Learning

The performance of an incremental model on streaming data is measured in three ways:

1. Cumulative metrics measure the performance since the start of incremental learning.
2. Window metrics measure the performance on a specified window of observations. The metrics are updated every time the model processes the specified window.
3. The `loss` function measures the performance on a specified batch of data only.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Create an incremental linear SVM model for binary classification. Configure it for `loss` by specifying the class names, prior class distribution (uniform), and arbitrary coefficient and bias values. Specify a metrics window size of 1000 observations.

```
p = size(X,2);
Beta = randn(p,1);
Bias = randn(1);
Mdl = incrementalClassificationLinear('Beta',Beta,'Bias',Bias,...
    'ClassNames',unique(Y),'Prior','uniform','MetricsWindowSize',1000);
```

`Mdl` is an `incrementalClassificationLinear` model. All its properties are read-only. Instead of specifying arbitrary values, you can take either of these actions to configure the model:

- Train an SVM model using `fitcsvm` or `fitclinear` on a subset of the data (if available), and then convert the model to an incremental learner by using `incrementalLearner`.
- Incrementally fit `Mdl` to data by using `fit`.

Simulate a data stream, and perform the following actions on each incoming chunk of 50 observations:

1. Call `updateMetrics` to measure the cumulative performance and the performance within a window of observations. Overwrite the previous incremental model with a new one to track performance metrics.
2. Call `loss` to measure the model performance on the incoming chunk.
3. Call `fit` to fit the model to the incoming chunk. Overwrite the previous incremental model with a new one fitted to the incoming observation.
4. Store all performance metrics to see how they evolve during incremental learning.

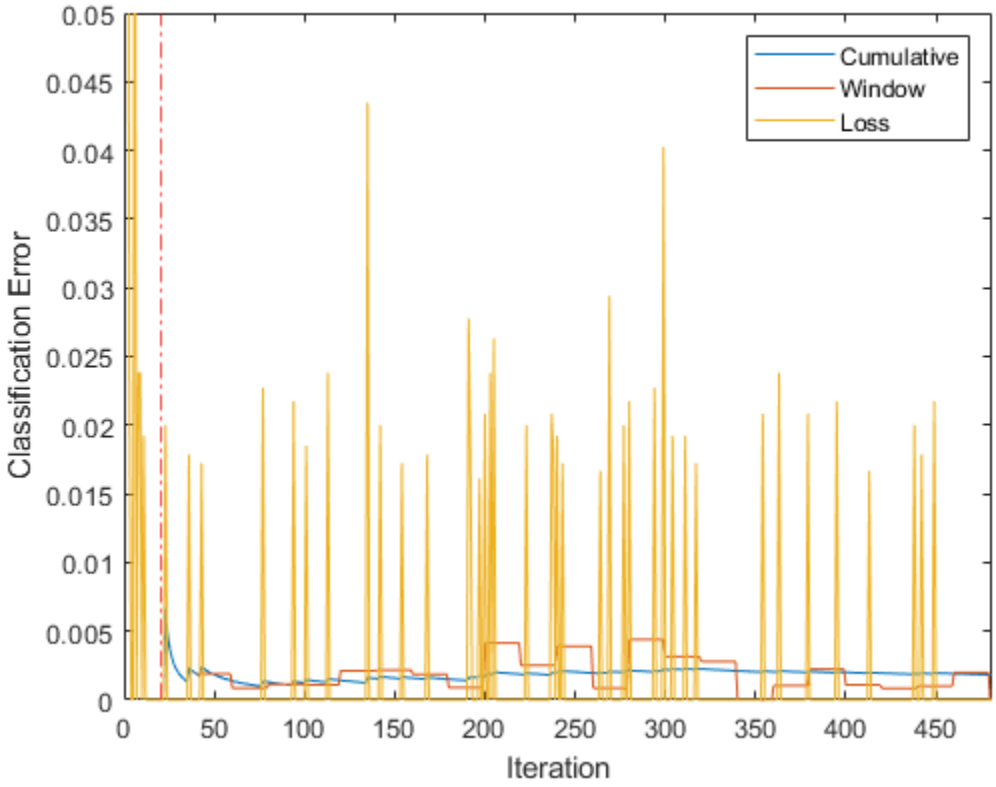
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,3),'VariableNames',["Cumulative" "Window" "Loss"]);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ce{j,["Cumulative" "Window"]} = Mdl.Metrics{"ClassificationError",:};
    ce{j,"Loss"} = loss(Mdl,X(idx,:),Y(idx));
    Mdl = fit(Mdl,X(idx,:),Y(idx));
end
```

`Mdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetrics` checks the performance of the model on the incoming observation, then and the `fit` function fits the model to that observation. `loss` is agnostic of the metrics warm-up period, so it measures the classification error for all iterations.

To see how the performance metrics evolved during training, plot them.

```
figure;
plot(ce.Variables);
xlim([0 nchunk]);
```

```
ylim([0 0.05])
ylabel('Classification Error')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(ce.Properties.VariableNames)
xlabel('Iteration')
```



During the metrics warm-up period (the area to the left of the red line), the yellow line represents the classification error on each incoming chunk of data. After the metrics warm-up period, `Mdl` tracks the cumulative and window metrics. The cumulative and batch losses converge as the `fit` function fits the incremental model to the incoming data.

Compute Custom Loss on Incoming Chunks of Data

Fit an incremental learning model for regression to streaming data, and compute the mean absolute deviation (MAD) on the incoming data batches.

Load the robot arm data set. Obtain the sample size `n` and the number of predictor variables `p`.

```
load robotarm
n = numel(ytrain);
p = size(Xtrain,2);
```

For details on the data set, enter `Description` at the command line.

Create an incremental linear model for regression. Configure the model as follows:

- Specify a metrics warm-up period of 1000 observations.
- Specify a metrics window size of 500 observations.
- Track the mean absolute deviation (MAD) to measure the performance of the model. Create an anonymous function that measures the absolute error of each new observation. Create a structure array containing the name `MeanAbsoluteError` and its corresponding function.
- Configure the model to predict responses by specifying that all regression coefficients and the bias are 0.

```
maefcn = @(z,zfit,w)(abs(z - zfit));
maemetric = struct("MeanAbsoluteError",maefcn);

Mdl = incrementalRegressionLinear('MetricsWarmupPeriod',1000,'MetricsWindowSize',500,...
    'Metrics',maemetric,'Beta',zeros(p,1),'Bias',0,'EstimationPeriod',0)
```

```
Mdl =
  incrementalRegressionLinear

               IsWarm: 0
              Metrics: [2x2 table]
    ResponseTransform: 'none'
                 Beta: [32x1 double]
                 Bias: 0
              Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalRegressionLinear` model object configured for incremental learning.

Perform incremental learning. At each iteration:

- Simulate a data stream by processing a chunk of 50 observations.
- Call `updateMetrics` to compute cumulative and window metrics on the incoming chunk of data. Overwrite the previous incremental model with a new one fitted to overwrite the previous metrics.
- Call `loss` to compute the MAD on the incoming chunk of data. Whereas the cumulative and window metrics require that custom losses return the loss for each observation, `loss` requires the loss on the entire chunk. Compute the mean of the absolute deviation.
- Call `fit` to fit the incremental model to the incoming chunk of data.
- Store the cumulative, window, and chunk metrics to see how they evolve during incremental learning.

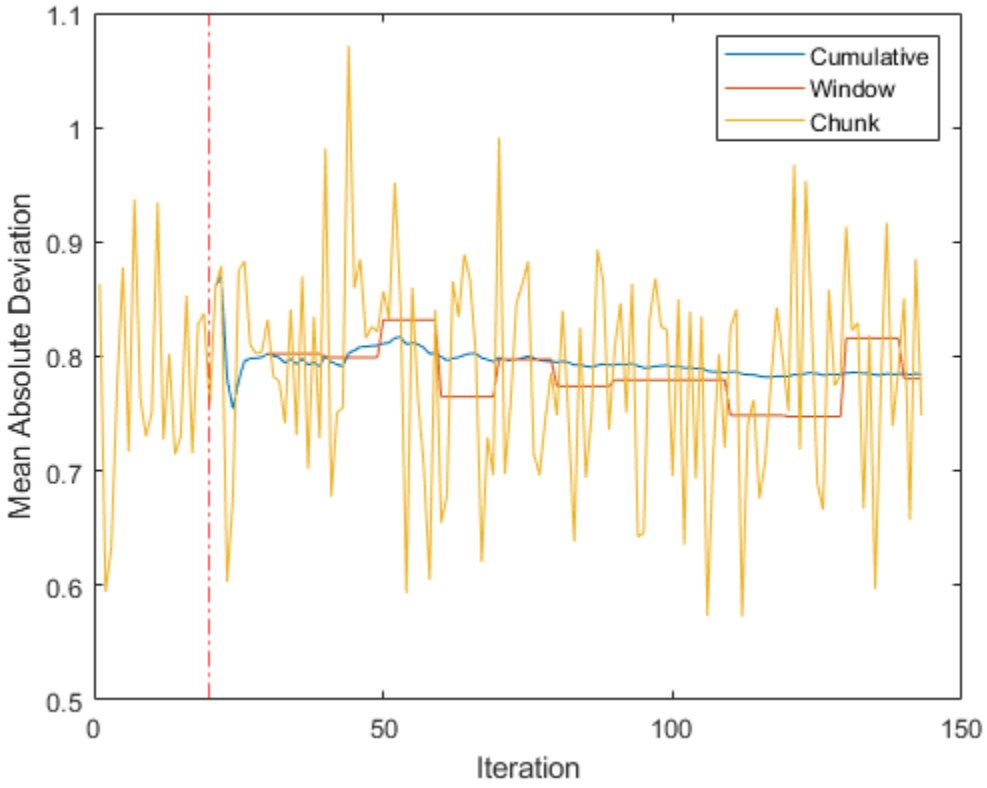
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
mae = array2table(zeros(nchunk,3),'VariableNames',["Cumulative" "Window" "Chunk"]);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,Xtrain(idx,:),ytrain(idx));
    mae{j,1:2} = Mdl.Metrics{"MeanAbsoluteError",:};
```

```
    mae{j,3} = loss(Mdl,Xtrain(idx,:),ytrain(idx),'LossFun',@(x,y,w)mean(maefcn(x,y,w)));
    Mdl = fit(Mdl,Xtrain(idx,:),ytrain(idx));
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetrics` checks the performance of the model on the incoming observation, and the `fit` function fits the model to that observation.

Plot the performance metrics to see how they evolved during incremental learning.

```
figure;
h = plot(mae.Variables);
ylabel('Mean Absolute Deviation')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
legend(h,mae.Properties.VariableNames)
```



The plot suggests the following:

- `updateMetrics` computes performance metrics after the metrics warm-up period only.
- `updateMetrics` computes the cumulative metrics during each iteration.
- `updateMetrics` computes the window metrics after processing 500 observations
- Because `Mdl` was configured to predict observations from the beginning of incremental learning, `loss` can compute the MAD on each incoming chunk of data.

Input Arguments

`Mdl` — Incremental learning model

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Incremental learning model, specified as an `incrementalClassificationLinear` or `incrementalRegressionLinear` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

You must configure `Mdl` to compute its loss on a batch of observations.

- If `Mdl` is a converted, traditionally trained model, you can compute its loss without any modifications.
- Otherwise, `Mdl` must satisfy the following criteria, which you can specify directly or by fitting `Mdl` to data using `fit` or `updateMetricsAndFit`.
  - If `Mdl` is an `incrementalRegressionLinear` model, its model coefficients `Mdl.Beta` and bias `Mdl.Bias` must be nonempty arrays.
  - If `Mdl` is an `incrementalClassificationLinear` model, its model coefficients `Mdl.Beta` and bias `Mdl.Bias` must be nonempty arrays, the class names `Mdl.ClassNames` must contain two classes, and the prior class distribution `Mdl.Prior` must contain known values.
  - Regardless of object type, if you configure the model so that functions standardize predictor data, the predictor means `Mdl.Mu` and standard deviations `Mdl.Sigma` must be nonempty arrays.

`X` — Batch of predictor data

floating-point matrix

Batch of predictor data with which to compute the loss, specified as a floating-point matrix of *n* observations and `Mdl.NumPredictors` predictor variables. The value of the `'ObservationsIn'` name-value pair argument determines the orientation of the variables and observations.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note `loss` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`Y` — Batch of labels

categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Batch of labels with which to compute the loss, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors for classification problems; or a floating-point vector for regression problems.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(j)` is the label of observation *j* (row or column) in `X`.

For classification problems:

- `loss` supports binary classification only.
- When the `ClassNames` property of the input model `Mdl` is nonempty, the following conditions apply:
  - If `Y` contains a label that is not a member of `Mdl.ClassNames`, `loss` issues an error.
  - The data type of `Y` and `Mdl.ClassNames` must be the same.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ObservationsIn','columns','Weights',W` specifies that the columns of the predictor matrix correspond to observations, and the vector `W` contains observation weights to apply.

LossFun — Loss function

string vector | function handle | cell vector | structure array | ...

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or function handle.

- **Classification problems**: The following table lists the available loss functions when `Mdl` is an `incrementalClassificationLinear` model. Specify one using its corresponding character vector or string scalar.

Name	Description
`"binodeviance"`	Binomial deviance
`"classiferror"` (default)	Misclassification rate in decimal
`"exponential"`	Exponential loss
`"hinge"`	Hinge loss
`"logit"`	Logistic loss
`"quadratic"`	Quadratic loss

  For more details, see “Classification Loss” on page 33-3760.

  Logistic regression learners return posterior probabilities as classification scores, but SVM learners do not (see `predict`).

  To specify a custom loss function, use function handle notation. The function must have this form:

  `lossval = lossfcn(C,S,W)`

  - The output argument `lossval` is an *n*-by-1 floating-point vector, where `lossval(j)` is the classification loss of observation *j*.

- You specify the function name (*lossfcn*).
- `C` is an *n*-by-2 logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in the `ClassNames` property. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-2 numeric matrix of predicted classification scores. `S` is similar to the `score` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in the `ClassNames` property. `S(p,q)` is the classification score of observation *p* being classified in class *q*.
- `W` is an *n*-by-1 numeric vector of observation weights.

- **Regression problems**: The following table lists the available loss functions when `Mdl` is an `incrementalRegressionLinear` model. Specify one using its corresponding character vector or string scalar.

Name	Description	Learners Supporting Metric
`"epsiloninsensitive"`	Epsilon insensitive loss	`'svm'`
`"mse"` (default)	Weighted mean squared error	`'svm'` and `'leastsquares'`

For more details, see “Regression Loss” on page 33-3761.

To specify a custom loss function, use function handle notation. The function must have this form:

```
lossval = lossfcn(Y,YFit,W)
```

- The output argument `lossval` is a floating-point scalar.
- You specify the function name (*lossfcn*).
- `Y` is a length *n* numeric vector of observed responses.
- `YFit` is a length *n* numeric vector of corresponding predicted responses.
- `W` is an *n*-by-1 numeric vector of observation weights.

Example: `'LossFun',"mse"`

Example: `'LossFun',@lossfcn`

Data Types: `char` | `string` | `function_handle`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`.

Data Types: `char` | `string`

`Weights` — Batch of observation weights
floating-point vector of positive values

Batch of observation weights, specified as the comma-separated pair consisting of `'Weights'` and a floating-point vector of positive values. `loss` weighs the observations in the input data with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in the input data.

By default, `Weights` is `ones(n,1)`.

For more details, see “Observation Weights” on page 33-3762.

Data Types: `double` | `single`

Output Arguments

L — Classification or regression loss

numeric scalar

Classification or regression loss, returned as a numeric scalar. The interpretation of `L` depends on `Weights` and `LossFun`.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

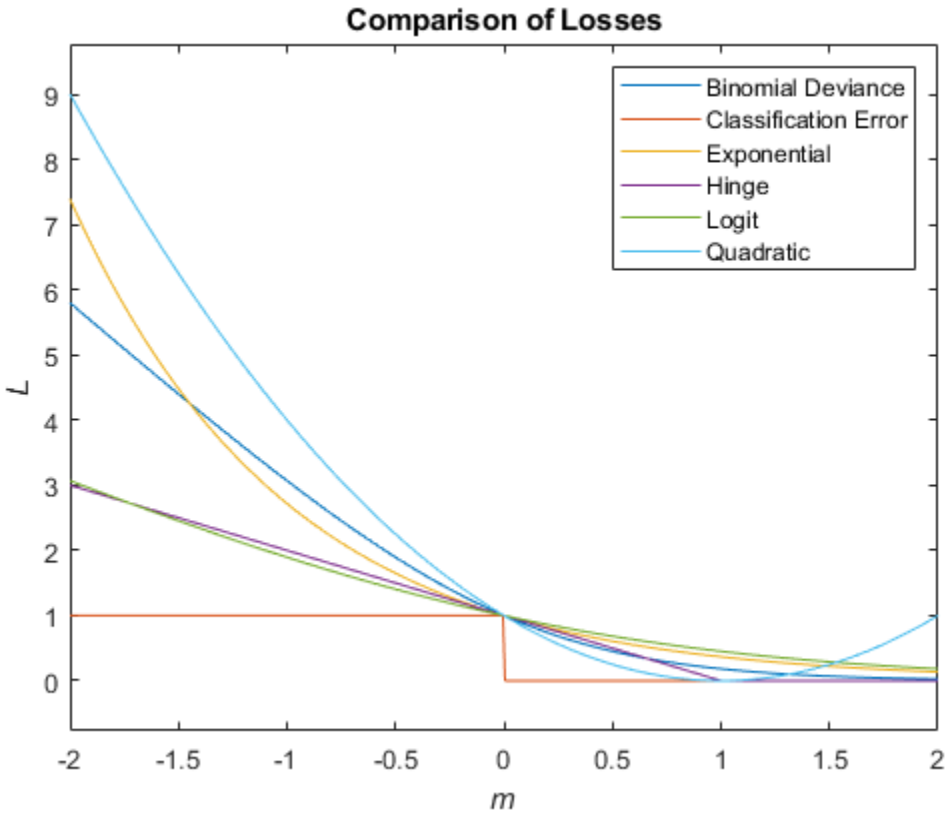
- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- The weight for observation j is w_j.

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of `LossFun`	Equation
Binomial deviance	`"binodeviance"`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Exponential loss	`"exponential"`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Misclassification rate in decimal	`"classiferror"`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.

Loss Function	Value of LossFun	Equation
Hinge loss	`"hinge"`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`"logit"`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Quadratic loss	`"quadratic"`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Regression Loss

Regression loss functions measure the predictive inaccuracy of regression models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.

- y_j is the observed response of observation j.
- $f(X_j) = \beta_0 + x_j\beta$ is the predicted value of observation j of the predictor data X, where β_0 is the bias and β is the vector of coefficients.
- The weight for observation j is w_j.

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Epsilon insensitive loss	`"epsiloninsensitive"`	$L = \max[0, \lvert y - f(x)\rvert - \varepsilon]$.
Mean squared error	`"mse"`	$L = [y - f(x)]^2$.

Algorithms

Observation Weights

For classification problems, if the prior class probability distribution is known (in other words, the prior distribution is not empirical), `loss` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that observation weights are the respective prior class probabilities by default.

For regression problems or if the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `loss`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `loss` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `loss` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for loss, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `loss`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For usage notes and limitations of the model object, see `incrementalClassificationLinear` or `incrementalRegressionLinear`.

Argument	Notes and Limitations
X	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `Y`. • The number of predictor variables must equal to `Mdl.NumPredictors`. • `X` must be `single` or `double`.
Y	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `X`. • For classification problems, all labels in `Y` must be represented in `Mdl.ClassNames`. • `Y` and `Mdl.ClassNames` must have the same data type.
`'LossFun'`	The specified function cannot be an anonymous function.

- If you configure `Mdl` to shuffle data (`Mdl.Shuffle` is `true`, or `Mdl.Solver` is `'sgd'` or `'asgd'`), the `loss` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB. Therefore, if you fit `Mdl` before computing the loss, the loss computed in MATLAB and those computed by the generated code might not be equal.
- Use a homogeneous data type for all floating-point input arguments and object properties, specifically, either `single` or `double`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Objects
`incrementalClassificationLinear` | `incrementalRegressionLinear`

Functions
`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8

Introduced in R2020b

loss

Loss of naive Bayes classification model for incremental learning on batch of data

Syntax

```
L = loss(Mdl,X,Y)
L = loss(Mdl,X,Y,Name,Value)
```

Description

`loss` returns the classification loss of a configured naive Bayes classification model for incremental learning model (`incrementalClassificationNaiveBayes` object).

To measure model performance on a data stream and store the results in the output model, call `updateMetrics` or `updateMetricsAndFit`.

`L = loss(Mdl,X,Y)` returns the minimal cost classification loss for the naive Bayes classification model for incremental learning `Mdl` using the batch of predictor data `X` and corresponding responses `Y`.

`L = loss(Mdl,X,Y,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify the classification loss function .

Examples

Measure Model Performance During Incremental Learning

The performance of an incremental model on streaming data is measured in three ways:

1. Cumulative metrics measure the performance since the start of incremental learning.
2. Window metrics measure the performance on a specified window of observations. The metrics are updated every time the model processes the specified window.
3. The `loss` function measures the performance on a specified batch of data only.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Create a naive Bayes classification model for incremental learning; specify the class names and a metrics window size of 1000 observations. Configure it for `loss` by fitting it to the first 10 observations.

```
Mdl = incrementalClassificationNaiveBayes('ClassNames',unique(Y),'MetricsWindowSize',1000);
initobs = 10;
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs));
canComputeLoss = (size(Mdl.DistributionParameters,2) == Mdl.NumPredictors) +...
    (size(Mdl.DistributionParameters,1) > 1) > 1
```

```
canComputeLoss = logical
   1
```

`Mdl` is an `incrementalClassificationLinear` model. All its properties are read-only.

Simulate a data stream, and perform the following actions on each incoming chunk of 50 observations:

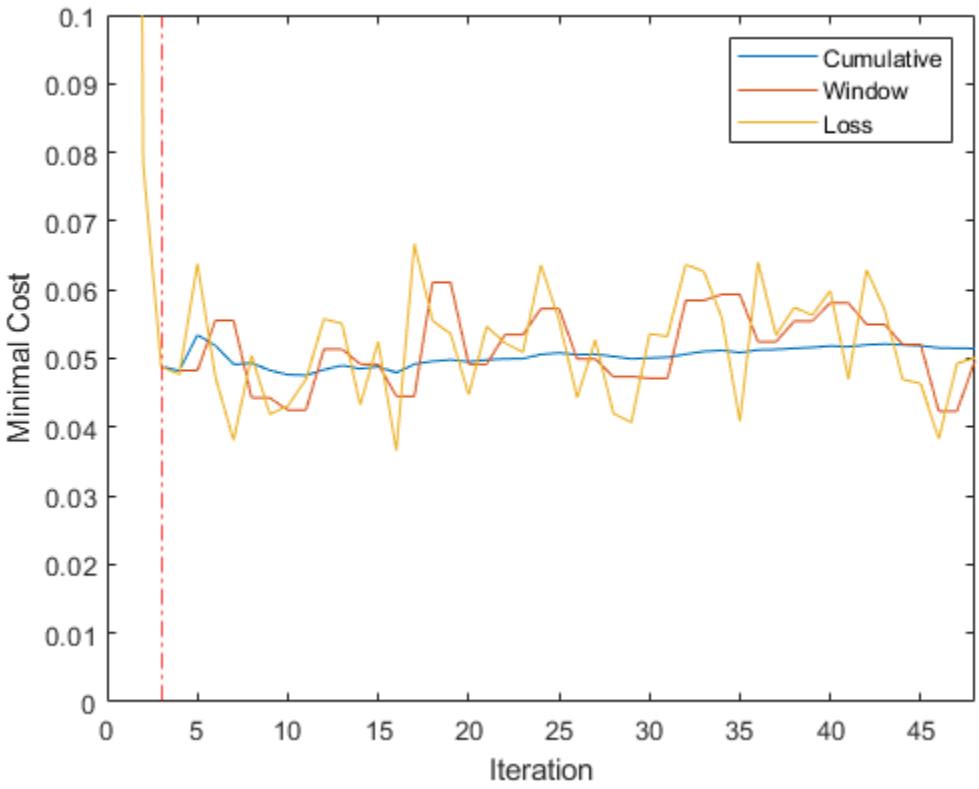
1. Call `updateMetrics` to measure the cumulative performance and the performance within a window of observations. Overwrite the previous incremental model with a new one to track performance metrics.
2. Call `loss` to measure the model performance on the incoming chunk.
3. Call `fit` to fit the model to the incoming chunk. Overwrite the previous incremental model with a new one fitted to the incoming observation.
4. Store all performance metrics to see how they evolve during incremental learning.

```
% Preallocation
numObsPerChunk = 500;
nchunk = floor((n - initobs)/numObsPerChunk);
mc = array2table(zeros(nchunk,3),'VariableNames',["Cumulative" "Window" "Loss"]);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    mc{j,["Cumulative" "Window"]} = Mdl.Metrics{"MinimalCost",:};
    mc{j,"Loss"} = loss(Mdl,X(idx,:),Y(idx));
    Mdl = fit(Mdl,X(idx,:),Y(idx));
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetrics` checks the performance of the model on the incoming observation, then and the `fit` function fits the model to that observation. `loss` is agnostic of the metrics warm-up period, so it measures the minimal cost for all iterations.

To see how the performance metrics evolved during training, plot them.

```
figure;
plot(mc.Variables);
xlim([0 nchunk]);
ylim([0 0.1])
ylabel('Minimal Cost')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk + 1,'r-.');
legend(mc.Properties.VariableNames)
xlabel('Iteration')
```



During the metrics warm-up period (the area to the left of the red line), the yellow line represents the minimal cost on each incoming chunk of data. After the metrics warm-up period, `Mdl` tracks the cumulative and window metrics. The cumulative and batch losses converge as the `fit` function fits the incremental model to the incoming data.

Compute Custom Loss on Incoming Chunks of Data

Fit a naive Bayes classification model for incremental learning to streaming data, and compute the multiclass cross entropy loss on the incoming chunks of data.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Create a naive Bayes classification model for incremental learning. Configure the model as follows:

- Specify the class names

- Specify a metrics warm-up period of 1000 observations.
- Specify a metrics window size of 2000 observations.
- Track multiclass cross entropy loss to measure the performance of the model. Create an anonymous function that measures the multiclass cross entropy loss of each new observation, include a tolerance for numerical stability. Create a structure array containing the name `CrossEntropy` and its corresponding function.
- Configure the model to compute classification loss by fitting the model to the first 10 observations.

```
tolerance = 1e-10;
crossentropy = @(z,zfit,w,cost)-log(max(zfit(z),tolerance));
ce = struct("CrossEntropy",crossentropy);

Mdl = incrementalClassificationNaiveBayes('ClassNames',unique(Y),'MetricsWarmupPeriod',1000,...
    'MetricsWindowSize',2000,'Metrics',ce);
initobs = 10;
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs));
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object configured for incremental learning.

Perform incremental learning. At each iteration:

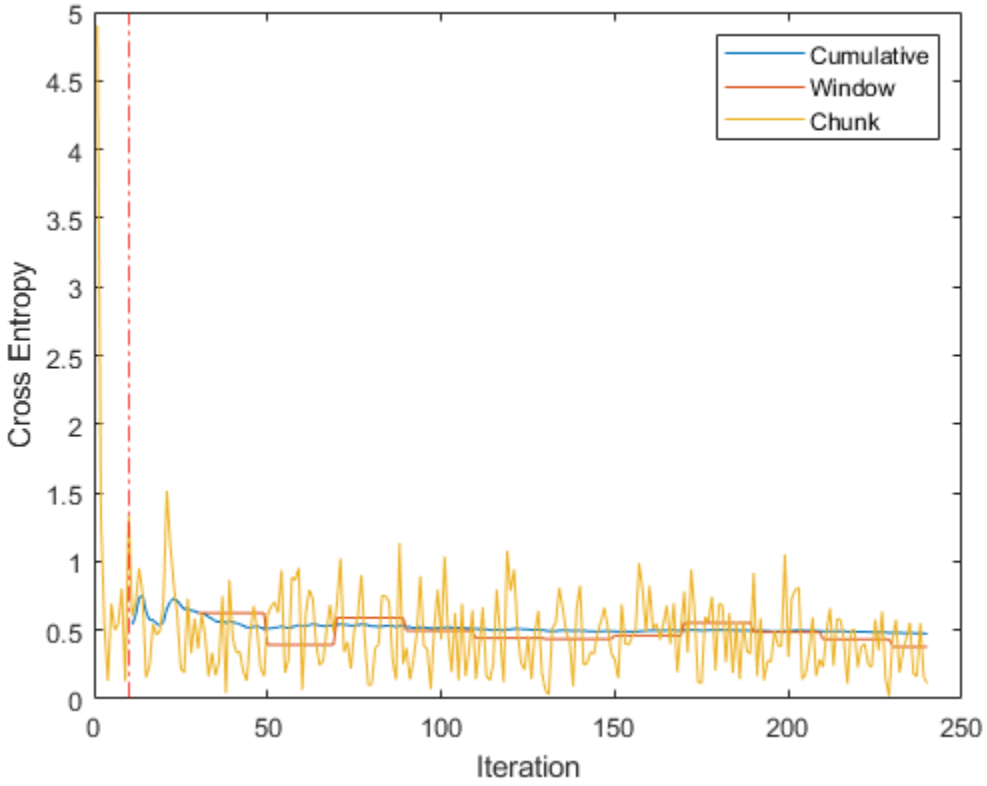
- Simulate a data stream by processing a chunk of 100 observations.
- Call `updateMetrics` to compute cumulative and window metrics on the incoming chunk of data. Overwrite the previous incremental model with a new one fitted to overwrite the previous metrics.
- Call `loss` to compute the cross entropy on the incoming chunk of data. Whereas the cumulative and window metrics require that custom losses return the loss for each observation, `loss` requires the loss on the entire chunk. Compute the mean of the losses within a chunk.
- Call `fit` to fit the incremental model to the incoming chunk of data.
- Store the cumulative, window, and chunk metrics to see how they evolve during incremental learning.

```
% Preallocation
numObsPerChunk = 100;
nchunk = floor((n - initobs)/numObsPerChunk);
tanloss = array2table(zeros(nchunk,3),'VariableNames',["Cumulative" "Window" "Chunk"]);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    tanloss{j,1:2} = Mdl.Metrics{"CrossEntropy",:};
    tanloss{j,3} = loss(Mdl,X(idx,:),Y(idx),'LossFun',@(z,zfit,w,cost)mean(crossentropy(z,zfit,w
    Mdl = fit(Mdl,X(idx,:),Y(idx));
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetrics` checks the performance of the model on the incoming observation, and the `fit` function fits the model to that observation.

Plot the performance metrics to see how they evolved during incremental learning.

```
figure;
h = plot(tanloss.Variables);
ylabel('Cross Entropy')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
legend(h,tanloss.Properties.VariableNames)
```



The plot suggests the following:

- `updateMetrics` computes performance metrics after the metrics warm-up period only.
- `updateMetrics` computes the cumulative metrics during each iteration.
- `updateMetrics` computes the window metrics after processing 100 observations
- Because `Mdl` was configured to predict observations from the beginning of incremental learning, `loss` can compute the cross entropy on each incoming chunk of data.

Input Arguments

`Mdl` — Naive Bayes classification model for incremental learning

`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

You must configure `Mdl` to compute its loss on a batch of observations.

- If `Mdl` is a converted, traditionally trained model, you can compute its loss without any modifications.
- Otherwise, you must fit the input model `Mdl` to data that contained all expected classes (`Mdl.DistributionParameters` must be a cell matrix with `Mdl.NumPredictors` columns and at least one row, where each row corresponds to each class name in `Mdl.ClassNames`).

`X` — Batch of predictor data
floating-point matrix

Batch of predictor data with which to compute the loss, specified as an *n*-by-`Mdl.NumPredictors` floating point matrix.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*`j`*`)` is the label of observation *j* (row or column) in `X`.

Note `loss` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`Y` — Batch of labels
categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Batch of labels with which to compute the loss, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors for classification problems; or a floating-point vector for regression problems.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*`j`*`)` is the label of observation *j* (row or column) in `X`.

- When the `ClassNames` property of the input model `Mdl` is nonempty, the following conditions apply:
  - If `Y` contains a label that is not a member of `Mdl.ClassNames`, `loss` issues an error.
  - The data type of `Y` and `Mdl.ClassNames` must be the same.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'LossFun','classiferror','Weights',W` specifies returning the misclassification error rate, and the observation weights `W`.

`LossFun` — Loss function
`'mincost'` (default) | string vector | function handle | cell vector | structure array | ...

Loss function, specified as a built-in loss function name or function handle.

The following table lists the built-in loss function names. You can specify more than one by using a string vector.

Name	Description
"binodeviance"	Binomial deviance
"classiferror"	Misclassification error rate
"exponential"	Exponential
"hinge"	Hinge
"logit"	Logistic
'"mincost"	Minimal expected misclassification cost
"quadratic"	Quadratic

For more details, see “Classification Loss” on page 33-3771.

To specify a custom loss function, use function handle notation. The function must have this form:

`lossval = lossfcn(C,S,W,Cost)`

- The output argument `lossval` is an *n*-by-1 floating-point vector, where `lossval(j)` is the classification loss of observation *j*.
- You specify the function name (*lossfcn*).
- `C` is an *n*-by-2 logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in the `ClassNames` property. Create `C` by setting `C(p,q) = 1`, if observation *p* is in class *q*, for each observation in the specified data. Set the other element in row *p* to `0`.
- `S` is an *n*-by-2 numeric matrix of predicted classification scores. `S` is similar to the `Posterior` output of `predict`, where rows correspond to observations in the data and the column order corresponds to the class order in the `ClassNames` property. `S(p,q)` is the classification score of observation *p* being classified in class *q*.
- `W` is an *n*-by-1 numeric vector of observation weights.
- `Cost` is a *K*-by-*K* numeric matrix of misclassification costs.

Example: `'LossFun',"classiferror"`

Example: `'LossFun',@lossfcn`

Data Types: `char` | `string` | `function_handle`

`Weights` — Chunk of observation weights
floating-point vector of positive values

Chunk of observation weights, specified as a floating-point vector of positive values. `loss` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(n,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-3774.

Data Types: `double` | `single`

Output Arguments

L — Classification loss

numeric scalar

Classification loss, returned as a numeric scalar. `L` is a measure of model quality. Its interpretation depends on the loss function and weighting scheme.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

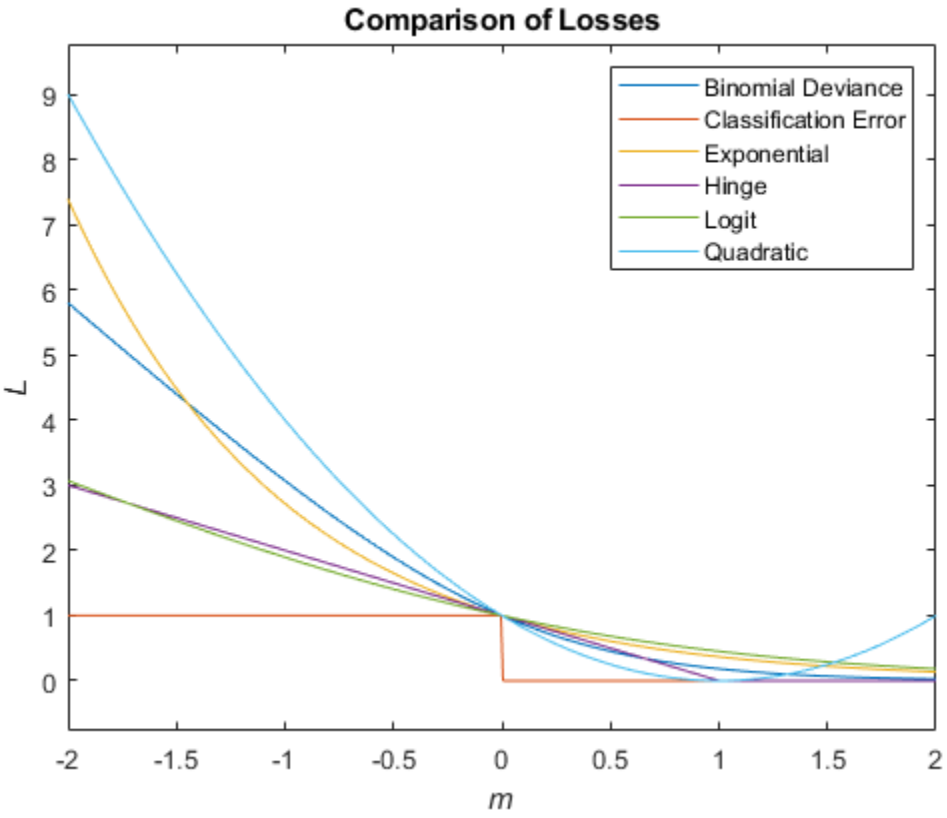
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Algorithms

Observation Weights

For each conditional predictor distribution, `loss` computes the weighted average and standard deviation.

If the prior class probability distribution is known (in other words, the prior distribution is not empirical), `loss` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that the default observation weights are the respective prior class probabilities.

If the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `loss`.

See Also

Objects

`incrementalClassificationNaiveBayes`

Functions

`fit` | `predict` | `updateMetrics` | `updateMetricsAndFit`

Topics

“Incremental Learning Overview” on page 26-2

“Configure Incremental Learning Model” on page 26-8

Introduced in R2021a

loss

Class: `FeatureSelectionNCAClassification`

Evaluate accuracy of learned feature weights on test data

Syntax

```
err = loss(mdl,X,Y)
err = loss(mdl,X,Y,Name,Value)
```

Description

`err = loss(mdl,X,Y)` computes the misclassification error of the model `mdl`, for the predictors in `X` and the class labels in `Y`.

`err = loss(mdl,X,Y,Name,Value)` computes the classification error with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl` — Neighborhood component analysis model for classification

`FeatureSelectionNCAClassification` object

Neighborhood component analysis model for classification, returned as a `FeatureSelectionNCAClassification` object.

`X` — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

`Y` — Class labels

categorical vector | logical vector | numeric vector | string array | cell array of character vectors of length *n* | character matrix with *n* rows

Class labels, specified as a categorical vector, logical vector, numeric vector, string array, cell array of character vectors of length *n*, or character matrix with *n* rows, where *n* is the number of observations. Element *i* or row *i* of `Y` is the class label corresponding to row *i* of `X` (observation *i*).

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFunction` — Loss function type

`'classiferror'` (default) | `'quadratic'`

Loss function type, specified as a comma-separated pair consisting of `'Loss Function'` and one of the following.

- `'classiferror'` — Misclassification rate in decimal, defined as

  $$\frac{1}{n}\sum_{i=1}^{n} I(k_i \neq t_i),$$

  where k_i is the predicted class and t_i is the true class for observation i. $I(k_i \neq t_i)$ is the indicator for when the k_i is not the same as t_i.

- `'quadratic'` — Quadratic loss function, defined as

  $$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} (p_{ik} - I(i,k))^2,$$

  where c is the number of classes, p_{ik} is the estimate probability that ith observation belongs to class k, and $I(i,k)$ is the indicator that ith observation belongs to class k.

Example: `'LossFunction','quadratic'`

Output Arguments

`err` — Smaller-the-better accuracy measure for learned feature weights

scalar value

Smaller-the-better accuracy measure for learned feature weights, returned as a scalar value. You can specify the measure of accuracy using the `LossFunction` name-value pair argument.

Examples

Tune NCA Model for Classification

Load the sample data.

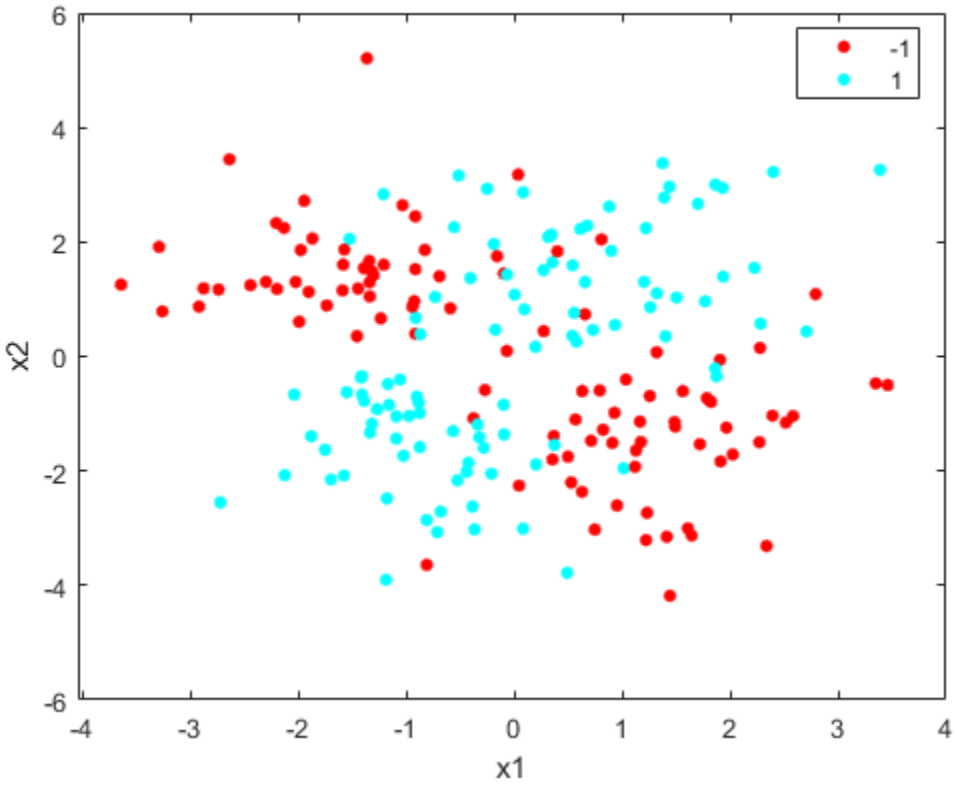
```
load('twodimclassdata.mat');
```

This data set is simulated using the scheme described in [1]. This is a two-class classification problem in two dimensions. Data from the first class (class –1) are drawn from two bivariate normal distributions $N(\mu_1, \Sigma)$ or $N(\mu_2, \Sigma)$ with equal probability, where $\mu_1 = [-0.75, -1.5]$, $\mu_2 = [0.75, 1.5]$, and $\Sigma = I_2$. Similarly, data from the second class (class 1) are drawn from two bivariate normal distributions $N(\mu_3, \Sigma)$ or $N(\mu_4, \Sigma)$ with equal probability, where $\mu_3 = [1.5, -1.5]$, $\mu_4 = [-1.5, 1.5]$, and $\Sigma = I_2$. The normal distribution parameters used to create this data set result in tighter clusters in data than the data used in [1].

Create a scatter plot of the data grouped by the class.

```
figure
gscatter(X(:,1),X(:,2),y)
```

```
xlabel('x1')
ylabel('x2')
```



Add 100 irrelevant features to X. First generate data from a Normal distribution with a mean of 0 and a variance of 20.

```
n = size(X,1);
rng('default')
XwithBadFeatures = [X,randn(n,100)*sqrt(20)];
```

Normalize the data so that all points are between 0 and 1.

```
XwithBadFeatures = bsxfun(@rdivide,...
    bsxfun(@minus,XwithBadFeatures,min(XwithBadFeatures,[],1)), ...
    range(XwithBadFeatures,1));
X = XwithBadFeatures;
```

Fit a neighborhood component analysis (NCA) model to the data using the default `Lambda` (regularization parameter, λ) value. Use the LBFGS solver and display the convergence information.

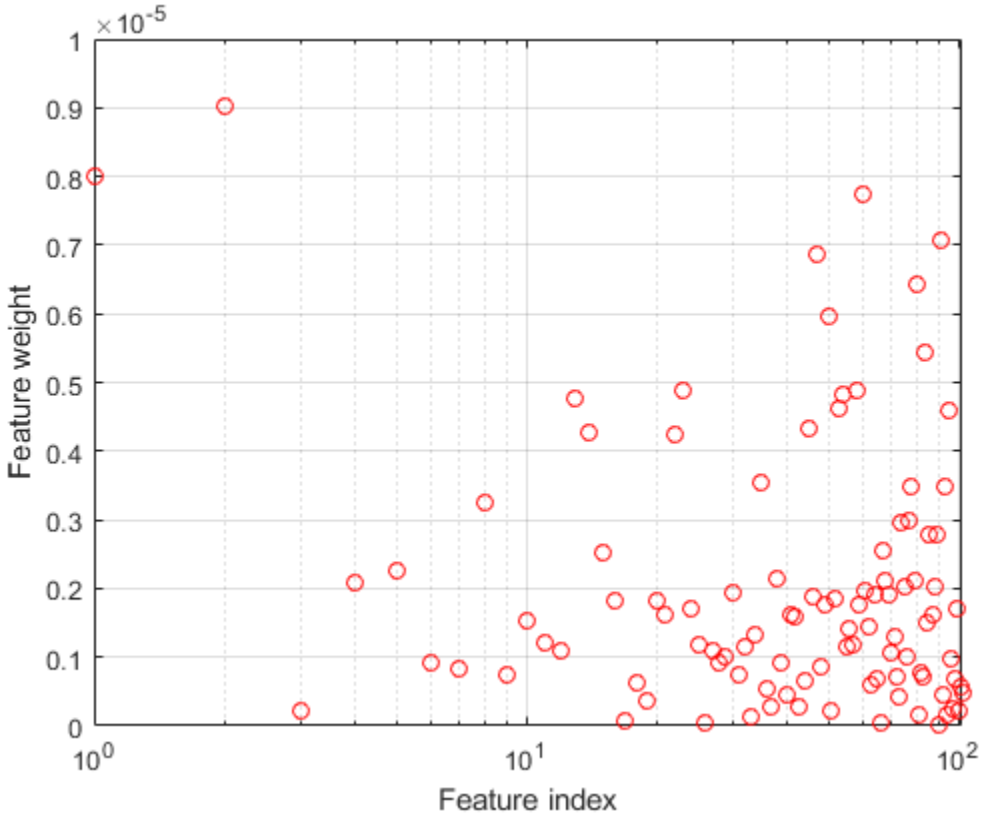
```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
              'Solver','lbfgs');
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
```

```
|=====================================================================================================
|         0 |  9.519258e-03 |   1.494e-02 |   0.000e+00 |          |   4.015e+01 |   0.000e+00 |
|         1 | -3.093574e-01 |   7.186e-03 |   4.018e+00 |    OK    |   8.956e+01 |   1.000e+00 |
|         2 | -4.809455e-01 |   4.444e-03 |   7.123e+00 |    OK    |   9.943e+01 |   1.000e+00 |
|         3 | -4.938877e-01 |   3.544e-03 |   1.464e+00 |    OK    |   9.366e+01 |   1.000e+00 |
|         4 | -4.964759e-01 |   2.901e-03 |   6.084e-01 |    OK    |   1.554e+02 |   1.000e+00 |
|         5 | -4.972077e-01 |   1.323e-03 |   6.129e-01 |    OK    |   1.195e+02 |   5.000e-01 |
|         6 | -4.974743e-01 |   1.569e-04 |   2.155e-01 |    OK    |   1.003e+02 |   1.000e+00 |
|         7 | -4.974868e-01 |   3.844e-05 |   4.161e-02 |    OK    |   9.835e+01 |   1.000e+00 |
|         8 | -4.974874e-01 |   1.417e-05 |   1.073e-02 |    OK    |   1.043e+02 |   1.000e+00 |
|         9 | -4.974874e-01 |   4.893e-06 |   1.781e-03 |    OK    |   1.530e+02 |   1.000e+00 |
|        10 | -4.974874e-01 |   9.404e-08 |   8.947e-04 |    OK    |   1.670e+02 |   1.000e+00 |

         Infinity norm of the final gradient = 9.404e-08
             Two norm of the final step     = 8.947e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 9.404e-08, TolFun = 1.000e-06
EXIT: Local minimum found.
```

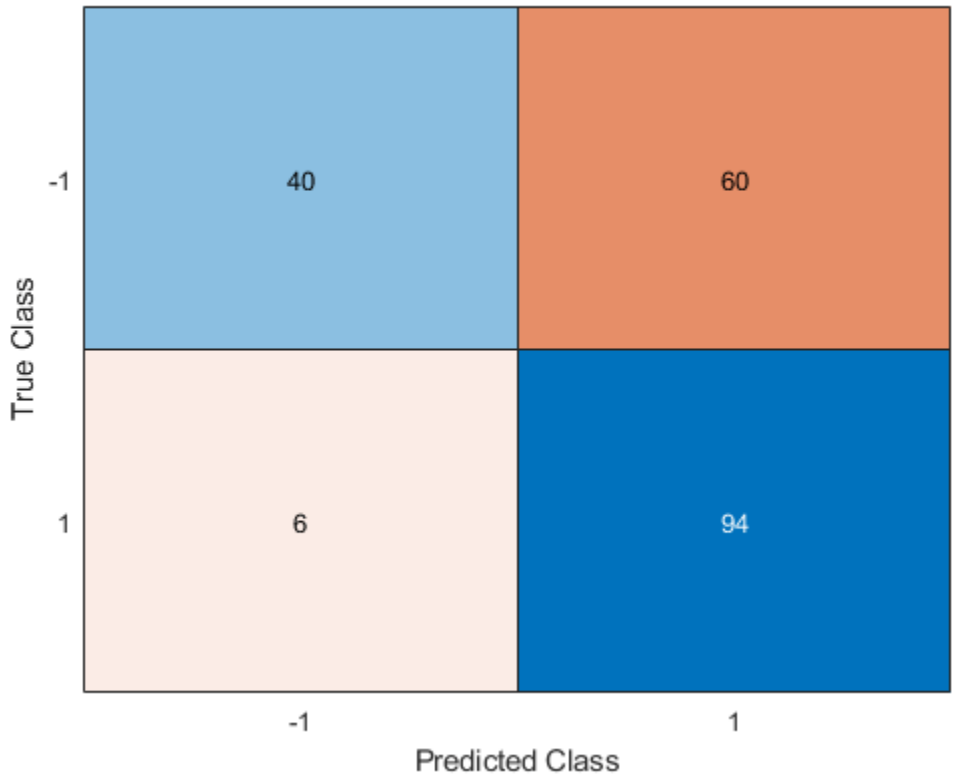
Plot the feature weights. The weights of the irrelevant features should be very close to zero.

```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



Predict the classes using the NCA model and compute the confusion matrix.

```
ypred = predict(ncaMdl,X);
confusionchart(y,ypred)
```



Confusion matrix shows that 40 of the data that are in class –1 are predicted as belonging to class –1. 60 of the data from class –1 are predicted to be in class 1. Similarly, 94 of the data from class 1 are predicted to be from class 1 and 6 of them are predicted to be from class –1. The prediction accuracy for class –1 is not good.

All weights are very close to zero, which indicates that the value of λ used in training the model is too large. When $\lambda \to \infty$, all features weights approach to zero. Hence, it is important to tune the regularization parameter in most cases to detect the relevant features.

Use five-fold cross-validation to tune λ for feature selection by using `fscnca`. Tuning λ means finding the λ value that will produce the minimum classification loss. To tune λ using cross-validation:

1. Partition the data into five folds. For each fold, `cvpartition` assigns four-fifths of the data as a training set and one-fifth of the data as a test set. Again for each fold, `cvpartition` creates a stratified partition, where each partition has roughly the same proportion of classes.

```
cvp = cvpartition(y,'kfold',5);
numtestsets = cvp.NumTestSets;
lambdavalues = linspace(0,2,20)/length(y);
lossvalues = zeros(length(lambdavalues),numtestsets);
```

2. Train the neighborhood component analysis (nca) model for each λ value using the training set in each fold.

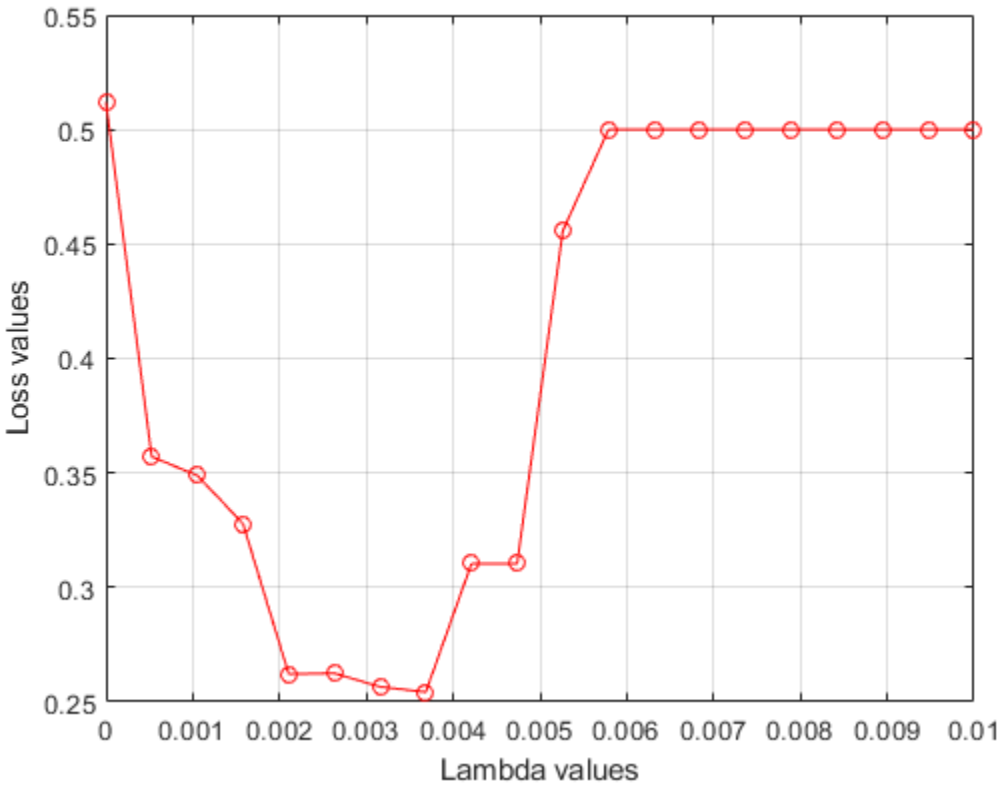
3. Compute the classification loss for the corresponding test set in the fold using the nca model. Record the loss value.

4. Repeat this process for all λ values.

```
for i = 1:length(lambdavalues)
    for k = 1:numtestsets

        % Extract the training set from the partition object
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);

        % Extract the test set from the partition object
        Xtest  = X(cvp.test(k),:);
        ytest  = y(cvp.test(k),:);

        % Train an NCA model for classification using the training set
        ncaMdl = fscnca(Xtrain,ytrain,'FitMethod','exact', ...
            'Solver','lbfgs','Lambda',lambdavalues(i));

        % Compute the classification loss for the test set using the NCA
        % model
        lossvalues(i,k) = loss(ncaMdl,Xtest,ytest, ...
            'LossFunction','quadratic');

    end
end
```

Plot the average loss values of the folds versus the λ values. If the λ value that corresponds to the minimum loss falls on the boundary of the tested λ values, the range of λ values should be reconsidered.

```
figure
plot(lambdavalues,mean(lossvalues,2),'ro-')
xlabel('Lambda values')
ylabel('Loss values')
grid on
```



Find the λ value that corresponds to the minimum average loss.

```
[~,idx] = min(mean(lossvalues,2)); % Find the index
bestlambda = lambdavalues(idx) % Find the best lambda value
```

```
bestlambda =

    0.0037
```

Fit the NCA model to all of the data using the best λ value. Use the LBFGS solver and display the convergence information.

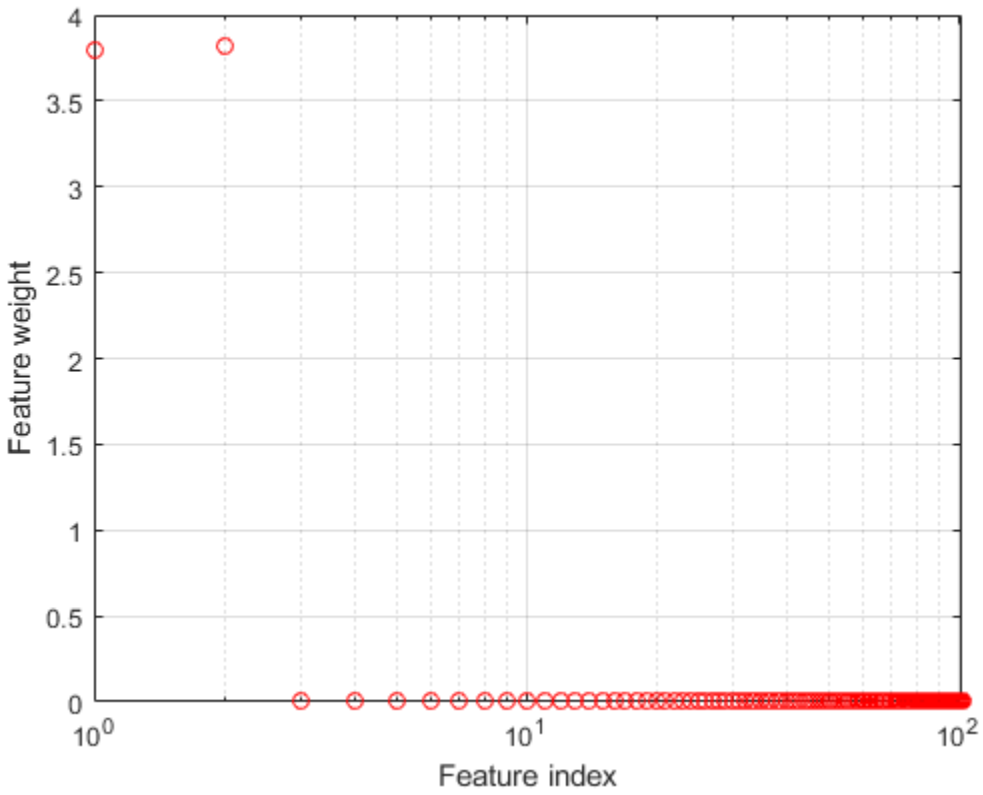
```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
        'Solver','lbfgs','Lambda',bestlambda);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 | -1.246913e-01 |   1.231e-02 |   0.000e+00 |        |   4.873e+01 |   0.000e+00 |
|        1 | -3.411330e-01 |   5.717e-03 |   3.618e+00 |    OK  |   1.068e+02 |   1.000e+00 |
|        2 | -5.226111e-01 |   3.763e-02 |   8.252e+00 |    OK  |   7.825e+01 |   1.000e+00 |
|        3 | -5.817731e-01 |   8.496e-03 |   2.340e+00 |    OK  |   5.591e+01 |   5.000e-01 |
|        4 | -6.132632e-01 |   6.863e-03 |   2.526e+00 |    OK  |   8.228e+01 |   1.000e+00 |
```

```
|        5 | -6.135264e-01 |   9.373e-03 |   7.341e-01 |    OK  |   3.244e+01 |   1.000e+00 |   Y
|        6 | -6.147894e-01 |   1.182e-03 |   2.933e-01 |    OK  |   2.447e+01 |   1.000e+00 |   Y
|        7 | -6.148714e-01 |   6.392e-04 |   6.688e-02 |    OK  |   3.195e+01 |   1.000e+00 |   Y
|        8 | -6.149524e-01 |   6.521e-04 |   9.934e-02 |    OK  |   1.236e+02 |   1.000e+00 |   Y
|        9 | -6.149972e-01 |   1.154e-04 |   1.191e-01 |    OK  |   1.171e+02 |   1.000e+00 |   Y
|       10 | -6.149990e-01 |   2.922e-05 |   1.983e-02 |    OK  |   7.365e+01 |   1.000e+00 |   Y
|       11 | -6.149993e-01 |   1.556e-05 |   8.354e-03 |    OK  |   1.288e+02 |   1.000e+00 |   Y
|       12 | -6.149994e-01 |   1.147e-05 |   7.256e-03 |    OK  |   2.332e+02 |   1.000e+00 |   Y
|       13 | -6.149995e-01 |   1.040e-05 |   6.781e-03 |    OK  |   2.287e+02 |   1.000e+00 |   Y
|       14 | -6.149996e-01 |   9.015e-06 |   6.265e-03 |    OK  |   9.974e+01 |   1.000e+00 |   Y
|       15 | -6.149996e-01 |   7.763e-06 |   5.206e-03 |    OK  |   2.919e+02 |   1.000e+00 |   Y
|       16 | -6.149997e-01 |   8.374e-06 |   1.679e-02 |    OK  |   6.878e+02 |   1.000e+00 |   Y
|       17 | -6.149997e-01 |   9.387e-06 |   9.542e-03 |    OK  |   1.284e+02 |   5.000e-01 |   Y
|       18 | -6.149997e-01 |   3.250e-06 |   5.114e-03 |    OK  |   1.225e+02 |   1.000e+00 |   Y
|       19 | -6.149997e-01 |   1.574e-06 |   1.275e-03 |    OK  |   1.808e+02 |   1.000e+00 |   Y

|=============================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=============================================================================================================
|       20 | -6.149997e-01 |   5.764e-07 |   6.765e-04 |    OK  |   2.905e+02 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 5.764e-07
             Two norm of the final step     = 6.765e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 5.764e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the feature weights.

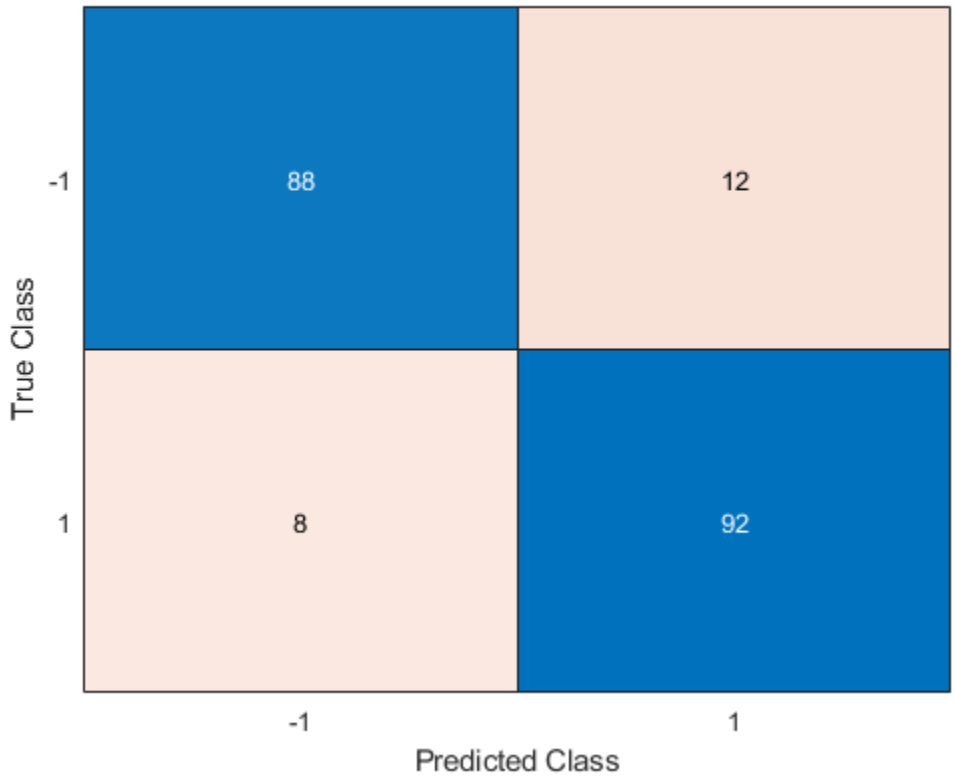
```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



`fscnca` correctly figures out that the first two features are relevant and that the rest are not. The first two features are not individually informative, but when taken together result in an accurate classification model.

Predict the classes using the new model and compute the accuracy.

```
ypred = predict(ncaMdl,X);
confusionchart(y,ypred)
```



Confusion matrix shows that prediction accuracy for class –1 has improved. 88 of the data from class –1 are predicted to be from –1, and 12 of them are predicted to be from class 1. 92 of the data from class 1 are predicted to be from class 1 and 8 of them are predicted to be from class –1.

References

[1] Yang, W., K. Wang, W. Zuo. "Neighborhood Component Feature Selection for High-Dimensional Data." *Journal of Computers*. Vol. 7, Number 1, January, 2012.

See Also

`FeatureSelectionNCAClassification` | `fscnca` | `predict` | `refit`

Introduced in R2016b

loss

Class: `FeatureSelectionNCARegression`

Evaluate accuracy of learned feature weights on test data

Syntax

```
err = loss(mdl,X,Y)
err = loss(mdl,X,Y,Name,Value)
```

Description

`err = loss(mdl,X,Y)` returns the mean squared error as the measure of accuracy in `err`, for the model `mdl`, predictor values in `X`, and response values in `Y`.

`err = loss(mdl,X,Y,Name,Value)` returns the measure of accuracy, `err`, with the additional option specified by the `Name,Value` pair argument.

Input Arguments

`mdl` — Neighborhood component analysis model for regression

`FeatureSelectionNCARegression` object

Neighborhood component analysis model for regression, specified as a `FeatureSelectionNCARegression` object.

`X` — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

`Y` — Response values

numeric real vector of length *n*

Response values, specified as a numeric real vector of length *n*, where *n* is the number of observations.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFunction` — Loss function type

`'mse'` (default) | `'mad'`

Loss function type, specified as a comma-separated pair consisting of `'Loss Function'` and one of the following.

Loss Function Type	Description
`'mse'`	Mean squared error
`'mad'`	Mean absolute deviation

Example: `'LossFunction','mse'`

Output Arguments

`err` — Smaller-the-better accuracy measure for learned feature weights

scalar value

Smaller-the-better accuracy measure for learned feature weights, returned as a scalar value. You can specify the measure of accuracy using the `LossFunction` name-value pair argument.

Examples

Tune NCA Model for Regression Using `loss` and `predict`

Load the sample data.

Download the housing data [1], from the UCI Machine Learning Repository [2]. The dataset has 506 observations. The first 13 columns contain the predictor values and the last column contains the response values. The goal is to predict the median value of owner-occupied homes in suburban Boston as a function of 13 predictors.

Load the data and define the response vector and the predictor matrix.

```
load('housing.data');
X = housing(:,1:13);
y = housing(:,end);
```

Divide the data into training and test sets using the 4th predictor as the grouping variable for a stratified partitioning. This ensures that each partition includes similar amount of observations from each group.

```
rng(1) % For reproducibility
cvp = cvpartition(X(:,4),'Holdout',56);
Xtrain = X(cvp.training,:);
ytrain = y(cvp.training,:);
Xtest  = X(cvp.test,:);
ytest  = y(cvp.test,:);
```

`cvpartition` randomly assigns 56 observations into a test set and the rest of the data into a training set.

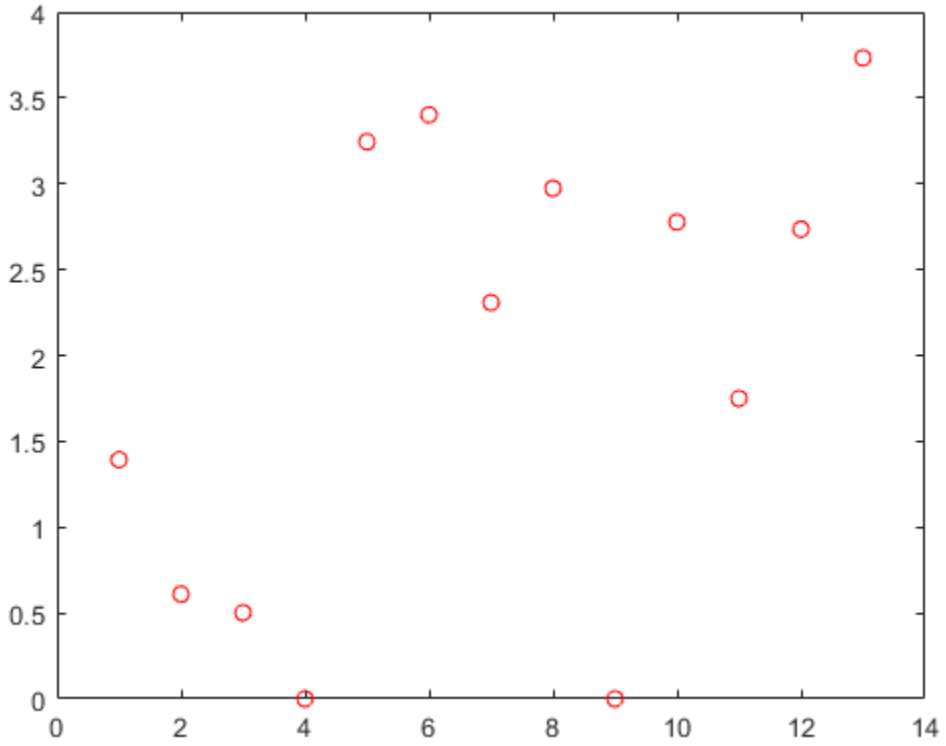
Perform Feature Selection Using Default Settings

Perform feature selection using NCA model for regression. Standardize the predictor values.

```
nca = fsrnca(Xtrain,ytrain,'Standardize',1);
```

Plot the feature weights.

```
figure()
plot(nca.FeatureWeights,'ro')
```



The weights of irrelevant features are expected to approach zero. `fsrnca` identifies two features as irrelevant.

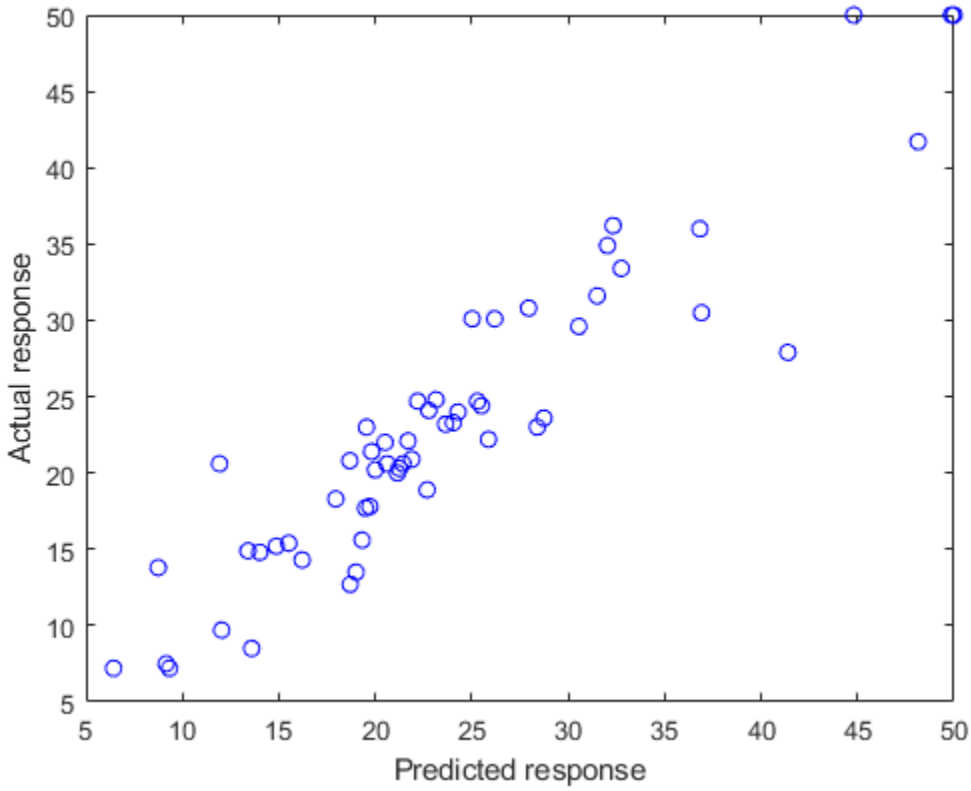
Compute the regression loss.

```
L = loss(nca,Xtest,ytest,'LossFunction','mad')
```

```
L = 2.5394
```

Compute the predicted response values for the test set and plot them versus the actual response.

```
ypred = predict(nca,Xtest);
figure()
plot(ypred,ytest,'bo')
xlabel('Predicted response')
ylabel('Actual response')
```



A perfect fit versus the actual values forms a 45 degree straight line. In this plot, the predicted and actual response values seem to be scattered around this line. Tuning λ (regularization parameter) value usually helps improve the performance.

Tune the regularization parameter using 10-fold cross-validation

Tuning λ means finding the λ value that will produce the minimum regression loss. Here are the steps for tuning λ using 10-fold cross-validation:

1. First partition the data into 10 folds. For each fold, `cvpartition` assigns 1/10th of the data as a training set, and 9/10th of the data as a test set.

```
n = length(ytrain);
cvp = cvpartition(Xtrain(:,4),'kfold',10);
numvalidsets = cvp.NumTestSets;
```

Assign the λ values for the search. Create an array to store the loss values.

```
lambdavals = linspace(0,2,30)*std(ytrain)/n;
lossvals = zeros(length(lambdavals),numvalidsets);
```

2. Train the neighborhood component analysis (nca) model for each λ value using the training set in each fold.

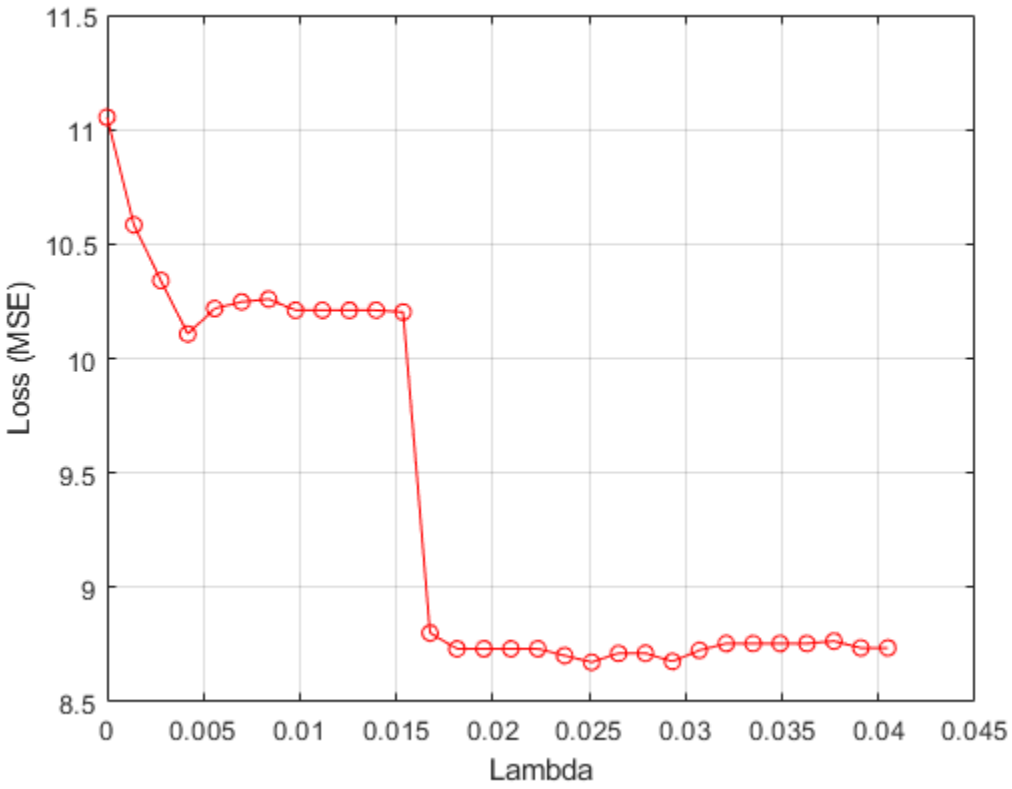
3. Fit a Gaussian process regression (gpr) model using the selected features. Next, compute the regression loss for the corresponding test set in the fold using the gpr model. Record the loss value.

4. Repeat this for each λ value and each fold.

```
for i = 1:length(lambdavals)
   for k = 1:numvalidsets
       X = Xtrain(cvp.training(k),:);
       y = ytrain(cvp.training(k),:);
       Xvalid  = Xtrain(cvp.test(k),:);
       yvalid  = ytrain(cvp.test(k),:);

       nca = fsrnca(X,y,'FitMethod','exact',...
            'Lambda',lambdavals(i),...
            'Standardize',1,'LossFunction','mad');

       % Select features using the feature weights and a relative
       % threshold.
       tol    = 1e-3;
       selidx = nca.FeatureWeights > tol*max(1,max(nca.FeatureWeights));

       % Fit a non-ARD GPR model using selected features.
       gpr = fitrgp(X(:,selidx),y,'Standardize',1,...
             'KernelFunction','squaredexponential','Verbose',0);


       lossvals(i,k) = loss(gpr,Xvalid(:,selidx),yvalid);

   end
end
```

Compute the average loss obtained from the folds for each λ value. Plot the mean loss versus the λ values.

```
meanloss = mean(lossvals,2);
figure;
plot(lambdavals,meanloss,'ro-');
xlabel('Lambda');
ylabel('Loss (MSE)');
grid on;
```



Find the λ value that produces the minimum loss value.

```
[~,idx] = min(meanloss);
bestlambda = lambdavals(idx)
```

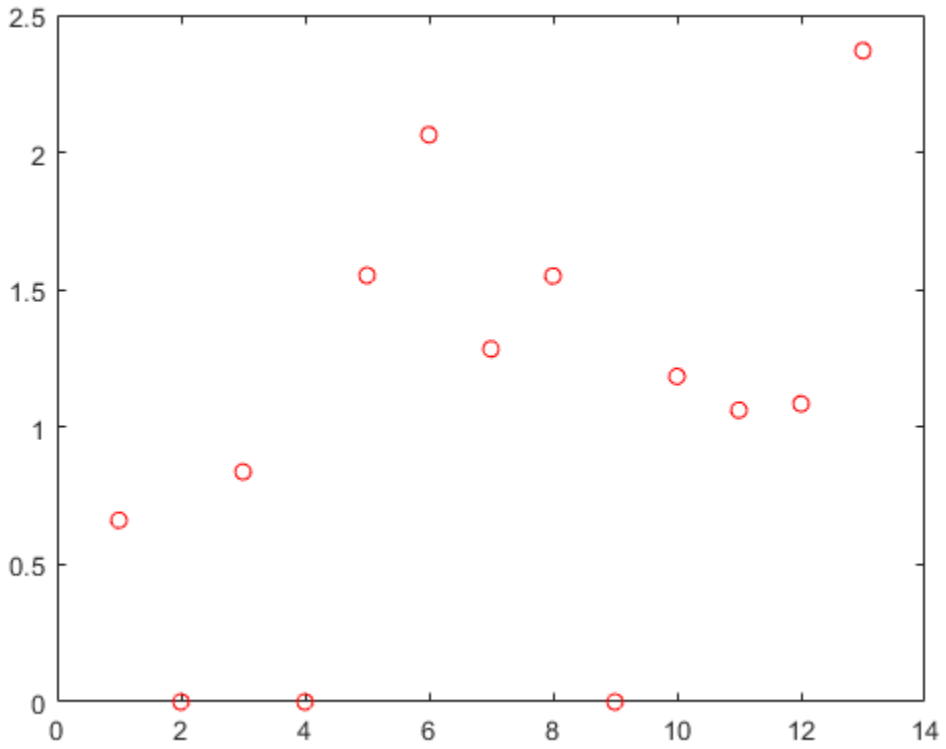
```
bestlambda = 0.0251
```

Perform feature selection for regression using the best λ value. Standardize the predictor values.

```
nca2 = fsrnca(Xtrain,ytrain,'Standardize',1,'Lambda',bestlambda,...
    'LossFunction','mad');
```

Plot the feature weights.

```
figure()
plot(nca.FeatureWeights,'ro')
```



Compute the loss using the new nca model on the test data, which is not used to select the features.

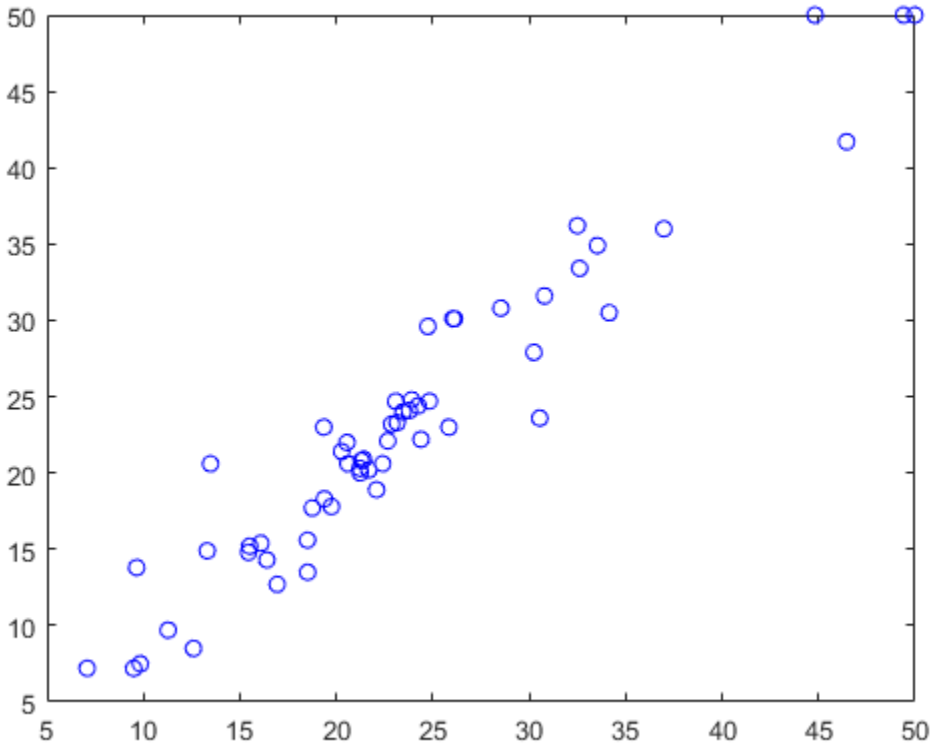
```
L2 = loss(nca2,Xtest,ytest,'LossFunction','mad')
```

```
L2 = 2.0560
```

Tuning the regularization parameter helps identify the relevant features and reduces the loss.

Plot the predicted versus the actual response values in the test set.

```
ypred = predict(nca2,Xtest);
figure;
plot(ypred,ytest,'bo');
```



The predicted response values seem to be closer to the actual values as well.

References

[1] Harrison, D. and D.L., Rubinfeld. "Hedonic prices and the demand for clean air." J. Environ. Economics & Management. Vol.5, 1978, pp. 81-102.

[2] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. https://archive.ics.uci.edu/ml.

See Also

`FeatureSelectionNCARegression` | `fsrnca` | `predict` | `refit`

Introduced in R2016b

loss

Package:

Regression loss for generalized additive model (GAM)

Syntax

```
L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)
L = loss(Mdl,X,Y)
L = loss(___,Name,Value)
```

Description

`L = loss(Mdl,Tbl,ResponseVarName)` returns the regression loss (`L`), a scalar representing how well the generalized additive model `Mdl` predicts the predictor data in `Tbl` compared to the true response values in `Tbl.ResponseVarName`.

The interpretation of `L` depends on the loss function (`'LossFun'`) and weighting scheme (`'Weights'`). In general, better models yield smaller loss values. The default `'LossFun'` value is `'mse'` (mean squared error).

`L = loss(Mdl,Tbl,Y)` uses the predictor data in table `Tbl` and the true response values in `Y`.

`L = loss(Mdl,X,Y)` uses the predictor data in matrix `X` and the true response values in `Y`.

`L = loss(___,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify the loss function and the observation weights.

Examples

Determine Test Sample Regression Loss

Determine the test sample regression loss (mean squared error) of a generalized additive model. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Randomly partition observations into a training set and a test set. Specify a 10% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(size(tbl,1),'HoldOut',0.10);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Train a univariate GAM that contains the linear terms for the predictors in `tbl`.

```
Mdl = fitrgam(tbl(trainInds,:),"Systolic");
```

Determine how well the algorithm generalizes by estimating the test sample regression loss. By default, the `loss` function of `RegressionGAM` estimates the mean squared error.

```
L = loss(Mdl,tbl(testInds,:))
```

```
L = 35.7540
```

Compare GAMs by Examining Regression Loss

Train a generalized additive model (GAM) that contains both linear and interaction terms for predictors, and estimate the regression loss (mean squared error, MSE) with and without interaction terms for the training data and test data. Specify whether to include interaction terms when estimating the regression loss.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 10 observations for the new test data set.

```
rng('default') % For reproducibility
n = size(X,1);
newInds = randsample(n,10);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a generalized additive model that contains all the available linear and interaction terms in `X`.

```
Mdl = fitrgam(X(inds,:),Y(inds),'Interactions','all');
```

`Mdl` is a `RegressionGAM` model object.

Compute the resubstitution MSEs (that is, the in-sample MSEs) both with and without interaction terms in `Mdl`. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
resubl = resubLoss(Mdl)
```

```
resubl = 0.0292
```

```
resubl_nointeraction = resubLoss(Mdl,'IncludeInteractions',false)
```

```
resubl_nointeraction = 4.7330
```

Compute the regression MSEs both with and without interaction terms for the test data set. Use a memory-efficient model object for the computation.

```
CMdl = compact(Mdl);
```

`CMdl` is a `CompactRegressionGAM` model object.

```
l = loss(CMdl,XNew,YNew)
```

```
l = 12.8604
```

```
l_nointeraction = loss(CMdl,XNew,YNew,'IncludeInteractions',false)
```

```
l_nointeraction = 15.6741
```

Including interaction terms achieves a smaller error for the training data set and test data set.

Input Arguments

`Mdl` — Generalized additive model

`RegressionGAM` model object | `CompactRegressionGAM` model object

Generalized additive model, specified as a `RegressionGAM` or `CompactRegressionGAM` model object.

- If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table (`Tbl`).
- If you trained `Mdl` using sample data contained in a matrix, then the input data for `loss` must also be in a matrix (`X`).

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`Tbl` must contain all of the predictors used to train `Mdl`. Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights.

- The response variable must be a numeric vector. If the response variable in `Tbl` has the same name as the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.
- The weight values must be a numeric vector. You must specify the observation weights in `Tbl` by using `'Weights'`.

If you trained `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`Y` — Response data
numeric column vector

Response data, specified as a numeric column vector. Each entry in `Y` is the response to the data in the corresponding row of `X` or `Tbl`.

Data Types: `single` | `double`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

If you trained `Mdl` using sample data contained in a matrix, then the input data for `loss` must also be in a matrix.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'IncludeInteractions',false,'Weights',w` specifies to exclude interaction terms from the model and to use the observation weights `w`.

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `'IncludeInteractions'` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`LossFun` — Loss function
`'mse'` (default) | function handle

Loss function, specified as `'mse'` or a function handle.

- `'mse'` — Weighted mean squared error.

- Function handle — To specify a custom loss function, use a function handle. The function must have this form:

  ```
  lossval = lossfun(Y,YFit,W)
  ```

  - The output argument `lossval` is a floating-point scalar.
  - You specify the function name (*`lossfun`*).
  - `Y` is a length *n* numeric vector of observed responses, where *n* is the number of observations in `Tbl` or `X`.
  - `YFit` is a length *n* numeric vector of corresponding predicted responses.
  - `W` is an *n*-by-1 numeric vector of observation weights.

Example: `'LossFun',@lossfun`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | vector of scalar values | name of variable in `Tbl`

Observation weights, specified as a vector of scalar values or the name of a variable in `Tbl`. The software weights the observations in each row of `X` or `Tbl` with the corresponding value in `Weights`. The size of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`.

`loss` normalizes the values of `Weights` to sum to 1.

Data Types: `single` | `double` | `char` | `string`

More About

Weighted Mean Squared Error

The weighted mean squared error measures the predictive inaccuracy of regression models. When you compare the same type of loss among many models, a lower error indicates a better predictive model.

The weighted mean squared error is calculated as follows:

$$\text{mse} = \frac{\sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j\right)^2}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of rows of data.
- x_j is the jth row of data.
- y_j is the true response to x_j.

- $f(x_j)$ is the response prediction of the model `Mdl` to x_j.
- w is the vector of observation weights.

See Also

`predict` | `resubLoss`

Topics

"Train Generalized Additive Model for Regression" on page 12-91

Introduced in R2021a

loss

Class: `RegressionLinear`

Regression loss for linear regression models

Syntax

```
L = loss(Mdl,X,Y)

L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(Mdl,X,Y)` returns the mean squared error (MSE) for the linear regression model `Mdl` using predictor data in `X` and corresponding responses in `Y`. `L` contains an MSE for each regularization strength in `Mdl`.

`L = loss(Mdl,Tbl,ResponseVarName)` returns the MSE for the predictor data in `Tbl` and the true responses in `Tbl.ResponseVarName`.

`L = loss(Mdl,Tbl,Y)` returns the MSE for the predictor data in table `Tbl` and the true responses in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, specify that columns in the predictor data correspond to observations or specify the regression loss function.

Input Arguments

`Mdl` — Linear regression model

`RegressionLinear` model object

Linear regression model, specified as a `RegressionLinear` model object. You can create a `RegressionLinear` model object using `fitrlinear`.

`X` — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix. This orientation of `X` indicates that rows correspond to individual observations, and columns correspond to individual predictor variables.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

`Y` — Response data
numeric vector

Response data, specified as an *n*-dimensional numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

`Tbl` — Sample data
table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function
`'mse'` (default) | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding value. Also, in the table, $f(x) = x\beta + b$.
  - β is a vector of p coefficients.
  - x is an observation from p predictor variables.
  - b is the scalar bias.

Value	Description
`'epsiloninsensitive'`	Epsilon-insensitive loss: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$
`'mse'`	MSE: $\ell[y, f(x)] = [y - f(x)]^2$

`'epsiloninsensitive'` is appropriate for SVM learners only.

- Specify your own function using function handle notation.

  Let *n* be the number of observations in `X`. Your function must have this signature

  `lossvalue = lossfun(Y,Yhat,W)`

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `Y` is an *n*-dimensional vector of observed responses. `loss` passes the input argument `Y` in for `Y`.
  - `Yhat` is an *n*-dimensional vector of predicted responses, which is similar to the output of `predict`.
  - `W` is an *n*-by-1 numeric vector of observation weights.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If you specify `Weights` as a numeric vector, then the size of `Weights` must be equal to the number of observations in `X` or `Tbl`.
- If you specify `Weights` as the name of a variable in `Tbl`, then the name must be a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, `loss` computes the weighted regression loss and normalizes `Weights` to sum to 1.

Data Types: `double` | `single`

Output Arguments

L — Regression losses

numeric scalar | numeric row vector

Regression losses, returned as a numeric scalar or row vector. The interpretation of `L` depends on `Weights` and `LossFun`.

`L` is the same size as `Mdl.Lambda`. `L(j)` is the regression loss of the linear regression model trained using the regularization strength `Mdl.Lambda(j)`.

Note If `Mdl.FittedLoss` is `'mse'`, then the loss term in the objective function is half of the MSE. `loss` returns the MSE by default. Therefore, if you use `loss` to check the resubstitution (training) error, then there is a discrepancy between the MSE and optimization results that `fitrlinear` returns.

Examples

Estimate Test-Sample Mean Squared Error

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model. Reserve 30% of the observations as a holdout sample.

```
CVMdl = fitrlinear(X,Y,'Holdout',0.3);
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0066
               Lambda: 1.4286e-04
              Learner: 'svm'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `RegressionLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Estimate the training- and test-sample MSE.

```
mseTrain = loss(Mdl,X(trainIdx,:),Y(trainIdx))
```

```
mseTrain = 0.1496
```

```
mseTest = loss(Mdl,X(testIdx,:),Y(testIdx))
```

```
mseTest = 0.1798
```

Because there is one regularization strength in `Mdl`, `mseTrain` and `mseTest` are numeric scalars.

Specify Custom Regression Loss

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
X = X'; % Put observations in columns for faster training
```

Train a linear regression model. Reserve 30% of the observations as a holdout sample.

```
CVMdl = fitrlinear(X,Y,'Holdout',0.3,'ObservationsIn','columns');
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0066
               Lambda: 1.4286e-04
              Learner: 'svm'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `RegressionLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Create an anonymous function that measures Huber loss ($\delta = 1$), that is,

$$L = \frac{1}{\sum w_j} \sum_{j=1}^{n} w_j \ell_j ,$$

where

$$\ell_j = \begin{cases} 0.5\hat{e}_j^2; & |\hat{e}_j| \le 1 \\ |\hat{e}_j| - 0.5; & |\hat{e}_j| > 1 \end{cases}.$$

$\hat{e}_j$ is the residual for observation j. Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `'LossFun'` name-value pair argument.

```
huberloss = @(Y,Yhat,W)sum(W.*((0.5*(abs(Y-Yhat)<=1).*(Y-Yhat).^2) + ...
    ((abs(Y-Yhat)>1).*abs(Y-Yhat)-0.5)))/sum(W);
```

Estimate the training set and test set regression loss using the Huber loss function.

```
eTrain = loss(Mdl,X(:,trainIdx),Y(trainIdx),'LossFun',huberloss,...
    'ObservationsIn','columns')
```

```
eTrain = -0.4186
```

```
eTest = loss(Mdl,X(:,testIdx),Y(testIdx),'LossFun',huberloss,...
    'ObservationsIn','columns')
```

```
eTest = -0.4010
```

Find Good Lasso Penalty Using Regression Loss

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e .$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-4} through 10^{-1}.

```
Lambda = logspace(-4,-1,15);
```

Hold out 30% of the data for testing. Identify the test-sample indices.

```
cvp = cvpartition(numel(Y),'Holdout',0.30);
idxTest = test(cvp);
```

Train a linear regression model using lasso penalties with the strengths in `Lambda`. Specify the regularization strengths, optimizing the objective function using SpaRSA, and the data partition. To increase execution speed, transpose the predictor data and specify that the observations are in columns.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Solver','sparsa','Regularization','lasso','CVPartition',cvp);
Mdl1 = CVMdl.Trained{1};
numel(Mdl1.Lambda)
```

```
ans = 15
```

`Mdl1` is a `RegressionLinear` model. Because `Lambda` is a 15-dimensional vector of regularization strengths, you can think of `Mdl1` as 15 trained models, one for each regularization strength.

Estimate the test-sample mean squared error for each regularized model.

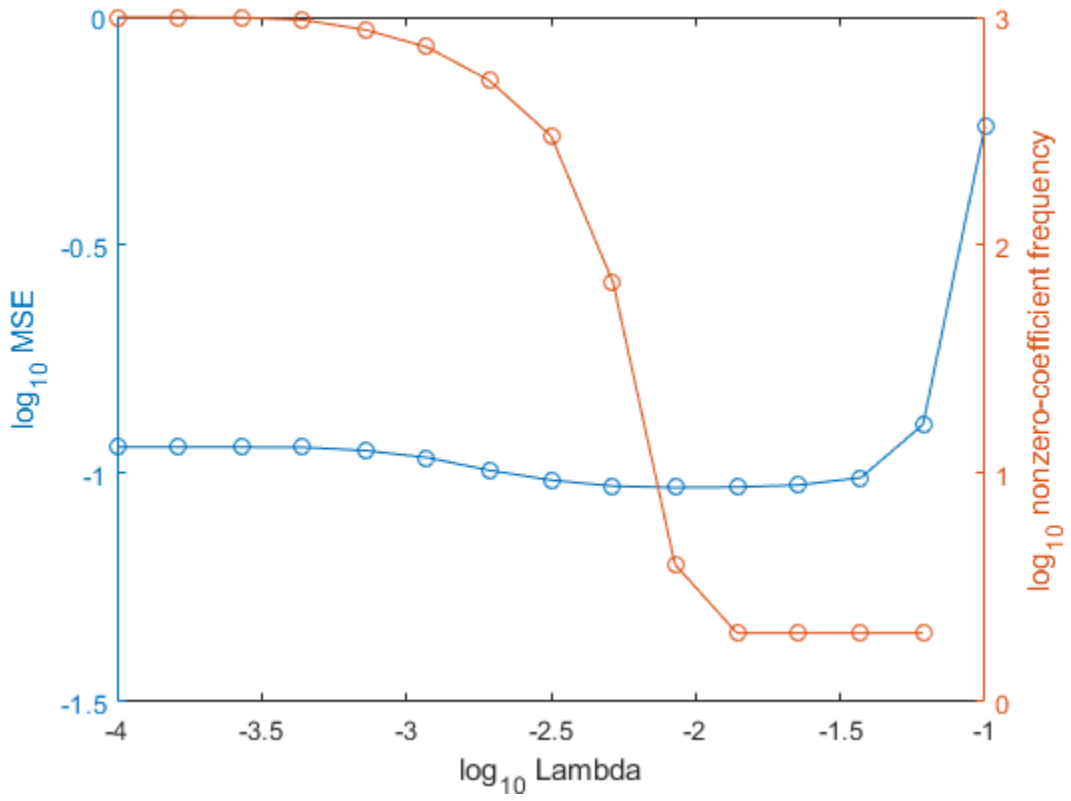
```
mse = loss(Mdl1,X(:,idxTest),Y(idxTest),'ObservationsIn','columns');
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. Retrain the model using the entire data set and all options used previously, except the data-partition specification. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Select the index or indices of `Lambda` that balance minimal classification error and predictor-variable sparsity (for example, `Lambda(11)`).

```
idx = 11;
MdlFinal = selectModels(Mdl,idx);
```

`MdlFinal` is a trained `RegressionLinear` model object that uses `Lambda(11)` as a regularization strength.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `loss` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`RegressionLinear` | `fitrlinear` | `predict`

Introduced in R2016a

lowerparams

Lower Pareto tail parameters

Syntax

```
params = lowerparams(pd)
```

Description

`params = lowerparams(pd)` returns the two-element vector `params`, which includes the shape and scale parameters of the generalized Pareto distribution (GPD) in the lower tail of `pd`.

`lowerparams` does not return the location parameter of the GPD. The location parameter is the quantile value corresponding to the lower tail cumulative probability. Use the `boundary` function to return the location parameter.

Examples

Parameters of Lower Pareto Tail

Generate a sample data set and fit a piecewise distribution with Pareto tails to the data by using `paretotails`. Find the distribution parameters of the lower Pareto tail by using the object function `lowerparams`.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and GPDs for the lower and upper 10% of the data set.

```
pd = paretotails(x,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.33251    (0 < p < 0.1): lower tail, GPD(-0.0063504,0.567017)
   -1.33251 < x < 1.80149  (0.1 < p < 0.9): interpolated empirical cdf
        1.80149 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.24874,3.00974)
```

Return the shape and scale parameters of the fitted GPD of the lower tail by using the `lowerparams` function.

```
params = lowerparams(pd)
```

```
params = 1×2

   -0.0064    0.5670
```

You can also get the lower Pareto tail parameters by using the `LowerParameters` property. Access the `LowerParameters` property by using dot notation.

```
pd.LowerParameters

ans = 1×2

   -0.0064    0.5670
```

The location parameter of the GPD is equal to the quantile value of the lower tail cumulative probability. Return the location parameter by using the `boundary` function.

```
[p,q] = boundary(pd)

p = 2×1

    0.1000
    0.9000


q = 2×1

   -1.3325
    1.8015
```

The values in `p` are the cumulative probabilities at the boundaries, and the values in `q` are the corresponding quantiles. `q(2)` is the location parameter of the GPD of the lower tail.

Use the `upperparams` function or the `UpperParameters` property to get the upper Pareto tail parameters.

Input Arguments

pd — Piecewise distribution with Pareto tails

`paretotails` object

Piecewise distribution with Pareto tails, specified as a `paretotails` object.

See Also

`boundary` | `gpfit` | `nsegments` | `paretotails` | `segment` | `upperparams`

Topics

“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

lt

Class: `qrandstream`

Less than relation for handles

Syntax

```
h1 < h2
```

Description

`h1 < h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise < result.

If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = lt(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`eq` | `ge` | `gt` | `le` | `ne` | `qrandstream`

lsline

Add least-squares line to scatter plot

Syntax

```
lsline
lsline(ax)
h = lsline( ___ )
```

Description

`lsline` superimposes a least-squares line on each scatter plot in the current axes.

`lsline` ignores data points that are connected with solid, dashed, or dash-dot lines (`'-'`, `'--'`, or `'.-'`) because it does not consider them to be scatter plots. To produce scatter plots, use the MATLAB `scatter` and `plot` functions.

`lsline(ax)` superimposes a least-squares line on the scatter plot in the axes specified by `ax` instead of the current axes (`gca`).

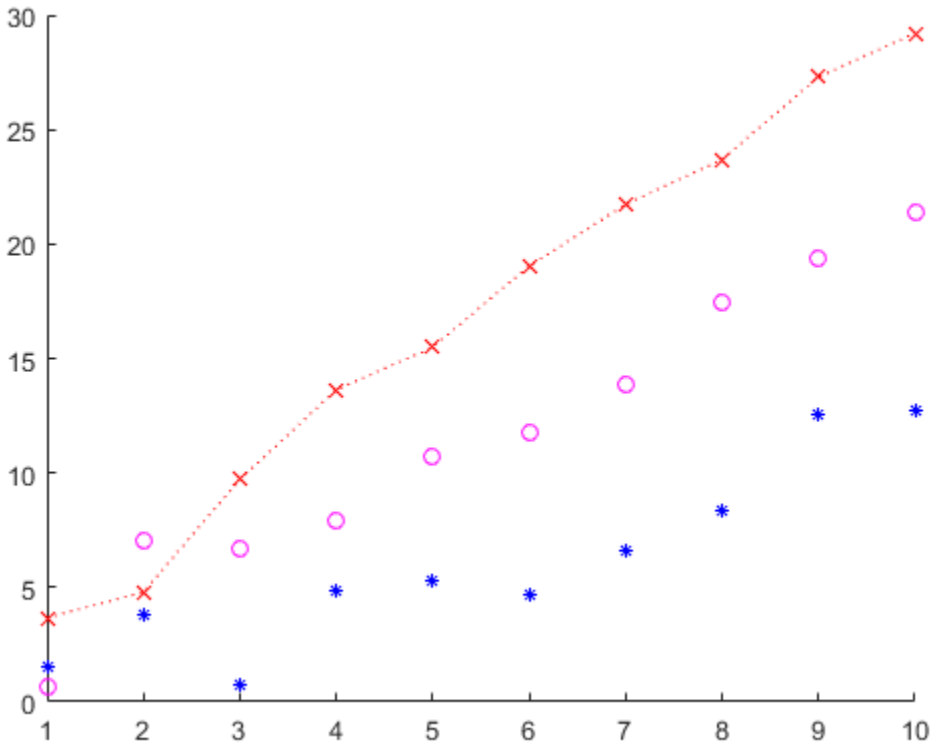
`h = lsline( ___ )` returns a column vector of least-squares line objects `h` using any of the previous syntaxes. Use `h` to modify the properties of a specific least-squares line after you create it. For a list of properties, see Line Properties.

Examples

Plot Least-Squares Line

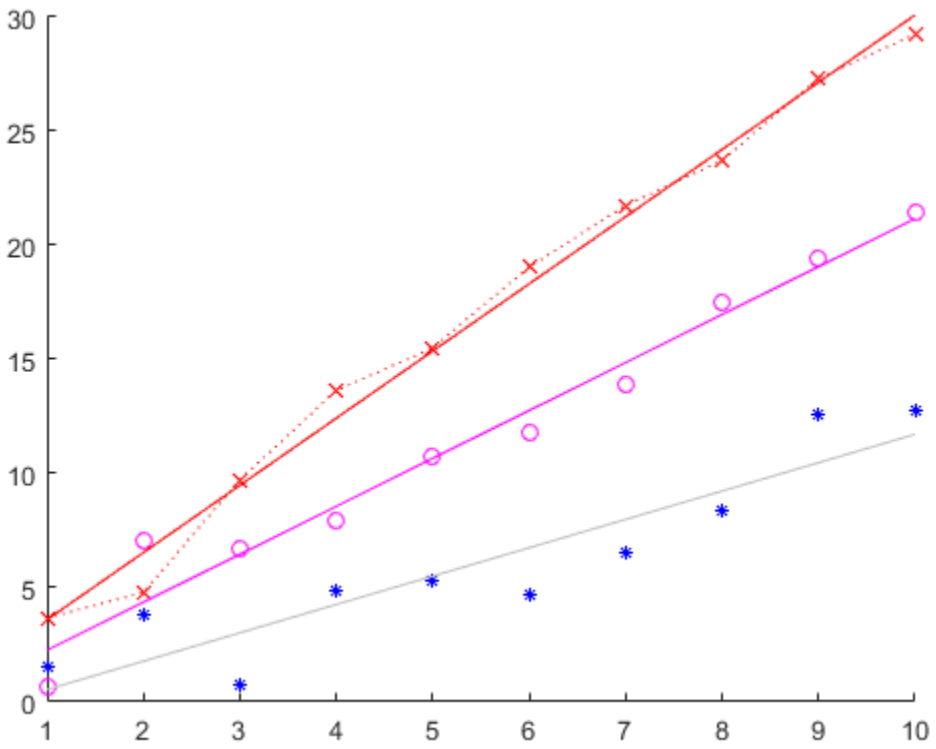
Generate three sets of sample data and plot each set on the same figure.

```
x = 1:10;
rng default;  % For reproducibility
figure;

y1 = x + randn(1,10);
scatter(x,y1,25,'b','*')
hold on

y2 = 2*x + randn(1,10);
plot(x,y2,'mo')

y3 = 3*x + randn(1,10);
plot(x,y3,'rx:')
```



Add a least-squares line for each set of sample data.

```
lsline
```



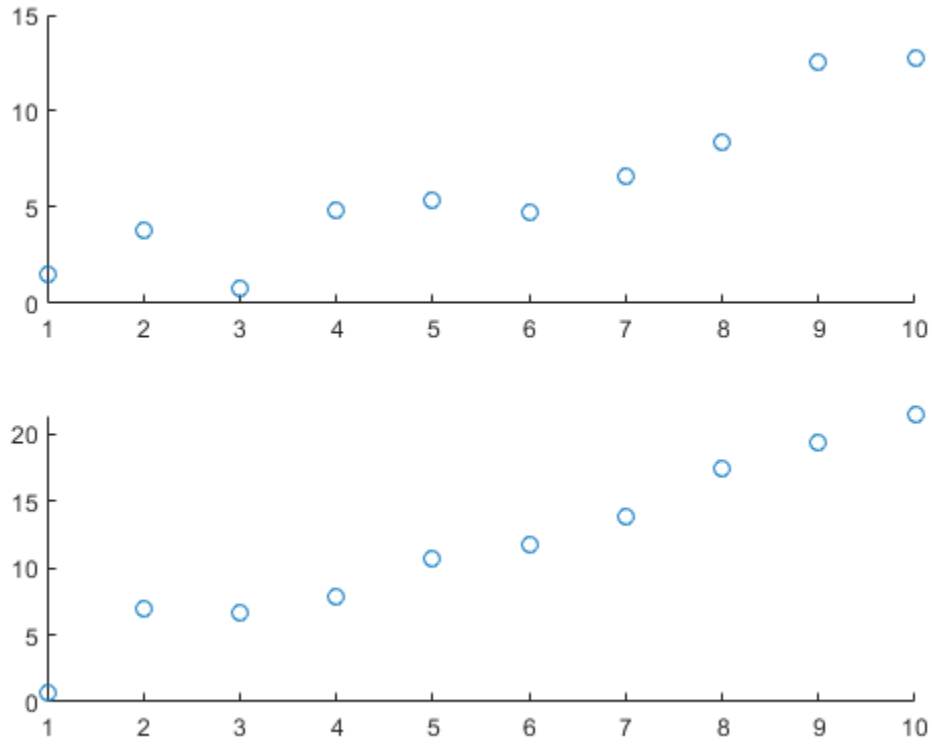
Specify Axes for Least-Squares and Reference Lines

Define the x-variable and two different y-variables to use for the plots.

```
rng default  % For reproducibility
x = 1:10;
y1 = x + randn(1,10);
y2 = 2*x + randn(1,10);
```

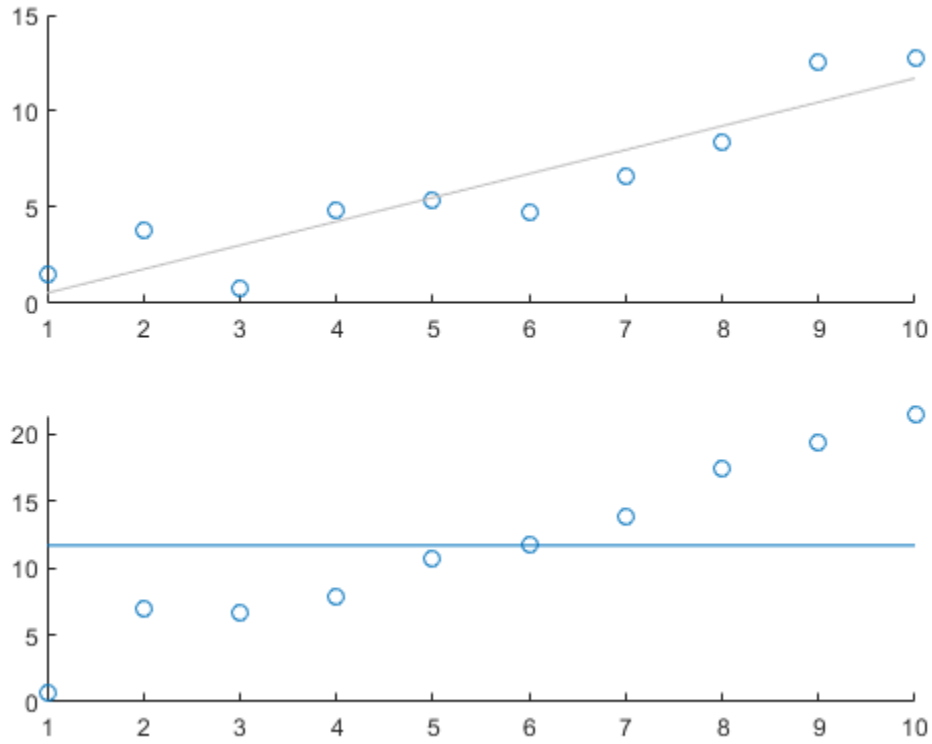
Define `ax1` as the top half of the figure, and `ax2` as the bottom half of the figure. Create the first scatter plot on the top axis using `y1`, and the second scatter plot on the bottom axis using `y2`.

```
figure
ax1 = subplot(2,1,1);
ax2 = subplot(2,1,2);

scatter(ax1,x,y1)
scatter(ax2,x,y2)
```



Superimpose a least-squares line on the top plot, and a reference line at the mean of the `y2` values in the bottom plot.

```
lsline(ax1) % This is equivalent to refline(ax1)

mu = mean(y2);
refline(ax2,[0 mu])
```



Use Least-Squares Line Object to Modify Line Properties

Define the x-variable and two different y-variables to use for the plots.

```
rng default  % For reproducibility
x = 1:10;
y1 = x + randn(1,10);
y2 = 2*x + randn(1,10);
```

Define `ax1` as the top half of the figure, and `ax2` as the bottom half of the figure. Create the first scatter plot on the top axis using `y1`, and the second scatter plot on the bottom axis using `y2`.

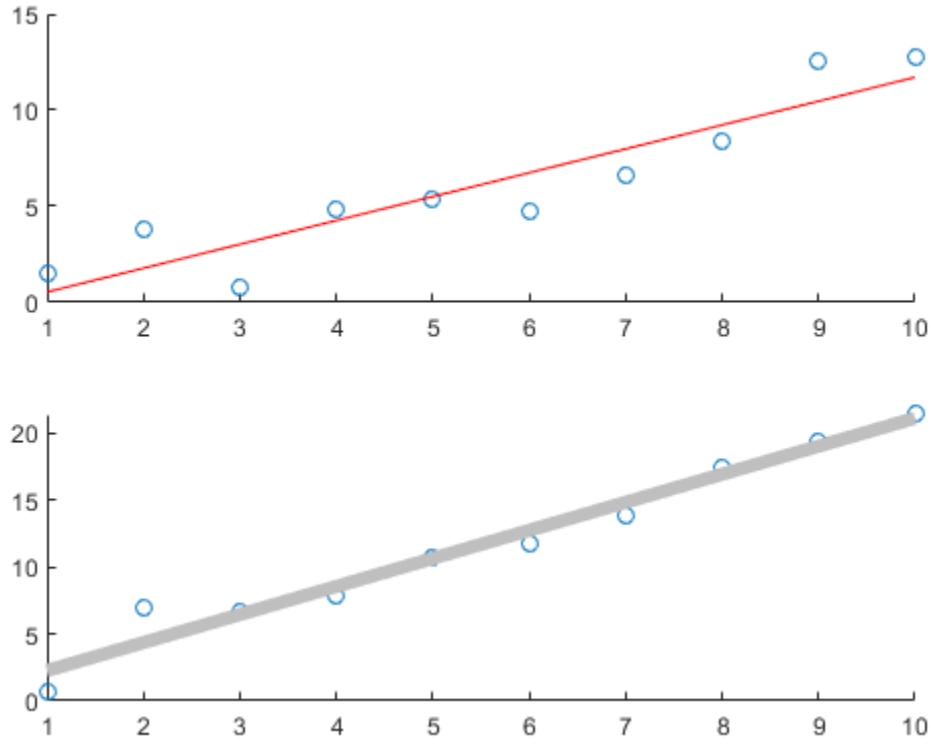
```
figure
ax1 = subplot(2,1,1);
ax2 = subplot(2,1,2);

scatter(ax1,x,y1)
scatter(ax2,x,y2)
```

Superimpose a least-squares line on the top plot. Then, use the least-squares line object `h1` to change the line color to red.

```
h1 = lsline(ax1);
h1.Color = 'r';
```

Superimpose a least-squares line on the bottom plot. Then, use the least-squares line object `h2` to increase the line width to 5.

```
h2 = lsline(ax2);
h2.LineWidth = 5;
```



Input Arguments

ax — Target axes

`gca` (default) | `axes` object

Target axes, specified as an `axes` object. If you do not specify the axes and if the current axes are Cartesian axes, then the `lsline` function uses the current axes.

Output Arguments

h — One or more least-squares line objects

scalar | vector

One or more least-squares line objects, returned as a scalar or a vector. These objects are unique identifiers, which you can use to query and modify properties of a specific least-squares line. For a list of properties, see Chart Line.

See Also

gca | gline | plot | refcurve | refline | scatter

Introduced before R2006a

mad

Mean or median absolute deviation

Syntax

```
y = mad(X)
y = mad(X,flag)
y = mad(X,flag,'all')
y = mad(X,flag,dim)
y = mad(X,flag,vecdim)
```

Description

`y = mad(X)` returns the mean absolute deviation of the values in `X`.

- If `X` is a vector, then `mad` returns the mean or median absolute deviation of the values in `X`.
- If `X` is a matrix, then `mad` returns a row vector containing the mean or median absolute deviation of each column of `X`.
- If `X` is a multidimensional array, then `mad` operates along the first nonsingleton dimension of `X`.

`y = mad(X,flag)` specifies whether to compute the mean absolute deviation (`flag = 0`, the default) or the median absolute deviation (`flag = 1`).

`y = mad(X,flag,'all')` returns the mean or median absolute deviation of all elements of `X`.

`y = mad(X,flag,dim)` returns the mean or median absolute deviation along the operating dimension `dim` of `X`.

`y = mad(X,flag,vecdim)` returns the mean or median absolute deviation over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `mad(X,0,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the mean absolute deviation of the elements on the corresponding page of `X`.

Examples

Compare Robustness of Scale Estimates

Compare the robustness of the standard deviation, mean absolute deviation, and median absolute deviation in the presence of outliers.

Create a data set `x` of normally distributed data. Create another data set `xo` that contains the elements of `x` and an additional outlier.

```
rng('default') % For reproducibility
x = normrnd(0,1,1,50);
xo = [x 10];
```

Compute the ratio of the standard deviations of the two data sets.

```
r1 = std(xo)/std(x)
```

```
r1 = 1.4633
```

Compute the ratio of the mean absolute deviations of the two data sets.

```
r2 = mad(xo)/mad(x)
```

```
r2 = 1.1833
```

Compute the ratio of the median absolute deviations of the two data sets.

```
r3 = mad(xo,1)/mad(x,1)
```

```
r3 = 1.0336
```

In this case, the median absolute deviation is less influenced by the outlier compared to the other two scale estimates.

Mean and Median Absolute Deviations of All Values

Find the mean and median absolute deviations of all the values in an array.

Create a 3-by-5-by-2 array `X` and add an outlier.

```
X = reshape(1:30,[3 5 2]);
X(6) = 100
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3   100     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the mean and median absolute deviations of the elements in `X`.

```
meandev = mad(X,0,'all')
```

```
meandev = 10.1178
```

```
mediandev = mad(X,1,'all')
```

```
mediandev = 7.5000
```

`meandev` is the mean absolute deviation of all the elements in `X`, and `mediandev` is the median absolute deviation of all the elements in `X`.

Find Median Absolute Deviation Along Given Dimension

Find the median absolute deviation along different dimensions for a multidimensional array.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 1-by-3-by-2 array of random numbers.

```
X = randn([1,3,2])
```

```
X =
X(:,:,1) =

    0.5377    1.8339   -2.2588


X(:,:,2) =

    0.8622    0.3188   -1.3077
```

Find the median absolute deviation of `X` along the default dimension.

```
Y2 = mad(X,1) % Flag is set to 1 for the median absolute deviation
```

```
Y2 =
Y2(:,:,1) =

    1.2962


Y2(:,:,2) =

    0.5434
```

By default, `mad` operates along the first dimension of `X` whose size does not equal 1. In this case, this dimension is the second dimension of `X`. Therefore, `Y2` is a 1-by-1-by-2 array.

Find the median absolute deviation of `X` along the third dimension.

```
Y3 = mad(X,1,3)
```

```
Y3 = 1×3

    0.1623    0.7576    0.4756
```

`Y3` is a 1-by-3 matrix.

Find Mean Absolute Deviation Along Vector of Dimensions

Find the mean absolute deviation over multiple dimensions by using the `vecdim` input argument.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4 3 2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the mean absolute deviation of each page of `X` by specifying the first and second dimensions.

```
ypage = mad(X,0,[1 2])
```

```
ypage =
ypage(:,:,1) =

    1.4626


ypage(:,:,2) =

    0.6652
```

For example, `ypage(:,:,2)` is the mean absolute deviation of all the elements in `X(:,:,2)`, and is equivalent to specifying `mad(X(:,:,2),0,'all')`.

Find the mean absolute deviation of the elements in each `X(:,i,:)` slice by specifying the first and third dimensions.

```
ycol = mad(X,0,[1 3])
```

```
ycol = 1×3

    0.8330    0.7872    1.5227
```

For example, `ycol(3)` is the mean absolute deviation of all the elements in `X(:,3,:)`, and is equivalent to specifying `mad(X(:,3,:),0,'all')`.

Input Arguments

X — Input data

vector | matrix | multidimensional array

Input data that represents a sample from a population, specified as a vector, matrix, or multidimensional array.

- If X is a vector, then mad returns the mean or median absolute deviation of the values in X.
- If X is a matrix, then mad returns a row vector containing the mean or median absolute deviation of each column of X.
- If X is a multidimensional array, then mad operates along the first nonsingleton dimension of X.

To specify the operating dimension when X is a matrix or an array, use the dim input argument.

mad treats NaNs as missing values and removes them.

Data Types: single | double

flag — Indicator for the type of deviation

0 (default) | 1

Indicator for the type of deviation, specified as 0 or 1.

- If flag is 0 (default), then mad computes the mean absolute deviation, mean(abs(X − mean(X))).
- If flag is 1, then mad computes the median absolute deviation, median(abs(X − median(X))).

Data Types: single | double | logical

dim — Dimension

positive integer

Dimension along which to operate, specified as a positive integer. If you do not specify a value for dim, then the default is the first dimension of X whose size does not equal 1.

Consider the mean absolute deviation of a matrix X:

- If dim is equal to 1, then mad(X) returns a row vector that contains the mean absolute deviation of each column in X.
- If dim is equal to 2, then mad(X) returns a column vector that contains the mean absolute deviation of each row in X.

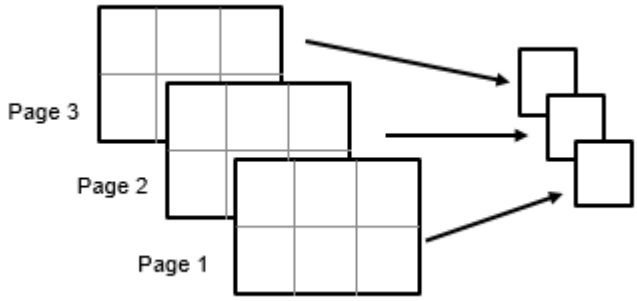
Data Types: single | double

vecdim — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of vecdim represents a dimension of the input array X. The output y has length 1 in the specified operating dimensions. The other dimension lengths are the same for X and y.

For example, if `X` is a 2-by-3-by-3 array, then `mad(X,0,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the mean absolute deviation of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

y — Mean or median absolute deviation

scalar | vector | matrix | multidimensional array

Mean or median absolute deviation, returned as a scalar, vector, matrix, or multidimensional array. If `flag` is 0 (default), then `y` is the mean absolute deviation of the values in `X`, `mean(abs(X - mean(X)))`. If `flag` is 1, then `y` is the median absolute deviation of the values in `X`, `median(abs(X - median(X)))`.

Tips

- For normally distributed data, multiply `mad` by one of the following factors to obtain an estimate of the normal scale parameter σ:
  - `sigma = 1.253 * mad(X,0)` — For mean absolute deviation
  - `sigma = 1.4826 * mad(X,1)` — For median absolute deviation

References

[1] Mosteller, F., and J. Tukey. *Data Analysis and Regression*. Upper Saddle River, NJ: Addison-Wesley, 1977.

[2] Sachs, L. *Applied Statistics: A Handbook of Techniques*. New York: Springer-Verlag, 1984, p. 253.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

- Empty `dim` input is not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also
`iqr` | `range` | `std`

Introduced before R2006a

mahal

Mahalanobis distance

Syntax

```
d2 = mahal(Y,X)
```

Description

`d2 = mahal(Y,X)` returns the squared Mahalanobis distance on page 33-3829 of each observation in `Y` to the reference samples in `X`.

Examples

Compare Mahalanobis and Squared Euclidean Distances

Generate a correlated bivariate sample data set.

```
rng('default') % For reproducibility
X = mvnrnd([0;0],[1 .9;.9 1],1000);
```

Specify four observations that are equidistant from the mean of `X` in Euclidean distance.

```
Y = [1 1;1 -1;-1 1;-1 -1];
```

Compute the Mahalanobis distance of each observation in `Y` to the reference samples in `X`.

```
d2_mahal = mahal(Y,X)
```

```
d2_mahal = 4×1

    1.1095
   20.3632
   19.5939
    1.0137
```

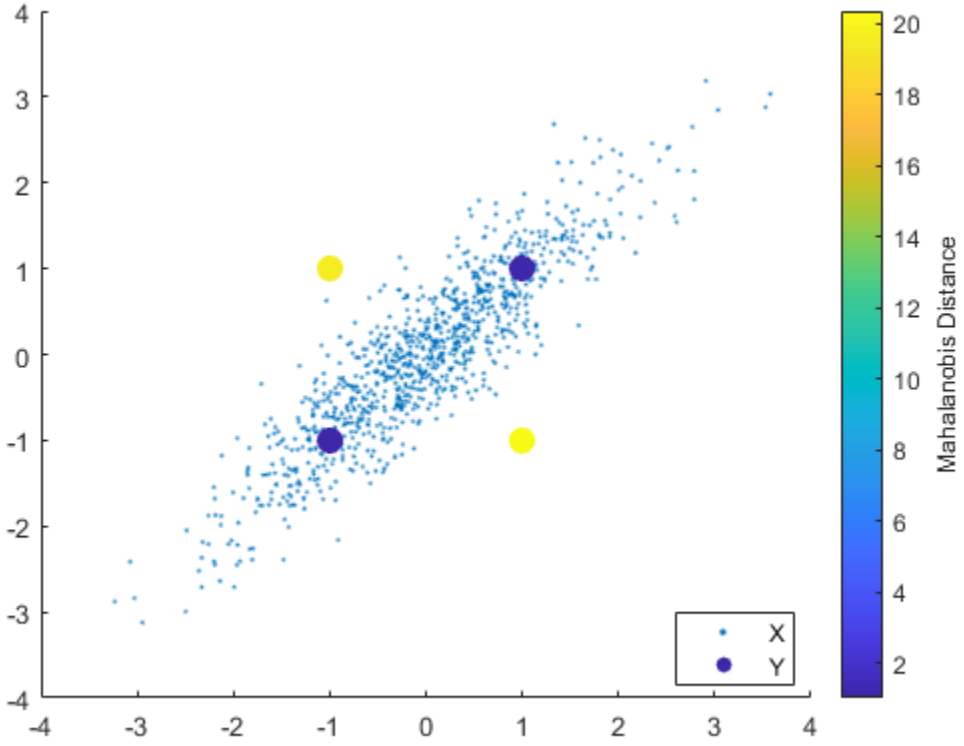
Compute the squared Euclidean distance of each observation in `Y` from the mean of `X` .

```
d2_Euclidean = sum((Y-mean(X)).^2,2)
```

```
d2_Euclidean = 4×1

    2.0931
    2.0399
    1.9625
    1.9094
```

Plot `X` and `Y` by using `scatter` and use marker color to visualize the Mahalanobis distance of `Y` to the reference samples in `X`.

```
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
scatter(Y(:,1),Y(:,2),100,d2_mahal,'o','filled')
hb = colorbar;
ylabel(hb,'Mahalanobis Distance')
legend('X','Y','Location','best')
```



All observations in Y (`[1,1]`, `[-1,-1,]`, `[1,-1]`, and `[-1,1]`) are equidistant from the mean of X in Euclidean distance. However, `[1,1]` and `[-1,-1]` are much closer to X than `[1,-1]` and `[-1,1]` in Mahalanobis distance. Because Mahalanobis distance considers the covariance of the data and the scales of the different variables, it is useful for detecting outliers.

Input Arguments

Y — Data

n-by-*m* numeric matrix

Data, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

X and Y must have the same number of columns, but can have different numbers of rows.

Data Types: `single` | `double`

X — Reference samples

p-by-*m* numeric matrix

Reference samples, specified as a *p*-by-*m* numeric matrix, where *p* is the number of samples and *m* is the number of variables in each sample.

X and Y must have the same number of columns, but can have different numbers of rows. X must have more rows than columns.

Data Types: `single` | `double`

Output Arguments

d2 — Squared Mahalanobis distance

n-by-1 numeric vector

Squared Mahalanobis distance on page 33-3829 of each observation in Y to the reference samples in X, returned as an *n*-by-1 numeric vector, where *n* is the number of observations in X.

More About

Mahalanobis Distance

The Mahalanobis distance is a measure between a sample point and a distribution.

The Mahalanobis distance from a vector y to a distribution with mean μ and covariance Σ is

$$d = \sqrt{(y-\mu)\sum^{-1}(y-\mu)'}.$$

This distance represents how far y is from the mean in number of standard deviations.

`mahal` returns the squared Mahalanobis distance d^2 from an observation in Y to the reference samples in X. In the `mahal` function, μ and Σ are the sample mean and covariance of the reference samples, respectively.

See Also

`mahal` | `pdist`

Introduced before R2006a

mahal

Mahalanobis distance to class means

Syntax

```
M = mahal(obj,X)
M = mahal(obj,X,Name,Value)
```

Description

`M = mahal(obj,X)` returns the squared Mahalanobis distances from observations in `X` to the class means in `obj`.

`M = mahal(obj,X,Name,Value)` computes the squared Mahalanobis distance with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier of class `ClassificationDiscriminant` or `CompactClassificationDiscriminant`, typically constructed with `fitcdiscr`.

`X`

Numeric matrix of size `n`-by-`p`, where `p` is the number of predictors in `obj`, and `n` is any positive integer. `mahal` computes the Mahalanobis distances from the rows of `X` to each of the `K` means of the classes in `obj`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`ClassLabels`

Class labels consisting of `n` elements of `obj.Y`, where `n` is the number of rows of `X`.

Output Arguments

`M`

Size and meaning of output `M` depends on whether the `ClassLabels` name-value pair is present:

- No `ClassLabels` — `M` is a numeric matrix of size `n`-by-`K`, where `K` is the number of classes in `obj`, and `n` is the number of rows in `X`. `M(i,j)` is the squared Mahalanobis distance from the `i`th row of `X` to the mean of class `j`.
- `ClassLabels` exists — `M` is a column vector with `n` elements. `M(i)` is the squared Mahalanobis distance from the `i`th row of `X` to the mean for the class of the `i`th element of `ClassLabels`.

Examples

Find the Mahalanobis distances from the mean of the Fisher iris data to the class means, using distinct covariance matrices for each class:

```
load fisheriris
obj = fitcdiscr(meas,species,...
    'DiscrimType','quadratic');
mahadist = mahal(obj,mean(meas))

mahadist =
  220.0667    5.0254   30.5804
```

More About

Mahalanobis Distance

The Mahalanobis distance $d(x,y)$ between n-dimensional points x and y, with respect to a given n-by-n covariance matrix S, is

$$d(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)}.$$

See Also

`CompactClassificationDiscriminant` | `fitcdiscr` | `gmdistribution` | `mahal`

Topics

“Discriminant Analysis Classification” on page 20-2

mahal

Mahalanobis distance to Gaussian mixture component

Syntax

```
d2 = mahal(gm,X)
```

Description

`d2 = mahal(gm,X)` returns the squared Mahalanobis distance of each observation in `X` to each Gaussian mixture component in `gm`.

Examples

Measure Mahalanobis Distance

Generate random variates that follow a mixture of two bivariate Gaussian distributions by using the `mvnrnd` function. Fit a Gaussian mixture model (GMM) to the generated data by using the `fitgmdist` function, and then compute Mahalanobis distances between the generated data and the mixture components of the fitted GMM.

Define the distribution parameters (means and covariances) of two bivariate Gaussian mixture components.

```
rng('default') % For reproducibility
mu1 = [1 2];           % Mean of the 1st component
sigma1 = [2 0; 0 .5]; % Covariance of the 1st component
mu2 = [-3 -5];         % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Covariance of the 2nd component
```

Generate an equal number of random variates from each component, and combine the two sets of random variates.

```
r1 = mvnrnd(mu1,sigma1,1000);
r2 = mvnrnd(mu2,sigma2,1000);
X = [r1; r2];
```

The combined data set `X` contains random variates following a mixture of two bivariate Gaussian distributions.

Fit a two-component GMM to `X`.

```
gm = fitgmdist(X,2)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:   -2.9617   -4.9727
```

```
Component 2:
Mixing proportion: 0.500000
Mean:    0.9539    2.0261
```

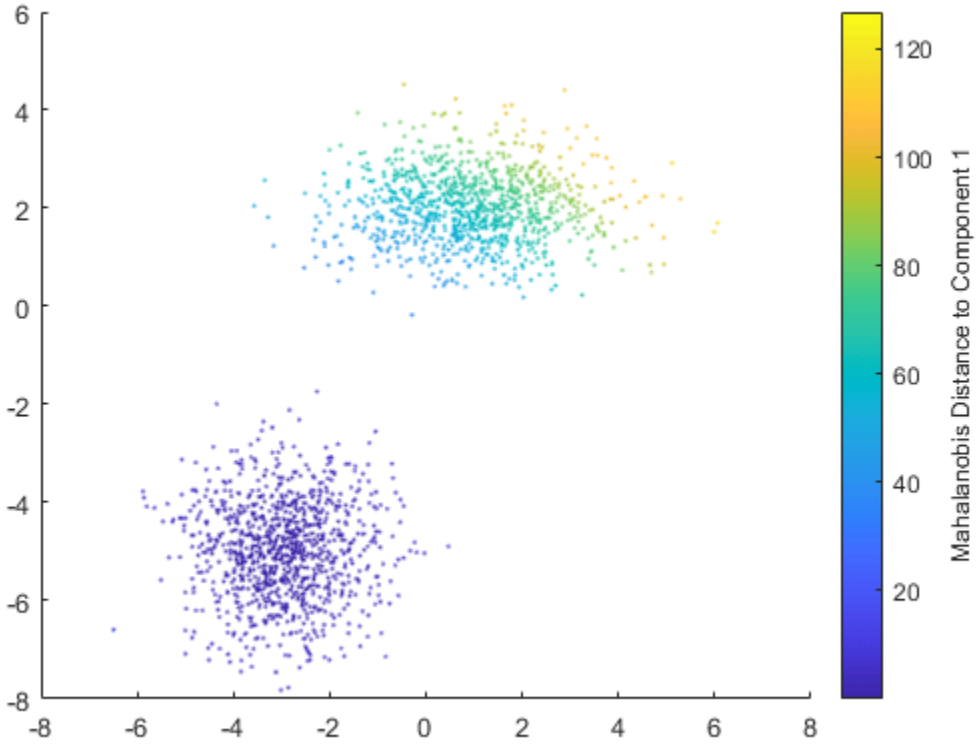
`fitgmdist` fits a GMM to `X` using two mixture components. The means of `Component 1` and `Component 2` are `[-2.9617,-4.9727]` and `[0.9539,2.0261]`, which are close to `mu2` and `mu1`, respectively.

Compute the Mahalanobis distance of each point in `X` to each component of `gm`.

```
d2 = mahal(gm,X);
```

Plot `X` by using `scatter` and use marker color to visualize the Mahalanobis distance to `Component 1`.

```
scatter(X(:,1),X(:,2),10,d2(:,1),'.') % Scatter plot with points of size 10
c = colorbar;
ylabel(c,'Mahalanobis Distance to Component 1')
```



Input Arguments

`gm` — Gaussian mixture distribution

`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a gmdistribution object using gmdistribution or fitgmdist. Use the gmdistribution function to create a gmdistribution object by specifying the distribution parameters. Use the fitgmdist function to fit a gmdistribution model to data given a fixed number of components.

X — Data
n-by-*m* numeric matrix

Data, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

If a row of X contains NaNs, then mahal excludes the row from the computation. The corresponding value in d2 is NaN.

Data Types: single | double

Output Arguments

d2 — Squared Mahalanobis distance
n-by-*k* numeric matrix

Squared Mahalanobis distance of each observation in X to each Gaussian mixture component in gm, returned as an *n*-by-*k* numeric matrix, where *n* is the number of observations in X and *k* is the number of mixture components in gm.

d2(i,j) is the squared distance of observation i to the jth Gaussian mixture component.

More About

Mahalanobis Distance

The Mahalanobis distance is a measure between a sample point and a distribution.

The Mahalanobis distance from a vector x to a distribution with mean μ and covariance Σ is

$$d = \sqrt{(x-\mu)\sum^{-1}(x-\mu)'}.$$

This distance represents how far x is from the mean in number of standard deviations.

mahal returns the squared Mahalanobis distance d^2 from an observation in X to a mixture component in gm.

See Also

cluster | fitgmdist | gmdistribution | mahal | posterior

Topics
“Cluster Using Gaussian Mixture Model” on page 16-39
“Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46
“Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52

Introduced in R2007b

maineffectsplot

Main effects plot for grouped data

Syntax

```
maineffectsplot(Y,GROUP)
maineffectsplot(Y,GROUP,param1,val1,param2,val2,...)
[figh,AXESH] = maineffectsplot(...)
```

Description

`maineffectsplot(Y,GROUP)` displays main effects plots for the group means of matrix `Y` with groups defined by entries in `GROUP`, which can be a cell array or a matrix. `Y` is a numeric matrix or vector. If `Y` is a matrix, the rows represent different observations and the columns represent replications of each observation. If `GROUP` is a cell array, then each cell of `GROUP` must contain a grouping variable that is a categorical variable, numeric vector, character matrix, string array, or single-column cell array of character vectors. If `GROUP` is a matrix, then its columns represent different grouping variables. Each grouping variable must have the same number of rows as `Y`. The number of grouping variables must be greater than 1.

The display has one subplot per grouping variable, with each subplot showing the group means of `Y` as a function of one grouping variable.

`maineffectsplot(Y,GROUP,param1,val1,param2,val2,...)` specifies one or more of the following name/value pairs:

- `'varnames'` — Grouping variable names in a character matrix, a string array, or a cell array of character vectors, one per grouping variable. Default names are `'X1'`, `'X2'`, ...
- `'statistic'` — Values that indicate whether the group mean or the group standard deviation should be plotted. Use `'mean'` or `'std'`. The default is `'mean'`. If the value is `'std'`, `Y` is required to have multiple columns.
- `'parent'` — A handle to the figure window for the plots. The default is the current figure window.

`[figh,AXESH] = maineffectsplot(...)` returns the handle `figh` to the figure window and an array of handles `AXESH` to the subplot axes.

Examples

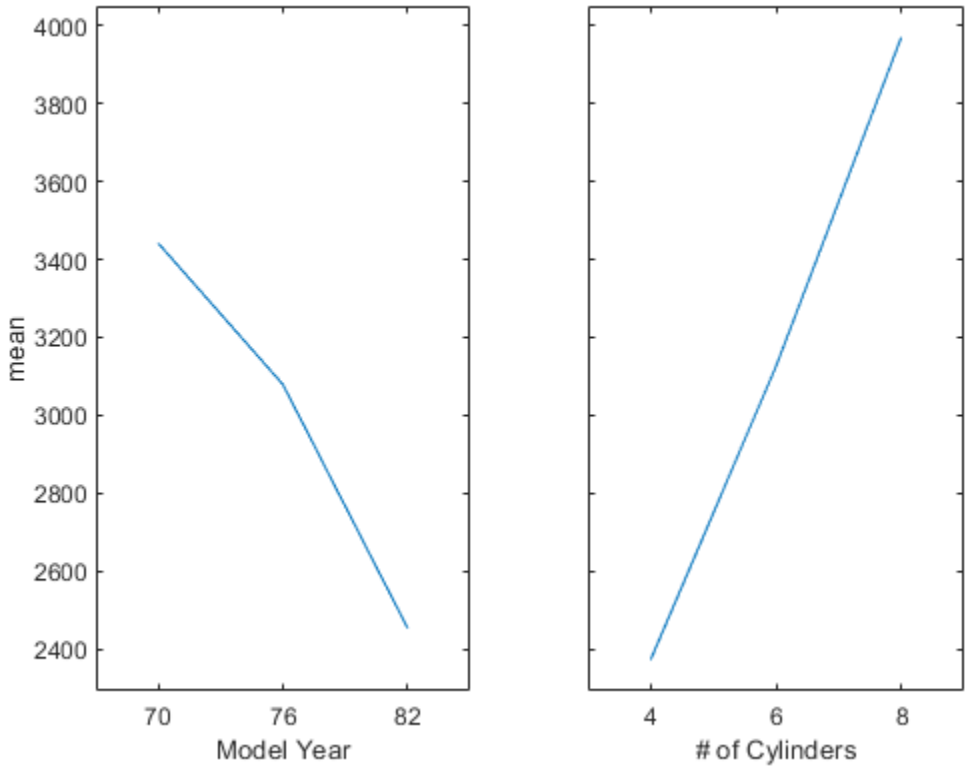
Main Effects Plot

Load the sample data.

```
load carsmall;
```

Display main effects plots for car weight with two grouping variables, model year and number of cylinders.

```
maineffectsplot(Weight,{Model_Year,Cylinders}, ...
                'varnames',{'Model Year','# of Cylinders'})
```



See Also

`interactionplot` | `multivarichart`

Topics

“Grouping Variables” on page 2-45

Introduced in R2006b

makecdiscr

Construct discriminant analysis classifier from parameters

Syntax

```
cobj = makecdiscr(Mu,Sigma)
cobj = makecdiscr(Mu,Sigma,Name,Value)
```

Description

`cobj = makecdiscr(Mu,Sigma)` constructs a compact discriminant analysis classifier from the class means `Mu` and covariance matrix `Sigma`.

`cobj = makecdiscr(Mu,Sigma,Name,Value)` constructs a compact classifier with additional options specified by one or more name-value pair arguments. For example, you can specify the cost of misclassification or the prior probabilities for each class.

Examples

Construct a Compact Linear Discriminant Analysis Classifier

Construct a compact linear discriminant analysis classifier from the means and covariances of the Fisher iris data.

```
load fisheriris
mu(1,:) = mean(meas(1:50,:));
mu(2,:) = mean(meas(51:100,:));
mu(3,:) = mean(meas(101:150,:));

mm1 = repmat(mu(1,:),50,1);
mm2 = repmat(mu(2,:),50,1);
mm3 = repmat(mu(3,:),50,1);
cc = meas;
cc(1:50,:) = cc(1:50,:) - mm1;
cc(51:100,:) = cc(51:100,:) - mm2;
cc(101:150,:) = cc(101:150,:) - mm3;
sigstar = cc' * cc / 147; % unbiased estimator of sigma
cpct = makecdiscr(mu,sigstar,...
   'ClassNames',{'setosa','versicolor','virginica'})
```

```
cpct =
  CompactClassificationDiscriminant
            PredictorNames: {'x1'  'x2'  'x3'  'x4'}
              ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
              DiscrimType: 'linear'
                       Mu: [3x4 double]
                   Coeffs: [3x3 struct]
```

```
Properties, Methods
```

Input Arguments

`Mu` — Class means

matrix of scalar values

Class means, specified as a `K`-by-`p` matrix of scalar values class means of size. `K` is the number of classes, and `p` is the number of predictors. Each row of `Mu` represents the mean of the multivariate normal distribution of the corresponding class. The class indices are in the `ClassNames` attribute.

Data Types: `single` | `double`

`Sigma` — Within-class covariance

matrix of scalar values

Within-class covariance, specified as a matrix of scalar values.

- For a linear discriminant, `Sigma` is a symmetric, positive semidefinite matrix of size `p`-by-`p`, where `p` is the number of predictors.
- For a quadratic discriminant, `Sigma` is an array of size `p`-by-`p`-by-`K`, where `K` is the number of classes. For each `i`, `Sigma(:,:,i)` is a symmetric, positive semidefinite matrix.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ClassNames',{'setosa' 'versicolor' 'virginica'}` specifies a discriminant analysis classifier that uses `'setosa'`, `'versicolor'`, and `'virginica'` as the grouping variables.

`ClassNames` — Class names

numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Class names as ordered in `Mu`, specified as the comma-separated pair consisting of `'ClassNames'` and an array containing grouping variables. Use any data type for a grouping variable, including numeric vector, categorical vector, logical vector, character array, string array, or cell array of character vectors.

The default is `1:K`, where `K` is the number of classes (the number of rows of `Mu`).

Example: `'ClassNames',{'setosa' 'versicolor' 'virginica'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`Cost` — Cost of misclassification

square matrix | structure

Cost of misclassification, specified as the comma-separated pair consisting of `'Cost'` and a square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`.

Alternatively, `Cost` can be a structure `S` having two fields: `S.ClassNames` containing the group names as a variable of the same type as `y`, and `S.ClassificationCosts` containing the cost matrix.

The default is `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`.

Data Types: `single` | `double` | `struct`

`PredictorNames` — Predictor variable names
`{'X1','X2',...}` (default) | string array | cell array of character vectors

Predictor variable names, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array or cell array of character vectors containing the names for the predictor variables, in the order in which they appear in `X`.

Data Types: `string` | `cell`

`Prior` — Prior probabilities
`'uniform'` (default) | vector of scalar values | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and one of the following:

- `'uniform'`, meaning all class prior probabilities are equal
- A vector containing one scalar value for each class
- A structure `S` with two fields:
  - `S.ClassNames` containing the class names as a variable of the same type as `ClassNames`
  - `S.ClassProbs` containing a vector of corresponding probabilities

Data Types: `char` | `string` | `single` | `double` | `struct`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as the comma-separated pair consisting of `'ResponseName'` and a character vector or string scalar containing the name of the response variable `y`.

Example: `'ResponseName','Response'`

Data Types: `char` | `string`

Output Arguments

`cobj` — Discriminant analysis classifier
discriminant analysis classifier object

Discriminant analysis classifier, returned as a discriminant analysis classifier object of class `CompactClassificationDiscriminant`. You can use the `predict` method to predict classification labels for new data.

Tips

- You can change the discriminant type using dot notation after constructing `cobj`:

```
cobj.DiscrimType = 'discrimType'
```

where *discrimType* is one of `'linear'`, `'quadratic'`, `'diagLinear'`, `'diagQuadratic'`, `'pseudoLinear'`, or `'pseudoQuadratic'`. You can change between linear types or between quadratic types, but cannot change between a linear and a quadratic type.

- `cobj` is a linear classifier when `Sigma` is a matrix. `cobj` is a quadratic classifier when `Sigma` is a three-dimensional array.

See Also

`CompactClassificationDiscriminant` | `compact` | `fitcdiscr` | `predict`

Topics

“Discriminant Analysis Classification” on page 20-2

Introduced in R2014a

makedist

Create probability distribution object

Syntax

```
pd = makedist(distname)
pd = makedist(distname,Name,Value)

list = makedist

makedist -reset
```

Description

`pd = makedist(distname)` creates a probability distribution object for the distribution `distname`, using the default parameter values.

`pd = makedist(distname,Name,Value)` creates a probability distribution object with one or more distribution parameter values specified by name-value pair arguments.

`list = makedist` returns a cell array `list` containing a list of the probability distributions that `makedist` can create.

`makedist -reset` resets the list of distributions by searching the path for files contained in a package named `prob` and implementing classes derived from `ProbabilityDistribution`. Use this syntax after you define a custom distribution function. For details, see “Define Custom Distributions Using the Distribution Fitter App” on page 5-81.

Examples

Create a Normal Distribution Object

Create a normal distribution object using the default parameter values.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Compute the interquartile range of the distribution.

```
r = iqr(pd)
```

```
r = 1.3490
```

Create a Gamma Distribution Object

Create a gamma distribution object using the default parameter values.

```
pd = makedist('Gamma')
```

```
pd =
  GammaDistribution

  Gamma distribution
    a = 1
    b = 1
```

Compute the mean of the gamma distribution.

```
mean = mean(pd)
```

```
mean = 1
```

Specify Parameters for a Normal Distribution Object

Create a normal distribution object with parameter values `mu = 75` and `sigma = 10`.

```
pd = makedist('Normal','mu',75,'sigma',10)
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75
    sigma = 10
```

Specify Parameters for a Gamma Distribution Object

Create a gamma distribution object with the parameter value `a = 3` and the default value `b = 1`.

```
pd = makedist('Gamma','a',3)
```

```
pd =
  GammaDistribution

  Gamma distribution
    a = 3
    b = 1
```

Input Arguments

`distname` — Distribution name

character vector | string scalar

Distribution name, specified as one of the following character vectors or string scalars. The distribution specified by `distname` determines the type of the returned probability distribution object.

Distribution Name	Description	Distribution Object
`'Beta'`	Beta distribution	`BetaDistribution`
`'Binomial'`	Binomial distribution	`BinomialDistribution`
`'BirnbaumSaunders'`	Birnbaum-Saunders distribution	`BirnbaumSaundersDistribution`
`'Burr'`	Burr distribution	`BurrDistribution`
`'Exponential'`	Exponential distribution	`ExponentialDistribution`
`'ExtremeValue'`	Extreme Value distribution	`ExtremeValueDistribution`
`'Gamma'`	Gamma distribution	`GammaDistribution`
`'GeneralizedExtremeValue'`	Generalized Extreme Value distribution	`GeneralizedExtremeValueDistribution`
`'GeneralizedPareto'`	Generalized Pareto distribution	`GeneralizedParetoDistribution`
`'HalfNormal'`	Half-normal distribution	`HalfNormalDistribution`
`'InverseGaussian'`	Inverse Gaussian distribution	`InverseGaussianDistribution`
`'Logistic'`	Logistic distribution	`LogisticDistribution`
`'Loglogistic'`	Loglogistic distribution	`LoglogisticDistribution`
`'Lognormal'`	Lognormal distribution	`LognormalDistribution`
`'Multinomial'`	Multinomial distribution	`MultinomialDistribution`
`'Nakagami'`	Nakagami distribution	`NakagamiDistribution`
`'NegativeBinomial'`	Negative Binomial distribution	`NegativeBinomialDistribution`
`'Normal'`	Normal distribution	`NormalDistribution`
`'PiecewiseLinear'`	Piecewise Linear distribution	`PiecewiseLinearDistribution`
`'Poisson'`	Poisson distribution	`PoissonDistribution`
`'Rayleigh'`	Rayleigh distribution	`RayleighDistribution`
`'Rician'`	Rician distribution	`RicianDistribution`
`'Stable'`	Stable distribution	`StableDistribution`
`'tLocationScale'`	*t* Location-Scale distribution	`tLocationScaleDistribution`
`'Triangular'`	Triangular distribution	`TriangularDistribution`

Distribution Name	Description	Distribution Object
'Uniform'	Uniform distribution	UniformDistribution
'Weibull'	Weibull distribution	WeibullDistribution

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

Example: makedist('Normal','mu',10) specifies a normal distribution with parameter mu equal to 10, and parameter sigma equal to the default value of 1.

Beta Distribution

a — First shape parameter
1 (default) | positive scalar value

Example: 'a',3

Data Types: single | double

b — Second shape parameter
1 (default) | positive scalar value

Example: 'b',5

Data Types: single | double

Binomial Distribution

N — Number of trials
1 (default) | positive integer value

Example: 'N',25

Data Types: single | double

p — Probability of success
0.5 (default) | scalar value in the range [0,1]

Example: 'p',0.25

Data Types: single | double

Birnbaum-Saunders Distribution

beta — Scale parameter
1 (default) | positive scalar value

Example: 'beta',2

Data Types: single | double

gamma — Shape parameter
1 (default) | nonnegative scalar value

Example: 'gamma',0

Data Types: single | double

Burr Distribution

`alpha` — Scale parameter
`1` (default) | positive scalar value

Example: `'alpha',2`

Data Types: `single` | `double`

`c` — First shape parameter
`1` (default) | positive scalar value

Example: `'c',2`

Data Types: `single` | `double`

`k` — Second shape parameter
`1` (default) | positive scalar value

Example: `'k',5`

Data Types: `single` | `double`

Exponential Distribution

`mu` — Mean parameter
`1` (default) | positive scalar value

Example: `'mu',5`

Data Types: `single` | `double`

Extreme Value Distribution

`mu` — Location parameter
`0` (default) | scalar value

Example: `'mu',-2`

Data Types: `single` | `double`

`sigma` — Scale parameter
`1` (default) | nonnegative scalar value

Example: `'sigma',2`

Data Types: `single` | `double`

Gamma Distribution

`a` — Shape parameter
`1` (default) | positive scalar value

Example: `'a',2`

Data Types: `single` | `double`

`b` — Scale parameter
`1` (default) | nonnegative scalar value

Example: `'b',0`

Data Types: `single` | `double`

Generalized Extreme Value Distribution

k — Shape parameter
0 (default) | scalar value

Example: 'k',0

Data Types: single | double

sigma — Scale parameter
1 (default) | nonnegative scalar value

Example: 'sigma',2

Data Types: single | double

mu — Location parameter
0 (default) | scalar value

Example: 'mu',1

Data Types: single | double

Generalized Pareto Distribution

k — Shape parameter
1 (default) | scalar value

Example: 'k',0

Data Types: single | double

sigma — Scale parameter
1 (default) | nonnegative scalar value

Example: 'sigma',2

Data Types: single | double

theta — Location parameter
1 (default) | scalar value

Example: 'theta',2

Data Types: single | double

Half-Normal Distribution

mu — Location parameter
0 (default) | scalar value

Example: 'mu',1

Data Types: single | double

sigma — Shape parameter
1 (default) | nonnegative scalar value

Example: 'sigma',2

Data Types: single | double

Inverse Gaussian Distribution

`mu` — Scale parameter
`1` (default) | positive scalar value

Example: `'mu',2`

Data Types: `single` | `double`

`lambda` — Shape parameter
`1` (default) | positive scalar value

Example: `'lambda',4`

Data Types: `single` | `double`

Logistic Distribution

`mu` — Mean
`0` (default) | scalar value

Example: `'mu',2`

Data Types: `single` | `double`

`sigma` — Scale parameter
`1` (default) | nonnegative scalar value

Example: `'sigma',4`

Data Types: `single` | `double`

Loglogistic Distribution

`mu` — Mean of logarithmic values
`0` (default) | scalar value

Example: `'mu',2`

Data Types: `single` | `double`

`sigma` — Scale parameter of logarithmic values
`1` (default) | nonnegative scalar value

Example: `'sigma',4`

Data Types: `single` | `double`

Lognormal Distribution

`mu` — Mean of logarithmic values
`0` (default) | scalar value

Example: `'mu',2`

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values
`1` (default) | nonnegative scalar value

Example: `'sigma',2`

Data Types: `single` | `double`

Multinomial Distribution

probabilities — Outcome probabilities

`[0.500 0.500]` (default) | vector of scalar values in the range [0,1]

Outcome probabilities, specified as a vector of scalar values in the range [0,1]. The probabilities sum to 1 and correspond to outcomes [1, 2, ..., k], where *k* is the number of elements in the probabilities vector.

Example: `'probabilities',[0.1 0.2 0.5 0.2]` gives the probabilities that the outcome is 1, 2, 3, or 4, respectively.

Data Types: `single` | `double`

Nakagami Distribution

mu — Shape parameter

`1` (default) | positive scalar value

Example: `'mu',5`

Data Types: `single` | `double`

omega — Scale parameter

`1` (default) | positive scalar value

Example: `'omega',5`

Data Types: `single` | `double`

Negative Binomial Distribution

R — Number of successes

`1` (default) | positive scalar value

Example: `'R',5`

Data Types: `single` | `double`

p — Probability of success

`0.5` (default) | scalar value in the range (0,1]

Example: `'p',0.1`

Data Types: `single` | `double`

Normal Distribution

mu — Mean

`0` (default) | scalar value

Example: `'mu',2`

Data Types: `single` | `double`

sigma — Standard deviation

`1` (default) | nonnegative scalar value

Example: `'sigma',2`

Data Types: `single` | `double`

Piecewise Linear Distribution

x — Data values
1 (default) | monotonically increasing vector of scalar values

Example: `'x',[1 2 3]`

Data Types: `single` | `double`

Fx — cdf values
1 (default) | monotonically increasing vector of scalar values that start at 0 and end at 1

Example: `'Fx',[0.2 0.5 1]`

Data Types: `single` | `double`

Poisson Distribution

lambda — Mean
1 (default) | nonnegative scalar value

Example: `'lambda',5`

Data Types: `single` | `double`

Rayleigh Distribution

b — Defining parameter
1 (default) | positive scalar value

Example: `'b',3`

Data Types: `single` | `double`

Rician Distribution

s — Noncentrality parameter
1 (default) | nonnegative scalar value

Example: `'s',0`

Data Types: `single` | `double`

sigma — Scale parameter
1 (default) | positive scalar value

Example: `'sigma',2`

Data Types: `single` | `double`

Stable Distribution

alpha — First shape parameter
2 (default) | scalar value in the range (0,2]

Example: `'alpha',1`

Data Types: `single` | `double`

beta — Second shape parameter
0 (default) | scalar value in the range [-1,1]

Example: `'beta',0.5`

Data Types: `single` | `double`

`gam` — Scale parameter
`1` (default) | scalar value in the range (0,∞)

Example: `'gam',2`

Data Types: `single` | `double`

`delta` — Location parameter
0 (default) | scalar value

Example: `'delta',5`

Data Types: `single` | `double`

t Location-Scale Distribution

`mu` — Location parameter
`0` (default) | scalar value

Example: `'mu',-2`

Data Types: `single` | `double`

`sigma` — Scale parameter
`1` (default) | positive scalar value

Example: `'sigma',2`

Data Types: `single` | `double`

`nu` — Degrees of freedom
`5` (default) | positive scalar value

Example: `'nu',20`

Data Types: `single` | `double`

Triangular Distribution

`a` — Lower limit
`0` (default) | scalar value

Example: `'a',-2`

Data Types: `single` | `double`

`b` — Peak location
`0.5` (default) | scalar value greater than or equal to `a`

Example: `'b',1`

Data Types: `single` | `double`

`c` — Upper limit
`1` (default) | scalar value greater than or equal to `b`

Example: `'c',5`

Data Types: `single` | `double`

Uniform Distribution

`lower` — Lower parameter
`0` (default) | scalar value

Example: `'lower',-4`

Data Types: `single` | `double`

`upper` — Upper parameter
`1` (default) | scalar value greater than `lower`

Example: `'upper',2`

Data Types: `single` | `double`

Weibull Distribution

`a` — Scale parameter
`1` (default) | positive scalar value

Example: `'a',2`

Data Types: `single` | `double`

`b` — Shape parameter
`1` (default) | positive scalar value

Example: `'b',5`

Data Types: `single` | `double`

Output Arguments

`pd` — Probability distribution
probability distribution object

Probability distribution, returned as a probability distribution object of the type specified by `distname`.

`list` — List of probability distributions
cell array of character vectors

List of probability distributions that `makedist` can create, returned as a cell array of character vectors.

Alternative Functionality

App

The **Distribution Fitter** app opens a graphical user interface for you to import data from the workspace and interactively fit a probability distribution to that data. You can then save the distribution to the workspace as a probability distribution object. Open the Distribution Fitter app using `distributionFitter`, or click Distribution Fitter on the Apps tab.

See Also
`distributionFitter` | `fitdist`

Topics

Introduced in R2013a

manova

Class: `RepeatedMeasuresModel`

Multivariate analysis of variance

Syntax

```
manovatbl = manova(rm)
manovatbl = manova(rm,Name,Value)
[manovatbl,A,C,D] = manova( ___ )
```

Description

`manovatbl = manova(rm)` returns the results of multivariate analysis of variance (manova) for the repeated measures model `rm`.

`manovatbl = manova(rm,Name,Value)` also returns manova results with additional options, specified by one or more `Name,Value` pair arguments.

`[manovatbl,A,C,D] = manova( ___ )` also returns arrays `A`, `C`, and `D` for the hypotheses tests of the form `A*B*C = D`, where `D` is zero.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`WithinModel` — Model specifying within-subjects hypothesis test

`'separatemeans'` (default) | model specification using formula | *r*-by-*nc* matrix

Model specifying the within-subjects hypothesis test, specified as one of the following:

- `'separatemeans'` — Compute a separate mean for each group, and test for equality among the means.
- Model specification — This is a model specification in the within-subject factors. Test each term in the model. In this case, `tbl` contains a separate manova for each term in the formula, with the multivariate response equal to the vector of coefficients of that term.
- An *r*-by-*nc* matrix, *C*, specifying *nc* contrasts among the *r* repeated measures. If *Y* represents the matrix of repeated measures you use in the repeated measures model `rm`, then the output `tbl` contains a separate manova for each column of *Y***C*.

Example: `'WithinModel','separatemeans'`

Data Types: `single` | `double` | `char` | `string`

By — Single between-subjects factor
character vector | string scalar

Single between-subjects factor, specified as the comma-separated pair consisting of `'By'` and a character vector or string scalar. `manova` performs a separate test of the within-subjects model for each value of this factor.

For example, if you have a between-subjects factor, Drug, then you can specify that factor to perform manova as follows.

Example: `'By','Drug'`

Data Types: `char` | `string`

Output Arguments

`manovatbl` — Results of multivariate analysis of variance
table

Results of multivariate analysis of variance for the repeated measures model `rm`, returned as a `table`.

`manova` uses these methods to measure the contributions of the model terms to the overall covariance:

- Wilks' Lambda
- Pillai's trace
- Hotelling-Lawley trace
- Roy's maximum root statistic

For details, see "Multivariate Analysis of Variance for Repeated Measures" on page 9-59.

`manova` returns the results for these tests for each group. `manovatbl` contains the following columns.

Column Name	Definition
`Within`	Within-subject terms
`Between`	Between-subject terms
`Statistic`	Name of the statistic computed
`Value`	Value of the corresponding statistic
`F`	*F*-statistic value
`RSquare`	Measure for variance explained
`df1`	Numerator degrees of freedom
`df2`	Denominator degrees of freedom
`pValue`	*p*-value for the corresponding *F*-statistic value

Data Types: `table`

`A` — Specification based on between-subjects model

matrix | cell array

Specification based on the between-subjects model, returned as a matrix or a cell array. It permits the hypothesis on the elements within given columns of `B` (within time hypothesis). If `manovatbl` contains multiple hypothesis tests, `A` might be a cell array.

Data Types: `single` | `double` | `cell`

`C` — Specification based on within-subjects model

matrix | cell array

Specification based on the within-subjects model, returned as a matrix or a cell array. It permits the hypotheses on the elements within given rows of `B` (between time hypotheses). If `manovatbl` contains multiple hypothesis tests, `C` might be a cell array.

Data Types: `single` | `double` | `cell`

`D` — Hypothesis value

0

Hypothesis value, returned as 0.

Examples

Perform Multivariate Analysis of Variance

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = table([1 2 3 4]','VariableNames',{'Measurements'});
```

Fit a repeated measures model where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform multivariate analysis of variance.

```
manova(rm)
```

```
ans=8×9 table
     Within       Between      Statistic      Value        F       RSquare    df1    df2      pVa
    ________    ____________    _________    _________    ______    _______    ___    ___    _____

    Constant    (Intercept)     Pillai         0.99013    4847.5    0.99013     3     145    3.788
    Constant    (Intercept)     Wilks        0.0098724    4847.5    0.99013     3     145    3.788
```

```
    Constant    (Intercept)    Hotelling       100.29    4847.5    0.99013     3     145    3.788
    Constant    (Intercept)    Roy             100.29    4847.5    0.99013     3     145    3.788
    Constant    species        Pillai         0.96909    45.749    0.48455     6     292     2.47
    Constant    species        Wilks         0.041153    189.92    0.79714     6     290     2.39
    Constant    species        Hotelling       23.051    555.17    0.92016     6     288    4.660
    Constant    species        Roy              23.04    1121.3     0.9584     3     146    1.477
```

Perform multivariate anova separately for each species.

```
manova(rm,'By','species')
```

```
ans=12×9 table
     Within          Between          Statistic     Value        F       RSquare    df1    df2
    ________    ___________________    _________    ________    ______    _______    ___    ___

    Constant    species=setosa         Pillai         0.9823    2682.7     0.9823     3     145
    Constant    species=setosa         Wilks        0.017698    2682.7     0.9823     3     145
    Constant    species=setosa         Hotelling      55.504    2682.7     0.9823     3     145
    Constant    species=setosa         Roy            55.504    2682.7     0.9823     3     145
    Constant    species=versicolor     Pillai           0.97    1562.8       0.97     3     145
    Constant    species=versicolor     Wilks        0.029999    1562.8       0.97     3     145
    Constant    species=versicolor     Hotelling      32.334    1562.8       0.97     3     145
    Constant    species=versicolor     Roy            32.334    1562.8       0.97     3     145
    Constant    species=virginica      Pillai        0.97261    1716.1    0.97261     3     145
    Constant    species=virginica      Wilks        0.027394    1716.1    0.97261     3     145
    Constant    species=virginica      Hotelling      35.505    1716.1    0.97261     3     145
    Constant    species=virginica      Roy            35.505    1716.1    0.97261     3     145
```

Return Arrays of the Hypothesis Test

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform multivariate analysis of variance. Also return the arrays for constructing the hypothesis test.

```
[manovatbl,A,C,D] = manova(rm)
```

```
manovatbl=8×9 table
     Within       Between      Statistic      Value        F       RSquare    df1    df2      pVa
```

```
    ________    ___________    _________    _________    ______    ________    ___    ___    _____

    Constant    (Intercept)    Pillai         0.99013    4847.5    0.99013     3     145    3.788
    Constant    (Intercept)    Wilks        0.0098724    4847.5    0.99013     3     145    3.788
    Constant    (Intercept)    Hotelling       100.29    4847.5    0.99013     3     145    3.788
    Constant    (Intercept)    Roy             100.29    4847.5    0.99013     3     145    3.788
    Constant    species        Pillai         0.96909    45.749    0.48455     6     292     2.47
    Constant    species        Wilks         0.041153    189.92    0.79714     6     290     2.39
    Constant    species        Hotelling       23.051    555.17    0.92016     6     288    4.666
    Constant    species        Roy              23.04    1121.3     0.9584     3     146    1.477
```

```
A=2×1 cell array
    {[   1 0 0]}
    {2x3 double}
```

```
C = 4×3

     1     0     0
    -1     1     0
     0    -1     1
     0     0    -1
```

```
D = 0
```

Index into matrix A.

```
A{1}
```

```
ans = 1×3

     1     0     0
```

```
A{2}
```

```
ans = 2×3

     0     1     0
     0     0     1
```

Tips

- The multivariate response for each observation (subject) is the vector of repeated measures.
- To test a more general hypothesis `A*B*C = D`, use `coeftest`.

See Also

`anova` | `coeftest` | `fitrm` | `ranova`

Topics

“Model Specification for Repeated Measures Models” on page 9-54
“Multivariate Analysis of Variance for Repeated Measures” on page 9-59

manova1

One-way multivariate analysis of variance

Syntax

```
d = manova1(X,group)
d = manova1(X,group,alpha)
[d,p] = manova1(...)
[d,p,stats] = manova1(...)
```

Description

`d = manova1(X,group)` performs a one-way Multivariate Analysis of Variance (MANOVA) for comparing the multivariate means of the columns of `X`, grouped by `group`. `X` is an *m*-by-*n* matrix of data values, and each row is a vector of measurements on *n* variables for a single observation. `group` is a grouping variable defined as a categorical variable, vector, character array, string array, or cell array of character vectors. Two observations are in the same group if they have the same value in the `group` array. The observations in each group represent a sample from a population.

The function returns `d`, an estimate of the dimension of the space containing the group means. `manova1` tests the null hypothesis that the means of each group are the same *n*-dimensional multivariate vector, and that any difference observed in the sample `X` is due to random chance. If `d = 0`, there is no evidence to reject that hypothesis. If `d = 1`, then you can reject the null hypothesis at the 5% level, but you cannot reject the hypothesis that the multivariate means lie on the same line. Similarly, if `d = 2` the multivariate means may lie on the same plane in *n*-dimensional space, but not on the same line.

`d = manova1(X,group,alpha)` gives control of the significance level, `alpha`. The return value `d` will be the smallest dimension having `p > alpha`, where `p` is a *p*-value for testing whether the means lie in a space of that dimension.

`[d,p] = manova1(...)` also returns a `p`, a vector of *p*-values for testing whether the means lie in a space of dimension 0, 1, and so on. The largest possible dimension is either the dimension of the space, or one less than the number of groups. There is one element of `p` for each dimension up to, but not including, the largest.

If the *i*th *p*-value is near zero, this casts doubt on the hypothesis that the group means lie on a space of *i*-1 dimensions. The choice of a critical *p*-value to determine whether the result is judged statistically significant is left to the researcher and is specified by the value of the input argument `alpha`. It is common to declare a result significant if the *p*-value is less than 0.05 or 0.01.

`[d,p,stats] = manova1(...)` also returns `stats`, a structure containing additional MANOVA results. The structure contains the following fields.

Field	Contents
`W`	Within-groups sum of squares and cross-products matrix
`B`	Between-groups sum of squares and cross-products matrix
`T`	Total sum of squares and cross-products matrix

Field	Contents
dfW	Degrees of freedom for W
dfB	Degrees of freedom for B
dfT	Degrees of freedom for T
lambda	Vector of values of Wilks' lambda test statistic for testing whether the means have dimension 0, 1, etc.
chisq	Transformation of lambda to an approximate chi-square distribution
chisqdf	Degrees of freedom for chisq
eigenval	Eigenvalues of $W^{-1}B$
eigenvec	Eigenvectors of $W^{-1}B$; these are the coefficients for the canonical variables C, and they are scaled so the within-group variance of the canonical variables is 1
canon	Canonical variables C, equal to XC*eigenvec, where XC is X with columns centered by subtracting their means
mdist	A vector of Mahalanobis distances from each point to the mean of its group
gmdist	A matrix of Mahalanobis distances between each pair of group means

The canonical variables C are linear combinations of the original variables, chosen to maximize the separation between groups. Specifically, C(:,1) is the linear combination of the X columns that has the maximum separation between groups. This means that among all possible linear combinations, it is the one with the most significant *F* statistic in a one-way analysis of variance. C(:,2) has the maximum separation subject to it being orthogonal to C(:,1), and so on.

You may find it useful to use the outputs from manova1 along with other functions to supplement your analysis. For example, you may want to start with a grouped scatter plot matrix of the original variables using gplotmatrix. You can use gscatter to visualize the group separation using the first two canonical variables. You can use manovacluster to graph a dendrogram showing the clusters among the group means.

Assumptions

The MANOVA test makes the following assumptions about the data in X:

- The populations for each group are normally distributed.
- The variance-covariance matrix is the same for each population.
- All observations are mutually independent.

Examples

you can use manova1 to determine whether there are differences in the averages of four car characteristics, among groups defined by the country where the cars were made.

```
load carbig
[d,p] = manova1([MPG Acceleration Weight Displacement],...
                Origin)
d =
   3
p =
      0
  0.0000
```

```
  0.0075
  0.1934
```

There are four dimensions in the input matrix, so the group means must lie in a four-dimensional space. `manova1` shows that you cannot reject the hypothesis that the means lie in a 3-D subspace.

References

[1] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

See Also

`anova1` | `canoncorr` | `gplotmatrix` | `gscatter` | `manovacluster`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

manovacluster

Dendrogram of group mean clusters following MANOVA

Syntax

```
manovacluster(stats)
manovacluster(stats,method)
H = manovacluster(stats,method)
```

Description

`manovacluster(stats)` generates a dendrogram plot of the group means after a multivariate analysis of variance (MANOVA). `stats` is the output `stats` structure from `manova1`. The clusters are computed by applying the single linkage method to the matrix of Mahalanobis distances between group means.

See `dendrogram` for more information on the graphical output from this function. The dendrogram is most useful when the number of groups is large.

`manovacluster(stats,method)` uses the specified method in place of single linkage. *`method`* can be one of these values, which identify the methods used to create the cluster hierarchy. (See `linkage` for additional information.)

Method	Description
`'single'`	Shortest distance (default)
`'complete'`	Largest distance
`'average'`	Average distance
`'centroid'`	Centroid distance
`'ward'`	Incremental sum of squares

`H = manovacluster(stats,method)` returns a vector of handles to the lines in the figure.

Examples

Dendrogram of Group Means After MANOVA

Load the sample data.

```
load carbig
```

Define the variable matrix.

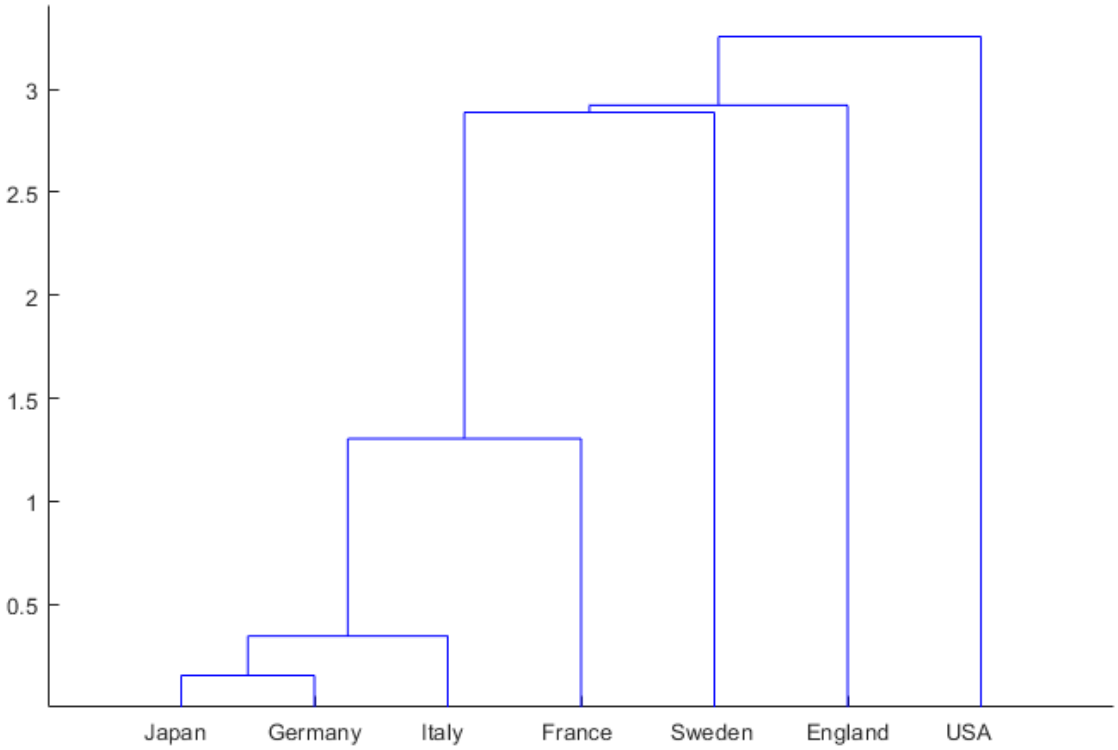
```
X = [MPG Acceleration Weight Displacement];
```

Perform one-way MANOVA to compare the means of MPG, Acceleration, Weight,and Displacement grouped by Origin.

```
[d,p,stats] = manova1(X,Origin);
```

Create a dendrogram plot of the group means.

```
manovacluster(stats)
```



See Also

`cluster` | `dendrogram` | `linkage` | `manova1`

Introduced before R2006a

margin

Package:

Classification margins for generalized additive model (GAM)

Syntax

```
m = margin(Mdl,Tbl,ResponseVarName)
m = margin(Mdl,Tbl,Y)
m = margin(Mdl,X,Y)
m = margin( ___ ,'IncludeInteractions',includeInteractions)
```

Description

`m = margin(Mdl,Tbl,ResponseVarName)` returns the "Classification Margin" on page 33-3869 (`m`) for the generalized additive model `Mdl` using the predictor data in `Tbl` and the true class labels in `Tbl.ResponseVarName`.

`m` is returned as an *n*-by-1 numeric column vector, where *n* is the number of observations in the predictor data.

`m = margin(Mdl,Tbl,Y)` uses the predictor data in table `Tbl` and the true class labels in `Y`.

`m = margin(Mdl,X,Y)` uses the predictor data in matrix `X` and the true class labels in `Y`.

`m = margin( ___ ,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. You can specify `includeInteractions` in addition to any of the input argument combinations in the previous syntaxes.

Examples

Estimate Test Sample Classification Margins and Edge

Estimate the test sample classification margins and edge of a generalized additive model. The test sample margins are the observed true class scores minus the false class scores, and the test sample edge is the mean of the margins.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains two sepal and two petal measurements for versicolor and virginica irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
X = meas(inds,:);
Y = species(inds,:);
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(XTrain,YTrain,'ClassNames',{'versicolor','virginica'});
```

`Mdl` is a `ClassificationGAM` model object.

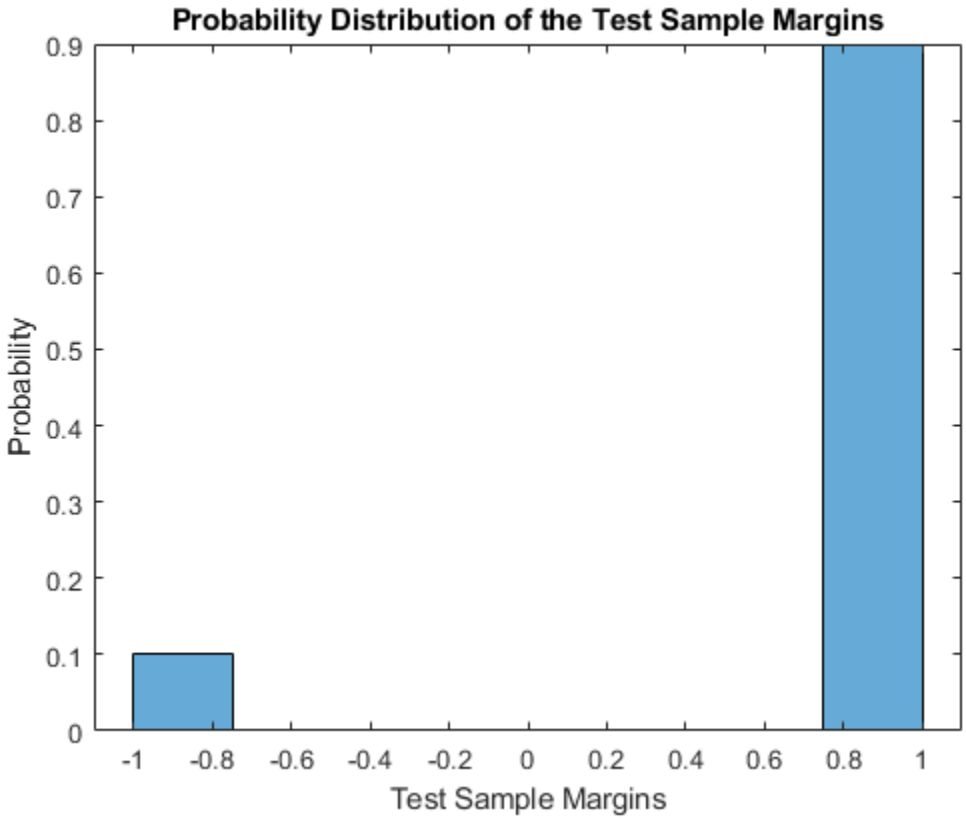
Estimate the test sample classification margins and edge.

```
m = margin(Mdl,XTest,YTest);
e = edge(Mdl,XTest,YTest)
```

```
e = 0.8000
```

Display the histogram of the test sample classification margins.

```
histogram(m,length(unique(m)),'Normalization','probability')
xlabel('Test Sample Margins')
ylabel('Probability')
title('Probability Distribution of the Test Sample Margins')
```



Compare GAMs by Examining Test Sample Margins and Edge

Compare a GAM with linear terms to a GAM with both linear and interaction terms by examining the test sample margins and edge. Based solely on this comparison, the classifier with the highest margins and edge is the best model.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'Holdout',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
```

```
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a GAM that contains both linear and interaction terms for predictors. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
Mdl = fitcgam(XTrain,YTrain,'Interactions','all','MaxPValue',0.05)
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 3.0398
             Interactions: [561x2 double]
          NumObservations: 246


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. `Mdl` includes all available interaction terms.

Estimate the test sample margins and edge for `Mdl`.

```
M = margin(Mdl,XTest,YTest);
E = edge(Mdl,XTest,YTest)
```

```
E = 0.7848
```

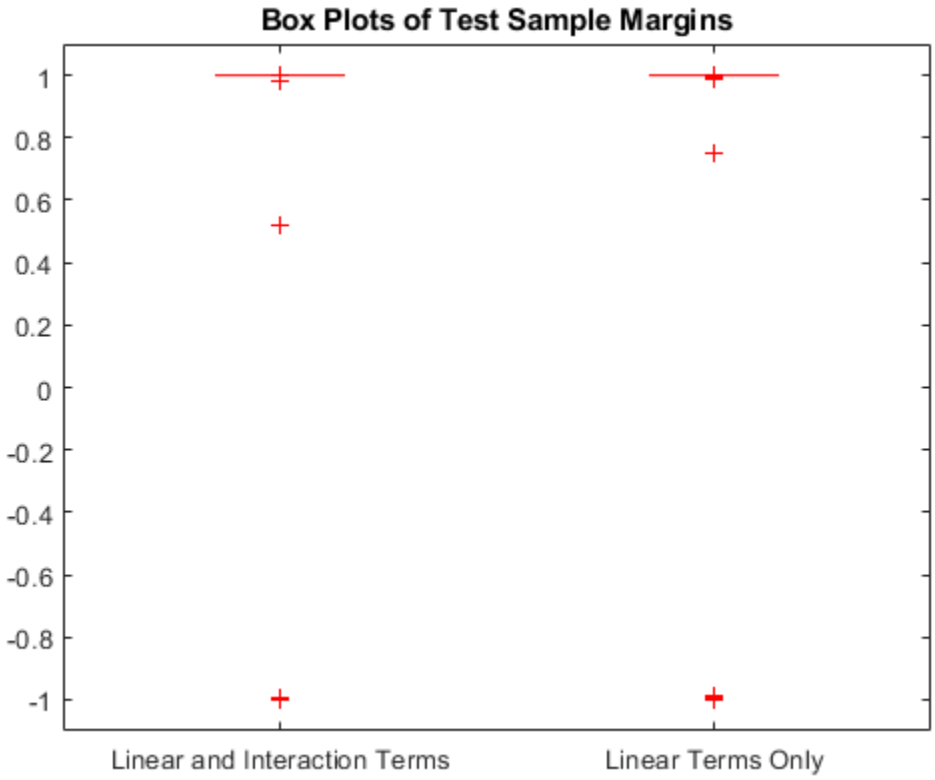
Estimate the test sample margins and edge for `Mdl` without including interaction terms.

```
M_nointeractions = margin(Mdl,XTest,YTest,'IncludeInteractions',false);
E_nointeractions = edge(Mdl,XTest,YTest,'IncludeInteractions',false)
```

```
E_nointeractions = 0.7871
```

Display the distributions of the margins using box plots.

```
boxplot([M M_nointeractions],'Labels',{'Linear and Interaction Terms','Linear Terms Only'})
title('Box Plots of Test Sample Margins')
```



The margins `M` and `M_nointeractions` have a similar distribution, but the test sample edge of the classifier with only linear terms is larger. Classifiers that yield relatively large margins are preferred.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `CompactClassificationGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `CompactClassificationGAM` model object.

- If you trained `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table (`Tbl`).
- If you trained `Mdl` using sample data contained in a matrix, then the input data for `margin` must also be in a matrix (`X`).

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

`Tbl` must contain all the predictors used to train `Mdl`. Optionally, `Tbl` can contain a column for the response variable and a column for the observation weights.

- The response variable must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.) If the response variable in `Tbl` has the same name as the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.
- The weight values must be a numeric vector. You must specify the observation weights in `Tbl` by using `'Weights'`.

If you trained `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as a character vector or string scalar containing the name of the response variable in `Tbl`. For example, if the response variable `Y` is stored in `Tbl.Y`, then specify it as `'Y'`.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X` or `Tbl`.

`Y` must have the same data type as `Mdl.Y`. (The software treats string arrays as cell arrays of character vectors.)

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

If you trained `Mdl` using sample data contained in a matrix, then the input data for `margin` must also be in a matrix.

Data Types: `single` | `double`

`includeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `includeInteractions` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

See Also

`edge` | `loss` | `predict` | `resubMargin`

Topics

"Train Generalized Additive Model for Binary Classification" on page 12-77

Introduced in R2021a

margin

Margin of *k*-nearest neighbor classifier

Syntax

```
m = margin(mdl,tbl,ResponseVarName)
m = margin(mdl,tbl,Y)
m = margin(mdl,X,Y)
```

Description

`m = margin(mdl,tbl,ResponseVarName)` returns the classification margins on page 33-3872 for `mdl` with data `tbl` and classification `tbl.ResponseVarName`. If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName`.

`m` is returned as a numeric vector of length `size(tbl,1)`. Each entry in `m` represents the margin for the corresponding row of `tbl` and the corresponding true class label in `tbl.ResponseVarName`, computed using `mdl`.

`m = margin(mdl,tbl,Y)` returns the classification margins for `mdl` with data `tbl` and classification `Y`.

`m = margin(mdl,X,Y)` returns the classification margins for `mdl` with data `X` and classification `Y`. `m` is returned as a numeric vector of length `size(X,1)`.

Examples

Margin Calculation

Create a *k*-nearest neighbor classifier for the Fisher iris data, where $k = 5$.

Load the Fisher iris data set.

```
load fisheriris
```

Create a classifier for five nearest neighbors.

```
mdl = fitcknn(meas,species,'NumNeighbors',5);
```

Examine the margin of the classifier for a mean observation classified as `'versicolor'`.

```
X = mean(meas);
Y = {'versicolor'};
m = margin(mdl,X,Y)
```

```
m = 1
```

All five nearest neighbors classify as `'versicolor'`.

Input Arguments

`mdl` — *k*-nearest neighbor classifier model

`ClassificationKNN` object

k-nearest neighbor classifier model, specified as a `ClassificationKNN` object.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `mdl` using sample data contained in a `table`, then the input data for `margin` must also be in a table.

Data Types: `table`

`ResponseVarName` — Response variable name

name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. If `tbl` contains the response variable used to train `mdl`, then you do not need to specify `ResponseVarName`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable is stored as `tbl.response`, then specify it as `'response'`. Otherwise, the software treats all columns of `tbl`, including `tbl.response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` represents one observation, and each column represents one variable.

Data Types: `single` | `double`

`Y` — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. Each row of `Y` represents the classification of the corresponding row of `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

More About

Margin

The classification margin for each observation is the difference between the classification score for the true class and the maximal classification score for the false classes.

Score

The score of a classification is the posterior probability of the classification. The posterior probability is the number of neighbors with that classification divided by the number of neighbors. For a more detailed definition that includes weights and prior probabilities, see “Posterior Probability” on page 33-4786.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.
- `margin` executes on a GPU in these cases only:
  - The input argument `X` is a `gpuArray`.
  - The input argument `tbl` contains `gpuArray` elements.
  - The input argument `mdl` was fitted with GPU array input arguments.

See Also
`ClassificationKNN` | `edge` | `fitcknn` | `loss`

Topics
“Classification Using Nearest Neighbors” on page 18-12

Introduced in R2012a

margin

Class: `ClassificationLinear`

Classification margins for linear classification models

Syntax

```
m = margin(Mdl,X,Y)
m = margin(Mdl,X,Y,'ObservationsIn',dimension)

m = margin(Mdl,Tbl,ResponseVarName)
m = margin(Mdl,Tbl,Y)
```

Description

`m = margin(Mdl,X,Y)` returns the classification margins on page 33-3880 for the binary, linear classification model `Mdl` using predictor data in `X` and corresponding class labels in `Y`. `m` contains classification margins for each regularization strength in `Mdl`.

`m = margin(Mdl,X,Y,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

`m = margin(Mdl,Tbl,ResponseVarName)` returns the classification margins for the trained linear classifier `Mdl` using the predictor data in table `Tbl` and the class labels in `Tbl.ResponseVarName`.

`m = margin(Mdl,Tbl,Y)` returns the classification margins for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

Input Arguments

`Mdl` — Binary, linear classification model

`ClassificationLinear` model object

Binary, linear classification model, specified as a `ClassificationLinear` model object. You can create a `ClassificationLinear` model object using `fitclinear`.

`X` — Predictor data

full matrix | sparse matrix

Predictor data, specified as an *n*-by-*p* full or sparse matrix. This orientation of `X` indicates that rows correspond to individual observations, and columns correspond to individual predictor variables.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

Y — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

dimension — Predictor data observation dimension

`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'columns'` or `'rows'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table.

ResponseVarName — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Output Arguments

`m` — Classification margins

numeric column vector | numeric matrix

Classification margins on page 33-3880, returned as a numeric column vector or matrix.

`m` is *n*-by-*L*, where *n* is the number of observations in `X` and *L* is the number of regularization strengths in `Mdl` (that is, `numel(Mdl.Lambda)`).

`m(i,j)` is the classification margin of observation *i* using the trained linear classification model that has regularization strength `Mdl.Lambda(j)`.

Examples

Estimate Test-Sample Margins

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. Specify to hold out 30% of the observations. Optimize the objective function using SpaRSA.

```
rng(1); % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','Holdout',0.30);
CMdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `ClassificationLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Estimate the training- and test-sample margins.

```
mTrain = margin(CMdl,X(trainIdx,:),Ystats(trainIdx));
mTest = margin(CMdl,X(testIdx,:),Ystats(testIdx));
```

Because there is one regularization strength in `CMdl`, `mTrain` and `mTest` are column vectors with lengths equal to the number of training and test observations, respectively.

Plot both sets of margins using box plots.

```
figure;
boxplot([mTrain; mTest],[zeros(size(mTrain,1),1); ones(size(mTest,1),1)], ...
    'Labels',{'Training set','Test set'});
h = gca;
h.YLim = [-5 60];
title 'Training- and Test-Set Margins'
```

The distributions of the margins between the training and test sets appear similar.

Feature Selection Using Test-Sample Margins

One way to perform feature selection is to compare test-sample margins from multiple models. Based solely on this criterion, the classifier with the larger margins is the better classifier.

Load the NLP data set. Preprocess the data as in “Estimate Test-Sample Margins” on page 33-3875.

```
load nlpdata
Ystats = Y == 'stats';
```

```
X = X';
rng(1); % For reproducibility
```

Create a data partition which holds out 30% of the observations for testing.

```
Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition); % Test-set indices
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

`Partition` is a `cvpartition` object that defines the data set partition.

Randomly choose 10% of the predictor variables.

```
p = size(X,1); % Number of predictors
idxPart = randsample(p,ceil(0.1*p));
```

Train two binary, linear classification models: one that uses the all of the predictors and one that uses the random 10%. Optimize the objective function using SpaRSA, and indicate that observations correspond to columns.

```
CVMdl = fitclinear(X,Ystats,'CVPartition',Partition,'Solver','sparsa',...
    'ObservationsIn','columns');
PCVMdl = fitclinear(X(idxPart,:),Ystats,'CVPartition',Partition,'Solver','sparsa',...
    'ObservationsIn','columns');
```

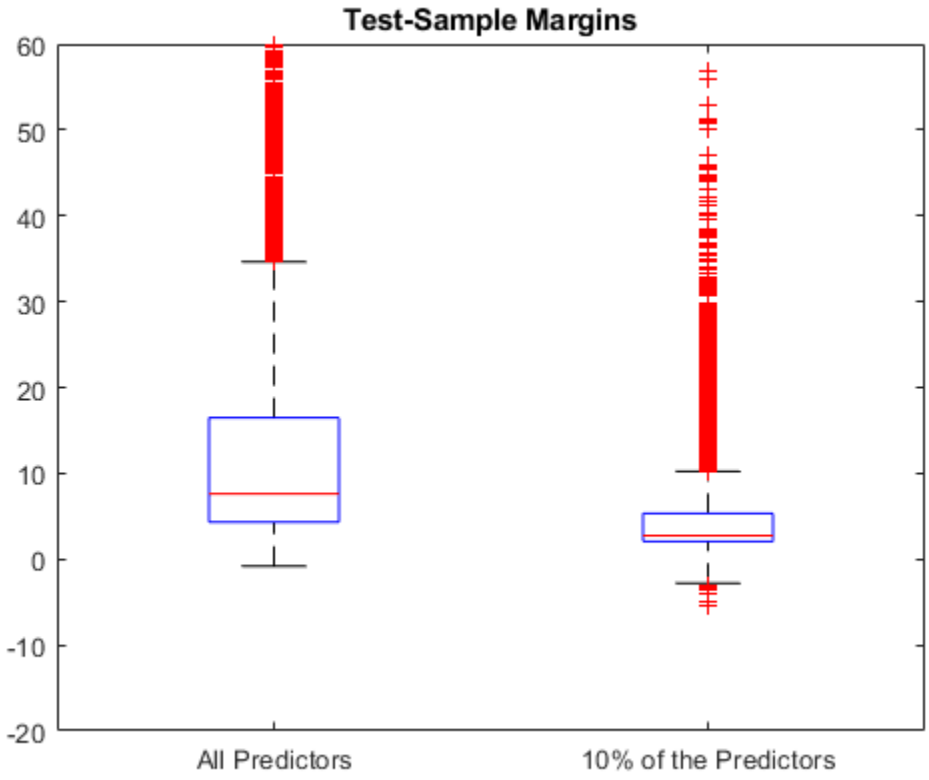
`CVMdl` and `PCVMdl` are `ClassificationPartitionedLinear` models.

Extract the trained `ClassificationLinear` models from the cross-validated models.

```
CMdl = CVMdl.Trained{1};
PCMdl = PCVMdl.Trained{1};
```

Estimate the test sample margins for each classifier. Plot the distribution of the margins sets using box plots.

```
fullMargins = margin(CMdl,XTest,YTest,'ObservationsIn','columns');
partMargins = margin(PCMdl,XTest(idxPart,:),YTest,...
    'ObservationsIn','columns');

figure;
boxplot([fullMargins partMargins],'Labels',...
    {'All Predictors','10% of the Predictors'});
h = gca;
h.YLim = [-20 60];
title('Test-Sample Margins')
```



The margin distribution of `CMdl` is situated higher than the margin distribution of `PCMdl`.

Find Good Lasso Penalty Using Margins

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare distributions of test-sample margins.

Load the NLP data set. Preprocess the data as in “Estimate Test-Sample Margins” on page 33-3875.

```
load nlpdata
Ystats = Y == 'stats';
X = X';

Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition);
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-8} through 10^{1}.

```
Lambda = logspace(-8,1,11);
```

Train binary, linear classification models that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
rng(10); % For reproducibility
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'CVPartition',Partition,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract the trained linear classification model.

```
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the test-sample margins.

```
m = margin(Mdl,X(:,testIdx),Ystats(testIdx),'ObservationsIn','columns');
size(m)
```

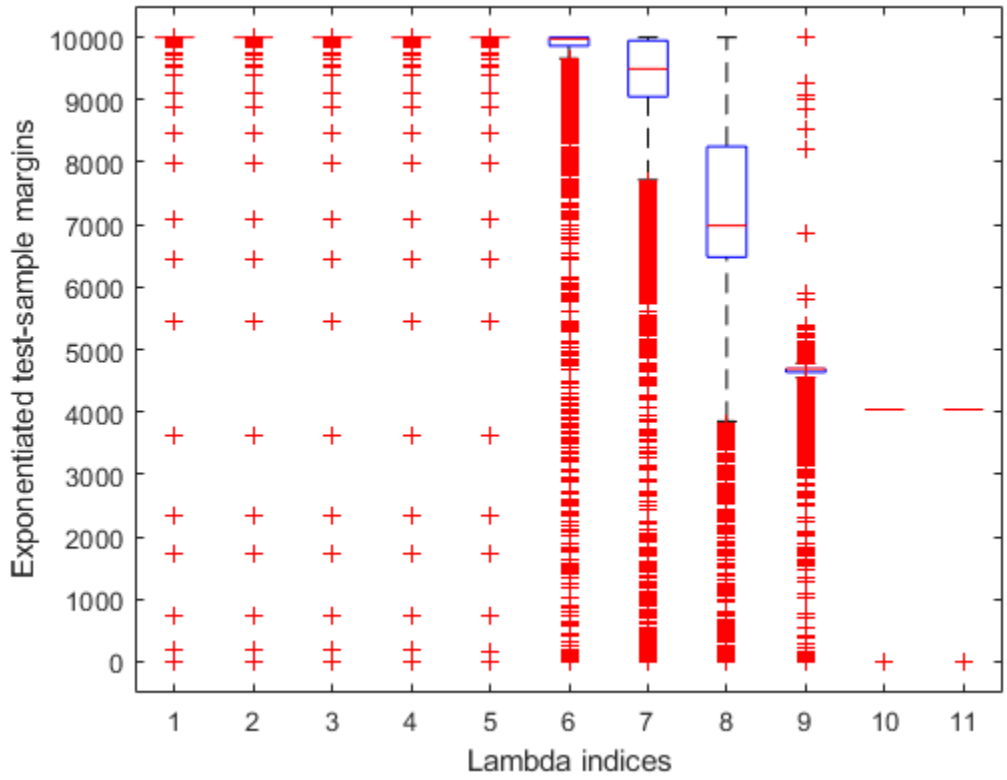
```
ans = 1×2

        9471          11
```

Because there are 11 regularization strengths, `m` has 11 columns.

Plot the test-sample margins for each regularization strength. Because logistic regression scores are in [0,1], margins are in [-1,1]. Rescale the margins to help identify the regularization strength that maximizes the margins over the grid.

```
figure;
boxplot(10000.^m)
```

```
ylabel('Exponentiated test-sample margins')
xlabel('Lambda indices')
```



Several values of `Lambda` yield margin distributions that are compacted near 10000[1]. Higher values of lambda lead to predictor variable sparsity, which is a good quality of a classifier.

Choose the regularization strength that occurs just before the centers of the margin distributions start decreasing.

```
LambdaFinal = Lambda(5);
```

Train a linear classification model using the entire data set and specify the desired regularization strength.

```
MdlFinal = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',LambdaFinal);
```

To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `margin` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationLinear` | `edge` | `fitclinear` | `predict`

Introduced in R2016a

margin

Classification margins

Syntax

```
m = margin(obj,X,Y)
```

Description

`m = margin(obj,X,Y)` returns the classification margins for the matrix of predictors `X` and class labels `Y`. For the definition, see “More About” on page 33-3883.

Input Arguments

`obj`

Discriminant analysis classifier of class `ClassificationDiscriminant` or `CompactClassificationDiscriminant`, typically constructed with `fitcdiscr`.

`X`

Matrix where each row represents an observation, and each column represents a predictor. The number of columns in `X` must equal the number of predictors in `obj`.

`Y`

Class labels, with the same data type as exists in `obj`. The number of elements of `Y` must equal the number of rows of `X`.

Output Arguments

`m`

Numeric column vector of length `size(X,1)`. Each entry in `m` represents the margin for the corresponding rows of `X` and (true class) `Y`, computed using `obj`.

Examples

Compute the classification margin for the Fisher iris data, trained on its first two columns of data, and view the last 10 entries:

```
load fisheriris
X = meas(:,1:2);
obj = fitcdiscr(X,species);
M = margin(obj,X,species);
M(end-10:end)

ans =
    0.6551
    0.4838
```

```
    0.6551
   -0.5127
    0.5659
    0.4611
    0.4949
    0.1024
    0.2787
   -0.1439
   -0.4444
```

The classifier trained on all the data is better:

```
obj = fitcdiscr(meas,species);
M = margin(obj,meas,species);
M(end-10:end)

ans =
    0.9983
    1.0000
    0.9991
    0.9978
    1.0000
    1.0000
    0.9999
    0.9882
    0.9937
    1.0000
    0.9649
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes.

The classification margin is a column vector with the same number of rows as in the matrix `X`. A high value of margin indicates a more reliable prediction than a low value.

Score (discriminant analysis)

For discriminant analysis, the score of a classification is the posterior probability of the classification. For the definition of posterior probability in discriminant analysis, see “Posterior Probability” on page 20-6.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`ClassificationDiscriminant` | `edge` | `fitcdiscr` | `loss` | `predict`

Topics

“Discriminant Analysis Classification” on page 20-2

margin

Package:

Classification margins for multiclass error-correcting output codes (ECOC) model

Syntax

```
m = margin(Mdl,tbl,ResponseVarName)
m = margin(Mdl,tbl,Y)

m = margin(Mdl,X,Y)

m = margin( ___ ,Name,Value)
```

Description

`m = margin(Mdl,tbl,ResponseVarName)` returns the classification margins on page 33-3893 (`m`) for the trained multiclass error-correcting output codes (ECOC) model `Mdl` using the predictor data in table `tbl` and the class labels in `tbl.ResponseVarName`.

`m = margin(Mdl,tbl,Y)` returns the classification margins for the classifier `Mdl` using the predictor data in table `tbl` and the class labels in vector `Y`.

`m = margin(Mdl,X,Y)` returns the classification margins for the classifier `Mdl` using the predictor data in matrix `X` and the class labels `Y`.

`m = margin( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify a decoding scheme, binary learner loss function, and verbosity level.

Examples

Test-Sample Classification Margins of ECOC Model

Calculate the test-sample classification margins of an ECOC model with SVM binary learners.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1)  % For reproducibility
```

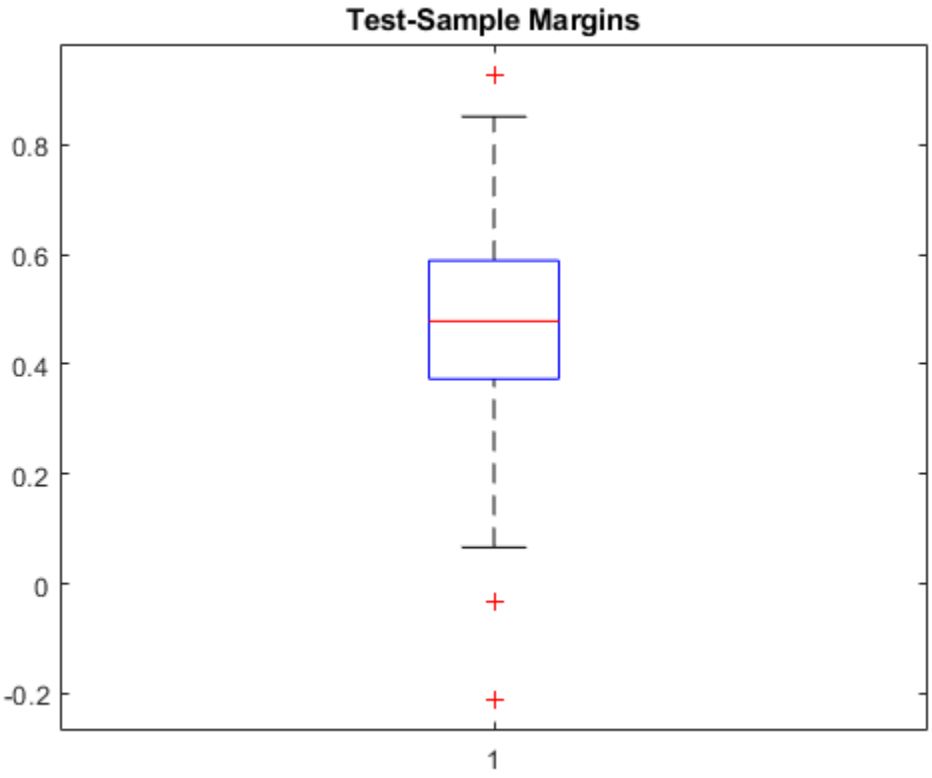
Train an ECOC model using SVM binary classifiers. Specify a 30% holdout sample, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.30,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};    % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

Calculate the test-sample classification margins. Display the distribution of the margins using a boxplot.

```
testInds = test(PMdl.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
m = margin(Mdl,XTest,YTest);

boxplot(m)
title('Test-Sample Margins')
```



The classification margin of an observation is the positive-class negated loss minus the maximum negative-class negated loss. Choose classifiers that yield relatively large margins.

Select ECOC Model Features by Examining Test-Sample Margins

Perform feature selection by comparing test-sample margins from multiple models. Based solely on this comparison, the model with the greatest margins is the best model.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Partition the data set into training and test sets. Specify a 30% holdout sample for testing.

```
Partition = cvpartition(Y,'Holdout',0.30);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`Partition` defines the data set partition.

Define these two data sets:

- `fullX` contains all four predictors.
- `partX` contains the sepal measurements only.

```
fullX = X;
partX = X(:,1:2);
```

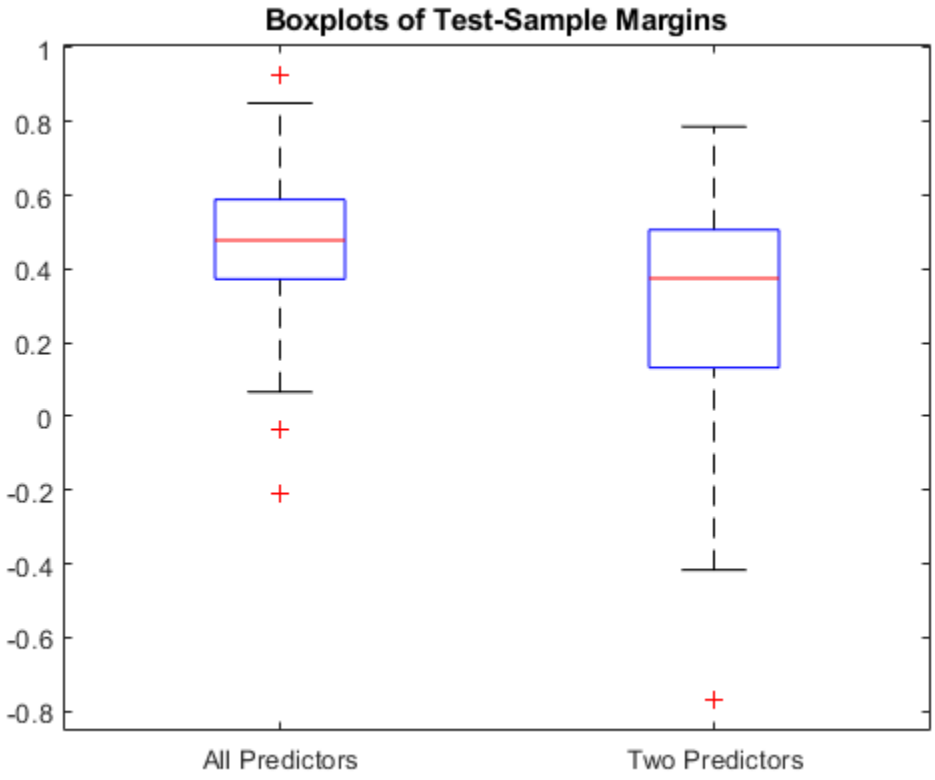
Train an ECOC model using SVM binary classifiers for each predictor set. Specify the partition definition, standardize the predictors using an SVM template, and define the class order.

```
t = templateSVM('Standardize',true);
fullPMdl = fitcecoc(fullX,Y,'CVPartition',Partition,'Learners',t,...
    'ClassNames',classOrder);
partPMdl = fitcecoc(partX,Y,'CVPartition',Partition,'Learners',t,...
    'ClassNames',classOrder);
fullMdl = fullPMdl.Trained{1};
partMdl = partPMdl.Trained{1};
```

`fullPMdl` and `partPMdl` are `ClassificationPartitionedECOC` models. Each model has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the corresponding training set.

Calculate the test-sample margins for each classifier. For each model, display the distribution of the margins using a boxplot.

```
fullMargins = margin(fullMdl,XTest,YTest);
partMargins = margin(partMdl,XTest(:,1:2),YTest);

boxplot([fullMargins partMargins],'Labels',{'All Predictors','Two Predictors'})
title('Boxplots of Test-Sample Margins')
```



The margin distribution of `fullMdl` is situated higher and has less variability than the margin distribution of `partMdl`.

Input Arguments

`Mdl` — Full or compact multiclass ECOC model

`ClassificationECOC` model object | `CompactClassificationECOC` model object

Full or compact multiclass ECOC model, specified as a `ClassificationECOC` or `CompactClassificationECOC` model object.

To create a full or compact ECOC model, see `ClassificationECOC` or `CompactClassificationECOC`.

`tbl` — Sample data

table

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `tbl` can contain additional columns for the response variable and observation weights. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you train `Mdl` using sample data contained in a `table`, then the input data for `margin` must also be in a table.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`. If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `tbl.y`, then specify `ResponseVarName` as `'y'`. Otherwise, the software treats all columns of `tbl`, including `tbl.y`, as predictors.

The response variable must be a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation, and each column corresponds to one variable. The variables in the columns of `X` must be the same as the variables that trained the classifier `Mdl`.

The number of rows in `X` must equal the number of rows in `Y`.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `double` | `single`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `Y` must have the same data type as `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The number of rows in `Y` must equal the number of rows in `tbl` or `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `margin(Mdl,tbl,'y','BinaryLoss','exponential')` specifies an exponential binary learner loss function.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-3892.

Example: `'Decoding','lossbased'`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`. `Mdl.BinaryLearners` must contain `ClassificationLinear` models.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, you can experience a significant reduction in execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`m` — Classification margins
numeric column vector | numeric matrix

Classification margins on page 33-3893, returned as a numeric column vector or numeric matrix.

If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `m` is an n-by-L vector, where n is the number of observations in `X` and L is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`). The value `m(i,j)` is the margin of observation `i` for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.

Otherwise, `m` is a column vector of length n.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} \left|m_{kj}\right| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|}.$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Tips

- To compare the margins or edges of several ECOC classifiers, use template objects to specify a common score transform function among the classifiers during training.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `margin` does not support tall `table` data when `Mdl` contains kernel or linear binary learners.

For more information, see “Tall Arrays”.

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `edge` | `fitcecoc` | `loss` | `predict` | `resubMargin`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

margin

Classification margins

Syntax

```
M = margin(ens,tbl,ResponseVarName)
M = margin(ens,tbl,Y)
M = margin(ens,X,Y)
M = margin(___,Name,Value)
```

Description

`M = margin(ens,tbl,ResponseVarName)` returns the classification margin for the predictions of `ens` on data `tbl`, when the true classifications are `tbl.ResponseVarName`.

`M = margin(ens,tbl,Y)` returns the classification margin for the predictions of `ens` on data `tbl`, when the true classifications are `Y`.

`M = margin(ens,X,Y)` returns the classification margin for the predictions of `ens` on data `X`, when the true classifications are `Y`.

`M = margin(___,Name,Value)` calculates margin with additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes.

Input Arguments

`ens`

Classification ensemble created with `fitcensemble`, or a compact classification ensemble created with `compact`.

`tbl`

Sample data, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all of the predictors used to train the model. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `ens` using sample data contained in a `table`, then the input data for this method must also be in a table.

`ResponseVarName`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `tbl`, including `Y`, as predictors when training the model.

X

Matrix of data to classify. Each row of X represents one observation, and each column represents one predictor. X must have the same number of columns as the data used to train ens. X should have the same number of rows as the number of elements in Y.

If you trained ens using sample data contained in a matrix, then the input data for this method must also be in a matrix.

Y

Class labels of observations in tbl or X. Y should be of the same type as the classification used to train ens, and its number of elements should equal the number of rows of tbl or X.

Name-Value Pair Arguments

Specify optional comma-separated pairs of Name,Value arguments. Name is the argument name and Value is the corresponding value. Name must appear inside quotes. You can specify several name and value pair arguments in any order as Name1,Value1,...,NameN,ValueN.

learners

Indices of weak learners in the ensemble ranging from 1 to ens.NumTrained. oobEdge uses only these learners for calculating loss.

Default: 1:NumTrained

UseObsForLearner

A logical matrix of size N-by-T, where:

- N is the number of rows of X.
- T is the number of weak learners in ens.

When UseObsForLearner(i,j) is true, learner j is used in predicting the class of row i of X.

Default: true(N,T)

Output Arguments

M

A numeric column vector with the same number of rows as tbl or X. Each row of M gives the classification margin for that row of tbl or X.

Examples

Find Classification Margin

Find the margin for classifying an average flower from the fisheriris data as 'versicolor'.

Load the Fisher iris data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Classify an average flower and find the classification margin.

```
flower = mean(meas);
predict(ens,flower)
```

```
ans = 1x1 cell array
    {'versicolor'}
```

```
margin(ens,flower,'versicolor')
```

```
ans = 3.2140
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `X`.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`edge` | `loss` | `predict`

margin

Classification margins for naive Bayes classifier

Syntax

```
m = margin(Mdl,tbl,ResponseVarName)
m = margin(Mdl,tbl,Y)
m = margin(Mdl,X,Y)
```

Description

`m = margin(Mdl,tbl,ResponseVarName)` returns the "Classification Margin" on page 33-3904 (`m`) for the trained naive Bayes classifier `Mdl` using the predictor data in table `tbl` and the class labels in `tbl.ResponseVarName`.

`m = margin(Mdl,tbl,Y)` returns the classification margins for `Mdl` using the predictor data in table `tbl` and the class labels in vector `Y`.

`m = margin(Mdl,X,Y)` returns the classification margins for `Mdl` using the predictor data in matrix `X` and the class labels in `Y`.

`m` is returned as a numeric vector with the same length as `Y`. The software estimates each entry of `m` using the trained naive Bayes classifier `Mdl`, the corresponding row of `X`, and the true class label `Y`.

Examples

Estimate Test Sample Classification Margins of Naive Bayes Classifier

Estimate the test sample classification margins of a naive Bayes classifier. An observation margin is the observed true class score minus the maximum false class score among all scores in the respective class.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 105
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

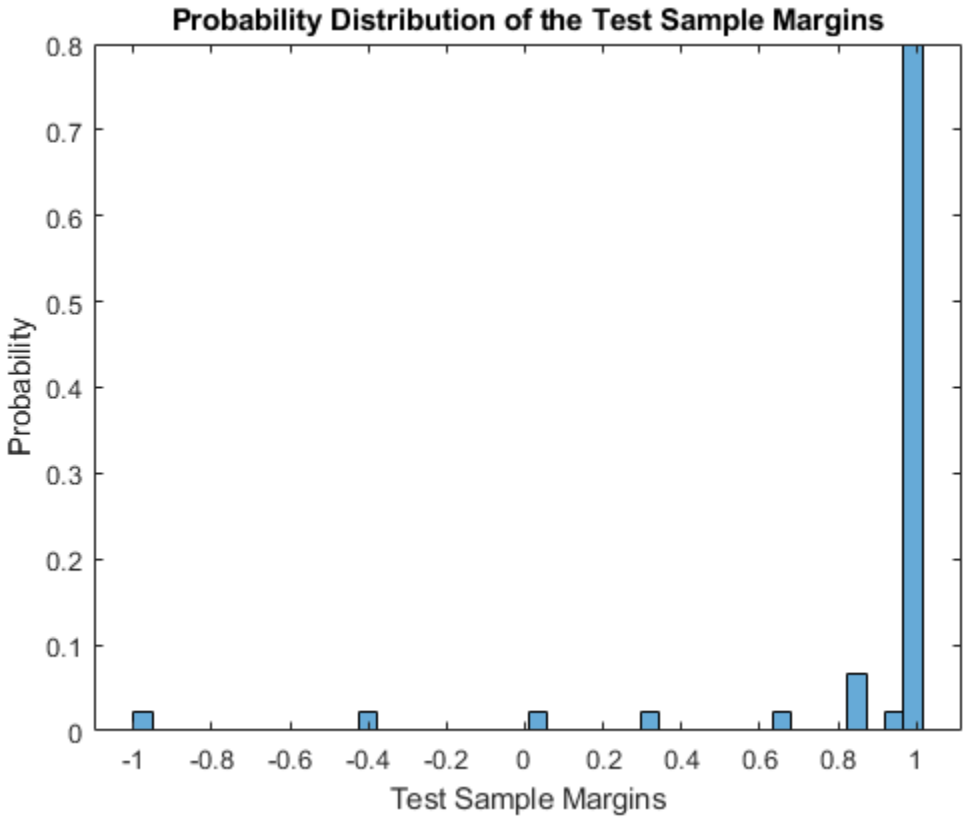
Estimate the test sample classification margins.

```
m = margin(Mdl,XTest,YTest);
median(m)
```

```
ans = 1.0000
```

Display the histogram of the test sample classification margins.

```
histogram(m,length(unique(m)),'Normalization','probability')
xlabel('Test Sample Margins')
ylabel('Probability')
title('Probability Distribution of the Test Sample Margins')
```



Classifiers that yield relatively large margins are preferred.

Select Naive Bayes Classifier Features by Examining Test Sample Margins

Perform feature selection by comparing test sample margins from multiple models. Based solely on this comparison, the classifier with the highest margins is the best model.

Load the `fisheriris` data set. Specify the predictors `X` and class labels `Y`.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing. `Partition` defines the data set partition.

```
cv = cvpartition(Y,'Holdout',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Define these two data sets:

- `fullX` contains all predictors.
- `partX` contains the last two predictors.

```
fullX = XTrain;
partX = XTrain(:,3:4);
```

Train a naive Bayes classifier for each predictor set.

```
fullMdl = fitcnb(fullX,YTrain);
partMdl = fitcnb(partX,YTrain);
```

`fullMdl` and `partMdl` are trained `ClassificationNaiveBayes` classifiers.

Estimate the test sample margins for each classifier.

```
fullM = margin(fullMdl,XTest,YTest);
median(fullM)
```

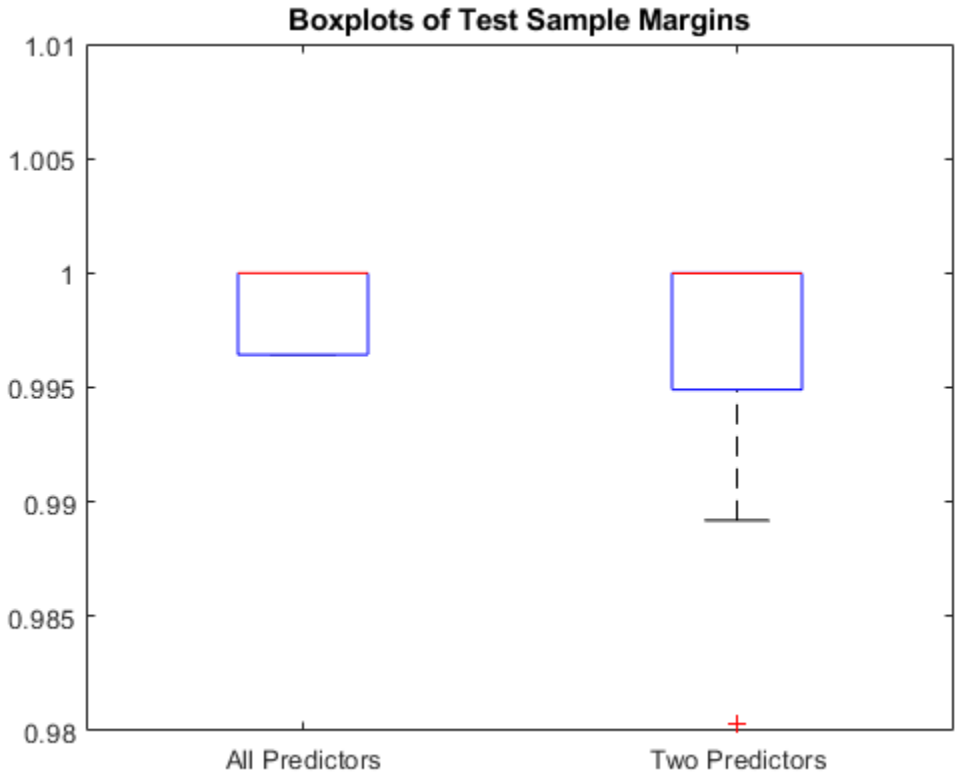
```
ans = 1.0000
```

```
partM = margin(partMdl,XTest(:,3:4),YTest);
median(partM)
```

```
ans = 1.0000
```

Display the distribution of the margins for each model using boxplots.

```
boxplot([fullM partM],'Labels',{'All Predictors','Two Predictors'})
ylim([0.98 1.01]) % Modify the y-axis limits to see the boxes
title('Boxplots of Test Sample Margins')
```



The margins for `fullMdl` (all predictors model) and `partMdl` (two predictors model) have a similar distribution with the same median. `partMdl` is less complex but has outliers.

Input Arguments

`Mdl` — Naive Bayes classification model

`ClassificationNaiveBayes` model object | `CompactClassificationNaiveBayes` model object

Naive Bayes classification model, specified as a `ClassificationNaiveBayes` model object or `CompactClassificationNaiveBayes` model object returned by `fitcnb` or `compact`, respectively.

`tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `tbl` corresponds to one observation, and each column corresponds to one predictor variable. `tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed. Optionally, `tbl` can contain additional columns for the response variable and observation weights.

If you train `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table.

`ResponseVarName` — Response variable name

name of a variable in `tbl`

Response variable name, specified as the name of a variable in `tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `y` is stored as `tbl.y`, then specify it as `'y'`. Otherwise, the software treats all columns of `tbl`, including `y`, as predictors.

If `tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an *instance* or *example*), and each column corresponds to one variable (also known as a *feature*). The variables in the columns of `X` must be the same as the variables that trained the `Mdl` classifier.

The length of `Y` and the number of rows of `X` must be equal.

Data Types: `double` | `single`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. `Y` must have the same data type as `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The length of `Y` must be equal to the number of rows of `tbl` or `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

If you supply weights, then the software normalizes them to sum to the prior probability of their respective class. The software uses the normalized weights to compute the weighted mean.

When choosing among multiple classifiers to perform a task such as feature section, choose the classifier that yields the highest edge.

Classification Margin

The classification margin for each observation is the difference between the score for the true class and the maximal score for the false classes. Margins provide a classification confidence measure; among multiple classifiers, those that yield larger margins (on the same scale) are better.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For naive Bayes, the posterior probability that a classification is k for a given observation $(x_1,...,x_P)$ is

$$\widehat{P}(Y = k \,|\, x_1, .., x_P) = \frac{P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)}{P(X_1, ..., X_P)},$$

where:

- $P(X_1, ..., X_P \,|\, y = k)$ is the conditional joint density of the predictors given they are in class k. `Mdl.DistributionNames` stores the distribution names of the predictors.
- $\pi(Y = k)$ is the class prior probability distribution. `Mdl.Prior` stores the prior distribution.
- $P(X_1, .., X_P)$ is the joint density of the predictors. The classes are discrete, so $P(X_1, ..., X_P) = \sum_{k=1}^{K} P(X_1, ..., X_P \,|\, y = k)\pi(Y = k)$.

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Score

The naive Bayes score is the class posterior probability given the observation.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`ClassificationNaiveBayes` | `CompactClassificationNaiveBayes` | `edge` | `fitcnb` | `loss` | `predict`

Topics

“Naive Bayes Classification” on page 21-2

Introduced in R2014b

margin

Package:

Classification margins for neural network classifier

Syntax

```
m = margin(Mdl,Tbl,ResponseVarName)
m = margin(Mdl,Tbl,Y)

m = margin(Mdl,X,Y)
m = margin(Mdl,X,Y,'ObservationsIn',dimension)
```

Description

`m = margin(Mdl,Tbl,ResponseVarName)` returns the classification margins on page 33-3912 for the trained neural network classifier `Mdl` using the predictor data in table `Tbl` and the class labels in the `ResponseVarName` table variable.

`m` is returned as a numeric vector, whose *i*th entry corresponds to the *i*th observation in `Tbl`.

`m = margin(Mdl,Tbl,Y)` returns the classification margins for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

`m = margin(Mdl,X,Y)` returns the classification margins for the trained neural network classifier `Mdl` using the predictor data `X` and the corresponding class labels in `Y`.

`m` is returned as a numeric vector, whose *i*th entry corresponds to the *i*th observation in `X`.

`m = margin(Mdl,X,Y,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'column'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

Examples

Test Set Classification Margins of Neural Network

Calculate the test set classification margins of a neural network classifier.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

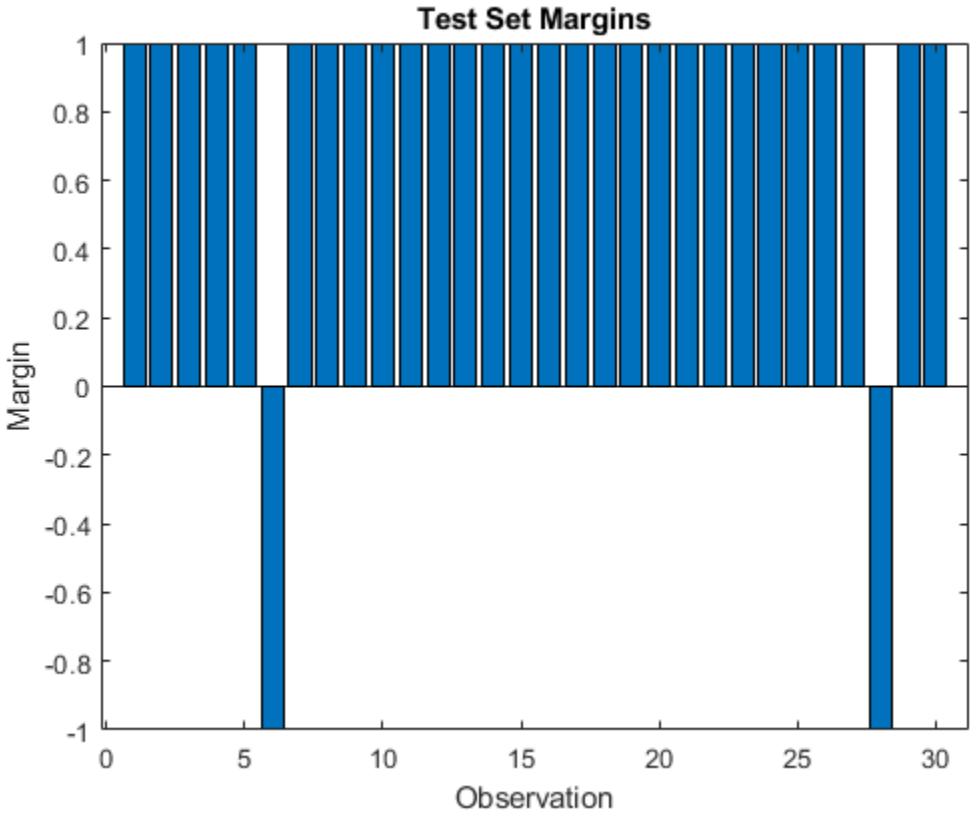
```
rng("default") % For reproducibility of the partition
c = cvpartition(tbl.Smoker,"Holdout",0.30);
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a neural network classifier using the training set. Specify the `Smoker` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitcnet(tblTrain,"Smoker", ...
    "Standardize",true);
```

Calculate the test set classification margins. Because the test set includes only 30 observations, display the margins using a bar graph.

```
m = margin(Mdl,tblTest,"Smoker");
bar(m)
xlabel("Observation")
ylabel("Margin")
title("Test Set Margins")
```



Only the sixth and twenty-eighth observations have negative margins, which indicates that the model performs well overall.

Select Features to Include in Neural Network Classifier

Perform feature selection by comparing test set classification margins, edges, errors, and predictions. Compare the test set metrics for a model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(fishertable.Species,"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

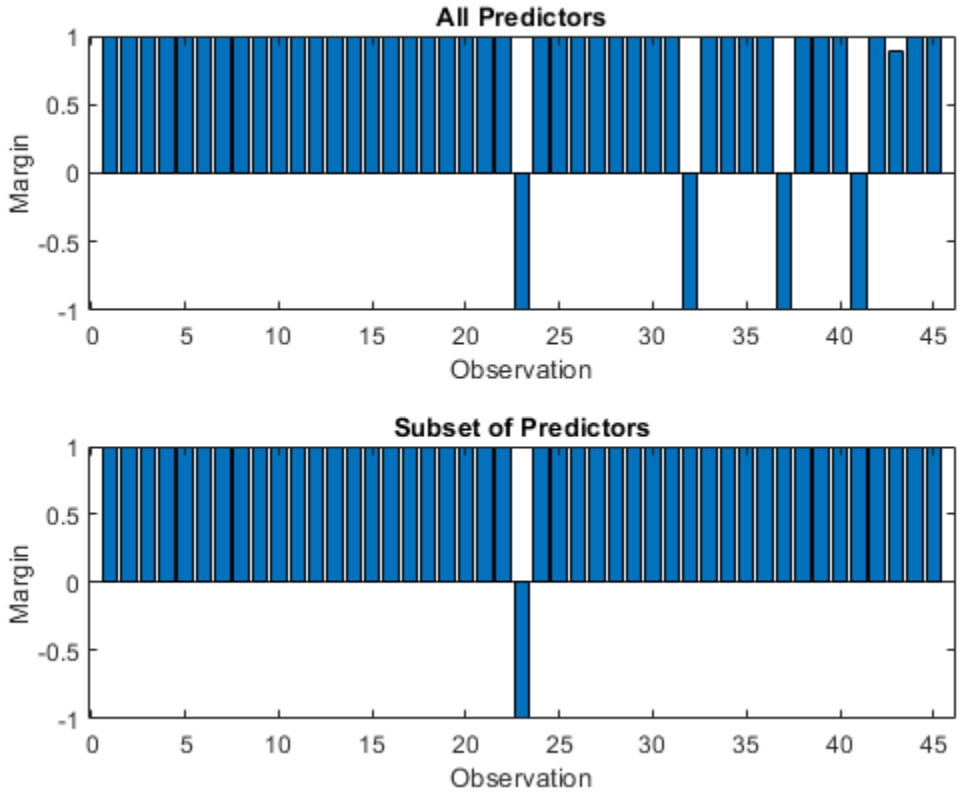
Train one neural network classifier using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `Species` as the response variable, and standardize the predictors.

```
allMdl = fitcnet(trainTbl,"Species","Standardize",true);
subsetMdl = fitcnet(trainTbl,"Species ~ SepalLength + SepalWidth + PetalLength", ...
    "Standardize",true);
```

Calculate the test set classification margins for the two models. Because the test set includes only 45 observations, display the margins using bar graphs.

For each observation, the classification margin is the difference between the classification score for the true class and the maximal score for the false classes. Because neural network classifiers return classification scores that are posterior probabilities, margin values close to 1 indicate confident classifications and negative margin values indicate misclassifications.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allMargins = margin(allMdl,testTbl);
bar(ax1,allMargins)
xlabel(ax1,"Observation")
ylabel(ax1,"Margin")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
subsetMargins = margin(subsetMdl,testTbl);
bar(ax2,subsetMargins)
xlabel(ax2,"Observation")
ylabel(ax2,"Margin")
title(ax2,"Subset of Predictors")
```



Compare the test set classification edge, or mean of the classification margins, of the two models.

```
allEdge = edge(allMdl,testTbl)
```

```
allEdge = 0.8198
```

```
subsetEdge = edge(subsetMdl,testTbl)
```

```
subsetEdge = 0.9556
```

Based on the test set classification margins and edges, the model trained on a subset of the predictors seems to outperform the model trained on all the predictors.

Compare the test set classification error of the two models.

```
allError = loss(allMdl,testTbl);
allAccuracy = 1-allError
```

```
allAccuracy = 0.9111
```

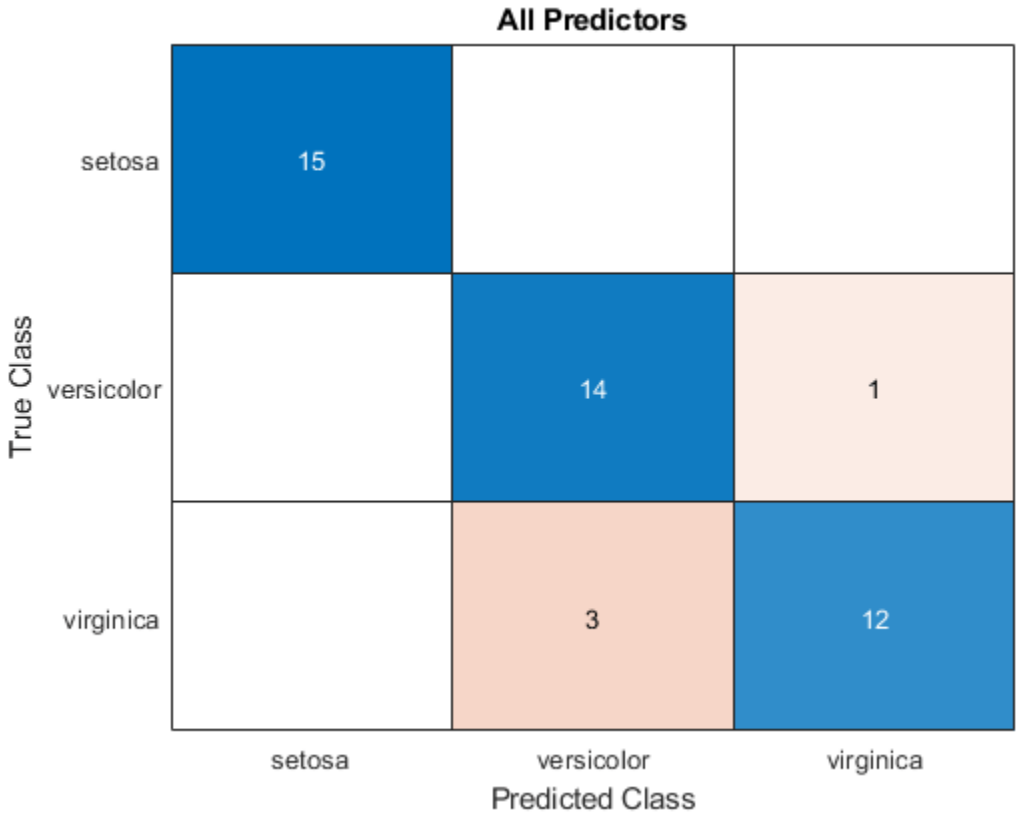
```
subsetError = loss(subsetMdl,testTbl);
subsetAccuracy = 1-subsetError
```

```
subsetAccuracy = 0.9778
```

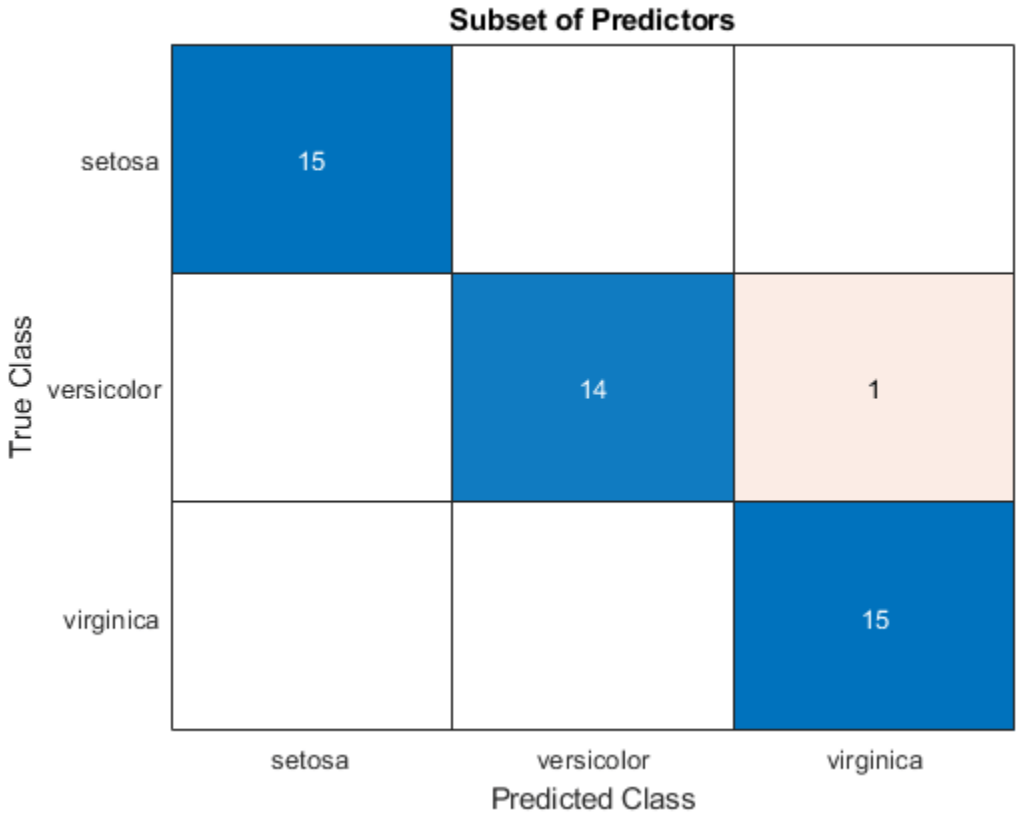
Again, the model trained using only a subset of the predictors seems to perform better than the model trained using all the predictors.

Visualize the test set classification results using confusion matrices.

```
allLabels = predict(allMdl,testTbl);
figure
confusionchart(testTbl.Species,allLabels)
title("All Predictors")
```




```
subsetLabels = predict(subsetMdl,testTbl);
figure
confusionchart(testTbl.Species,subsetLabels)
title("Subset of Predictors")
```



The model trained using all the predictors misclassifies four of the test set observations. The model trained using a subset of the predictors misclassifies only one of the test set observations.

Given the test set performance of the two models, consider using the model trained using all the predictors except `PetalWidth`.

Input Arguments

`Mdl` — Trained neural network classifier

`ClassificationNeuralNetwork` model object | `CompactClassificationNeuralNetwork` model object

Trained neural network classifier, specified as a `ClassificationNeuralNetwork` model object or `CompactClassificationNeuralNetwork` model object returned by `fitcnet` or `compact`, respectively.

`Tbl` — Sample data

table

Sample data, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain an additional column for the response variable. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.
- If you trained `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table.
- If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `table`

`ResponseVarName` — Response variable name
name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`X` — Predictor data
numeric matrix

Predictor data, specified as a numeric matrix. By default, `margin` assumes that each row of `X` corresponds to one observation, and each column corresponds to one predictor variable.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time.

The length of `Y` and the number of observations in `X` must be equal.

If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Data Types: `single` | `double`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

More About

Classification Edge

The classification edge is the mean of the classification margins.

One way to choose among multiple classifiers, for example, to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

See Also

`ClassificationNeuralNetwork` | `CompactClassificationNeuralNetwork` | `edge` | `fitcnet` | `loss` | `predict`

Topics

“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

margin

Package: `classreg.learning.classif`

Find classification margins for support vector machine (SVM) classifier

Syntax

```
m = margin(SVMModel,TBL,ResponseVarName)
m = margin(SVMModel,TBL,Y)

m = margin(SVMModel,X,Y)
```

Description

`m = margin(SVMModel,TBL,ResponseVarName)` returns the classification margins on page 33-3916 (`m`) for the trained support vector machine (SVM) classifier `SVMModel` using the sample data in table `TBL` and the class labels in `TBL.ResponseVarName`.

`m` is returned as a numeric vector with the same length as `Y`. The software estimates each entry of `m` using the trained SVM classifier `SVMModel`, the corresponding row of `X`, and the true class label `Y`.

`m = margin(SVMModel,TBL,Y)` returns the classification margins (`m`) for the trained SVM classifier `SVMModel` using the sample data in table `TBL` and the class labels in `Y`.

`m = margin(SVMModel,X,Y)` returns the classification margins for `SVMModel` using the predictor data in matrix `X` and the class labels in `Y`.

Examples

Estimate Test Sample Classification Margins of SVM Classifiers

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify a 15% holdout sample for testing, standardize the data, and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Holdout',0.15,'ClassNames',{'b','g'},...
    'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1}; ...
    % Extract the trained, compact classifier
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`CVSVMModel` is a `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Estimate the test sample classification margins.

```
m = margin(CompactSVMModel,XTest,YTest);
m(10:20)
```

```
ans = 11×1

    3.5457
    5.5941
    4.9948
    4.5614
   -4.7970
    5.5122
   -2.8774
    1.8671
    9.5002
    9.5035
      ⋮
```

An observation margin is the observed true class score minus the maximum false class score among all scores in the respective class. Classifiers that yield relatively large margins are preferred.

Select SVM Classifier Features by Examining Test Sample Margins

Perform feature selection by comparing test sample margins from multiple models. Based solely on this comparison, the model with the highest margins is the best model.

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Partition the data set into training and test sets. Specify a 15% holdout sample for testing.

```
Partition = cvpartition(Y,'Holdout',0.15);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`Partition` defines the data set partition.

Define these two data sets:

- `fullX` contains all predictors (except the removed column of 0s).
- `partX` contains the last 20 predictors.

```
fullX = X;
partX = X(:,end-20:end);
```

Train SVM classifiers for each predictor set. Specify the partition definition.

```
FullCVSVMModel = fitcsvm(fullX,Y,'CVPartition',Partition);
PartCVSVMModel = fitcsvm(partX,Y,'CVPartition',Partition);
FCSVMModel = FullCVSVMModel.Trained{1};
PCSVMModel = PartCVSVMModel.Trained{1};
```

`FullCVSVMModel` and `PartCVSVMModel` are `ClassificationPartitionedModel` classifiers. They contain the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Estimate the test sample margins for each classifier.

```
fullM = margin(FCSVMModel,XTest,YTest);
partM = margin(PCSVMModel,XTest(:,end-20:end),YTest);
n = size(XTest,1);
p = sum(fullM < partM)/n
```

```
p = 0.2500
```

Approximately 25% of the margins from the full model are less than those from the model with fewer predictors. This result suggests that the model trained with all the predictors is better.

Input Arguments

`SVMModel` — SVM classification model

`ClassificationSVM` model object | `CompactClassificationSVM` model object

SVM classification model, specified as a `ClassificationSVM` model object or `CompactClassificationSVM` model object returned by `fitcsvm` or `compact`, respectively.

`TBL` — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all of the predictors used to train `SVMModel`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName` or `Y`.

If you trained `SVMModel` using sample data contained in a table, then the input data for `margin` must also be in a table.

If you set `'Standardize',true` in `fitcsvm` when training `SVMModel`, then the software standardizes the columns of the predictor data using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `table`

`X` — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix.

Each row of `X` corresponds to one observation (also known as an instance or example), and each column corresponds to one variable (also known as a feature). The variables in the columns of `X` must be the same as the variables that trained the `SVMModel` classifier.

The length of `Y` and the number of rows in `X` must be equal.

If you set `'Standardize',true` in `fitcsvm` to train `SVMModel`, then the software standardizes the columns of `X` using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `double` | `single`

`ResponseVarName` — Response variable name
name of variable in `TBL`

Response variable name, specified as the name of a variable in `TBL`. If `TBL` contains the response variable used to train `SVMModel`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `TBL.Response`, then specify `ResponseVarName` as `'Response'`. Otherwise, the software treats all columns of `TBL`, including `TBL.Response`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. `Y` must be the same as the data type of `SVMModel.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

The length of `Y` must equal the number of rows in `TBL` or the number of rows in `X`.

More About

Classification Edge

The edge is the weighted mean of the classification margins.

The weights are the prior class probabilities. If you supply weights, then the software normalizes them to sum to the prior probabilities in the respective classes. The software uses the renormalized weights to compute the weighted mean.

One way to choose among multiple classifiers, for example, to perform feature selection, is to choose the classifier that yields the highest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

The positive class classification score $f(x)$ is the trained SVM classification function. $f(x)$ is also the numerical predicted response for x, or the score for predicting x into the positive class.

$$f(x) = \sum_{j=1}^{n} \alpha_j y_j G(x_j, x) + b,$$

where $(\alpha_1, ..., \alpha_n, b)$ are the estimated SVM parameters, $G(x_j, x)$ is the dot product in the predictor space between x and the support vectors, and the sum includes the training set observations. The negative class classification score for x, or the score for predicting x into the negative class, is $-f(x)$.

If $G(x_j,x) = x_j'x$ (the linear kernel), then the score function reduces to

$$f(x) = (x/s)'\beta + b\,.$$

s is the kernel scale and β is the vector of fitted linear coefficients.

For more details, see “Understanding Support Vector Machines” on page 18-150.

Algorithms

For binary classification, the software defines the margin for observation j, m_j, as

$$m_j = 2y_j f(x_j),$$

where $y_j \in \{-1,1\}$, and $f(x_j)$ is the predicted score of observation j for the positive class. However, $m_j = y_j f(x_j)$ is commonly used to define the margin.

References

[1] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationSVM` | `CompactClassificationSVM` | `edge` | `fitcsvm` | `loss` | `predict`

Introduced in R2014a

margin

Classification margins

Syntax

```
m = margin(tree,TBL,ResponseVarName)
m = margin(tree,TBL,Y)
m = margin(tree,X,Y)
```

Description

`m = margin(tree,TBL,ResponseVarName)` returns the classification margins for the table of predictors `TBL` and class labels `TBL.ResponseVarName`. For the definition, see “Margin” on page 33-3921.

`m = margin(tree,TBL,Y)` returns the classification margins for the table of predictors `TBL` and class labels `Y`.

`m = margin(tree,X,Y)` returns the classification margins for the matrix of predictors `X` and class labels `Y`.

Input Arguments

`tree` — Trained classification tree

`ClassificationTree` model object | `CompactClassificationTree` model object

Trained classification tree, specified as a `ClassificationTree` or `CompactClassificationTree` model object. That is, `tree` is a trained classification model returned by `fitctree` or `compact`.

`TBL` — Sample data

table

Sample data, specified as a table. Each row of `TBL` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `TBL` can contain additional columns for the response variable and observation weights. `TBL` must contain all the predictors used to train `tree`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `TBL` contains the response variable used to train `tree`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `tree` using sample data contained in a `table`, then the input data for this method must also be in a table.

Data Types: `table`

`X` — Data to classify

numeric matrix

Data to classify, specified as a numeric matrix. Each row of `X` represents one observation, and each column represents one predictor. `X` must have the same number of columns as the data used to train `tree`. `X` must have the same number of rows as the number of elements in `Y`.

Data Types: `single` | `double`

`ResponseVarName` — Response variable name
name of a variable in `TBL`

Response variable name, specified as the name of a variable in `TBL`. If `TBL` contains the response variable used to train `tree`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must do so as a character vector or string scalar. For example, if the response variable is stored as `TBL.Response`, then specify it as `'Response'`. Otherwise, the software treats all columns of `TBL`, including `TBL.ResponseVarName`, as predictors.

The response variable must be a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`Y` — Class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, a logical or numeric vector, or a cell array of character vectors. `Y` must be of the same type as the classification used to train `tree`, and its number of elements must equal the number of rows of `X`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Output Arguments

`m` — Margin
numeric column vector

Margin, returned as a numeric column vector of length `size(X,1)`. Each entry in `m` represents the margin for the corresponding rows of `X` and (true class) `Y`, computed using `tree`.

Examples

Compute the classification margin for the Fisher iris data, trained on its first two columns of data, and view the last 10 entries.

```
load fisheriris
X = meas(:,1:2);
tree = fitctree(X,species);
M = margin(tree,X,species);
M(end-10:end)
```

```
ans =
    0.1111
    0.1111
    0.1111
   -0.2857
    0.6364
    0.6364
    0.1111
```

```
    0.7500
    1.0000
    0.6364
    0.2000
```

The classification tree trained on all the data is better.

```
tree = fitctree(meas,species);
M = margin(tree,meas,species);
M(end-10:end)
```

```
ans =
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
    0.9565
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix X.

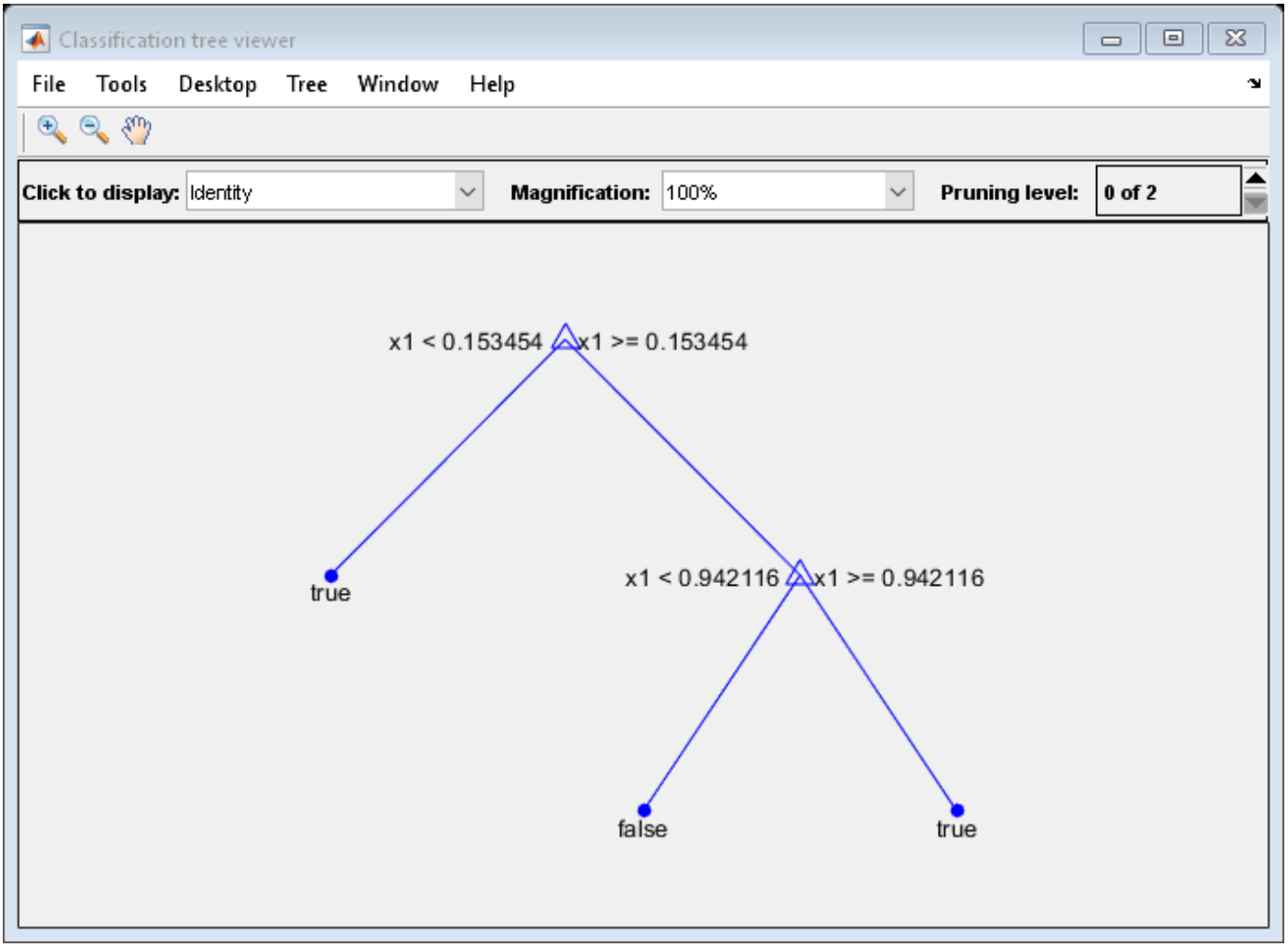
Score (tree)

For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor X as `true` when $X < 0.15$ or $X > 0.95$, and X is false otherwise.

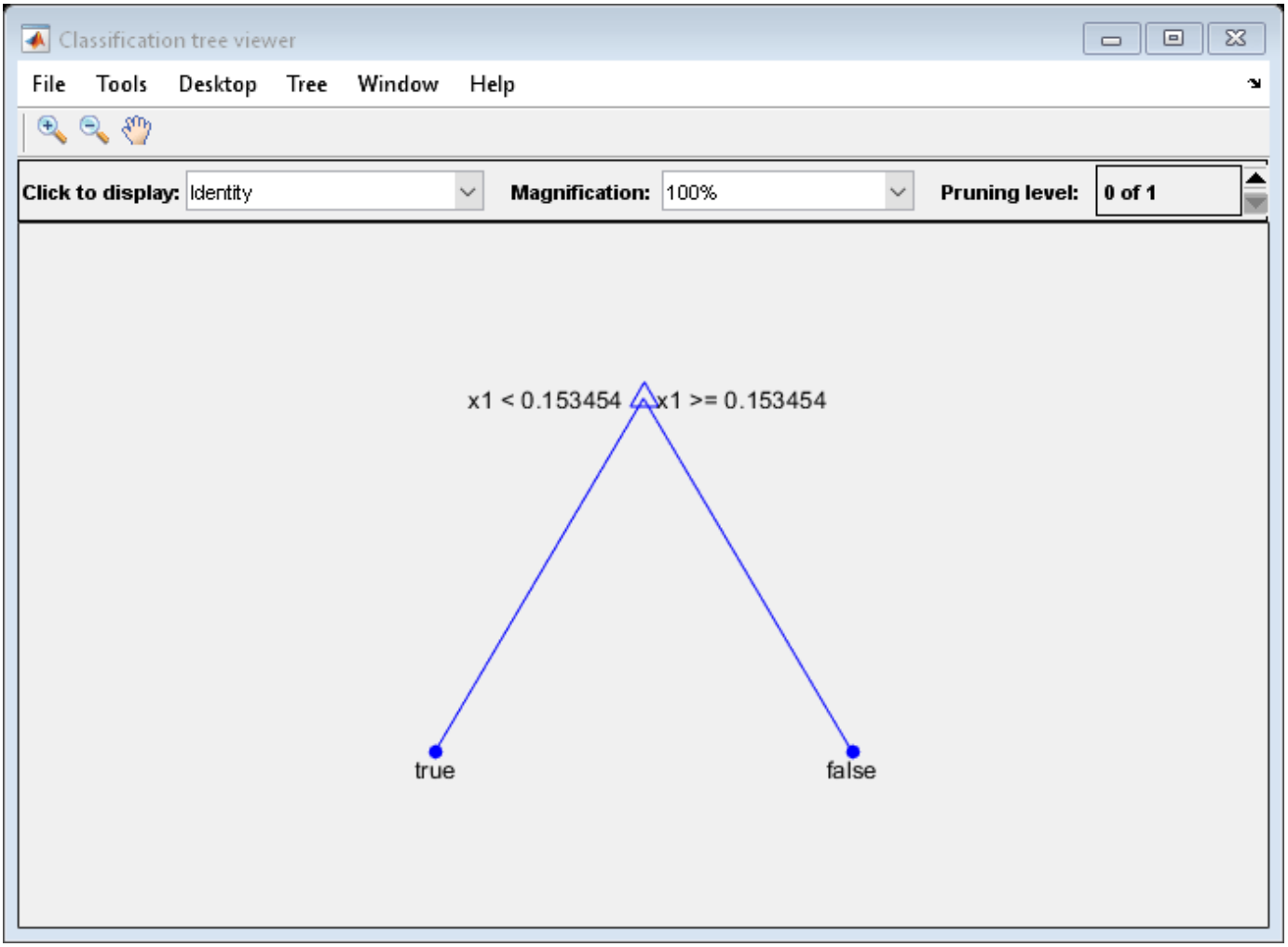
Generate 100 random points and classify them:

```
rng(0,'twister') % for reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`edge` | `fitctree` | `loss` | `predict`

margin

Class: `CompactTreeBagger`

Classification margin

Syntax

```
mar = margin(B,TBLnew,Ynew)
mar = margin(B,Xnew,Ynew)
mar = margin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
mar = margin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`mar = margin(B,TBLnew,Ynew)` computes the classification margins for the predictors contained in the table `TBLnew` given true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`mar = margin(B,Xnew,Ynew)` computes the classification margins for the predictors contained in the matrix `Xnew` given true response `Ynew`.

`Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector or logical vector. `mar` is a numeric array of size `Nobs`-by-`NTrees`, where `Nobs` is the number of rows of `TBLnew` and `Ynew`, and `NTrees` is the number of trees in the ensemble `B`. For observation `I` and tree `J`, `mar(I,J)` is the difference between the score for the true class and the largest score for other classes. This method is available for classification ensembles only.

`mar = margin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `mar = margin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	How the method computes errors. If set to `'cumulative'` (default), `margin` computes cumulative errors and `mar` is an `Nobs`-by-`NTrees` matrix, where the first column gives error from `trees(1)`, second column gives error from`trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is a `Nobs`-by-`NTrees` matrix, where each element is an error from each tree in the ensemble. If set to `'ensemble'`, `mar` a single column of length `Nobs` showing the cumulative margins for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives error from `trees(1)`, the second element gives error from `trees(1:2)` etc.

`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. The method uses these weights to combine output from the specified trees by taking a weighted average instead of the simple non-weighted majority vote. You cannot use this argument in the `'individual'` mode.
`'UseInstanceForTree'`	Logical matrix of size `Nobs`-by-`NTrees` indicating which trees should be used to make predictions for each observation. By default the method uses all trees for all observations.

See Also

`margin`

margin

Class: `TreeBagger`

Classification margin

Syntax

```
mar = margin(B,TBLnew,Ynew)
mar = margin(B,Xnew,Ynew)
mar = margin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
mar = margin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`mar = margin(B,TBLnew,Ynew)` computes the classification margins for the predictors contained in the table `TBLnew` given true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`mar = margin(B,Xnew,Ynew)` computes the classification margins for the predictors contained in the matrix `Xnew` given true response `Ynew`. If you trained `B` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

`Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector or logical vector. `mar` is a numeric array of size `Nobs`-by-`NTrees`, where `Nobs` is the number of rows of `TBLnew` and `Ynew`, and `NTrees` is the number of trees in the ensemble `B`. For observation `I` and tree `J`, `mar(I,J)` is the difference between the score for the true class and the largest score for other classes. This method is available for classification ensembles only.

`mar = margin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `mar = margin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	Character vector or string scalar indicating how the method computes errors. If set to `'cumulative'` (default), `margin` computes cumulative errors and `mar` is an `Nobs`-by-`NTrees` matrix, where the first column gives error from `trees(1)`, second column gives error from`trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is a `Nobs`-by-`NTrees` matrix, where each element is an error from each tree in the ensemble. If set to `'ensemble'`, `mar` a single column of length `Nobs` showing the cumulative margins for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives error from `trees(1)`, the second element gives error from `trees(1:2)` etc.

`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. The method uses these weights to combine output from the specified trees by taking a weighted average instead of the simple non-weighted majority vote. You cannot use this argument in the `'individual'` mode.
`'UseInstanceForTree'`	Logical matrix of size `Nobs`-by-`NTrees` indicating which trees should be used to make predictions for each observation. By default the method uses all trees for all observations.

See Also

`margin`

margmean

Class: `RepeatedMeasuresModel`

Estimate marginal means

Syntax

```
tbl = margmean(rm,vars)
tbl = margmean(rm,vars,'alpha',alpha)
```

Description

`tbl = margmean(rm,vars)` returns the estimated marginal means for the variables `vars`, in the table `tbl`.

`tbl = margmean(rm,vars,'alpha',alpha)` returns the 100*(1–`alpha`)% confidence intervals for the marginal means.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`vars` — Variables for which to compute the marginal means

character vector | string scalar | string array | cell array of character vectors

Variables for which to compute the marginal means, specified as a character vector or string scalar representing the name of a between or within-subjects factor in `rm`, or a string array or cell array of character vectors representing the names of multiple variables. Each between-subjects factor must be categorical.

For example, if you want to compute the marginal means for the variables Drug and Gender, then you can specify as follows.

Example: `{'Drug','Gender'}`

Data Types: `char` | `string` | `cell`

`alpha` — Significance level

0.05 (default) | scalar value in the range of 0 to 1

Significance level of the confidence intervals for population marginal means, specified as a scalar value in the range of 0 to 1. The confidence level is 100*(1–`alpha`)%.

For example, you can specify a 99% confidence level as follows.

Example: `'alpha',0.01`

Data Types: `double` | `single`

Output Arguments

`tbl` — Estimated marginal means

table

Estimated marginal means, returned as a table. `tbl` contains one row for each combination of the groups of the variables you specify in `vars`, one column for each variable, and the following columns.

Column name	Description
`Mean`	Estimated marginal means
`StdErr`	Standard errors of the estimates
`Lower`	Lower limit of a 95% confidence interval for the true population mean
`Upper`	Upper limit of a 95% confidence interval for the true population mean

Examples

Compute Marginal Means Grouped by Two Factors

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` to `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` to `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Compute the marginal means grouped by the factors `Group` and `Gender`.

```
M = margmean(rm,{'Group' 'Gender'})
```

```
M=6×6 table
    Group    Gender     Mean      StdErr     Lower       Upper
    _____    ______    _______    ______    ________    _______

      A      Female     15.946    5.6153      4.3009     27.592
      A      Male       8.0726    5.7236     -3.7973     19.943
      B      Female     11.758    5.7091    -0.08189     23.598
      B      Male       2.2858    5.6748      -9.483     14.055
      C      Female    -8.6183     5.871     -20.794     3.5574
      C      Male      -13.551    5.7283     -25.431    -1.6712
```

Display the description for table M.

```
M.Properties.Description
```

```
ans =
    'Estimated marginal means
     Means computed with Age=13.7, IQ=98.2667'
```

Compute Estimated Marginal Means and Confidence Intervals

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species, setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers, the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Compute the marginal means grouped by the factor species.

```
margmean(rm,'species')
```

```
ans=3×5 table
       species         Mean      StdErr     Lower     Upper
    _____________    ______    ________    ______    ______

    {'setosa'    }    2.5355    0.042807    2.4509    2.6201
    {'versicolor'}     3.573    0.042807    3.4884    3.6576
    {'virginica' }     4.285    0.042807    4.2004    4.3696
```

`StdError` field shows the standard errors of the estimated marginal means. The `Lower` and `Upper` fields show the lower and upper bounds for the 95% confidence intervals of the group marginal means, respectively. None of the confidence intervals overlap, which indicates that marginal means differ with species. You can also plot the estimated marginal means using the `plotprofile` method.

Compute the 99% confidence intervals for the marginal means.

```
margmean(rm,'species','alpha',0.01)
```

```
ans=3×5 table
       species         Mean      StdErr     Lower     Upper
    _____________    ______    ________    ______    ______

    {'setosa'    }    2.5355    0.042807    2.4238    2.6472
    {'versicolor'}     3.573    0.042807    3.4613    3.6847
    {'virginica' }     4.285    0.042807    4.1733    4.3967
```

See Also

`multcompare` | `fitrm` | `plotprofile`

mauchly

Class: `RepeatedMeasuresModel`

Mauchly's test for sphericity

Syntax

```
tbl = mauchly(rm)
tbl = mauchly(rm,C)
```

Description

`tbl = mauchly(rm)` returns the result of the Mauchly's test for sphericity for the repeated measures model `rm`.

It tests the null hypothesis that the sphericity assumption is true for the response variables in `rm`.

For more information, see "Mauchly's Test of Sphericity" on page 9-57.

`tbl = mauchly(rm,C)` returns the result of the Mauchly's test based on the contrast matrix `C`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`C` — Contrasts

matrix

Contrasts, specified as a matrix. The default value of `C` is the Q factor in a QR decomposition of the matrix M, where M is defined so that Y*M is the difference between all successive pairs of columns of the repeated measures matrix Y.

Data Types: `single` | `double`

Output Arguments

`tbl` — Results of Mauchly's test of sphericity

table

Results of Mauchly's test for sphericity for the repeated measures model `rm`, returned as a `table`.

`tbl` contains the following columns.

Column Name	Definition
`W`	Value of Mauchly's *W* statistic
`ChiStat`	Chi-square statistic value
`DF`	Degrees of freedom of the Chi-square statistic
`pValue`	*p*-value corresponding to the Chi-square statistic

Data Types: `table`

Examples

Perform Mauchly's Test

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform Mauchly's test to assess the sphericity assumption.

```
mauchly(rm)
```

```
ans=1×4 table
       W       ChiStat    DF      pValue
    _______    _______    __    __________

    0.55814    84.976     5     7.6149e-17
```

The small *p*-value (in the `pValue` field) indicates that the sphericity, hence the compound symmetry assumption, does not hold. You should use epsilon corrections to compute the *p*-values for a repeated measures anova. You can compute the epsilon corrections using the `epsilon` method and perform the repeated measures anova with the corrected *p*-values using the `ranova` method.

See Also

`epsilon` | `fitrm` | `ranova`

Topics

"Mauchly's Test of Sphericity" on page 9-57

“Compound Symmetry Assumption and Epsilon Corrections” on page 9-55

mat2dataset

(Not Recommended) Convert matrix to dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = mat2dataset(X)
ds = mat2dataset(X,Name,Value)
```

Description

`ds = mat2dataset(X)` converts a matrix to a `dataset` array.

`ds = mat2dataset(X,Name,Value)` performs the conversion using additional options specified by one or more `Name,Value` pair arguments.

Examples

Convert Matrix to Dataset Array

Convert a matrix to a dataset array using the default options.

Load sample data.

```
load('fisheriris')
X = meas;
size(X)
```

```
ans = 1×2

   150     4
```

Convert the matrix to a dataset array.

```
ds = mat2dataset(X);
size(ds)
```

```
ans = 1×2

   150     4
```

```
ds(1:5,:)
```

```
ans =
    X1     X2     X3     X4
    5.1    3.5    1.4    0.2
```

```
    4.9      3    1.4    0.2
    4.7    3.2    1.3    0.2
    4.6    3.1    1.5    0.2
      5    3.6    1.4    0.2
```

When you do not specify variable names, `mat2dataset` uses the matrix name and column numbers to create default variable names.

Convert Matrix to Dataset Array with Variable Names

Load sample data.

```
load('fisheriris')
X = meas;
size(X)
```

```
ans = 1×2

   150     4
```

Convert the matrix to a dataset array, providing a variable name for each of the four column of X.

```
ds = mat2dataset(X,'VarNames',{'SLength',...
'SWidth','PLength','PWidth'});
size(ds)
```

```
ans = 1×2

   150     4
```

```
ds(1:5,:)
```

```
ans =
    SLength    SWidth    PLength    PWidth
    5.1        3.5       1.4        0.2
    4.9          3       1.4        0.2
    4.7        3.2       1.3        0.2
    4.6        3.1       1.5        0.2
      5        3.6       1.4        0.2
```

Create a Dataset Array with Multicolumn Variables

Convert a matrix to a dataset array containing multicolumn variables.

Load sample data.

```
load('fisheriris')
X = meas;
size(X)
```

```
ans = 1×2

   150     4
```

Convert the matrix to a dataset array, combining the sepal measurements (the first two columns) into one variable named `SepalMeas`, and the petal measurements (third and fourth columns) into one variable names `PetalMeas`.

```
ds = mat2dataset(X,'NumCols',[2,2],...
'VarNames',{'SepalMeas','PetalMeas'});
ds(1:5,:)
```

```
ans =
    SepalMeas          PetalMeas
    5.1         3.5    1.4         0.2
    4.9           3    1.4         0.2
    4.7         3.2    1.3         0.2
    4.6         3.1    1.5         0.2
      5         3.6    1.4         0.2
```

The output dataset array has 150 observations and 2 variables.

```
size(ds)
```

```
ans = 1×2

   150     2
```

Input Arguments

X — Input matrix

matrix

Input matrix to convert to a dataset array, specified as an *M*-by-*N* numeric matrix. Each column of `X` becomes a variable in the output *M*-by-*N* dataset array.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumCols',[1,1,2,1]` specifies that the 3rd and 4th columns of the input matrix should be combined into a single variable.

VarNames — Variable names for output dataset array

string array | cell array of character vectors

Variable names for the output dataset array, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors. You must provide a variable name for each variable in `ds`. The names must be valid MATLAB identifiers, and must be unique.

Example: `'VarNames',{'myVar1','myVar2','myVar3'}`

`ObsNames` — Observation names for output dataset array
string array | cell array of character vectors

Observation names for the output dataset array, specified as the comma-separated pair consisting of `'ObsNames'` and a string array or cell array of character vectors. The names do not need to be valid MATLAB identifiers, but they must be unique.

`NumCols` — Number of columns for each variable
vector of nonnegative integers

Number of columns for each variable in `ds`, specified as the comma-separated pair consisting of `'NumCols'` and a vector of nonnegative integers. When the number of columns for a variable is greater than one, `mat2dataset` combines multiple columns in `X` into a single variable in `ds`. The vector you assign to `NumCols` must sum to `size(X,2)`.

For example, to convert a matrix with eight columns into a dataset array with five variables, specify a vector with five elements that sum to eight, such as `'NumCols',[1,1,3,1,2]`.

Output Arguments

`ds` — Output dataset array
dataset array

Output dataset array, returned by default with a variable for each column of `X`, and an observation for each row of `X`. If you specify `NumCols`, then the number of variables in `ds` is equal to the length of the specified vector of column numbers.

See Also

`cell2dataset` | `dataset` | `struct2dataset`

Topics
“Create a Dataset Array from Workspace Variables” on page 2-57
“Create a Dataset Array from a File” on page 2-62
“Dataset Arrays” on page 2-112

Introduced in R2012b

mdscale

Nonclassical multidimensional scaling

Syntax

```
Y = mdscale(D,p)
[Y,stress] = mdscale(D,p)
[Y,stress,disparities] = mdscale(D,p)
[...] = mdscale(D,p,'Name',value)
```

Description

`Y = mdscale(D,p)` performs nonmetric multidimensional scaling on the *n*-by-*n* dissimilarity matrix `D`, and returns `Y`, a configuration of *n* points (rows) in `p` dimensions (columns). The Euclidean distances between points in `Y` approximate a monotonic transformation of the corresponding dissimilarities in `D`. By default, `mdscale` uses Kruskal's normalized stress1 criterion.

You can specify `D` as either a full *n*-by-*n* matrix, or in upper triangle form such as is output by `pdist`. A full dissimilarity matrix must be real and symmetric, and have zeros along the diagonal and non-negative elements everywhere else. A dissimilarity matrix in upper triangle form must have real, non-negative entries. `mdscale` treats `NaN`s in `D` as missing values, and ignores those elements. `Inf` is not accepted.

You can also specify `D` as a full similarity matrix, with ones along the diagonal and all other elements less than one. `mdscale` transforms a similarity matrix to a dissimilarity matrix in such a way that distances between the points returned in `Y` approximate `sqrt(1-D)`. To use a different transformation, transform the similarities prior to calling `mdscale`.

`[Y,stress] = mdscale(D,p)` returns the minimized stress, i.e., the stress evaluated at `Y`.

`[Y,stress,disparities] = mdscale(D,p)` returns the disparities, that is, the monotonic transformation of the dissimilarities `D`.

`[...] = mdscale(D,p,'Name',value)` specifies one or more optional parameter name/value pairs that control further details of `mdscale`. Specify *`Name`* in single quotes. Available parameters are

- `Criterion`— The goodness-of-fit criterion to minimize. This also determines the type of scaling, either non-metric or metric, that `mdscale` performs. Choices for non-metric scaling are:
  - `'stress'` — Stress normalized by the sum of squares of the inter-point distances, also known as stress1. This is the default.
  - `'sstress'` — Squared stress, normalized with the sum of 4th powers of the inter-point distances.

  Choices for metric scaling are:
  - `'metricstress'` — Stress, normalized with the sum of squares of the dissimilarities.
  - `'metricsstress'` — Squared stress, normalized with the sum of 4th powers of the dissimilarities.

- `'sammon'` — Sammon's nonlinear mapping criterion. Off-diagonal dissimilarities must be strictly positive with this criterion.
- `'strain'` — A criterion equivalent to that used in classical multidimensional scaling.

- `Weights` — A matrix or vector the same size as `D`, containing nonnegative dissimilarity weights. You can use these to weight the contribution of the corresponding elements of `D` in computing and minimizing stress. Elements of `D` corresponding to zero weights are effectively ignored.

  Note When you specify weights as a full matrix, its diagonal elements are ignored and have no effect, since the corresponding diagonal elements of `D` do not enter into the stress calculation.

- `Start` — Method used to choose the initial configuration of points for Y. The choices are
  - `'cmdscale'` — Use the classical multidimensional scaling solution. This is the default. `'cmdscale'` is not valid when there are zero weights.
  - `'random'` — Choose locations randomly from an appropriately scaled p-dimensional normal distribution with uncorrelated coordinates.
  - An *n*-by-`p` matrix of initial locations, where n is the size of the matrix `D` and `p` is the number of columns of the output matrix `Y`. In this case, you can pass in `[]` for `p` and `mdscale` infers `p` from the second dimension of the matrix. You can also supply a 3-D array, implying a value for `'Replicates'` from the array's third dimension.
- `Replicates` — Number of times to repeat the scaling, each with a new initial configuration. The default is `1`.
- `Options` — Options for the iterative algorithm used to minimize the fitting criterion. Pass in an options structure created by `statset`. For example,

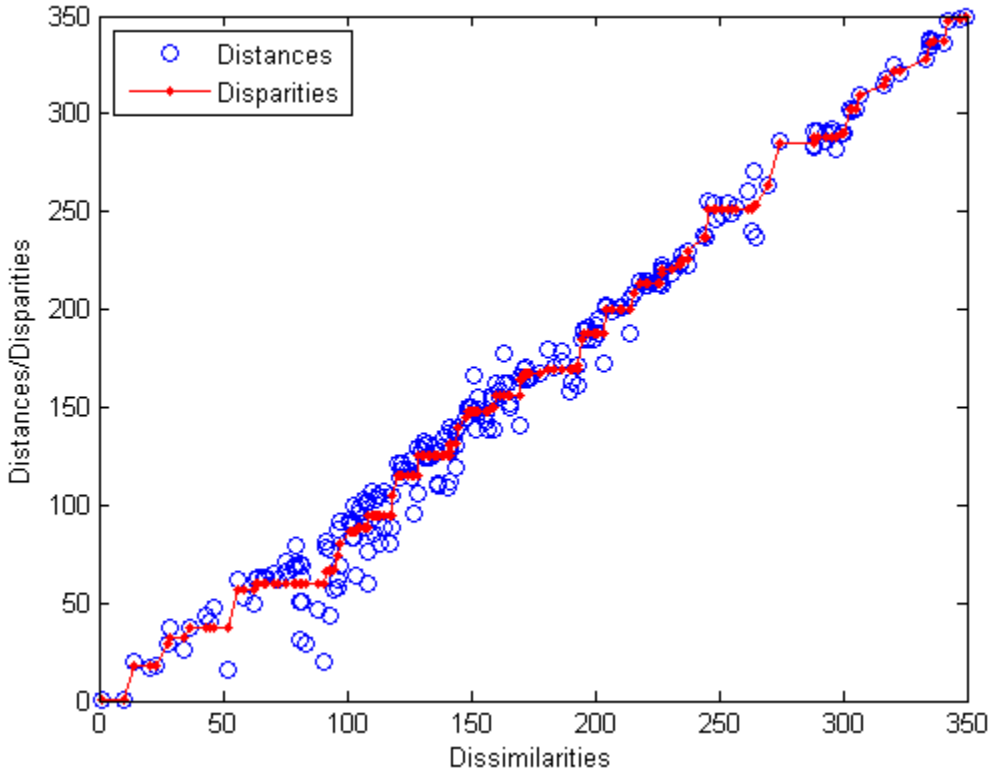
  ```
  opts = statset(param1,val1,param2,val2, ...);
  [...] = mdscale(...,'Options',opts)
  ```

  The choices of `statset` parameters are
  - `'Display'` — Level of display output. The choices are `'off'` (the default), `'iter'`, and `'final'`.
  - `'MaxIter'` — Maximum number of iterations allowed. The default is `200`.
  - `'TolFun'` — Termination tolerance for the stress criterion and its gradient. The default is `1e-4`.
  - `'TolX'`— Termination tolerance for the configuration location step size. The default is `1e-4`.

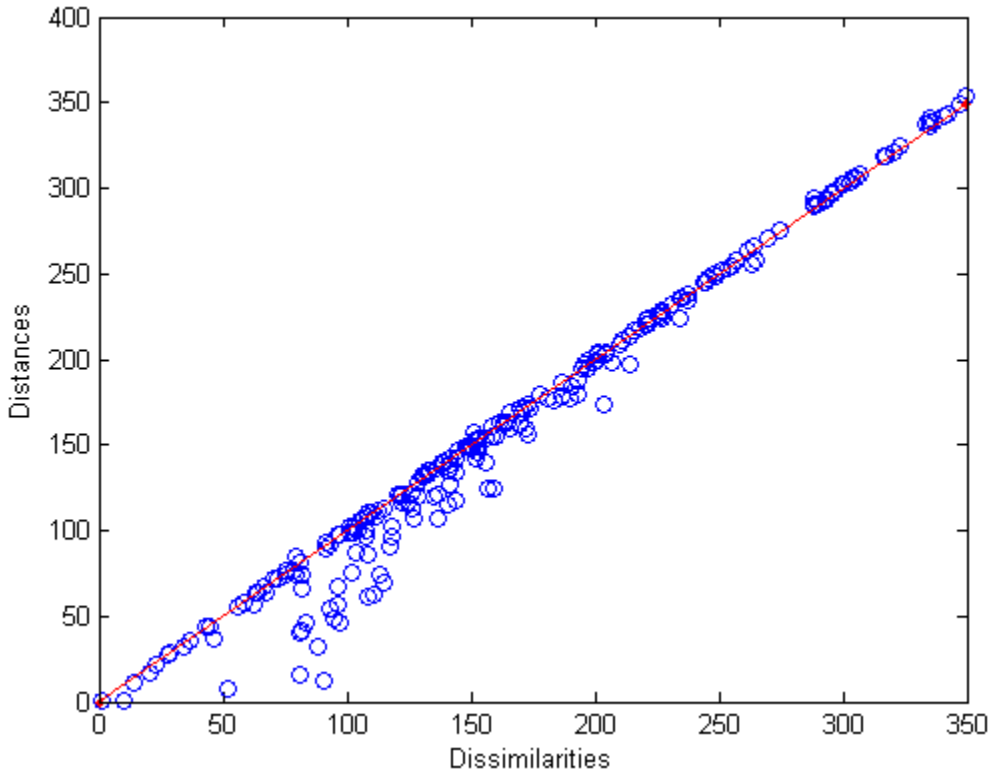
Examples

```
load cereal.mat
X = [Calories Protein Fat Sodium Fiber ...
     Carbo Sugars Shelf Potass Vitamins];

% Take a subset from a single manufacturer.
X = X(strcmp('K',cellstr(Mfg)),:);

% Create a dissimilarity matrix.
dissimilarities = pdist(X);

% Use non-metric scaling to recreate the data in 2D,
% and make a Shepard plot of the results.
```

```
[Y,stress,disparities] = mdscale(dissimilarities,2);
distances = pdist(Y);
[dum,ord] = sortrows([disparities(:) dissimilarities(:)]);
plot(dissimilarities,distances,'bo', ...
dissimilarities(ord),disparities(ord),'r.-');
xlabel('Dissimilarities'); ylabel('Distances/Disparities')
legend({'Distances' 'Disparities'},'Location','NW');
```




```
% Do metric scaling on the same dissimilarities.
figure
[Y,stress] = ...
mdscale(dissimilarities,2,'criterion','metricsstress');
distances = pdist(Y);
plot(dissimilarities,distances,'bo', ...
[0 max(dissimilarities)],[0 max(dissimilarities)],'r.-');
xlabel('Dissimilarities'); ylabel('Distances')
```



See Also

`cmdscale` | `pdist` | `statset`

Introduced before R2006a

mdsprox

Class: `CompactTreeBagger`

Multidimensional scaling of proximity matrix

Syntax

```
[SC,EIGEN] = mdsprox(B,X)
[SC,EIGEN] = mdsprox(B,X,'param1',val1,'param2',val2,...)
```

Description

`[SC,EIGEN] = mdsprox(B,X)` applies classical multidimensional scaling to the proximity matrix computed for the data in the matrix `X`, and returns scaled coordinates `SC` and eigenvalues `EIGEN` of the scaling transformation. The method applies multidimensional scaling to the matrix of distances defined as 1-`prox`, where `prox` is the proximity matrix returned by the `proximity` method.

You can supply the proximity matrix directly by using the `'Data'` parameter.

`[SC,EIGEN] = mdsprox(B,X,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Data'`	Flag indicating how the method treats the `X` input argument. If set to `'predictors'` (default), `mdsprox` assumes `X` to be a matrix of predictors and used for computation of the proximity matrix. If set to `'proximity'`, the method treats `X` as a proximity matrix returned by the `proximity` method.
`'Colors'`	If you supply this argument, `mdsprox` makes overlaid scatter plots of two scaled coordinates using specified colors for different classes. You must supply the colors as a character vector or a string scalar with one letter for each color. If there are more classes in the data than letters in the supplied value, `mdsprox` plots only the first `C` classes, where `C` is the number of letters in the supplied value. For regression or if you do not provide the vector of true class labels, the method uses the first color for all observations in `X`.
`'Labels'`	Vector of true class labels for a classification ensemble. True class labels can be a numeric vector, character matrix, string array, or cell array of character vectors. If supplied, this vector must have as many elements as there are observations (rows) in `X`. This argument has no effect unless you also supply the `'Colors'` argument.
`'MDSCoordinates'`	Indices of the two scaled coordinates to plot. By default, `mdsprox` makes a scatter plot of the first and second scaled coordinates which correspond to the two largest eigenvalues. You can specify any other two or three indices not exceeding the dimensionality of the scaled data. This argument has no effect unless you also supply the `'Colors'` argument.

See Also

`cmdscale` | `mdsprox` | `proximity`

mdsprox

Class: `TreeBagger`

Multidimensional scaling of proximity matrix

Syntax

```
[S,E] = mdsprox(B)
[S,E] = mdsprox(B,'param1',val1,'param2',val2,...)
```

Description

`[S,E] = mdsprox(B)` returns scaled coordinates, `S`, and eigenvalues, `E`, for the proximity matrix in the ensemble `B`. An earlier call to `fillprox(B)` must create the proximity matrix.

`[S,E] = mdsprox(B,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Keep'`	Array of indices of observations in the training data to use for multidimensional scaling. By default, this argument is set to `'all'`. If you provide numeric or logical indices, the method uses only the subset of the training data specified by these indices to compute the scaled coordinates and eigenvalues.
`'Colors'`	If you supply this argument, `mdsprox` makes overlaid scatter plots of two scaled coordinates using specified colors for different classes. You must supply the colors as a character vector or a string scalar with one letter for each color. If there are more classes in the data than letters in the supplied value, `mdsprox` plots only the first `C` classes, where `C` is the number of letters in the supplied value. For regression, the method uses the first color for all observations in `X`.
`'MDSCoordinates'`	Indices of the two scaled coordinates to plot. By default, `mdsprox` makes a scatter plot of the first and second scaled coordinates which correspond to the two largest eigenvalues. You can specify any other two or three indices not exceeding the dimensionality of the scaled data. This argument has no effect unless you also supply the `'Colors'` argument.

See Also

`cmdscale` | `fillprox` | `mdsprox`

mean

Package: prob

Mean of probability distribution

Syntax

```
m = mean(pd)
```

Description

`m = mean(pd)` returns the mean `m` of the probability distribution `pd`.

Examples

Mean of a Fitted Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Create a normal distribution object by fitting it to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the mean of the fitted distribution.

```
m = mean(pd)
```

```
m = 75.0083
```

The mean of the normal distribution is equal to the parameter `mu`.

Mean of a Skewed Distribution

Create a Weibull probability distribution object.

```
pd = makedist('Weibull','a',5,'b',2)
```

```
pd =
  WeibullDistribution
```

```
  Weibull distribution
    A = 5
    B = 2
```

Compute the mean of the distribution.

```
mean = mean(pd)
```

```
mean = 4.4311
```

Mean of a Uniform Distribution

Create a uniform distribution object

```
pd = makedist('Uniform','lower',-3,'upper',5)
```

```
pd =
  UniformDistribution

  Uniform distribution
    Lower = -3
    Upper =  5
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 1
```

Mean of a Kernel Distribution

Load the sample data. Create a probability distribution object by fitting a kernel distribution to the miles per gallon (MPG) data.

```
load carsmall;
pd = fitdist(MPG,'Kernel')
```

```
pd =
  KernelDistribution

    Kernel = normal
    Bandwidth = 4.11428
    Support = unbounded
```

Compute the mean of the distribution.

```
mean(pd)
```

```
ans = 23.7181
```

Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

m — Mean

scalar value

Mean of the probability distribution, returned as a scalar value.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `fitdist` | `makedist` | `median` | `std`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

meanMargin

Class: `CompactTreeBagger`

Mean classification margin

Syntax

```
mar = meanMargin(B,TBLnew,Ynew)
mar = meanMargin(B,Xnew,Ynew)
mar = meanMargin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
mar = meanMargin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`mar = meanMargin(B,TBLnew,Ynew)` computes average classification margins for the predictors contained in the table `TBLnew` given the true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`mar = meanMargin(B,Xnew,Ynew)` computes average classification margins for the predictors contained in the matrix `Xnew` given true response `Ynew`. If you trained `B` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

`Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector, or logical vector. `meanMargin` averages the margins over all observations (rows) in `TBLnew` or `Xnew` for each tree. `mar` is a matrix of size 1-by-`NTrees`, where `NTrees` is the number of trees in the ensemble `B`. This method is available for classification ensembles only.

`mar = meanMargin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `mar = meanMargin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	How `meanMargin` computes errors. If set to `'cumulative'` (default), is a vector of length `NTrees` where the first element gives mean margin from `trees(1)`, second column gives mean margins from `trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is a vector of length `NTrees`, where each element is a mean margin from each tree in the ensemble . If set to `'ensemble'`, `mar` is a scalar showing the cumulative mean margin for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives mean margin from `trees(1)`, the second element gives mean margin from `trees(1:2)` etc.

`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. `meanMargin` uses these weights to combine output from the specified trees by taking a weighted average instead of the simple nonweighted majority vote. You cannot use this argument in the `'individual'` mode.

See Also

`meanMargin`

meanMargin

Class: `TreeBagger`

Mean classification margin

Syntax

```
mar = meanMargin(B,TBLnew,Ynew)
mar = meanMargin(B,Xnew,Ynew)
mar = meanMargin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)
mar = meanMargin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)
```

Description

`mar = meanMargin(B,TBLnew,Ynew)` computes average classification margins for the predictors contained in the table `TBLnew` given the true response `Ynew`. You can omit `Ynew` if `TBLnew` contains the response variable. If you trained `B` using sample data contained in a table, then the input data for this method must also be in a table.

`mar = meanMargin(B,Xnew,Ynew)` computes average classification margins for the predictors contained in the matrix `Xnew` given true response `Ynew`. If you trained `B` using sample data contained in a matrix, then the input data for this method must also be in a matrix.

`Ynew` can be a numeric vector, character matrix, string array, cell array of character vectors, categorical vector or logical vector. `meanMargin` averages the margins over all observations (rows) in `TBLnew` or `Xnew` for each tree. `mar` is a matrix of size 1-by-`NTrees`, where `NTrees` is the number of trees in the ensemble `B`. This method is available for classification ensembles only.

`mar = meanMargin(B,TBLnew,Ynew,'param1',val1,'param2',val2,...)` or `mar = meanMargin(B,Xnew,Ynew,'param1',val1,'param2',val2,...)` specifies optional parameter name-value pairs:

`'Mode'`	Character vector or string scalar indicating how `meanMargin` computes errors. If set to `'cumulative'` (default), is a vector of length `NTrees` where the first element gives mean margin from `trees(1)`, second column gives mean margins from `trees(1:2)` etc, up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is a vector of length `NTrees`, where each element is a mean margin from each tree in the ensemble . If set to `'ensemble'`, `mar` is a scalar showing the cumulative mean margin for the entire ensemble .
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives mean margin from `trees(1)`, the second element gives mean margin from `trees(1:2)` etc.

`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. `meanMargin` uses these weights to combine output from the specified trees by taking a weighted average instead of the simple nonweighted majority vote. You cannot use this argument in the `'individual'` mode.
`'Weights'`	Vector of observation weights to use for margin averaging. By default the weight of every observation is set to 1. The length of this vector must be equal to the number of rows in X.
`'UseInstanceForTree'`	Logical matrix of size `Nobs`-by-`NTrees` indicating which trees to use to make predictions for each observation. By default the method uses all trees for all observations.

See Also

`meanMargin`

surrogateAssociation

Mean predictive measure of association for surrogate splits in classification tree

Syntax

```
ma = surrogateAssociation(tree)
ma = surrogateAssociation(tree,N)
```

Description

`ma = surrogateAssociation(tree)` returns a matrix of predictive measures of association for the predictors in `tree`.

`ma = surrogateAssociation(tree,N)` returns a matrix of predictive measures of association averaged over the nodes in vector `N`.

Input Arguments

`tree`

A classification tree constructed with `fitctree`, or a compact regression tree constructed with `compact`.

`N`

Vector of node numbers in `tree`.

Output Arguments

`ma`

- `ma = surrogateAssociation(tree)` returns a `P`-by-`P` matrix, where `P` is the number of predictors in `tree`. `ma(i,j)` is the predictive measure of association on page 33-3954 between the optimal split on variable `i` and a surrogate split on variable `j`. For more details, see "Algorithms" on page 33-3955.
- `ma = surrogateAssociation(tree,N)` returns a `P`-by-`P` representing the predictive measure of association between variables averaged over nodes in the vector `N`. `N` contains node numbers from `1` to `max(tree.NumNodes)`.

Examples

Estimate Predictive Measures of Association for Surrogate Splits

Load Fisher's iris data set.

```
load fisheriris
```

Grow a classification tree using `species` as the response. Specify to use surrogate splits for missing values.

```
tree = fitctree(meas,species,'surrogate','on');
```

Find the mean predictive measure of association between the predictor variables.

```
ma = surrogateAssociation(tree)
```

```
ma = 4×4

    1.0000         0         0         0
         0    1.0000         0         0
    0.4633    0.2500    1.0000    0.5000
    0.2065    0.1413    0.4022    1.0000
```

Find the mean predictive measure of association averaged over the odd-numbered nodes in `tree`.

```
N = 1:2:tree.NumNodes;
ma = surrogateAssociation(tree,N)
```

```
ma = 4×4

    1.0000         0         0         0
         0    1.0000         0         0
    0.7600    0.5000    1.0000    1.0000
    0.4130    0.2826    0.8043    1.0000
```

More About

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_jL_k} - P_{R_jR_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_jL_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.

- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in (–∞,1]. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Surrogate Decision Splits

A surrogate decision split is an alternative to the optimal decision split at a given node in a decision tree. The optimal split is found by growing the tree; the surrogate split uses a similar or correlated predictor variable and split criterion.

When the value of the optimal split predictor for an observation is missing, the observation is sent to the left or right child node using the best surrogate predictor. When the value of the best surrogate split predictor for the observation is also missing, the observation is sent to the left or right child node using the second-best surrogate predictor, and so on. Candidate splits are sorted in descending order by their predictive measure of association on page 33-2412.

Algorithms

Element `ma(i,j)` is the predictive measure of association averaged over surrogate splits on predictor `j` for which predictor `i` is the optimal split predictor. This average is computed by summing positive values of the predictive measure of association over optimal splits on predictor `i` and surrogate splits on predictor `j` and dividing by the total number of optimal splits on predictor `i`, including splits for which the predictive measure of association between predictors `i` and `j` is negative.

See Also

`ClassificationTree` | `fitctree`

surrogateAssociation

Mean predictive measure of association for surrogate splits in regression tree

Syntax

```
ma = surrogateAssociation(tree)
ma = surrogateAssociation(tree,N)
```

Description

`ma = surrogateAssociation(tree)` returns a matrix of predictive measures of association for the predictors in `tree`.

`ma = surrogateAssociation(tree,N)` returns a matrix of predictive measures of association averaged over the nodes in vector `N`.

Input Arguments

`tree`

A regression tree constructed with `fitrtree`, or a compact regression tree constructed with `compact`.

`N`

Vector of node numbers in `tree`.

Output Arguments

`ma`

- `ma = surrogateAssociation(tree)` returns a `P`-by-`P` matrix, where `P` is the number of predictors in `tree`. `ma(i,j)` is the predictive measure of association on page 33-3957 between the optimal split on variable `i` and a surrogate split on variable `j`. For more details, see "Algorithms" on page 33-3958.
- `ma = surrogateAssociation(tree,N)` returns a `P`-by-`P` representing the predictive measure of association between variables averaged over nodes in the vector `N`. `N` contains node numbers from `1` to `max(tree.NumNodes)`.

Examples

Estimate Predictive Measures of Association for Surrogate Splits

Load the `carsmall` data set. Specify `Displacement`, `Horsepower`, and `Weight` as predictor variables.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using `MPG` as the response. Specify to use surrogate splits for missing values.

```
tree = fitrtree(X,MPG,'surrogate','on');
```

Find the mean predictive measure of association between the predictor variables.

```
ma = surrogateAssociation(tree)
```

```
ma = 3×3

    1.0000    0.2167    0.5083
    0.4521    1.0000    0.3769
    0.2540    0.2659    1.0000
```

Find the mean predictive measure of association averaged over the odd-numbered nodes in `tree`.

```
N = 1:2:tree.NumNodes;
ma = surrogateAssociation(tree,N)
```

```
ma = 3×3

    1.0000    0.1250    0.6875
    0.5632    1.0000    0.5861
    0.3333    0.3148    1.0000
```

More About

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_j L_k} - P_{R_j R_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_j L_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
- $P_{R_j R_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in $(-\infty,1]$. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Surrogate Decision Splits

A surrogate decision split is an alternative to the optimal decision split at a given node in a decision tree. The optimal split is found by growing the tree; the surrogate split uses a similar or correlated predictor variable and split criterion.

When the value of the optimal split predictor for an observation is missing, the observation is sent to the left or right child node using the best surrogate predictor. When the value of the best surrogate split predictor for the observation is also missing, the observation is sent to the left or right child node using the second-best surrogate predictor, and so on. Candidate splits are sorted in descending order by their predictive measure of association on page 33-2412.

Algorithms

Element `ma(i,j)` is the predictive measure of association averaged over surrogate splits on predictor `j` for which predictor `i` is the optimal split predictor. This average is computed by summing positive values of the predictive measure of association over optimal splits on predictor `i` and surrogate splits on predictor `j` and dividing by the total number of optimal splits on predictor `i`, including splits for which the predictive measure of association between predictors `i` and `j` is negative.

See Also

`RegressionTree` | `fitrtree` | `prune`

median

Package: prob

Median of probability distribution

Syntax

```
m = median(pd)
```

Description

`m = median(pd)` returns the median `m` for the probability distribution `pd`

Examples

Median of a Fitted Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Create a normal distribution object by fitting it to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the median of the fitted distribution.

```
m = median(pd)
```

```
m = 75.0083
```

For a symmetrical distribution such as the normal distribution, the median is equal to the mean, `mu`.

Median of a Skewed Distribution

Create a Weibull probability distribution object.

```
pd = makedist('Weibull','a',5,'b',2)
```

```
pd =
  WeibullDistribution
```

```
  Weibull distribution
    A = 5
    B = 2
```

Compute the median of the distribution.

```
m = median(pd)
```

```
m = 4.1628
```

For a skewed distribution such as the Weibull distribution, the median and the mean may not be equal.

Calculate the mean of the Weibull distribution and compare it to the median.

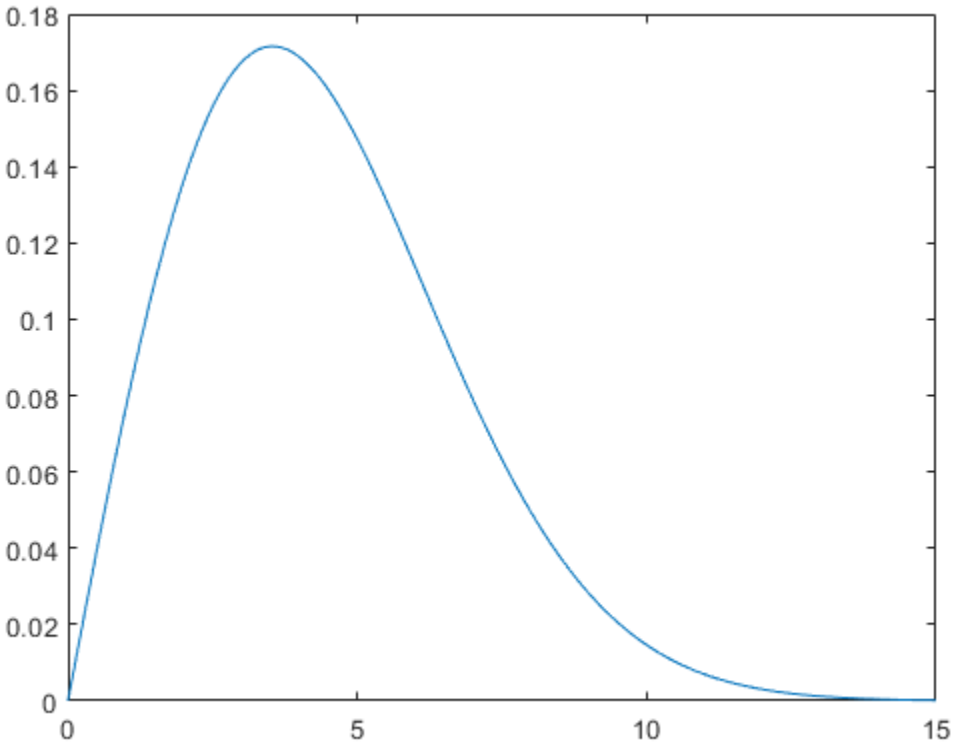
```
mean = mean(pd)
```

```
mean = 4.4311
```

The mean of the distribution is greater than the median.

Plot the pdf to visualize the distribution.

```
x = [0:.1:15];
pdf = pdf(pd,x);
plot(x,pdf)
```



Input Arguments

`pd` — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

`m` — Median

scalar value

Median of the probability distribution, returned as a scalar value. The value of `m` is the 50th percentile of the probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `fitdist` | `makedist` | `mean`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

mergelevels

(Not Recommended) Merge levels of nominal or ordinal arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Syntax

```
B = mergelevels(A,oldlevels)
B = mergelevels(A,oldlevels,newlevel)
```

Description

`B = mergelevels(A,oldlevels)` merges two or more levels of `A`.

- If `A` is a `nominal` array, `mergelevels` uses the first label in `oldlevels` as the new level.
- If `A` is an `ordinal` array, the levels specified by `oldlevels` must be consecutive, and `mergelevels` uses the label corresponding to the lowest level in `oldlevels` as the label for the new level.

`B = mergelevels(A,oldlevels,newlevel)` merges two or more levels into the new level with label `newlevel`.

Examples

Create New Category From Merged Levels

Create a nominal array from data in a cell array.

```
colors = nominal({'r','b','g';'g','r','b';'b','r','g'},...
                 {'blue','green','red'})
```

```
colors = 3x3 nominal
     red        blue      green
     green      red       blue
     blue       red       green
```

Merge the elements of the `'red'` and `'blue'` levels into a new level labeled `'purple'`.

```
colors = mergelevels(colors,{'red','blue'},'purple')
```

```
colors = 3x3 nominal
     purple     purple      green
     green      purple      purple
     purple     purple      green
```

Display the levels of `colors`.

```
getlevels(colors)

ans = 1x2 nominal
     purple      green
```

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

oldlevels — Levels to merge

string array | cell array of character vectors | 2-D character array

Levels to merge, specified as a string array, cell array of character vectors, or 2-D character array. For ordinal arrays, the levels in `oldlevels` must be consecutive.

Data Types: `char` | `string` | `cell`

newlevel — Level to create

character vector | string scalar

Level to create from the merged levels, specified as a character vector or string scalar that gives the label for the new level.

Data Types: `char` | `string`

Output Arguments

B — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, returned as a `nominal` or `ordinal` array object.

See Also

`addlevels` | `droplevels` | `nominal` | `ordinal` | `reorderlevels`

Topics

“Merge Category Levels” on page 2-16

Introduced in R2007a

mhsample

Metropolis-Hastings sample

Syntax

```
smpl = mhsample(start,nsamples,'pdf',pdf,'proppdf',proppdf,
'proprnd',proprnd)
smpl = mhsample(...,'symmetric',sym)
smpl = mhsample(...,'burnin',K)
smpl = mhsample(...,'thin',m)
smpl = mhsample(...,'nchain',n)
[smpl,accept] = mhsample(...)
```

Description

`smpl = mhsample(start,nsamples,'pdf',pdf,'proppdf',proppdf, 'proprnd',proprnd)` draws `nsamples` random samples from a target stationary distribution `pdf` using the Metropolis-Hastings algorithm.

`start` is a row vector containing the start value of the Markov Chain, `nsamples` is an integer specifying the number of samples to be generated, and `pdf`, `proppdf`, and `proprnd` are function handles created using `@`. `proppdf` defines the proposal distribution density, and `proprnd` defines the random number generator for the proposal distribution. `pdf` and `proprnd` take one argument as an input with the same type and size as `start`. `proppdf` takes two arguments as inputs with the same type and size as `start`.

`smpl` is a column vector or matrix containing the samples. If the log density function is preferred, `'pdf'` and `'proppdf'` can be replaced with `'logpdf'` and `'logproppdf'`. The density functions used in Metropolis-Hastings algorithm are not necessarily normalized.

The proposal distribution $q(x,y)$ gives the probability density for choosing x as the next point when y is the current point. It is sometimes written as $q(x|y)$.

If the `proppdf` or `logproppdf` satisfies $q(x,y) = q(y,x)$, that is, the proposal distribution is symmetric, `mhsample` implements Random Walk Metropolis-Hastings sampling. If the `proppdf` or `logproppdf` satisfies $q(x,y) = q(x)$, that is, the proposal distribution is independent of current values, `mhsample` implements Independent Metropolis-Hastings sampling.

`smpl = mhsample(...,'symmetric',sym)` draws `nsamples` random samples from a target stationary distribution `pdf` using the Metropolis-Hastings algorithm. `sym` is a logical value that indicates whether the proposal distribution is symmetric. The default value is false, which corresponds to the asymmetric proposal distribution. If `sym` is true, for example, the proposal distribution is symmetric, `proppdf` and `logproppdf` are optional.

`smpl = mhsample(...,'burnin',K)` generates a Markov chain with values between the starting point and the k^{th} point omitted in the generated sequence. Values beyond the k^{th} point are kept. `k` is a nonnegative integer with default value of `0`.

`smpl = mhsample(...,'thin',m)` generates a Markov chain with `m-1` out of `m` values omitted in the generated sequence. `m` is a positive integer with default value of `1`.

`smpl = mhsample(...,'nchain',n)` generates `n` Markov chains using the Metropolis-Hastings algorithm. `n` is a positive integer with a default value of 1. `smpl` is a matrix containing the samples. The last dimension contains the indices for individual chains.

`[smpl,accept] = mhsample(...)` also returns `accept`, the acceptance rate of the proposed distribution. `accept` is a scalar if a single chain is generated and is a vector if multiple chains are generated.

Examples

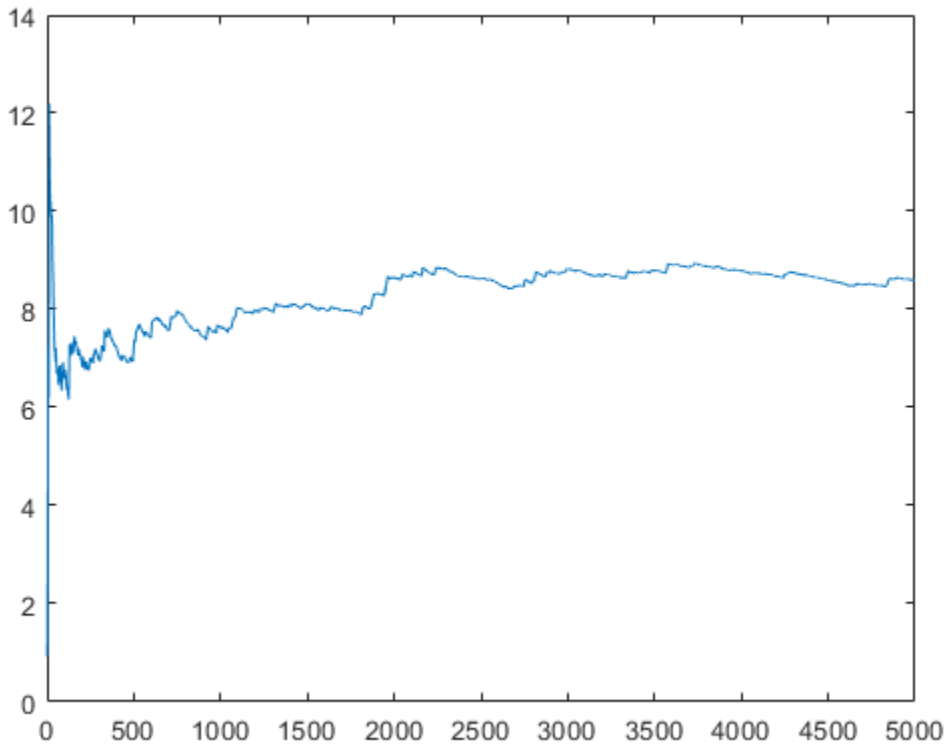
Estimate Moments Using Independent Metropolis-Hastings Sampling

Use Independent Metropolis-Hastings sampling to estimate the second order moment of a Gamma distribution.

```
rng default;  % For reproducibility
alpha = 2.43;
beta = 1;
pdf = @(x)gampdf(x,alpha,beta); % Target distribution
proppdf = @(x,y)gampdf(x,floor(alpha),floor(alpha)/alpha);
proprnd = @(x)sum(...
              exprnd(floor(alpha)/alpha,floor(alpha),1));
nsamples = 5000;
smpl = mhsample(1,nsamples,'pdf',pdf,'proprnd',proprnd,...
                'proppdf',proppdf);
```

Plot the results.

```
xxhat = cumsum(smpl.^2)./(1:nsamples)';
figure;
plot(1:nsamples,xxhat)
```



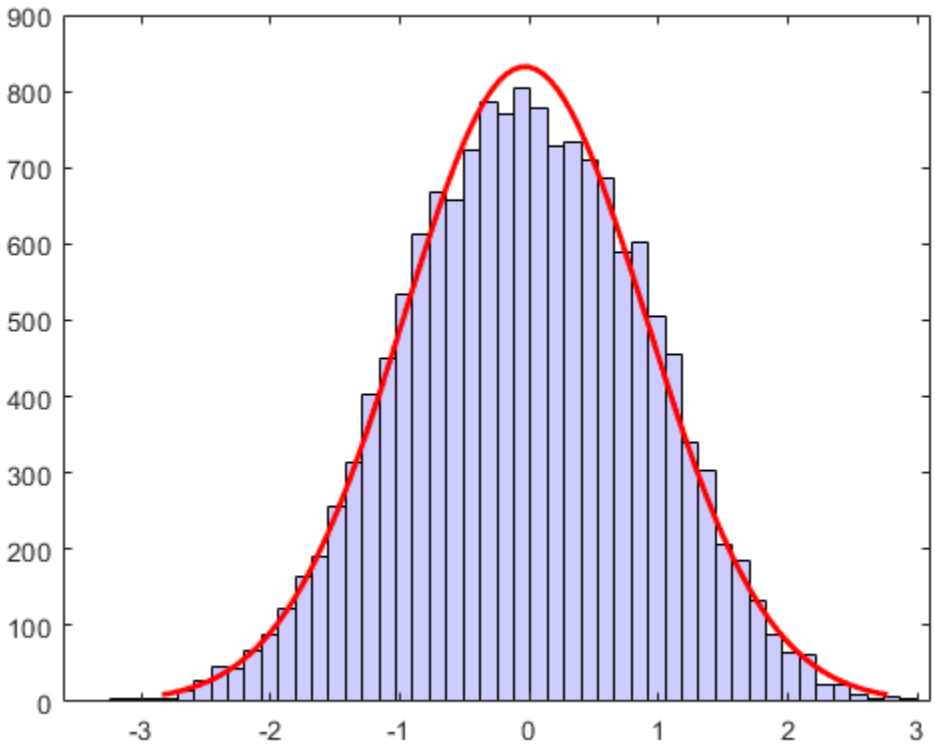
Random Walk Metropolis-Hastings Sampling

Use Random Walk Metropolis-Hastings sampling to generate sample data from a standard normal distribution.

```
rng default  % For reproducibility
delta = .5;
pdf = @(x) normpdf(x);
proppdf = @(x,y) unifpdf(y-x,-delta,delta);
proprnd = @(x) x + rand*2*delta - delta;
nsamples = 15000;
x = mhsample(1,nsamples,'pdf',pdf,'proprnd',proprnd,'symmetric',1);
```

Plot the sample data.

```
figure;
h = histfit(x,50);
h(1).FaceColor = [.8 .8 1];
```



See Also

`rand` | `slicesample`

Topics

“Using the Metropolis-Hastings Algorithm” on page 7-9
“Representing Sampling Distributions Using Markov Chain Samplers” on page 7-9

Introduced in R2006a

mle

Maximum likelihood estimates

Syntax

```
phat = mle(data)
phat = mle(data,'distribution',dist)

phat = mle(data,'pdf',pdf,'start',start)
phat = mle(data,'pdf',pdf,'start',start,'cdf',cdf)

phat = mle(data,'logpdf',logpdf,'start',start)
phat = mle(data,'logpdf',logpdf,'start',start,'logsf',logsf)

phat = mle(data,'nloglf',nloglf,'start',start)

phat = mle(___,Name,Value)
[phat,pci] = mle(___)
```

Description

`phat = mle(data)` returns maximum likelihood estimates (MLEs) for the parameters of a normal distribution, using the sample data in the vector `data`.

`phat = mle(data,'distribution',dist)` returns parameter estimates for a distribution specified by `dist`.

`phat = mle(data,'pdf',pdf,'start',start)` returns parameter estimates for a custom distribution specified by the probability density function `pdf`. You must also specify the initial parameter values, `start`.

`phat = mle(data,'pdf',pdf,'start',start,'cdf',cdf)` returns parameter estimates for a custom distribution specified by the probability density function `pdf` and custom cumulative distribution function `cdf`.

`phat = mle(data,'logpdf',logpdf,'start',start)` returns parameter estimates for a custom distribution specified by the log probability density function `logpdf`. You must also specify the initial parameter values, `start`.

`phat = mle(data,'logpdf',logpdf,'start',start,'logsf',logsf)` returns parameter estimates for a custom distribution specified by the log probability density function `logpdf` and custom log survival function on page 33-3981 `logsf`.

`phat = mle(data,'nloglf',nloglf,'start',start)` returns parameter estimates for the custom distribution specified by the negative loglikelihood function `nloglf`. You must also specify the initial parameter values, `start`.

`phat = mle(___,Name,Value)` specifies options using name-value pair arguments in addition to any of the input arguments in previous syntaxes. For example, you can specify the censored data, frequency of observations, and confidence level.

`[phat,pci] = mle( ___ )` also returns the 95% confidence intervals for the parameters.

Examples

Estimate Parameters of Burr Distribution

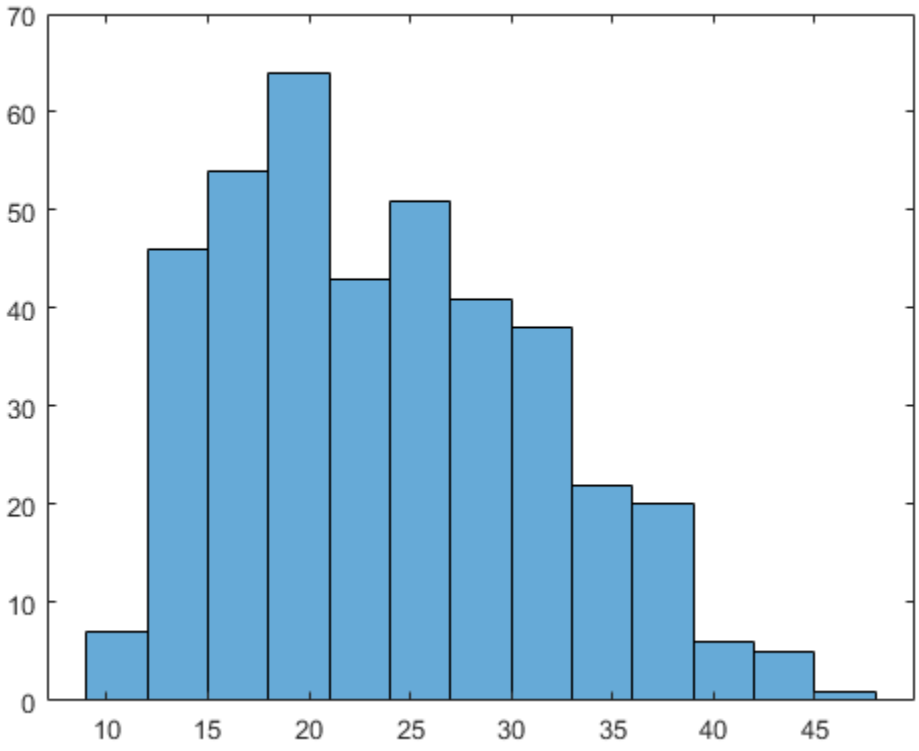
Load the sample data.

```
load carbig
```

The variable `MPG` has the miles per gallon for different models of cars.

Draw a histogram of `MPG` data.

```
histogram(MPG)
```



The distribution is somewhat right skewed. A symmetric distribution, such as normal distribution, might not be a good fit.

Estimate the parameters of the Burr Type XII distribution for the `MPG` data.

```
phat = mle(MPG,'distribution','burr')
```

```
phat = 1×3
```

```
   34.6447    3.7898    3.5722
```

The maximum likelihood estimates for the scale parameter α is 34.6447. The estimates for the two shape parameters *c* and *k* of the Burr Type XII distribution are 3.7898 and 3.5722, respectively.

Estimate Parameters of a Noncentral Chi-Square Distribution

Generate sample data of size 1000 from a noncentral chi-square distribution with degrees of freedom 8 and noncentrality parameter 3.

```
rng default % for reproducibility
x = ncx2rnd(8,3,1000,1);
```

Estimate the parameters of the noncentral chi-square distribution from the sample data. To do this, custom define the noncentral chi-square pdf using the `pdf` input argument.

```
[phat,pci] = mle(x,'pdf',@(x,v,d)ncx2pdf(x,v,d),'start',[1,1])
```

```
phat = 1×2

    8.1052    2.6693


pci = 2×2

    7.1120    1.6025
    9.0983    3.7362
```

The estimate for the degrees of freedom is 8.1052 and the noncentrality parameter is 2.6693. The 95% confidence interval for the degrees of freedom is (7.1121,9.0983) and the noncentrality parameter is (1.6025,3.7362). The confidence intervals include the true parameter values of 8 and 3, respectively.

Fit Custom Distribution to Censored Data

Load the sample data.

```
load('readmissiontimes.mat');
```

The data includes `ReadmissionTime`, which has readmission times for 100 patients. The column vector `Censored` has the censorship information for each patient, where 1 indicates a censored observation, and 0 indicates the exact readmission time is observed. This is simulated data.

Define a custom probability density and cumulative distribution function.

```
custpdf = @(data,lambda) lambda*exp(-lambda*data);
custcdf = @(data,lambda) 1-exp(-lambda*data);
```

Estimate the parameter, `lambda`, of the custom distribution for the censored sample data.

```
phat = mle(ReadmissionTime,'pdf',custpdf,'cdf',custcdf,'start',0.05,'Censoring',Censored)
```

```
phat = 0.1096
```

Fit Custom Log pdf and Survival Function

Load the sample data.

```
load('readmissiontimes.mat');
```

The data includes `ReadmissionTime`, which has readmission times for 100 patients. The column vector `Censored` has the censorship information for each patient, where 1 indicates a censored observation, and 0 indicates the exact readmission time is observed. This is simulated data.

Define a custom log probability density and survival function.

```
custlogpdf = @(data,lambda,k) log(k)-k*log(lambda)+(k-1)*log(data)-(data/lambda).^k;
custlogsf = @(data,lambda,k) -(data/lambda).^k;
```

Estimate the parameters, `lambda` and `k`, of the custom distribution for the censored sample data.

```
phat = mle(ReadmissionTime,'logpdf',custlogpdf,'logsf',custlogsf,...
    'start',[1,0.75],'Censoring',Censored)
```

```
phat = 1×2

    9.2090    1.4223
```

The scale and shape parameters of the custom-defined distribution are 9.2090 and 1.4223, respectively.

Fit Custom Log Negative Likelihood Function

Load the sample data.

```
load('readmissiontimes.mat')
```

The data includes `ReadmissionTime`, which has readmission times for 100 patients. This is simulated data.

Define a negative log likelihood function.

```
custnloglf = @(lambda,data,cens,freq) - length(data)*log(lambda) + sum(lambda*data,'omitnan');
```

Estimate the parameters of the defined distribution.

```
phat = mle(ReadmissionTime,'nloglf',custnloglf,'start',0.05)
```

```
phat = 0.1462
```

Estimate Probability of Success

Generate 100 random observations from a binomial distribution with the number of trials, $n = 20$, and the probability of success, $p = 0.75$.

```
data = binornd(20,0.75,100,1);
```

Estimate the probability of success and 95% confidence limits using the simulated sample data.

```
[phat,pci] = mle(data,'distribution','binomial','alpha',.05,'ntrials',20)
```

```
phat = 0.7615
```

```
pci = 2×1

    0.7422
    0.7800
```

The estimate of probability of success is 0.7615 and the lower and upper limits of the 95% confidence interval are 0.7422 and 0.78. This interval covers the true value used to simulate the data.

Fit Distribution with Known Parameter

Generate sample data of size 1000 from a noncentral chi-square distribution with degrees of freedom 10 and noncentrality parameter 5.

```
rng default % for reproducibility
x = ncx2rnd(10,5,1000,1);
```

Suppose the noncentrality parameter is fixed at the value 5. Estimate the degrees of freedom of the noncentral chi-square distribution from the sample data. To do this, custom define the noncentral chi-square pdf using the `pdf` input argument.

```
[phat,pci] = mle(x,'pdf',@(x,v,d)ncx2pdf(x,v,5),'start',1)
```

```
phat = 9.9307
```

```
pci = 2×1

    9.5626
   10.2989
```

The estimate for the noncentrality parameter is 9.9307, with a 95% confidence interval of 9.5626 and 10.2989. The confidence interval includes the true parameter value of 10.

Fit Rician Distribution with Known Scale Parameter

Generate sample data of size 1000 from a Rician distribution with noncentrality parameter of 8 and scale parameter of 5. First create the Rician distribution.

```
r = makedist('Rician','s',8,'sigma',5);
```

Now, generate sample data from the distribution you created above.

```
rng default % For reproducibility
x = random(r,1000,1);
```

Suppose the scale parameter is known, and estimate the noncentrality parameter from sample data. To do this using `mle`, you must custom define the Rician probability density function.

```
[phat,pci] = mle(x,'pdf',@(x,s,sigma) pdf('rician',x,s,5),'start',10)
```

```
phat = 7.8953

pci = 2×1

    7.5405
    8.2501
```

The estimate for the noncentrality parameter is 7.8953, with a 95% confidence interval of 7.5404 and 8.2501. The confidence interval includes the true parameter value of 8.

Fit a Distribution with Additional Parameter

Add a scale parameter to the chi-square distribution for adapting to the scale of data and fit it. First, generate sample data of size 1000 from a chi-square distribution with degrees of freedom 5, and scale it by the factor of 100.

```
rng default % For reproducibility
x = 100*chi2rnd(5,1000,1);
```

Estimate the degrees of freedom and the scaling factor. To do this, custom define the chi-square probability density function using the `pdf` input argument. The density function requires a $1/s$ factor for data scaled by s.

```
[phat,pci] = mle(x,'pdf',@(x,v,s)chi2pdf(x/s,v)/s,'start',[1,200])
```

```
phat = 1×2

    5.1079   99.1681


pci = 2×2

    4.6862   90.1215
    5.5297  108.2146
```

The estimate for the degrees of freedom is 5.1079 and the scale is 99.1681. The 95% confidence interval for the degrees of freedom is (4.6862,5.5279) and the scale parameter is (90.1215,108.2146). The confidence intervals include the true parameter values of 5 and 100, respectively.

Input Arguments

`data` — Sample data

vector

Sample data `mle` uses to estimate the distribution parameters, specified as a vector.

Data Types: `single` | `double`

`dist` — Distribution type

`'normal'` (default) | character vector or string scalar of distribution type

Distribution type to estimate parameters for, specified as one of the following.

`dist`	Description	Parameter 1	Parameter 2	Parameter 3	Parameter 4
`'Bernoulli'`	“Bernoulli Distribution” on page B-2	`p`: probability of success for each trial	—	—	—
`'Beta'`	“Beta Distribution” on page B-6	`a`: first shape parameter	`b`: second shape parameter	—	—
`'bino'` or `'Binomial'`	“Binomial Distribution” on page B-10	`n`: number of trials	`p`: probability of success for each trial	—	—
`'BirnbaumSaunders'`	“Birnbaum-Saunders Distribution” on page B-18	β: scale parameter	γ: shape parameter	—	—
`'Burr'`	“Burr Type XII Distribution” on page B-19	α: scale parameter	`c`: first shape parameter	`k`: second shape parameter	—
`'Discrete Uniform'` or `'unid'`	“Uniform Distribution (Discrete)” on page B-168	`n`: maximum observable value	—	—	—
`'exp'` or `'Exponential'`	“Exponential Distribution” on page B-33	μ: mean	—	—	—
`'ev'` or `'Extreme Value'`	“Extreme Value Distribution” on page B-40	μ: location parameter	σ: scale parameter	—	—
`'gam'` or `'Gamma'`	“Gamma Distribution” on page B-47	`a`: shape parameter	`b`: scale parameter	—	—
`'gev'` or `'Generalized Extreme Value'`	“Generalized Extreme Value Distribution” on page B-55	`k`: shape parameter	σ: scale parameter	μ: location parameter	—
`'gp'` or `'Generalized Pareto'`	“Generalized Pareto Distribution” on page B-59	`k`: tail index (shape) parameter	σ: scale parameter	θ: threshold (location) parameter	—

dist	Description	Parameter 1	Parameter 2	Parameter 3	Parameter 4
`'geo'` or `'Geometric'`	"Geometric Distribution" on page B-63	`p`: probability parameter	—	—	—
`'hn'` or `'Half Normal'`	"Half-Normal Distribution" on page B-68	μ: location parameter	σ: scale parameter	—	—
`'InverseGaussian'`	"Inverse Gaussian Distribution" on page B-75	μ: scale parameter	λ: shape parameter	—	—
`'Logistic'`	"Logistic Distribution" on page B-85	μ: mean	σ: scale parameter	—	—
`'LogLogistic'`	"Loglogistic Distribution" on page B-86	μ: mean of logarithmic values	σ: scale parameter of logarithmic values	—	—
`'logn'` or `'LogNormal'`	"Lognormal Distribution" on page B-88	μ: mean of logarithmic values	σ: standard deviation of logarithmic values	—	—
`'Nakagami'`	"Nakagami Distribution" on page B-108	μ: shape parameter	ω: scale parameter	—	—
`'nbin'` or `'Negative Binomial'`	"Negative Binomial Distribution" on page B-109	`r`: number of successes	`p`: probability of success in a single trial	—	—
`'norm'` or `'Normal'`	"Normal Distribution" on page B-119	μ: mean	σ: standard deviation	—	—
`'poiss'` or `'Poisson'`	"Poisson Distribution" on page B-131	λ: mean	—	—	—
`'rayl'` or `'Rayleigh'`	"Rayleigh Distribution" on page B-137	`b`: scale parameter	—	—	—
`'Rician'`	"Rician Distribution" on page B-139	`s`: noncentrality parameter	σ: scale parameter	—	—
`'Stable'`	"Stable Distribution" on page B-140	α: first shape parameter	β: second shape parameter	γ: scale parameter	δ: location parameter
`'tLocationScale'`	"t Location-Scale Distribution" on page B-156	μ: location parameter	σ: scale parameter	ν: shape parameter	—
`'unif'` or `'Uniform'`	"Uniform Distribution (Continuous)" on page B-163	`a`: lower endpoint (minimum)	`b`: upper endpoint (maximum)	—	—
`'wbl'` or `'Weibull'`	"Weibull Distribution" on page B-170	`a`: scale parameter	`b`: shape parameter	—	—

Example: `'rician'`

pdf — Custom probability density function
function handle

Custom probability distribution function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of probability density values.

For example, if the name of the custom probability density function is `newpdf`, then you can specify the function handle in `mle` as follows.

Example: `@newpdf`

Data Types: `function_handle`

cdf — Custom cumulative distribution function
function handle

Custom cumulative distribution function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of cumulative probability values.

You must define `cdf` with `pdf` if data is censored and you use the `'Censoring'` name-value pair argument. If `'Censoring'` is not present, you do not have to specify `cdf` while using `pdf`.

For example, if the name of the custom cumulative distribution function is `newcdf`, then you can specify the function handle in `mle` as follows.

Example: `@newcdf`

Data Types: `function_handle`

logpdf — Custom log probability density function
function handle

Custom log probability density function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of log probability values.

For example, if the name of the custom log probability density function is `customlogpdf`, then you can specify the function handle in `mle` as follows.

Example: `@customlogpdf`

Data Types: `function_handle`

logsf — Custom log survival function
function handle

Custom log survival function on page 33-3981, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of log survival probability values.

You must define `logsf` with `logpdf` if data is censored and you use the `'Censoring'` name-value pair argument. If `'Censoring'` is not present, you do not have to specify `logsf` while using `logpdf`.

For example, if the name of the custom log survival function is `logsurvival`, then you can specify the function handle in `mle` as follows.

Example: `@logsurvival`

Data Types: `function_handle`

`nloglf` — Custom negative loglikelihood function

function handle

Custom negative loglikelihood function, specified as a function handle created using `@`.

This custom function accepts the following input arguments.

`params`	Vector of distribution parameter values. `mle` detects the number of parameters from the number of elements in `start`.
`data`	Vector of data.
`cens`	Boolean vector of censored values.
`freq`	Vector of integer data frequencies.

`nloglf` must accept all four arguments even if you do not use the `'Censoring'` or `'Frequency'` name-value pair arguments. You can write `'nloglf'` to ignore `cens` and `freq` arguments in that case.

`nloglf` returns a scalar negative loglikelihood value and optionally, a negative loglikelihood gradient vector (see the `'GradObj'` field in `'Options'`).

If the name of the custom negative log likelihood function is `negloglik`, then you can specify the function handle in `mle` as follows.

Example: `@negloglik`

Data Types: `function_handle`

`start` — Initial parameter values

scalar | vector

Initial parameter values for the custom functions, specified as a scalar value or a vector of scalar values.

Use `start` when you fit custom distributions, that is, when you use `pdf` and `cdf`, `logpdf` and `logsf`, or `nloglf` input arguments.

Example: `0.05`

Example: `[100,2]`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Censoring',Cens,'Alpha',0.01,'Options',Opt` specifies that `mle` estimates the parameters for the distribution of censored data specified by array `Cens`, computes the 99%

confidence limits for the parameter estimates, and uses the algorithm control parameters specified by the structure `Opt`.

`Censoring` — Indicator for censoring
array of 0s (default) | array of 0s and 1s

Indicator for censoring, specified as the comma-separated pair consisting of `'Censoring'` and a Boolean array of the same size as `data`. Use 1 for observations that are right censored and 0 for observations that are fully observed. The default is all observations are fully observed.

For example, if the censored data information is in the binary array called `Censored`, then you can specify the censored data as follows.

Example: `'Censoring',Censored`

`mle` supports censoring for the following distributions:

Birnbaum-Saunders	Logistic
Burr	Lognormal
Exponential	Nakagami
Extreme Value	Normal
Gamma	Rician
Inverse Gaussian	t Location-Scale
Kernel	Weibull
Log-Logistic	

Data Types: `logical`

`Frequency` — Frequency of observations
array of 1s (default) | vector of nonnegative integer counts

Frequency of observations, specified as the comma-separated pair consisting of `'Frequency'` and an array containing nonnegative integer counts, which is the same size as `data`. The default is one observation per element of `data`.

For example, if the observation frequencies are stored in an array named `Freq`, you can specify the frequencies as follows.

Example: `'Frequency',Freq`

Data Types: `single` | `double`

`Alpha` — Significance level
0.05 (default) | scalar value in the range (0,1)

Significance level for the confidence interval of parameter estimates, `pci`, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). The confidence level of `pci` is `100(1-Alpha)%` . The default is `0.05` for 95% confidence.

For example, for 99% confidence limits, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`NTrials` — Number of trials
scalar value | vector

Number of trials for the corresponding element of `data`, specified as the comma-separated pair consisting of `'Ntrials'` and a scalar or a vector of the same size as `data`.

Applies only to binomial distribution.

Example: `'Ntrials',total`

Data Types: `single` | `double`

`mu` — Location parameter
0 (default) | scalar value

Location parameter for the half-normal distribution, specified as the comma-separated pair consisting of `'mu'` and a scalar value.

Applies only to half-normal distribution.

Example: `'mu',1`

Data Types: `single` | `double`

`Options` — Fitting algorithm control parameters
structure

Fitting algorithm control parameters, specified as the comma-separated pair consisting of `'Options'` and a structure returned by `statset`.

Not applicable to all distributions.

Use the `'Options'` name-value pair argument to control details of the maximum likelihood optimization when fitting a custom distribution. For parameter names and default values, type `statset('mlecustom')`. You can set the options under a new name and use that in the name-value pair argument. `mle` interprets the following `statset` parameters for custom distribution fitting.

Parameter	Value
`'GradObj'`	Default is `'off'`. `'on'` or `'off'`, indicating whether or not `fmincon` can expect the custom function provided with the `nloglf` input argument to return the gradient vector of the negative log-likelihood as a second output. `mle` ignores `'GradObj'` when using `fminsearch`.
`'DerivStep'`	Default is `eps^(1/3)`. The relative difference, specified as a scalar or a vector the same size as `start`, used in finite difference derivative approximations when using `fmincon`, and `'GradObj'` is `'off'`. `mle` ignores `'DerivStep'` when using `fminsearch`.

Parameter	Value
'FunValCheck'	Default is 'on'. 'on' or 'off', indicating whether or not mle should check the values returned by the custom distribution functions for validity. A poor choice of starting point can sometimes cause these functions to return NaNs, infinite values, or out-of-range values if they are written without suitable error checking.
'TolBnd'	Default is 1e-6. An offset for lower and upper bounds when using fmincon. mle treats lower and upper bounds as strict inequalities, that is, open bounds. With fmincon, this is approximated by creating closed bounds inset from the specified lower and upper bounds by TolBnd.

Example: 'Options',statset('mlecustom')

Data Types: struct

LowerBound — Lower bounds for distribution parameters
– ∞ (default) | vector

Lower bounds for distribution parameters, specified as the comma-separated pair consisting of 'Lowerbound' and a vector the same size as start.

This name-value pair argument is valid only when you use the pdf and cdf, logpdf and logcdf, or nloglf input arguments.

Example: 'Lowerbound',0

Data Types: single | double

UpperBound — Upper bounds for distribution parameters
∞ (default) | vector

Upper bounds for distribution parameters, specified as the comma-separated pair consisting of 'Upperbound' and a vector the same size as start.

This name-value pair argument is valid only when you use the pdf and cdf, logpdf and logsf, or nloglf input arguments.

Example: 'Upperbound',1

Data Types: single | double

OptimFun — Optimization function
'fminsearch' (default) | 'fmincon'

Optimization function mle uses in maximizing the likelihood, specified as the comma-separated pair consisting of 'Optimfun' and either 'fminsearch' or 'fmincon'.

Default is 'fminsearch'.

You can only specify 'fmincon' if Optimization Toolbox is available.

The `'Optimfun'` name-value pair argument is valid only when you fit custom distributions, that is, when you use the `pdf` and `cdf`, `logpdf` and `logsf`, or `nloglf` input arguments.

Example: `'Optimfun','fmincon'`

Output Arguments

`phat` — Parameter estimates

scalar value | row vector

Parameter estimates, returned as a scalar value or a row vector.

`pci` — Confidence intervals for parameter estimates

2-by-*k* matrix

Confidence intervals for parameter estimates, returned as a column vector or a matrix depending on the number of parameters, hence the size of `phat`.

`pci` is a 2-by-*k* matrix, where *k* is the number of parameters `mle` estimates. The first and second rows of the `pci` show the lower and upper confidence limits, respectively.

More About

Survival Function

The survival function is the probability of survival as a function of time. It is also called the survivor function. It gives the probability that the survival time of an individual exceeds a certain value. Since the cumulative distribution function, $F(t)$, is the probability that the survival time is less than or equal to a given point in time, the survival function for a continuous distribution, $S(t)$, is the complement of the cumulative distribution function: $S(t) = 1 - F(t)$.

Tips

When you supply distribution functions, `mle` computes the parameter estimates using an iterative maximization algorithm. With some models and data, a poor choice of starting point can cause `mle` to converge to a local optimum that is not the global maximizer, or to fail to converge entirely. Even in cases for which the log-likelihood is well-behaved near the global maximum, the choice of starting point is often crucial to convergence of the algorithm. In particular, if the initial parameter values are far from the MLEs, underflow in the distribution functions can lead to infinite log-likelihoods.

See Also

Distribution Fitter | `fitdist` | `mlecov` | `statset`

Topics

“Maximum Likelihood Estimation” on page 5-22
“What Is Survival Analysis?” on page 14-2
“Estimate Parameters of Three-Parameter Weibull Distribution” on page B-175

Introduced before R2006a

mlecov

Asymptotic covariance of maximum likelihood estimators

Syntax

```
acov = mlecov(params,data,'pdf',pdf)
acov = mlecov(params,data,'pdf',pdf,'cdf',cdf)

acov = mlecov(params,data,'logpdf',logpdf)
acov = mlecov(params,data,'logpdf',logpdf,'logsf',logsf)

acov = mlecov(params,data,'nloglf',nloglf)

acov = mlecov( ___ ,Name,Value)
```

Description

`acov = mlecov(params,data,'pdf',pdf)` returns an approximation to the asymptotic covariance matrix of the maximum likelihood estimators of the parameters for a distribution specified by the custom probability density function `pdf`.

`mlecov` computes a finite difference approximation to the Hessian of the log-likelihood at the maximum likelihood estimates `params`, given the observed `data`, and returns the negative inverse of that Hessian.

`acov = mlecov(params,data,'pdf',pdf,'cdf',cdf)` returns `acov` for a distribution specified by the custom probability density function `pdf` and cumulative distribution function `cdf`.

`acov = mlecov(params,data,'logpdf',logpdf)` returns `acov` for a distribution specified by the custom log probability density function `logpdf`.

`acov = mlecov(params,data,'logpdf',logpdf,'logsf',logsf)` returns `acov` for a distribution specified by the custom log probability density function `logpdf` and custom log survival function on page 33-3988 `logsf`.

`acov = mlecov(params,data,'nloglf',nloglf)` returns `acov` for a distribution specified by the custom negative loglikelihood function `nloglf`.

`acov = mlecov( ___ ,Name,Value)` specifies options using name-value pair arguments in addition to any of the input arguments in previous syntaxes. For example, you can specify the censored data and frequency of observations.

Examples

Custom Probability Density Function

Load the sample data.

```
load carbig
```

The vector `Weight` shows the weights of 406 cars.

In the MATLAB Editor, create a function that returns the probability density function (pdf) of a lognormal distribution. Save the file in your current working folder as `lognormpdf.m`.

```
function newpdf = lognormpdf(data,mu,sigma)
newpdf = exp((-(log(data)-mu).^2)/(2*sigma^2))./(data*sigma*sqrt(2*pi));
```

Estimate the parameters, `mu` and `sigma`, of the custom-defined distribution.

```
phat = mle(Weight,'pdf',@lognormpdf,'start',[4.5 0.3])
```

```
phat =

    7.9600    0.2804
```

Compute the approximate covariance matrix of the parameter estimates.

```
acov = mlecov(phat,Weight,'pdf',@lognormpdf)
```

```
acov =

   1.0e-03 *

    0.1937   -0.0000
   -0.0000    0.0968
```

Estimate the standard errors of estimates.

```
se = sqrt(diag(acov))
```

```
se =

    0.0139
    0.0098
```

The standard error of the estimates of mu and sigma are 0.0139 and 0.0098, respectively.

Custom Log Probability Density Function

In the MATLAB Editor, create a function that returns the log probability density function of a beta distribution. Save the file in your current working folder as `betalogpdf.m`.

```
function logpdf = betalogpdf(x,a,b)
logpdf = (a-1)*log(x)+(b-1)*log(1-x)-betaln(a,b);
```

Generate sample data from a beta distribution with parameters 1.23 and 3.45 and estimate the parameters using the simulated data.

```
rng('default')
x = betarnd(1.23,3.45,25,1);
phat = mle(x,'dist','beta')
```

```
phat =

    1.1213    2.7182
```

Compute the approximate covariance matrix of the parameter estimates.

```
acov = mlecov(phat,x,'logpdf',@betalogpdf)
```

```
acov =

    0.0810    0.1646
    0.1646    0.6074
```

Custom Log pdf and Survival Function

Load the sample data.

```
load('readmissiontimes.mat');
```

The sample data includes `ReadmissionTime`, which has readmission times for 100 patients. The column vector `Censored` has the censorship information for each patient, where 1 indicates a censored observation, and 0 indicates the exact readmission time is observed. This is simulated data.

Define a custom log probability density and survival function.

```
custlogpdf = @(data,lambda,k) log(k)-k*log(lambda)...
             +(k-1)*log(data)-(data/lambda).^k;
custlogsf = @(data,lambda,k) -(data/lambda).^k;
```

Estimate the parameters, `lambda` and `k`, of the custom distribution for the censored sample data.

```
phat = mle(ReadmissionTime,'logpdf',custlogpdf,...
'logsf',custlogsf,'start',[1,0.75],'Censoring',Censored)
```

```
phat = 1×2

    9.2090    1.4223
```

The scale and shape parameters of the custom-defined distribution are 9.2090 and 1.4223, respectively.

Compute the approximate covariance matrix of the parameter estimates.

```
acov = mlecov(phat,ReadmissionTime,...
'logpdf',custlogpdf,'logsf',custlogsf,'Censoring',Censored)
```

```
acov = 2×2

    0.5653    0.0102
    0.0102    0.0163
```

Custom Log Negative Likelihood Function

Load the sample data.

```
load('readmissiontimes.mat')
```

The sample data includes `ReadmissionTime`, which has readmission times for 100 patients. This is simulated data.

Define a negative log likelihood function.

```
custnloglf = @(lambda,data,cens,freq) -length(data)*log(lambda) ...
    + sum(lambda*data,'omitnan');
```

Estimate the parameters of the defined distribution.

```
phat = mle(ReadmissionTime,'nloglf',custnloglf,'start',0.05)
```

```
phat = 0.1462
```

Compute the variance of the parameter estimate.

```
acov = mlecov(phat,ReadmissionTime,'nloglf',custnloglf)
```

```
acov = 2.1374e-04
```

Compute the standard error.

```
sqrt(acov)
```

```
ans = 0.0146
```

Input Arguments

`params` — Parameter estimates

scalar value | vector

Parameter estimates, specified as a scalar value or vector of scalar values. These parameter estimates must be maximum likelihood estimates. For example, you can specify parameter estimates returned by `mle`.

Data Types: `single` | `double`

`data` — Sample data

vector

Sample data `mle` uses to estimate the distribution parameters, specified as a vector.

Data Types: `single` | `double`

`pdf` — Custom probability density function

function handle

Custom probability distribution function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of probability density values.

For example, if the name of the custom probability density function is `newpdf`, then you can specify the function handle in `mlecov` as follows.

Example: `@newpdf`

Data Types: `function_handle`

`cdf` — Custom cumulative distribution function
function handle

Custom cumulative distribution function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of cumulative probability values.

You must define `cdf` with `pdf` if data is censored and you use the `'Censoring'` name-value pair argument. If `'Censoring'` is not present, you do not have to specify `cdf` while using `pdf`.

For example, if the name of the custom cumulative distribution function is `newcdf`, then you can specify the function handle in `mlecov` as follows.

Example: `@newcdf`

Data Types: `function_handle`

`logpdf` — Custom log probability density function
function handle

Custom log probability density function, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of log probability values.

For example, if the name of the custom log probability density function is `customlogpdf`, then you can specify the function handle in `mlecov` as follows.

Example: `@customlogpdf`

Data Types: `function_handle`

`logsf` — Custom log survival function
function handle

Custom log survival function on page 33-3988, specified as a function handle created using `@`.

This custom function accepts the vector `data` and one or more individual distribution parameters as input parameters, and returns a vector of log survival probability values.

You must define `logsf` with `logpdf` if data is censored and you use the `'Censoring'` name-value pair argument. If `'Censoring'` is not present, you do not have to specify `logsf` while using `logpdf`.

For example, if the name of the custom log survival function is `logsurvival`, then you can specify the function handle in `mlecov` as follows.

Example: `@logsurvival`

Data Types: `function_handle`

`nloglf` — Custom negative loglikelihood function
function handle

Custom negative loglikelihood function, specified as a function handle created using `@`.

This custom function accepts the following input arguments.

`params`	Vector of distribution parameter values
`data`	Vector of data
`cens`	Boolean vector of censored values
`freq`	Vector of integer data frequencies

`nloglf` must accept all four arguments even if you do not use the `'Censoring'` or `'Frequency'` name-value pair arguments. You can write `'nloglf'` to ignore `cens` and `freq` arguments in that case.

`nloglf` returns a scalar negative loglikelihood value and optionally, a negative loglikelihood gradient vector (see the `'GradObj'` field in `'Options'`).

If the name of the custom negative log likelihood function is `negloglik`, then you can specify the function handle in `mlecov` as follows.

Example: `@negloglik`

Data Types: `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Censoring',cens,'Options',opt` specifies that `mlecov` reads the censored data information from the vector `cens` and performs according to the new options structure `opt`.

`Censoring` — Indicator for censoring

array of 0s (default) | array of 0s and 1s

Indicator for censoring, specified as the comma-separated pair consisting of `'Censoring'` and a Boolean array of the same size as `data`. Use 1 for observations that are right censored and 0 for observations that are fully observed. The default is all observations are fully observed.

For censored data, you must use `cdf` with `pdf`, or `logsf` with `logpdf`, or `nloglf` must be defined to account for censoring.

For example, if the censored data information is in the binary array called `Censored`, then you can specify the censored data as follows.

Example: `'Censoring',Censored`

Data Types: `logical`

`Frequency` — Frequency of observations

array of 1s (default) | vector of nonnegative integer counts

Frequency of observations, specified as the comma-separated pair consisting of `'Frequency'` and an array containing nonnegative integer counts, which is the same size as `data`. The default is one observation per element of `data`.

For example, if the observation frequencies are stored in an array named `Freq`, you can specify the frequencies as follows.

Example: `'Frequency',Freq`

Data Types: `single` | `double`

`Options` — Numerical options
structure

Numerical options for the finite difference Hessian calculation, specified as the comma-separated pair consisting of `'Options'` and a structure returned by `statset`.

You can set the options under a new name and use it in the name-value pair argument. The applicable `statset` parameters are as follows.

Parameter	Value
`'GradObj'`	Default is `'off'`. `'on'` or `'off'`, indicating whether or not the function provided with the `nloglf` input argument can return the gradient vector of the negative log-likelihood as a second output.
`'DerivStep'`	Default is `eps^(1/4)`. Relative step size used in finite difference for Hessian calculations. It can be a scalar, or the same size as `params`. A smaller value than the default might be appropriate if `'GradObj'` is `'on'`.

Example: `'Options',statset('mlecov')`

Data Types: `struct`

Output Arguments

`acov` — Approximation to asymptotic covariance matrix
p-by-*p* matrix

Approximation to asymptotic covariance matrix, returned as a *p*-by-*p* matrix, where *p* is the number of parameters in `params`.

More About

Survival Function

The survival function is the probability of survival as a function of time. It is also called the survivor function. It gives the probability that the survival time of an individual exceeds a certain value. Since the cumulative distribution function, $F(t)$, is the probability that the survival time is less than or equal to a given point in time, the survival function for a continuous distribution, $S(t)$, is the complement of the cumulative distribution function: $S(t) = 1 - F(t)$.

See Also
`mle`

Topics
“What Is Survival Analysis?” on page 14-2

Introduced before R2006a

mnpdf

Multinomial probability density function

Syntax

```
Y = mnpdf(X,PROB)
```

Description

`Y = mnpdf(X,PROB)` returns the pdf for the multinomial distribution with probabilities `PROB`, evaluated at each row of `X`. `X` and `PROB` are *m*-by-*k* matrices or 1-by-*k* vectors, where *k* is the number of multinomial bins or categories. Each row of `PROB` must sum to one, and the sample sizes for each observation (rows of `X`) are given by the row sums `sum(X,2)`. `Y` is an *m*-by-1 vector, and `mnpdf` computes each row of `Y` using the corresponding rows of the inputs, or replicates them if needed.

Examples

Compute the Multinomial Distribution pdf

Compute the pdf of a multinomial distribution with a sample size of `n = 10`. The probabilities are `p = 1/2` for outcome 1, `p = 1/3` for outcome 2, and `p = 1/6` for outcome 3.

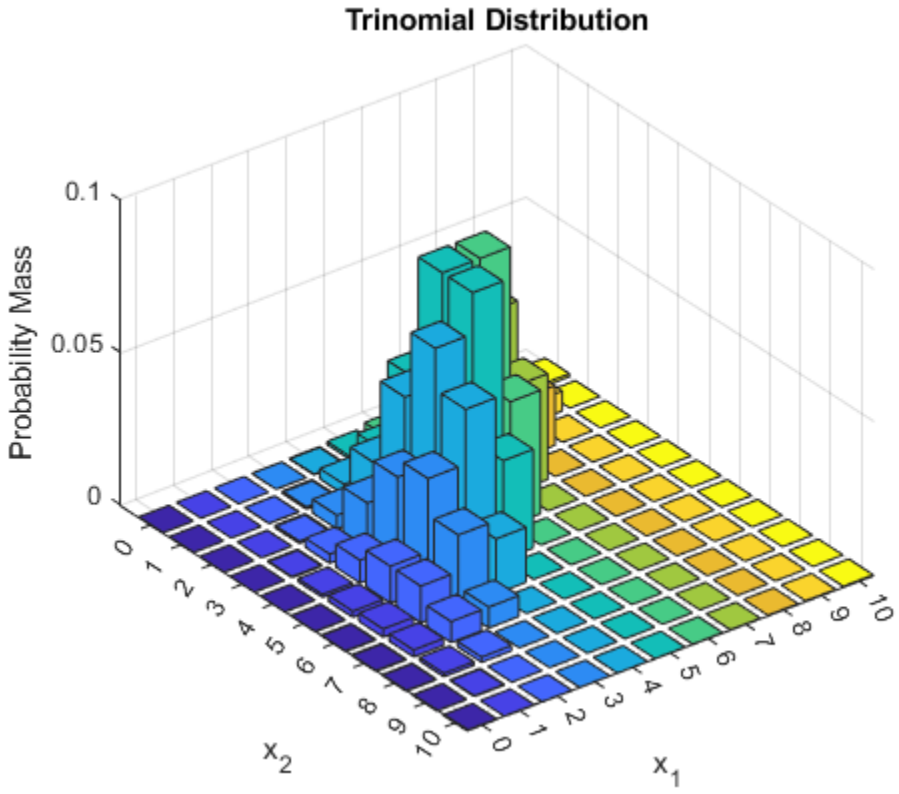
```
p = [1/2 1/3 1/6];
n = 10;
x1 = 0:n;
x2 = 0:n;
[X1,X2] = meshgrid(x1,x2);
X3 = n-(X1+X2);
```

Compute the pdf of the distribution.

```
Y = mnpdf([X1(:),X2(:),X3(:)],repmat(p,(n+1)^2,1));
```

Plot the pdf on a 3-dimensional figure.

```
Y = reshape(Y,n+1,n+1);
bar3(Y)
h = gca;
h.XTickLabel = [0:n];
h.YTickLabel = [0:n];
xlabel('x_1')
ylabel('x_2')
zlabel('Probability Mass')
title('Trinomial Distribution')
```



Note that the visualization does not show `x3`, which is determined by the constraint `x1 + x2 + x3 = n`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`mnrnd`

Topics

“Multinomial Distribution” on page B-96

Introduced in R2006b

mnrfit

Multinomial logistic regression

Syntax

```
B = mnrfit(X,Y)
B = mnrfit(X,Y,Name,Value)
[B,dev,stats] = mnrfit( ___ )
```

Description

`B = mnrfit(X,Y)` returns a matrix, `B`, of coefficient estimates for a multinomial logistic regression of the nominal responses in `Y` on the predictors in `X`.

`B = mnrfit(X,Y,Name,Value)` returns a matrix, `B`, of coefficient estimates for a multinomial model fit with additional options specified by one or more `Name,Value` pair arguments.

For example, you can fit a nominal, an ordinal, or a hierarchical model, or change the link function.

`[B,dev,stats] = mnrfit( ___ )` also returns the deviance of the fit, `dev`, and the structure `stats` for any of the previous input arguments. `stats` contains model statistics such as degrees of freedom, standard errors for coefficient estimates, and residuals.

Examples

Multinomial Regression for Nominal Responses

Fit a multinomial regression for nominal outcomes and interpret the results.

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species, setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers, the length and width of sepals and petals in centimeters, respectively.

Define the nominal response variable using a categorical array.

```
sp = categorical(species);
```

Fit a multinomial regression model to predict the species using the measurements.

```
[B,dev,stats] = mnrfit(meas,sp);
B
```

```
B = 5×2
10^3 ×

    2.5286    0.0426
```

```
    0.8435    0.0025
   -0.7099    0.0067
   -0.6480   -0.0094
   -3.4528   -0.0183
```

This is a nominal model for the response category relative risks, with separate slopes on all four predictors, that is, each category of `meas`. The first row of `B` contains the intercept terms for the relative risk of the first two response categories, setosa and versicolor versus the reference category, virginica. The last four rows contain the slopes for the models for the first two categories. `mnrfit` accepts the third category as the reference category.

The relative risk of an iris flower being species 2 (versicolor) versus species 3 (virginica) is the ratio of the two probabilities (the probability of being species 2 and the probability of being species 3). The model for the relative risk is

$$\ln\left(\frac{\pi_{\text{versicolor}}}{\pi_{\text{virginica}}}\right) = 42.6 + 2.5X_1 + 6.7X_2 - 9.4X_3 - 18.3X_4.$$

The coefficients express both the effects of the predictor variables on the relative risk and the log odds of being in one category versus the reference category. For example, the estimated coefficient 2.5 indicates that the relative risk of being species 2 (versicolor) versus species 3 (virginica) increases exp(2.5) times for each unit increase in X_1, the first measurement, given all else is equal. The relative log odds of being versicolor versus virginica increases 2.5 times with a one-unit increase in X_1, given all else is equal.

If the coefficients are converging toward infinity or negative infinity, the estimated coefficients can vary slightly depending on your operating system.

Check the statistical significance of the model coefficients.

```
stats.p
```

```
ans = 5×2

         0    0.0000
         0    0.0281
         0    0.0000
         0    0.0000
         0    0.0000
```

The small *p*-values indicate that all measures are significant on the relative risk of being a setosa versus a virginica (species 1 compared to species 3) and being a versicolor versus a virginica (species 2 compared to species 3).

Request the standard errors of coefficient estimates.

```
stats.se
```

```
ans = 5×2

   12.4038    5.2719
    3.5783    1.1228
    3.1760    1.4789
    3.5403    1.2934
```

```
    7.1203    2.0967
```

Calculate the 95% confidence limits for the coefficients.

```
LL = stats.beta - 1.96.*stats.se;
UL = stats.beta + 1.96.*stats.se;
```

Display the confidence intervals for the coefficients of the model for the relative risk of being a setosa versus a virginica (the first column of coefficients in B).

```
[LL(:,1) UL(:,1)]
```

```
ans = 5×2
10^3 ×

    2.5043    2.5529
    0.8365    0.8505
   -0.7162   -0.7037
   -0.6550   -0.6411
   -3.4667   -3.4388
```

Find the confidence intervals for the coefficients of the model for the relative risk of being a versicolor versus a virginica (the second column of coefficients in B).

```
[LL(:,2) UL(:,2)]
```

```
ans = 5×2

   32.3049   52.9707
    0.2645    4.6660
    3.7823    9.5795
  -11.9644   -6.8944
  -22.3957  -14.1766
```

Multinomial Regression for Ordinal Responses

Fit a multinomial regression model for categorical responses with natural ordering among categories.

Load the sample data and define the predictor variables.

```
load carbig
X = [Acceleration Displacement Horsepower Weight];
```

The predictor variables are the acceleration, engine displacement, horsepower, and weight of the cars. The response variable is miles per gallon (mpg).

Create an ordinal response variable categorizing MPG into four levels from 9 to 48 mpg by labeling the response values in the range 9-19 as 1, 20-29 as 2, 30-39 as 3, and 40-48 as 4.

```
miles = ordinal(MPG,{'1','2','3','4'},[],[9,19,29,39,48]);
```

Fit an ordinal response model for the response variable `miles`.

```
[B,dev,stats] = mnrfit(X,miles,'model','ordinal');
B
```

```
B = 7×1

  -16.6895
  -11.7208
   -8.0606
    0.1048
    0.0103
    0.0645
    0.0017
```

The first three elements of B are the intercept terms for the models, and the last four elements of B are the coefficients of the covariates, assumed common across all categories. This model corresponds to *parallel regression*, which is also called the *proportional odds* model, where there is a different intercept but common slopes among categories. You can specify this using the `'interactions','off'` name-value pair argument, which is the default for ordinal models.

```
[B(1:3)'; repmat(B(4:end),1,3)]
```

```
ans = 5×3

  -16.6895  -11.7208   -8.0606
    0.1048    0.1048    0.1048
    0.0103    0.0103    0.0103
    0.0645    0.0645    0.0645
    0.0017    0.0017    0.0017
```

The link function in the model is logit (`'link','logit'`), which is the default for an ordinal model. The coefficients express the relative risk or log odds of the mpg of a car being less than or equal to one value versus greater than that value.

The proportional odds model in this example is

$$\ln\left(\frac{P(mpg \le 19)}{P(mpg > 19)}\right) = -16.6895 + 0.1048X_A + 0.0103X_D + 0.0645X_H + 0.0017X_W$$

$$\ln\left(\frac{P(mpg \le 29)}{P(mpg > 29)}\right) = -11.7208 + 0.1048X_A + 0.0103X_D + 0.0645X_H + 0.0017X_W$$

$$\ln\left(\frac{P(mpg \le 39)}{P(mpg > 39)}\right) = -8.0606 + 0.1048X_A + 0.0103X_D + 0.0645X_H + 0.0017X_W$$

For example, the coefficient estimate of 0.1048 indicates that a unit change in acceleration would impact the odds of the mpg of a car being less than or equal to 19 versus more than 19, or being less than or equal to 29 versus greater than 29, or being less than or equal to 39 versus greater than 39, by a factor of exp(0.01048) given all else is equal.

Assess the significance of the coefficients.

```
stats.p
```

```
ans = 7×1

    0.0000
```

```
    0.0000
    0.0000
    0.1899
    0.0350
    0.0000
    0.0118
```

The *p*-values of 0.035, 0.0000, and 0.0118 for engine displacement, horsepower, and weight of a car, respectively, indicate that these factors are significant on the odds of mpg of a car being less than or equal to a certain value versus being greater than that value.

Hierarchical Multinomial Regression Model

Fit a hierarchical multinomial regression model.

Load the sample data.

```
load('smoking.mat');
```

The data set `smoking` contains five variables: sex, age, weight, and systolic and diastolic blood pressure. Sex is a binary variable where 1 indicates female patients, and 0 indicates male patients.

Define the response variable.

```
Y = categorical(smoking.Smoker);
```

The data in `Smoker` has four categories:

- 0: Nonsmoker, 0 cigarettes a day
- 1: Smoker, 1–5 cigarettes a day
- 2: Smoker, 6–10 cigarettes a day
- 3: Smoker, 11 or more cigarettes a day

Define the predictor variables.

```
X = [smoking.Sex smoking.Age smoking.Weight...
    smoking.SystolicBP smoking.DiastolicBP];
```

Fit a hierarchical multinomial model.

```
[B,dev,stats] = mnrfit(X,Y,'model','hierarchical');
B
```

```
B = 6×3

   43.8148    5.9571   44.0712
    1.8709   -0.0230    0.0662
    0.0188    0.0625    0.1335
    0.0046   -0.0072   -0.0130
   -0.2170    0.0416   -0.0324
   -0.2273   -0.1449   -0.4824
```

The first column of B includes the intercept and the coefficient estimates for the model of the relative risk of being a nonsmoker versus a smoker. The second column includes the parameter estimates for modeling the log odds of smoking 1–5 cigarettes a day versus more than five cigarettes a day given that a person is a smoker. Finally, the third column includes the parameter estimates for modeling the log odds of a person smoking 6–10 cigarettes a day versus more than 10 cigarettes a day given he/she smokes more than 5 cigarettes a day.

The coefficients differ across categories. You can specify this using the `'interactions','on'` name-value pair argument, which is the default for hierarchical models. So, the model in this example is

$$\ln\left(\frac{P(y=0)}{P(y>0)}\right) = 43.8148 + 1.8709X_S + 0.0188X_A + 0.0046X_W - 0.2170X_{SBP} - 0.2273X_{DBP}$$

$$\ln\left(\frac{P(1\le y\le 5)}{P(y>5)}\right) = 5.9571 - 0.0230X_S + 0.0625X_A - 0.0072X_W + 0.0416X_{SBP} - 0.1449X_{DBP}$$

$$\ln\left(\frac{P(6\le y\le 10)}{P(y>10)}\right) = 44.0712 + 0.0662X_S + 0.1335X_A - 0.0130X_W - 0.0324X_{SBP} - 0.4824X_{DBP}$$

For example, the coefficient estimate of 1.8709 indicates that the likelihood of being a smoker versus a nonsmoker increases by exp(1.8709) = 6.49 times as the gender changes from female to male given everything else held constant.

Assess the statistical significance of the terms.

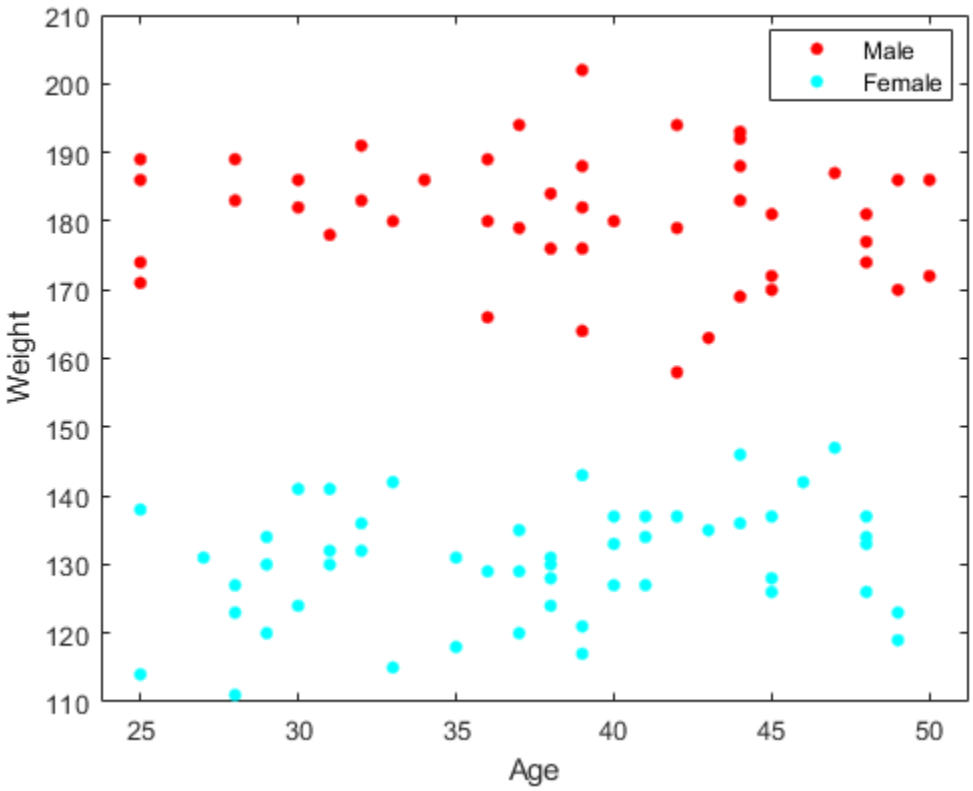
```
stats.p
```

```
ans = 6×3

    0.0000    0.5363    0.2149
    0.3549    0.9912    0.9835
    0.6850    0.2676    0.2313
    0.9032    0.8523    0.8514
    0.0009    0.5187    0.8165
    0.0004    0.0483    0.0545
```

Sex, age, or weight don’t appear significant on any level. The *p*-values of 0.0009 and 0.0004 indicate that both types of blood pressure are significant on the relative risk of a person being a smoker versus a nonsmoker. The *p*-value of 0.0483 shows that only diastolic blood pressure is significant on the odds of a person smoking 0–5 cigarettes a day versus more than 5 cigarettes a day. Similarly, the *p*-value of 0.0545 indicates that diastolic blood pressure is significant on the odds of a person smoking 6–10 cigarettes a day versus more than 10 cigarettes a day.

Check if any nonsignificant factors are correlated to each other. Draw a scatterplot of age versus weight grouped by sex.

```
figure()
gscatter(smoking.Age,smoking.Weight,smoking.Sex)
legend('Male','Female')
xlabel('Age')
ylabel('Weight')
```



The range of weight of an individual seems to differ according to gender. Age does not seem to have any obvious correlation with sex or weight. Age is insignificant and weight seems to be correlated with sex, so you can eliminate both and reconstruct the model.

Eliminate age and weight from the model and fit a hierarchical model with sex, systolic blood pressure, and diastolic blood pressure as the predictor variables.

```
X = double([smoking.Sex smoking.SystolicBP...
smoking.DiastolicBP]);
[B,dev,stats] = mnrfit(X,Y,'model','hierarchical');
B
```

```
B = 4×3

   44.8456    5.3230   25.0248
    1.6045    0.2330    0.4982
   -0.2161    0.0497    0.0179
   -0.2222   -0.1358   -0.3092
```

Here, a coefficient estimate of 1.6045 indicates that the likelihood of being a nonsmoker versus a smoker increases by exp(1.6045) = 4.97 times as sex changes from male to female. A unit increase in the systolic blood pressure indicates an exp(–.2161) = 0.8056 decrease in the likelihood of being a nonsmoker versus a smoker. Similarly, a unit increase in the diastolic blood pressure indicates an exp(–.2222) = 0.8007 decrease in the relative rate of being a nonsmoker versus being a smoker.

Assess the statistical significance of the terms.

```
stats.p

ans = 4×3

    0.0000    0.4715    0.2325
    0.0210    0.7488    0.6362
    0.0010    0.4107    0.8899
    0.0003    0.0483    0.0718
```

The *p*-values of 0.0210, 0.0010, and 0.0003 indicate that the terms sex and both types of blood pressure are significant on the relative risk of a person being a nonsmoker versus a smoker, given the other terms in the model. Based on the *p*-value of 0.0483, diastolic blood pressure appears significant on the relative risk of a person smoking 1–5 cigarettes versus more than 5 cigarettes a day, given that this person is a smoker. Because none of the *p*-values on the third column are less than 0.05, you can say that none of the variables are statistically significant on the relative risk of a person smoking from 6–10 cigarettes versus more than 10 cigarettes, given that this person smokes more than 5 cigarettes a day.

Input Arguments

X — Observations on predictor variables

n-by-*p* matrix

Observations on predictor variables, specified as an *n*-by-*p* matrix. `X` contains *n* observations for *p* predictors.

Note `mnrfit` automatically includes a constant term (intercept) in all models. Do not include a column of 1s in `X`.

Data Types: `single` | `double`

Y — Response values

n-by-*k* matrix | *n*-by-1 column vector

Response values, specified as a column vector or a matrix. `Y` can be one of the following:

- An *n*-by-*k* matrix, where $Y(i,j)$ is the number of outcomes of the multinomial category *j* for the predictor combinations given by $X(i,:)$. In this case, the number of observations are made at each predictor combination.
- An *n*-by-1 column vector of scalar integers from 1 to *k* indicating the value of the response for each observation. In this case, all sample sizes are 1.
- An *n*-by-1 categorical array indicating the nominal or ordinal value of the response for each observation. In this case, all sample sizes are 1.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Model','ordinal','Link','probit'` specifies an ordinal model with a probit link function.

`Model` — Type of model to fit
`'nominal'` (default) | `'ordinal'` | `'hierarchical'`

Type of model to fit, specified as the comma-separated pair consisting of `'Model'` and one of the following.

`'nominal'`	Default. There is no ordering among the response categories.
`'ordinal'`	There is a natural ordering among the response categories.
`'hierarchical'`	The choice of response category is sequential/nested.

Example: `'Model','ordinal'`

`Interactions` — Indicator for interaction between multinomial categories and coefficients
`'on'` | `'off'`

Indicator for an interaction between the multinomial categories and coefficients, specified as the comma-separated pair consisting of `'Interactions'` and one of the following.

`'on'`	Default for nominal and hierarchical models. Fit a model with different coefficients across categories.
`'off'`	Default for ordinal models. Fit a model with a common set of coefficients for the predictor variables, across all multinomial categories. This is often described as *parallel regression* or the *proportional odds model*.

In all cases, the model has different intercepts across categories. The choice of `'Interactions'` determines the dimensions of the output array `B`.

Example: `'Interactions','off'`

`Link` — Link function
`'logit'` (default) | `'probit'` | `'comploglog'` | `'loglog'`

Link function to use for ordinal and hierarchical models, specified as the comma-separated pair consisting of `'Link'` and one of the following.

`'logit'`	Default. $f(\gamma) = \ln(\gamma/(1 - \gamma))$
`'probit'`	$f(\gamma) = \Phi^{-1}(\gamma)$ — error term is normally distributed with variance 1
`'comploglog'`	Complementary log-log $f(\gamma) = \ln(-\ln(1 - \gamma))$
`'loglog'`	$f(\gamma) = \ln(-\ln(\gamma))$

The link function defines the relationship between response probabilities and the linear combination of predictors, $X\beta$. The link functions might be functions of cumulative or conditional probabilities based on whether the model is for an ordinal or a sequential/nested response. For example, for an ordinal model, γ represents the cumulative probability of being in categories 1 to j and the model with a logit link function as follows:

$$\ln\left(\frac{\gamma}{1-\gamma}\right) = \ln\left(\frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_k}\right) = \beta_{0j} + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

where k represents the last category.

You cannot specify the `'Link'` parameter for nominal models; these always use a multinomial logit link,

$$\ln\left(\frac{\pi_j}{\pi_r}\right) = \beta_{j0} + \beta_{j1}X_{j1} + \beta_{j2}X_{j2} + \cdots + \beta_{jp}X_{jp}, \quad j = 1, \ldots, k-1,$$

where π stands for a categorical probability, and r corresponds to the reference category. `mnrfit` uses the last category as the reference category for nominal models.

Example: `'Link','loglog'`

`EstDisp` — Indicator for estimating dispersion parameter
`'off'` (default) | `'on'`

Indicator for estimating a dispersion parameter, specified as the comma-separated pair consisting of `'EstDisp'` and one of the following.

`'off'`	Default. Use the theoretical dispersion value of 1.
`'on'`	Estimate a dispersion parameter for the multinomial distribution in computing standard errors.

Example: `'EstDisp','on'`

Output Arguments

`B` — Coefficient estimates
vector | matrix

Coefficient estimates for a multinomial logistic regression of the responses in `Y`, returned as a vector or a matrix.

- If `'Interaction'` is `'off'`, then `B` is a $k - 1 + p$ vector. The first $k - 1$ rows of `B` correspond to the intercept terms, one for each $k - 1$ multinomial categories, and the remaining p rows correspond to the predictor coefficients, which are common for all of the first $k - 1$ categories.
- If `'Interaction'` is `'on'`, then `B` is a $(p + 1)$-by-$(k - 1)$ matrix. Each column of `B` corresponds to the estimated intercept term and predictor coefficients, one for each of the first $k - 1$ multinomial categories.

The estimates for the kth category are taken to be zero as `mnrfit` takes the last category as the reference category.

`dev` — Deviance of the fit
scalar value

Deviance of the fit, returned as a scalar value. It is twice the difference between the maximum achievable log likelihood and that attained under the fitted model. This corresponds to the sum of deviance residuals,

$$dev = 2 * \sum_{i}^{n} \sum_{j}^{k} y_{ij} * \log\left(\frac{y_{ij}}{\widehat{\pi}_{ij} * m_i}\right) = \sum_{i}^{n} rd_i,$$

where rd_i are the deviance residuals. For deviance residuals see `stats`.

`stats` — Model statistics

structure

Model statistics, returned as a structure that contains the following fields.

`beta`	The coefficient estimates. These are the same as `B`.
`dfe`	Degrees of freedom for error • If `'Interactions'` is `'off'`, then degrees of freedom is $n*(k-1)-(k-1+p)$. • If `'Interactions'` is `'on'`, then degrees of freedom is $(n-p+1)*(k-1)$.
`sfit`	Estimated dispersion parameter.
`s`	Theoretical or estimated dispersion parameter. • If `'Estdisp'` is `'off'`, then `s` is the theoretical dispersion parameter, 1. • If `'Estdisp'` is `'on'`, then `s` is equal to the estimated dispersion parameter, `sfit`.
`estdisp`	Indicator for a theoretical or estimated dispersion parameter.
`se`	Standard errors of coefficient estimates, `B`.
`coeffcorr`	Estimated correlation matrix for `B`.
`covb`	Estimated covariance matrix for `B`.
`t`	*t* statistics for `B`.
`p`	*p*-values for `B`.
`resid`	Raw residuals. Observed minus fitted values, $r_{ij} = y_{ij} - \widehat{\pi}_{ij} * m_i, \quad \begin{cases} i = 1, \cdots, n \\ j = 1, \cdots, k \end{cases}$ where π_{ij} is the categorical, cumulative or conditional probability, and m_i is the corresponding sample size.
`residp`	Pearson residuals, which are the raw residuals scaled by the estimated standard deviation: $rp_{ij} = \frac{r_{ij}}{\widehat{\sigma}_{ij}} = \frac{y_{ij} - \widehat{\pi}_{ij} * m_i}{\sqrt{\widehat{\pi}_{ij} * \left(1 - \widehat{\pi}_{ij}\right) * m_i}}, \quad \begin{cases} i = 1, \cdots, n \\ j = 1, \cdots, k \end{cases}$ where π_{ij} is the categorical, cumulative, or conditional probability, and m_i is the corresponding sample size.
`residd`	Deviance residuals: $rd_i = 2 * \sum_j^k y_{ij} * \log\left(\frac{y_{ij}}{\widehat{\pi}_{ij} * m_i}\right), \quad i = 1, \cdots, n\,.$ where π_{ij} is the categorical, cumulative, or conditional probability, and m_i is the corresponding sample size.

Algorithms

`mnrfit` treats `NaN`s in either `X` or `Y` as missing values, and ignores them.

References

[1] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[2] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, 1997.

[3] Dobson, A. J., and A. G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC. Taylor & Francis Group, 2008.

See Also

`fitglm` | `glmfit` | `glmval` | `mnrval`

Topics

“Multinomial Distribution” on page B-96
“Multinomial Models for Nominal Responses” on page 12-2
“Multinomial Models for Ordinal Responses” on page 12-4
“Hierarchical Multinomial Models” on page 12-7

Introduced in R2006b

mnrnd

Multinomial random numbers

Syntax

```
r = mnrnd(n,p)
R = mnrnd(n,p,m)
R = mnrnd(N,P)
```

Description

`r = mnrnd(n,p)` returns random values `r` from the multinomial distribution with parameters `n` and `p`. `n` is a positive integer specifying the number of trials (sample size) for each multinomial outcome. `p` is a 1-by-*k* vector of multinomial probabilities, where *k* is the number of multinomial bins or categories. `p` must sum to one. (If `p` does not sum to one, `r` consists entirely of `NaN` values.) `r` is a 1-by-*k* vector, containing counts for each of the *k* multinomial bins.

`R = mnrnd(n,p,m)` returns `m` random vectors from the multinomial distribution with parameters `n` and `p`. `R` is a `m`-by-*k* matrix, where *k* is the number of multinomial bins or categories. Each row of `R` corresponds to one multinomial outcome.

`R = mnrnd(N,P)` generates outcomes from different multinomial distributions. `P` is a *m*-by-*k* matrix, where *k* is the number of multinomial bins or categories and each of the *m* rows contains a different set of multinomial probabilities. Each row of `P` must sum to one. (If any row of `P` does not sum to one, the corresponding row of `R` consists entirely of `NaN` values.) `N` is a *m*-by-1 vector of positive integers or a single positive integer (replicated by `mnrnd` to a *m*-by-1 vector). `R` is a `m`-by-*k* matrix. Each row of `R` is generated using the corresponding rows of `N` and `P`.

Examples

Generate 2 random vectors with the same probabilities:

```
n = 1e3;
p = [0.2,0.3,0.5];
R = mnrnd(n,p,2)
R =
   215   282   503
   194   303   503
```

Generate 2 random vectors with different probabilities:

```
n = 1e3;
P = [0.2, 0.3, 0.5; ...
     0.3, 0.4, 0.3;];
R = mnrnd(n,P)
R =
   186   290   524
   290   389   321
```

See Also

`mnpdf`

Topics

“Multinomial Distribution” on page B-96

Introduced in R2006b

mnrval

Multinomial logistic regression values

Syntax

```
pihat = mnrval(B,X)
[pihat,dlow,dhi] = mnrval(B,X,stats)
[pihat,dlow,dhi] = mnrval(B,X,stats,Name,Value)

yhat = mnrval(B,X,ssize)
[yhat,dlow,dhi] = mnrval(B,X,ssize,stats)
[yhat,dlow,dhi] = mnrval(B,X,ssize,stats,Name,Value)
```

Description

`pihat = mnrval(B,X)` returns the predicted probabilities for the multinomial logistic regression model with predictors, `X`, and the coefficient estimates, `B`.

`pihat` is an *n*-by-*k* matrix of predicted probabilities for each multinomial category. `B` is the vector or matrix that contains the coefficient estimates returned by `mnrfit`. And `X` is an *n*-by-*p* matrix which contains *n* observations for *p* predictors.

Note `mnrval` automatically includes a constant term in all models. Do not enter a column of 1s in `X`.

`[pihat,dlow,dhi] = mnrval(B,X,stats)` also returns 95% error bounds on the predicted probabilities, `pihat`, using the statistics in the structure, `stats`, returned by `mnrfit`.

The lower and upper confidence bounds for `pihat` are `pihat` minus `dlow` and `pihat` plus `dhi`, respectively. Confidence bounds are nonsimultaneous and only apply to the fitted curve, not to new observations.

`[pihat,dlow,dhi] = mnrval(B,X,stats,Name,Value)` returns the predicted probabilities and 95% error bounds on the predicted probabilities `pihat`, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the model type, link function, and the type of probabilities to return.

`yhat = mnrval(B,X,ssize)` returns the predicted category counts for sample sizes, `ssize`.

`[yhat,dlow,dhi] = mnrval(B,X,ssize,stats)` also computes 95% error bounds on the predicted counts `yhat`, using the statistics in the structure, `stats`, returned by `mnrfit`.

The lower and upper confidence bounds for `yhat` are `yhat` minus `dlo` and `yhat` plus `dhi`, respectively. Confidence bounds are nonsimultaneous and they apply to the fitted curve, not to new observations.

`[yhat,dlow,dhi] = mnrval(B,X,ssize,stats,Name,Value)` returns the predicted category counts and 95% error bounds on the predicted counts `yhat`, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the model type, link function, and the type of predicted counts to return.

Examples

Estimate Category Probabilities for Nominal Responses

Fit a multinomial regression for nominal outcomes and estimate the category probabilities.

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species, setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers, the length and width of sepals and petals in centimeters, respectively.

Define the nominal response variable.

```
sp = nominal(species);
sp = double(sp);
```

Now in `sp`, 1, 2, and 3 indicate the species setosa, versicolor, and virginica, respectively.

Fit a nominal model to estimate the species using the flower measurements as the predictor variables.

```
[B,dev,stats] = mnrfit(meas,sp);
```

Estimate the probability of being a certain kind of species for an iris flower having the measurements (6.3, 2.8, 4.9, 1.7).

```
x = [6.3, 2.8, 4.9, 1.7];
pihat = mnrval(B,x);
pihat
```

```
pihat = 1×3

         0    0.3977    0.6023
```

The probability of an iris flower having the measurements (6.3, 2.8, 4.9, 1.7) being a setosa is 0, a versicolor is 0.3977, and a virginica is 0.6023.

Estimate Upper and Lower Error Bounds for Probability Estimates of Ordinal Responses

Fit a multinomial regression model for categorical responses with natural ordering among categories. Then estimate the upper and lower confidence bounds for the category probability estimates.

Load the sample data and define the predictor variables.

```
load('carbig.mat')
X = [Acceleration Displacement Horsepower Weight];
```

The predictor variables are the acceleration, engine displacement, horsepower, and the weight of the cars. The response variable is miles per gallon (MPG).

Create an ordinal response variable categorizing MPG into four levels from 9 to 48 mpg.

```
miles = ordinal(MPG,{'1','2','3','4'},[],[9,19,29,39,48]);
miles = double(miles);
```

Now in miles, 1 indicates the cars with miles per gallon from 9 to 19, and 2 indicates the cars with miles per gallon from 20 to 29. Similarly, 3 and 4 indicate the cars with miles per gallon from 30 to 39 and 40 to 48, respectively.

Fit a multinomial regression model for the response variable `miles`. For an ordinal model, the default `'link'` is `logit` and the default `'interactions'` is `'off'`.

```
[B,dev,stats] = mnrfit(X,miles,'model','ordinal');
```

Compute the probability estimates and 95% error bounds for probability confidence intervals for miles per gallon of a car with x = (12, 113, 110, 2670).

```
x = [12,113,110,2670];
[pihat,dlow,hi] = mnrval(B,x,stats,'model','ordinal');
pihat
```

```
pihat = 1×4

    0.0615    0.8426    0.0932    0.0027
```

Calculate the confidence bounds for the category probability estimates.

```
LL = pihat - dlow;
UL = pihat + hi;
[LL;UL]
```

```
ans = 2×4

    0.0073    0.7829    0.0283   -0.0003
    0.1157    0.9022    0.1580    0.0057
```

Estimate Category Counts and Error Bounds for Nominal Responses

Fit a multinomial regression for nominal outcomes and estimate the category counts.

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species, setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers, the length and width of sepals and petals in centimeters, respectively.

Define the nominal response variable.

```
sp = nominal(species);
sp = double(sp);
```

Now in `sp`, 1, 2, and 3 indicate the species setosa, versicolor, and virginica, respectively.

Fit a nominal model to estimate the species based on the flower measurements.

```
[B,dev,stats] = mnrfit(meas,sp);
```

Estimate the number in each species category for a sample of 100 iris flowers all with the measurements (6.3, 2.8, 4.9, 1.7).

```
x = [6.3, 2.8, 4.9, 1.7];
yhat = mnrval(B,x,18)
```

```
yhat = 1×3

         0    7.1578   10.8422
```

Estimate the error bounds for the counts.

```
[yhat,dlow,hi] = mnrval(B,x,18,stats,'model','nominal');
```

Calculate the confidence bounds for the category probability estimates.

```
LL = yhat - dlow;
UL = yhat + hi;
[LL;UL]
```

```
ans = 2×3

         0    3.3019    6.9863
         0   11.0137   14.6981
```

Plot the Count Estimates

Create sample data with one predictor variable and a categorical response variable with three categories.

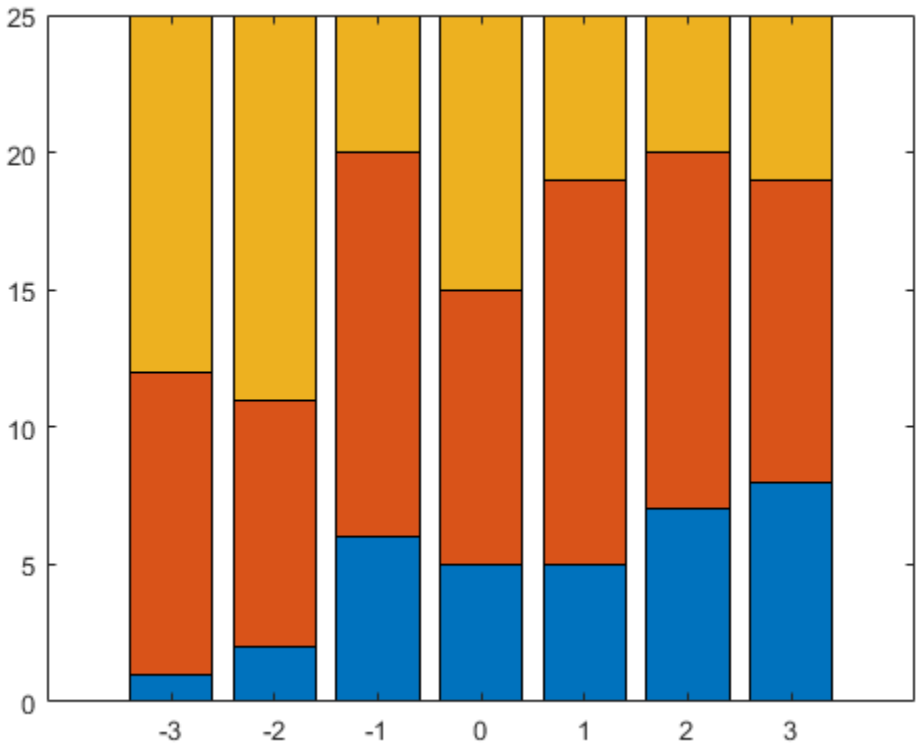
```
x = [-3 -2 -1 0 1 2 3]';
Y = [1 11 13; 2 9 14; 6 14 5; 5 10 10;...
         5 14 6; 7 13 5; 8 11 6];
[Y x]
```

```
ans = 7×4

     1    11    13    -3
     2     9    14    -2
     6    14     5    -1
     5    10    10     0
     5    14     6     1
     7    13     5     2
     8    11     6     3
```

There are observations on seven different values of the predictor variable x . The response variable Y has three categories and the data shows how many of the 25 individuals are in each category of Y for each observation of x. For example, when x is -3, 1 of 25 individuals is observed in category 1, 11 observed in category 2, and 13 observed in category 3. Similarly, when x is 1, 5 of the individuals are observed in category 1, 14 are observed in category 2, and 6 are observed in category 3.

Plot the number in each category versus the x values, on a stacked bar graph.

```
bar(x,Y,'stacked');
ylim([0 25]);
```



Fit a nominal model for the individual response category probabilities, with separate slopes on the single predictor variable, x, for each category.

```
betaHatNom = mnrfit(x,Y,'model','nominal',...
    'interactions','on')
```

```
betaHatNom = 2×2

   -0.6028    0.3832
    0.4068    0.1948
```

The first row of `betaHatOrd` contains the intercept terms for the first two response categories. The second row contains the slopes. `mnrfit` accepts the third category as the reference category and hence assumes the coefficients for the third category are zero.

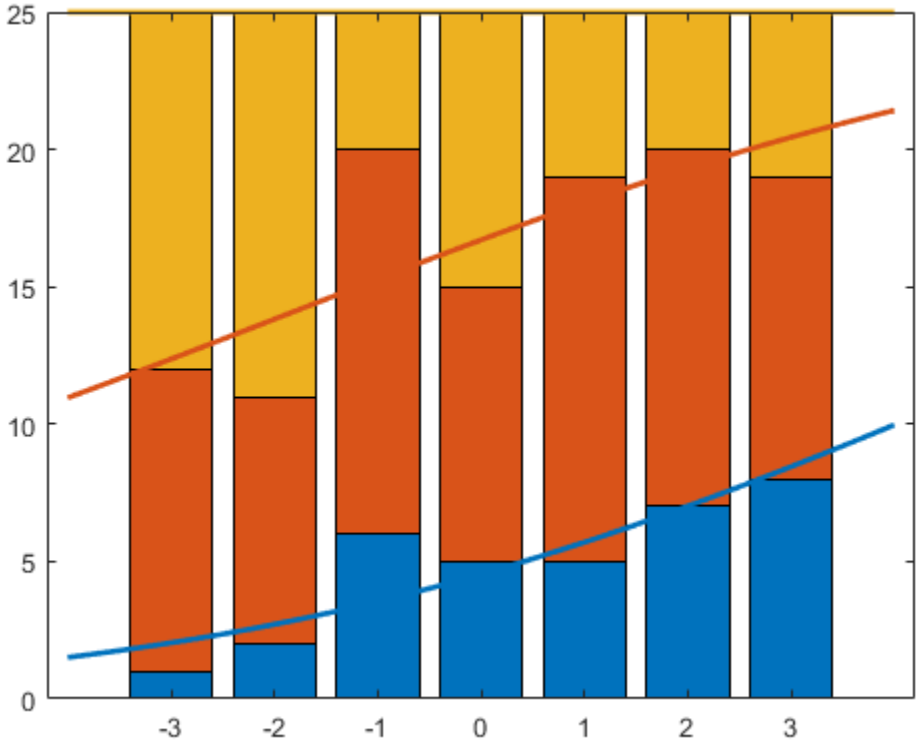
Compute the predicted probabilities for the three response categories.

```
xx = linspace(-4,4)';
piHatNom = mnrval(betaHatNom,xx,'model','nominal',...
    'interactions','on');
```

The probability of being in the third category is simply 1 - P(y = 1) - P(y = 2).

Plot the estimated cumulative number in each category on the bar graph.

```
line(xx,cumsum(25*piHatNom,2),'LineWidth',2);
```



The cumulative probability for the third category is always 1.

Now, fit a "parallel" ordinal model for the cumulative response category probabilities, with a common slope on the single predictor variable, `x`, across all categories:

```
betaHatOrd = mnrfit(x,Y,'model','ordinal',...
    'interactions','off')
```

```
betaHatOrd = 3×1

   -1.5001
    0.7266
    0.2642
```

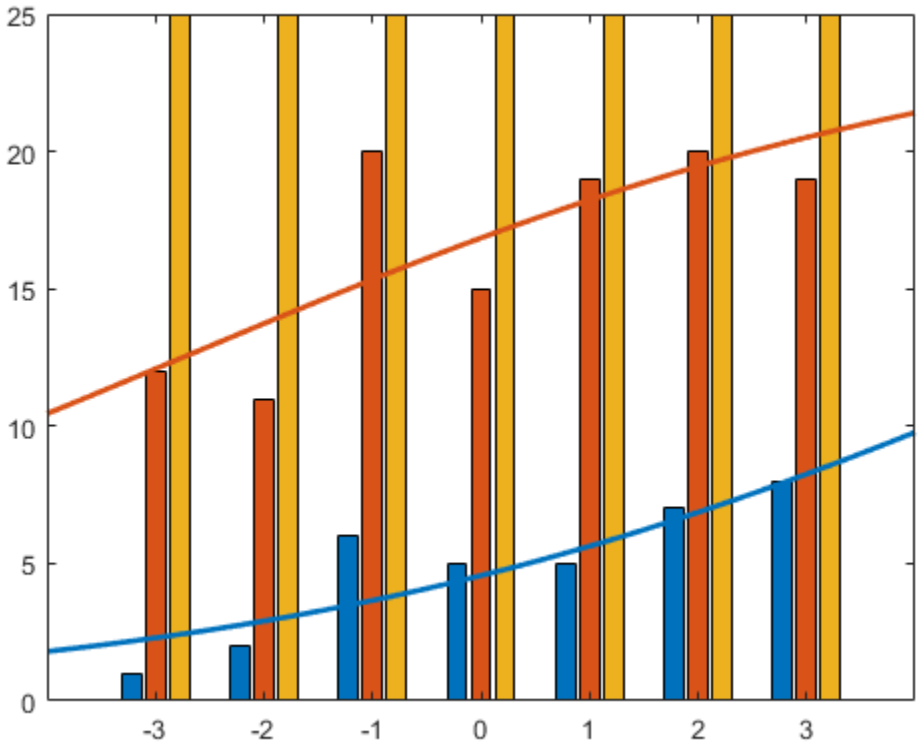
The first two elements of `betaHatOrd` are the intercept terms for the first two response categories. The last element of `betaHatOrd` is the common slope.

Compute the predicted cumulative probabilities for the first two response categories. The cumulative probability for the third category is always 1.

```
piHatOrd = mnrval(betaHatOrd,xx,'type','cumulative',...
    'model','ordinal','interactions','off');
```

Plot the estimated cumulative number on the bar graph of the observed cumulative number.

```
figure()
bar(x,cumsum(Y,2),'grouped');
ylim([0 25]);
line(xx,25*piHatOrd,'LineWidth',2);
```



Input Arguments

B — Coefficient estimates

vector or matrix returned by `mnrfit`

Coefficient estimates for the multinomial logistic regression model, specified as a vector or matrix returned by `mnrfit`. It is a vector or matrix depending on the model and interactions.

Example: `B = mnrfit(X,y); pihat = mnrval(B,X)`

Data Types: `single` | `double`

X — Sample data

matrix

Sample data on predictors, specified as an *n*-by-*p*. X contains *n* observations for *p* predictors.

Note `mnrval` automatically includes a constant term in all models. Do not enter a column of 1s in X.

Example: `pihat = mnrval(B,X)`

Data Types: `single` | `double`

`stats` — Model statistics

structure returned by `mnrfit`

Model statistics, specified as a structure returned by `mnrfit`. You must use the `stats` input argument in `mnrval` to compute the lower and upper error bounds on the category probabilities and counts.

Example: `[B,dev,stats] = mnrfit(X,y);[pihat,dlo,dhi] = mnrval(B,X,stats)`

`ssize` — Sample sizes

column vector of positive integers

Sample sizes to return the number of items in response categories for each combination of the predictor variables, specified as an *n*-by-1 column vector of positive integers.

For example, for a response variable having three categories, if an observation of the number of individuals in each category is y_1, y_2, and y_3, respectively, then the sample size, m, for that observation is $m = y_1 + y_2 + y_3$.

If the sample sizes for *n* observations are in vector `sample`, then you can enter the sample sizes as follows.

Example: `yhat = mnrval(B,X,sample)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'model','ordinal','link','probit','type','cumulative'` specifies that `mnrval` returns the estimates for cumulative probabilities for an ordinal model with a probit link function.

`model` — Type of multinomial model

`'nominal'` (default) | `'ordinal'` | `'hierarchical'`

Type of multinomial model fit by `mnrfit`, specified as the comma-separated pair consisting of `'model'` and one of the following.

`'nominal'`	Default. Specify when there is no ordering among the response categories.
`'ordinal'`	Specify when there is a natural ordering among the response categories.
`'hierarchical'`	Specify when the choice of response category is sequential.

Example: `'model','ordinal'`

`interactions` — Indicator for interaction between multinomial categories and coefficients

`'on'` | `'off'`

Indicator for an interaction between the multinomial categories and coefficients in the model fit by `mnrfit`, specified as the comma-separated pair consisting of `'interactions'` and one of the following.

`'on'`	Default for nominal and hierarchical models. Specify to fit a model with different intercepts and coefficients across categories.
`'off'`	Default for ordinal models. Specify to fit a model with different intercepts, but a common set of coefficients for the predictor variables, across all multinomial categories. This is often described as *parallel regression* or *proportional odds model*.

Example: `'interactions','off'`

Data Types: `logical`

`link` — Link function

`'logit'` (default) | `'probit'` | `'comploglog'` | `'loglog'`

Link function `mnrfit` uses for ordinal and hierarchical models, specified as the comma-separated pair consisting of `'link'` and one of the following.

`'logit'`	Default. $f(\gamma) = \ln(\gamma/(1 - \gamma))$
`'probit'`	$f(\gamma) = \Phi^{-1}(\gamma)$ — error term is normally distributed with variance 1
`'comploglog'`	Complementary log-log $f(\gamma) = \ln(-\ln(1 - \gamma))$
`'loglog'`	$f(\gamma) = \ln(-\ln(\gamma))$

The link function defines the relationship between response probabilities and the linear combination of predictors, $X\beta$.

γ might be cumulative or conditional probabilities based on whether the model is for an ordinal or a sequential/nested response.

You cannot specify the `'link'` parameter for nominal models; these always use a multinomial logit link,

$$\ln\left(\frac{\pi_j}{\pi_r}\right) = \beta_{j0} + \beta_{j1}X_{j1} + \beta_{j2}X_{j2} + \cdots + \beta_{jp}X_{jp}, \quad j = 1, \ldots, k-1,$$

where π stands for a categorical probability, and r corresponds to the reference category, k is the total number of response categories, p is the number of predictor variables. `mnrfit` uses the last category as the reference category for nominal models.

Example: `'link','loglog'`

`type` — Type of probabilities or counts to estimate

`'category'` (default) | `'cumulative'` | `'conditional'`

Type of probabilities or counts to estimate, specified as the comma-separated pair including `'type'` and one of the following.

`'category'`	Default. Specify to return predictions and error bounds for the probabilities (or counts) of the *k* multinomial categories.
`'cumulative'`	Specify to return predictions and confidence bounds for the cumulative probabilities (or counts) of the first *k* – 1 multinomial categories, as an *n*-by-(*k* – 1) matrix. The predicted cumulative probability for the *k*th category is always 1.
`'conditional'`	Specify to return predictions and error bounds in terms of the first *k* – 1 conditional category probabilities (counts), i.e., the probability (count) for category *j*, given an outcome in category *j* or higher. When `'type'` is `'conditional'`, and you supply the sample size argument `ssize`, the predicted counts at each row of `X` are conditioned on the corresponding element of `ssize`, across all categories.

Example: `'type','cumulative'`

`confidence` — Confidence level
0.95 (default) | scalar value in the range (0,1)

Confidence level for the error bounds, specified as the comma-separated pair consisting of `'confidence'` and a scalar value in the range (0,1).

For example, for 99% error bounds, you can specify the confidence as follows:

Example: `'confidence',0.99`

Data Types: `single` | `double`

Output Arguments

`pihat` — Probability estimates
n-by-(*k* – 1) matrix

Probability estimates for each multinomial category, returned as an *n*-by-(*k* – 1) matrix, where *n* is the number of observations, and *k* is the number of response categories.

`yhat` — Count estimates
n-by-*k*– 1 matrix

Count estimates for the number in each response category, returned as an *n*-by-*k* – 1 matrix, where *n* is the number of observations, and *k* is the number of response categories.

`dlow` — Lower error bound
column vector

Lower error bound to compute the lower confidence bound for `pihat` or `yhat`, returned as a column vector.

The lower confidence bound for `pihat` is `pihat` minus `dlow`. Similarly, the lower confidence bound for `yhat` is `yhat` minus `dlow`. Confidence bounds are nonsimultaneous and only apply to the fitted curve, not to new observations.

`dhi` — Upper error bound
column vector

Upper error bound to compute the upper confidence bound for `pihat` or `yhat`, returned as a column vector.

The upper confidence bound for `pihat` is `pihat` plus `dhi`. Similarly, the upper confidence bound for `yhat` is `yhat` plus `dhi`. Confidence bounds are nonsimultaneous and only apply to the fitted curve, not to new observations.

References

[1] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

See Also

`fitglm` | `glmfit` | `glmval` | `mnrfit`

Topics

“Multinomial Models for Nominal Responses” on page 12-2
“Multinomial Models for Ordinal Responses” on page 12-4
“Hierarchical Multinomial Models” on page 12-7

Introduced in R2006b

moment

Central moment

Syntax

```
m = moment(X,order)
m = moment(X,order,'all')
m = moment(X,order,dim)
m = moment(X,order,vecdim)
```

Description

`m = moment(X,order)` returns the central moment of `X` for the order specified by `order`.

- If `X` is a vector, then `moment(X,order)` returns a scalar value that is the *k*-order central moment of the elements in `X`.
- If `X` is a matrix, then `moment(X,order)` returns a row vector containing the *k*-order central moment of each column in `X`.
- If `X` is a multidimensional array, then `moment(X,order)` operates along the first nonsingleton dimension of `X`.

`m = moment(X,order,'all')` returns the central moment of the specified order for all elements of `X`.

`m = moment(X,order,dim)` takes the central moment along the operating dimension `dim` of `X`.

`m = moment(X,order,vecdim)` returns the central moment over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `moment(X,1,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the first-order central moment of the elements on the corresponding page of `X`.

Examples

Find Central Moment for Specified Order

Set the random seed for reproducibility of the results.

```
rng('default')
```

Generate a matrix with 6 rows and 5 columns.

```
X = randn(6,5)
```

```
X = 6×5

    0.5377   -0.4336    0.7254    1.4090    0.4889
    1.8339    0.3426   -0.0631    1.4172    1.0347
   -2.2588    3.5784    0.7147    0.6715    0.7269
    0.8622    2.7694   -0.2050   -1.2075   -0.3034
```

```
    0.3188   -1.3499   -0.1241    0.7172    0.2939
   -1.3077    3.0349    1.4897    1.6302   -0.7873
```

Find the third-order central moment of X.

```
m = moment(X,3)
```

```
m = 1×5

   -1.1143   -0.9973    0.1234   -1.1023   -0.1045
```

m is a row vector containing the third-order central moment of each column in X.

Find Central Moment Along Given Dimension

Find the central moment along different dimensions for a multidimensional array.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4,3,2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the fourth-order central moment of X along the default dimension.

```
m1 = moment(X,4)
```

```
m1 =
m1(:,:,1) =

   11.4427    0.3553   33.6733


m1(:,:,2) =
```

```
    0.0360    0.4902    2.3821
```

By default, `moment` operates along the first dimension of `X` whose size does not equal 1. In this case, this dimension is the first dimension of `X`. Therefore, `m1` is a 1-by-3-by-2 array.

Find the fourth-order central moment of `X` along the second dimension.

```
m2 = moment(X,4,2)
```

```
m2 =
m2(:,:,1) =

    7.3476
   13.8702
    0.4625
    2.7741


m2(:,:,2) =

    0.0341
    2.2389
    0.0171
    0.6766
```

`m2` is a 4-by-1-by-2 array.

Find the fourth-order central moment of `X` along the third dimension.

```
m3 = moment(X,4,3)
```

```
m3 = 4×3

    0.0001    0.0024    4.4627
    0.8093    3.8273   15.6340
    4.8866    0.7205    1.1412
    0.0811    0.0833    0.2433
```

`m3` is a 4-by-3 matrix.

Find Central Moment Along Vector of Dimensions

Find the central moment over multiple dimensions by using the `'all'` and `vecdim` input arguments.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4 3 2])
```

```
X =
X(:,:,1) =
```

```
    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the third-order central moment of `X`.

```
mall = moment(X,3,'all')
```

```
mall = 0.2431
```

`mall` is the third-order central moment of the entire input data set `X`.

Find the third-order moment of each page of `X` by specifying the first and second dimensions.

```
mpage = moment(X,3,[1 2])
```

```
mpage =
mpage(:,:,1) =

    0.6002


mpage(:,:,2) =

   -0.3475
```

For example, `mpage(1,1,2)` is the third-order central moment of the elements in `X(:,:,2)`.

Find the third-order moment of the elements in each `X(i,:,:)` slice by specifying the second and third dimensions.

```
mrow = moment(X,3,[2 3])
```

```
mrow = 4×1

    2.7552
    0.0443
   -0.7585
    0.5340
```

For example, `mrow(1)` is the third-order central moment of the elements in `X(1,:,:)`.

Input Arguments

`X` — Input data

vector | matrix | multidimensional array

Input data that represents a sample from a population, specified as a vector, matrix, or multidimensional array.

- If `X` is a vector, then `moment(X,order)` returns a scalar value that is the *k*-order central moment of the elements in `X`.
- If `X` is a matrix, then `moment(X,order)` returns a row vector containing the *k*-order central moment of each column in `X`.
- If `X` is a multidimensional array, then `moment(X,order)` operates along the first nonsingleton dimension of `X`.

To specify the operating dimension when `X` is a matrix or an array, use the `dim` input argument.

Data Types: `single` | `double`

`order` — Order of central moment

positive integer

Order of the central moment, specified as a positive integer.

Data Types: `single` | `double`

`dim` — Dimension

positive integer

Dimension along which to operate, specified as a positive integer. If you do not specify a value for `dim`, then the default is the first nonsingleton dimension of `X`.

Consider the third-order central moment of a matrix `X`:

- If `dim` is equal to 1, then `moment(X,3,1)` returns a row vector that contains the third-order central moment of each column in `X`.
- If `dim` is equal to 2, then `moment(X,3,2)` returns a column vector that contains the third-order central moment of each row in `X`.

If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `moment` returns an array of zeros the same size as `X`.

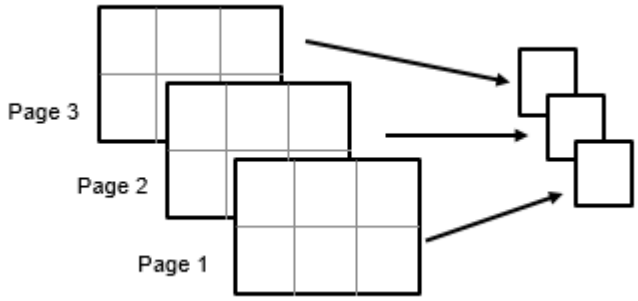
Data Types: `single` | `double`

`vecdim` — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `m` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `m`.

For example, if `X` is a 2-by-3-by-3 array, then `moment(X,1,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the first-order central moment of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

`m` — Central moments

scalar | vector | matrix | multidimensional array

Central moments, returned as a scalar, vector, matrix, or multidimensional array.

Algorithms

The central moment of order k for a distribution is defined as

$$m_k = E(x - \mu)^k,$$

where μ is the mean of x, and $E(t)$ represents the expected value of the quantity t. The `moment` function computes a sample version of this population value.

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k.$$

Note that the first-order central moment is zero, and the second-order central moment is the variance computed using a divisor of n rather than n – 1, where n is the length of the vector `x` or the number of rows in the matrix `X`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).
- If `order` is nonintegral and `X` is real, use `moment(complex(X),order)`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`kurtosis` | `mean` | `skewness` | `std` | `var`

Introduced before R2006a

multcompare

Multiple comparison test

Syntax

```
c = multcompare(stats)
c = multcompare(stats,Name,Value)
[c,m] = multcompare( ___ )
[c,m,h] = multcompare( ___ )
[c,m,h,gnames] = multcompare( ___ )
```

Description

`c = multcompare(stats)` returns a matrix `c` of the pairwise comparison results from a multiple comparison test using the information contained in the `stats` structure. `multcompare` also displays an interactive graph of the estimates and comparison intervals. Each group mean is represented by a symbol, and the interval is represented by a line extending out from the symbol. Two group means are significantly different if their intervals are disjoint; they are not significantly different if their intervals overlap. If you use your mouse to select any group, then the graph will highlight all other groups that are significantly different, if any.

`c = multcompare(stats,Name,Value)` returns a matrix of pairwise comparison results, `c`, using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence interval, or the type of critical value to use in the multiple comparison.

`[c,m] = multcompare( ___ )` also returns a matrix, `m`, which contains estimated values of the means (or whatever statistics are being compared) for each group and the corresponding standard errors. You can use any of the previous syntaxes.

`[c,m,h] = multcompare( ___ )` also returns a handle, `h`, to the comparison graph.

`[c,m,h,gnames] = multcompare( ___ )` also returns a cell array, `gnames`, which contains the names of the groups.

Examples

Multiple Comparison of Group Means

Load the sample data.

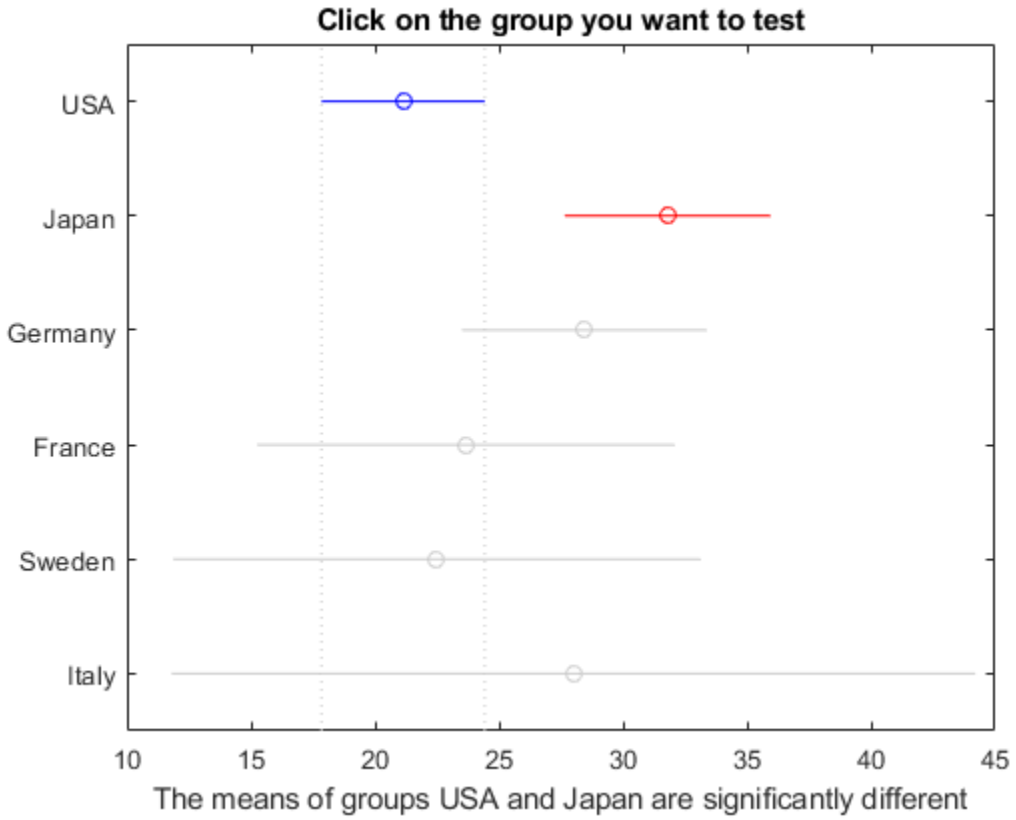
```
load carsmall
```

Perform a one-way analysis of variance (ANOVA) to see if there is any difference between the mileage of the cars by origin.

```
[p,t,stats] = anova1(MPG,Origin,'off');
```

Perform a multiple comparison of the group means.

```
[c,m,h,nms] = multcompare(stats);
```



`multcompare` displays the estimates with comparison intervals around them. You can click the graphs of each country to compare its mean to those of other countries.

Now display the mean estimates and the standard errors with the corresponding group names.

```
[nms num2cell(m)]
```

```
ans=6×3 cell array
    {'USA'    }    {[21.1328]}    {[0.8814]}
    {'Japan'  }    {[31.8000]}    {[1.8206]}
    {'Germany'}    {[28.4444]}    {[2.3504]}
    {'France' }    {[23.6667]}    {[4.0711]}
    {'Sweden' }    {[22.5000]}    {[4.9860]}
    {'Italy'  }    {[     28]}    {[7.0513]}
```

Multiple Comparisons for Two-Way ANOVA

Load the sample data.

```
load popcorn
popcorn
```

```
popcorn = 6×3
```

```
    5.5000    4.5000    3.5000
    5.5000    4.5000    4.0000
    6.0000    4.0000    3.0000
    6.5000    5.0000    4.0000
    7.0000    5.5000    5.0000
    7.0000    5.0000    4.5000
```

The data is from a study of popcorn brands and popper types (Hogg 1987). The columns of the matrix `popcorn` are brands (Gourmet, National, and Generic). The rows are popper types oil and air. In the study, researchers popped a batch of each brand three times with each popper. The values are the yield in cups of popped popcorn.

Perform a two-way ANOVA. Also compute the statistics that you need to perform a multiple comparison test on the main effects.

```
[~,~,stats] = anova2(popcorn,3,'off')
```

```
stats = struct with fields:
      source: 'anova2'
     sigmasq: 0.1389
    colmeans: [6.2500 4.7500 4]
        coln: 6
    rowmeans: [4.5000 5.5000]
        rown: 9
       inter: 1
        pval: 0.7462
          df: 12
```

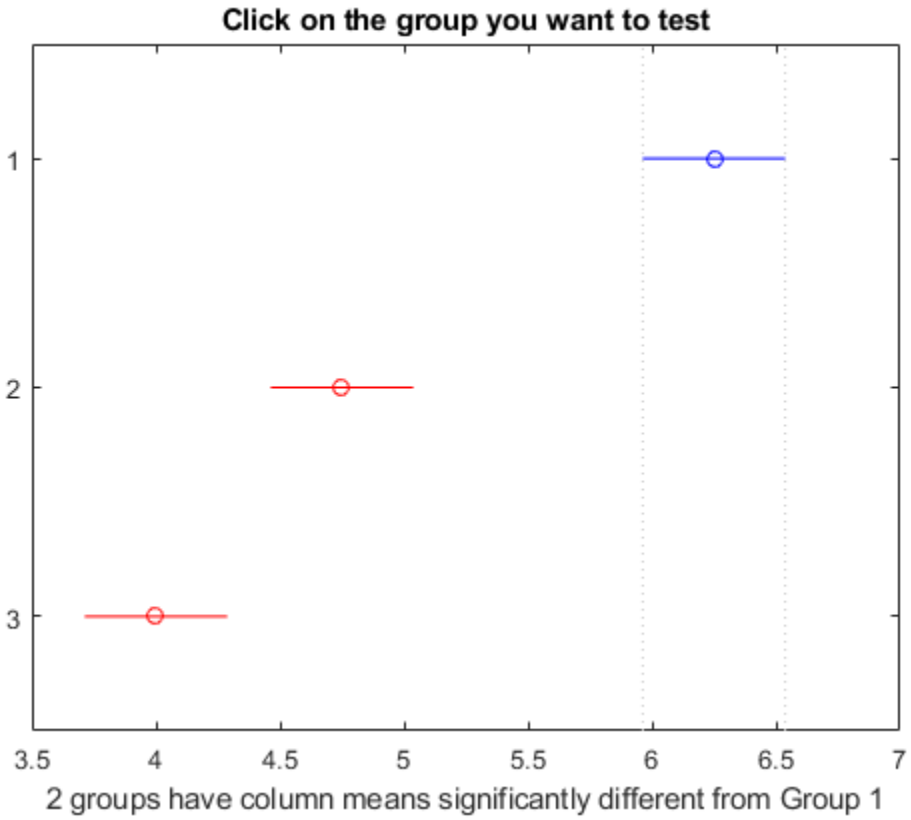
The `stats` structure includes

- The mean squared error (`sigmasq`)
- The estimates of the mean yield for each popcorn brand (`colmeans`)
- The number of observations for each popcorn brand (`coln`)
- The estimate of the mean yield for each popper type (`rowmeans`)
- The number of observations for each popper type (`rown`)
- The number of interactions (`inter`)
- The *p*-value that shows the significance level of the interaction term (`pval`)
- The error degrees of freedom (`df`).

Perform a multiple comparison test to see if the popcorn yield differs between pairs of popcorn brands (columns).

```
c = multcompare(stats)
```

```
Note: Your model includes an interaction term.  A test of main effects can be
difficult to interpret when the model includes interactions.
```




```
c = 3×6

    1.0000    2.0000    0.9260    1.5000    2.0740    0.0000
    1.0000    3.0000    1.6760    2.2500    2.8240    0.0000
    2.0000    3.0000    0.1760    0.7500    1.3240    0.0116
```

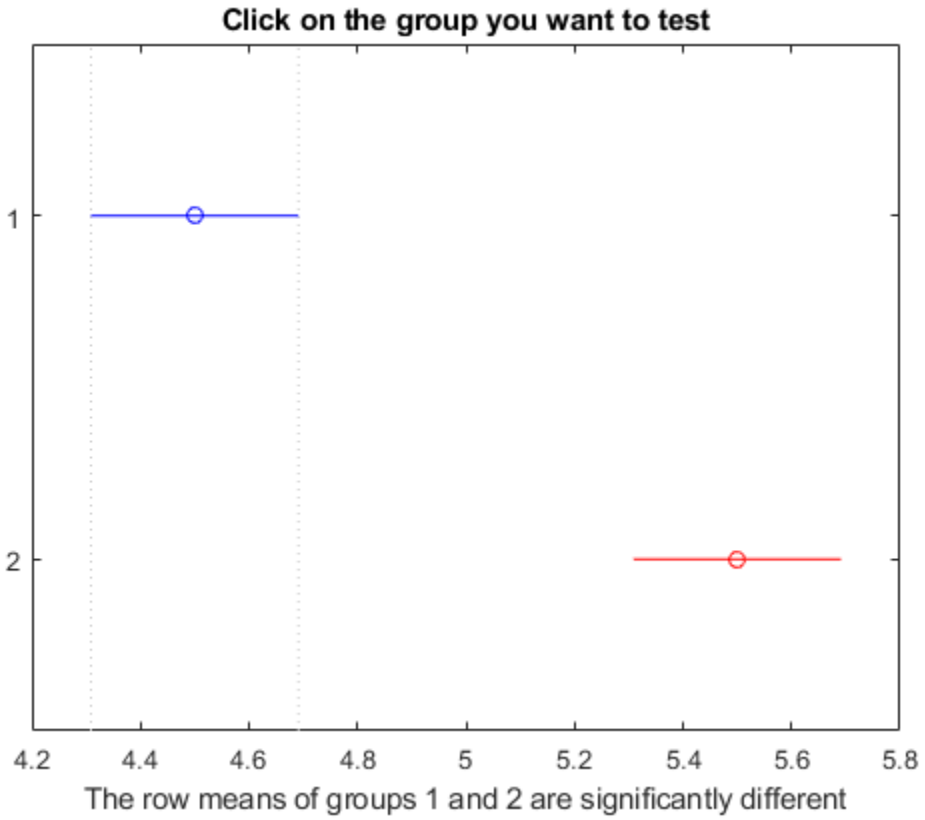
The first two columns of `c` show the groups that are compared. The fourth column shows the difference between the estimated group means. The third and fifth columns show the lower and upper limits for 95% confidence intervals for the true mean difference. The sixth column contains the *p*-value for a hypothesis test that the corresponding mean difference is equal to zero. All *p*-values (0, 0, and 0.0116) are very small, which indicates that the popcorn yield differs across all three brands.

The figure shows the multiple comparison of the means. By default, the group 1 mean is highlighted and the comparison interval is in blue. Because the comparison intervals for the other two groups do not intersect with the intervals for the group 1 mean, they are highlighted in red. This lack of intersection indicates that both means are different than group 1 mean. Select other group means to confirm that all group means are significantly different from each other.

Perform a multiple comparison test to see the popcorn yield differs between the two popper types (rows).

```
c = multcompare(stats,'Estimate','row')
```

```
Note: Your model includes an interaction term.  A test of main effects can be
difficult to interpret when the model includes interactions.
```




```
c = 1×6

    1.0000    2.0000   -1.3828   -1.0000   -0.6172    0.0001
```

The small *p*-value of 0.0001 indicates that the popcorn yield differs between the two popper types (air and oil). The figure shows the same results. The disjoint comparison intervals indicate that the group means are significantly different from each other.

Multiple Comparisons for Three-Way ANOVA

Load the sample data.

```
y = [52.7 57.5 45.9 44.5 53.0 57.0 45.9 44.0]';
g1 = [1 2 1 2 1 2 1 2];
g2 = {'hi';'hi';'lo';'lo';'hi';'hi';'lo';'lo'};
g3 = {'may';'may';'may';'may';'june';'june';'june';'june'};
```

`y` is the response vector and `g1`, `g2`, and `g3` are the grouping variables (factors). Each factor has two levels, and every observation in `y` is identified by a combination of factor levels. For example, observation `y(1)` is associated with level 1 of factor `g1`, level `'hi'` of factor `g2`, and level `'may'` of factor `g3`. Similarly, observation `y(6)` is associated with level 2 of factor `g1`, level `'hi'` of factor `g2`, and level `'june'` of factor `g3`.

Test if the response is the same for all factor levels. Also compute the statistics required for multiple comparison tests.

```
[~,~,stats] = anovan(y,{g1 g2 g3},'model','interaction',...
    'varnames',{'g1','g2','g3'});
```

Analysis of Variance

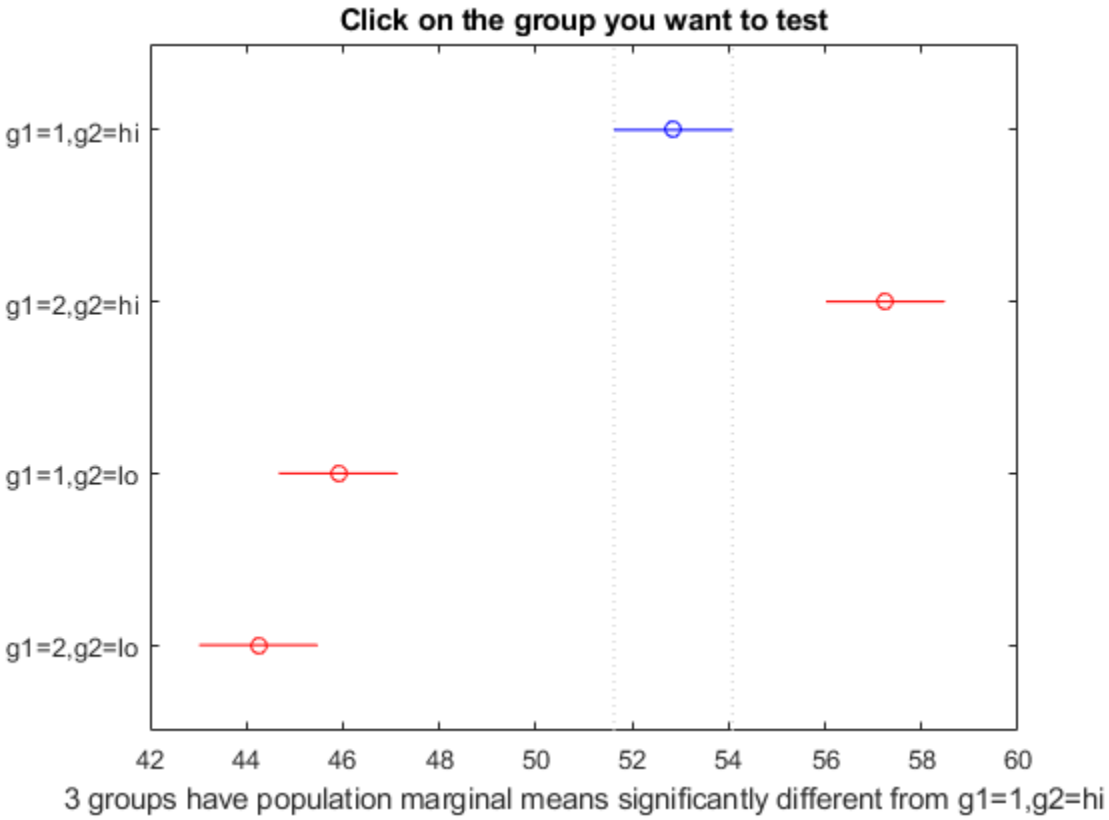
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
g1	3.781	1	3.781	336.11	0.0347
g2	199.001	1	199.001	17689	0.0048
g3	0.061	1	0.061	5.44	0.2578
g1*g2	18.301	1	18.301	1626.78	0.0158
g1*g3	0.211	1	0.211	18.78	0.1444
g2*g3	0.011	1	0.011	1	0.5
Error	0.011	1	0.011		
Total	221.379	7			

Constrained (Type III) sums of squares.

The *p*-value of 0.2578 indicates that the mean responses for levels `'may'` and `'june'` of factor `g3` are not significantly different. The *p*-value of 0.0347 indicates that the mean responses for levels `1` and `2` of factor `g1` are significantly different. Similarly, the *p*-value of 0.0048 indicates that the mean responses for levels `'hi'` and `'lo'` of factor `g2` are significantly different.

Perform multiple comparison tests to find out which groups of the factors `g1` and `g2` are significantly different.

```
results = multcompare(stats,'Dimension',[1 2])
```




```
results = 6×6

    1.0000    2.0000   -6.8604   -4.4000   -1.9396    0.0272
    1.0000    3.0000    4.4896    6.9500    9.4104    0.0170
    1.0000    4.0000    6.1396    8.6000   11.0604    0.0136
    2.0000    3.0000    8.8896   11.3500   13.8104    0.0101
    2.0000    4.0000   10.5396   13.0000   15.4604    0.0087
    3.0000    4.0000   -0.8104    1.6500    4.1104    0.0737
```

`multcompare` compares the combinations of groups (levels) of the two grouping variables, `g1` and `g2`. In the `results` matrix, the number 1 corresponds to the combination of level `1` of `g1` and level `hi` of `g2`, the number 2 corresponds to the combination of level 2 of `g1` and level `hi` of `g2`. Similarly, the number 3 corresponds to the combination of level `1` of `g1` and level `lo` of `g2`, and the number 4 corresponds to the combination of level 2 of `g1` and level `lo` of `g2`. The last column of the matrix contains the *p*-values.

For example, the first row of the matrix shows that the combination of level `1` of `g1` and level `hi` of `g2` has the same mean response values as the combination of level 2 of `g1` and level `hi` of `g2`. The *p*-value corresponding to this test is 0.0280, which indicates that the mean responses are significantly different. You can also see this result in the figure. The blue bar shows the comparison interval for the mean response for the combination of level `1` of `g1` and level `hi` of `g2`. The red bars are the comparison intervals for the mean response for other group combinations. None of the red bars overlap with the blue bar, which means the mean response for the combination of level `1` of `g1` and level `hi` of `g2` is significantly different from the mean response for other group combinations.

You can test the other groups by clicking on the corresponding comparison interval for the group. The bar you click on turns to blue. The bars for the groups that are significantly different are red. The bars for the groups that are not significantly different are gray. For example, if you click on the comparison interval for the combination of level `1` of `g1` and level `lo` of `g2`, the comparison interval for the combination of level `2` of `g1` and level `lo` of `g2` overlaps, and is therefore gray. Conversely, the other comparison intervals are red, indicating significant difference.

Input Arguments

`stats` — Test data

structure

Test data, specified as a structure. You can create a structure using one of the following functions:

- `anova1` — One-way analysis of variance.
- `anova2` — Two-way analysis of variance.
- `anovan` — *N*-way analysis of variance.
- `aoctool` — Interactive analysis of covariance tool.
- `friedman` — Friedman's test.
- `kruskalwallis` — Kruskal-Wallis test.

`multcompare` does not support multiple comparisons using `anovan` output for a model that includes random or nested effects. The calculations for a random effects model produce a warning that all effects are treated as fixed. Nested models are not accepted.

Data Types: `struct`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'CType','bonferroni','Display','off'` computes the Bonferroni critical values, conducts the hypothesis tests at the 1% significance level, and omits the interactive display.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the multiple comparison test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). The value specified for `'Alpha'` determines the 100 × (1 – α) confidence levels of the intervals returned in the matrix `c` and in the figure.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`CType` — Type of critical value

`'tukey-kramer'` (default) | `'hsd'` | `'lsd'` | `'bonferroni'` | `'dunn-sidak'` | `'scheffe'`

Type of critical value to use for the multiple comparison, specified as the comma-separated pair consisting of `'CType'` and one of the following.

Value	Description
`'tukey-kramer'` or `'hsd'`	Tukey's honest significant difference criterion on page 9-22
`'bonferroni'`	Bonferroni method on page 9-23
`'dunn-sidak'`	Dunn and Sidák's approach on page 9-23
`'lsd'`	Fisher's least significant difference procedure on page 9-23
`'scheffe'`	Scheffé's S procedure on page 9-24

Example: `'CType','bonferroni'`

`Display` — Display toggle
`'on'` (default) | `'off'`

Display toggle, specified as the comma-separated pair consisting of `'Display'` and either `'on'` or `'off'`. If you specify `'on'`, then `multcompare` displays a graph of the estimates and their comparison intervals. If you specify `'off'`, then `multcompare` omits the graph.

Example: `'Display','off'`

`Dimension` — Dimension over which to calculate marginal means
1 (default) | positive integer value | vector of positive integer values

A vector specifying the dimension or dimensions over which to calculate the population marginal means, specified as a positive integer value, or a vector of such values. Use the `'Dimension'` name-value pair only if you create the input structure `stats` using the function `anovan`.

For example, if you specify `'Dimension'` as `1`, then `multcompare` compares the means for each value of the first grouping variable, adjusted by removing effects of the other grouping variables as if the design were balanced. If you specify `'Dimension'`as `[1,3]`, then `multcompare` computes the population marginal means for each combination of the first and third grouping variables, removing effects of the second grouping variable. If you fit a singular model, some cell means may not be estimable and any population marginal means that depend on those cell means will have the value `NaN`.

Population marginal means are described by Milliken and Johnson (1992) and by Searle, Speed, and Milliken (1980). The idea behind population marginal means is to remove any effect of an unbalanced design by fixing the values of the factors specified by `'Dimension'`, and averaging out the effects of other factors as if each factor combination occurred the same number of times. The definition of population marginal means does not depend on the number of observations at each factor combination. For designed experiments where the number of observations at each factor combination has no meaning, population marginal means can be easier to interpret than simple means ignoring other factors. For surveys and other studies where the number of observations at each combination does have meaning, population marginal means may be harder to interpret.

Example: `'Dimension',[1,3]`

Data Types: `single` | `double`

`Estimate` — Estimates to be compared
`'column'` (default) | `'row'` | `'slope'` | `'intercept'` | `'pmm'`

Estimates to be compared, specified as the comma-separated pair consisting of `'Estimate'` and an allowable value. The allowable values for `'Estimate'` depend on the function used to generate the input structure `stats`, according to the following table.

Source	Values
`anova1`	None. This name-value pair is ignored, and `multcompare` always compares the group means.
`anova2`	Either `'column'` to compare column means, or `'row'` to compare row means.
`anovan`	None. This name-value pair is ignored, and `multcompare` always compares the population marginal means as specified by the `'Dimension'` name-value pair argument.
`aoctool`	Either `'slope'`, `'intercept'`, or `'pmm'` to compare slopes, intercepts, or population marginal means, respectively. If the analysis of covariance model did not include separate slopes, then `'slope'` is not allowed. If it did not include separate intercepts, then no comparisons are possible.
`friedman`	None. This name-value pair is ignored, and `multcompare` always compares the average column ranks.
`kruskalwallis`	None. This name-value pair is ignored, and `multcompare` always compares the average group ranks.

Example: `'Estimate','row'`

Output Arguments

`c` — Matrix of multiple comparison results

matrix of scalar values

Matrix of multiple comparison results, returned as an *p*-by-6 matrix of scalar values, where *p* is the number of pairs of groups. Each row of the matrix contains the result of one paired comparison test. Columns 1 and 2 contain the indices of the two samples being compared. Column 3 contains the lower confidence interval, column 4 contains the estimate, and column 5 contains the upper confidence interval. Column 6 contains the *p*-value for the hypothesis test that the corresponding mean difference is not equal to 0.

For example, suppose one row contains the following entries.

```
2.0000  5.0000  1.9442  8.2206  14.4971 0.0432
```

These numbers indicate that the mean of group 2 minus the mean of group 5 is estimated to be 8.2206, and a 95% confidence interval for the true difference of the means is [1.9442, 14.4971]. The *p*-value for the corresponding hypothesis test that the difference of the means of groups 2 and 5 is significantly different from zero is 0.0432.

In this example the confidence interval does not contain 0, so the difference is significant at the 5% significance level. If the confidence interval did contain 0, the difference would not be significant. The *p*-value of 0.0432 also indicates that the difference of the means of groups 2 and 5 is significantly different from 0.

`m` — Matrix of estimates

matrix of scalar values

Matrix of the estimates, returned as a matrix of scalar values. The first column of `m` contains the estimated values of the means (or whatever statistics are being compared) for each group, and the second column contains their standard errors.

`h` — Handle to the figure
handle

Handle to the figure containing the interactive graph, returned as a handle. The title of this graph contains instructions for interacting with the graph, and the *x*-axis label contains information about which means are significantly different from the selected mean. If you plan to use this graph for presentation, you may want to omit the title and the *x*-axis label. You can remove them using interactive features of the graph window, or you can use the following commands.

```
title('')
xlabel('')
```

`gnames` — Group names
cell array of character vectors

Group names, returned as a cell array of character vectors. Each row of `gnames` contains the name of a group.

More About

Multiple Comparison Tests

Analysis of variance compares the means of several groups to test the hypothesis that they are all equal, against the general alternative that they are not all equal. Sometimes this alternative may be too general. You may need information about which pairs of means are significantly different, and which are not. A *multiple comparison test* can provide this information.

When you perform a simple *t*-test of one group mean against another, you specify a significance level that determines the cutoff value of the *t*-statistic. For example, you can specify the value `alpha = 0.05` to insure that when there is no real difference, you will incorrectly find a significant difference no more than 5% of the time. When there are many group means, there are also many pairs to compare. If you applied an ordinary *t*-test in this situation, the `alpha` value would apply to each comparison, so the chance of incorrectly finding a significant difference would increase with the number of comparisons. Multiple comparison procedures are designed to provide an upper bound on the probability that *any* comparison will be incorrectly found significant.

References

[1] Hochberg, Y., and A. C. Tamhane. *Multiple Comparison Procedures*. Hoboken, NJ: John Wiley & Sons, 1987.

[2] Milliken, G. A., and D. E. Johnson. *Analysis of Messy Data, Volume I: Designed Experiments*. Boca Raton, FL: Chapman & Hall/CRC Press, 1992.

[3] Searle, S. R., F. M. Speed, and G. A. Milliken. “Population marginal means in the linear model: an alternative to least-squares means.” *American Statistician*. 1980, pp. 216–221.

See Also

`anova1` | `anova2` | `anovan` | `aoctool` | `friedman` | `kruskalwallis`

Introduced before R2006a

multcompare

Class: `RepeatedMeasuresModel`

Multiple comparison of estimated marginal means

Syntax

```
tbl = multcompare(rm,var)
tbl = multcompare(rm,var,Name,Value)
```

Description

`tbl = multcompare(rm,var)` returns multiple comparisons of the estimated marginal means based on the variable `var` in the repeated measures model `rm`.

`tbl = multcompare(rm,var,Name,Value)` returns multiple comparisons of the estimated marginal means with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the comparison type or which variable to group by.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`var` — Variables for which to compute marginal means

character vector | string scalar

Variables for which to compute the marginal means, specified as a character vector or string scalar representing the name of a between- or within-subjects factor in `rm`. If `var` is a between-subjects factor, it must be categorical.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range of 0 through 1

Significance level of the confidence intervals for population marginal means, specified as the comma-separated pair consisting of `'alpha'` and a scalar value in the range of 0 through 1. The confidence level is 100*(1–`alpha`)%.

Example: `'alpha',0.01`

Data Types: `double` | `single`

By — Factor to perform comparisons by
character vector | string scalar

Factor to do the comparisons by, specified as the comma-separated pair consisting of `'By'` and a character vector or string scalar. The comparison between levels of `var` occurs separately for each value of the factor you specify.

If you have more than one between-subjects factors, *A*, *B*, and *C*, and if you want to do the comparisons of *A* levels separately for each level of *C*, then specify *A* as the `var` argument and specify *C* using the `'By'` argument as follows.

Example: `'By',C`

Data Types: `char` | `string`

`ComparisonType` — Type of critical value to use
`'tukey-kramer'` (default) | `'dunn-sidak'` | `'bonferroni'` | `'scheffe'` | `'lsd'`

Type of critical value to use, specified as the comma-separated pair consisting of `'ComparisonType'` and one of the following.

Comparison Type	Definition
`'tukey-kramer'`	Default. Also called Tukey's Honest Significant Difference procedure. It is based on the Studentized range distribution. According to the unproven Tukey-Kramer conjecture, it is also accurate for problems where the quantities being compared are correlated, as in analysis of covariance with unbalanced covariate values.
`'dunn-sidak'`	Use critical values from the *t* distribution, after an adjustment for multiple comparisons that was proposed by Dunn and proved accurate by Sidák. The critical value is $$\lvert t\rvert = \frac{\lvert \bar{y}_i - \bar{y}_j \rvert}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > t_{1-\eta/2,\nu},$$ where $$\eta = 1 - (1-\alpha)^{1/\binom{k}{2}}$$ and *ng* is the number of groups (marginal means). This procedure is similar to, but less conservative than, the Bonferroni procedure.

Comparison Type	Definition
`'bonferroni'`	Use critical values from the t distribution, after a Bonferroni adjustment to compensate for multiple comparisons. The critical value is $t_{\alpha/2\binom{ng}{2},\nu}$, where ng is the number of groups (marginal means), and ν is the error degrees of freedom. This procedure is conservative, but usually less so than the Scheffé procedure.
`'scheffe'`	Use critical values from Scheffé's S procedure, derived from the F distribution. The critical value is $\sqrt{(ng-1)F_{\alpha,ng-1,\nu}}$, where ng is the number of groups (marginal means), and ν is the error degrees of freedom. This procedure provides a simultaneous confidence level for comparisons of all linear combinations of the means, and it is conservative for comparisons of simple differences of pairs.
`'lsd'`	Least significant difference. This option uses plain t-tests. The critical value is $t_{\alpha/2,\nu}$, where ν is the error degrees of freedom. It provides no protection against the multiple comparison problem.

Example: `'ComparisonType','dunn-sidak'`

Output Arguments

`tbl` — Results of multiple comparison

table

Results of multiple comparisons of estimated marginal means, returned as a table. `tbl` has the following columns.

Column Name	Description
`Difference`	Estimated difference between the corresponding two marginal means
`StdErr`	Standard error of the estimated difference between the corresponding two marginal means
`pValue`	p-value for a test that the difference between the corresponding two marginal means is 0
`Lower`	Lower limit of simultaneous 95% confidence intervals for the true difference

Column Name	Description
`Upper`	Upper limit of simultaneous 95% confidence intervals for the true difference

Examples

Multiple Comparison of Estimated Marginal Means

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform a multiple comparison of the estimated marginal means of species.

```
tbl = multcompare(rm,'species')
```

```
tbl=6×7 table
      species_1         species_2       Difference     StdErr       pValue       Lower       Upp
    ______________    ______________    __________    ________    __________    ________    ____

    {'setosa'    }    {'versicolor'}     -1.0375      0.060539    9.5606e-10     -1.1794    -0.8
    {'setosa'    }    {'virginica' }     -1.7495      0.060539    9.5606e-10     -1.8914    -1.6
    {'versicolor'}    {'setosa'    }      1.0375      0.060539    9.5606e-10      0.89562     1.1
    {'versicolor'}    {'virginica' }      -0.712      0.060539    9.5606e-10     -0.85388    -0.5
    {'virginica' }    {'setosa'    }      1.7495      0.060539    9.5606e-10       1.6076     1.8
    {'virginica' }    {'versicolor'}       0.712      0.060539    9.5606e-10      0.57012     0.8
```

The small *p*-values (in the `pValue` field) indicate that the estimated marginal means for the three species significantly differ from each other.

Perform Multiple Comparisons with Specified Options

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
R = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Perform a multiple comparison of the estimated marginal means based on the variable `Group`.

```
T = multcompare(R,'Group')
```

```
T=6×7 table
    Group_1    Group_2    Difference    StdErr     pValue       Lower      Upper
    _______    _______    __________    ______    _________    _______    _______

       A          B          4.9875      5.6271      0.65436    -9.1482     19.123
       A          C           23.094      5.9261    0.0021493     8.2074     37.981
       B          A          -4.9875      5.6271      0.65436    -19.123     9.1482
       B          C           18.107      5.8223     0.013588     3.4805     32.732
       C          A          -23.094      5.9261    0.0021493    -37.981    -8.2074
       C          B          -18.107      5.8223     0.013588    -32.732    -3.4805
```

The small *p*-value of 0.0021493 indicates that there is significant difference between the marginal means of groups A and C. The *p*-value of 0.65436 indicates that the difference between the marginal means for groups A and B is not significantly different from 0.

`multcompare` uses the Tukey-Kramer test statistic by default. Change the comparison type to the Scheffe procedure.

```
T = multcompare(R,'Group','ComparisonType','Scheffe')
```

```
T=6×7 table
    Group_1    Group_2    Difference    StdErr     pValue       Lower      Upper
    _______    _______    __________    ______    _________    _______    _______

       A          B          4.9875      5.6271      0.67981    -9.7795     19.755
       A          C           23.094      5.9261    0.0031072     7.5426     38.646
       B          A          -4.9875      5.6271      0.67981    -19.755     9.7795
       B          C           18.107      5.8223     0.018169     2.8273     33.386
       C          A          -23.094      5.9261    0.0031072    -38.646    -7.5426
       C          B          -18.107      5.8223     0.018169    -33.386    -2.8273
```

The Scheffe test produces larger *p*-values, but similar conclusions.

Perform multiple comparisons of estimated marginal means based on the variable `Group` for each gender separately.

```
T = multcompare(R,'Group','By','Gender')
```

```
T=12×8 table
    Gender    Group_1    Group_2    Difference    StdErr     pValue       Lower        Upper
    ______    _______    _______    __________    ______    _________    _________    _________
```

Female	A	B	4.1883	8.0177	0.86128	-15.953	24.329
Female	A	C	24.565	8.2083	0.017697	3.9449	45.184
Female	B	A	-4.1883	8.0177	0.86128	-24.329	15.953
Female	B	C	20.376	8.1101	0.049957	0.0033459	40.749
Female	C	A	-24.565	8.2083	0.017697	-45.184	-3.9449
Female	C	B	-20.376	8.1101	0.049957	-40.749	-0.0033459
Male	A	B	5.7868	7.9498	0.74977	-14.183	25.757
Male	A	C	21.624	8.1829	0.038022	1.0676	42.179
Male	B	A	-5.7868	7.9498	0.74977	-25.757	14.183
Male	B	C	15.837	8.0511	0.14414	-4.3881	36.062
Male	C	A	-21.624	8.1829	0.038022	-42.179	-1.0676
Male	C	B	-15.837	8.0511	0.14414	-36.062	4.3881

The results indicate that the difference between marginal means for groups A and B is not significant from 0 for either gender (corresponding *p*-values are 0.86128 for females and 0.74977 for males). The difference between marginal means for groups A and C is significant for both genders (corresponding *p*-values are 0.017697 for females and 0.038022 for males). While the difference between marginal means for groups B and C is significantly different from 0 for females (*p*-value is 0.049957), it is not significantly different from 0 for males (*p*-value is 0.14414).

References

[1] G. A. Milliken, and Johnson, D. E. *Analysis of Messy Data. Volume I: Designed Experiments*. New York, NY: Chapman & Hall, 1992.

See Also

`fitrm` | `margmean` | `plotprofile`

multivarichart

Multivari chart for grouped data

Syntax

```
multivarichart(y,GROUP)
multivarichart(Y)
multivarichart(...,param1,val1,param2,val2,...)
[charthandle,AXESH] = multivarichart(...)
```

Description

`multivarichart(y,GROUP)` displays the multivari chart for the vector `y` grouped by entries in `GROUP` that can be a cell array or a matrix. If `GROUP` is a cell array, then each cell in `GROUP` must contain a grouping variable that is a categorical vector, numeric vector, character matrix, string array, or single-column cell array of character vectors. If `GROUP` is a numeric matrix, then its columns represent different grouping variables. Each grouping variable must have the same number of elements as `y`. The number of grouping variables must be 2, 3, or 4.

Each subplot of the plot matrix contains a multivari chart for the first and second grouping variables. The *x*-axis in each subplot indicates values of the first grouping variable. The legend at the bottom of the figure window indicates values of the second grouping variable. The subplot at position (*i*,*j*) is the multivari chart for the subset of `y` at the *i*th level of the third grouping variable and the *j*th level of the fourth grouping variable. If the third or fourth grouping variable is absent, it is considered to have only one level.

`multivarichart(Y)` displays the multivari chart for a matrix `Y`. The data in different columns represent changes in one factor. The data in different rows represent changes in another factor.

`multivarichart(...,param1,val1,param2,val2,...)` specifies one or more of the following name/value pairs:

- `'varnames'` — Grouping variable names in a character matrix, a string array, or a cell array of character vectors, one per grouping variable. Default names are `'X1'`, `'X2'`, ...
- `'plotorder'` — `'sorted'` or a vector containing a permutation of the integers from 1 to the number of grouping variables.

  If `'plotorder'` is `'sorted'`, the grouping variables are rearranged in descending order according to the number of levels in each variable.

  If `'plotorder'` is a vector, it indicates the order in which each grouping variable should be plotted. For example, `[2,3,1,4]` indicates that the second grouping variable should be used as the *x*-axis of each subplot, the third grouping variable should be used as the legend, the first grouping variable should be used as the columns of the plot, and the fourth grouping variable should be used as the rows of the plot.

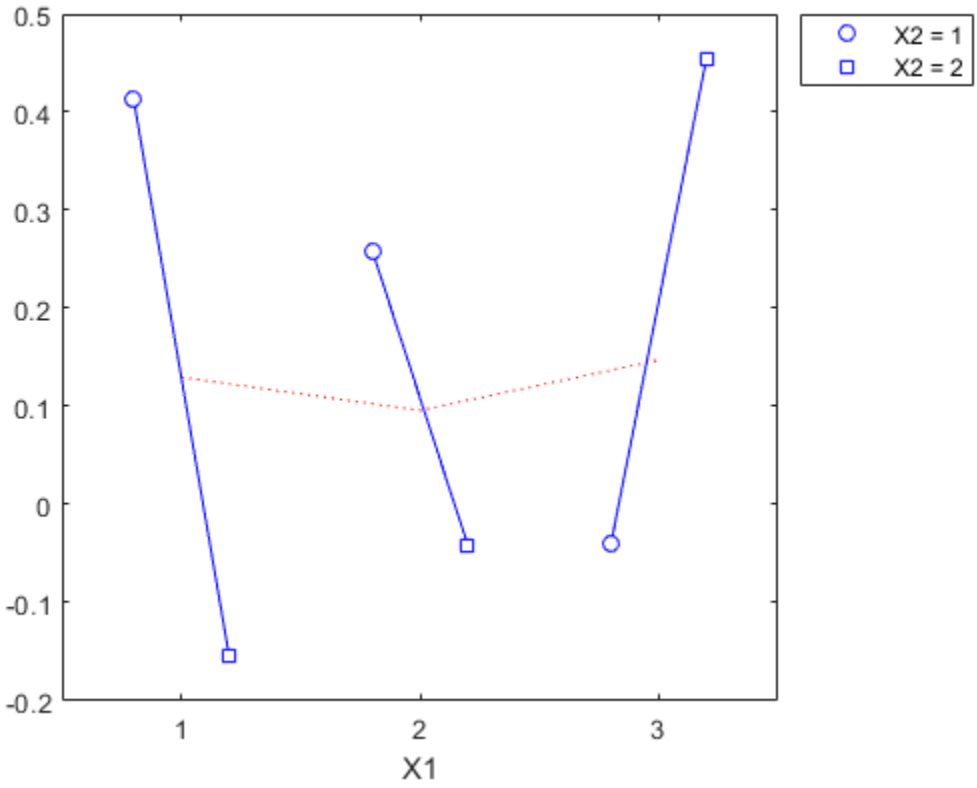
`[charthandle,AXESH] = multivarichart(...)` returns a handle `charthandle` to the figure window and a matrix `AXESH` of handles to the subplot axes.

Examples

Multivari Chart for Grouped Data

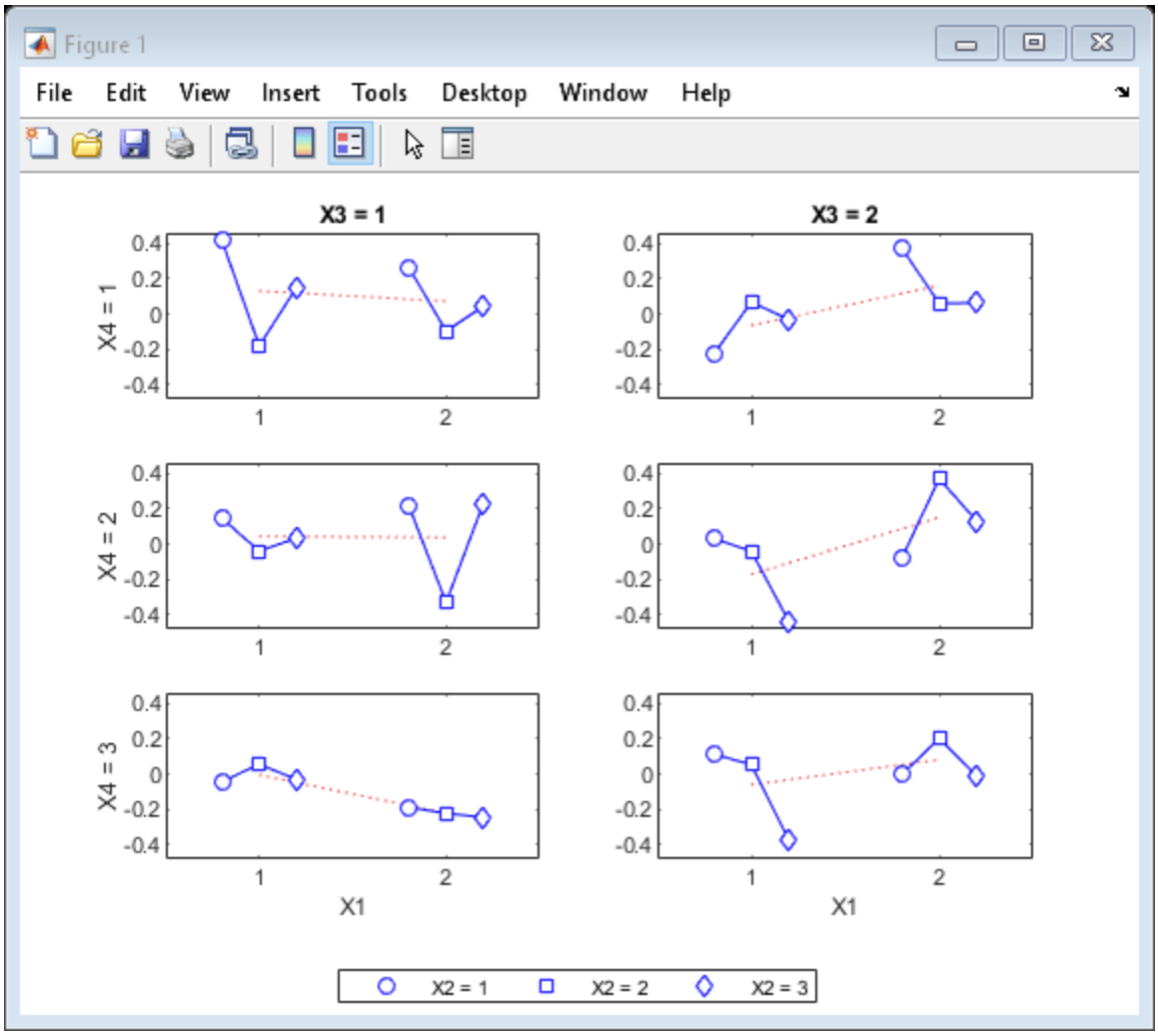
Display a multivari chart for data with two grouping variables.

```
rng default; % For reproducibility
y = randn(100,1); % Randomly generate response
group = [ceil(3*rand(100,1)) ceil(2*rand(100,1))];
multivarichart(y,group)
```



Display a multivari chart for data with four grouping variables.

```
y = randn(1000,1); % Randomly generate response
group = {ceil(2*rand(1000,1)),ceil(3*rand(1000,1)), ...
         ceil(2*rand(1000,1)),ceil(3*rand(1000,1))};
multivarichart(y,group)
```



See Also

`interactionplot` | `maineffectsplot`

Topics

“Grouping Variables” on page 2-45

Introduced in R2006b

mvksdensity

Kernel smoothing function estimate for multivariate data

Syntax

```
f = mvksdensity(x,pts,'Bandwidth',bw)
f = mvksdensity(x,pts,'Bandwidth',bw,Name,Value)
```

Description

`f = mvksdensity(x,pts,'Bandwidth',bw)` computes a probability density estimate of the sample data in the *n*-by-*d* matrix `x`, evaluated at the points in `pts` using the required name-value pair argument value `bw` for the bandwidth value. The estimation is based on a product Gaussian kernel function.

For univariate or bivariate data, use `ksdensity` instead.

`f = mvksdensity(x,pts,'Bandwidth',bw,Name,Value)` returns any of the previous output arguments, using additional options specified by one or more `Name,Value` pair arguments. For example, you can define the function type that `mvksdensity` evaluates, such as probability density, cumulative probability, or survivor function. You can also assign weights to the input values.

Examples

Estimate Multivariate Kernel Density

Load the Hald cement data.

```
load hald
```

The data measures the heat of hardening for 13 different cement compositions. The predictor matrix `ingredients` contains the percent composition for each of four cement ingredients. The response matrix `heat` contains the heat of hardening (in cal\g) after 180 days.

Estimate the kernel density for the first three observations in `ingredients`.

```
xi = ingredients(1:3,:);
f = mvksdensity(ingredients,xi,'Bandwidth',0.8);
```

Estimate Multivariate Kernel Density Using Grids

Load the Hald cement data.

```
load hald
```

The data measures the heat of hardening for 13 different cement compositions. The predictor matrix `ingredients` contains the percent composition for each of four cement ingredients. The response matrix `heat` contains the heat of hardening (in cal/g) after 180 days.

Create a array of points at which to estimate the density. First, define the range and spacing for each variable, using a similar number of points in each dimension.

```
gridx1 = 0:2:22;
gridx2 = 20:5:80;
gridx3 = 0:2:24;
gridx4 = 5:5:65;
```

Next, use `ndgrid` to generate a full grid of points using the defined range and spacing.

```
[x1,x2,x3,x4] = ndgrid(gridx1,gridx2,gridx3,gridx4);
```

Finally, transform and concatenate to create an array that contains the points at which to estimate the density. This array has one column for each variable.

```
x1 = x1(:,:)';
x2 = x2(:,:)';
x3 = x3(:,:)';
x4 = x4(:,:)';
xi = [x1(:) x2(:) x3(:) x4(:)];
```

Estimate the density.

```
f = mvksdensity(ingredients,xi,...
    'Bandwidth',[4.0579 10.7345 4.4185 11.5466],...
    'Kernel','normpdf');
```

View the size of `xi` and `f` to confirm that `mvksdensity` calculates the density at each point in `xi`.

```
size_xi = size(xi)
```

```
size_xi = 1×2

       26364           4
```

```
size_f = size(f)
```

```
size_f = 1×2

       26364           1
```

Input Arguments

x — Sample data

numeric matrix

Sample data for which `mvksdensity` returns the probability density estimate, specified as an *n*-by-*d* matrix of numeric values. *n* is the number of data points (rows) in `x`, and *d* is the number of dimensions (columns).

Data Types: `single` | `double`

pts — Points at which to evaluate f

matrix

Points at which to evaluate the probability density estimate `f`, specified as a matrix with the same number of columns as `x`. The returned estimate `f` and `pts` have the same number of rows.

Data Types: `single` | `double`

`bw` — Value for the bandwidth of the kernel smoothing window

scalar value | *d*-element vector

Value for the bandwidth of the kernel-smoothing window, specified as a scalar value or *d*-element vector. *d* is the number of dimensions (columns) in the sample data `x`. If `bw` is a scalar value, it applies to all dimensions.

If you specify `'BoundaryCorrection'` as `'log'`(default) and `'Support'` as either `'positive'` or a two-row matrix, `mvksdensity` converts bounded data to be unbounded by using log transformation. The value of `bw` is on the scale of the transformed values.

Silverman's rule of thumb for the bandwidth is

$$b_i = \sigma_i \left\{ \frac{4}{(d+2)n} \right\}^{1/(d+4)}, \quad i = 1, 2, \ldots, d,$$

where *d* is the number of dimensions, *n* is the number of observations, and σ_i is the standard deviation of the i^{th} variate [4].

Example: `'Bandwidth',0.8`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Kernel','triangle','Function,'cdf'` specifies that `mvksdensity` estimates the cdf of the sample data using the triangle kernel function.

`BoundaryCorrection` — Boundary correction method

'log' (default) | 'reflection'

Boundary correction method, specified as the comma-separated pair consisting of `'BoundaryCorrection'` and either `'log'` or `'reflection'`.

Value	Description
`'log'`	`mvksdensity` converts bounded data to be unbounded by using one of the following transformations. Then, it transforms back to the original bounded scale after density estimation. • If you specify `'Support','positive'`, then `mvksdensity` applies $\log(x_j)$ for each dimension, where x_j is the `j`th column of the input argument `x`. • If you specify `'Support'` as a two-row matrix consisting of the lower and upper limits for each dimension, then `mvksdensity` applies $\log((x_j-L_j)/(U_j-x_j))$ for each dimension, where L_j and U_j are the lower and upper limits of the `j`th dimension, respectively. The value of `bw` is on the scale of the transformed values.
`'reflection'`	`mvksdensity` augments bounded data by adding reflected data near the boundaries, then it returns estimates corresponding to the original support. For details, see “Reflection Method” on page 33-4050.

`mvksdensity` applies boundary correction only when you specify `'Support'` as a value other than `'unbounded'`.

Example: `'BoundaryCorrection','reflection'`

`Function` — Function to estimate
`'pdf'` (default) | `'cdf'` | `'survivor'`

Function to estimate, specified as the comma-separated pair consisting of `'Function'` and one of the following.

Value	Description
`'pdf'`	Probability density function
`'cdf'`	Cumulative distribution function
`'survivor'`	Survivor function

Example: `'Function','cdf'`

`Kernel` — Type of kernel smoother
`'normal'` (default) | `'box'` | `'triangle'` | `'epanechnikov'` | function handle | character vector | string scalar

Type of kernel smoother, specified as the comma-separated pair consisting of `'Kernel'` and one of the following.

Value	Description
`'normal'`	Normal (Gaussian) kernel
`'box'`	Box kernel
`'triangle'`	Triangular kernel
`'epanechnikov'`	Epanechnikov kernel

You can also specify a kernel function that is a custom or built-in function. Specify the function as a function handle (for example, `@myfunction` or `@normpdf`) or as a character vector or string scalar

(for example, `'myfunction'` or `'normpdf'`). The software calls the specified function with one argument that is an array of distances between data values and locations where the density is evaluated, normalized by the bandwidth in that dimension. The function must return an array of the same size containing the corresponding values of the kernel function.

`mvksdensity` applies the same kernel to each dimension.

Example: `'Kernel','box'`

`Support` — Support for the density
`'unbounded'` (default) | `'positive'` | 2-by-*d* matrix

Support for the density, specified as the comma-separated pair consisting of `'support'` and one of the following.

Value	Description
`'unbounded'`	Allow the density to extend over the whole real line
`'positive'`	Restrict the density to positive values
2-by-*d* matrix	Specify the finite lower and upper bounds for the support of the density. The first row contains the lower limits and the second row contains the upper limits. Each column contains the limits for one dimension of `x`.

`'Support'` can also be a combination of positive, unbounded, and bounded variables specified as `[0 -Inf L; Inf Inf U]`.

Example: `'Support','positive'`

Data Types: `single` | `double` | `char` | `string`

`Weights` — Weights for sample data
vector

Weights for sample data, specified as the comma-separated pair consisting of `'Weights'` and a vector of length `size(x,1)`, where `x` is the sample data.

Example: `'Weights',xw`

Data Types: `single` | `double`

Output Arguments

`f` — Estimated function values
vector

Estimated function values, returned as a vector. `f` and `pts` have the same number of rows.

More About

Multivariate Kernel Distribution

A multivariate kernel distribution is a nonparametric representation of the probability density function (pdf) of a random vector. You can use a kernel distribution when a parametric distribution

cannot properly describe the data, or when you want to avoid making assumptions about the distribution of the data. A multivariate kernel distribution is defined by a smoothing function and a bandwidth matrix, which control the smoothness of the resulting density curve.

The multivariate kernel density estimator is the estimated pdf of a random vector. Let $x = (x_1, x_2, \ldots, x_d)'$ be a d-dimensional random vector with a density function f and let $y_i = (y_{i1}, y_{i2}, \ldots, y_{id})'$ be a random sample drawn from f for $i = 1, 2, \ldots, n$, where n is the number of random samples. For any real vectors of x, the multivariate kernel density estimator is given by

$$\widehat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - y_i),$$

where $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$, $K(\cdot)$ is the kernel smoothing function, and H is the d-by-d bandwidth matrix.

`mvksdensity` uses a diagonal bandwidth matrix and a product kernel. That is, $H^{1/2}$ is a square diagonal matrix with the elements of vector $(h_1, h_2, \ldots, h_d)$ on the main diagonal. $K(x)$ takes the product form $K(x) = k(x_1)k(x_2)\cdots k(x_d)$, where $k(\cdot)$ is a one-dimensional kernel smoothing function. Then, the multivariate kernel density estimator becomes

$$\widehat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - y_i) = \frac{1}{nh_1h_2\cdots h_d}\sum_{i=1}^{n} K\left(\frac{x_1 - y_{i1}}{h_1}, \frac{x_2 - y_{i2}}{h_2}, \cdots, \frac{x_d - y_{id}}{h_d}\right)$$
$$= \frac{1}{nh_1h_2\cdots h_d}\sum_{i=1}^{n}\prod_{j=1}^{d} k\left(\frac{x_j - y_{ij}}{h_j}\right).$$

The kernel estimator for the cumulative distribution function (cdf), for any real vectors of x, is given by

$$\widehat{F}_H(x) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\cdots\int_{-\infty}^{x_d}\widehat{f}_H(t)dt_d\cdots dt_2dt_1 = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} G\left(\frac{x_j - y_{ij}}{h_j}\right),$$

where $G(x_j) = \int_{-\infty}^{x_j} k(t_j)dt_j$.

Reflection Method

The reflection method is a boundary correction method that accurately finds kernel density estimators when a random variable has bounded support. If you specify `'BoundaryCorrection','reflection'`, `mvksdensity` uses the reflection method.

If you additionally specify `'Support'` as a two-row matrix consisting of the lower and upper limits for each dimension, then `mvksdensity` finds the kernel estimator as follows.

- If `'Function'` is `'pdf'`, then the kernel density estimator is

  $$\widehat{f}_H(x) = \frac{1}{nh_1h_2\cdots h_d}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[k\left(\frac{x_j - y_{ij}^{-}}{h_j}\right) + k\left(\frac{x_j - y_{ij}}{h_j}\right) + k\left(\frac{x_j - y_{ij}^{+}}{h_j}\right)\right] \text{ for } L_j \le x_j \le U_j,$$

  where $y_{ij}^{-} = 2L_j - y_{ij}$, $y_{ij}^{+} = 2U_j - y_{ij}$, and y_{ij} is the `j`th element of the `i`th sample data corresponding to `x(i,j)` of the input argument `x`. L_j and U_j are the lower and upper limits of the `j`th dimension, respectively.

- If `'Function'` is `'cdf'`, then the kernel estimator for cdf is

$$\widehat{F}_H(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d}\left[G\left(\frac{x_j - y_{ij}^{-}}{h_j}\right) + G\left(\frac{x_j - y_{ij}}{h_j}\right) + G\left(\frac{x_j - y_{ij}^{+}}{h_j}\right) - G\left(\frac{L_j - y_{ij}^{-}}{h_j}\right) - G\left(\frac{L_j - y_{ij}}{h_j}\right) - G\left(\frac{L_j - y_{ij}^{+}}{h_j}\right)\right]$$

  for $L_j \le x_j \le U_j$.
- To obtain a kernel estimator for a survivor function (when `'Function'` is `'survivor'`), `mvksdensity` uses both $\widehat{f}_H(x)$ and $\widehat{F}_H(x)$.

If you additionally specify `'Support'` as `'positive'` or a matrix including `[0 inf]`, then `mvksdensity` finds the kernel density estimator by replacing $[L_j\ U_j]$ with `[0 inf]` in the above equations.

References

[1] Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press Inc., 1997.

[2] Hill, P. D. “Kernel estimation of a distribution function.” *Communications in Statistics – Theory and Methods*. Vol. 14, Issue 3, 1985, pp. 605-620.

[3] Jones, M. C. “Simple boundary correction for kernel density estimation.” *Statistics and Computing*. Vol. 3, Issue 3, 1993, pp. 135-146.

[4] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[5] Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Names in name-value pair arguments, including `'Bandwidth'`, must be compile-time constants.
- Values in the following name-value pair arguments must also be compile-time constants: `'BoundaryCorrection'`, `'Function'`, and `'Kernel'`. For example, to use the `'Function','cdf'` name-value pair argument in the generated code, include `{coder.Constant('Function'),coder.Constant('cdf')}` in the `-args` value of `codegen`.
- The value of the `'Kernel'` name-value pair argument cannot be a custom function handle. To specify a custom kernel function, use a character vector or string scalar.
- For the value of the `'Support'` name-value pair argument, the compile-time data type must match the runtime data type.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`ksdensity`

Topics

“Working with Probability Distributions” on page 5-3
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Supported Distributions” on page 5-14

Introduced in R2016a

mvncdf

Multivariate normal cumulative distribution function

Syntax

```
p = mvncdf(X)
p = mvncdf(X,mu,Sigma)

p = mvncdf(xl,xu,mu,Sigma)

p = mvncdf( ___ ,options)

[p,err] = mvncdf( ___ )
```

Description

`p = mvncdf(X)` returns the cumulative distribution function (cdf) of the multivariate normal distribution with zero mean and identity covariance matrix, evaluated at each row of `X`. For more information, see “Multivariate Normal Distribution” on page 33-4059.

`p = mvncdf(X,mu,Sigma)` returns the cdf of the multivariate normal distribution with mean `mu` and covariance `Sigma`, evaluated at each row of `X`.

Specify `[]` for `mu` to use its default value of zero when you want to specify only `Sigma`.

`p = mvncdf(xl,xu,mu,Sigma)` returns the multivariate normal cdf evaluated over the multidimensional rectangle with lower and upper limits defined by `xl` and `xu`, respectively.

`p = mvncdf( ___ ,options)` specifies control parameters for the numerical integration used to compute `p`, using any of the input argument combinations in the previous syntaxes. Create the `options` argument using the `statset` function with any combination of the parameters `'TolFun'`, `'MaxFunEvals'`, and `'Display'`.

`[p,err] = mvncdf( ___ )` additionally returns an estimate of the error in `p`. For more information, see “Algorithms” on page 33-4059.

Examples

Standard Multivariate Normal Distribution cdf

Evaluate the cdf of a standard four-dimensional multivariate normal distribution at points with increasing coordinates in every dimension.

Create a matrix `X` of five four-dimensional points with increasing coordinates.

```
firstDim = (-2:2)';
X = repmat(firstDim,1,4)
```

```
X = 5×4
```

```
    -2    -2    -2    -2
    -1    -1    -1    -1
     0     0     0     0
     1     1     1     1
     2     2     2     2
```

Evaluate the cdf at the points in `X`.

```
p = mvncdf(X)
```

```
p = 5×1

    0.0000
    0.0006
    0.0625
    0.5011
    0.9121
```

The cdf values increase because the coordinates of the points are increasing in every dimension.

Bivariate Normal Distribution cdf

Compute and plot the cdf of a bivariate normal distribution.

Define the mean vector `mu` and the covariance matrix `Sigma`.

```
mu = [1 -1];
Sigma = [.9 .4; .4 .3];
```

Create a grid of 625 evenly spaced points in two-dimensional space.

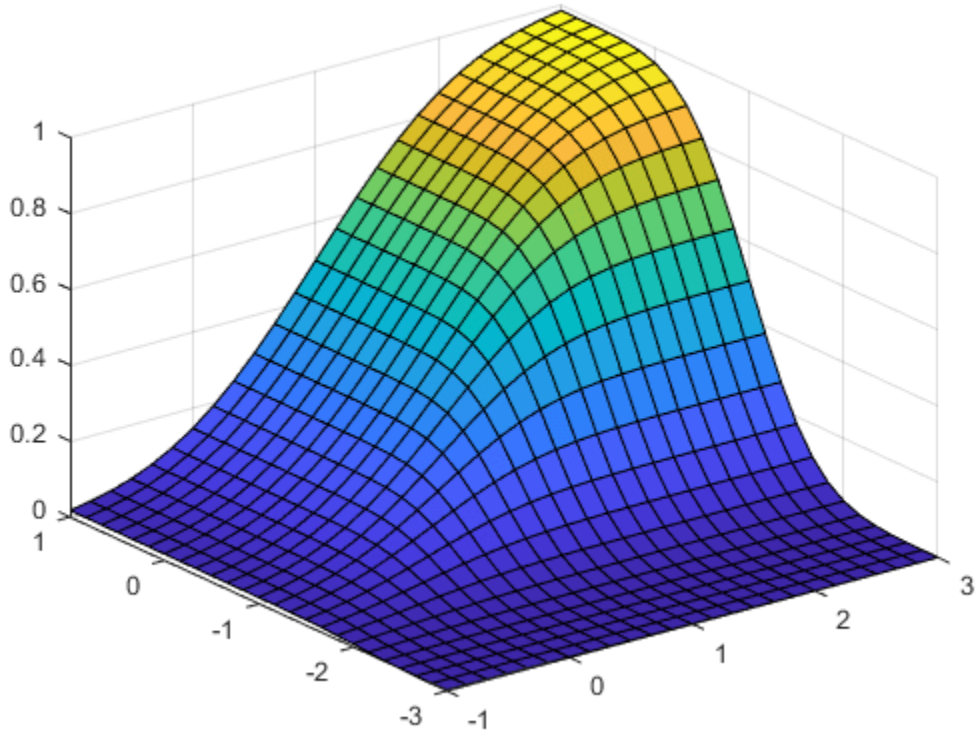
```
[X1,X2] = meshgrid(linspace(-1,3,25)',linspace(-3,1,25)');
X = [X1(:) X2(:)];
```

Evaluate the cdf of the normal distribution at the grid points.

```
p = mvncdf(X,mu,Sigma);
```

Plot the cdf values.

```
Z = reshape(p,25,25);
surf(X1,X2,Z)
```



Probability over Rectangular Region

Compute the probability over the unit square of a bivariate normal distribution, and create a contour plot of the results.

Define the bivariate normal distribution parameters `mu` and `Sigma`.

```
mu = [0 0];
Sigma = [0.25 0.3; 0.3 1];
```

Compute the probability over the unit square.

```
p = mvncdf([0 0],[1 1],mu,Sigma)
```

```
p = 0.2097
```

To visualize the result, first create a grid of evenly spaced points in two-dimensional space.

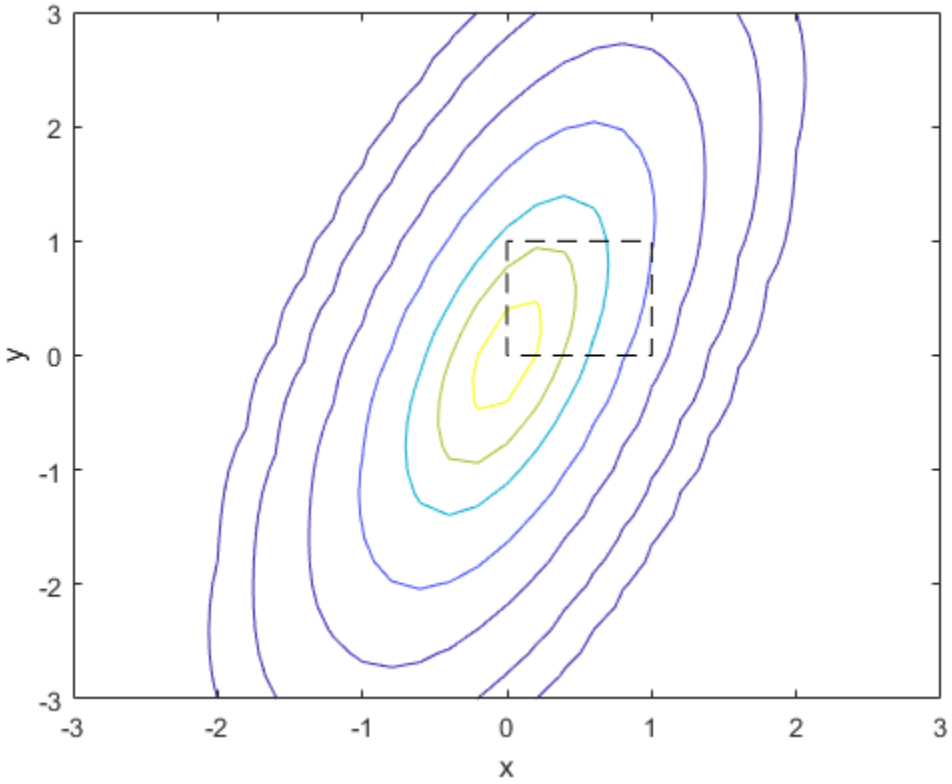
```
x1 = -3:.2:3;
x2 = -3:.2:3;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
```

Then, evaluate the pdf of the normal distribution at the grid points.

```
y = mvnpdf(X,mu,Sigma);
y = reshape(y,length(x2),length(x1));
```

Finally, create a contour plot of the multivariate normal distribution that includes the unit square.

```
contour(x1,x2,y,[0.0001 0.001 0.01 0.05 0.15 0.25 0.35])
xlabel('x')
ylabel('y')
line([0 0 1 1 0],[1 0 0 1 1],'Linestyle','--','Color','k')
```



Computing a multivariate cumulative probability requires significantly more work than computing a univariate probability. By default, the `mvncdf` function computes values to less than full machine precision, and returns an estimate of the error as an optional second output. View the error estimate in this case.

```
[p,err] = mvncdf([0 0],[1 1],mu,Sigma)
```

```
p = 0.2097
```

```
err = 1.0000e-08
```

Error Estimates in cdf Calculation

Evaluate the cdf of a multivariate normal distribution at random points, and display the error estimates associated with the cdf calculation.

Generate four random points from a five-dimensional multivariate normal distribution with mean vector `mu` and covariance matrix `Sigma`.

```
mu = [0.5 -0.3 0.2 0.1 -0.4];
Sigma = 0.5*eye(5);
rng('default')  % For reproducibility
X = mvnrnd(mu,Sigma,4);
```

Find the cdf values `p` at the points in `X` and the associated error estimates `err`. Display a summary of the numerical calculations.

```
[p,err] = mvncdf(X,mu,Sigma,statset('Display','final'))
```

```
Successfully satisfied error tolerance of 0.0001 in 8650 function evaluations.
Successfully satisfied error tolerance of 0.0001 in 8650 function evaluations.
Successfully satisfied error tolerance of 0.0001 in 8650 function evaluations.
Successfully satisfied error tolerance of 0.0001 in 8650 function evaluations.
```

```
p = 4×1

    0.1520
    0.0407
    0.0002
    0.1970


err = 4×1
10^-16 ×

    0.5949
    0.1487
         0
    0.1983
```

Input Arguments

`X` — Evaluation points

numeric matrix

Evaluation points, specified as an *n*-by-*d* numeric matrix, where *n* is a positive scalar integer and *d* is the dimension of a single multivariate normal distribution. The rows of `X` correspond to observations (or points), and the columns correspond to variables (or coordinates).

Data Types: `single` | `double`

`mu` — Mean vector of multivariate normal distribution

vector of zeros (default) | numeric vector | numeric scalar

Mean vector of a multivariate normal distribution, specified as a 1-by-*d* numeric vector or a numeric scalar, where *d* is the dimension of the multivariate normal distribution. If `mu` is a scalar, then `mvncdf` replicates the scalar to match the size of `X`.

Data Types: `single` | `double`

`Sigma` — Covariance matrix of multivariate normal distribution

identity matrix (default) | symmetric, positive definite matrix | numeric vector of diagonal entries

Covariance matrix of a multivariate normal distribution, specified as a *d*-by-*d* symmetric, positive definite matrix, where *d* is the dimension of the multivariate normal distribution. If the covariance matrix is diagonal, containing variances along the diagonal and zero covariances off it, then you can also specify `Sigma` as a 1-by-*d* vector containing just the diagonal entries.

Data Types: `single` | `double`

`xl` — Rectangle lower limit
numeric vector

Rectangle lower limit, specified as a 1-by-*d* numeric vector.

Data Types: `single` | `double`

`xu` — Rectangle upper limit
numeric vector

Rectangle upper limit, specified as a 1-by-*d* numeric vector.

Data Types: `single` | `double`

`options` — Numerical integration options
structure

Numerical integration options, specified as a structure. Create the `options` argument by calling the `statset` function with any combination of the following parameters:

- `'TolFun'` — Maximum absolute error tolerance. The default value is `1e-8` when $d < 4$, and `1e-4` when $d \geq 4$.
- `'MaxFunEvals'` — Maximum number of integrand evaluations allowed when $d \geq 4$. The default value is `1e7`. The function ignores `'MaxFunEvals'` when $d < 4$.
- `'Display'` — Level of display output. The choices are `'off'` (the default), `'iter'`, and `'final'`. The function ignores `'Display'` when $d < 4$.

Example: `statset('TolFun',1e-7,'Display','final')`

Data Types: `struct`

Output Arguments

`p` — cdf values
numeric vector | numeric scalar

cdf values, returned as either an *n*-by-1 numeric vector, where *n* is the number of rows in `X`, or a numeric scalar representing the probability over the rectangular region specified by `xl` and `xu`.

`err` — Absolute error tolerance
positive numeric scalar

Absolute error tolerance, returned as a positive numeric scalar. For bivariate and trivariate distributions, the default absolute error tolerance is `1e-8`. For four or more dimensions, the default absolute error tolerance is `1e-4`. For more information, see “Algorithms” on page 33-4059.

More About

Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the univariate normal distribution to two or more variables. It has two parameters, a mean vector μ and a covariance matrix Σ, that are analogous to the mean and variance parameters of a univariate normal distribution. The diagonal elements of Σ contain the variances for each variable, and the off-diagonal elements of Σ contain the covariances between variables.

The probability density function (pdf) of the d-dimensional multivariate normal distribution is

$$y = f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp\left(-\frac{1}{2}(x\text{-}\mu)\ \Sigma^{-1}(x\text{-}\mu)'\right)$$

where x and μ are 1-by-d vectors and Σ is a d-by-d symmetric, positive definite matrix. Only `mvnrnd` allows positive semi-definite Σ matrices, which can be singular. The pdf cannot have the same form when Σ is singular.

The multivariate normal cumulative distribution function (cdf) evaluated at x is the probability that a random vector v, distributed as multivariate normal, lies within the semi-infinite rectangle with upper limits defined by x:

$$\Pr\{v(1) \le x(1), v(2) \le x(2), \ldots, v(d) \le x(d)\}\ .$$

Although the multivariate normal cdf does not have a closed form, `mvncdf` can compute cdf values numerically.

Tips

- In the one-dimensional case, `Sigma` is the variance, not the standard deviation. For example, `mvncdf(1,0,4)` is the same as `normcdf(1,0,2)`, where `4` is the variance and `2` is the standard deviation.

Algorithms

For bivariate and trivariate distributions, `mvncdf` uses adaptive quadrature on a transformation of the t density, based on methods developed by Drezner and Wesolowsky [1] [2] and by Genz [3]. For four or more dimensions, `mvncdf` uses a quasi-Monte Carlo integration algorithm based on methods developed by Genz and Bretz [4] [5].

References

[1] Drezner, Z. “Computation of the Trivariate Normal Integral.” *Mathematics of Computation*. Vol. 63, 1994, pp. 289–294.

[2] Drezner, Z., and G. O. Wesolowsky. “On the Computation of the Bivariate Normal Integral.” *Journal of Statistical Computation and Simulation*. Vol. 35, 1989, pp. 101–107.

[3] Genz, A. “Numerical Computation of Rectangular Bivariate and Trivariate Normal and t Probabilities.” *Statistics and Computing*. Vol. 14, No. 3, 2004, pp. 251–260.

[4] Genz, A., and F. Bretz. “Numerical Computation of Multivariate t Probabilities with Application to Power Calculation of Multiple Contrasts.” *Journal of Statistical Computation and Simulation*. Vol. 63, 1999, pp. 361–378.

[5] Genz, A., and F. Bretz. “Comparison of Methods for the Computation of Multivariate t Probabilities.” *Journal of Computational and Graphical Statistics*. Vol. 11, No. 4, 2002, pp. 950–971.

[6] Kotz, S., N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions: Volume 1: Models and Applications.* 2nd ed. New York: John Wiley & Sons, Inc., 2000.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`mvnpdf` | `mvnrnd`

Topics

“Multivariate Normal Distribution” on page B-98

Introduced in R2006a

mvnpdf

Multivariate normal probability density function

Syntax

```
y = mvnpdf(X)
y = mvnpdf(X,mu)
y = mvnpdf(X,mu,Sigma)
```

Description

`y = mvnpdf(X)` returns an *n*-by-1 vector `y` containing the probability density function (pdf) values for the *d*-dimensional multivariate normal distribution with zero mean and identity covariance matrix, evaluated at each row of the *n*-by-*d* matrix `X`. For more information, see “Multivariate Normal Distribution” on page 33-4067.

`y = mvnpdf(X,mu)` returns pdf values of points in `X`, where `mu` determines the mean of each associated multivariate normal distribution.

`y = mvnpdf(X,mu,Sigma)` returns pdf values of points in `X`, where `Sigma` determines the covariance of each associated multivariate normal distribution.

Specify `[]` for `mu` to use its default value of zero when you want to specify only `Sigma`.

Examples

Standard Multivariate Normal pdf

Evaluate the pdf of a standard five-dimensional normal distribution at a set of random points.

Randomly sample eight points from the standard five-dimensional normal distribution.

```
mu = zeros(1,5);
Sigma = eye(5);
rng('default')  % For reproducibility
X = mvnrnd(mu,Sigma,8)
```

```
X = 8×5

    0.5377    3.5784   -0.1241    0.4889   -1.0689
    1.8339    2.7694    1.4897    1.0347   -0.8095
   -2.2588   -1.3499    1.4090    0.7269   -2.9443
    0.8622    3.0349    1.4172   -0.3034    1.4384
    0.3188    0.7254    0.6715    0.2939    0.3252
   -1.3077   -0.0631   -1.2075   -0.7873   -0.7549
   -0.4336    0.7147    0.7172    0.8884    1.3703
    0.3426   -0.2050    1.6302   -1.1471   -1.7115
```

Evaluate the pdf of the distribution at the points in `X`.

```
y = mvnpdf(X)
```

```
y = 8×1

    0.0000
    0.0000
    0.0000
    0.0000
    0.0054
    0.0011
    0.0015
    0.0003
```

Find the point in X with the greatest pdf value.

```
[maxpdf,idx] = max(y)
```

```
maxpdf = 0.0054
```

```
idx = 5
```

```
maxPoint = X(idx,:)
```

```
maxPoint = 1×5

    0.3188    0.7254    0.6715    0.2939    0.3252
```

The fifth point in X has a greater pdf value than any of the other randomly selected points.

Multivariate Normal pdfs Evaluated at Different Points

Create six three-dimensional normal distributions, each with a distinct mean. Evaluate the pdf of each distribution at a different random point.

Specify the means mu and covariances Sigma of the distributions. Each distribution has the same covariance matrix—the identity matrix.

```
firstDim = (1:6)';
mu = repmat(firstDim,1,3)
```

```
mu = 6×3

     1     1     1
     2     2     2
     3     3     3
     4     4     4
     5     5     5
     6     6     6
```

```
Sigma = eye(3)
```

```
Sigma = 3×3
```

```
     1     0     0
     0     1     0
     0     0     1
```

Randomly sample once from each of the six distributions.

```
rng('default')  % For reproducibility
X = mvnrnd(mu,Sigma)
```

```
X = 6×3

    1.5377    0.5664    1.7254
    3.8339    2.3426    1.9369
    0.7412    6.5784    3.7147
    4.8622    6.7694    3.7950
    5.3188    3.6501    4.8759
    4.6923    9.0349    7.4897
```

Evaluate the pdfs of the distributions at the points in `X`. The pdf of the first distribution is evaluated at the point `X(1,:)`, the pdf of the second distribution is evaluated at the point `X(2,:)`, and so on.

```
y = mvnpdf(X,mu)
```

```
y = 6×1

    0.0384
    0.0111
    0.0000
    0.0009
    0.0241
    0.0001
```

Multivariate Normal pdf

Evaluate the pdf of a two-dimensional normal distribution at a set of given points.

Specify the mean `mu` and covariance `Sigma` of the distribution.

```
mu = [1 -1];
Sigma = [0.9 0.4; 0.4 0.3];
```

Randomly sample from the distribution 100 times. Specify `X` as the matrix of sampled points.

```
rng('default')  % For reproducibility
X = mvnrnd(mu,Sigma,100);
```

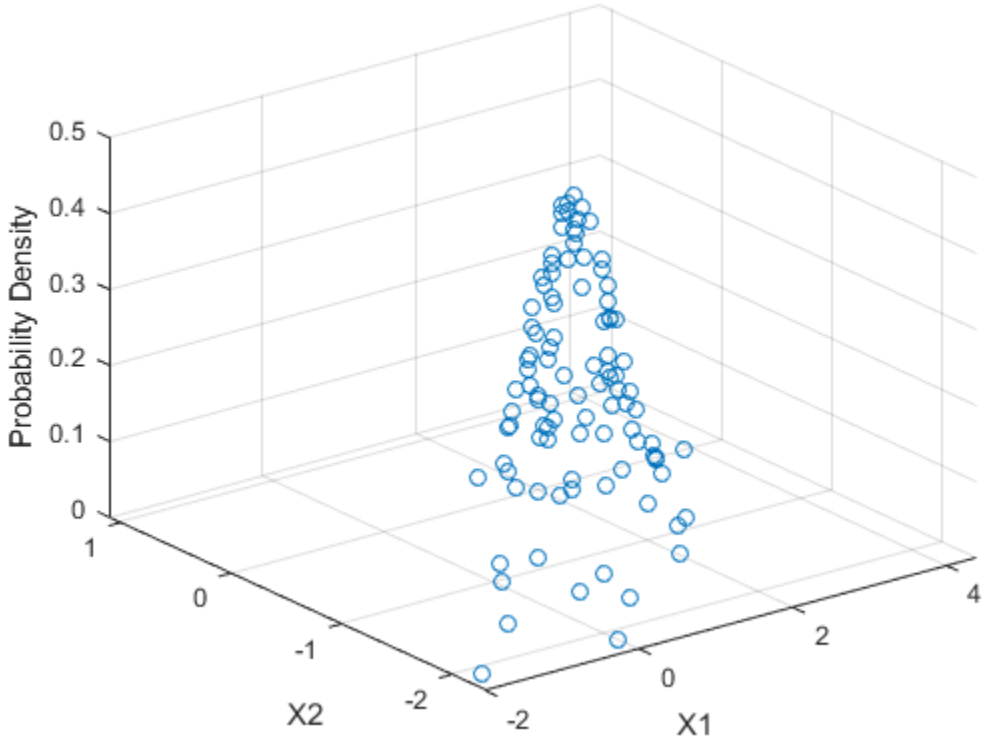
Evaluate the pdf of the distribution at the points in `X`.

```
y = mvnpdf(X,mu,Sigma);
```

Plot the probability density values.

```
scatter3(X(:,1),X(:,2),y)
xlabel('X1')
```

```
ylabel('X2')
zlabel('Probability Density')
```



Multivariate Normal pdfs Evaluated at Same Point

Create ten different five-dimensional normal distributions, and compare the values of their pdfs at a specified point.

Set the dimensions `n` and `d` equal to 10 and 5, respectively.

```
n = 10;
d = 5;
```

Specify the means `mu` and the covariances `Sigma` of the multivariate normal distributions. Let all the distributions have the same mean vector, but vary the covariance matrices.

```
mu = ones(1,d)
```

```
mu = 1×5

     1     1     1     1     1
```

```
mat = eye(d);
nMat = repmat(mat,1,1,n);
```

```
var = reshape(1:n,1,1,n);
Sigma = nMat.*var;
```

Display the first two covariance matrices in `Sigma`.

```
Sigma(:,:,1:2)
```

```
ans =
ans(:,:,1) =

     1     0     0     0     0
     0     1     0     0     0
     0     0     1     0     0
     0     0     0     1     0
     0     0     0     0     1


ans(:,:,2) =

     2     0     0     0     0
     0     2     0     0     0
     0     0     2     0     0
     0     0     0     2     0
     0     0     0     0     2
```

Set `x` to be a random point in five-dimensional space.

```
rng('default')  % For reproducibility
x = normrnd(0,1,1,5)
```

```
x = 1×5

    0.5377    1.8339   -2.2588    0.8622    0.3188
```

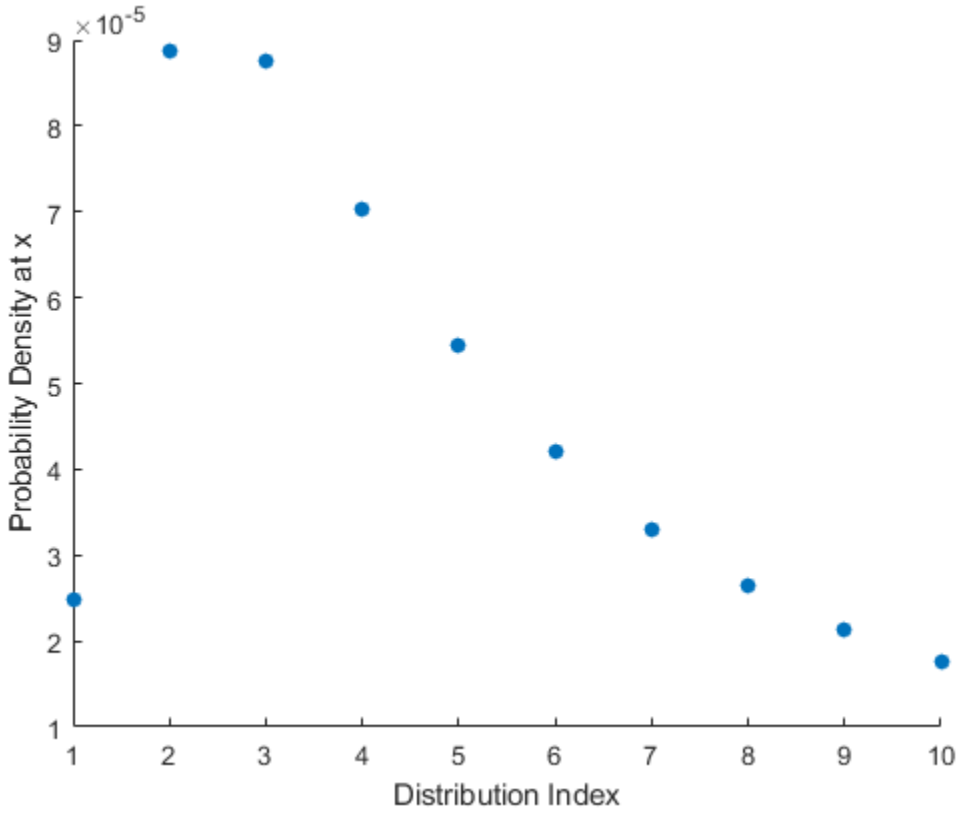
Evaluate the pdf at `x` for each of the ten distributions.

```
y = mvnpdf(x,mu,Sigma)
```

```
y = 10×1
10^-4 ×

    0.2490
    0.8867
    0.8755
    0.7035
    0.5438
    0.4211
    0.3305
    0.2635
    0.2134
    0.1753
```

Plot the results.

```
scatter(1:n,y,'filled')
xlabel('Distribution Index')
ylabel('Probability Density at x')
```



Input Arguments

X — Evaluation points

numeric vector | numeric matrix

Evaluation points, specified as a 1-by-*d* numeric vector or an *n*-by-*d* numeric matrix, where *n* is a positive scalar integer and *d* is the dimension of a single multivariate normal distribution. The rows of X correspond to observations (or points), and the columns correspond to variables (or coordinates).

If X is a vector, then `mvnpdf` replicates it to match the leading dimension of `mu` or the trailing dimension of `Sigma`.

Data Types: `single` | `double`

mu — Means of multivariate normal distributions

vector of zeros (default) | numeric vector | numeric matrix

Means of multivariate normal distributions, specified as a 1-by-*d* numeric vector or an *n*-by-*d* numeric matrix.

- If `mu` is a vector, then `mvnpdf` replicates the vector to match the trailing dimension of `Sigma`.
- If `mu` is a matrix, then each row of `mu` is the mean vector of a single multivariate normal distribution.

Data Types: `single` | `double`

`Sigma` — Covariances of multivariate normal distributions
identity matrix (default) | symmetric, positive definite matrix | numeric array

Covariances of multivariate normal distributions, specified as a *d*-by-*d* symmetric, positive definite matrix or a *d*-by-*d*-by-*n* numeric array.

- If `Sigma` is a matrix, then `mvnpdf` replicates the matrix to match the number of rows in `mu`.
- If `Sigma` is an array, then each page of `Sigma`, `Sigma(:,:,i)`, is the covariance matrix of a single multivariate normal distribution and, therefore, is a symmetric, positive definite matrix.

If the covariance matrices are diagonal, containing variances along the diagonal and zero covariances off it, then you can also specify `Sigma` as a `1`-by-*d* vector or a `1`-by-*d*-by-*n* array containing just the diagonal entries.

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
numeric vector

pdf values, returned as an *n*-by-`1` numeric vector, where *n* is one of the following:

- Number of rows in `X` if `X` is a matrix
- Number of times `X` is replicated if `X` is a vector

If `X` is a matrix, `mu` is a matrix, and `Sigma` is an array, then `mvnpdf` computes `y(i)` using `X(i,:)`, `mu(i,:)`, and `Sigma(:,:,i)`.

Data Types: `double`

More About

Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the univariate normal distribution to two or more variables. It has two parameters, a mean vector μ and a covariance matrix Σ, that are analogous to the mean and variance parameters of a univariate normal distribution. The diagonal elements of Σ contain the variances for each variable, and the off-diagonal elements of Σ contain the covariances between variables.

The probability density function (pdf) of the *d*-dimensional multivariate normal distribution is

$$y = f(x,\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}}\exp\left(-\frac{1}{2}(x\text{-}\mu)\,\Sigma^{-1}(x\text{-}\mu)'\right)$$

where x and μ are 1-by-*d* vectors and Σ is a *d*-by-*d* symmetric, positive definite matrix. Only `mvnrnd` allows positive semi-definite Σ matrices, which can be singular. The pdf cannot have the same form when Σ is singular.

The multivariate normal cumulative distribution function (cdf) evaluated at x is the probability that a random vector v, distributed as multivariate normal, lies within the semi-infinite rectangle with upper limits defined by x:

$$\Pr\{v(1) \le x(1), v(2) \le x(2), \ldots, v(d) \le x(d)\}.$$

Although the multivariate normal cdf does not have a closed form, `mvncdf` can compute cdf values numerically.

Tips

- In the one-dimensional case, `Sigma` is the variance, not the standard deviation. For example, `mvnpdf(1,0,4)` is the same as `normpdf(1,0,2)`, where `4` is the variance and `2` is the standard deviation.

References

[1] Kotz, S., N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions: Volume 1: Models and Applications.* 2nd ed. New York: John Wiley & Sons, Inc., 2000.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`mvncdf` | `mvnrnd` | `normpdf`

Topics

“Multivariate Normal Distribution” on page B-98

Introduced before R2006a

mvregress

Multivariate linear regression

Syntax

```
beta = mvregress(X,Y)
beta = mvregress(X,Y,Name,Value)
[beta,Sigma] = mvregress( ___ )
[beta,Sigma,E,CovB,logL] = mvregress( ___ )
```

Description

`beta = mvregress(X,Y)` returns the estimated coefficients for a multivariate normal regression on page 33-4080 of the *d*-dimensional responses in `Y` on the design matrices in `X`.

`beta = mvregress(X,Y,Name,Value)` returns the estimated coefficients using additional options specified by one or more name-value pair arguments. For example, you can specify the estimation algorithm, initial estimate values, or maximum number of iterations for the regression.

`[beta,Sigma] = mvregress( ___ )` also returns the estimated *d*-by-*d* variance-covariance matrix of `Y`, using any of the input arguments from the previous syntaxes.

`[beta,Sigma,E,CovB,logL] = mvregress( ___ )` also returns a matrix of residuals `E`, estimated variance-covariance matrix of the regression coefficients `CovB`, and the value of the log likelihood objective function after the last iteration `logL`.

Examples

Multivariate Regression Model for Panel Data with Different Intercepts

Fit a multivariate regression model to panel data, assuming different intercepts and common slopes.

Load the sample data.

```
load('flu')
```

The dataset array `flu` contains national CDC flu estimates, and nine separate regional estimates based on Google® query data.

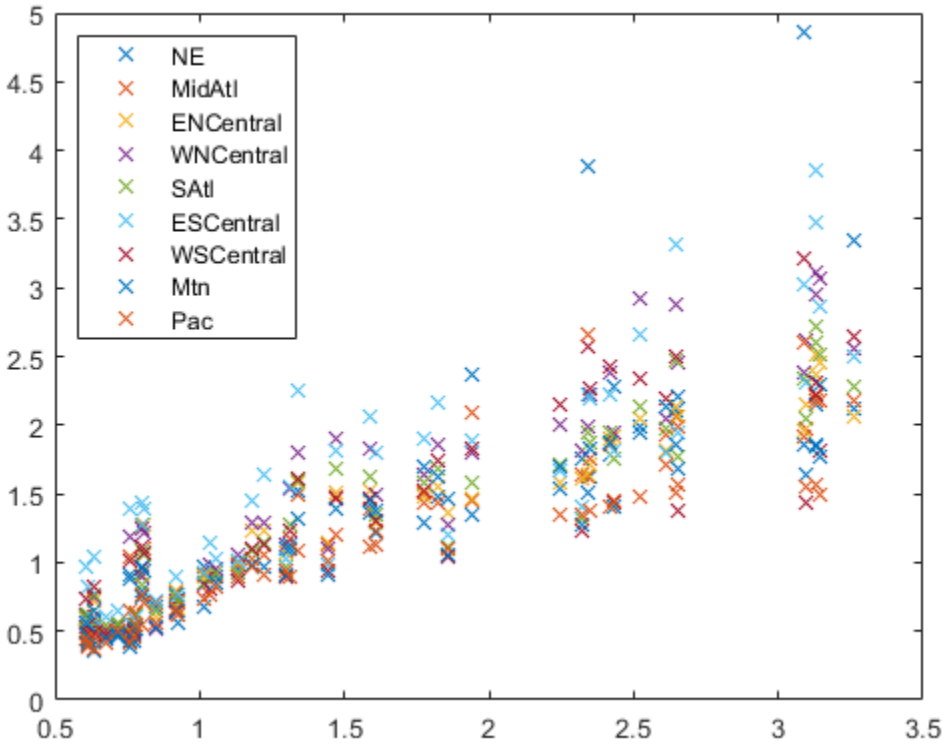
Extract the response and predictor data.

```
Y = double(flu(:,2:end-1));
[n,d] = size(Y);
x = flu.WtdILI;
```

The responses in `Y` are the nine regional flu estimates. Observations exist for every week over a one-year period, so $n = 52$. The dimension of the responses corresponds to the regions, so $d = 9$. The predictors in `x` are the weekly national flu estimates.

Plot the flu data, grouped by region.

```
figure;
regions = flu.Properties.VarNames(2:end-1);
plot(x,Y,'x')
legend(regions,'Location','NorthWest')
```



Fit the multivariate regression model $y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, d$, with between-region concurrent correlation $COV(\epsilon_{ij}, \epsilon_{ij}) = \sigma_{jj}$.

There are $K = 10$ regression coefficients to estimate: nine intercept terms and a common slope. The input argument `X` should be an n-element cell array of d -by- K design matrices.

```
X = cell(n,1);
for i = 1:n
    X{i} = [eye(d) repmat(x(i),d,1)];
end
[beta,Sigma] = mvregress(X,Y);
```

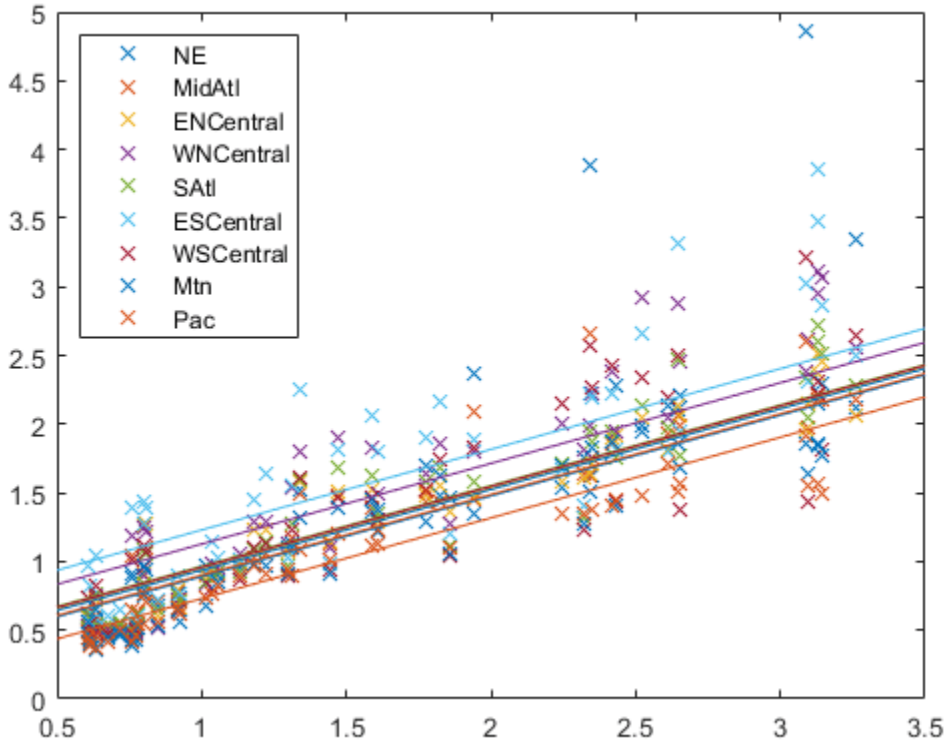
`beta` contains estimates of the K-dimensional coefficient vector $(\alpha_1, \alpha_2, \ldots, \alpha_9, \beta)'$.

`Sigma` contains estimates of the d -by- d variance-covariance matrix $(\sigma_{ij})_{d \times d}$, $i, j = 1, \ldots, d$ for the between-region concurrent correlations.

Plot the fitted regression model.

```
B = [beta(1:d)';repmat(beta(end),1,d)];
xx = linspace(.5,3.5)';
fits = [ones(size(xx)),xx]*B;
```

```
figure;
h = plot(x,Y,'x',xx,fits,'-');
for i = 1:d
    set(h(d+i),'color',get(h(i),'color'));
end
legend(regions,'Location','NorthWest');
```



The plot shows that each regression line has a different intercept but the same slope. Upon visual inspection, some regression lines appear to fit the data better than others.

Multivariate Regression for Panel Data with Different Slopes

Fit a multivariate regression model to panel data using least squares, assuming different intercepts and slopes.

Load the sample data.

```
load('flu');
```

The dataset array `flu` contains national CDC flu estimates, and nine separate regional estimates based on Google® queries.

Extract the response and predictor data.

```
Y = double(flu(:,2:end-1));
[n,d] = size(Y);
x = flu.WtdILI;
```

The responses in Y are the nine regional flu estimates. Observations exist for every week over a one-year period, so $n = 52$. The dimension of the responses corresponds to the regions, so $d = 9$. The predictors in x are the weekly national flu estimates.

Fit the multivariate regression model $y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, d$, with between-region concurrent correlation $COV(\epsilon_{ij}, \epsilon_{ij}) = \sigma_{jj}$.

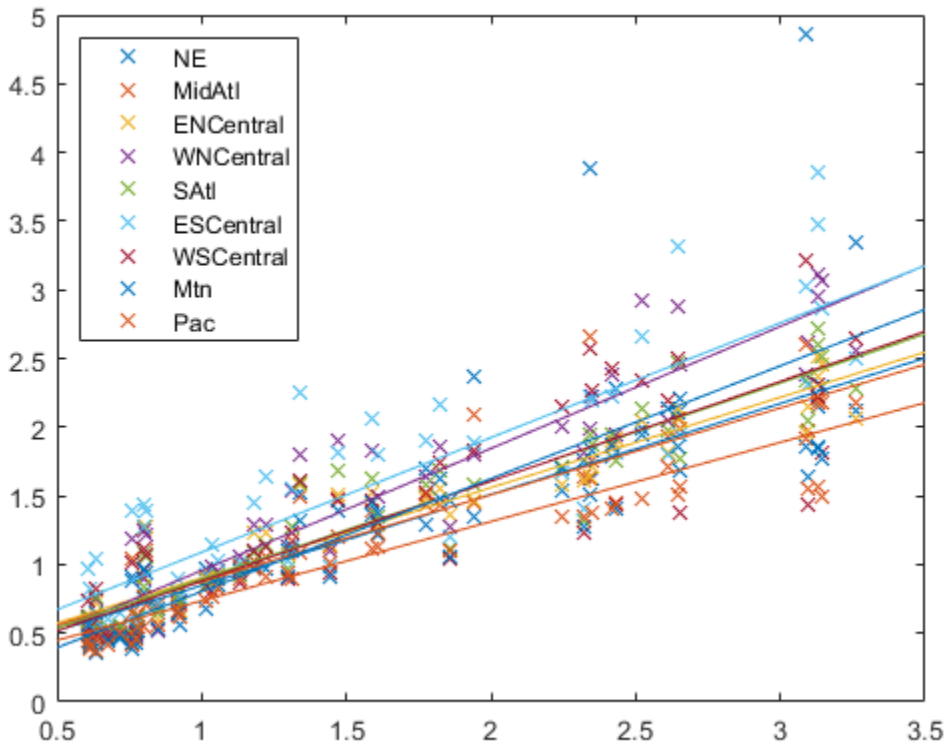
There are $K = 18$ regression coefficients to estimate: nine intercept terms, and nine slope terms. X is an n-element cell array of d -by- K design matrices.

```
X = cell(n,1);
for i = 1:n
    X{i} = [eye(d) x(i)*eye(d)];
end
[beta,Sigma] = mvregress(X,Y,'algorithm','cwls');
```

beta contains estimates of the K-dimensional coefficient vector $(\alpha_1, \alpha_2, \ldots, \alpha_9, \beta_1, \beta_2, \ldots, \beta_9)'$.

Plot the fitted regression model.

```
B = [beta(1:d)';beta(d+1:end)'];
xx = linspace(.5,3.5)';
fits = [ones(size(xx)),xx]*B;

figure;
h = plot(x,Y,'x',xx,fits,'-');
for i = 1:d
    set(h(d+i),'color',get(h(i),'color'));
end

regions = flu.Properties.VarNames(2:end-1);
legend(regions,'Location','NorthWest');
```



The plot shows that each regression line has a different intercept and slope.

Multivariate Regression With a Single Design Matrix

Fit a multivariate regression model using a single *n*-by-*P* design matrix for all response dimensions.

Load the sample data.

```
load('flu')
```

The dataset array `flu` contains national CDC flu estimates, and nine separate regional estimates based on Google® queries.

Extract the response and predictor data.

```
Y = double(flu(:,2:end-1));
[n,d] = size(Y);
x = flu.WtdILI;
```

The responses in `Y` are the nine regional flu estimates. Observations exist for every week over a one-year period, so $n = 52$. The dimension of the responses corresponds to the regions, so $d = 9$. The predictors in `x` are the weekly national flu estimates.

Create an n -by- P design matrix `X`. Add a column of ones to include a constant term in the regression.

```
X = [ones(size(x)),x];
```

Fit the multivariate regression model

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij},$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, d$, with between-region concurrent correlation

$$COV(\epsilon_{ij}, \epsilon_{ij}) = \sigma_{jj}.$$

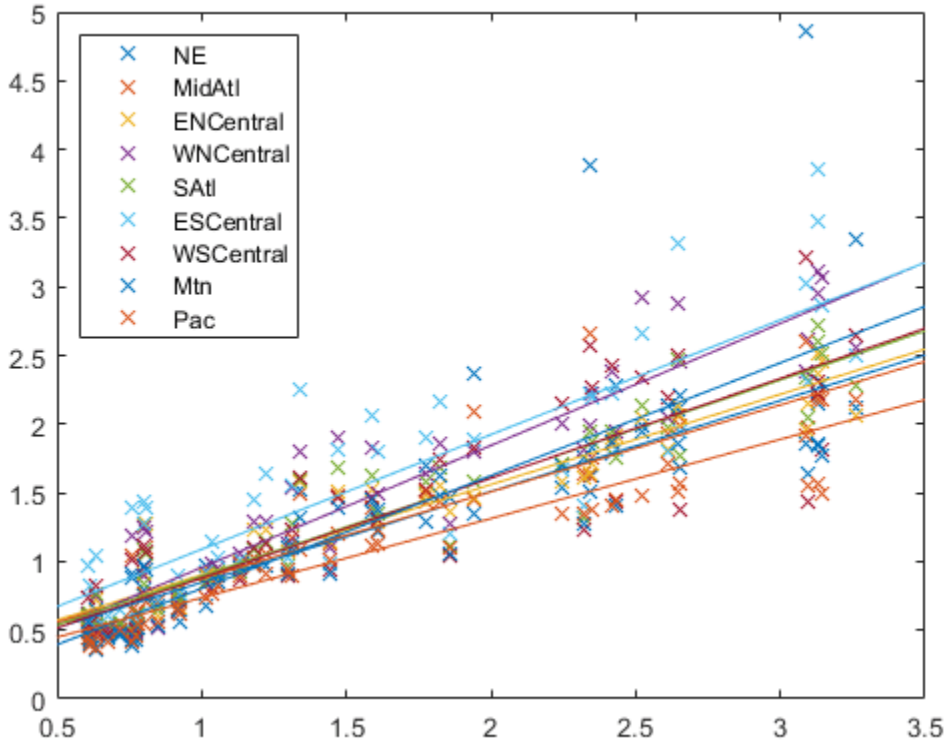
There are 18 regression coefficients to estimate: nine intercept terms, and nine slope terms.

```
[beta,Sigma,E,CovB,logL] = mvregress(X,Y);
```

`beta` contains estimates of the *P*-by-*d* coefficient matrix. `Sigma` contains estimates of the *d*-by-*d* variance-covariance matrix for the between-region concurrent correlations. `E` is a matrix of the residuals. `CovB` is the estimated variance-covariance matrix of the regression coefficients. `logL` is the value of the log likelihood objective function after the last iteration.

Plot the fitted regression model.

```
B = beta;
xx = linspace(.5,3.5)';
fits = [ones(size(xx)),xx]*B;

figure
h = plot(x,Y,'x', xx,fits,'-');
for i = 1:d
    set(h(d+i),'color',get(h(i),'color'))
end

regions = flu.Properties.VarNames(2:end-1);
legend(regions,'Location','NorthWest')
```



The plot shows that each regression line has a different intercept and slope.

Input Arguments

X — Design matrices

matrix | cell array of matrices

Design matrices for the multivariate regression, specified as a matrix or cell array of matrices. n is the number of observations in the data, K is the number of regression coefficients to estimate, p is the number of predictor variables, and d is the number of dimensions in the response variable matrix Y.

- If $d = 1$, then specify X as a single n-by-K design matrix.
- If $d > 1$ and all d dimensions have the same design matrix, then you can specify X as a single n-by-p design matrix (not in a cell array).
- If $d > 1$ and all n observations have the same design matrix, then you can specify X as a cell array containing a single d-by-K design matrix.
- If $d > 1$ and all n observations do not have the same design matrix, then specify X as a cell array of length n containing d-by-K design matrices.

To include a constant term in the regression model, each design matrix should contain a column of ones.

`mvregress` treats `NaN` values in X as missing values, and ignores rows in X with missing values.

Data Types: `single` | `double` | `cell`

Y — Response variables

matrix

Response variables, specified as an *n*-by-*d* matrix. *n* is the number of observations in the data, and *d* is the number of dimensions in the response. When *d* = 1, `mvregress` treats the values in `Y` like *n* independent response values.

`mvregress` treats `NaN` values in `Y` as missing values, and handles them according to the estimation algorithm specified using the name-value pair argument `algorithm`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'algorithm','cwls','covar0',C` specifies covariance-weighted least squares estimation using the covariance matrix `C`.

algorithm — Estimation algorithm

`'mvn'` | `'ecm'` | `'cwls'`

Estimation algorithm, specified as the comma-separated pair consisting of `'algorithm'` and one of the following.

`'mvn'`	Ordinary multivariate normal maximum likelihood estimation.
`'ecm'`	Maximum likelihood estimation via the ECM algorithm.
`'cwls'`	Covariance-weighted least squares estimation.

The default algorithm depends on the presence of missing data.

- For complete data, the default is `'mvn'`.
- If there are any missing responses (indicated by `NaN`), the default is `'ecm'`, provided the sample size is sufficient to estimate all parameters. Otherwise, the default algorithm is `'cwls'`.

Note If `algorithm` has the value `'mvn'`, then `mvregress` removes observations with missing response values before estimation.

Example: `'algorithm','ecm'`

beta0 — Initial estimates for regression coefficients

vector

Initial estimates for the regression coefficients, specified as the comma-separated pair consisting of `'beta0'` and a vector with *K* elements. The default value is a vector of 0s.

The `beta0` argument is not used if the estimation `algorithm` is `'mvn'`.

`covar0` — Initial estimate for variance-covariance matrix

matrix

Initial estimate for the variance-covariance matrix, `Sigma`, specified as the comma-separated pair consisting of `'covar0'` and a symmetric, positive definite, *d*-by-*d* matrix. The default value is the identity matrix.

If the estimation `algorithm` is `'cwls'`, then `mvregress` uses `covar0` as the weighting matrix at each iteration, without changing it.

`covtype` — Type of variance-covariance matrix

`'full'` (default) | `'diagonal'`

Type of variance-covariance matrix to estimate for `Y`, specified as the comma-separated pair consisting of `'covtype'` and one of the following.

`'full'`	Estimate all $d(d + 1)/2$ variance-covariance elements.
`'diagonal'`	Estimate only the d diagonal elements of the variance-covariance matrix.

Example: `'covtype','diagonal'`

`maxiter` — Maximum number of iterations

`100` (default) | positive integer

Maximum number of iterations for the estimation algorithm, specified as the comma-separated pair consisting of `'maxiter'` and a positive integer.

Iterations continue until estimates are within the convergence tolerances `tolbeta` and `tolobj`, or the maximum number of iterations specified by `maxiter` is reached. If both `tolbeta` and `tolobj` are 0, then `mvregress` performs `maxiter` iterations with no convergence tests.

Example: `'maxiter',50`

`outputfcn` — Function to evaluate each iteration

function handle

Function to evaluate at each iteration, specified as the comma-separated pair consisting of `'outputfcn'` and a function handle. The function must return a logical `true` or `false`. At each iteration, `mvregress` evaluates the function. If the result is `true`, iterations stop. Otherwise, iterations continue. For example, you could specify a function that plots or displays current iteration results, and returns `true` if you close the figure.

The function must accept three input arguments, in this order:

- Vector of current coefficient estimates
- Structure containing these three fields:

`Covar`	Current value of the variance-covariance matrix
`iteration`	Current iteration number
`fval`	Current value of the loglikelihood objective function

- Text that takes these three values:

`'init'`	When the function is called during initialization
`'iter'`	When the function is called after an iteration
`'done'`	When the function is called after completion

`tolbeta` — Convergence tolerance for regression coefficients
`sqrt(eps)` (default) | positive scalar value

Convergence tolerance for regression coefficients, specified as the comma-separated pair consisting of `'tolbeta'` and a positive scalar value.

Let $\mathbf{b}^t$ denote the estimate of the coefficient vector at iteration t, and τ_β be the tolerance specified by `tolbeta`. The convergence criterion for regression coefficient estimation is

$$\left\|\mathbf{b}^t - \mathbf{b}^{t-1}\right\| < \tau_\beta\sqrt{K}\left(1 + \left\|\mathbf{b}^t\right\|\right),$$

where K is the length of $\mathbf{b}^t$ and $\|\mathbf{v}\|$ is the norm of a vector $\mathbf{v}$.

Iterations continue until estimates are within the convergence tolerances `tolbeta` and `tolobj`, or the maximum number of iterations specified by `maxiter` is reached. If both `tolbeta` and `tolobj` are 0, then `mvregress` performs `maxiter` iterations with no convergence tests.

Example: `'tolbeta',1e-5`

`tolobj` — Convergence tolerance for loglikelihood objective function
`eps^(3/4)` (default) | positive scalar value

Convergence tolerance for the loglikelihood objective function, specified as the comma-separated pair consisting of `'tolobj'` and a positive scalar value.

Let L^t denote the value of the loglikelihood objective function at iteration t, and τ_ℓ be the tolerance specified by `tolobj`. The convergence criterion for the objective function is

$$\left|L^t - L^{t-1}\right| < \tau_\ell\left(1 + \left|L^t\right|\right).$$

Iterations continue until estimates are within the convergence tolerances `tolbeta` and `tolobj`, or the maximum number of iterations specified by `maxiter` is reached. If both `tolbeta` and `tolobj` are 0, then `mvregress` performs `maxiter` iterations with no convergence tests.

Example: `'tolobj',1e-5`

`varformat` — Format for parameter estimate variance-covariance matrix
`'beta'` (default) | `'full'`

Format for the parameter estimate variance-covariance matrix, `CovB`, specified as the comma-separated pair consisting of `'varformat'` and one of the following.

`'beta'`	Return the variance-covariance matrix for only the regression coefficient estimates, `beta`.
`'full'`	Return the variance-covariance matrix for both the regression coefficient estimates, `beta`, and the variance-covariance matrix estimate, `Sigma`.

Example: `'varformat','full'`

vartype — Type of variance-covariance matrix for parameter estimates
`'hessian'` (default) | `'fisher'`

Type of variance-covariance matrix for parameter estimates, specified as the comma-separated pair consisting of `'vartype'` and either `'hessian'` or `'fisher'`.

- If the value is `'hessian'`, then `mvregress` uses the Hessian, or observed information, matrix to compute `CovB`.
- If the value is `'fisher'`, then `mvregress` uses the complete-data Fisher, or expected information, matrix to compute `CovB`.

The `'hessian'` method takes into account the increase uncertainties due to missing data, while the `'fisher'` method does not.

Example: `'vartype','fisher'`

Output Arguments

beta — Estimated regression coefficients
column vector | matrix

Estimated regression coefficients, returned as a column vector or matrix.

- If you specify `X` as a single *n*-by-*K* design matrix, then `mvregress` returns `beta` as a column vector of length *K*. For example, if `X` is a 20-by-5 design matrix, then `beta` is a 5-by-1 column vector.
- If you specify `X` as a cell array containing one or more *d*-by-*K* design matrices, then `mvregress` returns `beta` as a column vector of length *K*. For example, if `X` is a cell array containing 2-by-10 design matrices, then `beta` is a 10-by-1 column vector.
- If you specify `X` as a single *n*-by-*p* design matrix (not in a cell array), and `Y` has dimension $d > 1$, then `mvregress` returns `beta` as a *p*-by-*d* matrix. For example, if `X` is a 20-by-5 design matrix, and `Y` has two dimensions such that $d = 2$, then `beta` is a 5-by-2 matrix, and the fitted `Y` values are `X` × `beta`.

Sigma — Estimated variance-covariance matrix
square matrix

Estimated variance-covariance matrix for the responses in `Y`, returned as a *d*-by-*d* square matrix.

Note The estimated variance-covariance matrix, `Sigma`, is not the sample covariance matrix of the residual matrix, `E`.

E — Residuals
matrix

Residuals for the fitted regression model, returned as an *n*-by-*d* matrix.

If `algorithm` has the value `'ecm'` or `'cwls'`, then `mvregress` computes the residual values corresponding to missing values in `Y` as the difference between the conditionally imputed values on page 33-4080 and the fitted values.

Note If `algorithm` has the value `'mvn'`, then `mvregress` removes observations with missing response values before estimation.

`CovB` — Parameter estimate variance-covariance matrix
square matrix

Parameter estimate variance-covariance matrix, returned as a square matrix.

- If `varformat` has the value `'beta'` (default), then `CovB` is the estimated variance-covariance matrix of the coefficient estimates in `beta`.
- If `varformat` has the value `'full'`, then `CovB` is the estimated variance-covariance matrix of the combined estimates in `beta` and `Sigma`.

`logL` — Loglikelihood objective function value
scalar value

Loglikelihood objective function value after the last iteration, returned as a scalar value.

More About

Multivariate Normal Regression

Multivariate normal regression is the regression of a d-dimensional response on a design matrix of predictor variables, with normally distributed errors. The errors can be heteroscedastic and correlated.

The model is

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{e}_i, \quad i = 1, \ldots, n,$$

where

- $\mathbf{y}_i$ is a d-dimensional vector of responses.
- $\mathbf{X}_i$ is a design matrix of predictor variables.
- β is vector or matrix of regression coefficients.
- $\mathbf{e}_i$ is a d-dimensional vector of error terms, with multivariate normal distribution

  $$\mathbf{e}_i \sim MVN_d(\mathbf{0}, \Sigma).$$

Conditionally Imputed Values

The expectation/conditional maximization (`'ecm'`) and covariance-weighted least squares (`'cwls'`) estimation algorithms include imputation of missing response values.

Let $\tilde{\mathbf{y}}$ denote missing observations. The conditionally imputed values are the expected value of the missing observation given the observed data, $E(\tilde{\mathbf{y}}|\mathbf{y})$.

The joint distribution of the missing and observed responses is a multivariate normal distribution,

$$\begin{pmatrix}\tilde{\mathbf{y}}\\ \mathbf{y}\end{pmatrix} \sim MVN\left\{\begin{pmatrix}\tilde{\mathbf{X}}\beta\\ \mathbf{X}\beta\end{pmatrix}, \begin{pmatrix}\Sigma_{\tilde{y}} & \Sigma_{\tilde{y}y}\\ \Sigma_{y\tilde{y}} & \Sigma_y\end{pmatrix}\right\}.$$

Using properties of the multivariate normal distribution, the imputed conditional expectation is given by

$$E(\tilde{\mathbf{y}}\,|\,\mathbf{y}) = \tilde{\mathbf{X}}\beta + \Sigma_{\tilde{y}y}\Sigma_{y}^{-1}(\mathbf{y} - \mathbf{X}\beta)\,.$$

Note `mvregress` only imputes missing response values. Observations with missing values in the design matrix are removed.

References

[1] Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2002.

[2] Meng, Xiao-Li, and Donald B. Rubin. "Maximum Likelihood Estimation via the ECM Algorithm." *Biometrika*. Vol. 80, No. 2, 1993, pp. 267–278.

[3] Sexton, Joe, and A. R. Swensen. "ECM Algorithms that Converge at the Rate of EM." *Biometrika*. Vol. 87, No. 3, 2000, pp. 651–662.

[4] Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*. Series B, Vol. 39, No. 1, 1977, pp. 1–37.

See Also

`manova1` | `mvregresslike`

Topics

"Set Up Multivariate Regression Problems" on page 15-11
"Multivariate General Linear Model" on page 15-20
"Fixed Effects Panel Model with Concurrent Correlation" on page 15-24
"Longitudinal Analysis" on page 15-30
"Multivariate Linear Regression" on page 15-2
"Estimation of Multivariate Regression Models" on page 15-5

Introduced in R2006b

mvregresslike

Negative log-likelihood for multivariate regression

Syntax

```
nlogL = mvregresslike(X,Y,b,SIGMA,alg)
[nlogL,COVB] = mvregresslike(...)
[nlogL,COVB] = mvregresslike(...,type,format)
```

Description

`nlogL = mvregresslike(X,Y,b,SIGMA,alg)` computes the negative log-likelihood `nlogL` for a multivariate regression of the *d*-dimensional multivariate observations in the *n*-by-*d* matrix `Y` on the predictor variables in the matrix or cell array `X`, evaluated for the *p*-by-1 column vector `b` of coefficient estimates and the *d*-by-*d* matrix `SIGMA` specifying the covariance of a row of `Y`. If $d = 1$, `X` can be an *n*-by-*p* design matrix of predictor variables. For any value of *d*, `X` can also be a cell array of length *n*, with each cell containing a *d*-by-*p* design matrix for one multivariate observation. If all observations have the same *d*-by-*p* design matrix, `X` can be a single cell.

`NaN` values in `X` or `Y` are taken as missing. Observations with missing values in `X` are ignored. Treatment of missing values in `Y` depends on the algorithm specified by *`alg`*.

`alg` should match the algorithm used by `mvregress` to obtain the coefficient estimates `b`, and must be one of the following:

- `'ecm'` — ECM algorithm
- `'cwls'` — Least squares conditionally weighted by `SIGMA`
- `'mvn'` — Multivariate normal estimates computed after omitting rows with any missing values in `Y`

`[nlogL,COVB] = mvregresslike(...)` also returns an estimated covariance matrix `COVB` of the parameter estimates `b`.

`[nlogL,COVB] = mvregresslike(...,type,format)` specifies the type and format of `COVB`.

`type` is either:

- `'hessian'` — To use the Hessian or observed information. This method takes into account the increased uncertainties due to missing data. This is the default.
- `'fisher'` — To use the Fisher or expected information. This method uses the complete data expected information, and does not include uncertainty due to missing data.

`format` is either:

- `'beta'` — To compute `COVB` for `b` only. This is the default.
- `'full'` — To compute `COVB` for both `b` and `SIGMA`.

See Also

`manova1` | `mvregress`

Topics
“Multivariate Normal Distribution” on page B-98

Introduced in R2007a

mvnrnd

Multivariate normal random numbers

Syntax

```
R = mvnrnd(mu,Sigma,n)
R = mvnrnd(mu,Sigma)
```

Description

`R = mvnrnd(mu,Sigma,n)` returns a matrix `R` of `n` random vectors chosen from the same multivariate normal distribution, with mean vector `mu` and covariance matrix `Sigma`. For more information, see “Multivariate Normal Distribution” on page 33-4087.

`R = mvnrnd(mu,Sigma)` returns an *m*-by-*d* matrix `R` of random vectors sampled from *m* separate *d*-dimensional multivariate normal distributions, with means and covariances specified by `mu` and `Sigma`, respectively. Each row of `R` is a single multivariate normal random vector.

Examples

Generate Multivariate Normal Random Numbers

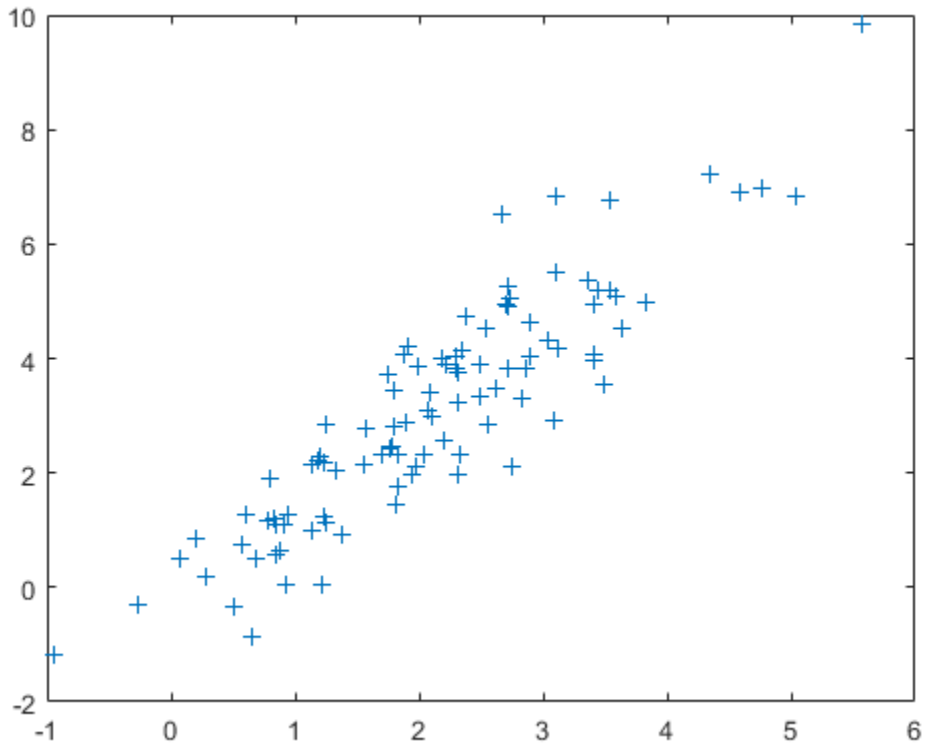
Generate random numbers from the same multivariate normal distribution.

Define `mu` and `Sigma`, and generate 100 random numbers.

```
mu = [2 3];
Sigma = [1 1.5; 1.5 3];
rng('default')  % For reproducibility
R = mvnrnd(mu,Sigma,100);
```

Plot the random numbers.

```
plot(R(:,1),R(:,2),'+')
```



Sample from Different Multivariate Normal Distributions

Randomly sample from five different three-dimensional normal distributions.

Specify the means `mu` and the covariances `Sigma` of the distributions. Let all the distributions share the same covariance matrix, but vary the mean vectors.

```
firstDim = (1:5)';
mu = repmat(firstDim,1,3)
```

```
mu = 5×3

     1     1     1
     2     2     2
     3     3     3
     4     4     4
     5     5     5
```

```
Sigma = eye(3)
```

```
Sigma = 3×3

     1     0     0
     0     1     0
```

```
     0     0     1
```

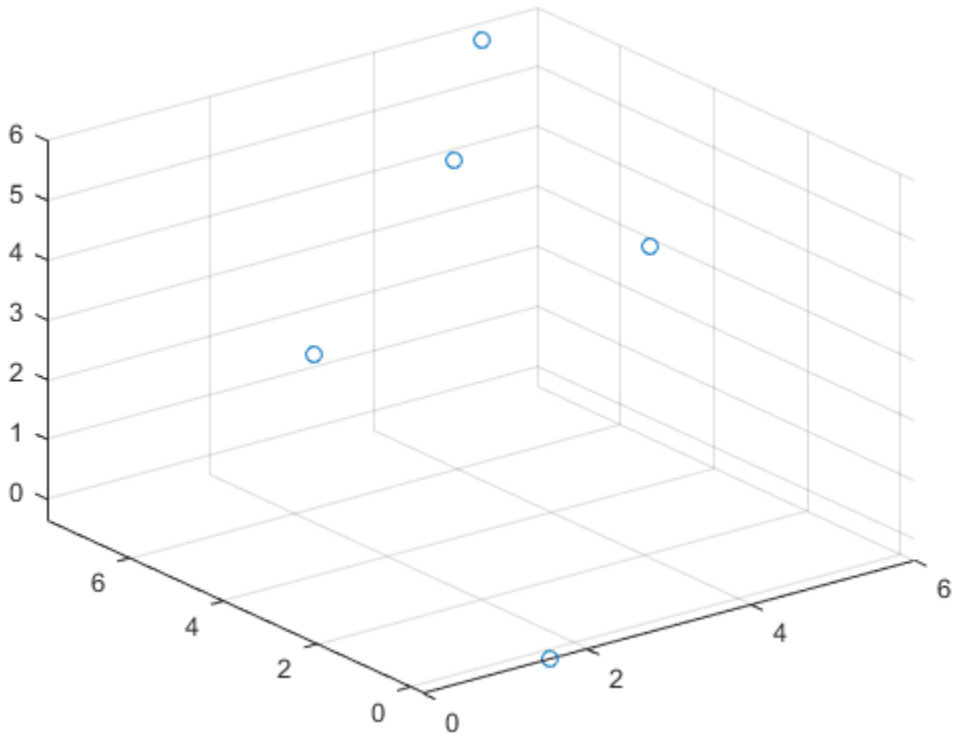
Randomly sample once from each of the five distributions.

```
rng('default')  % For reproducibility
R = mvnrnd(mu,Sigma)
```

```
R = 5×3

    1.5377   -0.3077   -0.3499
    3.8339    1.5664    5.0349
    0.7412    3.3426    3.7254
    4.8622    7.5784    3.9369
    5.3188    7.7694    5.7147
```

Plot the results.

```
scatter3(R(:,1),R(:,2),R(:,3))
```



Input Arguments

`mu` — Means of multivariate normal distributions

numeric vector | numeric matrix

Means of multivariate normal distributions, specified as a 1-by-*d* numeric vector or an *m*-by-*d* numeric matrix.

- If `mu` is a vector, then `mvnrnd` replicates the vector to match the trailing dimension of `Sigma`.
- If `mu` is a matrix, then each row of `mu` is the mean vector of a single multivariate normal distribution.

Data Types: `single` | `double`

`Sigma` — Covariances of multivariate normal distributions
symmetric, positive semi-definite matrix | numeric array

Covariances of multivariate normal distributions, specified as a *d*-by-*d* symmetric, positive semi-definite matrix or a *d*-by-*d*-by-*m* numeric array.

- If `Sigma` is a matrix, then `mvnrnd` replicates the matrix to match the number of rows in `mu`.
- If `Sigma` is an array, then each page of `Sigma`, `Sigma(:,:,i)`, is the covariance matrix of a single multivariate normal distribution and, therefore, is a symmetric, positive semi-definite matrix.

If the covariance matrices are diagonal, containing variances along the diagonal and zero covariances off it, then you can also specify `Sigma` as a 1-by-*d* vector or a 1-by-*d*-by-*m* array containing just the diagonal entries.

Data Types: `single` | `double`

`n` — Number of multivariate random numbers
positive scalar integer

Number of multivariate random numbers, specified as a positive scalar integer. `n` specifies the number of rows in `R`.

Data Types: `single` | `double`

Output Arguments

`R` — Multivariate normal random numbers
numeric matrix

Multivariate normal random numbers, returned as one of the following:

- *m*-by-*d* numeric matrix, where *m* and *d* are the dimensions specified by `mu` and `Sigma`
- `n`-by-*d* numeric matrix, where `n` is the specified input argument and *d* is the dimension specified by `mu` and `Sigma`

If `mu` is a matrix and `Sigma` is an array, then `mvnrnd` computes `R(i,:)` using `mu(i,:)` and `Sigma(:,:,i)`.

More About

Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the univariate normal distribution to two or more variables. It has two parameters, a mean vector μ and a covariance matrix Σ, that are analogous

to the mean and variance parameters of a univariate normal distribution. The diagonal elements of Σ contain the variances for each variable, and the off-diagonal elements of Σ contain the covariances between variables.

The probability density function (pdf) of the d-dimensional multivariate normal distribution is

$$y = f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp\left(-\frac{1}{2}(x\text{-}\mu)\ \Sigma^{-1}(x\text{-}\mu)'\right)$$

where x and μ are 1-by-d vectors and Σ is a d-by-d symmetric, positive definite matrix. Only `mvnrnd` allows positive semi-definite Σ matrices, which can be singular. The pdf cannot have the same form when Σ is singular.

The multivariate normal cumulative distribution function (cdf) evaluated at x is the probability that a random vector v, distributed as multivariate normal, lies within the semi-infinite rectangle with upper limits defined by x:

$$\Pr\{v(1) \le x(1), v(2) \le x(2), \ldots, v(d) \le x(d)\}\ .$$

Although the multivariate normal cdf does not have a closed form, `mvncdf` can compute cdf values numerically.

Tips

- `mvnrnd` requires the matrix `Sigma` to be symmetric. If `Sigma` has only minor asymmetry, you can use `(Sigma + Sigma')/2` instead to resolve the asymmetry.
- In the one-dimensional case, `Sigma` is the variance, not the standard deviation. For example, `mvnrnd(0,4)` is the same as `normrnd(0,2)`, where `4` is the variance and `2` is the standard deviation.

References

[1] Kotz, S., N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions: Volume 1: Models and Applications.* 2nd ed. New York: John Wiley & Sons, Inc., 2000.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`mvncdf` | `mvnpdf` | `normrnd`

Topics

“Multivariate Normal Distribution” on page B-98

Introduced before R2006a

mvtcdf

Multivariate *t* cumulative distribution function

Syntax

```
y = mvtcdf(X,C,DF)
y = mvtcdf(xl,xu,C,DF)
[y,err] = mvtcdf(...)
[...] = mvntdf(...,options)
```

Description

`y = mvtcdf(X,C,DF)` returns the cumulative probability of the multivariate *t* distribution with correlation parameters `C` and degrees of freedom `DF`, evaluated at each row of `X`. Rows of the *n*-by-*d* matrix `X` correspond to observations or points, and columns correspond to variables or coordinates. `y` is an `n`-by-`1` vector.

`C` is a symmetric, positive definite, *d*-by-*d* matrix, typically a correlation matrix. If its diagonal elements are not 1, `mvtcdf` scales `C` to correlation form. `mvtcdf` does not rescale `X`. `DF` is a scalar, or a vector with *n* elements.

The multivariate *t* cumulative probability at `X` is defined as the probability that a random vector `T`, distributed as multivariate *t*, will fall within the semi-infinite rectangle with upper limits defined by `X`, i.e., `Pr{T(1)≤X(1),T(2)≤X(2),...T(`*d*`)≤X(`*d*`)}`.

`y = mvtcdf(xl,xu,C,DF)` returns the multivariate *t* cumulative probability evaluated over the rectangle with lower and upper limits defined by `xl` and `xu`, respectively.

`[y,err] = mvtcdf(...)` returns an estimate of the error in `y`. For bivariate and trivariate distributions, `mvtcdf` uses adaptive quadrature on a transformation of the *t* density, based on methods developed by Genz, as described in the references. The default absolute error tolerance for these cases is `1e-8`. For four or more dimensions, `mvtcdf` uses a quasi-Monte Carlo integration algorithm based on methods developed by Genz and Bretz, as described in the references. The default absolute error tolerance for these cases is `1e-4`.

`[...] = mvntdf(...,options)` specifies control parameters for the numerical integration used to compute `y`. This argument can be created by a call to `statset`. Choices of `statset` parameters are:

- `'TolFun'` — Maximum absolute error tolerance. Default is `1e-8` when $d < 4$, or `1e-4` when $d \geq 4$.
- `'MaxFunEvals'` — Maximum number of integrand evaluations allowed when $d \geq 4$. Default is `1e7`. `'MaxFunEvals'` is ignored when $d < 4$.
- `'Display'` — Level of display output. Choices are `'off'` (the default), `'iter'`, and `'final'`. `'Display'` is ignored when $d < 4$.

Examples

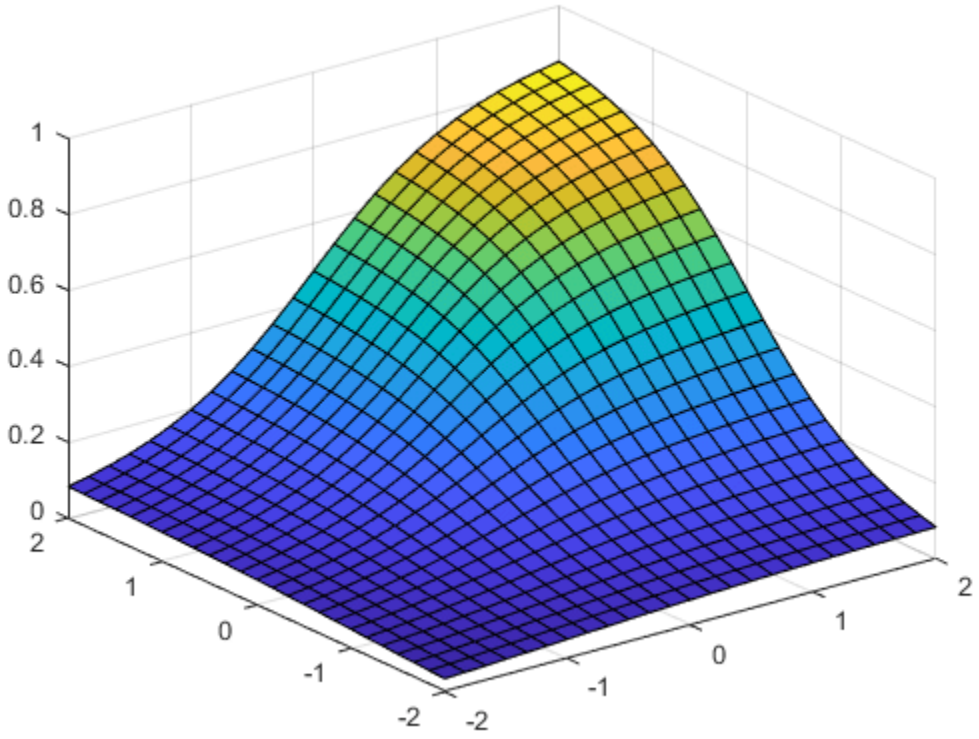
Compute the Multivariate *t* Distribution cdf

Compute the cdf of a multivariate *t* distribution with correlation parameters `C = [1 .4; .4 1]` and 2 degrees of freedom.

```
C = [1 .4; .4 1];
df = 2;
[X1,X2] = meshgrid(linspace(-2,2,25)',linspace(-2,2,25)');
X = [X1(:) X2(:)];
p = mvtcdf(X,C,df);
```

Plot the cdf.

```
figure;
surf(X1,X2,reshape(p,25,25));
```



References

[1] Genz, A. “Numerical Computation of Rectangular Bivariate and Trivariate Normal and t Probabilities.” *Statistics and Computing*. Vol. 14, No. 3, 2004, pp. 251–260.

[2] Genz, A., and F. Bretz. “Numerical Computation of Multivariate t Probabilities with Application to Power Calculation of Multiple Contrasts.” *Journal of Statistical Computation and Simulation*. Vol. 63, 1999, pp. 361–378.

[3] Genz, A., and F. Bretz. “Comparison of Methods for the Computation of Multivariate t Probabilities.” *Journal of Computational and Graphical Statistics*. Vol. 11, No. 4, 2002, pp. 950–971.

See Also

`mvtpdf` | `mvtrnd`

Topics

“Multivariate t Distribution” on page B-104

Introduced in R2006a

mvtpdf

Multivariate *t* probability density function

Syntax

```
y = mvtpdf(X,C,df)
```

Description

`y = mvtpdf(X,C,df)` returns the probability density of the multivariate *t* distribution with correlation parameters `C` and degrees of freedom `df`, evaluated at each row of `X`. Rows of the *n*-by-*d* matrix `X` correspond to observations or points, and columns correspond to variables or coordinates. `C` is a symmetric, positive definite, *d*-by-*d* matrix, typically a correlation matrix. If its diagonal elements are not 1, `mvtpdf` scales `C` to correlation form. `mvtcdf` does not rescale `X`. `df` is a scalar, or a vector with *n* elements. `y` is an *n*-by-1 vector.

Examples

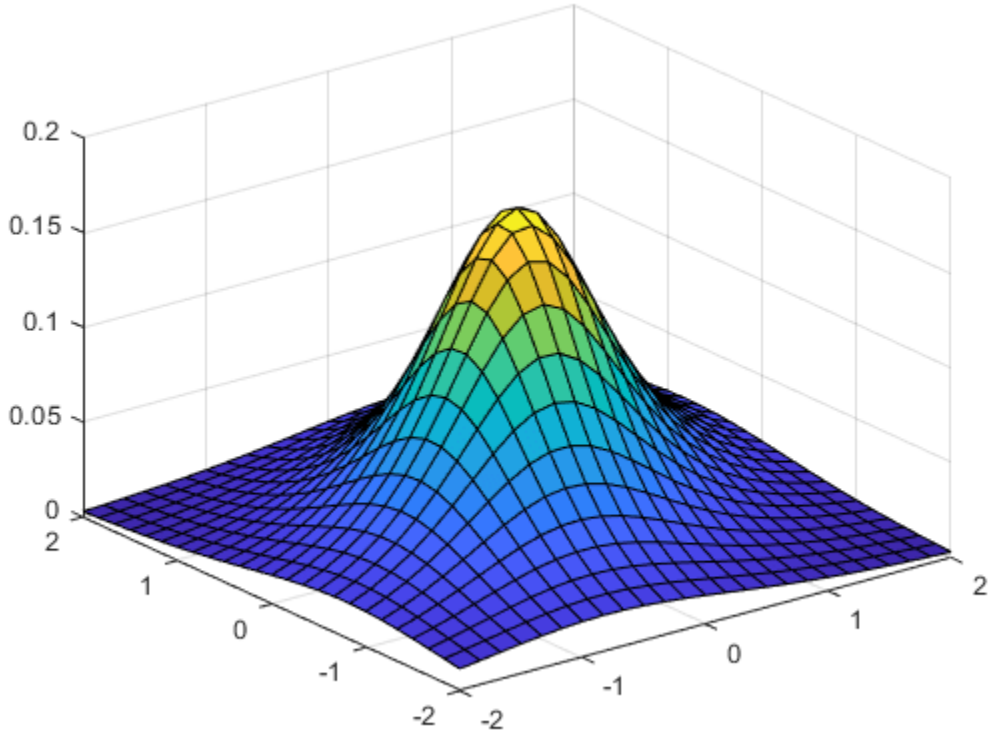
Compute the Multivariate *t* Distribution pdf

Compute the pdf of a multivariate *t* distribution with correlation parameters `C = [1 .4; .4 1]` and 2 degrees of freedom.

```
[X1,X2] = meshgrid(linspace(-2,2,25)',linspace(-2,2,25)');
X = [X1(:) X2(:)];
C = [1 .4; .4 1];
df = 2;
p = mvtpdf(X,C,df);
```

Plot the pdf.

```
figure;
surf(X1,X2,reshape(p,25,25))
```



See Also

`mvtcdf` | `mvtrnd`

Topics

“Multivariate t Distribution” on page B-104

Introduced in R2006b

mvtrnd

Multivariate *t* random numbers

Syntax

```
R = mvtrnd(C,df,cases)
R = mvtrnd(C,df)
```

Description

`R = mvtrnd(C,df,cases)` returns a matrix of random numbers chosen from the multivariate *t* distribution, where `C` is a correlation matrix. `df` is the degrees of freedom and is either a scalar or is a vector with `cases` elements. If `p` is the number of columns in `C`, then the output `R` has `cases` rows and `p` columns.

Let `t` represent a row of `R`. Then the distribution of `t` is that of a vector having a multivariate normal distribution with mean 0, variance 1, and covariance matrix `C`, divided by an independent chi-square random value having `df` degrees of freedom. The rows of `R` are independent.

`C` must be a square, symmetric and positive definite matrix. If its diagonal elements are not all 1 (that is, if `C` is a covariance matrix rather than a correlation matrix), `mvtrnd` rescales `C` to transform it to a correlation matrix before generating the random numbers.

`R = mvtrnd(C,df)` returns a single random number from the multivariate *t* distribution.

Examples

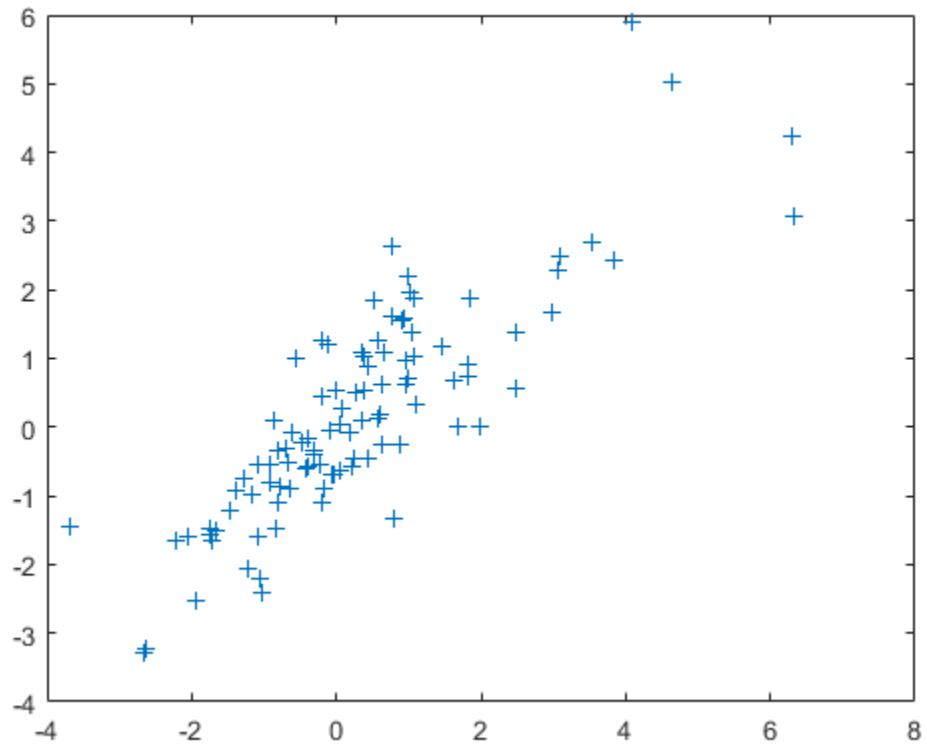
Generate Multivariate *t* Distribution Random Numbers

Generate random numbers from a multivariate *t* distribution with correlation parameters `SIGMA = [1 0.8;0.8 1]` and 3 degrees of freedom.

```
rng default;  % For reproducibility
SIGMA = [1 0.8;0.8 1];
R = mvtrnd(SIGMA,3,100);
```

Plot the random numbers.

```
figure;
plot(R(:,1),R(:,2),'+')
```



See Also

`mvtcdf` | `mvtpdf`

Topics

“Multivariate t Distribution” on page B-104

Introduced before R2006a

nancov

(Not recommended) Covariance ignoring NaN values

Note nancov is not recommended. Use the MATLAB® function cov instead. With the cov function, you can specify whether to include or omit NaN values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
Y = nancov(X)
Y = nancov(X1,X2)
Y = nancov(...,1)
Y = nancov(...,'pairwise')
```

Description

Y = nancov(X) is the covariance cov of X, computed after removing observations with NaN values.

For vectors x, nancov(x) is the sample variance of the remaining elements, once NaN values are removed. For matrices X, nancov(X) is the sample covariance of the remaining observations, once observations (rows) containing any NaN values are removed.

Y = nancov(X1,X2), where X1 and X2 are matrices with the same number of elements, is equivalent to nancov(X), where X = [X1(:) X2(:)].

nancov removes the mean from each variable (column for matrix X) before calculating Y. If n is the number of remaining observations after removing observations with NaN values, nancov normalizes Y by either $n - 1$ or n , depending on whether $n > 1$ or $n = 1$, respectively. To specify normalization by n, use Y = nancov(...,1).

Y = nancov(...,'pairwise') computes Y(i,j) using rows with no NaN values in columns i or j. The result Y may not be a positive definite matrix.

Examples

Generate random data for two variables (columns) with random missing values:

```
X = rand(10,2);
p = randperm(numel(X));
X(p(1:5)) = NaN
X =
    0.8147    0.1576
       NaN       NaN
    0.1270    0.9572
    0.9134       NaN
    0.6324       NaN
    0.0975    0.1419
    0.2785    0.4218
    0.5469    0.9157
```

```
    0.9575    0.7922
    0.9649       NaN
```

Establish a correlation between a third variable and the other two variables:

```
X(:,3) = sum(X,2)
X =
    0.8147    0.1576    0.9723
       NaN       NaN       NaN
    0.1270    0.9572    1.0842
    0.9134       NaN       NaN
    0.6324       NaN       NaN
    0.0975    0.1419    0.2394
    0.2785    0.4218    0.7003
    0.5469    0.9157    1.4626
    0.9575    0.7922    1.7497
    0.9649       NaN       NaN
```

Compute the covariance matrix for the three variables after removing observations (rows) with `NaN` values:

```
Y = nancov(X)
Y =
    0.1311    0.0096    0.1407
    0.0096    0.1388    0.1483
    0.1407    0.1483    0.2890
```

Compatibility Considerations

nancov is not recommended

Not recommended starting in R2020b

`nancov` is not recommended. Use the MATLAB function `cov` instead. There are no plans to remove `nancov`.

To update your code, change instances of the function name `nancov` to `cov`. Then specify the `'omitrows'` option for the `nanflag` input argument. The `'pairwise'` option of `nancov` corresponds to the `'partialrows'` option of `cov`.

The `cov` function offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, and C/C++ code generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- If the input is variable-size and is `[]` at run time, the generated code returns `[]` not `NaN`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

NaN | cov | var

Introduced before R2006a

nanmax

(Not recommended) Maximum, ignoring NaN values

Note nanmax is not recommended. Use the MATLAB® function max instead. With the max function, you can specify whether to include or omit NaN values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
y = nanmax(X)
y = nanmax(X,[],dim)
[y,indices] = nanmax( ___ )

y = nanmax(X,[],'all')
y = nanmax(X,[],vecdim)

Y = nanmax(X1,X2)
```

Description

y = nanmax(X) is the maximum max of X, computed after removing NaN values.

For vectors x, nanmax(x) is the maximum of the remaining elements, once NaN values are removed. For matrices X, nanmax(X) is a row vector of column maxima, once NaN values are removed. For multidimensional arrays X, nanmax operates along the first nonsingleton dimension.

y = nanmax(X,[],dim) operates along the dimension dim of X.

[y,indices] = nanmax(___) also returns the row indices of the maximum values for each column in the vector indices.

y = nanmax(X,[],'all') returns the maximum of all elements of X, computed after removing NaN values.

y = nanmax(X,[],vecdim) returns the maximum over the dimensions specified in the vector vecdim, computed after removing NaN values. Each element of vecdim represents a dimension of the input array X. The output y has length 1 in the specified operating dimensions. The other dimension lengths are the same for X and y. For example, if X is a 2-by-3-by-4 array, then nanmax(X, [],[1 2]) returns a 1-by-1-by-4 array. Each element of the output array is the maximum of the elements on the corresponding page of X.

Y = nanmax(X1,X2) returns an array Y the same size as X1 and X2 with Y(i,j) = nanmax(X1(i,j),X2(i,j)). Scalar inputs are expanded to an array of the same size as the other input.

Examples

Maximum of Matrix Data

Find the column maximum values and their indices for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
[y,indices] = nanmax(X)
```

```
y = 1×3

     4     5   NaN

indices = 1×3

     3     2     1
```

Maximum of All Values

Find the maximum of all the values in an array, ignoring missing values.

Create a 2-by-5-by-3 array X with some missing values.

```
X = reshape(1:30,[2 5 3]);
X([10:12 25]) = NaN
```

```
X =
X(:,:,1) =

     1     3     5     7     9
     2     4     6     8   NaN


X(:,:,2) =

   NaN    13    15    17    19
   NaN    14    16    18    20


X(:,:,3) =

    21    23   NaN    27    29
    22    24    26    28    30
```

Find the maximum of the elements of X.

```
y = nanmax(X,[],'all')
```

```
y = 30
```

Compatibility Considerations

nanmax is not recommended

Not recommended starting in R2020b

`nanmax` is not recommended. Use the MATLAB function `max` instead. There are no plans to remove `nanmax`.

To update your code, change instances of the function name `nanmax` to `max`. You do not need to change the input arguments. If you want to include `NaN` values, then specify the `'includenan'` option for the `nanflag` input argument.

The `max` function has these advantages over the `nanmax` function:

- `max` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.
- When you specify the `'linear'` option, `max` returns the linear index into the input array that corresponds to the maximum value.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NaN` | `max`

Introduced before R2006a

nanmean

(Not recommended) Mean, ignoring NaN values

Note nanmean is not recommended. Use the MATLAB® function mean instead. With the mean function, you can specify whether to include or omit NaN values for the calculation. For more information, see “Compatibility Considerations”.

Syntax

```
y = nanmean(X)
y = nanmean(X,'all')
y = nanmean(X,dim)
y = nanmean(X,vecdim)
```

Description

y = nanmean(X) returns the mean of the elements of X, computed after removing all NaN values.

- If X is a vector, then nanmean(X) is the mean of all the non-NaN elements of X.
- If X is a matrix, then nanmean(X) is a row vector of column means, computed after removing NaN values.
- If X is a multidimensional array, then nanmean operates along the first nonsingleton dimension of X. The size of this dimension becomes 1 while the sizes of all other dimensions remain the same. nanmean removes all NaN values.

For information on how nanmean treats arrays of all NaN values, see “Tips” on page 33-4106.

y = nanmean(X,'all') returns the mean of all elements of X, computed after removing NaN values.

y = nanmean(X,dim) returns the mean along the operating dimension dim of X, computed after removing NaN values.

y = nanmean(X,vecdim) returns the mean over the dimensions specified in the vector vecdim. The function computes the means after removing NaN values. For example, if X is a matrix, then nanmean(X,[1 2]) is the mean of all non-NaN elements of X because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

Examples

Mean of Matrix Data

Find the column means for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanmean(X)
```

```
y = 1×3

    3.5000    3.0000       NaN
```

Mean of All Values

Find the mean of all the values in an array, ignoring missing values.

Create a 2-by-5-by-3 array X with some missing values.

```
X = reshape(1:30,[2 5 3]);
X([10:12 25]) = NaN
```

```
X =
X(:,:,1) =

     1     3     5     7     9
     2     4     6     8   NaN


X(:,:,2) =

   NaN    13    15    17    19
   NaN    14    16    18    20


X(:,:,3) =

    21    23   NaN    27    29
    22    24    26    28    30
```

Find the mean of the elements of X.

```
y = nanmean(X,'all')
```

```
y = 15.6538
```

Mean Along Scalar Dimension

Find the row means for matrix data with missing values by specifying to compute the means along the second dimension.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanmean(X,2)
```

```
y = 3×1

     1
     4
     4
```

Mean Along Vector Dimension

Find the mean of a multidimensional array over multiple dimensions.

Create a 2-by-5-by-3 array X with some missing values.

```
X = reshape(1:30,[2 5 3]);
X([10:12 25]) = NaN
```

```
X =
X(:,:,1) =

     1     3     5     7     9
     2     4     6     8   NaN


X(:,:,2) =

   NaN    13    15    17    19
   NaN    14    16    18    20


X(:,:,3) =

    21    23   NaN    27    29
    22    24    26    28    30
```

Find the mean of each page of X by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = nanmean(X,[1 2])
```

```
ypage =
ypage(:,:,1) =

     5
```

```
ypage(:,:,2) =

   16.5000


ypage(:,:,3) =

   25.5556
```

For example, `ypage(1,1,1)` is the mean of the non-`NaN` elements in `X(:,:,1)`.

Find the mean of the elements in each `X(i,:,:)` slice by specifying dimensions 2 and 3 as the operating dimensions.

```
yrow = nanmean(X,[2 3])
```

```
yrow = 2×1

   14.5385
   16.7692
```

For example, `yrow(2)` is the mean of the non-`NaN` elements in `X(2,:,:)`.

Input Arguments

`X` — Input data

scalar | vector | matrix | multidimensional array

Input data, specified as a scalar, vector, matrix, or multidimensional array.

If `X` is an empty array, then `nanmean(X)` is `NaN`. For more details, see “Tips” on page 33-4106.

Data Types: `single` | `double`

`dim` — Dimension to operate along

positive integer scalar

Dimension to operate along, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension whose size does not equal 1.

`dim` indicates the dimension whose length reduces to 1. `size(y,dim)` is 1 while the sizes of all other dimensions remain the same.

Consider a two-dimensional array `X`:

- If `dim` is equal to 1, then `nanmean(X,1)` returns a row vector containing the mean for each column.
- If `dim` is equal to 2, then `nanmean(X,2)` returns a column vector containing the mean for each row.

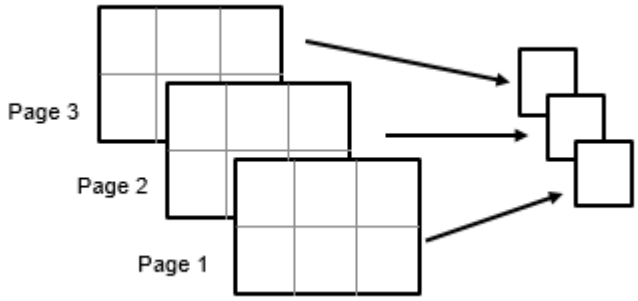
If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `nanmean` returns `X`.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`.

For example, if `X` is a 2-by-3-by-3 array, then `nanmean(X,[1 2])` returns a 1-by-1-by-3 array. Each element of the output is the mean of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

`y` — Mean values
scalar | vector | matrix | multidimensional array

Mean values, returned as a scalar, vector, matrix, or multidimensional array.

Tips

- When `nanmean` computes the mean of an array of all `NaN` values, the array is empty once the `NaN` values are removed and, therefore, the sum of the remaining elements is `0`. Because the mean calculation involves division by `0`, the mean value is `NaN`. The output `NaN` is not a mean of `NaN` values.

Compatibility Considerations

`nanmean` is not recommended
Not recommended starting in R2020b

`nanmean` is not recommended. Use the MATLAB function `mean` instead. There are no plans to remove `nanmean`.

To update your code, change instances of the function name `nanmean` to `mean`. Then specify the `'omitnan'` option for the `nanflag` input argument.

The `mean` function has these advantages over the `nanmean` function:

- `mean` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.

- `mean` returns an output value with a specified data type.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NaN` | `mean`

Introduced before R2006a

nanmedian

(Not recommended) Median, ignoring NaN values

Note nanmedian is not recommended. Use the MATLAB® function median instead. With the median function, you can specify whether to include or omit NaN values for the calculation. For more information, see “Compatibility Considerations”.

Syntax

```
y = nanmedian(X)
y = nanmedian(X,'all')
y = nanmedian(X,dim)
y = nanmedian(X,vecdim)
```

Description

y = nanmedian(X) is the median of X, computed after removing NaN values.

For vectors x, nanmedian(x) is the median of the remaining elements, once NaN values are removed. For matrices X, nanmedian(X) is a row vector of column medians, once NaN values are removed. For multidimensional arrays X, nanmedian operates along the first nonsingleton dimension.

y = nanmedian(X,'all') returns the median of all elements of X, computed after removing NaN values.

y = nanmedian(X,dim) takes the median along the operating dimension dim of X.

y = nanmedian(X,vecdim) returns the median over the dimensions specified in the vector vecdim, computed after removing NaN values. Each element of vecdim represents a dimension of the input array X. The output y has length 1 in the specified operating dimensions. The other dimension lengths are the same for X and y. For example, if X is a 2-by-3-by-4 array, then nanmedian(X,[1 2]) returns a 1-by-1-by-4 array. Each element of the output array is the median of the elements on the corresponding page of X.

Examples

Median of Matrix Data

Find the column medians for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanmedian(X)
```

```
y = 1×3

    3.5000    3.0000       NaN
```

Median Along Vector Dimension

Find the median of a multidimensional array over multiple dimensions.

Create a 3-by-5-by-2 array `X` with some missing values.

```
X = reshape(1:30,[3 5 2]);
X([10:12 25]) = NaN
```

```
X =
X(:,:,1) =

     1     4     7   NaN    13
     2     5     8   NaN    14
     3     6     9   NaN    15


X(:,:,2) =

    16    19    22   NaN    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the median of each page of `X` by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = nanmedian(X,[1 2])
```

```
ypage =
ypage(:,:,1) =

    6.5000


ypage(:,:,2) =

   22.5000
```

For example, `ypage(1,1,1)` is the median of the non-`NaN` elements in `X(:,:,1)`.

Find the median of the elements in each `X(:,i,:)` slice by specifying dimensions 1 and 3 as the operating dimensions.

```
ycol = nanmedian(X,[1 3])
```

```
ycol = 1×5

    9.5000   12.5000   15.5000   26.5000   21.5000
```

For example, `ycol(4)` is the median of the non-`NaN` elements in `X(:,4,:)`.

Compatibility Considerations

`nanmedian` is not recommended

Not recommended starting in R2020b

`nanmedian` is not recommended. Use the MATLAB function `median` instead. There are no plans to remove `nanmedian`.

To update your code, change instances of the function name `nanmedian` to `median`. Then specify the `'omitnan'` option for the `nanflag` input argument.

`median` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, and C/C++ code generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NaN` | `median`

Introduced before R2006a

nanmin

(Not recommended) Minimum, ignoring NaN values

Note nanmin is not recommended. Use the MATLAB® function min instead. With the min function, you can specify whether to include or omit NaN values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
y = nanmin(X)
y = nanmin(X,[],dim)
[y,indices] = nanmin( ___ )

y = nanmin(X,[],'all')
y = nanmin(X,[],vecdim)

Y = nanmin(X1,X2)
```

Description

y = nanmin(X) is the minimum min of X, computed after removing NaN values.

For vectors x, nanmin(x) is the minimum of the remaining elements, once NaN values are removed. For matrices X, nanmin(X) is a row vector of column minima, once NaN values are removed. For multidimensional arrays X, nanmin operates along the first nonsingleton dimension.

y = nanmin(X,[],dim) operates along the dimension dim of X.

[y,indices] = nanmin(___) also returns the row indices of the minimum values for each column in the vector indices.

y = nanmin(X,[],'all') returns the minimum of all elements of X, computed after removing NaN values.

y = nanmin(X,[],vecdim) returns the minimum over the dimensions specified in the vector vecdim, computed after removing NaN values. Each element of vecdim represents a dimension of the input array X. The output y has length 1 in the specified operating dimensions. The other dimension lengths are the same for X and y. For example, if X is a 2-by-3-by-4 array, then nanmin(X,[],[1 2]) returns a 1-by-1-by-4 array. Each element of the output array is the minimum of the elements on the corresponding page of X.

Y = nanmin(X1,X2) returns an array Y the same size as X1 and X2 with Y(i,j) = nanmin(X1(i,j),X2(i,j)). Scalar inputs are expanded to an array of the same size as the other input.

Examples

Minimum of Matrix Data

Find the column minimum values and their indices for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
[y,indices] = nanmin(X)
```

```
y = 1×3

     3     1   NaN


indices = 1×3

     2     1     1
```

Minimum of All Values

Find the minimum of all the values in an array, ignoring missing values.

Create a 2-by-5-by-3 array X with some missing values.

```
X = reshape(1:30,[2 5 3]);
X([10:12 25]) = NaN
```

```
X =
X(:,:,1) =

     1     3     5     7     9
     2     4     6     8   NaN


X(:,:,2) =

   NaN    13    15    17    19
   NaN    14    16    18    20


X(:,:,3) =

    21    23   NaN    27    29
    22    24    26    28    30
```

Find the minimum of the elements of X.

```
y = nanmin(X,[],'all')
```

```
y = 1
```

Compatibility Considerations

`nanmin` is not recommended

Not recommended starting in R2020b

`nanmin` is not recommended. Use the MATLAB function `min` instead. There are no plans to remove `nanmin`.

To update your code, change instances of the function name `nanmin` to `min`. You do not need to change the input arguments. If you want to include `NaN` values, then specify the `'includenan'` option for the `nanflag` input argument.

The `min` function has these advantages over the `nanmin` function:

- `min` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.
- When you specify the `'linear'` option, `min` returns the linear index into the input array that corresponds to the minimum value.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NaN` | `min`

Introduced before R2006a

nanstd

(Not recommended) Standard deviation, ignoring NaN values

Note `nanstd` is not recommended. Use the MATLAB® function `std` instead. With the `std` function, you can specify whether to include or omit `NaN` values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
y = nanstd(X)
y = nanstd(X,flag)
y = nanstd(X,flag,'all')
y = nanstd(X,flag,dim)
y = nanstd(X,flag,vecdim)
```

Description

`y = nanstd(X)` is the standard deviation `std` of `X`, computed after removing all `NaN` values.

- If `X` is a vector, then `nanstd(X)` is the sample standard deviation of all the non-`NaN` elements of `X`.
- If `X` is a matrix, then `nanstd(X)` is a row vector of column sample standard deviations, computed after removing `NaN` values.
- If `X` is a multidimensional array, then `nanstd` operates along the first nonsingleton dimension of `X`. The size of this dimension becomes 1 while the sizes of all other dimensions remain the same. `nanstd` removes all `NaN` values.
- By default, `nanstd` normalizes `y` by $n - 1$, where n is the number of remaining observations after removing observations with `NaN` values.

`y = nanstd(X,flag)` returns the standard deviation of `X` based on the normalization specified by `flag`. The `flag` is `0` (default) or `1` to specify normalization by $n - 1$ or n, respectively, where n is the number of remaining observations after removing observations with `NaN` values.

`y = nanstd(X,flag,'all')` returns the standard deviation of all elements of `X`, computed after removing `NaN` values.

`y = nanstd(X,flag,dim)` returns the standard deviation along the operating dimension `dim` of `X`, computed after removing `NaN` values.

`y = nanstd(X,flag,vecdim)` returns the standard deviation over the dimensions specified in the vector `vecdim`. The function computes the standard deviations after removing `NaN` values. For example, if `X` is a matrix, then `nanstd(X,0,[1 2])` is the sample standard deviation of all non-`NaN` elements of `X` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

Examples

Standard Deviation of Matrix Data

Find the column standard deviations for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanstd(X)
```

```
y = 1×3

    0.7071    2.8284       NaN
```

Standard Deviation for Population vs. Sample

Load the `carsmall` data set.

```
load carsmall
```

Compute the population and sample standard deviations for the `Horsepower` data. The `nanstd` function ignores the missing value in `Horsepower`.

```
y1 = nanstd(Horsepower,1)   % Population formula
```

```
y1 = 45.2963
```

```
y2 = nanstd(Horsepower,0)   % Sample formula
```

```
y2 = 45.5268
```

Standard Deviation of All Values

Find the standard deviation of all the values in an array, ignoring missing values.

Create a 3-by-4-by-2 array `X` with some missing values.

```
X = reshape(1:24,[3 4 2]);
X([8:10 18]) = NaN
```

```
X =
X(:,:,1) =

     1     4     7   NaN
     2     5   NaN    11
     3     6   NaN    12
```

```
X(:,:,2) =

    13    16    19    22
    14    17    20    23
    15   NaN    21    24
```

Find the sample standard deviation of the elements of X.

```
y = nanstd(X,0,'all')
```

```
y = 7.5385
```

Standard Deviation Along Scalar Dimension

Find the row standard deviations for matrix data with missing values. Specify to compute the sample standard deviations along the second dimension.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanstd(X,0,2)
```

```
y = 3×1

         0
    1.4142
         0
```

Standard Deviation Along Vector Dimension

Find the standard deviation of a multidimensional array over multiple dimensions.

Create a 3-by-4-by-2 array X with some missing values.

```
X = reshape(1:24,[3 4 2]);
X([8:10 18]) = NaN
```

```
X =
X(:,:,1) =

     1     4     7   NaN
     2     5   NaN    11
     3     6   NaN    12
```

```
X(:,:,2) =

    13    16    19    22
    14    17    20    23
    15   NaN    21    24
```

Find the sample standard deviation of each page of `X` by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = nanstd(X,0,[1 2])
```

```
ypage =
ypage(:,:,1) =

    3.8079


ypage(:,:,2) =

    3.7779
```

For example, `ypage(1,1,2)` is the sample standard deviation of the non-`NaN` elements in `X(:,:,2)`.

Find the sample standard deviation of the elements in each `X(i,:,:)` slice by specifying dimensions 2 and 3 as the operating dimensions.

```
yrow = nanstd(X,0,[2 3])
```

```
yrow = 3×1

    7.9102
    7.6904
    8.2158
```

For example, `yrow(3)` is the sample standard deviation of the non-`NaN` elements in `X(3,:,:)`.

Input Arguments

`X` — Input data

scalar | vector | matrix | multidimensional array

Input data, specified as a scalar, vector, matrix, or multidimensional array.

Data Types: `single` | `double`

`flag` — Indicator for normalization

`0` (default) | `1`

Indicator for the normalization used to compute the standard deviation, specified as `0` or `1`.

- If `flag` is `0` (default), then `nanstd` returns the sample standard deviation on page 33-4119 of `X`. `nanstd(X,0)` is the same as `nanstd(X)`.

- If `flag` is `1`, then `nanstd` returns the population standard deviation on page 33-4119 of `X`.

Data Types: `single` | `double`

`dim` — Dimension to operate along
positive integer scalar

Dimension to operate along, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension whose size does not equal 1.

`dim` indicates the dimension whose length reduces to 1. `size(y,dim)` is 1 while the sizes of all other dimensions remain the same.

Consider a two-dimensional array `X`:

- If `dim` is equal to 1, then `nanstd(X,0,1)` returns a row vector containing the sample standard deviation for each column.
- If `dim` is equal to 2, then `nanstd(X,0,2)` returns a column vector containing the sample standard deviation for each row.

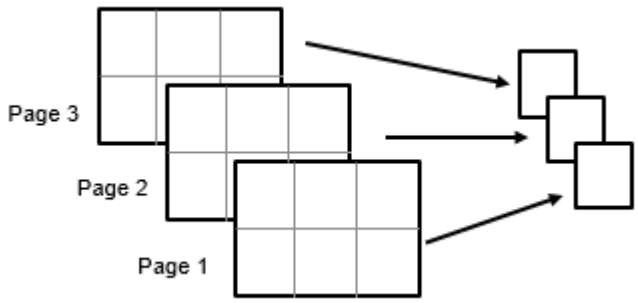
If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `nanstd` returns an array of zeros with the same dimensions and missing values as `X`.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`.

For example, if `X` is a 2-by-3-by-3 array, then `nanstd(X,0,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the sample standard deviation of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

`y` — Standard deviation values
scalar | vector | matrix | multidimensional array

Standard deviation values, returned as a scalar, vector, matrix, or multidimensional array.

More About

Sample Standard Deviation

The *sample standard deviation S* is given by

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n-1}}.$$

S is the square root of an unbiased estimator of the variance of the population from which X is drawn, as long as X consists of independent, identically distributed samples. $\bar{X}$ is the sample mean.

Notice that the denominator in this variance formula is n – 1.

Population Standard Deviation

If the data is the entire population of values, then you can use the *population standard deviation*,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}.$$

If X is a random sample from a population, then the mean μ is estimated by the sample mean, and σ is the biased maximum likelihood estimator of the population standard deviation.

Notice that the denominator in this variance formula is n.

Compatibility Considerations

`nanstd` is not recommended

Not recommended starting in R2020b

`nanstd` is not recommended. Use the MATLAB function `std` instead. There are no plans to remove `nanstd`.

To update your code, change instances of the function name `nanstd` to `std`. Then specify the `'omitnan'` option for the `nanflag` input argument.

`std` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.

- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also
`NaN` | `std`

Introduced before R2006a

nansum

(Not recommended) Sum, ignoring NaN values

Note nansum is not recommended. Use the MATLAB® function sum instead. With the sum function, you can specify whether to include or omit NaN values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
y = nansum(X)
y = nansum(X,'all')
y = nansum(X,dim)
y = nansum(X,vecdim)
```

Description

y = nansum(X) returns the sum of the elements of X, computed after removing all NaN values.

- If X is a vector, then nansum(X) is the sum of all the non-NaN elements of X.
- If X is a matrix, then nansum(X) is a row vector of column sums, computed after removing NaN values.
- If X is a multidimensional array, then nansum operates along the first nonsingleton dimension of X. The size of this dimension becomes 1 while the sizes of all other dimensions remain the same. nansum removes all NaN values.

For information on how nansum treats arrays of all NaN values, see "Tips" on page 33-4125.

y = nansum(X,'all') returns the sum of all elements of X, computed after removing NaN values.

y = nansum(X,dim) returns the sum along the operating dimension dim of X, computed after removing NaN values.

y = nansum(X,vecdim) returns the sum over the dimensions specified in the vector vecdim. The function computes the sums after removing NaN values. For example, if X is a matrix, then nansum(X,[1 2]) is the sum of all non-NaN elements of X because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

Examples

Sum of Matrix Data

Find the column sums for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3
```

```
   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nansum(X)
```

```
y = 1×3

     7     6     0
```

Sum of All Values

Find the sum of all the values in an array, ignoring missing values.

Create a 2-by-4-by-3 array X with some missing values.

```
X = reshape(1:24,[2 4 3]);
X([5:6 20]) = NaN
```

```
X =
X(:,:,1) =

     1     3   NaN     7
     2     4   NaN     8


X(:,:,2) =

     9    11    13    15
    10    12    14    16


X(:,:,3) =

    17    19    21    23
    18   NaN    22    24
```

Find the sum of the elements of X.

```
y = nansum(X,'all')
```

```
y = 269
```

Sum Along Scalar Dimension

Find the row sums for matrix data with missing values by specifying to compute the sums along the second dimension.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nansum(X,2)
```

```
y = 3×1

     1
     8
     4
```

Sum Along Vector Dimension

Find the sum of a multidimensional array over multiple dimensions.

Create a 2-by-4-by-3 array X with some missing values.

```
X = reshape(1:24,[2 4 3]);
X([5:6 20]) = NaN
```

```
X =
X(:,:,1) =

     1     3   NaN     7
     2     4   NaN     8


X(:,:,2) =

     9    11    13    15
    10    12    14    16


X(:,:,3) =

    17    19    21    23
    18   NaN    22    24
```

Find the sum of each page of X by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = nansum(X,[1 2])
```

```
ypage =
ypage(:,:,1) =

    25


ypage(:,:,2) =
```

```
   100


ypage(:,:,3) =

   144
```

For example, `ypage(1,1,1)` is the sum of the non-`NaN` elements in `X(:,:,1)`.

Find the sum of the elements in each `X(i,:,:)` slice by specifying dimensions 2 and 3 as the operating dimensions.

```
yrow = nansum(X,[2 3])

yrow = 2×1

   139
   130
```

For example, `yrow(2)` is the sum of the non-`NaN` elements in `X(2,:,:)`.

Input Arguments

`X` — Input data

scalar | vector | matrix | multidimensional array

Input data, specified as a scalar, vector, matrix, or multidimensional array.

If `X` is an empty array, then `nansum(X)` is `0`.

Data Types: `single` | `double`

`dim` — Dimension to operate along

positive integer scalar

Dimension to operate along, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension whose size does not equal 1.

`dim` indicates the dimension whose length reduces to 1. `size(y,dim)` is 1 while the sizes of all other dimensions remain the same.

Consider a two-dimensional array `X`:

- If `dim` is equal to 1, then `nansum(X,1)` returns a row vector containing the sum for each column.
- If `dim` is equal to 2, then `nansum(X,2)` returns a column vector containing the sum for each row.

If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `nansum` returns `X`, with `0` values in the place of any missing values.

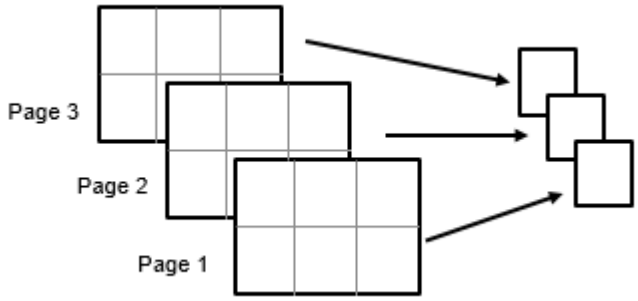
Data Types: `single` | `double`

`vecdim` — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`.

For example, if `X` is a 2-by-3-by-3 array, then `nansum(X,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the sum of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

y — Sum values

scalar | vector | matrix | multidimensional array

Sum values, returned as a scalar, vector, matrix, or multidimensional array.

Tips

- When `nansum` computes the sum of an array of all `NaN` values, the array is empty once the `NaN` values are removed and, therefore, the sum of the remaining elements is `0`. The output `0` is not a sum of `NaN` values.

Compatibility Considerations

`nansum` is not recommended

Not recommended starting in R2020b

`nansum` is not recommended. Use the MATLAB function `sum` instead. There are no plans to remove `nansum`.

To update your code, change instances of the function name `nansum` to `sum`. Then specify the `'omitnan'` option for the `nanflag` input argument.

The `sum` function has these advantages over the `nansum` function:

- `sum` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.
- `sum` returns an output value with a specified data type.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NaN` | `sum`

Introduced before R2006a

nanvar

(Not recommended) Variance, ignoring NaN values

Note `nanvar` is not recommended. Use the MATLAB® function `var` instead. With the `var` function, you can specify whether to include or omit `NaN` values for the calculation. For more information, see "Compatibility Considerations".

Syntax

```
y = nanvar(X)
y = nanvar(X,w)
y = nanvar(X,w,'all')
y = nanvar(X,w,dim)
y = nanvar(X,w,vecdim)
```

Description

`y = nanvar(X)` is the variance `var` of `X`, computed after removing `NaN` values.

For vectors `x`, `nanvar(x)` is the sample variance of the remaining elements, once `NaN` values are removed. For matrices `X`, `nanvar(X)` is a row vector of column sample variances, once `NaN` values are removed. For multidimensional arrays `X`, `nanvar` operates along the first nonsingleton dimension.

`nanvar` removes the mean from each variable (column for matrix `X`) before calculating `y`. If n is the number of remaining observations after removing observations with `NaN` values, `nanvar` normalizes `y` by either $n - 1$ or n, depending on whether $n > 1$ or $n = 1$, respectively.

`y = nanvar(X,w)` computes the variance of `X` according to the weighting scheme `w`. When `w` is `0` (default), `X` is normalized by $n - 1$, where n is the number of non-`NaN` observations. When `w` is `1`, `w` is normalized by the number of non-`NaN` observations. Otherwise, `w` can be a weight vector containing nonnegative elements. The length of `w` must equal the length of the dimension over which `nanvar` operates. Elements of `X` corresponding to `NaN` values of `w` are ignored.

`y = nanvar(X,w,'all')` returns the variance over all elements of `X` when `w = 0` or `w = 1`. The `nanvar` function computes the variance after removing `NaN` values.

`y = nanvar(X,w,dim)` returns the variance along the operating dimension `dim` of `X`.

`y = nanvar(X,w,vecdim)` returns the variance over the dimensions specified in the vector `vecdim`, computed after removing `NaN` values. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`. For example, if `X` is a 2-by-3-by-4 array, then `nanvar(X,[],[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the variance of the elements on the corresponding page of `X`. This syntax is supported when `w = 0` or `w = 1`.

Examples

Variance of Matrix Data

Find the column variances for matrix data with missing values.

```
X = magic(3);
X([1 6:9]) = NaN
```

```
X = 3×3

   NaN     1   NaN
     3     5   NaN
     4   NaN   NaN
```

```
y = nanvar(X)
```

```
y = 1×3

    0.5000    8.0000       NaN
```

Variance Along Vector Dimension

Find the variance of a multidimensional array over multiple dimensions.

Create a 3-by-4-by-2 array X with some missing values.

```
X = reshape(1:24,[3 4 2]);
X([8:10 18]) = NaN
```

```
X =
X(:,:,1) =

     1     4     7   NaN
     2     5   NaN    11
     3     6   NaN    12


X(:,:,2) =

    13    16    19    22
    14    17    20    23
    15   NaN    21    24
```

Find the sample variance of each page of X by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = nanvar(X,0,[1 2])
```

```
ypage =
ypage(:,:,1) =

   14.5000
```

```
ypage(:,:,2) =

   14.2727
```

For example, `ypage(1,1,2)` is the sample variance of the non-`NaN` elements in `X(:,:,2)`.

Find the sample variance of the elements in each `X(:,i,:)` slice by specifying dimensions 1 and 3 as the operating dimensions.

```
ycol = nanvar(X,0,[1 3])
```

```
ycol = 1×4

   44.0000   40.3000   42.9167   40.3000
```

For example, `ycol(4)` is the sample variance of the non-`NaN` elements in `X(:,4,:)`.

Compatibility Considerations

`nanvar` is not recommended

Not recommended starting in R2020b

`nanvar` is not recommended. Use the MATLAB function `var` instead. There are no plans to remove `nanvar`.

To update your code, change instances of the function name `nanvar` to `var`. Then specify the `'omitnan'` option for the `nanflag` input argument.

`var` offers more extended capabilities for supporting tall arrays, GPU arrays, distribution arrays, C/C++ code generation, and GPU code generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`NaN` | `var`

Introduced before R2006a

nearcorr

Compute nearest correlation matrix by minimizing Frobenius distance

Syntax

```
Y = nearcorr(A)
Y = nearcorr( ___ ,Name,Value)
```

Description

`Y = nearcorr(A)` returns the nearest correlation matrix `Y` by minimizing the Frobenius distance.

`Y = nearcorr( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to the input arguments in the previous syntax.

Examples

Compute the Nearest Correlation Matrix

Find the nearest correlation matrix in the Frobenius norm for a given nonpositive semidefinite matrix.

Specify an `N`-by-`N` symmetric matrix with all elements in the interval `[-1, 1]` and unit diagonal.

```
A =  [1.0000   0         0         0   -0.9360
      0    1.0000   -0.5500   -0.3645   -0.5300
      0   -0.5500    1.0000   -0.0351    0.0875
      0   -0.3645   -0.0351    1.0000    0.4557
     -0.9360   -0.5300    0.0875    0.4557    1.0000];
```

Compute the eigenvalues of `A` using `eig`.

```
eig(A)
```

```
ans = 5×1

   -0.1244
    0.3396
    1.0284
    1.4457
    2.3107
```

The smallest eigenvalue is less than `0`, which indicates that `A` is not a positive semidefinite matrix.

Compute the nearest correlation matrix using `nearcorr` with the default Newton algorithm.

```
B = nearcorr(A)
```

```
B = 5×5

    1.0000    0.0372    0.0100   -0.0219   -0.8478
```

```
    0.0372    1.0000   -0.5449   -0.3757   -0.4849
    0.0100   -0.5449    1.0000   -0.0381    0.0996
   -0.0219   -0.3757   -0.0381    1.0000    0.4292
   -0.8478   -0.4849    0.0996    0.4292    1.0000
```

Compute the eigenvalues of `B`.

```
eig(B)
```

```
ans = 5×1

    0.0000
    0.3266
    1.0146
    1.4113
    2.2475
```

All of the eigenvalues are greater than or equal to `0`, which means that `B` is a positive semidefinite matrix.

When you use `nearcorr`, you can specify the alternating projections algorithm by setting the name-value pair argument `'method'` to `'projection'`.

```
nearcorr(A,'method','projection')
```

```
ans = 5×5

    1.0000    0.0372    0.0100   -0.0219   -0.8478
    0.0372    1.0000   -0.5449   -0.3757   -0.4849
    0.0100   -0.5449    1.0000   -0.0381    0.0996
   -0.0219   -0.3757   -0.0381    1.0000    0.4292
   -0.8478   -0.4849    0.0996    0.4292    1.0000
```

You can also impose elementwise weights by specifying the `'Weights'` name-value pair argument. For more information on elementwise weights, see “'Weights'” on page 33-0 .

```
W = [0.0000  1.0000  0.1000  0.1500  0.2500
     1.0000  0.0000  0.0500  0.0250  0.1500
     0.1000  0.0500  0.0000  0.2500  1
     0.1500  0.0250  0.2500  0.0000  0.2500
     0.2500  0.1500  1  0.2500  0.0000];
nearcorr(A,'Weights',W)
```

```
ans = 5×5

    1.0000    0.0014    0.0287   -0.0222   -0.8777
    0.0014    1.0000   -0.4980   -0.7268   -0.4567
    0.0287   -0.4980    1.0000   -0.0358    0.0878
   -0.0222   -0.7268   -0.0358    1.0000    0.4465
   -0.8777   -0.4567    0.0878    0.4465    1.0000
```

In addition, you can impose `N`-by-`1` vectorized weights by specifying the `'Weights'` name-value pair argument. For more information on vectorized weights, see “'Weights'” on page 33-0 .

```
W = linspace(0.1,0.01,5)'
```

```
W = 5×1

    0.1000
    0.0775
    0.0550
    0.0325
    0.0100
```

```
C = nearcorr(A,'Weights', W)
```

```
C = 5×5

    1.0000    0.0051    0.0021   -0.0056   -0.8490
    0.0051    1.0000   -0.5486   -0.3684   -0.4691
    0.0021   -0.5486    1.0000   -0.0367    0.1119
   -0.0056   -0.3684   -0.0367    1.0000    0.3890
   -0.8490   -0.4691    0.1119    0.3890    1.0000
```

Compute the eigenvalues of `C`.

```
eig(C)
```

```
ans = 5×1

    0.0000
    0.3350
    1.0272
    1.4308
    2.2070
```

All of the eigenvalues are greater than or equal to `0`, which means that `C` is a positive semidefinite matrix.

Generate a Correlation Matrix for Stocks with Missing Values

Use `nearcorr` to create a positive semidefinite matrix for a correlation matrix for stocks with missing values.

Assume that you have stock values with missing values.

```
Stock_Missing = [59.875 42.734 47.938 60.359 NaN 69.625 61.500 62.125
                 53.188 49.000 39.500 64.813 34.750 56.625 83.000 44.500
                 55.750 50.000 38.938 62.875 30.188 43.375 NaN 29.938
                 65.500 51.063 45.563 69.313 48.250 62.375 85.250 46.875
                 69.938 47.000 52.313 71.016 37.500 59.359 61.188 48.219
                 61.500 44.188 NaN 57.000 35.313 55.813 51.500 62.188
                 59.230 48.210 62.190 61.390 54.310 70.170 61.750 91.080
                 NaN 48.700 60.300 68.580 61.250 70.340 61.590 90.350
                 52.900 52.690 54.230 61.670 68.170 NaN 57.870 88.640
                 57.370 59.040 59.870 62.090 61.620 66.470 65.370 85.840];
```

Use `corr` to compute the correlation matrix and then use `eig` to check if the correlation matrix is positive semidefinite.

```
A = corr(Stock_Missing, 'Rows','pairwise');
eig(A)
```

```
ans = 8×1

   -0.1300
   -0.0398
    0.0473
    0.2325
    0.6278
    1.6276
    1.7409
    3.8936
```

`A` has eigenvalues that are less than `0`, which indicates that the correlation matrix is not positive semidefinite.

Use `nearcorr` with this correlation matrix to generate a positive semidefinite matrix where all eigenvalues are greater than or equal to `0`.

```
B = nearcorr(A);
eigenvalues = eig(B)
```

```
eigenvalues = 8×1

    0.0000
    0.0000
    0.0180
    0.2205
    0.5863
    1.6026
    1.7258
    3.8469
```


Input Arguments

`A` — Input correlation matrix

matrix

Input correlation matrix, specified as an *N*-by-*N* symmetric approximate correlation matrix with all elements in the interval `[-1 1]` and unit diagonal. The `A` input may or may not be a positive semidefinite matrix.

Example: `A = [1.0000 0 0 0 -0.9360 0 1.0000 -0.5500 -0.3645 -0.5300 0 -0.5500 1.0000 -0.0351 0.0875 0 -0.3645 -0.0351 1.0000 0.4557 -0.9360 -0.5300 0.0875 0.4557 1.0000]`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `nearcorr(A,'Tolerance',1e-7,'MaxIterations',500,'Method','newton','Weights',weight_vector)` returns a nearest correlation matrix by minimizing the Frobenius distance.

`Tolerance` — Termination tolerance for algorithm

`1e-6` (default) | positive scalar

Termination tolerance for the algorithm, specified as the comma-separated pair consisting of `'Tolerance'` and a positive scalar.

Example: `'Tolerance',1e-7`

Data Types: `single` | `double`

`MaxIterations` — Maximum number of solver iterations

`200` (default) | positive integer

Maximum number of solver iterations, specified as the comma-separated pair consisting of `'MaxIterations'` and a positive integer.

Example: `'MaxIterations',500`

Data Types: `single` | `double`

`Method` — Method for solving nearest correlation matrix problem

`'newton'` (default) | `'projection'`

Method for solving nearest correlation matrix problem, specified as the comma-separated pair consisting of `'Method'` and one of the values in the following table.

Value	Description
`'newton'`	The Newton algorithm is quadratically convergent. If you specify the `'newton'` method, `Weights` can be either a symmetric matrix or an `N`-by-`1` vector.
`'projection'`	The alternating projections algorithm can converge to the nearest correlation matrix with high accuracy, at best linearly. If you specify the `'projection'` method, `Weights` must be an `N`-by-`1` vector.

Example: `'Method','projection'`

Data Types: `char` | `string`

`Weights` — Weights for confidence levels of entries in input matrix

`[ ]` (default) | `matrix` | `vector`

Weights for confidence levels of entries in the input matrix, specified as the comma-separated pair consisting of `'Weights'` and either a symmetric matrix or an `N`-by-`1` vector.

- Symmetric matrix — When you specify `Weights` as a symmetric matrix `W` with all elements `>= 0` to do elementwise weighting, the nearest correlation matrix `Y` is computed by minimizing the norm of ($W \circ (A-Y)$). Larger weight values place greater importance on the corresponding elements in `A`.
- `N`-by-`1` vector — When you specify `Weights` as an `N`-by-`1` vector `w` with positive numeric values, the nearest correlation matrix `Y` is computed by minimizing the norm of ($\mathtt{diag}(w)^{0.5} \times (A-Y) \times \mathtt{diag}(w)^{0.5}$).

Note Matrix weights put weight on individual entries of the correlation matrix. A full matrix must be specified, but you can control which entries are more important to match. Alternatively, vector weights put weight on a full column (and the corresponding row). Fewer weights need to be specified as compared to the matrix weights, but an entire column (and the corresponding row) is weighted by a single weight.

Example: `'Weights',W`

Data Types: `single` | `double`

Output Arguments

Y — Nearest correlation matrix to input A

positive semidefinite matrix

Nearest correlation matrix to the input `A`, returned as a positive semidefinite matrix.

References

[1] Higham, N. J. "Computing the Nearest Correlation Matrix — A Problem from Finance." *IMA Journal of Numerical Analysis.* Vol. 22, Issue 3, 2002.

[2] Qi, H. and D. Sun. "An Augmented Lagrangian Dual Approach for the H-Weighted Nearest Correlation Matrix Problem." *IMA Journal of Numerical Analysis.* Vol. 31, Issue 2, 2011.

[3] Pang, J. S., D. Sun, and J. Sun. "Semismooth Homeomorphisms and Strong Stability of Semidefinite and Lorentz Complementarity Problems." *Mathematics of Operation Research.* Vol. 28, Number 1, 2003.

See Also

`corrcoef` | `corrcov` | `partialcorr`

Introduced in R2019b

nbincdf

Negative binomial cumulative distribution function

Syntax

```
y = nbincdf(x,R,p)
y = nbincdf(x,R,p,'upper')
```

Description

`y = nbincdf(x,R,p)` computes the negative binomial cdf at each of the values in `x` using the corresponding number of successes, `R` and probability of success in a single trial, `p`. `x`, `R`, and `p` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `y`. A scalar input for `x`, `R`, or `p` is expanded to a constant array with the same dimensions as the other inputs.

`y = nbincdf(x,R,p,'upper')` returns the complement of the negative binomial cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The negative binomial cdf is

$$y = F(x \mid r, p) = \sum_{i=0}^{x} \binom{r+i-1}{i} p^r q^i I_{(0,\,1,\,\ldots)}(i)$$

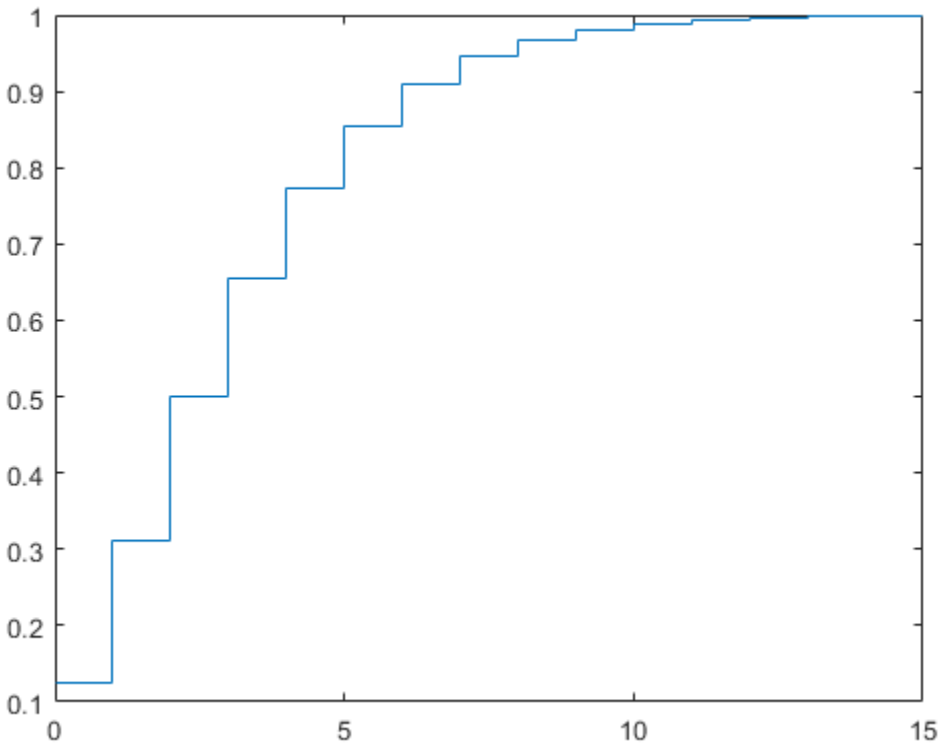
The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability `p` of success. The number of *extra* trials you must perform in order to observe a given number `R` of successes has a negative binomial distribution. However, consistent with a more general interpretation of the negative binomial, `nbincdf` allows `R` to be any positive value, including nonintegers. When `R` is noninteger, the binomial coefficient in the definition of the cdf is replaced by the equivalent expression

$$\frac{\Gamma(r+i)}{\Gamma(r)\Gamma(i+1)}$$

Examples

Compute Negative Binomial Distribution CDF

```
x = (0:15);
p = nbincdf(x,3,0.5);
stairs(x,p)
```



Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `nbinfit` | `nbininv` | `nbinpdf` | `nbinrnd` | `nbinstat`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

nbinfit

Negative binomial parameter estimates

Syntax

```
parmhat = nbinfit(data)
[parmhat,parmci] = nbinfit(data,alpha)
[...] = nbinfit(data,alpha,options)
```

Description

`parmhat = nbinfit(data)` returns the maximum likelihood estimates (MLEs) of the parameters of the negative binomial distribution given the data in the vector `data`.

`[parmhat,parmci] = nbinfit(data,alpha)` returns MLEs and `100(1-alpha)` percent confidence intervals. By default, `alpha = 0.05`, which corresponds to 95% confidence intervals.

`[...] = nbinfit(data,alpha,options)` accepts a structure, `options`, that specifies control parameters for the iterative algorithm the function uses to compute maximum likelihood estimates. The negative binomial fit function accepts an `options` structure which you can create using the function `statset`. Enter `statset('nbinfit')` to see the names and default values of the parameters that `nbinfit` accepts in the `options` structure. See the reference page for `statset` for more information about these options.

Note The variance of a negative binomial distribution is greater than its mean. If the sample variance of the data in `data` is less than its sample mean, `nbinfit` cannot compute MLEs. You should use the `poissfit` function instead.

See Also

`mle` | `nbincdf` | `nbininv` | `nbinpdf` | `nbinrnd` | `nbinstat` | `statset`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

nbininv

Negative binomial inverse cumulative distribution function

Syntax

```
X = nbininv(Y,R,P)
```

Description

`X = nbininv(Y,R,P)` returns the inverse of the negative binomial cdf with corresponding number of successes, `R` and probability of success in a single trial, `P`. Since the binomial distribution is discrete, `nbininv` returns the least integer `X` such that the negative binomial cdf evaluated at `X` equals or exceeds `Y`. `Y`, `R`, and `P` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `X`. A scalar input for `Y`, `R`, or `P` is expanded to a constant array with the same dimensions as the other inputs.

The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability `P` of success. The number of *extra* trials you must perform in order to observe a given number `R` of successes has a negative binomial distribution. However, consistent with a more general interpretation of the negative binomial, `nbininv` allows `R` to be any positive value, including nonintegers.

Examples

How many times would you need to flip a fair coin to have a 99% probability of having observed 10 heads?

```
flips = nbininv(0.99,10,0.5) + 10
flips =
  33
```

Note that you have to flip at least 10 times to get 10 heads. That is why the second term on the right side of the equals sign is a 10.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`icdf` | `nbincdf` | `nbinfit` | `nbinpdf` | `nbinrnd` | `nbinstat`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

nbinpdf

Negative binomial probability density function

Syntax

```
Y = nbinpdf(X,R,P)
```

Description

`Y = nbinpdf(X,R,P)` returns the negative binomial pdf at each of the values in `X` using the corresponding number of successes, `R` and probability of success in a single trial, `P`. `X`, `R`, and `P` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `Y`. A scalar input for `X`, `R`, or `P` is expanded to a constant array with the same dimensions as the other inputs. Note that the density function is zero unless the values in `X` are integers.

The negative binomial pdf is

$$y = f(x \mid r, p) = \binom{r + x - 1}{x} p^r q^x I_{(0,\,1,\,\ldots)}(x)$$

The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability `P` of success. The number of *extra* trials you must perform in order to observe a given number `R` of successes has a negative binomial distribution. However, consistent with a more general interpretation of the negative binomial, `nbinpdf` allows `R` to be any positive value, including nonintegers. When `R` is noninteger, the binomial coefficient in the definition of the pdf is replaced by the equivalent expression

$$\frac{\Gamma(r + x)}{\Gamma(r)\Gamma(x + 1)}$$

Examples

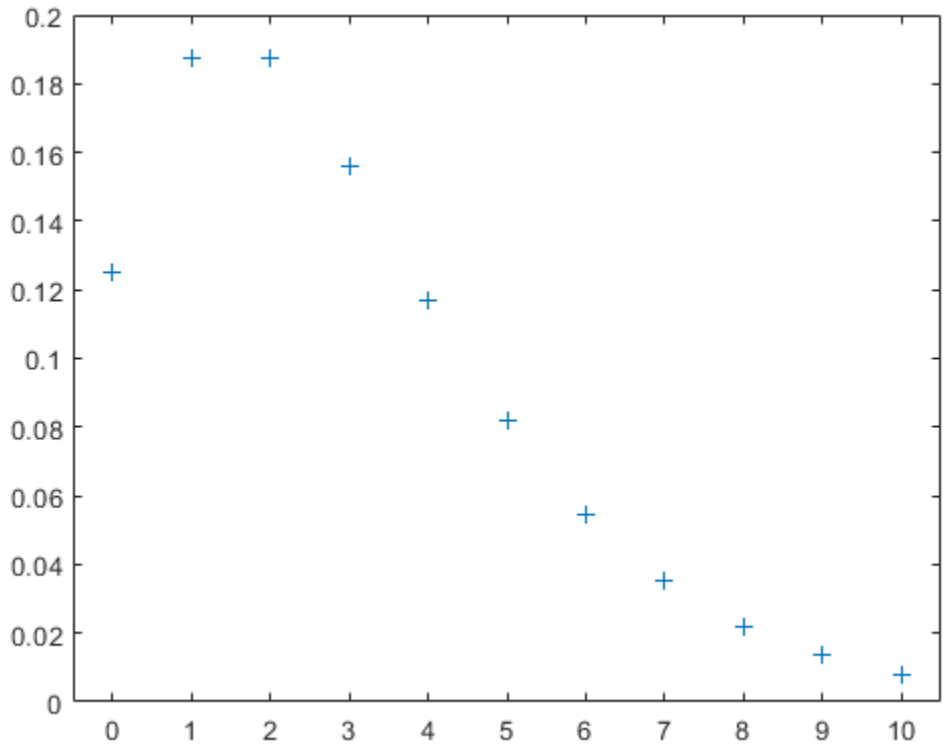
Compute the Negative Binomial Distribution pdf

Compute the pdf of a negative binomial distribution with parameters `R = 3` and `p = 0.5`.

```
x = (0:10);
y = nbinpdf(x,3,0.5);
```

Plot the pdf.

```
figure;
plot(x,y,'+')
xlim([-0.5,10.5])
```



Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`nbincdf` | `nbinfit` | `nbininv` | `nbinrnd` | `nbinstat` | `pdf`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

nbinrnd

Negative binomial random numbers

Syntax

```
RND = nbinrnd(R,P)
RND = nbinrnd(R,P,m,n,...)
RND = nbinrnd(R,P,[m,n,...])
```

Description

`RND = nbinrnd(R,P)` is a matrix of random numbers chosen from a negative binomial distribution with corresponding number of successes, `R` and probability of success in a single trial, `P`. `R` and `P` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `RND`. A scalar input for `R` or `P` is expanded to a constant array with the same dimensions as the other input.

`RND = nbinrnd(R,P,m,n,...)` or `RND = nbinrnd(R,P,[m,n,...])` generates an `m`-by-`n`-by-... array. The `R`, `P` parameters can each be scalars or arrays of the same size as `R`.

The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability `P` of success. The number of *extra* trials you must perform in order to observe a given number `R` of successes has a negative binomial distribution. However, consistent with a more general interpretation of the negative binomial, `nbinrnd` allows `R` to be any positive value, including nonintegers.

Examples

Suppose you want to simulate a process that has a defect probability of 0.01. How many units might Quality Assurance inspect before finding three defective items?

```
r = nbinrnd(3,0.01,1,6)+3
r =
  496  142  420  396  851  178
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`nbincdf` | `nbinfit` | `nbininv` | `nbinpdf` | `nbinstat` | `random`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

nbinstat

Negative binomial mean and variance

Syntax

```
[M,V] = nbinstat(R,P)
```

Description

`[M,V] = nbinstat(R,P)` returns the mean of and variance for the negative binomial distribution with corresponding number of successes, `R` and probability of success in a single trial, `P`. `R` and `P` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `M` and `V`. A scalar input for `R` or `P` is expanded to a constant array with the same dimensions as the other input.

The mean of the negative binomial distribution with parameters r and p is rq / p, where $q = 1 – p$. The variance is rq / p^2.

The simplest motivation for the negative binomial is the case of successive random trials, each having a constant probability `P` of success. The number of *extra* trials you must perform in order to observe a given number `R` of successes has a negative binomial distribution. However, consistent with a more general interpretation of the negative binomial, `nbinstat` allows `R` to be any positive value, including nonintegers.

Examples

```
p = 0.1:0.2:0.9;
r = 1:5;
[R,P] = meshgrid(r,p);
[M,V] = nbinstat(R,P)
M =
  9.0000  18.0000  27.0000  36.0000  45.0000
  2.3333  4.6667  7.0000  9.3333  11.6667
  1.0000  2.0000  3.0000  4.0000  5.0000
  0.4286  0.8571  1.2857  1.7143  2.1429
  0.1111  0.2222  0.3333  0.4444  0.5556

V =
  90.0000 180.0000 270.0000 360.0000 450.0000
  7.7778  15.5556  23.3333  31.1111  38.8889
  2.0000  4.0000  6.0000  8.0000  10.0000
  0.6122  1.2245  1.8367  2.4490  3.0612
  0.1235  0.2469  0.3704  0.4938  0.6173
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`nbincdf` | `nbinfit` | `nbininv` | `nbinpdf` | `nbinrnd`

Topics

“Negative Binomial Distribution” on page B-109

Introduced before R2006a

FeatureSelectionNCAClassification class

Feature selection for classification using neighborhood component analysis (NCA)

Description

`FeatureSelectionNCAClassification` object contains the data, fitting information, feature weights, and other parameters of a neighborhood component analysis (NCA) model. `fscnca` learns the feature weights using a diagonal adaptation of NCA and returns an instance of a `FeatureSelectionNCAClassification` object. The function achieves feature selection by regularizing the feature weights.

Construction

Create a `FeatureSelectionNCAClassification` object using `fscnca`.

Properties

`NumObservations` — Number of observations in the training data

scalar

Number of observations in the training data (`X` and `Y`) after removing `NaN` or `Inf` values, stored as a scalar.

Data Types: `double`

`ModelParameters` — Model parameters

structure

Model parameters used for training the model, stored as a structure.

You can access the fields of `ModelParameters` using dot notation.

For example, for a FeatureSelectionNCAClassification object named `mdl`, you can access the `LossFunction` value using `mdl.ModelParameters.LossFunction`.

Data Types: `struct`

`Lambda` — Regularization parameter

scalar

Regularization parameter used for training this model, stored as a scalar. For *n* observations, the best `Lambda` value that minimizes the generalization error of the NCA model is expected to be a multiple of 1/*n*.

Data Types: `double`

`FitMethod` — Name of fitting method

`'exact'` | `'none'` | `'average'`

Name of the fitting method used to fit this model, stored as one of the following:

- `'exact'` — Perform fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to `fscnca`.
- `'average'` — Divide the data into partitions (subsets), fit each partition using the `exact` method, and return the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

`Solver` — Name of the solver used to fit this model
`'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Name of the solver used to fit this model, stored as one of the following:

- `'lbfgs'` — Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm
- `'sgd'` — Stochastic gradient descent (SGD) algorithm
- `'minibatch-lbfgs'` — stochastic gradient descent with LBFGS algorithm applied to mini-batches

`GradientTolerance` — Relative convergence tolerance on gradient norm
positive scalar

Relative convergence tolerance on the gradient norm for the `'lbfgs'` and `'minibatch-lbfgs'` solvers, stored as a positive scalar value.

Data Types: `double`

`IterationLimit` — Maximum number of iterations for optimization
positive integer

Maximum number of iterations for optimization, stored as a positive integer value.

Data Types: `double`

`PassLimit` — Maximum number of passes
positive integer

Maximum number of passes for `'sgd'` and `'minibatch-lbfgs'` solvers. Every pass processes all of the observations in the data.

Data Types: `double`

`InitialLearningRate` — Initial learning rate
positive real scalar

Initial learning rate for the `'sgd'` and `'minibatch-lbfgs'` solvers, stored as a positive real scalar. The learning rate decays over iterations starting at the value specified for `InitialLearningRate`.

Use the `NumTuningIterations` and `TuningSubsetSize` name-value pair arguments to control the automatic tuning of initial learning rate in the call to `fscnca`.

Data Types: `double`

`Verbose` — Verbosity level indicator
nonnegative integer

Verbosity level indicator, stored as a nonnegative integer. Possible values are:

- 0 — No convergence summary
- 1 — Convergence summary, including norm of gradient and objective function value
- >1 — More convergence information, depending on the fitting algorithm. When you use the `'minibatch-lbfgs'` solver and verbosity level > 1, the convergence information includes the iteration log from intermediate minibatch LBFGS fits.

Data Types: `double`

`InitialFeatureWeights` — Initial feature weights
p-by-1 vector of positive real scalars

Initial feature weights, stored as a *p*-by-1 vector of positive real scalars, where *p* is the number of predictors in `X`.

Data Types: `double`

`FeatureWeights` — Feature weights
p-by-1 vector of real scalars

Feature weights, stored as a *p*-by-1 vector of real scalars, where *p* is the number of predictors in `X`.

If `FitMethod` is `'average'`, then `FeatureWeights` is a *p*-by-*m* matrix. *m* is the number of partitions specified via the `'NumPartitions'` name-value pair argument in the call to `fscnca`.

The absolute value of `FeatureWeights(k)` is a measure of the importance of predictor `k`. A `FeatureWeights(k)` value that is close to 0 indicates that predictor `k` does not influence the response in `Y`.

Data Types: `double`

`FitInfo` — Fit information
structure

Fit information, stored as a structure with the following fields.

Field Name	Meaning
`Iteration`	Iteration index
`Objective`	Regularized objective function for minimization
`UnregularizedObjective`	Unregularized objective function for minimization
`Gradient`	Gradient of regularized objective function for minimization

- For classification, `UnregularizedObjective` represents the negative of the leave-one-out accuracy of the NCA classifier on the training data.
- For regression, `UnregularizedObjective` represents the leave-one-out loss between the true response and the predicted response when using the NCA regression model.
- For the `'lbfgs'` solver, `Gradient` is the final gradient. For the `'sgd'` and `'minibatch-lbfgs'` solvers, `Gradient` is the final mini-batch gradient.
- If `FitMethod` is `'average'`, then `FitInfo` is an *m*-by-1 structure array, where *m* is the number of partitions specified via the `'NumPartitions'` name-value pair argument.

You can access the fields of `FitInfo` using dot notation. For example, for a FeatureSelectionNCAClassificationobject named `mdl`, you can access the `Objective` field using `mdl.FitInfo.Objective`.

Data Types: `struct`

`Mu` — Predictor means
p-by-1 vector | `[]`

Predictor means, stored as a *p*-by-1 vector for standardized training data. In this case, the `predict` method centers predictor matrix `X` by subtracting the respective element of `Mu` from every column.

If data is not standardized during training, then `Mu` is empty.

Data Types: `double`

`Sigma` — Predictor standard deviations
p-by-1 vector | `[]`

Predictor standard deviations, stored as a *p*-by-1 vector for standardized training data. In this case, the `predict` method scales predictor matrix `X` by dividing every column by the respective element of `Sigma` after centering the data using `Mu`.

If data is not standardized during training, then `Sigma` is empty.

Data Types: `double`

`X` — Predictor values
n-by-*p* matrix

Predictor values used to train this model, stored as an *n*-by-*p* matrix. *n* is the number of observations and *p* is the number of predictor variables in the training data.

Data Types: `double`

`Y` — Response values
numeric vector of size *n*

Response values used to train this model, stored as a numeric vector of size *n*, where n is the number of observations.

Data Types: `double`

`W` — Observation weights
numeric vector of size *n*

Observation weights used to train this model, stored as a numeric vector of size *n*. The sum of observation weights is *n*.

Data Types: `double`

Methods

loss	Evaluate accuracy of learned feature weights on test data
predict	Predict responses using neighborhood component analysis (NCA) classifier
refit	Refit neighborhood component analysis (NCA) model for classification

Examples

Explore `FeatureSelectionNCAClassification` Object

Load the sample data.

```
load ionosphere
```

The data set has 34 continuous predictors. The response variable is the radar returns, labeled as b (bad) or g (good).

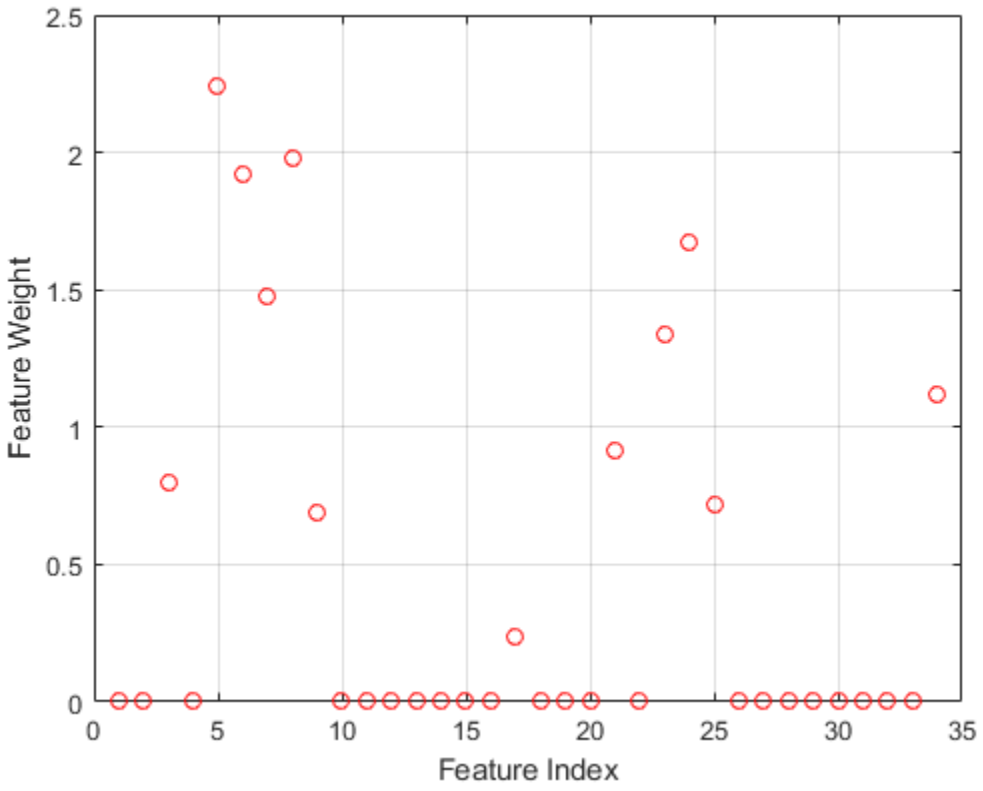
Fit a neighborhood component analysis (NCA) model for classification to detect the relevant features.

```
mdl = fscnca(X,Y);
```

The returned NCA model, `mdl`, is a `FeatureSelectionNCAClassification` object. This object stores information about the training data, model, and optimization. You can access the object properties, such as the feature weights, using dot notation.

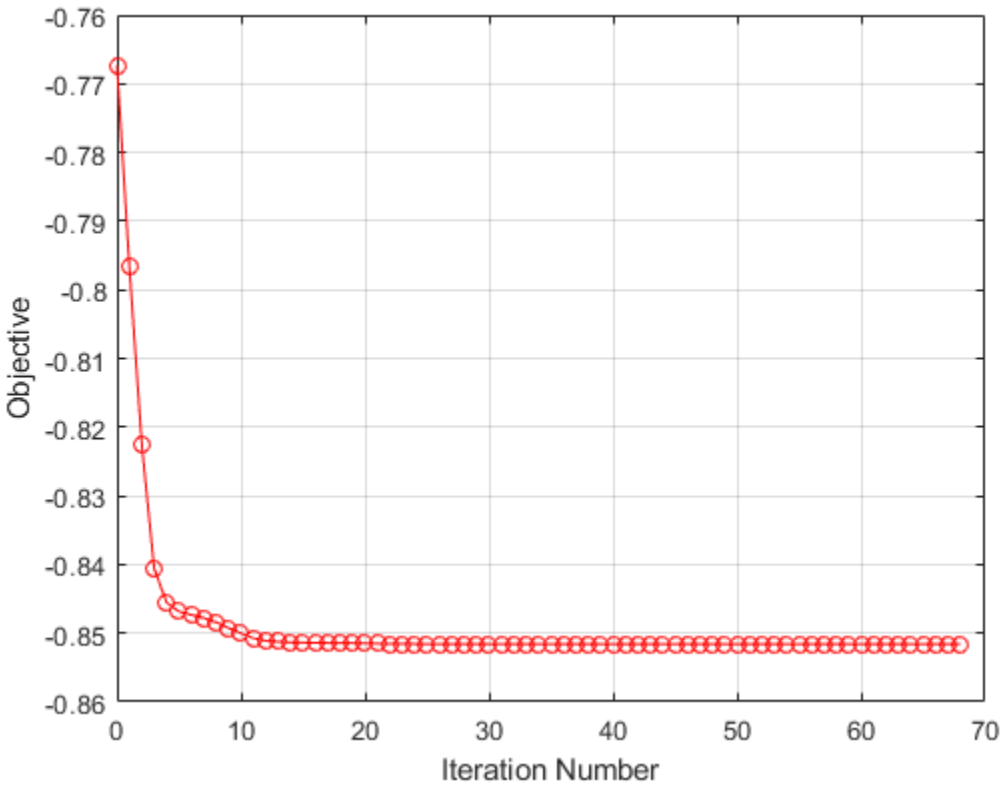
Plot the feature weights.

```
figure()
plot(mdl.FeatureWeights,'ro')
xlabel('Feature Index')
ylabel('Feature Weight')
grid on
```



The weights of the irrelevant features are zero. The `'Verbose',1` option in the call to `fscnca` displays the optimization information on the command line. You can also visualize the optimization process by plotting the objective function versus the iteration number.

```
figure
plot(mdl.FitInfo.Iteration,mdl.FitInfo.Objective,'ro-')
grid on
xlabel('Iteration Number')
ylabel('Objective')
```



The `ModelParameters` property is a `struct` that contains more information about the model. You can access the fields of this property using dot notation. For example, see if the data was standardized or not.

```
mdl.ModelParameters.Standardize
```

```
ans = logical
   0
```

`0` means that the data was not standardized before fitting the NCA model. You can standardize the predictors when they are on very different scales using the `'Standardize',1` name-value pair argument in the call to `fscnca` .

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

See Also

`fscnca` | `loss` | `predict` | `refit`

Topics
Class Attributes
Property Attributes

Introduced in R2016b

FeatureSelectionNCARegression class

Feature selection for regression using neighborhood component analysis (NCA)

Description

`FeatureSelectionNCARegression` contains the data, fitting information, feature weights, and other model parameters of a neighborhood component analysis (NCA) model. `fsrnca` learns the feature weights using a diagonal adaptation of NCA and returns an instance of `FeatureSelectionNCARegression` object. The function achieves feature selection by regularizing the feature weights.

Construction

Create a `FeatureSelectionNCAClassification` object using `fsrnca`.

Properties

`NumObservations` — Number of observations in the training data

scalar

Number of observations in the training data (`X` and `Y`) after removing `NaN` or `Inf` values, stored as a scalar.

Data Types: `double`

`ModelParameters` — Model parameters

structure

Model parameters used for training the model, stored as a structure.

You can access the fields of `ModelParameters` using dot notation.

For example, for a FeatureSelectionNCARegression object named `mdl`, you can access the `LossFunction` value using `mdl.ModelParameters.LossFunction`.

Data Types: `struct`

`Lambda` — Regularization parameter

scalar

Regularization parameter used for training this model, stored as a scalar. For *n* observations, the best `Lambda` value that minimizes the generalization error of the NCA model is expected to be a multiple of 1/*n*.

Data Types: `double`

`FitMethod` — Name of the fitting method used to fit this model

`'exact'` | `'none'` | `'average'`

Name of the fitting method used to fit this model, stored as one of the following:

- `'exact'` — Perform fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to `fsrnca`.
- `'average'` — The software divides the data into partitions (subsets), fits each partition using the `exact` method, and returns the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

`Solver` — Name of the solver used to fit this model
`'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Name of the solver used to fit this model, stored as one of the following:

- `'lbfgs'` — Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm
- `'sgd'` — Stochastic gradient descent (SGD) algorithm
- `'minibatch-lbfgs'` — stochastic gradient descent with LBFGS algorithm applied to mini-batches

`GradientTolerance` — Relative convergence tolerance on gradient norm
positive scalar

Relative convergence tolerance on the gradient norm for the `'lbfgs'` and `'minibatch-lbfgs'` solvers, stored as a positive scalar value.

Data Types: `double`

`IterationLimit` — Maximum number of iterations for optimization
positive integer

Maximum number of iterations for optimization, stored as a positive integer value.

Data Types: `double`

`PassLimit` — Maximum number of passes
positive integer

Maximum number of passes for `'sgd'` and `'minibatch-lbfgs'` solvers. Every pass processes all of the observations in the data.

Data Types: `double`

`InitialLearningRate` — Initial learning rate
positive real scalar

Initial learning rate for `'sgd'` and `'minibatch-lbfgs'` solvers. The learning rate decays over iterations starting at the value specified for `InitialLearningRate`.

Use the `NumTuningIterations` and `TuningSubsetSize` to control the automatic tuning of initial learning rate in the call to `fsrnca`.

Data Types: `double`

`Verbose` — Verbosity level indicator
nonnegative integer

Verbosity level indicator, stored as a nonnegative integer. Possible values are:

- 0 — No convergence summary
- 1 — Convergence summary, including norm of gradient and objective function value
- >1 — More convergence information, depending on the fitting algorithm. When you use the `'minibatch-lbfgs'` solver and verbosity level > 1, the convergence information includes the iteration log from intermediate minibatch LBFGS fits.

Data Types: `double`

`InitialFeatureWeights` — Initial feature weights
p-by-1 vector of positive real scalars

Initial feature weights, stored as a *p*-by-1 vector of positive real scalars, where *p* is the number of predictors in `X`.

Data Types: `double`

`FeatureWeights` — Feature weights
p-by-1 vector of real scalar values

Feature weights, stored as a *p*-by-1 vector of real scalar values, where *p* is the number of predictors in `X`.

For `'FitMethod'` equal to `'average'`, `FeatureWeights` is a *p*-by-*m* matrix, where *m* is the number of partitions specified via the `'NumPartitions'` name-value pair argument in the call to `fsrnca`.

The absolute value of `FeatureWeights(k)` is a measure of the importance of predictor `k`. If `FeatureWeights(k)` is close to 0, then this indicates that predictor `k` does not influence the response in `Y`.

Data Types: `double`

`FitInfo` — Fit information
structure

Fit information, stored as a structure with the following fields.

Field Name	Meaning
`Iteration`	Iteration index
`Objective`	Regularized objective function for minimization
`UnregularizedObjective`	Unregularized objective function for minimization
`Gradient`	Gradient of regularized objective function for minimization

- For classification, `UnregularizedObjective` represents the negative of the leave-one-out accuracy of the NCA classifier on the training data.
- For regression, `UnregularizedObjective` represents the leave-one-out loss between the true response and the predicted response when using the NCA regression model.
- For the `'lbfgs'` solver, `Gradient` is the final gradient. For the `'sgd'` and `'minibatch-lbfgs'` solvers, `Gradient` is the final mini-batch gradient.
- If `FitMethod` is `'average'`, then `FitInfo` is an *m*-by-1 structure array, where *m* is the number of partitions specified via the `'NumPartitions'` name-value pair argument.

You can access the fields of `FitInfo` using dot notation. For example, for a FeatureSelectionNCARegressionobject named `mdl`, you can access the `Objective` field using `mdl.FitInfo.Objective`.

Data Types: `struct`

`Mu` — Predictor means
p-by-1 vector | `[]`

Predictor means, stored as a *p*-by-1 vector for standardized training data. In this case, the `predict` method centers predictor matrix `X` by subtracting the respective element of `Mu` from every column.

If data is not standardized during training, then `Mu` is empty.

Data Types: `double`

`Sigma` — Predictor standard deviations
p-by-1 vector | `[]`

Predictor standard deviations, stored as a *p*-by-1 vector for standardized training data. In this case, the `predict` method scales predictor matrix `X` by dividing every column by the respective element of `Sigma` after centering the data using `Mu`.

If data is not standardized during training, then `Sigma` is empty.

Data Types: `double`

`X` — Predictor values
n-by-*p* matrix

Predictor values used to train this model, stored as an *n*-by-*p* matrix. *n* is the number of observations and *p* is the number of predictor variables in the training data.

Data Types: `double`

`Y` — Response values
numeric vector of size *n*

Response values used to train this model, stored as a numeric vector of size *n*, where n is the number of observations.

Data Types: `double`

`W` — Observation weights
numeric vector of size *n*

Observation weights used to train this model, stored as a numeric vector of size *n*. The sum of observation weights is *n*.

Data Types: `double`

Methods

loss	Evaluate accuracy of learned feature weights on test data
predict	Predict responses using neighborhood component analysis (NCA) regression model
refit	Refit neighborhood component analysis (NCA) model for regression

Examples

Explore `FeatureSelectionNCARegression` Object

Load the sample data.

```
load imports-85
```

The first 15 columns contain the continuous predictor variables, whereas the 16th column contains the response variable, which is the price of a car. Define the variables for the neighborhood component analysis model.

```
Predictors = X(:,1:15);
Y = X(:,16);
```

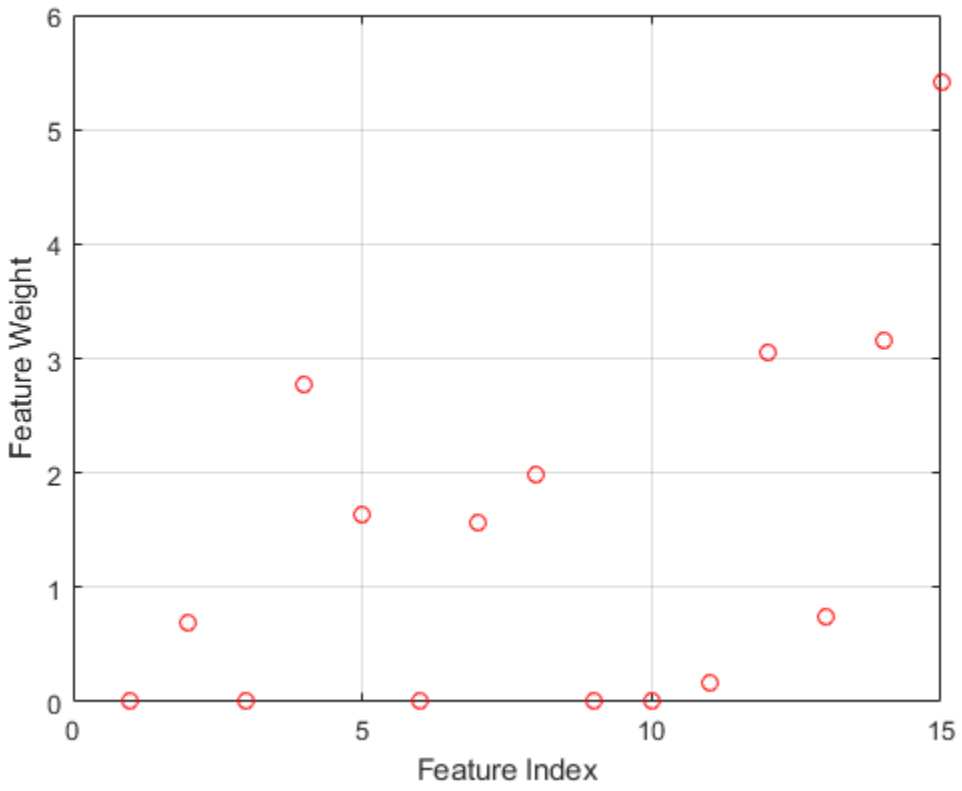
Fit a neighborhood component analysis (NCA) model for regression to detect the relevant features.

```
mdl = fsrnca(Predictors,Y);
```

The returned NCA model, `mdl`, is a `FeatureSelectionNCARegression` object. This object stores information about the training data, model, and optimization. You can access the object properties, such as the feature weights, using dot notation.

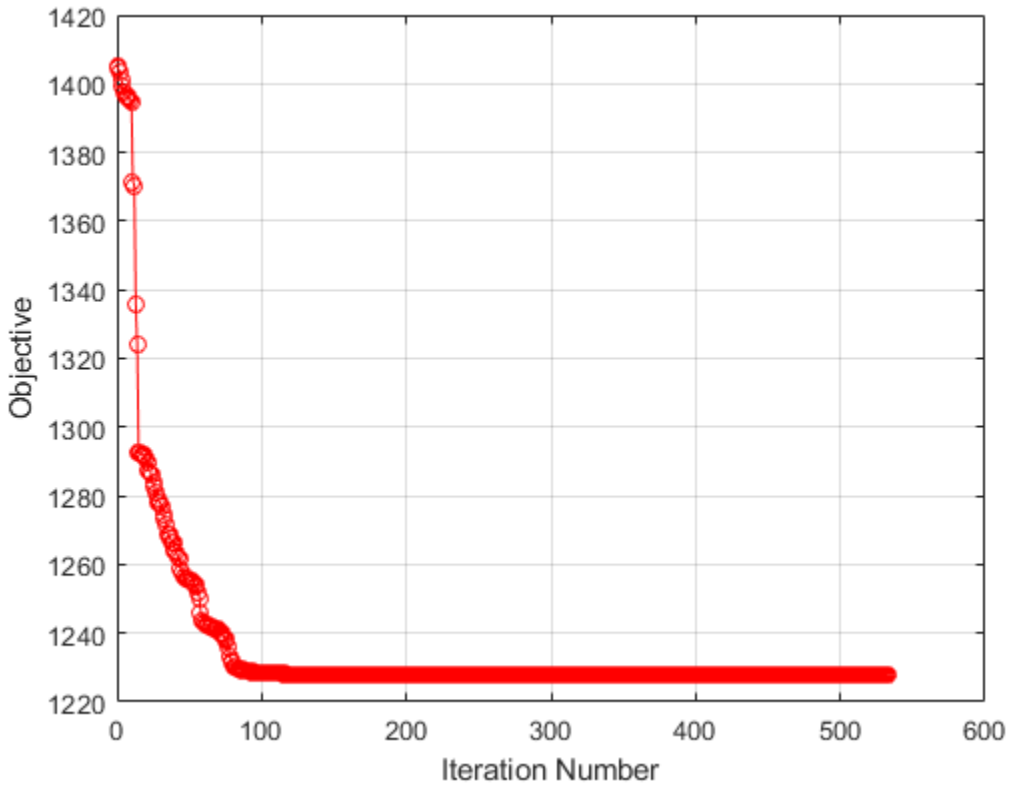
Plot the feature weights.

```
figure()
plot(mdl.FeatureWeights,'ro')
xlabel('Feature Index')
ylabel('Feature Weight')
grid on
```



The weights of the irrelevant features are zero. The `'Verbose',1` option in the call to `fsrnca` displays the optimization information on the command line. You can also visualize the optimization process by plotting the objective function versus the iteration number.

```
figure()
plot(mdl.FitInfo.Iteration,mdl.FitInfo.Objective,'ro-')
grid on
xlabel('Iteration Number')
ylabel('Objective')
```



The `ModelParameters` property is a `struct` that contains more information about the model. You can access the fields of this property using dot notation. For example, see if the data was standardized or not.

```
mdl.ModelParameters.Standardize
```

```
ans = logical
   0
```

`0` means that the data was not standardized before fitting the NCA model. You can standardize the predictors when they are on very different scales using the `'Standardize',1` name-value pair argument in the call to `fsrnca` .

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

See Also

`fsrnca` | `loss` | `predict` | `refit`

Topics
Class Attributes
Property Attributes

Introduced in R2016b

ncfcdf

Noncentral *F* cumulative distribution function

Syntax

```
p = ncfcdf(x,nu1,nu2,delta)
p = ncfcdf(x,nu1,nu2,delta,'upper')
```

Description

`p = ncfcdf(x,nu1,nu2,delta)` computes the noncentral *F* cdf at each value in `x` using the corresponding numerator degrees of freedom in `nu1`, denominator degrees of freedom in `nu2`, and positive noncentrality parameters in `delta`. `nu1`, `nu2`, and `delta` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `p`. A scalar input for `x`, `nu1`, `nu2`, or `delta` is expanded to a constant array with the same dimensions as the other inputs.

`p = ncfcdf(x,nu1,nu2,delta,'upper')` returns the complement of the noncentral *F* cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The noncentral *F* cdf is

$$F(x\left|\nu_1,\nu_2,\delta\right.) = \sum_{j=0}^{\infty}\left(\frac{\left(\frac{1}{2}\delta\right)^j}{j!}e^{\frac{-\delta}{2}}\right)I\left(\frac{\nu_1\cdot x}{\nu_2+\nu_1\cdot x}\middle|\frac{\nu_1}{2}+j,\frac{\nu_2}{2}\right)$$

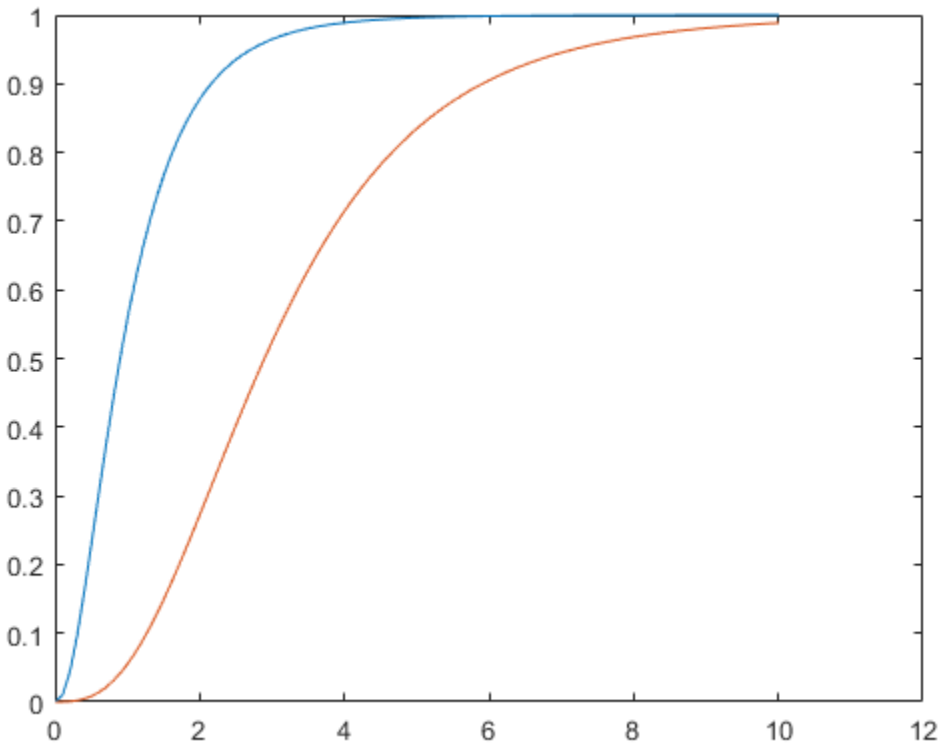
where $I(x|a,b)$ is the incomplete beta function with parameters a and b.

Examples

Compute Noncentral F Distribution cdf

Compare the noncentral *F* cdf with δ = 10 to the *F* cdf with the same number of numerator and denominator degrees of freedom (5 and 20 respectively).

```
x = (0.01:0.1:10.01)';
p1 = ncfcdf(x,5,20,10);
p = fcdf(x,5,20);
plot(x,p,'-',x,p1,'-')
```



References

[1] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 189–200.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `ncfinv` | `ncfpdf` | `ncfrnd` | `ncfstat`

Topics

“Noncentral F Distribution” on page B-115

Introduced before R2006a

ncfinv

Noncentral *F* inverse cumulative distribution function

Syntax

```
X = ncfinv(P,NU1,NU2,DELTA)
```

Description

`X = ncfinv(P,NU1,NU2,DELTA)` returns the inverse of the noncentral *F* cdf with numerator degrees of freedom `NU1`, denominator degrees of freedom `NU2`, and positive noncentrality parameter `DELTA` for the corresponding probabilities in `P`. `P`, `NU1`, `NU2`, and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `X`. A scalar input for `P`, `NU1`, `NU2`, or `DELTA` is expanded to a constant array with the same dimensions as the other inputs.

Examples

One hypothesis test for comparing two sample variances is to take their ratio and compare it to an *F* distribution. If the numerator and denominator degrees of freedom are 5 and 20 respectively, then you reject the hypothesis that the first variance is equal to the second variance if their ratio is less than that computed below.

```
critical = finv(0.95,5,20)
critical =
  2.7109
```

Suppose the truth is that the first variance is twice as big as the second variance. How likely is it that you would detect this difference?

```
prob = 1 - ncfcdf(critical,5,20,2)
prob =
  0.1297
```

If the true ratio of variances is 2, what is the typical (median) value you would expect for the *F* statistic?

```
ncfinv(0.5,5,20,2)
ans =
    1.2786
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. Hoboken, NJ: Wiley-Interscience, 2000.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 189–200.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`icdf` | `ncfcdf` | `ncfpdf` | `ncfrnd` | `ncfstat`

Topics

“Noncentral F Distribution” on page B-115

Introduced before R2006a

ncfpdf

Noncentral *F* probability density function

Syntax

```
Y = ncfpdf(X,NU1,NU2,DELTA)
```

Description

`Y = ncfpdf(X,NU1,NU2,DELTA)` computes the noncentral *F* pdf at each of the values in `X` using the corresponding numerator degrees of freedom in `NU1`, denominator degrees of freedom in `NU2`, and positive noncentrality parameters in `DELTA`. `X`, `NU1`, `N2`, and `B` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `Y`. A scalar input for `P`, `NU1`, `NU2`, or `DELTA` is expanded to a constant array with the same dimensions as the other inputs.

The *F* distribution is a special case of the noncentral *F* where $\delta = 0$. As δ increases, the distribution flattens like the plot in the example.

Examples

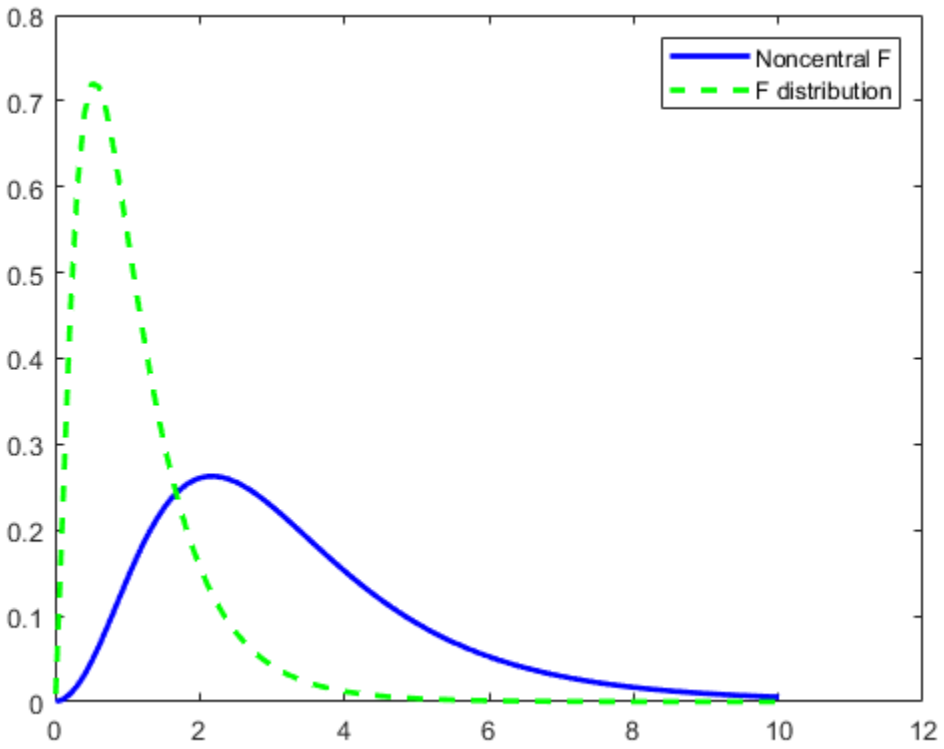
Compute Noncentral *F* Distribution pdf

Compute the pdf of a noncentral *F* distribution with degrees of freedom `NU1 = 5` and `NU2 = 20`, and noncentrality parameter `DELTA = 10`. For comparison, also compute the pdf of an *F* distribution with the same degrees of freedom.

```
x = (0.01:0.1:10.01)';
p1 = ncfpdf(x,5,20,10);
p = fpdf(x,5,20);
```

Plot the pdf of the noncentral *F* distribution and the pdf of the *F* distribution on the same figure.

```
figure;
plot(x,p1,'b-','LineWidth',2)
hold on
plot(x,p,'g--','LineWidth',2)
legend('Noncentral F','F distribution')
```



References

[1] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 189–200.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`ncfcdf` | `ncfinv` | `ncfrnd` | `ncfstat` | `pdf`

Topics

“Noncentral F Distribution” on page B-115

Introduced before R2006a

ncfrnd

Noncentral *F* random numbers

Syntax

```
R = ncfrnd(NU1,NU2,DELTA)
R = ncfrnd(NU1,NU2,DELTA,m,n,...)
R = ncfrnd(NU1,NU2,DELTA,[m,n,...])
```

Description

`R = ncfrnd(NU1,NU2,DELTA)` returns a matrix of random numbers chosen from the noncentral *F* distribution with corresponding numerator degrees of freedom in `NU1`, denominator degrees of freedom in `NU2`, and positive noncentrality parameters in `DELTA`. `NU1`, `NU2`, and `DELTA` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `R`. A scalar input for `NU1`, `NU2`, or `DELTA` is expanded to a constant matrix with the same dimensions as the other inputs.

`R = ncfrnd(NU1,NU2,DELTA,m,n,...)` or `R = ncfrnd(NU1,NU2,DELTA,[m,n,...])` generates an `m`-by-`n`-by-... array. The `NU1`, `NU2`, `DELTA` parameters can each be scalars or arrays of the same size as `R`.

Examples

Compute six random numbers from a noncentral *F* distribution with 10 numerator degrees of freedom, 100 denominator degrees of freedom and a noncentrality parameter, δ, of 4.0. Compare this to the *F* distribution with the same degrees of freedom.

```
r = ncfrnd(10,100,4,1,6)
r =
  2.5995  0.8824  0.8220  1.4485  1.4415  1.4864

r1 = frnd(10,100,1,6)
r1 =
  0.9826  0.5911  1.0967  0.9681  2.0096  0.6598
```

References

[1] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 189–200.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ncfcdf` | `ncfinv` | `ncfpdf` | `ncfstat` | `random`

Topics

“Noncentral F Distribution” on page B-115

Introduced before R2006a

ncfstat

Noncentral *F* mean and variance

Syntax

```
[M,V] = ncfstat(NU1,NU2,DELTA)
```

Description

`[M,V] = ncfstat(NU1,NU2,DELTA)` returns the mean of and variance for the noncentral *F* pdf with corresponding numerator degrees of freedom in `NU1`, denominator degrees of freedom in `NU2`, and positive noncentrality parameters in `DELTA`. `NU1`, `NU2`, and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `M` and `V`. A scalar input for `NU1`, `NU2`, or `DELTA` is expanded to a constant array with the same dimensions as the other input.

The mean of the noncentral F distribution with parameters ν_1, ν_2, and δ is

$$\frac{\nu_2(\delta+\nu_1)}{\nu_1(\nu_2-2)}$$

where $\nu_2 > 2$.

The variance is

$$2\left(\frac{\nu_2}{\nu_1}\right)^2\left[\frac{(\delta+\nu_1)^2+(2\delta+\nu_1)(\nu_2-2)}{(\nu_2-2)^2(\nu_2-4)}\right]$$

where $\nu_2 > 4$.

Examples

```
[m,v]= ncfstat(10,100,4)
m =
  1.4286
v =
  0.4252
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 73–74.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 189–200.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`ncfcdf` | `ncfinv` | `ncfpdf` | `ncfrnd`

Topics

“Noncentral F Distribution” on page B-115

Introduced before R2006a

nctcdf

Noncentral *t* cumulative distribution function

Syntax

```
p = nctcdf(x,nu,delta)
p = nctcdf(x,nu,delta,'upper')
```

Description

`p = nctcdf(x,nu,delta)` computes the noncentral *t* cdf at each value in `x` using the corresponding degrees of freedom in `nu` and noncentrality parameters in `delta`. `x`, `nu`, and `delta` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `p`. A scalar input for `x`, `nu`, or `delta` is expanded to a constant array with the same dimensions as the other inputs.

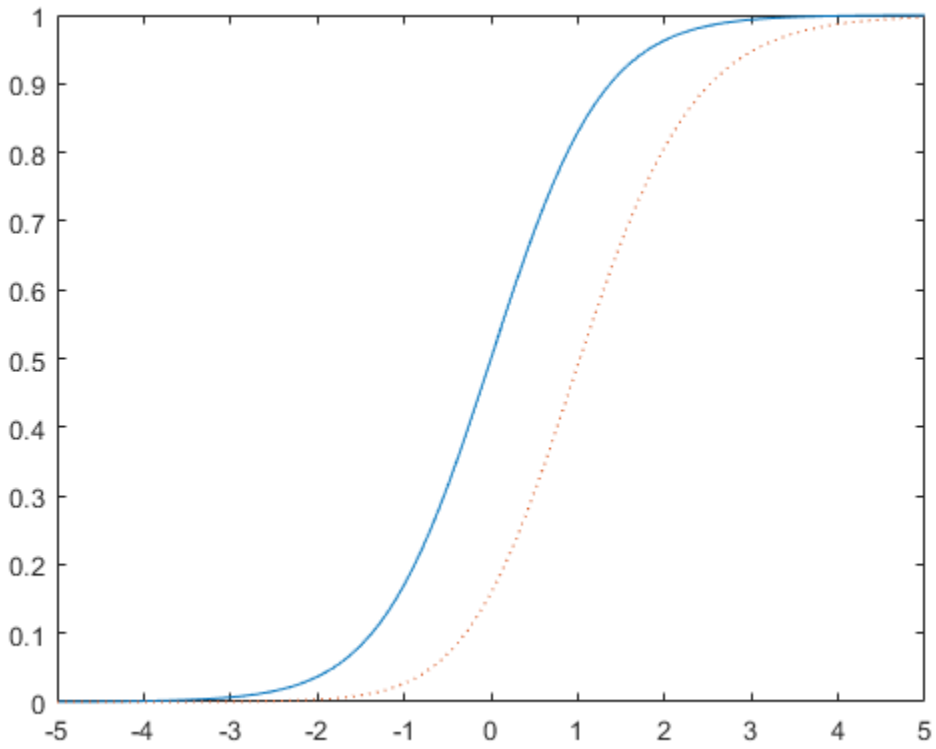
`p = nctcdf(x,nu,delta,'upper')` returns the complement of the noncentral *t* cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

Examples

Compute Noncentral t Distribution cdf

Compare the noncentral *t* cdf with `DELTA = 1` to the *t* cdf with the same number of degrees of freedom (10).

```
x = (-5:0.1:5)';
p1 = nctcdf(x,10,1);
p = tcdf(x,10);
plot(x,p,'-',x,p1,':')
```



References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 147–148.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 201–219.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `nctinv` | `nctpdf` | `nctrnd` | `nctstat`

Topics

“Noncentral t Distribution” on page B-117

Introduced before R2006a

nctinv

Noncentral *t* inverse cumulative distribution function

Syntax

```
X = nctinv(P,NU,DELTA)
```

Description

`X = nctinv(P,NU,DELTA)` returns the inverse of the noncentral *t* cdf with `NU` degrees of freedom and noncentrality parameter `DELTA` for the corresponding probabilities in `P`. `P`, `NU`, and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `X`. A scalar input for `P`, `NU`, or `DELTA` is expanded to a constant array with the same dimensions as the other inputs.

Examples

```
x = nctinv([0.1 0.2],10,1)
x =
  -0.2914  0.1618
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 147–148.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 201–219.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`icdf` | `nctcdf` | `nctpdf` | `nctrnd` | `nctstat`

Topics

“Noncentral t Distribution” on page B-117

Introduced before R2006a

nctpdf

Noncentral *t* probability density function

Syntax

```
Y = nctpdf(X,V,DELTA)
```

Description

`Y = nctpdf(X,V,DELTA)` computes the noncentral *t* pdf at each of the values in `X` using the corresponding degrees of freedom in `V` and noncentrality parameters in `DELTA`. Vector or matrix inputs for `X`, `V`, and `DELTA` must have the same size, which is also the size of `Y`. A scalar input for `X`, `V`, or `DELTA` is expanded to a constant matrix with the same dimensions as the other inputs.

Examples

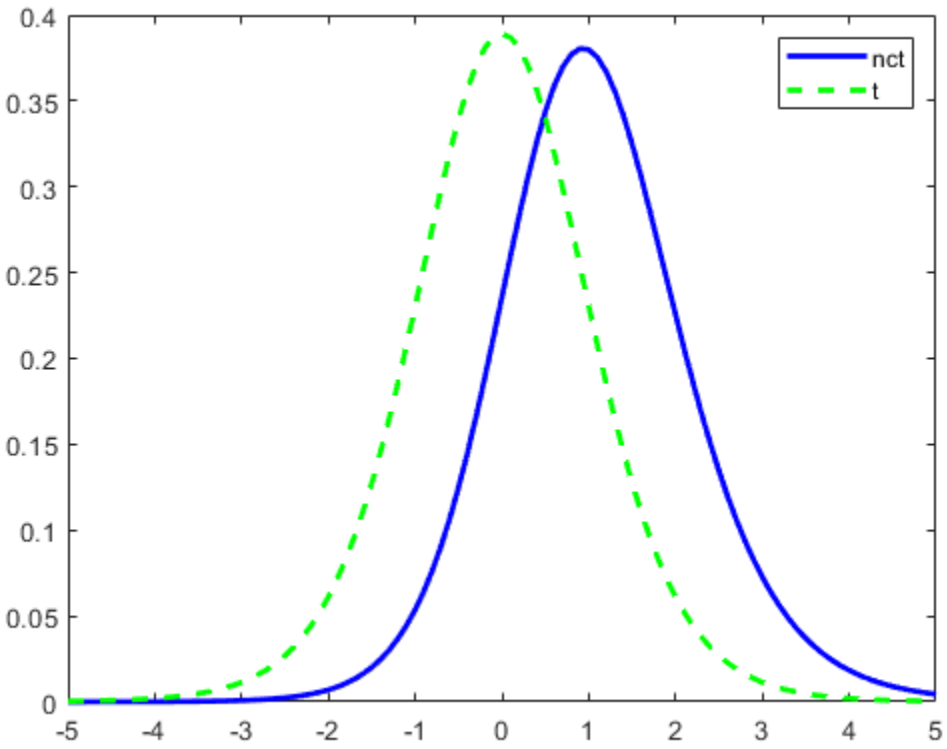
Compute Noncentral *t* Distribution pdf

Compute the pdf of a noncentral *t* distribution with degrees of freedom `V = 10` and noncentrality parameter `DELTA = 1`. For comparison, also compute the pdf of a *t* distribution with the same degrees of freedom.

```
x = (-5:0.1:5)';
nct = nctpdf(x,10,1);
t = tpdf(x,10);
```

Plot the pdf of the noncentral *t* distribution and the pdf of the *t* distribution on the same figure.

```
plot(x,nct,'b-','LineWidth',2)
hold on
plot(x,t,'g--','LineWidth',2)
legend('nct','t')
```



References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 147–148.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 201–219.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`nctcdf` | `nctinv` | `nctrnd` | `nctstat` | `pdf`

Topics

“Noncentral t Distribution” on page B-117

Introduced before R2006a

nctrnd

Noncentral *t* random numbers

Syntax

```
R = nctrnd(V,DELTA)
R = nctrnd(V,DELTA,m,n,...)
R = nctrnd(V,DELTA,[m,n,...])
```

Description

`R = nctrnd(V,DELTA)` returns a matrix of random numbers chosen from the noncentral T distribution using the corresponding degrees of freedom in `V` and noncentrality parameters in `DELTA`. `V` and `DELTA` can be vectors, matrices, or multidimensional arrays. A scalar input for `V` or `DELTA` is expanded to a constant array with the same dimensions as the other input.

`R = nctrnd(V,DELTA,m,n,...)` or `R = nctrnd(V,DELTA,[m,n,...])` generates an `m`-by-`n`-by-... array. The `V`, `DELTA` parameters can each be scalars or arrays of the same size as `R`.

Examples

```
nctrnd(10,1,5,1)
ans =
  1.6576
  1.0617
  1.4491
  0.2930
  3.6297
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 147–148.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 201–219.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.

- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`nctcdf` | `nctinv` | `nctpdf` | `nctstat` | `random`

Topics

“Noncentral t Distribution” on page B-117

Introduced before R2006a

nctstat

Noncentral t mean and variance

Syntax

```
[M,V] = nctstat(NU,DELTA)
```

Description

`[M,V] = nctstat(NU,DELTA)` returns the mean of and variance for the noncentral t pdf with `NU` degrees of freedom and noncentrality parameter `DELTA`. `NU` and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `M` and `V`. A scalar input for `NU` or `DELTA` is expanded to a constant array with the same dimensions as the other input.

The mean of the noncentral t distribution with parameters ν and δ is

$$\frac{\delta(\nu/2)^{1/2}\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}$$

where $\nu > 1$.

The variance is

$$\frac{\nu}{(\nu-2)}(1+\delta^2) - \frac{\nu}{2}\delta^2\left[\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right]^2$$

where $\nu > 2$.

Examples

```
[m,v] = nctstat(10,1)

m =
  1.0837

v =
  1.3255
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 147–148.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 201–219.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`nctcdf` | `nctinv` | `nctpdf` | `nctrnd`

Topics

“Noncentral t Distribution” on page B-117

Introduced before R2006a

ncx2cdf

Noncentral chi-square cumulative distribution function

Syntax

```
p = ncx2cdf(x,v,delta)
p = ncx2cdf(x,v,delta,'upper')
```

Description

`p = ncx2cdf(x,v,delta)` computes the noncentral chi-square cdf at each value in `x` using the corresponding degrees of freedom in `v` and positive noncentrality parameters in `delta`. `x`, `v`, and `delta` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `p`. A scalar input for `x`, `v`, or `delta` is expanded to a constant array with the same dimensions as the other inputs.

`p = ncx2cdf(x,v,delta,'upper')` returns the complement of the noncentral chi-square cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

Some texts refer to this distribution as the generalized Rayleigh, Rayleigh-Rice, or Rice distribution.

The noncentral chi-square cdf is

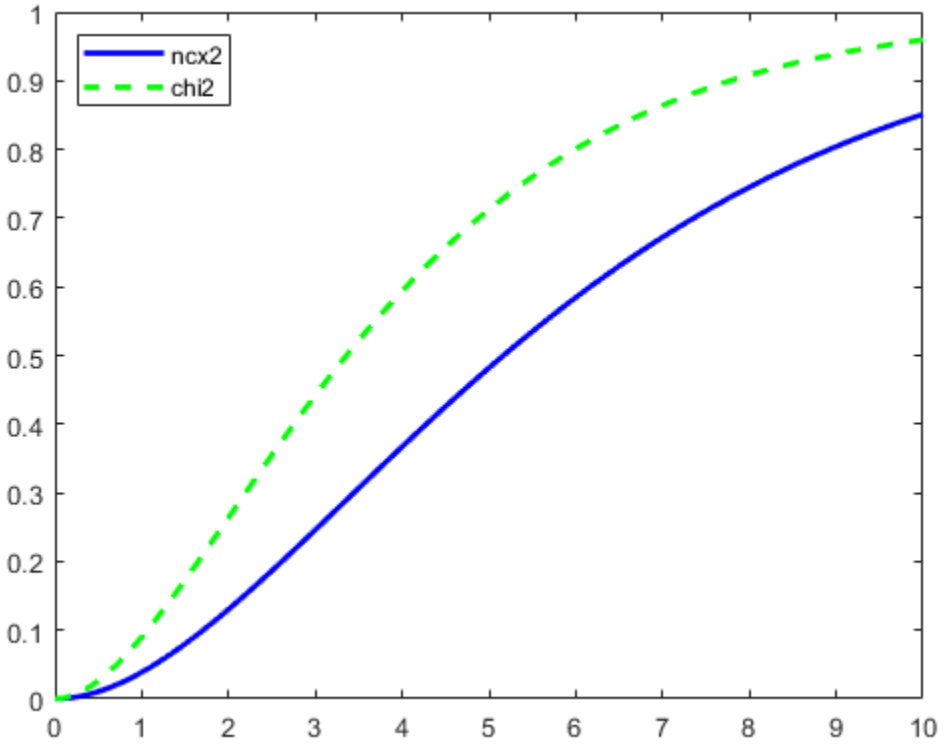
$$F(x|\nu,\delta) = \sum_{j=0}^{\infty}\left(\frac{\left(\frac{1}{2}\delta\right)^j}{j!}e^{\frac{-\delta}{2}}\right)\Pr\left[\chi^2_{\nu+2j} \le x\right]$$

Examples

Compute Noncentral Chi-Square cdf

Compare the noncentral chi-square cdf with `DELTA = 2` to the chi-square cdf with the same number of degrees of freedom (4):

```
x = (0:0.1:10)';
ncx2 = ncx2cdf(x,4,2);
chi2 = chi2cdf(x,4);

plot(x,ncx2,'b-','LineWidth',2)
hold on
plot(x,chi2,'g--','LineWidth',2)
legend('ncx2','chi2','Location','NW')
```



References

[1] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `ncx2inv` | `ncx2pdf` | `ncx2rnd` | `ncx2stat`

Topics

“Noncentral Chi-Square Distribution” on page B-113

Introduced before R2006a

ncx2inv

Noncentral chi-square inverse cumulative distribution function

Syntax

```
X = ncx2inv(P,V,DELTA)
```

Description

`X = ncx2inv(P,V,DELTA)` returns the inverse of the noncentral chi-square cdf using the corresponding degrees of freedom in `V` and positive noncentrality parameters in `DELTA`, at the corresponding probabilities in `P`. `P`, `V`, and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `X`. A scalar input for `P`, `V`, or `DELTA` is expanded to a constant array with the same dimensions as the other inputs.

Examples

```
ncx2inv([0.01 0.05 0.1],4,2)
ans =
  0.4858  1.1498  1.7066
```

Algorithms

`ncx2inv` uses Newton's method to converge to the solution.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 50–52.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148.

See Also

`icdf` | `ncx2cdf` | `ncx2pdf` | `ncx2rnd` | `ncx2stat`

Topics

“Noncentral Chi-Square Distribution” on page B-113

Introduced before R2006a

ncx2pdf

Noncentral chi-square probability density function

Syntax

```
Y = ncx2pdf(X,V,DELTA)
```

Description

`Y = ncx2pdf(X,V,DELTA)` computes the noncentral chi-square pdf at each of the values in `X` using the corresponding degrees of freedom in `V` and positive noncentrality parameters in `DELTA`. Vector or matrix inputs for `X`, `V`, and `DELTA` must have the same size, which is also the size of `Y`. A scalar input for `X`, `V`, or `DELTA` is expanded to a constant array with the same dimensions as the other inputs.

Some texts refer to this distribution as the generalized Rayleigh, Rayleigh-Rice, or Rice distribution.

Examples

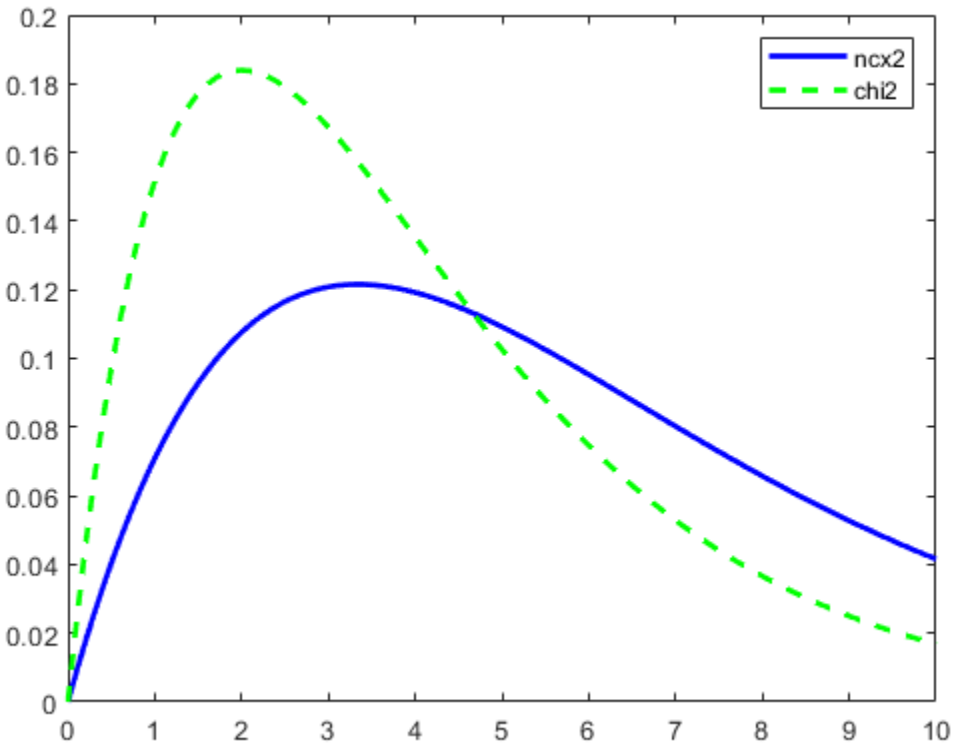
Compute Noncentral Chi-Square Distribution pdf

Compute the pdf of a noncentral chi-square distribution with degrees of freedom `V = 4` and noncentrality parameter `DELTA = 2`. For comparison, also compute the pdf of a chi-square distribution with the same degrees of freedom.

```
x = (0:0.1:10)';
ncx2 = ncx2pdf(x,4,2);
chi2 = chi2pdf(x,4);
```

Plot the pdf of the noncentral chi-square distribution on the same figure as the pdf of the chi-square distribution.

```
figure;
plot(x,ncx2,'b-','LineWidth',2)
hold on
plot(x,chi2,'g--','LineWidth',2)
legend('ncx2','chi2')
```



References

[1] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148.

See Also

`ncx2cdf` | `ncx2inv` | `ncx2rnd` | `ncx2stat` | `pdf`

Topics

“Noncentral Chi-Square Distribution” on page B-113

Introduced before R2006a

ncx2rnd

Noncentral chi-square random numbers

Syntax

```
R = ncx2rnd(V,DELTA)
R = ncx2rnd(V,DELTA,m,n,...)
R = ncx2rnd(V,DELTA,[m,n,...])
```

Description

`R = ncx2rnd(V,DELTA)` returns a matrix of random numbers chosen from the noncentral chi-square distribution using the corresponding degrees of freedom in `V` and positive noncentrality parameters in `DELTA`. `V` and `DELTA` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `R`. A scalar input for `V` or `DELTA` is expanded to a constant array with the same dimensions as the other input.

`R = ncx2rnd(V,DELTA,m,n,...)` or `R = ncx2rnd(V,DELTA,[m,n,...])` generates an `m`-by-`n`-by-... array. The `V`, `DELTA` parameters can each be scalars or arrays of the same size as `R`.

Examples

```
ncx2rnd(4,2,6,3)
ans =
  6.8552  5.9650  11.2961
  5.2631  4.2640  5.9495
  9.1939  6.7162  3.8315
 10.3100  4.4828  7.1653
  2.1142  1.9826  4.6400
  3.8852  5.3999  0.9282
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 50–52.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2.* Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ncx2cdf` | `ncx2inv` | `ncx2pdf` | `ncx2stat` | `random`

Topics

“Noncentral Chi-Square Distribution” on page B-113

Introduced before R2006a

ncx2stat

Noncentral chi-square mean and variance

Syntax

```
[M,V] = ncx2stat(NU,DELTA)
```

Description

`[M,V] = ncx2stat(NU,DELTA)` returns the mean of and variance for the noncentral chi-square pdf with `NU` degrees of freedom and noncentrality parameter `DELTA`. `NU` and `DELTA` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `M` and `V`. A scalar input for `NU` or `DELTA` is expanded to a constant array with the same dimensions as the other input.

The mean of the noncentral chi-square distribution with parameters ν and δ is $\nu+\delta$, and the variance is $2(\nu+2\delta)$.

Examples

```
[m,v] = ncx2stat(4,2)
m =
   6
v =
  16
```

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 50–52.

[2] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`ncx2cdf` | `ncx2inv` | `ncx2pdf` | `ncx2rnd`

Topics

“Noncentral Chi-Square Distribution” on page B-113

Introduced before R2006a

ndims

Class: `dataset`

(Not Recommended) Number of dimensions of dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
n = ndims(A)
```

Description

`n = ndims(A)` returns the number of dimensions in the dataset `A`. The number of dimensions in an array is always 2.

See Also

`size`

ne

Class: `qrandstream`

Not equal relation for handles

Syntax

```
h1 ~= h2
```

Description

Handles are equal if they are handles for the same object and are unequal otherwise.

`h1 ~= h2` performs element-wise comparisons between handle arrays `h1` and `h2`. `h1` and `h2` must be of the same dimensions unless one is a scalar. The result is a logical array of the same dimensions, where each element is an element-wise ~= result.

If one of `h1` or `h2` is scalar, scalar expansion is performed and the result will match the dimensions of the array that is not scalar.

`tf = ne(h1, h2)` stores the result in a logical array of the same dimensions.

See Also

`eq` | `ge` | `gt` | `le` | `lt` | `qrandstream`

negloglik

Package: prob

Negative loglikelihood of probability distribution

Syntax

```
nll = negloglik(pd)
```

Description

`nll = negloglik(pd)` returns the value of the negative loglikelihood function for the data used to fit the probability distribution `pd`.

Examples

Negative Log Likelihood for a Fitted Distribution

Load the sample data.

```
load carsmall
```

Create a Weibull distribution object by fitting it to the mile per gallon (`MPG`) data.

```
pd = fitdist(MPG,'Weibull')
```

```
pd =
  WeibullDistribution

  Weibull distribution
    A = 26.5079   [24.8333, 28.2954]
    B = 3.27193   [2.79441, 3.83104]
```

Compute the negative log likelihood for the fitted Weibull distribution.

```
wnll = negloglik(pd)
```

```
wnll = 327.4942
```

Negative Loglikelihood for a Kernel Distribution

Load the sample data. Fit a kernel distribution to the miles per gallon (`MPG`) data.

```
load carsmall;
pd = fitdist(MPG,'Kernel')
```

```
pd =
  KernelDistribution

    Kernel = normal
```

```
    Bandwidth = 4.11428
    Support = unbounded
```

Compute the negative loglikelihood.

```
nll = negloglik(pd)
```

```
nll = 327.3139
```

Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

nll — Negative loglikelihood

numeric value

Negative loglikelihood value for the data used to fit the distribution, returned as a numeric value.

See Also

Distribution Fitter | `fitdist` | `mle` | `paramci` | `proflik`

Topics

“Negative Loglikelihood Functions” on page 5-24
“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

net

Generate quasirandom point set

Syntax

```
X = net(p,n)
```

Description

`X = net(p,n)` returns the first `n` points from the point set `p`, which is either a `haltonset` or `sobolset` object. `X` is an `n`-by-*d* matrix, where *d* is the number of dimensions of the points in `p`.

The object `p` encapsulates properties of a specified quasirandom sequence. Values of the point set are generated whenever you access `p` using `net` or parenthesis indexing. Values are not stored within `p`.

Examples

Create Halton Point Set

Generate a three-dimensional Halton point set, skip the first 1000 values, and then retain every 101st point.

```
p = haltonset(3,'Skip',1e3,'Leap',1e2)
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : none
```

Apply reverse-radix scrambling by using `scramble`.

```
p = scramble(p,'RR2')
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : RR2
```

Generate the first four points by using `net`.

```
X0 = net(p,4)
```

```
X0 = 4×3
```

```
    0.0928    0.6950    0.0029
    0.6958    0.2958    0.8269
    0.3013    0.6497    0.4141
    0.9087    0.7883    0.2166
```

Generate every third point, up to the eleventh point, by using parenthesis indexing.

```
X = p(1:3:11,:)
```

```
X = 4×3
```

```
    0.0928    0.6950    0.0029
    0.9087    0.7883    0.2166
    0.3843    0.9840    0.9878
    0.6831    0.7357    0.7923
```

Input Arguments

p — Point set

`haltonset` object | `sobolset` object

Point set, specified as either a `haltonset` or `sobolset` object.

Example: `haltonset(4)`

n — Number of points to return

positive integer scalar

Number of points to return from the point set, specified as a positive integer scalar. `n` must be between `1` and `length(p)`, the number of points in `p`.

`net` always returns the first `n` points in `p`. To select a different set of `n` points from the quasirandom sequence, you can change `p` by using its `Leap` and `Skip` properties or the `scramble` object function. Alternatively, you can access points in `p` by using parenthesis indexing rather than the `net` object function.

Example: `1024`

Data Types: `single` | `double`

See Also

`haltonset` | `qrandstream` | `reduceDimensions` | `scramble` | `sobolset`

Introduced in R2008a

nLinearCoeffs

Number of nonzero linear coefficients

Syntax

```
ncoeffs = nLinearCoeffs(obj)
ncoeffs = nLinearCoeffs(obj,delta)
```

Description

`ncoeffs = nLinearCoeffs(obj)` returns the number of nonzero linear coefficients in the linear discriminant model `obj`.

`ncoeffs = nLinearCoeffs(obj,delta)` returns the number of nonzero linear coefficients for threshold parameter `delta`.

Input Arguments

`obj`

Discriminant analysis classifier, produced using `fitcdiscr`.

`delta`

Scalar or vector value of the `Delta` parameter. See “Gamma and Delta” on page 33-4197.

Output Arguments

`ncoeffs`

Nonnegative integer, the number of nonzero coefficients in the discriminant analysis model `obj`.

If you call `nLinearCoeffs` with a `delta` argument, `ncoeffs` is the number of nonzero linear coefficients for threshold parameter `delta`. If `delta` is a vector, `ncoeffs` is a vector with the same number of elements.

If `obj` is a quadratic discriminant model, `ncoeffs` is the number of predictors in `obj`.

Examples

Find the Number of Nonzero Coefficients in a Discriminant Analysis Classifier

Find the number of nonzero coefficients in a discriminant analysis classifier for various `Delta` values.

Create a discriminant analysis classifier from the `fishseriris` data.

```
load fisheriris
obj = fitcdiscr(meas,species);
```

Find the number of nonzero coefficients in `obj`.

```
ncoeffs = nLinearCoeffs(obj)
```

```
ncoeffs = 4
```

Find the number of nonzero coefficients for `delta` = 1, 2, 4, and 8.

```
delta = [1 2 4 8];
ncoeffs = nLinearCoeffs(obj,delta)
```

```
ncoeffs = 4×1

     4
     4
     3
     0
```

The `DeltaPredictor` property gives the values of `delta` where the number of nonzero coefficients changes.

```
ncoeffs2 = nLinearCoeffs(obj,obj.DeltaPredictor)
```

```
ncoeffs2 = 4×1

     4
     3
     1
     2
```

More About

Gamma and Delta

Regularization is the process of finding a small set of predictors that yield an effective predictive model. For linear discriminant analysis, there are two parameters, γ and δ, that control regularization as follows. `cvshrink` helps you select appropriate values of the parameters.

Let Σ represent the covariance matrix of the data X, and let $\widehat{X}$ be the centered data (the data X minus the mean by class). Define

$$D = \mathrm{diag}\left(\widehat{X}^T * \widehat{X}\right).$$

The regularized covariance matrix $\tilde{\Sigma}$ is

$$\tilde{\Sigma} = (1 - \gamma)\Sigma + \gamma D.$$

Whenever $\gamma \geq$ `MinGamma`, $\tilde{\Sigma}$ is nonsingular.

Let μ_k be the mean vector for those elements of X in class k, and let μ_0 be the global mean vector (the mean of the rows of X). Let C be the correlation matrix of the data X, and let $\tilde{C}$ be the regularized correlation matrix:

$$\tilde{C} = (1 - \gamma)C + \gamma I,$$

where I is the identity matrix.

The linear term in the regularized discriminant analysis classifier for a data point x is

$$(x - \mu_0)^T \tilde{\Sigma}^{-1}(\mu_k - \mu_0) = \left[(x - \mu_0)^T D^{-1/2}\right]\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right].$$

The parameter δ enters into this equation as a threshold on the final term in square brackets. Each component of the vector $\left[\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right]$ is set to zero if it is smaller in magnitude than the threshold δ. Therefore, for class k, if component j is thresholded to zero, component j of x does not enter into the evaluation of the posterior probability.

The `DeltaPredictor` property is a vector related to this threshold. When $\delta \geq$ `DeltaPredictor(i)`, all classes k have

$$\left|\tilde{C}^{-1} D^{-1/2}(\mu_k - \mu_0)\right| \leq \delta\,.$$

Therefore, when $\delta \geq$ `DeltaPredictor(i)`, the regularized classifier does not use predictor `i`.

See Also

`CompactClassificationDiscriminant` | `cvshrink` | `fitcdiscr`

Topics

“Discriminant Analysis Classification” on page 20-2

nlinfit

Nonlinear regression

Syntax

```
beta = nlinfit(X,Y,modelfun,beta0)
beta = nlinfit(X,Y,modelfun,beta0,options)
beta = nlinfit(___,Name,Value)
[beta,R,J,CovB,MSE,ErrorModelInfo] = nlinfit(___)
```

Description

`beta = nlinfit(X,Y,modelfun,beta0)` returns a vector of estimated coefficients for the nonlinear regression of the responses in `Y` on the predictors in `X` using the model specified by `modelfun`. The coefficients are estimated using iterative least squares estimation, with initial values specified by `beta0`.

`beta = nlinfit(X,Y,modelfun,beta0,options)` fits the nonlinear regression using the algorithm control parameters in the structure `options`. You can return any of the output arguments in the previous syntaxes.

`beta = nlinfit(___,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify observation weights or a nonconstant error model. You can use any of the input arguments in the previous syntaxes.

`[beta,R,J,CovB,MSE,ErrorModelInfo] = nlinfit(___)` additionally returns the residuals, `R`, the Jacobian of `modelfun`, `J`, the estimated variance-covariance matrix for the estimated coefficients, `CovB`, an estimate of the variance of the error term, `MSE`, and a structure containing details about the error model, `ErrorModelInfo`.

Examples

Nonlinear Regression Model Using Default Options

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the initial values in `beta0`.

```
beta = nlinfit(X,y,@hougen,beta0)
```

```
beta = 5×1

    1.2526
    0.0628
    0.0400
```

```
    0.1124
    1.1914
```

Nonlinear Regression Using Robust Options

Generate sample data from the nonlinear regression model $y = b_1 + b_2 \cdot exp\{ - b_3 x\} + \epsilon$, where b_1, b_2, and b_3 are coefficients, and the error term is normally distributed with mean 0 and standard deviation 0.1.

```
modelfun = @(b,x)(b(1)+b(2)*exp(-b(3)*x));

rng('default') % for reproducibility
b = [1;3;2];
x = exprnd(2,100,1);
y = modelfun(b,x) + normrnd(0,0.1,100,1);
```

Set robust fitting options.

```
opts = statset('nlinfit');
opts.RobustWgtFun = 'bisquare';
```

Fit the nonlinear model using the robust fitting options.

```
beta0 = [2;2;2];
beta = nlinfit(x,y,modelfun,beta0,opts)
```

```
beta = 3×1

    1.0041
    3.0997
    2.1483
```

Nonlinear Regression Using Observation Weights

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Specify a vector of known observation weights.

```
W = [8 2 1 6 12 9 12 10 10 12 2 10 8]';
```

Fit the Hougen-Watson model to the rate data using the specified observation weights.

```
[beta,R,J,CovB] = nlinfit(X,y,@hougen,beta0,'Weights',W);
beta
```

```
beta = 5×1
```

```
    2.2068
    0.1077
    0.0766
    0.1818
    0.6516
```

Display the coefficient standard errors.

```
sqrt(diag(CovB))
```

```
ans = 5×1

    2.5721
    0.1251
    0.0950
    0.2043
    0.7735
```

Nonlinear Regression Using Weights Function Handle

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Specify a function handle for observation weights. The function accepts the model fitted values as input, and returns a vector of weights.

```
 a = 1; b = 1;
 weights = @(yhat) 1./((a + b*abs(yhat)).^2);
```

Fit the Hougen-Watson model to the rate data using the specified observation weights function.

```
[beta,R,J,CovB] = nlinfit(X,y,@hougen,beta0,'Weights',weights);
beta
```

```
beta = 5×1

    0.8308
    0.0409
    0.0251
    0.0801
    1.8261
```

Display the coefficient standard errors.

```
sqrt(diag(CovB))
```

```
ans = 5×1

    0.5822
```

```
    0.0297
    0.0197
    0.0578
    1.2810
```

Nonlinear Regression Using Nonconstant Error Model

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the combined error model.

```
[beta,R,J,CovB,MSE,ErrorModelInfo] = nlinfit(X,y,@hougen,beta0,'ErrorModel','combined');
beta
```

```
beta = 5×1

    1.2526
    0.0628
    0.0400
    0.1124
    1.1914
```

Display the error model information.

```
ErrorModelInfo
```

```
ErrorModelInfo = struct with fields:
              ErrorModel: 'combined'
         ErrorParameters: [0.1517 5.6783e-08]
           ErrorVariance: [function_handle]
                     MSE: 1.6245
          ScheffeSimPred: 6
          WeightFunction: 0
            FixedWeights: 0
    RobustWeightFunction: 0
```

Input Arguments

X — Predictor variables

matrix

Predictor variables for the nonlinear regression function, specified as a matrix. Typically, `X` is a design matrix of predictor (independent variable) values, with one row for each value in `Y`, and one column for each predictor. However, `X` can be any array that `modelfun` can accept.

Data Types: `single` | `double`

Y — Response values
vector

Response values (dependent variable) for fitting the nonlinear regression function, specified as a vector with the same number of rows as `X`.

Data Types: `single` | `double`

`modelfun` — Nonlinear regression model function
function handle

Nonlinear regression model function, specified as a function handle. `modelfun` must accept two input arguments, a coefficient vector and an array `X`—in that order—and return a vector of fitted response values.

For example, to specify the `hougen` nonlinear regression function, use the function handle `@hougen`.

Data Types: `function_handle`

`beta0` — Initial coefficient values
vector

Initial coefficient values for the least squares estimation algorithm, specified as a vector.

Note Poor starting values can lead to a solution with large residual error.

Data Types: `single` | `double`

`options` — Estimation algorithm options
structure created using `statset`

Estimation algorithm options, specified as a structure you create using `statset`. The following `statset` parameters are applicable to `nlinfit`.

`DerivStep` — Relative difference for finite difference gradient
`eps^(1/3)` (default) | positive scalar value | vector

Relative difference for the finite difference gradient calculation, specified as a positive scalar value, or a vector the same size as `beta`. Use a vector to specify a different relative difference for each coefficient.

`Display` — Level of output display
`'off'` (default) | `'iter'` | `'final'`

Level of output display during estimation, specified as one of `'off'`, `'iter'`, or `'final'`. If you specify `'iter'`, output is displayed at each iteration. If you specify `'final'`, output is displayed after the final iteration.

`FunValCheck` — Indicator for whether to check for invalid values
`'on'` (default) | `'off'`

Indicator for whether to check for invalid values such as `NaN` or `Inf` from the objective function, specified as `'on'` or `'off'`.

`MaxIter` — Maximum number of iterations
`100` (default) | positive integer

Maximum number of iterations for the estimation algorithm, specified as a positive integer. Iterations continue until estimates are within the convergence tolerance, or the maximum number of iterations specified by `MaxIter` is reached.

`RobustWgtFun` — Weight function
character vector | string scalar | function handle | `[]`

Weight function for robust fitting, specified as a valid character vector, string scalar, or function handle.

Note `RobustWgtFun` must have value `[]` when you use observation weights, `W`.

The following table describes the possible character vectors and string scalars. Let r denote normalized residuals and w denote robust weights. The indicator function I[x] is equal to 1 if the expression x is true, and 0 otherwise.

Weight Function	Equation	Default Tuning Constant
`''` (default)	No robust fitting	—
`'andrews'`	$w = I[\lvert r\rvert < \pi] \times \sin(r)/r$	1.339
`'bisquare'`	$w = I[\lvert r\rvert < 1] \times \left(1 - r^2\right)^2$	4.685
`'cauchy'`	$w = 1/\left(1 + r^2\right)$	2.385
`'fair'`	$w = 1/(1 + \lvert r\rvert)$	1.400
`'huber'`	$w = 1/\max(1, \lvert r\rvert)$	1.345
`'logistic'`	$w = \tanh(r)/r$	1.205
`'talwar'`	$w = I[\lvert r\rvert < 1]$	2.795
`'welsch'`	$w = \exp\{-r^2\}$	2.985

You can alternatively specify a function handle that accepts a vector of normalized residuals as input, and returns a vector of robust weights as output. If you use a function handle, you must provide a `Tune` constant.

`Tune` — Tuning constant
positive scalar value

Tuning constant for robust fitting, specified as a positive scalar value. The tuning constant is used to normalize residuals before applying a robust weight function. The default tuning constant depends on the function specified by `RobustWgtFun`.

If you use a function handle to specify `RobustWgtFun`, then you must specify a value for `Tune`.

`TolFun` — Termination tolerance on residual sum of squares
`1e-8` (default) | positive scalar value

Termination tolerance for the residual sum of squares, specified as a positive scalar value. Iterations continue until estimates are within the convergence tolerance, or the maximum number of iterations specified by `MaxIter` is reached.

`TolX` — Termination tolerance on estimated coefficients
`1e-8` (default) | positive scalar value

Termination tolerance on the estimated coefficients, `beta`, specified as a positive scalar value. Iterations continue until estimates are within the convergence tolerance, or the maximum number of iterations specified by `MaxIter` is reached.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ErrorModel','proportional','ErrorParameters',0.5` specifies a proportional error model, with initial value 0.5 for the error parameter estimation

`ErrorModel` — Form of error term
`'constant'` (default) | `'proportional'` | `'combined'`

Form of the error term, specified as the comma-separated pair consisting of `'ErrorModel'` and `'constant'`, `'proportional'`, or `'combined'` indicating the error model. Each model defines the error using a standard mean-zero and unit-variance variable e in combination with independent components: the function value f, and one or two parameters a and b.

`'constant'` (default)	$y = f + ae$
`'proportional'`	$y = f + bfe$
`'combined'`	$y = f + (a + b\|f\|)e$

The only allowed error model when using `Weights` is `'constant'`.

Note `options.RobustWgtFun` must have value `[]` when using an error model other than `'constant'`.

`ErrorParameters` — Initial estimates for error model parameters
`1` or `[1,1]` (default) | scalar value | two-element vector

Initial estimates for the error model parameters in the chosen `ErrorModel`, specified as the comma-separated pair consisting of `'ErrorParameters'` and a scalar value or two-element vector.

Error Model	Parameters	Default Values
`'constant'`	a	1
`'proportional'`	b	1
`'combined'`	a, b	`[1,1]`

For example, if `'ErrorModel'` has the value `'combined'`, you can specify the starting value 1 for a and the starting value 2 for b as follows.

Example: `'ErrorParameters',[1,2]`

You can only use the `'constant'` error model when using `Weights`.

Note `options.RobustWgtFun` must have value `[]` when using an error model other than `'constant'`.

Data Types: `double` | `single`

`Weights` — Observation weights
vector | function handle

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of real positive weights or a function handle. You can use observation weights to down-weight the observations that you want to have less influence on the fitted model.

- If `W` is a vector, then it must be the same size as `Y`.
- If `W` is a function handle, then it must accept a vector of predicted response values as input, and return a vector of real positive weights as output.

Note `options.RobustWgtFun` must have value `[]` when you use observation weights.

Data Types: `double` | `single` | `function_handle`

Output Arguments

`beta` — Estimated regression coefficients
vector

Estimated regression coefficients, returned as a vector. The number of elements in `beta` equals the number of elements in `beta0`.

Let $f(X_i, \mathbf{b})$ denote the nonlinear function specified by `modelfun`, where $\mathbf{x}_i$ are the predictors for observation i, $i = 1,...,N$, and $\mathbf{b}$ are the regression coefficients. The vector of coefficients returned in `beta` minimizes the weighted least squares equation,

$$\sum_{i=1}^{N} w_i [y_i - f(\mathbf{x}_i, \mathbf{b})]^2 .$$

For unweighted nonlinear regression, all of the weight terms are equal to 1.

`R` — Residuals
vector

Residuals for the fitted model, returned as a vector.

- If you specify observation weights using the name-value pair argument `Weights`, then `R` contains weighted residuals on page 33-4207.
- If you specify an error model other than `'constant'` using the name-value pair argument `ErrorModel`, then you can no longer interpret `R` as model fit residuals.

`J` — Jacobian
matrix

Jacobian of the nonlinear regression model, `modelfun`, returned as an *N*-by-*p* matrix, where *N* is the number of observations and *p* is the number of estimated coefficients.

- If you specify observation weights using the name-value pair argument `Weights`, then `J` is the weighted model function Jacobian on page 33-4208.
- If you specify an error model other than `'constant'` using the name-value pair argument `ErrorModel`, then you can no longer interpret `J` as the model function Jacobian.

`CovB` — Estimated variance-covariance matrix
matrix

Estimated variance-covariance matrix for the fitted coefficients, `beta`, returned as a *p*-by-*p* matrix, where *p* is the number of estimated coefficients. If the model Jacobian, `J`, has full column rank, then `CovB = inv(J'*J)*MSE`, where `MSE` is the mean squared error.

`MSE` — Mean squared error
scalar value

Mean squared error (MSE) of the fitted model, returned as a scalar value. MSE is an estimate of the variance of the error term. If the model Jacobian, `J`, has full column rank, then `MSE = (R'*R)/(N-p)`, where `N` is the number of observations, and `p` is the number of estimated coefficients.

`ErrorModelInfo` — Information about error model fit
structure

Information about the error model fit, returned as a structure with the following fields:

`ErrorModel`	Chosen error model
`ErrorParameters`	Estimated error parameters
`ErrorVariance`	Function handle that accepts an *N*-by-*p* matrix, `X`, and returns an *N*-by-1 vector of error variances using the estimated error model
`MSE`	Mean squared error
`ScheffeSimPred`	Scheffé parameter for simultaneous prediction intervals when using the estimated error model
`WeightFunction`	Logical with value `true` if you used a custom weight function previously in `nlinfit`
`FixedWeights`	Logical with value `true` if you used fixed weights previously in `nlinfit`
`RobustWeightFunction`	Logical with value `true` if you used robust fitting previously in `nlinfit`

More About

Weighted Residuals

A weighted residual is a residual multiplied by the square root of the corresponding observation weight.

Given estimated regression coefficients, **b**, the residual for observation *i* is

$$r_i = y_i - f(\mathbf{x}_i, \mathbf{b}),$$

where y_i is the observed response and $f(\mathbf{x}_i, \mathbf{b})$ is the fitted response at predictors $\mathbf{x}_i$.

When you fit a weighted nonlinear regression with weights w_i, $i = 1,...,N$, `nlinfit` returns the weighted residuals,

$$r_i^* = \sqrt{w_i}(y_i - f(\mathbf{x}_i, \mathbf{b})).$$

Weighted Model Function Jacobian

The weighted model function Jacobian is the nonlinear model Jacobian multiplied by the square root of the observation weight matrix.

Given estimated regression coefficients, **b**, the estimated model Jacobian, **J**,for the nonlinear function $f(\mathbf{x}_i, \mathbf{b})$ has elements

$$\mathbf{J}_{ij} = \frac{\partial f(\mathbf{x}_i, \mathbf{b})}{\partial b_j},$$

where b_j is the jth element of **b**.

When you fit a weighted nonlinear regression with diagonal weights matrix **W**,`nlinfit` returns the weighted Jacobian matrix,

$$\mathbf{J}^* = \mathbf{W}^{1/2}\mathbf{J}.$$

Tips

- To produce error estimates on predictions, use the optional output arguments `R`, `J`, `CovB`, or `MSE` as inputs to `nlpredci`.
- To produce error estimates on the estimated coefficients, `beta`, use the optional output arguments `R`, `J`, `CovB`, or `MSE` as inputs to `nlparci`.
- If you use the robust fitting option, `RobustWgtFun`, you must use `CovB`—and might need `MSE`—as inputs to `nlpredci` or `nlparci` to ensure that the confidence intervals take the robust fit properly into account.

Algorithms

- `nlinfit` treats `NaN` values in `Y` or `modelfun(beta0,X)` as missing data, and ignores the corresponding observations.
- For nonrobust estimation, `nlinfit` uses the Levenberg-Marquardt nonlinear least squares algorithm [1].
- For robust estimation, `nlinfit` uses the algorithm of “Iteratively Reweighted Least Squares” on page 11-104 ([2], [3]). At each iteration, the robust weights are recalculated based on each observation’s residual from the previous iteration. These weights downweight outliers, so that their influence on the fit is decreased. Iterations continue until the weights converge.
- When you specify a function handle for observation weights, the weights depend on the fitted model. In this case, `nlinfit` uses an iterative generalized least squares algorithm to fit the nonlinear regression model.

References

[1] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

[2] DuMouchel, W. H., and F. L. O'Brien. “Integrating a Robust Option into a Multiple Regression Computing Environment.” *Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[3] Holland, P. W., and R. E. Welsch. “Robust Regression Using Iteratively Reweighted Least-Squares.” *Communications in Statistics: Theory and Methods*, *A6*, 1977, pp. 813–827.

See Also

`fitnlm` | `nlintool` | `nlparci` | `nlpredci`

Topics

“Nonlinear Regression” on page 13-2

Introduced before R2006a

nlintool

Interactive nonlinear regression

Syntax

```
nlintool(X,y,fun,beta0)
nlintool(X,y,fun,beta0,alpha)
nlintool(X,y,fun,beta0,alpha,'xname','yname')
```

Description

`nlintool(X,y,fun,beta0)` is a graphical user interface to the `nlinfit` function, and uses the same input arguments. The interface displays plots of the fitted response against each predictor, with the other predictors held fixed. The fixed values are in the text boxes below each predictor axis. Change the fixed values by typing in a new value or by dragging the vertical lines in the plots to new positions. When you change the value of a predictor, all plots update to display the model at the new point in predictor space. Dashed red curves show 95% simultaneous confidence bands for the function.

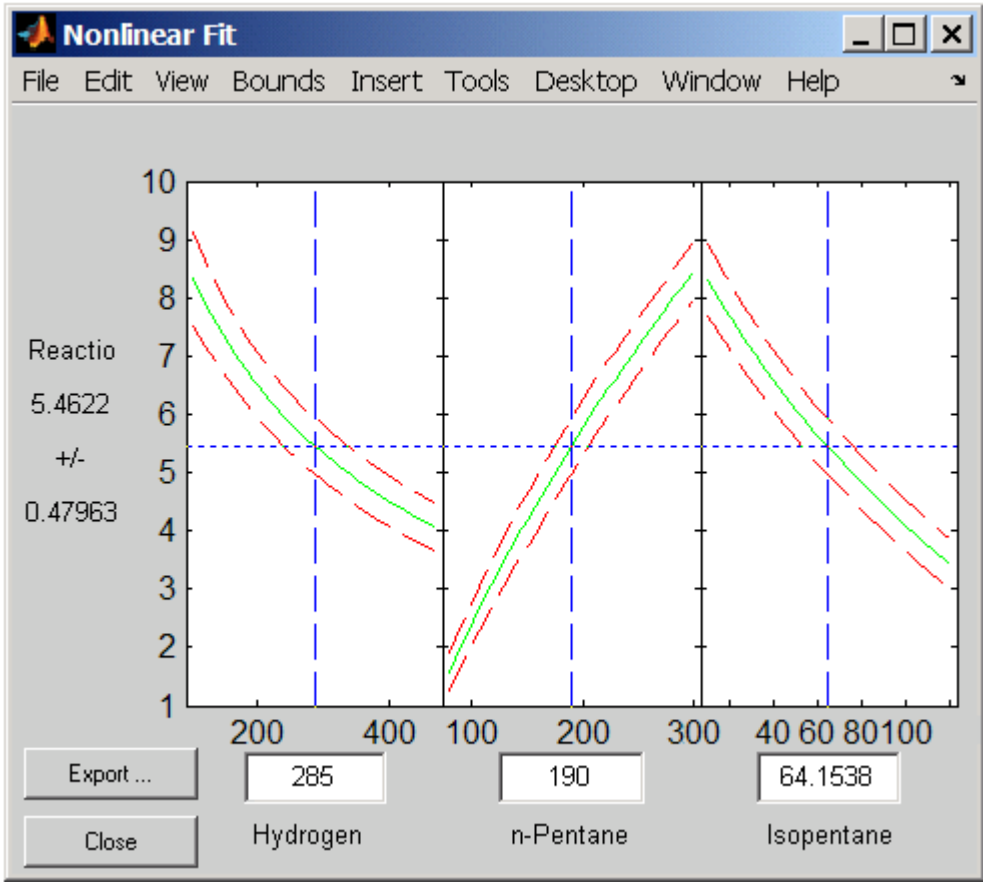
`nlintool(X,y,fun,beta0,alpha)` shows `100(1-alpha)`% confidence bands. These are simultaneous confidence bounds for the function value. Using the **Bounds** menu you can switch between simultaneous and non-simultaneous bounds, and between bounds on the function and bounds for predicting a new observation.

`nlintool(X,y,fun,beta0,alpha,'xname','yname')` labels the plots using the character matrix or string array `'xname'` for the predictors and the character vector or string scalar `'yname'` for the response.

Examples

The data in `reaction.mat` are partial pressures of three chemical reactants and the corresponding reaction rates. The function `hougen` implements the nonlinear Hougen-Watson model for reaction rates. The following fits the model to the data:

```
load reaction
nlintool(reactants,rate,@hougen,beta,0.01,xn,yn)
```



See Also

`nlinfit` | `polytool` | `rstool`

Introduced before R2006a

nlmefit

Nonlinear mixed-effects estimation

Syntax

```
beta = nlmefit(X,y,group,V,fun,beta0)
[beta,PSI] = nlmefit(X,y,group,V,fun,beta0)
[beta,PSI,stats] = nlmefit(X,y,group,V,fun,beta0)
[beta,PSI,stats,B] = nlmefit(X,y,group,V,fun,beta0)
[beta,PSI,stats,B] = nlmefit(X,y,group,V,fun,beta0,'Name',value)
```

Description

`beta = nlmefit(X,y,group,V,fun,beta0)` fits a nonlinear mixed-effects regression model and returns estimates of the fixed effects in `beta`. By default, `nlmefit` fits a model in which each parameter is the sum of a fixed and a random effect, and the random effects are uncorrelated (their covariance matrix is diagonal).

`X` is an *n*-by-*h* matrix of *n* observations on *h* predictors.

`y` is an *n*-by-1 vector of responses.

`group` is a grouping variable indicating *m* groups in the observations. `group` is a categorical variable, a numeric vector, a character matrix with rows for group names, a string array, or a cell array of character vectors. For more information on grouping variables, see “Grouping Variables” on page 2-45.

`V` is an *m*-by-*g* matrix or cell array of *g* group-specific predictors. These are predictors that take the same value for all observations in a group. The rows of `V` are assigned to groups using `grp2idx`, according to the order specified by `grp2idx(group)`. Use a cell array for `V` if group predictors vary in size across groups. Use `[]` for `V` if there are no group-specific predictors.

`fun` is a handle to a function that accepts predictor values and model parameters and returns fitted values. `fun` has the form

```
yfit = modelfun(PHI,XFUN,VFUN)
```

The arguments are:

- `PHI` — A 1-by-*p* vector of model parameters.
- `XFUN` — A *k*-by-*h* array of predictors, where:
  - *k* = 1 if `XFUN` is a single row of `X`.
  - *k* = n_i if `XFUN` contains the rows of `X` for a single group of size n_i.
  - *k* = *n* if `XFUN` contains all rows of `X`.
- `VFUN` — Group-specific predictors given by one of:
  - A 1-by-*g* vector corresponding to a single group and a single row of `V`.
  - An *n*-by-*g* array, where the *j*th row is V(`I`,:) if the *j*th observation is in group `I`.

If `V` is empty, `nlmefit` calls `modelfun` with only two inputs.

- `yfit` — A *k*-by-1 vector of fitted values

When either `PHI` or `VFUN` contains a single row, it corresponds to all rows in the other two input arguments.

Note If `modelfun` can compute `yfit` for more than one vector of model parameters per call, use the `'Vectorization'` parameter (described later) for improved performance.

`beta0` is a *q*-by-1 vector with initial estimates for *q* fixed effects. By default, *q* is the number of model parameters *p*.

`nlmefit` fits the model by maximizing an approximation to the marginal likelihood with random effects integrated out, assuming that:

- Random effects are multivariate normally distributed and independent between groups.
- Observation errors are independent, identically normally distributed, and independent of the random effects.

`[beta,PSI] = nlmefit(X,y,group,V,fun,beta0)` also returns `PSI`, an *r*-by-*r* estimated covariance matrix for the random effects. By default, *r* is equal to the number of model parameters *p*.

`[beta,PSI,stats] = nlmefit(X,y,group,V,fun,beta0)` also returns `stats`, a structure with fields:

- `dfe` — The error degrees of freedom for the model
- `logl` — The maximized loglikelihood for the fitted model
- `rmse` — The square root of the estimated error variance (computed on the log scale for the `exponential` error model)
- `errorparam` — The estimated parameters of the error variance model
- `aic` — The Akaike information criterion, calculated as `aic` = -2 * `logl` + 2 * `numParam`, where `numParam` is the number of fitting parameters, including the degree of freedom for covariance matrix of the random effects, the number of fixed effects and the number of parameters of the error model, and `logl` is a field in the `stats` structure
- `bic` — The Bayesian information criterion, calculated as `bic` = –2*`logl` + log(`M`) * `numParam`
  - `M` is the number of groups.
  - `numParam` and `logl` are defined as in `aic`.

  Note that some literature suggests that the computation of `bic` should be , `bic` = –2*`logl` + log(`N`) * `numParam`, where `N` is the number of observations.
- `covb` — The estimated covariance matrix of the parameter estimates
- `sebeta` — The standard errors for `beta`
- `ires` — The population residuals `(y-y_population)`, where `y_population` is the individual predicted values
- `pres` — The population residuals `(y-y_population)`, where `y_population` is the population predicted values
- `iwres` — The individual weighted residuals

- `pwres` — The population weighted residuals
- `cwres` — The conditional weighted residuals

`[beta,PSI,stats,B] = nlmefit(X,y,group,V,fun,beta0)` also returns `B`, an *r*-by-*m* matrix of estimated random effects for the *m* groups. By default, *r* is equal to the number of model parameters *p*.

`[beta,PSI,stats,B] = nlmefit(X,y,group,V,fun,beta0,'Name',value)` specifies one or more optional parameter name/value pairs. Specify *`Name`* inside single quotes.

Use the following parameters to fit a model different from the default. (The default model is obtained by setting both `FEConstDesign` and `REConstDesign` to `eye(p)`, or by setting both `FEParamsSelect` and `REParamsSelect` to `1:p`.) Use at most one parameter with an `'FE'` prefix and one parameter with an `'RE'` prefix. The `nlmefit` function requires you to specify at least one fixed effect and one random effect.

Parameter	Value
`FEParamsSelect`	A vector specifying which elements of the parameter vector `PHI` include a fixed effect, given as a numeric vector of indices from `1` to `p` or as a 1-by-*p* logical vector. If *q* is the specified number of elements, then the model includes *q* fixed effects.
`FEConstDesign`	A *p*-by-*q* design matrix `ADESIGN`, where `ADESIGN*beta` are the fixed components of the *p* elements of `PHI`.
`FEGroupDesign`	A *p*-by-*q*-by-*m* array specifying a different *p*-by-*q* fixed-effects design matrix for each of the *m* groups.
`FEObsDesign`	A *p*-by-*q*-by-*n* array specifying a different *p*-by-*q* fixed-effects design matrix for each of the *n* observations.
`REParamsSelect`	A vector specifying which elements of the parameter vector `PHI` include a random effect, given as a numeric vector of indices from `1` to `p` or as a 1-by-*p* logical vector. The model includes *r* random effects, where *r* is the specified number of elements.
`REConstDesign`	A *p*-by-*r* design matrix `BDESIGN`, where `BDESIGN*B` are the random components of the *p* elements of `PHI`.
`REGroupDesign`	A *p*-by-*r*-by-*m* array specifying a different *p*-by-*r* random-effects design matrix for each of *m* groups.
`REObsDesign`	A *p*-by-*r*-by-*n* array specifying a different *p*-by-*r* random-effects design matrix for each of *n* observations.

Use the following parameters to control the iterative algorithm for maximizing the likelihood:

Parameter	Value
`RefineBeta0`	Determines whether `nlmefit` makes an initial refinement of `beta0` by first fitting `modelfun` without random effects and replacing `beta0` with `beta`. Choices are `'on'` and `'off'`. The default value is `'on'`.

Parameter	Value
`ErrorModel`	A character vector or string scalar specifying the form of the error term. Default is `'constant'`. Each model defines the error using a standard normal (Gaussian) variable *e*, the function value *f*, and one or two parameters *a* and *b*. Choices are: • `'constant'`: $y = f + a*e$ • `'proportional'`: $y = f + b*f*e$ • `'combined'`: $y = f + (a+b*f)*e$ • `'exponential'`: $y = f*\exp(a*e)$, or equivalently $\log(y) = \log(f) + a*e$ If this parameter is given, the output `stats.errorparam` field has the value • *a* for `'constant'` and `'exponential'` • *b* for `'proportional'` • [*a b*] for `'combined'`
`ApproximationType`	The method used to approximate the likelihood of the model. Choices are: • `'LME'` — Use the likelihood for the linear mixed-effects model at the current conditional estimates of `beta` and `B`. This is the default. • `'RELME'` — Use the restricted likelihood for the linear mixed-effects model at the current conditional estimates of `beta` and `B`. • `'FO'` — First-order Laplacian approximation without random effects. • `'FOCE'` — First-order Laplacian approximation at the conditional estimates of `B`.

Parameter	Value
`Vectorization`	Indicates acceptable sizes for the `PHI`, `XFUN`, and `VFUN` input arguments to `modelfun`. Choices are: • `'SinglePhi'` — `modelfun` can only accept a single set of model parameters at a time, so `PHI` must be a single row vector in each call. `nlmefit` calls `modelfun` in a loop, if necessary, with a single `PHI` vector and with `XFUN` containing rows for a single observation or group at a time. `VFUN` may be a single row that applies to all rows of `XFUN`, or a matrix with rows corresponding to rows in `XFUN`. This is the default. • `'SingleGroup'` — `modelfun` can only accept inputs corresponding to a single group in the data, so `XFUN` must contain rows of `X` from a single group in each call. Depending on the model, `PHI` is a single row that applies to the entire group or a matrix with one row for each observation. `VFUN` is a single row. • `'Full'` — `modelfun` can accept inputs for multiple parameter vectors and multiple groups in the data. Either `PHI` or `VFUN` may be a single row that applies to all rows of `XFUN` or a matrix with rows corresponding to rows in `XFUN`. This option can improve performance by reducing the number of calls to `modelfun`, but may require `modelfun` to perform singleton expansion on `PHI` or `V`.
`CovParameterization`	Specifies the parameterization used internally for the scaled covariance matrix. Choices are `'chol'` for the Cholesky factorization or `'logm'` the matrix logarithm. The default is `'logm'`.
`CovPattern`	Specifies an *r*-by-*r* logical or numeric matrix `P` that defines the pattern of the random-effects covariance matrix `PSI`. `nlmefit` estimates the variances along the diagonal of `PSI` and the covariances specified by nonzeros in the off-diagonal elements of `P`. Covariances corresponding to zero off-diagonal elements in `P` are constrained to be zero. If `P` does not specify a row-column permutation of a block diagonal matrix, `nlmefit` adds nonzero elements to `P` as needed. The default value of `P` is `eye(r)`, corresponding to uncorrelated random effects. Alternatively, `P` may be a 1-by-*r* vector containing values in `1:r`, with equal values specifying groups of random effects. In this case, `nlmefit` estimates covariances only within groups, and constrains covariances across groups to be zero.

Parameter	Value
`ParamTransform`	A vector of *p*-values specifying a transformation function *f*() for each of the `P` parameters: `XB = ADESIGN*BETA + BDESIGN*B` `PHI = f(XB)`. Each element of the vector must be one of the following integer codes specifying the transformation for the corresponding value of PHI: • 0: `PHI = XB` (default for all parameters) • 1: log(`PHI`) = `XB` • 2: probit(`PHI`) = `XB` • 3: logit(`PHI`) = `XB`

`Options`	A structure of the form returned by `statset`. `nlmefit` uses the following `statset` parameters:
	• `'DerivStep'` — Relative difference used in finite difference gradient calculation. May be a scalar, or a vector whose length is the number of model parameters *p*. The default is `eps^(1/3)`.
	• `'Display'` — Level of iterative display during estimation. Choices are:
	  • `'off'` (default) — Displays no information
	  • `'final'` — Displays information after the final iteration
	  • `'iter'` — Displays information at each iteration
	• `'FunValCheck'` — Check for invalid values, such as `NaN` or `Inf`, from `modelfun`. Choices are `'on'` and `'off'`. The default is `'on'`.
	• `'MaxIter'` — Maximum number of iterations allowed. The default is `200`.
	• `'OutputFcn'` — Function handle specified using `@`, a cell array with function handles or an empty array (default). The solver calls all output functions after each iteration.
	• `'TolFun'` — Termination tolerance on the loglikelihood function. The default is `1e-4`.
	• `'TolX'` — Termination tolerance on the estimated fixed and random effects. The default is `1e-4`.

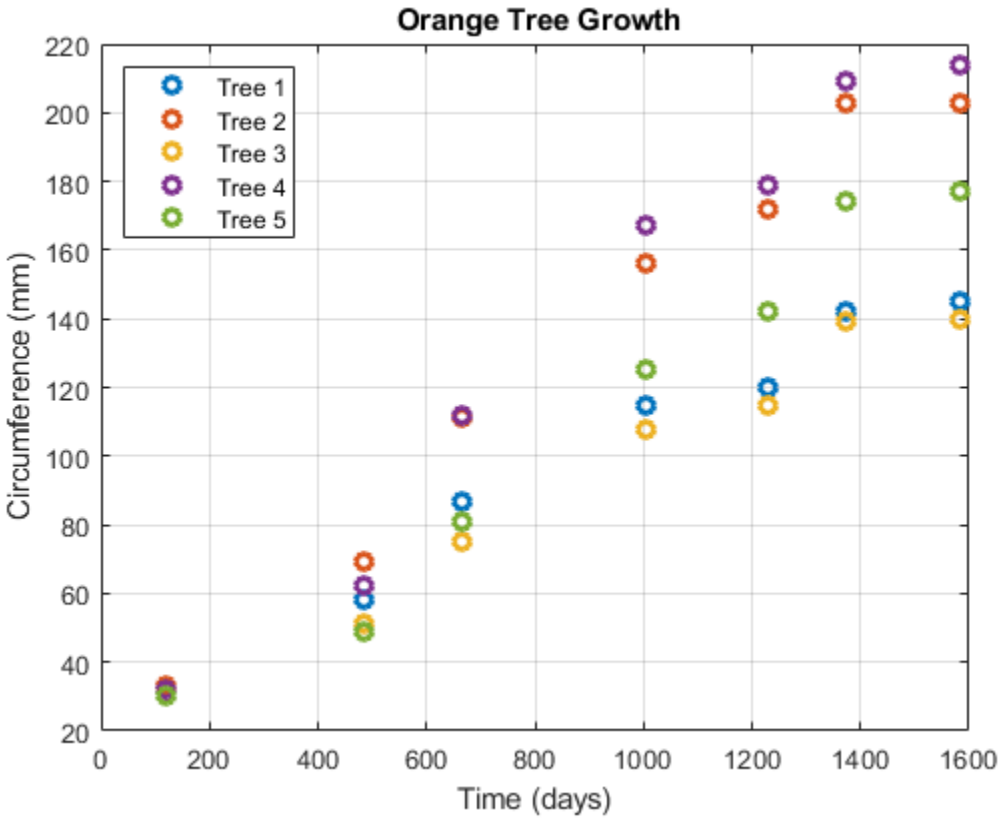
`OptimFun`	Specifies the optimization function used in maximizing the likelihood. Choices are `'fminsearch'` to use `fminsearch` or `'fminunc'` to use `fminunc`. The default is `'fminsearch'`. You can specify `'fminunc'` only if Optimization Toolbox software is installed.

Examples

Nonlinear Mixed-Effects Model

Enter and display data on the growth of five orange trees.

```
CIRC = [30 58 87 115 120 142 145;
        33 69 111 156 172 203 203;
        30 51 75 108 115 139 140;
        32 62 112 167 179 209 214;
        30 49 81 125 142 174 177];
time = [118 484 664 1004 1231 1372 1582];

h = plot(time,CIRC','o','LineWidth',2);
xlabel('Time (days)')
ylabel('Circumference (mm)')
title('{\bf Orange Tree Growth}')
legend([repmat('Tree ',5,1),num2str((1:5)')],...
       'Location','NW')
grid on
hold on
```



Use an anonymous function to specify a logistic growth model.

```
model = @(PHI,t)(PHI(:,1))./(1+exp(-(t-PHI(:,2))./PHI(:,3)));
```

Fit the model using `nlmefit` with default settings (that is, assuming each parameter is the sum of a fixed and a random effect, with no correlation among the random effects):

```
TIME = repmat(time,5,1);
NUMS = repmat((1:5)',size(time));

beta0 = [100 100 100];
```

```
[beta1,PSI1,stats1] = nlmefit(TIME(:),CIRC(:),NUMS(:),...
                              [],model,beta0)

beta1 = 3×1

  191.3189
  723.7608
  346.2517


PSI1 = 3×3

  962.1535         0         0
         0    0.0000         0
         0         0  297.9880


stats1 = struct with fields:
           dfe: 28
          logl: -131.5457
           mse: 59.7882
          rmse: 7.9016
    errorparam: 7.7323
           aic: 277.0913
           bic: 274.3574
          covb: [3x3 double]
        sebeta: [15.2249 33.1579 26.8235]
          ires: [35x1 double]
          pres: [35x1 double]
         iwres: [35x1 double]
         pwres: [35x1 double]
         cwres: [35x1 double]
```

The negligible variance of the second random effect, `PSI1(2,2)`, suggests that it can be removed to simplify the model.

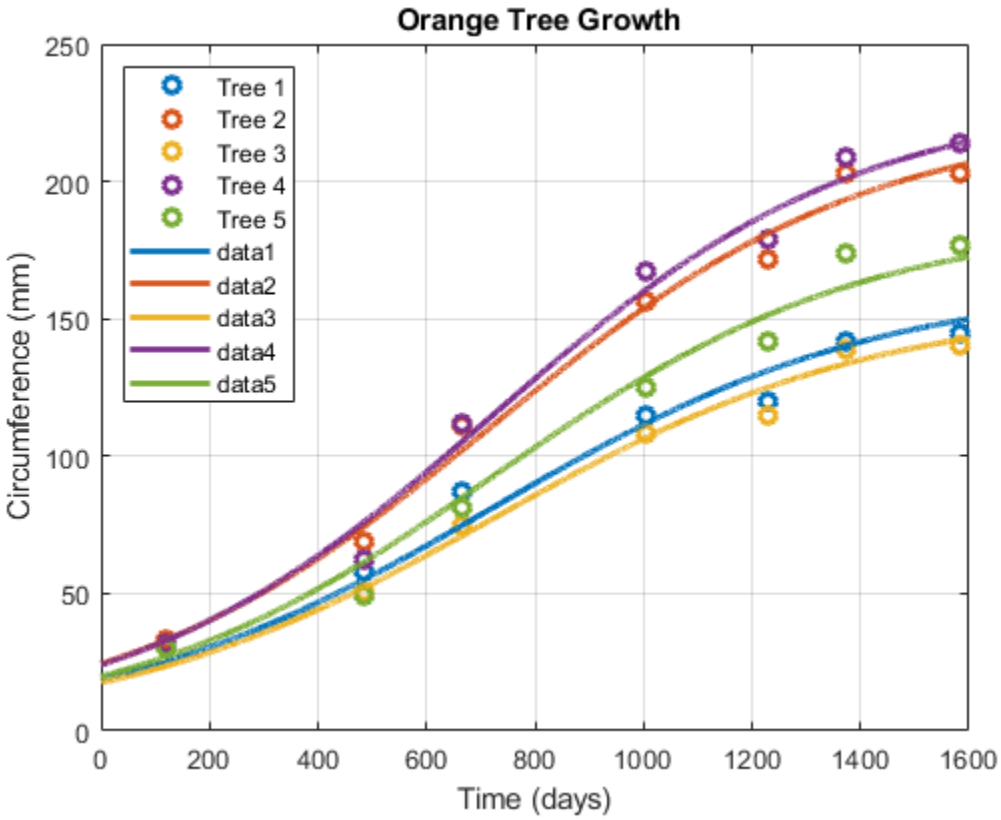
```
[beta2,PSI2,stats2,b2] = nlmefit(TIME(:),CIRC(:),...
    NUMS(:),[],model,beta0,'REParamsSelect',[1 3])

beta2 = 3×1

  191.3193
  723.7629
  346.2532


PSI2 = 2×2

  962.4847         0
         0  297.9930


stats2 = struct with fields:
           dfe: 29
          logl: -131.5456
           mse: 59.7847
          rmse: 7.7642
    errorparam: 7.7321
```

```
           aic: 275.0913
           bic: 272.7479
          covb: [3x3 double]
        sebeta: [15.2270 33.1573 26.8230]
          ires: [35x1 double]
          pres: [35x1 double]
         iwres: [35x1 double]
         pwres: [35x1 double]
         cwres: [35x1 double]


b2 = 2×5

  -28.5262   31.6066  -36.5078   39.0748   -5.6474
    9.9981   -0.7623    6.0046   -9.4579   -5.7824
```

The loglikelihood `logl` is unaffected, and both the Akaike and Bayesian information criteria (`aic` and `bic`) are reduced, supporting the decision to drop the second random effect from the model.

Use the estimated fixed effects in `beta2` and the estimated random effects for each tree in `b2` to plot the model through the data.

```
PHI = repmat(beta2,1,5) + ...          % Fixed effects
      [b2(1,:);zeros(1,5);b2(2,:)];    % Random effects

tplot = 0:0.1:1600;
for I = 1:5
  fitted_model=@(t)(PHI(1,I))./(1+exp(-(t-PHI(2,I))./ ...
       PHI(3,I)));
  plot(tplot,fitted_model(tplot),'Color',h(I).Color, ...
       'LineWidth',2)
end
```



References

[1] Lindstrom, M. J., and D. M. Bates. “Nonlinear mixed-effects models for repeated measures data.” *Biometrics*. Vol. 46, 1990, pp. 673–687.

[2] Davidian, M., and D. M. Giltinan. *Nonlinear Models for Repeated Measurements Data*. New York: Chapman & Hall, 1995.

[3] Pinheiro, J. C., and D. M. Bates. “Approximations to the log-likelihood function in the nonlinear mixed-effects model.” *Journal of Computational and Graphical Statistics*. Vol. 4, 1995, pp. 12–35.

[4] Demidenko, E. *Mixed Models: Theory and Applications*. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

See Also

`nlinfit` | `nlmefitsa` | `nlpredci`

Topics

“Mixed-Effects Models” on page 13-17
“Grouping Variables” on page 2-45

Introduced in R2008b

nlmefitsa

Fit nonlinear mixed-effects model with stochastic EM algorithm

Syntax

```
[BETA,PSI,STATS,B] = nlmefitsa(X,Y,GROUP,V,MODELFUN,BETA0)
[BETA,PSI,STATS,B] = nlmefitsa(X,Y,GROUP,V,MODELFUN,BETA0,'Name',Value)
```

Description

`[BETA,PSI,STATS,B] = nlmefitsa(X,Y,GROUP,V,MODELFUN,BETA0)` fits a nonlinear mixed-effects regression model and returns estimates of the fixed effects in `BETA`. By default, `nlmefitsa` fits a model where each model parameter is the sum of a corresponding fixed and random effect, and the covariance matrix of the random effects is diagonal, i.e., uncorrelated random effects.

The `BETA`, `PSI`, and other values this function returns are the result of a random (Monte Carlo) simulation designed to converge to the maximum likelihood estimates of the parameters. Because the results are random, it is advisable to examine the plot of simulation to results to be sure that the simulation has converged. It may also be helpful to run the function multiple times, using multiple starting values, or use the `'Replicates'` parameter to perform multiple simulations.

`[BETA,PSI,STATS,B] = nlmefitsa(X,Y,GROUP,V,MODELFUN,BETA0,'Name',Value)` accepts one or more comma-separated parameter name/value pairs. Specify *`Name`* inside single quotes.

Input Arguments

Definitions:

In the following list of arguments, the following variable definitions apply:

- *n* — number of observations
- *h* — number of predictor variables
- *m* — number of groups
- *g* — number of group-specific predictor variables
- *p* — number of parameters
- *f* — number of fixed effects

X

An *n*-by-*h* matrix of *n* observations on *h* predictor variables.

Y

An *n*-by-1 vector of responses.

GROUP

A grouping variable indicating to which of *m* groups each observation belongs. GROUP can be a categorical variable, a numeric vector, a character matrix with rows for group names, a string array, or a cell array of character vectors.

V

An *m*-by-*g* matrix of *g* group-specific predictor variables for each of the *m* groups in the data. These are predictor values that take on the same value for all observations in a group. Rows of V are ordered according to GRP2IDX(GROUP). Use an *m*-by-*g* cell array for V if any of the group-specific predictor values vary in size across groups. Specify [] for V if there are no group predictors.

MODELFUN

A handle to a function that accepts predictor values and model parameters, and returns fitted values. MODELFUN has the form YFIT = MODELFUN(PHI,XFUN,VFUN) with input arguments

- PHI — A 1-by-*p* vector of model parameters.
- XFUN — An *l*-by-*h* array of predictor variables where
  - *l* is 1 if XFUN is a single row of X
  - *l* is n_i if XFUN contains the rows of X for a single group of size n_i
  - *l* is *n* if XFUN contains all rows of X.
- VFUN — Either
  - A 1-by-*g* vector of group-specific predictors for a single group, corresponding to a single row of V
  - An *n*-by-*g* matrix, where the *k*-th row of VFUN is V(*i*,:) if the *k*-th observation is in group *i*.

  If V is empty, nlmefitsa calls MODELFUN with only two inputs.

MODELFUN returns an *l*-by-1 vector of fitted values YFIT. When either PHI or VFUN contains a single row, that one row corresponds to all rows in the other two input arguments. For improved performance, use the 'Vectorization' parameter name/value pair (described below) if MODELFUN can compute YFIT for more than one vector of model parameters in one call.

BETA0

An *f*-by-1 vector with initial estimates for the *f* fixed effects. By default, *f* is equal to the number of model parameters *p*. BETA0 can also be an *f*-by-REPS matrix, and the estimation is repeated REPS times using each column of BETA0 as a set of starting values.

Name-Value Pair Arguments

By default, nlmefitsa fits a model where each model parameter is the sum of a corresponding fixed and random effect. Use the following parameter name/value pairs to fit a model with a different number of or dependence on fixed or random effects. Use at most one parameter name with an 'FE' prefix and one parameter name with an 'RE' prefix. Note that some choices change the way nlmefitsa calls MODELFUN, as described further below.

`FEParamsSelect`

A vector specifying which elements of the model parameter vector `PHI` include a fixed effect, as a numeric vector with elements in 1:*p*, or as a 1-by-*p* logical vector. The model will include *f* fixed effects, where *f* is the specified number of elements.

`FEConstDesign`

A *p*-by-*f* design matrix `ADESIGN`, where `ADESIGN*BETA` are the fixed components of the *p* elements of `PHI`.

`FEGroupDesign`

A *p*-by-*f*-by-*m* array specifying a different *p*-by-*f* fixed effects design matrix for each of the *m* groups.

`REParamsSelect`

A vector specifying which elements of the model parameter vector `PHI` include a random effect, as a numeric vector with elements in 1:*p*, or as a 1-by-*p* logical vector. The model will include *r* random effects, where *r* is the specified number of elements.

`REConstDesign`

A *p*-by-*r* design matrix `BDESIGN`, where `BDESIGN*B` are the random components of the *p* elements of `PHI`. This matrix must consist of 0s and 1s, with at most one 1 per row.

The default model is equivalent to setting both `FEConstDesign` and `REConstDesign` to `eye(p)`, or to setting both `FEParamsSelect` and `REParamsSelect` to 1:*p*.

Additional optional parameter name/value pairs control the iterative algorithm used to maximize the likelihood:

`CovPattern`

Specifies an *r*-by-*r* logical or numeric matrix `PAT` that defines the pattern of the random effects covariance matrix `PSI`. `nlmefitsa` computes estimates for the variances along the diagonal of `PSI` as well as covariances that correspond to non-zeroes in the off-diagonal of `PAT`. `nlmefitsa` constrains the remaining covariances, i.e., those corresponding to off-diagonal zeroes in `PAT`, to be zero. `PAT` must be a row-column permutation of a block diagonal matrix, and `nlmefitsa` adds non-zero elements to `PAT` as needed to produce such a pattern. The default value of `PAT` is `eye(r)`, corresponding to uncorrelated random effects.

Alternatively, specify `PAT` as a 1-by-*r* vector containing values in 1:*r*. In this case, elements of `PAT` with equal values define groups of random effects, `nlmefitsa` estimates covariances only within groups, and constrains covariances across groups to be zero.

`Cov0`

Initial value for the covariance matrix `PSI`. Must be an *r*-by-*r* positive definite matrix. If empty, the default value depends on the values of `BETA0`.

`ComputeStdErrors`

`true` to compute standard errors for the coefficient estimates and store them in the output `STATS` structure, or `false` (default) to omit this computation.

`ErrorModel`

A character vector or string scalar specifying the form of the error term. Default is `'constant'`. Each model defines the error using a standard normal (Gaussian) variable *e*, the function value *f*, and one or two parameters *a* and *b*. Choices are

- `'constant'` — $y = f + a*e$
- `'proportional'` — $y = f + b*f*e$
- `'combined'` — $y = f + (a+b*f)*e$
- `'exponential'` — $y = f*\exp(a*e)$, or equivalently $\log(y) = \log(f) + a*e$

If this parameter is given, the output `STATS.errorparam` field has the value

- *a* for `'constant'` and `'exponential'`
- *b* for `'proportional'`
- [*a b*] for `'combined'`

`ErrorParameters`

A scalar or two-element vector specifying starting values for parameters of the error model. This specifies the *a*, *b*, or [*a b*] values depending on the `ErrorModel` parameter.

`LogLikMethod`

Specifies the method for approximating the loglikelihood. Choices are:

- `'is'` — Importance sampling
- `'gq'` — Gaussian quadrature
- `'lin'` — Linearization
- `'none'` — Omit the loglikelihood approximation (default)

`NBurnIn`

Number of initial burn-in iterations during which the parameter estimates are not recomputed. Default is 5.

`NChains`

Number *c* of "chains" simulated. Default is 1. Setting *c*>1 causes *c* simulated coefficient vectors to be computed for each group during each iteration. Default depends on the data, and is chosen to provide about 100 groups across all chains.

`NIterations`

Number of iterations. This can be a scalar or a three-element vector. Controls how many iterations are performed for each of three phases of the algorithm:

1 simulated annealing

2 full step size

3 reduced step size

Default is `[150 150 100]`. A scalar is distributed across the three phases in the same proportions as the default.

NMCMCIterations

Number of Markov Chain Monte Carlo (MCMC) iterations. This can be a scalar or a three-element vector. Controls how many of three different types of MCMC updates are performed during each phase of the main iteration:

1 full multivariate update

2 single coordinate update

3 multiple coordinate update

Default is `[2 2 2]`. A scalar value is treated as a three-element vector with all elements equal to the scalar.

OptimFun

Either `'fminsearch'` or `'fminunc'`, specifying the optimization function to be used during the estimation process. Default is `'fminsearch'`. Use of `'fminunc'` requires Optimization Toolbox.

Options

A structure created by a call to `statset`. `nlmefitsa` uses the following `statset` parameters:

- `'DerivStep'` — Relative difference used in finite difference gradient calculation. May be a scalar, or a vector whose length is the number of model parameters *p*. The default is `eps^(1/3)`.
- `Display` — Level of display during estimation.
  - `'off'` (default) — Displays no information
  - `'final'` — Displays information after the final iteration of the estimation algorithm
  - `'iter'` — Displays information at each iteration
- `FunValCheck`
  - `'on'` (default) — Check for invalid values (such as `NaN` or `Inf`) from `MODELFUN`
  - `'off'` — Skip this check
- `OutputFcn` — Function handle specified using `@`, a cell array with function handles or an empty array. `nlmefitsa` calls all output functions after each iteration. See `nlmefitoutputfcn.m` (the default output function for `nlmefitsa`) for an example of an output function.

ParamTransform

A vector of *p*-values specifying a transformation function `f()` for each of the *p* parameters:

```
XB = ADESIGN*BETA + BDESIGN*B
PHI = f(XB)
```

Each element of the vector must be one of the following integer codes specifying the transformation for the corresponding value of `PHI`:

- 0: `PHI = XB` (default for all parameters)
- 1: `log(PHI) = XB`
- 2: `probit(PHI) = XB`
- 3: `logit(PHI) = XB`

Replicates

Number REPS of estimations to perform starting from the starting values in the vector BETA0. If BETA0 is a matrix, REPS must match the number of columns in BETA0. Default is the number of columns in BETA0.

Vectorization

Determines the possible sizes of the PHI, XFUN, and VFUN input arguments to MODELFUN. Possible values are:

- 'SinglePhi' — MODELFUN is a function (such as an ODE solver) that can only compute YFIT for a single set of model parameters at a time, i.e., PHI must be a single row vector in each call. nlmefitsa calls MODELFUN in a loop if necessary using a single PHI vector and with XFUN containing rows for a single observation or group at a time. VFUN may be a single row that applies to all rows of XFUN, or a matrix with rows corresponding to rows in XFUN.
- 'SingleGroup' — MODELFUN can only accept inputs corresponding to a single group in the data, i.e., XFUN must contain rows of X from a single group in each call. Depending on the model, PHI is a single row that applies to the entire group, or a matrix with one row for each observation. VFUN is a single row.
- 'Full' — MODELFUN can accept inputs for multiple parameter vectors and multiple groups in the data. Either PHI or VFUN may be a single row that applies to all rows of XFUN, or a matrix with rows corresponding to rows in XFUN. Using this option can improve performance by reducing the number of calls to MODELFUN, but may require MODELFUN to perform singleton expansion on PHI or V.

The default for 'Vectorization' is 'SinglePhi'. In all cases, if V is empty, nlmefitsa calls MODELFUN with only two inputs.

Output Arguments

BETA

Estimates of the fixed effects

PSI

An *r*-by-*r* estimated covariance matrix for the random effects. By default, *r* is equal to the number of model parameters *p*.

STATS

A structure with the following fields:

- logl — The maximized loglikelihood for the fitted model; empty if the LogLikMethod parameter has its default value of 'none'
- rmse — The square root of the estimated error variance (computed on the log scale for the exponential error model)
- errorparam — The estimated parameters of the error variance model
- aic — The Akaike information criterion (empty if logl is empty), calculated as aic = –2 * logl + 2 * numParam, where

  - `logl` is the maximized loglikelihood.
  - `numParam` is the number of fitting parameters, including the degree of freedom for covariance matrix of the random effects, the number of fixed effects and the number of parameters of the error model.
- `bic` — The Bayesian information criterion (empty if `logl` is empty), calculated as `bic` = -2*`logl` + log(`M`) * `numParam`
  - `M` is the number of groups.
  - `logl` and `numParam` are defined as in `aic`.

  Note that some literature suggests that the computation of `bic` should be , `bic` = -2*`logl` + log(`N`) * `numParam`, where `N` is the number of observations. To adjust the value of the output you can redefine `bic` as follows: `bic` = `bic` - `numel`(`unique`(`group`)) + `numel`(`Y`)
- `sebeta` — The standard errors for BETA (empty if the `ComputeStdErrors` parameter has its default value of false)
- `covb` — The estimated covariance of the parameter estimates (empty if `ComputeStdErrors` is false)
- `dfe` — The error degrees of freedom
- `pres` — The population residuals `(y-y_population)`, where `y_population` is the population predicted values
- `ires` — The population residuals `(y-y_population)`, where `y_population` is the individual predicted values
- `pwres` — The population weighted residuals
- `cwres` — The conditional weighted residuals
- `iwres` — The individual weighted residuals

Examples

Nonlinear Mixed-Effects Model with Stochastic EM Algorithm

Load the sample data.

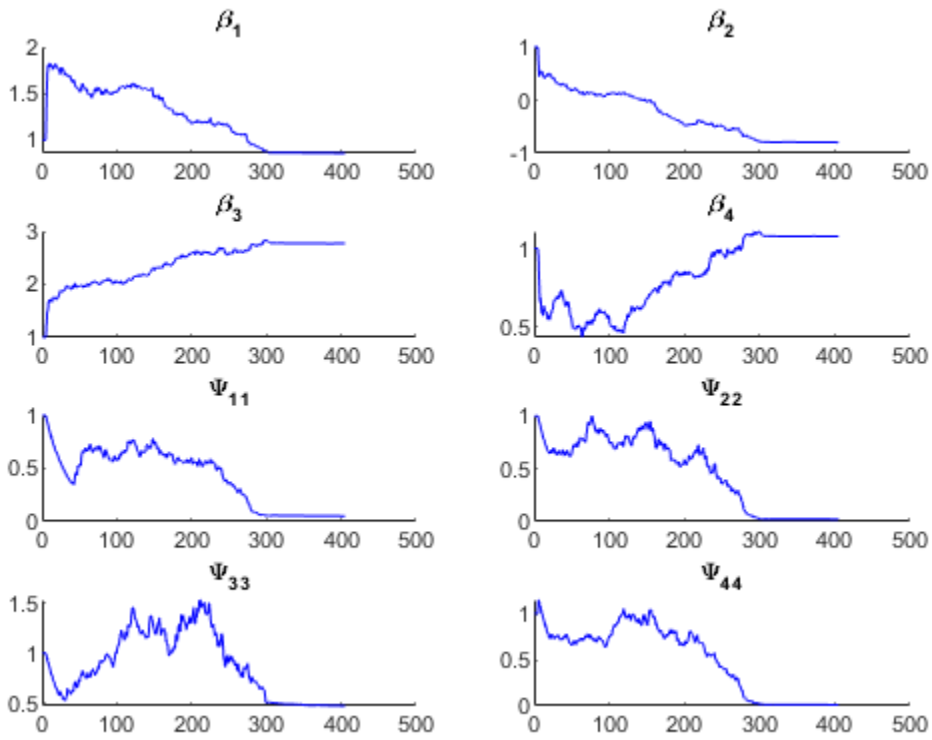
```
load indomethacin
```

Fit a model to data on concentrations of the drug indomethacin in the bloodstream of six subjects over eight hours.

```
model = @(phi,t)(phi(:,1).*exp(-phi(:,2).*t)+phi(:,3).*exp(-phi(:,4).*t));
phi0 = [1 1 1 1];
xform = [0 1 0 1]; % log transform for 2nd and 4th parameters
[beta,PSI,stats,br] = nlmefitsa(time,concentration,...
   subject,[],model,phi0,'ParamTransform',xform)
```

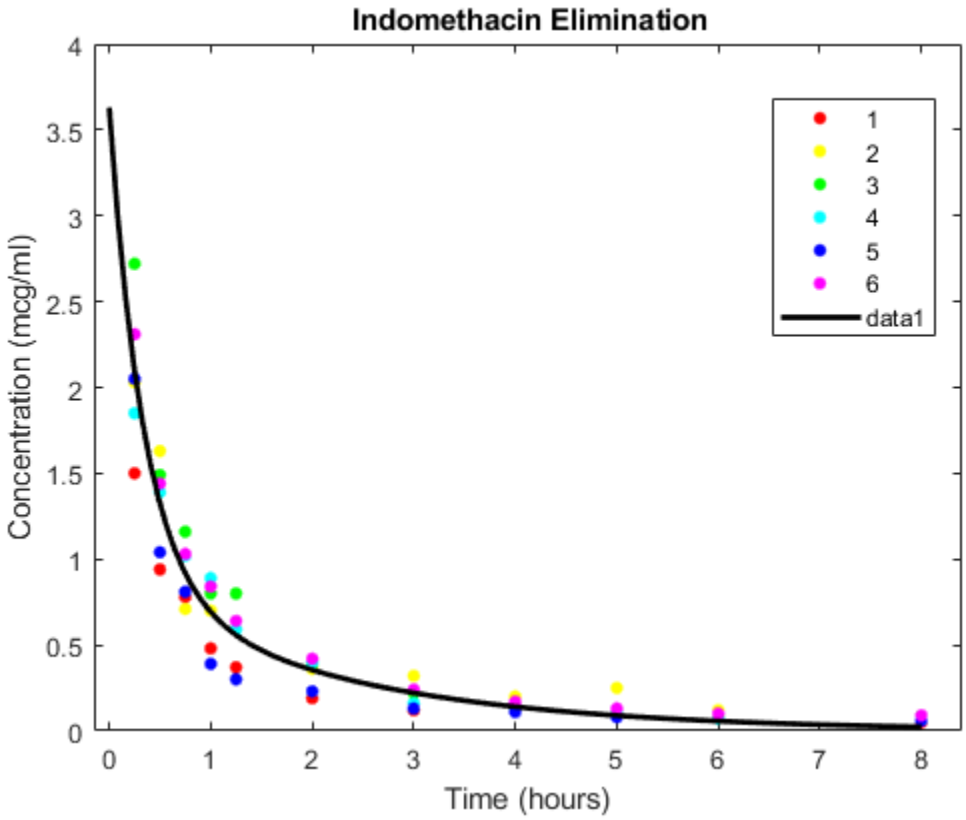
```
beta =

    0.8563
   -0.7950
    2.7744
    1.0772
```

```
PSI =

    0.0529         0         0         0
         0    0.0220         0         0
         0         0    0.4762         0
         0         0         0    0.0120


stats =

  struct with fields:

          logl: []
           aic: []
           bic: []
        sebeta: []
           dfe: 57
          covb: []
    errorparam: 0.0809
          rmse: 0.0775
          ires: [66x1 double]
          pres: [66x1 double]
         iwres: [66x1 double]
         pwres: [66x1 double]
         cwres: [66x1 double]


br =

   -0.2255    0.0063    0.1600    0.1773   -0.3269    0.1157
    0.0350   -0.1384    0.0058    0.0431    0.0093   -0.0453
   -0.7557   -0.0550    0.8736   -0.7875    0.5304    0.1727
   -0.0010   -0.0198    0.0137   -0.0757    0.0478   -0.0076
```



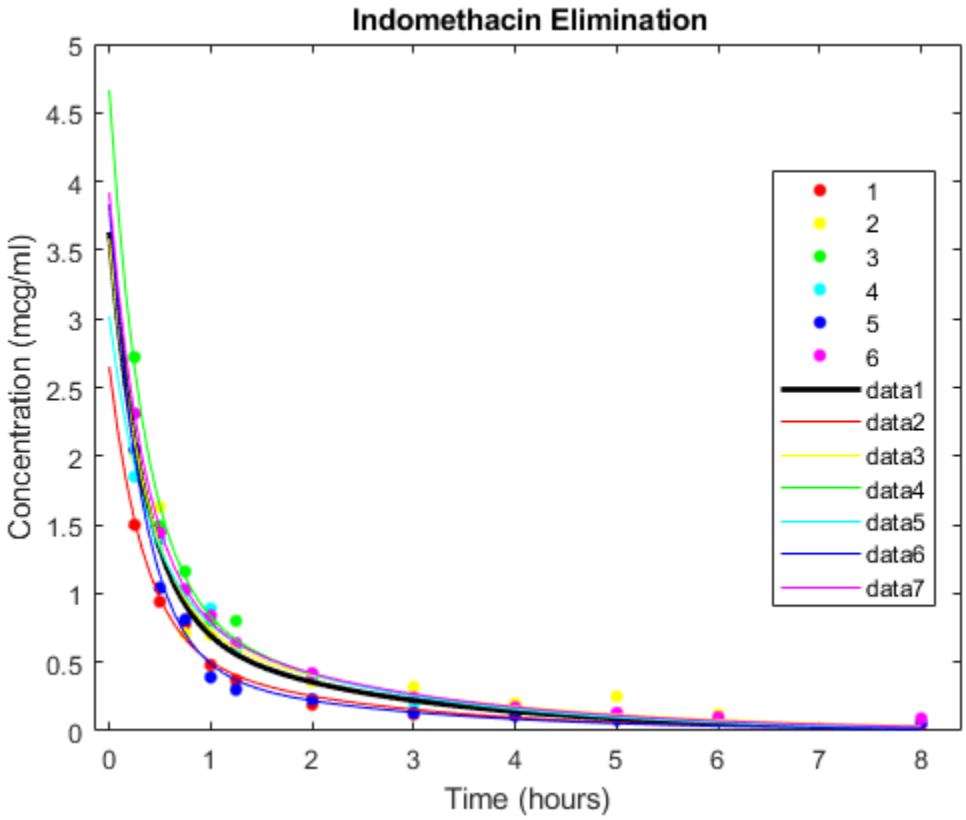
Plot the data along with an overall population fit

```
clf
phi = [beta(1), exp(beta(2)), beta(3), exp(beta(4))];
h = gscatter(time,concentration,subject);
xlabel('Time (hours)')
ylabel('Concentration (mcg/ml)')
title('{\bf Indomethacin Elimination}')
xx = linspace(0,8);
line(xx,model(phi,xx),'linewidth',2,'color','k')
```



Plot individual curves based on random-effect estimates.

```
for j=1:6
    phir = [beta(1)+br(1,j), exp(beta(2)+br(2,j)), ...
            beta(3)+br(3,j), exp(beta(4)+br(4,j))];
    line(xx,model(phir,xx),'color',get(h(j),'color'))
end
```



Algorithms

In order to estimate the parameters of a nonlinear mixed effects model, we would like to choose the parameter values that maximize a likelihood function. These values are called the maximum likelihood estimates. The likelihood function can be written in the form

$$p\left(y\left|\beta,\sigma^2,\Sigma\right.\right)=\int p\left(y\left|\beta,b,\sigma^2\right.\right)p(b\,|\,\Sigma)db$$

where

- y is the response data
- β is the vector of population coefficients
- σ^2 is the residual variance
- $\sum$ is the covariance matrix for the random effects
- b is the set of unobserved random effects

Each $p()$ function on the right-hand-side is a normal (Gaussian) likelihood function that may depend on covariates.

Since the integral does not have a closed form, it is difficult to find parameters that maximize it. Delyon, Lavielle, and Moulines [1] proposed to find the maximum likelihood estimates using an Expectation-Maximization (EM) algorithm in which the E step is replaced by a stochastic procedure. They called their algorithm SAEM, for Stochastic Approximation EM. They demonstrated that this

algorithm has desirable theoretical properties, including convergence under practical conditions and convergence to a local maximum of the likelihood function. Their proposal involves three steps:

1. Simulation: Generate simulated values of the random effects b from the posterior density $p(b|\Sigma)$ given the current parameter estimates.
2. Stochastic approximation: Update the expected value of the loglikelihood function by taking its value from the previous step, and moving part way toward the average value of the loglikelihood calculated from the simulated random effects.
3. Maximization step: Choose new parameter estimates to maximize the loglikelihood function given the simulated values of the random effects.

References

[1] Delyon, B., M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm." *Annals of Statistics,* 27, 94-128, 1999.

[2] Mentré, F., and M. Lavielle, "Stochastic EM Algorithms in Population PKPD analyses." *American Conference on Pharmacometrics,* 2008.

See Also

`nlinfit` | `nlmefit` | `nlpredci`

Topics

“Mixed-Effects Models” on page 13-17
“Grouping Variables” on page 2-45

Introduced in R2010a

nlparci

Nonlinear regression parameter confidence intervals

Syntax

```
ci = nlparci(beta,resid,'covar',sigma)
ci = nlparci(beta,resid,'jacobian',J)
ci = nlparci(...,'alpha',alpha)
```

Description

`ci = nlparci(beta,resid,'covar',sigma)` returns the 95% confidence intervals `ci` for the nonlinear least squares parameter estimates `beta`. Before calling `nlparci`, use `nlinfit` to fit a nonlinear regression model and get the coefficient estimates `beta`, residuals `resid`, and estimated coefficient covariance matrix `sigma`.

`ci = nlparci(beta,resid,'jacobian',J)` is an alternative syntax that also computes 95% confidence intervals. `J` is the Jacobian computed by `nlinfit`. If the `'robust'` option is used with `nlinfit`, use the `'covar'` input rather than the `'jacobian'` input so that the required `sigma` parameter takes the robust fitting into account.

`ci = nlparci(...,'alpha',alpha)` returns `100(1-alpha)`% confidence intervals.

`nlparci` treats `NaN`s in `resid` or `J` as missing values, and ignores the corresponding observations.

The confidence interval calculation is valid for systems where the length of `resid` exceeds the length of `beta` and `J` has full column rank. When `J` is ill-conditioned, confidence intervals may be inaccurate.

Examples

Fit to Exponential Decay

Suppose you have data, and want to fit a model of the form

$y_i = a_1 + a_2\exp(-a_3 x_i) + \epsilon_i$.

a_i are the parameters you want to estimate, x_i are the data points, y_i are the responses, and ϵ_i are noise terms.

Write a function handle that represents the model:

```
mdl = @(a,x)(a(1) + a(2)*exp(-a(3)*x));
```

Generate synthetic data with parameters `a = [1;3;2]`, with the `x` data points distributed exponentially with parameter `2`, and normally distributed noise with standard deviation `0.1`:

```
rng(9845,'twister') % for reproducibility
a = [1;3;2];
x = exprnd(2,100,1);
```

```
epsn = normrnd(0,0.1,100,1);
y = mdl(a,x) + epsn;
```

Fit the model to data starting from the arbitrary guess `a0 = [2;2;2]`:

```
a0 = [2;2;2];
[ahat,r,J,cov,mse] = nlinfit(x,y,mdl,a0);
ahat
```

```
ahat = 3×1

    1.0153
    3.0229
    2.1070
```

Check whether `[1;3;2]` is in a 95% confidence interval using the Jacobian argument in `nlparci`:

```
ci = nlparci(ahat,r,'Jacobian',J)
```

```
ci = 3×2

    0.9869    1.0438
    2.9401    3.1058
    1.9963    2.2177
```

You can obtain the same result using the covariance argument:

```
ci = nlparci(ahat,r,'covar',cov)
```

```
ci = 3×2

    0.9869    1.0438
    2.9401    3.1058
    1.9963    2.2177
```

See Also

`nlinfit` | `nlpredci`

Introduced before R2006a

nlpredci

Nonlinear regression prediction confidence intervals

Syntax

```
[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Covar',CovB)
[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Covar',CovB,Name,Value)

[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Jacobian',J)
[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Jacobian',J,Name,Value)
```

Description

`[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Covar',CovB)` returns predictions, `Ypred`, and 95% confidence interval half-widths, `delta`, for the nonlinear regression model `modelfun` at input values `X`. Before calling `nlpredci`, use `nlinfit` to fit `modelfun` and get the estimated coefficients, `beta`, residuals, `R`, and variance-covariance matrix, `CovB`.

`[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Covar',CovB,Name,Value)` uses additional options specified by one or more name-value pair arguments.

`[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Jacobian',J)` returns predictions, `Ypred`, and 95% confidence interval half-widths, `delta`, for the nonlinear regression model `modelfun` at input values `X`. Before calling `nlpredci`, use `nlinfit` to fit `modelfun` and get the estimated coefficients, `beta`, residuals, `R`, and Jacobian, `J`.

If you use a robust option with `nlinfit`, then you should use the `Covar` syntax rather than the `Jacobian` syntax. The variance-covariance matrix, `CovB`, is required to properly take the robust fitting into account.

`[Ypred,delta] = nlpredci(modelfun,X,beta,R,'Jacobian',J,Name,Value)` uses additional options specified by one or more name-value pair arguments.

Examples

Confidence Interval for Nonlinear Regression Curve

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the initial values in `beta0`.

```
[beta,R,J] = nlinfit(X,y,@hougen,beta0);
```

Obtain the predicted response and 95% confidence interval half-width for the value of the curve at average reactant levels.

```
[ypred,delta] = nlpredci(@hougen,mean(X),beta,R,'Jacobian',J)
```

```
ypred = 5.4622
```

```
delta = 0.1921
```

Compute the 95% confidence interval for the value of the curve.

```
[ypred-delta,ypred+delta]
```

```
ans = 1×2

    5.2702    5.6543
```

Prediction Interval for New Observation

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the initial values in `beta0`.

```
[beta,R,J] = nlinfit(X,y,@hougen,beta0);
```

Obtain the predicted response and 95% prediction interval half-width for a new observation with reactant levels `[100,100,100]`.

```
[ypred,delta] = nlpredci(@hougen,[100,100,100],beta,R,'Jacobian',J,...
                         'PredOpt','observation')
```

```
ypred = 1.8346
```

```
delta = 0.5101
```

Compute the 95% prediction interval for the new observation.

```
[ypred-delta,ypred+delta]
```

```
ans = 1×2

    1.3245    2.3447
```

Simultaneous Confidence Intervals for Robust Fit Curve

Generate sample data from the nonlinear regression model $y = b_1 + b_2 \cdot exp\{b_3 x\} + \epsilon$, where b_1, b_2, and b_3 are coefficients, and the error term is normally distributed with mean 0 and standard deviation 0.5.

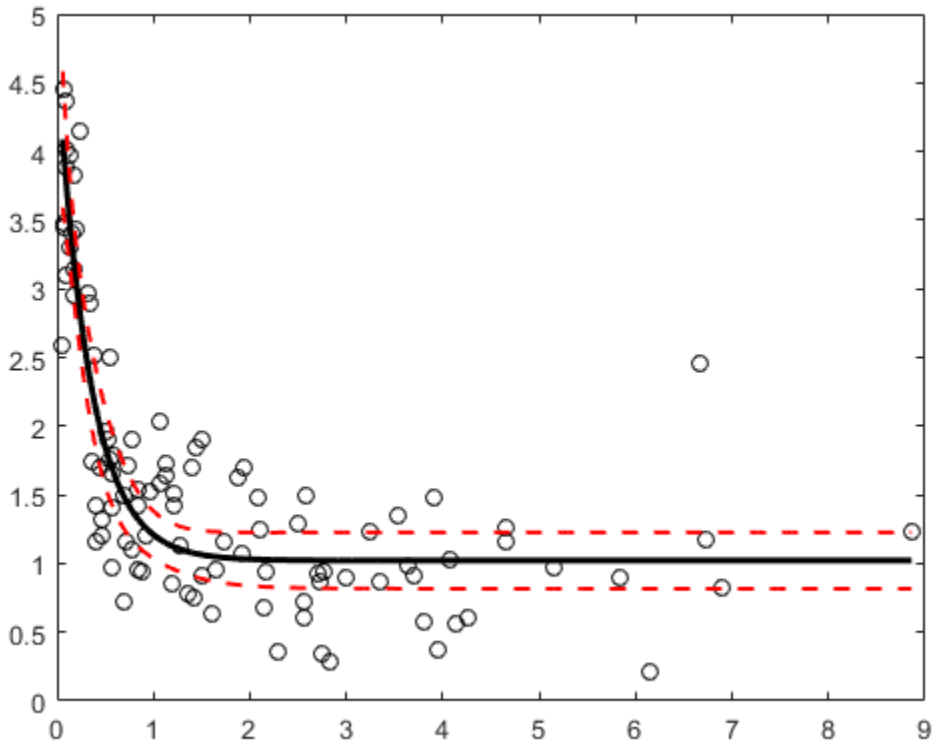
```
modelfun = @(b,x)(b(1)+b(2)*exp(-b(3)*x));

rng('default') % for reproducibility
b = [1;3;2];
x = exprnd(2,100,1);
y = modelfun(b,x) + normrnd(0,0.5,100,1);
```

Fit the nonlinear model using robust fitting options.

```
opts = statset('nlinfit');
opts.RobustWgtFun = 'bisquare';
beta0 = [2;2;2];
[beta,R,J,CovB,MSE] = nlinfit(x,y,modelfun,beta0,opts);
```

Plot the fitted regression model and simultaneous 95% confidence bounds.

```
xrange = min(x):.01:max(x);
[ypred,delta] = nlpredci(modelfun,xrange,beta,R,'Covar',CovB,...
                         'MSE',MSE,'SimOpt','on');
lower = ypred - delta;
upper = ypred + delta;

figure()
plot(x,y,'ko') % observed data
hold on
plot(xrange,ypred,'k','LineWidth',2)
plot(xrange,[lower;upper],'r--','LineWidth',1.5)
```



Confidence Interval Using Observation Weights

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Specify a function handle for observation weights, then fit the Hougen-Watson model to the rate data using the specified observation weights function.

```
a = 1; b = 1;
weights = @(yhat) 1./((a + b*abs(yhat)).^2);
[beta,R,J,CovB] = nlinfit(X,y,@hougen,beta0,'Weights',weights);
```

Compute the 95% prediction interval for a new observation with reactant levels `[100,100,100]` using the observation weight function.

```
[ypred,delta] = nlpredci(@hougen,[100,100,100],beta,R,'Jacobian',J,...
                         'PredOpt','observation','Weights',weights);
[ypred-delta,ypred+delta]
```

```
ans = 1×2

    1.5264    2.1033
```

Confidence Interval Using Nonconstant Error Model

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Fit the Hougen-Watson model to the rate data using the combined error variance model.

```
[beta,R,J,CovB,MSE,S] = nlinfit(X,y,@hougen,beta0,'ErrorModel','combined');
```

Compute the 95% prediction interval for a new observation with reactant levels `[100,100,100]` using the fitted error variance model.

```
[ypred,delta] = nlpredci(@hougen,[100,100,100],beta,R,'Jacobian',J,...
                         'PredOpt','observation','ErrorModelInfo',S);
[ypred-delta,ypred+delta]
```

```
ans = 1×2

    1.3245    2.3447
```

Input Arguments

`modelfun` — Nonlinear regression model function

function handle

Nonlinear regression model function, specified as a function handle. `modelfun` must accept two input arguments, a coefficient vector and an array `X`—in that order—and return a vector of fitted response values.

For example, to specify the `hougen` nonlinear regression function, use the function handle `@hougen`.

Data Types: `function_handle`

`X` — Input values for predictions

matrix

Input values for predictions, specified as a matrix. `nlpredci` makes a prediction for the covariates in each row of `X`. There should be a column in `X` for each coefficient in the model.

Data Types: `single` | `double`

`beta` — Estimated regression coefficients

vector returned by `nlinfit`

Estimated regression coefficients, specified as the vector of fitted coefficients returned by a previous call to `nlinfit`.

Data Types: `single` | `double`

`R` — Residuals

vector returned by `nlinfit`

Residuals for the fitted `modelfun`, specified as the vector of residuals returned by a previous call to `nlinfit`.

`CovB` — Estimated variance-covariance matrix

matrix returned by `nlinfit`

Estimated variance-covariance matrix for the fitted coefficients, `beta`, specified as the variance-covariance matrix returned by a previous call to `nlinfit`.

`J` — Estimated Jacobian

matrix returned by `nlinfit`

Estimated Jacobian of the nonlinear regression model, `modelfun`, specified as the Jacobian matrix returned by a previous call to `nlinfit`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.1,'PredOpt','observation'` specifies 90% prediction intervals for new observations.

`Alpha` — Significance level
`0.05` (default) | scalar value in the range (0,1)

Significance level for the confidence interval, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). If `Alpha` has value α, then `nlpredci` returns intervals with $100\times(1–\alpha)$% confidence level.

The default confidence level is 95% (α = 0.05).

Example: `'Alpha',0.1`

Data Types: `single` | `double`

`ErrorModelInfo` — Information about error model fit
structure returned by `nlinfit`

Information about the error model fit, specified as the comma-separated pair consisting of `'ErrorModelInfo'` and a structure returned by a previous call to `nlinfit`.

`ErrorModelInfo` only has an effect on the returned prediction interval when `PredOpt` has the value `'observation'`. If you do not use `ErrorModelInfo`, then `nlpredci` assumes the error variance model is `'constant'`.

The error model structure returned by `nlinfit` has the following fields:

`ErrorModel`	Chosen error model
`ErrorParameters`	Estimated error parameters
`ErrorVariance`	Function handle that accepts an *N*-by-*p* matrix, `X`, and returns an *N*-by-1 vector of error variances using the estimated error model
`MSE`	Mean squared error
`ScheffeSimPred`	Scheffé parameter for simultaneous prediction intervals when using the estimated error model
`WeightFunction`	Logical with value `true` if you used a custom weight function previously in `nlinfit`
`FixedWeights`	Logical with value `true` if you used fixed weights previously in `nlinfit`
`RobustWeightFunction`	Logical with value `true` if you used robust fitting previously in `nlinfit`

`MSE` — Mean squared error
MSE returned by `nlinfit`

Mean squared error (MSE) for the fitted nonlinear regression model, specified as the comma-separated pair consisting of `'MSE'` and the MSE value returned by a previous call to `nlinfit`.

If you use a robust option with `nlinfit`, then you must specify the MSE when predicting new observations to properly take the robust fitting into account. If you do not specify the MSE, then `nlpredci` computes the MSE from the residuals, `R`, and does not take the robust fitting into account.

For example, if `mse` is the MSE value returned by `nlinfit`, then you can specify `'MSE',mse`.

Data Types: `single` | `double`

`PredOpt` — Prediction interval to compute
`'curve'` (default) | `'observation'`

Prediction interval to compute, specified as the comma-separated pair consisting of `'PredOpt'` and either `'curve'` or `'observation'`.

- If you specify the value `'curve'`, then `nlpredci` returns confidence intervals for the estimated curve (function value) at the observations `X`.
- If you specify the value `'observation'`, then `nlpredci` returns prediction intervals for new observations at `X`.

If you specify `'observation'` after using a robust option with `nlinfit`, then you must also specify a value for `MSE` to provide the robust estimate of the mean squared error.

Example: `'PredOpt','observation'`

`SimOpt` — Indicator for specifying simultaneous bounds
`'off'` (default) | `'on'`

Indicator for specifying simultaneous bounds, specified as the comma-separated pair consisting of `'SimOpt'` and either `'off'` or `'on'`. Use the value `'off'` to compute nonsimultaneous bounds, and `'on'` for simultaneous bounds.

`Weights` — Observation weights
vector | function handle

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of positive scalar values or a function handle. The default is no weights.

- If you specify a vector of weights, then it must have the same number of elements as the number of observations (rows) in `X`.
- If you specify a function handle for the weights, then it must accept a vector of predicted response values as input, and return a vector of real positive weights as output.

Given weights, `W`, `nlpredci` estimates the error variance at observation `i` by `mse*(1/W(i))`, where `mse` is the mean squared error value specified using `MSE`.

Example: `'Weights',@WFun`

Data Types: `double` | `single` | `function_handle`

Output Arguments

`Ypred` — Predicted responses
vector

Predicted responses, returned as a vector with the same number of rows as `X`.

`delta` — Confidence interval half-widths
vector

Confidence interval half-widths, returned as a vector with the same number of rows as `X`. By default, `delta` contains the half-widths for nonsimultaneous 95% confidence intervals for `modelfun` at the observations in `X`. You can compute the lower and upper bounds of the confidence intervals as `Ypred-delta` and `Ypred+delta`, respectively.

If `'PredOpt'` has value `'observation'`, then `delta` contains the half-widths for prediction intervals of new observations at the values in X.

More About

Confidence Intervals for Estimable Predictions

When the estimated model Jacobian is not of full rank, then it might not be possible to construct sensible confidence intervals at all prediction points. In this case, `nlpredci` still tries to construct confidence intervals for any estimable prediction points.

For example, suppose you fit the linear function $f(\mathbf{x}_i,\beta) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$ at the points in the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}.$$

The estimated Jacobian at the values in $\mathbf{X}$ is the design matrix itself, $\mathbf{J} = \mathbf{X}$. Thus, the Jacobian is not of full rank:

```
rng('default') % For reproducibility
y = randn(6,1);

linfun = @(b,x) x*b;
beta0 = [1;1;1];
X = [repmat([1 1 0],3,1); repmat([1 0 1],3,1)];

[beta,R,J]  = nlinfit(X,y,linfun,beta0);
```

```
Warning: The Jacobian at the solution is ill-conditioned, and
some model parameters may not be estimated well (they are not
identifiable).  Use caution in making predictions.
> In nlinfit at 283
```

In this example, `nlpredci` can only compute prediction intervals at points that satisfy the linear relationship

$$x_{i1} = x_{i2} + x_{i3}.$$

If you try to compute confidence intervals for predictions at nonidentifiable points, `nlpredci` returns `NaN` for the corresponding interval half-widths:

```
xpred = [1 1 1;0 1 -1;2 1 1];
[ypred,delta] = nlpredci(linfun,xpred,beta,R,'Jacobian',J)
```

```
ypred =

   -0.0035
    0.0798
   -0.0047
```

```
delta =

       NaN
    3.8102
    3.8102
```

Here, the first element of `delta` is `NaN` because the first row in `xpred` does not satisfy the required linear dependence, and is therefore not an estimable contrast.

Tips

- To compute confidence intervals for complex parameters or data, you need to split the problem into its real and imaginary parts. When calling `nlinfit`:

  1. Define your parameter vector `beta` as the concatenation of the real and imaginary parts of the original parameter vector.
  2. Concatenate the real and imaginary parts of the response vector `Y` as a single vector.
  3. Modify your model function `modelfun` to accept `X` and the purely real parameter vector, and return a concatenation of the real and imaginary parts of the fitted values.

  With the problem formulated this way, `nlinfit` computes real estimates, and confidence intervals are feasible.

Algorithms

- `nlpredci` treats `NaN` values in the residuals, `R`, or the Jacobian, `J`, as missing values, and ignores the corresponding observations.
- If the Jacobian, `J`, does not have full column rank, then some of the model parameters might be nonidentifiable. In this case, `nlpredci` tries to construct confidence intervals for estimable predictions on page 33-4244, and returns `NaN` for those that are not.

References

[1] Lane, T. P. and W. H. DuMouchel. “Simultaneous Confidence Intervals in Multiple Regression.” *The American Statistician*. Vol. 48, No. 4, 1994, pp. 315–321.

[2] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

See Also

`NonLinearModel` | `nlinfit` | `nlparci`

Introduced before R2006a

nnmf

Nonnegative matrix factorization

Syntax

```
[W,H] = nnmf(A,k)
[W,H] = nnmf(A,k,Name,Value)
[W,H,D] = nnmf(___)
```

Description

`[W,H] = nnmf(A,k)` factors the *n*-by-*m* matrix `A` into nonnegative factors `W` (*n*-by-`k`) and `H` (`k`-by-*m*). The factorization is not exact; `W*H` is a lower-rank approximation to `A`. The factors `W` and `H` minimize the root mean square residual `D` between `A` and `W*H`.

```
D = norm(A - W*H,'fro')/sqrt(n*m)
```

The factorization uses an iterative algorithm starting with random initial values for `W` and `H`. Because the root mean square residual `D` might have local minima, repeated factorizations might yield different `W` and `H`. Sometimes the algorithm converges to a solution of lower rank than *k*, which can indicate that the result is not optimal.

`[W,H] = nnmf(A,k,Name,Value)` modifies the factorization using one or more name-value pair arguments. For example, you can request repeated factorizations by setting `'Replicates'` to an integer value greater than 1.

`[W,H,D] = nnmf(___)` also returns the root mean square residual `D` using any of the input argument combinations in the previous syntaxes.

Examples

Nonnegative Rank-Two Approximation and Biplot

Load the sample data.

```
load fisheriris
```

Compute a nonnegative rank-two approximation of the measurements of the four variables in Fisher's iris data.

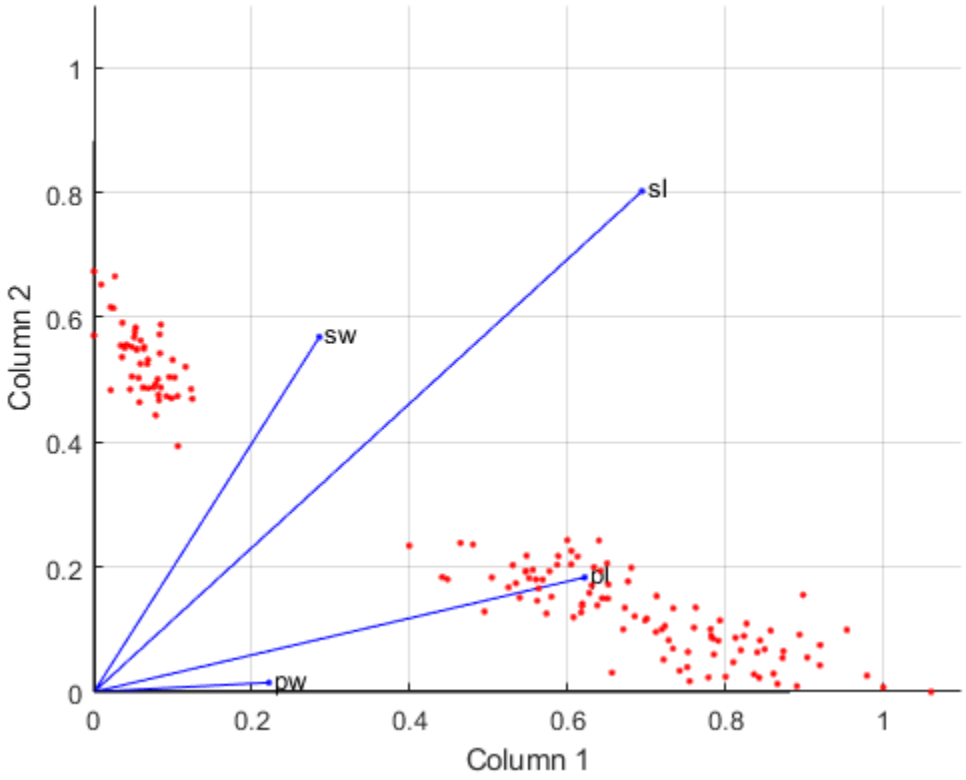
```
rng(1) % For reproducibility
[W,H] = nnmf(meas,2);
H
```

```
H = 2×4

    0.6945    0.2856    0.6220    0.2218
    0.8020    0.5683    0.1834    0.0149
```

The first and third variables in `meas` (sepal length and petal length, with coefficients 0.6945 and 0.6220, respectively) provide relatively strong weights to the first column of `W` . The first and second variables in `meas` (sepal length and sepal width, with coefficients 0.8020 and 0.5683, respectively) provide relatively strong weights to the second column of `W` .

Create a `biplot` of the data and the variables in `meas` in the column space of `W` .

```
biplot(H','Scores',W,'VarLabels',{'sl','sw','pl','pw'});
axis([0 1.1 0 1.1])
xlabel('Column 1')
ylabel('Column 2')
```



Change Algorithm

Starting from a random array X with rank 20, try a few iterations at several replicates using the multiplicative algorithm.

```
rng default % For reproducibility
X = rand(100,20)*rand(20,50);
opt = statset('MaxIter',5,'Display','final');
[W0,H0] = nnmf(X,5,'Replicates',10,...
                   'Options',opt,...
                   'Algorithm','mult');
```

```
    rep       iteration       rms resid      |delta x|
      1           5        0.560887       0.0245182
      2           5         0.66418       0.0364471
      3           5        0.609125       0.0358355
      4           5        0.608894       0.0415491
      5           5        0.619291       0.0455135
      6           5        0.621549       0.0299965
      7           5        0.640549       0.0438758
      8           5        0.673015       0.0366856
      9           5        0.606835       0.0318931
     10           5        0.633526       0.0319591
Final root mean square residual = 0.560887
```

Continue with more iterations from the best of these results using alternating least squares.

```
opt = statset('Maxiter',1000,'Display','final');
[W,H] = nnmf(X,5,'W0',W0,'H0',H0,...
                 'Options',opt,...
                 'Algorithm','als');
```

```
    rep       iteration       rms resid      |delta x|
      1          24        0.257336      0.00271859
Final root mean square residual = 0.257336
```

Input Arguments

`A` — Matrix to factorize

real matrix

Matrix to factorize, specified as a real matrix.

Example: `rand(20,30)`

Data Types: `single` | `double`

`k` — Rank of factors

positive integer

Rank of factors, specified as a positive integer. The resulting factors `W` and `H` have `k` columns and rows, respectively.

Example: 3

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `[W,H] = nnmf(A,k,'Algorithm','mult','Replicates',10)` chooses the multiplicative update algorithm and ten replicates to improve the result

`Algorithm` — Factorization algorithm

`'als'` (default) | `'mult'`

Factorization algorithm, specified as the comma-separated pair consisting of `'Algorithm'` and `'als'` (alternating least squares) or `'mult'` (a multiplicative update algorithm).

The `'als'` algorithm typically is more stable and converges in fewer iterations. Each iteration takes longer. Therefore, the default maximum is 50, which usually gives satisfactory results in internal testing.

The `'mult'` algorithm typically has faster iterations and requires more of them. The default maximum is 100. This algorithm tends to be more sensitive to starting values and, therefore, seems to benefit more from running multiple replications.

Example: `'Algorithm','mult'`

Data Types: `char` | `string`

`W0` — Initial value of `W`
n-by-k matrix

Initial value of `W`, specified as the comma-separated pair consisting of `'W0'` and an *n*-by-k matrix, where *n* is the number of rows of `A`, and `k` is the second input argument of `nnmf`.

Data Types: `single` | `double`

`H0` — Initial value of `H`
k-by-*m* matrix

Initial value of `H`, specified as the comma-separated pair consisting of `'H0'` and a k-by-*m* matrix, where `k` is the second input argument of `nnmf`, and *m* is the number of columns of `A`.

Data Types: `single` | `double`

`Options` — Algorithm options
`[]` (default) | structure returned by `statset`

Algorithm options, specified as the comma-separated pair consisting of `'Options'` and a structure returned by the `statset` function. `nnmf` uses the following fields of the options structure.

Field	Description	Values
`Display`	Level of iterative display	• `'off'` (default) — No display • `'final'` — Display of final result • `'iter'` — Iterative display of intermediate results
`MaxIter`	Maximum number of iterations	Positive integer. The default is `50` for the `'als'` algorithm and `100` for the `'mult'` algorithm. Unlike in optimization settings, reaching `MaxIter` iterations is treated as convergence.
`TolFun`	Termination tolerance on the change in size of the residual	Nonnegative value. The default is `1e-4`.
`TolX`	Termination tolerance on the relative change in the elements of `W` and `H`	Nonnegative value. The default is `1e-4`.

Field	Description	Values
`UseParallel`	Indication to compute in parallel	Logical value. The default `false` indicates not to compute in parallel, and `true` indicates to compute in parallel. Computing in parallel requires a Parallel Computing Toolbox license.
`UseSubstreams`	Type of reproducibility when computing in parallel	• `false` (default) — Do not compute reproducibly • `'mlfg6331_64'` • `'mrg32k3a'` For details, see "Reproducibility in Parallel Statistical Computations" on page 31-13.
`Streams`	A `RandStream` object or cell array of such objects	• If you do not specify `Streams`, `nnmf` uses the default stream or streams. • If `UseParallel` is `true` and `UseSubstreams` is `false`, specify a cell array of `RandStream` objects the same size as the Parallel pool. Otherwise, specify a single `RandStream` object.

Example: `'Options',statset('Display','iter','MaxIter',50)`

Data Types: `struct`

`Replicates` — Number of times to repeat factorization
1 (default) | positive integer

Number of times to repeat the factorization, specified as the comma-separated pair consisting of `'Replicates'` and a positive integer. The algorithm chooses new random starting values for `W` and `H` at each replication, except at the first replication if you specify `'W0'` and `'H0'`. If you specify a value greater than `1`, you can obtain better results by setting `Algorithm` to `'mult'`. See "Change Algorithm" on page 33-4247.

Example: 10

Data Types: `single` | `double`

Output Arguments

`W` — Nonnegative left factor of `A`
n-by-k matrix

Nonnegative left factor of `A`, returned as an *n*-by-k matrix. *n* is the number of rows of `A`, and `k` is the second input argument of `nnmf`.

`W` and `H` are normalized so that the rows of `H` have unit length. The columns of `W` are ordered by decreasing length.

`H` — Nonnegative right factor of `A`
k-by-*m* matrix

Nonnegative right factor of `A`, returned as a `k`-by-*m* matrix. `k` is the second input argument of `nnmf`, and *m* is the number of columns of `A`.

`W` and `H` are normalized so that the rows of `H` have unit length. The columns of `W` are ordered by decreasing length.

`D` — Root mean square residual
nonnegative scalar

Root mean square residual, returned as a nonnegative scalar.

```
D = norm(A - W*H,'fro')/sqrt(n*m)
```

References

[1] Berry, Michael W., Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. "Algorithms and Applications for Approximate Nonnegative Matrix Factorization." *Computational Statistics & Data Analysis* 52, no. 1 (September 2007): 155–73. https://doi.org/10.1016/j.csda.2006.11.006.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see "Run MATLAB Functions with Automatic Parallel Support" (Parallel Computing Toolbox).

See Also

`biplot` | `factoran` | `pca` | `statset`

Topics

"Perform Nonnegative Matrix Factorization" on page 15-66
"Quick Start Parallel Computing for Statistics and Machine Learning Toolbox" on page 31-2

Introduced in R2008a

nominal

(Not Recommended) Arrays for nominal data

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Description

Nominal data are discrete, nonnumeric values that do not have a natural ordering. `nominal` array objects provide efficient storage and convenient manipulation of such data, while also maintaining meaningful labels for the values.

You can manipulate `nominal` arrays like ordinary numeric arrays, by subscripting, concatenating, and reshaping. Use `nominal` arrays as grouping variables when the elements indicate the group to which an observation belongs.

Creation

Syntax

```
B = nominal(X)
B = nominal(X,labels)
B = nominal(X,labels,levels)

B = nominal(X,labels,[],edges)
```

Description

`B = nominal(X)` creates a nominal array `B` from the array `X`. `nominal` creates the levels of `B` from the sorted unique values in `X`, and creates default labels for the levels.

`B = nominal(X,labels)` labels the levels in `B` according to `labels`.

`B = nominal(X,labels,levels)` creates a nominal array with possible levels defined by `levels`.

`B = nominal(X,labels,[],edges)` creates a nominal array by binning the numeric array `X` with bin edges given by the numeric vector `edges`.

Input Arguments

`X` — Input array
numeric | logical | character | string | categorical | cell array of character vectors

Input array to convert to `nominal`, specified as a numeric, logical, character, string, or categorical array, or a cell array of character vectors. The levels of the resulting `nominal` array correspond to the sorted unique values in `X`.

`labels` — Labels for discrete levels
character array | string array | cell array of character vectors

Labels for the discrete levels, specified as a character array, string array, or cell array of character vectors. By default, `nominal` assigns labels to the levels in `B` in order of the sorted unique values in `X`.

You can include duplicate labels in `labels` to merge multiple values in `X` into a single level in `B`.

Data Types: `char` | `string` | `cell`

`levels` — Possible nominal levels
vector

Possible nominal levels for the output `nominal` array, specified as a vector whose values can be compared to those in `X` using the equality operator. `nominal` assigns labels to each level from the corresponding elements of `labels`. If `X` contains any values not present in `levels`, the levels of the corresponding elements of `B` are undefined.

`edges` — Bin edges
numeric vector

Bin edges used to create the nominal array by binning a numeric array, specified as a numeric vector. The uppermost bin includes values equal to the rightmost edge. `nominal` assigns labels to each level in the resulting nominal array from the corresponding elements of `labels`. When you specify the `edges` input argument, it must have one more element than `labels`.

Output Arguments

`B` — Nominal array
`nominal` array object

Nominal array, returned as a `nominal` array object.

By default, an element of `B` is undefined if the corresponding element of `X` is `NaN` (when `X` is numeric), an empty character vector (when `X` is a character), an empty or missing string (when `X` is a string), or undefined (when `X` is categorical). `nominal` treats such elements as undefined or missing and does not include entries for them among the possible levels. To create an explicit level for such elements instead of treating them as undefined, use the `levels` input argument and include `NaN`, the empty character vector, the empty or missing string, or an undefined element.

Object Functions

addlevels	(Not Recommended) Add levels to nominal or ordinal arrays
droplevels	(Not Recommended) Drop levels from a nominal or ordinal array
getlabels	(Not Recommended) Access nominal or ordinal array labels
getlevels	(Not Recommended) Access nominal or ordinal array levels
islevel	(Not Recommended) Determine if levels are in nominal or ordinal array
levelcounts	(Not Recommended) Element counts by level of a nominal or ordinal array
mergelevels	(Not Recommended) Merge levels of nominal or ordinal arrays
reorderlevels	(Not Recommended) Reorder levels of nominal or ordinal arrays
setlabels	(Not Recommended) Assign labels to levels of nominal or ordinal arrays

The following is a partial list of the many other MATLAB array functions you can use with nominal arrays. For a complete list, see “Other MATLAB Functions Supporting Nominal and Ordinal Arrays” on page 2-2.

double	Double-precision arrays
histogram	Histogram plot
isequal	Determine array equality
isundefined	Find undefined elements in categorical array
pie	Pie chart
summary	Print summary of table, timetable, or categorical array
times	Multiplication

Examples

Create and Label Nominal Arrays

Create nominal arrays from a cell array of character vectors and from integer data. Provide explicit labels.

Create a nominal array from a cell array of character vectors with values `'r'`, `'g'`, and `'b'`. Label these levels `'red'`, `'green'`, and `'blue'`, respectively. `nominal` assigns the labels according to the sorted (alphabetical) order of the elements in `X`.

```
X = {'r' 'b' 'g';'g' 'r' 'b';'b' 'r' 'g'}
```

```
X = 3x3 cell
    {'r'}    {'b'}    {'g'}
    {'g'}    {'r'}    {'b'}
    {'b'}    {'r'}    {'g'}
```

```
labels = {'blue','green','red'};
B = nominal(X,labels)
```

```
B = 3x3 nominal
     red        blue      green
     green      red       blue
     blue       red       green
```

Create a nominal array from integer data with values 1 to 4, merging odd and even values into two nominal levels with the labels `'odd'` and `'even'`. Merge the values by duplicating the labels.

```
X = randi([1 4],5,2)
```

```
X = 5×2

     4     1
     4     2
     1     3
     4     4
     3     4
```

```
labels = {'odd','even','odd','even'};
B = nominal(X,labels)
```

```
B = 5x2 nominal
     even      odd
     even      even
     odd       odd
     even      even
     odd       even
```

Create and Manipulate Nominal Arrays

Create a nominal array from data in a cell array.

```
X = {'r','b','g';'g','r','b';'b','r','g'};
labels = {'blue','green','red'};
colors = nominal(X,labels)
```

```
colors = 3x3 nominal
     red       blue      green
     green     red       blue
     blue      red       green
```

Identify the elements in `colors` that are members of the level `'red'`. A value of `1` in the resulting array indicates that the corresponding element of `colors` is a member of `'red'`.

```
colors == 'red'
```

```
ans = 3x3 logical array

   1   0   0
   0   1   0
   0   1   0
```

Identify the elements in `colors` that are members of either the level `'red'` or `'blue'`.

```
ismember(colors,{'red','blue'})
```

```
ans = 3x3 logical array

   1   1   0
   0   1   1
   1   1   0
```

Merge the elements of the `'red'` and `'blue'` levels into a new level labeled `'purple'`.

```
colors = mergelevels(colors,{'red','blue'},'purple')
```

```
colors = 3x3 nominal
     purple    purple    green
     green     purple    purple
     purple    purple    green
```

Display the levels of `colors`.

```
getlevels(colors)
```

```
ans = 1x2 nominal
     purple      green
```

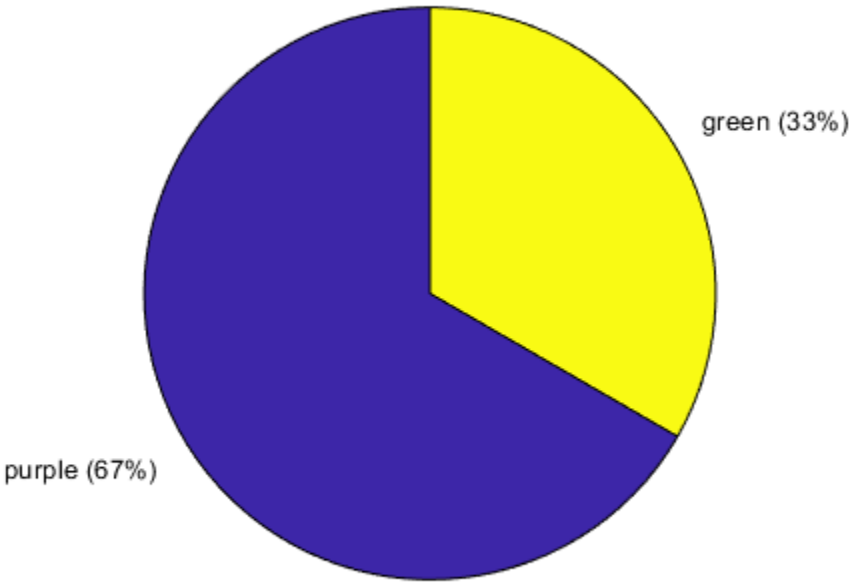
Summarize the number of elements in each level. By default, `summary` returns counts for each column of the input array.

```
summary(colors)
```

```
     purple      2      3      1
     green       1      0      2
```

Create a pie chart for the data in `colors`.

```
pie(colors)
```



See Also

`ordinal`

Topics

“Create Nominal and Ordinal Arrays” on page 2-3
“Add and Drop Category Levels” on page 2-18
“Plot Data Grouped by Category” on page 2-21
“Summary Statistics Grouped by Category” on page 2-32

“Advantages of Using Nominal and Ordinal Arrays” on page 2-38
“Index and Search Using Nominal and Ordinal Arrays” on page 2-41
“Grouping Variables” on page 2-45

Introduced in R2007a

notify

Class: `qrandstream`

Notify listeners of event

Syntax

```
notify(h,'eventname')
notify(h,'eventname',data)
```

Description

`notify(h,'eventname')` notifies listeners added to the event named `eventname` on handle object array `h` that the event is taking place. `h` is the array of handles to objects triggering the event, and `eventname` must be a character vector.

`notify(h,'eventname',data)` provides a way of encapsulating information about an event which can then be accessed by each registered listener. `data` must belong to the `event.eventdata` class.

See Also

`addlistener` | `event.EventData` | `events` | `qrandstream`

NonLinearModel class

Nonlinear regression model class

Description

An object comprising training data, model description, diagnostic information, and fitted coefficients for a nonlinear regression. Predict model responses with the `predict` or `feval` methods.

Construction

Create a `NonLinearModel` object using `fitnlm`.

Properties

`CoefficientCovariance` — Covariance matrix of coefficient estimates

numeric matrix

This property is read-only.

Covariance matrix of coefficient estimates, specified as a *p*-by-*p* matrix of numeric values. *p* is the number of coefficients in the fitted model.

For details, see “Coefficient Standard Errors and Confidence Intervals” on page 11-58.

Data Types: `single` | `double`

`CoefficientNames` — Coefficient names

cell array of character vectors

This property is read-only.

Coefficient names, specified as a cell array of character vectors, each containing the name of the corresponding term.

Data Types: `cell`

`Coefficients` — Coefficient values

table

This property is read-only.

Coefficient values, specified as a table. `Coefficients` contains one row for each coefficient and these columns:

- `Estimate` — Estimated coefficient value
- `SE` — Standard error of the estimate
- `tStat` — *t*-statistic for a test that the coefficient is zero
- `pValue` — *p*-value for the *t*-statistic

Use `anova` (only for a linear regression model) or `coefTest` to perform other tests on the coefficients. Use `coefCI` to find the confidence intervals of the coefficient estimates.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the estimated coefficient vector in the model `mdl`:

```
beta = mdl.Coefficients.Estimate
```

Data Types: `table`

`Diagnostics` — Diagnostic information
table

This property is read-only.

Diagnostic information for the model, specified as a table. Diagnostics can help identify outliers and influential observations. `Diagnostics` contains the following fields.

Field	Meaning	Utility
`Leverage`	Diagonal elements of `HatMatrix`	Leverage indicates to what extent the predicted value for an observation is determined by the observed value for that observation. A value close to `1` indicates that the prediction is largely determined by that observation, with little contribution from the other observations. A value close to `0` indicates the fit is largely determined by the other observations. For a model with `P` coefficients and `N` observations, the average value of `Leverage` is `P/N`. An observation with `Leverage` larger than `2*P/N` can be regarded as having high leverage.
`CooksDistance`	Cook's measure of scaled change in fitted values	`CooksDistance` is a measure of scaled change in fitted values. An observation with `CooksDistance` larger than three times the mean Cook's distance can be an outlier.
`HatMatrix`	Projection matrix to compute fitted from observed responses	`HatMatrix` is an `N`-by-`N` matrix such that `Fitted = HatMatrix*Y`, where `Y` is the response vector and `Fitted` is the vector of fitted response values.

Data Types: `table`

`DFE` — Degrees of freedom for error
positive integer

This property is read-only.

Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients, specified as a positive integer.

Data Types: `double`

`Fitted` — Fitted response values based on input data
numeric vector

This property is read-only.

Fitted (predicted) values based on the input data, specified as a numeric vector. `fitnlm` attempts to make `Fitted` as close as possible to the response data.

Data Types: `single` | `double`

`Formula` — Model information

`NonLinearFormula` object

This property is read-only.

Model information, specified as a `NonLinearFormula` object.

Display the formula of the fitted model `mdl` by using dot notation.

```
mdl.Formula
```

`Iterative` — Information about fitting process

structure

This property is read-only.

Information about the fitting process, specified as a structure with the following fields:

- `InitialCoefs` — Initial coefficient values (the `beta0` vector)
- `IterOpts` — Options included in the `Options` name-value pair argument for `fitnlm`.

Data Types: `struct`

`LogLikelihood` — Loglikelihood

numeric value

This property is read-only.

Loglikelihood of the model distribution at the response values, specified as a numeric value. The mean is fitted from the model, and other parameters are estimated as part of the model fit.

Data Types: `single` | `double`

`ModelCriterion` — Criterion for model comparison

structure

This property is read-only.

Criterion for model comparison, specified as a structure with these fields:

- `AIC` — Akaike information criterion. `AIC = –2*logL + 2*m`, where `logL` is the loglikelihood and `m` is the number of estimated parameters.
- `AICc` — Akaike information criterion corrected for the sample size. `AICc = AIC + (2*m*(m + 1))/(n – m – 1)`, where `n` is the number of observations.
- `BIC` — Bayesian information criterion. `BIC = –2*logL + m*log(n)`.
- `CAIC` — Consistent Akaike information criterion. `CAIC = –2*logL + m*(log(n) + 1)`.

Information criteria are model selection tools that you can use to compare multiple models fit to the same data. These criteria are likelihood-based measures of model fit that include a penalty for complexity (specifically, the number of parameters). Different information criteria are distinguished by the form of the penalty.

When you compare multiple models, the model with the lowest information criterion value is the best-fitting model. The best-fitting model can vary depending on the criterion used for model comparison.

To obtain any of the criterion values as a scalar, index into the property using dot notation. For example, obtain the AIC value `aic` in the model `mdl`:

```
aic = mdl.ModelCriterion.AIC
```

Data Types: `struct`

`MSE` — Mean squared error
numeric value

This property is read-only.

Mean squared error, specified as a numeric value. The mean squared error is an estimate of the variance of the error term in the model.

Data Types: `single` | `double`

`NumCoefficients` — Number of model coefficients
positive integer

This property is read-only.

Number of coefficients in the fitted model, specified as a positive integer. `NumCoefficients` is the same as `NumEstimatedCoefficients` for `NonLinearModel` objects. `NumEstimatedCoefficients` is equal to the degrees of freedom for regression.

Data Types: `double`

`NumEstimatedCoefficients` — Number of estimated coefficients
positive integer

This property is read-only.

Number of estimated coefficients in the fitted model, specified as a positive integer. `NumEstimatedCoefficients` is the same as `NumCoefficients` for `NonLinearModel` objects. `NumEstimatedCoefficients` is equal to the degrees of freedom for regression.

Data Types: `double`

`NumPredictors` — Number of predictor variables
positive integer

This property is read-only.

Number of predictor variables used to fit the model, specified as a positive integer.

Data Types: `double`

`NumVariables` — Number of variables
positive integer

This property is read-only.

Number of variables in the input data, specified as a positive integer. `NumVariables` is the number of variables in the original table or dataset, or the total number of columns in the predictor matrix and response vector.

`NumVariables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `double`

`ObservationInfo` — Observation information

table

This property is read-only.

Observation information, specified as an *n*-by-4 table, where *n* is equal to the number of rows of input data. `ObservationInfo` contains the columns described in this table.

Column	Description
`Weights`	Observation weights, specified as a numeric value. The default value is `1`.
`Excluded`	Indicator of excluded observations, specified as a logical value. The value is `true` if you exclude the observation from the fit by using the `'Exclude'` name-value pair argument.
`Missing`	Indicator of missing observations, specified as a logical value. The value is `true` if the observation is missing.
`Subset`	Indicator of whether or not the fitting function uses the observation, specified as a logical value. The value is `true` if the observation is not excluded or missing, meaning the fitting function uses the observation.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the weight vector `w` of the model `mdl`:

```
w = mdl.ObservationInfo.Weights
```

Data Types: `table`

`ObservationNames` — Observation names

cell array of character vectors

This property is read-only.

Observation names, specified as a cell array of character vectors containing the names of the observations used in the fit.

- If the fit is based on a table or dataset containing observation names, `ObservationNames` uses those names.
- Otherwise, `ObservationNames` is an empty cell array.

Data Types: `cell`

`PredictorNames` — Names of predictors used to fit model

cell array of character vectors

This property is read-only.

Names of predictors used to fit the model, specified as a cell array of character vectors.

Data Types: `cell`

`Residuals` — Residuals for fitted model
table

This property is read-only.

Residuals for the fitted model, specified as a table that contains one row for each observation and the columns described in this table.

Column	Description
`Raw`	Observed minus fitted values
`Pearson`	Raw residuals divided by the root mean squared error (RMSE)
`Standardized`	Raw residuals divided by their estimated standard deviation
`Studentized`	Raw residual divided by an independent estimate of the residual standard deviation. The residual for observation *i* is divided by an estimate of the error standard deviation based on all observations except observation *i*.

Use `plotResiduals` to create a plot of the residuals. For details, see “Residuals” on page 11-80.

Rows not used in the fit because of missing values (in `ObservationInfo.Missing`) or excluded values (in `ObservationInfo.Excluded`) contain `NaN` values.

To obtain any of these columns as a vector, index into the property using dot notation. For example, obtain the raw residual vector `r` in the model `mdl`:

```
r = mdl.Residuals.Raw
```

Data Types: `table`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`RMSE` — Root mean squared error
numeric value

This property is read-only.

Root mean squared error, specified as a numeric value. The root mean squared error is an estimate of the standard deviation of the error term in the model.

Data Types: `single` | `double`

`Robust` — Robust fit information
structure

This property is read-only.

Robust fit information, specified as a structure with the following fields:

Field	Description
`WgtFun`	Robust weighting function, such as `'bisquare'` (see `robustfit`)
`Tune`	Value specified for tuning parameter (can be `[]`)
`Weights`	Vector of weights used in final iteration of robust fit

This structure is empty unless `fitnlm` constructed the model using robust regression.

Data Types: `struct`

Rsquared — R-squared value for model

structure

This property is read-only.

R-squared value for the model, specified as a structure with two fields:

- `Ordinary` — Ordinary (unadjusted) R-squared
- `Adjusted` — R-squared adjusted for the number of coefficients

The R-squared value is the proportion of the total sum of squares explained by the model. The ordinary R-squared value relates to the `SSR` and `SST` properties:

```
Rsquared = SSR/SST,
```

where `SST` is the total sum of squares, and `SSR` is the regression sum of squares.

For details, see “Coefficient of Determination (R-Squared)” on page 11-61.

To obtain either of these values as a scalar, index into the property using dot notation. For example, obtain the adjusted R-squared value in the model `mdl`:

```
r2 = mdl.Rsquared.Adjusted
```

Data Types: `struct`

SSE — Sum of squared errors

numeric value

This property is read-only.

Sum of squared errors (residuals), specified as a numeric value.

Data Types: `single` | `double`

SSR — Regression sum of squares

numeric value

This property is read-only.

Regression sum of squares, specified as a numeric value. The regression sum of squares is equal to the sum of squared deviations of the fitted values from their mean.

Data Types: `single` | `double`

SST — Total sum of squares

numeric value

This property is read-only.

Total sum of squares, specified as a numeric value. The total sum of squares is equal to the sum of squared deviations of the response vector `y` from the `mean(y)`.

Data Types: `single` | `double`

`VariableInfo` — Information about variables

table

This property is read-only.

Information about variables contained in `Variables`, specified as a table with one row for each variable and the columns described in this table.

Column	Description
`Class`	Variable class, specified as a cell array of character vectors, such as `'double'` and `'categorical'`
`Range`	Variable range, specified as a cell array of vectors • Continuous variable — Two-element vector `[min,max]`, the minimum and maximum values • Categorical variable — Vector of distinct variable values
`InModel`	Indicator of which variables are in the fitted model, specified as a logical vector. The value is `true` if the model includes the variable.
`IsCategorical`	Indicator of categorical variables, specified as a logical vector. The value is `true` if the variable is categorical.

`VariableInfo` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

`VariableNames` — Names of variables

cell array of character vectors

This property is read-only.

Names of variables, specified as a cell array of character vectors.

- If the fit is based on a table or dataset, this property provides the names of the variables in the table or dataset.
- If the fit is based on a predictor matrix and response vector, `VariableNames` contains the values specified by the `'VarNames'` name-value pair argument of the fitting method. The default value of `'VarNames'` is `{'x1','x2',...,'xn','y'}`.

`VariableNames` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `cell`

`Variables` — Input data

table

This property is read-only.

Input data, specified as a table. `Variables` contains both predictor and response values. If the fit is based on a table or dataset array, `Variables` contains all the data from the table or dataset array. Otherwise, `Variables` is a table created from the input data matrix `X` and the response vector `y`.

`Variables` also includes any variables that are not used to fit the model as predictors or as the response.

Data Types: `table`

Object Functions

coefCI	Confidence intervals of coefficient estimates of nonlinear regression model
coefTest	Linear hypothesis test on nonlinear regression model coefficients
feval	Evaluate nonlinear regression model prediction
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
plotDiagnostics	Plot diagnostics of nonlinear regression model
plotResiduals	Plot residuals of nonlinear regression model
plotSlice	Plot of slices through fitted nonlinear regression surface
predict	Predict response of nonlinear regression model
random	Simulate responses for nonlinear regression model

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Fit a Nonlinear Regression Model

Fit a nonlinear regression model for auto mileage based on the `carbig` data. Predict the mileage of an average car.

Load the sample data. Create a matrix `X` containing the measurements for the horsepower (`Horsepower`) and weight (`Weight`) of each car. Create a vector `y` containing the response values in miles per gallon (`MPG`).

```
load carbig
X = [Horsepower,Weight];
y = MPG;
```

Fit a nonlinear regression model.

```
modelfun = @(b,x)b(1) + b(2)*x(:,1).^b(3) + ...
    b(4)*x(:,2).^b(5);
beta0 = [-50 500 -1 500 -1];
mdl = fitnlm(X,y,modelfun,beta0)
```

```
mdl =
Nonlinear regression model:
    y ~ b1 + b2*x1^b3 + b4*x2^b5

Estimated Coefficients:
          Estimate      SE        tStat       pValue
          ________    _______    ________    ________

    b1     -49.383     119.97    -0.41164     0.68083
    b2      376.43     567.05     0.66384     0.50719
    b3    -0.78193    0.47168     -1.6578    0.098177
    b4      422.37     776.02     0.54428     0.58656
    b5    -0.24127    0.48325    -0.49926     0.61788


Number of observations: 392, Error degrees of freedom: 387
Root Mean Squared Error: 3.96
R-Squared: 0.745,  Adjusted R-Squared 0.743
F-statistic vs. constant model: 283, p-value = 1.79e-113
```

Find the predicted mileage of an average car. Because the sample data contains some missing (`NaN`) observations, compute the mean using `mean` with the `'omitnan'` option.

```
Xnew = mean(X,'omitnan')
```

```
Xnew = 1×2
10^3 ×

    0.1051    2.9794
```

```
MPGnew = predict(mdl,Xnew)
```

```
MPGnew = 21.8073
```

More About

Hat Matrix

The hat matrix H is defined in terms of the data matrix X and the Jacobian matrix J:

$$J_{i,j} = \left.\frac{\partial f}{\partial \beta_j}\right|_{x_i,\beta}$$

Here f is the nonlinear model function, and β is the vector of model coefficients.

The Hat Matrix H is

$H = J(J^TJ)^{-1}J^T$.

The diagonal elements H_{ii} satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where n is the number of observations (rows of X), and p is the number of coefficients in the regression model.

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

Cook's Distance

The Cook's distance D_i of observation i is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\,MSE},$$

where

- $\hat{y}_j$ is the jth fitted response value.
- $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook's distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p\,MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right),$$

where e_i is the ith residual.

See Also

`GeneralizedLinearModel` | `LinearModel` | `fitnlm` | `nlinfit` | `predict`

Topics

"Nonlinear Regression Workflow" on page 13-12
"Nonlinear Regression" on page 13-2

normcdf

Normal cumulative distribution function

Syntax

```
p = normcdf(x)
p = normcdf(x,mu)
p = normcdf(x,mu,sigma)

[p,pLo,pUp] = normcdf(x,mu,sigma,pCov)
[p,pLo,pUp] = normcdf(x,mu,sigma,pCov,alpha)

___ = normcdf(___,'upper')
```

Description

`p = normcdf(x)` returns the cumulative distribution function (cdf) of the standard normal distribution, evaluated at the values in `x`.

`p = normcdf(x,mu)` returns the cdf of the normal distribution with mean `mu` and unit standard deviation, evaluated at the values in `x`.

`p = normcdf(x,mu,sigma)` returns the cdf of the normal distribution with mean `mu` and standard deviation `sigma`, evaluated at the values in `x`.

`[p,pLo,pUp] = normcdf(x,mu,sigma,pCov)` also returns the 95% confidence bounds `[pLo,pUp]` of `p` when `mu` and `sigma` are estimates. `pCov` is the covariance matrix of the estimated parameters.

`[p,pLo,pUp] = normcdf(x,mu,sigma,pCov,alpha)` specifies the confidence level for the confidence interval `[pLo,pUp]` to be `100(1–alpha)`%.

`___ = normcdf(___,'upper')` returns the complement of the cdf, evaluated at the values in `x`, using an algorithm that more accurately computes the extreme upper-tail probabilities. `'upper'` can follow any of the input arguments in the previous syntaxes.

Examples

Standard Normal Distribution cdf

Compute the probability that an observation from a standard normal distribution falls on the interval `[–1 1]`.

```
p = normcdf([-1 1]);
p(2)-p(1)
```

```
ans = 0.6827
```

About 68% of the observations from a normal distribution fall within one standard deviation of the mean 0.

Normal Distribution cdf

Compute the cdf values evaluated at the values in `x` for the normal distribution with mean `mu` and standard deviation `sigma`.

```
x = [-2,-1,0,1,2];
mu = 2;
sigma = 1;
p = normcdf(x,mu,sigma)
```

```
p = 1×5

    0.0000    0.0013    0.0228    0.1587    0.5000
```

Compute the cdf values evaluated at zero for various normal distributions with different mean parameters.

```
mu = [-2,-1,0,1,2];
sigma = 1;
p = normcdf(0,mu,sigma)
```

```
p = 1×5

    0.9772    0.8413    0.5000    0.1587    0.0228
```

Confidence Interval of Normal cdf Value

Find the maximum likelihood estimates (MLEs) of the normal distribution parameters, and then find the confidence interval of the corresponding cdf value.

Generate 1000 normal random numbers from the normal distribution with mean 5 and standard deviation 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = normrnd(5,2,n,1);
```

Find the MLEs for the distribution parameters (mean and standard deviation) by using `mle`.

```
phat = mle(x)
```

```
phat = 1×2

    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `normlike`. The function `normlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = normlike([muHat,sigmaHat],x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the cdf value at zero and its 95% confidence interval.

```
[p,pLo,pUp] = normcdf(0,muHat,sigmaHat,pCov)
```

```
p = 0.0067

pLo = 0.0047

pUp = 0.0095
```

`p` is the cdf value using the normal distribution with the parameters `muHat` and `sigmaHat`. The interval `[pLo,pUp]` is the 95% confidence interval of the cdf evaluated at 0, considering the uncertainty of `muHat` and `sigmaHat` using `pCov`. The 95% confidence interval means the probability that `[pLo,pUp]` contains the true cdf value is 0.95.

Complementary cdf (Tail Distribution)

Determine the probability that an observation from a standard normal distribution will fall on the interval `[10,Inf]`.

```
p1 = 1 - normcdf(10)
```

```
p1 = 0
```

`normcdf(10)` is nearly 1, so `p1` becomes 0. Specify `'upper'` so that `normcdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = normcdf(10,'upper')
```

```
p2 = 7.6199e-24
```

You can also use `'upper'` to compute a right-tailed *p*-value.

Test for Normal Distribution Using Function Handle

Use the probability distribution function `normcdf` as a function handle in the chi-square goodness-of-fit test (`chi2gof`).

Test the null hypothesis that the sample data in the input vector `x` comes from a normal distribution with parameters μ and σ equal to the mean (`mean`) and standard deviation (`std`) of the sample data, respectively.

```
rng('default') % For reproducibility
x = normrnd(50,5,100,1);
h = chi2gof(x,'cdf',{@normcdf,mean(x),std(x)})
```

```
h = 0
```

The returned result `h = 0` indicates that `chi2gof` does not reject the null hypothesis at the default 5% significance level.

Input Arguments

`x` — Values at which to evaluate cdf

scalar value | array of scalar values

Values at which to evaluate the cdf, specified as a scalar value or an array of scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `x` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[-1,0,3,4]`

Data Types: `single` | `double`

`mu` — Mean

`0` (default) | scalar value | array of scalar values

Mean of the normal distribution, specified as a scalar value or an array of scalar values.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `mu` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation

`1` (default) | nonnegative scalar value | array of nonnegative scalar values

Standard deviation of the normal distribution, specified as a nonnegative scalar value or an array of nonnegative scalar values.

If `sigma` is zero, then the output `p` is either 0 or 1. `p` is 0 if `x` is smaller than `mu`, or 1 otherwise.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `sigma` must be a scalar value.

To evaluate the cdf at multiple values, specify `x` using an array. To evaluate the cdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normcdf` expands each scalar

input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

`pCov` — Covariance of estimates
2-by-2 numeric matrix

Covariance of the estimates `mu` and `sigma`, specified as a 2-by-2 matrix.

If you specify `pCov` to compute the confidence interval `[pLo,pUp]`, then `x`, `mu`, and `sigma` must be scalar values.

You can estimate `mu` and `sigma` by using `mle`, and estimate the covariance of `mu` and `sigma` by using `normlike`. For an example, see “Confidence Interval of Normal cdf Value” on page 33-4271.

Data Types: `single` | `double`

`alpha` — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence interval does not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

Output Arguments

`p` — cdf values
scalar value | array of scalar values

cdf values, evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `p` is the same size as `x`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `p` is the cdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

`pLo` — Lower confidence bound for `p`
scalar value | array of scalar values

Lower confidence bound for `p`, returned as a scalar value or an array of scalar values. `pLo` has the same size as `p`.

`pUp` — Upper confidence bound for `p`
scalar value | array of scalar values

Upper confidence bound for `p`, returned as a scalar value or an array of scalar values. `pUp` has the same size as `p`.

More About

Normal Distribution

The normal distribution is a two-parameter family of curves. The first parameter, μ, is the mean. The second parameter, σ, is the standard deviation.

The standard normal distribution has zero mean and unit standard deviation.

The normal cumulative distribution function (cdf) is

$$p = F(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt, \quad \text{for } x \in \mathbb{R}.$$

p is the probability that a single observation from a normal distribution with parameters μ and σ falls in the interval (-∞,x].

Algorithms

- The `normcdf` function uses the complementary error function `erfc`. The relationship between `normcdf` and `erfc` is

  $$\text{normcdf}(x) = \frac{1}{2}\text{erfc}\left(-\frac{x}{\sqrt{2}}\right).$$

  The complementary error function `erfc(x)` is defined as

  $$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt.$$

- The `normcdf` function computes confidence bounds for `p` by using the delta method. `normcdf(x,mu,sigma)` is equivalent to `normcdf((x–mu)/sigma,0,1)`. Therefore, the `normcdf` function estimates the variance of `(x–mu)/sigma` using the covariance matrix of `mu` and `sigma` by the delta method, and finds the confidence bounds of `(x–mu)/sigma` using the estimates of this variance. Then, the function transforms the bounds to the scale of `p`. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pCov` from large samples.

Alternative Functionality

- `normcdf` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, create a `NormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `normcdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

References

[1] Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1964.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`NormalDistribution` | `cdf` | `erfc` | `normfit` | `norminv` | `normlike` | `normpdf`

Topics

“Normal Distribution” on page B-119

Introduced before R2006a

normfit

Normal parameter estimates

Syntax

```
[muHat,sigmaHat] = normfit(x)

[muHat,sigmaHat,muCI,sigmaCI] = normfit(x)
[muHat,sigmaHat,muCI,sigmaCI] = normfit(x,alpha)

[ ___ ] = normfit(x,alpha,censoring)
[ ___ ] = normfit(x,alpha,censoring,freq)
[ ___ ] = normfit(x,alpha,censoring,freq,options)
```

Description

`[muHat,sigmaHat] = normfit(x)` returns estimates of normal distribution parameters (the mean `muHat` and standard deviation `sigmaHat`), given the sample data in `x`. `muHat` is the sample mean, and `sigmaHat` is the square root of the unbiased estimator of the variance.

`[muHat,sigmaHat,muCI,sigmaCI] = normfit(x)` also returns 95% confidence intervals for the parameter estimates on the mean and standard deviation in the arrays `muCI` and `sigmaCI`, respectively.

`[muHat,sigmaHat,muCI,sigmaCI] = normfit(x,alpha)` specifies the confidence level for the confidence intervals to be `100(1–alpha)`%.

`[ ___ ] = normfit(x,alpha,censoring)` specifies whether each value in `x` is right-censored or not. Use the logical vector `censoring` in which 1 indicates observations that are right-censored and 0 indicates observations that are fully observed. With censoring, `muHat` and `sigmaHat` are the maximum likelihood estimates (MLEs).

`[ ___ ] = normfit(x,alpha,censoring,freq)` specifies the frequency or weights of observations.

`[ ___ ] = normfit(x,alpha,censoring,freq,options)` specifies optimization options for the iterative algorithm `normfit` to use to compute MLEs with censoring. Create `options` by using the function `statset`.

You can pass in `[]` for `alpha`, `censoring`, and `freq` to use their default values.

Examples

Estimate Parameters and Confidence Intervals

Generate 1000 normal random numbers from the normal distribution with mean 3 and standard deviation 5.

```
rng('default') % For reproducibility
x = normrnd(3,5,[1000,1]);
```

Find the parameter estimates and the 99% confidence intervals.

```
[muHat,sigmaHat,muCI,sigmaCI] = normfit(x,0.01)
```

```
muHat = 2.8368

sigmaHat = 4.9948

muCI = 2×1

    2.4292
    3.2445


sigmaCI = 2×1

    4.7218
    5.2989
```

`muHat` is the sample mean, and `sigmaHat` is the square root of the unbiased estimator of the variance. `muCI` and `sigmaCI` contain the 99% confidence intervals of the mean and standard deviation parameters, respectively. The first row is the lower bound, and the second row is the upper bound.

Change Algorithm Options

Find the MLEs of a data set with censoring by using `normfit`. Use `statset` to specify the iterative algorithm options that `normfit` uses to compute MLEs for censored data, and then find the MLEs again.

Load the sample data.

```
load lightbulb
```

The first column of the data contains the lifetime (in hours) of two types of bulbs. The second column contains the binary variable indicating whether the bulb is fluorescent or incandescent. 1 indicates that the bulb is fluorescent, and 0 indicates that the bulb is incandescent. The third column contains the censorship information, where 0 indicates the bulb is observed until failure, and 1 indicates the item (bulb) is censored.

Find the indices for fluorescent bulbs.

```
idx = find(lightbulb(:,2) == 0);
```

Assume that the lifetime follows the normal distribution, and find the MLEs of the normal distribution parameters. The second input argument of `normfit` specifies the confidence level. Pass in `[]` to use its default value 0.05. The third input argument specifies the censorship information.

```
censoring = lightbulb(idx,3) == 1;
[muHat1,sigmaHat1] = normfit(lightbulb(idx,1),[],censoring)
```

```
muHat1 = 9.4966e+03

sigmaHat1 = 3.0640e+03
```

Display the default algorithm parameters that `normfit` uses to estimate the normal distribution parameters.

```
statset('normfit')
```

```
ans = struct with fields:
          Display: 'off'
      MaxFunEvals: 200
          MaxIter: 100
           TolBnd: 1.0000e-06
           TolFun: 1.0000e-08
       TolTypeFun: []
             TolX: 1.0000e-08
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

Save the options using a different name. Change how the results are displayed (`Display`) and the termination tolerance for the objective function (`TolFun`).

```
options = statset('normfit');
options.Display = 'final';
options.TolFun = 1e-10;
```

Alternatively, you can specify algorithm parameters by using the name-value pair arguments of the function `statset`.

```
options = statset('Display','final','TolFun',1e-10);
```

Find the MLEs with the new algorithm parameters.

```
[muHat2,sigmaHat2] = normfit(lightbulb(idx,1),[],censoring,[],options)
```

```
Successful convergence: Norm of gradient less than OPTIONS.TolFun

muHat2 = 9.4966e+03

sigmaHat2 = 3.0640e+03
```

`normfit` displays a report on the final iteration.

Convert Unbiased Estimator to MLE

The function `normfit` finds the sample mean and the square root of the unbiased estimator of the variance with no censoring. The sample mean is equal to the MLE of the mean parameter, but the

square root of the unbiased estimator of the variance is not equal to the MLE of the standard deviation parameter.

Find the normal distribution parameters by using `normfit`, convert them into MLEs, and then compare the negative log likelihoods of the estimates by using `normlike`.

Generate 100 normal random numbers from the standard normal distribution.

```
rng('default') % For reproducibility
n = 100;
x = normrnd(0,1,[n,1]);
```

Find the sample mean and the square root of the unbiased estimator of the variance.

```
[muHat,sigmaHat] = normfit(x)
```

```
muHat = 0.1231
```

```
sigmaHat = 1.1624
```

Convert the square root of the unbiased estimator of the variance into the MLE of the standard deviation parameter.

```
sigmaHat_MLE = sqrt((n-1)/n)*sigmaHat
```

```
sigmaHat_MLE = 1.1566
```

The difference between `sigmaHat` and `sigmaHat_MLE` is negligible for large `n`.

Alternatively, you can find the MLEs by using the function `mle`.

```
phat = mle(x)
```

```
phat = 1×2

    0.1231    1.1566
```

`phat(1)` and `phat(2)` are the MLEs of the mean and the standard deviation parameter, respectively.

Confirm that the log likelihood of the MLEs (`muHat` and `sigmaHat_MLE`) is greater than the log likelihood of the unbiased estimators (`muHat` and `sigmaHat`) by using the `normlike` function.

```
logL = -normlike([muHat,sigmaHat],x)
```

```
logL = -156.4424
```

```
logL_MLE = -normlike([muHat,sigmaHat_MLE],x)
```

```
logL_MLE = -156.4399
```

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

`alpha` — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence intervals, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence intervals do not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

`censoring` — Indicator for censoring
array of 0s (default) | logical vector

Indicator for the censoring of each value in `x`, specified as a logical vector of the same size as `x`. Use 1 for observations that are right-censored and 0 for observations that are fully observed.

The default is an array of 0s, meaning that all observations are fully observed.

Data Types: `logical`

`freq` — Frequency or weights of observations
array of 1s (default) | nonnegative vector

Frequency or weights of observations, specified as a nonnegative vector that is the same size as `x`. The `freq` input argument typically contains nonnegative integer counts for the corresponding elements in `x`, but can contain any nonnegative values.

To obtain the weighted MLEs for a data set with censoring, specify weights of observations, normalized to the number of observations in `x`.

The default is an array of 1s, meaning one observation per element of `x`.

Data Types: `single` | `double`

`options` — Optimization options
`statset('normfit')` (default) | structure

Optimization options, specified as a structure. `options` determines the control parameters for the iterative algorithm that `normfit` uses to compute MLEs for censored data.

Create `options` by using the function `statset` or by creating a structure array containing the fields and values described in this table.

Field Name	Value	Default Value
`Display`	Amount of information displayed by the algorithm. • `'off'` — Displays no information. • `'final'` — Displays the final output.	`'off'`
`MaxFunEvals`	Maximum number of objective function evaluations allowed, specified as a positive integer.	`200`

Field Name	Value	Default Value
`MaxIter`	Maximum number of iterations allowed, specified as a positive integer.	`100`
`TolBnd`	Lower bound of the standard deviation parameter estimate, specified as a positive scalar. The bounds for the mean and standard deviation parameter estimates are `[-Inf,Inf]` and `[TolBnd,Inf]`, respectively.	`1e-6`
`TolFun`	Termination tolerance for the objective function value, specified as a positive scalar.	`1e-8`
`TolX`	Termination tolerance for the parameters, specified as a positive scalar.	`1e-8`

You can also enter `statset('normfit')` in the Command Window to see the names and default values of the fields that `normfit` accepts in the `options` structure.

Example: `statset('Display','final','MaxIter',1000)` specifies to display the final information of the iterative algorithm results, and change the maximum number of iterations allowed to 1000.

Data Types: `struct`

Output Arguments

`muHat` — Estimate of mean

scalar

Estimate of the mean parameter of the normal distribution, returned as a scalar.

- With no censoring, `muHat` is the sample mean.
- With censoring, `muHat` is the MLE. To compute the weighted MLE, specify the weights of observations by using `freq`.

`sigmaHat` — Estimate of standard deviation

scalar

Estimate of the standard deviation parameter of the normal distribution, returned as a scalar.

- With no censoring, `sigmaHat` is the square root of the unbiased estimator of the variance. To compute the MLE with no censoring, use the `mle` function.
- With censoring, `sigmaHat` is the MLE. To compute the weighted MLE, specify the weights of observations by using `freq`.

`muCI` — Confidence interval for mean

2-by-1 column vector

Confidence interval for the mean parameter of the normal distribution, returned as a 2-by-1 column vector containing the lower and upper bounds of the `100(1-alpha)`% confidence interval.

The first and second rows correspond to the lower and upper bounds of the confidence intervals, respectively.

`sigmaCI` — Confidence interval for standard deviation
2-by-1 column vector

Confidence interval for the standard deviation parameter of the normal distribution, returned as a 2-by-1 column vector containing the lower and upper bounds of the `100(1–alpha)`% confidence interval.

The first and second rows correspond to the lower and upper bounds of the confidence intervals, respectively.

Algorithms

To compute the confidence intervals, `normfit` uses the exact method for uncensored data and the Wald method for censored data. The exact method provides exact coverage for uncensored samples based on *t* and chi-square distributions.

Alternative Functionality

`normfit` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic functions `mle`, `fitdist`, and `paramci` and the **Distribution Fitter** app, which support various probability distributions.

- `mle` returns MLEs and the confidence intervals of MLEs for the parameters of various probability distributions. You can specify the probability distribution name or a custom probability density function.
- Create a `NormalDistribution` probability distribution object by fitting the distribution to data using the `fitdist` function or the **Distribution Fitter** app. The object properties `mu` and `sigma` store the parameter estimates. To obtain the confidence intervals for the parameter estimates, pass the object to `paramci`.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[3] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

Distribution Fitter | `NormalDistribution` | `fitdist` | `mle` | `normcdf` | `norminv` | `normlike` | `paramci` | `statset`

Topics

“Normal Distribution” on page B-119

Introduced before R2006a

norminv

Normal inverse cumulative distribution function

Syntax

```
x = norminv(p)
x = norminv(p,mu)
x = norminv(p,mu,sigma)

[x,xLo,xUp] = norminv(p,mu,sigma,pCov)
[x,xLo,xUp] = norminv(p,mu,sigma,pCov,alpha)
```

Description

`x = norminv(p)` returns the inverse of the standard normal cumulative distribution function (cdf), evaluated at the probability values in `p`.

`x = norminv(p,mu)` returns the inverse of the normal cdf with mean `mu` and the unit standard deviation, evaluated at the probability values in `p`.

`x = norminv(p,mu,sigma)` returns the inverse of the normal cdf with mean `mu` and standard deviation `sigma`, evaluated at the probability values in `p`.

`[x,xLo,xUp] = norminv(p,mu,sigma,pCov)` also returns the 95% confidence bounds [`xLo`,`xUp`] of `x` when `mu` and `sigma` are estimates. `pCov` is the covariance matrix of the estimated parameters.

`[x,xLo,xUp] = norminv(p,mu,sigma,pCov,alpha)` specifies the confidence level for the confidence interval [`xLo`,`xUp`] to be `100(1–alpha)`%.

Examples

Inverse of Standard Normal cdf

Find an interval that contains 95% of the values from a standard normal distribution.

```
x = norminv([0.025 0.975])
```

```
x = 1×2

   -1.9600    1.9600
```

Note that the interval `x` is not the only such interval, but it is the shortest. Find another interval.

```
xl = norminv([0.01 0.96])
```

```
xl = 1×2

   -2.3263    1.7507
```

The interval `x1` also contains 95% of the probability, but it is longer than `x`.

Inverse of Normal Distribution cdf

Compute the inverse of cdf values evaluated at the probability values in `p` for the normal distribution with mean `mu` and standard deviation `sigma`.

```
p = 0:0.25:1;
mu = 2;
sigma = 1;
x = norminv(p,mu,sigma)
```

```
x = 1×5

      -Inf    1.3255    2.0000    2.6745       Inf
```

Compute the inverse of cdf values evaluated at 0.5 for various normal distributions with different mean parameters.

```
mu = [-2,-1,0,1,2];
sigma = 1;
x = norminv(0.5,mu,sigma)
```

```
x = 1×5

    -2    -1     0     1     2
```

Confidence Interval of Inverse Normal cdf Value

Find the maximum likelihood estimates (MLEs) of the normal distribution parameters, and then find the confidence interval of the corresponding inverse cdf value.

Generate 1000 normal random numbers from the normal distribution with mean 5 and standard deviation 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = normrnd(5,2,[n,1]);
```

Find the MLEs for the distribution parameters (mean and standard deviation) by using `mle`.

```
phat = mle(x)
```

```
phat = 1×2

    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `normlike`. The function `normlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = normlike([muHat,sigmaHat],x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the inverse cdf value at 0.5 and its 99% confidence interval.

```
[x,xLo,xUp] = norminv(0.5,muHat,sigmaHat,pCov,0.01)
```

```
x = 4.9347

xLo = 4.7721

xUp = 5.0974
```

`x` is the inverse cdf value using the normal distribution with the parameters `muHat` and `sigmaHat`. The interval `[xLo,xUp]` is the 99% confidence interval of the inverse cdf value evaluated at 0.5, considering the uncertainty of `muHat` and `sigmaHat` using `pCov`. The 99% confidence interval means the probability that `[xLo,xUp]` contains the true inverse cdf value is 0.99.

Input Arguments

`p` — Probability values at which to evaluate inverse of cdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the inverse of the cdf (icdf), specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `p` must be a scalar value.

To evaluate the icdf at multiple values, specify `p` using an array. To evaluate the icdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `p`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `norminv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

Example: `[0.1,0.5,0.9]`

Data Types: `single` | `double`

`mu` — Mean

`0` (default) | scalar value | array of scalar values

Mean of the normal distribution, specified as a scalar value or an array of scalar values.

If you specify `pCov` to compute the confidence interval `[xLo,xUp]`, then `mu` must be a scalar value.

To evaluate the icdf at multiple values, specify `p` using an array. To evaluate the icdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `p`, `mu`, and

sigma are arrays, then the array sizes must be the same. In this case, norminv expands each scalar input into a constant array of the same size as the array inputs. Each element in x is the icdf value of the distribution specified by the corresponding elements in mu and sigma, evaluated at the corresponding element in p.

Example: [0 1 2; 0 1 2]

Data Types: single | double

sigma — Standard deviation
1 (default) | positive scalar value | array of positive scalar values

Standard deviation of the normal distribution, specified as a positive scalar value or an array of positive scalar values.

If you specify pCov to compute the confidence interval [xLo,xUp], then sigma must be a scalar value.

To evaluate the icdf at multiple values, specify p using an array. To evaluate the icdfs of multiple distributions, specify mu and sigma using arrays. If one or more of the input arguments p, mu, and sigma are arrays, then the array sizes must be the same. In this case, norminv expands each scalar input into a constant array of the same size as the array inputs. Each element in x is the icdf value of the distribution specified by the corresponding elements in mu and sigma, evaluated at the corresponding element in p.

Example: [1 1 1; 2 2 2]

Data Types: single | double

pCov — Covariance of estimates
2-by-2 numeric matrix

Covariance of the estimates mu and sigma, specified as a 2-by-2 matrix.

If you specify pCov to compute the confidence interval [xLo,xUp], then p, mu, and sigma must be scalar values.

You can estimate mu and sigma by using mle, and estimate the covariance of mu and sigma by using normlike. For an example, see “Confidence Interval of Inverse Normal cdf Value” on page 33-4286.

Data Types: single | double

alpha — Significance level
0.05 (default) | scalar in the range (0,1)

Significance level for the confidence interval, specified as a scalar in the range (0,1). The confidence level is 100(1–alpha)%, where alpha is the probability that the confidence interval does not contain the true value.

Example: 0.01

Data Types: single | double

Output Arguments

x — icdf values
scalar value | array of scalar values

icdf values, evaluated at the probability values in `p`, returned as a scalar value or an array of scalar values. `x` is the same size as `p`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `x` is the icdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `p`.

`xLo` — Lower confidence bound for `x`
scalar value | array of scalar values

Lower confidence bound for `x`, returned as a scalar value or an array of scalar values. `xLo` has the same size as `x`.

`xUp` — Upper confidence bound for `x`
scalar value | array of scalar values

Upper confidence bound for `x`, returned as a scalar value or an array of scalar values. `xUp` has the same size as `x`.

More About

Normal Distribution

The normal distribution is a two-parameter family of curves. The first parameter, μ, is the mean. The second parameter, σ, is the standard deviation.

The standard normal distribution has zero mean and unit standard deviation.

The normal inverse function is defined in terms of the normal cdf as

$$x = F^{-1}(p\left|\mu,\sigma\right.) = \left\{x : F(x\left|\mu,\sigma\right.) = p\right\},$$

where

$$p = F(x\left|\mu,\sigma\right.) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt\,.$$

The result x is the solution of the integral equation where you supply the desired probability p.

Algorithms

- The `norminv` function uses the inverse complementary error function `erfcinv`. The relationship between `norminv` and `erfcinv` is

  $$\text{norminv}(p) = -\sqrt{2}\,\text{erfcinv}(2p)$$

  The inverse complementary error function `erfcinv(x)` is defined as `erfcinv(erfc(x))=x`, and the complementary error function `erfc(x)` is defined as

  $$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{\infty} e^{-t^2} dt\,.$$

- The `norminv` function computes confidence bounds for `x` by using the delta method. `norminv(p,mu,sigma)` is equivalent to `mu + sigma*norminv(p,0,1)`. Therefore, the `norminv` function estimates the variance of `mu + sigma*norminv(p,0,1)` using the covariance

matrix of `mu` and `sigma` by the delta method, and finds the confidence bounds using the estimates of this variance. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pCov` from large samples.

Alternative Functionality

- `norminv` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use `icdf`, create a `NormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `norminv` is faster than the generic function `icdf`.

References

[1] Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1964.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`NormalDistribution` | `erfcinv` | `icdf` | `normcdf` | `normfit` | `normlike` | `normspec`

Topics

“Normal Distribution” on page B-119

Introduced before R2006a

normlike

Normal negative loglikelihood

Syntax

```
nlogL = normlike(params,x)
nlogL = normlike(params,x,censoring)
nlogL = normlike(params,x,censoring,freq)

[nlogL,aVar] = normlike( ___ )
```

Description

`nlogL = normlike(params,x)` returns the normal negative loglikelihood of the distribution parameters (`params`) given the sample data (`x`). `params(1)` and `params(2)` correspond to the mean and standard deviation of the normal distribution, respectively.

`nlogL = normlike(params,x,censoring)` specifies whether each value in `x` is right-censored or not. Use the logical vector `censoring` in which 1 indicates observations that are right-censored and 0 indicates observations that are fully observed.

`nlogL = normlike(params,x,censoring,freq)` specifies the frequency or weights of observations. To specify `freq` without specifying `censoring`, you can pass `[]` for `censoring`.

`[nlogL,aVar] = normlike( ___ )` also returns the inverse of the Fisher information matrix `aVar`, using any of the input argument combinations in the previous syntaxes. If values in `params` are the maximum likelihood estimates (MLEs) of the parameters, `aVar` is an approximation to the asymptotic covariance matrix.

Examples

Negative Loglikelihood of MLEs

Find the MLEs of a data set with censoring by using `normfit`, and then find the negative loglikelihood of the MLEs by using `normlike`.

Load the sample data.

```
load lightbulb
```

The first column of the data contains the lifetime (in hours) of two types of bulbs. The second column contains the binary variable indicating whether the bulb is fluorescent or incandescent. 1 indicates that the bulb is fluorescent, and 0 indicates that the bulb is incandescent. The third column contains the censorship information, where 0 indicates the bulb is observed until failure, and 1 indicates the item (bulb) is censored.

Find the indices for fluorescent bulbs.

```
idx = find(lightbulb(:,2) == 0);
```

Find the MLEs of the normal distribution parameters. The second input argument of `normfit` specifies the confidence level. Pass in `[]` to use its default value 0.05. The third input argument specifies the censorship information.

```
censoring = lightbulb(idx,3) == 1;
[muHat,sigmaHat] = normfit(lightbulb(idx,1),[],censoring)
```

```
muHat = 9.4966e+03
```

```
sigmaHat = 3.0640e+03
```

Find the negative loglikelihood of the MLEs.

```
nlogL = normlike([muHat,sigmaHat],lightbulb(idx,1),censoring)
```

```
nlogL = 376.2305
```

Convert Unbiased Estimator to MLE

The function `normfit` finds the sample mean and the square root of the unbiased estimator of the variance with no censoring. The sample mean is equal to the MLE of the mean parameter, but the square root of the unbiased estimator of the variance is not equal to the MLE of the standard deviation parameter.

Find the normal distribution parameters by using `normfit`, convert them into MLEs, and then compare the negative log likelihoods of the estimates by using `normlike`.

Generate 100 normal random numbers from the standard normal distribution.

```
rng('default') % For reproducibility
n = 100;
x = normrnd(0,1,[n,1]);
```

Find the sample mean and the square root of the unbiased estimator of the variance.

```
[muHat,sigmaHat] = normfit(x)
```

```
muHat = 0.1231
```

```
sigmaHat = 1.1624
```

Convert the square root of the unbiased estimator of the variance into the MLE of the standard deviation parameter.

```
sigmaHat_MLE = sqrt((n-1)/n)*sigmaHat
```

```
sigmaHat_MLE = 1.1566
```

The difference between `sigmaHat` and `sigmaHat_MLE` is negligible for large `n`.

Alternatively, you can find the MLEs by using the function `mle`.

```
phat = mle(x)
```

```
phat = 1×2
```

```
    0.1231    1.1566
```

`phat(1)` and `phat(2)` are the MLEs of the mean and the standard deviation parameter, respectively.

Confirm that the log likelihood of the MLEs (`muHat` and `sigmaHat_MLE`) is greater than the log likelihood of the unbiased estimators (`muHat` and `sigmaHat`) by using the `normlike` function.

```
logL = -normlike([muHat,sigmaHat],x)
```

```
logL = -156.4424
```

```
logL_MLE = -normlike([muHat,sigmaHat_MLE],x)
```

```
logL_MLE = -156.4399
```

Confidence Interval of Inverse Normal cdf Value

Find the maximum likelihood estimates (MLEs) of the normal distribution parameters, and then find the confidence interval of the corresponding inverse cdf value.

Generate 1000 normal random numbers from the normal distribution with mean 5 and standard deviation 2.

```
rng('default') % For reproducibility
n = 1000; % Number of samples
x = normrnd(5,2,[n,1]);
```

Find the MLEs for the distribution parameters (mean and standard deviation) by using `mle`.

```
phat = mle(x)
```

```
phat = 1×2

    4.9347    1.9969
```

```
muHat = phat(1);
sigmaHat = phat(2);
```

Estimate the covariance of the distribution parameters by using `normlike`. The function `normlike` returns an approximation to the asymptotic covariance matrix if you pass the MLEs and the samples used to estimate the MLEs.

```
[~,pCov] = normlike([muHat,sigmaHat],x)
```

```
pCov = 2×2

    0.0040   -0.0000
   -0.0000    0.0020
```

Find the inverse cdf value at 0.5 and its 99% confidence interval.

```
[x,xLo,xUp] = norminv(0.5,muHat,sigmaHat,pCov,0.01)
```

```
x = 4.9347
```

```
xLo = 4.7721

xUp = 5.0974
```

`x` is the inverse cdf value using the normal distribution with the parameters `muHat` and `sigmaHat`. The interval `[xLo,xUp]` is the 99% confidence interval of the inverse cdf value evaluated at 0.5, considering the uncertainty of `muHat` and `sigmaHat` using `pCov`. The 99% confidence interval means the probability that `[xLo,xUp]` contains the true inverse cdf value is 0.99.

Input Arguments

`params` — Normal distribution parameters

vector of two numeric values

Normal distribution parameters consisting of the mean and standard deviation, specified as a vector of two numeric values. `params(1)` and `params(2)` correspond to the mean and standard deviation of the normal distribution, respectively. `params(2)` must be positive.

Example: `[0,1]`

Data Types: `single` | `double`

`x` — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

`censoring` — Indicator for censoring

array of 0s (default) | logical vector

Indicator for the censoring of each value in `x`, specified as a logical vector of the same size as `x`. Use 1 for observations that are right-censored and 0 for observations that are fully observed.

The default is an array of 0s, meaning that all observations are fully observed.

Data Types: `logical`

`freq` — Frequency or weights of observations

array of 1s (default) | nonnegative vector

Frequency or weights of observations, specified as a nonnegative vector that is the same size as `x`. The `freq` input argument typically contains nonnegative integer counts for the corresponding elements in `x`, but can contain any nonnegative values.

To obtain the weighted negative loglikelihood for a data set with censoring, specify weights of observations, normalized to the number of observations in `x`.

The default is an array of 1s, meaning one observation per element of `x`.

Data Types: `single` | `double`

Output Arguments

`nlogL` — Negative loglikelihood

numeric scalar

Negative loglikelihood value of the distribution parameters (`params`) given the sample data (`x`), returned as a numeric scalar.

`aVar` — Inverse of Fisher information matrix

numeric matrix

Inverse of the Fisher information matrix, returned as a 2-by-2 numeric matrix. `aVar` is based on the observed Fisher information given the observed data (`x`), not the expected information.

If values in `params` are the MLEs of the parameters, `aVar` is an approximation to the asymptotic variance-covariance matrix (also known as the asymptotic covariance matrix). To find the MLEs, use `mle`.

Alternative Functionality

`normlike` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic functions `mlecov`, `fitdist`, `negloglik`, and `proflik` and the **Distribution Fitter** app, which support various probability distributions.

- `mlecov` returns the asymptotic covariance matrix of the MLEs of the parameters for a distribution specified by a custom probability density function. For example, `mlecov(params,x,'pdf',@normpdf)` returns the asymptotic covariance matrix of the MLEs for the normal distribution.
- Create a `NormalDistribution` probability distribution object by fitting the distribution to data using the `fitdist` function or the **Distribution Fitter** app. The object property `ParameterCovariance` stores the covariance matrix of the parameter estimates. To obtain the negative loglikelihood of the parameter estimates and the profile of the likelihood function, pass the object to `negloglik` and `proflik`, respectively.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[2] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[3] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

Distribution Fitter | `NormalDistribution` | `mle` | `mlecov` | `negloglik` | `normcdf` | `normfit` | `norminv` | `proflik`

Topics

“Negative Loglikelihood Functions” on page 5-24
“Normal Distribution” on page B-119

Introduced before R2006a

normpdf

Normal probability density function

Syntax

```
y = normpdf(x)
y = normpdf(x,mu)
y = normpdf(x,mu,sigma)
```

Description

`y = normpdf(x)` returns the probability density function (pdf) of the standard normal distribution, evaluated at the values in `x`.

`y = normpdf(x,mu)` returns the pdf of the normal distribution with mean `mu` and the unit standard deviation, evaluated at the values in `x`.

`y = normpdf(x,mu,sigma)` returns the pdf of the normal distribution with mean `mu` and standard deviation `sigma`, evaluated at the values in `x`.

Examples

Standard Normal Distribution pdf

Compute the pdf values for the standard normal distribution at the values in `x`.

```
x = [-2,-1,0,1,2];
y = normpdf(x)
```

```
y = 1×5

    0.0540    0.2420    0.3989    0.2420    0.0540
```

Normal Distribution pdf

Compute the pdf values evaluated at the values in `x` for the normal distribution with mean `mu` and standard deviation `sigma`.

```
x = [-2,-1,0,1,2];
mu = 2;
sigma = 1;
y = normpdf(x,mu,sigma)
```

```
y = 1×5

    0.0001    0.0044    0.0540    0.2420    0.3989
```

Compute the pdf values evaluated at zero for various normal distributions with different mean parameters.

```
mu = [-2,-1,0,1,2];
sigma = 1;
y = normpdf(0,mu,sigma)
```

```
y = 1×5

    0.0540    0.2420    0.3989    0.2420    0.0540
```

Input Arguments

x — Values at which to evaluate pdf

scalar value | array of scalar values

Values at which to evaluate the pdf, specified as a scalar value or an array of scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[-1,0,3,4]`

Data Types: `single` | `double`

mu — Mean

`0` (default) | scalar value | array of scalar values

Mean of the normal distribution, specified as a scalar value or an array of scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

sigma — Standard deviation

`1` (default) | positive scalar value | array of positive scalar values

Standard deviation of the normal distribution, specified as a positive scalar value or an array of positive scalar values.

To evaluate the pdf at multiple values, specify `x` using an array. To evaluate the pdfs of multiple distributions, specify `mu` and `sigma` using arrays. If one or more of the input arguments `x`, `mu`, and `sigma` are arrays, then the array sizes must be the same. In this case, `normpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of

the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

Output Arguments

y — pdf values

scalar value | array of scalar values

pdf values, evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x`, `mu`, and `sigma` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `mu` and `sigma`, evaluated at the corresponding element in `x`.

More About

Normal Distribution

The normal distribution is a two-parameter family of curves. The first parameter, μ, is the mean. The second parameter, σ, is the standard deviation.

The standard normal distribution has zero mean and unit standard deviation.

The normal probability density function (pdf) is

$$y = f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in \mathbb{R}.$$

The *likelihood function* is the pdf viewed as a function of the parameters. The maximum likelihood estimates (MLEs) are the parameter estimates that maximize the likelihood function for fixed values of `x`.

Alternative Functionality

- `normpdf` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create a `NormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `normpdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`NormalDistribution` | `mvnpdf` | `normcdf` | `norminv` | `normrnd` | `normspec` | `pdf`

Topics

“Normal Distribution” on page B-119

Introduced before R2006a

normplot

Normal probability plot

Syntax

```
normplot(x)
normplot(ax,x)
h = normplot( ___ )
```

Description

`normplot(x)` creates a normal probability plot comparing the distribution of the data in `x` to the normal distribution.

`normplot` plots each data point in `x` using plus sign (`'+'`) markers and draws two reference lines that represent the theoretical distribution. A solid reference line connects the first and third quartiles of the data, and a dashed reference line extends the solid line to the ends of the data. If the sample data has a normal distribution, then the data points appear along the reference line. A distribution other than normal introduces curvature in the data plot.

`normplot(ax,x)` adds a normal probability plot into the axes specified by `ax`.

`h = normplot( ___ )` returns graphics handles corresponding to the plotted lines, using any of the previous syntaxes.

Examples

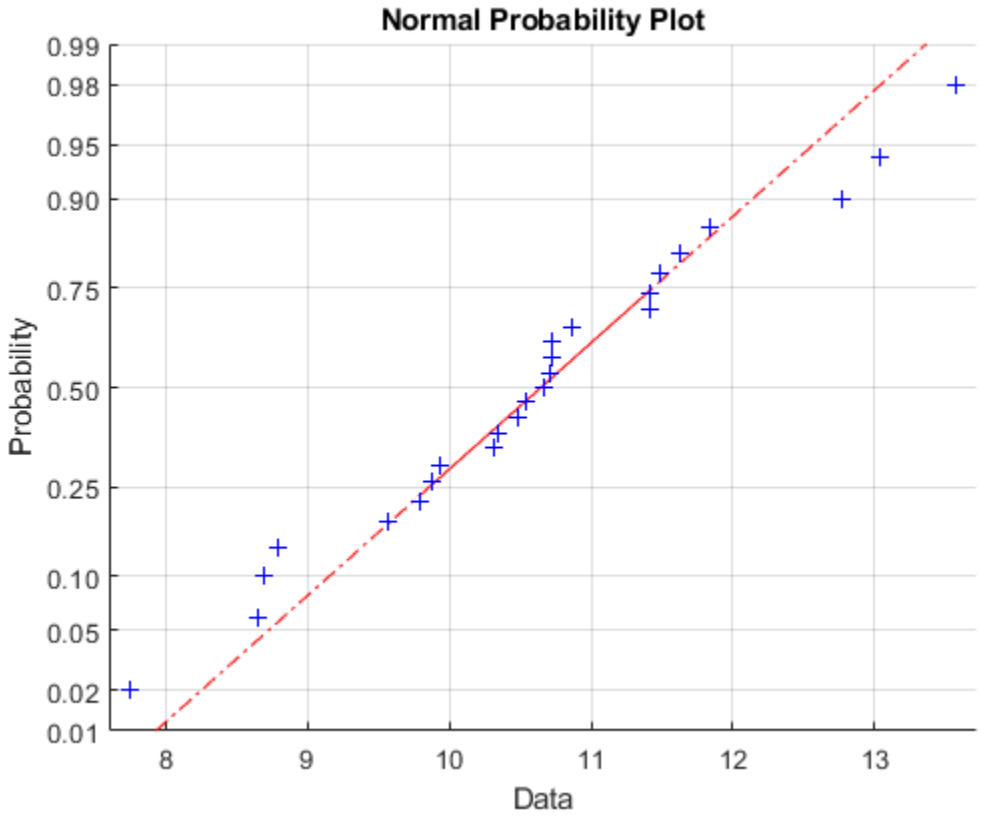
Generate a Normal Probability Plot

Generate random sample data from a normal distribution with `mu = 10` and `sigma = 1`.

```
rng default;  % For reproducibility
x = normrnd(10,1,25,1);
```

Create a normal probability plot of the sample data.

```
figure;
normplot(x)
```



The plot indicates that the data follows a normal distribution.

Assess Normality Using a Normal Probability Plot

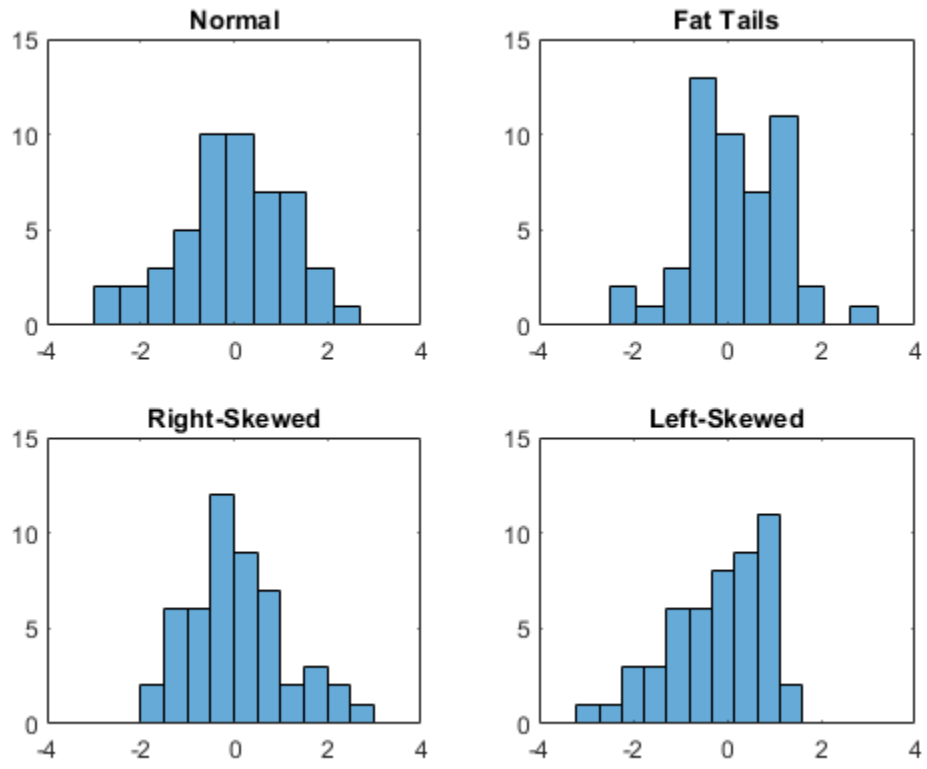
Generate 50 random numbers from each of four different distributions: A standard normal distribution; a Student's-t distribution with five degrees of freedom (a "fat-tailed" distribution); a set of Pearson random numbers with `mu` equal to 0, `sigma` equal to 1, skewness equal to 0.5, and kurtosis equal to 3 (a "right-skewed" distribution); and a set of Pearson random numbers with `mu` equal to 0, `sigma` equal to 1, skewness equal to -0.5, and kurtosis equal to 3 (a "left-skewed" distribution).

```
rng(11)  % For reproducibility
x1 = normrnd(0,1,[50,1]);
x2 = trnd(5,[50,1]);
x3 = pearsrnd(0,1,0.5,3,[50,1]);
x4 = pearsrnd(0,1,-0.5,3,[50,1]);
```

Plot four histograms on the same figure for a visual comparison of the pdf of each distribution.

```
figure
subplot(2,2,1)
histogram(x1,10)
title('Normal')
axis([-4,4,0,15])
```

```
subplot(2,2,2)
histogram(x2,10)
title('Fat Tails')
axis([-4,4,0,15])

subplot(2,2,3)
histogram(x3,10)
title('Right-Skewed')
axis([-4,4,0,15])

subplot(2,2,4)
histogram(x4,10)
title('Left-Skewed')
axis([-4,4,0,15])
```

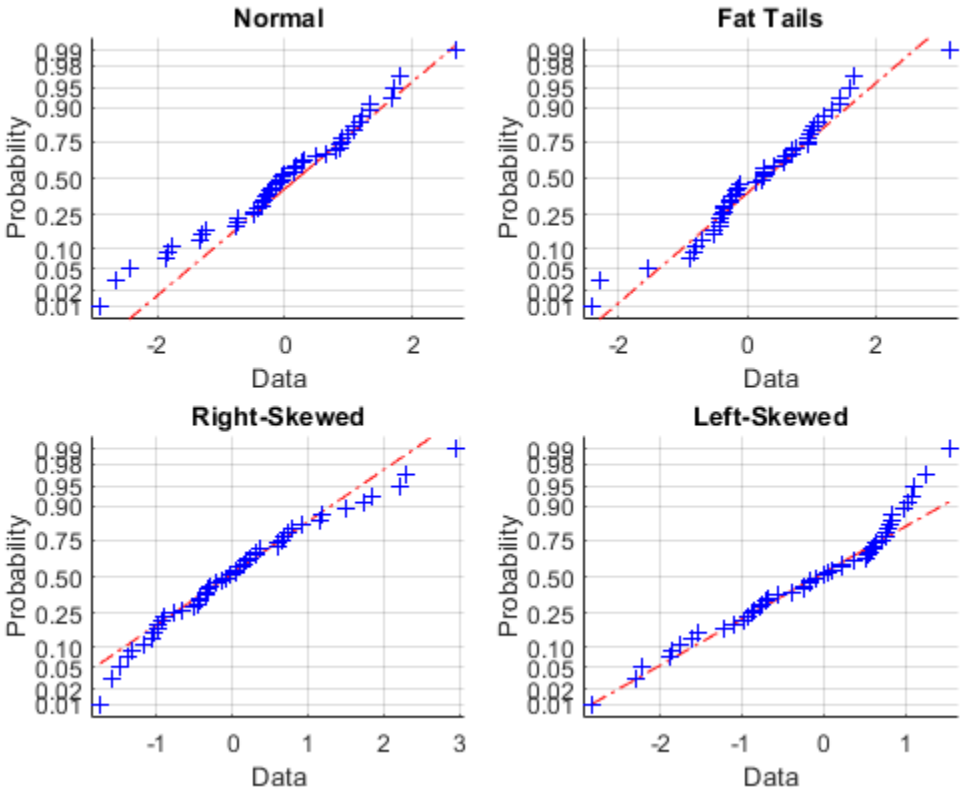


The histograms show how each sample differs from the normal distribution.

Create a normal probability plot for each sample.

```
figure
subplot(2,2,1)
normplot(x1)
title('Normal')

subplot(2,2,2)
normplot(x2)
title('Fat Tails')
```

```
subplot(2,2,3)
normplot(x3)
title('Right-Skewed')

subplot(2,2,4)
normplot(x4)
title('Left-Skewed')
```



Adjust Normal Probability Plot Line Properties

Create a 50-by-2 matrix containing 50 random numbers from each of two different distributions: A standard normal distribution in column 1, and a set of Pearson random numbers with `mu` equal to 0, `sigma` equal to 1, skewness equal to 0.5, and kurtosis equal to 3 (a "right-skewed" distribution) in column 2.

```
rng default  % For reproducibility
x = [normrnd(0,1,[50,1]) pearsrnd(0,1,0.5,3,[50,1])];
```

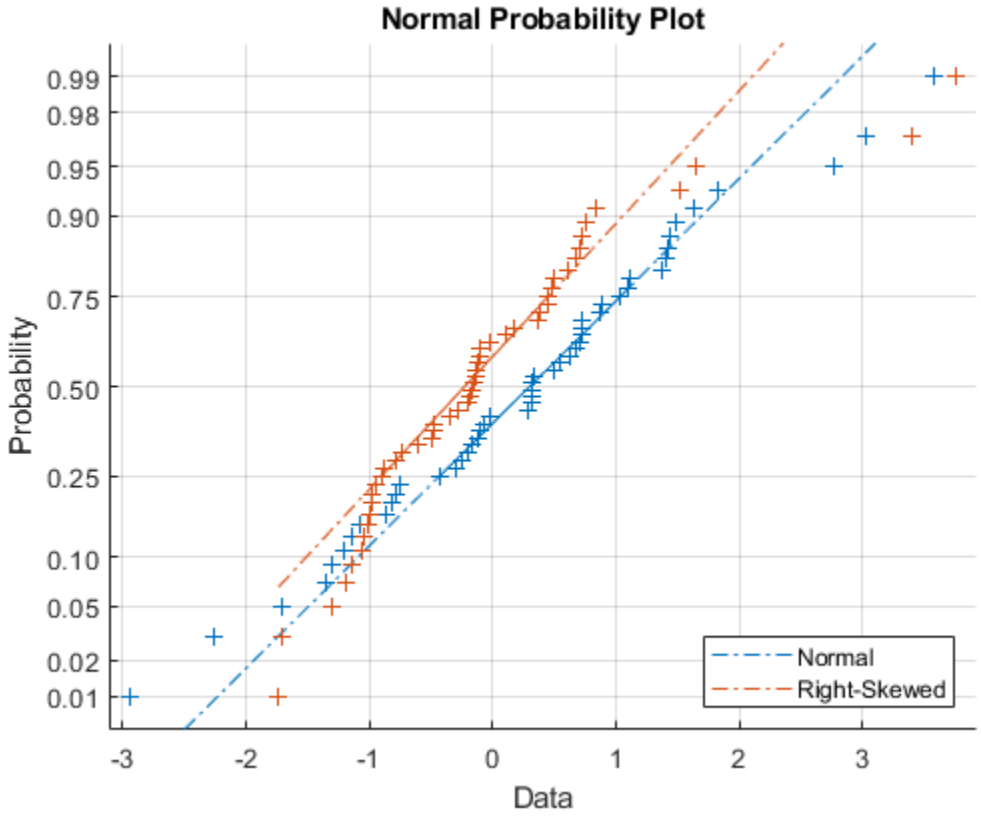
Create a normal probability plot for both samples on the same figure. Return the plot line graphic handles.

```
figure
h = normplot(x)
```

```
h =
  6x1 Line array:
```

```
  Line
  Line
  Line
  Line
  Line
  Line
```

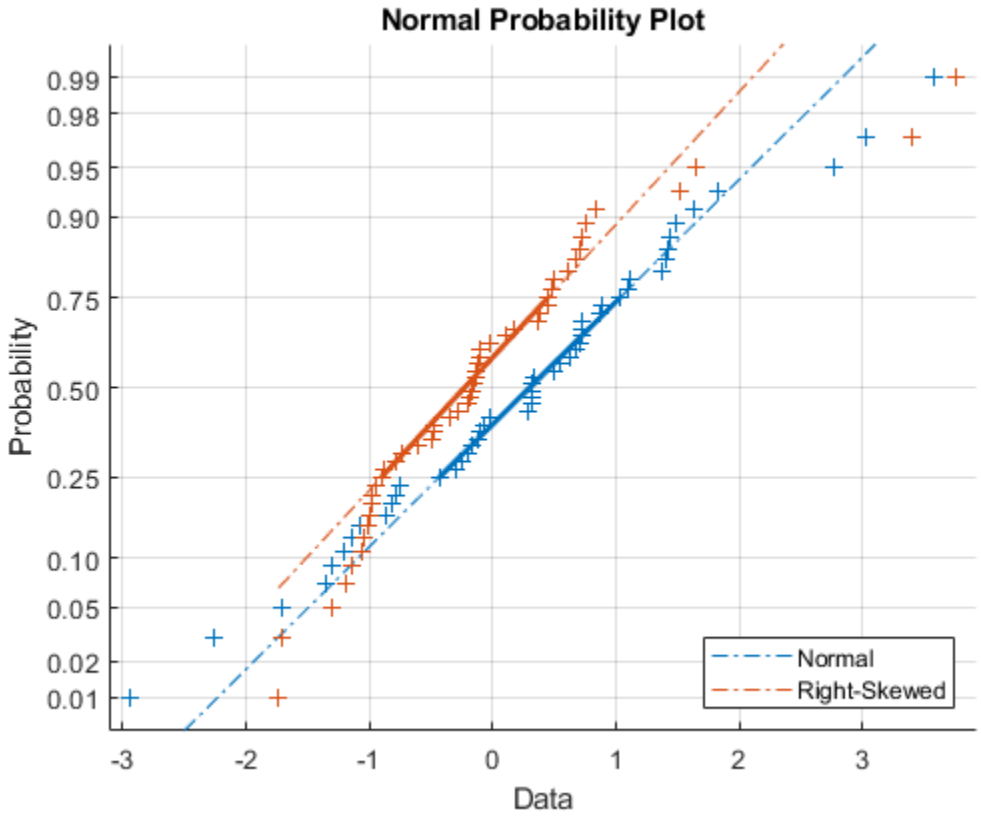
```
legend({'Normal','Right-Skewed'},'Location','southeast')
```



The handles h(1) and h(2) correspond to the data points for the normal and skewed distributions, respectively. The handles h(3) and h(4) correspond to the second and third quartile line fit to the sample data. The handles h(5) and h(6) correspond to the extrapolated line that extends to the minimum and maximum of each set of sample data.

To illustrate, increase the line width of the second and third quartile line for the normally distributed data sample (represented by h(3)) to 2.

```
h(3).LineWidth = 2;
h(4).LineWidth = 2;
```



Input Arguments

x — Sample data

numeric vector | numeric matrix

Sample data, specified as a numeric vector or numeric matrix. `normplot` displays each value in `x` using the symbol `'+'`. If `x` is a matrix, then `normplot` displays a separate line for each column of `x`.

Data Types: `single` | `double`

ax — Target axes

`Axes` object | `UIAxes` object

Target axes, specified as an `Axes` object or a `UIAxes` object. `normplot` adds an additional plot into the axes specified by `ax`. For details, see Axes Properties and UIAxes Properties.

Use `gca` to return the current axes for the current figure.

Output Arguments

h — Graphics handles for line objects

vector of `Line` graphics handles

Graphics handles for line objects, returned as a vector of `Line` graphics handles. Graphics handles are unique identifiers that you can use to query and modify the properties of a specific line on the plot. For each column of `x`, `normplot` returns three handles:

- The line representing the data points. `normplot` represents each data point in `x` using plus sign (`'+'`) markers.
- The line joining the first and third quartiles of each column of `x`, represented as a solid line.
- The extrapolation of the quartile line, extended to the minimum and maximum values of `x`, represented as a dashed line.

To view and set properties of line objects, use dot notation. For information on using dot notation, see "Access Property Values". For information on the `Line` properties that you can set, see Primitive Line.

Algorithms

`normplot` matches the quantiles of sample data to the quantiles of a normal distribution. The sample data is sorted and plotted on the x-axis. The y-axis represents the quantiles of the normal distribution, converted into probability values. Therefore, the y-axis scaling is not linear.

Where the x-axis value is the ith sorted value from a sample of size N, the y-axis value is the midpoint between evaluation points of the empirical cumulative distribution function of the data. The midpoint is equal to $\frac{(i - 0.5)}{N}$.

`normplot` superimposes a reference line to assess the linearity of the plot. The line goes through the first and third quartiles of the data.

Alternative Functionality

You can use the `probplot` function to create a probability plot. The `probplot` function enables you to indicate censored data and specify the distribution for a probability plot.

See Also

`cdfplot` | `ecdf` | `probplot` | `wblplot`

Topics

"Distribution Plots" on page 4-7
"Hypothesis Testing" on page 8-5
"Normal Distribution" on page B-119

Introduced before R2006a

normrnd

Normal random numbers

Syntax

```
r = normrnd(mu,sigma)
r = normrnd(mu,sigma,sz1,...,szN)
r = normrnd(mu,sigma,sz)
```

Description

`r = normrnd(mu,sigma)` generates a random number from the normal distribution with mean parameter `mu` and standard deviation parameter `sigma`.

`r = normrnd(mu,sigma,sz1,...,szN)` generates an array of normal random numbers, where `sz1,...,szN` indicates the size of each dimension.

`r = normrnd(mu,sigma,sz)` generates an array of normal random numbers, where vector `sz` specifies `size(r)`.

Examples

Generate Normal Random Number

Generate a single random value from the standard normal distribution.

```
rng('default') % For reproducibility
r = normrnd(0,1)
```

```
r = 0.5377
```

Reset Random Number Generator

Save the current state of the random number generator. Then create a 1-by-5 vector of normal random numbers from the normal distribution with mean 3 and standard deviation 10.

```
s = rng;
r = normrnd(3,10,[1,5])
```

```
r = 1×5

    8.3767   21.3389  -19.5885   11.6217    6.1877
```

Restore the state of the random number generator to `s`, and then create a new 1-by-5 vector of random numbers. The values are the same as before.

```
rng(s);
r1 = normrnd(3,10,[1,5])
```

```
r1 = 1×5

    8.3767   21.3389  -19.5885   11.6217    6.1877
```

Clone Size from Existing Array

Create a matrix of normally distributed random numbers with the same size as an existing array.

```
A = [3 2; -2 1];
sz = size(A);
R = normrnd(0,1,sz)
```

```
R = 2×2

    0.5377   -2.2588
    1.8339    0.8622
```

You can combine the previous two lines of code into a single line.

```
R = normrnd(1,0,size(A));
```

Input Arguments

`mu` — Mean

scalar value | array of scalar values

Mean of the normal distribution, specified as a scalar value or an array of scalar values.

To generate random numbers from multiple distributions, specify `mu` and `sigma` using arrays. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `normrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation

nonnegative scalar value | array of nonnegative scalar values

Standard deviation of the normal distribution, specified as a nonnegative scalar value or an array of nonnegative scalar values.

If `sigma` is zero, then the output `r` is always equal to `mu`.

To generate random numbers from multiple distributions, specify `mu` and `sigma` using arrays. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `normrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as integers. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the probability distribution.

If either `mu` or `sigma` is an array, then the specified dimensions `sz1,...,szN` must match the common dimensions of `mu` and `sigma` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `normrnd` ignores trailing dimensions with a size of 1. For example, specifying `3,1,1,1` produces a 3-by-1 vector of random numbers.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5,3,2]` generates a 5-by-3-by-2 array of random numbers from the probability distribution.

If either `mu` or `sigma` is an array, then the specified dimensions `sz` must match the common dimensions of `mu` and `sigma` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value `[sz1]`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `normrnd` ignores trailing dimensions with a size of 1. For example, specifying `[3,1,1,1]` produces a 3-by-1 vector of random numbers.

Example: `[5,3,2]`

Data Types: `single` | `double`

Output Arguments

`r` — Normal random numbers
scalar value | array of scalar values

Normal random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `mu` and `sigma`.

Alternative Functionality

- `normrnd` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use

`random`, create a `NormalDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `normrnd` is faster than the generic function `random`.

- Use `randn` to generate random numbers from the standard normal distribution.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

References

[1] Marsaglia, G, and W. W. Tsang. "A Fast, Easily Implemented Method for Sampling from Decreasing or Symmetric Unimodal Density Functions." *SIAM Journal on Scientific and Statistical Computing*. Vol. 5, Number 2, 1984, pp. 349–359.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`NormalDistribution` | `lognrnd` | `mvnrnd` | `normcdf` | `normpdf` | `randn` | `random`

Topics

"Random Number Generation" on page 5-27
"Normal Distribution" on page B-119

Introduced before R2006a

normspec

Normal density plot shading between specifications

Syntax

```
p = normspec(specs)
p = normspec(specs,mu,sigma)
p = normspec(specs,mu,sigma,region)
[p,h] = normspec( ___ )
```

Description

`p = normspec(specs)` plots the standard normal density, shading the portion inside the specification limits given by the two-element vector `specs`, and returns the probability `p` of the shaded area. If `spec` has no lower limit, then set `specs(1)` to `−Inf`; if `spec` has no upper limit, then set `specs(2)` to `Inf`.

`p = normspec(specs,mu,sigma)` uses a normal density with parameters `mu` and `sigma`.

`p = normspec(specs,mu,sigma,region)` specifies the shading region as either `'inside'` or `'outside'` the specification limits. The default is `'inside'`.

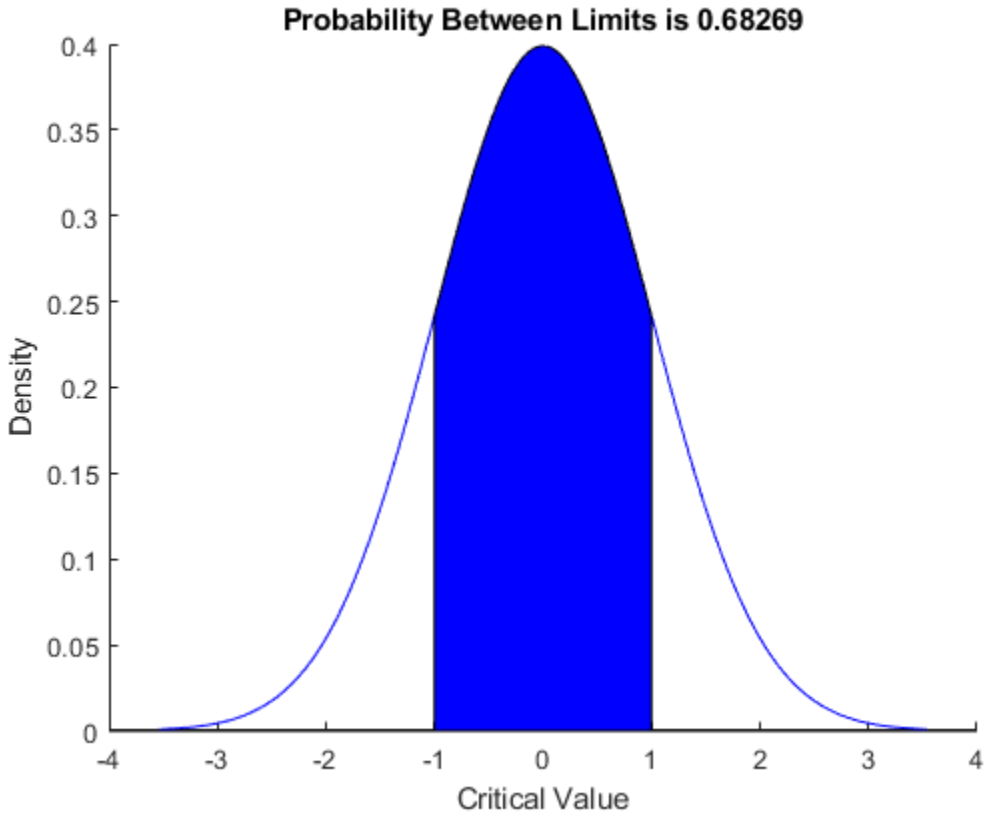
`[p,h] = normspec( ___ )` also returns a column vector of chart line objects using any of the input arguments in the previous syntaxes. Use `h` to modify properties of a specific chart line after you create it. For a list of properties, see Line.

Examples

Create Standard Normal Density Plot Shading Inside of Limits

Create a standard normal density plot, shading the portion inside the specification limits `[-1,1]`.

```
p = normspec([-1,1])
```



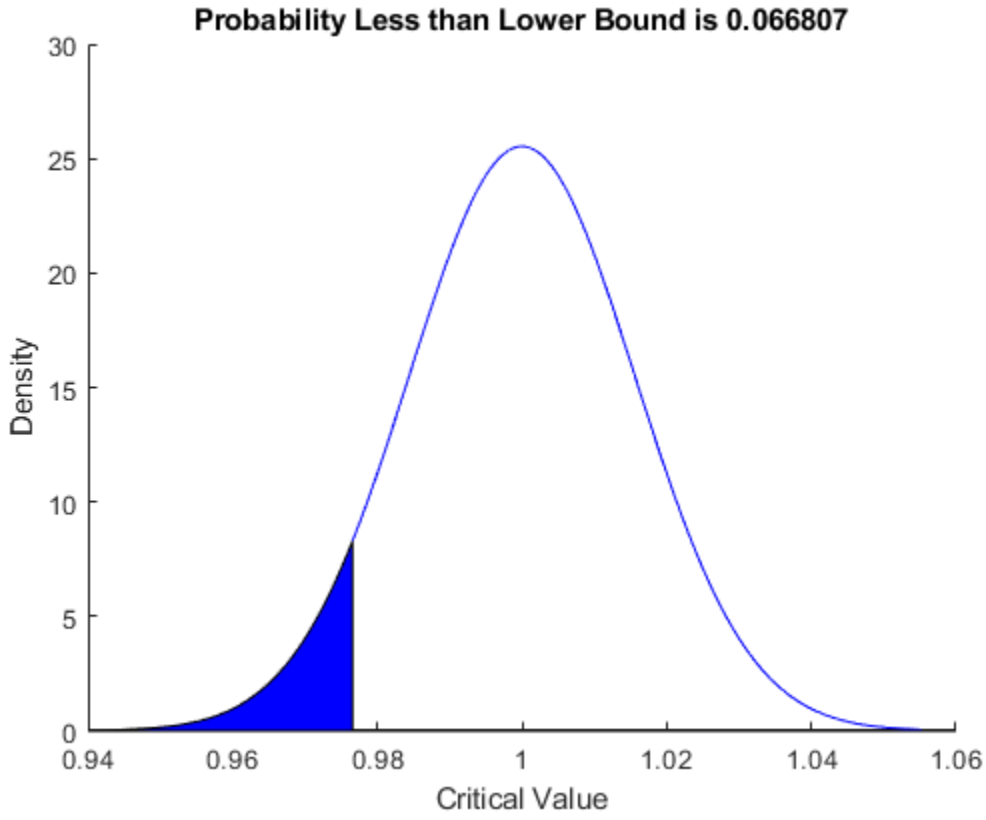

```
p = 0.6827
```

Create Normal Density Plot Shading Outside of Limits

Create a normal density plot, shading the portion outside the specification limits.

For example, consider a production process that fills cans of paint. The average amount of paint in any can is 1 gallon, but variability in the process produces a standard deviation of 2 ounces (2/128 gallons). Create a normal density plot, shading the portion corresponding to the probability that the cans will be filled under specification by 3 or more ounces.

```
p = normspec([1-3/128,Inf],1,2/128,'outside')
```




```
p = 0.0668
```

Input Arguments

`specs` — Specification limits

two-element numeric vector

Specification limits, specified as a two-element numeric vector. Set `specs(1)` to `–Inf` if there is no lower limit; set `specs(2)` to `Inf` if there is no upper limit.

Data Types: `single` | `double`

`mu` — Mean

`0` (default) | scalar value

Mean of the normal distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Standard deviation

`1` (default) | positive scalar value

Standard deviation of the normal distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`region` — Shading region
`'inside'` (default) | `'outside'`

Shading region, specified as either `'inside'` or `'outside'`. The function `normspec` shades the region either `'inside'` or `'outside'` the specification limits.

Output Arguments

`p` — Probability of shaded area
nonnegative scalar value

Probability of the shaded area, returned as a nonnegative scalar value.

`h` — One or more chart line objects
scalar | vector

One or more chart line objects, returned as a scalar or vector. These objects are unique identifiers that you use to query and modify properties of a specific chart line. For a list of properties, see Chart Line.

See Also
`capaplot` | `histfit`

Topics
“Normal Distribution” on page B-119

Introduced before R2006a

normstat

Normal mean and variance

Syntax

```
[m,v] = normstat(mu,sigma)
```

Description

`[m,v] = normstat(mu,sigma)` returns the mean and variance of the normal distribution with mean `mu` and standard deviation `sigma`.

The mean of the normal distribution with parameters μ and σ is μ, and the variance is σ^2.

Examples

Compute Mean and Variance

Compute the mean and variance of the normal distribution with parameters `mu` and `sigma`.

```
mu = 1;
sigma = 1:5;
[m,v] = normstat(mu,sigma)
```

```
m = 1×5

     1     1     1     1     1


v = 1×5

     1     4     9    16    25
```

Input Arguments

`mu` — Mean

scalar value | array of scalar values

Mean of the normal distribution, specified as a scalar value or an array of scalar values.

To compute the means and variances of multiple distributions, specify distribution parameters using an array of scalar values. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `normstat` expands the scalar argument into a constant array of the same size as the other argument. Each element in `m` and `v` is the mean and variance of the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[0 1 2; 0 1 2]`

Data Types: `single` | `double`

`sigma` — Standard deviation
positive scalar value | array of positive scalar values

Standard deviation of the normal distribution, specified as a positive scalar value or an array of positive scalar values.

To compute the means and variances of multiple distributions, specify distribution parameters using an array of scalar values. If both `mu` and `sigma` are arrays, then the array sizes must be the same. If either `mu` or `sigma` is a scalar, then `normstat` expands the scalar argument into a constant array of the same size as the other argument. Each element in `m` and `v` is the mean and variance of the distribution specified by the corresponding elements in `mu` and `sigma`.

Example: `[1 1 1; 2 2 2]`

Data Types: `single` | `double`

Output Arguments

`m` — Mean
scalar value | array of scalar values

Mean of the normal distribution, returned as a scalar value or an array of scalar values. `m` is the same size as `mu` and `sigma` after any necessary scalar expansion. Each element in `m` is the mean of the normal distribution specified by the corresponding elements in `mu` and `sigma`.

`v` — Variance
scalar value | array of scalar values

Variance of the normal distribution, returned as a scalar value or an array of scalar values. `v` is the same size as `mu` and `sigma` after any necessary scalar expansion. Each element in `v` is the variance of the normal distribution specified by the corresponding elements in `mu` and `sigma`.

Alternative Functionality

- `normstat` is a function specific to normal distribution. Statistics and Machine Learning Toolbox also offers generic functions to compute summary statistics, including mean (`mean`), median (`median`), interquartile range (`iqr`), variance (`var`), and standard deviation (`std`). These generic functions support various probability distributions. To use these functions, create a `NormalDistribution` probability distribution object and pass the object as an input argument.

References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`NormalDistribution` | `mean` | `normcdf` | `normpdf` | `normrnd` | `std` | `var`

Topics

“Normal Distribution” on page B-119

Introduced before R2006a

nsegments

Number of segments in piecewise distribution

Syntax

```
n = nsegments(pd)
```

Description

`n = nsegments(pd)` returns the number of segments in the piecewise distribution object `pd`.

Examples

Number of Segments in `paretotails` Object

Generate a sample data set and fit a piecewise distribution with a Pareto tail to the data by using `paretotails`. Find the number of segments in the fitted distribution by using the object function `nsegments`.

Generate a sample data set containing 10% outliers in the right tail.

```
rng('default');  % For reproducibility
right_tail = exprnd(5,100,1);
center = randn(900,1);
x = [center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities. Pass in 0 and 0.9 so that a fitted object does not contain a lower tail segment, and consists of the empirical distribution for the lower 90% of the data set and a generalized Pareto distribution (GPD) for the upper 10% of the data set.

```
pd = paretotails(x,0,0.9)
```

```
pd =
Piecewise distribution with 2 segments
   -Inf < x < 1.73931  (0 < p < 0.9): interpolated empirical cdf
    1.73931 < x < Inf  (0.9 < p < 1): upper tail, GPD(0.643752,1.62246)
```

Return the number of segments in `pd` by using the `nsegments` function.

```
n = nsegments(pd)
```

```
n = 2
```

You can also get the number of segments by using the `NumSegments` property. Access the `NumSegments` property by using dot notation.

```
pd.NumSegments
```

```
ans = 2
```

Input Arguments

pd — Piecewise distribution with Pareto tails

`paretotails` object

Piecewise distribution with Pareto tails, specified as a `paretotails` object.

See Also

`boundary` | `lowerparams` | `paretotails` | `segment` | `upperparams`

Topics

“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

numel

Class: `dataset`

(Not Recommended) Number of elements in dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
n = numel(A)
n = numel(A, varargin)
```

Description

`n = numel(A)` returns 1. To find the number of elements, `n`, in the dataset array `A`, use `prod(size(A))` or `numel(A,':',':')`.

`n = numel(A, varargin)` returns the number of subscripted elements, `n`, in `A(index1, index2, ..., indexn)`, where `varargin` is a string array or cell array whose elements are `index1, index2, ... indexn`.

See Also

`length` | `size`

optimalleaforder

Optimal leaf ordering for hierarchical clustering

Syntax

```
leafOrder = optimalleaforder(tree,D)
leafOrder = optimalleaforder(tree,D,Name,Value)
```

Description

`leafOrder = optimalleaforder(tree,D)` returns an optimal leaf ordering for the hierarchical binary cluster tree, `tree`, using the distances, `D`. An optimal leaf ordering of a binary tree maximizes the sum of the similarities between adjacent leaves by flipping tree branches without dividing the clusters.

`leafOrder = optimalleaforder(tree,D,Name,Value)` returns the optimal leaf ordering using one or more name-value pair arguments.

Examples

Plot Dendrogram With Optimal Leaf Order

Create a hierarchical binary cluster tree using `linkage`. Then, compare the dendrogram plot with the default ordering to a dendrogram with an optimal leaf ordering.

Generate sample data.

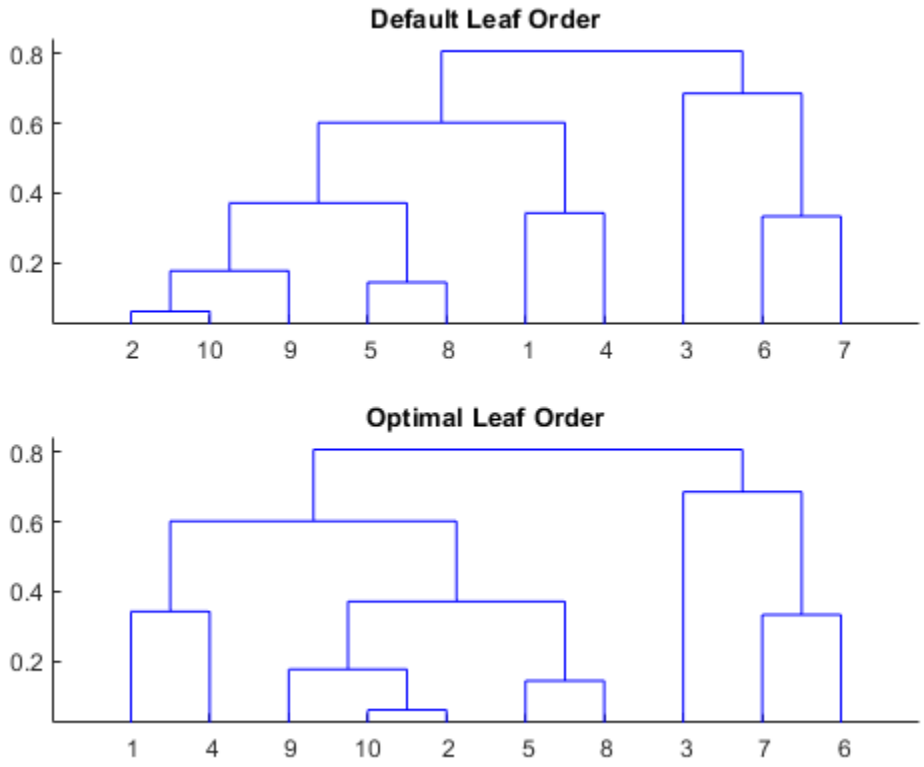
```
rng('default') % For reproducibility
X = rand(10,2);
```

Create a distance vector and a hierarchical binary clustering tree. Use the distances and clustering tree to determine an optimal leaf order.

```
D = pdist(X);
tree = linkage(D,'average');
leafOrder = optimalleaforder(tree,D);
```

Plot the dendrogram with the default ordering and the dendrogram with the optimal leaf ordering.

```
figure()
subplot(2,1,1)
dendrogram(tree)
title('Default Leaf Order')

subplot(2,1,2)
dendrogram(tree,'reorder',leafOrder)
title('Optimal Leaf Order')
```



The order of the leaves in the bottom figure corresponds to the elements in `leafOrder`.

```
leafOrder
```

```
leafOrder = 1×10

     1     4     9    10     2     5     8     3     7     6
```

Optimal Leaf Order Using Inverse Distance Similarity

Generate sample data.

```
rng('default') % For reproducibility
X = rand(10,2);
```

Create a distance vector and a hierarchical binary clustering tree.

```
D = pdist(X);
tree = linkage(D,'average');
```

Use the inverse distance similarity transformation to determine an optimal leaf order.

```
leafOrder = optimalleaforder(tree,D,'Transformation','inverse')
```

```
leafOrder = 1×10

     1     4     9    10     2     5     8     3     7     6
```

Input Arguments

`tree` — Hierarchical binary cluster tree

matrix returned by `linkage`

Hierarchical binary cluster tree, specified as an (M – 1)-by-3 matrix that you generate using `linkage`, where M is the number of leaves.

`D` — Distances

matrix | vector

Distances for determining similarities between leaves, specified as a matrix or vector of distances. For example, you can generate distances using `pdist`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Criteria','group','Transformation','inverse'` specifies that the sum of similarities be maximized between every leaf and all other leaves in adjacent clusters, using an inverse similarity transformation.

`Criteria` — Optimization criterion

`'adjacent'` (default) | `'group'`

Optimization criterion for determining an optimal leaf ordering, specified as the comma-separated pair consisting of `'criteria'` and one of these values:

`'adjacent'`	Maximize the sum of similarities between adjacent leaves.
`'group'`	Maximize the sum of similarities between every leaf and all other leaves in the adjacent clusters at the same level of the dendrogram.

Example: `'Criteria','group'`

`Transformation` — Method for transforming distances to similarities

`'linear'` (default) | `'inverse'` | function handle

Method for transforming distances to similarities, specified as the comma-separated pair consisting of `'Transformation'` and one of `'linear'`, `'inverse'`, or a function handle.

Let $d_{i,j}$ and $Sim_{i,j}$ denote the distance and similarity between leaves i and j, respectively. The included similarity transformations are:

`'linear'`	$Sim_{i,j} = \max_{i,j}(d_{i,j}) - d_{i,j}$
`'inverse'`	$Sim_{i,j} = 1/d_{i,j}$

To use a custom transformation function, specify a handle to a function that accepts a matrix of distances, `D`, and returns a matrix of similarities, `S`. The function should be monotonic decreasing in the range of distance values. `S` must have the same size as `D`, with `S(i,j)` being the similarity computed based on `D(i,j)`.

Example: `'Transformation',@myTransform`

Output Arguments

`leafOrder` — Optimal leaf order

vector

Optimal leaf order, returned as a length-*M* vector, where *M* is the number of leaves. `leafOrder` is a permutation of the vector `1:M`, giving an optimal leaf ordering based on the specified distances and similarity transformation.

References

[1] Bar-Joseph, Z., Gifford, D.K., and Jaakkola, T.S. (2001). "Fast optimal leaf ordering for hierarchical clustering." *Bioinformatics* Vol. 17, Suppl 1:S22–9. PMID: 11472989.

See Also

`dendrogram` | `linkage` | `pdist`

Introduced in R2012b

oobEdge

Out-of-bag classification edge

Syntax

```
edge = oobEdge(ens)
edge = oobEdge(ens,Name,Value)
```

Description

`edge = oobEdge(ens)` returns out-of-bag classification edge for `ens`.

`edge = oobEdge(ens,Name,Value)` computes classification edge with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A classification bagged ensemble, constructed with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `oobEdge` uses only these learners for calculating loss.

Default: `1:NumTrained`

`mode`

Character vector or string scalar representing the meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

edge

Classification edge, a weighted average of the classification margin.

Examples

Estimate Out-of-Bag Edge

Load Fisher's iris data set.

```
load fisheriris
```

Train an ensemble of 100 bagged classification trees using the entire data set.

```
Mdl = fitcensemble(meas,species,'Method','Bag');
```

Estimate the out-of-bag edge.

```
edge = oobEdge(Mdl)
```

```
edge = 0.8767
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `ens.Prior`.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `ens.X`.

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitcensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitcensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` take an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

See Also

`oobLoss` | `oobMargin` | `oobPredict`

oobError

Class: `TreeBagger`

Out-of-bag error

Syntax

```
err = oobError(B)
err = oobError(B,'param1',val1,'param2',val2,...)
```

Description

`err = oobError(B)` computes the misclassification probability (for classification trees) or mean squared error (for regression trees) for out-of-bag observations in the training data, using the trained bagger `B`. `err` is a vector of length `NTrees`, where `NTrees` is the number of trees in the ensemble.

`err = oobError(B,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Mode'`	Character vector or string scalar indicating how `oobError` computes errors. If set to `'cumulative'` (default), the method computes cumulative errors and `err` is a vector of length `NTrees`, where the first element gives error from `trees(1)`, second element gives error from `trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `err` is a vector of length `NTrees`, where each element is an error from each tree in the ensemble. If set to `'ensemble'`, `err` is a scalar showing the cumulative error for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives error from `trees(1)`, the second element gives error from `trees(1:2)` etc.
`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. `oobError` uses these weights to combine output from the specified trees by taking a weighted average instead of the simple nonweighted majority vote. You cannot use this argument in the `'individual'` mode.

Algorithms

`oobError` estimates the weighted ensemble error for out-of-bag observations. That is, `oobError` applies `error` to the training data stored in the input `TreeBagger` model `B`, and selects the out-of-bag observations for each tree to compose the ensemble error.

- `B.X` and `B.Y` are the training data predictors and responses, respectively.

- `B.OOBIndices` specifies which observations are out-of-bag for each tree in the ensemble.
- `B.W` specifies the observation weights.
- Optionally:
  - Using the `'Mode'` name-value pair argument, you can specify to return the individual, weighted ensemble error for each tree, or the entire, weighted ensemble error. By default, `oobError` returns the cumulative, weighted ensemble error.
  - Using the `'Trees'` name-value pair argument, you can choose which trees to use in the ensemble error calculations.
  - Using the `'TreeWeights'` name-value pair argument, you can attribute each tree with a weight.

`oobError` applies the algorithms described below. For more details, see `error` and `predict`.

For regression problems, `oobError` returns the weighted MSE.

1. `oobError` predicts responses for all out-of-bag observations.
2. The MSE estimate depends on the value of `'Mode'`.
   - If you specify `'Mode','Individual'`, then `oobError` sets any in bag observations within a selected tree to the weighted sample average of the observed, training data responses. Then, `oobError` computes the weighted MSE for each selected tree.
   - If you specify `'Mode','Cumulative'`, then `ooError` returns a vector of cumulative, weighted MSEs, where MSE_t is the cumulative, weighted MSE for selected tree t. To compute MSE_t, for each observation that is out of bag for at least one tree through tree t, `oobError` computes the cumulative, weighted mean of the predicted responses through tree t. `oobError` sets observations that are in bag for all selected trees through tree t to the weighted sample average of the observed, training data responses. Then, `oobError` computes MSE_t.
   - If you specify `'Mode','Ensemble'`, then, for each observation that is out of bag for at least one tree, `oobError` computes the weighted mean over all selected trees. `oobError` sets observations that are in bag for all selected trees to the weighted sample average of the observed, training data responses. Then, `oobError` computes the weighted MSE, which is the same as the final, cumulative, weighted MSE.

In classification problems, `oobError` returns the weighted misclassification rate.

1. `oobError` predicts classes for all out-of-bag observations.
2. The weighted misclassification rate estimate depends on the value of `'Mode'`.
   - If you specify `'Mode','Individual'`, then `oobError` sets any in bag observations within a selected tree to the predicted, weighted, most popular class over all training responses. If there are multiple most popular classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular. Then, `oobError` computes the weighted misclassification rate for each selected tree.
   - If you specify `'Mode','Cumulative'`, then `ooError` returns a vector of cumulative, weighted misclassification rates, where e_t^* is the cumulative, weighted misclassification rate for selected tree t. To compute e_t^*, for each observation that is out of bag for at least one tree through tree t, `oobError` finds the predicted, cumulative, weighted most popular class through tree t. `oobError` sets observations that are in bag for all selected trees through tree t to the weighted, most popular class over all training responses. If there are multiple most

popular classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular. Then, `oobError` computes e_t^*.

- If you specify `'Mode','Ensemble'`, then, for each observation that is out of bag for at least one tree, `oobError` computes the weighted, most popular class over all selected trees. `oobError` sets observations that are in bag for all selected trees through tree t to the predicted, weighted, most popular class over all training responses. If there are multiple most popular classes, `error` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular. Then, `oobError` computes the weighted misclassification rate , which is the same as the final, cumulative, weighted misclassification rate.

See Also

`TreeBagger` | `error` | `oobPredict` | `oobQuantileError` | `predict`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

oobLoss

Out-of-bag classification error

Syntax

```
L = oobLoss(ens)
L = oobLoss(ens,Name,Value)
```

Description

`L = oobLoss(ens)` returns the classification error for `ens` computed for out-of-bag data.

`L = oobLoss(ens,Name,Value)` computes error with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A classification bagged ensemble, constructed with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `oobLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

`lossfun`

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss

Value	Description
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

`'mincost'` is appropriate for classification scores that are posterior probabilities. Bagged ensembles return posterior probabilities as classification scores by default.

- Specify your own function using function handle notation.

  Suppose that `n` be the number of observations in `X` and `K` be the number of distinct classes (`numel(ens.ClassNames)`, `ens` is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `ens.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `ens.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.
  - `Cost` is a *K*-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see “Classification Loss” on page 33-4334.

Default: `'classiferror'`

`mode`

Character vector or string scalar representing the meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

L

Classification loss on page 33-4334 of the out-of-bag observations, a scalar. `L` can be a vector, or can represent a different quantity, depending on the name-value settings.

Examples

Estimate Out-Of-Bag Error

Load Fisher's iris data set.

```
load fisheriris
```

Grow a bag of 100 classification trees.

```
ens = fitcensemble(meas,species,'Method','Bag');
```

Estimate the out-of-bag classification error.

```
L = oobLoss(ens)
```

```
L = 0.0400
```

More About

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitcensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitcensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` take an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.

- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

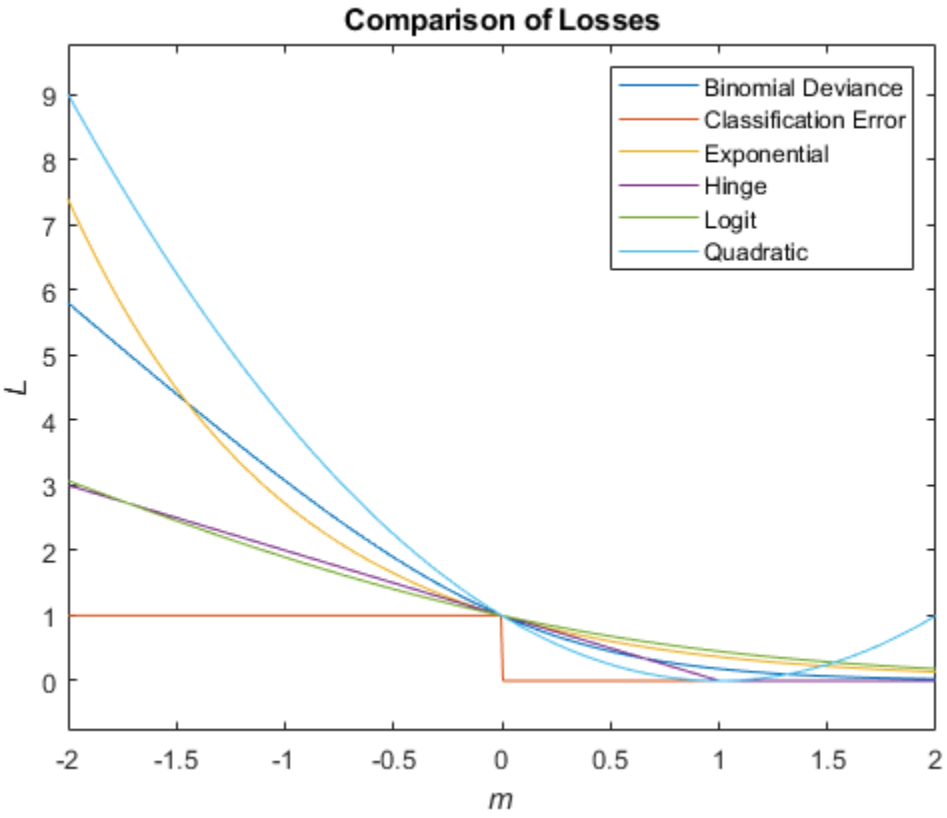
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}$.
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.

Loss Function	Value of LossFun	Equation
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $$L = -\sum_{j=1}^{n}\frac{\tilde{w}_j \log(m_j)}{Kn},$$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`loss` | `oobEdge` | `oobMargin` | `oobPredict`

oobLoss

Out-of-bag regression error

Syntax

```
L = oobLoss(ens)
L = oobLoss(ens,Name,Value)
```

Description

`L = oobLoss(ens)` returns the mean squared error for `ens` computed for out-of-bag data.

`L = oobLoss(ens,Name,Value)` computes error with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression bagged ensemble, constructed with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `oobLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

`lossfun`

Function handle for loss function, or `'mse'`, meaning mean squared error. If you pass a function handle `fun`, `oobLoss` calls it as

```
FUN(Y,Yfit,W)
```

where `Y`, `Yfit`, and `W` are numeric vectors of the same length. `Y` is the observed response, `Yfit` is the predicted response, and `W` is the observation weights.

Default: `'mse'`

`mode`

Character vector or string scalar representing the meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

`L`

Mean squared error of the out-of-bag observations, a scalar. `L` can be a vector, or can represent a different quantity, depending on the name-value settings.

Examples

Find Out-of-Bag Regression Error

Compute the out-of-bag error for the `carsmall` data.

Load the `carsmall` data set and select engine displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train an ensemble of bagged regression trees.

```
ens = fitrensemble(X,MPG,'Method','Bag');
```

Find the out-of-bag error.

```
L = oobLoss(ens)
```

```
L = 16.9551
```

More About

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitrensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitrensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` takes an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses

against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

See Also

`loss` | `oobPredict`

oobMargin

Out-of-bag classification margins

Syntax

```
margin = oobMargin(ens)
margin = oobMargin(ens,Name,Value)
```

Description

`margin = oobMargin(ens)` returns out-of-bag classification margins.

`margin = oobMargin(ens,Name,Value)` calculates margins with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A classification bagged ensemble, constructed with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `oobEdge` uses only these learners for calculating loss.

Default: `1:NumTrained`

Output Arguments

`margin`

A numeric column vector of length `size(ens.X,1)`.

Examples

Find Out-of-Bag Classification Margins

Find the out-of-bag margins for a bagged ensemble from the Fisher iris data.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of bagged classification trees.

```
ens = fitcensemble(meas,species,'Method','Bag');
```

Find the number of out-of-bag margins that are equal to 1.

```
margin = oobMargin(ens);
sum(margin == 1)
```

```
ans = 109
```

More About

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitrensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitrensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` takes an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `ens.X`.

See Also

`margin` | `oobEdge` | `oobLoss` | `oobPredict`

oobMargin

Class: `TreeBagger`

Out-of-bag margins

Syntax

```
mar = oobMargin(B)
mar = oobMargin(B,'param1',val1,'param2',val2,...)
```

Description

`mar = oobMargin(B)` computes an `Nobs`-by-`NTrees` matrix of classification margins for out-of-bag observations in the training data, using the trained bagger `B`.

`mar = oobMargin(B,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Mode'`	Character vector or string scalar indicating how `oobMargin` computes errors. If set to `'cumulative'` (default), the method computes cumulative margins and `mar` is an `Nobs`-by-`NTrees` matrix, where the first column gives margins from `trees(1)`, second column gives margins from `trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is an `Nobs`-by-`NTrees` matrix, where each column gives margins from each tree in the ensemble. If set to `'ensemble'`, `mar` is a single column of length `Nobs` showing the cumulative margins for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns an `Nobs`-by-`NTrees` matrix for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a single column for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first column gives margins from `trees(1)`, the second column gives margins from `trees(1:2)` etc.
`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. `oobMargin` uses these weights to combine output from the specified trees by taking a weighted average instead of the simple nonweighted majority vote. You cannot use this argument in the `'individual'` mode.

See Also

`margin`

oobMeanMargin

Class: `TreeBagger`

Out-of-bag mean margins

Syntax

```
mar = oobMeanMargin(B)
mar = oobMeanMargin(B,'param1',val1,'param2',val2,...)
```

Description

`mar = oobMeanMargin(B)` computes average classification margins for out-of-bag observations in the training data, using the trained bagger `B`. `oobMeanMargin` averages the margins over all out-of-bag observations. `mar` is a row-vector of length `NTrees`, where `NTrees` is the number of trees in the ensemble.

`mar = oobMeanMargin(B,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Mode'`	Character vector or string scalar indicating how `oobMeanMargin` computes errors. If set to `'cumulative'` (default), is a vector of length `NTrees` where the first element gives mean margin from `trees(1)`, second column gives mean margins from `trees(1:2)` etc., up to `trees(1:NTrees)`. If set to `'individual'`, `mar` is a vector of length `NTrees`, where each element is a mean margin from each tree in the ensemble. If set to `'ensemble'`, `mar` is a scalar showing the cumulative mean margin for the entire ensemble.
`'Trees'`	Vector of indices indicating what trees to include in this calculation. By default, this argument is set to `'all'` and the method uses all trees. If `'Trees'` is a numeric vector, the method returns a vector of length `NTrees` for `'cumulative'` and `'individual'` modes, where `NTrees` is the number of elements in the input vector, and a scalar for `'ensemble'` mode. For example, in the `'cumulative'` mode, the first element gives mean margin from `trees(1)`, the second element gives mean margin from `trees(1:2)` etc.
`'TreeWeights'`	Vector of tree weights. This vector must have the same length as the `'Trees'` vector. `oobMeanMargin` uses these weights to combine output from the specified trees by taking a weighted average instead of the simple nonweighted majority vote. You cannot use this argument in the `'individual'` mode.

See Also

`meanMargin`

oobPermutedPredictorImportance

Predictor importance estimates by permutation of out-of-bag predictor observations for random forest of classification trees

Syntax

```
Imp = oobPermutedPredictorImportance(Mdl)
Imp = oobPermutedPredictorImportance(Mdl,Name,Value)
```

Description

`Imp = oobPermutedPredictorImportance(Mdl)` returns a vector of out-of-bag, predictor importance estimates by permutation on page 33-4351 using the random forest of classification trees `Mdl`. `Mdl` must be a `ClassificationBaggedEnsemble` model object.

`Imp = oobPermutedPredictorImportance(Mdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, you can speed up computation using parallel computing or indicate which trees to use in the predictor importance estimation.

Input Arguments

`Mdl` — Random forest of classification trees

`ClassificationBaggedEnsemble` model object

Random forest of classification trees, specified as a `ClassificationBaggedEnsemble` model object created by `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Learners` — Indices of learners to use in predictor importance estimation

`1:Mdl.NumTrained` (default) | numeric vector of positive integers

Indices of learners to use in predictor importance estimation, specified as the comma-separated pair consisting of `'Learners'` and a numeric vector of positive integers. Values must be at most `Mdl.NumTrained`. When `oobPermutedPredictorImportance` estimates the predictor importance, it includes the learners in `Mdl.Trained(`*`learners`*`)` only, where *`learners`* is the value of `'Learners'`.

Example: `'Learners',[1:2:Mdl.NumTrained]`

`Options` — Parallel computing options

`[]` (default) | structure array returned by `statset`

Parallel computing options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. `'Options'` requires a Parallel Computing Toolbox license.

`oobPermutedPredictorImportance` uses the `'UseParallel'` field only. `statset('UseParallel',true)` invokes a pool of workers.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

`Imp` — Out-of-bag, predictor importance estimates by permutation

numeric vector

Out-of-bag, predictor importance estimates by permutation on page 33-4351, returned as a 1-by-*p* numeric vector. *p* is the number of predictor variables in the training data (`size(Mdl.X,2)`). `Imp(`*j*`)` is the predictor importance of the predictor `Mdl.PredictorNames(`*j*`)`.

Examples

Estimate Importance of Predictors

Load the `census1994` data set. Consider a model that predicts a person's salary category given their age, working class, education level, martial status, race, sex, capital gain and loss, and number of working hours per week.

```
load census1994
X = adultdata(:,{'age','workClass','education_num','marital_status','race',...
    'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

You can train a random forest of 50 classification trees using the entire data set.

```
Mdl = fitcensemble(X,'salary','Method','Bag','NumLearningCycles',50);
```

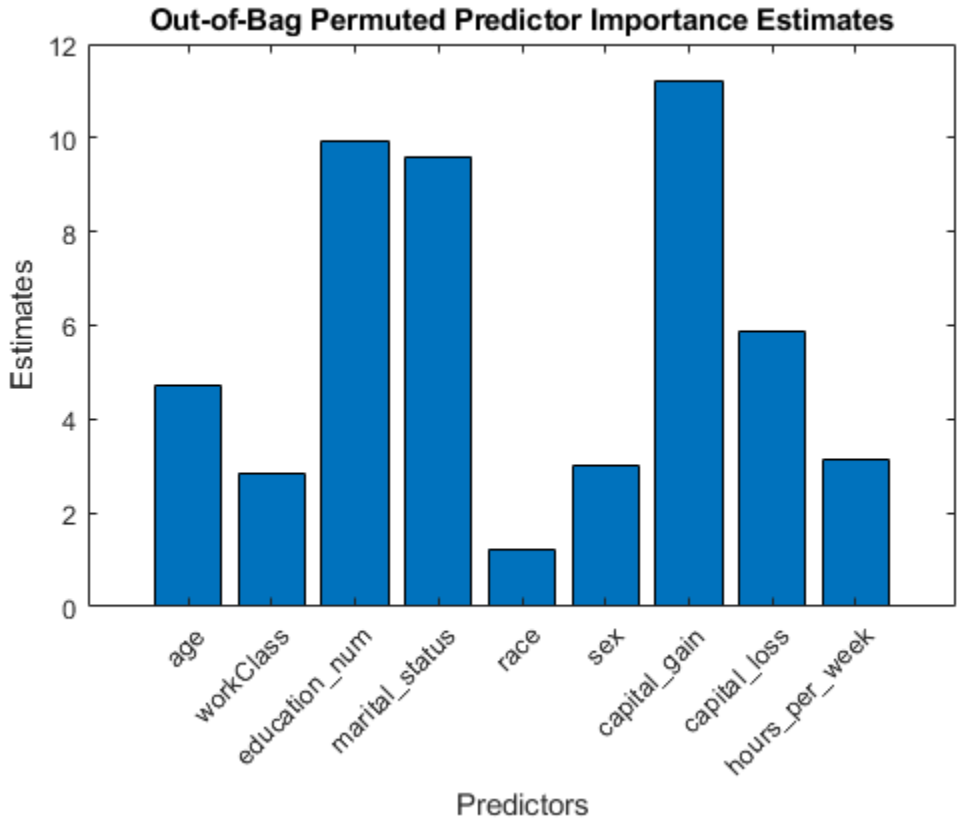
`fitcensemble` uses a default template tree object `templateTree()` as a weak learner when `'Method'` is `'Bag'`. In this example, for reproducibility, specify `'Reproducible',true` when you create a tree template object, and then use the object as a weak learner.

```
rng('default') % For reproducibility
t = templateTree('Reproducible',true); % For reproducibiliy of random predictor selections
Mdl = fitcensemble(X,'salary','Method','Bag','NumLearningCycles',50,'Learners',t);
```

`Mdl` is a `ClassificationBaggedEnsemble` model.

Estimate predictor importance measures by permuting out-of-bag observations. Compare the estimates using a bar graph.

```
imp = oobPermutedPredictorImportance(Mdl);

figure;
bar(imp);
title('Out-of-Bag Permuted Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



`imp` is a 1-by-9 vector of predictor importance estimates. Larger values indicate predictors that have a greater influence on predictions. In this case, `marital_status` is the most important predictor, followed by `capital_gain`.

Unbiased Estimates of Predictor Importance Using Parallel Computing

Load the `census1994` data set. Consider a model that predicts a person's salary category given their age, working class, education level, martial status, race, sex, capital gain and loss, and number of working hours per week.

```
load census1994
X = adultdata(:,{'age','workClass','education_num','marital_status','race', ...
    'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

Display the number of categories represented in the categorical variables using `summary`.

```
summary(X)
```

```
Variables:

    age: 32561×1 double

        Values:

            Min        17
```

```
        Median     37
        Max        90

workClass: 32561×1 categorical

    Values:

        Federal-gov              960
        Local-gov               2093
        Never-worked               7
        Private                22696
        Self-emp-inc            1116
        Self-emp-not-inc        2541
        State-gov               1298
        Without-pay               14
        NumMissing              1836

education_num: 32561×1 double

    Values:

        Min              1
        Median          10
        Max             16

marital_status: 32561×1 categorical

    Values:

        Divorced                       4443
        Married-AF-spouse                23
        Married-civ-spouse            14976
        Married-spouse-absent           418
        Never-married                 10683
        Separated                      1025
        Widowed                         993

race: 32561×1 categorical

    Values:

        Amer-Indian-Eskimo      311
        Asian-Pac-Islander     1039
        Black                  3124
        Other                   271
        White                 27816

sex: 32561×1 categorical

    Values:

        Female    10771
        Male      21790

capital_gain: 32561×1 double

    Values:
```

```
            Min               0
            Median            0
            Max           99999

    capital_loss: 32561×1 double

        Values:

            Min               0
            Median            0
            Max            4356

    hours_per_week: 32561×1 double

        Values:

            Min               1
            Median           40
            Max              99

    salary: 32561×1 categorical

        Values:

            <=50K     24720
            >50K       7841
```

Because there are few categories represented in the categorical variables compared to levels in the continuous variables, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over the categorical variables.

Train a random forest of 50 classification trees using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits. To reproduce random predictor selections, set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

```
rng('default') % For reproducibility
t = templateTree('PredictorSelection','curvature','Surrogate','on', ...
    'Reproducible',true); % For reproducibility of random predictor selections
Mdl = fitcensemble(X,'salary','Method','bag','NumLearningCycles',50, ...
    'Learners',t);
```

Estimate predictor importance measures by permuting out-of-bag observations. Perform calculations in parallel.

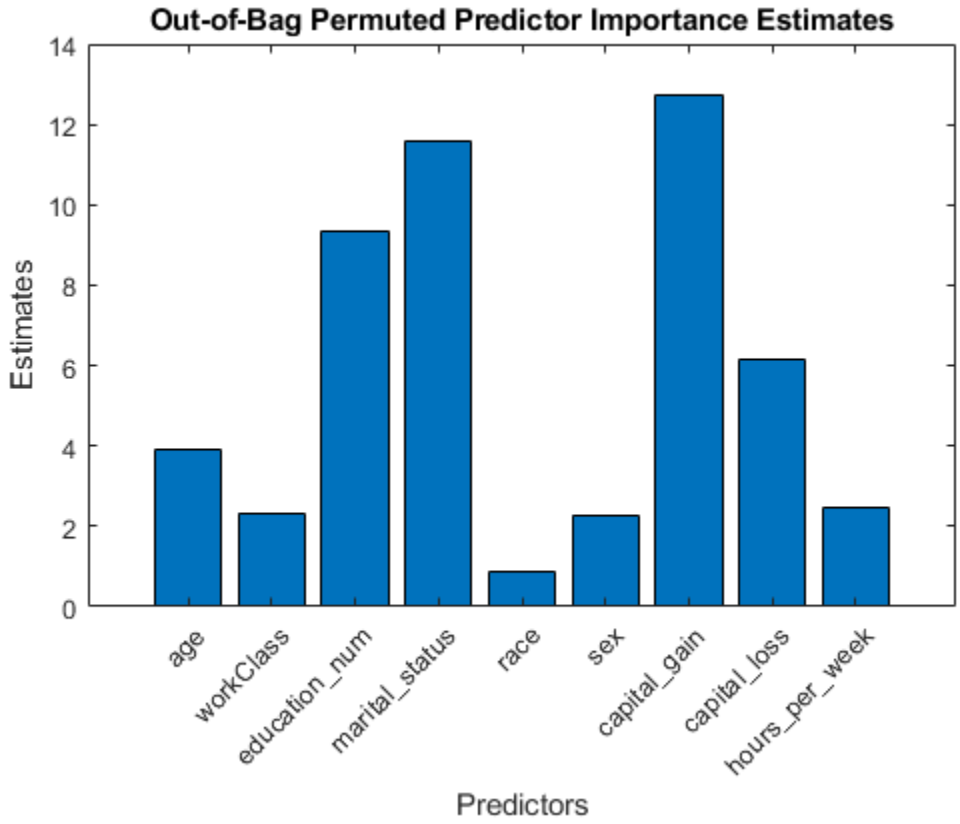
```
options = statset('UseParallel',true);
imp = oobPermutedPredictorImportance(Mdl,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Compare the estimates using a bar graph.

```
figure
bar(imp)
title('Out-of-Bag Permuted Predictor Importance Estimates')
ylabel('Estimates')
xlabel('Predictors')
```

```
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `capital_gain` is the most important predictor, followed by `martial_status`. Compare these results to the results in “Estimate Importance of Predictors” on page 33-4347.

More About

Out-of-Bag, Predictor Importance Estimates by Permutation

Out-of-bag, predictor importance estimates by permutation measure how influential the predictor variables in the model are at predicting the response. The influence of a predictor increases with the value of this measure.

If a predictor is influential in prediction, then permuting its values should affect the model error. If a predictor is not influential, then permuting its values should have little to no effect on the model error.

The following process describes the estimation of out-of-bag predictor importance values by permutation. Suppose that R is a random forest of T learners and p is the number of predictors in the training data.

1 For tree t, $t = 1,...,T$:

a Identify the out-of-bag observations and the indices of the predictor variables that were split to grow tree t, $s_t \subseteq \{1,...,p\}$.

b Estimate the out-of-bag error ε_t.

c For each predictor variable x_j, $j \in s_t$:

i Randomly permute the observations of x_j.

ii Estimate the model error, ε_{tj}, using the out-of-bag observations containing the permuted values of x_j.

iii Take the difference $d_{tj} = \varepsilon_{tj} - \varepsilon_t$. Predictor variables not split when growing tree t are attributed a difference of 0.

2 For each predictor variable in the training data, compute the mean, $\bar{d}_j$, and standard deviation, σ_j, of the differences over the learners, $j = 1,...,p$.

3 The out-of-bag predictor importance by permutation for x_j is $\bar{d}_j/\sigma_j$.

Tips

When growing a random forest using `fitcensemble`:

- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [3]. If the predictor data set is heterogeneous, or if there are predictors that have relatively fewer distinct values than other variables, then consider specifying the curvature or interaction test.
- Trees grown using standard CART are not sensitive to predictor variable interactions. Also, such trees are less likely to identify important variables in the presence of many irrelevant predictors than the application of the interaction test. Therefore, to account for predictor interactions and identify importance variables in the presence of many irrelevant variables, specify the interaction test [2].
- If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` of the `templateTree` function as `'all'` for the tree learners of the ensemble. Otherwise, the software might not select some predictors, underestimating their importance.

For more details, see `templateTree` and “Choose Split Predictor Selection Technique” on page 19-14.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[2] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[3] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationBaggedEnsemble` | `fitcensemble` | `predictorImportance`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Select Predictors for Random Forests” on page 18-60
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

Introduced in R2016b

oobPermutedPredictorImportance

Predictor importance estimates by permutation of out-of-bag predictor observations for random forest of regression trees

Syntax

```
Imp = oobPermutedPredictorImportance(Mdl)
Imp = oobPermutedPredictorImportance(Mdl,Name,Value)
```

Description

`Imp = oobPermutedPredictorImportance(Mdl)` returns a vector of out-of-bag, predictor importance estimates by permutation on page 33-4358 using the random forest of regression trees `Mdl`. `Mdl` must be a `RegressionBaggedEnsemble` model object.

`Imp = oobPermutedPredictorImportance(Mdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, you can speed up computation using parallel computing or indicate which trees to use in the predictor importance estimation.

Input Arguments

`Mdl` — Random forest of regression trees

`RegressionBaggedEnsemble` model object

Random forest of regression trees, specified as a `RegressionBaggedEnsemble` model object created by `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Learners` — Indices of learners to use in predictor importance estimation

`1:Mdl.NumTrained` (default) | numeric vector of positive integers

Indices of learners to use in predictor importance estimation, specified as the comma-separated pair consisting of `'Learners'` and a numeric vector of positive integers. Values must be at most `Mdl.NumTrained`. When `oobPermutedPredictorImportance` estimates the predictor importance, it includes the learners in `Mdl.Trained(`*`learners`*`)` only, where *`learners`* is the value of `'Learners'`.

Example: `'Learners',[1:2:Mdl.NumTrained]`

`Options` — Parallel computing options

`[]` (default) | structure array returned by `statset`

Parallel computing options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. `'Options'` requires a Parallel Computing Toolbox license.

`oobPermutedPredictorImportance` uses the `'UseParallel'` field only. `statset('UseParallel',true)` invokes a pool of workers.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

`Imp` — Out-of-bag, predictor importance estimates by permutation

numeric vector

Out-of-bag, predictor importance estimates by permutation on page 33-4358, returned as a 1-by-*p* numeric vector. *p* is the number of predictor variables in the training data (`size(Mdl.X,2)`). `Imp(`*j*`)` is the predictor importance of the predictor `Mdl.PredictorNames(`*j*`)`.

Examples

Estimate Importance of Predictors

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
```

You can train a random forest of 500 regression trees using the entire data set.

```
Mdl = fitrensemble(X,'MPG','Method','Bag','NumLearningCycles',500);
```

`fitrensemble` uses a default template tree object `templateTree()` as a weak learner when `'Method'` is `'Bag'`. In this example, for reproducibility, specify `'Reproducible',true` when you create a tree template object, and then use the object as a weak learner.

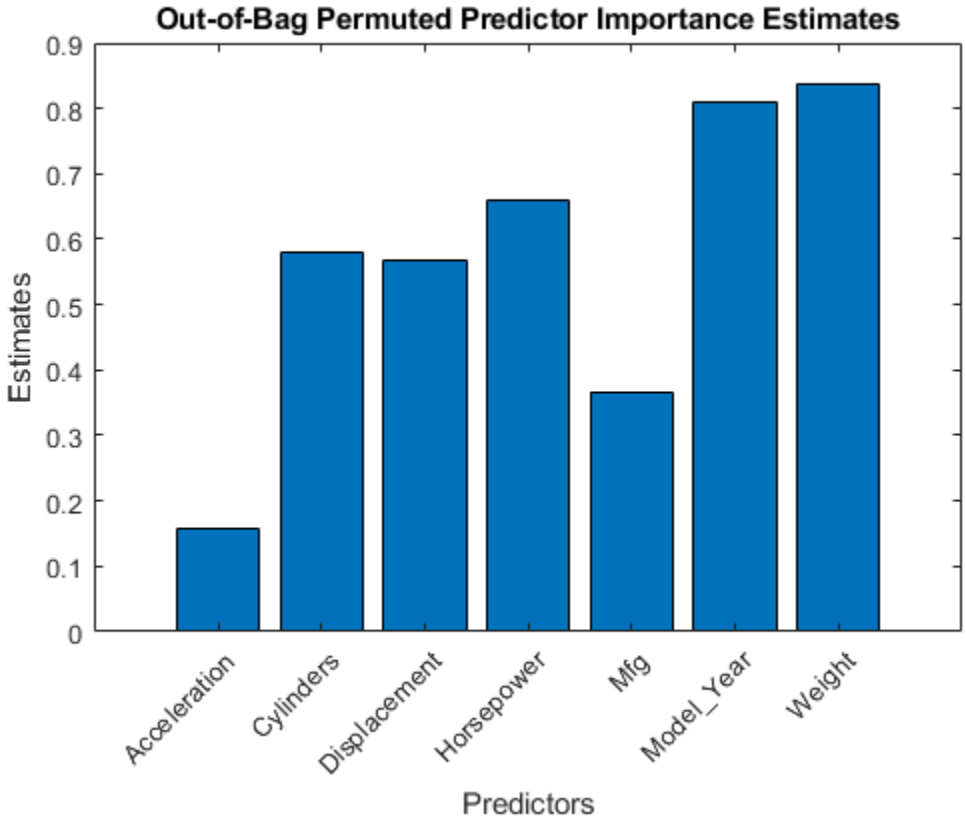
```
rng('default') % For reproducibility
t = templateTree('Reproducible',true); % For reproducibiliy of random predictor selections
Mdl = fitrensemble(X,'MPG','Method','Bag','NumLearningCycles',500,'Learners',t);
```

`Mdl` is a `RegressionBaggedEnsemble` model.

Estimate predictor importance measures by permuting out-of-bag observations. Compare the estimates using a bar graph.

```
imp = oobPermutedPredictorImportance(Mdl);

figure;
bar(imp);
title('Out-of-Bag Permuted Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
```

```
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



`imp` is a 1-by-7 vector of predictor importance estimates. Larger values indicate predictors that have a greater influence on predictions. In this case, `Weight` is the most important predictor, followed by `Model_Year`.

Unbiased Estimates of Predictor Importance Using Parallel Computing

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a random forest of 500 regression trees using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits. To reproduce random predictor selections, set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

```
rng('default'); % For reproducibility
t = templateTree('PredictorSelection','curvature','Surrogate','on', ...
    'Reproducible',true); % For reproducibility of random predictor selections
Mdl = fitrensemble(X,'MPG','Method','bag','NumLearningCycles',500, ...
    'Learners',t);
```

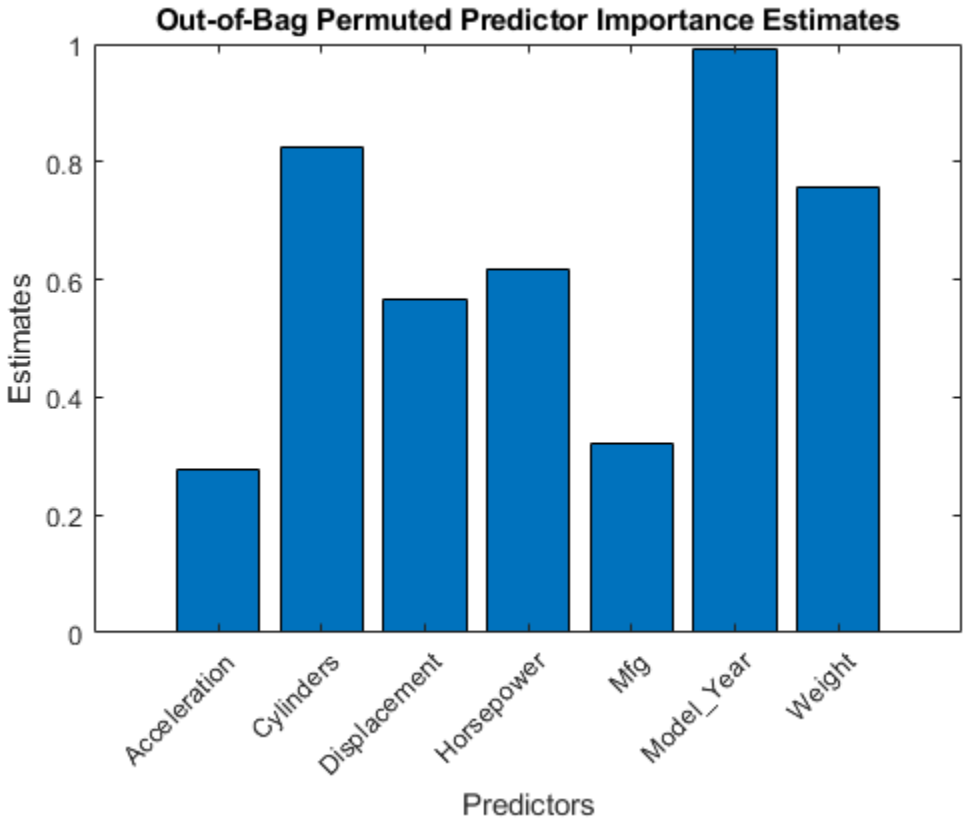
Estimate predictor importance measures by permuting out-of-bag observations. Perform calculations in parallel.

```
options = statset('UseParallel',true);
imp = oobPermutedPredictorImportance(Mdl,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Compare the estimates using a bar graph.

```
figure;
bar(imp);
title('Out-of-Bag Permuted Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Model_Year` is the most important predictor, followed by `Cylinders`. Compare these results to the results in “Estimate Importance of Predictors” on page 33-4355.

More About

Out-of-Bag, Predictor Importance Estimates by Permutation

Out-of-bag, predictor importance estimates by permutation measure how influential the predictor variables in the model are at predicting the response. The influence of a predictor increases with the value of this measure.

If a predictor is influential in prediction, then permuting its values should affect the model error. If a predictor is not influential, then permuting its values should have little to no effect on the model error.

The following process describes the estimation of out-of-bag predictor importance values by permutation. Suppose that R is a random forest of T learners and p is the number of predictors in the training data.

1. For tree t, $t = 1,...,T$:
   a. Identify the out-of-bag observations and the indices of the predictor variables that were split to grow tree t, $s_t \subseteq \{1,...,p\}$.
   b. Estimate the out-of-bag error ε_t.

c For each predictor variable x_j, $j \in s_t$:

i Randomly permute the observations of x_j.

ii Estimate the model error, ε_{tj}, using the out-of-bag observations containing the permuted values of x_j.

iii Take the difference $d_{tj} = \varepsilon_{tj} - \varepsilon_t$. Predictor variables not split when growing tree t are attributed a difference of 0.

2 For each predictor variable in the training data, compute the mean, $\bar{d}_j$, and standard deviation, σ_j, of the differences over the learners, $j = 1,...,p$.

3 The out-of-bag predictor importance by permutation for x_j is $\bar{d}_j/\sigma_j$.

Tips

When growing a random forest using `fitrensemble`:

- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [3]. If the predictor data set is heterogeneous, or if there are predictors that have relatively fewer distinct values than other variables, then consider specifying the curvature or interaction test.
- Trees grown using standard CART are not sensitive to predictor variable interactions. Also, such trees are less likely to identify important variables in the presence of many irrelevant predictors than the application of the interaction test. Therefore, to account for predictor interactions and identify importance variables in the presence of many irrelevant variables, specify the interaction test [2].
- If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` of the `templateTree` function as `'all'` for the tree learners of the ensemble. Otherwise, the software might not select some predictors, underestimating their importance.

For more details, see `templateTree` and “Choose Split Predictor Selection Technique” on page 19-14.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[2] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[3] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionBaggedEnsemble` | `fitrensemble` | `plotPartialDependence` | `predictorImportance`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Select Predictors for Random Forests” on page 18-60
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

Introduced in R2016b

oobPredict

Predict out-of-bag response of ensemble

Syntax

```
[label,score] = oobPredict(ens)
[label,score] = oobPredict(ens,Name,Value)
```

Description

`[label,score] = oobPredict(ens)` returns class labels and scores for `ens` for out-of-bag data.

`[label,score] = oobPredict(ens,Name,Value)` computes labels and scores with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A classification bagged ensemble, constructed with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `ens.NumTrained`. `oobEdge` uses only these learners for calculating loss.

Default: `1:NumTrained`

Output Arguments

`label`

Classification labels of the same data type as the training data `Y`. (The software treats string arrays as cell arrays of character vectors.) There are `N` elements or rows, where `N` is the number of training observations. The label is the class with the highest score. In case of a tie, the label is earliest in `ens.ClassNames`.

`score`

An `N`-by-`K` numeric matrix for `N` observations and `K` classes. A high score indicates that an observation is likely to come from this class. Scores are in the range `0` to `1`.

Examples

Find Out-of-Bag Response of Classification Ensemble

Find the out-of-bag predictions and scores for the Fisher iris data. Find the scores with notable uncertainty in the resulting classifications.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of bagged classification trees.

```
ens = fitcensemble(meas,species,'Method','Bag');
```

Find the out-of-bag predictions and scores.

```
[label,score] = oobPredict(ens);
```

Find the scores in the range `(0.2,0.8)`. These scores have notable uncertainty in the resulting classifications.

```
unsure = ((score > .2) & (score < .8));
sum(sum(unsure))  % Number of uncertain predictions
```

```
ans = 16
```

More About

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitrensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitrensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` takes an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

See Also

`oobEdge` | `oobLoss` | `oobMargin` | `predict`

oobPredict

Predict out-of-bag response of ensemble

Syntax

```
Yfit = oobPredict(ens)
Yfit = oobPredict(ens,Name,Value)
```

Description

`Yfit = oobPredict(ens)` returns the predicted responses for the out-of-bag data in `ens`.

`Yfit = oobPredict(ens,Name,Value)` predicts responses with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A regression bagged ensemble, constructed with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `oobLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

Output Arguments

`Yfit`

A vector of predicted responses for out-of-bag data. `Yfit` has `size(ens.X,1)` elements.

You can find the indices of out-of-bag observations for weak learner `L` with the command

```
~ens.UseObsForLearner(:,L)
```

Examples

Find Out-of-Bag Response of Regression Ensemble

Compute the out-of-bag predictions for the `carsmall` data set. Display the first three terms of the fit.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train an ensemble of bagged regression trees.

```
ens = fitrensemble(X,MPG,'Method','Bag');
```

Find the out-of-bag predictions, and display the first three terms of the fit.

```
Yfit = oobPredict(ens);
Yfit(1:3) % First three terms
```

```
ans = 3×1

   15.5200
   14.5558
   15.0231
```

More About

Out of Bag

Bagging, which stands for “bootstrap aggregation”, is a type of ensemble learning. To bag a weak learner such as a decision tree on a dataset, `fitrensemble` generates many bootstrap replicas of the dataset and grows decision trees on these replicas. `fitrensemble` obtains each bootstrap replica by randomly selecting `N` observations out of `N` with replacement, where `N` is the dataset size. To find the predicted response of a trained ensemble, `predict` takes an average over predictions from individual trees.

Drawing `N` out of `N` observations with replacement omits on average 37% (1/*e*) of observations for each decision tree. These are "out-of-bag" observations. For each observation, `oobLoss` estimates the out-of-bag prediction by averaging over predictions from all trees in the ensemble for which this observation is out of bag. It then compares the computed prediction against the true response for this observation. It calculates the out-of-bag error by comparing the out-of-bag predicted responses against the true responses for all observations used for training. This out-of-bag average is an unbiased estimator of the true ensemble error.

See Also

`oobLoss` | `predict`

oobPredict

Class: `TreeBagger`

Ensemble predictions for out-of-bag observations

Syntax

```
Y = oobPredict(B)
Y = oobPredict(B,Name,Value)
[Y,stdevs] = oobPredict( ___ )
[Y,scores] = oobPredict( ___ )
[Y,scores,stdevs] = oobPredict( ___ )
```

Description

`Y = oobPredict(B)` computes predicted responses using the trained bagger `B` for out-of-bag observations in the training data. The output has one prediction for each observation in the training data. The returned Y is a cell array of character vectors for classification and a numeric array for regression.

`Y = oobPredict(B,Name,Value)` specifies additional options using one or both name-value pair arguments:

- `'Trees'` — Array of tree indices to use for computation of responses. The default is `'all'`.
- `'TreeWeights'` — Array of `NTrees` weights for weighting votes from the specified trees, where `NTrees` is the number of trees in the ensemble.

For regression, `[Y,stdevs] = oobPredict( ___ )` also returns standard deviations of the computed responses over the ensemble of the grown trees using any of the input argument combinations in previous syntaxes.

For classification, `[Y,scores] = oobPredict( ___ )` also returns scores for all classes. `scores` is a matrix with one row per observation and one column per class. For each out-of-bag observation and each class, the score generated by each tree is the probability of the observation originating from the class, computed as the fraction of observations of the class in a tree leaf. `oobPredict` averages these scores over all trees in the ensemble.

`[Y,scores,stdevs] = oobPredict( ___ )` also returns standard deviations of the computed scores for classification. `stdevs` is a matrix with one row per observation and one column per class, with standard deviations taken over the ensemble of the grown trees.

Algorithms

`oobPredict` and `predict` similarly predict classes and responses.

- In regression problems:
  - For each observation that is out of bag for at least one tree, `oobPredict` composes the weighted mean by selecting responses of trees in which the observation is out of bag. For this computation, the `'TreeWeights'` name-value pair argument specifies the weights.

    - For each observation that is in bag for all trees, the predicted response is the weighted mean of all of the training responses. For this computation, the `W` property of the `TreeBagger` model (i.e., the observation weights) specify the weights.
- In classification problems:
    - For each observation that is out of bag for at least one tree, `oobPredict` composes the weighted mean of the class posterior probabilities by selecting the trees in which the observation is out of bag. Consequently, the predicted class is the class corresponding to the largest weighted mean. For this computation, the `'TreeWeights'` name-value pair argument specifies the weights.
    - For each observation that is in bag for all trees, the predicted class is the weighted, most popular class over all training responses. For this computation, the `W` property of the `TreeBagger` model (i.e., the observation weights) specify the weights. If there are multiple most popular classes, `oobPredict` considers the one listed first in the `ClassNames` property of the `TreeBagger` model the most popular.

See Also

`TreeBagger` | `compact` | `oobQuantilePredict` | `predict`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

oobQuantileError

Class: `TreeBagger`

Out-of-bag quantile loss of bag of regression trees

Syntax

```
err = oobQuantileError(Mdl)
err = oobQuantileError(Mdl,Name,Value)
```

Description

`err = oobQuantileError(Mdl)` returns half of the out-of-bag on page 33-4372 mean absolute deviation (MAD) from comparing the true responses in `Mdl.Y` to the predicted, out-of-bag medians at `Mdl.X`, the predictor data, and using the bag of regression trees `Mdl`. `Mdl` must be a `TreeBagger` model object.

`err = oobQuantileError(Mdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify quantile probabilities, the error type, or which trees to include in the quantile-regression-error estimation.

Input Arguments

`Mdl` — Bag of regression trees

`TreeBagger` model object (default)

Bag of regression trees, specified as a `TreeBagger` model object created by `TreeBagger`.

- The value of `Mdl.Method` must be `regression`.
- When you train `Mdl` using `TreeBagger`, you must specify the name-value pair `'OOBPrediction','on'`. Consequently, `TreeBagger` saves required out-of-bag observation index matrix in `Mdl.OOBIndices`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Mode` — Ensemble error type

`'ensemble'` (default) | `'cumulative'` | `'individual'`

Ensemble error type, specified as the comma-separated pair consisting of `'Mode'` and a value in this table. Suppose *`tau`* is the value of `Quantile`.

Value	Description
`'cumulative'`	`err` is a `Mdl.NumTrees`-by-`numel(tau)` numeric matrix of cumulative quantile regression errors. `err(j,k)` is the `tau(k)` quantile regression error using the learners in `Mdl.Trees(1:j)` only.
`'ensemble'`	`err` is a 1-by-`numel(tau)` numeric vector of cumulative quantile regression errors for the entire ensemble. `err(k)` is the `tau(k)` ensemble quantile regression error.
`'individual'`	`err` is a `Mdl.NumTrees`-by-`numel(tau)` numeric matrix of quantile regression errors from individual learners. `err(j,k)` is the `tau(k)` quantile regression error using the learner in `Mdl.Trees(j)` only.

For `'cumulative'` and `'individual'`, if you choose to include fewer trees in quantile estimation using `Trees`, then this action affects the number of rows in `err` and corresponding row indices.

Example: `'Mode','cumulative'`

`Quantile` — Quantile probability
`0.5` (default) | numeric vector containing values in [0,1]

Quantile probability, specified as the comma-separated pair consisting of `'Quantile'` and a numeric vector containing values in the interval [0,1]. For each observation (row) in `Mdl.X`, `oobQuantileError` estimates corresponding quantiles for all probabilities in `Quantile`.

Example: `'Quantile',[0 0.25 0.5 0.75 1]`

Data Types: `single` | `double`

`Trees` — Indices of trees to use in response estimation
`'all'` (default) | numeric vector of positive integers

Indices of trees to use in response estimation, specified as the comma-separated pair consisting of `'Trees'` and `'all'` or a numeric vector of positive integers. Indices correspond to the cells of `Mdl.Trees`; each cell therein contains a tree in the ensemble. The maximum value of `Trees` must be less than or equal to the number of trees in the ensemble (`Mdl.NumTrees`).

For `'all'`, `oobQuantileError` uses all trees in the ensemble (that is, the indices `1:Mdl.NumTrees`).

Values other than the default can affect the number of rows in `err`.

Example: `'Trees',[1 10 Mdl.NumTrees]`

Data Types: `char` | `string` | `single` | `double`

`TreeWeights` — Weights to attribute to responses from individual trees
`ones(Mdl.NumTrees,1)` (default) | numeric vector of nonnegative values

Weights to attribute to responses from individual trees, specified as the comma-separated pair consisting of `'TreeWeights'` and a numeric vector of `numel(trees)` nonnegative values. `trees` is the value of `Trees`.

If you specify `'Mode','individual'`, then `oobQuantileError` ignores `TreeWeights`.

Data Types: `single` | `double`

Output Arguments

`err` — Half of out-of-bag quantile regression error

numeric scalar | numeric matrix

Half of the out-of-bag quantile regression error on page 33-4372, returned as a numeric scalar or *T*-by-`numel(tau)` matrix. *`tau`* is the value of `Quantile`.

T depends on the values of `Mode`, `Trees`, and `Quantile`. Suppose that you specify `'Quantile',tau` and `'Trees',trees`.

- For `'Mode','cumulative'`, `err` is a `numel(trees)`-by-`numel(tau)` numeric matrix. `err(j,k)` is the `tau(k)` cumulative, out-of-bag quantile regression error using the learners in `Mdl.Trees(trees(1:j))`.
- For `'Mode','ensemble'`, `err` is a 1-by-`numel(tau)` numeric vector. `err(k)` is the `tau(k)` cumulative, out-of-bag quantile regression error using the learners in `Mdl.Trees(trees)`.
- For `'Mode','individual'`, `err` is a `numel(trees)`-by-`numel(tau)` numeric matrix. `err(j,k)` is the `tau(k)` out-of-bag quantile regression error using the learner in `Mdl.Trees(trees(j))`.

Examples

Estimate Out-of-Bag Quantile Regression Error

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement, weight, and number of cylinders. Consider `Cylinders` a categorical variable.

```
load carsmall
Cylinders = categorical(Cylinders);
X = table(Displacement,Weight,Cylinders,MPG);
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners and save the out-of-bag indices.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,X,'MPG','Method','regression','OOBPrediction','on');
```

`Mdl` is a `TreeBagger` ensemble.

Perform quantile regression, and out-of-bag estimate the MAD of the entire ensemble using the predicted conditional medians.

```
oobErr = oobQuantileError(Mdl)
```

```
oobErr = 1.5349
```

`oobErr` is an unbiased estimate of the quantile regression error for the entire ensemble.

Find Appropriate Ensemble Size Using Out-of-Bag Quantile Regression Error

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement, weight, and number of cylinders.

```
load carsmall
X = table(Displacement,Weight,Cylinders,MPG);
```

Train an ensemble of bagged regression trees using the entire data set. Specify 250 weak learners and save the out-of-bag indices.

```
rng('default'); % For reproducibility
Mdl = TreeBagger(250,X,'MPG','Method','regression',...
    'OOBPrediction','on');
```

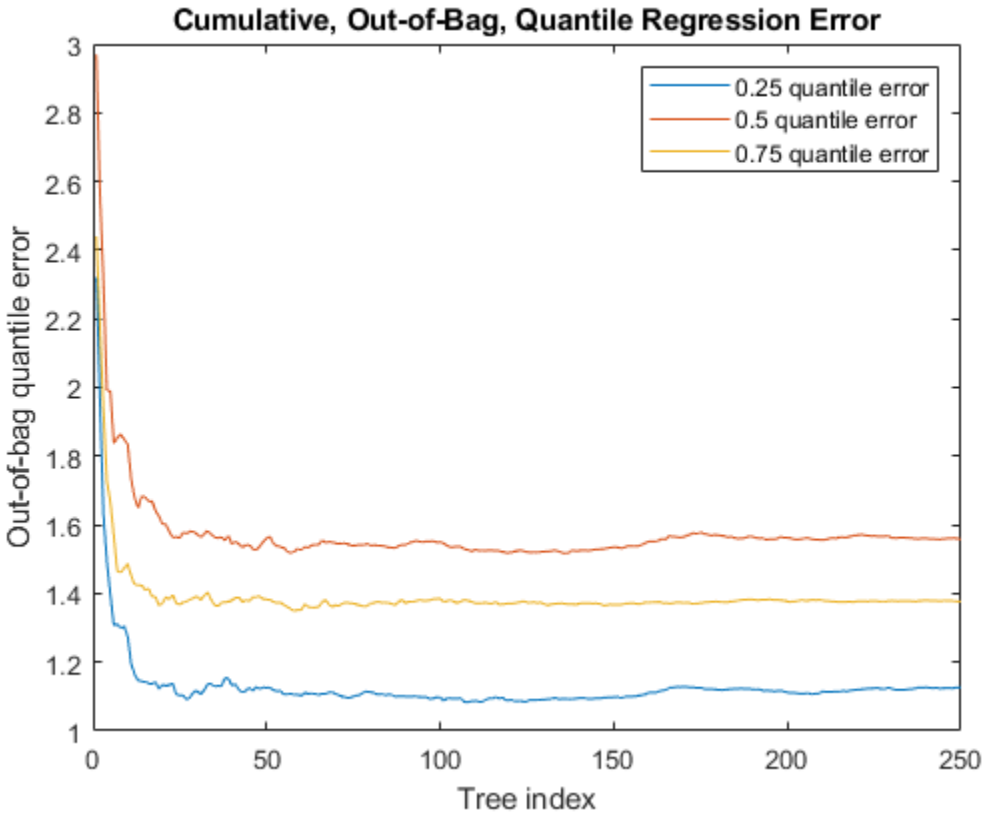
Estimate the cumulative; out-of-bag; 0.25, 0.5, and 0.75 quantile regression errors.

```
err = oobQuantileError(Mdl,'Quantile',[0.25 0.5 0.75],'Mode','cumulative');
```

`err` is an 250-by-3 matrix of cumulative, out-of-bag, quantile regression errors. Columns correspond to quantile probabilities and rows correspond to trees in the ensemble. The errors are cumulative, so they incorporate aggregated predictions from previous trees.

Plot the cumulative, out-of-bag, quantile errors on the same plot.

```
figure;
plot(err);
legend('0.25 quantile error','0.5 quantile error','0.75 quantile error');
ylabel('Out-of-bag quantile error');
xlabel('Tree index');
title('Cumulative, Out-of-Bag, Quantile Regression Error')
```



All quantile error curves appear to level off after training about 50 trees. So, training 50 trees appears to be sufficient to achieve minimal quantile error for the three quantile probabilities.

More About

Out-of-Bag

In a bagged ensemble, observations are out-of-bag when they are left out of the training sample for a particular learner. Observations are in-bag when they are used to train a particular learner.

When bagging learners, a practitioner takes a bootstrap sample (that is, a random sample with replacement) of size n for each learner, and then trains the learners using their respective bootstrap samples. Drawing n out of n observations with replacement omits on average about 37% of observations for each learner.

The out-of-bag ensemble error, the ensemble error estimated using out-of-bag observations only, is an unbiased estimator of the true ensemble error.

Quantile Regression Error

The quantile regression error of a model given observed predictor data and responses is the weighted mean absolute deviation (MAD). If the model under-predicts the response, then deviation weights are τ, the quantile probability. If the model over-predicts, then deviation weights are $1 - \tau$.

That is, the τ quantile regression error is

$$L_\tau = \tau \frac{\sum_{\{j: y_j \geq \widehat{y}_{\tau,j}\}} w_j \left(y_j - \widehat{y}_{\tau,j}\right)}{\sum_{j=1}^{n} w_j} + (1-\tau) \frac{\sum_{\{j: y_j < \widehat{y}_{\tau,j}\}} w_j \left(\widehat{y}_{\tau,j} - y_j\right)}{\sum_{j=1}^{n} w_j}.$$

y_j is true response j, $\widehat{y}_{\tau,j}$ is the τ quantile that the model predicts, and w_j is observation weight j.

Tip

The out-of-bag ensemble error estimator is unbiased for the true ensemble error. So, to tune parameters of a random forest, estimate the out-of-bag ensemble error instead of implementing cross-validation.

References

[1] Breiman, L. "Random Forests." *Machine Learning* 45, pp. 5–32, 2001.

[2] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

See Also

`TreeBagger` | `error` | `oobQuantilePredict` | `quantileError`

Topics

“Tune Random Forest Using Quantile Error and Bayesian Optimization” on page 18-145

Introduced in R2016b

oobQuantilePredict

Class: `TreeBagger`

Quantile predictions for out-of-bag observations from bag of regression trees

Syntax

```
YFit = oobQuantilePredict(Mdl)
YFit = oobQuantilePredict(Mdl,Name,Value)
[YFit,YW] = oobQuantilePredict( ___ )
```

Description

`YFit = oobQuantilePredict(Mdl)` returns a vector of medians of the predicted responses at all out-of-bag on page 33-4381 observations in `Mdl.X`, the predictor data, and using `Mdl`, which is a bag of regression trees. `Mdl` must be a `TreeBagger` model object and `Mdl.OOBIndices` must be nonempty.

`YFit = oobQuantilePredict(Mdl,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify quantile probabilities or trees to include for quantile estimation.

`[YFit,YW] = oobQuantilePredict( ___ )` also returns a sparse matrix of response weights on page 33-4382 using any of the previous syntaxes.

Input Arguments

`Mdl` — Bag of regression trees

`TreeBagger` model object (default)

Bag of regression trees, specified as a `TreeBagger` model object created by `TreeBagger`.

- The value of `Mdl.Method` must be `regression`.
- When you train `Mdl` using `TreeBagger`, you must specify the name-value pair `'OOBPrediction','on'`. Consequently, `TreeBagger` saves required out-of-bag observation index matrix in `Mdl.OOBIndices`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Quantile` — Quantile probability

`0.5` (default) | numeric vector containing values in [0,1]

Quantile probability, specified as the comma-separated pair consisting of `'Quantile'` and a numeric vector containing values in the interval [0,1]. For each observation (row) in `Mdl.X`, `oobQuantilePredict` estimates corresponding quantiles for all probabilities in `Quantile`.

Example: `'Quantile',[0 0.25 0.5 0.75 1]`

Data Types: `single` | `double`

`Trees` — Indices of trees to use in response estimation
`'all'` (default) | numeric vector of positive integers

Indices of trees to use in response estimation, specified as the comma-separated pair consisting of `'Trees'` and `'all'` or a numeric vector of positive integers. Indices correspond to the cells of `Mdl.Trees`; each cell therein contains a tree in the ensemble. The maximum value of `Trees` must be less than or equal to the number of trees in the ensemble (`Mdl.NumTrees`).

For `'all'`, `oobQuantilePredict` uses the indices `1:Mdl.NumTrees`.

Example: `'Trees',[1 10 Mdl.NumTrees]`

Data Types: `char` | `string` | `single` | `double`

`TreeWeights` — Weights to attribute to responses from individual trees
numeric vector of nonnegative values

Weights to attribute to responses from individual trees, specified as the comma-separated pair consisting of `'TreeWeights'` and a numeric vector of `numel(`*`trees`*`)` nonnegative values. *`trees`* is the value of the `Trees` name-value pair argument.

The default is `ones(size(`*`trees`*`))`.

Data Types: `single` | `double`

Output Arguments

`YFit` — Estimated quantiles
numeric matrix

Estimated quantiles for out-of-bag observations, returned as an *n*-by-`numel(`*`tau`*`)` numeric matrix. *n* is the number of observations in the training data (`numel(Mdl.Y)`) and *`tau`* is the value of the `Quantile` name-value pair argument. That is, `YFit(`*`j`*`,`*`k`*`)` is the estimated `100*`*`tau`*`(`*`k`*`)` percentile of the response distribution given `X(`*`j`*`,:)` and using `Mdl`.

`YW` — Response weights
sparse matrix

Response weights on page 33-4382, returned as an *n*-by-*n* sparse matrix. *n* is the number of responses in the training data (`numel(Mdl.Y)`). `YW(:,`*`j`*`)` specifies the response weights for the observation in `Mdl.X(`*`j`*`,:)`.

`oobQuantilePredict` predicts quantiles using linear interpolation of the empirical cumulative distribution function (cdf). For a particular observation, you can use its response weights to estimate quantiles using alternative methods, such as approximating the cdf using kernel smoothing on page B-78.

Examples

Predict Out-of-Bag Medians Using Quantile Regression

Load the `carsmall` data set. Consider a model that predicts the fuel economy (in MPG) of a car given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners and save out-of-bag indices.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression',...
    'OOBPrediction','on');
```

`Mdl` is a `TreeBagger` ensemble.

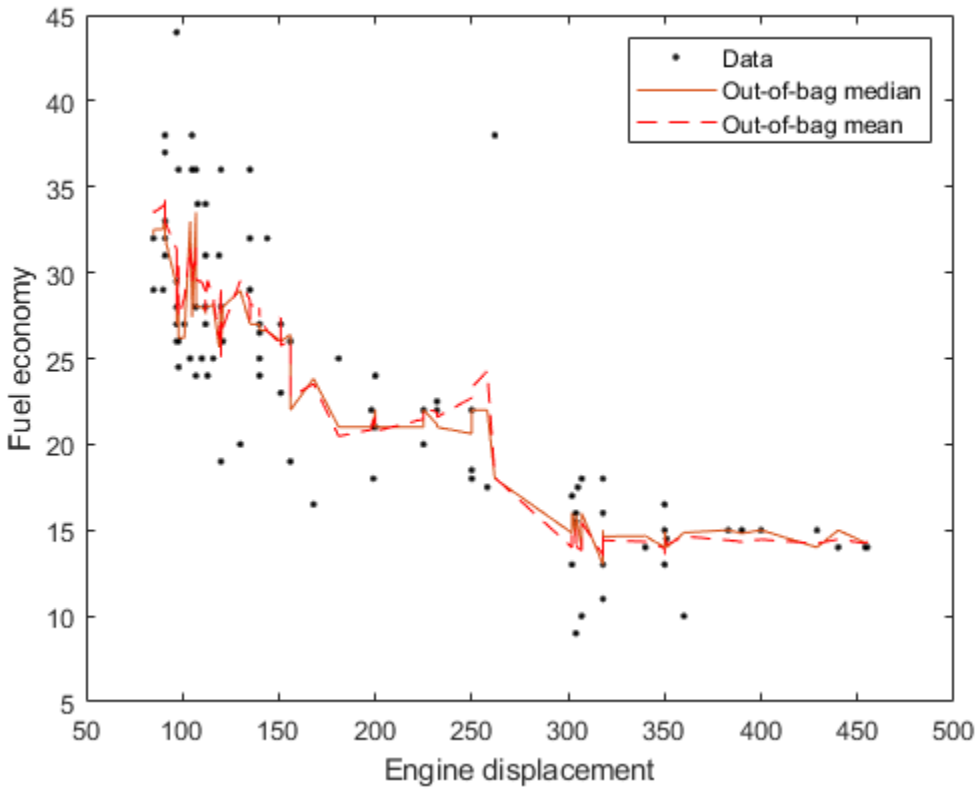
Perform quantile regression to predict the out-of-bag median fuel economy for all training observations.

```
oobMedianMPG = oobQuantilePredict(Mdl);
```

`oobMedianMPG` is an `n`-by-1 numeric vector of medians corresponding to the conditional distribution of the response given the sorted observations in `Mdl.X`. `n` is the number of observations, `size(Mdl.X,1)`.

Sort the observations in ascending order. Plot the observations and the estimated medians on the same figure. Compare the out-of-bag median and mean responses.

```
[sX,idx] = sort(Mdl.X);
oobMeanMPG = oobPredict(Mdl);

figure;
plot(Displacement,MPG,'k.');
hold on
plot(sX,oobMedianMPG(idx));
plot(sX,oobMeanMPG(idx),'r--');
ylabel('Fuel economy');
xlabel('Engine displacement');
legend('Data','Out-of-bag median','Out-of-bag mean');
hold off;
```



Estimate Out-of-Bag Prediction Intervals Using Percentiles

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car (in MPG) given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners and save out-of-bag indices.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression',...
    'OOBPrediction','on');
```

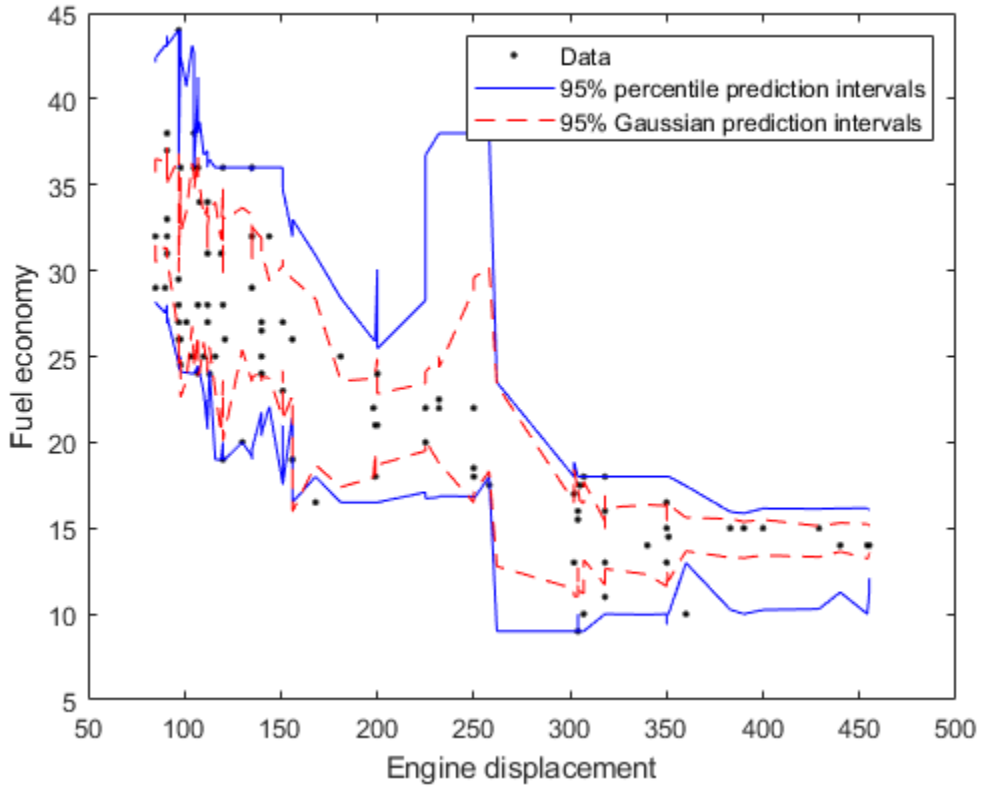
Perform quantile regression to predict the out-of-bag 2.5% and 97.5% percentiles.

```
oobQuantPredInts = oobQuantilePredict(Mdl,'Quantile',[0.025,0.975]);
```

`oobQuantPredInts` is an `n`-by-2 numeric matrix of prediction intervals corresponding to the out-of-bag observations in `Mdl.X`. `n` is number of observations, `size(Mdl.X,1)`. The first column contains the 2.5% percentiles and the second column contains the 97.5% percentiles.

Plot the observations and the estimated medians on the same figure. Compare the percentile prediction intervals and the 95% prediction intervals, assuming the conditional distribution of `MPG` is Gaussian.

```
[oobMeanMPG,oobSTEMeanMPG] = oobPredict(Mdl);
STDNPredInts = oobMeanMPG + [-1 1]*norminv(0.975).*oobSTEMeanMPG;
[sX,idx] = sort(Mdl.X);

figure;
h1 = plot(Displacement,MPG,'k.');
hold on
h2 = plot(sX,oobQuantPredInts(idx,:),'b');
h3 = plot(sX,STDNPredInts(idx,:),'r--');
ylabel('Fuel economy');
xlabel('Engine displacement');
legend([h1,h2(1),h3(1)],{'Data','95% percentile prediction intervals',...
    '95% Gaussian prediction intervals'});
hold off;
```



Estimate Out-of-Bag Conditional Cumulative Distribution Using Quantile Regression

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car (in MPG) given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners and save the out-of-bag indices.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression',...
    'OOBPrediction','on');
```

Estimate the out-of-bag response weights.

```
[~,YW] = oobQuantilePredict(Mdl);
```

`YW` is an n-by-n sparse matrix containing the response weights. `n` is the number of training observations, `numel(Y)`. The response weights for the observation in `Mdl.X(j,:)` are in `YW(:,j)`. Response weights are independent of any specified quantile probabilities.

Estimate the out-of-bag, conditional cumulative distribution function (ccdf) of the responses by:

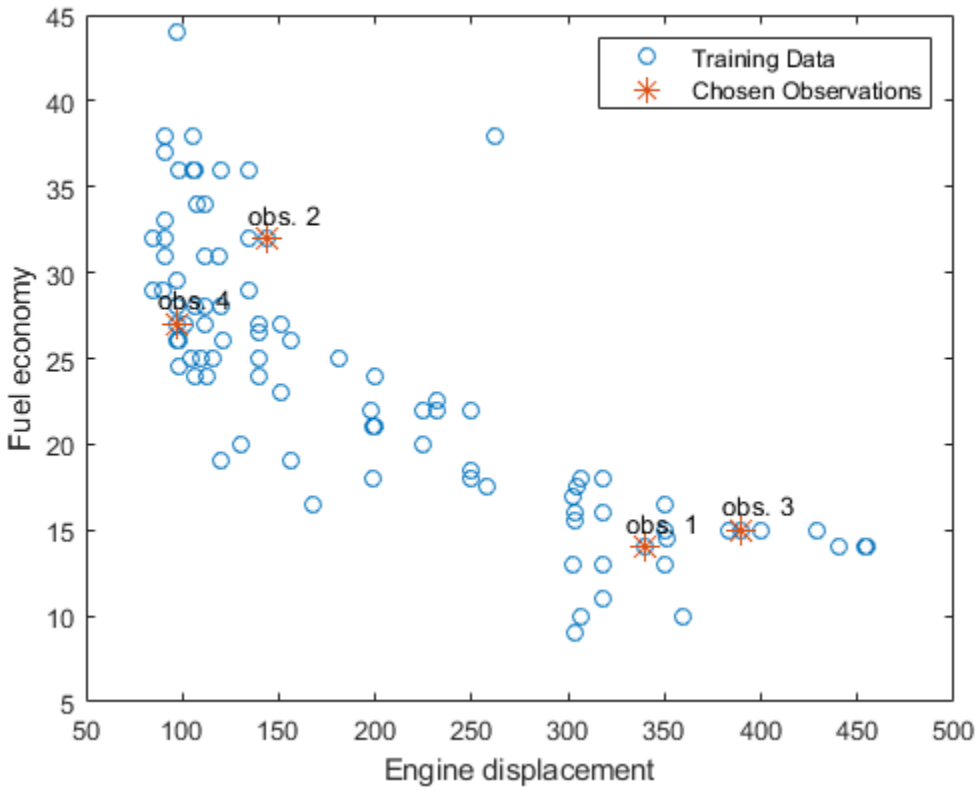
1. Sorting the responses is ascending order, and then sorting the response weights using the indices induced by sorting the responses.
2. Computing the cumulative sums over each column of the sorted response weights.

```
[sortY,sortIdx] = sort(Mdl.Y);
cpdf = full(YW(sortIdx,:));
ccdf = cumsum(cpdf);
```

`ccdf(:,j)` is the empirical out-of-bag ccdf of the response, given observation `j`.

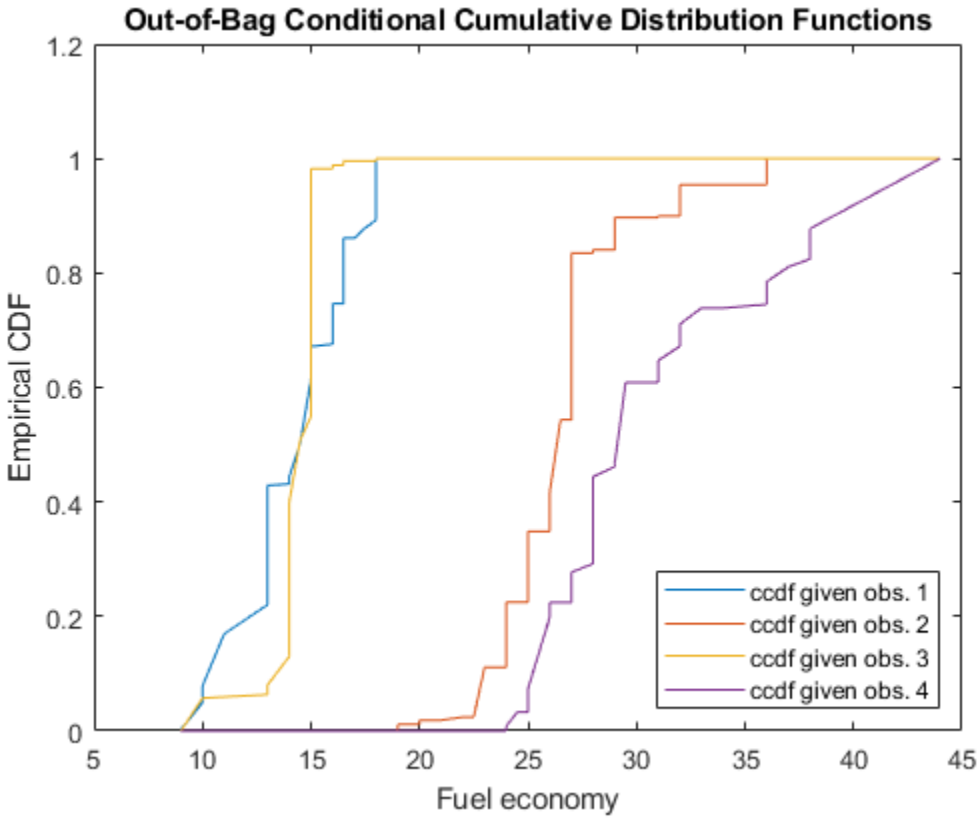
Choose a random sample of four training observations. Plot the training sample and identify the chosen observations.

```
[randX,idx] = datasample(Mdl.X,4);
figure;
plot(Mdl.X,Mdl.Y,'o');
hold on
plot(randX,Mdl.Y(idx),'*','MarkerSize',10);
text(randX-10,Mdl.Y(idx)+1.5,{'obs. 1' 'obs. 2' 'obs. 3' 'obs. 4'});
legend('Training Data','Chosen Observations');
xlabel('Engine displacement')
ylabel('Fuel economy')
hold off
```



Plot the out-of-bag ccdf for the four chosen responses in the same figure.

```
figure;
plot(sortY,ccdf(:,idx));
legend('ccdf given obs. 1','ccdf given obs. 2',...
    'ccdf given obs. 3','ccdf given obs. 4',...
    'Location','SouthEast')
title('Out-of-Bag Conditional Cumulative Distribution Functions')
xlabel('Fuel economy')
ylabel('Empirical CDF')
```



More About

Out-of-Bag

In a bagged ensemble, observations are out-of-bag when they are left out of the training sample for a particular learner. Observations are in-bag when they are used to train a particular learner.

When bagging learners, a practitioner takes a bootstrap sample (that is, a random sample with replacement) of size n for each learner, and then trains the learners using their respective bootstrap samples. Drawing n out of n observations with replacement omits on average about 37% of observations for each learner.

The out-of-bag ensemble error, the ensemble error estimated using out-of-bag observations only, is an unbiased estimator of the true ensemble error.

Quantile Random Forest

Quantile random forest [2] is a quantile-regression method that uses a random forest [1] of regression trees to model the conditional distribution of a response variable, given the value of predictor variables. You can use a fitted model to estimate quantiles in the conditional distribution of the response.

Besides quantile estimation, you can use quantile regression to estimate prediction intervals or detect outliers. For example:

- To estimate 95% quantile prediction intervals, estimate the 0.025 and 0.975 quantiles.
- To detect outliers, estimate the 0.01 and 0.99 quantiles. All observations smaller than the 0.01 quantile and larger than the 0.99 quantile are outliers. All observations that are outside the interval $[L,U]$ can be considered outliers:

  $$L = Q_1 - 1.5 * IQR$$

  and

  $$U = Q_3 + 1.5 * IQR,$$

  where:

  - Q_1 is the 0.25 quantile.
  - Q_3 is the 0.75 quantile.
  - $IQR = Q_3 - Q_1$ (the interquartile range).

Response Weights

Response weights are scalars that represent the conditional distribution of the response given a value in the predictor space. The observations in the bootstrap samples and the leaves that the training and test observations share induce response weights.

Given the observation x, the response weight for observation j in the training sample using tree t in the ensemble is

$$w_{tj}(x) = \frac{I\left\{X_j \in S_t(x)\right\}}{\sum_{k=1}^{n_{\text{train}}} I\{X_k \in S_t(x)\}},$$

where:

- $I\{h\}$ is the indicator function.
- $S_t(x)$ is the leaf of tree t containing x.
- n_{train} is the number of training observations.

In other words, the response weights of a particular tree form the conditional relative frequency distribution of the response.

The response weights for the entire ensemble are averaged over the trees:

$$w_j^*(x) = \frac{1}{T}\sum_{t=1}^{T} w_{tj}(x)\,.$$

Algorithms

`oobQuantilePredict` estimates out-of-bag quantiles by applying `quantilePredict` to all observations in the training data (`Mdl.X`). For each observation, the method uses only the trees for which the observation is out-of-bag.

For observations that are in-bag for all trees in the ensemble, `oobQuantilePredict` assigns the sample quantile of the response data. In other words, `oobQuantilePredict` does not use quantile

regression for out-of-bag observations. Instead, it assigns `quantile(Mdl.Y,tau)`, where *`tau`* is the value of the `Quantile` name-value pair argument.

References

[1] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

[2] Breiman, L. “Random Forests.” *Machine Learning*. Vol. 45, 2001, pp. 5–32.

See Also

`TreeBagger` | `oobQuantileError` | `predict` | `quantilePredict`

Topics

“Conditional Quantile Estimation Using Kernel Smoothing” on page 18-142

Introduced in R2016b

optimizableVariable

Variable description for `bayesopt` or other optimizers

Description

Create variables for optimizers.

Creation

Syntax

```
variable = optimizableVariable(Name,Range)
variable = optimizableVariable(Name,Range,Name,Value)
```

Description

`variable = optimizableVariable(Name,Range)` creates a variable with the specified name and range of values.

`variable = optimizableVariable(Name,Range,Name,Value)` sets properties on page 33-4384 using name-value pair arguments. For example, `optimizableVariable('xvar',[1 1000],'Type','integer')` creates an integer variable from 1 to 1000. You can specify multiple name-value pair arguments. Enclose each property name in quotes.

Properties

`Name` — Variable name
character vector | string scalar

Variable name, specified as a character vector or string scalar. The name must be unique, meaning different from those of other variables in the optimization.

Note

- There are two names associated with an `optimizableVariable`:
  - The MATLAB workspace variable name
  - The name of the variable in the optimization

  For example,

  ```
  xvar = optimizableVariable('spacevar',[1,100]);
  ```

  `xvar` is the MATLAB workspace variable, and `'spacevar'` is the variable in the optimization.

  Use these names as follows:

- Use `xvar` as an element in the vector of variables you pass to `bayesopt`. For example,

  ```
  results = bayesopt(fun,[xvar,tvar])
  ```

- Use `'spacevar'` as the name of the variable in the optimization. For example, in an objective function,

  ```
  function objective = mysvmfun(x,cdata,grp)
  SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
      'BoxConstraint',x.spacevar,...
      'KernelScale',x.tvar);
  objective = kfoldLoss(crossval(SVMModel));
  ```

Example: `'X1'`

Data Types: `char` | `string`

Range — Variable range
2-element increasing real vector | string array or cell array of names of categorical variables

Variable range, specified as a 2-element finite increasing real vector, or as a string array or cell array of names of categorical variables:

- For real or integer variables, `Range` gives the lower bound and upper bound of that variable.
- For categorical variables, `Range` gives the possible values.

Example: `[-10,1]`

Example: `{'red','blue','black'}`

Data Types: `double` | `string` | `cell`

Type — Variable type
`'real'` (default) | `'integer'` | `'categorical'`

Variable type, specified as `'real'` (real variable), `'integer'` (integer variable), or `'categorical'` (categorical variable).

Note The MATLAB data type of both `'real'` and `'integer'` variables is the standard double-precision floating point number. The data type of `'categorical'` variables is categorical. So, for example, to read a value of a categorical variable named `'colorv'` in a table of variables named `x`, use the command `char(x.colorv)`. For an example, see the objective function in “Custom Output Functions” on page 10-19.

Example: `'Type','categorical'`

`Transform` — Transform applied to variable
`'none'` (default) | `'log'`

Transform applied to variable, specified as `'none'` (no transform) or `'log'` (logarithmic transform).

For `'log'`, the variable must be `'real'` or `'integer'` and positive. The variable is searched and modeled on a log scale.

Example: `'Transform','log'`

`Optimize` — Indication to use variable in optimization
`true` (default) | `false`

Indication to use variable in optimization, specified as `true` (use the variable) or `false` (do not use the variable).

Example: `'Optimize',false`

Data Types: `logical`

Note You can use dot notation to change the following properties after creation.

- `Range` of real or integer variables. For example,

```
xvar = optimizableVariable('x',[-10,10]);
% Modify the range:
xvar.Range = [1,5];
```

- `Type` between `'integer'` and `'real'`. For example,

```
xvar.Type = 'integer';
```

- `Transform` of real or integer variables between `'log'` and `'none'`. For example,

```
xvar.Transform = 'log';
```

You can use this flexibility, for example, to tweak an optimization that you want to continue. Update the range or transform using dot notation and then call `resume`.

Object Functions

bayesopt Select optimal machine learning hyperparameters using Bayesian optimization

Examples

Variables for Optimization Examples

Real variable from 0 to 1:

```
var1 = optimizableVariable('xvar',[0 1])
```

```
var1 =
  optimizableVariable with properties:

         Name: 'xvar'
        Range: [0 1]
         Type: 'real'
    Transform: 'none'
     Optimize: 1
```

Integer variable from 1 to 1000 on a log scale:

```
var2 = optimizableVariable('ivar',[1 1000],'Type','integer','Transform','log')
```

```
var2 =
  optimizableVariable with properties:
```

```
        Name: 'ivar'
       Range: [1 1000]
        Type: 'integer'
   Transform: 'log'
    Optimize: 1
```

Categorical variable of rainbow colors:

```
var3 = optimizableVariable('rvar',{'r' 'o' 'y' 'g' 'b' 'i' 'v'},'Type','categorical')
```

```
var3 =
  optimizableVariable with properties:

        Name: 'rvar'
       Range: {'r'  'o'  'y'  'g'  'b'  'i'  'v'}
        Type: 'categorical'
   Transform: 'none'
    Optimize: 1
```

See Also

`BayesianOptimization` | `bayesopt`

Topics

“Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45
“Bayesian Optimization Using bayesopt” on page 10-26

Introduced in R2016b

ordinal

(Not Recommended) Arrays for ordinal data

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the “Categorical Arrays” data type instead.

Description

Ordinal data are discrete, nonnumeric values that have a natural ordering. `ordinal` array objects provide efficient storage and convenient manipulation of such data, while also maintaining meaningful labels for the values.

You can manipulate `ordinal` arrays like ordinary numeric arrays, by subscripting, concatenating, and reshaping. Use `ordinal` arrays as grouping variables when the elements indicate the group to which an observation belongs.

Creation

Syntax

```
B = ordinal(X)
B = ordinal(X,labels)
B = ordinal(X,labels,levels)

B = ordinal(X,labels,[],edges)
```

Description

`B = ordinal(X)` creates an ordinal array `B` from the array `X`. `ordinal` creates the levels of `B` from the sorted unique values in `X`, and creates default labels for the levels.

`B = ordinal(X,labels)` labels the levels in `B` according to `labels`.

`B = ordinal(X,labels,levels)` creates an ordinal array with possible levels defined by `levels`.

`B = ordinal(X,labels,[],edges)` creates an ordinal array by binning the numeric array `X` with bin edges given by the numeric vector `edges`.

Input Arguments

X — Input array

numeric | logical | character | string | categorical | cell array of character vectors

Input array to convert to `ordinal`, specified as a numeric, logical, character, string, or categorical array, or a cell array of character vectors. The levels of the resulting `ordinal` array correspond to the sorted unique values in `X`.

`labels` — Labels for discrete levels
character array | string array | cell array of character vectors

Labels for the discrete levels, specified as a character array, string array, or cell array of character vectors. By default, `ordinal` assigns labels to the levels in `B` in order of the sorted unique values in `X`.

You can include duplicate labels in `labels` to merge multiple values in `X` into a single level in `B`.

Data Types: `char` | `string` | `cell`

`levels` — Possible ordinal levels
vector

Possible ordinal levels for the output `ordinal` array, specified as a vector whose values can be compared to those in `X` using the equality operator. `ordinal` assigns labels to each level from the corresponding elements of `labels`. If `X` contains any values not present in `levels`, the levels of the corresponding elements of `B` are undefined.

`edges` — Bin edges
numeric vector

Bin edges used to create the ordinal array by binning a numeric array, specified as a numeric vector. The uppermost bin includes values equal to the rightmost edge. `ordinal` assigns labels to each level in the resulting ordinal array from the corresponding elements of `labels`. When you specify the `edges` input argument, it must have one more element than `labels`.

Output Arguments

`B` — Ordinal array
`ordinal` array object

Ordinal array, returned as an `ordinal` array object.

By default, an element of `B` is undefined if the corresponding element of `X` is `NaN` (when `X` is numeric), an empty character vector (when `X` is a character), an empty or missing string (when `X` is a string), or undefined (when `X` is categorical). `ordinal` treats such elements as undefined or missing and does not include entries for them among the possible levels. To create an explicit level for such elements instead of treating them as undefined, use the `levels` input argument and include `NaN`, the empty character vector, the empty or missing string, or an undefined element.

Object Functions

addlevels	(Not Recommended) Add levels to nominal or ordinal arrays
droplevels	(Not Recommended) Drop levels from a nominal or ordinal array
getlabels	(Not Recommended) Access nominal or ordinal array labels
getlevels	(Not Recommended) Access nominal or ordinal array levels
islevel	(Not Recommended) Determine if levels are in nominal or ordinal array
levelcounts	(Not Recommended) Element counts by level of a nominal or ordinal array
mergelevels	(Not Recommended) Merge levels of nominal or ordinal arrays
reorderlevels	(Not Recommended) Reorder levels of nominal or ordinal arrays
setlabels	(Not Recommended) Assign labels to levels of nominal or ordinal arrays

The following is a partial list of the many other MATLAB array functions you can use with ordinal arrays. For a complete list, see “Other MATLAB Functions Supporting Nominal and Ordinal Arrays” on page 2-2.

double	Double-precision arrays
histogram	Histogram plot
isequal	Determine array equality
isundefined	Find undefined elements in categorical array
pie	Pie chart
summary	Print summary of table, timetable, or categorical array
times	Multiplication

Examples

Create and Label Ordinal Arrays

Create an ordinal array from integer data, providing explicit labels.

```
quality = ordinal([1 2 3 3 2 1 2 1 3],...
   {'low' 'medium' 'high'})
```

```
quality = 1x9 ordinal
  Columns 1 through 7

     low      medium      high      high      medium      low      medium

  Columns 8 through 9

     low      high
```

Show that the first element is less than the second element (low is less than medium).

```
quality(1) < quality(2)
```

```
ans = logical
   1
```

Create an ordinal array by binning values between 0 and 1 into thirds with labels `'small'`, `'medium'`, and `'large'`.

```
X = rand(5,2)
```

```
X = 5×2

    0.8147    0.0975
    0.9058    0.2785
    0.1270    0.5469
    0.9134    0.9575
    0.6324    0.9649
```

```
A = ordinal(X,{'small' 'medium' 'large'},[],[0 1/3 2/3 1])
```

```
A = 5x2 ordinal
     large       small
```

```
     large       small
     small       medium
     large       large
     medium      large
```

Create and Manipulate Ordinal Arrays

Create an ordinal array from integer data.

```
quality = ordinal([1 2 3; 3 2 1; 2 1 3],{'low' 'medium' 'high'})
```

```
quality = 3x3 ordinal
     low         medium      high
     high        medium      low
     medium      low         high
```

Identify the elements in `quality` that are members of a level greater than or equal to `'medium'`. A value of `1` in the resulting array indicates that the corresponding element of `quality` is a member of this level.

```
quality >= 'medium'
```

```
ans = 3x3 logical array

   0   1   1
   1   1   0
   1   0   1
```

Identify the elements in `quality` that are members of either the level `'low'` or `'high'`.

```
ismember(quality,{'low' 'high'})
```

```
ans = 3x3 logical array

   1   0   1
   1   0   1
   0   1   1
```

Merge the elements of the `'medium'` and `'high'` levels into a new level labeled `'ok'`.

```
quality = mergelevels(quality,{'medium','high'},'ok')
```

```
quality = 3x3 ordinal
     low      ok       ok
     ok       ok       low
     ok       low      ok
```

Display the levels of `quality`.

```
getlevels(quality)
```

```
ans = 1x2 ordinal
     low      ok
```

Summarize the number of elements in each level. By default, `summary` returns counts for each column of the input array.

```
summary(quality)
```

```
     low      1      1      1
     ok       2      2      2
```

See Also

`nominal`

Topics

“Create Nominal and Ordinal Arrays” on page 2-3
“Reorder Category Levels” on page 2-9
“Categorize Numeric Data” on page 2-13
“Merge Category Levels” on page 2-16
“Sort Ordinal Arrays” on page 2-34
“Advantages of Using Nominal and Ordinal Arrays” on page 2-38
“Index and Search Using Nominal and Ordinal Arrays” on page 2-41
“Grouping Variables” on page 2-45

Introduced in R2007a

outlierMeasure

Class: `CompactTreeBagger`

Outlier measure for data

Syntax

```
out = outlierMeasure(B,X)
out = outlierMeasure(B,X,'param1',val1,'param2',val2,...)
```

Description

`out = outlierMeasure(B,X)` computes outlier measures for predictors `X` using trees in the ensemble `B`. The method computes the outlier measure for a given observation by taking an inverse of the average squared proximity between this observation and other observations. `outlierMeasure` then normalizes these outlier measures by subtracting the median of their distribution, taking the absolute value of this difference, and dividing by the median absolute deviation. A high value of the outlier measure indicates that this observation is an outlier.

You can supply the proximity matrix directly by using the `'Data'` parameter.

`out = outlierMeasure(B,X,'param1',val1,'param2',val2,...)` specifies optional parameter name/value pairs:

`'Data'`	Flag indicating how to treat the `X` input argument. If set to `'predictors'` (default), the method assumes `X` is a matrix of predictors and uses it for computation of the proximity matrix. If set to `'proximity'`, the method treats `X` as a proximity matrix returned by the `proximity` method. If you do not supply the proximity matrix, `outlierMeasure` computes it internally. If you use the `proximity` method to compute a proximity matrix, supplying it as input to `outlierMeasure` reduces computing time.
`'Labels'`	Vector of true class labels. True class labels can be a numeric vector, character matrix, string array, or cell array of character vectors. When you supply this parameter, the method performs the outlier calculation for any observations using only other observations from the same class. This parameter must specify one label for each observation (row) in `X`.

See Also

`proximity`

parallelcoords

Parallel coordinates plot

Syntax

```
parallelcoords(x)
parallelcoords(x,Name,Value)

parallelcoords(ax, ___ )

h = parallelcoords( ___ )
```

Description

`parallelcoords(x)` creates a parallel coordinates plot of the multivariate data in the matrix `x`. Use a parallel coordinates plot to visualize high dimensional data, where each observation is represented by the sequence of its coordinate values plotted against their coordinate indices.

`parallelcoords(x,Name,Value)` creates a parallel coordinates plot with additional options specified by one or more `Name,Value` pair arguments. For example, you can standardize the data in `x` or label the coordinate tick marks along the horizontal axis of the plot.

`parallelcoords(ax, ___ )` creates a parallel coordinates plot using the axes specified by the axes graphic object `ax`, using any of the previous syntaxes.

`h = parallelcoords( ___ )` returns a column vector of handles to the `Line` objects created by `parallelcoords`, with one handle for each row of `x`.

Examples

Parallel Coordinates Plot for Grouped Data

Load the Fisher iris sample data.

```
load fisheriris
```

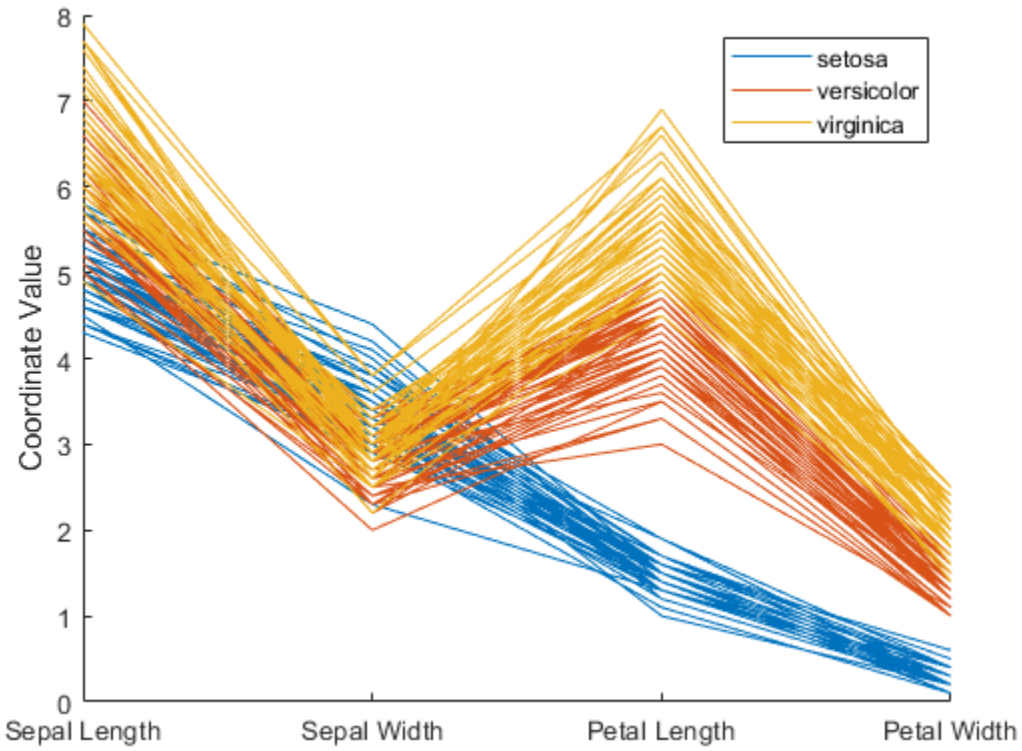
The data contains four measurements (sepal length, sepal width, petal length, and petal width) from three species of iris flowers. The matrix `meas` contains all four measurements for each of 150 flowers. The cell array `species` contains the species name for each of the 150 flowers.

Create a cell array that contains the name of each measurement variable in the sample data.

```
labels = {'Sepal Length','Sepal Width','Petal Length','Petal Width'};
```

Create a parallel coordinate plot using the measurement data in `meas`. Use a different color for each group as identified in `species`, and label the horizontal axis using the variable names.

```
parallelcoords(meas,'Group',species,'Labels',labels)
```



The resulting plot contains one line for each observation (flower). The color of each line indicates the flower species.

Parallel Coordinates Plot with Quantile Values

Load the Fisher iris sample data.

```
load fisheriris
```

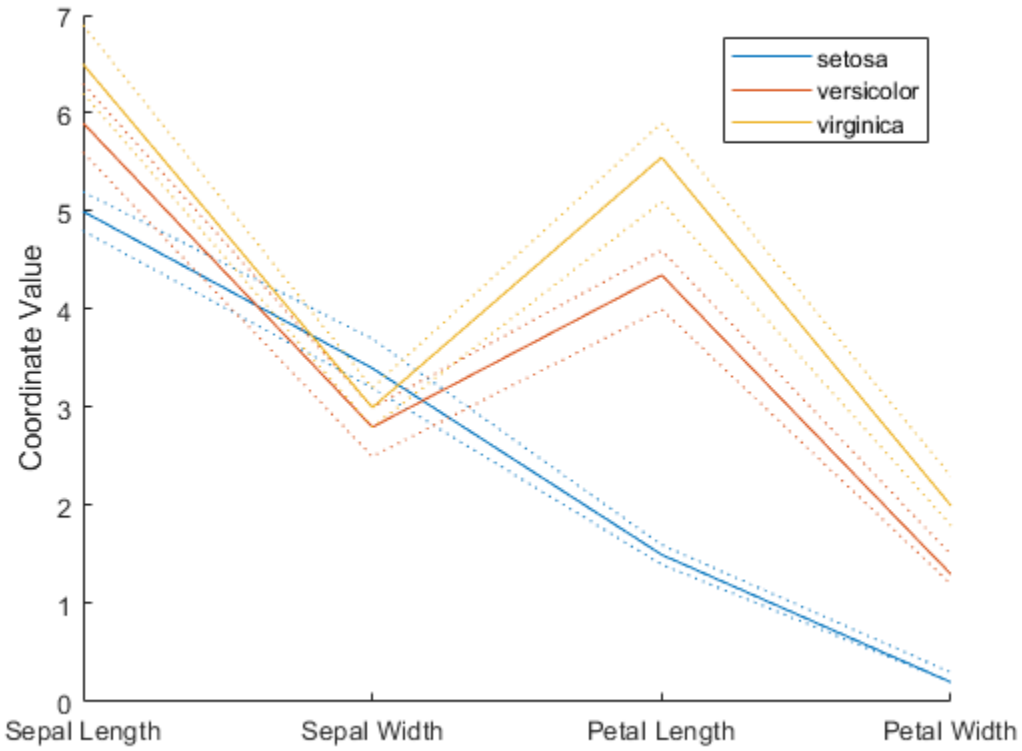
The data contains four measurements (sepal length, sepal width, petal length, and petal width) from three species of iris flowers. The matrix `meas` contains all four measurements for each of 150 flowers. The cell array `species` contains the species name for each of the 150 flowers.

Create a cell array that contains the name of each measurement variable in the sample data.

```
labels = {'Sepal Length','Sepal Width','Petal Length','Petal Width'};
```

Create a parallel coordinates plot using the measurement data in `meas`. Plot only the median, 25 percent, and 75 percent quartile values for each group identified in `species`. Label the horizontal axis using the variable names.

```
parallelcoords(meas,'group',species,'labels',labels,...
               'quantile',.25)
```



The plot shows the median values for each group as a solid line and the quartile values as dotted lines of the same color. For example, the solid blue line shows the median value measured for each variable on `setosa` irises. The dotted blue line below the solid blue line shows the 25th percentile of measurements for each variable on `setosa` irises. The dotted blue line above the solid blue line shows the 75th percentile of measurements for each variable on `setosa` irises.

Adjust Line Properties in Parallel Coordinates Plot

Load the Fisher iris sample data.

```
load fisheriris
```

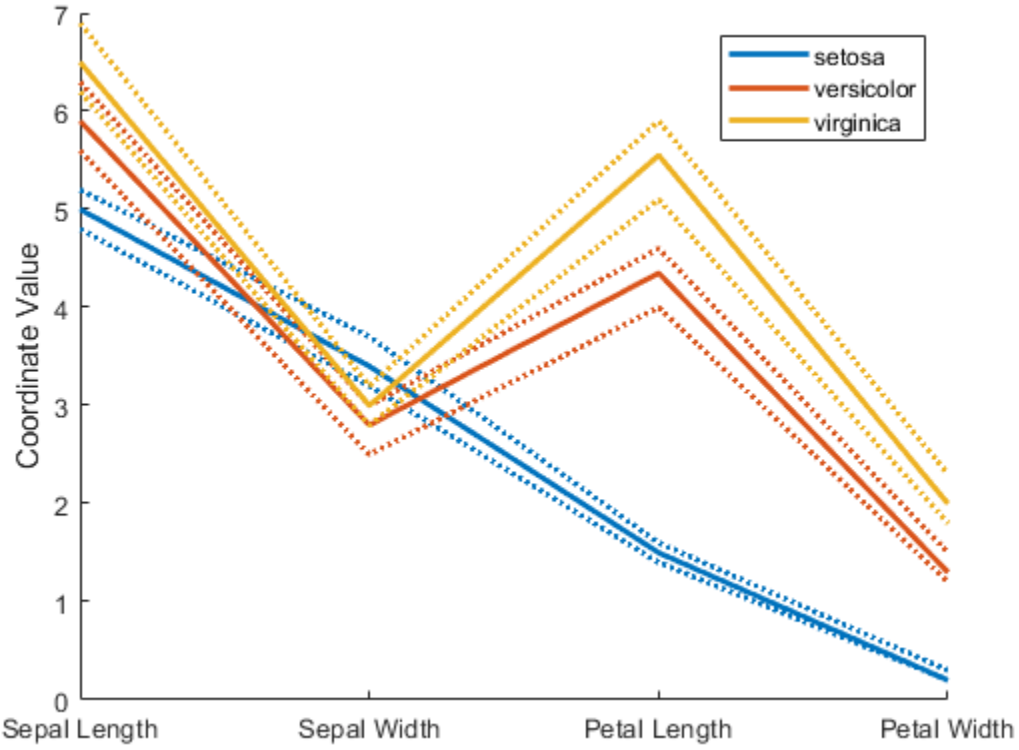
The data contains four measurements (sepal length, sepal width, petal length, and petal width) from three species of iris flowers. The matrix `meas` contains all four measurements for each of 150 flowers. The cell array `species` contains the species name for each of the 150 flowers.

Create a cell array that contains the name of each measurement variable in the sample data.

```
labels = {'Sepal Length','Sepal Width','Petal Length','Petal Width'};
```

Create a parallel coordinates plot using the measurement data in `meas`. Plot only the median, 25 percent, and 75 percent quartile values for each group identified in `species`. Label the horizontal axis using the variable names. Set the line width to 2.

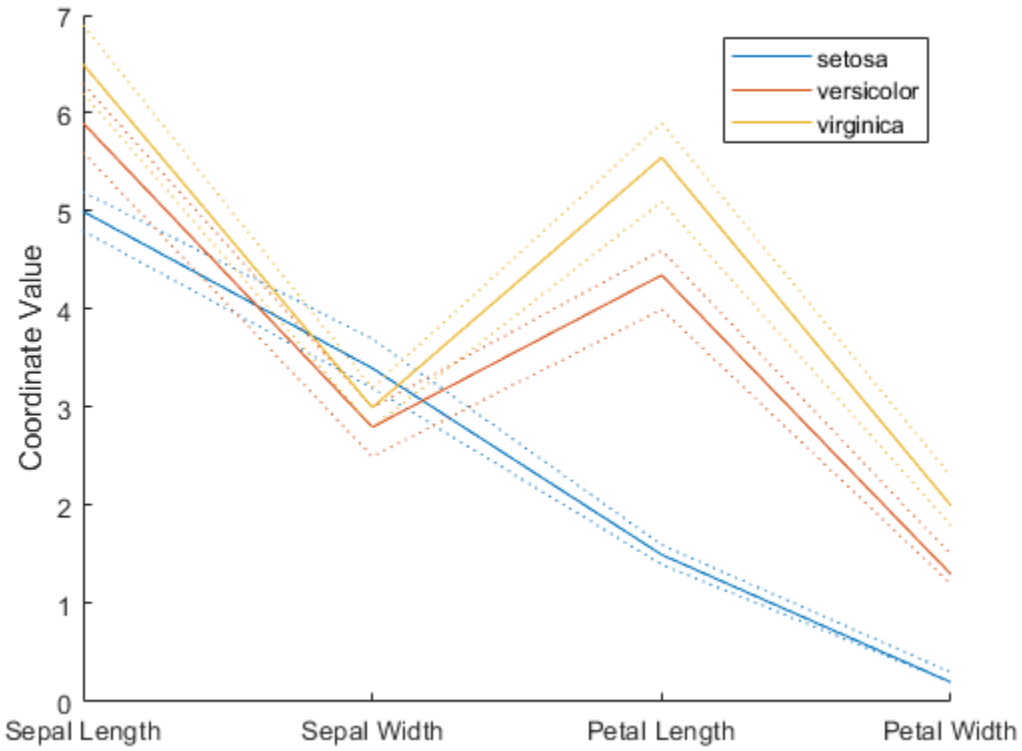
```
parallelcoords(meas,'group',species,'labels',labels,...
               'quantile',.25,'LineWidth',2)
```



Specifying `'LineWidth'` in this way sets the width of every line in the plot to 2.

Recreate the parallel coordinates plot, but this time, use handles to increase the width of only the line representing the median value for each measurement made on irises in the `setosa` group.

```
h = parallelcoords(meas,'group',species,'labels',labels,...
               'quantile',.25)
```

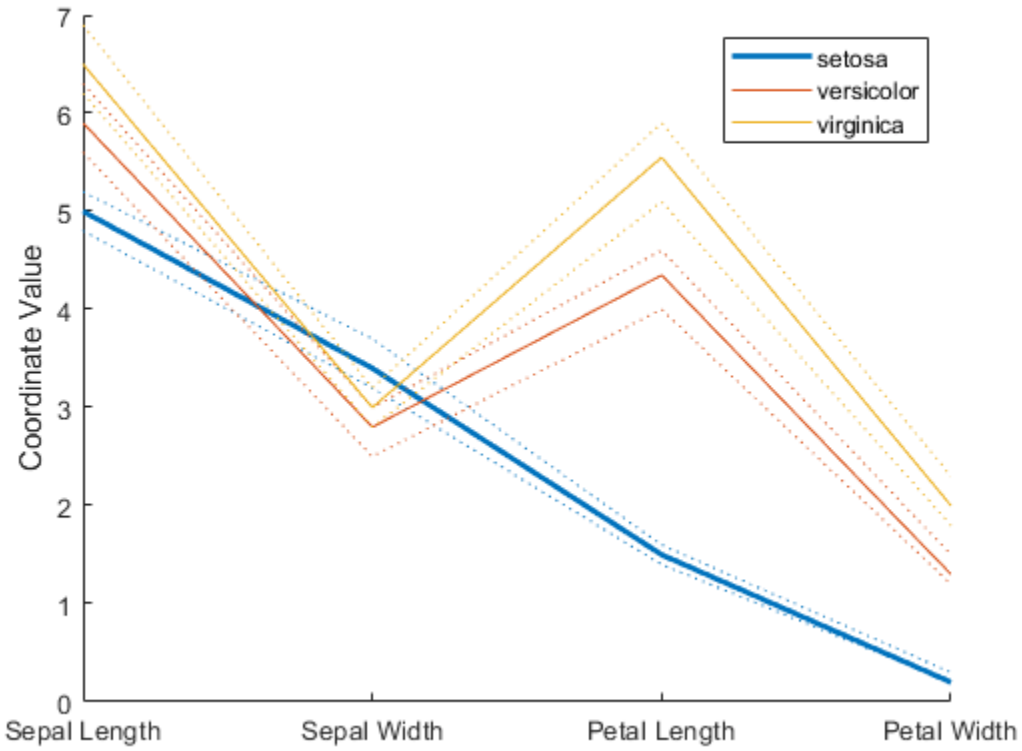



```
h =
  9x1 Line array:

  Line    (median)
  Line    (lower quantile)
  Line    (upper quantile)
  Line    (median)
  Line    (lower quantile)
  Line    (upper quantile)
  Line    (median)
  Line    (lower quantile)
  Line    (upper quantile)
```

The returned column vector `h` contains handles that correspond to each line object created by `parallelcoords`. For example, h(1) corresponds to the median line for the first grouping variable (`setosa`).

Use dot notation to increase the width of the line showing the median value for each measurement made on irises in the `setosa` group.

```
h(1).LineWidth = 2;
```



Input Arguments

x — Multivariate input data

numeric matrix

Multivariate input data, specified as an *n*-by-*p* matrix of numeric values. *n* is the number of rows of `x`, and each row corresponds to an observation in `x`. *p* is the number of columns in `x`, and each column corresponds to a variable in `x`.

`parallelcoords` treats `NaN` values in `x` as missing values and does not plot those coordinate values.

Data Types: `single` | `double`

ax — Axes for plot

axes graphic object

Axes for plot, specified as an axes graphic object. If you do not specify `ax`, then `parallelcoords` creates the plot using the current axis. For more information on creating an axes graphic object, see `axes` and Axes.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Group',species,'Quantile',.25` plots the median, 25 percent, and 75 percent quartile values for the input data, using a different color for each group identified in the variable `species`.

`Group` — Grouping variable for input data
numeric array | categorical variable | character matrix | string array | cell array

Grouping variable for input data, specified as the comma-separated pair consisting of `'Group'` and a numeric array containing a group index for each observation. Alternatively, the array can be a categorical variable, character matrix, string array, or cell array containing a group name for each observation.

Data Types: `single` | `double` | `categorical` | `char` | `string` | `cell`

`Labels` — Horizontal axis labels
character array | string array | cell array

Horizontal axis labels, specified as the comma-separated pair consisting of `'Labels'` and a character array, string array, or cell array containing the label names.

Example: `'Labels',{'Sepal Width','Sepal Length'}`

Data Types: `char` | `string` | `cell`

`Quantile` — Quantiles of input data to plot
numeric value in the range (0,1)

Quantiles of input data to plot, specified as the comma-separated pair consisting of `'Quantile'` and a numeric value in the range (0,1). If you specify a value *alpha* for `'Quantile'`, then `parallelcoords` plots only the median, *alpha*, and 1 – *alpha* quantiles for each of the variables (columns) in `x`.

The quantile plot option provides a useful summary of the data when `x` contains many observations.

Example: `'Quantile',.25`

Data Types: `single` | `double`

`Standardize` — Method to standardize input data
`'on'` | `'PCA'` | `'PCAStd'`

Method to standardize input data, specified as the comma-separated pair consisting of `'Standardize'` and one of the following.

`'on'`	Scale each column of `x` to have a mean equal to 0 and a standard deviation equal to 1 before plotting.
`'PCA'`	Create plot from the principal component scores of `x`, in order of decreasing eigenvalues. `parallelcoords` removes rows of `x` containing missing values (`NaN`) for PCA standardization.
`'PCAStd'`	Create plot using the standardized principal component scores.

Example: `'Standardize','on'`

Tips

- You can modify certain aspects of the plot lines by specifying a property name and value for any of the properties listed in Primitive Line. However, this approach applies the modification to all the

lines in the plot. To modify only certain plot lines, use the syntax that returns graphics handles and use dot notation to adjust each line property individually. For an illustration, see “Adjust Line Properties in Parallel Coordinates Plot” on page 33-4396.

Output Arguments

h — Graphic handles for line objects

vector of `Line` graphic handles

Graphic handles for line objects, returned as a vector of `Line` graphic handles. Graphic handles are unique identifiers that you can use to query and modify the properties of a specific line on the plot. To view and set properties of line objects, use dot notation. For information on using dot notation, see“Access Property Values”. For information on the `Line` properties that you can set, see Primitive Line.

If you use the `'Quantile'` name-value pair argument, then `h` contains one handle for each of the three lines objects created. If you use both the `'Quantile'` and the `'Group'` name-value pair arguments, then `h` contains three handles for each group.

Alternative Functionality

Alternatively, you can create a `ParallelCoordinatesPlot` object by using the `parallelplot` function.

- Unlike the `parallelcoords` function, `parallelplot` allows you to plot tabular data that includes categorical variables.
- `parallelplot` does not support the plotting of quantiles for numeric data. However, the `ParallelCoordinatesPlot` object contains the `DataNormalization` property, which provides several data normalization methods for coordinates with numeric values.

To control the appearance and behavior of the object, change the ParallelCoordinatesPlot.

See Also

`andrewsplot` | `glyphplot` | `parallelplot`

Topics

“Grouping Variables” on page 2-45
“Access Property Values”

Introduced before R2006a

paramci

Package: prob

Confidence intervals for probability distribution parameters

Syntax

```
ci = paramci(pd)
ci = paramci(pd,Name,Value)
```

Description

`ci = paramci(pd)` returns the array `ci` containing the lower and upper boundaries of the 95% confidence interval for each parameter in probability distribution `pd`.

`ci = paramci(pd,Name,Value)` returns confidence intervals with additional options specified by one or more name-value pair arguments. For example, you can specify a different percentage for the confidence interval, or compute confidence intervals only for selected parameters.

Examples

Parameter Confidence Intervals

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Fit a normal distribution object to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

You can also obtain these intervals by using the function `paramci`.

```
ci = paramci(pd)
```

```
ci = 2×2

   73.4321    7.7391
   76.5846    9.9884
```

Column 1 of `ci` contains the lower and upper 95% confidence interval boundaries for the mu parameter, and column 2 contains the boundaries for the sigma parameter.

Change Parameter Confidence Intervals

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Fit a normal distribution object to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the 99% confidence interval for the distribution parameters.

```
ci = paramci(pd,'Alpha',.01)
```

```
ci = 2×2

   72.9245    7.4627
   77.0922   10.4403
```

Column 1 of `ci` contains the lower and upper 99% confidence interval boundaries for the mu parameter, and column 2 contains the boundaries for the sigma parameter.

Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01` specifies a 99% confidence interval.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level for the confidence interval, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1). The confidence level of `ci` is `100(1–Alpha)`%. The default value `0.05` corresponds to a 95% confidence interval.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Parameter` — Parameter list

character vector | string array | cell array of character vectors

Parameter list for which to compute confidence intervals, specified as the comma-separated pair consisting of `'Parameter'` and a character vector, string array, or cell array of character vectors containing the parameter names. By default, `paramci` computes confidence intervals for all distribution parameters.

Example: `'Parameter','mu'`

Data Types: `char` | `string` | `cell`

Type — Computation method

`'exact'` | `'Wald'` | `'lr'`

Computation method for the confidence intervals, specified as the comma-separated pair consisting of `'Type'` and `'exact'`, `'Wald'`, or `'lr'`.

`'exact'` computes the confidence intervals using an exact method, and is available for the following distributions.

Distribution	Computation Method
Binomial	Compute using the Clopper-Pearson method based on exact probability calculations. This method does not provide exact coverage probabilities.
Exponential	Compute using a method based on a chi-square distribution. This method provides exact coverage for complete and Type 2 censored samples.
Normal	Computation method based on *t* and chi-square distributions for uncensored samples provides exact coverage for uncensored samples. For censored samples, `paramci` uses the Wald method if `Type` is `exact`.

Distribution	Computation Method
Lognormal	Computation method based on *t* and chi-square distributions for uncensored samples provides exact coverage. For censored samples, `paramci` uses the Wald method if `Type` is `exact`.
Poisson	Computation method based on a chi-square distribution provides exact coverage. For large degrees of freedom, the chi-square is approximated by a normal distribution for numerical efficiency.
Rayleigh	Computation method based on a chi-square distribution provides exact coverage probabilities.

`'exact'` is the default when it is available. Alternatively, you can specify `'Wald'` to compute the confidence intervals using the Wald method, or `'lr'` to compute the confidence intervals using the likelihood ratio method.

Example: `'Type','Wald'`

`LogFlag` — Boolean flag for log scale

vector

Boolean flag for the log scale, specified as the comma-separated pair consisting of `'LogFlag'` and a vector containing Boolean values corresponding to each distribution parameter. The flag specifies which Wald intervals to compute on a log scale. The default values depend on the distribution.

Example: `'LogFlag',[0,1]`

Data Types: `logical`

Output Arguments

`ci` — Confidence interval

array

Confidence interval, returned as a *p*-by-2 array containing the lower and upper bounds of the `100(1–Alpha)`% confidence interval for each distribution parameter. *p* is the number of distribution parameters.

If you create `pd` by using `makedist` and specifying the distribution parameters, the lower and upper bounds are equal to the specified parameters.

See Also

Distribution Fitter | `fitdist` | `makedist` | `mle` | `negloglik` | `proflik`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

paretotails

Piecewise distribution with Pareto tails

Description

A `paretotails` object is a piecewise distribution with generalized Pareto distributions (GPDs) in the tails.

A `paretotails` object consists of one or two GPDs in the tails and another distribution in the center. You can specify the distribution type for the center by using the `cdffun` argument of `paretotails` when you create an object. Valid values are `'ecdf'`, `'kernel'`, and a function handle.

`paretotails` fits a distribution of type `cdffun` to the observations (`x`) and finds the quantiles corresponding to the lower and upper tail cumulative probabilities (`pl` and `pu`, respectively). Then, `paretotails` fits two GPDs to the lower `100*pl` percent of the observations and the upper `100*(1–pu)` percent of the observations, respectively. If `x` does not have at least two distinct observations in a tail, then `paretotails` does not create the corresponding tail segment.

Use the object functions `boundary`, `segment`, `upperparams`, and `lowerparams` to find distribution characteristics. `lowerparams` and `upperparams` return the parameters of the GPDs in the tails. `boundary` returns the boundary points between piecewise distribution segments, `segment` returns the segment of a piecewise distribution containing input values, and `nsegments` returns the number of segments in an object.

Use the object functions `cdf`, `icdf`, `pdf`, and `random` to evaluate the distribution. These functions are well suited to copula and other Monte Carlo simulations. `pdf` returns the GPD density in the tails and the slope of the cumulative distribution function (cdf) in the center. These probability density function (pdf) values in the center are generally not good estimates of the underlying density of the original data.

Creation

Create a piecewise distribution object using `paretotails`.

Syntax

```
pd = paretotails(x,pl,pu)
pd = paretotails(x,pl,pu,cdffun)
```

Description

`pd = paretotails(x,pl,pu)` returns the piecewise distribution object `pd`, which consists of the empirical distribution in the center and generalized Pareto distributions in the tails. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities `pl` and `pu`, respectively.

`pd = paretotails(x,pl,pu,cdffun)` specifies the type of center distribution segment using `cdffun`.

Input Arguments

x — Input data
numeric vector

Input data, specified as a numeric vector.

Data Types: `double`

pl — Lower tail cumulative probability
numeric scalar in the range `[0,1]`

Lower tail cumulative probability, specified as a numeric scalar in the range `[0,1]`. The quantile of `pl` is the boundary of the lower tail observations.

If `pl` is `0` or `x` does not have at least two distinct observations in the lower tail, then `paretotails` divides the input data in `x` into two groups, center and upper tail. In this case, the fitted piecewise distribution object `pd` consists of two segments: the empirical distribution in the center and GPD in the upper tail.

Example: `0.1`

Data Types: `single` | `double`

pu — Upper tail cumulative probability
numeric scalar in the range `[0,1]`

Upper tail cumulative probability, specified as a numeric scalar in the range `[0,1]`. The quantile of `pu` is the boundary of the upper tail observations.

If `pu` is `1` or `x` does not have at least two distinct observations in the upper tail, then `paretotails` divides the input data in `x` into two groups, center and lower tail. In this case, the fitted piecewise distribution object `pd` consists of two segments: the empirical distribution in the center and GPD in the lower tail.

Example: `0.9`

Data Types: `single` | `double`

cdffun — Type of center distribution segment
`'ecdf'` (default) | `'kernel'` | function handle

Type of center distribution segment, specified as `'ecdf'`, `'kernel'`, or a function handle.

Value	Description
`'ecdf'`	Interpolated empirical cdf. `paretotails` uses values in `x` as the midpoints in the vertical steps of the empirical cdf, and computes the estimates for the points between the values in `x` by linear interpolation. For details about how to find the interpolated empirical cdf, see A Piecewise Linear Nonparametric CDF Estimate on page 5-181.

Value	Description
`'kernel'`	Interpolated kernel smoothing estimate of the cdf. `paretotails` uses the `ksdensity` function to find cdf estimates for 100 points in the range of `x`, and uses linear interpolation to compute the estimates for the points between the 100 points. `'kernel'` is equivalent to specifying a function handle `fun = @(x)ksdensity(x,'function','cdf');`.
function handle	Interpolated estimates using a specified function. `paretotails` uses a handle to a function of the form `[p,xi] = fun(x)` that accepts the input data vector `x` and returns a vector `p` of cdf values and a vector `xi` of evaluation points. Values in `xi` must be sorted and distinct but do not have to equal the values in `x`. The `paretotails` function computes the cdf estimates for the points between the values in `xi` by linear interpolation.

`paretotails` uses `cdffun` to compute the quantiles corresponding to `pl` and `pu`.

Example: `'kernel'`

Properties

`NumSegments` — Number of segments

3 | 2 | 1

This property is read-only.

Number of segments, including the center segment and tail segments in a `paretotail` object, specified as a scalar. `NumSegments` is 3, 2, or 1 if the number of the tail segments in the object is 2, 1, or 0, respectively.

Data Types: `double`

`LowerParameters` — Lower tail GPD parameters

numeric vector

This property is read-only.

Lower tail GPD parameters, fit to the lower extreme observations in `x`, specified as a numeric vector. The first value is the shape parameter and the second value is the scale parameter of the GPD.

The location parameter of the lower tail GPD is equal to the quantile of `pl`. Use the `boundary` function to return the location parameter. For example, run `[p,q] = boundary(pd)`, where `pd` is a `paretotails` object. `q(1)` is the location parameter.

Data Types: `single` | `double`

`UpperParameters` — Upper tail GPD parameters

numeric vector

This property is read-only.

Upper tail GPD parameters, fit to the upper extreme observations in `x`, specified as a numeric vector. The first value is the shape parameter and the second value is the scale parameter of the GPD.

The location parameter of the upper tail GPD is equal to the quantile of `pu`. Use the `boundary` function to return the location parameter. For example, run `[p,q] = boundary(pd)`, where `pd` is a `paretotails` object. `q(2)` is the location parameter.

Data Types: `single` | `double`

Object Functions

boundary	Piecewise distribution boundaries
cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
lowerparams	Lower Pareto tail parameters
nsegments	Number of segments in piecewise distribution
pdf	Probability density function
random	Random numbers
segment	Piecewise distribution segments containing input values
upperparams	Upper Pareto tail parameters

Examples

Create `paretotails` with Empirical Distribution

Generate a sample data set and fit a piecewise distribution with Pareto tails to the data. Specify an empirical distribution for the center by using `paretotails` with its default settings.

Generate a sample data set containing 100 random numbers from a *t* distribution with 3 degrees of freedom.

```
rng('default');  % For reproducibility
t = trnd(3,100,1);
```

Create a `paretotails` object by fitting a piecewise distribution to `t`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and GPDs for the lower and upper 10% of the data set.

```
pd = paretotails(t,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.84875    (0 < p < 0.1): lower tail, GPD(0.183032,1.00347)
   -1.84875 < x < 2.07662  (0.1 < p < 0.9): interpolated empirical cdf
       2.07662 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.333239,1.19705)
```

Each line of the object display shows the summary of each segment, including the GPD parameters (shape and scale parameters) and the boundary values in the quantiles and cumulative probabilities. Use the object functions `boundary`, `lowerparams`, and `upperparams` to return these values.

You can use the `nsegments` function to return the number of segments and the `segment` function to return the segment that contains input values.

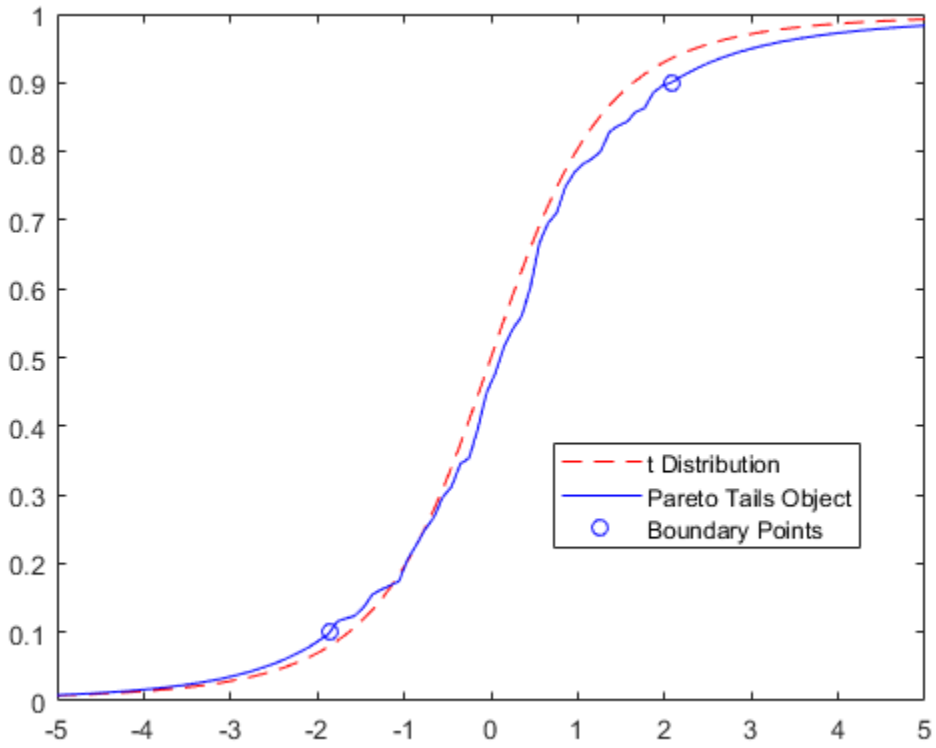
You can also use the distribution functions `cdf`, `icdf`, `pdf`, and `random` to evaluate the distribution and generate random samples.

Plot the cdf of the *t* distribution and the cdf of the `paretotails` object on the same figure.

```
x = linspace(-5,5);
plot(x,tcdf(x,3),'r--')
hold on
plot(x,cdf(pd,x),'b-')
```

Find the boundary points between the segments of the `paretotails` object by using `boundary`, and mark the points on the figure.

```
[p,q] = boundary(pd);
plot(q,p,'bo')
legend('t Distribution','Pareto Tails Object','Boundary Points','Location','best')
hold off
```



Create `paretotails` with Function Handle

Generate a sample data set and fit a piecewise distribution with Pareto tails to the data. Fit a center segment by using `paretotails` with a function handle.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Define a function handle using `ksdensity` to specify a nondefault value of the bandwidth.

```
myfun1 = @(x)ksdensity(x,'Bandwidth',.1,'Function','cdf');
```

Create a `paretotails` object by fitting a piecewise distribution with the specified kernel smoothing estimator to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the kernel estimator for the middle 80% of the data set and GPDs for the lower and upper 10% of the data set.

```
pd1 = paretotails(x,0.1,0.9,myfun1)
```

```
pd1 =
Piecewise distribution with 3 segments
      -Inf < x < -1.35204    (0 < p < 0.1): lower tail, GPD(0.0104112,0.54947)
   -1.35204 < x < 1.80824  (0.1 < p < 0.9): function: @(x)ksdensity(x,'Bandwidth',.1,'Function',
        1.80824 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.227542,3.10586)
```

You can also use a parametric distribution for the center segment. Define a function that fits a normal distribution to data and returns the cdf values, and pass the function handle when you create a `paretotails` object.

```
pd2 = paretotails(x,0.1,0.9,@myfun2)
```

```
pd2 =
Piecewise distribution with 3 segments
      -Inf < x < -2.70875    (0 < p < 0.1): lower tail, GPD(-0.358104,0.831855)
   -2.70875 < x < 3.52195  (0.1 < p < 0.9): function: myfun2
        3.52195 < x < Inf    (0.9 < p < 1): upper tail, GPD(-0.0661815,5.04694)
```

```
function [p,xi] = myfun2(x)
    pd = fitdist(x,'Normal');
    xi = linspace(min(x),max(x),length(x)*2);
    p = cdf(pd,xi);
end
```

See Also

`ecdf` | `gpfit` | `ksdensity`

Topics

“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

partialcorr

Linear or rank partial correlation coefficients

Syntax

```
rho = partialcorr(x)
rho = partialcorr(x,z)
rho = partialcorr(x,y,z)
rho = partialcorr( ___ ,Name,Value)
[rho,pval] = partialcorr( ___ )
```

Description

`rho = partialcorr(x)` returns the sample linear partial correlation coefficients between pairs of variables in `x`, controlling for the remaining variables in `x`.

`rho = partialcorr(x,z)` returns the sample linear partial correlation coefficients between pairs of variables in `x`, controlling for the variables in `z`.

`rho = partialcorr(x,y,z)` returns the sample linear partial correlation coefficients between pairs of variables in `x` and `y`, controlling for the variables in `z`.

`rho = partialcorr( ___ ,Name,Value)` returns the sample linear partial correlation coefficients with additional options specified by one or more name-value pair arguments, using input arguments from any of the previous syntaxes. For example, you can specify whether to use Pearson or Spearman partial correlations, or specify how to treat missing values.

`[rho,pval] = partialcorr( ___ )` also returns a matrix `pval` of *p*-values for testing the hypothesis of no partial correlation against the one- or two-sided alternative that there is a nonzero partial correlation.

Examples

Compute Partial Correlation Coefficients

Compute partial correlation coefficients between pairs of variables in the input matrix.

Load the sample data. Convert the genders in `hospital.Sex` to numeric group identifiers.

```
load hospital;
hospital.SexID = grp2idx(hospital.Sex);
```

Create an input matrix containing the sample data.

```
x = [hospital.SexID hospital.Age hospital.Smoker hospital.Weight];
```

Each row in `x` contains a patient's gender, age, smoking status, and weight.

Compute partial correlation coefficients between pairs of variables in `x`, while controlling for the effects of the remaining variables in `x`.

```
rho = partialcorr(x)
```

```
rho = 4×4

    1.0000   -0.0105    0.0273    0.9421
   -0.0105    1.0000    0.0419    0.0369
    0.0273    0.0419    1.0000    0.0451
    0.9421    0.0369    0.0451    1.0000
```

The matrix `rho` indicates, for example, a correlation of 0.9421 between gender and weight after controlling for all other variables in `x`. You can return the *p*-values as a second output, and examine them to confirm whether these correlations are statistically significant.

For a clearer display, create a table with appropriate variable and row labels.

```
rho = array2table(rho, ...
    'VariableNames',{'SexID','Age','Smoker','Weight'},...
    'RowNames',{'SexID','Age','Smoker','Weight'});

disp('Partial Correlation Coefficients')
```

```
Partial Correlation Coefficients
```

```
disp(rho)
```

```
                SexID         Age          Smoker       Weight
              ________    _________    ________    ________

    SexID           1     -0.01052     0.027324      0.9421
    Age       -0.01052            1     0.041945    0.036873
    Smoker    0.027324     0.041945            1    0.045106
    Weight      0.9421     0.036873     0.045106           1
```

Test for Partial Correlations with Controlled Variables

Test for partial correlation between pairs of variables in the input matrix, while controlling for the effects of a second set of variables.

Load the sample data. Convert the genders in `hospital.Sex` to numeric group identifiers.

```
load hospital;
hospital.SexID = grp2idx(hospital.Sex);
```

Create two matrices containing the sample data.

```
x = [hospital.Age hospital.BloodPressure];
z = [hospital.SexID hospital.Smoker hospital.Weight];
```

The `x` matrix contains the variables to test for partial correlation. The `z` matrix contains the variables to control for. The measurements for `BloodPressure` are contained in two columns: The first column contains the upper (systolic) number, and the second column contains the lower (diastolic) number. `partialcorr` treats each column as a separate variable.

Test for partial correlation between pairs of variables in `x`, while controlling for the effects of the variables in `z`. Compute the correlation coefficients.

```
[rho,pval] = partialcorr(x,z)

rho = 3×3

    1.0000    0.1300    0.0462
    0.1300    1.0000    0.0012
    0.0462    0.0012    1.0000


pval = 3×3

         0    0.2044    0.6532
    0.2044         0    0.9903
    0.6532    0.9903         0
```

The large values in `pval` indicate that there is no significant correlation between age and either blood pressure measurement after controlling for gender, smoking status, and weight.

For a clearer display, create tables with appropriate variable and row labels.

```
rho = array2table(rho, ...
    'VariableNames',{'Age','BPTop','BPBottom'},...
    'RowNames',{'Age','BPTop','BPBottom'});

pval = array2table(pval, ...
    'VariableNames',{'Age','BPTop','BPBottom'},...
    'RowNames',{'Age','BPTop','BPBottom'});

disp('Partial Correlation Coefficients')

Partial Correlation Coefficients

disp(rho)

                Age          BPTop       BPBottom
              _______    _________    _________

    Age             1         0.13     0.046202
    BPTop        0.13            1    0.0012475
    BPBottom 0.046202    0.0012475            1

disp('p-values')

p-values

disp(pval)

                Age         BPTop      BPBottom
              _______    ________    ________

    Age             0    0.20438     0.65316
    BPTop     0.20438          0     0.99032
    BPBottom  0.65316    0.99032           0
```

Test for Paired Partial Correlation Coefficients

Test for partial correlation between pairs of variables in the `x` and `y` input matrices, while controlling for the effects of a third set of variables.

Load the sample data. Convert the genders in `hospital.Sex` to numeric group identifiers.

```
load hospital;
hospital.SexID = grp2idx(hospital.Sex);
```

Create three matrices containing the sample data.

```
x = [hospital.BloodPressure];
y = [hospital.Weight hospital.Age];
z = [hospital.SexID hospital.Smoker];
```

`partialcorr` can test for partial correlation between the pairs of variables in `x` (the systolic and diastolic blood pressure measurements) and `y` (weight and age), while controlling for the variables in `z` (gender and smoking status). The measurements for `BloodPressure` are contained in two columns: The first column contains the upper (systolic) number, and the second column contains the lower (diastolic) number. `partialcorr` treats each column as a separate variable.

Test for partial correlation between pairs of variables in `x` and `y`, while controlling for the effects of the variables in `z`. Compute the correlation coefficients.

```
[rho,pval] = partialcorr(x,y,z)
```

```
rho = 2×2

   -0.0257    0.1289
    0.0292    0.0472


pval = 2×2

    0.8018    0.2058
    0.7756    0.6442
```

The results in `pval` indicate that, after controlling for gender and smoking status, there is no significant correlation between either of a patient's blood pressure measurements and that patient's weight or age.

For a clearer display, create tables with appropriate variable and row labels.

```
rho = array2table(rho, ...
    'RowNames',{'BPTop','BPBottom'},...
    'VariableNames',{'Weight','Age'});

pval = array2table(pval, ...
    'RowNames',{'BPTop','BPBottom'},...
    'VariableNames',{'Weight','Age'});

disp('Partial Correlation Coefficients')
```

```
Partial Correlation Coefficients
```

```
disp(rho)
```

```
                Weight       Age
               ________    ________

    BPTop       -0.02568     0.12893
    BPBottom    0.029168    0.047226
```

```
disp('p-values')
```

```
p-values
```

```
disp(pval)
```

```
                Weight       Age
               _______    _______

    BPTop       0.80182     0.2058
    BPBottom    0.77556    0.64424
```

One-Tailed Partial Correlation Test

Test the hypothesis that pairs of variables have no correlation, against the alternative hypothesis that the correlation is greater than 0.

Load the sample data. Convert the genders in `hospital.Sex` to numeric group identifiers.

```
load hospital;
hospital.SexID = grp2idx(hospital.Sex);
```

Create three matrices containing the sample data.

```
x = [hospital.BloodPressure];
y = [hospital.Weight hospital.Age];
z = [hospital.SexID hospital.Smoker];
```

`partialcorr` can test for partial correlation between the pairs of variables in `x` (the systolic and diastolic blood pressure measurements) and `y` (weight and age), while controlling for the variables in `z` (gender and smoking status). The measurements for `BloodPressure` are contained in two columns: The first column contains the upper (systolic) number, and the second column contains the lower (diastolic) number. `partialcorr` treats each column as a separate variable.

Compute the correlation coefficients using a right-tailed test.

```
[rho,pval] = partialcorr(x,y,z,'Tail','right')
```

```
rho = 2×2

   -0.0257    0.1289
    0.0292    0.0472


pval = 2×2

    0.5991    0.1029
    0.3878    0.3221
```

The results in `pval` indicate that `partialcorr` does not reject the null hypothesis of nonzero correlations between the variables in `x` and `y`, after controlling for the variables in `z`, when the alternative hypothesis is that the correlations are greater than 0.

For a clearer display, create tables with appropriate variable and row labels.

```
rho = array2table(rho, ...
    'RowNames',{'BPTop','BPBottom'},...
    'VariableNames',{'Weight','Age'});

pval = array2table(pval, ...
    'RowNames',{'BPTop','BPBottom'},...
    'VariableNames',{'Weight','Age'});

disp('Partial Correlation Coefficients')
```

```
Partial Correlation Coefficients
```

```
disp(rho)
```

```
                Weight       Age
               ________    ________

    BPTop       -0.02568     0.12893
    BPBottom    0.029168    0.047226
```

```
disp('p-values')
```

```
p-values
```

```
disp(pval)
```

```
               Weight       Age
               _______    _______

    BPTop      0.59909     0.1029
    BPBottom   0.38778    0.32212
```

Input Arguments

x — Data matrix

matrix

Data matrix, specified as an n-by-p_x matrix. The rows of `x` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

y — Data matrix

matrix

Data matrix, specified as an n-by-p_y matrix. The rows of `y` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

z — Data matrix

matrix

Data matrix, specified as an n-by-p_z matrix. The rows of `z` correspond to observations, and columns correspond to variables.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Type','Spearman','Rows','complete'` computes Spearman partial correlations using only the data in rows that contain no missing values.

Type — Type of partial correlations

`'Pearson'` (default) | `'Spearman'`

Type of partial correlations to compute, specified as the comma-separated pair consisting of `'Type'` and one of the following.

`'Pearson'`	Compute Pearson (linear) partial correlations.
`'Spearman'`	Compute Spearman (rank) partial correlations.

Example: `'Type','Spearman'`

Rows — Rows to use in computation

`'all'` (default) | `'complete'` | `'pairwise'`

Rows to use in computation, specified as the comma-separated pair consisting of `'Rows'` and one of the following.

`'all'`	Use all rows of the input regardless of missing values (`NaN`s).
`'complete'`	Use only rows of the input with no missing values.
`'pairwise'`	Compute `rho(i,j)` using rows with no missing values in column `i` or `j`.

Example: `'Rows','complete'`

Tail — Alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Alternative hypothesis to test against, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternative hypothesis that the correlation is not 0.
`'right'`	Test the alternative hypothesis that the correlation is greater than 0.
`'left'`	Test the alternative hypothesis that the correlation is less than 0.

Example: `'Tail','right'`

Output Arguments

rho — Sample linear partial correlation coefficients

matrix

Sample linear partial correlation coefficients, returned as a matrix.

- If you input only an `x` matrix, `rho` is a symmetric p_x-by-p_x matrix. The (i,j)th entry is the sample linear partial correlation between the i-th and j-th columns in `x`.
- If you input `x` and `z` matrices, `rho` is a symmetric p_x-by-p_x matrix. The (i,j)th entry is the sample linear partial correlation between the ith and jth columns in `x`, controlled for the variables in `z`.
- If you input `x`, `y`, and `z` matrices, `rho` is a p_x-by-p_y matrix, where the (i,j)th entry is the sample linear partial correlation between the ith column in `x` and the jth column in `y`, controlled for the variables in `z`.

If the covariance matrix of `[x,z]` is

$$S = \begin{pmatrix} S_{xx} & S_{xz} \\ S_{xz}{}^T & S_{zz} \end{pmatrix},$$

then the partial correlation matrix of `x`, controlling for `z`, can be defined formally as a normalized version of the covariance matrix: $S_{xx} - (S_{xz}S_{zz}{}^{-1}S_{xz}{}^T)$.

`pval` — *p*-values
matrix

p-values, returned as a matrix. Each element of `pval` is the *p*-value for the corresponding element of `rho`.

If `pval(i,j)` is small, then the corresponding partial correlation `rho(i,j)` is statistically significantly different from 0.

`partialcorr` computes *p*-values for linear and rank partial correlations using a Student's *t* distribution for a transformation of the correlation. This is exact for linear partial correlation when `x` and `z` are normal, but is a large-sample approximation otherwise.

References

[1] Stuart, Alan, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics*. 6th edition, Volume 2A, Chapter 28, Wiley, 2004.

[2] Fisher, Ronald A. "The Distribution of the Partial Correlation Coefficient." *Metron* 3 (1924): 329-332

See Also

`corr` | `corrcoef` | `partialcorri` | `tiedrank`

Introduced before R2006a

partialcorri

Partial correlation coefficients adjusted for internal variables

Syntax

```
rho = partialcorri(y,x)
rho = partialcorri(y,x,z)
rho = partialcorri( ___ ,Name,Value)
[rho,pval] = partialcorri( ___ )
```

Description

`rho = partialcorri(y,x)` returns the sample linear partial correlation coefficients between pairs of variables in `y` and `x`, adjusting for the remaining variables in `x`.

`rho = partialcorri(y,x,z)` returns the sample linear partial correlation coefficients between pairs of variables in `y` and `x`, adjusting for the remaining variables in `x`, after first controlling both `x` and `y` for the variables in `z`.

`rho = partialcorri( ___ ,Name,Value)` returns the sample linear partial correlation coefficients with additional options specified by one or more name-value pair arguments, using input arguments from any of the previous syntaxes. For example, you can specify whether to use Pearson or Spearman partial correlations, or specify how to treat missing values.

`[rho,pval] = partialcorri( ___ )` also returns a matrix `pval` of *p*-values for testing the hypothesis of no partial correlation against the one- or two-sided alternative that there is a nonzero partial correlation.

Examples

Compute Partial Correlation Coefficients

Compute partial correlation coefficients for each pair of variables in the `x` and `y` input matrices, while controlling for the effects of the remaining variables in `x`.

Load the sample data.

```
load carsmall;
```

The data contains measurements from cars manufactured in 1970, 1976, and 1982. It includes `MPG` and `Acceleration` as performance measures, and `Displacement`, `Horsepower`, and `Weight` as design variables. `Acceleration` is the time required to accelerate from 0 to 60 miles per hour, so a high value for `Acceleration` corresponds to a vehicle with low acceleration.

Define the input matrices. The `y` matrix includes the performance measures, and the `x` matrix includes the design variables.

```
y = [MPG,Acceleration];
x = [Displacement,Horsepower,Weight];
```

Compute the correlation coefficients. Include only rows with no missing values in the computation.

```
rho = partialcorri(y,x,'Rows','complete')
```

```
rho = 2×3

   -0.0537   -0.1520   -0.4856
   -0.3994   -0.4008    0.4912
```

The results suggest, for example, a 0.4912 correlation between weight and acceleration after controlling for the effects of displacement and horsepower. You can return the *p*-values as a second output, and examine them to confirm whether these correlations are statistically significant.

For a clearer display, create a table with appropriate variable and row labels.

```
rho = array2table(rho, ...
   'VariableNames',{'Displacement','Horsepower','Weight'}, ...
   'RowNames',{'MPG','Acceleration'});

disp('Partial Correlation Coefficients')
```

```
Partial Correlation Coefficients
```

```
disp(rho)
```

```
                    Displacement    Horsepower     Weight
                    ____________    __________    ________

    MPG               -0.053684       -0.15199    -0.48563
    Acceleration       -0.39941       -0.40075     0.49123
```

Test Partial Correlations While Controlling for Additional Variables

Test for partial correlation between pairs of variables in the `x` and `y` input matrices, while controlling for the effects of the remaining variables in `x` plus additional variables in matrix `z`.

Load the sample data.

```
load carsmall;
```

The data contains measurements from cars manufactured in 1970, 1976, and 1982. It includes `MPG` and `Acceleration` as performance measures, and `Displacement`, `Horsepower`, and `Weight` as design variables. `Acceleration` is the time required to accelerate from 0 to 60 miles per hour, so a high value for `Acceleration` corresponds to a vehicle with low acceleration.

Create a new variable `Headwind`, and randomly generate data to represent the notion of an average headwind along the performance measurement route.

```
rng('default');  % For reproducibility
Headwind = (10:-0.2:-9.8)' + 5*randn(100,1);
```

Since headwind can affect the performance measures, control for its effects when testing for partial correlation between the remaining variables.

Define the input matrices. The `y` matrix includes the performance measures, and the `x` matrix includes the design variables. The `z` matrix contains additional variables to control for when computing the partial correlations, such as headwind.

```
y = [MPG,Acceleration];
x = [Displacement,Horsepower,Weight];
z = Headwind;
```

Compute the partial correlation coefficients. Include only rows with no missing values in the computation.

```
[rho,pval] = partialcorri(y,x,z,'Rows','complete')
```

```
rho = 2×3

    0.0572   -0.1055   -0.5736
   -0.3845   -0.3966    0.4674


pval = 2×3

    0.5923    0.3221    0.0000
    0.0002    0.0001    0.0000
```

The small returned *p*-value of 0.001 in `pval` indicates, for example, a significant negative correlation between horsepower and acceleration, after controlling for displacement, weight, and headwind.

For a clearer display, create tables with appropriate variable and row labels.

```
rho = array2table(rho, ...
   'VariableNames',{'Displacement','Horsepower','Weight'}, ...
   'RowNames',{'MPG','Acceleration'});

pval = array2table(pval, ...
   'VariableNames',{'Displacement','Horsepower','Weight'}, ...
   'RowNames',{'MPG','Acceleration'});

disp('Partial Correlation Coefficients, Accounting for Headwind')
```

```
Partial Correlation Coefficients, Accounting for Headwind
```

```
disp(rho)
```

```
                    Displacement    Horsepower     Weight
                    ____________    __________    ________

    MPG                0.057197       -0.10555    -0.57358
    Acceleration       -0.38452       -0.39658      0.4674
```

```
disp('p-values, Accounting for Headwind')
```

```
p-values, Accounting for Headwind
```

```
disp(pval)
```

```
                    Displacement    Horsepower      Weight
                    ____________    __________    __________
```

```
MPG                0.59233        0.32212    3.4401e-09
Acceleration    0.00018272     0.00010902    3.4091e-06
```

Input Arguments

x — Data matrix

matrix

Data matrix, specified as an n-by-p_x matrix. The rows of `x` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

y — Data matrix

matrix

Data matrix, specified as an n-by-p_y matrix. The rows of `y` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

z — Data matrix

matrix

Data matrix, specified as an n-by-p_z matrix. The rows of `z` correspond to observations, and the columns correspond to variables.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Type','Spearman','Rows','complete'` computes Spearman partial correlations using only the data in rows that contain no missing values.

Type — Type of partial correlations

`'Pearson'` (default) | `'Spearman'`

Type of partial correlations to compute, specified as the comma-separated pair consisting of `'Type'` and either `'Pearson'` or `'Spearman'`. `Pearson` computes the Pearson (linear) partial correlations. `Spearman` computes the Spearman (rank) partial correlations.

Example: `'Type','Spearman'`

Rows — Rows to use in computation

`'all'` (default) | `'complete'` | `'pairwise'`

Rows to use in computation, specified as the comma-separated pair consisting of `'Rows'` and one of the following.

`'all'`	Use all rows regardless of missing `(NaN)` values.
`'complete'`	Use only rows with no missing values.

`'pairwise'`	Use all available values in each column of `y` when computing the partial correlation coefficients and *p*-values corresponding to that column. For each column of `y`, rows will be dropped corresponding to missing values in `x` (and/or `z`, if supplied). However, remaining rows with valid values in that column of `y` are used, even if there are missing values in other columns of `y`.

Example: `'Rows','complete'`

`Tail` — Alternative hypothesis
`'both'` (default) | `'right'` | `'left'`

Alternative hypothesis to test against, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternative hypothesis that the correlation is not zero.
`'right'`	Test the alternative hypothesis that the correlation is greater than 0.
`'left'`	Test the alternative hypothesis that the correlation is less than 0.

Example: `'Tail','right'`

Output Arguments

`rho` — Sample linear partial correlation coefficients
matrix

Sample linear partial correlation coefficients, returned as a p_y-by-p_x matrix.

- If you input `x` and `y` matrices, the (*i*,*j*)th entry is the sample linear partial correlation between the *i*th column in `y` and the *j*th column in `x`, controlled for all the columns of `x` except column *j*.
- If you input `x`, `y`, and `z` matrices, the (*i*,*j*)th entry is the sample linear partial correlation between the *i*th column in `y` and the *j*th column in `x`, adjusted for all the columns of `x` except column *j*, after first controlling both `x` and `y` for the variables in `z`.

`pval` — *p*-values
matrix

p-values, returned as a matrix. Each element of `pval` is the *p*-value for the corresponding element of `rho`. If `pval(i,j)` is small, then the corresponding partial correlation `rho(i,j)` is statistically significantly different from zero.

`partialcorri` computes *p*-values for linear and rank partial correlations using a Student's *t* distribution for a transformation of the correlation. This is exact for linear partial correlation when `x` and `z` are normal, but is a large-sample approximation otherwise.

References

[1] Stuart, Alan, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics*. 6th edition, Volume 2A, Chapter 28, Wiley, 2004.

[2] Fisher, Ronald A. "The Distribution of the Partial Correlation Coefficient." *Metron* 3 (1924): 329-332

See Also

`corr` | `partialcorr`

Introduced in R2013b

partialDependence

Package:

Compute partial dependence

Syntax

```
pd = partialDependence(RegressionMdl,Vars)

pd = partialDependence(ClassificationMdl,Vars,Labels)

pd = partialDependence( ___ ,Data)
pd = partialDependence( ___ ,Name,Value)
[pd,x,y] = partialDependence( ___ )
```

Description

`pd = partialDependence(RegressionMdl,Vars)` computes the partial dependence `pd` between the predictor variables listed in `Vars` and the responses predicted by using the regression model `RegressionMdl`, which contains predictor data.

`pd = partialDependence(ClassificationMdl,Vars,Labels)` computes the partial dependence `pd` between the predictor variables listed in `Vars` and the scores for the classes specified in `Labels` by using the classification model `ClassificationMdl`, which contains predictor data.

`pd = partialDependence( ___ ,Data)` uses new predictor data in `Data`. You can specify `Data` in addition to any of the input argument combinations in the previous syntaxes.

`pd = partialDependence( ___ ,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, if you specify `'UseParallel','true'`, the `partialDependence` function uses parallel computing to perform the partial dependence calculations.

`[pd,x,y] = partialDependence( ___ )` also returns `x` and `y`, which contain the query points of the first and second predictor variables in `Vars`, respectively. If you specify one variable in `Vars`, then `partialDependence` returns an empty matrix (`[]`) for `y`.

Examples

Compute and Plot Partial Dependence on One Variable

Train a naive Bayes classification model with the `fisheriris` data set, and compute partial dependence values that show the relationship between the predictor variable and the predicted scores (posterior probabilities) for multiple classes.

Load the `fisheriris` data set, which contains species (`species`) and measurements (`meas`) on sepal length, sepal width, petal length, and petal width for 150 iris specimens. The data set contains 50 specimens from each of three species: setosa, versicolor, and virginica.

```
load fisheriris
```

Train a naive Bayes classification model with `species` as the response and `meas` as predictors.

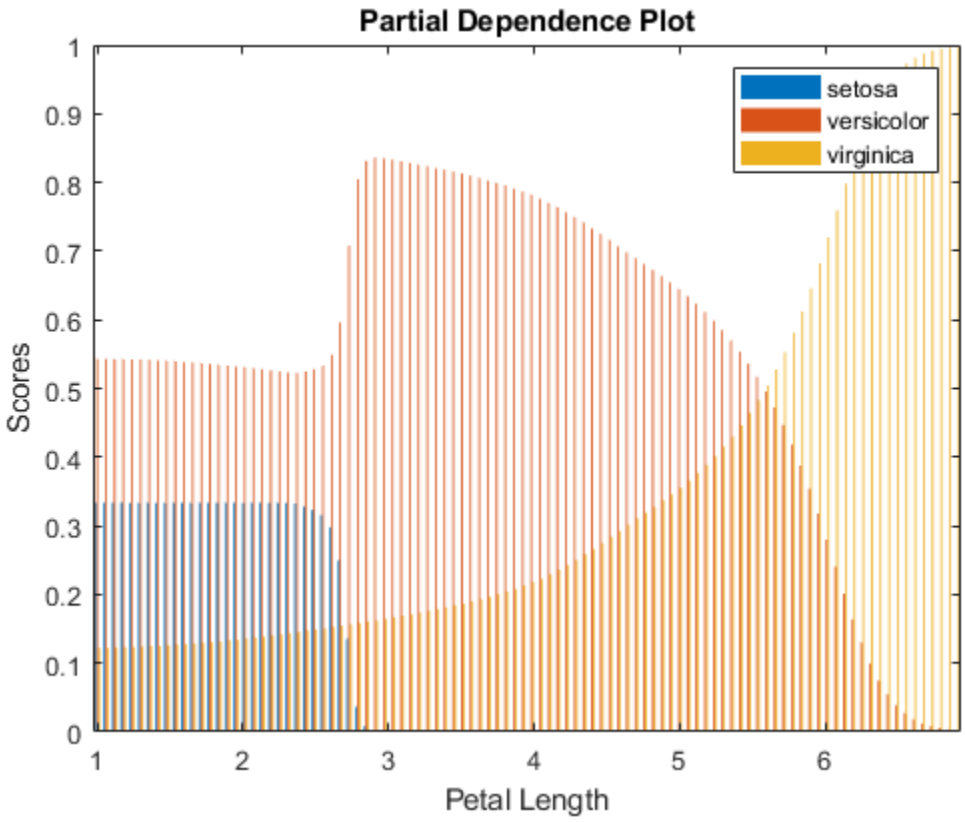
```
Mdl = fitcnb(meas,species,'PredictorNames',["Sepal Length","Sepal Width","Petal Length","Petal W
```

Compute partial dependence values on the third predictor variable (petal length) of the scores predicted by `Mdl` for all three classes of `species`. Specify the class labels by using the `ClassNames` property of `Mdl`.

```
[pd,x] = partialDependence(Mdl,3,Mdl.ClassNames);
```

pd contains the partial dependence values for the query points x. You can plot the computed partial dependence values by using plotting functions such as `plot` and `bar`. Plot `pd` against `x` by using the `bar` function.

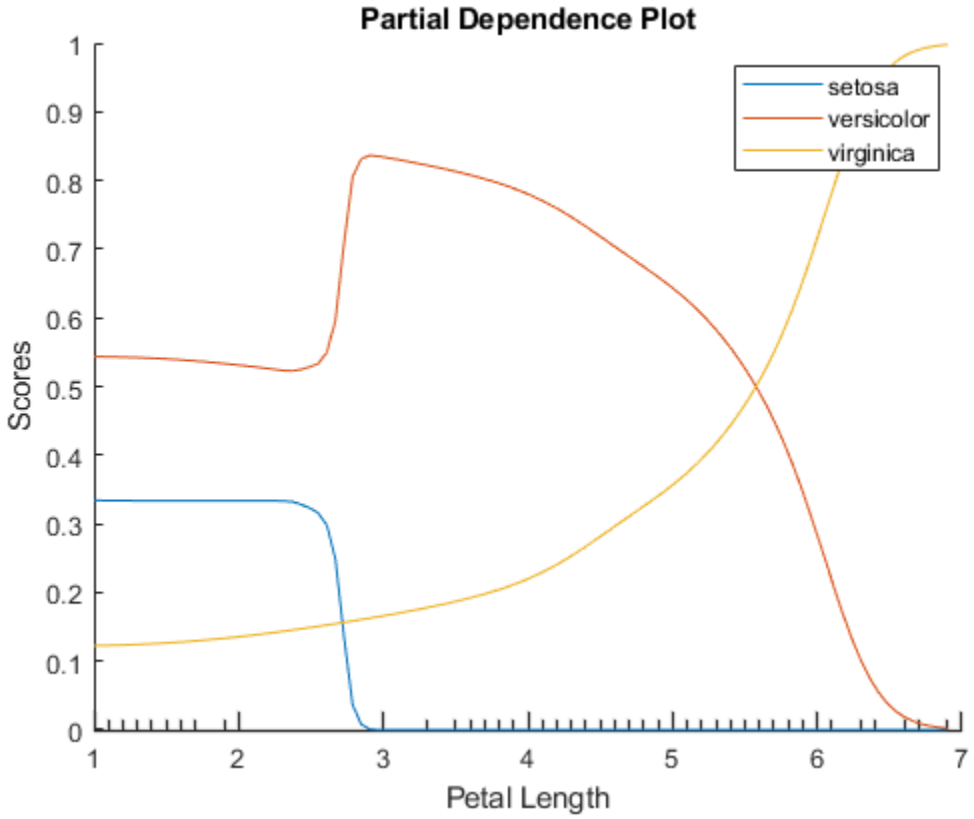
```
bar(x,pd)
legend(Mdl.ClassNames)
xlabel("Petal Length")
ylabel("Scores")
title("Partial Dependence Plot")
```



According to this model, the probability of `virginica` increases with petal length. The probability of `setosa` is about 0.33, from where petal length is 0 to around 2.5, and then the probability drops to almost 0.

Alternatively, you can use the `plotPartialDependence` function to compute and plot partial dependence values.

```
plotPartialDependence(Mdl,3,Mdl.ClassNames)
```



Compute and Plot Partial Dependence on Two Variables for Multiple Classes

Train an ensemble of classification models and compute partial dependence values on two variables for multiple classes. Then plot the partial dependence values for each class.

Load the `census1994` data set, which contains US yearly salary data, categorized as `<=50K` or `>50K`, and several demographic variables.

```
load census1994
```

Extract a subset of variables to analyze from the table `adultdata`.

```
X = adultdata(1:500,{'age','workClass','education_num','marital_status','race', ...
    'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

Train a random forest of classification trees by using `fitcensemble` and specifying `'Method'` as `'Bag'`. For reproducibility, use a template of trees created by using `templateTree` with the `'Reproducible'` option.

```
rng('default')
t = templateTree('Reproducible',true);
Mdl = fitcensemble(X,'salary','Method','Bag','Learners',t);
```

Inspect the class names in `Mdl`.

```
Mdl.ClassNames
```

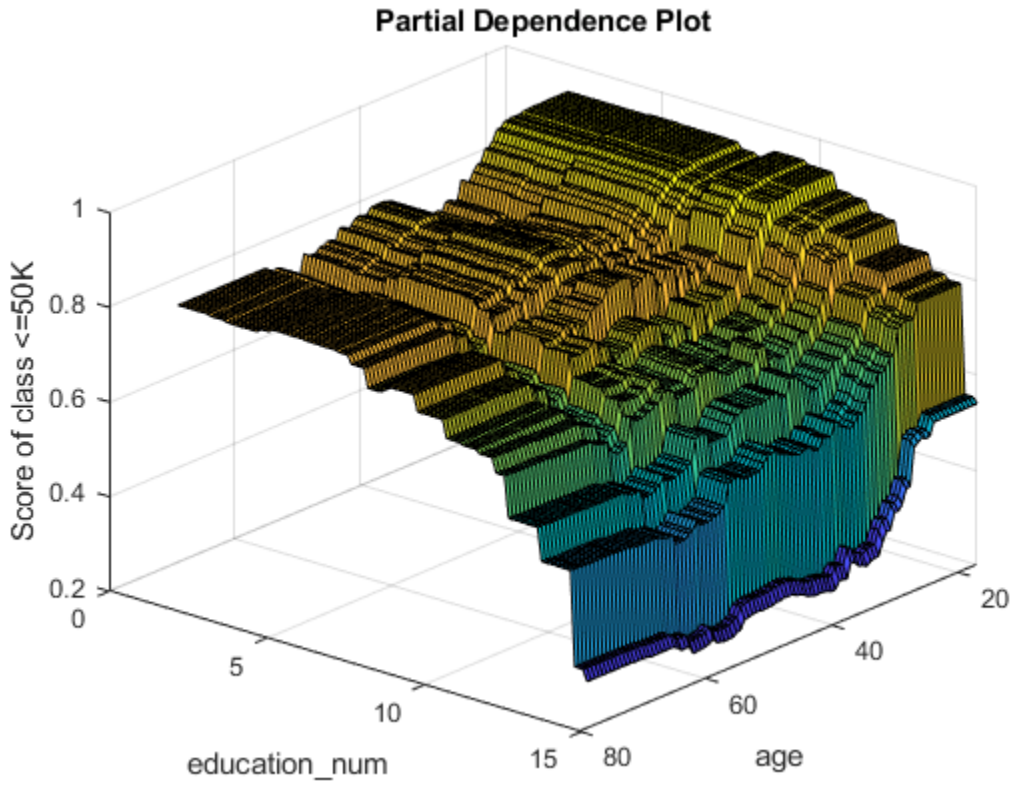
```
ans = 2x1 categorical
     <=50K
     >50K
```

Compute partial dependence values of the scores on the predictors `age` and `education_num` for both classes (`<=50K` and `>50K`). Specify the number of observations to sample as 100.

```
[pd,x,y] = partialDependence(Mdl,{'age','education_num'},Mdl.ClassNames,'NumObservationsToSample
```

Create a surface plot of the partial dependence values for the first class (`<=50K`) by using the `surfl` function.

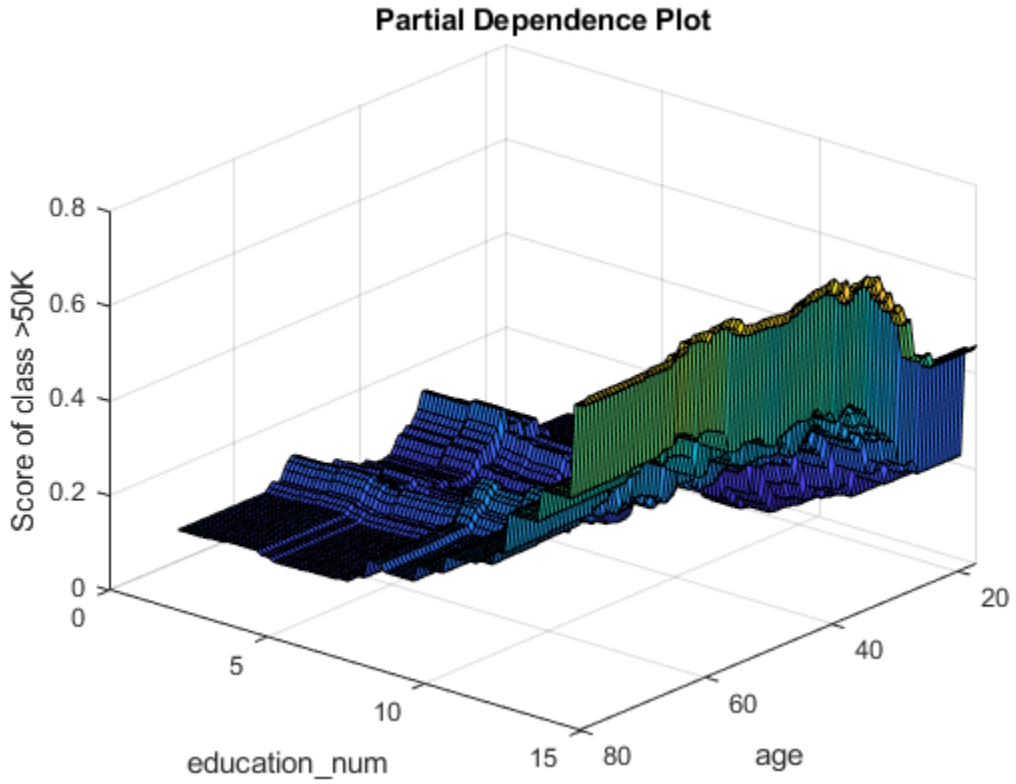
```
figure
surf(x,y,squeeze(pd(1,:,:)))
xlabel('age')
ylabel('education\_num')
zlabel('Score of class <=50K')
title('Partial Dependence Plot')
view([130 30]) % Modify the viewing angle
```



Create a surface plot of the partial dependence values for the second class (`>50K`).

```
figure
surf(x,y,squeeze(pd(2,:,:)))
```

```
xlabel('age')
ylabel('education\_num')
zlabel('Score of class >50K')
title('Partial Dependence Plot')
view([130 30]) % Modify the viewing angle
```



The two plots show different partial dependence patterns depending on the class.

Compute and Plot Partial Dependence on Multiple Variables for Regression

Train a support vector machine (SVM) regression model using the `carsmall` data set, and compute the partial dependence on two predictor variables. Then, create a figure that shows the partial dependence on the two variables along with the histogram on each variable.

Load the `carsmall` data set.

```
load carsmall
```

Create a table that contains `Weight`, `Cylinders`, `Displacement`, and `Horsepower`.

```
Tbl = table(Weight,Cylinders,Displacement,Horsepower);
```

Train an SVM regression model using the predictor variables in `Tbl` and the response variable `MPG`. Use a Gaussian kernel function with an automatic kernel scale.

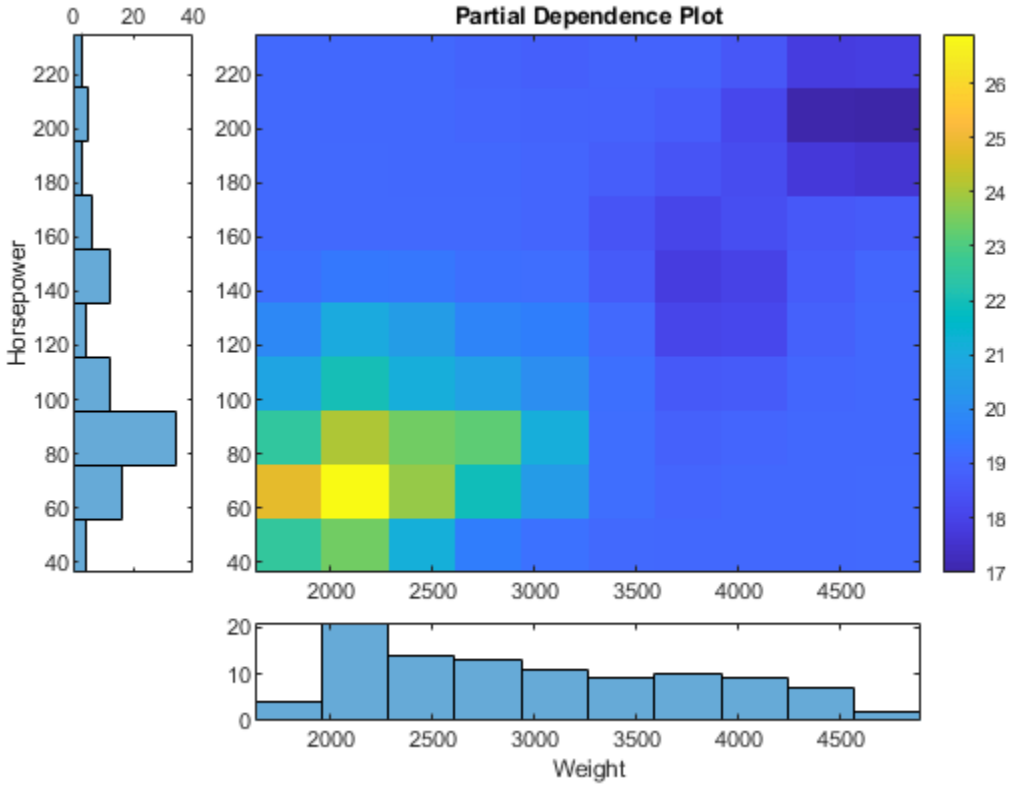
```
Mdl = fitrsvm(Tbl,MPG,'ResponseName','MPG', ...
    'CategoricalPredictors','Cylinders','Standardize',true, ...
    'KernelFunction','gaussian','KernelScale','auto');
```

Compute the partial dependence of the predicted response (`MPG`) on the predictor variables `Weight` and `Horsepower`. Specify query points to compute the partial dependence by using the `'QueryPoints'` name-value pair argument.

```
numPoints = 10;
ptX = linspace(min(Weight),max(Weight),numPoints)';
ptY = linspace(min(Horsepower),max(Horsepower),numPoints)';
[pd,x,y] = partialDependence(Mdl,{'Weight','Horsepower'},'QueryPoints',[ptX ptY]);
```

Create a figure that contains a 5-by-5 tiled chart layout. Plot the partial dependence on the two variables by using the `imagesc` function. Then draw the histogram for each variable by using the `histogram` function. Specify the edges of the histograms so that the centers of the histogram bars align with the query points. Change the axes properties to align the axes of the plots.

```
t = tiledlayout(5,5,'TileSpacing','compact');

ax1 = nexttile(2,[4,4]);
imagesc(x,y,pd)
title('Partial Dependence Plot')
colorbar('eastoutside')
ax1.YDir = 'normal';

ax2 = nexttile(22,[1,4]);
dX = diff(ptX(1:2));
edgeX = [ptX-dX/2;ptX(end)+dX];
histogram(Weight,edgeX);
xlabel('Weight')
xlim(ax1.XLim);

ax3 = nexttile(1,[4,1]);
dY = diff(ptY(1:2));
edgeY = [ptY-dY/2;ptY(end)+dY];
histogram(Horsepower,edgeY)
xlabel('Horsepower')
xlim(ax1.YLim);
ax3.XDir = 'reverse';
camroll(-90)
```



Each element of `pd` specifies the color for one pixel of the image plot. The histograms aligned with the axes of the image show the distribution of the predictors.

Input Arguments

`RegressionMdl` — Regression model

regression model object

Regression model, specified as a full or compact regression model object, as given in the following tables of supported models.

Model	Full or Compact Model Object
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`
Generalized linear mixed-effect model	`GeneralizedLinearMixedModel`
Linear regression	`LinearModel`, `CompactLinearModel`
Linear mixed-effect model	`LinearMixedModel`
Nonlinear regression	`NonLinearModel`
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`

Model	Full or Compact Model Object
Generalized additive model (GAM)	`RegressionGAM`, `CompactRegressionGAM`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`
Linear regression for high-dimensional data	`RegressionLinear`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`
Support vector machine (SVM) regression	`RegressionSVM`, `CompactRegressionSVM`
Regression tree	`RegressionTree`, `CompactRegressionTree`
Bootstrap aggregation for ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`

If `RegressionMdl` is a model object that does not contain predictor data (for example, a compact model), you must provide the input argument `Data`.

`partialDependence` does not support a model object trained with a sparse matrix. When you train a model, use a full numeric matrix or table for predictor data where rows correspond to individual observations.

`ClassificationMdl` — Classification model
classification model object

Classification model, specified as a full or compact classification model object, as given in the following tables of supported models.

Model	Full or Compact Model Object
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`
Generalized additive model (GAM)	`ClassificationGAM`, `CompactClassificationGAM`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`
k-nearest neighbor classifier	`ClassificationKNN`
Linear classification model	`ClassificationLinear`

Model	Full or Compact Model Object
Multiclass naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`
Support vector machine (SVM) classifier for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`
Bagged ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`

If `ClassificationMdl` is a model object that does not contain predictor data (for example, a compact model), you must provide the input argument `Data`.

`partialDependence` does not support a model object trained with a sparse matrix. When you train a model, use a full numeric matrix or table for predictor data where rows correspond to individual observations.

`Vars` — Predictor variables

vector of positive integers | character vector | string scalar | string array | cell array of character vectors

Predictor variables, specified as a vector of positive integers, character vector, string scalar, string array, or cell array of character vectors. You can specify one or two predictor variables, as shown in the following tables.

One Predictor Variable

Value	Description
positive integer	Index value corresponding to the column of the predictor data.
character vector or string scalar	Name of a predictor variable. The name must match the entry in `RegressionMdl.PredictorNames` or `ClassificationMdl.PredictorNames`.

Two Predictor Variables

Value	Description
vector of two positive integers	Index values corresponding to the columns of the predictor data.
string array or cell array of character vectors	Names of predictor variables. Each element in the array is the name of a predictor variable. The names must match the entries in `RegressionMdl.PredictorNames` or `ClassificationMdl.PredictorNames`.

Example: `{'x1','x3'}`

Data Types: `single` | `double` | `char` | `string` | `cell`

`Labels` — Class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. The values and data types in `Labels` must match those of the class names in the `ClassNames` property of `ClassificationMdl` (`ClassificationMdl.ClassNames`).

You can specify one or multiple class labels.

This argument is valid only when `ClassificationMdl` is a classification model object.

Example: `{'red','blue'}`

Example: `ClassificationMdl.ClassNames([1 3])` specifies `Labels` as the first and third classes in `ClassificationMdl`.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Data` — Predictor data

numeric matrix | table

Predictor data, specified as a numeric matrix or table. Each row of `Data` corresponds to one observation, and each column corresponds to one variable.

`Data` must be consistent with the predictor data that trained the model (`RegressionMdl` or `ClassificationMdl`), stored in either the `X` or `Variables` property.

- If you trained the model using a numeric matrix, then `Data` must be a numeric matrix. The variables making up the columns of `Data` must have the same number and order as the predictor variables that trained the model.
- If you trained the model using a table (for example, `Tbl`), then `Data` must be a table. All predictor variables in `Data` must have the same variable names and data types as the names and types in `Tbl`. However, the column order of `Data` does not need to correspond to the column order of `Tbl`.
- `partialDependence` does not support a sparse matrix.

If `RegressionMdl` or `ClassificationMdl` is a model object that does not contain predictor data, you must provide `Data`. If the model is a full model object that contains predictor data and you specify this argument, then `partialDependence` does not use the predictor data in the model and uses `Data` only.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `partialDependence(Mdl,Vars,Data,'NumObservationsToSample',100,'UseParallel',true)` computes the partial dependence values by using 100 sampled observations in `Data` and executing `for`-loop iterations in parallel.

`IncludeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the generalized additive model (GAM) in the partial dependence computation, specified as `true` or `false`. This argument is valid only for a GAM. That is, you can specify this argument only when `RegressionMdl` is `RegressionGAM` or `CompactRegressionGAM`, or `ClassificationMdl` is `ClassificationGAM` or `CompactClassificationGAM`.

The default `'IncludeInteractions'` value is `true` if the model contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`IncludeIntercept` — Flag to include intercept term
`true` (default) | `false`

Flag to include an intercept term of the generalized additive model (GAM) in the partial dependence computation, specified as `true` or `false`. This argument is valid only for a GAM. That is, you can specify this argument only when `RegressionMdl` is `RegressionGAM` or `CompactRegressionGAM`, or `ClassificationMdl` is `ClassificationGAM` or `CompactClassificationGAM`.

Example: `'IncludeIntercept',false`

Data Types: `logical`

`NumObservationsToSample` — Number of observations to sample
number of total observations (default) | positive integer

Number of observations to sample, specified as a positive integer. The default value is the number of total observations in `Data` or the model (`RegressionMdl` or `ClassificationMdl`). If you specify a value larger than the number of total observations, then `partialDependence` uses all observations.

`partialDependence` samples observations without replacement by using the `datasample` function and uses the sampled observations to compute partial dependence.

Example: `'NumObservationsToSample',100`

Data Types: `single` | `double`

`QueryPoints` — Points to compute partial dependence
numeric column vector | numeric two-column matrix | cell array of two numeric column vectors

Points to compute partial dependence for numeric predictors, specified as a numeric column vector, a numeric two-column matrix, or a cell array of two numeric column vectors.

- If you select one predictor variable in `Vars`, use a numeric column vector.
- If you select two predictor variables in `Vars`:
  - Use a numeric two-column matrix to specify the same number of points for each predictor variable.
  - Use a cell array of two numeric column vectors to specify a different number of points for each predictor variable.

The default value is a numeric column vector or a numeric two-column matrix, depending on the number of selected predictor variables. Each column contains 100 evenly spaced points between the minimum and maximum values of the sampled observations for the corresponding predictor variable.

You cannot modify `'QueryPoints'` for a categorical variable. The `partialDependence` function uses all categorical values in the selected variable.

If you select one numeric variable and one categorical variable, you can specify `'QueryPoints'` for a numeric variable by using a cell array consisting of a numeric column vector and an empty array.

Example: `'QueryPoints',{pt,[]}`

Data Types: `single` | `double` | `cell`

`UseParallel` — Flag to run in parallel
`false` (default) | `true`

Flag to run in parallel, specified as `true` or `false`. If you specify `'UseParallel',true`, the `partialDependence` function executes `for`-loop iterations in parallel by using `parfor` when predicting responses or scores for each observation and averaging them. This option requires Parallel Computing Toolbox.

Example: `'UseParallel',true`

Data Types: `logical`

Output Arguments

`pd` — Partial dependence values
numeric array

Partial dependence values, returned as a `numX`-by-`numY` numeric matrix (for a regression model) or `numLabels`-by-`numX`-by-`numY` numeric array (for a classification model). `numX` and `numY` are the number of query points of the first and second variables in `Vars`, respectively. `numLabels` is the number of class labels in `Labels`.

The value in `pd(i,j,k)` is the partial dependence value of the query point `x(j)` and `y(k)` for the `i`th class label. `x(j)` is the `j`th query point of the first predictor variable, and `y(k)` is the `k`th query point of the second predictor variable.

`x` — Query points of first predictor variable
numeric column vector | categorical column vector

Query points of the first predictor variable in `Vars`, returned as a numeric or categorical column vector.

If the predictor variable is numeric, then you can specify the query points by using the `'QueryPoints'` name-value pair argument.

Data Types: `single` | `double` | `categorical`

`y` — Query points of second predictor variable
numeric column vector | categorical column vector

Query points of the second predictor variable in `Vars`, returned as a numeric or categorical column vector.

If the predictor variable is numeric, then you can specify the query points by using the `'QueryPoints'` name-value pair argument.

Data Types: `single` | `double` | `categorical`

More About

Partial Dependence for Regression Models

Partial dependence[1] represents the relationships between predictor variables and predicted responses in a trained regression model. `partialDependence` computes the partial dependence of predicted responses on a subset of predictor variables by marginalizing over the other variables.

Consider partial dependence on a subset X^S of the whole predictor variable set $X = \{x_1, x_2, ..., x_m\}$. A subset X^S includes either one variable or two variables: $X^S = \{x_{S1}\}$ or $X^S = \{x_{S1}, x_{S2}\}$. Let X^C be the complementary set of X^S in X. A predicted response $f(X)$ depends on all variables in X:

$f(X) = f(X^S, X^C)$.

The partial dependence of predicted responses on X^S is defined by the expectation of predicted responses with respect to X^C:

$$f^S\left(X^S\right) = E_C\left[f\left(X^S, X^C\right)\right] = \int f\left(X^S, X^C\right) p_C\left(X^C\right) dX^C,$$

where $p_C(X^C)$ is the marginal probability of X^C, that is, $p_C\left(X^C\right) \approx \int p\left(X^S, X^C\right) dX^S$. Assuming that each observation is equally likely, and the dependence between X^S and X^C and the interactions of X^S and X^C in responses is not strong, `partialDependence` estimates the partial dependence by using observed predictor data as follows:

$$f^S\left(X^S\right) \approx \frac{1}{N} \sum_{i=1}^{N} f\left(X^S, X_i{}^C\right), \tag{33-1}$$

where N is the number of observations and $X_i = (X_i{}^S, X_i{}^C)$ is the ith observation.

When you call the `partialDependence` function, you can specify a trained model ($f(\cdot)$) and select variables (X^S) by using the input arguments `RegressionMdl` and `Vars`, respectively. `partialDependence` computes the partial dependence at 100 evenly spaced points of X^S or the points that you specify by using the `'QueryPoints'` name-value pair argument. You can specify the number (N) of observations to sample from given predictor data by using the `'NumObservationsToSample'` name-value pair argument.

Partial Dependence Classification Models

In the case of classification models, `partialDependence` computes the partial dependence in the same way as for regression models, with one exception: instead of using the predicted responses from the model, the function uses the predicted scores for the classes specified in `Labels`.

Weighted Traversal Algorithm

The weighted traversal algorithm[1] is a method to estimate partial dependence for a tree-based model. The estimated partial dependence is the weighted average of response or score values corresponding to the leaf nodes visited during the tree traversal.

Let X^S be a subset of the whole variable set X and X^C be the complementary set of X^S in X. For each X^S value to compute partial dependence, the algorithm traverses a tree from the root (beginning) node down to leaf (terminal) nodes and finds the weights of leaf nodes. The traversal starts by assigning a weight value of one at the root node. If a node splits by X^S, the algorithm traverses to the appropriate child node depending on the X^S value. The weight of the child node becomes the same value as its

parent node. If a node splits by X^C, the algorithm traverses to both child nodes. The weight of each child node becomes a value of its parent node multiplied by the fraction of observations corresponding to each child node. After completing the tree traversal, the algorithm computes the weighted average by using the assigned weights.

For an ensemble of bagged trees, the estimated partial dependence is an average of the weighted averages over the individual trees.

Algorithms

`partialDependence` uses a `predict` function to predict responses or scores. `partialDependence` chooses the proper `predict` function according to the model (`RegressionMdl` or `ClassificationMdl`) and runs `predict` with its default settings. For details about each `predict` function, see the `predict` functions in the following two tables. If the specified model is a tree-based model (not including a boosted ensemble of trees), then `partialDependence` uses the weighted traversal algorithm instead of the `predict` function. For details, see “Weighted Traversal Algorithm” on page 33-4438.

Regression Model Object

Model Type	Full or Compact Regression Model Object	Function to Predict Responses
Bootstrap aggregation for ensemble of decision trees	`CompactTreeBagger`	`predict`
Bootstrap aggregation for ensemble of decision trees	`TreeBagger`	`predict`
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`	`predict`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`	`predict`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`	`predict`
Generalized additive model	`RegressionGAM`, `CompactRegressionGAM`	`predict`
Generalized linear mixed-effect model	`GeneralizedLinearMixedModel`	`predict`
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`	`predict`
Linear mixed-effect model	`LinearMixedModel`	`predict`
Linear regression	`LinearModel`, `CompactLinearModel`	`predict`
Linear regression for high-dimensional data	`RegressionLinear`	`predict`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`	`predict`
Nonlinear regression	`NonLinearModel`	`predict`
Regression tree	`RegressionTree`, `CompactRegressionTree`	`predict`
Support vector machine	`RegressionSVM`, `CompactRegressionSVM`	`predict`

Classification Model Object

Model Type	Full or Compact Classification Model Object	Function to Predict Labels and Scores
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`	`predict`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`	`predict`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`	`predict`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`	`predict`
Generalized additive model	`ClassificationGAM`, `CompactClassificationGAM`	`predict`
k-nearest neighbor model	`ClassificationKNN`	`predict`
Linear classification model	`ClassificationLinear`	`predict`
Naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`	`predict`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`	`predict`
Support vector machine for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	`predict`
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`	`predict`
Bagged ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`	`predict`

Alternative Functionality

- `plotPartialDependence` computes and plots partial dependence values. The function can also create individual conditional expectation on page 33-4659 (ICE) plots.

References

[1] Friedman, Jerome. H. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29, no. 5 (2001): 1189-1232.

[2] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. New York*, NY: Springer New York, 2009.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` name-value argument to `true` in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function fully supports GPU arrays for a regression or classification model specified as one of the following objects:
  - `LinearModel` object
  - `CompactLinearModel` object
  - `GeneralizedLinearModel` object
  - `CompactGeneralizedLinearModel` object
  - `ClassificationKNN` object
- The function also supports these objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`lime` | `oobPermutedPredictorImportance` | `plotPartialDependence` | `predictorImportance (RegressionEnsemble)` | `predictorImportance (RegressionTree)` | `relieff` | `sequentialfs` | `shapley`

Topics

“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

Introduced in R2020b

pca

Principal component analysis of raw data

Syntax

```
coeff = pca(X)
coeff = pca(X,Name,Value)
[coeff,score,latent] = pca(___)
[coeff,score,latent,tsquared] = pca(___)
[coeff,score,latent,tsquared,explained,mu] = pca(___)
```

Description

`coeff = pca(X)` returns the principal component coefficients, also known as loadings, for the *n*-by-*p* data matrix `X`. Rows of `X` correspond to observations and columns correspond to variables. The coefficient matrix is *p*-by-*p*. Each column of `coeff` contains coefficients for one principal component, and the columns are in descending order of component variance. By default, `pca` centers the data and uses the singular value decomposition (SVD) algorithm.

`coeff = pca(X,Name,Value)` returns any of the output arguments in the previous syntaxes using additional options for computation and handling of special data types, specified by one or more `Name,Value` pair arguments.

For example, you can specify the number of principal components `pca` returns or an algorithm other than SVD to use.

`[coeff,score,latent] = pca(___)` also returns the principal component scores in `score` and the principal component variances in `latent`. You can use any of the input arguments in the previous syntaxes.

Principal component scores are the representations of `X` in the principal component space. Rows of `score` correspond to observations, and columns correspond to components.

The principal component variances are the eigenvalues of the covariance matrix of `X`.

`[coeff,score,latent,tsquared] = pca(___)` also returns the Hotelling's T-squared statistic for each observation in `X`.

`[coeff,score,latent,tsquared,explained,mu] = pca(___)` also returns `explained`, the percentage of the total variance explained by each principal component and `mu`, the estimated mean of each variable in `X`.

Examples

Principal Components of a Data Set

Load the sample data set.

```
load hald
```

The ingredients data has 13 observations for 4 variables.

Find the principal components for the ingredients data.

```
 coeff = pca(ingredients)
```

```
coeff = 4×4

   -0.0678   -0.6460    0.5673    0.5062
   -0.6785   -0.0200   -0.5440    0.4933
    0.0290    0.7553    0.4036    0.5156
    0.7309   -0.1085   -0.4684    0.4844
```

The rows of `coeff` contain the coefficients for the four ingredient variables, and its columns correspond to four principal components.

PCA in the Presence of Missing Data

Find the principal component coefficients when there are missing values in a data set.

Load the sample data set.

```
load imports-85
```

Data matrix `X` has 13 continuous variables in columns 3 to 15: wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, and highway-mpg. The variables bore and stroke are missing four values in rows 56 to 59, and the variables horsepower and peak-rpm are missing two values in rows 131 and 132.

Perform principal component analysis.

```
coeff = pca(X(:,3:15));
```

By default, `pca` performs the action specified by the `'Rows','complete'` name-value pair argument. This option removes the observations with `NaN` values before calculation. Rows of `NaN`s are reinserted into `score` and `tsquared` at the corresponding locations, namely rows 56 to 59, 131, and 132.

Use `'pairwise'` to perform the principal component analysis.

```
coeff = pca(X(:,3:15),'Rows','pairwise');
```

In this case, `pca` computes the (*i*,*j*) element of the covariance matrix using the rows with no `NaN` values in the columns *i* or *j* of `X`. Note that the resulting covariance matrix might not be positive definite. This option applies when the algorithm `pca` uses is eigenvalue decomposition. When you don't specify the algorithm, as in this example, `pca` sets it to `'eig'`. If you require `'svd'` as the algorithm, with the `'pairwise'` option, then `pca` returns a warning message, sets the algorithm to `'eig'` and continues.

If you use the `'Rows','all'` name-value pair argument, `pca` terminates because this option assumes there are no missing values in the data set.

```
coeff = pca(X(:,3:15),'Rows','all');
```

```
Error using pca (line 180)
Raw data contains NaN missing value while 'Rows' option is set to 'all'. Consider using 'complete
```

Weighted PCA

Use the inverse variable variances as weights while performing the principal components analysis.

Load the sample data set.

```
load hald
```

Perform the principal component analysis using the inverse of variances of the ingredients as variable weights.

```
 [wcoeff,~,latent,~,explained] = pca(ingredients,...
'VariableWeights','variance')
```

```
wcoeff = 4×4

   -2.7998    2.9940   -3.9736    1.4180
   -8.7743   -6.4411    4.8927    9.9863
    2.5240   -3.8749   -4.0845    1.7196
    9.1714    7.5529    3.2710   11.3273


latent = 4×1

    2.2357
    1.5761
    0.1866
    0.0016


explained = 4×1

   55.8926
   39.4017
    4.6652
    0.0406
```

Note that the coefficient matrix, `wcoeff`, is not orthonormal.

Calculate the orthonormal coefficient matrix.

```
coefforth = inv(diag(std(ingredients)))* wcoeff
```

```
coefforth = 4×4

   -0.4760    0.5090   -0.6755    0.2411
   -0.5639   -0.4139    0.3144    0.6418
    0.3941   -0.6050   -0.6377    0.2685
    0.5479    0.4512    0.1954    0.6767
```

Check orthonormality of the new coefficient matrix, `coefforth`.

```
 coefforth*coefforth'
```

```
ans = 4×4

    1.0000    0.0000    0.0000         0
    0.0000    1.0000         0    0.0000
    0.0000         0    1.0000    0.0000
         0    0.0000    0.0000    1.0000
```

PCA Using ALS for Missing Data

Find the principal components using the alternating least squares (ALS) algorithm when there are missing values in the data.

Load the sample data.

```
load hald
```

The ingredients data has 13 observations for 4 variables.

Perform principal component analysis using the ALS algorithm and display the component coefficients.

```
[coeff,score,latent,tsquared,explained] = pca(ingredients);
coeff
```

```
coeff = 4×4

   -0.0678   -0.6460    0.5673    0.5062
   -0.6785   -0.0200   -0.5440    0.4933
    0.0290    0.7553    0.4036    0.5156
    0.7309   -0.1085   -0.4684    0.4844
```

Introduce missing values randomly.

```
y = ingredients;
rng('default'); % for reproducibility
ix = random('unif',0,1,size(y))<0.30;
y(ix) = NaN
```

```
y = 13×4

     7    26     6   NaN
     1    29    15    52
   NaN   NaN     8    20
    11    31   NaN    47
     7    52     6    33
   NaN    55   NaN   NaN
   NaN    71   NaN     6
     1    31   NaN    44
     2   NaN   NaN    22
    21    47     4    26
     ⋮
```

Approximately 30% of the data has missing values now, indicated by `NaN`.

Perform principal component analysis using the ALS algorithm and display the component coefficients.

```
[coeff1,score1,latent,tsquared,explained,mu1] = pca(y,...
'algorithm','als');
coeff1
```

```
coeff1 = 4×4

   -0.0362    0.8215   -0.5252    0.2190
   -0.6831   -0.0998    0.1828    0.6999
    0.0169    0.5575    0.8215   -0.1185
    0.7292   -0.0657    0.1261    0.6694
```

Display the estimated mean.

```
mu1
```

```
mu1 = 1×4

    8.9956   47.9088    9.0451   28.5515
```

Reconstruct the observed data.

```
t = score1*coeff1' + repmat(mu1,13,1)
```

```
t = 13×4

    7.0000   26.0000    6.0000   51.5250
    1.0000   29.0000   15.0000   52.0000
   10.7819   53.0230    8.0000   20.0000
   11.0000   31.0000   13.5500   47.0000
    7.0000   52.0000    6.0000   33.0000
   10.4818   55.0000    7.8328   17.9362
    3.0982   71.0000   11.9491    6.0000
    1.0000   31.0000   -0.5161   44.0000
    2.0000   53.7914    5.7710   22.0000
   21.0000   47.0000    4.0000   26.0000
       ⋮
```

The ALS algorithm estimates the missing values in the data.

Another way to compare the results is to find the angle between the two spaces spanned by the coefficient vectors. Find the angle between the coefficients found for complete data and data with missing values using ALS.

```
subspace(coeff,coeff1)
```

```
ans = 4.8688e-16
```

This is a small value. It indicates that the results if you use `pca` with `'Rows','complete'` name-value pair argument when there is no missing data and if you use `pca` with `'algorithm','als'` name-value pair argument when there is missing data are close to each other.

Perform the principal component analysis using `'Rows','complete'` name-value pair argument and display the component coefficients.

```
[coeff2,score2,latent,tsquared,explained,mu2] = pca(y,...
'Rows','complete');
coeff2
```

```
coeff2 = 4×3

   -0.2054    0.8587    0.0492
   -0.6694   -0.3720    0.5510
    0.1474   -0.3513   -0.5187
    0.6986   -0.0298    0.6518
```

In this case, `pca` removes the rows with missing values, and `y` has only four rows with no missing values. `pca` returns only three principal components. You cannot use the `'Rows','pairwise'` option because the covariance matrix is not positive semidefinite and `pca` returns an error message.

Find the angle between the coefficients found for complete data and data with missing values using listwise deletion (when `'Rows','complete'`).

```
subspace(coeff(:,1:3),coeff2)
```

```
ans = 0.3576
```

The angle between the two spaces is substantially larger. This indicates that these two results are different.

Display the estimated mean.

```
mu2
```

```
mu2 = 1×4

    7.8889   46.9091    9.8750   29.6000
```

In this case, the mean is just the sample mean of `y`.

Reconstruct the observed data.

```
score2*coeff2'
```

```
ans = 13×4

       NaN       NaN       NaN       NaN
   -7.5162  -18.3545    4.0968   22.0056
       NaN       NaN       NaN       NaN
       NaN       NaN       NaN       NaN
   -0.5644    5.3213   -3.3432    3.6040
       NaN       NaN       NaN       NaN
       NaN       NaN       NaN       NaN
       NaN       NaN       NaN       NaN
       NaN       NaN       NaN       NaN
   12.8315   -0.1076   -6.3333   -3.7758
      ⋮
```

This shows that deleting rows containing `NaN` values does not work as well as the ALS algorithm. Using ALS is better when the data has too many missing values.

Principal Component Coefficients, Scores, and Variances

Find the coefficients, scores, and variances of the principal components.

Load the sample data set.

```
load hald
```

The ingredients data has 13 observations for 4 variables.

Find the principal component coefficients, scores, and variances of the components for the ingredients data.

```
[coeff,score,latent] = pca(ingredients)
```

```
coeff = 4×4

   -0.0678   -0.6460    0.5673    0.5062
   -0.6785   -0.0200   -0.5440    0.4933
    0.0290    0.7553    0.4036    0.5156
    0.7309   -0.1085   -0.4684    0.4844


score = 13×4

   36.8218   -6.8709   -4.5909    0.3967
   29.6073    4.6109   -2.2476   -0.3958
  -12.9818   -4.2049    0.9022   -1.1261
   23.7147   -6.6341    1.8547   -0.3786
   -0.5532   -4.4617   -6.0874    0.1424
  -10.8125   -3.6466    0.9130   -0.1350
  -32.5882    8.9798   -1.6063    0.0818
   22.6064   10.7259    3.2365    0.3243
   -9.2626    8.9854   -0.0169   -0.5437
   -3.2840  -14.1573    7.0465    0.3405
      ⋮


latent = 4×1

  517.7969
   67.4964
   12.4054
    0.2372
```

Each column of `score` corresponds to one principal component. The vector, `latent`, stores the variances of the four principal components.

Reconstruct the centered ingredients data.

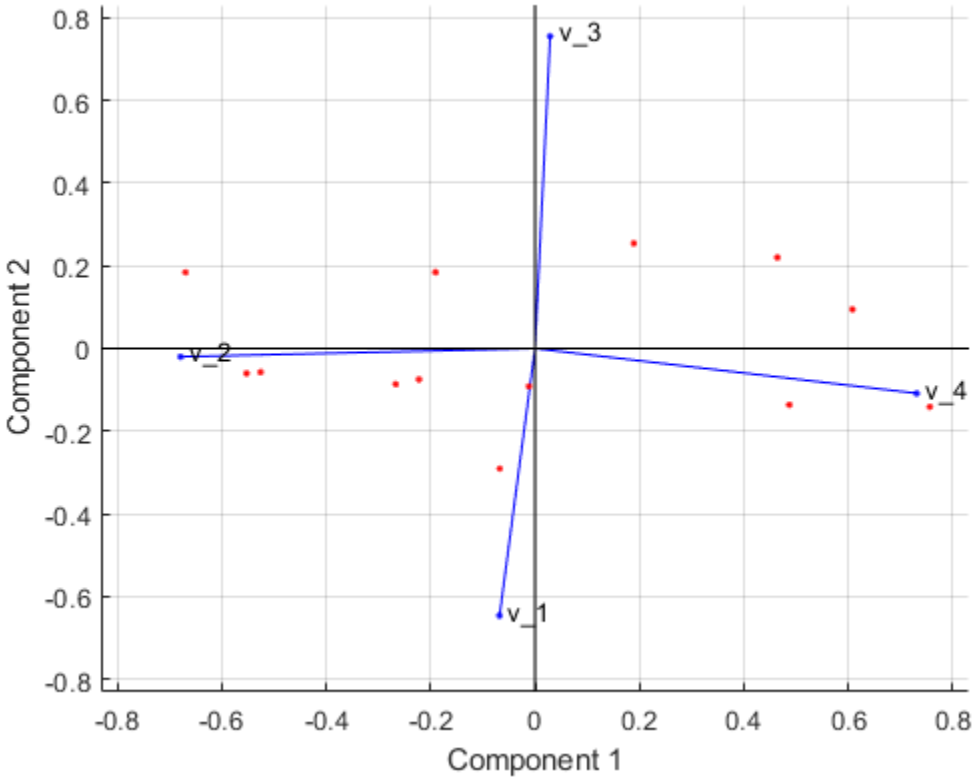
```
Xcentered = score*coeff'
```

```
Xcentered = 13×4

   -0.4615  -22.1538   -5.7692   30.0000
   -6.4615  -19.1538    3.2308   22.0000
    3.5385    7.8462   -3.7692  -10.0000
    3.5385  -17.1538   -3.7692   17.0000
   -0.4615    3.8462   -5.7692    3.0000
    3.5385    6.8462   -2.7692   -8.0000
   -4.4615   22.8462    5.2308  -24.0000
   -6.4615  -17.1538   10.2308   14.0000
   -5.4615    5.8462    6.2308   -8.0000
   13.5385   -1.1538   -7.7692   -4.0000
      ⋮
```

The new data in `Xcentered` is the original ingredients data centered by subtracting the column means from corresponding columns.

Visualize both the orthonormal principal component coefficients for each variable and the principal component scores for each observation in a single plot.

```
biplot(coeff(:,1:2),'scores',score(:,1:2),'varlabels',{'v_1','v_2','v_3','v_4'});
```



All four variables are represented in this biplot by a vector, and the direction and length of the vector indicate how each variable contributes to the two principal components in the plot. For example, the first principal component, which is on the horizontal axis, has positive coefficients for the third and

fourth variables. Therefore, vectors v_3 and v_4 are directed into the right half of the plot. The largest coefficient in the first principal component is the fourth, corresponding to the variable v_4.

The second principal component, which is on the vertical axis, has negative coefficients for the variables v_1, v_2, and v_4, and a positive coefficient for the variable v_3.

This 2-D biplot also includes a point for each of the 13 observations, with coordinates indicating the score of each observation for the two principal components in the plot. For example, points near the left edge of the plot have the lowest scores for the first principal component. The points are scaled with respect to the maximum score value and maximum coefficient length, so only their relative locations can be determined from the plot.

T-Squared Statistic

Find the Hotelling's T-squared statistic values.

Load the sample data set.

```
load hald
```

The ingredients data has 13 observations for 4 variables.

Perform the principal component analysis and request the T-squared values.

```
[coeff,score,latent,tsquared] = pca(ingredients);
tsquared
```

```
tsquared = 13×1

    5.6803
    3.0758
    6.0002
    2.6198
    3.3681
    0.5668
    3.4818
    3.9794
    2.6086
    7.4818
      ⋮
```

Request only the first two principal components and compute the T-squared values in the reduced space of requested principal components.

```
[coeff,score,latent,tsquared] = pca(ingredients,'NumComponents',2);
tsquared
```

```
tsquared = 13×1

    5.6803
    3.0758
    6.0002
    2.6198
```

```
    3.3681
    0.5668
    3.4818
    3.9794
    2.6086
    7.4818
      ⋮
```

Note that even when you specify a reduced component space, `pca` computes the T-squared values in the full space, using all four components.

The T-squared value in the reduced space corresponds to the Mahalanobis distance in the reduced space.

```
tsqreduced = mahal(score,score)
```

```
tsqreduced = 13×1

    3.3179
    2.0079
    0.5874
    1.7382
    0.2955
    0.4228
    3.2457
    2.6914
    1.3619
    2.9903
      ⋮
```

Calculate the T-squared values in the discarded space by taking the difference of the T-squared values in the full space and Mahalanobis distance in the reduced space.

```
tsqdiscarded = tsquared - tsqreduced
```

```
tsqdiscarded = 13×1

    2.3624
    1.0679
    5.4128
    0.8816
    3.0726
    0.1440
    0.2362
    1.2880
    1.2467
    4.4915
      ⋮
```

Percent Variability Explained by Principal Components

Find the percent variability explained by the principal components. Show the data representation in the principal components space.

Load the sample data set.

```
load imports-85
```

Data matrix X has 13 continuous variables in columns 3 to 15: wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, and highway-mpg.

Find the percent variability explained by principal components of these variables.

```
[coeff,score,latent,tsquared,explained] = pca(X(:,3:15));
```

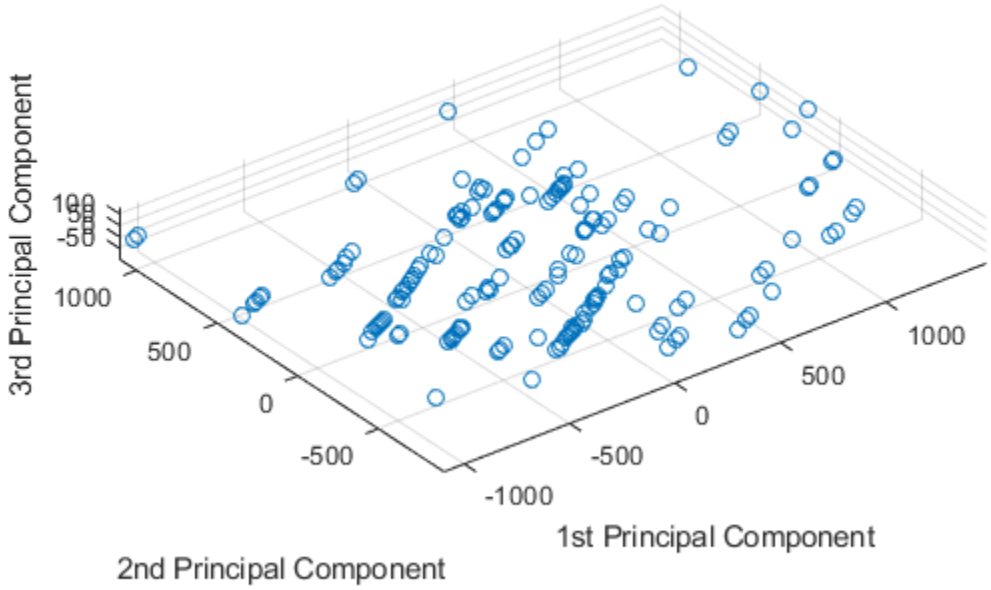
```
explained
```

```
explained = 13×1

   64.3429
   35.4484
    0.1550
    0.0379
    0.0078
    0.0048
    0.0013
    0.0011
    0.0005
    0.0002
      ⋮
```

The first three components explain 99.95% of all variability.

Visualize the data representation in the space of the first three principal components.

```
scatter3(score(:,1),score(:,2),score(:,3))
axis equal
xlabel('1st Principal Component')
ylabel('2nd Principal Component')
zlabel('3rd Principal Component')
```



The data shows the largest variability along the first principal component axis. This is the largest possible variance among all possible choices of the first axis. The variability along the second principal component axis is the largest among all possible remaining choices of the second axis. The third principal component axis has the third largest variability, which is significantly smaller than the variability along the second principal component axis. The fourth through thirteenth principal component axes are not worth inspecting, because they explain only 0.05% of all variability in the data.

To skip any of the outputs, you can use ~ instead in the corresponding element. For example, if you don't want to get the T-squared values, specify

```
[coeff,score,latent,~,explained] = pca(X(:,3:15));
```

Apply PCA to New Data and Generate C/C++ Code

Find the principal components for one data set and apply the PCA to another data set. This procedure is useful when you have a training data set and a test data set for a machine learning model. For example, you can preprocess the training data set by using PCA and then train a model. To test the trained model using the test data set, you need to apply the PCA transformation obtained from the training data to the test data set.

This example also describes how to generate C/C++ code. Because `pca` supports code generation, you can generate code that performs PCA using a training data set and applies the PCA to a test data

set. Then deploy the code to a device. In this workflow, you must pass training data, which can be of considerable size. To save memory on the device, you can separate training and prediction. Use `pca` in MATLAB® and apply PCA to new data in the generated code on the device.

Generating C/C++ code requires MATLAB® Coder™.

Apply PCA to New Data

Load the data set into a table by using `readtable`. The data set is in the file `CreditRating_Historical.dat`, which contains the historical credit rating data.

```
creditrating = readtable('CreditRating_Historical.dat');
creditrating(1:5,:)
```

```
ans=5×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    _______

    62394    0.013    0.104     0.036      0.447      0.142        3       {'BB' }
    48608    0.232    0.335     0.062      1.969      0.281        8       {'A'  }
    42444    0.311    0.367     0.074      1.935      0.366        1       {'A'  }
    48631    0.194    0.263     0.062      1.017      0.228        4       {'BBB'}
    43768    0.121    0.413     0.057      3.647      0.466       12       {'AAA'}
```

The first column is an ID of each observation, and the last column is a rating. Specify the second to seventh columns as predictor data and specify the last column (`Rating`) as the response.

```
X = table2array(creditrating(:,2:7));
Y = creditrating.Rating;
```

Use the first 100 observations as test data and the rest as training data.

```
XTest = X(1:100,:);
XTrain = X(101:end,:);
YTest = Y(1:100);
YTrain = Y(101:end);
```

Find the principal components for the training data set `XTrain`.

```
[coeff,scoreTrain,~,~,explained,mu] = pca(XTrain);
```

This code returns four outputs: `coeff`, `scoreTrain`, `explained`, and `mu`. Use `explained` (percentage of total variance explained) to find the number of components required to explain at least 95% variability. Use `coeff` (principal component coefficients) and `mu` (estimated means of `XTrain`) to apply the PCA to a test data set. Use `scoreTrain` (principal component scores) instead of `XTrain` when you train a model.

Display the percent variability explained by the principal components.

```
explained
```

```
explained = 6×1

   58.2614
   41.2606
    0.3875
    0.0632
```

```
    0.0269
    0.0005
```

The first two components explain more than 95% of all variability. Find the number of components required to explain at least 95% variability.

```
idx = find(cumsum(explained)>95,1)
```

```
idx = 2
```

Train a classification tree using the first two components.

```
scoreTrain95 = scoreTrain(:,1:idx);
mdl = fitctree(scoreTrain95,YTrain);
```

`mdl` is a `ClassificationTree` model.

To use the trained model for the test set, you need to transform the test data set by using the PCA obtained from the training data set. Obtain the principal component scores of the test data set by subtracting `mu` from `XTest` and multiplying by `coeff`. Only the scores for the first two components are necessary, so use the first two coefficients `coeff(:,1:idx)`.

```
scoreTest95 = (XTest-mu)*coeff(:,1:idx);
```

Pass the trained model `mdl` and the transformed test data set `scoreTest` to the `predict` function to predict ratings for the test set.

```
YTest_predicted = predict(mdl,scoreTest95);
```

Generate Code

Generate code that applies PCA to data and predicts ratings using the trained model. Note that generating C/C++ code requires MATLAB® Coder™.

Save the classification model to the file `myMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(mdl,'myMdl');
```

Define an entry-point function named `myPCAPredict` that accepts a test data set (`XTest`) and PCA information (`coeff` and `mu`) and returns the ratings of the test data.

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would cause errors during code generation.

```
function label = myPCAPredict(XTest,coeff,mu) %#codegen
% Transform data using PCA
scoreTest = bsxfun(@minus,XTest,mu)*coeff;

% Load trained classification model
mdl = loadLearnerForCoder('myMdl');
% Predict ratings using the loaded model
label = predict(mdl,scoreTest);
```

`myPCAPredict` applies PCA to new data using `coeff` and `mu`, and then predicts ratings using the transformed data. In this way, you do not pass training data, which can be of considerable size.

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate code by using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and exact input array size, pass a MATLAB® expression that represents the set of values with a certain data type and array size by using the `-args` option. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45.

```
codegen myPCAPredict -args {coder.typeof(XTest,[Inf,6],[1,0]),coeff(:,1:idx),mu}
```

```
Code generation successful.
```

`codegen` generates the MEX function `myPCAPredict_mex` with a platform-dependent extension.

Verify the generated code.

```
YTest_predicted_mex = myPCAPredict_mex(XTest,coeff(:,1:idx),mu);
isequal(YTest_predicted,YTest_predicted_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal. The comparison confirms that the `predict` function of `mdl` and the `myPCAPredict_mex` function return the same ratings.

For more information on code generation, see “Introduction to Code Generation” on page 32-2and “Code Generation and Classification Learner App” on page 32-31. The latter describes how to perform PCA and train a model by using the Classification Learner app, and how to generate C/C++ code that predicts labels for new data based on the trained model.

Input Arguments

X — Input data

matrix

Input data for which to compute the principal components, specified as an *n*-by-*p* matrix. Rows of `X` correspond to observations and columns to variables.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Algorithm','eig','Centered',false,'Rows','all','NumComponents',3` specifies that `pca` uses eigenvalue decomposition algorithm, not center the data, use all of the observations, and return only the first three principal components.

`Algorithm` — Principal component algorithm
`'svd'` (default) | `'eig'` | `'als'`

Principal component algorithm that `pca` uses to perform the principal component analysis, specified as the comma-separated pair consisting of `'Algorithm'` and one of the following.

Value	Description
`'svd'`	Default. Singular value decomposition (SVD) of `X`.
`'eig'`	Eigenvalue decomposition (EIG) of the covariance matrix. The EIG algorithm is faster than SVD when the number of observations, *n*, exceeds the number of variables, *p*, but is less accurate because the condition number of the covariance is the square of the condition number of `X`.
`'als'`	Alternating least squares (ALS) algorithm. This algorithm finds the best rank-*k* approximation by factoring `X` into a *n*-by-*k* left factor matrix, L, and a *p*-by-*k* right factor matrix, R, where *k* is the number of principal components. The factorization uses an iterative method starting with random initial values. ALS is designed to better handle missing values. It is preferable to pairwise deletion (`'Rows','pairwise'`) and deals with missing values without listwise deletion (`'Rows','complete'`). It can work well for data sets with a small percentage of missing data at random, but might not perform well on sparse data sets.

Example: `'Algorithm','eig'`

`Centered` — Indicator for centering columns
`true` (default) | `false`

Indicator for centering the columns, specified as the comma-separated pair consisting of `'Centered'` and one of these logical expressions.

Value	Description
`true`	Default. `pca` centers `X` by subtracting column means before computing singular value decomposition or eigenvalue decomposition. If `X` contains `NaN` missing values, `mean(X,'omitnan')` is used to find the mean with any available data. You can reconstruct the centered data using `score*coeff'`.
`false`	In this case `pca` does not center the data. You can reconstruct the original data using `score*coeff'`.

Example: `'Centered',false`

Data Types: `logical`

`Economy` — Indicator for economy size output
`true` (default) | `false`

Indicator for the economy size output when the degrees of freedom on page 33-4462, *d*, is smaller than the number of variables, *p*, specified as the comma-separated pair consisting of `'Economy'` and one of these logical expressions.

Value	Description
`true`	Default. `pca` returns only the first *d* elements of `latent` and the corresponding columns of `coeff` and `score`. This option can be significantly faster when the number of variables *p* is much larger than *d*.
`false`	`pca` returns all elements of `latent`. The columns of `coeff` and `score` corresponding to zero elements in `latent` are zeros.

Note that when $d < p$, `score(:,d+1:p)` and `latent(d+1:p)` are necessarily zero, and the columns of `coeff(:,d+1:p)` define directions that are orthogonal to `X`.

Example: `'Economy',false`

Data Types: `logical`

`NumComponents` — Number of components requested
number of variables (default) | scalar integer

Number of components requested, specified as the comma-separated pair consisting of `'NumComponents'` and a scalar integer *k* satisfying $0 < k \leq p$, where *p* is the number of original variables in `X`. When specified, `pca` returns the first *k* columns of `coeff` and `score`.

Example: `'NumComponents',3`

Data Types: `single` | `double`

`Rows` — Action to take for `NaN` values
`'complete'` (default) | `'pairwise'` | `'all'`

Action to take for `NaN` values in the data matrix `X`, specified as the comma-separated pair consisting of `'Rows'` and one of the following.

Value	Description
`'complete'`	Default. Observations with `NaN` values are removed before calculation. Rows of `NaN`s are reinserted into `score` and `tsquared` at the corresponding locations.
`'pairwise'`	This option only applies when the algorithm is `'eig'`. If you don't specify the algorithm along with `'pairwise'`, then `pca` sets it to `'eig'`. If you specify `'svd'` as the algorithm, along with the option `'Rows','pairwise'`, then `pca` returns a warning message, sets the algorithm to `'eig'` and continues. When you specify the `'Rows','pairwise'` option, `pca` computes the (*i*,*j*) element of the covariance matrix using the rows with no `NaN` values in the columns *i* or *j* of `X`. Note that the resulting covariance matrix might not be positive definite. In that case, `pca` terminates with an error message.
`'all'`	`X` is expected to have no missing values. `pca` uses all of the data and terminates if any `NaN` value is found.

Example: `'Rows','pairwise'`

`Weights` — Observation weights
ones (default) | row vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of length *n* containing all positive elements.

Data Types: `single` | `double`

`VariableWeights` — Variable weights
row vector | `'variance'`

Variable weights on page 33-4462, specified as the comma-separated pair consisting of `'VariableWeights'` and one of the following.

Value	Description
row vector	Vector of length *p* containing all positive elements.
`'variance'`	The variable weights are the inverse of sample variance. If you also assign weights to observations using `'Weights'`, then the variable weights become the inverse of weighted sample variance. If `'Centered'` is set to `true` at the same time, the data matrix `X` is centered and standardized. In this case, `pca` returns the principal components based on the correlation matrix.

Example: `'VariableWeights','variance'`

Data Types: `single` | `double` | `char` | `string`

`Coeff0` — Initial value for coefficients
matrix of random values (default) | *p*-by-*k* matrix

Initial value for the coefficient matrix `coeff`, specified as the comma-separated pair consisting of `'Coeff0'` and a *p*-by-*k* matrix, where *p* is the number of variables, and *k* is the number of principal components requested.

Note You can use this name-value pair only when `'algorithm'` is `'als'`.

Data Types: `single` | `double`

`Score0` — Initial value for scores
matrix of random values (default) | *k*-by-*m* matrix

Initial value for scores matrix `score`, specified as a comma-separated pair consisting of `'Score0'` and an *n*-by-*k* matrix, where *n* is the number of observations and *k* is the number of principal components requested.

Note You can use this name-value pair only when `'algorithm'` is `'als'`.

Data Types: `single` | `double`

`Options` — Options for iterations
structure

Options for the iterations, specified as a comma-separated pair consisting of `'Options'` and a structure created by the `statset` function. `pca` uses the following fields in the options structure.

Field Name	Description
'Display'	Level of display output. Choices are 'off', 'final', and 'iter'.
'MaxIter'	Maximum number steps allowed. The default is 1000. Unlike in optimization settings, reaching the MaxIter value is regarded as convergence.
'TolFun'	Positive number giving the termination tolerance for the cost function. The default is 1e-6.
'TolX'	Positive number giving the convergence threshold for the relative change in the elements of the left and right factor matrices, L and R, in the ALS algorithm. The default is 1e-6.

Note You can use this name-value pair only when 'algorithm' is 'als'.

You can change the values of these fields and specify the new structure in pca using the 'Options' name-value pair argument.

Example: `opt = statset('pca'); opt.MaxIter = 2000; coeff = pca(X,'Options',opt);`

Data Types: struct

Output Arguments

coeff — Principal component coefficients

matrix

Principal component coefficients, returned as a *p*-by-*p* matrix. Each column of coeff contains coefficients for one principal component. The columns are in the order of descending component variance, latent.

score — Principal component scores

matrix

Principal component scores, returned as a matrix. Rows of score correspond to observations, and columns to components.

latent — Principal component variances

column vector

Principal component variances, that is the eigenvalues of the covariance matrix of X, returned as a column vector.

tsquared — Hotelling's T-squared statistic

column vector

"Hotelling's T-Squared Statistic" on page 33-4462, which is the sum of squares of the standardized scores for each observation, returned as a column vector.

explained — Percentage of total variance explained

column vector

Percentage of the total variance explained by each principal component, returned as a column vector.

`mu` — Estimated means
row vector

Estimated means of the variables in `X`, returned as a row vector when `Centered` is set to `true`. When `Centered` is `false`, the software does not compute the means and returns a vector of zeros.

More About

Hotelling's T-Squared Statistic

Hotelling's T-squared statistic is a statistical measure of the multivariate distance of each observation from the center of the data set.

Even when you request fewer components than the number of variables, `pca` uses all principal components to compute the T-squared statistic (computes it in the full space). If you want the T-squared statistic in the reduced or the discarded space, do one of the following:

- For the T-squared statistic in the reduced space, use `mahal(score,score)`.
- For the T-squared statistic in the discarded space, first compute the T-squared statistic using `[coeff,score,latent,tsquared] = pca(X,'NumComponents',k,...)`, compute the T-squared statistic in the reduced space using `tsqreduced = mahal(score,score)`, and then take the difference: `tsquared - tsqreduced`.

Degrees of Freedom

The degrees of freedom, d, is equal to n – 1, if data is centered and n otherwise, where:

- n is the number of rows without any `NaN`s if you use `'Rows','complete'`.
- n is the number of rows without any `NaN`s in the column pair that has the maximum number of rows without `NaN`s if you use `'Rows','pairwise'`.

Variable Weights

Note that when variable weights are used, the coefficient matrix is not orthonormal. Suppose the variable weights vector you used is called `varwei`, and the principal component coefficients vector `pca` returned is `wcoeff`. You can then calculate the orthonormal coefficients using the transformation `diag(sqrt(varwei))*wcoeff`.

Algorithms

The `pca` function imposes a sign convention, forcing the element with the largest magnitude in each column of `coefs` to be positive. Changing the sign of a coefficient vector does not change its meaning.

References

[1] Jolliffe, I. T. *Principal Component Analysis*. 2nd ed., Springer, 2002.

[2] Krzanowski, W. J. *Principles of Multivariate Analysis*. Oxford University Press, 1988.

[3] Seber, G. A. F. *Multivariate Observations*. Wiley, 1984.

[4] Jackson, J. E. A. *User's Guide to Principal Components*. Wiley, 1988.

[5] Roweis, S. “EM Algorithms for PCA and SPCA.” *In Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*. Vol.10 (NIPS 1997), Cambridge, MA, USA: MIT Press, 1998, pp. 626–632.

[6] Ilin, A., and T. Raiko. “Practical Approaches to Principal Component Analysis in the Presence of Missing Values.” *J. Mach. Learn. Res.*. Vol. 11, August 2010, pp. 1957–2000.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with some limitations.

- `pca` works directly with tall arrays by computing the covariance matrix and using the in-memory `pcacov` function to compute the principle components.
- Supported syntaxes are:
  - `coeff = pca(X)`
  - `[coeff,score,latent] = pca(X)`
  - `[coeff,score,latent,explained] = pca(X)`
  - `[coeff,score,latent,tsquared] = pca(X)`
  - `[coeff,score,latent,tsquared,explained] = pca(X)`
- Name-value pair arguments are not supported.

For more information, see “Tall Arrays for Out-of-Memory Data”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- When `'Algorithm'` is `'als'`, the `'Display'` value for `'Options'` is ignored.
- The values for the `'Weights'` and `'VariableWeights'` name-value pair arguments must be real.
- The value for the `'Economy'` name-value pair argument must be a compile-time constant. For example, to use the `'Economy',false` name-value pair argument in the generated code, include `{coder.Constant('Economy'),coder.Constant(false)}` in the `-args` value of `codegen`.
- Names in name-value pair arguments must be compile-time constants.
- The generated code always returns the fifth output `explained` as a column vector.
- The generated code always returns the sixth output `mu` as a row vector.
- If `mu` is empty, `pca` returns `mu` as a 1-by-0 array. `pca` does not convert `mu` to a 0-by-0 empty array.
- The generated code does not treat an input matrix `X` that has all `NaN` values as a special case. The output dimensions are commensurate with corresponding finite inputs.
- To save memory on the device to which you deploy generated code, you can separate training (constructing PCA components from input data) and prediction (performing PCA transformation). Construct PCA components in MATLAB. Then, define an entry-point function that performs PCA

transformation using the principal component coefficients (`coeff`) and estimated means (`mu`), which are the outputs of `pca`. Finally, generate code for the entry-point function. For an example, see “Apply PCA to New Data and Generate C/C++ Code” on page 33-4454.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- You cannot specify the name-value argument `'Algorithm'` as `'als'`.
- The default principal component algorithm is SVD (name-value argument `'Algorithm','svd'`). The SVD algorithm is rarely faster when executed on a GPU instead of a CPU. Specify the name-value argument `'Algorithm'` as `'eig'` to speed up calculation on a GPU.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`barttest` | `biplot` | `canoncorr` | `factoran` | `pcacov` | `pcares` | `ppca` | `rotatefactors`

Topics

“Analyze Quality of Life in U.S. Cities Using PCA” on page 15-69
“Principal Component Analysis (PCA)” on page 15-68

Introduced in R2012b

pcacov

Principal component analysis on covariance matrix

Syntax

```
coeff = pcacov(V)
[coeff,latent] = pcacov(V)
[coeff,latent,explained] = pcacov(V)
```

Description

`coeff = pcacov(V)` performs principal component analysis on the square covariance matrix `V` and returns the principal component coefficients, also known as loadings.

`pcacov` does not standardize `V` to have unit variances. To perform principal component analysis on standardized variables, use the correlation matrix `R = V./(SD*SD')`, where `SD = sqrt(diag(V))`, in place of `V`. To perform principal component analysis directly on the data matrix, use `pca`.

`[coeff,latent] = pcacov(V)` also returns a vector containing the principal component variances, meaning the eigenvalues of `V`.

`[coeff,latent,explained] = pcacov(V)` also returns a vector containing the percentage of the total variance explained by each principal component.

Examples

Perform Principal Component Analysis on Covariance Matrix

Create a covariance matrix from the `hald` dataset.

```
load hald
covx = cov(ingredients);
```

Perform principal component analysis on the `covx` variable.

```
[coeff,latent,explained] = pcacov(covx)
```

```
coeff = 4×4

   -0.0678   -0.6460    0.5673    0.5062
   -0.6785   -0.0200   -0.5440    0.4933
    0.0290    0.7553    0.4036    0.5156
    0.7309   -0.1085   -0.4684    0.4844


latent = 4×1

  517.7969
   67.4964
   12.4054
```

```
    0.2372


explained = 4×1

   86.5974
   11.2882
    2.0747
    0.0397
```

The first component explains over 85% of the total variance. The first two components explain nearly 98% of the total variance.

Input Arguments

`V` — Covariance matrix

square, symmetric, positive semidefinite matrix

Covariance matrix, specified as a square, symmetric, positive semidefinite matrix.

Data Types: `single` | `double`

Output Arguments

`coeff` — Principal component coefficients

matrix

Principal component coefficients, returned as a matrix the same size as `V`. Each column of `coeff` contains coefficients for one principal component. The columns are in order of decreasing component variance.

`latent` — Principal component variances

vector

Principal component variances, returned as a vector with length equal to `size(coeff,1)`. The vector `latent` contains the eigenvalues of `V`.

`explained` — Percentage of total variance explained by each principal component

vector

Percentage of the total variance explained by each principal component, returned as a vector the same size as `latent`. The entries in `explained` range from 0 (none of the variance is explained) to 100 (all of the variance is explained).

References

[1] Jackson, J. E. *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley and Sons, 1991.

[2] Jolliffe, I. T. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag, 2002.

[3] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

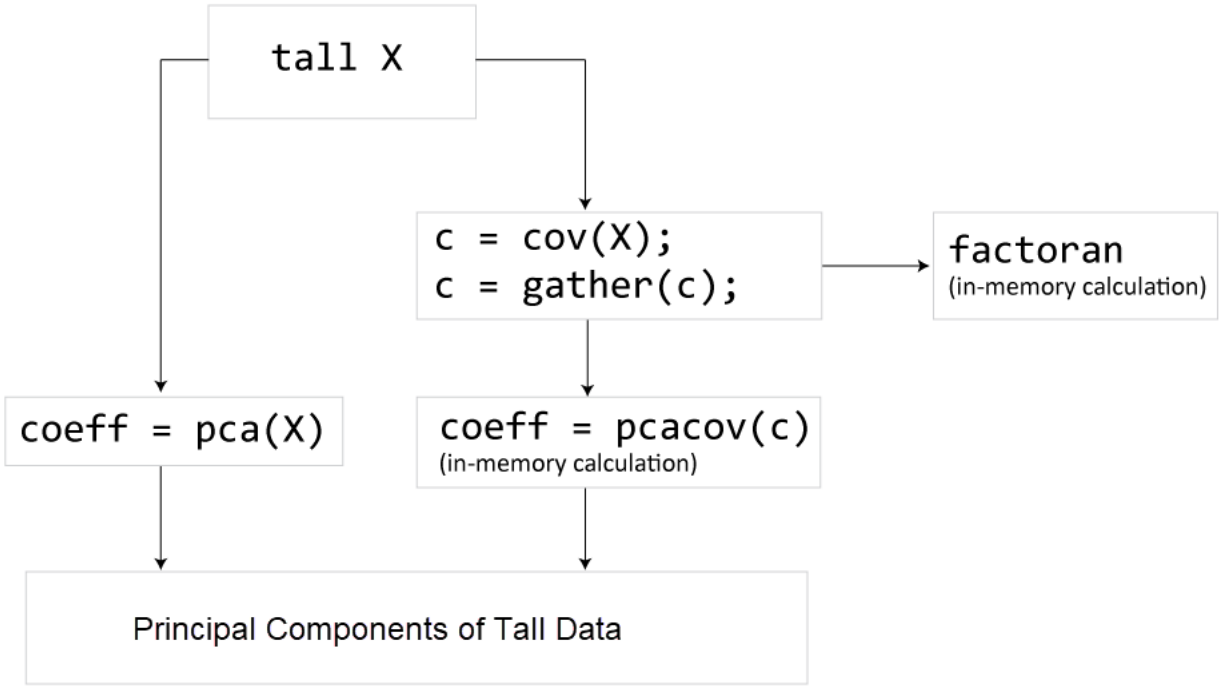
[4] Seber, G. A. F. *Multivariate Observations*, Wiley, 1984.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

`pcacov` and `factoran` do not work directly on tall arrays. Instead, use `C = gather(cov(X))` to compute the covariance matrix of a tall array. Then, you can use `pcacov` or `factoran` to work on the in-memory covariance matrix. Alternatively, you can use `pca` directly on a tall array.



For more information, see “Tall Arrays for Out-of-Memory Data”.

See Also

`barttest` | `biplot` | `factoran` | `pca` | `pcares`

Introduced before R2006a

pcares

Residuals from principal component analysis

Syntax

```
residuals = pcares(X,ndim)
[residuals,reconstructed] = pcares(X,ndim)
```

Description

`residuals = pcares(X,ndim)` returns the `residuals` obtained by retaining `ndim` principal components of the n-by-p matrix `X`. Rows of `X` correspond to observations, columns to variables. `ndim` is a scalar and must be less than or equal to p. `residuals` is a matrix of the same size as `X`. Use the data matrix, *not* the covariance matrix, with this function.

`pcares` does not normalize the columns of X. To perform the principal components analysis based on standardized variables, that is, based on correlations, use `pcares(zscore(X), ndim)`. You can perform principal components analysis directly on a covariance or correlation matrix, but without constructing residuals, by using `pcacov`.

`[residuals,reconstructed] = pcares(X,ndim)` returns the reconstructed observations; that is, the approximation to `X` obtained by retaining its first `ndim` principal components.

Examples

This example shows the drop in the residuals from the first row of the Hald data as the number of component dimensions increases from one to three.

```
load hald
r1 = pcares(ingredients,1);
r2 = pcares(ingredients,2);
r3 = pcares(ingredients,3);

r11 = r1(1,:)
r11 =
  2.0350  2.8304  -6.8378  3.0879

r21 = r2(1,:)
r21 =
  -2.4037  2.6930  -1.6482  2.3425

r31 = r3(1,:)
r31 =
  0.2008  0.1957  0.2045  0.1921
```

References

[1] Jackson, J. E., *A User's Guide to Principal Components*, John Wiley and Sons, 1991.

[2] Jolliffe, I. T., *Principal Component Analysis*, 2nd Edition, Springer, 2002.

[3] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[4] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

See Also

`factoran` | `pca` | `pcacov`

Introduced before R2006a

ppca

Probabilistic principal component analysis

Syntax

```
[coeff,score,pcvar] = ppca(Y,K)
[coeff,score,pcvar] = ppca(Y,K,Name,Value)
[coeff,score,pcvar,mu] = ppca( ___ )
[coeff,score,pcvar,mu,v,S] = ppca( ___ )
```

Description

`[coeff,score,pcvar] = ppca(Y,K)` returns the principal component coefficients for the *n*-by-*p* data matrix `Y` based on a probabilistic principal component analysis on page 33-4476 (PPCA). It also returns the principal component scores, which are the representations of `Y` in the principal component space, and the principal component variances, which are the eigenvalues of the covariance matrix of `Y`, in `pcvar`.

Each column of `coeff` contains coefficients for one principal component, and the columns are in descending order of component variance. Rows of `score` correspond to observations, and columns correspond to components. Rows of `Y` correspond to observations and columns correspond to variables.

Probabilistic principal component analysis might be preferable to other algorithms that handle missing data, such as the alternating least squares algorithm when any data vector has one or more missing values. It assumes that the values are missing at random through the data set. An expectation-maximization algorithm is used for both complete and missing data.

`[coeff,score,pcvar] = ppca(Y,K,Name,Value)` returns the principal component coefficients, scores, and variances using additional options for computation and handling of special data types, specified by one or more `Name,Value` pair arguments.

For example, you can introduce initial values for the residual variance, `v`, or change the termination criteria.

`[coeff,score,pcvar,mu] = ppca( ___ )` also returns the estimated mean of each variable in `Y`. You can use any of the input arguments in the previous syntaxes.

`[coeff,score,pcvar,mu,v,S] = ppca( ___ )` also returns the isotropic residual variance in `v` and the final results at convergence in structure `S`.

Examples

Perform Probabilistic Principal Component Analysis

Load the sample data.

```
load fisheriris
```

The double matrix `meas` consists of four types of measurements on the flowers, which, respectively, are the length and width of sepals and petals.

Introduce missing values randomly.

```
y = meas;
rng('default'); % for reproducibility
ix = random('unif',0,1,size(y))<0.20;
y(ix) = NaN;
```

Now, approximately 20% of the data is missing, indicated by `NaN`.

Perform probabilistic principal component analysis and request the component coefficients and variances.

```
[coeff,score,pcvar,mu] = ppca(y,3);
coeff
```

```
coeff = 4×3

    0.3562    0.6709   -0.5518
   -0.0765    0.7120    0.6332
    0.8592   -0.1597    0.0596
    0.3592   -0.1318    0.5395
```

```
pcvar
```

```
pcvar = 3×1

    4.0914
    0.2125
    0.0617
```

Perform principal component analysis using the alternating least squares algorithm and request the component coefficients and variances.

```
[coeff2,score2,pcvar2,mu2] = pca(y,'algorithm','als',...
'NumComponents',3);
coeff2
```

```
coeff2 = 4×3

    0.3376    0.4952    0.7406
   -0.0731    0.8609   -0.4476
    0.8657   -0.1168   -0.1233
    0.3623   -0.0086   -0.4857
```

```
pcvar2
```

```
pcvar2 = 3×1

    4.0733
    0.2652
    0.1222
```

The coefficients and the variances of the first two principal components are similar.

Another way to compare the results is to find the angle between the two spaces spanned by the coefficient vectors.

```
subspace(coeff,coeff2)
```

```
ans = 0.0884
```

The angle between the two spaces is pretty small. This indicates that these two results are close to each other.

Change the Termination Criteria for Probabilistic Principal Component Analysis

Load the sample data set.

```
load imports-85
```

Data matrix X has 13 continuous variables in columns 3 to 15: wheel-base, length, width, height, curb-weight, engine-size, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, and highway-mpg. The variables bore and stroke are missing four values in rows 56 to 59, and the variables horsepower and peak-rpm are missing two values in rows 131 and 132.

Perform probabilistic principal component analysis and display the first three principal components.

```
[coeff,score,pcvar] = ppca(X(:,3:15),3);
```

```
Warning: Maximum number of iterations 1000 reached.
```

Change the termination tolerance for the cost function to 0.01.

```
opt = statset('ppca');
opt.TolFun = 0.01;
```

Perform probabilistic principal component analysis.

```
[coeff,score,pcvar] = ppca(X(:,3:15),3,'Options',opt);
```

```
Warning: Maximum number of iterations 1000 reached.
```

`ppca` now terminates before the maximum number of iterations is reached because it meets the tolerance for the cost function.

Reconstruct Observations

Load the sample data.

```
load hald
y = ingredients;
```

The ingredients data has 13 observations for 4 variables.

Introduce missing values to the data.

```
y(2:16:end) = NaN;
```

Every 16th value is `NaN`. This corresponds to 7.69% of the data.

Find the first three principal components of data using PPCA and display the reconstructed observations.

```
[coeff,score,pcvar,mu,v,S] = ppca(y,3);
```

```
Warning: Maximum number of iterations 1000 reached.
```

```
S.Recon
```

```
ans = 13×4

    6.8536   25.8700    5.8389   59.8730
    1.0433   28.9710   14.9654   51.9738
   11.5770   56.5067    8.6352   20.5076
   11.0835   31.0722    8.0920   47.0748
    7.0679   52.2556    6.0748   33.0598
   11.0486   55.0430    9.0534   22.0423
    2.8493   70.8691   16.8339    5.8656
    1.0333   31.0281   19.6907   44.0306
    2.0400   54.0354   18.0440   22.0349
   20.7822   46.8091    3.7603   25.8081
       ⋮
```

You can also reconstruct the observations using the principal components and the estimated mean.

```
t = score*coeff' + repmat(mu,13,1);
```

Results at Convergence

Load the data.

```
load hald
```

Here, `ingredients` is a real-valued matrix of predictor variables.

Perform the probabilistic principal components analysis and display coefficients.

```
[coeff,score,pcvariance,mu,v,S] = ppca(ingredients,3);
```

```
Warning: Maximum number of iterations 1000 reached.
```

```
coeff
```

```
coeff = 4×3

   -0.0693   -0.6459    0.5673
   -0.6786   -0.0184   -0.5440
    0.0308    0.7552    0.4036
    0.7306   -0.1102   -0.4684
```

Display the algorithm results at convergence of the PPCA.

```
S
```

```
S = struct with fields:
           W: [4x3 double]
        Xexp: [13x3 double]
       Recon: [13x4 double]
           v: 0.2372
     NumIter: 1000
    RMSResid: 0.2340
      nloglk: 149.3388
```

Display the matrix `W`.

```
S.W
```

```
ans = 4×3

    0.5624    2.0279    5.4075
    4.8320  -10.3894    5.9202
   -3.7521   -3.0555   -4.1552
   -1.5144   11.7122   -7.2564
```

Orthogonalizing `W` recovers the coefficients.

```
orth(S.W)
```

```
ans = 4×3

   -0.0693    0.6459    0.5673
   -0.6786    0.0184   -0.5440
    0.0308   -0.7552    0.4036
    0.7306    0.1102   -0.4684
```

Input Arguments

`Y` — Input data

n-by-*p* matrix

Input data for which to compute the principal components, specified as an *n*-by-*p* matrix. Rows of `Y` correspond to observations and columns correspond to variables.

Data Types: `single` | `double`

`K` — Number of principal components

positive integer value less than rank

Number of principal components to return, specified as an integer value less than the rank of data. The maximum possible rank is min(*n*,*p*), where *n* is the number of observations and *p* is the number of variables. However, if the data is correlated, the rank might be smaller than min(*n*,*p*).

`ppca` orders the components based on their variance.

If `K` is min(*n*,*p*), `ppca` sets `K` equal to min(*n*,*p*) – 1, and `'W0'` is truncated to min(*p*,*n*) – 1 columns if you specify a *p*-by-*p* `W0` matrix.

For example, you can request only the first three components, based on the component variance as follows.

Example: `coeff = ppca(Y,3)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'W0',init,'Options',opt` specifies that the initial values for `'W0'` are in matrix `init` and `ppca` uses the options defined by `opt`.

`W0` — Initial value of *W*

matrix of random values (default) | *p*-by-*k* matrix

Initial value of *W* in the probabilistic principal component analysis on page 33-4476 algorithm, specified as a comma-separated pair consisting of `'W0'` and a *p*-by-*k* matrix.

Data Types: `single` | `double`

`v0` — Initial value of residual variance

random number (default) | positive scalar value

Initial value of residual variance, specified as the comma-separated pair consisting of `'v0'` and a positive scalar value.

Data Types: `single` | `double`

`Options` — Options for iterations

structure

Options for the iterations, specified as a comma-separated pair `'Options'` and a structure created by the `statset` function. `ppca` uses the following fields in the options structure.

`'Display'`	Level of display output. Choices are `'off'`, `'final'`, and `'iter'`.
`'MaxIter'`	Maximum number of steps allowed. The default is 1000. Unlike in optimization settings, reaching the `MaxIter` value is regarded as convergence.
`'TolFun'`	Positive integer stating the termination tolerance for the cost function. The default is 1e-6.
`'TolX'`	Positive integer stating the convergence threshold for the relative change in the elements of *W*. The default is 1e-6.

You can change the values of these fields and specify the new structure in `ppca` using the `'Options'` name-value pair argument.

Example: `opt = statset('ppca'); opt.MaxIter = 2000; coeff = ppca(Y,3,'Options',opt);`

Data Types: `struct`

Output Arguments

`coeff` — Principal component coefficients

p-by-*k* matrix

Principal component coefficients, returned as a *p*-by-*k* matrix. Each column of `coeff` contains coefficients for one principal component. The columns are in the order of descending component variance, `pcvar`.

`score` — Principal component scores

n-by-*k* matrix

Principal component scores, returned as an *n*-by-*k* matrix. Rows of `score` correspond to observations, and columns correspond to components.

`pcvar` — Principal component variances

column vector

Principal component variances, which are the eigenvalues of the covariance matrix of `Y`, returned as a column vector.

`mu` — Estimated mean

row vector

Estimated mean of each variable in `Y`, returned as a row vector.

`v` — Isotropic residual variance

scalar value

Isotropic residual variance, returned as a scalar value.

`S` — Final results at convergence

structure

Final results at convergence, returned as a structure containing the following fields.

`W`	*W* at convergence.
`Xexp`	Conditional expectation of the estimated latent variable *x*.
`Recon`	Reconstructed observations using *k* principal components. This is a low dimension approximation of the input data `Y`, and is equal to `mu + score*coeff'`.
`v`	Residual variance.
`RMSResid`	Root mean square of residuals.
`NumIter`	Number of iteration counts.
`nloglk`	Negative loglikelihood function value.

More About

Probabilistic Principal Component Analysis

Probabilistic principal component analysis (PPCA) is a method to estimate the principal axes when any data vector has one or more missing values.

PPCA is based on an isotropic error model. It seeks to relate a p-dimensional observation vector y to a corresponding k-dimensional vector of latent (or unobserved) variable x, which is normal with mean zero and covariance I(k). The relationship is

$$y^T = W * x^T + \mu + \varepsilon,$$

where y is the row vector of observed variable, x is the row vector of latent variables, and ε is the isotropic error term. ε is Gaussian with mean zero and covariance of v*I(k), where v is the residual variance. Here, k needs to be smaller than the rank for the residual variance to be greater than 0 (v>0). Standard principal component analysis, where the residual variance is zero, is the limiting case of PPCA. The observed variables, y, are conditionally independent given the values of the latent variables, x. So, the latent variables explain the correlations between the observation variables and the error explains the variability unique to a particular y_i. The p-by-k matrix W relates the latent and observation variables, and the vector μ permits the model to have a nonzero mean. PPCA assumes that the values are missing at random through the data set. This means that whether a data value is missing or not does not depend on the latent variable given the observed data values.

Under this model,

$$y \sim N\left(\mu, W * W^T + v * I(k)\right).$$

There is no closed-form analytical solution for W and v, so their estimates are determined by iterative maximization of the corresponding loglikelihood using an expectation-maximization (EM) algorithm. This EM algorithm handles missing values by treating them as additional latent variables. At convergence, the columns of W spans the subspace, but they are not orthonormal. `ppca` obtains the orthonormal coefficients, `coeff`, for the components by orthogonalization of W.

References

[1] Tipping, M. E., and C. M. Bishop. Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 61, No.3, 1999, pp. 611–622.

[2] Roweis, S. “EM Algorithms for PCA and SPCA.” *In Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*. Vol.10 (NIPS 1997), Cambridge, MA, USA: MIT Press, 1998, pp. 626–632.

[3] Ilin, A., and T. Raiko. “Practical Approaches to Principal Component Analysis in the Presence of Missing Values.” *J. Mach. Learn. Res.*. Vol. 11, August, 2010, pp. 1957–2000.

See Also

`barttest` | `biplot` | `canoncorr` | `factoran` | `pca` | `pcacov` | `pcares` | `rotatefactors`

Introduced in R2013a

pdf

Package: prob

Probability density function

Syntax

```
y = pdf('name',x,A)
y = pdf('name',x,A,B)
y = pdf('name',x,A,B,C)
y = pdf('name',x,A,B,C,D)

y = pdf(pd,x)
```

Description

`y = pdf('name',x,A)` returns the probability density function (pdf) for the one-parameter distribution family specified by `'name'` and the distribution parameter `A`, evaluated at the values in `x`.

`y = pdf('name',x,A,B)` returns the pdf for the two-parameter distribution family specified by `'name'` and the distribution parameters `A` and `B`, evaluated at the values in `x`.

`y = pdf('name',x,A,B,C)` returns the pdf for the three-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, and `C`, evaluated at the values in `x`.

`y = pdf('name',x,A,B,C,D)` returns the pdf for the four-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, `C`, and `D`, evaluated at the values in `x`.

`y = pdf(pd,x)` returns the pdf of the probability distribution object `pd`, evaluated at the values in `x`.

Examples

Compute the Normal Distribution pdf

Create a standard normal distribution object with the mean μ equal to 0 and the standard deviation σ equal to 1.

```
mu = 0;
sigma = 1;
pd = makedist('Normal','mu',mu,'sigma',sigma);
```

Define the input vector x to contain the values at which to calculate the pdf.

```
x = [-2 -1 0 1 2];
```

Compute the pdf values for the standard normal distribution at the values in x.

```
y = pdf(pd,x)
```

```
y = 1×5

    0.0540    0.2420    0.3989    0.2420    0.0540
```

Each value in y corresponds to a value in the input vector x. For example, at the value x equal to 1, the corresponding pdf value y is equal to 0.2420.

Alternatively, you can compute the same pdf values without creating a probability distribution object. Use the `pdf` function, and specify a standard normal distribution using the same parameter values for μ and σ.

```
y2 = pdf('Normal',x,mu,sigma)
```

```
y2 = 1×5

    0.0540    0.2420    0.3989    0.2420    0.0540
```

The pdf values are the same as those computed using the probability distribution object.

Compute the Poisson Distribution pdf

Create a Poisson distribution object with the rate parameter, λ, equal to 2.

```
lambda = 2;
pd = makedist('Poisson','lambda',lambda);
```

Define the input vector x to contain the values at which to calculate the pdf.

```
x = [0 1 2 3 4];
```

Compute the pdf values for the Poisson distribution at the values in x.

```
y = pdf(pd,x)
```

```
y = 1×5

    0.1353    0.2707    0.2707    0.1804    0.0902
```

Each value in y corresponds to a value in the input vector x. For example, at the value x equal to 3, the corresponding pdf value in y is equal to 0.1804.

Alternatively, you can compute the same pdf values without creating a probability distribution object. Use the `pdf` function, and specify a Poisson distribution using the same value for the rate parameter, λ.

```
y2 = pdf('Poisson',x,lambda)
```

```
y2 = 1×5

    0.1353    0.2707    0.2707    0.1804    0.0902
```

The pdf values are the same as those computed using the probability distribution object.

Plot the pdf of a Standard Normal Distribution

Create a standard normal distribution object.

```
pd = makedist('Normal')
```

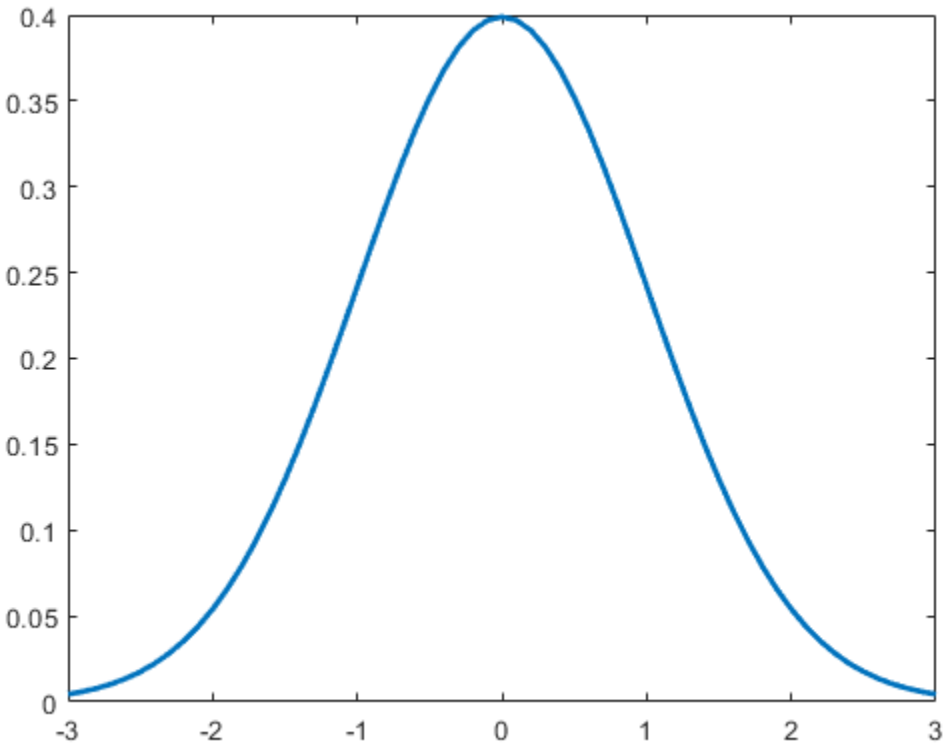
```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Specify the x values and compute the pdf.

```
x = -3:.1:3;
pdf_normal = pdf(pd,x);
```

Plot the pdf.

```
plot(x,pdf_normal,'LineWidth',2)
```



Plot the pdf of a Weibull Distribution

Create a Weibull probability distribution object.

```
pd = makedist('Weibull','a',5,'b',2)
```

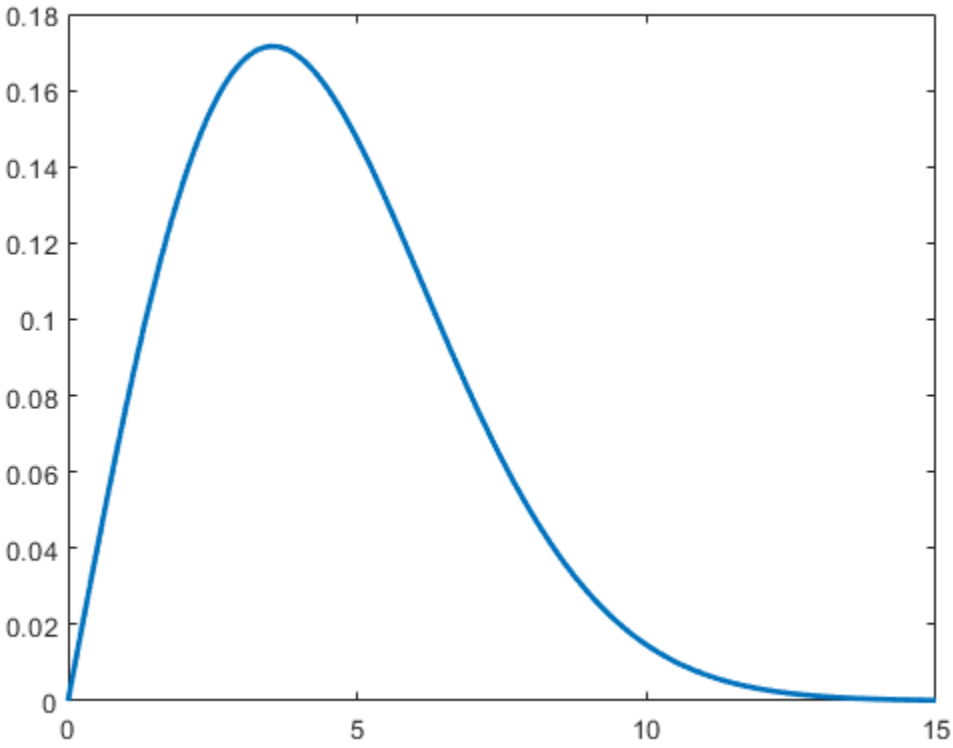
```
pd =
  WeibullDistribution

  Weibull distribution
    A = 5
    B = 2
```

Specify the x values and compute the pdf.

```
x = 0:.1:15;
y = pdf(pd,x);
```

Plot the pdf.

```
plot(x,y,'LineWidth',2)
```



Input Arguments

`'name'` — Probability distribution name

character vector or string scalar of probability distribution name

Probability distribution name, specified as one of the probability distribution names in this table.

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Beta'	“Beta Distribution” on page B-6	a first shape parameter	b second shape parameter	—	—
'Binomial'	“Binomial Distribution” on page B-10	n number of trials	p probability of success for each trial	—	—
'BirnbaumSaunders'	“Birnbaum-Saunders Distribution” on page B-18	β scale parameter	γ shape parameter	—	—
'Burr'	“Burr Type XII Distribution” on page B-19	α scale parameter	c first shape parameter	k second shape parameter	—
'Chisquare'	“Chi-Square Distribution” on page B-28	ν degrees of freedom	—	—	—
'Exponential'	“Exponential Distribution” on page B-33	μ mean	—	—	—
'Extreme Value'	“Extreme Value Distribution” on page B-40	μ location parameter	σ scale parameter	—	—
'F'	“F Distribution” on page B-45	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	—	—
'Gamma'	“Gamma Distribution” on page B-47	a shape parameter	b scale parameter	—	—
'Generalized Extreme Value'	“Generalized Extreme Value Distribution” on page B-55	k shape parameter	σ scale parameter	μ location parameter	—
'Generalized Pareto'	“Generalized Pareto Distribution” on page B-59	k tail index (shape) parameter	σ scale parameter	μ threshold (location) parameter	—
'Geometric'	“Geometric Distribution” on page B-63	p probability parameter	—	—	—
'HalfNormal'	“Half-Normal Distribution” on page B-68	μ location parameter	σ scale parameter	—	—
'Hypergeometric'	“Hypergeometric Distribution” on page B-73	m size of the population	k number of items with the desired characteristic in the population	n number of samples drawn	—
'InverseGaussian'	“Inverse Gaussian Distribution” on page B-75	μ scale parameter	λ shape parameter	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Logistic'	"Logistic Distribution" on page B-85	μ mean	σ scale parameter	—	—
'LogLogistic'	"Loglogistic Distribution" on page B-86	μ mean of logarithmic values	σ scale parameter of logarithmic values	—	—
'Lognormal'	"Lognormal Distribution" on page B-88	μ mean of logarithmic values	σ standard deviation of logarithmic values	—	—
'Nakagami'	"Nakagami Distribution" on page B-108	μ shape parameter	ω scale parameter	—	—
'Negative Binomial'	"Negative Binomial Distribution" on page B-109	r number of successes	p probability of success in a single trial	—	—
'Noncentral F'	"Noncentral F Distribution" on page B-115	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	δ noncentrality parameter	—
'Noncentral t'	"Noncentral t Distribution" on page B-117	ν degrees of freedom	δ noncentrality parameter	—	—
'Noncentral Chi-square'	"Noncentral Chi-Square Distribution" on page B-113	ν degrees of freedom	δ noncentrality parameter	—	—
'Normal'	"Normal Distribution" on page B-119	μ mean	σ standard deviation	—	—
'Poisson'	"Poisson Distribution" on page B-131	λ mean	—	—	—
'Rayleigh'	"Rayleigh Distribution" on page B-137	b scale parameter	—	—	—
'Rician'	"Rician Distribution" on page B-139	s noncentrality parameter	σ scale parameter	—	—
'Stable'	"Stable Distribution" on page B-140	α first shape parameter	β second shape parameter	γ scale parameter	δ location parameter
'T'	"Student's t Distribution" on page B-149	ν degrees of freedom	—	—	—
'tLocationScale'	"t Location-Scale Distribution" on page B-156	μ location parameter	σ scale parameter	ν shape parameter	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Uniform'	"Uniform Distribution (Continuous)" on page B-163	*a* lower endpoint (minimum)	*b* upper endpoint (maximum)	—	—
'Discrete Uniform'	"Uniform Distribution (Discrete)" on page B-168	*n* maximum observable value	—	—	—
'Weibull'	"Weibull Distribution" on page B-170	*a* scale parameter	*b* shape parameter	—	—

Example: 'Normal'

x — Values at which to evaluate pdf
scalar value | array of scalar values

Values at which to evaluate the pdf, specified as a scalar value or an array of scalar values.

If one or more of the input arguments x, A, B, C, and D are arrays, then the array sizes must be the same. In this case, pdf expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Example: [-1,0,3,4]

Data Types: single | double

A — First probability distribution parameter
scalar value | array of scalar values

First probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments x, A, B, C, and D are arrays, then the array sizes must be the same. In this case, pdf expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Data Types: single | double

B — Second probability distribution parameter
scalar value | array of scalar values

Second probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments x, A, B, C, and D are arrays, then the array sizes must be the same. In this case, pdf expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Data Types: single | double

C — Third probability distribution parameter
scalar value | array of scalar values

Third probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `pdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`D` — Fourth probability distribution parameter
scalar value | array of scalar values

Fourth probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `x`, `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `pdf` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`pd` — Probability distribution
probability distribution object

Probability distribution, specified as a probability distribution object created with a function or app in this table.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.
`paretotails`	Create a piecewise distribution object that has generalized Pareto distributions in the tails.

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values, returned as a scalar value or an array of scalar values. `y` is the same size as `x` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution, specified by the corresponding elements in the distribution parameters (`A`, `B`, `C`, and `D`) or specified by the probability distribution object (`pd`), evaluated at the corresponding element in `x`.

Alternative Functionality

- `pdf` is a generic function that accepts either a distribution by its name `'name'` or a probability distribution object `pd`. It is faster to use a distribution-specific function, such as `normpdf` for the normal distribution and `binopdf` for the binomial distribution. For a list of distribution-specific functions, see “Supported Distributions” on page 5-14.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `'name'` must be a compile-time constant. For example, to use the normal distribution, include `coder.Constant('Normal')` in the `-args` value of `codegen`.
- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `cdf` | `fitdist` | `icdf` | `makedist` | `mle` | `paretotails` | `random`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

pdf

Probability density function for Gaussian mixture distribution

Syntax

```
y = pdf(gm,X)
```

Description

`y = pdf(gm,X)` returns the probability density function (pdf) of the Gaussian mixture distribution `gm`, evaluated at the values in `X`.

Examples

Compute pdf Values

Create a `gmdistribution` object and compute its pdf values.

Define the distribution parameters (means and covariances) of a two-component bivariate Gaussian mixture distribution.

```
mu = [1 2;-3 -5];
sigma = [1 1]; % shared diagonal covariance matrix
```

Create a `gmdistribution` object by using the `gmdistribution` function. By default, the function creates an equal proportion mixture.

```
gm = gmdistribution(mu,sigma)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:     1     2

Component 2:
Mixing proportion: 0.500000
Mean:    -3    -5
```

Compute the pdf values of `gm`.

```
X = [0 0;1 2;3 3;5 3];
pdf(gm,X)
```

```
ans = 4×1

    0.0065
    0.0796
    0.0065
    0.0000
```

Plot pdf

Create a `gmdistribution` object and plot its pdf.

Define the distribution parameters (means, covariances, and mixing proportions) of two bivariate Gaussian mixture components.

```
p = [0.4 0.6];               % Mixing proportions
mu = [1 2;-3 -5];            % Means
sigma = cat(3,[2 .5],[1 1])  % Covariances 1-by-2-by-2 array
```

```
sigma =
sigma(:,:,1) =

    2.0000    0.5000


sigma(:,:,2) =

     1     1
```

The `cat` function concatenates the covariances along the third array dimension. The defined covariance matrices are diagonal matrices. `sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.

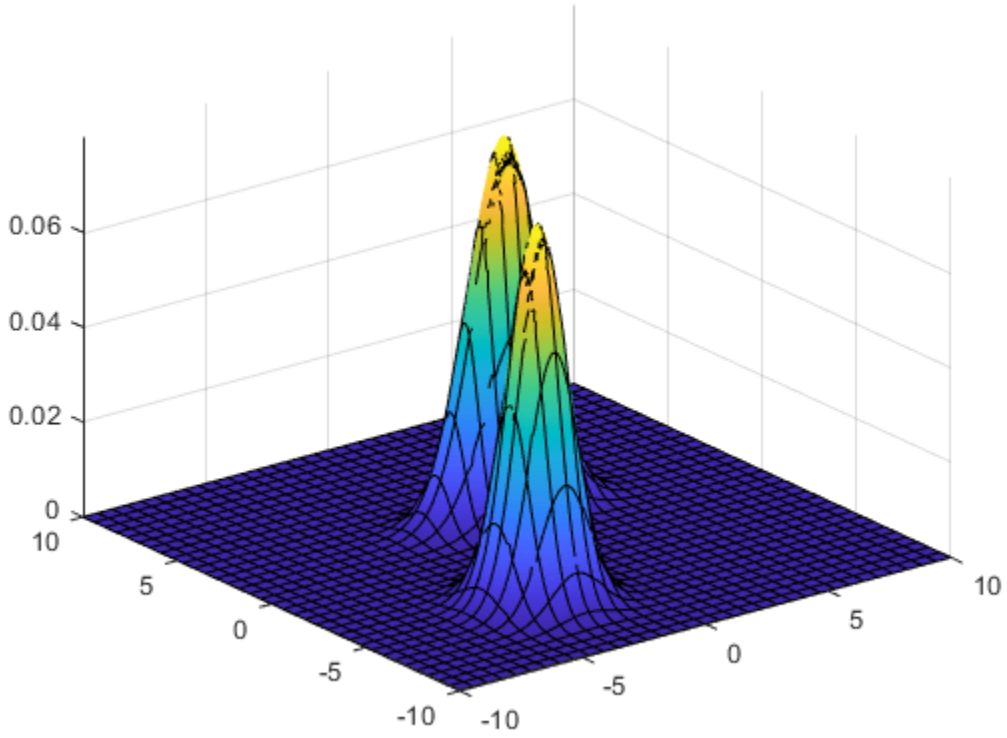
Create a `gmdistribution` object by using the `gmdistribution` function.

```
gm = gmdistribution(mu,sigma)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:     1     2

Component 2:
Mixing proportion: 0.500000
Mean:    -3    -5
```

Plot the pdf of the Gaussian mixture distribution by using `fsurf`.

```
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fsurf(gmPDF,[-10 10])
```



Input Arguments

gm — Gaussian mixture distribution
`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a `gmdistribution` object using `gmdistribution` or `fitgmdist`. Use the `gmdistribution` function to create a `gmdistribution` object by specifying the distribution parameters. Use the `fitgmdist` function to fit a `gmdistribution` model to data given a fixed number of components.

X — Values at which to evaluate pdf
n-by-*m* numeric matrix

Values at which to evaluate the pdf, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

Data Types: `single` | `double`

Output Arguments

y — pdf values
n-by-1 numeric vector

pdf values of the Gaussian mixture distribution `gm`, evaluated at `X`, returned as an *n*-by-1 numeric vector, where *n* is the number of observations in `X`.

The `pdf` function computes the pdf values by using the likelihood of each component given each observation and the component probabilities.

$$y(i) = \sum_{j=1}^{k} L(C_j \mid O_i)\mathrm{P}(\mathrm{C}_j),$$

where $L(C_j|O_i)$ is the likelihood of component `j` given observation `i`, and $P(C_j)$ is the probability of component `j`. The `pdf` function computes the likelihood term by using the multivariate normal pdf of the `j`th Gaussian mixture component evaluated at observation `i`. The component probabilities are the mixing proportions of mixture components, the `ComponentProportion` property of `gm`.

See Also

`cdf` | `fitgmdist` | `gmdistribution` | `mvnpdf` | `random`

Topics

“Create Gaussian Mixture Model” on page 5-112
“Fit Gaussian Mixture Model to Data” on page 5-115
“Simulate Data from Gaussian Mixture Model” on page 5-119
“Cluster Using Gaussian Mixture Model” on page 16-39

Introduced in R2007b

pdist

Pairwise distance between pairs of observations

Syntax

```
D = pdist(X)
D = pdist(X,Distance)
D = pdist(X,Distance,DistParameter)
```

Description

`D = pdist(X)` returns the Euclidean distance between pairs of observations in `X`.

`D = pdist(X,Distance)` returns the distance by using the method specified by `Distance`.

`D = pdist(X,Distance,DistParameter)` returns the distance by using the method specified by `Distance` and `DistParameter`. You can specify `DistParameter` only when `Distance` is `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

Examples

Compute Euclidean Distance and Convert Distance Vector to Matrix

Compute the Euclidean distance between pairs of observations, and convert the distance vector to a matrix using `squareform`.

Create a matrix with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
```

Compute the Euclidean distance.

```
D = pdist(X)
```

```
D = 1×3

    0.2954    1.0670    0.9448
```

The pairwise distances are arranged in the order (2,1), (3,1), (3,2). You can easily locate the distance between observations `i` and `j` by using `squareform`.

```
Z = squareform(D)
```

```
Z = 3×3

         0    0.2954    1.0670
    0.2954         0    0.9448
    1.0670    0.9448         0
```

`squareform` returns a symmetric matrix where `Z(i,j)` corresponds to the pairwise distance between observations `i` and `j`. For example, you can find the distance between observations 2 and 3.

```
Z(2,3)
```

```
ans = 0.9448
```

Pass `Z` to the `squareform` function to reproduce the output of the `pdist` function.

```
y = squareform(Z)
```

```
y = 1×3

    0.2954    1.0670    0.9448
```

The outputs `y` from `squareform` and `D` from `pdist` are the same.

Compute Minkowski Distance

Create a matrix with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
```

Compute the Minkowski distance with the default exponent 2.

```
D1 = pdist(X,'minkowski')
```

```
D1 = 1×3

    0.2954    1.0670    0.9448
```

Compute the Minkowski distance with an exponent of 1, which is equal to the city block distance.

```
D2 = pdist(X,'minkowski',1)
```

```
D2 = 1×3

    0.3721    1.5036    1.3136
```

```
D3 = pdist(X,'cityblock')
```

```
D3 = 1×3

    0.3721    1.5036    1.3136
```

Compute Pairwise Distance with Missing Elements Using a Custom Distance Function

Define a custom distance function that ignores coordinates with `NaN` values, and compute pairwise distance by using the custom distance function.

Create a matrix with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
```

Assume that the first element of the first observation is missing.

```
X(1,1) = NaN;
```

Compute the Euclidean distance.

```
D1 = pdist(X)
```

```
D1 = 1×3

       NaN       NaN    0.9448
```

If observation `i` or `j` contains `NaN` values, the function `pdist` returns `NaN` for the pairwise distance between `i` and `j`. Therefore, D1(1) and D1(2), the pairwise distances (2,1) and (3,1), are `NaN` values.

Define a custom distance function `naneucdist` that ignores coordinates with `NaN` values and returns the Euclidean distance.

```
function D2 = naneucdist(XI,XJ)
%NANEUCDIST Euclidean distance ignoring coordinates with NaNs
n = size(XI,2);
sqdx = (XI-XJ).^2;
nstar = sum(~isnan(sqdx),2); % Number of pairs that do not contain NaNs
nstar(nstar == 0) = NaN; % To return NaN if all pairs include NaNs
D2squared = sum(sqdx,2,'omitnan').*n./nstar; % Correction for missing coordinates
D2 = sqrt(D2squared);
```

Compute the distance with `naneucdist` by passing the function handle as an input argument of `pdist`.

```
D2 = pdist(X,@naneucdist)
```

```
D2 = 1×3

    0.3974    1.1538    0.9448
```

Input Arguments

X — Input data

numeric matrix

Input data, specified as a numeric matrix of size *m*-by-*n*. Rows correspond to individual observations, and columns correspond to individual variables.

Data Types: `single` | `double`

Distance — Distance metric

character vector | string scalar | function handle

Distance metric, specified as a character vector, string scalar, or function handle, as described in the following table.

Value	Description
`'euclidean'`	Euclidean distance (default).
`'squaredeuclidean'`	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`. Use `DistParameter` to specify another value for `S`.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use `DistParameter` to specify another value for `C`, where the matrix `C` is symmetric and positive definite.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. Use `DistParameter` to specify a different exponent `P`, where `P` is a positive scalar value of the exponent.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For definitions, see “Distance Metrics” on page 33-4495.

When you use `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`, you can specify an additional input argument `DistParameter` to control these metrics. You can also use these metrics in the same way as the other metrics with a default value of `DistParameter`.

Example: `'minkowski'`

`DistParameter` — Distance metric parameter values
positive scalar | numeric vector | numeric matrix

Distance metric parameter values, specified as a positive scalar, numeric vector, or numeric matrix. This argument is valid only when you specify `Distance` as `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

- If `Distance` is `'seuclidean'`, `DistParameter` is a vector of scaling factors for each dimension, specified as a positive vector. The default value is `std(X,'omitnan')`.
- If `Distance` is `'minkowski'`, `DistParameter` is the exponent of Minkowski distance, specified as a positive scalar. The default value is 2.
- If `Distance` is `'mahalanobis'`, `DistParameter` is a covariance matrix, specified as a numeric matrix. The default value is `cov(X,'omitrows')`. `DistParameter` must be symmetric and positive definite.

Example: `'minkowski',3`

Data Types: `single` | `double`

Output Arguments

`D` — Pairwise distances
numeric row vector

Pairwise distances, returned as a numeric row vector of length $m(m–1)/2$, corresponding to pairs of observations, where m is the number of observations in `X`.

The distances are arranged in the order (2,1), (3,1), ..., (m,1), (3,2), ..., (m,2), ..., (m,m–1), i.e., the lower-left triangle of the m-by-m distance matrix in column order. The pairwise distance between observations i and j is in $D((i-1)*(m-i/2)+j-i)$ for $i \le j$.

You can convert `D` into a symmetric matrix by using the `squareform` function. `Z = squareform(D)` returns an m-by-m matrix where `Z(i,j)` corresponds to the pairwise distance between observations i and j.

If observation i or j contains `NaN`s, then the corresponding value in `D` is `NaN` for the built-in distance functions.

`D` is commonly used as a dissimilarity matrix in clustering or multidimensional scaling. For details, see “Hierarchical Clustering” on page 16-6 and the function reference pages for `cmdscale`, `cophenet`, `linkage`, `mdscale`, and `optimalleaforder`. These functions take `D` as an input argument.

More About

Distance Metrics

A distance metric is a function that defines a distance between two observations. `pdist` supports various distance metrics: Euclidean distance, standardized Euclidean distance, Mahalanobis distance,

city block distance, Minkowski distance, Chebychev distance, cosine distance, correlation distance, Hamming distance, Jaccard distance, and Spearman distance.

Given an m-by-n data matrix X, which is treated as m (1-by-n) row vectors x_1, x_2, ..., x_m, the various distances between the vector x_s and x_t are defined as follows:

- Euclidean distance

  $$d_{st}^2 = (x_s - x_t)(x_s - x_t)' .$$

  The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.

- Standardized Euclidean distance

  $$d_{st}^2 = (x_s - x_t)V^{-1}(x_s - x_t)',$$

  where V is the n-by-n diagonal matrix whose jth diagonal element is $(S(j))^2$, where S is a vector of scaling factors for each dimension.

- Mahalanobis distance

  $$d_{st}^2 = (x_s - x_t)C^{-1}(x_s - x_t)',$$

  where C is the covariance matrix.

- City block distance

  $$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right| .$$

  The city block distance is a special case of the Minkowski distance, where $p = 1$.

- Minkowski distance

  $$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left| x_{sj} - x_{tj} \right|^p} .$$

  For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.

- Chebychev distance

  $$d_{st} = \max_j \left\{ \left| x_{sj} - x_{tj} \right| \right\} .$$

  The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- Cosine distance

  $$d_{st} = 1 - \frac{x_s x'_t}{\sqrt{(x_s x'_s)(x_t x'_t)}} .$$

- Correlation distance

  $$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(x_t - \bar{x}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(x_t - \bar{x}_t)(x_t - \bar{x}_t)'}},$$

where

$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$ and $\bar{x}_t = \frac{1}{n}\sum_j x_{tj}$.

- Hamming distance

  $d_{st} = \left(\#\left(x_{sj} \neq x_{tj}\right)/n \right)$.

- Jaccard distance

  $$d_{st} = \frac{\#\left[\left(x_{sj} \neq x_{tj}\right) \cap \left(\left(x_{sj} \neq 0\right) \cup \left(x_{tj} \neq 0\right)\right)\right]}{\#\left[\left(x_{sj} \neq 0\right) \cup \left(x_{tj} \neq 0\right)\right]}.$$

- Spearman distance

  $$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

  where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...x_{mj}, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and x_t, i.e., $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$.
  - $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.
  - $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The distance input argument value (`Distance`) must be a compile-time constant. For example, to use the Minkowski distance, include `coder.Constant('Minkowski')` in the `-args` value of `codegen`.
- The distance input argument value (`Distance`) cannot be a custom distance function.
- The generated code of `pdist` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Code Generation

Generate CUDA® code for NVIDIA® GPUs using GPU Coder™.

Usage notes and limitations:

- The supported distance input argument values (`Distance`) for optimized CUDA code are `'euclidean'`, `'squaredeuclidean'`, `'seuclidean'`, `'cityblock'`, `'minkowski'`, `'chebychev'`, `'cosine'`, `'correlation'`, `'hamming'`, and `'jaccard'`.
- `Distance` cannot be a custom distance function.
- `Distance` must be a compile-time constant.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The distance input argument value (`Distance`) cannot be a custom distance function.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cluster` | `clusterdata` | `cmdscale` | `cophenet` | `dendrogram` | `inconsistent` | `linkage` | `pdist2` | `silhouette` | `squareform`

Topics
“Choose Cluster Analysis Method” on page 16-2
“Hierarchical Clustering” on page 16-6

Introduced before R2006a

pdist2

Pairwise distance between two sets of observations

Syntax

```
D = pdist2(X,Y,Distance)

D = pdist2(X,Y,Distance,DistParameter)

D = pdist2(___,Name,Value)
[D,I] = pdist2(___,Name,Value)
```

Description

`D = pdist2(X,Y,Distance)` returns the distance between each pair of observations in `X` and `Y` using the metric specified by `Distance`.

`D = pdist2(X,Y,Distance,DistParameter)` returns the distance using the metric specified by `Distance` and `DistParameter`. You can specify `DistParameter` only when `Distance` is `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

`D = pdist2(___,Name,Value)` specifies an additional option using one of the name-value pair arguments `'Smallest'` or `'Largest'` in addition to any of the arguments in the previous syntaxes.

For example,

- `D = pdist2(X,Y,Distance,'Smallest',K)` computes the distance using the metric specified by `Distance` and returns the `K` smallest pairwise distances to observations in `X` for each observation in `Y` in ascending order.
- `D = pdist2(X,Y,Distance,DistParameter,'Largest',K)` computes the distance using the metric specified by `Distance` and `DistParameter` and returns the `K` largest pairwise distances in descending order.

`[D,I] = pdist2(___,Name,Value)` also returns the matrix `I`. The matrix `I` contains the indices of the observations in `X` corresponding to the distances in `D`.

Examples

Compute Euclidean Distance

Create two matrices with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
Y = rand(3,2);
```

Compute the Euclidean distance. The default value of the input argument `Distance` is `'euclidean'`. When computing the Euclidean distance without using a name-value pair argument, you do not need to specify `Distance`.

```
D = pdist2(X,Y)
```

```
D = 3×3

    0.5387    0.8018    0.1538
    0.7100    0.5951    0.3422
    0.8805    0.4242    1.2050
```

`D(i,j)` corresponds to the pairwise distance between observation `i` in `X` and observation `j` in `Y`.

Compute Minkowski Distance

Create two matrices with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
Y = rand(3,2);
```

Compute the Minkowski distance with the default exponent 2.

```
D1 = pdist2(X,Y,'minkowski')
```

```
D1 = 3×3

    0.5387    0.8018    0.1538
    0.7100    0.5951    0.3422
    0.8805    0.4242    1.2050
```

Compute the Minkowski distance with an exponent of 1, which is equal to the city block distance.

```
D2 = pdist2(X,Y,'minkowski',1)
```

```
D2 = 3×3

    0.5877    1.0236    0.2000
    0.9598    0.8337    0.3899
    1.0189    0.4800    1.7036
```

```
D3 = pdist2(X,Y,'cityblock')
```

```
D3 = 3×3

    0.5877    1.0236    0.2000
    0.9598    0.8337    0.3899
    1.0189    0.4800    1.7036
```

Find the Two Smallest Pairwise Distances

Create two matrices with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
Y = rand(3,2);
```

Find the two smallest pairwise Euclidean distances to observations in X for each observation in Y.

```
[D,I] = pdist2(X,Y,'euclidean','Smallest',2)
```

```
D = 2×3

    0.5387    0.4242    0.1538
    0.7100    0.5951    0.3422


I = 2×3

     1     3     1
     2     2     2
```

For each observation in Y, `pdist2` finds the two smallest distances by computing and comparing the distance values to all the observations in X. The function then sorts the distances in each column of D in ascending order. I contains the indices of the observations in X corresponding to the distances in D.

Compute Pairwise Distance with Missing Elements Using a Custom Distance Function

Define a custom distance function that ignores coordinates with NaN values, and compute pairwise distance by using the custom distance function.

Create two matrices with three observations and three variables.

```
rng('default') % For reproducibility
X = rand(3,3)
Y = [X(:,1:2) rand(3,1)]
```

```
X =

    0.8147    0.9134    0.2785
    0.9058    0.6324    0.5469
    0.1270    0.0975    0.9575


Y =

    0.8147    0.9134    0.9649
    0.9058    0.6324    0.1576
    0.1270    0.0975    0.9706
```

The first two columns of X and Y are identical. Assume that `X(1,1)` is missing.

```
X(1,1) = NaN
```

```
X =
```

```
       NaN    0.9134    0.2785
    0.9058    0.6324    0.5469
    0.1270    0.0975    0.9575
```

Compute the Hamming distance.

```
D1 = pdist2(X,Y,'hamming')
```

```
D1 =

       NaN       NaN       NaN
    1.0000    0.3333    1.0000
    1.0000    1.0000    0.3333
```

If observation `i` in `X` or observation `j` in `Y` contains `NaN` values, the function `pdist2` returns `NaN` for the pairwise distance between `i` and `j`. Therefore, D1(1,1), D1(1,2), and D1(1,3) are `NaN` values.

Define a custom distance function `nanhamdist` that ignores coordinates with `NaN` values and computes the Hamming distance. When working with a large number of observations, you can compute the distance more quickly by looping over coordinates of the data.

```
function D2 = nanhamdist(XI,XJ)
%NANHAMDIST Hamming distance ignoring coordinates with NaNs
[m,p] = size(XJ);
nesum = zeros(m,1);
pstar = zeros(m,1);
for q = 1:p
    notnan = ~(isnan(XI(q)) | isnan(XJ(:,q)));
    nesum = nesum + ((XI(q) ~= XJ(:,q)) & notnan);
    pstar = pstar + notnan;
end
D2 = nesum./pstar;
```

Compute the distance with `nanhamdist` by passing the function handle as an input argument of `pdist2`.

```
D2 = pdist2(X,Y,@nanhamdist)
```

```
D2 =

    0.5000    1.0000    1.0000
    1.0000    0.3333    1.0000
    1.0000    1.0000    0.3333
```

Assign New Data to Existing Clusters and Generate C/C++ Code

`kmeans` performs *k*-means clustering to partition data into *k* clusters. When you have a new data set to cluster, you can create new clusters that include the existing data and the new data by using

`kmeans`. The `kmeans` function supports C/C++ code generation, so you can generate code that accepts training data and returns clustering results, and then deploy the code to a device. In this workflow, you must pass training data, which can be of considerable size. To save memory on the device, you can separate training and prediction by using `kmeans` and `pdist2`, respectively.

Use `kmeans` to create clusters in MATLAB® and use `pdist2` in the generated code to assign new data to existing clusters. For code generation, define an entry-point function that accepts the cluster centroid positions and the new data set, and returns the index of the nearest cluster. Then, generate code for the entry-point function.

Generating C/C++ code requires MATLAB® Coder™.

Perform *k*-Means Clustering

Generate a training data set using three distributions.

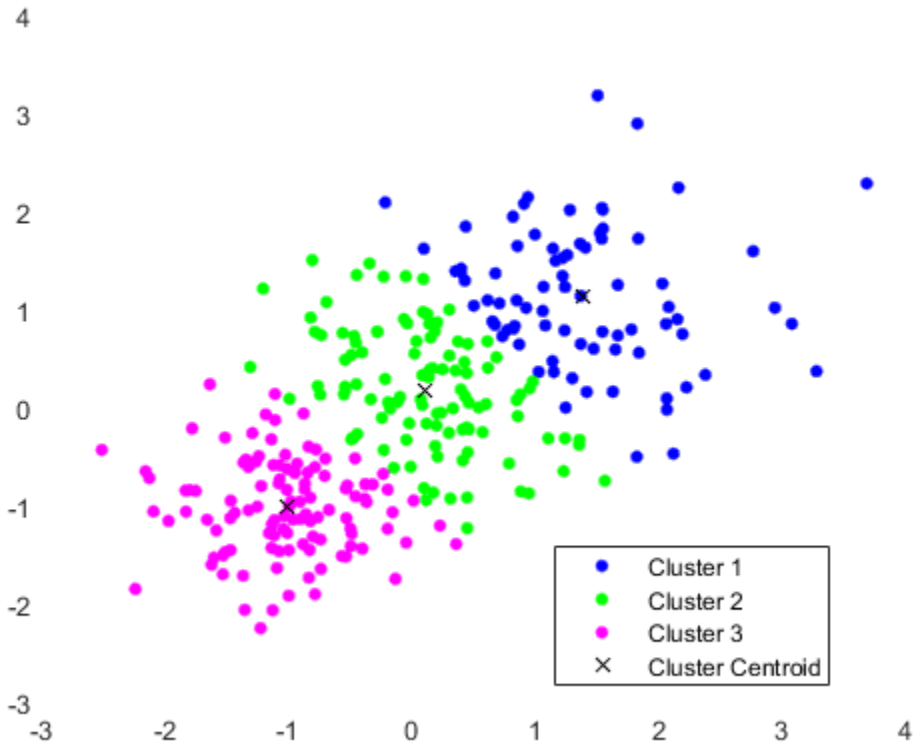
```
rng('default') % For reproducibility
X = [randn(100,2)*0.75+ones(100,2);
     randn(100,2)*0.5-ones(100,2);
     randn(100,2)*0.75];
```

Partition the training data into three clusters by using `kmeans`.

```
[idx,C] = kmeans(X,3);
```

Plot the clusters and the cluster centroids.

```
figure
gscatter(X(:,1),X(:,2),idx,'bgm')
hold on
plot(C(:,1),C(:,2),'kx')
legend('Cluster 1','Cluster 2','Cluster 3','Cluster Centroid')
```



Assign New Data to Existing Clusters

Generate a test data set.

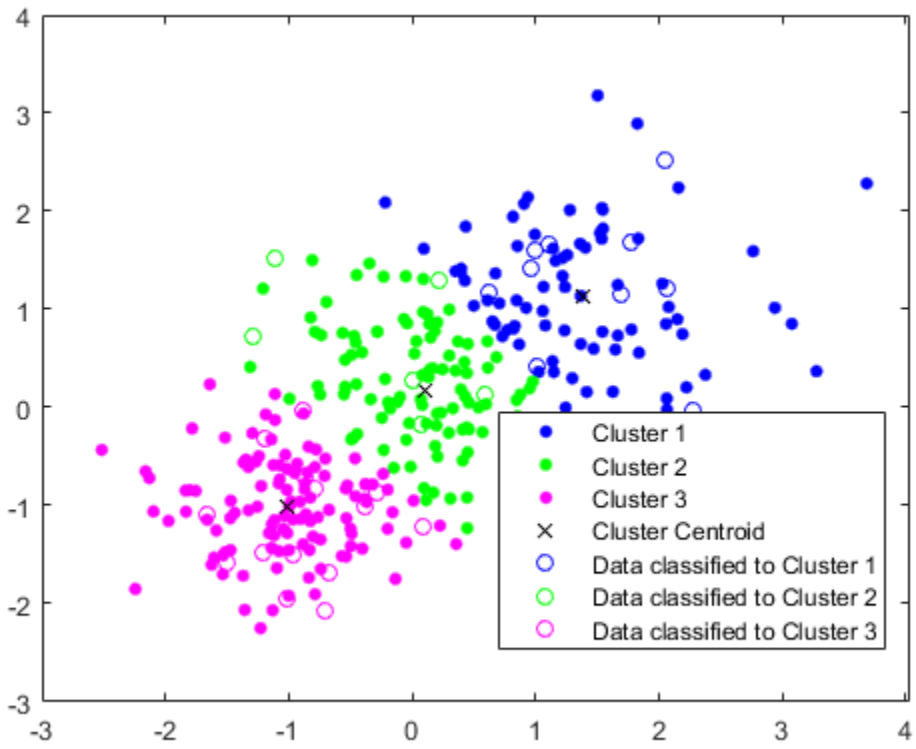
```
Xtest = [randn(10,2)*0.75+ones(10,2);
    randn(10,2)*0.5-ones(10,2);
    randn(10,2)*0.75];
```

Classify the test data set using the existing clusters. Find the nearest centroid from each test data point by using `pdist2`.

```
[~,idx_test] = pdist2(C,Xtest,'euclidean','Smallest',1);
```

Plot the test data and label the test data using `idx_test` by using `gscatter`.

```
gscatter(Xtest(:,1),Xtest(:,2),idx_test,'bgm','ooo')
legend('Cluster 1','Cluster 2','Cluster 3','Cluster Centroid', ...
    'Data classified to Cluster 1','Data classified to Cluster 2', ...
    'Data classified to Cluster 3')
```



Generate Code

Generate C code that assigns new data to the existing clusters. Note that generating C/C++ code requires MATLAB® Coder™.

Define an entry-point function named `findNearestCentroid` that accepts centroid positions and new data, and then find the nearest cluster by using `pdist2`.

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would cause errors during code generation.

```
type findNearestCentroid % Display contents of findNearestCentroid.m
```

```
function idx = findNearestCentroid(C,X) %#codegen
[~,idx] = pdist2(C,X,'euclidean','Smallest',1); % Find the nearest centroid
```

Note: If you click the button located in the upper-right section of this page and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate code by using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and array size of the inputs of `findNearestCentroid`, pass a MATLAB expression that represents the set of values with a certain data type and array size by using

the `-args` option. For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45.

```
codegen findNearestCentroid -args {C,Xtest}
```

```
Code generation successful.
```

`codegen` generates the MEX function `findNearestCentroid_mex` with a platform-dependent extension.

Verify the generated code.

```
myIndx = findNearestCentroid(C,Xtest);
myIndex_mex = findNearestCentroid_mex(C,Xtest);
verifyMEX = isequal(idx_test,myIndx,myIndex_mex)
```

```
verifyMEX = logical
   1
```

`isequal` returns logical 1 (`true`), which means all the inputs are equal. The comparison confirms that the `pdist2` function, the `findNearestCentroid` function, and the MEX function return the same index.

You can also generate optimized CUDA® code using GPU Coder™.

```
cfg = coder.gpuConfig('mex');
codegen -config cfg findNearestCentroid -args {C,Xtest}
```

For more information on code generation, see “General Code Generation Workflow” on page 32-5. For more information on GPU coder, see “Get Started with GPU Coder” (GPU Coder) and “Supported Functions” (GPU Coder).

Input Arguments

`X,Y` — Input data

numeric matrix

Input data, specified as a numeric matrix. `X` is an *mx*-by-*n* matrix and `Y` is an *my*-by-*n* matrix. Rows correspond to individual observations, and columns correspond to individual variables.

Data Types: `single` | `double`

`Distance` — Distance metric

character vector | string scalar | function handle

Distance metric, specified as a character vector, string scalar, or function handle, as described in the following table.

Value	Description
`'euclidean'`	Euclidean distance (default).
`'squaredeuclidean'`	Squared Euclidean distance. (This option is provided for efficiency only. It does not satisfy the triangle inequality.)

Value	Description
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation, `S = std(X,'omitnan')`. Use `DistParameter` to specify another value for `S`.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use `DistParameter` to specify another value for `C`, where the matrix `C` is symmetric and positive definite.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is 2. Use `DistParameter` to specify a different exponent `P`, where `P` is a positive scalar value of the exponent.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors).
`'correlation'`	One minus the sample correlation between points (treated as sequences of values).
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For definitions, see “Distance Metrics” on page 33-4509.

When you use `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`, you can specify an additional input argument `DistParameter` to control these metrics. You can also use these metrics in the same way as the other metrics with a default value of `DistParameter`.

Example: `'minkowski'`

`DistParameter` — Distance metric parameter values
positive scalar | numeric vector | numeric matrix

Distance metric parameter values, specified as a positive scalar, numeric vector, or numeric matrix. This argument is valid only when you specify `Distance` as `'seuclidean'`, `'minkowski'`, or `'mahalanobis'`.

- If `Distance` is `'seuclidean'`, `DistParameter` is a vector of scaling factors for each dimension, specified as a positive vector. The default value is `std(X,'omitnan')`.
- If `Distance` is `'minkowski'`, `DistParameter` is the exponent of Minkowski distance, specified as a positive scalar. The default value is 2.
- If `Distance` is `'mahalanobis'`, `DistParameter` is a covariance matrix, specified as a numeric matrix. The default value is `cov(X,'omitrows')`. `DistParameter` must be symmetric and positive definite.

Example: `'minkowski',3`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: Either `'Smallest',K` or `'Largest',K`. You cannot use both `'Smallest'` and `'Largest'`.

`Smallest` — Number of smallest distances to find
positive integer

Number of smallest distances to find, specified as the comma-separated pair consisting of `'Smallest'` and a positive integer. If you specify `'Smallest'`, then `pdist2` sorts the distances in each column of `D` in ascending order.

Example: `'Smallest',3`

Data Types: `single` | `double`

`Largest` — Number of largest distances to find
positive integer

Number of largest distances to find, specified as the comma-separated pair consisting of `'Largest'` and a positive integer. If you specify `'Largest'`, then `pdist2` sorts the distances in each column of `D` in descending order.

Example: `'Largest',3`

Data Types: `single` | `double`

Output Arguments

`D` — Pairwise distances
numeric matrix

Pairwise distances, returned as a numeric matrix.

If you do not specify either `'Smallest'` or `'Largest'`, then `D` is an *mx*-by-*my* matrix, where *mx* and *my* are the number of observations in `X` and `Y`, respectively. `D(i,j)` is the distance between

observation `i` in `X` and observation `j` in `Y`. If observation *i* in `X` or observation *j* in `Y` contains `NaN`, then `D(i,j)` is `NaN` for the built-in distance functions.

If you specify either `'Smallest'` or `'Largest'` as `K`, then `D` is a `K`-by-*my* matrix. `D` contains either the `K` smallest or `K` largest pairwise distances to observations in `X` for each observation in `Y`. For each observation in `Y`, `pdist2` finds the `K` smallest or largest distances by computing and comparing the distance values to all the observations in `X`. If `K` is greater than *mx*, `pdist2` returns an *mx*-by-*my* matrix.

`I` — Sort index
positive integer matrix

Sort index, returned as a positive integer matrix. `I` is the same size as `D`. `I` contains the indices of the observations in `X` corresponding to the distances in `D`.

More About

Distance Metrics

A distance metric is a function that defines a distance between two observations. `pdist2` supports various distance metrics: Euclidean distance, standardized Euclidean distance, Mahalanobis distance, city block distance, Minkowski distance, Chebychev distance, cosine distance, correlation distance, Hamming distance, Jaccard distance, and Spearman distance.

Given an *mx*-by-*n* data matrix *X*, which is treated as *mx* (1-by-*n*) row vectors x_1, x_2, ..., x_{mx}, and an *my*-by-*n* data matrix *Y*, which is treated as *my* (1-by-*n*) row vectors y_1, y_2, ...,y_{my}, the various distances between the vector x_s and y_t are defined as follows:

- Euclidean distance

  $$d_{st}^2 = (x_s - y_t)(x_s - y_t)' .$$

  The Euclidean distance is a special case of the Minkowski distance, where $p = 2$.
- Standardized Euclidean distance

  $$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)',$$

  where *V* is the *n*-by-*n* diagonal matrix whose *j*th diagonal element is $(S(j))^2$, where *S* is a vector of scaling factors for each dimension.
- Mahalanobis distance

  $$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)',$$

  where *C* is the covariance matrix.
- City block distance

  $$d_{st} = \sum_{j=1}^{n} \left| x_{sj} - y_{tj} \right| .$$

  The city block distance is a special case of the Minkowski distance, where $p = 1$.
- Minkowski distance

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} \left|x_{sj} - y_{tj}\right|^p}.$$

For the special case of $p = 1$, the Minkowski distance gives the city block distance. For the special case of $p = 2$, the Minkowski distance gives the Euclidean distance. For the special case of $p = \infty$, the Minkowski distance gives the Chebychev distance.

- Chebychev distance

  $$d_{st} = \max_j\left\{\left|x_{sj} - y_{tj}\right|\right\}.$$

  The Chebychev distance is a special case of the Minkowski distance, where $p = \infty$.

- Cosine distance

  $$d_{st} = \left(1 - \frac{x_s y'_t}{\sqrt{(x_s x'_s)(y_t y'_t)}}\right).$$

- Correlation distance

  $$d_{st} = 1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}},$$

  where

  $$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

  and

  $$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}.$$

- Hamming distance

  $$d_{st} = \left(\#\left(x_{sj} \neq y_{tj}\right)/n\right).$$

- Jaccard distance

  $$d_{st} = \frac{\#\left[\left(x_{sj} \neq y_{tj}\right) \cap \left(\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right)\right]}{\#\left[\left(x_{sj} \neq 0\right) \cup \left(y_{tj} \neq 0\right)\right]}.$$

- Spearman distance

  $$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}},$$

  where

  - r_{sj} is the rank of x_{sj} taken over x_{1j}, x_{2j}, ...$x_{mx,j}$, as computed by `tiedrank`.
  - r_{tj} is the rank of y_{tj} taken over y_{1j}, y_{2j}, ...$y_{my,j}$, as computed by `tiedrank`.
  - r_s and r_t are the coordinate-wise rank vectors of x_s and y_t, that is, $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.
  - $\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$.

- $\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- The first input `X` must be a tall array. Input `Y` cannot be a tall array.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The distance input argument value (`Distance`) must be a compile-time constant. For example, to use the Minkowski distance, include `coder.Constant('Minkowski')` in the `-args` value of `codegen`.
- The distance input argument value (`Distance`) cannot be a custom distance function.
- Names in name-value pair arguments must be compile-time constants. For example, to use the `'Smallest'` name-value pair argument in the generated code, include `{coder.Constant('Smallest'),0}` in the `-args` value of `codegen`.
- The sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision.
- The generated code of `pdist2` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.
- Starting in R2020a, `pdist2` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for strict single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Code Generation

Generate CUDA® code for NVIDIA® GPUs using GPU Coder™.

Usage notes and limitations:

- The supported distance input argument values (`Distance`) for optimized CUDA code are `'euclidean'`, `'squaredeuclidean'`, `'seuclidean'`, `'cityblock'`, `'minkowski'`, `'chebychev'`, `'cosine'`, `'correlation'`, `'hamming'`, and `'jaccard'`.

- `Distance` cannot be a custom distance function.
- `Distance` must be a compile-time constant.
- Names in name-value pair arguments must be compile-time constants.
- The sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The distance input argument value (`Distance`) cannot be a custom distance function.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`ExhaustiveSearcher` | `KDTreeSearcher` | `createns` | `knnsearch` | `pdist`

Introduced in R2010a

pearsrnd

Pearson system random numbers

Syntax

```
r = pearsrnd(mu,sigma,skew,kurt,m,n)
r = pearsrnd(mu,sigma,skew,kurt)
r = pearsrnd(mu,sigma,skew,kurt,m,n,...)
r = pearsrnd(mu,sigma,skew,kurt,[m,n,...])
[r,type] = pearsrnd(...)
[r,type,coefs] = pearsrnd(...)
```

Description

`r = pearsrnd(mu,sigma,skew,kurt,m,n)` returns an `m`-by-`n` matrix of random numbers drawn from the distribution in the Pearson system with mean `mu`, standard deviation `sigma`, skewness `skew`, and kurtosis `kurt`. The parameters `mu`, `sigma`, `skew`, and `kurt` must be scalars.

Note Because `r` is a random sample, its sample moments, especially the skewness and kurtosis, typically differ somewhat from the specified distribution moments.

`pearsrnd` uses the definition of kurtosis for which a normal distribution has a kurtosis of 3. Some definitions of kurtosis subtract 3, so that a normal distribution has a kurtosis of 0. The `pearsrnd` function does not use this convention.

Some combinations of moments are not valid; in particular, the kurtosis must be greater than the square of the skewness plus 1. The kurtosis of the normal distribution is defined to be 3.

`r = pearsrnd(mu,sigma,skew,kurt)` returns a scalar value.

`r = pearsrnd(mu,sigma,skew,kurt,m,n,...)` or `r = pearsrnd(mu,sigma,skew,kurt,[m,n,...])` returns an `m`-by-`n`-by-... array.

`[r,type] = pearsrnd(...)` returns the type of the specified distribution within the Pearson system. `type` is a scalar integer from `0` to `7`. Set `m` and `n` to `0` to identify the distribution type without generating any random values.

The seven distribution types in the Pearson system correspond to the following distributions:

- `0` — Normal on page B-119 distribution
- `1` — Four-parameter beta on page B-6 distribution
- `2` — Symmetric four-parameter beta on page B-6 distribution
- `3` — Three-parameter gamma on page B-47 distribution
- `4` — Not related to any standard distribution. The density is proportional to:

  $(1 + ((x - a)/b)^2)^{-c} \exp(-d \arctan((x - a)/b))$.

- 5 — Inverse gamma on page B-47 location-scale distribution
- 6 — *F* on page B-45 location-scale distribution
- 7 — Student's *t* location-scale on page B-156 distribution

`[r,type,coefs] = pearsrnd(...)` returns the coefficients `coefs` of the quadratic polynomial that defines the distribution via the differential equation

$$\frac{d}{dx}\log(p(x)) = \frac{-(a+x)}{c(0)+c(1)x+c(2)x^2}.$$

Examples

Generate random values from the standard normal distribution:

```
r = pearsrnd(0,1,0,3,100,1);  % Equivalent to randn(100,1)
```

Determine the distribution type:

```
[r,type] = pearsrnd(0,1,1,4,0,0);
r =
     []
type =
     1
```

References

[1] Johnson, N.L., S. Kotz, and N. Balakrishnan (1994) Continuous Univariate Distributions, Volume 1, Wiley-Interscience, Pg 15, Eqn 12.33.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code matches MATLAB only when generated output `r` is scalar.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`johnsrnd` | `random`

Topics

“Generating Data Using Flexible Families of Distributions” on page 7-20

Introduced in R2006a

perfcurve

Receiver operating characteristic (ROC) curve or other performance curve for classifier output

Syntax

```
[X,Y] = perfcurve(labels,scores,posclass)
[X,Y,T] = perfcurve(labels,scores,posclass)
[X,Y,T,AUC] = perfcurve(labels,scores,posclass)
[X,Y,T,AUC,OPTROCPT] = perfcurve(labels,scores,posclass)
[X,Y,T,AUC,OPTROCPT,SUBY] = perfcurve(labels,scores,posclass)
[X,Y,T,AUC,OPTROCPT,SUBY,SUBYNAMES] = perfcurve(labels,scores,posclass)
[___] = perfcurve(labels,scores,posclass,Name,Value)
```

Description

`[X,Y] = perfcurve(labels,scores,posclass)` returns the `X` and `Y` coordinates of an ROC curve for a vector of classifier predictions, `scores`, given true class labels, `labels`, and the positive class label, `posclass`. You can visualize the performance curve using `plot(X,Y)`.

`[X,Y,T] = perfcurve(labels,scores,posclass)` returns an array of thresholds on classifier scores for the computed values of `X` and `Y`.

`[X,Y,T,AUC] = perfcurve(labels,scores,posclass)` returns the area under the curve for the computed values of `X` and `Y`.

`[X,Y,T,AUC,OPTROCPT] = perfcurve(labels,scores,posclass)` returns the optimal operating point of the ROC curve.

`[X,Y,T,AUC,OPTROCPT,SUBY] = perfcurve(labels,scores,posclass)` returns the `Y` values for negative subclasses.

`[X,Y,T,AUC,OPTROCPT,SUBY,SUBYNAMES] = perfcurve(labels,scores,posclass)` returns the negative class names.

`[___] = perfcurve(labels,scores,posclass,Name,Value)` returns the coordinates of a ROC curve and any other output argument from the previous syntaxes, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can provide a list of negative classes, change the `X` or `Y` criterion, compute pointwise confidence bounds on page 33-4537 using cross validation or bootstrap, specify the misclassification cost, or compute the confidence bounds in parallel.

Examples

Plot ROC Curve for Classification by Logistic Regression

Load the sample data.

```
load fisheriris
```

Use only the first two features as predictor variables. Define a binary classification problem by using only the measurements that correspond to the species versicolor and virginica.

```
pred = meas(51:end,1:2);
```

Define the binary response variable.

```
resp = (1:100)'>50;  % Versicolor = 0, virginica = 1
```

Fit a logistic regression model.

```
mdl = fitglm(pred,resp,'Distribution','binomial','Link','logit');
```

Compute the ROC curve. Use the probability estimates from the logistic regression model as scores.

```
scores = mdl.Fitted.Probability;
[X,Y,T,AUC] = perfcurve(species(51:end,:),scores,'virginica');
```

`perfcurve` stores the threshold values in the array `T`.

Display the area under the curve.

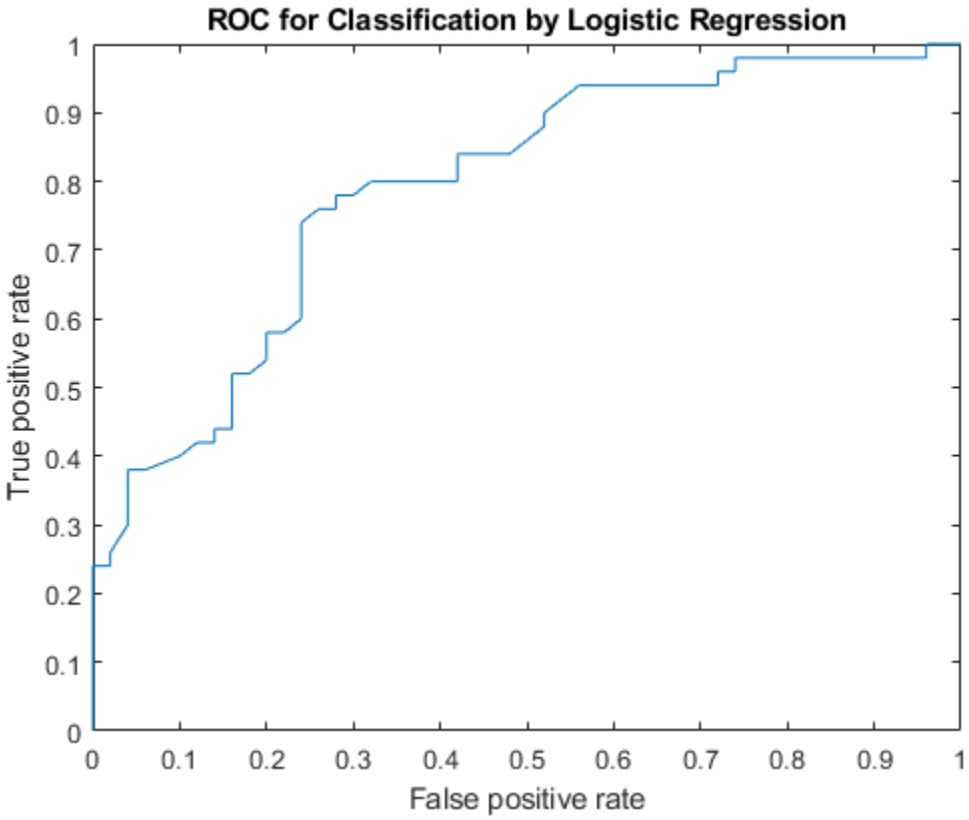
```
AUC
```

```
AUC = 0.7918
```

The area under the curve is 0.7918. The maximum AUC is 1, which corresponds to a perfect classifier. Larger AUC values indicate better classifier performance.

Plot the ROC curve.

```
plot(X,Y)
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC for Classification by Logistic Regression')
```



Compare Classification Methods Using ROC Curve

Load the sample data.

```
load ionosphere
```

X is a 351x34 real-valued matrix of predictors. Y is a character array of class labels: `'b'` for bad radar returns and `'g'` for good radar returns.

Reformat the response to fit a logistic regression. Use the predictor variables 3 through 34.

```
resp = strcmp(Y,'b'); % resp = 1, if Y = 'b', or 0 if Y = 'g'
pred = X(:,3:34);
```

Fit a logistic regression model to estimate the posterior probabilities for a radar return to be a bad one.

```
mdl = fitglm(pred,resp,'Distribution','binomial','Link','logit');
score_log = mdl.Fitted.Probability; % Probability estimates
```

Compute the standard ROC curve using the probabilities for scores.

```
[Xlog,Ylog,Tlog,AUClog] = perfcurve(resp,score_log,'true');
```

Train an SVM classifier on the same sample data. Standardize the data.

```
mdlSVM = fitcsvm(pred,resp,'Standardize',true);
```

Compute the posterior probabilities (scores).

```
mdlSVM = fitPosterior(mdlSVM);
[~,score_svm] = resubPredict(mdlSVM);
```

The second column of `score_svm` contains the posterior probabilities of bad radar returns.

Compute the standard ROC curve using the scores from the SVM model.

```
[Xsvm,Ysvm,Tsvm,AUCsvm] = perfcurve(resp,score_svm(:,mdlSVM.ClassNames),'true');
```

Fit a naive Bayes classifier on the same sample data.

```
mdlNB = fitcnb(pred,resp);
```

Compute the posterior probabilities (scores).

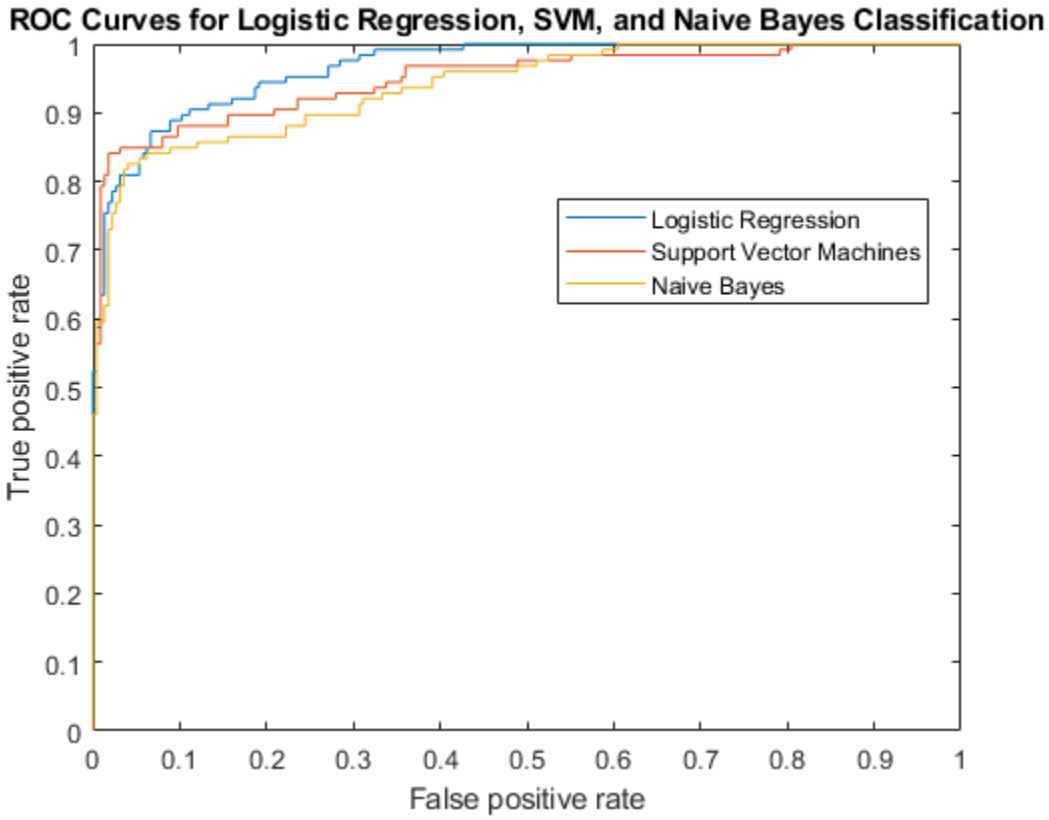
```
[~,score_nb] = resubPredict(mdlNB);
```

Compute the standard ROC curve using the scores from the naive Bayes classification.

```
[Xnb,Ynb,Tnb,AUCnb] = perfcurve(resp,score_nb(:,mdlNB.ClassNames),'true');
```

Plot the ROC curves on the same graph.

```
plot(Xlog,Ylog)
hold on
plot(Xsvm,Ysvm)
plot(Xnb,Ynb)
legend('Logistic Regression','Support Vector Machines','Naive Bayes','Location','Best')
xlabel('False positive rate'); ylabel('True positive rate');
title('ROC Curves for Logistic Regression, SVM, and Naive Bayes Classification')
hold off
```



Although SVM produces better ROC values for higher thresholds, logistic regression is usually better at distinguishing the bad radar returns from the good ones. The ROC curve for naive Bayes is generally lower than the other two ROC curves, which indicates worse in-sample performance than the other two classifier methods.

Compare the area under the curve for all three classifiers.

```
AUClog
```

```
AUClog = 0.9659
```

```
AUCsvm
```

```
AUCsvm = 0.9489
```

```
AUCnb
```

```
AUCnb = 0.9393
```

Logistic regression has the highest AUC measure for classification and naive Bayes has the lowest. This result suggests that logistic regression has better in-sample average performance for this sample data.

Determine the Parameter Value for Custom Kernel Function

This example shows how to determine the better parameter value for a custom kernel function in a classifier using the ROC curves.

Generate a random set of points within the unit circle.

```
rng(1);  % For reproducibility
n = 100; % Number of points per quadrant

r1 = sqrt(rand(2*n,1));                     % Random radii
t1 = [pi/2*rand(n,1); (pi/2*rand(n,1)+pi)]; % Random angles for Q1 and Q3
X1 = [r1.*cos(t1) r1.*sin(t1)];             % Polar-to-Cartesian conversion

r2 = sqrt(rand(2*n,1));
t2 = [pi/2*rand(n,1)+pi/2; (pi/2*rand(n,1)-pi/2)]; % Random angles for Q2 and Q4
X2 = [r2.*cos(t2) r2.*sin(t2)];
```

Define the predictor variables. Label points in the first and third quadrants as belonging to the positive class, and those in the second and fourth quadrants in the negative class.

```
pred = [X1; X2];
resp = ones(4*n,1);
resp(2*n + 1:end) = -1; % Labels
```

Create the function `mysigmoid.m` , which accepts two matrices in the feature space as inputs, and transforms them into a Gram matrix using the sigmoid kernel.

```
function G = mysigmoid(U,V)
% Sigmoid kernel function with slope gamma and intercept c
gamma = 1;
c = -1;
G = tanh(gamma*U*V' + c);
end
```

Train an SVM classifier using the sigmoid kernel function. It is good practice to standardize the data.

```
SVMModel1 = fitcsvm(pred,resp,'KernelFunction','mysigmoid',...
                'Standardize',true);
SVMModel1 = fitPosterior(SVMModel1);
[~,scores1] = resubPredict(SVMModel1);
```

Set `gamma = 0.5` ; within `mysigmoid.m` and save as `mysigmoid2.m`. And, train an SVM classifier using the adjusted sigmoid kernel.

```
function G = mysigmoid2(U,V)
% Sigmoid kernel function with slope gamma and intercept c
gamma = 0.5;
c = -1;
G = tanh(gamma*U*V' + c);
end
```

```
SVMModel2 = fitcsvm(pred,resp,'KernelFunction','mysigmoid2',...
                'Standardize',true);
```

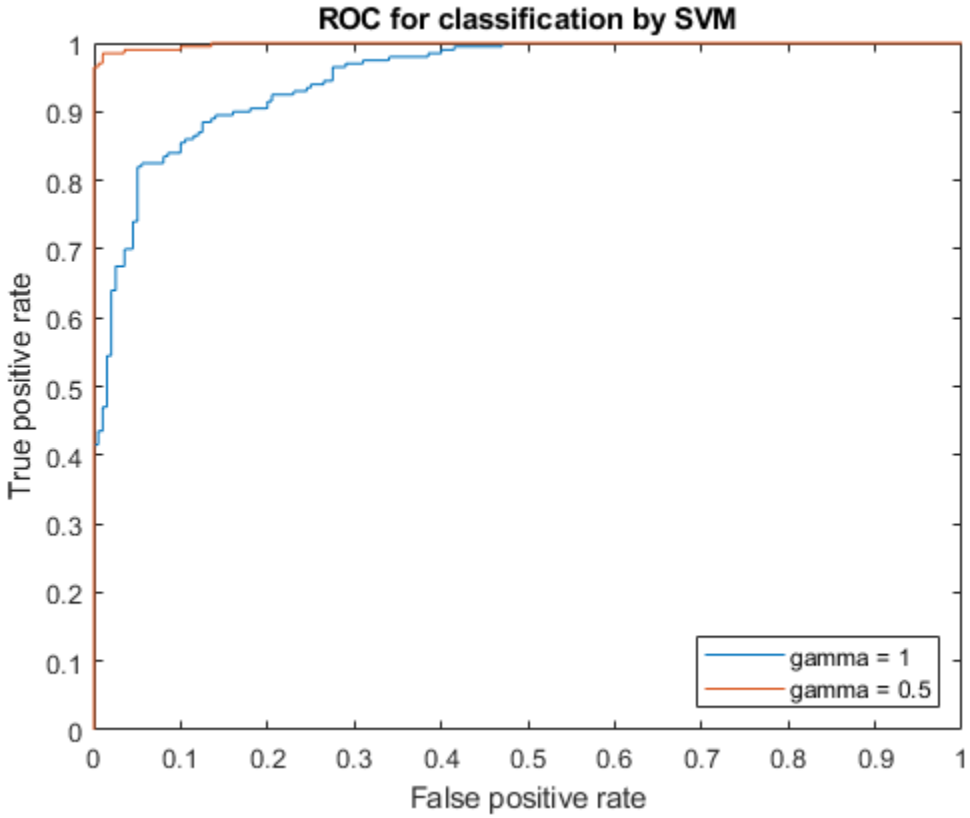
```
SVMModel2 = fitPosterior(SVMModel2);
[~,scores2] = resubPredict(SVMModel2);
```

Compute the ROC curves and the area under the curve (AUC) for both models.

```
[x1,y1,~,auc1] = perfcurve(resp,scores1(:,2),1);
[x2,y2,~,auc2] = perfcurve(resp,scores2(:,2),1);
```

Plot the ROC curves.

```
plot(x1,y1)
hold on
plot(x2,y2)
hold off
legend('gamma = 1','gamma = 0.5','Location','SE');
xlabel('False positive rate'); ylabel('True positive rate');
title('ROC for classification by SVM');
```



The kernel function with the gamma parameter set to 0.5 gives better in-sample results.

Compare the AUC measures.

```
auc1
auc2
```

```
auc1 =
```

```
    0.9518

auc2 =

    0.9985
```

The area under the curve for gamma set to 0.5 is higher than that for gamma set to 1. This also confirms that gamma parameter value of 0.5 produces better results. For visual comparison of the classification performance with these two gamma parameter values, see “Train SVM Classifier Using Custom Kernel” on page 18-159.

Plot ROC Curve for Classification Tree

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: sepal length, sepal width, petal length, and petal width. All measures are in centimeters.

Train a classification tree using the sepal length and width as the predictor variables. It is a good practice to specify the class names.

```
Model = fitctree(meas(:,1:2),species, ...
    'ClassNames',{'setosa','versicolor','virginica'});
```

Predict the class labels and scores for the species based on the tree `Model`.

```
[~,score] = resubPredict(Model);
```

The scores are the posterior probabilities that an observation (a row in the data matrix) belongs to a class. The columns of `score` correspond to the classes specified by `'ClassNames'`. So, the first column corresponds to setosa, the second corresponds to versicolor, and the third column corresponds to virginica.

Compute the ROC curve for the predictions that an observation belongs to versicolor, given the true class labels `species`. Also compute the optimal operating point and *y* values for negative subclasses. Return the names of the negative classes.

Because this is a multiclass problem, you cannot merely supply `score(:,2)` as input to `perfcurve`. Doing so would not give `perfcurve` enough information about the scores for the two negative classes (setosa and virginica). This problem is unlike a binary classification problem, where knowing the scores of one class is enough to determine the scores of the other class. Therefore, you must supply `perfcurve` with a function that factors in the scores of the two negative classes. One such function is $score(:,2) - \max(score(:,1), score(:,3))$.

```
diffscore = score(:,2) - max(score(:,1),score(:,3));
[X,Y,T,~,OPTROCPT,suby,subnames] = perfcurve(species,diffscore,'versicolor');
```

`X`, by default, is the false positive rate (fallout or 1-specificity) and `Y`, by default, is the true positive rate (recall or sensitivity). The positive class label is `versicolor`. Because a negative class is not

defined, `perfcurve` assumes that the observations that do not belong to the positive class are in one class. The function accepts it as the negative class.

`OPTROCPT`

```
OPTROCPT = 1×2

    0.1000    0.8000
```

`suby`

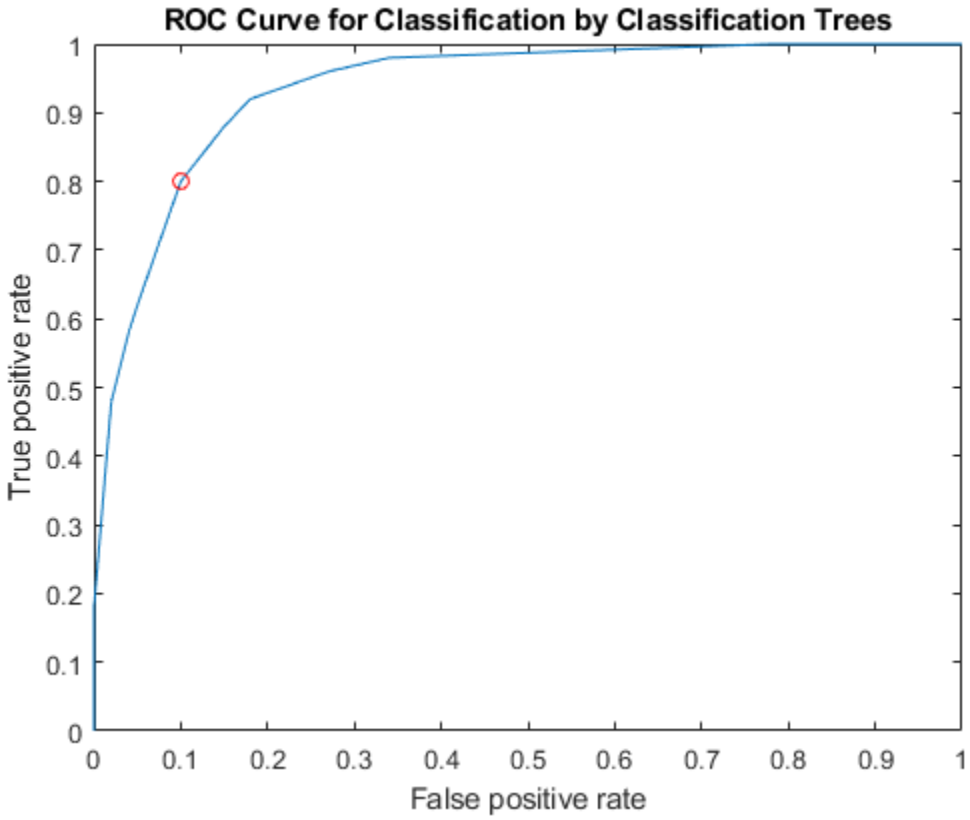
```
suby = 12×2

         0         0
    0.1800    0.1800
    0.4800    0.4800
    0.5800    0.5800
    0.6200    0.6200
    0.8000    0.8000
    0.8800    0.8800
    0.9200    0.9200
    0.9600    0.9600
    0.9800    0.9800
      ⋮
```

`subnames`

```
subnames = 1x2 cell
    {'setosa'}    {'virginica'}
```

Plot the ROC curve and the optimal operating point on the ROC curve.

```
plot(X,Y)
hold on
plot(OPTROCPT(1),OPTROCPT(2),'ro')
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve for Classification by Classification Trees')
hold off
```



Find the threshold that corresponds to the optimal operating point.

```
T((X==OPTROCPT(1))&(Y==OPTROCPT(2)))
```

```
ans = 0.2857
```

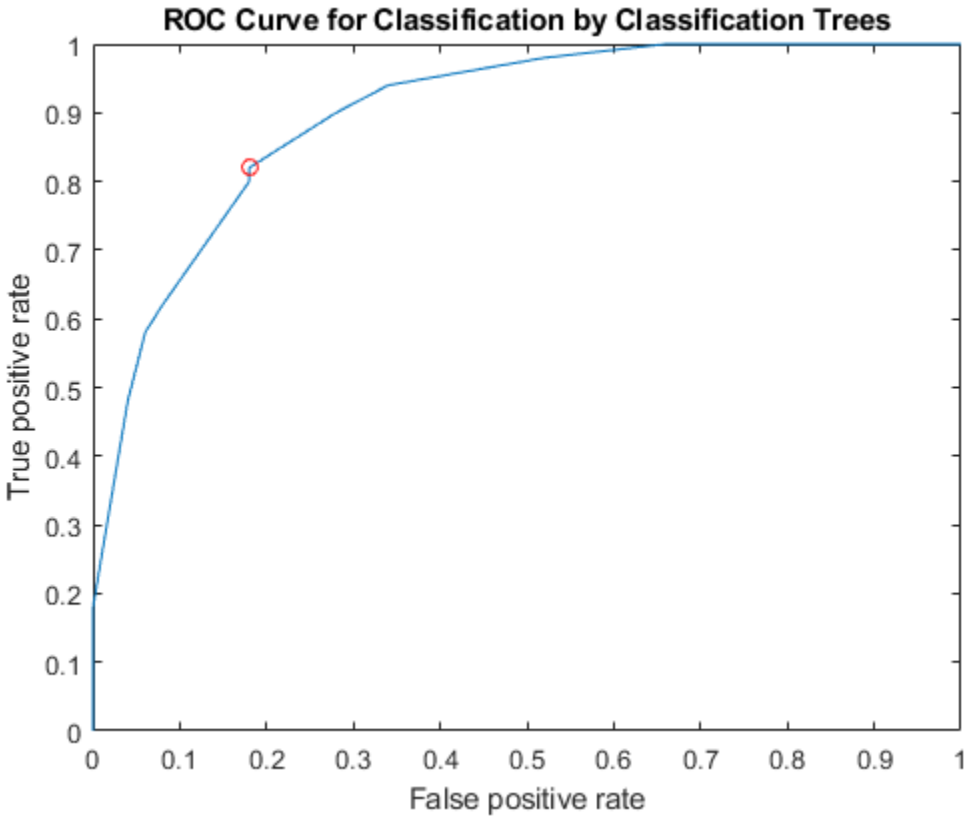
Specify `virginica` as the negative class and compute and plot the ROC curve for `versicolor`.

Again, you must supply `perfcurve` with a function that factors in the scores of the negative class. An example of a function to use is *score*(:, 2) − *score*(:, 3).

```
diffscore = score(:,2) - score(:,3);
[X,Y,~,~,OPTROCPT] = perfcurve(species,diffscore,'versicolor', ...
    'negClass','virginica');
OPTROCPT
```

```
OPTROCPT = 1×2

    0.1800    0.8200
```

```
figure, plot(X,Y)
hold on
plot(OPTROCPT(1),OPTROCPT(2),'ro')
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve for Classification by Classification Trees')
hold off
```



Compute Pointwise Confidence Intervals for ROC Curve

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: sepal length, sepal width, petal length, and petal width. All measures are in centimeters.

Use only the first two features as predictor variables. Define a binary problem by using only the measurements that correspond to the versicolor and virginica species.

```
pred = meas(51:end,1:2);
```

Define the binary response variable.

```
resp = (1:100)'>50;  % Versicolor = 0, virginica = 1
```

Fit a logistic regression model.

```
mdl = fitglm(pred,resp,'Distribution','binomial','Link','logit');
```

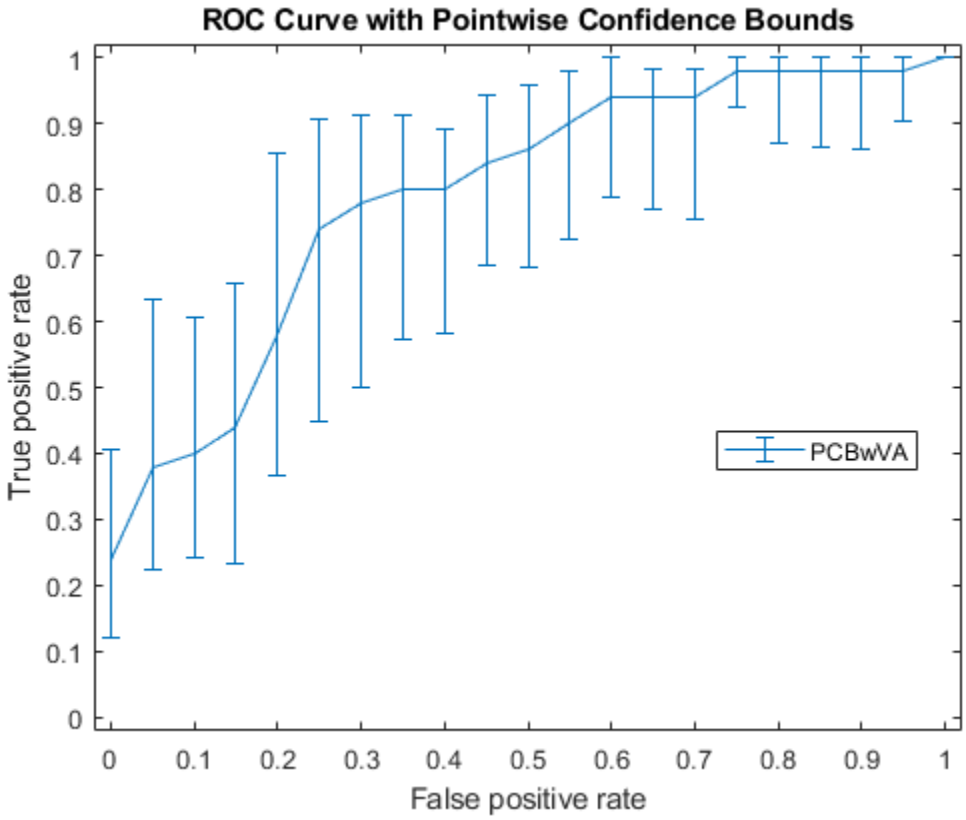
Compute the pointwise confidence intervals on the true positive rate (TPR) by vertical averaging (VA) and sampling using bootstrap.

```
[X,Y,T] = perfcurve(species(51:end,:),mdl.Fitted.Probability,...
       'virginica','NBoot',1000,'XVals',[0:0.05:1]);
```

`'NBoot',1000` sets the number of bootstrap replicas to 1000. `'XVals','All'` prompts `perfcurve` to return X, Y, and T values for all scores, and average the Y values (true positive rate) at all X values (false positive rate) using vertical averaging. If you do not specify `XVals`, then `perfcurve` computes the confidence bounds using threshold averaging by default.

Plot the pointwise confidence intervals.

```
errorbar(X,Y(:,1),Y(:,1)-Y(:,2),Y(:,3)-Y(:,1));
xlim([-0.02,1.02]); ylim([-0.02,1.02]);
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve with Pointwise Confidence Bounds')
legend('PCBwVA','Location','Best')
```



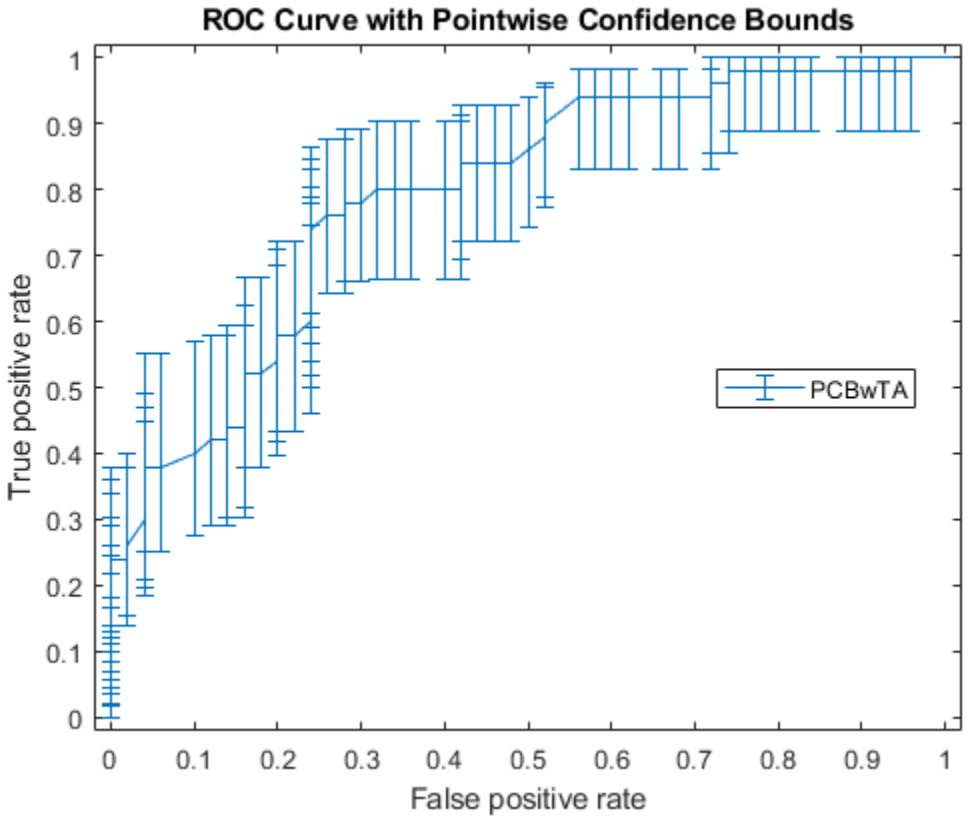
It might not always be possible to control the false positive rate (FPR, the X value in this example). So you might want to compute the pointwise confidence intervals on true positive rates (TPR) by threshold averaging.

```
[X1,Y1,T1] = perfcurve(species(51:end,:),mdl.Fitted.Probability,...
    'virginica','NBoot',1000);
```

If you set `'TVals'` to `'All'`, or if you do not specify `'TVals'` or `'Xvals'`, then `perfcurve` returns X, Y, and T values for all scores and computes pointwise confidence bounds for X and Y using threshold averaging.

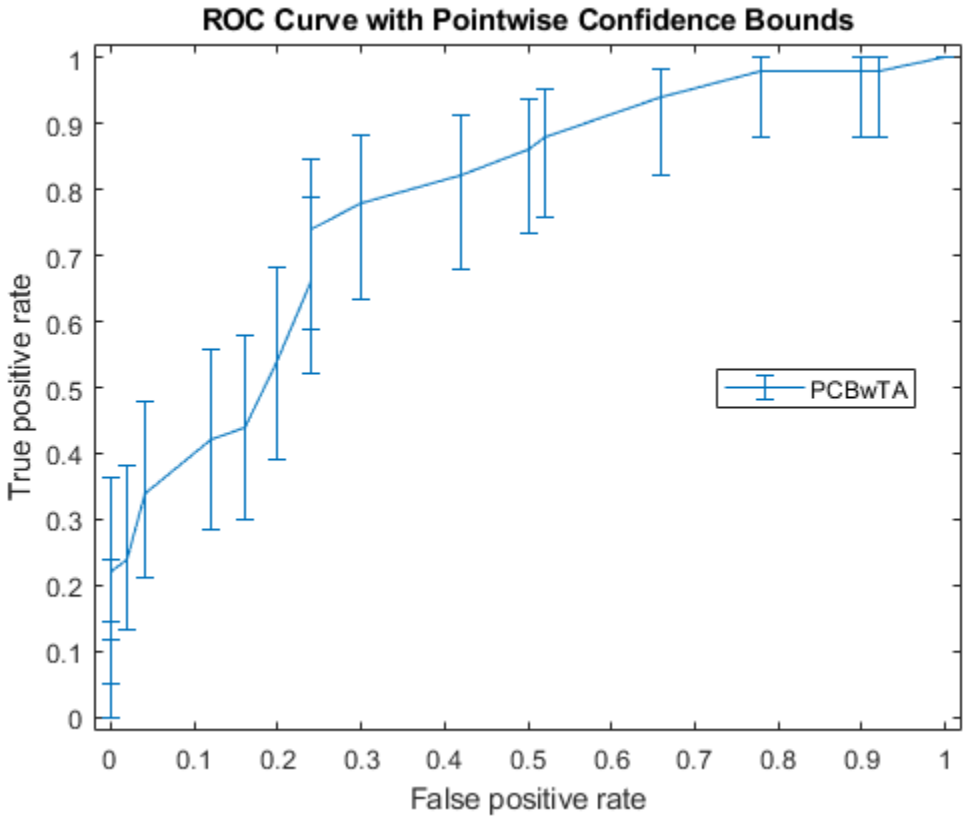
Plot the confidence bounds.

```
figure()
errorbar(X1(:,1),Y1(:,1),Y1(:,1)-Y1(:,2),Y1(:,3)-Y1(:,1));
xlim([-0.02,1.02]); ylim([-0.02,1.02]);
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve with Pointwise Confidence Bounds')
legend('PCBwTA','Location','Best')
```



Specify the threshold values to fix and compute the ROC curve. Then plot the curve.

```
[X1,Y1,T1] = perfcurve(species(51:end,:),mdl.Fitted.Probability,...
    'virginica','NBoot',1000,'TVals',0:0.05:1);
figure()
errorbar(X1(:,1),Y1(:,1),Y1(:,1)-Y1(:,2),Y1(:,3)-Y1(:,1));
xlim([-0.02,1.02]); ylim([-0.02,1.02]);
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve with Pointwise Confidence Bounds')
legend('PCBwTA','Location','Best')
```



Input Arguments

`labels` — True class labels

numeric vector | logical vector | character matrix | string array | cell array of character vectors | categorical array

True class labels, specified as a numeric vector, logical vector, character matrix, string array, cell array of character vectors, or categorical array. For more information, see “Grouping Variables” on page 2-45.

Example: `{'hi','mid','hi','low',...,'mid'}`

Example: `['H','M','H','L',...,'M']`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`scores` — Scores returned by a classifier

vector of floating points

Scores returned by a classifier for some sample data, specified as a vector of floating points. `scores` must have the same number of elements as `labels`.

Data Types: `single` | `double`

`posclass` — Positive class label

numeric scalar | logical scalar | character vector | string scalar | cell containing a character vector | categorical scalar

Positive class label, specified as a numeric scalar, logical scalar, character vector, string scalar, cell containing a character vector, or categorical scalar. The positive class must be a member of the input labels. The value of `posclass` that you can specify depends on the value of `labels`.

labels value	**posclass value**
Numeric vector	Numeric scalar
Logical vector	Logical scalar
Character matrix	Character vector
String array	String scalar
Cell array of character vectors	Character vector or cell containing character vector
Categorical vector	Categorical scalar

For example, in a cancer diagnosis problem, if a malignant tumor is the positive class, then specify `posclass` as `'malignant'`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NegClass','versicolor','XCrit','fn','NBoot',1000,'BootType','per'` specifies the species versicolor as the negative class, the criterion for the X-coordinate as false negative, the number of bootstrap samples as 1000. It also specifies that the pointwise confidence bounds are computed using the percentile method.

`NegClass` — List of negative classes

`'all'` (default) | numeric array | categorical array | string array | cell array of character vectors

List of negative classes, specified as the comma-separated pair consisting of `'NegClass'`, and a numeric array, a categorical array, a string array, or a cell array of character vectors. By default, `perfcurve` sets `NegClass` to `'all'` and considers all nonpositive classes found in the input array of labels to be negative.

If `NegClass` is a subset of the classes found in the input array of labels, then `perfcurve` discards the instances with labels that do not belong to either positive or negative classes.

Example: `'NegClass',{'versicolor','setosa'}`

Data Types: `single` | `double` | `categorical` | `char` | `string` | `cell`

`XCrit` — Criterion to compute for `X`

`'fpr'` (default) | `'fnr'` | `'tnr'` | `'ppv'` | `'ecost'` | ...

Criterion to compute for `X`, specified as the comma-separated pair consisting of `'XCrit'` and one of the following.

Criterion	**Description**
`tp`	Number of true positive instances

Criterion	Description
`fn`	Number of false negative instances.
`fp`	Number of false positive instances.
`tn`	Number of true negative instances.
`tp+fp`	Sum of true positive and false positive instances.
`rpp`	Rate of positive predictions. `rpp = (tp+fp)/(tp+fn+fp+tn)`
`rnp`	Rate of negative predictions. `rnp = (tn+fn)/(tp+fn+fp+tn)`
`accu`	Accuracy. `accu = (tp+tn)/(tp+fn+fp+tn)`
`tpr`, or `sens`, or `reca`	True positive rate, or sensitivity, or recall. `tpr= sens = reca = tp/(tp+fn)`
`fnr`, or `miss`	False negative rate, or miss. `fnr = miss = fn/(tp+fn)`
`fpr`, or `fall`	False positive rate, or fallout, or 1 – specificity. `fpr = fall = fp/(tn+fp)`
`tnr`, or `spec`	True negative rate, or specificity. `tnr = spec = tn/(tn+fp)`
`ppv`, or `prec`	Positive predictive value, or precision. `ppv = prec = tp/(tp+fp)`
`npv`	Negative predictive value. `npv = tn/(tn+fn)`
`ecost`	Expected cost. `ecost = (tp*Cost(P|P)+fn*Cost(N|P)+fp* Cost(P|N)+tn*Cost(N|N))/(tp+fn+fp+tn)`
Custom criterion	A custom-defined function with the input arguments `(C,scale,cost)`, where `C` is a 2-by-2 confusion matrix, `scale` is a 2-by-1 array of class scales, and `cost` is a 2-by-2 misclassification cost matrix.

Caution Some of these criteria return `NaN` values at one of the two special thresholds, `'reject all'` and `'accept all'`.

Example: `'XCrit','ecost'`

`YCrit` — Criterion to compute for `Y`
`'tpr'` (default) | same criteria options for `X`

Criterion to compute for `Y`, specified as the comma-separated pair consisting of `'YCrit'` and one of the same criteria options as for `X`. This criterion does not have to be a monotone function of the positive class score.

Example: `'YCrit','ecost'`

`XVals` — Values for the X criterion

`'all'` (default) | numeric array

Values for the X criterion, specified as the comma-separated pair consisting of `'XVals'` and a numeric array.

- If you specify `XVals`, then `perfcurve` computes X and Y and the pointwise confidence bounds on page 33-4537 for Y (when applicable) only for the specified `XVals`.
- If you do not specify `XVals`, then `perfcurve`, computes X and Y and the values for all scores by default.

Note You cannot set `XVals` and `TVals` at the same time.

Example: `'XVals',[0:0.05:1]`

Data Types: `single` | `double` | `char` | `string`

`TVals` — Thresholds for the positive class score

`'all'` (default) | numeric array

Thresholds for the positive class score, specified as the comma-separated pair consisting of `'TVals'` and either `'all'` or a numeric array.

- If `TVals` is set to `'all'` or not specified, and `XVals` is not specified, then `perfcurve` returns X, Y, and T values for all scores and computes pointwise confidence bounds on page 33-4537 for X and Y using threshold averaging.
- If `TVals` is set to a numeric array, then `perfcurve` returns X, Y, and T values for the specified thresholds and computes pointwise confidence bounds for X and Y at these thresholds using threshold averaging.

Note You cannot set `XVals` and `TVals` at the same time.

Example: `'TVals',[0:0.05:1]`

Data Types: `single` | `double` | `char` | `string`

`UseNearest` — Indicator to use the nearest values in the data

`'on'` (default) | `'off'`

Indicator to use the nearest values in the data instead of the specified numeric `XVals` or `TVals`, specified as the comma-separated pair consisting of `'UseNearest'` and either `'on'` or `'off'`.

- If you specify numeric `XVals` and set `UseNearest` to `'on'`, then `perfcurve` returns the nearest unique X values found in the data, and it returns the corresponding values of Y and T.
- If you specify numeric `XVals` and set `UseNearest` to `'off'`, then `perfcurve` returns the sorted `XVals`.
- If you compute confidence bounds by cross validation or bootstrap, then this parameter is always `'off'`.

Example: `'UseNearest','off'`

`ProcessNaN` — `perfcurve` method for processing `NaN` scores
`'ignore'` (default) | `'addtofalse'`

`perfcurve` method for processing `NaN` scores, specified as the comma-separated pair consisting of `'ProcessNaN'` and `'ignore'` or `'addtofalse'`.

- If `ProcessNaN` is `'ignore'`, then `perfcurve` removes observations with `NaN` scores from the data.
- If `ProcessNaN` is `'addtofalse'`, then `perfcurve` adds instances with `NaN` scores to false classification counts in the respective class. That is, `perfcurve` always counts instances from the positive class as false negative (FN), and it always counts instances from the negative class as false positive (FP).

Example: `'ProcessNaN','addtofalse'`

`Prior` — Prior probabilities for positive and negative classes
`'empirical'` (default) | `'uniform'` | array with two elements

Prior probabilities for positive and negative classes, specified as the comma-separated pair consisting of `'Prior'` and `'empirical'`, `'uniform'`, or an array with two elements.

If `Prior` is `'empirical'`, then `perfcurve` derives prior probabilities from class frequencies.

If `Prior` is `'uniform'` , then `perfcurve` sets all prior probabilities to be equal.

Example: `'Prior',[0.3,0.7]`

Data Types: `single` | `double` | `char` | `string`

`Cost` — Misclassification costs
`[0 0.5;0.5 0]` (default) | 2-by-2 matrix

Misclassification costs, specified as the comma-separated pair consisting of `'Cost'` and a 2-by-2 matrix, containing `[Cost(P|P),Cost(N|P);Cost(P|N),Cost(N|N)]`.

`Cost(N|P)` is the cost of misclassifying a positive class as a negative class. `Cost(P|N)` is the cost of misclassifying a negative class as a positive class. Usually, `Cost(P|P)` = 0 and `Cost(N|N)` = 0, but `perfcurve` allows you to specify nonzero costs for correct classification as well.

Example: `'Cost',[0 0.7;0.3 0]`

Data Types: `single` | `double`

`Alpha` — Significance level
0.05 (default) | scalar value in the range 0 through 1

Significance level for the confidence bounds, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 through 1. `perfcurve` computes 100*(1 – α) percent pointwise confidence bounds on page 33-4537 for `X`, `Y`, `T`, and `AUC` for a confidence level of 1 – α.

Example: `'Alpha',0.01` specifies 99% confidence bounds

Data Types: `single` | `double`

`Weights` — Observation weights
vector of nonnegative scalar values | cell array of vectors of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of nonnegative scalar values. This vector must have as many elements as `scores` or `labels` do.

If `scores` and `labels` are in cell arrays and you need to supply `Weights`, the weights must be in a cell array as well. In this case, every element in `Weights` must be a numeric vector with as many elements as the corresponding element in `scores`. For example, `numel(weights{1}) == numel(scores{1})`.

When `perfcurve` computes the `X`, `Y` and `T` or confidence bounds using cross-validation, it uses these observation weights instead of observation counts.

When `perfcurve` computes confidence bounds using bootstrap, it samples *N* out of *N* observations with replacement, using these weights as multinomial sampling probabilities.

The default is a vector of 1s or a cell array in which each element is a vector of 1s.

Data Types: `single` | `double` | `cell`

`NBoot` — Number of bootstrap replicas
0 (default) | positive integer

Number of bootstrap replicas for computation of confidence bounds, specified as the comma-separated pair consisting of `'NBoot'` and a positive integer. The default value 0 means the confidence bounds are not computed.

If `labels` and `scores` are cell arrays, this parameter must be 0 because `perfcurve` can use either cross-validation or bootstrap to compute confidence bounds.

Example: `'NBoot',500`

Data Types: `single` | `double`

`BootType` — Confidence interval type for `bootci`
`'bca'` (default) | `'norm` | `'per'` | `'cper'` | `'stud'`

Confidence interval type for `bootci` to use to compute confidence bounds, specified as the comma-separated pair consisting of `'BootType'` and one of the following:

- `'bca'` — Bias corrected and accelerated percentile method
- `'norm` or `'normal'` — Normal approximated interval with bootstrapped bias and standard error
- `'per'` or `'percentile'` — Percentile method
- `'cper'` or `'corrected percentile'` — Bias corrected percentile method
- `'stud'` or `'student'` — Studentized confidence interval

Example: `'BootType','cper'`

`BootArg` — Optional input arguments for `bootci`
`[ ]` (default) | `{'Nbootstd',nbootstd}` | `{'Stderr',stderr}`

Optional input arguments for `bootci` to compute confidence bounds, specified as the comma-separated pair consisting of `'BootArg'` and `{'Nbootstd',nbootstd}` or `{'Stderr',stderr}`, name-value pair arguments of `bootci`.

When you compute the studentized bootstrap confidence intervals (`'BootType'` is `'student'`), you can additionally specify `'Nbootstd'` or `'Stderr'`, name-value pair arguments of `bootci`, by using `'BootArg'`.

- `'BootArg',{'Nbootstd',nbootstd}` estimates the standard error of the bootstrap statistics using bootstrap with `nbootstd` data samples. `nbootstd` is a positive integer and its default is 100.
- `'BootArg',{'Stderr',stderr}` evaluates the standard error of the bootstrap statistics by a user-defined function `stderr` that takes `[1:numel(scores)]` as an input argument. `stderr` is a function handle.

Example: `'BootArg',{'Nbootstd',nbootstd}`

Data Types: `cell`

`Options` — Options for controlling the computation of confidence intervals
`[]` (default) | structure array returned by `statset`

Options for controlling the computation of confidence intervals, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. These options require Parallel Computing Toolbox. `perfcurve` uses this argument for computing pointwise confidence bounds only. To compute these bounds, you must pass cell arrays for `labels` and `scores` or set `NBoot` to a positive integer.

This table summarizes the available options.

Option	Description
`'UseParallel'`	• `false` — Serial computation (default). • `true` — Parallel computation. You need Parallel Computing Toolbox for this option to work.
`'UseSubstreams'`	• `false` — Do not use a separate substream for each iteration (default). • `true` — Use a separate substream for each iteration to compute in parallel in a reproducible fashion. To compute reproducibly, set `Streams` to a type allowing substreams: `'mlfg6331_64'` or `'mrg32k3a'`.
`'Streams'`	A `RandStream` object, or a cell array of such objects. If you specify `Streams`, use a single object, except when: • `UseParallel` is `true`. • `UseSubstreams` is `false`. In that case, use a cell array of the same size as the parallel pool. If a parallel pool is not open, then `Streams` must supply a single random number stream.

If `'UseParallel'` is `true` and `'UseSubstreams'` is `false`, then the length of `'Streams'` must equal the number of workers used by `perfcurve`. If a parallel pool is already open, then the length of `'Streams'` is the size of the parallel pool. If a parallel pool is not already open, then MATLAB might open a pool for you, depending on your installation and preferences. To ensure more predictable results, use `parpool` and explicitly create a parallel pool before invoking `perfcurve` and setting `'Options',statset('UseParallel',true)`.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Output Arguments

X — *x*-coordinates for the performance curve

vector, `fpr` (default) | *m*-by-3 matrix

x-coordinates for the performance curve, returned as a vector or an *m*-by-3 matrix. By default, `X` values are the false positive rate, FPR (fallout or 1 – specificity). To change `X`, use the `XCrit` name-value pair argument.

- If `perfcurve` does not compute the pointwise confidence bounds on page 33-4537, or if it computes them using vertical averaging, then `X` is a vector.
- If `perfcurve` computes the confidence bounds using threshold averaging, then `X` is an *m*-by-3 matrix, where *m* is the number of fixed threshold values. The first column of `X` contains the mean value. The second and third columns contain the lower bound and the upper bound, respectively, of the pointwise confidence bounds.

Y — *y*-coordinates for the performance curve

vector, `tpr` (default) | *m*-by-3 matrix

y-coordinates for the performance curve, returned as a vector or an *m*-by-3 matrix. By default, `Y` values are the true positive rate, TPR (recall or sensitivity). To change `Y`, use `YCrit` name-value pair argument.

- If `perfcurve` does not compute the pointwise confidence bounds on page 33-4537, then `Y` is a vector.
- If `perfcurve` computes the confidence bounds, then `Y` is an *m*-by-3 matrix, where *m* is the number of fixed `X` values or thresholds (`T` values). The first column of `Y` contains the mean value. The second and third columns contain the lower bound and the upper bound, respectively, of the pointwise confidence bounds.

T — Thresholds on classifier scores

vector | *m*-by-3 matrix

Thresholds on classifier scores for the computed values of `X` and `Y`, returned as a vector or *m*-by-3 matrix.

- If `perfcurve` does not compute the pointwise confidence bounds on page 33-4537, or computes them using threshold averaging, then `T` is a vector.
- If `perfcurve` computes the confidence bounds using vertical averaging, `T` is an *m*-by-3 matrix, where *m* is the number of fixed `X` values. The first column of `T` contains the mean value. The second and third columns contain the lower bound, and the upper bound, respectively, of the pointwise confidence bounds.

For each threshold, TP is the count of true positive observations with scores greater than or equal to this threshold, and FP is the count of false positive observations with scores greater than or equal to this threshold. perfcurve defines negative counts, TN and FN, in a similar way. The function then sorts the thresholds in the descending order that corresponds to the ascending order of positive counts.

For the *m* distinct thresholds found in the array of scores, perfcurve returns the X, Y and T arrays with *m* + 1 rows. perfcurve sets elements T(2:*m*+1) to the distinct thresholds, and T(1) replicates T(2). By convention, T(1) represents the highest 'reject all' threshold, and perfcurve computes the corresponding values of X and Y for TP = 0 and FP = 0. The T(end) value is the lowest 'accept all' threshold for which TN = 0 and FN = 0.

AUC — Area under the curve
scalar value | 3-by-1 vector

Area under the curve (AUC) for the computed values of X and Y, returned as a scalar value or a 3-by-1 vector.

- If perfcurve does not compute the pointwise confidence bounds on page 33-4537, AUC is a scalar value.
- If perfcurve computes the confidence bounds using vertical averaging, AUC is a 3-by-1 vector. The first column of AUC contains the mean value. The second and third columns contain the lower bound and the upper bound, respectively, of the confidence bound.

For a perfect classifier, AUC = 1. For a classifier that randomly assigns observations to classes, AUC = 0.5.

If you set XVals to 'all' (default), then perfcurve computes AUC using the returned X and Y values.

If XVals is a numeric array, then perfcurve computes AUC using X and Y values from all distinct scores in the interval, which are specified by the smallest and largest elements of XVals. More precisely, perfcurve finds X values for all distinct thresholds as if XVals were set to 'all', and then uses a subset of these (with corresponding Y values) between min(XVals) and max(XVals) to compute AUC.

perfcurve uses trapezoidal approximation to estimate the area. If the first or last value of X or Y are NaNs, then perfcurve removes them to allow calculation of AUC. This takes care of criteria that produce NaNs for the special 'reject all' or 'accept all' thresholds, for example, positive predictive value (PPV) or negative predictive value (NPV).

OPTROCPT — Optimal operating point of the ROC curve
1-by-2 array

Optimal operating point of the ROC curve, returned as a 1-by-2 array with false positive rate (FPR) and true positive rate (TPR) values for the optimal ROC operating point.

perfcurve computes OPTROCPT for the standard ROC curve only, and sets to NaNs otherwise. To obtain the optimal operating point for the ROC curve, perfcurve first finds the slope, *S*, using

$$S = \frac{\text{Cost}(P|N) - \text{Cost}(N|N)}{\text{Cost}(N|P) - \text{Cost}(P|P)} * \frac{N}{P}$$

- Cost(*N*|*P*) is the cost of misclassifying a positive class as a negative class. Cost(*P*|*N*) is the cost of misclassifying a negative class as a positive class.

- $P = TP + FN$ and $N = TN + FP$. They are the total instance counts in the positive and negative class, respectively.

`perfcurve` then finds the optimal operating point by moving the straight line with slope S from the upper left corner of the ROC plot (`FPR = 0`, `TPR = 1`) down and to the right, until it intersects the ROC curve.

`SUBY` — Values for negative subclasses
array

Values for negative subclasses, returned as an array.

- If you specify only one negative class, then `SUBY` is identical to `Y`.
- If you specify k negative classes, then `SUBY` is a matrix of size m-by-k, where m is the number of returned values for `X` and `Y`, and k is the number of negative classes. `perfcurve` computes `Y` values by summing counts over all negative classes.

`SUBY` gives values of the `Y` criterion for each negative class separately. For each negative class, `perfcurve` places a new column in `SUBY` and fills it with `Y` values for true negative (TN) and false positive (FP) counted just for this class.

`SUBYNAMES` — Negative class names
cell array

Negative class names, returned as a cell array.

- If you provide an input array of negative class names, , `NegClass`, then `perfcurve` copies names into `SUBYNAMES`.
- If you do not provide `NegClass`, then `perfcurve` extracts `SUBYNAMES` from the input labels. The order of `SUBYNAMES` is the same as the order of columns in `SUBY`. That is, `SUBY(:,1)` is for negative class `SUBYNAMES{1}`, `SUBY(:,2)` is for negative class `SUBYNAMES{2}`, and so on.

Algorithms

Pointwise Confidence Bounds

If you supply cell arrays for `labels` and `scores`, or if you set `NBoot` to a positive integer, then `perfcurve` returns pointwise confidence bounds for `X`,`Y`,`T`, and `AUC`. You cannot supply cell arrays for `labels` and `scores` and set `NBoot` to a positive integer at the same time.

`perfcurve` resamples data to compute confidence bounds using either cross validation or bootstrap.

- Cross-validation — If you supply cell arrays for `labels` and `scores`, then `perfcurve` uses cross-validation and treats elements in the cell arrays as cross-validation folds. `labels` can be a cell array of numeric vectors, logical vectors, character matrices, cell arrays of character vectors, or categorical vectors. All elements in `labels` must have the same type. `scores` can be a cell array of numeric vectors. The cell arrays for `labels` and `scores` must have the same number of elements. The number of labels in cell j of `labels` must be equal to the number of scores in cell j of `scores` for any j in the range from 1 to the number of elements in `scores`.
- Bootstrap — If you set `NBoot` to a positive integer n, `perfcurve` generates n bootstrap replicas to compute pointwise confidence bounds. If you use `XCrit` or `YCrit` to set the criterion for `X` or `Y` to an anonymous function, `perfcurve` can compute confidence bounds only using bootstrap.

`perfcurve` estimates the confidence bounds using one of two methods:

- Vertical averaging (VA) — `perfcurve` estimates confidence bounds on `Y` and `T` at fixed values of `X`. That is, `perfcurve` takes samples of the ROC curves for fixed `X` values, averages the corresponding `Y` and `T` values, and computes the standard errors. You can use the `XVals` name-value pair argument to fix the `X` values for computing confidence bounds. If you do not specify `XVals`, then `perfcurve` computes the confidence bounds at all `X` values.
- Threshold averaging (TA) — `perfcurve` takes samples of the ROC curves at fixed thresholds `T` for the positive class score, averages the corresponding `X` and `Y` values, and estimates the confidence bounds. You can use the `TVals` name-value pair argument to use this method for computing confidence bounds. If you set `TVals` to `'all'` or do not specify `TVals` or `XVals`, then `perfcurve` returns `X`, `Y`, and `T` values for all scores and computes pointwise confidence bounds for `Y` and `X` using threshold averaging.

When you compute the confidence bounds, `Y` is an *m*-by-3 array, where *m* is the number of fixed `X` values or thresholds (`T` values). The first column of `Y` contains the mean value. The second and third columns contain the lower bound and the upper bound, respectively, of the pointwise confidence bounds. `AUC` is a row vector with three elements, following the same convention. If `perfcurve` computes the confidence bounds using VA, then `T` is an *m*-by-3 matrix, and `X` is a column vector. If `perfcurve` uses TA, then `X` is an *m*-by-3 matrix and `T` is a column-vector.

`perfcurve` returns pointwise confidence bounds. It does not return a simultaneous confidence band for the entire curve.

References

[1] T. Fawcett. “ROC Graphs: Notes and Practical Considerations for Researchers”, 2004.

[2] Zweig, M., and G. Campbell. “Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine.” *Clin. Chem*. 1993, 39/4, pp. 561–577 .

[3] Davis, J., and M. Goadrich. “The Relationship Between Precision-Recall and ROC Curves.” *Proceedings of ICML ’06*, 2006, pp. 233–240.

[4] Moskowitz, C., and M. Pepe. “Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes.” *Biostatistics*, 2004, 5, pp. 113–127.

[5] Huang, Y., M. Pepe, and Z. Feng. “Evaluating the Predictiveness of a Continuous Marker.” *U. Washington Biostatistics Paper Series*, 2006, 250–261.

[6] Briggs, W., and R. Zaretzki. “The Skill Plot: A Graphical Technique for Evaluating Continuous Diagnostic Tests.” *Biometrics*, 2008, 63, pp. 250 – 261.

[7] R. Bettinger. “Cost-Sensitive Classifier Selection Using the ROC Convex Hull Method.” *SAS Institute*.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`bootci` | `classify` | `fitcnb` | `fitctree` | `fitrtree` | `glmfit` | `mnrfit`

Topics

“Performance Curves” on page 17-3

Introduced in R2009a

plot

Class: `clustering.evaluation.ClusterCriterion`
Package: `clustering.evaluation`

Plot clustering evaluation object criterion values

Syntax

```
plot(eva)
h = plot(eva)
```

Description

`plot(eva)` displays a plot of the criterion values versus the number of clusters, based on the values stored in the clustering evaluation object `eva`.

`h = plot(eva)` returns a handle to the plot line.

Input Arguments

eva — Clustering evaluation data

clustering evaluation object

Clustering evaluation data, specified as a clustering evaluation object. Create a clustering evaluation object using `evalclusters`.

Output Arguments

h — Handle to plot line

scalar value

Handle to the plot line, returned as a scalar value.

Examples

Plot the Clustering Evaluation Criterion Values

Plot the criterion values versus the number of clusters for each clustering solution stored in a clustering evaluation object.

Load the sample data.

```
load fisheriris
```

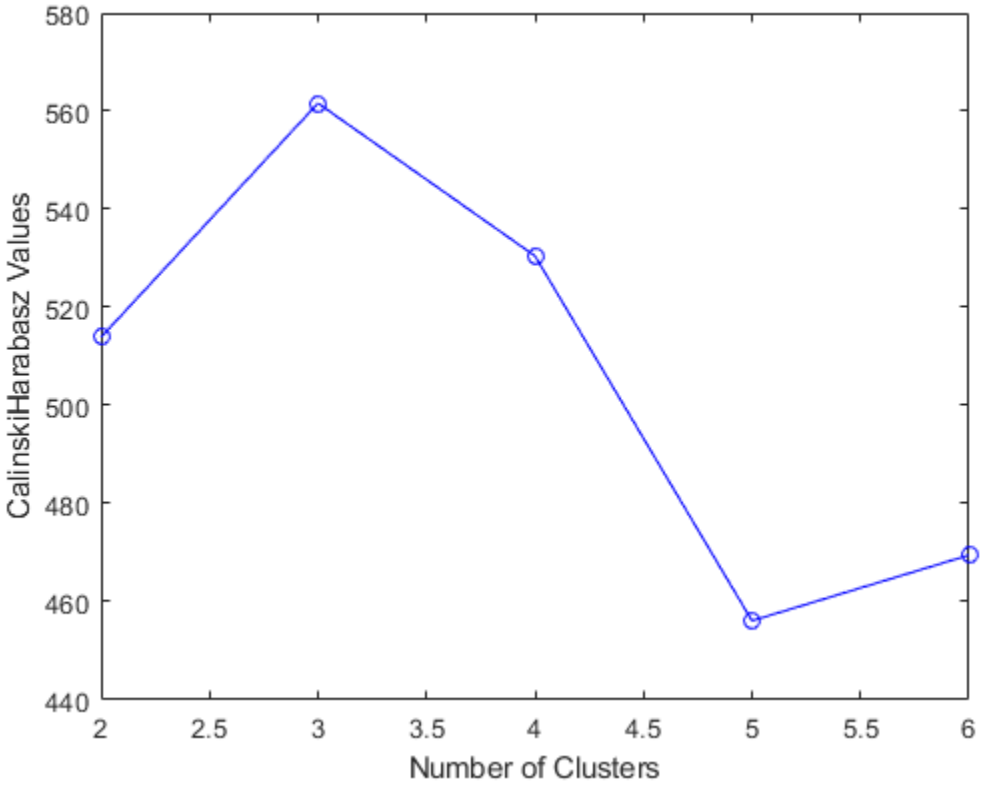
The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Create a clustering evaluation object. Cluster the data using `kmeans`, and evaluate the optimal number of clusters using the Calinski-Harabasz criterion.

```
rng('default');  % For reproducibility
eva = evalclusters(meas,'kmeans','CalinskiHarabasz','KList',[1:6]);
```

Plot the Calinski-Harabasz criterion values for each number of clusters tested.

```
figure;
plot(eva);
```



The plot shows that the highest Calinski-Harabasz value occurs at three clusters, suggesting that the optimal number of clusters is three.

See Also

`evalclusters`

plot

Plot results of local interpretable model-agnostic explanations (LIME)

Syntax

```
f = plot(results)
```

Description

`f = plot(results)` visualizes the LIME results in the `lime` object `results`. The function returns the `Figure` object `f`. Use `f` to query or modify Figure Properties of the figure after it is created.

- The figure contains a horizontal bar graph that shows the coefficient values of a linear simple model or predictor importance values of a decision tree simple model, depending on the simple model in `results` (`SimpleModel` property of `results`).
- The figure displays two predictions for the query point computed using the machine learning model and the simple model, respectively. These values correspond to the `BlackboxFitted` property and the `SimpleModelFitted` property of `results`.

Examples

Explain Prediction with Decision Tree Simple Model

Train a classification model and create a `lime` object that uses a decision tree simple model. When you create a `lime` object, specify a query point and the number of important predictors so that the software generates samples of a synthetic data set and fits a simple model for the query point with important predictors. Then display the estimated predictor importance in the simple model by using the object function `plot`.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    ______

    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

Create a table of predictor variables by removing the columns of customer IDs and ratings from `tbl`.

```
tblX = removevars(tbl,["ID","Rating"]);
```

Train a blackbox model of credit ratings by using the `fitcecoc` function.

```
blackbox = fitcecoc(tblX,tbl.Rating,'CategoricalPredictors','Industry');
```

Create a `lime` object that explains the prediction for the last observation using a decision tree simple model. Specify `'NumImportantPredictors'` as six to find at most 6 important predictors. If you specify the `'QueryPoint'` and `'NumImportantPredictors'` values when you create a `lime` object, then the software generates samples of a synthetic data set and fits a simple interpretable model to the synthetic data set.

```
queryPoint = tblX(end,:)
```

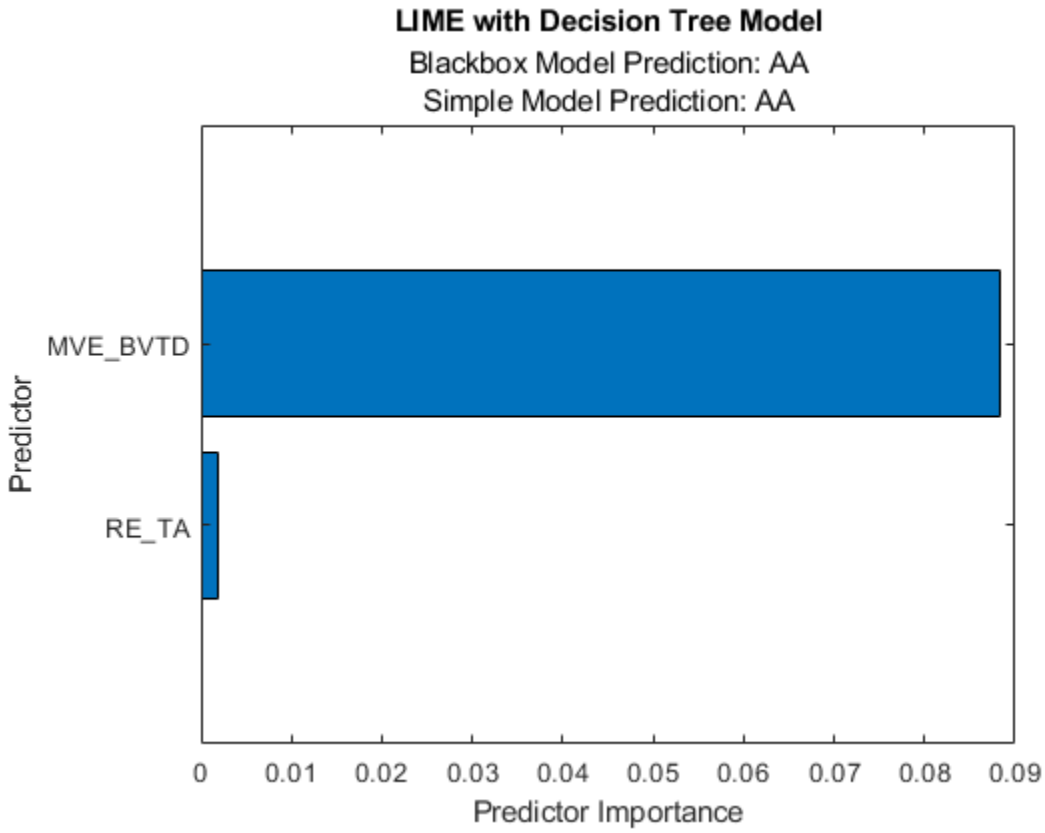
```
queryPoint=1×6 table
    WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA    Industry
    _____    _____    _______    ________    ____    ________

    0.239    0.463     0.065      2.924      0.34       2
```

```
rng('default') % For reproducibility
results = lime(blackbox,'QueryPoint',queryPoint,'NumImportantPredictors',6, ...
    'SimpleModelType','tree')
```

```
results =
  lime with properties:

             BlackboxModel: [1x1 ClassificationECOC]
              DataLocality: 'global'
     CategoricalPredictors: 6
                      Type: 'classification'
                         X: [3932x6 table]
                QueryPoint: [1x6 table]
    NumImportantPredictors: 6
          NumSyntheticData: 5000
             SyntheticData: [5000x6 table]
                    Fitted: {5000x1 cell}
               SimpleModel: [1x1 ClassificationTree]
       ImportantPredictors: [2x1 double]
            BlackboxFitted: {'AA'}
         SimpleModelFitted: {'AA'}
```

Plot the `lime` object `results` by using the object function `plot`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```

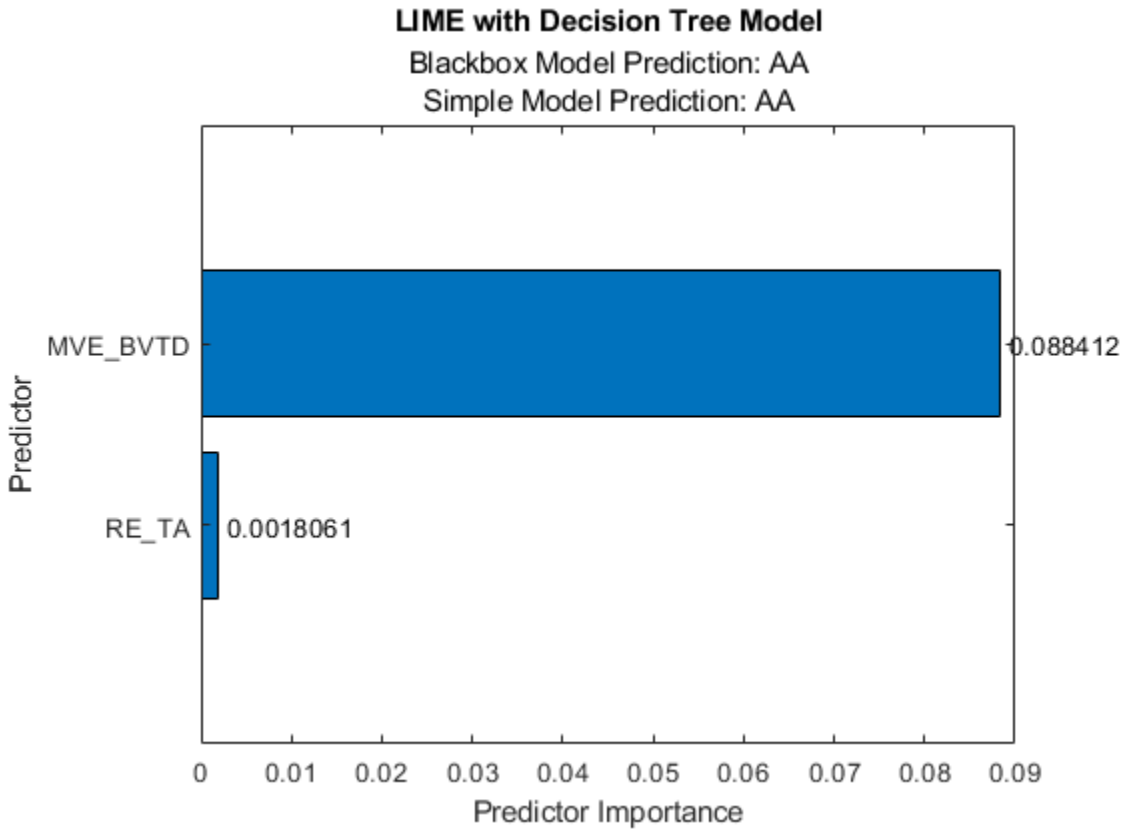


The plot displays two predictions for the query point, which correspond to the “BlackboxFitted” on page 33-0 property and the “SimpleModelFitted” on page 33-0 property of `results`.

The horizontal bar graph shows the sorted predictor importance values. `lime` finds the financial ratio variables `EBIT_TA` and `WC_TA` as important predictors for the query point.

You can read the bar lengths by using data tips or Bar Properties. For example, you can find `Bar` objects by using the `findobj` function and add labels to the ends of the bars by using the `text` function.

```
b = findobj(f,'Type','bar');
text(b.YEndPoints+0.001,b.XEndPoints,string(b.YData))
```



Alternatively, you can display the coefficient values in a table with the predictor variable names.

```
imp = b.YData;
flipud(array2table(imp', ...
    'RowNames',f.CurrentAxes.YTickLabel,'VariableNames',{'Predictor Importance'}))
```

```
ans=2×1 table
               Predictor Importance
               ____________________

    MVE_BVTD          0.088412
    RE_TA            0.0018061
```

Explain Prediction with Linear Simple Model

Train a regression model and create a `lime` object that uses a linear simple model. When you create a `lime` object, if you do not specify a query point and the number of important predictors, then the software generates samples of a synthetic data set but does not fit a simple model. Use the object function `fit` to fit a simple model for a query point. Then display the coefficients of the fitted linear simple model by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Create a table of predictor variables by removing the response variable from `tbl`.

```
tblX = removevars(tbl,'MPG');
```

Train a blackbox model of `MPG` by using the `fitrkernel` function.

```
rng('default') % For reproducibility
mdl = fitrkernel(tblX,tbl.MPG,'CategoricalPredictors',[2 5]);
```

Create a `lime` object. Specify a predictor data set because `mdl` does not contain predictor data.

```
results = lime(mdl,tblX)
```

```
results =
  lime with properties:

             BlackboxModel: [1x1 RegressionKernel]
              DataLocality: 'global'
     CategoricalPredictors: [2 5]
                      Type: 'regression'
                         X: [392x6 table]
                QueryPoint: []
    NumImportantPredictors: []
          NumSyntheticData: 5000
             SyntheticData: [5000x6 table]
                    Fitted: [5000x1 double]
               SimpleModel: []
       ImportantPredictors: []
            BlackboxFitted: []
          SimpleModelFitted: []
```

`results` contains the generated synthetic data set. The `SimpleModel` property is empty (`[]`).

Fit a linear simple model for the first observation in `tblX`. Specify the number of important predictors to find as 3.

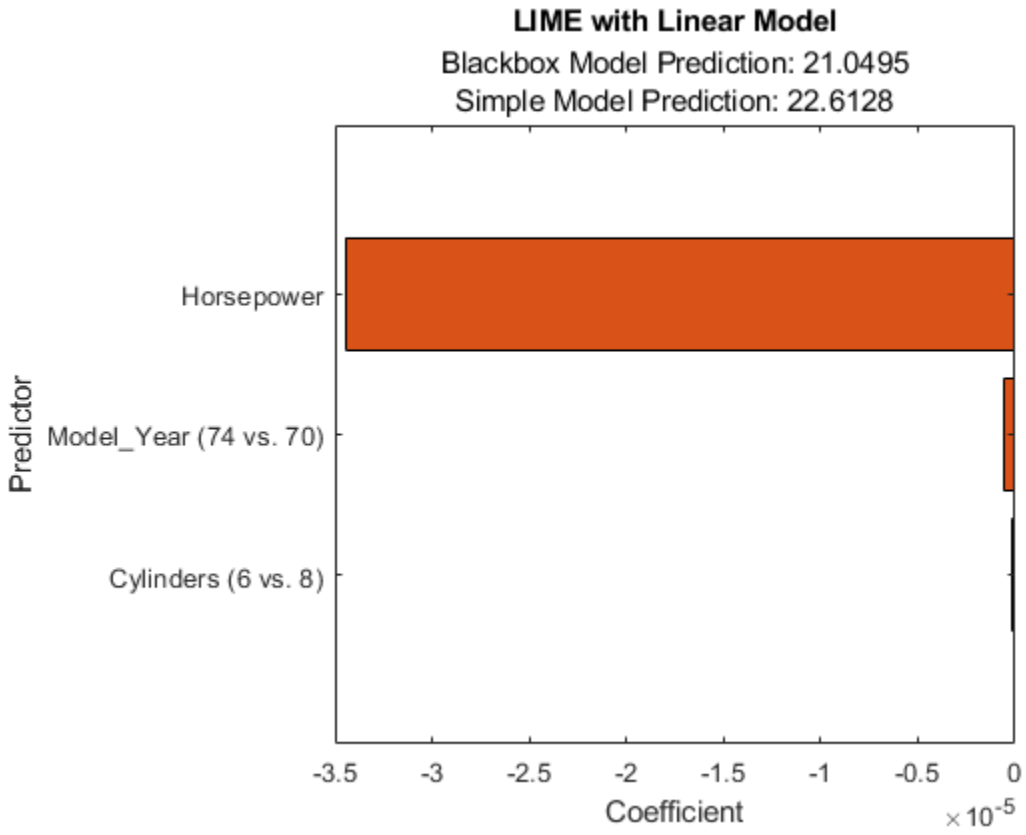
```
queryPoint = tblX(1,:)
```

```
queryPoint=1×6 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight
    ____________    _________    ____________    __________    __________    ______

         12             8            307            130            70         3504
```

```
results = fit(results,queryPoint,3);
```

Plot the `lime` object `results` by using the object function `plot`. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = plot(results);
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The plot displays two predictions for the query point, which correspond to the “BlackboxFitted” on page 33-0 property and the “SimpleModelFitted” on page 33-0 property of `results`.

The horizontal bar graph shows the coefficient values of the simple model, sorted by their absolute values. LIME finds `Horsepower`, `Model_Year`, and `Cylinders` as important predictors for the query point.

`Model_Year` and `Cylinders` are categorical predictors that have multiple categories. For a linear simple model, the software creates one less dummy variable than the number of categories for each categorical predictor. The bar graph displays only the most important dummy variable. You can check the coefficients of the other dummy variables using the `SimpleModel` property of `results`. Display the sorted coefficient values, including all categorical dummy variables.

```
[~,I] = sort(abs(results.SimpleModel.Beta),'descend');
table(results.SimpleModel.ExpandedPredictorNames(I)',results.SimpleModel.Beta(I), ...
    'VariableNames',{'Exteded Predictor Name','Coefficient'})
```

```
ans=17×2 table
      Exteded Predictor Name      Coefficient
    _____________________________    ___________
```

```
{'Horsepower'            }    -3.4485e-05
{'Model_Year (74 vs. 70)'}    -6.1279e-07
{'Model_Year (80 vs. 70)'}     -4.015e-07
{'Model_Year (81 vs. 70)'}     3.4176e-07
{'Model_Year (82 vs. 70)'}    -2.2483e-07
{'Cylinders (6 vs. 8)'   }    -1.9024e-07
{'Model_Year (76 vs. 70)'}     1.8136e-07
{'Cylinders (5 vs. 8)'   }     1.7461e-07
{'Model_Year (71 vs. 70)'}      1.558e-07
{'Model_Year (75 vs. 70)'}     1.5456e-07
{'Model_Year (77 vs. 70)'}      1.521e-07
{'Model_Year (78 vs. 70)'}     1.4272e-07
{'Model_Year (72 vs. 70)'}     6.7001e-08
{'Model_Year (73 vs. 70)'}     4.7214e-08
{'Cylinders (4 vs. 8)'   }     4.5118e-08
{'Model_Year (79 vs. 70)'}    -2.2598e-08
  ⋮
```

Input Arguments

`results` — LIME results

`lime` object

LIME results, specified as a `lime` object. The `SimpleModel` property of `results` must contain a fitted simple model.

References

[1] Ribeiro, Marco Tulio, S. Singh, and C. Guestrin. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. San Francisco, California: ACM, 2016.

See Also

`fit` | `lime` | `plotPartialDependence`

Topics

“Interpret Machine Learning Models” on page 18-256

Introduced in R2020b

plot

Scatter plot or added variable plot of linear regression model

Syntax

```
plot(mdl)
plot(ax,mdl)
h = plot( ___ )
```

Description

`plot(mdl)` creates a plot of the linear regression model `mdl`. The plot type depends on the number of predictor variables.

- If `mdl` includes multiple predictor variables, `plot` creates an "Added Variable Plot" on page 33-4554 for the whole model except the constant (intercept) term, equivalent to `plotAdded(mdl)`.
- If `mdl` includes a single predictor variable, `plot` creates a scatter plot of the data along with a fitted curve and confidence bounds.
- If `mdl` does not include a predictor, `plot` creates a histogram of the residuals, equivalent to `plotResiduals(mdl)`.

`plot(ax,mdl)` creates the plot in the axes specified by `ax` instead of the current axes.

`h = plot( ___ )` returns graphics objects for the lines or patch in the plot, using any of the input argument combinations in the previous syntaxes. Use `h` to modify the properties of a specific line or patch after you create the plot. For a list of properties, see Chart Line and Patch Properties.

Examples

Create Added Variable Plot

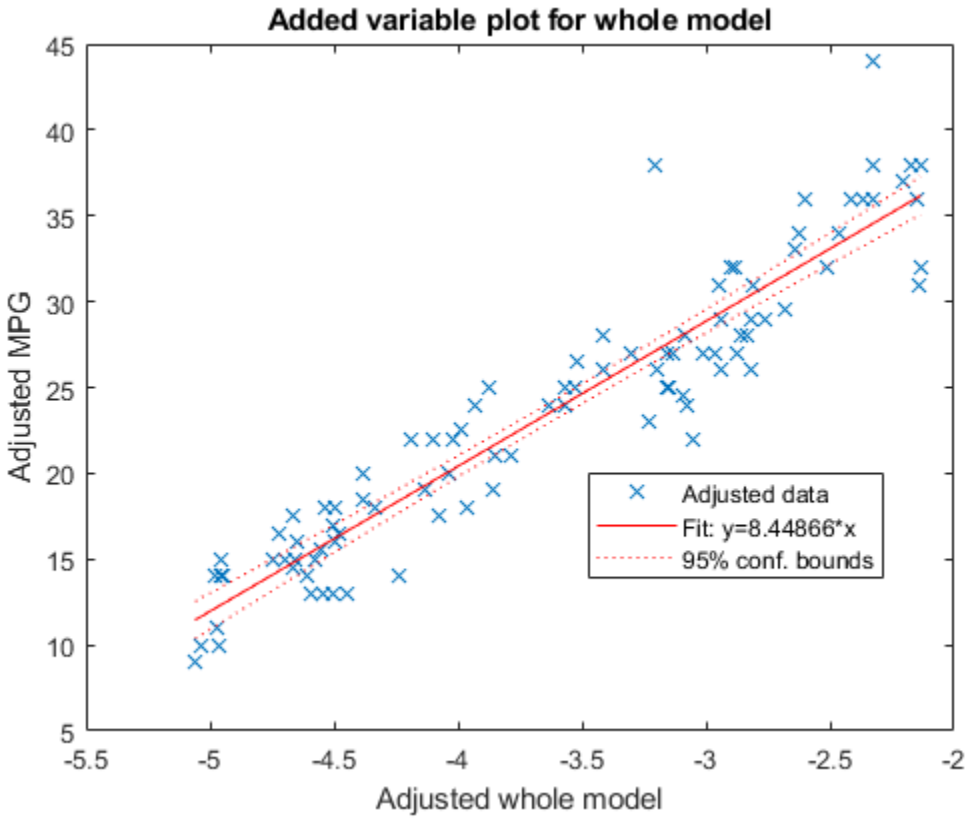
Create a linear regression model of car mileage as a function of weight and model year. Then create an added variable plot to see the significance of the model.

Create a linear regression model of mileage from the `carsmall` data set.

```
load carsmall
Year = categorical(Model_Year);
tbl = table(MPG,Weight,Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Create an added variable plot of the model.

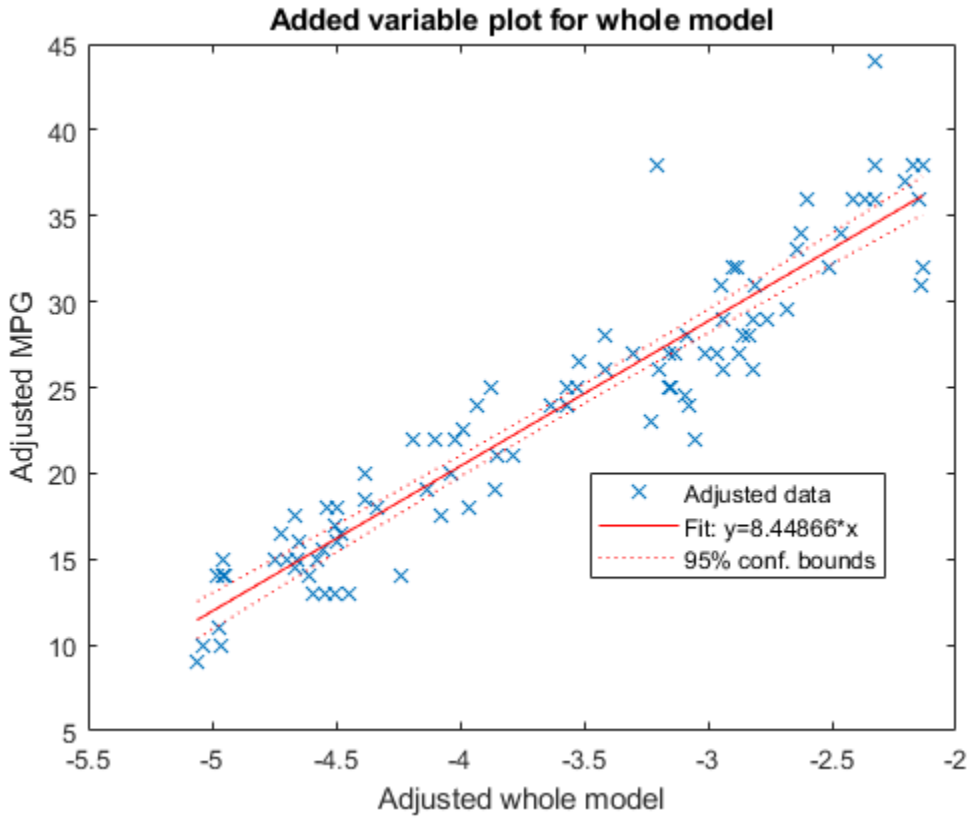
```
plot(mdl)
```



The plot illustrates that the model is significant because a horizontal line does not fit between the confidence bounds.

Create the same plot by using the `plotAdded` function.

```
plotAdded(mdl)
```



Create Scatter Plot for Simple Linear Regression

Create a scatter plot of data along with a fitted curve and confidence bounds for a simple linear regression model. A simple linear regression model includes only one predictor variable.

Create a simple linear regression model of mileage from the `carsmall` data set.

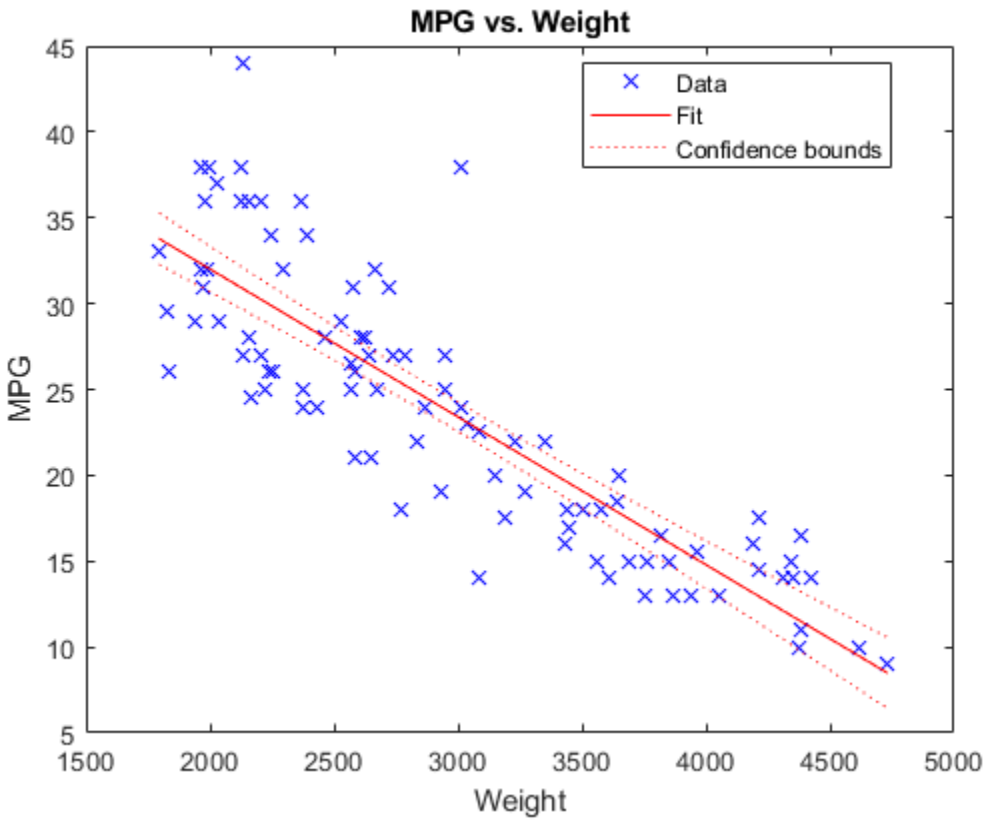
```
load carsmall
tbl = table(MPG,Weight);
mdl = fitlm(tbl,'MPG ~ Weight')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Weight

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ________    _________    _______    __________

    (Intercept)       49.238       1.6411     30.002    2.7015e-49
    Weight         -0.0086119    0.0005348    -16.103    1.6434e-28


Number of observations: 94, Error degrees of freedom: 92
Root Mean Squared Error: 4.13
```

```
R-squared: 0.738,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 259, p-value = 1.64e-28
```

`pValue` of the `Weight` variable is very small, which means that the variable is statistically significant in the model. Visualize this result by creating a scatter plot of the data, along with a fitted curve and its 95% confidence bounds, using the `plot` function.

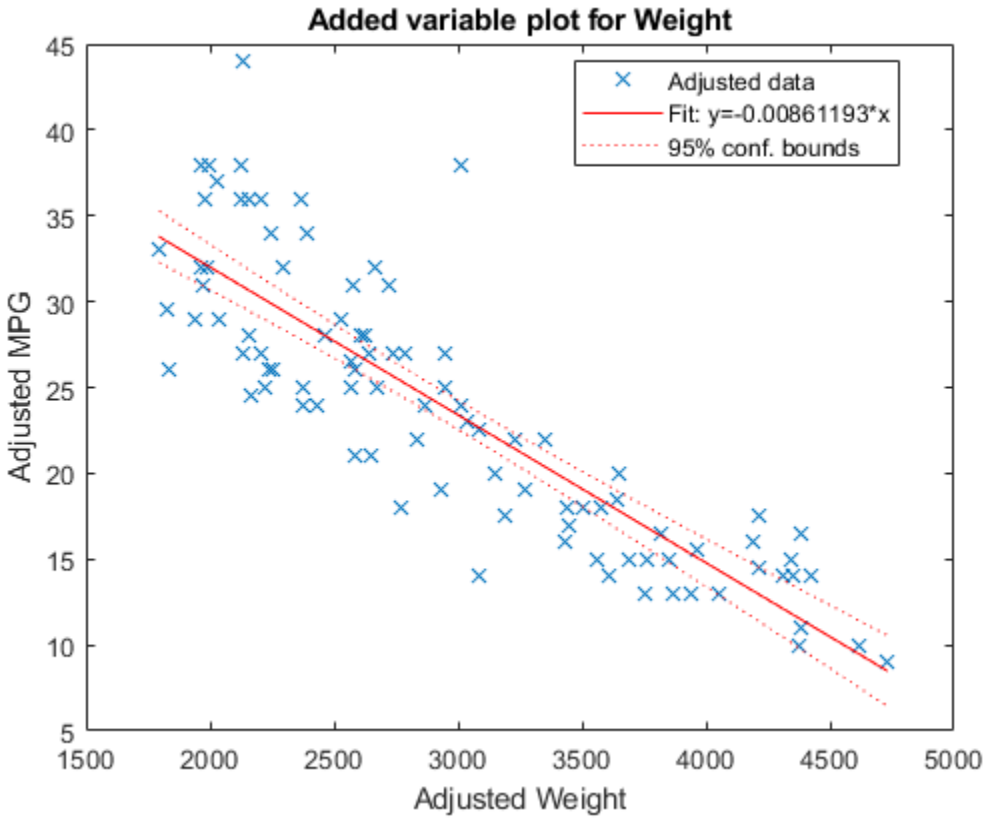
```
plot(mdl)
```



The plot illustrates that the model is significant because a horizontal line does not fit between the confidence bounds, which is consistent with the `pValue` result.

Create the same plot by using the `plotAdded` function.

```
plotAdded(mdl)
```



When a model includes only one term in addition to the constant term, an adjusted value is equivalent to its original value. Therefore, this added variable plot is the same as the scatter plot created by the `plot` function.

Input Arguments

`mdl` — Linear regression model
`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`ax` — Target axes
`Axes` object

Target axes, specified as an `Axes` object.

If you do not specify the axes and the current axes are Cartesian, then `plot` uses the current axes (`gca`). For more information on creating an `Axes` object, see `axes` and `gca`.

Output Arguments

`h` — Graphics objects
graphics array

Graphics objects corresponding to the lines or patch in the plot, returned as a graphics array. Use dot notation to query and set properties of graphics objects. For details, see Line Properties and Patch Properties.

If `mdl` includes one or more predictors, then `h(1)`, `h(2)`, `h(3)`, and `h(4)` correspond to adjusted data points, the fitted line, and the lower and upper bounds of the fitted line, respectively.

If `mdl` does not include a predictor, then `h` corresponds to the histogram of residuals.

More About

Added Variable Plot

An added variable plot, also known as a partial regression leverage plot, illustrates the incremental effect on the response of specified terms caused by removing the effects of all other terms.

An added variable plot created by `plotAdded` with a single selected term corresponding to a single predictor variable includes these plots:

- Scatter plot of adjusted response values against adjusted predictor variable values
- Fitted line for adjusted response values as a function of adjusted predictor variable values
- 95% confidence bounds of the fitted line

The adjusted values are equal to the average of the variable plus the residuals of the variable fit to all predictors except the selected predictor. For example, consider an added variable plot for the first predictor variable x_1. Fit the response variable y and the selected predictor variable x_1 to all predictors except x_1 as follows:

$$y_i = g_y(x_{2i}, x_{3i}, \ldots, x_{pi}) + r_{yi},$$
$$x_{1i} = g_x(x_{2i}, x_{3i}, \ldots, x_{pi}) + r_{xi},$$

where g_y and g_x are the fit of y and x_1, respectively, against all predictors except the selected predictor (x_1). r_y and r_x are the corresponding residual vectors. The subscript i represents the observation number. The adjusted value is the sum of the average value and the residual for each observation.

$$\tilde{y}_i = \bar{y} + r_{yi},$$
$$\tilde{x}_{1i} = \bar{x}_1 + r_{xi},$$

where $\bar{x}_1$ and $\bar{y}$ represent the average of x_1 and y, respectively.

`plotAdded` plots a scatter plot of $(\tilde{x}_{1i}, \tilde{y}_i)$, a fitted line for $\tilde{y}$ as a function of $\tilde{x}_1$ (that is, $\beta_1\tilde{x}_1$), and the 95% confidence bounds of the fitted line. The coefficient β_1 is the same as the coefficient estimate of x_1 in the full model, which includes all predictors.

r_{yi} represents the part of the response values unexplained by the predictors (except x_1), and r_{xi} represents the part of the x_1 values unexplained by the other predictors. Therefore, the fitted line represents how the new information introduced by adding x_1 can explain the unexplained part of the response values. If the slope of the fitted line is close to zero and the confidence bounds can include a horizontal line, then the plot indicates that the new information from x_1 does not explain the unexplained part of the response values well. That is, x_1 is not significant in the model fit.

`plotAdded` also supports an extension of the added variable plot so that you can select multiple terms instead of a single term. Therefore, you can also specify a categorical predictor, all terms that

involve a specific predictor, or the model as a whole (except a constant (intercept) term). Consider a set of predictors X with a coefficient vector β, where β_i is the coefficient estimate of x_i in the full model if you specify the ith coefficient for an added variable plot; otherwise, β_i is zero. Define a unit direction vector u as $u = \beta/s$ where $s = \text{norm}(\beta)$. Then, $X\beta = (Xu)s$. Treat Xu as a single predictor with a coefficient s, and create an added variable plot for Xu in the same way as creating the plot for a single term. The coefficient of the fitted line in the added variable plot corresponds to s.

`plot` creates an added variable plot for the model as a whole (except a constant term) if the model includes multiple terms.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the x-axis and y-axis values for the selected point, along with the observation name or number.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.
- The `plot` function creates an added variable plot for the model as a whole (except a constant term) if the model includes multiple terms. Use `plotAdded` to select particular predictors for an added variable plot.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `plotAdded` | `plotResiduals`

Topics

“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2012a

plot

Class: `RepeatedMeasuresModel`

Plot data with optional grouping

Syntax

```
plot(rm)
plot(rm,Name,Value)
H = plot( ___ )
```

Description

`plot(rm)` plots the measurements in the repeated measures model `rm` for each subject as a function of time. If there is a single numeric within-subjects factor, `plot` uses the values of that factor as the time values. Otherwise, `plot` uses the discrete values 1 through *r* as the time values, where *r* is the number of repeated measurements.

`plot(rm,Name,Value)` also plots the measurements in the repeated measures model `rm`, with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the factors to group by or change the line colors.

`H = plot( ___ )` returns handles, `H`, to the plotted lines.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Group` — Name of between-subject factor or factors

character vector | string array | cell array of character vectors

Name of between-subject factor or factors, specified as the comma-separated pair consisting of `'Group'` and a character vector, string array, or cell array of character vectors. This name-value pair argument groups the lines according to the factor values.

For example, if you have two between-subject factors, drug and sex, and you want to group the lines in the plot according to them, you can specify these factors as follows.

Example: `'Group',{'Drug','Sex'}`

Data Types: `char` | `string` | `cell`

`Marker` — Marker to use for each group
string array | cell array of character vectors

Marker to use for each group, specified as the comma-separated pair consisting of `'Marker'` and a string array or cell array of character vectors.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify `o` as the marker for the groups of drug and `x` as the marker for the groups of sex as follows.

Example: `'Marker',{'o','o','x','x'}`

Data Types: `string` | `cell`

`Color` — Color for each group
character vector | string array | cell array of character vectors | rows of a three-column RGB matrix

Color for each group, specified as the comma-separated pair consisting of `'Color'` and a character vector, string array, cell array of character vectors, or rows of a three-column RGB matrix.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify red as the color for the groups of drug and blue as the color for the groups of sex as follows.

Example: `'Color','rrbb'`

Data Types: `single` | `double` | `char` | `string` | `cell`

`LineStyle` — Line style for each group
string array | cell array of character vectors

Line style for each group, specified as the comma-separated pair consisting of `'LineStyle'` and a string array or cell array of character vectors.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify `-` as the line style of one group and `:` as the line style for the other group as follows.

Example: `'LineStyle',{'-' ':' '-' ':'}`

Data Types: `string` | `cell`

Output Arguments

`H` — Handle to plotted lines
handle

Handle to plotted lines, returned as a handle.

Examples

Plot Data by Group

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

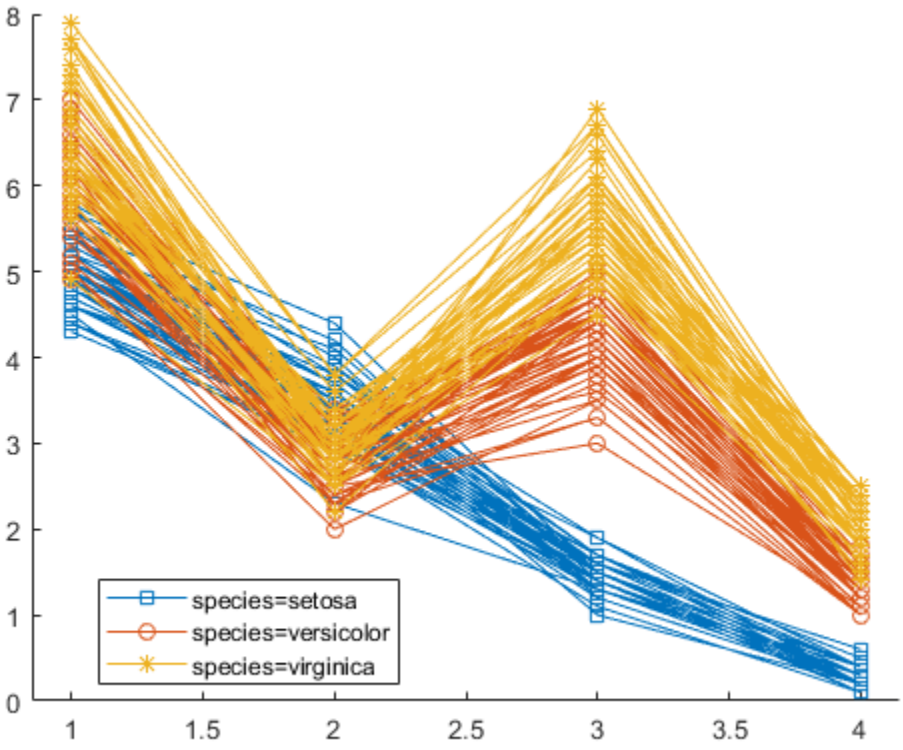
```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

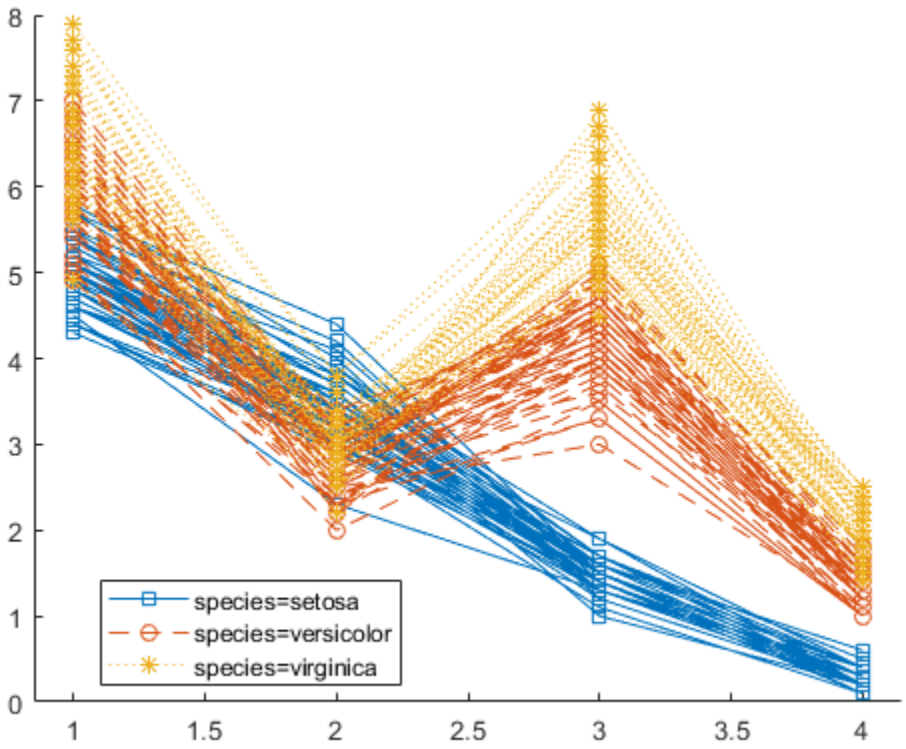
Plot data grouped by the factor species.

```
plot(rm,'group','species')
```



Change the line style for each group.

```
plot(rm,'group','species','LineStyle',{'-','--',':'})
```



Plot Data Grouped by Two Factors

Load the sample data.

```
load repeatedmeas
```

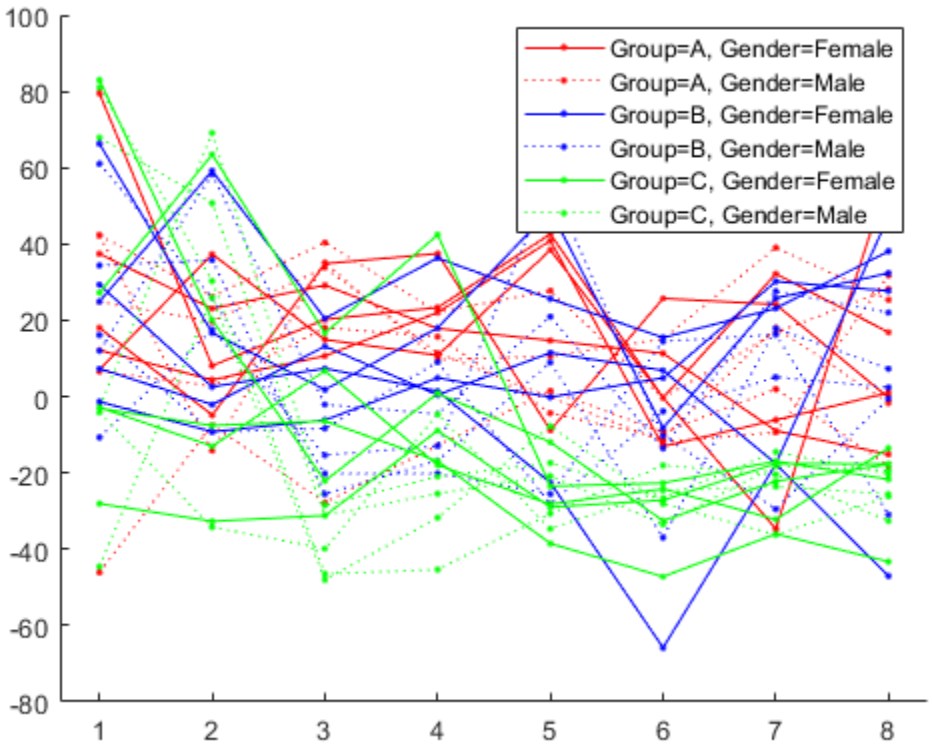
The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Plot data with `Group` coded by color and `Gender` coded by line type.

```
plot(rm,'group',{'Group' 'Gender'},'Color','rrbbgg',...
             'LineStyle',{'-' ':' '-' ':' '-' ':'},'Marker','.')
```



See Also

`fitrm` | `multcompare` | `plotprofile`

plot

Plot Shapley values

Syntax

```
plot(explainer)
plot(explainer,Name,Value)
b = plot( ___ )
```

Description

`plot(explainer)` creates a horizontal bar graph of the Shapley values of the `shapley` object `explainer`. These values are stored in the object's `ShapleyValues` property. Each bar shows the Shapley value of each feature in the blackbox model (`explainer.BlackboxModel`) for the query point (`explainer.QueryPoint`).

`plot(explainer,Name,Value)` specifies additional options using one or more name-value arguments. For example, specify `'NumImportantPredictors',5` to plot the Shapley values of the five features with the highest absolute Shapley values.

`b = plot( ___ )` returns a bar graph object `b` using any of the input argument combinations in the previous syntaxes. Use `b` to query or modify Bar Properties of the bar graph after it is created.

Examples

Plot Shapley Values for All Classes

Train a classification model and create a `shapley` object. Then plot the Shapley values by using the object function `plot`.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    ______

    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

Train a blackbox model of credit ratings by using the `fitcecoc` function. Use the variables from the second through seventh columns in `tbl` as the predictor variables. A recommended practice is to specify the class names to set the order of the classes.

```
blackbox = fitcecoc(tbl,'Rating', ...
    'PredictorNames',tbl.Properties.VariableNames(2:7), ...
    'CategoricalPredictors','Industry', ...
    'ClassNames',{'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'});
```

Create a `shapley` object that explains the prediction for the last observation. For faster computation, subsample 25% of the observations from `tbl` with stratification and use the samples to compute the Shapley values.

```
queryPoint = tbl(end,:)
```

```
queryPoint=1×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA    Industry    Rating
    _____    _____    _____    _______    ________    ____    ________    ______

    73104    0.239    0.463     0.065      2.924      0.34       2        {'AA'}
```

```
rng('default') % For reproducibility
c = cvpartition(tbl.Rating,'Holdout',0.25);
tbl_s = tbl(test(c),:);
explainer = shapley(blackbox,tbl_s,'QueryPoint',queryPoint);
```

For a classification model, `shapley` computes Shapley values using the predicted class score for each class. Display the values in the `ShapleyValues` property.

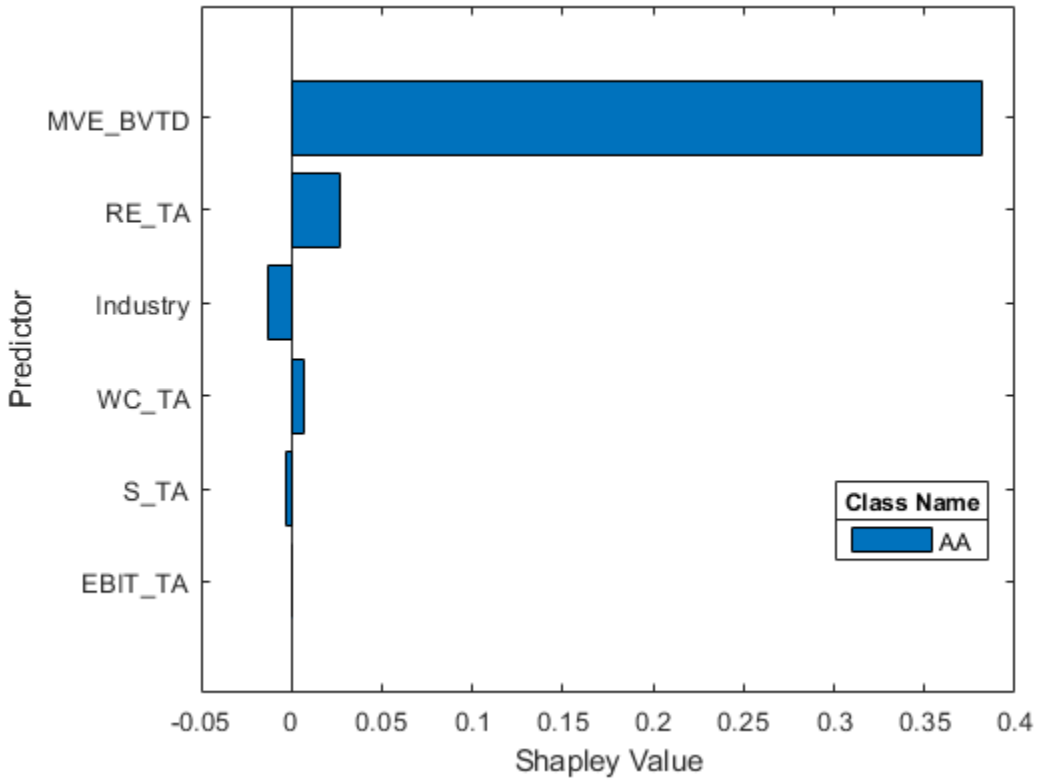
```
explainer.ShapleyValues
```

```
ans=6×8 table
    Predictor        AAA           AA           A            BBB           BB            B
    __________    __________    __________    ___________    __________    ___________    ____

    "WC_TA"         0.014583     0.0064712      0.0027454    0.00045582     -0.0079591     -0.0
    "RE_TA"         0.047796      0.027069       0.015173    -0.0031936      -0.025054     -0.0
    "EBIT_TA"     0.00034325    0.00015238     0.00012385    3.5202e-05     -0.00019141    -0.000
    "MVE_BVTD"       0.38221       0.38228        0.19382    -0.0079011       -0.15755      -0.2
    "S_TA"        -0.0035662    -0.0025999    -0.00021203    -0.0010166     -2.0954e-05    0.000
    "Industry"     -0.028314     -0.013387     0.00088939      0.022877        0.025637     0.02
```

The `ShapleyValues` property contains the Shapley values of all features for each class.

Plot the Shapley values for the predicted class by using the `plot` function. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

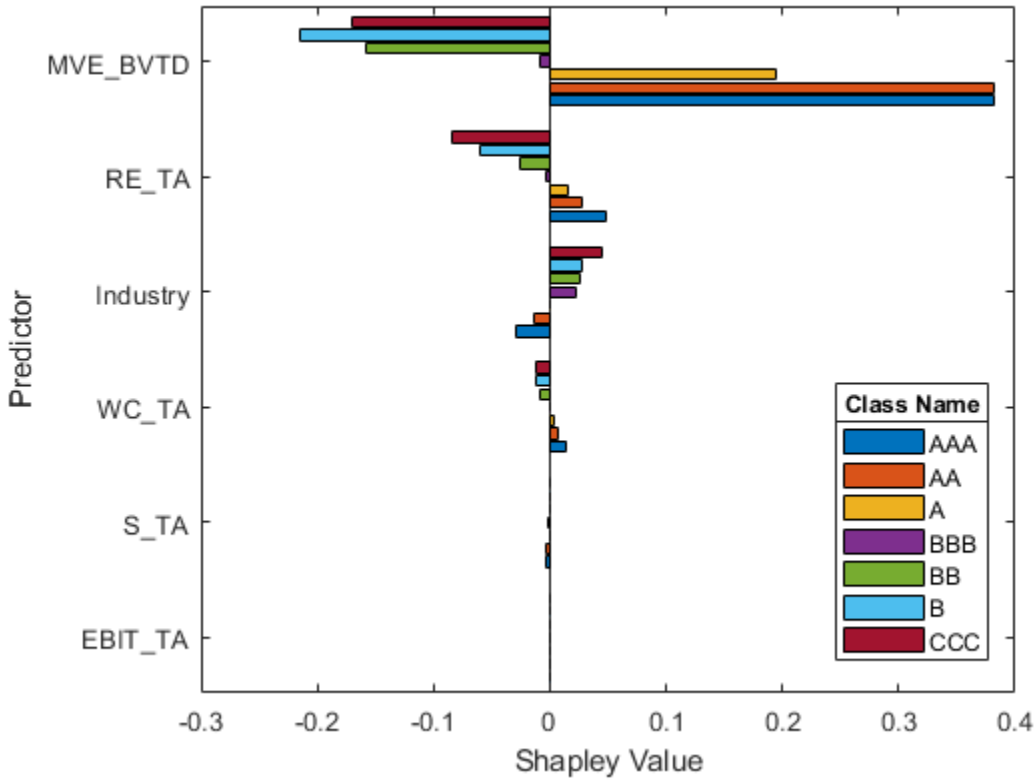
```
f = figure;
plot(explainer);
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The horizontal bar graph shows the Shapley values for all variables, sorted by their absolute values. Each Shapley value explains the deviation of the score for the query point from the average score of the predicted class, due to the corresponding variable.

Plot the Shapley values for all classes by specifying all class names in `explainer.BlackboxModel`.

```
f = figure;
plot(explainer,'ClassNames',explainer.BlackboxModel.ClassNames)
f.CurrentAxes.TickLabelInterpreter = 'none';
```



Specify Number of Important Predictors to Plot

Train a regression model and create a `shapley` object. Use the object function `fit` to compute the Shapley values for the specified query point. Then plot the Shapley values of the predictors by using the object function `plot`. Specify the number of important predictors to plot when you call the `plot` function.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Train a blackbox model of `MPG` by using the `fitrkernel` function

```
rng('default') % For reproducibility
mdl = fitrkernel(tbl,'MPG','CategoricalPredictors',[2 5]);
```

Create a `shapley` object. Specify the data set `tbl`, because `mdl` does not contain training data.

```
explainer = shapley(mdl,tbl)
```

```
explainer =
  shapley with properties:

            BlackboxModel: [1x1 RegressionKernel]
               QueryPoint: []
           BlackboxFitted: []
            ShapleyValues: []
               NumSubsets: 64
                        X: [392x7 table]
    CategoricalPredictors: [2 5]
                   Method: 'interventional-kernel'
```

`explainer` stores the training data `tbl` in the `X` property.

Compute the Shapley values of all predictor variables for the first observation in `tbl`.

```
queryPoint = tbl(1,:)
```

```
queryPoint=1×7 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight    MPG
    ____________    _________    ____________    __________    __________    ______    ___

         12             8            307            130            70         3504     18
```

```
explainer = fit(explainer,queryPoint);
```

For a regression model, `shapley` computes Shapley values using the predicted response, and stores them in the `ShapleyValues` property. Display the values in the `ShapleyValues` property.

```
explainer.ShapleyValues
```

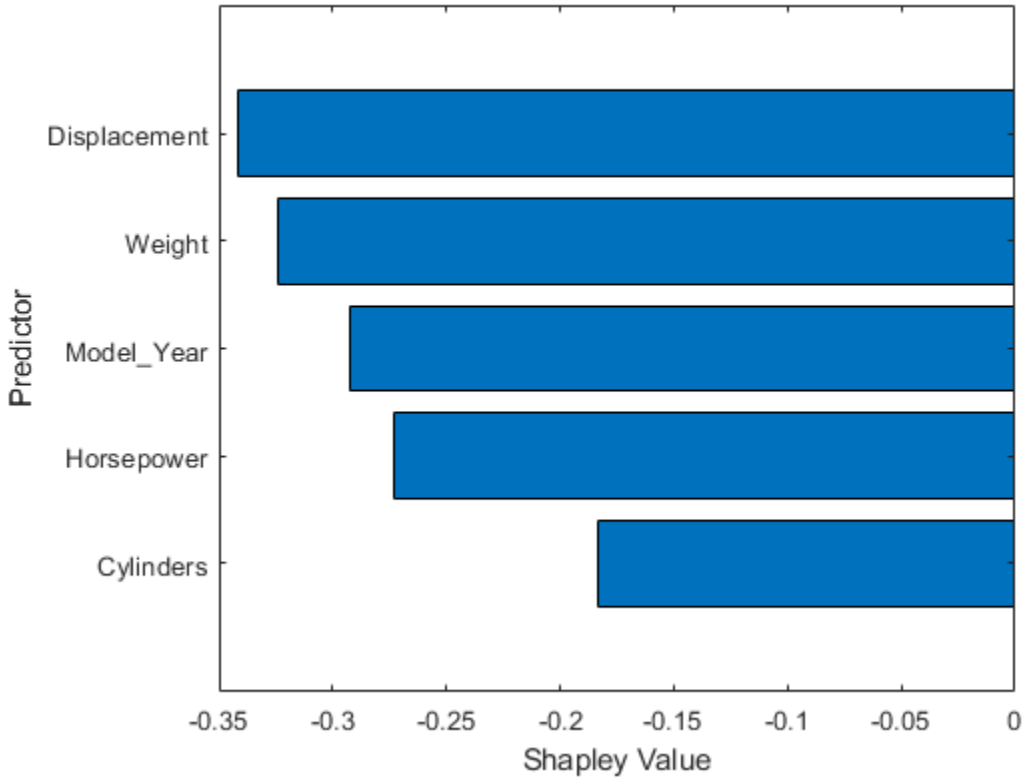
```
ans=6×2 table
      Predictor       ShapleyValue
    _______________    ____________

    "Acceleration"       -0.1561
    "Cylinders"         -0.18306
    "Displacement"      -0.34203
    "Horsepower"        -0.27291
    "Model_Year"         -0.2926
    "Weight"            -0.32402
```

Display the predicted response for the query point, and plot the Shapley values for the query point by using the `plot` function. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`. Specify `'NumImportantPredictors',5` to plot only the five most important predictors for the predicted response.

```
explainer.BlackboxFitted
```

```
ans = 21.0495
```

```
f = figure;
plot(explainer,'NumImportantPredictors',5)
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The horizontal bar graph shows the Shapley values for the five most important predictors, sorted by their absolute values. Each Shapley value explains the deviation of the prediction for the query point from the average, due to the corresponding variable.

Input Arguments

`explainer` — Object explaining blackbox model

`shapley` object

Object explaining the blackbox model, specified as a `shapley` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `plot(explainer,'NumImportantPredictors',5,'ClassNames',c)` creates a bar graph containing the Shapley values of the five most important predictors for the class `c`.

`NumImportantPredictors` — Number of important predictors to plot

`min(M,10)` where `M` is the number of predictors (default) | positive integer

Number of important predictors to plot, specified as a positive integer. The `plot` function plots the Shapley values of the specified number of predictors with the highest absolute Shapley values.

Example: `'NumImportantPredictors',5` specifies to plot the five most important predictors. The `plot` function determines the order of importance by using the absolute Shapley values.

Data Types: `single` | `double`

`ClassNames` — Class labels to plot
`explainer.BlackboxFitted` (default) | categorical array | character array | logical vector | numeric vector | cell array of character vectors

Class labels to plot, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. The values and data types in the `'ClassNames'` value must match those of the class names in the `ClassNames` property of the machine learning model in `explainer` (`explainer.BlackboxModel.ClassNames`).

You can specify one or more labels. If you specify multiple class labels, the function plots multiple bars for each feature with different colors.

The default value is the predicted class for the query point (the `BlackboxFitted` property of `explainer`).

This argument is valid only when the machine learning model (`BlackboxModel`) in `explainer` is a classification model.

Example: `'ClassNames',{'red','blue'}`

Example: `'ClassNames',explainer.BlackboxModel.ClassNames` specifies `'ClassNames'` as all classes in `BlackboxModel`.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

More About

Shapley Values

In game theory, the Shapley value of a player is the average marginal contribution of the player in a cooperative game. In the context of machine learning prediction, the Shapley value of a feature for a query point explains the contribution of the feature to a prediction (response for regression or score of each class for classification) at the specified query point.

The Shapley value corresponds to the deviation of the prediction for the query point from the average prediction, due to the feature. For a query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average.

For more details, see “Shapley Values for Machine Learning Model” on page 18-272.

References

[1] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[2] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

See Also

`fit` | `shapley`

Topics

“Shapley Values for Machine Learning Model” on page 18-272
“Interpret Machine Learning Models” on page 18-256

Introduced in R2021a

plotAdded

Added variable plot of linear regression model

Syntax

```
plotAdded(mdl)
plotAdded(mdl,coef)
plotAdded(mdl,coef,Name,Value)

plotAdded(ax,___)

h = plotAdded(___)
```

Description

`plotAdded(mdl)` creates an added variable plot on page 33-4579 for the whole model `mdl` except the constant (intercept) term.

`plotAdded(mdl,coef)` creates an added variable plot for the specified terms `coef`.

`plotAdded(mdl,coef,Name,Value)` specifies graphical properties of adjusted data points using one or more name-value pair arguments. For example, you can specify the marker symbol and size for the data points.

`plotAdded(ax,___)` creates the plot in the axes specified by `ax` instead of the current axes, using any of the input argument combinations in the previous syntaxes.

`h = plotAdded(___)` returns line objects for the plot. Use `h` to modify the properties of a specific line after you create the plot. For a list of properties, see Line Properties.

Examples

Create Added Variable Plot

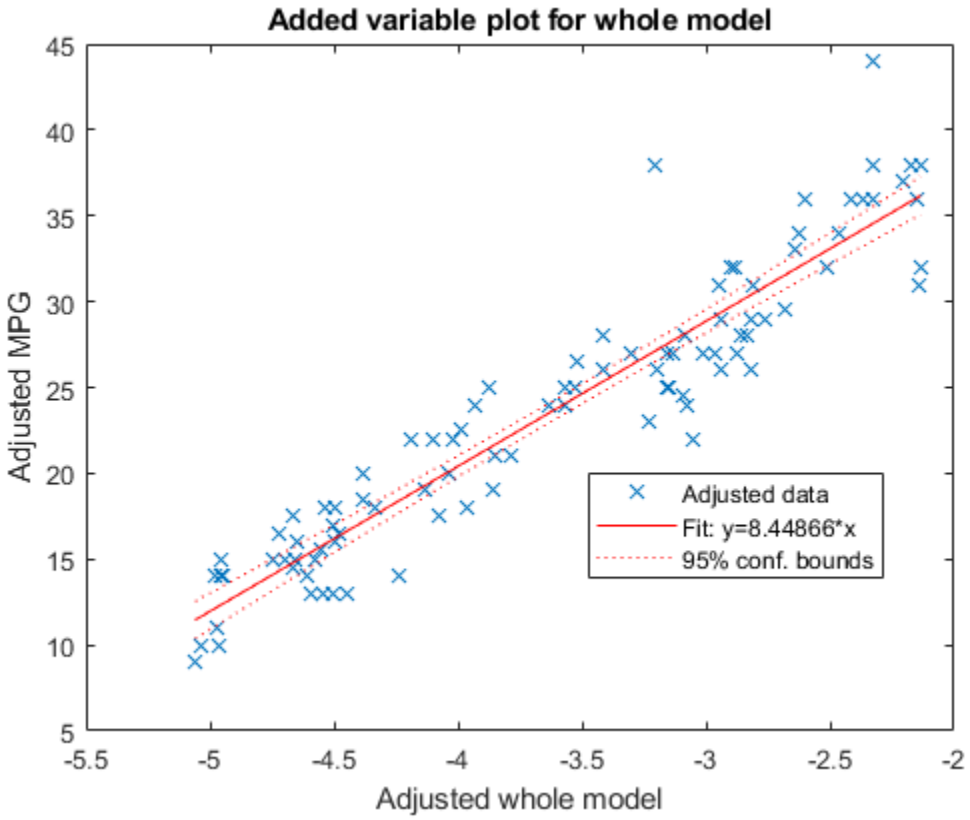
Create a linear regression model of car mileage as a function of weight and model year. Then create an added variable plot to see the significance of the model.

Create a linear regression model of mileage from the `carsmall` data set.

```
load carsmall
Year = categorical(Model_Year);
tbl = table(MPG,Weight,Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Create an added variable plot of the model.

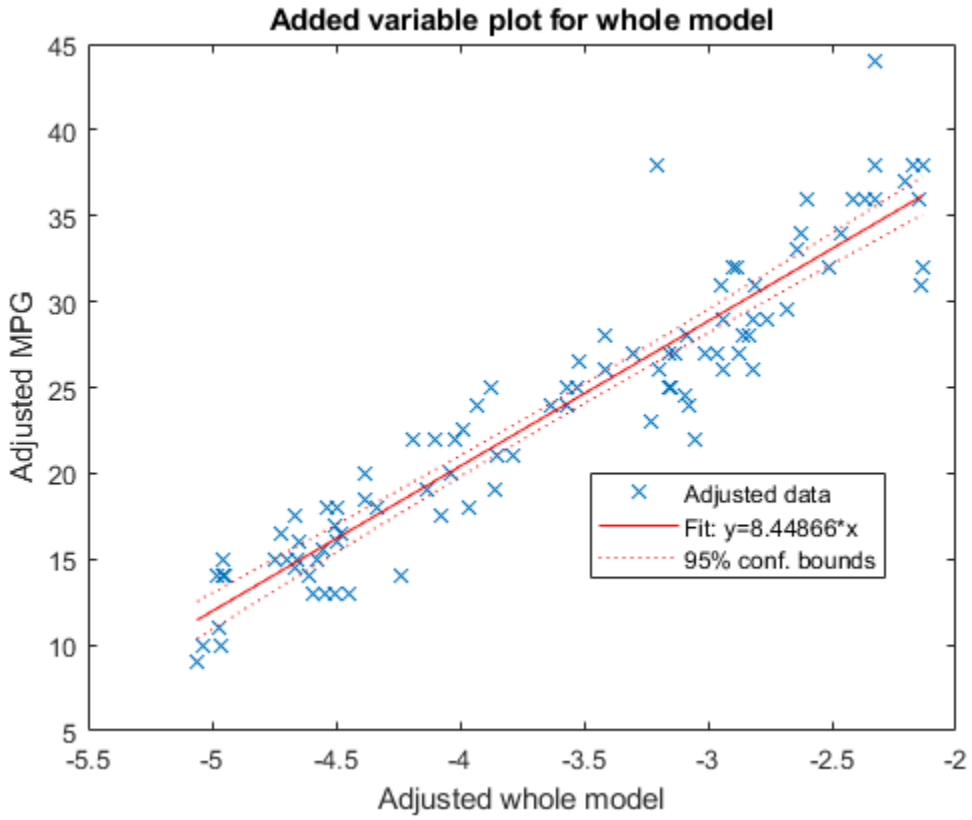
```
plot(mdl)
```



The plot illustrates that the model is significant because a horizontal line does not fit between the confidence bounds.

Create the same plot by using the `plotAdded` function.

```
plotAdded(mdl)
```



Create Added Variable Plot for Specified Variables

Create a linear regression model of car mileage as a function of weight and model year. Then create an added variable plot to see the effect of the weight terms (`Weight` and `Weight^2`).

Create the linear regression model using the `carsmall` data set.

```
load carsmall
Year = categorical(Model_Year);
tbl = table(MPG,Weight,Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Find the terms in the model corresponding to `Weight` and `Weight^2`.

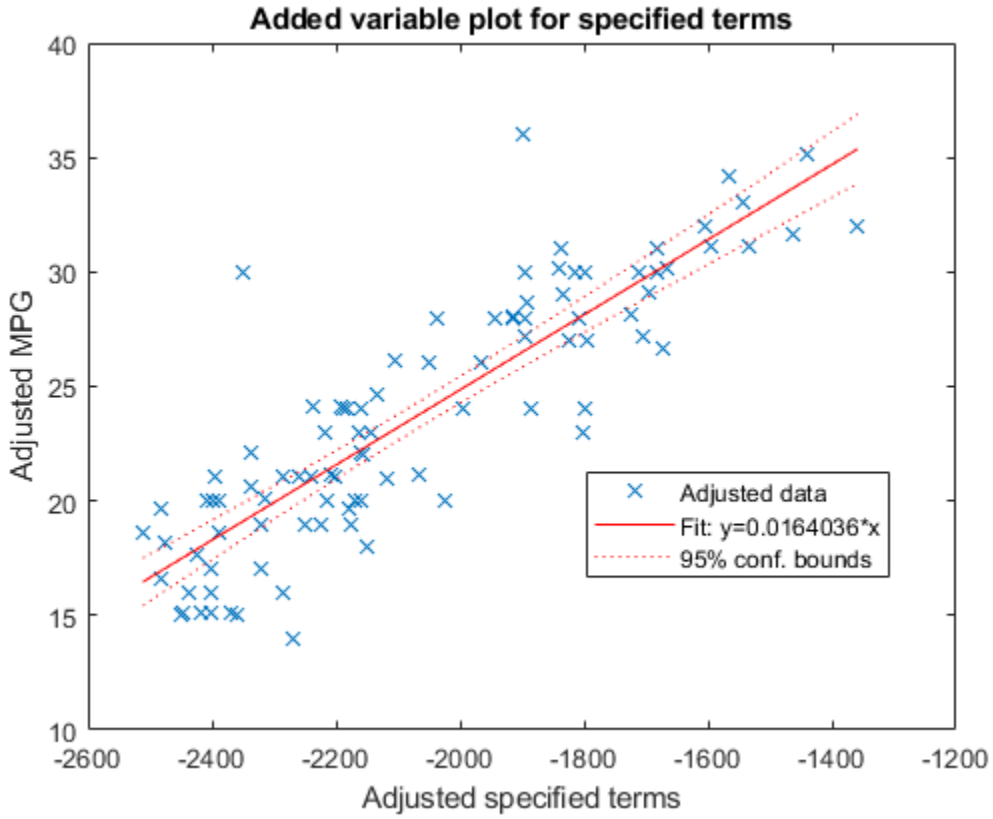
```
mdl.CoefficientNames
```

```
ans = 1x5 cell
    {'(Intercept)'}    {'Weight'}    {'Year_76'}    {'Year_82'}    {'Weight^2'}
```

The weight terms are 2 and 5.

Create an added variable plot with the weight terms.

```
coef = [2 5];
plotAdded(mdl,coef)
```



The plot illustrates that the weight terms are significant because a horizontal line does not fit between the confidence bounds.

Create Scatter Plot for Simple Linear Regression

Create a scatter plot of data along with a fitted curve and confidence bounds for a simple linear regression model. A simple linear regression model includes only one predictor variable.

Create a simple linear regression model of mileage from the `carsmall` data set.

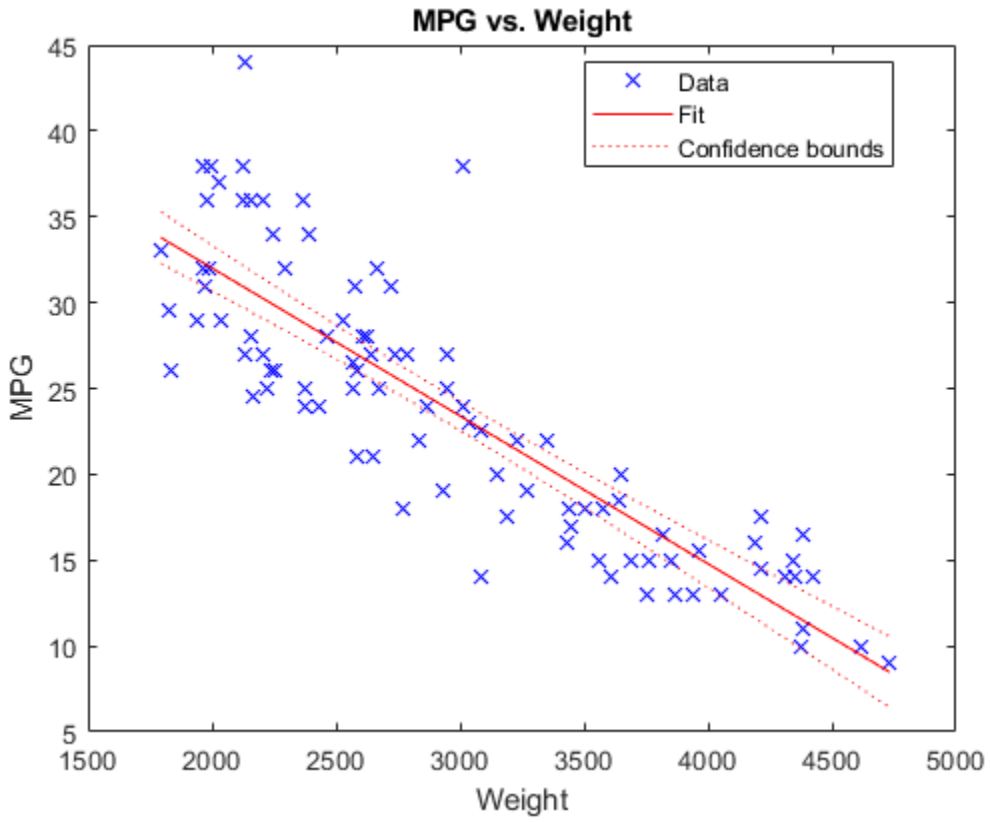
```
load carsmall
tbl = table(MPG,Weight);
mdl = fitlm(tbl,'MPG ~ Weight')
```

```
mdl =
Linear regression model:
    MPG ~ 1 + Weight

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   _________    _________    _______    _________
```

```
    (Intercept)        49.238       1.6411     30.002    2.7015e-49
    Weight         -0.0086119    0.0005348    -16.103    1.6434e-28


Number of observations: 94, Error degrees of freedom: 92
Root Mean Squared Error: 4.13
R-squared: 0.738,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 259, p-value = 1.64e-28
```

`pValue` of the `Weight` variable is very small, which means that the variable is statistically significant in the model. Visualize this result by creating a scatter plot of the data, along with a fitted curve and its 95% confidence bounds, using the `plot` function.

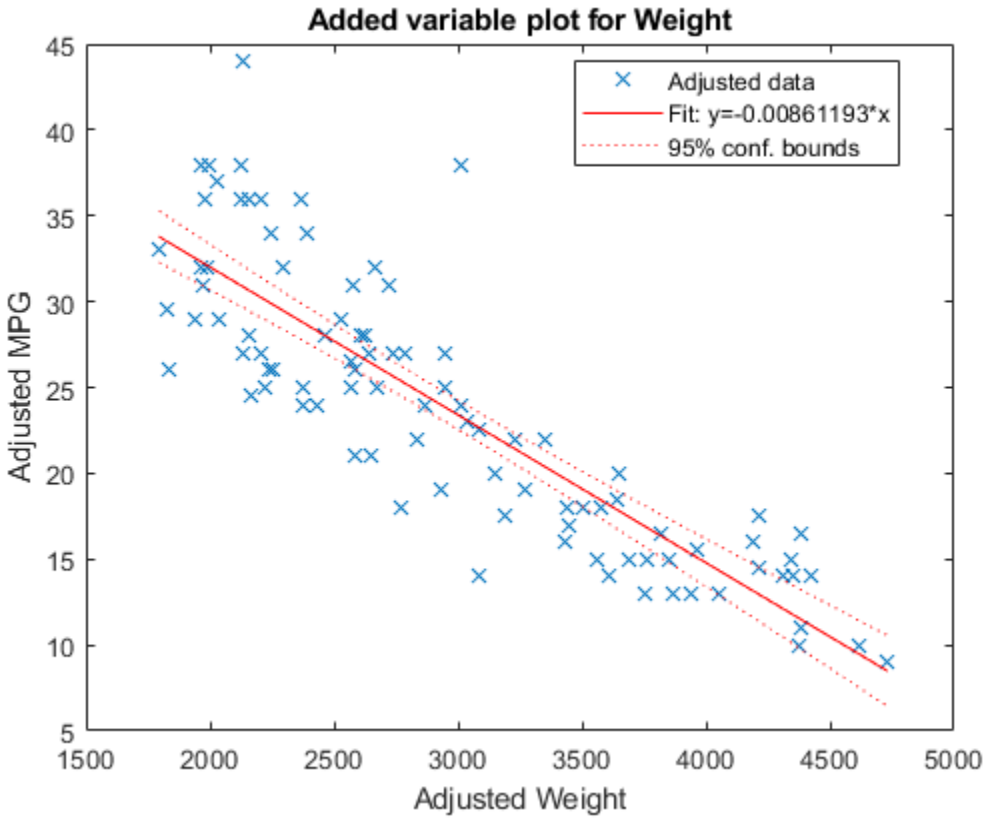
```
plot(mdl)
```



The plot illustrates that the model is significant because a horizontal line does not fit between the confidence bounds, which is consistent with the `pValue` result.

Create the same plot by using the `plotAdded` function.

```
plotAdded(mdl)
```



When a model includes only one term in addition to the constant term, an adjusted value is equivalent to its original value. Therefore, this added variable plot is the same as the scatter plot created by the `plot` function.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`coef` — Coefficients in regression model

character vector | string scalar | vector of positive integers

Coefficients in the regression model `mdl`, specified as one of the following:

- Character vector or string scalar of a single coefficient name in `mdl.CoefficientNames` (`CoefficientNames` property of `mdl`).
- Vector of positive integers representing the indexes of coefficients in `mdl.CoefficientNames`. Use a vector to specify multiple coefficients.

Data Types: `char` | `string` | `single` | `double`

`ax` — Target axes

`Axes` object

Target axes, specified as an `Axes` object.

If you do not specify the axes and the current axes are Cartesian, then `plotAdded` uses the current axes (`gca`). For more information on creating an `Axes` object, see `axes` and `gca`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','*'`

Note The graphical properties listed here are only a subset. For a complete list, see Line Properties. The specified properties determine the appearance of adjusted data points.

`Color` — Line color

RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

Example: `'Color','blue'`

`LineWidth` — Line width
positive value

Line width, specified as the comma-separated pair consisting of `'LineWidth'` and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: `'LineWidth',0.75`

`Marker` — Marker symbol
`'o'` | `'+'` | `'*'` | `'.'` | `'x'` | ...

Marker symbol, specified as the comma-separated pair consisting of `'Marker'` and one of the values in this table.

Value	Description
`'o'`	Circle
`'+'`	Plus sign
`'*'`	Asterisk
`'.'`	Point
`'x'`	Cross
`'_'`	Horizontal line
`'|'`	Vertical line
`'square'` or `'s'`	Square
`'diamond'` or `'d'`	Diamond
`'^'`	Upward-pointing triangle
`'v'`	Downward-pointing triangle

Value	Description
'>'	Right-pointing triangle
'<'	Left-pointing triangle
'pentagram' or 'p'	Five-pointed star (pentagram)
'hexagram' or 'h'	Six-pointed star (hexagram)
'none'	No markers

Example: 'Marker','+'

MarkerEdgeColor — Marker outline color
'auto' (default) | 'none' | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of 'MarkerEdgeColor' and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the Color name-value pair argument.

The default value of 'auto' uses the same color specified by using 'Color'.

Example: 'MarkerEdgeColor','blue'

MarkerFaceColor — Marker fill color
'none' (default) | 'auto' | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of 'MarkerFaceColor' and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the Color name-value pair argument.

The 'auto' value uses the same color specified by using 'Color'.

Example: 'MarkerFaceColor','blue'

MarkerSize — Marker size
6 (default) | positive value

Marker size, specified as the comma-separated pair consisting of 'MarkerSize' and a positive value in points.

Example: 'MarkerSize',2

Output Arguments

h — Line objects
vector

Line objects, returned as a 3-by-1 vector. h(1), h(2), and h(3) correspond to the adjusted data points, fitted line, and 95% confidence bounds of the fitted line, respectively. Use dot notation to query and set properties of the line objects. For details, see Line Properties.

You can use name-value pair arguments to specify the appearance of adjusted data points corresponding to the first graphics object h(1).

More About

Added Variable Plot

An added variable plot, also known as a partial regression leverage plot, illustrates the incremental effect on the response of specified terms caused by removing the effects of all other terms.

An added variable plot created by `plotAdded` with a single selected term corresponding to a single predictor variable includes these plots:

- Scatter plot of adjusted response values against adjusted predictor variable values
- Fitted line for adjusted response values as a function of adjusted predictor variable values
- 95% confidence bounds of the fitted line

The adjusted values are equal to the average of the variable plus the residuals of the variable fit to all predictors except the selected predictor. For example, consider an added variable plot for the first predictor variable x_1. Fit the response variable y and the selected predictor variable x_1 to all predictors except x_1 as follows:

$$y_i = g_y(x_{2i}, x_{3i}, \ldots, x_{pi}) + r_{yi},$$
$$x_{1i} = g_x(x_{2i}, x_{3i}, \ldots, x_{pi}) + r_{xi},$$

where g_y and g_x are the fit of y and x_1, respectively, against all predictors except the selected predictor (x_1). r_y and r_x are the corresponding residual vectors. The subscript i represents the observation number. The adjusted value is the sum of the average value and the residual for each observation.

$$\tilde{y}_i = \bar{y} + r_{yi},$$
$$\tilde{x}_{1i} = \bar{x}_1 + r_{xi},$$

where $\bar{x}_1$ and $\bar{y}$ represent the average of x_1 and y, respectively.

`plotAdded` plots a scatter plot of $(\tilde{x}_{1i}, \tilde{y}_i)$, a fitted line for $\tilde{y}$ as a function of $\tilde{x}_1$ (that is, $\beta_1\tilde{x}_1$), and the 95% confidence bounds of the fitted line. The coefficient β_1 is the same as the coefficient estimate of x_1 in the full model, which includes all predictors.

r_{yi} represents the part of the response values unexplained by the predictors (except x_1), and r_{xi} represents the part of the x_1 values unexplained by the other predictors. Therefore, the fitted line represents how the new information introduced by adding x_1 can explain the unexplained part of the response values. If the slope of the fitted line is close to zero and the confidence bounds can include a horizontal line, then the plot indicates that the new information from x_1 does not explain the unexplained part of the response values well. That is, x_1 is not significant in the model fit.

`plotAdded` also supports an extension of the added variable plot so that you can select multiple terms instead of a single term. Therefore, you can also specify a categorical predictor, all terms that involve a specific predictor, or the model as a whole (except a constant (intercept) term). Consider a set of predictors X with a coefficient vector β, where β_i is the coefficient estimate of x_i in the full model if you specify the ith coefficient for an added variable plot; otherwise, β_i is zero. Define a unit direction vector u as $u = \beta/s$ where $s = \text{norm}(\beta)$. Then, $X\beta = (Xu)s$. Treat Xu as a single predictor with a coefficient s, and create an added variable plot for Xu in the same way as creating the plot for a single term. The coefficient of the fitted line in the added variable plot corresponds to s.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.
- `plotAdded` shows the incremental effect on the response of specified terms by removing the effects of the other terms, whereas `plotAdjustedResponse` shows the effect of a selected predictor in the model fit with the other predictors averaged out by averaging the fitted values. Note that the definitions of adjusted values in `plotAdded` and `plotAdjustedResponse` are not the same.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `plot`

Topics

“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2012a

plotAdjustedResponse

Adjusted response plot of linear regression model

Syntax

```
plotAdjustedResponse(mdl,var)
plotAdjustedResponse(mdl,var,Name,Value)
h = plotAdjustedResponse( ___ )
```

Description

`plotAdjustedResponse(mdl,var)` creates an adjusted response on page 33-4586 plot for the variable `var` in the linear regression model `mdl`.

`plotAdjustedResponse(mdl,var,Name,Value)` specifies graphical properties of adjusted response data points using one or more name-value pair arguments. For example, you can specify the marker symbol and size for the data points.

`h = plotAdjustedResponse( ___ )` returns line objects using any of the input argument combinations in the previous syntaxes. Use `h` to modify the properties of a specific line after you create the plot. For a list of properties, see Line Properties.

Examples

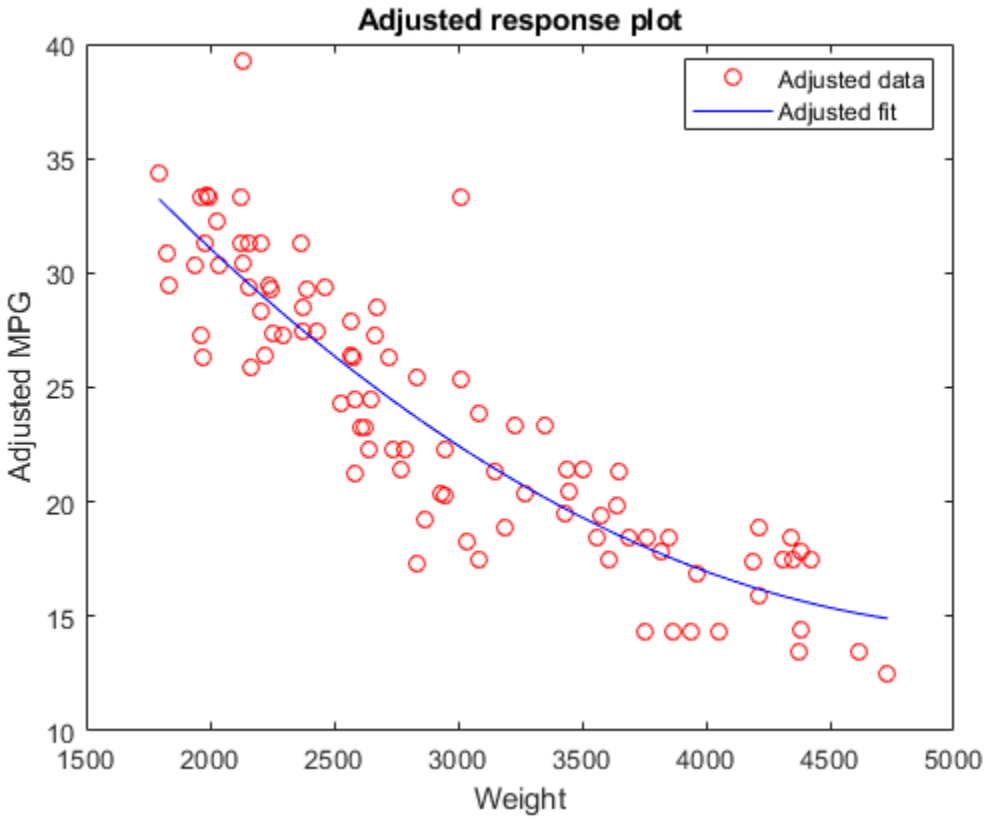
Plot Adjusted Responses

Load the `carsmall` data set and fit a linear model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

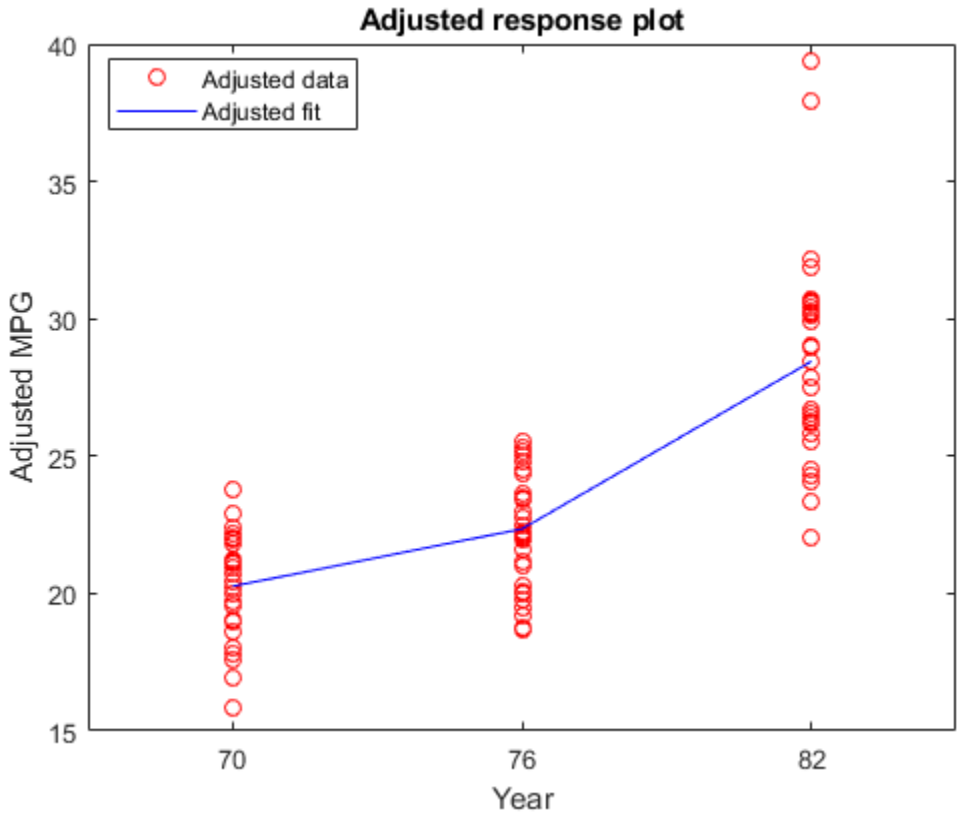
Plot the effect of `Weight` averaged over `Year`.

```
plotAdjustedResponse(mdl,'Weight')
```



Plot the effect of `Year` averaged over `Weight`.

```
plotAdjustedResponse(mdl,'Year');
```



Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`var` — Variable for adjusted response plot

character vector | string array | positive integer

Variable for the adjusted response plot, specified as a character vector or string array of the variable name in `mdl.VariableNames`, or a positive integer representing the index of a variable in `mdl.VariableNames`.

Data Types: `char` | `string` | `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','*'`

Note The graphical properties listed here are only a subset. For a complete list, see Line Properties. The specified properties determine the appearance of adjusted response data points.

`Color` — Line color

RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	

RGB Triplet	Hexadecimal Color Code	Appearance
[0.4660 0.6740 0.1880]	'#77AC30'	
[0.3010 0.7450 0.9330]	'#4DBEEE'	
[0.6350 0.0780 0.1840]	'#A2142F'	

Example: 'Color','blue'

LineWidth — Line width
positive value

Line width, specified as the comma-separated pair consisting of 'LineWidth' and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: 'LineWidth',0.75

Marker — Marker symbol
'o' | '+' | '*' | '.' | 'x' | ...

Marker symbol, specified as the comma-separated pair consisting of 'Marker' and one of the values in this table.

Value	Description
'o'	Circle
'+'	Plus sign
'*'	Asterisk
'.'	Point
'x'	Cross
'_'	Horizontal line
'|'	Vertical line
'square' or 's'	Square
'diamond' or 'd'	Diamond
'^'	Upward-pointing triangle
'v'	Downward-pointing triangle
'>'	Right-pointing triangle
'<'	Left-pointing triangle
'pentagram' or 'p'	Five-pointed star (pentagram)
'hexagram' or 'h'	Six-pointed star (hexagram)
'none'	No markers

Example: 'Marker','+'

MarkerEdgeColor — Marker outline color
'auto' (default) | 'none' | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of 'MarkerEdgeColor' and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the Color name-value pair argument.

The default value of `'auto'` uses the same color specified by using `'Color'`.

Example: `'MarkerEdgeColor','blue'`

`MarkerFaceColor` — Marker fill color
`'none'` (default) | `'auto'` | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of `'MarkerFaceColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The `'auto'` value uses the same color specified by using `'Color'`.

Example: `'MarkerFaceColor','blue'`

`MarkerSize` — Marker size
6 (default) | positive value

Marker size, specified as the comma-separated pair consisting of `'MarkerSize'` and a positive value in points.

Example: `'MarkerSize',2`

Output Arguments

`h` — Line objects
vector

Line objects, returned as a 2-by-1 vector. `h(1)` corresponds to the adjusted response data points, and `h(2)` corresponds to the adjusted response function. Use dot notation to query and set properties of the line objects. For details, see Line Properties.

You can use name-value pair arguments to specify the appearance of adjusted response data points corresponding to the first graphics object `h(1)`.

More About

Adjusted Response

An adjusted response function describes the relationship between the fitted response and a single predictor, with the other predictors averaged out by averaging the fitted values over the data used in the fit.

A regression model for the predictor variables (x_1, x_2, ..., x_p) and the response variable y has the form

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{pi}) + r_i,$$

where f is a fitted regression function and r is a residual. The subscript i represents the observation number.

The adjusted response function for the first predictor variable x_1, for example, is defined as

$$g(x_1) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_1, x_{2i}, x_{3i}, \ldots, x_{pi}\right),$$

where n is the number of observations. The adjusted response data value is the sum of the adjusted fitted value and the residual for each observation.

$$\tilde{y}_i = g(x_{1i}) + r_i .$$

`plotAdjustedResponse` plots the adjusted response function and the adjusted response data values for a selected predictor variable.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.
- `plotPartialDependence` creates either a line plot or a surface plot of predicted responses against a single feature or a pair of features, respectively, by marginalizing over the other variables. A line plot for a single feature from `plotPartialDependence` and an adjusted response function plot from `plotAdjustedResponse` are the same within numerical precision.
- `plotEffects` creates a summary plot that shows separate effects for all predictors.
- `plotAdded` shows the incremental effect on the response of specified terms by removing the effects of the other terms, whereas `plotAdjustedResponse` shows the effect of a selected predictor in the model fit with the other predictors averaged out by averaging the fitted values. Note that the definitions of adjusted values in `plotAdded` and `plotAdjustedResponse` are not the same.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `plotAdded` | `plotEffects` | `plotInteraction` | `plotPartialDependence`

Topics

“Linear Regression Workflow” on page 11-35

“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9

Introduced in R2012a

plot

Plot Bayesian optimization results

Syntax

```
plot(results,'all')
plot(results,plotFcn1,plotFcn2,...)
```

Description

`plot(results,'all')` calls all predefined plot functions on `results`.

`plot(results,plotFcn1,plotFcn2,...)` calls the listed plot functions on `results`.

Examples

Plot After Optimization

This example shows how to plot the error model and the best objective trace after the optimization has finished. The objective function for this example throws an error for points with norm larger than 2.

```
function f = makeanerror(x)
f = x.x1 - x.x2 - sqrt(4-x.x1^2-x.x2^2);
```

```
fun = @makeanerror;
```

Create the variables for optimization.

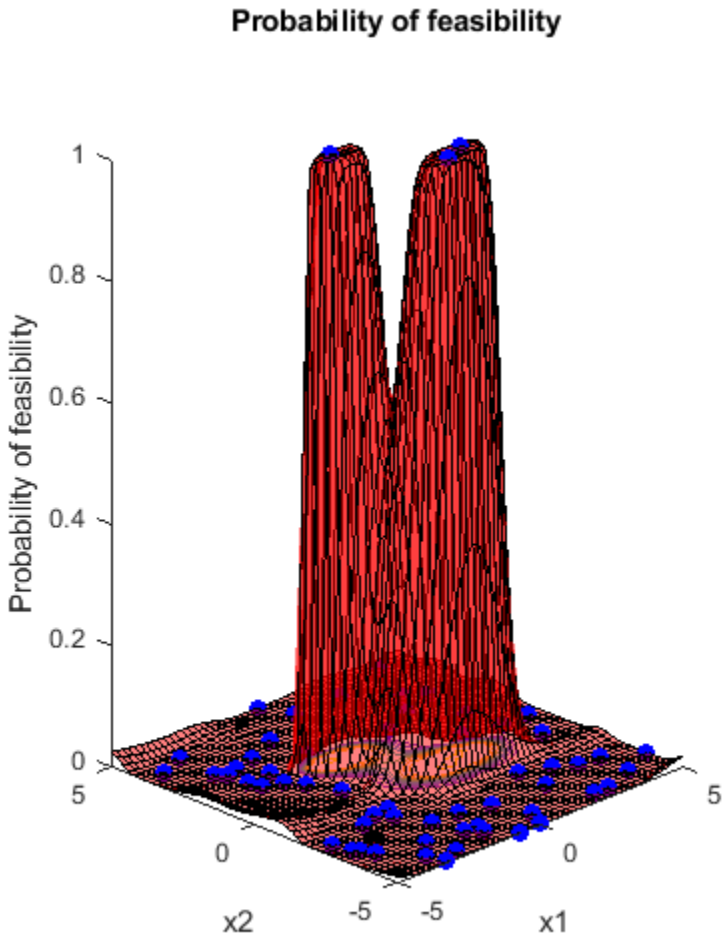
```
var1 = optimizableVariable('x1',[-5,5]);
var2 = optimizableVariable('x2',[-5,5]);
vars = [var1,var2];
```

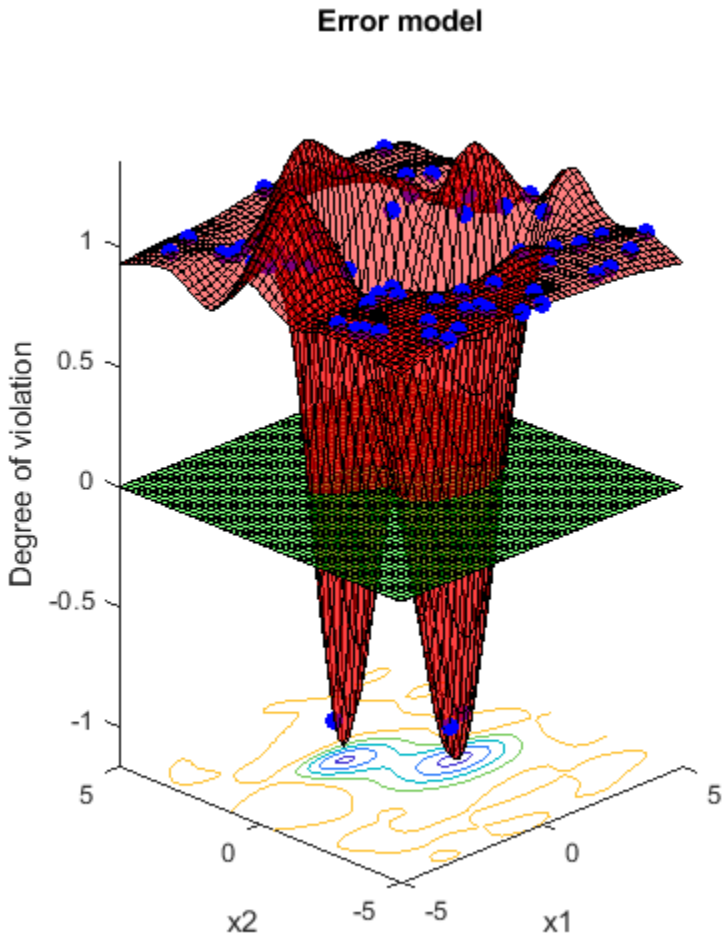
Run the optimization without any plots. For reproducibility, set the random seed and use the 'expected-improvement-plus' acquisition function. Optimize for 60 iterations so the error model becomes well-trained.

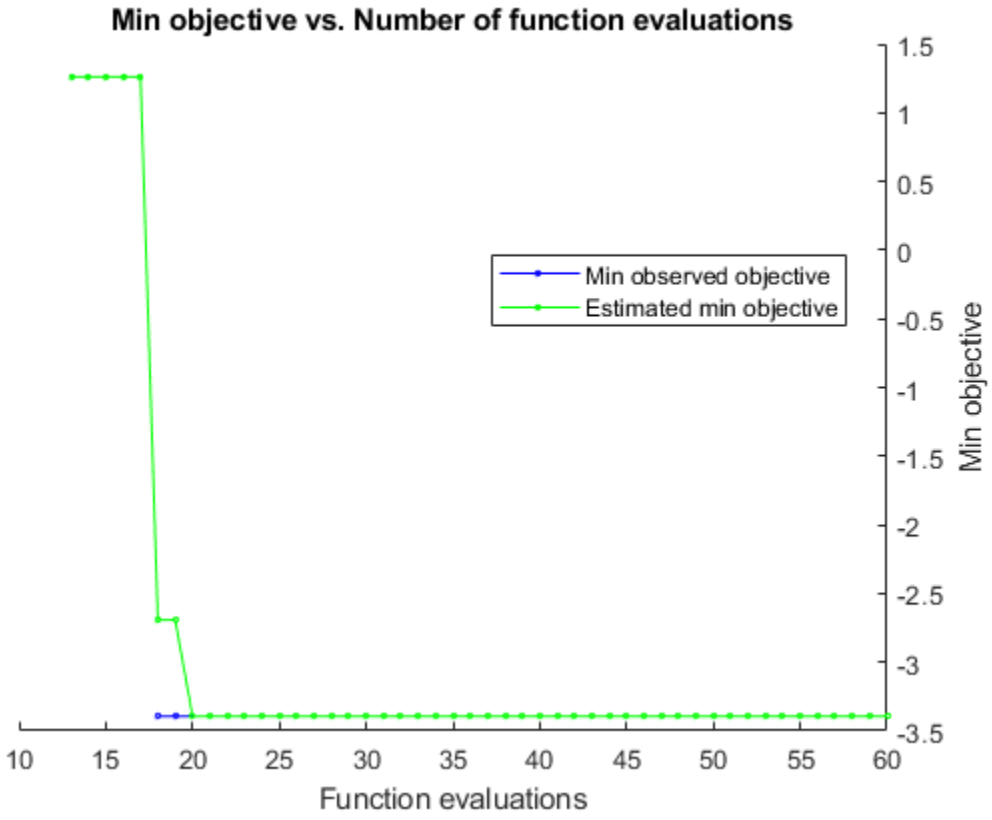
```
rng default
results = bayesopt(fun,vars,'MaxObjectiveEvaluations',60,...
    'AcquisitionFunctionName','expected-improvement-plus',...
    'PlotFcn',[],'Verbose',0);
```

Plot the error model and the best objective trace.

```
plot(results,@plotConstraintModels,@plotMinObjective)
```


Probability of feasibility
Probability of feasibility
1
0.8
0.6
0.4
0.2
0
5
0
-5
x2
-5
0
5
x1


Error model
Degree of violation
1
0.5
0
-0.5
-1
5
0
-5
x2
-5
0
5
x1



Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

`plotFcn` — Plot function

function handle

Plot function, specified as a function handle.

There are several built-in plot functions:

Model Plots — Apply When D ≤ 2	Description
`@plotAcquisitionFunction`	Plot the acquisition function surface.

Model Plots — Apply When D ≤ 2	Description
`@plotConstraintModels`	Plot each constraint model surface. Negative values indicate feasible points. Also plot a *P*(feasible) surface. Also plot the error model, if it exists, which ranges from –1 to 1. Negative values mean that the model probably does not error, positive values mean that it probably does error. The model is: Plotted error = 2*Probability(error) – 1.
`@plotObjectiveEvaluationTimeModel`	Plot the objective function evaluation time model surface.
`@plotObjectiveModel`	Plot the `fun` model surface, the estimated location of the minimum, and the location of the next proposed point to evaluate. For one-dimensional problems, plot envelopes one credible interval above and below the mean function, and envelopes one noise standard deviation above and below the mean.

Trace Plots — Apply to All D	Description
`@plotObjective`	Plot each observed function value versus the number of function evaluations.
`@plotObjectiveEvaluationTime`	Plot each observed function evaluation run time versus the number of function evaluations.
`@plotMinObjective`	Plot the minimum observed and estimated function values versus the number of function evaluations.
`@plotElapsedTime`	Plot three curves: the total elapsed time of the optimization, the total function evaluation time, and the total modeling and point selection time, all versus the number of function evaluations.

You can include a handle to your own plot functions. For details, see “Bayesian Optimization Plot Functions” on page 10-11.

Example: `@plotObjective`

Data Types: `function_handle`

Alternative Functionality

You can specify plot functions in the `bayesopt` `PlotFcn` name-value pair. This allows you to monitor the progress of the optimization.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

plotDiagnostics

Plot observation diagnostics of generalized linear regression model

Syntax

```
plotDiagnostics(mdl)
plotDiagnostics(mdl,plottype)
plotDiagnostics(mdl,plottype,Name,Value)
h = plotDiagnostics( ___ )
```

Description

`plotDiagnostics` creates a plot of observation diagnostics, such as leverage and Cook's distance, to identify outliers and influential observations.

`plotDiagnostics(mdl)` creates a leverage plot of the generalized linear regression model (`mdl`) observations. A dotted line in the plot represents the recommended threshold values.

`plotDiagnostics(mdl,plottype)` specifies the type of observation diagnostics `plottype`.

`plotDiagnostics(mdl,plottype,Name,Value)` specifies the graphical properties of diagnostic data points using one or more name-value pair arguments. For example, you can specify the marker symbol and size for the data points.

`h = plotDiagnostics( ___ )` returns graphics objects for the lines or contour in the plot using any of the input argument combinations in the previous syntaxes. Use `h` to modify the properties of a specific line or contour after you create the plot. For a list of properties, see Line Properties and Contour Properties.

Examples

Find Outliers Using Leverage and Cook's Distance

Create leverage and Cook's distance plots of a fitted generalized linear model, and find the outliers.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

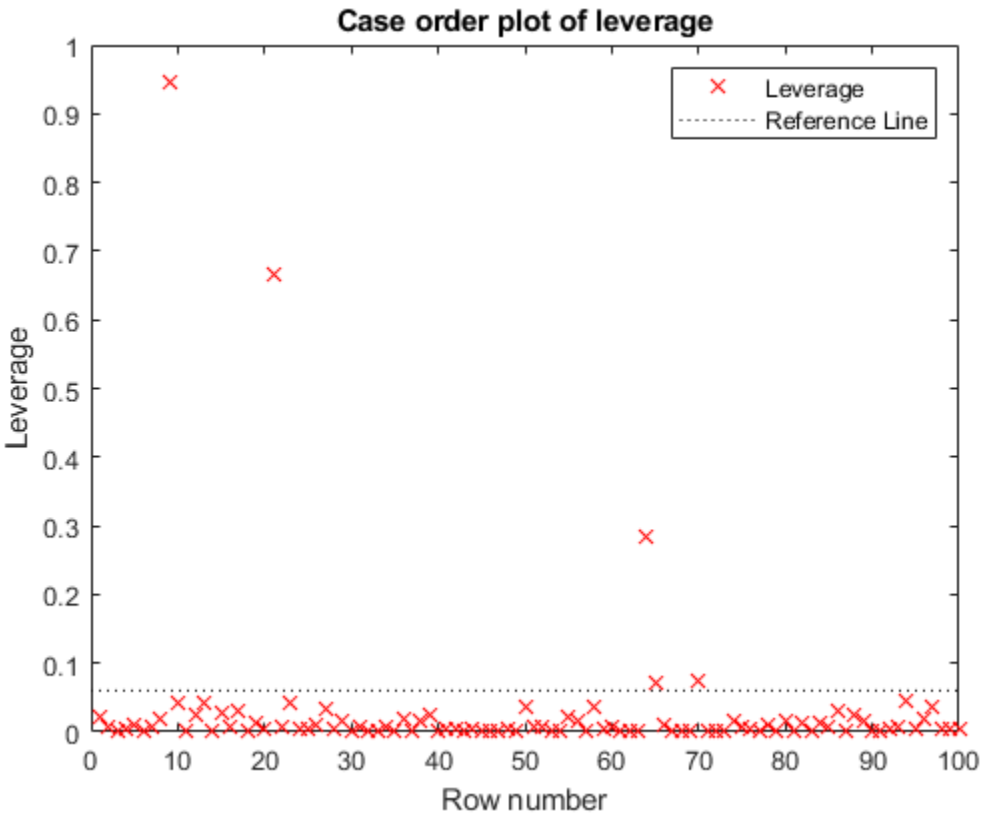
```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

Create a leverage plot.

```
plotDiagnostics(mdl)
legend('show') % Show the legend
```



The dotted line represents the recommended threshold value 2*p/n, where p is the number of coefficients, and n is the number of observations. Find the threshold value using the `NumCoefficients` and `NumObservations` properties.

```
t_leverage = 2*mdl.NumCoefficients/mdl.NumObservations
```

```
t_leverage = 0.0600
```

Find the observations with leverage values that exceed the threshold value.

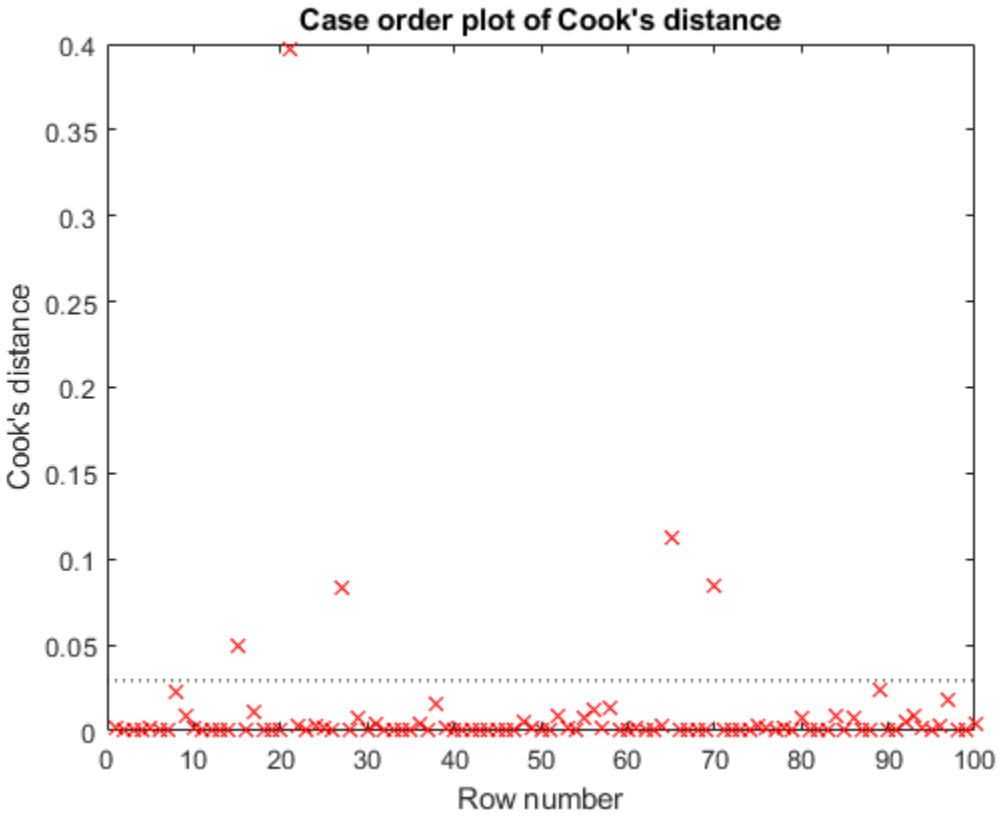
```
find(mdl.Diagnostics.Leverage > t_leverage)
```

```
ans = 5×1

     9
    21
    64
    65
    70
```

You can also find an observation number by using a data tip. Select the data points above the threshold line to display their data tips. The data tip includes the x-axis and y-axis values for the selected point, along with the observation number.

Plot the Cook's distance values.

```
plotDiagnostics(mdl,'cookd')
```



The dotted line represents the recommended threshold value. Compute the threshold value `t_cookd`.

```
t_cookd = 3*mean(mdl.Diagnostics.CooksDistance')
```

```
t_cookd = 0.0294
```

Find the observations with the Cook's distance values that exceed the threshold value.

```
find(mdl.Diagnostics.CooksDistance > t_cookd)
```

```
ans = 5×1

    15
    21
    27
    65
    70
```

Three observations (21, 65, and 70) are outliers by both measures, but some points (9, 15, 27, and 64) are outliers by only one measure.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

`plottype` — Type of plot

`'leverage'` (default) | `'contour'` | `'cookd'`

Type of plot, specified as one of the values in this table.

Value	Plot Type	Dotted Reference Line in Plot	Purpose
`'contour'`	Residual vs. leverage with overlaid contours of Cook's distance	Contours of Cook's distance	Identify observations with large residual values, high leverage, and large Cook's distance values.
`'cookd'`	Cook's distance	Recommended threshold, computed by `3*mean(mdl.Diagnostics.CooksDistance)`	Identify observations with large Cook's distance values.
`'leverage'`	Leverage	Recommended threshold, computed by `2*p/n`, where `p` is the number of coefficients (`mdl.NumCoefficients`) and `n` is the number of observations (`mdl.NumObservations`)	Identify high leverage observations.

For `'cookd'` and `'leverage'`, the *x*-axis is the row number (case order) of observations.

The `Diagnostics` property of `mdl` contains the diagnostic values used by `plotDiagnostics` to create plots.

For more information about observation diagnostics, see “Cook’s Distance” on page 33-4600 and “Leverage” on page 33-4601.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','o'`

Note The graphical properties listed here are only a subset. For a complete list, see Line Properties. The specified properties determine the appearance of diagnostic data points.

`Color` — Line color

RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	

RGB Triplet	Hexadecimal Color Code	Appearance
[0.4660 0.6740 0.1880]	'#77AC30'	
[0.3010 0.7450 0.9330]	'#4DBEEE'	
[0.6350 0.0780 0.1840]	'#A2142F'	

Example: 'Color','blue'

LineWidth — Line width
positive value

Line width, specified as the comma-separated pair consisting of 'LineWidth' and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: 'LineWidth',0.75

Marker — Marker symbol
'o' | '+' | '*' | '.' | 'x' | ...

Marker symbol, specified as the comma-separated pair consisting of 'Marker' and one of the values in this table.

Value	Description
'o'	Circle
'+'	Plus sign
'*'	Asterisk
'.'	Point
'x'	Cross
'_'	Horizontal line
'|'	Vertical line
'square' or 's'	Square
'diamond' or 'd'	Diamond
'^'	Upward-pointing triangle
'v'	Downward-pointing triangle
'>'	Right-pointing triangle
'<'	Left-pointing triangle
'pentagram' or 'p'	Five-pointed star (pentagram)
'hexagram' or 'h'	Six-pointed star (hexagram)
'none'	No markers

Example: 'Marker','+'

MarkerEdgeColor — Marker outline color
'auto' (default) | 'none' | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of 'MarkerEdgeColor' and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the Color name-value pair argument.

The default value of `'auto'` uses the same color specified by using `'Color'`.

Example: `'MarkerEdgeColor','blue'`

`MarkerFaceColor` — Marker fill color
`'none'` (default) | `'auto'` | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of `'MarkerFaceColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The `'auto'` value uses the same color specified by using `'Color'`.

Example: `'MarkerFaceColor','blue'`

`MarkerSize` — Marker size
6 (default) | positive value

Marker size, specified as the comma-separated pair consisting of `'MarkerSize'` and a positive value in points.

Example: `'MarkerSize',2`

Output Arguments

`h` — Graphics objects
graphics array

Graphics objects corresponding to the lines or contour in the plot, returned as a graphics array. Use dot notation to query and set properties of the graphics objects. For details, see Line Properties and Contour Properties.

You can use name-value pair arguments to specify the appearance of diagnostic data points corresponding to the first graphics object `h(1)`.

More About

Cook's Distance

Cook's distance is the scaled change in fitted values, which is useful for identifying outliers in the observations for predictor variables. Cook's distance shows the influence of each observation on the fitted response values. An observation with Cook's distance larger than three times the mean Cook's distance might be an outlier.

The Cook's distance D_i of observation i is

$$D_i = w_i \frac{e_i^2}{p\widehat{\varphi}} \frac{h_{ii}}{(1-h_{ii})^2},$$

where

- $\widehat{\varphi}$ is the dispersion parameter (estimated or theoretical).
- e_i is the linear predictor residual, $g(y_i) - x_i\widehat{\beta}$, where

- g is the link function.
- y_i is the observed response.
- x_i is the observation.
- $\widehat{\beta}$ is the estimated coefficient vector.

- p is the number of coefficients in the regression model.
- h_{ii} is the ith diagonal element of the Hat Matrix on page 33-4601 H.

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

Hat Matrix

The hat matrix is a projection matrix that projects the vector of response observations onto the vector of predictions.

The hat matrix H is defined in terms of the data matrix X and a diagonal weight matrix W:

$$H = X(X^TWX)^{-1}X^TW^T.$$

W has diagonal elements w_i:

$$w_i = \frac{g'(\mu_i)}{\sqrt{V(\mu_i)}},$$

where

- g is the link function mapping y_i to $x_i b$.
- g' is the derivative of the link function g.
- V is the variance function.
- μ_i is the ith mean.

The diagonal elements H_{ii} satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where n is the number of observations (rows of X), and p is the number of coefficients in the regression model.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.
- Use `legend('show')` to show the pre-populated legend.

Alternative Functionality

A `GeneralizedLinearModel` object provides multiple plotting functions.

- When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
- After fitting a model, use `plotPartialDependence` to understand the effect of a particular predictor. Also, use `plotSlice` to plot slices through the prediction surface.

References

[1] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, Fourth Edition. Chicago: McGraw-Hill Irwin, 1996.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `plotPartialDependence` | `plotResiduals` | `plotSlice`

Topics

“Diagnostic Plots” on page 12-17
“Generalized Linear Models” on page 12-9

Introduced in R2012a

plotDiagnostics

Plot observation diagnostics of linear regression model

Syntax

```
plotDiagnostics(mdl)
plotDiagnostics(mdl,plottype)
plotDiagnostics(mdl,plottype,Name,Value)
h = plotDiagnostics( ___ )
```

Description

`plotDiagnostics` creates a plot of observation diagnostics such as leverage, Cook's distance, and delete-1 statistics to identify outliers and influential observations.

`plotDiagnostics(mdl)` creates a leverage plot of the linear regression model (`mdl`) observations. A dotted line in the plot represents the recommended threshold values.

`plotDiagnostics(mdl,plottype)` specifies the type of observation diagnostics `plottype`.

`plotDiagnostics(mdl,plottype,Name,Value)` specifies the graphical properties of diagnostic data points using one or more name-value pair arguments. For example, you can specify the marker symbol and size for the data points.

`h = plotDiagnostics( ___ )` returns graphics objects for the lines or contour in the plot using any of the input argument combination in the previous syntaxes. Use `h` to modify the properties of a specific line or contour after you create the plot. For a list of properties, see Line Properties and Contour Properties.

Examples

Find Outliers Using Leverage and Cook's Distance

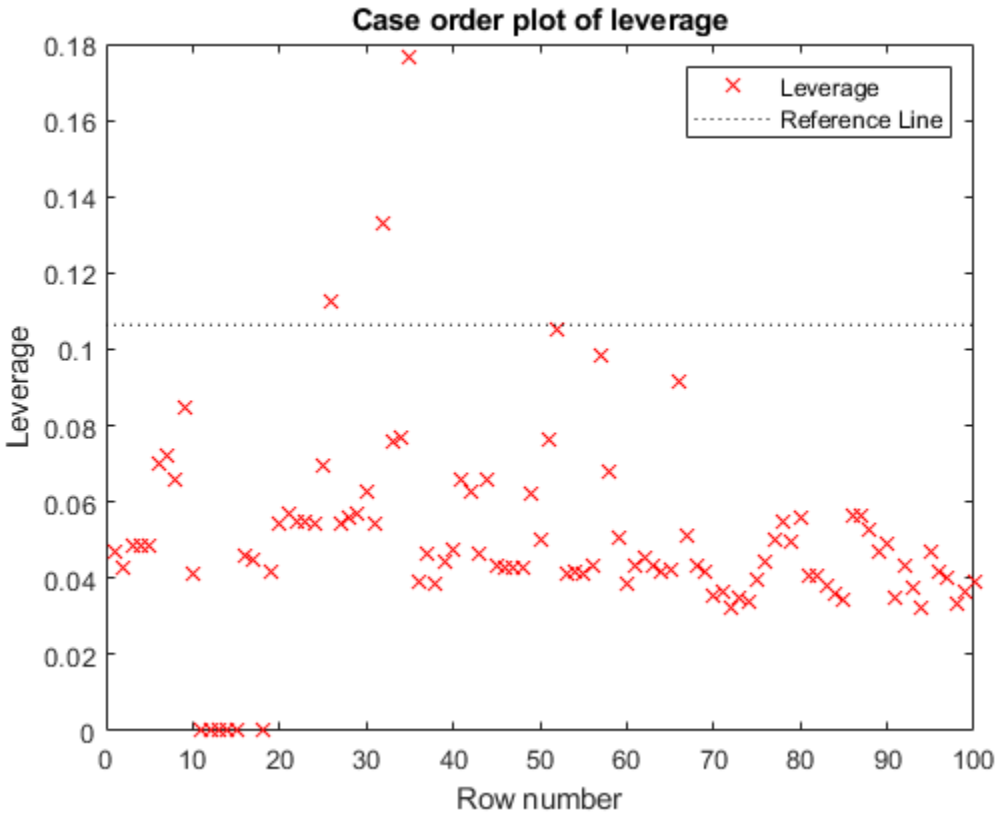
Plot the leverage values and Cook's distances of observations and find the outliers.

Load the `carsmall` data set and fit a linear regression model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Plot the leverage values.

```
plotDiagnostics(mdl)
legend('show') % Show the legend
```



The dotted line represents the recommended threshold value 2*p/n, where p is the number of coefficients, and n is the number of observations. Find the threshold value using the `NumCoefficients` and `NumObservations` properties.

```
t_leverage = 2*mdl.NumCoefficients/mdl.NumObservations
```

```
t_leverage = 0.1064
```

Find the observations with leverage values that exceed the threshold value.

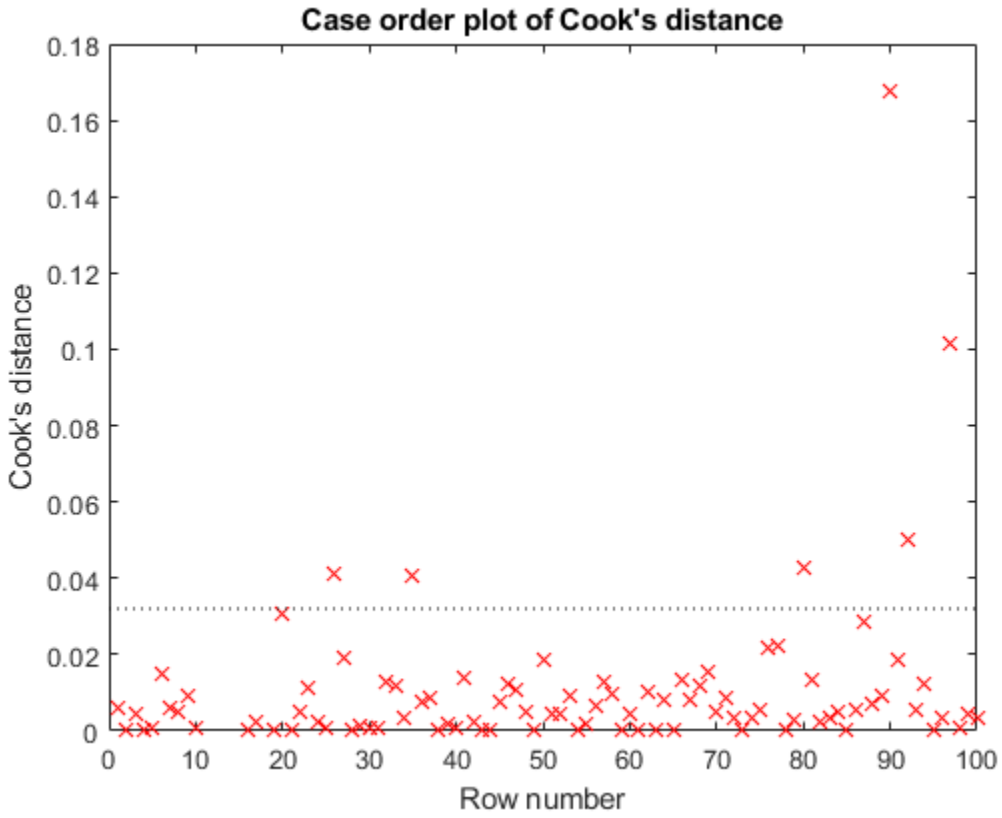
```
find(mdl.Diagnostics.Leverage > t_leverage)
```

```
ans = 3×1

    26
    32
    35
```

You can also find an observation number by using a data tip. Select the data points above the threshold line to display their data tips. The data tip includes the x-axis and y-axis values for the selected point, along with the observation number.

Plot the Cook's distance values.

```
plotDiagnostics(mdl,'cookd')
```



The dotted line represents the recommended threshold value. Compute the threshold value `t_cookd`.

```
t_cookd = 3*mean(mdl.Diagnostics.CooksDistance,'omitnan')
```

```
t_cookd = 0.0320
```

Find the observations with the Cook's distance values that exceed the threshold value.

```
find(mdl.Diagnostics.CooksDistance > t_cookd)
```

```
ans = 6×1

    26
    35
    80
    90
    92
    97
```

Two observations (26 and 35) are outliers by both measures, but some points (32, 80, 90, 92, and 97) are outliers by only one measure.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`plottype` — Type of plot
`'leverage'` (default) | `'contour'` | `'cookd'` | `'covratio'` | `'dfbetas'` | `'dffits'` | `'s2_i'`

Type of plot, specified as one of the values in this table.

Value	Plot Type	Dotted Reference Line in Plot	Purpose
`'contour'`	Residual vs. leverage with overlaid contours of Cook's distance	Contours of Cook's distance	Identify observations with large residual values, high leverage, and large Cook's distance values.
`'cookd'`	Cook's distance	Recommended threshold, computed by `3*mean(mdl.Diagnostics.CooksDistance)`	Identify observations with large Cook's distance values.
`'covratio'`	Delete-1 ratio of determinant of covariance	Recommended thresholds, computed by `1±3*p/n`, where `p` is the number of coefficients (`mdl.NumCoefficients`) and `n` is the number of observations (`mdl.NumObservations`)	Identify observations where the delete-1 statistic value is not in the range of the recommended thresholds.
`'dfbetas'`	Delete-1 scaled differences in coefficient estimates	Recommended threshold, computed by `3/sqrt(n)`	Identify observations with large delete-1 statistic values.
`'dffits'`	Delete-1 scaled differences in fitted values	Recommended threshold, computed by `2*sqrt(p/n)` in an absolute value	Identify observations with large delete-1 statistic values in an absolute value.
`'leverage'`	Leverage	Recommended threshold, computed by `2*p/n`	Identify high leverage observations.
`'s2_i'`	Delete-1 variance	Mean squared error (`mdl.MSE`)	Compare the delete-1 variance with the mean squared error.

For all plot types except `'contour'`, the *x*-axis is the row number (case order) of observations.

The `Diagnostics` property of `mdl` contains the diagnostic values used by `plotDiagnostics` to create plots.

For more information about observation diagnostics, see “Cook’s Distance” on page 33-4609, “Delete-1 Statistics” on page 33-4610, and “Leverage” on page 33-4610.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','o'`

Note The graphical properties listed here are only a subset. For a complete list, see Line Properties. The specified properties determine the appearance of diagnostic data points.

`Color` — Line color

RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	

RGB Triplet	Hexadecimal Color Code	Appearance
[0.9290 0.6940 0.1250]	'#EDB120'	
[0.4940 0.1840 0.5560]	'#7E2F8E'	
[0.4660 0.6740 0.1880]	'#77AC30'	
[0.3010 0.7450 0.9330]	'#4DBEEE'	
[0.6350 0.0780 0.1840]	'#A2142F'	

Example: 'Color','blue'

LineWidth — Line width
positive value

Line width, specified as the comma-separated pair consisting of 'LineWidth' and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: 'LineWidth',0.75

Marker — Marker symbol
'o' | '+' | '*' | '.' | 'x' | ...

Marker symbol, specified as the comma-separated pair consisting of 'Marker' and one of the values in this table.

Value	Description
'o'	Circle
'+'	Plus sign
'*'	Asterisk
'.'	Point
'x'	Cross
'_'	Horizontal line
'|'	Vertical line
'square' or 's'	Square
'diamond' or 'd'	Diamond
'^'	Upward-pointing triangle
'v'	Downward-pointing triangle
'>'	Right-pointing triangle
'<'	Left-pointing triangle
'pentagram' or 'p'	Five-pointed star (pentagram)
'hexagram' or 'h'	Six-pointed star (hexagram)
'none'	No markers

Example: 'Marker','+'

MarkerEdgeColor — Marker outline color
'auto' (default) | 'none' | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of `'MarkerEdgeColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The default value of `'auto'` uses the same color specified by using `'Color'`.

Example: `'MarkerEdgeColor','blue'`

`MarkerFaceColor` — Marker fill color
`'none'` (default) | `'auto'` | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of `'MarkerFaceColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The `'auto'` value uses the same color specified by using `'Color'`.

Example: `'MarkerFaceColor','blue'`

`MarkerSize` — Marker size
`6` (default) | positive value

Marker size, specified as the comma-separated pair consisting of `'MarkerSize'` and a positive value in points.

Example: `'MarkerSize',2`

Output Arguments

`h` — Graphics objects
graphics array

Graphics objects corresponding to the lines or contour in the plot, returned as a graphics array. Use dot notation to query and set properties of the graphics objects. For details, see Line Properties and Contour Properties.

You can use name-value pair arguments to specify the appearance of diagnostic data points corresponding to the first graphics object `h(1)`. If `plottype` is `'dfbetas'`, the plot includes a line object for each coefficient. Name-value pair arguments specify the line object properties of all coefficients. You can modify the properties of each coefficient separately by using the corresponding graphics object.

More About

Cook’s Distance

Cook’s distance is the scaled change in fitted values, which is useful for identifying outliers in the X values (observations for predictor variables). Cook’s distance shows the influence of each observation on the fitted response values. An observation with Cook’s distance larger than three times the mean Cook’s distance might be an outlier.

Each element in the Cook's distance D is the normalized change in the fitted response values due to the deletion of an observation. The Cook’s distance of observation i is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\,MSE},$$

where

- $\hat{y}_j$ is the jth fitted response value.
- $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook's distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p\,MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right),$$

where r_i is the ith residual, and h_{ii} is the ith leverage value.

For more details, see "Cook's Distance" on page 11-55.

Delete-1 Statistics

Delete-1 statistics are useful for finding the influence of each observation. These statistics capture the changes that would result from excluding each observation in turn from the fit. If the delete-1 statistics differ significantly from the model using all observations, then the observation is influential.

See "Delete-1 Statistics" on page 11-63 for the definitions and usages of the delete-1 statistics.

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. The hat matrix H is defined in terms of the data matrix X:

$H = X(X^TX)^{-1}X^T$.

The hat matrix is also known as the *projection matrix* because it projects the vector of observations y onto the vector of predictions $\hat{y}$, thus putting the "hat" on y.

Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

For more details, see "Hat Matrix and Leverage" on page 11-77.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the x-axis and y-axis values for the selected point, along with the observation name or number.
- Use `legend('show')` to show the pre-populated legend.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

References

[1] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, Fourth Edition. Chicago: McGraw-Hill Irwin, 1996.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `plotResiduals`

Topics

“Cook’s Distance” on page 11-55
“Delete-1 Statistics” on page 11-63
“Hat Matrix and Leverage” on page 11-77
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

plotDiagnostics

Class: `NonLinearModel`

Plot diagnostics of nonlinear regression model

Syntax

```
plotDiagnostics(mdl)
plotDiagnostics(mdl,plottype)
h = plotDiagnostics(...)
h = plotDiagnostics(mdl,plottype,Name,Value)
```

Description

`plotDiagnostics(mdl)` plots diagnostics from the `mdl` linear model using leverage as the plot type.

`plotDiagnostics(mdl,plottype)` plots diagnostics in a plot of type `plottype`.

`h = plotDiagnostics(...)` returns handles to the lines in the plot.

`h = plotDiagnostics(mdl,plottype,Name,Value)` plots with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`plottype`

Character vector or string scalar specifying the type of plot:

`'contour'`	Residual vs. leverage with overlaid Cook's contours
`'cookd'`	Cook's distance
`'leverage'`	Leverage (diagonal of Hat matrix)

Default: `'leverage'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note The plot property name-value pairs apply to the first returned handle `h(1)`.

`Color`

Color of the line or marker, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

`LineStyle`

Type of line, a Chart Line specification. For details, see `linespec`.

`LineWidth`

Width of the line or edges of filled area, in points, a positive scalar. One point is 1/72 inch.

Default: `0.5`

`MarkerEdgeColor`

Marker outline color, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

`MarkerFaceColor`

Fill color for filled markers, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

`MarkerSize`

Size of the marker in points, a strictly positive scalar. One point is 1/72 inch.

Output Arguments

`h`

Vector of handles to lines or patches in the plot.

Examples

Nonlinear Model Leverage Plot

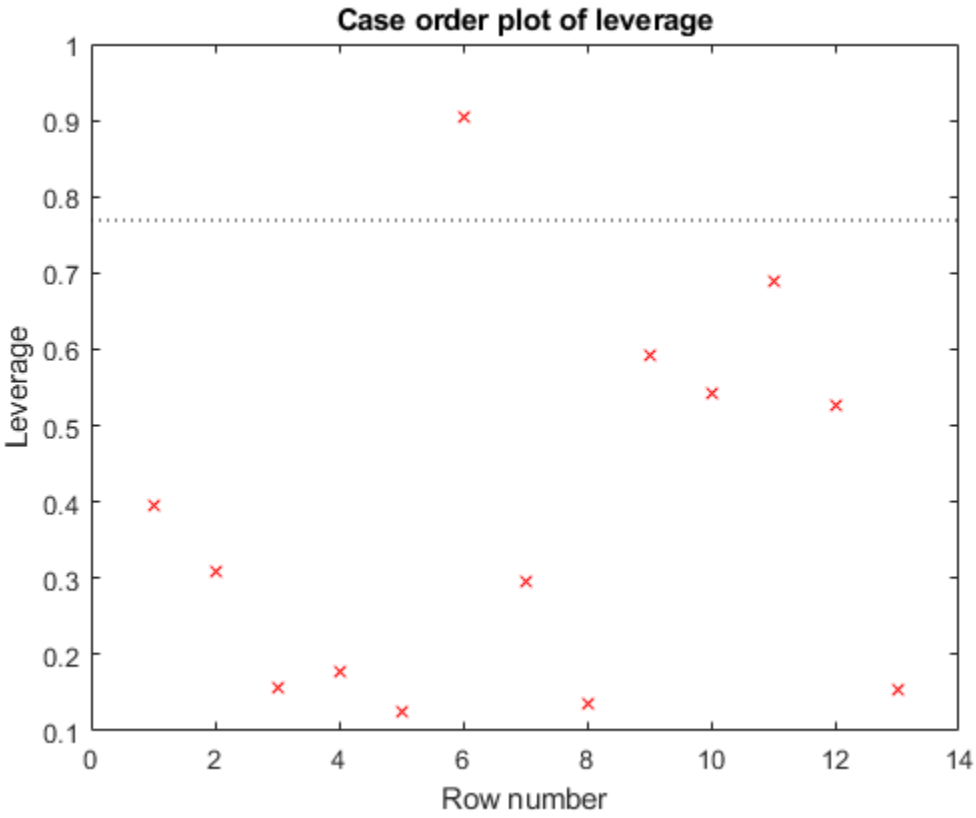
Create a leverage plot of a fitted nonlinear model, and find the points with high leverage.

Load the reaction data and fit a model of the reaction rate as a function of reactants.

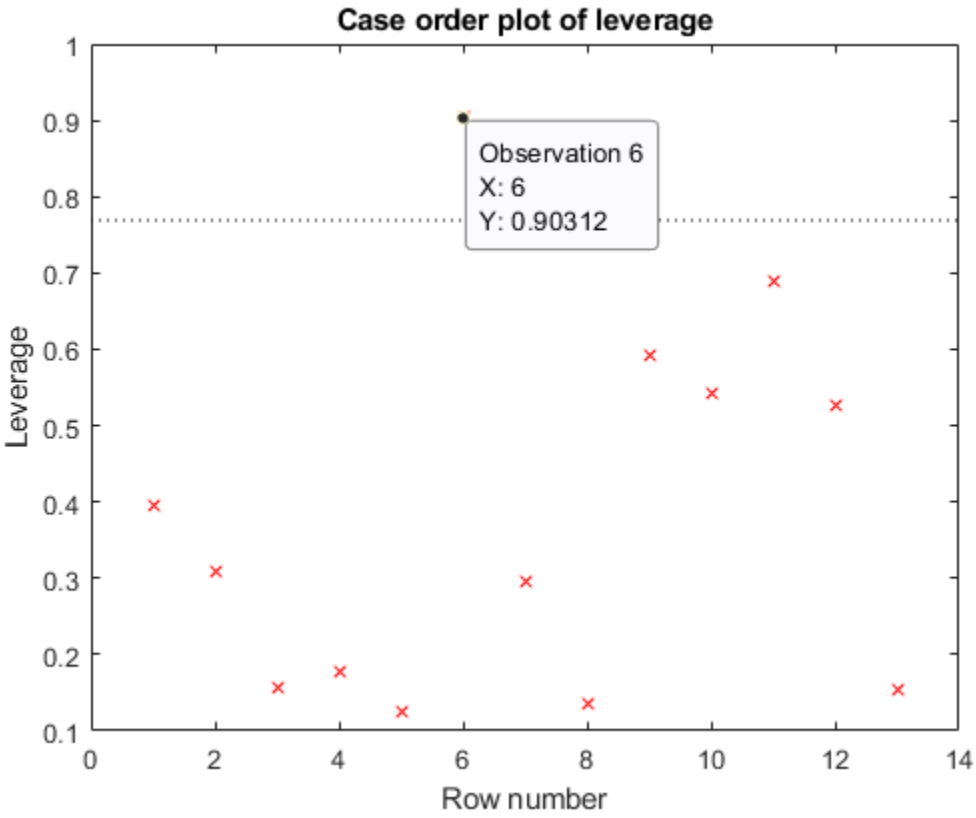
```
load reaction
mdl = fitnlm(reactants,rate,@hougen,[1 .05 .02 .1 2]);
```

Create a leverage plot of the fitted model.

```
plotDiagnostics(mdl)
```



Use data tips to examine the observation with high leverage. A data tip appears when you hover over a data point.



Alternatively, find the high-leverage observation at the command line.

```
find(mdl.Diagnostics.Leverage > 0.8)
```

```
ans =

     6
```

More About

Hat Matrix

The hat matrix H is defined in terms of the data matrix X and the Jacobian matrix J:

$$J_{i,j} = \left.\frac{\partial f}{\partial \beta_j}\right|_{x_i,\beta}$$

Here f is the nonlinear model function, and β is the vector of model coefficients.

The Hat Matrix H is

$$H = J(J^TJ)^{-1}J^T.$$

The diagonal elements H_{ii} satisfy

$$0 \le h_{ii} \le 1$$

$$\sum_{i=1}^{n} h_{ii} = p,$$

where n is the number of observations (rows of X), and p is the number of coefficients in the regression model.

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

Cook's Distance

The Cook's distance D_i of observation i is

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\,MSE},$$

where

- $\hat{y}_j$ is the jth fitted response value.
- $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook's distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p\,MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right),$$

where e_i is the ith residual.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the x-axis and y-axis values for the selected point, along with the observation name or number.

References

[1] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, Fourth Edition. Irwin, Chicago, 1996.

See Also

`NonLinearModel` | `plotResiduals`

Topics

“Examine Quality and Adjust the Fitted Nonlinear Model” on page 13-6
“Nonlinear Regression Workflow” on page 13-12
“Nonlinear Regression” on page 13-2

plotEffects

Package:

Plot main effects of predictors in linear regression model

Syntax

```
plotEffects(mdl)
h = plotEffects(mdl)
```

Description

`plotEffects(mdl)` creates an effects plot of the predictors in the linear regression model `mdl`. An effects plot shows the estimated main effect on page 33-4621 on the response from changing each predictor value, averaging out the effects of the other predictors. A horizontal line through an effect value indicates the 95% confidence interval for the effect value.

`h = plotEffects(mdl)` returns line objects. Use `h` to modify the properties of a specific line after you create the plot. For a list of properties, see Line Properties.

Examples

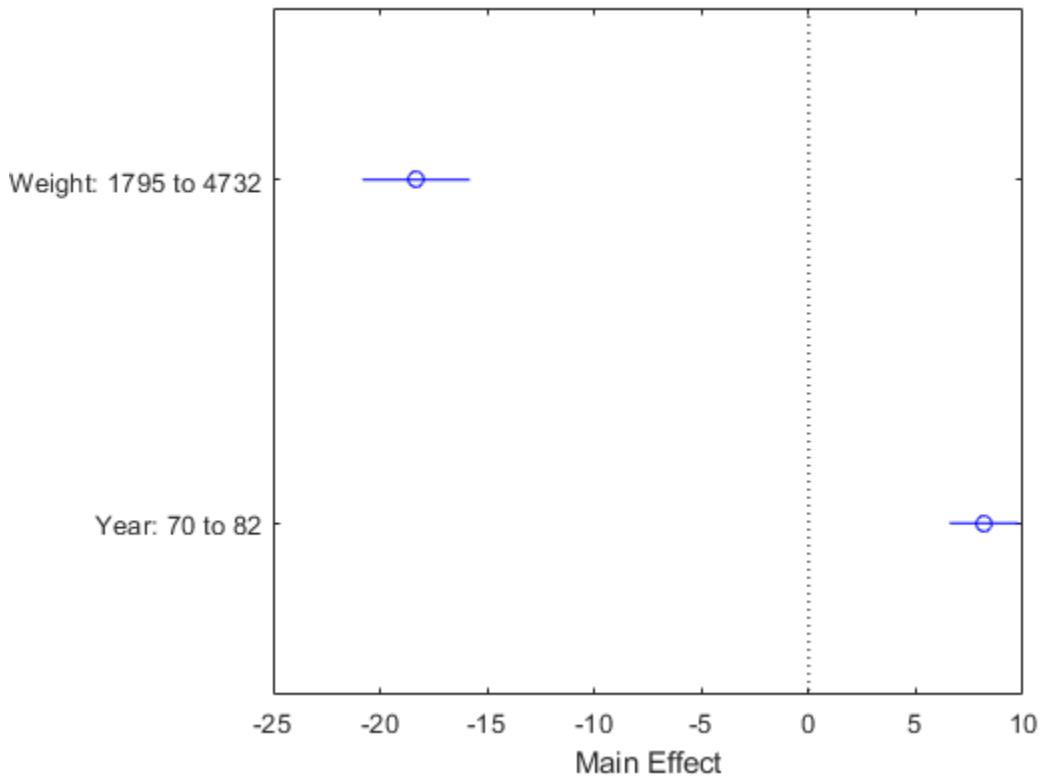
Effects Plot for Linear Regression Model

Load the `carsmall` data set and fit a linear regression model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Create an effects plot.

```
plotEffects(mdl)
```



The length of each horizontal line in the figure shows a 95% confidence interval for the effect on the response of the change shown for each predictor. For example, the estimated effect of changing `Year` from `70` to `82` is an increase of about 8, and is between 6 and 10 with 95% confidence.

Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

Output Arguments

`h` — Line objects

vector

Line objects, returned as a vector. `h(1)` corresponds to the circles that represent the effect estimates, and `h(j+1)` corresponds to the 95% confidence interval for the effect of predictor `j`. Use dot notation to query and set properties of line objects. For details, see Line Properties.

More About

Main Effect

An effect, or main effect, of a predictor represents an effect of one predictor on the response from changing the predictor value while averaging out the effects of the other predictors.

For a predictor variable x_s, the effect is defined by

$$g(x_{si}) - g(x_{sj}) ,$$

where g is an "Adjusted Response" on page 33-4621 function. The `plotEffects` function chooses the observations i and j as follows. For a categorical variable that is not ordinal, x_{si} and x_{sj} are the predictor values that produce the maximum and minimum adjusted responses, respectively, so that the effect value is always positive. For a numeric variable or an ordinal categorical variable, the function chooses two predictor values that produce the minimum and maximum adjusted responses where $x_{si} < x_{sj}$.

`plotEffects` plots the effect value and the 95% confidence interval of the effect value for each predictor variable.

Adjusted Response

An adjusted response function describes the relationship between the fitted response and a single predictor, with the other predictors averaged out by averaging the fitted values over the data used in the fit.

A regression model for the predictor variables (x_1, x_2, ..., x_p) and the response variable y has the form

$$y_i = f(x_{1i}, x_{2i}, ..., x_{pi}) + r_i,$$

where f is a fitted regression function and r is a residual. The subscript i represents the observation number.

The adjusted response function for the first predictor variable x_1, for example, is defined as

$$g(x_1) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_1, x_{2i}, x_{3i}, ..., x_{pi}\right),$$

where n is the number of observations. The adjusted response data value is the sum of the adjusted fitted value and the residual for each observation.

$$\tilde{y}_i = g(x_{1i}) + r_i .$$

`plotAdjustedResponse` plots the adjusted response function and the adjusted response data values for a selected predictor variable.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the x-axis and y-axis values for the selected point. Use the x-axis values to view an estimated effect value and its confidence bounds.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.

- When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
- When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
- After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `plotAdjustedResponse` | `plotInteraction`

Topics

“Linear Regression with Interaction Effects” on page 11-44
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

plotInteraction

Package:

Plot interaction effects of two predictors in linear regression model

Syntax

```
plotInteraction(mdl,var1,var2)
plotInteraction(mdl,var1,var2,ptype)
h = plotInteraction(___)
```

Description

`plotInteraction(mdl,var1,var2)` creates a plot of the main effects on page 33-4630 of the two selected predictors `var1` and `var2` and their conditional effects on page 33-4631 in the linear regression model `mdl`. Horizontal lines through the effect values indicate their 95% confidence intervals.

`plotInteraction(mdl,var1,var2,ptype)` specifies the plot type `ptype`. For example, if `ptype` is `'predictions'`, then `plotInteraction` plots the adjusted response function as a function of the second predictor, with the first predictor fixed at specific values. For details, see “Conditional Effect” on page 33-4631.

`h = plotInteraction(___)` returns line objects using any of the input argument combinations in the previous syntaxes. Use `h` to modify the properties of a specific line after you create the plot. For a list of properties, see Line Properties.

Examples

Interaction Plot of Main Effects and Conditional Effects

Fit a model with an interaction term and create an interaction plot that shows the main effects and conditional effects.

Using the data in the `carsmall` data set, create response values that include an interaction term. First, load the data set and normalize the predictor data.

```
load carsmall
Acceleration = normalize(Acceleration);
Horsepower = normalize(Horsepower);
Displacement = normalize(Displacement);
```

Define a response variable that includes the interaction term `Acceleration*Horsepower`.

```
y = Acceleration + 4*Horsepower + Acceleration.*Horsepower + Displacement;
```

Add some noise to the response values.

```
rng('default') % For reproducibility
y = y + normrnd(10,0.25*std(y,'omitnan'),size(y));
```

Create a table that includes the predictor data and response values.

```
tbl = table(Acceleration,Horsepower,Displacement,y);
```

Fit a linear regression model.

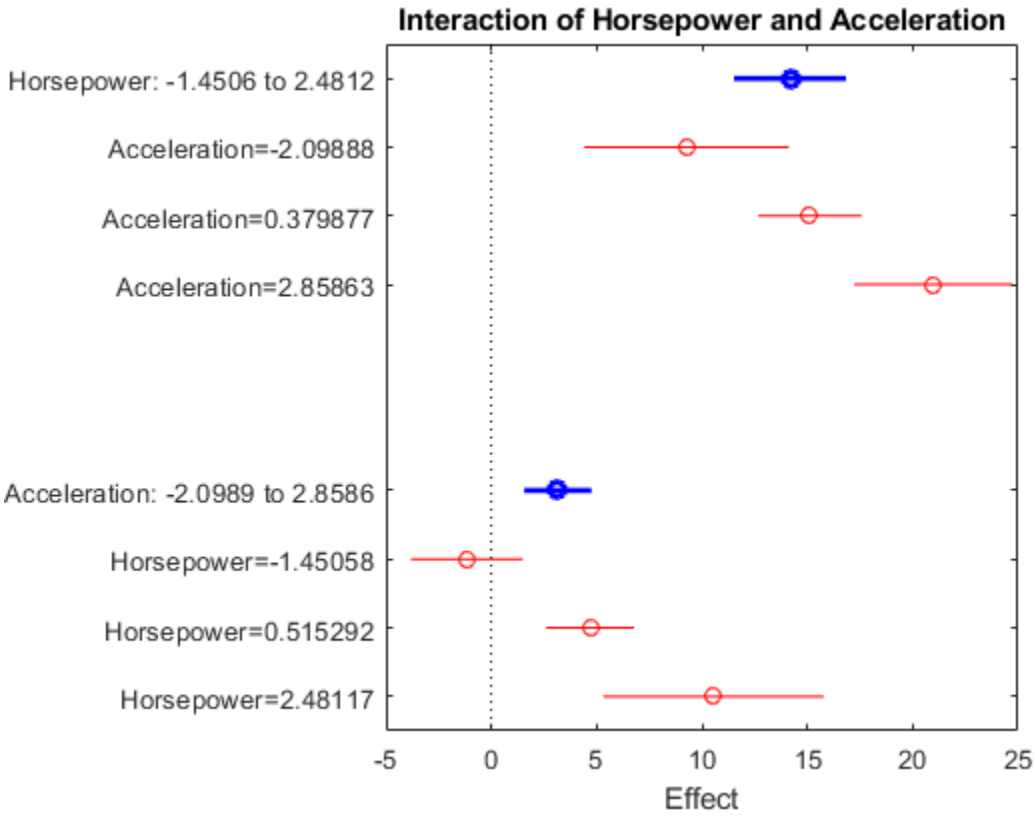
```
mdl = fitlm(tbl,'y ~ Acceleration + Horsepower + Acceleration*Horsepower + Displacement + Horsepo
```

```
mdl =
Linear regression model:
    y ~ 1 + Acceleration*Horsepower + Horsepower*Displacement

Estimated Coefficients:
                               Estimate       SE         tStat        pValue
                              __________    _______    _________    __________

    (Intercept)                  9.8652     0.16177       60.982     8.587e-77
    Acceleration                0.63726      0.1626       3.9191    0.00016967
    Horsepower                   3.6168        0.34       10.638     9.273e-18
    Displacement                0.95032     0.31828       2.9858     0.0036144
    Acceleration:Horsepower     0.60108      0.1851       3.2473     0.0016209
    Horsepower:Displacement   -0.0096069     0.20947    -0.045863       0.96352


Number of observations: 99, Error degrees of freedom: 93
Root Mean Squared Error: 1.07
R-squared: 0.93,  Adjusted R-Squared: 0.927
F-statistic vs. constant model: 249, p-value = 3.3e-52
```

`pValue` of the interaction term `Acceleration*Horsepower` is very small, meaning that the interaction term is statistically significant.

Create an interaction plot that shows the main effects and conditional effects of `Horsepower` and `Acceleration`.

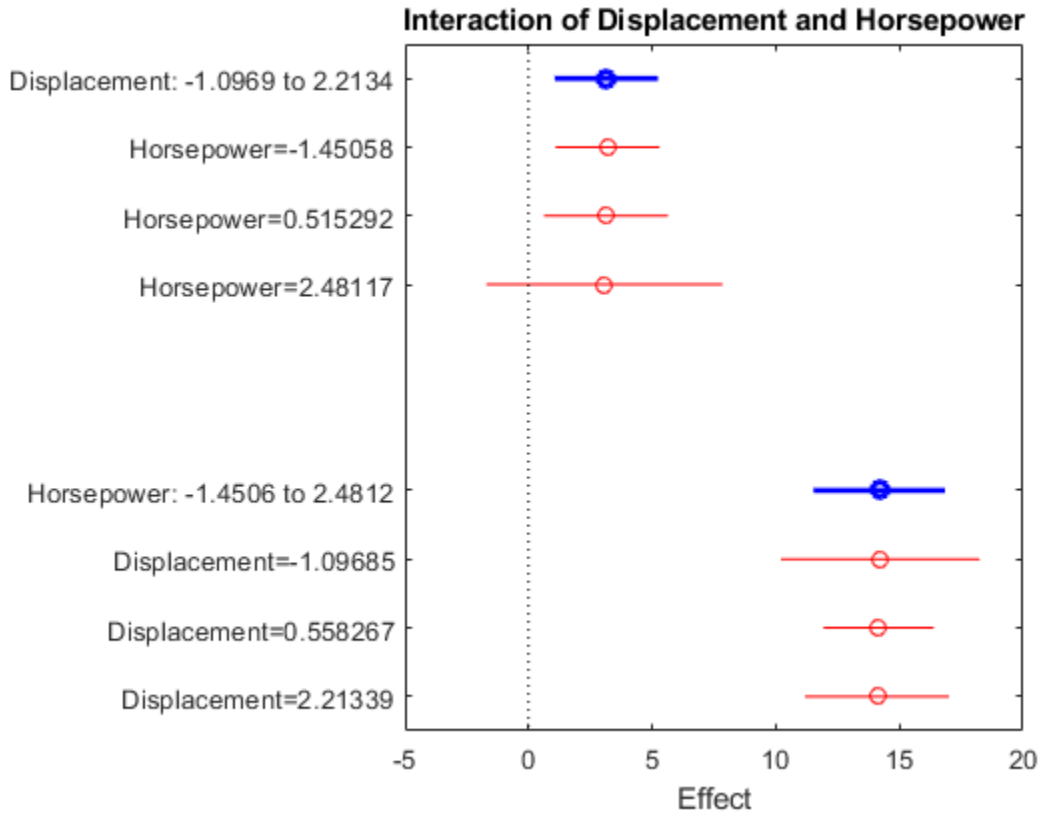
```
plotInteraction(mdl,'Horsepower','Acceleration')
```



For each predictor, the main effect point and its conditional effect points are not vertically aligned. Therefore, you cannot find any vertical lines that pass through the confidence intervals of the main and conditional effect points for each predictor. This plot indicates the existence of interaction effects on the response variable.

For comparison, create an interaction plot for `Displacement` and `Horsepower`. This *p*-value of this interaction term (`Displacement*Horsepower`) is large, meaning that the interaction term is not statistically significant.

```
plotInteraction(mdl,'Displacement','Horsepower')
```



For each predictor, the main effect point and its conditional effect points are aligned vertically. This plot indicates no interaction.

Interaction Plot of Adjusted Response Curve

Fit a model with an interaction term and create an interaction plot of adjusted response curves.

Using the data in the `carsmall` data set, create response values that include an interaction term. First, load the data set and normalize the predictor data.

```
load carsmall
Acceleration = normalize(Acceleration);
Horsepower = normalize(Horsepower);
Displacement = normalize(Displacement);
```

Define a response variable that includes the interaction term `Acceleration*Horsepower`.

```
y = Acceleration + 4*Horsepower + Acceleration.*Horsepower + Displacement;
```

Add some noise to the response values.

```
rng('default') % For reproducibility
y = y + normrnd(10,0.25*std(y,'omitnan'),size(y));
```

Create a table that includes the predictor data and response values.

```
tbl = table(Acceleration,Horsepower,Displacement,y);
```

Fit a linear regression model.

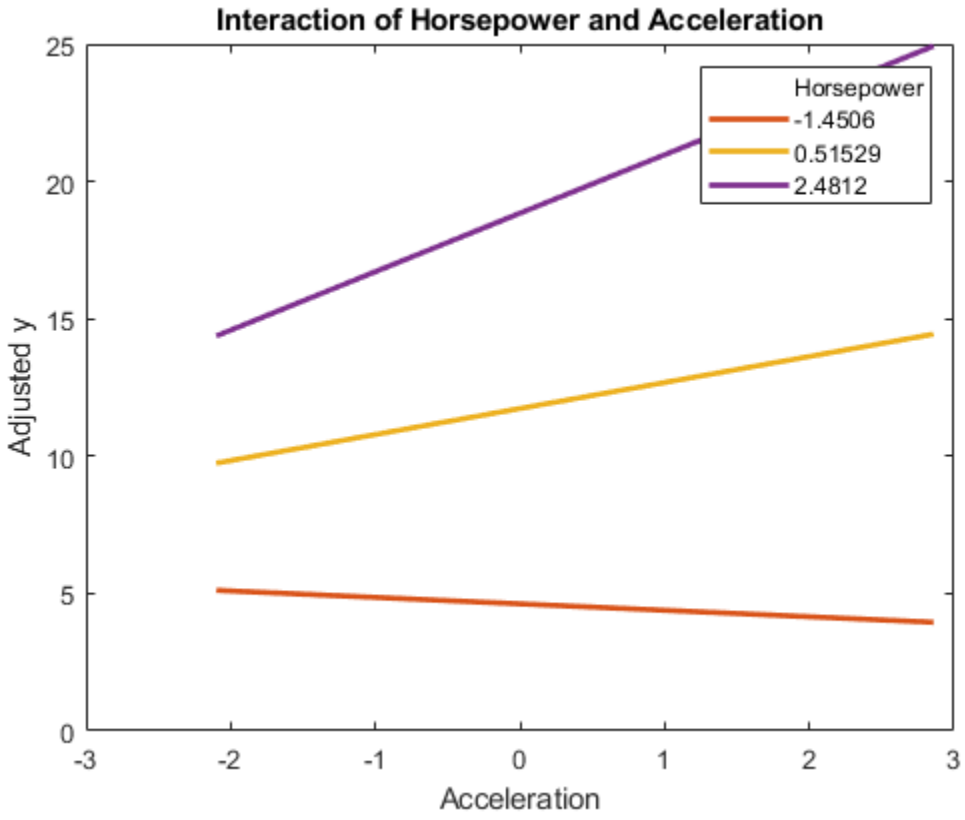
```
mdl = fitlm(tbl,'y ~ Acceleration + Horsepower + Acceleration*Horsepower + Displacement + Horsepo
```

```
mdl =
Linear regression model:
    y ~ 1 + Acceleration*Horsepower + Horsepower*Displacement

Estimated Coefficients:
                               Estimate       SE         tStat        pValue
                              __________    _______    _________    __________

    (Intercept)                  9.8652     0.16177       60.982     8.587e-77
    Acceleration                0.63726      0.1626       3.9191    0.00016967
    Horsepower                   3.6168        0.34       10.638     9.273e-18
    Displacement                0.95032     0.31828       2.9858     0.0036144
    Acceleration:Horsepower     0.60108      0.1851       3.2473     0.0016209
    Horsepower:Displacement   -0.0096069     0.20947    -0.045863       0.96352


Number of observations: 99, Error degrees of freedom: 93
Root Mean Squared Error: 1.07
R-squared: 0.93,  Adjusted R-Squared: 0.927
F-statistic vs. constant model: 249, p-value = 3.3e-52
```

`pValue` of the interaction term `Acceleration*Horsepower` is very small, meaning that the interaction term is statistically significant.

Create an interaction plot that shows the adjusted response function as a function of `Acceleration`, with `Horsepower` fixed at specific values.

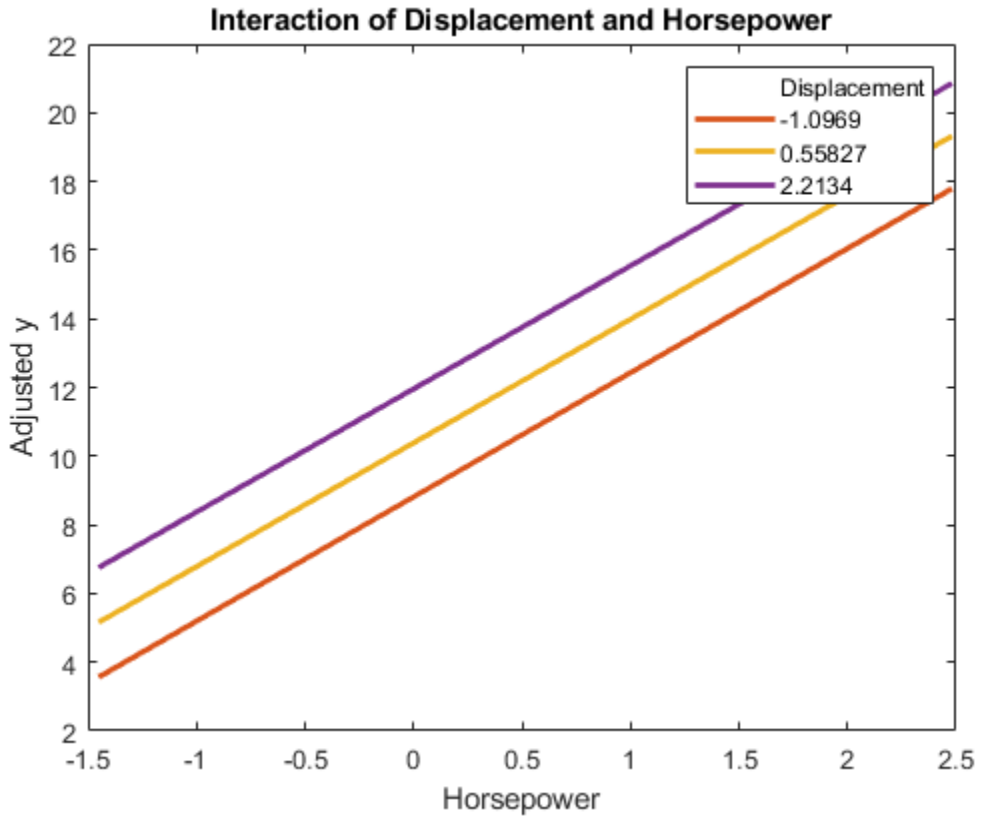
```
plotInteraction(mdl,'Horsepower','Acceleration','predictions')
```



The curves are not parallel. This plot indicates interactions between the predictors.

For comparison, create an interaction plot for the `Displacement` and `Horsepower`. The *p*-value of this interaction term (`Displacement*Horsepower`) is large, meaning that the interaction term is not statistically significant.

```
plotInteraction(mdl,'Displacement','Horsepower','predictions')
```



The curves are parallel, indicating no interaction.

Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`var1` — First variable for plot

character vector | string array | positive integer

First variable for the plot, specified as a character vector or string array of the variable name in `mdl.VariableNames` (`VariableNames` property of `mdl`), or a positive integer representing the index of a variable in `mdl.VariableNames`.

Data Types: `char` | `string` | `single` | `double`

`var2` — Second variable for plot

character vector | string array | positive integer

Second variable for the plot, specified as a character vector or string array of the variable name in `mdl.VariableNames` (`VariableNames` property of `mdl`), or a positive integer representing the index of a variable in `mdl.VariableNames`.

Data Types: `char` | `string` | `single` | `double`

`ptype` — Plot type
`'effects'` (default) | `'predictions'`

Plot type, specified as one of these values:

- `'effects'` — `plotInteraction` creates a plot of the main effects of the two selected predictors `var1` and `var2` and their conditional effects. Horizontal lines through the effect values indicate their 95% confidence intervals.
- `'predictions'` — `plotInteraction` plots the adjusted response function as a function of `var2`, with `var1` fixed at specific values.

For details, see “Main Effect” on page 33-4630 and “Conditional Effect” on page 33-4631.

Output Arguments

`h` — Line objects
vector

Line objects, returned as a vector. Use dot notation to query and set properties of the line objects. For details, see Line Properties.

If the plot type is `'effects'` (default), `h(1)` corresponds to the circles that represent the main effect estimates, and `h(2)` and `h(3)` correspond to the 95% confidence intervals for the two main effects. The remaining entries in `h` correspond to the conditional effects and their confidence intervals. The line objects associated with the main effects have the tag `'main'`. The line objects associated with the conditional effects of `var1` and `var2` have the tags `'conditional1'` and `'conditional2'`, respectively.

If the plot type is `'predictions'`, each entry in `h` corresponds to each curve on the plot.

More About

Main Effect

An effect, or main effect, of a predictor represents an effect of one predictor on the response from changing the predictor value while averaging out the effects of the other predictors.

For a predictor variable x_s, the effect is defined by

$$g(x_{si}) - g(x_{sj}),$$

where g is an “Adjusted Response” on page 33-4631 function. The `plotEffects` function chooses the observations i and j as follows. For a categorical variable that is not ordinal, x_{si} and x_{sj} are the predictor values that produce the maximum and minimum adjusted responses, respectively, so that the effect value is always positive. For a numeric variable or an ordinal categorical variable, the function chooses two predictor values that produce the minimum and maximum adjusted responses where $x_{si} < x_{sj}$.

`plotEffects` plots the effect value and the 95% confidence interval of the effect value for each predictor variable.

Adjusted Response

An adjusted response function describes the relationship between the fitted response and a single predictor, with the other predictors averaged out by averaging the fitted values over the data used in the fit.

A regression model for the predictor variables (x_1, x_2, ..., x_p) and the response variable y has the form

$$y_i = f(x_{1i}, x_{2i}, ..., x_{pi}) + r_i,$$

where f is a fitted regression function and r is a residual. The subscript i represents the observation number.

The adjusted response function for the first predictor variable x_1, for example, is defined as

$$g(x_1) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_1, x_{2i}, x_{3i}, ..., x_{pi}\right),$$

where n is the number of observations. The adjusted response data value is the sum of the adjusted fitted value and the residual for each observation.

$$\tilde{y}_i = g(x_{1i}) + r_i\,.$$

`plotAdjustedResponse` plots the adjusted response function and the adjusted response data values for a selected predictor variable.

Conditional Effect

When a model contains an interaction term, the main effect of one predictor depends on the value of another predictor that interacts with it. In this case, a conditional effect of one predictor given a specific value of another is helpful in understanding the actual effect of both predictors. You can examine whether the effect of one predictor depends on the value of another by using conditional effect values.

To define a conditional effect, define the adjusted response function as a function of two predictor variables. For example, the adjusted response function of x_1 and x_2 is

$$h(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_1, x_2, x_{3i}, ..., x_{pi}\right),$$

where f is a fitted regression function, and n is the number of observations.

The conditional effect of one predictor (x_2) given a specific value of another predictor (x_{1k}) is defined by

$$h(x_{1k},x_{2i}) - h(x_{1k},x_{2j}).$$

To compute conditional effect values, `plotInteraction` chooses the observations i and j of x_2 in the same way as when the function computes the "Main Effect" on page 33-4630 and chooses the x_{1k} values. If x_1 is a categorical variable, then `plotInteraction` computes the conditional effect for all levels of x_1. If x_1 is a numeric variable, then `plotInteraction` computes the conditional effect for three values of x_1: the minimum value of x_1, the maximum value of x_1, and the average value of the minimum and maximum.

If the plot type is `'effects'` (default), `plotInteraction` plots the main effects of the two selected predictors, their conditional effects, and the 95% confidence bounds for the effect values.

If the plot type is `'predictions'`, `plotInteraction` plots the adjusted response function as a function of the second predictor, with the first predictor fixed at specific values. For example,

`plotInteraction(mdl,'x1','x2','predictions')` plots the curve of $h(x_{1k}, x_2)$ for each x_{1k} value.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `plotAdjustedResponse` | `plotEffects`

Topics

“Linear Regression with Interaction Effects” on page 11-44
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

plotLocalEffects

Package:

Plot local effects of terms in generalized additive model (GAM)

Syntax

```
plotLocalEffects(Mdl,queryPoint)
plotLocalEffects(Mdl,queryPoint,Name,Value)
plotLocalEffects( ___ )
```

Description

`plotLocalEffects(Mdl,queryPoint)` creates a bar graph showing the local effects of the terms in the generalized additive model `Mdl` on the prediction at the specified query point `queryPoint`.

`plotLocalEffects(Mdl,queryPoint,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'IncludeIntercept',true` specifies to include an intercept term in the bar graph.

`b = plotLocalEffects( ___ )` returns a bar graph object `b` using any of the input argument combinations in the previous syntaxes. Use `b` to query or modify Bar Properties of the bar graph after it is created.

Examples

Plot Local Effects

Train a univariate generalized additive classification model, which contains linear terms for predictors. Classify a new observation using a memory-efficient model object. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a univariate GAM that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

```
Mdl = fitcgam(X,Y);
```

`Mdl` is a `ClassificationGAM` model object.

Conserve memory by reducing the size of the trained model.

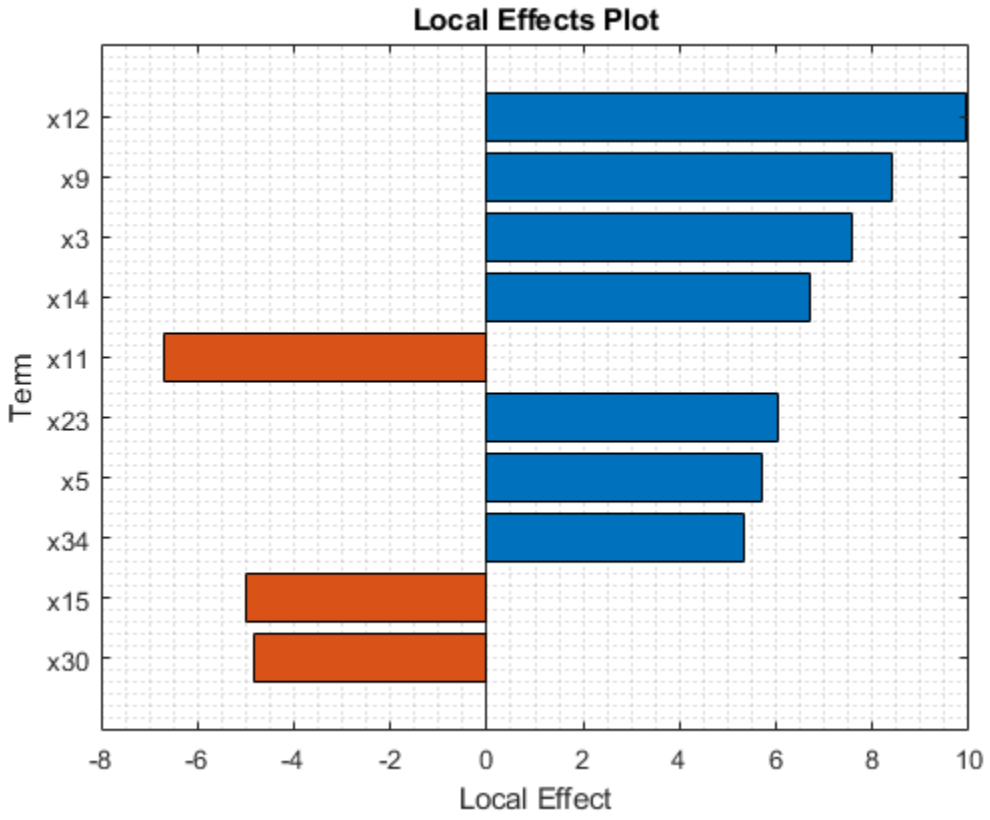
```
CMdl = compact(Mdl);
```

Classify the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction.

```
label = predict(CMdl,X(1,:))
```

```
label = 1x1 cell array
    {'g'}
```

```
plotLocalEffects(CMdl,X(1,:))
```



The `predict` function classifies the first observation `X(1,:)` as `'g'`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the 10 most important terms on the prediction. Each local effect value shows the contribution of each term to the classification score for `'g'`, which is the logit of the posterior probability that the classification is `'g'` for the observation.

Compare Local Effects in GAMs

Train a GAM for binary classification with both linear and interaction terms for predictors. Create local effects plot using both linear and interaction terms in the model, and then create a plot using only linear terms in the model. Specify whether to include interaction terms when creating the local effects plot.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

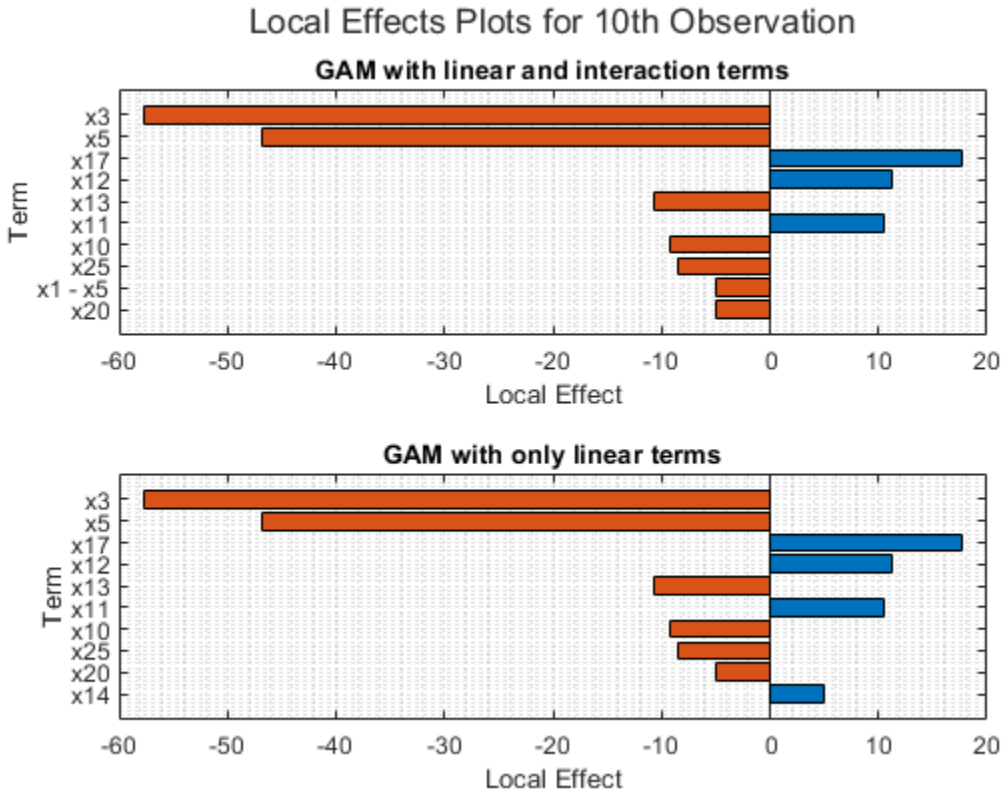
Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X,Y,'ClassNames',{'b','g'},'Interactions',10);
```

`Mdl` is a `ClassificationGAM` model object.

Create local effects plots for the 10th observation. Use both the linear and interaction terms in `Mdl` for the first plot, and use only the linear terms in `Mdl` for the second plot. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
t = tiledlayout(2,1);
title(t,'Local Effects Plots for 10th Observation')
nexttile
plotLocalEffects(Mdl,X(10,:))
title('GAM with linear and interaction terms')
nexttile
plotLocalEffects(Mdl,X(10,:),'IncludeInteractions',false)
title('GAM with only linear terms')
```



The plots display the 10 most important terms. Both plots include nine common terms and one uncommon term. The first plot includes the interaction term for `x1` and `x5`, whereas the second plot includes the linear term for `x14`.

Include Intercept Term in Local Effects Plot

Train a univariate GAM for regression, which contains linear terms for predictors. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 10 variables with information on the sales of properties in New York City in 2015. This example uses these variables to analyze the sale prices (`SALEPRICE`).

Preprocess the data set. Remove outliers, convert the `datetime` array (`SALEDATE`) to the month numbers, and move the response variable (`SALEPRICE`) to the last column.

```
idx = isoutlier(NYCHousing2015.SALEPRICE);
NYCHousing2015(idx,:) = [];
NYCHousing2015.SALEDATE = month(NYCHousing2015.SALEDATE);
NYCHousing2015 = movevars(NYCHousing2015,'SALEPRICE','After','SALEDATE');
```

Display the first three rows of the table.

```
head(NYCHousing2015,3)
```

```
ans=3×10 table
    BOROUGH    NEIGHBORHOOD       BUILDINGCLASSCATEGORY        RESIDENTIALUNITS    COMMERCIALUNI
    _______    ____________    ___________________________    ________________    _____________

       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   1
```

Train a univariate GAM for the sale prices. Specify the variables for `BOROUGH`, `NEIGHBORHOOD`, `BUILDINGCLASSCATEGORY`, and `SALEDATE` as categorical predictors.

```
Mdl = fitrgam(NYCHousing2015,'SALEPRICE','CategoricalPredictors',[1 2 3 9]);
```

`Mdl` is a `RegressionGAM` model object.

Display the estimated intercept (constant) term of `Mdl`.

```
Mdl.Intercept
```

```
ans = 3.7518e+05
```

The intercept term value is close to the average of the response variable in a regression GAM if the training data does not include `NaN` values. Compute average of the response variable.

```
mean(NYCHousing2015.SALEPRICE)
```

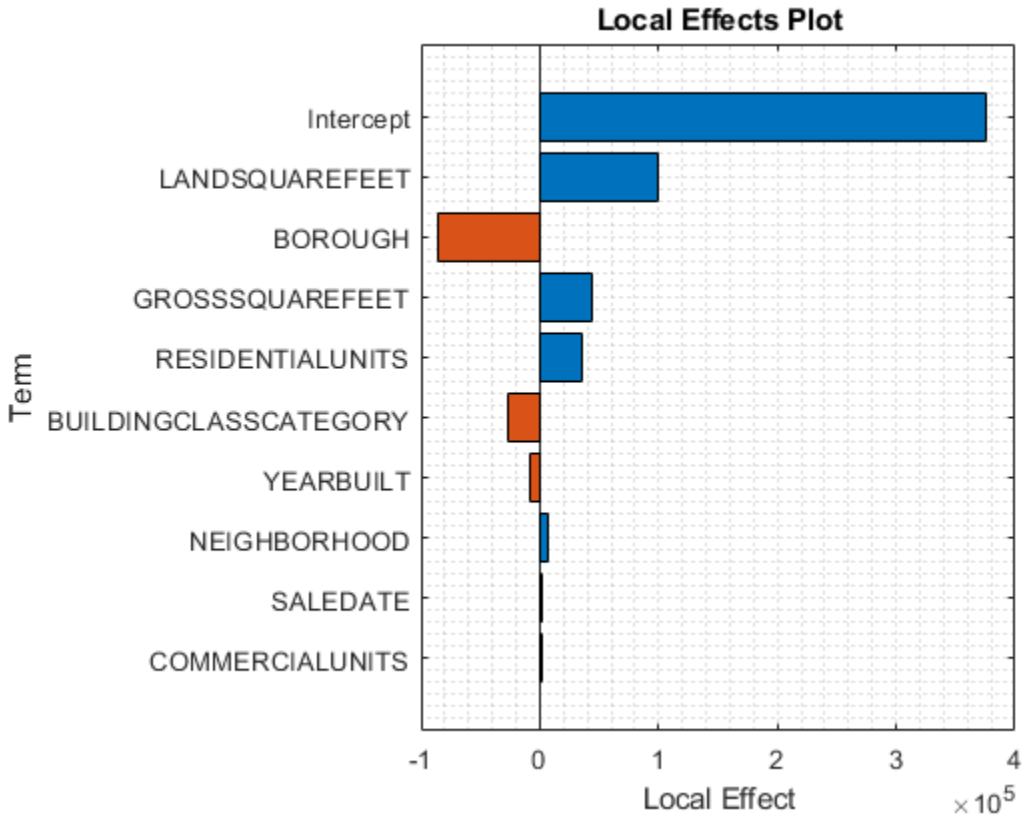
```
ans = 3.7518e+05
```

Predict the sale price for the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction. Specify `'IncludeIntercept',true` to include the intercept term in the plot.

```
yFit = predict(Mdl,NYCHousing2015(1,:))
```

```
yFit = 4.4421e+05
```

```
plotLocalEffects(Mdl,NYCHousing2015(1,:),'IncludeIntercept',true)
```



The `predict` function predicts the sale price for the first observation as `4.4421e5`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the terms in `Mdl` on the prediction. Each local effect value shows the contribution of each term to the predicted sale price.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `CompactClassificationGAM` model object | `RegressionGAM` model object | `CompactRegressionGAM` model object

Generalized additive model, specified as a `ClassificationGAM`, `CompactClassificationGAM`, `RegressionGAM`, or `CompactRegressionGAM` model object.

`queryPoint` — Query point

row vector of numeric values | single-row table

Query point at which `plotLocalEffects` plots the local effects, specified as a row vector of numeric values or a single-row table.

- For a row vector of numeric values:

- The variables that makes up the columns of `queryPoint` must have the same order as the predictor variables that trained `Mdl`.
- If you trained `Mdl` using a table (for example, `Tbl`), then `queryPoint` can be a numeric matrix if `Tbl` contains all numeric variables.

- For a single-row table:
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `queryPoint` must have the same variable names and data types as those in `Tbl`. However, the column order of `queryPoint` does not need to correspond to the column order of `Tbl`.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and the corresponding predictor variable names in `queryPoint` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `queryPoint` must be numeric vectors.
  - `queryPoint` can contain additional variables (response variables, observation weights, and so on), but `plotLocalEffects` ignores them.
  - `plotLocalEffects` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `plotLocalEffects(Mdl,queryPoint,'IncludeInteractions',false,'NumTerms',5)` specifies to create a bar plot containing the five most important linear terms for predictors in `Mdl` excluding the interaction terms in `Mdl`.

`IncludeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model in the plot, specified as `true` or `false`.

The default `'IncludeInteractions'` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`IncludeIntercept` — Flag to include intercept term

`false` (default) | `true`

Flag to include an intercept term of the model in the plot, specified as `true` or `false`.

Example: `'IncludeIntercept',true`

Data Types: `logical`

`NumTerms` — Number of terms to plot

`min(M,10)` where `M` is the number of terms in `Mdl` (default) | positive integer scalar

Number of terms to plot, specified as a positive integer scalar. `plotLocalEffects` plots the specified number of terms with the highest absolute local effect values.

Example: `'NumTerms',5` specifies to plot the five most important terms. `plotLocalEffects` determines the order of importance by using the absolute local effect values.

Data Types: `single` | `double`

See Also

`ClassificationGAM` | `RegressionGAM` | `plotPartialDependence`

Topics

“Train Generalized Additive Model for Binary Classification” on page 12-77
“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

plotPartialDependence

Package:

Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots

Syntax

```
plotPartialDependence(RegressionMdl,Vars)

plotPartialDependence(ClassificationMdl,Vars,Labels)

plotPartialDependence( ___ ,Data)
plotPartialDependence( ___ ,Name,Value)
ax = plotPartialDependence( ___ )
```

Description

`plotPartialDependence(RegressionMdl,Vars)` computes and plots the partial dependence between the predictor variables listed in `Vars` and the responses predicted by using the regression model `RegressionMdl`, which contains predictor data.

- If you specify one variable in `Vars`, the function creates a line plot of the partial dependence against the variable.
- If you specify two variables in `Vars`, the function creates a surface plot of the partial dependence against the two variables.

`plotPartialDependence(ClassificationMdl,Vars,Labels)` computes and plots the partial dependence between the predictor variables listed in `Vars` and the scores for the classes specified in `Labels` by using the classification model `ClassificationMdl`, which contains predictor data.

- If you specify one variable in `Vars` and one class in `Labels`, the function creates a line plot of the partial dependence against the variable for the specified class.
- If you specify one variable in `Vars` and multiple classes in `Labels`, the function creates a line plot for each class on one figure.
- If you specify two variables in `Vars` and one class in `Labels`, the function creates a surface plot of the partial dependence against the two variables.

`plotPartialDependence( ___ ,Data)` uses new predictor data `Data`. You can specify `Data` in addition to any of the input argument combinations in the previous syntaxes.

`plotPartialDependence( ___ ,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, if you specify `'Conditional','absolute'`, the `plotPartialDependence` function creates a figure including a PDP, a scatter plot of the selected predictor variable and predicted responses or scores, and an ICE plot for each observation.

`ax = plotPartialDependence( ___ )` returns the axes of the plot.

Examples

Create Partial Dependence Plot

Train a regression tree using the `carsmall` data set, and create a PDP that shows the relationship between a feature and the predicted responses in the trained regression tree.

Load the `carsmall` data set.

```
load carsmall
```

Specify `Weight`, `Cylinders`, and `Horsepower` as the predictor variables (`X`), and `MPG` as the response variable (`Y`).

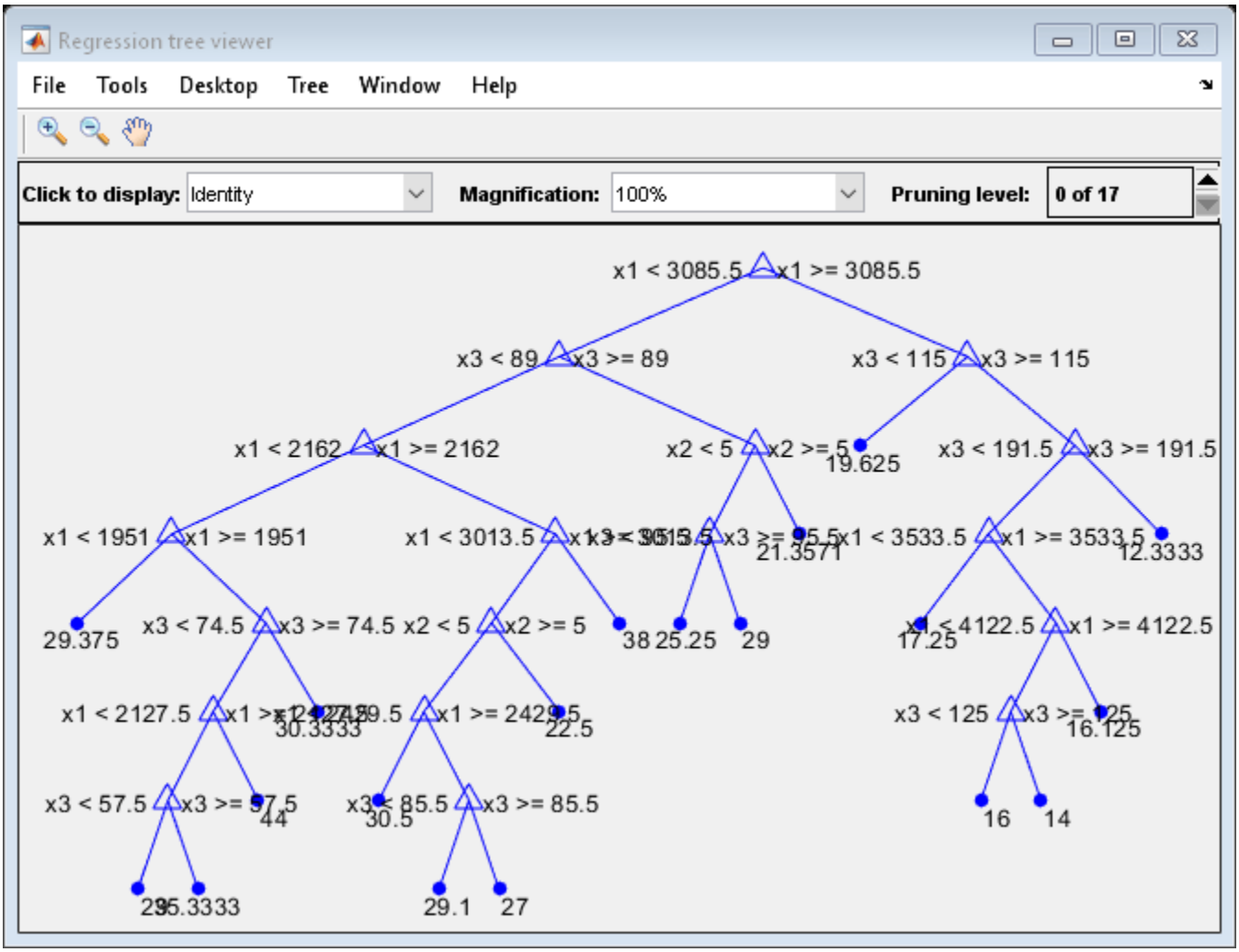
```
X = [Weight,Cylinders,Horsepower];
Y = MPG;
```

Train a regression tree using `X` and `Y`.

```
Mdl = fitrtree(X,Y);
```

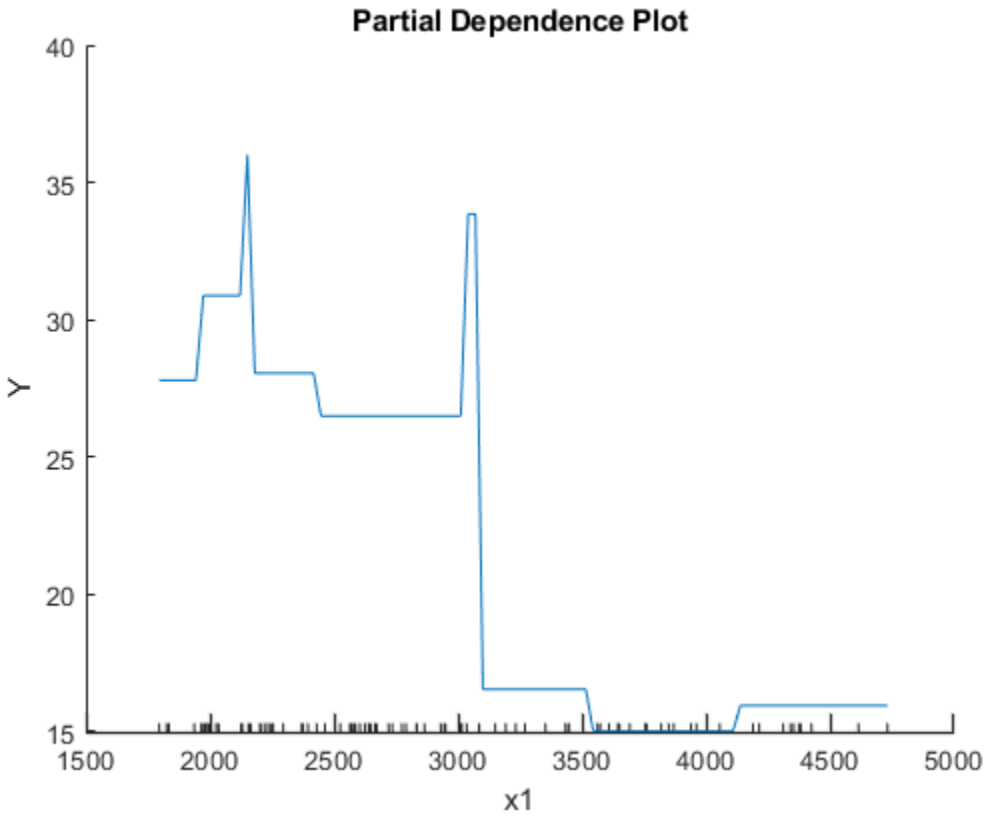
View a graphical display of the trained regression tree.

```
view(Mdl,'Mode','graph')
```



Create a PDP of the first predictor variable, `Weight`.

```
plotPartialDependence(Mdl,1)
```



The plotted line represents averaged partial relationships between `Weight` (labeled as `x1`) and `MPG` (labeled as `Y`) in the trained regression tree `Mdl`. The *x*-axis minor ticks represent the unique values in `x1`.

The regression tree viewer shows that the first decision is whether `x1` is smaller than 3085.5. The PDP also shows a large change near `x1` = 3085.5. The tree viewer visualizes each decision at each node based on predictor variables. You can find several nodes split based on the values of `x1`, but determining the dependence of `Y` on `x1` is not easy. However, the `plotPartialDependence` plots average predicted responses against `x1`, so you can clearly see the partial dependence of `Y` on `x1`.

The labels `x1` and `Y` are the default values of the predictor names and the response name. You can modify these names by specifying the name-value pair arguments `'PredictorNames'` and `'ResponseName'` when you train `Mdl` using `fitrtree`. You can also modify axis labels by using the `xlabel` and `ylabel` functions.

Create Partial Dependence Plot for Multiple Classes

Train a naive Bayes classification model with the `fisheriris` data set, and create a PDP that shows the relationship between the predictor variable and the predicted scores (posterior probabilities) for multiple classes.

Load the `fisheriris` data set, which contains species (`species`) and measurements (`meas`) on sepal length, sepal width, petal length, and petal width for 150 iris specimens. The data set contains 50 specimens from each of three species: setosa, versicolor, and virginica.

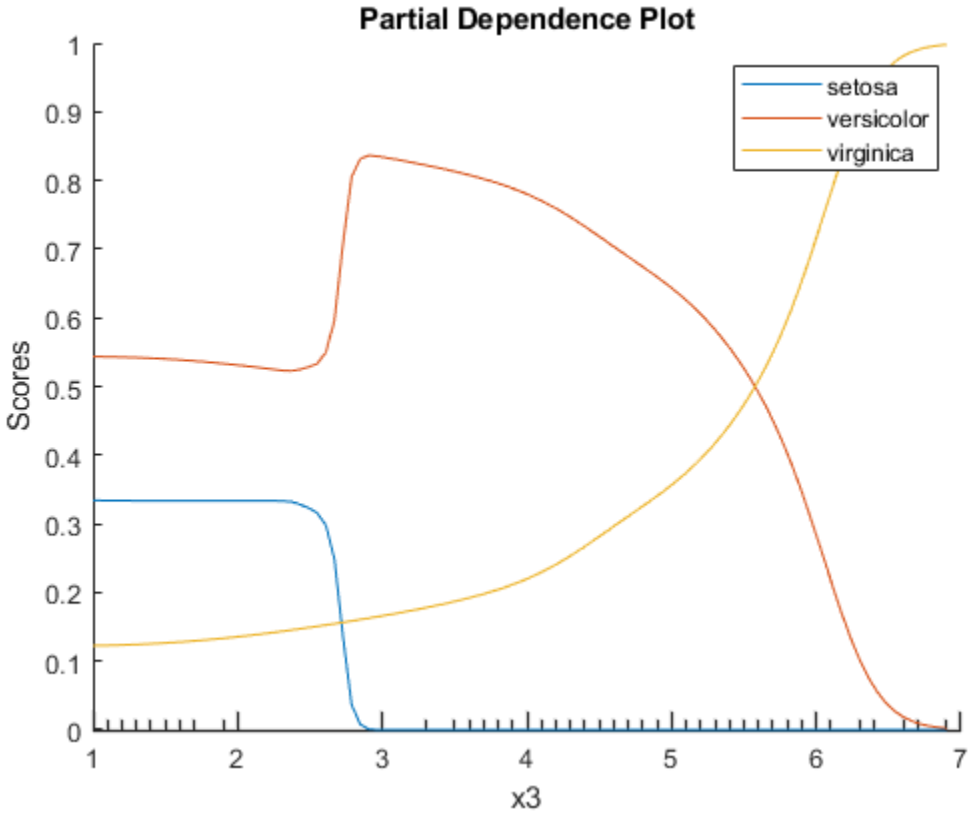
```
load fisheriris
```

Train a naive Bayes classification model with `species` as the response and `meas` as predictors.

```
Mdl = fitcnb(meas,species);
```

Create a PDP of the scores predicted by `Mdl` for all three classes of `species` against the third predictor variable `x3`. Specify the class labels by using the `ClassNames` property of `Mdl`.

```
plotPartialDependence(Mdl,3,Mdl.ClassNames);
```



According to this model, the probability of `virginica` increases with `x3`. The probability of `setosa` is about 0.33, from where `x3` is 0 to around 2.5, and then the probability drops to almost 0.

Create Individual Conditional Expectation Plots

Train a Gaussian process regression model using generated sample data where a response variable includes interactions between predictor variables. Then, create ICE plots that show the relationship between a feature and the predicted responses for each observation.

Generate sample predictor data `x1` and `x2`.

```
rng('default') % For reproducibility
n = 200;
x1 = rand(n,1)*2-1;
x2 = rand(n,1)*2-1;
```

Generate response values that include interactions between `x1` and `x2`.

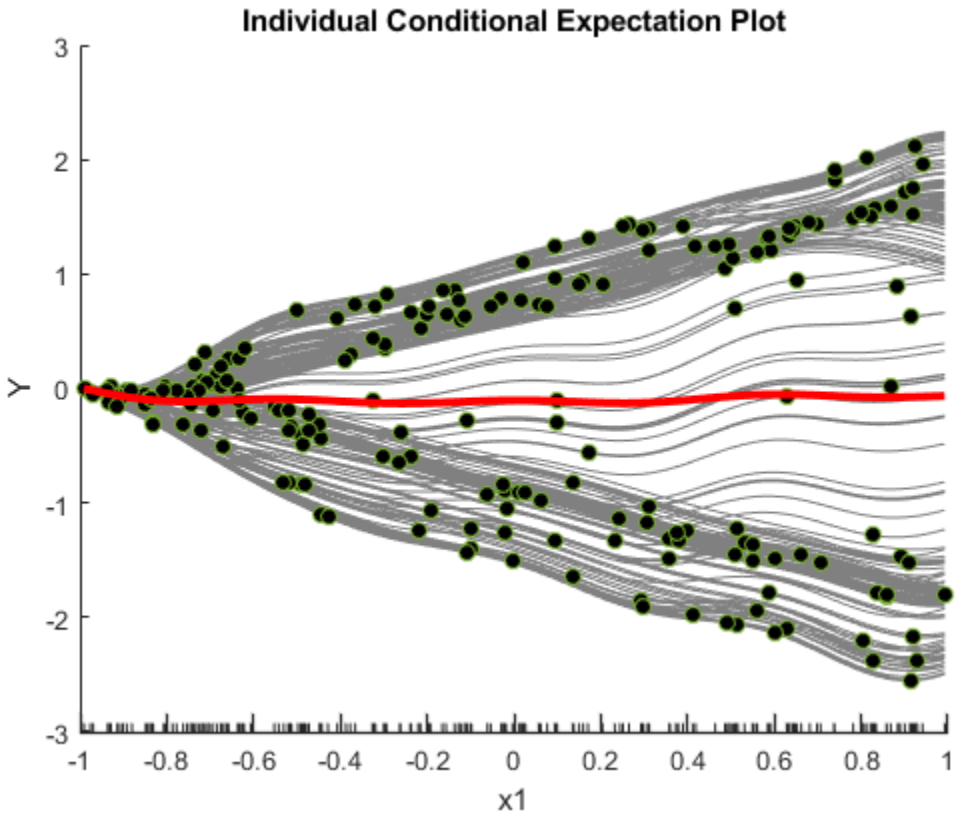
```
Y = x1-2*x1.*(x2>0)+0.1*rand(n,1);
```

Create a Gaussian process regression model using `[x1 x2]` and `Y`.

```
Mdl = fitrgp([x1 x2],Y);
```

Create a figure including a PDP (red line) for the first predictor `x1`, a scatter plot (circle markers) of `x1` and predicted responses, and a set of ICE plots (gray lines) by specifying `'Conditional'` as `'centered'`.

```
plotPartialDependence(Mdl,1,'Conditional','centered')
```



When `'Conditional'` is `'centered'`, `plotPartialDependence` offsets plots so that all plots start from zero, which is helpful in examining the cumulative effect of the selected feature.

A PDP finds averaged relationships, so it does not reveal hidden dependencies especially when responses include interactions between features. However, the ICE plots clearly show two different dependencies of responses on `x1`.

Use New Predictor Data for Partial Dependence Plot

Train an ensemble of classification models and create two PDPs, one using the training data set and the other using a new data set.

Load the `census1994` data set, which contains US yearly salary data, categorized as `<=50K` or `>50K`, and several demographic variables.

```
load census1994
```

Extract a subset of variables to analyze from the tables `adultdata` and `adulttest`.

```
X = adultdata(:,{'age','workClass','education_num','marital_status','race', ...
   'sex','capital_gain','capital_loss','hours_per_week','salary'});
Xnew = adulttest(:,{'age','workClass','education_num','marital_status','race', ...
   'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

Train an ensemble of classifiers with `salary` as the response and the remaining variables as predictors by using the function `fitcensemble`. For binary classification, `fitcensemble` aggregates 100 classification trees using the `LogitBoost` method.

```
Mdl = fitcensemble(X,'salary');
```

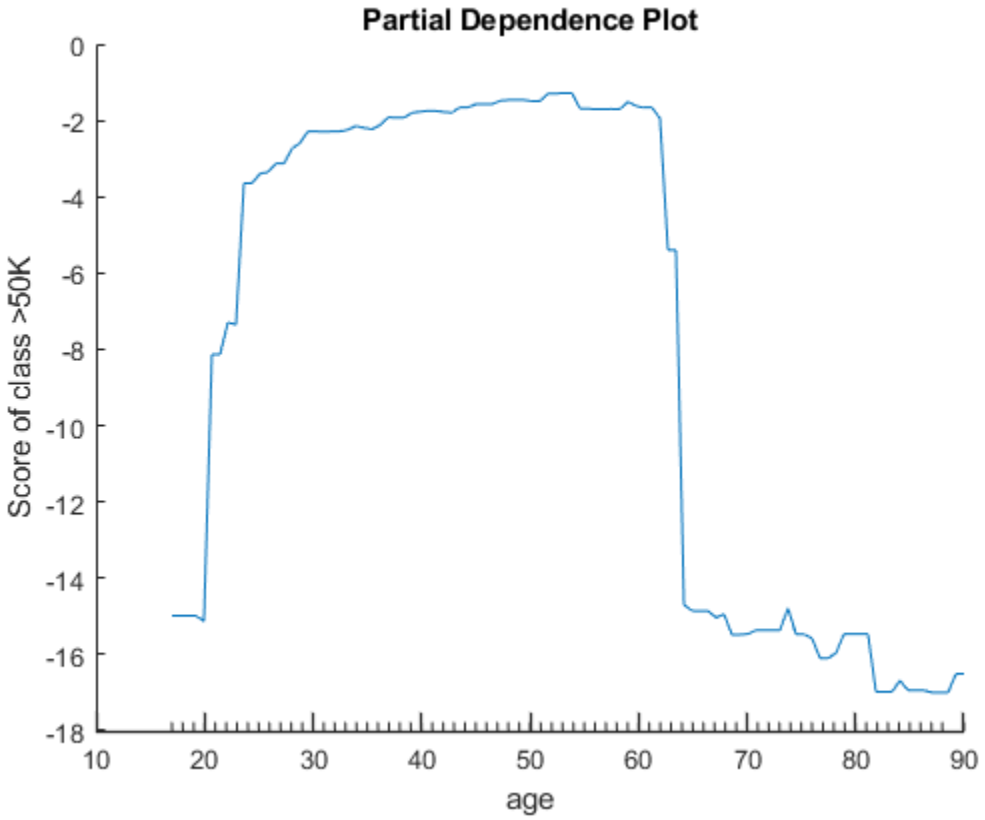
Inspect the class names in `Mdl`.

```
Mdl.ClassNames
```

```
ans = 2x1 categorical
     <=50K
     >50K
```

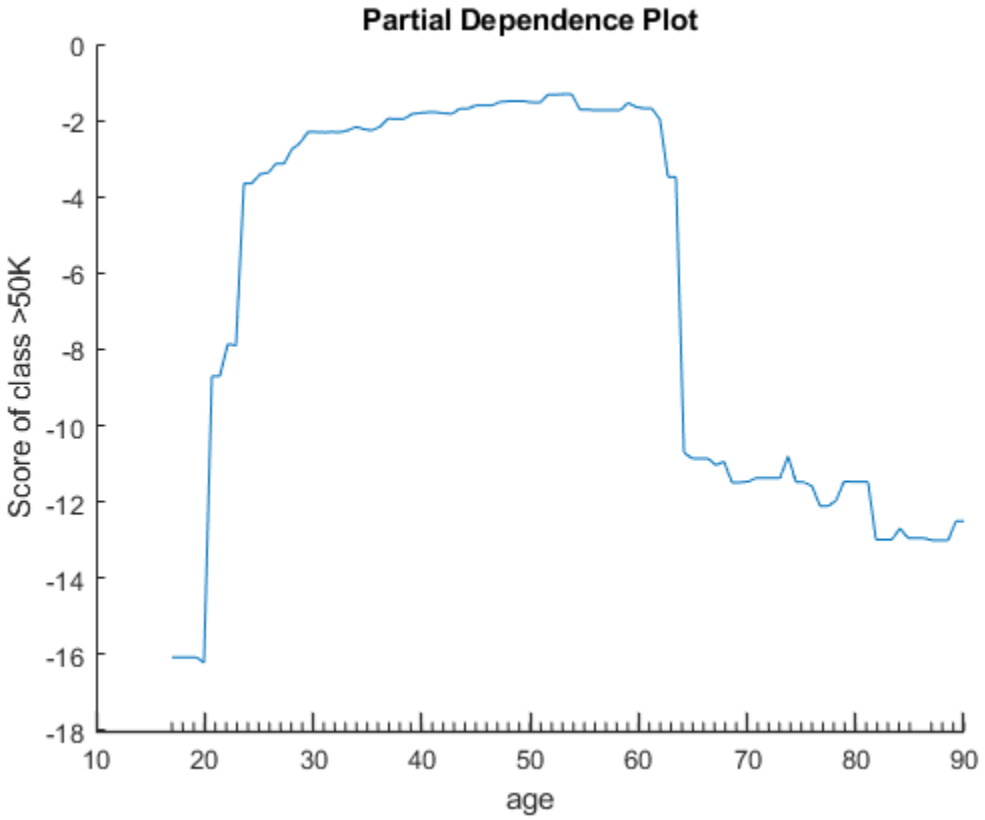
Create a partial dependence plot of the scores predicted by `Mdl` for the second class of `salary` (`>50K`) against the predictor `age` using the training data.

```
plotPartialDependence(Mdl,'age',Mdl.ClassNames(2))
```



Create a PDP of the scores for class `>50K` against `age` using new predictor data from the table `Xnew`.

```
plotPartialDependence(Mdl,'age',Mdl.ClassNames(2),Xnew)
```



The two plots show similar shapes for the partial dependence of the predicted score of high `salary` (`>50K`) on `age`. Both plots indicate that the predicted score of high salary rises fast until the age of 30, then stays almost flat until the age of 60, and then drops fast. However, the plot based on the new data produces slightly higher scores for ages over 65.

Compare Importance of Predictor Variables

Train a regression ensemble using the `carsmall` data set, and create a PDP plot and ICE plots for each predictor variable using a new data set, `carbig`. Then, compare the figures to analyze the importance of predictor variables. Also, compare the results with the estimates of predictor importance returned by the `predictorImportance` function.

Load the `carsmall` data set.

```
load carsmall
```

Specify `Weight`, `Cylinders`, `Horsepower`, and `Model_Year` as the predictor variables (`X`), and `MPG` as the response variable (`Y`).

```
X = [Weight,Cylinders,Horsepower,Model_Year];
Y = MPG;
```

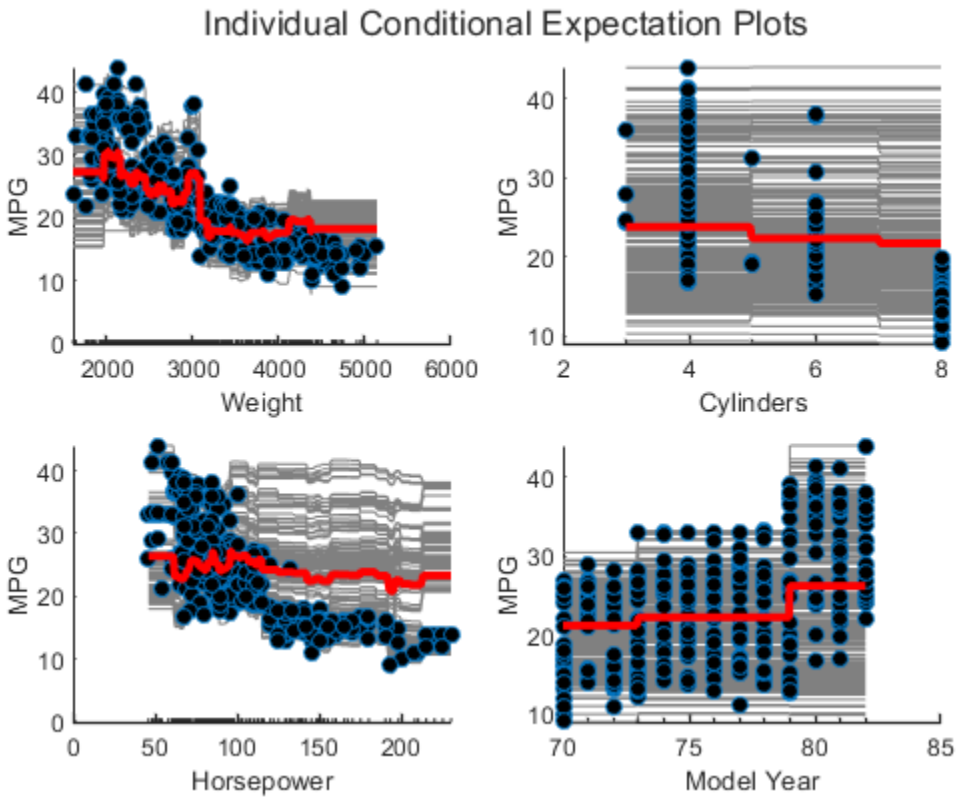
Train a regression ensemble using `X` and `Y`.

```
Mdl = fitrensemble(X,Y, ...
    'PredictorNames',{'Weight','Cylinders','Horsepower','Model Year'}, ...
    'ResponseName','MPG');
```

Create the importance of predictor variables by using the `plotPartialDependence` and `predictorImportance` functions. The `plotPartialDependence` function visualizes the relationships between a selected predictor and predicted responses. `predictorImportance` summarizes the importance of a predictor with a single value.

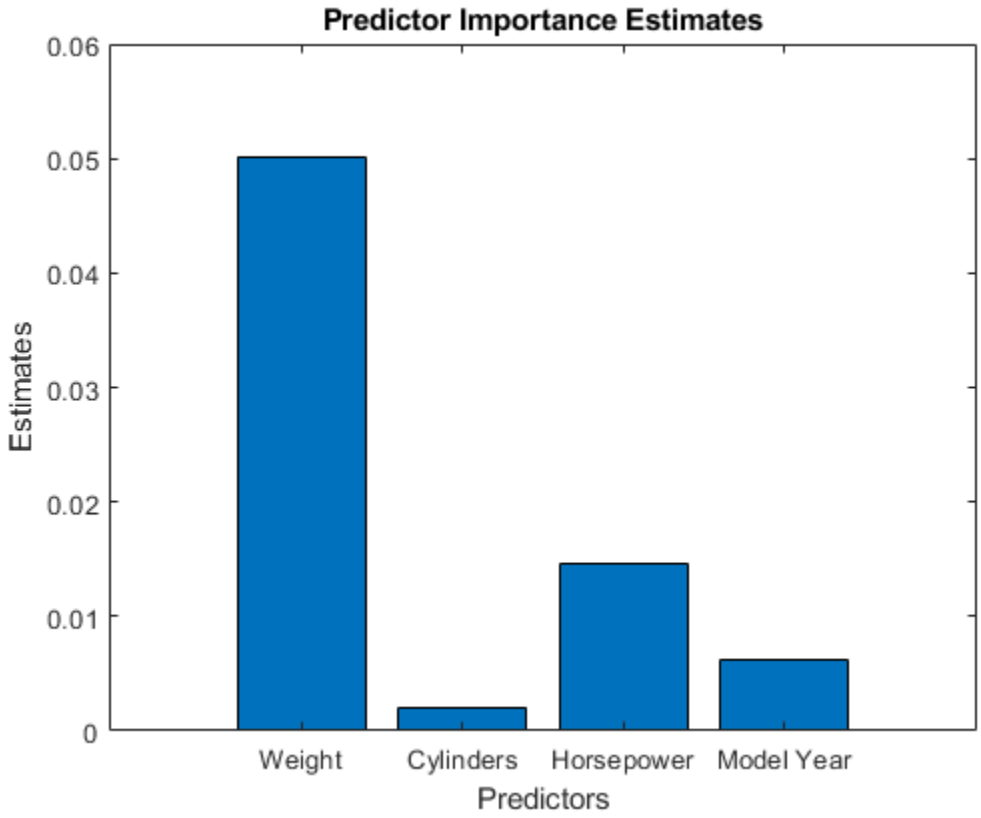
Create a figure including a PDP plot (red line) and ICE plots (gray lines) for each predictor by using `plotPartialDependence` and specifying `'Conditional','absolute'`. Each figure also includes a scatter plot (circle markers) of the selected predictor and predicted responses. Also, load the `carbig` data set and use it as new predictor data, `Xnew`. When you provide `Xnew`, the `plotPartialDependence` function uses `Xnew` instead of the predictor data in `Mdl`.

```
load carbig
Xnew = [Weight,Cylinders,Horsepower,Model_Year];

figure
t = tiledlayout(2,2,'TileSpacing','compact');
title(t,'Individual Conditional Expectation Plots')

for i = 1 : 4
    nexttile
    plotPartialDependence(Mdl,i,Xnew,'Conditional','absolute')
    title('')
end
```



Compute estimates of predictor importance by using `predictorImportance`. This function sums changes in the mean squared error (MSE) due to splits on every predictor, and then divides the sum by the number of branch nodes.

```
imp = predictorImportance(Mdl);
figure
bar(imp)
title('Predictor Importance Estimates')
ylabel('Estimates')
xlabel('Predictors')
ax = gca;
ax.XTickLabel = Mdl.PredictorNames;
```



The variable `Weight` has the most impact on `MPG` according to predictor importance. The PDP of `Weight` also shows that `MPG` has high partial dependence on `Weight`. The variable `Cylinders` has the least impact on `MPG` according to predictor importance. The PDP of `Cylinders` also shows that `MPG` does not change much depending on `Cylinders`.

Compare Partial Dependence of Generalized Additive Model

Train a generalized additive model (GAM) with both linear and interaction terms for predictors. Then, create a PDP with both linear and interaction terms and a PDP with only linear terms. Specify whether to include interaction terms when creating the PDPs.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X,Y,'ClassNames',{'b','g'},'Interactions',10);
```

`Mdl` is a `ClassificationGAM` model object.

List the interaction terms in `Mdl`.

```
Mdl.Interactions
```

```
ans = 10×2

     1     5
     7     8
     6     7
     5     6
     5     7
     5     8
     3     5
     4     7
     1     7
     4     5
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term.

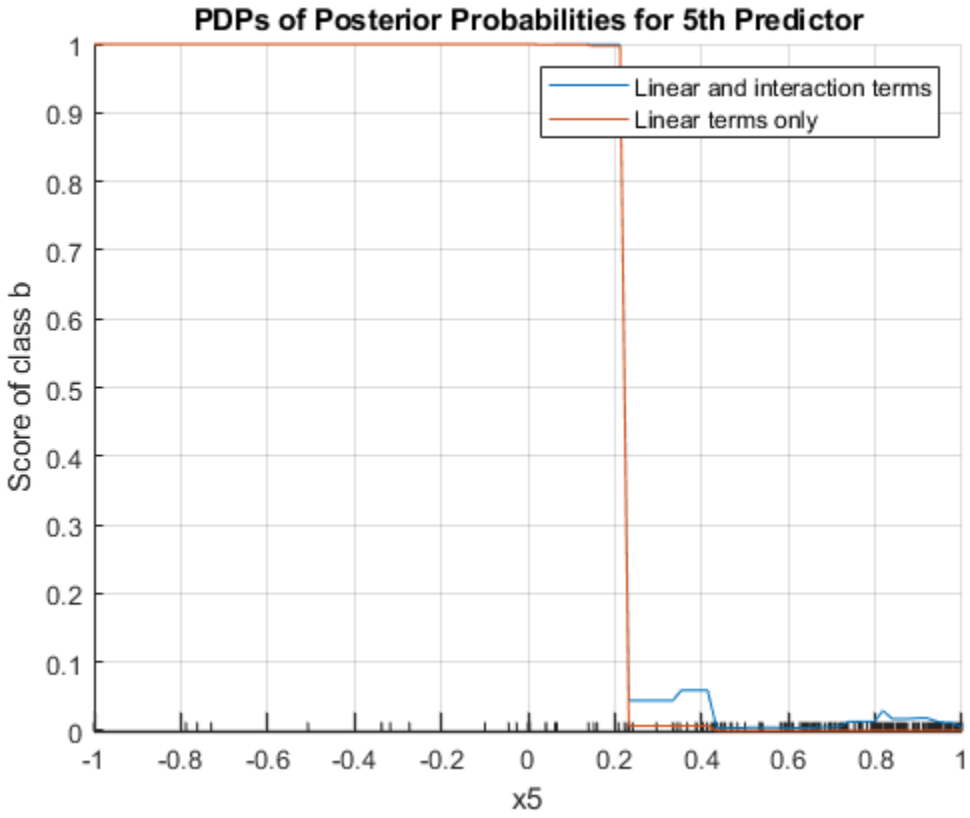
Find the most frequent predictor in the interaction terms.

```
mode(Mdl.Interactions,'all')
```

```
ans = 5
```

The most frequent predictor in the interaction terms is the 5th predictor (`x5`). Create PDPs for the 5th predictor. To exclude interaction terms from the computation, specify `'IncludeInteractions',false` for the second PDP.

```
plotPartialDependence(Mdl,5,Mdl.ClassNames(1))
hold on
plotPartialDependence(Mdl,5,Mdl.ClassNames(1),'IncludeInteractions',false)
grid on
legend('Linear and interaction terms','Linear terms only')
title('PDPs of Posterior Probabilities for 5th Predictor')
hold off
```



The plot shows that the partial dependence of the scores (posterior probabilities) on `x5` varies depending on whether the model includes the interaction terms, especially where `x5` is between 0.2 and 0.45.

Extract Partial Dependence Estimates from Plots

Train a support vector machine (SVM) regression model using the `carsmall` data set, and create a PDP for two predictor variables. Then, extract partial dependence estimates from the output of `plotPartialDependence`. Alternatively, you can get the partial dependence values by using the `partialDependence` function.

Load the `carsmall` data set.

```
load carsmall
```

Specify `Weight`, `Cylinders`, `Displacement`, and `Horsepower` as the predictor variables (`Tbl`).

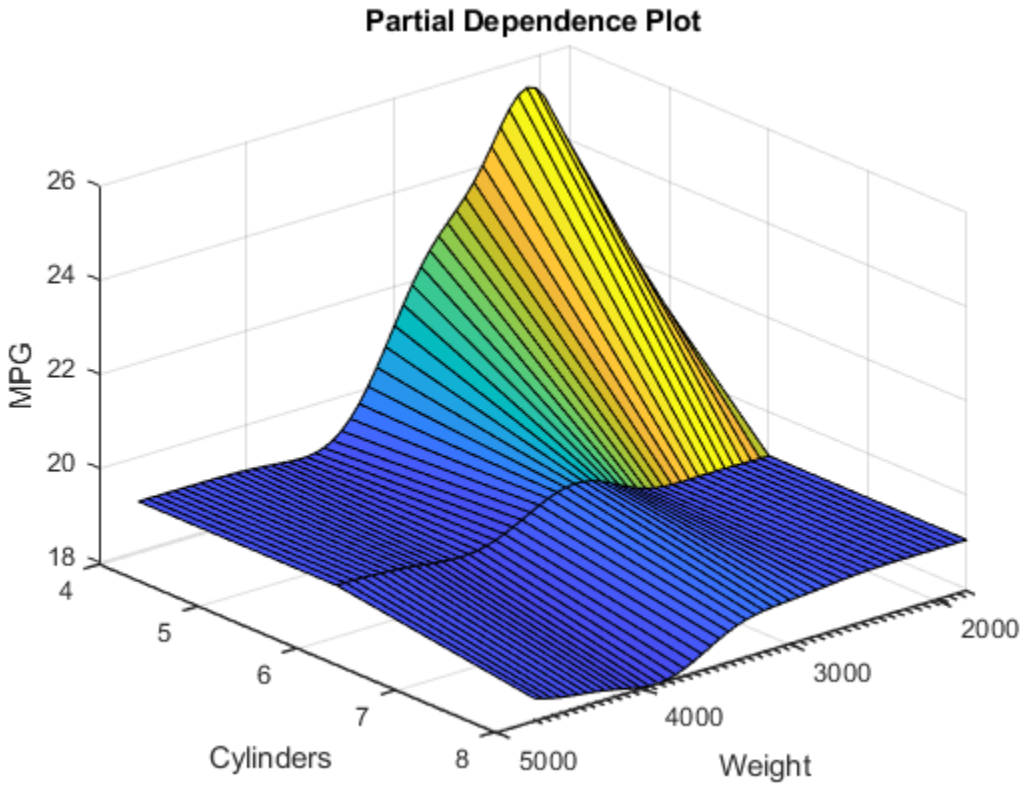
```
Tbl = table(Weight,Cylinders,Displacement,Horsepower);
```

Construct an SVM regression model using `Tbl` and the response variable `MPG`. Use a Gaussian kernel function with an automatic kernel scale.

```
Mdl = fitrsvm(Tbl,MPG,'ResponseName','MPG', ...
    'CategoricalPredictors','Cylinders','Standardize',true, ...
    'KernelFunction','gaussian','KernelScale','auto');
```

Create a PDP that visualizes partial dependence of predicted responses (`MPG`) on the predictor variables `Weight` and `Cylinders`. Specify query points to compute the partial dependence for `Weight` by using the `'QueryPoints'` name-value pair argument. You cannot specify the `'QueryPoints'` value for `Cylinders` because it is a categorical variable. `plotPartialDependence` uses all categorical values.

```
pt = linspace(min(Weight),max(Weight),50)';
ax = plotPartialDependence(Mdl,{'Weight','Cylinders'},'QueryPoints',{pt,[]});
view(140,30) % Modify the viewing angle
```



The PDP shows an interaction effect between `Weight` and `Cylinders`. The partial dependence of `MPG` on `Weight` changes depending on the value of `Cylinders`.

Extract the estimated partial dependence of `MPG` on `Weight` and `Cylinders`. The `XData`, `YData`, and `ZData` values of `ax.Children` are x-axis values (the first selected predictor values), y-axis values (the second selected predictor values), and z-axis values (the corresponding partial dependence values), respectively.

```
xval = ax.Children.XData;
yval = ax.Children.YData;
zval = ax.Children.ZData;
```

Alternatively, you can get the partial dependence values by using the `partialDependence` function.

```
[pd,x,y] = partialDependence(Mdl,{'Weight','Cylinders'},'QueryPoints',{pt,[]});
```

`pd` contains the partial dependence values for the query points `x` and `y`.

If you specify `'Conditional'` as `'absolute'`, `plotPartialDependence` creates a figure including a PDP, a scatter plot, and a set of ICE plots. `ax.Children(1)` and `ax.Children(2)` correspond to the PDP and scatter plot, respectively. The remaining elements of `ax.Children` correspond to the ICE plots. The `XData` and `YData` values of `ax.Children(i)` are x-axis values (the selected predictor values) and y-axis values (the corresponding partial dependence values), respectively.

Input Arguments

`RegressionMdl` — Regression model

regression model object

Regression model, specified as a full or compact regression model object, as given in the following tables of supported models.

Model	Full or Compact Model Object
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`
Generalized linear mixed-effect model	`GeneralizedLinearMixedModel`
Linear regression	`LinearModel`, `CompactLinearModel`
Linear mixed-effect model	`LinearMixedModel`
Nonlinear regression	`NonLinearModel`
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`
Generalized additive model (GAM)	`RegressionGAM`, `CompactRegressionGAM`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`
Linear regression for high-dimensional data	`RegressionLinear`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`
Support vector machine (SVM) regression	`RegressionSVM`, `CompactRegressionSVM`
Regression tree	`RegressionTree`, `CompactRegressionTree`
Bootstrap aggregation for ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`

If `RegressionMdl` is a model object that does not contain predictor data (for example, a compact model), you must provide the input argument `Data`.

`plotPartialDependence` does not support a model object trained with a sparse matrix. When you train a model, use a full numeric matrix or table for predictor data where rows correspond to individual observations.

`ClassificationMdl` — Classification model

classification model object

Classification model, specified as a full or compact classification model object, as given in the following tables of supported models.

Model	Full or Compact Model Object
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`
Generalized additive model (GAM)	`ClassificationGAM`, `CompactClassificationGAM`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`
k-nearest neighbor classifier	`ClassificationKNN`
Linear classification model	`ClassificationLinear`
Multiclass naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`
Support vector machine (SVM) classifier for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`
Bagged ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`

If `ClassificationMdl` is a model object that does not contain predictor data (for example, a compact model), you must provide the input argument `Data`.

`plotPartialDependence` does not support a model object trained with a sparse matrix. When you train a model, use a full numeric matrix or table for predictor data where rows correspond to individual observations.

`Vars` — Predictor variables

vector of positive integers | character vector | string scalar | string array | cell array of character vectors

Predictor variables, specified as a vector of positive integers, character vector, string scalar, string array, or cell array of character vectors. You can specify one or two predictor variables, as shown in the following tables.

One Predictor Variable

Value	Description
positive integer	Index value corresponding to the column of the predictor data.
character vector or string scalar	Name of a predictor variable. The name must match the entry in `RegressionMdl.PredictorNames` or `ClassificationMdl.PredictorNames`.

Two Predictor Variables

Value	Description
vector of two positive integers	Index values corresponding to the columns of the predictor data.
string array or cell array of character vectors	Names of predictor variables. Each element in the array is the name of a predictor variable. The names must match the entries in `RegressionMdl.PredictorNames` or `ClassificationMdl.PredictorNames`.

Example: `{'x1','x3'}`

Data Types: `single` | `double` | `char` | `string` | `cell`

`Labels` — Class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. The values and data types in `Labels` must match those of the class names in the `ClassNames` property of `ClassificationMdl` (`ClassificationMdl.ClassNames`).

- You can specify multiple class labels only when you specify one variable in `Vars` and specify `'Conditional'` as `'none'` (default).
- Use `partialDependence` if you want to compute the partial dependence for multiple variables and multiple class labels in one function call.

This argument is valid only when you specify a classification model object `ClassificationMdl`.

Example: `{'red','blue'}`

Example: `ClassificationMdl.ClassNames([1 3])` specifies `Labels` as the first and third classes in `ClassificationMdl`.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`Data` — Predictor data

numeric matrix | table

Predictor data, specified as a numeric matrix or table. Each row of `Data` corresponds to one observation, and each column corresponds to one variable.

`Data` must be consistent with the predictor data that trained the model (`RegressionMdl` or `ClassificationMdl`), stored in either the `X` or `Variables` property.

- If you trained the model using a numeric matrix, then `Data` must be a numeric matrix. The variables making up the columns of `Data` must have the same number and order as the predictor variables that trained the model.

- If you trained the model using a table (for example, `Tbl`), then `Data` must be a table. All predictor variables in `Data` must have the same variable names and data types as the names and types in `Tbl`. However, the column order of `Data` does not need to correspond to the column order of `Tbl`.
- `plotPartialDependence` does not support a sparse matrix.

If `RegressionMdl` or `ClassificationMdl` is a model object that does not contain predictor data, you must provide `Data`. If the model is a full model object that contains predictor data and you specify this argument, then `plotPartialDependence` does not use the predictor data in the model and uses `Data` only.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `plotPartialDependence(Mdl,Vars,Data,'NumObservationsToSample',100,'UseParallel',true)` creates a PDP by using 100 sampled observations in `Data` and executing `for`-loop iterations in parallel.

`Conditional` — Plot type

`'none'` (default) | `'absolute'` | `'centered'`

Plot type, specified as `'none'`, `'absolute'`, or `'centered'`.

Value	Description
`'none'`	`plotPartialDependence` creates a PDP. The plot type depends on the number of predictor variables specified in `Vars` and the number of class labels specified in `Labels` (for a classification model). • One predictor variable and one class label — `plotPartialDependence` computes partial dependence at the query points, and creates a 2-D line plot of the partial dependence. • One predictor variable and multiple class labels — `plotPartialDependence` creates one figure containing multiple 2-D line plots for the selected classes. • Two predictor variables and one class label — `plotPartialDependence` creates a surface plot of partial dependence against the two variables.
`'absolute'`	`plotPartialDependence` creates a figure including the following three types of plots: • PDP with a red line • Scatter plot of the selected predictor variable and predicted responses or scores with circle markers • ICE plot for each observation with a gray line This value is valid when you select only one predictor variable in `Vars` and one class label in `Labels` (for a classification model).

Value	Description
`'centered'`	`plotPartialDependence` creates a figure including the same three types of plots as `'absolute'`. The function offsets plots so that all plots start from zero. This value is valid when you select only one predictor variable in `Vars` and one class label in `Labels` (for a classification model).

Example: `'Conditional','absolute'`

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the generalized additive model (GAM) in the partial dependence computation, specified as `true` or `false`. This argument is valid only for a GAM. That is, you can specify this argument only when `RegressionMdl` is `RegressionGAM` or `CompactRegressionGAM`, or `ClassificationMdl` is `ClassificationGAM` or `CompactClassificationGAM`.

The default `'IncludeInteractions'` value is `true` if the model contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`IncludeIntercept` — Flag to include intercept term
`true` (default) | `false`

Flag to include an intercept term of the generalized additive model (GAM) in the partial dependence computation, specified as `true` or `false`. This argument is valid only for a GAM. That is, you can specify this argument only when `RegressionMdl` is `RegressionGAM` or `CompactRegressionGAM`, or `ClassificationMdl` is `ClassificationGAM` or `CompactClassificationGAM`.

Example: `'IncludeIntercept',false`

Data Types: `logical`

`NumObservationsToSample` — Number of observations to sample
number of total observations (default) | positive integer

Number of observations to sample, specified as a positive integer. The default value is the number of total observations in `Data` or the model (`RegressionMdl` or `ClassificationMdl`). If you specify a value larger than the number of total observations, then `plotPartialDependence` uses all observations.

`plotPartialDependence` samples observations without replacement by using the `datasample` function and uses the sampled observations to compute partial dependence.

`plotPartialDependence` displays minor tick marks at the unique values of the sampled observations.

If you specify `'Conditional'` as either `'absolute'` or `'centered'`, `plotPartialDependence` creates a figure including an ICE plot for each sampled observation.

Example: `'NumObservationsToSample',100`

Data Types: `single` | `double`

`Parent` — Axes in which to plot
`gca` (default) | axes object

Axes in which to plot, specified as an axes object. If you do not specify the axes and if the current axes are Cartesian, then `plotPartialDependence` uses the current axes (`gca`). If axes do not exist, `plotPartialDependence` plots in a new figure.

Example: `'Parent',ax`

`QueryPoints` — Points to compute partial dependence
numeric column vector | numeric two-column matrix | cell array of two numeric column vectors

Points to compute partial dependence for numeric predictors, specified as a numeric column vector, a numeric two-column matrix, or a cell array of two numeric column vectors.

- If you select one predictor variable in `Vars`, use a numeric column vector.
- If you select two predictor variables in `Vars`:
  - Use a numeric two-column matrix to specify the same number of points for each predictor variable.
  - Use a cell array of two numeric column vectors to specify a different number of points for each predictor variable.

The default value is a numeric column vector or a numeric two-column matrix, depending on the number of selected predictor variables. Each column contains 100 evenly spaced points between the minimum and maximum values of the sampled observations for the corresponding predictor variable.

If `'Conditional'` is `'absolute'` or `'centered'`, then the software adds the predictor data values (`Data` or predictor data in `RegressionMdl` or `ClassificationMdl`) of the selected predictors to the query points.

You cannot modify `'QueryPoints'` for a categorical variable. The `plotPartialDependence` function uses all categorical values in the selected variable.

If you select one numeric variable and one categorical variable, you can specify `'QueryPoints'` for a numeric variable by using a cell array consisting of a numeric column vector and an empty array.

Example: `'QueryPoints',{pt,[]}`

Data Types: `single` | `double` | `cell`

`UseParallel` — Flag to run in parallel
`false` (default) | `true`

Flag to run in parallel, specified as `true` or `false`. If you specify `'UseParallel',true`, the `plotPartialDependence` function executes `for`-loop iterations in parallel by using `parfor` when predicting responses or scores for each observation and averaging them. This option requires Parallel Computing Toolbox.

Example: `'UseParallel',true`

Data Types: `logical`

Output Arguments

ax — Axes of the plot

axes object

Axes of the plot, returned as an axes object. For details on how to modify the appearance of the axes and extract data from plots, see “Axes Appearance” and “Extract Partial Dependence Estimates from Plots” on page 33-4651.

More About

Partial Dependence for Regression Models

Partial dependence[1] represents the relationships between predictor variables and predicted responses in a trained regression model. `plotPartialDependence` computes the partial dependence of predicted responses on a subset of predictor variables by marginalizing over the other variables.

Consider partial dependence on a subset X^S of the whole predictor variable set $X = \{x_1, x_2, \ldots, x_m\}$. A subset X^S includes either one variable or two variables: $X^S = \{x_{S1}\}$ or $X^S = \{x_{S1}, x_{S2}\}$. Let X^C be the complementary set of X^S in X. A predicted response $f(X)$ depends on all variables in X:

$f(X) = f(X^S, X^C)$.

The partial dependence of predicted responses on X^S is defined by the expectation of predicted responses with respect to X^C:

$$f^S\left(X^S\right) = E_C\left[f\left(X^S, X^C\right)\right] = \int f\left(X^S, X^C\right) p_C\left(X^C\right) dX^C,$$

where $p_C(X^C)$ is the marginal probability of X^C, that is, $p_C\left(X^C\right) \approx \int p\left(X^S, X^C\right) dX^S$. Assuming that each observation is equally likely, and the dependence between X^S and X^C and the interactions of X^S and X^C in responses is not strong, `plotPartialDependence` estimates the partial dependence by using observed predictor data as follows:

$$f^S\left(X^S\right) \approx \frac{1}{N} \sum_{i=1}^{N} f\left(X^S, X_i{}^C\right), \tag{33-2}$$

where N is the number of observations and $X_i = (X_i{}^S, X_i{}^C)$ is the ith observation.

When you call the `plotPartialDependence` function, you can specify a trained model ($f(\cdot)$) and select variables (X^S) by using the input arguments `RegressionMdl` and `Vars`, respectively. `plotPartialDependence` computes the partial dependence at 100 evenly spaced points of X^S or the points that you specify by using the `'QueryPoints'` name-value pair argument. You can specify the number (N) of observations to sample from given predictor data by using the `'NumObservationsToSample'` name-value pair argument.

Individual Conditional Expectation for Regression Models

An individual conditional expectation (ICE) [2], as an extension of partial dependence, represents the relationship between a predictor variable and the predicted responses for each observation. While partial dependence shows the averaged relationship between predictor variables and predicted responses, a set of ICE plots disaggregates the averaged information and shows an individual dependence for each observation.

`plotPartialDependence` creates an ICE plot for each observation. A set of ICE plots is useful to investigate heterogeneities of partial dependence originating from different observations. `plotPartialDependence` can also create ICE plots with any predictor data provided through the input argument `Data`. You can use this feature to explore predicted response space.

Consider an ICE plot for a selected predictor variable x_S with a given observation X_i^C, where $X^S = \{x_S\}$, X^C is the complementary set of X^S in the whole variable set X, and $X_i = (X_i^S, X_i^C)$ is the ith observation. The ICE plot corresponds to the summand of the summation in "Equation 33-2":

$$f^S{}_i\left(X^S\right) = f\left(X^S, X_i{}^C\right).$$

`plotPartialDependence` plots $f^S{}_i\left(X^S\right)$ for each observation i when you specify `'Conditional'` as `'absolute'`. If you specify `'Conditional'` as `'centered'`, `plotPartialDependence` draws all plots after removing level effects due to different observations:

$$f^S{}_{i,\,\text{centered}}\left(X^S\right) = f\left(X^S, X_i{}^C\right) - f\left(\min\left(X^S\right), X_i{}^C\right).$$

This subtraction ensures that each plot starts from zero, so that you can examine the cumulative effect of X^S and the interactions between X^S and X^C.

Partial Dependence and ICE for Classification Models

In the case of classification models, `plotPartialDependence` computes the partial dependence and individual conditional expectation in the same way as for regression models, with one exception: instead of using the predicted responses from the model, the function uses the predicted scores for the classes specified in `Labels`.

Weighted Traversal Algorithm

The weighted traversal algorithm[1] is a method to estimate partial dependence for a tree-based model. The estimated partial dependence is the weighted average of response or score values corresponding to the leaf nodes visited during the tree traversal.

Let X^S be a subset of the whole variable set X and X^C be the complementary set of X^S in X. For each X^S value to compute partial dependence, the algorithm traverses a tree from the root (beginning) node down to leaf (terminal) nodes and finds the weights of leaf nodes. The traversal starts by assigning a weight value of one at the root node. If a node splits by X^S, the algorithm traverses to the appropriate child node depending on the X^S value. The weight of the child node becomes the same value as its parent node. If a node splits by X^C, the algorithm traverses to both child nodes. The weight of each child node becomes a value of its parent node multiplied by the fraction of observations corresponding to each child node. After completing the tree traversal, the algorithm computes the weighted average by using the assigned weights.

For an ensemble of bagged trees, the estimated partial dependence is an average of the weighted averages over the individual trees.

Algorithms

`plotPartialDependence` uses a `predict` function to predict responses or scores. `plotPartialDependence` chooses the proper `predict` function according to the model (`RegressionMdl` or `ClassificationMdl`) and runs `predict` with its default settings. For details about each `predict` function, see the `predict` functions in the following two tables. If the specified

model is a tree-based model (not including a boosted ensemble of trees) and `'Conditional'` is `'none'`, then `plotPartialDependence` uses the weighted traversal algorithm instead of the `predict` function. For details, see “Weighted Traversal Algorithm” on page 33-4660.

Regression Model Object

Model Type	Full or Compact Regression Model Object	Function to Predict Responses
Bootstrap aggregation for ensemble of decision trees	`CompactTreeBagger`	`predict`
Bootstrap aggregation for ensemble of decision trees	`TreeBagger`	`predict`
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`	`predict`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`	`predict`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`	`predict`
Generalized additive model	`RegressionGAM`, `CompactRegressionGAM`	`predict`
Generalized linear mixed-effect model	`GeneralizedLinearMixedModel`	`predict`
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`	`predict`
Linear mixed-effect model	`LinearMixedModel`	`predict`
Linear regression	`LinearModel`, `CompactLinearModel`	`predict`
Linear regression for high-dimensional data	`RegressionLinear`	`predict`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`	`predict`
Nonlinear regression	`NonLinearModel`	`predict`
Regression tree	`RegressionTree`, `CompactRegressionTree`	`predict`
Support vector machine	`RegressionSVM`, `CompactRegressionSVM`	`predict`

Classification Model Object

Model Type	Full or Compact Classification Model Object	Function to Predict Labels and Scores
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`	`predict`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`	`predict`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`	`predict`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`	`predict`
Generalized additive model	`ClassificationGAM`, `CompactClassificationGAM`	`predict`
k-nearest neighbor model	`ClassificationKNN`	`predict`
Linear classification model	`ClassificationLinear`	`predict`
Naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`	`predict`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`	`predict`
Support vector machine for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	`predict`
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`	`predict`
Bagged ensemble of decision trees	`TreeBagger`, `CompactTreeBagger`	`predict`

Alternative Functionality

- `partialDependence` computes partial dependence without visualization. The function can compute partial dependence for two variables and multiple classes in one function call.

References

[1] Friedman, Jerome. H. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29, no. 5 (2001): 1189-1232.

[2] Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. “Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation.” *Journal of Computational and Graphical Statistics* 24, no. 1 (January 2, 2015): 44–65.

[3] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. New York*, NY: Springer New York, 2001.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` name-value argument to `true` in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function fully supports GPU arrays for a regression or classification model specified as one of the following objects:
  - `LinearModel` object
  - `CompactLinearModel` object
  - `GeneralizedLinearModel` object
  - `CompactGeneralizedLinearModel` object
  - `ClassificationKNN` object
- The function also supports these objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`lime` | `oobPermutedPredictorImportance` | `partialDependence` | `predictorImportance (RegressionEnsemble)` | `predictorImportance (RegressionTree)` | `relieff` | `sequentialfs` | `shapley`

Topics

“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

Introduced in R2017b

plotprofile

Class: `RepeatedMeasuresModel`

Plot expected marginal means with optional grouping

Syntax

```
plotprofile(rm,X)
plotprofile(rm,Name,Value)
H = plotprofile( ___ )
```

Description

`plotprofile(rm,X)` plots the expected marginal means computed from the repeated measures model `rm` as a function of the variable `X`.

`plotprofile(rm,Name,Value)` plots the expected marginal means computed from the repeated measures model `rm` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the factors to group by or change the line colors.

`H = plotprofile( ___ )` returns handles, `H`, to the plotted lines.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`X` — Name of between-subjects or within-subjects factor

character vector | string scalar

Name of a between-subjects or within-subjects factor, specified as a character vector or string scalar.

For example, if you want to plot the marginal means as a function of the groups of a between-subjects variable drug, you can specify it as follows.

Example: `'Drug'`

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Group` — Name of between-subject factor or factors

character vector | string array | cell array of character vectors

Name of between-subject factor or factors, specified as the comma-separated pair consisting of `'Group'` and a character vector, string array, or cell array of character vectors. This name-value pair argument groups the lines according to the factor values.

For example, if you have two between-subject factors, drug and sex, and you want to group the lines in the plot according to them, you can specify these factors as follows.

Example: `'Group',{'Drug','Sex'}`

Data Types: `char` | `string` | `cell`

`Marker` — Marker to use for each group
string array | cell array of character vectors

Marker to use for each group, specified as the comma-separated pair consisting of `'Marker'` and a string array or cell array of character vectors.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify `o` as the marker for the groups of drug and `x` as the marker for the groups of sex as follows.

Example: `'Marker',{'o','o','x','x'}`

Data Types: `string` | `cell`

`Color` — Color for each group
character vector | string array | cell array of character vectors | rows of a three-column RGB matrix

Color for each group, specified as the comma-separated pair consisting of `'Color'` and a character vector, string array, cell array of character vectors, or rows of a three-column RGB matrix.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify red as the color for the groups of drug and blue as the color for the groups of sex as follows.

Example: `'Color','rrbb'`

Data Types: `single` | `double` | `char` | `string` | `cell`

`LineStyle` — Line style for each group
string array | cell array of character vectors

Line style for each group, specified as the comma-separated pair consisting of `'LineStyle'` and a string array or cell array of character vectors.

For example, if you have two between-subject factors, drug and sex, with each having two groups, you can specify `-` as the line style of one group and `:` as the line style for the other group as follows.

Example: `'LineStyle',{'-' ':' '-' ':'}`

Data Types: `string` | `cell`

Output Arguments

`H` — Handle to plotted lines
handle

Handle to plotted lines, returned as a handle.

Examples

Plot Expected Marginal Means

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

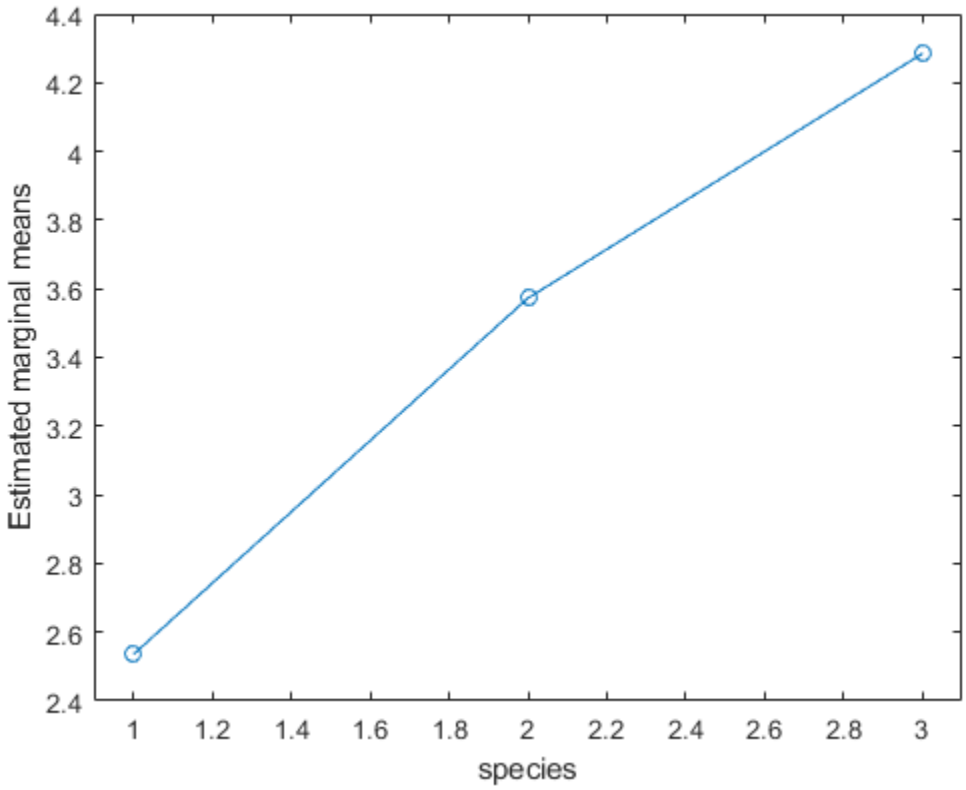
```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform data grouped by the factor species.

```
plotprofile(rm,'species')
```



The estimated marginal means seem to differ with group. You can compute the standard error and the 95% confidence intervals for the marginal means using the `margmean` method.

Plot Marginal Means for Two Groups

Load the sample data.

```
load repeatedmeas
```

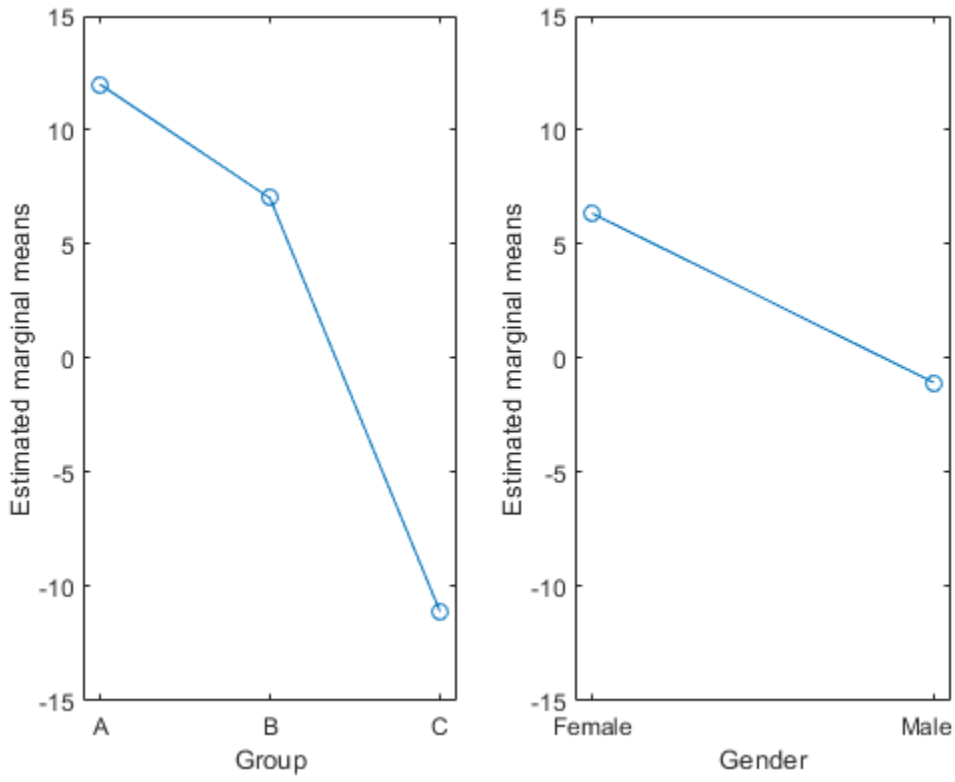
The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table `within` includes the within-subject variables `w1` and `w2`. This is simulated data.

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

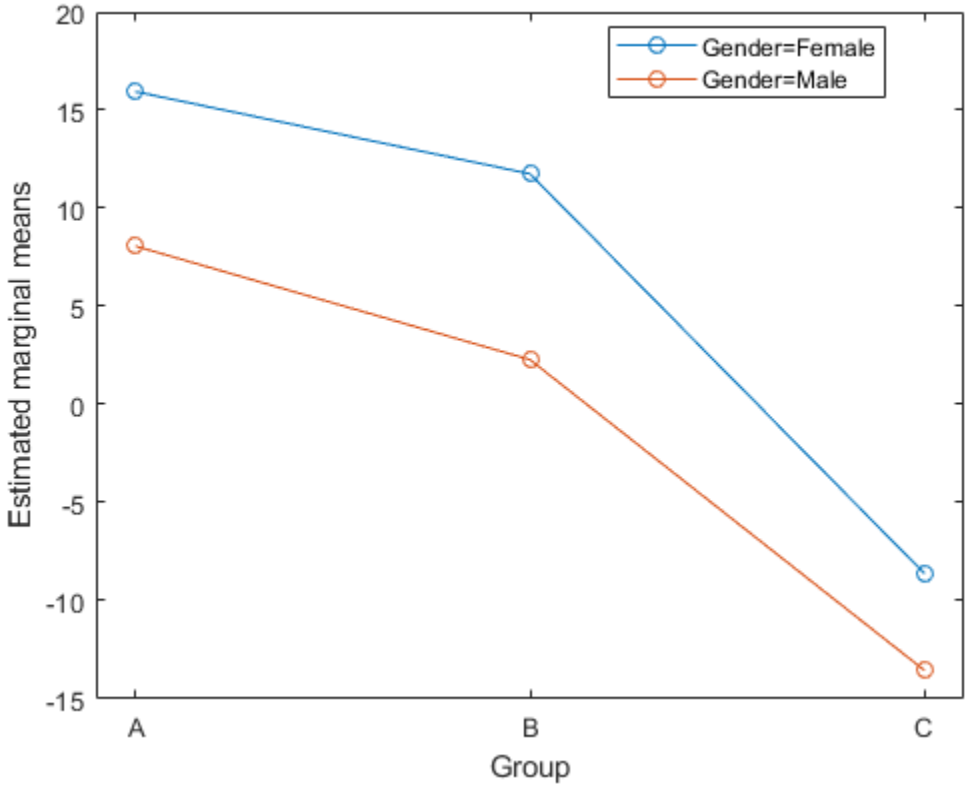
Plot the estimated marginal means based on the factors `Group` and `Gender`.

```
ax1 = subplot(1,2,1);
plotprofile(rm,'Group')
ax2 = subplot(1,2,2);
plotprofile(rm,'Gender')
linkaxes([ax1 ax2],'y')
```



Plot the estimated marginal means based on the factor `Group` and grouped by `Gender`.

```
figure()
plotprofile(rm,'Group','Group','Gender')
```



See Also

`fitrm` | `margmean` | `plot`

plotResiduals

Plot residuals of generalized linear regression model

Syntax

```
plotResiduals(mdl)
plotResiduals(mdl,plottype)
plotResiduals(mdl,plottype,Name,Value)
h = plotResiduals( ___ )
```

Description

`plotResiduals(mdl)` creates a histogram plot of the generalized linear regression model (`mdl`) residuals.

`plotResiduals(mdl,plottype)` specifies the residual plot type `plottype`.

`plotResiduals(mdl,plottype,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the residual type and the graphical properties of residual data points.

`h = plotResiduals( ___ )` returns graphics objects for the lines or patch in the plot using any of the input argument combinations in the previous syntaxes. Use `h` to modify the properties of a specific line or patch after you create the plot. For a list of properties, see Chart Line and Patch Properties.

Examples

Create Residual Plots for Generalized Linear Regression Model

Create three plots of a fitted generalized linear regression model: a histogram of raw residuals, a normal probability plot of raw residuals, a normal probability plot of Anscombe type residuals.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

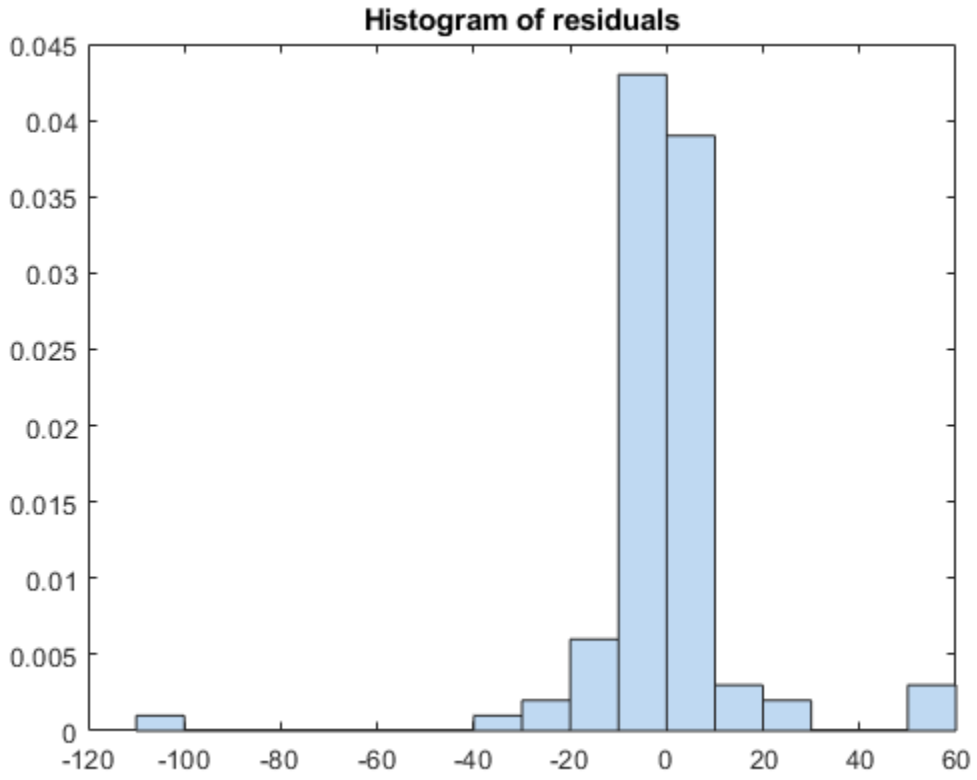
```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

Create a histogram of the raw residuals using probability density function scaling.

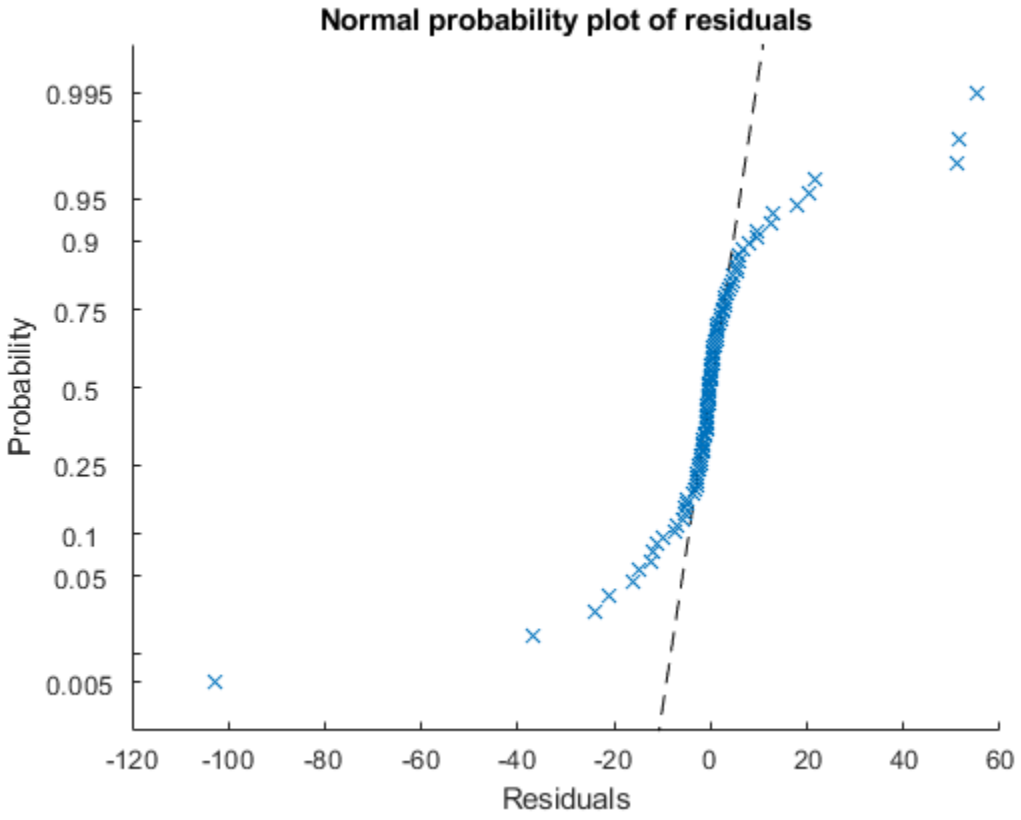
```
plotResiduals(mdl)
```



The area of each bar is the relative number of observations. The sum of the bar areas is equal to 1.

Create a normal probability plot of the raw residuals.

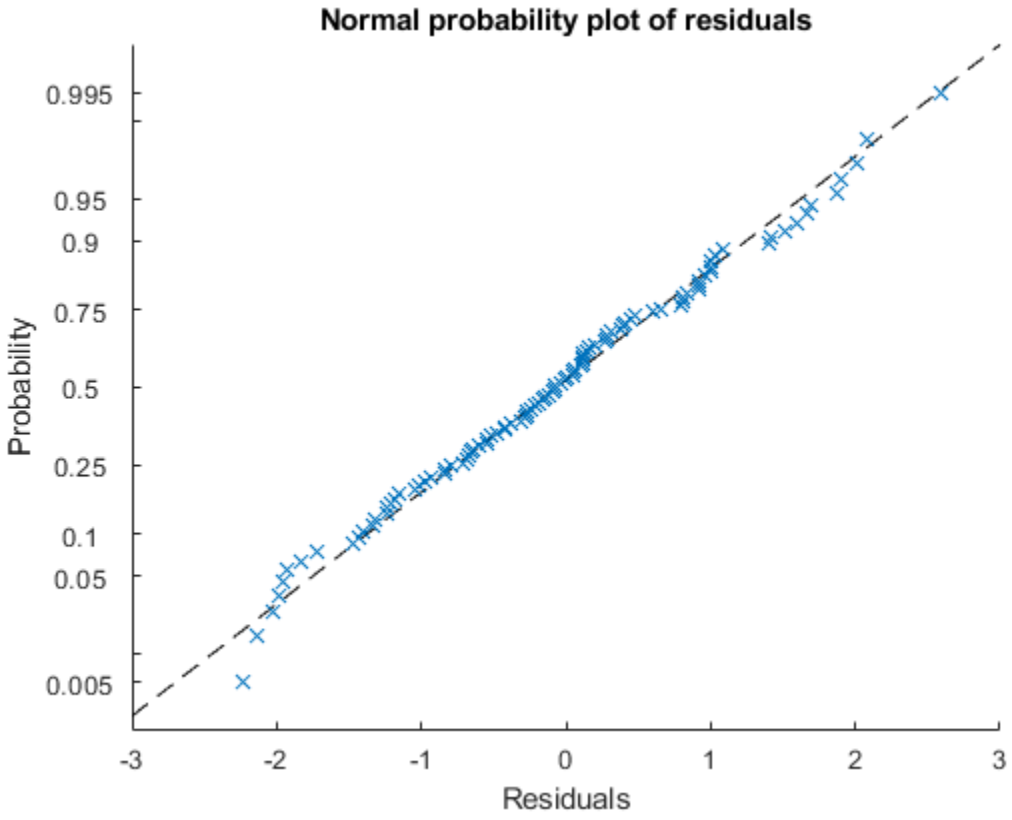
```
plotResiduals(mdl,'probability')
```



The residuals do not match a normal distribution in the tails because the residuals are more spread out.

Create a normal probability plot of the Anscombe type residuals.

```
plotResiduals(mdl,'probability','ResidualType','Anscombe')
```



The Anscombe type residuals match a normal distribution.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

`plottype` — Plot type

`'histogram'` (default) | `'caseorder'` | `'fitted'` | `'lagged'` | `'probability'` | `'symmetry'`

Plot type, specified as one of the values in this table.

Value	Description
`'caseorder'`	Residuals vs. case order (row number)
`'fitted'`	Residuals vs. fitted values
`'histogram'`	Histogram of residuals using probability density function scaling. The area of each bar is the relative number of observations. The sum of the bar areas is equal to 1.
`'lagged'`	Residuals vs. lagged residuals ($r(t)$ vs. $r(t-1)$)

Value	Description
`'probability'`	Normal probability plot of residuals. For details, see `probplot`.
`'symmetry'`	Symmetry plot of residuals around their median (residuals in upper tail – median vs. median – residuals in lower tail). This plot includes a dotted reference line of $y = x$ to examine the symmetry of residuals.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','o'`

Note The graphical properties listed here are only a subset. For a complete list, see Chart Line for lines and Patch Properties for histogram. The specified properties apply to the appearance of residual data points or the appearance of the histogram.

`ResidualType` — Type of residual
`'Raw'` (default) | `'Anscombe'` | `'Deviance'` | `'LinearPredictor'` | `'Pearson'`

Type of residual used in the plot, specified as the comma-separated pair consisting of `'ResidualType'` and one of the values in this table.

Value	Description
`'Raw'`	Observed minus fitted values
`'LinearPredictor'`	Residuals on the linear predictor scale, equal to the adjusted response value minus the fitted linear combination of the predictors
`'Pearson'`	Raw residuals divided by the estimated standard deviation of the response
`'Anscombe'`	Residuals defined on transformed data with the transformation selected to remove skewness
`'Deviance'`	Residuals based on the contribution of each observation to the deviance

The `Residuals` property of `mdl` contains the residual values used by `plotResiduals` to create plots.

Example: `'ResidualType','Pearson'`

`Color` — Line color
RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

Example: `'Color','blue'`

`LineWidth` — Line width
positive value

Line width, specified as the comma-separated pair consisting of `'LineWidth'` and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: `'LineWidth',0.75`

`Marker` — Marker symbol
`'o'` | `'+'` | `'*'` | `'.'` | `'x'` | ...

Marker symbol, specified as the comma-separated pair consisting of `'Marker'` and one of the values in this table.

Value	Description
`'o'`	Circle
`'+'`	Plus sign
`'*'`	Asterisk
`'.'`	Point
`'x'`	Cross
`'_'`	Horizontal line
`'|'`	Vertical line
`'square'` or `'s'`	Square
`'diamond'` or `'d'`	Diamond
`'^'`	Upward-pointing triangle
`'v'`	Downward-pointing triangle
`'>'`	Right-pointing triangle
`'<'`	Left-pointing triangle
`'pentagram'` or `'p'`	Five-pointed star (pentagram)
`'hexagram'` or `'h'`	Six-pointed star (hexagram)
`'none'`	No markers

Example: `'Marker','+'`

`MarkerEdgeColor` — Marker outline color
`'auto'` (default) | `'none'` | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of `'MarkerEdgeColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The default value of `'auto'` uses the same color specified by using `'Color'`.

Example: `'MarkerEdgeColor','blue'`

`MarkerFaceColor` — Marker fill color
`'none'` (default) | `'auto'` | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of `'MarkerFaceColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The `'auto'` value uses the same color specified by using `'Color'`.

Example: `'MarkerFaceColor','blue'`

`MarkerSize` — Marker size
6 (default) | positive value

Marker size, specified as the comma-separated pair consisting of `'MarkerSize'` and a positive value in points.

Example: `'MarkerSize',2`

Output Arguments

`h` — Graphics objects
graphics array

Graphics objects corresponding to the lines or patch in the plot, returned as a graphics array. Use dot notation to query and set properties of the graphics objects. For details, see Line Properties and Patch Properties.

You can use name-value pair arguments to specify the appearance of residual data points or the appearance of the histogram, corresponding to the first graphics object `h(1)`.

More About

Deviance

Deviance is twice the loglikelihood of the model. Because this overall loglikelihood is a sum of loglikelihoods for each observation, a residual plot with the deviance type shows the loglikelihood per observation.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

Alternative Functionality

A `GeneralizedLinearModel` object provides multiple plotting functions.

- When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
- After fitting a model, use `plotPartialDependence` to understand the effect of a particular predictor. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `plotDiagnostics` | `plotPartialDependence` | `plotSlice`

Topics

“Residuals — Model Quality for Training Data” on page 12-19
“Generalized Linear Models” on page 12-9

Introduced in R2012a

plotResiduals

Class: `GeneralizedLinearMixedModel`

Plot residuals of generalized linear mixed-effects model

Syntax

```
plotResiduals(glme,plottype)
plotResiduals(glme,plottype,Name,Value)

h = plotResiduals( ___ )
```

Description

`plotResiduals(glme,plottype)` plots the raw conditional residuals of the generalized linear mixed-effects model `glme` in a plot of the type specified by `plottype`.

`plotResiduals(glme,plottype,Name,Value)` plots the conditional residuals of `glme` using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify to plot the Pearson residuals.

`h = plotResiduals( ___ )` returns a handle, `h`, to the lines or patches in the plot of residuals.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`plottype` — Type of residual plot

`'histogram'` (default) | `'caseorder'` | `'fitted'` | `'lagged'` | `'probability'` | `'symmetry'`

Type of residual plot, specified as one of the following.

Value	Description
`'histogram'`	Histogram of residuals
`'caseorder'`	Residuals versus case order. Case order is the same as the row order used in the input data `tbl` when fitting the model using `fitglme`.
`'fitted'`	Residuals versus fitted values
`'lagged'`	Residuals versus lagged residual ($r(t)$ versus $r(t-1)$)
`'probability'`	Normal probability plot
`'symmetry'`	Symmetry plot

Example: `plotResiduals(glme,'lagged')`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`ResidualType` — Residual type
`'raw'` (default) | `'Pearson'`

Residual type, specified by the comma-separated pair consisting of `ResidualType` and one of the following.

Residual Type	Formula
`'raw'`	$r_{ci} = y_i - g^{-1}\left(x_i^T\widehat{\beta} + z_i^T\widehat{b} + \delta_i\right)$
`'Pearson'`	$r_{ci}^{pearson} = \frac{r_{ci}}{\sqrt{\frac{\widehat{\sigma^2}}{w_i}v_i\left(\mu_i\left(\widehat{\beta}, \widehat{b}\right)\right)}}$

In each of these equations:

- y_i is the ith element of the n-by-1 response vector, y, where $i = 1, ..., n$.
- g^{-1} is the inverse link function for the model.
- x_i^T is the ith row of the fixed-effects design matrix X.
- z_i^T is the ith row of the random-effects design matrix Z.
- δ_i is the ith offset value.
- σ^2 is the dispersion parameter.
- w_i is the ith observation weight.
- v_i is the variance term for the ith observation.
- μ_i is the mean of the response for the ith observation.
- $\widehat{\beta}$ and $\widehat{b}$ are estimated values of β and b.

Raw residuals from a generalized linear mixed-effects model have nonconstant variance. Pearson residuals are expected to have an approximately constant variance, and are generally used for analysis.

Example: `'ResidualType','Pearson'`

Output Arguments

`h` — Handle to residual plot
graphics object

Handle to the residual plot, returned as a graphics object. You can use dot notation to change certain property values of the object, including face color for a histogram, and marker style and color for a scatterplot. For more information, see “Access Property Values”.

Examples

Create Plots of Residuals

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution:

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.

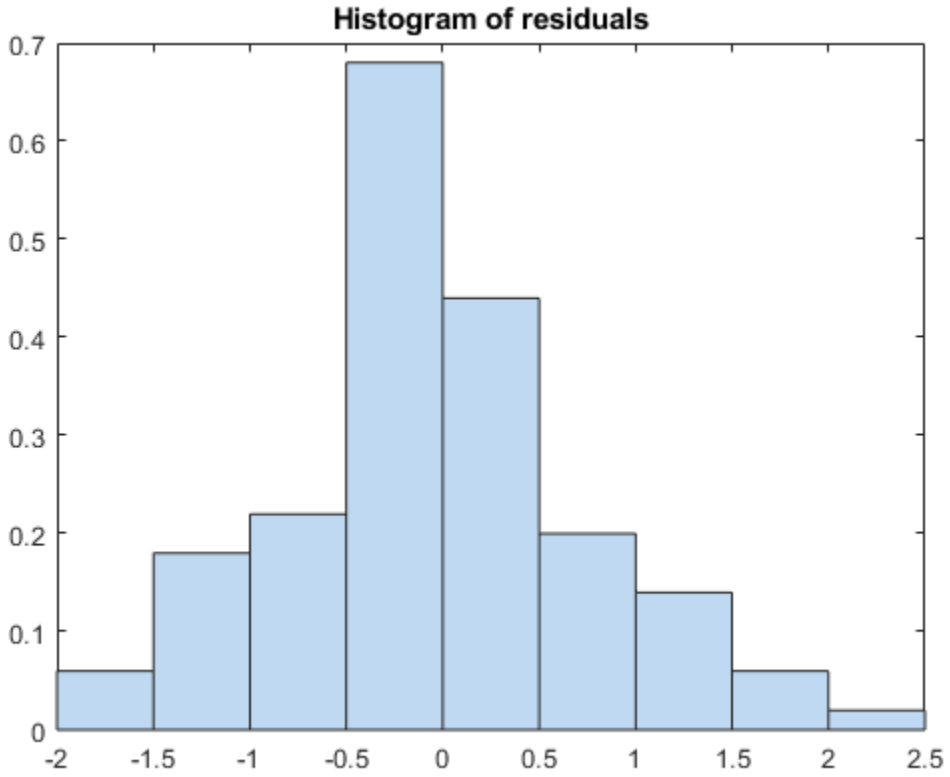
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Create diagnostic plots using Pearson residuals to test the model assumptions.

Plot a histogram to visually confirm that the mean of the Pearson residuals is equal to 0. If the model is correct, we expect the Pearson residuals to be centered at 0.

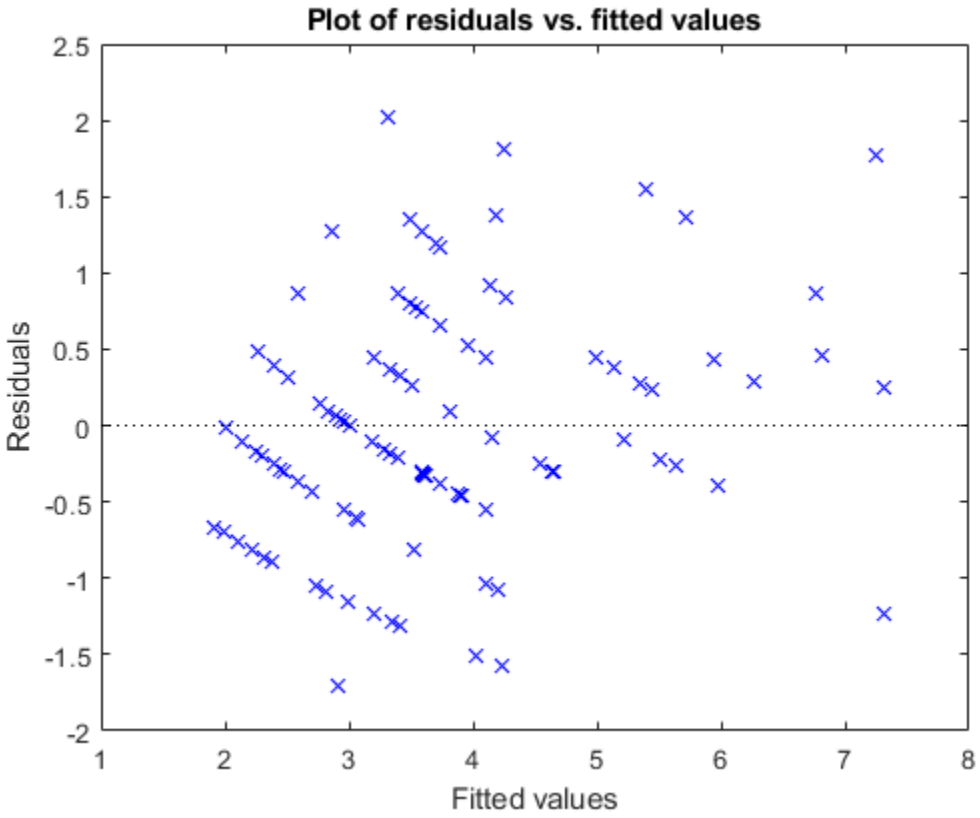
```
plotResiduals(glme,'histogram','ResidualType','Pearson')
```



The histogram shows that the Pearson residuals are centered at 0.

Plot the Pearson residuals versus the fitted values, to check for signs of nonconstant variance among the residuals (heteroscedasticity). We expect the conditional Pearson residuals to have a constant variance. Therefore, a plot of conditional Pearson residuals versus conditional fitted values should not reveal any systematic dependence on the conditional fitted values.

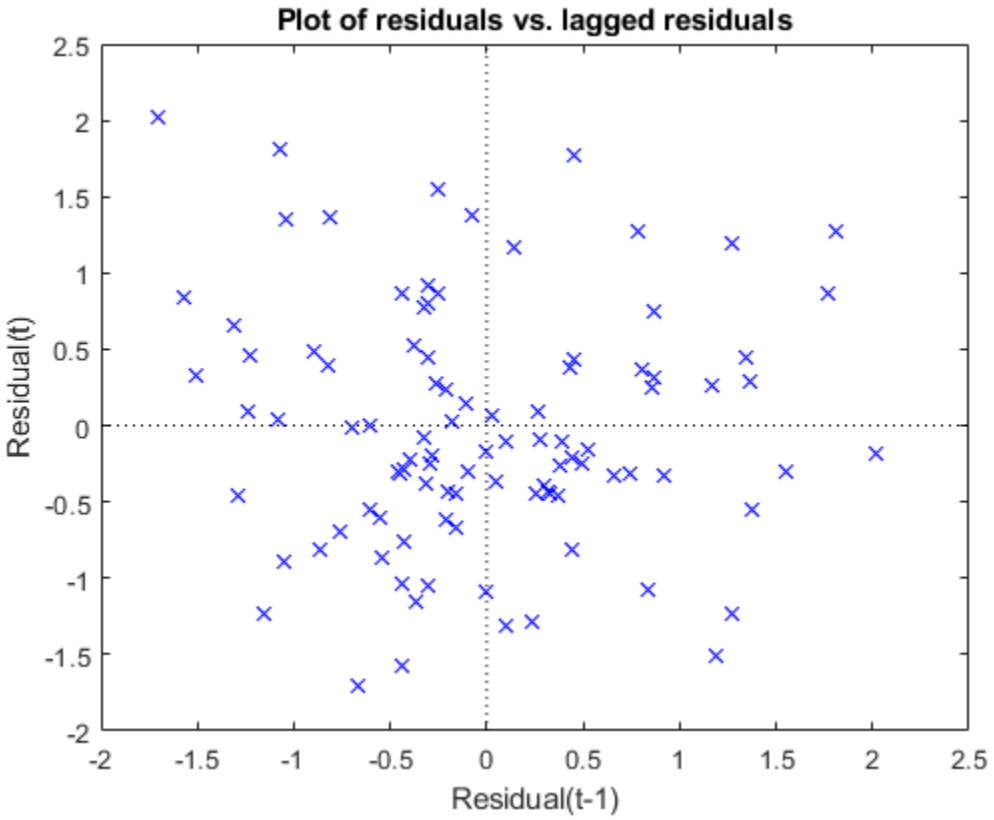
```
plotResiduals(glme,'fitted','ResidualType','Pearson')
```



The plot does not show a systematic dependence on the fitted values, so there are no signs of nonconstant variance among the residuals.

Plot the Pearson residuals versus lagged residuals, to check for correlation among the residuals. The conditional independence assumption in GLME implies that the conditional Pearson residuals are approximately uncorrelated.

```
plotResiduals(glme,'lagged','ResidualType','Pearson')
```



There is no pattern to the plot, so there are no signs of correlation among the residuals.

See Also

`GeneralizedLinearMixedModel` | `fitglme` | `fitted` | `plot` | `residuals`

plotResiduals

Plot residuals of linear regression model

Syntax

```
plotResiduals(mdl)
plotResiduals(mdl,plottype)
plotResiduals(mdl,plottype,Name,Value)

plotResiduals(ax,___)

h = plotResiduals(___)
```

Description

`plotResiduals(mdl)` creates a histogram plot of the linear regression model (`mdl`) residuals.

`plotResiduals(mdl,plottype)` specifies the residual plot type `plottype`.

`plotResiduals(mdl,plottype,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the residual type and the graphical properties of residual data points.

`plotResiduals(ax,___)` creates the plot in the axes specified by `ax` instead of the current axes, using any of the input argument combinations in the previous syntaxes.

`h = plotResiduals(___)` returns graphics objects for the lines or patch in the plot. Use `h` to modify the properties of a specific line or patch after you create the plot. For a list of properties, see Chart Line and Patch Properties.

Examples

Histogram of Residuals

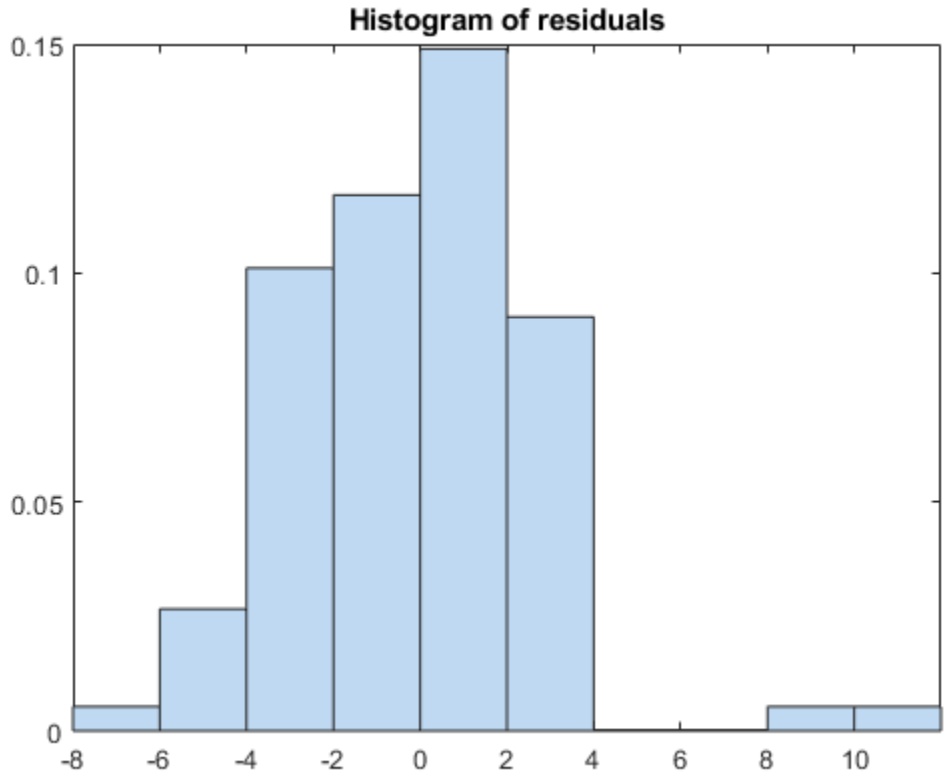
Plot a histogram of the residuals of a fitted linear regression model.

Load the `carsmall` data set and fit a linear regression model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
tbl = table(MPG,Weight);
tbl.Year = categorical(Model_Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Create a histogram of the raw residuals using probability density function scaling.

```
plotResiduals(mdl)
```



The area of each bar is the relative number of observations. The sum of the bar areas is equal to 1.

Normal Probability Plot of Residuals

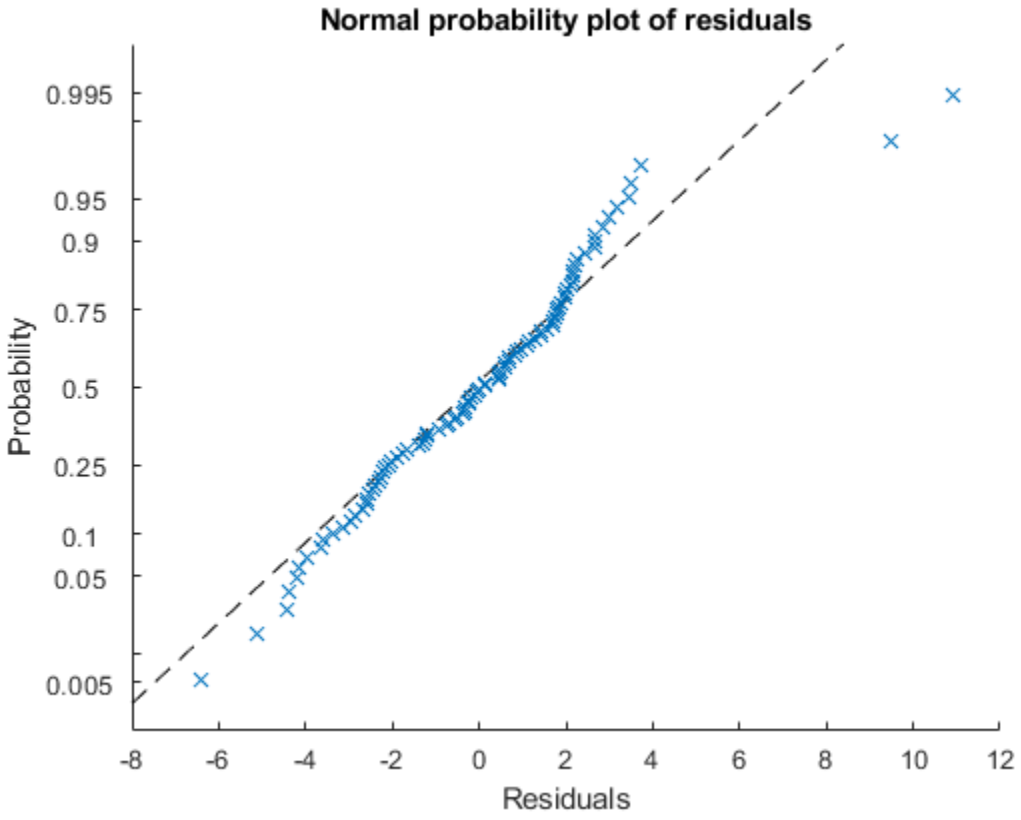
Create a normal probability plot of the residuals of a fitted linear regression model.

Load the `carsmall` data set and fit a linear regression model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
X = [Weight,Model_Year];
mdl = fitlm(X,MPG,...
    'y ~ x2 + x1^2','Categorical',2);
```

Create a normal probability plot of the residuals of the fitted model.

```
plotResiduals(mdl,'probability')
```



Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`plottype` — Plot type

`'histogram'` (default) | `'caseorder'` | `'fitted'` | `'lagged'` | `'probability'` | `'symmetry'`

Plot type, specified as one of the values in this table.

Value	Description
`'caseorder'`	Residuals vs. case order (row number)
`'fitted'`	Residuals vs. fitted values
`'histogram'`	Histogram of residuals using probability density function scaling. The area of each bar is the relative number of observations. The sum of the bar areas is equal to 1.
`'lagged'`	Residuals vs. lagged residuals ($r(t)$ vs. $r(t-1)$)
`'probability'`	Normal probability plot of residuals. For details, see `probplot`.

Value	Description
`'symmetry'`	Symmetry plot of residuals around their median (residuals in upper tail – median vs. median – residuals in lower tail). This plot includes a dotted reference line of $y = x$ to examine the symmetry of residuals.

ax — Target axes
`Axes` object

Target axes, specified as an `Axes` object.

If you do not specify the axes and the current axes are Cartesian, then `plotResiduals` uses the current axes (`gca`). For more information on creating an `Axes` object, see `axes` and `gca`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Color','blue','Marker','o'`

Note The graphical properties listed here are only a subset. For a complete list, see Chart Line for lines and Patch Properties for histogram. The specified properties apply to the appearance of residual data points or the appearance of the histogram.

`ResidualType` — Type of residual
`'raw'` (default) | `'pearson'` | `'standardized'` | `'studentized'`

Type of residual used in the plot, specified as the comma-separated pair consisting of `'ResidualType'` and one of these values:

Value	Description
`'raw'`	Observed minus fitted values
`'pearson'`	Raw residuals divided by the root mean squared error (RMSE)
`'standardized'`	Raw residuals divided by their estimated standard deviation
`'studentized'`	Raw residuals divided by an independent (delete-1) estimate of their standard deviation

The `Residuals` property of `mdl` contains the residual values used by `plotResiduals` to create plots.

For details, see “Residuals” on page 11-80.

Example: `'ResidualType','Pearson'`

`Color` — Line color
RGB triplet | hexadecimal color code | color name | short name

Line color, specified as the comma-separated pair consisting of `'Color'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

The `'Color'` name-value pair argument also determines marker outline color and marker fill color if `'MarkerEdgeColor'` is `'auto'` (default) and `'MarkerFaceColor'` is `'auto'`.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	
`'none'`	Not applicable	Not applicable	Not applicable	No color

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

Example: `'Color','blue'`

`LineWidth` — Line width
positive value

Line width, specified as the comma-separated pair consisting of `'LineWidth'` and a positive value in points. If the line has markers, then the line width also affects the marker edges.

Example: `'LineWidth',0.75`

`Marker` — Marker symbol
`'o'` | `'+'` | `'*'` | `'.'` | `'x'` | ...

Marker symbol, specified as the comma-separated pair consisting of `'Marker'` and one of the values in this table.

Value	Description
`'o'`	Circle
`'+'`	Plus sign
`'*'`	Asterisk
`'.'`	Point
`'x'`	Cross
`'_'`	Horizontal line
`'|'`	Vertical line
`'square'` or `'s'`	Square
`'diamond'` or `'d'`	Diamond
`'^'`	Upward-pointing triangle
`'v'`	Downward-pointing triangle
`'>'`	Right-pointing triangle
`'<'`	Left-pointing triangle
`'pentagram'` or `'p'`	Five-pointed star (pentagram)
`'hexagram'` or `'h'`	Six-pointed star (hexagram)
`'none'`	No markers

Example: `'Marker','+'`

`MarkerEdgeColor` — Marker outline color
`'auto'` (default) | `'none'` | RGB triplet | hexadecimal color code | color name | short name

Marker outline color, specified as the comma-separated pair consisting of `'MarkerEdgeColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The default value of `'auto'` uses the same color specified by using `'Color'`.

Example: `'MarkerEdgeColor','blue'`

`MarkerFaceColor` — Marker fill color
`'none'` (default) | `'auto'` | RGB triplet | hexadecimal color code | color name | short name

Marker fill color, specified as the comma-separated pair consisting of `'MarkerFaceColor'` and an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

The `'auto'` value uses the same color specified by using `'Color'`.

Example: `'MarkerFaceColor','blue'`

`MarkerSize` — Marker size
6 (default) | positive value

Marker size, specified as the comma-separated pair consisting of `'MarkerSize'` and a positive value in points.

Example: `'MarkerSize',2`

Output Arguments

`h` — Graphics objects
graphics array

Graphics objects corresponding to the lines or patch in the plot, returned as a graphics array. Use dot notation to query and set properties of the graphics objects. For details, see Line Properties and Patch Properties.

You can use name-value pair arguments to specify the appearance of residual data points or the appearance of the histogram, corresponding to the first graphics object `h(1)`.

Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

Alternative Functionality

- A `LinearModel` object provides multiple plotting functions.
  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`LinearModel` | `plotDiagnostics`

Topics

“Residuals” on page 11-80
“Compare large and small stepwise models” on page 11-99
“Compare Results of Standard and Robust Least-Squares Fit” on page 11-105
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

plotResiduals

Class: `LinearMixedModel`

Plot residuals of linear mixed-effects model

Syntax

```
plotResiduals(lme,plottype)
plotResiduals(lme,plottype,Name,Value)

h = plotResiduals( ___ )
```

Description

`plotResiduals(lme,plottype)` plots the raw conditional residuals of the linear mixed-effects model `lme` in a plot of the type specified by `plottype`.

`plotResiduals(lme,plottype,Name,Value)` also plots the residuals of the linear mixed-effects model `lme` with additional options specified by one or more name-value pair arguments. For example, you can specify the residual type to plot.

`plotResiduals` also accepts some other name-value pair arguments that specify the properties of the primary line in the plot. For those name-value pairs, see `plot`.

`h = plotResiduals( ___ )` returns a handle, `h`, to the lines or patches in the plot of residuals.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`plottype` — Type of residual plot

`'histogram'` (default) | `'caseorder'` | `'fitted'` | `'lagged'` | `'probability'` | `'symmetry'`

Type of residual plot, specified as one of the following.

`'histogram'`	Default. Histogram of residuals
`'caseorder'`	Residuals versus case (row) order
`'fitted'`	Residuals versus fitted values
`'lagged'`	Residuals versus lagged residual ($r(t)$ versus $r(t-1)$)
`'probability'`	Normal probability plot
`'symmetry'`	Symmetry plot

Example: `plotResiduals(lme,'lagged')`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

ResidualType — Residual type

`'Raw'` (default) | `'Pearson'` | `'Standardized'`

Residual type, specified by the comma-separated pair consisting of `ResidualType` and one of the following.

Residual Type	Conditional	Marginal
`'Raw'`	$r_i^C = \left[y - X\widehat{\beta} - Z\widehat{b}\right]_i$	$r_i^M = \left[y - X\widehat{\beta}\right]_i$
`'Pearson'`	$pr_i^C = \frac{r_i^C}{\sqrt{\left[\widehat{Var}_{y,b}(y - X\beta - Zb)\right]_{ii}}}$	$pr_i^M = \frac{r_i^M}{\sqrt{\left[\widehat{Var}_y(y - X\beta)\right]_{ii}}}$
`'Standardized'`	$st_i^C = \frac{r_i^C}{\sqrt{\left[\widehat{Var}_y(r^C)\right]_{ii}}}$	$st_i^M = \frac{r_i^M}{\sqrt{\left[\widehat{Var}_y(r^M)\right]_{ii}}}$

For more information on the conditional and marginal residuals and residual variances, see `Definitions` at the end of this page.

Example: `'ResidualType','Standardized'`

Output Arguments

h — Handle to residual plot

handle

Handle to the residual plot, returned as a handle.

Examples

Examine Residuals

Load the sample data.

```
load('weight.mat')
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

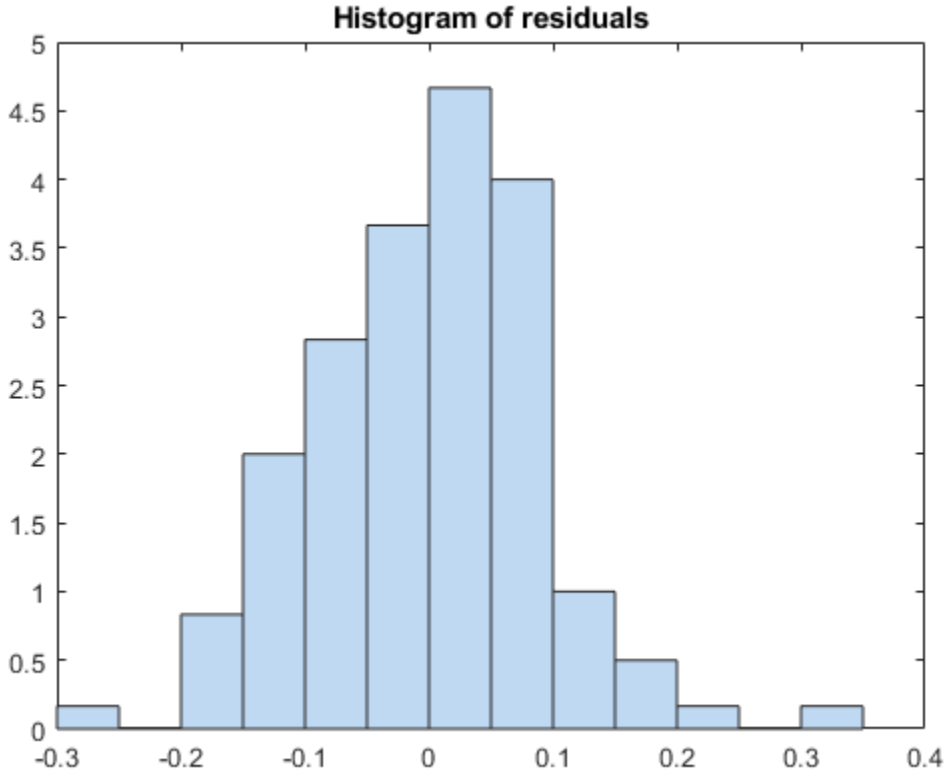
```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = categorical(tbl.Subject);
tbl.Program = categorical(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

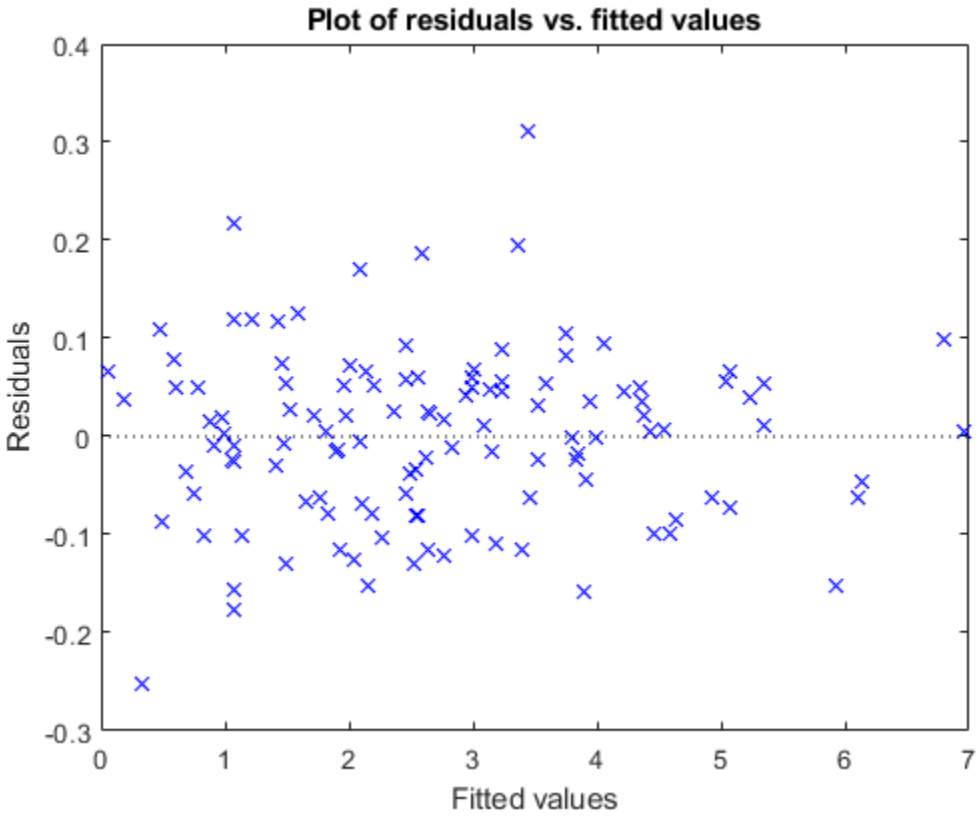
Plot the histogram of the raw residuals.

```
plotResiduals(lme)
```



Plot the residuals versus the fitted values.

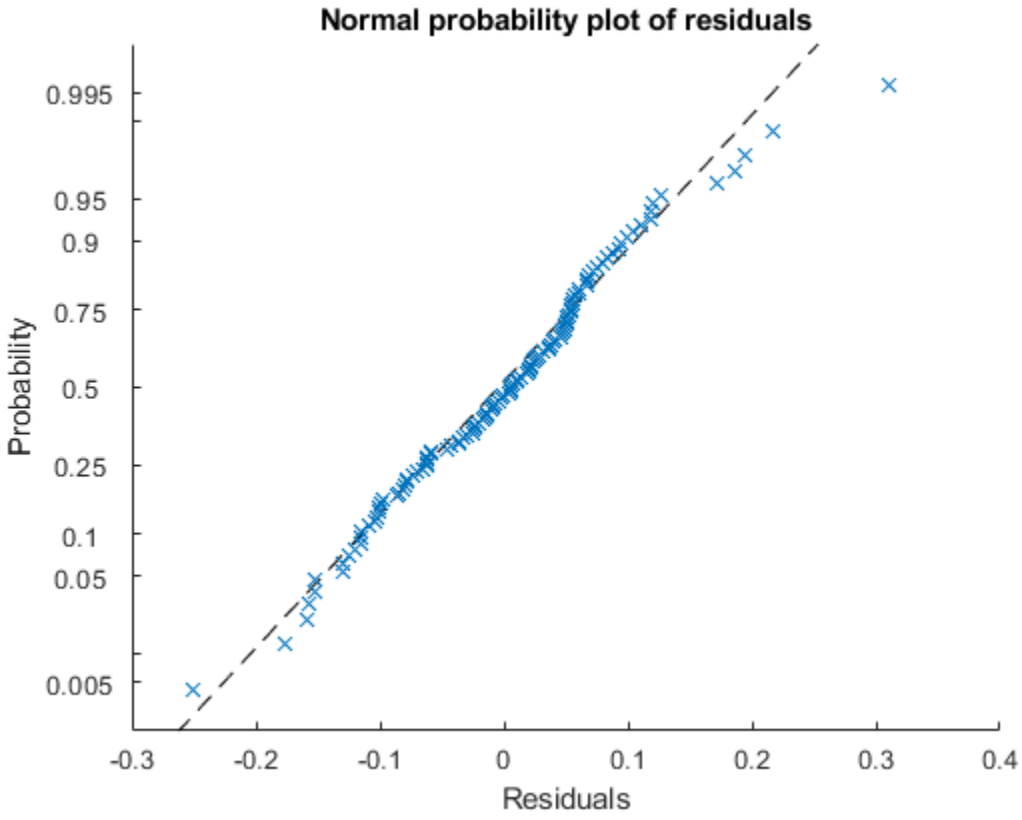
```
plotResiduals(lme,'fitted')
```



There is no obvious pattern, so there are no immediate signs of heteroscedasticity.

Create the normal probability plot of residuals.

```
plotResiduals(lme,'probability')
```



Data appears to be normal.

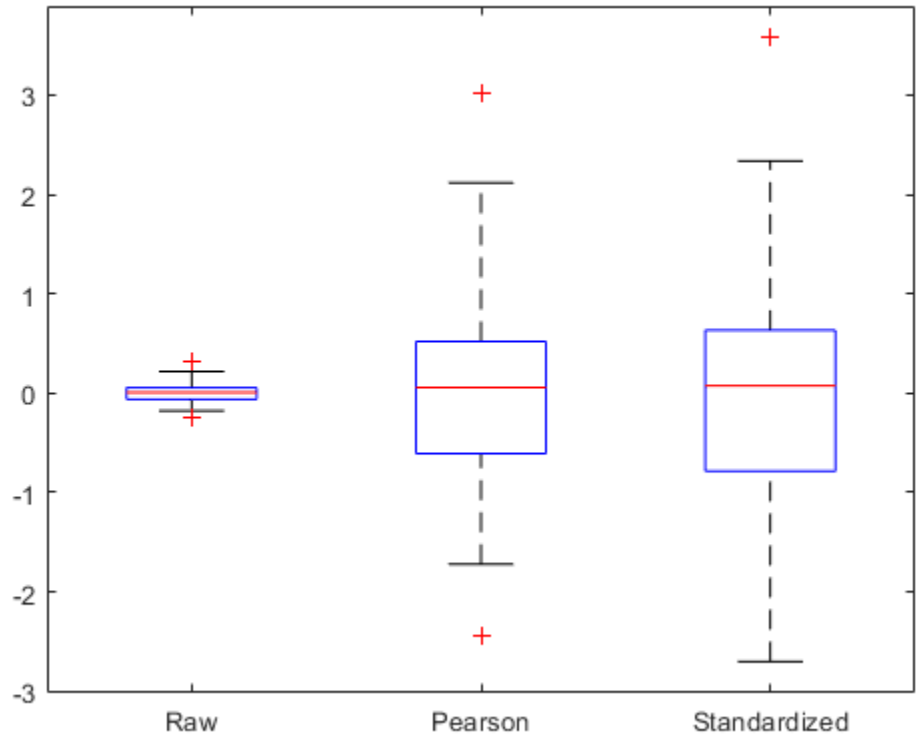
Find the observation number for the data that appears to be an outlier to the right of the plot.

```
find(residuals(lme)>0.25)
```

```
ans = 101
```

Create a box plot of the raw, Pearson, and standardized residuals.

```
r = residuals(lme);
pr = residuals(lme,'ResidualType','Pearson');
st = residuals(lme,'ResidualType','Standardized');
X = [r pr st];
boxplot(X,'labels',{'Raw','Pearson','Standardized'})
```



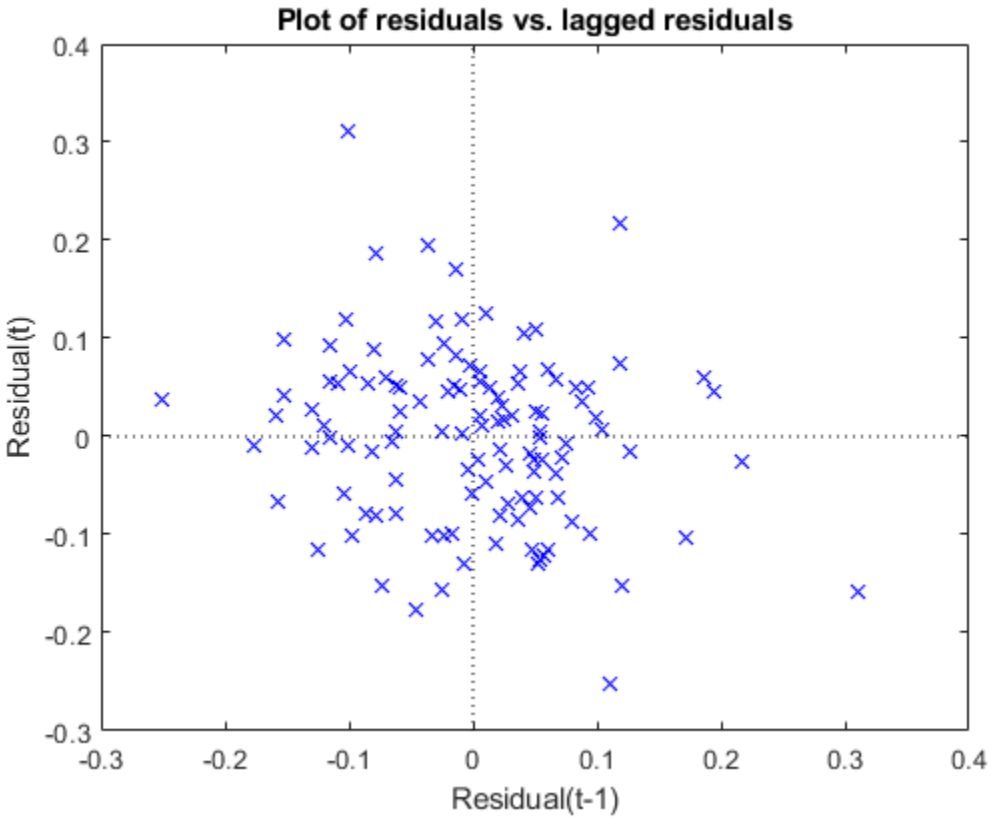
All three box plots point out the outlier on the right tail of the distribution. The box plots of raw and Pearson residuals also point out a second possible outlier on the left tail. Find the corresponding observation number.

```
find(pr<-2)
```

```
ans = 10
```

Plot the raw residuals versus lagged residuals.

```
plotResiduals(lme,'lagged')
```



There is no obvious pattern in the graph. The residuals do not appear to be correlated.

See Also

`LinearMixedModel` | `fitted` | `residuals`

plotResiduals

Class: `NonLinearModel`

Plot residuals of nonlinear regression model

Syntax

```
plotResiduals(mdl)
plotResiduals(mdl,plottype)
h = plotResiduals(...)
h = plotResiduals(mdl,plottype,Name,Value)
```

Description

`plotResiduals(mdl)` gives a histogram plot of the residuals of the `mdl` nonlinear model.

`plotResiduals(mdl,plottype)` plots residuals in a plot of type `plottype`.

`h = plotResiduals(...)` returns handles to the lines in the plot.

`h = plotResiduals(mdl,plottype,Name,Value)` plots with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`plottype`

Character vector or string scalar specifying the type of plot:

`'caseorder'`	Residuals vs. case (row) order
`'fitted'`	Residuals vs. fitted values
`'histogram'`	Histogram
`'lagged'`	Residuals vs. lagged residual ($r(t)$ vs. $r(t–1)$)
`'probability'`	Normal probability plot
`'symmetry'`	Symmetry plot

Default: `'histogram'`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note The plot property name-value pairs apply to the first returned handle `h(1)`.

`Color`

Color of the line or marker, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the following table.

For a custom color, specify an RGB triplet or a hexadecimal color code.

- An RGB triplet is a three-element row vector whose elements specify the intensities of the red, green, and blue components of the color. The intensities must be in the range `[0,1]`; for example, `[0.4 0.6 0.7]`.
- A hexadecimal color code is a character vector or a string scalar that starts with a hash symbol (`#`) followed by three or six hexadecimal digits, which can range from `0` to `F`. The values are not case sensitive. Thus, the color codes `'#FF8800'`, `'#ff8800'`, `'#F80'`, and `'#f80'` are equivalent.

Alternatively, you can specify some common colors by name. This table lists the named color options, the equivalent RGB triplets, and hexadecimal color codes.

Color Name	Short Name	RGB Triplet	Hexadecimal Color Code	Appearance
`'red'`	`'r'`	`[1 0 0]`	`'#FF0000'`	
`'green'`	`'g'`	`[0 1 0]`	`'#00FF00'`	
`'blue'`	`'b'`	`[0 0 1]`	`'#0000FF'`	
`'cyan'`	`'c'`	`[0 1 1]`	`'#00FFFF'`	
`'magenta'`	`'m'`	`[1 0 1]`	`'#FF00FF'`	
`'yellow'`	`'y'`	`[1 1 0]`	`'#FFFF00'`	
`'black'`	`'k'`	`[0 0 0]`	`'#000000'`	
`'white'`	`'w'`	`[1 1 1]`	`'#FFFFFF'`	

Here are the RGB triplets and hexadecimal color codes for the default colors MATLAB uses in many types of plots.

RGB Triplet	Hexadecimal Color Code	Appearance
`[0 0.4470 0.7410]`	`'#0072BD'`	
`[0.8500 0.3250 0.0980]`	`'#D95319'`	
`[0.9290 0.6940 0.1250]`	`'#EDB120'`	
`[0.4940 0.1840 0.5560]`	`'#7E2F8E'`	
`[0.4660 0.6740 0.1880]`	`'#77AC30'`	
`[0.3010 0.7450 0.9330]`	`'#4DBEEE'`	
`[0.6350 0.0780 0.1840]`	`'#A2142F'`	

`LineStyle`

Type of line, a Chart Line specification. For details, see `linespec`.

`LineWidth`

Width of the line or edges of filled area, in points, a positive scalar. One point is 1/72 inch.

Default: `0.5`

`MarkerEdgeColor`

Marker outline color, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

`MarkerFaceColor`

Fill color for filled markers, specified as an RGB triplet, hexadecimal color code, color name, or short name for one of the color options listed in the `Color` name-value pair argument.

`MarkerSize`

Size of the marker in points, a strictly positive scalar. One point is 1/72 inch.

`ResidualType`

Type of residual used in the plot:

`'Raw'`	Observed minus fitted values
`'Pearson'`	Raw residuals divided by RMSE
`'Standardized'`	Raw residuals divided by their estimated standard deviation
`'Studentized'`	Raw residuals divided by an independent (delete-1) estimate of their standard deviation

Default: `'Raw'`

Output Arguments

`h`

Vector of handles to lines or patches in the plot.

Examples

Residual Plot

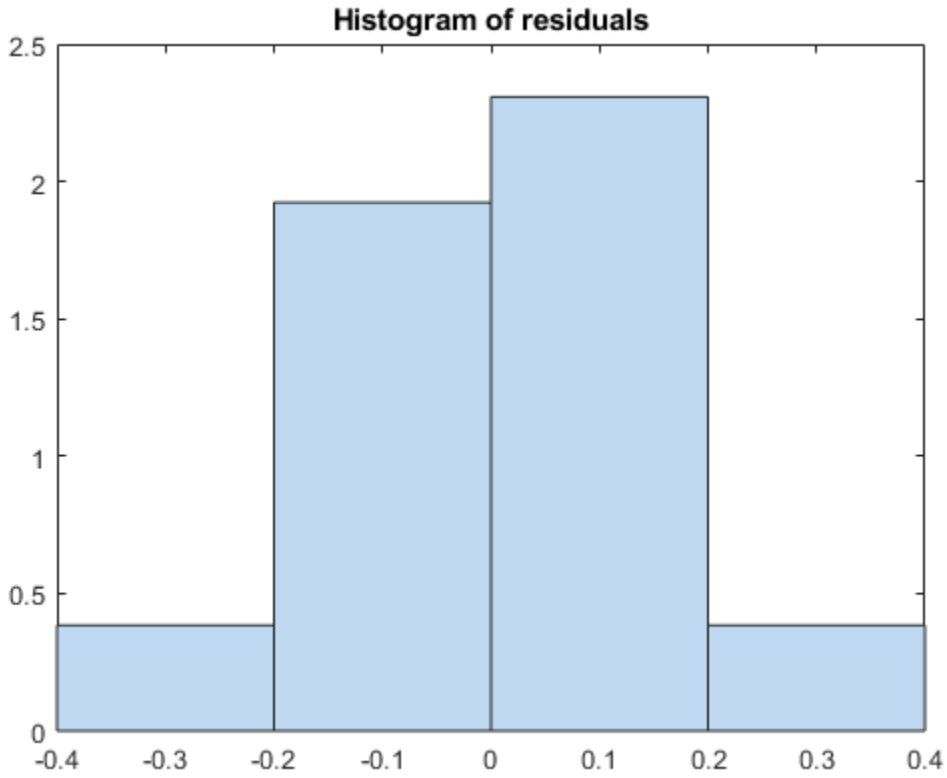
Plot the residuals of a fitted nonlinear model.

Load the `reaction` data and fit a model of the reaction rate as a function of reactants.

```
load reaction
mdl = fitnlm(reactants,rate,@hougen,[1 .05 .02 .1 2]);
```

Plot the residuals of the fitted model.

```
plotResiduals(mdl)
```



Residual Probability Plot

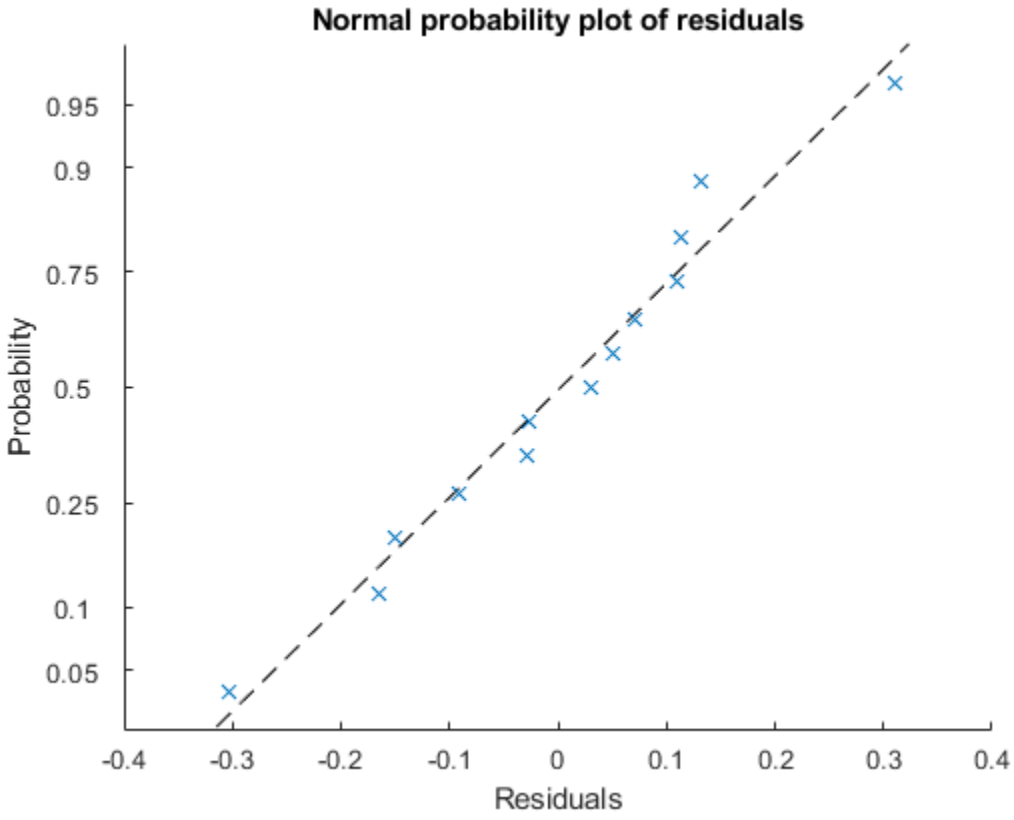
Create a normal probability plot of the residuals of a fitted nonlinear model.

Load the `reaction` data and fit a model of the reaction rate as a function of reactants.

```
load reaction
mdl = fitnlm(reactants,rate,@hougen,[1 .05 .02 .1 2]);
```

Create a normal probability plot of the residuals of the fitted model.

```
plotResiduals(mdl,'probability')
```



Tips

- The data cursor displays the values of the selected plot point in a data tip (small text box located next to the data point). The data tip includes the *x*-axis and *y*-axis values for the selected point, along with the observation name or number.

See Also

`NonLinearModel` | `plotDiagnostics`

Topics

“Examine Quality and Adjust the Fitted Nonlinear Model” on page 13-6
“Nonlinear Regression Workflow” on page 13-12
“Nonlinear Regression” on page 13-2

plotSlice

Package:

Plot of slices through fitted generalized linear regression surface

Syntax

```
plotSlice(mdl)
```

Description

`plotSlice(mdl)` creates a figure containing one or more plots, each representing a slice through the regression surface predicted by `mdl`. Each plot shows the fitted response values as a function of a single predictor variable, with the other predictor variables held constant.

`plotSlice` also displays the 95% confidence bounds for the response values. Use the **Bounds** menu to choose the type of confidence bounds, and use the **Predictors** menu to select which predictors to use for each slice plot. For details, see “Tips” on page 33-4706.

Examples

Slice Plot for Generalized Linear Regression Model

Plot slices through a fitted generalized linear regression model surface.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

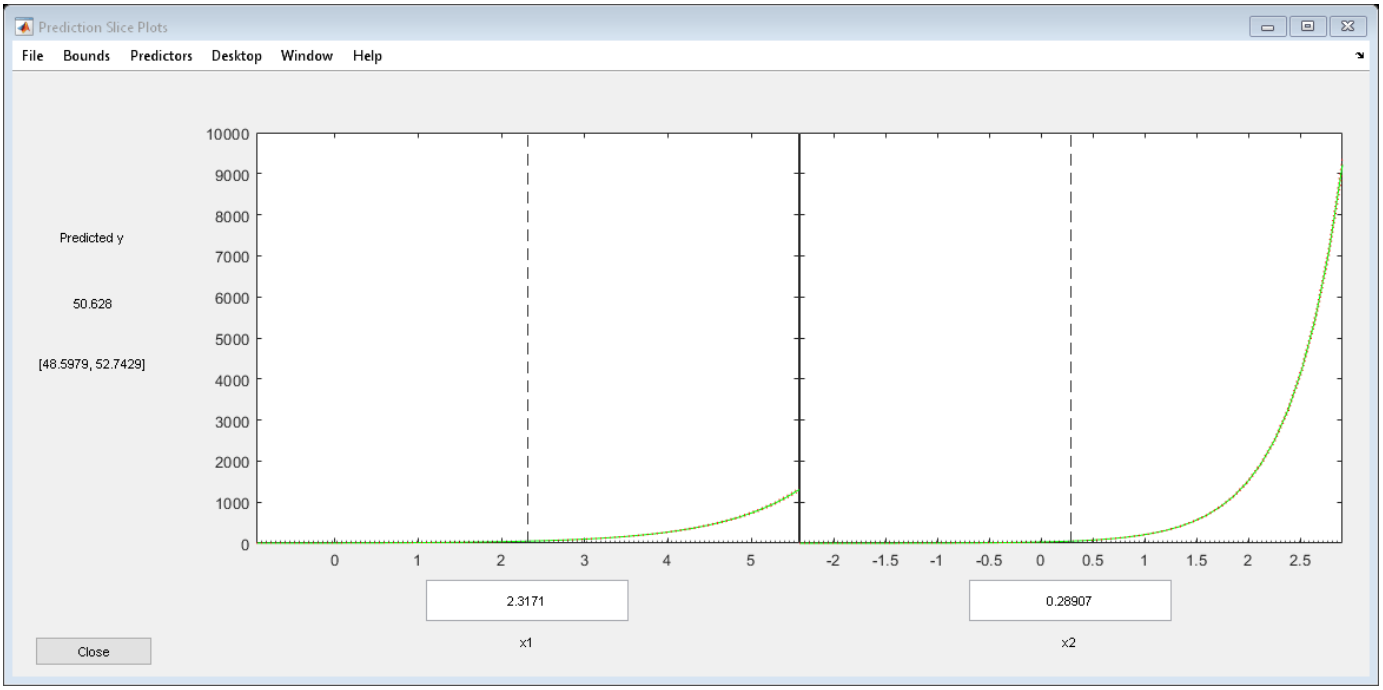
```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

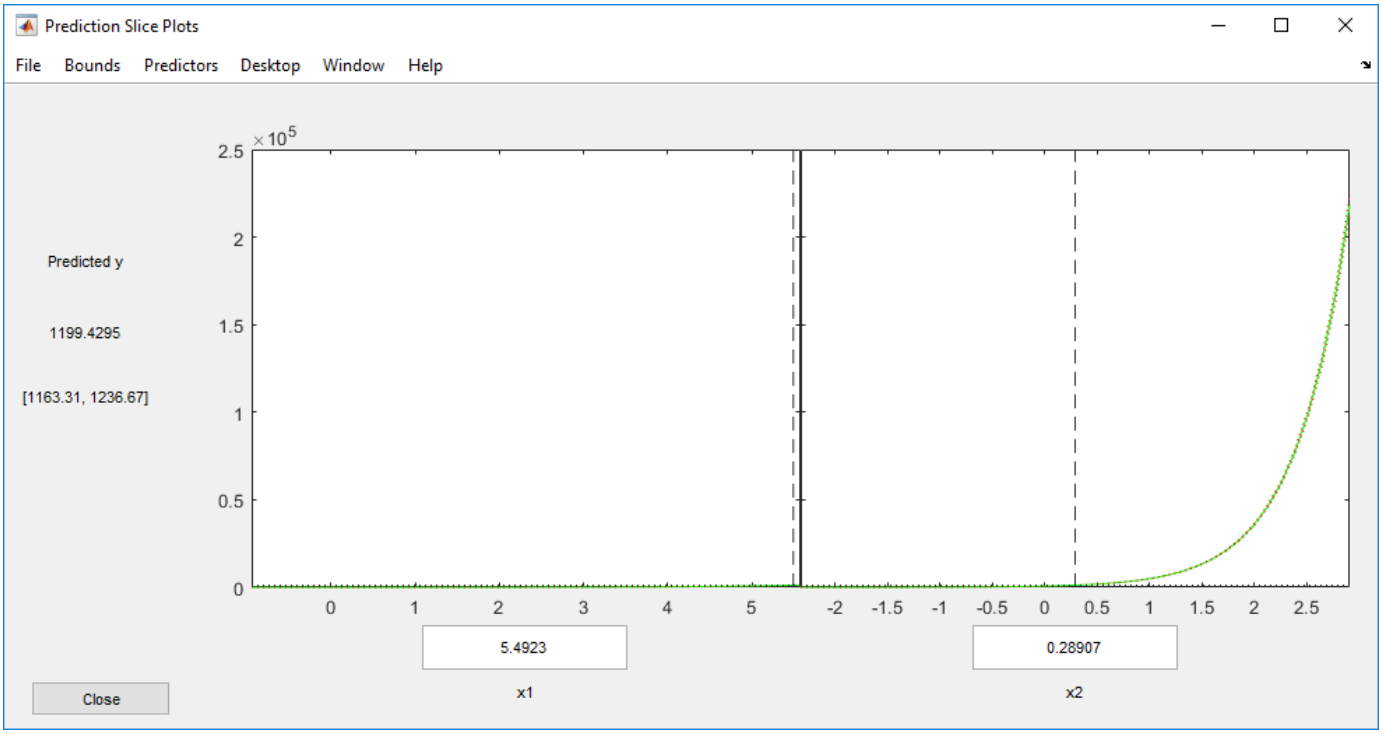
Create a slice plot.

```
plotSlice(mdl)
```



The green line in each plot represents the predicted response values as a function of a single predictor variable, with the other predictor variables held constant. The red dotted lines are the 95% confidence bounds. The *y*-axis label includes the predicted response value and the corresponding confidence bound for the point selected by the vertical and horizontal lines. The *x*-axis label shows the predictor variable name and the predictor value for the selected point.

Move the vertical line in the `x1` plot to the right and observe the change in the *y*-axis label and the changes in the `x2` plot.



Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

Tips

- Use the **Bounds** menu in the figure window to choose the type of confidence bounds. You can choose **Simultaneous** or **Non-Simultaneous**. You can also choose **No Bounds** to have no confidence bounds.
  - Simultaneous (default) — `plotSlice` computes confidence bounds for the curve of the response values using Scheffe's method. The range between the upper and lower confidence bounds contains the curve consisting of true response values with 95% confidence.
  - Non-Simultaneous — `plotSlice` computes confidence bounds for the response value at each observation. The confidence interval for a response value at a specific predictor value contains the true response value with 95% confidence.

  Simultaneous bounds are wider than separate bounds, because requiring the entire curve of response values to be within the bounds is stricter than requiring the response value at a single predictor value to be within the bounds.
- Use the **Predictors** menu in the figure window to select which predictors to use for each slice plot. If the regression model `mdl` includes more than eight predictors, `plotSlice` creates plots for the first five predictors by default.

Alternative Functionality

- Use `predict` to return the predicted response values and confidence bounds. You can also specify the confidence level for confidence bounds by using the `'Alpha'` name-value pair argument of the `predict` function. Note that `predict` finds nonsimultaneous bounds by default, whereas `plotSlice` finds simultaneous bounds by default.
- A `GeneralizedLinearModel` object provides multiple plotting functions.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotPartialDependence` to understand the effect of a particular predictor. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `plotDiagnostics` | `plotPartialDependence` | `plotResiduals` | `predict`

Topics

“Diagnostic Plots” on page 12-17
“Plots to Understand Predictor Effects and How to Modify a Model” on page 12-21
“Generalized Linear Models” on page 12-9

Introduced in R2012a

plotSlice

Package:

Plot of slices through fitted linear regression surface

Syntax

```
plotSlice(mdl)
```

Description

`plotSlice(mdl)` creates a figure containing one or more plots, each representing a slice through the regression surface predicted by `mdl`. Each plot shows the fitted response values as a function of a single predictor variable, with the other predictor variables held constant.

`plotSlice` also displays the 95% confidence bounds for the response values. Use the **Bounds** menu to choose the type of confidence bounds, and use the **Predictors** menu to select which predictors to use for each slice plot. For details, see “Tips” on page 33-4710.

Examples

Slice Plot for Linear Regression Model

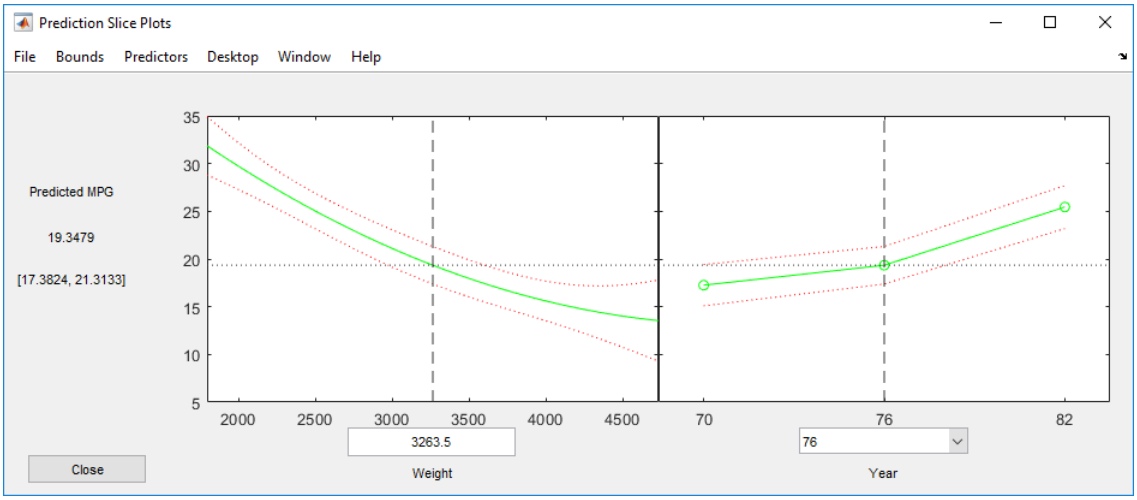
Plot slices through a fitted linear regression model surface.

Load the `carsmall` data set and fit a linear regression model of the mileage as a function of model year, weight, and weight squared.

```
load carsmall
Year = categorical(Model_Year);
tbl = table(MPG,Weight,Year);
mdl = fitlm(tbl,'MPG ~ Year + Weight^2');
```

Create a slice plot.

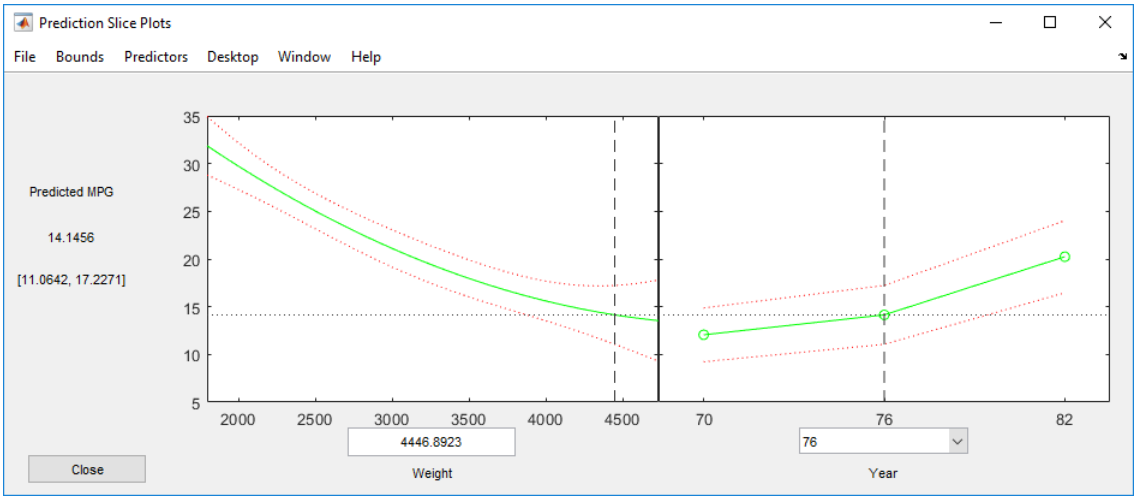
```
plotSlice(mdl)
```



The green line in each plot represents the predicted response values as a function of a single predictor variable, with the other predictor variables held constant. The red dotted lines are the 95% confidence bounds. The *y*-axis label includes the predicted response value and the corresponding confidence bound for the point selected by the vertical and horizontal lines. The *x*-axis label shows the predictor variable name and the predictor value for the selected point.

Note that `mdl` includes both the `Weight` and `Weight^2` terms, but `plotSlice` creates only one plot for the `Weight` term.

Move the vertical line in the `Weight` plot to the right and observe the change in the *y*-axis label and the changes in the `Year` plot.



Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

Tips

- Use the **Bounds** menu in the figure window to choose the type of confidence bounds. You can choose **Simultaneous** or **Non-Simultaneous**, and **Curve** or **Observation**. You can also choose **No Bounds** to have no confidence bounds.

  - **Simultaneous** or **Non-Simultaneous**

    - Simultaneous (default) — `plotSlice` computes confidence bounds for the curve of the response values using Scheffe's method. The range between the upper and lower confidence bounds contains the curve consisting of true response values with 95% confidence.
    - Non-Simultaneous — `plotSlice` computes confidence bounds for the response value at each observation. The confidence interval for a response value at a specific predictor value contains the true response value with 95% confidence.

    Simultaneous bounds are wider than separate bounds, because requiring the entire curve of response values to be within the bounds is stricter than requiring the response value at a single predictor value to be within the bounds.

  - **Curve** or **Observation**

    A regression model for the predictor variables X and the response variable y has the form
    $y = f(X) + \varepsilon$,
    where f is a function of X and ε is a random noise term.

    - **Curve** (default) — `plotSlice` predicts confidence bounds for the fitted responses $f(X)$.
    - **Observation** — `plotSlice` predicts confidence bounds for the response observations y.

    The bounds for y are wider than the bounds for $f(X)$ because of the additional variability of the noise term.

- Use the **Predictors** menu in the figure window to select which predictors to use for each slice plot. If the regression model `mdl` includes more than eight predictors, `plotSlice` creates plots for the first five predictors by default.

Alternative Functionality

- Use `predict` to return the predicted response values and confidence bounds. You can also specify the confidence level for confidence bounds by using the `'Alpha'` name-value pair argument of the `predict` function. Note that `predict` finds nonsimultaneous bounds by default whereas `plotSlice` finds simultaneous bounds by default.
- A `LinearModel` object provides multiple plotting functions.

  - When creating a model, use `plotAdded` to understand the effect of adding or removing a predictor variable.
  - When verifying a model, use `plotDiagnostics` to find questionable data and to understand the effect of each observation. Also, use `plotResiduals` to analyze the residuals of the model.
  - After fitting a model, use `plotAdjustedResponse`, `plotPartialDependence`, and `plotEffects` to understand the effect of a particular predictor. Use `plotInteraction` to understand the interaction effect between two predictors. Also, use `plotSlice` to plot slices through the prediction surface.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`CompactLinearModel` | `LinearModel` | `predict`

Topics

“Linear Regression with Interaction Effects” on page 11-44
“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35
“Linear Regression” on page 11-9

Introduced in R2012a

plotSlice

Class: `NonLinearModel`

Plot of slices through fitted nonlinear regression surface

Syntax

```
plotSlice(mdl)
h = plotSlice(mdl)
```

Description

`plotSlice(mdl)` creates a new figure containing a series of plots, each representing a slice through the regression surface predicted by `mdl`. For each plot, the surface slice is shown as a function of a single predictor variable, with the other predictor variables held constant.

`h = plotSlice(mdl)` returns handles to the lines in the plot.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

Output Arguments

`h`

Vector of handles to lines or patches in the plot.

Examples

Slice Plot

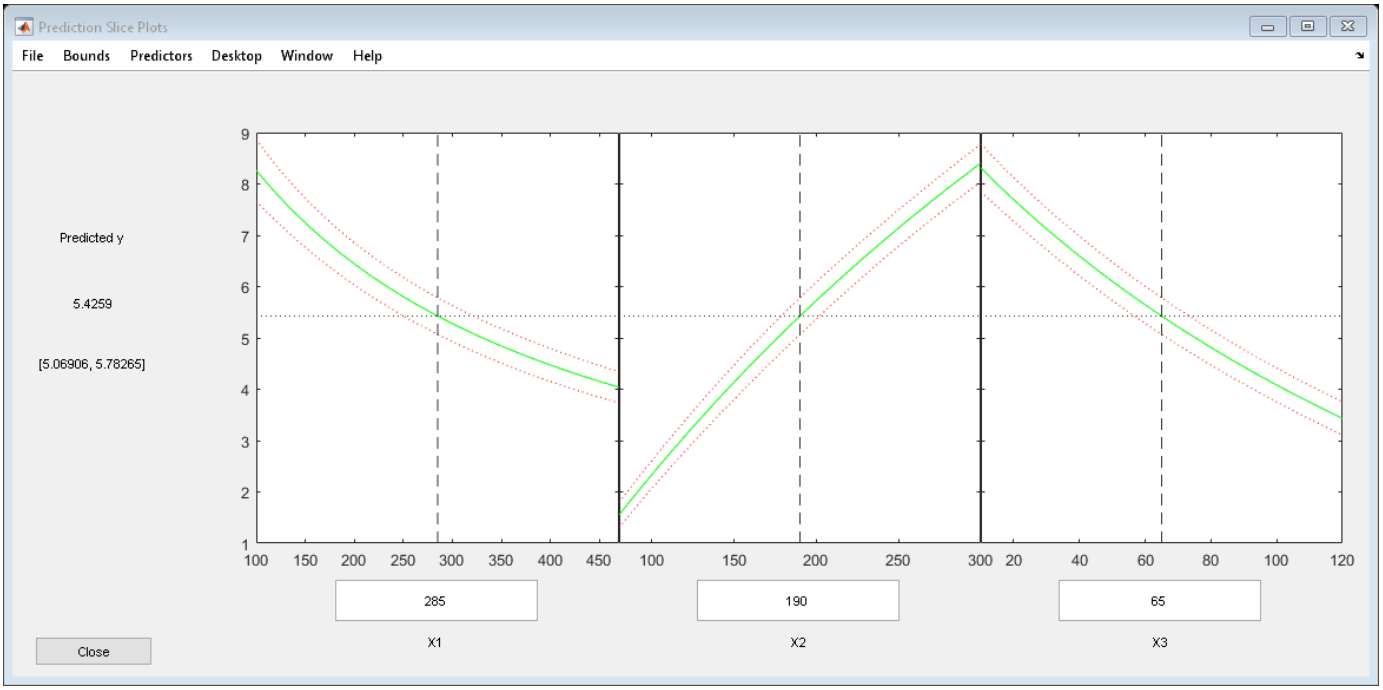
Plot slices of a fitted nonlinear model.

Load the `reaction` data and fit a model of the reaction rate as a function of reactants.

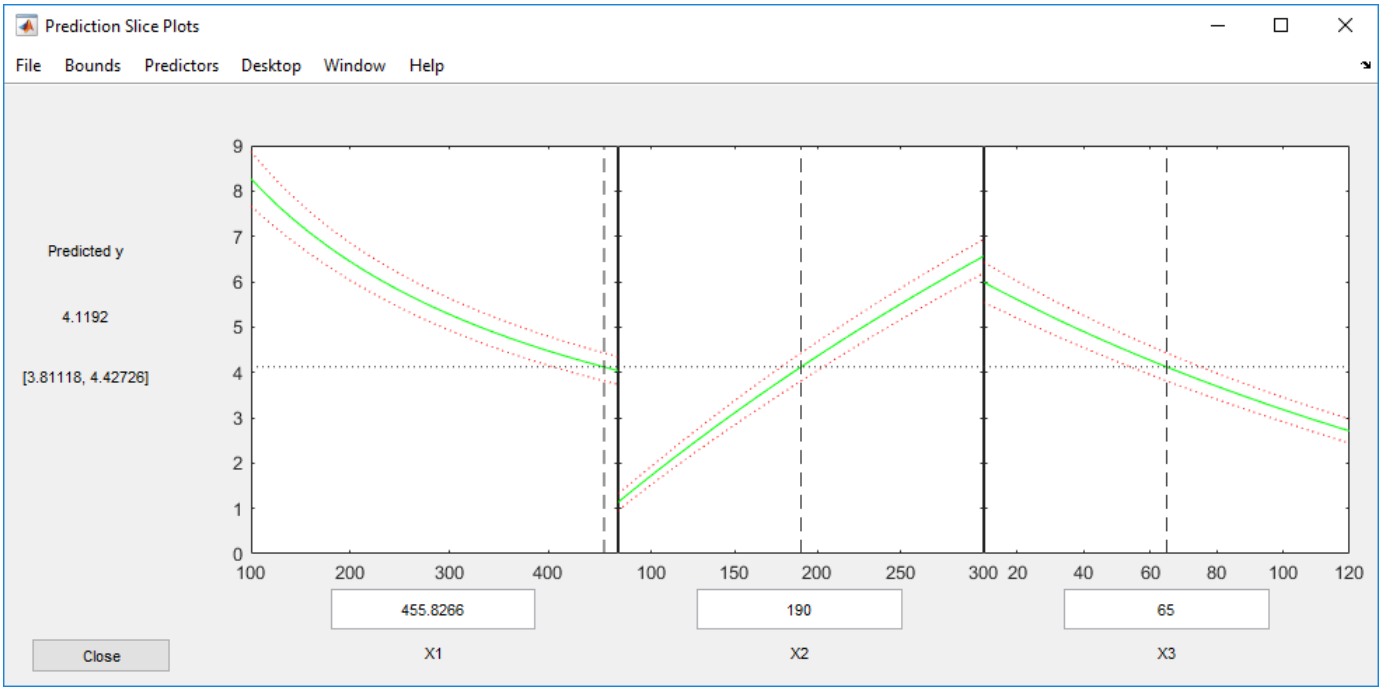
```
load reaction
mdl = fitnlm(reactants,rate,@hougen,[1 .05 .02 .1 2]);
```

Create a slice plot.

```
plotSlice(mdl)
```



Drag the X1 prediction line to the right, and observe the change in the predicted response y and in the predicted response curves to X2 and X3.



Tips

- If there are more than eight predictors, `plotSlice` selects the first five for plotting. Use the **Predictors** menu to control which predictors are plotted.
- The **Bounds** menu lets you choose between simultaneous or non-simultaneous bounds, and between bounds on the function or bounds on a new observation.

See Also

`NonLinearModel` | `predict`

Topics

“Examine Quality and Adjust the Fitted Nonlinear Model” on page 13-6
“Predict or Simulate Responses Using a Nonlinear Model” on page 13-9
“Nonlinear Regression Workflow” on page 13-12
“Nonlinear Regression” on page 13-2

plotSurvival

Plot survival function of Cox proportional hazards model

Syntax

```
plotSurvival(coxMdl)
plotSurvival(coxMdl,X)
plotSurvival(coxMdl,X,Stratification)
plotSurvival(coxMdl,ax, ___ )
plotSurvival( ___ ,'Time',T)
graphics = plotSurvival( ___ )
```

Description

`plotSurvival(coxMdl)` plots the baseline survival function of a Cox proportional hazards model `coxMdl`. The survival function at time `t` is the estimated probability of survival until time `t`. The term baseline refers to the survival function at the determined baseline of the predictors. This value is stored in `coxMdl.Baseline`, and the default value is the mean of the data set used for training.

`plotSurvival(coxMdl,X)` plots the survival function when the predictors have the values in `X`. The plot includes one line for each row of `X`.

`plotSurvival(coxMdl,X,Stratification)` plots the survival function for the given value of the stratification variable `Stratification`. You must have one row in `Stratification` for each row in `X`.

Note When you train `coxMdl` using stratification variables and pass predictor variables `X`, `plotSurvival` also requires you to pass stratification variables.

`plotSurvival(coxMdl,ax, ___ )` plots in the specified graphics axes `ax` using any of the input argument combinations in the previous syntaxes.

`plotSurvival( ___ ,'Time',T)` plots the survival function at times listed in `T`.

`graphics = plotSurvival( ___ )` returns an array of `Stair` graphics objects. See Stair Properties.

Examples

Plot Survival

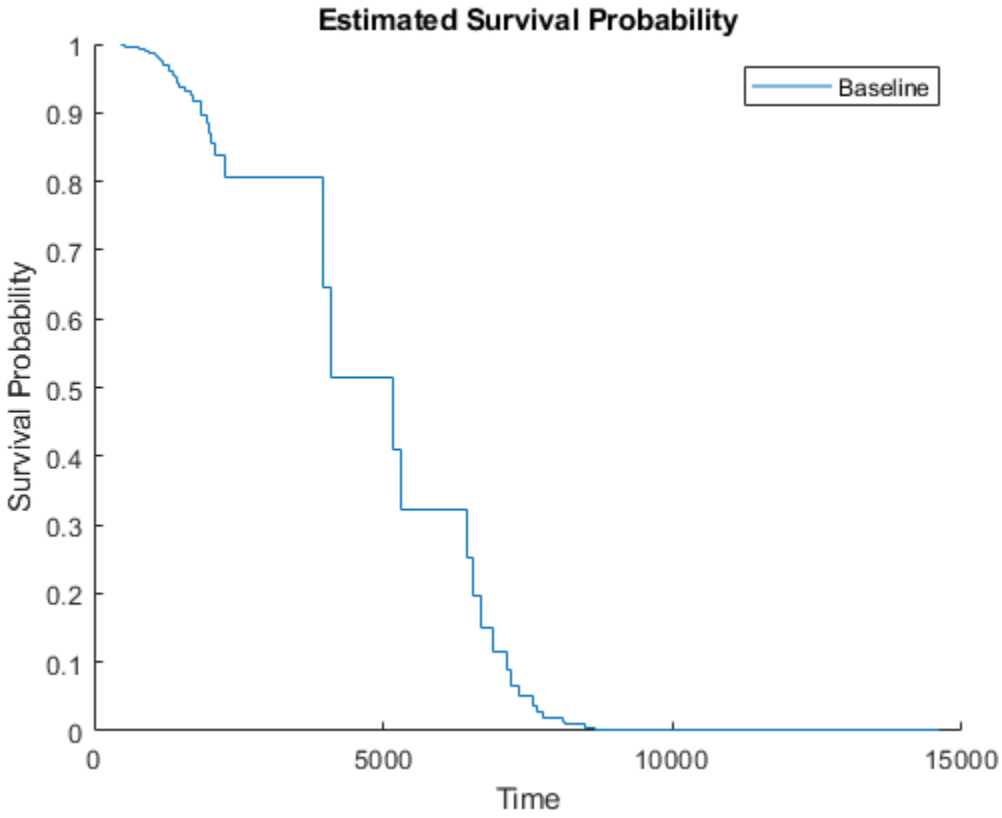
Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

```
load lightbulb
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3));
```

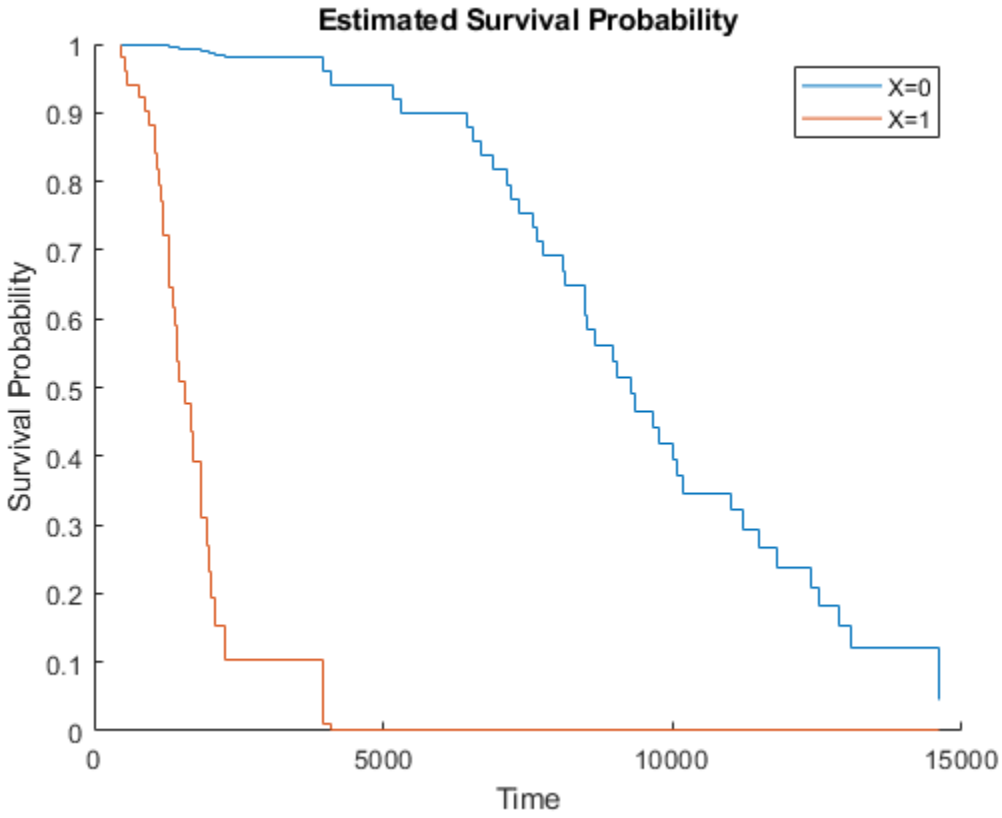
Plot the baseline survival function as a function of time `t`, meaning the probability that a lightbulb fails after time `t`. By default, the baseline is calculated for the mean of the predictor, which in this case is `mean(lightbulb(:,2)) = 0.5`.

```
plotSurvival(coxMdl)
```



Plot the survival for fluorescent bulbs (predictor = 0) and incandescent bulbs (predictor = 1).

```
plotSurvival(coxMdl,[0;1])
```



To calculate the survival without plotting, use `survival`.

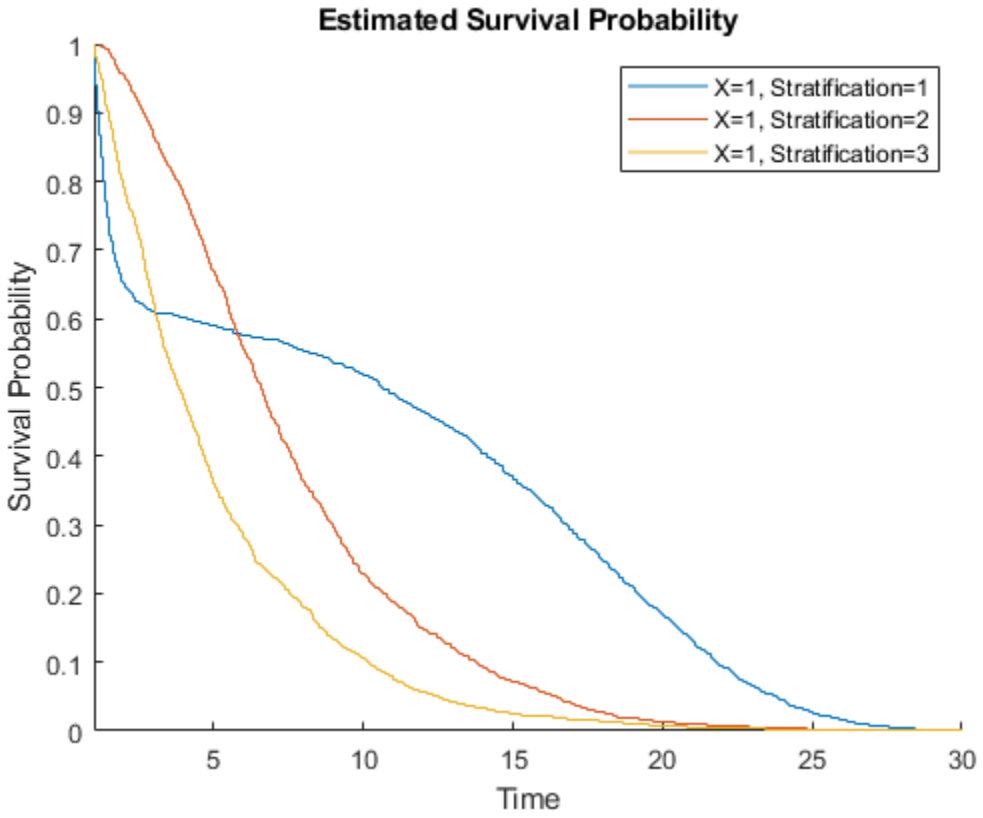
Survival Plot with Stratification Variables

Load the `coxModel` data. (This simulated data is generated in the example “Cox Proportional Hazards Model Object” on page 14-39.) The model named `coxMdl` has three stratification levels (1, 2, and 3) and a predictor `X` with three categorical values (1, 1/20, and 1/100).

```
load coxModel
```

Plot the survival for `X = 1` at the three stratification levels.

```
c1 = categorical(1);
X = [c1;c1;c1];
stratification = [1;2;3];
plotSurvival(coxMdl,X,stratification)
xlim([1,30])
```



Log-Scaled Survival Plot

Load the `coxModel` data. (This simulated data is generated in the example “Cox Proportional Hazards Model Object” on page 14-39.) The model named `coxMdl` has three stratification levels (1, 2, and 3) and a predictor X with three categorical values (1, 1/20, and 1/100).

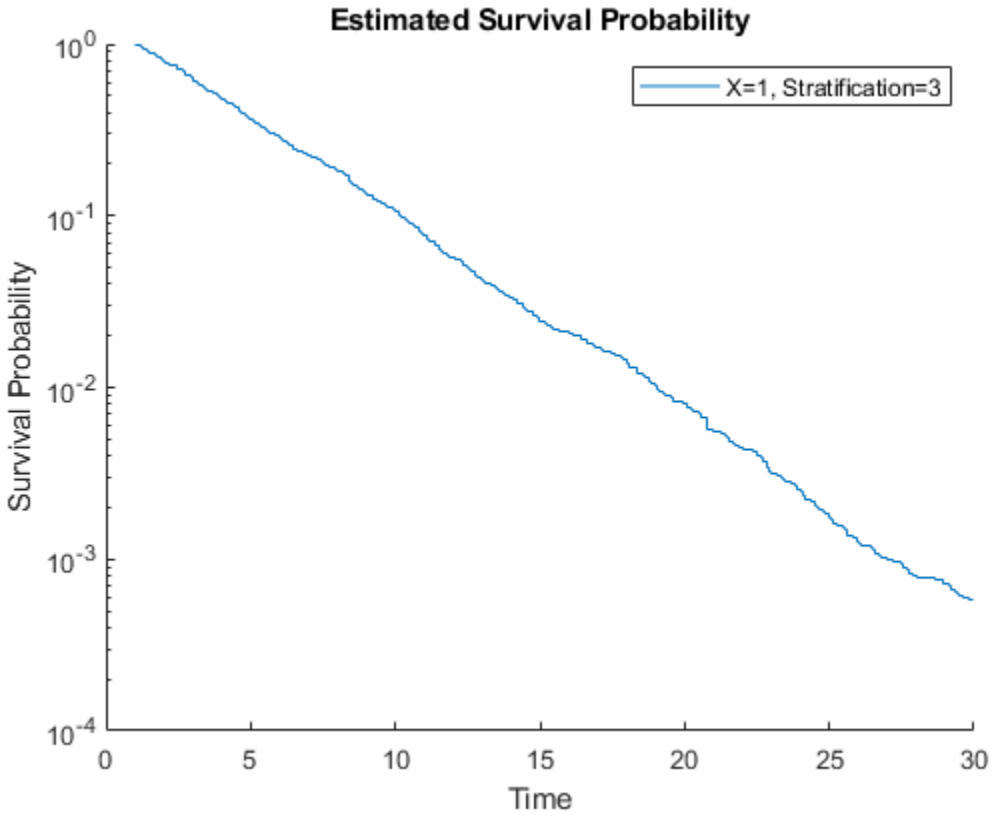
```
load coxModel
```

To enable programmatic editing of a survival plot, create an axes.

```
h = figure;
axes1 = axes('Parent',h);
```

Plot the survival function for the X predictor value `categorical(1)` and stratification level 3. This stratification level represents a constant hazard rate. When log-scaled, the resulting survival plot should, therefore, be close to a straight line. Plot for times 1 through 30.

```
oo = categorical(1);
plotSurvival(coxMdl,axes1,oo,3,'Time',linspace(1,30,300));
axes1.YScale = 'log';
```



Input Arguments

`coxMdl` — Fitted Cox proportional hazards model

`CoxModel` object

Fitted Cox proportional hazards model, specified as a `CoxModel` object. Create `coxMdl` using `fitcox`.

`X` — Predictors for model

mean of predictors used for training, but `0` for all categorical predictors (default) | array of predictors of type used for training

Predictors for the model, specified as an array of predictors of the same type used for training `coxMdl`. Each row of `X` represents one set of predictors.

Data Types: `double` | `table` | `categorical`

`Stratification` — Stratification level

variable or variables of type used for training

Stratification level, specified as a variable or variables of the same type used for training `coxMdl`. Specify the same number of rows in `Stratification` as in `X`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `table` | `cell` | `categorical`

ax — Axes for plotting
graphics axes object

Axes for plotting, specified as a graphics axes object.

T — Times for survival estimates
`coxMdl.Hazard(:,1)` (default) | real vector

Times for survival estimates, specified as a real vector. `plotSurvival` sorts the specified times and converts them to a column vector, if necessary. The resulting values are linearly interpolated from times in the training data.

Example: `0:40`

Data Types: `double`

See Also

`CoxModel` | `fitcox` | `hazardratio` | `survival`

Topics
“Cox Proportional Hazards Model Object” on page 14-39

Introduced in R2021a

plsregress

Partial least-squares (PLS) regression

Syntax

```
[XL,YL] = plsregress(X,Y,ncomp)
[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] = plsregress(X,Y,ncomp)
[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] = plsregress( ___ ,Name,Value)
```

Description

`[XL,YL] = plsregress(X,Y,ncomp)` returns the predictor and response loadings `XL` and `YL`, respectively, for a partial least-squares (PLS) regression of the responses in matrix `Y` on the predictors in matrix `X`, using `ncomp` PLS components.

`[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] = plsregress(X,Y,ncomp)` also returns:

- The predictor scores `XS`. Predictor scores are PLS components that are linear combinations of the variables in `X`.
- The response scores `YS`. Response scores are linear combinations of the responses with which the PLS components `XS` have maximum covariance.
- A matrix `BETA` of coefficient estimates for PLS regression. `plsregress` adds a column of ones in the matrix `X` to compute coefficient estimates for a model with constant terms (intercept).
- The percentage of variance `PCTVAR` explained by the regression model.
- The estimated mean squared errors `MSE` for PLS models with `ncomp` components.
- A structure `stats` that contains the PLS weights, T^2 statistic, and predictor and response residuals.

`[XL,YL,XS,YS,BETA,PCTVAR,MSE,stats] = plsregress( ___ ,Name,Value)` specifies options using one or more name-value arguments in addition to any of the input argument combinations in previous syntaxes. The name-value arguments specify `MSE` calculation parameters. For example, `'cv',5` calculates the `MSE` using 5-fold cross-validation.

Examples

Perform Partial Least-Squares Regression

Load the `spectra` data set. Create the predictor `X` as a numeric matrix that contains the near infrared (NIR) spectral intensities of 60 samples of gasoline at 401 wavelengths. Create the response `y` as a numeric vector that contains the corresponding octane ratings.

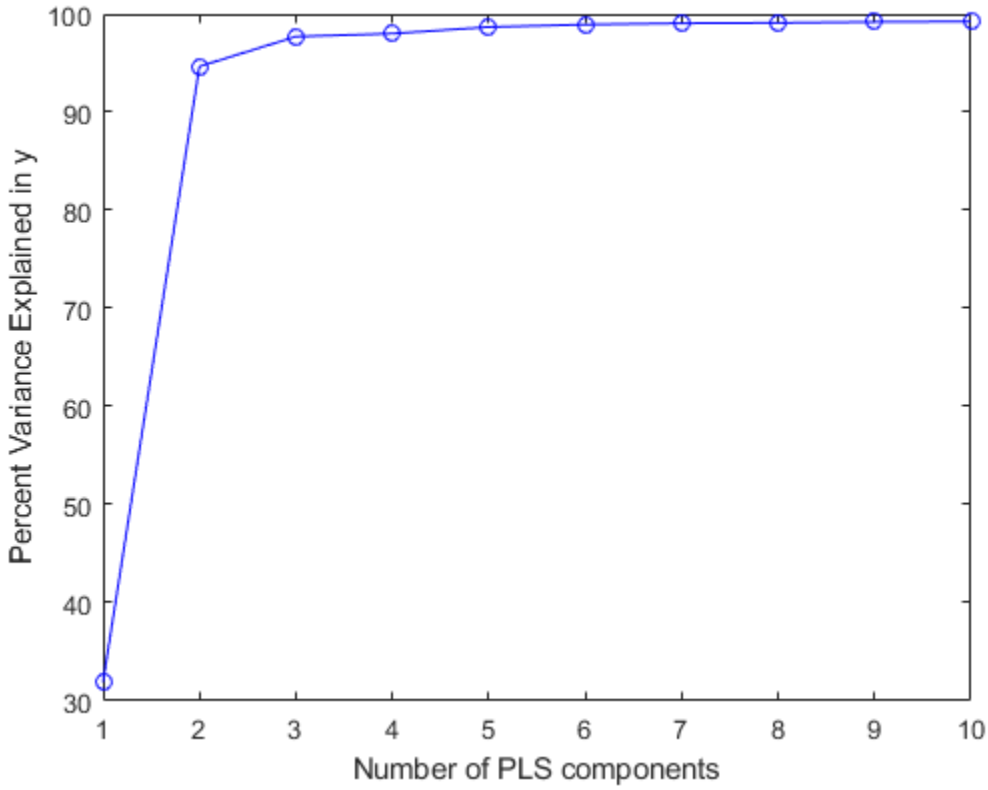
```
load spectra
X = NIR;
y = octane;
```

Perform PLS regression with 10 components of the responses in `y` on the predictors in `X`.

```
[XL,yl,XS,YS,beta,PCTVAR] = plsregress(X,y,10);
```

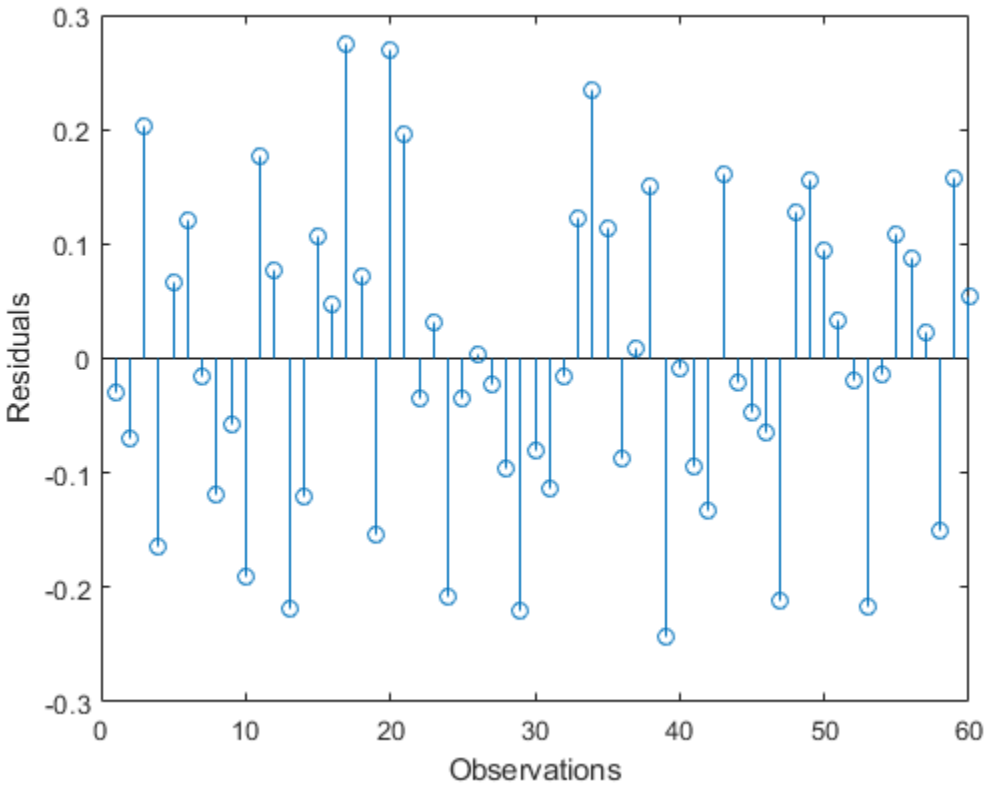
Plot the percent of variance explained in the response variable (PCTVAR) as a function of the number of components.

```
plot(1:10,cumsum(100*PCTVAR(2,:)),'-bo');
xlabel('Number of PLS components');
ylabel('Percent Variance Explained in y');
```



Compute the fitted response and display the residuals.

```
yfit = [ones(size(X,1),1) X]*beta;
residuals = y - yfit;
stem(residuals)
xlabel('Observations');
ylabel('Residuals');
```



Calculate Variable Importance in Projection for PLS Regression

Calculate variable importance in projection (VIP) scores for a partial least-squares (PLS) regression model. You can use VIP to select predictor variables when multicollinearity exists among variables. Variables with a VIP score greater than 1 are considered important for the projection of the PLS regression model [3].

Load the `spectra` data set. Create the predictor `X` as a numeric matrix that contains the near infrared (NIR) spectral intensities of 60 samples of gasoline at 401 wavelengths. Create the response `y` as a numeric vector that contains the corresponding octane ratings. Specify the number of components `ncomp`.

```
load spectra
X = NIR;
y = octane;
ncomp = 10;
```

Perform PLS regression with 10 components of the responses in `y` on the predictors in `X`.

```
[XL,yl,XS,YS,beta,PCTVAR,MSE,stats] = plsregress(X,y,ncomp);
```

Calculate the normalized PLS weights.

```
W0 = stats.W ./ sqrt(sum(stats.W.^2,1));
```

Calculate the VIP scores for `ncomp` components.

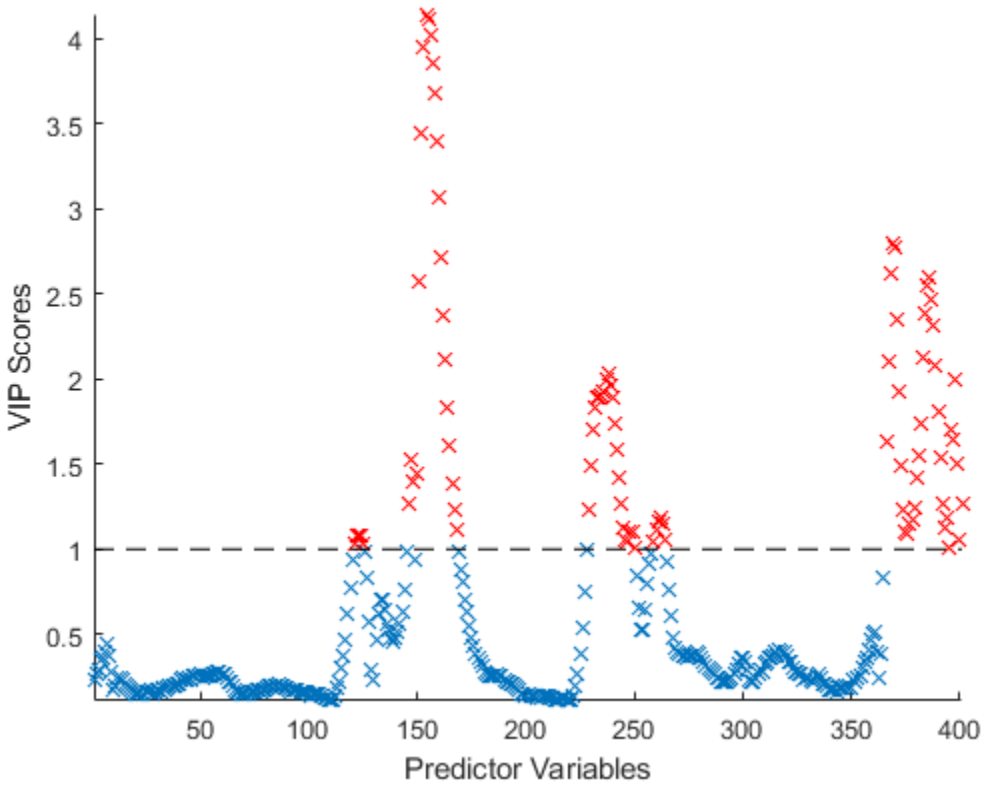
```
p = size(XL,1);
sumSq = sum(XS.^2,1).*sum(yl.^2,1);
vipScore = sqrt(p* sum(sumSq.*(W0.^2),2) ./ sum(sumSq,2));
```

Find variables with a VIP score greater than or equal to 1.

```
indVIP = find(vipScore >= 1);
```

Plot the VIP scores.

```
scatter(1:length(vipScore),vipScore,'x')
hold on
scatter(indVIP,vipScore(indVIP),'rx')
plot([1 length(vipScore)],[1 1],'--k')
hold off
axis tight
xlabel('Predictor Variables')
ylabel('VIP Scores')
```



Input Arguments

X — Predictor variables

numeric matrix

Predictor variables, specified as a numeric matrix. X is an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each row of X represents one observation, and each column represents one variable. X must have the same number of rows as Y.

Data Types: `single` | `double`

Y — Response variables

numeric matrix

Response variables, specified as a numeric matrix. Y is an *n*-by-*m* matrix, where *n* is the number of observations and *m* is the number of response variables. Each row of Y represents one observation, and each column represents one variable. Each row in Y is the response for the corresponding row in X.

Data Types: `single` | `double`

`ncomp` — Number of components

numeric vector

Number of components, specified as a numeric vector. If you do not specify `ncomp`, the default value is `min(size(X,1) − 1,size(X,2))`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'cv',10,'Options',statset('UseParallel',true)` calculates the MSE using 10-fold cross-validation, where computations run in parallel.

`cv` — MSE calculation method

`'resubstitution'` (default) | positive integer | `cvpartition` object

MSE calculation method, specified as `'resubstitution'`, a positive integer, or a `cvpartition` object.

- Specify `'cv'` as `'resubstitution'` to use both X and Y to fit the model and estimate the mean squared errors, without cross-validation.
- Specify `'cv'` as a positive integer k to use k-fold cross-validation.
- Specify `'cv'` as a `cvpartition` object to specify another type of cross-validation partition.

Example: `'cv',5`

Example: `'cv',cvpartition(n,'Holdout',0.3)`

Data Types: `single` | `double` | `char` | `string`

`mcreps` — Number of Monte Carlo repetitions

1 (default) | positive integer

Number of Monte Carlo repetitions for cross-validation, specified as a positive integer. If you specify `'cv'` as `'resubstitution'`, then `'mcreps'` must be 1.

Example: `'mcreps',5`

Data Types: `single` | `double`

`Options` — Options for running in parallel and setting random streams
structure

Options for running computations in parallel and setting random streams, specified as a structure. Create the `Options` structure with `statset`. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to run computations in parallel.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or a cell array consisting of one such object.	If you do not specify `Streams`, then `plsregress` uses the default stream.

Note You need Parallel Computing Toolbox to run computations in parallel.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

Output Arguments

`XL` — Predictor loadings
numeric matrix

Predictor loadings, returned as a numeric matrix. `XL` is a *p*-by-`ncomp` matrix, where *p* is the number of predictor variables and `ncomp` is the number of PLS components. Each row of `XL` contains coefficients that define a linear combination of PLS components approximating the original predictor variables.

Data Types: `single` | `double`

`YL` — Response loadings
numeric matrix

Response loadings, returned as a numeric matrix. `YL` is an *m*-by-`ncomp` matrix, where *m* is the number of response variables and `ncomp` is the number of PLS components. Each row of `YL` contains coefficients that define a linear combination of PLS components approximating the original response variables.

Data Types: `single` | `double`

XS — Predictor scores

numeric matrix

Predictor scores, returned as a numeric matrix. XS is an *n*-by-ncomp orthonormal matrix, where *n* is the number of observations and ncomp is the number of PLS components. Each row of XS corresponds to one observation, and each column corresponds to one component.

Data Types: single | double

YS — Response scores

numeric matrix

Response scores, returned as a numeric matrix. YS is an *n*-by-ncomp matrix, where *n* is the number of observations and ncomp is the number of PLS components. Each row of YS corresponds to one observation, and each column corresponds to one component. YS is not orthogonal or normalized.

Data Types: single | double

BETA — Coefficient estimates for PLS regression

numeric matrix

Coefficient estimates for PLS regression, returned as a numeric matrix. BETA is a (p + 1)-by-m matrix, where p is the number of predictor variables and m is the number of response variables. The first row of BETA contains coefficient estimates for the constant terms.

Data Types: single | double

PCTVAR — Percentage of variance

numeric matrix

Percentage of variance explained by the model, returned as a numeric matrix. PCTVAR is a 2-by-ncomp matrix, where ncomp is the number of PLS components. The first row of PCTVAR contains the percentage of variance explained in X by each PLS component, and the second row contains the percentage of variance explained in Y.

Data Types: single | double

MSE — Mean squared error

numeric matrix

Mean squared error, returned as a numeric matrix. MSE is a 2-by-(ncomp + 1) matrix, where ncomp is the number of PLS components. MSE contains the estimated mean squared errors for a PLS model with ncomp components. The first row of MSE contains mean squared errors for the predictor variables in X, and the second row contains mean squared errors for the response variables in Y. The column *j* of MSE contains mean squared errors for *j – 1* components.

Data Types: single | double

stats — Model statistics

structure

Model statistics, returned as a structure with the fields described in this table.

Field	Description
W	*p*-by-ncomp matrix of PLS weights so that XS = X0*W

Field	Description
T2	T^2 statistic for each point in XS
Xresiduals	Predictor residuals, X0 – XS*XL'
Yresiduals	Response residuals, Y0 – XS*YL'

For more information about the centered predictor and response variables X0 and Y0, see “Algorithms” on page 33-4728.

Algorithms

plsregress uses the SIMPLS algorithm [1]. The function first centers X and Y by subtracting the column means to get the centered predictor and response variables X0 and Y0, respectively. However, the function does not rescale the columns. To perform PLS regression with standardized variables, use zscore to normalize X and Y (columns of X0 and Y0 are centered to have mean 0 and scaled to have standard deviation 1).

After centering X and Y, plsregress computes the singular value decomposition (SVD) on X0'*Y0. The predictor and response loadings XL and YL are the coefficients obtained from regressing X0 and Y0 on the predictor score XS. You can reconstruct the centered data X0 and Y0 using XS*XL' and XS*YL', respectively.

plsregress initially computes YS as YS = Y0*YL. By convention [1], however, plsregress then orthogonalizes each column of YS with respect to preceding columns of XS, so that XS'*YS is a lower triangular matrix.

References

[1] de Jong, Sijmen. “SIMPLS: An Alternative Approach to Partial Least Squares Regression.” *Chemometrics and Intelligent Laboratory Systems* 18, no. 3 (March 1993): 251–63. https://doi.org/10.1016/0169-7439(93)85002-X.

[2] Rosipal, Roman, and Nicole Kramer. "Overview and Recent Advances in Partial Least Squares." *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS 2005), Revised Selected Papers (Lecture Notes in Computer Science 3940)*. Berlin, Germany: Springer-Verlag, 2006, vol. 3940, pp. 34–51. https://doi.org/10.1007/11752790_2.

[3] Chong, Il-Gyo, and Chi-Hyuck Jun. “Performance of Some Variable Selection Methods When Multicollinearity Is Present.” *Chemometrics and Intelligent Laboratory Systems* 78, no. 1–2 (July 2005) 103–12. https://doi.org/10.1016/j.chemolab.2004.12.011.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the 'Options' name-value argument in the call to this function and set the 'UseParallel' field of the options structure to true using statset.

For example: 'Options',statset('UseParallel',true)

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`pca` | `regress` | `sequentialfs`

Topics

“Partial Least Squares” on page 11-126
“Partial Least Squares Regression and Principal Components Regression” on page 11-188

Introduced in R2008a

PointSet property

Class: `qrandstream`

Point set from which stream is drawn

Description

The `PointSet` property contains a copy of the point set from which the stream is providing points. The point set is specified during construction of a quasi-random stream and cannot subsequently be altered.

Examples

```
Q = qrandstream('sobol', 5, 'Skip', 8);
% Create a new stream based on the same sequence as that in Q
Q2 = qrandstream(Q.PointSet);
u1 = qrand(Q, 10)
u2 = qrand(Q2, 10) % contains exactly the same values as u1
```

poisscdf

Poisson cumulative distribution function

Syntax

```
y = poisscdf(x,lambda)
y = poisscdf(x,lambda,'upper')
```

Description

`y = poisscdf(x,lambda)` computes the Poisson cumulative distribution function at each of the values in `x` using the rate parameters in `lambda`.

`x` and `lambda` can be scalars, vectors, matrices, or multidimensional arrays that all have the same size. If only one argument is a scalar, `poisscdf` expands it to a constant array with the same dimensions as the other argument.

`y = poisscdf(x,lambda,'upper')` returns the complement of the Poisson cumulative distribution function at each value in `x`, using an algorithm that computes the extreme upper tail probabilities more accurately.

Examples

Compute and Plot Poisson Cumulative Distribution Function

Compute and plot the Poisson cumulative distribution function for the specified range of integer values and average rate.

A computer hard disk manufacturing facility performs random tests of individual hard disks. The policy is to shut down the manufacturing process if an inspector finds more than four bad sectors on a disk. Assuming that on average a disk has two bad sectors, find the probability of a manufacturing process shutdown after the first inspection.

```
1 - poisscdf(4,2)
```

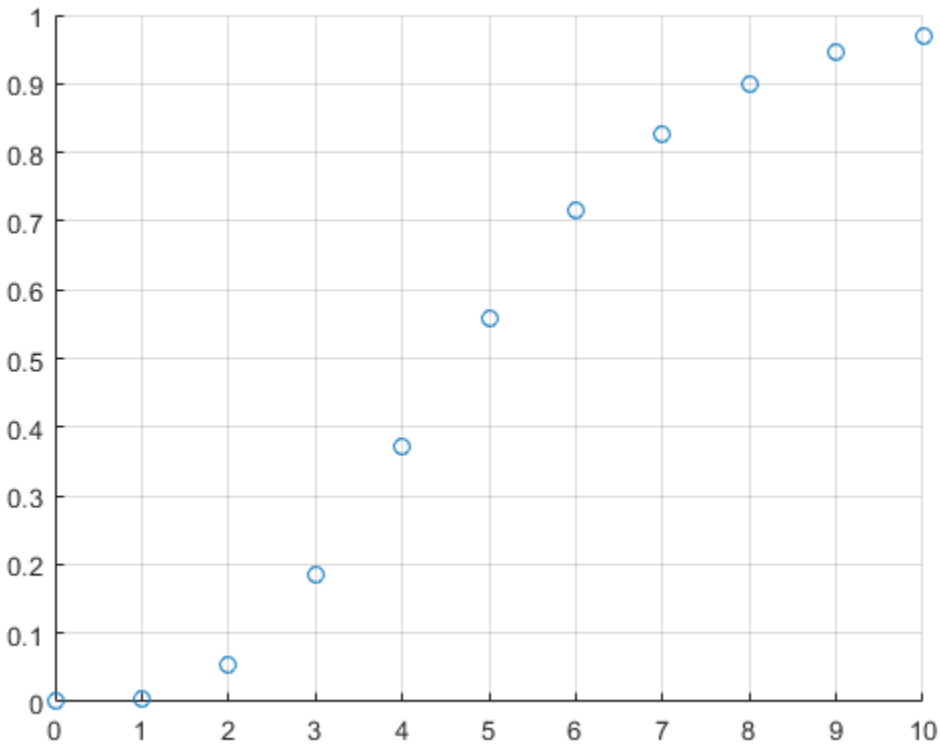
```
ans = 0.0527
```

Compute the probabilities a manufacturing process shutdown after the first inspection if on average a disk has 0, 1, 2, ..., 10 bad sectors.

```
lambda = 0:10;
y = 1 - poisscdf(4,lambda);
```

Plot the results.

```
scatter(lambda,y,'Marker',"o")
grid on
```



Compute Extreme Upper Tail Probabilities

Compute the complement of the Poisson cumulative distribution function with more accurate upper tail probabilities.

A computer hard disk manufacturing facility performs random tests of individual hard disks. Assuming that on average a disk has 10 bad sectors, find the probability that a disk has more than 100 bad sectors.

```
format long
1 - poisscdf(100,10)
```

```
ans =
     0
```

This result shows that `poisscdf(100,10)` is so close to 1 (within `eps`) that subtracting it from 1 gives 0. To approximate the extreme upper tail probabilities better, compute the complement of the Poisson cumulative distribution function directly instead of computing the difference.

```
poisscdf(100,10,'upper')
```

```
ans =
     5.339405460719755e-64
```

Input Arguments

x — Values at which to evaluate Poisson cdf

scalar value | array of scalar values

Values at which to evaluate the Poisson cdf, specified as a scalar value or array of scalar values.

Example: `[0,1,3,4]`

Data Types: `single` | `double`

lambda — Rate parameters

positive value | array of positive values

Rate parameters, specified as a positive value or array of positive values. The rate parameter indicates the average number of events in a given time interval.

Example: 2

Data Types: `single` | `double`

Output Arguments

y — Poisson cdf values

scalar value | array of scalar values

Poisson cdf values, returned as a scalar value or array of scalar values. Each element in `y` is the Poisson cdf value of the distribution evaluated at the corresponding element in `x`.

More About

Poisson Cumulative Distribution Function

The Poisson cumulative distribution function lets you obtain the probability of an event occurring within a given time or space interval less than or equal to x times if on average the event occurs λ times within that interval.

The Poisson cumulative distribution function for the given values x and λ is

$$p = F(x|\lambda) = e^{-\lambda}\sum_{i=0}^{floor(x)}\frac{\lambda^i}{i!}.$$

Alternative Functionality

- `poisscdf` is a function specific to Poisson distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, specify the probability distribution name and its parameters. Alternatively, create a `PoissonDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `poisscdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`PoissonDistribution` | `cdf` | `poissfit` | `poissinv` | `poisspdf` | `poissrnd` | `poisstat`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

poissfit

Poisson parameter estimates

Syntax

```
lambdahat = poissfit(data)
[lambdahat,lambdaci] = poissfit(data)
[lambdahat,lambdaci] = poissfit(data,alpha)
```

Description

`lambdahat = poissfit(data)` returns the maximum likelihood estimate (MLE) of the parameter of the Poisson distribution, λ, given the data `data`.

`[lambdahat,lambdaci] = poissfit(data)` also gives 95% confidence intervals in `lambdaci`.

`[lambdahat,lambdaci] = poissfit(data,alpha)` gives `100(1 - alpha)`% confidence intervals. For example `alpha = 0.001` yields 99.9% confidence intervals.

The sample mean is the MLE of λ.

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Examples

```
r = poissrnd(5,10,2);
[l,lci] = poissfit(r)
l =
    7.4000    6.3000
lci =
    5.8000    4.8000
    9.1000    7.9000
```

See Also

`mle` | `poisscdf` | `poissinv` | `poisspdf` | `poissrnd` | `poisstat`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

poissinv

Poisson inverse cumulative distribution function

Syntax

```
X = poissinv(P,lambda)
```

Description

`X = poissinv(P,lambda)` returns the smallest value `X` such that the Poisson cdf evaluated at `X` equals or exceeds `P`, using mean parameters in `lambda`. `P` and `lambda` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other input.

Examples

If the average number of defects (λ) is two, what is the 95th percentile of the number of defects?

```
poissinv(0.95,2)
ans =
   5
```

What is the median number of defects?

```
median_defects = poissinv(0.50,2)
median_defects =
   2
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`icdf` | `poisscdf` | `poissfit` | `poisspdf` | `poissrnd` | `poisstat`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

poisspdf

Poisson probability density function

Syntax

```
y = poisspdf(x,lambda)
```

Description

`y = poisspdf(x,lambda)` computes the Poisson probability density function at each of the values in `x` using the rate parameters in `lambda`.

`x` and `lambda` can be scalars, vectors, matrices, or multidimensional arrays that all have the same size. If only one argument is a scalar, `poisspdf` expands it to a constant array with the same dimensions as the other argument.

Examples

Compute and Plot Poisson Probability Density Function

Compute and plot the Poisson probability density function for the specified range of integer values and average rate.

In the computer hard disk manufacturing process, flaws occur randomly. Assuming that on average a 4 GB hard disk has two flaws, compute the probability that a disk has no flaws.

```
poisspdf(0,2)
```

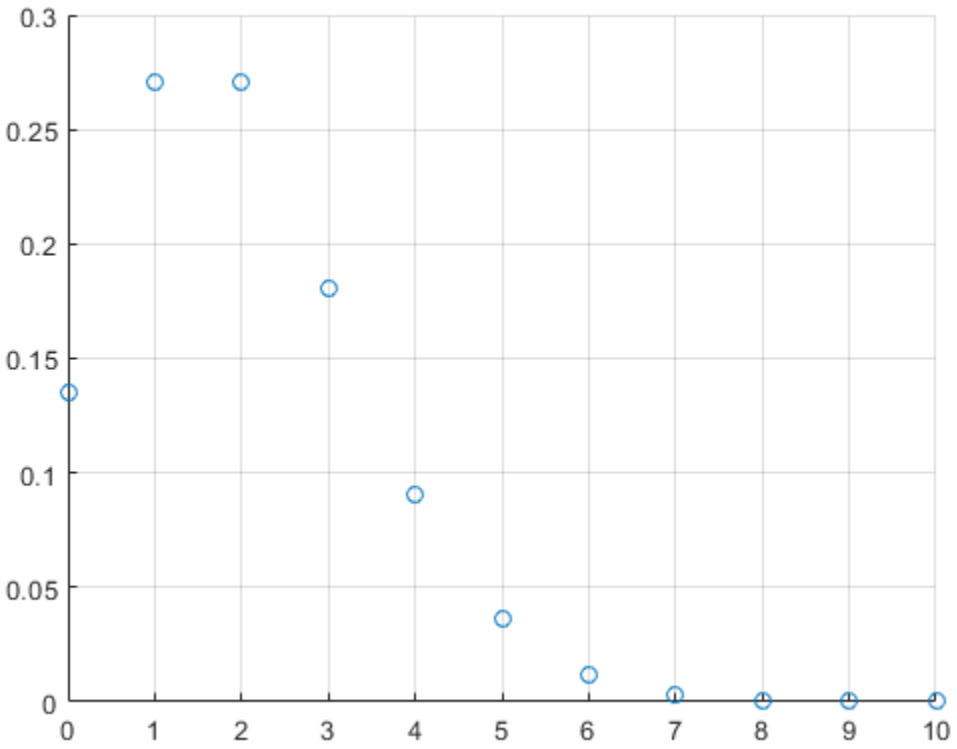
```
ans = 0.1353
```

Compute the Poisson probability density function values at each value from 0 to 10. These values correspond to the probabilities that a disk has 0, 1, 2, ..., 10 flaws.

```
flaws = 0:10;
y = poisspdf(flaws,2);
```

Plot the resulting probability values.

```
scatter(flaws,y,'Marker',"o")
grid on
```



Input Arguments

x — Values at which to evaluate Poisson pdf

scalar value | array of scalar values

Values at which to evaluate the Poisson pdf, specified as a scalar value or array of scalar values. For noninteger values `x`, the Poisson probability density function is zero.

Example: `[0,1,3,4]`

Data Types: `single` | `double`

lambda — Rate parameters

positive value | array of positive values

Rate parameters, specified as a positive value or array of positive values. The rate parameter indicates the average number of events in a given time interval.

Example: 2

Data Types: `single` | `double`

Output Arguments

y — Poisson pdf values

scalar value | array of scalar values

Poisson pdf values, returned as a scalar value or array of scalar values. Each element in `y` is the Poisson pdf value of the distribution evaluated at the corresponding element in `x`.

Data Types: `single` | `double`

More About

Poisson Probability Density Function

The Poisson probability density function lets you obtain the probability of an event occurring within a given time or space interval exactly x times if on average the event occurs λ times within that interval.

The Poisson probability density function for the given values x and λ is

$$f(x\left|\lambda\right.) = \frac{\lambda^x}{x!}e^{-\lambda};\ x = 0, 1, 2, \ldots, \infty\ .$$

Alternative Functionality

- `poisspdf` is a function specific to Poisson distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, specify the probability distribution name and its parameters. Alternatively, create a `PoissonDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `poisspdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`PoissonDistribution` | `pdf` | `poisscdf` | `poissfit` | `poissinv` | `poissrnd` | `poisstat`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

poissrnd

Random numbers from Poisson distribution

Syntax

```
r = poissrnd(lambda)
r = poissrnd(lambda,sz1,...,szN)
r = poissrnd(lambda,sz)
```

Description

`r = poissrnd(lambda)` generates random numbers from the Poisson distribution specified by the rate parameter `lambda`.

`lambda` can be a scalar, vector, matrix, or multidimensional array.

`r = poissrnd(lambda,sz1,...,szN)` generates an array of random numbers from the Poisson distribution with the scalar rate parameter `lambda`, where `sz1,...,szN` indicates the size of each dimension.

`r = poissrnd(lambda,sz)` generates an array of random numbers from the Poisson distribution with the scalar rate parameter `lambda`, where vector `sz` specifies `size(r)`.

Examples

Array of Random Numbers from Several Poisson Distributions

Generate an array of random numbers from the Poisson distributions. Specify the average rate for each distribution.

```
lambda = 10:2:20
```

```
lambda = 1×6

    10    12    14    16    18    20
```

Generate random numbers from the Poisson distributions.

```
r = poissrnd(lambda)
```

```
r = 1×6

    14    13    14     9    14    31
```

Array of Random Numbers from One Poisson Distribution

Generate an array of random numbers from one Poisson distribution. Here, the distribution parameter `lambda` is a scalar.

Use the `poissrnd` function to generate random numbers from the Poisson distribution with the average rate 20. The function returns one number.

```
r_scalar = poissrnd(20)
```

```
r_scalar = 9
```

Generate a 2-by-3 array of random numbers from the same distribution by specifying the required array dimensions.

```
r_array = poissrnd(20,2,3)
```

```
r_array = 2×3

    13    14    18
    26    16    21
```

Alternatively, specify the required array dimensions as a vector.

```
r_array = poissrnd(20,[2 3])
```

```
r_array = 2×3

    22    27    22
    25    19    21
```

Input Arguments

`lambda` — Rate parameters

positive value | array of positive values

Rate parameters, specified as a positive value or array of positive values. The rate parameter indicates the average number of events in a given time interval.

Example: 2

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)

integers

Size of each dimension, specified as separate arguments of integers. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the Poisson probability distribution.

If `lambda` is an array, then the specified dimensions `sz1,...,szN` must match the dimensions of `lambda`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.

- Beyond the second dimension, `poissrnd` ignores trailing dimensions with a size of 1. For example, `poissrnd(5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the Poisson distribution with rate parameter 5.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5 3 2]` generates a 5-by-3-by-2 array of random numbers from the Poisson probability distribution.

If `lambda` is an array, then the specified dimensions `sz` must match the dimensions of `lambda`.

- If you specify a single value `[sz1]`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `poissrnd` ignores trailing dimensions with a size of 1. For example, `poissrnd(5,[3,1,1,1])` produces a 3-by-1 vector of random numbers from the Poisson distribution with rate parameter 5.

Example: `[5 3 2]`

Data Types: `single` | `double`

Output Arguments

`r` — Random numbers from Poisson distribution
scalar value | array of scalar values

Random numbers from the Poisson distribution, returned as a scalar value or an array of scalar values.

Data Types: `single` | `double`

Alternative Functionality

- `poissrnd` is a function specific to Poisson distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Alternatively, create a `PoissonDistribution` probability distribution object and pass the object as an input argument. Note that the distribution-specific function `poissrnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`poisscdf` | `poissfit` | `poissinv` | `poisspdf` | `poisstat` | `random`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

poisstat

Poisson mean and variance

Syntax

```
M = poisstat(lambda)
[M,V] = poisstat(lambda)
```

Description

`M = poisstat(lambda)` returns the mean of the Poisson distribution using mean parameters in `lambda`. The size of `M` is the size of `lambda`.

`[M,V] = poisstat(lambda)` also returns the variance `V` of the Poisson distribution.

For the Poisson distribution with parameter λ, both the mean and variance are equal to λ.

Examples

Find the mean and variance for the Poisson distribution with λ = 2.

```
[m,v] = poisstat([1 2; 3 4])
m =
   1   2
   3   4
v =
   1   2
   3   4
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`poisscdf` | `poissfit` | `poissinv` | `poisspdf` | `poissrnd`

Topics

“Poisson Distribution” on page B-131

Introduced before R2006a

polyconf

Polynomial confidence intervals

Syntax

```
Y = polyconf(p,X)
[Y,DELTA] = polyconf(p,X,S)
[Y,DELTA] = polyconf(p,X,S,param1,val1,param2,val2,...)
```

Description

`Y = polyconf(p,X)` evaluates the polynomial `p` at the values in `X`. `p` is a vector of coefficients in descending powers.

`[Y,DELTA] = polyconf(p,X,S)` takes outputs `p` and `S` from `polyfit` and generates 95% prediction intervals `Y ± DELTA` for new observations at the values in `X`.

`[Y,DELTA] = polyconf(p,X,S,param1,val1,param2,val2,...)` specifies optional parameter name/value pairs chosen from the following list.

Parameter	Value
`'alpha'`	A value between 0 and 1 specifying a confidence level of `100*(1-alpha)`%. The default is `0.05`.
`'mu'`	A two-element vector containing centering and scaling parameters. With this option, `polyconf` uses `(X-mu(1))/mu(2)` in place of `X`.
`'predopt'`	Either `'observation'` (the default) to compute prediction intervals for new observations at the values in `X`, or `'curve'` to compute confidence intervals for the fit evaluated at the values in `X`. See below.
`'simopt'`	Either `'off'` (the default) for nonsimultaneous bounds, or `'on'` for simultaneous bounds. See below.

The `'predopt'` and `'simopt'` parameters can be understood in terms of the following functions:

- $p(x)$ — the unknown mean function estimated by the fit
- $l(x)$ — the lower confidence bound
- $u(x)$ — the upper confidence bound

Suppose you make a new observation y_{n+1} at x_{n+1}, so that

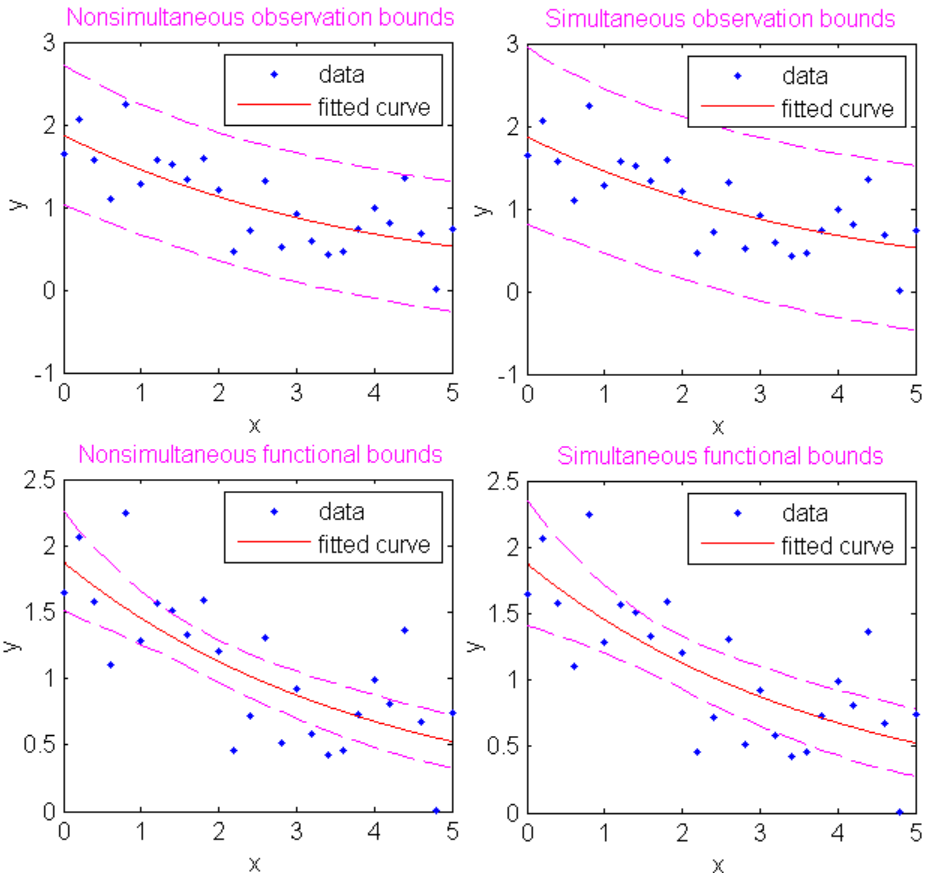
$$y_{n+1}(x_{n+1}) = p(x_{n+1}) + \varepsilon_{n+1}$$

By default, the interval $[l_{n+1}(x_{n+1}), u_{n+1}(x_{n+1})]$ is a 95% confidence bound on $y_{n+1}(x_{n+1})$.

The following combinations of the `'predopt'` and `'simopt'` parameters allow you to specify other bounds.

'simopt'	'predopt'	Bounded Quantity
'off'	'observation'	$y_{n+1}(x_{n+1})$ (default)
'off'	'curve'	$p(x_{n+1})$
'on'	'observation'	$y_{n+1}(x)$, for all x
'on'	'curve'	$p(x)$, for all x

In general, 'observation' intervals are wider than 'curve' intervals, because of the additional uncertainty of predicting a new response value (the curve plus random errors). Likewise, simultaneous intervals are wider than nonsimultaneous intervals, because of the additional uncertainty of bounding values for all predictors x.



Examples

Plot Polynomial Fit and Prediction Intervals

Fit a polynomial to a sample data set, and estimate the 95% prediction intervals and the roots of the fitted polynomial. Plot the data and the estimations, and display the fitted polynomial expression using the helper function polystr, whose code appears at the end of this example.

Generate sample data points (x,y) with a quadratic trend.

```
rng('default') % For reproducibility
x = -5:5;
y = x.^2 - 20*x - 3 + 5*randn(size(x));
```

Fit a second degree polynomial to the data by using `polyfit`.

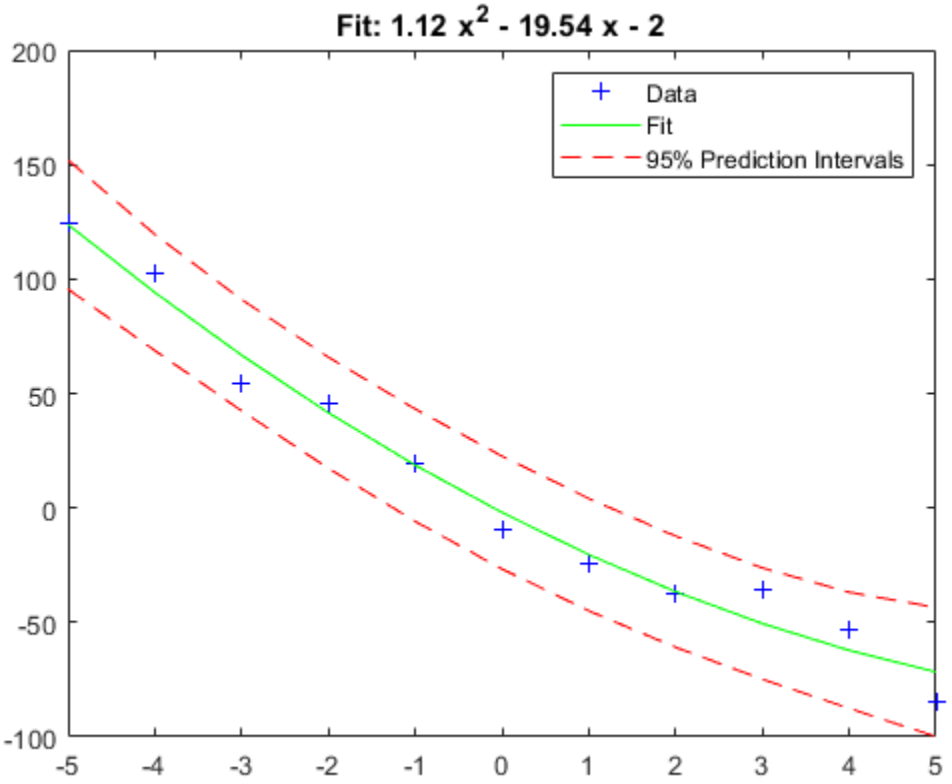
```
degree = 2; % Degree of the fit
[p,S] = polyfit(x,y,degree);
```

Estimate the 95% prediction intervals by using `polyconf`.

```
alpha = 0.05; % Significance level
[yfit,delta] = polyconf(p,x,S,'alpha',alpha);
```

Plot the data, fitted polynomial, and prediction intervals. Display the fitted polynomial expression using the helper function `polystr`.

```
plot(x,y,'b+')
hold on
plot(x,yfit,'g-')
plot(x,yfit-delta,'r--',x,yfit+delta,'r--')
legend('Data','Fit','95% Prediction Intervals')
title(['Fit: ',texlabel(polystr(round(p,2)))])
hold off
```



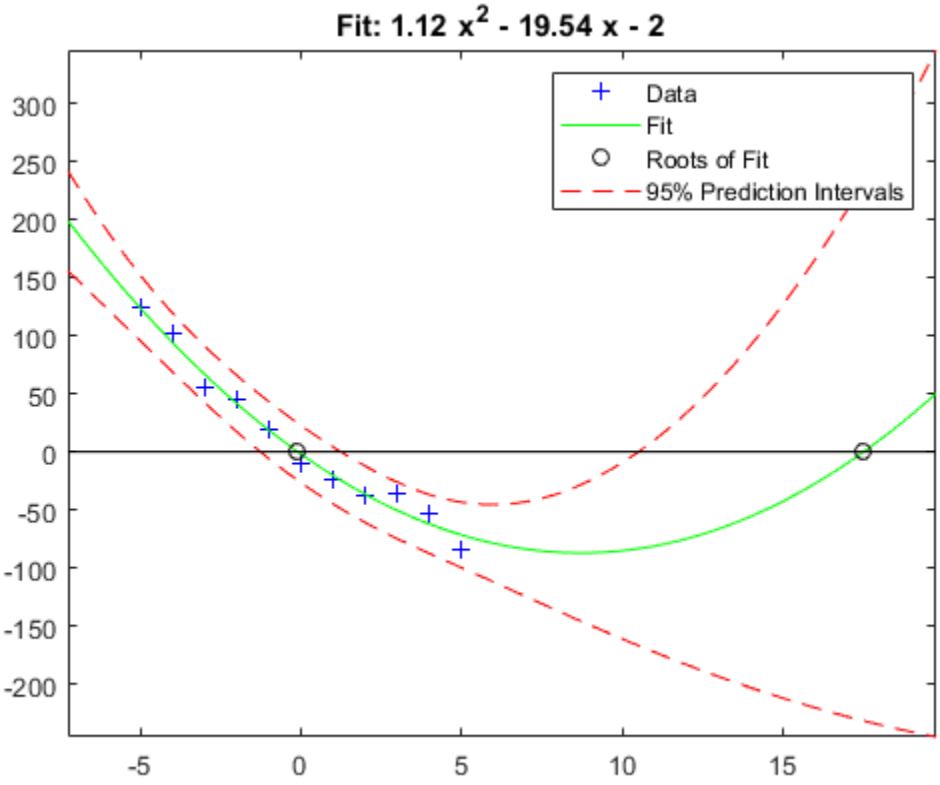
Find the roots of the polynomial `p`.

```
r = roots(p)
```

```
r = 2×1

   17.5152
   -0.1017
```

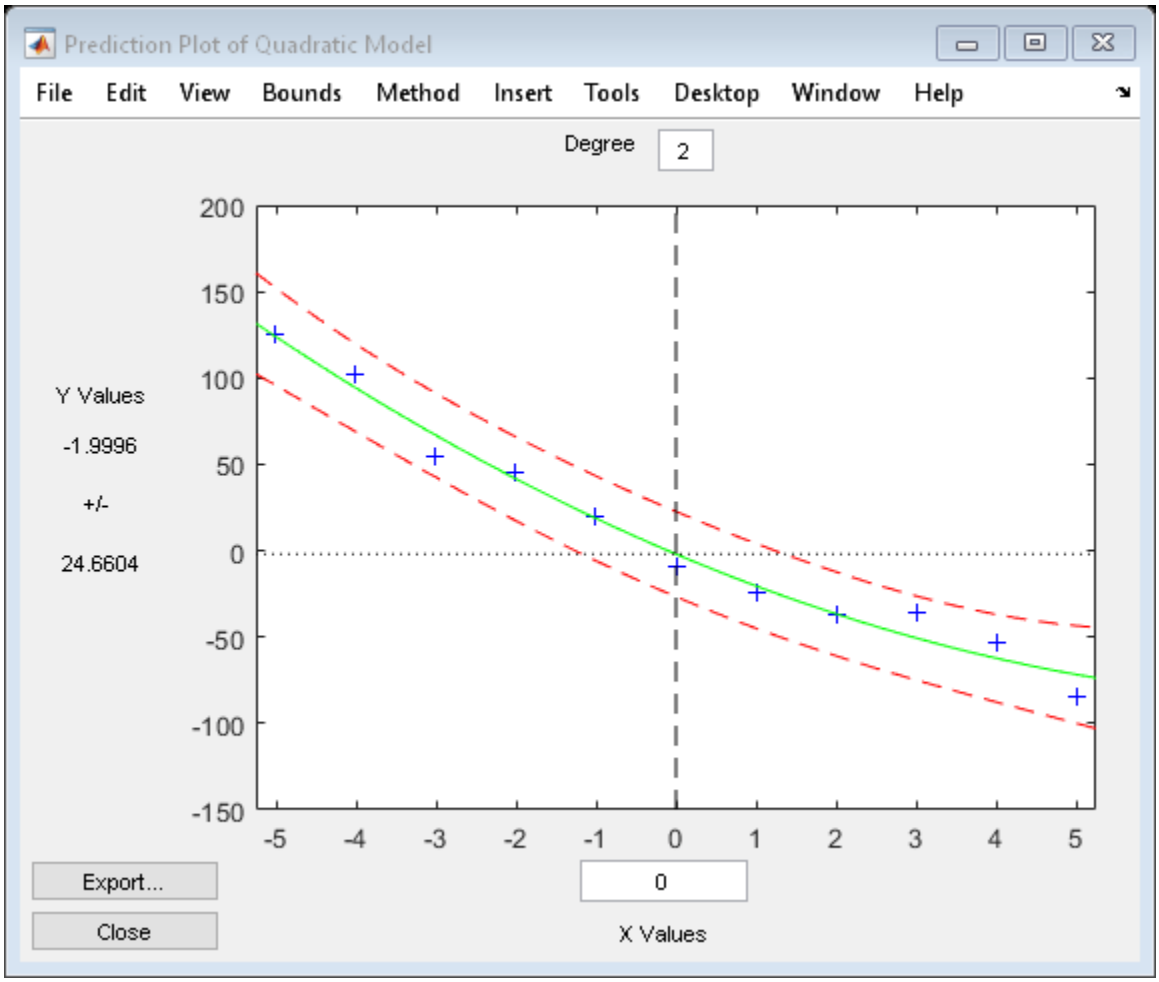
Because the roots are real values, you can plot them as well. Estimate the fitted values and prediction intervals for the *x* interval that includes the roots. Then, plot the roots and the estimations.

```
if isreal(r)
    xmin = min([r(:);x(:)]);
    xrange = range([r(:);x(:)]);
    xExtended = linspace(xmin - 0.1*xrange, xmin + 1.1*xrange,1000);
    [yfitExtended,deltaExtended] = polyconf(p,xExtended,S,'alpha',alpha);

    plot(x,y,'b+')
    hold on
    plot(xExtended,yfitExtended,'g-')
    plot(r,zeros(size(r)),'ko')
    plot(xExtended,yfitExtended-deltaExtended,'r--')
    plot(xExtended,yfitExtended+deltaExtended,'r--')
    plot(xExtended,zeros(size(xExtended)),'k-')
    legend('Data','Fit','Roots of Fit','95% Prediction Intervals')
    title(['Fit: ',texlabel(polystr(round(p,2)))])
    axis tight
    hold off
end
```



Alternatively, you can use `polytool` for interactive polynomial fitting.

```
polytool(x,y,degree,alpha)
```



Helper Function

The `polystr.m` file defines the `polystr` helper function.

```
type polystr.m % Display contents of polystr.m file
```

```
function s = polystr(p)
% POLYSTR Converts a vector of polynomial coefficients to a character string.
% S is the string representation of P.

if all(p == 0) % All coefficients are 0.
    s = '0';
else
    d = length(p) - 1; % Degree of polynomial.
    s = []; % Initialize s.
    for a = p
        if a ~= 0 % Coefficient is nonzero.
            if ~isempty(s) % String is not empty.
                if a > 0
```

```
                    s = [s ' + ']; % Add next term.
                else
                    s = [s ' - ']; % Subtract next term.
                    a = -a; % Value to subtract.
                end
            end
            if a ~= 1 || d == 0 % Add coefficient if it is ~=1 or polynomial is constant.
                s = [s num2str(a)];
                if d > 0 % For nonconstant polynomials, add *.
                    s = [s '*'];
                end
            end
            if d >= 2 % For terms of degree > 1, add power of x.
                s = [s 'x^' int2str(d)];
            elseif d == 1 % No power on x term.
                s = [s 'x'];
            end
        end
        d = d - 1; % Increment loop: Add term of next lowest degree.
    end
end
end
```

See Also

polyfit | polytool | polyval

Introduced before R2006a

polytool

Interactive polynomial fitting

Syntax

```
polytool(x,y)
polytool(x,y,n)
polytool(x,y,n,alpha)
polytool(x,y,n,alpha,xname,yname)
h = polytool(...)
```

Description

`polytool(x,y)` fits a line to the vectors `x` and `y` and displays an interactive plot of the result in a graphical interface. You can use the interface to explore the effects of changing the parameters of the fit and to export fit results to the workspace.

`polytool(x,y,n)` initially fits a polynomial of degree `n`. The default is `1`, which produces a linear fit.

`polytool(x,y,n,alpha)` initially plots `100(1 - alpha)`% confidence intervals on the predicted values. The default is `0.05` which results in 95% confidence intervals.

`polytool(x,y,n,alpha,xname,yname)` labels the `x` and `y` values on the graphical interface using `xname` and `yname`. Specify `n` and `alpha` as `[]` to use their default values.

`h = polytool(...)` outputs a vector of handles, `h`, to the line objects in the plot. The handles are returned in the degree: data, fit, lower bounds, upper bounds.

Examples

Interactive polynomial fitting

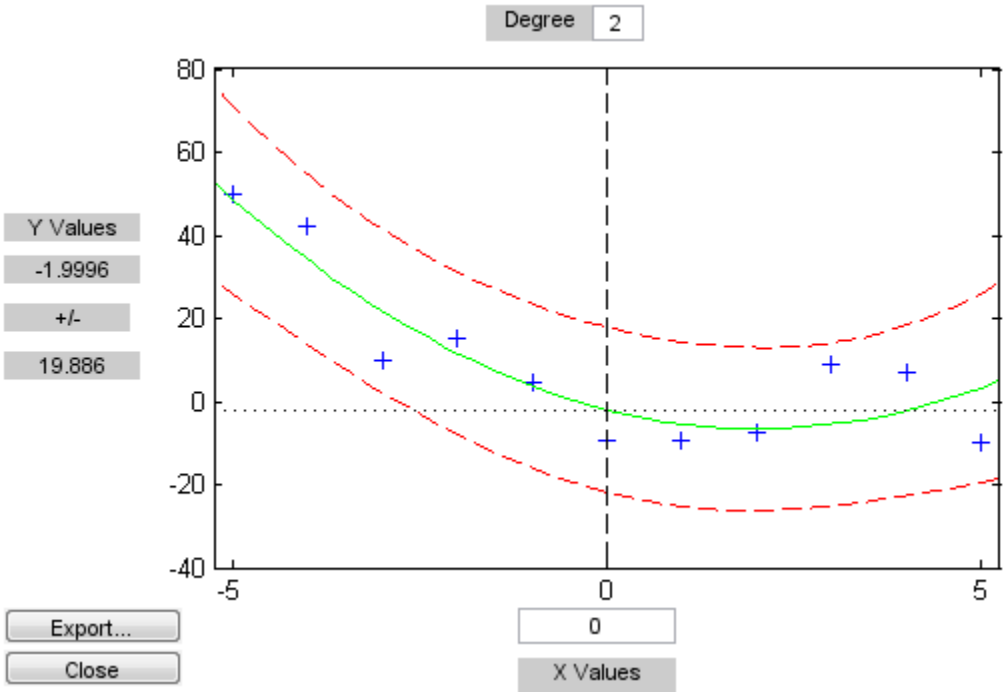
This example shows how to start an interactive fitting session with `polytool`.

Generate data from a quadratic curve with added noise.

```
rng('default') % for reproducibility
x = -5:5;
y = x.^2 - 5*x - 3 + 5*randn(size(x));
```

Fit a quadratic (degree-2) model with `0.90` confidence intervals.

```
n = 2;
alpha = 0.1;
polytool(x,y,n,alpha)
```



See Also

`invpred` | `polyconf` | `polyfit` | `polyval`

Introduced before R2006a

posterior

Posterior probability of Gaussian mixture component

Syntax

```
P = posterior(gm,X)
[P,nlogL] = posterior(gm,X)
```

Description

`P = posterior(gm,X)` returns the posterior probability of each Gaussian mixture component in `gm` given each observation in `X`.

`[P,nlogL] = posterior(gm,X)` also returns the negative loglikelihood of the Gaussian mixture model `gm` given the data `X`.

Examples

Compute Posterior Probabilities

Generate random variates that follow a mixture of two bivariate Gaussian distributions by using the `mvnrnd` function. Fit a Gaussian mixture model (GMM) to the generated data by using the `fitgmdist` function, and then compute the posterior probabilities of the mixture components.

Define the distribution parameters (means and covariances) of two bivariate Gaussian mixture components.

```
mu1 = [2 2];          % Mean of the 1st component
sigma1 = [2 0; 0 1];  % Covariance of the 1st component
mu2 = [-2 -1];        % Mean of the 2nd component
sigma2 = [1 0; 0 1];  % Covariance of the 2nd component
```

Generate an equal number of random variates from each component, and combine the two sets of random variates.

```
rng('default') % For reproducibility
r1 = mvnrnd(mu1,sigma1,1000);
r2 = mvnrnd(mu2,sigma2,1000);
X = [r1; r2];
```

The combined data set `X` contains random variates following a mixture of two bivariate Gaussian distributions.

Fit a two-component GMM to `X`.

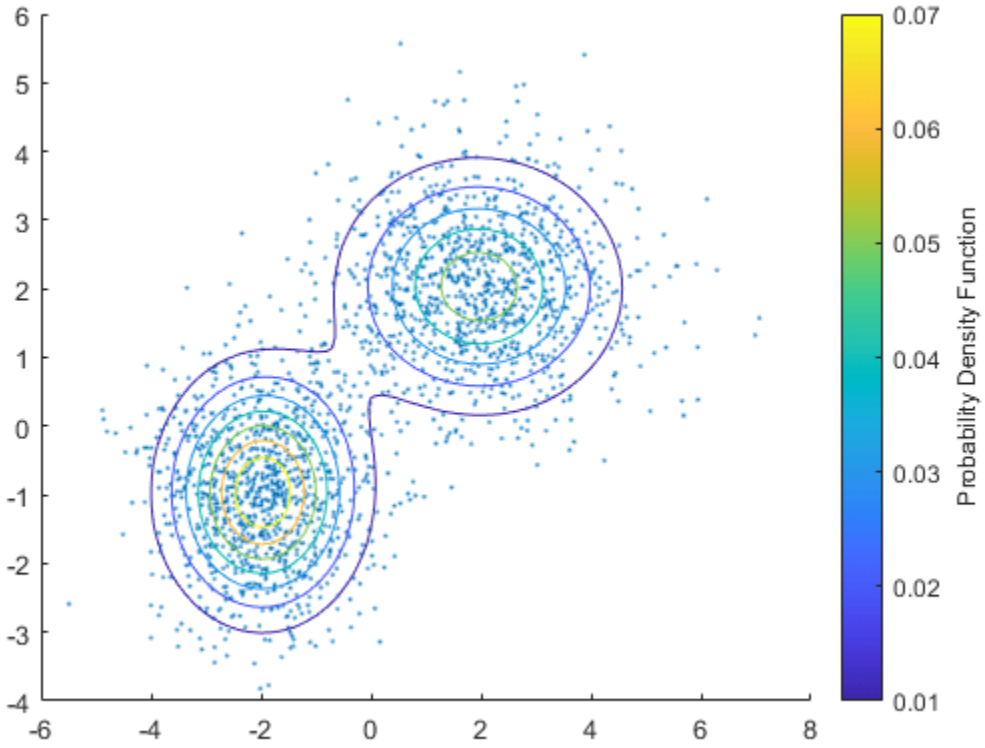
```
gm = fitgmdist(X,2)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
```

```
Mixing proportion: 0.500765
Mean:   -1.9675   -0.9654

Component 2:
Mixing proportion: 0.499235
Mean:    1.9657    2.0342
```

Plot `X` by using `scatter`. Visualize the fitted model `gm` by using `pdf` and `fcontour`.

```
figure
scatter(X(:,1),X(:,2),10,'.') % Scatter plot with points of size 10
hold on
gmPDF = @(x,y) arrayfun(@(x0,y0) pdf(gm,[x0 y0]),x,y);
fcontour(gmPDF,[-6 8 -4 6])
c1 = colorbar;
ylabel(c1,'Probability Density Function')
```



Compute the posterior probabilities of the components.

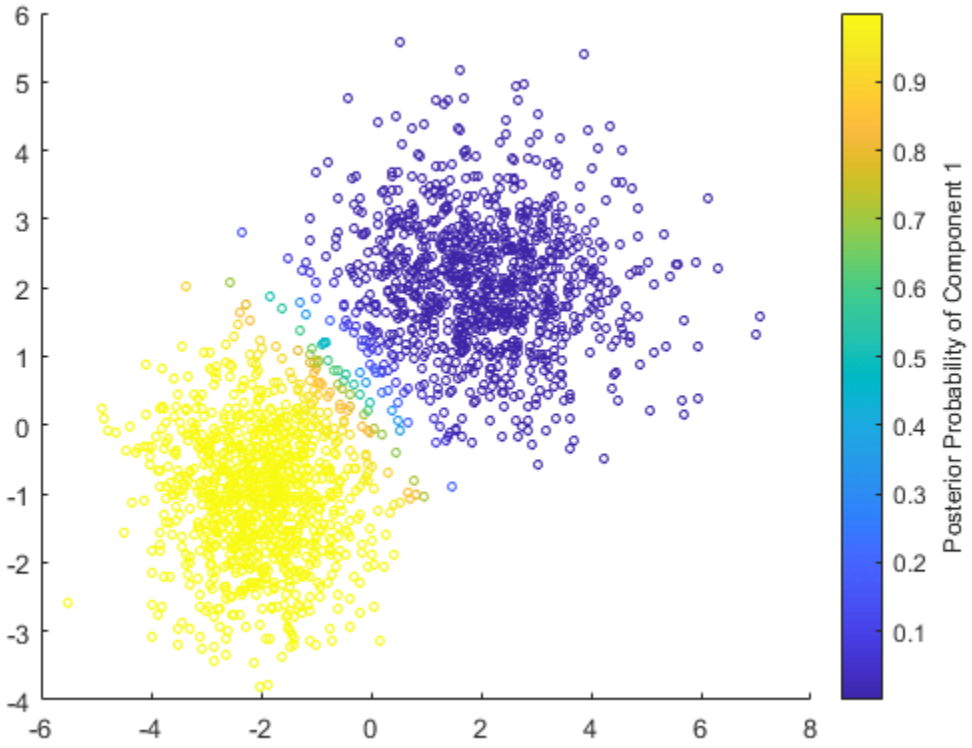
```
P = posterior(gm,X);
```

`P(i,j)` is the posterior probability of the `j`th Gaussian mixture component given observation `i`.

Plot the posterior probabilities of `Component 1` by using the `scatter` function. Use the circle colors to visualize the posterior probability values.

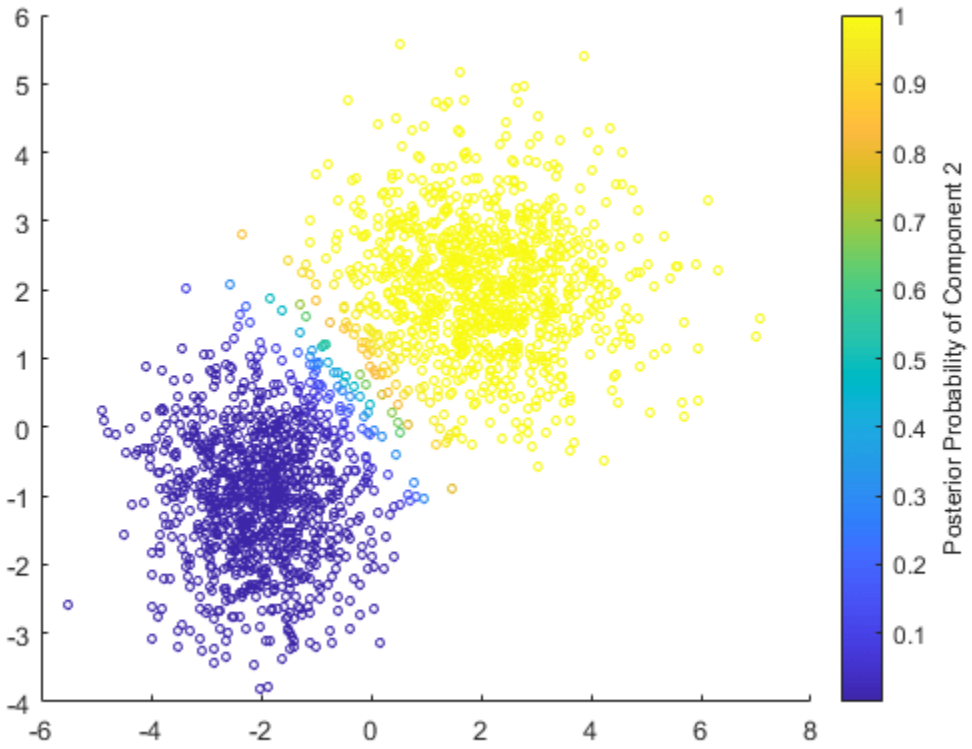
```
figure
scatter(X(:,1),X(:,2),10,P(:,1))
```

```
c2 = colorbar;
ylabel(c2,'Posterior Probability of Component 1')
```



Plot the posterior probabilities of `Component 2`.

```
figure
scatter(X(:,1),X(:,2),10,P(:,2))
c3 = colorbar;
ylabel(c3,'Posterior Probability of Component 2')
```



Input Arguments

`gm` — Gaussian mixture distribution

`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a `gmdistribution` object using `gmdistribution` or `fitgmdist`. Use the `gmdistribution` function to create a `gmdistribution` object by specifying the distribution parameters. Use the `fitgmdist` function to fit a `gmdistribution` model to data given a fixed number of components.

`X` — Data

n-by-*m* numeric matrix

Data, specified as an *n*-by-*m* numeric matrix, where *n* is the number of observations and *m* is the number of variables in each observation.

If a row of `X` contains `NaNs`, then `posterior` excludes the row from the computation. The corresponding value in `P` is `NaN`.

Data Types: `single` | `double`

Output Arguments

P — Posterior probability

n-by-*k* numeric vector

Posterior probability of each Gaussian mixture component in `gm` given each observation in `X`, returned as an *n*-by-*k* numeric vector, where *n* is the number of observations in `X` and *k* is the number of mixture components in `gm`.

`P(i,j)` is the posterior probability of the `j`th Gaussian mixture component given observation `i`, Probability(component `j` | observation `i`).

nlogL — Negative loglikelihood

numeric value

Negative loglikelihood value of the Gaussian mixture model `gm` given the data `X`, returned as a numeric value.

See Also

`cluster` | `fitgmdist` | `gmdistribution` | `mahal`

Topics

“Cluster Using Gaussian Mixture Model” on page 16-39
“Cluster Gaussian Mixture Data Using Hard Clustering” on page 16-46
“Cluster Gaussian Mixture Data Using Soft Clustering” on page 16-52

Introduced in R2007b

postFitStatistics

Class: `RegressionGP`

Compute post-fit statistics for the exact Gaussian process regression model

Syntax

```
loores = postFitStatistics(gprMdl)
[loores,neff] = postFitStatistics(gprMdl)
```

Description

`loores = postFitStatistics(gprMdl)` returns the leave-one-out residuals, `loores`, for the trained Gaussian process regression (GPR) model.

`[loores,neff] = postFitStatistics(gprMdl)` also returns the number of effective parameters, `neff`.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` object.

Output Arguments

`loores` — Leave-one-out residuals

n-by-1 matrix

Leave-one-out residuals, returned as an *n*-by-1 matrix, where *n* is the number of observations in the training data.

`neff` — Number of effective parameters

n-by-1 matrix

Number of effective parameters, returned as an *n*-by-1 matrix, where *n* is the number of observations in the training data.

Examples

Compute Post-Fit Statistics

Generate sample data.

```
rng(0,'twister'); % For reproducibility
n = 1500;
x = linspace(-10,10,n)';
y = sin(3*x).*cos(3*x) + sin(2*x).*cos(2*x) + sin(x) + cos(x) + 0.2*randn(n,1);
```

Fit a GPR model using the exact method for fitting and prediction.

```
gprMdl = fitrgp(x,y,'Basis','linear','FitMethod','exact',...
'PredictMethod','exact','KernelFunction','matern52');
```

Compute the leave-one-out residuals and the effective number of parameters in the trained model.

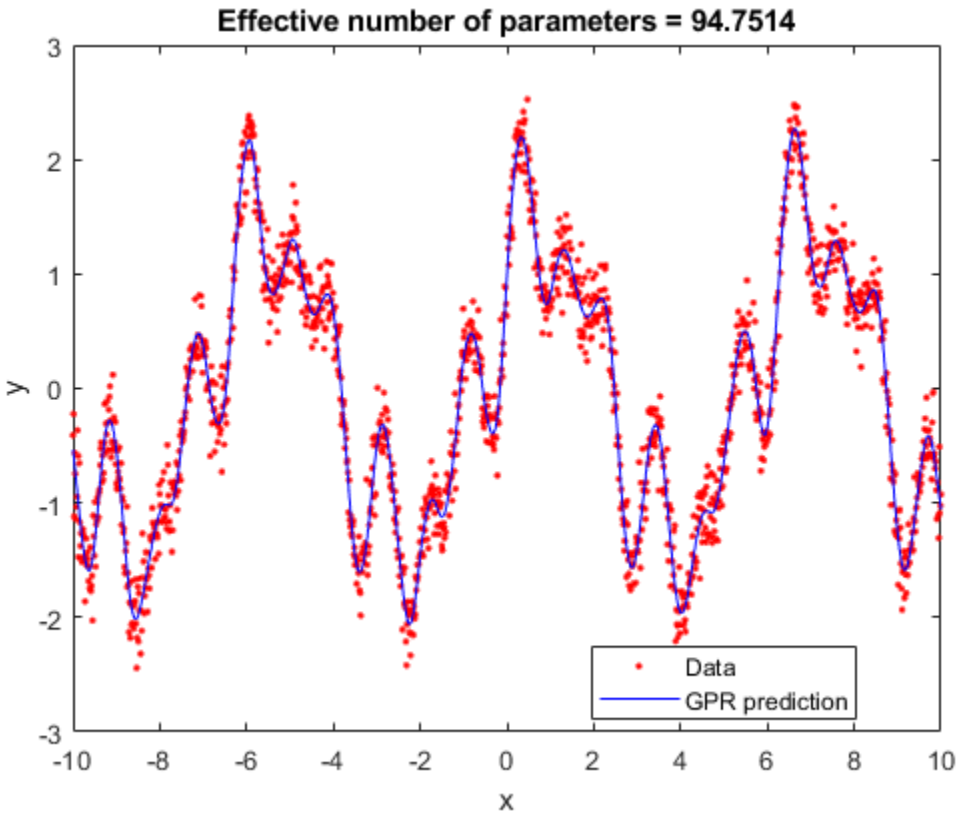
```
[loores,neff] = postFitStatistics(gprMdl);
```

Predict the responses using the trained model.

```
ypred = resubPredict(gprMdl);
```

Plot the true and predicted responses, and display effective number of parameters in the fit.

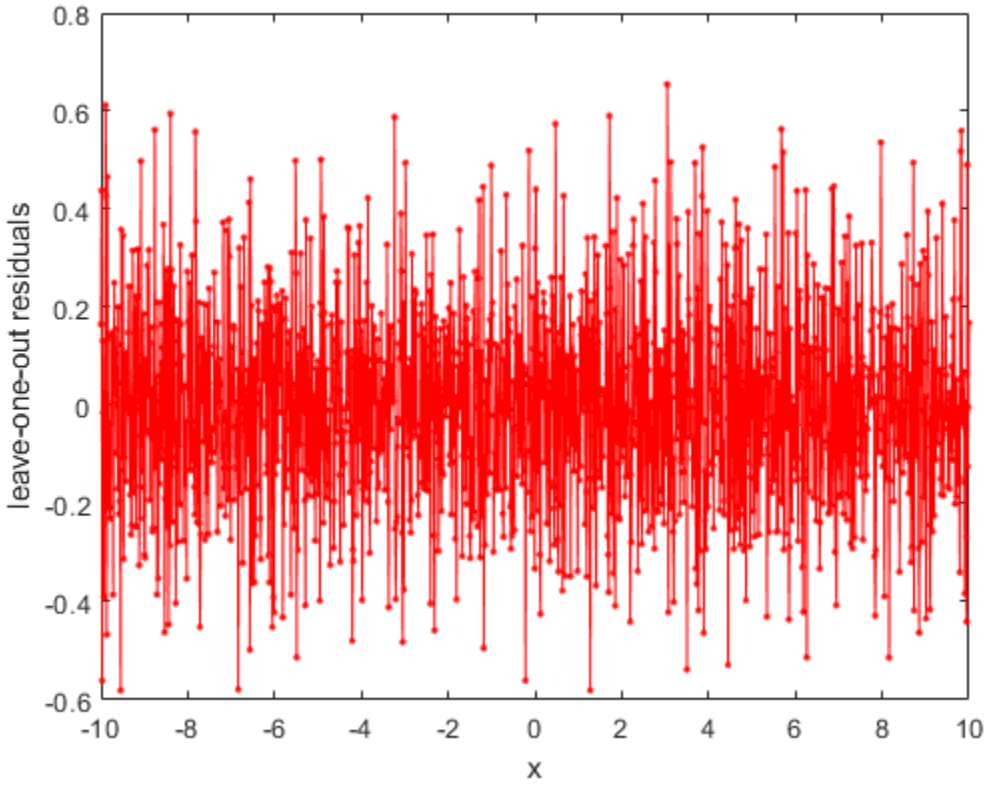
```
figure()
plot(x,y,'r.');
hold on;
plot(x,ypred,'b');
xlabel('x');
ylabel('y');
legend('Data','GPR prediction','Location','Best');
title(['Effective number of parameters = ',num2str(neff)]);
hold off
```



Plot leave-one-out residuals.

```
figure()
plot(x,loores,'r.-');
```

```
xlabel('x');
ylabel('leave-one-out residuals');
```



Tips

- You can only compute the post-fit statistics when `PredictMethod` is `'exact'`.
- If `FitMethod` is `'exact'`, then `postFitStatistics` accounts for the fact that the fixed basis function coefficients are estimated from the data.
- If `FitMethod` is different than `'exact'`, then `postFitStatistics` treats the fixed basis function coefficients as known.
- For all `PredictMethod` and `FitMethod` options, `postFitStatistics` treats the estimated kernel parameters and noise standard deviation as known.

See Also

`RegressionGP` | `fitrgp`

Introduced in R2015b

prctile

Percentiles of a data set

Syntax

```
Y = prctile(X,p)
Y = prctile(X,p,'all')
Y = prctile(X,p,dim)
Y = prctile(X,p,vecdim)
Y = prctile( ___ ,'Method',method)
```

Description

`Y = prctile(X,p)` returns percentiles of the elements in a data vector or array `X` for the percentages `p` in the interval [0,100].

- If `X` is a vector, then `Y` is a scalar or a vector with the same length as the number of percentiles requested (`length(p)`). `Y(i)` contains the `p(i)` percentile.
- If `X` is a matrix, then `Y` is a row vector or a matrix, where the number of rows of `Y` is equal to the number of percentiles requested (`length(p)`). The `i`th row of `Y` contains the `p(i)` percentiles of each column of `X`.
- For multidimensional arrays on page 33-4769, `prctile` operates along the first nonsingleton dimension on page 33-4769 of `X`.

`Y = prctile(X,p,'all')` returns percentiles of all the elements of `X`.

`Y = prctile(X,p,dim)` returns percentiles along the operating dimension `dim`.

`Y = prctile(X,p,vecdim)` returns percentiles over the dimensions specified in the vector `vecdim`. For example, if `X` is a matrix, then `prctile(X,50,[1 2])` returns the 50th percentile of all the elements of `X` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

`Y = prctile( ___ ,'Method',method)` returns either exact or approximate percentiles based on the value of `method`, using any of the input argument combinations in the previous syntaxes.

Examples

Percentiles of a Data Vector

Generate a data set of size 10.

```
rng('default'); % for reproducibility
x = normrnd(5,2,1,10)
```

```
x = 1×10

    6.0753    8.6678    0.4823    6.7243    5.6375    2.3846    4.1328    5.6852   12.1568   10.5
```

Calculate the 42nd percentile.

```
Y = prctile(x,42)
```

```
Y = 5.6709
```

Percentiles of All Values

Find the percentiles of all the values in an array.

Create a 3-by-5-by-2 array X.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

Find the 40th and 60th percentiles of the elements of X.

```
Y = prctile(X,[40 60],'all')
```

```
Y = 2×1

   12.5000
   18.5000
```

`Y(1)` is the 40th percentile of X, and `Y(2)` is the 60th percentile of X.

Percentiles of a Data Matrix

Calculate the percentiles along the columns and rows of a data matrix for specified percentages.

Generate a 5-by-5 data matrix.

```
X = (1:5)'*(2:6)
```

```
X = 5×5

     2     3     4     5     6
     4     6     8    10    12
```

```
     6     9    12    15    18
     8    12    16    20    24
    10    15    20    25    30
```

Calculate the 25th, 50th, and 75th percentiles along the columns of X.

```
Y = prctile(X,[25 50 75],1)
```

```
Y = 3×5

    3.5000    5.2500    7.0000    8.7500   10.5000
    6.0000    9.0000   12.0000   15.0000   18.0000
    8.5000   12.7500   17.0000   21.2500   25.5000
```

The rows of Y correspond to the percentiles of columns of X. For example, the 25th, 50th, and 75th percentiles of the third column of X with elements (4, 8, 12, 16, 20) are 7, 12, and 17, respectively. `Y = prctile(X,[25 50 75])` returns the same percentile matrix.

Calculate the 25th, 50th, and 75th percentiles along the rows of X.

```
Y = prctile(X,[25 50 75],2)
```

```
Y = 5×3

    2.7500    4.0000    5.2500
    5.5000    8.0000   10.5000
    8.2500   12.0000   15.7500
   11.0000   16.0000   21.0000
   13.7500   20.0000   26.2500
```

The rows of Y correspond to the percentiles of rows of X. For example, the 25th, 50th, and 75th percentiles of the first row of X with elements (2, 3, 4, 5, 6) are 2.75, 4, and 5.25, respectively.

Percentiles of Multidimensional Array

Find the percentiles of a multidimensional array along multiple dimensions simultaneously.

Create a 3-by-5-by-2 array X.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =

     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
```

```
    18    21    24    27    30
```

Calculate the 40th and 60th percentiles for each page of `X` by specifying dimensions 1 and 2 as the operating dimensions.

```
Ypage = prctile(X,[40 60],[1 2])
```

```
Ypage =
Ypage(:,:,1) =

    6.5000
    9.5000


Ypage(:,:,2) =

   21.5000
   24.5000
```

For example, `Ypage(1,1,1)` is the 40th percentile of the first page of `X`, and `Ypage(2,1,1)` is the 60th percentile of the first page of `X`.

Calculate the 40th and 60th percentiles of the elements in each `X(:,i,:)` slice by specifying dimensions 1 and 3 as the operating dimensions.

```
Ycol = prctile(X,[40 60],[1 3])
```

```
Ycol = 2×5

    2.9000    5.9000    8.9000   11.9000   14.9000
   16.1000   19.1000   22.1000   25.1000   28.1000
```

For example, `Ycol(1,4)` is the 40th percentile of the elements in `X(:,4,:)`, and `Ycol(2,4)` is the 60th percentile of the elements in `X(:,4,:)`.

Percentiles of Tall Vector for Given Percentage

Calculate exact and approximate percentiles of a tall column vector for a given percentage.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore for the `airlinesmall` data set. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Specify to work with the `ArrTime` variable.

```
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames','ArrTime');
```

Create a tall table on top of the datastore, and extract the data from the tall table into a tall vector.

```
t = tall(ds) % Tall table
```

```
t =

  Mx1 tall table

    ArrTime
    _______

      735
     1124
     2218
     1431
      746
     1547
     1052
     1134
       :
       :
```

```
x = t{:,:}   % Tall vector
```

```
x =

  Mx1 tall double column vector

        735
       1124
       2218
       1431
        746
       1547
       1052
       1134
        :
        :
```

Calculate the exact 50th percentile of `x`. Because `x` is a tall column vector and `p` is a scalar, `prctile` returns the exact percentile value by default.

```
p = 50;
yExact = prctile(x,p)
```

```
yExact =

  tall double

    ?
```

Calculate the approximate 50th percentile of x. Specify `'Method','approximate'` to use an approximation algorithm based on “T-Digest” on page 33-4770 for computing the percentile.

```
yApprox = prctile(x,p,'Method','approximate')
```

```
yApprox =

  MxNx... tall double array

    ?    ?    ?    ...
```

```
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[yExact,yApprox] = gather(yExact,yApprox)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 1.9 sec
- Pass 2 of 4: Completed in 0.79 sec
- Pass 3 of 4: Completed in 1.2 sec
- Pass 4 of 4: Completed in 1.1 sec
Evaluation completed in 6.4 sec

yExact = 1522

yApprox = 1.5220e+03
```

The values of the approximate percentile and the exact percentile are the same to the four digits shown.

Percentiles of Tall Matrix Along Different Dimensions

Calculate exact and approximate percentiles of a tall matrix for specified percentages along different dimensions.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a tall matrix `X` containing a subset of variables from the `airlinesmall` data set. See “Percentiles of Tall Vector for Given Percentage” on page 33-4764 for details about the steps to extract data from a tall array.

```
varnames = {'ArrDelay','ArrTime','DepTime','ActualElapsedTime'}; % Subset of variables in the da
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames',varnames); % Datastore
t = tall(ds);     % Tall table
X = t{:,varnames} % Tall matrix
```

```
X =

  Mx4 tall double matrix

           8         735         642          53
           8        1124        1021          63
          21        2218        2055          83
          13        1431        1332          59
           4         746         629          77
          59        1547        1446          61
```

```
     3        1052         928          84
    11        1134         859         155
    :          :            :           :
    :          :            :           :
```

When operating along a dimension that is not 1, the `prctile` function calculates the exact percentiles only, so that it can perform the computation efficiently using a sorting-based algorithm (see “Algorithms” on page 33-4771) instead of an approximation algorithm based on “T-Digest” on page 33-4770.

Calculate the exact 25th, 50th, and 75th percentiles of `X` along the second dimension.

```
p = [25 50 75]; % Vector of percentages
Yexact = prctile(X,p,2)
```

```
Yexact =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

When the function operates along the first dimension and `p` is a vector of percentages, you must use the approximation algorithm based on t-digest to compute the percentiles. Using the sorting-based algorithm to find the percentiles along the first dimension of a tall array is computationally intensive.

Calculate the approximate 25th, 50th, and 75th percentiles of `X` along the first dimension. Because the default dimension is 1, you do not need to specify a value for `dim`.

```
Yapprox = prctile(X,p,'Method','approximate')
```

```
Yapprox =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[Yexact,Yapprox] = gather(Yexact,Yapprox);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 4.5 sec
Evaluation completed in 6 sec
```

Show the first five rows of the exact 25th, 50th, and 75th percentiles along the second dimension of `X` .

```
Yexact(1:5,:)
```

```
ans = 5×3
10^3 ×
```

```
    0.0305    0.3475    0.6885
    0.0355    0.5420    1.0725
    0.0520    1.0690    2.1365
    0.0360    0.6955    1.3815
    0.0405    0.3530    0.6875
```

Each row of the matrix `Yexact` contains the three percentiles of the corresponding row in `X`. For example, `30.5`, `347.5`, and `688.5` are the 25th, 50th, and 75th percentiles, respectively, of the first row in `X`.

Show the approximate 25th, 50th, and 75th percentiles of `X` along the first dimension.

```
Yapprox
```

```
Yapprox = 3×4
10^3 ×

   -0.0070    1.1150    0.9321    0.0700
         0    1.5220    1.3350    0.1020
    0.0110    1.9180    1.7400    0.1510
```

Each column of the matrix `Yapprox` corresponds to the three percentiles for each column of the matrix `X`. For example, the first column of `Yapprox` with elements (–7, 0, 11) contains the percentiles for the first column of `X`.

Input Arguments

`X` — Input data

vector | array

Input data, specified as a vector or array.

Data Types: `double` | `single`

`p` — Percentages

scalar | vector

Percentages for which to compute percentiles, specified as a scalar or vector of scalars from 0 to 100.

Example: 25

Example: [25, 50, 75]

Data Types: `double` | `single`

`dim` — Dimension

positive integer

Dimension along which the percentiles of `X` are requested, specified as a positive integer. For example, for a matrix `X`, when `dim` = 1, `prctile` returns the percentile(s) of the columns of `X`; when `dim` = 2, `prctile` returns the percentile(s) of the rows of `X`. For a multidimensional array `X`, the length of the `dim`th dimension of `Y` is equal to the length of `p`.

Data Types: `double` | `single`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `Y` has length `length(p)` in the smallest specified operating dimension (that is, dimension `min(vecdim)`) and has length 1 in each of the remaining operating dimensions. The other dimension lengths are the same for `X` and `Y`.

For example, consider a 2-by-3-by-3 array `X` with `p = [20 40 60 80]`. In this case, `prctile(X,p,[1 2])` returns an array, where each page of the array contains the 20th, 40th, 60th, and 80th percentiles of the elements of the corresponding page of `X`. Because 1 and 2 are the operating dimensions, with `min([1 2]) = 1` and `length(p) = 4`, the output is a 4-by-1-by-3 array.

Data Types: `single` | `double`

`method` — Method for calculating percentiles
`'exact'` (default) | `'approximate'`

Method for calculating percentiles, specified as `'exact'` or `'approximate'`. By default, `prctile` returns the exact percentiles by implementing an algorithm on page 33-4771 that uses sorting. You can specify `'method','approximate'` for `prctile` to return approximate percentiles by implementing an algorithm that uses T-Digest on page 33-4770.

Data Types: `char` | `string`

Output Arguments

`Y` — Percentiles
scalar | array

Percentiles of a data vector or array, returned as a scalar or array for one or more percentage values.

- If `X` is a vector, then `Y` is a scalar or a vector with the same length as the number of percentiles requested (`length(p)`). `Y(i)` contains the `p(i)`th percentile.
- If `X` is an array of dimension *d*, then `Y` is an array with the length of the smallest operating dimension equal to the number of percentiles requested (`length(p)`).

More About

Multidimensional Array

A *multidimensional array* is an array with more than two dimensions. For example, if X is a 1-by-3-by-4 array, then X is a 3-D array.

Nonsingleton Dimension

A *nonsingleton dimension* of an array is a dimension whose size is not equal to 1. A *first nonsingleton dimension* of an array is the first dimension that satisfies the nonsingleton condition. For example, if `X` is a 1-by-1-by-2-by-4 array, then the third dimension is the first nonsingleton dimension of `X`.

Linear Interpolation

Linear interpolation uses linear polynomials to find $y_i = f(x_i)$, the values of the underlying function $Y = f(X)$ at the points in the vector or array x. Given the data points (x_1, y_1) and (x_2, y_2), where $y_1 = f(x_1)$ and $y_2 = f(x_2)$, linear interpolation finds $y = f(x)$ for a given x between x_1 and x_2 as follows:

$$y = f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1).$$

Similarly, if the $100(1.5/n)$th percentile is $y_{1.5/n}$ and the $100(2.5/n)$th percentile is $y_{2.5/n}$, then linear interpolation finds the $100(2.3/n)$th percentile, $y_{2.3/n}$ as:

$$y_{\frac{2.3}{n}} = y_{\frac{1.5}{n}} + \frac{\left(\frac{2.3}{n} - \frac{1.5}{n}\right)}{\left(\frac{2.5}{n} - \frac{1.5}{n}\right)}\left(y_{\frac{2.5}{n}} - y_{\frac{1.5}{n}}\right).$$

T-Digest

T-digest[2] on page 33-4772 is a probabilistic data structure that is a sparse representation of the empirical cumulative distribution function (CDF) of a data set. T-digest is useful for computing approximations of rank-based statistics (such as percentiles and quantiles) from online or distributed data in a way that allows for controllable accuracy, particularly near the tails of the data distribution.

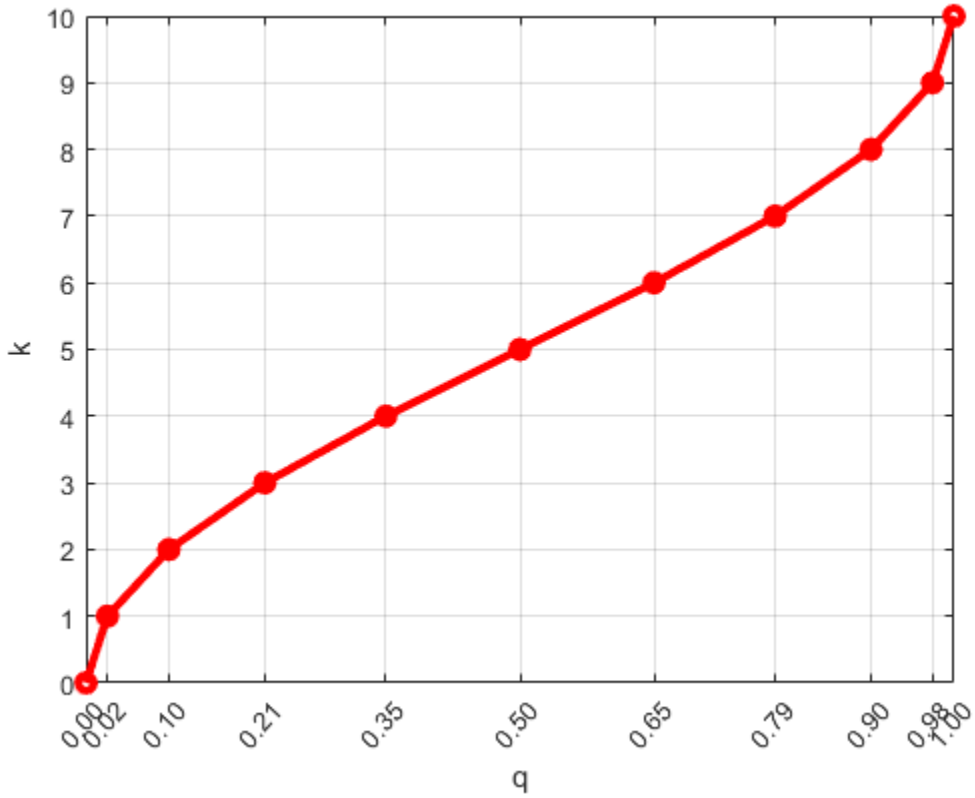
For data that is distributed in different partitions, t-digest computes quantile estimates (and percentile estimates) for each data partition separately, and then combines the estimates while maintaining a constant-memory bound and constant relative accuracy of computation ($q(1 - q)$ for the qth quantile). For these reasons, t-digest is practical for working with tall arrays.

To estimate quantiles of an array that is distributed in different partitions, first build a t-digest in each partition of the data. A t-digest clusters the data in the partition and summarizes each cluster by a centroid value and an accumulated weight that represents the number of samples contributing to the cluster. T-digest uses large clusters (widely spaced centroids) to represent areas of the CDF that are near `q = 0.5` and uses small clusters (tightly spaced centroids) to represent areas of the CDF that are near `q = 0` or `q = 1`.

T-digest controls the cluster size by using a scaling function that maps a quantile q to an index k with a compression parameter δ. That is,

$$k(q, \delta) = \delta \cdot \left(\frac{\sin^{-1}(2q - 1)}{\pi} + \frac{1}{2}\right),$$

where the mapping k is monotonic with minimum value $k(0,\delta) = 0$ and maximum value $k(1,\delta) = \delta$. The following figure shows the scaling function for $\delta = 10$.



The scaling function translates the quantile q to the scaling factor k in order to give variable size steps in q. As a result, cluster sizes are unequal (larger around the center quantiles and smaller near $q = 0$ or $q = 1$). The smaller clusters allow for better accuracy near the edges of the data.

To update a t-digest with a new observation that has a weight and location, find the cluster closest to the new observation. Then, add the weight and update the centroid of the cluster based on the weighted average, provided that the updated weight of the cluster does not exceed the size limitation.

You can combine independent t-digests from each partition of the data by taking a union of the t-digests and merging their centroids. To combine t-digests, first sort the clusters from all the independent t-digests in decreasing order of cluster weights. Then, merge neighboring clusters, when they meet the size limitation, to form a new t-digest.

Once you form a t-digest that represents the complete data set, you can estimate the end-points (or boundaries) of each cluster in the t-digest and then use interpolation between the end-points of each cluster to find accurate quantile estimates.

Algorithms

For an n-element vector `X`, `prctile` returns percentiles by using a sorting-based algorithm as follows:

1 The sorted elements in `X` are taken as the 100(0.5/n)th, 100(1.5/n)th, ..., 100([n – 0.5]/n)th percentiles. For example:

- For a data vector of five elements such as {6, 3, 2, 10, 1}, the sorted elements {1, 2, 3, 6, 10} respectively correspond to the 10th, 30th, 50th, 70th, and 90th percentiles.
- For a data vector of six elements such as {6, 3, 2, 10, 8, 1}, the sorted elements {1, 2, 3, 6, 8, 10} respectively correspond to the (50/6)th, (150/6)th, (250/6)th, (350/6)th, (450/6)th, and (550/6)th percentiles.

2 `prctile` uses linear interpolation on page 33-4770 to compute percentiles for percentages between 100(0.5/*n*) and 100([*n* – 0.5]/*n*).

3 `prctile` assigns the minimum or maximum values of the elements in `X` to the percentiles corresponding to the percentages outside that range.

`prctile` treats `NaN`s as missing values and removes them.

References

[1] Langford, E. “Quartiles in Elementary Statistics”, *Journal of Statistics Education*. Vol. 14, No. 3, 2006.

[2] Dunning, T., and O. Ertl. “Computing Extremely Accurate Quantiles Using T-Digests.” August 2017.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `Y = prctile(X,p)` returns the exact percentiles (using a sorting-based algorithm on page 33-4771) only if `X` is a tall column vector.
- `Y = prctile(X,p,dim)` returns the exact percentiles only when *one* of these conditions exists:
  - `X` is a tall column vector.
  - `X` is a tall array and `dim` is not `1`. For example, `prctile(X,p,2)` returns the exact percentiles along the rows of the tall array `X`.

  If `X` is a tall array and `dim` is `1`, then you must specify `'Method','approximate'` to use an approximation algorithm based on T-Digest on page 33-4770 for computing the percentiles. For example, `prctile(X,p,1,'Method','approximate')` returns the approximate percentiles along the columns of the tall array `X`.
- `Y = prctile(X,p,vecdim)` returns the exact percentiles only when *one* of these conditions exists:
  - `X` is a tall column vector.
  - `X` is a tall array and `vecdim` does not include `1`. For example, if `X` is a 3-by-5-by-2 array, then `prctile(X,p,[2,3])` returns the exact percentiles of the elements in each `X(i,:,:)` slice.
  - `X` is a tall array and `vecdim` includes `1` and all the nonsingleton dimensions on page 33-4769 of `X`. For example, if `X` is a 10-by-1-by-4 array, then `prctile(X,p,[1 3])` returns the exact percentiles of the elements in `X(:,1,:)`.

  If `X` is a tall array and `vecdim` includes `1` but does not include all the nonsingleton dimensions of `X`, then you must specify `'Method','approximate'` to use the approximation algorithm. For

example, if `X` is a 10-by-1-by-4 array, you can use `prctile(X,p,[1 2],'Method','approximate')` to find the approximate percentiles of each page of `X`.

For more information, see “Tall Arrays”.

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `'Method'` name-value pair argument is not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).
- If the output `Y` is a vector, the orientation of `Y` differs from MATLAB when all of the following are true:
  - You do not supply `dim`.
  - `X` is a variable-size array, and not a variable-size vector, at compile time, but `X` is a vector at run time.
  - The orientation of the vector `X` does not match the orientation of the vector `p`.

  In this case, the output `Y` matches the orientation of `X`, not the orientation of `p`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `'Method'` name-value pair argument is not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`iqr` | `median` | `quantile`

Topics
“Quantiles and Percentiles” on page 3-6

Introduced before R2006a

predict

Package:

Classify observations using generalized additive model (GAM)

Syntax

```
label = predict(Mdl,X)
label = predict(Mdl,X,'IncludeInteractions',includeInteractions)
[label,score] = predict( ___ )
```

Description

`label = predict(Mdl,X)` returns a vector of "Predicted Class Labels" on page 33-4781 for the predictor data in the table or matrix `X`, based on the generalized additive model `Mdl` for binary classification. The trained model can be either full or compact.

For each observation in `X`, the predicted class label corresponds to the minimum "Expected Misclassification Cost" on page 33-4781.

`label = predict(Mdl,X,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations.

`[label,score] = predict( ___ )` also returns classification scores using any of the input argument combinations in the previous syntaxes.

Examples

Label Test Sample Observations of GAM

Train a generalized additive model using training samples, and then label the test samples.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains sepal and petal measurements for versicolor and virginica irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
X = meas(inds,:);
Y = species(inds,:);
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a generalized additive model using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(XTrain,YTrain,'ClassNames',{'versicolor','virginica'})
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'versicolor'  'virginica'}
           ScoreTransform: 'logit'
                Intercept: -1.1090
          NumObservations: 70


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object.

Predict the test sample labels.

```
label = predict(Mdl,XTest);
```

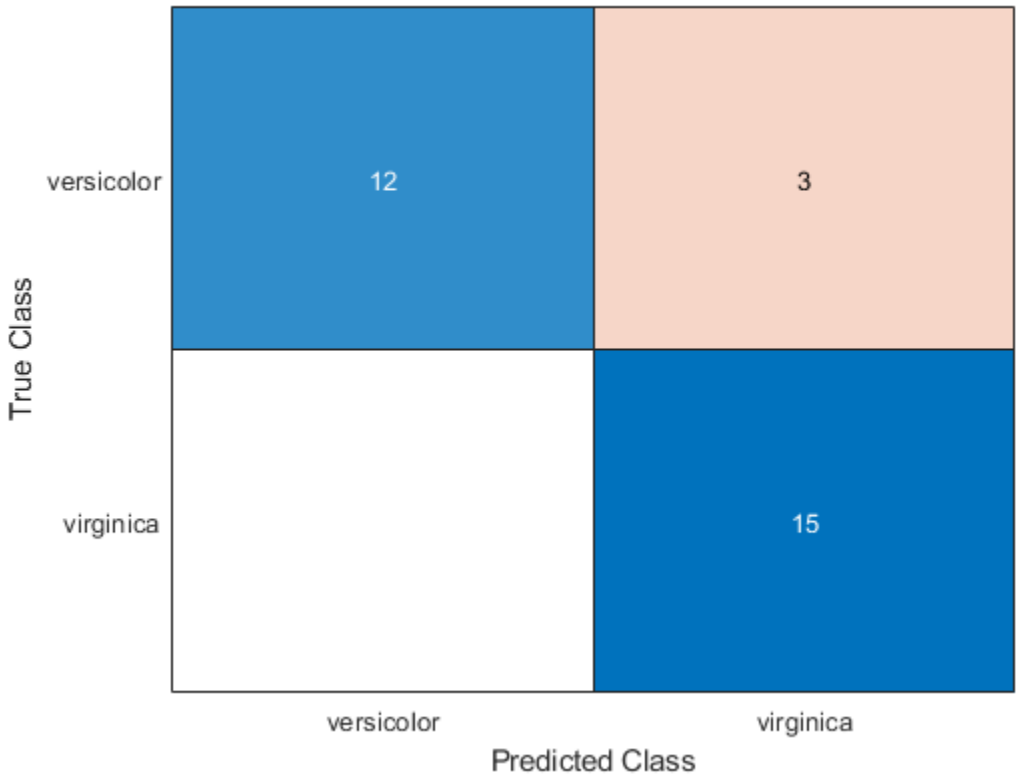
Create a table containing the true labels and predicted labels. Display the table for a random set of 10 observations.

```
t = table(YTest,label,'VariableNames',{'True Label','Predicted Label'});
idx = randsample(sum(testInds),10);
t(idx,:)
```

```
ans=10×2 table
      True Label      Predicted Label
    _______________    _______________

    {'virginica' }    {'virginica' }
    {'virginica' }    {'virginica' }
    {'versicolor'}    {'virginica' }
    {'virginica' }    {'virginica' }
    {'virginica' }    {'virginica' }
    {'versicolor'}    {'versicolor'}
    {'versicolor'}    {'versicolor'}
    {'versicolor'}    {'versicolor'}
    {'versicolor'}    {'versicolor'}
    {'virginica' }    {'virginica' }
```

Create a confusion chart from the true labels `YTest` and the predicted labels `label`.

```
cm = confusionchart(YTest,label);
```



Compare Logit of Posterior Probabilities

Estimate the logit of posterior probabilities for new observations using a classification GAM that contains both linear and interaction terms for predictors. Classify new observations using a memory-efficient model object. Specify whether to include interaction terms when classifying new observations.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 10 observations for the new test data set.

```
rng('default') % For reproducibility
n = size(X,1);
newInds = randsample(n,10);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X(inds,:),Y(inds),'ClassNames',{'b','g'},'Interactions',10);
```

`Mdl` is a `ClassificationGAM` model object.

Conserve memory by reducing the size of the trained model.

```
CMdl = compact(Mdl);
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class                                                  Attri

  CMdl      1x1             1081082  classreg.learning.classif.CompactClassificationGAM
  Mdl       1x1             1282640  ClassificationGAM
```

`CMdl` is a `CompactClassificationGAM` model object.

Predict the labels using both linear and interaction terms, and then using only linear terms. To exclude interaction terms, specify `'IncludeInteractions',false`. Estimate the logit of posterior probabilities by specifying the `ScoreTransform` property as `'none'`.

```
CMdl.ScoreTransform = 'none';
[labels,scores] = predict(CMdl,XNew);
[labels_nointeraction,scores_nointeraction] = predict(CMdl,XNew,'IncludeInteractions',false);
t = table(YNew,labels,scores,labels_nointeraction,scores_nointeraction, ...
    'VariableNames',{'True Labels','Predicted Labels','Scores' ...
    'Predicted Labels Without Interactions','Scores Without Interactions'})
```

```
t=10×5 table
    True Labels    Predicted Labels          Scores          Predicted Labels Without Interactio
    ___________    ________________    ____________________    ___________________________________

       {'g'}           {'g'}            -40.23      40.23                  {'g'}
       {'g'}           {'g'}           -41.215     41.215                  {'g'}
       {'g'}           {'g'}           -44.413     44.413                  {'g'}
       {'g'}           {'b'}            3.0658    -3.0658                  {'b'}
       {'g'}           {'g'}           -84.637     84.637                  {'g'}
       {'g'}           {'g'}            -27.44      27.44                  {'g'}
       {'g'}           {'g'}           -62.989     62.989                  {'g'}
       {'g'}           {'g'}           -77.109     77.109                  {'g'}
       {'g'}           {'g'}           -48.519     48.519                  {'g'}
       {'g'}           {'g'}           -56.256     56.256                  {'g'}
```

The predicted labels for the test data `Xnew` do not vary depending on the inclusion of interaction terms, but the estimated score values are different.

Plot Posterior Probability Regions

Train a generalized additive model, and then plot the posterior probability regions using the probability values of the first class.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains two petal measurements for versicolor and virginica irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
inds = strcmp(species,'versicolor') | strcmp(species,'virginica');
X = meas(inds,3:4);
Y = species(inds,:);
```

Train a generalized additive model using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names.

```
Mdl = fitcgam(X,Y,'ClassNames',{'versicolor','virginica'});
```

`Mdl` is a `ClassificationGAM` model object.

Define a grid of values in the observed predictor space.

```
xMax = max(X);
xMin = min(X);
x1 = linspace(xMin(1),xMax(1),250);
x2 = linspace(xMin(2),xMax(2),250);
[x1Grid,x2Grid] = meshgrid(x1,x2);
```

Predict the posterior probabilities for each instance in the grid.

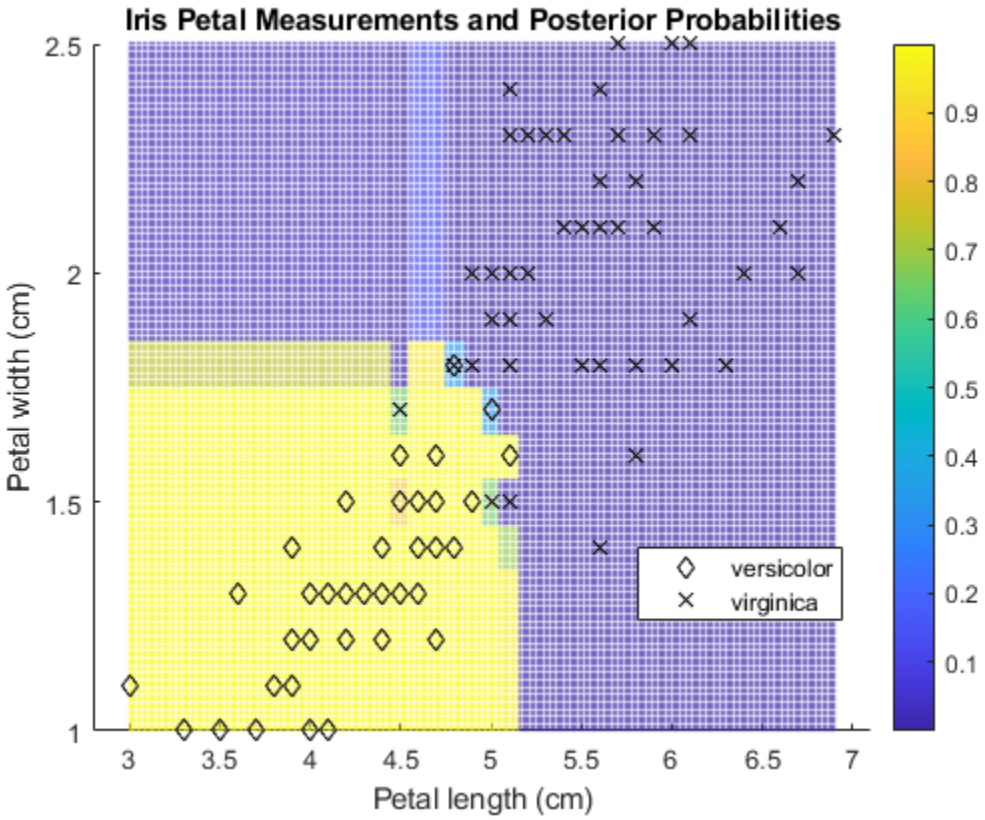
```
[~,PosteriorRegion] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
```

Plot the posterior probability regions using the probability values of the first class `'versicolor'`.

```
h = scatter(x1Grid(:),x2Grid(:),1,PosteriorRegion(:,1));
h.MarkerEdgeAlpha = 0.3;
```

Plot the training data.

```
hold on
gh = gscatter(X(:,1),X(:,2),Y,'k','dx');
title('Iris Petal Measurements and Posterior Probabilities')
xlabel('Petal length (cm)')
ylabel('Petal width (cm)')
legend(gh,'Location','Best')
colorbar
hold off
```



Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `CompactClassificationGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `CompactClassificationGAM` model object.

`X` — Predictor data

numeric matrix | table

Predictor data, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that make up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table, then `X` can be a numeric matrix if the table contains all numeric predictor variables.
- For a table:

- If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those in `Tbl`. However, the column order of `X` does not need to correspond to the column order of `Tbl`.
- If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and the corresponding predictor variable names in `X` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `X` must be numeric vectors.
- `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
- `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

Data Types: `table` | `double` | `single`

`includeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `includeInteractions` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

"Predicted Class Labels" on page 33-4781, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

If `Mdl.ScoreTransform` is `'logit'`(default), then each entry of `label` corresponds to the class with the minimal "Expected Misclassification Cost" on page 33-4781 for the corresponding row of `X`. Otherwise, each entry corresponds to the class with the maximal score.

`label` has the same data type as the observed class labels that trained `Mdl`, and its length is equal to the number of rows in `X`. (The software treats string arrays as cell arrays of character vectors.)

`score` — Predicted posterior probabilities or class scores
two-column numeric matrix

Predicted posterior probabilities or class scores, returned as a two-column numeric matrix with the same number of rows as `X`. The first and second columns of `score` contain the first class (or negative class, `Mdl.ClassNames(1)`) and second class (or positive class, `Mdl.ClassNames(2)`) score values for the corresponding observations, respectively.

If `Mdl.ScoreTransform` is `'logit'`(default), then the score values are posterior probabilities. If `Mdl.ScoreTransform` is `'none'`, then the score values are the logit of posterior probabilities. The software provides several built-in score transformation functions. For more details, see the `ScoreTransform` property of `Mdl`.

You can change the score transformation by specifying the `'ScoreTransform'` argument of `fitcgam` during training, or by changing the `ScoreTransform` property after training.

More About

Predicted Class Labels

`predict` classifies by minimizing the expected misclassification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{j=1}^{K} \widehat{P}(j|x) C(y|j),$$

where:

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(j|x)$ is the posterior probability of class j for observation x.
- $C(y|j)$ is the cost of classifying an observation as y when its true class is j.

Expected Misclassification Cost

The expected misclassification cost per observation is an averaged cost of classifying the observation into each class.

Suppose you have `Nobs` observations that you want to classify with a trained classifier, and you have `K` classes. You place the observations into a matrix `X` with one observation per row.

The expected cost matrix `CE` has size `Nobs`-by-`K`. Each row of `CE` contains the expected (average) cost of classifying the observation into each of the `K` classes. `CE(n,k)` is

$$\sum_{i=1}^{K} \widehat{P}(i|X(n)) C(k|i),$$

where:

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability of class i for observation $X(n)$.
- $C(k|i)$ is the true misclassification cost of classifying an observation as k when its true class is i.

True Misclassification Cost

The true misclassification cost is the cost of classifying an observation into an incorrect class.

You can set the true misclassification cost per class by using the `'Cost'` name-value argument when you create the classifier. `Cost(i,j)` is the cost of classifying an observation into class `j` when its true class is `i`. By default, `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

See Also

`edge` | `loss` | `margin` | `resubPredict`

Topics

“Train Generalized Additive Model for Binary Classification” on page 12-77

Introduced in R2021a

predict

Predict labels using *k*-nearest neighbor classification model

Syntax

```
label = predict(mdl,X)
[label,score,cost] = predict(mdl,X)
```

Description

`label = predict(mdl,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the trained *k*-nearest neighbor classification model `mdl`. See “Predicted Class Label” on page 33-4785.

`[label,score,cost] = predict(mdl,X)` also returns:

- A matrix of classification scores (`score`) indicating the likelihood that a label comes from a particular class. For *k*-nearest neighbor, scores are posterior probabilities. See “Posterior Probability” on page 33-4786.
- A matrix of expected classification cost (`cost`). For each observation in `X`, the predicted class label corresponds to the minimum expected classification costs among all classes. See “Expected Cost” on page 33-4786.

Examples

k-Nearest Neighbor Classification Predictions

Create a *k*-nearest neighbor classifier for Fisher's iris data, where $k = 5$. Evaluate some model predictions on new data.

Load the Fisher iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create a classifier for five nearest neighbors. Standardize the noncategorical predictor data.

```
mdl = fitcknn(X,Y,'NumNeighbors',5,'Standardize',1);
```

Predict the classifications for flowers with minimum, mean, and maximum characteristics.

```
Xnew = [min(X);mean(X);max(X)];
[label,score,cost] = predict(mdl,Xnew)
```

```
label = 3x1 cell
    {'versicolor'}
    {'versicolor'}
    {'virginica' }
```

```
score = 3×3

    0.4000    0.6000         0
         0    1.0000         0
         0         0    1.0000


cost = 3×3

    0.6000    0.4000    1.0000
    1.0000         0    1.0000
    1.0000    1.0000         0
```

The second and third rows of the score and cost matrices have binary values, which means all five nearest neighbors of the mean and maximum flower measurements have identical classifications.

Input Arguments

`mdl` — *k*-nearest neighbor classifier model

`ClassificationKNN` object

k-nearest neighbor classifier model, specified as a `ClassificationKNN` object.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that make up the columns of `X` must have the same order as the predictor variables used to train `mdl`.
  - If you train `mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. *k*-nearest neighbor classification requires homogeneous predictors. Therefore, to treat all numeric predictors in `Tbl` as categorical during training, set `'CategoricalPredictors','all'` when you train using `fitcknn`. If `Tbl` contains heterogeneous predictors (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables and cell arrays other than cell arrays of character vectors.
  - If you train `mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those used to train `mdl` (stored in `mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Both `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `mdl` using a numeric matrix, then the predictor names in `mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitcknn`. All

predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

If you set `'Standardize',true` in `fitcknn` to train `mdl`, then the software standardizes the columns of `X` using the corresponding means in `mdl.Mu` and standard deviations in `mdl.Sigma`.

Data Types: `double` | `single` | `table`

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical vector | vector of numeric values | cell array of character vectors

Predicted class labels for the observations (rows) in `X`, returned as a categorical array, character array, logical vector, vector of numeric values, or cell array of character vectors. `label` has length equal to the number of rows in `X`. The label is the class with minimal expected cost. See “Predicted Class Label” on page 33-4785.

`score` — Predicted class scores or posterior probabilities

numeric matrix

Predicted class scores or posterior probabilities, returned as a numeric matrix of size *n*-by-*K*. *n* is the number of observations (rows) in `X`, and *K* is the number of classes (in `mdl.ClassNames`). `score(i,j)` is the posterior probability that observation `i` in `X` is of class `j` in `mdl.ClassNames`. See “Posterior Probability” on page 33-4786.

Data Types: `single` | `double`

`cost` — Expected classification costs

numeric matrix

Expected classification costs, returned as a numeric matrix of size *n*-by-*K*. *n* is the number of observations (rows) in `X`, and *K* is the number of classes (in `mdl.ClassNames`). `cost(i,j)` is the cost of classifying row `i` of `X` as class `j` in `mdl.ClassNames`. See “Expected Cost” on page 33-4786.

Data Types: `single` | `double`

Algorithms

Predicted Class Label

`predict` classifies by minimizing the expected misclassification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{j=1}^{K} \widehat{P}(j|x) C(y|j),$$

where:

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(j|x)$ is the posterior probability of class j for observation x.

- $C(y|j)$ is the cost of classifying an observation as y when its true class is j.

Posterior Probability

Consider a vector (single query point) `xnew` and a model `mdl`.

- k is the number of nearest neighbors used in prediction, `mdl.NumNeighbors`.
- `nbd(mdl,xnew)` specifies the k nearest neighbors to `xnew` in `mdl.X`.
- `Y(nbd)` specifies the classifications of the points in `nbd(mdl,xnew)`, namely `mdl.Y(nbd)`.
- `W(nbd)` specifies the weights of the points in `nbd(mdl,xnew)`.
- `prior` specifies the priors of the classes in `mdl.Y`.

If the model contains a vector of prior probabilities, then the observation weights `W` are normalized by class to sum to the priors. This process might involve a calculation for the point `xnew`, because weights can depend on the distance from `xnew` to the points in `mdl.X`.

The posterior probability $p(j|\texttt{xnew})$ is

$$p(j|\text{xnew}) = \frac{\sum_{i \in \text{nbd}} W(i) 1_{Y(X(i)) = j}}{\sum_{i \in \text{nbd}} W(i)}.$$

Here, $1_{Y(X(i)) = j}$ is `1` when `mdl.Y(i) = j`, and `0` otherwise.

True Misclassification Cost

Two costs are associated with KNN classification: the true misclassification cost per class and the expected misclassification cost per observation.

You can set the true misclassification cost per class by using the `'Cost'` name-value pair argument when you run `fitcknn`. The value `Cost(i,j)` is the cost of classifying an observation into class `j` if its true class is `i`. By default, `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Expected Cost

Two costs are associated with KNN classification: the true misclassification cost per class and the expected misclassification cost per observation. The third output of `predict` is the expected misclassification cost per observation.

Suppose you have `Nobs` observations that you want to classify with a trained classifier `mdl`, and you have `K` classes. You place the observations into a matrix `Xnew` with one observation per row. The command

```
[label,score,cost] = predict(mdl,Xnew)
```

returns a matrix `cost` of size `Nobs`-by-`K`, among other outputs. Each row of the `cost` matrix contains the expected (average) cost of classifying the observation into each of the `K` classes. `cost(n,j)` is

$$\sum_{i=1}^{K} \widehat{P}(i|Xnew(n))C(j|i),$$

where

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability on page 33-4786 of class i for observation $Xnew(n)$.
- $C(j|i)$ is the true misclassification cost of classifying an observation as j when its true class is i.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations

`mdl`	• A `ClassificationKNN` model object is a full object that does not have a corresponding compact object. For this model, `saveLearnerForCoder` saves a compact version that does not include the hyperparameter optimization properties.
	• If `mdl` is a model trained using the *k*d-tree search algorithm, and the code generation build type is a MEX function, then `codegen` generates a MEX function using Intel Threading Building Blocks (TBB) for parallel computation. Otherwise, `codegen` generates code using `parfor`.
	  • MEX function for the *k*d-tree search algorithm — `codegen` generates an optimized MEX function using Intel TBB for parallel computation on multicore platforms. You can use the MEX function to accelerate MATLAB algorithms. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.
	  If you generate the MEX function to test the generated code of the `parfor` version, you can disable the usage of Intel TBB. Set the `ExtrinsicCalls` property of the MEX configuration object to `false`. For details, see `coder.MexCodeConfig`.
	  • MEX function for the exhaustive search algorithm and standalone C/C++ code for both algorithms — The generated code of `predict` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.
	• For the usage notes and limitations of the model object, see “Code Generation” on page 33-411 of the `ClassificationKNN` object.

Argument	Notes and Limitations

X	• X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in X can be a variable size, but the number of columns in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	  For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

For more information, see "Introduction to Code Generation" on page 32-2.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also
`ClassificationKNN` | `fitcknn`

Topics
"Predict Classification Using KNN Classifier" on page 18-29
"Classification Using Nearest Neighbors" on page 18-12

Introduced in R2012a

predict

Class: `ClassificationLinear`

Predict labels for linear classification models

Syntax

```
Label = predict(Mdl,X)
Label = predict(Mdl,X,'ObservationsIn',dimension)
[Label,Score] = predict( ___ )
```

Description

`Label = predict(Mdl,X)` returns predicted class labels for each observation in the predictor data `X` based on the trained, binary, linear classification model `Mdl`. `Label` contains class labels for each regularization strength in `Mdl`.

`Label = predict(Mdl,X,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

`[Label,Score] = predict( ___ )` also returns classification scores on page 33-4800 for both classes using any of the input argument combinations in the previous syntaxes. `Score` contains classification scores for each regularization strength in `Mdl`.

Input Arguments

`Mdl` — Binary, linear classification model

`ClassificationLinear` model object

Binary, linear classification model, specified as a `ClassificationLinear` model object. You can create a `ClassificationLinear` model object using `fitclinear`.

`X` — Predictor data to be classified

full numeric matrix | sparse numeric matrix | table

Predictor data to be classified, specified as a full or sparse numeric matrix or a table.

By default, each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`) and `Tbl` contains only numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value pair argument of `fitclinear`. If `Tbl` contains heterogeneous predictor variables (for

example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.

- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` must be the same as the corresponding predictor variable names in `X`. To specify predictor names during training, use the `PredictorNames` name-value pair argument of `fitclinear`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `table` | `double` | `single`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'columns'` or `'rows'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Output Arguments

`Label` — Predicted class labels
categorical array | character array | logical matrix | numeric matrix | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric matrix, or cell array of character vectors.

In most cases, `Label` is an *n*-by-*L* array of the same data type as the observed class labels (`Y`) used to train `Mdl`. (The software treats string arrays as cell arrays of character vectors.) *n* is the number of observations in `X` and *L* is the number of regularization strengths in `Mdl.Lambda`. That is, `Label(i,j)` is the predicted class label for observation `i` using the linear classification model that has regularization strength `Mdl.Lambda(j)`.

If `Y` is a character array and $L > 1$, then `Label` is a cell array of class labels.

`Score` — Classification scores
numeric array

Classification scores on page 33-4800, returned as a *n*-by-2-by-*L* numeric array. *n* is the number of observations in `X` and *L* is the number of regularization strengths in `Mdl.Lambda`. `Score(i,k,j)` is the score for classifying observation `i` into class `k` using the linear classification model that has regularization strength `Mdl.Lambda(j)`. `Mdl.ClassNames` stores the order of the classes.

If `Mdl.Learner` is `'logistic'`, then classification scores are posterior probabilities.

Examples

Predict Training-Sample Labels

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

The models should identify whether the word counts in a web page are from the Statistics and Machine Learning Toolbox™ documentation. So, identify the labels that correspond to the Statistics and Machine Learning Toolbox™ documentation web pages.

```
Ystats = Y == 'stats';
```

Train a binary, linear classification model using the entire data set, which can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation.

```
rng(1); % For reproducibility
Mdl = fitclinear(X,Ystats);
```

`Mdl` is a `ClassificationLinear` model.

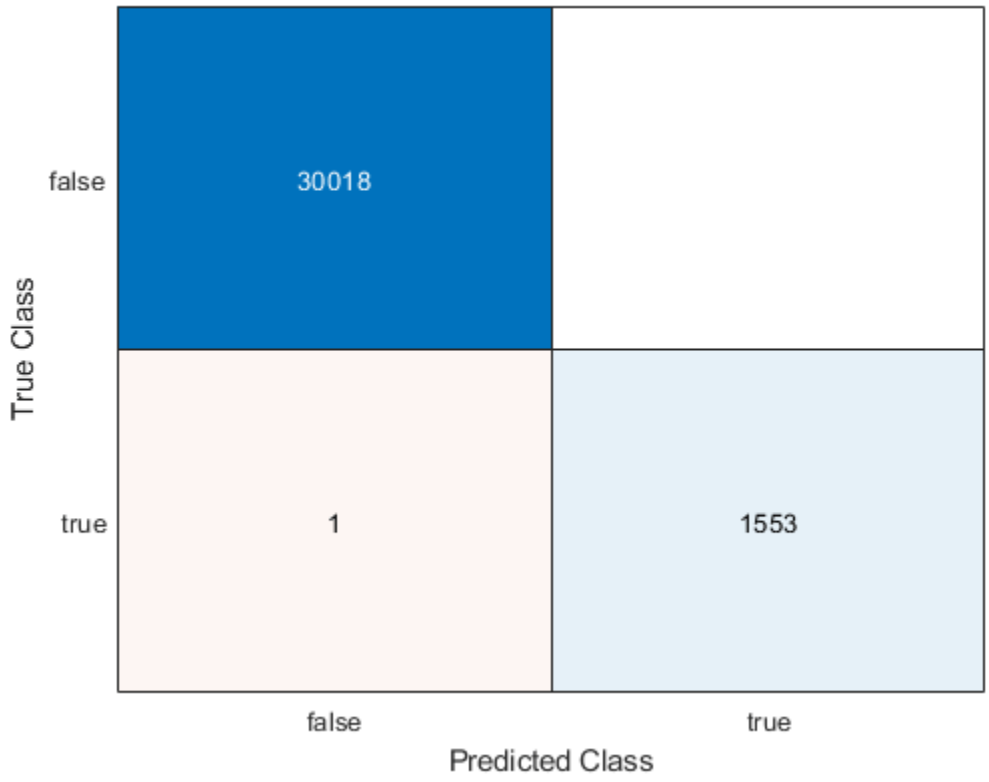
Predict the training-sample, or resubstitution, labels.

```
label = predict(Mdl,X);
```

Because there is one regularization strength in `Mdl`, `label` is column vectors with lengths equal to the number of observations.

Construct a confusion matrix.

```
ConfusionTrain = confusionchart(Ystats,label);
```



The model misclassifies only one `'stats'` documentation page as being outside of the Statistics and Machine Learning Toolbox documentation.

Predict Test-Sample Labels

Load the NLP data set and preprocess it as in “Predict Training-Sample Labels” on page 33-4792. Transpose the predictor data matrix.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Train a binary, linear classification model that can identify whether the word counts in a documentation web page are from the Statistics and Machine Learning Toolbox™ documentation. Specify to hold out 30% of the observations. Optimize the objective function using SpaRSA.

```
rng(1) % For reproducibility
CVMdl = fitclinear(X,Ystats,'Solver','sparsa','Holdout',0.30,...
    'ObservationsIn','columns');
Mdl = CVMdl.Trained{1};
```

`CVMdl` is a `ClassificationPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `ClassificationLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

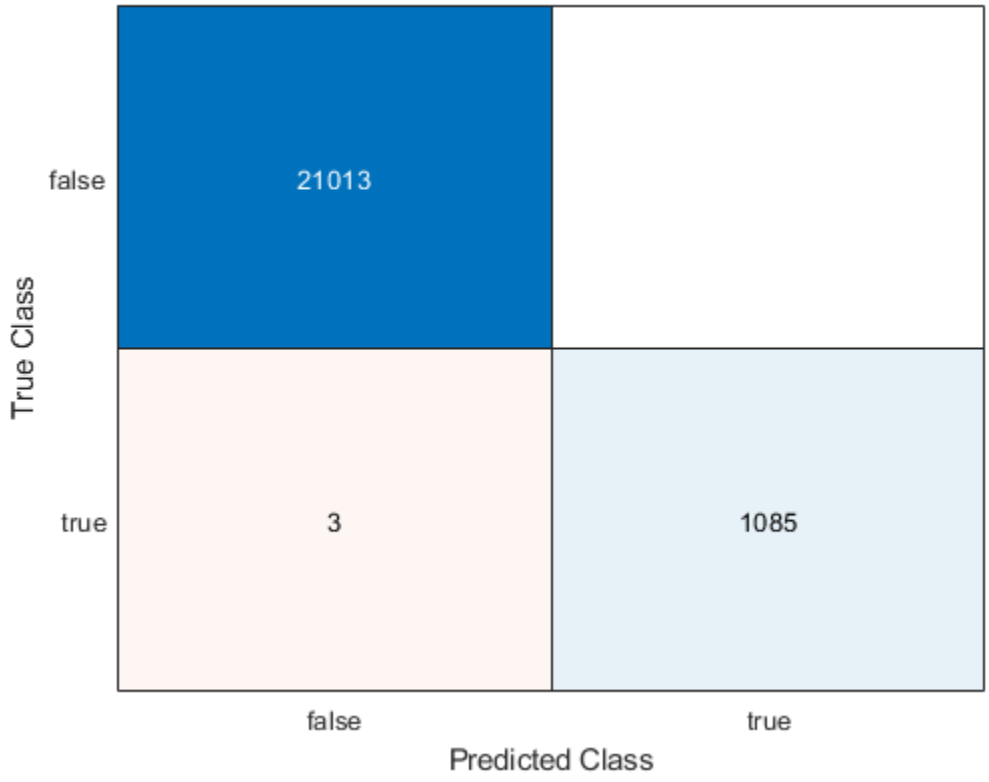
Predict the training- and test-sample labels.

```
labelTrain = predict(Mdl,X(:,trainIdx),'ObservationsIn','columns');
labelTest = predict(Mdl,X(:,testIdx),'ObservationsIn','columns');
```

Because there is one regularization strength in `Mdl`, `labelTrain` and `labelTest` are column vectors with lengths equal to the number of training and test observations, respectively.

Construct a confusion matrix for the training data.

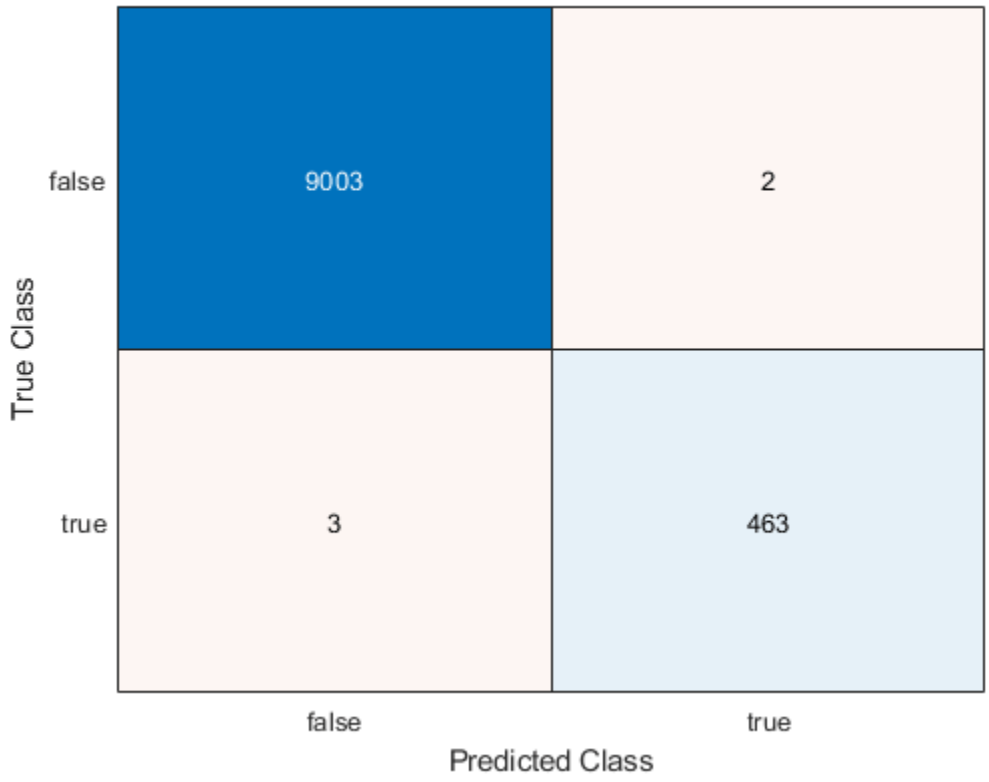
```
ConfusionTrain = confusionchart(Ystats(trainIdx),labelTrain);
```



The model misclassifies only three documentation pages as being outside of Statistics and Machine Learning Toolbox documentation.

Construct a confusion matrix for the test data.

```
ConfusionTest = confusionchart(Ystats(testIdx),labelTest);
```



The model misclassifies three documentation pages as being outside the Statistics and Machine Learning Toolbox, and two pages as being inside.

Estimate Posterior Class Probabilities

Estimate test-sample, posterior class probabilities, and determine the quality of the model by plotting a ROC curve. Linear classification models return posterior probabilities for logistic regression learners only.

Load the NLP data set and preprocess it as in “Predict Test-Sample Labels” on page 33-4793.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Randomly partition the data into training and test sets by specifying a 30% holdout sample. Identify the test-set indices.

```
cvp = cvpartition(Ystats,'Holdout',0.30);
idxTest = test(cvp);
```

Train a binary linear classification model. Fit logistic regression learners using SpaRSA. To hold out the test set, specify the partitioned model.

```
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns','CVPartition',cvp,...
    'Learner','logistic','Solver','sparsa');
Mdl = CVMdl.Trained{1};
```

`Mdl` is a `ClassificationLinear` model trained using the training set specified in the partition `cvp` only.

Predict the test-sample posterior class probabilities.

```
[~,posterior] = predict(Mdl,X(:,idxTest),'ObservationsIn','columns');
```

Because there is one regularization strength in `Mdl`, `posterior` is a matrix with 2 columns and rows equal to the number of test-set observations. Column *i* contains posterior probabilities of `Mdl.ClassNames(i)` given a particular observation.

Obtain false and true positive rates, and estimate the AUC. Specify that the second class is the positive class.

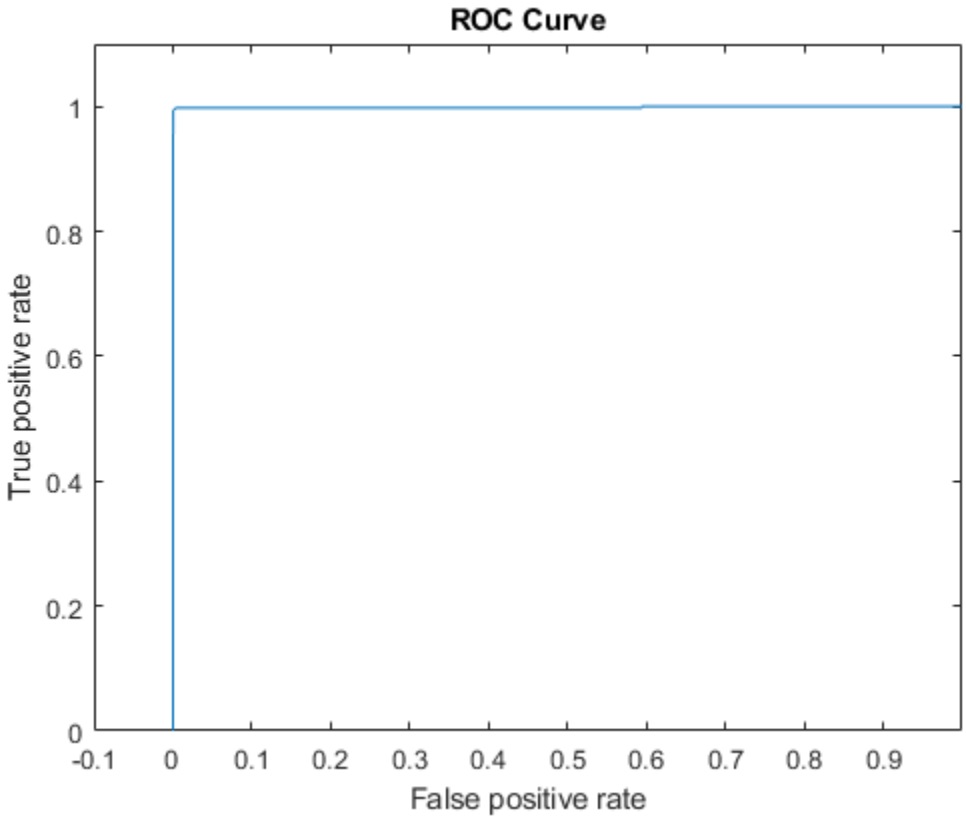
```
[fpr,tpr,~,auc] = perfcurve(Ystats(idxTest),posterior(:,2),Mdl.ClassNames(2));
auc
```

```
auc = 0.9986
```

The AUC is 1, which indicates a model that predicts well.

Plot an ROC curve.

```
figure;
plot(fpr,tpr)
h = gca;
h.XLim(1) = -0.1;
h.YLim(2) = 1.1;
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve')
```



The ROC curve and AUC indicate that the model classifies the test-sample observations almost perfectly.

Find Good Lasso Penalty Using AUC

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare test-sample values of the AUC.

Load the NLP data set. Preprocess the data as in “Predict Test-Sample Labels” on page 33-4793.

```
load nlpdata
Ystats = Y == 'stats';
X = X';
```

Create a data partition that specifies to holdout 10% of the observations. Extract test-sample indices.

```
rng(10); % For reproducibility
Partition = cvpartition(Ystats,'Holdout',0.10);
testIdx = test(Partition);
XTest = X(:,testIdx);
n = sum(testIdx)
```

```
n = 3157
```

```
YTest = Ystats(testIdx);
```

There are 3157 observations in the test sample.

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Train binary, linear classification models that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'CVPartition',Partition,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract the trained linear classification model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'


  Properties, Methods
```

`Mdl` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the test-sample predicted labels and posterior class probabilities.

```
[label,posterior] = predict(Mdl1,XTest,'ObservationsIn','columns');
Mdl1.ClassNames;
posterior(3,1,5)
```

```
ans = 1.0000
```

`label` is a 3157-by-11 matrix of predicted labels. Each column corresponds to the predicted labels of the model trained using the corresponding regularization strength. `posterior` is a 3157-by-2-by-11

matrix of posterior class probabilities. Columns correspond to classes and pages correspond to regularization strengths. For example, `posterior(3,1,5)` indicates that the posterior probability that the first class (label `0`) is assigned to observation 3 by the model that uses `Lambda(5)` as a regularization strength is 1.0000.

For each model, compute the AUC. Designate the second class as the positive class.

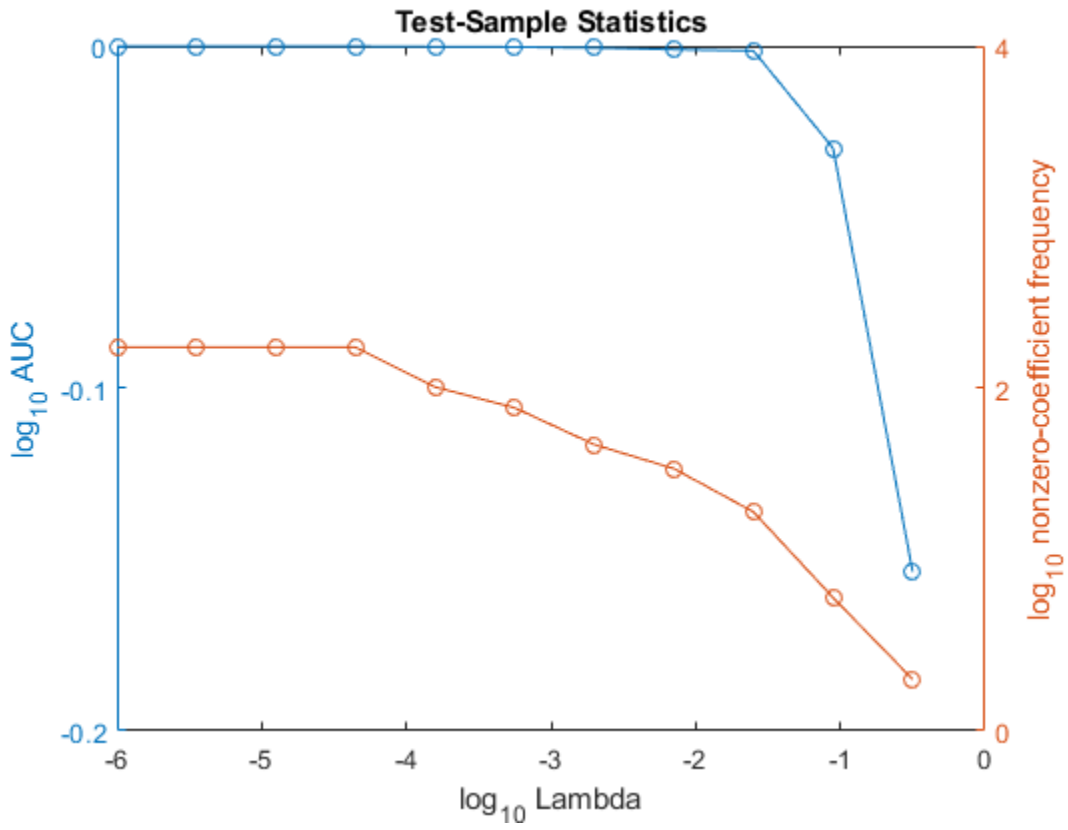
```
auc = 1:numel(Lambda);  % Preallocation
for j = 1:numel(Lambda)
    [~,~,~,auc(j)] = perfcurve(YTest,posterior(:,2,j),Mdl1.ClassNames(2));
end
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you trained the model. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the test-sample error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(auc),...
    log10(Lambda),log10(numNZCoeff + 1));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} AUC')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
title('Test-Sample Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and high AUC. In this case, a value between 10^{-2} to 10^{-1} should suffice.

```
idxFinal = 9;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

More About

Classification Score

For linear classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f_j(x) = x\beta_j + b_j.$$

For the model with regularization strength j, β_j is the estimated column vector of coefficients (the model property `Beta(:,j)`) and b_j is the estimated, scalar bias (the model property `Bias(j)`).

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `predict` does not support tall `table` data.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-419 of the `ClassificationLinear` object.

Argument	Notes and Limitations

X	• For general code generation, X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, X must be a single-precision or double-precision matrix.
	• The number of observations in X can be a variable size, but the number of variables in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

Name-value pair arguments	• Names in name-value pair arguments must be compile-time constants. • The value for the `'ObservationsIn'` name-value pair argument must be a compile-time constant. For example, to use the `'ObservationsIn','columns'` name-value pair argument in the generated code, include `{coder.Constant('ObservationsIn'),coder.Constant('columns')}` in the `-args` value of `codegen`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationLinear` | `confusionchart` | `fitclinear` | `loss` | `perfcurve` | `testcholdout`

Introduced in R2016a

predict

Predict labels using discriminant analysis classification model

Syntax

```
label = predict(Mdl,X)
[label,score,cost] = predict(Mdl,X)
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the trained discriminant analysis classification model `Mdl`.

`[label,score,cost] = predict(Mdl,X)` also returns:

- A matrix of classification scores (`score`) indicating the likelihood that a label comes from a particular class. For discriminant analysis, scores are posterior probabilities on page 33-4807.
- A matrix of expected classification cost on page 33-4808 (`cost`). For each observation in `X`, the predicted class label corresponds to the minimum expected classification cost among all classes.

Input Arguments

`Mdl` — Discriminant analysis classification model

`ClassificationDiscriminant` model object | `CompactClassificationDiscriminant` model object

Discriminant analysis classification model, specified as a `ClassificationDiscriminant` or `CompactClassificationDiscriminant` model object returned by `fitcdiscr`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable. All predictor variables in `X` must be numeric vectors.

- For a numeric matrix, the variables that compose the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
- For a table:
  - `predict` does not support multicolumn variables and cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

- If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitcdiscr`. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

Data Types: `table` | `double` | `single`

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical vector | vector of numeric values | cell array of character vectors

Predicted class labels on page 33-4808, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label`:

- Is the same data type as the observed class labels (`Y`) that trained `Mdl`. (The software treats string arrays as cell arrays of character vectors.)
- Has length equal to the number of rows of `X`.

`score` — Predicted class posterior probabilities

numeric matrix

Predicted class posterior probabilities on page 33-4807, returned as a numeric matrix of size `N`-by-`K`. `N` is the number of observations (rows) in `X`, and `K` is the number of classes (in `Mdl.ClassNames`). `score(i,j)` is the posterior probability that observation `i` in `X` is of class `j` in `Mdl.ClassNames`.

`cost` — Expected classification costs

numeric matrix

Expected classification costs on page 33-4808, returned as a matrix of size `N`-by-`K`. `N` is the number of observations (rows) in `X`, and `K` is the number of classes (in `Mdl.ClassNames`). `cost(i,j)` is the cost of classifying row `i` of `X` as class `j` in `Mdl.ClassNames`.

Examples

Predict Class Labels Using Discriminant Analysis Model

Load Fisher's iris data set. Determine the sample size.

```
load fisheriris
N = size(meas,1);
```

Partition the data into training and test sets. Hold out 10% of the data for testing.

```
rng(1); % For reproducibility
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Store the training data in a table.

```
tblTrn = array2table(meas(idxTrn,:));
tblTrn.Y = species(idxTrn);
```

Train a discriminant analysis model using the training set and default options.

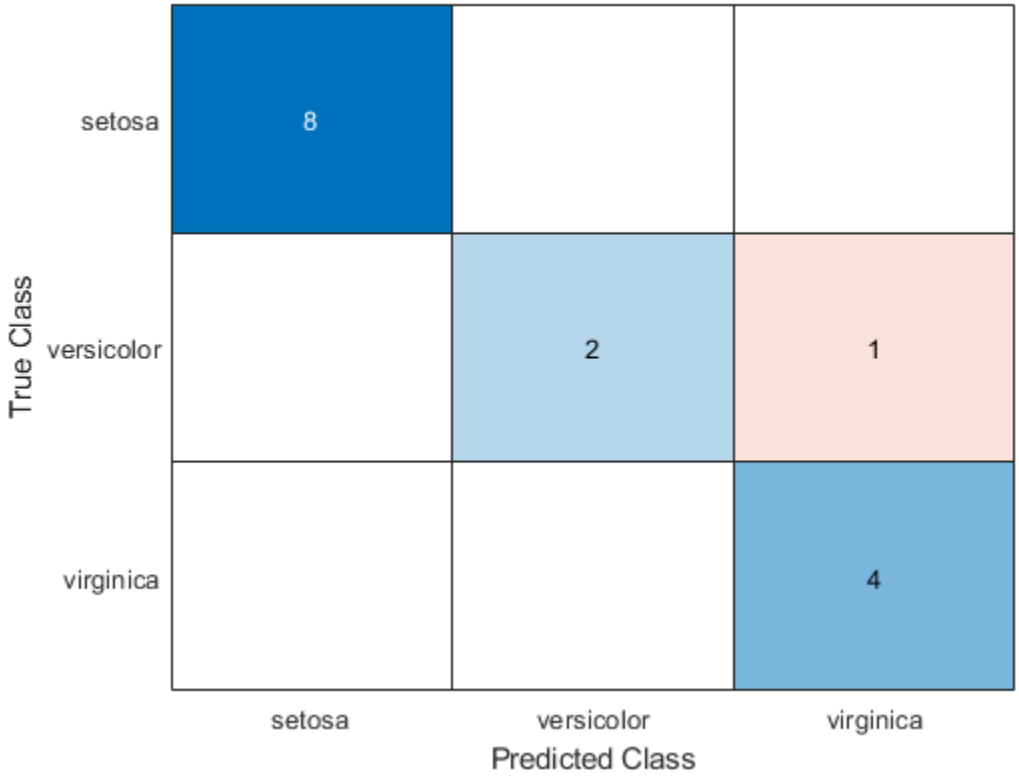
```
Mdl = fitcdiscr(tblTrn,'Y');
```

Predict labels for the test set. You trained `Mdl` using a table of data, but you can predict labels using a matrix.

```
labels = predict(Mdl,meas(idxTest,:));
```

Construct a confusion matrix for the test set.

```
confusionchart(species(idxTest),labels)
```



`Mdl` misclassifies one versicolor iris as virginica in the test set.

Plot Class Posterior Probability Regions

Load Fisher's iris data set. Consider training using the petal lengths and widths only.

```
load fisheriris
X = meas(:,3:4);
```

Train a quadratic discriminant analysis model using the entire data set.

```
Mdl = fitcdiscr(X,species,'DiscrimType','quadratic');
```

Define a grid of values in the observed predictor space. Predict the posterior probabilities for each instance in the grid.

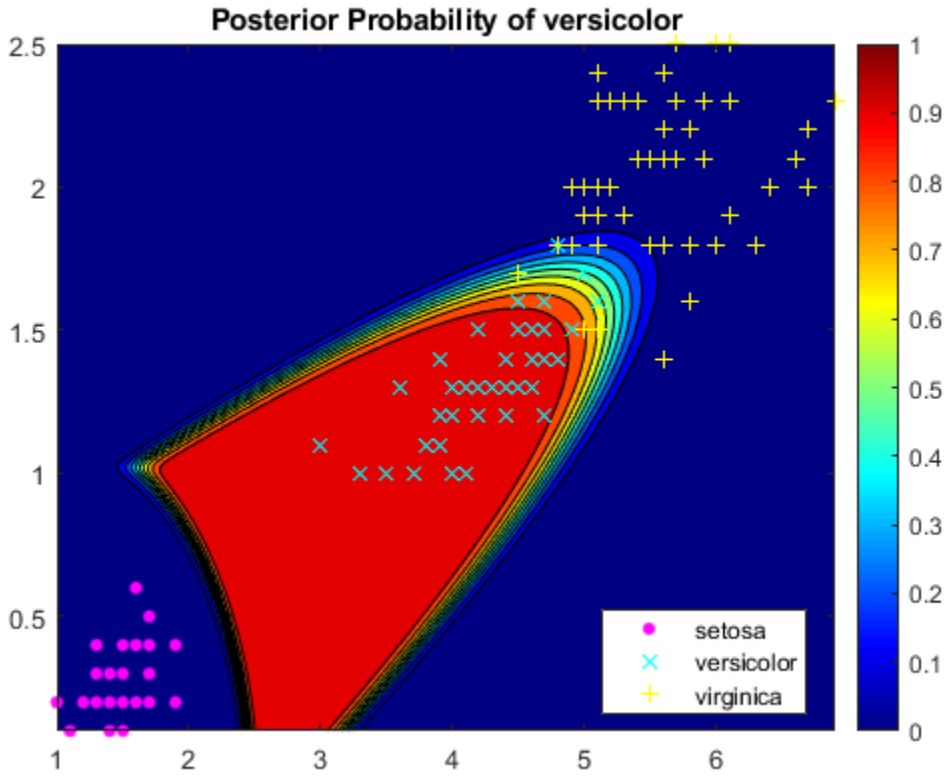
```
xMax = max(X);
xMin = min(X);
d = 0.01;
[x1Grid,x2Grid] = meshgrid(xMin(1):d:xMax(1),xMin(2):d:xMax(2));

[~,score] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

`score` is a matrix of class posterior probabilities. The columns correspond to the classes in `Mdl.ClassNames`. For example, `score(j,1)` is the posterior probability that observation `j` is a setosa iris.

Plot the posterior probability of versicolor classification for each observation in the grid and plot the training data.

```
figure;
contourf(x1Grid,x2Grid,reshape(score(:,2),size(x1Grid,1),size(x1Grid,2)));
h = colorbar;
caxis([0 1]);
colormap jet;
hold on
gscatter(X(:,1),X(:,2),species,'mcy','.x+');
axis tight
title('Posterior Probability of versicolor');
hold off
```



The posterior probability region exposes a portion of the decision boundary.

More About

Posterior Probability

The posterior probability that a point x belongs to class k is the product of the prior probability and the multivariate normal density. The density function of the multivariate normal with 1-by-d mean μ_k and d-by-d covariance Σ_k at a 1-by-d point x is

$$P(x|k) = \frac{1}{\left((2\pi)^d|\Sigma_k|\right)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^T\right),$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix.

Let $P(k)$ represent the prior probability of class k. Then the posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)},$$

where $P(x)$ is a normalization constant, the sum over k of $P(x|k)P(k)$.

Prior Probability

The prior probability is one of three choices:

- `'uniform'` — The prior probability of class `k` is one over the total number of classes.
- `'empirical'` — The prior probability of class `k` is the number of training samples of class `k` divided by the total number of training samples.
- Custom — The prior probability of class `k` is the `k`th element of the `prior` vector. See `fitcdiscr`.

After creating a classification model (`Mdl`) you can set the prior using dot notation:

```
Mdl.Prior = v;
```

where `v` is a vector of positive elements representing the frequency with which each element occurs. You do not need to retrain the classifier when you set a new prior.

Cost

The matrix of expected costs per observation is defined in “Cost” on page 20-7.

Predicted Class Label

`predict` classifies so as to minimize the expected classification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{k = 1}^{K} \widehat{P}(k|x) C(y|k),$$

where

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(k|x)$ is the posterior probability on page 20-6 of class k for observation x.
- $C(y|k)$ is the cost on page 20-7 of classifying an observation as y when its true class is k.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.

- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-748 of the `CompactClassificationDiscriminant` object.

`X`	• `X` must be a single-precision or double-precision matrix or a table containing numeric variables.
	• The number of rows, or observations, in `X` can be a variable size, but the number of columns in `X` must be fixed.
	• If you want to specify `X` as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`ClassificationDiscriminant` | `CompactClassificationDiscriminant` | `edge` | `fitcdiscr` | `loss` | `margin`

Topics

"Discriminant Analysis Classification" on page 20-2

Introduced in R2011b

predict

Package:

Classify observations using multiclass error-correcting output codes (ECOC) model

Syntax

```
label = predict(Mdl,X)
label = predict(Mdl,X,Name,Value)
[label,NegLoss,PBScore] = predict( ___ )
[label,NegLoss,PBScore,Posterior] = predict( ___ )
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels (`label`) for the predictor data in the table or matrix `X`, based on the trained multiclass error-correcting output codes (ECOC) model `Mdl`. The trained ECOC model can be either full or compact.

`label = predict(Mdl,X,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify the posterior probability estimation method, decoding scheme, and verbosity level.

`[label,NegLoss,PBScore] = predict( ___ )` uses any of the input argument combinations in the previous syntaxes and additionally returns:

- An array of negated average binary losses on page 33-4822 (`NegLoss`). For each observation in `X`, `predict` assigns the label of the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).
- An array of positive-class scores (`PBScore`) for the observations classified by each binary learner.

`[label,NegLoss,PBScore,Posterior] = predict( ___ )` additionally returns posterior class probability estimates for the observations (`Posterior`).

To obtain posterior class probabilities, you must set `'FitPosterior',true` when training the ECOC model using `fitcecoc`. Otherwise, `predict` throws an error.

Examples

Predict Test-Sample Labels of Training Data Using ECOC Model

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Specify a 30% holdout sample, standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.30,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

Predict the test-sample labels. Print a random subset of true and predicted labels.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
labels = predict(Mdl,XTest);

idx = randsample(sum(testInds),10);
table(YTest(idx),labels(idx),...
    'VariableNames',{'TrueLabels','PredictedLabels'})
```

```
ans=10×2 table
    TrueLabels    PredictedLabels
    __________    _______________

    setosa          setosa
    versicolor      virginica
    setosa          setosa
    virginica       virginica
    versicolor      versicolor
    setosa          setosa
    virginica       virginica
    virginica       virginica
    setosa          setosa
    setosa          setosa
```

`Mdl` correctly labels all except one of the test-sample observations with indices `idx`.

Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers and specify a 30% holdout sample. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
PMdl = fitcecoc(X,Y,'Holdout',0.30,'Learners',t,'ClassNames',classOrder);
Mdl = PMdl.Trained{1};           % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

SVM scores are signed distances from the observation to the decision boundary. Therefore, $(-\infty, \infty)$ is the domain. Create a custom binary loss function that does the following:

- Map the coding design matrix (M) and positive-class classification scores (s) for each learner to the binary loss for each observation.
- Use linear loss.
- Aggregate the binary learner loss using the median.

You can create a separate function for the binary loss function, and then save it on the MATLAB® path. Or, you can specify an anonymous binary loss function. In this case, create a function handle (`customBL`) to an anonymous binary loss function.

```
customBL = @(M,s) median(1 - bsxfun(@times,M,s),2,'omitnan')/2;
```

Predict test-sample labels and estimate the median binary loss per class. Print the median negative binary losses per class for a random set of 10 test-sample observations.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
[label,NegLoss] = predict(Mdl,XTest,'BinaryLoss',customBL);

idx = randsample(sum(testInds),10);
classOrder
```

```
classOrder = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
table(YTest(idx),label(idx),NegLoss(idx,:),'VariableNames',...
    {'TrueLabel','PredictedLabel','NegLoss'})
```

```
ans=10×3 table
    TrueLabel     PredictedLabel                NegLoss
    __________    ______________    ____________________________________

    setosa          versicolor        0.1858       1.9877      -3.6735
    versicolor      virginica        -1.3315     -0.12343    -0.045018
    setosa          versicolor       0.13891       1.9262      -3.5651
    virginica       virginica         -1.513     -0.38289      0.39594
    versicolor      versicolor      -0.87221      0.74785      -1.3756
    setosa          versicolor       0.48413        1.997      -3.9811
    virginica       virginica         -1.936      -0.6755       1.1115
    virginica       virginica        -1.5786     -0.83372      0.91236
    setosa          versicolor       0.51027       2.1206      -4.1309
    setosa          versicolor       0.36128       2.0594      -3.9207
```

The order of the columns corresponds to the elements of `classOrder`. The software predicts the label based on the maximum negated loss. The results indicate that the median of the linear losses might not perform as well as other losses.

Estimate Posterior Probabilities Using ECOC Classifier

Train an ECOC classifier using SVM binary learners. First predict the training-sample labels and class posterior probabilities. Then predict the maximum class posterior probability at each point in a grid. Visualize the results.

Load Fisher's iris data set. Specify the petal dimensions as the predictors and the species names as the response.

```
load fisheriris
X = meas(:,3:4);
Y = species;
rng(1); % For reproducibility
```

Create an SVM template. Standardize the predictors, and specify the Gaussian kernel.

```
t = templateSVM('Standardize',true,'KernelFunction','gaussian');
```

`t` is an SVM template. Most of its properties are empty. When the software trains the ECOC classifier, it sets the applicable properties to their default values.

Train the ECOC classifier using the SVM template. Transform classification scores to class posterior probabilities (which are returned by `predict` or `resubPredict`) using the `'FitPosterior'` name-value pair argument. Specify the class order using the `'ClassNames'` name-value pair argument. Display diagnostic messages during training by using the `'Verbose'` name-value pair argument.

```
Mdl = fitcecoc(X,Y,'Learners',t,'FitPosterior',true,...
    'ClassNames',{'setosa','versicolor','virginica'},...
    'Verbose',2);

Training binary learner 1 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 2
Positive class indices: 1

Fitting posterior probabilities for learner 1 (SVM).
Training binary learner 2 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 1

Fitting posterior probabilities for learner 2 (SVM).
Training binary learner 3 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 2

Fitting posterior probabilities for learner 3 (SVM).
```

`Mdl` is a `ClassificationECOC` model. The same SVM template applies to each binary learner, but you can adjust options for each binary learner by passing in a cell vector of templates.

Predict the training-sample labels and class posterior probabilities. Display diagnostic messages during the computation of labels and class posterior probabilities by using the `'Verbose'` name-value pair argument.

```
[label,~,~,Posterior] = resubPredict(Mdl,'Verbose',1);
```

```
Predictions from all learners have been computed.
Loss for all observations has been computed.
Computing posterior probabilities...
```

```
Mdl.BinaryLoss
```

```
ans =
'quadratic'
```

The software assigns an observation to the class that yields the smallest average binary loss. Because all binary learners are computing posterior probabilities, the binary loss function is `quadratic`.

Display a random set of results.

```
idx = randsample(size(X,1),10,1);
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
table(Y(idx),label(idx),Posterior(idx,:),...
    'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

```
ans=10×3 table
      TrueLabel         PredLabel                    Posterior
    ______________    ______________    ____________________________________

    {'virginica' }    {'virginica' }     0.0039319     0.0039866       0.99208
    {'virginica' }    {'virginica' }      0.017066      0.018262       0.96467
    {'virginica' }    {'virginica' }      0.014947      0.015855        0.9692
    {'versicolor'}    {'versicolor'}    2.2197e-14       0.87318       0.12682
    {'setosa'    }    {'setosa'    }         0.999    0.00025091    0.00074639
    {'versicolor'}    {'virginica' }    2.2195e-14      0.059427       0.94057
    {'versicolor'}    {'versicolor'}    2.2194e-14       0.97002      0.029984
    {'setosa'    }    {'setosa'    }         0.999     0.0002499    0.00074741
    {'versicolor'}    {'versicolor'}     0.0085638       0.98259     0.0088482
    {'setosa'    }    {'setosa'    }         0.999    0.00025013    0.00074718
```

The columns of `Posterior` correspond to the class order of `Mdl.ClassNames`.

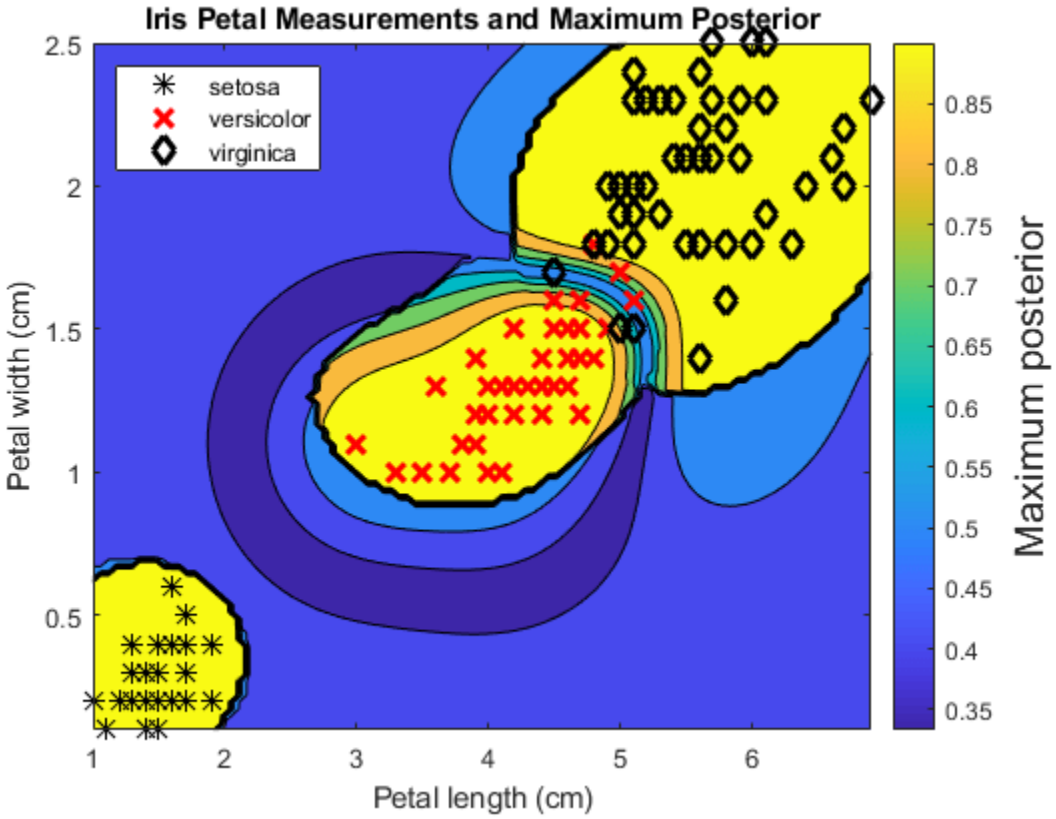
Define a grid of values in the observed predictor space. Predict the posterior probabilities for each instance in the grid.

```
xMax = max(X);
xMin = min(X);

x1Pts = linspace(xMin(1),xMax(1));
x2Pts = linspace(xMin(2),xMax(2));
[x1Grid,x2Grid] = meshgrid(x1Pts,x2Pts);
```

```
[~,~,~,PosteriorRegion] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
```

For each coordinate on the grid, plot the maximum class posterior probability among all classes.

```
contourf(x1Grid,x2Grid,...
        reshape(max(PosteriorRegion,[],2),size(x1Grid,1),size(x1Grid,2)));
h = colorbar;
h.YLabel.String = 'Maximum posterior';
h.YLabel.FontSize = 15;

hold on
gh = gscatter(X(:,1),X(:,2),Y,'krk','*xd',8);
gh(2).LineWidth = 2;
gh(3).LineWidth = 2;

title('Iris Petal Measurements and Maximum Posterior')
xlabel('Petal length (cm)')
ylabel('Petal width (cm)')
axis tight
legend(gh,'Location','NorthWest')
hold off
```



Estimate Test-Sample Posterior Probabilities Using Parallel Computing

Train a multiclass ECOC model and estimate posterior probabilities using parallel computing.

Load the `arrhythmia` data set. Examine the response data `Y`, and determine the number of classes.

```
load arrhythmia
Y = categorical(Y);
tabulate(Y)
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

```
K = numel(unique(Y));
```

Several classes are not represented in the data, and many of the other classes have low relative frequencies.

Specify an ensemble learning template that uses the GentleBoost method and 50 weak classification tree learners.

```
t = templateEnsemble('GentleBoost',50,'Tree');
```

`t` is a template object. Most of its properties are empty (`[]`). The software uses default values for all empty properties during training.

Because the response variable contains many classes, specify a sparse random coding design.

```
rng(1); % For reproducibility
Coding = designecoc(K,'sparserandom');
```

Train an ECOC model using parallel computing. Specify a 15% holdout sample, and fit posterior probabilities.

```
pool = parpool;                    % Invokes workers
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
options = statset('UseParallel',true);
PMdl = fitcecoc(X,Y,'Learner',t,'Options',options,'Coding',Coding,...
    'FitPosterior',true,'Holdout',0.15);
Mdl = PMdl.Trained{1};             % Extract trained, compact classifier
```

`PMdl` is a `ClassificationPartitionedECOC` model. It has the property `Trained`, a 1-by-1 cell array containing the `CompactClassificationECOC` model that the software trained using the training set.

The pool invokes six workers, although the number of workers might vary among systems.

Estimate posterior probabilities, and display the posterior probability of being classified as not having arrhythmia (class 1) given the data for a random set of test-sample observations.

```
testInds = test(PMdl.Partition);  % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
[~,~,~,posterior] = predict(Mdl,XTest,'Options',options);

idx = randsample(sum(testInds),10);
table(idx,YTest(idx),posterior(idx,1),...
    'VariableNames',{'TestSampleIndex','TrueLabel','PosteriorNoArrhythmia'})
```

```
ans=10×3 table
    TestSampleIndex    TrueLabel    PosteriorNoArrhythmia
    _______________    _________    _____________________

          11               6               0.60631
          41               4               0.23674
          51               2               0.13802
          33              10               0.43831
          12               1               0.94332
           8               1               0.97278
          37               1               0.62807
          24              10               0.96876
          56              16               0.29375
          30               1               0.64512
```

Input Arguments

`Mdl` — Full or compact multiclass ECOC model

`ClassificationECOC` model object | `CompactClassificationECOC` model object

Full or compact multiclass ECOC model, specified as a `ClassificationECOC` or `CompactClassificationECOC` model object.

To create a full or compact ECOC model, see `ClassificationECOC` or `CompactClassificationECOC`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

By default, each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that constitute the columns of `X` must have the same order as the predictor variables that train `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical

during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitcecoc`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.

- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the predictor variables that train `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Both `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and the corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitcecoc`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Note If `Mdl.BinaryLearners` contains linear classification models (`ClassificationLinear`), then you can orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`. However, you cannot specify `'ObservationsIn','columns'` for predictor data in a table.

When training `Mdl`, assume that you set `'Standardize',true` for a template object specified in the `'Learners'` name-value pair argument of `fitcecoc`. In this case, for the corresponding binary learner `j`, the software standardizes the columns of the new predictor data using the corresponding means in `Mdl.BinaryLearner{j}.Mu` and standard deviations in `Mdl.BinaryLearner{j}.Sigma`.

Data Types: `table` | `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `predict(Mdl,X,'BinaryLoss','quadratic','Decoding','lossbased')` specifies a quadratic binary learner loss function and a loss-based decoding scheme for aggregating the binary losses.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	$[0,1]$ or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_js_j)/2$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	$[0,1]$	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.
  - `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
  - K is the number of classes.
  - L is the number of binary learners.

  For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-4822.

Example: `'Decoding','lossbased'`

`NumKLInitializations` — Number of random initial values
`0` (default) | nonnegative integer scalar

Number of random initial values for fitting posterior probabilities by Kullback-Leibler divergence minimization, specified as the comma-separated pair consisting of `'NumKLInitializations'` and a nonnegative integer scalar.

If you do not request the fourth output argument (`Posterior`) and set `'PosteriorMethod','kl'` (the default), then the software ignores the value of `NumKLInitializations`.

For more details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-4824.

Example: `'NumKLInitializations',5`

Data Types: `single` | `double`

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`. `Mdl.BinaryLearners` must contain `ClassificationLinear` models.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, you can experience a significant reduction in execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`PosteriorMethod` — Posterior probability estimation method
`'kl'` (default) | `'qp'`

Posterior probability estimation method, specified as the comma-separated pair consisting of `'PosteriorMethod'` and `'kl'` or `'qp'`.

- If `PosteriorMethod` is `'kl'`, then the software estimates multiclass posterior probabilities by minimizing the Kullback-Leibler divergence between the predicted and expected posterior probabilities returned by binary learners. For details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-4824.
- If `PosteriorMethod` is `'qp'`, then the software estimates multiclass posterior probabilities by solving a least-squares problem using quadratic programming. You need an Optimization Toolbox license to use this option. For details, see “Posterior Estimation Using Quadratic Programming” on page 33-4825.
- If you do not request the fourth output argument (`Posterior`), then the software ignores the value of `PosteriorMethod`.

Example: `'PosteriorMethod','qp'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical array | numeric array | cell array of character vectors

Predicted class labels, returned as a categorical, character, logical, or numeric array, or a cell array of character vectors. The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`label` has the same data type as the class labels used to train `Mdl` and has the same number of rows as `X`. (The software treats string arrays as cell arrays of character vectors.)

If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `label` is an *m*-by-*L* matrix, where *m* is the number of observations in `X`, and *L* is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`). The value `label(i,j)` is the predicted label of observation `i` for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.

Otherwise, `label` is a column vector of length *m*.

`NegLoss` — Negated average binary losses
numeric matrix | numeric array

Negated average binary losses on page 33-4822, returned as a numeric matrix or array.

- If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `NegLoss` is an *m*-by-*K*-by-*L* array.

  - *m* is the number of observations in `X`.
  - *K* is the number of distinct classes in the training data (`numel(Mdl.ClassNames)`).
  - *L* is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`).

  `NegLoss(i,k,j)` is the negated average binary loss for observation `i`, corresponding to class `Mdl.ClassNames(k)`, for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.
- Otherwise, `NegLoss` is an *m*-by-*K* matrix.

`PBScore` — Positive-class scores
numeric matrix | numeric array

Positive-class scores for each binary learner, returned as a numeric matrix or array.

- If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `PBScore` is an *m*-by-*B*-by-*L* array.

  - *m* is the number of observations in `X`.
  - *B* is the number of binary learners (`numel(Mdl.BinaryLearners)`).
  - *L* is the number of regularization strengths in the linear classification models (`numel(Mdl.BinaryLearners{1}.Lambda)`).

  `PBScore(i,b,j)` is the positive-class score for observation `i`, using binary learner `b`, for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.
- Otherwise, `PBScore` is an *m*-by-*B* matrix.

`Posterior` — Posterior class probabilities
numeric matrix | numeric array

Posterior class probabilities, returned as a numeric matrix or array.

- If `Mdl.BinaryLearners` contains `ClassificationLinear` models, then `Posterior` is an *m*-by-*K*-by-*L* array. For dimension definitions, see `NegLoss`. `Posterior(i,k,j)` is the posterior probability that observation `i` comes from class `Mdl.ClassNames(k)`, for the model trained using regularization strength `Mdl.BinaryLearners{1}.Lambda(j)`.
- Otherwise, `Posterior` is an *m*-by-*K* matrix.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).

- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Algorithms

The software can estimate class posterior probabilities by minimizing the Kullback-Leibler divergence or by using quadratic programming. For the following descriptions of the posterior estimation algorithms, assume that:

- m_{kj} is the element (k,j) of the coding design matrix M.
- I is the indicator function.
- $\widehat{p}_k$ is the class posterior probability estimate for class k of an observation, $k = 1,...,K$.
- r_j is the positive-class posterior probability for binary learner j. That is, r_j is the probability that binary learner j classifies an observation into the positive class, given the training data.

Posterior Estimation Using Kullback-Leibler Divergence

By default, the software minimizes the Kullback-Leibler divergence to estimate class posterior probabilities. The Kullback-Leibler divergence between the expected and observed positive-class posterior probabilities is

$$\Delta(r,\widehat{r}) = \sum_{j=1}^{L} w_j \left[r_j \log\frac{r_j}{\widehat{r}_j} + (1-r_j)\log\frac{1-r_j}{1-\widehat{r}_j} \right],$$

where $w_j = \sum_{S_j} w_i^*$ is the weight for binary learner j.

- S_j is the set of observation indices on which binary learner j is trained.
- w_i^* is the weight of observation i.

The software minimizes the divergence iteratively. The first step is to choose initial values $\widehat{p}_k^{(0)};\ k = 1,...,K$ for the class posterior probabilities.

- If you do not specify `'NumKLIterations'`, then the software tries both sets of deterministic initial values described next, and selects the set that minimizes Δ.
  - $\widehat{p}_k^{(0)} = 1/K;\ k = 1,...,K$.
  - $\widehat{p}_k^{(0)};\ k = 1,...,K$ is the solution of the system

    $$M_{01}\widehat{p}^{(0)} = r,$$

    where M_{01} is M with all $m_{kj} = -1$ replaced with 0, and r is a vector of positive-class posterior probabilities returned by the L binary learners [Dietterich et al.] on page 18-279. The software uses `lsqnonneg` to solve the system.
- If you specify `'NumKLIterations',c`, where `c` is a natural number, then the software does the following to choose the set $\widehat{p}_k^{(0)};\ k = 1,...,K$, and selects the set that minimizes Δ.
  - The software tries both sets of deterministic initial values as described previously.
  - The software randomly generates `c` vectors of length K using `rand`, and then normalizes each vector to sum to 1.

At iteration t, the software completes these steps:

1 Compute

$$\widehat{r}_j^{(t)} = \frac{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1)}{\sum_{k=1}^{K} \widehat{p}_k^{(t)} I(m_{kj} = +1 \cup m_{kj} = -1)}.$$

2 Estimate the next class posterior probability using

$$\widehat{p}_k^{(t+1)} = \widehat{p}_k^{(t)} \frac{\sum_{j=1}^{L} w_j \left[r_j I\left(m_{kj} = +1\right) + \left(1 - r_j\right) I\left(m_{kj} = -1\right) \right]}{\sum_{j=1}^{L} w_j \left[\widehat{r}_j^{(t)} I\left(m_{kj} = +1\right) + \left(1 - \widehat{r}_j^{(t)}\right) I\left(m_{kj} = -1\right) \right]}.$$

3 Normalize $\widehat{p}_k^{(t+1)};\ k = 1, \ldots, K$ so that they sum to 1.

4 Check for convergence.

For more details, see [Hastie et al.] on page 18-280 and [Zadrozny] on page 18-281.

Posterior Estimation Using Quadratic Programming

Posterior probability estimation using quadratic programming requires an Optimization Toolbox license. To estimate posterior probabilities for an observation using this method, the software completes these steps:

1 Estimate the positive-class posterior probabilities, r_j, for binary learners $j = 1,\ldots,L$.

2 Using the relationship between r_j and $\widehat{p}_k$ [Wu et al.] on page 18-281, minimize

$$\sum_{j=1}^{L} \left[-r_j \sum_{k=1}^{K} \widehat{p}_k I\left(m_{kj} = -1\right) + \left(1 - r_j\right) \sum_{k=1}^{K} \widehat{p}_k I\left(m_{kj} = +1\right) \right]^2$$

with respect to $\widehat{p}_k$ and the restrictions

$$0 \le \widehat{p}_k \le 1$$

$$\sum_k \widehat{p}_k = 1.$$

The software performs minimization using `quadprog`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recognition*. Vol. 30, Issue 3, 2009, pp. 285–297.

[5] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[6] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[7] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `predict` does not support tall `table` data when `Mdl` contains kernel or linear binary learners.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-757 of the `CompactClassificationECOC` object.

Argument	Notes and Limitations

X	• For general code generation, X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, X must be a single-precision or double-precision matrix.
	• The number of observations in X can be a variable size, but the number of variables in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	  For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

`Posterior`	This output argument is not supported.
Name-value pair arguments	Names in name-value pair arguments must be compile-time constants.
`BinaryLoss`	• The value for the `'BinaryLoss'` name-value pair argument must be a compile-time constant. For example, to use the `'BinaryLoss','logit'` name-value pair argument in the generated code, include `{coder.Constant('BinaryLoss'),coder.Constant('logit')}` in the `-args` value of `codegen`. • To set the `'BinaryLoss'` name-value pair argument to a custom binary loss function in the generated code, define a custom function on the MATLAB search path, and specify the name of the custom function instead of its function handle. The custom function name must be a compile-time constant. For example, if you define a custom function named `customFunction`, then include `{coder.Constant('BinaryLoss'),coder.Constant('customFunction')}` in the `-args` value of `codegen`.
`NumKLInitializations`	This name-value pair argument is not supported.
`ObservationsIn`	The value for the `'ObservationsIn'` name-value pair argument must be a compile-time constant. For example, to use the `'ObservationsIn','columns'` name-value pair argument in the generated code, include `{coder.Constant('ObservationsIn'),coder.Constant('columns')}` in the `-args` value of `codegen`.

Argument	Notes and Limitations
`Options`	This name-value pair argument is not supported.
`PosteriorMethod`	This name-value pair argument is not supported.
`Verbose`	If you plan to generate a MEX file without using a coder configurer, then you can specify `Verbose`. Otherwise, `codegen` does not support `Verbose`.

For more information, see “Introduction to Code Generation” on page 32-2.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `CompactClassificationECOC` | `fitcecoc` | `loss` | `quadprog` | `resubPredict` | `statset`

Topics
“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

predict

Classify observations using ensemble of classification models

Syntax

```
labels = predict(Mdl,X)
labels = predict(Mdl,X,Name,Value)
[labels,scores] = predict( ___ )
```

Description

`labels = predict(Mdl,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the full or compact, trained classification ensemble `Mdl`.

`labels = predict(Mdl,X,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments.

`[labels,scores] = predict( ___ )` also returns a matrix of classification scores on page 33-4832 (`scores`), indicating the likelihood that a label comes from a particular class, using any of the input arguments in the previous syntaxes. For each observation in `X`, the predicted class label corresponds to the maximum score among all classes.

Input Arguments

`Mdl`

A classification ensemble created by `fitcensemble` or a compact classification ensemble created by `compact`.

`X`

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitcensemble`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

- If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and be of the same data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
- If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitcensemble`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Learners`

Indices of weak learners `predict` uses for computation of responses, a numeric vector.

Default: `1:T`, where `T` is the number of weak learners in `Mdl`

`UseObsForLearner`

A logical matrix of size `N`-by-`T`, where:

- `N` is the number of rows of `X`.
- `T` is the number of weak learners in `Mdl`.

When `UseObsForLearner(i,j)` is `true`, learner `j` is used in predicting the class of row `i` of `X`.

Default: `true(N,T)`

Output Arguments

`labels`

Vector of classification labels. `labels` has the same data type as the labels used in training `Mdl`. (The software treats string arrays as cell arrays of character vectors.)

`scores`

A matrix with one row per observation and one column per class. For each observation and each class, the score represents the confidence that the observation originates from that class. A higher score indicates a higher confidence. For more information, see “Score (ensemble)” on page 33-4832.

Examples

Predict Class Labels Using Classification Ensemble

Load Fisher's iris data set. Determine the sample size.

```
load fisheriris
N = size(meas,1);
```

Partition the data into training and test sets. Hold out 10% of the data for testing.

```
rng(1); % For reproducibility
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Store the training data in a table.

```
tblTrn = array2table(meas(idxTrn,:));
tblTrn.Y = species(idxTrn);
```

Train a classification ensemble using AdaBoostM2 and the training set. Specify tree stumps as the weak learners.

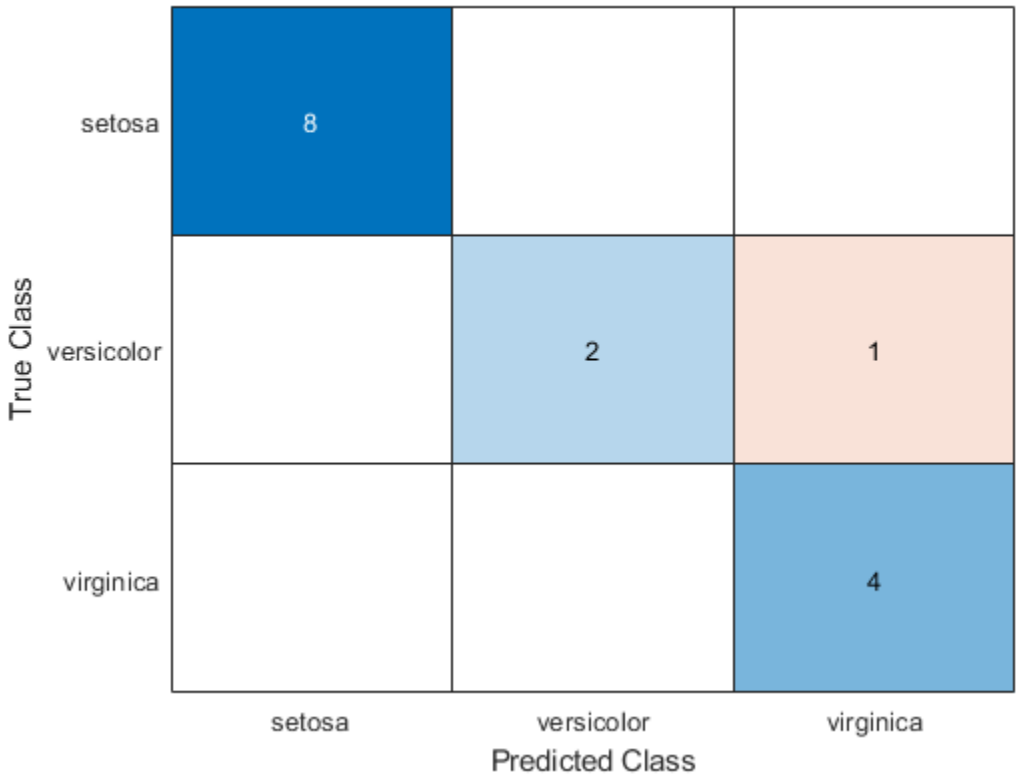
```
t = templateTree('MaxNumSplits',1);
Mdl = fitcensemble(tblTrn,'Y','Method','AdaBoostM2','Learners',t);
```

Predict labels for the test set. You trained model using a table of data, but you can predict labels using a matrix.

```
labels = predict(Mdl,meas(idxTest,:));
```

Construct a confusion matrix for the test set.

```
confusionchart(species(idxTest),labels)
```



`Mdl` misclassifies one versicolor iris as virginica in the test set.

More About

Score (ensemble)

For ensembles, a classification score represents the confidence that an observation originates from a specific class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `Bag` scores range from `0` to `1`. You can interpret these scores as probabilities averaged over all the trees in the ensemble.
- `AdaBoostM1`, `GentleBoost`, and `LogitBoost` scores range from –∞ to ∞. You can convert these scores to probabilities by setting the `ScoreTransform` property of `Mdl` to `'doublelogit'` before passing `Mdl` to `predict`:

  ```
  Mdl.ScoreTransform = 'doublelogit';
  [labels,scores] = predict(Mdl,X);
  ```

  Alternatively, you can specify `'ScoreTransform','doublelogit'` in the call to `fitcensemble` when you create `Mdl`.

For more information on the different ensemble algorithms and how they compute scores, see “Ensemble Algorithms” on page 18-39.

Alternative Functionality

Simulink Block

To integrate the prediction of an ensemble into Simulink, you can use the ClassificationEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see “Predict Class Labels Using ClassificationEnsemble Predict Block” on page 32-130 and “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- Generating fixed-point code for `predict` includes propagating data types for individual learners and, therefore, can be time consuming.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-763 of the `CompactClassificationEnsemble` object.

`X`	• For general code generation, `X` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• For fixed-point code generation, `X` must be a fixed-point matrix.
	• The number of rows, or observations, in `X` can be a variable size, but the number of columns in `X` must be fixed.
	• If you want to specify `X` as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

Argument	Notes and Limitations
Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow user-defined indices up to 5 weak learners in the generated code, include `{coder.Constant('Learners'),coder.typeof(0,[1,5],[0,1])}` in the `-args` value of `codegen`.
`'Learners'`	For fixed-point code generation, the `'Learners'` value must have an integer data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationBaggedEnsemble` | `ClassificationEnsemble` | `CompactClassificationEnsemble` | `edge` | `fitcensemble` | `loss` | `margin`

Introduced in R2011a

predict

Classify observations using naive Bayes classifier

Syntax

```
label = predict(Mdl,X)
[label,Posterior,Cost] = predict(Mdl,X)
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the trained naive Bayes classification model `Mdl`. The trained naive Bayes model can either be full or compact.

`[label,Posterior,Cost] = predict(Mdl,X)` also returns the "Posterior Probability" on page 33-4842 (`Posterior`) and predicted (expected) "Misclassification Cost" on page 33-4842 (`Cost`) corresponding to the observations (rows) in `Mdl.X`. For each observation in `X`, the predicted class label corresponds to the minimum expected classification cost among all classes.

Examples

Label Test Sample Observations of Naive Bayes Classifier

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Randomly partition observations into a training set and a test set with stratification, using the class information in `Y`. Specify a 30% holdout sample for testing.

```
cv = cvpartition(Y,'HoldOut',0.30);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Specify the training and test data sets.

```
XTrain = X(trainInds,:);
YTrain = Y(trainInds);
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a naive Bayes classifier using the predictors `XTrain` and class labels `YTrain`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(XTrain,YTrain,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 105
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Predict the test sample labels.

```
idx = randsample(sum(testInds),10);
label = predict(Mdl,XTest);
```

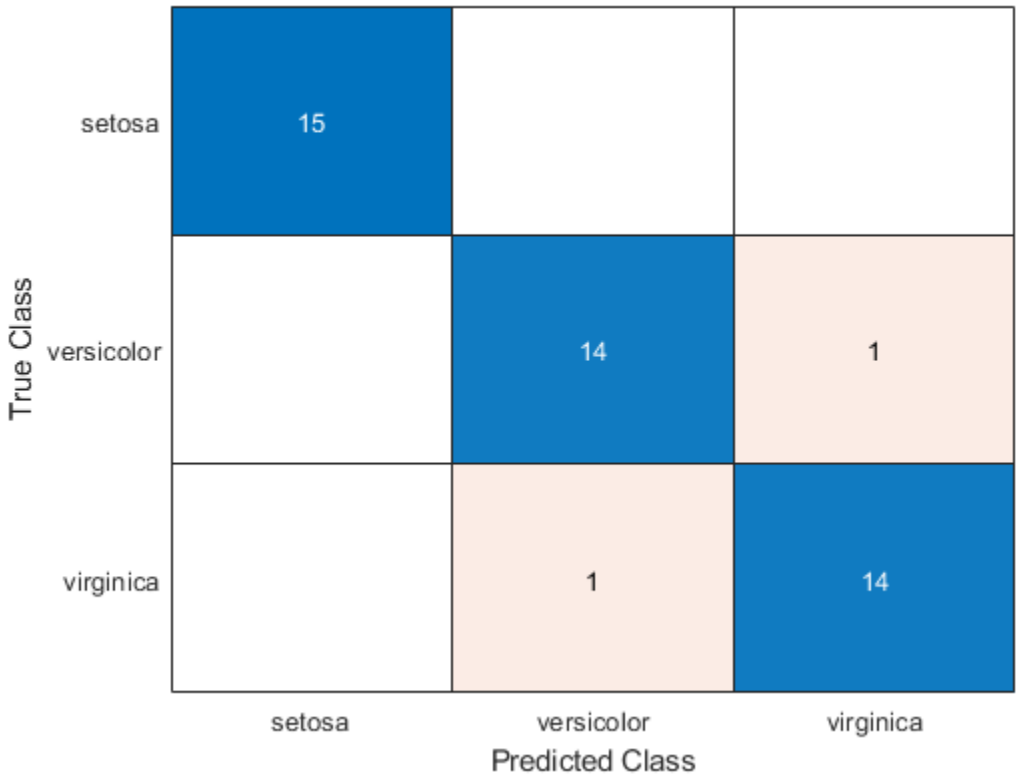
Display the results for a random set of 10 observations in the test sample.

```
table(YTest(idx),label(idx),'VariableNames',...
    {'TrueLabel','PredictedLabel'})
```

```
ans=10×2 table
      TrueLabel       PredictedLabel
    ______________    ______________

    {'virginica' }    {'virginica' }
    {'versicolor'}    {'versicolor'}
    {'versicolor'}    {'versicolor'}
    {'virginica' }    {'virginica' }
    {'setosa'    }    {'setosa'    }
    {'virginica' }    {'virginica' }
    {'setosa'    }    {'setosa'    }
    {'versicolor'}    {'versicolor'}
    {'versicolor'}    {'virginica' }
    {'versicolor'}    {'versicolor'}
```

Create a confusion chart from the true labels `YTest` and the predicted labels `label`.

```
cm = confusionchart(YTest,label);
```



Estimate Posterior Probabilities and Misclassification Costs

Estimate posterior probabilities and misclassification costs for new observations using a naive Bayes classifier. Classify new observations using a memory-efficient pretrained classifier.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default')  % for reproducibility
```

Partition the data set into two sets: one contains the training set, and the other contains new, unobserved data. Reserve 10 observations for the new data set.

```
n = size(X,1);
newInds = randsample(n,10);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a naive Bayes classifier using the predictors X and class labels Y. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X(inds,:),Y(inds),...
    'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Conserve memory by reducing the size of the trained naive Bayes classifier.

```
CMdl = compact(Mdl);
whos('Mdl','CMdl')
```

```
  Name      Size            Bytes  Class                                                                A

  CMdl      1x1              5406  classreg.learning.classif.CompactClassificationNaiveBayes
  Mdl       1x1             12707  ClassificationNaiveBayes
```

`CMdl` is a `CompactClassificationNaiveBayes` classifier. It uses less memory than `Mdl` because `Mdl` stores the data.

Display the class names of `CMdl` using dot notation.

```
CMdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

Predict the labels. Estimate the posterior probabilities and expected class misclassification costs.

```
[labels,PostProbs,MisClassCost] = predict(CMdl,XNew);
```

Compare the true labels with the predicted labels.

```
table(YNew,labels,PostProbs,MisClassCost,'VariableNames',...
    {'TrueLabels','PredictedLabels',...
    'PosteriorProbabilities','MisclassificationCosts'})
```

```
ans=10×4 table
      TrueLabels      PredictedLabels            PosteriorProbabilities                  Mis
    _____________    _______________    ___________________________________________    _________

    {'virginica' }    {'virginica' }    4.0832e-268    4.6422e-09             1             1
    {'setosa'    }    {'setosa'    }              1    3.0706e-18    4.6719e-25    3.0706e-18
    {'virginica' }    {'virginica' }    1.0007e-246    5.8758e-10             1             1
    {'versicolor'}    {'versicolor'}     1.2022e-61       0.99995    4.9859e-05             1
    {'virginica' }    {'virginica' }     2.687e-226    1.7905e-08             1             1
    {'versicolor'}    {'versicolor'}     3.3431e-76       0.99971    0.00028983             1
    {'virginica' }    {'virginica' }      4.05e-166     0.0028527       0.99715             1
    {'setosa'    }    {'setosa'    }              1    1.1272e-14    2.0308e-23    1.1272e-14
    {'virginica' }    {'virginica' }    1.3292e-228    8.3604e-10             1             1
    {'setosa'    }    {'setosa'    }              1    4.5023e-17    2.1724e-24    4.5023e-17
```

`PostProbs` and `MisClassCost` are `10`-by-`3` numeric matrices, where each row corresponds to a new observation and each column corresponds to a class. The order of the columns corresponds to the order of `CMdl.ClassNames`.

Plot Posterior Probability Regions for Naive Bayes Classifier

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four petal measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas(:,3:4);
Y = species;
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'});
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Define a grid of values in the observed predictor space.

```
xMax = max(X);
xMin = min(X);
h = 0.01;
[x1Grid,x2Grid] = meshgrid(xMin(1):h:xMax(1),xMin(2):h:xMax(2));
```

Predict the posterior probabilities for each instance in the grid.

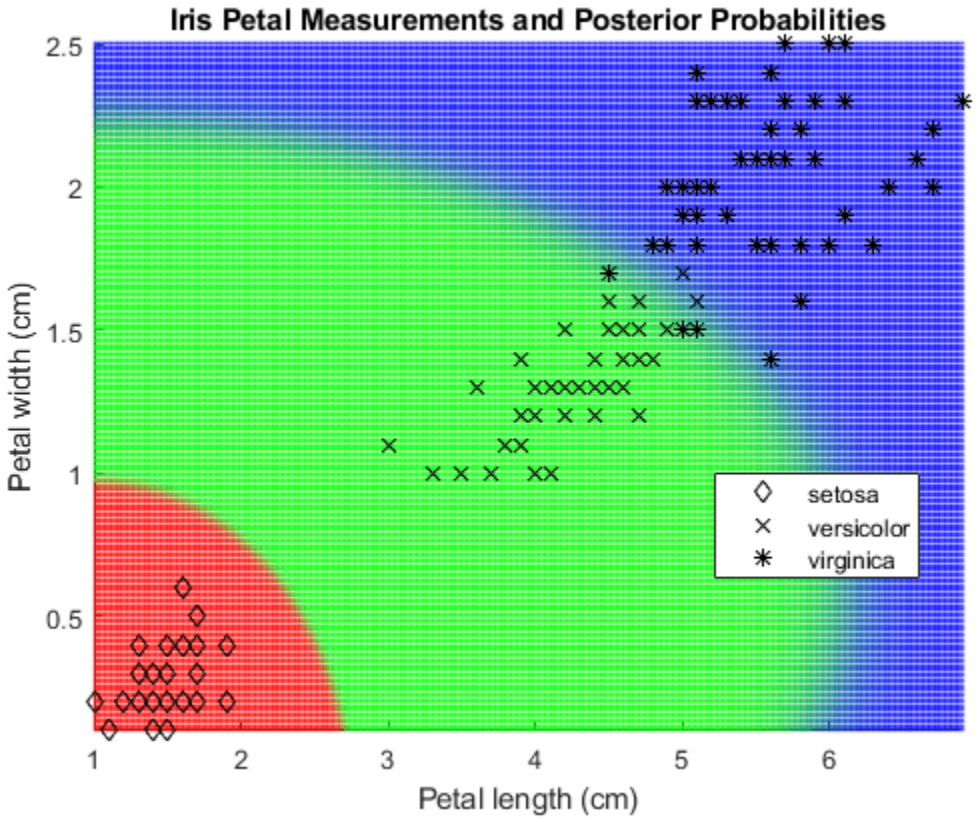
```
[~,PosteriorRegion] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
```

Plot the posterior probability regions and the training data.

```
h = scatter(x1Grid(:),x2Grid(:),1,PosteriorRegion);
h.MarkerEdgeAlpha = 0.3;
```

Plot the data.

```
hold on
gh = gscatter(X(:,1),X(:,2),Y,'k','dx*');
title 'Iris Petal Measurements and Posterior Probabilities';
xlabel 'Petal length (cm)';
ylabel 'Petal width (cm)';
axis tight
legend(gh,'Location','Best')
hold off
```



Input Arguments

`Mdl` — Naive Bayes classification model

`ClassificationNaiveBayes` model object | `CompactClassificationNaiveBayes` model object

Naive Bayes classification model, specified as a `ClassificationNaiveBayes` model object or `CompactClassificationNaiveBayes` model object returned by `fitcnb` or `compact`, respectively.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that make up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains only numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `'CategoricalPredictors'` name-value pair argument of `fitcnb`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.

- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` must be the same as the corresponding predictor variable names in `X`. To specify predictor names during training, use the `'PredictorNames'` name-value pair argument of `fitcnb`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Data Types: `table` | `double` | `single`

Notes:

- If `Mdl.DistributionNames` is `'mn'`, then the software returns `NaN`s corresponding to rows of `X` that contain at least one `NaN`.
- If `Mdl.DistributionNames` is not `'mn'`, then the software ignores `NaN` values when estimating misclassification costs and posterior probabilities. Specifically, the software computes the conditional density of the predictors given the class by leaving out the factors corresponding to missing predictor values.
- For predictor distribution specified as `'mvmn'`, if `X` contains levels that are not represented in the training data (that is, not in `Mdl.CategoricalLevels` for that predictor), then the conditional density of the predictors given the class is 0. For those observations, the software returns the corresponding value of `Posterior` as a `NaN`. The software determines the class label for such observations using the class prior probability stored in `Mdl.Prior`.

Output Arguments

`label` — Predicted class labels

categorical vector | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical vector, character array, logical or numeric vector, or cell array of character vectors.

The predicted class labels have the following:

- Same data type as the observed class labels (`Mdl.Y`). (The software treats string arrays as cell arrays of character vectors.)
- Length equal to the number of rows of `Mdl.X`.
- Class yielding the lowest expected misclassification cost (`Cost`).

`Posterior` — Class posterior probability

numeric matrix

Class “Posterior Probability” on page 33-4842, returned as a numeric matrix. `Posterior` has rows equal to the number of rows of `Mdl.X` and columns equal to the number of distinct classes in the training data (`size(Mdl.ClassNames,1)`).

`Posterior(j,k)` is the predicted posterior probability of class `k` (in class `Mdl.ClassNames(k)`) given the observation in row `j` of `Mdl.X`.

`Cost` — Expected misclassification costs
numeric matrix

Expected “Misclassification Cost” on page 33-4842, returned as a numeric matrix. `Cost` has rows equal to the number of rows of `Mdl.X` and columns equal to the number of distinct classes in the training data (`size(Mdl.ClassNames,1)`).

`Cost(j,k)` is the expected misclassification cost of the observation in row `j` of `Mdl.X` predicted into class `k` (in class `Mdl.ClassNames(k)`).

More About

Misclassification Cost

A misclassification cost is the relative severity of a classifier labeling an observation into the wrong class.

There are two types of misclassification costs: true and expected. Let K be the number of classes.

- True misclassification cost — A K-by-K matrix, where element (i,j) indicates the misclassification cost of predicting an observation into class j if its true class is i. The software stores the misclassification cost in the property `Mdl.Cost`, and uses it in computations. By default, `Mdl.Cost(i,j) = 1` if `i ≠ j`, and `Mdl.Cost(i,j) = 0` if `i = j`. In other words, the cost is `0` for correct classification and `1` for any incorrect classification.
- Expected misclassification cost — A K-dimensional vector, where element k is the weighted average misclassification cost of classifying an observation into class k, weighted by the class posterior probabilities.

  $$c_k = \sum_{j=1}^{K} \widehat{P}(Y = j \,|\, x_1, \ldots, x_P) Cost_{jk}\,.$$

  In other words, the software classifies observations to the class corresponding with the lowest expected misclassification cost.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For naive Bayes, the posterior probability that a classification is k for a given observation $(x_1,\ldots,x_P)$ is

$$\widehat{P}(Y = k \,|\, x_1, .., x_P) = \frac{P(X_1, \ldots, X_P \,|\, y = k)\pi(Y = k)}{P(X_1, \ldots, X_P)},$$

where:

- $P(X_1, ..., X_P | y = k)$ is the conditional joint density of the predictors given they are in class k. `Mdl.DistributionNames` stores the distribution names of the predictors.
- $\pi(Y = k)$ is the class prior probability distribution. `Mdl.Prior` stores the prior distribution.
- $P(X_1, .., X_P)$ is the joint density of the predictors. The classes are discrete, so

$$P(X_1, ..., X_P) = \sum_{k=1}^{K} P(X_1, ..., X_P \Big| y = k)\pi(Y = k) .$$

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-787 of the `CompactClassificationNaiveBayes` object.

Argument	Notes and Limitations

X	• X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in X can be a variable size, but the number of columns in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	  For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationNaiveBayes` | `CompactClassificationNaiveBayes` | `fitcnb` | `loss` | `resubPredict`

Topics

“Naive Bayes Classification” on page 21-2
“Grouping Variables” on page 2-45
“Plot Posterior Classification Probabilities” on page 21-5

Introduced in R2014b

predict

Package:

Classify observations using neural network classifier

Syntax

```
label = predict(Mdl,X)
label = predict(Mdl,X,'ObservationsIn',dimension)
[label,Score] = predict( ___ )
```

Description

`label = predict(Mdl,X)` returns predicted class labels for the predictor data in the table or matrix `X` using the trained neural network classification model `Mdl`.

`label = predict(Mdl,X,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

`[label,Score] = predict( ___ )` also returns a matrix of classification scores on page 33-4853 indicating the likelihood that a label comes from a particular class, using any of the input argument combinations in the previous syntaxes. For each observation in `X`, the predicted class label corresponds to the maximum score among all classes.

Examples

Classify Test Set Observations Using Neural Network

Predict labels for test set observations using a neural network classifier.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Smoker` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Diastolic,Systolic,Gender,Height,Weight,Age,Smoker);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

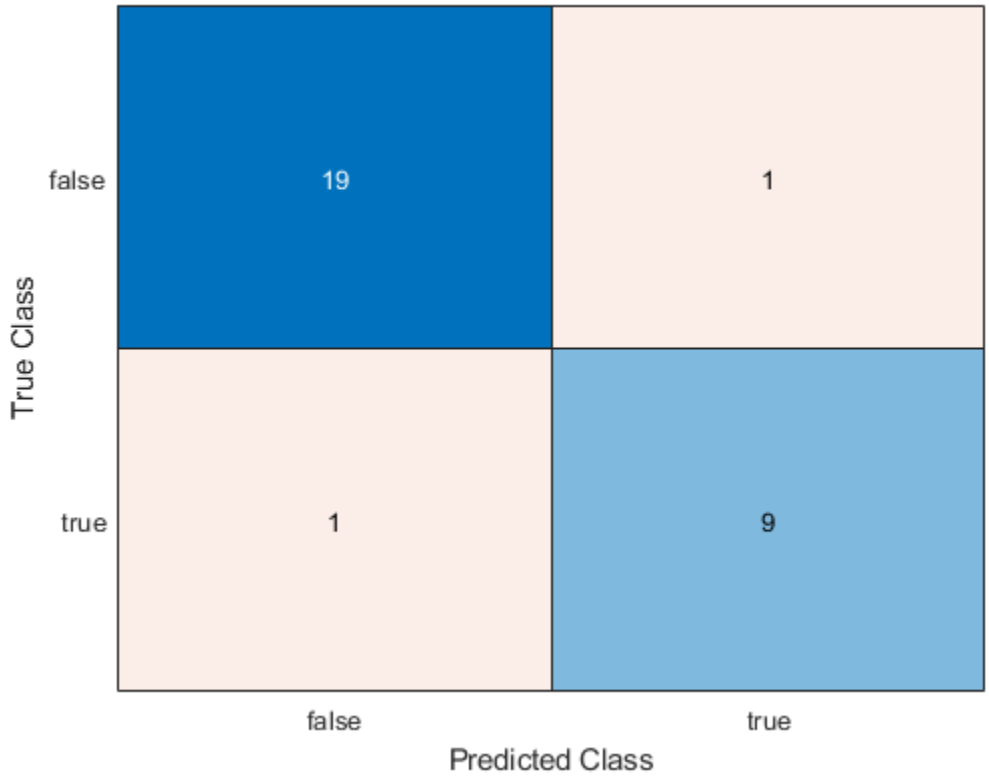
```
rng("default") % For reproducibility of the partition
c = cvpartition(tbl.Smoker,"Holdout",0.30);
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a neural network classifier using the training set. Specify the `Smoker` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors.

```
Mdl = fitcnet(tblTrain,"Smoker", ...
    "Standardize",true);
```

Classify the test set observations. Visualize the results using a confusion matrix.

```
label = predict(Mdl,tblTest);
confusionchart(tblTest.Smoker,label)
```



The neural network model correctly classifies all but two of the test set observations.

Select Features to Include in Neural Network Classifier

Perform feature selection by comparing test set classification margins, edges, errors, and predictions. Compare the test set metrics for a model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a stratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(fishertable.Species,"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

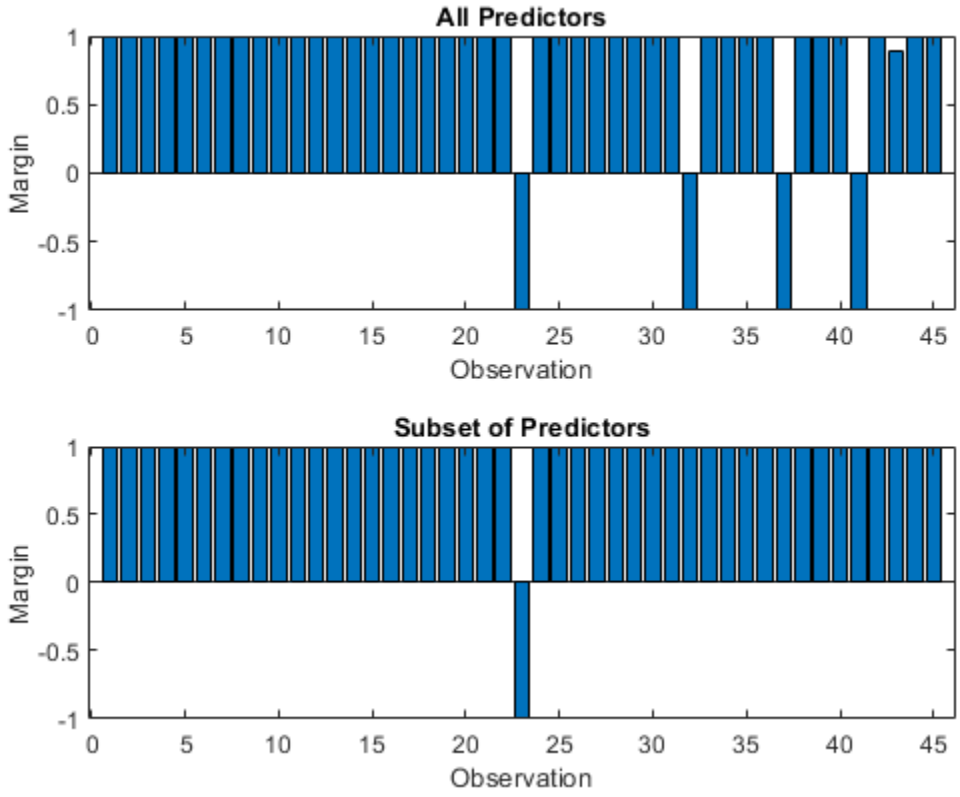
Train one neural network classifier using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `Species` as the response variable, and standardize the predictors.

```
allMdl = fitcnet(trainTbl,"Species","Standardize",true);
subsetMdl = fitcnet(trainTbl,"Species ~ SepalLength + SepalWidth + PetalLength", ...
    "Standardize",true);
```

Calculate the test set classification margins for the two models. Because the test set includes only 45 observations, display the margins using bar graphs.

For each observation, the classification margin is the difference between the classification score for the true class and the maximal score for the false classes. Because neural network classifiers return classification scores that are posterior probabilities, margin values close to 1 indicate confident classifications and negative margin values indicate misclassifications.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allMargins = margin(allMdl,testTbl);
bar(ax1,allMargins)
xlabel(ax1,"Observation")
ylabel(ax1,"Margin")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
subsetMargins = margin(subsetMdl,testTbl);
bar(ax2,subsetMargins)
xlabel(ax2,"Observation")
ylabel(ax2,"Margin")
title(ax2,"Subset of Predictors")
```



Compare the test set classification edge, or mean of the classification margins, of the two models.

```
allEdge = edge(allMdl,testTbl)
```

```
allEdge = 0.8198
```

```
subsetEdge = edge(subsetMdl,testTbl)
```

```
subsetEdge = 0.9556
```

Based on the test set classification margins and edges, the model trained on a subset of the predictors seems to outperform the model trained on all the predictors.

Compare the test set classification error of the two models.

```
allError = loss(allMdl,testTbl);
allAccuracy = 1-allError
```

```
allAccuracy = 0.9111
```

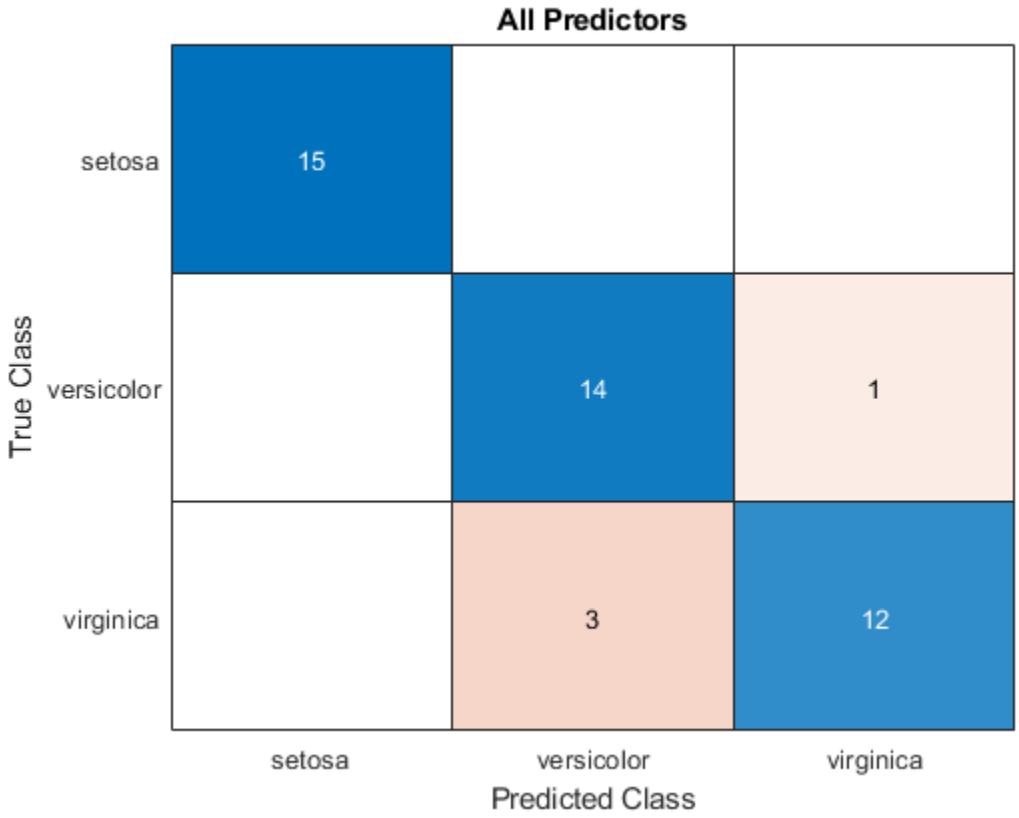
```
subsetError = loss(subsetMdl,testTbl);
subsetAccuracy = 1-subsetError
```

```
subsetAccuracy = 0.9778
```

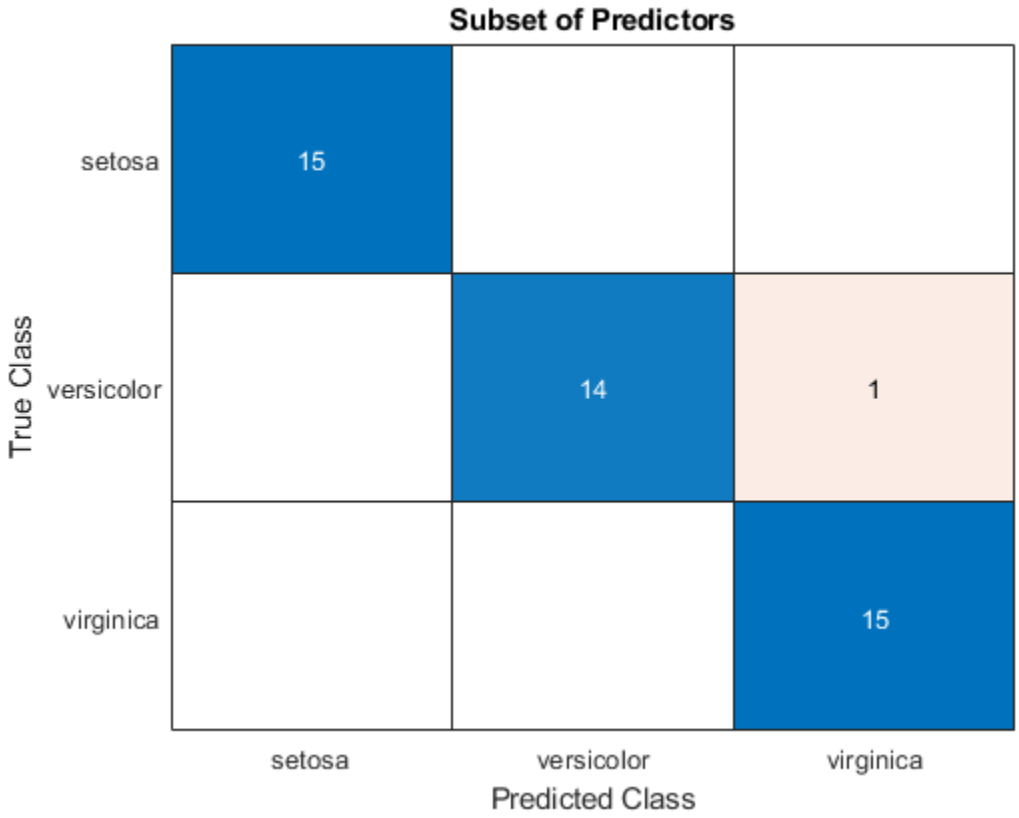
Again, the model trained using only a subset of the predictors seems to perform better than the model trained using all the predictors.

Visualize the test set classification results using confusion matrices.

```
allLabels = predict(allMdl,testTbl);
figure
confusionchart(testTbl.Species,allLabels)
title("All Predictors")
```




```
subsetLabels = predict(subsetMdl,testTbl);
figure
confusionchart(testTbl.Species,subsetLabels)
title("Subset of Predictors")
```



The model trained using all the predictors misclassifies four of the test set observations. The model trained using a subset of the predictors misclassifies only one of the test set observations.

Given the test set performance of the two models, consider using the model trained using all the predictors except `PetalWidth`.

Predict Using Layer Structure of Neural Network Classifier

See how the layers of a neural network classifier work together to predict the label and classification scores for a single observation.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Train a neural network classifier using the data set. Specify the `Species` column of `fishertable` as the response variable.

```
Mdl = fitcnet(fishertable,"Species");
```

Select the fifteenth observation from the data set. See how the layers of the neural network classifier take the observation and return a predicted class label `newPointLabel` and classification scores `newPointScores`.

```
newPoint = Mdl.X{15,:}

newPoint = 1×4

    5.8000    4.0000    1.2000    0.2000


firstFCStep = (Mdl.LayerWeights{1})*newPoint' + Mdl.LayerBiases{1};
reluStep = max(firstFCStep,0);

finalFCStep = (Mdl.LayerWeights{end})*reluStep + Mdl.LayerBiases{end};
finalSoftmaxStep = softmax(finalFCStep);

[~,classIdx] = max(finalSoftmaxStep);
newPointLabel = Mdl.ClassNames{classIdx}

newPointLabel =
'setosa'

newPointScores = finalSoftmaxStep'

newPointScores = 1×3

    1.0000    0.0000    0.0000
```

Check that the predictions match those returned by the `predict` object function.

```
[predictedLabel,predictedScores] = predict(Mdl,newPoint)

predictedLabel = 1×1 cell array
    {'setosa'}


predictedScores = 1×3

    1.0000    0.0000    0.0000
```

Input Arguments

`Mdl` — Trained neural network classifier

`ClassificationNeuralNetwork` model object | `CompactClassificationNeuralNetwork` model object

Trained neural network classifier, specified as a `ClassificationNeuralNetwork` model object or `CompactClassificationNeuralNetwork` model object returned by `fitcnet` or `compact`, respectively.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

By default, each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`) and `Tbl` contains only numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value argument of `fitcnet`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` must be the same as the corresponding predictor variable names in `X`. To specify predictor names during training, use the `PredictorNames` name-value argument of `fitcnet`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

If you set `'Standardize',true` in `fitcnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `single` | `double` | `table`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

Output Arguments

`label` — Predicted class labels
numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Predicted class labels, returned as a numeric, categorical, or logical vector; a character or string array; or a cell array of character vectors. The software predicts the classification of an observation by assigning the observation to the class yielding the largest classification score or posterior probability.

`label` has the same data type as the observed class labels that trained `Mdl`, and its length is equal to the number of observations in `X`. (The software treats string arrays as cell arrays of character vectors.)

`Score` — Classification scores
numeric matrix

Classification scores on page 33-4853, returned as an *n*-by-*K* matrix, where *n* is the number of observations in `X` and *K* is the number of unique classes. The classification score `Score(i,j)` represents the posterior probability that the `i`th observation belongs to class `j`.

More About

Classification Scores

The classification scores for a neural network classifier are computed using the softmax activation function that follows the final fully connected layer in the network. The scores correspond to posterior probabilities.

The posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{\sum_{j=1}^{K} P(x|j)P(j)} = \frac{\exp(a_k(x))}{\sum_{j=1}^{K} \exp(a_j(x))}$$

where

- $P(x|k)$ is the conditional probability of x given class k.
- $P(k)$ is the prior probability for class k.
- K is the number of classes in the response variable.
- $a_k(x)$ is the k output from the final fully connected layer for observation x.

See Also

`ClassificationNeuralNetwork` | `CompactClassificationNeuralNetwork` | `edge` | `fitcnet` | `loss` | `margin`

Topics
“Assess Neural Network Classifier Performance” on page 18-177

Introduced in R2021a

predict

Package: `classreg.learning.classif`

Classify observations using support vector machine (SVM) classifier

Syntax

```
label = predict(SVMModel,X)
[label,score] = predict(SVMModel,X)
```

Description

`label = predict(SVMModel,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the trained support vector machine (SVM) classification model `SVMModel`. The trained SVM model can either be full or compact.

`[label,score] = predict(SVMModel,X)` also returns a matrix of scores (`score`) indicating the likelihood that a label comes from a particular class. For SVM, likelihood measures are either classification scores on page 33-4858 or class posterior probabilities on page 33-4859. For each observation in `X`, the predicted class label corresponds to the maximum score among all classes.

Examples

Label Test Sample Observations of SVM Classifiers

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify a 15% holdout sample for testing, standardize the data, and specify that `'g'` is the positive class.

```
CVSVMModel = fitcsvm(X,Y,'Holdout',0.15,'ClassNames',{'b','g'},...
    'Standardize',true);
CompactSVMModel = CVSVMModel.Trained{1}; % Extract trained, compact classifier
testInds = test(CVSVMModel.Partition);   % Extract the test indices
XTest = X(testInds,:);
YTest = Y(testInds,:);
```

`CVSVMModel` is a `ClassificationPartitionedModel` classifier. It contains the property `Trained`, which is a 1-by-1 cell array holding a `CompactClassificationSVM` classifier that the software trained using the training set.

Label the test sample observations. Display the results for the first 10 observations in the test sample.

```
[label,score] = predict(CompactSVMModel,XTest);
table(YTest(1:10),label(1:10),score(1:10,2),'VariableNames',...
    {'TrueLabel','PredictedLabel','Score'})
```

```
ans=10×3 table
    TrueLabel    PredictedLabel     Score
    _________    ______________    ________

      {'b'}          {'b'}          -1.7176
      {'g'}          {'g'}           2.0002
      {'b'}          {'b'}          -9.6851
      {'g'}          {'g'}           2.5618
      {'b'}          {'b'}          -1.5481
      {'g'}          {'g'}           2.0983
      {'b'}          {'b'}          -2.7012
      {'b'}          {'b'}         -0.66312
      {'g'}          {'g'}           1.6045
      {'g'}          {'g'}           1.7729
```

Predict Labels and Posterior Probabilities of SVM Classifiers

Label new observations using an SVM classifier.

Load the ionosphere data set. Assume that the last 10 observations become available after you train the SVM classifier.

```
load ionosphere
rng(1); % For reproducibility
n = size(X,1);       % Training sample size
isInds = 1:(n-10);   % In-sample indices
oosInds = (n-9):n;   % Out-of-sample indices
```

Train an SVM classifier. Standardize the data and specify that `'g'` is the positive class. Conserve memory by reducing the size of the trained SVM classifier.

```
SVMModel = fitcsvm(X(isInds,:),Y(isInds),'Standardize',true,...
    'ClassNames',{'b','g'});
CompactSVMModel = compact(SVMModel);
whos('SVMModel','CompactSVMModel')
```

```
  Name                Size             Bytes  Class

  CompactSVMModel     1x1              30482  classreg.learning.classif.CompactClassificationSVM
  SVMModel            1x1             137582  ClassificationSVM
```

The `CompactClassificationSVM` classifier (`CompactSVMModel`) uses less space than the `ClassificationSVM` classifier (`SVMModel`) because `SVMModel` stores the data.

Estimate the optimal score-to-posterior-probability transformation function.

```
CompactSVMModel = fitPosterior(CompactSVMModel,...
    X(isInds,:),Y(isInds))
```

```
CompactSVMModel =
  CompactClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: '@(S)sigmoid(S,-1.968445e+00,3.121617e-01)'
```

```
                    Alpha: [88x1 double]
                     Bias: -0.2142
        KernelParameters: [1x1 struct]
                       Mu: [1x34 double]
                    Sigma: [1x34 double]
          SupportVectors: [88x34 double]
      SupportVectorLabels: [88x1 double]


  Properties, Methods
```

The optimal score transformation function (`CompactSVMModel.ScoreTransform`) is the sigmoid function because the classes are inseparable.

Predict the out-of-sample labels and positive class posterior probabilities. Because true labels are available, compare them with the predicted labels.

```
[labels,PostProbs] = predict(CompactSVMModel,X(oosInds,:));
table(Y(oosInds),labels,PostProbs(:,2),'VariableNames',...
    {'TrueLabels','PredictedLabels','PosClassPosterior'})
```

```
ans=10×3 table
    TrueLabels    PredictedLabels    PosClassPosterior
    __________    _______________    _________________

      {'g'}            {'g'}              0.98419
      {'g'}            {'g'}              0.95545
      {'g'}            {'g'}              0.67789
      {'g'}            {'g'}              0.94447
      {'g'}            {'g'}              0.98744
      {'g'}            {'g'}               0.9248
      {'g'}            {'g'}               0.9711
      {'g'}            {'g'}              0.96986
      {'g'}            {'g'}              0.97803
      {'g'}            {'g'}               0.9436
```

`PostProbs` is a 10-by-2 matrix, where the first column is the negative class posterior probabilities, and the second column is the positive class posterior probabilities corresponding to the new observations.

Input Arguments

`SVMModel` — SVM classification model

`ClassificationSVM` model object | `CompactClassificationSVM` model object

SVM classification model, specified as a `ClassificationSVM` model object or `CompactClassificationSVM` model object returned by `fitcsvm` or `compact`, respectively.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `SVMModel`.
  - If you trained `SVMModel` using a table (for example, `Tbl`) and `Tbl` contains all numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value pair argument of `fitcsvm`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `SVMModel` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `SVMModel` (stored in `SVMModel.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you trained `SVMModel` using a numeric matrix, then the predictor names in `SVMModel.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitcsvm`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

If you set `'Standardize',true` in `fitcsvm` to train `SVMModel`, then the software standardizes the columns of `X` using the corresponding means in `SVMModel.Mu` and the standard deviations in `SVMModel.Sigma`.

Data Types: `table` | `double` | `single`

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label` has the same data type as the observed class labels (`Y`) that trained `SVMModel`, and its length is equal to the number of rows in `X`. (The software treats string arrays as cell arrays of character vectors.)

For one-class learning, `label` is the one class represented in the observed class labels.

`score` — Predicted class scores or posterior probabilities

numeric column vector | numeric matrix

Predicted class scores on page 33-4858 or posterior probabilities on page 33-4859, returned as a numeric column vector or numeric matrix.

- For one-class learning, `score` is a column vector with the same number of rows as the observations (`X`). The elements of `score` are the positive class scores for the corresponding observations. You cannot obtain posterior probabilities for one-class learning.
- For two-class learning, `score` is a two-column matrix with the same number of rows as `X`.
  - If you fit the optimal score-to-posterior-probability transformation function using `fitPosterior` or `fitSVMPosterior`, then `score` contains class posterior probabilities. That is, if the value of `SVMModel.ScoreTransform` is not `none`, then the first and second columns of `score` contain the negative class (`SVMModel.ClassNames{1}`) and positive class (`SVMModel.ClassNames{2}`) posterior probabilities for the corresponding observations, respectively.
  - Otherwise, the first column contains the negative class scores and the second column contains the positive class scores for the corresponding observations.

If `SVMModel.KernelParameters.Function` is `'linear'`, then the classification score for the observation x is

$$f(x) = (x/s)'\beta + b\,.$$

`SVMModel` stores β, b, and s in the properties `Beta`, `Bias`, and `KernelParameters.Scale`, respectively.

To estimate classification scores manually, you must first apply any transformations to the predictor data that were applied during training. Specifically, if you specify `'Standardize',true` when using `fitcsvm`, then you must standardize the predictor data manually by using the mean `SVMModel.Mu` and standard deviation `SVMModel.Sigma`, and then divide the result by the kernel scale in `SVMModel.KernelParameters.Scale`.

All SVM functions, such as `resubPredict` and `predict`, apply any required transformation before estimation.

If `SVMModel.KernelParameters.Function` is not `'linear'`, then `Beta` is empty (`[]`).

More About

Classification Score

The SVM classification score for classifying observation x is the signed distance from x to the decision boundary ranging from -∞ to +∞. A positive score for a class indicates that x is predicted to be in that class. A negative score indicates otherwise.

The positive class classification score $f(x)$ is the trained SVM classification function. $f(x)$ is also the numerical predicted response for x, or the score for predicting x into the positive class.

$$f(x) = \sum_{j=1}^{n} \alpha_j y_j G(x_j, x) + b,$$

where $(\alpha_1, ..., \alpha_n, b)$ are the estimated SVM parameters, $G(x_j, x)$ is the dot product in the predictor space between x and the support vectors, and the sum includes the training set observations. The negative class classification score for x, or the score for predicting x into the negative class, is $-f(x)$.

If $G(x_j,x) = x_j'x$ (the linear kernel), then the score function reduces to

$$f(x) = (x/s)'\beta + b\,.$$

s is the kernel scale and β is the vector of fitted linear coefficients.

For more details, see “Understanding Support Vector Machines” on page 18-150.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For SVM, the posterior probability is a function of the score $P(s)$ that observation j is in class $k = \{-1,1\}$.

- For separable classes, the posterior probability is the step function

$$P(s_j) = \begin{cases} 0; & s < \max\limits_{y_k = -1} s_k \\ \pi; & \max\limits_{y_k = -1} s_k \le s_j \le \min\limits_{y_k = +1} s_k \\ 1; & s_j > \min\limits_{y_k = +1} s_k \end{cases},$$

  where:

  - s_j is the score of observation j.
  - +1 and –1 denote the positive and negative classes, respectively.
  - π is the prior probability that an observation is in the positive class.
- For inseparable classes, the posterior probability is the sigmoid function

$$P(s_j) = \frac{1}{1 + \exp(As_j + B)},$$

  where the parameters A and B are the slope and intercept parameters, respectively.

Prior Probability

The prior probability of a class is the assumed relative frequency with which observations from that class occur in a population.

Tips

- If you are using a linear SVM model for classification and the model has many support vectors, then using `predict` for the prediction method can be slow. To efficiently classify observations based on a linear SVM model, remove the support vectors from the model object by using `discardSupportVectors`.

Algorithms

- By default and irrespective of the model kernel function, MATLAB uses the dual representation of the score function to classify observations based on trained SVM models, specifically

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\alpha}_j y_j G(x, x_j) + \widehat{b}\,.$$

This prediction method requires the trained support vectors and α coefficients (see the `SupportVectors` and `Alpha` properties of the SVM model).

- By default, the software computes optimal posterior probabilities using Platt's method [1]:

  1. Perform 10-fold cross-validation.
  2. Fit the sigmoid function parameters to the scores returned from the cross-validation.
  3. Estimate the posterior probabilities by entering the cross-validation scores into the fitted sigmoid function.

- The software incorporates prior probabilities in the SVM objective function during training.
- For SVM, `predict` and `resubPredict` classify observations into the class yielding the largest score (the largest posterior probability). The software accounts for misclassification costs by applying the average-cost correction before training the classifier. That is, given the class prior vector P, misclassification cost matrix C, and observation weight vector w, the software defines a new vector of observation weights (W) such that

$$W_j = w_j P_j \sum_{k=1}^{K} C_{jk}.$$

Alternative Functionality

Simulink Block

To integrate the prediction of an SVM classification model into Simulink, you can use the ClassificationSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see "Predict Class Labels Using ClassificationSVM Predict Block" on page 32-111 and "Predict Class Labels Using MATLAB Function Block" on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

References

[1] Platt, J. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in Large Margin Classifiers*. MIT Press, 1999, pages 61–74.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see "Tall Arrays".

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`SVMModel`	• If you use `saveLearnerForCoder` to save an SVM model that is equipped to predict posterior probabilities, and use `loadLearnerForCoder` to load the model, then `loadLearnerForCoder` cannot restore the ScoreTransform property into the MATLAB Workspace. However, `loadLearnerForCoder` can load the model, including the ScoreTransform property, at compile time for code generation, within an entry-point function. • For the usage notes and limitations of the model object, see “Code Generation” on page 33-807 of the `CompactClassificationSVM` object.

Argument	Notes and Limitations

X	• For general code generation, X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, X must be a single-precision or double-precision matrix.
	• For fixed-point code generation, X must be a fixed-point matrix.
	• The number of rows, or observations, in X can be a variable size, but the number of columns in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

For more information, see “Introduction to Code Generation” on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationSVM` | `CompactClassificationSVM` | `fitSVMPosterior` | `fitcsvm` | `loss` | `resubPredict`

Topics

“Support Vector Machines for Binary Classification” on page 18-150
“Analyze Images Using Linear Support Vector Machines” on page 18-172
“Train SVM Classifier Using Custom Kernel” on page 18-159
“Plot Posterior Probability Regions for SVM Classification Models” on page 18-170

Introduced in R2014a

predict

Predict labels using classification tree

Syntax

```
label = predict(Mdl,X)
label = predict(Mdl,X,Name,Value)
[label,score,node,cnum] = predict( ___ )
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels for the predictor data in the table or matrix `X`, based on the trained, full or compact classification tree `Mdl`.

`label = predict(Mdl,X,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, you can specify to prune `Mdl` to a particular level before predicting labels.

`[label,score,node,cnum] = predict( ___ )` uses any of the input argument in the previous syntaxes and additionally returns:

- A matrix of classification scores (`score`) indicating the likelihood that a label comes from a particular class. For classification trees, scores are posterior probabilities. For each observation in `X`, the predicted class label on page 33-4868 corresponds to the minimum expected misclassification cost on page 33-4871 among all classes.
- A vector of predicted node numbers for the classification (`node`).
- A vector of predicted class number for the classification (`cnum`).

Input Arguments

`Mdl` — Trained classification tree

`ClassificationTree` model object | `CompactClassificationTree` model object

Trained classification tree, specified as a `ClassificationTree` or `CompactClassificationTree` model object. That is, `Mdl` is a trained classification model returned by `fitctree` or `compact`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical

during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitctree`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.

- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitctree`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

Data Types: `table` | `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level

0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(Mdl.PruneList)`. `0` indicates the full, unpruned tree and `max(Mdl.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `predict` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(Mdl.PruneList)`.

`predict` prunes `Mdl` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `Mdl` must be nonempty. In other words, grow `Mdl` by setting `'Prune','on'`, or by pruning `Mdl` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`label` — Predicted class labels

vector | array

Predicted class labels on page 33-4868, returned as a vector or array. Each entry of `label` corresponds to the class with minimal expected cost for the corresponding row of `X`.

Suppose `Subtrees` is a numeric vector containing `T` elements (for `'all'`, see `Subtrees`), and `X` has `N` rows.

- If the response data type is `char` and:
  - `T` = 1, then `label` is a character matrix containing `N` rows. Each row contains the predicted label produced by subtree `Subtrees`.
  - `T` > 1, then `label` is an `N`-by-`T` cell array.
- Otherwise, `label` is an `N`-by-`T` array having the same data type as the response. (The software treats string arrays as cell arrays of character vectors.)

In the latter two cases, column *j* of `label` contains the vector of predicted labels produced by subtree `Subtrees(`*j*`)`.

`score` — Posterior probabilities
numeric matrix

Posterior probabilities, returned as a numeric matrix of size `N`-by-`K`, where `N` is the number of observations (rows) in `X`, and `K` is the number of classes (in `Mdl.ClassNames`). `score(i,j)` is the posterior probability that row `i` of `X` is of class `j`.

If `Subtrees` has `T` elements, and `X` has `N` rows, then `score` is an `N`-by-`K`-by-`T` array, and `node` and `cnum` are `N`-by-`T` matrices.

`node` — Node numbers
numeric vector

Node numbers for the predicted classes, returned as a numeric vector. Each entry corresponds to the predicted node in `Mdl` for the corresponding row of `X`.

`cnum` — Class numbers
numeric vector

Class numbers corresponding to the predicted `labels`, returned as a numeric vector. Each entry of `cnum` corresponds to a predicted class number for the corresponding row of `X`.

Examples

Predict Labels Using a Classification Tree

Examine predictions for a few rows in a data set left out of training.

Load Fisher's iris data set.

```
load fisheriris
```

Partition the data into training (50%) and validation (50%) sets.

```
n = size(meas,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
```

```
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a classification tree using the training set.

```
Mdl = fitctree(meas(idxTrn,:),species(idxTrn));
```

Predict labels for the validation data. Count the number of misclassified observations.

```
label = predict(Mdl,meas(idxVal,:));
label(randsample(numel(label),5)) % Display several predicted labels
```

```
ans = 5x1 cell
    {'setosa'    }
    {'setosa'    }
    {'setosa'    }
    {'virginica' }
    {'versicolor'}
```

```
numMisclass = sum(~strcmp(label,species(idxVal)))
```

```
numMisclass = 3
```

The software misclassifies three out-of-sample observations.

Estimate Class Posterior Probabilities Using a Classification Tree

Load Fisher's iris data set.

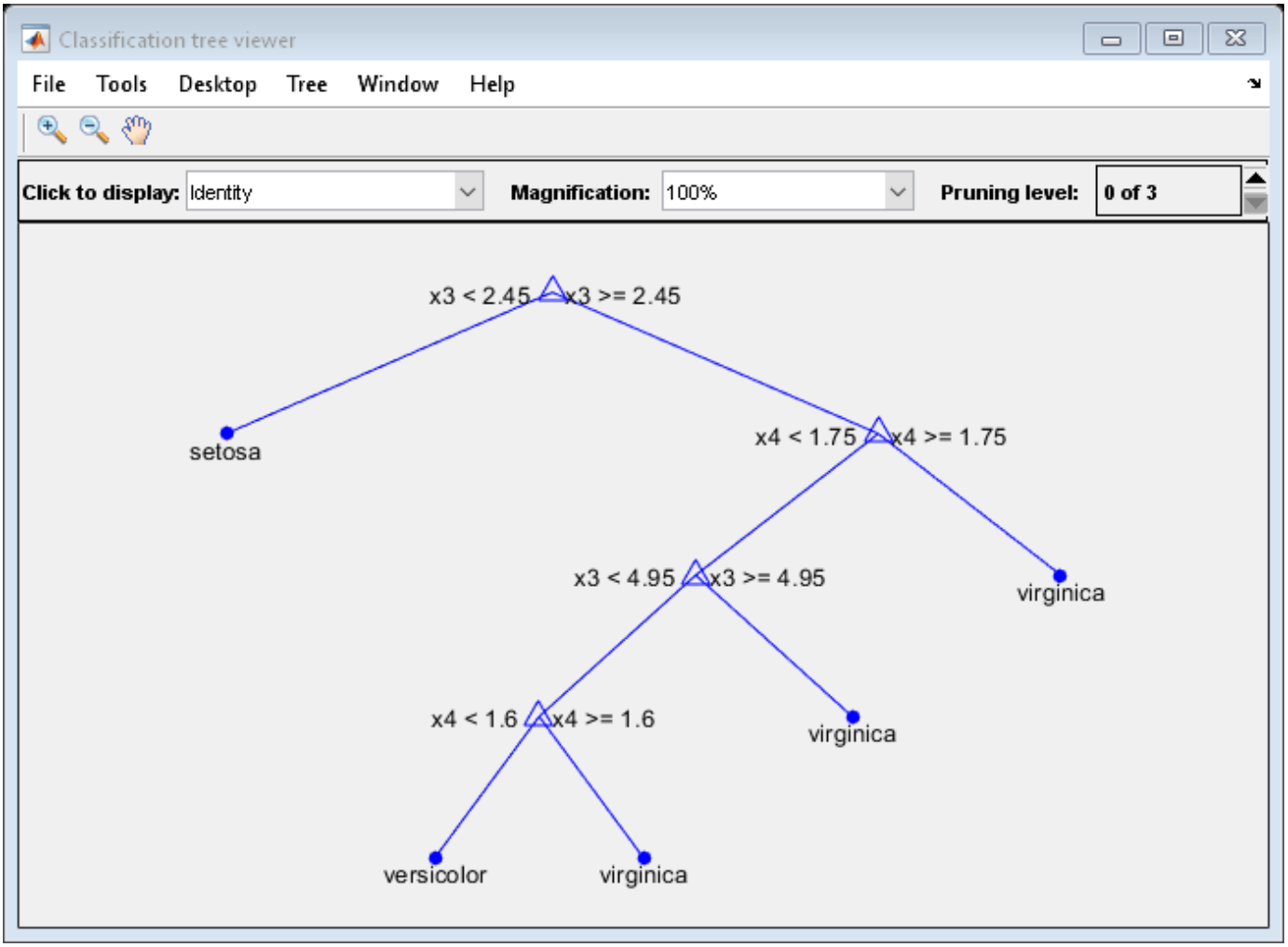
```
load fisheriris
```

Partition the data into training (50%) and validation (50%) sets.

```
n = size(meas,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a classification tree using the training set, and then view it.

```
Mdl = fitctree(meas(idxTrn,:),species(idxTrn));
view(Mdl,'Mode','graph')
```



The resulting tree has four levels.

Estimate posterior probabilities for the test set using subtrees pruned to levels 1 and 3.

```
[~,Posterior] = predict(Mdl,meas(idxVal,:),'SubTrees',[1 3]);
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Posterior(randsample(size(Posterior,1),5),:,:),...
    % Display several posterior probabilities
```

```
ans =
ans(:,:,1) =

    1.0000         0         0
    1.0000         0         0
    1.0000         0         0
```

```
         0         0    1.0000
         0    0.8571    0.1429


ans(:,:,2) =

    0.3733    0.3200    0.3067
    0.3733    0.3200    0.3067
    0.3733    0.3200    0.3067
    0.3733    0.3200    0.3067
    0.3733    0.3200    0.3067
```

The elements of `Posterior` are class posterior probabilities:

- Rows correspond to observations in the validation set.
- Columns correspond to the classes as listed in `Mdl.ClassNames`.
- Pages correspond to the subtrees.

The subtree pruned to level 1 is more sure of its predictions than the subtree pruned to level 3 (i.e., the root node).

More About

Predicted Class Label

`predict` classifies by minimizing the expected misclassification cost:

$$\widehat{y} = \underset{y = 1, \ldots, K}{\operatorname{argmin}} \sum_{j=1}^{K} \widehat{P}(j|x) C(y|j),$$

where:

- $\widehat{y}$ is the predicted classification.
- K is the number of classes.
- $\widehat{P}(j|x)$ is the posterior probability of class j for observation x.
- $C(y|j)$ is the cost of classifying an observation as y when its true class is j.

Score (tree)

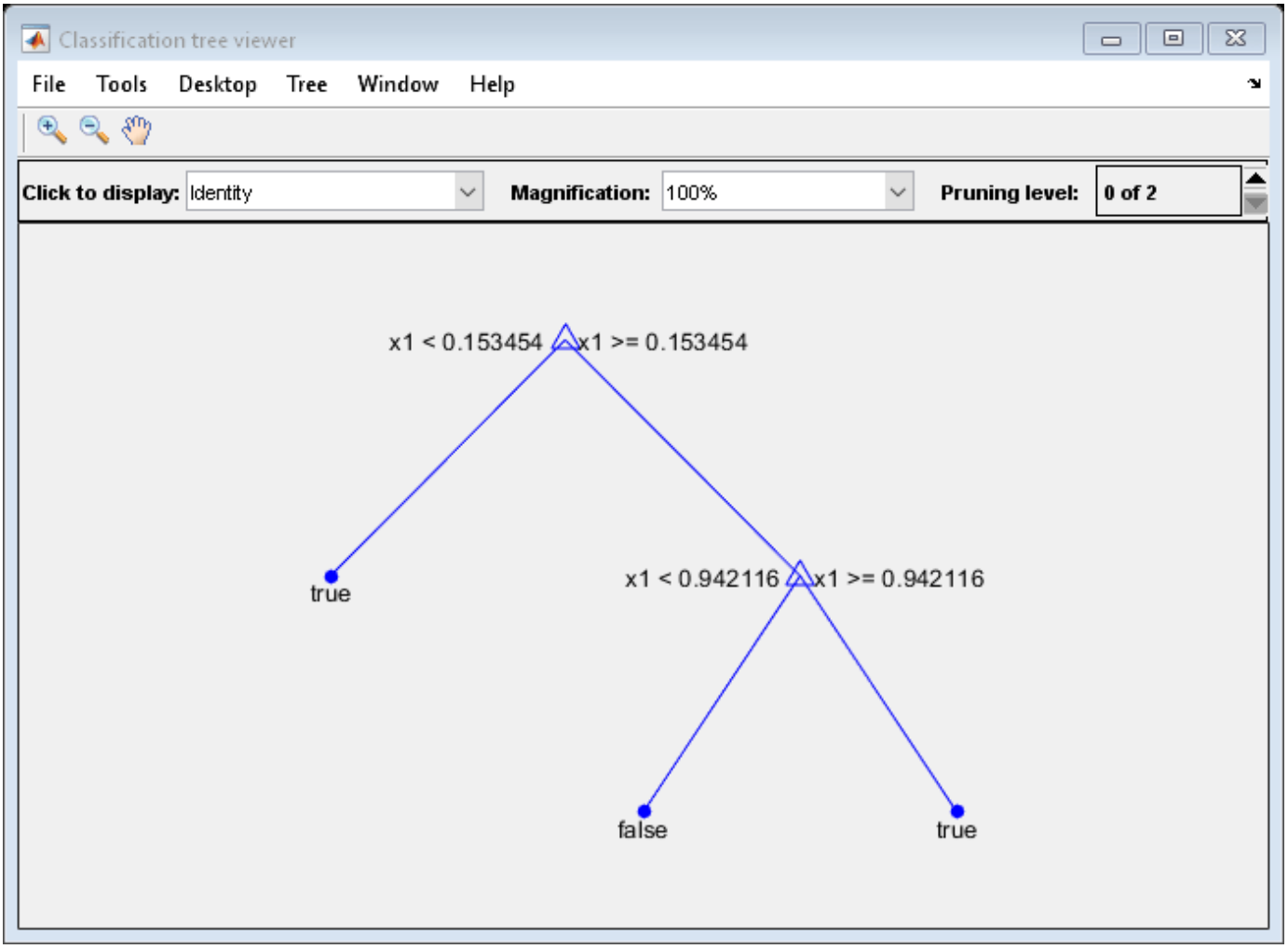
For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor `X` as `true` when `X < 0.15` or `X > 0.95`, and `X` is false otherwise.

Generate 100 random points and classify them:

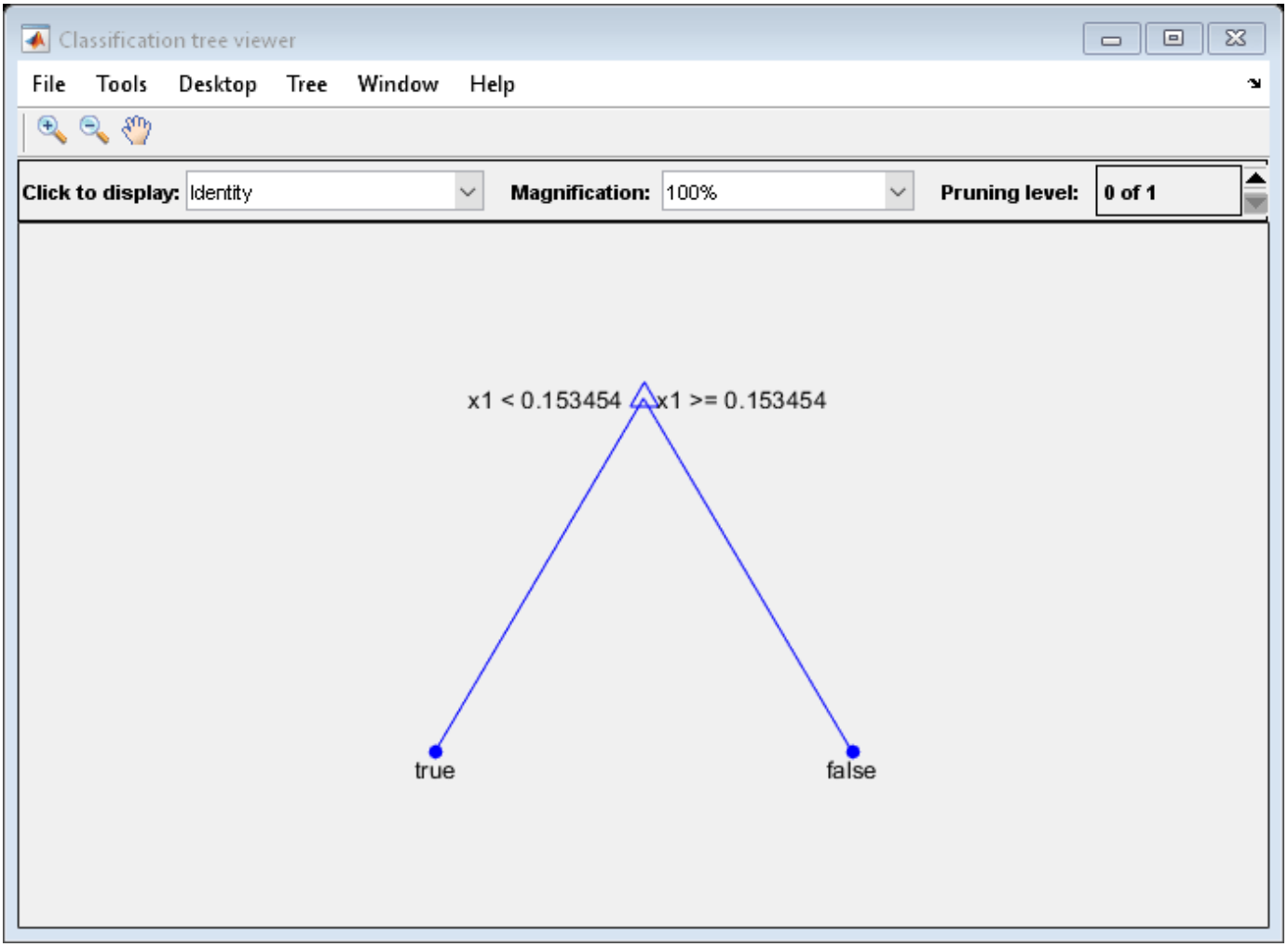
```
rng(0,'twister') % for reproducibility
X = rand(100,1);
```

```
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

True Misclassification Cost

The true misclassification cost is the cost of classifying an observation into an incorrect class.

You can set the true misclassification cost per class by using the `'Cost'` name-value argument when you create the classifier. `Cost(i,j)` is the cost of classifying an observation into class `j` when its true class is `i`. By default, `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Expected Cost

The expected misclassification cost per observation is an averaged cost of classifying the observation into each class.

Suppose you have `Nobs` observations that you want to classify with a trained classifier, and you have `K` classes. You place the observations into a matrix `X` with one observation per row.

The expected cost matrix `CE` has size `Nobs`-by-`K`. Each row of `CE` contains the expected (average) cost of classifying the observation into each of the `K` classes. `CE(n,k)` is

$$\sum_{i=1}^{K} \widehat{P}(i|X(n))C(k|i),$$

where:

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability of class i for observation $X(n)$.
- $C(k|i)$ is the true misclassification cost of classifying an observation as k when its true class is i.

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_j L_k} - P_{R_j R_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.

- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_jL_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in (–∞,1]. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Algorithms

`predict` generates predictions by following the branches of `Mdl` until it reaches a leaf node or a missing value. If `predict` reaches a leaf node, it returns the classification of that node.

If `predict` reaches a node with a missing value for a predictor, its behavior depends on the setting of the `Surrogate` name-value pair when `fitctree` constructs `Mdl`.

- **`Surrogate = 'off'`** (default) — `predict` returns the label with the largest number of training samples that reach the node.
- **`Surrogate = 'on'`** — `predict` uses the best surrogate split at the node. If all surrogate split variables with positive predictive measure of association are missing, `predict` returns the label with the largest number of training samples that reach the node. For a definition, see “Predictive Measure of Association” on page 33-4871.

Alternative Functionality

Simulink Block

To integrate the prediction of a classification tree model into Simulink, you can use the ClassificationTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see “Predict Class Labels Using ClassificationTree Predict Block” on page 32-121 and “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-816 of the `CompactClassificationTree` object.

Argument	Notes and Limitations

X	• For general code generation, X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, X must be a single-precision or double-precision matrix.
	• For fixed-point code generation, X must be a fixed-point matrix.
	• The number of rows, or observations, in X can be a variable size, but the number of columns in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

`label`	If the response data type is `char` and `codegen` cannot determine that the value of `Subtrees` is a scalar, then `label` is a cell array of character vectors.
`'Subtrees'`	• Names in name-value pair arguments must be compile-time constants. For example, to allow user-defined pruning levels in the generated code, include `{coder.Constant('Subtrees'),coder.typeof(0,[1,n],[0,1])}` in the `-args` value of `codegen`, where `n` is `max(Mdl.PruneList)`. • The `'Subtrees'` name-value pair argument is not supported in the coder configurer workflow. • For fixed-point code generation, the `'Subtrees'` value must be `coder.Constant('all')` or have an integer data type.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationTree` | `CompactClassificationTree` | `compact` | `edge` | `fitctree` | `loss` | `margin` | `prune`

Topics

“Specify Variable-Size Arguments for Code Generation” on page 32-45

Introduced in R2011a

predict

Predict responses using ensemble of regression models

Syntax

```
Yfit = predict(Mdl,X)
Yfit = predict(Mdl,X,Name,Value)
```

Description

`Yfit = predict(Mdl,X)` returns predicted responses to the predictor data in the table or matrix `X`, based on the regression ensemble model `Mdl`.

`Yfit = predict(Mdl,X,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`Mdl`

Regression ensemble created by `fitrensemble`, or by the `compact` method.

`X`

Predictor data used to generate responses, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitrensemble`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names

during training, see the `PredictorNames` name-value pair argument of `fitrensemble`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`, where `NumTrained` is the number of weak learners.

Default: `1:NumTrained`

`UseObsForLearner`

A logical matrix of size `N`-by-`NumTrained`, where `N` is the number of observations in `X`, and `NumTrained` is the number of weak learners. When `UseObsForLearner(I,J)` is `true`, `predict` uses learner `J` in predicting observation `I`.

Default: `true(N,NumTrained)`

Output Arguments

`Yfit`

A numeric column vector with the same number of rows as `TBLdata` or `Xdata`. Each row of `Yfit` gives the predicted response to the corresponding row of `TBLdata` or `Xdata`, based on the `ens` regression model.

Examples

Predict Responses Based on Regression Ensemble

Find the predicted mileage for a car based on regression ensemble trained on the `carsmall` data.

Load the `carsmall` data set and select the number of cylinders, engine displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Cylinders Displacement Horsepower Weight];
```

Train an ensemble of regression trees and predict `MPG` for a four-cylinder car, with 200 cubic inch engine displacement, 150 horsepower, weighing 3000 lbs.

```
rens = fitrensemble(X,MPG);
Mileage = predict(rens,[4 200 150 3000])
```

```
Mileage = 25.6467
```

Alternative Functionality

Simulink Block

To integrate the prediction of an ensemble into Simulink, you can use the RegressionEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see “Predict Responses Using RegressionEnsemble Predict Block” on page 32-137 and “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- Generating fixed-point code for `predict` includes propagating data types for individual learners and, therefore, can be time consuming.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-841 of the `CompactRegressionEnsemble` object.

`X`	• For general code generation, `X` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• For fixed-point code generation, `X` must be a fixed-point matrix.
	• The number of rows, or observations, in `X` can be a variable size, but the number of columns in `X` must be fixed.
	• If you want to specify `X` as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	  For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow user-defined indices up to 5 weak learners in the generated code, include `{coder.Constant('Learners'),coder.typeof(0,[1,5],[0,1])}` in the `-args` value of `codegen`.
`'Learners'`	For fixed-point code generation, the `'Learners'` value must have an integer data type.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`CompactRegressionEnsemble` | `RegressionBaggedEnsemble` | `RegressionEnsemble` | `fitrensemble` | `loss`

predict

Predict response of Gaussian process regression model

Syntax

```
ypred = predict(gprMdl,Xnew)
[ypred,ysd] = predict(gprMdl,Xnew)
[ypred,ysd,yint] = predict(gprMdl,Xnew)
[ypred,ysd,yint] = predict(gprMdl,Xnew,Name,Value)
```

Description

`ypred = predict(gprMdl,Xnew)` returns the predicted responses `ypred` for the full or compact Gaussian process regression (GPR) model, `gprMdl`, and the predictor values in `Xnew`.

`[ypred,ysd] = predict(gprMdl,Xnew)` also returns the estimated standard deviations for the new responses at the predictor values in `Xnew` from a trained GPR model.

`[ypred,ysd,yint] = predict(gprMdl,Xnew)` also returns the 95% prediction intervals, `yint`, for the true responses corresponding to each row of `Xnew`.

`[ypred,ysd,yint] = predict(gprMdl,Xnew,Name,Value)` also returns the prediction intervals with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence level of the prediction interval.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object | `CompactRegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` (full) or `CompactRegressionGP` (compact) object.

`Xnew` — New values for the predictors

table | *m*-by-*d* matrix

New values for the predictors that `fitrgp` uses in training the GPR model, specified as a `table` or an *m*-by-*d* matrix. *m* is the number of observations and *d* is the number of predictor variables in the training data.

If you trained `gprMdl` on a `table`, then `Xnew` must be a `table` that contains all the predictor variables used to train `gprMdl`.

If you trained `gprMdl` on a matrix, then `Xnew` must be a numeric matrix with *d* columns.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level
0.05 (default) | scalar value in the range from 0 to 1

Significance level for the prediction intervals, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range from 0 to 1.

Example: `'Alpha',0.01` specifies 99% prediction intervals.

Data Types: `single` | `double`

Output Arguments

`ypred` — Predicted response values
n-by-1 vector

Predicted response values, returned as an *n*-by-1 vector.

`ysd` — Estimated standard deviation of the new response values
n-by-1 vector

Estimated standard deviation of the new response values, returned as an *n*-by-1 vector , where `ysd`(*i*), *i* = 1, 2, ..., *n*, contains the estimated standard deviation of the new response corresponding to the predictor values at the i^{th} row of `Xnew` from the trained GPR model.

`yint` — Prediction intervals for the true response values
n-by-2 matrix

Prediction intervals for the true response values corresponding to each row of `Xnew`, returned as an *n*-by-2 matrix. The first column of `yint` contains the lower limits and the second column contains the upper limits of the prediction intervals.

Examples

Compute Predicted Responses

Generate the sample data.

```
n = 10000;
rng(1) % For reproducibility
x = linspace(0.5,2.5,n)';
y = sin(10*pi.*x) ./ (2.*x)+(x-1).^4 + 1.5*rand(n,1);
```

Fit a GPR model using the Matern 3/2 kernel function with separate length scale for each predictor and an active set size of 100. Use the subset of regressors approximation method for parameter estimation and fully independent conditional method for prediction.

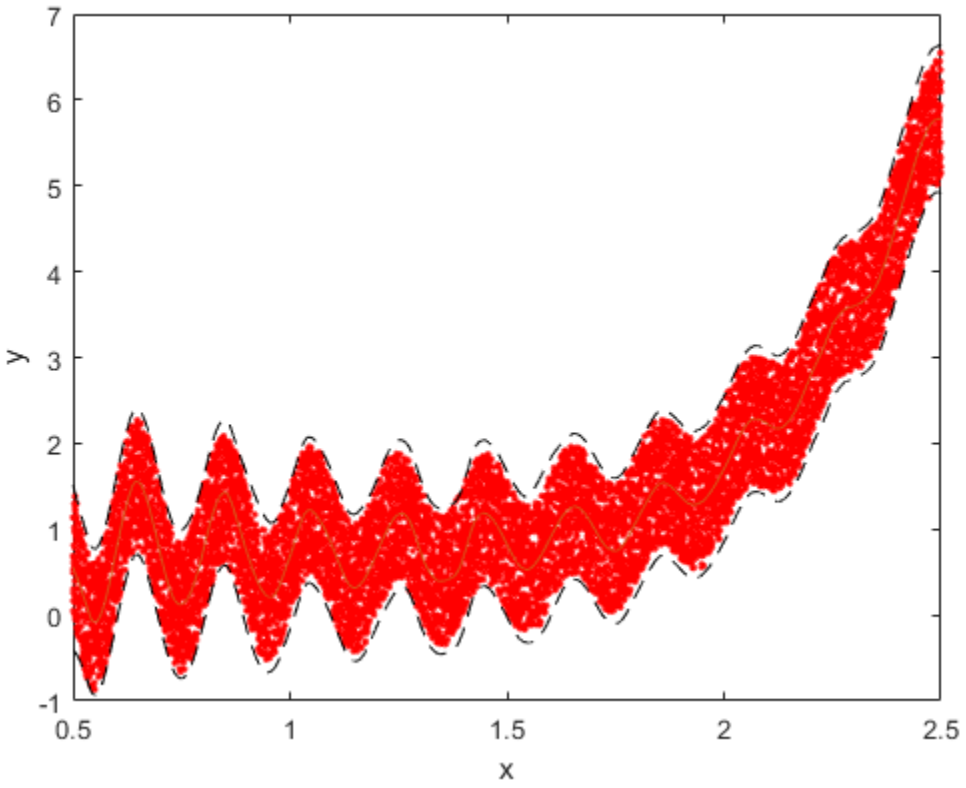
```
gprMdl = fitrgp(x,y,'KernelFunction','ardmatern32',...
'ActiveSetSize',100,'FitMethod','sr','PredictMethod','fic');
```

Compute the predictions.

```
[ypred,~,yci] = predict(gprMdl,x);
```

Plot the data along with the predictions and prediction intervals.

```
plot(x,y,'r.');
hold on
plot(x,ypred);
plot(x,yci(:,1),'k--');
plot(x,yci(:,2),'k--');
xlabel('x');
ylabel('y');
```



Compute Predictions When Data in Table

Load the sample data and store in a `table`.

```
load fisheriris
tbl = table(meas(:,1),meas(:,2),meas(:,3),meas(:,4),species,...
'VariableNames',{'meas1','meas2','meas3','meas4','species'});
```

Fit a GPR model using the first measurement as the response and the other variables as the predictors.

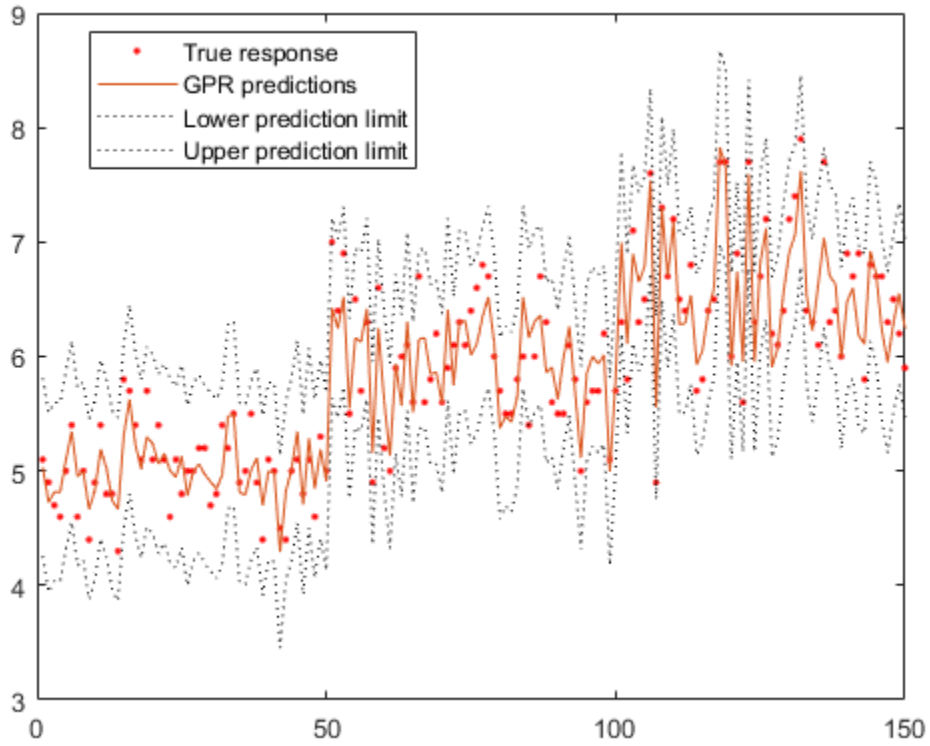
```
mdl = fitrgp(tbl,'meas1');
```

Compute the predictions and the 99% confidence intervals.

```
[ypred,~,yci] = predict(mdl,tbl,'Alpha',0.01);
```

Plot the true response and the predictions along with the prediction intervals.

```
figure();
plot(mdl.Y,'r.');
hold on;
plot(ypred);
plot(yci(:,1),'k:');
plot(yci(:,2),'k:');
legend('True response','GPR predictions',...
'Lower prediction limit','Upper prediction limit',...
'Location','Best');
```



Plot Predicted Response for Test Data

Load the sample data.

```
load('gprdata.mat');
```

The data contains training and test data. There are 500 observations in training data and 100 observations in test data. The data has 8 predictor variables. This is simulated data.

Fit a GPR model using the squared exponential kernel function with a separate length scale for each predictor. Standardize predictors in the training data. Use the exact fitting and prediction methods.

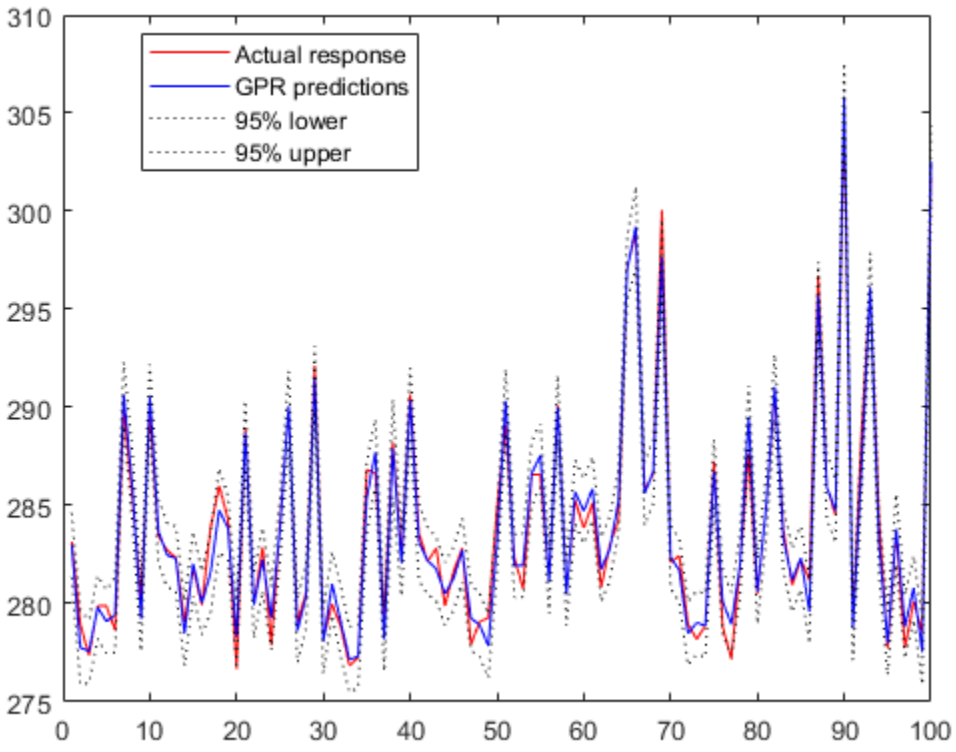
```
gprMdl = fitrgp(Xtrain,ytrain,'Basis','constant',...
'FitMethod','exact','PredictMethod','exact',...
'KernelFunction','ardsquaredexponential','Standardize',1);
```

Predict the responses for test data.

```
[ytestpred,~,ytestci] = predict(gprMdl,Xtest);
```

Plot the test response along with the predictions.

```
figure;
plot(ytest,'r');
hold on;
plot(ytestpred,'b');
plot(ytestci(:,1),'k:');
plot(ytestci(:,2),'k:');
legend('Actual response','GPR predictions',...
'95% lower','95% upper','Location','Best');
hold off
```



Tips

- You can choose the prediction method while training the GPR model using the `PredictMethod` name-value pair argument in `fitrgp`. The default prediction method is `'exact'` for $n \le 10000$,

where *n* is the number of observations in the training data, and `'bcd'` (block coordinate descent), otherwise.

- Computation of standard deviations, `ysd`, and prediction intervals, `yint`, is not supported when `PredictMethod` is `'bcd'`.
- If `gprMdl` is a `CompactRegressionGP` object, you cannot compute standard deviations, `ysd`, or prediction intervals, `yint`, for `PredictMethod` equal to `'sr'` or `'fic'`. To compute `ysd` and `yint` for `PredictMethod` equal to `'sr'` or `'fic'`, use the full regression (`RegressionGP`) object.

Alternatives

You can use `resubPredict` to compute the predicted responses for the trained GPR model at the observations in the training data.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-850 of the `CompactRegressionGP` object.

Argument	Notes and Limitations

Xnew	• Xnew must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in Xnew can be a variable size, but the number of columns in Xnew must be fixed.
	• If you want to specify Xnew as a table, then your model must be trained using a table, and you must ensure that your entry-point function for prediction:
	  • Accepts data as arrays
	  • Creates a table from the data input arguments and specifies the variable names in the table
	  • Passes the table to `predict`
	  For an example of this table workflow, see “Generate Code to Classify Data in Table” on page 32-100. For more information on using tables in code generation, see “Code Generation for Tables” (MATLAB Coder) and “Table Limitations for Code Generation” (MATLAB Coder).

Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined significance level in the generated code, include `{coder.Constant('Alpha'),0}` in the `-args` value of `codegen`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`CompactRegressionGP` | `RegressionGP` | `compact` | `fitrgp` | `loss` | `resubPredict`

Introduced in R2015b

predict

Package:

Predict responses using regression neural network

Syntax

```
yfit = predict(Mdl,X)
yfit = predict(Mdl,X,'ObservationsIn',dimension)
```

Description

`yfit = predict(Mdl,X)` returns predicted response values for the predictor data in the table or matrix `X` using the trained regression neural network model `Mdl`.

`yfit` is returned as a numeric vector, whose *i*th entry corresponds to the *i*th observation in `X`.

`yfit = predict(Mdl,X,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

Examples

Predict Test Set Response Using Regression Neural Network

Predict test set response values by using a trained regression neural network model.

Load the `patients` data set. Create a table from the data set. Each row corresponds to one patient, and each column corresponds to a diagnostic variable. Use the `Systolic` variable as the response variable, and the rest of the variables as predictors.

```
load patients
tbl = table(Age,Diastolic,Gender,Height,Smoker,Weight,Systolic);
```

Separate the data into a training set `tblTrain` and a test set `tblTest` by using a nonstratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default") % For reproducibility of the partition
c = cvpartition(size(tbl,1),"Holdout",0.30);
trainingIndices = training(c);
testIndices = test(c);
tblTrain = tbl(trainingIndices,:);
tblTest = tbl(testIndices,:);
```

Train a regression neural network model using the training set. Specify the `Systolic` column of `tblTrain` as the response variable. Specify to standardize the numeric predictors. By default, the neural network model has one fully connected layer with 10 outputs, excluding the final fully connected layer.

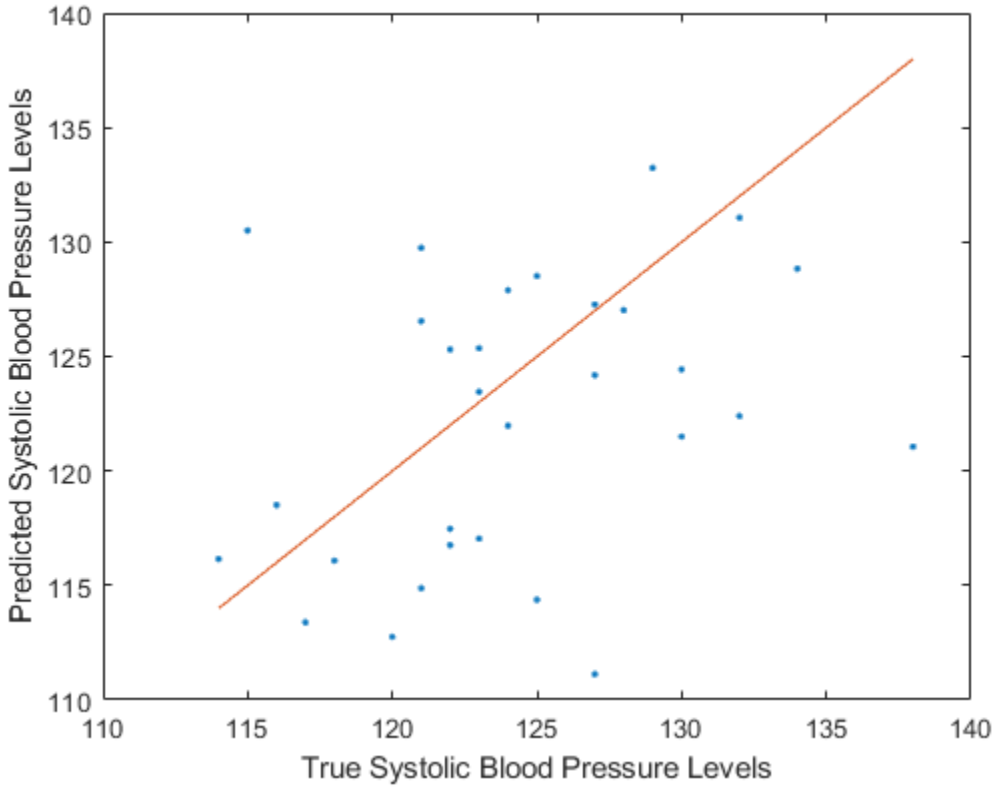
```
Mdl = fitrnet(tblTrain,"Systolic", ...
    "Standardize",true);
```

Predict the systolic blood pressure levels for patients in the test set.

```
predictedY = predict(Mdl,tblTest);
```

Visualize the results by using a scatter plot with a reference line. Plot the predicted values along the vertical axis and the true response values along the horizontal axis. Points on the reference line indicate correct predictions.

```
plot(tblTest.Systolic,predictedY,".")
hold on
plot(tblTest.Systolic,tblTest.Systolic)
hold off
xlabel("True Systolic Blood Pressure Levels")
ylabel("Predicted Systolic Blood Pressure Levels")
```



Because many of the points are far from the reference line, the default neural network model with a fully connected layer of size 10 does not seem to be a great predictor of systolic blood pressure levels.

Select Features to Include in Regression Neural Network

Perform feature selection by comparing test set losses and predictions. Compare the test set metrics for a regression neural network model trained using all the predictors to the test set metrics for a model trained using only a subset of the predictors.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table.

```
fishertable = readtable('fisheriris.csv');
```

Separate the data into a training set `trainTbl` and a test set `testTbl` by using a nonstratified holdout partition. The software reserves approximately 30% of the observations for the test data set and uses the rest of the observations for the training data set.

```
rng("default")
c = cvpartition(size(fishertable,1),"Holdout",0.3);
trainTbl = fishertable(training(c),:);
testTbl = fishertable(test(c),:);
```

Train one regression neural network model using all the predictors in the training set, and train another classifier using all the predictors except `PetalWidth`. For both models, specify `PetalLength` as the response variable, and standardize the predictors.

```
allMdl = fitrnet(trainTbl,"PetalLength","Standardize",true);
subsetMdl = fitrnet(trainTbl,"PetalLength ~ SepalLength + SepalWidth + Species", ...
    "Standardize",true);
```

Compare the test set mean squared error (MSE) of the two models. Smaller MSE values indicate better performance.

```
allMSE = loss(allMdl,testTbl)
```

```
allMSE = 0.0834
```

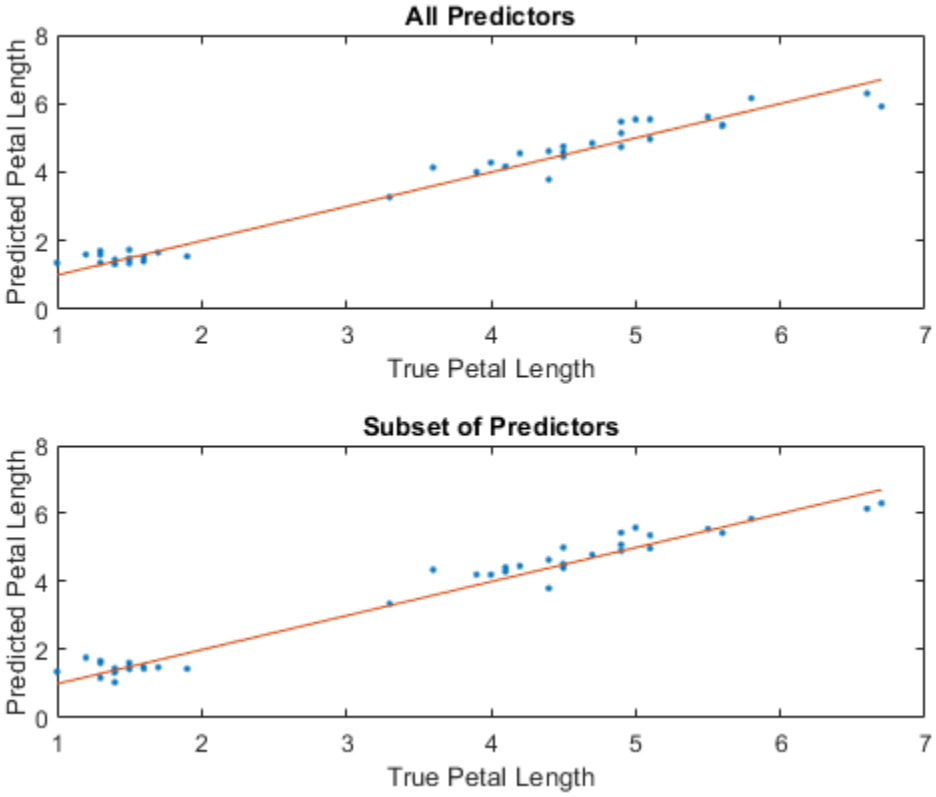
```
subsetMSE = loss(subsetMdl,testTbl)
```

```
subsetMSE = 0.0887
```

For each model, compare the test set predicted petal lengths to the true petal lengths. Plot the predicted petal lengths along the vertical axis and the true petal lengths along the horizontal axis. Points on the reference line indicate correct predictions.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
allPredictedY = predict(allMdl,testTbl);
plot(ax1,testTbl.PetalLength,allPredictedY,".")
hold on
plot(ax1,testTbl.PetalLength,testTbl.PetalLength)
hold off
xlabel(ax1,"True Petal Length")
ylabel(ax1,"Predicted Petal Length")
title(ax1,"All Predictors")

% Bottom axes
ax2 = nexttile;
```

```
subsetPredictedY = predict(subsetMdl,testTbl);
plot(ax2,testTbl.PetalLength,subsetPredictedY,".")
hold on
plot(ax2,testTbl.PetalLength,testTbl.PetalLength)
hold off
xlabel(ax2,"True Petal Length")
ylabel(ax2,"Predicted Petal Length")
title(ax2,"Subset of Predictors")
```



Because both models seems to perform well, with predictions scattered near the reference line, consider using the model trained using all predictors except `PetalWidth`.

Predict Using Layer Structure of Regression Neural Network Model

See how the layers of a regression neural network model work together to predict the response value for a single observation.

Load the sample file `fisheriris.csv`, which contains iris data including sepal length, sepal width, petal length, petal width, and species type. Read the file into a table, and display the first few rows of the table.

```
fishertable = readtable('fisheriris.csv');
head(fishertable)
```

```
ans=8×5 table
    SepalLength    SepalWidth    PetalLength    PetalWidth     Species
    ___________    __________    ___________    __________    __________

        5.1           3.5           1.4           0.2        {'setosa'}
        4.9             3           1.4           0.2        {'setosa'}
        4.7           3.2           1.3           0.2        {'setosa'}
        4.6           3.1           1.5           0.2        {'setosa'}
          5           3.6           1.4           0.2        {'setosa'}
        5.4           3.9           1.7           0.4        {'setosa'}
        4.6           3.4           1.4           0.3        {'setosa'}
          5           3.4           1.5           0.2        {'setosa'}
```

Train a regression neural network model using the data set. Specify the `PetalLength` variable as the response and use the other numeric variables as predictors.

```
Mdl = fitrnet(fishertable,"PetalLength ~ SepalLength + SepalWidth + PetalWidth");
```

Select the fifteenth observation from the data set. See how the layers of the neural network take the observation and return a predicted response value `newPointResponse`.

```
newPoint = Mdl.X{15,:}
```

```
newPoint = 1×3

    5.8000    4.0000    0.2000
```

```
firstFCStep = (Mdl.LayerWeights{1})*newPoint' + Mdl.LayerBiases{1};
reluStep = max(firstFCStep,0);

finalFCStep = (Mdl.LayerWeights{end})*reluStep + Mdl.LayerBiases{end};

newPointResponse = finalFCStep
```

```
newPointResponse = 1.6716
```

Check that the prediction matches the one returned by the `predict` object function.

```
predictedY = predict(Mdl,newPoint)
```

```
predictedY = 1.6716
```

```
isequal(newPointResponse,predictedY)
```

```
ans = logical
   1
```

The two results match.

Input Arguments

`Mdl` — Trained regression neural network

`RegressionNeuralNetwork` model object | `CompactRegressionNeuralNetwork` model object

Trained regression neural network, specified as a `RegressionNeuralNetwork` model object or `CompactRegressionNeuralNetwork` model object returned by `fitrnet` or `compact`, respectively.

`X` — Predictor data used to generate responses
numeric matrix | table

Predictor data used to generate responses, specified as a numeric matrix or table.

By default, each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`) and `Tbl` contains only numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value argument of `fitrnet`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` must be the same as the corresponding predictor variable names in `X`. To specify predictor names during training, use the `PredictorNames` name-value argument of `fitrnet`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

If you set `'Standardize',true` in `fitrnet` when training `Mdl`, then the software standardizes the numeric columns of the predictor data using the corresponding means and standard deviations.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `single` | `double` | `table`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'rows'` or `'columns'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in computation time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `char` | `string`

See Also

`CompactRegressionNeuralNetwork` | `RegressionNeuralNetwork` | `fitrnet` | `loss`

Topics

“Assess Regression Neural Network Performance” on page 18-184

Introduced in R2021a

predict

Predict responses using support vector machine regression model

Syntax

```
yfit = predict(Mdl,X)
```

Description

`yfit = predict(Mdl,X)` returns a vector of predicted responses for the predictor data in the table or matrix `X`, based on the full or compact, trained support vector machine (SVM) regression model `Mdl`.

Input Arguments

`Mdl` — SVM regression model

`RegressionSVM` object | `CompactRegressionSVM` object

SVM regression model, specified as a `RegressionSVM` model or a `CompactRegressionSVM` model, returned by `fitrsvm` or `compact`, respectively.

`X` — Predictor data used to generate responses

numeric matrix | table

Predictor data used to generate responses, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitrsvm`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names

during training, see the `PredictorNames` name-value pair argument of `fitrsvm`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

If you set `'Standardize',true` in `fitrsvm` to train `Mdl`, then the software standardizes the columns of `X` using the corresponding means in `Mdl.Mu` and standard deviations in `Mdl.Sigma`.

Data Types: `table` | `double` | `single`

Output Arguments

`yfit` — Predicted responses

vector

Predicted responses, returned as a vector of length *n*, where *n* is the number of observations in the training data.

For details about how to predict responses, see “Equation 25-1” and “Equation 25-2” in “Understanding Support Vector Machine Regression” on page 25-2.

Examples

Predict Test Sample Response for SVM Regression Model

Load the `carsmall` data set. Consider a model that predicts a car's fuel efficiency given its horsepower and weight. Determine the sample size.

```
load carsmall
tbl = table(Horsepower,Weight,MPG);
N = size(tbl,1);
```

Partition the data into training and test sets. Hold out 10% of the data for testing.

```
rng(10); % For reproducibility
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Train a linear SVM regression model. Standardize the data.

```
Mdl = fitrsvm(tbl(idxTrn,:),'MPG','Standardize',true);
```

`Mdl` is a `RegressionSVM` model.

Predict responses for the test set.

```
YFit = predict(Mdl,tbl(idxTest,:));
```

Create a table containing the observed response values and the predicted response values side by side.

```
table(tbl.MPG(idxTest),YFit,'VariableNames',...
    {'ObservedValue','PredictedValue'})
```

```
ans=10×2 table
    ObservedValue    PredictedValue
```

```
    ____________    _____________

      14              9.4833
      27              28.938
      10               7.765
      28              27.155
      22              21.054
      29              31.484
    24.5              30.306
    18.5               19.12
      32              28.225
      28              26.632
```

Tips

- If `mdl` is a cross-validated `RegressionPartitionedSVM` model, use `kfoldPredict` instead of `predict` to predict new response values.

Alternative Functionality

Simulink Block

To integrate the prediction of an SVM regression model into Simulink, you can use the RegressionSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see “Predict Responses Using RegressionSVM Predict Block” on page 32-115 and “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.

- Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
- Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.

- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-862 of the `CompactRegressionSVM` object.

`X`	• For general code generation, `X` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, `X` must be a single-precision or double-precision matrix.
	• For fixed-point code generation, `X` must be a fixed-point matrix.
	• The number of rows, or observations, in `X` can be a variable size, but the number of columns in `X` must be fixed.
	• If you want to specify `X` as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`CompactRegressionSVM` | `RegressionSVM` | `fitrsvm` | `kfoldPredict`

Introduced in R2015b

predict

Predict responses using regression tree

Syntax

```
Yfit = predict(Mdl,X)
Yfit = predict(Mdl,X,Name,Value)
[Yfit,node] = predict( ___ )
```

Description

`Yfit = predict(Mdl,X)` returns a vector of predicted responses for the predictor data in the table or matrix `X`, based on the full or compact regression tree `Mdl`.

`Yfit = predict(Mdl,X,Name,Value)` predicts response values with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify to prune `Mdl` to a particular level before predicting responses.

`[Yfit,node] = predict( ___ )` also returns a vector of predicted node numbers for the responses, using any of the input arguments in the previous syntaxes.

Input Arguments

`Mdl` — Trained regression tree

`RegressionTree` model object | `CompactRegressionTree` model object

Trained classification tree, specified as a `RegressionTree` or `CompactRegressionTree` model object. That is, `Mdl` is a trained classification model returned by `fitrtree` or `compact`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitrtree`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

- If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
- If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitrtree`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

Data Types: `table` | `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level
0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(Mdl.PruneList)`. `0` indicates the full, unpruned tree and `max(Mdl.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `predict` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(Mdl.PruneList)`.

`predict` prunes `Mdl` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `Mdl` must be nonempty. In other words, grow `Mdl` by setting `'Prune','on'`, or by pruning `Mdl` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`Yfit` — Predicted response values
numeric column vector

Predicted response values, returned as a numeric column vector with the same number of rows as `X`. Each row of `Yfit` gives the predicted response to the corresponding row of `X`, based on the `Mdl`.

`node` — Node numbers
numeric vector

Node numbers for the predictions, specified as a numeric vector. Each entry corresponds to the predicted leaf node in `Mdl` for the corresponding row of `X`.

Examples

Predict a Response Using a Regression Tree

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using the entire data set.

```
Mdl = fitrtree(X,MPG);
```

Predict the MPG for a car with 200 cubic inch engine displacement, 150 horsepower, and that weighs 3000 lbs.

```
X0 = [200 150 3000];
MPG0 = predict(Mdl,X0)
```

```
MPG0 = 21.9375
```

The regression tree predicts the car's efficiency to be 21.94 mpg.

Alternative Functionality

Simulink Block

To integrate the prediction of a regression tree model into Simulink, you can use the RegressionTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function. For examples, see “Predict Responses Using RegressionTree Predict Block” on page 32-127 and “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding which approach to use, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. You can use models trained on either in-memory or tall data with this function.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.
  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- You can also generate fixed-point C/C++ code for `predict`. Fixed-point code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function on page 33-2631 generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see “Code Generation” on page 33-868 of the `CompactRegressionTree` object.

Argument	Notes and Limitations

X	• For general code generation, X must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, X must be a single-precision or double-precision matrix.
	• For fixed-point code generation, X must be a fixed-point matrix.
	• The number of rows, or observations, in X can be a variable size, but the number of columns in X must be fixed.
	• If you want to specify X as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

`Subtrees`	• Names in name-value pair arguments must be compile-time constants. For example, to allow user-defined pruning levels in the generated code, include `{coder.Constant('Subtrees'),coder.typeof(0,[1,n],[0,1])}` in the `-args` value of `codegen`, where `n` is `max(Mdl.PruneList)`. • The `'Subtrees'` name-value pair argument is not supported in the coder configurer workflow. • For fixed-point code generation, the `'Subtrees'` value must be `coder.Constant('all')` or have an integer data type.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`CompactRegressionTree` | `RegressionTree` | `compact` | `fitrtree` | `loss`

Topics

"Predict Out-of-Sample Responses of Subtrees" on page 19-10
"Decision Trees" on page 19-2
"Prediction Using Classification and Regression Trees" on page 19-9
"Specify Variable-Size Arguments for Code Generation" on page 32-45

Introduced in R2011a

predict

Predict responses for new observations from linear model for incremental learning

Syntax

```
label = predict(Mdl,X)
label = predict(Mdl,X,'ObservationsIn',dimension)

[label,score] = predict( ___ )
```

Description

`label = predict(Mdl,X)` returns the predicted responses or labels `label` of the observations in the predictor data `X` from the incremental learning model `Mdl`.

`label = predict(Mdl,X,'ObservationsIn',dimension)` specifies the observation dimension of the predictor data, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that observations in the predictor data are oriented along the columns of `X`.

`[label,score] = predict( ___ )` also returns classification scores on page 33-4910 for all classes when `Mdl` is an incremental learning model for classification, using any of the input argument combinations in the previous syntaxes.

Examples

Predict Class Labels

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = actid > 2;
```

Fit a linear classification model to the entire data set.

```
TTMdl = fitclinear(feat,Y)
```

```
TTMdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60×1 double]
              Bias: -0.2005
```

```
           Lambda: 4.1537e-05
          Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear classification model.

Convert the traditionally trained linear classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

           IsWarm: 1
          Metrics: [1×2 table]
       ClassNames: [0 1]
    ScoreTransform: 'none'
             Beta: [60×1 double]
             Bias: -0.2005
          Learner: 'svm'


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object prepared for incremental learning using SVM.

- The `incrementalLearner` function initializes the incremental learner by passing learned coefficients to it, along with other information `TTMdl` learned from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.
- The `incrementalLearner` configures the model to be trained using the adaptive scale-invariant solver, whereas `fitclinear` trained `TTMdl` using the BFGS solver

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict class labels for all observations using both models.

```
ttlabels = predict(TTMdl,feat);
illables = predict(IncrementalMdl,feat);
sameLabels = sum(ttlabels ~= illables) == 0
```

```
sameLabels = logical
   1
```

Both models predict the same labels for each observation.

Specify Observation Orientation in Data

If you orient the observations along the columns of the predictor data matrix, you can experience an efficiency boost during incremental learning.

Load and shuffle the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015

rng(1) % For reproducibility
n = size(NYCHousing2015,1);
shuffidx = randsample(n,n);
NYCHousing2015 = NYCHousing2015(shuffidx,:);
```

Extract the response variable `SALEPRICE` from the table. Apply the log transform to `SALEPRICE`.

```
Y = log(NYCHousing2015.SALEPRICE + 1); % Add 1 to avoid log of 0
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data, and transpose the data to speed up computations.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}]';
```

Configure a linear regression model for incremental learning with no estimation period.

```
Mdl = incrementalRegressionLinear('Learner','leastsquares','EstimationPeriod',0);
```

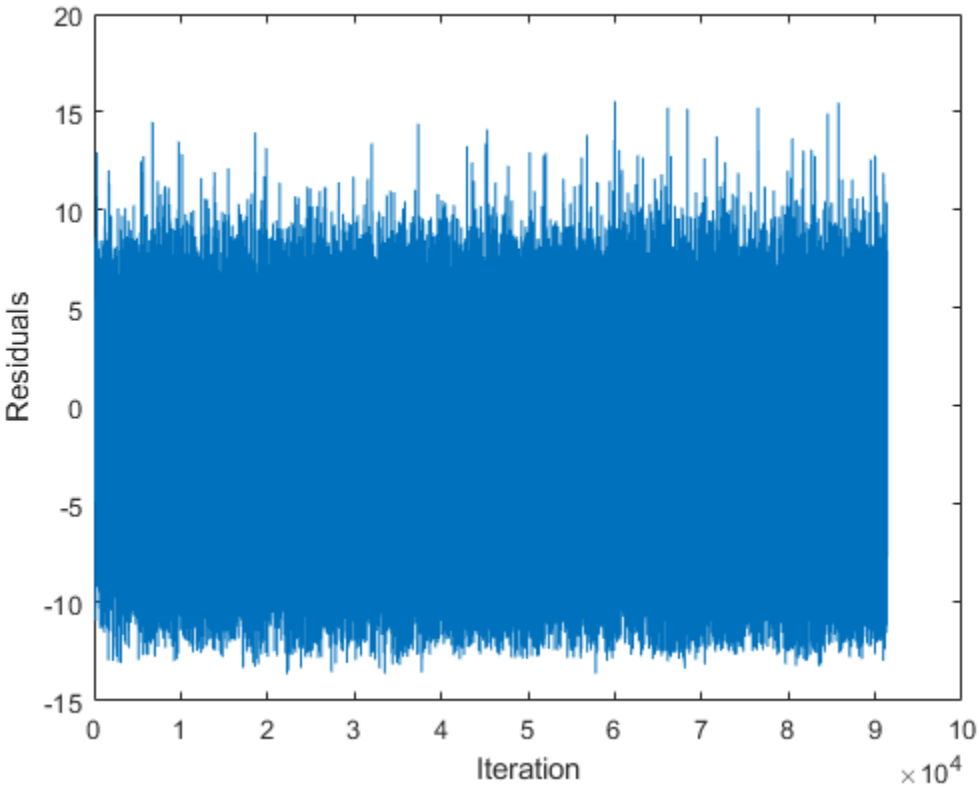
`Mdl` is an `incrementalRegressionLinear` model object.

Perform incremental learning and prediction by following this procedure for each iteration:

- Simulate a data stream by processing a chunk of 100 observations at a time.
- Fit the model to the incoming chunk of data. Specify that the observations are oriented along the columns of the data. Overwrite the previous incremental model with the new model.
- Predict responses using the fitted model and the incoming chunk of data. Specify that the observations are oriented along the columns of the data.

```
% Preallocation
numObsPerChunk = 100;
n = numel(Y);
nchunk = floor(n/numObsPerChunk);
r = nan(n,1);

figure
h = plot(r);
h.YDataSource = 'r';
```

```
ylabel('Residuals')
xlabel('Iteration')

% Incremental fitting
for j = 2:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = fit(Mdl,X(:,idx),Y(idx),'ObservationsIn','columns');
    yhat = predict(Mdl,X(:,idx),'ObservationsIn','columns');
    r(idx) = Y(idx) - yhat;
    refreshdata
    drawnow
end
```



`Mdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

The residuals appear symmetrically spread around 0 throughout incremental learning.

Compute Posterior Class Probabilities

To compute posterior class probabilities, specify a logistic regression incremental learner.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(10); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Create an incremental logistic regression model for binary classification. Prepare it for `predict` by specifying the class names and arbitrary coefficient and bias values.

```
p = size(X,2);
Beta = randn(p,1);
Bias = randn(1);
Mdl = incrementalClassificationLinear('Learner','logistic','Beta',Beta,...
    'Bias',Bias,'ClassNames',unique(Y));
```

`Mdl` is an `incrementalClassificationLinear` model. All its properties are read-only. Instead of specifying arbitrary values, you can take either of these actions to prepare the model:

- Train a logistic regression model for binary classification using `fitclinear` on a subset of the data (if available), and then convert the model to an incremental learner by using `incrementalLearner`.
- Incrementally fit `Mdl` to data by using `fit`.

Simulate a data stream, and perform the following actions on each incoming chunk of 50 observations.

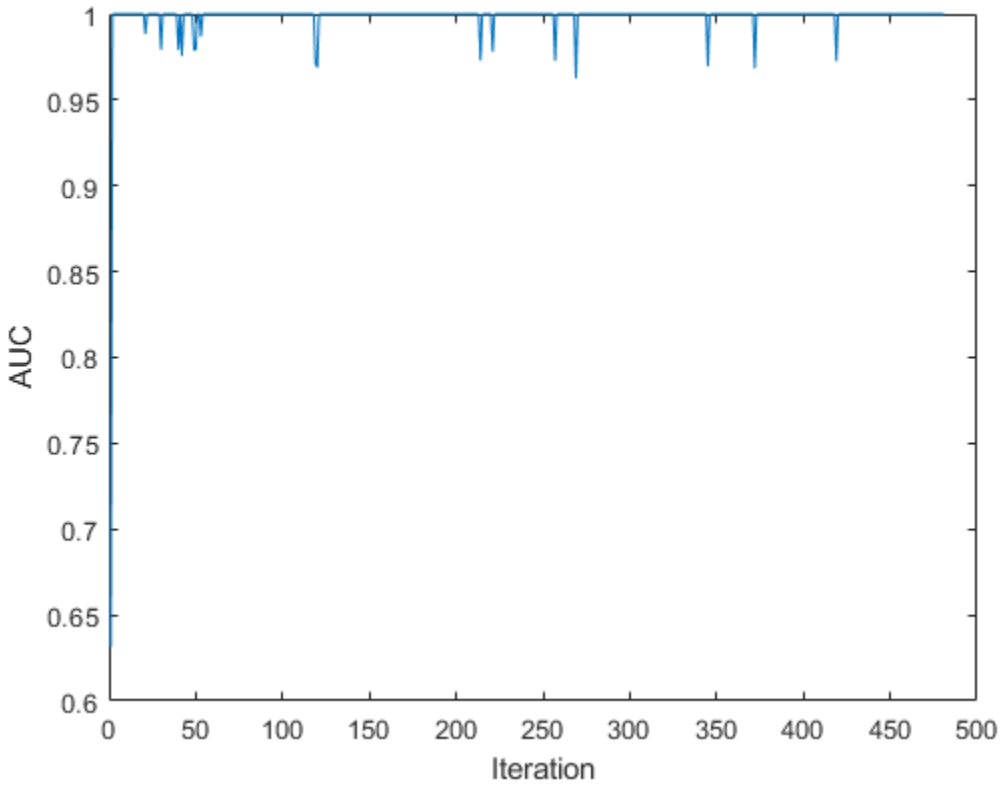
1. Call `predict` to predict classification scores for the observations in the incoming chunk of data. The classification scores are posterior class probabilities for logistic regression learners.
2. Call `perfcurve` to compute the area under the ROC curve (AUC) using the incoming chunk of data, and store the result.
3. Call `fit` to fit the model to the incoming chunk. Overwrite the previous incremental model with a new one fitted to the incoming observation.

```
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
auc = zeros(nchunk,1);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    [~,posteriorProb] = predict(Mdl,X(idx,:));
    [~,~,~,auc(j)] = perfcurve(Y(idx),posteriorProb(:,2),Mdl.ClassNames(2));
    Mdl = fit(Mdl,X(idx,:),Y(idx));
end
```

`Mdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

Plot the AUC on the incoming chunks of data.

```
plot(auc)
ylabel('AUC')
xlabel('Iteration')
```



The AUC suggests that the classifier correctly predicts moving subjects well.

Input Arguments

`Mdl` — Incremental learning model

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Incremental learning model, specified as an `incrementalClassificationLinear` or `incrementalRegressionLinear` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

You must configure `Mdl` to predict labels for a batch of observations.

- If `Mdl` is a converted, traditionally trained model, you can predict labels without any modifications.
- Otherwise, `Mdl` must satisfy the following criteria, which you can specify directly or by fitting `Mdl` to data using `fit` or `updateMetricsAndFit`.

- If `Mdl` is an `incrementalRegressionLinear` model, its model coefficients `Mdl.Beta` and bias `Mdl.Bias` must be nonempty arrays.
- If `Mdl` is an `incrementalClassificationLinear` model, its model coefficients `Mdl.Beta` and bias `Mdl.Bias` must be nonempty arrays and the class names in `Mdl.ClassNames` must contain two classes.
- Regardless of object type, if you configure the model so that functions standardize predictor data, the predictor means `Mdl.Mu` and standard deviations `Mdl.Sigma` must be nonempty arrays.

`X` — Batch of predictor data
floating-point matrix

Batch of predictor data for which to predict labels, specified as a floating-point matrix of *n* observations and `Mdl.NumPredictors` predictor variables. The value of `dimension` determines the orientation of the variables and observations.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note `predict` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'columns'` or `'rows'`.

Example: `'ObservationsIn','columns'`

Data Types: `char` | `string`

Output Arguments

`label` — Predicted responses
categorical array | character array | string vector | logical vector | cell array of character vectors | floating-point vector

Predicted responses (or labels), returned as a categorical or character array; floating-point, logical, or string vector; or cell array of character vectors with *n* rows. *n* is the number of observations in `X`, and `label(`*j*`)` is the predicted response for observation *j*.

- For classification problems, `label` has the same data type as the class names stored in `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- For regression problems, `label` is a floating-point vector.

`score` — Classification scores
floating-point matrix

Classification scores on page 33-4910, returned as an *n*-by-2 floating-point matrix when `Mdl` is an `incrementalClassificationLinear` model. *n* is the number of observations in `X`. `score(j,k)` is the score for classifying observation *j* into class *k*. `Mdl.ClassNames` specifies the order of the classes.

If `Mdl.Learner` is `'svm'`, `predict` returns raw classification scores. If `Mdl.Learner` is `'logistic'`, classification scores are posterior probabilities.

More About

Classification Score

For linear incremental learning models for binary classification, the raw classification score for classifying the observation *x*, a row vector, into the positive class is

$$f(x) = \beta_0 + x\beta,$$

where

- β_0 is the scalar bias `Mdl.Bias`.
- β is the column vector of coefficients `Mdl.Beta`.

The raw classification score for classifying *x* into the negative class is –*f*(*x*). The software classifies observations into the class that yields the positive score.

If the linear classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For usage notes and limitations of the model object, see `incrementalClassificationLinear` or `incrementalRegressionLinear`.

Argument	Notes and Limitations
X	• Batch-to-batch, the number of observations can be a variable size. • The number of predictor variables must equal to `Mdl.NumPredictors`. • `X` must be `single` or `double`.

- The following restrictions apply:
  - If you configure `Mdl` to shuffle data (`Mdl.Shuffle` is `true`, or `Mdl.Solver` is `'sgd'` or `'asgd'`), the `predict` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB. Therefore, if you fit `Mdl` before generating predictions, the predictions computed in MATLAB and those computed by the generated code might not be equal.
  - Use a homogeneous data type for all floating-point input arguments and object properties, specifically, either `single` or `double`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Objects
`incrementalClassificationLinear` | `incrementalRegressionLinear`

Functions
`fit` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8

Introduced in R2020b

predict

Predict responses for new observations from naive Bayes classification model for incremental learning

Syntax

```
label = predict(Mdl,X)
```

```
[label,Posterior,Cost] = predict(Mdl,X)
```

Description

`label = predict(Mdl,X)` returns the predicted responses or labels `label` of the observations in the predictor data `X` from the naive Bayes classification model for incremental learning `Mdl`.

`[label,Posterior,Cost] = predict(Mdl,X)` also returns the posterior probabilities on page 33-4917 (`Posterior`) and predicted (expected) misclassification costs on page 33-4916 (`Cost`) corresponding to the observations (rows) in `X`. For each observation in `X`, the predicted class label corresponds to the minimum expected classification cost among all classes.

Examples

Predict Class Labels

Load the human activity data set.

```
load humanactivity
```

For details on the data set, enter `Description` at the command line.

Fit a naive Bayes classification model to the entire data set.

```
TTMdl = fitcnb(feat,actid)
```

```
TTMdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 24075
        DistributionNames: {1×60 cell}
   DistributionParameters: {5×60 cell}


  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained model.

Convert the traditionally trained model to a naive Bayes classification model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 1
                  Metrics: [1×2 table]
               ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object prepared for incremental learning.

- The `incrementalLearner` function initializes the incremental learner by passing learned conditional predictor distribution parameters to it, along with other information `TTMdl` learned from the training data.
- `IncrementalMdl` is warm (`IsWarm` is `1`), which means that incremental learning functions can start tracking performance metrics.

An incremental learner created from converting a traditionally trained model can generate predictions without further processing.

Predict class labels for all observations using both models.

```
ttlabels = predict(TTMdl,feat);
illables = predict(IncrementalMdl,feat);
sameLabels = sum(ttlabels ~= illables) == 0
```

```
sameLabels = logical
   1
```

Both models predict the same labels for each observation.

Compute Posterior Class Probabilities

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(10); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Create a naive Bayes classification model for incremental learning; specify the class names. Prepare it for `predict` by fitting the model to the first 10 observations.

```
Mdl = incrementalClassificationNaiveBayes('ClassNames',unique(Y));
initobs = 10;
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs));
canPredict = size(Mdl.DistributionParameters,1) == numel(Mdl.ClassNames)
```

```
canPredict = logical
   1
```

`Mdl` is an `incrementalClassificationNaiveBayes` model. All its properties are read-only. The model is configured to generate predictions.

Simulate a data stream, and perform the following actions on each incoming chunk of 100 observations.

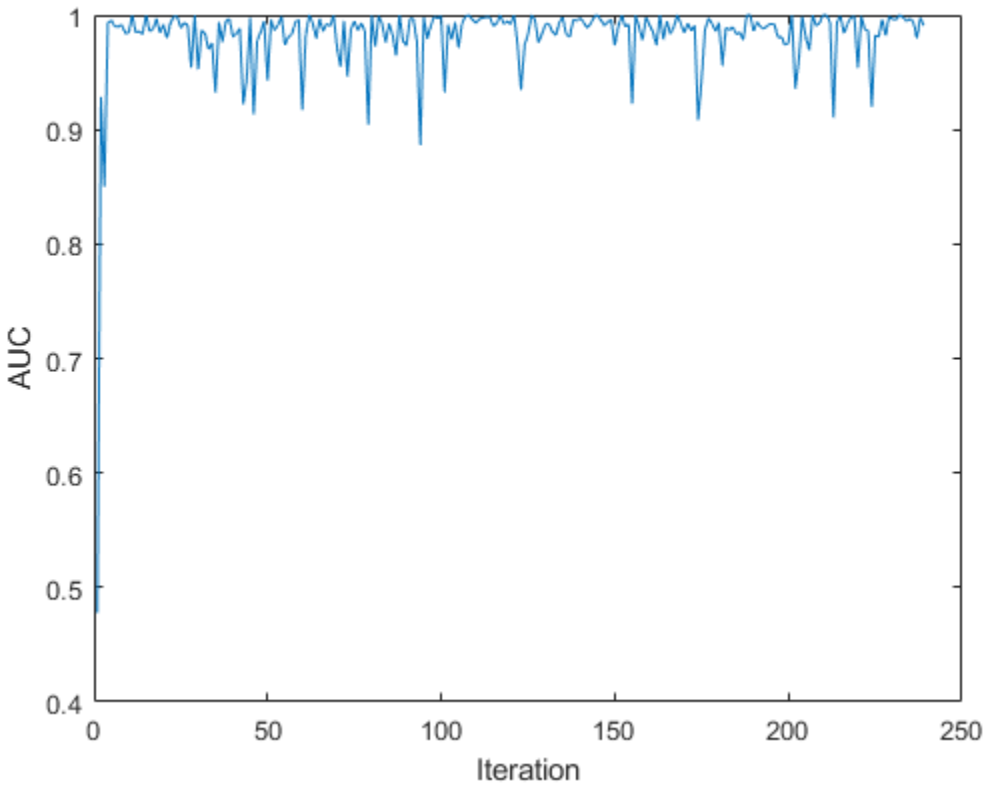
1. Call `predict` to compute class posterior probabilities for each observation in the incoming chunk of data.
2. Consider incrementally measuring how well the model predicts whether a subject is dancing (Y is 5). You can accomplish this by computing the AUC of an ROC curve created by passing, for each observation in the chunk, the difference between the posterior probability of class 5 and the maximum posterior probability among the other classes to `perfcurve`.
3. Call `fit` to fit the model to the incoming chunk. Overwrite the previous incremental model with a new one fitted to the incoming observation.

```
numObsPerChunk = 100;
nchunk = floor((n - initobs)/numObsPerChunk) - 1;
Posterior = zeros(nchunk,numel(Mdl.ClassNames));
auc = zeros(nchunk,1);
classauc = 5;

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    [~,Posterior(idx,:)] = predict(Mdl,X(idx,:));
    diffscore = Posterior(idx,classauc) - max(Posterior(idx,setdiff(Mdl.ClassNames,classauc)),[]
    [~,~,~,auc(j)] = perfcurve(Y(idx),diffscore,Mdl.ClassNames(classauc));
    Mdl = fit(Mdl,X(idx,:),Y(idx));
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

Plot the AUC on the incoming chunks of data.

```
plot(auc)
ylabel('AUC')
xlabel('Iteration')
```



The AUC suggests that the classifier correctly predicts dancing subjects well during incremental learning.

Input Arguments

`Mdl` — Naive Bayes classification model for incremental learning

`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

You must configure `Mdl` to predict labels for a batch of observations.

- If `Mdl` is a converted, traditionally trained model, you can predict labels without any modifications.
- Otherwise, `Mdl.DistributionParameters` must be a cell matrix with `Mdl.NumPredictors` > 0 columns and at least one row, where each row corresponds to each class name in `Mdl.ClassNames`.

`X` — Batch of predictor data

floating-point matrix

Batch of predictor data for which to predict labels, specified as an *n*-by-`Mdl.NumPredictors` floating-point matrix.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note `predict` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see "Dummy Variables" on page 2-48.

Data Types: `single` | `double`

Output Arguments

`label` — Predicted responses

categorical array | character array | string vector | logical vector | cell array of character vectors | floating-point vector

Predicted responses (or labels), returned as a categorical or character array; floating-point, logical, or string vector; or cell array of character vectors with *n* rows. *n* is the number of observations in `X`, and `label(`*j*`)` is the predicted response for observation *j*.

`label` has the same data type as the class names stored in `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)

`Posterior` — Class posterior probabilities

floating-point matrix

Class posterior probabilities on page 33-4917, returned as an *n*-by-2 floating-point matrix. `Posterior(`*j*`,`*k*`)` is the posterior probability that observation *j* is in class *k*. `Mdl.ClassNames` specifies the order of the classes.

`Cost` — Expected misclassification costs

floating-point matrix

Expected misclassification costs on page 33-4916, returned as an *n*-by-`numel(Mdl.ClassNames)` floating-point matrix.

`Cost(`*j*`,`*k*`)` is the expected misclassification cost of the observation in row *j* of `X` predicted into class *k* (`Mdl.ClassNames(`*k*`)`).

More About

Misclassification Cost

A misclassification cost is the relative severity of a classifier labeling an observation into the wrong class.

There are two types of misclassification costs: true and expected. Let *K* be the number of classes.

- True misclassification cost — A K-by-K matrix, where element (i,j) indicates the misclassification cost of predicting an observation into class j if its true class is i. The software stores the misclassification cost in the property `Mdl.Cost`, and uses it in computations. By default, `Mdl.Cost(i,j)` = 1 if `i` ≠ `j`, and `Mdl.Cost(i,j)` = 0 if `i` = `j`. In other words, the cost is `0` for correct classification and `1` for any incorrect classification.
- Expected misclassification cost — A K-dimensional vector, where element k is the weighted average misclassification cost of classifying an observation into class k, weighted by the class posterior probabilities.

  $$c_k = \sum_{j=1}^{K} \widehat{P}(Y = j \,|\, x_1, \ldots, x_P) Cost_{jk}\,.$$

  In other words, the software classifies observations to the class corresponding with the lowest expected misclassification cost.

Posterior Probability

The posterior probability is the probability that an observation belongs in a particular class, given the data.

For naive Bayes, the posterior probability that a classification is k for a given observation $(x_1,\ldots,x_P)$ is

$$\widehat{P}(Y = k \,|\, x_1, .., x_P) = \frac{P(X_1, \ldots, X_P \,|\, y = k)\pi(Y = k)}{P(X_1, \ldots, X_P)},$$

where:

- $P(X_1, \ldots, X_P \,|\, y = k)$ is the conditional joint density of the predictors given they are in class k. `Mdl.DistributionNames` stores the distribution names of the predictors.
- $\pi(Y = k)$ is the class prior probability distribution. `Mdl.Prior` stores the prior distribution.
- $P(X_1, .., X_P)$ is the joint density of the predictors. The classes are discrete, so

  $$P(X_1, \ldots, X_P) = \sum_{k=1}^{K} P(X_1, \ldots, X_P \,|\, y = k)\pi(Y = k)\,.$$

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `updateMetrics` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8

Introduced in R2021a

predict

Package:

Predict responses using generalized additive model (GAM)

Syntax

```
yFit = predict(Mdl,X)
yFit = predict(Mdl,X,'IncludeInteractions',includeInteractions)
```

Description

`yFit = predict(Mdl,X)` returns a vector of predicted responses for the predictor data in the table or matrix `X`, based on the generalized additive model `Mdl` for regression. The trained model can be either full or compact.

`yFit = predict(Mdl,X,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations.

Examples

Predict Test Sample Response

Train a generalized additive model using training samples, and then predict the test sample responses.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Randomly partition observations into a training set and a test set. Specify a 10% holdout sample for testing.

```
rng('default') % For reproducibility
cv = cvpartition(size(tbl,1),'HoldOut',0.10);
```

Extract the training and test indices.

```
trainInds = training(cv);
testInds = test(cv);
```

Train a univariate GAM that contains the linear terms for the predictors in `tbl`.

```
Mdl = fitrgam(tbl(trainInds,:),'Systolic')
```

```
Mdl =
  RegressionGAM
```

```
           PredictorNames: {1x6 cell}
             ResponseName: 'Systolic'
    CategoricalPredictors: [3 5 6]
        ResponseTransform: 'none'
                Intercept: 122.7444
          NumObservations: 90


  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object.

Predict responses for the test set.

```
yFit = predict(Mdl,tbl(testInds,:));
```

Create a table containing the observed response values and the predicted response values.

```
table(tbl.Systolic(testInds),yFit, ...
    'VariableNames',{'Observed Value','Predicted Value'})
```

```
ans=10×2 table
    Observed Value    Predicted Value
    ______________    _______________

         124              126.58
         121              123.95
         130              116.72
         115              117.35
         121              117.45
         116               118.5
         123              126.16
         132              124.14
         125              127.36
         124              115.99
```

Compare Predicted Responses

Predict responses for new observations using a generalized additive model that contains both linear and interaction terms for predictors. Use a memory-efficient model object, and specify whether to include interaction terms when predicting responses.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 10 observations for the new test data set.

```
rng('default')
n = size(X,1);
newInds = randsample(n,10);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a GAM that contains all the available linear and interaction terms in `X`.

```
Mdl = fitrgam(X(inds,:),Y(inds),'Interactions','all');
```

`Mdl` is a `RegressionGAM` model object.

Conserve memory by reducing the size of the trained model.

```
CMdl = compact(Mdl);
whos('Mdl','CMdl')
```

```
  Name      Size              Bytes  Class                                                Attributes

  CMdl      1x1             1228122  classreg.learning.regr.CompactRegressionGAM
  Mdl       1x1             1262143  RegressionGAM
```

`CMdl` is a `CompactRegressionGAM` model object.

Predict the responses using both linear and interaction terms, and then using only linear terms. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
yFit = predict(CMdl,XNew);
yFit_nointeraction = predict(CMdl,XNew,'IncludeInteractions',false);
```

Create a table containing the observed response values and the predicted response values.

```
t = table(YNew,yFit,yFit_nointeraction, ...
    'VariableNames',{'Observed Response', ...
    'Predicted Response','Predicted Response Without Interactions'})
```

```
t=10×3 table
    Observed Response    Predicted Response    Predicted Response Without Interactions
    _________________    __________________    ______________________________________

          27.9                 23.04                        23.649
           NaN                37.163                        35.779
           NaN                25.876                        21.978
            13                12.786                        14.141
            36                28.889                        27.281
          19.9                22.199                        18.451
          24.2                23.995                        24.885
            12                14.247                        13.982
            38                33.797                        33.528
            13                12.225                        11.127
```

Input Arguments

`Mdl` — Generalized additive model

`RegressionGAM` model object | `CompactRegressionGAM` model object

Generalized additive model, specified as a `RegressionGAM` or a `CompactRegressionGAM` model object.

`X` — Predictor data
numeric matrix | table

Predictor data, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that make up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table, then `X` can be a numeric matrix if the table contains all numeric predictor variables.
- For a table:
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those in `Tbl`. However, the column order of `X` does not need to correspond to the column order of `Tbl`.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and the corresponding predictor variable names in `X` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `X` must be numeric vectors.
  - `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

Data Types: `table` | `double` | `single`

`includeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`.

The default `includeInteractions` value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

Output Arguments

`yFit` — Predicted responses
vector

Predicted responses, returned as a vector of length *n*, where *n* is the number of observations in the predictor data `X`.

See Also
`loss` | `resubPredict`

Topics
“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

predict

Class: `RegressionLinear`

Predict response of linear regression model

Syntax

```
YHat = predict(Mdl,X)
YHat = predict(Mdl,X,'ObservationsIn',dimension)
```

Description

`YHat = predict(Mdl,X)` returns predicted responses for each observation in the predictor data `X` based on the trained linear regression model `Mdl`. `YHat` contains responses for each regularization strength in `Mdl`.

`YHat = predict(Mdl,X,'ObservationsIn',dimension)` specifies the predictor data observation dimension, either `'rows'` (default) or `'columns'`. For example, specify `'ObservationsIn','columns'` to indicate that columns in the predictor data correspond to observations.

Input Arguments

`Mdl` — Linear regression model

`RegressionLinear` model object

Linear regression model, specified as a `RegressionLinear` model object. You can create a `RegressionLinear` model object using `fitrlinear`.

`X` — Predictor data used to generate responses

full numeric matrix | sparse numeric matrix | table

Predictor data used to generate responses, specified as a full or sparse numeric matrix or a table.

By default, each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you train `Mdl` using a table (for example, `Tbl`) and `Tbl` contains only numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value pair argument of `fitrlinear`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:

- `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
- If you train `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as the variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
- If you train `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` must be the same as the corresponding predictor variable names in `X`. To specify predictor names during training, use the `PredictorNames` name-value pair argument of `fitrlinear`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Data Types: `double` | `single` | `table`

`dimension` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as `'columns'` or `'rows'`.

Note If you orient your predictor matrix so that observations correspond to columns and specify `'ObservationsIn','columns'`, then you might experience a significant reduction in optimization execution time. You cannot specify `'ObservationsIn','columns'` for predictor data in a table.

Output Arguments

`YHat` — Predicted responses
numeric matrix

Predicted responses, returned as a *n*-by-*L* numeric matrix. *n* is the number of observations in `X` and *L* is the number of regularization strengths in `Mdl.Lambda`. `YHat(i,j)` is the response for observation `i` using the linear regression model that has regularization strength `Mdl.Lambda(j)`.

The predicted response using the model with regularization strength j is $\widehat{y}_j = x\beta_j + b_j$.

- x is an observation from the predictor data matrix `X`, and is row vector.
- β_j is the estimated column vector of coefficients. The software stores this vector in `Mdl.Beta(:,j)`.
- b_j is the estimated, scalar bias, which the software stores in `Mdl.Bias(j)`.

Examples

Predict Test-Sample Responses

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model. Reserve 30% of the observations as a holdout sample.

```
CVMdl = fitrlinear(X,Y,'Holdout',0.3);
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0066
               Lambda: 1.4286e-04
              Learner: 'svm'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedLinear` model. It contains the property `Trained`, which is a 1-by-1 cell array holding a `RegressionLinear` model that the software trained using the training set.

Extract the training and test data from the partition definition.

```
trainIdx = training(CVMdl.Partition);
testIdx = test(CVMdl.Partition);
```

Predict the training- and test-sample responses.

```
yHatTrain = predict(Mdl,X(trainIdx,:));
yHatTest = predict(Mdl,X(testIdx,:));
```

Because there is one regularization strength in `Mdl`, `yHatTrain` and `yHatTest` are numeric vectors.

Predict from Best-Performing Model

Predict responses from the best-performing, linear regression model that uses a lasso-penalty and least squares.

Simulate 10000 observations as in “Predict Test-Sample Responses” on page 33-4924.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-5} through 10^{-1}.

```
Lambda = logspace(-5,-1,15);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Optimize the objective function using SpaRSA.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','KFold',5,'Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');

numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `RegressionPartitionedLinear` model. Because `fitrlinear` implements 5-fold cross-validation, `CVMdl` contains 5 `RegressionLinear` models that the software trains on each fold.

Display the first trained linear regression model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 = 
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x15 double]
                 Bias: [1x15 double]
               Lambda: [1x15 double]
              Learner: 'leastsquares'


  Properties, Methods
```

`Mdl1` is a `RegressionLinear` model object. `fitrlinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 11 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated MSE.

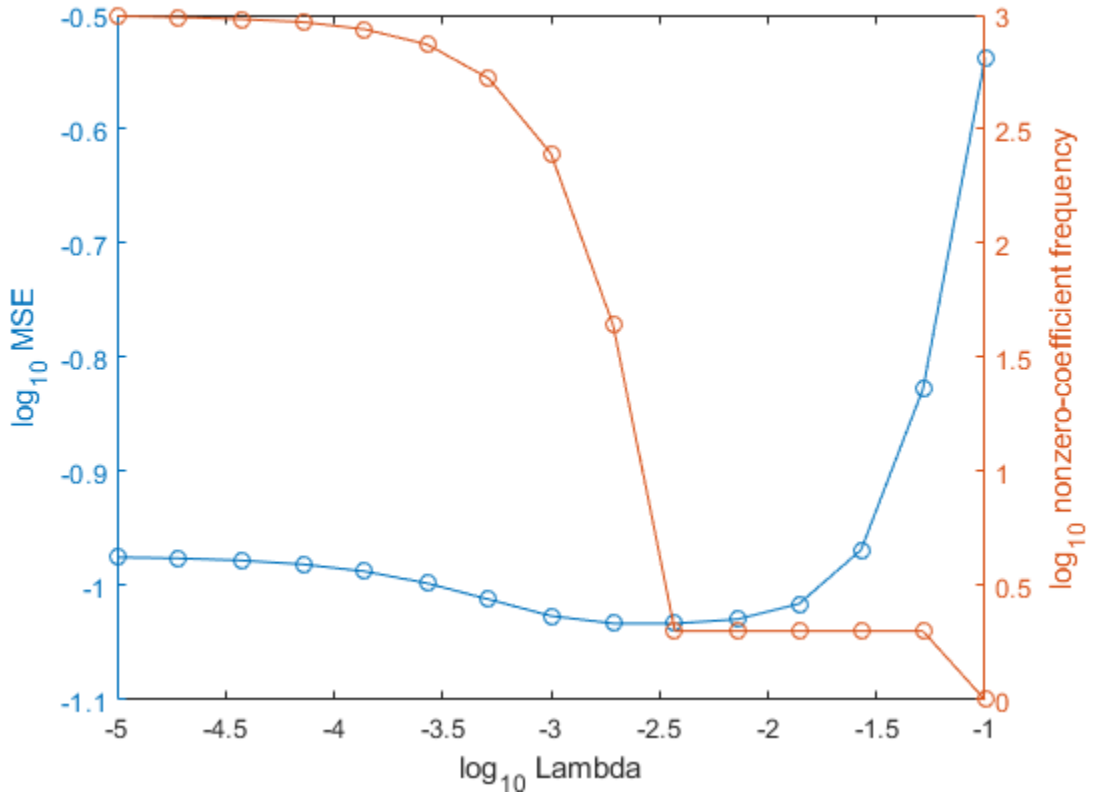
```
mse = kfoldLoss(CVMdl);
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. For each regularization strength, train a linear regression model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low MSE (for example, `Lambda(10)`).

```
idxFinal = 10;
```

Extract the model with corresponding to the minimal MSE.

```
MdlFinal = selectModels(Mdl,idxFinal)
```

```
MdlFinal =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0050
```

```
               Lambda: 0.0037
              Learner: 'leastsquares'


  Properties, Methods
```

```
idxNZCoeff = find(MdlFinal.Beta~=0)
```

```
idxNZCoeff = 2×1

   100
   200
```

```
EstCoeff = Mdl.Beta(idxNZCoeff)
```

```
EstCoeff = 2×1

    1.0051
    1.9965
```

`MdlFinal` is a `RegressionLinear` model with one regularization strength. The nonzero coefficients `EstCoeff` are close to the coefficients that simulated the data.

Simulate 10 new observations, and predict corresponding responses using the best-performing model.

```
XNew = sprandn(d,10,nz);
YHat = predict(MdlFinal,XNew,'ObservationsIn','columns');
```

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `predict` does not support tall `table` data.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You can generate C/C++ code for both `predict` and `update` by using a coder configurer. Or, generate code only for `predict` by using `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`.
  - Code generation for `predict` and `update` — Create a coder configurer by using `learnerCoderConfigurer` and then generate code by using `generateCode`. Then you can update model parameters in the generated code without having to regenerate the code.

  - Code generation for `predict` — Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.

- To generate single-precision C/C++ code for predict, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-5330 of the `RegressionLinear` object.

`X`	• For general code generation, `X` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• In the coder configurer workflow, `X` must be a single-precision or double-precision matrix.
	• The number of observations in `X` can be a variable size, but the number of variables in `X` must be fixed.
	• If you want to specify `X` as a table, then your model must be trained using a table, and your entry-point function for prediction must:
	  • Accept data as arrays.
	  • Create a table from the data input arguments and specifies the variable names in the table.
	  • Pass the table to `predict`.
	For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

Name-value pair arguments	• Names in name-value pair arguments must be compile-time constants. • The value for the `'ObservationsIn'` name-value pair argument must be a compile-time constant. For example, to use the `'ObservationsIn','columns'` name-value pair argument in the generated code, include `{coder.Constant('ObservationsIn'),coder.Constant('columns')}` in the `-args` value of `codegen`.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`RegressionLinear` | `fitrlinear`

Introduced in R2016a

predict

Class: `CompactTreeBagger`

Predict responses using ensemble of bagged decision trees

Syntax

```
Yfit = predict(B,X)
Yfit = predict(B,X,Name,Value)
[Yfit,stdevs] = predict( ___ )
[Yfit,scores] = predict( ___ )
[Yfit,scores,stdevs] = predict( ___ )
```

Description

`Yfit = predict(B,X)` returns a vector of predicted responses for the predictor data in the table or matrix `X`, based on the compact ensemble of bagged decision trees `B`. `Yfit` is a cell array of character vectors for classification and a numeric array for regression. By default, `predict` takes a democratic (nonweighted) average vote from all trees in the ensemble.

`B` is a trained `CompactTreeBagger` model object, that is, a model returned by `compact`.

`X` is a table or matrix of predictor data used to generate responses. Rows represent observations and columns represent variables.

- If `X` is a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `B`.
  - If you trained `B` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `TreeBagger`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- If `X` is a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `B` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and be of the same data types as those that trained `B` (stored in `B.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
  - If you trained `B` using a numeric matrix, then the predictor names in `B.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `TreeBagger`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

`Yfit = predict(B,X,Name,Value)` specifies additional options using one or more name-value pair arguments:

- `'Trees'` — Array of tree indices to use for computation of responses. The default is `'all'`.
- `'TreeWeights'` — Array of `NTrees` weights for weighting votes from the specified trees, where `NTrees` is the number of trees in the ensemble.
- `'UseInstanceForTree'` — Logical matrix of size `Nobs`-by-`NTrees` indicating which trees to use to make predictions for each observation, where `Nobs` is the number of observations. By default all trees are used for all observations.

For regression, `[Yfit,stdevs] = predict( ___ )` also returns standard deviations of the computed responses over the ensemble of the grown trees using any of the input argument combinations in previous syntaxes.

For classification, `[Yfit,scores] = predict( ___ )` also returns scores for all classes. `scores` is a matrix with one row per observation and one column per class. For each observation and each class, the score generated by each tree is the probability of the observation originating from the class, computed as the fraction of observations of the class in a tree leaf. `predict` averages these scores over all trees in the ensemble.

`[Yfit,scores,stdevs] = predict( ___ )` also returns standard deviations of the computed scores for classification. `stdevs` is a matrix with one row per observation and one column per class, with standard deviations taken over the ensemble of the grown trees.

Algorithms

- For regression problems, the predicted response for an observation is the weighted average of the predictions using selected trees only. That is,

$$\widehat{y}_{\text{bag}} = \frac{1}{\sum_{t=1}^{T} \alpha_t I(t \in S)} \sum_{t=1}^{T} \alpha_t \widehat{y}_t I(t \in S) .$$

  - $\widehat{y}_t$ is the prediction from tree t in the ensemble.
  - S is the set of indices of selected trees that comprise the prediction (see `'Trees'` and `'UseInstanceForTree'`). $I(t \in S)$ is 1 if t is in the set S, and 0 otherwise.
  - α_t is the weight of tree t (see `'TreeWeights'`).
- For classification problems, the predicted class for an observation is the class that yields the largest weighted average of the class posterior probabilities (i.e., classification scores) computed using selected trees only. That is,

  1. For each class $c \in C$ and each tree $t = 1,...,T$, `predict` computes $\widehat{P}_t(c|x)$, which is the estimated posterior probability of class c given observation x using tree t. C is the set of all distinct classes in the training data. For more details on classification tree posterior probabilities, see `fitctree` and `predict`.
  2. `predict` computes the weighted average of the class posterior probabilities over the selected trees.

$$\widehat{P}_{\text{bag}}(c|x) = \frac{1}{\sum_{t=1}^{T} \alpha_t I(t \in S)} \sum_{t=1}^{T} \alpha_t \widehat{P}_t(c|x) I(t \in S) .$$

3 The predicted class is the class that yields the largest weighted average.

$$\widehat{y}_{\text{bag}} = \underset{c \in C}{\operatorname{argmax}} \left\{ \widehat{P}_{\text{bag}}(c \mid x) \right\}.$$

See Also

`CompactTreeBagger` | `error`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

predict

Package:

Predict responses of generalized linear regression model

Syntax

```
ypred = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew,Name,Value)
```

Description

`ypred = predict(mdl,Xnew)` returns the predicted response values of the generalized linear regression model `mdl` to the points in `Xnew`.

`[ypred,yci] = predict(mdl,Xnew)` also returns confidence intervals for the responses at `Xnew`.

`[ypred,yci] = predict(mdl,Xnew,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the confidence level of the confidence interval.

Examples

Predict Response Values

Create a generalized linear regression model, and predict its response to new data.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

Create data points for prediction.

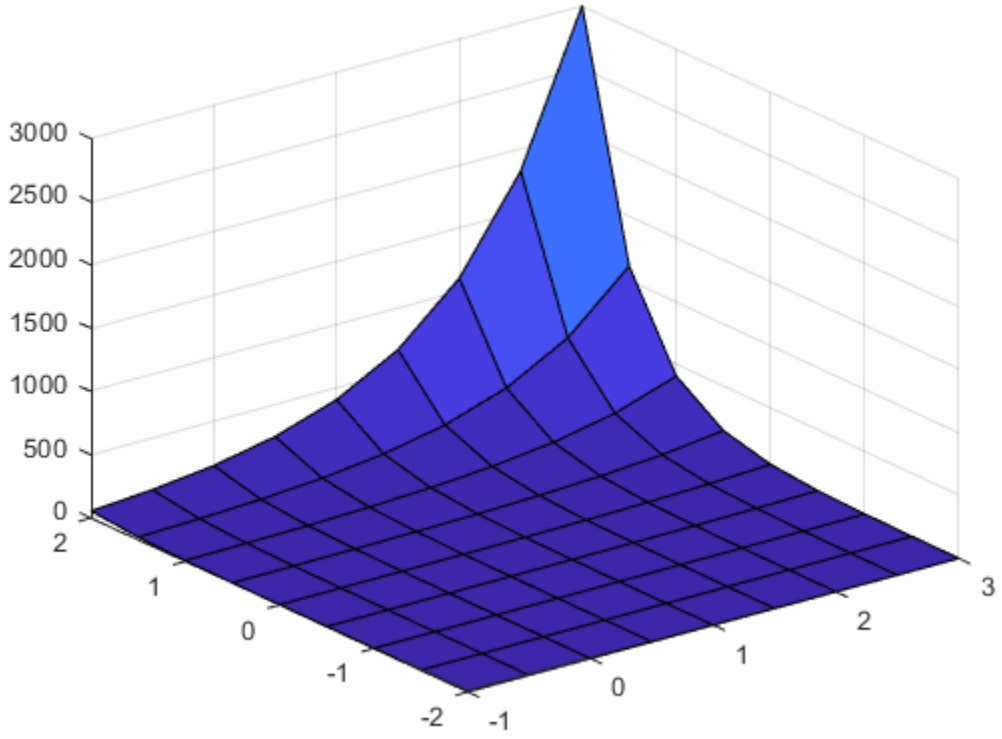
```
[Xtest1,Xtest2] = meshgrid(-1:.5:3,-2:.5:2);
Xnew = [Xtest1(:),Xtest2(:)];
```

Predict responses at the data points.

```
ypred = predict(mdl,Xnew);
```

Plot the predictions.

```
surf(Xtest1,Xtest2,reshape(ypred,9,9))
```



Generate C/C++ Code for Prediction

Fit a generalized linear regression model, and then save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the fitted model. Then use `codegen` (MATLAB Coder) to generate C/C++ code. Note that generating C/C++ code requires MATLAB® Coder™.

This example briefly explains the code generation workflow for the prediction of linear regression models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. For details, see “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22.

Train Model

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model. Specify the Poisson distribution for the response.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson');
```

Save Model

Save the fitted generalized linear regression model to the file `GLMMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(mdl,'GLMMdl');
```

Define Entry-Point Function

In your current folder, define an entry-point function named `mypredictGLM.m` that does the following:

- Accept new predictor input and valid name-value pair arguments.
- Load the fitted generalized linear regression model in `GLMMdl.mat` by using `loadLearnerForCoder`.
- Return predictions and confidence interval bounds.

```
function [yhat,ci] = mypredictGLM(x,varargin) %#codegen
%MYPREDICTGLM Predict responses using GLM model
%   MYPREDICTGLM predicts responses for the n observations in the n-by-1
%   vector x using the GLM model stored in the MAT-file GLMMdl.mat,
%   and then returns the predictions in the n-by-1 vector yhat.
%   MYPREDICTGLM also returns confidence interval bounds for the
%   predictions in the n-by-2 vector ci.
CompactMdl = loadLearnerForCoder('GLMMdl');
narginchk(1,Inf);
[yhat,ci] = predict(CompactMdl,x,varargin{:});
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and exact input array size, pass a MATLAB® expression that represents the set of values with a certain data type and array size. Use `coder.Constant` (MATLAB Coder) for the names of name-value pair arguments.

Create points for prediction.

```
[Xtest1,Xtest2] = meshgrid(-1:.5:3,-2:.5:2);
Xnew = [Xtest1(:),Xtest2(:)];
```

Generate code and specify returning 90% simultaneous confidence intervals on the predictions.

```
codegen mypredictGLM -args {Xnew,coder.Constant('Alpha'),0.1,coder.Constant('Simultaneous'),true
```

```
Code generation successful.
```

`codegen` generates the MEX function `mypredictGLM_mex` with a platform-dependent extension.

If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see "Specify Variable-Size Arguments for Code Generation" on page 32-45 and "Specify Properties of Entry-Point Function Inputs" (MATLAB Coder).

Verify Generated Code

Compare predictions and confidence intervals using `predict` and `mypredictGLM_mex`. Specify name-value pair arguments in the same order as in the `-args` argument in the call to `codegen`.

```
[yhat1,ci1] = predict(mdl,Xnew,'Alpha',0.1,'Simultaneous',true);
[yhat2,ci2] = mypredictGLM_mex(Xnew,'Alpha',0.1,'Simultaneous',true);
```

The returned values from `mypredictGLM_mex` might include round-off differences compared to the values from `predict`. In this case, compare the values allowing a small tolerance.

```
find(abs(yhat1-yhat2) > 1e-6)

ans =

  0x1 empty double column vector

find(abs(ci1-ci2) > 1e-6)

ans =

  0x1 empty double column vector
```

The comparison confirms that the returned values are equal within the tolerance `1e–6`.

Input Arguments

`mdl` — Generalized linear regression model
`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

`Xnew` — New predictor input values
table | dataset array | matrix

New predictor input values, specified as a table, dataset array, or matrix. Each row of `Xnew` corresponds to one observation, and each column corresponds to one variable.

- If `Xnew` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
- If `Xnew` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating `mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `[ypred,yci] = predict(Mdl,Xnew,'Alpha',0.01,'Simultaneous',true)` returns the confidence interval `yci` with a 99% confidence level, computed simultaneously for all predictor values.

`Alpha` — Significance level

0.05 (default) | numeric value in the range [0,1]

Significance level for the confidence interval, specified as the comma-separated pair consisting of `'Alpha'` and a numeric value in the range [0,1]. The confidence level of `yci` is equal to 100(1 – `Alpha`)%. `Alpha` is the probability that the confidence interval does not contain the true value.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`BinomialSize` — Number of trials for binomial distribution

1 (default) | scalar | vector

Number of trials for the binomial distribution, specified as the comma-separated pair consisting of `'BinomialSize'` and a scalar or vector of the same length as the response. `predict` expands the scalar input into a constant array of the same size as the response. The scalar input means that all observations have the same number of trials.

The meaning of the output values in `ypred` depends on the value of `'BinomialSize'`.

- If `'BinomialSize'` is 1 (default), then each value in the output `ypred` is the probability of success.
- If `'BinomialSize'` is not 1, then each value in the output `ypred` is the predicted number of successes in the trials.

Data Types: `single` | `double`

`Offset` — Offset value

`zeros(size(Xnew,1))` (default) | scalar | vector

Offset value for each row in `Xnew`, specified as the comma-separated pair consisting of `'Offset'` and a scalar or vector with the same length as the response. `predict` expands the scalar input into a constant array of the same size as the response.

Note that the default value of this argument is a vector of zeros even if you specify the `'Offset'` name-value pair argument when fitting a model. If you specify `'Offset'` for fitting, the software treats the offset as an additional predictor with a coefficient value fixed at 1. In other words, the formula for fitting is

$$f(\mu) = \text{Offset} + X*b,$$

where f is the link function, μ is the mean response, and $X*b$ is the linear combination of predictors X. The `Offset` predictor has coefficient `1`.

Data Types: `single` | `double`

`Simultaneous` — Flag to compute simultaneous confidence bounds
`false` (default) | `true`

Flag to compute simultaneous confidence bounds, specified as the comma-separated pair consisting of `'Simultaneous'` and either true or false.

- `true` — `predict` computes confidence bounds for the curve of response values corresponding to all predictor values in `Xnew`, using Scheffe's method. The range between the upper and lower bounds contains the curve consisting of true response values with 100(1 – α)% confidence.
- `false` — `predict` computes confidence bounds for the response value at each observation in `Xnew`. The confidence interval for a response value at a specific predictor value contains the true response value with 100(1 – α)% confidence.

Simultaneous bounds are wider than separate bounds, because requiring the entire curve of response values to be within the bounds is stricter than requiring the response value at a single predictor value to be within the bounds.

Example: `'Simultaneous',true`

Output Arguments

`ypred` — Predicted response values
numeric vector

Predicted response values at `Xnew`, returned as a numeric vector.

For a binomial model, the meaning of the output values in `ypred` depends on the value of the `'BinomialSize'` name-value pair argument.

- If `'BinomialSize'` is 1 (default), then each value in the output `ypred` is the probability of success.
- If `'BinomialSize'` is not 1, then each value in the output `ypred` is the predicted number of successes in the trials.

For a model with an offset, specify the offset value by using the `'Offset'` name-value pair argument. Otherwise, `predict` uses `0` as the offset value.

`yci` — Confidence intervals for responses
two-column numeric matrix

Confidence intervals for the responses, returned as a two-column matrix with each row providing one interval. The meaning of the confidence interval depends on the settings of the name-value pair arguments `'Alpha'` and `'Simultaneous'`.

Alternative Functionality

- `feval` returns the same predictions as `predict`. The `feval` function does not support the `'Offset'` and `'BinomialSize'` name-value pair arguments . `feval` uses 0 as the offset value, and the output values in `ypred` are predicted probabilities. The `feval` function can take multiple input arguments for new predictor input values, with one input for each predictor variable, which is simpler to use with a model created from a table or dataset array. Note that the `feval` function does not give confidence intervals on its predictions.

- `random` returns predictions with added noise.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-835 of the `CompactGeneralizedLinearModel` object.

`Xnew`	• `Xnew` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in `Xnew` can be a variable size, but the number of columns in `Xnew` must be fixed.
	• If you want to specify `Xnew` as a table, then your model must be trained using a table, and you must ensure that your entry-point function for prediction:
	  • Accepts data as arrays
	  • Creates a table from the data input arguments and specifies the variable names in the table
	  • Passes the table to `predict`
	For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined significance level in the generated code, include `{coder.Constant('Alpha'),0}` in the `-args` value of `codegen`.

For more information, see "Introduction to Code Generation" on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `feval` | `fitglm` | `random`

Topics

“Predict or Simulate Responses to New Data” on page 12-23
“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2012a

predict

Class: `GeneralizedLinearMixedModel`

Predict response of generalized linear mixed-effects model

Syntax

```
ypred = predict(glme)
ypred = predict(glme,tblnew)
ypred = predict(___,Name,Value)
[ypred,ypredCI] = predict(___)
[ypred,ypredCI,DF] = predict(___)
```

Description

`ypred = predict(glme)` returns the predicted conditional means of the response, `ypred`, using the original predictor values used to fit the generalized linear mixed-effects model `glme`.

`ypred = predict(glme,tblnew)` returns the predicted conditional means using the new predictor values specified in `tblnew`.

If a grouping variable in `tblnew` has levels that are not in the original data, then the random effects for that grouping variable do not contribute to the `'Conditional'` prediction at observations where the grouping variable has new levels.

`ypred = predict(___,Name,Value)` returns the predicted conditional means of the response using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence level, simultaneous confidence bounds, or contributions from only fixed effects. You can use any of the input arguments in the previous syntaxes.

`[ypred,ypredCI] = predict(___)` also returns 95% point-wise confidence intervals, `ypredCI`, for each predicted value.

`[ypred,ypredCI,DF] = predict(___)` also returns the degrees of freedom, `DF`, used to compute the confidence intervals.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`tblnew` — New input data

table | `dataset` array

New input data, which includes the response variable, predictor variables, and grouping variables on page 2-45, specified as a table or dataset array. The predictor variables can be continuous or

grouping variables. `tblnew` must have the same variables as the original table or dataset array used in `fitglme` to fit the generalized linear mixed-effects model `glme`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range [0,1]

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range [0,1]. For a value α, the confidence level is 100 × (1 – α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Conditional` — Indicator for conditional predictions

`true` (default) | `false`

Indicator for conditional predictions, specified as the comma-separated pair consisting of `'Conditional'` and one of the following.

Value	Description
`true`	Contributions from both fixed effects and random effects (conditional)
`false`	Contribution from only fixed effects (marginal)

Example: `'Conditional',false`

`DFMethod` — Method for computing approximate degrees of freedom

`'residual'` (default) | `'none'`

Method for computing approximate degrees of freedom, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom value is assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	The degrees of freedom is set to infinity.

Example: `'DFMethod','none'`

`Offset` — Model offset

`zeros(m,1)` (default) | *m*-by-1 vector of scalar values

Model offset, specified as a vector of scalar values of length m, where m is the number of rows in `tblnew`. The offset is used as an additional predictor and has a coefficient value fixed at `1`.

`Simultaneous` — Type of confidence bounds
`false` (default) | `true`

Type of confidence bounds, specified as the comma-separated pair consisting of `'Simultaneous'` and either `false` or `true`.

- If `'Simultaneous'` is `false`, then `predict` computes nonsimultaneous confidence bounds.
- If `'Simultaneous'` is `true`, `predict` returns simultaneous confidence bounds.

Example: `'Simultaneous',true`

Output Arguments

`ypred` — Predicted responses
vector

Predicted responses, returned as a vector. If the `'Conditional'` name-value pair argument is specified as `true`, `ypred` contains predictions for the conditional means of the responses given the random effects. Conditional predictions include contributions from both fixed and random effects. Marginal predictions include only contributions from fixed effects.

To compute marginal predictions, `predict` computes conditional predictions, but substitutes a vector of zeros in place of the empirical Bayes predictors (EBPs) of the random effects.

`ypredCI` — Point-wise confidence intervals
two-column matrix

Point-wise confidence intervals for the predicted values, returned as a two-column matrix. The first column of `ypredCI` contains the lower bound, and the second column contains the upper bound. By default, `ypredCI` contains the 95% nonsimultaneous confidence intervals for the predictions. You can change the confidence level using the `Alpha` name-value pair argument, and make them simultaneous using the `Simultaneous` name-value pair argument.

When fitting a GLME model using `fitglme` and one of the maximum likelihood fit methods (`'Laplace'` or `'ApproximateLaplace'`), `predict` computes the confidence intervals using the conditional mean squared error of prediction (CMSEP) approach conditional on the estimated covariance parameters and the observed response. Alternatively, you can interpret the confidence intervals as approximate Bayesian credible intervals conditional on the estimated covariance parameters and the observed response.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `predict` bases the computations on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

`DF` — Degrees of freedom
vector | scalar value

Degrees of freedom used in computing the confidence intervals, returned as a vector or a scalar value.

- If `'Simultaneous'` is `false`, then `DF` is a vector.
- If `'Simultaneous'` is `true`, then `DF` is a scalar value.

Examples

Predict Responses at Original Design Values

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution:

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.

- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Predict the response values at the original design values. Display the first ten predictions along with the observed response values.

```
ypred = predict(glme);
[ypred(1:10),mfr.defects(1:10)]
```

```
ans = 10×2

    4.9883    6.0000
    5.9423    7.0000
    5.1318    6.0000
    5.6295    5.0000
    5.3499    6.0000
    5.2134    5.0000
    4.6430    4.0000
    4.5342    4.0000
    5.3903    9.0000
    4.6529    4.0000
```

Column 1 contains the predicted response values at the original design values. Column 2 contains the observed response values.

Predict Responses at Values in New Table

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution:

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Predict the response values at the original design values.

```
ypred = predict(glme);
```

Create a new table by copying the first 10 rows of `mfr` into `tblnew`.

```
tblnew = mfr(1:10,:);
```

The first 10 rows of `mfr` include data collected from trials 1 through 5 for factories 1 and 2. Both factories used the old process for all of their trials during the experiment, so `newprocess = 0` for all 10 observations.

Change the value of `newprocess` to `1` for the observations in `tblnew`.

```
tblnew.newprocess = ones(height(tblnew),1);
```

Compute predicted response values and nonsimultaneous 99% confidence intervals using `tblnew`. Display the first 10 rows of the predicted values based on `tblnew`, the predicted values based on `mfr`, and the observed response values.

```
[ypred_new,ypredCI] = predict(glme,tblnew,'Alpha',0.01);
[ypred_new,ypred(1:10),mfr.defects(1:10)]
```

```
ans = 10×3

    3.4536    4.9883    6.0000
    4.1142    5.9423    7.0000
    3.5530    5.1318    6.0000
    3.8976    5.6295    5.0000
    3.7040    5.3499    6.0000
    3.6095    5.2134    5.0000
    3.2146    4.6430    4.0000
    3.1393    4.5342    4.0000
    3.7320    5.3903    9.0000
    3.2214    4.6529    4.0000
```

Column 1 contains predicted response values based on the data in `tblnew`, where `newprocess = 1`. Column 2 contains predicted response values based on the original data in `mfr`, where `newprocess = 0`. Column 3 contains the observed response values in `mfr`. Based on these results, if all other predictors retain their original values, the predicted number of defects appears to be smaller when using the new process.

Display the 99% confidence intervals for rows 1 through 10 corresponding to the new predicted response values.

```
ypredCI(1:10,1:2)
```

```
ans = 10×2

    1.6983    7.0235
    1.9191    8.8201
    1.8735    6.7380
    2.0149    7.5395
    1.9034    7.2079
    1.8918    6.8871
    1.6776    6.1597
    1.5404    6.3976
    1.9574    7.1154
    1.6892    6.1436
```

References

[1] Booth, J.G., and J.P. Hobert. “Standard Errors of Prediction in Generalized Linear Mixed Models.” *Journal of the American Statistical Association*, Vol. 93, 1998, pp. 262–272.

See Also

`GeneralizedLinearMixedModel` | `fitglme` | `fitted` | `random`

predict

Package:

Predict responses of linear regression model

Syntax

```
ypred = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew,Name,Value)
```

Description

`ypred = predict(mdl,Xnew)` returns the predicted response values of the linear regression model `mdl` to the points in `Xnew`.

`[ypred,yci] = predict(mdl,Xnew)` also returns confidence intervals for the responses at `Xnew`.

`[ypred,yci] = predict(mdl,Xnew,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the confidence level of the confidence interval and the prediction type.

Examples

Predict Response Values

Create a quadratic model of car mileage as a function of weight from the `carsmall` data set.

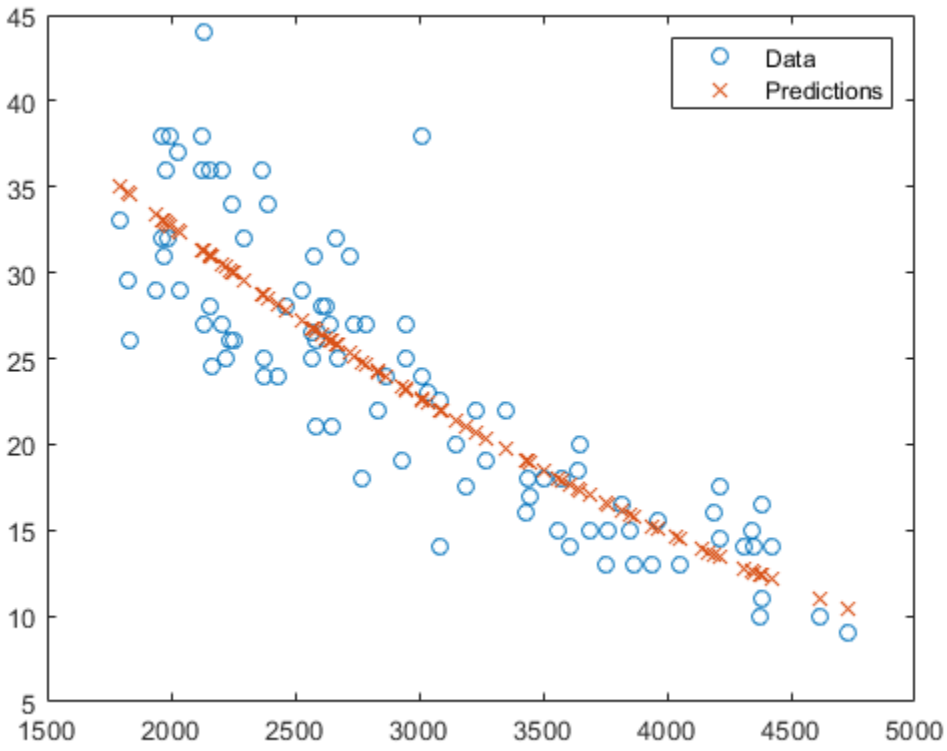
```
load carsmall
X = Weight;
y = MPG;
mdl = fitlm(X,y,'quadratic');
```

Create predicted responses to the data.

```
ypred = predict(mdl,X);
```

Plot the original responses and the predicted responses to see how they differ.

```
plot(X,y,'o',X,ypred,'x')
legend('Data','Predictions')
```



Generate C/C++ Code for Prediction

Fit a linear regression model, and then save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the fitted model. Then use `codegen` (MATLAB Coder) to generate C/C++ code. Note that generating C/C++ code requires MATLAB® Coder™.

This example briefly explains the code generation workflow for the prediction of linear regression models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. For details, see “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22.

Train Model

Load the `carsmall` data set, and then fit the quadratic regression model.

```
load carsmall
X = Weight;
y = MPG;
mdl = fitlm(X,y,'quadratic');
```

Save Model

Save the fitted quadratic model to the file `QLMMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(mdl,'QLMMdl');
```

Define Entry-Point Function

Define an entry-point function named `mypredictQLM` that does the following:

- Accept measurements corresponding to X and optional, valid name-value pair arguments.
- Load the fitted quadratic model in `QLMMdl.mat`.
- Return predictions and confidence interval bounds.

```
function [yhat,ci] = mypredictQLM(x,varargin) %#codegen
%MYPREDICTQLM Predict response using linear model
%   MYPREDICTQLM predicts responses for the n observations in the n-by-1
%   vector x using the linear model stored in the MAT-file QLMMdl.mat, and
%   then returns the predictions in the n-by-1 vector yhat. MYPREDICTQLM
%   also returns confidence interval bounds for the predictions in the
%   n-by-2 vector ci.
CompactMdl = loadLearnerForCoder('QLMMdl');
[yhat,ci] = predict(CompactMdl,x,varargin{:});
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and exact input array size, pass a MATLAB® expression that represents the set of values with a certain data type and array size. Use `coder.Constant` (MATLAB Coder) for the names of name-value pair arguments.

If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

```
codegen mypredictQLM -args {X,coder.Constant('Alpha'),0.1,coder.Constant('Simultaneous'),true}
```

```
Code generation successful.
```

`codegen` generates the MEX function `mypredictQLM_mex` with a platform-dependent extension.

Verify Generated Code

Compare predictions and confidence intervals using `predict` and `mypredictQLM_mex`. Specify name-value pair arguments in the same order as in the `-args` argument in the call to `codegen`.

```
Xnew = sort(X);
[yhat1,ci1] = predict(mdl,Xnew,'Alpha',0.1,'Simultaneous',true);
[yhat2,ci2] = mypredictQLM_mex(Xnew,'Alpha',0.1,'Simultaneous',true);
```

The returned values from `mypredictQLM_mex` might include round-off differences compared to the values from `predict`. In this case, compare the values allowing a small tolerance.

```
find(abs(yhat1-yhat2) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

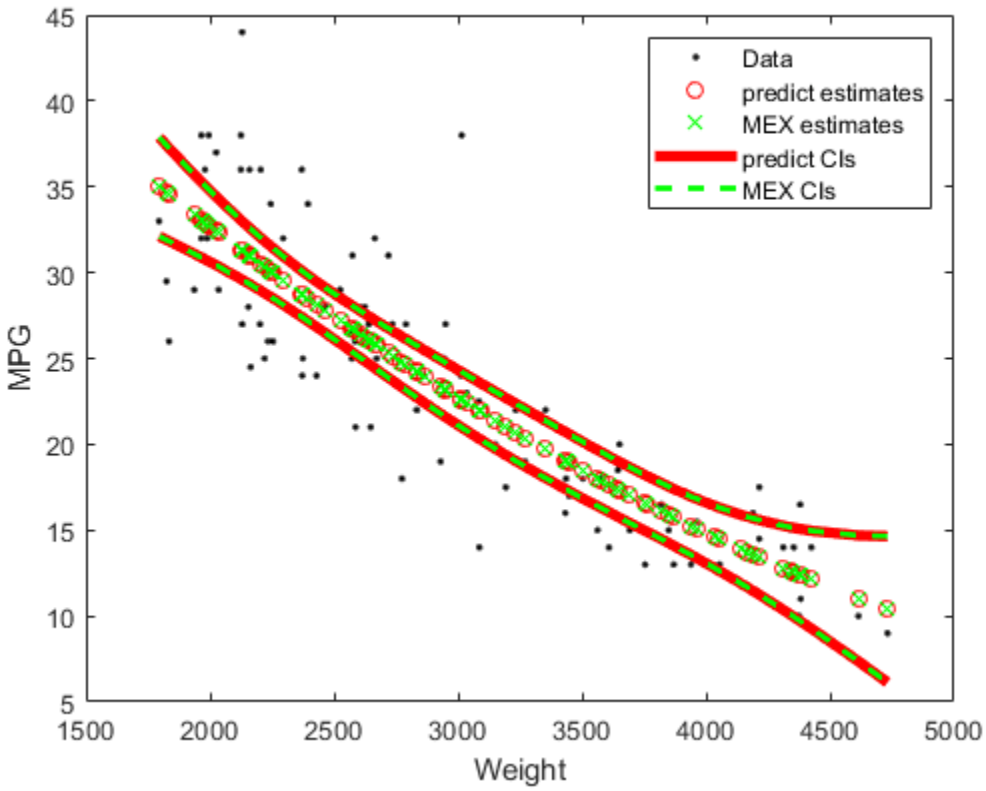
```
find(abs(ci1-ci2) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that the returned values are equal within the tolerance `1e–6`.

Plot the returned values for comparison.

```
h1 = plot(X,y,'k.');
hold on
h2 = plot(Xnew,yhat1,'ro',Xnew,yhat2,'gx');
h3 = plot(Xnew,ci1,'r-','LineWidth',4);
h4 = plot(Xnew,ci2,'g--','LineWidth',2);
legend([h1; h2; h3(1); h4(1)], ...
    {'Data','predict estimates','MEX estimates','predict CIs','MEX CIs'});
xlabel('Weight');
ylabel('MPG');
```



Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`Xnew` — New predictor input values

table | dataset array | matrix

New predictor input values, specified as a table, dataset array, or matrix. Each row of `Xnew` corresponds to one observation, and each column corresponds to one variable.

- If `Xnew` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
- If `Xnew` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating `mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `[ypred,yci] = predict(Mdl,Xnew,'Alpha',0.01,'Simultaneous',true)` returns the confidence interval `yci` with a 99% confidence level, computed simultaneously for all predictor values.

`Alpha` — Significance level

0.05 (default) | numeric value in the range [0,1]

Significance level for the confidence interval, specified as the comma-separated pair consisting of `'Alpha'` and a numeric value in the range [0,1]. The confidence level of `yci` is equal to 100(1 – `Alpha`)%. `Alpha` is the probability that the confidence interval does not contain the true value.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Prediction` — Prediction type

`'curve'` (default) | `'observation'`

Prediction type, specified as the comma-separated pair consisting of `'Prediction'` and either `'curve'` or `'observation'`.

A regression model for the predictor variables X and the response variable y has the form

$$y = f(X) + \varepsilon,$$

where f is a fitted regression function and ε is a random noise term.

- If `'Prediction'` is `'curve'`, then `predict` predicts confidence bounds for $f(Xnew)$, the fitted responses at `Xnew`.
- If `'Prediction'` is `'observation'`, then `predict` predicts confidence bounds for y, the response observations at `Xnew`.

The bounds for y are wider than the bounds for $f(X)$ because of the additional variability of the noise term.

Example: `'Prediction','observation'`

`Simultaneous` — Flag to compute simultaneous confidence bounds

`false` (default) | `true`

Flag to compute simultaneous confidence bounds, specified as the comma-separated pair consisting of `'Simultaneous'` and either true or false.

- `true` — `predict` computes confidence bounds for the curve of response values corresponding to all predictor values in `Xnew`, using Scheffe's method. The range between the upper and lower bounds contains the curve consisting of true response values with 100(1 – α)% confidence.
- `false` — `predict` computes confidence bounds for the response value at each observation in `Xnew`. The confidence interval for a response value at a specific predictor value contains the true response value with 100(1 – α)% confidence.

Simultaneous bounds are wider than separate bounds, because requiring the entire curve of response values to be within the bounds is stricter than requiring the response value at a single predictor value to be within the bounds.

Example: `'Simultaneous',true`

Output Arguments

ypred — Predicted response values

numeric vector

Predicted response values evaluated at `Xnew`, returned as a numeric vector.

yci — Confidence intervals for responses

two-column numeric matrix

Confidence intervals for the responses, returned as a two-column matrix with each row providing one interval. The meaning of the confidence interval depends on the settings of the name-value pair arguments `'Alpha'`, `'Prediction'`, and `'Simultaneous'`.

Alternative Functionality

- `feval` returns the same predictions as `predict`. The `feval` function can take multiple input arguments, with one input for each predictor variable, which is simpler to use with a model created from a table or dataset array. Note that the `feval` function does not give confidence intervals on its predictions.
- `random` returns predictions with added noise.
- Use `plotSlice` to create a figure containing a series of plots, each representing a slice through the predicted regression surface. Each plot shows the fitted response values as a function of a single predictor variable, with the other predictor variables held constant.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `predict` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `predict` function. Then use `codegen` to generate code for the entry-point function.
- This table contains notes about the arguments of `predict`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`mdl`	• Suppose you train a linear model by using `fitlm` and specifying `'RobustOpts'` as a structure with an anonymous function handle for the `RobustWgtFun` field, use `saveLearnerForCoder` to save the model, and then use `loadLearnerForCoder` to load the model. In this case, `loadLearnerForCoder` cannot restore the Robust property into the MATLAB Workspace. However, `loadLearnerForCoder` can load the model at compile time within an entry-point function for code generation. • For the usage notes and limitations of the model object, see "Code Generation" on page 33-824 of the `CompactLinearModel` object.

`Xnew`	• `Xnew` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in `Xnew` can be a variable size, but the number of columns in `Xnew` must be fixed.
	• If you want to specify `Xnew` as a table, then your model must be trained using a table, and you must ensure that your entry-point function for prediction:
	  • Accepts data as arrays
	  • Creates a table from the data input arguments and specifies the variable names in the table
	  • Passes the table to `predict`
	  For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined significance level in the generated code, include `{coder.Constant('Alpha'),0}` in the `-args` value of `codegen`.

For more information, see "Introduction to Code Generation" on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactLinearModel` | `LinearModel` | `feval` | `plotSlice` | `random`

Topics

"Predict or Simulate Responses to New Data" on page 11-31
"Linear Regression Workflow" on page 11-35
"Interpret Linear Regression Results" on page 11-50
"Linear Regression" on page 11-9

Introduced in R2012a

predict

Class: `LinearMixedModel`

Predict response of linear mixed-effects model

Syntax

```
ypred = predict(lme)
ypred = predict(lme,tblnew)
ypred = predict(lme,Xnew,Znew)
ypred = predict(lme,Xnew,Znew,Gnew)
ypred = predict( ___ ,Name,Value)
[ypred,ypredCI] = predict( ___ )
[ypred,ypredCI,DF] = predict( ___ )
```

Description

`ypred = predict(lme)` returns a vector of conditional predicted responses on page 33-4967 `ypred` at the original predictors used to fit the linear mixed-effects model `lme`.

`ypred = predict(lme,tblnew)` returns a vector of conditional predicted responses `ypred` from the fitted linear mixed-effects model `lme` at the values in the new table or dataset array `tblnew`. Use a table or dataset array for `predict` if you use a table or dataset array for fitting the model `lme`.

If a particular grouping variable in `tblnew` has levels that are not in the original data, then the random effects for that grouping variable do not contribute to the `'Conditional'` prediction at observations where the grouping variable has new levels.

`ypred = predict(lme,Xnew,Znew)` returns a vector of conditional predicted responses `ypred` from the fitted linear mixed-effects model `lme` at the values in the new fixed- and random-effects design matrices, `Xnew` and `Znew`, respectively. `Znew` can also be a cell array of matrices. In this case, the grouping variable `G` is `ones(n,1)`, where *n* is the number of observations used in the fit.

Use the matrix format for `predict` if using design matrices for fitting the model `lme`.

`ypred = predict(lme,Xnew,Znew,Gnew)` returns a vector of conditional predicted responses `ypred` from the fitted linear mixed-effects model `lme` at the values in the new fixed- and random-effects design matrices, `Xnew` and `Znew`, respectively, and the grouping variable `Gnew`.

`Znew` and `Gnew` can also be cell arrays of matrices and grouping variables, respectively.

`ypred = predict( ___ ,Name,Value)` returns a vector of predicted responses `ypred` from the fitted linear mixed-effects model `lme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the confidence level, simultaneous confidence bounds, or contributions from only fixed effects.

`[ypred,ypredCI] = predict( ___ )` also returns confidence intervals `ypredCI` for the predictions `ypred` for any of the input arguments in the previous syntaxes.

`[ypred,ypredCI,DF] = predict( ___ )` also returns the degrees of freedom `DF` used in computing the confidence intervals for any of the input arguments in the previous syntaxes.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`tblnew` — New input data

table | `dataset` array

New input data, which includes the response variable, predictor variables, and grouping variables on page 2-45, specified as a table or dataset array. The predictor variables can be continuous or grouping variables. `tblnew` must have the same variables as in the original table or dataset array used to fit the linear mixed-effects model `lme`.

`Xnew` — New fixed-effects design matrix

n-by-*p* matrix

New fixed-effects design matrix, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of fixed predictor variables. Each row of `X` corresponds to one observation and each column of `X` corresponds to one variable.

Data Types: `single` | `double`

`Znew` — New random-effects design

n-by-*q* matrix | cell array of length *R*

New random-effects design, specified as an *n*-by-*q* matrix or a cell array of *R* design matrices `Z{r}`, where $r = 1, 2, \ldots, R$. If `Znew` is a cell array, then each `Z{r}` is an *n*-by-*q*(*r*) matrix, where *n* is the number of observations, and *q*(*r*) is the number of random predictor variables.

Data Types: `single` | `double` | `cell`

`Gnew` — New grouping variable or variables

vector | cell array of grouping variables of length *R*

New grouping variable or variables on page 2-45, specified as a vector or a cell array, of length *R*, of grouping variables with the same levels or groups as the original grouping variables used to fit the linear mixed-effects model `lme`.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Conditional` — Indicator for conditional predictions
`true` (default) | `false`

Indicator for conditional prediction on page 33-4967s, specified as the comma-separated pair consisting of `'Conditional'` and one of the following.

`true`	Contributions from both fixed effects and random effects (conditional)
`false`	Contribution from only fixed effects (marginal)

Example: `'Conditional,false`

`DFMethod` — Method for computing approximate degrees of freedom
`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing approximate degrees of freedom to use in the confidence interval computation, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

`Simultaneous` — Type of confidence bounds
`false` (default) | `true`

Type of confidence bounds, specified as the comma-separated pair consisting of `'Simultaneous'` and one of the following.

`false`	Default. Nonsimultaneous bounds.
`true`	Simultaneous bounds.

Example: `'Simultaneous',true`

`Prediction` — Type of prediction
`'curve'` (default) | `'observation'`

Type of prediction, specified as the comma-separated pair consisting of `'Prediction'` and one of the following.

`'curve'`	Default. Confidence bounds for the predictions based on the fitted function.
`'observation'`	Variability due to observation error for the new observations is also included in the confidence bound calculations and this results in wider bounds.

Example: `'Prediction','observation'`

Output Arguments

`ypred` — Predicted responses

vector

Predicted responses, returned as a vector. `ypred` can contain the conditional or marginal responses, depending on the value choice of the `'Conditional'` name-value pair argument. Conditional predictions include contributions from both fixed and random effects.

`ypredCI` — Point-wise confidence intervals

two-column matrix

Point-wise confidence intervals for the predicted values, returned as a two-column matrix. The first column of `yCI` contains the lower bounds, and the second column contains the upper bound. By default, `yCI` contains the 95% confidence intervals for the predictions. You can change the confidence level using the `Alpha` name-value pair argument, make them simultaneous using the `Simultaneous` name-value pair argument, and also make them for a new observation rather than for the curve using the `Prediction` name-value pair argument.

`DF` — Degrees of freedom

vector | scalar value

Degrees of freedom used in computing the confidence intervals, returned as a vector or a scalar value.

- If the `'Simultaneous'` name-value pair argument is `false`, then `DF` is a vector.
- If the `'Simultaneous'` name-value pair argument is `true`, then `DF` is a scalar value.

Examples

Predict Responses at the Original Design Values

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Predict the response values at the original design values. Display the first five predictions with the observed response values.

```
yhat = predict(lme);
[yhat(1:5) ds.Yield(1:5)]
```

```
ans = 5×2

  115.4788  104.0000
  135.1455  136.0000
  152.8121  158.0000
  160.4788  174.0000
   58.0839   57.0000
```

Plot Predictions vs. Observed Responses

Load the sample data.

```
load carsmall
```

Fit a linear mixed-effects model, with a fixed effect for `Weight`, and a random intercept grouped by `Model_Year`. First, store the data in a table.

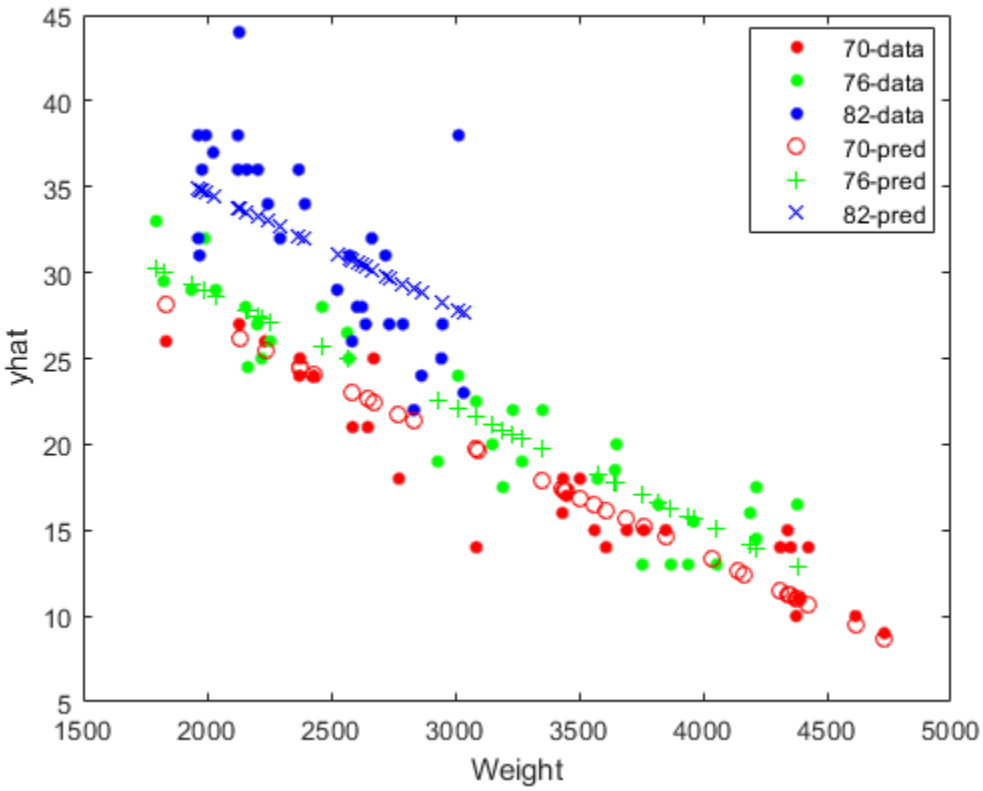
```
tbl = table(MPG,Weight,Model_Year);
lme = fitlme(tbl,'MPG ~ Weight + (1|Model_Year)');
```

Create predicted responses to the data.

```
yhat = predict(lme,tbl);
```

Plot the original responses and the predicted responses to see how they differ. Group them by model year.

```
figure()
gscatter(Weight,MPG,Model_Year)
hold on
gscatter(Weight,yhat,Model_Year,[],'o+x')
legend('70-data','76-data','82-data','70-pred','76-pred','82-pred')
hold off
```



Predict Responses at Values in a New Dataset Array

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Create a new dataset array with design values. The new dataset array must have the same variables as the original dataset array you use for fitting the model `lme`.

```
dsnew = dataset();
dsnew.Soil = nominal({'Sandy';'Silty'});
dsnew.Tomato = nominal({'Cherry';'Vine'});
dsnew. Fertilizer = nominal([2;2]);
```

Predict the conditional and marginal responses at the original design points.

```
yhatC = predict(lme,dsnew);
yhatM = predict(lme,dsnew,'Conditional',false);
[yhatC yhatM]
```

```
ans = 2×2

   92.7505  111.6667
   87.5891   82.6667
```

Predict Responses at the Values in New Design Matrices

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and cylinders, and potentially correlated random effects for intercept and acceleration grouped by model year.

First, prepare the design matrices for fitting the linear mixed-effects model.

```
X = [ones(406,1) Acceleration Horsepower];
Z = [ones(406,1) Acceleration];
Model_Year = nominal(Model_Year);
G = Model_Year;
```

Now, fit the model using `fitlmematrix` with the defined design matrices and grouping variables.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'}},'RandomEffectGroups',{'Model_Year'});
```

Create the design matrices that contain the data at which to predict the response values. `Xnew` must have three columns as in `X`. The first column must be a column of 1s. And the values in the last two columns must correspond to `Acceleration` and `Horsepower`, respectively. The first column of `Znew` must be a column of 1s, and the second column must contain the same `Acceleration` values as in `Xnew`. The original grouping variable in `G` is the model year. So, `Gnew` must contain values for the model year. Note that `Gnew` must contain nominal values.

```
Xnew = [1,13.5,185; 1,17,205; 1,21.2,193];
Znew = [1,13.5; 1,17; 1,21.2]; % alternatively Znew = Xnew(:,1:2);
Gnew = nominal([73 77 82]);
```

Predict the responses for the data in the new design matrices.

```
yhat = predict(lme,Xnew,Znew,Gnew)
```

```
yhat = 3×1

    8.7063
    5.4423
   12.5384
```

Now, repeat the same for a linear mixed-effects model with uncorrelated random-effects terms for intercept and acceleration. First, change the original random effects design and the random effects grouping variables. Then, refit the model.

```
Z = {ones(406,1),Acceleration};
G = {Model_Year,Model_Year};

lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept'},{'Acceleration'}},'RandomEffectGroups',{'Model_Year','Model_Year'});
```

Now, recreate the new random effects design, `Znew`, and the grouping variable design, `Gnew`, using which to predict the response values.

```
Znew = {[1;1;1],[13.5;17;21.2]};
MY = nominal([73 77 82]);
Gnew = {MY,MY};
```

Predict the responses using the new design matrices.

```
yhat = predict(lme,Xnew,Znew,Gnew)
```

```
yhat = 3×1

    8.6365
    5.9199
   12.1247
```

Compute Confidence Intervals for Predictions

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and cylinders, and potentially correlated random effects for intercept and acceleration grouped by model year. First, store the variables in a table.

```
tbl = table(MPG,Acceleration,Horsepower,Model_Year);
```

Now, fit the model using `fitlme` with the defined design matrices and grouping variables.

```
lme = fitlme(tbl,'MPG ~ Acceleration + Horsepower + (Acceleration|Model_Year)');
```

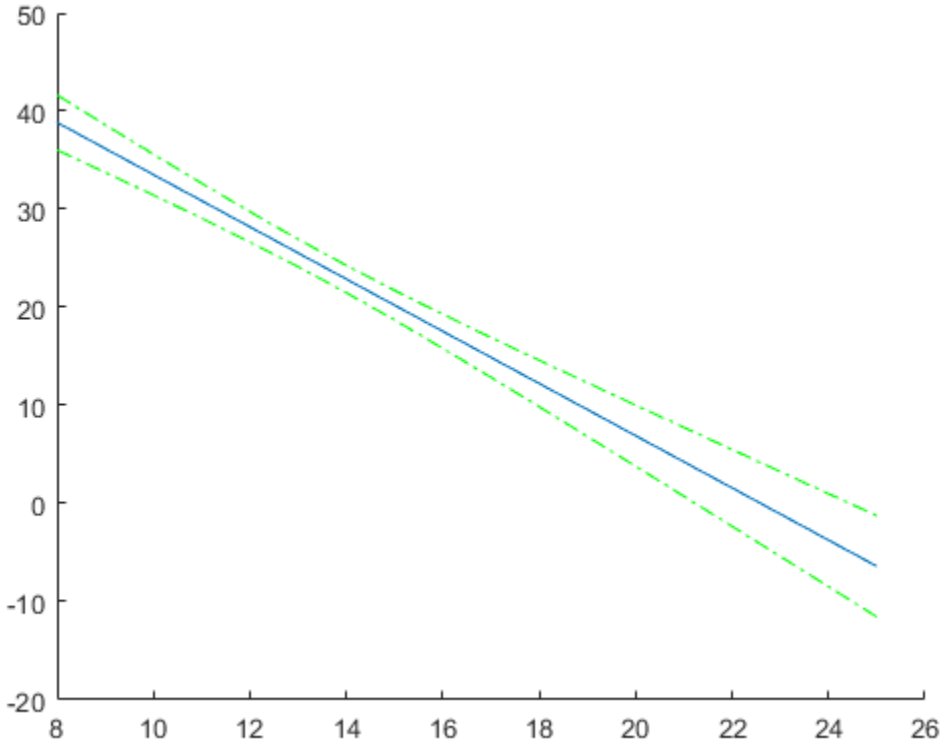
Create the new data and store it in a new table.

```
tblnew = table();
tblnew.Acceleration = linspace(8,25)';
tblnew.Horsepower = linspace(min(Horsepower),max(Horsepower))';
tblnew.Model_Year = repmat(70,100,1);
```

`linspace` creates 100 equally distanced values between the lower and the upper input limits. `Model_Year` is fixed at 70. You can repeat this for any model year.

Compute and plot the predicted values and 95% confidence limits (nonsimultaneous).

```
[ypred,yCI,DF] = predict(lme,tblnew);
figure();
h1 = line(tblnew.Acceleration,ypred);
hold on;
h2 = plot(tblnew.Acceleration,yCI,'g-.');
```



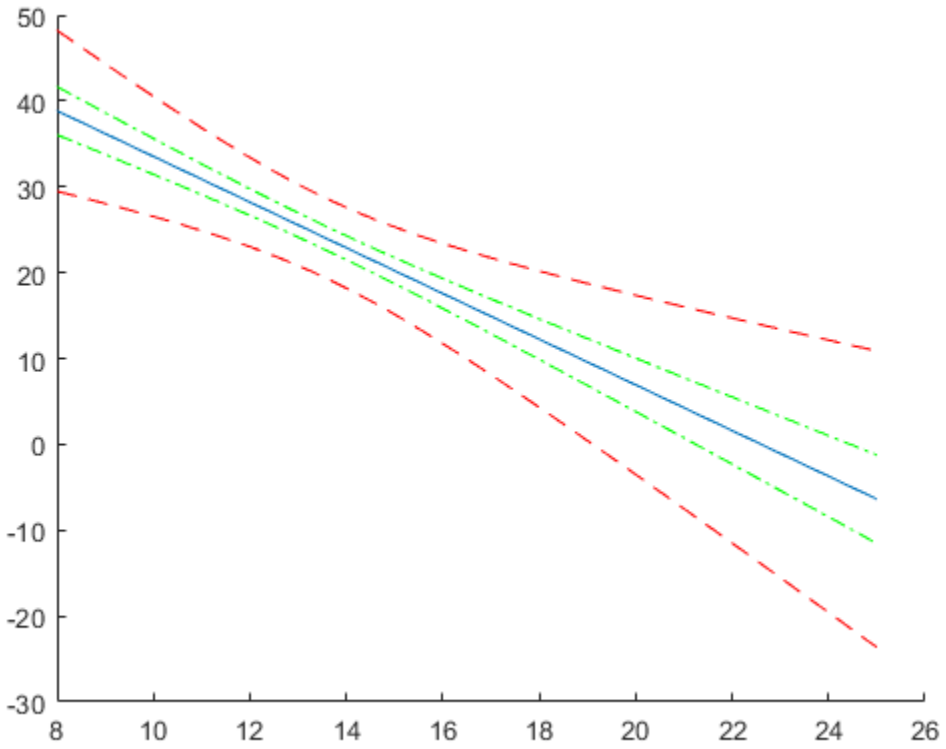
Display the degrees of freedom.

```
DF(1)
```

```
ans = 389
```

Compute and plot the simultaneous confidence bounds.

```
[ypred,yCI,DF] = predict(lme,tblnew,'Simultaneous',true);
h3 = plot(tblnew.Acceleration,yCI,'r--');
```



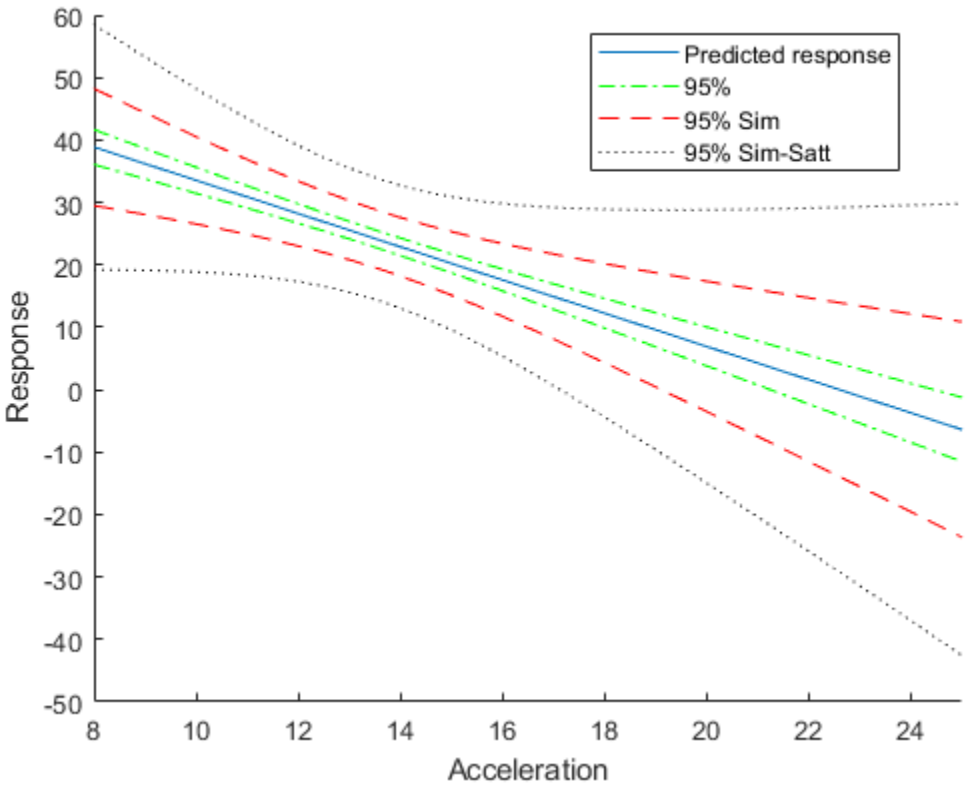
Display the degrees of freedom.

```
DF
```

```
DF = 389
```

Compute the simultaneous confidence bounds using the Satterthwaite method to compute the degrees of freedom.

```
[ypred,yCI,DF] = predict(lme,tblnew,'Simultaneous',true,'DFMethod','satterthwaite');
h4 = plot(tblnew.Acceleration,yCI,'k:');
hold off
xlabel('Acceleration')
ylabel('Response')
ylim([-50,60])
xlim([8,25])
legend([h1,h2(1),h3(1),h4(1)],'Predicted response','95%','95% Sim',...
'95% Sim-Satt','Location','Best')
```



Display the degrees of freedom.

```
DF
```

```
DF = 3.6001
```

More About

Conditional and Marginal Predictions

A conditional prediction includes contributions from both fixed and random effects, whereas a marginal model includes contribution from only fixed effects.

Suppose the linear mixed-effects model `lme` has an n-by-p fixed-effects design matrix `X` and an n-by-q random-effects design matrix `Z`. Also, suppose the estimated p-by-1 fixed-effects vector is $\widehat{\beta}$, and the q-by-1 estimated best linear unbiased predictor (BLUP) vector of random effects is $\widehat{b}$. The predicted conditional response is

$$\widehat{y}_{Cond} = X\widehat{\beta} + Z\widehat{b},$$

which corresponds to the `'Conditional','true'` name-value pair argument.

The predicted marginal response is

$$\widehat{y}_{Mar} = X\widehat{\beta},$$

which corresponds to the `'Conditional','false'` name-value pair argument.

When making predictions, if a particular grouping variable has new levels (1s that were not in the original data), then the random effects for the grouping variable do not contribute to the `'Conditional'` prediction at observations where the grouping variable has new levels.

See Also

`LinearMixedModel` | `fitted` | `random`

predict

Class: `FeatureSelectionNCAClassification`

Predict responses using neighborhood component analysis (NCA) classifier

Syntax

```
[labels,postprobs,classnames] = predict(mdl,X)
```

Description

`[labels,postprobs,classnames] = predict(mdl,X)` computes the predicted labels, `labels`, corresponding to the rows of `X`, using the model `mdl`.

Input Arguments

`mdl` — Neighborhood component analysis model for classification

`FeatureSelectionNCAClassification` object

Neighborhood component analysis model for classification, specified as a `FeatureSelectionNCAClassification` object.

`X` — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

Output Arguments

`labels` — Predicted class labels

categorical vector | logical vector | numeric vector | cell array of character vectors | character array

Predicted class labels corresponding to the rows of `X`, returned as a categorical, logical, or numeric vector, a cell array of character vectors of length *n*, or a character array with *n* rows. *n* is the number of observations. The type of `labels` is the same as `Y` used in training.

`postprobs` — Posterior probabilities

n-by-*c* matrix

Posterior probabilities, returned as an *n*-by-*c* matrix, where *n* is the number of observations and *c* is the number of classes. A posterior probability, `postprobs(i,:)`, represents the membership of an observation in `X(i,:)` in classes 1 through *c*.

`classnames` — Class names

cell array of character vectors

Class names corresponding to posterior probabilities, returned as a cell array of character vectors. Each character vector is the class name corresponding to a column of `postprobs`.

Examples

Tune NCA Model for Classification

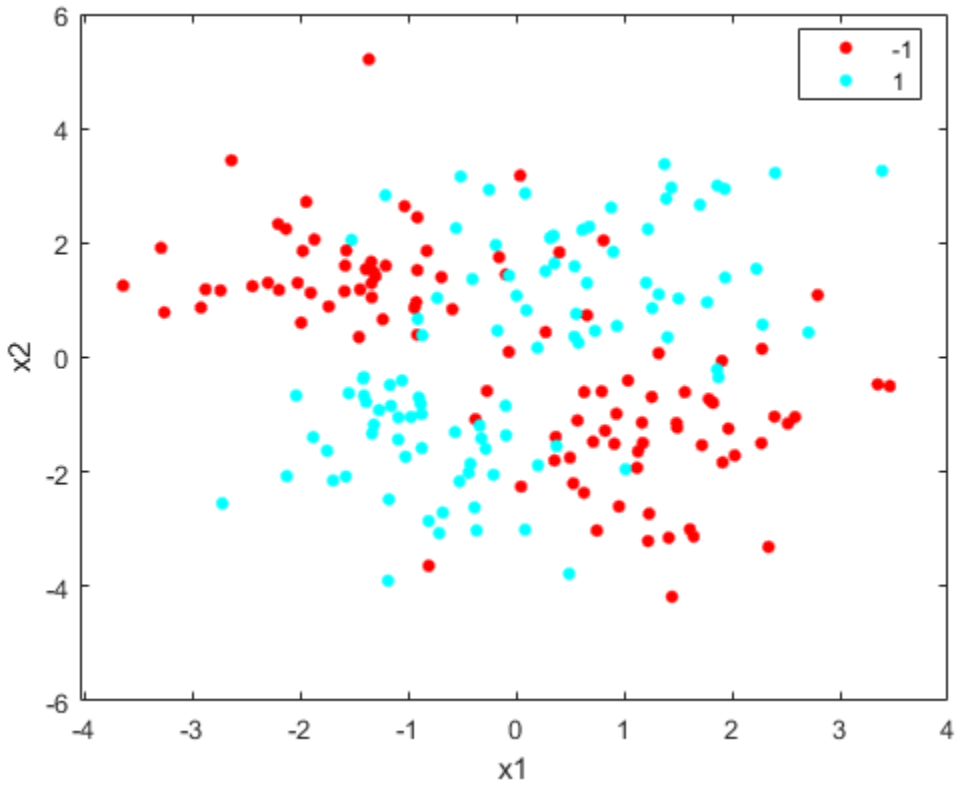
Load the sample data.

```
load('twodimclassdata.mat');
```

This data set is simulated using the scheme described in [1]. This is a two-class classification problem in two dimensions. Data from the first class (class –1) are drawn from two bivariate normal distributions $N(\mu_1, \Sigma)$ or $N(\mu_2, \Sigma)$ with equal probability, where $\mu_1 = [-0.75, -1.5]$, $\mu_2 = [0.75, 1.5]$, and $\Sigma = I_2$. Similarly, data from the second class (class 1) are drawn from two bivariate normal distributions $N(\mu_3, \Sigma)$ or $N(\mu_4, \Sigma)$ with equal probability, where $\mu_3 = [1.5, -1.5]$, $\mu_4 = [-1.5, 1.5]$, and $\Sigma = I_2$. The normal distribution parameters used to create this data set result in tighter clusters in data than the data used in [1].

Create a scatter plot of the data grouped by the class.

```
figure
gscatter(X(:,1),X(:,2),y)
xlabel('x1')
ylabel('x2')
```



Add 100 irrelevant features to X. First generate data from a Normal distribution with a mean of 0 and a variance of 20.

```
n = size(X,1);
rng('default')
XwithBadFeatures = [X,randn(n,100)*sqrt(20)];
```

Normalize the data so that all points are between 0 and 1.

```
XwithBadFeatures = bsxfun(@rdivide,...
    bsxfun(@minus,XwithBadFeatures,min(XwithBadFeatures,[],1)), ...
    range(XwithBadFeatures,1));
X = XwithBadFeatures;
```

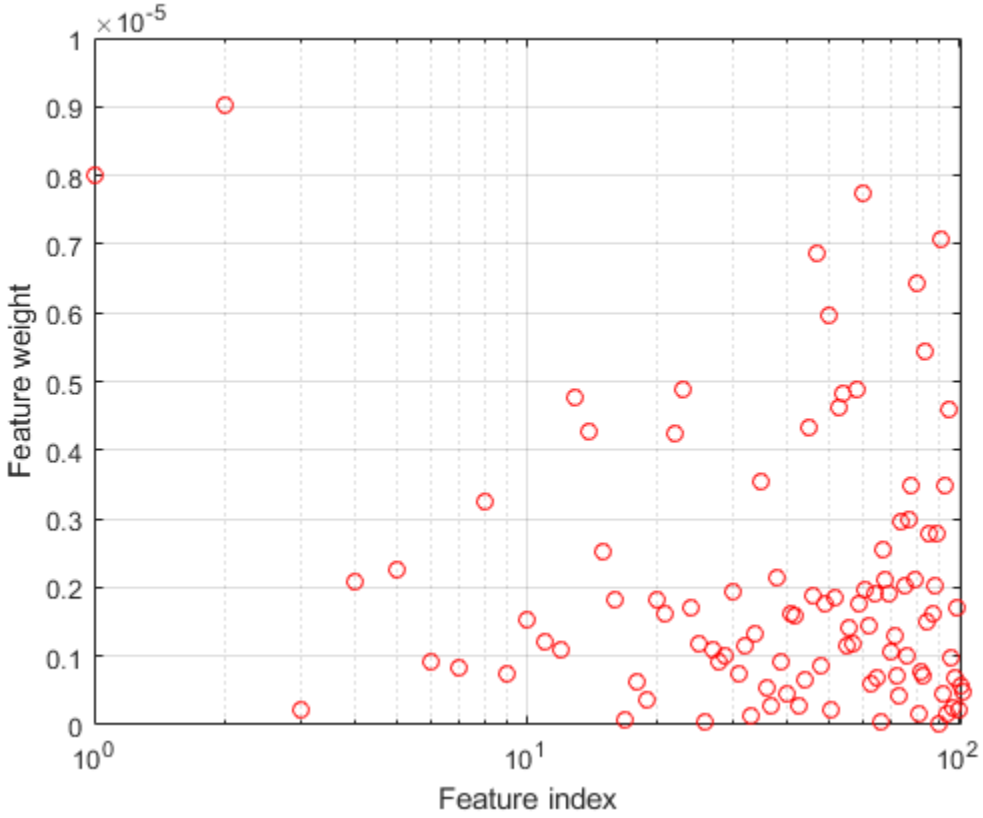
Fit a neighborhood component analysis (NCA) model to the data using the default `Lambda` (regularization parameter, λ) value. Use the LBFGS solver and display the convergence information.

```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
               'Solver','lbfgs');
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |    FUN VALUE    |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 |  9.519258e-03 |   1.494e-02 |   0.000e+00 |        |   4.015e+01 |   0.000e+00 |
|        1 | -3.093574e-01 |   7.186e-03 |   4.018e+00 |    OK  |   8.956e+01 |   1.000e+00 |
|        2 | -4.809455e-01 |   4.444e-03 |   7.123e+00 |    OK  |   9.943e+01 |   1.000e+00 |
|        3 | -4.938877e-01 |   3.544e-03 |   1.464e+00 |    OK  |   9.366e+01 |   1.000e+00 |
|        4 | -4.964759e-01 |   2.901e-03 |   6.084e-01 |    OK  |   1.554e+02 |   1.000e+00 |
|        5 | -4.972077e-01 |   1.323e-03 |   6.129e-01 |    OK  |   1.195e+02 |   5.000e-01 |
|        6 | -4.974743e-01 |   1.569e-04 |   2.155e-01 |    OK  |   1.003e+02 |   1.000e+00 |
|        7 | -4.974868e-01 |   3.844e-05 |   4.161e-02 |    OK  |   9.835e+01 |   1.000e+00 |
|        8 | -4.974874e-01 |   1.417e-05 |   1.073e-02 |    OK  |   1.043e+02 |   1.000e+00 |
|        9 | -4.974874e-01 |   4.893e-06 |   1.781e-03 |    OK  |   1.530e+02 |   1.000e+00 |
|       10 | -4.974874e-01 |   9.404e-08 |   8.947e-04 |    OK  |   1.670e+02 |   1.000e+00 |

         Infinity norm of the final gradient = 9.404e-08
              Two norm of the final step     = 8.947e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 9.404e-08, TolFun = 1.000e-06
EXIT: Local minimum found.
```

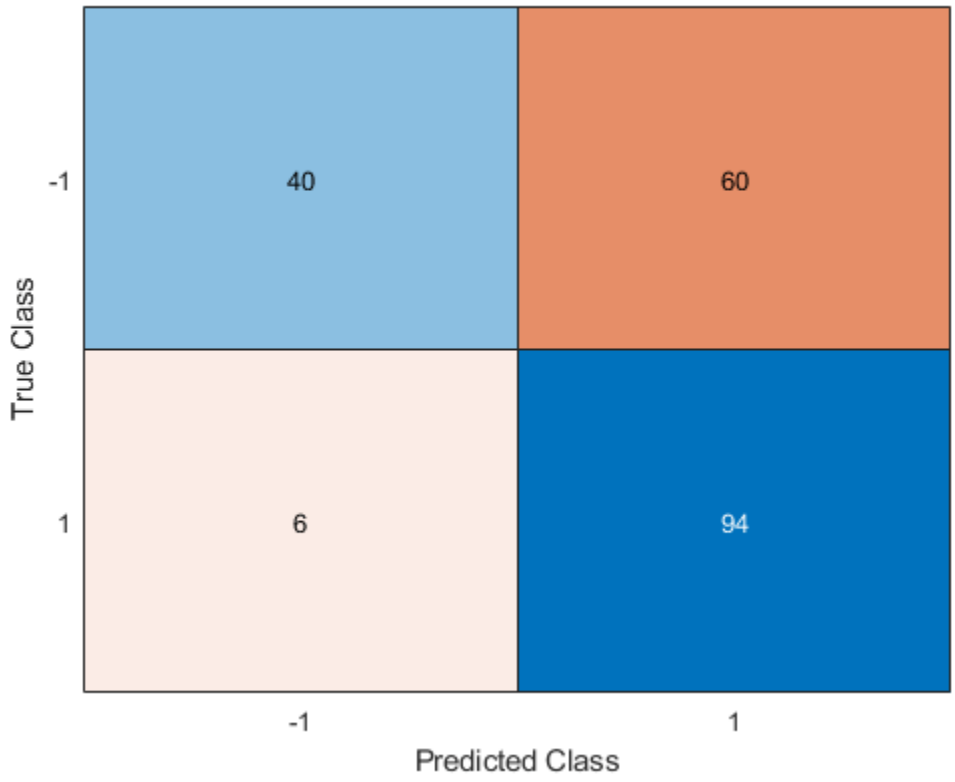
Plot the feature weights. The weights of the irrelevant features should be very close to zero.

```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



Predict the classes using the NCA model and compute the confusion matrix.

```
ypred = predict(ncaMdl,X);
confusionchart(y,ypred)
```



Confusion matrix shows that 40 of the data that are in class –1 are predicted as belonging to class –1. 60 of the data from class –1 are predicted to be in class 1. Similarly, 94 of the data from class 1 are predicted to be from class 1 and 6 of them are predicted to be from class –1. The prediction accuracy for class –1 is not good.

All weights are very close to zero, which indicates that the value of λ used in training the model is too large. When $\lambda \to \infty$, all features weights approach to zero. Hence, it is important to tune the regularization parameter in most cases to detect the relevant features.

Use five-fold cross-validation to tune λ for feature selection by using `fscnca`. Tuning λ means finding the λ value that will produce the minimum classification loss. To tune λ using cross-validation:

1. Partition the data into five folds. For each fold, `cvpartition` assigns four-fifths of the data as a training set and one-fifth of the data as a test set. Again for each fold, `cvpartition` creates a stratified partition, where each partition has roughly the same proportion of classes.

```
cvp = cvpartition(y,'kfold',5);
numtestsets = cvp.NumTestSets;
lambdavalues = linspace(0,2,20)/length(y);
lossvalues = zeros(length(lambdavalues),numtestsets);
```

2. Train the neighborhood component analysis (nca) model for each λ value using the training set in each fold.

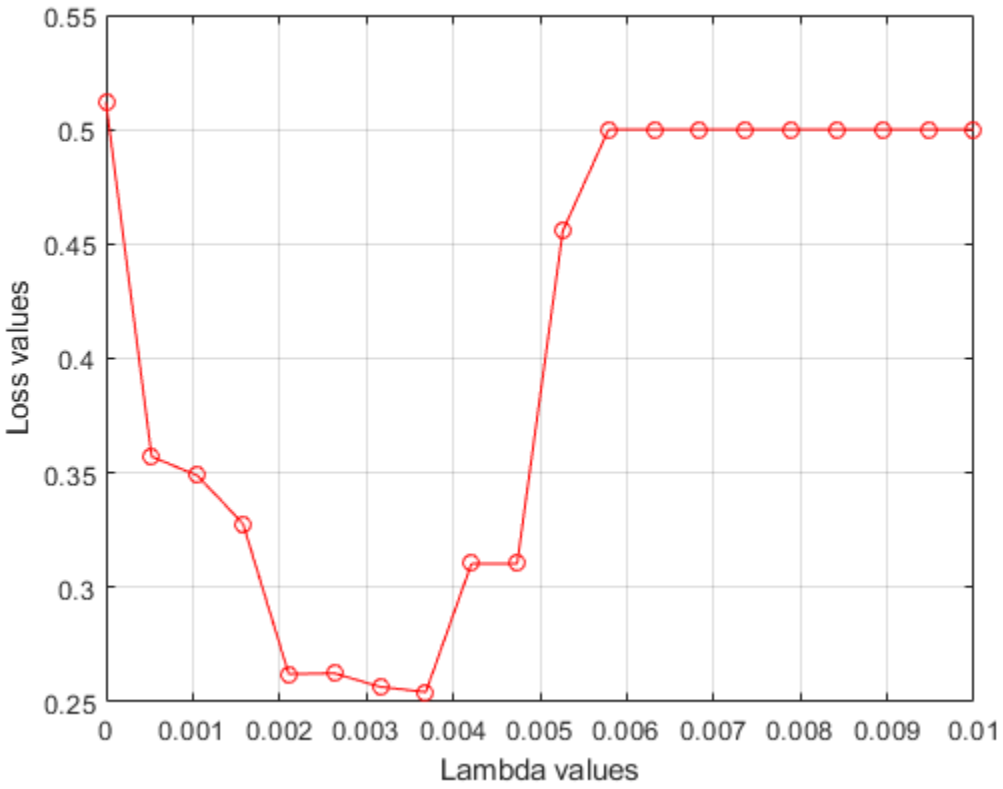
3. Compute the classification loss for the corresponding test set in the fold using the nca model. Record the loss value.

4. Repeat this process for all all λ values.

```
for i = 1:length(lambdavalues)
    for k = 1:numtestsets

        % Extract the training set from the partition object
        Xtrain = X(cvp.training(k),:);
        ytrain = y(cvp.training(k),:);

        % Extract the test set from the partition object
        Xtest  = X(cvp.test(k),:);
        ytest  = y(cvp.test(k),:);

        % Train an NCA model for classification using the training set
        ncaMdl = fscnca(Xtrain,ytrain,'FitMethod','exact', ...
            'Solver','lbfgs','Lambda',lambdavalues(i));

        % Compute the classification loss for the test set using the NCA
        % model
        lossvalues(i,k) = loss(ncaMdl,Xtest,ytest, ...
            'LossFunction','quadratic');

    end
end
```

Plot the average loss values of the folds versus the λ values. If the λ value that corresponds to the minimum loss falls on the boundary of the tested λ values, the range of λ values should be reconsidered.

```
figure
plot(lambdavalues,mean(lossvalues,2),'ro-')
xlabel('Lambda values')
ylabel('Loss values')
grid on
```



Find the λ value that corresponds to the minimum average loss.

```
[~,idx] = min(mean(lossvalues,2)); % Find the index
bestlambda = lambdavalues(idx) % Find the best lambda value
```

```
bestlambda =

    0.0037
```

Fit the NCA model to all of the data using the best λ value. Use the LBFGS solver and display the convergence information.

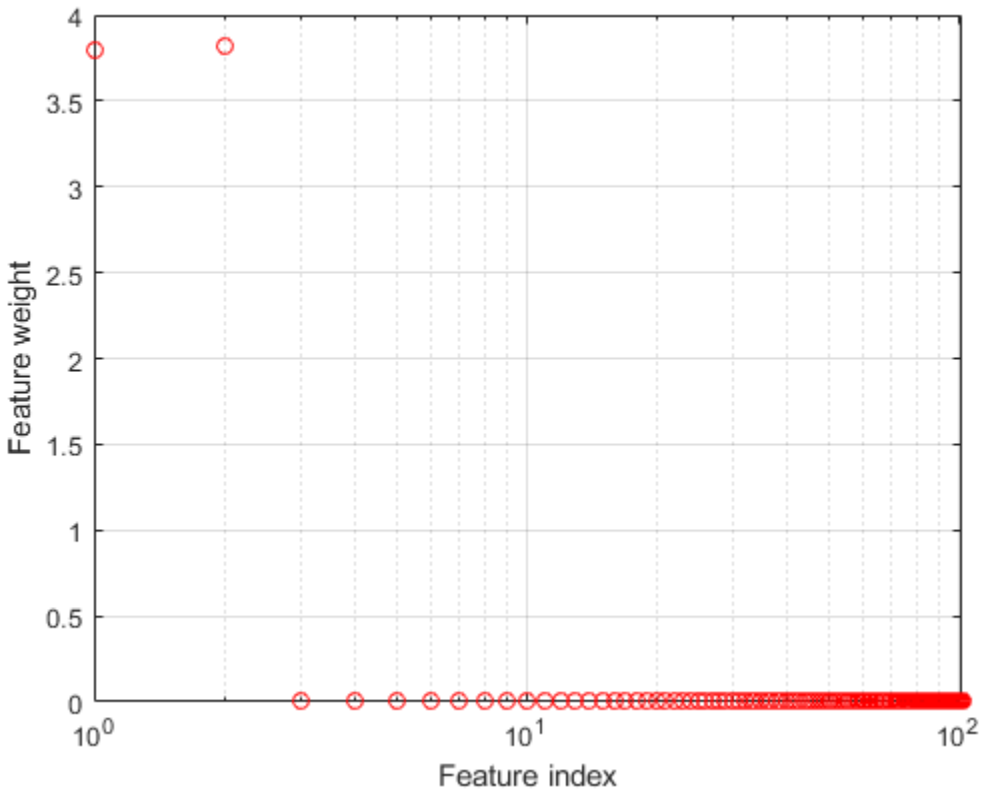
```
ncaMdl = fscnca(X,y,'FitMethod','exact','Verbose',1, ...
        'Solver','lbfgs','Lambda',bestlambda);
```

```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=====================================================================================================
|        0 | -1.246913e-01 |   1.231e-02 |   0.000e+00 |        |   4.873e+01 |   0.000e+00 |
|        1 | -3.411330e-01 |   5.717e-03 |   3.618e+00 |    OK  |   1.068e+02 |   1.000e+00 |
|        2 | -5.226111e-01 |   3.763e-02 |   8.252e+00 |    OK  |   7.825e+01 |   1.000e+00 |
|        3 | -5.817731e-01 |   8.496e-03 |   2.340e+00 |    OK  |   5.591e+01 |   5.000e-01 |
|        4 | -6.132632e-01 |   6.863e-03 |   2.526e+00 |    OK  |   8.228e+01 |   1.000e+00 |
```

```
|        5 | -6.135264e-01 |   9.373e-03 |   7.341e-01 |    OK  |   3.244e+01 |   1.000e+00 |   Y
|        6 | -6.147894e-01 |   1.182e-03 |   2.933e-01 |    OK  |   2.447e+01 |   1.000e+00 |   Y
|        7 | -6.148714e-01 |   6.392e-04 |   6.688e-02 |    OK  |   3.195e+01 |   1.000e+00 |   Y
|        8 | -6.149524e-01 |   6.521e-04 |   9.934e-02 |    OK  |   1.236e+02 |   1.000e+00 |   Y
|        9 | -6.149972e-01 |   1.154e-04 |   1.191e-01 |    OK  |   1.171e+02 |   1.000e+00 |   Y
|       10 | -6.149990e-01 |   2.922e-05 |   1.983e-02 |    OK  |   7.365e+01 |   1.000e+00 |   Y
|       11 | -6.149993e-01 |   1.556e-05 |   8.354e-03 |    OK  |   1.288e+02 |   1.000e+00 |   Y
|       12 | -6.149994e-01 |   1.147e-05 |   7.256e-03 |    OK  |   2.332e+02 |   1.000e+00 |   Y
|       13 | -6.149995e-01 |   1.040e-05 |   6.781e-03 |    OK  |   2.287e+02 |   1.000e+00 |   Y
|       14 | -6.149996e-01 |   9.015e-06 |   6.265e-03 |    OK  |   9.974e+01 |   1.000e+00 |   Y
|       15 | -6.149996e-01 |   7.763e-06 |   5.206e-03 |    OK  |   2.919e+02 |   1.000e+00 |   Y
|       16 | -6.149997e-01 |   8.374e-06 |   1.679e-02 |    OK  |   6.878e+02 |   1.000e+00 |   Y
|       17 | -6.149997e-01 |   9.387e-06 |   9.542e-03 |    OK  |   1.284e+02 |   5.000e-01 |   Y
|       18 | -6.149997e-01 |   3.250e-06 |   5.114e-03 |    OK  |   1.225e+02 |   1.000e+00 |   Y
|       19 | -6.149997e-01 |   1.574e-06 |   1.275e-03 |    OK  |   1.808e+02 |   1.000e+00 |   Y

|=============================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | ACC
|=============================================================================================================
|       20 | -6.149997e-01 |   5.764e-07 |   6.765e-04 |    OK  |   2.905e+02 |   1.000e+00 |   Y

         Infinity norm of the final gradient = 5.764e-07
             Two norm of the final step     = 6.765e-04, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 5.764e-07, TolFun = 1.000e-06
EXIT: Local minimum found.
```

Plot the feature weights.

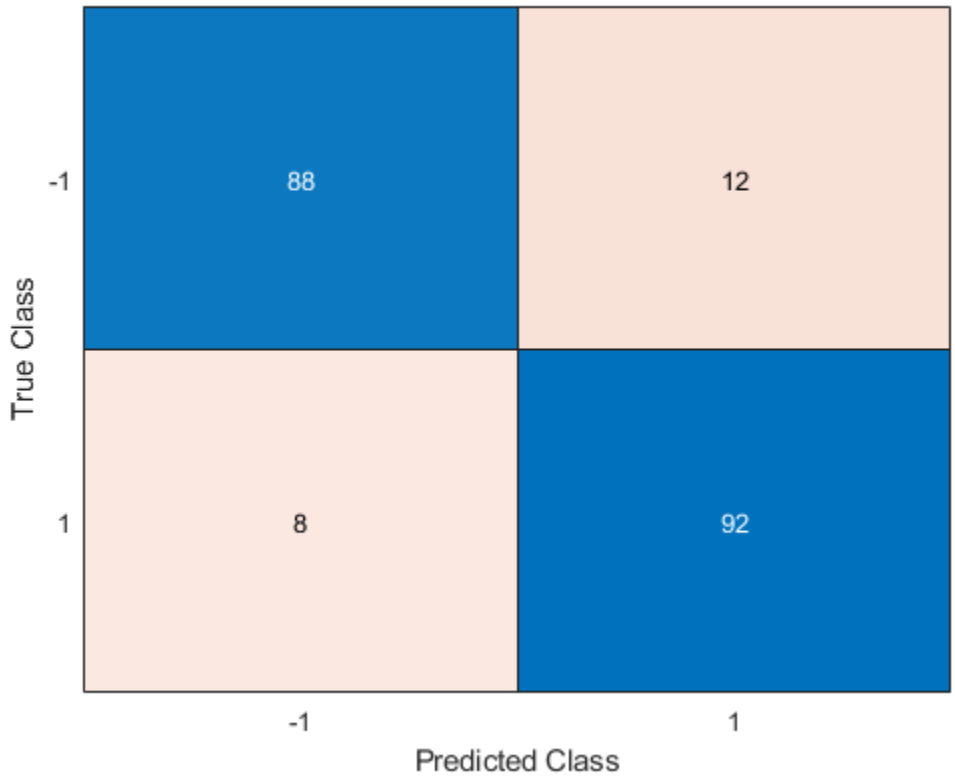
```
figure
semilogx(ncaMdl.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



`fscnca` correctly figures out that the first two features are relevant and that the rest are not. The first two features are not individually informative, but when taken together result in an accurate classification model.

Predict the classes using the new model and compute the accuracy.

```
ypred = predict(ncaMdl,X);
confusionchart(y,ypred)
```



Confusion matrix shows that prediction accuracy for class –1 has improved. 88 of the data from class –1 are predicted to be from –1, and 12 of them are predicted to be from class 1. 92 of the data from class 1 are predicted to be from class 1 and 8 of them are predicted to be from class –1.

References

[1] Yang, W., K. Wang, W. Zuo. "Neighborhood Component Feature Selection for High-Dimensional Data." *Journal of Computers*. Vol. 7, Number 1, January, 2012.

See Also

`FeatureSelectionNCAClassification` | `fscnca` | `loss` | `refit`

Introduced in R2016b

predict

Class: `FeatureSelectionNCARegression`

Predict responses using neighborhood component analysis (NCA) regression model

Syntax

```
ypred = predict(mdl,X)
```

Description

`ypred = predict(mdl,X)` computes the predicted response values, `ypred`, corresponding to rows of `X`, using the model `mdl`.

Input Arguments

`mdl` — Neighborhood component analysis model for regression

`FeatureSelectionNCARegression` object

Neighborhood component analysis model for regression, specified as a `FeatureSelectionNCARegression` object.

`X` — Predictor variable values

n-by-*p* matrix

Predictor variable values, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables.

Data Types: `single` | `double`

Output Arguments

`ypred` — Predicted response values

n-by-1 vector

Predicted response values, specified as an *n*-by-1 vector, where *n* is the number of observations.

Examples

Tune NCA Model for Regression Using `loss` and `predict`

Load the sample data.

Download the housing data [1], from the UCI Machine Learning Repository [2]. The dataset has 506 observations. The first 13 columns contain the predictor values and the last column contains the response values. The goal is to predict the median value of owner-occupied homes in suburban Boston as a function of 13 predictors.

Load the data and define the response vector and the predictor matrix.

```
load('housing.data');
X = housing(:,1:13);
y = housing(:,end);
```

Divide the data into training and test sets using the 4th predictor as the grouping variable for a stratified partitioning. This ensures that each partition includes similar amount of observations from each group.

```
rng(1) % For reproducibility
cvp = cvpartition(X(:,4),'Holdout',56);
Xtrain = X(cvp.training,:);
ytrain = y(cvp.training,:);
Xtest  = X(cvp.test,:);
ytest  = y(cvp.test,:);
```

`cvpartition` randomly assigns 56 observations into a test set and the rest of the data into a training set.

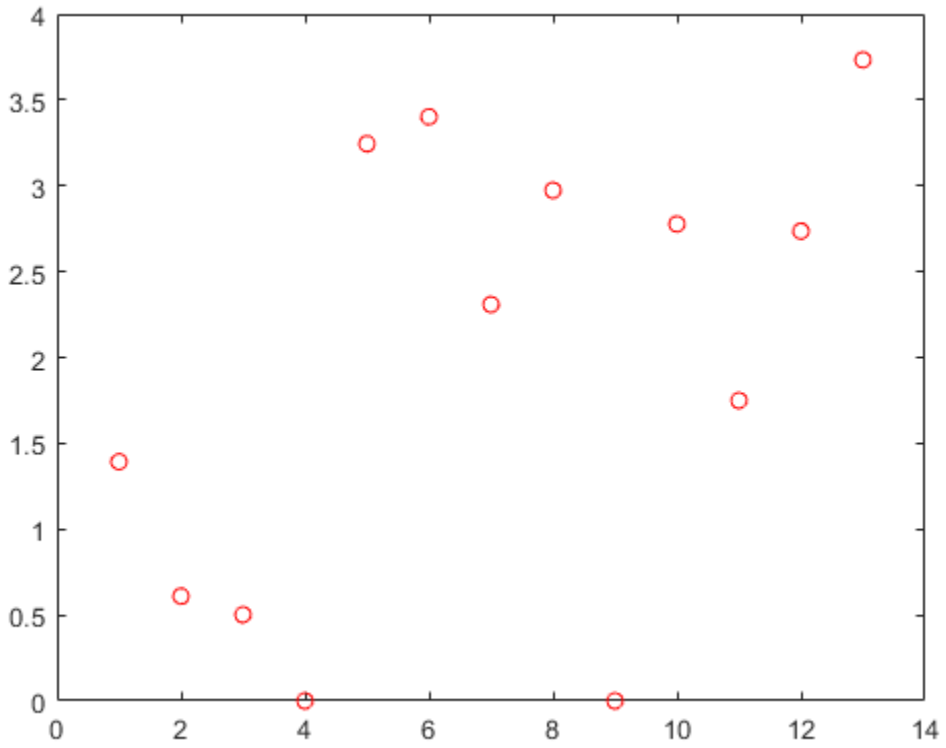
Perform Feature Selection Using Default Settings

Perform feature selection using NCA model for regression. Standardize the predictor values.

```
nca = fsrnca(Xtrain,ytrain,'Standardize',1);
```

Plot the feature weights.

```
figure()
plot(nca.FeatureWeights,'ro')
```



The weights of irrelevant features are expected to approach zero. `fsrnca` identifies two features as irrelevant.

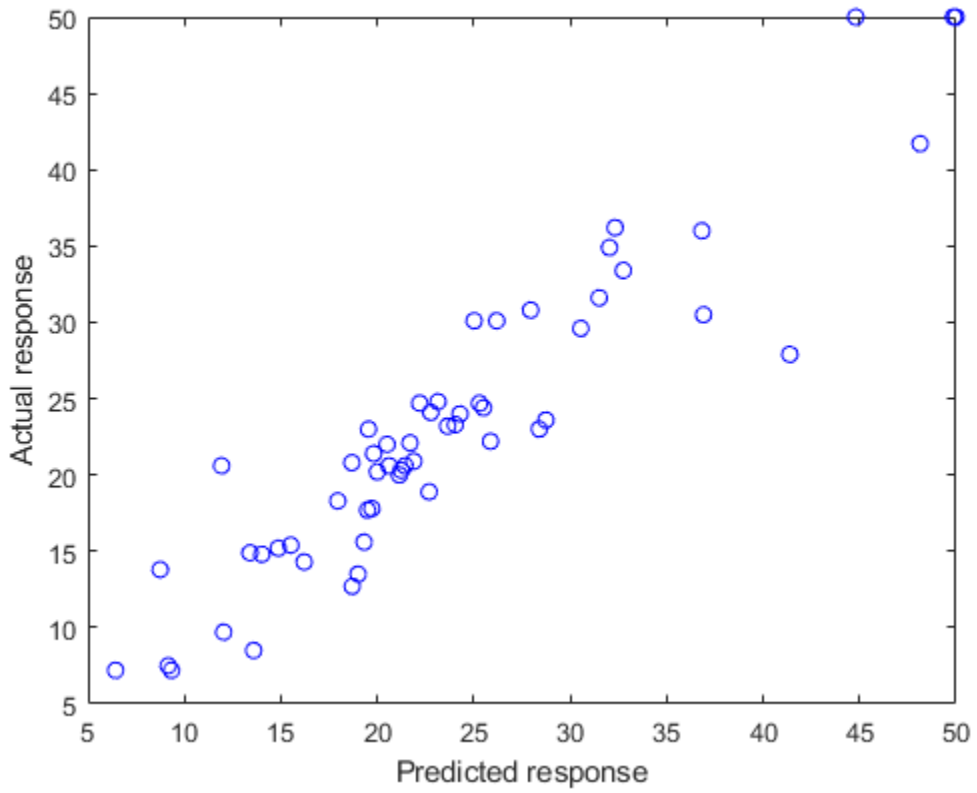
Compute the regression loss.

```
L = loss(nca,Xtest,ytest,'LossFunction','mad')
```

```
L = 2.5394
```

Compute the predicted response values for the test set and plot them versus the actual response.

```
ypred = predict(nca,Xtest);
figure()
plot(ypred,ytest,'bo')
xlabel('Predicted response')
ylabel('Actual response')
```



A perfect fit versus the actual values forms a 45 degree straight line. In this plot, the predicted and actual response values seem to be scattered around this line. Tuning λ (regularization parameter) value usually helps improve the performance.

Tune the regularization parameter using 10-fold cross-validation

Tuning λ means finding the λ value that will produce the minimum regression loss. Here are the steps for tuning λ using 10-fold cross-validation:

1. First partition the data into 10 folds. For each fold, `cvpartition` assigns 1/10th of the data as a training set, and 9/10th of the data as a test set.

```
n = length(ytrain);
cvp = cvpartition(Xtrain(:,4),'kfold',10);
numvalidsets = cvp.NumTestSets;
```

Assign the λ values for the search. Create an array to store the loss values.

```
lambdavals = linspace(0,2,30)*std(ytrain)/n;
lossvals = zeros(length(lambdavals),numvalidsets);
```

2. Train the neighborhood component analysis (nca) model for each λ value using the training set in each fold.

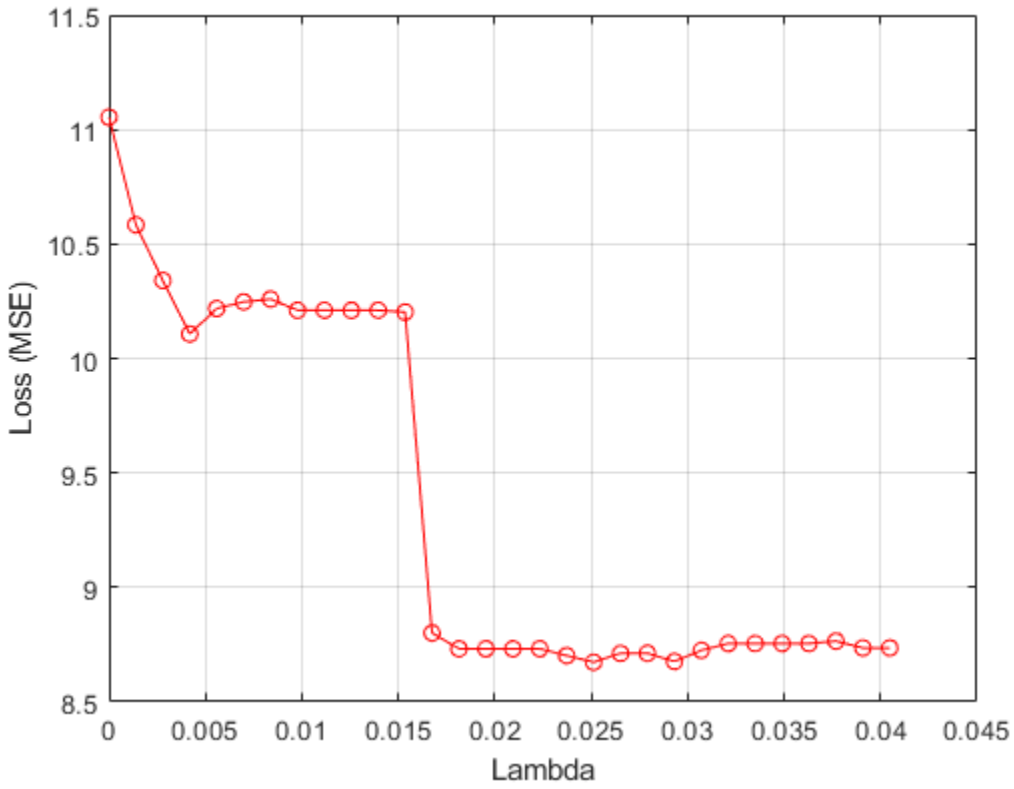
3. Fit a Gaussian process regression (gpr) model using the selected features. Next, compute the regression loss for the corresponding test set in the fold using the gpr model. Record the loss value.

4. Repeat this for each λ value and each fold.

```
 for i = 1:length(lambdavals)
    for k = 1:numvalidsets
        X = Xtrain(cvp.training(k),:);
        y = ytrain(cvp.training(k),:);
        Xvalid  = Xtrain(cvp.test(k),:);
        yvalid  = ytrain(cvp.test(k),:);

        nca = fsrnca(X,y,'FitMethod','exact',...
             'Lambda',lambdavals(i),...
             'Standardize',1,'LossFunction','mad');

        % Select features using the feature weights and a relative
        % threshold.
        tol    = 1e-3;
        selidx = nca.FeatureWeights > tol*max(1,max(nca.FeatureWeights));

        % Fit a non-ARD GPR model using selected features.
        gpr = fitrgp(X(:,selidx),y,'Standardize',1,...
              'KernelFunction','squaredexponential','Verbose',0);


        lossvals(i,k) = loss(gpr,Xvalid(:,selidx),yvalid);

    end
 end
```

Compute the average loss obtained from the folds for each λ value. Plot the mean loss versus the λ values.

```
meanloss = mean(lossvals,2);
figure;
plot(lambdavals,meanloss,'ro-');
xlabel('Lambda');
ylabel('Loss (MSE)');
grid on;
```



Find the λ value that produces the minimum loss value.

```
[~,idx] = min(meanloss);
bestlambda = lambdavals(idx)
```

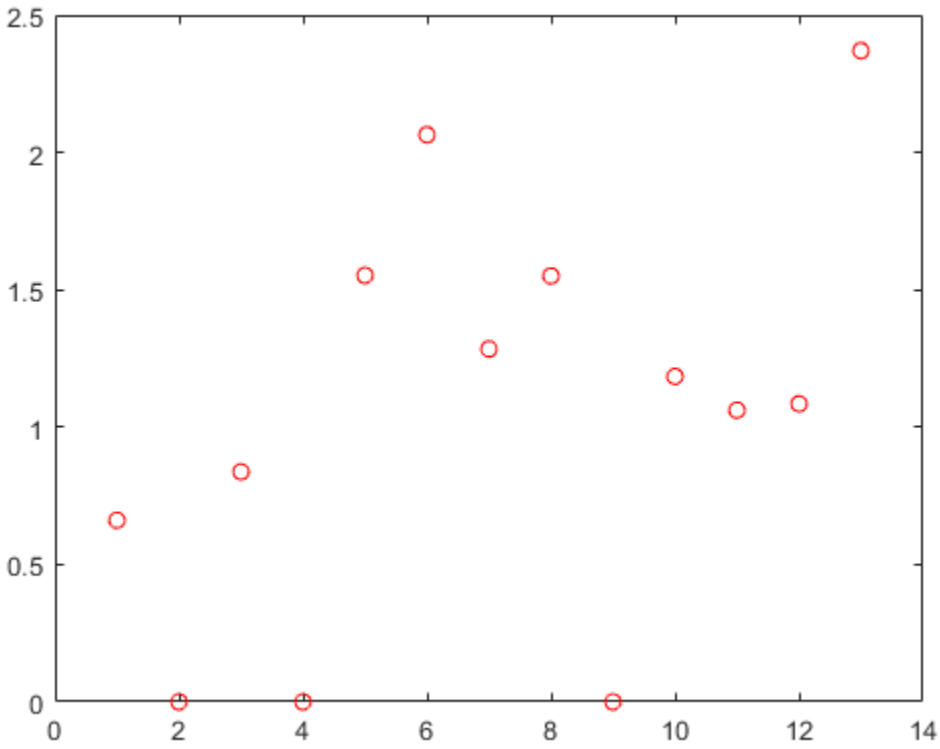
```
bestlambda = 0.0251
```

Perform feature selection for regression using the best λ value. Standardize the predictor values.

```
nca2 = fsrnca(Xtrain,ytrain,'Standardize',1,'Lambda',bestlambda,...
    'LossFunction','mad');
```

Plot the feature weights.

```
figure()
plot(nca.FeatureWeights,'ro')
```



Compute the loss using the new nca model on the test data, which is not used to select the features.

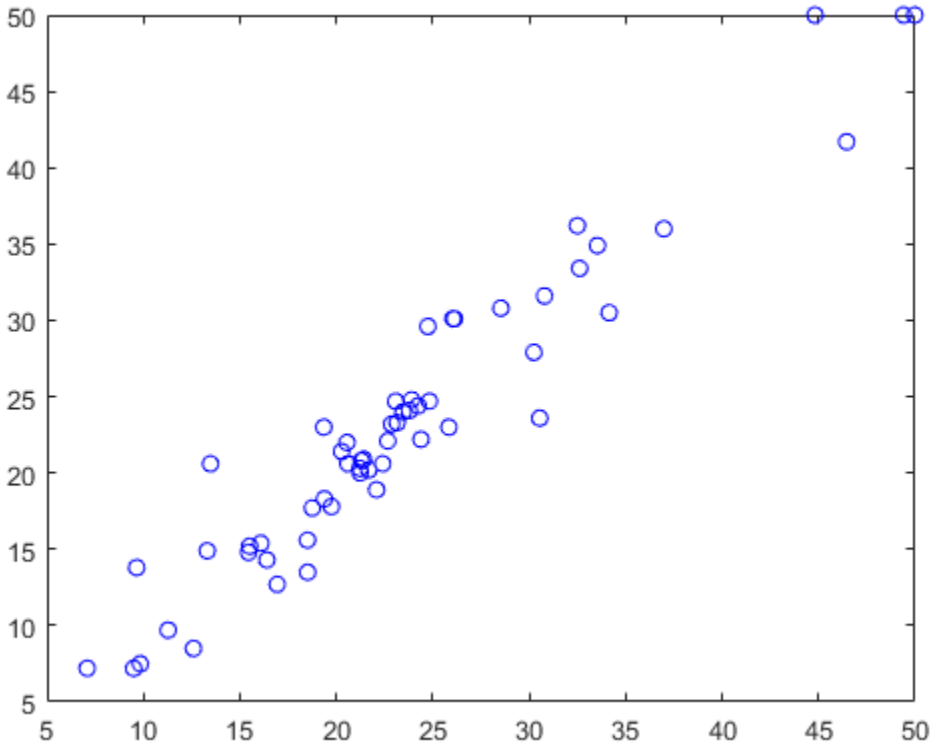
```
L2 = loss(nca2,Xtest,ytest,'LossFunction','mad')
```

```
L2 = 2.0560
```

Tuning the regularization parameter helps identify the relevant features and reduces the loss.

Plot the predicted versus the actual response values in the test set.

```
ypred = predict(nca2,Xtest);
figure;
plot(ypred,ytest,'bo');
```



The predicted response values seem to be closer to the actual values as well.

References

[1] Harrison, D. and D.L., Rubinfeld. "Hedonic prices and the demand for clean air." J. Environ. Economics & Management. Vol.5, 1978, pp. 81-102.

[2] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. https://archive.ics.uci.edu/ml.

See Also

`FeatureSelectionNCARegression` | `fsrnca` | `loss` | `refit`

Introduced in R2016b

predict

Class: NonLinearModel

Predict response of nonlinear regression model

Syntax

```
ypred = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew)
[ypred,yci] = predict(mdl,Xnew,Name,Value)
```

Description

`ypred = predict(mdl,Xnew)` returns the predicted response of the `mdl` nonlinear regression model to the points in `Xnew`.

`[ypred,yci] = predict(mdl,Xnew)` returns confidence intervals for the true mean responses.

`[ypred,yci] = predict(mdl,Xnew,Name,Value)` predicts responses with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`Xnew`

Points at which `mdl` predicts responses.

- If `Xnew` is a table or dataset array, it must contain the predictor names in `mdl`.
- If `Xnew` is a numeric matrix, it must have the same number of variables (columns) as was used to create `mdl`. Furthermore, all variables used in creating `mdl` must be numeric.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha`

Positive scalar from `0` to `1`. Confidence level of `yci` is 100(1 – `alpha`)%.

Default: `0.05`, meaning a 95% confidence interval.

`Prediction`

Type of prediction:

- `'curve'` — `predict` predicts confidence bounds for the fitted mean values.
- `'observation'` — `predict` predicts confidence bounds for the new observations. This results in wider bounds because the error in a new observation is equal to the error in the estimated mean value, plus the variability in the observation from the true mean.

For details, see `polyconf`.

Default: `'curve'`

`Simultaneous`

Logical value specifying whether the confidence bounds are for all predictor values simultaneously (`true`), or hold for each individual predictor value (`false`). Simultaneous bounds are wider than separate bounds, because it is more stringent to require that the entire curve be within the bounds than to require that the curve at a single predictor value be within the bounds.

For details, see `polyconf`.

Default: `false`

`Weights`

Vector of real, positive value weights or a function handle.

- If you specify a vector, then it must have the same number of elements as the number of observations (or rows) in `Xnew`.
- If you specify a function handle, then the function must accept a vector of predicted response values as input, and return a vector of real positive weights as output.

Given weights, `W`, `predict` estimates the error variance at observation `i` by `MSE*(1/W(i))`, where MSE is the mean squared error.

Default: No weights

Output Arguments

`ypred`

Predicted mean values at `Xnew`. `ypred` is the same size as each component of `Xnew`.

`yci`

Confidence intervals, a two-column matrix with each row providing one interval. The meaning of the confidence interval depends on the settings of the name-value pairs.

Examples

Predict Responses

Create a nonlinear model of car mileage as a function of weight, and predict the response.

Create an exponential model of car mileage as a function of weight from the `carsmall` data. Scale the weight by a factor of 1000 so all the variables are roughly equal in size.

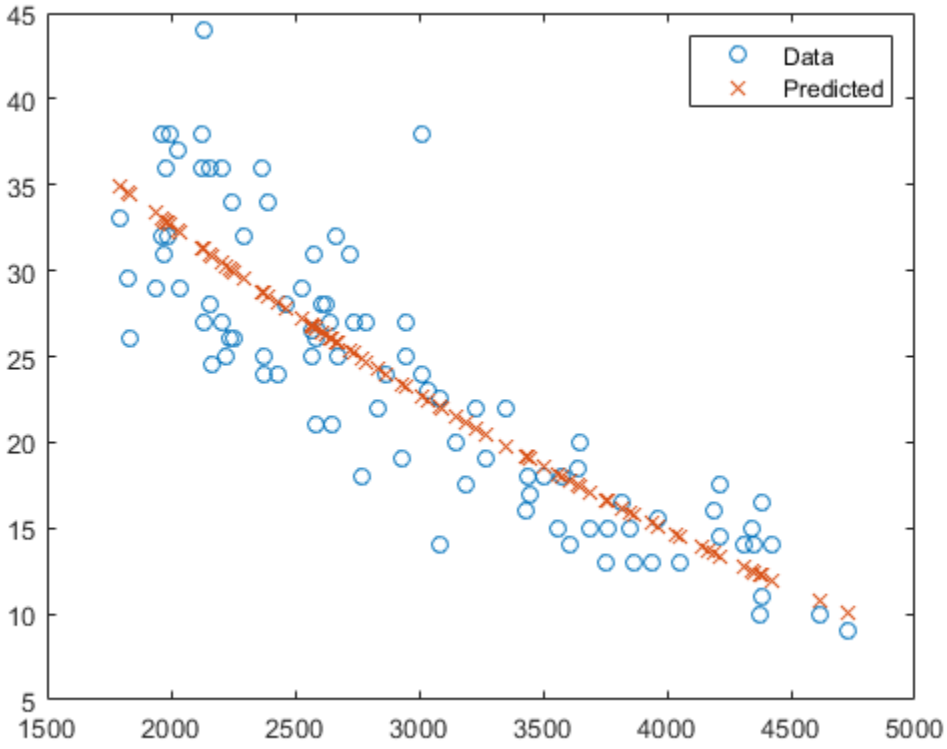
```
load carsmall
X = Weight;
y = MPG;
modelfun = 'y ~ b1 + b2*exp(-b3*x/1000)';
beta0 = [1 1 1];
mdl = fitnlm(X,y,modelfun,beta0);
```

Create predicted responses to the data.

```
Xnew = X;
ypred = predict(mdl,Xnew);
```

Plot the original responses and the predicted responses to see how they differ.

```
plot(X,y,'o',X,ypred,'x')
legend('Data','Predicted')
```



Confidence Intervals for Predictions

Create a nonlinear model of car mileage as a function of weight, and examine confidence intervals of some responses.

Create an exponential model of car mileage as a function of weight from the `carsmall` data. Scale the weight by a factor of 1000 so all the variables are roughly equal in size.

```
load carsmall
X = Weight;
y = MPG;
modelfun = 'y ~ b1 + b2*exp(-b3*x/1000)';
beta0 = [1 1 1];
mdl = fitnlm(X,y,modelfun,beta0);
```

Create predicted responses to the smallest, mean, and largest data points.

```
Xnew = [min(X);mean(X);max(X)];
[ypred,yci] = predict(mdl,Xnew)
```

```
ypred = 3×1

   34.9469
   22.6868
   10.0617


yci = 3×2

   32.5212   37.3726
   21.4061   23.9674
    7.0148   13.1086
```

Simultaneous Confidence Intervals for Robust Fit Curve

Generate sample data from the nonlinear regression model

$$y = b_1 + b_2\exp(-b_3x) + \epsilon$$

where b_1, b_2, and b_3 are coefficients, and the error term ϵ is normally distributed with mean 0 and standard deviation 0.5.

```
modelfun = @(b,x)(b(1)+b(2)*exp(-b(3)*x));

rng('default') % For reproducibility
b = [1;3;2];
x = exprnd(2,100,1);
y = modelfun(b,x) + normrnd(0,0.5,100,1);
```

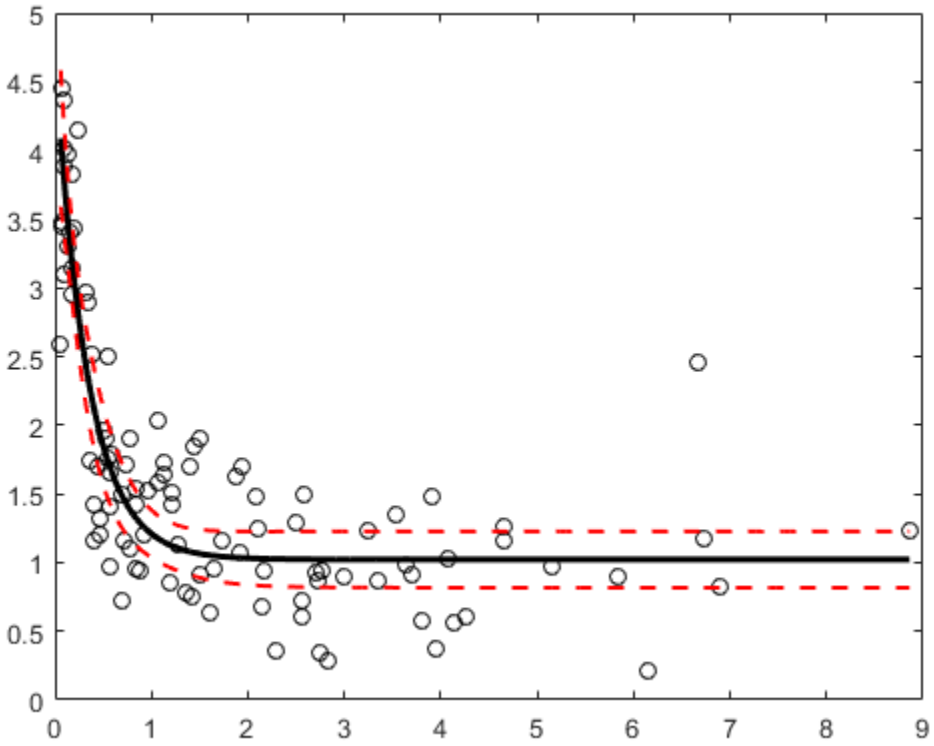
Fit the nonlinear model using robust fitting options.

```
opts = statset('nlinfit');
opts.RobustWgtFun = 'bisquare';
b0 = [2;2;2];
mdl = fitnlm(x,y,modelfun,b0,'Options',opts);
```

Plot the fitted regression model and simultaneous 95% confidence bounds.

```
xrange = [min(x):.01:max(x)]';
[ypred,yci] = predict(mdl,xrange,'Simultaneous',true);

figure()
plot(x,y,'ko') % observed data
```

```
hold on
plot(xrange,ypred,'k','LineWidth',2)
plot(xrange,yci','r--','LineWidth',1.5)
```



Confidence Interval Using Observation Weights

Load sample data.

```
S = load('reaction');
X = S.reactants;
y = S.rate;
beta0 = S.beta;
```

Specify a function handle for observation weights, then fit the Hougen-Watson model to the rate data using the specified observation weights function.

```
a = 1; b = 1;
weights = @(yhat) 1./((a + b*abs(yhat)).^2);
mdl = fitnlm(X,y,@hougen,beta0,'Weights',weights);
```

Compute the 95% prediction interval for a new observation with reactant levels [100,100,100] using the observation weight function.

```
[ypred,yci] = predict(mdl,[100,100,100],'Prediction','observation', ...
    'Weights',weights)
```

```
ypred = 1.8149

yci = 1×2

    1.5264    2.1033
```

Tips

- For predictions with added noise, use `random`.
- For a syntax that can be easier to use with models created from tables or dataset arrays, try `feval`.

References

[1] Lane, T. P. and W. H. DuMouchel. “Simultaneous Confidence Intervals in Multiple Regression.” *The American Statistician*. Vol. 48, No. 4, 1994, pp. 315–321.

[2] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

See Also

`NonLinearModel` | `random`

Topics

“Predict or Simulate Responses Using a Nonlinear Model” on page 13-9
“Nonlinear Regression Workflow” on page 13-12
“Nonlinear Regression” on page 13-2

predict

Class: `RepeatedMeasuresModel`

Compute predicted values given predictor values

Syntax

```
ypred = predict(rm,tnew)
ypred = predict(rm,tnew,Name,Value)
[ypred,yci] = predict( ___ )
```

Description

`ypred = predict(rm,tnew)` returns the predicted values from the repeated measures model `rm` using the predictor values from the table `t`.

`ypred = predict(rm,tnew,Name,Value)` returns the predicted values from the repeated measures model `rm` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify the within-subjects design matrix.

`[ypred,yci] = predict( ___ )` also returns the 95% confidence interval for the predicted values.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`tnew` — New data

table used to create `rm` (default) | table

New data including the values of the response variables and the between-subject factors used as predictors in the repeated measures model, `rm`, specified as a table. `tnew` must contain all of the between-subject factors used to create `rm`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range of 0 through 1

Significance level of the confidence intervals for the predicted values, specified as the comma-separated pair consisting of `'alpha'` and a scalar value in the range of 0 to 1. The confidence level is 100*(1–`alpha`)%.

Example: `'alpha',0.01`

Data Types: `double` | `single`

`WithinModel` — Model for within-subject factors
`'separatemeans'` | `'orthogonalcontrats'` | character vector | string scalar

Model for the within-subject factors, specified as the comma-separated pair consisting of `'WithinModel'` and one of the following:

- `'separatemeans'` — Compute a separate mean for each group.
- `'orthogonalcontrasts'` — Valid when the within-subject design consists of a single numeric factor *T*. This specifies a model consisting of orthogonal polynomials up to order $T^{(r-1)}$, where *r* is the number of repeated measures.
- A character vector or string scalar that defines a model specification in the within-subject factors.

Example: `'WithinModel','orthogonalcontrasts'`

Data Types: `char` | `string`

`WithinDesign` — Design for within-subject factors
vector | matrix | table

Design for within-subject factors, specified as the comma-separated pair consisting of `'WithinDesign'` and a vector, matrix, or a table. It provides the values of the within-subject factors in the same form as the `RM.WithinDesign` property.

Example: `'WithinDesign','Time'`

Data Types: `single` | `double` | `table`

Output Arguments

`ypred` — Predicted values
n-by-*r* matrix

Predicted values from the repeated measures model `rm`, returned as an *n*-by-*r* matrix, where *n* is the number of rows in `tnew` and *r* is the number of repeated measures in `rm`.

`yci` — Confidence intervals for predicted values
n-by-*r*-by-2 matrix

Confidence intervals for predicted values from the repeated measures model `rm`, returned as an *n*-by-*r*-by-2 matrix.

These are nonsimultaneous intervals for predicting the mean response at the specified predictor values. For predicted value `ypred(i,j)`, the lower limit of the interval is `yci(i,j,1)` and the upper limit is `yci(i,j,2)`.

Examples

Predict Response Values

Load the sample data.

```
load fisheriris
```

The column vector, `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4), ...
    'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Predict responses for the three species.

```
Y = predict(rm,t([1 51 101],:))
```

```
Y = 3×4

    5.0060    3.4280    1.4620    0.2460
    5.9360    2.7700    4.2600    1.3260
    6.5880    2.9740    5.5520    2.0260
```

Predict Response Values and Plot Predictions

Load the sample data.

```
load longitudinalData
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (time = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual, and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Define a variable that stores gender information.

```
Gender = ['F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']';
```

Store the data in a proper table array format to perform repeated measures analysis.

```
t = table(Gender,Y(:,1),Y(:,2),Y(:,3),Y(:,4),Y(:,5), ...
    'VariableNames',{'Gender','t0','t2','t4','t6','t8'});
```

Define the within-subjects variable.

```
Time = [0 2 4 6 8]';
```

Fit a repeated measures model, where the blood levels are the responses and gender is the predictor variable.

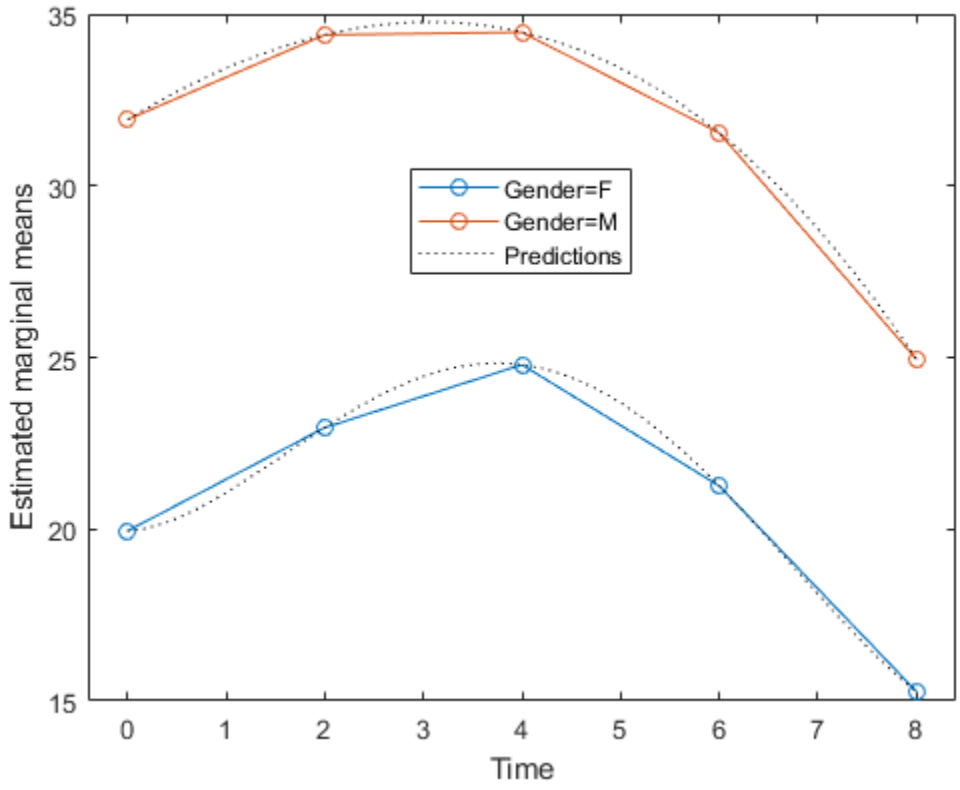
```
rm = fitrm(t,'t0-t8 ~ Gender','WithinDesign',Time);
```

Predict the responses at intermediate times.

```
time = linspace(0,8)';
Y = predict(rm,t([1 5 8 12],:), ...
    'WithinModel','orthogonalcontrasts','WithinDesign',time);
```

Plot the predictions along with the estimated marginal means.

```
plotprofile(rm,'Time','Group',{'Gender'})
hold on;
plot(time,Y,'Color','k','LineStyle',':');
legend('Gender=F','Gender=M','Predictions')
hold off
```



Compute and Plot Confidence Intervals

Load the sample data.

```
load longitudinalData
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (time = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual, and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Define a variable that stores gender information.

```
Gender = ['F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']';
```

Store the data in a proper table array format to perform repeated measures analysis.

```
t = table(Gender,Y(:,1),Y(:,2),Y(:,3),Y(:,4),Y(:,5), ...
    'VariableNames',{'Gender','t0','t2','t4','t6','t8'});
```

Define the within-subjects variable.

```
Time = [0 2 4 6 8]';
```

Fit a repeated measures model, where the blood levels are the responses and gender is the predictor variable.

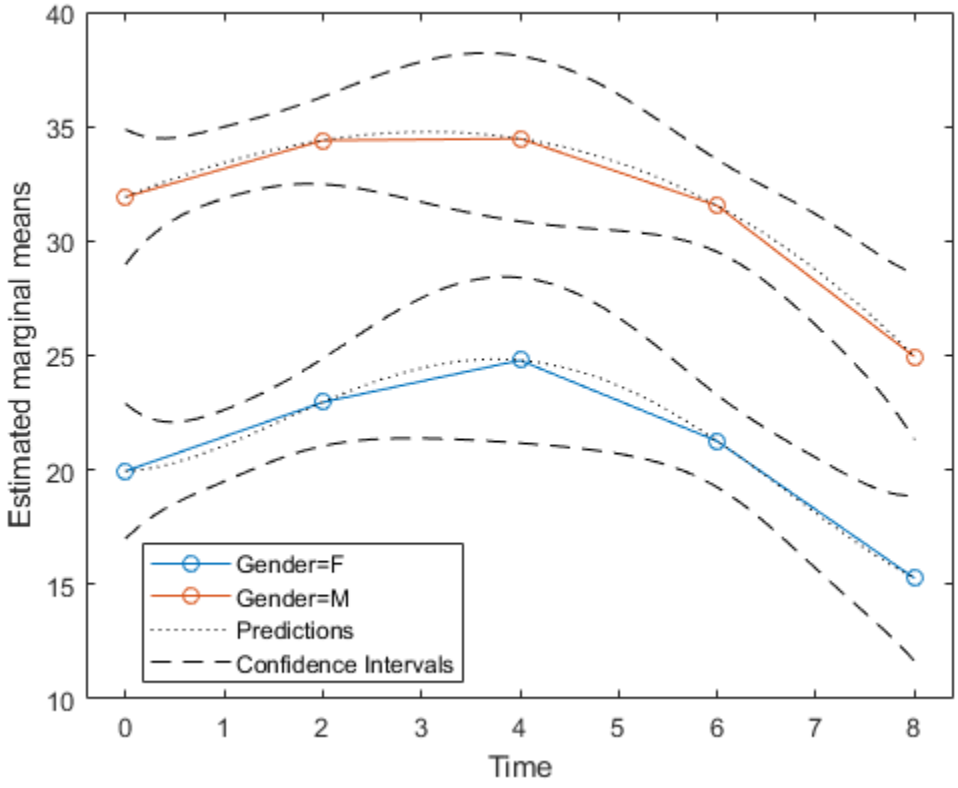
```
rm = fitrm(t,'t0-t8 ~ Gender','WithinDesign',Time);
```

Predict the responses at intermediate times.

```
time = linspace(0,8)';
[ypred,ypredci] = predict(rm,t([1 5 8 12],:), ...
    'WithinModel','orthogonalcontrasts','WithinDesign',time);
```

Plot the predictions and the confidence intervals for predictions along with the estimated marginal means.

```
p1 = plotprofile(rm,'Time','Group',{'Gender'});
hold on;
p2 = plot(time,ypred,'Color','k','LineStyle',':');
p3 = plot(time,ypredci(:,:,1),'k--');
p4 = plot(time,ypredci(:,:,2),'k--');
legend([p1;p2(1);p3(1)],'Gender=F','Gender=M','Predictions','Confidence Intervals')
hold off
```



See Also

`fitrm` | `random`

predict

Label new data using semi-supervised graph-based classifier

Syntax

```
label = predict(Mdl,X)
[label,score] = predict(Mdl,X)
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels for the data in the table or matrix `X`, based on the semi-supervised graph-based classifier `Mdl`.

`[label,score] = predict(Mdl,X)` also returns a matrix of scores indicating the likelihood that a label comes from a particular class. For each observation in `X`, the predicted class label corresponds to the maximum score among all classes.

Examples

Classify New Data Using Model Trained on Labeled and Unlabeled Data

Use both labeled and unlabeled data to train a `SemiSupervisedGraphModel` object. Label new data using the trained model.

Randomly generate 15 observations of labeled data, with 5 observations in each of three classes.

```
rng('default') % For reproducibility
labeledX = [randn(5,2)*0.25 + ones(5,2);
            randn(5,2)*0.25 - ones(5,2);
            randn(5,2)*0.5];
Y = [ones(5,1); ones(5,1)*2; ones(5,1)*3];
```

Randomly generate 300 additional observations of unlabeled data, with 100 observations per class.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
```

Fit labels to the unlabeled data by using a semi-supervised graph-based method. Specify label spreading as the labeling algorithm, and use an automatically selected kernel scale factor. The function `fitsemigraph` returns a `SemiSupervisedGraphModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemigraph(labeledX,Y,unlabeledX,'Method','labelspreading', ...
    'KernelScale','auto')
```

```
Mdl =
  SemiSupervisedGraphModel with properties:
```

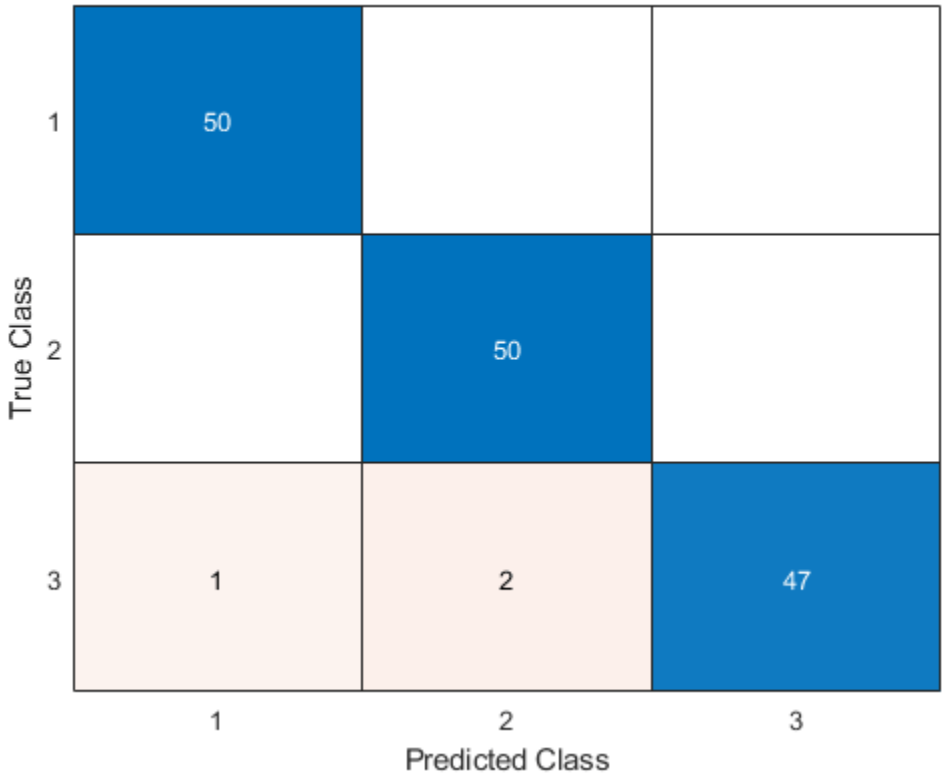
```
            FittedLabels: [300x1 double]
             LabelScores: [300x3 double]
              ClassNames: [1 2 3]
            ResponseName: 'Y'
  CategoricalPredictors: []
                  Method: 'labelspreading'


  Properties, Methods
```

Randomly generate 150 observations of new data, with 50 observations per class. For the purposes of validation, keep track of the true labels for the new data.

```
newX = [randn(50,2)*0.25 + ones(50,2);
        randn(50,2)*0.25 - ones(50,2);
        randn(50,2)*0.5];
trueLabels = [ones(50,1); ones(50,1)*2; ones(50,1)*3];
```

Predict the labels for the new data by using the `predict` function of the `SemiSupervisedGraphModel` object. Compare the true labels to the predicted labels by using a confusion matrix.

```
predictedLabels = predict(Mdl,newX);
confusionchart(trueLabels,predictedLabels)
```



Only 3 of the 150 observations in `newX` are mislabeled.

Input Arguments

`Mdl` — Semi-supervised graph-based classifier

`SemiSupervisedGraphModel` object

Semi-supervised graph-based classifier, specified as a `SemiSupervisedGraphModel` object returned by `fitsemigraph`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table. Each row of `X` corresponds to one observation, and each column corresponds to one variable.

If you trained `Mdl` using matrix data (`X` and `UnlabeledX` in the call to `fitsemigraph`), then specify `X` as a numeric matrix.

- The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
- The software treats the predictors in `X` whose indices match `Mdl.CategoricalPredictors` as categorical predictors.

If you trained `Mdl` using tabular data (`Tbl` and `UnlabeledTbl` in the call to `fitsemigraph`), then specify `X` as a table.

- All predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (for example, response variables), but `predict` ignores them.
- `predict` does not support multicolumn variables, cell arrays other than cell arrays of character vectors, or ordinal categorical variables.

If you set `'Standardize',true` in `fitsemigraph` to train `Mdl`, then the software standardizes the columns of `X` using the corresponding means and standard deviations computed on the training data.

Data Types: `single` | `double` | `table`

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors. `label` has the same data type as the fitted class labels `Mdl.FittedLabels`, and its length is equal to the number of rows in `X`.

For more information on how `predict` predicts class labels, see “Algorithms” on page 33-5001.

`score` — Predicted class scores

numeric matrix

Predicted class scores, returned as a numeric matrix. `score` has size *m*-by-*K*, where *m* is the number of observations (or rows) in `X` and *K* is the number of classes in `Mdl.ClassNames`.

`score(m,k)` is the likelihood that observation `m` in `X` belongs to class `k`, where a higher score value indicates a higher likelihood.

For more information on how `predict` predicts class scores, see “Algorithms” on page 33-5001.

More About

Similarity Graph

A similarity graph models the local neighborhood relationships between observations in the predictor data, both labeled and unlabeled, as an undirected graph. The nodes in the graph represent observations, and the edges, which are directionless, represent the connections between the observations.

If the pairwise distance $Dist_{i,j}$ between any two nodes i and j is positive (or larger than a certain threshold), then the similarity graph connects the two nodes using an edge. The edge between the two nodes is weighted by the pairwise similarity $S_{i,j}$, where $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$, for a specified kernel scale σ value.

Similarity Matrix

A similarity matrix is a matrix representation of a similarity graph on page 33-5001. The n-by-n matrix $S = (S_{i,j})_{i,j = 1,\ldots,n}$ contains pairwise similarity values between connected nodes in the similarity graph. The similarity matrix of a graph is also called an adjacency matrix.

The similarity matrix is symmetric because the edges of the similarity graph are directionless. A value of $S_{i,j} = 0$ means that nodes i and j of the similarity graph are not connected.

Algorithms

To fit labels to unlabeled training data, `fitsemigraph` constructs a similarity graph with both labeled and unlabeled observations as nodes, and distributes the label information from labeled observations to unlabeled observations by using either label propagation or label spreading. The resulting `SemiSupervisedGraphModel` object stores the fitted labels and label scores for the unlabeled data in its `FittedLabels` and `LabelScores` properties, respectively.

To predict the label of a new observation x, the `predict` function uses a weighted average of neighboring observation scores to compute the label scores for x, namely $F_x = \frac{\sum_{j=1}^{n} S(x, x_j) F_{x_j}}{\sum_{j=1}^{n} S(x, x_j)}$.

- n is the number of observations in the training data.
- F_{x_j} is the row vector of label scores for the training observation x_j (or node j). For more information on the computation of label scores for training observations, see “Algorithms” on page 33-2062.
- $S(x,x_j)$ is the pairwise similarity between the new observation x and the training observation x_j, where $S(x_i,x_j) = S_{i,j}$ is as defined in “Similarity Graph” on page 33-5001.

The column with the maximum score in F_x corresponds to the predicted class label for x. For more information, see [1].

References

[1] Delalleau, Olivier, Yoshua Bengio, and Nicolas Le Roux. “Efficient Non-Parametric Function Induction in Semi-Supervised Learning.” *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*. 2005.

See Also

`SemiSupervisedGraphModel` | `fitsemigraph`

Introduced in R2020b

predict

Label new data using semi-supervised self-trained classifier

Syntax

```
label = predict(Mdl,X)
[label,score] = predict(Mdl,X)
```

Description

`label = predict(Mdl,X)` returns a vector of predicted class labels for the data in the table or matrix `X`, based on the semi-supervised self-trained classifier `Mdl`.

`[label,score] = predict(Mdl,X)` also returns a matrix of scores indicating the likelihood that a label comes from a particular class. For each observation in `X`, the predicted class label corresponds to the maximum score among all classes.

Examples

Classify New Data Using Model Trained on Labeled and Unlabeled Data

Use both labeled and unlabeled data to train a `SemiSupervisedSelfTrainingModel` object. Label new data using the trained model.

Randomly generate 15 observations of labeled data, with 5 observations in each of three classes.

```
rng('default') % For reproducibility
labeledX = [randn(5,2)*0.25 + ones(5,2);
            randn(5,2)*0.25 - ones(5,2);
            randn(5,2)*0.5];
Y = [ones(5,1); ones(5,1)*2; ones(5,1)*3];
```

Randomly generate 300 additional observations of unlabeled data, with 100 observations per class.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
```

Fit labels to the unlabeled data by using a semi-supervised self-training method. The function `fitsemiself` returns a `SemiSupervisedSelfTrainingModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemiself(labeledX,Y,unlabeledX)
```

```
Mdl =
  SemiSupervisedSelfTrainingModel with properties:

             FittedLabels: [300x1 double]
              LabelScores: [300x3 double]
                ClassNames: [1 2 3]
```

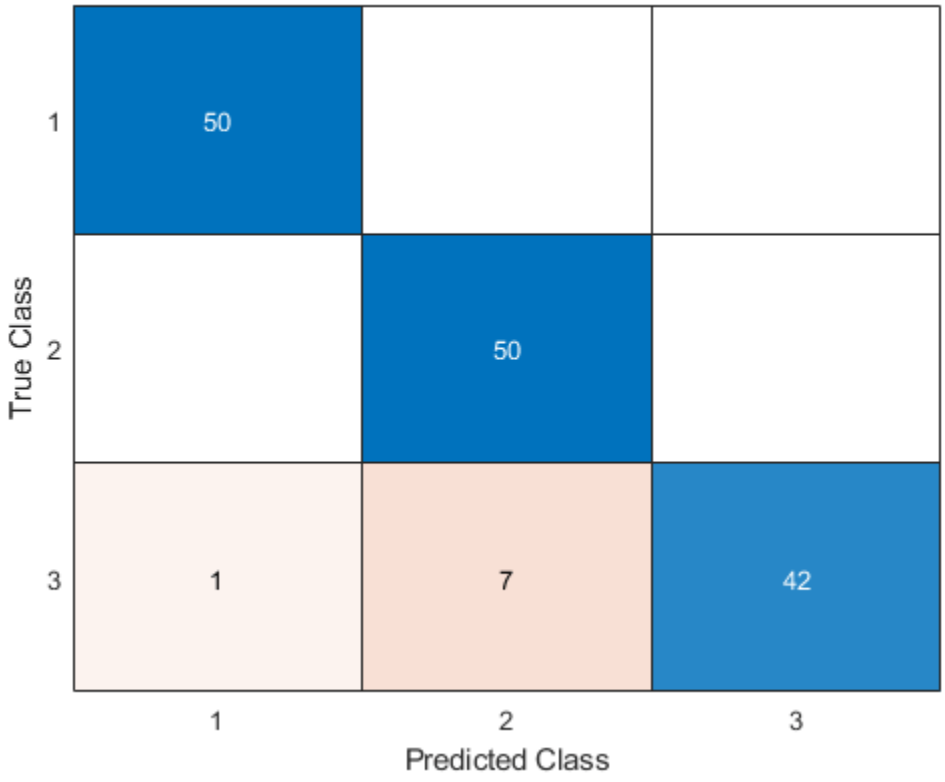
```
          ResponseName: 'Y'
CategoricalPredictors: []
              Learner: [1x1 classreg.learning.classif.CompactClassificationECOC]


  Properties, Methods
```

Randomly generate 150 observations of new data, with 50 observations per class. For the purposes of validation, keep track of the true labels for the new data.

```
newX = [randn(50,2)*0.25 + ones(50,2);
        randn(50,2)*0.25 - ones(50,2);
        randn(50,2)*0.5];
trueLabels = [ones(50,1); ones(50,1)*2; ones(50,1)*3];
```

Predict the labels for the new data by using the `predict` function of the `SemiSupervisedSelfTrainingModel` object. Compare the true labels to the predicted labels by using a confusion matrix.

```
predictedLabels = predict(Mdl,newX);
confusionchart(trueLabels,predictedLabels)
```



Only 8 of the 150 observations in `newX` are mislabeled.

Input Arguments

`Mdl` — Semi-supervised self-training classifier

`SemiSupervisedSelfTrainingModel` object

Semi-supervised self-training classifier, specified as a `SemiSupervisedSelfTrainingModel` object returned by `fitsemiself`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table. Each row of `X` corresponds to one observation, and each column corresponds to one variable.

If you trained `Mdl` using matrix data (`X` and `UnlabeledX` in the call to `fitsemiself`), then specify `X` as a numeric matrix.

- The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
- The software treats the predictors in `X` whose indices match `Mdl.CategoricalPredictors` as categorical predictors.

If you trained `Mdl` using tabular data (`Tbl` and `UnlabeledTbl` in the call to `fitsemiself`), then specify `X` as a table.

- All predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (for example, response variables), but `predict` ignores them.
- `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

Data Types: `single` | `double` | `table`

Output Arguments

`label` — Predicted class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors. `label` has the same data type as the fitted class labels `Mdl.FittedLabels`, and its length is equal to the number of rows in `X`.

`score` — Predicted class scores

numeric matrix

Predicted class scores, returned as a numeric matrix. `score` has size *m*-by-*K*, where *m* is the number of observations (or rows) in `X` and *K* is the number of classes in `Mdl.ClassNames`.

`score(m,k)` is the likelihood that observation `m` in `X` belongs to class `k`, where a higher score value indicates a higher likelihood. The range of score values depends on the underlying classifier `Mdl.Learner`.

See Also

`SemiSupervisedSelfTrainingModel` | `fitsemiself`

Introduced in R2020b

predict

Class: `TreeBagger`

Predict responses using ensemble of bagged decision trees

Syntax

```
Yfit = predict(B,X)
Yfit = predict(B,X,Name,Value)
[Yfit,stdevs] = predict( ___ )
[Yfit,scores] = predict( ___ )
[Yfit,scores,stdevs] = predict( ___ )
```

Description

`Yfit = predict(B,X)` returns a vector of predicted responses for the predictor data in the table or matrix `X`, based on the ensemble of bagged decision trees `B`. `Yfit` is a cell array of character vectors for classification and a numeric array for regression. By default, `predict` takes a democratic (nonweighted) average vote from all trees in the ensemble.

`B` is a trained `TreeBagger` model object, that is, a model returned by `TreeBagger`.

`X` is a table or matrix of predictor data used to generate responses. Rows represent observations and columns represent variables.

- If `X` is a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `B`.
  - If you trained `B` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `TreeBagger`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- If `X` is a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `B` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and be of the same data types as those that trained `B` (stored in `B.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.
  - If you trained `B` using a numeric matrix, then the predictor names in `B.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `TreeBagger`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `predict` ignores them.

`Yfit = predict(B,X,Name,Value)` specifies additional options using one or more name-value pair arguments:

- `'Trees'` — Array of tree indices to use for computation of responses. The default is `'all'`.
- `'TreeWeights'` — Array of `NTrees` weights for weighting votes from the specified trees, where `NTrees` is the number of trees in the ensemble.
- `'UseInstanceForTree'` — Logical matrix of size `Nobs`-by-`NTrees` indicating which trees to use to make predictions for each observation, where `Nobs` is the number of observations. By default all trees are used for all observations.

For regression, `[Yfit,stdevs] = predict( ___ )` also returns standard deviations of the computed responses over the ensemble of the grown trees using any of the input argument combinations in previous syntaxes.

For classification, `[Yfit,scores] = predict( ___ )` also returns scores for all classes. `scores` is a matrix with one row per observation and one column per class. For each observation and each class, the score generated by each tree is the probability of the observation originating from the class, computed as the fraction of observations of the class in a tree leaf. `predict` averages these scores over all trees in the ensemble.

`[Yfit,scores,stdevs] = predict( ___ )` also returns standard deviations of the computed scores for classification. `stdevs` is a matrix with one row per observation and one column per class, with standard deviations taken over the ensemble of the grown trees.

Algorithms

- For regression problems, the predicted response for an observation is the weighted average of the predictions using selected trees only. That is,

$$\widehat{y}_{\text{bag}} = \frac{1}{\sum_{t=1}^{T} \alpha_t I(t \in S)} \sum_{t=1}^{T} \alpha_t \widehat{y}_t I(t \in S) .$$

  - $\widehat{y}_t$ is the prediction from tree t in the ensemble.
  - S is the set of indices of selected trees that comprise the prediction (see `'Trees'` and `'UseInstanceForTree'`). $I(t \in S)$ is 1 if t is in the set S, and 0 otherwise.
  - α_t is the weight of tree t (see `'TreeWeights'`).
- For classification problems, the predicted class for an observation is the class that yields the largest weighted average of the class posterior probabilities (i.e., classification scores) computed using selected trees only. That is,

  1. For each class $c \in C$ and each tree $t = 1,...,T$, `predict` computes $\widehat{P}_t(c|x)$, which is the estimated posterior probability of class c given observation x using tree t. C is the set of all distinct classes in the training data. For more details on classification tree posterior probabilities, see `fitctree` and `predict`.
  2. `predict` computes the weighted average of the class posterior probabilities over the selected trees.

$$\widehat{P}_{\text{bag}}(c|x) = \frac{1}{\sum_{t=1}^{T} \alpha_t I(t \in S)} \sum_{t=1}^{T} \alpha_t \widehat{P}_t(c|x) I(t \in S) .$$

3 The predicted class is the class that yields the largest weighted average.

$$\widehat{y}_{\text{bag}} = \underset{c \in C}{\operatorname{argmax}} \left\{ \widehat{P}_{\text{bag}}(c \mid x) \right\}.$$

See Also

`TreeBagger` | `error` | `oobPredict` | `predict` | `quantilePredict`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124

predictConstraints

Predict coupled constraint violations at a set of points

Syntax

```
ConstraintViolations = predictConstraints(results,XTable)
[ConstraintViolations,sigma] = predictConstraints(results,XTable)
```

Description

`ConstraintViolations = predictConstraints(results,XTable)` returns the coupled constraint function violations at the points in `XTable`.

`[ConstraintViolations,sigma] = predictConstraints(results,XTable)` also returns the standard deviations of the coupled constraint functions.

Examples

Predict Coupled Constraints

This example shows how to predict the coupled constraints of an optimized SVM model. For details of this model, see “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
redpts = zeros(100,2);
grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
cdata = [grnpts;redpts];
grp = ones(200,1);
grp(101:200) = -1;
c = cvpartition(200,'KFold',10);
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
```

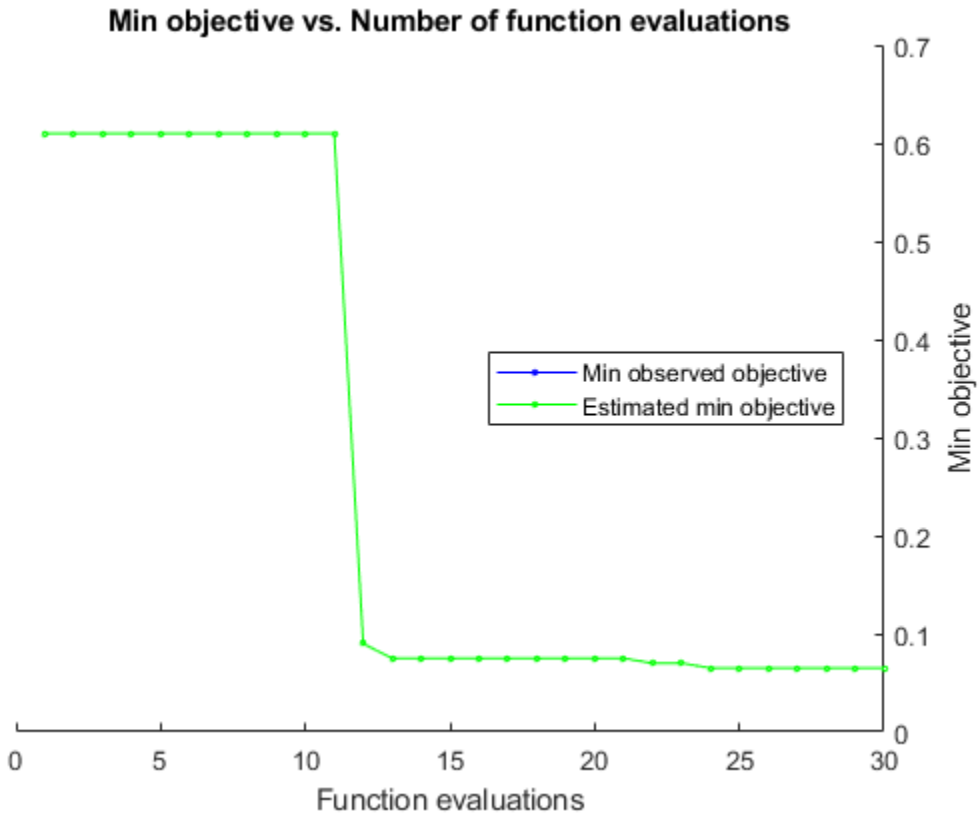
The objective function is the cross-validation loss of the SVM model for the partition `c`. The coupled constraint is the number of support vectors in the model minus 100. The model has 200 data points, so the coupled constraint values range from -100 to 100. Positive values mean the constraint is not satisfied.

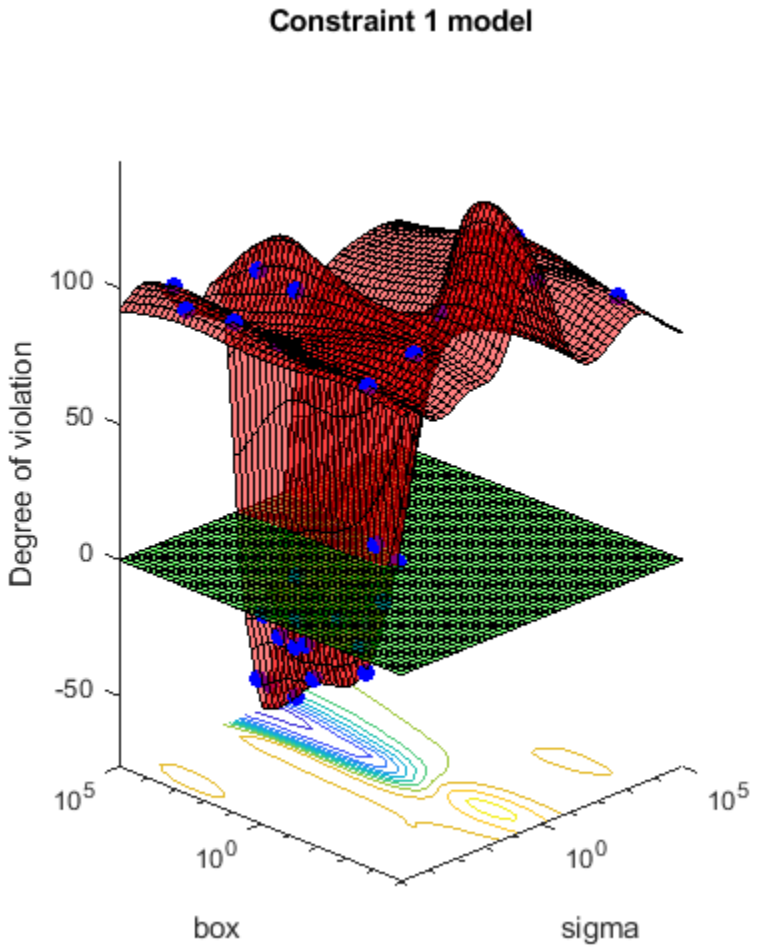
```
function [objective,constraint] = mysvmfun(x,cdata,grp,c)
SVMModel = fitcsvm(cdata,grp,'KernelFunction','rbf',...
    'BoxConstraint',x.box,...
    'KernelScale',x.sigma);
cvModel = crossval(SVMModel,'CVPartition',c);
```

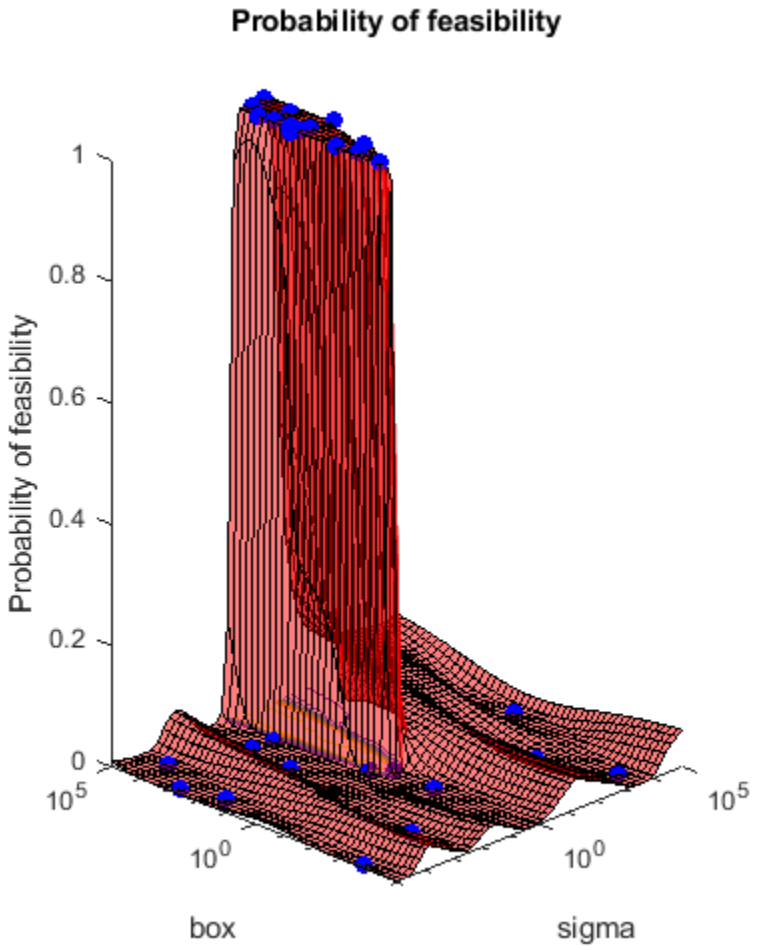
```
objective = kfoldLoss(cvModel);
constraint = sum(SVMModel.IsSupportVector)-100.5;
```

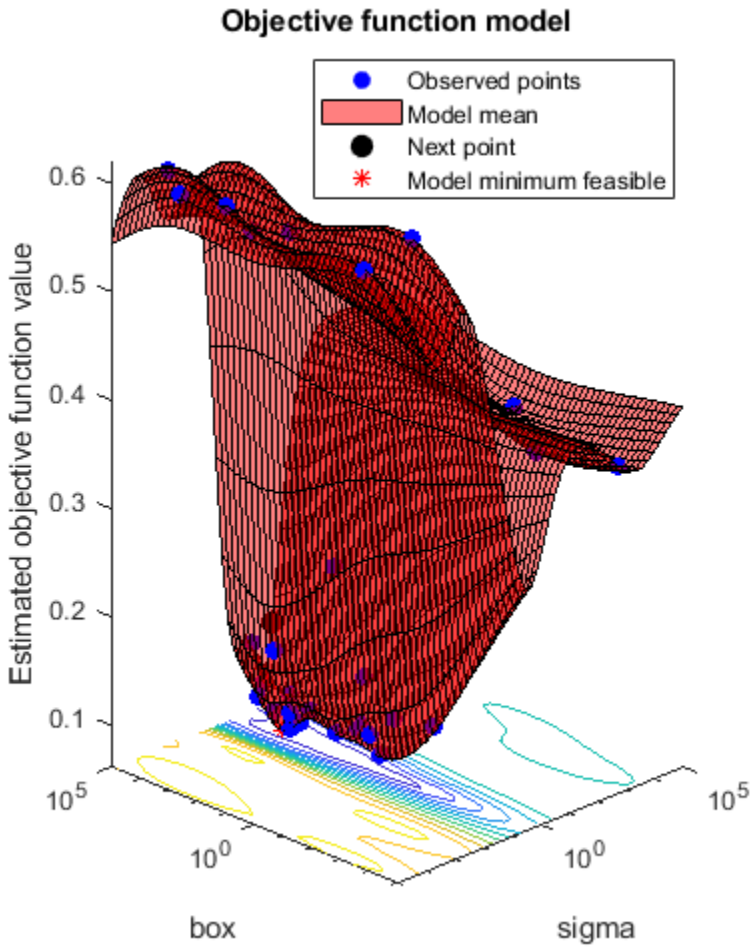
Call the optimizer using this function and its one coupled constraint.

```
fun = @(x)mysvmfun(x,cdata,grp,c);
results = bayesopt(fun,[sigma,box],'IsObjectiveDeterministic',true,...
    'NumCoupledConstraints',1,'PlotFcn',...
    {@plotMinObjective,@plotConstraintModels,@plotObjectiveModel},...
    'AcquisitionFunctionName','expected-improvement-plus','Verbose',0);
```




Constraint 1 model
Degree of violation
100
50
0
-50
10^5
10^0
box
10^5
10^0
sigma


Probability of feasibility
Probability of feasibility
1
0.8
0.6
0.4
0.2
0
10^5
10^0
box
10^0
10^5
sigma



The constraint model plot shows that most parameters in the range are infeasible, and are feasible only for relatively high values of the `box` parameter and a small range of the `sigma` parameter. Predict the coupled constraint values for several values of the control variables `box` and `sigma`.

```
sigma = logspace(-2,2,11)';
box = logspace(0,5,11)';
XTable = table(sigma,box);
cons = predictConstraints(results,XTable);
[XTable,table(cons)]
```

```
ans =

  11x3 table

    sigma       box       cons
    _____    ______    _______

     0.01         1     99.443
 0.025119    3.1623     106.49
 0.063096        10     94.468
```

```
    0.15849    31.623     25.134
    0.39811       100    -38.732
          1    316.23    -55.156
     2.5119      1000    -34.181
     6.3096    3162.3     5.0153
     15.849     10000     39.465
     39.811     31623       60.9
        100     1e+05     71.906
```

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

`XTable` — Prediction points

table with D columns

Prediction points, specified as a table with D columns, where D is the number of variables in the problem. The function performs its predictions on these points.

Data Types: `table`

Output Arguments

`ConstraintViolations` — Constraint violations

N-by-K matrix

Constraint violations, returned as an N-by-K matrix, where there are `N` rows in `XTable` and `K` coupled constraints. The constraint violations are the posterior means of the Gaussian process model of the coupled constraints at the points in `XTable`.

`sigma` — Constraint standard deviations

N-by-K matrix

Constraint standard deviations, returned as an N-by-K matrix, where there are `N` rows in `XTable` and `K` coupled constraints. The standard deviations represent those of the posterior distribution at the points in `XTable`.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

predictError

Predict error value at a set of points

Syntax

```
error = predictError(results,XTable)
[error,sigma] = predictError(results,XTable)
```

Description

`error = predictError(results,XTable)` returns the posterior mean of the error coupled constraint at the points in `XTable`.

`[error,sigma] = predictError(results,XTable)` also returns the posterior standard deviations.

Examples

Error Prediction

This example shows optimizing a function that throws an error when the evaluation point has norm larger than 2. The error model for the objective function learns this behavior.

Create variables named `x1` and `x2` that range from `-5` to `5`.

```
var1 = optimizableVariable('x1',[-5,5]);
var2 = optimizableVariable('x2',[-5,5]);
vars = [var1,var2];
```

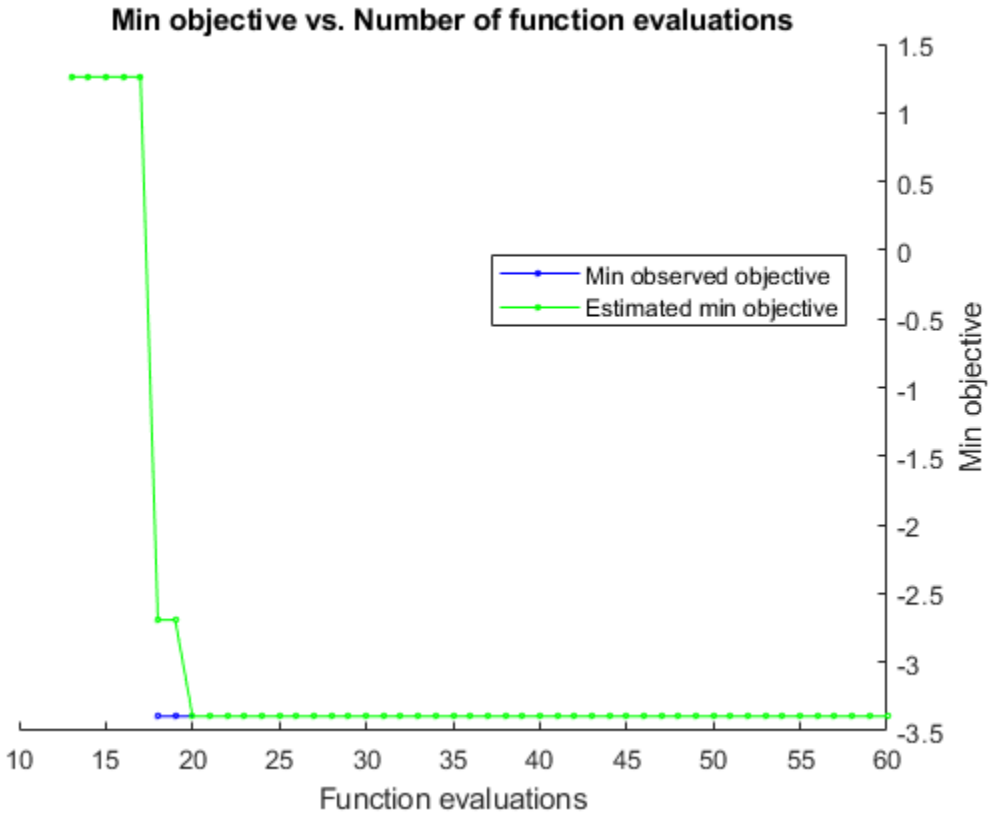
The following objective function throws an error when the norm of `x = [x1,x2]` exceeds 2:

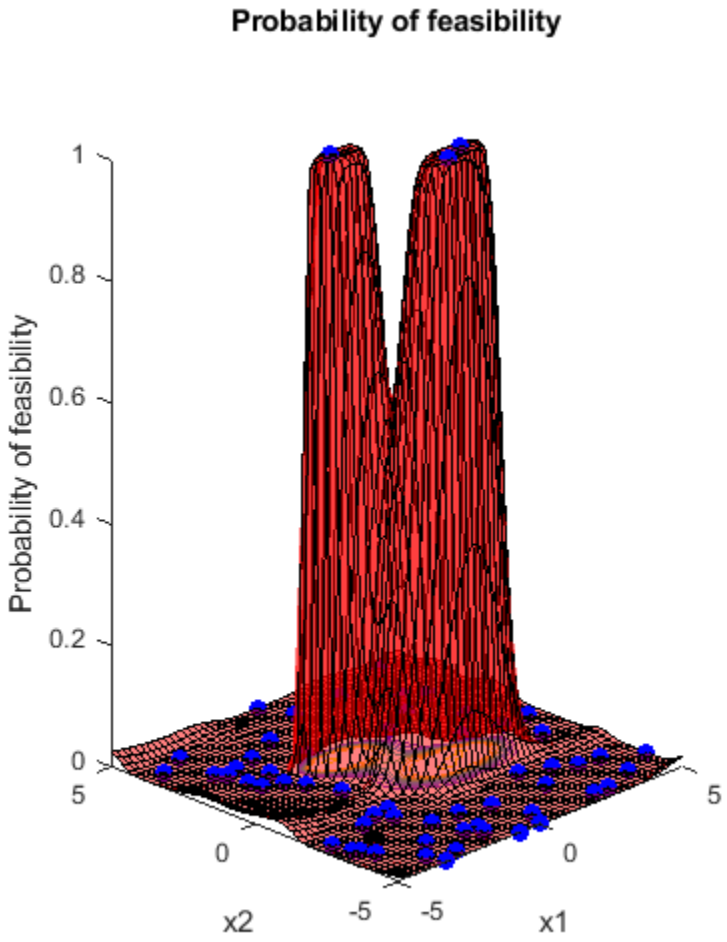
```
function f = makeanerror(x)
f = x.x1 - x.x2 - sqrt(4-x.x1^2-x.x2^2);
```

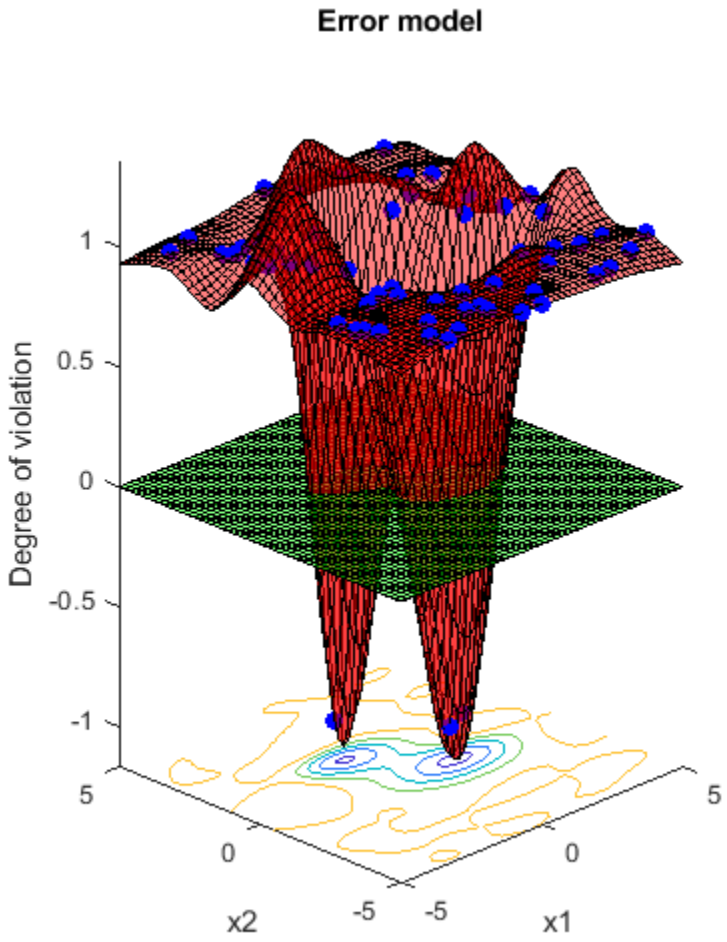
```
fun = @makeanerror;
```

Plot the error model and minimum objective as the optimization proceeds. Optimize for 60 iterations so the error model becomes well-trained. For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng default
results = bayesopt(fun,vars,'Verbose',0,'MaxObjectiveEvaluations',60,...
    'AcquisitionFunctionName','expected-improvement-plus',...
    'PlotFcn',{@plotMinObjective,@plotConstraintModels});
```


Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Min objective
Function evaluations
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
10
15
20
25
30
35
40
45
50
55
60


Probability of feasibility
Probability of feasibility
1
0.8
0.6
0.4
0.2
0
5
0
-5
x2
-5
0
5
x1



Predict the error at points on the line `x1 = x2`. If the error model were perfect, it would have value `-1` at every point where the norm of `x` is no more than 2, and value 1 at all other points.

```
x1 = (-5:0.5:5)';
x2 = x1;
XTable = table(x1,x2);
error = predictError(results,XTable);
normx = sqrt(x1.^2 + x2.^2);
[XTable,table(normx,error)]
```

```
ans =

  21x4 table

    x1      x2      normx       error
    ____    ____    _______    _________

     -5      -5     7.0711      0.94663
   -4.5    -4.5      6.364      0.97396
     -4      -4     5.6569      0.99125
```

```
    -3.5    -3.5     4.9497       1.0033
      -3      -3     4.2426       1.0018
    -2.5    -2.5     3.5355      0.99627
      -2      -2     2.8284       1.0043
    -1.5    -1.5     2.1213      0.89886
      -1      -1     1.4142       0.4746
    -0.5    -0.5    0.70711    0.0042389
       0       0          0     -0.16004
     0.5     0.5    0.70711    -0.012397
       1       1     1.4142      0.30187
     1.5     1.5     2.1213      0.88588
       2       2     2.8284       1.0872
     2.5     2.5     3.5355        0.997
       3       3     4.2426      0.99861
     3.5     3.5     4.9497      0.98894
       4       4     5.6569      0.98941
     4.5     4.5      6.364      0.98956
       5       5     7.0711      0.95549
```

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

`XTable` — Prediction points

table with D columns

Prediction points, specified as a table with D columns, where D is the number of variables in the problem. The function performs its predictions on these points.

Data Types: `table`

Output Arguments

`error` — Mean of error coupled constraint

N-by-1 vector

Mean of error coupled constraint, returned as an N-by-1 vector, where `N` is the number of rows of `XTable`. The mean is the posterior mean of the error coupled constraint at the points in `XTable`.

`bayesopt` deems your objective function to return an error if it returns anything other than a finite real scalar. See “Objective Function Errors” on page 10-36.

`sigma` — Standard deviation of error coupled constraint

N-by-1 vector

Standard deviation of error coupled constraint, returned as an N-by-1 vector, where `N` is the number of rows of `XTable`.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

predictObjective

Predict objective function at a set of points

Syntax

```
objective = predictObjective(results,XTable)
[objective,sigma] = predictObjective(results,XTable)
```

Description

`objective = predictObjective(results,XTable)` returns the estimated objective function value at the points in `XTable`.

`[objective,sigma] = predictObjective(results,XTable)` also returns estimated standard deviations.

Examples

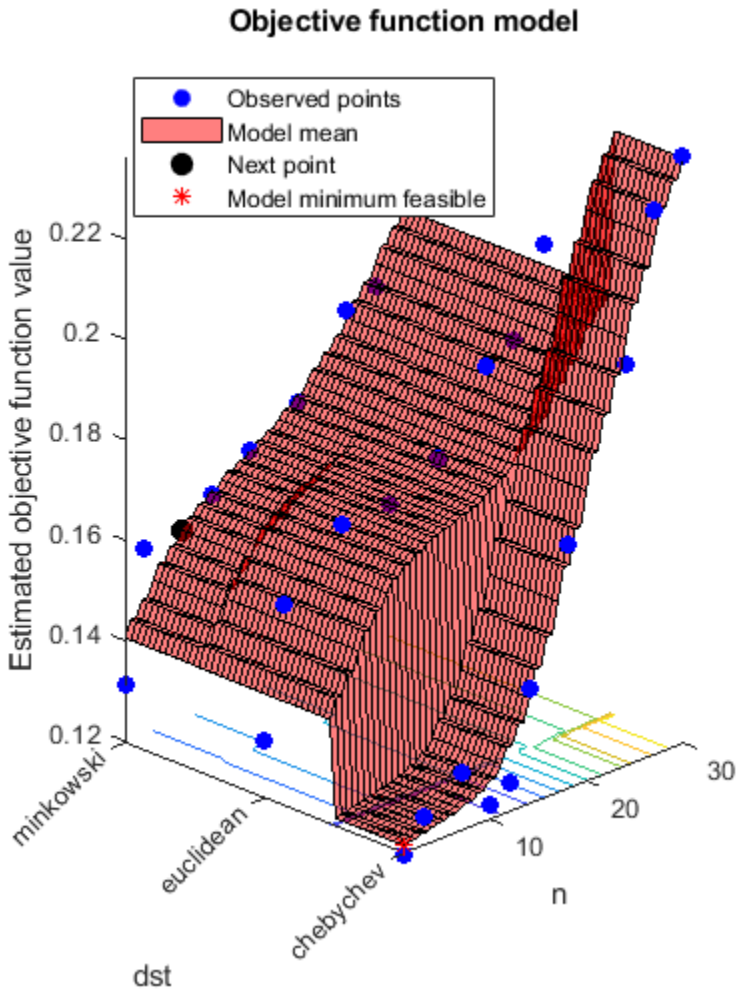
Predict Cross-Validation Loss of an Optimized Classifier

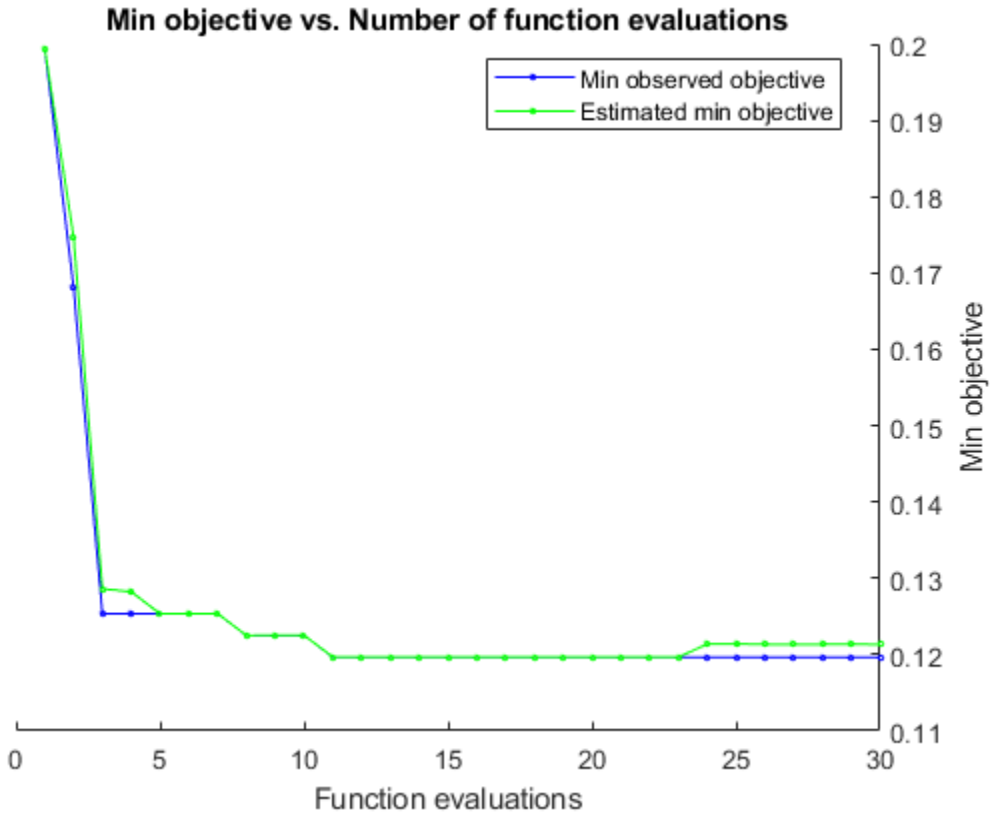
This example shows how to estimate the cross-validation loss of an optimized classifier.

Optimize a KNN classifier for the `ionosphere` data, meaning find parameters that minimize the cross-validation loss. Minimize over nearest-neighborhood sizes from 1 to 30, and over the distance functions `'chebychev'`, `'euclidean'`, and `'minkowski'`.

For reproducibility, set the random seed, and set the `AcquisitionFunctionName` option to `'expected-improvement-plus'`.

```
load ionosphere
rng default
num = optimizableVariable('n',[1,30],'Type','integer');
dst = optimizableVariable('dst',{'chebychev','euclidean','minkowski'},'Type','categorical');
c = cvpartition(351,'Kfold',5);
fun = @(x)kfoldLoss(fitcknn(X,Y,'CVPartition',c,'NumNeighbors',x.n,...
    'Distance',char(x.dst),'NSMethod','exhaustive'));
results = bayesopt(fun,[num,dst],'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus');
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.22
0.2
0.18
0.16
0.14
0.12
minkowski
euclidean
chebychev
dst
10
20
30
n



Create a table of points to estimate.

```
b = categorical({'chebychev','euclidean','minkowski'});
n = [1;1;1;4;2;2];
dst = [b(1);b(2);b(3);b(1);b(1);b(3)];
XTable = table(n,dst);
```

Estimate the objective and standard deviation of the objective at these points.

```
[objective,sigma] = predictObjective(results,XTable);
[XTable,table(objective,sigma)]
```

```
ans=6×4 table
    n       dst       objective      sigma
    _    _________    _________    _________

    1    chebychev     0.12132     0.0068029
    1    euclidean     0.14052     0.0079128
    1    minkowski     0.14057     0.0079117
    4    chebychev      0.1227     0.0068805
    2    chebychev     0.12176     0.0066739
    2    minkowski      0.1437     0.0075448
```

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

`XTable` — Prediction points

table with D columns

Prediction points, specified as a table with D columns, where D is the number of variables in the problem. The function performs its predictions on these points.

Data Types: `table`

Output Arguments

`objective` — Objective estimates

N-by-1 vector

Objective estimates, returned as an N-by-1 vector, where `N` is the number of rows of `XTable`. The estimates are the mean values of the posterior distribution of the Gaussian process model of the objective function.

`sigma` — Standard deviations of objective function

N-by-1 vector

Standard deviations of objective function, returned as an N-by-1 vector, where `N` is the number of rows of `XTable`. The standard deviations are those of the posterior distribution of the Gaussian process model of the objective function.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

predictObjectiveEvaluationTime

Predict objective function run times at a set of points

Syntax

```
time = predictObjectiveEvaluationTime(results,XTable)
```

Description

`time = predictObjectiveEvaluationTime(results,XTable)` returns estimated objective evaluation times at the points in `XTable`.

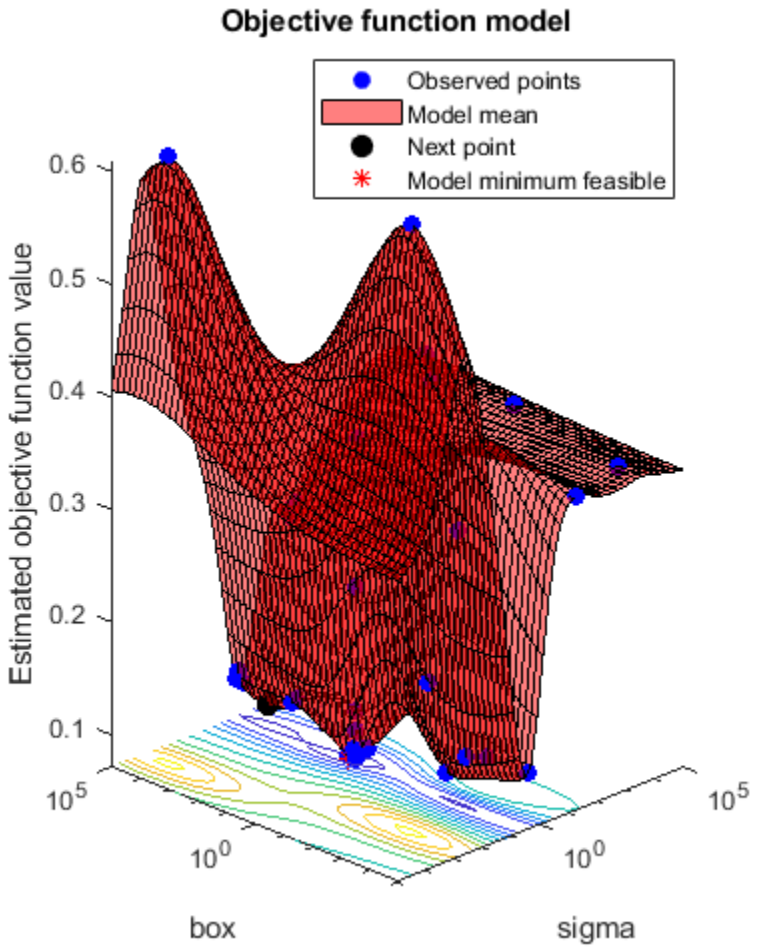
Examples

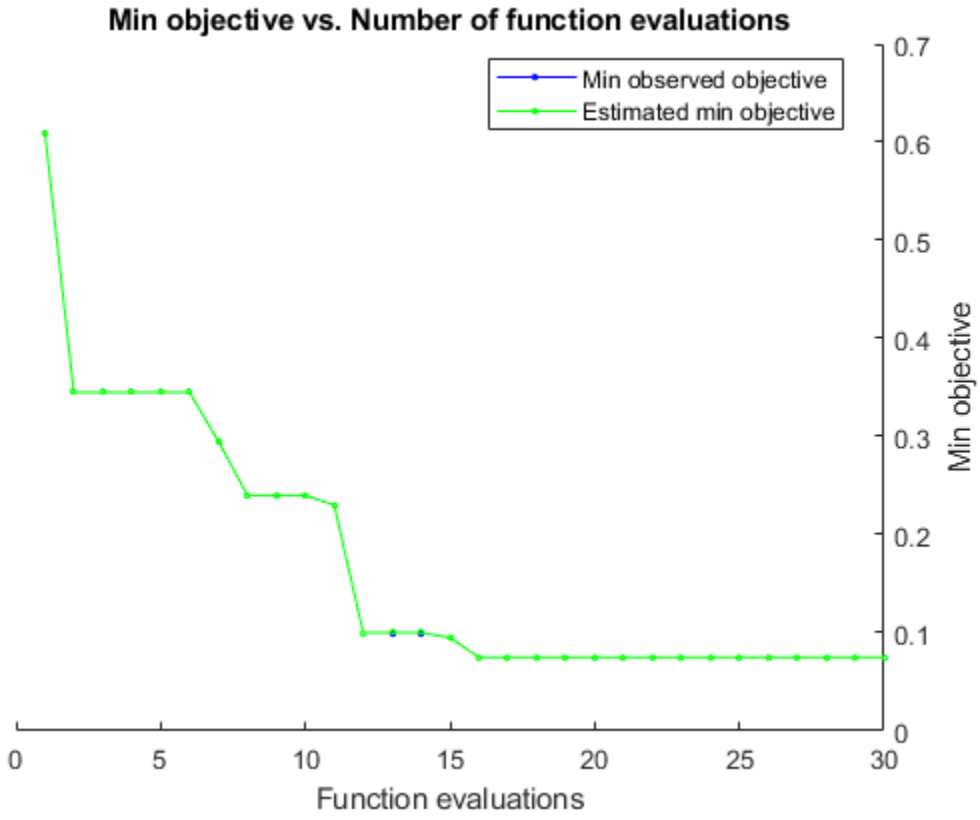
Predict Evaluation Time of Objective In an Optimized Model

This example shows how to estimate the objective function evaluation time in an optimized Bayesian model of SVM classification.

Create an optimized SVM model. For details of this model, see “Optimize a Cross-Validated SVM Classifier Using bayesopt” on page 10-45.

```
rng default
grnpop = mvnrnd([1,0],eye(2),10);
redpop = mvnrnd([0,1],eye(2),10);
redpts = zeros(100,2);
grnpts = redpts;
for i = 1:100
    grnpts(i,:) = mvnrnd(grnpop(randi(10),:),eye(2)*0.02);
    redpts(i,:) = mvnrnd(redpop(randi(10),:),eye(2)*0.02);
end
cdata = [grnpts;redpts];
grp = ones(200,1);
grp(101:200) = -1;
c = cvpartition(200,'KFold',10);
sigma = optimizableVariable('sigma',[1e-5,1e5],'Transform','log');
box = optimizableVariable('box',[1e-5,1e5],'Transform','log');
minfn = @(z)kfoldLoss(fitcsvm(cdata,grp,'CVPartition',c,...
    'KernelFunction','rbf','BoxConstraint',z.box,...
    'KernelScale',z.sigma));
results = bayesopt(minfn,[sigma,box],'IsObjectiveDeterministic',true,...
    'AcquisitionFunctionName','expected-improvement-plus','Verbose',0);
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
Estimated objective function value
0.6
0.5
0.4
0.3
0.2
0.1
10^5
10^0
box
10^5
10^0
sigma



Predict the evaluation time for various points.

```
sigma = logspace(-5,5,11)';
box = 1e5*ones(size(sigma));
XTable = table(sigma,box);
time = predictObjectiveEvaluationTime(results,XTable);
[XTable,table(time)]
```

```
ans=11×3 table
    sigma      box      time
    _____    _____    _______

     1e-05    1e+05    0.35146
    0.0001    1e+05    0.35726
     0.001    1e+05    0.34691
      0.01    1e+05    0.33154
       0.1    1e+05    0.36187
         1    1e+05    0.72805
        10    1e+05     2.5234
       100    1e+05     1.1003
      1000    1e+05    0.29006
     10000    1e+05    0.26345
     1e+05    1e+05    0.26789
```

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

`XTable` — Prediction points

table with D columns

Prediction points, specified as a table with D columns, where D is the number of variables in the problem. The function performs its predictions on these points.

Data Types: `table`

Output Arguments

`time` — Estimated objective evaluation times

N-by-1 vector

Estimated objective evaluation times, returned as an `N`-by-`1` vector, where `N` is the number of rows of `XTable`. The estimated values are the means of the posterior distribution of the Gaussian process model of the evaluation times of the objective function.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

predictorImportance

Estimates of predictor importance for classification ensemble of decision trees

Syntax

```
imp = predictorImportance(ens)
[imp,ma] = predictorImportance(ens)
```

Description

`imp = predictorImportance(ens)` computes estimates of predictor importance for `ens` by summing these estimates over all weak learners in the ensemble. `imp` has one element for each input predictor in the data used to train this ensemble. A high value indicates that this predictor is important for `ens`.

`[imp,ma] = predictorImportance(ens)` returns a `P`-by-`P` matrix with predictive measures of association for `P` predictors, when the learners in `ens` contain surrogate splits. See “More About” on page 33-5032.

Input Arguments

`ens`

A classification ensemble of decision trees, created by `fitcensemble`, or by the `compact` method.

Output Arguments

`imp`

A row vector with the same number of elements as the number of predictors (columns) in `ens.X`. The entries are the estimates of predictor importance, with `0` representing the smallest possible importance.

`ma`

A `P`-by-`P` matrix of predictive measures of association for `P` predictors. Element `ma(I,J)` is the predictive measure of association averaged over surrogate splits on predictor `J` for which predictor `I` is the optimal split predictor. `predictorImportance` averages this predictive measure of association over all trees in the ensemble.

Examples

Estimate Predictor Importance

Estimate the predictor importance for all variables in the Fisher iris data.

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification ensemble using AdaBoostM2. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Estimate the predictor importance for all predictor variables.

```
imp = predictorImportance(ens)
```

```
imp = 1×4

    0.0004    0.0016    0.1266    0.0324
```

The first two predictors are not very important in the ensemble.

Predictor Importance and Surrogate Splits

Estimate the predictor importance for all variables in the Fisher iris data for an ensemble where the trees contain surrogate splits.

Load Fisher's iris data set.

```
load fisheriris
```

Grow an ensemble of 100 classification trees using AdaBoostM2. Specify tree stumps as the weak learners, and also identify surrogate splits.

```
t = templateTree('MaxNumSplits',1,'Surrogate','on');
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Estimate the predictor importance and predictive measures of association for all predictor variables.

```
[imp,ma] = predictorImportance(ens)
```

```
imp = 1×4

    0.0674    0.0417    0.1582    0.1537


ma = 4×4

    1.0000         0         0         0
    0.0115    1.0000    0.0022    0.0054
    0.3186    0.2137    1.0000    0.6391
    0.0392    0.0073    0.1137    1.0000
```

The first two predictors show much more importance than the analysis in “Estimate Predictor Importance” on page 33-5030.

More About

Predictor Importance

`predictorImportance` estimates predictor importance for each tree learner in the ensemble `ens` and returns the weighted average `imp` computed using `ens.TrainedWeight`. The output `imp` has one element for each predictor.

`predictorImportance` computes importance measures of the predictors in a tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. For example, if a tree splits a parent node (for example, node 1) into two child nodes (for example, nodes 2 and 3), then `predictorImportance` increases the importance of the split predictor by

$$(R_1 - R_2 - R_3)/N_{\text{branch}},$$

where R_i is the node risk of node i, and N_{branch} is the total number of branch nodes. A node risk is defined as a node error or node impurity weighted by the node probability:

$$R_i = P_i E_i,$$

where P_i is the node probability of node i, and E_i is either the node error (for a tree grown by minimizing the twoing criterion) or node impurity (for a tree grown by minimizing an impurity criterion, such as the Gini index or deviance) of node i.

The estimates of predictor importance depend on whether you use surrogate splits for training.

- If you use surrogate splits, `predictorImportance` sums the changes in the node risk over all splits at each branch node, including surrogate splits. If you do not use surrogate splits, then the function takes the sum over the best splits found at each branch node.
- Estimates of predictor importance do not depend on the order of predictors if you use surrogate splits, but do depend on the order if you do not use surrogate splits.

Impurity and Node Error

A decision tree splits nodes based on either impurity or node error.

Impurity means one of several things, depending on your choice of the `SplitCriterion` name-value pair argument:

- Gini's Diversity Index (`gdi`) — The Gini index of a node is

  $$1 - \sum_i p^2(i),$$

  where the sum is over the classes i at the node, and $p(i)$ is the observed fraction of classes with class i that reach the node. A node with just one class (a pure node) has Gini index `0`; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.
- Deviance (`'deviance'`) — With $p(i)$ defined the same as for the Gini index, the deviance of a node is

  $$-\sum_i p(i)\log_2 p(i)\,.$$

  A pure node has deviance `0`; otherwise, the deviance is positive.
- Twoing rule (`'twoing'`) — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class i in the left child

node after a split, and $R(i)$ denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

$$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2,$$

where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.

- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $1 – p(j)$.

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_jL_k} - P_{R_jR_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_jL_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.
- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in (–∞,1]. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Algorithms

Element `ma(i,j)` is the predictive measure of association averaged over surrogate splits on predictor `j` for which predictor `i` is the optimal split predictor. This average is computed by summing positive values of the predictive measure of association over optimal splits on predictor `i` and surrogate splits on predictor `j` and dividing by the total number of optimal splits on predictor `i`, including splits for which the predictive measure of association between predictors `i` and `j` is negative.

See Also

`predictorImportance (ClassificationTree)` | `templateTree`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

predictorImportance

Estimates of predictor importance for classification tree

Syntax

```
imp = predictorImportance(tree)
```

Description

`imp = predictorImportance(tree)` computes estimates of predictor importance for `tree` by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes.

Input Arguments

`tree`

A classification tree created by `fitctree`, or by the `compact` method.

Output Arguments

`imp`

A row vector with the same number of elements as the number of predictors (columns) in `tree.X`. The entries are the estimates of predictor importance, with `0` representing the smallest possible importance.

Examples

Estimate Predictor Importance Values

Load Fisher's iris data set.

```
load fisheriris
```

Grow a classification tree.

```
Mdl = fitctree(meas,species);
```

Compute predictor importance estimates for all predictor variables.

```
imp = predictorImportance(Mdl)
```

```
imp = 1×4

         0         0    0.0907    0.0682
```

The first two elements of `imp` are zero. Therefore, the first two predictors do not enter into `Mdl` calculations for classifying irises.

Estimates of predictor importance do not depend on the order of predictors if you use surrogate splits, but do depend on the order if you do not use surrogate splits.

Permute the order of the data columns in the previous example, grow another classification tree, and then compute predictor importance estimates.

```
measPerm  = meas(:,[4 1 3 2]);
MdlPerm = fitctree(measPerm,species);
impPerm = predictorImportance(MdlPerm)
```

```
impPerm = 1×4

    0.1515         0    0.0074         0
```

The estimates of predictor importance are not a permutation of `imp`.

Surrogate Splits and Predictor Importance

Load Fisher's iris data set.

```
load fisheriris
```

Grow a classification tree. Specify usage of surrogate splits.

```
Mdl = fitctree(meas,species,'Surrogate','on');
```

Compute predictor importance estimates for all predictor variables.

```
imp = predictorImportance(Mdl)
```

```
imp = 1×4

    0.0791    0.0374    0.1530    0.1529
```

All predictors have some importance. The first two predictors are less important than the final two.

Permute the order of the data columns in the previous example, grow another classification tree specifying usage of surrogate splits, and then compute predictor importance estimates.

```
measPerm  = meas(:,[4 1 3 2]);
MdlPerm = fitctree(measPerm,species,'Surrogate','on');
impPerm = predictorImportance(MdlPerm)
```

```
impPerm = 1×4

    0.1529    0.0791    0.1530    0.0374
```

The estimates of predictor importance are a permutation of `imp`.

Unbiased Predictor Importance Estimates

Load the `census1994` data set. Consider a model that predicts a person's salary category given their age, working class, education level, martial status, race, sex, capital gain and loss, and number of working hours per week.

```
load census1994
X = adultdata(:,{'age','workClass','education_num','marital_status','race',...
    'sex','capital_gain','capital_loss','hours_per_week','salary'});
```

Display the number of categories represented in the categorical variables using `summary`.

```
summary(X)
```

```
Variables:

    age: 32561x1 double

        Values:

            Min          17
            Median       37
            Max          90

    workClass: 32561x1 categorical

        Values:

            Federal-gov            960
            Local-gov             2093
            Never-worked             7
            Private              22696
            Self-emp-inc          1116
            Self-emp-not-inc      2541
            State-gov             1298
            Without-pay             14
            NumMissing            1836

    education_num: 32561x1 double

        Values:

            Min           1
            Median       10
            Max          16

    marital_status: 32561x1 categorical

        Values:

            Divorced                  4443
            Married-AF-spouse           23
            Married-civ-spouse        14976
            Married-spouse-absent       418
            Never-married             10683
            Separated                  1025
            Widowed                     993
```

```
race: 32561x1 categorical

    Values:

        Amer-Indian-Eskimo       311
        Asian-Pac-Islander      1039
        Black                   3124
        Other                    271
        White                  27816

sex: 32561x1 categorical

    Values:

        Female     10771
        Male       21790

capital_gain: 32561x1 double

    Values:

        Min            0
        Median         0
        Max        99999

capital_loss: 32561x1 double

    Values:

        Min            0
        Median         0
        Max         4356

hours_per_week: 32561x1 double

    Values:

        Min            1
        Median        40
        Max           99

salary: 32561x1 categorical

    Values:

        <=50K     24720
        >50K       7841
```

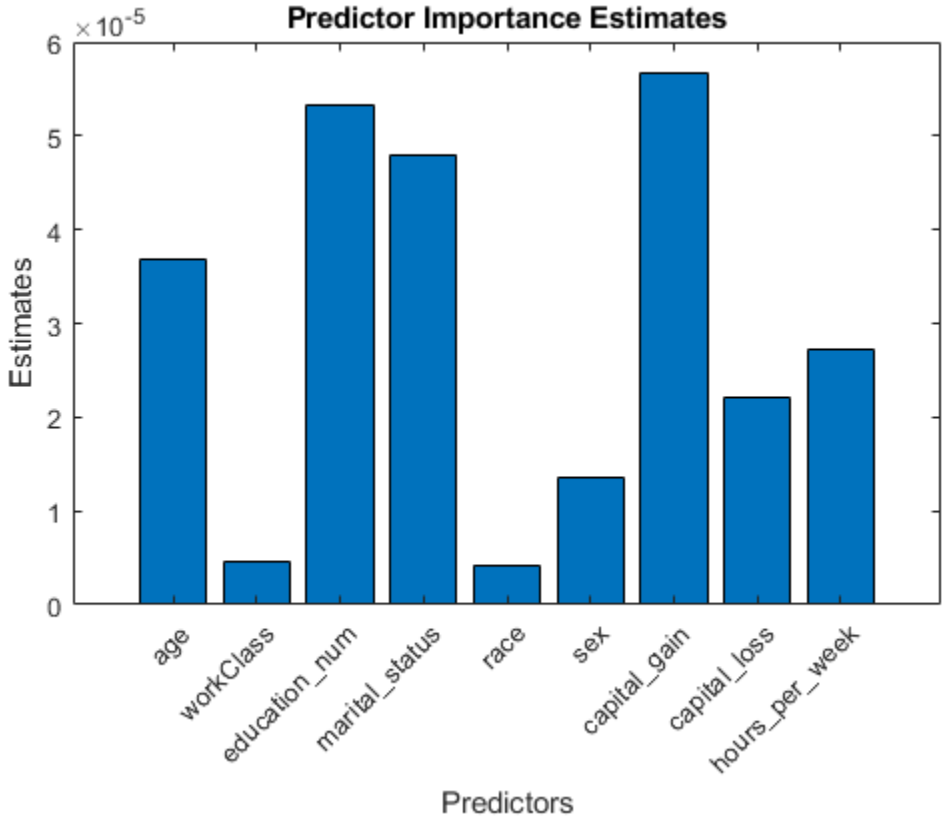
Because there are few categories represented in the categorical variables compared to levels in the continuous variables, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over the categorical variables.

Train a classification tree using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing observations in the data, specify usage of surrogate splits.

```
Mdl = fitctree(X,'salary','PredictorSelection','curvature',...
    'Surrogate','on');
```

Estimate predictor importance values by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. Compare the estimates using a bar graph.

```
imp = predictorImportance(Mdl);

figure;
bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `capital_gain` is the most important predictor, followed by `education_num`.

More About

Predictor Importance

`predictorImportance` computes importance measures of the predictors in a tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. For example, if a tree splits a parent node (for example,

node 1) into two child nodes (for example, nodes 2 and 3), then `predictorImportance` increases the importance of the split predictor by

$$(R_1 - R_2 - R_3)/N_{\text{branch}},$$

where R_i is the node risk of node i, and N_{branch} is the total number of branch nodes. A node risk is defined as a node error or node impurity weighted by the node probability:

$$R_i = P_i E_i,$$

where P_i is the node probability of node i, and E_i is either the node error (for a tree grown by minimizing the twoing criterion) or node impurity (for a tree grown by minimizing an impurity criterion, such as the Gini index or deviance) of node i.

The estimates of predictor importance depend on whether you use surrogate splits for training.

- If you use surrogate splits, `predictorImportance` sums the changes in the node risk over all splits at each branch node, including surrogate splits. If you do not use surrogate splits, then the function takes the sum over the best splits found at each branch node.
- Estimates of predictor importance do not depend on the order of predictors if you use surrogate splits, but do depend on the order if you do not use surrogate splits.
- If you use surrogate splits, `predictorImportance` computes estimates before the tree is reduced by pruning (or merging leaves). If you do not use surrogate splits, `predictorImportance` computes estimates after the tree is reduced by pruning. Therefore, pruning affects the predictor importance for a tree grown without surrogate splits, and does not affect the predictor importance for a tree grown with surrogate splits.

Impurity and Node Error

A decision tree splits nodes based on either impurity or node error.

Impurity means one of several things, depending on your choice of the `SplitCriterion` name-value pair argument:

- Gini's Diversity Index (`gdi`) — The Gini index of a node is

  $$1 - \sum_i p^2(i),$$

  where the sum is over the classes i at the node, and $p(i)$ is the observed fraction of classes with class i that reach the node. A node with just one class (a pure node) has Gini index `0`; otherwise the Gini index is positive. So the Gini index is a measure of node impurity.
- Deviance (`'deviance'`) — With $p(i)$ defined the same as for the Gini index, the deviance of a node is

  $$-\sum_i p(i)\log_2 p(i)\,.$$

  A pure node has deviance `0`; otherwise, the deviance is positive.
- Twoing rule (`'twoing'`) — Twoing is not a purity measure of a node, but is a different measure for deciding how to split a node. Let $L(i)$ denote the fraction of members of class i in the left child node after a split, and $R(i)$ denote the fraction of members of class i in the right child node after a split. Choose the split criterion to maximize

  $$P(L)P(R)\left(\sum_i |L(i) - R(i)|\right)^2,$$

where $P(L)$ and $P(R)$ are the fractions of observations that split to the left and right respectively. If the expression is large, the split made each child node purer. Similarly, if the expression is small, the split made each child node similar to each other, and therefore similar to the parent node. The split did not increase node purity.

- Node error — The node error is the fraction of misclassified classes at a node. If j is the class with the largest number of training samples at a node, the node error is

  $1 - p(j)$.

See Also

`fitcensemble` | `fitctree` | `oobPermutedPredictorImportance` | `predictorImportance (ClassificationEnsemble)`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

predictorImportance

Estimates of predictor importance for regression ensemble

Syntax

```
imp = predictorImportance(ens)
[imp,ma] = predictorImportance(ens)
```

Description

`imp = predictorImportance(ens)` computes estimates of predictor importance for `ens` by summing these estimates over all weak learners in the ensemble. `imp` has one element for each input predictor in the data used to train this ensemble. A high value indicates that this predictor is important for `ens`.

`[imp,ma] = predictorImportance(ens)` returns a P-by-P matrix with predictive measures of association for P predictors.

Input Arguments

`ens`

A regression ensemble, created by `fitrensemble`, or by the `compact` method.

Output Arguments

`imp`

A row vector with the same number of elements as the number of predictors (columns) in `ens.X`. The entries are the estimates of predictor importance, with `0` representing the smallest possible importance.

`ma`

A P-by-P matrix of predictive measures of association for P predictors. Element `ma(I,J)` is the predictive measure of association averaged over surrogate splits on predictor `J` for which predictor `I` is the optimal split predictor. `predictorImportance` averages this predictive measure of association over all trees in the ensemble.

Examples

Estimate Predictor Importance

Estimate the predictor importance for all predictor variables in the data.

Load the `carsmall` data set.

```
load carsmall
```

Grow an ensemble of 100 regression trees for `MPG` using `Acceleration`, `Cylinders`, `Displacement`, `Horsepower`, `Model_Year`, and `Weight` as predictors. Specify tree stumps as the weak learners.

```
X = [Acceleration Cylinders Displacement Horsepower Model_Year Weight];
t = templateTree('MaxNumSplits',1);
ens = fitrensemble(X,MPG,'Method','LSBoost','Learners',t);
```

Estimate the predictor importance for all predictor variables.

```
imp = predictorImportance(ens)
```

```
imp = 1×6

    0.0150         0    0.0066    0.1111    0.0437    0.5181
```

`Weight`, the last predictor, has the most impact on mileage. The second predictor has importance 0, which means that the number of cylinders has no impact on predictions made with `ens`.

Predictor Importance and Surrogate Splits

Estimate the predictor importance for all variables in the data and where the regression tree ensemble contains surrogate splits.

Load the `carsmall` data set.

```
load carsmall
```

Grow an ensemble of 100 regression trees for `MPG` using `Acceleration`, `Cylinders`, `Displacement`, `Horsepower`, `Model_Year`, and `Weight` as predictors. Specify tree stumps as the weak learners, and also identify surrogate splits.

```
X = [Acceleration Cylinders Displacement Horsepower Model_Year Weight];
t = templateTree('MaxNumSplits',1,'Surrogate','on');
ens = fitrensemble(X,MPG,'Method','LSBoost','Learners',t);
```

Estimate the predictor importance and predictive measures of association for all predictor variables.

```
[imp,ma] = predictorImportance(ens)
```

```
imp = 1×6

    0.2141    0.3798    0.4369    0.6498    0.3728    0.5700


ma = 6×6

    1.0000    0.0098    0.0102    0.0098    0.0033    0.0067
         0    1.0000         0         0         0         0
    0.0056    0.0084    1.0000    0.0078    0.0022    0.0084
    0.3537    0.4769    0.5834    1.0000    0.1612    0.5827
    0.0061    0.0070    0.0063    0.0064    1.0000    0.0056
    0.0154    0.0296    0.0533    0.0447    0.0070    1.0000
```

Comparing `imp` to the results in “Estimate Predictor Importance” on page 33-5042, `Horsepower` has the greatest impact on mileage, with `Weight` having the second greatest impact.

More About

Predictor Importance

`predictorImportance` estimates predictor importance of the predictors for each tree learner in the ensemble `ens` and returns the weighted average `imp` computed using `ens.TrainedWeight`. The output `imp` has one element for each predictor.

`predictorImportance` computes importance measures of the predictors in a tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. For example, if a tree splits a parent node (for example, node 1) into two child nodes (for example, nodes 2 and 3), then `predictorImportance` increases the importance of the split predictor by

$$(R_1 - R_2 - R_3)/N_{\text{branch}},$$

where R_i is node risk of node i, and N_{branch} is the total number of branch nodes. A node risk is defined as a node error weighted by the node probability:

$$R_i = P_i E_i,$$

where P_i is the node probability of node i, and E_i is the mean squared error of node i.

The estimates of predictor importance depend on whether you use surrogate splits for training.

- If you use surrogate splits, `predictorImportance` sums the changes in the node risk over all splits at each branch node, including surrogate splits. If you do not use surrogate splits, then the function takes the sum over the best splits found at each branch node.
- Estimates of predictor importance do not depend on the order of predictors if you use surrogate splits, but do depend on the order if you do not use surrogate splits.

Predictive Measure of Association

The predictive measure of association is a value that indicates the similarity between decision rules that split observations. Among all possible decision splits that are compared to the optimal split (found by growing the tree), the best surrogate decision split on page 33-1928 yields the maximum predictive measure of association. The second-best surrogate split has the second-largest predictive measure of association.

Suppose x_j and x_k are predictor variables j and k, respectively, and $j \neq k$. At node t, the predictive measure of association between the optimal split $x_j < u$ and a surrogate split $x_k < v$ is

$$\lambda_{jk} = \frac{\min(P_L, P_R) - \left(1 - P_{L_j L_k} - P_{R_j R_k}\right)}{\min(P_L, P_R)}.$$

- P_L is the proportion of observations in node t, such that $x_j < u$. The subscript L stands for the left child of node t.
- P_R is the proportion of observations in node t, such that $x_j \geq u$. The subscript R stands for the right child of node t.
- $P_{L_j L_k}$ is the proportion of observations at node t, such that $x_j < u$ and $x_k < v$.

- $P_{R_jR_k}$ is the proportion of observations at node t, such that $x_j \geq u$ and $x_k \geq v$.
- Observations with missing values for x_j or x_k do not contribute to the proportion calculations.

λ_{jk} is a value in (–∞,1]. If $\lambda_{jk} > 0$, then $x_k < v$ is a worthwhile surrogate split for $x_j < u$.

Algorithms

Element `ma(i,j)` is the predictive measure of association averaged over surrogate splits on predictor `j` for which predictor `i` is the optimal split predictor. This average is computed by summing positive values of the predictive measure of association over optimal splits on predictor `i` and surrogate splits on predictor `j` and dividing by the total number of optimal splits on predictor `i`, including splits for which the predictive measure of association between predictors `i` and `j` is negative.

See Also

`plotPartialDependence` | `predictorImportance (RegressionTree)` | `templateTree`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

predictorImportance

Estimates of predictor importance for regression tree

Syntax

```
imp = predictorImportance(tree)
```

Description

`imp = predictorImportance(tree)` computes estimates of predictor importance for `tree` by summing changes in the mean squared error due to splits on every predictor and dividing the sum by the number of branch nodes.

Input Arguments

`tree`

A regression tree created by `fitrtree`, or by the `compact` method.

Output Arguments

`imp`

A row vector with the same number of elements as the number of predictors (columns) in `tree.X`. The entries are the estimates of predictor importance, with `0` representing the smallest possible importance.

Examples

Estimate Predictor Importance

Estimate the predictor importance for all predictor variables in the data.

Load the `carsmall` data set.

```
load carsmall
```

Grow a regression tree for `MPG` using `Acceleration`, `Cylinders`, `Displacement`, `Horsepower`, `Model_Year`, and `Weight` as predictors.

```
X = [Acceleration Cylinders Displacement Horsepower Model_Year Weight];
tree = fitrtree(X,MPG);
```

Estimate the predictor importance for all predictor variables.

```
imp = predictorImportance(tree)
```

```
imp = 1×6
```

```
    0.0647    0.1068    0.1155    0.1411    0.3348    2.6565
```

`Weight`, the last predictor, has the most impact on mileage. The predictor with the minimal impact on making predictions is the first variable, which is `Acceleration`.

Predictor Importance and Surrogate Splits

Estimate the predictor importance for all variables in the data and where the regression tree contains surrogate splits.

Load the `carsmall` data set.

```
load carsmall
```

Grow a regression tree for `MPG` using `Acceleration`, `Cylinders`, `Displacement`, `Horsepower`, `Model_Year`, and `Weight` as predictors. Specify to identify surrogate splits.

```
X = [Acceleration Cylinders Displacement Horsepower Model_Year Weight];
tree = fitrtree(X,MPG,'Surrogate','on');
```

Estimate the predictor importance for all predictor variables.

```
imp = predictorImportance(tree)
```

```
imp = 1×6

    1.0449    2.4560    2.5570    2.5788    2.0832    2.8938
```

Comparing `imp` to the results in “Estimate Predictor Importance” on page 33-5046, `Weight` still has the most impact on mileage, but `Cylinders` is the fourth most important predictor.

Unbiased Predictor Importance Estimates

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

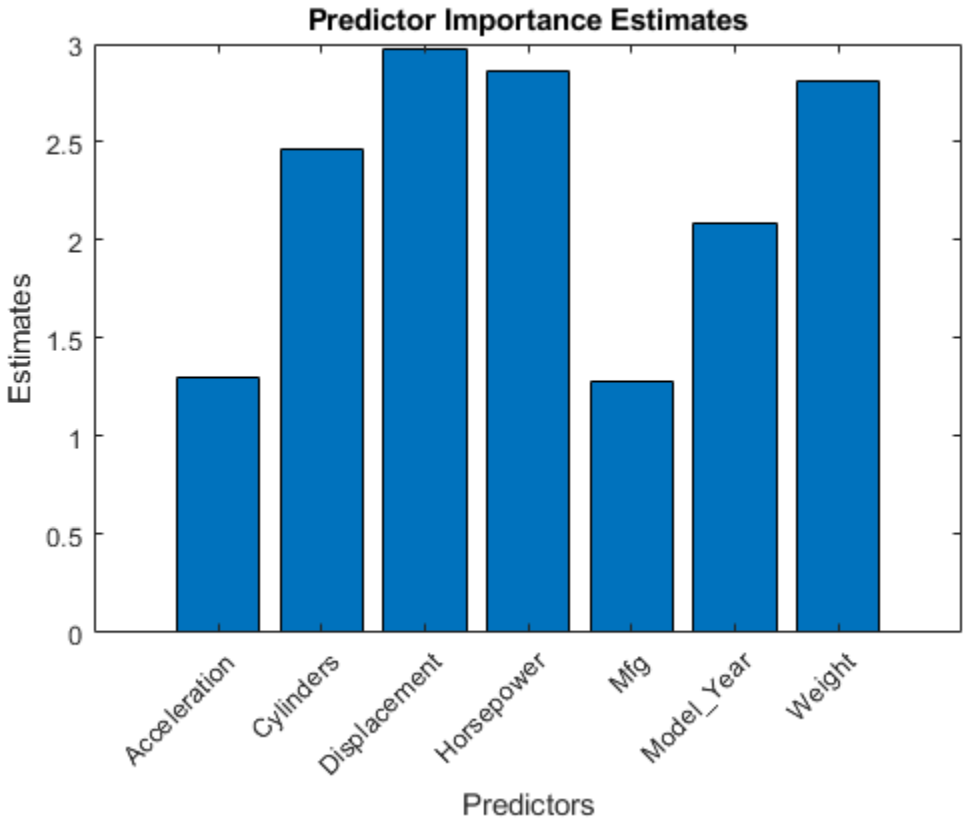
Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a regression tree using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits.

```
Mdl = fitrtree(X,'MPG','PredictorSelection','curvature','Surrogate','on');
```

Estimate predictor importance values by summing changes in the risk due to splits on every predictor and dividing the sum by the number of branch nodes. Compare the estimates using a bar graph.

```
imp = predictorImportance(Mdl);

figure;
bar(imp);
title('Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Displacement` is the most important predictor, followed by `Horsepower`.

More About

Predictor Importance

`predictorImportance` computes importance measures of the predictors in a tree by summing changes in the node risk due to splits on every predictor, and then dividing the sum by the total number of branch nodes. The change in the node risk is the difference between the risk for the parent node and the total risk for the two children. For example, if a tree splits a parent node (for example, node 1) into two child nodes (for example, nodes 2 and 3), then `predictorImportance` increases the importance of the split predictor by

$$(R_1 - R_2 - R_3)/N_{\text{branch}},$$

where R_i is node risk of node i, and N_{branch} is the total number of branch nodes. A node risk is defined as a node error weighted by the node probability:

$$R_i = P_i E_i,$$

where P_i is the node probability of node i, and E_i is the mean squared error of node i.

The estimates of predictor importance depend on whether you use surrogate splits for training.

- If you use surrogate splits, `predictorImportance` sums the changes in the node risk over all splits at each branch node, including surrogate splits. If you do not use surrogate splits, then the function takes the sum over the best splits found at each branch node.
- Estimates of predictor importance do not depend on the order of predictors if you use surrogate splits, but do depend on the order if you do not use surrogate splits.
- If you use surrogate splits, `predictorImportance` computes estimates before the tree is reduced by pruning (or merging leaves). If you do not use surrogate splits, `predictorImportance` computes estimates after the tree is reduced by pruning. Therefore, pruning affects the predictor importance for a tree grown without surrogate splits, and does not affect the predictor importance for a tree grown with surrogate splits.

See Also

`fitrensemble` | `fitrtree` | `oobPermutedPredictorImportance` | `plotPartialDependence` | `predictorImportance (RegressionEnsemble)`

Topics

“Choose Split Predictor Selection Technique” on page 19-14
“Introduction to Feature Selection” on page 15-49
“Interpret Machine Learning Models” on page 18-256

probplot

Probability plots

Syntax

```
probplot(y)
probplot(y,cens)
probplot(y,cens,freq)
probplot(dist,___)

probplot(ax,___)
probplot(ax,pd)
probplot(ax,fun,params)

probplot(___,'noref')

h = probplot(___)
```

Description

`probplot(y)` creates a normal probability plot comparing the distribution of the data in `y` to the normal distribution.

`probplot` plots each data point in `y` using marker symbols and draws a reference line that represents the theoretical distribution. If the sample data has a normal distribution, then the data points appear along the reference line. The reference line connects the first and third quartiles of the data and extends to the ends of the data. A distribution other than normal introduces curvature in the data plot.

`probplot(y,cens)` creates a probability plot using the censoring data in `cens`.

`probplot(y,cens,freq)` creates a probability plot using the censoring data in `cens` and the frequency data in `freq`.

`probplot(dist,___)` creates a probability plot for the distribution specified by `dist`, using any of the input arguments in the previous syntaxes.

`probplot(ax,___)` adds a probability plot into the existing probability plot axes specified by `ax`, using any of the input arguments in the previous syntaxes.

`probplot(ax,pd)` adds a fitted line on the existing probability plot axes specified by `ax` to represent the probability distribution `pd`.

`probplot(ax,fun,params)` adds a fitted line on the existing probability plot axes specified by `ax` to represent the function `fun` with the parameters `params`.

`probplot(___,'noref')` omits the reference line from the plot.

`h = probplot(___)` returns graphics handles corresponding to the plotted lines.

Examples

Create Weibull Probability Plot

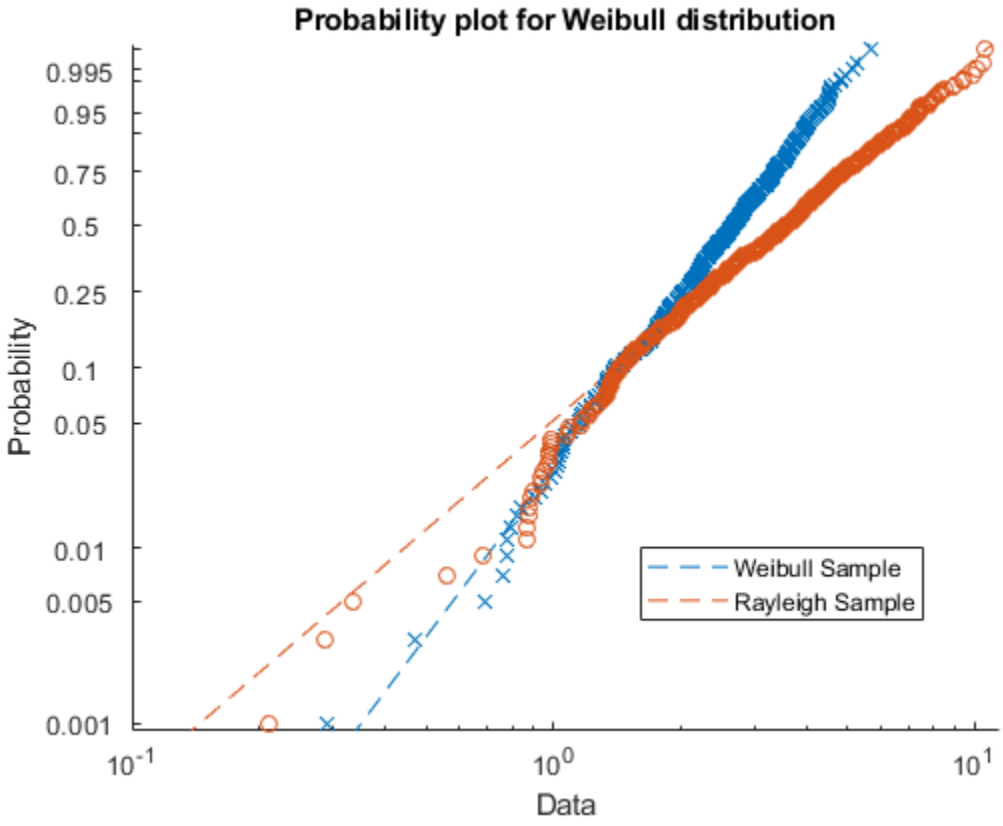
Generate sample data and create a probability plot.

Generate sample data. The sample `x1` contains 500 random numbers from a Weibull distribution with scale parameter `A = 3` and shape parameter `B = 3`. The sample `x2` contains 500 random numbers from a Rayleigh distribution with scale parameter `B = 3`.

```
rng('default');  % For reproducibility
x1 = wblrnd(3,3,[500,1]);
x2 = raylrnd(3,[500,1]);
```

Create a probability plot to assess whether the data in `x1` and `x2` comes from a Weibull distribution.

```
figure
probplot('weibull',[x1 x2])
legend('Weibull Sample','Rayleigh Sample','Location','best')
```



The probability plot shows that the data in `x1` comes from a Weibull distribution, while the data in `x2` does not.

Alternatively, you can use `wblplot` to create a Weibull probability plot.

Add Fitted Line to Probability Plot

Create a probability plot and an additional fitted line on the same figure.

Generate sample data containing about 20% outliers in the tails. The left tail of the sample data contains 10 values randomly generated from an exponential distribution with parameter `mu = 1`. The right tail contains 10 values randomly generated from an exponential distribution with parameter `mu = 5`. The center of the sample data contains 80 values randomly generated from a standard normal distribution.

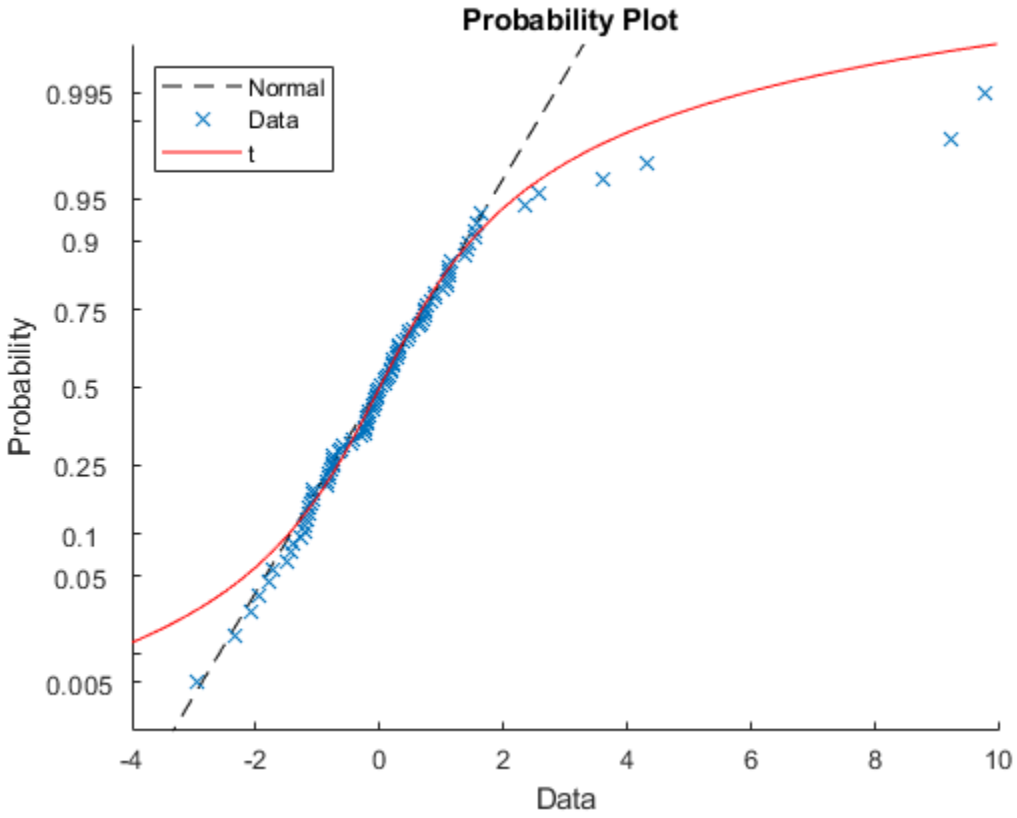
```
rng('default')  % For reproducibility
left_tail = -exprnd(1,10,1);
right_tail = exprnd(5,10,1);
center = randn(80,1);
data = [left_tail;center;right_tail];
```

Create a probability plot to assess whether the sample data comes from a normal distribution.

```
probplot(data)
```

Plot a *t* location-scale curve on the same figure to compare with `data`.

```
p = mle(data,'distribution','tLocationScale');
t = @(data,mu,sig,df)cdf('tLocationScale',data,mu,sig,df);
h = probplot(gca,t,p);
h.Color = 'r';
h.LineStyle = '-';
title('{\bf Probability Plot}')
legend('Normal','Data','t','Location','NW')
```



The plot shows that neither the normal line nor the *t* location-scale curve fits the tails very well because of the outliers.

Identify Significant Effects with Half-Normal Probability Plot

Create a half-normal probability distribution plot to identify significant effects in an experiment to study factors that might influence flow rate in a chemical manufacturing process. The four factors are reactants `A`, `B`, `C`, and `D`. Each factor is present at two levels (high and low concentration). The experiment contains only one replication at each factor level.

Load the sample data.

```
load flowrate
```

The first four columns of the table `flowrate` contain the design matrix for the factors and their interactions. The design matrix is coded to use `1` for the high factor level and `-1` for the low factor level. The fifth column of `flowrate` contains the measured flow rate.

Fit a linear regression model using `rate` as the response variable. Use predictor variables `A`, `B`, `C`, `D`, and all of their interaction terms.

```
mdl = fitlm(flowrate,'rate ~ A*B*C*D');
```

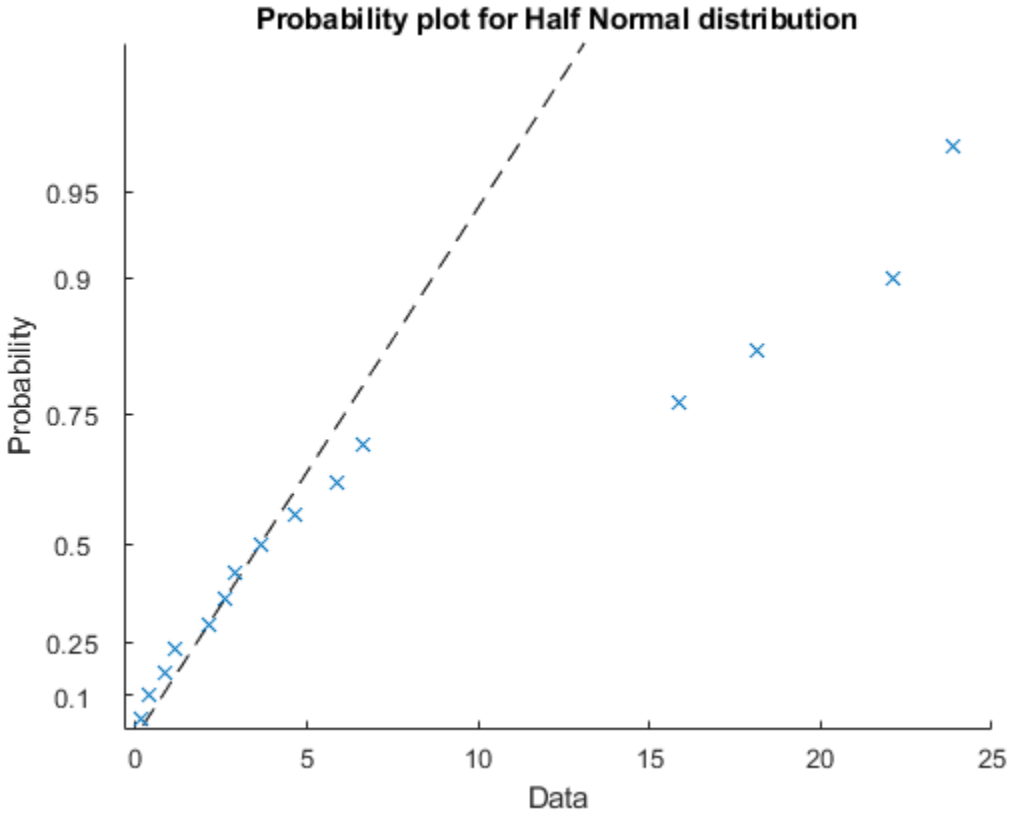
Calculate and store the absolute value of the factor effect estimates. To obtain the factor effect estimates, multiply the coefficient estimates obtained during the model fitting by two. This step is

necessary because the regression coefficients measure the effect of a one-unit change in `x` on the mean of `y`. However, the effects estimates measure a two-unit change in `x` due to the design matrix coding of -1 and 1. Exclude the baseline measurement. Note that the factor order in `mdl` may be different from the order in the original design matrix.

```
effects = abs(mdl.Coefficients{2:end,1}*2);
```

Create a half-normal probability plot using the absolute value of the effects estimates, excluding the baseline.

```
figure
h = probplot('halfnormal',effects);
```

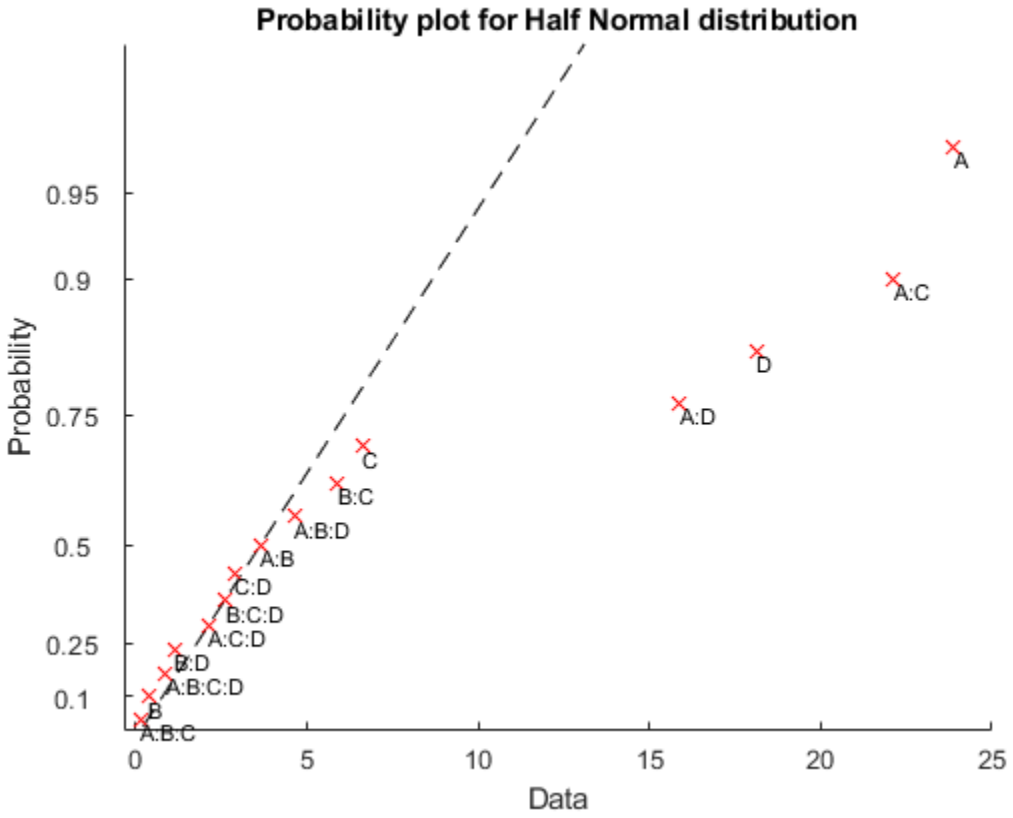


Label the points and format the plot. First, return the index values for the sorted effects estimates (from lowest to highest). Then use these index values to sort the probability values stored in the graphics handle (`h(1).YData`).

```
[b,i] = sort(effects);
prob(i) = h(1).YData;
```

Add text labels to the plot at each point. For each point, the x-value is the effects estimate and the y-value is the corresponding probability.

```
text(effects,prob,mdl.CoefficientNames(2:end),'FontSize',8,...
    'VerticalAlignment','top')
h(1).Color = 'r';
```



The points located far from the reference line represent the significant effects.

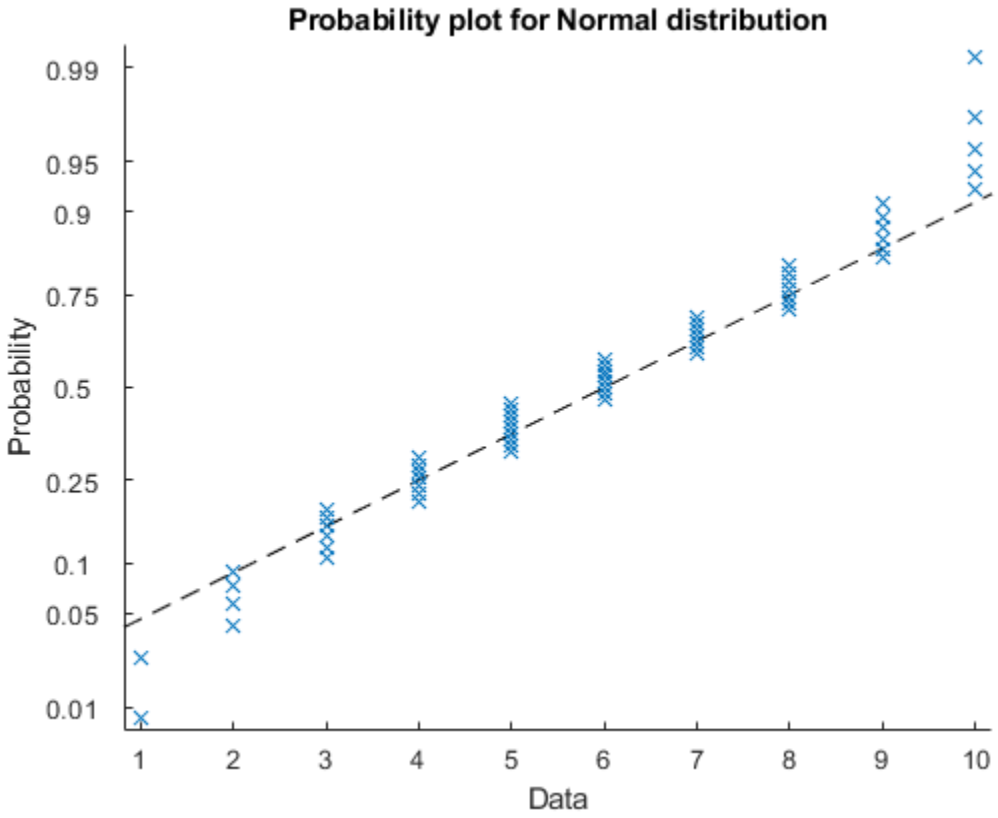
Create a Normal Probability Plot Using Frequency Data

Generate simulated frequency data.

```
y = 1:10;
freq = [2 4 6 7 9 8 7 7 6 5];
```

Create a normal probability plot using the frequency data.

```
probplot(y,[],freq)
```



The normal probability plot shows that the data do not have a normal distribution.

Input Arguments

y — Sample data

numeric vector | numeric matrix

Sample data, specified as a numeric vector or numeric matrix. `probplot` displays each value in `y` using marker symbols including `'x'` and `'o'`. If `y` is a matrix, then `probplot` displays a separate line for each column of `y`.

Not all distributions are appropriate for all data sets. `probplot` errors if the data set is inappropriate for a specified distribution. See `dist` for appropriate data ranges for each distribution.

dist — Distribution for probability plot

probability distribution object | `'normal'` | `'exponential'` | `'extreme value'` | `'half normal'` | `'lognormal'` | ...

Distribution for probability plot, specified as a probability distribution object or one of the following distribution names:

Name	Plot Type	Data Range
`'normal'`	Normal probability plot	All values

Name	Plot Type	Data Range
`'exponential'`	Exponential probability plot	Nonnegative values
`'extreme value'`	Extreme value probability plot	All values
`'half normal'`	Half-normal probability plot	All values
`'lognormal'`	Lognormal probability plot	Positive values
`'logistic'`	Logistic probability plot	All values
`'loglogistic'`	Loglogistic probability plot	Positive values
`'rayleigh'`	Rayleigh probability plot	Positive values
`'weibull'`	Weibull probability plot	Positive values

The default is `'normal'` if you create a probability plot in a new figure. If you add a probability plot to a figure that already includes one by using the `ax` input argument, then the default is the plot type of the existing probability plot.

You can create a probability distribution object with specified parameter values using `makedist`. Alternatively, fit a probability distribution object to sample data using `fitdist`. For more information on probability distribution objects, see “Working with Probability Distributions” on page 5-3.

The *y*-axis scale is based on the selected distribution. The *x*-axis has a log scale for the Weibull, loglogistic, and lognormal distributions, and a linear scale for the others.

Not all distributions are appropriate for all data sets. `probplot` errors if the data set is inappropriate for a specified distribution.

Example: `'weibull'`

`cens` — Censoring data
numeric vector

Censoring data, specified as a numeric vector. `cens` must be the same length as `y`, and contain a `1` value for observations that are right-censored and a `0` value for observations that are measured exactly.

Data Types: `single` | `double`

`freq` — Frequency data
vector of integer values

Frequency data, specified as a vector of integer values. `freq` must be the same length as `y`. `freq` contains the integer frequencies for the corresponding elements in `y`.

To create a probability plot using frequency data but not censoring data, specify empty brackets (`[]`) for `cens`.

Data Types: `single` | `double`

`ax` — Target axes
`Axes` object | `UIAxes` object

Target axes, specified as an `Axes` object or a `UIAxes` object. `probplot` adds an additional plot into the axes specified by `ax`. For details, see Axes Properties and UIAxes Properties.

Use `gca` to return the current axes for the current figure.

pd — Probability distribution for reference line
probability distribution object

Probability distribution for reference line, specified as a probability distribution object. `probplot` adds a fitted line to the axes specified by `ax` to represent the probability distribution specified by `pd`.

Create a probability distribution object with specified parameter values using `makedist`. Alternatively, fit a probability distribution object to sample data using `fitdist`. For more information on probability distribution objects, see “Working with Probability Distributions” on page 5-3.

`fun` — Function for reference line
function handle

Function for reference line, specified as a function handle. `probplot` adds a fitted line to the axes specified by `ax` to represent the function specified by `fun`, evaluated at the parameters specified by `params`.

`fun` is a function handle to a cdf function, specified using the function handle operator `@`. The function must accept a vector of input values as its first argument, and return a vector containing the cdf evaluated at each input value. Specify the parameter values required to evaluate `fun` using the `params` argument. For more information on function handles, see “Create Function Handle”.

Example: `@wblpdf`

Data Types: `function_handle`

`params` — Reference line function parameters
vector of numeric values | cell array

Reference line function parameters, specified as a vector of numeric values or a cell array. `probplot` adds a fitted line to the axes specified by `ax` to represent the function specified by `fun`, evaluated at the parameters specified by `params`.

`fun` is a function handle to a cdf function, specified using the function handle operator `@`. The function must accept a vector of values as its first argument, and return a vector of cdf values evaluated at each value. Specify the parameter values required to evaluate `fun` using the `params` argument. For more information on function handles, see “Create Function Handle”.

Output Arguments

`h` — Graphic handles for line objects
vector of `Line` graphic handles

Graphic handles for line objects, returned as a vector of Line graphic handles. Graphic handles are unique identifiers that you can use to query and modify the properties of a specific line on the plot. For each column of `y`, `probplot` returns two handles:

- The line representing the data points. `probplot` represents each data point in `y` using marker symbols such as `'+'` and `'o'`.
- The line showing the theoretical distribution for the probability plot, represented as a dashed line.

To view and set properties of line objects, use dot notation. For information on using dot notation, see “Access Property Values”. For information on the `Line` properties that you can set, see Primitive Line.

Algorithms

`probplot` matches the quantiles of sample data to the quantiles of a given probability distribution. The sample data is sorted, scaled according to the choice of `dist`, and plotted on the x-axis. When `dist` is `'lognormal'`, `'loglogistic'`, or `'weibull'`, the scaling is logarithmic. Otherwise, the scaling is linear. The y-axis represents the quantiles of the distribution specified in `dist`, converted into probability values. The scaling depends on the given distribution and is not linear.

Where the x-axis value is the ith sorted value from a sample of size N, the y-axis value is the midpoint between evaluation points of the empirical cumulative distribution function of the data. In the case of uncensored data, the midpoint is equal to $\frac{(i-0.5)}{N}$.

`probplot` superimposes a reference line to assess the linearity of the plot. If the data is uncensored, then the line goes through the first and third quartiles of the data. If the data is censored, then the line shifts accordingly. If the data is uncensored and `dist` is `'half normal'`, then `probplot` uses the zeroth and second quartiles instead.

See Also

`ecdf` | `normplot` | `wblplot`

Topics

“Distribution Plots” on page 4-7
“Hypothesis Testing” on page 8-5

Introduced before R2006a

procrustes

Procrustes analysis

Syntax

```
d = procrustes(X,Y)
[d,Z] = procrustes(X,Y)
[d,Z,transform] = procrustes(X,Y)
[...] = procrustes(...,'scaling',flag)
[...] = procrustes(...,'reflection',flag)
```

Description

`d = procrustes(X,Y)` determines a linear transformation (translation, reflection, orthogonal rotation, and scaling) of the points in matrix `Y` to best conform them to the points in matrix `X`. The goodness-of-fit criterion is the sum of squared errors. `procrustes` returns the minimized value of this dissimilarity measure in `d`. `d` is standardized by a measure of the scale of `X`, given by:

```
sum(sum((X-repmat(mean(X,1),size(X,1),1)).^2,1))
```

That is, the sum of squared elements of a centered version of `X`. However, if `X` comprises repetitions of the same point, the sum of squared errors is not standardized.

`X` and `Y` must have the same number of points (rows), and `procrustes` matches `Y(i)` to `X(i)`. Points in `Y` can have smaller dimension (number of columns) than those in `X`. In this case, `procrustes` adds columns of zeros to `Y` as necessary.

`[d,Z] = procrustes(X,Y)` also returns the transformed `Y` values.

`[d,Z,transform] = procrustes(X,Y)` also returns the transformation that maps `Y` to `Z`. `transform` is a structure array with fields:

- `c` — Translation component
- `T` — Orthogonal rotation and reflection component
- `b` — Scale component

That is:

```
c = transform.c;
T = transform.T;
b = transform.b;

Z = b*Y*T + c;
```

`[...] = procrustes(...,'scaling',flag)`, when *`flag`* is `false`, allows you to compute the transformation without a scale component (that is, with `b` equal to `1`). The default *`flag`* is `true`.

`[...] = procrustes(...,'reflection',flag)`, when *`flag`* is `false`, allows you to compute the transformation without a reflection component (that is, with `det(T)` equal to `1`). The default *`flag`* is `'best'`, which computes the best-fitting transformation, whether or not it includes a

reflection component. A *`flag`* of `true` forces the transformation to be computed with a reflection component (that is, with `det(T)` equal to `-1`)

Examples

Procrustes Analysis

Generate the sample data in two dimensions.

```
rng('default')
n = 10;
X = normrnd(0,1,[n 2]);
```

Rotate, scale, translate, and add some noise to sample points.

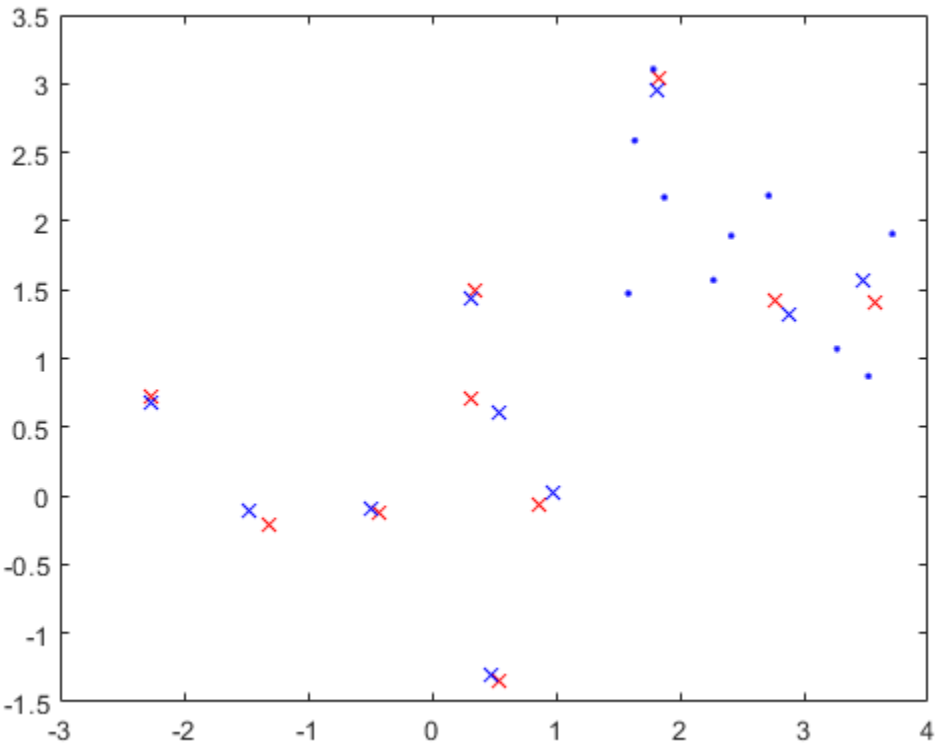
```
S = [0.5 -sqrt(3)/2; sqrt(3)/2 0.5];
Y = normrnd(0.5*X*S+2,0.05,n,2);
```

Conform Y to X using procrustes analysis.

```
[d,Z,tr] = procrustes(X,Y);
```

Plot the original X and Y with the transformed Y .

```
plot(X(:,1),X(:,2),'rx',Y(:,1),Y(:,2),'b.',Z(:,1),Z(:,2),'bx');
```



References

[1] Kendall, David G. “A Survey of the Statistical Theory of Shape.” *Statistical Science*. Vol. 4, No. 2, 1989, pp. 87–99.

[2] Bookstein, Fred L. *Morphometric Tools for Landmark Data*. Cambridge, UK: Cambridge University Press, 1991.

[3] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

See Also

`cmdscale` | `factoran`

Introduced before R2006a

proflik

Package: prob

Profile likelihood function for probability distribution

Syntax

```
[ll,param] = proflik(pd,pnum)
[ll,param] = proflik(pd,pnum,'Display',display)
[ll,param] = proflik(pd,pnum,setparam)
[ll,param] = proflik(pd,pnum,setparam,'Display',display)
[ll,param,other] = proflik( ___ )
```

Description

`[ll,param] = proflik(pd,pnum)` returns a vector `ll` of loglikelihood values and a vector `param` of corresponding parameter values for the parameter in the position indicated by `pnum`.

`[ll,param] = proflik(pd,pnum,'Display',display)` returns the loglikelihood values and corresponding parameter values, and plots the profile likelihood overlaid on an approximation of the loglikelihood.

`[ll,param] = proflik(pd,pnum,setparam)` returns the loglikelihood values and corresponding parameter values as specified by `setparam`.

`[ll,param] = proflik(pd,pnum,setparam,'Display',display)` returns the loglikelihood values and corresponding parameter values as specified by `setparam`, and plots the profile likelihood overlaid on an approximation of the loglikelihood.

`[ll,param,other] = proflik( ___ )` also returns a matrix `other` containing the values of the other parameters that maximize the likelihood, using any of the input arguments from the previous syntaxes.

Examples

Profile Likelihood of a Distribution Parameter

Load the sample data. Create a probability distribution object by fitting a Weibull distribution to the miles per gallon (MPG) data.

```
load carsmall
pd = fitdist(MPG,'Weibull')
```

```
pd =
  WeibullDistribution

  Weibull distribution
    A = 26.5079   [24.8333, 28.2954]
    B = 3.27193   [2.79441, 3.83104]
```

View the parameter names for the distribution.

```
pd.ParameterNames
```

```
ans = 1x2 cell
    {'A'}    {'B'}
```

For the Weibull distribution, A is in position 1, and B is in position 2.

Compute the profile likelihood for B, which is in position `pnum = 2`.

```
[ll,param] = proflik(pd,2);
```

Display the loglikelihood values for the estimated values of B.

```
[ll',param']
```

```
ans = 21×2

 -329.9688    2.7132
 -329.4312    2.7748
 -328.9645    2.8365
 -328.5661    2.8981
 -328.2340    2.9597
 -327.9658    3.0213
 -327.7596    3.0830
 -327.6135    3.1446
 -327.5256    3.2062
 -327.4943    3.2678
      ⋮
```

These results show that the profile log likelihood is maximized between the estimated B values of 3.2678 and 3.3295, which correspond to loglikelihood values -327.4943 and -327.5178. From the earlier fit, the MLE of B is 3.27193, which is in this interval as expected.

Profile Likelihood With Restricted Parameter Values

Load the sample data. Create a probability distribution object by fitting a generalized extreme value distribution to the miles per gallon (MPG) data.

```
load carsmall
pd = fitdist(MPG,'GeneralizedExtremeValue')
```

```
pd =
  GeneralizedExtremeValueDistribution

  Generalized Extreme Value distribution
        k = -0.207765   [-0.381674, -0.0338563]
    sigma =   7.49674   [6.31755, 8.89604]
       mu =   20.6233   [18.8859, 22.3606]
```

View the parameter names for the distribution.

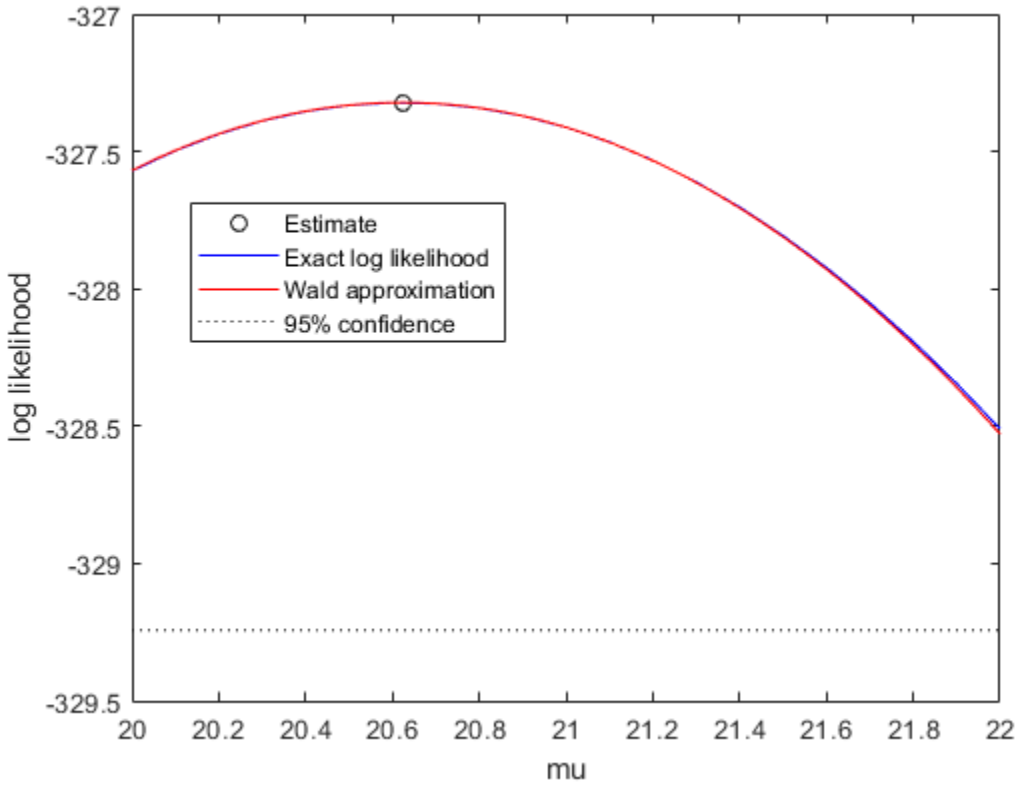
```
pd.ParameterNames
```

```
ans = 1x3 cell
    {'k'}    {'sigma'}    {'mu'}
```

For the generalized extreme value distribution, `k` is in position 1, `sigma` is in position 2, and `mu` is in position 3.

Compute the profile likelihood for `mu`, which is in position `pnum = 3`. Restrict the computation to parameter values from 20 to 22, and display the plot.

```
[ll,param,other] = proflik(pd,3,20:.1:22,'display','on');
```



The plot shows the estimated value for the parameter `mu` that maximizes the loglikelihood.

Display the loglikelihood values for the estimated values of `mu`, and the values of the other distribution parameters that maximize the corresponding loglikelihood.

```
[ll',param',other]
```

```
ans = 21×4

  -327.5706   20.0000   -0.1803    7.4087
  -327.4971   20.1000   -0.1846    7.4218
  -327.4364   20.2000   -0.1890    7.4354
  -327.3887   20.3000   -0.1934    7.4493
  -327.3538   20.4000   -0.1978    7.4636
```

```
-327.3317   20.5000   -0.2023    7.4783
-327.3223   20.6000   -0.2067    7.4932
-327.3257   20.7000   -0.2112    7.5084
-327.3418   20.8000   -0.2156    7.5240
-327.3706   20.9000   -0.2201    7.5399
     ⋮
```

The first column contains the log likelihood value that corresponds to the estimate of `mu` in the second column. The log likelihood is maximized between the parameter values 20.6000 and 20.7000, corresponding to log likelihood values -327.3223 and -327.3257. The third column contains the value of `k` that maximizes the corresponding log likelihood for `mu`. The fourth column contains the value of `sigma` that maximizes the corresponding log likelihood for `mu`.

Input Arguments

`pd` — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

`pnum` — Parameter number

positive integer value

Parameter number for which to compute the profile likelihood, specified as a positive integer value corresponding to the position of the desired parameter in the parameter name vector. For example, a Weibull distribution has a parameter name vector `{'A','B'}`, so specify `pnum` as `2` to compute the profile likelihood for `B`.

Data Types: `single` | `double`

`setparam` — Parameter value restriction

scalar value | vector of scalar values

Parameter value restriction, specified as a scalar value or a vector of such values. If you do not specify `setparam`, `proflik` chooses the values for output vector `param` based on the default confidence interval method for the probability distribution `pd`. If the parameter can take only restricted values, and if the confidence interval violates that restriction, you can use `setparam` to specify valid values.

Example: `[3,3.5,4]`

`display` — Display toggle

`'off'` (default) | `'on'`

Display toggle, specified as either `'on'` or `'off'`. Specify `'on'` to display the profile of the exact loglikelihood overlaid on the Wald approximation of the loglikelihood. Specify `'off'` to omit the

display. The Wald approximation is based on a Taylor series expansion around the estimated parameter value, as a function of the parameter in position `pnum` or its logarithm. The intersection of the curves with the horizontal dotted line marks the endpoints of 95% confidence intervals.

Output Arguments

`ll` — Loglikelihood values

vector

Loglikelihood values, returned as a vector. The loglikelihood is the value of the likelihood with the parameter in position `pnum` set to the values in `param`, maximized over the remaining parameters.

`param` — Parameter values

vector

Parameter values corresponding to the loglikelihood values in `ll`, returned as a vector. If you specify parameter values using `setparam`, then `param` is equal to `setparam`.

`other` — Other parameter values

matrix

Other parameter values that maximize the likelihood, returned as a matrix. Each row of `other` contains the values for all parameters except the parameter in position `pnum`.

See Also

Distribution Fitter | `fitdist` | `mle` | `negloglik` | `paramci`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

proximity

Class: `CompactTreeBagger`

Proximity matrix for data

Syntax

```
prox = proximity(B,X)
```

Description

`prox = proximity(B,X)` computes a numeric matrix of size `Nobs`-by-`Nobs` of proximities for data `X`, where `Nobs` is the number of observations (rows) in `X`. Proximity between any two observations in the input data is defined as a fraction of trees in the ensemble `B` for which these two observations land on the same leaf. This is a symmetric matrix with ones on the diagonal and off-diagonal elements ranging from 0 to 1.

prune

Class: `ClassificationTree`

Produce sequence of classification subtrees by pruning

Syntax

```
tree1 = prune(tree)
tree1 = prune(tree,Name,Value)
```

Description

`tree1 = prune(tree)` creates a copy of the classification tree `tree` with its optimal pruning sequence filled in.

`tree1 = prune(tree,Name,Value)` creates a pruned tree with additional options specified by one `Name,Value` pair argument. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree`

A classification tree created with `fitctree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha`

A numeric scalar. `prune` prunes `tree` to the specified value of the pruning cost.

`Level`

A numeric scalar from `0` (no pruning) to the largest pruning level of this tree `max(tree.PruneList)`. `prune` returns the tree pruned to this level.

`Nodes`

A numeric vector with elements from `1` to `tree.NumNodes`. Any `tree` branch nodes listed in `nodes` become leaf nodes in `tree1`, unless their parent nodes are also pruned.

Output Arguments

`tree1`

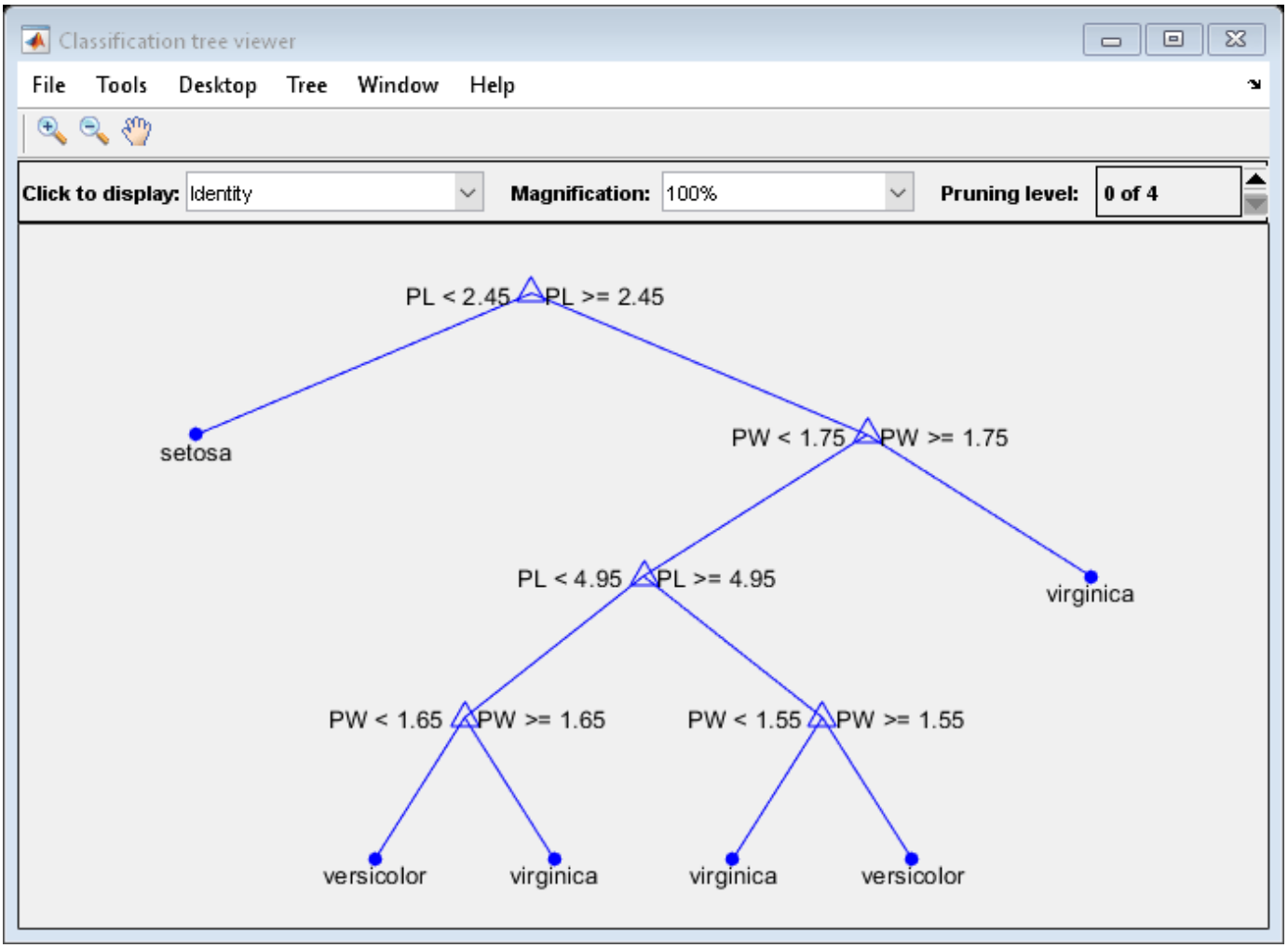
A classification tree.

Examples

Prune and Display a Classification Tree

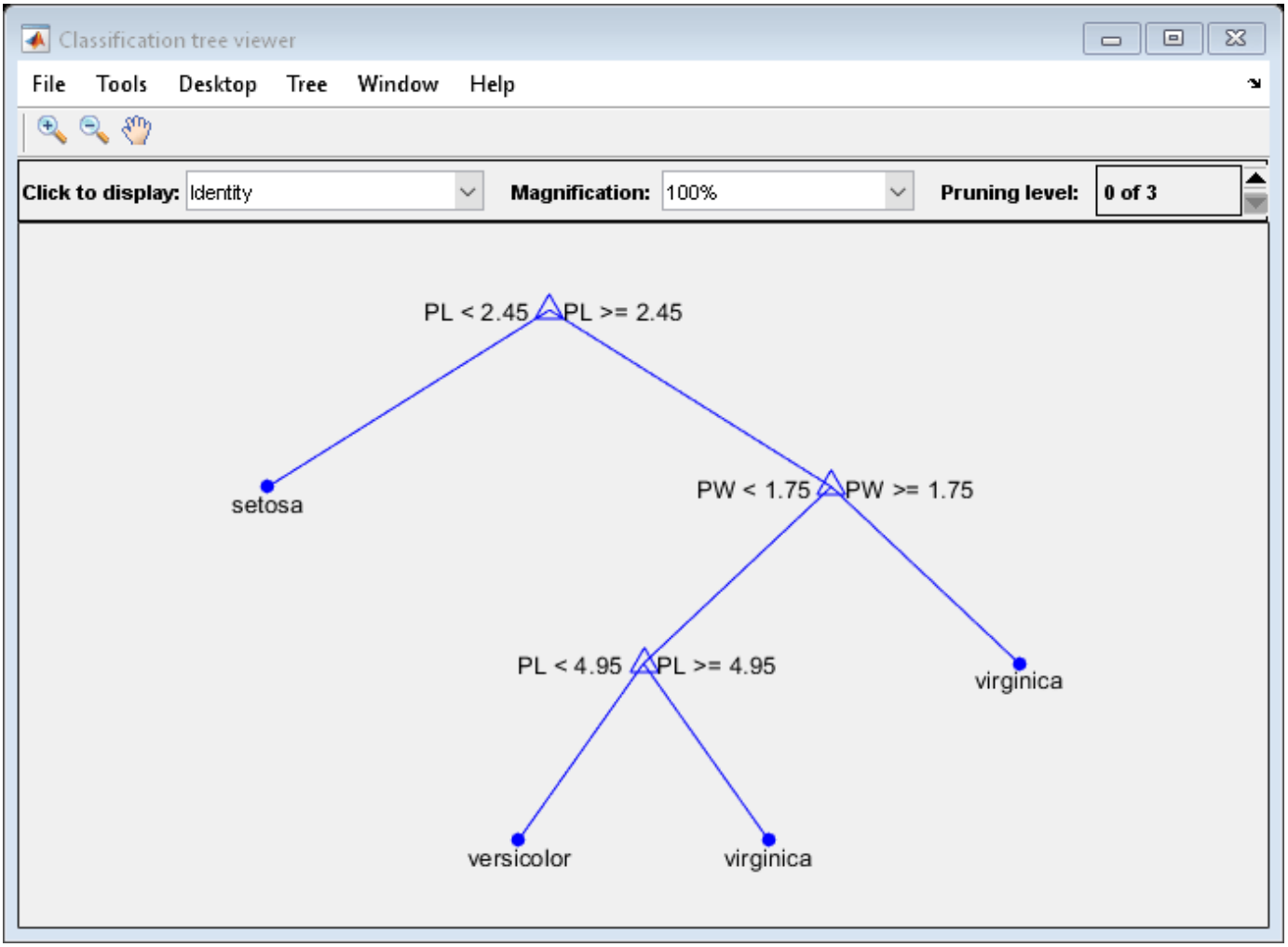
Construct and display a full classification tree for Fisher's iris data.

```
load fisheriris;
varnames = {'SL','SW','PL','PW'};
t1 = fitctree(meas,species,'MinParentSize',5,'PredictorNames',varnames);
view(t1,'Mode','graph');
```



Construct and display the next largest tree from the optimal pruning sequence.

```
t2 = prune(t1,'Level',1);
view(t2,'Mode','graph');
```



Tips

- `tree1 = prune(tree)` returns the decision tree `tree1` that is the full, unpruned `tree`, but with optimal pruning information added. This is useful only if you created `tree` by pruning another tree, or by using the `fitctree` function with pruning set `'off'`. If you plan to prune a tree multiple times along the optimal pruning sequence, it is more efficient to create the optimal pruning sequence first.

See Also

`fitctree`

prune

Class: `RegressionTree`

Produce sequence of regression subtrees by pruning

Syntax

```
tree1 = prune(tree)
tree1 = prune(tree,Name,Value)
```

Description

`tree1 = prune(tree)` creates a copy of the regression tree `tree` with its optimal pruning sequence filled in.

`tree1 = prune(tree,Name,Value)` creates a pruned tree with additional options specified by one `Name,Value` pair argument. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree`

A regression tree created with `fitrtree`.

Name-Value Pair Arguments

Optional comma-separated pair of `Name,Value` arguments, where `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside single quotes (`''`). You can specify only one name-value pair argument.

`Alpha`

A numeric scalar from `0` (no pruning) to `1` (prune to one node). Prunes to minimize the sum of (`Alpha` times the number of leaf nodes) and a cost (mean squared error).

`Level`

A numeric scalar from `0` (no pruning) to the largest pruning level of this tree `max(tree.PruneList)`. `prune` returns the tree pruned to this level.

`Nodes`

A numeric vector with elements from `1` to `tree.NumNodes`. Any `tree` branch nodes listed in `Nodes` become leaf nodes in `tree1`, unless their parent nodes are also pruned.

Output Arguments

tree1

A regression tree.

Examples

Prune Regression Tree

Load the `carsmall` data set. Consider `Horsepower` and `Weight` as predictor variables.

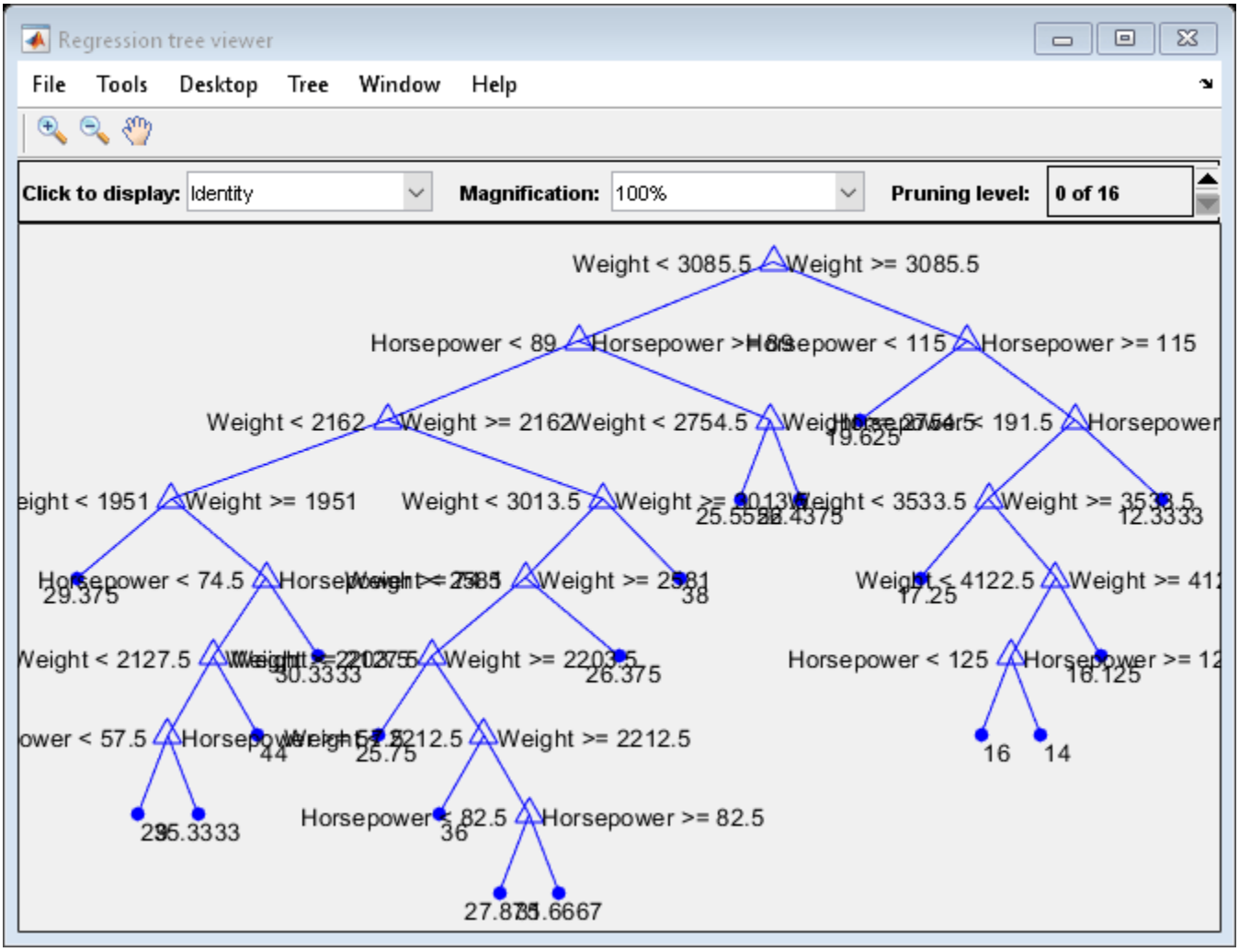
```
load carsmall;
X = [Weight Horsepower];
varNames = {'Weight' 'Horsepower'};
```

Grow a regression tree using the entire data set. View the tree.

```
Mdl = fitrtree(X,MPG,'PredictorNames',varNames)
```

```
Mdl =
  RegressionTree
             PredictorNames: {'Weight'  'Horsepower'}
               ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94


  Properties, Methods
```

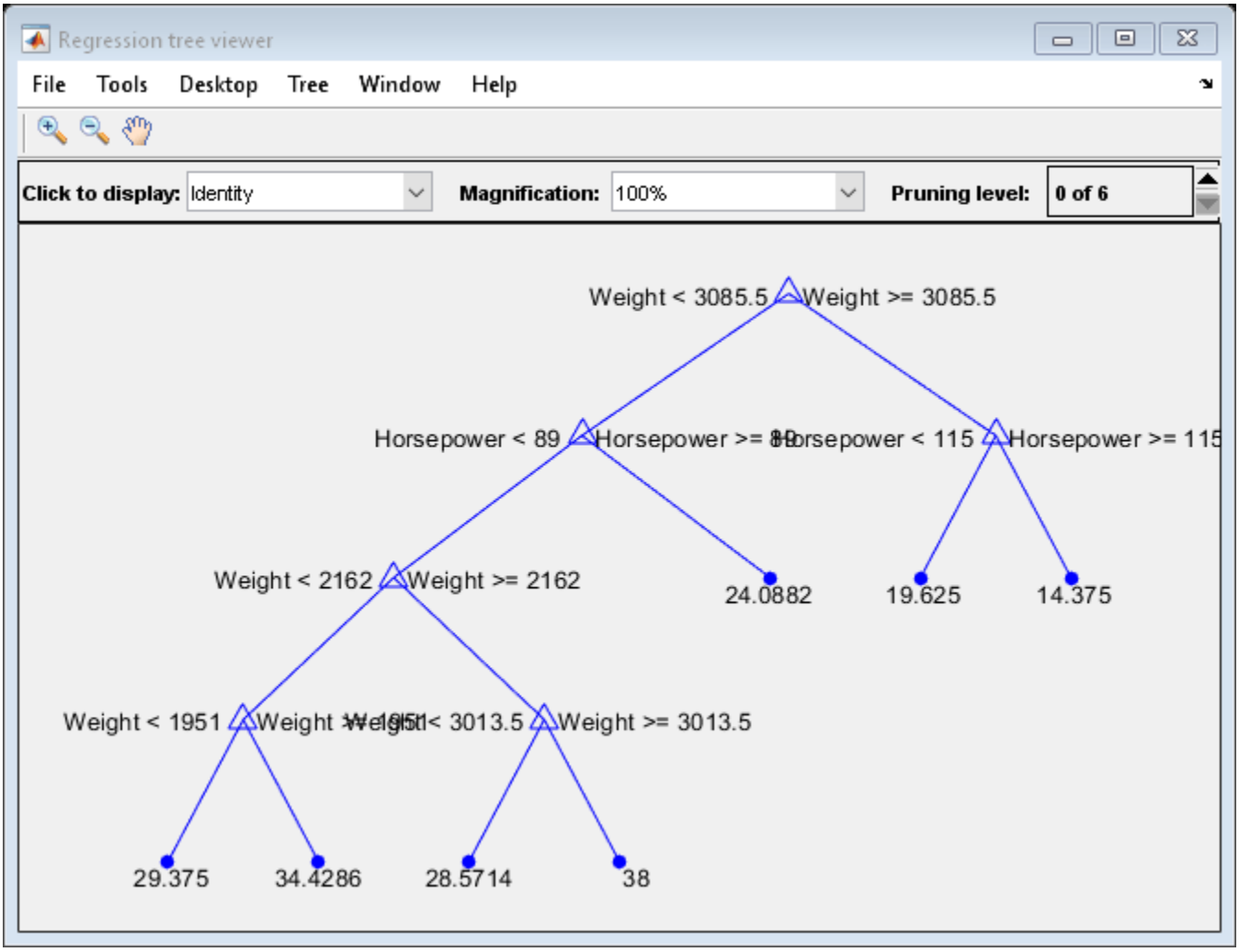
```
view(Mdl,'Mode','graph');
```



The regression tree has 16 pruning levels.

Prune the regression tree to pruning-level 10. View the pruned tree.

```
MdlPruned = prune(Mdl,'Level',10);
view(MdlPruned,'Mode','graph');
```



The pruned tree has six pruning levels.

Alternatively, you can use the pruning-level field in the Regression tree viewer to prune the tree.

Tips

- `tree1 = prune(tree)` returns the decision tree `tree1` that is the full, unpruned `tree`, but with optimal pruning information added. This is useful only if you created `tree` by pruning another tree, or by using `fitrtree` with pruning set `'off'`. If you plan to prune a tree multiple times along the optimal pruning sequence, it is more efficient to create the optimal pruning sequence first.

See Also

`fitrtree`

qrand

Class: `qrandstream`

Generate quasi-random points from stream

Syntax

```
x = qrand(q)
X = qrand(q,n)
```

Description

`x = qrand(q)` returns the next value `x` in the quasi-random number stream `q` of the `qrandstream` on page 33-5078 class. `x` is a 1-by-*d* vector, where *d* is the dimension of the stream. The command sets `q.State` to the index in the underlying point set of the next value to be returned.

`X = qrand(q,n)` returns the next `n` values `X` in an `n`-by-*d* matrix.

Objects `q` of the `qrandstream` class encapsulate properties of a specified quasi-random number stream. Values of the stream are not generated and stored in memory until `q` is accessed using `qrand`.

Examples

Use `qrandstream` to construct a 3-D Halton stream, based on a point set that skips the first 1000 values and then retains every 101st point:

```
q = qrandstream('halton',3,'Skip',1e3,'Leap',1e2)
q =
   Halton quasi-random stream in 3 dimensions
   Point set properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : none

nextIdx = q.State
nextIdx =
     1
```

Use `qrand` to generate two samples of size four:

```
X1 = qrand(q,4)
X1 =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
nextIdx = q.State
nextIdx =
     5

X2 = qrand(q,4)
```

```
X2 =
    0.2446    0.0238    0.8102
    0.5298    0.7540    0.0438
    0.3843    0.5112    0.2758
    0.8335    0.2245    0.4694
nextIdx = q.State
nextIdx =
     9
```

Use `reset` to reset the stream, then generate another sample:

```
reset(q)
nextIdx = q.State
nextIdx =
     1

X = qrand(q,4)
X =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
```

See Also

`qrandstream` | `reset`

qrandstream class

Quasi-random number streams

Construction

qrandstream	Construct quasi-random number stream

Methods

addlistener	Add listener for event
delete	Delete handle object
disp	Display `qrandstream` object
eq	Test handle equality
findobj	Find objects matching specified conditions
findprop	Find property of MATLAB handle object
ge	Greater than or equal relation for handles
gt	Greater than relation for handles
isvalid	Test handle validity
le	Less than or equal relation for handles
lt	Less than relation for handles
ne	Not equal relation for handles
notify	Notify listeners of event
qrand	Generate quasi-random points from stream
rand	Generate quasi-random points from stream
reset	Reset state

Properties

PointSet	Point set from which stream is drawn
State	Current state of the stream

Copy Semantics

Handle. To learn how this affects your use of the class, see Comparing Handle and Value Classes in the MATLAB Object-Oriented Programming documentation.

qrandstream

Class: `qrandstream`

Construct quasi-random number stream

Syntax

```
q = qrandstream(type,d)
q = qrandstream(type,d,prop1,val1,prop2,val2,...)
q = qrandstream(p)
```

Description

`q = qrandstream(type,d)` constructs a `d`-dimensional quasi-random number stream `q` of the `qrandstream` on page 33-5078 class, of type specified by *`type`*. *`type`* is either `'halton'` or `'sobol'`, and `q` is based on a point set from either the `haltonset` class or `sobolset` class, respectively, with default property settings.

`q = qrandstream(type,d,prop1,val1,prop2,val2,...)` specifies property name/value pairs for the point set on which the stream is based. Applicable properties depend on *`type`*.

`q = qrandstream(p)` constructs a stream based on the specified point set `p`. `p` must be a point set from either the `haltonset` class or `sobolset` class.

Examples

Construct a 3-D Halton stream, based on a point set that skips the first 1000 values and then retains every 101st point:

```
q = qrandstream('halton',3,'Skip',1e3,'Leap',1e2)
q =
   Halton quasi-random stream in 3 dimensions
   Point set properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : none

nextIdx = q.State
nextIdx =
     1
```

Use `qrand` to generate two samples of size four:

```
X1 = qrand(q,4)
X1 =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
nextIdx = q.State
nextIdx =
```

```
     5

X2 = qrand(q,4)
X2 =
    0.2446    0.0238    0.8102
    0.5298    0.7540    0.0438
    0.3843    0.5112    0.2758
    0.8335    0.2245    0.4694
nextIdx = q.State
nextIdx =
     9
```

Use `reset` to reset the stream, and then generate another sample:

```
reset(q)
nextIdx = q.State
nextIdx =
     1

X = qrand(q,4)
X =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
```

See Also

`haltonset` | `qrand` | `reset` | `sobolset`

Introduced in R2008a

qqplot

Quantile-quantile plot

Syntax

```
qqplot(x)
qqplot(x,pd)

qqplot(x,y)

qqplot( ___ ,pvec)
qqplot(ax, ___ )
h = qqplot( ___ )
```

Description

`qqplot(x)` displays a quantile-quantile plot of the quantiles of the sample data `x` versus the theoretical quantile values from a normal distribution. If the distribution of `x` is normal, then the data plot appears linear.

`qqplot` plots each data point in `x` using plus sign (`'+'`) markers and draws two reference lines that represent the theoretical distribution. A solid reference line connects the first and third quartiles of the data, and a dashed reference line extends the solid line to the ends of the data.

`qqplot(x,pd)` displays a quantile-quantile plot of the quantiles of the sample data `x` versus the theoretical quantiles of the distribution specified by the probability distribution object `pd`. If the distribution of `x` is the same as the distribution specified by `pd`, then the plot appears linear.

`qqplot(x,y)` displays a quantile-quantile plot of the quantiles of the sample data `x` versus the quantiles of the sample data `y`. If the samples come from the same distribution, then the plot appears linear.

`qqplot( ___ ,pvec)` displays a quantile-quantile plot with the quantiles specified in the vector `pvec`. Specify `pvec` after any of the input argument combinations in the previous syntaxes.

`qqplot(ax, ___ )` uses the plot axes specified by the `Axes` object `ax`. The option `ax` can precede any of the input argument combinations in the previous syntaxes.

`h = qqplot( ___ )` returns the handles (`h`) to the lines in the quantile-quantile plot.

Examples

Quantile-Quantile Plot for Normal Distribution

Use a quantile-quantile plot to determine whether gas prices in Massachusetts follow a normal distribution.

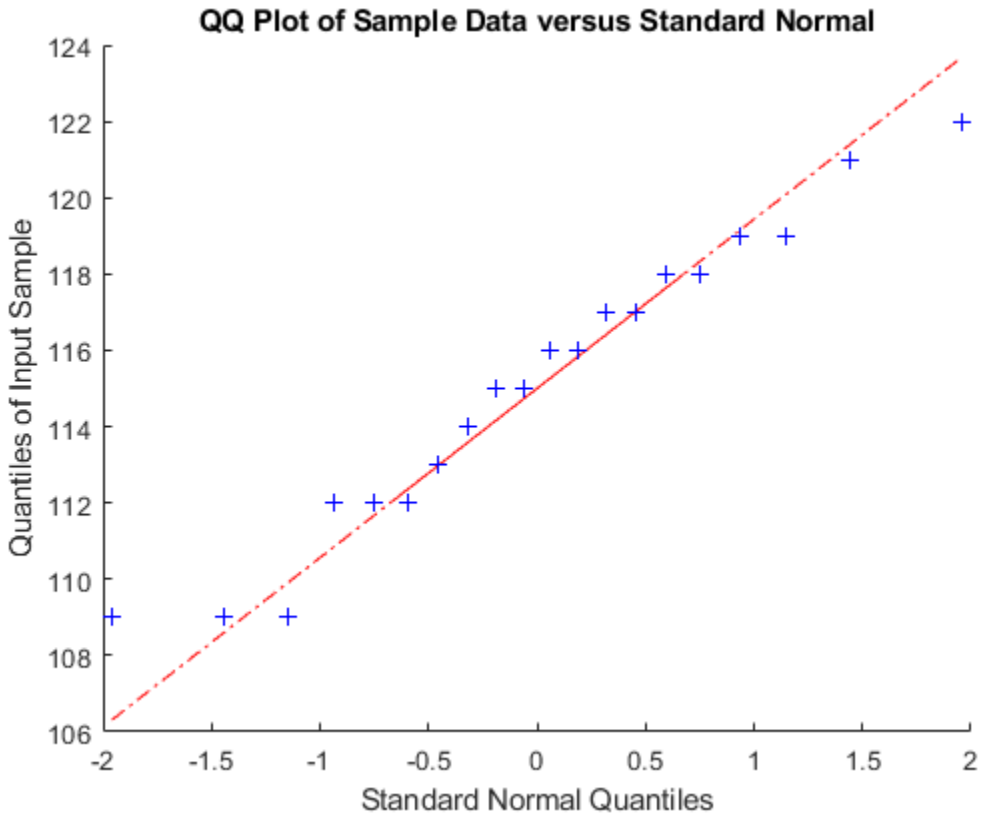
Load the sample data.

```
load gas
```

The sample data in `price1` and `price2` represent gasoline prices at 20 different gas stations in Massachusetts. The samples were collected during two different months.

Create a quantile-quantile plot to determine if the gas prices in `price1` follow a normal distribution.

```
figure
qqplot(price1)
```



The plot produces an approximately straight line, suggesting that the gas prices follow a normal distribution.

Quantile-Quantile Plot With Two Samples

Use a quantile-quantile plot to determine whether two sets of sample data come from the same distribution.

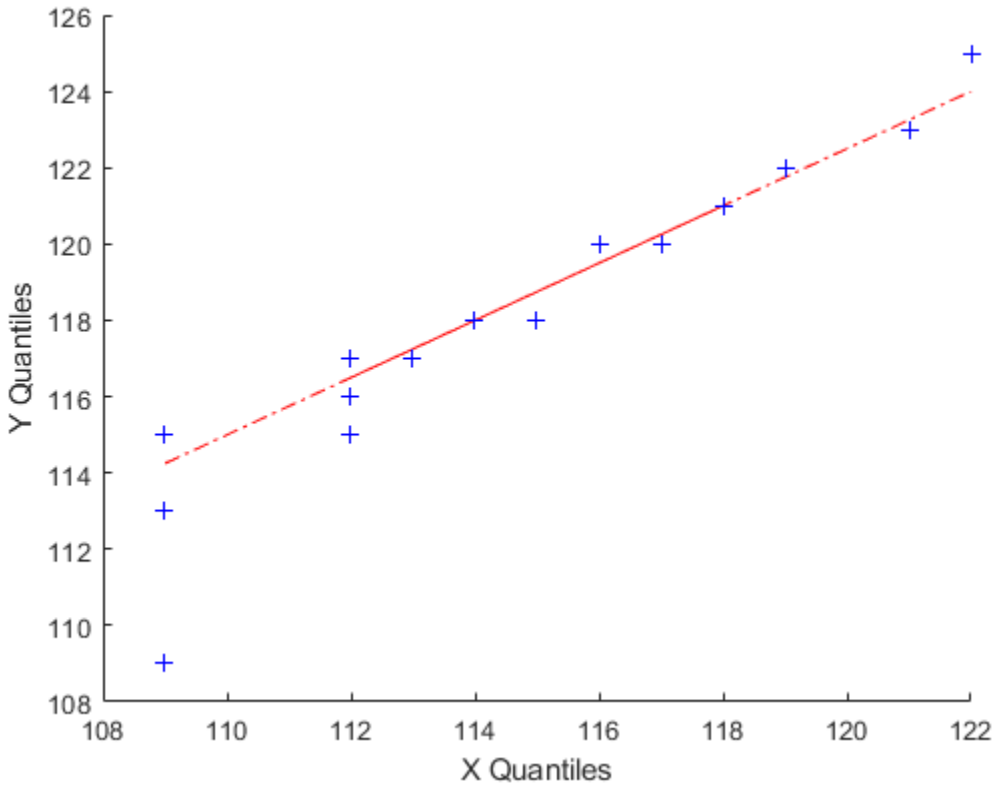
Load the sample data.

```
load gas
```

The sample data in `price1` and `price2` represent gasoline prices at 20 different gas stations in Massachusetts. The samples were collected during two different months.

Create a quantile-quantile plot using both sets of sample data, to assess whether prices at different times have the same distribution.

```
qqplot(price1,price2);
```



The plot produces an approximately straight line, suggesting that the two sets of sample data have the same distribution.

Quantile-Quantile Plot for Weibull Distribution

Use a quantile-quantile plot to determine whether sample data comes from a Weibull distribution.

Load the sample data.

```
load lightbulb
```

The first column of the data has the lifetime (in hours) of two types of light bulbs. The second column has information about the type of light bulb. 1 indicates fluorescent bulbs whereas 0 indicates the incandescent bulbs. The third column has censoring information. 1 indicates censored data, and 0 indicates the exact failure time. This is simulated data.

Remove the censored data.

```
lightbulb = [lightbulb(lightbulb(:,3)==0,1),...
    lightbulb(lightbulb(:,3)==0,2)];
```

Create a variable for each light bulb type. Include only uncensored data.

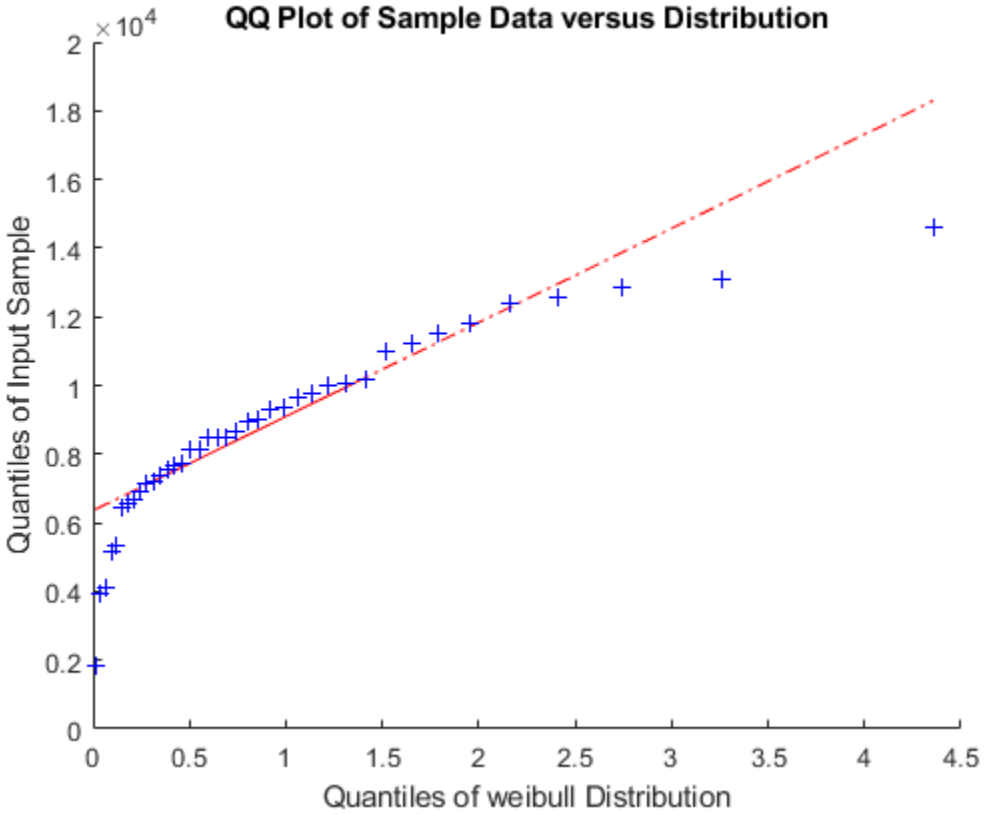
```
fluo = [lightbulb(lightbulb(:,2)==0,1)];
insc = [lightbulb(lightbulb(:,2)==1,1)];
```

Create a Weibull probability distribution object using the default parameters of `A = 1` and `B = 1`.

```
pd = makedist('Weibull');
```

Create a q-q plot to determine whether the lifetime of fluorescent bulbs has a Weibull distribution.

```
figure
qqplot(fluo,pd)
```



The plot is not a straight line, suggesting that the lifetime data for fluorescent bulbs does not follow a Weibull distribution.

Specify Axes for Quantile-Quantile Plots

Display a side-by-side pair of quantile-quantile plots using the `tiledlayout` and `nexttile` functions.

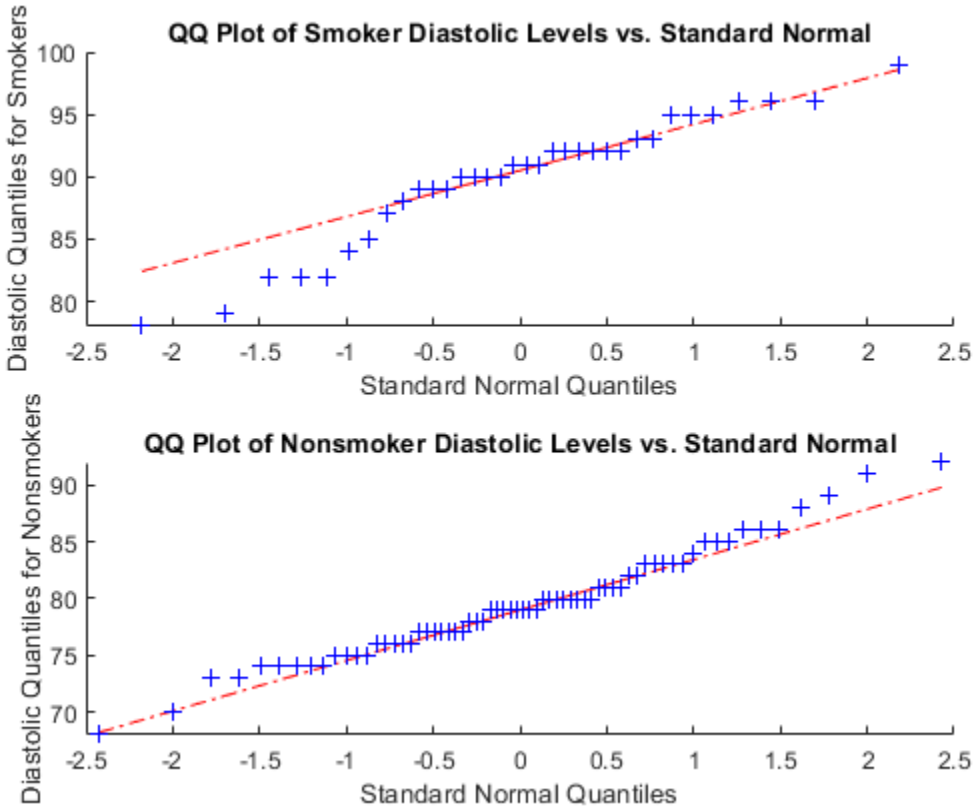
Load the `patients` data set. Separate the patient diastolic blood pressure levels into two data sets: one containing the diastolic blood pressure levels of smokers and one containing the diastolic levels of nonsmokers.

```
load patients
```

```
smokerIndices = (Smoker == 1);
nonsmokerIndices = (Smoker == 0);

smokerDiastolic = Diastolic(smokerIndices);
nonsmokerDiastolic = Diastolic(nonsmokerIndices);
```

Create a 2-by-1 tiled chart layout using the `tiledlayout` function. Create the first set of axes `ax1` within the chart layout by calling the `nexttile` function. In the axes, display a q-q plot to determine whether the diastolic blood pressure levels of smokers come from a normal distribution. Create the second set of axes `ax2` within the tiled chart layout by calling the `nexttile` function. In the axes, display a q-q plot to determine whether the diastolic blood pressure levels of nonsmokers come from a normal distribution.

```
tiledlayout(2,1)

% Top axes
ax1 = nexttile;
qqplot(ax1,smokerDiastolic)
ylabel(ax1,'Diastolic Quantiles for Smokers')
title(ax1,'QQ Plot of Smoker Diastolic Levels vs. Standard Normal')

% Bottom axes
ax2 = nexttile;
qqplot(ax2,nonsmokerDiastolic)
ylabel(ax2,'Diastolic Quantiles for Nonsmokers')
title(ax2,'QQ Plot of Nonsmoker Diastolic Levels vs. Standard Normal')
```



The second plot more closely follows a straight line, suggesting that the sample of nonsmoker blood pressure values has an approximately normal distribution. In contrast, the first plot has points below the line to the left, suggesting a heavier tail (more outliers) than a normal distribution.

Input Arguments

x — Sample data

numeric vector | numeric matrix

Sample data, specified as a numeric vector or numeric matrix. If `x` is a matrix, then `qqplot` displays a separate line for each column.

`qqplot` displays the sample data using the plot symbol `'+'`. A line joining the first and third quartiles of each distribution is superimposed on the plot. The line represents a robust linear fit of the order statistics for the data in `x`. This line is extrapolated out to the minimum and maximum values in `x` to help evaluate the linearity of the data.

Data Types: `single` | `double`

y — Second set of sample data

numeric vector | numeric matrix

Second set of sample data, specified as a numeric vector or numeric matrix. `x` and `y` do not need to be the same length. However, if `x` and `y` are matrices, they must contain the same number of columns. If `x` and `y` are matrices, then `qqplot` displays a separate line for each pair of columns.

`qqplot` selects the quantiles to plot based on the size of the smaller data set.

Data Types: `single` | `double`

pd — Hypothesized probability distribution

probability distribution object

Hypothesized probability distribution, specified as a probability distribution object. `qqplot` plots the quantiles of the input data `x` versus the theoretical quartiles of the distribution specified by `pd`.

Create a probability distribution object with specified parameter values using `makedist`, or fit a probability distribution object to data using `fitdist`.

pvec — Quantiles for plot

numeric value in the range [0,100] | vector of numeric values in the range [0,100]

Quantiles for plot, specified as a numeric value, or vector of numeric values, in the range [0,100].

For a single set of sample data (`x`), `qqplot` uses the quantiles in `x`. For two sets of sample data (`x` and `y`), `qqplot` uses the quantiles in the smaller of the two data sets.

Data Types: `single` | `double`

ax — Axes for plot

`Axes` object

Axes for the plot, specified as an `Axes` object. If you do not specify `ax`, then `qqplot` creates the plot using the current axes. For more information on creating an `Axes` object, see `axes`.

Output Arguments

h — Graphics handles for line objects

vector of `Line` graphics handles

Graphics handles for line objects, returned as a vector of `Line` graphics handles. Graphics handles are unique identifiers that you can use to query and modify the properties of a specific line on the plot. For each column of `x`, `qqplot` returns three handles:

- The line representing the data points. `qqplot` represents each data point in `x` using plus sign (`'+'`) markers.
- The line joining the first and third quartiles of each column of `x`, represented as a solid line.
- The extrapolation of the quartile line, extended to the minimum and maximum values of `x`, represented as a dashed line.

To view and set properties of line objects, use dot notation. For information on using dot notation, see "Access Property Values". For information on the `Line` properties that you can set, see Primitive Line.

More About

Quantile-Quantile Plot

A *quantile-quantile plot* (also called a *q-q plot*) visually assesses whether sample data comes from a specified distribution. Alternatively, a q-q plot assesses whether two sets of sample data come from the same distribution.

A q-q plot orders the sample data values from smallest to largest, then plots these values against the expected value for the specified distribution at each quantile in the sample data. The quantile values of the input sample appear along the *y*-axis, and the theoretical values of the specified distribution at the same quantiles appear along the *x*-axis. If the resulting plot is linear, then the sample data likely comes from the specified distribution.

The q-q plot selects quantiles based on the number of values in the sample data. If the sample data contains *n* values, then the plot uses *n* quantiles. Plot the *i*th ordered value (also called the *i*th *order statistic*) against the $\frac{i-0.5}{n}$th quantile of the specified distribution.

A q-q plot can also assess whether two sets of sample data have the same distribution, even if you do not know the underlying distribution. The quantile values for the first data set appear on the *x*-axis and the corresponding quantile values for the second data set appear on the *y*-axis. Since q-q plots rely on quantiles, the number of data points in the two samples does not need to be equal. If the sample sizes are unequal, the q-q plot chooses the quantiles based on the smaller data set. If the resulting plot is linear, then the two sets of sample data likely come from the same distribution.

See Also

`normplot` | `probplot`

Topics

"Distribution Plots" on page 4-7

Introduced before R2006a

quantile

Quantiles of a data set

Syntax

```
Y = quantile(X,p)
Y = quantile(X,N)

Y = quantile( ___ ,'all')
Y = quantile( ___ ,dim)
Y = quantile( ___ ,vecdim)

Y = quantile( ___ ,'Method',method)
```

Description

`Y = quantile(X,p)` returns quantiles of the elements in data vector or array `X` for the cumulative probability or probabilities `p` in the interval [0,1].

- If `X` is a vector, then `Y` is a scalar or a vector having the same length as `p`.
- If `X` is a matrix, then `Y` is a row vector or a matrix where the number of rows of `Y` is equal to the length of `p`.
- For multidimensional arrays on page 33-5101, `quantile` operates along the first nonsingleton dimension on page 33-5101 of `X`.

`Y = quantile(X,N)` returns quantiles for `N` evenly spaced cumulative probabilities (1/(`N` + 1), 2/(`N` + 1), ..., `N`/(`N` + 1)) for integer `N`>1.

- If `X` is a vector, then `Y` is a scalar or a vector with length `N`.
- If `X` is a matrix, then `Y` is a matrix where the number of rows of `Y` is equal to `N`.
- For multidimensional arrays on page 33-5101, `quantile` operates along the first nonsingleton dimension on page 33-5101 of `X`.

`Y = quantile( ___ ,'all')` returns quantiles of all the elements of `X` for either of the first two syntaxes.

`Y = quantile( ___ ,dim)` returns quantiles along the operating dimension `dim` for either of the first two syntaxes.

`Y = quantile( ___ ,vecdim)` returns quantiles over the dimensions specified in the vector `vecdim` for either of the first two syntaxes. For example, if `X` is a matrix, then `quantile(X,0.5,[1 2])` returns the 0.5 quantile of all the elements of `X` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

`Y = quantile( ___ ,'Method',method)` returns either exact or approximate quantiles based on the value of `method`, using any of the input argument combinations in the previous syntaxes.

Examples

Quantiles for Given Probabilities

Calculate the quantiles of a data set for specified probabilities.

Generate a data set of size 10.

```
rng('default'); % for reproducibility
x = normrnd(0,1,1,10)
```

```
x = 1×10

    0.5377    1.8339   -2.2588    0.8622    0.3188   -1.3077   -0.4336    0.3426    3.5784    2.7
```

Calculate the 0.3 quantile.

```
y = quantile(x,0.30)
```

```
y = -0.0574
```

Calculate the quantiles for the cumulative probabilities 0.025, 0.25, 0.5, 0.75, and 0.975.

```
y = quantile(x,[0.025 0.25 0.50 0.75 0.975])
```

```
y = 1×5

   -2.2588   -0.4336    0.4401    1.8339    3.5784
```

Quantiles for N Evenly Spaced Cumulative Probabilities

Calculate the quantiles of a data set for a given number of quantiles.

Generate a data set of size 10.

```
rng('default'); % for reproducibility
x = normrnd(0,1,1,10)
```

```
x = 1×10

    0.5377    1.8339   -2.2588    0.8622    0.3188   -1.3077   -0.4336    0.3426    3.5784    2.7
```

Calculate four evenly spaced quantiles.

```
y = quantile(x,4)
```

```
y = 1×4

   -0.8706    0.3307    0.6999    2.3017
```

Using `y = quantile(x,[0.2,0.4,0.6,0.8])` is another way to return the four evenly spaced quantiles.

Quantiles of a Matrix for Given Probabilities

Calculate the quantiles along the columns and rows of a data matrix for specified probabilities.

Generate a 4-by-6 data matrix.

```
rng default  % For reproducibility
X = normrnd(0,1,4,6)
```

```
X = 4×6

    0.5377    0.3188    3.5784    0.7254   -0.1241    0.6715
    1.8339   -1.3077    2.7694   -0.0631    1.4897   -1.2075
   -2.2588   -0.4336   -1.3499    0.7147    1.4090    0.7172
    0.8622    0.3426    3.0349   -0.2050    1.4172    1.6302
```

Calculate the 0.3 quantile for each column of `X` (`dim` = 1).

```
y = quantile(X,0.3,1)
```

```
y = 1×6

   -0.3013   -0.6958    1.5336   -0.1056    0.9491    0.1078
```

`quantile` returns a row vector `y` when calculating one quantile for each column of a matrix. For example, `-0.3013` is the 0.3 quantile of the first column of `X` with elements (0.5377, 1.8339, -2.2588, 0.8622). Because the default value of `dim` is 1, you can return the same result with `y = quantile(X,0.3)`.

Calculate the 0.3 quantile for each row of `X` (`dim` = 2).

```
y = quantile(X,0.3,2)
```

```
y = 4×1

    0.3844
   -0.8642
   -1.0750
    0.4985
```

`quantile` returns a column vector `y` when calculating one quantile for each row of a matrix. For example `0.3844` is the 0.3 quantile of the first row of `X` with elements (0.5377, 0.3188, 3.5784, 0.7254, -0.1241, 0.6715).

Quantiles of a Matrix for Given Number of Quantiles

Calculate the *N* evenly spaced quantiles along the columns and rows of a data matrix.

Generate a 6-by-10 data matrix.

```
rng('default');  % for reproducibility
X = unidrnd(10,6,7)
```

```
X = 6×7

     9     3    10     8     7     8     7
    10     6     5    10     8     1     4
     2    10     9     7     8     3    10
    10    10     2     1     4     1     1
     7     2     5     9     7     1     5
     1    10    10    10     2     9     4
```

Calculate three evenly spaced quantiles for each column of `X` (`dim` = 1).

```
y = quantile(X,3,1)
```

```
y = 3×7

    2.0000    3.0000    5.0000    7.0000    4.0000    1.0000    4.0000
    8.0000    8.0000    7.0000    8.5000    7.0000    2.0000    4.5000
   10.0000   10.0000   10.0000   10.0000    8.0000    8.0000    7.0000
```

Each column of matrix `y` corresponds to the three evenly spaced quantiles of each column of matrix `X`. For example, the first column of `y` with elements (2, 8, 10) has the quantiles for the first column of `X` with elements (9, 10, 2, 10, 7, 1). `y = quantile(X,3)` returns the same answer because the default value of `dim` is 1.

Calculate three evenly spaced quantiles for each row of `X` (`dim` = 2).

```
y = quantile(X,3,2)
```

```
y = 6×3

    7.0000    8.0000    8.7500
    4.2500    6.0000    9.5000
    4.0000    8.0000    9.7500
    1.0000    2.0000    8.5000
    2.7500    5.0000    7.0000
    2.5000    9.0000   10.0000
```

Each row of matrix `y` corresponds to the three evenly spaced quantiles of each row of matrix `X`. For example, the first row of `y` with elements (7, 8, 8.75) has the quantiles for the first row of `X` with elements (9, 3, 10, 8, 7, 8, 7).

Quantiles of Multidimensional Array for Given Probabilities

Calculate the quantiles of a multidimensional array for specified probabilities by using the `'all'` and `vecdim` input arguments.

Create a 3-by-5-by-2 array `X`. Specify the vector of probabilities `p`.

```
X = reshape(1:30,[3 5 2])
```

```
X =
X(:,:,1) =
```

```
     1     4     7    10    13
     2     5     8    11    14
     3     6     9    12    15


X(:,:,2) =

    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
```

```
p = [0.25 0.75];
```

Calculate the 0.25 and 0.75 quantiles of all the elements in `X`.

```
Yall = quantile(X,p,'all')
```

```
Yall = 2×1

     8
    23
```

`Yall(1)` is the 0.25 quantile of `X`, and `Yall(2)` is the 0.75 quantile of `X`.

Calculate the 0.25 and 0.75 quantiles for each page of `X` by specifying dimensions 1 and 2 as the operating dimensions.

```
Ypage = quantile(X,p,[1 2])
```

```
Ypage =
Ypage(:,:,1) =

    4.2500
   11.7500


Ypage(:,:,2) =

   19.2500
   26.7500
```

For example, `Ypage(1,1,1)` is the 0.25 quantile of the first page of `X`, and `Ypage(2,1,1)` is the 0.75 quantile of the first page of `X`.

Calculate the 0.25 and 0.75 quantiles of the elements in each `X(i,:,:)` slice by specifying dimensions 2 and 3 as the operating dimensions.

```
Yrow = quantile(X,p,[2 3])
```

```
Yrow = 3×2

     7    22
     8    23
     9    24
```

For example, `Yrow(3,1)` is the 0.25 quantile of the elements in `X(3,:,:)`, and `Yrow(3,2)` is the 0.75 quantile of the elements in `X(3,:,:)`.

Median and Quartiles for Even Number of Data Elements

Find median and quartiles of a vector, `x`, with even number of elements.

Enter the data.

```
x = [2 5 6 10 11 13]
```

```
x = 1×6

     2     5     6    10    11    13
```

Calculate the median of `x`.

```
y = quantile(x,0.50)
```

```
y = 8
```

Calculate the quartiles of `x`.

```
y = quantile(x,[0.25, 0.5, 0.75])
```

```
y = 1×3

     5     8    11
```

Using `y = quantile(x,3)` is another way to compute the quartiles of `x`.

These results might be different than the textbook definitions because `quantile` uses “Linear Interpolation” on page 33-5101 to find the median and quartiles.

Median and Quartiles for Odd Number of Data Elements

Find median and quartiles of a vector, `x`, with odd number of elements.

Enter the data.

```
x = [2 4 6 8 10 12 14]
```

```
x = 1×7

     2     4     6     8    10    12    14
```

Find the median of `x`.

```
y = quantile(x,0.50)
```

```
y = 8
```

Find the quartiles of x.

```
y = quantile(x,[0.25, 0.5, 0.75])
```

```
y = 1×3

    4.5000    8.0000   11.5000
```

Using `y = quantile(x,3)` is another way to compute the quartiles of x.

These results might be different than the textbook definitions because `quantile` uses “Linear Interpolation” on page 33-5101 to find the median and quartiles.

Quantiles of Tall Vector for Given Probability

Calculate exact and approximate quantiles of a tall column vector for a given probability.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore for the `airlinesmall` data set. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Specify to work with the `ArrTime` variable.

```
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames','ArrTime');
```

Create a tall table on top of the datastore, and extract the data from the tall table into a tall vector.

```
t = tall(ds) % Tall table
```

```
t =

  Mx1 tall table

    ArrTime
    _______

      735
     1124
     2218
     1431
      746
     1547
     1052
     1134
       :
       :
```

```
x = t{:,:}   % Tall vector
```

```
x =
```

```
  Mx1 tall double column vector

        735
       1124
       2218
       1431
        746
       1547
       1052
       1134
        :
        :
```

Calculate the exact quantile of x for `p` = 0.5. Because `X` is a tall column vector and `p` is a scalar, `quantile` returns the exact quantile value by default.

```
p = 0.5; % Cumulative probability
yExact = quantile(x,p)
```

```
yExact =

  tall double

    ?
```

Calculate the approximate quantile of x for `p` = 0.5. Specify `'Method','approximate'` to use an approximation algorithm based on “T-Digest” on page 33-5102 for computing the quantiles.

```
yApprox = quantile(x,p,'Method','approximate')
```

```
yApprox =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[yExact,yApprox] = gather(yExact,yApprox)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 4: Completed in 2.1 sec
- Pass 2 of 4: Completed in 0.91 sec
- Pass 3 of 4: Completed in 1.4 sec
- Pass 4 of 4: Completed in 1.2 sec
Evaluation completed in 7.4 sec

yExact = 1522

yApprox = 1.5220e+03
```

The values of the approximate quantile and the exact quantile are the same to the four digits shown.

Quantiles of Tall Matrix Along Different Dimensions

Calculate exact and approximate quantiles of a tall matrix for specified cumulative probabilities along different dimensions.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a tall matrix `X` containing a subset of variables from the `airlinesmall` data set. See “Quantiles of Tall Vector for Given Probability” on page 33-5094 for details about the steps to extract data from a tall array.

```
varnames = {'ArrDelay','ArrTime','DepTime','ActualElapsedTime'}; % Subset of variables in the da
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames',varnames); % Datastore
t = tall(ds);     % Tall table
X = t{:,varnames} % Tall matrix
```

```
X =

  Mx4 tall double matrix

           8         735         642          53
           8        1124        1021          63
          21        2218        2055          83
          13        1431        1332          59
           4         746         629          77
          59        1547        1446          61
           3        1052         928          84
          11        1134         859         155
          :           :           :           :
          :           :           :           :
```

When operating along a dimension that is not 1, the `quantile` function calculates the exact quantiles only, so that it can perform the computation efficiently using a sorting-based algorithm (see “Algorithms” on page 33-5103) instead of an approximation algorithm based on “T-Digest” on page 33-5102.

Calculate the exact quantiles of `X` along the second dimension for the cumulative probabilities 0.25, 0.5, and 0.75.

```
p = [0.25 0.50 0.75]; % Vector of cumulative probabilities
Yexact = quantile(X,p,2)
```

```
Yexact =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

When the function operates along the first dimension and `p` is a vector of cumulative probabilities, you must use the approximation algorithm based on t-digest to compute the quantiles. Using the sorting-based algorithm to find the quantiles along the first dimension of a tall array is computationally intensive.

Calculate the approximate quantiles of `X` along the first dimension for the cumulative probabilities 0.25, 0.5, and 0.75. Because the default dimension is 1, you do not need to specify a value for `dim`.

```
Yapprox = quantile(X,p,'Method','approximate')
```

```
Yapprox =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[Yexact,Yapprox] = gather(Yexact,Yapprox);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 5.2 sec
Evaluation completed in 6.7 sec
```

Show the first five rows of the exact quantiles of `X` (along the second dimension) for the cumulative probabilities 0.25, 0.5, and 0.75.

```
Yexact(1:5,:)
```

```
ans = 5×3
10^3 ×

    0.0305    0.3475    0.6885
    0.0355    0.5420    1.0725
    0.0520    1.0690    2.1365
    0.0360    0.6955    1.3815
    0.0405    0.3530    0.6875
```

Each row of the matrix `Yexact` contains the three quantiles of the corresponding row in `X`. For example, `30.5`, `347.5`, and `688.5` are the 0.25, 0.5, and 0.75 quantiles, respectively, of the first row in `X`.

Show the approximate quantiles of `X` (along the first dimension) for the cumulative probabilities 0.25, 0.5, and 0.75.

```
Yapprox
```

```
Yapprox = 3×4
10^3 ×

   -0.0070    1.1149    0.9322    0.0700
         0    1.5220    1.3350    0.1020
    0.0110    1.9180    1.7400    0.1510
```

Each column of the matrix `Yapprox` corresponds to the three quantiles for each column of the matrix `X`. For example, the first column of `Yapprox` with elements (–7, 0, 11) contains the quantiles for the first column of `X`.

Quantiles of Tall Matrix for N Evenly Spaced Probabilities

Calculate exact and approximate quantiles along different dimensions of a tall matrix for `N` evenly spaced cumulative probabilities.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a tall matrix `X` containing a subset of variables from the `airlinesmall` data set. See “Quantiles of Tall Vector for Given Probability” on page 33-5094 for details about the steps to extract data from a tall array.

```
varnames = {'ArrDelay','ArrTime','DepTime','ActualElapsedTime'}; % Subset of variables in the da
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames',varnames); % Datastore
t = tall(ds); % Tall table
X = t{:,varnames}
```

```
X =

  Mx4 tall double matrix

           8         735         642          53
           8        1124        1021          63
          21        2218        2055          83
          13        1431        1332          59
           4         746         629          77
          59        1547        1446          61
           3        1052         928          84
          11        1134         859         155
          :           :           :           :
          :           :           :           :
```

To find evenly spaced quantiles along the first dimension, you must use the approximation algorithm based on “T-Digest” on page 33-5102. Using the sorting-based algorithm (see “Algorithms” on page 33-5103) to find quantiles along the first dimension of a tall array is computationally intensive.

Calculate three evenly spaced quantiles along the first dimension of `X`. Because the default dimension is 1, you do not need to specify a value for `dim`. Specify `'Method','approximate'` to use the approximation algorithm.

```
N = 3; % Number of quantiles
Yapprox = quantile(X,N,'Method','approximate')
```

```
Yapprox =
```

```
  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

To find evenly spaced quantiles along any other dimension (`dim` is not `1`), `quantile` calculates the exact quantiles only, so that it can perform the computation efficiently by using the sorting-based algorithm.

Calculate three evenly spaced quantiles along the second dimension of `X`. Because `dim` is not 1, `quantile` returns the exact quantiles by default.

```
Yexact = quantile(X,N,2)
```

```
Yexact =

  MxNx... tall double array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[Yapprox,Yexact] = gather(Yapprox,Yexact);
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 4.1 sec
Evaluation completed in 5.4 sec
```

Show the approximate quantiles of `X` (along the first dimension) for the three evenly spaced cumulative probabilities.

```
Yapprox
```

```
Yapprox = 3×4
10^3 ×

   -0.0070    1.1150    0.9322    0.0700
         0    1.5220    1.3350    0.1020
    0.0110    1.9180    1.7400    0.1510
```

Each column of the matrix `Yapprox` corresponds to the three evenly spaced quantiles for each column of the matrix `X`. For example, the first column of `Yapprox` with elements (–7, 0, 11) contains the quantiles for the first column of `X`.

Show the first five rows of the exact quantiles of `X` (along the second dimension) for the three evenly spaced cumulative probabilities.

```
Yexact(1:5,:)
```

```
ans = 5×3
10^3 ×
```

```
    0.0305    0.3475    0.6885
    0.0355    0.5420    1.0725
    0.0520    1.0690    2.1365
    0.0360    0.6955    1.3815
    0.0405    0.3530    0.6875
```

Each row of the matrix `Yexact` contains the three evenly spaced quantiles of the corresponding row in `X`. For example, `30.5`, `347.5`, and `688.5` are the 0.25, 0.5, and 0.75 quantiles, respectively, of the first row in `X`.

Input Arguments

`X` — Input data

vector | array

Input data, specified as a vector or array.

Data Types: `double` | `single`

`p` — Cumulative probabilities

scalar | vector

Cumulative probabilities for which to compute the quantiles, specified as a scalar or vector of scalars from 0 to 1.

Example: 0.3

Example: [0.25, 0.5, 0.75]

Example: (0:0.25:1)

Data Types: `double` | `single`

`N` — Number of quantiles

positive integer

Number of quantiles to compute, specified as a positive integer. `quantile` returns `N` quantiles that divide the data set into evenly distributed `N`+1 segments.

Data Types: `double` | `single`

`dim` — Dimension

positive integer

Dimension along which the quantiles of a matrix `X` are requested, specified as a positive integer. For example, for a matrix `X`, when `dim` = 1, `quantile` returns the quantile(s) of the columns of `X`; when `dim` = 2, `quantile` returns the quantile(s) of the rows of `X`. For a multidimensional array `X`, the length of the `dim`th dimension of `Y` is the same as the length of `p`.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions

positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. In the smallest specified operating dimension (that is, dimension

`min(vecdim)`), the output `Y` has length equal to the number of quantiles requested (either `N` or `length(p)`). In each of the remaining operating dimensions, `Y` has length 1. The other dimension lengths are the same for `X` and `Y`.

For example, consider a 2-by-3-by-3 array `X` with `p = [0.2 0.4 0.6 0.8]`. In this case, `quantile(X,p,[1 2])` returns an array, where each page of the array contains the 0.2, 0.4, 0.6, and 0.8 quantiles of the elements on the corresponding page of `X`. Because 1 and 2 are the operating dimensions, with `min([1 2]) = 1` and `length(p) = 4`, the output is a 4-by-1-by-3 array.

Data Types: `single` | `double`

`method` — Method for calculating quantiles
`'exact'` (default) | `'approximate'`

Method for calculating quantiles, specified as `'exact'` or `'approximate'`. By default, `quantile` returns the exact quantiles by implementing an algorithm on page 33-5103 that uses sorting. You can specify `'method','approximate'` for `quantile` to return approximate quantiles by implementing an algorithm that uses T-Digest on page 33-5102.

Data Types: `char` | `string`

Output Arguments

`Y` — Quantiles
scalar | array

Quantiles of a data vector or array, returned as a scalar or array for one or multiple values of cumulative probabilities.

- If `X` is a vector, then `Y` is a scalar or a vector with the same length as the number of quantiles requested (`N` or `length(p)`). `Y(i)` contains the `p(i)` quantile.
- If `X` is an array of dimension *d*, then `Y` is an array with the length of the smallest operating dimension equal to the number of quantiles requested (`N` or `length(p)`).

More About

Multidimensional Array

A *multidimensional array* is an array with more than two dimensions. For example, if `X` is a 1-by-3-by-4 array, then `X` is a 3-D array.

Nonsingleton Dimension

A *nonsingleton dimension* of an array is a dimension whose size is not equal to 1. A *first nonsingleton dimension* of an array is the first dimension that satisfies the nonsingleton condition. For example, if `X` is a 1-by-1-by-2-by-4 array, then the third dimension is the first nonsingleton dimension of `X`.

Linear Interpolation

Linear interpolation uses linear polynomials to find $y_i = f(x_i)$, the values of the underlying function $Y = f(X)$ at the points in the vector or array x. Given the data points (x_1, y_1) and (x_2, y_2), where $y_1 = f(x_1)$ and $y_2 = f(x_2)$, linear interpolation finds $y = f(x)$ for a given x between x_1 and x_2 as follows:

$$y = f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1).$$

Similarly, if the $1.5/n$ quantile is $y_{1.5/n}$ and the $2.5/n$ quantile is $y_{2.5/n}$, then linear interpolation finds the $2.3/n$ quantile $y_{2.3/n}$ as

$$y_{\frac{2.3}{n}} = y_{\frac{1.5}{n}} + \frac{\left(\frac{2.3}{n} - \frac{1.5}{n}\right)}{\left(\frac{2.5}{n} - \frac{1.5}{n}\right)}\left(y_{\frac{2.5}{n}} - y_{\frac{1.5}{n}}\right).$$

T-Digest

T-digest[2] on page 33-5104 is a probabilistic data structure that is a sparse representation of the empirical cumulative distribution function (CDF) of a data set. T-digest is useful for computing approximations of rank-based statistics (such as percentiles and quantiles) from online or distributed data in a way that allows for controllable accuracy, particularly near the tails of the data distribution.

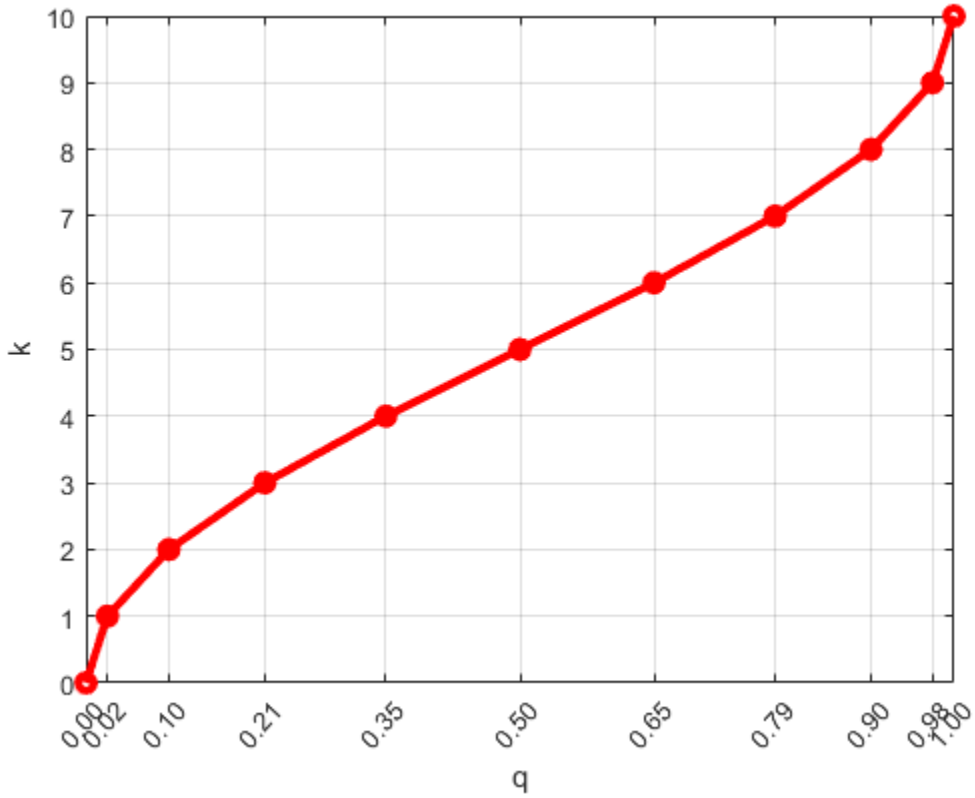
For data that is distributed in different partitions, t-digest computes quantile estimates (and percentile estimates) for each data partition separately, and then combines the estimates while maintaining a constant-memory bound and constant relative accuracy of computation ($q(1 - q)$ for the qth quantile). For these reasons, t-digest is practical for working with tall arrays.

To estimate quantiles of an array that is distributed in different partitions, first build a t-digest in each partition of the data. A t-digest clusters the data in the partition and summarizes each cluster by a centroid value and an accumulated weight that represents the number of samples contributing to the cluster. T-digest uses large clusters (widely spaced centroids) to represent areas of the CDF that are near `q = 0.5` and uses small clusters (tightly spaced centroids) to represent areas of the CDF that are near `q = 0` or `q = 1`.

T-digest controls the cluster size by using a scaling function that maps a quantile q to an index k with a compression parameter δ. That is,

$$k(q, \delta) = \delta \cdot \left(\frac{\sin^{-1}(2q - 1)}{\pi} + \frac{1}{2}\right),$$

where the mapping k is monotonic with minimum value $k(0,\delta) = 0$ and maximum value $k(1,\delta) = \delta$. The following figure shows the scaling function for $\delta = 10$.



The scaling function translates the quantile q to the scaling factor k in order to give variable size steps in q. As a result, cluster sizes are unequal (larger around the center quantiles and smaller near $q = 0$ or $q = 1$). The smaller clusters allow for better accuracy near the edges of the data.

To update a t-digest with a new observation that has a weight and location, find the cluster closest to the new observation. Then, add the weight and update the centroid of the cluster based on the weighted average, provided that the updated weight of the cluster does not exceed the size limitation.

You can combine independent t-digests from each partition of the data by taking a union of the t-digests and merging their centroids. To combine t-digests, first sort the clusters from all the independent t-digests in decreasing order of cluster weights. Then, merge neighboring clusters, when they meet the size limitation, to form a new t-digest.

Once you form a t-digest that represents the complete data set, you can estimate the end-points (or boundaries) of each cluster in the t-digest and then use interpolation between the end-points of each cluster to find accurate quantile estimates.

Algorithms

For an n-element vector `X`, `quantile` computes quantiles by using a sorting-based algorithm as follows:

1 The sorted elements in `X` are taken as the $(0.5/n)$, $(1.5/n)$, ..., $([n - 0.5]/n)$ quantiles. For example:

- For a data vector of five elements such as {6, 3, 2, 10, 1}, the sorted elements {1, 2, 3, 6, 10} respectively correspond to the 0.1, 0.3, 0.5, 0.7, 0.9 quantiles.

- For a data vector of six elements such as {6, 3, 2, 10, 8, 1}, the sorted elements {1, 2, 3, 6, 8, 10} respectively correspond to the (0.5/6), (1.5/6), (2.5/6), (3.5/6), (4.5/6), (5.5/6) quantiles.

2 `quantile` uses "Linear Interpolation" on page 33-5101 to compute quantiles for probabilities between (0.5/*n*) and ([*n* – 0.5]/*n*).

3 For the quantiles corresponding to the probabilities outside that range, `quantile` assigns the minimum or maximum values of the elements in `X`.

`quantile` treats `NaN`s as missing values and removes them.

References

[1] Langford, E. "Quartiles in Elementary Statistics", *Journal of Statistics Education*. Vol. 14, No. 3, 2006.

[2] Dunning, T., and O. Ertl. "Computing Extremely Accurate Quantiles Using T-Digests." August 2017.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `Y = quantile(X,p)` and `Y = quantile(X,N)` return the exact quantiles (using a sorting-based algorithm on page 33-5103) only if `X` is a tall column vector.
- `Y = quantile(__,dim)` returns the exact quantiles only when *one* of these conditions exists:
  - `X` is a tall column vector.
  - `X` is a tall array and `dim` is not `1`. For example, `quantile(X,p,2)` returns the exact quantiles along the rows of the tall array `X`.

  If `X` is a tall array and `dim` is `1`, then you must specify `'Method','approximate'` to use an approximation algorithm based on T-Digest on page 33-5102 for computing the quantiles. For example, `quantile(X,p,1,'Method','approximate')` returns the approximate quantiles along the columns of the tall array `X`.
- `Y = quantile(__,vecdim)` returns the exact quantiles only when *one* of these conditions exists:
  - `X` is a tall column vector.
  - `X` is a tall array and `vecdim` does not include `1`. For example, if `X` is a 3-by-5-by-2 array, then `quantile(X,p,[2,3])` returns the exact quantiles of the elements in each `X(i,:,:)` slice.
  - `X` is a tall array and `vecdim` includes `1` and all the nonsingleton dimensions on page 33-5101 of `X`. For example, if `X` is a 10-by-1-by-4 array, then `quantile(X,p,[1 3])` returns the exact quantiles of the elements in `X(:,1,:)`.

  If `X` is a tall array and `vecdim` includes `1` but does not include all the nonsingleton dimensions of `X`, then you must specify `'Method','approximate'` to use the approximation algorithm. For example, if `X` is a 10-by-1-by-4 array, you can use `quantile(X,p,[1 2],'Method','approximate')` to find the approximate quantiles of each page of `X`.

For more information, see “Tall Arrays”.

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `'Method'` name-value pair argument is not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see “Automatic dimension restriction” (MATLAB Coder).
- If the output `Y` is a vector, the orientation of `Y` differs from MATLAB when all of the following are true:
  - You do not supply `dim`.
  - `X` is a variable-size array, and not a variable-size vector, at compile time, but `X` is a vector at run time.
  - The orientation of the vector `X` does not match the orientation of the vector `p`.

  In this case, the output `Y` matches the orientation of `X`, not the orientation of `p`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `'Method'` name-value pair argument is not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also
`iqr` | `median` | `prctile`

Topics
“Quantiles and Percentiles” on page 3-6

Introduced before R2006a

rand

Class: `qrandstream`

Generate quasi-random points from stream

Syntax

```
rand
rand(q,n)
rand(q)
rand(q,m,n)
rand(q,[m,n])
rand(q,m,n,p,...)
rand(q,[m,n,p,...])
```

Description

`rand` returns a matrix of quasi-random values and is intended to allow objects of the `qrandstream` on page 33-5078 class to be used in code that contains calls to the `rand` method of the MATLAB pseudo-random `randstream` class. Due to the multidimensional nature of quasi-random numbers, only some syntaxes of `rand` are supported by the `qrandstream` class.

`rand(q,n)` returns an `n`-by-`n` matrix only when `n` is equal to the number of dimensions. Any other value of `n` produces an error.

`rand(q)` returns a scalar only when the stream is in one dimension. Having more than one dimension in `q` produces an error.

`rand(q,m,n)` or `rand(q,[m,n])` returns an `m`-by-`n` matrix only when `n` is equal to the number of dimensions in the stream. Any other value of `n` produces an error.

`rand(q,m,n,p,...)` or `rand(q,[m,n,p,...])` produces an error unless `p` and all following dimensions sizes are equal to one.

Examples

Generate the first 256 points from a 5-D Sobol sequence:

```
q = qrandstream('sobol',5);
X = rand(q,256,5);
```

See Also

`qrand` | `qrandstream` | `rand`

quantileError

Class: `TreeBagger`

Quantile loss using bag of regression trees

Syntax

```
err = quantileError(Mdl,X)
err = quantileError(Mdl,X,ResponseVarName)
err = quantileError(Mdl,X,Y)
err = quantileError( ___ ,Name,Value)
```

Description

`err = quantileError(Mdl,X)` returns half of the mean absolute deviation (MAD) from comparing the true responses in the table `X` to the predicted medians resulting from applying the bag of regression trees `Mdl` to the observations of the predictor data in `X`.

- `Mdl` must be a `TreeBagger` model object.
- The response variable name in `X` must have the same name as the response variable in the table containing the training data.

`err = quantileError(Mdl,X,ResponseVarName)` uses the true response and predictor variables contained in the table `X`. `ResponseVarName` is the name of the response variable and `Mdl.PredictorNames` contain the names of the predictor variables.

`err = quantileError(Mdl,X,Y)` uses the predictor data in the table or matrix `X` and the response data in the vector `Y`.

`err = quantileError( ___ ,Name,Value)` uses any of the previous syntaxes and additional options specified by one or more `Name,Value` pair arguments. For example, specify quantile probabilities, the error type, or which trees to include in the quantile-regression-error estimation.

Input Arguments

`Mdl` — Bag of regression trees

`TreeBagger` model object (default)

Bag of regression trees, specified as a `TreeBagger` model object created by `TreeBagger`. The value of `Mdl.Method` must be `regression`.

`X` — Sample data

numeric matrix | table

Sample data used to estimate quantiles, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable. If you specify `Y`, then the number of rows in `X` must be equal to the length of `Y`.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl` (stored in `Mdl.PredictorNames`).
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types), then `quantileError` throws an error.
  - Specify `Y` for the true responses.
- For a table:
  - `quantileError` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.).
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `TreeBagger`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.).
  - If `X` contains the response variable:
    - If the response variable has the same name as the response variable that trained `Mdl`, then you do not have to supply the response variable name or vector of true responses. `quantileError` uses that variable for the true responses by default.
    - You can specify `ResponseVarName` or `Y` for the true responses.

Data Types: `table` | `double` | `single`

`ResponseVarName` — Response variable name
character vector | string scalar

Response variable name, specified as a character vector or string scalar. `ResponseVarName` must be the name of the response variable in the table of sample data `X`.

If the table `X` contains the response variable, and it has the same name as the response variable used to train `Mdl`, then you do not have to specify `ResponseVarName`. `quantileError` uses that variable for the true responses by default.

Data Types: `char` | `string`

`Y` — True responses
numeric vector

True responses, specified as a numeric vector. The number of rows in `X` must be equal to the length of `Y`.

Data Types: `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Mode` — Ensemble error type

`'ensemble'` (default) | `'cumulative'` | `'individual'`

Ensemble error type, specified as the comma-separated pair consisting of `'Mode'` and a value in this table. Suppose *`tau`* is the value of `Quantile`.

Value	Description
`'cumulative'`	`err` is a `Mdl.NumTrees`-by-`numel(`*`tau`*`)` numeric matrix of cumulative quantile regression errors. `err(`*`j`*`,`*`k`*`)` is the *`tau`*`(`*`k`*`)` quantile regression error using the learners in `Mdl.Trees(1:`*`j`*`)` only.
`'ensemble'`	`err` is a 1-by-`numel(`*`tau`*`)` numeric vector of cumulative quantile regression errors for the entire ensemble. `err(`*`k`*`)` is the *`tau`*`(`*`k`*`)` ensemble quantile regression error.
`'individual'`	`err` is a `Mdl.NumTrees`-by-`numel(`*`tau`*`)` numeric matrix of quantile regression errors from individual learners. `err(`*`j`*`,`*`k`*`)` is the *`tau`*`(`*`k`*`)` quantile regression error using the learner in `Mdl.Trees(`*`j`*`)` only.

For `'cumulative'` and `'individual'`, if you include fewer trees in quantile estimation using `Trees` or `UseInstanceForTree`, then the number of rows in `err` decreases from `Mdl.NumTrees`.

Example: `'Mode','cumulative'`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | numeric vector of positive values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector of positive values with length equal to `size(X,1)`. `quantileError` uses `Weights` to compute the weighted average of the deviations when estimating the quantile regression error.

By default, `quantileError` attributes a weight of `1` to each observation, which yields an unweighted average of the deviations.

`Quantile` — Quantile probability

`0.5` (default) | numeric vector containing values in [0,1]

Quantile probability, specified as the comma-separated pair consisting of `'Quantile'` and a numeric vector containing values in the interval [0,1]. For each element in `Quantile`, `quantileError` returns corresponding quantile regression errors for all probabilities in `Quantile`.

Example: `'Quantile',[0 0.25 0.5 0.75 1]`

Data Types: `single` | `double`

`Trees` — Indices of trees to use in response estimation

`'all'` (default) | numeric vector of positive integers

Indices of trees to use in response estimation, specified as the comma-separated pair consisting of `'Trees'` and `'all'` or a numeric vector of positive integers. Indices correspond to the cells of

`Mdl.Trees`; each cell therein contains a tree in the ensemble. The maximum value of `Trees` must be less than or equal to the number of trees in the ensemble (`Mdl.NumTrees`).

For `'all'`, `quantileError` uses all trees in the ensemble (that is, the indices `1:Mdl.NumTrees`).

Values other than the default can affect the number of rows in `err`.

Example: `'Trees',[1 10 Mdl.NumTrees]`

Data Types: `char` | `string` | `single` | `double`

`TreeWeights` — Weights to attribute to responses from individual trees
`ones(Mdl.NumTrees,1)` (default) | numeric vector of nonnegative values

Weights to attribute to responses from individual trees, specified as the comma-separated pair consisting of `'TreeWeights'` and a numeric vector of `numel(`*`trees`*`)` nonnegative values. *`trees`* is the value of `Trees`.

If you specify `'Mode','individual'`, then `quantileError` ignores `TreeWeights`.

Data Types: `single` | `double`

`UseInstanceForTree` — Indicators specifying which trees to use to make predictions for each observation
`'all'` (default) | logical matrix

Indicators specifying which trees to use to make predictions for each observation, specified as the comma-separated pair consisting of `'UseInstanceForTree'` and an *n*-by-`Mdl.Trees` logical matrix. *n* is the number of observations (rows) in `X`. Rows of `UseInstanceForTree` correspond to observations and columns correspond to learners in `Mdl.Trees`. `'all'` indicates to use all trees for all observations when estimating the quantiles.

If `UseInstanceForTree(`*`j`*`,`*`k`*`) = true`, then `quantileError` uses the tree in `Mdl.Trees(`*`k`*`)` when it predicts the response for the observation `X(`*`j`*`,:)`.

You can estimate quantiles using the response data in `Mdl.Y` directly instead of using the predictions from the random forest by specifying a row composed entirely of `false` values. For example, to estimate the quantile for observation *`j`* using the response data, and to use the predictions from the random forest for all other observations, specify this matrix:

```
UseInstanceForTree = true(size(Mdl.X,2),Mdl.NumTrees);
UseInstanceForTree(j,:) = false(1,Mdl.NumTrees);
```

Values other than the default can affect the number of rows in `err`. Also, the value of `Trees` affects the value of `UseInstanceForTree`. Suppose that *`U`* is the value of `UseInstanceForTree`. `quantileError` ignores the columns of *`U`* corresponding to trees not being used in estimation from the specification of `Trees`. That is, `quantileError` resets the value of `'UseInstanceForTree'` to `U(:,`*`trees`*`)`, where *`trees`* is the value of `'Trees'`.

Data Types: `char` | `string` | `logical`

Output Arguments

`err` — Half of quantile regression error
numeric scalar | numeric matrix

Half of the quantile regression error on page 33-5113, returned as a numeric scalar or *T*-by-`numel(tau)` matrix. *`tau`* is the value of `Quantile`.

T depends on the values of `Mode`, `Trees`, `UseInstanceForTree`, and `Quantile`. Suppose that you specify `'Trees',trees` and you use the default value of `'UseInstanceForTree'`.

- For `'Mode','cumulative'`, `err` is a `numel(trees)`-by-`numel(tau)` numeric matrix. `err(j,k)` is the `tau(k)` cumulative quantile regression error using the learners in `Mdl.Trees(trees(1:j))`.
- For `'Mode','ensemble'`, `err` is a 1-by-`numel(tau)` numeric vector. `err(k)` is the `tau(k)` cumulative quantile regression error using the learners in `Mdl.Trees(trees)`.
- For `'Mode','individual'`, `err` is a `numel(trees)`-by-`numel(tau)` numeric matrix. `err(j,k)` is the `tau(k)` quantile regression error using the learner in `Mdl.Trees(trees(j))`.

Examples

Estimate In-Sample Quantile Regression Error

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement, weight, and number of cylinders. Consider `Cylinders` a categorical variable.

```
load carsmall
Cylinders = categorical(Cylinders);
X = table(Displacement,Weight,Cylinders,MPG);
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,X,'MPG','Method','regression');
```

`Mdl` is a `TreeBagger` ensemble.

Perform quantile regression, and estimate the MAD of the entire ensemble using the predicted conditional medians.

```
err = quantileError(Mdl,X)
```

```
err = 1.2339
```

Because `X` is a table containing the response and commensurate variable names, you do not have to specify the response variable name or data. However, you can specify the response using this syntax.

```
err = quantileError(Mdl,X,'MPG')
```

```
err = 1.2339
```

Find Appropriate Ensemble Size Using Quantile Regression Error

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement, weight, and number of cylinders.

```
load carsmall
X = table(Displacement,Weight,Cylinders,MPG);
```

Randomly split the data into two sets: 75% training and 25% testing. Extract the subset indices.

```
rng(1); % For reproducibility
cvp = cvpartition(size(X,1),'Holdout',0.25);
idxTrn = training(cvp);
idxTest = test(cvp);
```

Train an ensemble of bagged regression trees using the training set. Specify 250 weak learners.

```
Mdl = TreeBagger(250,X(idxTrn,:),'MPG','Method','regression');
```

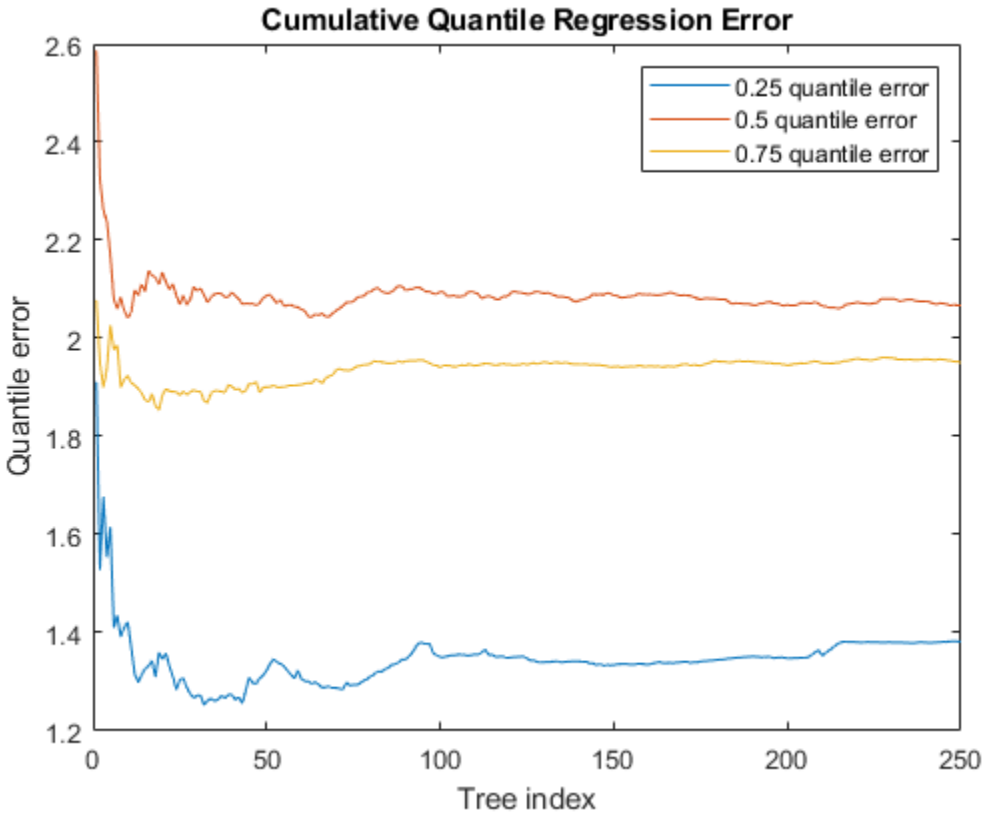
Estimate the cumulative 0.25, 0.5, and 0.75 quantile regression errors for the test set. Pass the predictor data in as a numeric matrix, and the response data in as a vector.

```
err = quantileError(Mdl,X{idxTest,1:3},MPG(idxTest),'Quantile',[0.25 0.5 0.75],...
    'Mode','cumulative');
```

`err` is a 250-by-3 matrix of cumulative quantile regression errors. Columns correspond to quantile probabilities and rows correspond to trees in the ensemble. The errors are cumulative, so they incorporate aggregated predictions from previous trees. Although, `Mdl` was trained using a table, if all predictor variables in the table are numeric, then you can supply a matrix of predictor data instead.

Plot the cumulative quantile errors on the same plot.

```
figure;
plot(err);
legend('0.25 quantile error','0.5 quantile error','0.75 quantile error');
ylabel('Quantile error');
xlabel('Tree index');
title('Cumulative Quantile Regression Error')
```



Training using about 60 trees appears to be enough for the first two quartiles, but the third quartile requires about 150 trees.

More About

Quantile Regression Error

The quantile regression error of a model given observed predictor data and responses is the weighted mean absolute deviation (MAD). If the model under-predicts the response, then deviation weights are τ, the quantile probability. If the model over-predicts, then deviation weights are 1 – τ.

That is, the τ quantile regression error is

$$L_\tau = \tau \frac{\sum_{\{j:y_j \geq \hat{y}_{\tau,j}\}} w_j(y_j - \hat{y}_{\tau,j})}{\sum_{j=1}^{n} w_j} + (1-\tau)\frac{\sum_{\{j:y_j < \hat{y}_{\tau,j}\}} w_j(\hat{y}_{\tau,j} - y_j)}{\sum_{j=1}^{n} w_j}.$$

y_j is true response j, $\hat{y}_{\tau,j}$ is the τ quantile that the model predicts, and w_j is observation weight j.

Tips

- To tune the number of trees in the ensemble, set `'Mode','cumulative'` and plot the quantile regression errors with respect to tree indices. The maximal number of required trees is the tree index where the quantile regression error appears to level off.
- To investigate the performance of a model when the training sample is small, use `oobQuantileError` instead.

References

[1] Breiman, L. *Random Forests.* Machine Learning 45, pp. 5–32, 2001.

[2] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

See Also

`TreeBagger` | `error` | `oobQuantileError` | `quantilePredict`

Introduced in R2016b

quantilePredict

Class: `TreeBagger`

Predict response quantile using bag of regression trees

Syntax

```
YFit = quantilePredict(Mdl,X)
YFit = quantilePredict(Mdl,X,Name,Value)
[YFit,YW] = quantilePredict( ___ )
```

Description

`YFit = quantilePredict(Mdl,X)` returns a vector of medians of the predicted responses at `X`, a table or matrix of predictor data, and using the bag of regression trees `Mdl`. `Mdl` must be a `TreeBagger` model object.

`YFit = quantilePredict(Mdl,X,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, specify quantile probabilities or which trees to include for quantile estimation.

`[YFit,YW] = quantilePredict( ___ )` also returns a sparse matrix of response weights on page 33-5122.

Input Arguments

`Mdl` — Bag of regression trees

`TreeBagger` model object (default)

Bag of regression trees, specified as a `TreeBagger` model object created by `TreeBagger`. The value of `Mdl.Method` must be `regression`.

`X` — Predictor data

numeric matrix | table

Predictor data used to estimate quantiles, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables making up the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`), then `X` can be a numeric matrix if `Tbl` contains all numeric predictor variables. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `quantilePredict` throws an error.
- For a table:

- `quantilePredict` does not support multicolumn variables and cell arrays other than cell arrays of character vectors.
- If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those variables that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. `Tbl` and `X` can contain additional variables (response variables, observation weights, etc.), but `quantilePredict` ignores them.
- If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `TreeBagger`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, etc.), but `quantilePredict` ignores them.

Data Types: `table` | `double` | `single`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Quantile` — Quantile probability
`0.5` (default) | numeric vector containing values in [0,1]

Quantile probability, specified as the comma-separated pair consisting of `'Quantile'` and a numeric vector containing values in the interval [0,1]. For each observation (row) in `X`, `quantilePredict` returns corresponding quantiles for all probabilities in `Quantile`.

Example: `'Quantile',[0 0.25 0.5 0.75 1]`

Data Types: `single` | `double`

`Trees` — Indices of trees to use in response estimation
`'all'` (default) | numeric vector of positive integers

Indices of trees to use in response estimation, specified as the comma-separated pair consisting of `'Trees'` and `'all'` or a numeric vector of positive integers. Indices correspond to the cells of `Mdl.Trees`; each cell therein contains a tree in the ensemble. The maximum value of `Trees` must be less than or equal to the number of trees in the ensemble (`Mdl.NumTrees`).

For `'all'`, `quantilePredict` uses the indices `1:Mdl.NumTrees`.

Example: `'Trees',[1 10 Mdl.NumTrees]`

Data Types: `char` | `string` | `single` | `double`

`TreeWeights` — Weights to attribute to responses from individual trees
numeric vector of nonnegative values

Weights to attribute to responses from individual trees, specified as the comma-separated pair consisting of `'TreeWeights'` and a numeric vector of `numel(`*`trees`*`)` nonnegative values. *`trees`* is the value of the `Trees` name-value pair argument.

The default is `ones(size(`*`trees`*`))`.

Data Types: `single` | `double`

`UseInstanceForTree` — Indicators specifying which trees to use to make predictions for each observation
`'all'` (default) | logical matrix

Indicators specifying which trees to use to make predictions for each observation, specified as the comma-separated pair consisting of `'UseInstanceForTree'` and an *n*-by-`Mdl.Trees` logical matrix. *n* is the number of observations (rows) in `X`. Rows of `UseInstanceForTree` correspond to observations and columns correspond to learners in `Mdl.Trees`. `'all'` indicates to use all trees for all observations when estimating the quantiles.

If `UseInstanceForTree(j,k) = true`, then `quantilePredict` uses the tree in `Mdl.Trees(trees(k))` when it predicts the response for the observation `X(j,:)`.

You can estimate the quantile using the response data in `Mdl.Y` directly instead of using the predictions from the random forest by specifying a row composed entirely of `false` values. For example, to estimate the quantile for observation *j* using the response data, and to use the predictions from the random forest for all other observations, specify this matrix:

```
UseInstanceForTree = true(size(Mdl.X,2),Mdl.NumTrees);
UseInstanceForTree(j,:) = false(1,Mdl.NumTrees);
```

Data Types: `char` | `string` | `logical`

Output Arguments

`YFit` — Estimated quantiles
numeric matrix

Estimated quantiles, returned as an *n*-by-`numel(tau)` numeric matrix. *n* is the number of observations in `X` (`size(X,1)`) and *tau* is the value of `Quantile`. That is, `YFit(j,k)` is the estimated `100*tau(k)`% percentile of the response distribution given `X(j,:)` and using `Mdl`.

`YW` — Response weights
sparse matrix

Response weights on page 33-5122, returned as an n_{train}-by-*n* sparse matrix. n_{train} is the number of responses in the training data (`numel(Mdl.Y)`) and *n* is the number of observations in `X` (`size(X,1)`).

`quantilePredict` predicts quantiles using linear interpolation of the empirical cumulative distribution function (C.D.F.). For a particular observation, you can use its response weights to estimate quantiles using alternative methods, such as approximating the C.D.F. using kernel smoothing.

Note `quantilePredict` derives response weights by passing an observation through the trees in the ensemble. If you specify `UseInstanceForTree` and you compose row *j* entirely of `false` values, then `YW(:,j) = Mdl.W` instead, that is, the observation weights.

Examples

Predict Training Sample Medians

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression');
```

`Mdl` is a `TreeBagger` ensemble.

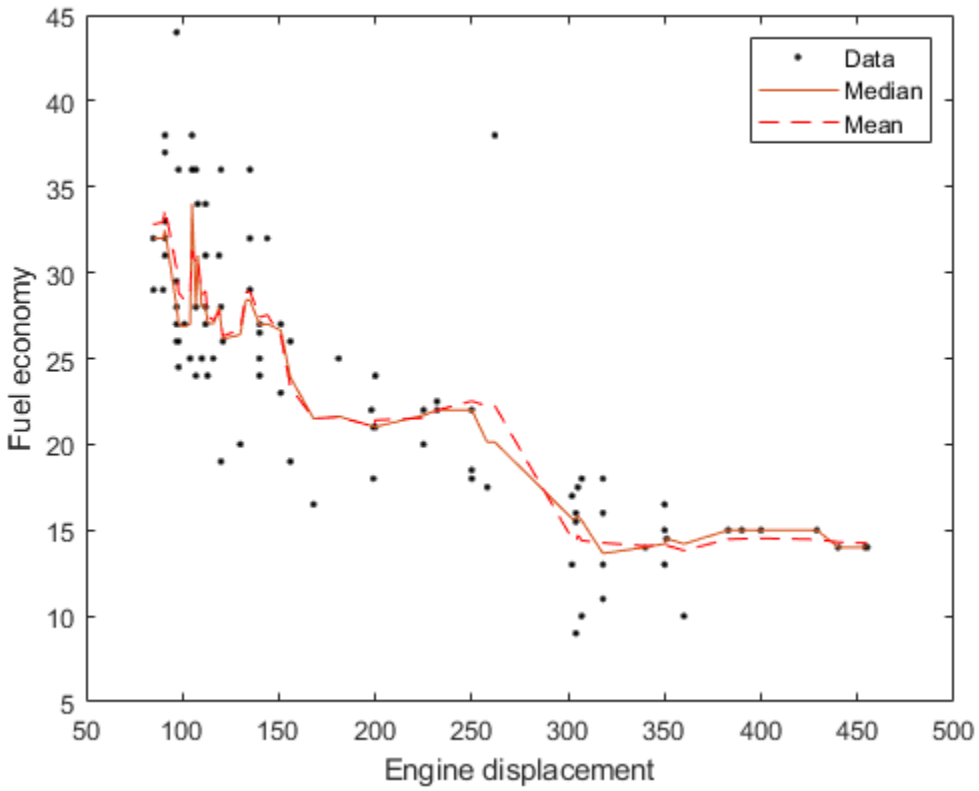
Perform quantile regression to predict the median MPG for all sorted training observations.

```
medianMPG = quantilePredict(Mdl,sort(Displacement));
```

`medianMPG` is an `n`-by-1 numeric vector of medians corresponding to the conditional distribution of the response given the sorted observations in `Displacement`. `n` is the number of observations in `Displacement`.

Plot the observations and the estimated medians on the same figure. Compare the median and mean responses.

```
meanMPG = predict(Mdl,sort(Displacement));

figure;
plot(Displacement,MPG,'k.');
hold on
plot(sort(Displacement),medianMPG);
plot(sort(Displacement),meanMPG,'r--');
ylabel('Fuel economy');
xlabel('Engine displacement');
legend('Data','Median','Mean');
hold off;
```



Estimate Prediction Intervals Using Percentiles

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression');
```

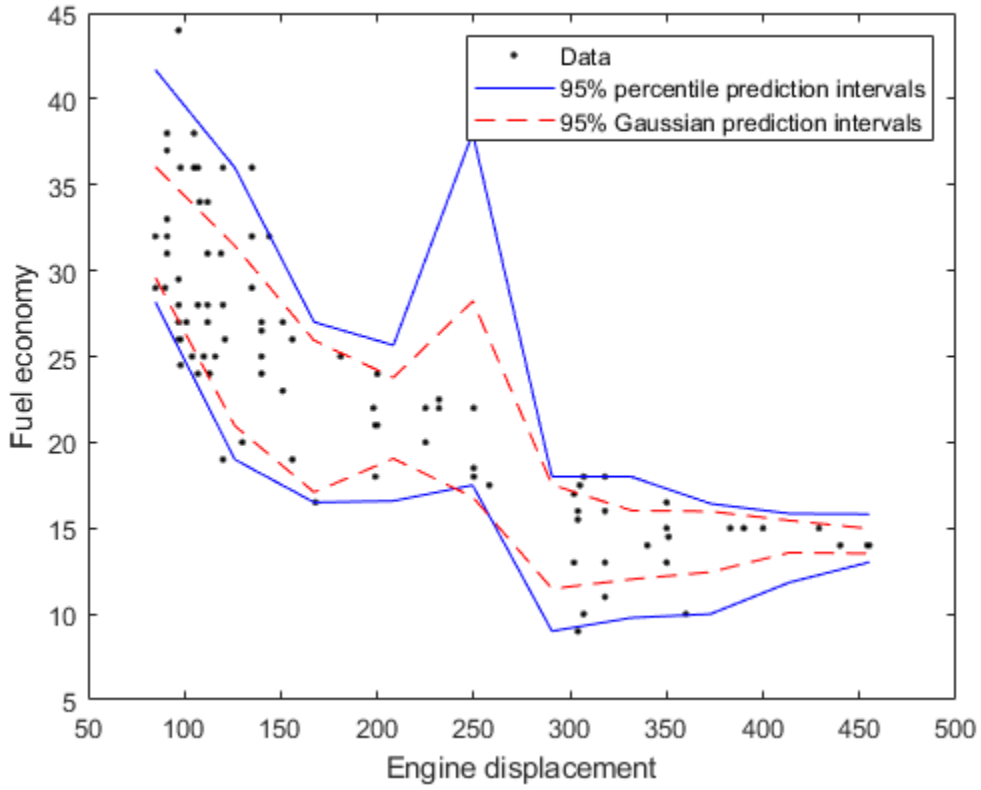
Perform quantile regression to predict the 2.5% and 97.5% percentiles for ten equally-spaced engine displacements between the minimum and maximum in-sample displacement.

```
predX = linspace(min(Displacement),max(Displacement),10)';
quantPredInts = quantilePredict(Mdl,predX,'Quantile',[0.025,0.975]);
```

`quantPredInts` is a 10-by-2 numeric matrix of prediction intervals corresponding to the observations in `predX`. The first column contains the 2.5% percentiles and the second column contains the 97.5% percentiles.

Plot the observations and the estimated medians on the same figure. Compare the percentile prediction intervals and the 95% prediction intervals assuming the conditional distribution of `MPG` is Gaussian.

```
[meanMPG,steMeanMPG] = predict(Mdl,predX);
stndPredInts = meanMPG + [-1 1]*norminv(0.975).*steMeanMPG;

figure;
h1 = plot(Displacement,MPG,'k.');
hold on
h2 = plot(predX,quantPredInts,'b');
h3 = plot(predX,stndPredInts,'r--');
ylabel('Fuel economy');
xlabel('Engine displacement');
legend([h1,h2(1),h3(1)],{'Data','95% percentile prediction intervals',...
    '95% Gaussian prediction intervals'});
hold off;
```



Estimate Conditional Cumulative Distribution Using Quantile Regression

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement.

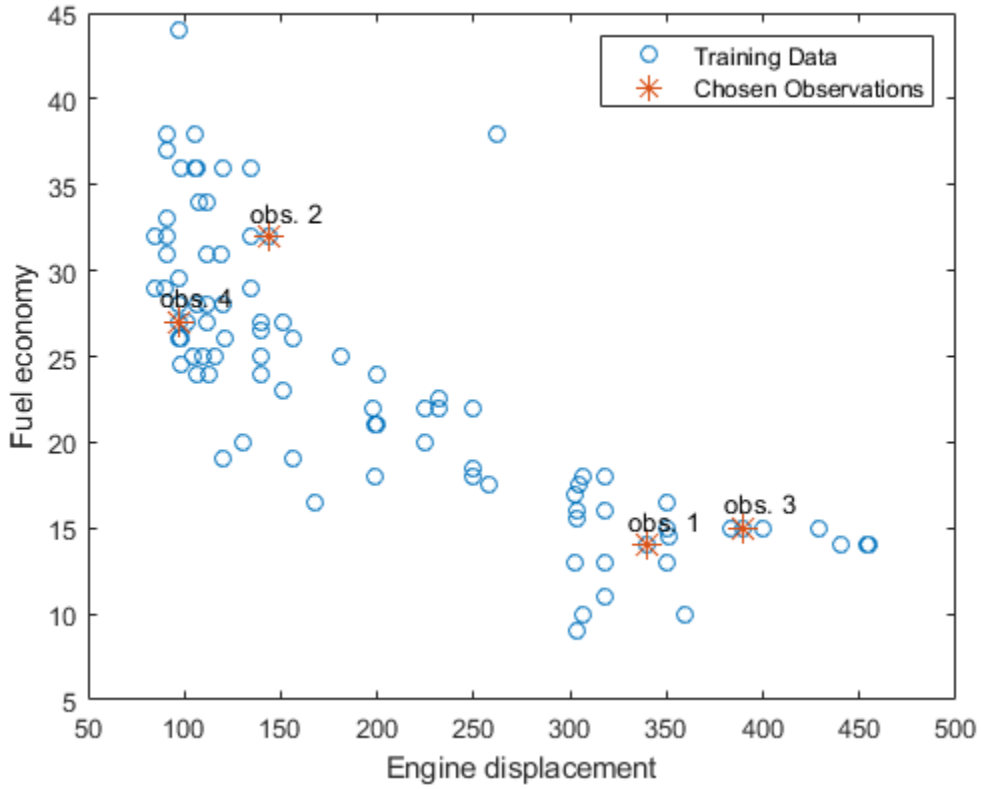
```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression');
```

Estimate the response weights for a random sample of four training observations. Plot the training sample and identify the chosen observations.

```
[predX,idx] = datasample(Mdl.X,4);
[~,YW] = quantilePredict(Mdl,predX);
n = numel(Mdl.Y);

figure;
plot(Mdl.X,Mdl.Y,'o');
hold on
plot(predX,Mdl.Y(idx),'*','MarkerSize',10);
text(predX-10,Mdl.Y(idx)+1.5,{'obs. 1' 'obs. 2' 'obs. 3' 'obs. 4'});
legend('Training Data','Chosen Observations');
xlabel('Engine displacement')
ylabel('Fuel economy')
hold off
```



`YW` is an `n`-by-4 sparse matrix containing the response weights. Columns correspond to test observations and rows correspond to responses in the training sample. Response weights are independent of the specified quantile probability.

Estimate the conditional cumulative distribution function (C.C.D.F.) of the responses by:

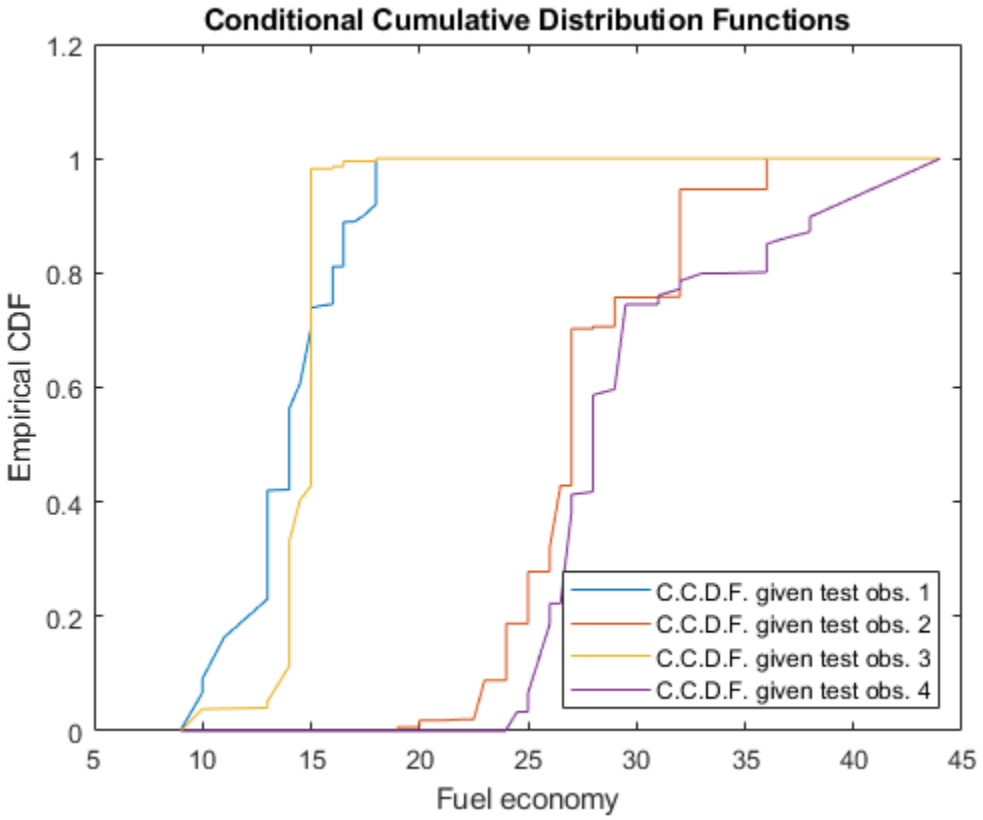
1. Sorting the responses is ascending order, and then sorting the response weights using the indices induced by sorting the responses.
2. Computing the cumulative sums over each column of the sorted response weights.

```
[sortY,sortIdx] = sort(Mdl.Y);
cpdf = full(YW(sortIdx,:));
ccdf = cumsum(cpdf);
```

`ccdf(:,j)` is the empirical C.C.D.F. of the response given test observation `j`.

Plot the four empirical C.C.D.F. in the same figure.

```
figure;
plot(sortY,ccdf);
legend('C.C.D.F. given test obs. 1','C.C.D.F. given test obs. 2',...
    'C.C.D.F. given test obs. 3','C.C.D.F. given test obs. 4',...
    'Location','SouthEast')
title('Conditional Cumulative Distribution Functions')
xlabel('Fuel economy')
ylabel('Empirical CDF')
```



More About

Response Weights

Response weights are scalars that represent the conditional distribution of the response given a value in the predictor space. The observations in the bootstrap samples and the leaves that the training and test observations share induce response weights.

Given the observation x, the response weight for observation j in the training sample using tree t in the ensemble is

$$w_{tj}(x) = \frac{I\{X_j \in S_t(x)\}}{\sum_{k=1}^{n_{\text{train}}} I\{X_k \in S_t(x)\}},$$

where:

- $I\{h\}$ is the indicator function.
- $S_t(x)$ is the leaf of tree t containing x.
- n_{train} is the number of training observations.

In other words, the response weights of a particular tree form the conditional relative frequency distribution of the response.

The response weights for the entire ensemble are averaged over the trees:

$$w_j^*(x) = \frac{1}{T}\sum_{t=1}^{T} w_{tj}(x).$$

Quantile Random Forest

Quantile random forest [2] is a quantile-regression method that uses a random forest [1] of regression trees to model the conditional distribution of a response variable, given the value of predictor variables. You can use a fitted model to estimate quantiles in the conditional distribution of the response.

Besides quantile estimation, you can use quantile regression to estimate prediction intervals or detect outliers. For example:

- To estimate 95% quantile prediction intervals, estimate the 0.025 and 0.975 quantiles.
- To detect outliers, estimate the 0.01 and 0.99 quantiles. All observations smaller than the 0.01 quantile and larger than the 0.99 quantile are outliers. All observations that are outside the interval $[L,U]$ can be considered outliers:

  $L = Q_1 - 1.5*IQR$

  and

  $U = Q_3 + 1.5*IQR,$

  where:

  - Q_1 is the 0.25 quantile.
  - Q_3 is the 0.75 quantile.
  - $IQR = Q_3 - Q_1$ (the interquartile range).

Tip

`quantilePredict` estimates the conditional distribution of the response using the training data every time you call it. To predict many quantiles efficiently, or quantiles for many observations

efficiently, you should pass X as a matrix or table of observations and specify all quantiles in a vector using the `Quantile` name-value pair argument. That is, avoid calling `quantilePredict` within a loop.

Algorithms

- `TreeBagger` grows a random forest of regression trees using the training data. Then, to implement quantile random forest on page 33-5123, `quantilePredict` predicts quantiles using the empirical conditional distribution of the response given an observation from the predictor variables. To obtain the empirical conditional distribution of the response:

  1. `quantilePredict` passes all the training observations in `Mdl.X` through all the trees in the ensemble, and stores the leaf nodes of which the training observations are members.
  2. `quantilePredict` similarly passes each observation in X through all the trees in the ensemble.
  3. For each observation in X, `quantilePredict`:

     a. Estimates the conditional distribution of the response by computing response weights on page 33-5122 for each tree.

     b. For observation k in X, aggregates the conditional distributions for the entire ensemble:

$$\widehat{F}(y|X=x_k)=\sum_{j=1}^{n}\sum_{t=1}^{T}\frac{1}{T}w_{tj}(x_k)I\{Y_j\le y\}.$$

        n is the number of training observations (`size(Y,1)`) and T is the number of trees in the ensemble (`Mdl.NumTrees`).
  4. For observation k in X, the τ quantile or, equivalently, the 100τ% percentile, is $Q_\tau(x_k)=\inf\{y:\widehat{F}(y|X=x_k)\ge\tau\}$.

- This process describes how `quantilePredict` uses all specified weights.

  1. For all training observations $j=1,\ldots,n$ and all chosen trees $t=1,\ldots,T$,

     `quantilePredict` attributes the product $v_{tj}=b_{tj}w_{j,\text{obs}}$ to training observation j (stored in `Mdl.X(j,:)` and `Mdl.Y(j)`). b_{tj} is the number of times observation j is in the bootstrap sample for tree t. $w_{j,\text{obs}}$ is the observation weight in `Mdl.W(j)`.
  2. For each chosen tree, `quantilePredict` identifies the leaves in which each training observation falls. Let $S_t(x_j)$ be the set of all observations contained in the leaf of tree t of which observation j is a member.
  3. For each chosen tree, `quantilePredict` normalizes all weights within a particular leaf to sum to 1, that is,

$$v_{tj}^{*}=\frac{v_{tj}}{\sum_{i\in S_t(x_j)}v_{ti}}.$$

  4. For each training observation and tree, `quantilePredict` incorporates tree weights ($w_{t,\text{tree}}$) specified by `TreeWeights`, that is, $w^{*}_{tj,\text{tree}}=w_{t,\text{tree}}v_{tj}^{*}$Trees not chosen for prediction have 0 weight.
  5. For all test observations $k=1,\ldots,K$ in X and all chosen trees $t=1,\ldots,T$`quantilePredict` predicts the unique leaves in which the observations fall, and then identifies all training

observations within the predicted leaves. `quantilePredict` attributes the weight u_{tj} such that

$$u_{tj} = \begin{cases} w^*_{tj,\text{tree}}; \text{ if } x_k \in S_t(x_j) \\ 0; \text{ otherwise} \end{cases}.$$

6 `quantilePredict` sums the weights over all chosen trees, that is,

$$u_j = \sum_{t=1}^{T} u_{tj}.$$

7 `quantilePredict` creates response weights by normalizing the weights so that they sum to 1, that is,

$$w^*_j = \frac{u_j}{\sum_{j=1}^{n} u_j}.$$

References

[1] Breiman, L. "Random Forests." *Machine Learning* 45, pp. 5–32, 2001.

[2] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

See Also

`TreeBagger` | `oobQuantilePredict` | `predict` | `quantileError`

Topics

“Detect Outliers Using Quantile Regression” on page 18-137

Introduced in R2016b

randg

Gamma random numbers with unit scale

Syntax

```
Y = randg
Y = randg(A)
Y = randg(A,m)
Y = randg(A,m,n,p,...)
Y = randg(A,[m,n,p,...])
Y = randg(...,classname)
Y = randg(...,'like',X)
Y = randg(...,'like',classname)
```

Description

`Y = randg` returns a scalar random value chosen from a gamma distribution with unit scale and shape.

`Y = randg(A)` returns a matrix of random values chosen from gamma distributions with unit scale. `Y` is the same size as `A`, and `randg` generates each element of `Y` using a shape parameter equal to the corresponding element of `A`.

`Y = randg(A,m)` returns an `m`-by-`m` matrix of random values chosen from gamma distributions with shape parameters `A`. `A` is either an `m`-by-`m` matrix or a scalar. If `A` is a scalar, `randg` uses that single shape parameter value to generate all elements of `Y`.

`Y = randg(A,m,n,p,...)` or `Y = randg(A,[m,n,p,...])` returns an `m`-by-`n`-by-`p`-by-`...` array of random values chosen from gamma distributions with shape parameters `A`. `A` is either an `m`-by-`n`-by-`p`-by-`...` array or a scalar.

`Y = randg(...,classname)` returns an array of random values chosen from gamma distributions of the specified class. `classname` can be `double` or `single`.

`Y = randg(...,'like',X)` or `Y = randg(...,'like',classname)` returns an array of random values chosen from gamma distributions of the same class as `X` or `classname`, respectively. `X` is a numeric array.

`randg` produces pseudo-random numbers using the MATLAB functions `rand` and `randn`. The sequence of numbers generated is determined by the settings of the uniform random number generator that underlies `rand` and `randn`. Control that shared random number generator using `rng`. See the `rng` documentation for more information.

Note To generate gamma random numbers and specify both the scale and shape parameters, you should call `gamrnd`.

Examples

Example 1

Generate a 100-by-1 array of values drawn from a gamma distribution with shape parameter 3.

```
r = randg(3,100,1);
```

Example 2

Generate a 100-by-2 array of values drawn from gamma distributions with shape parameters 3 and 2.

```
A = [ones(100,1)*3,ones(100,1)*2];
r = randg(A,[100,2]);
```

Example 3

To create reproducible output from `randg`, reset the random number generator used by `rand` and `randn` to its default startup settings. This way `randg` produces the same random numbers as if you restarted MATLAB.

```
rng('default')
randg(3,1,5)

ans =

    6.9223    4.3369    1.0505    3.2662   11.3269
```

Example 4

Save the settings for the random number generator used by `rand` and `randn`, generate 5 values from `randg`, restore the settings, and repeat those values.

```
s = rng; % Obtain the current state of the random stream
r1 = randg(10,1,5)

r1 =

    9.4719    9.0433   15.0774   14.7763    6.3775

rng(s); % Reset the stream to the previous state
r2 = randg(10,1,5)

r2 =

    9.4719    9.0433   15.0774   14.7763    6.3775
```

`r2` contains exactly the same values as `r1`.

Example 5

Reinitialize the random number generator used by `rand` and `randn` with a seed based on the current time. `randg` returns different values each time you do this. Note that it is usually not necessary to do this more than once per MATLAB session.

```
rng('shuffle');
randg(2,1,5);
```

References

[1] Marsaglia, G., and W. W. Tsang. “A Simple Method for Generating Gamma Variables.” *ACM Transactions on Mathematical Software*. Vol. 26, 2000, pp. 363–372.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`gamrnd`

Introduced before R2006a

random

Package: prob

Random numbers

Syntax

```
R = random('name',A)
R = random('name',A,B)
R = random('name',A,B,C)
R = random('name',A,B,C,D)

R = random(pd)

R = random(___,sz1,...,szN)
R = random(___,sz)
```

Description

`R = random('name',A)` returns a random number from the one-parameter distribution family specified by `'name'` and the distribution parameter `A`.

`R = random('name',A,B)` returns a random number from the two-parameter distribution family specified by `'name'` and the distribution parameters `A` and `B`.

`R = random('name',A,B,C)` returns a random number from the three-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, and `C`.

`R = random('name',A,B,C,D)` returns a random number from the four-parameter distribution family specified by `'name'` and the distribution parameters `A`, `B`, `C`, and `D`.

`R = random(pd)` returns a random number from the probability distribution object `pd`.

`R = random(___,sz1,...,szN)` generates an array of random numbers from the specified probability distribution using input arguments from any of the previous syntaxes, where `sz1,...,szN` indicates the size of each dimension.

`R = random(___,sz)` generates an array of random numbers from the specified probability distribution using input arguments from any of the previous syntaxes, where vector `sz` specifies `size(r)`.

Examples

Generate One Random Number

Create a standard normal probability distribution object.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution
```

```
  Normal distribution
       mu = 0
    sigma = 1
```

Generate one random number from the distribution.

```
rng('default') % For reproducibility
r1 = random(pd)
```

```
r1 = 0.5377
```

Alternatively, you can generate a standard normal random number by specifying its name and parameters.

```
r2 = random('Normal',0,1)
```

```
r2 = 1.8339
```

Reset Random Number Generator

Save the current state of the random number generator. Then generate a random number from the Poisson distribution with rate parameter 5.

```
s = rng;
r = random('Poisson',5)
```

```
r = 5
```

Restore the state of the random number generator to s, and then create a new random number. The value is the same as before.

```
rng(s);
r1 = random('Poisson',5)
```

```
r1 = 5
```

Clone Size from Existing Array

Create a matrix of random numbers with the same size as an existing array. Use the stable distribution with shape parameters 2 and 0, scale parameter 1, and location parameter 0.

```
A = [3 2; -2 1];
sz = size(A);
R = random('Stable',2,0,1,0,sz)
```

```
R = 2×2

    0.7604   -3.1945
    2.5935    1.2193
```

You can combine the previous two lines of code into a single line.

```
R = random('Stable',2,0,1,0,size(A))
```

```
R = 2×2

    0.4508   -0.6132
   -1.8494    0.4845
```

Generate Multiple Random Numbers

Create a Weibull probability distribution object using the default parameter values.

```
pd = makedist('Weibull')
```

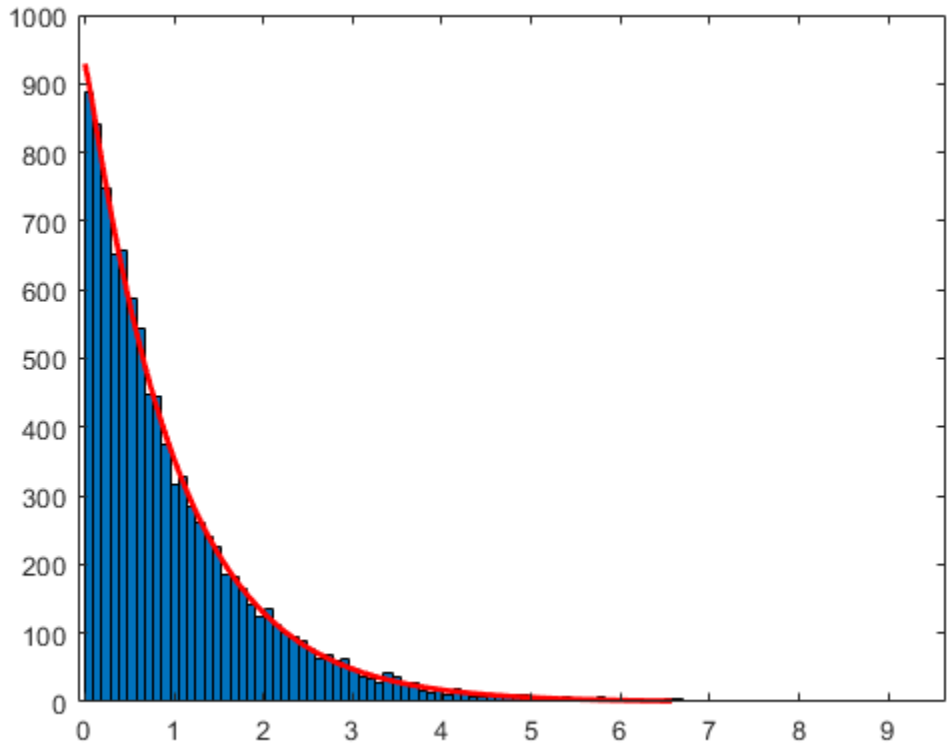
```
pd =
  WeibullDistribution

  Weibull distribution
    A = 1
    B = 1
```

Generate random numbers from the distribution.

```
rng('default')  % For reproducibility
r = random(pd,10000,1);
```

Construct a histogram using 100 bins with a Weibull distribution fit.

```
histfit(r,100,'weibull')
```



Generate Multidimensional Array of Random Numbers

Create a standard normal probability distribution object.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Generate a 2-by-3-by-2 array of random numbers from the distribution.

```
r = random(pd,[2,3,2])
```

```
r =
r(:,:,1) =

    0.5377   -2.2588    0.3188
    1.8339    0.8622   -1.3077
```

```
r(:,:,2) =

   -0.4336    3.5784   -1.3499
    0.3426    2.7694    3.0349
```

Input Arguments

`'name'` — Probability distribution name

character vector or string scalar of probability distribution name

Probability distribution name, specified as one of the probability distribution names in this table.

`'name'`	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
`'Beta'`	“Beta Distribution” on page B-6	a first shape parameter	b second shape parameter	—	—
`'Binomial'`	“Binomial Distribution” on page B-10	n number of trials	p probability of success for each trial	—	—
`'BirnbaumSaunders'`	“Birnbaum-Saunders Distribution” on page B-18	β scale parameter	γ shape parameter	—	—
`'Burr'`	“Burr Type XII Distribution” on page B-19	α scale parameter	c first shape parameter	k second shape parameter	—
`'Chisquare'`	“Chi-Square Distribution” on page B-28	ν degrees of freedom	—	—	—
`'Exponential'`	“Exponential Distribution” on page B-33	μ mean	—	—	—
`'Extreme Value'`	“Extreme Value Distribution” on page B-40	μ location parameter	σ scale parameter	—	—
`'F'`	“F Distribution” on page B-45	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	—	—
`'Gamma'`	“Gamma Distribution” on page B-47	a shape parameter	b scale parameter	—	—
`'Generalized Extreme Value'`	“Generalized Extreme Value Distribution” on page B-55	k shape parameter	σ scale parameter	μ location parameter	—
`'Generalized Pareto'`	“Generalized Pareto Distribution” on page B-59	k tail index (shape) parameter	σ scale parameter	μ threshold (location) parameter	—
`'Geometric'`	“Geometric Distribution” on page B-63	p probability parameter	—	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'HalfNormal'	“Half-Normal Distribution” on page B-68	μ location parameter	σ scale parameter	—	—
'Hypergeometric'	“Hypergeometric Distribution” on page B-73	m size of the population	k number of items with the desired characteristic in the population	n number of samples drawn	—
'InverseGaussian'	“Inverse Gaussian Distribution” on page B-75	μ scale parameter	λ shape parameter	—	—
'Logistic'	“Logistic Distribution” on page B-85	μ mean	σ scale parameter	—	—
'LogLogistic'	“Loglogistic Distribution” on page B-86	μ mean of logarithmic values	σ scale parameter of logarithmic values	—	—
'Lognormal'	“Lognormal Distribution” on page B-88	μ mean of logarithmic values	σ standard deviation of logarithmic values	—	—
'Nakagami'	“Nakagami Distribution” on page B-108	μ shape parameter	ω scale parameter	—	—
'Negative Binomial'	“Negative Binomial Distribution” on page B-109	r number of successes	p probability of success in a single trial	—	—
'Noncentral F'	“Noncentral F Distribution” on page B-115	ν_1 numerator degrees of freedom	ν_2 denominator degrees of freedom	δ noncentrality parameter	—
'Noncentral t'	“Noncentral t Distribution” on page B-117	ν degrees of freedom	δ noncentrality parameter	—	—
'Noncentral Chi-square'	“Noncentral Chi-Square Distribution” on page B-113	ν degrees of freedom	δ noncentrality parameter	—	—
'Normal'	“Normal Distribution” on page B-119	μ mean	σ standard deviation	—	—
'Poisson'	“Poisson Distribution” on page B-131	λ mean	—	—	—
'Rayleigh'	“Rayleigh Distribution” on page B-137	b scale parameter	—	—	—
'Rician'	“Rician Distribution” on page B-139	s noncentrality parameter	σ scale parameter	—	—

'name'	Distribution	Input Parameter A	Input Parameter B	Input Parameter C	Input Parameter D
'Stable'	“Stable Distribution” on page B-140	α first shape parameter	β second shape parameter	γ scale parameter	δ location parameter
'T'	“Student's t Distribution” on page B-149	ν degrees of freedom	—	—	—
'tLocationScale'	“t Location-Scale Distribution” on page B-156	μ location parameter	σ scale parameter	ν shape parameter	—
'Uniform'	“Uniform Distribution (Continuous)” on page B-163	a lower endpoint (minimum)	b upper endpoint (maximum)	—	—
'Discrete Uniform'	“Uniform Distribution (Discrete)” on page B-168	n maximum observable value	—	—	—
'Weibull'	“Weibull Distribution” on page B-170	a scale parameter	b shape parameter	—	—

Example: 'Normal'

A — First probability distribution parameter
scalar value | array of scalar values

First probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments A, B, C, and D are arrays, then the array sizes must be the same. In this case, random expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Data Types: single | double

B — Second probability distribution parameter
scalar value | array of scalar values

Second probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments A, B, C, and D are arrays, then the array sizes must be the same. In this case, random expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Data Types: single | double

C — Third probability distribution parameter
scalar value | array of scalar values

Third probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments A, B, C, and D are arrays, then the array sizes must be the same. In this case, random expands each scalar input into a constant array of the same size as the array inputs. See 'name' for the definitions of A, B, C, and D for each distribution.

Data Types: single | double

`D` — Fourth probability distribution parameter
scalar value | array of scalar values

Fourth probability distribution parameter, specified as a scalar value or an array of scalar values.

If one or more of the input arguments `A`, `B`, `C`, and `D` are arrays, then the array sizes must be the same. In this case, `random` expands each scalar input into a constant array of the same size as the array inputs. See `'name'` for the definitions of `A`, `B`, `C`, and `D` for each distribution.

Data Types: `single` | `double`

`pd` — Probability distribution
probability distribution object

Probability distribution, specified as a probability distribution object created with a function or app in this table.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.
`paretotails`	Create a piecewise distribution object that has generalized Pareto distributions in the tails.

`sz1,...,szN` — Size of each dimension (as separate arguments)
integer values

Size of each dimension, specified as integer values. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the specified probability distribution.

If one or more of the input arguments `A`, `B`, `C`, and `D` are arrays, then the specified dimensions `sz1,...,szN` must match the common dimensions of `A`, `B`, `C`, and `D` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `R` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `R` is an empty array.
- Beyond the second dimension, `random` ignores trailing dimensions with a size of 1. For example, specifying `3,1,1,1` produces a 3-by-1 vector of random numbers.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5 3 2]` generates a 5-by-3-by-2 array of random numbers from the specified probability distribution.

If one or more of the input arguments `A`, `B`, `C`, and `D` are arrays, then the specified dimensions `sz` must match the common dimensions of `A`, `B`, `C`, and `D` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value `[sz1]`, then `R` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `R` is an empty array.
- Beyond the second dimension, `random` ignores trailing dimensions with a size of 1. For example, specifying `[3 1 1 1]` produces a 3-by-1 vector of random numbers.

Example: `[5 3 2]`

Data Types: `single` | `double`

Output Arguments

R — Random number

scalar value | array of scalar values

Random number generated from the specified probability distribution, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`.

If you specify distribution parameters `A`, `B`, `C`, or `D`, then each element in `R` is the random number generated from the distribution specified by the corresponding elements in `A`, `B`, `C`, and `D`.

Alternative Functionality

- `random` is a generic function that accepts either a distribution by its name `'name'` or a probability distribution object `pd`. It is faster to use a distribution-specific function, such as `randn` and `normrnd` for the normal distribution and `binornd` for the binomial distribution. For a list of distribution-specific functions, see “Supported Distributions” on page 5-14.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `'name'` must be a compile-time constant. For example, to use the normal distribution, include `coder.Constant('Normal')` in the `-args` value of `codegen`.
- Code generation does not support the probability distribution object (`pd`) input argument.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

Distribution Fitter | `cdf` | `fitdist` | `icdf` | `makedist` | `mle` | `paretotails` | `pdf`

Topics

“Random Number Generation” on page 5-27
“Generate Random Numbers Using the Triangular Distribution” on page 5-47
“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

random

Package:

Simulate responses with random noise for generalized linear regression model

Syntax

```
ysim = random(mdl,Xnew)
ysim = random(mdl,Xnew,Name,Value)
```

Description

`ysim = random(mdl,Xnew)` simulates responses to the predictor data in `Xnew` using the generalized linear regression model `mdl`, adding random noise.

`ysim = random(mdl,Xnew,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the number of trials for binomial distribution or the offset value used for fitting.

Examples

Simulate Response Data with Random Noise

Create a generalized linear regression model, and simulate its response with random noise to new data.

Generate sample data using Poisson random numbers with one underlying predictor `X`.

```
rng('default') % For reproducibility
X = rand(20,1);
mu = exp(1 + 2*X);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1','Distribution','poisson');
```

Create data points for prediction.

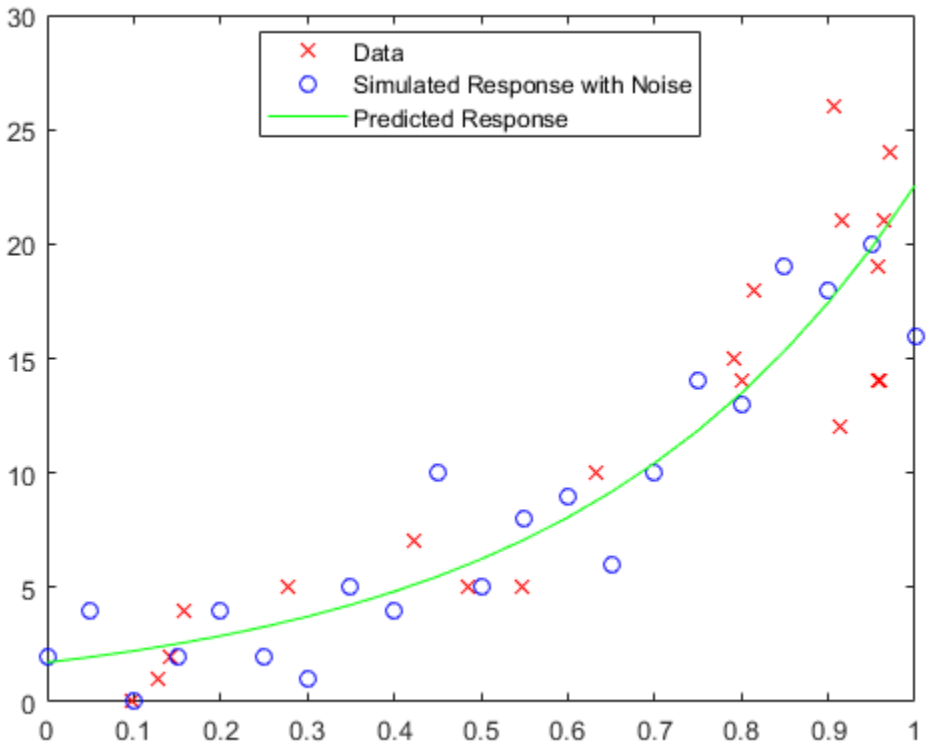
```
Xnew = (0:.05:1)';
```

Simulate responses with random noise at the data points.

```
ysim = random(mdl,Xnew);
```

Plot the simulated values and the original values.

```
plot(X,y,'rx',Xnew,ysim,'bo',Xnew,feval(mdl,Xnew),'g-')
legend('Data','Simulated Response with Noise','Predicted Response', ...
    'Location','best')
```



Generate C/C++ Code That Simulates Responses

Fit a generalized linear regression model, and then save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the fitted model. Then use `codegen` (MATLAB Coder) to generate C/C++ code. Note that generating C/C++ code requires MATLAB® Coder™.

This example briefly explains the code generation workflow for the prediction of linear regression models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. For details, see “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22.

Train Model

Generate sample data of the predictor `x` and response `y` with the following distributions:

- $x \sim N(1, 0.5^2)$.
- $\beta_0 = 1$ and $\beta = -2$.
- $y \sim Binomial\left(10, \frac{exp(1 + x\beta)}{1 + exp(1 + x\beta)}\right)$.

```
rng('default') % For reproducibility
x = 1 + randn(100,1)*0.5;
beta = -2;
p = exp(1 + x*beta)./(1 + exp(1 + x*beta)); % Inverse logit
n = 10;
y = binornd(n,p,100,1);
```

Create a generalized linear regression model of binomial data. Specify a binomial sample size of 10.

```
mdl = fitglm(x,y,'y ~ x1','Distribution','Binomial','BinomialSize',n);
```

Save Model

Save the fitted generalized linear regression model to the file `GLMMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(mdl,'GLMMdl');
```

Define Entry-Point Function

In your current folder, define an entry-point function named `myrandomGLM.m` that does the following:

- Accept new predictor input and valid name-value pair arguments.
- Load the fitted generalized linear regression model in `GLMMdl.mat` by using `loadLearnerForCoder`.
- Simulate responses from the loaded GLM model.

```
function y = myrandomGLM(x,varargin) %#codegen
%MYRANDOMGLM Simulate responses using GLM model
%   MYRANDOMGLM simulates responses for the n observations in the n-by-1
%   vector x using the GLM model stored in the MAT-file GLMMdl.mat, and
%   then returns the simulations in the n-by-1 vector y.
CompactMdl = loadLearnerForCoder('GLMMdl');
narginchk(1,Inf);
y = random(CompactMdl,x,varargin{:});
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. To specify the data type and exact input array size, pass a MATLAB® expression that represents the set of values with a certain data type and array size. Use `coder.Constant` (MATLAB Coder) for the names of name-value pair arguments.

Specify the predictor data `x` and binomial parameter `n`.

```
codegen -config:mex myrandomGLM -args {x,coder.Constant('BinomialSize'),coder.Constant(n)}
```

```
Code generation successful.
```

`codegen` generates the MEX function `myrandomGLM_mex` with a platform-dependent extension.

If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

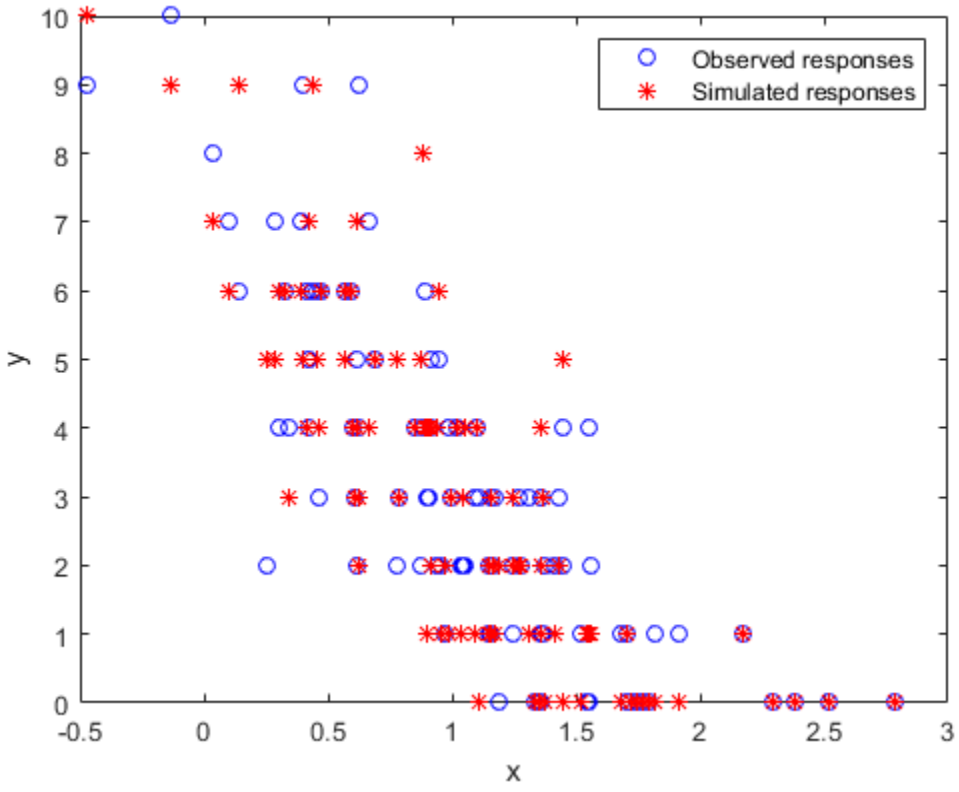
Verify Generated Code

Simulate responses using the MEX function. Specify the predictor data `x` and binomial parameter `n`.

```
ysim = myrandomGLM_mex(x,'BinomialSize',n);
```

Plot the simulated values and the data in the same figure.

```
figure
plot(x,y,'bo',x,ysim,'r*')
legend('Observed responses','Simulated responses')
xlabel('x')
ylabel('y')
```



The observed and simulated responses appear to be similarly distributed.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object | `CompactGeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`, or a `CompactGeneralizedLinearModel` object created using `compact`.

`Xnew` — New predictor input values
table | dataset array | matrix

New predictor input values, specified as a table, dataset array, or matrix. Each row of `Xnew` corresponds to one observation, and each column corresponds to one variable.

- If `Xnew` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
- If `Xnew` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating `mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `ysim = random(Mdl,Xnew,'BinomialSize',50)` returns the numbers of success, perturbed by random noise, using the number of trials specified by `'BinomialSize'`.

`BinomialSize` — Number of trials for binomial distribution
1 (default) | scalar | vector

Number of trials for the binomial distribution, specified as the comma-separated pair consisting of `'BinomialSize'` and a scalar or vector of the same length as the response. `random` expands the scalar input into a constant array of the same size as the response. The scalar input means that all observations have the same number of trials.

The meaning of the output values in `ysim` depends on the value of `'BinomialSize'`.

- If `'BinomialSize'` is 1 (default), then each value in the output `ysim` is the probability of success.
- If `'BinomialSize'` is not 1, then each value in the output `ysim` is the predicted number of successes in the trials.

Data Types: `single` | `double`

`Offset` — Offset value
`zeros(size(Xnew,1))` (default) | scalar | vector

Offset value for each row in `Xnew`, specified as the comma-separated pair consisting of `'Offset'` and a scalar or vector with the same length as the response. `random` expands the scalar input into a constant array of the same size as the response.

Note that the default value of this argument is a vector of zeros even if you specify the `'Offset'` name-value pair argument when fitting a model. If you specify `'Offset'` for fitting, the software

treats the offset as an additional predictor with a coefficient value fixed at 1. In other words, the formula for fitting is

$$f(\mu) = \texttt{Offset} + X*b,$$

where f is the link function, μ is the mean response, and $X*b$ is the linear combination of predictors X. The `Offset` predictor has coefficient `1`.

Data Types: `single` | `double`

Output Arguments

`ysim` — Simulated response values

numeric vector

Simulated response values, returned as a numeric vector. The simulated values are the predicted response values at `Xnew` perturbed by random noise with the distribution given by the fitted model. The values in `ysim` are independent, conditional on the predictors. For binomial and Poisson fits, `random` generates `ysim` with the specified distribution and no adjustment for any estimated dispersion.

- If `'BinomialSize'` is 1 (default), then each value in the output `ysim` is the probability of success.
- If `'BinomialSize'` is not 1, then each value in the output `ysim` is the predicted number of successes in the trials.

Alternative Functionality

For predictions without random noise, use `predict` or `feval`.

- `predict` accepts a single input argument containing all predictor variables, and gives confidence intervals on its predictions.
- `feval` accepts multiple input arguments with one input for each predictor variable, which is simpler to use with a model created from a table or dataset array. The `feval` function does not support the name-value pair arguments `'Offset'` and `'BinomialSize'`. The function uses 0 as the offset value, and the output values are predicted probabilities.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `random` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `random` function. Then use `codegen` to generate code for the entry-point function.
- `random` can return a different sequence of numbers than MATLAB if either of the following is true:
  - The output is nonscalar.

- An input parameter is invalid for the distribution.

- This table contains notes about the arguments of `random`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`mdl`	For the usage notes and limitations of the model object, see "Code Generation" on page 33-835 of the `CompactGeneralizedLinearModel` object.

`Xnew`	• `Xnew` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in `Xnew` can be a variable size, but the number of columns in `Xnew` must be fixed.
	• If you want to specify `Xnew` as a table, then your model must be trained using a table, and you must ensure that your entry-point function for prediction:
	  • Accepts data as arrays
	  • Creates a table from the data input arguments and specifies the variable names in the table
	  • Passes the table to `predict`
	  For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to use the `'BinomialSize'` name-value pair argument in the generated code, include `{coder.Constant('BinomialSize'),0}` in the `-args` value of `codegen`.

For more information, see "Introduction to Code Generation" on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactGeneralizedLinearModel` | `GeneralizedLinearModel` | `feval` | `predict`

Topics

"Predict or Simulate Responses to New Data" on page 12-23
"Generalized Linear Models" on page 12-9

Introduced in R2012a

random

Class: `GeneralizedLinearMixedModel`

Generate random responses from fitted generalized linear mixed-effects model

Syntax

```
ysim = random(glme)
ysim = random(glme,tblnew)
ysim = random( ___ ,Name,Value)
```

Description

`ysim = random(glme)` returns simulated responses, `ysim`, from the fitted generalized linear mixed-effects model `glme`, at the original design points.

`ysim = random(glme,tblnew)` returns simulated responses using new input values specified in the table or dataset array, `tblnew`.

`ysim = random( ___ ,Name,Value)` returns simulated responses using additional options specified by one or more `Name,Value` pair arguments, using any of the previous syntaxes. For example, you can specify observation weights, binomial sizes, or offsets for the model.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`tblnew` — New input data

table | dataset array

New input data, which includes the response variable, predictor variables, and grouping variables on page 2-45, specified as a table or dataset array. The predictor variables can be continuous or grouping variables. `tblnew` must contain the same variables as the original table or dataset array, `tbl`, used to fit the generalized linear mixed-effects model `glme`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`BinomialSize` — Number of trials for binomial distribution

`ones(m,1)` (default) | *m*-by-1 vector of positive integer values

Number of trials for binomial distribution, specified as the comma-separated pair consisting of `'BinomialSize'` and an *m*-by-1 vector of positive integer values, where *m* is the number of rows in

`tblnew`. The `'BinomialSize'` name-value pair applies only to the binomial distribution. The value specifies the number of binomial trials when generating the random response values.

Data Types: `single` | `double`

`Offset` — Model offset
`zeros(m,1)` (default) | vector of scalar values

Model offset, specified as a vector of scalar values of length *m*, where *m* is the number of rows in `tblnew`. The offset is used as an additional predictor and has a coefficient value fixed at `1`.

`Weights` — Observation weights
m-by-1 vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *m*-by-1 vector of nonnegative scalar values, where *m* is the number of rows in `tblnew`. If the response distribution is binomial or Poisson, then `'Weights'` must be a vector of positive integers.

Data Types: `single` | `double`

Output Arguments

`ysim` — Simulated response values
m-by-1 vector

Simulated response values, returned as an *m*-by-1 vector, where *m* is the number of rows in `tblnew`. `random` creates `ysim` by first generating the random-effects vector based on its fitted prior distribution. `random` then generates `ysim` from its fitted conditional distribution given the random effects. `random` takes into account the effect of observation weights specified when fitting the model using `fitglme`, if any.

Examples

Simulate Random Responses From a GLME Model

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)

- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Use `random` to simulate a new response vector from the fitted model.

```
rng(0,'twister');  % For reproducibility
ynew = random(glme);
```

Display the first 10 rows of the simulated response vector.

```
ynew(1:10)
```

```
ans = 10×1

     3
     3
     1
     7
     5
```

```
     8
     7
     9
     5
     9
```

Simulate a new response vector using new input values. Create a new table by copying the first 10 rows of `mfr` into `tblnew`.

```
tblnew = mfr(1:10,:);
```

The first 10 rows of `mfr` include data collected from trials 1 through 5 for factories 1 and 2. Both factories used the old process for all of their trials during the experiment, so `newprocess = 0` for all 10 observations.

Change the value of `newprocess` to `1` for the observations in `tblnew`.

```
tblnew.newprocess = ones(height(tblnew),1);
```

Simulate new responses using the new input values in `tblnew`.

```
ynew2 = random(glme,tblnew)
```

```
ynew2 = 10×1

     2
     3
     5
     4
     2
     2
     2
     1
     2
     0
```

More About

Conditional Distribution Method

`random` generates random data from the fitted generalized linear mixed-effects model as follows:

- Sample $b_{sim} \sim P\left(b \middle| \widehat{\theta}, \widehat{\sigma}^2\right)$, where $P\left(b \middle| \widehat{\theta}, \widehat{\sigma}^2\right)$ is the estimated prior distribution of random effects, and $\widehat{\theta}$ is a vector of estimated covariance parameters, and $\widehat{\sigma}^2$ is the estimated dispersion parameter.
- Given b_{sim}, for $i = 1$ to m, sample $y_{sim\_i} \sim P\left(y_{new\_i} \middle| b_{sim}, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}^2\right)$, where $P\left(y_{new\_i} \middle| b_{sim}, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}^2\right)$ is the conditional distribution of the ith new response y_{new_i} given b_{sim} and the model parameters.

See Also

`GeneralizedLinearMixedModel` | `fitglme` | `fitted` | `predict`

random

Random variate from Gaussian mixture distribution

Syntax

```
Y = random(gm)
Y = random(gm,n)
[Y,compIdx] = random( ___ )
```

Description

`Y = random(gm)` generates a 1-by-*m* random variate from the *m*-dimensional Gaussian mixture distribution `gm`.

`Y = random(gm,n)` returns `n` random variates. Each row of `Y` is a random variate generated from the *m*-dimensional Gaussian mixture distribution `gm`.

`[Y,compIdx] = random( ___ )` also returns an `n`-by-1 index vector `compIdx` for any of the input arguments in previous syntaxes. `compIdx(i)` indicates the mixture component used to generate the `i`th random variate `Y(i,:)`.

Examples

Generate Random Variates

Create a `gmdistribution` object and generate random variates.

Define the distribution parameters (means and covariances) of a two-component bivariate Gaussian mixture distribution.

```
mu = [1 2;-3 -5];
sigma = [1 1]; % shared diagonal covariance matrix
```

Create a `gmdistribution` object by using the `gmdistribution` function. By default, the function creates an equal proportion mixture.

```
gm = gmdistribution(mu,sigma)
```

```
gm =

Gaussian mixture distribution with 2 components in 2 dimensions
Component 1:
Mixing proportion: 0.500000
Mean:     1     2

Component 2:
Mixing proportion: 0.500000
Mean:    -3    -5
```

Generate 1000 random variates.

```
rng('default'); % For reproducibility
[Y,compIdx] = random(gm,1000);
```

`compIdx(i)` indicates the mixture component used to generate the `i`th random variate `Y(i,:)`. Count the number of random variates generated by `Component 1`.

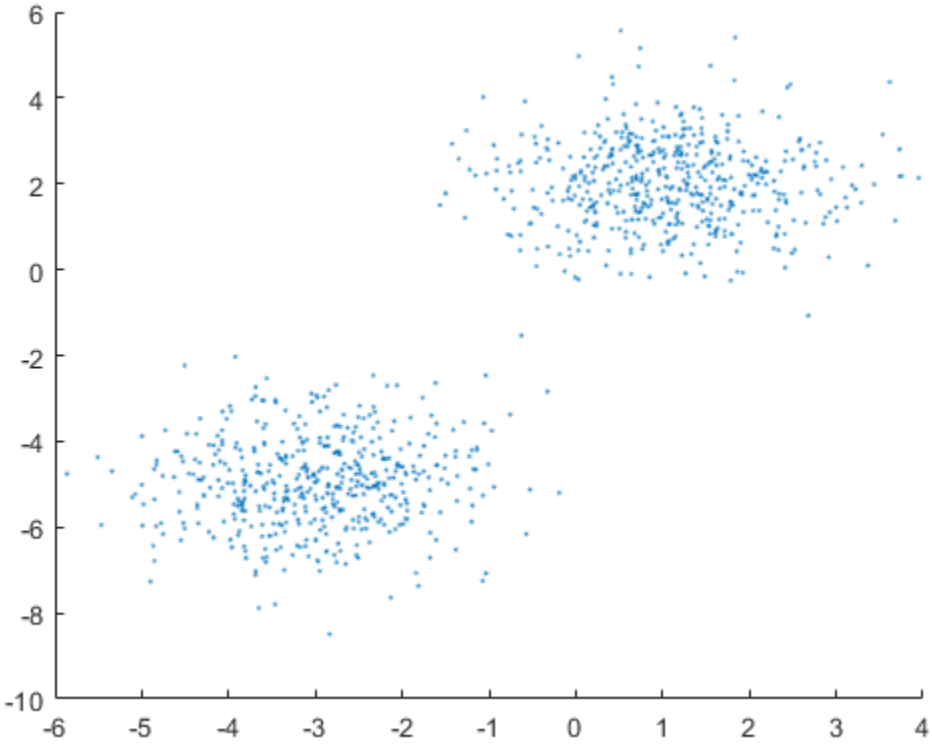
```
numIdx1 = sum(compIdx == 1)
```

```
numIdx1 = 512
```

`random` generates about half of the random variates using `Component 1` because `gm` has equal mixing proportions.

Plot the generated random variates by using `scatter`.

```
scatter(Y(:,1),Y(:,2),10,'.') % Scatter plot with points of size 10
```



Reset Random Number Generator

Reset the random number generator to generate the same random variate.

Define the distribution parameters (means, covariances, and mixing proportions) of two bivariate Gaussian mixture components.

```
p = [0.4 0.6];               % Mixing proportions
mu = [1 2;-3 -5];            % Means
sigma = cat(3,[2 .5],[1 1])  % Covariances 1-by-2-by-2 array
```

```
sigma =
sigma(:,:,1) =

    2.0000    0.5000


sigma(:,:,2) =

     1     1
```

The `cat` function concatenates the covariances along the third array dimension. The defined covariance matrices are diagonal matrices. `sigma(1,:,i)` contains the diagonal elements of the covariance matrix of component `i`.

Create a `gmdistribution` object by using the `gmdistribution` function.

```
gm = gmdistribution(mu,sigma);
```

Save the current state of the random number generator, and then generate a random variate using `gm`.

```
s = rng;
r = random(gm)
```

```
r = 1×2

   -1.1661   -7.2588
```

Restore the state of the random number generator to `s`, and then generate a random variate using `gm`. The values are the same as before.

```
rng(s);
r1 = random(gm)
```

```
r1 = 1×2

   -1.1661   -7.2588
```

Input Arguments

`gm` — Gaussian mixture distribution

`gmdistribution` object

Gaussian mixture distribution, also called Gaussian mixture model (GMM), specified as a `gmdistribution` object.

You can create a `gmdistribution` object using `gmdistribution` or `fitgmdist`. Use the `gmdistribution` function to create a `gmdistribution` object by specifying the distribution parameters. Use the `fitgmdist` function to fit a `gmdistribution` model to data given a fixed number of components.

`n` — Number of random variates

1 (default) | positive integer

Number of random variates to generate, specified as a positive integer.

Data Types: `single` | `double`

Output Arguments

`Y` — Random variate

1-by-*m* numeric vector | `n`-by-*m* numeric matrix

Random variate, returned as a 1-by-*m* numeric vector or an `n`-by-*m* numeric matrix. Each row of `Y` is a random variate generated from the *m*-dimensional Gaussian mixture distribution `gm`.

`compIdx` — Component index

positive integer | *n*-by-1 numeric vector

Component index, returned as a positive integer or an `n`-by-1 index vector, where `compIdx(i)` indicates the mixture component used to generate the `i`th random variate `Y(i,:)`.

See Also

`cdf` | `fitgmdist` | `gmdistribution` | `mvnrnd` | `pdf`

Topics

“Create Gaussian Mixture Model” on page 5-112
“Fit Gaussian Mixture Model to Data” on page 5-115
“Simulate Data from Gaussian Mixture Model” on page 5-119
“Cluster Using Gaussian Mixture Model” on page 16-39

Introduced in R2007b

random

Package:

Simulate responses with random noise for linear regression model

Syntax

```
ysim = random(mdl,Xnew)
```

Description

`ysim = random(mdl,Xnew)` simulates responses to the predictor data in `Xnew` using the linear model `mdl`, adding random noise.

Examples

Simulate Response Data with Random Noise

Create a quadratic model of car mileage as a function of weight from the `carsmall` data set.

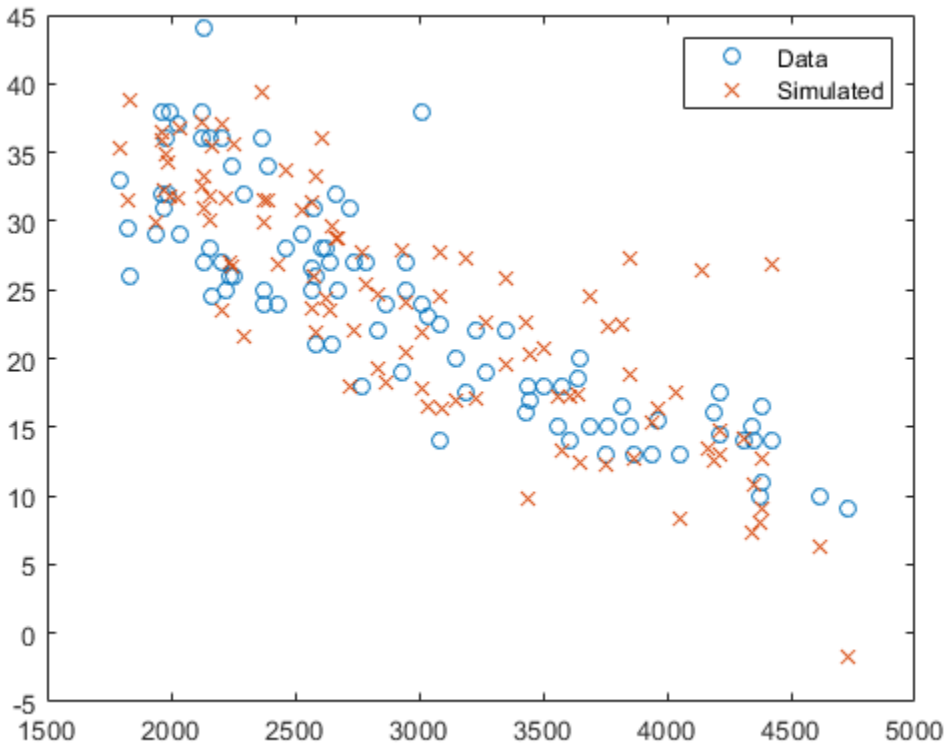
```
load carsmall
X = Weight;
y = MPG;
mdl = fitlm(X,y,'quadratic');
```

Create simulated responses to the data with random noise.

```
ysim = random(mdl,X);
```

Plot the original responses and the simulated responses to see how they differ.

```
plot(X,y,'o',X,ysim,'x')
legend('Data','Simulated')
```



Input Arguments

`mdl` — Linear regression model object

`LinearModel` object | `CompactLinearModel` object

Linear regression model object, specified as a `LinearModel` object created by using `fitlm` or `stepwiselm`, or a `CompactLinearModel` object created by using `compact`.

`Xnew` — New predictor input values

table | dataset array | matrix

New predictor input values, specified as a table, dataset array, or matrix. Each row of `Xnew` corresponds to one observation, and each column corresponds to one variable.

- If `Xnew` is a table or dataset array, it must contain predictors that have the same predictor names as in the `PredictorNames` property of `mdl`.
- If `Xnew` is a matrix, it must have the same number of variables (columns) in the same order as the predictor input used to create `mdl`. Note that `Xnew` must also contain any predictor variables that are not used as predictors in the fitted model. Also, all variables used in creating `mdl` must be numeric. To treat numerical predictors as categorical, identify the predictors using the `'CategoricalVars'` name-value pair argument when you create `mdl`.

Data Types: `single` | `double` | `table`

Output Arguments

`ysim` — Simulated response values

numeric vector

Simulated response value, returned as a numeric vector. The simulated value is the predicted response values at `Xnew` perturbed by random noise. The noise is independent and normally distributed, with mean equal to zero and variance equal to the estimated error variance of the model.

Alternative Functionality

For predictions without random noise, use `predict` or `feval`. These two functions give the same predictions.

- `predict` accepts a single input argument containing all predictor variables, and gives confidence intervals on its predictions.
- `feval` accepts multiple input arguments with one input for each predictor variable.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `random` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `random` function. Then use `codegen` to generate code for the entry-point function.
- `random` can return a different sequence of numbers than MATLAB if either of the following is true:
  - The output is nonscalar.
  - An input parameter is invalid for the distribution.
- This table contains notes about the arguments of `random`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`mdl`	• Suppose you train a linear model by using `fitlm` and specifying `'RobustOpts'` as a structure with an anonymous function handle for the `RobustWgtFun` field, use `saveLearnerForCoder` to save the model, and then use `loadLearnerForCoder` to load the model. In this case, `loadLearnerForCoder` cannot restore the Robust property into the MATLAB Workspace. However, `loadLearnerForCoder` can load the model at compile time within an entry-point function for code generation. • For the usage notes and limitations of the model object, see “Code Generation” on page 33-824 of the `CompactLinearModel` object.

Argument	Notes and Limitations

`Xnew`	• `Xnew` must be a single-precision or double-precision matrix or a table containing numeric variables, categorical variables, or both.
	• The number of rows, or observations, in `Xnew` can be a variable size, but the number of columns in `Xnew` must be fixed.
	• If you want to specify `Xnew` as a table, then your model must be trained using a table, and you must ensure that your entry-point function for prediction:
	  • Accepts data as arrays
	  • Creates a table from the data input arguments and specifies the variable names in the table
	  • Passes the table to `predict`
	  For an example of this table workflow, see "Generate Code to Classify Data in Table" on page 32-100. For more information on using tables in code generation, see "Code Generation for Tables" (MATLAB Coder) and "Table Limitations for Code Generation" (MATLAB Coder).

For more information, see "Introduction to Code Generation" on page 32-2.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

- This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).
- This function supports model objects fitted with GPU array input arguments.

See Also

`CompactLinearModel` | `LinearModel` | `feval` | `predict`

Topics

"Predict or Simulate Responses to New Data" on page 11-31
"Linear Regression Workflow" on page 11-35
"Interpret Linear Regression Results" on page 11-50
"Linear Regression" on page 11-9

Introduced in R2012a

random

Class: `LinearMixedModel`

Generate random responses from fitted linear mixed-effects model

Syntax

```
ysim = random(lme)
ysim = random(lme,tblnew)
ysim = random(lme,Xnew,Znew)
ysim = random(lme,Xnew,Znew,Gnew)
```

Description

`ysim = random(lme)` returns a vector of simulated responses `ysim` from the fitted linear mixed-effects model `lme` at the original fixed- and random-effects design points, used to fit `lme`.

`random` simulates new random-effects vector and new observation errors. So, the simulated response is

$$y_{sim} = X\widehat{\beta} + Z\widehat{b} + \varepsilon,$$

where $\widehat{\beta}$ is the estimated fixed-effects coefficients, $\widehat{b}$ is the new random effects, and ε is the new observation error.

`random` also accounts for the effect of observation weights, if you use any when fitting the model.

`ysim = random(lme,tblnew)` returns a vector of simulated responses `ysim` from the fitted linear mixed-effects model `lme` at the values in the new table or dataset array `tblnew`. Use a table or dataset array for `random` if you use a table or dataset array for fitting the model `lme`.

`ysim = random(lme,Xnew,Znew)` returns a vector of simulated responses `ysim` from the fitted linear mixed-effects model `lme` at the values in the new fixed- and random-effects design matrices, `Xnew` and `Znew`, respectively. `Znew` can also be a cell array of matrices. Use the matrix format for `random` if you use design matrices for fitting the model `lme`.

`ysim = random(lme,Xnew,Znew,Gnew)` returns a vector of simulated responses `ysim` from the fitted linear mixed-effects model `lme` at the values in the new fixed- and random-effects design matrices, `Xnew` and `Znew`, respectively, and the grouping variable `Gnew`.

`Znew` and `Gnew` can also be cell arrays of matrices and grouping variables, respectively.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

`tblnew` — New input data
table | `dataset` array

New input data, which includes the response variable, predictor variables, and grouping variables on page 2-45, specified as a table or dataset array. The predictor variables can be continuous or grouping variables. `tblnew` must have the same variables as in the original table or dataset array used to fit the linear mixed-effects model `lme`.

`Xnew` — New fixed-effects design matrix
n-by-*p* matrix

New fixed-effects design matrix, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of fixed predictor variables. Each row of `X` corresponds to one observation and each column of `X` corresponds to one variable.

Data Types: `single` | `double`

`Znew` — New random-effects design
n-by-*q* matrix | cell array of length *R*

New random-effects design, specified as an *n*-by-*q* matrix or a cell array of *R* design matrices `Z{r}`, where *r* = 1, 2, ..., *R*. If `Znew` is a cell array, then each `Z{r}` is an *n*-by-*q*(*r*) matrix, where *n* is the number of observations, and *q*(*r*) is the number of random predictor variables.

Data Types: `single` | `double` | `cell`

`Gnew` — New grouping variable or variables
vector | cell array of grouping variables of length *R*

New grouping variable or variables on page 2-45, specified as a vector or a cell array, of length *R*, of grouping variables used to fit the linear mixed-effects model, `lme`.

`random` treats all levels of each grouping variable as new levels. It draws an independent random effects vector for each level of each grouping variable.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

Output Arguments

`ysim` — Simulated response values
n-by-1 vector

Simulated response values, returned as an *n*-by-1 vector, where *n* is the number of observations.

Examples

Generate Random Responses at the Original Design Values

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five

different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Generate random response values at the original design points. Display the first five values.

```
rng(123,'twister') % For reproducibility
ysim = random(lme);
ysim(1:5)
```

```
ans = 5×1

  114.8785
  134.2018
  154.2818
  169.7554
   84.6089
```

Plot Randomly Generated vs. Observed Response Values

Load the sample data.

```
load carsmall
```

Fit a linear mixed-effects model, with a fixed-effects for `Weight`, and a random intercept grouped by `Model_Year`. First, store the data in a table.

```
tbl = table(MPG,Weight,Model_Year);
lme = fitlme(tbl,'MPG ~ Weight + (1|Model_Year)');
```

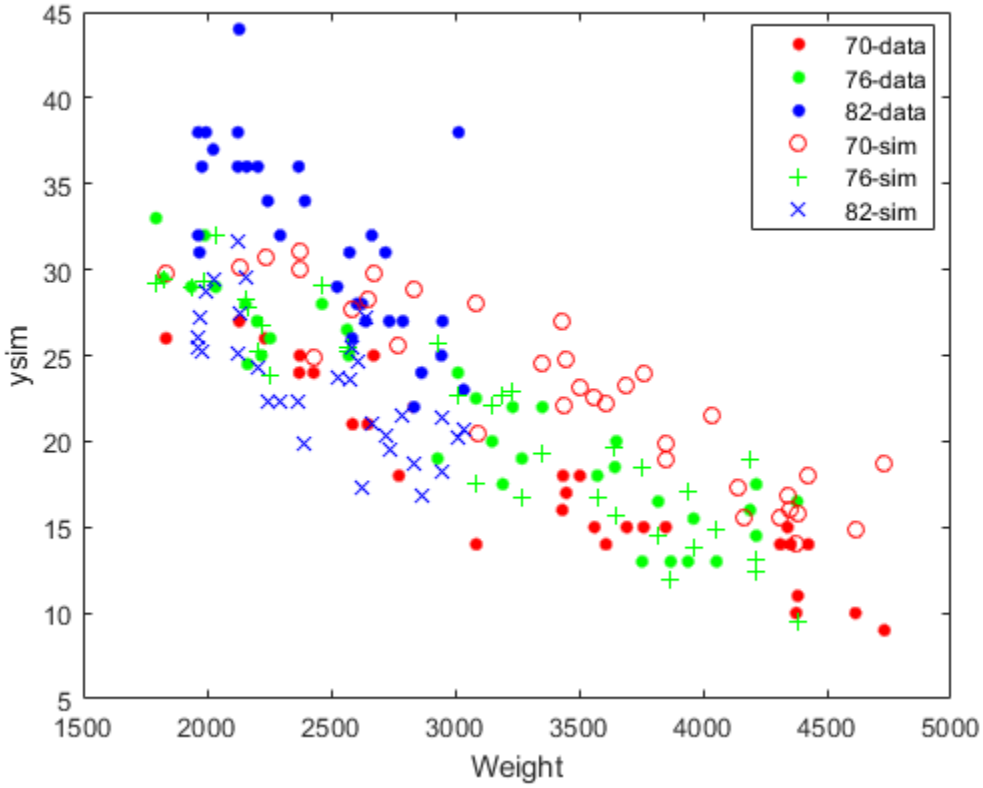
Randomly generate responses using the original data.

```
rng(123,'twister') % For reproducibility
ysim = random(lme,tbl);
```

Plot the original and the randomly generated responses to see how they differ. Group them by model year.

```
figure()
gscatter(Weight,MPG,Model_Year)
```

```
hold on
gscatter(Weight,ysim,Model_Year,[],'o+x')
legend('70-data','76-data','82-data','70-sim','76-sim','82-sim')
hold off
```



Note that the simulated random response values for year 82 are lower than the original data for that year. This might be due to a lower simulated random effect for year 82 than the estimated random effect in the original data.

Generate Responses Using a New Dataset Array

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Create a new dataset array with design values. The new dataset array must have the same variables as the original dataset array you use for fitting the model `lme`.

```
dsnew = dataset();
dsnew.Soil = nominal({'Sandy';'Silty';'Silty'});
dsnew.Tomato = nominal({'Cherry';'Vine';'Plum'});
dsnew.Fertilizer = nominal([2;2;4]);
```

Generate random responses at the new points.

```
rng(123,'twister') % For reproducibility
ysim = random(lme,dsnew)
```

```
ysim = 3×1

   99.6006
  101.9911
  161.4026
```

Generate Random Responses Using New Design Matrices

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and cylinders, and potentially correlated random effect for intercept and acceleration grouped by model year.

First, prepare the design matrices for fitting the linear mixed-effects model.

```
X = [ones(406,1) Acceleration Horsepower];
Z = [ones(406,1) Acceleration];
Model_Year = nominal(Model_Year);
G = Model_Year;
```

Now, fit the model using `fitlmematrix` with the defined design matrices and grouping variables.

```
lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept','Acceleration'}},'RandomEffectGroups',{'Model_Year'});
```

Create the design matrices that contain the data at which to predict the response values. `Xnew` must have three columns as in `X`. The first column must be a column of 1s. And the values in the last two

columns must correspond to `Acceleration` and `Horsepower`, respectively. The first column of `Znew` must be a column of 1s, and the second column must contain the same `Acceleration` values as in `Xnew`. The original grouping variable in `G` is the model year. So, `Gnew` must contain values for the model year. Note that `Gnew` must contain nominal values.

```
Xnew = [1,13.5,185; 1,17,205; 1,21.2,193];
Znew = [1,13.5; 1,17; 1,21.2];
Gnew = nominal([73 77 82]);
```

Generate random responses for the data in the new design matrices.

```
rng(123,'twister') % For reproducibility
ysim = random(lme,Xnew,Znew,Gnew)
```

```
ysim = 3×1

   15.7416
   10.6085
    6.8796
```

Now, repeat the same for a linear mixed-effects model with uncorrelated random-effects terms for intercept and acceleration. First, change the original random effects design and the random effects grouping variables. Then, fit the model.

```
Z = {ones(406,1),Acceleration};
G = {Model_Year,Model_Year};

lme = fitlmematrix(X,MPG,Z,G,'FixedEffectPredictors',....
{'Intercept','Acceleration','Horsepower'},'RandomEffectPredictors',...
{{'Intercept'},{'Acceleration'}},'RandomEffectGroups',{'Model_Year','Model_Year'});
```

Now, recreate the new random effects design, `Znew`, and the grouping variable design, `Gnew`, using which to predict the response values.

```
Znew = {[1;1;1],[13.5;17;21.2]};
MY = nominal([73 77 82]);
Gnew = {MY,MY};
```

Generate random responses using the new design matrices.

```
rng(123,'twister') % For reproducibility
ysim = random(lme,Xnew,Znew,Gnew)
```

```
ysim = 3×1

   16.8280
   10.4375
    4.1027
```

See Also

`LinearMixedModel` | `fitlme` | `fitlmematrix` | `predict`

random

Class: NonLinearModel

Simulate responses for nonlinear regression model

Syntax

```
ysim = random(mdl)
ysim = random(mdl,Xnew)
ysim = random(mdl,Xnew,'Weights',W)
```

Description

`ysim = random(mdl)` simulates responses from the fitted nonlinear model `mdl` at the original design points.

`ysim = random(mdl,Xnew)` simulates responses from the fitted nonlinear model `mdl` to the data in `Xnew`, adding random noise.

`ysim = random(mdl,Xnew,'Weights',W)` simulates responses using the observation weights, `W`.

Input Arguments

`mdl`

Nonlinear regression model, constructed by `fitnlm`.

`Xnew`

Points at which `mdl` predicts responses.

- If `Xnew` is a table or dataset array, it must contain the predictor names in `mdl`.
- If `Xnew` is a numeric matrix, it must have the same number of variables (columns) as was used to create `mdl`. Furthermore, all variables used in creating `mdl` must be numeric.

`W`

Vector of real, positive value weights or a function handle.

- If you specify a vector, then it must have the same number of elements as the number of observations (or rows) in `Xnew`.
- If you specify a function handle, the function must accept a vector of predicted response values as input, and returns a vector of real positive weights as output.

Given weights, `W`, `random` estimates the error variance at observation `i` by `MSE*(1/W(i))`, where MSE is the mean squared error.

Default: No weights

Output Arguments

`ysim`

Vector of predicted mean values at `Xnew`, perturbed by random noise. The noise is independent, normally distributed, with mean zero, and variance equal to the estimated error variance of the model.

Examples

Simulate Responses

Create a nonlinear model of car mileage as a function of weight, and simulate the response.

Create an exponential model of car mileage as a function of weight from the `carsmall` data. Scale the weight by a factor of 1000 so all the variables are roughly equal in size.

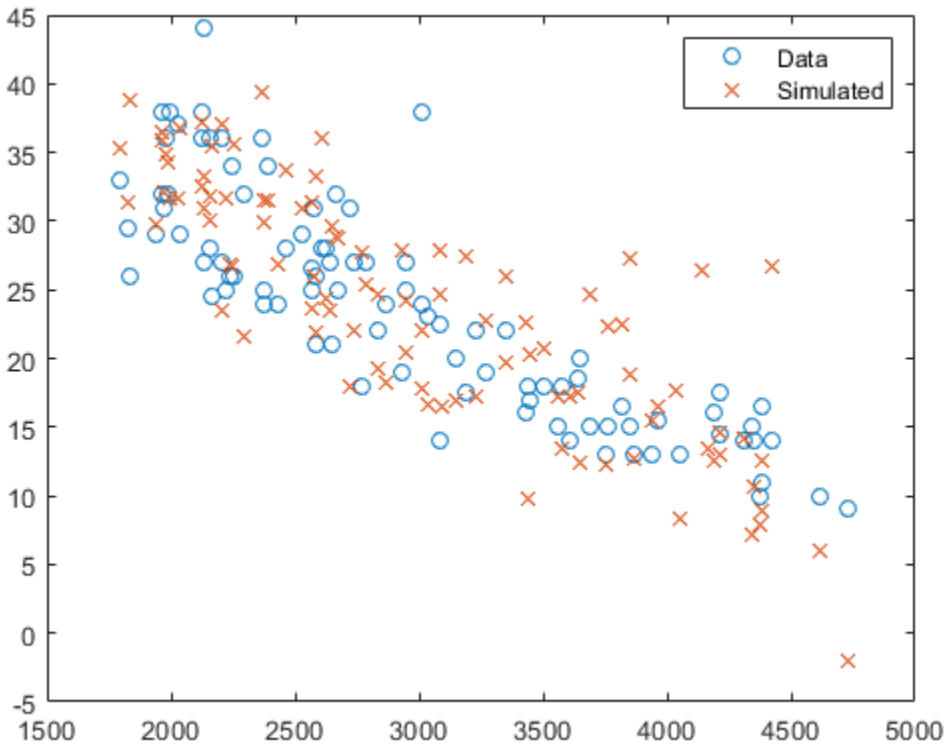
```
load carsmall
X = Weight;
y = MPG;
modelfun = 'y ~ b1 + b2*exp(-b3*x/1000)';
beta0 = [1 1 1];
mdl = fitnlm(X,y,modelfun,beta0);
```

Create simulated responses to the data.

```
Xnew = X;
ysim = random(mdl,Xnew);
```

Plot the original responses and the simulated responses to see how they differ.

```
plot(X,y,'o',X,ysim,'x')
legend('Data','Simulated')
```



Alternatives

For predictions without added noise, use `predict`.

See Also

`NonLinearModel` | `feval` | `predict`

Topics

“Predict or Simulate Responses Using a Nonlinear Model” on page 13-9
“Nonlinear Regression” on page 13-2

random

Class: `RepeatedMeasuresModel`

Generate new random response values given predictor values

Syntax

```
ysim = random(rm,tnew)
```

Description

`ysim = random(rm,tnew)` generates random response values from the repeated measures model `rm` using the predictor variables from table `tnew`.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`tnew` — New data

table used to create `rm` (default) | table

New data including the values of the response variables and the between-subject factors used as predictors in the repeated measures model, `rm`, specified as a table. `tnew` must contain all of the between-subject factors used to create `rm`.

Output Arguments

`ysim` — Random response values

n-by-*r* matrix

Random response values random generates, returned as an *n*-by-*r* matrix, where *n* is the number of rows in `tnew`, and *r* is the number of repeated measures in `rm`.

Examples

Randomly Generate New Response Values

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, and virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = dataset([1 2 3 4]','VarNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the `species` is the predictor variable.

```
  rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Randomly generate new response values.

```
ysim = random(rm);
```

`random` uses the predictor values in the original sample data you use to fit the repeated measures model `rm` in table `t`.

Randomly Generate Response Values Using New Data

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures *y*1 through *y*8 as responses. The table `within` includes the within-subject variables *w*1 and *w*2. This is simulated data.

Fit a repeated measures model, where the repeated measures *y*1 through *y*8 are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Define a table with new values for the predictor variables.

```
tnew = table(16,93,{'B'},{'Male'},'VariableNames',{'Age','IQ','Group','Gender'})
```

```
tnew=1×4 table
    Age    IQ    Group     Gender
    ___    __    _____    ________

    16     93    {'B'}    {'Male'}
```

Randomly generate new response values using the values in the new table `tnew`.

```
ysim = random(rm,tnew)
```

```
ysim = 1×8

   46.2252   66.8003  -40.4987   -1.9930   27.5213  -37.9809    4.8905   -3.7568
```

Algorithms

`random` computes `ysim` by creating predicted values and adding random noise values. For each row, the noise has a multivariate normal distribution with covariance the same as `rm.Covariance`.

See Also

`fitrm` | `predict`

randomEffects

Class: `GeneralizedLinearMixedModel`

Estimates of random effects and related statistics

Syntax

```
B = randomEffects(glme)
[B,BNames] = randomEffects(glme)
[B,BNames,stats] = randomEffects(glme)
[B,BNames,stats] = randomEffects(glme,Name,Value)
```

Description

`B = randomEffects(glme)` returns the estimates of the empirical Bayes predictors (EPBs) of random effects in the generalized linear mixed-effects model `glme` conditional on the estimated covariance parameters and the observed response.

`[B,BNames] = randomEffects(glme)` also returns the names of the coefficients, `BNames`. Each name corresponds to a coefficient in `B`.

`[B,BNames,stats] = randomEffects(glme)` also returns related statistics, `stats`, for the estimated EBPs of random effects in `glme`.

`[B,BNames,stats] = randomEffects(glme,Name,Value)` returns any of the above output arguments using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence interval level, or the method for computing the approximate degrees of freedom.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range [0,1]

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range [0,1]. For a value α, the confidence level is 100 × (1 – α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`DFMethod` — Method for computing approximate degrees of freedom
`'residual'` (default) | `'none'`

Method for computing approximate degrees of freedom, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

Value	Description
`'residual'`	The degrees of freedom value is assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'none'`	The degrees of freedom is set to infinity.

Example: `'DFMethod','none'`

Output Arguments

`B` — Estimated empirical Bayes predictors for the random effects
column vector

Estimated empirical Bayes predictors (EBPs) for the random effects in the generalized linear mixed-effects model `glme`, returned as a column vector. The EBPs in `B` are approximated by the mode of the empirical posterior distribution of the random effects given the estimated covariance parameters and the observed response.

Suppose `glme` has R grouping variables g_1, g_2, ..., g_R, with levels m_1, m_2, ..., m_R, respectively. Also suppose q_1, q_2, ..., q_R are the lengths of the random-effects vectors that are associated with g_1, g_2, ..., g_R, respectively. Then, `B` is a column vector of length $q_1*m_1 + q_2*m_2 + ... + q_R*m_R$.

`randomEffects` creates `B` by concatenating the empirical Bayes predictors of random-effects vectors corresponding to each level of each grouping variable as $[g_1\text{level}_1; g_1\text{level}_2; ...; g_1\text{level}_{m_1}; g_2\text{level}_1; g_2\text{level}_2; ...; g_2\text{level}_{m_2}; ...; g_R\text{level}_1; g_R\text{level}_2; ...; g_R\text{level}_{m_R}]'$.

`BNames` — Names of random-effects coefficients
table

Names of random-effects coefficients in `B`, returned as a table.

`stats` — Estimated empirical Bayes predictors and related statistics
table

Estimated empirical Bayes predictors (EBPs) and related statistics for the random effects in the generalized linear mixed-effects model `glme`, returned as a table. `stats` has one row for each of the random effects, and one column for each of the following statistics.

Column Name	Description
`Group`	Grouping variable associated with the random effect

Column Name	Description
`Level`	Level within the grouping variable corresponding to the random effect
`Name`	Name of the random-effect coefficient
`Estimate`	Empirical Bayes predictor (EBP) of random effect
`SEPred`	Square root of the conditional mean squared error of prediction (CMSEP) given covariance parameters and response
`tStat`	*t*-statistic for a test that the random-effects coefficient is equal to 0
`DF`	Estimated degrees of freedom for the *t*-statistic
`pValue`	*p*-value for the *t*-statistic
`Lower`	Lower limit of a 95% confidence interval for the random-effects coefficient
`Upper`	Upper limit of a 95% confidence interval for the random-effects coefficient

`randomEffects` computes the confidence intervals using the conditional mean squared error of prediction (CMSEP) approach conditional on the estimated covariance parameters and the observed response. An alternative interpretation of the confidence intervals is that they are approximate Bayesian credible intervals conditional on the estimated covariance parameters and the observed response.

When fitting a GLME model using `fitglme` and one of the pseudo likelihood fit methods (`'MPL'` or `'REMPL'`), `randomEffects` computes confidence intervals and related statistics based on the fitted linear mixed-effects model from the final pseudo likelihood iteration.

Examples

Compute and Plot Estimated Random Effects

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)

- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Compute and display the names and estimated values of the empirical Bayes predictors (EBPs) for the random effects.

```
[B,BNames] = randomEffects(glme)
```

```
B = 20×1

    0.2913
    0.1542
   -0.2633
   -0.4257
    0.5453
   -0.1069
    0.3040
   -0.1653
   -0.1458
```

```
   -0.0816
      ⋮


BNames=20×3 table
       Group       Level          Name
    ___________    ______    _______________

    {'factory'}    {'1' }    {'(Intercept)'}
    {'factory'}    {'2' }    {'(Intercept)'}
    {'factory'}    {'3' }    {'(Intercept)'}
    {'factory'}    {'4' }    {'(Intercept)'}
    {'factory'}    {'5' }    {'(Intercept)'}
    {'factory'}    {'6' }    {'(Intercept)'}
    {'factory'}    {'7' }    {'(Intercept)'}
    {'factory'}    {'8' }    {'(Intercept)'}
    {'factory'}    {'9' }    {'(Intercept)'}
    {'factory'}    {'10'}    {'(Intercept)'}
    {'factory'}    {'11'}    {'(Intercept)'}
    {'factory'}    {'12'}    {'(Intercept)'}
    {'factory'}    {'13'}    {'(Intercept)'}
    {'factory'}    {'14'}    {'(Intercept)'}
    {'factory'}    {'15'}    {'(Intercept)'}
    {'factory'}    {'16'}    {'(Intercept)'}
      ⋮
```

Each row of `B` contains the estimated EPB for the random-effects coefficient named in the corresponding row of `Bnames`. For example, the value –0.2633 in row 3 of `B` is the estimated EPB for `'(Intercept)'` for level `'3'` of `factory`.

Compute 99% Confidence Intervals for Random Effects

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Compute and display the 99% confidence intervals for the random-effects coefficients.

```
[B,BNames,stats] = randomEffects(glme,'Alpha',0.01);
stats
```

```
stats =
    Random effect coefficients: DFMethod = 'residual', Alpha = 0.01

    Group              Level         Name                  Estimate     SEPred
    {'factory'}        {'1' }        {'(Intercept)'}         0.29131    0.19163
    {'factory'}        {'2' }        {'(Intercept)'}         0.15423    0.19216
    {'factory'}        {'3' }        {'(Intercept)'}        -0.26325    0.21249
    {'factory'}        {'4' }        {'(Intercept)'}        -0.42568    0.21667
    {'factory'}        {'5' }        {'(Intercept)'}          0.5453    0.17963
    {'factory'}        {'6' }        {'(Intercept)'}        -0.10692    0.20133
    {'factory'}        {'7' }        {'(Intercept)'}         0.30404    0.18397
    {'factory'}        {'8' }        {'(Intercept)'}        -0.16527    0.20505
```

```
{'factory'}        {'9' }        {'(Intercept)'}         -0.14577      0.203
{'factory'}        {'10'}        {'(Intercept)'}        -0.081632    0.20256
{'factory'}        {'11'}        {'(Intercept)'}         0.014529    0.21421
{'factory'}        {'12'}        {'(Intercept)'}          0.17706    0.20721
{'factory'}        {'13'}        {'(Intercept)'}          0.24872    0.20522
{'factory'}        {'14'}        {'(Intercept)'}          0.21145    0.20678
{'factory'}        {'15'}        {'(Intercept)'}           0.2777    0.20345
{'factory'}        {'16'}        {'(Intercept)'}         -0.25175    0.22568
{'factory'}        {'17'}        {'(Intercept)'}         -0.13507    0.22301
{'factory'}        {'18'}        {'(Intercept)'}          -0.1627    0.22269
{'factory'}        {'19'}        {'(Intercept)'}         -0.32083    0.23294
{'factory'}        {'20'}        {'(Intercept)'}         0.058418    0.21481


tStat       DF    pValue       Lower       Upper
  1.5202    94      0.13182    -0.21251    0.79514
 0.80259    94      0.42423      -0.351    0.65946
 -1.2389    94      0.21846    -0.82191    0.29541
 -1.9646    94     0.052408    -0.99534    0.14398
  3.0356    94    0.0031051    0.073019     1.0176
-0.53105    94      0.59664    -0.63625    0.42241
  1.6527    94      0.10173    -0.17964    0.78771
-0.80597    94      0.42229    -0.70438    0.37385
-0.71806    94       0.4745    -0.67949    0.38795
  -0.403    94      0.68786    -0.61419    0.45093
0.067826    94      0.94607    -0.54866    0.57772
 0.85446    94      0.39502    -0.36774    0.72185
   1.212    94      0.22857    -0.29083    0.78827
  1.0226    94      0.30913    -0.33221    0.75511
   1.365    94      0.17552    -0.25719    0.81259
 -1.1156    94      0.26746    -0.84509    0.34158
-0.60568    94      0.54619     -0.7214    0.45125
-0.73061    94      0.46684    -0.74817    0.42278
 -1.3773    94      0.17168    -0.93325    0.29159
 0.27195    94      0.78626    -0.50635    0.62319
```

The first three columns of `stats` contain the group name, level, and random-effects coefficient name. Column 4 contains the estimated EBP of the random-effects coefficient. The last two columns of `stats`, `Lower` and `Upper`, contain the lower and upper bounds of the 99% confidence interval, respectively. For example, for the coefficient for `'(Intercept)'` for level `3` of `factory`, the estimated EBP is -0.26325, and the 99% confidence interval is [-0.82191,0.29541].

References

[1] Booth, J.G., and J.P. Hobert. “Standard Errors of Prediction in Generalized Linear Mixed Models.” *Journal of the American Statistical Association*, Vol. 93, 1998, pp. 262–272.

See Also

`GeneralizedLinearMixedModel` | `coefCI` | `coefTest` | `fixedEffects`

randomEffects

Class: `LinearMixedModel`

Estimates of random effects and related statistics

Syntax

```
B = randomEffects(lme)
[B,Bnames] = randomEffects(lme)
[B,Bnames,stats] = randomEffects(lme)
[B,Bnames,stats] = randomEffects(lme,Name,Value)
```

Description

`B = randomEffects(lme)` returns the estimates of the best linear unbiased predictors (BLUPs) of random effects in the linear mixed-effects model `lme`.

`[B,Bnames] = randomEffects(lme)` also returns the names of the coefficients in `Bnames`. Each name corresponds to a coefficient in `B`.

`[B,Bnames,stats] = randomEffects(lme)` also returns the estimated BLUPs of random effects in the linear mixed-effects model `lme` and related statistics.

`[B,Bnames,stats] = randomEffects(lme,Name,Value)` also returns the BLUPs of random effects in the linear mixed-effects model `lme` and related statistics with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. For a value α, the confidence level is 100*(1–α)%.

For example, for 99% confidence intervals, you can specify the confidence level as follows.

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`DFMethod` — Method for computing approximate degrees of freedom
`'residual'` (default) | `'satterthwaite'` | `'none'`

Method for computing approximate degrees of freedom for the *t*-statistics that test the random-effects coefficients against 0, specified as the comma-separated pair consisting of `'DFMethod'` and one of the following.

`'residual'`	Default. The degrees of freedom are assumed to be constant and equal to $n - p$, where n is the number of observations and p is the number of fixed effects.
`'satterthwaite'`	Satterthwaite approximation.
`'none'`	All degrees of freedom are set to infinity.

For example, you can specify the Satterthwaite approximation as follows.

Example: `'DFMethod','satterthwaite'`

Output Arguments

`B` — Estimated best linear unbiased predictors of random effects
column vector

Estimated best linear unbiased predictors of random effects of linear mixed-effects model `lme`, returned as a column vector.

Suppose `lme` has R grouping variables g_1, g_2, ..., g_R, with levels m_1, m_2, ..., m_R, respectively. Also suppose q_1, q_2, ..., q_R are the lengths of the random-effects vectors that are associated with g_1, g_2, ..., g_R, respectively. Then, `B` is a column vector of length $q_1*m_1 + q_2*m_2 + ... + q_R*m_R$.

`randomEffects` creates `B` by concatenating the best linear unbiased predictors of random-effects vectors corresponding to each level of each grouping variable as $[g_1level_1; g_1level_2; ...; g_1level_{m_1}; g_2level_1; g_2level_2; ...; g_2level_{m_2}; ...; g_Rlevel_1; g_Rlevel_2; ...; g_Rlevel_{m_R}]'$.

`Bnames` — Names of random-effects coefficients
table

Names of random-effects coefficients in `B`, returned as a table.

`stats` — Estimates of random effects BLUPs and related statistics
dataset array

Estimates of random effects BLUPs and related statistics, returned as a dataset array that has one row for each of the fixed effects and one column for each of the following statistics.

`Group`	Grouping variable associated with the random effect
`Level`	Level within the grouping variable corresponding to the random effect

`Name`	Name of the random-effect coefficient
`Estimate`	Best linear unbiased predictor (BLUP) of random effect
`SEPred`	Standard error of the estimate (BLUP minus random effect)
`tStat`	*t*-statistic for a test that the random effect is zero
`DF`	Estimated degrees of freedom for the *t*-statistic
`pValue`	*p*-value for the *t*-statistic
`Lower`	Lower limit of a 95% confidence interval for the random effect
`Upper`	Upper limit of a 95% confidence interval for the random effect

Examples

Display Random-Effects Estimates and Coefficient Names

Load the sample data.

```
load carbig
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration and horsepower, and potentially correlated random effects for intercept and acceleration, grouped by the model year. First, store the data in a table.

```
tbl = table(Acceleration,Horsepower,Model_Year,MPG);
```

Fit the model.

```
lme = fitlme(tbl, 'MPG ~ Acceleration + Horsepower + (Acceleration|Model_Year)');
```

Compute the BLUPs of the random-effects coefficients and display the names of the corresponding random effects.

```
[B,Bnames] = randomEffects(lme)
```

```
B = 26×1

    3.1270
   -0.2426
   -1.6532
   -0.0086
    1.2075
   -0.2179
    4.4107
   -0.4887
   -1.3103
   -0.0208
      ⋮


Bnames=26×3 table
        Group         Level           Name
```

```
    _____________    ______    _________________

    {'Model_Year'}    {'70'}    {'(Intercept)' }
    {'Model_Year'}    {'70'}    {'Acceleration'}
    {'Model_Year'}    {'71'}    {'(Intercept)' }
    {'Model_Year'}    {'71'}    {'Acceleration'}
    {'Model_Year'}    {'72'}    {'(Intercept)' }
    {'Model_Year'}    {'72'}    {'Acceleration'}
    {'Model_Year'}    {'73'}    {'(Intercept)' }
    {'Model_Year'}    {'73'}    {'Acceleration'}
    {'Model_Year'}    {'74'}    {'(Intercept)' }
    {'Model_Year'}    {'74'}    {'Acceleration'}
    {'Model_Year'}    {'75'}    {'(Intercept)' }
    {'Model_Year'}    {'75'}    {'Acceleration'}
    {'Model_Year'}    {'76'}    {'(Intercept)' }
    {'Model_Year'}    {'76'}    {'Acceleration'}
    {'Model_Year'}    {'77'}    {'(Intercept)' }
    {'Model_Year'}    {'77'}    {'Acceleration'}
      ⋮
```

Since intercept and acceleration have potentially correlated random effects, grouped by model year of the cars, `randomEffects` creates a separate row for intercept and acceleration at each level of the grouping variable.

Compute the covariance parameters of the random effects.

```
[~,~,stats] = covarianceParameters(lme)
```

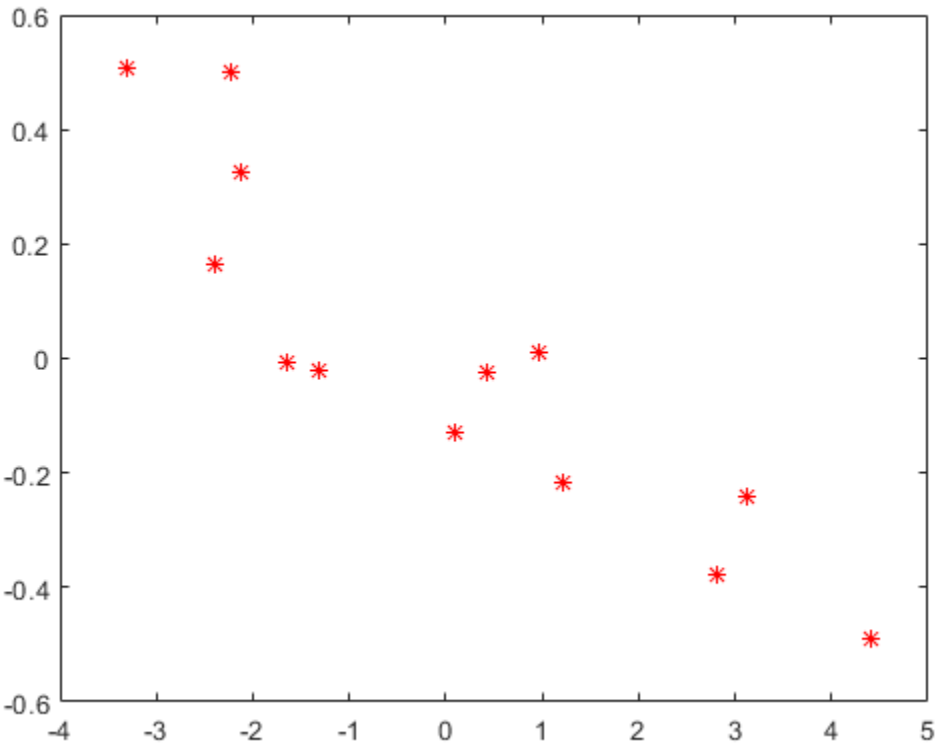
```
stats=2×1 cell array
    {3x7 classreg.regr.lmeutils.titleddataset}
    {1x5 classreg.regr.lmeutils.titleddataset}
```

```
stats{1}
```

```
ans =
    Covariance Type: FullCholesky

    Group         Name1                  Name2                  Type
    Model_Year    {'(Intercept)' }       {'(Intercept)' }       {'std' }
    Model_Year    {'Acceleration'}       {'(Intercept)' }       {'corr'}
    Model_Year    {'Acceleration'}       {'Acceleration'}       {'std' }


    Estimate    Lower       Upper
      3.3475      1.2862      8.7119
    -0.87971    -0.98501    -0.29675
     0.33789      0.1825     0.62558
```

The correlation value suggests that random effects seem negatively correlated. Plot the random effects for intercept versus acceleration to confirm this.

```
plot(B(1:2:end),B(2:2:end),'r*')
```



Compute Random-Effects Estimates and Related Statistics

Load the sample data.

```
load('fertilizer.mat');
```

The dataset array includes data from a split-plot experiment, where soil is divided into three blocks based on the soil type: sandy, silty, and loamy. Each block is divided into five plots, where five different types of tomato plants (cherry, heirloom, grape, vine, and plum) are randomly assigned to these plots. The tomato plants in the plots are then divided into subplots, where each subplot is treated by one of four fertilizers. This is simulated data.

Store the data in a dataset array called `ds`, for practical purposes, and define `Tomato`, `Soil`, and `Fertilizer` as categorical variables.

```
ds = fertilizer;
ds.Tomato = nominal(ds.Tomato);
ds.Soil = nominal(ds.Soil);
ds.Fertilizer = nominal(ds.Fertilizer);
```

Fit a linear mixed-effects model, where `Fertilizer` and `Tomato` are the fixed-effects variables, and the mean yield varies by the block (soil type), and the plots within blocks (tomato types within soil types) independently.

```
lme = fitlme(ds,'Yield ~ Fertilizer * Tomato + (1|Soil) + (1|Soil:Tomato)');
```

Compute the BLUPs and related statistics for random effects.

```
[~,~,stats] = randomEffects(lme)
```

```
stats =
    Random effect coefficients: DFMethod = 'Residual', Alpha = 0.05

    Group                  Level                    Name
    {'Soil'       }        {'Loamy'         }       {'(Intercept)'}
    {'Soil'       }        {'Sandy'         }       {'(Intercept)'}
    {'Soil'       }        {'Silty'         }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Loamy Cherry'  }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Loamy Grape'   }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Loamy Heirloom'}       {'(Intercept)'}
    {'Soil:Tomato'}        {'Loamy Plum'    }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Loamy Vine'    }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Sandy Cherry'  }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Sandy Grape'   }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Sandy Heirloom'}       {'(Intercept)'}
    {'Soil:Tomato'}        {'Sandy Plum'    }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Sandy Vine'    }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Silty Cherry'  }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Silty Grape'   }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Silty Heirloom'}       {'(Intercept)'}
    {'Soil:Tomato'}        {'Silty Plum'    }       {'(Intercept)'}
    {'Soil:Tomato'}        {'Silty Vine'    }       {'(Intercept)'}


    Estimate    SEPred    tStat       DF    pValue      Lower      Upper
     1.0061     2.3374     0.43044    40     0.66918     -3.718     5.7303
    -1.5236     2.3374    -0.65181    40     0.51825    -6.2477     3.2006
    0.51744     2.3374     0.22137    40     0.82593    -4.2067     5.2416
      12.46     7.1765      1.7362    40    0.090224    -2.0443     26.964
    -2.6429     7.1765    -0.36827    40     0.71461    -17.147     11.861
     16.681     7.1765      2.3244    40    0.025269     2.1766     31.185
    -5.0172     7.1765    -0.69911    40     0.48853    -19.522     9.4872
    -4.6874     7.1765    -0.65316    40     0.51739    -19.192     9.8169
    -17.393     7.1765     -2.4235    40    0.019987    -31.897    -2.8882
    -7.3679     7.1765     -1.0267    40     0.31075    -21.872     7.1364
     -8.621     7.1765     -1.2013    40     0.23671    -23.125     5.8833
      7.669     7.1765      1.0686    40     0.29165    -6.8353     22.173
    0.28246     7.1765    0.039359    40      0.9688    -14.222     14.787
     4.9326     7.1765     0.68732    40     0.49585    -9.5718     19.437
     10.011     7.1765      1.3949    40     0.17073    -4.4935     24.515
    -8.0599     7.1765     -1.1231    40      0.2681    -22.564     6.4444
    -2.6519     7.1765    -0.36952    40     0.71369    -17.156     11.852
      4.405     7.1765      0.6138    40     0.54282    -10.099     18.909
```

The first three rows contain the random-effects estimates and the statistics for the three levels, `Loamy`, `Sandy`, and `Silty` of the grouping variable `Soil`. The corresponding *p*-values 0.66918, 0.51825, and 0.82593 indicate that these random-effects are not significantly different from 0. The following 15 rows include the BLUPS of random-effects estimates for the intercept, grouped by the variable Tomato nested in Soil, i.e. interaction of `Tomato` and `Soil`.

Compute Confidence Intervals with Specified Options

Load the sample data.

```
load shift
```

Fit a linear mixed-effects model with a random intercept grouped by operator, to assess if there is a significant difference in the performance according to the time of the shift. Use the restricted maximum likelihood method.

```
lme = fitlme(shift,'QCDev ~ Shift + (1|Operator)');
```

Compute the 99% confidence intervals for random effects using the residuals option to compute the degrees of freedom. This is the default method.

```
[~,~,stats] = randomEffects(lme,'Alpha',0.01)
```

```
stats =
    Random effect coefficients: DFMethod = 'Residual', Alpha = 0.01

    Group               Level        Name                    Estimate    SEPred
    {'Operator'}        {'1'}        {'(Intercept)'}         0.57753     0.90378
    {'Operator'}        {'2'}        {'(Intercept)'}          1.1757     0.90378
    {'Operator'}        {'3'}        {'(Intercept)'}         -2.1715     0.90378
    {'Operator'}        {'4'}        {'(Intercept)'}          2.3655     0.90378
    {'Operator'}        {'5'}        {'(Intercept)'}         -1.9472     0.90378


    tStat      DF    pValue      Lower       Upper
    0.63902    12     0.53482     -2.1831     3.3382
     1.3009    12     0.21772     -1.5849     3.9364
    -2.4027    12    0.033352     -4.9322    0.58909
     2.6174    12    0.022494    -0.39511     5.1261
    -2.1546    12    0.052216     -4.7079    0.81337
```

Compute the 99% confidence intervals for random effects using the Satterthwaite approximation to compute the degrees of freedom.

```
[~,~,stats] = randomEffects(lme,'DFMethod','satterthwaite','Alpha',0.01)
```

```
stats =
    Random effect coefficients: DFMethod = 'Satterthwaite', Alpha = 0.01

    Group               Level        Name                    Estimate    SEPred
    {'Operator'}        {'1'}        {'(Intercept)'}         0.57753     0.90378
    {'Operator'}        {'2'}        {'(Intercept)'}          1.1757     0.90378
    {'Operator'}        {'3'}        {'(Intercept)'}         -2.1715     0.90378
    {'Operator'}        {'4'}        {'(Intercept)'}          2.3655     0.90378
    {'Operator'}        {'5'}        {'(Intercept)'}         -1.9472     0.90378


    tStat      DF        pValue      Lower       Upper
    0.63902    6.4253      0.5449      -2.684     3.839
     1.3009    6.4253     0.23799     -2.0858    4.4372
    -2.4027    6.4253    0.050386      -5.433      1.09
     2.6174    6.4253    0.037302    -0.89598     5.627
    -2.1546    6.4253    0.071626     -5.2087    1.3142
```

The Satterthwaite method usually produces smaller values for the degrees of freedom (`DF`), which results in larger p-values (`pValue`) and larger confidence intervals (`Lower` and `Upper`) for the random-effects estimates.

See Also

`LinearMixedModel` | `coefCI` | `coefTest` | `fitlme` | `fixedEffects`

randsample

Random sample

Syntax

```
y = randsample(n,k)
y = randsample(population,k)
y = randsample( ___ ,replacement)
y = randsample(n,k,true,w)
y = randsample(population,k,true,w)
y = randsample(s, ___ )
```

Description

`y = randsample(n,k)` returns `k` values sampled uniformly at random, without replacement, from the integers `1` to `n`.

`y = randsample(population,k)` returns a vector of `k` values sampled uniformly at random, without replacement, from the values in the vector `population`.

`y = randsample( ___ ,replacement)` returns a sample taken with replacement if `replacement` is `true`, or without replacement if `replacement` is `false`. Specify `replacement` following any of the input argument combinations in the previous syntaxes.

`y = randsample(n,k,true,w)` uses a vector of non-negative weights, `w`, whose length is `n`, to determine the probability that an integer `i` is selected as an entry for `y`.

`y = randsample(population,k,true,w)` uses a vector of nonnegative weights, `w`, of the same length as the vector `population`, to determine the probability that a value `population(i)` is selected as an entry for `y`.

`y = randsample(s, ___ )` uses the stream `s` for random number generation. The option `s` can precede any of the input arguments in the previous syntaxes. `s` is a member of the `RandStream` class.

Examples

Sample Unique Value from Range

Draw a single value from the integers `1` through `10`.

```
n = 10;
x = randsample(n,1)
```

```
x = 9
```

Sample from Population Vector

Create the random seed for reproducibility of the results.

```
s = RandStream('mlfg6331_64');
```

Draw a single value from the vector `[10:20]`.

```
population = 10:20;
y = randsample(s,population,1)
```

```
y = 17
```

Generate Random Sequence for Specified Probabilities

Create the random number stream for reproducibility.

```
s = RandStream('mlfg6331_64');
```

Choose 48 characters randomly and with replacement from the sequence `ACGT`, according to the specified probabilities.

```
R = randsample(s,'ACGT',48,true,[0.15 0.35 0.35 0.15])
```

```
R =
'GGCGGCGCAAGGCGCCGGACCTGGCTGCACGCCGTTCCCTGCTACTCG'
```

Set Random Number Stream

Create the random number stream for reproducibility.

```
s = RandStream('mlfg6331_64');
```

Draw five values with replacement from the integers `1:10`.

```
y = randsample(s,10,5,true)
```

```
y = 5×1

     7
     8
     5
     7
     8
```

Input Arguments

n — Upper limit of range

positive integer

Upper limit of the range (`1` to `n`) from which to sample, specified as a positive integer. By default, `randsample` samples uniformly at random, without replacement, from the values in the range `1` to `n`.

Data Types: `single` | `double`

`population` — Input data
vector

Input data from which to sample, specified as a vector. By default, `randsample` samples uniformly at random, without replacement, from the values in `population`. The orientation of `y` (row or column) is the same as that of `population`.

If `population` is a numeric vector containing only nonnegative integer values, and `population` can have the length 1, then use `y = population(randsample(length(population),k))` instead of `y = randsample(population,k)`.

Example: `y = randsample([50:100],20)` returns a vector of `20` values sampled uniformly at random, without replacement, from the `population` vector consisting of integers from `50` to `100`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `categorical`

`k` — Number of samples
positive integer

Number of samples, specified as a positive integer.

Example: `randsample(20,10)` returns a vector of `10` values sampled uniformly at random, without replacement, from the integers `1` to `20`.

Data Types: `single` | `double`

`replacement` — Indicator for sampling with replacement
false (default) | true

Indicator for sampling with replacement, specified as either `false` or `true`.

Example: `randsample(10,2,true)` returns two values with replacement from the integers `1` to `10`.

Data Types: `logical`

`w` — Sampling weights
`ones(n,1)` (default) | vector of nonnegative scalar values

Sampling weights, specified as a vector of nonnegative scalar values. The length of `w` must be equal to the range of integers to sample or the length of `population`. The vector `w` must have at least one positive value. If `w` contains negative values or `NaN` values, `randsample` displays an error message. The `randsample` function samples with probability proportional to `w(i)/sum(w)`. Usually, `w` is a vector of probabilities. The `randsample` function supports specifying weights only for sampling with replacement.

Example: `[0.1 0.5 0.35 0.46]`

Data Types: `single` | `double`

`s` — Random number stream
MATLAB default random number stream (default) | `RandStream`

Random number stream, specified as the MATLAB default random number stream or `RandStream`. For details, see “Creating and Controlling a Random Number Stream”.

Example: `s = RandStream('mlfg6331_64')` creates a random number stream that uses the multiplicative lagged Fibonacci generator algorithm.

Output Arguments

y — Sample

vector | scalar

Sample, returned as a vector or scalar.

- If `k = 1`, then `y` is a scalar.
- If `k > 1`, then `y` is a *k*-by-1 vector.

Tips

To sample data randomly, with or without replacement, use `datasample`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- When you sample without replacement, the order of the output values might not match the order in MATLAB.
- Code generation does not support the random number stream input argument `s`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`RandStream` | `datasample` | `rand` | `randperm`

Introduced before R2006a

randtool

Interactive random number generation

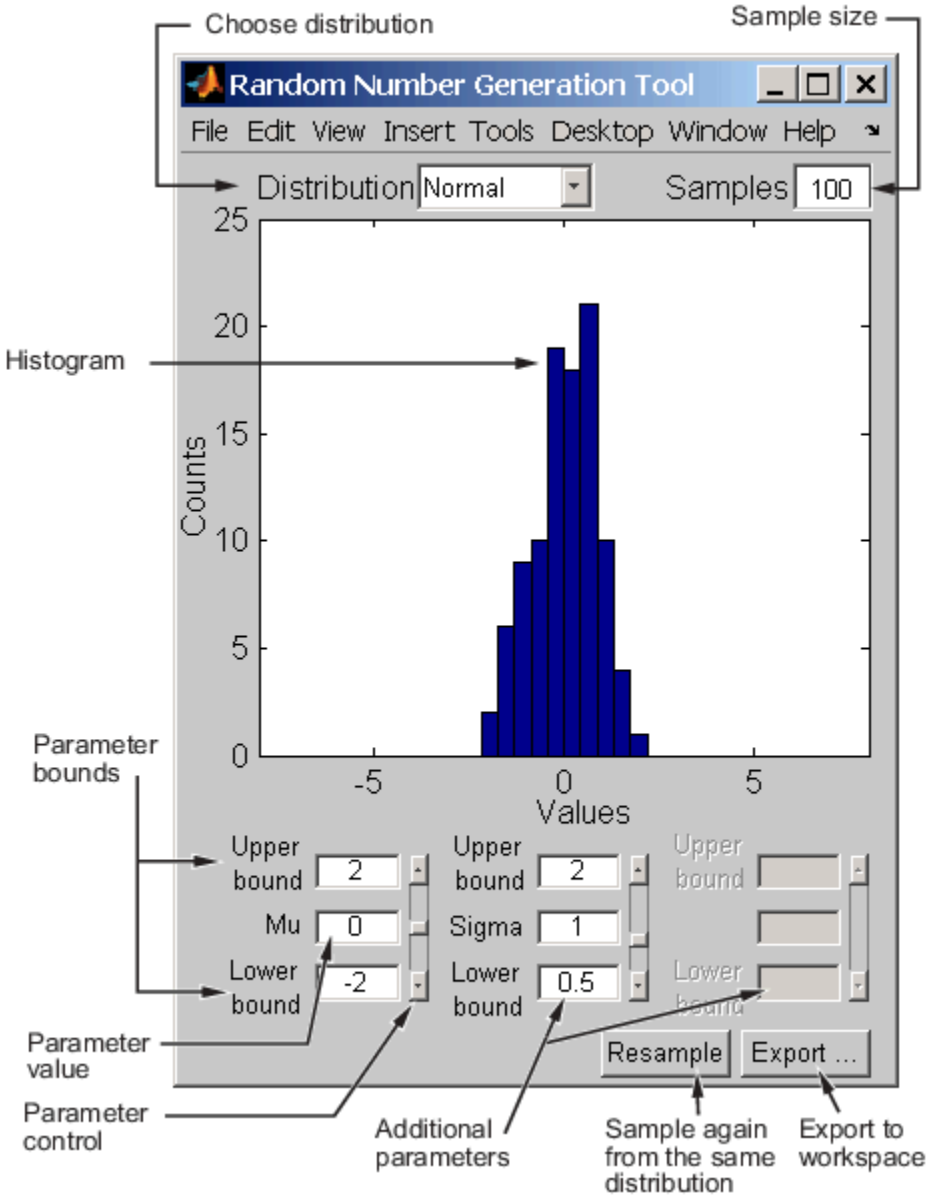
Syntax

```
randtool
```

Description

`randtool` opens the Random Number Generation Tool.

The Random Number Generation Tool is a graphical user interface that generates random samples from specified probability distributions and displays the samples as histograms. Use the tool to explore the effects of changing parameters and sample size on the distributions.



Start by selecting a distribution, then enter the desired sample size.

You can also

- Use the controls at the bottom of the window to set parameter values for the distribution and to change their upper and lower bounds.
- Draw another sample from the same distribution, with the same size and parameters.
- Export the current sample to your workspace. A dialog box enables you to provide a name for the sample.

`randttool` does not provide printing functionality in MATLAB Online.

Introduced before R2006a

range

Range of values

Syntax

```
y = range(X)
y = range(X,'all')
y = range(X,dim)
y = range(X,vecdim)
```

Description

`y = range(X)` returns the difference between the maximum and minimum values of sample data in `X`.

- If `X` is a vector, then `range(X)` is the range of the values in `X`.
- If `X` is a matrix, then `range(X)` is a row vector containing the range of each column in `X`.
- If `X` is a multidimensional array, then `range` operates along the first nonsingleton dimension of `X`, treating the values as vectors. The size of this dimension becomes 1 while the sizes of all other dimensions remain the same. If `X` is an empty array with first dimension 0, then `range(X)` returns an empty array with the same size as `X`.

`y = range(X,'all')` returns the range of all elements in `X`.

`y = range(X,dim)` returns the range along the operating dimension `dim` of `X`. For example, if `X` is a matrix, then `range(X,2)` is a column vector containing the range value of each row.

`y = range(X,vecdim)` returns the range over the dimensions specified in the vector `vecdim`. For example, if `X` is a matrix, then `range(X,[1 2])` is the range of all elements in `X` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

Examples

Range of Standard Normal Random Numbers

Generate five large samples of standard normal random numbers.

```
rng('default') % For reproducibility
rv = normrnd(0,1,1000,5);
```

Find the range values of the samples.

```
near6 = range(rv)
```

```
near6 = 1×5

    6.8104    6.6420    6.9578    6.0860    6.8165
```

The range value is approximately 6 for each sample.

Range of Exam Scores

Find the range of exam scores over the course of five exams.

Load the `examgrades` data set.

```
load examgrades
```

Find the range of all exam scores.

```
y = range(grades,'all')
```

```
y = 46
```

Range of Exam Scores by Student

Find the range of exam scores for each student over the course of five exams.

Load the `examgrades` data set.

```
load examgrades
```

Find the range of exam scores for the first 10 students. For example, the difference between the eighth student's best and worst exam scores is 7 points.

```
X = grades(1:10,:);
y = range(X,2)
```

```
y = 10×1

    12
    13
    10
    12
     8
    16
    14
     7
    12
    10
```

Range Along Vector Dimension

Find the range of a multidimensional array over multiple dimensions.

Create a 3-by-5-by-2 array of normal random numbers with mean `mu = 2` and standard deviation `sigma = 7`.

```
rng('default') % For reproducibility
mu = 2;
```

```
sigma = 7;
X = normrnd(mu,sigma,[3 5 2])

X =
X(:,:,1) =

    5.7637    8.0352   -1.0351   21.3861    7.0778
   14.8372    4.2314    4.3984   -7.4492    1.5586
  -13.8119   -7.1538   27.0488   23.2445    7.0032


X(:,:,2) =

    0.5652   11.8632   -6.4524    5.4223   -0.1241
    1.1310   11.9203    7.0207    9.2429    4.0571
   12.4279    6.7005   13.4116    7.0882   -3.5110
```

Find the range of each page of `X` by specifying dimensions 1 and 2 as the operating dimensions.

```
ypage = range(X,[1 2])

ypage =
ypage(:,:,1) =

   40.8607


ypage(:,:,2) =

   19.8641
```

For example, `ypage(1,1,2)` is the range of all the elements in `X(:,:,2)`.

Find the range of the elements in each `X(i,:,:)` slice by specifying dimensions 2 and 3 as the operating dimensions.

```
yrow = range(X,[2 3])

yrow = 3×1

   27.8385
   22.2864
   40.8607
```

For example, `yrow(3)` is the range of all the elements in `X(3,:,:)`.

Input Arguments

X — Data sample

scalar | vector | matrix | multidimensional array

Data sample, specified as a scalar, vector, matrix, or multidimensional array.

- If `X` is a scalar, then `range(X)` is 0.
- If `X` is a 0-by-0 empty array, then `range(X)` is also an empty array.

Data Types: `single` | `double` | `logical` | `datetime` | `duration`

`dim` — Dimension to operate along
positive integer scalar

Dimension to operate along, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension whose size does not equal 1.

`dim` indicates the dimension whose length reduces to 1. `size(y,dim)` is 1, while the sizes of all other dimensions remain the same unless `size(X,dim)` is 0. If `size(X,dim)` is 0, then `range(X,dim)` returns an empty array of the same size as `X`.

Consider a two-dimensional data sample `X`:

- If `dim` is equal to 1, then `range(X,1)` returns a row vector containing the range for each column.
- If `dim` is equal to 2, then `range(X,2)` returns a column vector containing the range for each row.

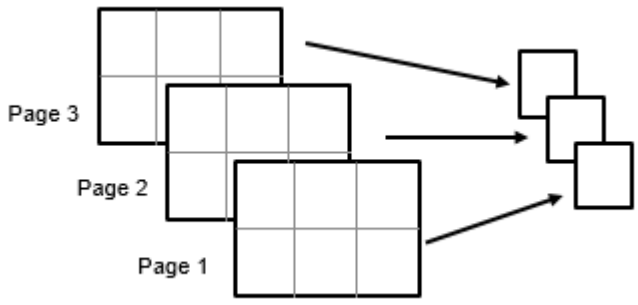
If `dim` is greater than `ndims(X)`, `range` returns an array of zeros with the same dimensions and missing values as `X`.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`.

For example, if `X` is a 2-by-3-by-3 array, then `range(X,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the range of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

`y` — Difference between maximum and minimum values
scalar | vector | matrix | multidimensional array

Difference between the maximum and minimum values, returned as a scalar, vector, matrix, or multidimensional array.

Tips

- `range` treats `NaN`s as missing values and ignores them.
- `range` provides an easily calculated estimate of the spread of a sample. Avoid using `range` with data that has outliers because they have an undue influence on this statistic.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`iqr` | `mad` | `std`

Introduced before R2006a

rangesearch

Find all neighbors within specified distance using searcher object

Syntax

```
Idx = rangesearch(Mdl,Y,r)
Idx = rangesearch(Mdl,Y,r,Name,Value)
[Idx,D] = rangesearch( ___ )
```

Description

`Idx = rangesearch(Mdl,Y,r)` searches for all neighbors (i.e., points, rows, or observations) in `Mdl.X` within radius `r` of each point (i.e., row or observation) in the query data `Y` using an exhaustive search or a *K*d-tree. `rangesearch` returns `Idx`, which is a column vector of the indices of `Mdl.X` within `r` units.

`Idx = rangesearch(Mdl,Y,r,Name,Value)` returns the indices of the observation in `Mdl.X` within radius `r` of each observation in `Y` with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify to use a different distance metric than is stored in `Mdl.Distance` or a different distance metric parameter than is stored in `Mdl.DistParameter`.

`[Idx,D] = rangesearch( ___ )` additionally returns the matrix `D` using any of the input arguments in the previous syntaxes. `D` contains the distances between the observations in `Mdl.X` within radius `r` of each observation in `Y`. By default, the function arranges the columns of `D` in ascending order by closeness, with respect to the distance metric.

Examples

Search for Neighbors Within a Radius Using *K*d-tree and Exhaustive Search

`rangesearch` accepts `ExhaustiveSearcher` or `KDTreeSearcher` model objects to search the training data for the nearest neighbors to the query data. An `ExhaustiveSearcher` model invokes the exhaustive searcher algorithm, and a `KDTreeSearcher` model defines a *K*d-tree, which `rangesearch` uses to search for nearest neighbors.

Load Fisher's iris data set. Randomly reserve five observations from the data for query data. Focus on the petal dimensions.

```
load fisheriris
rng(1); % For reproducibility
n = size(meas,1);
idx = randsample(n,5);
X = meas(~ismember(1:n,idx),3:4); % Training data
Y = meas(idx,3:4);                % Query data
```

Grow a default two-dimensional *K*d-tree.

```
MdlKDT = KDTreeSearcher(X)
```

```
MdlKDT =
  KDTreeSearcher with properties:

      BucketSize: 50
        Distance: 'euclidean'
    DistParameter: []
                X: [145x2 double]
```

`MdlKDT` is a `KDTreeSearcher` model object. You can alter its writable properties using dot notation.

Prepare an exhaustive nearest neighbor searcher.

```
MdlES = ExhaustiveSearcher(X)
```

```
MdlES =
  ExhaustiveSearcher with properties:

        Distance: 'euclidean'
    DistParameter: []
                X: [145x2 double]
```

`MdlES` is an `ExhaustiveSearcher` model object. It contains the options, such as the distance metric, to use to find nearest neighbors.

Alternatively, you can grow a *K*d-tree or prepare an exhaustive nearest neighbor searcher using `createns`.

Search training data for the nearest neighbor indices that correspond to each query observation that are within a 0.5 cm radius. Conduct both types of searches and use the default settings.

```
r = 0.15; % Search radius
IdxKDT = rangesearch(MdlKDT,Y,r);
IdxES = rangesearch(MdlES,Y,r);
[IdxKDT IdxES]
```

```
ans=5×2 cell array
    {1x27 double}    {1x27 double}
    {[       13]}    {[       13]}
    {1x27 double}    {1x27 double}
    {[    64 66]}    {[    64 66]}
    {1x0  double}    {1x0  double}
```

`IdxKDT` and `IdxES` are cell arrays of vectors corresponding to the indices of `X` that are within 0.15 cm of the observations in `Y`. Each row of the index matrices corresponds to a query observation.

Compare the results between the methods.

```
cellfun(@isequal,IdxKDT,IdxES)
```

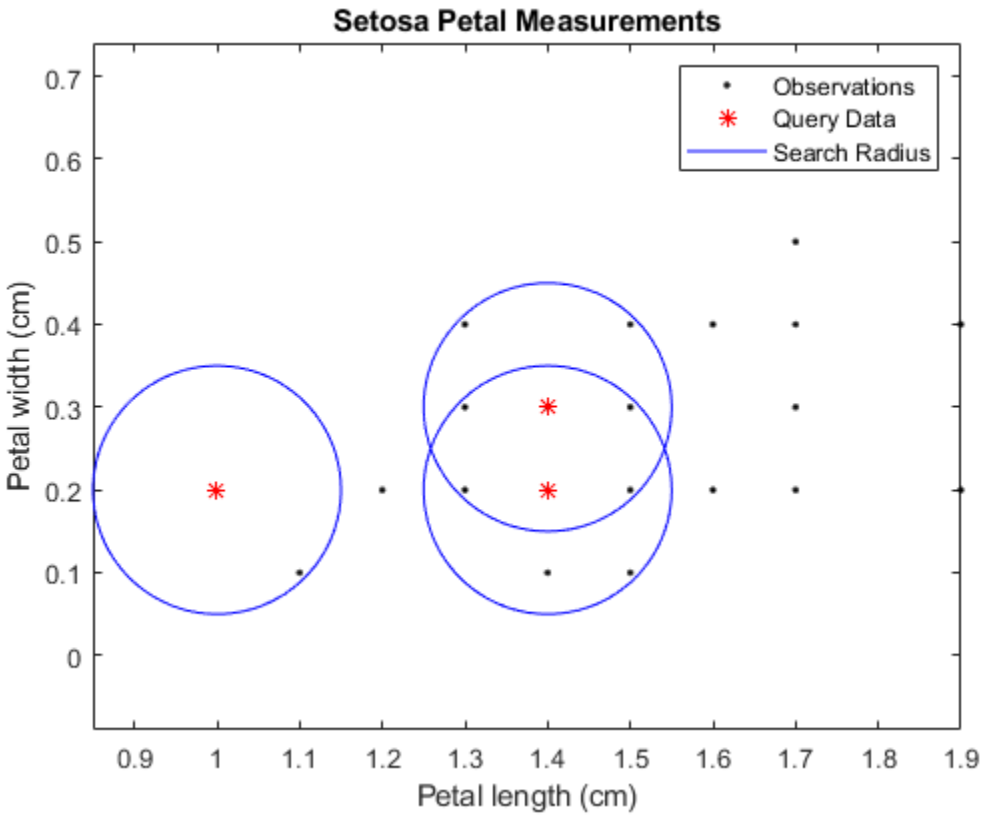
```
ans = 5x1 logical array

   1
   1
   1
   1
```

```
1
```

In this case, the results are the same.

Plot the results for the setosa irises.

```
setosaIdx = strcmp(species(~ismember(1:n,idx)),'setosa');
XSetosa = X(setosaIdx,:);
ySetosaIdx = strcmp(species(idx),'setosa');
YSetosa = Y(ySetosaIdx,:);

figure;
plot(XSetosa(:,1),XSetosa(:,2),'.k');
hold on;
plot(YSetosa(:,1),YSetosa(:,2),'*r');
for j = 1:sum(ySetosaIdx)
    c = YSetosa(j,:);
    circleFun = @(x1,x2)r^2 - (x1 - c(1)).^2 - (x2 - c(2)).^2;
    fimplicit(circleFun,[c(1) + [-1 1]*r, c(2) + [-1 1]*r],'b-')
end
xlabel 'Petal length (cm)';
ylabel 'Petal width (cm)';
title 'Setosa Petal Measurements';
legend('Observations','Query Data','Search Radius');
axis equal
hold off
```



Search for Neighbors Within a Radius Using the Mahalanobis Distance

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng(1);                      % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Prepare a default exhaustive nearest neighbor searcher.

```
Mdl = ExhaustiveSearcher(X)
```

```
Mdl =
  ExhaustiveSearcher with properties:

         Distance: 'euclidean'
    DistParameter: []
                X: [145x4 double]
```

`Mdl` is an `ExhaustiveSearcher` model.

Find the indices of the training data (`X`) that are within 0.15 cm of each point in the query data (`Y`). Specify that the distances are with respect to the Mahalanobis metric.

```
r = 1;
Idx = rangesearch(Mdl,Y,r,'Distance','mahalanobis')
```

```
Idx=5×1 cell array
    {[26 38 7 17 47 4 27 46 25 10 39 20 21 2 33]}
    {[                              6 21 25 4 19]}
    {[                           1 34 33 22 24 2]}
    {[                                       84]}
    {[                                       69]}
```

```
Idx{3}
```

```
ans = 1×6

     1    34    33    22    24     2
```

Each cell of `Idx` corresponds to a query data observation and contains in `X` a vector of indices of the neighbors within 0.15cm of the query data. `rangesearch` arranges the indices in ascending order by distance. For example, using the Mahalanobis distance, the second nearest neighbor of `Y(3,:)` is `X(34,:)`.

Compute Distances of Neighbors Within a Radius

Load Fisher's iris data set.

```
load fisheriris
```

Remove five irises randomly from the predictor data to use as a query set.

```
rng(4);                      % For reproducibility
n = size(meas,1);            % Sample size
qIdx = randsample(n,5);      % Indices of query data
X = meas(~ismember(1:n,qIdx),:);
Y = meas(qIdx,:);
```

Grow a four-dimensional *K*d-tree using the training data. Specify to use the Minkowski distance for finding nearest neighbors.

```
Mdl = KDTreeSearcher(X);
```

`Mdl` is a `KDTreeSearcher` model. By default, the distance metric for finding nearest neighbors is the Euclidean metric.

Find the indices of the training data (`X`) that are within 0.5 cm from each point in the query data (`Y`).

```
r = 0.5;
[Idx,D] = rangesearch(Mdl,Y,r);
```

`Idx` and `D` are five-element cell arrays of vectors. The vector values in `Idx` are the indices in `X`. The `X` indices represent the observations that are within 0.5 cm of the query data, `Y`. `D` contains the distances that correspond to the observations.

Display the results for query observation 3.

```
Idx{3}
```

```
ans = 1×2

   127   122
```

```
D{3}
```

```
ans = 1×2

    0.2646    0.4359
```

The closest observation to `Y(3,:)` is `X(127,:)`, which is `0.2646` cm away. The next closest is `X(122,:)`, which is `0.4359` cm away. All other observations are greater than `0.5` cm away from `Y(5,:)`.

Input Arguments

`Mdl` — Nearest neighbor searcher

`ExhaustiveSearcher` model object | `KDTreeSearcher` model object

Nearest neighbor searcher, specified as an `ExhaustiveSearcher` or `KDTreeSearcher` model object, respectively.

If `Mdl` is an `ExhaustiveSearcher` model, then `rangesearch` searches for nearest neighbors using an exhaustive search. Otherwise, `rangesearch` uses the grown *K*d-tree to search for nearest neighbors.

Y — Query data

numeric matrix

Query data, specified as a numeric matrix.

`Y` is an *m*-by-*K* matrix. Rows of `Y` correspond to observations (i.e., examples), and columns correspond to predictors (i.e., variables or features). `Y` must have the same number of columns as the training data stored in `Mdl.X`.

Data Types: `single` | `double`

r — Search radius

nonnegative scalar

Search radius around each point in the query data, specified as a nonnegative scalar.

`rangesearch` finds all observations in `Mdl.X` that are within distance `r` of each observation in `Y`. The property `Mdl.Distance` stores the distance.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Distance','minkowski','P',3` specifies to find all observations in `Mdl.X` within distance `r` of each observation in `Y`, using the Minkowski distance metric with exponent `3`.

For Both Nearest Neighbor Searchers

Distance — Distance metric

`Mdl.Distance` (default) | `'cityblock'` | `'euclidean'` | `'mahalanobis'` | `'minkowski'` | `'seuclidean'` | function handle | ...

Distance metric used to find neighbors of the training data to the query observations, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle.

For both types of nearest neighbor searchers, `rangesearch` supports these distance metrics.

Value	Description
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cityblock'`	City block distance.
`'euclidean'`	Euclidean distance.

Value	Description
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different exponent, use the 'P' name-value pair argument.

If Mdl is an ExhaustiveSearcher model object, then rangesearch also supports these distance metrics.

Value	Description
'correlation'	One minus the sample linear correlation between observations (treated as sequences of values).
'cosine'	One minus the cosine of the included angle between observations (treated as row vectors).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
'mahalanobis'	Mahalanobis distance, computed using a positive definite covariance matrix. To change the value of the covariance matrix, use the 'Cov' name-value pair argument.
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in Mdl.X and the query matrix is scaled by dividing by the corresponding element of the standard deviation computed from Mdl.X. To specify another scaling, use the 'Scale' name-value pair argument.
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

If Mdl is an ExhaustiveSearcher model object, then you can also specify a function handle for a custom distance metric by using @ (for example, @distfun). The custom distance function must:

- Have the form function D2 = distfun(ZI,ZJ).
- Take as arguments:
  - A 1-by-*K* vector ZI containing a single row from Mdl.X or Y, where *K* is the number of columns of Mdl.X.
  - An *m*-by-*K* matrix ZJ containing multiple rows of Mdl.X or Y, where *m* is a positive integer.
- Return an *m*-by-1 vector of distances D2, where D2(*j*) is the distance between the observations ZI and ZJ(*j*,:).

For more details, see “Distance Metrics” on page 18-12.

Example: 'Distance','minkowski'

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

`SortIndices` — Flag to sort returned indices according to distance
`true` (`1`) (default) | `false` (`0`)

Flag to sort returned indices according to distance, specified as the comma-separated pair consisting of `'SortIndices'` and either `true` (`1`) or `false` (`0`).

For faster performance when `Y` contains many observations that have many nearest points, you can set `SortIndices` to `false`. In this case, `rangesearch` returns the indices of the nearest points in no particular order. When `SortIndices` is `true`, the function arranges the indices of the nearest points in ascending order by distance.

Example: `'SortIndices',false`

Data Types: `logical`

For Exhaustive Nearest Neighbor Searchers

`Cov` — Covariance matrix for Mahalanobis distance metric
`cov(Mdl.X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix. `Cov` is a K-by-K matrix, where K is the number of columns of `Mdl.X`. If you specify `Cov` and do not specify `'Distance','mahalanobis'`, then `rangesearch` returns an error message.

Example: `'Cov',eye(3)`

Data Types: `single` | `double`

`Scale` — Scale parameter value for standardized Euclidean distance metric
`std(Mdl.X,'omitnan')` (default) | nonnegative numeric vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative numeric vector. `Scale` has length K, where K is the number of columns of `Mdl.X`.

The software scales each difference between the training and query data using the corresponding element of `Scale`. If you specify `Scale` and do not specify `'Distance','seuclidean'`, then `rangesearch` returns an error message.

Example: `'Scale',quantile(Mdl.X,0.75) - quantile(Mdl.X,0.25)`

Data Types: `single` | `double`

Note If you specify `'Distance'`, `'Cov'`, `'P'`, or `'Scale'`, then `Mdl.Distance` and `Mdl.DistParameter` do not change value.

Output Arguments

`Idx` — Training data indices of nearest neighbors

cell array of numeric vectors

Training data indices of nearest neighbors, returned as a cell array of numeric vectors.

`Idx` is an *m*-by-`1` cell array such that cell `j` (`Idx{j}`) contains an m_j-dimensional vector of indices of the observations in `Mdl.X` that are within `r` units to the query observation `Y(j,:)`. If `SortIndices` is `true`, then `rangesearch` arranges the elements of the vectors in ascending order by distance.

`D` — Distances of nearest neighbors to the query data

cell array of numeric vectors

Distances of the neighbors to the query data, returned as a numeric matrix or cell array of numeric vectors.

`D` is an *m*-by-`1` cell array such that cell `j` (`D{j}`) contains an m_j-dimensional vector of the distances that the observations in `Mdl.X` are from the query observation `Y(j,:)`. All elements of the vector are less than `r`. If `SortIndices` is `true`, then `rangesearch` arranges the elements of the vectors in ascending order.

Tips

`knnsearch` finds the *k* (positive integer) points in `Mdl.X` that are *k*-nearest for each `Y` point. In contrast, `rangesearch` finds all the points in `Mdl.X` that are within distance `r` (positive scalar) of each `Y` point.

Alternative Functionality

`rangesearch` is an object function that requires an `ExhaustiveSearcher` or a `KDTreeSearcher` model object, query data, and a distance. Under equivalent conditions, `rangesearch` returns the same results as `rangesearch` when you specify the name-value pair argument `'NSMethod','exhaustive'` or `'NSMethod','kdtree'`, respectively.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- This table contains notes about the arguments of `rangesearch`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	There are two ways to use `Mdl` in code generation. For an example, see “Code Generation for Nearest Neighbor Searcher” on page 32-19. • Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `rangesearch` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `rangesearch` function. Then use `codegen` to generate code for the entry-point function. • Include `coder.Constant(Mdl)` in the `-args` value of `codegen`. If `Mdl` is a `KDTreeSearcher` object, and the code generation build type is a MEX function, then `codegen` generates a MEX function using Intel Threading Building Blocks (TBB) for parallel computation. Otherwise, `codegen` generates code using `parfor`. • MEX function for the *k*d-tree search algorithm — `codegen` generates an optimized MEX function using Intel TBB for parallel computation on multicore platforms. You can use the MEX function to accelerate MATLAB algorithms. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb. If you generate the MEX function to test the generated code of the `parfor` version, you can disable the usage of Intel TBB. Set the `ExtrinsicCalls` property of the MEX configuration object to `false`. For details, see `coder.MexCodeConfig`. • MEX function for the exhaustive search algorithm and standalone C/C++ code for both algorithms — The generated code of `rangesearch` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.
`'Distance'`	• Cannot be a custom distance function. • Must be a compile-time constant; its value cannot change in the generated code.
`'SortIndices'`	Not supported. The output arguments are always sorted.
Name-value pair arguments	Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined exponent for the Minkowski distance in the generated code, include `{coder.Constant('Distance'),coder.Constant('Minkowski'),coder.Constant('P'),0}` in the `-args` value of `codegen`.

Argument	Notes and Limitations
`Idx`	• The sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision. • Starting in R2020a, `rangesearch` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for strict single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Nearest Neighbor Searcher” on page 32-19.

See Also

`ExhaustiveSearcher` | `KDTreeSearcher` | `createns` | `knnsearch` | `rangesearch`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2011b

rangesearch

Find all neighbors within specified distance using input data

Syntax

```
Idx = rangesearch(X,Y,r)
[Idx,D] = rangesearch(X,Y,r)
[Idx,D] = rangesearch(X,Y,r,Name,Value)
```

Description

`Idx = rangesearch(X,Y,r)` finds all the `X` points that are within distance `r` of the `Y` points. The rows of `X` and `Y` correspond to observations, and the columns correspond to variables.

`[Idx,D] = rangesearch(X,Y,r)` also returns the distances between the `Y` points and the `X` points that are within a distance of `r`.

`[Idx,D] = rangesearch(X,Y,r,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the nearest neighbor search method and the distance metric used in the search.

Examples

Find All Points Within Specified Distance

Find the `X` points that are within a Euclidean distance `1.5` of each `Y` point. Both `X` and `Y` are samples of five-dimensional normally distributed variables.

```
rng('default') % For reproducibility
X = randn(100,5);
Y = randn(10,5);
[Idx,D] = rangesearch(X,Y,1.5)
```

```
Idx=10×1 cell array
    {[             25 62 33 99 87 92 16]}
    {[                               92 25]}
    {[ 93 42 31 73 60 28 78 83 48 89 85]}
    {[                               92 41]}
    {[44 7 28 78 75 42 69 31 1 26 83 93]}
    {[       15 31 89 41 27 17 29 60 34]}
    {[                                  89]}
    {1x0 double                            }
    {1x0 double                            }
    {1x0 double                            }


D=10×1 cell array
    {[              0.9546 1.0987 1.2730 1.3981 1.4140 1.4249 1.4822]}
    {[                                                 1.4203 1.4558]}
    {1x11 double                                                      }
    {[                                                 1.1244 1.4672]}
```

```
    {1x12 double                                                          }
    {[1.2824 1.2843 1.3342 1.3469 1.4154 1.4237 1.4625 1.4626 1.4744]}
    {[                                                                1.1739]}
    {1x0  double                                                          }
    {1x0  double                                                          }
    {1x0  double                                                          }
```

In this case, the last three `Y` points are more than `1.5` distance away from any `X` point. `X(89,:)` is `1.1739` distance away from `Y(7,:)`, and no other `X` point is within distance `1.5` of `Y(7,:)`. `X` contains `12` points within distance `1.5` of `Y(5,:)`.

Find Nearest Points in Clustered Data

Generate 5000 random points from each of three distinct multivariate normal distributions. Shift the means of the distributions so that the randomly generated points are likely to form three separate clusters.

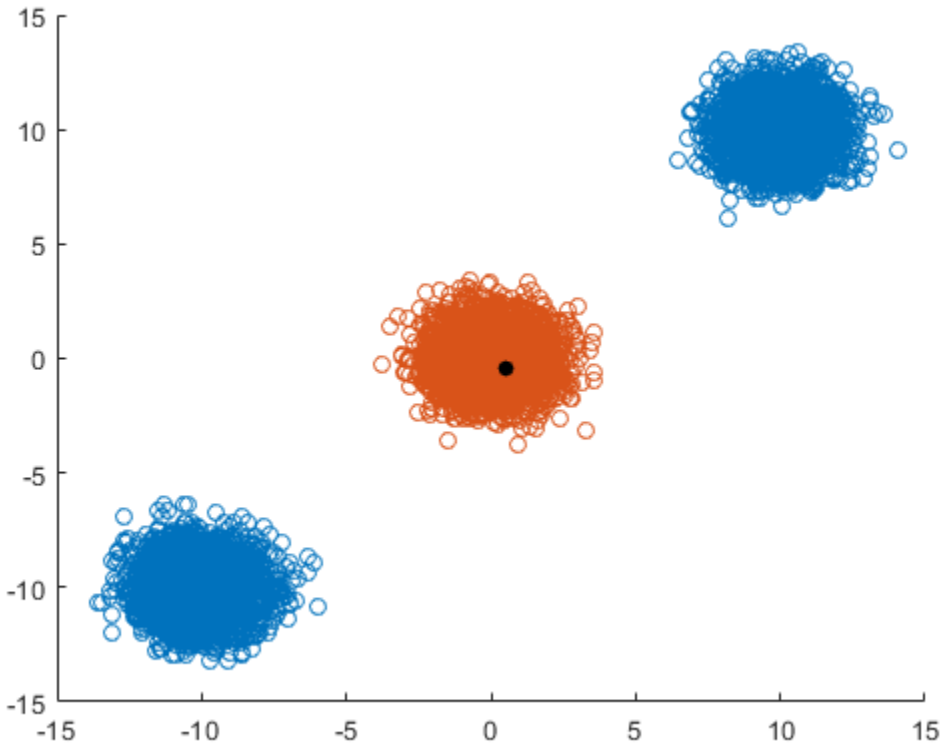
```
rng('default')  % For reproducibility
N = 5000;
dist = 10;
X = [mvnrnd([0 0],eye(2),N);
     mvnrnd(dist*[1 1],eye(2),N);
     mvnrnd(dist*[-1 -1],eye(2),N)];
```

For each point in `X`, find the points in `X` that are within a radius `dist` away from the point. For faster computation, specify to keep the indices of the nearest neighbors unsorted. Select the first point in `X`, and find its nearest neighbors.

```
Idx = rangesearch(X,X,dist,'SortIndices',false);
x = X(1,:);
nearestPoints = X(Idx{1},:);
```

Find the values in `X` that are not the nearest neighbors of `x`. Display those points in one color and the nearest neighbors of `x` in a different color. Label the point `x` with a black, filled circle.

```
nonNearestIdx = true(size(X,1),1);
nonNearestIdx(Idx{1}) = false;

scatter(X(nonNearestIdx,1),X(nonNearestIdx,2))
hold on
scatter(nearestPoints(:,1),nearestPoints(:,2))
scatter(x(1),x(2),'black','filled')
hold off
```



Find Nearest Points Using Custom Distance Function

Find the patients in the `patients` data set that are within a certain age and weight range of the patients in `Y`.

Load the `patients` data set. The `Age` values are in years, and the `Weight` values are in pounds.

```
load patients
X = [Age Weight];
Y = [20 162; 30 169; 40 168];   % New patients
```

Create a custom distance function `distfun` that determines the distance between patients in terms of age and weight. For example, according to `distfun`, two patients that are one year apart in age and have the same weight are one distance unit apart. Similarly, two patients that have the same age and are five pounds apart in weight are also one distance unit apart.

```
type distfun.m % Display contents of distfun.m file
```

```
function D2 = distfun(ZI,ZJ)
ageDifference = abs(ZI(1)-ZJ(:,1));
weightDifference = abs(ZI(2)-ZJ(:,2));
D2 = ageDifference + 0.2*weightDifference;
end
```

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the function file `distfun.m`.

Find the patients in `X` that are within the distance `2` of the patients in `Y`.

```
[Idx,D] = rangesearch(X,Y,2,'Distance',@distfun)
```

```
Idx=3×1 cell array
    {1x0 double}
    {1x0 double}
    {[      41]}


D=3×1 cell array
    {1x0 double}
    {1x0 double}
    {[  1.8000]}
```

The third patient in `Y` is the only one to have a patient in `X` within a distance of `2`.

Display the `Age` and `Weight` values for the nearest patient in `X` to the patient with age `40` and weight `168`.

```
X(Idx{3},:)
```

```
ans = 1×2

    39   164
```

Input Arguments

`X` — Input data

numeric matrix

Input data, specified as an *mx*-by-*n* numeric matrix, where each row represents one *n*-dimensional point. The number of columns *n* must equal the number of columns in `Y`.

Data Types: `single` | `double`

`Y` — Query points

numeric matrix

Query points, specified as an *my*-by-*n* numeric matrix, where each row represents one *n*-dimensional point. The number of columns *n* must equal the number of columns in `X`.

Data Types: `single` | `double`

`r` — Search radius

nonnegative scalar

Search radius around each query point, specified as a nonnegative scalar. `rangesearch` finds all `X` points (rows) that are within distance `r` of each `Y` point. The meaning of distance depends on the `'Distance'` name-value pair argument.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `rangesearch(X,Y,1.4,'Distance','seuclidean','Scale',iqr(X))` specifies to find all the observations in `X` within distance `1.4` of each observation in `Y`, using a standardized Euclidean distance scaled by the interquartile range of `X`.

`NSMethod` — Nearest neighbor search method

`'kdtree'` | `'exhaustive'`

Nearest neighbor search method, specified as the comma-separated pair consisting of `'NSMethod'` and one of these values.

Value	Description
`'kdtree'`	Create and use a *K*d-tree to find nearest neighbors. `'kdtree'` is valid only when the distance metric is one of these options: • `'chebychev'` • `'cityblock'` • `'euclidean'` • `'minkowski'`
`'exhaustive'`	Use the exhaustive search algorithm. The software computes the distances from all `X` points to each `Y` point to find nearest neighbors.

`'kdtree'` is the default value when the number of columns in `X` is less than or equal to `10`, `X` is not sparse, and the distance metric is one of the valid `'kdtree'` metrics. Otherwise, the default value is `'exhaustive'`.

Example: `'NSMethod','exhaustive'`

`Distance` — Distance metric

`'euclidean'` (default) | `'seuclidean'` | `'mahalanobis'` | `'cityblock'` | `'minkowski'` | `'chebychev'` | function handle | ...

Distance metric that `rangesearch` uses, specified as the comma-separated pair consisting of `'Distance'` and one of the values in this table.

Value	Description
`'euclidean'`	Euclidean distance.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between a row in `X` and a query point is scaled by dividing by the corresponding element of the standard deviation computed from `X`, `std(X,'omitnan')`. To specify another scaling, use the `'Scale'` name-value pair argument.

Value	Description
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix `C`. The default value of `C` is the sample covariance matrix of `X`, as computed by `cov(X,'omitrows')`. To specify a different value for `C`, use the `'Cov'` name-value pair argument.
`'cityblock'`	City block distance.
`'minkowski'`	Minkowski distance. The default exponent is `2`. To specify a different exponent, use the `'P'` name-value pair argument.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors).
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'hamming'`	Hamming distance, the percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a 1-by-*n* vector containing one row of `X` or `Y`.
	• `ZJ` is an *m*-by-*n* matrix containing multiple rows of `X` or `Y`.
	• `D2` is an *m*-by-1 vector of distances, and `D2(j)` is the distance between the observations `ZI` and `ZJ(j,:)`.

For more information, see “Distance Metrics” on page 18-12.

Example: `'Distance','minkowski'`

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar. This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',4`

Data Types: `single` | `double`

Cov — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix. This argument is valid only when `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(4)`

Data Types: `single` | `double`

`Scale` — Scale parameter value for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | nonnegative vector

Scale parameter value for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a nonnegative vector. `Scale` has length equal to the number of columns in `X`. Each coordinate difference between a row in `X` and a query point is scaled by the corresponding element of `Scale`. This argument is valid only when `'Distance'` is `'seuclidean'`.

Example: `'Scale',iqr(X)`

Data Types: `single` | `double`

`BucketSize` — Maximum number of data points in leaf node of *K*d-tree
`50` (default) | positive integer scalar

Maximum number of data points in the leaf node of the *K*d-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer scalar. This argument is valid only when `NSMethod` is `'kdtree'`.

Example: `'BucketSize',20`

Data Types: `single` | `double`

`SortIndices` — Flag to sort returned indices according to distance
`true` (`1`) (default) | `false` (`0`)

Flag to sort returned indices according to distance, specified as the comma-separated pair consisting of `'SortIndices'` and either `true` (`1`) or `false` (`0`).

For faster performance when `Y` contains many observations that have many nearest points in `X`, you can set `SortIndices` to `false`. In this case, `rangesearch` returns the indices of the nearest points in no particular order. When `SortIndices` is `true`, the function arranges the indices of the nearest points in ascending order by distance.

Example: `'SortIndices',false`

Data Types: `logical`

Output Arguments

`Idx` — Indices of nearest points
cell array of numeric vectors

Indices of nearest points, returned as a cell array of numeric vectors.

`Idx` is an *my*-by-`1` cell array, where *my* is the number of rows in `Y`. The vector `Idx{j}` contains the indices of points (rows) in `X` whose distances to `Y(j,:)` are not greater than `r`. If `SortIndices` is `true`, then `rangesearch` arranges the indices in ascending order by distance.

`D` — Distances of nearest points to query points
cell array of numeric vectors

Distances of the nearest points to the query points, returned as a cell array of numeric vectors.

`D` is an *my*-by-`1` cell array, where *my* is the number of rows in `Y`. `D{j}` contains the distance values between `Y(j,:)` and the points (rows) in `X(Idx{j},:)`. If `SortIndices` is `true`, then `rangesearch` arranges the distances in ascending order.

Tips

- For a fixed positive real value `r`, `rangesearch` finds all the `X` points that are within a distance `r` of each `Y` point. To find the *k* points in `X` that are nearest to each `Y` point, for a fixed positive integer *k*, use `knnsearch`.
- `rangesearch` does not save a search object. To create a search object, use `createns`.

Algorithms

- For an overview of the *k*d-tree algorithm, see “k-Nearest Neighbor Search Using a Kd-Tree” on page 18-15.
- The exhaustive search algorithm finds the distance from each point in `X` to each point in `Y`.

Alternative Functionality

If you set the `rangesearch` function `'NSMethod'` name-value pair argument to the appropriate value (`'exhaustive'` for an exhaustive search algorithm or `'kdtree'` for a *K*d-tree algorithm), then the search results are equivalent to the results obtained by conducting a distance search using the `rangesearch` object function. Unlike the `rangesearch` function, the `rangesearch` object function requires an `ExhaustiveSearcher` or `KDTreeSearcher` model object.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- If `X` is a tall array, then `Y` cannot be a tall array. Similarly, if `Y` is a tall array, then `X` cannot be a tall array.

For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- For code generation, the default value of the `'NSMethod'` name-value pair argument is `'exhaustive'` when the number of columns in `X` is greater than 7.
- The value of the `'Distance'` name-value pair argument must be a compile-time constant and cannot be a custom distance function.
- The `'SortIndices'` name-value pair argument is not supported. The output arguments are always sorted.
- Names in name-value pair arguments must be compile-time constants. For example, to allow a user-defined exponent for the Minkowski distance in the generated code, include

`{coder.Constant('Distance'),coder.Constant('Minkowski'),coder.Constant('P'),0}` in the `-args` value of `codegen`.

- The sorted order of tied distances in the generated code can be different from the order in MATLAB due to numerical precision.
- When `rangesearch` uses the *k*d-tree search algorithm, and the code generation build type is a MEX function, `codegen` generates a MEX function using Intel Threading Building Blocks (TBB) for parallel computation. Otherwise, `codegen` generates code using `parfor`.
  - MEX function for the *k*d-tree search algorithm — `codegen` generates an optimized MEX function using Intel TBB for parallel computation on multicore platforms. You can use the MEX function to accelerate MATLAB algorithms. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

    If you generate the MEX function to test the generated code of the `parfor` version, you can disable the usage of Intel TBB. Set the `ExtrinsicCalls` property of the MEX configuration object to `false`. For details, see `coder.MexCodeConfig`.
  - MEX function for the exhaustive search algorithm and standalone C/C++ code for both algorithms — The generated code of `rangesearch` uses `parfor` to create loops that run in parallel on supported shared-memory multicore platforms in the generated code. If your compiler does not support the Open Multiprocessing (OpenMP) application interface or you disable OpenMP library, MATLAB Coder treats the `parfor`-loops as `for`-loops. To find supported compilers, see `https://www.mathworks.com/support/compilers/current_release/`. To disable OpenMP library, set the `EnableOpenMP` property of the configuration object to `false`. For details, see `coder.CodeConfig`.
- Starting in R2020a, `rangesearch` returns integer-type (`int32`) indices, rather than double-precision indices, in generated standalone C/C++ code. Therefore, the function allows for strict single-precision support when you use single-precision inputs. For MEX code generation, the function still returns double-precision indices to match the MATLAB behavior.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`ExhaustiveSearcher` | `KDTreeSearcher` | `createns` | `knnsearch` | `pdist2` | `rangesearch`

Topics

“k-Nearest Neighbor Search and Radius Search” on page 18-14
“Distance Metrics” on page 18-12

Introduced in R2011b

ranksum

Wilcoxon rank sum test

Syntax

```
p = ranksum(x,y)
[p,h] = ranksum(x,y)
[p,h,stats] = ranksum(x,y)
[___] = ranksum(x,y,Name,Value)
```

Description

`p = ranksum(x,y)` returns the *p*-value of a two-sided Wilcoxon rank sum test on page 33-5220. `ranksum` tests the null hypothesis that data in `x` and `y` are samples from continuous distributions with equal medians, against the alternative that they are not. The test assumes that the two samples are independent. `x` and `y` can have different lengths.

This test is equivalent to a Mann-Whitney U-test.

`[p,h] = ranksum(x,y)` also returns a logical value indicating the test decision. The result `h = 1` indicates a rejection of the null hypothesis, and `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level.

`[p,h,stats] = ranksum(x,y)` also returns the structure `stats` with information about the test statistic.

`[___] = ranksum(x,y,Name,Value)` returns any of the output arguments in the previous syntaxes, for a rank sum test with additional options specified by one or more `Name,Value` pair arguments.

Examples

Test for Equal Median of Two Populations

Test the hypothesis of equal medians for two independent unequal-sized samples.

Generate sample data.

```
rng('default') % for reproducibility
x = unifrnd(0,1,10,1);
y = unifrnd(0.25,1.25,15,1);
```

These samples come from populations with identical distributions except for a shift of 0.25 in the location.

Test the equality of medians of `x` and `y`.

```
p = ranksum(x,y)
```

```
p = 0.0375
```

The *p*-value of 0.0375 indicates that `ranksum` rejects the null hypothesis of equal medians at the default 5% significance level.

Statistics of the Test for Two Population Medians

Obtain the statistics of the test for the equality of two population medians.

Load the sample data.

```
load mileage
```

Test if the mileage per gallon is the same for the first and second type of cars.

```
[p,h,stats] = ranksum(mileage(:,1),mileage(:,2))
```

```
p = 0.0043
```

```
h = logical
   1
```

```
stats = struct with fields:
    ranksum: 21.5000
```

Both the *p*-value, 0.043, and `h = 1` indicate the rejection of the null hypothesis of equal medians at the default 5% significance level. Because the sample sizes are small (six each), `ranksum` calculates the *p*-value using the exact method. The structure `stats` includes only the value of the rank sum test statistic.

Increase in the Median

Test the hypothesis of an increase in the population median.

Load the sample data.

```
load('weather.mat');
```

The weather data shows the daily high temperatures taken in the same month in two consecutive years.

Perform a left-sided test to assess the increase in the median at the 1% significance level.

```
[p,h,stats] = ranksum(year1,year2,'alpha',0.01,...
'tail','left')
```

```
p = 0.1271
```

```
h = logical
   0
```

```
stats = struct with fields:
       zval: -1.1403
```

```
    ranksum: 837.5000
```

Both the *p*-value of 0.1271 and `h = 0` indicate that there is not enough evidence to reject the null hypothesis and conclude that there is a positive shift in the median of observed high temperatures in the same month from year 1 to year 2 at the 1% significance level. Notice that `ranksum` uses the approximate method to calculate the *p*-value due to the large sample sizes.

Use the exact method to calculate the *p*-value.

```
[p,h,stats] = ranksum(year1,year2,'alpha',0.01,...
'tail','left','method','exact')
```

```
p = 0.1273
```

```
h = logical
   0
```

```
stats = struct with fields:
    ranksum: 837.5000
```

The results of the approximate and exact methods are consistent with each other.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

y — Sample data

vector

Sample data, specified as a vector. The length of `y` does not have to be the same as the length of `x`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'alpha',0.01,'method','approximate','tail','right'` specifies a right-tailed rank sum test with 1% significance level, which returns the approximate p-value.

alpha — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level of the decision of a hypothesis test, specified as the comma-separated pair consisting of `'alpha'` and a scalar value in the range 0 to 1. The significance level of `h` is 100 * `alpha`%.

Example: `'alpha',0.01`

Data Types: `double` | `single`

`method` — Computation method of the *p*-value
`'exact'` | `'approximate'`

Computation method of the *p*-value, `p`, specified as the comma-separated pair consisting of `'method'` and one of the following:

`'exact'`	Exact computation of the *p*-value, `p`.
`'approximate'`	Normal approximation while computing the *p*-value, `p`.

When `'method'` is unspecified, the default is:

- `'exact'` if $\min(n_x,n_y) < 10$ and $n_x + n_y < 20$
- `'approximate'` otherwise

n_x and n_y are the sizes of the samples in `x` and `y`, respectively.

Example: `'method','exact'`

`tail` — Type of test
`'both'` (default) | `'right'` | `'left'`

Type of test, specified as the comma-separated pair consisting of `'tail'` and one of the following:

`'both'`	Two-sided hypothesis test, where the alternative hypothesis states that `x` and `y` have different medians. Default test type if `'tail'` is not specified.
`'right'`	Right-tailed hypothesis test, where the alternative hypothesis states that the median of `x` is greater than the median of `y`.
`'left'`	Left-tailed hypothesis test, where the alternative hypothesis states that the median of `x` is less than the median of `y`.

Example: `'tail','left'`

Output Arguments

`p` — *p*-value of the test
nonnegative scalar

p-value of the test, returned as a positive scalar from 0 to 1. `p` is the probability of observing a test statistic as or more extreme than the observed value under the null hypothesis. `ranksum` computes the two-sided *p*-value by doubling the most significant one-sided value.

`h` — Result of the hypothesis test
1 | 0

Result of the hypothesis test, returned as a logical value.

- If `h` = 1, this indicates rejection of the null hypothesis at the 100 * `alpha`% significance level.
- If `h` = 0, this indicates a failure to reject the null hypothesis at the 100 * `alpha`% significance level.

`stats` — Test statistics
structure

Test statistics, returned as a structure. The test statistics stored in `stats` are:

- `ranksum` : Value of the rank sum test statistic
- `zval`: Value of the z-statistic on page 33-5220 (computed when `'method'` is `'approximate'`)

More About

Wilcoxon Rank Sum Test

The Wilcoxon rank sum test is a nonparametric test for two populations when samples are independent. If `X` and `Y` are independent samples with different sample sizes, the test statistic which `ranksum` returns is the rank sum of the first sample.

The Wilcoxon rank sum test is equivalent to the Mann-Whitney U-test. The Mann-Whitney U-test is a nonparametric test for equality of population medians of two independent samples `X` and `Y`.

The Mann-Whitney U-test statistic, U, is the number of times a y precedes an x in an ordered arrangement of the elements in the two independent samples `X` and `Y`. It is related to the Wilcoxon rank sum statistic in the following way: If `X` is a sample of size n_X, then

$$U = W - \frac{n_X(n_X + 1)}{2}.$$

z-Statistic

For large samples, `ranksum` uses a z-statistic to compute the approximate p-value of the test.

If `X` and `Y` are two independent samples of size n_X and n_Y, where $n_X < n_Y$ the z-statistic is

$$z = \frac{W - E(W)}{\sqrt{V(W)}} = \frac{W - \left[\frac{n_X n_Y + n_X(n_X + 1)}{2}\right] - 0.5 * sign(W - E(W))}{\sqrt{\frac{n_X n_Y (n_X + n_Y + 1) - tiescor}{12}}},$$

with continuity correction and tie adjustment. Here *tiescor* is given by

$$tiescor = \frac{2 * tieadj}{(n_X + n_Y)(n_X + n_Y - 1)},$$

where `ranksum` uses `[ranks,tieadj] = tiedrank(x,y)` to obtain tie adjustments. The standard normal distribution gives the p-value for this z-statistic.

Algorithms

`ranksum` treats `NaN`s in `x` and `y` as missing values and ignores them.

For a two-sided test of medians with unequal sample sizes, the test statistic that `ranksum` returns is the rank sum of the first sample.

References

[1] Gibbons, J. D., and S. Chakraborti. *Nonparametric Statistical Inference*, 5th Ed., Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group, 2011.

[2] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

See Also

`kruskalwallis` | `signrank` | `signtest` | `ttest2`

Introduced before R2006a

ranova

Class: `RepeatedMeasuresModel`

Repeated measures analysis of variance

Syntax

```
ranovatbl = ranova(rm)
ranovatbl = ranova(rm,'WithinModel',WM)
[ranovatbl,A,C,D] = ranova( ___ )
```

Description

`ranovatbl = ranova(rm)` returns the results of repeated measures analysis of variance for a repeated measures model `rm` in table `ranovatbl`.

`ranovatbl = ranova(rm,'WithinModel',WM)` returns the results of repeated measures analysis of variance using the responses specified by the within-subject model `WM`.

`[ranovatbl,A,C,D] = ranova( ___ )` also returns arrays `A`, `C`, and `D` for the hypotheses tests of the form `A*B*C = D`, where `D` is zero.

Input Arguments

`rm` — Repeated measures model

`RepeatedMeasuresModel` object

Repeated measures model, returned as a `RepeatedMeasuresModel` object.

For properties and methods of this object, see `RepeatedMeasuresModel`.

`WM` — Model specifying responses

`'separatemeans'` (default) | *r*-by-*nc* contrast matrix | character vector or string scalar that defines a model specification

Model specifying the responses, specified as one of the following:

- `'separatemeans'` — Compute a separate mean for each group.
- `C` — *r*-by-*nc* contrast matrix specifying the *nc* contrasts among the *r* repeated measures. If *Y* represents a matrix of repeated measures, `ranova` tests the hypothesis that the means of *Y***C* are zero.
- A character vector or string scalar that defines a model specification in the within-subject factors. You can define the model based on the rules for the `terms` in the `modelspec` argument of `fitrm`. Also see “Model Specification for Repeated Measures Models” on page 9-54.

For example, if there are three within-subject factors `w1`, `w2`, and `w3`, then you can specify a model for the within-subject factors as follows.

Example: `'WithinModel','w1+w2+w2*w3'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`ranovatbl` — Results of repeated measures anova

table

Results of repeated measures anova, returned as a `table`.

`ranovatbl` includes a term representing all differences across the within-subjects factors. This term has either the name of the within-subjects factor if specified while fitting the model, or the name `Time` if the name of the within-subjects factor is not specified while fitting the model or there are more than one within-subjects factors. `ranovatbl` also includes all interactions between the terms in the within-subject model and all between-subject model terms. It contains the following columns.

Column Name	Definition
`SumSq`	Sum of squares.
`DF`	Degrees of freedom.
`MeanSq`	Mean squared error.
`F`	*F*-statistic.
`pValue`	*p*-value for the corresponding *F*-statistic. A small *p*-value indicates significant term effect.
`pValueGG`	*p*-value with Greenhouse-Geisser adjustment.
`pValueHF`	*p*-value with Huynh-Feldt adjustment.
`pValueLB`	*p*-value with Lower bound adjustment.

The last three *p*-values are the adjusted *p*-values for use when the compound symmetry assumption is not satisfied. For details, see “Compound Symmetry Assumption and Epsilon Corrections” on page 9-55. The `mauchy` method tests for sphericity (hence, compound symmetry) and `epsilon` method returns the epsilon adjustment values.

`A` — Specification based on between-subjects model

matrix | cell array

Specification based on the between-subjects model, returned as a matrix or a cell array. It permits the hypothesis on the elements within given columns of `B` (within time hypothesis). If `ranovatbl` contains multiple hypothesis tests, `A` might be a cell array.

Data Types: `single` | `double` | `cell`

`C` — Specification based on within-subjects model

matrix | cell array

Specification based on the within-subjects model, returned as a matrix or a cell array. It permits the hypotheses on the elements within given rows of `B` (between time hypotheses). If `ranovatbl` contains multiple hypothesis tests, `C` might be a cell array.

Data Types: `single` | `double` | `cell`

`D` — Hypothesis value

0

Hypothesis value, returned as 0.

Examples

Repeated Measures Analysis of Variance

Load the sample data.

```
load fisheriris
```

The column vector `species` consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = table([1 2 3 4]','VariableNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas);
```

Perform repeated measures analysis of variance.

```
ranovatbl = ranova(rm)
```

```
ranovatbl=3×8 table
                              SumSq     DF      MeanSq       F          pValue          pValueGG
                              ______    ___    ________    ______    ____________    ____________

    (Intercept):Measurements    1656.3     3      552.09    6873.3              0    9.4491e-279
    species:Measurements        282.47     6      47.078     586.1    1.4271e-206    4.9313e-150
    Error(Measurements)         35.423   441    0.080324
```

There are four measurements, three types of species, and 150 observations. So, degrees of freedom for measurements is (4–1) = 3, for species-measurements interaction it is (4–1)*(3–1) = 6, and for error it is (150–3)*(4–1) = 441. `ranova` computes the last three *p*-values using Greenhouse-Geisser, Huynh-Feldt, and Lower bound corrections, respectively. You can check the compound symmetry (sphericity) assumption using the `mauchly` method, and display the epsilon corrections using the `epsilon` method.

Longitudinal Data

Load the sample data.

```
load('longitudinalData.mat');
```

The matrix Y contains response data for 16 individuals. The response is the blood level of a drug measured at five time points (time = 0, 2, 4, 6, and 8). Each row of Y corresponds to an individual,

and each column corresponds to a time point. The first eight subjects are female, and the second eight subjects are male. This is simulated data.

Define a variable that stores gender information.

```
Gender = ['F' 'F' 'F' 'F' 'F' 'F' 'F' 'F' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M']';
```

Store the data in a proper table array format to do repeated measures analysis.

```
t = table(Gender,Y(:,1),Y(:,2),Y(:,3),Y(:,4),Y(:,5),...
'VariableNames',{'Gender','t0','t2','t4','t6','t8'});
```

Define the within-subjects variable.

```
Time = [0 2 4 6 8]';
```

Fit a repeated measures model, where the blood levels are the responses and gender is the predictor variable.

```
rm = fitrm(t,'t0-t8 ~ Gender','WithinDesign',Time);
```

Perform repeated measures analysis of variance.

```
ranovatbl = ranova(rm)
```

```
ranovatbl=3×8 table
                       SumSq     DF    MeanSq       F         pValue       pValueGG      pValueH
                       ______    __    ______    _______    __________    __________    _______

    (Intercept):Time    881.7     4    220.43     37.539    3.0348e-15    4.7325e-09    2.4439e-
    Gender:Time         17.65     4    4.4125    0.75146      0.56126        0.4877      0.501
    Error(Time)         328.83    56    5.872
```

There are 5 time points, 2 genders, and 16 observations. So, the degrees of freedom for time is (5–1) = 4, for gender-time interaction it is (5–1)*(2–1) = 4, and for error it is (16–2)*(5–1) = 56. The small *p*-value of 2.6198e–05 indicates that there is a significant effect of time on blood pressure. The *p* -value of 0.40063 indicates that there is no significant gender-time interaction.

Specify the Within-Subjects Model

Load the sample data.

```
load repeatedmeas
```

The table `between` includes the between-subject variables age, IQ, group, gender, and eight repeated measures `y1` through `y8` as responses. The table within includes the within-subject variables `w1` and `w2`. This is simulated data. Hypothetically, the response can be results of a memory test. The within-subject variable `w1` can be the type of exercise the subject does before the test and `w2` can be the different points in the day the subject takes the memory test. So, one subject does two different type of exercises A and B before taking the test and takes the test at four different times on different days. For each subject, the measurements are taken under these conditions:

Exercise to perform before the test: A B A B A B A B

Test time: 1 1 2 2 3 3 4 4

Fit a repeated measures model, where the repeated measures `y1` through `y8` are the responses, and age, IQ, group, gender, and the group-gender interaction are the predictor variables. Also specify the within-subject design matrix.

```
rm = fitrm(between,'y1-y8 ~ Group*Gender + Age + IQ','WithinDesign',within);
```

Perform repeated measures analysis of variance.

```
ranovatbl = ranova(rm)
```

```
ranovatbl=7×8 table
                        SumSq     DF     MeanSq       F        pValue      pValueGG    pValueHF
                        ______    ___    ______    _______    _________    ________    ________

    (Intercept):Time     6645.2     7    949.31     2.2689     0.031674    0.071235     0.05625
    Age:Time             5824.3     7    832.05     1.9887     0.059978     0.10651     0.09012
    IQ:Time              5188.3     7    741.18     1.7715     0.096749     0.14492      0.1289
    Group:Time            15800    14    1128.6     2.6975    0.0014425    0.011884    0.006434
    Gender:Time          4455.8     7    636.55     1.5214      0.16381     0.20533      0.1925
    Group:Gender:Time    4247.3    14    303.38    0.72511      0.74677       0.663      0.6918
    Error(Time)           64433   154    418.39
```

Specify the model for the within-subject factors. Also display the matrices used in the hypothesis test.

```
[ranovatbl,A,C,D] = ranova(rm,'WithinModel','w1+w2')
```

```
ranovatbl=21×8 table
                      SumSq     DF    MeanSq       F        pValue      pValueGG     pValueHF
                      ______    __    ______    _______    _________    _________    _________

    (Intercept)       3141.7     1    3141.7     2.5034      0.12787      0.12787      0.12787
    Age               537.48     1    537.48    0.42828      0.51962      0.51962      0.51962
    IQ                2975.9     1    2975.9     2.3712      0.13785      0.13785      0.13785
    Group              20836     2     10418     8.3012    0.0020601    0.0020601    0.0020601
    Gender            3036.3     1    3036.3     2.4194      0.13411      0.13411      0.13411
    Group:Gender       211.8     2     105.9    0.084385      0.91937      0.91937      0.91937
    Error              27609    22      1255          1          0.5          0.5          0.5
    (Intercept):w1    146.75     1    146.75    0.23326      0.63389      0.63389      0.63389
    Age:w1            942.02     1    942.02     1.4974      0.23402      0.23402      0.23402
    IQ:w1             11.563     1    11.563    0.01838      0.89339      0.89339      0.89339
    Group:w1          4481.9     2    2240.9      3.562     0.045697     0.045697     0.045697
    Gender:w1         270.65     1    270.65     0.4302      0.51869      0.51869      0.51869
    Group:Gender:w1   240.37     2    120.19    0.19104      0.82746      0.82746      0.82746
    Error(w1)          13841    22    629.12          1          0.5          0.5          0.5
    (Intercept):w2    3663.8     3    1221.3     3.8381     0.013513     0.020339     0.01575
    Age:w2            1199.9     3    399.95     1.2569       0.2964      0.29645      0.29662
      ⋮


A=6×1 cell array
    {[1 0 0 0 0 0 0 0]}
    {[0 1 0 0 0 0 0 0]}
    {[0 0 1 0 0 0 0 0]}
    {2x8 double       }
    {[0 0 0 0 0 1 0 0]}
```

```
    {2x8 double       }


C=1×3 cell array
    {8x1 double}    {8x1 double}    {8x3 double}


D = 0
```

Display the contents of `A`.

```
[A{1};A{2};A{3};A{4};A{5};A{6}]
```

```
ans = 8×8

     1     0     0     0     0     0     0     0
     0     1     0     0     0     0     0     0
     0     0     1     0     0     0     0     0
     0     0     0     1     0     0     0     0
     0     0     0     0     1     0     0     0
     0     0     0     0     0     1     0     0
     0     0     0     0     0     0     1     0
     0     0     0     0     0     0     0     1
```

Display the contents of `C`.

```
[C{1} C{2} C{3}]
```

```
ans = 8×5

     1     1     1     0     0
     1     1     0     1     0
     1     1     0     0     1
     1     1    -1    -1    -1
     1    -1     1     0     0
     1    -1     0     1     0
     1    -1     0     0     1
     1    -1    -1    -1    -1
```

Algorithms

`ranova` computes the regular *p*-value (in the `pValue` column of the `rmanova` table) using the *F*-statistic cumulative distribution function:

p-value = 1 – fcdf(F,v_1,v_2).

When the compound symmetry assumption is not satisfied, `ranova` uses a correction factor epsilon, ε, to compute the corrected *p*-values as follows:

p-value_corrected = 1 – fcdf(F,ε*v_1,ε*v_2).

The `mauchly` method tests for sphericity (hence, compound symmetry) and `epsilon` method returns the epsilon adjustment values.

See Also

`anova` | `epsilon` | `fitrm` | `manova` | `mauchly`

Topics

“Model Specification for Repeated Measures Models” on page 9-54
“Compound Symmetry Assumption and Epsilon Corrections” on page 9-55
“Mauchly’s Test of Sphericity” on page 9-57

raylcdf

Rayleigh cumulative distribution function

Syntax

```
p = raylcdf(x,b)
p = raylcdf(x,b,'upper')
```

Description

`p = raylcdf(x,b)` returns the Rayleigh cdf at each value in `x` using the corresponding scale parameter, `b`. `x` and `b` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input for `x` or `b` is expanded to a constant array with the same dimensions as the other input.

`p = raylcdf(x,b,'upper')` returns the complement of the Rayleigh cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The Rayleigh cdf is

$$y = F(x|b) = \int_0^x \frac{t}{b^2} e\left(\frac{-t^2}{2b^2}\right) dt$$

Examples

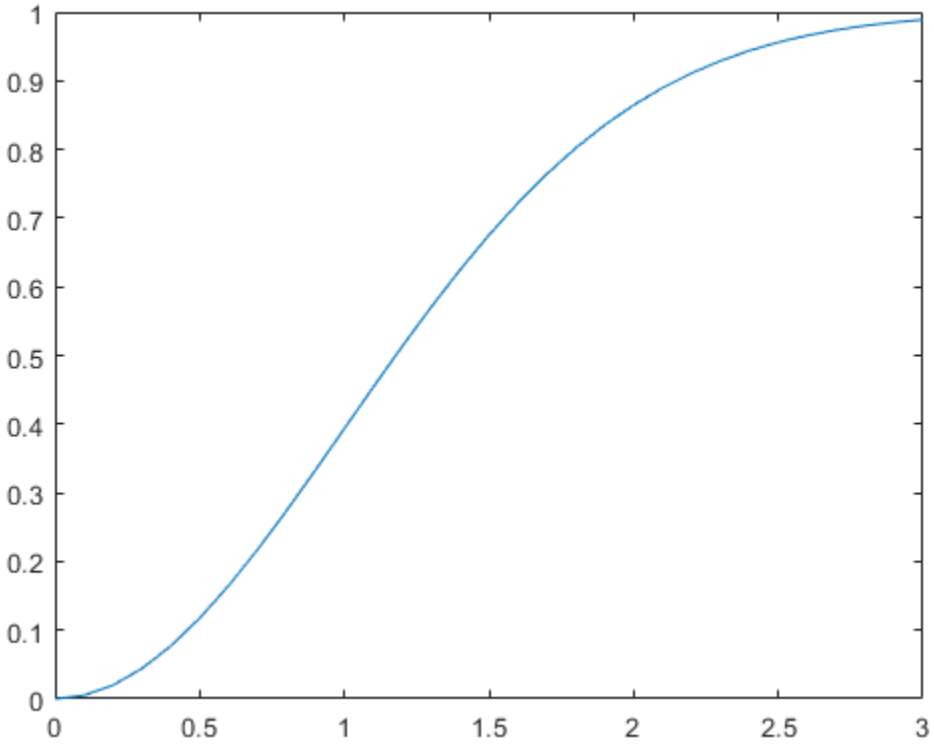
Compute and Plot Rayleigh Distribution cdf

Compute the cdf of a Rayleigh distribution with parameter `B = 1`.

```
x = 0:0.1:3;
p = raylcdf(x,1);
```

Plot the cdf.

```
figure;
plot(x,p)
```



References

[1] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. Hoboken, NJ: Wiley-Interscience, 2000. pp. 134–136.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `raylfit` | `raylinv` | `raylpdf` | `raylrnd` | `raylstat`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

raylfit

Rayleigh parameter estimates

Syntax

```
raylfit(data,alpha)
[phat,pci] = raylfit(data,alpha)
```

Description

`raylfit(data,alpha)` returns the maximum likelihood estimates of the parameter of the Rayleigh distribution given the data in the vector `data`.

`[phat,pci] = raylfit(data,alpha)` returns the maximum likelihood estimate and 100(1 - `alpha`)% confidence interval given the data. The default value of the optional parameter `alpha` is 0.05, corresponding to 95% confidence intervals.

See Also

`mle` | `raylcdf` | `raylinv` | `raylpdf` | `raylrnd` | `raylstat`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

raylinv

Rayleigh inverse cumulative distribution function

Syntax

```
X = raylinv(P,B)
```

Description

`X = raylinv(P,B)` returns the inverse of the Rayleigh cumulative distribution function using the corresponding scale parameter, `B` at the corresponding probabilities in `P`. `P` and `B` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input for `P` or `B` is expanded to a constant array with the same dimensions as the other input.

Examples

```
x = raylinv(0.9,1)
x =
  2.1460
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`raylcdf` | `raylpdf` | `raylrnd` | `raylstat`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

raylpdf

Rayleigh probability density function

Syntax

```
Y = raylpdf(X,B)
```

Description

`Y = raylpdf(X,B)` computes the Rayleigh pdf at each of the values in `X` using the corresponding scale parameter, `B`. `X` and `B` can be vectors, matrices, or multidimensional arrays that all have the same size, which is also the size of `Y`. A scalar input for `X` or `B` is expanded to a constant array with the same dimensions as the other input.

The Rayleigh pdf is

$$y = f(x \mid b) = \frac{x}{b^2} e\left(\frac{-x^2}{2b^2}\right)$$

Examples

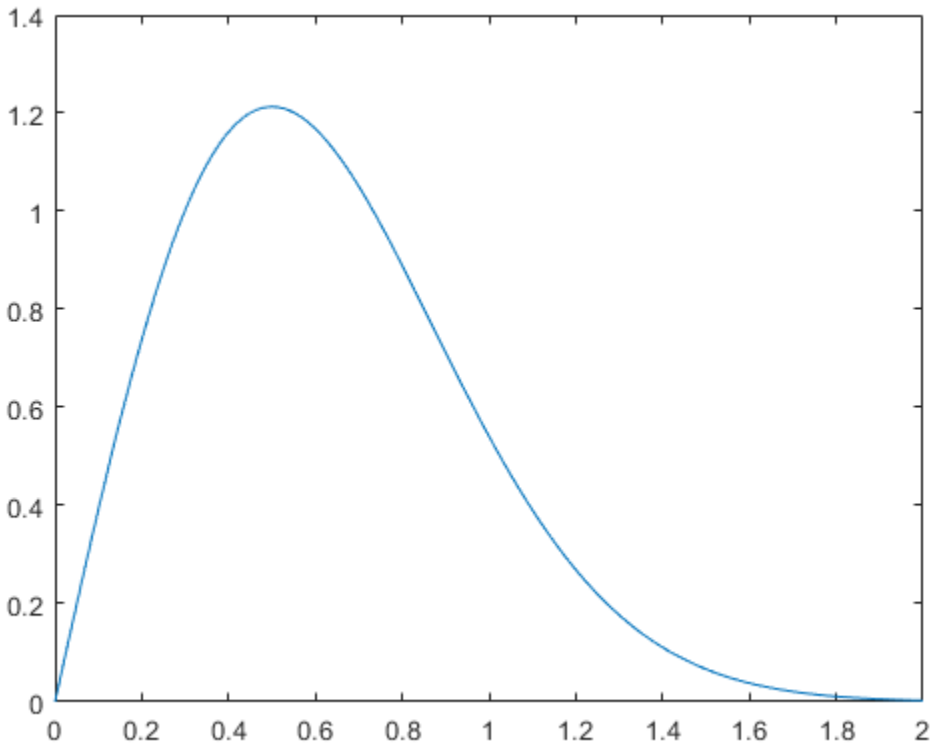
Compute and Plot Rayleigh Distribution pdf

Compute the pdf of a Rayleigh distribution with parameter `B = 0.5`.

```
x = [0:0.01:2];
p = raylpdf(x,0.5);
```

Plot the pdf.

```
figure;
plot(x,p)
```



Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`pdf` | `raylcdf` | `raylfit` | `raylinv` | `raylrnd` | `raylstat`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

raylrnd

Rayleigh random numbers

Syntax

```
R = raylrnd(B)
R = raylrnd(B,v)
R = raylrnd(B,m,n)
```

Description

`R = raylrnd(B)` returns a matrix of random numbers chosen from the Rayleigh distribution with scale parameter, `B`. `B` can be a vector, a matrix, or a multidimensional array. The size of `R` is the size of `B`.

`R = raylrnd(B,v)` returns a matrix of random numbers chosen from the Rayleigh distribution with parameter `B`, where `v` is a row vector. If `v` is a 1-by-2 vector, `R` is a matrix with `v(1)` rows and `v(2)` columns. If `v` is 1-by-n, `R` is an n-dimensional array.

`R = raylrnd(B,m,n)` returns a matrix of random numbers chosen from the Rayleigh distribution with parameter `B`, where scalars `m` and `n` are the row and column dimensions of `R`.

Examples

```
r = raylrnd(1:5)
r =
  1.7986  0.8795  3.3473  8.9159  3.5182
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`random` | `raylcdf` | `raylfit` | `raylinv` | `raylpdf` | `raylstat`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

raylstat

Rayleigh mean and variance

Syntax

```
[M,V] = raylstat(B)
```

Description

`[M,V] = raylstat(B)` returns the mean of and variance for the Rayleigh distribution with scale parameter `B`.

The mean of the Rayleigh distribution with parameter b is $b\sqrt{\pi/2}$and the variance is

$$\frac{4-\pi}{2}b^2$$

Examples

```
[mn,v] = raylstat(1)
mn =
  1.2533
v =
  0.4292
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`raylcdf` | `raylfit` | `raylinv` | `raylpdf` | `raylrnd`

Topics

“Rayleigh Distribution” on page B-137

Introduced before R2006a

rcoplot

Residual case order plot

Syntax

```
rcoplot(r,rint)
```

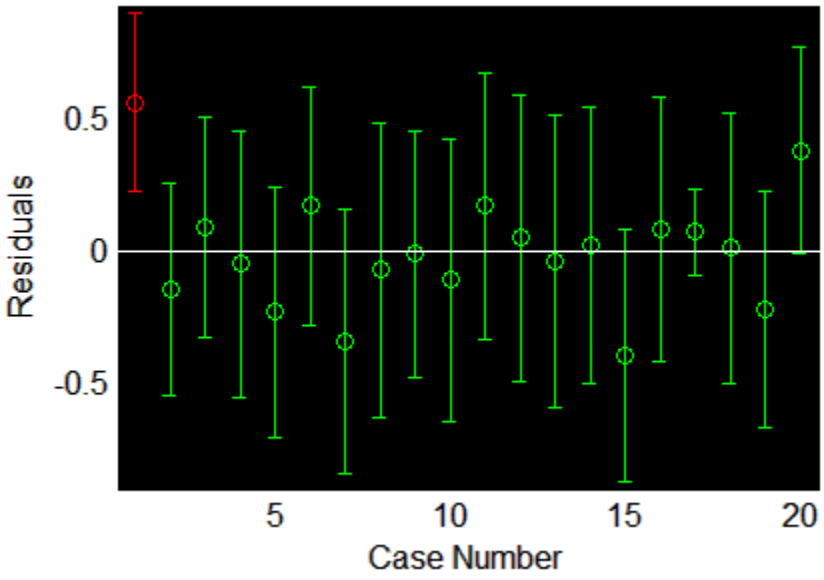
Description

`rcoplot(r,rint)` displays an error bar plot of the confidence intervals on the residuals from a regression. The residuals appear in the plot in case order. Inputs `r` and `rint` are outputs from the `regress` function.

Examples

The following plots residuals and prediction intervals from a regression of a linearly additive model to the data in `moore.mat`:

```
load moore
X = [ones(size(moore,1),1) moore(:,1:5)];
y = moore(:,6);
alpha = 0.05;
[betahat,Ibeta,res,Ires,stats] = regress(y,X,alpha);
rcoplot(res,Ires)
```



The interval around the first residual, shown in red, does not contain zero. This indicates that the residual is larger than expected in 95% of new observations, and suggests the data point is an outlier.

See Also

`regress`

Introduced before R2006a

ReconstructionICA

Feature extraction by reconstruction ICA

Description

`ReconstructionICA` applies reconstruction independent component analysis (RICA) to learn a transformation that maps input predictors to new predictors.

Creation

Create a `ReconstructionICA` object by using the `rica` function.

Properties

`FitInfo` — Fitting history
structure

This property is read-only.

Fitting history, returned as a structure with two fields:

- `Iteration` — Iteration numbers from 0 through the final iteration.
- `Objective` — Objective function value at each corresponding iteration. Iteration 0 corresponds to the initial values, before any fitting.

Data Types: `struct`

`InitialTransformWeights` — Initial feature transformation weights
`p`-by-`q` matrix

This property is read-only.

Initial feature transformation weights, returned as a `p`-by-`q` matrix, where `p` is the number of predictors passed in `X` and `q` is the number of features that you want. These weights are the initial weights passed to the creation function. The data type is single when the training data `X` is single.

Data Types: `single` | `double`

`ModelParameters` — Parameters for training model
structure

This property is read-only.

Parameters for training the model, returned as a structure. The structure contains a subset of the fields that correspond to the `rica` name-value pairs that were in effect during model creation:

- `IterationLimit`
- `VerbosityLevel`

- `Lambda`
- `Standardize`
- `ContrastFcn`
- `GradientTolerance`
- `StepTolerance`

For details, see the `rica` `Name,Value` on page 33-5704 pairs.

Data Types: `struct`

`Mu` — Predictor means when standardizing
`p`-by-1 vector

This property is read-only.

Predictor means when standardizing, returned as a `p`-by-`1` vector. This property is nonempty when the `Standardize` name-value pair is `true` at model creation. The value is the vector of predictor means in the training data. The data type is single when the training data `X` is single.

Data Types: `single` | `double`

`NonGaussianityIndicator` — Non-Gaussianity of sources
length-`q` vector of ±1

This property is read-only.

Non-Gaussianity of sources, returned as a length-`q` vector of ±1.

- `NonGaussianityIndicator(k) = 1` means `rica` models the `k`th source as sub-Gaussian.
- `NonGaussianityIndicator(k) = -1` means `rica` models the `k`th source as super-Gaussian, with a sharp peak at 0.

Data Types: `double`

`NumLearnedFeatures` — Number of output features
positive integer

This property is read-only.

Number of output features, returned as a positive integer. This value is the `q` argument passed to the creation function, which is the requested number of features to learn.

Data Types: `double`

`NumPredictors` — Number of input predictors
positive integer

This property is read-only.

Number of input predictors, returned as a positive integer. This value is the number of predictors passed in `X` to the creation function.

Data Types: `double`

`Sigma` — Predictor standard deviations when standardizing
`p`-by-1 vector

This property is read-only.

Predictor standard deviations when standardizing, returned as a p-by-1 vector. This property is nonempty when the `Standardize` name-value pair is `true` at model creation. The value is the vector of predictor standard deviations in the training data. The data type is single when the training data X is single.

Data Types: `single` | `double`

`TransformWeights` — Feature transformation weights
p-by-q matrix

This property is read-only.

Feature transformation weights, returned as a p-by-q matrix, where p is the number of predictors passed in X and q is the number of features that you want. The data type is single when the training data X is single.

Data Types: `single` | `double`

Object Functions

transform Transform predictors into extracted features

Examples

Create Reconstruction ICA Object

Create a `ReconstructionICA` object by using the `rica` function.

Load the `SampleImagePatches` image patches.

```
data = load('SampleImagePatches');
size(data.X)
```

```
ans = 1×2

        5000         363
```

There are 5,000 image patches, each containing 363 features.

Extract 100 features from the data.

```
rng default % For reproducibility
q = 100;
Mdl = rica(data.X,q,'IterationLimit',100)
```

```
Warning: Solver LBFGS was not able to converge to a solution.

Mdl =
  ReconstructionICA
           ModelParameters: [1x1 struct]
             NumPredictors: 363
         NumLearnedFeatures: 100
                        Mu: []
```

```
                    Sigma: []
                  FitInfo: [1x1 struct]
          TransformWeights: [363x100 double]
    InitialTransformWeights: []
    NonGaussianityIndicator: [100x1 double]


  Properties, Methods
```

`rica` issues a warning because it stopped due to reaching the iteration limit, instead of reaching a step-size limit or a gradient-size limit. You can still use the learned features in the returned object by calling the `transform` function.

See Also

`SparseFiltering` | `rica` | `sparsefilt` | `transform`

Topics

“Feature Extraction Workflow” on page 15-135
“Extract Mixed Signals” on page 15-164
“Feature Extraction” on page 15-130

Introduced in R2017a

refcurve

Add reference curve to plot

Syntax

```
refcurve(p)
refcurve
refcurve(ax,p)
hcurve = refcurve(...)
```

Description

`refcurve(p)` adds a polynomial reference curve with coefficients `p` to the current axes. If `p` is a vector with `n+1` elements, the curve is:

```
y = p(1)*x^n + p(2)*x^(n-1) + ... + p(n)*x + p(n+1)
```

`refcurve` with no input arguments adds a line along the *x* axis.

`refcurve(ax,p)` uses the plot axes specified in `ax`, an `Axes` object. For more information, see `axes`.

`hcurve = refcurve(...)` returns the handle `hcurve` to the curve using any of the input argument combinations in the previous syntaxes.

Examples

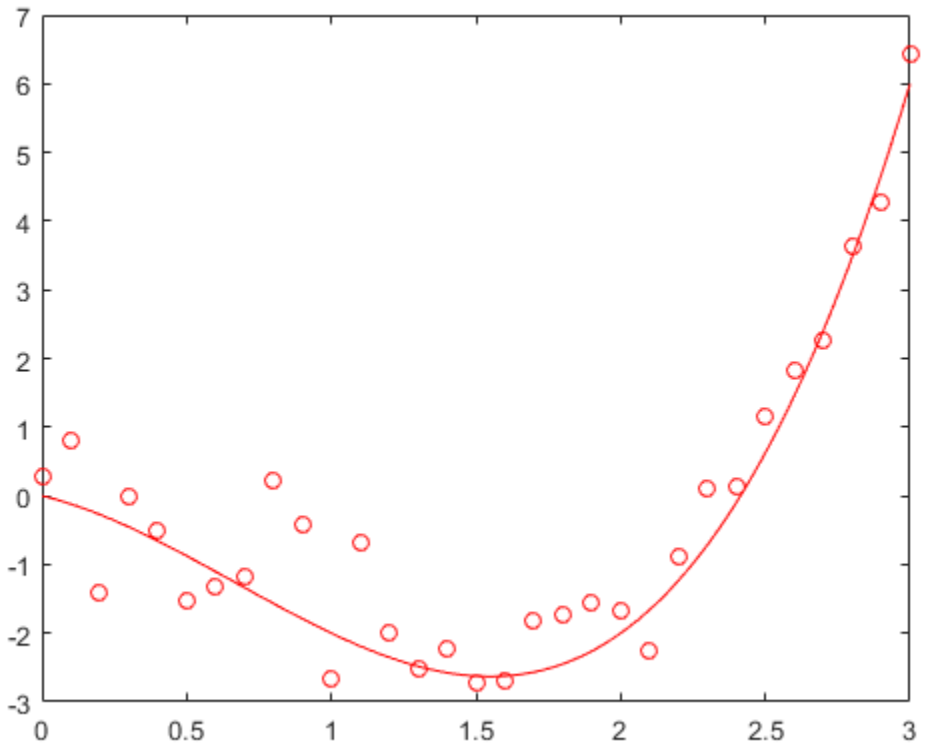
Add Population and Fitted Mean Functions

Generate data with a polynomial trend.

```
p = [1 -2 -1 0];
t = 0:0.1:3;
rng default  % For reproducibility
y = polyval(p,t) + 0.5*randn(size(t));
```

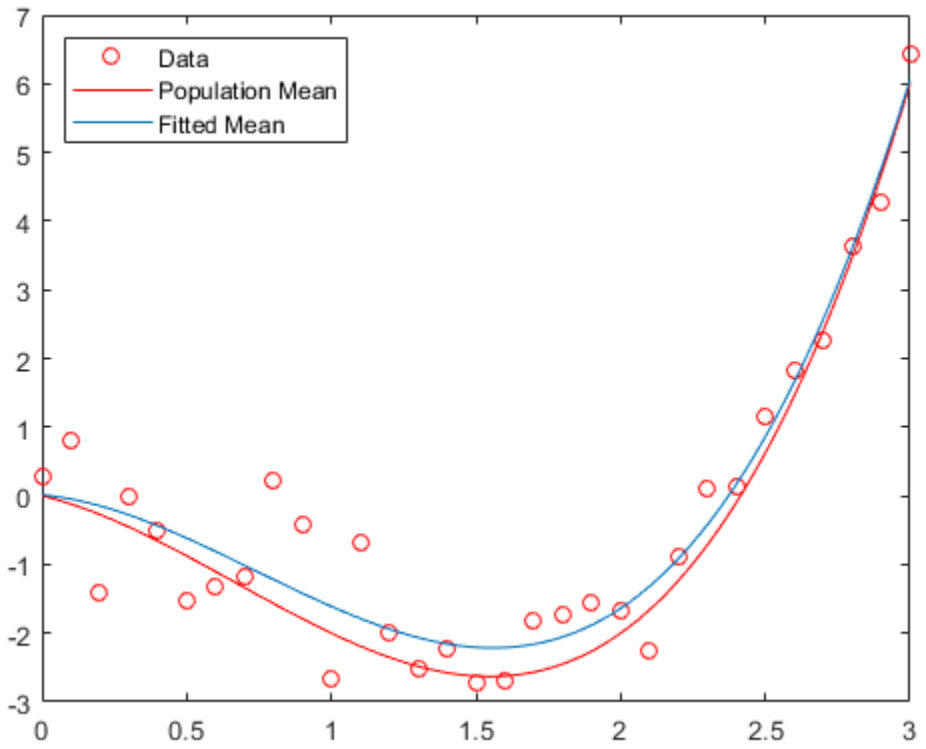
Plot data and add the population mean function using `refcurve` .

```
plot(t,y,'ro')
h = refcurve(p);
h.Color = 'r';
```



Also add the fitted mean function.

```
q = polyfit(t,y,3);
refcurve(q)
legend('Data','Population Mean','Fitted Mean',...
       'Location','NW')
```



Plot Trajectories of a Batted Baseball Using `refcurve`

Introduce the relevant physical constants.

```
M = 0.145;      % Mass (kg)
R = 0.0366;     % Radius (m)
A = pi*R^2;     % Area (m^2)
rho = 1.2;      % Density of air (kg/m^3)
C = 0.5;        % Drag coefficient
D = rho*C*A/2;  % Drag proportional to the square of the speed
g = 9.8;        % Acceleration due to gravity (m/s^2)
```

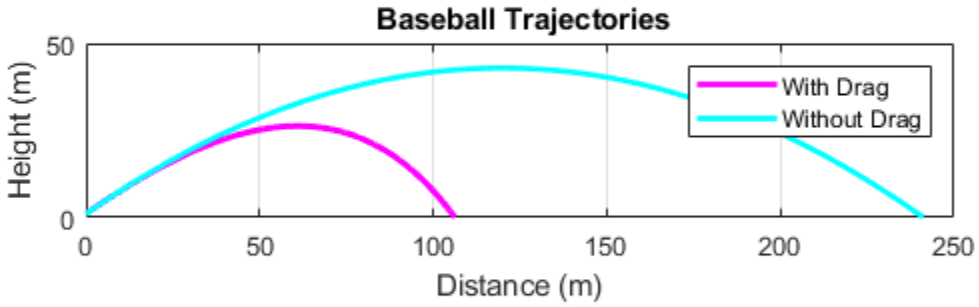
Simulate the trajectory with drag proportional to the square of the speed, assuming constant acceleration in each time interval.

```
dt = 1e-2;      % Simulation time interval (s)
r0 = [0 1];     % Initial position (m)
s0 = 50;        % Initial speed (m/s)
alpha0 = 35;    % Initial angle (deg)
v0 = s0*[cosd(alpha0) sind(alpha0)]; % Initial velocity (m/s)

r = r0;
v = v0;
trajectory = r0;
while r(2) > 0
```

```
    a = [0 -g] - (D/M)*norm(v)*v;
    v = v + a*dt;
    r = r + v*dt + (1/2)*a*(dt^2);
    trajectory = [trajectory;r];
end
```

Plot `trajectory` and use `refcurve` to add the drag-free parabolic trajectory (found analytically) to the plot of trajectory.

```
figure
plot(trajectory(:,1),trajectory(:,2),'m','LineWidth',2)
xlim([0,250])
h = refcurve([-g/(2*v0(1)^2),...
    (g*r0(1)/v0(1)^2) + (v0(2)/v0(1)),...
    (-g*r0(1)^2/(2*v0(1)^2)) - (v0(2)*r0(1)/v0(1)) + r0(2)]);
h.Color = 'c';
h.LineWidth = 2;
axis equal
ylim([0,50])
grid on
xlabel('Distance (m)')
ylabel('Height (m)')
title('{\bf Baseball Trajectories}')
legend('With Drag','Without Drag')
```



See Also

`gline` | `lsline` | `polyfit` | `refline`

Introduced before R2006a

refit

Class: `GeneralizedLinearMixedModel`

Refit generalized linear mixed-effects model

Syntax

```
glmenew = refit(glme,ynew)
```

Description

`glmenew = refit(glme,ynew)` returns a refitted generalized linear mixed-effects model, `glmenew`, based on the input model `glme`, using a new response vector, `ynew`.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

`ynew` — New response vector

n-by-1 vector of scalar values

New response vector, specified as an *n*-by-1 vector of scalar values, where *n* is the number of observations used to fit `glme`.

For an observation i with prior weights w_i^p and binomial size n_i (when applicable), the response values y_i contained in `ynew` can have the following values.

Distribution	Permitted Values	Notes
`Binomial`	$\left\{0, \frac{1}{w_i^p n_i}, \frac{2}{w_i^p n_i}, \ldots, 1\right\}$	w_i^p and n_i are integer values > 0
`Poisson`	$\left\{0, \frac{1}{w_i^p}, \frac{2}{w_i^p}, \cdots, 1\right\}$	w_i^p is an integer value > 0
`Gamma`	$(0,\infty)$	$w_i^p \geq 0$
`InverseGaussian`	$(0,\infty)$	$w_i^p \geq 0$
`Normal`	$(-\infty,\infty)$	$w_i^p \geq 0$

You can access the prior weights property w_i^p using dot notation.

```
glme.ObservationInfo.Weights
```

Data Types: `single` | `double`

Output Arguments

`glmenew` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, returned as a `GeneralizedLinearMixedModel` object. `glmenew` is an updated version of the generalized linear mixed-effects model `glme`, refit to the values in the response vector `ynew`.

For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Examples

Refit Model to New Response Vector

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- $defects_{ij}$ is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- $newprocess_{ij}$, $time\_dev_{ij}$, and $temp\_dev_{ij}$ are the measurements for each variable that correspond to factory i during batch j. For example, $newprocess_{ij}$ indicates whether the batch produced by factory i during batch j used the new process.
- $supplier\_C_{ij}$ and $supplier\_B_{ij}$ are dummy variables that use effects (sum-to-zero) coding to indicate whether company C or B, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)','Dis
```

Use `random` to simulate a new response vector from the fitted model.

```
rng(0,'twister');  % For reproducibility
ynew = random(glme);
```

Refit the model using the new response vector.

```
glme = refit(glme,ynew)
```

```
glme =
Generalized linear mixed-effects model fit by ML

Model information:
    Number of observations             100
    Fixed effects coefficients           6
    Random effects coefficients         20
    Covariance parameters                1
    Distribution                   Poisson
    Link                           Log
    FitMethod                      Laplace

Formula:
    defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1 | factory)

Model fit statistics:
    AIC       BIC       LogLikelihood    Deviance
    469.24    487.48    -227.62          455.24

Fixed effects coefficients (95% CIs):
    Name                   Estimate    SE          tStat       DF    pValue
    {'(Intercept)'}          1.5738     0.18674      8.4276    94    4.0158e-13
    {'newprocess' }         -0.21089     0.2306    -0.91455    94       0.36277
    {'time_dev'   }         -0.13769    0.77477    -0.17772    94       0.85933
    {'temp_dev'   }          0.24339    0.84657      0.2875    94       0.77436
    {'supplier_C' }         -0.12102    0.07323     -1.6526    94       0.10175
    {'supplier_B' }         0.098254    0.066943      1.4677    94       0.14551


    Lower        Upper
```

```
        1.203      1.9445
     -0.66875     0.24696
       -1.676      1.4006
      -1.4375      1.9243
     -0.26642    0.024381
    -0.034662     0.23117

Random effects covariance parameters:
Group: factory (20 Levels)
    Name1                  Name2                  Type           Estimate
    {'(Intercept)'}        {'(Intercept)'}        {'std'}        0.46587

Group: Error
    Name                       Estimate
    {'sqrt(Dispersion)'}       1
```

Tips

- You can use `refit` and `random` to conduct a simulated likelihood ratio test or parametric bootstrap.

See Also

`GeneralizedLinearMixedModel` | `designMatrix` | `fitted` | `residuals`

refit

Class: `FeatureSelectionNCAClassification`

Refit neighborhood component analysis (NCA) model for classification

Syntax

```
mdlrefit = refit(mdl,Name,Value)
```

Description

`mdlrefit = refit(mdl,Name,Value)` refits the model `mdl`, with modified parameters specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl` — Neighborhood component analysis model for classification

`FeatureSelectionNCAClassification` object

Neighborhood component analysis model or classification, specified as a `FeatureSelectionNCAClassification` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Fitting Options

`FitMethod` — Method for fitting the model

`mdl.FitMethod` (default) | `'exact'` | `'none'` | `'average'`

Method for fitting the model, specified as the comma-separated pair consisting of `'FitMethod'` and one of the following.

- `'exact'` — Performs fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to `fscnca`.
- `'average'` — The function divides the data into partitions (subsets), fits each partition using the `exact` method, and returns the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

Example: `'FitMethod','none'`

`Lambda` — Regularization parameter

`mdl.Lambda` (default) | non-negative scalar value

Regularization parameter, specified as the comma-separated pair consisting of `'Lambda'` and a non-negative scalar value.

For *n* observations, the best `Lambda` value that minimizes the generalization error of the NCA model is expected to be a multiple of 1/*n*

Example: `'Lambda',0.01`

Data Types: `double` | `single`

`Solver` — Solver type
`mdl.Solver` (default) | `'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Solver type for estimating feature weights, specified as the comma-separated pair consisting of `'Solver'` and one of the following.

- `'lbfgs'` — Limited memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm (LBFGS algorithm)
- `'sgd'` — Stochastic gradient descent
- `'minibatch-lbfgs'` — Stochastic gradient descent with LBFGS algorithm applied to mini-batches

Example: `'solver','minibatch-lbfgs'`

`InitialFeatureWeights` — Initial feature weights
`mdl.InitialFeatureWeights` (default) | *p*-by-1 vector of real positive scalar values

Initial feature weights, specified as the comma-separated pair consisting of `'InitialFeatureWeights'` and a *p*-by-1 vector of real positive scalar values.

Data Types: `double` | `single`

`Verbose` — Indicator for verbosity level
`mdl.Verbose` (default) | 0 | 1 | >1

Indicator for verbosity level for the convergence summary display, specified as the comma-separated pair consisting of `'Verbose'` and one of the following.

- 0 — No convergence summary
- 1 — Convergence summary including iteration number, norm of the gradient, and objective function value.
- >1 — More convergence information depending on the fitting algorithm

  When using solver `'minibatch-lbfgs'` and verbosity level >1, the convergence information includes iteration log from intermediate minibatch LBFGS fits.

Example: `'Verbose',2`

Data Types: `double` | `single`

LBFGS or Mini-Batch LBFGS Options

`GradientTolerance` — Relative convergence tolerance
`mdl.GradientTolerance` (default) | positive real scalar value

Relative convergence tolerance on the gradient norm for solver `lbfgs`, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive real scalar value.

Example: `'GradientTolerance',0.00001`

Data Types: `double` | `single`

SGD or Mini-Batch LBFGS Options

`InitialLearningRate` — Initial learning rate for solver `sgd`

`mdl.InitialLearningRate` (default) | positive real scalar value

Initial learning rate for solver `sgd`, specified as the comma-separated pair consisting of `'InitialLearningRate'` and a positive scalar value.

When using solver type `'sgd'`, the learning rate decays over iterations starting with the value specified for `'InitialLearningRate'`.

Example: `'InitialLearningRate',0.8`

Data Types: `double` | `single`

`PassLimit` — Maximum number of passes for solver `'sgd'`

`mdl.PassLimit` (default) | positive integer value

Maximum number of passes for solver `'sgd'` (stochastic gradient descent), specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer. Every pass processes `size(mdl.X,1)` observations.

Example: `'PassLimit',10`

Data Types: `double` | `single`

SGD or LBFGS or Mini-Batch LBFGS Options

`IterationLimit` — Maximum number of iterations

`mdl.IterationLimit` (default) | positive integer value

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

Example: `'IterationLimit',250`

Data Types: `double` | `single`

Output Arguments

`mdlrefit` — Neighborhood component analysis model for classification

`FeatureSelectionNCAClassification` object

Neighborhood component analysis model for classification, returned as a `FeatureSelectionNCAClassification` object. You can either save the results as a new model or update the existing model as `mdl = refit(mdl,Name,Value)`.

Examples

Refit NCA Model for Classification with Modified Settings

Generate checkerboard data using the `generateCheckerBoardData.m` function.

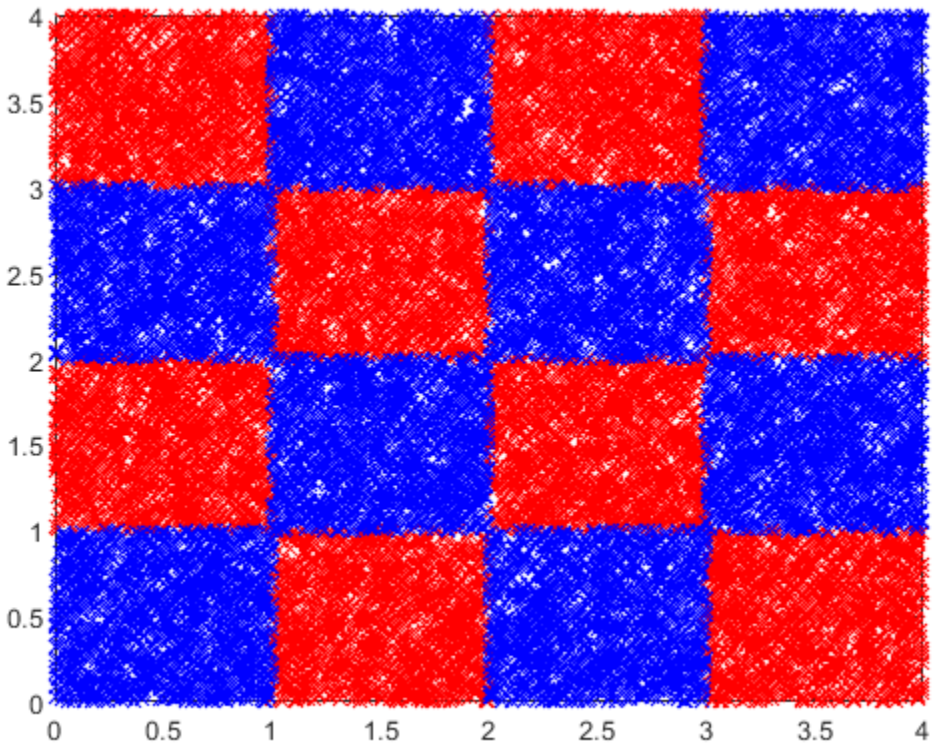
```
rng(2016,'twister'); % For reproducibility
pps = 1375;
```

```
[X,y] = generateCheckerBoardData(pps);
X = X + 2;
```

Plot the data.

```
figure
plot(X(y==1,1),X(y==1,2),'rx')
hold on
plot(X(y==-1,1),X(y==-1,2),'bx')

[n,p] = size(X)
```

```
n =

       22000


p =

     2
```



Add irrelevant predictors to the data.

```
Q = 98;
Xrnd = unifrnd(0,4,n,Q);
Xobs = [X,Xrnd];
```

This piece of code creates 98 additional predictors, all uniformly distributed between 0 and 4.

Partition the data into training and test sets. To create stratified partitions, so that each partition has similar proportion of classes, use `y` instead of `length(y)` as the partitioning criteria.

```
cvp = cvpartition(y,'holdout',2000);
```

`cvpartition` randomly chooses 2000 of the observations to add to the test set and the rest of the data to add to the training set. Create the training and validation sets using the assignments stored in the `cvpartition` object `cvp` .

```
Xtrain = Xobs(cvp.training(1),:);
ytrain = y(cvp.training(1),:);

Xval = Xobs(cvp.test(1),:);
yval = y(cvp.test(1),:);
```

Compute the misclassification error without feature selection.

```
nca = fscnca(Xtrain,ytrain,'FitMethod','none','Standardize',true, ...
    'Solver','lbfgs');
loss_nofs = loss(nca,Xval,yval)
```

```
loss_nofs =

    0.5165
```

`'FitMethod','none'` option uses the default weights (all 1s), which means all features are equally important.

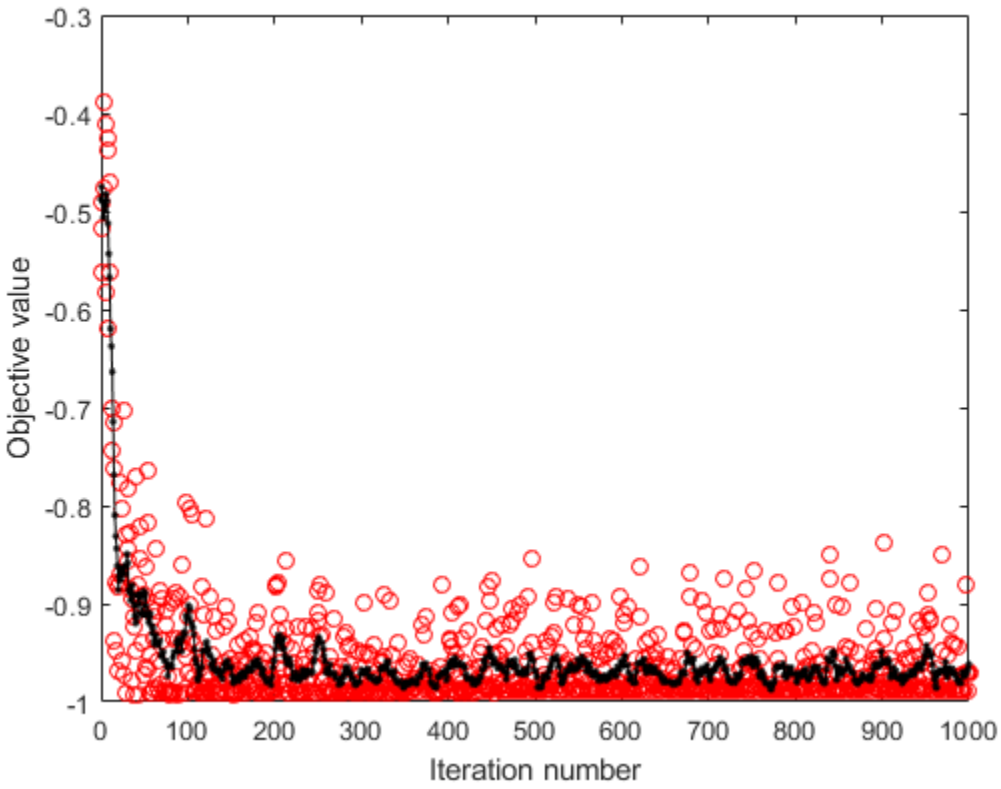
This time, perform feature selection using neighborhood component analysis for classification, with $\lambda = 1/n$.

```
w0 = rand(100,1);
n = length(ytrain)
lambda = 1/n;
nca = refit(nca,'InitialFeatureWeights',w0,'FitMethod','exact', ...
       'Lambda',lambda,'solver','sgd');
```

```
n =

       20000
```

Plot the objective function value versus the iteration number.

```
figure()
plot(nca.FitInfo.Iteration,nca.FitInfo.Objective,'ro')
hold on
plot(nca.FitInfo.Iteration,movmean(nca.FitInfo.Objective,10),'k.-')
xlabel('Iteration number')
ylabel('Objective value')
```



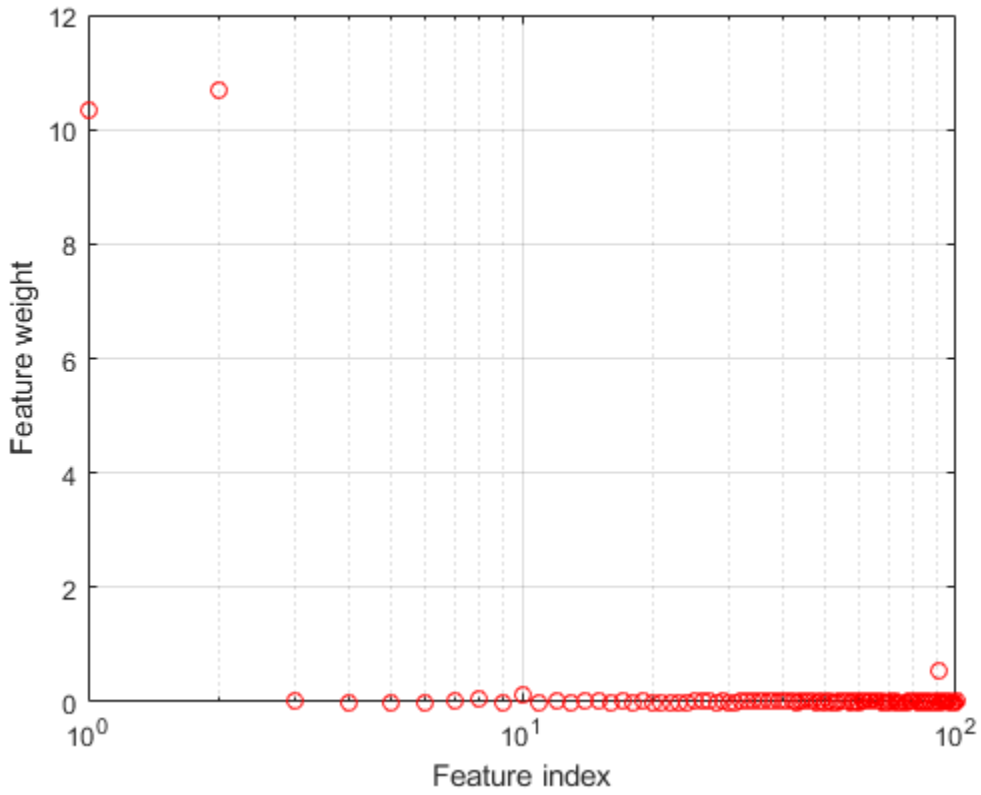
Compute the misclassification error with feature selection.

```
loss_withfs = loss(nca,Xval,yval)
```

```
loss_withfs =

    0.0115
```

Plot the selected features.

```
figure
semilogx(nca.FeatureWeights,'ro')
xlabel('Feature index')
ylabel('Feature weight')
grid on
```



Select features using the feature weights and a relative threshold.

```
tol = 0.15;
selidx = find(nca.FeatureWeights > tol*max(1,max(nca.FeatureWeights)))
```

```
selidx =

     1
     2
```

Feature selection improves the results and `fscnca` detects the correct two features as relevant.

See Also

`FeatureSelectionNCAClassification` | `fscnca` | `loss` | `predict`

Introduced in R2016b

refit

Class: `FeatureSelectionNCARegression`

Refit neighborhood component analysis (NCA) model for regression

Syntax

```
mdlrefit = refit(mdl,Name,Value)
```

Description

`mdlrefit = refit(mdl,Name,Value)` refits the model `mdl`, with modified parameters specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl` — Neighborhood component analysis model for regression

`FeatureSelectionNCARegression` object

Neighborhood component analysis model or classification, specified as a `FeatureSelectionNCARegression` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Fitting Options

`FitMethod` — Method for fitting the model

`mdl.FitMethod` (default) | `'exact'` | `'none'` | `'average'`

Method for fitting the model, specified as the comma-separated pair consisting of `'FitMethod'` and one of the following.

- `'exact'` — Performs fitting using all of the data.
- `'none'` — No fitting. Use this option to evaluate the generalization error of the NCA model using the initial feature weights supplied in the call to `fsrnca`.
- `'average'` — The function divides the data into partitions (subsets), fits each partition using the `exact` method, and returns the average of the feature weights. You can specify the number of partitions using the `NumPartitions` name-value pair argument.

Example: `'FitMethod','none'`

`Lambda` — Regularization parameter

`mdl.Lambda` (default) | non-negative scalar value

Regularization parameter, specified as the comma-separated pair consisting of `'Lambda'` and a non-negative scalar value.

For *n* observations, the best `Lambda` value that minimizes the generalization error of the NCA model is expected to be a multiple of 1/*n*

Example: `'Lambda',0.01`

Data Types: `double` | `single`

`Solver` — Solver type
`mdl.Solver` (default) | `'lbfgs'` | `'sgd'` | `'minibatch-lbfgs'`

Solver type for estimating feature weights, specified as the comma-separated pair consisting of `'Solver'` and one of the following.

- `'lbfgs'` — Limited memory BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm (LBFGS algorithm)
- `'sgd'` — Stochastic gradient descent
- `'minibatch-lbfgs'` — Stochastic gradient descent with LBFGS algorithm applied to mini-batches

Example: `'solver','minibatch-lbfgs'`

`InitialFeatureWeights` — Initial feature weights
`mdl.InitialFeatureWeights` (default) | *p*-by-1 vector of real positive scalar values

Initial feature weights, specified as the comma-separated pair consisting of `'InitialFeatureWeights'` and a *p*-by-1 vector of real positive scalar values.

Data Types: `double` | `single`

`Verbose` — Indicator for verbosity level
`mdl.Verbose` (default) | 0 | 1 | >1

Indicator for verbosity level for the convergence summary display, specified as the comma-separated pair consisting of `'Verbose'` and one of the following.

- 0 — No convergence summary
- 1 — Convergence summary including iteration number, norm of the gradient, and objective function value.
- >1 — More convergence information depending on the fitting algorithm

  When using solver `'minibatch-lbfgs'` and verbosity level >1, the convergence information includes iteration log from intermediate minibatch LBFGS fits.

Example: `'Verbose',2`

Data Types: `double` | `single`

LBFGS or Mini-Batch LBFGS Options

`GradientTolerance` — Relative convergence tolerance
`mdl.GradientTolerance` (default) | positive real scalar value

Relative convergence tolerance on the gradient norm for solver `lbfgs`, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive real scalar value.

Example: `'GradientTolerance',0.00001`

Data Types: `double` | `single`

SGD or Mini-Batch LBFGS Options

`InitialLearningRate` — Initial learning rate for solver `sgd`
`mdl.InitialLearningRate` (default) | positive real scalar value

Initial learning rate for solver `sgd`, specified as the comma-separated pair consisting of `'InitialLearningRate'` and a positive scalar value.

When using solver type `'sgd'`, the learning rate decays over iterations starting with the value specified for `'InitialLearningRate'`.

Example: `'InitialLearningRate',0.8`

Data Types: `double` | `single`

`PassLimit` — Maximum number of passes for solver `'sgd'`
`mdl.PassLimit` (default) | positive integer value

Maximum number of passes for solver `'sgd'` (stochastic gradient descent), specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer. Every pass processes `size(mdl.X,1)` observations.

Example: `'PassLimit',10`

Data Types: `double` | `single`

SGD or LBFGS or Mini-Batch LBFGS Options

`IterationLimit` — Maximum number of iterations
`mdl.IterationLimit` (default) | positive integer value

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

Example: `'IterationLimit',250`

Data Types: `double` | `single`

Output Arguments

`mdlrefit` — Neighborhood component analysis model for regression
`FeatureSelectionNCARegression` object

Neighborhood component analysis model or classification, returned as a `FeatureSelectionNCARegression` object. You can either save the results as a new model or update the existing model as `mdl = refit(mdl,Name,Value)`.

Examples

Refit NCA Model for Regression with Modified Settings

Load the sample data.

```
load('robotarm.mat')
```

The `robotarm` (pumadyn32nm) dataset is created using a robot arm simulator with 7168 training and 1024 test observations with 32 features [1], [2]. This is a preprocessed version of the original data set. Data are preprocessed by subtracting off a linear regression fit followed by normalization of all features to unit variance.

Compute the generalization error without feature selection.

```
nca = fsrnca(Xtrain,ytrain,'FitMethod','none','Standardize',1);
L = loss(nca,Xtest,ytest)
```

```
L = 0.9017
```

Now, refit the model and compute the prediction loss with feature selection, with $\lambda = 0$ (no regularization term) and compare to the previous loss value, to determine feature selection seems necessary for this problem. For the settings that you do not change, `refit` uses the settings of the initial model `nca`. For example, it uses the feature weights found in `nca` as the initial feature weights.

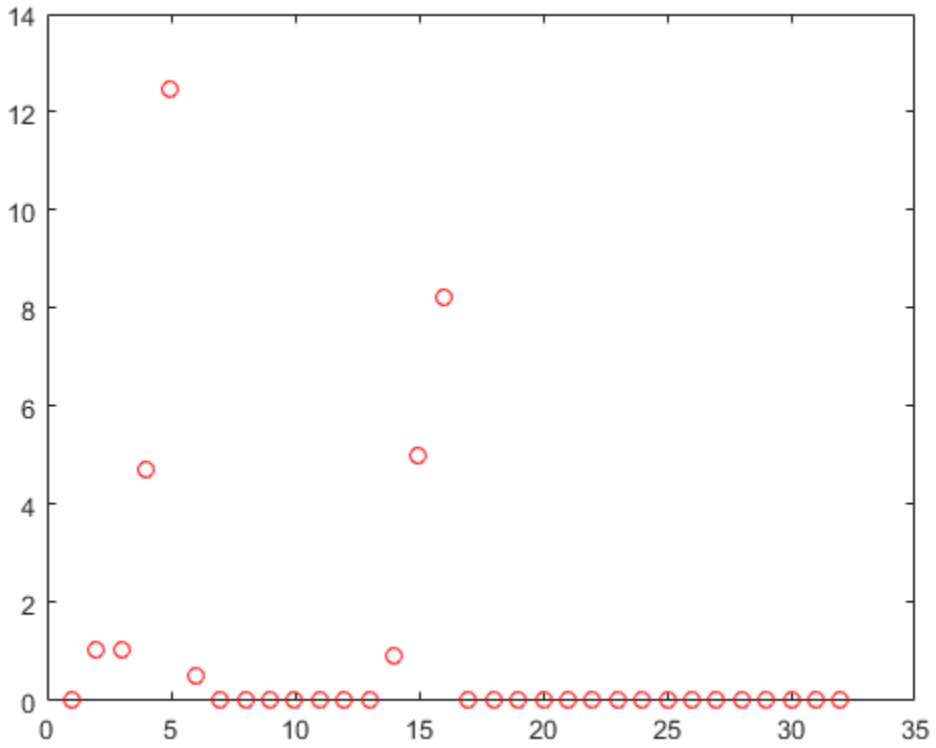
```
nca2 = refit(nca,'FitMethod','exact','Lambda',0);
L2 = loss(nca2,Xtest,ytest)
```

```
L2 = 0.1088
```

The decrease in the loss suggests that feature selection is necessary.

Plot the feature weights.

```
figure()
plot(nca2.FeatureWeights,'ro')
```



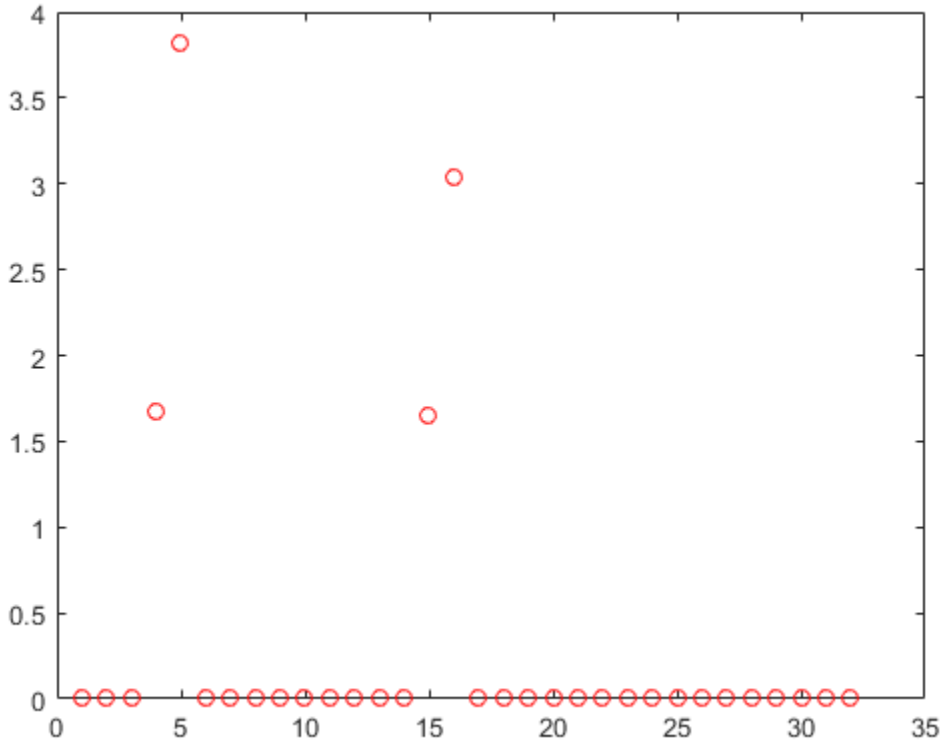
Tuning the regularization parameter usually improves the results. Suppose that, after tuning λ using cross-validation as in "Tune Regularization Parameter in NCA for Regression" on page 33-2503, the best λ value found is 0.0035. Refit the `nca` model using this λ value and stochastic gradient descent as the solver. Compute the prediction loss.

```
nca3 = refit(nca2,'FitMethod','exact','Lambda',0.0035,...
          'Solver','sgd');
L3 = loss(nca3,Xtest,ytest)
```

```
L3 = 0.0573
```

Plot the feature weights.

```
figure()
plot(nca3.FeatureWeights,'ro')
```



After tuning the regularization parameter, the loss decreased even more and the software identified four of the features as relevant.

References

[1] Rasmussen, C. E., R. M. Neal, G. E. Hinton, D. van Campand, M. Revow, Z. Ghahramani, R. Kustra, R. Tibshirani. The DELVE Manual, 1996, http://mlg.eng.cam.ac.uk/pub/pdf/RasNeaHinetal96.pdf

[2] https://www.cs.toronto.edu/~delve/data/datasets.html

See Also

`FeatureSelectionNCARegression` | `fsrnca` | `loss` | `predict`

Introduced in R2016b

reduceDimensions

Reduce dimensions of Sobol point set

Syntax

```
pr = reduceDimensions(p,d)
```

Description

`pr = reduceDimensions(p,d)` reduces the Sobol quasirandom point set `p` to the first `d` dimensions. `d` must be less than or equal to the number of dimensions in `p`.

The reduced point set `pr` is a `sobolset` object.

Examples

Reduce Dimensions of Sobol Point Set

Generate a seven-dimensional Sobol point set and scramble the points.

```
p = sobolset(7);
ps = scramble(p,'MatousekAffineOwen')
```

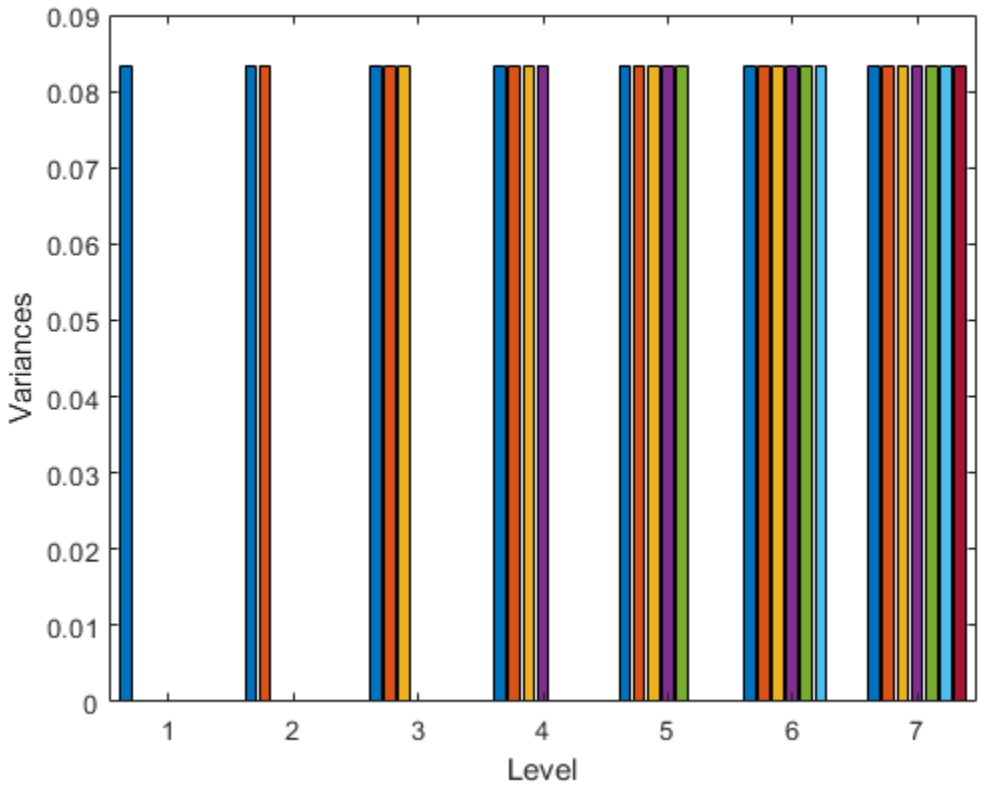
```
ps =
Sobol point set in 7 dimensions (9007199254740992 points)

Properties:
              Skip : 0
              Leap : 0
    ScrambleMethod : MatousekAffineOwen
        PointOrder : standard
```

Split the first 7168 points in `ps` into seven levels of 1024 points each. Reduce the first 1024 points to be one-dimensional, the second 1024 points to be two-dimensional, and so on. For each level, compute the variance of the point values in each dimension.

```
variance = NaN(7);
for level = 1:7
    pr = reduceDimensions(ps,level);
    pr.Skip = (level-1)*1024;
    pts = pr(1:1024,:);
    variance(level,1:level) = var(pts);
end
```

Plot the variances. The dark blue bars show the variance of the points in the first dimension, the dark orange bars show the variance of the points in the second dimension, and so on.

```
bar(variance)
xlabel('Level')
ylabel('Variances')
```



Input Arguments

p — Sobol point set

`sobolset` object

Sobol point set, specified as a `sobolset` object.

d — Number of dimensions to retain

positive integer scalar

Number of dimensions to retain from the point set `p`, specified as a positive integer scalar between 1 and the number of dimensions in `p`. The function always retains the first `d` dimensions of `p`.

Data Types: `single` | `double`

See Also

`net` | `scramble` | `sobolset`

Introduced in R2019a

refline

Add reference line to plot

Syntax

```
refline(m,b)
refline(coeffs)
refline
refline(ax, ___ )
hline = refline( ___ )
```

Description

`refline(m,b)` adds a reference line with slope `m` and intercept `b` to the current axes.

`refline(coeffs)` adds the line defined by the elements of the vector `coeffs` to the figure.

`refline` with no input arguments is equivalent to `lsline`.

`refline(ax, ___ )` adds a reference line to the plot in the axis specified by `ax`, using any of the input arguments in the previous syntaxes.

`hline = refline( ___ )` returns the reference line object `hline` using any of the input arguments in the previous syntaxes. Use `hline` to modify properties of a specific reference line after you create it. For a list of properties, see Line Properties.

Examples

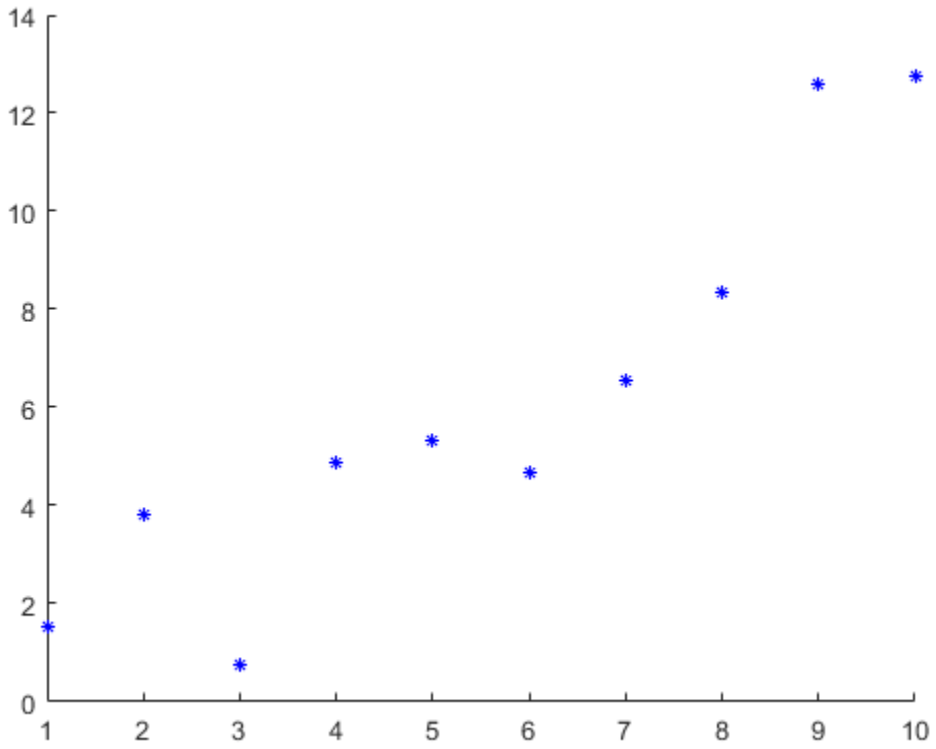
Add Reference Line at the Mean

Generate sample data for an independent variable `x` and a dependent variable `y` .

```
x = 1:10;
y = x + randn(1,10);
```

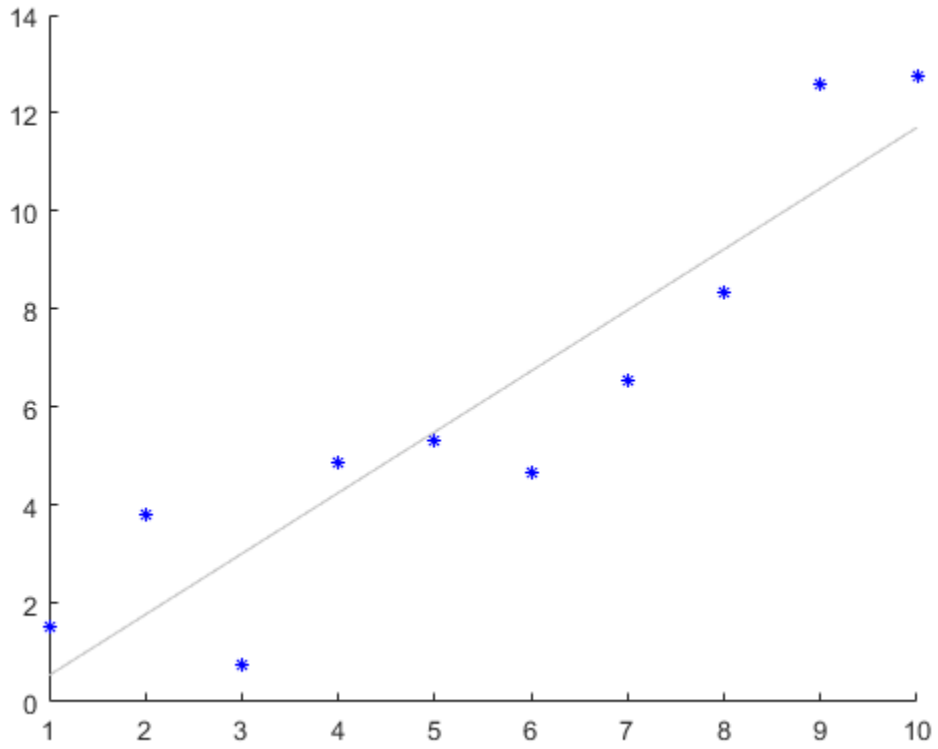
Create a scatter plot of `x` and `y` .

```
scatter(x,y,25,'b','*')
```



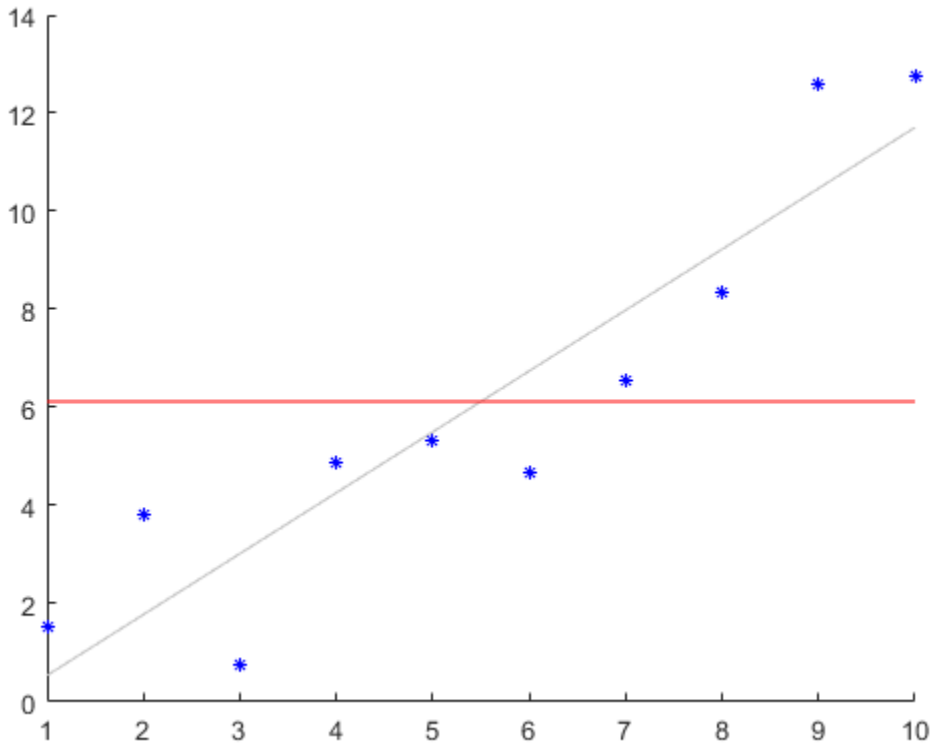
Superimpose a least-squares line on the scatter plot.

```
refline
```



Add a reference line at the mean of the scatter plot.

```
mu = mean(y);
hline = refline([0 mu]);
hline.Color = 'r';
```



The red line is the reference line at the mean of the data.

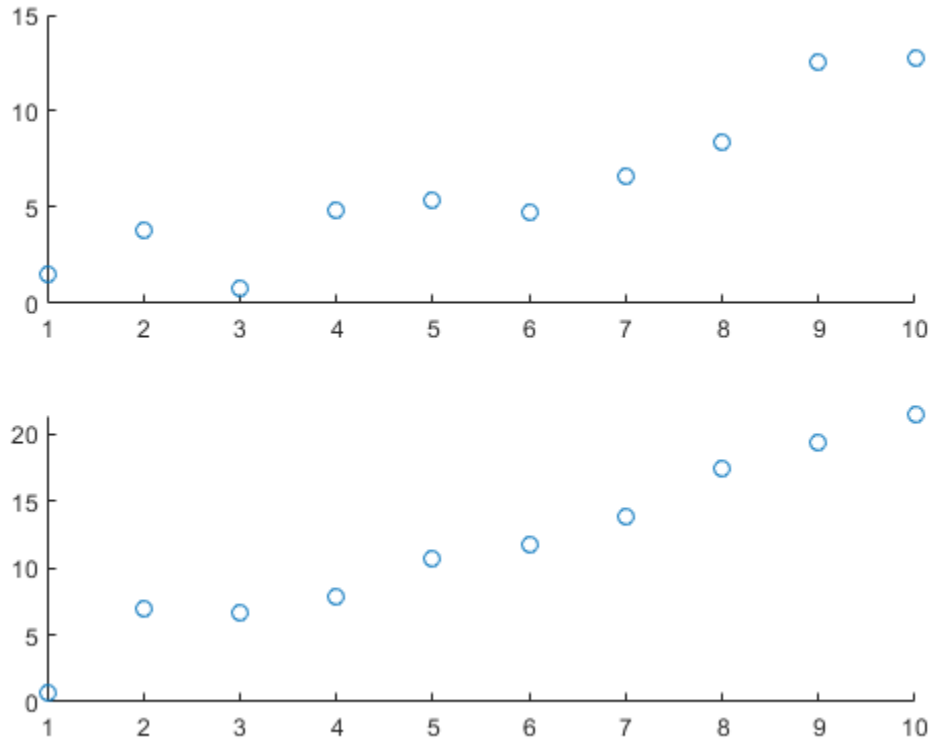
Specify Axes for Least-Squares and Reference Lines

Define the x-variable and two different y-variables to use for the plots.

```
rng default  % For reproducibility
x = 1:10;
y1 = x + randn(1,10);
y2 = 2*x + randn(1,10);
```

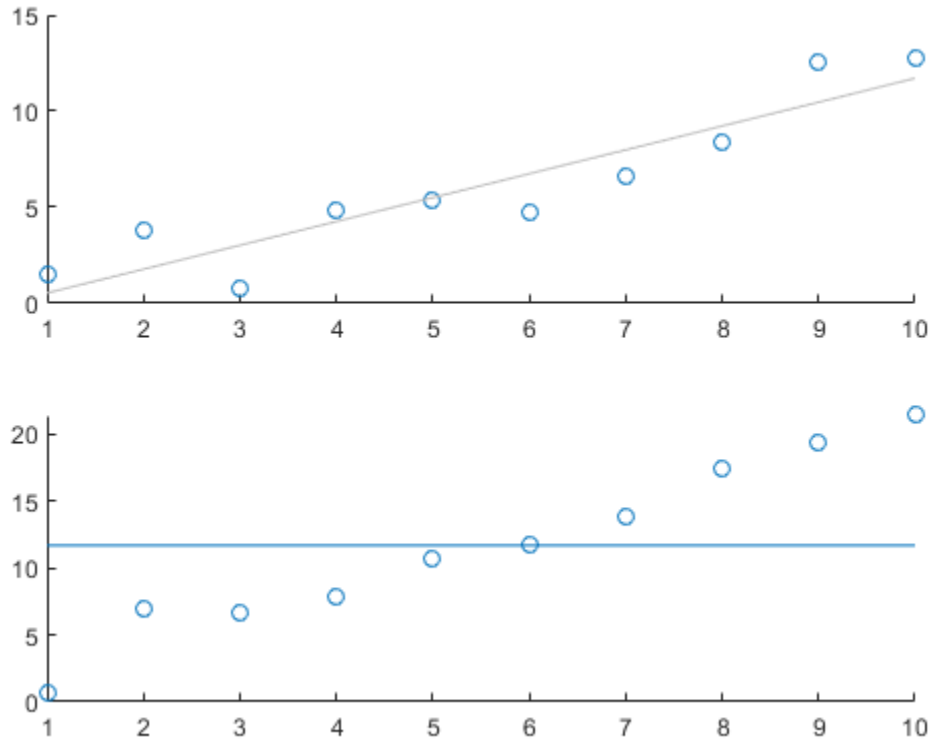
Define `ax1` as the top half of the figure, and `ax2` as the bottom half of the figure. Create the first scatter plot on the top axis using `y1`, and the second scatter plot on the bottom axis using `y2`.

```
figure
ax1 = subplot(2,1,1);
ax2 = subplot(2,1,2);

scatter(ax1,x,y1)
scatter(ax2,x,y2)
```



Superimpose a least-squares line on the top plot, and a reference line at the mean of the `y2` values in the bottom plot.

```
lsline(ax1) % This is equivalent to refline(ax1)

mu = mean(y2);
refline(ax2,[0 mu])
```



Input Arguments

`m` — Slope of reference line

numeric scalar

Slope of the reference line, specified as a numeric scalar. The function uses `m` to define the line

```
y = m*x + b.
```

Example: `refline(-1,1)`

Data Types: `single` | `double`

`b` — Intercept of reference line

numeric scalar

Intercept of the reference line, specified as a numeric scalar. The function uses `b` to define the line

```
y = m*x + b.
```

Example: `refline(2,-10)`

Data Types: `single` | `double`

`coeffs` — Linear coefficients

length-two numeric vector

Linear coefficients, specified as a length-two numeric vector. `coeffs` contains the coefficients of a line defined as

```
y = coeffs(1)*x + coeffs(2).
```

Example: `refline([-1,2])`

Data Types: `single` | `double`

`ax` — Target axes
`gca` (default) | `axes` object

Target axes, specified as an `axes` object. If you do not specify the axes and if the current axes are Cartesian axes, then the `refline` function uses the current axes.

Output Arguments

`hline` — One or more reference line objects
scalar | vector

One or more reference line objects, returned as a scalar or a vector. These objects are unique identifiers, which you can use to query and modify properties of a specific reference line. For a list of properties, see Chart Line.

See Also

`gline` | `lsline` | `refcurve`

Introduced before R2006a

regress

Multiple linear regression

Syntax

```
b = regress(y,X)
[b,bint] = regress(y,X)
[b,bint,r] = regress(y,X)
[b,bint,r,rint] = regress(y,X)
[b,bint,r,rint,stats] = regress(y,X)
[___] = regress(y,X,alpha)
```

Description

`b = regress(y,X)` returns a vector `b` of coefficient estimates for a multiple linear regression of the responses in vector `y` on the predictors in matrix `X`. To compute coefficient estimates for a model with a constant term (intercept), include a column of ones in the matrix `X`.

`[b,bint] = regress(y,X)` also returns a matrix `bint` of 95% confidence intervals for the coefficient estimates.

`[b,bint,r] = regress(y,X)` also returns an additional vector `r` of residuals.

`[b,bint,r,rint] = regress(y,X)` also returns a matrix `rint` of intervals that can be used to diagnose outliers.

`[b,bint,r,rint,stats] = regress(y,X)` also returns a vector `stats` that contains the R^2 statistic, the F-statistic and its p-value, and an estimate of the error variance. The matrix `X` must include a column of ones for the software to compute the model statistics correctly.

`[___] = regress(y,X,alpha)` uses a `100*(1-alpha)`% confidence level to compute `bint` and `rint`. Specify any of the output argument combinations in the previous syntaxes.

Examples

Estimate Multiple Linear Regression Coefficients

Load the `carsmall` data set. Identify weight and horsepower as predictors and mileage as the response.

```
load carsmall
x1 = Weight;
x2 = Horsepower;    % Contains NaN data
y = MPG;
```

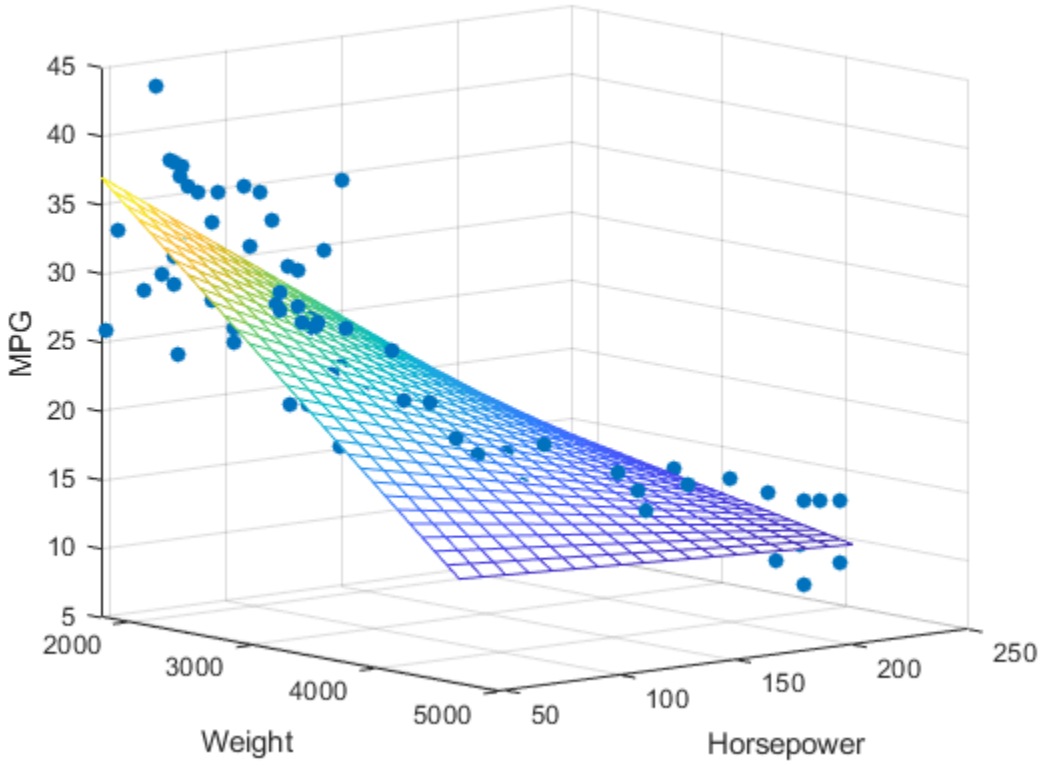
Compute the regression coefficients for a linear model with an interaction term.

```
X = [ones(size(x1)) x1 x2 x1.*x2];
b = regress(y,X)    % Removes NaN data
```

```
b = 4×1

   60.7104
   -0.0102
   -0.1882
    0.0000
```

Plot the data and the model.

```
scatter3(x1,x2,y,'filled')
hold on
x1fit = min(x1):100:max(x1);
x2fit = min(x2):10:max(x2);
[X1FIT,X2FIT] = meshgrid(x1fit,x2fit);
YFIT = b(1) + b(2)*X1FIT + b(3)*X2FIT + b(4)*X1FIT.*X2FIT;
mesh(X1FIT,X2FIT,YFIT)
xlabel('Weight')
ylabel('Horsepower')
zlabel('MPG')
view(50,10)
hold off
```



Diagnose Outliers Using Residuals

Load the `examgrades` data set.

```
load examgrades
```

Use the last exam scores as response data and the first two exam scores as predictor data.

```
y = grades(:,5);
X = [ones(size(grades(:,1))) grades(:,1:2)];
```

Perform multiple linear regression with alpha = 0.01.

```
[~,~,r,rint] = regress(y,X,0.01);
```

Diagnose outliers by finding the residual intervals `rint` that do not contain 0.

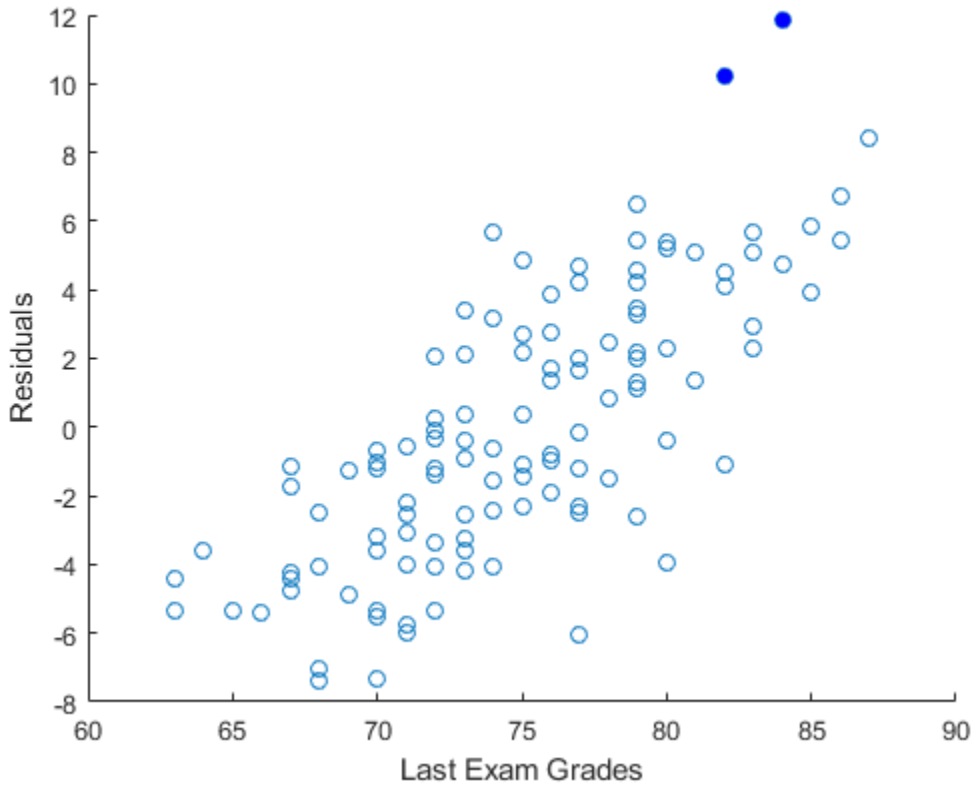
```
contain0 = (rint(:,1)<0 & rint(:,2)>0);
idx = find(contain0==false)
```

```
idx = 2×1

    53
    54
```

Observations 53 and 54 are possible outliers.

Create a scatter plot of the residuals. Fill in the points corresponding to the outliers.

```
hold on
scatter(y,r)
scatter(y(idx),r(idx),'b','filled')
xlabel("Last Exam Grades")
ylabel("Residuals")
hold off
```



Determine Significance of Linear Regression Relationship

Load the `hald` data set. Use `heat` as the response variable and `ingredients` as the predictor data.

```
load hald
y = heat;
X1 = ingredients;
x1 = ones(size(X1,1),1);
X = [x1 X1];    % Includes column of ones
```

Perform multiple linear regression and generate model statistics.

```
[~,~,~,~,stats] = regress(y,X)
```

```
stats = 1×4

    0.9824  111.4792    0.0000    5.9830
```

Because the R^2 value of `0.9824` is close to 1, and the p-value of `0.0000` is less than the default significance level of 0.05, a significant linear regression relationship exists between the response `y` and the predictor variables in `X`.

Input Arguments

y — Response data

numeric vector

Response data, specified as an *n*-by-1 numeric vector. Rows of `y` correspond to different observations. `y` must have the same number of rows as `X`.

Data Types: `single` | `double`

X — Predictor data

numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix. Rows of `X` correspond to observations, and columns correspond to predictor variables. `X` must have the same number of rows as `y`.

Data Types: `single` | `double`

alpha — Significance level

`0.05` (default) | positive scalar

Significance level, specified as a positive scalar. `alpha` must be between 0 and 1.

Data Types: `single` | `double`

Output Arguments

b — Coefficient estimates for multiple linear regression

numeric vector

Coefficient estimates for multiple linear regression, returned as a numeric vector. `b` is a *p*-by-1 vector, where *p* is the number of predictors in `X`. If the columns of `X` are linearly dependent, `regress` sets the maximum number of elements of `b` to zero.

Data Types: `double`

bint — Lower and upper confidence bounds for coefficient estimates

numeric matrix

Lower and upper confidence bounds for coefficient estimates, returned as a numeric matrix. `bint` is a *p*-by-2 matrix, where *p* is the number of predictors in `X`. The first column of `bint` contains lower confidence bounds for each of the coefficient estimates; the second column contains upper confidence bounds. If the columns of `X` are linearly dependent, `regress` returns zeros in elements of `bint` corresponding to the zero elements of `b`.

Data Types: `double`

r — Residuals

numeric vector

Residuals, returned as a numeric vector. `r` is an *n*-by-1 vector, where *n* is the number of observations, or rows, in `X`.

Data Types: `single` | `double`

rint — Intervals to diagnose outliers

numeric matrix

Intervals to diagnose outliers, returned as a numeric matrix. `rint` is an *n*-by-2 matrix, where *n* is the number of observations, or rows, in `X`. If the interval `rint(i,:)` for observation `i` does not contain zero, the corresponding residual is larger than expected in `100*(1-alpha)`% of new observations, suggesting an outlier. For more information, see “Algorithms” on page 33-5281.

Data Types: `single` | `double`

`stats` — Model statistics
numeric vector

Model statistics, returned as a numeric vector including the R^2 statistic, the F-statistic and its p-value, and an estimate of the error variance.

- `X` must include a column of ones so that the model contains a constant term. The F-statistic and its p-value are computed under this assumption and are not correct for models without a constant.
- The F-statistic is the test statistic of the F-test on the regression model. The F-test looks for a significant linear regression relationship between the response variable and the predictor variables.
- The R^2 statistic can be negative for models without a constant, indicating that the model is not appropriate for the data.

Data Types: `single` | `double`

Tips

- `regress` treats `NaN` values in `X` or `y` as missing values. `regress` omits observations with missing values from the regression fit.

Algorithms

Residual Intervals

In a linear model, observed values of `y` and their residuals are random variables. Residuals have normal distributions with zero mean but with different variances at different values of the predictors. To put residuals on a comparable scale, `regress` “Studentizes” the residuals. That is, `regress` divides the residuals by an estimate of their standard deviation that is independent of their value. Studentized residuals have t-distributions with known degrees of freedom. The intervals returned in `rint` are shifts of the `100*(1-alpha)`% confidence intervals of these t-distributions, centered at the residuals.

Alternative Functionality

`regress` is useful when you simply need the output arguments of the function and when you want to repeat fitting a model multiple times in a loop. If you need to investigate a fitted regression model further, create a linear regression model object `LinearModel` by using `fitlm` or `stepwiselm`. A `LinearModel` object provides more features than `regress`.

- Use the properties of `LinearModel` to investigate a fitted linear regression model. The object properties include information about coefficient estimates, summary statistics, fitting method, and input data.
- Use the object functions of `LinearModel` to predict responses and to modify, evaluate, and visualize the linear regression model.

- Unlike `regress`, the `fitlm` function does not require a column of ones in the input data. A model created by `fitlm` always includes an intercept term unless you specify not to include it by using the `'Intercept'` name-value pair argument.
- You can find the information in the output of `regress` using the properties and object functions of `LinearModel`.

Output of regress	Equivalent Values in LinearModel
`b`	See the `Estimate` column of the `Coefficients` property.
`bint`	Use the `coefCI` function.
`r`	See the `Raw` column of the `Residuals` property.
`rint`	Not supported. Instead, use studentized residuals (`Residuals` property) and observation diagnostics (`Diagnostics` property) to find outliers.
`stats`	See the model display in the Command Window. You can find the statistics in the model properties (`MSE` and `Rsquared`) and by using the `anova` function.

References

[1] Chatterjee, S., and A. S. Hadi. “Influential Observations, High Leverage Points, and Outliers in Linear Regression.” *Statistical Science*. Vol. 1, 1986, pp. 379–416.

See Also

`LinearModel` | `fitlm` | `mvregress` | `rcoplot` | `stepwiselm`

Topics

“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35

Introduced before R2006a

RegressionBaggedEnsemble

Package: `classreg.learning.regr`
Superclasses: `RegressionEnsemble`

Regression ensemble grown by resampling

Description

`RegressionBaggedEnsemble` combines a set of trained weak learner models and data on which these learners were trained. It can predict ensemble response for new data by aggregating predictions from its weak learners.

Construction

Create a bagged regression ensemble object using `fitrensemble`. Set the name-value pair argument `'Method'` of `fitrensemble` to `'Bag'` to use bootstrap aggregation (bagging, for example, random forest).

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
```

```
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CombineWeights`

A character vector describing how the ensemble combines learner predictions.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`FitInfo`

A numeric array of fit information. The `FitInfoDescription` property describes the content of this array.

`FitInfoDescription`

Character vector describing the meaning of the `FitInfo` array.

`FResample`

A numeric scalar between `0` and `1`. `FResample` is the fraction of training data `fitrensemble` resampled at random for every weak learner when constructing the ensemble.

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`LearnerNames`

Cell array of character vectors with names of the weak learners in the ensemble. The name of each learner appears just once. For example, if you have an ensemble of 100 trees, `LearnerNames` is `{'Tree'}`.

Method

A character vector with the name of the algorithm `fitrensemble` used for training the ensemble.

ModelParameters

Parameters used in training `ens`.

NumObservations

Numeric scalar containing the number of observations in the training data.

NumTrained

Number of trained learners in the ensemble, a positive scalar.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in `X`.

ReasonForTermination

A character vector describing the reason `fitrensemble` stopped adding weak learners to the ensemble.

Regularization

A structure containing the result of the `regularize` method. Use `Regularization` with `shrink` to lower resubstitution error and shrink the ensemble.

Replace

Boolean flag indicating if training data for weak learners in this ensemble were sampled with replacement. `Replace` is `true` for sampling with replacement, `false` otherwise.

ResponseName

A character vector with the name of the response variable `Y`.

ResponseTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`.

Add or change a `ResponseTransform` function using dot notation:

```
ens.ResponseTransform = @function
```

Trained

The trained learners, a cell array of compact regression models.

TrainedWeights

A numeric vector of weights the ensemble assigns to its learners. The ensemble computes predicted response by aggregating weighted predictions from its learners.

`UseObsForLearner`

A logical matrix of size `N`-by-`NumTrained`, where `N` is the number of rows (observations) in the training data `X`, and `NumTrained` is the number of trained weak learners. `UseObsForLearner(I,J)` is `true` if observation `I` was used for training learner `J`, and is `false` otherwise.

`W`

The scaled `weights`, a vector with length `n`, the number of rows in `X`. The sum of the elements of `W` is `1`.

`X`

The matrix or table of predictor values that trained the ensemble. Each column of `X` represents one variable, and each row represents one observation.

`Y`

The numeric column vector with the same number of rows as `X` that trained the ensemble. Each entry in `Y` is the response to the data in the corresponding row of `X`.

Object Functions

compact	Create compact regression ensemble
crossval	Cross validate ensemble
cvshrink	Cross validate shrinking (pruning) ensemble
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error
oobLoss	Out-of-bag regression error
oobPermutedPredictorImportance	Predictor importance estimates by permutation of out-of-bag predictor observations for random forest of regression trees
oobPredict	Predict out-of-bag response of ensemble
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using ensemble of regression models
predictorImportance	Estimates of predictor importance for regression ensemble
regularize	Find weights to minimize resubstitution error plus penalty term
removeLearners	Remove members of compact regression ensemble
resubLoss	Regression error by resubstitution
resubPredict	Predict response of ensemble by resubstitution
resume	Resume training ensemble
shapley	Shapley values
shrink	Prune ensemble

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Bagged Ensemble of Regression Trees

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

Train a bagged ensemble of 100 regression trees using all measurements.

```
Mdl = fitrensemble(X,Y,'Method','bag')
```

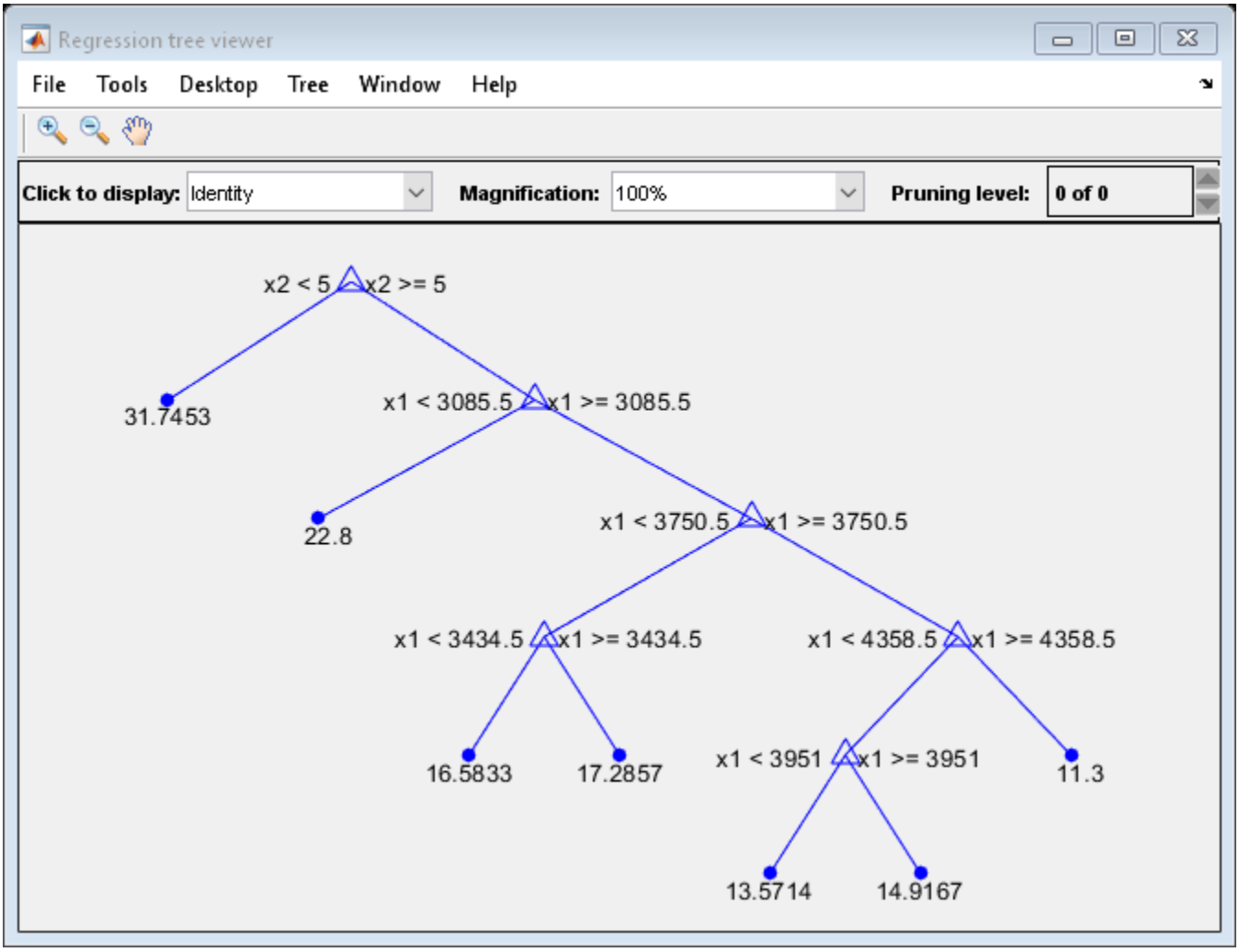
```
Mdl =
  RegressionBaggedEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94
                NumTrained: 100
                    Method: 'Bag'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: []
       FitInfoDescription: 'None'
           Regularization: []
                 FResample: 1
                   Replace: 1
          UseObsForLearner: [94x100 logical]


  Properties, Methods
```

`Mdl` is a `RegressionBaggedEnsemble` model object.

`Mdl.Trained` is the property that stores a 100-by-1 cell vector of the trained, compact regression trees (`CompactRegressionTree` model objects) that compose the ensemble.

Plot a graph of the first trained regression tree.

```
view(Mdl.Trained{1},'Mode','graph')
```



By default, `fitrensemble` grows deep trees for bags of trees.

Estimate the in-sample mean-squared error (MSE).

```
L = resubLoss(Mdl)
```

```
L = 12.4048
```

Tip

For a bagged ensemble of regression trees, the `Trained` property of `ens` stores a cell vector of `ens.NumTrained` `CompactRegressionTree` model objects. For a textual or graphical display of tree *`t`* in the cell vector, enter

```
view(ens.Trained{t})
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the RegressionEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitrensemble`, code generation limitations for regression trees also apply to ensembles of regression trees. For more details, see "Code Generation" on page 33-868 of the `CompactRegressionTree` class.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`RegressionEnsemble` | `fitrensemble` | `view`

Introduced in R2011a

RegressionEnsemble

Package: `classreg.learning.regr`
Superclasses: `CompactRegressionEnsemble`

Ensemble regression

Description

`RegressionEnsemble` combines a set of trained weak learner models and data on which these learners were trained. It can predict ensemble response for new data by aggregating predictions from its weak learners.

Construction

Create a regression ensemble object using `fitrensemble`.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CombineWeights`

A character vector describing how the ensemble combines learner predictions.

`ExpandedPredictorNames`

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

`FitInfo`

A numeric array of fit information. The `FitInfoDescription` property describes the content of this array.

`FitInfoDescription`

Character vector describing the meaning of the `FitInfo` array.

`LearnerNames`

Cell array of character vectors with names of the weak learners in the ensemble. The name of each learner appears just once. For example, if you have an ensemble of 100 trees, `LearnerNames` is `{'Tree'}`.

`HyperparameterOptimizationResults`

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`Method`

A character vector with the name of the algorithm `fitrensemble` used for training the ensemble.

`ModelParameters`

Parameters used in training `ens`.

NumObservations

Numeric scalar containing the number of observations in the training data.

NumTrained

Number of trained learners in the ensemble, a positive scalar.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in `X`.

ReasonForTermination

A character vector describing the reason `fitrensemble` stopped adding weak learners to the ensemble.

Regularization

A structure containing the result of the `regularize` method. Use `Regularization` with `shrink` to lower resubstitution error and shrink the ensemble.

ResponseName

A character vector with the name of the response variable `Y`.

ResponseTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`.

Add or change a `ResponseTransform` function using dot notation:

```
ens.ResponseTransform = @function
```

Trained

The trained learners, a cell array of compact regression models.

TrainedWeights

A numeric vector of weights the ensemble assigns to its learners. The ensemble computes predicted response by aggregating weighted predictions from its learners.

W

The scaled `weights`, a vector with length `n`, the number of rows in `X`. The sum of the elements of `W` is `1`.

X

The matrix or table of predictor values that trained the ensemble. Each column of `X` represents one variable, and each row represents one observation.

Y

The numeric column vector with the same number of rows as X that trained the ensemble. Each entry in Y is the response to the data in the corresponding row of X.

Object Functions

compact	Create compact regression ensemble
crossval	Cross validate ensemble
cvshrink	Cross validate shrinking (pruning) ensemble
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using ensemble of regression models
predictorImportance	Estimates of predictor importance for regression ensemble
regularize	Find weights to minimize resubstitution error plus penalty term
removeLearners	Remove members of compact regression ensemble
resubLoss	Regression error by resubstitution
resubPredict	Predict response of ensemble by resubstitution
resume	Resume training ensemble
shapley	Shapley values
shrink	Prune ensemble

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Boosted Regression Ensemble

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

Train a boosted ensemble of 100 regression trees using the `LSBoost` method. Specify that `Cylinders` is a categorical variable.

```
Mdl = fitrensemble(X,Y,'Method','LSBoost',...
    'PredictorNames',{'W','C'},'CategoricalPredictors',2)
```

```
Mdl =
  RegressionEnsemble
            PredictorNames: {'W'  'C'}
              ResponseName: 'Y'
    CategoricalPredictors: 2
        ResponseTransform: 'none'
          NumObservations: 94
```

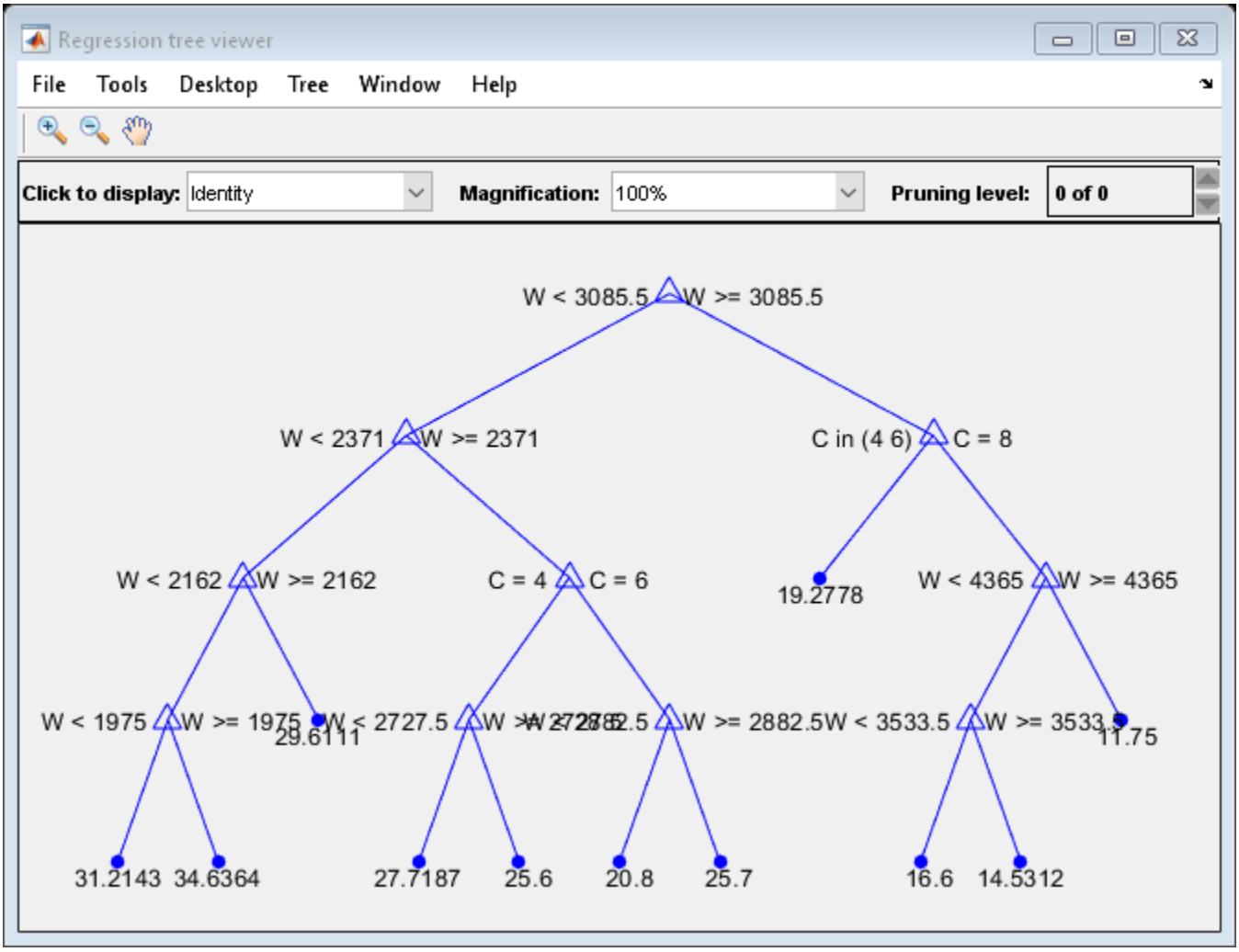
```
               NumTrained: 100
                   Method: 'LSBoost'
              LearnerNames: {'Tree'}
    ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

`Mdl` is a `RegressionEnsemble` model object that contains the training data, among other things.

`Mdl.Trained` is the property that stores a 100-by-1 cell vector of the trained regression trees (`CompactRegressionTree` model objects) that compose the ensemble.

Plot a graph of the first trained regression tree.

```
view(Mdl.Trained{1},'Mode','graph')
```



By default, `fitrensemble` grows shallow trees for boosted ensembles of trees.

Predict the fuel economy of 4,000 pound cars with 4, 6, and 8 cylinders.

```
XNew = [4000*ones(3,1) [4; 6; 8]];
mpgNew = predict(Mdl,XNew)
```

```
mpgNew = 3×1

   19.5926
   18.6388
   15.4810
```

Tip

For an ensemble of regression trees, the `Trained` property contains a cell vector of `ens.NumTrained` `CompactRegressionTree` model objects. For a textual or graphical display of tree `t` in the cell vector, enter

```
view(ens.Trained{t})
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.
- To integrate the prediction of an ensemble into Simulink, you can use the RegressionEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an ensemble by using `fitrensemble`, code generation limitations for regression trees also apply to ensembles of regression trees. For more details, see "Code Generation" on page 33-868 of the `CompactRegressionTree` class.

For more information, see "Introduction to Code Generation" on page 32-2.

See Also

`ClassificationEnsemble` | `CompactRegressionEnsemble` | `fitrensemble` | `templateTree` | `view`

Introduced in R2011a

RegressionEnsemble Predict

Predict responses using ensemble of decision trees for regression
Library: Statistics and Machine Learning Toolbox / Regression

Description

The RegressionEnsemble Predict block predicts responses using an ensemble of decision trees (`RegressionEnsemble`, `RegressionBaggedEnsemble`, or `CompactRegressionEnsemble`).

Import a trained regression object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **yfit** returns a predicted response for the observation.

Ports

Input

x — Predictor data

row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the model specified by **`Select trained machine learning model`**.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

yfit — Predicted response

scalar

Predicted response, returned as a scalar.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

Select trained machine learning model — Regression ensemble model

`ensMdl` (default) | `RegressionEnsemble` object | `RegressionBaggedEnsemble` object | `CompactRegressionEnsemble` object

Specify the name of a workspace variable that contains a `RegressionEnsemble` object, `RegressionBaggedEnsemble` object, or `CompactRegressionEnsemble` object.

When you train the model by using `fitrensemble`, the following restrictions apply:

- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ResponseTransform'` name-value argument must be `'none'` (default).
- You cannot use surrogate splits for tree weak learners, that is, the value of the `'Surrogate'` name-value argument must be `'off'` (default) when you define tree weak learners by using the `templateTree` function.

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `RegressionEnsemble` object | `RegressionBaggedEnsemble` object | `CompactRegressionEnsemble` object
Default: `'ensMdl'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations
`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use

Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action
`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use
Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type
`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use
Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Output data type` — Data type of yfit output
`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **yfit** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use

Block Parameter: `OutDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Output minimum` — Minimum value of yfit output for range checking
`[]` (default) | scalar

Lower value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output minimum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `OutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Output maximum` — Maximum value of yfit output for range checking

`[]` (default) | scalar

Upper value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output maximum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `OutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Weak learner data type` — Data type of weak learner outputs

`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the outputs from weak learners. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use

Block Parameter: `WeakLearnerDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Weak learner minimum` — Minimum value of weak learner outputs for range checking

`[]` (default) | scalar

Lower value of the weak learner output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Weak learner minimum** parameter does not saturate or clip the actual weak learner output signals.

Programmatic Use

Block Parameter: `WeakLearnerOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Weak learner maximum` — Maximum value of weak learner outputs for range checking

`[]` (default) | scalar

Upper value of the weak learner output range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Weak learner maximum** parameter does not saturate or clip the actual weak learner output signals.

Programmatic Use

Block Parameter: `WeakLearnerOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of an ensemble of decision trees (`RegressionEnsemble`, `RegressionBaggedEnsemble`, or `CompactRegressionEnsemble`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the RegressionEnsemble Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion

Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks

ClassificationEnsemble Predict | RegressionSVM Predict | RegressionTree Predict

Objects

`CompactRegressionEnsemble` | `RegressionBaggedEnsemble` | `RegressionEnsemble`

Functions

`fitrensemble` | `predict`

Topics

“Predict Responses Using RegressionSVM Predict Block” on page 32-115
“Predict Responses Using RegressionTree Predict Block” on page 32-127

“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2021a

RegressionGAM

Generalized additive model (GAM) for regression

Description

A `RegressionGAM` object is a generalized additive model on page 33-5314 (GAM) object for regression. It is an interpretable model that explains a response variable using a sum of univariate and bivariate shape functions.

You can predict responses for new observations by using the `predict` function, and plot the effect of each shape function on the prediction (response value) for an observation by using the `plotLocalEffects` function. For the full list of object functions for `RegressionGAM`, see “Object Functions” on page 33-5308.

Creation

Create a `RegressionGAM` object by using `fitrgam`. You can specify both linear terms and interaction terms for predictors to include univariate shape functions (predictor trees) and bivariate shape functions (interaction trees) in a trained model, respectively.

You can update a trained model by using `resume` or `addInteractions`.

- The `resume` function resumes training for the existing terms in a model.
- The `addInteractions` function adds interaction terms to a model that contains only linear terms.

Properties

GAM Properties

`BinEdges` — Bin edges for numeric predictors
cell array of numeric vectors | `[ ]`

This property is read-only.

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
```

```
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Data Types: `cell`

`Interactions` — Interaction term indices
two-column matrix of positive integers | `[]`

This property is read-only.

Interaction term indices, specified as a `t`-by-2 matrix of positive integers, where `t` is the number of interaction terms in the model. Each row of the matrix represents one interaction term and contains the column indexes of the predictor data `X` for the interaction term. If the model does not include an interaction term, then this property is empty (`[]`).

The software adds interaction terms to the model in the order of importance based on the *p*-values. Use this property to check the order of the interaction terms added to the model.

Data Types: `double`

`Intercept` — Intercept term of model
numeric scalar

This property is read-only.

Intercept (constant) term of the model, which is the sum of the intercept terms in the predictor trees and interaction trees, specified as a numeric scalar.

Data Types: `single` | `double`

`ModelParameters` — Parameters used to train model
model parameter object

This property is read-only.

Parameters used to train the model, specified as a model parameter object. `ModelParameters` contains parameter values such as those for the name-value arguments used to train the model. `ModelParameters` does not contain estimated parameters.

Access the fields of `ModelParameters` by using dot notation. For example, access the maximum number of decision splits per interaction tree by using `Mdl.ModelParameters.MaxNumSplitsPerInteraction`.

`PairDetectionBinEdges` — Bin edges for interaction term detection
cell array of numeric vectors

This property is read-only.

Bin edges for interaction term detection for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

To speed up the interaction term detection process, the software bins numeric predictors into at most 8 equiprobable bins. The number of bins can be less than 8 if a predictor has fewer than 8 unique values.

Data Types: `cell`

`ReasonForTermination` — Reason training stops
structure

This property is read-only.

Reason training the model stops, specified as a structure with two fields, `PredictorTrees` and `InteractionTrees`.

Use this property to check if the model contains the specified number of trees for each linear term (`'NumTreesPerPredictor'`) and for each interaction term (`'NumTreesPerInteraction'`). If the `fitrgam` function terminates training before adding the specified number of trees, this property contains the reason for the termination.

Data Types: `struct`

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, specified as a cell array of character vectors.

`ExpandedPredictorNames` is the same as `PredictorNames` for a generalized additive model.

Data Types: `cell`

`NumObservations` — Number of observations
numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

Data Types: `double`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

`RowsUsed` — Rows used in fitting
`[]` | logical vector

This property is read-only.

Rows of the original training data used in fitting the `RegressionGAM` model, specified as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to train the model, specified as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `'Weights'` name-value argument so that the elements of `W` sum up to 1.

Data Types: `double`

`X` — Predictors
numeric matrix | table

This property is read-only.

Predictors used to train the model, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

`Y` — Response
numeric vector

This property is read-only.

Response, specified as a numeric vector.

Each row of `Y` represents the observed response of the corresponding row of `X`.

Data Types: `single` | `double`

Object Functions

Create `CompactRegressionGAM`

compact	Reduce size of machine learning model

Create `RegressionPartitionedGAM`

crossval	Cross-validate machine learning model

Update GAM

addInteractions	Add interaction terms to univariate generalized additive model (GAM)
resume	Resume training of generalized additive model (GAM)

Interpret Prediction

lime	Local interpretable model-agnostic explanations (LIME)
partialDependence	Compute partial dependence
plotLocalEffects	Plot local effects of terms in generalized additive model (GAM)
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
shapley	Shapley values

Assess Predictive Performance on New Observations

predict	Predict responses using generalized additive model (GAM)
loss	Regression loss for generalized additive model (GAM)

Assess Predictive Performance on Training Data

resubPredict	Predict responses for training data using trained regression model
resubLoss	Resubstitution regression loss

Examples

Train Generalized Additive Model

Train a univariate GAM, which contains linear terms for predictors. Then, interpret the prediction for a specified data instance by using the `plotLocalEffects` function.

Load the data set `NYCHousing2015`.

```
load NYCHousing2015
```

The data set includes 10 variables with information on the sales of properties in New York City in 2015. This example uses these variables to analyze the sale prices (`SALEPRICE`).

Preprocess the data set. Remove outliers, convert the `datetime` array (`SALEDATE`) to the month numbers, and move the response variable (`SALEPRICE`) to the last column.

```
idx = isoutlier(NYCHousing2015.SALEPRICE);
NYCHousing2015(idx,:) = [];
NYCHousing2015.SALEDATE = month(NYCHousing2015.SALEDATE);
NYCHousing2015 = movevars(NYCHousing2015,'SALEPRICE','After','SALEDATE');
```

Display the first three rows of the table.

```
head(NYCHousing2015,3)
```

```
ans=3×10 table
    BOROUGH    NEIGHBORHOOD       BUILDINGCLASSCATEGORY        RESIDENTIALUNITS    COMMERCIALUNI
    _______    ____________    ___________________________    ________________    _____________

       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   0
       2       {'BATHGATE'}    {'01  ONE FAMILY DWELLINGS'}          1                   1
```

Train a univariate GAM for the sale prices. Specify the variables for `BOROUGH`, `NEIGHBORHOOD`, `BUILDINGCLASSCATEGORY`, and `SALEDATE` as categorical predictors.

```
Mdl = fitrgam(NYCHousing2015,'SALEPRICE','CategoricalPredictors',[1 2 3 9])
```

```
Mdl =
  RegressionGAM
             PredictorNames: {1x9 cell}
               ResponseName: 'SALEPRICE'
    CategoricalPredictors: [1 2 3 9]
        ResponseTransform: 'none'
                Intercept: 3.7518e+05
```

```
          NumObservations: 83517

  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object. The model display shows a partial list of the model properties. To view the full list of properties, double-click the variable name `Mdl` in the Workspace. The Variables editor opens for `Mdl`. Alternatively, you can display the properties in the Command Window by using dot notation. For example, display the estimated intercept (constant) term of `Mdl`.

```
Mdl.Intercept
```

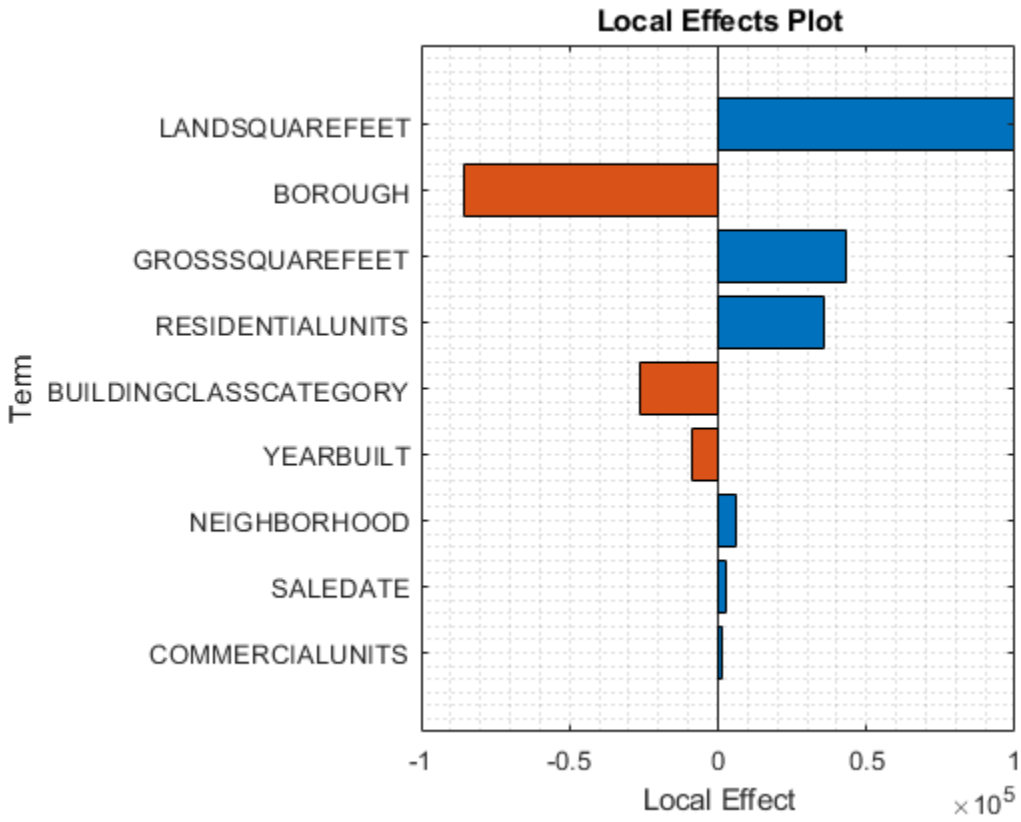
```
ans = 3.7518e+05
```

Predict the sale price for the first observation of the training data, and plot the local effects of the terms in `Mdl` on the prediction.

```
yFit = predict(Mdl,NYCHousing2015(1,:))
```

```
yFit = 4.4421e+05
```

```
plotLocalEffects(Mdl,NYCHousing2015(1,:))
```



The `predict` function predicts the sale price for the first observation as `4.4421e5`. The `plotLocalEffects` function creates a horizontal bar graph that shows the local effects of the terms in `Mdl` on the prediction. Each local effect value shows the contribution of each term to the predicted sale price.

Train GAM with Interaction Terms

Train a generalized additive model that contains linear and interaction terms for predictors in three different ways:

- Specify the interaction terms using the `formula` input argument.
- Specify the `'Interactions'` name-value argument.
- Build a model with linear terms first and add interaction terms to the model by using the `addInteractions` function.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table that contains the predictor variables (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and the response variable (`MPG`).

```
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Specify `formula`

Train a GAM that contains the four linear terms (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and two interaction terms (`Acceleration*Displacement` and `Displacement*Horsepower`). Specify the terms using a formula in the form `'Y ~ terms'`.

```
Mdl1 = fitrgam(tbl,'MPG ~ Acceleration + Displacement + Horsepower + Weight + Acceleration:Displa
```

The function adds interaction terms to the model in the order of importance. You can use the `Interactions` property to check the interaction terms in the model and the order in which `fitrgam` adds them to the model. Display the `Interactions` property.

```
Mdl1.Interactions
```

```
ans = 2×2

     2     3
     1     2
```

Each row of `Interactions` represents one interaction term and contains the column indexes of the predictor variables for the interaction term.

Specify `'Interactions'`

Pass the training data (`tbl`) and the name of the response variable in `tbl` to `fitrgam`, so that the function includes the linear terms for all the other variables as predictors. Specify the `'Interactions'` name-value argument using a logical matrix to include the two interaction terms, `x1*x2` and `x2*x3`.

```
Mdl2 = fitrgam(tbl,'MPG','Interactions',logical([1 1 0 0; 0 1 1 0]));
Mdl2.Interactions
```

```
ans = 2×2
```

```
     2     3
     1     2
```

You can also specify `'Interactions'` as the number of interaction terms or as `'all'` to include all available interaction terms. Among the specified interaction terms, `fitrgam` identifies those whose *p*-values are not greater than the `'MaxPValue'` value and adds them to the model. The default `'MaxPValue'` is 1 so that the function adds all specified interaction terms to the model.

Specify `'Interactions','all'` and set the `'MaxPValue'` name-value argument to 0.05.

```
Mdl3 = fitrgam(tbl,'MPG','Interactions','all','MaxPValue',0.05);
```

```
Warning: Model does not include interaction terms because all interaction terms have p-values gre
```

```
Mdl3.Interactions
```

```
ans =

  0x2 empty double matrix
```

`Mdl3` includes no interaction terms, which implies one of the following: all interaction terms have *p*-values greater than 0.05, or adding the interaction terms does not improve the model fit.

Use `addInteractions` Function

Train a univariate GAM that contains linear terms for predictors, and then add interaction terms to the trained model by using the `addInteractions` function. Specify the second input argument of `addInteractions` in the same way you specify the `'Interactions'` name-value argument of `fitrgam`. You can specify the list of interaction terms using a logical matrix, the number of interaction terms, or `'all'`.

Specify the number of interaction terms as 3 to add the three most important interaction terms to the trained model.

```
Mdl4 = fitrgam(tbl,'MPG');
UpdatedMdl4 = addInteractions(Mdl4,3);
UpdatedMdl4.Interactions
```

```
ans = 3×2

     2     3
     1     2
     3     4
```

`Mdl4` is a univariate GAM, and `UpdatedMdl4` is an updated GAM that contains all the terms in `Mdl4` and three additional interaction terms.

Resume Training Interaction Trees in GAM

Train a regression GAM that contains both linear and interaction terms. Specify to train the interaction terms for a small number of iterations. After training the interaction terms for more iterations, compare the resubstitution loss.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a GAM that includes all available linear and interaction terms in `X`. Specify the number of trees per interaction term as 2. `fitrgam` iterates the boosting algorithm 300 times (default) for linear terms, and iterates the algorithm the specified number of iterations for interaction terms. For each boosting iteration, the function adds one tree per linear term or one tree per interaction term. Specify `'Verbose'` as 1 to display diagnostic messages at every 10 iterations.

```
Mdl = fitrgam(X,Y,'Interactions','all','NumTreesPerInteraction',2,'Verbose',1);
```

```
|=========================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=========================================================|
|    1D|         0|  2.4432e+05|      -     |      -     |
|    1D|         1|      9507.4|         Inf|           1|
|    1D|        10|      4470.6|  0.00025206|           1|
|    1D|        20|      3895.3|  0.00011448|           1|
|    1D|        30|      3617.7|  3.5365e-05|           1|
|    1D|        40|      3402.5|  3.7992e-05|           1|
|    1D|        50|      3257.1|  2.4983e-05|           1|
|    1D|        60|      3131.8|  2.3873e-05|           1|
|    1D|        70|      3019.8|  2.2967e-05|           1|
|    1D|        80|      2925.9|  2.8071e-05|           1|
|    1D|        90|      2845.3|  1.6811e-05|           1|
|    1D|       100|      2772.7|   1.852e-05|           1|
|    1D|       110|      2707.8|  1.6754e-05|           1|
|    1D|       120|      2649.8|   1.651e-05|           1|
|    1D|       130|      2596.6|  1.1723e-05|           1|
|    1D|       140|      2547.4|   1.813e-05|           1|
|    1D|       150|      2501.1|  1.8659e-05|           1|
|    1D|       160|      2455.7|   1.386e-05|           1|
|    1D|       170|      2416.9|  1.0615e-05|           1|
|    1D|       180|      2377.2|   8.534e-06|           1|
|    1D|       190|        2339|  7.6771e-06|           1|
|    1D|       200|      2303.3|  9.5866e-06|           1|
|    1D|       210|      2270.7|  8.4276e-06|           1|
|    1D|       220|      2240.1|  8.5778e-06|           1|
|    1D|       230|      2209.2|  9.6761e-06|           1|
|    1D|       240|      2178.7|  7.0622e-06|           1|
|    1D|       250|      2150.3|  8.3082e-06|           1|
|    1D|       260|      2122.3|  7.9542e-06|           1|
|    1D|       270|      2097.7|  7.6328e-06|           1|
|    1D|       280|      2070.4|  9.4322e-06|           1|
|    1D|       290|      2044.3|  7.5722e-06|           1|
|    1D|       300|      2019.7|  6.6719e-06|           1|
|=========================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=========================================================|
|    2D|         0|      2019.7|      -     |      -     |
|    2D|         1|      1795.5|   0.0005975|           1|
|    2D|         2|      1523.4|   0.0010079|           1|
```

To check whether `fitrgam` trains the specified number of trees, display the `ReasonForTermination` property of the trained model and view the displayed messages.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Terminated after training the requested number of trees.'
```

Compute the regression loss for the training data.

```
resubLoss(Mdl)
```

```
ans = 3.8277
```

Resume training the model for another 100 iterations. Because `Mdl` contains both linear and interaction terms, the `resume` function resumes training for the interaction terms and adds more trees for them (interaction trees).

```
UpdatedMdl = resume(Mdl,100);
```

```
|============================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|============================================================|
|    2D|         0|      1523.4|           -|           -|
|    2D|         1|      1363.9|  0.00039695|           1|
|    2D|        10|      594.04|  8.0295e-05|           1|
|    2D|        20|      359.44|  4.3201e-05|           1|
|    2D|        30|      238.51|  2.6869e-05|           1|
|    2D|        40|      153.98|  2.6271e-05|           1|
|    2D|        50|      91.464|  8.0936e-06|           1|
|    2D|        60|      61.882|  3.8528e-06|           1|
|    2D|        70|      43.206|  5.9888e-06|           1|
```

```
UpdatedMdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Unable to improve the model fit.'
```

`resume` terminates training when adding more trees does not improve the deviance of the model fit.

Compute the regression loss using the updated model.

```
resubLoss(UpdatedMdl)
```

```
ans = 0.0944
```

The regression loss decreases after `resume` updates the model with more iterations.

More About

Generalized Additive Model (GAM) for Regression

A generalized additive model (GAM) is an interpretable model that explains a response variable using a sum of univariate and bivariate shape functions of predictors.

`fitrgam` uses a boosted tree as a shape function for each predictor and, optionally, each pair of predictors; therefore, the function can capture a nonlinear relation between a predictor and the response variable. Because contributions of individual shape functions to the prediction (response value) are well separated, the model is easy to interpret.

The standard GAM uses a univariate shape function for each predictor.

$$y \sim N\left(\mu, \sigma^2\right)$$
$$g(\mu) = \mu = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where y is a response variable that follows the normal distribution with mean μ and standard deviation σ. $g(\mu)$ is an identity link function, and c is an intercept (constant) term. $f_i(x_i)$ is a univariate shape function for the ith predictor, which is a boosted tree for a linear term for the predictor (predictor tree).

You can include interactions between predictors in a model by adding bivariate shape functions of important interaction terms to the model.

$$\mu = c + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \sum_{i,\, j \in \{1, 2, \cdots, p\}} f_{ij}(x_i x_j),$$

where $f_{ij}(x_i x_j)$ is a bivariate shape function for the ith and jth predictors, which is a boosted tree for an interaction term for the predictors (interaction tree).

`fitrgam` finds important interaction terms based on the p-values of F-tests. For details, see “Interaction Term Detection” on page 33-2135.

References

[1] Lou, Yin, Rich Caruana, and Johannes Gehrke. "Intelligible Models for Classification and Regression." *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’12).* Beijing, China: ACM Press, 2012, pp. 150–158.

[2] Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate Intelligible Models with Pairwise Interactions." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’13)* Chicago, Illinois, USA: ACM Press, 2013, pp. 623–631.

See Also

`CompactRegressionGAM` | `RegressionPartitionedGAM` | `addInteractions` | `fitrgam` | `resume`

Topics

“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

RegressionGP class

Superclasses: `CompactRegressionGP`

Gaussian process regression model class

Description

`RegressionGP` is a Gaussian process regression (GPR) model. You can train a GPR model, using `fitrgp`. Using the trained model, you can

- Predict responses for training data using `resubPredict` or new predictor data using `predict`. You can also compute the prediction intervals.
- Compute the regression loss for training data using `resubLoss` or new data using `loss`.

Construction

Create a `RegressionGP` object by using `fitrgp`.

Properties

Fitting

`FitMethod` — Method used to estimate the parameters

`'none'` | `'exact'` | `'sd'` | `'sr'` | `'fic'`

Method used to estimate the basis function coefficients, β; noise standard deviation, σ; and kernel parameters, θ, of the GPR model, stored as a character vector. It can be one of the following.

Fit Method	Description
`'none'`	No estimation. `fitrgp` uses the initial parameter values as the parameter values.
`'exact'`	Exact Gaussian process regression.
`'sd'`	Subset of data points approximation.
`'sr'`	Subset of regressors approximation.
`'fic'`	Fully independent conditional approximation.

`BasisFunction` — Explicit basis function

`'none'` | `'constant'` | `'linear'` | `'pureQuadratic'` | function handle

Explicit basis function used in the GPR model, stored as a character vector or a function handle. It can be one of the following. If n is the number of observations, the basis function adds the term $\boldsymbol{H}*\boldsymbol{\beta}$ to the model, where $\boldsymbol{H}$ is the basis matrix and $\boldsymbol{\beta}$ is a p-by-1 vector of basis coefficients.

Explicit Basis	Basis Matrix
`'none'`	Empty matrix.

Explicit Basis	Basis Matrix
`'constant'`	$H = 1$ (n-by-1 vector of 1s, where n is the number of observations)
`'linear'`	$H = [1, X]$
`'pureQuadratic'`	$H = [1, X, X_2]$, where $X_2 = \begin{bmatrix} x_{11}^2 & x_{12}^2 & \cdots & x_{1d}^2 \\ x_{21}^2 & x_{22}^2 & \cdots & x_{2d}^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1}^2 & x_{n2}^2 & \cdots & x_{nd}^2 \end{bmatrix}$.
Function handle	Function handle, `hfcn`, that `fitrgp` calls as: $H = hfcn(X)$, where X is an n-by-d matrix of predictors and H is an n-by-p matrix of basis functions.

Data Types: `char` | `function_handle`

`Beta` — Estimated coefficients
vector

Estimated coefficients for the explicit basis functions, stored as a vector. You can define the explicit basis function by using the `BasisFunction` name-value pair argument in `fitrgp`.

Data Types: `double`

`Sigma` — Estimated noise standard deviation
scalar value

Estimated noise standard deviation of the GPR model, stored as a scalar value.

Data Types: `double`

`CategoricalPredictors` — Indices of categorical predictors
vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table

This property is read-only.

Cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty if the

`'OptimizeHyperparameters'` name-value pair argument is nonempty when you create the model. The value of `HyperparameterOptimizationResults` depends on the setting of the `Optimizer` field in the `HyperparameterOptimizationOptions` structure when you create the model.

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`LogLikelihood` — Maximized marginal log likelihood
scalar value | `[]`

Maximized marginal log likelihood of the GPR model, stored as a scalar value if the `FitMethod` is different from `'none'`. If `FitMethod` is `'none'`, then `LogLikelihood` is empty.

If `FitMethod` is `'sd'`, `'sr'`, or `'fic'`, then `LogLikelihood` is the maximized approximation of the marginal log likelihood of the GPR model.

Data Types: `double`

`ModelParameters` — Parameters used for training
`GPParams` object

Parameters used for training the GPR model, stored as a `GPParams` object.

Kernel Function

`KernelFunction` — Form of the covariance function
`'squaredExponential'` | `'matern32'` | `'matern52'` | `'ardsquaredexponential'` | `'ardmatern32'` | `'ardmatern52'` | function handle

Form of the covariance function used in the GPR model, stored as a character vector containing the name of the built-in covariance function or a function handle. It can be one of the following.

Function	Description
`'squaredexponential'`	Squared exponential kernel.
`'matern32'`	Matern kernel with parameter 3/2.
`'matern52'`	Matern kernel with parameter 5/2.
`'ardsquaredexponential'`	Squared exponential kernel with a separate length scale per predictor.
`'ardmatern32'`	Matern kernel with parameter 3/2 and a separate length scale per predictor.
`'ardmatern52'`	Matern kernel with parameter 5/2 and a separate length scale per predictor.

Function	Description
Function handle	A function handle that `fitrgp` can call like this: `Kmn = kfcn(Xm,Xn,theta)` where `Xm` is an *m*-by-*d* matrix, `Xn` is an *n*-by-*d* matrix and `Kmn` is an *m*-by-*n* matrix of kernel products such that `Kmn`(*i*,*j*) is the kernel product between `Xm`(*i*,:) and `Xn`(*j*,:). `theta` is the *r*-by-1 unconstrained parameter vector for `kfcn`.

Data Types: `char` | `function_handle`

`KernelInformation` — Information about the parameters of the kernel function

structure

Information about the parameters of the kernel function used in the GPR model, stored as a structure with the following fields.

Field Name	Description
`Name`	Name of the kernel function
`KernelParameters`	Vector of the estimated kernel parameters
`KernelParameterNames`	Names associated with the elements of `KernelParameters`.

Data Types: `struct`

Prediction

`PredictMethod` — Method used to make predictions

`'exact'` | `'bcd'` | `'sd'` | `'sr'` | `'fic'`

Method that `predict` uses to make predictions from the GPR model, stored as a character vector. It can be one of the following.

`PredictMethod`	Description
`'exact'`	Exact Gaussian process regression
`'bcd'`	Block Coordinate Descent
`'sd'`	Subset of Data points approximation
`'sr'`	Subset of Regressors approximation
`'fic'`	Fully Independent Conditional approximation

`Alpha` — Weights

numeric vector

Weights used to make predictions from the trained GPR model, stored as a numeric vector. `predict` computes the predictions for a new predictor matrix `Xnew` by using the product

$$K(X_{new}, A) * \alpha \ .$$

$K(X_{new}, A)$ is the matrix of kernel products between X_{new} and active set vector $\boldsymbol{A}$ and α is a vector of weights.

Data Types: `double`

`BCDInformation` — Information on BCD-based computation of `Alpha`
structure | `[]`

Information on block coordinate descent (BCD)-based computation of `Alpha` when `PredictMethod` is `'bcd'`, stored as a structure containing the following fields.

Field Name	Description
`Gradient`	*n*-by-1 vector containing the gradient of the BCD objective function at convergence.
`Objective`	Scalar containing the BCD objective function at convergence.
`SelectionCounts`	*n*-by-1 integer vector indicating the number of times each point was selected into a block during BCD.

`Alpha` property contains the `Alpha` vector computed from BCD.

If `PredictMethod` is not `'bcd'`, then `BCDInformation` is empty.

Data Types: `struct`

`ResponseTransform` — Transformation applied to predicted response
`'none'` (default)

Transformation applied to the predicted response, stored as a character vector describing how the response values predicted by the model are transformed. In `RegressionGP`, `ResponseTransform` is `'none'` by default, and `RegressionGP` does not use `ResponseTransform` when making predictions.

Active Set Selection

`ActiveSetVectors` — Subset of training data
matrix

Subset of training data used to make predictions from the GPR model, stored as a matrix.

`predict` computes the predictions for a new predictor matrix `Xnew` by using the product

$$K(X_{new}, A) * \alpha \ .$$

$K(X_{new}, A)$ is the matrix of kernel products between X_{new} and active set vector $\boldsymbol{A}$ and α is a vector of weights.

`ActiveSetVectors` is equal to the training data `X` for exact GPR fitting and a subset of the training data `X` for sparse GPR methods. When there are categorical predictors in the model, `ActiveSetVectors` contains dummy variables for the corresponding predictors.

Data Types: `double`

`ActiveSetHistory` — History of active set selection and parameter estimation
structure

History of interleaved active set selection and parameter estimation on page 33-5323 for `FitMethod` equal to `'sd'`, `'sr'`, or `'fic'`, stored as a structure with the following fields.

Field Name	Description
`ParameterVector`	Cell array containing the parameter vectors: basis function coefficients, β, kernel function parameters θ, and noise standard deviation σ.
`ActiveSetIndices`	Cell array containing the active set indices.
`Loglikelihood`	Vector containing the maximized log likelihoods.
`CriterionProfile`	Cell array containing the active set selection criterion values as the active set grows from size 0 to its final size.

Data Types: `struct`

`ActiveSetMethod` — Method used to select the active set
`'sgma'` | `'entropy'` | `'likelihood'` | `'random'`

Method used to select the active set for sparse methods (`'sd'`,`'sr'`, or `'fic'`), stored as a character vector. It can be one of the following.

`ActiveSetMethod`	Description
`'sgma'`	Sparse greedy matrix approximation
`'entropy'`	Differential entropy-based selection
`'likelihood'`	Subset of regressors log likelihood-based selection
`'random'`	Random selection

The selected active set is used in parameter estimation or prediction, depending on the choice of `FitMethod` and `PredictMethod` in `fitrgp`.

`ActiveSetSize` — Size of the active set
integer value

Size of the active set for sparse methods (`'sd'`,`'sr'`, or `'fic'`), stored as an integer value.

Data Types: `double`

`IsActiveSetVector` — Indicators for selected active set
logical vector

Indicators for selected active set for making predictions from the trained GPR model, stored as a logical vector. These indicators mark the subset of training data that `fitrgp` selects as the active set. For example, if `X` is the original training data, then `ActiveSetVectors = X(IsActiveSetVector,:)`.

Data Types: `logical`

Training Data

`NumObservations` — Number of observations in training data
scalar value

Number of observations in training data, stored as a scalar value.

Data Types: `double`

X — Training data
n-by-*d* table | *n*-by-*d* matrix

Training data, stored as an *n*-by-*d* table or matrix, where *n* is the number of observations and *d* is the number of predictor variables (columns) in the training data. If the GPR model is trained on a table, then `X` is a table. Otherwise, `X` is a matrix.

Data Types: `double` | `table`

Y — Observed response values
n-by-1 vector

Observed response values used to train the GPR model, stored as an *n*-by-1 vector, where *n* is the number of observations.

Data Types: `double`

`PredictorNames` — Names of predictors
cell array of character vectors

Names of predictors used in the GPR model, stored as a cell array of character vectors. Each name (cell) corresponds to a column in `X`.

Data Types: `cell`

`ExpandedPredictorNames` — Names of expanded predictors
cell array of character vectors

Names of expanded predictors for the GPR model, stored as a cell array of character vectors. Each name (cell) corresponds to a column in `ActiveSetVectors`.

If the model uses dummy variables for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ResponseName` — Name of the response variable
character vector

Name of the response variable in the GPR model, stored as a character vector.

Data Types: `char`

`PredictorLocation` — Means of predictors
1-by-*d* vector | `[]`

Means of predictors used for training the GPR model if the training data is standardized, stored as a 1-by-*d* vector. If the training data is not standardized, `PredictorLocation` is empty.

If `PredictorLocation` is not empty, then the `predict` method centers the predictor values by subtracting the respective element of `PredictorLocation` from every column of `X`.

If there are categorical predictors, then `PredictorLocation` includes a 0 for each dummy variable corresponding to those predictors. The dummy variables are not centered or scaled.

Data Types: `double`

`PredictorScale` — Standard deviations of predictors
1-by-*d* vector | `[ ]`

Standard deviations of predictors used for training the GPR model if the training data is standardized, stored as a 1-by-*d* vector. If the training data is not standardized, `PredictorScale` is empty.

If `PredictorScale` is not empty, the `predict` method scales the predictors by dividing every column of `X` by the respective element of `PredictorScale` (after centering using `PredictorLocation`).

If there are categorical predictors, then `PredictorLocation` includes a 1 for each dummy variable corresponding to those predictors. The dummy variables are not centered or scaled.

Data Types: `double`

`RowsUsed` — Indicators for rows used in training
logical vector | `[ ]`

Indicators for rows used in training the GPR model, stored as a logical vector. If all rows are used in training the model, then `RowsUsed` is empty.

Data Types: `logical`

Object Functions

compact	Create compact Gaussian process regression model
crossval	Cross-validate Gaussian process regression model
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error for Gaussian process regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
postFitStatistics	Compute post-fit statistics for the exact Gaussian process regression model
predict	Predict response of Gaussian process regression model
resubLoss	Resubstitution loss for a trained Gaussian process regression model
resubPredict	Resubstitution prediction from a trained Gaussian process regression model
shapley	Shapley values

More About

Active Set Selection and Parameter Estimation

For subset of data, subset of regressors, or fully independent conditional approximation fitting methods (`FitMethod` equal to `'sd'`, `'sr'`, or `'fic'`), if you do not provide the active set, `fitrgp` selects the active set and computes the parameter estimates in a series of iterations.

In the first iteration, the software uses the initial parameter values in vector $\eta_0 = [\beta_0,\sigma_0,\theta_0]$ to select an active set A_1. It maximizes the GPR marginal log likelihood or its approximation using η_0 as the initial values and A_1 to compute the new parameter estimates η_1. Next, it computes the new log likelihood L_1 using η_1 and A_1.

In the second iteration, the software selects the active set A_2 using the parameter values in η_1. Then, using η_1 as the initial values and A_2, it maximizes the GPR marginal log likelihood or its approximation and estimates the new parameter values η_2. Then using η_2 and A_2, computes the new log likelihood value L_2.

The following table summarizes the iterations and what is computed at each iteration.

Iteration Number	Active Set	Parameter Vector	Log Likelihood
1	A_1	η_1	L_1
2	A_2	η_2	L_2
3	A_3	η_3	L_3
…	…	…	…

The software iterates similarly for a specified number of repetitions. You can specify the number of replications for active set selection using the `NumActiveSetRepeats` name-value pair argument.

Tips

- You can access the properties of this class using dot notation. For example, `KernelInformation` is a structure holding the kernel parameters and their names. Hence, to access the kernel function parameters of the trained model `gprMdl`, use `gprMdl.KernelInformation.KernelParameters`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` function supports code generation.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`CompactRegressionGP` | `compact` | `fitrgp`

Topics

Class Attributes
Property Attributes

Introduced in R2015b

RegressionLinear class

Linear regression model for high-dimensional data

Description

`RegressionLinear` is a trained linear model object for regression; the linear model is a support vector machine regression (SVM) or linear regression model. `fitrlinear` fits a `RegressionLinear` model by minimizing the objective function using techniques that reduce computation time for high-dimensional data sets (e.g., stochastic gradient descent). The regression loss plus the regularization term compose the objective function.

Unlike other regression models, and for economical memory usage, `RegressionLinear` model objects do not store the training data. However, they do store, for example, the estimated linear model coefficients, estimated coefficients, and the regularization strength.

You can use trained `RegressionLinear` models to predict responses for new data. For details, see `predict`.

Construction

Create a `RegressionLinear` object by using `fitrlinear`.

Properties

Linear Regression Properties

`Epsilon` — Half of width of epsilon-insensitive band

nonnegative scalar

Half of the width of the epsilon-insensitive-band, specified as a nonnegative scalar.

If `Learner` is not `'svm'`, then `Epsilon` is an empty array (`[]`).

Data Types: `single` | `double`

`Lambda` — Regularization term strength

nonnegative scalar | vector of nonnegative values

Regularization term strength, specified as a nonnegative scalar or vector of nonnegative values.

Data Types: `double` | `single`

`Learner` — Linear regression model type

`'leastsquares'` | `'svm'`

Linear regression model type, specified as `'leastsquares'` or `'svm'`.

In this table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.

- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Loss function	FittedLoss Value
'leastsquares'	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	'mse'
'svm'	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	'epsiloninsensitive'

Beta — Linear coefficient estimates
numeric vector

Linear coefficient estimates, specified as a numeric vector with length equal to the number of predictors.

Data Types: double

Bias — Estimated bias term
numeric scalar

Estimated bias term or model intercept, specified as a numeric scalar.

Data Types: double

FittedLoss — Loss function used to fit the linear model
'epsiloninsensitive' | 'mse'

Loss function used to fit the model, specified as 'epsiloninsensitive' or 'mse'.

Value	Algorithm	Loss function	Learner Value
'epsiloninsensitive'	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	'svm'
'mse'	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	'leastsquares'

Regularization — Complexity penalty type
'lasso (L1)' | 'ridge (L2)'

Complexity penalty type, specified as 'lasso (L1)' or 'ridge (L2)'.

The software composes the objective function for minimization from the sum of the average loss function (see FittedLoss) and a regularization value from this table.

Value	Description
`'lasso (L1)'`	Lasso (L_1) penalty: $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$
`'ridge (L2)'`	Ridge (L_2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

λ specifies the regularization term strength (see `Lambda`).

The software excludes the bias term (β_0) from the regularization penalty.

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ModelParameters` — Parameters used for training model

structure

Parameters used for training the `RegressionLinear` model, specified as a structure.

Access fields of `ModelParameters` using dot notation. For example, access the relative tolerance on the linear coefficients and the bias term by using `Mdl.ModelParameters.BetaTolerance`.

Data Types: `struct`

`PredictorNames` — Predictor names

cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`ExpandedPredictorNames` — Expanded predictor names

cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ResponseName` — Response variable name

character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @function`, where `function` accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

Object Functions

incrementalLearner	Convert linear regression model to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression loss for linear regression models
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict response of linear regression model
selectModels	Select fitted regularized linear regression models
shapley	Shapley values
update	Update model parameters for code generation

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Linear Regression Model

Train a linear regression model using SVM, dual SGD, and ridge regularization.

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model. By default, `fitrlinear` uses support vector machines with a ridge penalty, and optimizes using dual SGD for SVM. Determine how well the optimization algorithm fit the model to the data by extracting a fit summary.

```
[Mdl,FitInfo] = fitrlinear(X,Y)
```

```
Mdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0056
               Lambda: 1.0000e-04
              Learner: 'svm'


  Properties, Methods


FitInfo = struct with fields:
                    Lambda: 1.0000e-04
                 Objective: 0.2725
                 PassLimit: 10
                 NumPasses: 10
                BatchLimit: []
             NumIterations: 100000
              GradientNorm: NaN
         GradientTolerance: 0
       RelativeChangeInBeta: 0.4907
             BetaTolerance: 1.0000e-04
             DeltaGradient: 1.5816
    DeltaGradientTolerance: 0
           TerminationCode: 0
         TerminationStatus: {'Iteration limit exceeded.'}
                     Alpha: [10000x1 double]
                   History: []
                   FitTime: 0.1318
                    Solver: {'dual'}
```

`Mdl` is a `RegressionLinear` model. You can pass `Mdl` and the training or new data to `loss` to inspect the in-sample mean-squared error. Or, you can pass `Mdl` and new predictor data to `predict` to predict responses for new observations.

`FitInfo` is a structure array containing, among other things, the termination status (`TerminationStatus`) and how long the solver took to fit the model to the data (`FitTime`). It is good practice to use `FitInfo` to determine whether optimization-termination measurements are satisfactory. In this case, `fitrlinear` reached the maximum number of iterations. Because training time is fast, you can retrain the model, but increase the number of passes through the data. Or, try another solver, such as LBFGS.

Predict Responses Using Linear Regression Model

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Hold out 5% of the data.

```
rng(1); % For reproducibility
cvp = cvpartition(n,'Holdout',0.05)
```

```
cvp =
Hold-out cross validation partition
   NumObservations: 10000
       NumTestSets: 1
         TrainSize: 9500
          TestSize: 500
```

`cvp` is a `CVPartition` object that defines the random partition of n data into training and test sets.

Train a linear regression model using the training set. For faster training time, orient the predictor data matrix so that the observations are in columns.

```
idxTrain = training(cvp); % Extract training set indices
X = X';
Mdl = fitrlinear(X(:,idxTrain),Y(idxTrain),'ObservationsIn','columns');
```

Predict observations and the mean squared error (MSE) for the hold out sample.

```
idxTest = test(cvp); % Extract test set indices
yHat = predict(Mdl,X(:,idxTest),'ObservationsIn','columns');
L = loss(Mdl,X(:,idxTest),Y(idxTest),'ObservationsIn','columns')
```

```
L = 0.1851
```

The hold-out sample MSE is 0.1852.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- When you train a linear regression model by using `fitrlinear`, the following restrictions apply.
  - If the predictor data input argument value is a matrix, it must be a full, numeric matrix. Code generation does not support sparse data.

- You can specify only one regularization strength, either `'auto'` or a nonnegative scalar for the `'Lambda'` name-value pair argument.
- The value of the `'ResponseTransform'` name-value pair argument cannot be an anonymous function.
- Code generation with a coder configurer does not support categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`RegressionPartitionedLinear` | `fitrlinear` | `predict`

Introduced in R2016a

RegressionLinearCoderConfigurer

Coder configurer for linear regression model with high-dimensional data

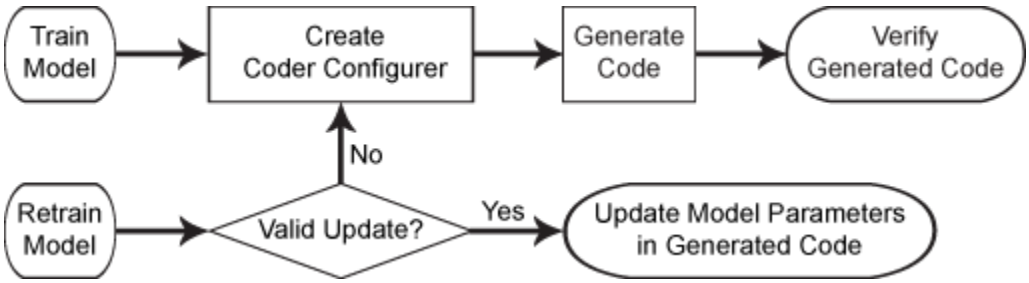
Description

A `RegressionLinearCoderConfigurer` object is a coder configurer of a linear regression model (`RegressionLinear`) with high-dimensional data.

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes for linear model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the linear regression model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the linear model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of a linear regression model, see the Code Generation sections of `RegressionLinear`, `predict`, and `update`.

Creation

After training a linear regression model by using `fitrlinear`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of the `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data

`LearnerCoderInput` object

Coder attributes of the predictor data to pass to the generated C/C++ code for the `predict` function of the linear regression model, specified as a `LearnerCoderInput` on page 33-5343 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
  - If the `Value` attribute of the `ObservationsIn` property for the `RegressionLinearCoderConfigurer` is `'rows'`, then this `SizeVector` value is `[n p]`, where `n` corresponds to the number of observations and `p` corresponds to the number of predictors.
  - If the `Value` attribute of the `ObservationsIn` property for the `RegressionLinearCoderConfigurer` is `'columns'`, then this `SizeVector` value is `[p n]`.

  To switch the elements of `SizeVector` (for example, to change `[n p]` to `[p n]`), modify the `Value` attribute of the `ObservationsIn` property for the `RegressionLinearCoderConfigurer` accordingly. You cannot modify the `SizeVector` value directly.
- `VariableDimensions` — The default value is `[0 0]`, which indicates that the array size is fixed as specified in `SizeVector`.

  You can set this value to `[1 0]` if the `SizeVector` value is `[n p]` or to `[0 1]` if it is `[p n]`, which indicates that the array has variable-size rows and fixed-size columns. For example, `[1 0]` specifies that the first value of `SizeVector` (`n`) is the upper bound for the number of rows, and the second value of `SizeVector` (`p`) is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations (in rows) of three predictor variables (in columns), specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of

observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

`ObservationsIn` — Coder attributes of predictor data observation dimension
`EnumeratedInput` object

Coder attributes of the predictor data observation dimension (`'ObservationsIn'` name-value pair argument of `predict`), specified as an `EnumeratedInput` on page 33-5343 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the `'ObservationsIn'` name-value pair argument determines the default values of the `EnumeratedInput` coder attributes:

- `Value` — The default value is the predictor data observation dimension you use when creating the coder configurer, specified as `'rows'` or `'columns'`. If you do not specify `'ObservationsIn'` when creating the coder configurer, the default value is `'rows'`.
- `SelectedOption` — This value is always `'Built-in'`. This attribute is read-only.
- `BuiltInOptions` — Cell array of `'rows'` and `'columns'`. This attribute is read-only.
- `IsConstant` — This value must be `true`.
- `Tunability` — The default value is `false` if you specify `'ObservationsIn','rows'` when creating the coder configurer, and `true` if you specify `'ObservationsIn','columns'`. If you set `Tunability` to `false`, the software sets `Value` to `'rows'`. If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

`NumOutputs` — Number of outputs in `predict`
1 (default)

Number of output arguments to return from the generated C/C++ code for the `predict` function of the linear regression model, specified as 1. `predict` returns `YHat` (predicted responses) in the generated C/C++ code.

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of a linear regression model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-5343 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Beta` — Coder attributes of linear predictor coefficients
`LearnerCoderInput` object

Coder attributes of the linear predictor coefficients (`Beta` of a linear regression model), specified as a `LearnerCoderInput` on page 33-5343 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[p 1]`, where `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

`Bias` — Coder attributes of bias term
`LearnerCoderInput` object

Coder attributes of the bias term (`Bias` of a linear regression model), specified as a `LearnerCoderInput` on page 33-5343 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code
`'RegressionLinearModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `RegressionLinearCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level
`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — List of tunable input arguments of `predict`
cell array

This property is read-only.

List of tunable input arguments of the entry-point function `predict.m` for code generation, specified as a cell array. The cell array contains another cell array that includes `coder.PrimitiveType` objects and `coder.Constant` objects.

If you modify the coder attributes of `predict` arguments on page 33-5332, then the software updates the corresponding objects accordingly. If you specify the `Tunability` attribute as `false`, then the software removes the corresponding objects from the `PredictInputs` list.

The cell array in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each `coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-5334 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Simulate 10,000 observations from the model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10,000-by-1000 numeric matrix with standard normal elements.
- e is a random normal error with mean 0 and standard deviation 0.3.

```
rng('default') % For reproducibility
n = 10000;
p = 1000;
X = randn(n,p);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model using the simulated data. Pass the transposed predictor matrix `Xnew` to `fitrlinear`, and use the `'ObservationsIn'` name-value pair argument to specify that the columns of `Xnew` correspond to observations.

```
Xnew = X';
Mdl = fitrlinear(Xnew,Y,'ObservationsIn','columns');
```

`Mdl` is a `RegressionLinear` object.

Create a coder configurer for the `RegressionLinear` model by using `learnerCoderConfigurer`. Specify the predictor data `Xnew`, and use the `'ObservationsIn'` name-value pair argument to specify the observation dimension of `Xnew`. The `learnerCoderConfigurer` function uses these input arguments to configure the coder attributes of the corresponding input arguments of `predict`.

```
configurer = learnerCoderConfigurer(Mdl,Xnew,'ObservationsIn','columns')
```

```
configurer =
  RegressionLinearCoderConfigurer with properties:

   Update Inputs:
              Beta: [1x1 LearnerCoderInput]
              Bias: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
    ObservationsIn: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'RegressionLinearModel'


  Properties, Methods
```

`configurer` is a `RegressionLinearCoderConfigurer` object, which is a coder configurer of a `RegressionLinear` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the linear regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionLinearModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionLinearModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionLinearModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:17
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:17
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:17
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('RegressionLinearModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Beta
        %                       Bias
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X, ObservationsIn
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of Linear Regression Model in Generated Code

Train a linear regression model using a partial data set, and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the linear regression model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Simulate 10,000 observations from the model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = x_1, \ldots, x_{1000}$ is a 10,000-by-1000 numeric matrix with standard normal elements.
- e is a random normal error with mean 0 and standard deviation 0.3.

```
rng('default') % For reproducibility
n = 10000;
p = 1000;
X = randn(n,p);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Train a linear regression model using the first 500 observations. Transpose the predictor data, and use the `'ObservationsIn'` name-value pair argument to specify that the columns of `XTrain` correspond to observations.

```
XTrain = X(1:500,:)';
YTrain = Y(1:500);
Mdl = fitrlinear(XTrain,YTrain,'ObservationsIn','columns');
```

`Mdl` is a `RegressionLinear` object.

Create Coder Configurer

Create a coder configurer for the `RegressionLinear` model by using `learnerCoderConfigurer`. Specify the predictor data `XTrain`, and use the `'ObservationsIn'` name-value pair argument to specify the observation dimension of `XTrain`. The `learnerCoderConfigurer` function uses these input arguments to configure the coder attributes of the corresponding input arguments of `predict`.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'ObservationsIn','columns');
```

`configurer` is a `RegressionLinearCoderConfigurer` object, which is a coder configurer of a `RegressionLinear` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the linear regression model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of the predictor data that you want to pass to the generated code.

Specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The

`SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [1000 Inf];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   0   1
```

The size of the first dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 1000 predictors, the value of the `SizeVector` attribute must be 1000 and the value of the `VariableDimensions` attribute must be `0`.

The size of the second dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The order of the dimensions in `SizeVector` and `VariableDimensions` depends on the coder attributes of `ObservationsIn`.

```
configurer.ObservationsIn
```

```
ans =
  EnumeratedInput with properties:

             Value: 'columns'
    SelectedOption: 'Built-in'
    BuiltInOptions: {'rows'  'columns'}
        IsConstant: 1
        Tunability: 1
```

When the `Value` attribute of the `ObservationsIn` property is `'columns'`, the first dimension of the `SizeVector` and `VariableDimensions` attributes of `X` corresponds to the number of predictors, and the second dimension corresponds to the number of observations. When the `Value` attribute of `ObservationsIn` is `'rows'`, the order of the dimensions is switched.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the linear regression model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionLinearModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionLinearModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionLinearModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
YHat = predict(Mdl,XTrain,'ObservationsIn','columns');
YHat_mex = RegressionLinearModel('predict',XTrain,'ObservationsIn','columns');
```

Compare `YHat` and `YHat_mex`.

```
max(abs(YHat-YHat_mex))
```

```
ans = 0
```

In general, `YHat_mex` might include round-off differences compared to `YHat`. In this case, the comparison confirms that `YHat` and `YHat_mex` are equal.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrlinear(X',Y,'ObservationsIn','columns');
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionLinearModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
YHat = predict(retrainedMdl,X','ObservationsIn','columns');
YHat_mex = RegressionLinearModel('predict',X','ObservationsIn','columns');
max(abs(YHat-YHat_mex))
```

```
ans = 0
```

The comparison confirms that `YHat` and `YHat_mex` are equal.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the data type of the bias term `Bias` of the coder configurer `configurer`:

```
configurer.Bias.DataType = 'single';
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

`EnumeratedInput` Object

A coder configurer uses an `EnumeratedInput` object to specify the coder attributes of `predict` input arguments that have a finite set of available values.

An `EnumeratedInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description

`Value`	Value of the `predict` argument in the generated code, specified as a character vector or a `LearnerCoderInput` on page 33-5343 object.
	• Character vector in `BuiltInOptions` — You can specify one of the `BuiltInOptions` using either the option name or its index value. For example, to choose the first option, specify `Value` as either the first character vector in `BuiltInOptions` or `1`.
	• Character vector designating a custom function name — To use a custom option, define a custom function on the MATLAB search path, and specify `Value` as the name of the custom function.
	• `LearnerCoderInput` on page 33-5343 object — If you set `IsConstant` to `false` (logical 0), then the software changes `Value` to a `LearnerCoderInput` on page 33-5343 object with the following read-only coder attribute values. These values indicate that the input in the generated code is a variable-size, tunable character vector that is one of the available values in `BuiltInOptions`.
	  • `SizeVector` — `[1 c]`, indicating the upper bound of the array size, where `c` is the length of the longest available character vector in `Option`
	  • `VariableDimensions` — `[0 1]`, indicating that the array is a variable-size vector
	  • `DataType` — `'char'`
	  • `Tunability` — 1
	The default value of `Value` is consistent with the default value of the corresponding `predict` argument, which is one of the character vectors in `BuiltInOptions`.

`SelectedOption`	Status of the selected option, specified as `'Built-in'`, `'Custom'`, or `'NonConstant'`. The software sets `SelectedOption` according to `Value`: • `'Built-in'`(default) — When `Value` is one of the character vectors in `BuiltInOptions` • `'Custom'` — When `Value` is a character vector that is not in `BuiltInOptions` • `'NonConstant'` — When `Value` is a `LearnerCoderInput` on page 33-5343 object This attribute is read-only.
`BuiltInOptions`	List of available character vectors for the corresponding `predict` argument, specified as a cell array. This attribute is read-only.

Attribute Name	Description
`IsConstant`	Indicator specifying whether or not the array value is a compile-time constant (`coder.Constant`) in the generated code, specified as `true` (logical 1, default) or `false` (logical 0). If you set this value to `false`, then the software changes `Value` to a `LearnerCoderInput` on page 33-5343 object.
`Tunability`	Indicator specifying whether or not `predict` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0, default). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of `ObservationsIn` of the coder configurer `configurer`:

```
configurer.ObservationsIn.Value = 'columns';
```

See Also

`RegressionLinear` | `learnerCoderConfigurer` | `predict` | `update`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2019b

RegressionNeuralNetwork

Neural network model for regression

Description

A `RegressionNeuralNetwork` object is a trained, feedforward, and fully connected neural network for regression. The first fully connected layer of the neural network has a connection from the network input (predictor data `X`), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix (`LayerWeights`) and then adds a bias vector (`LayerBiases`). An activation function follows each fully connected layer, excluding the last (`Activations` and `OutputLayerActivation`). The final fully connected layer produces the network's output, namely predicted response values. For more information, see “Neural Network Structure” on page 33-2239.

Creation

Create a `RegressionNeuralNetwork` object by using `fitrnet`.

Properties

Neural Network Properties

`LayerSizes` — Sizes of fully connected layers
positive integer vector

This property is read-only.

Sizes of the fully connected layers in the neural network model, returned as a positive integer vector. The *i*th element of `LayerSizes` is the number of outputs in the *i*th fully connected layer of the neural network model.

`LayerSizes` does not include the size of the final fully connected layer. This layer always has one output.

Data Types: `single` | `double`

`LayerWeights` — Learned layer weights
cell array

This property is read-only.

Learned layer weights for fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer weights for the *i*th fully connected layer. For example, `Mdl.LayerWeights{1}` returns the weights for the first fully connected layer of the model `Mdl`.

`LayerWeights` includes the weights for the final fully connected layer.

Data Types: `cell`

`LayerBiases` — Learned layer biases

cell array

This property is read-only.

Learned layer biases for fully connected layers, returned as a cell array. The *i*th entry in the cell array corresponds to the layer biases for the *i*th fully connected layer. For example, `Mdl.LayerBiases{1}` returns the biases for the first fully connected layer of the model `Mdl`.

`LayerBiases` includes the biases for the final fully connected layer.

Data Types: `cell`

`Activations` — Activation functions for fully connected layers

`'relu'` | `'tanh'` | `'sigmoid'` | `'none'` | cell array of character vectors

This property is read-only.

Activation functions for the fully connected layers of the neural network model, returned as a character vector or cell array of character vectors with values from this table.

Value	Description
`'relu'`	Rectified linear unit (ReLU) function — Performs a threshold operation on each element of the input, where any value less than zero is set to zero, that is, $$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$
`'tanh'`	Hyperbolic tangent (tanh) function — Applies the `tanh` function to each input element
`'sigmoid'`	Sigmoid function — Performs the following operation on each input element: $$f(x) = \frac{1}{1 + e^{-x}}$$
`'none'`	Identity function — Returns each input element without performing any transformation, that is, $f(x) = x$

- If `Activations` contains only one activation function, then it is the activation function for every fully connected layer of the neural network model, excluding the final fully connected layer, which does not have an activation function (`OutputLayerActivation`).
- If `Activations` is an array of activation functions, then the *i*th element is the activation function for the *i*th layer of the neural network model.

Data Types: `char` | `cell`

`OutputLayerActivation` — Activation function for final fully connected layer

`'none'`

This property is read-only.

Activation function for final fully connected layer, returned as `'none'`.

`ModelParameters` — Parameter values used to train model

`NeuralNetworkParams` object

This property is read-only.

Parameter values used to train the `RegressionNeuralNetwork` model, returned as a `NeuralNetworkParams` object. `ModelParameters` contains parameter values such as the name-value arguments used to train the regression neural network model.

Access the properties of `ModelParameters` by using dot notation. For example, access the function used to initialize the fully connected layer weights of a model `Mdl` by using `Mdl.ModelParameters.LayerWeightsInitializer`.

Convergence Control Properties

`ConvergenceInfo` — Convergence information

structure array

This property is read-only.

Convergence information, returned as a structure array.

Field	Description
`Iterations`	Number of training iterations used to train the neural network model
`TrainingLoss`	Training mean squared error (MSE) for the returned model, or `resubLoss(Mdl)` for model `Mdl`
`Gradient`	Gradient of the loss function with respect to the weights and biases at the iteration corresponding to the returned model
`Step`	Step size at the iteration corresponding to the returned model
`Time`	Total time spent across all iterations (in seconds)
`ValidationLoss`	Validation MSE for the returned model
`ValidationChecks`	Maximum number of times in a row that the validation loss was greater than or equal to the minimum validation loss
`ConvergenceCriterion`	Criterion for convergence
`History`	See `TrainingHistory`

Data Types: `struct`

`TrainingHistory` — Training history

table

This property is read-only.

Training history, returned as a table.

Column	Description
`Iteration`	Training iteration
`TrainingLoss`	Training mean squared error (MSE) for the model at this iteration
`Gradient`	Gradient of the loss function with respect to the weights and biases at this iteration
`Step`	Step size at this iteration
`Time`	Time spent during this iteration (in seconds)
`ValidationLoss`	Validation MSE for the model at this iteration
`ValidationChecks`	Running total of times that the validation loss is greater than or equal to the minimum validation loss

Data Types: `table`

`Solver` — Solver used to train neural network model
`'LBFGS'`

This property is read-only.

Solver used to train the neural network model, returned as `'LBFGS'`. To create a `RegressionNeuralNetwork` model, `fitrnet` uses a limited-memory Broyden-Flecter-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) as its loss function minimization technique, where the software minimizes the mean squared error (MSE).

Predictor Properties

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, returned as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, returned as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

This property is read-only.

Expanded predictor names, returned as a cell array of character vectors. If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`X` — Unstandardized predictors
numeric matrix | table

This property is read-only.

Unstandardized predictors used to train the neural network model, returned as a numeric matrix or table. `X` retains its original orientation, with observations in rows or columns depending on the value of the `ObservationsIn` name-value argument in the call to `fitrnet`.

Data Types: `single` | `double` | `table`

Response Properties

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, returned as a character vector.

Data Types: `char`

`Y` — Response values
numeric vector

This property is read-only.

Response values used to train the model, returned as a numeric vector. Each row of `Y` represents the response value of the corresponding observation in `X`.

Data Types: `single` | `double`

`ResponseTransform` — Response transformation function
`'none'`

This property is read-only.

Response transformation function, returned as `'none'`. The software does not transform the raw response values.

Other Data Properties

`NumObservations` — Number of observations
positive numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, returned as a positive numeric scalar.

Data Types: `double`

RowsUsed — Rows used in fitting

[] | logical vector

This property is read-only.

Rows of the original training data used in fitting the model, returned as a logical vector. This property is empty if all rows are used.

Data Types: `logical`

W — Observation weights

numeric vector

This property is read-only.

Observation weights used to train the model, returned as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `Weights` name-value argument so that the elements of `W` sum up to 1.

Data Types: `single` | `double`

Object Functions

compact	Reduce size of machine learning model
crossval	Cross-validate machine learning model
loss	Loss for regression neural network
resubLoss	Resubstitution regression loss
resubPredict	Predict responses for training data using trained regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using regression neural network

Examples

Train Neural Network Regression Model

Train a neural network regression model, and assess the performance of the model on a test set.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a table containing the predictor variables `Acceleration`, `Displacement`, and so on, as well as the response variable `MPG`.

```
load carbig
cars = table(Acceleration,Displacement,Horsepower, ...
    Model_Year,Origin,Weight,MPG);
```

Partition the data into training and test sets. Use approximately 80% of the observations to train a neural network model, and 20% of the observations to test the performance of the trained model on new data. Use `cvpartition` to partition the data.

```
rng("default") % For reproducibility of the data partition
c = cvpartition(length(MPG),"Holdout",0.20);
```

```
trainingIdx = training(c); % Training set indices
carsTrain = cars(trainingIdx,:);
testIdx = test(c); % Test set indices
carsTest = cars(testIdx,:);
```

Train a neural network regression model by passing the `carsTrain` training data to the `fitrnet` function. For better results, specify to standardize the predictor data.

```
Mdl = fitrnet(carsTrain,"MPG","Standardize",true)
```

```
Mdl =
  RegressionNeuralNetwork
           PredictorNames: {'Acceleration'  'Displacement'  'Horsepower'  'Model_Year'  'Origin'
             ResponseName: 'MPG'
    CategoricalPredictors: 5
        ResponseTransform: 'none'
          NumObservations: 314
               LayerSizes: 10
              Activations: 'relu'
    OutputLayerActivation: 'linear'
                   Solver: 'LBFGS'
         ConvergenceInfo: [1×1 struct]
         TrainingHistory: [1000×7 table]


  Properties, Methods
```

`Mdl` is a trained `RegressionNeuralNetwork` model. You can use dot notation to access the properties of `Mdl`. For example, you can specify `Mdl.TrainingHistory` to get more information about the training history of the neural network model.

Evaluate the performance of the regression model on the test set by computing the test mean squared error (MSE). Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,carsTest,"MPG")
```

```
testMSE = 16.6154
```

Specify Neural Network Regression Model Architecture

Specify the structure of the neural network regression model, including the size of the fully connected layers.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s. Create a matrix `X` containing the predictor variables `Acceleration`, `Cylinders`, and so on. Store the response variable `MPG` in the variable `Y`.

```
load carbig
X = [Acceleration Cylinders Displacement Weight];
Y = MPG;
```

Partition the data into training data (`XTrain` and `YTrain`) and test data (`XTest` and `YTest`). Reserve approximately 20% of the observations for testing, and use the rest of the observations for training.

```
rng("default") % For reproducibility of the partition
c = cvpartition(length(Y),"Holdout",0.20);
trainingIdx = training(c); % Indices for the training set
XTrain = X(trainingIdx,:);
YTrain = Y(trainingIdx);
testIdx = test(c); % Indices for the test set
XTest = X(testIdx,:);
YTest = Y(testIdx);
```

Train a neural network regression model. Specify to standardize the predictor data, and to have 30 outputs in the first fully connected layer and 10 outputs in the second fully connected layer. By default, both layers use a rectified linear unit (ReLU) activation function. You can change the activation functions for the fully connected layers by using the `Activations` name-value argument.

```
Mdl = fitrnet(XTrain,YTrain,"Standardize",true, ...
    "LayerSizes",[30 10])
```

```
Mdl =
  RegressionNeuralNetwork
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 318
              LayerSizes: [30 10]
             Activations: 'relu'
    OutputLayerActivation: 'linear'
                  Solver: 'LBFGS'
          ConvergenceInfo: [1×1 struct]
          TrainingHistory: [1000×7 table]


  Properties, Methods
```

Access the weights and biases for the fully connected layers of the trained model by using the `LayerWeights` and `LayerBiases` properties of `Mdl`. The first two elements of each property correspond to the values for the first two fully connected layers, and the third element corresponds to the values for the final fully connected layer for regression. For example, display the weights and biases for the first fully connected layer.

```
Mdl.LayerWeights{1}
```

```
ans = 30×4

   -1.0617    0.1287    0.0797    0.4648
   -0.6497   -1.4565   -2.6026    2.6962
   -0.6420    0.2744   -0.0234   -0.0252
   -1.9727   -0.4665   -0.5833    0.9371
   -0.4373    0.1607    0.3930    0.7859
    0.5091   -0.0032   -0.6503   -1.6694
    0.0123   -0.2624   -2.2928   -1.0965
   -0.1386    1.2747    0.4085    0.5395
   -0.1755    1.5641   -3.1896   -1.1336
    0.4401    0.4942    1.8957   -1.1617
      ⋮
```

```
Mdl.LayerBiases{1}
```

```
ans = 30×1

   -1.3086
   -1.6205
   -0.7815
    1.5382
   -0.5256
    1.2394
   -2.3078
   -1.0709
   -1.8898
    1.9443
      ⋮
```

The final fully connected layer has one output. The number of layer outputs corresponds to the first dimension of the layer weights and layer biases.

```
size(Mdl.LayerWeights{end})
```

```
ans = 1×2

     1    10
```

```
size(Mdl.LayerBiases{end})
```

```
ans = 1×2

     1     1
```

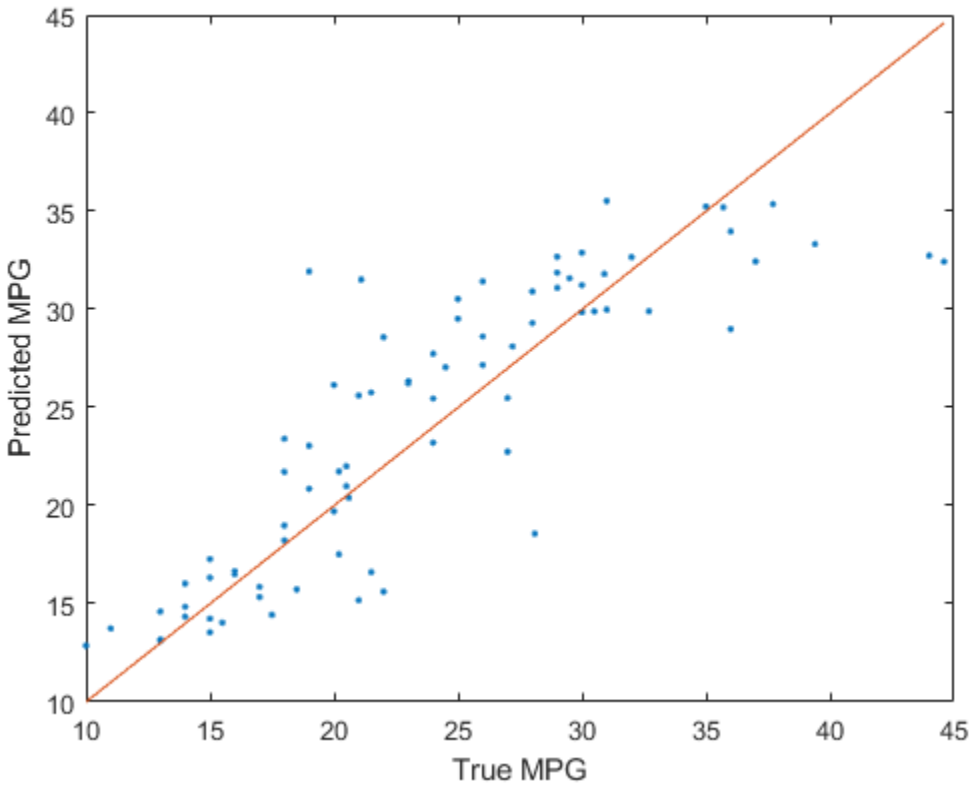
To estimate the performance of the trained model, compute the test set mean squared error (MSE) for `Mdl`. Smaller MSE values indicate better performance.

```
testMSE = loss(Mdl,XTest,YTest)
```

```
testMSE = 17.2022
```

Compare the predicted test set response values to the true response values. Plot the predicted miles per gallon (MPG) along the vertical axis and the true MPG along the horizontal axis. Points on the reference line indicate correct predictions. A good model produces predictions that are scattered near the line.

```
testPredictions = predict(Mdl,XTest);
plot(YTest,testPredictions,".")
hold on
plot(YTest,YTest)
hold off
xlabel("True MPG")
ylabel("Predicted MPG")
```



See Also

`CompactRegressionNeuralNetwork` | `RegressionPartitionedModel` | `fitrnet` | `loss` | `predict`

Topics

"Assess Regression Neural Network Performance" on page 18-184

Introduced in R2021a

RegressionPartitionedEnsemble

Package: `classreg.learning.partition`
Superclasses: `RegressionPartitionedModel`

Cross-validated regression ensemble

Description

`RegressionPartitionedEnsemble` is a set of regression ensembles trained on cross-validated folds. Estimate the quality of classification by cross validation using one or more "kfold" methods: `kfoldfun`, `kfoldLoss`, or `kfoldPredict`. Every "kfold" method uses models trained on in-fold observations to predict response for out-of-fold observations. For example, suppose you cross validate using five folds. In this case, every training fold contains roughly 4/5 of the data and every test fold contains roughly 1/5 of the data. The first model stored in `Trained{1}` was trained on `X` and `Y` with the first 1/5 excluded, the second model stored in `Trained{2}` was trained on `X` and `Y` with the second 1/5 excluded, and so on. When you call `kfoldPredict`, it computes predictions for the first 1/5 of the data using the first model, for the second 1/5 of data using the second model and so on. In short, response for every observation is computed by `kfoldPredict` using the model trained without this observation.

Construction

`cvens = crossval(ens)` creates a cross-validated ensemble from `ens`, a regression ensemble. For syntax details, see the `crossval` method reference page.

`cvens = fitrensemble(X,Y,Name,Value)` creates a cross-validated ensemble when `Name` is one of `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`. For syntax details, see the `fitrensemble` function reference page.

Input Arguments

`ens`

A regression ensemble constructed with `fitrensemble`.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CrossValidatedModel`

Name of the cross-validated model, a character vector.

`Kfold`

Number of folds used in a cross-validated tree, a positive integer.

`ModelParameters`

Object holding parameters of `tree`.

`NumObservations`

Numeric scalar containing the number of observations in the training data.

`NumTrainedPerFold`

Vector of `Kfold` elements. Each entry contains the number of trained learners in this cross-validation fold.

`Partition`

The partition of class `cvpartition` used in creating the cross-validated ensemble.

`PredictorNames`

A cell array of names for the predictor variables, in the order in which they appear in `X`.

ResponseName

Name of the response variable `Y`, a character vector.

ResponseTransform

Function handle for transforming scores, or character vector representing a built-in transformation function. `'none'` means no transformation; equivalently, `'none'` means `@(x)x`.

Add or change a `ResponseTransform` function using dot notation:

```
ens.ResponseTransform = @function
```

Trainable

Cell array of ensembles trained on cross-validation folds. Every ensemble is full, meaning it contains its training data and weights.

Trained

Cell array of compact ensembles trained on cross-validation folds.

W

The scaled `weights`, a vector with length `n`, the number of rows in `X`.

X

A matrix or table of predictor values. Each column of `X` represents one variable, and each row represents one observation.

Y

A numeric column vector with the same number of rows as `X`. Each entry in `Y` is the response to the data in the corresponding row of `X`.

Object Functions

kfoldLoss	Loss for cross-validated partitioned regression model
kfoldPredict	Predict responses for observations in cross-validated regression model
kfoldfun	Cross-validate function for regression
resume	Resume training ensemble

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Construct Partitioned Regression Ensemble

Construct a partitioned regression ensemble, and examine the cross-validation losses for the folds.

Load the `carsmall` data set.

```
load carsmall;
```

Create a subset of variables.

```
XX = [Cylinders Displacement Horsepower Weight];
YY = MPG;
```

Construct the ensemble model.

```
rens = fitrensemble(XX,YY);
```

Create a cross-validated ensemble from `rens`.

```
rng(10,'twister') % For reproducibility
cvrens = crossval(rens);
```

Examine the cross-validation losses.

```
L = kfoldLoss(cvrens,'mode','individual')
```

```
L = 10×1

   21.4489
   48.4388
   28.2560
   17.5354
   29.9441
   49.5254
   51.2372
   31.0152
   31.6388
    8.9607
```

L is a vector containing the cross-validation loss for each trained learner in the ensemble.

See Also

`ClassificationPartitionedEnsemble` | `RegressionEnsemble` | `RegressionPartitionedModel` | `fitrtree`

RegressionPartitionedGAM

Cross-validated generalized additive model (GAM) for regression

Description

`RegressionPartitionedGAM` is a set of generalized additive models trained on cross-validated folds. Estimate the quality of the cross-validated regression by using one or more *kfold* functions: `kfoldPredict`, `kfoldLoss`, and `kfoldfun`.

Every *kfold* object function uses models trained on training-fold (in-fold) observations to predict the response for validation-fold (out-of-fold) observations. For example, suppose you cross-validate using five folds. The software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (roughly 4/5 of the data), and the validation fold contains the other group (roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) by using the observations in the last four groups, and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) by using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar manner for the third, fourth, and fifth models.

If you validate by using `kfoldPredict`, the software computes predictions for the observations in group *i* by using the *i*th model. In short, the software estimates a response for every observation by using the model trained without that observation.

Creation

You can create a `RegressionPartitionedGAM` model in two ways:

- Create a cross-validated model from a GAM object `RegressionGAM` by using the `crossval` object function.
- Create a cross-validated model by using the `fitrgam` function and specifying one of the name-value arguments `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
`'GAM'`

This property is read-only.

Cross-validated model name, specified as `'GAM'`.

`KFold` — Number of cross-validated folds
positive integer

This property is read-only.

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

This property is read-only.

Cross-validation parameter values, specified as an object. The parameter values correspond to the values of the name-value arguments used to cross-validate the generalized additive model. `ModelParameters` does not contain estimated parameters.

You can access the properties of `ModelParameters` using dot notation.

`Partition` — Data partition
`cvpartition` model

This property is read-only.

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Compact models trained on cross-validation folds
cell array of `CompactRegressionGAM` models

This property is read-only.

Compact models trained on cross-validation folds, specified as a cell array of `CompactRegressionGAM` model objects. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`NumObservations` — Number of observations
numeric scalar

This property is read-only.

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

Data Types: `double`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the training data.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to train the model, specified as an *n*-by-1 numeric vector. *n* is the number of observations (`NumObservations`).

The software normalizes the observation weights specified in the `'Weights'` name-value argument so that the elements of `W` sum up to 1.

Data Types: `double`

`X` — Predictors
numeric matrix | table

This property is read-only.

Predictors used to cross-validate the model, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double` | `table`

`Y` — Response
numeric vector

This property is read-only.

Response used to cross-validate the model, specified as a numeric vector.

Each row of `Y` represents the observed response of the corresponding row of `X`.

Data Types: `single` | `double`

Object Functions

kfoldPredict	Predict responses for observations in cross-validated regression model
kfoldLoss	Loss for cross-validated partitioned regression model
kfoldfun	Cross-validate function for regression

Examples

Create Cross-Validated GAM Using `fitrgam`

Train a cross-validated GAM with 10 folds, which is the default cross-validation option, by using `fitrgam`. Then, use `kfoldPredict` to predict responses for validation-fold observations using a model trained on training-fold observations.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table that contains the predictor variables (`Acceleration`, `Displacement`, `Horsepower`, and `Weight`) and the response variable (`MPG`).

```
tbl = table(Acceleration,Displacement,Horsepower,Weight,MPG);
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`.

```
rng('default') % For reproducibility
CVMdl = fitrgam(tbl,'MPG','CrossVal','on')
```

```
CVMdl =
  RegressionPartitionedGAM
    CrossValidatedModel: 'GAM'
         PredictorNames: {1x4 cell}
           ResponseName: 'MPG'
        NumObservations: 398
                  KFold: 10
              Partition: [1x1 cvpartition]
      NumTrainedPerFold: [1x1 struct]
      ResponseTransform: 'none'


  Properties, Methods
```

The `fitrgam` function creates a `RegressionPartitionedGAM` model object `CVMdl` with 10 folds. During cross-validation, the software completes these steps:

1. Randomly partition the data into 10 sets.
2. For each set, reserve the set as validation data, and train the model using the other 9 sets.
3. Store the 10 compact, trained models a in a 10-by-1 cell vector in the `Trained` property of the cross-validated model object `RegressionPartitionedGAM`.

You can override the default cross-validation setting by using the `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value argument.

Predict responses for the observations in `tbl` by using `kfoldPredict`. The function predicts responses for every observation using the model trained without that observation.

```
yHat = kfoldPredict(CVMdl);
```

`yHat` is a numeric vector. Display the first five predicted responses.

```
yHat(1:5)
```

```
ans = 5×1

   19.4848
   15.7203
   15.5742
   15.3185
   17.8223
```

Compute the regression loss (mean squared error).

```
L = kfoldLoss(CVMdl)
```

```
L = 17.7248
```

`kfoldLoss` returns the average mean squared error over 10 folds.

Create Cross-Validated Regression GAM Using `crossval`

Train a regression generalized additive model (GAM) by using `fitrgam`, and create a cross-validated GAM by using `crossval` and the holdout option. Then, use `kfoldPredict` to predict responses for validation-fold observations using a model trained on training-fold observations.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Train a GAM that contains linear terms for predictors.

```
Mdl = fitrgam(tbl,'Systolic');
```

`Mdl` is a `RegressionGAM` model object.

Cross-validate the model by specifying a 30% holdout sample.

```
rng('default') % For reproducibility
CVMdl = crossval(Mdl,'Holdout',0.3)
```

```
CVMdl =
  RegressionPartitionedGAM
      CrossValidatedModel: 'GAM'
           PredictorNames: {1x6 cell}
    CategoricalPredictors: [3 5 6]
             ResponseName: 'Systolic'
          NumObservations: 100
                    KFold: 1
                Partition: [1x1 cvpartition]
        NumTrainedPerFold: [1x1 struct]
        ResponseTransform: 'none'


  Properties, Methods
```

The `crossval` function creates a `RegressionPartitionedGAM` model object `CVMdl` with the holdout option. During cross-validation, the software completes these steps:

1. Randomly select and reserve 30% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model object `RegressionPartitionedGAM`.

You can choose a different cross-validation setting by using the `'CrossVal'`, `'CVPartition'`, `'KFold'`, or `'Leaveout'` name-value argument.

Predict responses for the validation-fold observations by using `kfoldPredict`. The function predicts responses for the validation-fold observations by using the model trained on the training-fold observations. The function assigns `NaN` to the training-fold observations.

```
yFit = kfoldPredict(CVMdl);
```

Find the validation-fold observation indexes, and create a table containing the observation index, observed response values, and predicted response values. Display the first eight rows of the table.

```
idx = find(~isnan(yFit));
t = table(idx,tbl.Systolic(idx),yFit(idx), ...
    'VariableNames',{'Obseraction Index','Observed Value','Predicted Value'});
head(t)
```

```
ans=8×3 table
    Obseraction Index    Observed Value    Predicted Value
    _________________    ______________    _______________

            1                124              130.22
            6                121              124.38
            7                130              125.26
           12                115              117.05
           20                125              121.82
           22                123              116.99
           23                114                 107
```

```
           24                128              122.52
```

Compute the regression error (mean squared error) for the validation-fold observations.

```
L = kfoldLoss(CVMdl)
```

```
L = 43.8715
```

Find Optimal Number of Trees for GAM Using kfoldLoss

Train a cross-validated generalized additive model (GAM) with 10 folds. Then, use `kfoldLoss` to compute the cumulative cross-validation regression loss (mean squared errors). Use the errors to determine the optimal number of trees per predictor (linear term for predictor) and the optimal number of trees per interaction term.

Alternatively, you can find optimal values of `fitrgam` name-value arguments by using the `bayesopt` function. For an example, see “Optimize Cross-Validated GAM Using bayesopt” on page 33-2120.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, and `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Create a cross-validated GAM by using the default cross-validation option. Specify the `'CrossVal'` name-value argument as `'on'`. Also, specify to include 5 interaction terms.

```
rng('default') % For reproducibility
CVMdl = fitrgam(tbl,'Systolic','CrossVal','on','Interactions',5);
```

If you specify `'Mode'` as `'cumulative'` for `kfoldLoss`, then the function returns cumulative errors, which are the average errors across all folds obtained using the same number of trees for each fold. Display the number of trees for each fold.

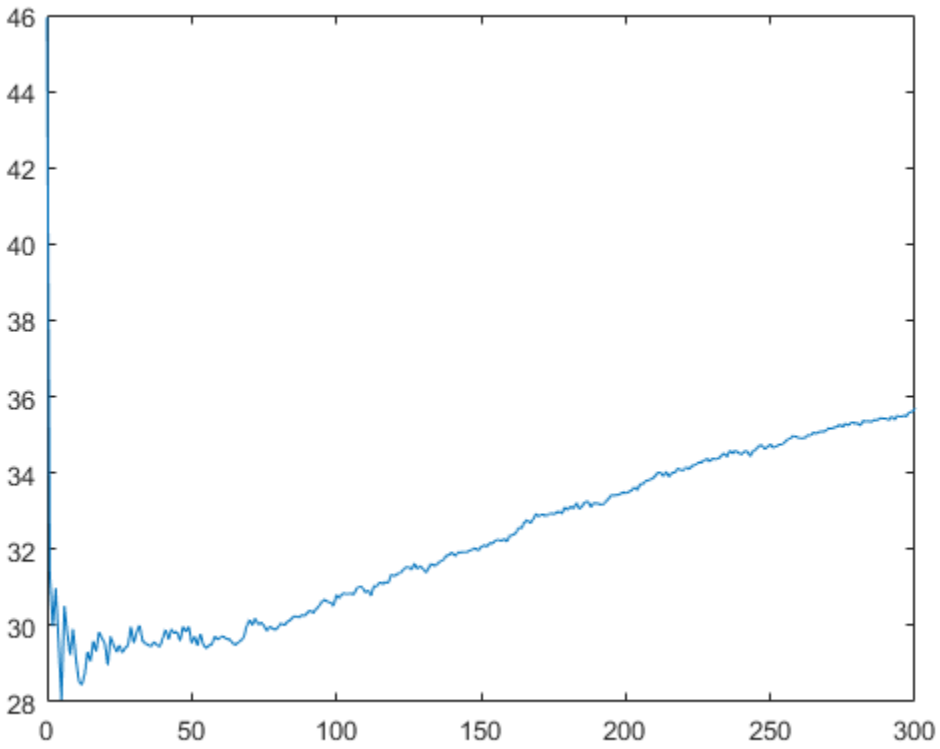
```
CVMdl.NumTrainedPerFold
```

```
ans = struct with fields:
      PredictorTrees: [300 300 300 300 300 300 300 300 300 300]
    InteractionTrees: [76 100 100 100 100 42 100 100 59 100]
```

`kfoldLoss` can compute cumulative errors using up to 300 predictor trees and 42 interaction trees.

Plot the cumulative, 10-fold cross-validated, mean squared errors. Specify `'IncludeInteractions'` as `false` to exclude interaction terms from the computation.

```
L_noInteractions = kfoldLoss(CVMdl,'Mode','cumulative','IncludeInteractions',false);
figure
plot(0:min(CVMdl.NumTrainedPerFold.PredictorTrees),L_noInteractions)
```



The first element of `L_noInteractions` is the average error over all folds obtained using only the intercept (constant) term. The (`J`+1)th element of `L_noInteractions` is the average error obtained using the intercept term and the first `J` predictor trees per linear term. Plotting the cumulative loss allows you to monitor how the error changes as the number of predictor trees in the GAM increases.

Find the minimum error and the number of predictor trees used to achieve the minimum error.

```
[M,I] = min(L_noInteractions)
```

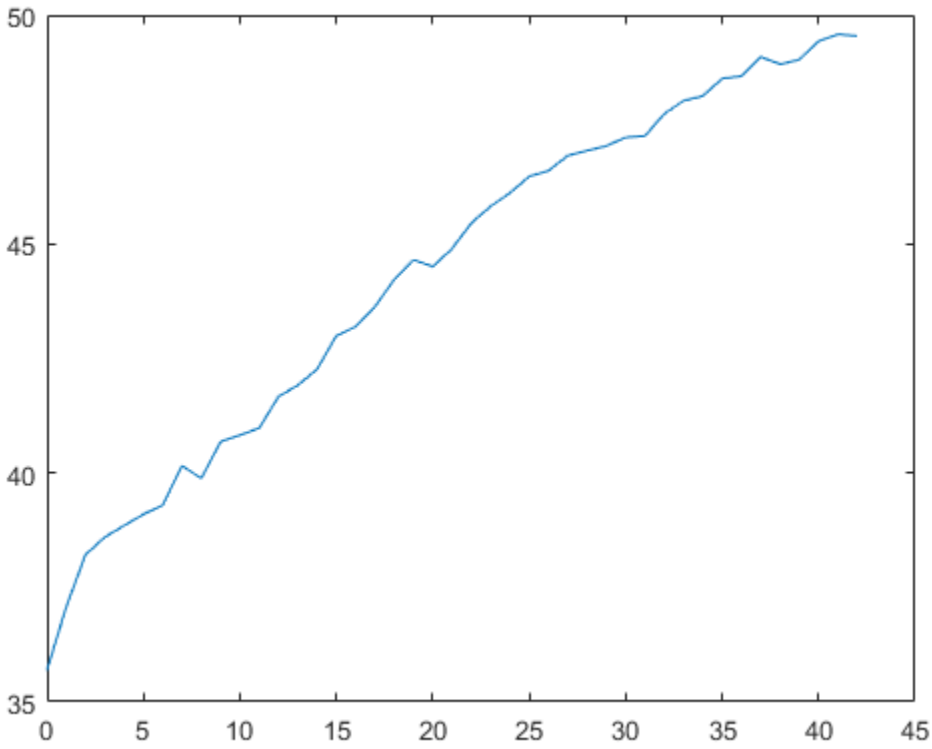
```
M = 28.0506
```

```
I = 6
```

The GAM achieves the minimum error when it includes 5 predictor trees.

Compute the cumulative mean squared error using both linear terms and interaction terms.

```
L = kfoldLoss(CVMdl,'Mode','cumulative');
figure
plot(0:min(CVMdl.NumTrainedPerFold.InteractionTrees),L)
```



The first element of `L` is the average error over all folds obtained using the intercept (constant) term and all predictor trees per linear term. The `(J+1)`th element of `L` is the average error obtained using the intercept term, all predictor trees per linear term, and the first `J` interaction trees per interaction term. The plot shows that the error increases when interaction terms are added.

If you are satisfied with the error when the number of predictor trees is 5, you can create a predictive model by training the univariate GAM again and specifying `'NumTreesPerPredictor',5` without cross-validation.

See Also

`RegressionGAM` | `crossval`

Topics

“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

RegressionPartitionedLinear

Package: `classreg.learning.partition`
Superclasses: `RegressionPartitionedModel`

Cross-validated linear regression model for high-dimensional data

Description

`RegressionPartitionedLinear` is a set of linear regression models trained on cross-validated folds. To obtain a cross-validated, linear regression model, use `fitrlinear` and specify one of the cross-validation options. You can estimate the predictive quality of the model, or how well the linear regression model generalizes, using one or more of these "kfold" methods: `kfoldPredict` and `kfoldLoss`.

Every "kfold" method uses models trained on in-fold observations to predict the response for out-of-fold observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five roughly equally sized groups. The training fold contains four of the groups (that is, roughly 4/5 of the data) and the test fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third through fifth models.

If you validate by calling `kfoldPredict`, it computes predictions for the observations in group 1 using the first model, group 2 for the second model, and so on. In short, the software estimates a response for every observation using the model trained without that observation.

Note Unlike other cross-validated, regression models, `RegressionPartitionedLinear` model objects do not store the predictor data set.

Construction

`CVMdl = fitrlinear(X,Y,Name,Value)` creates a cross-validated, linear regression model when `Name` is either `'CrossVal'`, `'CVPartition'`, `'Holdout'`, or `'KFold'`. For more details, see `fitrlinear`.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name

character vector

Cross-validated model name, specified as a character vector.

For example, `'Linear'` specifies a cross-validated linear model for binary classification or regression.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer

Number of cross-validated folds, specified as a positive integer.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

Cross-validation parameter values, e.g., the name-value pair argument values used to cross-validate the linear model, specified as an object. `ModelParameters` does not contain estimated parameters.

Access properties of `ModelParameters` using dot notation.

`NumObservations` — Number of observations
positive numeric scalar

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Linear regression models trained on cross-validation folds
cell array of `RegressionLinear` model objects

Linear regression models trained on cross-validation folds, specified as a cell array of `RegressionLinear` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

`Y` — Observed responses
numeric vector

Observed responses used to cross-validate the model, specified as a numeric vector containing `NumObservations` elements.

Each row of `Y` represents the observed response of the corresponding observation in the predictor data.

Data Types: `single` | `double`

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices

vector of positive integers | `[]`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`PredictorNames` — Predictor names

cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of variables in the training data `X` or `Tbl` used as predictor variables.

Data Types: `cell`

`ResponseName` — Response variable name

character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function

`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

Methods

kfoldLoss	Regression loss for observations not used in training
kfoldPredict	Predict responses for observations not used for training

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Create Cross-Validated Linear Regression Model

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Cross-validate a linear regression model. To increase execution speed, transpose the predictor data and specify that the observations are in columns.

```
X = X';
CVMdl = fitrlinear(X,Y,'CrossVal','on','ObservationsIn','columns');
```

`CVMdl` is a `RegressionPartitionedLinear` cross-validated model. Because `fitrlinear` implements 10-fold cross-validation by default, `CVMdl.Trained` contains a cell vector of ten `RegressionLinear` models. Each cell contains a linear regression model trained on nine folds, and then tested on the remaining fold.

Predict responses for out-of-fold observations and estimate the generalization error by passing `CVMdl` to `kfoldPredict` and `kfoldLoss`, respectively.

```
oofYHat = kfoldPredict(CVMdl);
ge = kfoldLoss(CVMdl)
```

```
ge = 0.1748
```

The estimated, generalization, mean squared error is 0.1748.

Find Good Lasso Penalty Using Cross-Validation

To determine a good lasso-penalty strength for a linear regression model that uses least squares, implement 5-fold cross-validation.

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.

- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-5} through 10^{-1}.

```
Lambda = logspace(-5,-1,15);
```

Cross-validate the models. To increase execution speed, transpose the predictor data and specify that the observations are in columns. Optimize the objective function using SpaRSA.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','KFold',5,'Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');

numCLModels = numel(CVMdl.Trained)
```

```
numCLModels = 5
```

`CVMdl` is a `RegressionPartitionedLinear` model. Because `fitrlinear` implements 5-fold cross-validation, `CVMdl` contains 5 `RegressionLinear` models that the software trains on each fold.

Display the first trained linear regression model.

```
Mdl1 = CVMdl.Trained{1}
```

```
Mdl1 = 
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x15 double]
                 Bias: [1x15 double]
               Lambda: [1x15 double]
              Learner: 'leastsquares'


  Properties, Methods
```

`Mdl1` is a `RegressionLinear` model object. `fitrlinear` constructed `Mdl1` by training on the first four folds. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl1` as 15 models, one for each regularization strength in `Lambda`.

Estimate the cross-validated MSE.

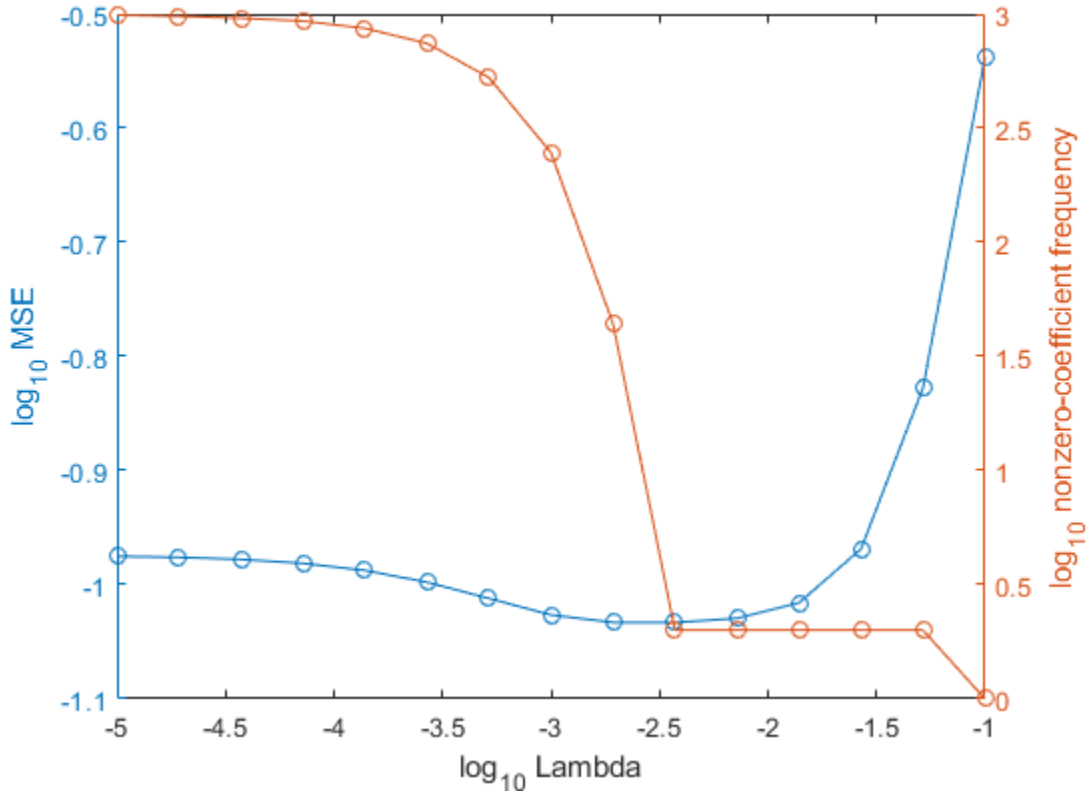
```
mse = kfoldLoss(CVMdl);
```

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. For each regularization strength, train a linear regression model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Learner','leastsquares','Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the cross-validated MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low MSE (for example, `Lambda(10)`).

```
idxFinal = 10;
```

Extract the model with corresponding to the minimal MSE.

```
MdlFinal = selectModels(Mdl,idxFinal)
```

```
MdlFinal =
  RegressionLinear
```

```
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [1000x1 double]
                 Bias: -0.0050
               Lambda: 0.0037
              Learner: 'leastsquares'


  Properties, Methods
```

```
idxNZCoeff = find(MdlFinal.Beta~=0)
```

```
idxNZCoeff = 2×1

   100
   200
```

```
EstCoeff = Mdl.Beta(idxNZCoeff)
```

```
EstCoeff = 2×1

    1.0051
    1.9965
```

`MdlFinal` is a `RegressionLinear` model with one regularization strength. The nonzero coefficients `EstCoeff` are close to the coefficients that simulated the data.

See Also

`RegressionLinear` | `fitrlinear` | `kfoldLoss` | `kfoldPredict`

Introduced in R2016a

RegressionPartitionedModel

Package: `classreg.learning.partition`

Cross-validated regression model

Description

`RegressionPartitionedModel` is a set of regression models trained on cross-validated folds. Estimate the quality of regression by cross validation using one or more "kfold" methods: `kfoldPredict`, `kfoldLoss`, and `kfoldfun`. Every "kfold" method uses models trained on in-fold observations to predict response for out-of-fold observations. For example, suppose you cross validate using five folds. In this case, every training fold contains roughly 4/5 of the data and every test fold contains roughly 1/5 of the data. The first model stored in `Trained{1}` was trained on `X` and `Y` with the first 1/5 excluded, the second model stored in `Trained{2}` was trained on `X` and `Y` with the second 1/5 excluded, and so on. When you call `kfoldPredict`, it computes predictions for the first 1/5 of the data using the first model, for the second 1/5 of data using the second model and so on. In short, response for every observation is computed by `kfoldPredict` using the model trained without this observation.

Construction

`CVMdl = crossval(Mdl)` creates a cross-validated regression model from a regression model (`Mdl`).

Alternatively:

- `CVNetMdl = fitrnet(X,Y,Name,Value)`
- `CVTreeMdl = fitrtree(X,Y,Name,Value)`

create a cross-validated model when `Name` is either `'CrossVal'`, `'KFold'`, `'Holdout'`, `'Leaveout'`, or `'CVPartition'`. For syntax details, see `fitrnet` and `fitrtree`.

Input Arguments

`Mdl`

A regression model, specified as one of the following:

- A neural network regression model trained using `fitrnet`
- A regression tree trained using `fitrtree`

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CrossValidatedModel`

Name of the cross-validated model, a character vector.

`Kfold`

Number of folds used in the cross-validated model, a positive integer.

`ModelParameters`

Object holding parameters of `Mdl`.

`NumObservations`

Number of observations in the training data stored in `X` and `Y`, specified as a numeric scalar.

`Partition`

The partition of class `cvpartition` used in the cross-validated model.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in X.

ResponseName

Name of the response variable Y, a character vector.

ResponseTransform

Function handle for transforming the raw response values (mean squared error). The function handle should accept a matrix of response values and return a matrix of the same size. The default character vector `'none'` means `@(x)x`, or no transformation.

Add or change a `ResponseTransform` function using dot notation:

```
CVMdl.ResponseTransform = @function
```

Trained

The trained learners, a cell array of compact regression models.

W

The scaled `weights`, a vector with length `n`, the number of observations in X.

X

A matrix or table of predictor values.

Y

A numeric column vector. Each entry in Y is the response value of the corresponding observation in X.

Object Functions

kfoldLoss	Loss for cross-validated partitioned regression model
kfoldPredict	Predict responses for observations in cross-validated regression model
kfoldfun	Cross-validate function for regression

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Evaluate Cross-Validation Error

Load the sample data. Create a variable X containing the `Horsepower` and `Weight` data.

```
load carsmall
X = [Horsepower Weight];
```

Construct a regression tree using the sample data.

```
cvtree = fitrtree(X,MPG,'crossval','on');
```

Evaluate the cross-validation error of the `carsmall` data using `Horsepower` and `Weight` as predictor variables for mileage (`MPG`).

```
L = kfoldLoss(cvtree)
```

```
L = 25.5338
```

See Also

`ClassificationPartitionedModel` | `RegressionNeuralNetwork` | `RegressionPartitionedEnsemble` | `RegressionTree` | `fitrnet` | `fitrtree`

RegressionPartitionedSVM

Package: `classreg.learning.partition`
Superclasses: `RegressionPartitionedModel`

Cross-validated support vector machine regression model

Description

`RegressionPartitionedSVM` is a set of support vector machine (SVM) regression models trained on cross-validated folds.

Construction

`CVMdl = crossval(mdl)` returns a cross-validated (partitioned) support vector machine regression model, `CVMdl`, from a trained SVM regression model, `mdl`.

`CVMdl = crossval(mdl,Name,Value)` returns a cross-validated model with additional options specified by one or more `Name,Value` pair arguments. `Name` can also be a property name on page 33-5380 and `Value` is the corresponding value. `Name` must appear inside single quotes (`''`). You can specify several name-value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Properties

`CategoricalPredictors` — Indices of categorical predictors

vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`CrossValidatedModel` — Name of the cross-validated model

character vector

Name of the cross-validated model, stored as a character vector.

Data Types: `char`

`KFold` — Number of cross-validation folds

positive integer value

Number of cross-validation folds, stored as a positive integer value.

Data Types: `single` | `double`

`ModelParameters` — Cross-validation parameters
object

Cross-validation parameters, stored as an object.

`NumObservations` — Number of observations
positive integer value

Number of observations in the training data, stored as a positive integer value.

Data Types: `single` | `double`

`Partition` — Data partition for cross-validation
`cvpartition` object

Data partition for cross-validation, stored as a `cvpartition` object.

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names, stored as a cell array of character vectors containing the name of each predictor in the order in which they appear in `X`. `PredictorNames` has a length equal to the number of columns in `X`.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

Response variable name, stored as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

`Trained` — Trained, compact regression models
cell array of `CompactRegressionSVM` models

Trained, compact regression models, stored as a cell array of `CompactRegressionSVM` models.

Data Types: `cell`

`W` — Observation weights
vector of numeric values

Observation weights used to train the model, stored as a numeric vector containing `NumObservation` number of elements. `fitrsvm` normalizes the weights used for training so that they sum to 1.

Data Types: `single` | `double`

`X` — Predictor values
matrix of numeric values | table of numeric values

Predictor values used to train the model, stored as a matrix of numeric values if the model is trained on a matrix, or a table if the model is trained on a table. `X` has size *n*-by-*p*, where *n* is the number of rows and *p* is the number of predictor variables or columns in the training data.

Data Types: `single` | `double` | `table`

`Y` — Observed responses
numeric vector

Observed responses used to cross-validate the model, specified as a numeric vector containing `NumObservations` elements.

Each row of `Y` represents the observed classification of the corresponding row of `X`.

Data Types: `single` | `double`

Object Functions

kfoldLoss	Loss for cross-validated partitioned regression model
kfoldPredict	Predict responses for observations in cross-validated regression model
kfoldfun	Cross-validate function for regression

Examples

Train Cross-Validated SVM Regression Model Using `crossval`

This example shows how to train a cross-validated SVM regression model using `crossval`.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with a kernel scale equal to 2.2. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale',2.2,'Standardize',true);
```

`mdl` is a trained `RegressionSVM` regression model.

Cross validate the model using 10-fold cross validation.

```
CVMdl = crossval(mdl)

CVMdl =

  classreg.learning.partition.RegressionPartitionedSVM
      CrossValidatedModel: 'SVM'
           PredictorNames: {1x8 cell}
    CategoricalPredictors: 1
             ResponseName: 'Var9'
          NumObservations: 4177
                    KFold: 10
                Partition: [1x1 cvpartition]
        ResponseTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedSVM` cross-validated regression model. The software:

1. Randomly partitions the data into ten equally-sized sets.

2. Trains an SVM regression model on nine of the ten sets.

3. Repeats steps 1 and 2 $k = 10$ times. It leaves out one of the partitions each time, and trains on the other nine partitions.

4. Combines generalization statistics for each fold.

Display the first of the 10 trained models.

```
FirstModel = CVMdl.Trained{1}

FirstModel =

  classreg.learning.regr.CompactRegressionSVM
       PredictorNames: {1x8 cell}
         ResponseName: 'Var9'
    ResponseTransform: 'none'
                Alpha: [3553x1 double]
                 Bias: 11.0623
     KernelParameters: [1x1 struct]
                   Mu: [0 0 0 0.5242 0.4080 0.1393 0.8300 0.3599 0.1811 0.2392]
                Sigma: [1 1 1 0.1205 0.0995 0.0392 0.4907 0.2217 0.1103 0.1392]
        SupportVectors: [3553x10 double]


  Properties, Methods
```

`FirstModel` is the first of the 10 trained `CompactRegressionSVM` models.

Specify Cross-Validation Holdout Proportion for SVM Regression

This example shows how to specify a holdout proportion for training a cross-validated SVM regression model.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with a kernel scale equal to 2.2. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale',2.2,'Standardize',true);
```

`mdl` is a trained `RegressionSVM` regression model.

Cross validate the regression model by specifying a 10% holdout sample.

```
CVMdl = crossval(mdl,'Holdout',0.1)
```

```
CVMdl =

  classreg.learning.partition.RegressionPartitionedSVM
      CrossValidatedModel: 'SVM'
           PredictorNames: {1x8 cell}
    CategoricalPredictors: 1
             ResponseName: 'Var9'
          NumObservations: 4177
                    KFold: 1
                Partition: [1x1 cvpartition]
       ResponseTransform: 'none'


  Properties, Methods
```

`CVMdl` is a `RegressionPartitionedSVM` model object.

Extract and display the trained, compact SVM regression model from `CVMdl`.

```
CVMdl.Trained{1}
```

```
TrainedModel =

  classreg.learning.regr.CompactRegressionSVM
       PredictorNames: {1x8 cell}
         ResponseName: 'Var9'
    ResponseTransform: 'none'
                Alpha: [3530x1 double]
                 Bias: 11.2646
     KernelParameters: [1x1 struct]
                   Mu: [0 0 0 0.5244 0.4080 0.1393 0.8282 0.3595 0.1805 0.2386]
                Sigma: [1 1 1 0.1198 0.0989 0.0388 0.4891 0.2218 0.1093 0.1390]
       SupportVectors: [3530x10 double]


  Properties, Methods
```

`TrainedModel` is a `CompactRegressionSVM` regression model that was trained using 90% of the data.

Alternatives

You can create a `RegressionPartitionedSVM` model using the following techniques:

- Use the training function `fitrsvm` and specify one of the `'CrossVal'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value pairs.
- Train a model using `fitrsvm`, then cross validate the model using the `crossval` method.
- Create a cross validation partition using `cvpartition`, then pass the resulting partition object to `fitrsvm` during training using the `'CVPartition'` name-value pair.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

See Also

`RegressionSVM` | `crossval` | `cvpartition` | `fitrsvm`

Introduced in R2015b

RegressionSVM class

Superclasses: `CompactRegressionSVM`

Support vector machine regression model

Description

`RegressionSVM` is a support vector machine (SVM) regression model. Train a `RegressionSVM` model using `fitrsvm` and the sample data.

`RegressionSVM` models store data, parameter values, support vectors, and algorithmic implementation information. You can use these models to:

- Estimate resubstitution predictions. For details, see `resubPredict`.
- Predict values for new data. For details, see `predict`.
- Compute resubstitution loss. For details, see `resubLoss`.
- Compute the mean square error or epsilon-insensitive loss. For details, see `loss`.

Construction

Create a `RegressionSVM` object by using `fitrsvm`.

Properties

`Alpha` — Dual problem coefficients

vector of numeric values

Dual problem coefficients, specified as a vector of numeric values. `Alpha` contains *m* elements, where *m* is the number of support vectors in the trained SVM regression model. The dual problem introduces two Lagrange multipliers for each support vector. The values of `Alpha` are the differences between the two estimated Lagrange multipliers for the support vectors. For more details, see “Understanding Support Vector Machine Regression” on page 25-2.

If you specified to remove duplicates using `RemoveDuplicates`, then, for a particular set of duplicate observations that are support vectors, `Alpha` contains one coefficient corresponding to the entire set. That is, MATLAB attributes a nonzero coefficient to one observation from the set of duplicates and a coefficient of `0` to all other duplicate observations in the set.

Data Types: `single` | `double`

`Beta` — Primal linear problem coefficients

vector of numeric values | `'[]'`

Primal linear problem coefficients, stored as a numeric vector of length *p*, where *p* is the number of predictors in the SVM regression model.

The values in `Beta` are the linear coefficients for the primal optimization problem.

If the model is obtained using a kernel function other than `'linear'`, this property is empty (`'[]'`).

The `predict` method computes predicted response values for the model as `YFIT = (X/S)×Beta + Bias`, where `S` is the value of the kernel scale stored in the `KernelParameters.Scale` property.

Data Types: `single` | `double`

`Bias` — Bias term

scalar value

Bias term in the SVM regression model, stored as a scalar value.

Data Types: `single` | `double`

`BoxConstraints` — Box constraints for dual problem coefficients

vector of numeric values

Box constraints for dual problem alpha coefficients, stored as a numeric vector containing *n* elements, where *n* is the number of observations in `X` (`Mdl.NumObservations`).

The absolute value of the dual coefficient `Alpha` for observation *i* cannot exceed `BoxConstraints(i)`.

If you specify removing duplicates using `'RemoveDuplicates'`, then for a given set of duplicate observations, MATLAB sums the box constraints, and then attributes the sum to one observation and box constraints of `0` to all other observations in the set.

Data Types: `single` | `double`

`CacheInfo` — Caching information

structure

Caching information, stored as a structure with the following fields.

Field	Description
`'Size'`	Positive scalar value indicating the cache size (in MB) that the software reserves to store entries of the Gram matrix. Set the cache size by using the `'CacheSize'` name-value pair argument in `fitrsvm`.
`'Algorithm'`	Character vector containing the name of the algorithm used to remove entries from the cache when its capacity is exceeded. Currently, the only available caching algorithm is `'Queue'`. You cannot set the caching algorithm.

Data Types: `struct`

`CategoricalPredictors` — Indices of categorical predictors

vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ConvergenceInfo` — Convergence information
structure

Convergence information, stored as a structure with the following fields.

Field	Description
`Converged`	Logical flag indicating whether the algorithm converged. A value of `1` indicates convergence.
`ReasonForConvergence`	Character vector indicating the criterion the software used to detect convergence.
`Gap`	Scalar feasibility gap between the dual and primal objective functions.
`GapTolerance`	Scalar tolerance for the feasibility gap. You can set this tolerance using the `'GapTolerance'` name-value pair argument in `fitrsvm`.
`DeltaGradient`	Scalar gradient difference between upper and lower violators.
`DeltaGradientTolerance`	Scalar tolerance for the gradient difference. You can set this tolerance using the `DeltaGradientTolerance` name-value pair argument in `fitrsvm`.
`LargestKKTViolation`	Maximal scalar Karush-Kuhn-Tucker (KKT) violation value.
`KKTTolerance`	Scalar tolerance for the largest KKT violation. You can set this tolerance using the `'KKTTolerance'` name-value pair argument in `fitrsvm`.
`History`	Structure containing convergence information recorded at periodic intervals during the model training process. This structure contains the following fields: • `NumIterations` — Array of iteration indices at which the software recorded convergence criteria. • `Gap` — `Gap` values at these iterations. • `DeltaGradient` — `DeltaGradient` values at these iterations. • `LargestKKTViolation` — `LargestKKTViolation` values at these iterations. • `NumSupportVectors` — Number of support vectors at these iterations. • `Objective` — `Objective` values at these iterations.
`Objective`	Numeric value of the dual objective.

Data Types: `struct`

`Epsilon` — Half the width of the epsilon-insensitive band
nonnegative scalar value

Half the width of the epsilon-insensitive band, stored as a nonnegative scalar value.

Data Types: `single` | `double`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`Gradient` — Gradient values in training data
vector of numeric values

Gradient values in training data, stored as a numeric vector containing 2*n* elements, where *n* is the number of observations in the training data.

Element *i* of `Gradient` contains the gradient value for the `Alpha` coefficient that corresponds to the upper boundary of the epsilon-insensitive band at observation *i* at the end of the optimization.

Element *i* + *NumObservations* of `Gradient` contains the gradient value for the `Alpha` coefficient that corresponds to the lower boundary of the epsilon-insensitive band at observation *i* at the end of the optimization.

Data Types: `single` | `double`

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table

This property is read-only.

Cross-validation optimization of hyperparameters, specified as a `BayesianOptimization` object or a table of hyperparameters and associated values. This property is nonempty if the `'OptimizeHyperparameters'` name-value pair argument is nonempty when you create the model. The value of `HyperparameterOptimizationResults` depends on the setting of the `Optimizer` field in the `HyperparameterOptimizationOptions` structure when you create the model.

Value of `Optimizer` Field	**Value of `HyperparameterOptimizationResults`**
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

`IsSupportVector` — Flag indicating whether observation is support vector
logical vector

Flag indicating whether an observation is a support vector, stored as an *n*-by-1 logical vector. *n* is the number of observations in `X` (see `NumObservations`). A value of `1` indicates that the corresponding observation in the training data is a support vector.

If you specify removing duplicates using `RemoveDuplicates`, then for a given set of duplicate observations that are support vectors, `IsSupportVector` flags only one as a support vector.

Data Types: `logical`

`KernelParameters` — Kernel function parameters
structure

Kernel function parameters, stored as a structure with the following fields.

Field	Description
`Function`	Kernel function name (a character vector).
`Scale`	Numeric scale factor used to divide predictor values.

You can specify values for `KernelParameters.Function` and `KernelParameters.Scale` by using the `KernelFunction` and `KernelScale` name-value pair arguments in `fitrsvm`, respectively.

Data Types: `struct`

`ModelParameters` — Parameter values
model parameter object

Parameter values used to train the SVM regression model, stored as a model parameter object. Access the properties of `ModelParameters` using dot notation. For example, access the value of `Epsilon` used to train the model as `Mdl.ModelParameters.Epsilon`.

`Mu` — Predictor means
vector of numeric values | `'[]'`

Predictor means, stored as a vector of numeric values.

If the predictors are standardized, then `Mu` is a numeric vector of length *p*, where *p* is the number of predictors used to train the model. In this case, the `predict` method centers predictor matrix `X` by subtracting the corresponding element of `Mu` from each column.

If the predictors are not standardized, then `Mu` is empty (`'[]'`).

If the data contains categorical predictors, then `Mu` includes elements for the dummy variables for those predictors. The corresponding entries in `Mu` are 0 because dummy variables are not centered or scaled.

Data Types: `single` | `double`

`NumIterations` — Number of iterations required for convergence
positive integer value

Number of iterations required for the optimization routine to reach convergence, stored as a positive integer value.

To set a limit on the number of iterations, use the `'IterationLimit'` name-value pair argument of `fitrsvm`.

Data Types: `single` | `double`

`NumObservations` — Number of observations
positive integer value

Number of observations in the training data, stored as a positive integer value.

Data Types: `single` | `double`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names, stored as a cell array of character vectors containing the name of each predictor in the order they appear in `X`. `PredictorNames` has a length equal to the number of columns in `X`.

Data Types: `cell`

`OutlierFraction` — Expected fraction of outliers
scalar value in the range [0,1]

Expected fraction of outliers in the training set, stored as a scalar value in the range [0,1]. You can specify the expected fraction of outliers using the `'OutlierFraction'` name-value pair argument in `fitrsvm`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, stored as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

`ShrinkagePeriod` — Number of iterations between reductions of active set
nonnegative integer value

Number of iterations between reductions of the active set during optimization, stored as a nonnegative integer value.

You can set the shrinkage period by using the `'ShrinkagePeriod'` name-value pair argument in `fitrsvm`.

Data Types: `single` | `double`

`Sigma` — Predictor standard deviations
vector of numeric values | `'[]'`

Predictor standard deviations, stored as a vector of numeric values.

If the predictors are standardized, then `Sigma` is a numeric vector of length *p*, where *p* is the number of predictors used to train the model. In this case, the `predict` method scales the predictor matrix `X` by dividing each column by the corresponding element of `Sigma`, after centering each element using `Mu`.

If the predictors are not standardized, then `Sigma` is empty (`'[]'`).

If the data contains categorical predictors, `Sigma` includes elements for the dummy variables for those predictors. The corresponding entries in `Sigma` are 1, because dummy variables are not centered or scaled.

Data Types: `single` | `double`

`Solver` — Name of solver algorithm
character vector

Name of the solver algorithm used to solve the optimization problem, stored as a value in this table.

Value	Description
`'SMO'`	Sequential Minimal Optimization
`'ISDA'`	Iterative Single Data Algorithm
`'L1QP'`	L1 soft-margin minimization by quadratic programming (requires an Optimization Toolbox license).

`SupportVectors` — Support vectors
matrix of numeric values

Support vectors, stored as an *m*-by-*p* matrix of numeric values. *m* is the number of support vectors (`sum(Mdl.IsSupportVector)`), and *p* is the number of predictors in `X`.

If you specified to remove duplicates using `RemoveDuplicates`, then for a given set of duplicate observations that are support vectors, `SupportVectors` contains one unique support vector.

Data Types: `single` | `double`

`W` — Observation weights
vector of numeric values

Observation weights used to train the model, stored as a numeric vector containing `NumObservation` number of elements. `fitrsvm` normalizes the weights used for training so that they sum to 1.

Data Types: `single` | `double`

`X` — Predictor values
matrix of numeric values | table of numeric values

Predictor values used to train the model, stored as a matrix of numeric values if the model is trained on a matrix, or a table if the model is trained on a table. `X` has size *n*-by-*p*, where *n* is the number of rows and *p* is the number of predictor variables or columns in the training data.

Data Types: `single` | `double` | `table`

Y — Observed response values
vector of numeric values

Observed response values, stored as a numeric vector containing `NumObservations` number of elements.

Data Types: `single` | `double`

Object Functions

compact	Compact support vector machine regression model
crossval	Cross-validated support vector machine regression model
discardSupportVectors	Discard support vectors
incrementalLearner	Convert support vector machine (SVM) regression model to incremental learner
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error for support vector machine regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using support vector machine regression model
resubLoss	Resubstitution loss for support vector machine regression model
resubPredict	Predict resubstitution response of support vector machine regression model
resume	Resume training support vector machine regression model
shapley	Shapley values

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Train Linear Support Vector Machine Regression Model

This example shows how to train a linear support vector machine (SVM) regression model using sample data stored in matrices.

Load the `carsmall` data set.

```
load carsmall
rng default  % for reproducibility
```

Specify `Horsepower` and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Horsepower,Weight];
Y = MPG;
```

Train a linear SVM regression model.

```
Mdl = fitrsvm(X,Y)
```

```
Mdl =
  RegressionSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                    Alpha: [75x1 double]
                     Bias: 57.3800
         KernelParameters: [1x1 struct]
          NumObservations: 93
           BoxConstraints: [93x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [93x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

The Command Window shows that `Mdl` is a trained `RegressionSVM` model and a list of its properties.

Check the model for convergence.

```
Mdl.ConvergenceInfo.Converged
```

```
ans = logical
   0
```

`0` indicates that the model did not converge.

```
MdlStd = fitrsvm(X,Y,'Standardize',true)
```

```
MdlStd =
  RegressionSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
                    Alpha: [77x1 double]
                     Bias: 22.9131
         KernelParameters: [1x1 struct]
                       Mu: [109.3441 2.9625e+03]
                    Sigma: [45.3545 805.9668]
          NumObservations: 93
           BoxConstraints: [93x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [93x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

Check the model for convergence.

```
MdlStd.ConvergenceInfo.Converged
```

```
ans = logical
   1
```

`1` indicates that the model did converge.

Compute the resubstitution mean squared error for the new model.

```
lStd = resubLoss(MdlStd)
```

```
lStd = 17.0256
```

Train Support Vector Machine Regression Model

Train a support vector machine regression model using the abalone data from the UCI Machine Learning Repository.

Download the data and save it in your current folder with the name `'abalone.csv'`.

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data';
websave('abalone.csv',url);
```

Read the data into a table. Specify the variable names.

```
varnames = {'Sex'; 'Length'; 'Diameter'; 'Height'; 'Whole_weight';...
    'Shucked_weight'; 'Viscera_weight'; 'Shell_weight'; 'Rings'};
Tbl = readtable('abalone.csv','Filetype','text','ReadVariableNames',false);
Tbl.Properties.VariableNames = varnames;
```

The sample data contains 4177 observations. All the predictor variables are continuous except for `Sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings (stored in `Rings`) on the abalone and determine its age using physical measurements.

Train an SVM regression model, using a Gaussian kernel function with an automatic kernel scale. Standardize the data.

```
rng default  % For reproducibility
Mdl = fitrsvm(Tbl,'Rings','KernelFunction','gaussian','KernelScale','auto',...
    'Standardize',true)
```

```
Mdl =

  RegressionSVM
           PredictorNames: {1×8 cell}
             ResponseName: 'Rings'
    CategoricalPredictors: 1
        ResponseTransform: 'none'
                    Alpha: [3635×1 double]
                     Bias: 10.8144
         KernelParameters: [1×1 struct]
                       Mu: [1×10 double]
                    Sigma: [1×10 double]
          NumObservations: 4177
```

```
          BoxConstraints: [4177×1 double]
         ConvergenceInfo: [1×1 struct]
         IsSupportVector: [4177×1 logical]
                  Solver: 'SMO'
```

The Command Window shows that `Mdl` is a trained `RegressionSVM` model and displays a property list.

Display the properties of `Mdl` using dot notation. For example, check to confirm whether the model converged and how many iterations it completed.

```
conv = Mdl.ConvergenceInfo.Converged
iter = Mdl.NumIterations
```

```
conv =

  logical

   1


iter =

        2759
```

The returned results indicate that the model converged after 2759 iterations.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of an SVM regression model into Simulink, you can use the RegressionSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train an SVM regression model by using `fitrsvm`, the following restrictions apply.
  - The value of the `'ResponseTransform'` name-value pair argument must be `'none'` (default).
  - For fixed-point code generation, the value of the `'KernelFunction'` name-value pair argument must be `'gaussian'`, `'linear'`, or `'polynomial'`.
  - Fixed-point code generation and code generation with a coder configurer do not support categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`CompactRegressionSVM` | `RegressionPartitionedSVM` | `fitrsvm`

Topics

“Understanding Support Vector Machine Regression” on page 25-2

Introduced in R2015b

RegressionSVMCoderConfigurer

Coder configurer for support vector machine (SVM) regression model

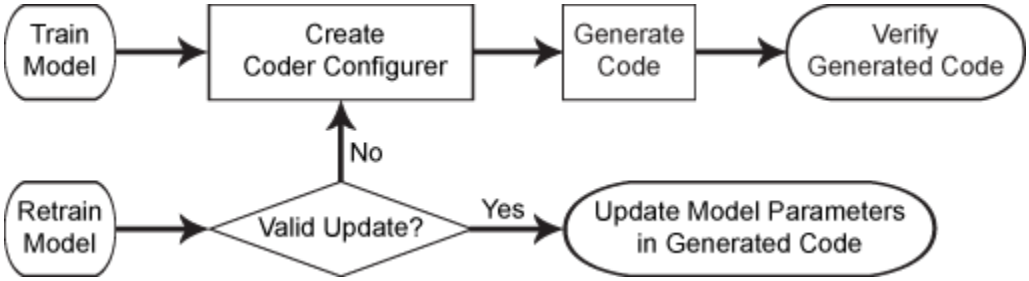
Description

A `RegressionSVMCoderConfigurer` object is a coder configurer of an SVM regression model (`RegressionSVM` or `CompactRegressionSVM`).

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes for SVM model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the SVM regression model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the SVM model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of an SVM regression model, see the Code Generation sections of `CompactRegressionSVM`, `predict`, and `update`.

Creation

After training an SVM regression model by using `fitrsvm`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data

`LearnerCoderInput` object

Coder attributes of predictor data to pass to the generated C/C++ code for the `predict` function of the SVM regression model, specified as a `LearnerCoderInput` on page 33-5409 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations of three predictor variables, specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

NumOutputs — Number of outputs in predict

1 (default)

Number of output arguments to return from the generated C/C++ code for the `predict` function of the SVM regression model, specified as 1. `predict` returns `yfit` (predicted responses) in the generated C/C++ code.

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the

`predict` and `update` functions of an SVM regression model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `single` | `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-5409 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Alpha` — Coder attributes of dual problem coefficients

`LearnerCoderInput` object

Coder attributes of the dual problem coefficients (`Alpha` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[s,1]`, where `s` is the number of support vectors in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function and discard support vectors by using `discardSupportVectors`, this value must be `false`. Otherwise, this value must be `true`.

`Beta` — Coder attributes of primal linear problem coefficients

`LearnerCoderInput` object

Coder attributes of the primal linear problem coefficients (`Beta` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[p 1]`, where `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function and discard support vectors by using `discardSupportVectors`, this value must be `true`. Otherwise, this value must be `false`.

`Bias` — Coder attributes of bias term

`LearnerCoderInput` object

Coder attributes of the bias term (`Bias` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

`Mu` — Coder attributes of predictor means

`LearnerCoderInput` object

Coder attributes of the predictor means (`Mu` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, this value must be `[1,p]`, where `p` is the number of predictors in `Mdl`. Otherwise, this value must be `[0,0]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, the default value is `true`. Otherwise, this value must be `false`.

`Scale` — Coder attributes of kernel scale parameter

`LearnerCoderInput` object

Coder attributes of the kernel scale parameter (`KernelParameters.Scale` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — This value must be `[1 1]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — The default value is `true`.

`Sigma` — Coder attributes of predictor standard deviations
`LearnerCoderInput` object

Coder attributes of the predictor standard deviations (`Sigma` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, this value must be `[1,p]`, where `p` is the number of predictors in `Mdl`. Otherwise, this value must be `[0,0]`.
- `VariableDimensions` — This value must be `[0 0]`, indicating that the array size is fixed as specified in `SizeVector`.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train `Mdl` using standardized predictor data by specifying `'Standardize',true`, the default value is `true`. Otherwise, this value must be `false`.

`SupportVectors` — Coder attributes of support vectors
`LearnerCoderInput` object

Coder attributes of the support vectors (`SupportVectors` of an SVM regression model), specified as a `LearnerCoderInput` on page 33-5409 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[s,p]`, where `s` is the number of support vectors, and `p` is the number of predictors in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — If you train a model with a linear kernel function and discard support vectors by using `discardSupportVectors`, this value must be `false`. Otherwise, this value must be `true`.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code
`'RegressionSVMModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `RegressionSVMCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level

`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments

cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — Input argument of `predict`
cell array of a `coder.PrimitiveType` object

This property is read-only.

Input argument of the entry-point function `predict.m` for code generation, specified as a cell array of a `coder.PrimitiveType` object. The `coder.PrimitiveType` object includes the coder attributes of the predictor data stored in the `X` property.

If you modify the coder attributes of the predictor data, then the software updates the `coder.PrimitiveType` object accordingly.

The `coder.PrimitiveType` object in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each `coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-5400 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer

validatedUpdateInputs Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `carsmall` data set and train a support vector machine (SVM) regression model.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X,Y);
```

`Mdl` is a `RegressionSVM` object.

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  RegressionSVMCoderConfigurer with properties:

   Update Inputs:
             Alpha: [1x1 LearnerCoderInput]
    SupportVectors: [1x1 LearnerCoderInput]
             Scale: [1x1 LearnerCoderInput]
              Bias: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 1
    OutputFileName: 'RegressionSVMModel'


  Properties, Methods
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionSVMModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionSVMModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:01
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('RegressionSVMModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Alpha
        %                       SupportVectors
        %                       Scale
        %                       Bias
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
```

```
end
end
```

Update Parameters of SVM Regression Model in Generated Code

Train a support vector machine (SVM) model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `carsmall` data set and train an SVM regression model using the first 50 observations.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `RegressionSVM` object.

Create Coder Configurer

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:));
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM regression model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM regression model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 2];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains two predictors, so the value of the `SizeVector` attribute must be two and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 2];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `RegressionSVMModel` for the two entry-point functions in the `codegen\mex\RegressionSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
yfit = predict(Mdl,X);
yfit_mex = RegressionSVMModel('predict',X);
```

`yfit_mex` might include round-off differences compared with `yfit`. In this case, compare `yfit` and `yfit_mex`, allowing a small tolerance.

```
find(abs(yfit-yfit_mex) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e–6`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
yfit = predict(retrainedMdl,X);
yfit_mex = RegressionSVMModel('predict',X);
find(abs(yfit-yfit_mex) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e-6`.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.

Attribute Name	Description
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of the coefficients `Alpha` of the coder configurer `configurer` as follows:

```
configurer.Alpha.SizeVector = [100 1];
configurer.Alpha.VariableDimensions = [1 0];
configurer.Alpha.DataType = 'double';
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

See Also

`CompactRegressionSVM` | `RegressionSVM` | `learnerCoderConfigurer` | `predict` | `update`

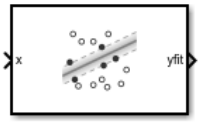
Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

RegressionSVM Predict

Predict responses using support vector machine (SVM) regression model
Library: Statistics and Machine Learning Toolbox / Regression



Description

The RegressionSVM Predict block predicts responses using an SVM regression object (`RegressionSVM` or `CompactRegressionSVM`).

Import a trained SVM regression object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **yfit** returns a predicted response for the observation.

Ports

Input

x — Predictor data

row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the SVM model specified by **`Select trained machine learning model`**.
- If you set `'Standardize',true` in `fitrsvm` when training the SVM model, then the RegressionSVM Predict block standardizes the values of **x** using the means and standard deviations in the `Mu` and `Sigma` properties (respectively) of the SVM model.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

yfit — Predicted response

scalar

Predicted response, returned as a scalar.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

`Select trained machine learning model` — SVM regression model

`svmMdl` (default) | `RegressionSVM` object | `CompactRegressionSVM` object

Specify the name of a workspace variable that contains a `RegressionSVM` object or `CompactRegressionSVM` object.

When you train the SVM model by using `fitrsvm`, the following restrictions apply:

- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ResponseTransform'` name-value argument must be `'none'` (default).
- The value of the `'KernelFunction'` name-value argument must be `'gaussian'`, `'linear'` (default), or `'polynomial'`.

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `RegressionSVM` object | `CompactRegressionSVM` object
Default: `'svmMdl'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations

`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use

Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action

`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use
Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type
`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use
Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Output data type` — Data type of yfit output
`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **yfit** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use

Block Parameter: `OutDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Output minimum` — Minimum value of yfit output for range checking
`[]` (default) | scalar

Lower value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output minimum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `OutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Output maximum` — Maximum value of yfit output for range checking

`[]` (default) | scalar

Upper value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output maximum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `OutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Kernel data type` — Kernel computation data type

`double` (default) | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `int64` | `uint64` | `uint32` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type of a parameter for kernel computation. The type can be specified directly or expressed as a data type object such as `Simulink.NumericType`.

The **Kernel data type** parameter specifies the data type of a different parameter depending on the type of kernel function of the specified SVM model. You specify the `'KernelFunction'` name-value argument when training the SVM model.

`'KernelFunction'` value	Data Type
`'gaussian'` or `'rbf'`	**Kernel data type** specifies the data type of the squared distance $D^2 = \|x - s\|^2$ for the Gaussian kernel $G(x,s) = \exp\left(-D^2\right)$, where x is the predictor data for an observation and s is a support vector.
`'linear'`	**Kernel data type** specifies the data type for the output of the linear kernel function $G(x,s) = xs'$, where x is the predictor data for an observation and s is a support vector.

'KernelFunction' value	Data Type
'polynomial'	**Kernel data type** specifies the data type for the output of the polynomial kernel function $G(x,s) = (1 + xs')^p$, where x is the predictor data for an observation, s is a support vector, and p is a polynomial kernel function order.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use
Block Parameter: `KernelDataTypeStr`
Type: character vector
Values: `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'uint64'` | `'int64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'double'`

`Kernel minimum` — Minimum kernel computation value for range checking

`[ ]` (default) | scalar

Lower value of the kernel computation internal variable range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Kernel minimum** parameter does not saturate or clip the actual kernel computation value signal.

Programmatic Use
Block Parameter: `KernelOutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Kernel maximum` — Maximum kernel computation value for range checking

`[ ]` (default) | scalar

Upper value of the kernel computation internal variable range that Simulink checks.

Simulink uses the maximum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Kernel maximum** parameter does not saturate or clip the actual kernel computation value signal.

Programmatic Use

Block Parameter: `KernelOutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

Tips

- If you are using a linear SVM model and it has many support vectors, then prediction can be slow. To efficiently predict responses based on a linear SVM model, remove the support vectors from the `RegressionSVM` or `CompactRegressionSVM` object by using `discardSupportVectors`.

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of an SVM regression object (`RegressionSVM` or `CompactRegressionSVM`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the RegressionSVM Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.
- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Compatibility Considerations

Specify Kernel data type as a data type name or data type object

Behavior changed in R2021a

Starting in R2021a, the **Kernel data type** parameter does not support inherited options. You can specify **Kernel data type** as a supported data type name or data type object.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion

Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks

ClassificationSVM Predict | RegressionEnsemble Predict | RegressionTree Predict

Objects

`CompactRegressionSVM` | `RegressionSVM`

Functions

`fitrsvm` | `predict`

Topics

“Predict Responses Using RegressionTree Predict Block” on page 32-127
“Predict Responses Using RegressionEnsemble Predict Block” on page 32-137
“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2020b

RegressionTree class

Superclasses: `CompactRegressionTree`

Regression tree

Description

A decision tree with binary splits for regression. An object of class `RegressionTree` can predict responses for new data with the `predict` method. The object contains the data used for training, so can compute resubstitution predictions.

Construction

Create a `RegressionTree` object by using `fitrtree`.

Properties

`BinEdges`

Bin edges for numeric predictors, specified as a cell array of *p* numeric vectors, where *p* is the number of predictors. Each vector includes the bin edges for a numeric predictor. The element in the cell array for a categorical predictor is empty because the software does not bin categorical predictors.

The software bins numeric predictors only if you specify the `'NumBins'` name-value argument as a positive integer scalar when training a model with tree learners. The `BinEdges` property is empty if the `'NumBins'` value is empty (default).

You can reproduce the binned predictor data `Xbinned` by using the `BinEdges` property of the trained model `mdl`.

```
X = mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are 0 for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

`CategoricalPredictors`

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

`CategoricalSplits`

An *n*-by-2 cell array, where `n` is the number of categorical splits in `tree`. Each row in `CategoricalSplits` gives left and right values for a categorical split. For each branch node with categorical split `j` based on a categorical predictor variable `z`, the left child is chosen if `z` is in `CategoricalSplits(j,1)` and the right child is chosen if `z` is in `CategoricalSplits(j,2)`. The splits are in the same order as nodes of the tree. Nodes for these splits can be found by running `cuttype` and selecting `'categorical'` cuts from top to bottom.

`Children`

An *n*-by-2 array containing the numbers of the child nodes for each node in `tree`, where *n* is the number of nodes. Leaf nodes have child node `0`.

`CutCategories`

An *n*-by-2 cell array of the categories used at branches in `tree`, where *n* is the number of nodes. For each branch node `i` based on a categorical predictor variable `x`, the left child is chosen if `x` is among the categories listed in `CutCategories{i,1}`, and the right child is chosen if `x` is among those listed in `CutCategories{i,2}`. Both columns of `CutCategories` are empty for branch nodes based on continuous predictors and for leaf nodes.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

`CutPoint`

An *n*-element vector of the values used as cut points in `tree`, where *n* is the number of nodes. For each branch node `i` based on a continuous predictor variable `x`, the left child is chosen if `x<CutPoint(i)` and the right child is chosen if `x>=CutPoint(i)`. `CutPoint` is `NaN` for branch nodes based on categorical predictors and for leaf nodes.

`CutType`

An *n*-element cell array indicating the type of cut at each node in `tree`, where *n* is the number of nodes. For each node `i`, `CutType{i}` is:

- `'continuous'` — If the cut is defined in the form `x < v` for a variable `x` and cut point `v`.
- `'categorical'` — If the cut is defined by whether a variable `x` takes a value in a set of categories.
- `''` — If `i` is a leaf node.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

CutPredictor

An *n*-element cell array of the names of the variables used for branching in each node in `tree`, where *n* is the number of nodes. These variables are sometimes known as *cut variables*. For leaf nodes, `CutPredictor` contains an empty character vector.

`CutPoint` contains the cut points for `'continuous'` cuts, and `CutCategories` contains the set of categories.

CutPredictorIndex

An *n*-element array of numeric indices for the variables used for branching in each node in `tree`, where *n* is the number of nodes. For more information, see `CutPredictor`.

ExpandedPredictorNames

Expanded predictor names, stored as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

HyperparameterOptimizationResults

Description of the cross-validation optimization of hyperparameters, stored as a `BayesianOptimization` object or a table of hyperparameters and associated values. Nonempty when the `OptimizeHyperparameters` name-value pair is nonempty at creation. Value depends on the setting of the `HyperparameterOptimizationOptions` name-value pair at creation:

- `'bayesopt'` (default) — Object of class `BayesianOptimization`
- `'gridsearch'` or `'randomsearch'` — Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

IsBranchNode

An *n*-element logical vector `ib` that is `true` for each branch node and `false` for each leaf node of `tree`.

ModelParameters

Object holding parameters of `tree`.

NumObservations

Number of observations in the training data, a numeric scalar. `NumObservations` can be less than the number of rows of input data `X` when there are missing values in `X` or response `Y`.

NodeError

An *n*-element vector `e` of the errors of the nodes in `tree`, where *n* is the number of nodes. `e(i)` is the mean squared error for node `i`.

NodeMean

An *n*-element numeric array with mean values in each node of `tree`, where *n* is the number of nodes in the tree. Every element in `NodeMean` is the average of the true `Y` values over all observations in the node.

NodeProbability

An *n*-element vector `p` of the probabilities of the nodes in `tree`, where *n* is the number of nodes. The probability of a node is computed as the proportion of observations from the original data that satisfy the conditions for the node.

NodeRisk

An *n*-element vector of the risk of the nodes in the tree, where *n* is the number of nodes. The risk for each node is the node error weighted by the node probability.

NodeSize

An *n*-element vector `sizes` of the sizes of the nodes in `tree`, where *n* is the number of nodes. The size of a node is defined as the number of observations from the data used to create the tree that satisfy the conditions for the node.

NumNodes

The number of nodes `n` in `tree`.

Parent

An *n*-element vector `p` containing the number of the parent node for each node in `tree`, where *n* is the number of nodes. The parent of the root node is `0`.

PredictorNames

A cell array of names for the predictor variables, in the order in which they appear in `X`.

PruneAlpha

Numeric vector with one element per pruning level. If the pruning level ranges from 0 to *M*, then `PruneAlpha` has *M* + 1 elements sorted in ascending order. `PruneAlpha(1)` is for pruning level 0 (no pruning), `PruneAlpha(2)` is for pruning level 1, and so on.

PruneList

An *n*-element numeric vector with the pruning levels in each node of `tree`, where *n* is the number of nodes. The pruning levels range from 0 (no pruning) to *M*, where *M* is the distance between the deepest leaf and the root node.

ResponseName

A character vector that specifies the name of the response variable (`Y`).

ResponseTransform

Function handle for transforming the raw response values (mean squared error). The function handle must accept a matrix of response values and return a matrix of the same size. The default `'none'` means `@(x)x`, or no transformation.

Add or change a `ResponseTransform` function using dot notation:

```
tree.ResponseTransform = @function
```

RowsUsed

An *n*-element logical vector indicating which rows of the original predictor data (`X`) were used in fitting. If the software uses all rows of `X`, then `RowsUsed` is an empty array (`[]`).

SurrogateCutCategories

An *n*-element cell array of the categories used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutCategories{k}` is a cell array. The length of `SurrogateCutCategories{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutCategories{k}` is either an empty character vector for a continuous surrogate predictor, or is a two-element cell array with categories for a categorical surrogate predictor. The first element of this two-element cell array lists categories assigned to the left child by this surrogate split, and the second element of this two-element cell array lists categories assigned to the right child by this surrogate split. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutCategories` contains an empty cell.

SurrogateCutFlip

An *n*-element cell array of the numeric cut assignments used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutFlip{k}` is a numeric vector. The length of `SurrogateCutFlip{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutFlip{k}` is either zero for a categorical surrogate predictor, or a numeric cut assignment for a continuous surrogate predictor. The numeric cut assignment can be either –1 or +1. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z < C$ and the cut assignment for this surrogate split is +1, or if $Z \geq C$ and the cut assignment for this surrogate split is –1. Similarly, the right child is chosen if $Z \geq C$ and the cut assignment for this surrogate split is +1, or if $Z < C$ and the cut assignment for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables in `SurrogateCutPredictor`. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, `SurrogateCutFlip` contains an empty array.

SurrogateCutPoint

An *n*-element cell array of the numeric values used for surrogate splits in `tree`, where *n* is the number of nodes in `tree`. For each node `k`, `SurrogateCutPoint{k}` is a numeric vector. The length of `SurrogateCutPoint{k}` is equal to the number of surrogate predictors found at this node. Every element of `SurrogateCutPoint{k}` is either `NaN` for a categorical surrogate predictor, or a numeric cut for a continuous surrogate predictor. For every surrogate split with a numeric cut C based on a continuous predictor variable Z, the left child is chosen if $Z<C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z \geq C$ and `SurrogateCutFlip` for this surrogate split is –1. Similarly, the right child is chosen if $Z \geq C$ and `SurrogateCutFlip` for this surrogate split is +1, or if $Z < C$ and

SurrogateCutFlip for this surrogate split is –1. The order of the surrogate split variables at each node is matched to the order of variables returned by SurrCutPredictor. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, SurrogateCutPoint contains an empty cell.

SurrogateCutType

An *n*-element cell array indicating types of surrogate splits at each node in tree, where *n* is the number of nodes in tree. For each node k, SurrogateCutType{k} is a cell array with the types of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The order of the surrogate split variables at each node is matched to the order of variables in SurrogateCutPredictor. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, SurrogateCutType contains an empty cell. A surrogate split type can be either 'continuous' if the cut is defined in the form Z < V for a variable Z and cut point V or 'categorical' if the cut is defined by whether Z takes a value in a set of categories.

SurrogateCutPredictor

An *n*-element cell array of the names of the variables used for surrogate splits in each node in tree, where *n* is the number of nodes in tree. Every element of SurrogateCutPredictor is a cell array with the names of the surrogate split variables at this node. The variables are sorted by the predictive measure of association with the optimal predictor in the descending order, and only variables with the positive predictive measure are included. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, SurrogateCutPredictor contains an empty cell.

SurrogatePredictorAssociation

An *n*-element cell array of the predictive measures of association for surrogate splits in tree, where *n* is the number of nodes in tree. For each node k, SurrogatePredictorAssociation{k} is a numeric vector. The length of SurrogatePredictorAssociation{k} is equal to the number of surrogate predictors found at this node. Every element of SurrogatePredictorAssociation{k} gives the predictive measure of association between the optimal split and this surrogate split. The order of the surrogate split variables at each node is the order of variables in SurrogateCutPredictor. The optimal-split variable at this node does not appear. For nonbranch (leaf) nodes, SurrogatePredictorAssociation contains an empty cell.

W

The scaled weights, a vector with length n, the number of rows in X.

X

A matrix or table of predictor values. Each column of X represents one variable, and each row represents one observation.

Y

A numeric column vector with the same number of rows as X. Each entry in Y is the response to the data in the corresponding row of X.

Object Functions

compact	Compact regression tree

crossval	Cross-validated decision tree
cvloss	Regression error by cross validation
lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression error
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using regression tree
predictorImportance	Estimates of predictor importance for regression tree
prune	Produce sequence of regression subtrees by pruning
resubLoss	Regression error by resubstitution
resubPredict	Predict resubstitution response of tree
shapley	Shapley values
surrogateAssociation	Mean predictive measure of association for surrogate splits in regression tree
view	View regression tree

Copy Semantics

Value. To learn how value classes affect copy operations, see Copying Objects.

Examples

Construct Regression Tree

Load the sample data.

```
load carsmall
```

Construct a regression tree using the sample data. The response variable is miles per gallon, MPG.

```
tree = fitrtree([Weight, Cylinders],MPG,...
                'CategoricalPredictors',2,'MinParentSize',20,...
                'PredictorNames',{'W','C'})
```

```
tree =
  RegressionTree
           PredictorNames: {'W'  'C'}
             ResponseName: 'Y'
    CategoricalPredictors: 2
        ResponseTransform: 'none'
          NumObservations: 94


  Properties, Methods
```

Predict the mileage of 4,000-pound cars with 4, 6, and 8 cylinders.

```
MPG4Kpred = predict(tree,[4000 4; 4000 6; 4000 8])
```

```
MPG4Kpred = 3×1

   19.2778
```

```
19.2778
14.3889
```

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `predict` and `update` functions support code generation.
- To integrate the prediction of a regression tree model into Simulink, you can use the RegressionTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function.
- When you train a regression tree model by using `fitrtree`, the following restrictions apply.
  - The value of the `'ResponseTransform'` name-value pair argument must be `'none'` (default).
  - You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value pair argument must be `'off'`.
  - Fixed-point code generation and code generation with a coder configurer do not support categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). You cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess them by using `dummyvar` before fitting the model.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`ClassificationTree` | `CompactRegressionTree` | `RegressionEnsemble` | `fitrtree` | `predict`

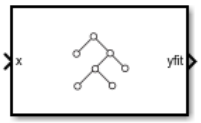
Topics

“Decision Trees” on page 19-2

Introduced in R2011a

RegressionTree Predict

Predict responses using regression tree model
Library: Statistics and Machine Learning Toolbox / Regression



Description

The RegressionTree Predict block predicts responses using a regression tree object (`RegressionTree` or `CompactRegressionTree`).

Import a trained regression object into the block by specifying the name of a workspace variable that contains the object. The input port **x** receives an observation (predictor data), and the output port **yfit** returns a predicted response for the observation.

Ports

Input

x — Predictor data
row vector | column vector

Predictor data, specified as a column vector or row vector of one observation.

Dependencies

- The variables in **x** must have the same order as the predictor variables that trained the model specified by **`Select trained machine learning model`**.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Output

yfit — Predicted response
scalar

Predicted response, returned as a scalar.

Data Types: `single` | `double` | `half` | `int8` | `int16` | `int32` | `int64` | `uint8` | `uint16` | `uint32` | `uint64` | `Boolean` | `fixed point`

Parameters

Main

Select trained machine learning model — Regression tree model
`treeMdl` (default) | `RegressionTree` object | `CompactRegressionTree` object

Specify the name of a workspace variable that contains a `RegressionTree` object or `CompactRegressionTree` object.

When you train the model by using `fitrtree`, the following restrictions apply:

- The predictor data cannot include categorical predictors (`logical`, `categorical`, `char`, `string`, or `cell`). If you supply training data in a table, the predictors must be numeric (`double` or `single`). Also, you cannot use the `'CategoricalPredictors'` name-value argument. To include categorical predictors in a model, preprocess the categorical predictors by using `dummyvar` before fitting the model.
- The value of the `'ResponseTransform'` name-value argument must be `'none'` (default).
- You cannot use surrogate splits, that is, the value of the `'Surrogate'` name-value argument must be `'off'` (default).

Programmatic Use

Block Parameter: `TrainedLearner`
Type: workspace variable
Values: `RegressionTree` object | `CompactRegressionTree` object
Default: `'treeMdl'`

Data Types

Fixed-Point Operational Parameters

`Integer rounding mode` — Rounding mode for fixed-point operations

`Floor` (default) | `Ceiling` | `Convergent` | `Nearest` | `Round` | `Simplest` | `Zero`

Specify the rounding mode for fixed-point operations. For more information, see “Rounding” (Fixed-Point Designer).

Block parameters always round to the nearest representable value. To control the rounding of a block parameter, enter an expression using a MATLAB rounding function into the mask field.

Programmatic Use

Block Parameter: `RndMeth`
Type: character vector
Values: `'Ceiling' | 'Convergent' | 'Floor' | 'Nearest' | 'Round' | 'Simplest' | 'Zero'`
Default: `'Floor'`

`Saturate on integer overflow` — Method of overflow action

`off` (default) | `on`

Specify whether overflows saturate or wrap.

Action	Rationale	Impact on Overflows	Example
Select this check box (`on`).	Your model has possible overflow, and you want explicit saturation protection in the generated code.	Overflows saturate to either the minimum or maximum value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box selected, the block output saturates at 127. Similarly, the block output saturates at a minimum output value of –128.
Clear this check box (`off`).	You want to optimize efficiency of your generated code. You want to avoid overspecifying how a block handles out-of-range signals. For more information, see "Troubleshoot Signal Range Errors" (Simulink).	Overflows wrap to the appropriate value that the data type can represent.	The maximum value that the `int8` (signed 8-bit integer) data type can represent is 127. Any block operation result greater than this maximum value causes overflow of the 8-bit integer. With the check box cleared, the software interprets the overflow causing value as `int8`, which can produce an unintended result. For example, a block result of 130 (binary 1000 0010) expressed as `int8` is –126.

Programmatic Use
Block Parameter: `SaturateOnIntegerOverflow`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

`Lock output data type setting against changes by the fixed-point tools` — Prevent fixed-point tools from overriding data type
`off` (default) | `on`

Select this parameter to prevent the fixed-point tools from overriding the data type you specify for the block. For more information, see "Use Lock Output Data Type Setting" (Fixed-Point Designer).

Programmatic Use
Block Parameter: `LockScale`
Type: character vector
Values: `'off' | 'on'`
Default: `'off'`

Data Type

`Output data type` — Data type of yfit output
`Inherit: auto` (default) | `double` | `single` | `half` | `int8` | `uint8` | `int16` | `uint16` | `int32` | `uint32` | `int64` | `uint64` | `boolean` | `fixdt(1,16)` | `fixdt(1,16,0)` | `fixdt(1,16,2^0,0)` | `<data type expression>`

Specify the data type for the **yfit** output. The type can be inherited, specified directly, or expressed as a data type object such as `Simulink.NumericType`.

When you select `Inherit: auto`, the block uses a rule that inherits a data type.

For more information about data types, see “Control Signal Data Types” (Simulink).

Click the **Show data type assistant** button  to display the **Data Type Assistant**, which helps you set the data type attributes. For more information, see “Specify Data Types Using Data Type Assistant” (Simulink).

Programmatic Use

Block Parameter: `OutDataTypeStr`
Type: character vector
Values: `'Inherit: auto'` | `'double'` | `'single'` | `'half'` | `'int8'` | `'uint8'` | `'int16'` | `'uint16'` | `'int32'` | `'uint32'` | `'int64'` | `'uint64'` | `'boolean'` | `'fixdt(1,16)'` | `'fixdt(1,16,0)'` | `'fixdt(1,16,2^0,0)'` | `'<data type expression>'`
Default: `'Inherit: auto'`

`Output minimum` — Minimum value of yfit output for range checking
`[]` (default) | scalar

Lower value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output minimum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use

Block Parameter: `OutMin`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

`Output maximum` — Maximum value of yfit output for range checking
`[]` (default) | scalar

Upper value of the **yfit** output range that Simulink checks.

Simulink uses the minimum value to perform:

- Parameter range checking (see “Specify Minimum and Maximum Values for Block Parameters” (Simulink)) for some blocks.
- Simulation range checking (see “Specify Signal Ranges” (Simulink) and “Enable Simulation Range Checking” (Simulink)).
- Automatic scaling of fixed-point data types.
- Optimization of the code that you generate from the model. This optimization can remove algorithmic code and affect the results of some simulation modes such as SIL or external mode. For more information, see “Optimize using the specified minimum and maximum values” (Embedded Coder).

Note The **Output maximum** parameter does not saturate or clip the actual **yfit** signal. Use the Saturation block instead.

Programmatic Use
Block Parameter: `OutMax`
Type: character vector
Values: `'[]'` | scalar
Default: `'[]'`

Block Characteristics

Data Types	`Boolean` | `double` | `fixed point` | `half` | `integer` | `single`
Direct Feedthrough	`yes`
Multidimensional Signals	`no`
Variable-Size Signals	`no`
Zero-Crossing Detection	`no`

Alternative Functionality

You can use a MATLAB Function block with the `predict` object function of a regression tree object (`RegressionTree` or `CompactRegressionTree`). For an example, see “Predict Class Labels Using MATLAB Function Block” on page 32-40.

When deciding whether to use the RegressionTree Predict block in the Statistics and Machine Learning Toolbox library or a MATLAB Function block with the `predict` function, consider the following:

- If you use the Statistics and Machine Learning Toolbox library block, you can use the **Fixed-Point Tool** to convert a floating-point model to fixed point.

- Support for variable-size arrays must be enabled for a MATLAB Function block with the `predict` function.
- If you use a MATLAB Function block, you can use MATLAB functions for preprocessing or post-processing before or after predictions in the same MATLAB Function block.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using Simulink® Coder™.

Fixed-Point Conversion
Design and simulate fixed-point systems using Fixed-Point Designer™.

See Also

Blocks
ClassificationTree Predict | RegressionEnsemble Predict | RegressionSVM Predict

Objects
`CompactRegressionTree` | `RegressionTree`

Functions
`fitrtree` | `predict`

Topics
“Predict Responses Using RegressionSVM Predict Block” on page 32-115
“Predict Responses Using RegressionEnsemble Predict Block” on page 32-137
“Predict Class Labels Using MATLAB Function Block” on page 32-40

Introduced in R2021a

RegressionTreeCoderConfigurer

Coder configurer of binary decision tree model for regression

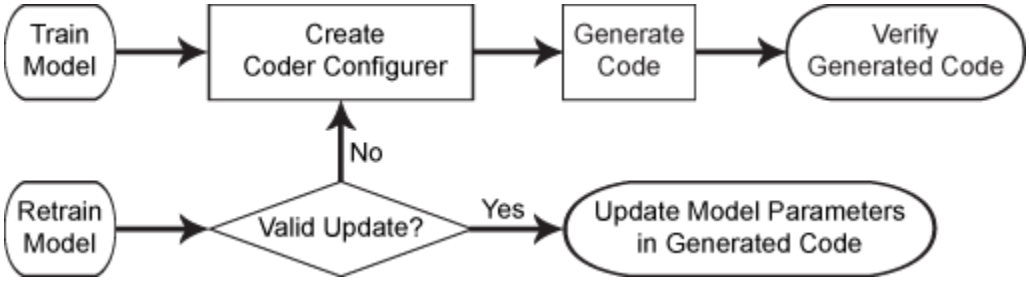
Description

A `RegressionTreeCoderConfigurer` object is a coder configurer of a binary decision tree model for regression (`RegressionTree` or `CompactRegressionTree`).

A coder configurer offers convenient features to configure code generation options, generate C/C++ code, and update model parameters in the generated code.

- Configure code generation options and specify the coder attributes for tree model parameters by using object properties.
- Generate C/C++ code for the `predict` and `update` functions of the regression tree model by using `generateCode`. Generating C/C++ code requires MATLAB Coder.
- Update model parameters in the generated C/C++ code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain the tree model with new data or settings. Before updating model parameters, use `validatedUpdateInputs` to validate and extract the model parameters to update.

This flow chart shows the code generation workflow using a coder configurer.



For the code generation usage notes and limitations of a regression tree model, see the Code Generation sections of `CompactRegressionTree`, `predict`, and `update`.

Creation

After training a regression tree model by using `fitrtree`, create a coder configurer for the model by using `learnerCoderConfigurer`. Use the properties of a coder configurer to specify the coder attributes of the `predict` and `update` arguments. Then, use `generateCode` to generate C/C++ code based on the specified coder attributes.

Properties

`predict` Arguments

The properties listed in this section specify the coder attributes of the `predict` function arguments in the generated code.

X — Coder attributes of predictor data
`LearnerCoderInput` object

Coder attributes of the predictor data to pass to the generated C/C++ code for the `predict` function of the regression tree model, specified as a `LearnerCoderInput` on page 33-5445 object.

When you create a coder configurer by using the `learnerCoderConfigurer` function, the input argument `X` determines the default values of the `LearnerCoderInput` coder attributes:

- `SizeVector` — The default value is the array size of the input `X`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `single` or `double`. The default data type depends on the data type of the input `X`.
- `Tunability` — This value must be `true`, meaning that `predict` in the generated C/C++ code always includes predictor data as an input.

You can modify the coder attributes by using dot notation. For example, to generate C/C++ code that accepts predictor data with 100 observations of three predictor variables, specify these coder attributes of `X` for the coder configurer `configurer`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [0 0];
```

`[0 0]` indicates that the first and second dimensions of `X` (number of observations and number of predictor variables, respectively) have fixed sizes.

To allow the generated C/C++ code to accept predictor data with up to 100 observations, specify these coder attributes of `X`:

```
configurer.X.SizeVector = [100 3];
configurer.X.DataType = 'double';
configurer.X.VariableDimensions = [1 0];
```

`[1 0]` indicates that the first dimension of `X` (number of observations) has a variable size and the second dimension of `X` (number of predictor variables) has a fixed size. The specified number of observations, 100 in this example, becomes the maximum allowed number of observations in the generated C/C++ code. To allow any number of observations, specify the bound as `Inf`.

NumOutputs — Number of outputs in `predict`
1 (default) | 2

Number of output arguments to return from the generated C/C++ code for the `predict` function of the regression tree model, specified as 1 or 2.

The output arguments of `predict` are `Yfit` (predicted responses) and `node` (node numbers for predictions), in that order. `predict` in the generated C/C++ code returns the first `n` outputs of the `predict` function, where `n` is the `NumOutputs` value.

After creating the coder configurer `configurer`, you can specify the number of outputs by using dot notation.

```
configurer.NumOutputs = 2;
```

The `NumOutputs` property is equivalent to the `'-nargout'` compiler option of `codegen`. This option specifies the number of output arguments in the entry-point function of code generation. The object function `generateCode` generates two entry-point functions—`predict.m` and `update.m` for the `predict` and `update` functions of a regression tree model, respectively—and generates C/C++ code for the two entry-point functions. The specified value for the `NumOutputs` property corresponds to the number of output arguments in the entry-point function `predict.m`.

Data Types: `double`

`update` Arguments

The properties listed in this section specify the coder attributes of the `update` function arguments in the generated code. The `update` function takes a trained model and new model parameters as input arguments, and returns an updated version of the model that contains the new parameters. To enable updating the parameters in the generated code, you need to specify the coder attributes of the parameters before generating code. Use a `LearnerCoderInput` on page 33-5445 object to specify the coder attributes of each parameter. The default attribute values are based on the model parameters in the input argument `Mdl` of `learnerCoderConfigurer`.

`Children` — Coder attributes of child nodes for each node

`LearnerCoderInput` object

Coder attributes of the child nodes for each node in the tree (`Children` of a regression tree model), specified as a `LearnerCoderInput` on page 33-5445 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 2]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `CutPoint`, `CutPredictorIndex`, and `NodeMean`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`CutPoint` — Coder attributes of cut point for each node

`LearnerCoderInput` object

Coder attributes of the cut point for each node in the tree (`CutPoint` of a regression tree model), specified as a `LearnerCoderInput` on page 33-5445 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 1]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `CutPredictorIndex`, and `NodeMean`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`CutPredictorIndex` — Coder attributes of cut predictor index for each node
`LearnerCoderInput` object

Coder attributes of the cut predictor index for each node in the tree (`CutPredictorIndex` of a regression tree model), specified as a `LearnerCoderInput` on page 33-5445 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 1]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `CutPoint`, and `NodeMean`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

`NodeMean` — Coder attributes of mean response value for each node
`LearnerCoderInput` object

Coder attributes of the mean response value for each node in the tree (`NodeMean` of a regression tree model), specified as a `LearnerCoderInput` on page 33-5445 object.

The default attribute values of the `LearnerCoderInput` object are based on the input argument `Mdl` of `learnerCoderConfigurer`:

- `SizeVector` — The default value is `[nd 1]`, where `nd` is the number of nodes in `Mdl`.
- `VariableDimensions` — This value is `[0 0]`(default) or `[1 0]`.
  - `[0 0]` indicates that the array size is fixed as specified in `SizeVector`.
  - `[1 0]` indicates that the array has variable-size rows and fixed-size columns. In this case, the first value of `SizeVector` is the upper bound for the number of rows, and the second value of `SizeVector` is the number of columns.
- `DataType` — This value is `'single'` or `'double'`. The default data type is consistent with the data type of the training data you use to train `Mdl`.
- `Tunability` — This value must be `true`.

If you modify the first dimension of `SizeVector` to be `newnd`, then the software modifies the first dimension of the `SizeVector` attribute to be `newnd` for the properties `Children`, `CutPoint`, and `CutPredictorIndex`. Similarly, if you modify the first dimension of `VariableDimensions` to be `1`, then the software modifies the first dimension of the `VariableDimensions` attribute to be `1` for these properties.

Other Configurer Options

`OutputFileName` — File name of generated C/C++ code
`'RegressionTreeModel'` (default) | character vector

File name of the generated C/C++ code, specified as a character vector.

The object function `generateCode` of `RegressionTreeCoderConfigurer` generates C/C++ code using this file name.

The file name must not contain spaces because they can lead to code generation failures in certain operating system configurations. Also, the name must be a valid MATLAB function name.

After creating the coder configurer `configurer`, you can specify the file name by using dot notation.

```
configurer.OutputFileName = 'myModel';
```

Data Types: `char`

`Verbose` — Verbosity level
`true` (logical 1) (default) | `false` (logical 0)

Verbosity level, specified as `true` (logical 1) or `false` (logical 0). The verbosity level controls the display of notification messages at the command line.

Value	Description
`true` (logical 1)	The software displays notification messages when your changes to the coder attributes of a parameter result in changes for other dependent parameters.

Value	Description
`false` (logical 0)	The software does not display notification messages.

To enable updating machine learning model parameters in the generated code, you need to configure the coder attributes of the parameters before generating code. The coder attributes of parameters are dependent on each other, so the software stores the dependencies as configuration constraints. If you modify the coder attributes of a parameter by using a coder configurer, and the modification requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters. The verbosity level determines whether or not the software displays notification messages for these subsequent changes.

After creating the coder configurer `configurer`, you can modify the verbosity level by using dot notation.

```
configurer.Verbose = false;
```

Data Types: `logical`

Options for Code Generation Customization

To customize the code generation workflow, use the `generateFiles` function and the following three properties with `codegen`, instead of using the `generateCode` function.

After generating the two entry-point function files (`predict.m` and `update.m`) by using the `generateFiles` function, you can modify these files according to your code generation workflow. For example, you can modify the `predict.m` file to include data preprocessing, or you can add these entry-point functions to another code generation project. Then, you can generate C/C++ code by using the `codegen` function and the `codegen` arguments appropriate for the modified entry-point functions or code generation project. Use the three properties described in this section as a starting point to set the `codegen` arguments.

`CodeGenerationArguments` — `codegen` arguments
cell array

This property is read-only.

`codegen` arguments, specified as a cell array.

This property enables you to customize the code generation workflow. Use the `generateCode` function if you do not need to customize your workflow.

Instead of using `generateCode` with the coder configurer `configurer`, you can generate C/C++ code as follows:

```
generateFiles(configurer)
cgArgs = configurer.CodeGenerationArguments;
codegen(cgArgs{:})
```

If you customize the code generation workflow, modify `cgArgs` accordingly before calling `codegen`.

If you modify other properties of `configurer`, the software updates the `CodeGenerationArguments` property accordingly.

Data Types: `cell`

`PredictInputs` — Input argument of `predict`
cell array of a `coder.PrimitiveType` object

This property is read-only.

Input argument of the entry-point function `predict.m` for code generation, specified as a cell array of a `coder.PrimitiveType` object. The `coder.PrimitiveType` object includes the coder attributes of the predictor data stored in the `X` property.

If you modify the coder attributes of the predictor data, then the software updates the `coder.PrimitiveType` object accordingly.

The `coder.PrimitiveType` object in `PredictInputs` is equivalent to `configurer.CodeGenerationArguments{6}` for the coder configurer `configurer`.

Data Types: `cell`

`UpdateInputs` — List of tunable input arguments of `update`
cell array of a structure including `coder.PrimitiveType` objects

This property is read-only.

List of the tunable input arguments of the entry-point function `update.m` for code generation, specified as a cell array of a structure including `coder.PrimitiveType` objects. Each `coder.PrimitiveType` object includes the coder attributes of a tunable machine learning model parameter.

If you modify the coder attributes of a model parameter by using the coder configurer properties (`update` Arguments on page 33-5435 properties), then the software updates the corresponding `coder.PrimitiveType` object accordingly. If you specify the `Tunability` attribute of a machine learning model parameter as `false`, then the software removes the corresponding `coder.PrimitiveType` object from the `UpdateInputs` list.

The structure in `UpdateInputs` is equivalent to `configurer.CodeGenerationArguments{3}` for the coder configurer `configurer`.

Data Types: `cell`

Object Functions

generateCode	Generate C/C++ code using coder configurer
generateFiles	Generate MATLAB files for code generation using coder configurer
validatedUpdateInputs	Validate and extract machine learning model parameters to update

Examples

Generate Code Using Coder Configurer

Train a machine learning model, and then generate code for the `predict` and `update` functions of the model by using a coder configurer.

Load the `carbig` data set, which contains car data, and train a regression tree model.

```
load carbig
X = [Displacement Horsepower Weight];
```

```
Y = MPG;
Mdl = fitrtree(X,Y);
```

`Mdl` is a `RegressionTree` object.

Create a coder configurer for the `RegressionTree` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X)
```

```
configurer =
  RegressionTreeCoderConfigurer with properties:

   Update Inputs:
             Children: [1x1 LearnerCoderInput]
             NodeMean: [1x1 LearnerCoderInput]
             CutPoint: [1x1 LearnerCoderInput]
    CutPredictorIndex: [1x1 LearnerCoderInput]

   Predict Inputs:
                   X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
           NumOutputs: 1
       OutputFileName: 'RegressionTreeModel'


  Properties, Methods
```

`configurer` is a `RegressionTreeCoderConfigurer` object, which is a coder configurer of a `RegressionTree` object.

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the regression tree model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionTreeModel` folder.
- Copy the MEX function to the current folder.

Display the contents of the `predict.m`, `update.m`, and `initialize.m` files by using the `type` function.

```
type predict.m
```

```
function varargout = predict(X,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:29
[varargout{1:nargout}] = initialize('predict',X,varargin{:});
end
```

```
type update.m
```

```
function update(varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:29
initialize('update',varargin{:});
end
```

```
type initialize.m
```

```
function [varargout] = initialize(command,varargin) %#codegen
% Autogenerated by MATLAB, 25-Feb-2021 14:01:29
coder.inline('always')
persistent model
if isempty(model)
    model = loadLearnerForCoder('RegressionTreeModel.mat');
end
switch(command)
    case 'update'
        % Update struct fields: Children
        %                       NodeMean
        %                       CutPoint
        %                       CutPredictorIndex
        model = update(model,varargin{:});
    case 'predict'
        % Predict Inputs: X
        X = varargin{1};
        if nargin == 2
            [varargout{1:nargout}] = predict(model,X);
        else
            PVPairs = cell(1,nargin-2);
            for i = 1:nargin-2
                PVPairs{1,i} = varargin{i+1};
            end
            [varargout{1:nargout}] = predict(model,X,PVPairs{:});
        end
end
end
```

Update Parameters of Regression Tree Model in Generated Code

Train a regression tree using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Load the `carbig` data set, and train a regression tree model using half of the observations.

```
load carbig
X = [Displacement Horsepower Weight];
Y = MPG;

rng('default') % For reproducibility
n = length(Y);
idxTrain = randsample(n,n/2);
XTrain = X(idxTrain,:);
YTrain = Y(idxTrain);

Mdl = fitrtree(XTrain,YTrain);
```

`Mdl` is a `RegressionTree` object.

Create Coder Configurer

Create a coder configurer for the `RegressionTree` model by using `learnerCoderConfigurer`. Specify the predictor data `XTrain`. The `learnerCoderConfigurer` function uses the input `XTrain` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted responses and node numbers for the predictions.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'NumOutputs',2);
```

`configurer` is a `RegressionTreeCoderConfigurer` object, which is a coder configurer of a `RegressionTree` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the regression tree model parameters so that you can update the parameters in the generated code after retraining the model.

Specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 3];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

The size of the first dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 3 predictors, the value of the `SizeVector` attribute must be `3` and the value of the `VariableDimensions` attribute must be `0`.

If you retrain the tree model using new data or different settings, the number of nodes in the tree can vary. Therefore, specify the first dimension of the `SizeVector` attribute of one of these properties so

that you can update the number of nodes in the generated code: `Children`, `CutPoint`, `CutPredictorIndex`, or `NodeMean`. The software then modifies the other properties automatically.

For example, set the first value of the `SizeVector` attribute of the `NodeMean` property to `Inf`. The software modifies the `SizeVector` and `VariableDimensions` attributes of `Children`, `CutPoint`, and `CutPredictorIndex` to match the new upper bound on the number of nodes in the tree. Additionally, the first value of the `VariableDimensions` attribute of `NodeMean` changes to `1`.

```
configurer.NodeMean.SizeVector = [Inf 1];
```

```
SizeVector attribute for Children has been modified to satisfy configuration constraints.
SizeVector attribute for CutPoint has been modified to satisfy configuration constraints.
SizeVector attribute for CutPredictorIndex has been modified to satisfy configuration constraints
VariableDimensions attribute for Children has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPoint has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPredictorIndex has been modified to satisfy configuration co
```

```
configurer.NodeMean.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the regression tree model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionTreeModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[Yfit,node] = predict(Mdl,XTrain);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',XTrain);
```

Compare `Yfit` to `Yfit_mex` and `node` to `node_mex`.

```
max(abs(Yfit-Yfit_mex),[],'all')

ans = 0

isequal(node,node_mex)

ans = logical
   1
```

In general, `Yfit_mex` might include round-off differences compared to `Yfit`. In this case, the comparison confirms that `Yfit` and `Yfit_mex` are equal.

`isequal` returns logical 1 (`true`) if all the input arguments are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same node numbers.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrtree(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionTreeModel('update',params)
```

Verify Generated Code

Compare the output arguments from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[Yfit,node] = predict(retrainedMdl,X);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',X);

max(abs(Yfit-Yfit_mex),[],'all')

ans = 0

isequal(node,node_mex)

ans = logical
   1
```

The comparison confirms that the predicted responses and node numbers are equal.

More About

`LearnerCoderInput` Object

A coder configurer uses a `LearnerCoderInput` object to specify the coder attributes of `predict` and `update` input arguments.

A `LearnerCoderInput` object has the following attributes to specify the properties of an input argument array in the generated code.

Attribute Name	Description
`SizeVector`	Array size if the corresponding `VariableDimensions` value is `false`. Upper bound of the array size if the corresponding `VariableDimensions` value is `true`. To allow an unbounded array, specify the bound as `Inf`.
`VariableDimensions`	Indicator specifying whether each dimension of the array has a variable size or fixed size, specified as `true` (logical 1) or `false` (logical 0): • A value of `true` (logical 1) means that the corresponding dimension has a variable size. • A value of `false` (logical 0) means that the corresponding dimension has a fixed size.
`DataType`	Data type of the array
`Tunability`	Indicator specifying whether or not `predict` or `update` includes the argument as an input in the generated code, specified as `true` (logical 1) or `false` (logical 0). If you specify other attribute values when `Tunability` is `false`, the software sets `Tunability` to `true`.

After creating a coder configurer, you can modify the coder attributes by using dot notation. For example, specify the coder attributes of the `CutPoint` property of the coder configurer `configurer`:

```
configurer.CutPoint.SizeVector = [40 1];
configurer.CutPoint.VariableDimensions = [1 0];
```

If you specify the verbosity level (`Verbose`) as `true` (default), then the software displays notification messages when you modify the coder attributes of a machine learning model parameter and the modification changes the coder attributes of other dependent parameters.

See Also

`CompactRegressionTree` | `RegressionTree` | `learnerCoderConfigurer` | `predict` | `update`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2019b

regstats

Regression diagnostics

Syntax

```
regstats(y,X,model)
stats = regstats(...)
stats = regstats(y,X,model,whichstats)
```

Description

`regstats(y,X,model)` performs a multilinear regression of the responses in `y` on the predictors in `X`. `X` is an *n*-by-*p* matrix of *p* predictors at each of *n* observations. `y` is an *n*-by-1 vector of observed responses.

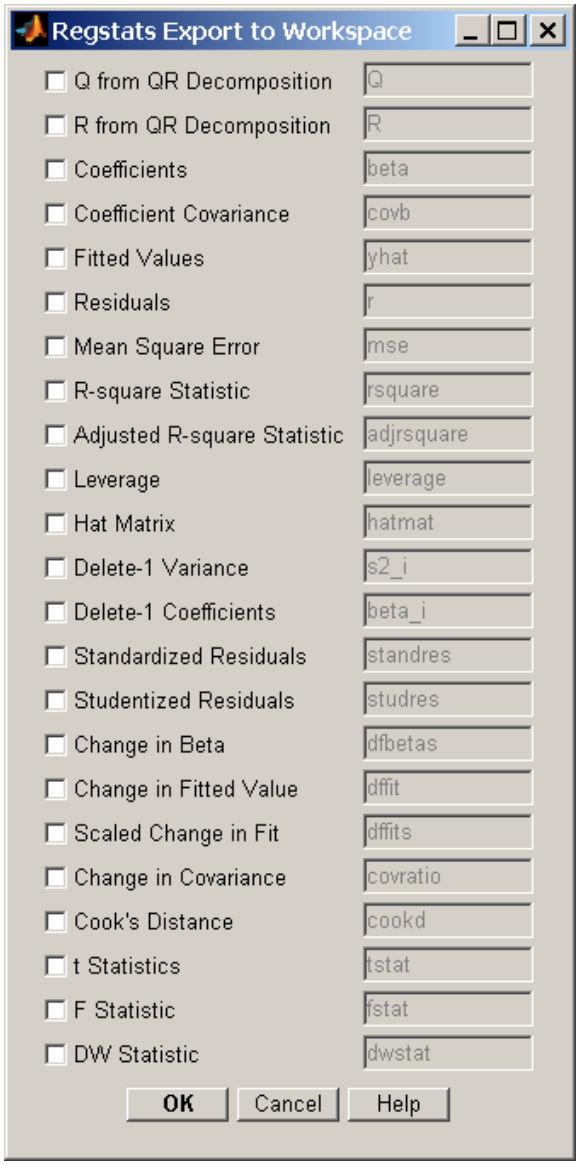
Note By default, `regstats` adds a first column of 1s to `X`, corresponding to a constant term in the model. Do not enter a column of 1s directly into `X`.

The optional input `model` controls the regression model. By default, `regstats` uses a linear additive model with a constant term. `model` can be any one of the following:

- `'linear'` — Constant and linear terms (the default)
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

Alternatively, `model` can be a matrix of model terms accepted by the `x2fx` function. See `x2fx` for a description of this matrix and for a description of the order in which terms appear. You can use this matrix to specify other models including ones without a constant term.

With this syntax, the function displays a graphical user interface (GUI) with a list of diagnostic statistics, as shown in the following figure.



When you select check boxes corresponding to the statistics you want to compute and click **OK**, `regstats` returns the selected statistics to the MATLAB workspace. The names of the workspace variables are displayed on the right-hand side of the interface. You can change the name of the workspace variable to any valid MATLAB variable name.

`stats = regstats(...)` creates the structure `stats`, whose fields contain all of the diagnostic statistics for the regression. This syntax does not open the GUI. The fields of `stats` are listed in the following table.

Field	Description
`Q`	*Q* from the *QR* decomposition of the design matrix
`R`	*R* from the *QR* decomposition of the design matrix
`beta`	Regression coefficients

Field	Description
`covb`	Covariance of regression coefficients
`yhat`	Fitted values of the response data
`r`	Residuals
`mse`	Mean squared error
`rsquare`	R^2 statistic
`adjrsquare`	Adjusted R^2 statistic
`leverage`	Leverage
`hatmat`	Hat matrix
`s2_i`	Delete-1 variance
`beta_i`	Delete-1 coefficients
`standres`	Standardized residuals
`studres`	Studentized residuals
`dfbetas`	Scaled change in regression coefficients
`dffit`	Change in fitted values
`dffits`	Scaled change in fitted values
`covratio`	Change in covariance
`cookd`	Cook's distance
`tstat`	*t* statistics and *p*-values for coefficients
`fstat`	*F* statistic and *p*-value
`dwstat`	Durbin-Watson statistic and *p*-value

Note that the fields names of `stats` correspond to the names of the variables returned to the MATLAB workspace when you use the GUI. For example, `stats.beta` corresponds to the variable `beta` that is returned when you select **Coefficients** in the GUI and click **OK**.

`stats = regstats(y,X,model,whichstats)` returns only the statistics that you specify in *`whichstats`*. *`whichstats`* can be a single character vector such as `'leverage'`, a string array such as `["leverage","standres","studres"]`, or a cell array of character vectors such as `{'leverage','standres','studres'}`. Set *`whichstats`* to `'all'` to return all of the statistics.

Note The *F* statistic is computed under the assumption that the model contains a constant term. It is not correct for models without a constant. The R^2 statistic can be negative for models without a constant, which indicates that the model is not appropriate for the data.

Examples

Open the `regstats` GUI using data from `hald.mat`:

```
load hald
regstats(heat,ingredients,'linear');
```

Select **Fitted Values** and **Residuals** in the GUI:

Click **OK** to export the fitted values and residuals to the MATLAB workspace in variables named `yhat` and `r`, respectively.

You can create the same variables using the `stats` output, without opening the GUI:

```
whichstats = {'yhat','r'};
stats = regstats(heat,ingredients,'linear',whichstats);
yhat = stats.yhat;
r = stats.r;
```

Tips

- `regstats` treats `NaN` values in `X` or `y` as missing values. `regstats` omits observations with missing values from the regression fit.

References

[1] Belsley, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics*. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

[2] Chatterjee, S., and A. S. Hadi. “Influential Observations, High Leverage Points, and Outliers in Linear Regression.” *Statistical Science*. Vol. 1, 1986, pp. 379–416.

[3] Cook, R. D., and S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman & Hall/CRC Press, 1983.

[4] Goodall, C. R. “Computation Using the QR Decomposition.” *Handbook in Statistics.* Vol. 9, Amsterdam: Elsevier/North-Holland, 1993.

See Also

`LinearModel` | `fitlm` | `stepwiselm`

Topics

“Interpret Linear Regression Results” on page 11-50
“Linear Regression Workflow” on page 11-35

Introduced before R2006a

regularize

Find weights to minimize resubstitution error plus penalty term

Syntax

```
ens1 = regularize(ens)
ens1 = regularize(ens,Name,Value)
```

Description

`ens1 = regularize(ens)` finds optimal weights for learners in `ens` by lasso regularization. `regularize` returns a regression ensemble identical to `ens`, but with a populated `Regularization` property.

`ens1 = regularize(ens,Name,Value)` computes optimal weights with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression ensemble, created by `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`lambda`

Vector of nonnegative regularization parameter values for lasso. For the default setting of `lambda`, `regularize` calculates the smallest value `lambda_max` for which all optimal weights for learners are `0`. The default value of `lambda` is a vector including `0` and nine exponentially-spaced numbers from `lambda_max/1000` to `lambda_max`.

Default: `[0 logspace(log10(lambda_max/1000),log10(lambda_max),9)]`

`MaxIter`

Maximum number of iterations allowed, specified as a positive integer. If the algorithm executes `MaxIter` iterations before reaching the convergence tolerance, then the function stops iterating and returns a warning message. The function can return more than one warning when either `npass` or the number of `lambda` values is greater than 1.

Default: `1e3`

`npass`

Maximal number of passes for lasso optimization, a positive integer.

Default: `10`

`reltol`

Relative tolerance on the regularized loss for lasso, a numeric positive scalar.

Default: `1e-3`

`verbose`

Verbosity level, either `0` or `1`. When set to `1`, `regularize` displays more information as it runs.

Default: `0`

Output Arguments

`ens1`

A regression ensemble. Usually you set `ens1` to the same name as `ens`.

Examples

Regularize Ensemble of Bagged Trees

Regularize an ensemble of bagged trees.

Generate sample data.

```
rng(10,'twister') % For reproducibility
X = rand(2000,20);
Y = repmat(-1,2000,1);
Y(sum(X(:,1:5),2)>2.5) = 1;
```

You can create a bagged classification ensemble of 300 trees from the sample data.

```
bag = fitrensemble(X,Y,'Method','Bag','NumLearningCycles',300);
```

`fitrensemble` uses a default template tree object `templateTree()` as a weak learner when `'Method'` is `'Bag'`. In this example, for reproducibility, specify `'Reproducible',true` when you create a tree template object, and then use the object as a weak learner.

```
t = templateTree('Reproducible',true); % For reproducibiliy of random predictor selections
bag = fitrensemble(X,Y,'Method','Bag','NumLearningCycles',300,'Learners',t);
```

Regularize the ensemble of bagged regression trees.

```
bag = regularize(bag,'lambda',[0.001 0.1],'verbose',1);
```

```
Starting lasso minimization for Lambda=0.001. Initial MSE=0.109923.
    Lasso minimization completed pass 1 for Lambda=0.001
        MSE = 0.086912
        Relative change in MSE = 0.264768
        Number of learners with non-zero weights = 15
    Lasso minimization completed pass 2 for Lambda=0.001
        MSE = 0.0670602
        Relative change in MSE = 0.296029
```

```
        Number of learners with non-zero weights = 34
    Lasso minimization completed pass 3 for Lambda=0.001
        MSE = 0.0623931
        Relative change in MSE = 0.0748019
        Number of learners with non-zero weights = 51
    Lasso minimization completed pass 4 for Lambda=0.001
        MSE = 0.0605444
        Relative change in MSE = 0.0305348
        Number of learners with non-zero weights = 70
    Lasso minimization completed pass 5 for Lambda=0.001
        MSE = 0.0599666
        Relative change in MSE = 0.00963517
        Number of learners with non-zero weights = 94
    Lasso minimization completed pass 6 for Lambda=0.001
        MSE = 0.0598835
        Relative change in MSE = 0.00138719
        Number of learners with non-zero weights = 105
    Lasso minimization completed pass 7 for Lambda=0.001
        MSE = 0.0598608
        Relative change in MSE = 0.000379227
        Number of learners with non-zero weights = 113
    Lasso minimization completed pass 8 for Lambda=0.001
        MSE = 0.0598586
        Relative change in MSE = 3.72856e-05
        Number of learners with non-zero weights = 115
    Lasso minimization completed pass 9 for Lambda=0.001
        MSE = 0.0598587
        Relative change in MSE = 6.42954e-07
        Number of learners with non-zero weights = 115
    Lasso minimization completed pass 10 for Lambda=0.001
        MSE = 0.0598587
        Relative change in MSE = 4.53658e-08
        Number of learners with non-zero weights = 115
    Completed lasso minimization for Lambda=0.001.
    Resubstitution MSE changed from 0.109923 to 0.0598587.
    Number of learners reduced from 300 to 115.
Starting lasso minimization for Lambda=0.1. Initial MSE=0.109923.
    Lasso minimization completed pass 1 for Lambda=0.1
        MSE = 0.104917
        Relative change in MSE = 0.0477191
        Number of learners with non-zero weights = 12
    Lasso minimization completed pass 2 for Lambda=0.1
        MSE = 0.0851031
        Relative change in MSE = 0.232821
        Number of learners with non-zero weights = 30
    Lasso minimization completed pass 3 for Lambda=0.1
        MSE = 0.081245
        Relative change in MSE = 0.0474877
        Number of learners with non-zero weights = 40
    Lasso minimization completed pass 4 for Lambda=0.1
        MSE = 0.0796749
        Relative change in MSE = 0.0197067
        Number of learners with non-zero weights = 53
    Lasso minimization completed pass 5 for Lambda=0.1
        MSE = 0.0788411
        Relative change in MSE = 0.0105746
        Number of learners with non-zero weights = 64
    Lasso minimization completed pass 6 for Lambda=0.1
```

```
        MSE = 0.0784959
        Relative change in MSE = 0.00439793
        Number of learners with non-zero weights = 81
    Lasso minimization completed pass 7 for Lambda=0.1
        MSE = 0.0784429
        Relative change in MSE = 0.000676468
        Number of learners with non-zero weights = 88
    Lasso minimization completed pass 8 for Lambda=0.1
        MSE = 0.078447
        Relative change in MSE = 5.24449e-05
        Number of learners with non-zero weights = 88
    Completed lasso minimization for Lambda=0.1.
    Resubstitution MSE changed from 0.109923 to 0.078447.
    Number of learners reduced from 300 to 88.
```

`regularize` reports on its progress.

Inspect the resulting regularization structure.

```
bag.Regularization
```

```
ans = struct with fields:
               Method: 'Lasso'
       TrainedWeights: [300x2 double]
               Lambda: [1.0000e-03 0.1000]
    ResubstitutionMSE: [0.0599 0.0784]
       CombineWeights: @classreg.learning.combiner.WeightedSum
```

Check how many learners in the regularized ensemble have positive weights. These are the learners included in a shrunken ensemble.

```
sum(bag.Regularization.TrainedWeights > 0)
```

```
ans = 1×2

   115    88
```

Shrink the ensemble using the weights from `Lambda = 0.1`.

```
cmp = shrink(bag,'weightcolumn',2)
```

```
cmp =
  CompactRegressionEnsemble
             ResponseName: 'Y'
    CategoricalPredictors: []
        ResponseTransform: 'none'
               NumTrained: 88


  Properties, Methods
```

The compact ensemble contains `87` members, less than 1/3 of the original `300`.

More About

Lasso

The lasso algorithm finds an optimal set of learner weights α_t that minimize

$$\sum_{n=1}^{N} w_n g\left(\left(\sum_{t=1}^{T} \alpha_t h_t(x_n)\right), y_n\right) + \lambda \sum_{t=1}^{T} |\alpha_t| .$$

Here

- $\lambda \geq 0$ is a parameter you provide, called the lasso parameter.
- h_t is a weak learner in the ensemble trained on N observations with predictors x_n, responses y_n, and weights w_n.
- $g(f,y) = (f - y)^2$ is the squared error.

See Also

`cvshrink` | `lasso` | `shrink`

Topics

“Ensemble Regularization” on page 18-70

relieff

Rank importance of predictors using ReliefF or RReliefF algorithm

Syntax

```
[idx,weights] = relieff(X,y,k)
[idx,weights] = relieff(X,y,k,Name,Value)
```

Description

`[idx,weights] = relieff(X,y,k)` ranks predictors using either the ReliefF or RReliefF algorithm with `k` nearest neighbors. The input matrix `X` contains predictor variables, and the vector `y` contains a response vector. The function returns `idx`, which contains the indices of the most important predictors, and `weights`, which contains the weights of the predictors.

If `y` is numeric, `relieff` performs RReliefF analysis for regression by default. Otherwise, `relieff` performs ReliefF analysis for classification using `k` nearest neighbors per class. For more information on ReliefF and RReliefF, see “Algorithms” on page 33-5461.

`[idx,weights] = relieff(X,y,k,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, `'updates',10` sets the number of observations randomly selected for computing weights to 10.

Examples

Determine Important Predictors

Load the sample data.

```
load fisheriris
```

Find the important predictors using 10 nearest neighbors.

```
[idx,weights] = relieff(meas,species,10)
```

```
idx = 1×4

     4     3     1     2


weights = 1×4

    0.1399    0.1226    0.3590    0.3754
```

`idx` shows the predictor numbers listed according to their ranking. The fourth predictor is the most important, and the second predictor is the least important. `weights` gives the weight values in the same order as the predictors. The first predictor has a weight of 0.1399, and the fourth predictor has a weight of 0.3754.

Rank Predictors by Importance

Load the sample data.

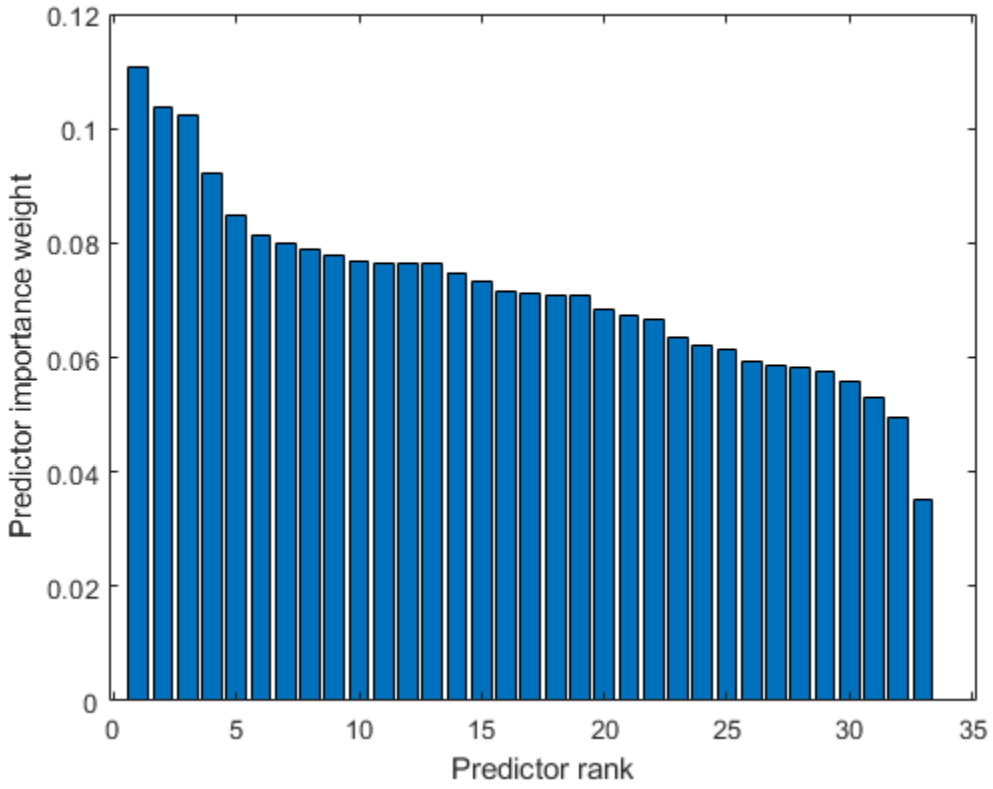
```
load ionosphere
```

Rank the predictors based on importance using 10 nearest neighbors.

```
[idx,weights] = relieff(X,Y,10);
```

Create a bar plot of predictor importance weights.

```
bar(weights(idx))
xlabel('Predictor rank')
ylabel('Predictor importance weight')
```



Select the top 5 most important predictors. Find the columns of these predictors in X.

```
idx(1:5)
```

```
ans = 1×5

    24     3     8     5    14
```

The 24th column of X is the most important predictor of Y.

Determine Important Categorical Predictors

Rank categorical predictors using `relieff`.

Load the sample data.

```
load carbig
```

Convert the categorical predictor variables `Mfg`, `Model`, and `Origin` to numerical values, and combine them into an input matrix. Specify the response variable `MPG`.

```
X = [grp2idx(Mfg) grp2idx(Model) grp2idx(Origin)];
y = MPG;
```

Find the ranks and weights of the predictor variables using 10 nearest neighbors and treating the data in `X` as categorical.

```
[idx,weights] = relieff(X,y,10,'categoricalx','on')
```

```
idx = 1×3

     2     3     1


weights = 1×3

   -0.0019    0.0501    0.0114
```

The `Model` predictor is the most important in predicting `MPG`. The `Mfg` variable has a negative weight, indicating it is not a good predictor of `MPG`.

Input Arguments

X — Predictor data

numeric matrix

Predictor data, specified as a numeric matrix. Each row of `X` corresponds to one observation, and each column corresponds to one variable.

Data Types: `single` | `double`

y — Response data

numeric vector | categorical vector | logical vector | character array | string array | cell array of character vectors

Response data, specified as a numeric vector, categorical vector, logical vector, character array, string array, or cell array of character vectors.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

k — Number of nearest neighbors

positive integer scalar

Number of nearest neighbors, specified as a positive integer scalar.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `relieff(X,y,5,'method','classification','categoricalx','on')` specifies 5 nearest neighbors and treats the response variable and predictor data as categorical.

`method` — Method for computing weights

`'regression'` | `'classification'`

Method for computing weights, specified as the comma-separated pair consisting of `'method'` and either `'regression'` or `'classification'`. If `y` is numeric, `'regression'` is the default method. Otherwise, `'classification'` is the default.

Example: `'method','classification'`

`prior` — Prior probabilities for each class

`'empirical'` (default) | `'uniform'` | numeric vector | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'prior'` and a value in this table.

Value	Description
`'empirical'`	The class probabilities are determined from class frequencies in `y`.
`'uniform'`	All class probabilities are equal.
numeric vector	One value exists for each distinct group name.
structure	A structure `S` with two fields: • `S.group` contains the group names as a variable of the same type as `y`. • `S.prob` contains a vector of corresponding probabilities.

Example: `'prior','uniform'`

Data Types: `single` | `double` | `char` | `string` | `struct`

`updates` — Number of observations for computing weights

`'all'` (default) | positive integer scalar

Number of observations to select at random for computing weights, specified as the comma-separated pair consisting of `'updates'` and either `'all'` or a positive integer scalar. By default, `relieff` uses all observations.

Example: `'updates',25`

Data Types: `single` | `double` | `char` | `string`

`categoricalx` — Categorical predictors flag
`'off'` (default) | `'on'`

Categorical predictors flag, specified as the comma-separated pair consisting of `'categoricalx'` and either `'on'` or `'off'`. If you specify `'on'`, then `relieff` treats all predictors in `X` as categorical. Otherwise, it treats all predictors in `X` as numeric. You cannot mix numeric and categorical predictors.

Example: `'categoricalx','on'`

`sigma` — Distance scaling factor
numeric positive scalar

Distance scaling factor, specified as the comma-separated pair consisting of `'sigma'` and a numeric positive scalar. For observation i, influence on the predictor weight from its nearest neighbor j is multiplied by $e^{-(\text{rank}(i,j)/\text{sigma})^2}$. rank($i$,$j$) is the position of the jth observation among the nearest neighbors of the ith observation, sorted by distance. The default is `Inf` for classification (all nearest neighbors have the same influence) and 50 for regression.

Example: `'sigma',20`

Data Types: `single` | `double`

Output Arguments

`idx` — Indices of predictors ordered by predictor importance
numeric vector

Indices of predictors in `X` ordered by predictor importance, returned as a numeric vector. For example, if `idx(3)` is `5`, then the third most important predictor is the fifth column in `X`.

Data Types: `double`

`weights` — Weights of predictors
numeric vector

Weights of the predictors, returned as a numeric vector. The values in `weights` have the same order as the predictors in `X`. `weights` range from `−1` to `1`, with large positive weights assigned to important predictors.

Data Types: `double`

Tips

- Predictor ranks and weights usually depend on `k`. If you set `k` to 1, then the estimates can be unreliable for noisy data. If you set `k` to a value comparable with the number of observations (rows) in `X`, `relieff` can fail to find important predictors. You can start with `k = 10` and investigate the stability and reliability of `relieff` ranks and weights for various values of `k`.
- `relieff` removes observations with `NaN` values.

Algorithms

ReliefF

ReliefF finds the weights of predictors in the case where `y` is a multiclass categorical variable. The algorithm penalizes the predictors that give different values to neighbors of the same class, and rewards predictors that give different values to neighbors of different classes.

ReliefF first sets all predictor weights W_j to 0. Then, the algorithm iteratively selects a random observation x_r, finds the k-nearest observations to x_r for each class, and updates, for each nearest neighbor x_q, all the weights for the predictors F_j as follows:

If x_r and x_q are in the same class,

$$W_j{}^i = W_j{}^{i-1} - \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq}\,.$$

If x_r and x_q are in different classes,

$$W_j{}^i = W_j{}^{i-1} + \frac{p_{y_q}}{1 - p_{y_r}} \cdot \frac{\Delta_j(x_r, x_q)}{m} \cdot d_{rq}\,.$$

- W_j^i is the weight of the predictor F_j at the ith iteration step.
- p_{y_r} is the prior probability of the class to which x_r belongs, and p_{y_q} is the prior probability of the class to which x_q belongs.
- m is the number of iterations specified by `'updates'`.
- $\Delta_j(x_r, x_q)$ is the difference in the value of the predictor F_j between observations x_r and x_q. Let x_{rj} denote the value of the jth predictor for observation x_r, and let x_{qj} denote the value of the jth predictor for observation x_q.
  - For discrete F_j,

    $$\Delta_j(x_r, x_q) = \begin{cases} 0, & x_{rj} = x_{qj} \\ 1, & x_{rj} \neq x_{qj} \end{cases}.$$

  - For continuous F_j,

    $$\Delta_j(x_r, x_q) = \frac{\left|x_{rj} - x_{qj}\right|}{\max(F_j) - \min(F_j)}\,.$$
- d_{rq} is a distance function of the form

  $$d_{rq} = \frac{\tilde{d}_{rq}}{\sum_{l=1}^{k} \tilde{d}_{rl}}\,.$$

  The distance is subject to the scaling

  $$\tilde{d}_{rq} = e^{-(\mathrm{rank}(r,\,q)/\mathrm{sigma})^2}$$

where rank(r,q) is the position of the qth observation among the nearest neighbors of the rth observation, sorted by distance. k is the number of nearest neighbors, specified by `k`. You can change the scaling by specifying `'sigma'`.

RReliefF

RReliefF works with continuous `y`. Similar to ReliefF, RReliefF also penalizes the predictors that give different values to neighbors with the same response values, and rewards predictors that give different values to neighbors with different response values. However, RReliefF uses intermediate weights to compute the final predictor weights.

Given two nearest neighbors, assume the following:

- W_{dy} is the weight of having different values for the response y.
- W_{dj} is the weight of having different values for the predictor F_j.
- $W_{dy \wedge dj}$ is the weight of having different response values and different values for the predictor F_j.

RReliefF first sets the weights W_{dy}, W_{dj}, $W_{dy \wedge dj}$, and W_j equal to 0. Then, the algorithm iteratively selects a random observation x_r, finds the k-nearest observations to x_r, and updates, for each nearest neighbor x_q, all the intermediate weights as follows:

$$W_{dy}^{i} = W_{dy}^{i-1} + \Delta_y(x_r, x_q) \cdot d_{rq}\,.$$

$$W_{dj}^{i} = W_{dj}^{i-1} + \Delta_j(x_r, x_q) \cdot d_{rq}\,.$$

$$W_{dy \wedge dj}^{i} = W_{dy \wedge dj}^{i-1} + \Delta_y(x_r, x_q) \cdot \Delta_j(x_r, x_q) \cdot d_{rq}\,.$$

- The i and i-1 superscripts denote the iteration step number. m is the number of iterations specified by `'updates'`.
- $\Delta_y(x_r, x_q)$ is the difference in the value of the continuous response y between observations x_r and x_q. Let y_r denote the value of the response for observation x_r, and let y_q denote the value of the response for observation x_q.

$$\Delta_y(x_r, x_q) = \frac{|y_r - y_q|}{\max(y) - \min(y)}\,.$$

- The $\Delta_j(x_r, x_q)$ and d_{rq} functions are the same as for “ReliefF” on page 33-5461.

RReliefF calculates the predictor weights W_j after fully updating all the intermediate weights.

$$W_j = \frac{W_{dy \wedge dj}}{W_{dy}} - \frac{W_{dj} - W_{dy \wedge dj}}{m - W_{dy}}\,.$$

For more information, see [2].

References

[1] Kononenko, I., E. Simec, and M. Robnik-Sikonja. (1997). “Overcoming the myopia of inductive learning algorithms with RELIEFF.” Retrieved from CiteSeerX: `https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.4740`

[2] Robnik-Sikonja, M., and I. Kononenko. (1997). “An adaptation of Relief for attribute estimation in regression.” Retrieved from CiteSeerX: `https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.8381`

[3] Robnik-Sikonja, M., and I. Kononenko. (2003). “Theoretical and empirical analysis of ReliefF and RReliefF.” *Machine Learning*, 53, 23–69.

See Also

`fscmrmr` | `fscnca` | `fsrnca` | `fsulaplacian` | `knnsearch` | `pdist2` | `plotPartialDependence` | `sequentialfs`

Topics

“Introduction to Feature Selection” on page 15-49
“Sequential Feature Selection” on page 15-61

Introduced in R2010b

removeLearners

Remove members of compact classification ensemble

Syntax

```
cens1 = removeLearners(cens,idx)
```

Description

`cens1 = removeLearners(cens,idx)` creates a compact classification ensemble identical to `cens` only without the ensemble members in the `idx` vector.

Input Arguments

cens

Compact classification ensemble, constructed with `compact`.

idx

Vector of positive integers with entries from `1` to `cens.NumTrained`, where `cens.NumTrained` is the number of members in `cens`. `cens1` contains all members of `cens` except those with indices in `idx`.

Typically, you set `idx = j:cens.NumTrained` for some positive integer `j`.

Output Arguments

cens1

Compact classification ensemble, identical to `cens` except `cens1` does not contain those members of `cens` with indices in `idx`.

Examples

Remove Learners from an Ensemble

Create a compact classification ensemble. Compact it further by removing members of the ensemble.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a classification ensemble for the `ionosphere` data using AdaBoostM1. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
```

Create a compact classification ensemble `cens` from `ens`.

```
cens = compact(ens);
```

Remove the last 50 members of the ensemble.

```
idx = cens.NumTrained-49:cens.NumTrained;
cens1 = removeLearners(cens,idx);
```

Tips

- Typically, set `cens1` equal to `cens` to retain just one ensemble.
- Removing learners reduces the memory used by the ensemble and speeds up its predictions.

See Also

`CompactClassificationEnsemble`

Topics

“Classification with Imbalanced Data” on page 18-79

removeLearners

Remove members of compact regression ensemble

Syntax

```
cens1 = removeLearners(cens,idx)
```

Description

`cens1 = removeLearners(cens,idx)` creates a compact regression ensemble identical to `cens` only without the ensemble members in the `idx` vector.

Input Arguments

cens

Compact regression ensemble, constructed with `compact`.

idx

Vector of positive integers with entries from `1` to `cens.NumTrained`, where `cens.NumTrained` is the number of members in `cens`. `cens1` contains the members of `cens` except those with indices in `idx`.

Typically, you set `idx = j:cens.NumTrained` for some positive integer `j`.

Output Arguments

cens1

Compact regression ensemble, identical to `cens` except `cens1` does not contain members of `cens` with indices in `idx`.

Examples

Remove Learners from an Ensemble

Create a compact regression ensemble. Compact it further by removing members of the ensemble.

Load the `carsmall` data set and select `Weight` and `Cylinders` as predictors.

```
load carsmall
X = [Weight Cylinders];
```

Train a regression ensemble using LSBoost. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitrensemble(X,MPG,'Method','LSBoost','Learners',t,...
    'CategoricalPredictors',2);
```

Create a compact classification ensemble `cens` from `ens`.

```
cens = compact(ens);
```

Remove the last 50 members of the ensemble.

```
idx = cens.NumTrained-49:cens.NumTrained;
cens1 = removeLearners(cens,idx);
```

Tips

- Typically, set `cens1` equal to `cens` to retain just one ensemble.
- Removing learners reduces the memory used by the ensemble and speeds up its predictions.

See Also

`CompactRegressionEnsemble`

removeTerms

Remove terms from generalized linear regression model

Syntax

```
NewMdl = removeTerms(mdl,terms)
```

Description

`NewMdl = removeTerms(mdl,terms)` returns a generalized linear regression model fitted using the input data and settings in `mdl` with the terms `terms` removed.

Examples

Remove Terms from Generalized Linear Regression Model

Create a generalized linear regression model using two predictors, and then remove one predictor.

Generate sample data using Poisson random numbers with two underlying predictors `X(:,1)` and `X(:,2)`.

```
rng('default') % For reproducibility
rndvars = randn(100,2);
X = [2 + rndvars(:,1),rndvars(:,2)];
mu = exp(1 + X*[1;2]);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data.

```
mdl = fitglm(X,y,'y ~ x1 + x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     1.0405      0.022122    47.034      0
    x1              0.9968      0.003362    296.49      0
    x2               1.987     0.0063433    313.24      0


100 observations, 97 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 2.95e+05, p-value = 0
```

Remove the second predictor from the model.

```
mdl1 = removeTerms(mdl,'x2')
```

```
mdl1 =
Generalized linear regression model:
    log(y) ~ 1 + x1
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat     pValue
                   ________    _________    ______    ______

    (Intercept)     2.7784      0.014043    197.85      0
    x1              1.1732     0.0033653     348.6      0


100 observations, 98 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.25e+05, p-value = 0
```

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

`terms` — Terms to remove from regression model

character vector or string scalar formula in Wilkinson notation | *t*-by-*p* terms matrix

Terms to remove from the regression model `mdl`, specified as one of the following:

- Character vector or string scalar formula in "Wilkinson Notation" on page 33-5470 representing one or more terms. The variable names in the formula must be valid MATLAB identifiers.
- Terms matrix `T` of size *t*-by-*p*, where *t* is the number of terms and *p* is the number of predictor variables in `mdl`. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

  For example, suppose `mdl` has three variables `A`, `B`, and `C` in that order. Each row of `T` represents one term:

  - `[0 0 0]` — Constant term or intercept
  - `[0 1 0]` — `B`; equivalently, `A^0 * B^1 * C^0`
  - `[1 0 1]` — `A*C`
  - `[2 0 0]` — `A^2`
  - `[0 1 2]` — `B*(C^2)`

`removeTerms` treats a group of indicator variables for a categorical predictor as a single variable. Therefore, you cannot specify an indicator variable to remove from the model. If you specify a categorical predictor to remove from the model, `removeTerms` removes a group of indicator variables for the predictor in one step.

Output Arguments

`NewMdl` — Generalized linear regression model with fewer terms

`GeneralizedLinearModel` object

Generalized linear regression model with fewer terms, returned as a `GeneralizedLinearModel` object. `NewMdl` is a newly fitted model that uses the input data and settings in `mdl` with the terms specified in `terms` removed from `mdl`.

To overwrite the input argument `mdl`, assign the newly fitted model to `mdl`:

```
mdl = removeTerms(mdl,terms);
```

More About

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.
- − means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
−x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 − x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- `removeTerms` treats a categorical predictor as follows:
  - A model with a categorical predictor that has L levels (categories) includes L – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `removeTerms` treats the group of L – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Alternative Functionality

- Use `stepwiseglm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `addTerms` to add specific terms to a model.
- Use `step` to optimally improve a model by adding or removing terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `addTerms` | `step` | `stepwiseglm`

Topics

“Plots to Understand Predictor Effects and How to Modify a Model” on page 12-21
“Generalized Linear Model Workflow” on page 12-28
“Generalized Linear Models” on page 12-9

Introduced in R2012a

removeTerms

Remove terms from linear regression model

Syntax

```
NewMdl = removeTerms(mdl,terms)
```

Description

`NewMdl = removeTerms(mdl,terms)` returns a linear regression model fitted using the input data and settings in `mdl` with the terms `terms` removed.

Examples

Remove Terms from Linear Regression Model

Create a linear regression model using the `hald` data set. Remove terms that have high *p*-values.

Load the data set.

```
load hald
X = ingredients; % predictor variables
y = heat; % response variable
```

Fit a linear regression model to the data.

```
mdl = fitlm(X,y)
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    ________

    (Intercept)      62.405     70.071      0.8906     0.39913
    x1               1.5511    0.74477      2.0827    0.070822
    x2              0.51017    0.72379     0.70486      0.5009
    x3              0.10191    0.75471     0.13503     0.89592
    x4             -0.14406    0.70905    -0.20317     0.84407


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 2.45
R-squared: 0.982,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 111, p-value = 4.76e-07
```

Remove the `x3` and `x4` terms because their *p*-values are high.

```
terms = 'x3 + x4'; % terms to remove
NewMdl = removeTerms(mdl,terms)
```

```
NewMdl =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     52.577       2.2862    22.998    5.4566e-10
    x1              1.4683       0.1213    12.105    2.6922e-07
    x2             0.66225     0.045855    14.442     5.029e-08


Number of observations: 13, Error degrees of freedom: 10
Root Mean Squared Error: 2.41
R-squared: 0.979,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 230, p-value = 4.41e-09
```

`NewMdl` has the same adjusted R-squared value (0.974) as the previous model, meaning the fit is as good in the new model. All the terms in the new model have extremely low *p*-values.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

`terms` — Terms to remove from regression model

character vector or string scalar formula in Wilkinson notation | *t*-by-*p* terms matrix

Terms to remove from the regression model `mdl`, specified as one of the following:

- Character vector or string scalar formula in "Wilkinson Notation" on page 33-5475 representing one or more terms. The variable names in the formula must be valid MATLAB identifiers.
- Terms matrix `T` of size *t*-by-*p*, where *t* is the number of terms and *p* is the number of predictor variables in `mdl`. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

  For example, suppose `mdl` has three variables `A`, `B`, and `C` in that order. Each row of `T` represents one term:

  - `[0 0 0]` — Constant term or intercept
  - `[0 1 0]` — `B`; equivalently, `A^0 * B^1 * C^0`
  - `[1 0 1]` — `A*C`
  - `[2 0 0]` — `A^2`
  - `[0 1 2]` — `B*(C^2)`

`removeTerms` treats a group of indicator variables for a categorical predictor as a single variable. Therefore, you cannot specify an indicator variable to remove from the model. If you specify a categorical predictor to remove from the model, `removeTerms` removes a group of indicator variables for the predictor in one step. See "Modify Linear Regression Model Using step" on page 33-5990 for an example that describes how to create indicator variables manually and treat each one as a separate variable.

Output Arguments

`NewMdl` — Linear regression model with fewer terms

`LinearModel` object

Linear regression model with fewer terms, returned as a `LinearModel` object. `NewMdl` is a newly fitted model that uses the input data and settings in `mdl` with the terms specified in `terms` removed from `mdl`.

To overwrite the input argument `mdl`, assign the newly fitted model to `mdl`:

```
mdl = removeTerms(mdl,terms);
```

More About

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.
- − means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
−x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 − x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- `removeTerms` treats a categorical predictor as follows:
  - A model with a categorical predictor that has L levels (categories) includes L – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see "Automatic Creation of Dummy Variables" on page 2-49.
  - `removeTerms` treats the group of L – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Alternative Functionality

- Use `stepwiselm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `addTerms` to add specific terms to a model.
- Use `step` to optimally improve a model by adding or removing terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`LinearModel` | `addTerms` | `step` | `stepwiselm`

Topics

"Linear Regression Workflow" on page 11-35
"Interpret Linear Regression Results" on page 11-50
"Linear Regression" on page 11-9

“Stepwise Regression” on page 11-99

Introduced in R2012a

reorderlevels

(Not Recommended) Reorder levels of nominal or ordinal arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the "Categorical Arrays" data type instead.

Syntax

```
B = reorderlevels(A,newlevels)
```

Description

`B = reorderlevels(A,newlevels)` returns a `nominal` or `ordinal` array object of the same type as `A`, but with levels in the new order specified by `newlevels`.

For ordinal arrays, the order of the levels has significance for relational operators, finding minimum and maximum values, and sorting.

Input Arguments

A — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

newlevels — New order of levels

string array | cell array of character vectors | 2-D character matrix

New order of levels, specified as a string array, cell array of character vectors, or 2-D character matrix. `newlevels` must be a reordering of the labels returned by `getlabels`.

Data Types: `char` | `string` | `cell`

Output Arguments

B — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, returned as a `nominal` or `ordinal` array object.

See Also

`addlevels` | `droplevels` | `getlabels` | `nominal` | `ordinal` | `reorderlevels`

Topics

"Reorder Category Levels" on page 2-9
"Sort Ordinal Arrays" on page 2-34

Introduced in R2007a

repartition

Repartition data for cross-validation

Syntax

```
cnew = repartition(c)
cnew = repartition(c,s)
```

Description

`cnew = repartition(c)` creates a `cvpartition` object `cnew` that defines a random partition of the same type as `c`, where `c` is also a `cvpartition` object. That is, `repartition` takes the same observations in `c` and repartitions them into new training and test sets.

`cnew = repartition(c,s)` uses the `RandStream` object `s` as the random number generator for the new partition.

Examples

Repartition Data for Cross-Validation

Repartition observations in a `cvpartition` object. The type of validation partition remains the same.

Partition 100 observations for 3-fold cross-validation.

```
c = cvpartition(100,'KFold',3)
```

```
c =
K-fold cross validation partition
   NumObservations: 100
       NumTestSets: 3
         TrainSize: 67  66  67
          TestSize: 33  34  33
```

Repartition the observations.

```
cnew = repartition(c)
```

```
cnew =
K-fold cross validation partition
   NumObservations: 100
       NumTestSets: 3
         TrainSize: 67  66  67
          TestSize: 33  34  33
```

Notice that the set of observations in the first test set (fold) of `c` is not the same as the set of observations in the first test set of `cnew`.

```
isequal(test(c,1),test(cnew,1))
```

```
ans = logical
   0
```

View the validation partition type of `c` and `cnew`. Both `c` and `cnew` are validation partitions of the same type, `'kfold'`.

```
isequal(c.Type,cnew.Type)
```

```
ans = logical
   1
```

```
c.Type
```

```
ans =
'kfold'
```

Input Arguments

c — Validation partition

`cvpartition` object

Validation partition, specified as a `cvpartition` object. The validation partition type of `c`, `c.Type`, is the same as the validation partition type of the new partition `cnew`.

s — Random number generator

`RandStream` object

Random number generator for the new partition, specified as a `RandStream` object.

Tips

- Repartitioning is useful for Monte Carlo repetitions of cross-validation analyses. `crossval` calls `repartition` when you specify the `'MCReps'` name-value pair argument.

See Also

`RandStream` | `crossval` | `cvpartition`

Introduced in R2008a

RepeatedMeasuresModel class

Repeated measures model class

Description

A `RepeatedMeasuresModel` object represents a model fitted to data with multiple measurements per subject. The object comprises data, fitted coefficients, covariance parameters, design matrix, error degrees of freedom, and between- and within-subjects factor names for a repeated measures model. You can predict model responses using the `predict` method and generate random data at new design points using the `random` method.

Construction

You can fit a repeated measures model using `fitrm(t,modelspec)`.

Input Arguments

`t` — Input data

table

Input data, which includes the values of the response variables and the between-subject factors to use as predictors in the repeated measures model, specified as a table.

Data Types: `table`

`modelspec` — Formula for model specification

character vector or string scalar of the form `'y1-yk ~ terms'`

Formula for model specification, specified as a character vector or string scalar of the form `'y1-yk ~ terms'`. Specify the terms using Wilkinson notation. `fitrm` treats the variables used in model terms as categorical if they are categorical (nominal or ordinal), logical, character arrays, string arrays, or a cell array of character vectors.

Example: `'y1-y4 ~ x1 + x2 * x3'`

Data Types: `char` | `string`

Properties

`BetweenDesign` — Design for between-subject factors

table

Design for between-subject factors and values of repeated measures, stored as a table.

Data Types: `table`

`BetweenModel` — Model for between-subjects factors

character vector

Model for between-subjects factors, stored as a character vector. This character vector is the text representation to the right of the tilde in the model specification you provide when fitting the repeated measures model using `fitrm`.

Data Types: `char`

`BetweenFactorNames` — Names of variables used as between-subject factors
cell array of character vectors

Names of variables used as between-subject factors in the repeated measures model, `rm`, stored as a cell array of character vectors.

Data Types: `cell`

`ResponseNames` — Names of variables used as response variables
cell array of character vectors

Names of variables used as response variables in the repeated measures model, `rm`, stored as a cell array of character vectors.

Data Types: `cell`

`WithinDesign` — Values of within-subject factors
table

Values of the within-subject factors, stored as a table.

Data Types: `table`

`WithinModel` — Model for within-subjects factors
character vector

Model for within-subjects factors, stored as a character vector.

You can specify `WithinModel` as a character vector or a string scalar using dot notation: `Mdl.WithinModel = newWithinModelValue`.

`WithinFactorNames` — Names of within-subject factors
cell array of character vectors

Names of the within-subject factors, stored as a cell array of character vectors.

Data Types: `cell`

`Coefficients` — Values of estimated coefficients
table

Values of the estimated coefficients for fitting the repeated measures as a function of the terms in the between-subjects model, stored as a table.

`fitrm'` defines the coefficients for a categorical term using 'effects' coding, which means coefficients sum to 0. There is one coefficient for each level except the first. The implied coefficient for the first level is the sum of the other coefficients for the term.

You can display the coefficient values as a matrix rather than a table using `coef = r.Coefficients{:,:}`.

You can display marginal means for all levels using the `margmean` method.

Data Types: `table`

`Covariance` — Estimated response covariances
table

Estimated response covariances, that is, covariance of the repeated measures, stored as a table. `fitrm` computes the covariances around the mean returned by the fitted repeated measures model `rm`.

You can display the covariance values as a matrix rather than a table using `coef = r.Covariance{:,:}`.

Data Types: `table`

`DFE` — Error degrees of freedom
scalar value

Error degrees of freedom, stored as a scalar value. `DFE` is the number of observations minus the number of estimated coefficients in the between-subjects model.

Data Types: `double`

Methods

anova	Analysis of variance for between-subject effects
epsilon	Epsilon adjustment for repeated measures anova
grpstats	Compute descriptive statistics of repeated measures data by group
manova	Multivariate analysis of variance
margmean	Estimate marginal means
mauchly	Mauchly's test for sphericity
multcompare	Multiple comparison of estimated marginal means
plot	Plot data with optional grouping
plotprofile	Plot expected marginal means with optional grouping
predict	Compute predicted values given predictor values
random	Generate new random response values given predictor values
ranova	Repeated measures analysis of variance

Examples

Fit a Repeated Measures Model

Load the sample data.

```
load fisheriris
```

The column vector, `species`, consists of iris flowers of three different species: setosa, versicolor, virginica. The double matrix `meas` consists of four types of measurements on the flowers: the length and width of sepals and petals in centimeters, respectively.

Store the data in a table array.

```
t = table(species,meas(:,1),meas(:,2),meas(:,3),meas(:,4),...
'VariableNames',{'species','meas1','meas2','meas3','meas4'});
Meas = table([1 2 3 4]','VariableNames',{'Measurements'});
```

Fit a repeated measures model, where the measurements are the responses and the species is the predictor variable.

```
rm = fitrm(t,'meas1-meas4~species','WithinDesign',Meas)
```

```
rm =
  RepeatedMeasuresModel with properties:

   Between Subjects:
         BetweenDesign: [150x5 table]
         ResponseNames: {'meas1'  'meas2'  'meas3'  'meas4'}
    BetweenFactorNames: {'species'}
         BetweenModel: '1 + species'

   Within Subjects:
         WithinDesign: [4x1 table]
     WithinFactorNames: {'Measurements'}
          WithinModel: 'separatemeans'

   Estimates:
         Coefficients: [3x4 table]
           Covariance: [4x4 table]
```

Display the coefficients.

```
rm.Coefficients
```

```
ans=3×4 table
                          meas1       meas2      meas3      meas4
                        ________    ________    ______    ________

    (Intercept)           5.8433      3.0573     3.758      1.1993
    species_setosa       -0.83733     0.37067   -2.296     -0.95333
    species_versicolor    0.092667   -0.28733    0.502      0.12667
```

`fitrm` uses the `'effects'` contrasts, which means that the coefficients sum to 0. The `rm.DesignMatrix` has one column of 1s for the intercept, and two other columns `species_setosa` and `species_versicolor`, which are as follows:

$$\text{species\_setosa} = \begin{cases} 1, & \textit{if setosa} \\ 0, & \textit{if versicolor} \\ -1, & \textit{if virginica} \end{cases}$$

and

$$\text{species\_versicolor} = \begin{cases} 0, & \textit{if setosa} \\ 1, & \textit{if versicolor} \\ -1, & \textit{if virginica} \end{cases}.$$

Display the covariance matrix.

```
rm.Covariance
```

```
ans=4×4 table
              meas1       meas2       meas3       meas4
            ________    ________    ________    ________

    meas1     0.26501    0.092721     0.16751    0.038401
    meas2    0.092721     0.11539    0.055244     0.03271
    meas3     0.16751    0.055244     0.18519    0.042665
    meas4    0.038401     0.03271    0.042665    0.041882
```

Display the error degrees of freedom.

```
rm.DFE
```

```
ans = 147
```

The error degrees of freedom is the number of observations minus the number of estimated coefficients in the between-subjects model, e.g. 150 – 3 = 147.

More About

Wilkinson Notation

Wilkinson notation describes the factors present in models. It does not describe the multipliers (coefficients) of those factors.

Use these rules to specify the responses in `modelspec`.

Wilkinson Notation	Description
`Y1,Y2,Y3`	Specific list of variables
`Y1-Y5`	All table variables from `Y1` through `Y5`

Use these rules to specify terms in `modelspec`.

Wilkinson Notation	Factors in Standard Notation
`1`	Constant (intercept) term
`X^k`, where `k` is a positive integer	X, X^2, ..., X^k
`X1 + X2`	X1, X2
`X1*X2`	X1, X2, X1*X2
`X1:X2`	X1*X2 only
`-X2`	Do not include X2
`X1*X2 + X3`	X1, X2, X3, X1*X2
`X1 + X2 + X3 + X1:X2`	X1, X2, X3, X1*X2
`X1*X2*X3 - X1:X2:X3`	X1, X2, X3, X1*X2, X1*X3, X2*X3
`X1*(X2 + X3)`	X1, X2, X3, X1*X2, X1*X3

Statistics and Machine Learning Toolbox notation always includes a constant term unless you explicitly remove the term using `-1`.

See Also

`fitrm`

Topics

Class Attributes
Property Attributes

replacedata

Class: `dataset`

(Not Recommended) Replace dataset variables

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = replacedata(A,X)
B = replacedata(A,X,vars)
B = replacedata(A,fun)
B = replacedata(A,fun,vars)
```

Description

`B = replacedata(A,X)` creates a dataset array `B` with the same variables as the dataset array `A`, but with the data for those variables replaced by the data in the array `X`. `replacedata` creates each variable in `B` using one or more columns from `X`, in order. `X` must have as many columns as the total number of columns in all of the variables in `A`, and as many rows as `A` has observations.

`B = replacedata(A,X,vars)` creates a dataset array `B` with the same variables as the dataset array `A`, but with the data for the variables specified in `vars` replaced by the data in the array `X`. The remaining variables in `B` are copies of the corresponding variables in `A`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. Each variable in `B` has as many columns as the corresponding variable in `A`. `X` must have as many columns as the total number of columns in all the variables specified in `vars`.

`B = replacedata(A,fun)` or `B = replacedata(A,fun,vars)` creates a dataset array `B` by applying the function `fun` to the values in `A`'s variables. `replacedata` first horizontally concatenates `A`'s variables into a single array, then applies the function `fun`. The specified variables in `A` must have types and sizes compatible with the concatenation. `fun` is a function handle that accepts a single input array and returns an array with the same number of rows and columns as the input.

Examples

```
data = dataset({rand(3,3),'Var1','Var2','Var3'})

% Use ZSCORE to normalize each variable in a dataset array
% separately, by explicitly extracting and transforming the
% data, and then replacing it.
X = double(data);
X = zscore(X);
data = replacedata(data,X)

% Equivalently, provide a handle to ZSCORE.
data = replacedata(data,@zscore)
```

```
% Use ZSCORE to normalize each observation in a dataset
% array separately by creating an anonymous function.
data = replacedata(data,@(x) zscore(x,[],2))
```

See Also

dataset

replaceWithMissing

Class: `dataset`

(Not Recommended) Insert missing data indicators into a dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds2 = replaceWithMissing(ds,Name,Value)
```

Description

`ds2 = replaceWithMissing(ds,Name,Value)` replaces specified values in a dataset array with standard missing data indicators using options specified by one or more `Name,Value` pair arguments. Use `replaceWithMissing` to specify:

- Which numeric missing value indicators to replace with `NaN`.
- Which character missing value indicators to replace with `''`.
- Which categorical levels to replace with `<undefined>`.

Input Arguments

`ds`

`dataset` array.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NumericValues`

Vector of numeric values that `replaceWithMissing` replaces with `NaN`.

`CategoricalLevels`

Character vector or cell array of character vectors naming the categorical levels that `replaceWithMissing` replaces with `<undefined>`.

`Strings`

Character vector or cell array of character vectors containing the text that `replaceWithMissing` replaces with `''`.

`DataVars`

Specified set of variables in `ds` for which `replaceWithMissing` replaces values. You can specify a positive integer or vector of positive integers indicating the variable column numbers, a variable name or a cell array of variables names, or a logical vector indicating which variables to replace missing values in.

Default: All variables in `ds`.

Output Arguments

`ds2`

`dataset` array that has the specified missing value indicators, in the specified variables of `ds`, replaced with standard missing value indicators.

Examples

Replace Nonstandard Missing Value Indicators

Replace nonstandard missing value indicators with standard missing value indicators.

Replace numeric missing values coded `99` with `NaN`, and character missing values coded `'.'` with `''`.

```
ds = replaceWithMissing(ds,'NumericValues',99,'Strings','.');
```

See Also

`dataset` | `ismissing`

Topics

“Clean Messy and Missing Data” on page 2-97
“Dataset Arrays” on page 2-112

reset

Class: `qrandstream`

Reset state

Syntax

```
reset(q)
```

Description

`reset(q)` resets the state of the quasi-random number stream `q` of the `qrandstream` on page 33-5078 class back to its initial state, `1`. Subsequent points drawn from the stream will be the same as those drawn from a new stream. The command is equivalent to `q.State = 1`.

Examples

Use `qrandstream` to construct a 3-D Halton stream, based on a point set that skips the first 1000 values and then retains every 101st point:

```
q = qrandstream('halton',3,'Skip',1e3,'Leap',1e2)
q =
   Halton quasi-random stream in 3 dimensions
   Point set properties:
               Skip : 1000
               Leap : 100
    ScrambleMethod : none

nextIdx = q.State
nextIdx =
     1
```

Use `qrand` to generate two samples of size four:

```
X1 = qrand(q,4)
X1 =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
nextIdx = q.State
nextIdx =
     5

X2 = qrand(q,4)
X2 =
    0.2446    0.0238    0.8102
    0.5298    0.7540    0.0438
    0.3843    0.5112    0.2758
    0.8335    0.2245    0.4694
nextIdx = q.State
nextIdx =
     9
```

Use `reset` to reset the stream, then generate another sample:

```
reset(q)
nextIdx = q.State
nextIdx =
     1

X = qrand(q,4)
X =
    0.0928    0.3475    0.0051
    0.6958    0.2035    0.2371
    0.3013    0.8496    0.4307
    0.9087    0.5629    0.6166
```

See Also

`qrand` | `qrandstream`

residuals

Class: `GeneralizedLinearMixedModel`

Residuals of fitted generalized linear mixed-effects model

Syntax

```
r = residuals(glme)
r = residuals(glme,Name,Value)
```

Description

`r = residuals(glme)` returns the raw conditional residuals from a fitted generalized linear mixed-effects model `glme`.

`r = residuals(glme,Name,Value)` returns the residuals using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify to return Pearson residuals for the model.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Conditional` — Indicator for conditional residuals

`true` (default) | `false`

Indicator for conditional residuals, specified as the comma-separated pair consisting of `'Conditional'` and one of the following.

Value	Description
`true`	Contributions from both fixed effects and random effects (conditional)
`false`	Contribution from only fixed effects (marginal)

Conditional residuals include contributions from both fixed- and random-effects predictors. Marginal residuals include contribution from only fixed effects. To obtain marginal residual values, `residuals` computes the conditional mean of the response with the empirical Bayes predictor vector of random effects, b, set to 0.

Example: `'Conditional',false`

`ResidualType` — Residual type
`'raw'` (default) | `'Pearson'`

Residual type, specified as the comma-separated pair consisting of `'ResidualType'` and one of the following.

Residual Type	Conditional	Marginal
`'raw'`	$r_{ci} = y_i - g^{-1}\left(x_i^T\widehat{\beta} + z_i^T\widehat{b} + \delta_i\right)$	$r_{mi} = y_i - g^{-1}\left(x_i^T\widehat{\beta} + \delta_i\right)$
`'Pearson'`	$r_{ci}^{pearson} = \frac{r_{ci}}{\sqrt{\frac{\widehat{\sigma^2}}{w_i}v_i\left(\mu_i\left(\widehat{\beta},\widehat{b}\right)\right)}}$	$r_{mi}^{pearson} = \frac{r_{mi}}{\sqrt{\frac{\widehat{\sigma^2}}{w_i}v_i\left(\mu_i\left(\widehat{\beta},0\right)\right)}}$

In each of these equations:

- y_i is the ith element of the n-by-1 response vector, y, where $i = 1, \ldots, n$.
- g^{-1} is the inverse link function for the model.
- x_i^T is the ith row of the fixed-effects design matrix X.
- z_i^T is the ith row of the random-effects design matrix Z.
- δ_i is the ith offset value.
- σ^2 is the dispersion parameter.
- w_i is the ith observation weight.
- v_i is the variance term for the ith observation.
- μ_i is the mean of the response for the ith observation.
- $\widehat{\beta}$ and $\widehat{b}$ are estimated values of β and b.

Raw residuals from a generalized linear mixed-effects model have nonconstant variance. Pearson residuals are expected to have an approximately constant variance, and are generally used for analysis.

Example: `'ResidualType','Pearson'`

Output Arguments

`r` — Residuals
n-by-1 vector

Residuals of the fitted generalized linear mixed-effects model `glme` returned as an n-by-1 vector, where n is the number of observations.

Examples

Plot Residuals Versus Fitted Values

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Generate the conditional Pearson residuals and the conditional fitted values from the model.

```
r = residuals(glme,'ResidualType','Pearson');
mufit = fitted(glme);
```

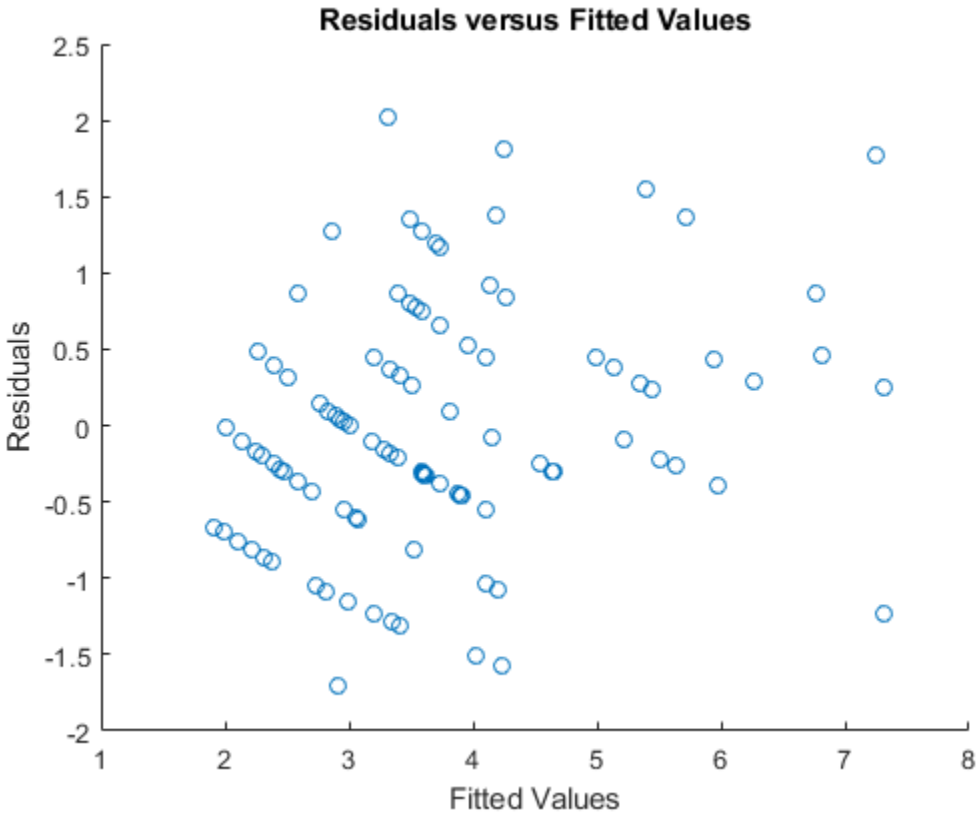
Display the first ten rows of the Pearson residuals.

```
r(1:10)
```

```
ans = 10×1

    0.4530
    0.4339
    0.3833
   -0.2653
    0.2811
   -0.0935
   -0.2984
   -0.2509
    1.5547
   -0.3027
```

Plot the Pearson residuals versus the fitted values, to check for signs of nonconstant variance among the residuals (heteroscedasticity).

```
figure
scatter(mufit,r)
title('Residuals versus Fitted Values')
xlabel('Fitted Values')
ylabel('Residuals')
```



The plot does not show a systematic dependence on the fitted values, so there are no signs of nonconstant variance among the residuals.

See Also

`GeneralizedLinearMixedModel` | `designMatrix` | `fitted` | `response`

residuals

Class: `LinearMixedModel`

Residuals of fitted linear mixed-effects model

Syntax

```
R = residuals(lme)
R = residuals(lme,Name,Value)
```

Description

`R = residuals(lme)` returns the raw conditional residuals from a fitted linear mixed-effects model `lme`.

`R = residuals(lme,Name,Value)` returns the residuals from the linear mixed-effects model `lme` with additional options specified by one or more `Name,Value` pair arguments.

For example, you can specify Pearson or standardized residuals, or residuals with contributions from only fixed effects.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Conditional` — Indicator for conditional residuals

`True` (default) | `False`

Indicator for conditional residuals, specified as the comma-separated pair consisting of `'Conditional'` and one of the following.

`True`	Contribution from both fixed effects and random effects (conditional)
`False`	Contribution from only fixed effects (marginal)

Example: `'Conditional,'False'`

`ResidualType` — Residual type

`'Raw'` (default) | `'Pearson'` | `'Standardized'`

Residual type, specified by the comma-separated pair consisting of `ResidualType` and one of the following.

Residual Type	Conditional	Marginal
`'Raw'`	$r_i^C = \left[y - X\widehat{\beta} - Z\widehat{b}\right]_i$	$r_i^M = \left[y - X\widehat{\beta}\right]_i$
`'Pearson'`	$pr_i^C = \frac{r_i^C}{\sqrt{\left[\widehat{Var}_{y,b}(y - X\beta - Zb)\right]_{ii}}}$	$pr_i^M = \frac{r_i^M}{\sqrt{\left[\widehat{Var}_y(y - X\beta)\right]_{ii}}}$
`'Standardized'`	$st_i^C = \frac{r_i^C}{\sqrt{\left[\widehat{Var}_y(r^C)\right]_{ii}}}$	$st_i^M = \frac{r_i^M}{\sqrt{\left[\widehat{Var}_y(r^M)\right]_{ii}}}$

For more information on the conditional and marginal residuals and residual variances, see `Definitions` at the end of this page.

Example: `'ResidualType','Standardized'`

Output Arguments

R — Residuals

n-by-1 vector

Residuals of the fitted linear mixed-effects model `lme`returned as an *n*-by-1 vector, where *n* is the number of observations.

Examples

Plot Residuals vs. Fitted Values

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over six 2-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

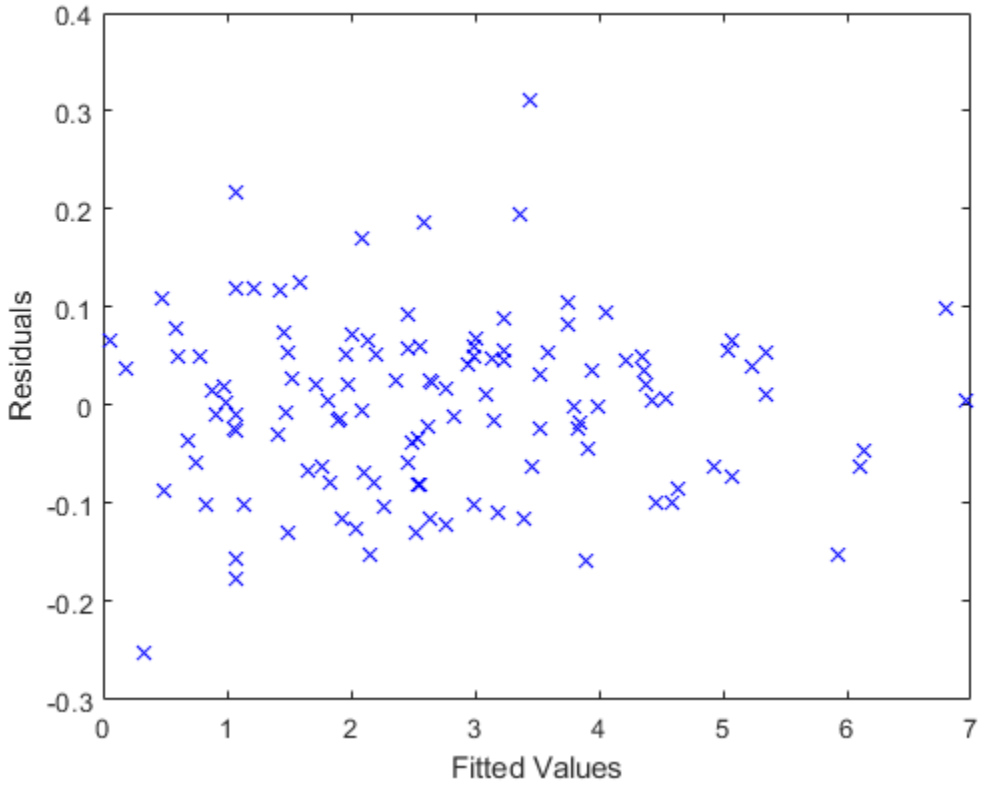
```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

Compute the fitted values and raw residuals.

```
F = fitted(lme);
R = residuals(lme);
```

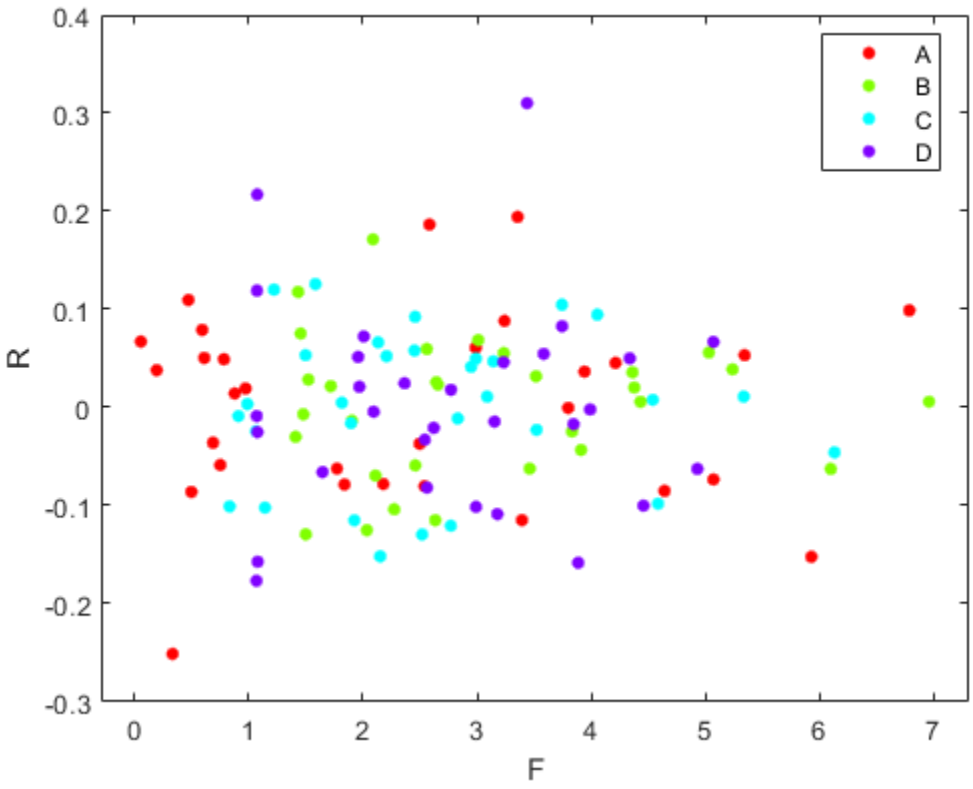
Plot the residuals versus the fitted values.

```
plot(F,R,'bx')
xlabel('Fitted Values')
ylabel('Residuals')
```



Now, plot the residuals versus the fitted values, grouped by program.

```
figure();
gscatter(F,R,Program)
```



The residuals seem to behave similarly across levels of the program as expected.

Compute Conditional and Marginal Pearson Residuals

Load the sample data.

```
load carbig
```

Store the variables for miles per gallon (MPG), acceleration, horsepower, cylinders, and model year in a table.

```
tbl = table(MPG,Acceleration,Horsepower,Cylinders,Model_Year);
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and the cylinders, and potentially correlated random effects for intercept and acceleration grouped by model year.

```
lme = fitlme(tbl,'MPG ~ Acceleration + Horsepower + Cylinders + (Acceleration|Model_Year)');
```

Compute the conditional Pearson residuals and display the first five residuals.

```
PR = residuals(lme,'ResidualType','Pearson');
PR(1:5)
```

```
ans = 5×1
```

```
   -0.0533
    0.0652
    0.3655
   -0.0106
   -0.3340
```

Compute the marginal Pearson residuals and display the first five residuals.

```
PRM = residuals(lme,'ResidualType','Pearson','Conditional',false);
PRM(1:5)
```

```
ans = 5×1

   -0.1250
    0.0130
    0.3242
   -0.0861
   -0.3006
```

Examine Residuals

Load the sample data.

```
load carbig
```

Store the variables for miles per gallon (MPG), acceleration, horsepower, cylinders, and model year in a table.

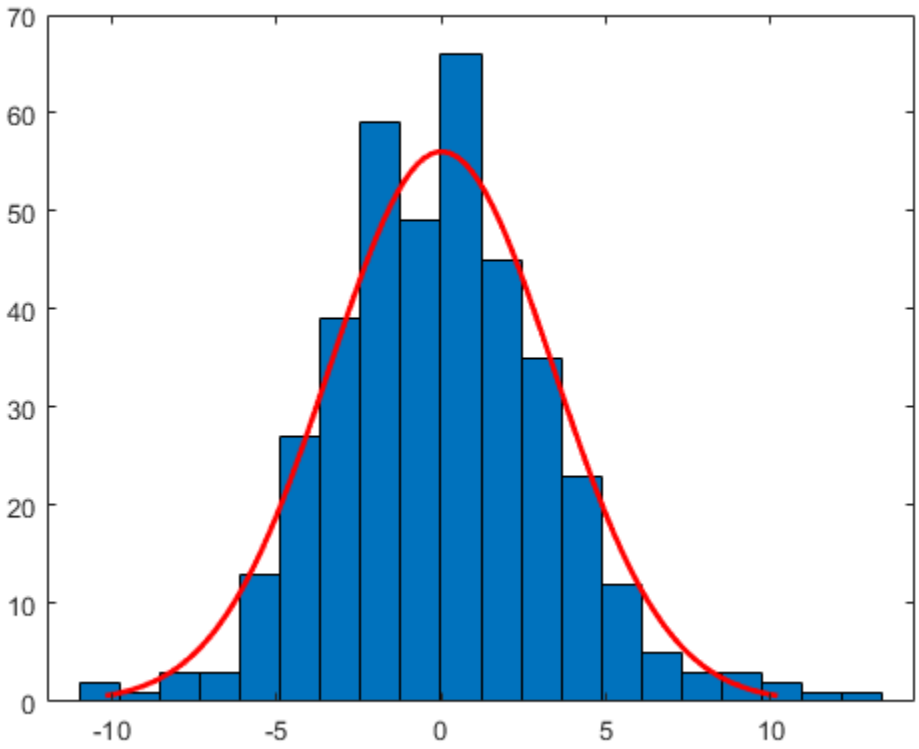
```
tbl = table(MPG,Acceleration,Horsepower,Cylinders,Model_Year);
```

Fit a linear mixed-effects model for miles per gallon (MPG), with fixed effects for acceleration, horsepower, and the cylinders, and potentially correlated random effects for intercept and acceleration grouped by model year.

```
lme = fitlme(tbl,'MPG ~ Acceleration + Horsepower + Cylinders + (Acceleration|Model_Year)');
```

Draw a histogram of the raw residuals with a normal fit.

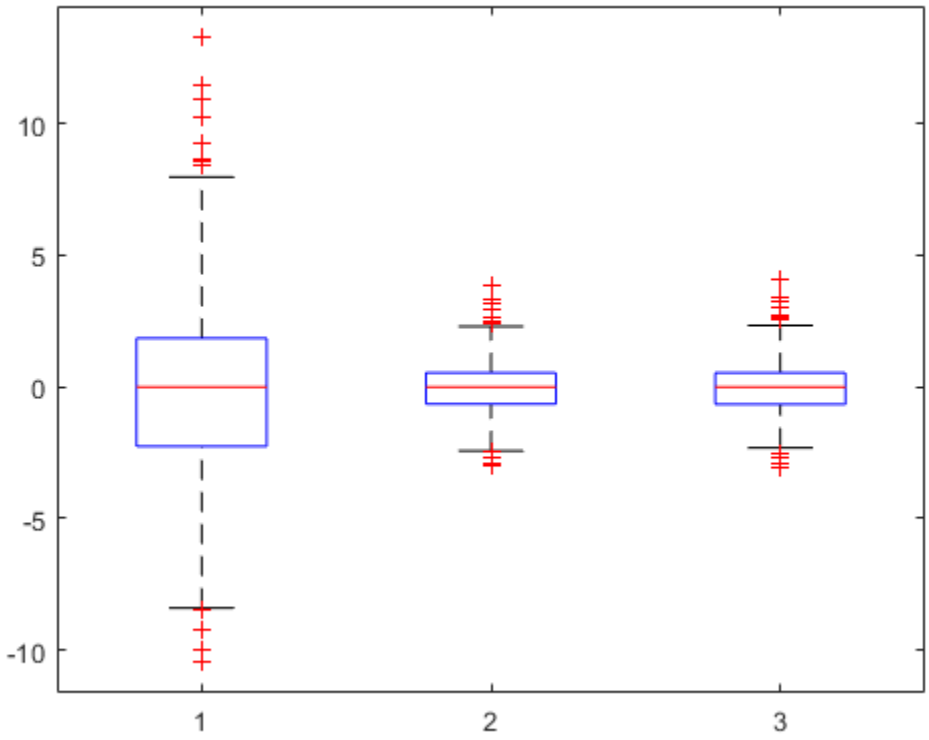
```
r = residuals(lme);
histfit(r)
```



Normal distribution seems to be a good fit for the residuals.

Compute the conditional Pearson and standardized residuals and create box plots of all three types of residuals.

```
pr = residuals(lme,'ResidualType','Pearson');
st = residuals(lme,'ResidualType','Standardized');
X = [r pr st];
boxplot(X)
```



Red plus signs show the observations with residuals above or below $q3 + 1.5(q3 - q1)$ and $q1 - 1.5(q3 - q1)$, where $q1$ and $q3$ are the 25th and 75th percentiles, respectively.

Find the observations with residuals that are 2.5 standard deviations above and below the mean.

```
find(r > mean(r,'omitnan') + 2.5*std(r,'omitnan'))
```

```
ans = 7×1

    62
   252
   255
   330
   337
   341
   396
```

```
find(r < mean(r,'omitnan') - 2.5*std(r,'omitnan'))
```

```
ans = 3×1

   119
   324
   375
```

More About

Conditional and Marginal Residuals

Conditional residuals include contributions from both fixed and random effects, whereas marginal residuals include contribution from only fixed effects.

Suppose the linear mixed-effects model `lme`has an n-by-p fixed-effects design matrix X and an n-by-q random-effects design matrix Z. Also, suppose the p-by-1 estimated fixed-effects vector is $\widehat{\beta}$, and the q-by-1 estimated best linear unbiased predictor (BLUP) vector of random effects is $\widehat{b}$. The fitted conditional response is

$$\widehat{y}_{Cond} = X\widehat{\beta} + Z\widehat{b},$$

and the fitted marginal response is

$$\widehat{y}_{Mar} = X\widehat{\beta},$$

`residuals` can return three types of residuals: raw, Pearson, and standardized. For any type, you can compute the conditional or the marginal residuals. For example, the conditional raw residual is

$$r_{Cond} = y - X\widehat{\beta} - Z\widehat{b},$$

and the marginal raw residual is

$$r_{Mar} = y - X\widehat{\beta}.$$

For more information on other types of residuals, see the `ResidualType` name-value pair argument.

See Also

`LinearMixedModel` | `fitted` | `plotResiduals` | `response`

response

Class: `GeneralizedLinearMixedModel`

Response vector of generalized linear mixed-effects model

Syntax

```
y = response(glme)
[y,binomialsize] = response(glme)
```

Description

`y = response(glme)` returns the response vector `y` used to fit the generalized linear mixed effects model `glme`.

`[y,binomialsize] = response(glme)` also returns the binomial size associated with each element of `y` if the conditional distribution of response given the random effects is binomial.

Input Arguments

`glme` — Generalized linear mixed-effects model

`GeneralizedLinearMixedModel` object

Generalized linear mixed-effects model, specified as a `GeneralizedLinearMixedModel` object. For properties and methods of this object, see `GeneralizedLinearMixedModel`.

Output Arguments

`y` — Response values

n-by-1 vector

Response values, specified as an *n*-by-1 vector, where *n* is the number of observations.

For an observation i with prior weights w_i^p and binomial size n_i (when applicable), the response values y_i can have the following values.

Distribution	Permitted Values	Notes
`Binomial`	$\left\{0, \frac{1}{w_i^p n_i}, \frac{2}{w_i^p n_i}, \ldots, 1\right\}$	w_i^p and n_i are integer values > 0
`Poisson`	$\left\{0, \frac{1}{w_i^p}, \frac{2}{w_i^p}, \ldots\right\}$	w_i^p is an integer value > 0
`Gamma`	$(0,\infty)$	$w_i^p \geq 0$
`InverseGaussian`	$(0,\infty)$	$w_i^p \geq 0$
`normal`	$(-\infty,\infty)$	$w_i^p \geq 0$

You can access the prior weights property w_i^p using dot notation. For example, to access the prior weights property for a model `glme`:

```
glme.ObservationInfo.Weights
```

`binomialsize` — Binomial size
vector

Binomial size associated with each element of `y`, returned as an *n*-by-1 vector, where *n* is the number of observations. `response` only returns `binomialsize` if the conditional distribution of response given the random effects is binomial. `binomialsize` is empty for other distributions.

Examples

Plot Response Versus Fitted Values

Load the sample data.

```
load mfr
```

This simulated data is from a manufacturing company that operates 50 factories across the world, with each factory running a batch process to create a finished product. The company wants to decrease the number of defects in each batch, so it developed a new manufacturing process. To test the effectiveness of the new process, the company selected 20 of its factories at random to participate in an experiment: Ten factories implemented the new process, while the other ten continued to run the old process. In each of the 20 factories, the company ran five batches (for a total of 100 batches) and recorded the following data:

- Flag to indicate whether the batch used the new process (`newprocess`)
- Processing time for each batch, in hours (`time`)
- Temperature of the batch, in degrees Celsius (`temp`)
- Categorical variable indicating the supplier (`A`, `B`, or `C`) of the chemical used in the batch (`supplier`)
- Number of defects in the batch (`defects`)

The data also includes `time_dev` and `temp_dev`, which represent the absolute deviation of time and temperature, respectively, from the process standard of 3 hours at 20 degrees Celsius.

Fit a generalized linear mixed-effects model using `newprocess`, `time_dev`, `temp_dev`, and `supplier` as fixed-effects predictors. Include a random-effects term for intercept grouped by `factory`, to account for quality differences that might exist due to factory-specific variations. The response variable `defects` has a Poisson distribution, and the appropriate link function for this model is log. Use the Laplace fit method to estimate the coefficients. Specify the dummy variable encoding as `'effects'`, so the dummy variable coefficients sum to 0.

The number of defects can be modeled using a Poisson distribution

$$\text{defects}_{ij} \sim \text{Poisson}(\mu_{ij})$$

This corresponds to the generalized linear mixed-effects model

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{newprocess}_{ij} + \beta_2 \text{time\_dev}_{ij} + \beta_3 \text{temp\_dev}_{ij} + \beta_4 \text{supplier\_C}_{ij} + \beta_5 \text{supplier\_B}_{ij} + b_i,$$

where

- defects_{ij} is the number of defects observed in the batch produced by factory i during batch j.
- μ_{ij} is the mean number of defects corresponding to factory i (where $i = 1, 2, \ldots, 20$) during batch j (where $j = 1, 2, \ldots, 5$).
- newprocess_{ij}, time_dev_{ij}, and temp_dev_{ij} are the measurements for each variable that correspond to factory i during batch j. For example, newprocess_{ij} indicates whether the batch produced by factory i during batch j used the new process.
- supplier_C_{ij} and supplier_B_{ij} are dummy variables that use effects (sum-to-zero) coding to indicate whether company `C` or `B`, respectively, supplied the process chemicals for the batch produced by factory i during batch j.
- $b_i \sim N(0, \sigma_b^2)$ is a random-effects intercept for each factory i that accounts for factory-specific variation in quality.

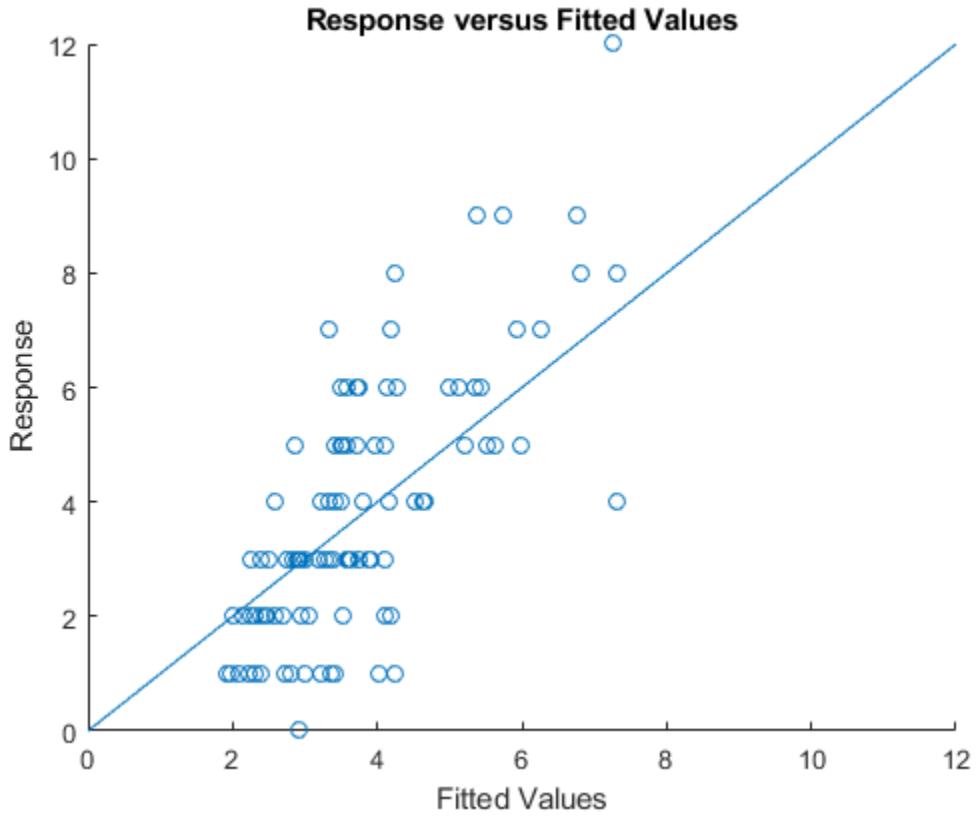
```
glme = fitglme(mfr,'defects ~ 1 + newprocess + time_dev + temp_dev + supplier + (1|factory)',...
    'Distribution','Poisson','Link','log','FitMethod','Laplace','DummyVarCoding','effects');
```

Extract the observed response values for the model, then use `fitted` to generate the fitted conditional mean values.

```
y = response(glme);   % Observed response values
yfit = fitted(glme);  % Fitted response values
```

Create a scatterplot of the observed response values versus fitted values. Add a reference line to improve the visualization.

```
figure
scatter(yfit,y)
xlim([0,12])
ylim([0,12])
refline(1,0)
title('Response versus Fitted Values')
xlabel('Fitted Values')
ylabel('Response')
```



The plot shows a positive correlation between the fitted values and the observed response values.

References

[1] Hox, J. *Multilevel Analysis, Techniques and Applications*. Lawrence Erlbaum Associates, Inc., 2002.

See Also

`GeneralizedLinearMixedModel` | `fitted` | `residuals`

response

Class: `LinearMixedModel`

Response vector of the linear mixed-effects model

Syntax

```
y = response(lme)
```

Description

`y = response(lme)` returns the response vector `y` used to fit the linear mixed-effects model `lme`.

Input Arguments

`lme` — Linear mixed-effects model

`LinearMixedModel` object

Linear mixed-effects model, specified as a `LinearMixedModel` object constructed using `fitlme` or `fitlmematrix`.

Output Arguments

`y` — Response values

n-by-1 vector

Response values, specified as an *n*-by-1 vector, where *n* is the number of observations.

Data Types: `single` | `double`

Examples

Plot Response versus Fitted Values

Load the sample data.

```
load('weight.mat');
```

`weight` contains data from a longitudinal study, where 20 subjects are randomly assigned to 4 exercise programs, and their weight loss is recorded over two-week time periods. This is simulated data.

Store the data in a table. Define `Subject` and `Program` as categorical variables.

```
tbl = table(InitialWeight,Program,Subject,Week,y);
tbl.Subject = nominal(tbl.Subject);
tbl.Program = nominal(tbl.Program);
```

Fit a linear mixed-effects model where the initial weight, type of program, week, and the interaction between the week and type of program are the fixed effects. The intercept and week vary by subject.

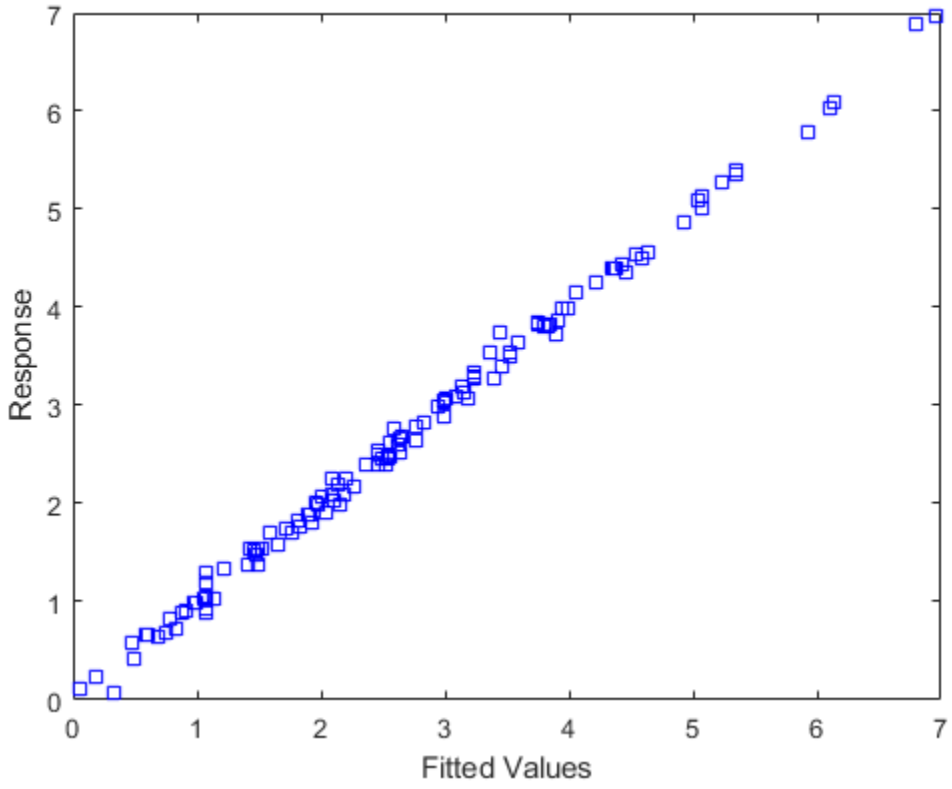
```
lme = fitlme(tbl,'y ~ InitialWeight + Program*Week + (Week|Subject)');
```

Compute the fitted values and the response.

```
F = fitted(lme);
y = response(lme);
```

Plot the response versus the fitted values.

```
plot(F,y,'bs')
xlabel('Fitted Values')
ylabel('Response')
```



See Also

`LinearMixedModel` | `fitted` | `residuals`

resubEdge

Resubstitution classification edge

Syntax

```
e = resubEdge(Mdl)
e = resubEdge(Mdl,'IncludeInteractions',includeInteractions)
```

Description

`e = resubEdge(Mdl)` returns the weighted resubstitution "Classification Edge" on page 33-1294 (`e`) for the trained classification model `Mdl` using the predictor data stored in `Mdl.X`, the corresponding true class labels stored in `Mdl.Y`, and the observation weights stored in `Mdl.W`.

`e = resubEdge(Mdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

Examples

Estimate Resubstitution Edge of SVM Classifiers

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a support vector machine (SVM) classifier. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X,Y,'Standardize',true,'ClassNames',{'b','g'});
```

`SVMModel` is a trained `ClassificationSVM` classifier.

Estimate the resubstitution edge, which is the mean of the training sample margins.

```
e = resubEdge(SVMModel)
```

```
e = 5.1000
```

Select Naive Bayes Classifier Features by Comparing In-Sample Edges

The classifier edge measures the average of the classifier margins. One way to perform feature selection is to compare training sample edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the `ionosphere` data set. Remove the first two predictors for stability.

```
load ionosphere
X = X(:,3:end);
```

Define these two data sets:

- `fullX` contains all predictors.
- `partX` contains the 10 most important predictors.

```
fullX = X;
idx = fscmrmr(X,Y);
partX = X(:,idx(1:10));
```

Train a naive Bayes classifier for each predictor set.

```
FullMdl = fitcnb(fullX,Y);
PartMdl = fitcnb(partX,Y);
```

`FullMdl` and `PartMdl` are trained `ClassificationNaiveBayes` classifiers.

Estimate the training sample edge for each classifier.

```
fullEdge = resubEdge(FullMdl)
```

```
fullEdge = 0.6554
```

```
partEdge = resubEdge(PartMdl)
```

```
partEdge = 0.7796
```

The edge of the classifier trained on the 10 most important predictors is larger. This result suggests that the classifier trained using only those predictors has a better in-sample fit.

Compare GAMs by Examining Training Sample Margins and Edge

Compare a generalized additive model (GAM) with linear terms to a GAM with both linear and interaction terms by examining the training sample margins and edge. Based solely on this comparison, the classifier with the highest margins and edge is the best model.

Load the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

```
load census1994
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. To reduce the running time for this example, subsample 500 training observations from `adultdata` by using the `datasample` function.

```
rng('default') % For reproducibility
NumSamples = 5e2;
adultdata = datasample(adultdata,NumSamples,'Replace',false);
```

Train a GAM that contains both linear and interaction terms for predictors. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
Mdl = fitcgam(adultdata,'salary','Interactions','all','MaxPValue',0.05)
```

```
Mdl =
  ClassificationGAM
             PredictorNames: {1x14 cell}
               ResponseName: 'salary'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
             ScoreTransform: 'logit'
                  Intercept: -32.0842
               Interactions: [82x2 double]
           NumObservations: 500


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. `Mdl` includes 82 interaction terms.

Estimate the training sample margins and edge for `Mdl`.

```
M = resubMargin(Mdl);
E = resubEdge(Mdl)
```

```
E = 1.0000
```

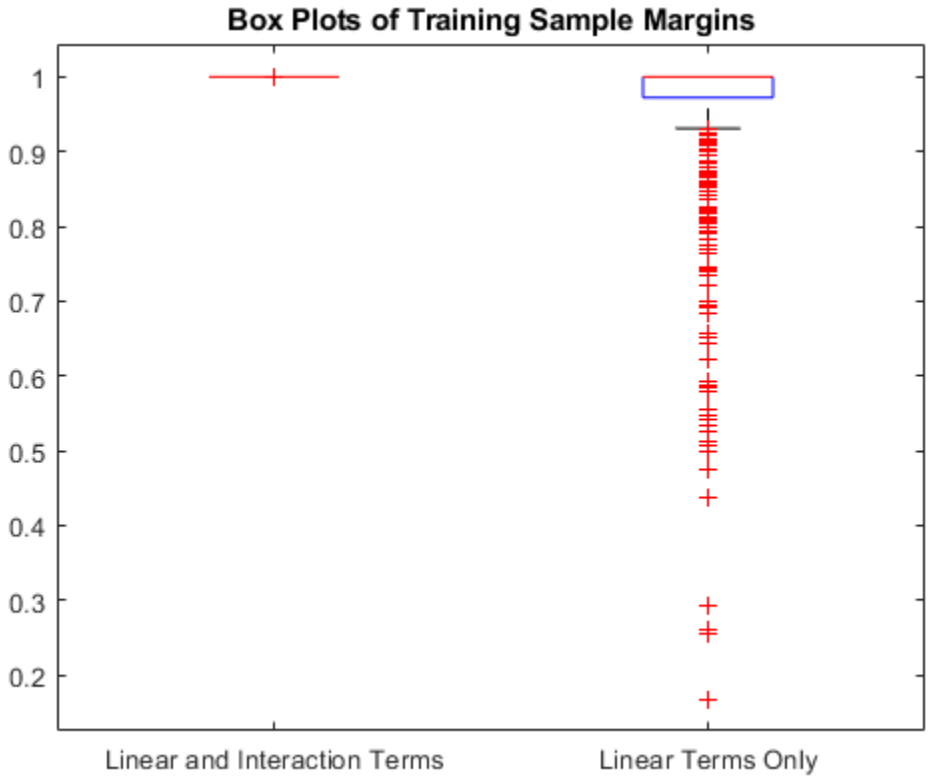
Estimate the training sample margins and edge for `Mdl` without including interaction terms.

```
M_nointeractions = resubMargin(Mdl,'IncludeInteractions',false);
E_nointeractions = resubEdge(Mdl,'IncludeInteractions',false)
```

```
E_nointeractions = 0.9516
```

Display the distributions of the margins using box plots.

```
boxplot([M M_nointeractions],'Labels',{'Linear and Interaction Terms','Linear Terms Only'})
title('Box Plots of Training Sample Margins')
```



When you include the interaction terms in the computation, all the resubstitution margin values for `Mdl` are 1, and the resubstitution edge value (average of the margins) is 1. The margins and edge decrease when you do not include the interaction terms in `Mdl`.

Input Arguments

`Mdl` — Classification machine learning model

full classification model object

Classification machine learning model, specified as a full classification model object, as given in the following table of supported models.

Model	Classification Model Object
Generalized additive model	`ClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

`includeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `ClassificationGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal classification score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Algorithms

`resubEdge` computes the classification edge according to the corresponding `edge` function of the object (`Mdl`). For a model-specific description, see the `edge` function reference pages in the following table.

Model	Classification Model Object (`Mdl`)	edge Object Function
Generalized additive model	`ClassificationGAM`	`edge`
k-nearest neighbor model	`ClassificationKNN`	`edge`
Naive Bayes model	`ClassificationNaiveBayes`	`edge`
Neural network model	`ClassificationNeuralNetwork`	`edge`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`edge`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports `ClassificationKNN` and `ClassificationSVM` objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`resubLoss` | `resubMargin` | `resubPredict`

Introduced in R2012a

resubEdge

Class: `ClassificationDiscriminant`

Classification edge by resubstitution

Syntax

```
edge = resubEdge(obj)
```

Description

`edge = resubEdge(obj)` returns the classification edge obtained by `obj` on its training data.

Input Arguments

obj

Discriminant analysis classifier, produced using `fitcdiscr`.

Output Arguments

edge

Classification edge obtained by resubstituting the training data into the calculation of edge.

Examples

Estimate the Resubstitution Edge of Discriminant Analysis Classifiers

Estimate the quality of a discriminant analysis classifier for Fisher's iris data by resubstitution.

Load Fisher's iris data set.

```
load fisheriris
```

Train a discriminant analysis classifier.

```
Mdl = fitcdiscr(meas,species);
```

Compute the resubstitution edge.

```
redge = resubEdge(Mdl)
```

```
redge = 0.9454
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are class prior probabilities. If you supply additional weights, those weights are normalized to sum to the prior probabilities in the respective classes, and are then used to compute the weighted average.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes.

The classification margin is a column vector with the same number of rows as in the matrix `X`. A high value of margin indicates a more reliable prediction than a low value.

Score

For discriminant analysis, the score of a classification is the posterior probability of the classification. For the definition of posterior probability in discriminant analysis, see “Posterior Probability” on page 20-6.

See Also

`ClassificationDiscriminant` | `edge` | `fitcdiscr` | `resubMargin`

Topics

“Discriminant Analysis Classification” on page 20-2

resubEdge

Resubstitution classification edge for multiclass error-correcting output codes (ECOC) model

Syntax

```
e = resubEdge(Mdl)
e = resubEdge(Mdl,Name,Value)
```

Description

`e = resubEdge(Mdl)` returns the resubstitution classification edge on page 33-5525 (`e`) for the multiclass error-correcting output codes (ECOC) model `Mdl` using the training data stored in `Mdl.X` and the corresponding class labels stored in `Mdl.Y`.

The classification edge is a scalar value that represents the weighted mean of the classification margins on page 33-5525.

`e = resubEdge(Mdl,Name,Value)` computes the resubstitution classification edge with additional options specified by one or more name-value pair arguments. For example, you can specify a decoding scheme, binary learner loss function, and verbosity level.

Examples

Resubstitution Edge of ECOC Model

Compute the resubstitution edge for an ECOC model with SVM binary learners.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
classOrder = unique(Y)
```

```
classOrder = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

Compute the resubstitution edge, which is the mean of the training-sample margins.

```
e = resubEdge(Mdl)
```

```
e = 0.4960
```

Select ECOC Model Features by Comparing Training-Sample Edges

Perform feature selection by comparing training-sample edges from multiple models. Based solely on this comparison, the classifier with the greatest edge is the best classifier.

Load Fisher's iris data set. Define two data sets:

- `fullX` contains all four predictors.
- `partX` contains the sepal measurements only.

```
load fisheriris
X = meas;
fullX = X;
partX = X(:,1:2);
Y = species;
```

Train an ECOC model using SVM binary learners for each predictor set. Standardize the predictors using an SVM template, specify the class order, and compute the posterior probabilities.

```
t = templateSVM('Standardize',true);
classOrder = unique(Y)
```

```
classOrder = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
FullMdl = fitcecoc(fullX,Y,'Learners',t,'ClassNames',classOrder,...
    'FitPosterior',true);
PartMdl = fitcecoc(partX,Y,'Learners',t,'ClassNames',classOrder,...
    'FitPosterior',true);
```

The default SVM score is the distance from the decision boundary. If you specify to compute posterior probabilities, then the software uses posterior probabilities as scores.

Compute the resubstitution edge for each classifier. The quadratic loss function operates on scores in the domain [0,1]. Specify to use quadratic loss when aggregating the binary learners for both models.

```
fullEdge = resubEdge(FullMdl,'BinaryLoss','quadratic')
```

```
fullEdge = 0.9896
```

```
partEdge = resubEdge(PartMdl,'BinaryLoss','quadratic')
```

```
partEdge = 0.5059
```

The edge for the classifier trained on the complete data set is greater, suggesting that the classifier trained with all the predictors has a better training-sample fit.

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubEdge(Mdl,'BinaryLoss','quadratic')` specifies a quadratic binary learner loss function.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.
  - `s` is the 1-by-L row vector of classification scores.

- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-5525.

Example: `'Decoding','lossbased'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Tips

- To compare the margins or edges of several ECOC classifiers, use template objects to specify a common score transform function among the classifiers during training.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `edge` | `fitcecoc` | `predict` | `resubLoss` | `resubMargin` | `resubPredict`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

resubEdge

Classification edge by resubstitution

Syntax

```
edge = resubEdge(ens)
edge = resubEdge(ens,Name,Value)
```

Description

`edge = resubEdge(ens)` returns the classification edge obtained by `ens` on its training data.

`edge = resubEdge(ens,Name,Value)` calculates edge with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A classification ensemble created with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `resubEdge` uses only these learners for calculating edge.

Default: `1:NumTrained`

`mode`

Character vector or string scalar representing the meaning of the output `edge`:

- `'ensemble'` — `edge` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `edge` is a vector with one element per trained learner.
- `'cumulative'` — `edge` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

edge

Classification edge obtained by `ens` by resubstituting the training data into the calculation of edge. Classification edge is classification margin averaged over the entire data. `edge` can be a scalar or vector, depending on the setting of the `mode` name-value pair.

Examples

Find Classification Edge by Resubstitution of Training Data

Find the resubstitution edge for an ensemble that classifies the Fisher iris data.

Load the sample data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Find the resubstitution edge.

```
edge = resubEdge(ens)
```

```
edge = 3.2486
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `ens.Prior`.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `ens.X`.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

See Also

resubEdge | resubLoss | resubMargin | resubPredict

resubEdge

Class: `ClassificationTree`

Classification edge by resubstitution

Syntax

```
edge = resubEdge(tree)
```

Description

`edge = resubEdge(tree)` returns the classification edge obtained by `tree` on its training data.

Input Arguments

`tree`

A classification tree created using `fitctree`.

Output Arguments

`edge`

Classification edge obtained by resubstituting the training data into the calculation of edge.

Examples

Estimate the quality of a classification tree for the Fisher iris data by resubstitution.

```
load fisheriris
tree = fitctree(meas,species);
redge = resubEdge(tree)
```

```
redge =
    0.9384
```

More About

Edge

The edge is the weighted mean value of the classification margin. The weights are the class probabilities in `tree.Prior`.

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix X.

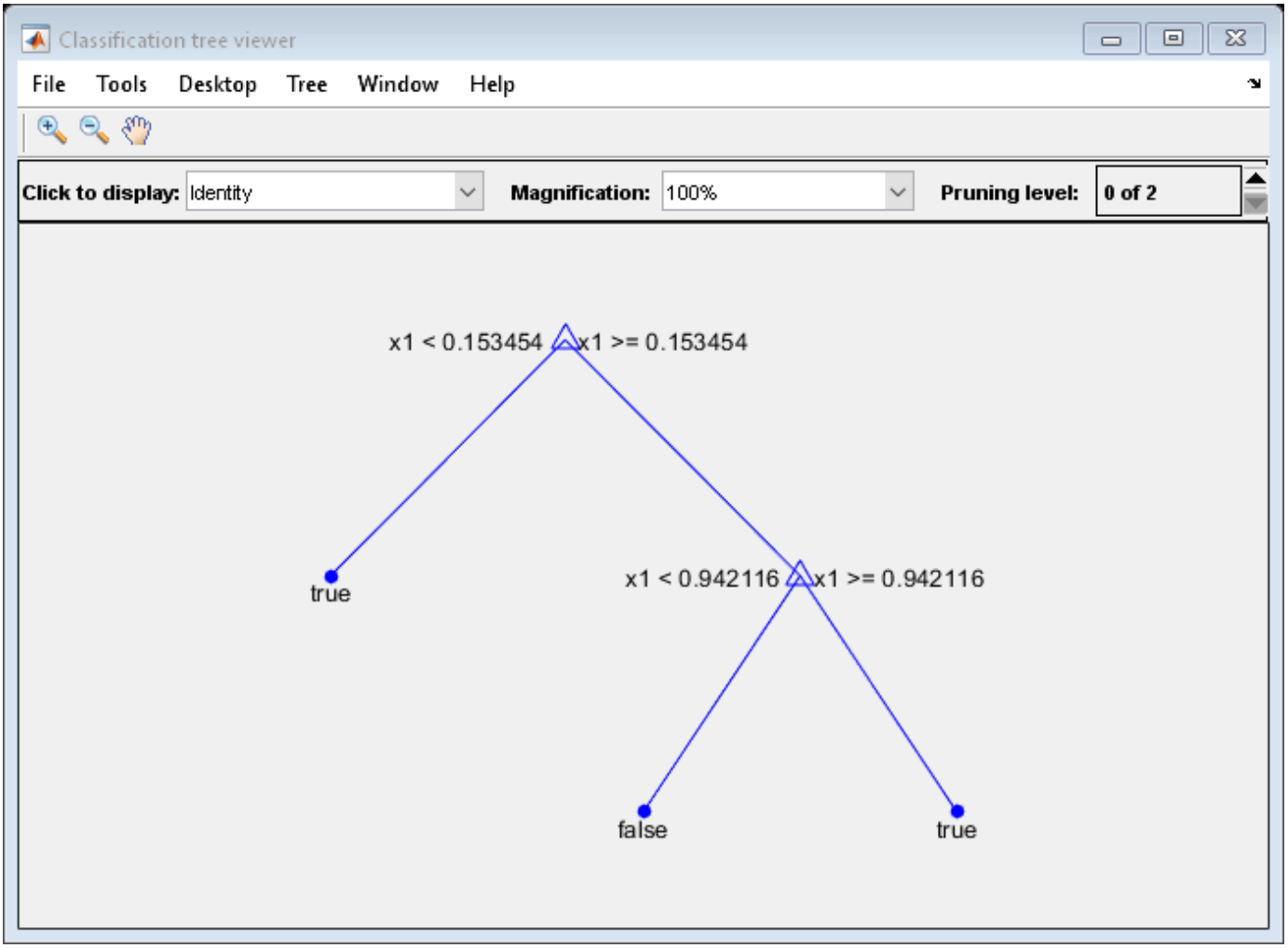
Score (tree)

For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor `X` as `true` when `X < 0.15` or `X > 0.95`, and `X` is false otherwise.

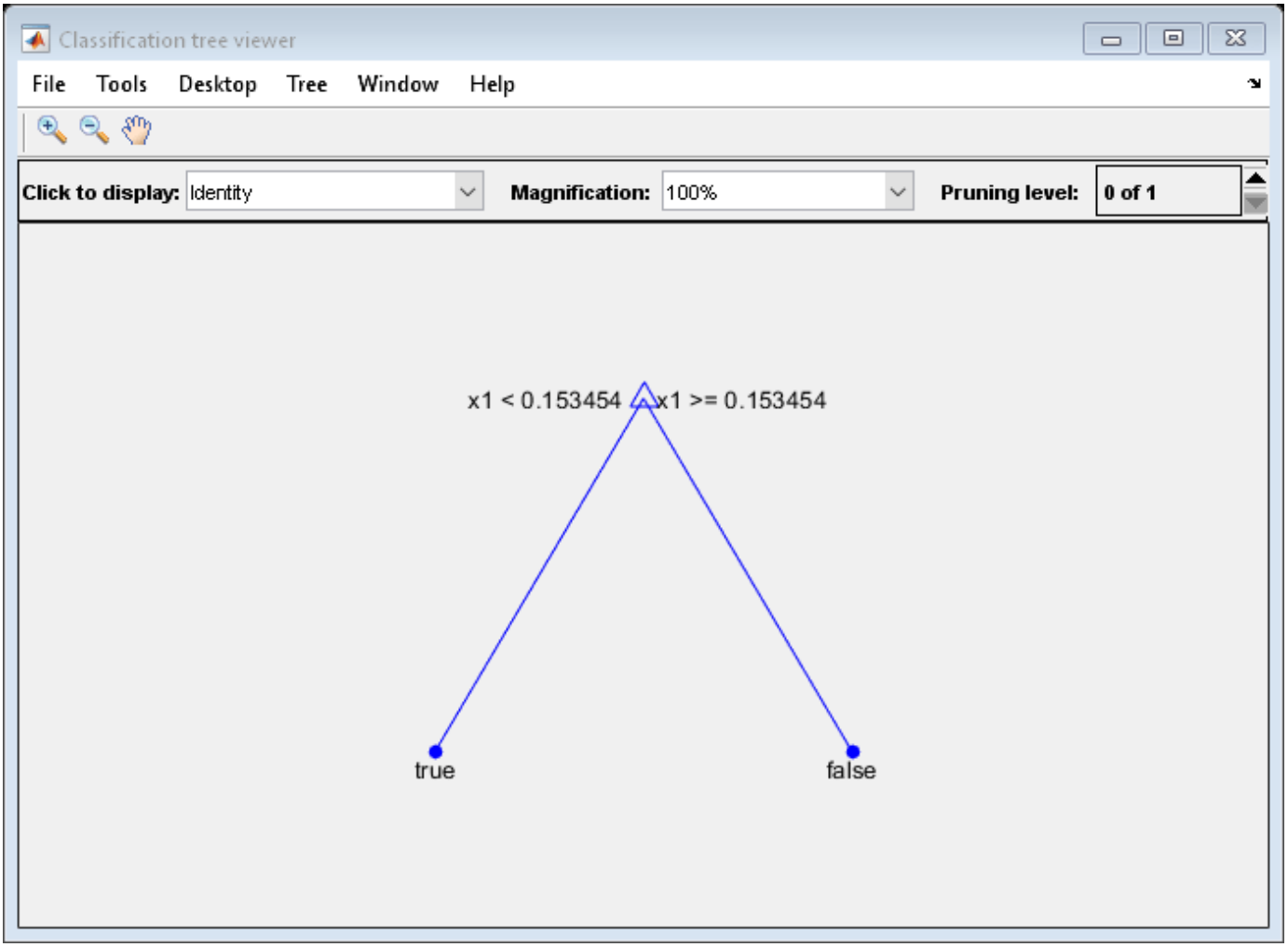
Generate 100 random points and classify them:

```
rng(0,'twister') % for reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

See Also

`edge` | `fitctree` | `resubLoss` | `resubMargin` | `resubPredict`

resubLoss

Class: `ClassificationDiscriminant`

Classification error by resubstitution

Syntax

```
L = resubLoss(obj)
L = resubLoss(obj,Name,Value)
```

Description

`L = resubLoss(obj)` returns the resubstitution loss, meaning the loss computed for the data that `fitcdiscr` used to create `obj`.

`L = resubLoss(obj,Name,Value)` returns loss statistics with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`obj`

Discriminant analysis classifier, produced using `fitcdiscr`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function

`'mincost'` (default) | `'binodeviance'` | `'classiferror'` | `'exponential'` | `'hinge'` | `'logit'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using the corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss

Value	Description
'mincost'	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
'quadratic'	Quadratic loss

'mincost' is appropriate for classification scores that are posterior probabilities. Discriminant analysis models return posterior probabilities as classification scores by default (see predict).

- Specify your own function using function handle notation.

  Suppose that n be the number of observations in X and K be the number of distinct classes (numel(obj.ClassNames)). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument lossvalue is a scalar.
  - You choose the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in obj.ClassNames.

    Construct C by setting C(p,q) = 1 if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in obj.ClassNames. S is a matrix of classification scores, similar to the output of predict.
  - W is an n-by-1 numeric vector of observation weights. If you pass W, the software normalizes them to sum to 1.
  - Cost is a *K*-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) - eye(K) specifies a cost of 0 for correct classification, and 1 for misclassification.

  Specify your function using 'LossFun',@*lossfun*.

For more details on loss functions, see “Classification Loss” on page 33-5537.

Data Types: char | string | function_handle

Output Arguments

L

Classification error, a scalar. The meaning of the error depends on the values in weights and lossfun. See “Classification Loss” on page 33-5537.

Examples

Compute the resubstituted classification error for the Fisher iris data:

```
load fisheriris
obj = fitcdiscr(meas,species);
```

```
L = resubLoss(obj)

L =
    0.0200
```

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

  $$\sum_{j=1}^{n} w_j = 1.$$

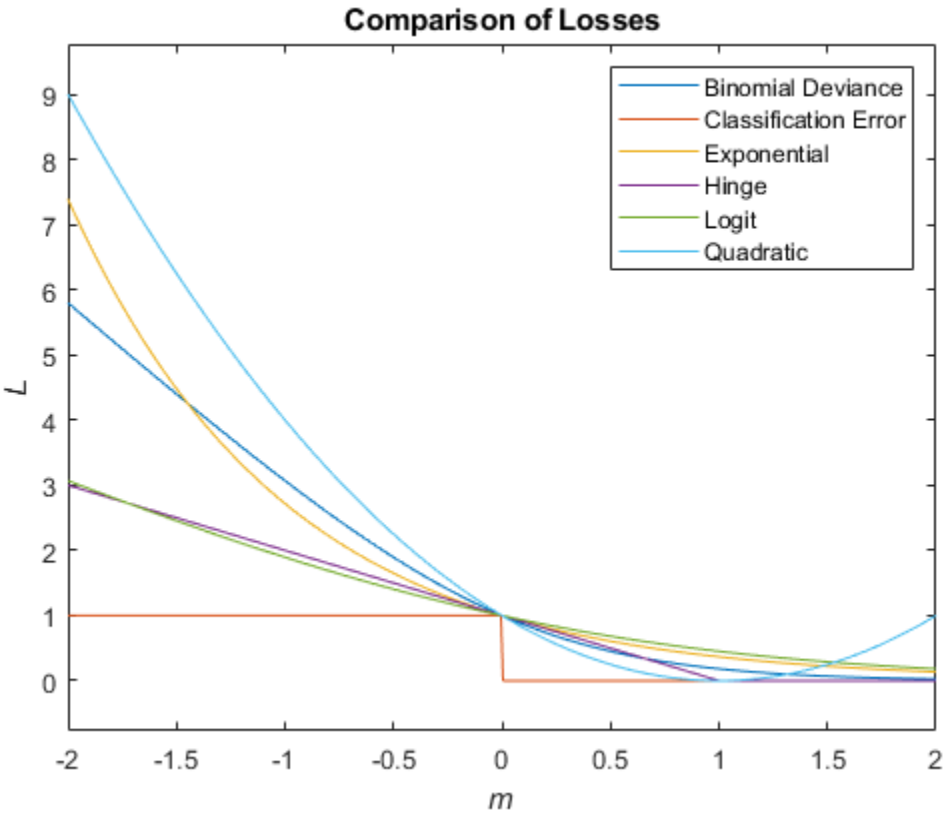
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$

Loss Function	Value of LossFun	Equation
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Posterior Probability

The posterior probability that a point x belongs to class k is the product of the prior probability and the multivariate normal density. The density function of the multivariate normal with 1-by-d mean μ_k and d-by-d covariance Σ_k at a 1-by-d point x is

$$P(x|k) = \frac{1}{\left((2\pi)^d|\Sigma_k|\right)^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^T\right),$$

where $|\Sigma_k|$ is the determinant of Σ_k, and Σ_k^{-1} is the inverse matrix.

Let $P(k)$ represent the prior probability of class k. Then the posterior probability that an observation x is of class k is

$$\widehat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)},$$

where $P(x)$ is a normalization constant, the sum over k of $P(x|k)P(k)$.

Prior Probability

The prior probability is one of three choices:

- `'uniform'` — The prior probability of class `k` is one over the total number of classes.
- `'empirical'` — The prior probability of class `k` is the number of training samples of class `k` divided by the total number of training samples.

- Custom — The prior probability of class `k` is the `k`th element of the `prior` vector. See `fitcdiscr`.

After creating a classification model (`Mdl`) you can set the prior using dot notation:

```
Mdl.Prior = v;
```

where `v` is a vector of positive elements representing the frequency with which each element occurs. You do not need to retrain the classifier when you set a new prior.

Cost

The matrix of expected costs per observation is defined in “Cost” on page 20-7.

See Also

`ClassificationDiscriminant` | `fitcdiscr` | `loss`

Topics

“Discriminant Analysis Classification” on page 20-2

resubLoss

Resubstitution classification loss for multiclass error-correcting output codes (ECOC) model

Syntax

```
L = resubLoss(Mdl)
L = resubLoss(Mdl,Name,Value)
```

Description

`L = resubLoss(Mdl)` returns the classification loss by resubstitution (`L`) for the multiclass error-correcting output codes (ECOC) model `Mdl` using the training data stored in `Mdl.X` and the corresponding class labels stored in `Mdl.Y`. By default, `resubLoss` uses the classification error on page 33-5546 to compute `L`.

The classification loss (`L`) is a generalization or resubstitution quality measure. Its interpretation depends on the loss function and weighting scheme, but in general, better classifiers yield smaller classification loss values.

`L = resubLoss(Mdl,Name,Value)` returns the classification loss with additional options specified by one or more name-value pair arguments. For example, you can specify the loss function, decoding scheme, and verbosity level.

Examples

Resubstitution Loss of ECOC Model

Compute the resubstitution loss for an ECOC model with SVM binary learners.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
classOrder = unique(Y)
```

```
classOrder = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

Estimate the resubstitution classification error, which is the default classification loss.

```
L = resubLoss(Mdl)
```

```
L = 0.0267
```

The ECOC model misclassifies 2.67% of the training-sample irises.

Determine ECOC Model Quality Using Custom Resubstitution Loss

Determine the quality of an ECOC model by using a custom loss function that considers the minimal binary loss for each observation.

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y)  % Class order
```

```
classOrder = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
rng(1); % For reproducibility
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

Create a function that takes the minimal loss for each observation, then averages the minimal losses for all observations. `S` corresponds to the `NegLoss` output of `resubPredict`.

```
lossfun = @(~,S,~,~)mean(min(-S,[],2));
```

Compute the custom classification loss for the training data.

```
resubLoss(Mdl,'LossFun',lossfun)
```

```
ans = 0.0065
```

The average minimal binary loss for the training data is `0.0065`.

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubLoss(Mdl,'BinaryLoss','hamming','LossFun',@lossfun)` specifies `'hamming'` as the binary learner loss function and the custom function handle `@lossfun` as the overall loss function.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

  where:

  - `M` is the K-by-L coding matrix stored in `Mdl.CodingMatrix`.

- `s` is the 1-by-*L* row vector of classification scores.
- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-5547.

Example: `'Decoding','lossbased'`

`LossFun` — Loss function
`'classiferror'` (default) | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'classiferror'` or a function handle.

- Specify the built-in function `'classiferror'`. In this case, the loss function is the classification error on page 33-5546, which is the proportion of misclassified observations.
- Or, specify your own function using function handle notation.

  Assume that `n = size(X,1)` is the sample size and `K` is the number of classes. Your function must have the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

- The output argument `lossvalue` is a scalar.
- You specify the function name (*`lossfun`*).
- `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. The column order corresponds to the class order in `Mdl.ClassNames`.

  Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
- `S` is an `n`-by-`K` numeric matrix of negated loss values for the classes. Each row corresponds to an observation. The column order corresponds to the class order in `Mdl.ClassNames`. The input `S` resembles the output argument `NegLoss` of `resubPredict`.
- `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes its elements to sum to `1`.
- `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of 0 for correct classification and 1 for misclassification.

Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

More About

Classification Error

The classification error is a binary classification error measure that has the form

$$L = \frac{\sum_{j=1}^{n} w_j e_j}{\sum_{j=1}^{n} w_j},$$

where:

- w_j is the weight for observation j. The software renormalizes the weights to sum to 1.
- $e_j = 1$ if the predicted class of observation j differs from its true class, and 0 otherwise.

In other words, the classification error is the proportion of observations misclassified by the classifier.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	$(-\infty,\infty)$	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	$(-\infty,\infty)$	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	$[0,1]$ or $(-\infty,\infty)$	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	$(-\infty,\infty)$	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	$(-\infty,\infty)$	$(1 - y_js_j)/2$

Value	Description	Score Domain	$g(y_j,s_j)$
`'logit'`	Logistic	$(-\infty,\infty)$	$\log[1 + \exp(-y_j s_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `fitcecoc` | `loss` | `predict` | `resubPredict`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

resubLoss

Classification error by resubstitution

Syntax

```
L = resubLoss(ens)
L = resubLoss(ens,Name,Value)
```

Description

`L = resubLoss(ens)` returns the resubstitution loss, meaning the loss computed for the data that `fitcensemble` used to create `ens`.

`L = resubLoss(ens,Name,Value)` calculates loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A classification ensemble created with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `resubLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

`lossfun`

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss

Value	Description
'hinge'	Hinge loss
'logit'	Logistic loss
'mincost'	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
'quadratic'	Quadratic loss

'mincost' is appropriate for classification scores that are posterior probabilities.

- Bagged and subspace ensembles return posterior probabilities by default (ens.Method is 'Bag' or 'Subspace').
- If the ensemble method is 'AdaBoostM1', 'AdaBoostM2', GentleBoost, or 'LogitBoost', then, to use posterior probabilities as classification scores, you must specify the double-logit score transform by entering

  ```
  ens.ScoreTransform = 'doublelogit';
  ```
- For all other ensemble methods, the software does not support posterior probabilities as classification scores.

- Specify your own function using function handle notation.

  Suppose that n be the number of observations in X and K be the number of distinct classes (numel(ens.ClassNames), ens is the input model). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument lossvalue is a scalar.
  - You choose the function name (*lossfun*).
  - C is an n-by-K logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in ens.ClassNames.

    Construct C by setting C(p,q) = 1 if observation p is in class q, for each row. Set all other elements of row p to 0.
  - S is an n-by-K numeric matrix of classification scores. The column order corresponds to the class order in ens.ClassNames. S is a matrix of classification scores, similar to the output of predict.
  - W is an n-by-1 numeric vector of observation weights. If you pass W, the software normalizes them to sum to 1.
  - Cost is a *K*-by-K numeric matrix of misclassification costs. For example, Cost = ones(K) - eye(K) specifies a cost of 0 for correct classification, and 1 for misclassification.

  Specify your function using 'LossFun',@*lossfun*.

For more details on loss functions, see “Classification Loss” on page 33-5551.

Default: 'classiferror'

`mode`

Character vector or string scalar representing the meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

`L`

Classification loss on page 33-5551, by default the fraction of misclassified data. `L` can be a vector, and can mean different things, depending on the name-value pair settings.

Examples

Estimate Classification Error for Training Observations

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification ensemble of 100 decision trees using AdaBoostM2. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Estimate the resubstitution classification error.

```
loss = resubLoss(ens)
```

```
loss = 0.0333
```

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:

- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

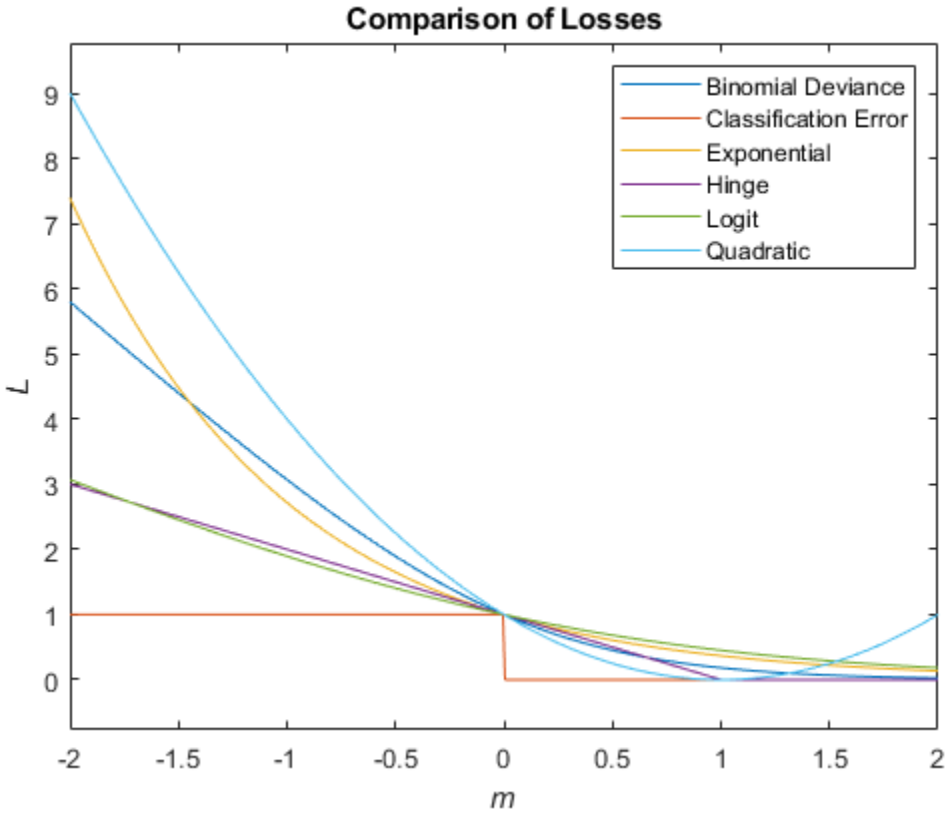
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\hat{y}_j \neq y_j\right\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.

Loss Function	Value of LossFun	Equation
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = (f(X_j)'C)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\hat{y}_j = \underset{k=1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j (1 - m_j)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



See Also

`resubEdge` | `resubLoss` | `resubMargin` | `resubPredict`

resubLoss

Resubstitution classification loss

Syntax

```
L = resubLoss(Mdl)
L = resubLoss(Mdl,Name,Value)
```

Description

`L = resubLoss(Mdl)` returns the "Classification Loss" on page 33-5559 by resubstitution (L), or the in-sample classification loss, for the trained classification model `Mdl` using the training data stored in `Mdl.X` and the corresponding class labels stored in `Mdl.Y`.

The interpretation of `L` depends on the loss function (`'LossFun'`) and weighting scheme (`Mdl.W`). In general, better classifiers yield smaller classification loss values. The default `'LossFun'` value varies depending on the model object `Mdl`.

`L = resubLoss(Mdl,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'LossFun','binodeviance'` sets the loss function to the binomial deviance function.

Examples

Determine Resubstitution Loss of Naive Bayes Classifier

Determine the in-sample classification error (resubstitution loss) of a naive Bayes classifier. In general, a smaller loss indicates a better classifier.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
```

```
    DistributionParameters: {3x4 cell}

  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Estimate the in-sample classification error.

```
L = resubLoss(Mdl)
```

```
L = 0.0400
```

The naive Bayes classifier misclassifies 4% of the training observations.

Determine Resubstitution Hinge Loss of SVM Classifier

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a support vector machine (SVM) classifier. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X,Y,'ClassNames',{'b','g'},'Standardize',true);
```

`SVMModel` is a trained `ClassificationSVM` classifier.

Estimate the in-sample hinge loss.

```
L = resubLoss(SVMModel,'LossFun','hinge')
```

```
L = 0.1603
```

The hinge loss is `0.1603`. Classifiers with hinge losses close to 0 are preferred.

Compare GAMs by Examining Classification Loss

Train a generalized additive model (GAM) that contains both linear and interaction terms for predictors, and estimate the classification loss with and without interaction terms. Specify whether to include interaction terms when estimating the classification loss for training and test data.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 50 observations for the new test data set.

```
rng('default') % For reproducibility
n = size(X,1);
```

```
newInds = randsample(n,50);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X(inds,:),Y(inds),'ClassNames',{'b','g'},'Interactions',10)
```

```
Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 2.0026
             Interactions: [10x2 double]
          NumObservations: 301


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object.

Compute the resubstitution classification loss both with and without interaction terms in `Mdl`. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
resubl = resubLoss(Mdl)
```

```
resubl = 0
```

```
resubl_nointeraction = resubLoss(Mdl,'IncludeInteractions',false)
```

```
resubl_nointeraction = 0
```

Estimate the classification loss both with and without interaction terms in `Mdl`.

```
l = loss(Mdl,XNew,YNew)
```

```
l = 0.0615
```

```
l_nointeraction = loss(Mdl,XNew,YNew,'IncludeInteractions',false)
```

```
l_nointeraction = 0.0615
```

Including interaction terms does not change the classification loss for `Mdl`. The trained model classifies all training samples correctly and misclassifies approximately 6% of the test samples.

Input Arguments

`Mdl` — Classification machine learning model

full classification model object

Classification machine learning model, specified as a full classification model object, as given in the following table of supported models.

Model	Classification Model Object
Generalized additive model	`ClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubLoss(Mdl,'LossFun','logit')` estimates the logit resubstitution loss.

`IncludeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `ClassificationGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

`LossFun` — Loss function
`'binodeviance'` | `'classiferror'` | `'crossentropy'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'crossentropy'`	Cross-entropy loss (for neural networks only)
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  The default value depends on the trained model (`Mdl`).

- The default value is `'classiferror'` if `Mdl` is a `ClassificationGAM`, `ClassificationNeuralNetwork`, or `ClassificationSVM` object.
- The default value is `'mincost'` if `Mdl` is a `ClassificationKNN` or `ClassificationNaiveBayes` object.

For more details on loss functions, see “Classification Loss” on page 33-5559.

- To specify a custom loss function, use function handle notation. The function must have this form:

  `lossvalue = lossfun(C,S,W,Cost)`

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (`lossfun`).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. `n` is the number of observations in `Tbl` or `X`, and `K` is the number of distinct classes (`numel(Mdl.ClassNames)`). The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(p,q) = 1`, if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights.
  - `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

Example: `'LossFun','binodeviance'`

Data Types: `char` | `string` | `function_handle`

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):

- y_j^* is a vector of K – 1 zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
- $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
- $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.

- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

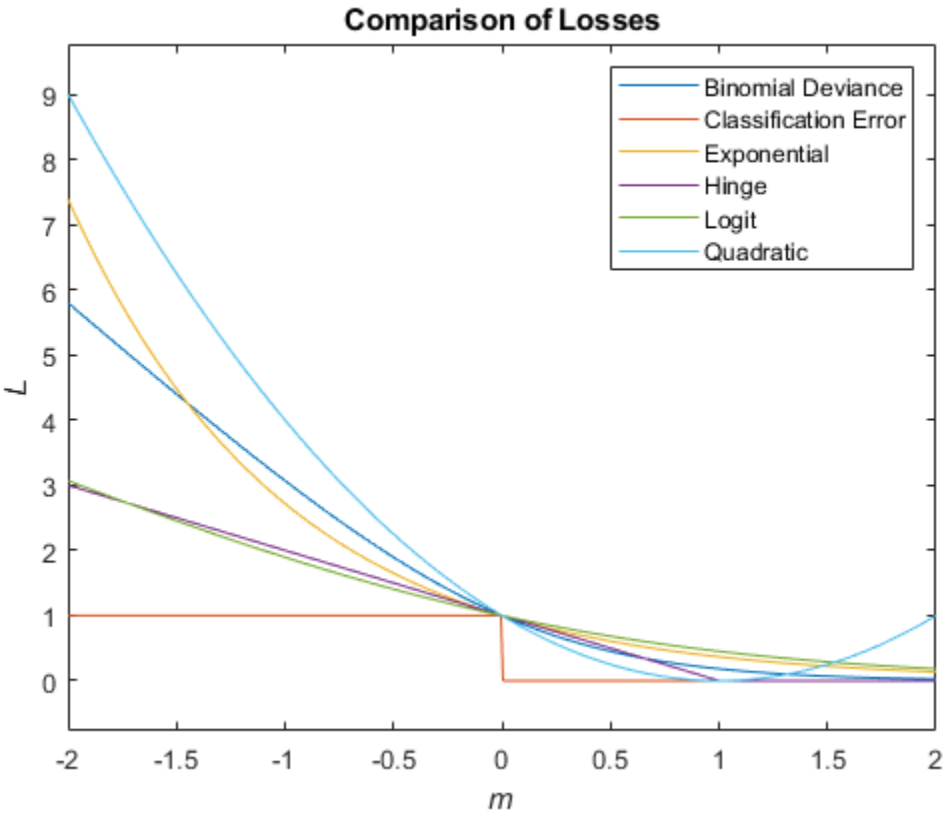
$$\sum_{j=1}^{n} w_j = 1.$$

Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\left\{\widehat{y}_j \neq y_j\right\}.$ $\widehat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn},$ where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp\left(-m_j\right).$
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\left\{0, 1 - m_j\right\}.$
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log\left(1 + \exp\left(-m_j\right)\right).$

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Algorithms

`resubLoss` computes the classification loss according to the corresponding `loss` function of the object (`Mdl`). For a model-specific description, see the `loss` function reference pages in the following table.

Model	Classification Model Object (`Mdl`)	`loss` Object Function
Generalized additive model	`ClassificationGAM`	`loss`
k-nearest neighbor model	`ClassificationKNN`	`loss`
Naive Bayes model	`ClassificationNaiveBayes`	`loss`
Neural network model	`ClassificationNeuralNetwork`	`loss`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`loss`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports `ClassificationKNN` and `ClassificationSVM` objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`resubEdge` | `resubMargin` | `resubPredict`

Introduced in R2012a

resubLoss

Class: `ClassificationTree`

Classification error by resubstitution

Syntax

```
L = resubLoss(tree)
L = resubLoss(tree,Name,Value)
L = resubLoss(tree,'Subtrees',subtreevector)
[L,se] = resubLoss(tree,'Subtrees',subtreevector)
[L,se,NLeaf] = resubLoss(tree,'Subtrees',subtreevector)
[L,se,NLeaf,bestlevel] = resubLoss(tree,'Subtrees',subtreevector)
[L,...] = resubLoss(tree,'Subtrees',subtreevector,Name,Value)
```

Description

`L = resubLoss(tree)` returns the resubstitution loss, meaning the loss computed for the data that `fitctree` used to create `tree`.

`L = resubLoss(tree,Name,Value)` returns the loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

`L = resubLoss(tree,'Subtrees',subtreevector)` returns a vector of classification errors for the trees in the pruning sequence `subtreevector`.

`[L,se] = resubLoss(tree,'Subtrees',subtreevector)` returns the vector of standard errors of the classification errors.

`[L,se,NLeaf] = resubLoss(tree,'Subtrees',subtreevector)` returns the vector of numbers of leaf nodes in the trees of the pruning sequence.

`[L,se,NLeaf,bestlevel] = resubLoss(tree,'Subtrees',subtreevector)` returns the best pruning level as defined in the `TreeSize` name-value pair. By default, `bestlevel` is the pruning level that gives loss within one standard deviation of minimal loss.

`[L,...] = resubLoss(tree,'Subtrees',subtreevector,Name,Value)` returns loss statistics with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree`

A classification tree constructed by `fitctree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function
`'mincost'` (default) | `'binodeviance'` | `'classiferror'` | `'exponential'` | `'hinge'` | `'logit'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. Classification trees return posterior probabilities as classification scores by default (see `predict`).

- Specify your own function using function handle notation.

  Suppose that `n` be the number of observations in `X` and `K` be the number of distinct classes (`numel(tree.ClassNames)`). Your function must have this signature

  ```
  lossvalue = lossfun(C,S,W,Cost)
  ```

  where:

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs. The column order corresponds to the class order in `tree.ClassNames`.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `tree.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes them to sum to `1`.

- `Cost` is a *K*-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification, and `1` for misclassification.

Specify your function using `'LossFun',@lossfun`.

For more details on loss functions, see "Classification Loss" on page 33-5570.

Data Types: `char` | `string` | `function_handle`

`Name,Value` arguments associated with pruning subtrees:

`Subtrees` — Pruning level
0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `resubLoss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`resubLoss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as the comma-separated pair consisting of `'TreeSize'` and one of the following values:

- `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L+se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Output Arguments

`L`

Classification loss on page 33-5570, a vector the length of `Subtrees`. The meaning of the error depends on the values in `Weights` and `LossFun`.

`se`

Standard error of loss, a vector the length of `Subtrees`.

`NLeaf`

Number of leaves (terminal nodes) in the pruned subtrees, a vector the length of `Subtrees`.

`bestlevel`

A scalar whose value depends on `TreeSize`:

- `TreeSize` = `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L+se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `TreeSize` = `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Examples

Compute the In-Sample Classification Error

Compute the resubstitution classification error for the `ionosphere` data.

```
load ionosphere
tree = fitctree(X,Y);
L = resubLoss(tree)
```

```
L = 0.0114
```

Examine the Classification Error for Each Subtree

Unpruned decision trees tend to overfit. One way to balance model complexity and out-of-sample performance is to prune a tree (or restrict its growth) so that in-sample and out-of-sample performance are satisfactory.

Load Fisher's iris data set. Partition the data into training (50%) and validation (50%) sets.

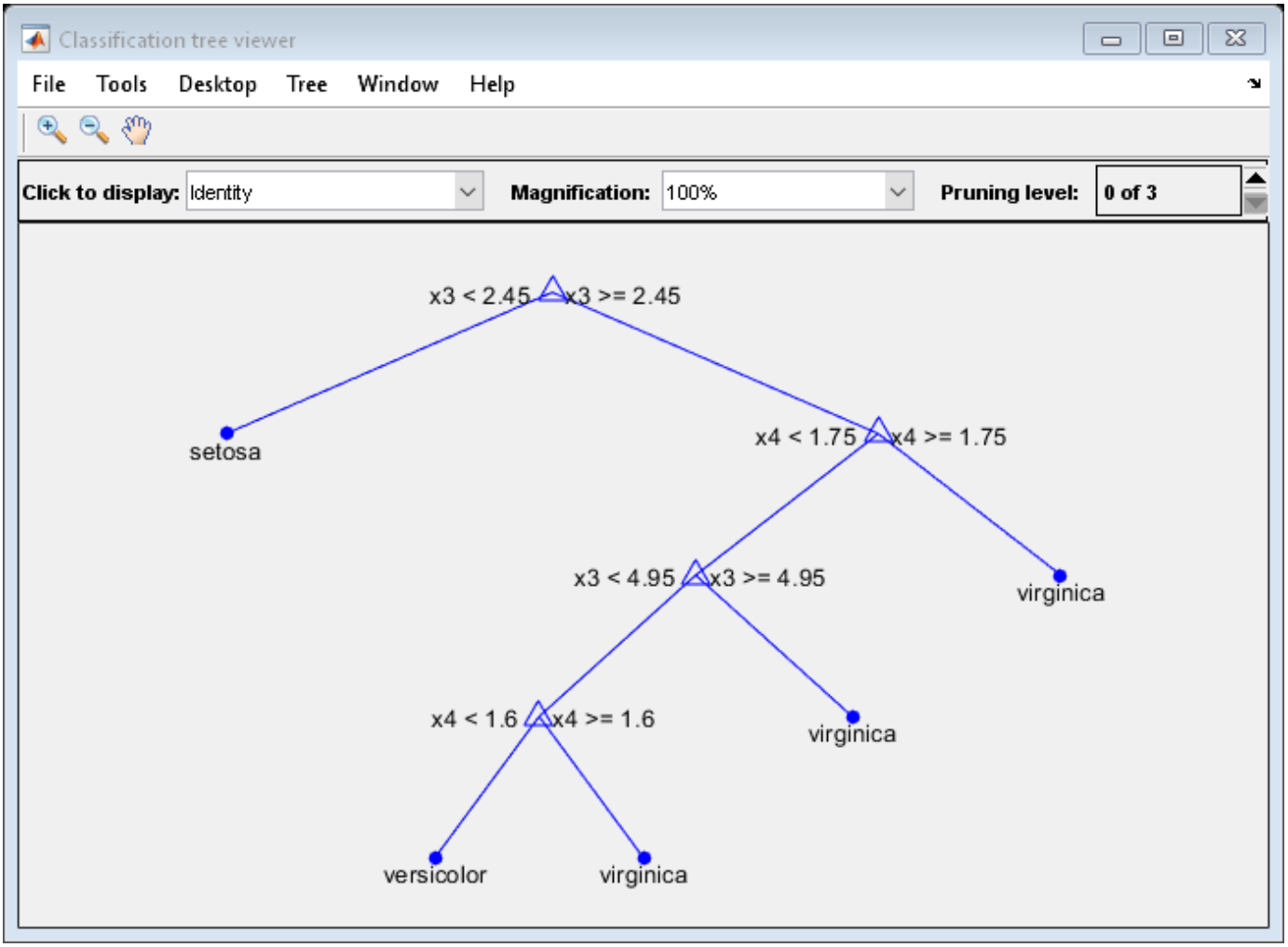
```
load fisheriris
n = size(meas,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a classification tree using the training set.

```
Mdl = fitctree(meas(idxTrn,:),species(idxTrn));
```

View the classification tree.

```
view(Mdl,'Mode','graph');
```



The classification tree has four pruning levels. Level 0 is the full, unpruned tree (as displayed). Level 3 is just the root node (i.e., no splits).

Examine the training sample classification error for each subtree (or pruning level) excluding the highest level.

```
m = max(Mdl.PruneList) - 1;
trnLoss = resubLoss(Mdl,'SubTrees',0:m)
```

```
trnLoss = 3×1

    0.0267
    0.0533
    0.3067
```

- The full, unpruned tree misclassifies about 2.7% of the training observations.
- The tree pruned to level 1 misclassifies about 5.3% of the training observations.
- The tree pruned to level 2 (i.e., a stump) misclassifies about 30.6% of the training observations.

Examine the validation sample classification error at each level excluding the highest level.

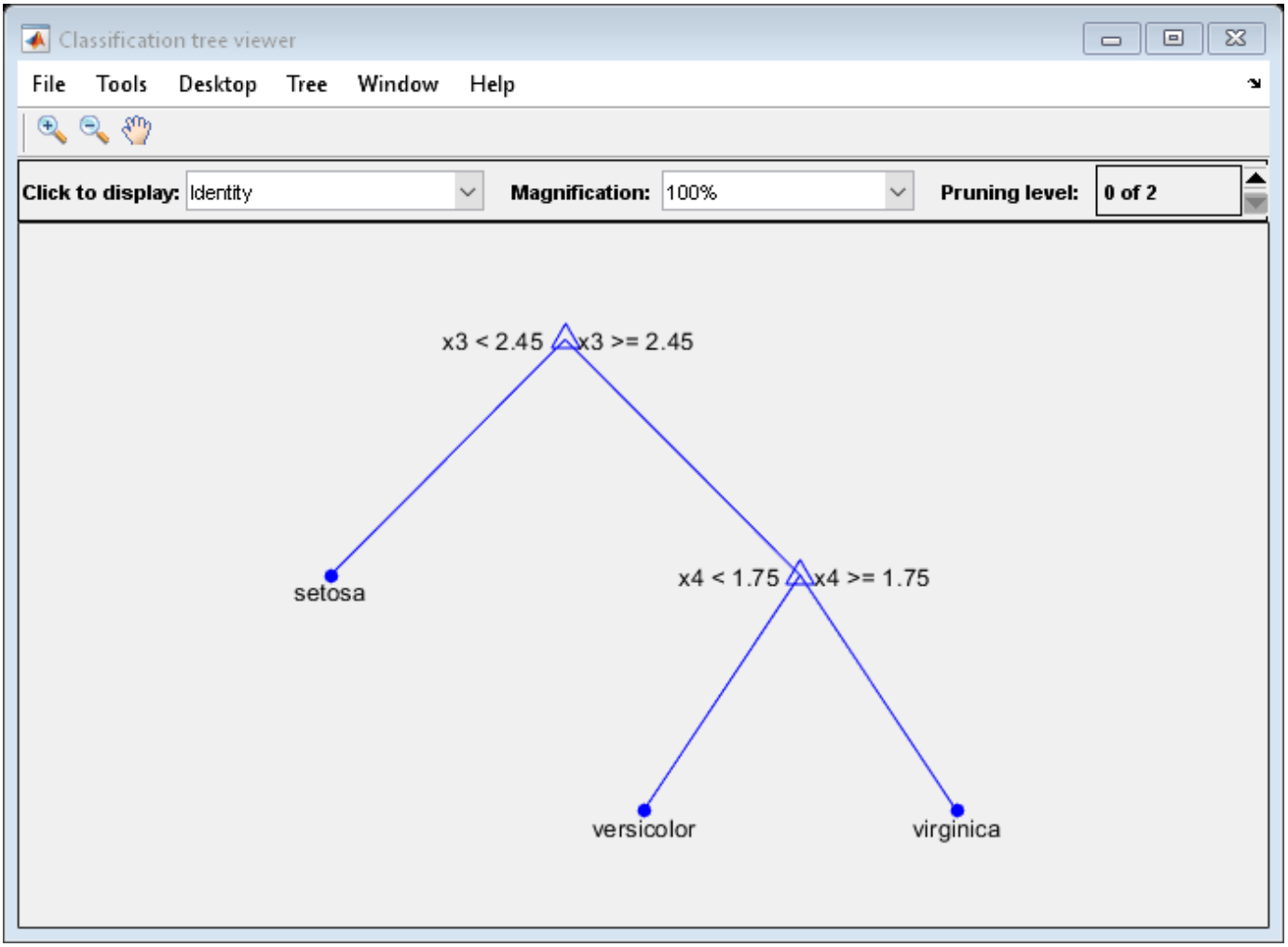
```
valLoss = loss(Mdl,meas(idxVal,:),species(idxVal),'SubTrees',0:m)
```

```
valLoss = 3×1

    0.0369
    0.0237
    0.3067
```

- The full, unpruned tree misclassifies about 3.7% of the validation observations.
- The tree pruned to level 1 misclassifies about 2.4% of the validation observations.
- The tree pruned to level 2 (i.e., a stump) misclassifies about 30.7% of the validation observations.

To balance model complexity and out-of-sample performance, consider pruning `Mdl` to level 1.

```
pruneMdl = prune(Mdl,'Level',1);
view(pruneMdl,'Mode','graph')
```



More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Consider the following scenario.

- L is the weighted average classification loss.
- n is the sample size.
- For binary classification:
  - y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
  - $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
  - $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.
- For algorithms that support multiclass classification (that is, $K \geq 3$):
  - y_j^* is a vector of $K – 1$ zeros, with 1 in the position corresponding to the true, observed class y_j. For example, if the true class of the second observation is the third class and $K = 4$, then $y_2^* = [0\ 0\ 1\ 0]'$. The order of the classes corresponds to the order in the `ClassNames` property of the input model.
  - $f(X_j)$ is the length K vector of class scores for observation j of the predictor data X. The order of the scores corresponds to the order of the classes in the `ClassNames` property of the input model.
  - $m_j = y_j^{*\prime} f(X_j)$. Therefore, m_j is the scalar classification score that the model predicts for the true, observed class.
- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so they sum to 1. Therefore,

$$\sum_{j=1}^{n} w_j = 1.$$

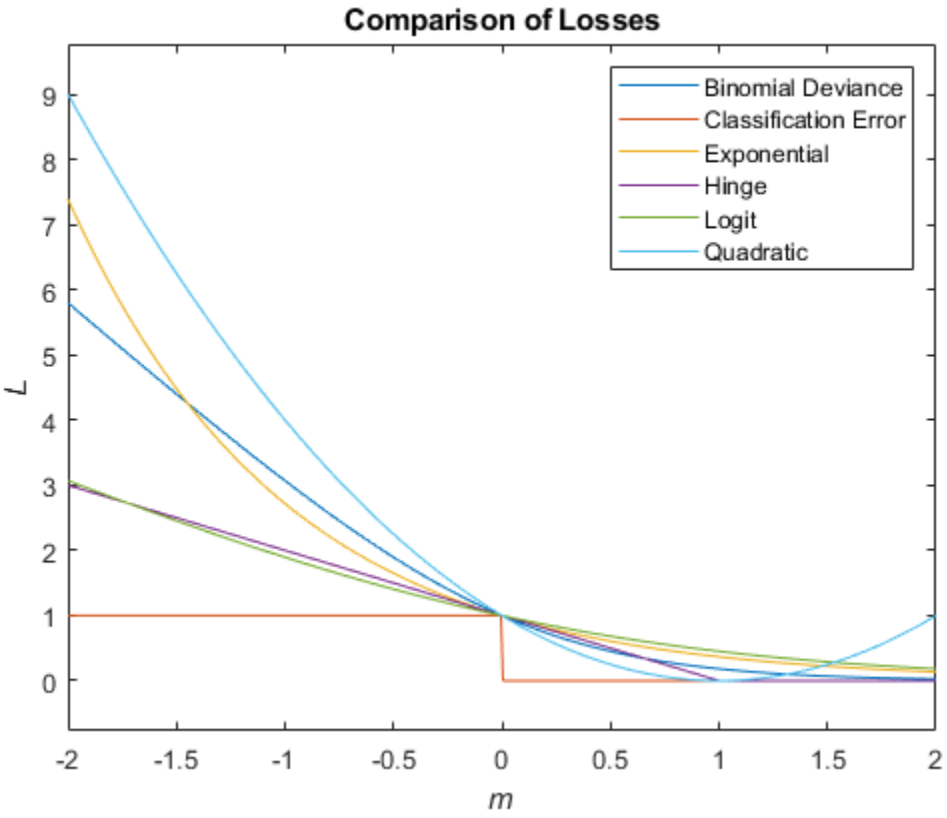
Given this scenario, the following table describes the supported loss functions that you can specify by using the `'LossFun'` name-value pair argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\left\{1 + \exp\left[-2m_j\right]\right\}.$

Loss Function	Value of LossFun	Equation
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}$. $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Cross-entropy loss	`'crossentropy'`	`'crossentropy'` is appropriate only for neural network models. The weighted cross-entropy loss is $L = -\sum_{j=1}^{n} \frac{\tilde{w}_j \log(m_j)}{Kn}$, where the weights $\tilde{w}_j$ are normalized to sum to n instead of 1.
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j)$.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}$.
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j))$.

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'crossentropy'` and `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



True Misclassification Cost

The true misclassification cost is the cost of classifying an observation into an incorrect class.

You can set the true misclassification cost per class by using the `'Cost'` name-value argument when you create the classifier. `Cost(i,j)` is the cost of classifying an observation into class `j` when its true class is `i`. By default, `Cost(i,j)=1` if `i~=j`, and `Cost(i,j)=0` if `i=j`. In other words, the cost is `0` for correct classification and `1` for incorrect classification.

Expected Misclassification Cost

The expected misclassification cost per observation is an averaged cost of classifying the observation into each class.

Suppose you have `Nobs` observations that you want to classify with a trained classifier, and you have `K` classes. You place the observations into a matrix `X` with one observation per row.

The expected cost matrix `CE` has size `Nobs`-by-`K`. Each row of `CE` contains the expected (average) cost of classifying the observation into each of the `K` classes. `CE(n,k)` is

$$\sum_{i=1}^{K}\widehat{P}(i|X(n))C(k|i),$$

where:

- K is the number of classes.
- $\widehat{P}(i|X(n))$ is the posterior probability of class i for observation $X(n)$.
- $C(k|i)$ is the true misclassification cost of classifying an observation as k when its true class is i.

See Also

`fitctree` | `loss` | `resubEdge` | `resubMargin` | `resubPredict`

resubLoss

Regression error by resubstitution

Syntax

```
L = resubLoss(ens)
L = resubLoss(ens,Name,Value)
```

Description

`L = resubLoss(ens)` returns the resubstitution loss, meaning the mean squared error computed for the data that `fitrensemble` used to create `ens`.

`L = resubLoss(ens,Name,Value)` calculates loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `resubLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

`lossfun`

Function handle for loss function, or `'mse'`, meaning mean squared error. If you pass a function handle `fun`, `resubLoss` calls it as

```
FUN(Y,Yfit,W)
```

where `Y`, `Yfit`, and `W` are numeric vectors of the same length. `Y` is the observed response, `Yfit` is the predicted response, and `W` is the observation weights.

Default: `'mse'`

`mode`

Character vector or string scalar representing the meaning of the output `L`:

- `'ensemble'` — `L` is a scalar value, the loss for the entire ensemble.
- `'individual'` — `L` is a vector with one element per trained learner.
- `'cumulative'` — `L` is a vector in which element `J` is obtained by using learners `1:J` from the input list of learners.

Default: `'ensemble'`

Output Arguments

L

Loss, by default the mean squared error. `L` can be a vector, and can mean different things, depending on the name-value pair settings.

Examples

Estimate Resubstitution Loss

Find the mean-squared difference between resubstitution predictions and training data.

Load the `carsmall` data set and select horsepower and vehicle weight as predictors.

```
load carsmall
X = [Horsepower Weight];
```

Train an ensemble of regression trees, and find the mean-squared difference of predictions from the training data.

```
ens = fitrensemble(X,MPG);
MSE = resubLoss(ens)
```

```
MSE = 0.5836
```

See Also

`loss` | `resubLoss` | `resubPredict`

resubLoss

Resubstitution regression loss

Syntax

```
L = resubLoss(Mdl)
L = resubLoss(Mdl,Name,Value)
```

Description

`L = resubLoss(Mdl)` returns the regression loss by resubstitution (L), or the in-sample regression loss, for the trained regression model `Mdl` using the training data stored in `Mdl.X` and the corresponding responses stored in `Mdl.Y`.

The interpretation of `L` depends on the loss function (`'LossFun'`) and weighting scheme (`Mdl.W`). In general, better models yield smaller loss values. The default `'LossFun'` value is `'mse'` (mean squared error).

`L = resubLoss(Mdl,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'IncludeInteractions',false` specifies to exclude interaction terms from a generalized additive model `Mdl`.

Examples

Resubstitution Loss

Train a generalized additive model (GAM), then calculate the resubstitution loss using the mean squared error (MSE).

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Train a univariate GAM that contains the linear terms for the predictors in `tbl`.

```
Mdl = fitrgam(tbl,"Systolic")
```

```
Mdl =
  RegressionGAM
             PredictorNames: {1x6 cell}
               ResponseName: 'Systolic'
    CategoricalPredictors: [3 5 6]
        ResponseTransform: 'none'
                Intercept: 122.7800
          NumObservations: 100
```

```
  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object.

Calculate the resubstitution loss using the mean squared error (MSE).

```
L = resubLoss(Mdl)
```

```
L = 4.1957
```

Compare GAMs by Examining Regression Loss

Train a generalized additive model (GAM) that contains both linear and interaction terms for predictors, and estimate the regression loss (mean squared error, MSE) with and without interaction terms for the training data and test data. Specify whether to include interaction terms when estimating the regression loss.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Partition the data set into two sets: one containing training data, and the other containing new, unobserved test data. Reserve 10 observations for the new test data set.

```
rng('default') % For reproducibility
n = size(X,1);
newInds = randsample(n,10);
inds = ~ismember(1:n,newInds);
XNew = X(newInds,:);
YNew = Y(newInds);
```

Train a generalized additive model that contains all the available linear and interaction terms in `X`.

```
Mdl = fitrgam(X(inds,:),Y(inds),'Interactions','all');
```

`Mdl` is a `RegressionGAM` model object.

Compute the resubstitution MSEs (that is, the in-sample MSEs) both with and without interaction terms in `Mdl`. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
resubl = resubLoss(Mdl)
```

```
resubl = 0.0292
```

```
resubl_nointeraction = resubLoss(Mdl,'IncludeInteractions',false)
```

```
resubl_nointeraction = 4.7330
```

Compute the regression MSEs both with and without interaction terms for the test data set. Use a memory-efficient model object for the computation.

```
CMdl = compact(Mdl);
```

`CMdl` is a `CompactRegressionGAM` model object.

```
l = loss(CMdl,XNew,YNew)
```

```
l = 12.8604
```

```
l_nointeraction = loss(CMdl,XNew,YNew,'IncludeInteractions',false)
```

```
l_nointeraction = 15.6741
```

Including interaction terms achieves a smaller error for the training data set and test data set.

Input Arguments

`Mdl` — Regression machine learning model

full regression model object

Regression machine learning model, specified as a full regression model object, as given in the following table of supported models.

Model	Regression Model Object
Generalized additive model	`RegressionGAM`
Neural network model	`RegressionNeuralNetwork`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubLoss(Mdl,'IncludeInteractions',false)` excludes interaction terms from a generalized additive model `Mdl`.

`IncludeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `RegressionGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Example: `'IncludeInteractions',false`

Data Types: `logical`

`LossFun` — Loss function

`'mse'` (default) | function handle

Loss function, specified as `'mse'` or a function handle.

- `'mse'` — Weighted mean squared error.
- Function handle — To specify a custom loss function, use a function handle. The function must have this form:

  ```
  lossval = lossfun(Y,YFit,W)
  ```

  - The output argument `lossval` is a floating-point scalar.
  - You specify the function name (*`lossfun`*).
  - `Y` is a length *n* numeric vector of observed responses, where *n* is the number of observations in `Tbl` or `X`.
  - `YFit` is a length *n* numeric vector of corresponding predicted responses.
  - `W` is an *n*-by-1 numeric vector of observation weights.

Example: `'LossFun',@lossfun`

Data Types: `char` | `string` | `function_handle`

More About

Weighted Mean Squared Error

The weighted mean squared error measures the predictive inaccuracy of regression models. When you compare the same type of loss among many models, a lower error indicates a better predictive model.

The weighted mean squared error is calculated as follows:

$$\text{mse} = \frac{\sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j\right)^2}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of rows of data.
- x_j is the jth row of data.
- y_j is the true response to x_j.
- $f(x_j)$ is the response prediction of the model `Mdl` to x_j.
- w is the vector of observation weights.

Algorithms

`resubLoss` computes the regression loss according to the corresponding `loss` function of the object (`Mdl`). For a model-specific description, see the `loss` function reference pages in the following table.

Model	Regression Model Object (`Mdl`)	`loss` Object Function
Generalized additive model	`RegressionGAM`	`loss`
Neural network model	`RegressionNeuralNetwork`	`loss`

See Also

`resubPredict`

Introduced in R2021a

resubLoss

Class: `RegressionGP`

Resubstitution loss for a trained Gaussian process regression model

Syntax

```
L = resubLoss(gprMdl)
L = resubLoss(gprMdl,Name,Value)
```

Description

`L = resubLoss(gprMdl)` returns the resubstitution mean squared error for the Gaussian process regression (GPR) model, `gprMdl`.

`L = resubLoss(gprMdl,Name,Value)` returns the resubstitution loss for the GPR model, `gprMdl`, with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify a custom loss function or the observation weights.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`lossfun` — Loss function

`'mse'` (default) | function handle

Loss function for the trained GPR model, specified as `'mse'` or a function handle. `'mse'` stands for the mean squared error.

If you pass a function handle, say `fun`, `resubLoss` calls it as : `fun(Y,Ypred,W)`, where `Y`, `Ypred`, and `W` are numeric vectors of length *n*, and *n* is the number of observations in the training data. `Y` is the observed response, `Ypred` is the predicted response, and `W` is the observation weights.

Example: `'lossfun',Fct` calls the loss function `Fct`.

Data Types: `char` | `string` | `function_handle`

`weights` — Observation weights

vector of 1s (default) | *n*-by-1 vector

Observation weights, specified as an *n*-by-1 vector, where *n* is the number of observations in the training data. By default, weight of each observation is 1.

Data Types: `double` | `single`

Output Arguments

L — Resubstitution error

scalar value

Resubstitution error for the GPR model, returned as a scalar value.

Examples

Compute the Resubstitution Loss

This example uses "Housing" data set [1] from the UCI machine learning archive [2] described in http://archive.ics.uci.edu/ml/datasets/Housing. Download the data and save it in your current directory as a data file named `housing.data`.

The dataset has 506 observations. The first 13 columns contain the predictor values and the last column contains the response values. The goal is to predict the median value of owner-occupied homes in the Boston suburb area as a function of 13 predictors.

Load the data and define the response vector and predictor matrix.

```
load('housing.data');
X = housing(:,1:13);
y = housing(:,end);
```

Fit a GPR model using subset of regressors (`'sr'`) approximation method with Matern 3/2 (`'Matern32'`) kernel function. Predict using the fully independent conditional (`'fic'`) method.

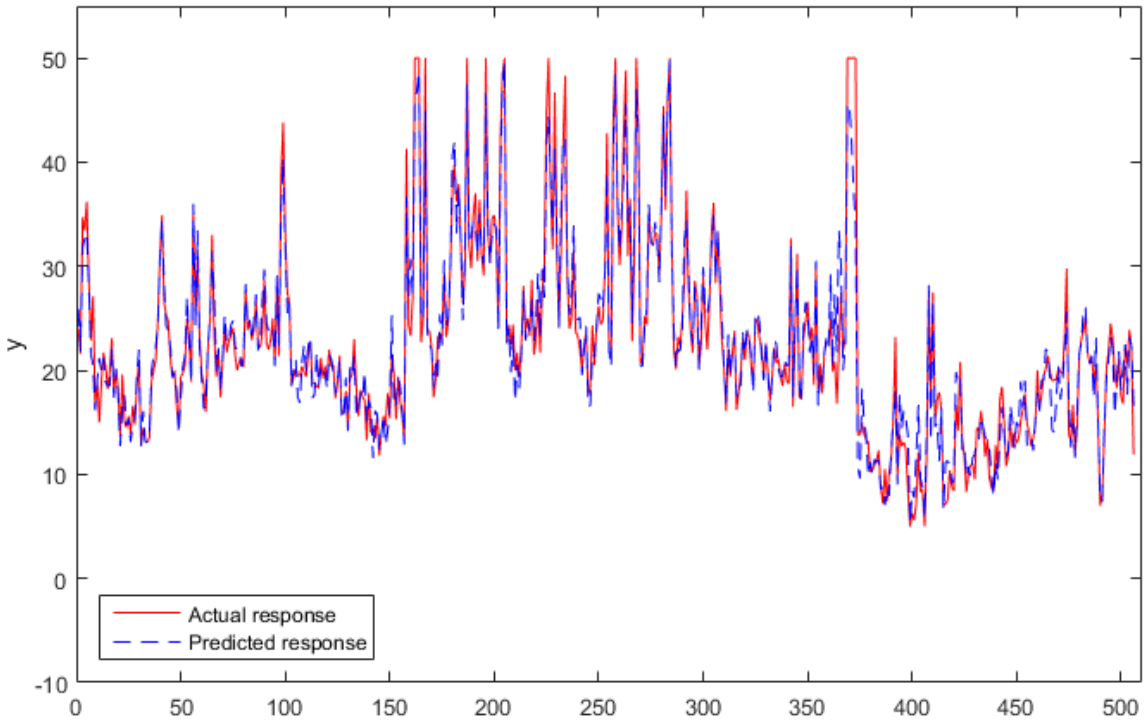
```
gprMdl = fitrgp(X,y,'KernelFunction','Matern32',...
'FitMethod','sr','PredictMethod','fic');
```

Compute the resubstitution predictions.

```
ypred = resubPredict(gprMdl);
```

Plot the predicted response values along with the actual response values.

```
plot(y,'r'); % Plot original response values
hold on;
plot(ypred,'b--'); % Plot predicted response values
ylabel('y');
legend('Actual response','Predicted response','Location','SouthWest');
axis([0 510 -10 55]);
hold off
```



Compute the resubstitution loss.

```
L = resubLoss(gprMdl)
```

```
L =

    4.8478
```

Manually compute the regression loss.

```
n = length(y);
L = (y-ypred)'*(y-ypred)/n
```

```
L =

    4.8478
```

Compute Custom Resubstitution Loss

Load the sample data and store in a `table`.

```
load fisheriris
tbl = table(meas(:,1),meas(:,2),meas(:,3),meas(:,4),species,...
'VariableNames',{'meas1','meas2','meas3','meas4','species'});
```

Fit a GPR model using the first measurement as the response and the other variables as the predictors.

```
mdl = fitrgp(tbl,'meas1');
```

Predict the responses using the trained model.

```
ypred = predict(mdl,tbl);
```

Compute the mean absolute error.

```
n = height(tbl);
y = tbl.meas1;
fun = @(y,ypred,w) sum(abs(y-ypred))/n;
L = resubLoss(mdl,'lossfun',fun)
```

```
L = 0.2345
```

Alternatives

To compute the regression error for new data, use `loss`.

References

[1] Harrison, D. and D.L., Rubinfeld. "Hedonic prices and the demand for clean air." *J. Environ. Economics & Management*. Vol.5, 1978, pp. 81-102.

[2] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

See Also

`RegressionGP` | `fitrgp` | `loss` | `resubPredict`

Introduced in R2015b

resubLoss

Class: RegressionSVM

Resubstitution loss for support vector machine regression model

Syntax

```
L = resubLoss(mdl)
L = resubLoss(mdl,Name,Value)
```

Description

`L = resubLoss(mdl)` returns the resubstitution loss for the support vector machine (SVM) regression model `mdl`, using the training data stored in `mdl.X` and corresponding response values stored in `mdl.Y`.

`L = resubLoss(mdl,Name,Value)` returns the resubstitution loss with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the loss function or observation weights.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function

`'mse'` (default) | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'mse'`, `'epsiloninsensitive'`, or a function handle.

- The following table lists the available loss functions. Specify one using its corresponding value.

Value	Loss Function
`'mse'`	“Mean Squared Error” on page 33-5588
`'epsiloninsensitive'`	“Epsilon-Insensitive Loss Function” on page 33-5588

- Specify your own function using function handle notation.

  Suppose that `n = size(X,1)` is the sample size. Your function must have the signature `lossvalue = lossfun(Y,Yfit,W)`, where:

- The output argument `lossvalue` is a numeric value.
- You choose the function name (*`lossfun`*).
- `Y` is an *n*-by-1 numeric vector of observed response values.
- `Yfit` is an *n*-by-1 numeric vector of predicted response values, calculated using the corresponding predictor values in `X` (similar to the output of `predict`).
- `W` is an *n*-by-1 numeric vector of observation weights.

Specify your function using `'LossFun',@`*`lossfun`*.

Example: `'LossFun','epsiloninsensitive'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector. `Weights` must be the same length as the number of rows in `X`. The software weighs the observations in each row of `X` using the corresponding weight value in `Weights`.

Data Types: `single` | `double`

Output Arguments

`L` — Resubstitution loss
scalar value

Resubstitution loss, returned as a scalar value.

The resubstitution loss is the loss calculated between the response training data and the model's predicted response values based on the input training data.

Resubstitution loss can be an overly optimistic estimate of the predictive error on new data. If the resubstitution loss is high, the model's predictions are not likely to be very good. However, having a low resubstitution loss does not guarantee good predictions for new data.

To better assess the predictive accuracy of your model, cross validate the model using `crossval`.

Examples

Resubstitution Loss for SVM Regression Model

This example shows how to train an SVM regression model, then calculate the resubstitution loss using mean square error (MSE) and epsilon-insensitive loss.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`.

Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model to the data, using a Gaussian kernel function with an automatic kernel scale. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale','auto','Standardize',true);
```

Calculate the resubstitution loss using mean square error (MSE).

```
mse_loss = resubLoss(mdl)
```

```
mse_loss =

    4.0603
```

Calculate the epsilon-insensitive loss.

```
eps_loss = resubLoss(mdl,'LossFun','epsiloninsensitive')
```

```
eps_loss =

    1.1027
```

More About

Mean Squared Error

The weighted mean squared error is calculated as follows:

$$\text{mse} = \frac{\sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j\right)^2}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of rows of data
- x_j is the jth row of data
- y_j is the true response to x_j
- $f(x_j)$ is the response prediction of the SVM regression model `mdl` to x_j
- w is the vector of weights.

The weights in w are all equal to one by default. You can specify different values for weights using the `'Weights'` name-value pair argument. If you specify weights, each value is divided by the sum of all weights, such that the normalized weights add to one.

Epsilon-Insensitive Loss Function

The epsilon-insensitive loss function ignores errors that are within the distance epsilon (ε) of the function value. It is formally described as:

$$Loss_\varepsilon = \begin{cases} 0\,,\ if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon\,,\ otherwise\,. \end{cases}$$

The mean epsilon-insensitive loss is calculated as follows:

$$Loss = \frac{\sum_{j=1}^{n} w_j \max\left(0, \left|y_j - f\left(x_j\right)\right| - \varepsilon\right)}{\sum_{j=1}^{n} w_j},$$

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. Adams. "A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

See Also

`RegressionSVM` | `fitrsvm` | `loss` | `resubPredict`

Introduced in R2015b

resubLoss

Class: RegressionTree

Regression error by resubstitution

Syntax

```
L = resubLoss(tree)
L = resubLoss(tree,Name,Value)
L = resubLoss(tree,'Subtrees',subtreevector)
[L,se] = resubLoss(tree,'Subtrees',subtreevector)
[L,se,NLeaf] = resubLoss(tree,'Subtrees',subtreevector)
[L,se,NLeaf,bestlevel] = resubLoss(tree,'Subtrees',subtreevector)
[L,...] = resubLoss(tree,'Subtrees',subtreevector,Name,Value)
```

Description

`L = resubLoss(tree)` returns the resubstitution loss, meaning the loss computed for the data that `fitrtree` used to create `tree`.

`L = resubLoss(tree,Name,Value)` returns the loss with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

`L = resubLoss(tree,'Subtrees',subtreevector)` returns a vector of mean squared errors for the trees in the pruning sequence `subtreevector`.

`[L,se] = resubLoss(tree,'Subtrees',subtreevector)` returns the vector of standard errors of the classification errors.

`[L,se,NLeaf] = resubLoss(tree,'Subtrees',subtreevector)` returns the vector of numbers of leaf nodes in the trees of the pruning sequence.

`[L,se,NLeaf,bestlevel] = resubLoss(tree,'Subtrees',subtreevector)` returns the best pruning level as defined in the `TreeSize` name-value pair. By default, `bestlevel` is the pruning level that gives loss within one standard deviation of minimal loss.

`[L,...] = resubLoss(tree,'Subtrees',subtreevector,Name,Value)` returns loss statistics with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`tree` — Regression tree

`RegressionTree` model object

A regression tree (`RegressionTree` model object) constructed using `fitrtree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`LossFun` — Loss function
`'mse'` (default) | function handle

Loss function, specified as a function handle or `'mse'` meaning mean squared error.

You can write your own loss function in the syntax described in “Loss Functions” on page 33-5595.

Data Types: `char` | `string` | `function_handle`

`Name,Value` arguments associated with pruning subtrees:

`Subtrees` — Pruning level
0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `resubLoss` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`resubLoss` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

`TreeSize` — Tree size
`'se'` (default) | `'min'`

Tree size, specified as one of the following:

- `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L` + `se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Output Arguments

`L` — Regression loss
numeric vector of positive values

Regression loss (mean squared error), a vector the length of `Subtrees`. The meaning of the error depends on the values in `Weights` and `LossFun`.

`se` — Standard error of loss
numeric vector of positive values

Standard error of loss, a vector the length of `Subtrees`.

`NLeaf` — Number of leaves
numeric vector of nonnegative integers

Number of leaves (terminal nodes) in the pruned subtrees, a vector the length of `Subtrees`.

`bestlevel` — Optimal pruning level
nonnegative numeric scalar

A scalar whose value depends on `TreeSize`:

- `TreeSize` = `'se'` — `loss` returns the highest pruning level with loss within one standard deviation of the minimum (`L` + `se`, where `L` and `se` relate to the smallest value in `Subtrees`).
- `TreeSize` = `'min'` — `loss` returns the element of `Subtrees` with smallest loss, usually the smallest element of `Subtrees`.

Examples

Compute the In-Sample MSE

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using all observations.

```
Mdl = fitrtree(X,MPG);
```

Compute the resubstitution MSE.

```
resubLoss(Mdl)
```

```
ans = 4.8952
```

Examine the MSE for Each Subtree

Unpruned decision trees tend to overfit. One way to balance model complexity and out-of-sample performance is to prune a tree (or restrict its growth) so that in-sample and out-of-sample performance are satisfactory.

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
Y = MPG;
```

Partition the data into training (50%) and validation (50%) sets.

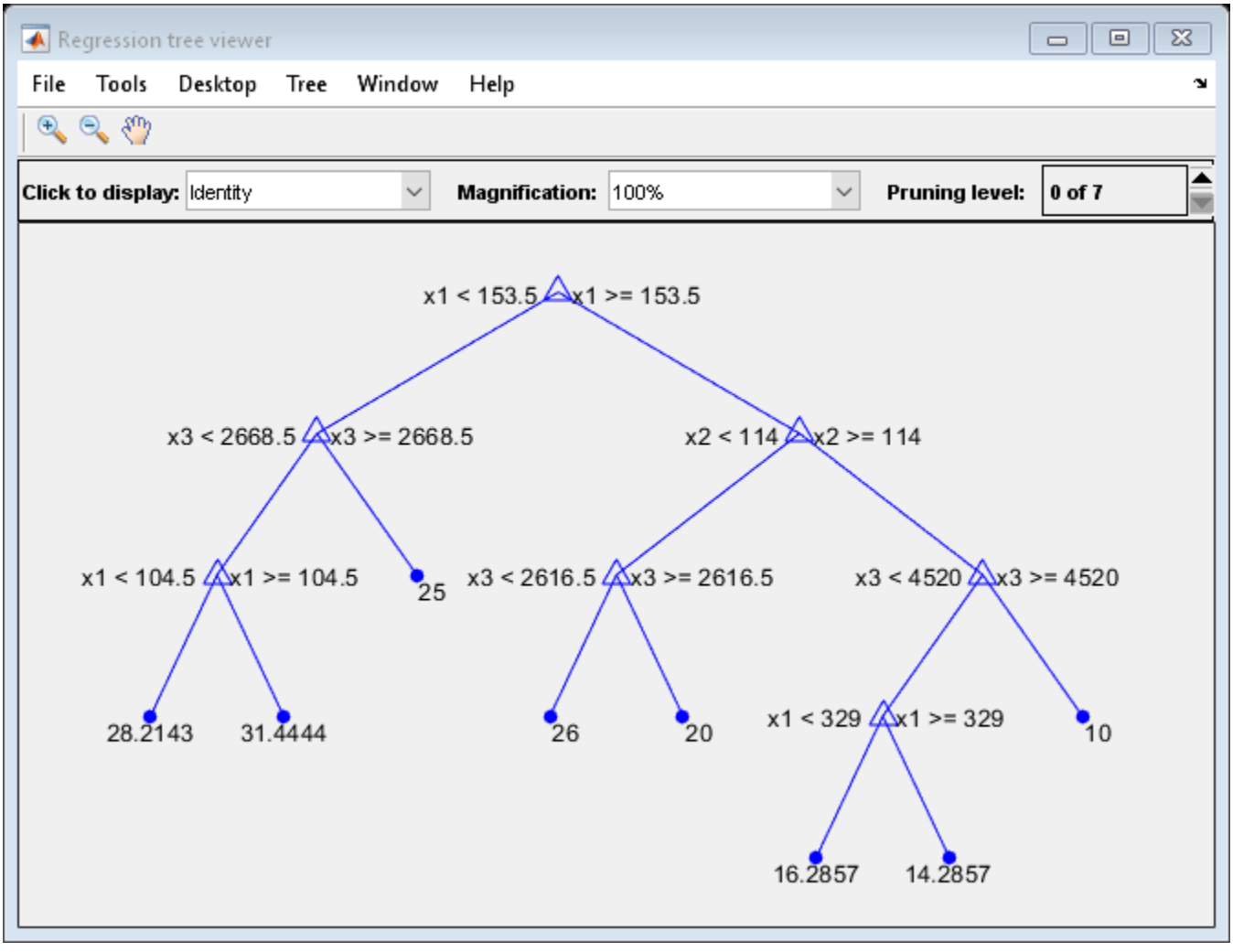
```
n = size(X,1);
rng(1) % For reproducibility
idxTrn = false(n,1);
idxTrn(randsample(n,round(0.5*n))) = true; % Training set logical indices
idxVal = idxTrn == false;                  % Validation set logical indices
```

Grow a regression tree using the training set.

```
Mdl = fitrtree(X(idxTrn,:),Y(idxTrn));
```

View the regression tree.

```
view(Mdl,'Mode','graph');
```



The regression tree has seven pruning levels. Level 0 is the full, unpruned tree (as displayed). Level 7 is just the root node (i.e., no splits).

Examine the training sample MSE for each subtree (or pruning level) excluding the highest level.

```
m = max(Mdl.PruneList) - 1;
trnLoss = resubLoss(Mdl,'SubTrees',0:m)
```

```
trnLoss = 7×1

    5.9789
    6.2768
    6.8316
    7.5209
    8.3951
   10.7452
   14.8445
```

- The MSE for the full, unpruned tree is about 6 units.
- The MSE for the tree pruned to level 1 is about 6.3 units.
- The MSE for the tree pruned to level 6 (i.e., a stump) is about 14.8 units.

Examine the validation sample MSE at each level excluding the highest level.

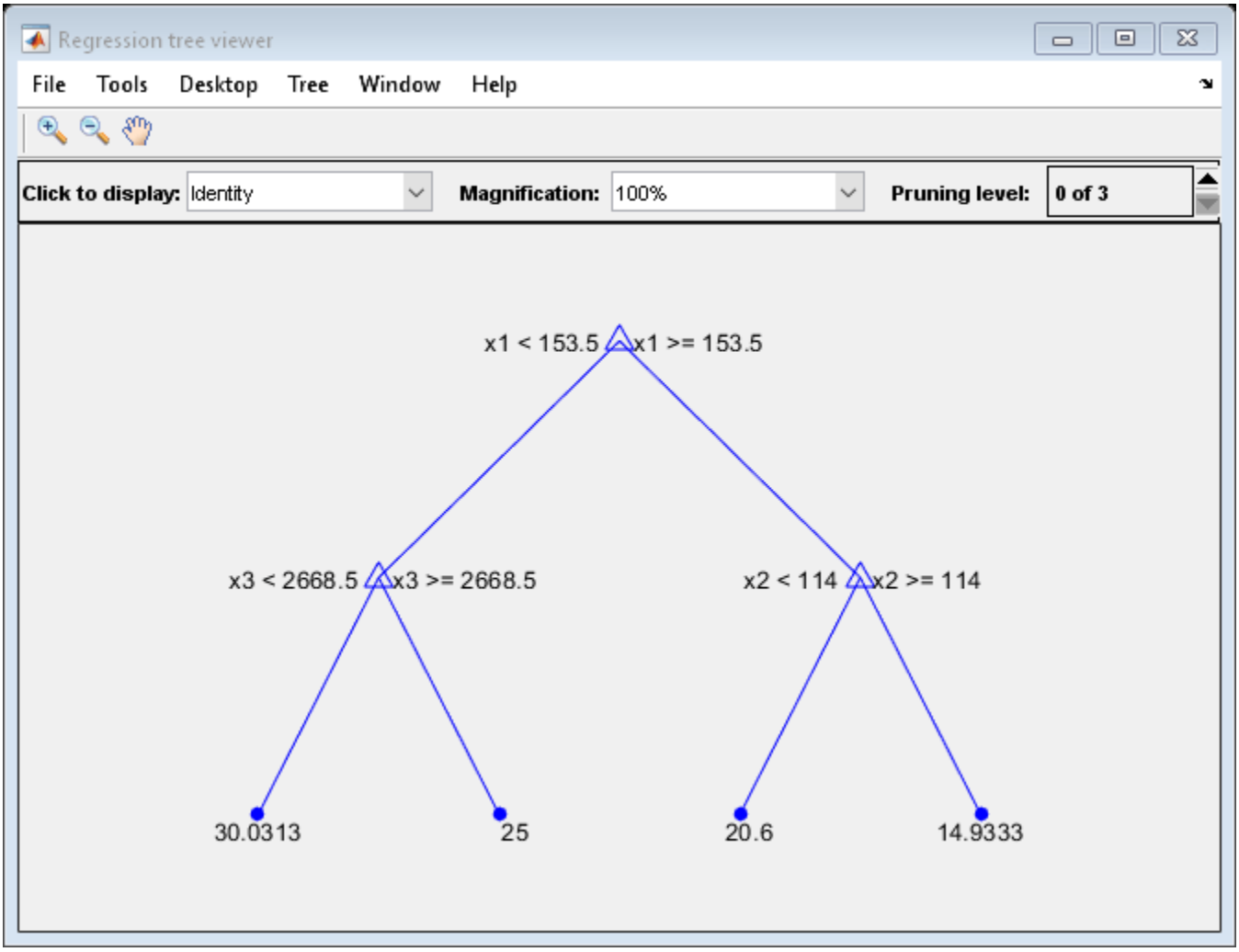
```
valLoss = loss(Mdl,X(idxVal,:),Y(idxVal),'SubTrees',0:m)
```

```
valLoss = 7×1

   32.1205
   31.5035
   32.0541
   30.8183
   26.3535
   30.0137
   38.4695
```

- The MSE for the full, unpruned tree (level 0) is about 32.1 units.
- The MSE for the tree pruned to level 4 is about 26.4 units.
- The MSE for the tree pruned to level 5 is about 30.0 units.
- The MSE for the tree pruned to level 6 (i.e., a stump) is about 38.5 units.

To balance model complexity and out-of-sample performance, consider pruning `Mdl` to level 4.

```
pruneMdl = prune(Mdl,'Level',4);
view(pruneMdl,'Mode','graph')
```



More About

Loss Functions

The built-in loss function is `'mse'`, meaning mean squared error.

To write your own loss function, create a function file of the form

`function loss = lossfun(Y,Yfit,W)`

- `N` is the number of rows of `tree.X`.
- `Y` is an `N`-element vector representing the observed response.
- `Yfit` is an `N`-element vector representing the predicted responses.
- `W` is an `N`-element vector representing the observation weights.
- The output `loss` should be a scalar.

Pass the function handle `@lossfun` as the value of the `LossFun` name-value pair.

See Also

`fitrtree` | `loss` | `resubPredict`

resubMargin

Resubstitution classification margin

Syntax

```
m = resubMargin(Mdl)
m = resubMargin(Mdl,'IncludeInteractions',includeInteractions)
```

Description

`m = resubMargin(Mdl)` returns the resubstitution "Classification Margin" on page 33-5602 (`m`) for the trained classification model `Mdl` using the predictor data stored in `Mdl.X` and the corresponding true class labels stored in `Mdl.Y`.

`m` is returned as an *n*-by-1 numeric column vector, where *n* is the number of observations in the predictor data.

`m = resubMargin(Mdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

Examples

Estimate Resubstitution Classification Margins of Naive Bayes Classifier

Estimate the resubstitution (in-sample) classification margins of a naive Bayes classifier. An observation margin is the observed true class score minus the maximum false class score among all scores in the respective class.

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
        DistributionNames: {'normal'  'normal'  'normal'  'normal'}
   DistributionParameters: {3x4 cell}
```

```
  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

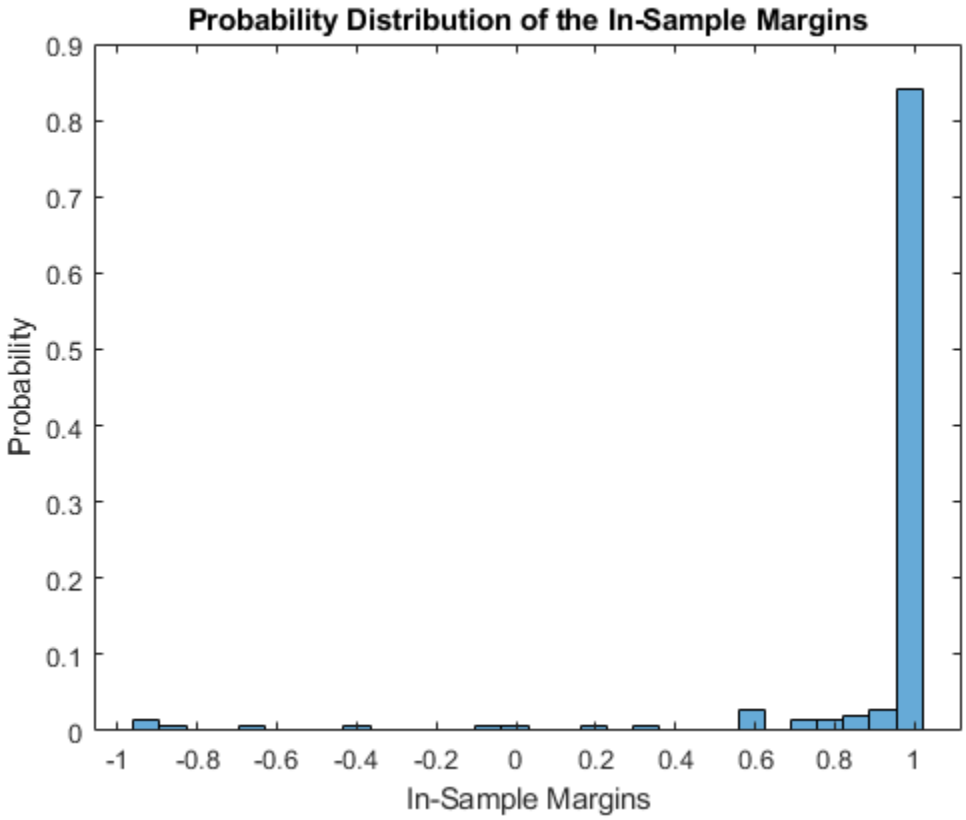
Estimate the resubstitution classification margins.

```
m = resubMargin(Mdl);
median(m)
```

```
ans = 1.0000
```

Display the histogram of the in-sample classification margins.

```
histogram(m,30,'Normalization','probability')
xlabel('In-Sample Margins')
ylabel('Probability')
title('Probability Distribution of the In-Sample Margins')
```



Classifiers that yield relatively large margins are preferred.

Select SVM Classifier Features by Examining In-Sample Margins

Perform feature selection by comparing in-sample margins from multiple models. Based solely on this comparison, the model with the highest margins is the best model.

Load the `ionosphere` data set. Define two data sets:

- `fullX` contains all predictors (except the removed column of 0s).
- `partX` contains the last 20 predictors.

```
load ionosphere
fullX = X;
partX = X(:,end-20:end);
```

Train a support vector machine (SVM) classifier for each predictor set.

```
FullSVMModel = fitcsvm(fullX,Y);
PartSVMModel = fitcsvm(partX,Y);
```

Estimate the in-sample margins for each classifier.

```
fullMargins = resubMargin(FullSVMModel);
partMargins = resubMargin(PartSVMModel);
n = size(X,1);
p = sum(fullMargins < partMargins)/n
```

```
p = 0.2222
```

Approximately 22% of the margins from the full model are less than those from the model with fewer predictors. This result suggests that the model trained with all the predictors is better.

Compare GAMs by Examining Training Sample Margins and Edge

Compare a generalized additive model (GAM) with linear terms to a GAM with both linear and interaction terms by examining the training sample margins and edge. Based solely on this comparison, the classifier with the highest margins and edge is the best model.

Load the 1994 census data stored in `census1994.mat`. The data set consists of demographic data from the US Census Bureau to predict whether an individual makes over $50,000 per year. The classification task is to fit a model that predicts the salary category of people given their age, working class, education level, marital status, race, and so on.

```
load census1994
```

`census1994` contains the training data set `adultdata` and the test data set `adulttest`. To reduce the running time for this example, subsample 500 training observations from `adultdata` by using the `datasample` function.

```
rng('default') % For reproducibility
NumSamples = 5e2;
adultdata = datasample(adultdata,NumSamples,'Replace',false);
```

Train a GAM that contains both linear and interaction terms for predictors. Specify to include all available interaction terms whose *p*-values are not greater than 0.05.

```
Mdl = fitcgam(adultdata,'salary','Interactions','all','MaxPValue',0.05)
```

```
Mdl =
  ClassificationGAM
             PredictorNames: {1x14 cell}
               ResponseName: 'salary'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
             ScoreTransform: 'logit'
                  Intercept: -32.0842
               Interactions: [82x2 double]
           NumObservations: 500


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object. `Mdl` includes 82 interaction terms.

Estimate the training sample margins and edge for `Mdl`.

```
M = resubMargin(Mdl);
E = resubEdge(Mdl)
```

```
E = 1.0000
```

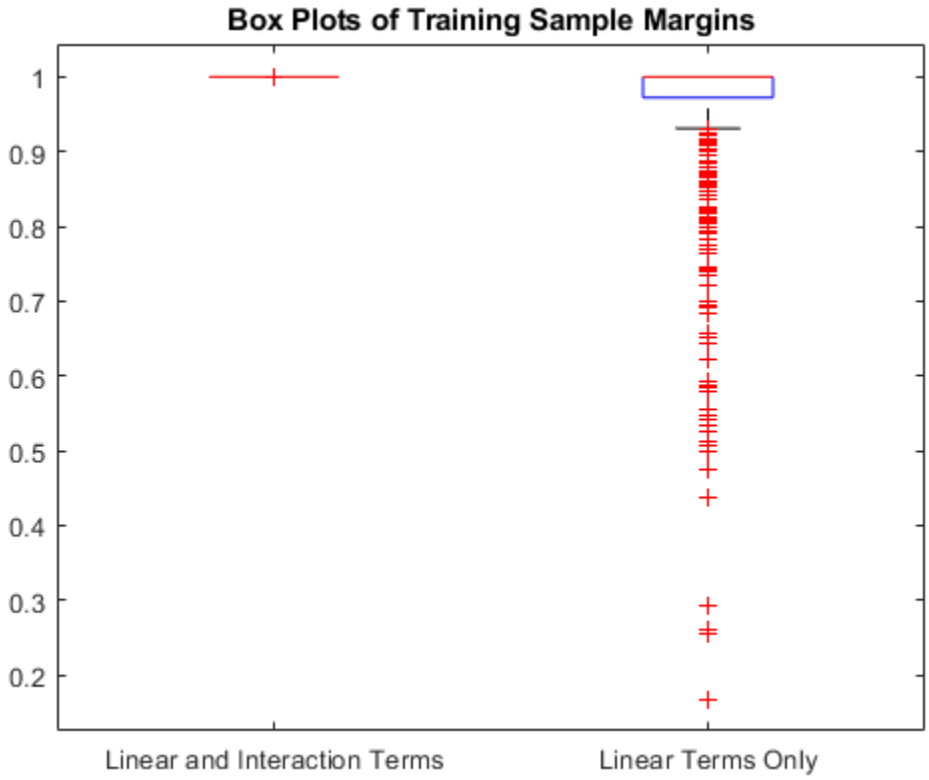
Estimate the training sample margins and edge for `Mdl` without including interaction terms.

```
M_nointeractions = resubMargin(Mdl,'IncludeInteractions',false);
E_nointeractions = resubEdge(Mdl,'IncludeInteractions',false)
```

```
E_nointeractions = 0.9516
```

Display the distributions of the margins using box plots.

```
boxplot([M M_nointeractions],'Labels',{'Linear and Interaction Terms','Linear Terms Only'})
title('Box Plots of Training Sample Margins')
```



When you include the interaction terms in the computation, all the resubstitution margin values for `Mdl` are 1, and the resubstitution edge value (average of the margins) is 1. The margins and edge decrease when you do not include the interaction terms in `Mdl`.

Input Arguments

`Mdl` — Classification machine learning model

full classification model object

Classification machine learning model, specified as a full classification model object, as given in the following table of supported models.

Model	Classification Model Object
Generalized additive model	`ClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

`includeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `ClassificationGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class. The classification margin for multiclass classification is the difference between the classification score for the true class and the maximal classification score for the false classes.

If the margins are on the same scale (that is, the score values are based on the same score transformation), then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Algorithms

`resubMargin` computes the classification margin according to the corresponding `margin` function of the object (`Mdl`). For a model-specific description, see the `margin` function reference pages in the following table.

Model	Classification Model Object (`Mdl`)	`margin` Object Function
Generalized additive model	`ClassificationGAM`	`margin`
k-nearest neighbor model	`ClassificationKNN`	`margin`
Naive Bayes model	`ClassificationNaiveBayes`	`margin`
Neural network model	`ClassificationNeuralNetwork`	`margin`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`margin`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports `ClassificationKNN` and `ClassificationSVM` objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`resubEdge` | `resubLoss` | `resubPredict`

Introduced in R2012a

resubMargin

Class: `ClassificationDiscriminant`

Classification margins by resubstitution

Syntax

```
M = resubMargin(obj)
```

Description

`M = resubMargin(obj)` returns resubstitution classification margins for `obj`.

Input Arguments

obj

Discriminant analysis classifier, produced using `fitcdiscr`.

Output Arguments

M

Numeric column-vector of length `size(obj.X,1)` containing the classification margins.

Examples

Estimate Resubstitution Margins for Discriminant Analysis Classifiers

Find the margins for a discriminant analysis classifier for Fisher's iris data by resubstitution. Examine several entries.

Load Fisher's iris data set.

```
load fisheriris
```

Train a discriminant analysis classifier.

```
Mdl = fitcdiscr(meas,species);
```

Compute the resubstitution margins, and display several of them.

```
m = resubMargin(Mdl);
m(1:25:end)
```

```
ans = 6×1

    1.0000
    1.0000
```

```
    0.9998
    0.9998
    1.0000
    0.9946
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes.

The classification margin is a column vector with the same number of rows as in the matrix `X`. A high value of margin indicates a more reliable prediction than a low value.

Score

For discriminant analysis, the score of a classification is the posterior probability of the classification. For the definition of posterior probability in discriminant analysis, see “Posterior Probability” on page 20-6.

See Also

`ClassificationDiscriminant` | `fitcdiscr` | `margin`

Topics

“Discriminant Analysis Classification” on page 20-2

resubMargin

Resubstitution classification margins for multiclass error-correcting output codes (ECOC) model

Syntax

```
m = resubMargin(Mdl)
m = resubMargin(Mdl,Name,Value)
```

Description

`m = resubMargin(Mdl)` returns the resubstitution classification margins on page 33-5611 (`m`) for the multiclass error-correcting output codes (ECOC) model `Mdl` using the training data stored in `Mdl.X` and the corresponding class labels stored in `Mdl.Y`.

`m` is returned as a numeric column vector with the same length as `Mdl.Y`. The software estimates each entry of `m` using the trained ECOC model `Mdl`, the corresponding row of `Mdl.X`, and the true class label `Mdl.Y`.

`m = resubMargin(Mdl,Name,Value)` returns the classification margins with additional options specified by one or more name-value pair arguments. For example, you can specify a decoding scheme, binary learner loss function, and verbosity level.

Examples

Resubstitution Classification Margins of ECOC Model

Calculate the resubstitution classification margins for an ECOC model with SVM binary learners.

Load Fisher's iris data set. Specify the predictor data `X` and the response data `Y`.

```
load fisheriris
X = meas;
Y = species;
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
classOrder = unique(Y)
```

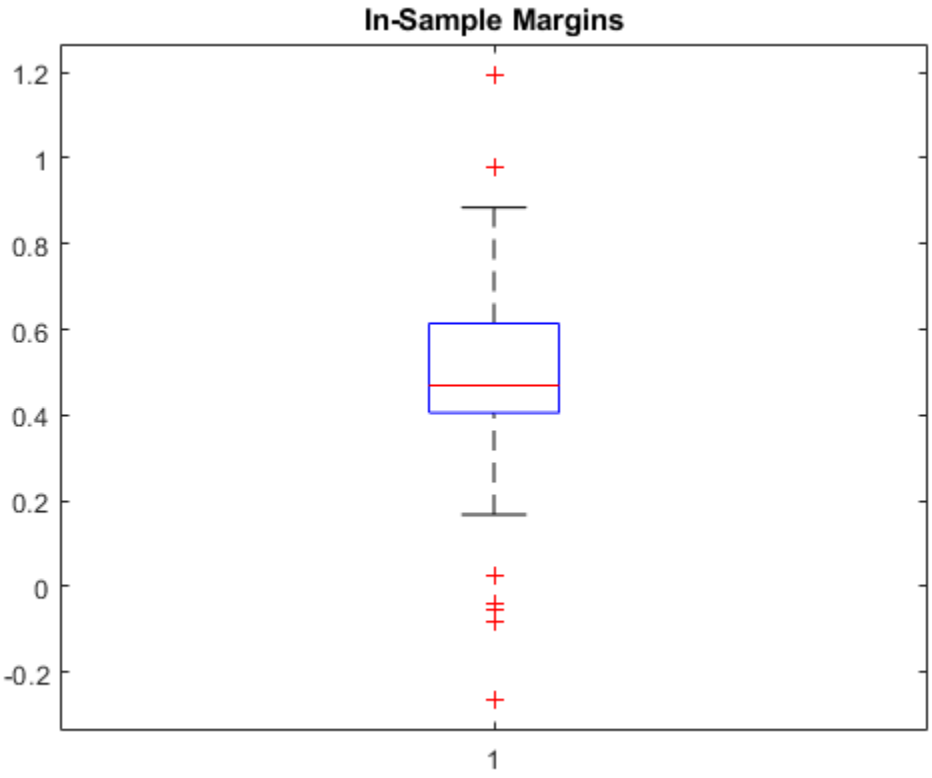
```
classOrder = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

Calculate the classification margins for the observations used to train `Mdl`. Display the distribution of the margins using a boxplot.

```
m = resubMargin(Mdl);

boxplot(m)
title('In-Sample Margins')
```



The classification margin of an observation is the positive-class negated loss minus the maximum negative-class negated loss. Choose classifiers that yield relatively large margins.

Select ECOC Model Features by Examining Training-Sample Margins

Perform feature selection by comparing training-sample margins from multiple models. Based solely on this comparison, the model with the greatest margins is the best model.

Load Fisher's iris data set. Define two data sets:

- `fullX` contains all four predictors.
- `partX` contains the sepal measurements only.

```
load fisheriris
X = meas;
fullX = X;
```

```
partX = X(:,1:2);
Y = species;
```

Train an ECOC model using SVM binary learners for each predictor set. Standardize the predictors using an SVM template, specify the class order, and compute posterior probabilities.

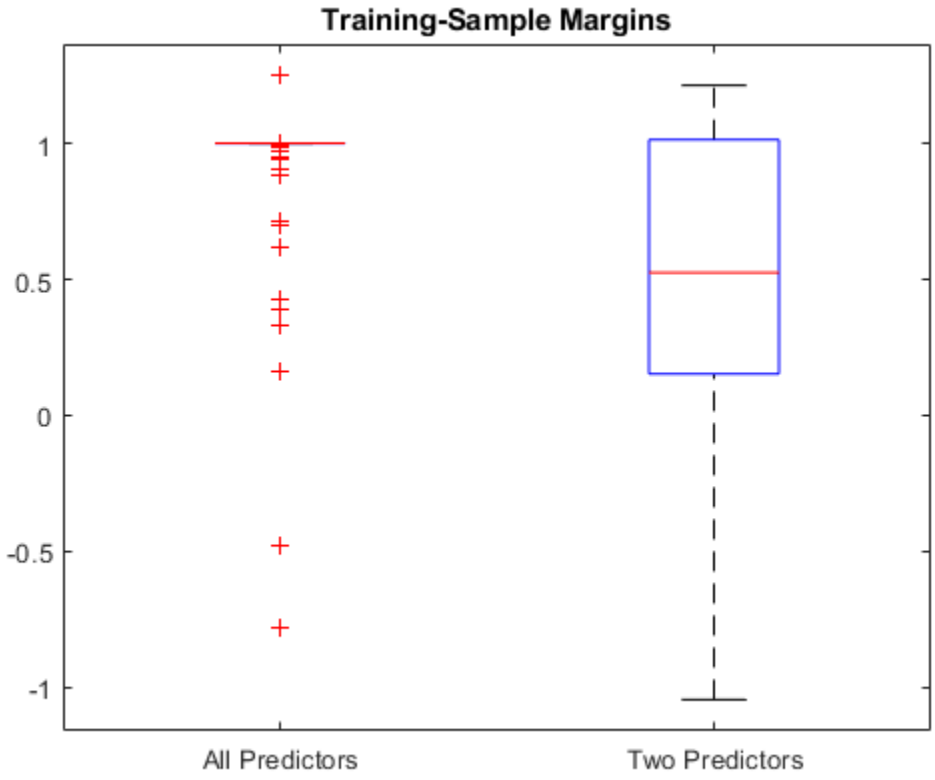
```
t = templateSVM('Standardize',true);
classOrder = unique(Y)
```

```
classOrder = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
FullMdl = fitcecoc(fullX,Y,'Learners',t,'ClassNames',classOrder,...
    'FitPosterior',true);
PartMdl = fitcecoc(partX,Y,'Learners',t,'ClassNames',classOrder,...
    'FitPosterior',true);
```

Compute the resubstitution margins for each classifier. For each model, display the distribution of the margins using a boxplot.

```
fullMargins = resubMargin(FullMdl);
partMargins = resubMargin(PartMdl);

boxplot([fullMargins partMargins],'Labels',{'All Predictors','Two Predictors'})
title('Training-Sample Margins')
```



The margin distribution of `FullMdl` is situated higher and has less variability than the margin distribution of `PartMdl`. This result suggests that the model trained with all the predictors fits the training data better.

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubMargin(Mdl,'Verbose',1)` specifies to display diagnostic messages in the Command Window.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

  `customFunction` has this form:

  ```
  bLoss = customFunction(M,s)
  ```

where:

- `M` is the *K*-by-*L* coding matrix stored in `Mdl.CodingMatrix`.
- `s` is the 1-by-*L* row vector of classification scores.
- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-5611.

Example: `'Decoding','lossbased'`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.

- Specify `'Options',statset('UseParallel',true)`.

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

More About

Classification Margin

The classification margin is, for each observation, the difference between the negative loss for the true class and the maximal negative loss among the false classes. If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|} .$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Tips

- To compare the margins or edges of several ECOC classifiers, use template objects to specify a common score transform function among the classifiers during training.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[3] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `fitcecoc` | `margin` | `predict` | `resubEdge` | `resubLoss` | `resubPredict`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

resubMargin

Classification margins by resubstitution

Syntax

```
margin = resubMargin(ens)
margin = resubMargin(ens,Name,Value)
```

Description

`margin = resubMargin(ens)` returns the classification margin obtained by `ens` on its training data.

`margin = resubMargin(ens,Name,Value)` calculates margins with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A classification ensemble created with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `resubMargin` uses only these learners for calculating margin.

Default: `1:NumTrained`

Output Arguments

`margin`

A numeric column-vector of length `size(ens.X,1)` containing the classification margins.

Examples

Compute Resubstitution Margins for Classification Ensemble

Find the resubstitution margins for an ensemble that classifies the Fisher iris data.

Load the Fisher iris data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Find the resubstitution margins.

```
margin = resubMargin(ens);
[min(margin) mean(margin) max(margin)]
```

```
ans = 1×3

   -0.5674    3.2486    4.6245
```

More About

Margin

The classification margin is the difference between the classification score for the true class and maximal classification score for the false classes. Margin is a column vector with the same number of rows as in the matrix `ens.X`.

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

See Also

`resubEdge` | `resubLoss` | `resubMargin` | `resubPredict`

resubMargin

Class: `ClassificationTree`

Classification margins by resubstitution

Syntax

```
M = resubMargin(tree)
```

Description

`M = resubMargin(tree)` returns resubstitution classification margins for `tree`.

Input Arguments

`tree`

A classification tree created by `fitctree`.

Output Arguments

`M`

A numeric column-vector of length `size(tree.X,1)` containing the classification margins.

Examples

Find the margins for a classification tree for the Fisher iris data by resubstitution. Examine several entries:

```
load fisheriris
tree = fitctree(meas,species);
M = resubMargin(tree);
M(1:25:end)
```

```
ans =
    1.0000
    1.0000
    1.0000
    1.0000
    0.9565
    0.9565
```

More About

Margin

Classification margin is the difference between classification score for the true class and maximal classification score for the false classes. A high value of margin indicates a more reliable prediction than a low value.

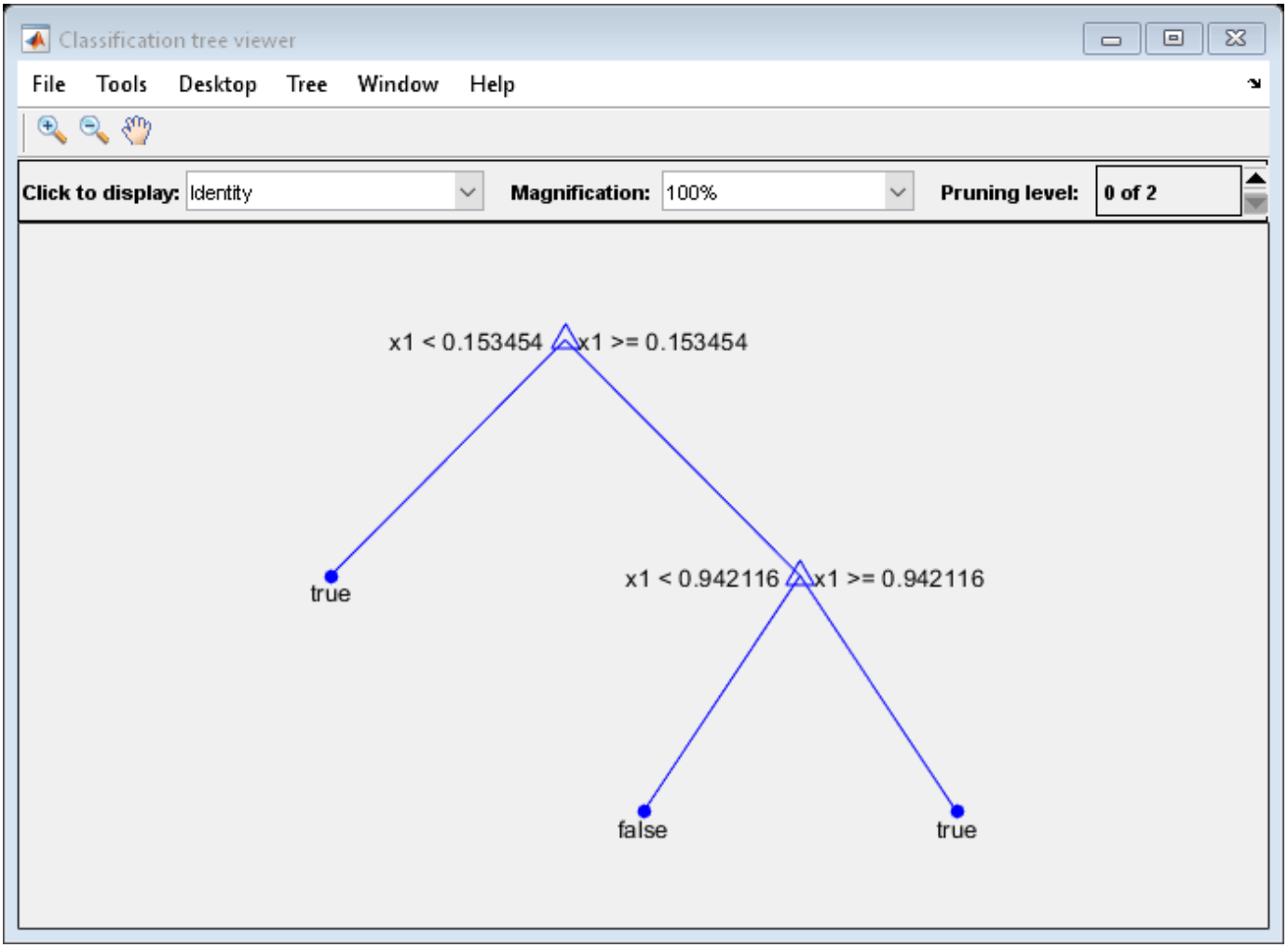
Score (tree)

For trees, the score of a classification of a leaf node is the posterior probability of the classification at that node. The posterior probability of the classification at a node is the number of training sequences that lead to that node with the classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor `X` as `true` when `X < 0.15` or `X > 0.95`, and `X` is false otherwise.

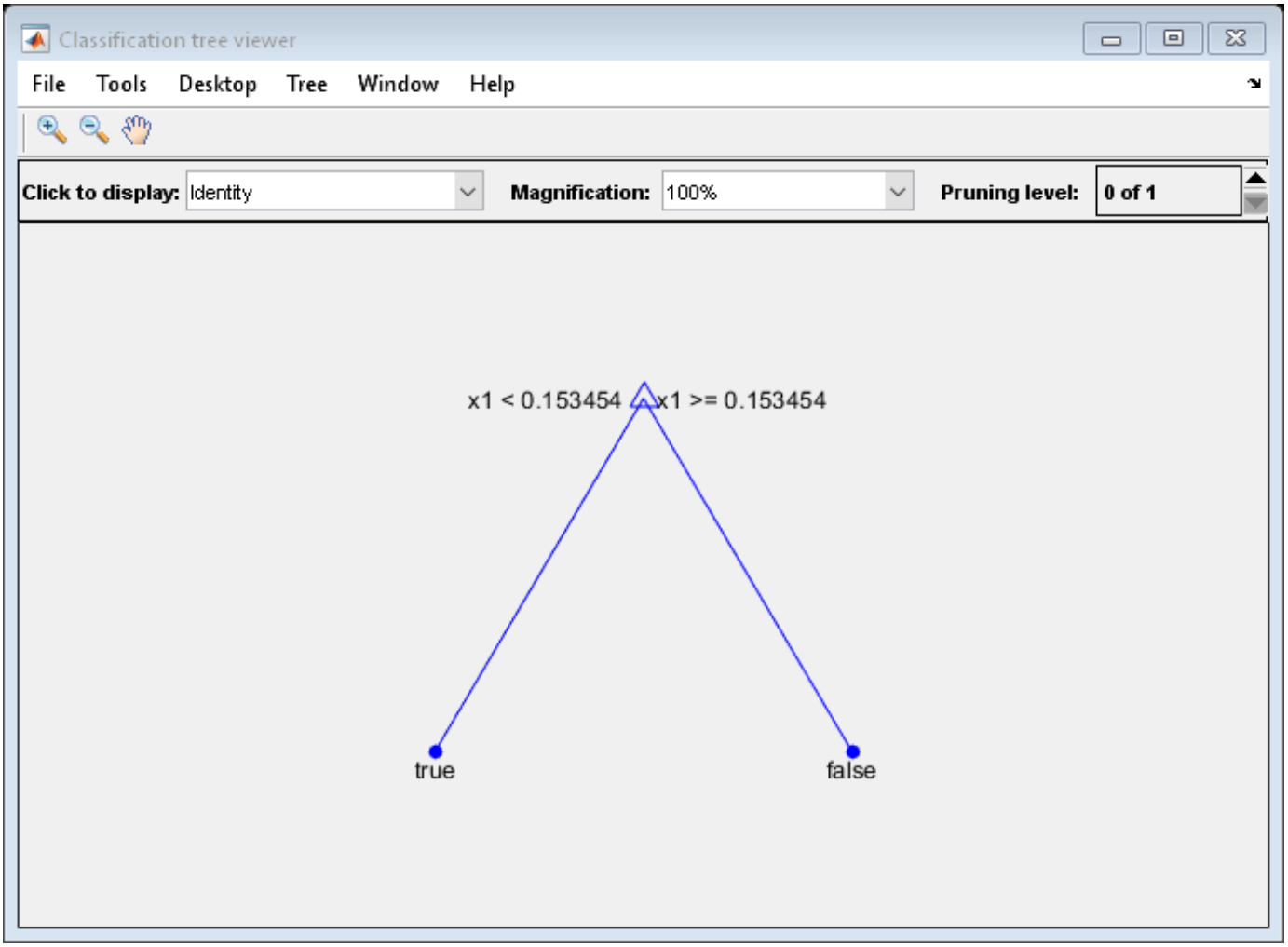
Generate 100 random points and classify them:

```
rng(0,'twister') % for reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','Graph')
```



Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','Graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations from .15 to .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore, the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

Compute the prediction scores for the first 10 rows of X:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the right-most column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.91` and `0.09`. The difference (score `0.09` instead of the expected `.06`) is due to a statistical fluctuation: there are `8` observations in `X` in the range `(.95,1)` instead of the expected `5` observations.

See Also

`fitctree` | `margin` | `resubEdge` | `resubLoss` | `resubPredict`

resubPredict

Class: `ClassificationDiscriminant`

Predict resubstitution labels of discriminant analysis classification model

Syntax

```
label = resubPredict(obj)
[label,posterior] = resubPredict(obj)
[label,posterior,cost] = resubPredict(obj)
```

Description

`label = resubPredict(obj)` returns the labels `obj` predicts for the data `obj.X`. `label` is the predictions of `obj` on the data that `fitcdiscr` used to create `obj`.

`[label,posterior] = resubPredict(obj)` returns the posterior class probabilities for the predictions.

`[label,posterior,cost] = resubPredict(obj)` returns the predicted misclassification costs per class for the resubstituted data.

Input Arguments

`obj`

Discriminant analysis classifier, produced using `fitcdiscr`.

Output Arguments

`label`

Response `obj` predicts for the training data. `label` is the same data type as the training response data `obj.Y`. The predicted class labels are those with minimal expected misclassification cost; see “Prediction Using Discriminant Analysis Models” on page 20-6.

`posterior`

`N`-by-`K` matrix of posterior probabilities for classes `obj` predicts, where `N` is the number of observations and `K` is the number of classes.

`cost`

`N`-by-`K` matrix of predicted misclassification costs. Each cost is the average misclassification cost with respect to the posterior probability.

Examples

Find the total number of misclassifications of the Fisher iris data for a discriminant analysis classifier:

```
load fisheriris
obj = fitcdiscr(meas,species);
Ypredict = resubPredict(obj); % the predictions
Ysame = strcmp(Ypredict,species); % true when ==
sum(~Ysame) % how many are different?

ans =
     3
```

More About

Posterior Probability

`posterior(i,k)` is the posterior probability of class `k` for observation `i`. For the mathematical definition, see “Posterior Probability” on page 20-6.

See Also

`ClassificationDiscriminant` | `fitcdiscr` | `predict`

Topics

“Discriminant Analysis Classification” on page 20-2

resubPredict

Classify observations in multiclass error-correcting output codes (ECOC) model

Syntax

```
label = resubPredict(Mdl)
label = resubPredict(Mdl,Name,Value)
[label,NegLoss,PBScore] = resubPredict( ___ )
[label,NegLoss,PBScore,Posterior] = resubPredict( ___ )
```

Description

`label = resubPredict(Mdl)` returns a vector of predicted class labels (`label`) for the trained multiclass error-correcting output codes (ECOC) model `Mdl` using the predictor data stored in `Mdl.X`.

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`label = resubPredict(Mdl,Name,Value)` returns predicted class labels with additional options specified by one or more name-value pair arguments. For example, specify the posterior probability estimation method, decoding scheme, or verbosity level.

`[label,NegLoss,PBScore] = resubPredict( ___ )` uses any of the input argument combinations in the previous syntaxes and additionally returns the negated average binary loss on page 33-5632 per class (`NegLoss`) for observations, and the positive-class scores (`PBScore`) for the observations classified by each binary learner.

`[label,NegLoss,PBScore,Posterior] = resubPredict( ___ )` additionally returns posterior class probability estimates for observations (`Posterior`).

To obtain posterior class probabilities, you must set `'FitPosterior',true` when training the ECOC model using `fitcecoc`. Otherwise, `resubPredict` throws an error.

Examples

Predict Labels of Training Data Using ECOC Model

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y);
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

Predict the labels of the training data. Print a random subset of true and predicted labels.

```
labels = resubPredict(Mdl);

rng(1); % For reproducibility
n = numel(Y); % Sample size
idx = randsample(n,10);
table(Y(idx),labels(idx),'VariableNames',{'TrueLabels','PredictedLabels'})
```

```
ans=10×2 table
    TrueLabels    PredictedLabels
    __________    _______________

    setosa          setosa
    versicolor      versicolor
    virginica       virginica
    setosa          setosa
    versicolor      versicolor
    setosa          setosa
    versicolor      versicolor
    versicolor      versicolor
    setosa          setosa
    setosa          setosa
```

`Mdl` correctly labels the observations with indices `idx`.

Predict Resubstitution Labels of ECOC Model Using Custom Binary Loss Function

Load Fisher's iris data set. Specify the predictor data `X`, the response data `Y`, and the order of the classes in `Y`.

```
load fisheriris
X = meas;
Y = categorical(species);
classOrder = unique(Y); % Class order
```

Train an ECOC model using SVM binary classifiers. Standardize the predictors using an SVM template, and specify the class order.

```
t = templateSVM('Standardize',true);
Mdl = fitcecoc(X,Y,'Learners',t,'ClassNames',classOrder);
```

`t` is an SVM template object. During training, the software uses default values for empty properties in `t`. `Mdl` is a `ClassificationECOC` model.

SVM scores are signed distances from the observation to the decision boundary. Therefore, the domain is $(-\infty, \infty)$. Create a custom binary loss function that does the following:

- Map the coding design matrix (M) and positive-class classification scores (s) for each learner to the binary loss for each observation.

- Use linear loss.
- Aggregate the binary learner loss using the median.

You can create a separate function for the binary loss function, and then save it on the MATLAB® path. Or, you can specify an anonymous binary loss function. In this case, create a function handle (`customBL`) to an anonymous binary loss function.

```
customBL = @(M,s)nanmedian(1 - bsxfun(@times,M,s),2)/2;
```

Predict labels for the training data and estimate the median binary loss per class. Print the median negative binary losses per class for a random set of 10 observations.

```
[label,NegLoss] = resubPredict(Mdl,'BinaryLoss',customBL);

rng(1); % For reproducibility
n = numel(Y); % Sample size
idx = randsample(n,10);
classOrder
```

```
classOrder = 3x1 categorical
     setosa
     versicolor
     virginica
```

```
table(Y(idx),label(idx),NegLoss(idx,:),'VariableNames',...
    {'TrueLabel','PredictedLabel','NegLoss'})
```

```
ans=10×3 table
    TrueLabel     PredictedLabel                NegLoss
    __________    ______________    ___________________________________

    setosa          versicolor       0.1237       1.957     -3.5807
    versicolor      versicolor       -1.017     0.62917     -1.1122
    virginica       virginica       -1.9082    -0.21802     0.62618
    setosa          versicolor      0.43842      2.2443     -4.1827
    versicolor      versicolor      -1.0733     0.39627    -0.82294
    setosa          versicolor      0.26668      2.2004      -3.967
    versicolor      versicolor      -1.1234     0.69883     -1.0754
    versicolor      versicolor      -1.2709     0.51788    -0.74697
    setosa          versicolor      0.35181       2.068     -3.9198
    setosa          versicolor      0.23355      2.1886     -3.9221
```

The order of the columns corresponds to the elements of `classOrder`. The software predicts the label based on the maximum negated loss. The results indicate that the median of the linear losses might not perform as well as other losses.

Estimate Posterior Probabilities Using ECOC Classifier

Train an ECOC classifier using SVM binary learners. First predict the training-sample labels and class posterior probabilities. Then predict the maximum class posterior probability at each point in a grid. Visualize the results.

Load Fisher's iris data set. Specify the petal dimensions as the predictors and the species names as the response.

```
load fisheriris
X = meas(:,3:4);
Y = species;
rng(1); % For reproducibility
```

Create an SVM template. Standardize the predictors, and specify the Gaussian kernel.

```
t = templateSVM('Standardize',true,'KernelFunction','gaussian');
```

`t` is an SVM template. Most of its properties are empty. When the software trains the ECOC classifier, it sets the applicable properties to their default values.

Train the ECOC classifier using the SVM template. Transform classification scores to class posterior probabilities (which are returned by `predict` or `resubPredict`) using the `'FitPosterior'` name-value pair argument. Specify the class order using the `'ClassNames'` name-value pair argument. Display diagnostic messages during training by using the `'Verbose'` name-value pair argument.

```
Mdl = fitcecoc(X,Y,'Learners',t,'FitPosterior',true,...
    'ClassNames',{'setosa','versicolor','virginica'},...
    'Verbose',2);

Training binary learner 1 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 2
Positive class indices: 1

Fitting posterior probabilities for learner 1 (SVM).
Training binary learner 2 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 1

Fitting posterior probabilities for learner 2 (SVM).
Training binary learner 3 (SVM) out of 3 with 50 negative and 50 positive observations.
Negative class indices: 3
Positive class indices: 2

Fitting posterior probabilities for learner 3 (SVM).
```

`Mdl` is a `ClassificationECOC` model. The same SVM template applies to each binary learner, but you can adjust options for each binary learner by passing in a cell vector of templates.

Predict the training-sample labels and class posterior probabilities. Display diagnostic messages during the computation of labels and class posterior probabilities by using the `'Verbose'` name-value pair argument.

```
[label,~,~,Posterior] = resubPredict(Mdl,'Verbose',1);

Predictions from all learners have been computed.
Loss for all observations has been computed.
Computing posterior probabilities...
```

```
Mdl.BinaryLoss

ans =
'quadratic'
```

The software assigns an observation to the class that yields the smallest average binary loss. Because all binary learners are computing posterior probabilities, the binary loss function is `quadratic`.

Display a random set of results.

```
idx = randsample(size(X,1),10,1);
Mdl.ClassNames
```

```
ans = 3x1 cell
    {'setosa'    }
    {'versicolor'}
    {'virginica' }
```

```
table(Y(idx),label(idx),Posterior(idx,:),...
    'VariableNames',{'TrueLabel','PredLabel','Posterior'})
```

```
ans=10×3 table
      TrueLabel         PredLabel                    Posterior
    ______________    ______________    ____________________________________

    {'virginica' }    {'virginica' }     0.0039319     0.0039866       0.99208
    {'virginica' }    {'virginica' }      0.017066      0.018262       0.96467
    {'virginica' }    {'virginica' }      0.014947      0.015855        0.9692
    {'versicolor'}    {'versicolor'}    2.2197e-14       0.87318       0.12682
    {'setosa'    }    {'setosa'    }         0.999    0.00025091    0.00074639
    {'versicolor'}    {'virginica' }    2.2195e-14      0.059427       0.94057
    {'versicolor'}    {'versicolor'}    2.2194e-14       0.97002      0.029984
    {'setosa'    }    {'setosa'    }         0.999     0.0002499    0.00074741
    {'versicolor'}    {'versicolor'}     0.0085638       0.98259     0.0088482
    {'setosa'    }    {'setosa'    }         0.999    0.00025013    0.00074718
```

The columns of `Posterior` correspond to the class order of `Mdl.ClassNames`.

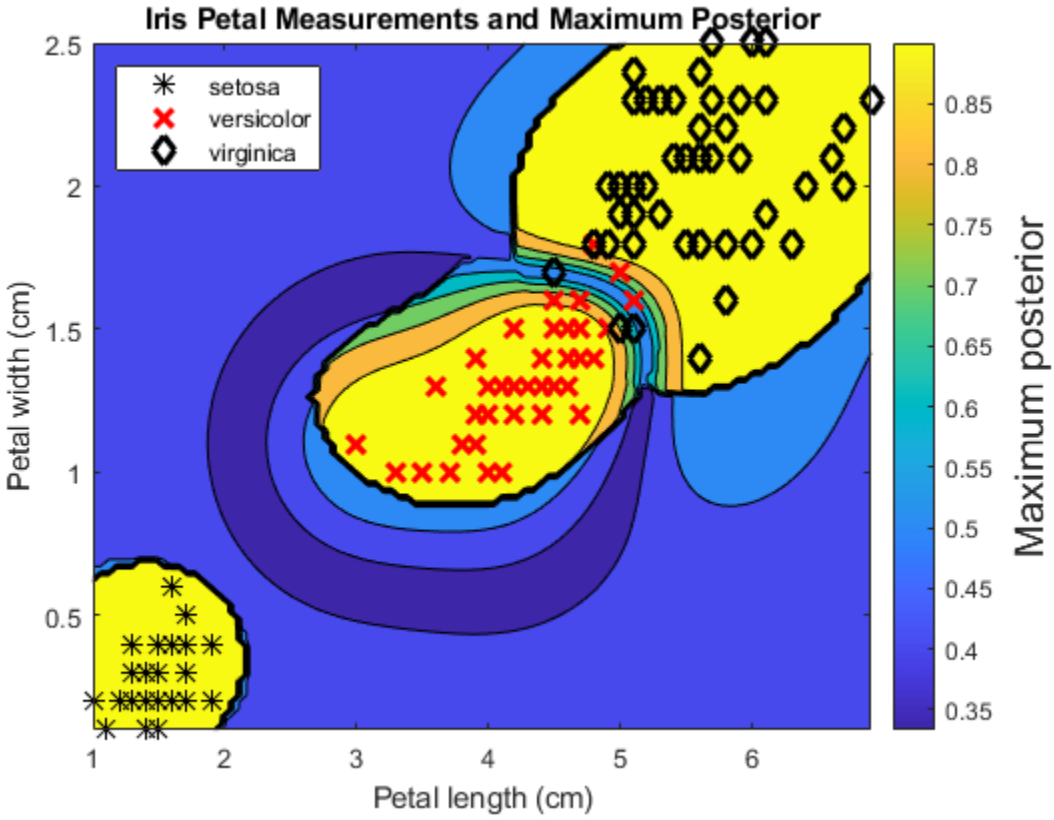
Define a grid of values in the observed predictor space. Predict the posterior probabilities for each instance in the grid.

```
xMax = max(X);
xMin = min(X);

x1Pts = linspace(xMin(1),xMax(1));
x2Pts = linspace(xMin(2),xMax(2));
[x1Grid,x2Grid] = meshgrid(x1Pts,x2Pts);

[~,~,~,PosteriorRegion] = predict(Mdl,[x1Grid(:),x2Grid(:)]);
```

For each coordinate on the grid, plot the maximum class posterior probability among all classes.

```
contourf(x1Grid,x2Grid,...
        reshape(max(PosteriorRegion,[],2),size(x1Grid,1),size(x1Grid,2)));
h = colorbar;
h.YLabel.String = 'Maximum posterior';
h.YLabel.FontSize = 15;

hold on
gh = gscatter(X(:,1),X(:,2),Y,'krk','*xd',8);
gh(2).LineWidth = 2;
gh(3).LineWidth = 2;
```

```
title('Iris Petal Measurements and Maximum Posterior')
xlabel('Petal length (cm)')
ylabel('Petal width (cm)')
axis tight
legend(gh,'Location','NorthWest')
hold off
```



Estimate Posterior Probabilities Using Parallel Computing

Train a multiclass ECOC model and estimate the posterior probabilities using parallel computing.

Load the `arrhythmia` data set. Examine the response data `Y`, and determine the number of classes.

```
load arrhythmia
Y = categorical(Y);
tabulate(Y)
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
```

```
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

```
K = numel(unique(Y));
```

Several classes are not represented in the data, and many other classes have low relative frequencies.

Specify an ensemble learning template that uses the GentleBoost method and 50 weak classification tree learners.

```
t = templateEnsemble('GentleBoost',50,'Tree');
```

`t` is a template object. Most of its properties are empty (`[]`). The software uses default values for all empty properties during training.

Because the response variable contains many classes, specify a sparse random coding design.

```
rng(1); % For reproducibility
Coding = designecoc(K,'sparserandom');
```

Train an ECOC model using parallel computing. Specify to fit posterior probabilities.

```
pool = parpool;                     % Invokes workers
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

```
options = statset('UseParallel',true);
Mdl = fitcecoc(X,Y,'Learner',t,'Options',options,'Coding',Coding,...
    'FitPosterior',true);
```

`Mdl` is a `ClassificationECOC` model. You can access its properties using dot notation.

The pool invokes six workers, although the number of workers might vary among systems.

Estimate posterior probabilities, and display the posterior probability of being classified as not having arrhythmia (class 1) given a random subset of the training data.

```
[~,~,~,posterior] = resubPredict(Mdl);

n = numel(Y);
idx = randsample(n,10,1);
table(idx,Y(idx),posterior(idx,1),...
    'VariableNames',{'ObservationIndex','TrueLabel','PosteriorNoArrythmia'})
```

```
ans=10×3 table
    ObservationIndex    TrueLabel    PosteriorNoArrythmia
    ________________    _________    ____________________

           79               1               0.93436
          248               1               0.95574
          398               10             0.032378
          207               1               0.97965
```

```
    340             1               0.93656
    206             1               0.97795
    345             10             0.015642
    296             2               0.13433
    391             1                0.9648
    406             1               0.94861
```

Input Arguments

`Mdl` — Full, trained multiclass ECOC model

`ClassificationECOC` model

Full, trained multiclass ECOC model, specified as a `ClassificationECOC` model trained with `fitcecoc`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resubPredict(Mdl,'BinaryLoss','linear','Decoding','lossbased')` specifies a linear binary learner loss function and a loss-based decoding scheme for aggregating the binary losses.

`BinaryLoss` — Binary learner loss function

`'hamming'` | `'linear'` | `'logit'` | `'exponential'` | `'binodeviance'` | `'hinge'` | `'quadratic'` | function handle

Binary learner loss function, specified as the comma-separated pair consisting of `'BinaryLoss'` and a built-in loss function name or function handle.

- This table describes the built-in functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$ is the binary loss formula.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

  The software normalizes binary losses so that the loss is 0.5 when $y_j = 0$. Also, the software calculates the mean binary loss for each class.

- For a custom binary loss function, for example `customFunction`, specify its function handle `'BinaryLoss',@customFunction`.

`customFunction` has this form:

```
bLoss = customFunction(M,s)
```

where:

- `M` is the *K*-by-*L* coding matrix stored in `Mdl.CodingMatrix`.
- `s` is the 1-by-*L* row vector of classification scores.
- `bLoss` is the classification loss. This scalar aggregates the binary losses for every learner in a particular class. For example, you can use the mean binary loss to aggregate the loss over the learners for each class.
- *K* is the number of classes.
- *L* is the number of binary learners.

For an example of passing a custom binary loss function, see “Predict Test-Sample Labels of ECOC Model Using Custom Binary Loss Function” on page 33-4811.

The default `BinaryLoss` value depends on the score ranges returned by the binary learners. This table describes some default `BinaryLoss` values based on the given assumptions.

Assumption	Default Value
All binary learners are SVMs or either linear or kernel classification models of SVM learners.	`'hinge'`
All binary learners are ensembles trained by `AdaboostM1` or `GentleBoost`.	`'exponential'`
All binary learners are ensembles trained by `LogitBoost`.	`'binodeviance'`
All binary learners are linear or kernel classification models of logistic regression learners. Or, you specify to predict class posterior probabilities by setting `'FitPosterior',true` in `fitcecoc`.	`'quadratic'`

To check the default value, use dot notation to display the `BinaryLoss` property of the trained model at the command line.

Example: `'BinaryLoss','binodeviance'`

Data Types: `char` | `string` | `function_handle`

`Decoding` — Decoding scheme
`'lossweighted'` (default) | `'lossbased'`

Decoding scheme that aggregates the binary losses, specified as the comma-separated pair consisting of `'Decoding'` and `'lossweighted'` or `'lossbased'`. For more information, see “Binary Loss” on page 33-5632.

Example: `'Decoding','lossbased'`

`NumKLInitializations` — Number of random initial values
`0` (default) | nonnegative integer scalar

Number of random initial values for fitting posterior probabilities by Kullback-Leibler divergence minimization, specified as the comma-separated pair consisting of `'NumKLInitializations'` and a nonnegative integer scalar.

If you do not request the fourth output argument (`Posterior`) and set `'PosteriorMethod','kl'` (the default), then the software ignores the value of `NumKLInitializations`.

For more details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-5633.

Example: `'NumKLInitializations',5`

Data Types: `single` | `double`

`Options` — Estimation options
`[]` (default) | structure array returned by `statset`

Estimation options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`.

To invoke parallel computing:

- You need a Parallel Computing Toolbox license.
- Specify `'Options',statset('UseParallel',true)`.

`PosteriorMethod` — Posterior probability estimation method
`'kl'` (default) | `'qp'`

Posterior probability estimation method, specified as the comma-separated pair consisting of `'PosteriorMethod'` and `'kl'` or `'qp'`.

- If `PosteriorMethod` is `'kl'`, then the software estimates multiclass posterior probabilities by minimizing the Kullback-Leibler divergence between the predicted and expected posterior probabilities returned by binary learners. For details, see “Posterior Estimation Using Kullback-Leibler Divergence” on page 33-5633.
- If `PosteriorMethod` is `'qp'`, then the software estimates multiclass posterior probabilities by solving a least-squares problem using quadratic programming. You need an Optimization Toolbox license to use this option. For details, see “Posterior Estimation Using Quadratic Programming” on page 33-5635.
- If you do not request the fourth output argument (`Posterior`), then the software ignores the value of `PosteriorMethod`.

Example: `'PosteriorMethod','qp'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0` or `1`. `Verbose` controls the number of diagnostic messages that the software displays in the Command Window.

If `Verbose` is `0`, then the software does not display diagnostic messages. Otherwise, the software displays diagnostic messages.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label` has the same data type as `Mdl.ClassNames` and has the same number of rows as `Mdl.X`.

The software predicts the classification of an observation by assigning the observation to the class yielding the largest negated average binary loss (or, equivalently, the smallest average binary loss).

`NegLoss` — Negated average binary losses
numeric matrix

Negated average binary losses on page 33-5632, returned as a numeric matrix. `NegLoss` is an *n*-by-*K* matrix, where *n* is the number of observations (`size(Mdl.X,1)`) and *K* is the number of unique classes (`size(Mdl.ClassNames,1)`).

`PBScore` — Positive-class scores
numeric matrix

Positive-class scores for each binary learner, returned as a numeric matrix. `PBScore` is an *n*-by-*L* matrix, where *n* is the number of observations (`size(Mdl.X,1)`) and *L* is the number of binary learners (`size(Mdl.CodingMatrix,2)`).

`Posterior` — Posterior class probabilities
numeric matrix

Posterior class probabilities, returned as a numeric matrix. `Posterior` is an *n*-by-*K* matrix, where *n* is the number of observations (`size(Mdl.X,1)`) and *K* is the number of unique classes (`size(Mdl.ClassNames,1)`).

To request `Posterior`, you must set `'FitPosterior',true` when training the ECOC model using `fitcecoc`. Otherwise, the software throws an error.

More About

Binary Loss

A binary loss is a function of the class and classification score that determines how well a binary learner classifies an observation into the class.

Suppose the following:

- m_{kj} is element (k,j) of the coding design matrix M (that is, the code corresponding to class k of binary learner j).
- s_j is the score of binary learner j for an observation.
- g is the binary loss function.
- $\widehat{k}$ is the predicted class for the observation.

In loss-based decoding [Escalera et al.] on page 18-279, the class producing the minimum sum of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\arg\min} \sum_{j=1}^{L} \left|m_{kj}\right| g(m_{kj}, s_j) .$$

In loss-weighted decoding [Escalera et al.] on page 18-279, the class producing the minimum average of the binary losses over binary learners determines the predicted class of an observation, that is,

$$\widehat{k} = \underset{k}{\operatorname{argmin}} \frac{\sum_{j=1}^{L} |m_{kj}| g(m_{kj}, s_j)}{\sum_{j=1}^{L} |m_{kj}|}.$$

Allwein et al. on page 18-279 suggest that loss-weighted decoding improves classification accuracy by keeping loss values for all classes in the same dynamic range.

This table summarizes the supported loss functions, where y_j is a class label for a particular binary learner (in the set {–1,1,0}), s_j is the score for observation j, and $g(y_j,s_j)$.

Value	Description	Score Domain	$g(y_j,s_j)$
`'binodeviance'`	Binomial deviance	(–∞,∞)	$\log[1 + \exp(-2y_js_j)]/[2\log(2)]$
`'exponential'`	Exponential	(–∞,∞)	$\exp(-y_js_j)/2$
`'hamming'`	Hamming	[0,1] or (–∞,∞)	$[1 - \text{sign}(y_js_j)]/2$
`'hinge'`	Hinge	(–∞,∞)	$\max(0,1 - y_js_j)/2$
`'linear'`	Linear	(–∞,∞)	$(1 - y_js_j)/2$
`'logit'`	Logistic	(–∞,∞)	$\log[1 + \exp(-y_js_j)]/[2\log(2)]$
`'quadratic'`	Quadratic	[0,1]	$[1 - y_j(2s_j - 1)]^2/2$

The software normalizes binary losses such that the loss is 0.5 when $y_j = 0$, and aggregates using the average of the binary learners [Allwein et al.] on page 18-279.

Do not confuse the binary loss with the overall classification loss (specified by the `'LossFun'` name-value pair argument of the `loss` and `predict` object functions), which measures how well an ECOC classifier performs as a whole.

Algorithms

The software can estimate class posterior probabilities by minimizing the Kullback-Leibler divergence or by using quadratic programming. For the following descriptions of the posterior estimation algorithms, assume that:

- m_{kj} is the element (k,j) of the coding design matrix M.
- I is the indicator function.
- $\widehat{p}_k$ is the class posterior probability estimate for class k of an observation, $k = 1,...,K$.
- r_j is the positive-class posterior probability for binary learner j. That is, r_j is the probability that binary learner j classifies an observation into the positive class, given the training data.

Posterior Estimation Using Kullback-Leibler Divergence

By default, the software minimizes the Kullback-Leibler divergence to estimate class posterior probabilities. The Kullback-Leibler divergence between the expected and observed positive-class posterior probabilities is

$$\Delta(r,\widehat{r}) = \sum_{j=1}^{L} w_j\left[r_j\log\frac{r_j}{\widehat{r}_j} + (1-r_j)\log\frac{1-r_j}{1-\widehat{r}_j}\right],$$

where $w_j = \sum_{S_j} w_i^*$ is the weight for binary learner j.

- S_j is the set of observation indices on which binary learner j is trained.
- w_i^* is the weight of observation i.

The software minimizes the divergence iteratively. The first step is to choose initial values $\widehat{p}_k^{(0)};\ k = 1,\ldots,K$ for the class posterior probabilities.

- If you do not specify `'NumKLIterations'`, then the software tries both sets of deterministic initial values described next, and selects the set that minimizes Δ.
  - $\widehat{p}_k^{(0)} = 1/K;\ k = 1,\ldots,K$.
  - $\widehat{p}_k^{(0)};\ k = 1,\ldots,K$ is the solution of the system

    $$M_{01}\widehat{p}^{(0)} = r,$$

    where M_{01} is M with all $m_{kj} = -1$ replaced with 0, and r is a vector of positive-class posterior probabilities returned by the L binary learners [Dietterich et al.] on page 18-279. The software uses `lsqnonneg` to solve the system.
- If you specify `'NumKLIterations',c`, where `c` is a natural number, then the software does the following to choose the set $\widehat{p}_k^{(0)};\ k = 1,\ldots,K$, and selects the set that minimizes Δ.
  - The software tries both sets of deterministic initial values as described previously.
  - The software randomly generates `c` vectors of length K using `rand`, and then normalizes each vector to sum to 1.

At iteration t, the software completes these steps:

1 Compute

$$\widehat{r}_j^{(t)} = \frac{\sum_{k=1}^{K}\widehat{p}_k^{(t)}I(m_{kj} = +1)}{\sum_{k=1}^{K}\widehat{p}_k^{(t)}I(m_{kj} = +1 \cup m_{kj} = -1)}.$$

2 Estimate the next class posterior probability using

$$\widehat{p}_k^{(t+1)} = \widehat{p}_k^{(t)}\frac{\sum_{j=1}^{L} w_j\left[r_j I\left(m_{kj} = +1\right) + \left(1-r_j\right)I\left(m_{kj} = -1\right)\right]}{\sum_{j=1}^{L} w_j\left[\widehat{r}_j^{(t)}I\left(m_{kj} = +1\right) + \left(1-\widehat{r}_j^{(t)}\right)I\left(m_{kj} = -1\right)\right]}.$$

3 Normalize $\widehat{p}_k^{(t+1)};\ k = 1,\ldots,K$ so that they sum to 1.

4 Check for convergence.

For more details, see [Hastie et al.] on page 18-280 and [Zadrozny] on page 18-281.

Posterior Estimation Using Quadratic Programming

Posterior probability estimation using quadratic programming requires an Optimization Toolbox license. To estimate posterior probabilities for an observation using this method, the software completes these steps:

1. Estimate the positive-class posterior probabilities, r_j, for binary learners $j = 1,...,L$.
2. Using the relationship between r_j and $\widehat{p}_k$ [Wu et al.] on page 18-281, minimize

$$\sum_{j=1}^{L}\left[-r_j\sum_{k=1}^{K}\widehat{p}_k I\left(m_{kj}=-1\right)+\left(1-r_j\right)\sum_{k=1}^{K}\widehat{p}_k I\left(m_{kj}=+1\right)\right]^2$$

with respect to $\widehat{p}_k$ and the restrictions

$$0\le\widehat{p}_k\le 1$$

$$\sum_k\widehat{p}_k=1.$$

The software performs minimization using `quadprog`.

References

[1] Allwein, E., R. Schapire, and Y. Singer. “Reducing multiclass to binary: A unifying approach for margin classifiers.” *Journal of Machine Learning Research*. Vol. 1, 2000, pp. 113–141.

[2] Dietterich, T., and G. Bakiri. “Solving Multiclass Learning Problems Via Error-Correcting Output Codes.” *Journal of Artificial Intelligence Research*. Vol. 2, 1995, pp. 263–286.

[3] Escalera, S., O. Pujol, and P. Radeva. “On the decoding process in ternary error-correcting output codes.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 32, Issue 7, 2010, pp. 120–134.

[4] Escalera, S., O. Pujol, and P. Radeva. “Separability of ternary codes for sparse designs of error-correcting output codes.” *Pattern Recogn*. Vol. 30, Issue 3, 2009, pp. 285–297.

[5] Hastie, T., and R. Tibshirani. “Classification by Pairwise Coupling.” *Annals of Statistics*. Vol. 26, Issue 2, 1998, pp. 451–471.

[6] Wu, T. F., C. J. Lin, and R. Weng. “Probability Estimates for Multi-Class Classification by Pairwise Coupling.” *Journal of Machine Learning Research*. Vol. 5, 2004, pp. 975–1005.

[7] Zadrozny, B. “Reducing Multiclass to Binary by Coupling Probability Estimates.” *NIPS 2001: Proceedings of Advances in Neural Information Processing Systems 14*, 2001, pp. 1041–1048.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationECOC` | `fitcecoc` | `predict` | `quadprog` | `resubLoss` | `statset`

Topics

“Quick Start Parallel Computing for Statistics and Machine Learning Toolbox” on page 31-2
“Reproducibility in Parallel Statistical Computations” on page 31-13
“Concepts of Parallel Computing in Statistics and Machine Learning Toolbox” on page 31-8

Introduced in R2014b

resubPredict

Classify observations in ensemble of classification models

Syntax

```
label = resubPredict(ens)
[label,score] = resubPredict(ens)
[label,score] = resubPredict(ens,Name,Value)
```

Description

`label = resubPredict(ens)` returns the labels `ens` predicts for the data `ens.X`. `label` is the predictions of `ens` on the data that `fitcensemble` used to create `ens`.

`[label,score] = resubPredict(ens)` also returns scores for all classes.

`[label,score] = resubPredict(ens,Name,Value)` finds resubstitution predictions with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A classification ensemble created with `fitcensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `oobLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

Output Arguments

`label`

The response `ens` predicts for the training data. `label` is the same data type as the training response data `ens.Y`, and has the same number of entries as the number of rows in `ens.X`.

`score`

An `N`-by-`K` matrix, where `N` is the number of rows in `ens.X`, and `K` is the number of classes in `ens`. High score value indicates that an observation likely comes from this class.

Examples

Find Number of Ensemble Misclassifications

Find the total number of misclassifications of the `fisheriris` data for a classification ensemble.

Load the Fisher iris data set.

```
load fisheriris
```

Train an ensemble of 100 boosted classification trees using AdaBoostM2.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
ens = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

Find the total number of misclassifications.

```
Ypredict = resubPredict(ens); % The predictions
Ysame = strcmp(Ypredict,species); % True when Ypredict and species are equal
sum(~Ysame) % Number of different predictions
```

```
ans = 5
```

More About

Score (ensemble)

For ensembles, a classification score represents the confidence of a classification into a class. The higher the score, the higher the confidence.

Different ensemble algorithms have different definitions for their scores. Furthermore, the range of scores depends on ensemble type. For example:

- `AdaBoostM1` scores range from –∞ to ∞.
- `Bag` scores range from `0` to `1`.

See Also

`resubEdge` | `resubLoss` | `resubMargin` | `resubPredict`

resubPredict

Classify training data using trained classifier

Syntax

```
label = resubPredict(Mdl)
label = resubPredict(Mdl,'IncludeInteractions',includeInteractions)
[label,Score] = resubPredict( ___ )
[label,Score,Cost] = resubPredict(Mdl)
```

Description

`label = resubPredict(Mdl)` returns a vector of predicted class labels (`label`) for the trained classification model `Mdl` using the predictor data stored in `Mdl.X`.

`label = resubPredict(Mdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

`[label,Score] = resubPredict( ___ )` also returns classification scores using any of the input argument combinations in the previous syntaxes.

`[label,Score,Cost] = resubPredict(Mdl)` also returns the expected misclassification cost. This syntax applies only to *k*-nearest neighbor and naive Bayes models.

Examples

Label Training Sample Observations of Naive Bayes Classifier

Load the `fisheriris` data set. Create `X` as a numeric matrix that contains four measurements for 150 irises. Create `Y` as a cell array of character vectors that contains the corresponding iris species.

```
load fisheriris
X = meas;
Y = species;
rng('default') % For reproducibility
```

Train a naive Bayes classifier using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. `fitcnb` assumes that each predictor is conditionally and normally distributed.

```
Mdl = fitcnb(X,Y,'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationNaiveBayes
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
```

```
         DistributionNames: {'normal'  'normal'  'normal'  'normal'}
    DistributionParameters: {3x4 cell}


  Properties, Methods
```

`Mdl` is a trained `ClassificationNaiveBayes` classifier.

Predict the training sample labels.

```
label = resubPredict(Mdl);
```

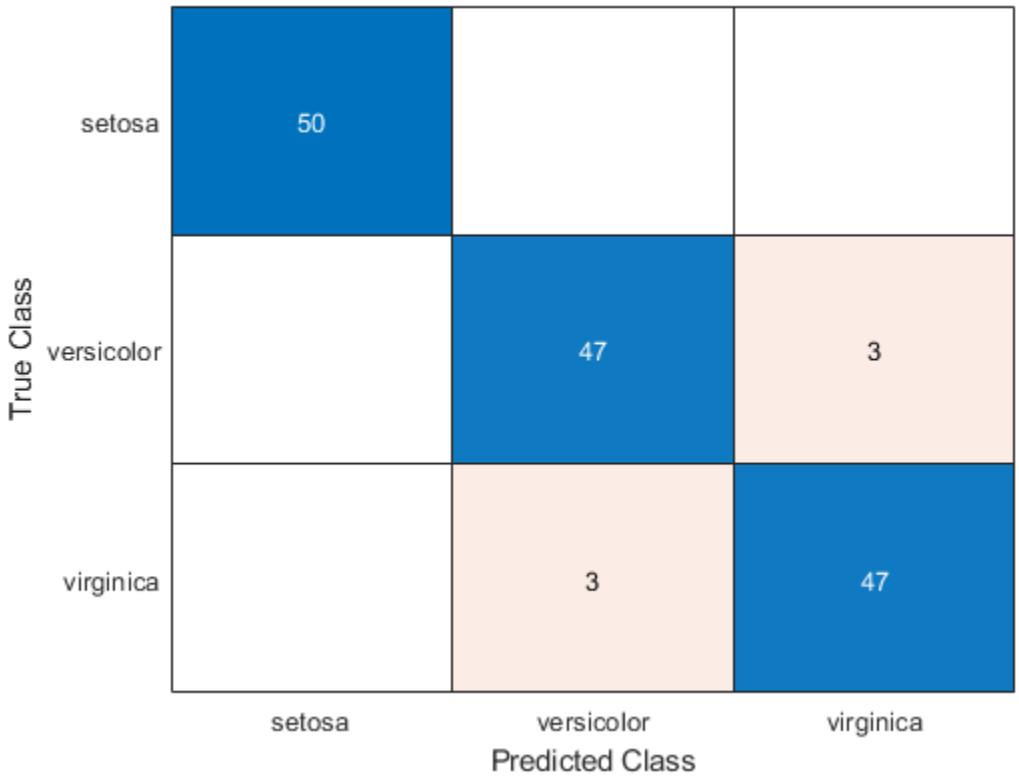
Display the results for a random set of 10 observations.

```
idx = randsample(size(X,1),10);
table(Y(idx),label(idx),'VariableNames', ...
    {'True Label','Predicted Label'})
```

```
ans=10×2 table
      True Label      Predicted Label
    _____________    _______________

    {'virginica' }    {'virginica' }
    {'setosa'    }    {'setosa'    }
    {'virginica' }    {'virginica' }
    {'versicolor'}    {'versicolor'}
    {'virginica' }    {'virginica' }
    {'versicolor'}    {'versicolor'}
    {'virginica' }    {'virginica' }
    {'setosa'    }    {'setosa'    }
    {'virginica' }    {'virginica' }
    {'setosa'    }    {'setosa'    }
```

Create a confusion chart from the true labels `Y` and the predicted labels `label`.

```
cm = confusionchart(Y,label);
```



Estimate In-Sample Posterior Probabilities of SVM Classifier

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a support vector machine (SVM) classifier. Standardize the data and specify that `'g'` is the positive class.

```
SVMModel = fitcsvm(X,Y,'ClassNames',{'b','g'},'Standardize',true);
```

`SVMModel` is a `ClassificationSVM` classifier.

Fit the optimal score-to-posterior-probability transformation function.

```
rng(1); % For reproducibility
ScoreSVMModel = fitPosterior(SVMModel)
```

```
ScoreSVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
          ScoreTransform: '@(S)sigmoid(S,-9.481840e-01,-1.218721e-01)'
```

```
          NumObservations: 351
                    Alpha: [90x1 double]
                     Bias: -0.1343
        KernelParameters: [1x1 struct]
                      Mu: [1x34 double]
                    Sigma: [1x34 double]
           BoxConstraints: [351x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [351x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

Because the classes are inseparable, the score transformation function (`ScoreSVMModel.ScoreTransform`) is the sigmoid function.

Estimate scores and positive class posterior probabilities for the training data. Display the results for the first 10 observations.

```
[label,scores] = resubPredict(SVMModel);
[~,postProbs] = resubPredict(ScoreSVMModel);
table(Y(1:10),label(1:10),scores(1:10,2),postProbs(1:10,2),'VariableNames',...
    {'TrueLabel','PredictedLabel','Score','PosteriorProbability'})
```

```
ans=10×4 table
    TrueLabel    PredictedLabel     Score     PosteriorProbability
    _________    ______________    _______    ____________________

      {'g'}          {'g'}          1.4861           0.82215
      {'b'}          {'b'}         -1.0002           0.30439
      {'g'}          {'g'}          1.8686           0.86917
      {'b'}          {'b'}         -2.6456          0.084197
      {'g'}          {'g'}          1.2806           0.79185
      {'b'}          {'b'}         -1.4617           0.22026
      {'g'}          {'g'}          2.1671           0.89814
      {'b'}          {'b'}         -5.7089         0.0050106
      {'g'}          {'g'}          2.4796           0.92223
      {'b'}          {'b'}         -2.7812          0.074801
```

Compare GAMs by Examining Logit of Posterior Probabilities

Estimate the logit of posterior probabilities (classification scores) for training data using a classification generalized additive model (GAM) that contains both linear and interaction terms for predictors. Specify whether to include interaction terms when computing the classification scores.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a GAM using the predictors `X` and class labels `Y`. A recommended practice is to specify the class names. Specify to include the 10 most important interaction terms.

```
Mdl = fitcgam(X,Y,'ClassNames',{'b','g'},'Interactions',10)

Mdl =
  ClassificationGAM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'logit'
                Intercept: 3.2565
             Interactions: [10x2 double]
          NumObservations: 351


  Properties, Methods
```

`Mdl` is a `ClassificationGAM` model object.

Predict the labels using both linear and interaction terms, and then using only linear terms. To exclude interaction terms, specify `'IncludeInteractions',false`. Estimate the logit of posterior probabilities by specifying the `ScoreTransform` property as `'none'`.

```
Mdl.ScoreTransform = 'none';
[labels,scores] = resubPredict(Mdl);
[labels_nointeraction,scores_nointeraction] = resubPredict(Mdl,'IncludeInteractions',false);
```

Create a table containing the true labels, predicted labels, and scores. Display the first eight rows of the table.

```
t = table(Y,labels,scores,labels_nointeraction,scores_nointeraction, ...
    'VariableNames',{'True Labels','Predicted Labels','Scores' ...
    'Predicted Labels Without Interactions','Scores Without Interactions'});
head(t)
```

```
ans=8×5 table
    True Labels    Predicted Labels          Scores          Predicted Labels Without Interactio
    ___________    ________________    ___________________    ___________________________________

       {'g'}           {'g'}          -51.628     51.628                  {'g'}
       {'b'}           {'b'}           37.433    -37.433                  {'b'}
       {'g'}           {'g'}          -62.061     62.061                  {'g'}
       {'b'}           {'b'}           37.666    -37.666                  {'b'}
       {'g'}           {'g'}          -47.361     47.361                  {'g'}
       {'b'}           {'b'}           106.48    -106.48                  {'b'}
       {'g'}           {'g'}          -62.665     62.665                  {'g'}
       {'b'}           {'b'}           201.46    -201.46                  {'b'}
```

The predicted labels for the training data `X` do not vary depending on the inclusion of interaction terms, but the estimated score values are different.

Input Arguments

`Mdl` — Classification machine learning model

full classification model object

Classification machine learning model, specified as a full classification model object, as given in the following table of supported models.

Model	Classification Model Object
Generalized additive model	`ClassificationGAM`
k-nearest neighbor model	`ClassificationKNN`
Naive Bayes model	`ClassificationNaiveBayes`
Neural network model	`ClassificationNeuralNetwork`
Support vector machine for one-class and binary classification	`ClassificationSVM`

`includeInteractions` — Flag to include interaction terms
`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `ClassificationGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

Output Arguments

`label` — Predicted class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric vector, or cell array of character vectors.

`label` has the same data type as the observed class labels that trained `Mdl`, and its length is equal to the number of observations in `Mdl.X`. (The software treats string arrays as cell arrays of character vectors.)

`Score` — Class scores
numeric matrix

Class scores, returned as a numeric matrix. `Score` has rows equal to the number of observations in `Mdl.X` and columns equal to the number of distinct classes in the training data (`size(Mdl.ClassNames,1)`).

`Cost` — Expected misclassification costs
numeric matrix

Expected misclassification costs, returned as a numeric matrix. This output applies only to *k*-nearest neighbor and naive Bayes models. That is, `resubPredict` returns `Cost` only when `Mdl` is `ClassificationKNN` or `ClassificationNaiveBayes`.

`Cost` has rows equal to the number of observations in `Mdl.X` and columns equal to the number of distinct classes in the training data (`size(Mdl.ClassNames,1)`).

`Cost(j,k)` is the expected misclassification cost of the observation in row `j` of `Mdl.X` predicted into class `k` (in class `Mdl.ClassNames(k)`).

Algorithms

`resubPredict` computes predictions according to the corresponding `predict` function of the object (`Mdl`). For a model-specific description, see the `predict` function reference pages in the following table.

Model	Classification Model Object (`Mdl`)	`predict` Object Function
Generalized additive model	`ClassificationGAM`	`predict`
k-nearest neighbor model	`ClassificationKNN`	`predict`
Naive Bayes model	`ClassificationNaiveBayes`	`predict`
Neural network model	`ClassificationNeuralNetwork`	`predict`
Support vector machine for one-class and binary classification	`ClassificationSVM`	`predict`

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports `ClassificationKNN` and `ClassificationSVM` objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`resubEdge` | `resubLoss` | `resubMargin`

Introduced in R2012a

resubPredict

Class: `ClassificationTree`

Predict resubstitution labels of classification tree

Syntax

```
label = resubPredict(tree)
[label,posterior] = resubPredict(tree)
[label,posterior,node] = resubPredict(tree)
[label,posterior,node,cnum] = resubPredict(tree)
[label,...] = resubPredict(tree,Name,Value)
```

Description

`label = resubPredict(tree)` returns the labels `tree` predicts for the data `tree.X`. `label` is the predictions of `tree` on the data that `fitctree` used to create `tree`.

`[label,posterior] = resubPredict(tree)` returns the posterior class probabilities for the predictions.

`[label,posterior,node] = resubPredict(tree)` returns the node numbers of `tree` for the resubstituted data.

`[label,posterior,node,cnum] = resubPredict(tree)` returns the predicted class numbers for the predictions.

`[label,...] = resubPredict(tree,Name,Value)` returns resubstitution predictions with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`tree`

A classification tree constructed by `fitctree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level

0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `resubPredict` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`resubPredict` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`label`

The response `tree` predicts for the training data. `label` is the same data type as the training response data `tree.Y`.

If the `Subtrees` name-value argument contains `m>1` entries, `label` has `m` columns, each of which represents the predictions of the corresponding subtree. Otherwise, `label` is a vector.

`posterior`

Matrix or array of posterior probabilities for classes `tree` predicts.

If the `Subtrees` name-value argument is a scalar or is missing, `posterior` is an `n`-by-`k` matrix, where `n` is the number of rows in the training data `tree.X`, and `k` is the number of classes.

If `Subtrees` contains `m>1` entries, `posterior` is an `n`-by-`k`-by-`m` array, where the matrix for each `m` gives posterior probabilities for the corresponding subtree.

`node`

The node numbers of `tree` where each data row resolves.

If the `Subtrees` name-value argument is a scalar or is missing, `node` is a numeric column vector with `n` rows, the same number of rows as `tree.X`.

If `Subtrees` contains `m>1` entries, `node` is a `n`-by-`m` matrix. Each column represents the node predictions of the corresponding subtree.

`cnum`

The class numbers that `tree` predicts for the resubstituted data.

If the `Subtrees` name-value argument is a scalar or is missing, `cnum` is a numeric column vector with `n` rows, the same number of rows as `tree.X`.

If `Subtrees` contains `m>1` entries, `cnum` is a `n`-by-`m` matrix. Each column represents the class predictions of the corresponding subtree.

Examples

Compute Number of Misclassified Observations

Find the total number of misclassifications of the Fisher iris data for a classification tree.

```
load fisheriris
tree = fitctree(meas,species);
Ypredict = resubPredict(tree);    % The predictions
Ysame = strcmp(Ypredict,species); % True when ==
sum(~Ysame) % How many are different?
```

```
ans = 3
```

Compare In-Sample Posterior Probabilities for Each Subtree

Load Fisher's iris data set. Partition the data into training (50%)

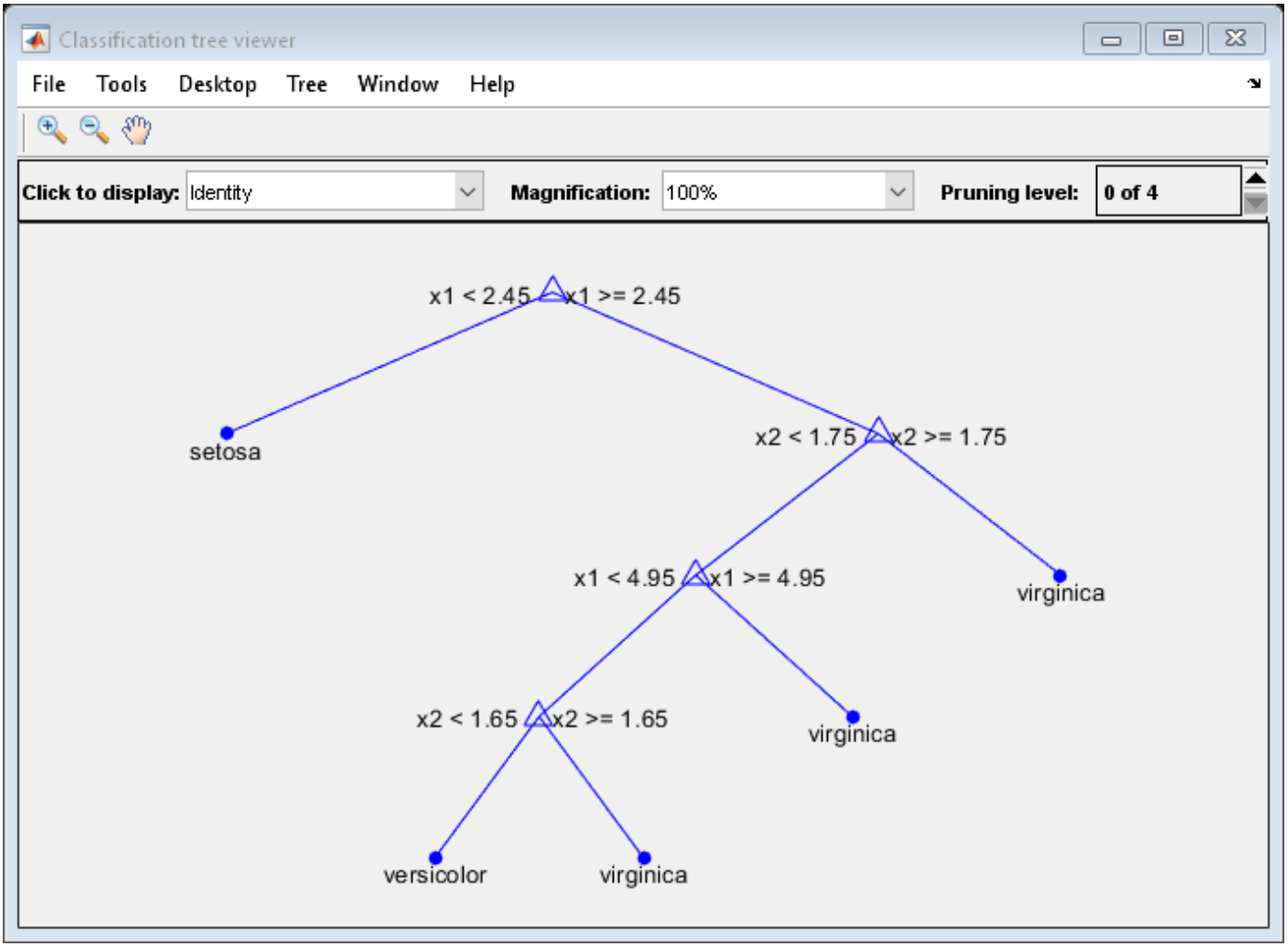
```
load fisheriris
```

Grow a classification tree using the all petal measurements.

```
Mdl = fitctree(meas(:,3:4),species);
n = size(meas,1); % Sample size
K = numel(Mdl.ClassNames); % Number of classes
```

View the classification tree.

```
view(Mdl,'Mode','graph');
```



The classification tree has four pruning levels. Level 0 is the full, unpruned tree (as displayed). Level 4 is just the root node (i.e., no splits).

Estimate the posterior probabilities for each class using the subtrees pruned to levels 1 and 3.

```
[~,Posterior] = resubPredict(Mdl,'SubTrees',[1 3]);
```

`Posterior` is an n-by- K-by- 2 array of posterior probabilities. Rows of `Posterior` correspond to observations, columns correspond to the classes with order `Mdl.ClassNames`, and pages correspond to pruning level.

Display the class posterior probabilities for iris 125 using each subtree.

```
Posterior(125,:,:)
```

```
ans =
ans(:,:,1) =

         0    0.0217    0.9783


ans(:,:,2) =
```

```
         0    0.5000    0.5000
```

The decision stump (page 2 of `Posterior`) has trouble predicting whether iris 125 is versicolor or virginica.

More About

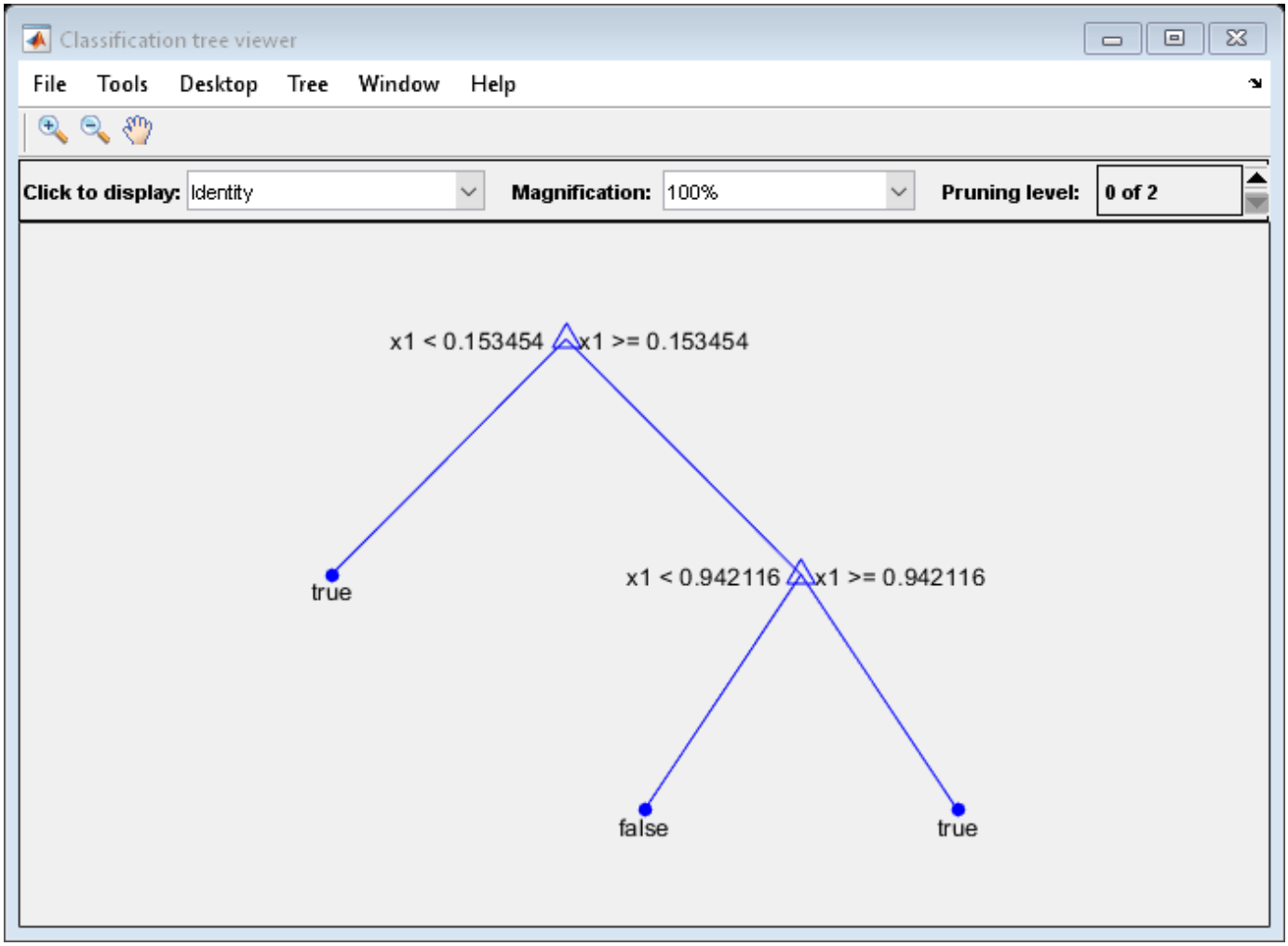
Posterior Probability

The posterior probability of the classification at a node is the number of training sequences that lead to that node with this classification, divided by the number of training sequences that lead to that node.

For example, consider classifying a predictor `X` as `true` when `X<0.15` or `X>0.95`, and `X` is false otherwise.

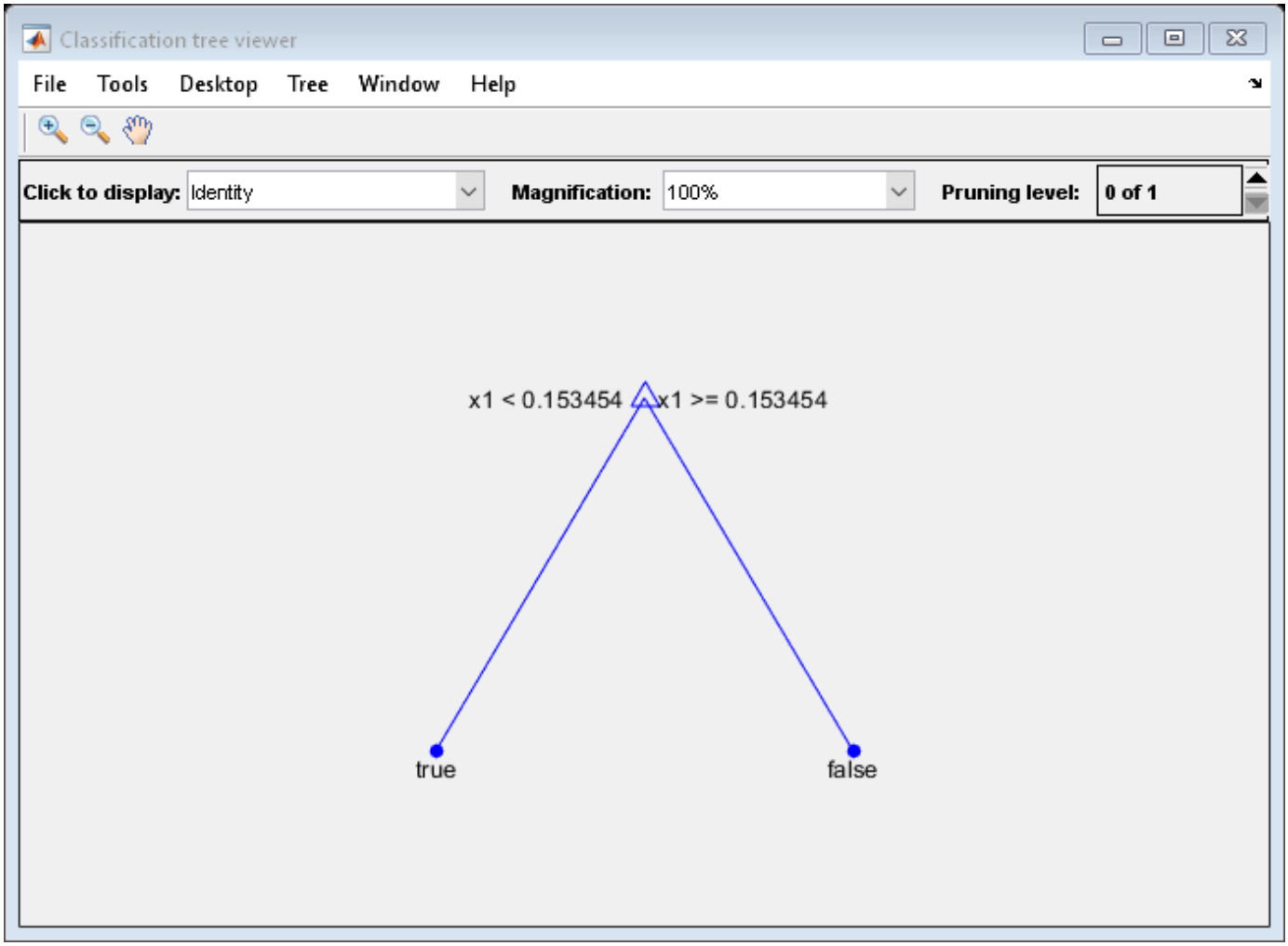
1 Generate 100 random points and classify them:

```
rng(0) % For reproducibility
X = rand(100,1);
Y = (abs(X - .55) > .4);
tree = fitctree(X,Y);
view(tree,'Mode','graph')
```



2 Prune the tree:

```
tree1 = prune(tree,'Level',1);
view(tree1,'Mode','graph')
```



The pruned tree correctly classifies observations that are less than 0.15 as `true`. It also correctly classifies observations between .15 and .94 as `false`. However, it incorrectly classifies observations that are greater than .94 as `false`. Therefore the score for observations that are greater than .15 should be about .05/.85=.06 for `true`, and about .8/.85=.94 for `false`.

3 Compute the prediction scores for the first 10 rows of `X`:

```
[~,score] = predict(tree1,X(1:10));
[score X(1:10,:)]
```

```
ans = 10×3

    0.9059    0.0941    0.8147
    0.9059    0.0941    0.9058
         0    1.0000    0.1270
    0.9059    0.0941    0.9134
    0.9059    0.0941    0.6324
         0    1.0000    0.0975
    0.9059    0.0941    0.2785
    0.9059    0.0941    0.5469
    0.9059    0.0941    0.9575
```

```
    0.9059    0.0941    0.9649
```

Indeed, every value of `X` (the rightmost column) that is less than 0.15 has associated scores (the left and center columns) of `0` and `1`, while the other values of `X` have associated scores of `0.94` and `0.06`.

See Also

`fitctree` | `predict` | `resubEdge` | `resubLoss` | `resubMargin`

resubPredict

Predict response of ensemble by resubstitution

Syntax

```
Yfit = resubPredict(ens)
Yfit = resubPredict(ens,Name,Value)
```

Description

`Yfit = resubPredict(ens)` returns the response `ens` predicts for the data `ens.X`. `Yfit` is the predictions of `ens` on the data that `fitrensemble` used to create `ens`.

`Yfit = resubPredict(ens,Name,Value)` predicts responses with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`learners`

Indices of weak learners in the ensemble ranging from `1` to `NumTrained`. `oobLoss` uses only these learners for calculating loss.

Default: `1:NumTrained`

Output Arguments

`Yfit`

A vector of predicted responses to the training data, with `ens.X` elements.

Examples

Find Mean-Squared Error of Resubstitution Predictions

Find the resubstitution predictions of mileage from the `carsmall` data, and look at their mean-squared difference from the training data.

Load the `carsmall` data set and select horsepower and vehicle weight as predictors.

```
load carsmall
X = [Horsepower Weight];
```

Train an ensemble of regression trees.

```
ens = fitrensemble(X,MPG,'Method','LSBoost','Learners','Tree');
```

Find the resubstitution predictions of `MPG`.

```
Yfit = resubPredict(ens);
```

Calculate the mean-squared difference of the resubstitution predictions from the training data.

```
MSE = mean((Yfit - ens.Y).^2)
```

```
MSE = 0.5836
```

Confirm that the result is the same as the result of `resubLoss`.

```
resubLoss(ens)
```

```
ans = 0.5836
```

See Also

`predict` | `resubLoss` | `resubPredict`

resubPredict

Predict responses for training data using trained regression model

Syntax

```
yFit = resubPredict(Mdl)
yFit = resubPredict(Mdl,'IncludeInteractions',includeInteractions)
```

Description

`yFit = resubPredict(Mdl)` returns a vector of predicted responses for the trained regression model `Mdl` using the predictor data stored in `Mdl.X`.

`yFit = resubPredict(Mdl,'IncludeInteractions',includeInteractions)` specifies whether to include interaction terms in computations. This syntax applies only to generalized additive models.

Examples

Resubstitution Predictions

Train a generalized additive model (GAM), then predict responses for the training data.

Load the `patients` data set.

```
load patients
```

Create a table that contains the predictor variables (`Age`, `Diastolic`, `Smoker`, `Weight`, `Gender`, `SelfAssessedHealthStatus`) and the response variable (`Systolic`).

```
tbl = table(Age,Diastolic,Smoker,Weight,Gender,SelfAssessedHealthStatus,Systolic);
```

Train a univariate GAM that contains the linear terms for the predictors in `tbl`.

```
Mdl = fitrgam(tbl,"Systolic")
```

```
Mdl =
  RegressionGAM
             PredictorNames: {1x6 cell}
               ResponseName: 'Systolic'
    CategoricalPredictors: [3 5 6]
        ResponseTransform: 'none'
                Intercept: 122.7800
          NumObservations: 100


  Properties, Methods
```

`Mdl` is a `RegressionGAM` model object.

Predict responses for the training set.

```
yFit = resubPredict(Mdl);
```

Create a table containing the observed response values and the predicted response values. Display the first eight rows of the table.

```
t = table(tbl.Systolic,yFit, ...
    'VariableNames',{'Observed Value','Predicted Value'});
head(t)
```

```
ans=8×2 table
    Observed Value    Predicted Value
    ______________    _______________

         124               124.75
         109               109.48
         125               122.89
         117               115.87
         122               121.61
         121               122.02
         130               126.39
         115               115.95
```

Compare GAMs by Examining Resubstitution Predictions

Predict responses for a training data set using a generalized additive model (GAM) that contains both linear and interaction terms for predictors. Specify whether to include interaction terms when predicting responses.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a generalized additive model that contains all the available linear and interaction terms in `X`.

```
Mdl = fitrgam(X,Y,'Interactions','all');
```

`Mdl` is a `RegressionGAM` model object.

Predict the responses using both linear and interaction terms, and then using only linear terms. To exclude interaction terms, specify `'IncludeInteractions',false`.

```
yFit = resubPredict(Mdl);
yFit_nointeraction = resubPredict(Mdl,'IncludeInteractions',false);
```

Create a table containing the observed response values and the predicted response values. Display the first eight rows of the table.

```
t = table(Mdl.Y,yFit,yFit_nointeraction, ...
    'VariableNames',{'Observed Response', ...
    'Predicted Response','Predicted Response Without Interactions'});
head(t)
```

```
ans=8×3 table
    Observed Response    Predicted Response    Predicted Response Without Interactions
    _________________    __________________    _______________________________________

           18                 18.026                          17.22
           15                 15.003                         15.791
           18                 17.663                          16.18
           16                 16.178                         15.536
           17                 17.107                         17.361
           15                 14.943                         14.424
           14                 14.119                         14.981
           14                 13.864                         13.498
```

Input Arguments

`Mdl` — Regression machine learning model

full regression model object

Regression machine learning model, specified as a full regression model object, as given in the following table of supported models.

Model	Regression Model Object
Generalized additive model	`RegressionGAM`
Neural network model	`RegressionNeuralNetwork`

`includeInteractions` — Flag to include interaction terms

`true` | `false`

Flag to include interaction terms of the model, specified as `true` or `false`. This argument is valid only for a generalized additive model (GAM). That is, you can specify this argument only when `Mdl` is `RegressionGAM`.

The default value is `true` if `Mdl` contains interaction terms. The value must be `false` if the model does not contain interaction terms.

Data Types: `logical`

Output Arguments

`yFit` — Predicted responses

vector

Predicted responses, returned as a vector of length *n*, where *n* is the number of observations in the predictor data (`Mdl.X`).

Algorithms

`resubPredict` predicts responses according to the corresponding `predict` function of the object (`Mdl`). For a model-specific description, see the `predict` function reference pages in the following table.

Model	Regression Model Object (`Mdl`)	`predict` Object Function
Generalized additive model	`RegressionGAM`	`predict`
Neural network model	`RegressionNeuralNetwork`	`predict`

See Also

`resubLoss`

Introduced in R2021a

resubPredict

Class: RegressionGP

Resubstitution prediction from a trained Gaussian process regression model

Syntax

```
ypred = resubPredict(gprMdl)
[ypred,ysd] = resubPredict(gprMdl)
[ypred,ysd,yint] = resubPredict(gprMdl)
[ypred,ysd,yint] = resubPredict(gprMdl,Name,Value)
```

Description

`ypred = resubPredict(gprMdl)` returns the predicted responses, `ypred`, for the trained Gaussian process regression (GPR) model, `gprMdl`.

`[ypred,ysd] = resubPredict(gprMdl)` also returns the estimated standard deviations of the predicted responses corresponding to the rows of `gprMdl.X`.

`[ypred,ysd,yint] = resubPredict(gprMdl)` also returns the 95% prediction intervals, `yint`, for the true responses corresponding to each row of training data, `gprMdl.X`.

`[ypred,ysd,yint] = resubPredict(gprMdl,Name,Value)` returns the prediction intervals with additional options, specified by one or more `Name,Value` pair arguments. For example, you can specify the confidence level of the prediction interval.

Input Arguments

`gprMdl` — Gaussian process regression model

`RegressionGP` object

Gaussian process regression model, specified as a `RegressionGP` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Alpha` — Significance level

0.05 (default) | scalar value in the range from 0 to 1

Significance level for the prediction intervals, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range from 0 to 1.

Example: `'Alpha',0.01` specifies 99% prediction intervals.

Data Types: `single` | `double`

Output Arguments

`ypred` — Predicted response values

n-by-1 vector

Predicted response values, returned as an *n*-by-1 vector, where *n* is the number of observations in the training data.

`ysd` — Standard deviation of the predicted response values

n-by-1 vector

Standard deviation of the predicted response values corresponding to the rows of `gprMdl.X`, returned as an *n*-by-1 vector. `ysd`(*i*), *i* = 1, 2, ..., *n*, contains the estimated standard deviation of the new response corresponding to the predictor values at the i^{th} observation in the training data.

`yint` — Prediction intervals for the true response values

n-by-2 matrix

Prediction intervals for the true response values corresponding to the rows of `gprMdl.X`, returned as an *n*-by-2 matrix, where *n* is the number of observations in the training data. The first column of `yint` contains the lower limits and the second column contains the upper limits of the prediction intervals.

Examples

Plot Predicted Response and Prediction Intervals

This example uses "Housing" data set [1] from the UCI machine learning archive [2] described in http://archive.ics.uci.edu/ml/datasets/Housing. Download the data and save it in your current directory as a data file named `housing.data`.

The dataset has 506 observations. The first 13 columns contain the predictor values and the last column contains the response values. The goal is to predict the median value of owner-occupied homes in the Boston suburb area as a function of 13 predictors.

Load the data and define the response vector and predictor matrix.

```
load('housing.data');
X = housing(:,1:13);
y = housing(:,end);
```

Train a GPR model using subset of regressors (`'sr'`) approximation method with Matern 3/2 (`'Matern32'`) kernel function. Predict using the fully independent conditional (`'fic'`) method.

```
gprMdl = fitrgp(X,y,'KernelFunction','Matern32',...
'FitMethod','sr','PredictMethod','fic');
```

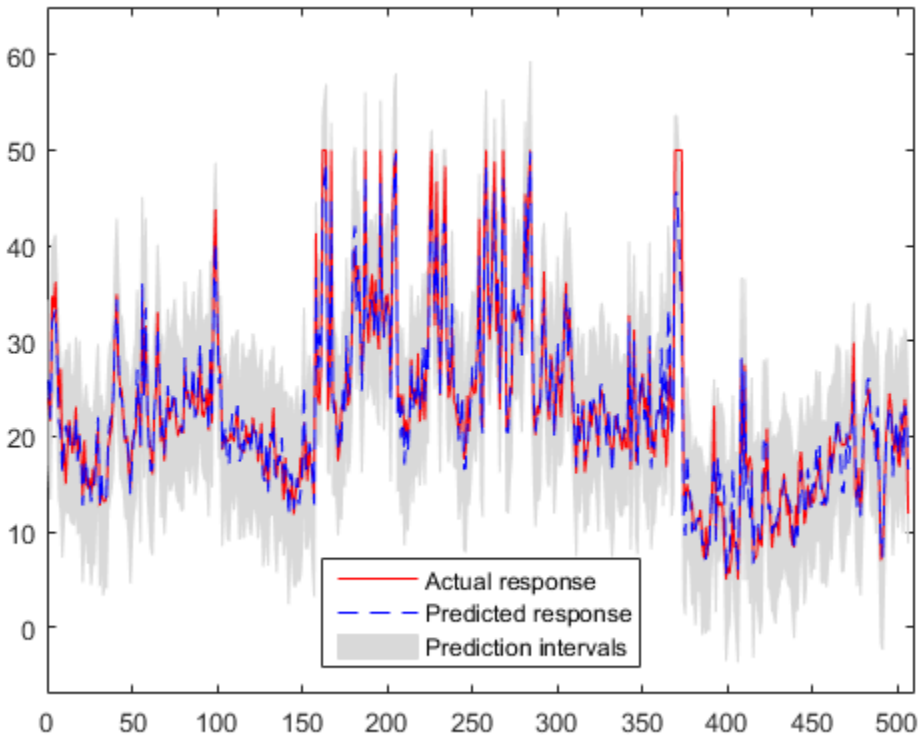
Predict the responses using the trained GPR model. Compute the 99% prediction intervals.

```
[ypred,~,yint] = resubPredict(gprMdl,'Alpha',0.01);
```

Plot the actual response values along with predictions from the GPR model.

```
figure;
h1 = area([yint(:,1) yint(:,2)-yint(:,1)],-8,...
```

```
'FaceColor',[0.85,0.85,0.85],'EdgeColor',[0.85,0.85,0.85]);
hold on;
h1(1).FaceColor = 'none'; % remove color from bottom area
h1(1).EdgeColor = 'none';
h2 = plot(y,'r'); % Plot original response values
h3 = plot(ypred,'b--'); % Plot predicted response values
legend([h2 h3 h1(2)],'Actual response','Predicted response',...
'Prediction intervals','Location','South');
axis([0 510 -7 65]);
hold off
```



The gray area shows the 99% prediction intervals.

Tips

- You can choose the prediction method while training the GPR model using the `PredictMethod` name-value pair argument in `fitrgp`. The default prediction method is `'exact'` for $n \le 10000$, where n is the number of observations in the training data, and `'bcd'` (block coordinate descent), otherwise.
- Computation of standard deviations, `ysd`, and prediction intervals, `yint`, is not supported when `PredictMethod` is `'bcd'`.

Alternatives

To compute the predicted responses for new data, use `predict`.

References

[1] Harrison, D. and D.L., Rubinfeld. "Hedonic prices and the demand for clean air." *J. Environ. Economics & Management*. Vol.5, 1978, pp. 81-102.

[2] Lichman, M. UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2013. http://archive.ics.uci.edu/ml.

See Also

`RegressionGP` | `fitrgp` | `predict` | `resubLoss`

Introduced in R2015b

resubPredict

Class: `RegressionSVM`

Predict resubstitution response of support vector machine regression model

Syntax

```
yfit = resubPredict(mdl)
```

Description

`yfit = resubPredict(mdl)` returns a vector of predicted response values, `yfit`, for the trained support vector machine (SVM) regression model `mdl` using the predictor data stored in `mdl.X`.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model returned by `fitrsvm`.

Output Arguments

`yfit` — Predicted response

vector of numeric values

Predicted responses, returned as a vector of numeric values. The length of `yfit` is equal to the number of observations in the training data, `mdl.NumObservations`.

For details about how to predict responses, see “Equation 25-1” and “Equation 25-2” in “Understanding Support Vector Machine Regression” on page 25-2.

Examples

Resubstitution Predictions for SVM Regression Model

This example shows how to train an SVM regression model, then use the model to generate predicted response values from the training data.

This example uses the abalone data from the UCI Machine Learning Repository. Download the data and save it in your current directory with the name `'abalone.data'`. Read the data into a `table`.

```
tbl = readtable('abalone.data','Filetype','text','ReadVariableNames',false);
rng default  % for reproducibility
```

The sample data contains 4177 observations. All of the predictor variables are continuous except for `sex`, which is a categorical variable with possible values `'M'` (for males), `'F'` (for females), and `'I'` (for infants). The goal is to predict the number of rings on the abalone, and thereby determine its age, using physical measurements.

Train an SVM regression model to the data, using a Gaussian kernel function with an automatic kernel scale. Standardize the data.

```
mdl = fitrsvm(tbl,'Var9','KernelFunction','gaussian','KernelScale','auto','Standardize',true);
```

Use the trained model to predict response values based on the original data.

```
yfit = resubPredict(mdl);
```

Display the first ten predicted responses alongside the actual response values.

```
[mdl.Y(1:10),yfit(1:10)]
```

```
ans =

   15.0000    8.1836
    7.0000    8.3545
    9.0000   10.9383
   10.0000    9.3446
    7.0000    6.4042
    8.0000    7.7910
   20.0000   13.8275
   16.0000   11.7959
    9.0000    9.5724
   19.0000   13.6909
```

The left column shows the actual response and the right column shows the corresponding predicted response.

References

[1] Nash, W.J., T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait." Sea Fisheries Division, Technical Report No. 48, 1994.

[2] Waugh, S. "Extending and Benchmarking Cascade-Correlation: Extensions to the Cascade-Correlation Architecture and Benchmarking of Feed-forward Supervised Artificial Neural Networks." *University of Tasmania Department of Computer Science thesis*, 1995.

[3] Clark, D., Z. Schreter, A. "Adams. A Quantitative Comparison of Dystal and Backpropagation." submitted to the Australian Conference on Neural Networks, 1996.

[4] Lichman, M. *UCI Machine Learning Repository*, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

See Also

`RegressionSVM` | `fitrsvm` | `predict` | `resubLoss`

Introduced in R2015b

resubPredict

Class: `RegressionTree`

Predict resubstitution response of tree

Syntax

```
Yfit = resubPredict(tree)
[Yfit,node] = resubPredict(tree)
[Yfit,node] = resubPredict(tree,Name,Value)
```

Description

`Yfit = resubPredict(tree)` returns the responses `tree` predicts for the data `tree.X`. `Yfit` is the predictions of `tree` on the data that `fitrtree` used to create `tree`.

`[Yfit,node] = resubPredict(tree)` returns the node numbers of `tree` for the resubstituted data.

`[Yfit,node] = resubPredict(tree,Name,Value)` predicts with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`tree`

A regression tree constructed using `fitrtree`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Subtrees` — Pruning level

0 (default) | vector of nonnegative integers | `'all'`

Pruning level, specified as the comma-separated pair consisting of `'Subtrees'` and a vector of nonnegative integers in ascending order or `'all'`.

If you specify a vector, then all elements must be at least `0` and at most `max(tree.PruneList)`. `0` indicates the full, unpruned tree and `max(tree.PruneList)` indicates the completely pruned tree (i.e., just the root node).

If you specify `'all'`, then `resubPredict` operates on all subtrees (i.e., the entire pruning sequence). This specification is equivalent to using `0:max(tree.PruneList)`.

`resubPredict` prunes `tree` to each level indicated in `Subtrees`, and then estimates the corresponding output arguments. The size of `Subtrees` determines the size of some output arguments.

To invoke `Subtrees`, the properties `PruneList` and `PruneAlpha` of `tree` must be nonempty. In other words, grow `tree` by setting `'Prune','on'`, or by pruning `tree` using `prune`.

Example: `'Subtrees','all'`

Data Types: `single` | `double` | `char` | `string`

Output Arguments

`Yfit`

The response `tree` predicts for the training data.

If the `Subtrees` name-value argument is a scalar or is missing, `label` is the same data type as the training response data `tree.Y`.

If `Subtrees` contains `m>1` entries, `label` has `m` columns, each of which represents the predictions of the corresponding subtree.

`node`

The `tree` node numbers where `tree` sends each data row.

If the `Subtrees` name-value argument is a scalar or is missing, `node` is a numeric column vector with `n` rows, the same number of rows as `tree.X`.

If `Subtrees` contains `m>1` entries, `node` is a `n`-by-`m` matrix. Each column represents the node predictions of the corresponding subtree.

Examples

Compute the In-Sample MSE

Load the `carsmall` data set. Consider `Displacement`, `Horsepower`, and `Weight` as predictors of the response `MPG`.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Grow a regression tree using all observations.

```
Mdl = fitrtree(X,MPG);
```

Compute the resubstitution MSE.

```
Yfit = resubPredict(Mdl);
mean((Yfit - Mdl.Y).^2)
```

```
ans = 4.8952
```

You can get the same result using `resubLoss`.

```
resubLoss(Mdl)
```

```
ans = 4.8952
```

Estimate In-Sample Responses For Each Subtree

Load the `carsmall` data set. Consider `Weight` as a predictor of the response `MPG`.

```
load carsmall
idxNaN = isnan(MPG + Weight);
X = Weight(~idxNaN);
Y = MPG(~idxNaN);
n = numel(X);
```

Grow a regression tree using all observations.

```
Mdl = fitrtree(X,Y);
```

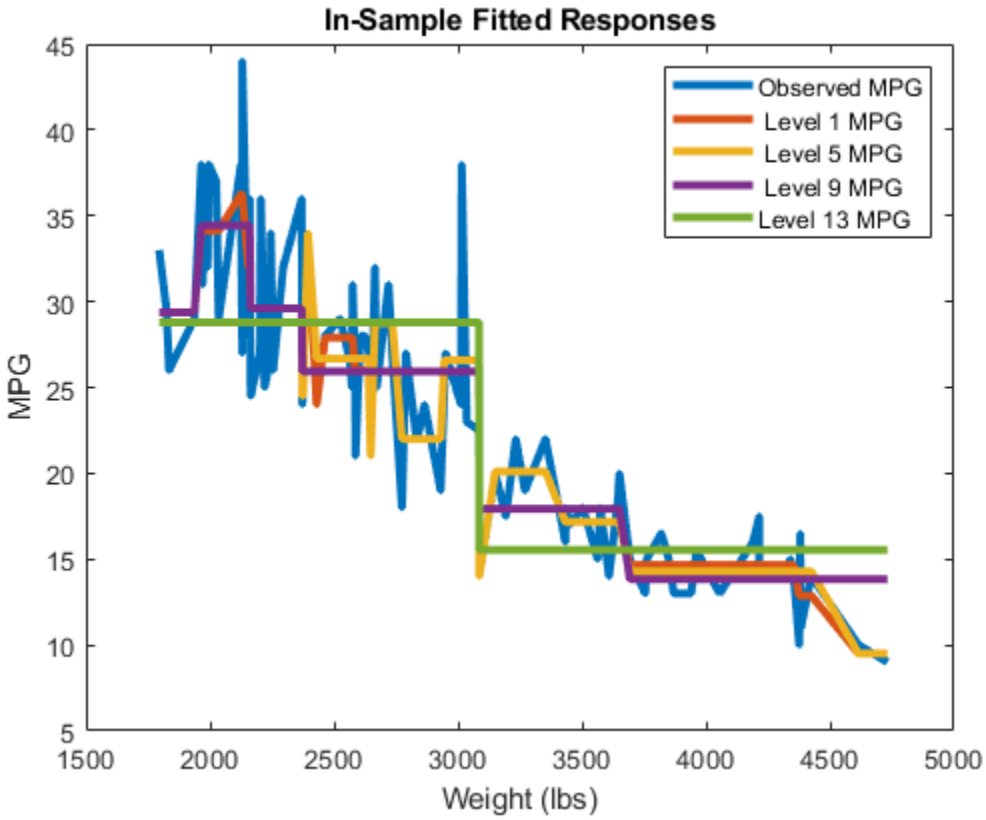
Compute resubstitution fitted values for the subtrees at several pruning levels.

```
m = max(Mdl.PruneList);
pruneLevels = 1:4:m; % Pruning levels to consider
z = numel(pruneLevels);
Yfit = resubPredict(Mdl,'SubTrees',pruneLevels);
```

`Yfit` is an `n`-by- `z` matrix of fitted values in which the rows correspond to observations and the columns correspond to a subtree.

Plot several columns of `Yfit` and `Y` against `X`.

```
figure;
sortDat = sortrows([X Y Yfit],1); % Sort all data with respect to X
plot(repmat(sortDat(:,1),1,size(Yfit,2) + 1),sortDat(:,2:end))...
    % Vectorize for efficiency
lev = cellstr(num2str((pruneLevels)','Level %d MPG'));
legend(['Observed MPG'; lev])
title 'In-Sample Fitted Responses'
xlabel 'Weight (lbs)';
ylabel 'MPG';
h = findobj(gcf);
set(h(4:end),'LineWidth',3) % Widen all lines
```



The values of `Yfit` for lower pruning levels tend to follow the data more closely than higher levels. Higher pruning levels tend to be flat for large `X` intervals.

See Also

`fitrtree` | `predict` | `resubLoss`

resume

Resume a Bayesian optimization

Syntax

```
newresults = resume(results,Name,Value)
```

Description

`newresults = resume(results,Name,Value)` resumes the optimization that produced `results` with additional options specified by one or more `Name,Value` pair arguments.

Examples

Resume a Bayesian Optimization

This example shows how to resume a Bayesian optimization. The optimization is for a deterministic function known as Rosenbrock's function, which is a well-known test case for nonlinear optimization. The function has a global minimum value of `0` at the point `[1,1]`.

Create two real variables bounded by `-5` and `5`.

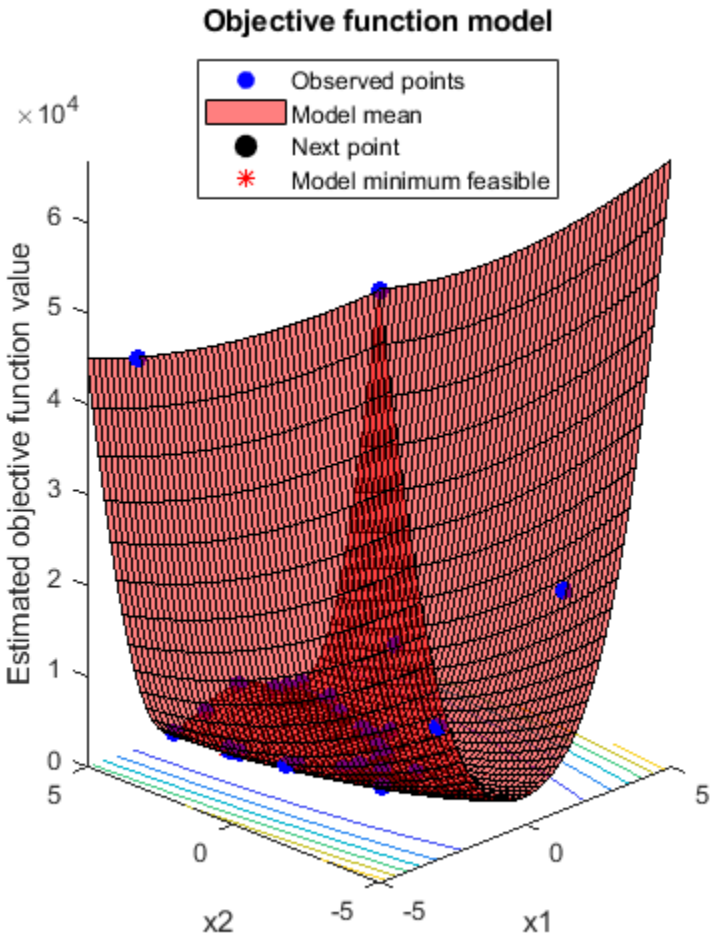
```
x1 = optimizableVariable('x1',[-5,5]);
x2 = optimizableVariable('x2',[-5,5]);
vars = [x1,x2];
```

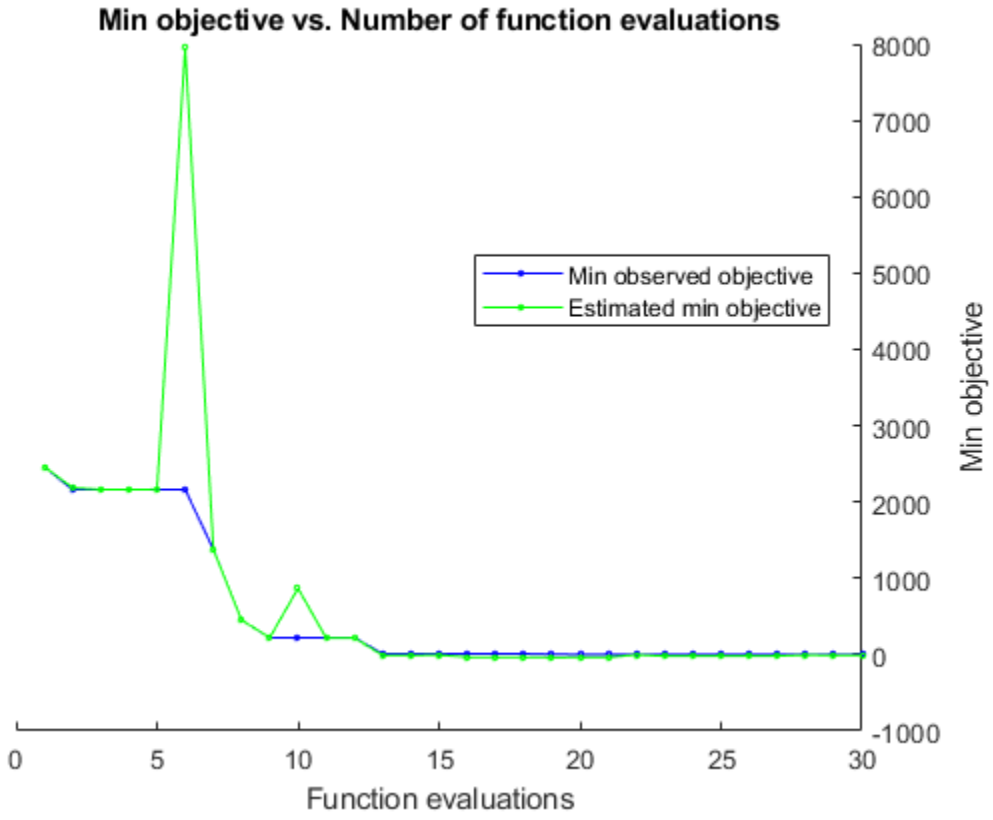
Create the objective function.

```
function f = rosenbrocks(x)

f = 100*(x.x2 - x.x1^2)^2 + (1 - x.x1)^2;


fun = @rosenbrocks;
```

For reproducibility, set the random seed, and set the acquisition function to `'expected-improvement-plus'` in the optimization.

```
rng default
results = bayesopt(fun,vars,'Verbose',0,...
    'AcquisitionFunctionName','expected-improvement-plus');
```


Objective function model
Observed points
Model mean
Next point
Model minimum feasible
×10^4
Estimated objective function value
6
5
4
3
2
1
0
5
0
-5
x2
-5
0
5
x1



View the best point found and the best modeled objective.

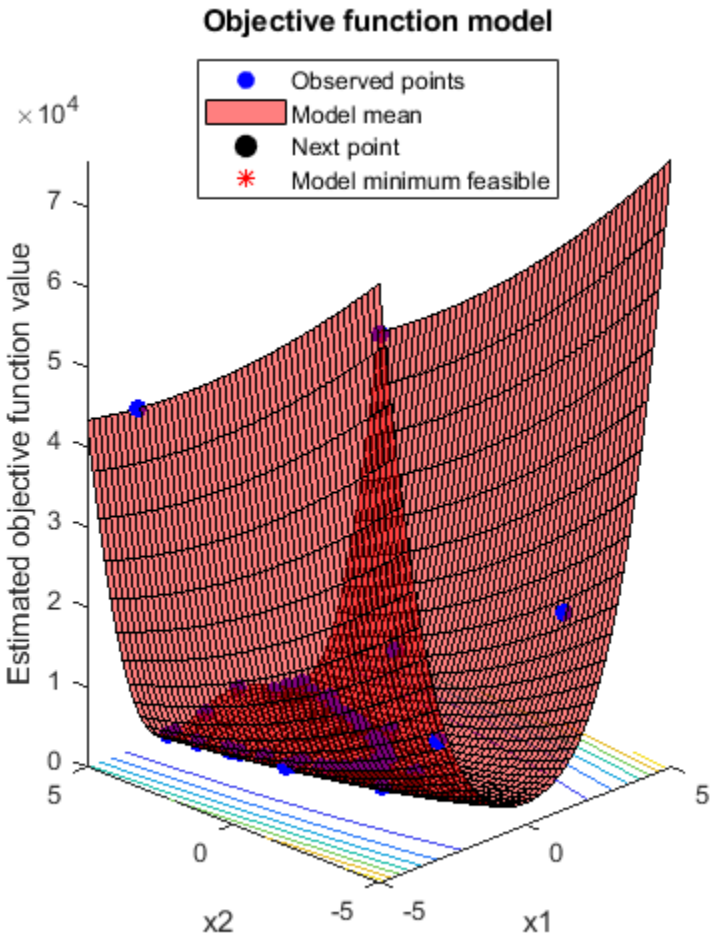
```
results.XAtMinObjective
results.MinEstimatedObjective
```

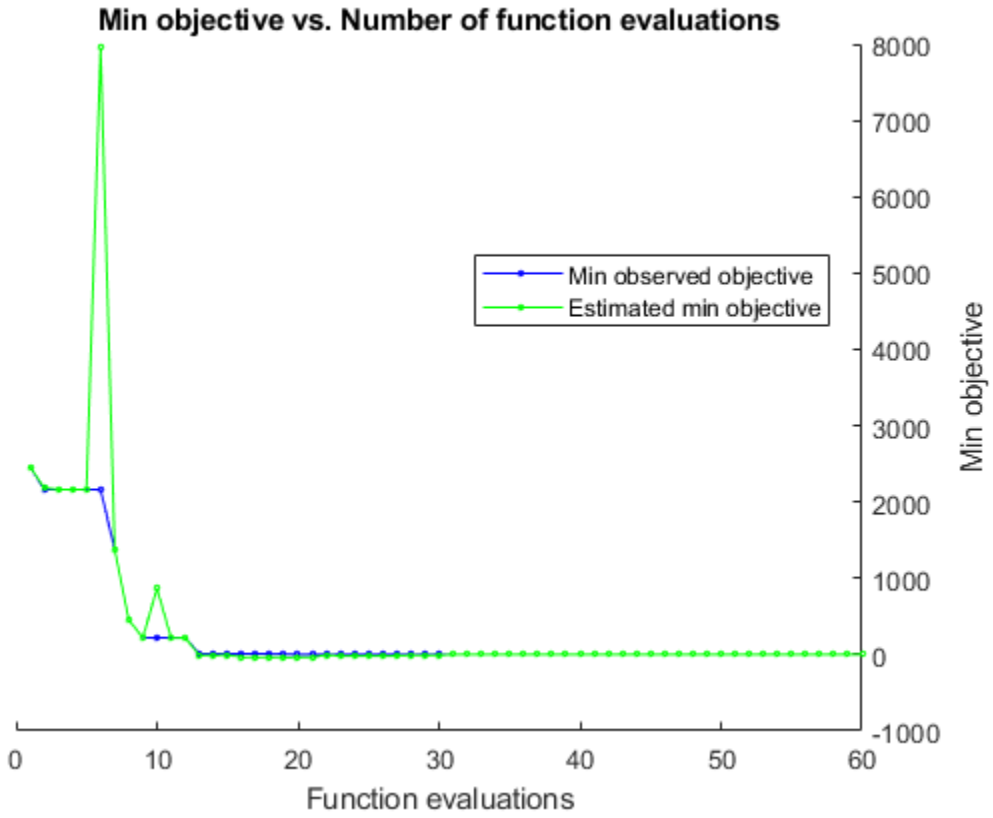
```
ans =

  1x2 table

      x1         x2
    _____    _____

    1.2421    1.5299


ans =

   -9.5402
```

The best point is somewhat close to the optimum, but the function model is inaccurate. Resume the optimization for 30 more points (a total of 60 points), this time telling the optimizer that the objective function is deterministic.

```
newresults = resume(results,'IsObjectiveDeterministic',true,'MaxObjectiveEvaluations',30);
newresults.XAtMinObjective
newresults.MinEstimatedObjective
```

```
ans =

  1x2 table

      x1         x2
    _______    _______

    1.0473     1.1002


ans =

   -0.0066
```





The objective function model is much closer to the true function this time. The best point is closer to the true optimum.

Input Arguments

`results` — Bayesian optimization results

`BayesianOptimization` object

Bayesian optimization results, specified as a `BayesianOptimization` object.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

You can use any name-value pair accepted by `bayesopt` except for those beginning with `Initial`. See the `bayesopt` “Input Arguments” on page 33-132.

Note The `MaxObjectiveEvaluations` and `MaxTime` name-value pairs mean *additional* time or evaluations, above the numbers stored in `results`. So, for example, the default number of evaluations is `30` in addition to the original specification.

Additionally, you can use the following name-value pair.

Example: `resume(results,'MaxObjectiveEvaluations',60)`

`VariableDescriptions` — Modify variable
`OptimizableVariable` object

Modify variable, specified as an `OptimizableVariable` object.

You can change only the following properties of a variable in an optimization.

- `Range` of real or integer variables. For example,

```
xvar = optimizableVariable('x',[-10,10]);
% Modify the range:
xvar.Range = [1,5];
```

- `Type` between `'integer'` and `'real'`. For example,

```
xvar.Type = 'integer';
```

- `Transform` of real or integer variables between `'log'` and `'none'`. For example,

```
xvar.Transform = 'log';
```

Output Arguments

`newresults` — Optimization results
`BayesianOptimization` object

Optimization results, returned as a `BayesianOptimization` object.

See Also

`BayesianOptimization` | `bayesopt`

Introduced in R2016b

resume

Resume training ensemble

Syntax

```
ens1 = resume(ens,nlearn)
ens1 = resume(ens,nlearn,Name,Value)
```

Description

`ens1 = resume(ens,nlearn)` trains `ens` for `nlearn` more cycles. `resume` uses the same training options `fitcensemble` used to create `ens`, except for parallel training options. If you want to resume training in parallel, pass the `'Options'` name-value pair.

Note You cannot `resume` training when `ens` is a `Subspace` ensemble created with `'AllPredictorCombinations'` number of learners.

`ens1 = resume(ens,nlearn,Name,Value)` trains `ens` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A classification ensemble, created with `fitcensemble`.

`nlearn`

A positive integer, the number of cycles for additional training of `ens`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NPrint`

Printout frequency, a positive integer scalar or `'off'` (no printouts). When `NPrint` is a positive integer, displays a message to the command line after training `NPrint` weak learners.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Default: `'off'`

`Options`

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

You can use the same parallel options for `resume` as you used for the original training. However, you can change the parallel options as needed. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `resume` uses the default stream or streams.

For dual-core systems and above, `resume` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

`ens1`

The classification ensemble `ens`, augmented with additional training.

Examples

Train Classification Ensemble for Additional Cycles

Train a classification ensemble for three cycles, and compare the resubstitution error obtained after training the ensemble for more cycles.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a classification ensemble for three cycles and examine the resubstitution error.

```
ens = fitcensemble(X,Y,'Method','GentleBoost','NumLearningCycles',3);
L = resubLoss(ens)
```

```
L = 0.0085
```

Train for three more cycles and examine the new resubstitution error.

```
ens1 = resume(ens,3);
L = resubLoss(ens1)
```

```
L = 0
```

The resubstitution error is much lower in the new ensemble than the original.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`resume` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

See Also

`ClassificationBaggedEnsemble` | `ClassificationEnsemble` | `fitcensemble`

resume

Resume training learners on cross-validation folds

Syntax

```
ens1 = resume(ens,nlearn)
ens1 = resume(ens,nlearn,Name,Value)
```

Description

`ens1 = resume(ens,nlearn)` trains `ens` in every fold for `nlearn` more cycles. `resume` uses the same training options `fitcensemble` used to create `ens`, except for parallel training options. If you want to resume training in parallel, pass the `'Options'` name-value pair.

`ens1 = resume(ens,nlearn,Name,Value)` trains `ens` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A cross-validated classification ensemble. `ens` is the result of either:

- The `fitcensemble` function with a cross-validation name-value pair. The names are `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.
- The `crossval` method applied to a classification ensemble.

`nlearn`

A positive integer, the number of cycles for additional training of `ens`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NPrint`

Printout frequency, a positive integer scalar or `'off'` (no printouts). When `NPrint` is a positive integer, displays a message to the command line after training `NPrint` folds.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Default: `'off'`

Options

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

You can use the same parallel options for `resume` as you used for the original training. However, you can change the parallel options as needed. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `resume` uses the default stream or streams.

For dual-core systems and above, `resume` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

ens1

The cross-validated classification ensemble `ens`, augmented with additional training.

Examples

Train Partitioned Classification Ensemble for More Cycles

Train a partitioned classification ensemble for 10 cycles, and compare the classification loss obtained after training the ensemble for more cycles.

Load the `ionosphere` data set.

```
load ionosphere
```

Train a partitioned classification ensemble for 10 cycles and examine the error.

```
t = templateTree('MaxNumSplits',1); % Weak learner template tree object
cvens = fitcensemble(X,Y,'Method','GentleBoost','NumLearningCycles',10,'Learners',t,'crossval','
rng(10,'twister') % For reproducibility
L = kfoldLoss(cvens)
```

```
L = 0.0940
```

Train for 10 more cycles and examine the new error.

```
cvens = resume(cvens,10);
L = kfoldLoss(cvens)
```

```
L = 0.0712
```

The cross-validation error is lower in the ensemble after training for 10 more cycles.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`resume` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

See Also

`ClassificationPartitionedEnsemble` | `kfoldEdge` | `kfoldLoss` | `kfoldMargin` | `kfoldPredict`

resume

Resume training support vector machine (SVM) classifier

Syntax

```
UpdatedSVMModel = resume(SVMModel,numIter)
UpdatedSVMModel = resume(SVMModel,numIter,Name,Value)
```

Description

`UpdatedSVMModel = resume(SVMModel,numIter)` returns an updated support vector machine (SVM) classifier `UpdatedSVMModel` by training the SVM classifier `SVMModel` for `numIter` more iterations. Like `SVMModel`, the updated SVM classifier is a `ClassificationSVM` classifier.

`resume` continues applying the training options set when `SVMModel` was trained with `fitcsvm`.

`UpdatedSVMModel = resume(SVMModel,numIter,Name,Value)` returns `UpdatedSVMModel` with additional options specified by one or more name-value pair arguments. For example, you can specify the verbosity level.

Examples

Resume Training SVM Classifier

Train an SVM classifier and intentionally cause the solver to fail to converge onto a solution. Then resume training the classifier without having to restart the entire learning process.

Load the `ionosphere` data set.

```
load ionosphere
rng(1); % For reproducibility
```

Train an SVM classifier. Specify that the optimization routine uses at most 50 iterations.

```
SVMModel = fitcsvm(X,Y,'IterationLimit',50);
DidConverge = SVMModel.ConvergenceInfo.Converged
```

```
DidConverge = logical
   0
```

```
Reason = SVMModel.ConvergenceInfo.ReasonForConvergence
```

```
Reason =
'NoConvergence'
```

`DidConverge = 0` indicates that the optimization routine did not converge onto a solution. `Reason` states the reason why the routine did not converge. Therefore, `SVMModel` is a partially trained SVM classifier.

Resume training the SVM classifier for another `1500` iterations.

```
UpdatedSVMModel = resume(SVMModel,1500);
DidConverge = UpdatedSVMModel.ConvergenceInfo.Converged
```

```
DidConverge = logical
   1
```

```
Reason = UpdatedSVMModel.ConvergenceInfo.ReasonForConvergence
```

```
Reason =
'DeltaGradient'
```

`DidConverge` indicates that the optimization routine converged onto a solution. `Reason` indicates that the gradient difference (`DeltaGradient`) reached its tolerance level (`DeltaGradientTolerance`). Therefore, `SVMModel` is a fully trained SVM classifier.

Monitor Training of SVM Classifier

Train an SVM classifier and intentionally cause the solver to fail to converge onto a solution. Then resume training the classifier without having to restart the entire learning process. Compare values of the resubstitution loss for the partially trained classifier and the fully trained classifier.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an SVM classifier. Specify that the optimization routine uses at most 100 iterations. Monitor the algorithm specifying that the software prints diagnostic information every `50` iterations.

```
SVMModel = fitcsvm(X,Y,'IterationLimit',100,'Verbose',1,'NumPrint',50);
```

```
|=============================================================================================================
|   Iteration  | Set  |   Set Size   |  Feasibility  |     Delta     |      KKT      |  Number o
|              |      |              |      Gap      |    Gradient   |   Violation   |  Supp. Ve
|=============================================================================================================
|           0 |active|          351 |  9.971591e-01 |  2.000000e+00 |  1.000000e+00 |
|          50 |active|          351 |  8.064425e-01 |  3.736929e+00 |  2.161317e+00 |

SVM optimization did not converge to the required tolerance.
```

The software prints an iterative display to the Command Window. The printout indicates that the optimization routine has not converged onto a solution.

Estimate the resubstitution loss of the partially trained SVM classifier.

```
partialLoss = resubLoss(SVMModel)
```

```
partialLoss = 0.1054
```

The training sample misclassification error is approximately 12%.

Resume training the classifier for another `1500` iterations. Specify that the software print diagnostic information every `250` iterations.

```
UpdatedSVMModel = resume(SVMModel,1500,'NumPrint',250)
```

```
|=============================================================================================================
|   Iteration   | Set  |   Set Size    |  Feasibility  |     Delta     |      KKT      |  Number o
|               |      |               |      Gap      |    Gradient   |   Violation   |  Supp. Ve
|=============================================================================================================
|           250 |active|           351 |  1.441556e-01 |  1.701201e+00 |  1.015454e+00 |           1
|           500 |active|           351 |  3.277736e-03 |  9.155364e-02 |  4.830095e-02 |           1
|           750 |active|           351 |  3.928360e-04 |  1.367091e-02 |  9.155316e-03 |           1
|          1000 |active|           351 |  4.802547e-05 |  1.551900e-03 |  7.765843e-04 |           1
|          1044 |active|           351 |  3.602828e-05 |  9.382457e-04 |  5.182592e-04 |           1

 Exiting Active Set upon convergence due to DeltaGradient.

UpdatedSVMModel =
  ClassificationSVM
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'b'  'g'}
           ScoreTransform: 'none'
          NumObservations: 351
                    Alpha: [103x1 double]
                     Bias: -3.8828
         KernelParameters: [1x1 struct]
           BoxConstraints: [351x1 double]
          ConvergenceInfo: [1x1 struct]
          IsSupportVector: [351x1 logical]
                   Solver: 'SMO'


  Properties, Methods
```

The software resumes at iteration `1000` and uses the same verbosity level as the one set when you trained the model using `fitcsvm`. The printout indicates that the algorithm converged. Therefore, `UpdatedSVMModel` is a fully trained `ClassificationSVM` classifier.

```
updatedLoss = resubLoss(UpdatedSVMModel)

updatedLoss = 0.0769
```

The training sample misclassification error of the fully trained classifier is approximately 8%.

Input Arguments

`SVMModel` — Full, trained SVM classifier

`ClassificationSVM` classifier

Full, trained SVM classifier, specified as a `ClassificationSVM` model trained with `fitcsvm`.

`numIter` — Number of iterations

positive integer

Number of iterations to continue training the SVM classifier, specified as a positive integer.

Data Types: `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `resume(SVMModel,500,'Verbose',2)` trains `SVMModel` for `500` more iterations and specifies displaying diagnostic messages and saving convergence criteria at every iteration.

`Verbose` — Verbosity level

`0` | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. `Verbose` controls the amount of optimization information displayed in the Command Window and saved as a structure to `SVMModel.ConvergenceInfo.History`.

This table summarizes the verbosity level values.

Value	Description
`0`	The software does not display or save convergence information.
`1`	The software displays diagnostic messages and saves convergence criteria every *`numprint`* iterations, where *`numprint`* is the value of the `'NumPrint'` name-value pair argument.
`2`	The software displays diagnostic messages and saves convergence criteria at every iteration.

By default, `Verbose` is the value that `fitcsvm` uses to train `SVMModel`.

Example: `'Verbose',1`

Data Types: `single`

`NumPrint` — Number of iterations between diagnostic message printouts

nonnegative integer

Number of iterations between diagnostic message printouts, specified as the comma-separated pair consisting of `'NumPrint'` and a nonnegative integer.

If you set `'Verbose',1` and `'NumPrint',`*`numprint`*, then the software displays all optimization diagnostic messages from SMO [1] and ISDA [2] every *`numprint`* iterations in the Command Window.

By default, `NumPrint` is the value that `fitcsvm` uses to train `SVMModel`.

Example: `'NumPrint',500`

Data Types: `single`

Tips

If optimization does not converge and the solver is `'SMO'` or `'ISDA'`, then try to resume training the SVM classifier.

References

[1] Fan, R.-E., P.-H. Chen, and C.-J. Lin. “Working set selection using second order information for training support vector machines.” *Journal of Machine Learning Research*, Vol. 6, 2005, pp. 1889–1918.

[2] Kecman V., T. -M. Huang, and M. Vogt. “Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance.” *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ClassificationSVM` | `fitcsvm`

Introduced in R2014a

resume

Resume training of generalized additive model (GAM)

Syntax

```
UpdatedMdl = resume(Mdl,numTrees)
UpdatedMdl = resume(Mdl,numTrees,Name,Value)
```

Description

`UpdatedMdl = resume(Mdl,numTrees)` returns an updated generalized additive model `UpdatedMdl` by training `Mdl` for `numTrees` more iterations with the same options used to train `Mdl`.

For each iteration, `resume` trains one predictor tree per linear term or one interaction tree per interaction term.

- If `Mdl` contains only linear terms for predictors (predictor trees), then `resume` trains an additional `numTrees` number of trees per predictor.
- If `Mdl` contains both linear and interaction terms for predictors (predictor trees and interaction trees), then `resume` trains an additional `numTrees` number of trees per interaction term.

`resume` does not add new terms to the model. If you want to add interaction terms to a model that contains only linear terms, use the `addInteractions` function.

`UpdatedMdl = resume(Mdl,numTrees,Name,Value)` specifies additional options using one or more name-value arguments. For example, `'Verbose',2` specifies the verbosity level as 2 to display diagnostic messages at every iteration.

Examples

Resume Training Predictor Trees in GAM

Train a univariate classification GAM (which contains only linear terms) for a small number of iterations. After training the model for more iterations, compare the resubstitution loss.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a univariate GAM that identifies whether the radar return is bad (`'b'`) or good (`'g'`). Specify the number of trees per linear term as 2. `fitcgam` iterates the boosting algorithm for the specified number of iterations. For each boosting iteration, the function adds one tree per linear term. Specify `'Verbose'` as 2 to display diagnostic messages at every iteration.

```
Mdl = fitcgam(X,Y,'NumTreesPerPredictor',2,'Verbose',2);
```

```
|=====================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
```

```
|=========================================================|
|    1D|         0|      486.59|      -     |      -     |
|    1D|         1|      166.71|         Inf|           1|
|    1D|         2|      78.336|     0.58205|           1|
```

To check whether `fitcgam` trains the specified number of trees, display the `ReasonForTermination` property of the trained model and view the displayed message.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: ''
```

Compute the classification loss for the training data.

```
resubLoss(Mdl)
```

```
ans = 0.0142
```

Resume training the model for another 100 iterations. Because `Mdl` contains only linear terms, the `resume` function resumes training for the linear terms and adds more trees for them (predictor trees). Specify `'Verbose'` and `'NumPrint'` to display diagnostic messages at every 10 iterations.

```
UpdatedMdl = resume(Mdl,100,'Verbose',1,'NumPrint',10);
```

```
|=========================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=========================================================|
|    1D|         0|      78.336|      -     |      -     |
|    1D|         1|      38.364|     0.17429|           1|
|    1D|        10|     0.16311|    0.011894|           1|
|    1D|        20|  0.00035693|   0.0025178|           1|
|    1D|        30|  8.1191e-07|   0.0011006|           1|
|    1D|        40|  1.7978e-09|  0.00074607|           1|
|    1D|        50|  3.6113e-12|  0.00034404|           1|
|    1D|        60|  1.7497e-13|  0.00016541|           1|
```

```
UpdatedMdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Unable to improve the model fit.'
    InteractionTrees: ''
```

`resume` terminates training when adding more trees does not improve the deviance of the model fit.

Compute the classification loss using the updated model.

```
resubLoss(UpdatedMdl)
```

```
ans = 0
```

The classification loss decreases after `resume` updates the model with more iterations.

Resume Training Interaction Trees in GAM

Train a regression GAM that contains both linear and interaction terms. Specify to train the interaction terms for a small number of iterations. After training the interaction terms for more iterations, compare the resubstitution loss.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Specify `Acceleration`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a GAM that includes all available linear and interaction terms in `X`. Specify the number of trees per interaction term as 2. `fitrgam` iterates the boosting algorithm 300 times (default) for linear terms, and iterates the algorithm the specified number of iterations for interaction terms. For each boosting iteration, the function adds one tree per linear term or one tree per interaction term. Specify `'Verbose'` as 1 to display diagnostic messages at every 10 iterations.

```
Mdl = fitrgam(X,Y,'Interactions','all','NumTreesPerInteraction',2,'Verbose',1);
```

```
|=========================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate   |
|=========================================================|
|    1D|         0|  2.4432e+05|      -     |      -      |
|    1D|         1|      9507.4|         Inf|            1|
|    1D|        10|      4470.6|  0.00025206|            1|
|    1D|        20|      3895.3|  0.00011448|            1|
|    1D|        30|      3617.7|  3.5365e-05|            1|
|    1D|        40|      3402.5|  3.7992e-05|            1|
|    1D|        50|      3257.1|  2.4983e-05|            1|
|    1D|        60|      3131.8|  2.3873e-05|            1|
|    1D|        70|      3019.8|  2.2967e-05|            1|
|    1D|        80|      2925.9|  2.8071e-05|            1|
|    1D|        90|      2845.3|  1.6811e-05|            1|
|    1D|       100|      2772.7|   1.852e-05|            1|
|    1D|       110|      2707.8|  1.6754e-05|            1|
|    1D|       120|      2649.8|   1.651e-05|            1|
|    1D|       130|      2596.6|  1.1723e-05|            1|
|    1D|       140|      2547.4|   1.813e-05|            1|
|    1D|       150|      2501.1|  1.8659e-05|            1|
|    1D|       160|      2455.7|   1.386e-05|            1|
|    1D|       170|      2416.9|  1.0615e-05|            1|
|    1D|       180|      2377.2|   8.534e-06|            1|
|    1D|       190|        2339|  7.6771e-06|            1|
|    1D|       200|      2303.3|  9.5866e-06|            1|
|    1D|       210|      2270.7|  8.4276e-06|            1|
|    1D|       220|      2240.1|  8.5778e-06|            1|
|    1D|       230|      2209.2|  9.6761e-06|            1|
|    1D|       240|      2178.7|  7.0622e-06|            1|
|    1D|       250|      2150.3|  8.3082e-06|            1|
|    1D|       260|      2122.3|  7.9542e-06|            1|
|    1D|       270|      2097.7|  7.6328e-06|            1|
|    1D|       280|      2070.4|  9.4322e-06|            1|
|    1D|       290|      2044.3|  7.5722e-06|            1|
```

```
|    1D|       300|      2019.7|  6.6719e-06|           1|
|=====================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=====================================================|
|    2D|         0|      2019.7|          -     |      -     |
|    2D|         1|      1795.5|   0.0005975|           1|
|    2D|         2|      1523.4|   0.0010079|           1|
```

To check whether `fitrgam` trains the specified number of trees, display the `ReasonForTermination` property of the trained model and view the displayed messages.

```
Mdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Terminated after training the requested number of trees.'
```

Compute the regression loss for the training data.

```
resubLoss(Mdl)
```

```
ans = 3.8277
```

Resume training the model for another 100 iterations. Because `Mdl` contains both linear and interaction terms, the `resume` function resumes training for the interaction terms and adds more trees for them (interaction trees).

```
UpdatedMdl = resume(Mdl,100);
```

```
|=====================================================|
| Type | NumTrees |  Deviance  |   RelTol   | LearnRate  |
|=====================================================|
|    2D|         0|      1523.4|          -     |      -     |
|    2D|         1|      1363.9|  0.00039695|           1|
|    2D|        10|      594.04|  8.0295e-05|           1|
|    2D|        20|      359.44|  4.3201e-05|           1|
|    2D|        30|      238.51|  2.6869e-05|           1|
|    2D|        40|      153.98|  2.6271e-05|           1|
|    2D|        50|      91.464|  8.0936e-06|           1|
|    2D|        60|      61.882|  3.8528e-06|           1|
|    2D|        70|      43.206|  5.9888e-06|           1|
```

```
UpdatedMdl.ReasonForTermination
```

```
ans = struct with fields:
      PredictorTrees: 'Terminated after training the requested number of trees.'
    InteractionTrees: 'Unable to improve the model fit.'
```

`resume` terminates training when adding more trees does not improve the deviance of the model fit.

Compute the regression loss using the updated model.

```
resubLoss(UpdatedMdl)
```

```
ans = 0.0944
```

The regression loss decreases after `resume` updates the model with more iterations.

Input Arguments

`Mdl` — Generalized additive model

`ClassificationGAM` model object | `RegressionGAM` model object

Generalized additive model, specified as a `ClassificationGAM` or `RegressionGAM` model object.

`numTrees` — Number of trees to add

positive integer scalar

Number of trees to add, specified as a positive integer scalar.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Verbose',1,'NumPrint',100` specifies to print diagnostic messages in the Command Window every 100 iterations.

`NumPrint` — Number of iterations between diagnostic message printouts

`Mdl.ModelParameters.NumPrint` (default) | nonnegative integer scalar

Number of iterations between diagnostic message printouts, specified as a nonnegative integer scalar. This argument is valid only when you specify `'Verbose'` as 1.

If you specify `'Verbose',1` and `'NumPrint',numPrint`, then the software displays diagnostic messages every `numPrint` iterations in the Command Window.

The default value is `Mdl.ModelParameters.NumPrint`, which is the `NumPrint` value that you specify when creating the GAM object `Mdl`.

Example: `'NumPrint',500`

Data Types: `single` | `double`

`Verbose` — Verbosity level

`Mdl.ModelParameters.VerbosityLevel` (default) | `0` | `1` | `2`

Verbosity level, specified as `0`, `1`, or `2`. The `Verbose` value controls the amount of information that the software displays in the Command Window.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software displays no information.
`1`	The software displays diagnostic messages every `numPrint` iterations, where `numPrint` is the `'NumPrint'` value.
`2`	The software displays diagnostic messages at every iteration.

Each line of the diagnostic messages shows the information about each boosting iteration and includes the following columns:

- `Type` — Type of trained trees, `1D` (predictor trees, or boosted trees for linear terms for predictors) or `2D` (interaction trees, or boosted trees for interaction terms for predictors)
- `NumTrees` — Number of trees per linear term or interaction term that `resume` added to the model so far
- `Deviance` — "Deviance" on page 33-5692 of the model
- `RelTol` — Relative change of model predictions: $(\widehat{y}_k - \widehat{y}_{k-1})'(\widehat{y}_k - \widehat{y}_{k-1})/\widehat{y}_k'\widehat{y}_k$, where $\widehat{y}_k$ is a column vector of model predictions at iteration k
- `LearnRate` — Learning rate used for the current iteration

The default value is `Mdl.ModelParameters.VerbosityLevel`, which is the `Verbose` value that you specify when creating the GAM object `Mdl`.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`UpdatedMdl` — Updated generalized additive model

`ClassificationGAM` model object | `RegressionGAM` model object

Updated generalized additive model, returned as a `ClassificationGAM` or `RegressionGAM` model object. `UpdatedMdl` has the same object type as the input model `Mdl`.

To overwrite the input argument `Mdl`, assign the output of `resume` to `Mdl`:

```
Mdl = resume(Mdl,numTrees);
```

More About

Deviance

Deviance is a generalization of the residual sum of squares. It measures the goodness of fit compared to the saturated model.

The deviance of a fitted model is twice the difference between the loglikelihoods of the model and the saturated model:

$$-2(\log L - \log L_s),$$

where L and L_s are the likelihoods of the fitted model and the saturated model, respectively. The saturated model is the model with the maximum number of parameters that you can estimate.

`resume` uses the deviance to measure the goodness of model fit and finds a learning rate that reduces the deviance at each iteration. Specify `'Verbose'` as 1 or 2 to display the deviance and learning rate in the Command Window.

See Also

`ClassificationGAM` | `RegressionGAM` | `addInteractions`

Topics
“Train Generalized Additive Model for Binary Classification” on page 12-77
“Train Generalized Additive Model for Regression” on page 12-91

Introduced in R2021a

resume

Resume training ensemble

Syntax

```
ens1 = resume(ens,nlearn)
ens1 = resume(ens,nlearn,Name,Value)
```

Description

`ens1 = resume(ens,nlearn)` trains `ens` for `nlearn` more cycles. `resume` uses the same training options `fitrensemble` used to create `ens`, except for parallel training options. If you want to resume training in parallel, pass the `'Options'` name-value pair.

`ens1 = resume(ens,nlearn,Name,Value)` trains `ens` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A regression ensemble, created with `fitrensemble`.

`nlearn`

A positive integer, the number of cycles for additional training of `ens`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NPrint`

Printout frequency, a positive integer scalar or `'off'` (no printouts). When `NPrint` is a positive integer, displays a message to the command line after training `NPrint` weak learners.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Default: `'off'`

`Options`

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

You can use the same parallel options for `resume` as you used for the original training. However, you can change the parallel options as needed. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `resume` uses the default stream or streams.

For dual-core systems and above, `resume` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

`ens1`

The regression ensemble `ens`, augmented with additional training.

Examples

Train Regression Ensemble for Additional Cycles

Train a regression ensemble for 50 cycles, and compare the resubstitution error obtained after training the ensemble for more cycles.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train a regression ensemble for 50 cycles and examine the resubstitution error.

```
ens = fitrensemble(X,MPG,'NumLearningCycles',50);
L = resubLoss(ens)
```

```
L = 0.5563
```

Train for 50 more cycles and examine the new resubstitution error.

```
ens = resume(ens,50);
L = resubLoss(ens)
```

```
L = 0.3463
```

The resubstitution error is lower in the new ensemble than in the original.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`resume` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

See Also

`RegressionBaggedEnsemble` | `RegressionEnsemble` | `fitrensemble`

resume

Resume training ensemble

Syntax

```
ens1 = resume(ens,nlearn)
ens1 = resume(ens,nlearn,Name,Value)
```

Description

`ens1 = resume(ens,nlearn)` trains `ens` in every fold for `nlearn` more cycles. `resume` uses the same training options `fitrensemble` used to create `ens`, except for parallel training options. If you want to resume training in parallel, pass the `'Options'` name-value pair.

`ens1 = resume(ens,nlearn,Name,Value)` trains `ens` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`ens`

A cross-validated regression ensemble. `ens` is the result of either:

- The `fitrensemble` function with a cross-validation name-value pair. The names are `'crossval'`, `'kfold'`, `'holdout'`, `'leaveout'`, or `'cvpartition'`.
- The `crossval` method applied to a regression ensemble.

`nlearn`

A positive integer, the number of cycles for additional training of `ens`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`NPrint`

Printout frequency, a positive integer scalar or `'off'` (no printouts). When `NPrint` is a positive integer, displays a message to the command line after training `NPrint` folds.

Tip For fastest training of some boosted decision trees, set `NPrint` to the default value `'off'`. This tip holds when the classification `Method` is `'AdaBoostM1'`, `'AdaBoostM2'`, `'GentleBoost'`, or `'LogitBoost'`, or when the regression `Method` is `'LSBoost'`.

Default: `'off'`

Options

Options for computing in parallel and setting random numbers, specified as a structure. Create the `Options` structure with `statset`.

Note You need Parallel Computing Toolbox to compute in parallel.

You can use the same parallel options for `resume` as you used for the original training. However, you can change the parallel options as needed. This table lists the option fields and their values.

Field Name	Value	Default
`UseParallel`	Set this value to `true` to compute in parallel. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.	`false`
`UseSubstreams`	Set this value to `true` to run computations in parallel in a reproducible manner. To compute reproducibly, set `Streams` to a type that allows substreams: `'mlfg6331_64'` or `'mrg32k3a'`.	`false`
`Streams`	Specify this value as a `RandStream` object or cell array of such objects. Use a single object except when the `UseParallel` value is `true` and the `UseSubstreams` value is `false`. In that case, use a cell array that has the same size as the parallel pool.	If you do not specify `Streams`, then `resume` uses the default stream or streams.

For dual-core systems and above, `resume` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, specifying the `UseParallel` option as `true` might not provide a significant speedup on a single computer. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

Example: `'Options',statset('UseParallel',true)`

Output Arguments

ens1

The cross-validated regression ensemble `ens`, augmented with additional training.

Examples

Cross-Validate Regression Ensemble Augmented with Additional Training

Examine the cross-validation error after training a regression ensemble for more cycles.

Load the `carsmall` data set and select displacement, horsepower, and vehicle weight as predictors.

```
load carsmall
X = [Displacement Horsepower Weight];
```

Train a regression ensemble for 50 cycles.

```
ens = fitrensemble(X,MPG,'NumLearningCycles',50);
```

Cross-validate the ensemble and examine the cross-validation error.

```
rng(10,'twister') % For reproducibility
cvens = crossval(ens);
L = kfoldLoss(cvens)
```

```
L = 27.9435
```

Train for 50 more cycles and examine the new cross-validation error.

```
cvens = resume(cvens,50);
L = kfoldLoss(cvens)
```

```
L = 28.7114
```

The additional training did not improve the cross-validation error.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

`resume` supports parallel training using the `'Options'` name-value argument. Create options using `statset`, such as `options = statset('UseParallel',true)`. Parallel ensemble training requires you to set the `'Method'` name-value argument to `'Bag'`. Parallel training is available only for tree learners, the default type for `'Bag'`.

See Also

`RegressionPartitionedEnsemble` | `fitrensemble` | `kfoldLoss`

resume

Class: RegressionSVM

Resume training support vector machine regression model

Syntax

```
updatedMdl = resume(mdl,numIter)
updatedMdl = resume(mdl,numIter,Name,Value)
```

Description

`updatedMdl = resume(mdl,numIter)` returns an updated support vector machine (SVM) regression model, `updatedMdl`, by training the model for an additional number of iterations as specified by `numIter`.

`resume` applies the same training options to `updatedMdl` that you set when using `fitrsvm` to train `mdl`.

`updatedMdl = resume(mdl,numIter,Name,Value)` returns an updated SVM regression model with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`mdl` — Full, trained SVM regression model

`RegressionSVM` model

Full, trained SVM regression model, specified as a `RegressionSVM` model trained using `fitrsvm`.

`numIter` — Number of iterations

positive integer value

Number of iterations to continue training the SVM regression model, specified as a positive integer value.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Verbose` — Verbosity level

`0` | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `0`, `1`, or `2`. `Verbose` controls the amount of optimization information that the software displays to the Command Window and is saved in the model as `mdl.ModelParameters.VerbosityLevel`.

By default, `Verbose` is the value that `fitrsvm` used to train `mdl`.

Example: `'Verbose',1`

Data Types: `single` | `double`

`NumPrint` — Number of iterations between diagnostic message printouts
nonnegative integer value

Number of iterations between diagnostic message printouts, specified as the comma-separated pair consisting of `'NumPrint'` and a nonnegative integer.

If you set `'Verbose',1` and `'NumPrint',`*`numprint`*, then the software displays optimization diagnostic messages to the Command Window every *numprint* number of iterations .

By default, `NumPrint` is the value that `fitrsvm` used to train `mdl`.

Example: `'NumPrint',500`

Data Types: `single` | `double`

Output Arguments

`updatedMdl` — Updated SVM regression model
`RegressionSVM` model

Updated SVM regression model, returned as a `RegressionSVM` model.

Examples

Resume Training an SVM Regression Model

This example shows how to resume training an SVM regression model that failed to converge without restarting the entire learning process.

Load the `carsmall` data set.

```
load carsmall
rng default  % for reproducibility
```

Specify `Acceleration`, `Cylinders`, `Displacement`, `Horsepower`, and `Weight` as the predictor variables (`X`) and `MPG` as the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Train a linear SVM regression model. For illustration purposes, set the iteration limit to 50. Standardize the data.

```
mdl = fitrsvm(X,Y,'IterationLimit',50,'Standardize',true);
```

Check to confirm whether the model converged.

```
mdl.ConvergenceInfo.Converged
```

```
ans =

     0
```

The returned value of `0` indicates that the model did not converge.

Resume training the model for up to an additional 100 iterations.

```
updatedMdl = resume(mdl,100);
```

Check to confirm whether the updated model converged.

```
updatedMdl.ConvergenceInfo.Converged
```

```
ans =

     1
```

The returned value of `1` indicates that the updated model did converge.

Check the reason for convergence and the total number of iterations required.

```
updatedMdl.ConvergenceInfo.ReasonForConvergence
updatedMdl.NumIterations
```

```
ans =

FeasibilityGap


ans =

    97
```

The model converged because the feasibility gap reached its tolerance value after 97 iterations.

Tips

If optimization has not converged and `'Solver'` is set to `'SMO'` or `'ISDA'`, then try to resume training the SVM regression model.

See Also

`RegressionSVM` | `fitrsvm`

Introduced in R2015b

rica

Feature extraction by using reconstruction ICA

Syntax

```
Mdl = rica(X,q)
Mdl = rica(X,q,Name,Value)
```

Description

`Mdl = rica(X,q)` returns a reconstruction independent component analysis (RICA) model object that contains the results from applying RICA to the table or matrix of predictor data `X` containing *p* variables. `q` is the number of features to extract from `X`, therefore `rica` learns a *p*-by-`q` matrix of transformation weights. For undercomplete or overcomplete feature representations, `q` can be less than or greater than the number of predictor variables, respectively.

- To access the learned transformation weights, use `Mdl.TransformWeights`.
- To transform `X` to the new set of features by using the learned transformation, pass `Mdl` and `X` to `transform`.

`Mdl = rica(X,q,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, you can standardize the predictor data or specify the value of the penalty coefficient in the reconstruction term of the objective function.

Examples

Create Reconstruction ICA Object

Create a `ReconstructionICA` object by using the `rica` function.

Load the `SampleImagePatches` image patches.

```
data = load('SampleImagePatches');
size(data.X)
```

```
ans = 1×2

        5000         363
```

There are 5,000 image patches, each containing 363 features.

Extract 100 features from the data.

```
rng default % For reproducibility
q = 100;
Mdl = rica(data.X,q,'IterationLimit',100)
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

```
Mdl =
  ReconstructionICA
            ModelParameters: [1x1 struct]
              NumPredictors: 363
          NumLearnedFeatures: 100
                         Mu: []
                      Sigma: []
                    FitInfo: [1x1 struct]
           TransformWeights: [363x100 double]
    InitialTransformWeights: []
    NonGaussianityIndicator: [100x1 double]


  Properties, Methods
```

`rica` issues a warning because it stopped due to reaching the iteration limit, instead of reaching a step-size limit or a gradient-size limit. You can still use the learned features in the returned object by calling the `transform` function.

Input Arguments

X — Predictor data

numeric matrix | table

Predictor data, specified as an *n*-by-*p* numeric matrix or table. Rows correspond to individual observations and columns correspond to individual predictor variables. If `X` is a table, then all of its variables must be numeric vectors.

Data Types: `single` | `double` | `table`

q — Number of features to extract

positive integer

Number of features to extract from the predictor data, specified as a positive integer.

`rica` stores a *p*-by-`q` transform weight matrix in `Mdl.TransformWeights`. Therefore, setting very large values for `q` can result in greater memory consumption and increased computation time.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `Mdl = rica(X,q,'IterationLimit',200,'Standardize',true)` runs `rica` with optimization iterations limited to 200 and standardized predictor data.

IterationLimit — Maximum number of iterations

`1000` (default) | positive integer

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

Example: `'IterationLimit',1e6`

Data Types: `single` | `double`

`VerbosityLevel` — Verbosity level

`0` (default) | nonnegative integer

Verbosity level for monitoring algorithm convergence, specified as the comma-separated pair consisting of `'VerbosityLevel'` and a value in this table.

Value	Description
`0`	`rica` does not display convergence information at the command line.
Positive integer	`rica` displays convergence information at the command line.

Convergence Information

Heading	Meaning
`FUN VALUE`	Objective function value.
`NORM GRAD`	Norm of the gradient of the objective function.
`NORM STEP`	Norm of the iterative step, meaning the distance between the previous point and the current point.
`CURV`	`OK` means the weak Wolfe condition is satisfied. This condition is a combination of sufficient decrease of the objective function and a curvature condition.
`GAMMA`	Inner product of the step times the gradient difference, divided by the inner product of the gradient difference with itself. The gradient difference is the gradient at the current point minus the gradient at the previous point. Gives diagnostic information on the objective function curvature.
`ALPHA`	Step direction multiplier, which differs from `1` when the algorithm performed a line search.
`ACCEPT`	`YES` means the algorithm found an acceptable step to take.

Example: `'VerbosityLevel',1`

Data Types: `single` | `double`

`Lambda` — Regularization coefficient value

`1` (default) | positive numeric scalar

Regularization coefficient value for the transform weight matrix, specified as the comma-separated pair consisting of `'Lambda'` and a positive numeric scalar. If you specify `0`, then there is no regularization term in the objective function.

Example: `'Lambda',0.1`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data

`false` (default) | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and `true` (`1`) or `false` (`0`).

If `Standardize` is `true`, then:

- `rica` centers and scales each column of the predictor data (`X`) by the column mean and standard deviation, respectively.
- `rica` extracts new features by using the standardized predictor matrix, and stores the predictor variable means and standard deviations in properties `Mu` and `Sigma` of `Mdl`.

Example: `'Standardize',true`

Data Types: `logical`

`ContrastFcn` — Contrast function
`'logcosh'` (default) | `'exp'` | `'sqrt'`

Contrast function, specified as `'logcosh'`, `'exp'`, or `'sqrt'`. The contrast function is a smooth function that is similar to an absolute value function. The `rica` objective function contains a term

$$\sum_{j=1}^{q}\frac{1}{n}\sum_{i=1}^{n}g\left(w_j^T\tilde{x}_i\right),$$

where g represents the contrast function, the w_j are the variables over which the optimization takes place, and the $\tilde{x}_i$ are data.

The three available contrast functions are:

- `'logcosh'` — $g=\frac{1}{2}\log(\cosh(2x))$
- `'exp'` — $g=-\exp\left(-\frac{x^2}{2}\right)$
- `'sqrt'` — $g=\sqrt{x^2+10^{-8}}$

Example: `'ContrastFcn','exp'`

`InitialTransformWeights` — Transformation weights that initialize optimization
`randn(p,q)` (default) | numeric matrix

Transformation weights that initialize optimization, specified as the comma-separated pair consisting of `'InitialTransformWeights'` and a *p*-by-`q` numeric matrix. *p* must be the number of columns or variables in `X` and `q` is the value of `q`.

Tip You can continue optimizing a previously returned transform weight matrix by passing it as an initial value in another call to `rica`. The output model object `Mdl` stores a learned transform weight matrix in the `TransformWeights` property.

Example: `'InitialTransformWeights',Mdl.TransformWeights`

Data Types: `single` | `double`

`NonGaussianityIndicator` — Non-Gaussianity of sources
`ones(q,1)` (default) | length-`q` vector of ±1

Non-Gaussianity of sources, specified as a length-`q` vector of ±1.

- `NonGaussianityIndicator(k) = 1` means `rica` models the `k`th source as super-Gaussian, with a sharp peak at 0.
- `NonGaussianityIndicator(k) = -1` means `rica` models the `k`th source as sub-Gaussian.

Data Types: `single` | `double`

`GradientTolerance` — Relative convergence tolerance on gradient norm
`1e-6` (default) | positive numeric scalar

Relative convergence tolerance on gradient norm, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive numeric scalar. This gradient is the gradient of the objective function.

Example: `'GradientTolerance',1e-4`

Data Types: `single` | `double`

`StepTolerance` — Absolute convergence tolerance on step size
`1e-6` (default) | positive numeric scalar

Absolute convergence tolerance on the step size, specified as the comma-separated pair consisting of `'StepTolerance'` and a positive numeric scalar.

Example: `'StepTolerance',1e-4`

Data Types: `single` | `double`

Output Arguments

`Mdl` — Learned reconstruction ICA model
`ReconstructionICA` model object

Learned reconstruction ICA model, returned as a `ReconstructionICA` model object.

To access properties of `Mdl`, use dot notation. For example:

- To access the learned transform weights, use `Mdl.TransformWeights`.
- To access the structure of fitting information, use `Mdl.FitInfo`.

Algorithms

The `rica` function creates a linear transformation of input features to output features. The transformation is based on optimizing a nonlinear objective function that roughly balances statistical independence of the output features versus the ability to reconstruct the input data using the output features.

For details, see “Reconstruction ICA Algorithm” on page 15-132.

See Also
`ReconstructionICA` | `sparsefilt` | `transform`

Topics
“Feature Extraction Workflow” on page 15-135
“Extract Mixed Signals” on page 15-164
“Feature Extraction” on page 15-130

Introduced in R2017a

ridge

Ridge regression

Syntax

```
B = ridge(y,X,k)
B = ridge(y,X,k,scaled)
```

Description

`B = ridge(y,X,k)` returns coefficient estimates for ridge regression models on page 33-5713 of the predictor data `X` and the response `y`. Each column of `B` corresponds to a particular ridge parameter `k`. By default, the function computes `B` after centering and scaling the predictors to have mean 0 and standard deviation 1. Because the model does not include a constant term, do not add a column of 1s to `X`.

`B = ridge(y,X,k,scaled)` specifies the scaling for the coefficient estimates in `B`. When `scaled` is `1` (default), `ridge` does not restore the coefficients to the original data scale. When `scaled` is `0`, `ridge` restores the coefficients to the scale of the original data. For more information, see “Coefficient Scaling” on page 33-5714.

Examples

Ridge Regression

Perform ridge regression for a range of ridge parameters and observe how the coefficient estimates change.

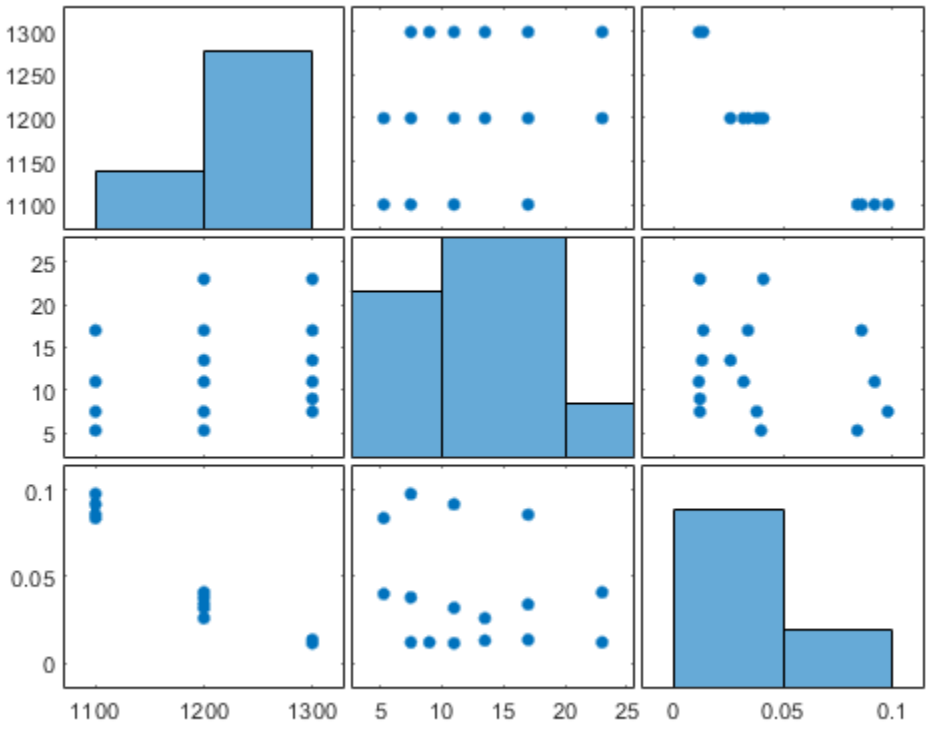
Load the `acetylene` data set.

```
load acetylene
```

`acetylene` contains observations for the predictor variables `x1`, `x2`, and `x3`, and the response variable `y`.

Plot the predictor variables against each other. Observe any correlation between the variables.

```
plotmatrix([x1 x2 x3])
```



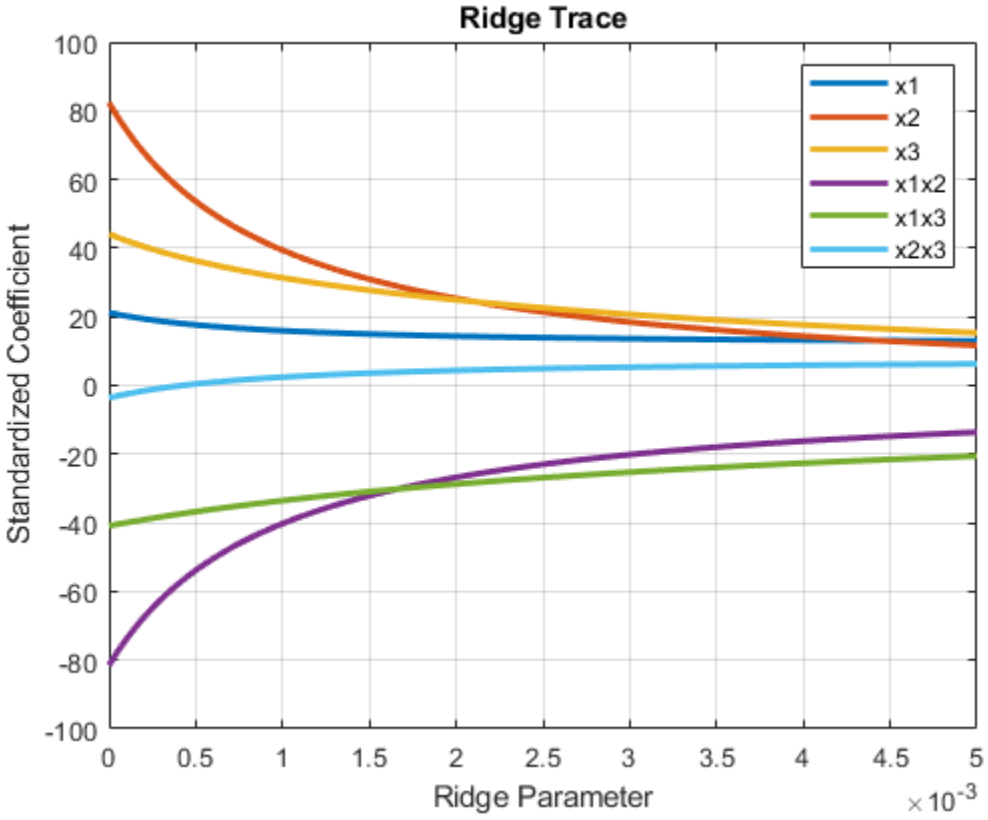
For example, note the linear correlation between `x1` and `x3`.

Compute coefficient estimates for a multilinear model with interaction terms, for a range of ridge parameters. Use `x2fx` to create interaction terms and `ridge` to perform ridge regression.

```
X = [x1 x2 x3];
D = x2fx(X,'interaction');
D(:,1) = []; % No constant term
k = 0:1e-5:5e-3;
B = ridge(y,D,k);
```

Plot the ridge trace.

```
figure
plot(k,B,'LineWidth',2)
ylim([-100 100])
grid on
xlabel('Ridge Parameter')
ylabel('Standardized Coefficient')
title('Ridge Trace')
legend('x1','x2','x3','x1x2','x1x3','x2x3')
```



The estimates stabilize to the right of the plot. Note that the coefficient of the `x2x3` interaction term changes sign at a value of the ridge parameter $\approx 5*10^{-4}$.

Predict Values Using Ridge Regression

Predict miles per gallon (MPG) values using ridge regression.

Load the `carbig` data set.

```
load carbig
X = [Acceleration Weight Displacement Horsepower];
y = MPG;
```

Split the data into training and test sets.

```
n = length(y);
rng('default') % For reproducibility
c = cvpartition(n,'HoldOut',0.3);
idxTrain = training(c,1);
idxTest = ~idxTrain;
```

Find the coefficients of a ridge regression model (with k = 5).

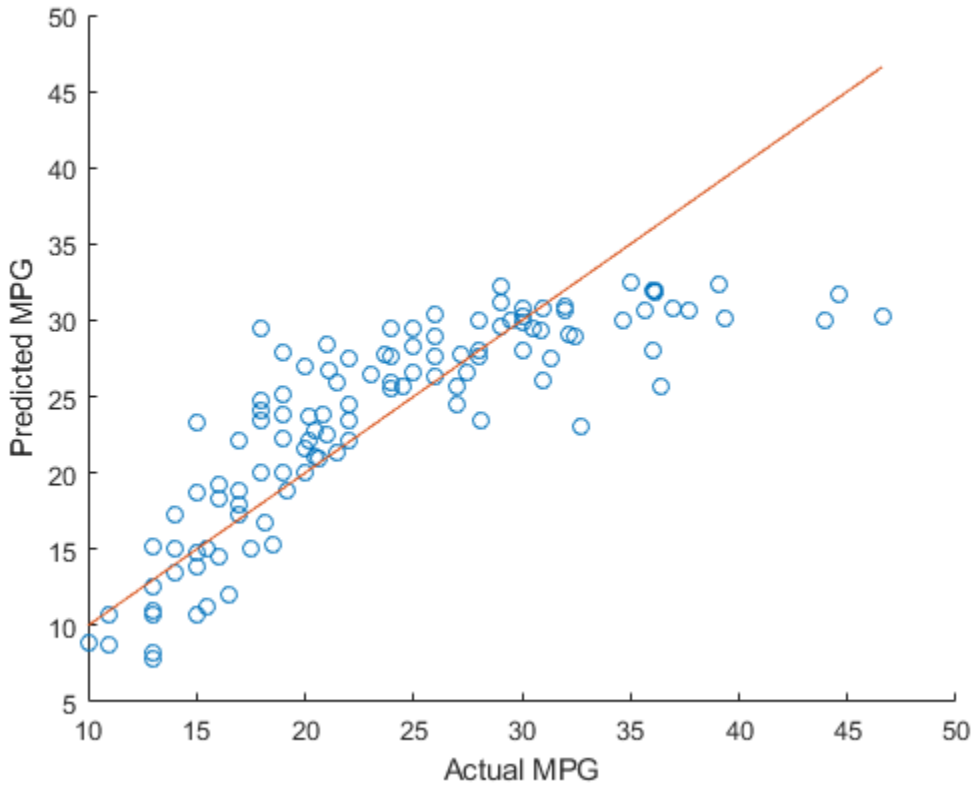
```
k = 5;
b = ridge(y(idxTrain),X(idxTrain,:),k,0);
```

Predict MPG values for the test data using the model.

```
yhat = b(1) + X(idxTest,:)*b(2:end);
```

Compare the predicted values to the actual miles per gallon (MPG) values using a reference line.

```
scatter(y(idxTest),yhat)
hold on
plot(y(idxTest),y(idxTest))
xlabel('Actual MPG')
ylabel('Predicted MPG')
hold off
```



Input Arguments

y — Response data

numeric vector

Response data, specified as an *n*-by-1 numeric vector, where *n* is the number of observations.

Data Types: `single` | `double`

X — Predictor data

numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix. The rows of X correspond to the *n* observations, and the columns of X correspond to the *p* predictors.

Data Types: `single` | `double`

k — Ridge parameters
numeric vector

Ridge parameters, specified as a numeric vector.

Example: `[0.2 0.3 0.4 0.5]`

Data Types: `single` | `double`

`scaled` — Scaling flag
`1` (default) | `0`

Scaling flag that determines whether the coefficient estimates in `B` are restored to the scale of the original data, specified as either `0` or `1`. If `scaled` is `0`, then `ridge` performs this additional transformation. In this case, `B` contains $p+1$ coefficients for each value of `k`, with the first row of `B` corresponding to a constant term in the model. If `scaled` is `1`, then the software omits the additional transformation, and `B` contains p coefficients without a constant term coefficient.

Output Arguments

`B` — Coefficient estimates
numeric matrix

Coefficient estimates, returned as a numeric matrix. The rows of `B` correspond to the predictors in `X`, and the columns of `B` correspond to the ridge parameters `k`.

If `scaled` is `1`, then `B` is a p-by-m matrix, where m is the number of elements in `k`. If `scaled` is `0`, then `B` is a $(p+1)$-by-m matrix.

More About

Ridge Regression

Ridge regression is a method for estimating coefficients of linear models that include linearly correlated predictors.

Coefficient estimates for multiple linear regression models rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the matrix $(X^TX)^{-1}$ is close to singular. Therefore, the least-squares estimate

$$\widehat{\beta} = (X^TX)^{-1}X^Ty$$

is highly sensitive to random errors in the observed response y, producing a large variance. This situation of multicollinearity can arise, for example, when you collect data without an experimental design.

Ridge regression addresses the problem of multicollinearity by estimating regression coefficients using

$$\widehat{\beta} = (X^TX + kI)^{-1}X^Ty$$

where k is the ridge parameter and I is the identity matrix. Small, positive values of k improve the conditioning of the problem and reduce the variance of the estimates. While biased, the reduced

variance of ridge estimates often results in a smaller mean squared error when compared to least-squares estimates.

Coefficient Scaling

The scaling of the coefficient estimates for the ridge regression models depends on the value of the `scaled` input argument.

Suppose the ridge parameter k is equal to 0. The coefficients returned by `ridge`, when `scaled` is equal to `1`, are estimates of the b_i^1 in the multilinear model

$$y - \mu_y = b_1^1 z_1 + \ldots + b_p^1 z_p + \varepsilon$$

where $z_i = (x_i - \mu_i)/\sigma_i$ are the centered and scaled predictors, $y - \mu_y$ is the centered response, and ε is an error term. You can rewrite the model as

$$y = b_0^0 + b_1^0 x_1 + \ldots + b_p^0 x_p + \varepsilon$$

with $b_0^0 = \mu_y - \sum_{i=1}^{p} \frac{b_i^1 \mu_i}{\sigma_i}$ and $b_i^0 = \frac{b_i^1}{\sigma_i}$. The b_i^0 terms correspond to the coefficients returned by `ridge` when `scaled` is equal to `0`.

More generally, for any value of `k`, if `B1 = ridge(y,X,k,1)`, then

```
m = mean(X);
s = std(X,0,1)';
B1_scaled = B1./s;
B0 = [mean(y)-m*B1_scaled; B1_scaled]
```

where `B0 = ridge(y,X,k,0)`.

Tips

- `ridge` treats `NaN` values in `X` or `y` as missing values. `ridge` omits observations with missing values from the ridge regression fit.
- In general, set `scaled` equal to `1` to produce plots where the coefficients are displayed on the same scale. See “Ridge Regression” on page 33-5709 for an example using a ridge trace plot, where the regression coefficients are displayed as a function of the ridge parameter. When making predictions, set `scaled` equal to `0`. For an example, see “Predict Values Using Ridge Regression” on page 33-5711.

Alternative Functionality

- Ridge, lasso, and elastic net regularization are all methods for estimating the coefficients of a linear model while penalizing large coefficients. The type of penalty depends on the method (see “More About” on page 33-3431 for more details). To perform lasso or elastic net regularization, use `lasso` instead.
- If you have high-dimensional full or sparse predictor data, you can use `fitrlinear` instead of `ridge`. When using `fitrlinear`, specify the `'Regularization','ridge'` name-value pair argument. Set the value of the `'Lambda'` name-value pair argument to a vector of the ridge parameters of your choice. `fitrlinear` returns a trained linear model `Mdl`. You can access the coefficient estimates stored in the `Beta` property of the model by using `Mdl.Beta`.

References

[1] Hoerl, A. E., and R. W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*. Vol. 12, No. 1, 1970, pp. 55–67.

[2] Hoerl, A. E., and R. W. Kennard. “Ridge Regression: Applications to Nonorthogonal Problems.” *Technometrics*. Vol. 12, No. 1, 1970, pp. 69–82.

[3] Marquardt, D. W. “Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation.” *Technometrics*. Vol. 12, No. 3, 1970, pp. 591–612.

[4] Marquardt, D. W., and R. D. Snee. “Ridge Regression in Practice.” *The American Statistician*. Vol. 29, No. 1, 1975, pp. 3–20.

See Also

`fitrlinear` | `lasso` | `regress` | `stepwise`

Introduced before R2006a

robustcov

Robust multivariate covariance and mean estimate

Syntax

```
sig = robustcov(x)
[sig,mu] = robustcov(x)
[sig,mu,mah] = robustcov(x)
[sig,mu,mah,outliers] = robustcov(x)
[sig,mu,mah,outliers,s] = robustcov(x)
[___] = robustcov(x,Name,Value)
```

Description

`sig = robustcov(x)` returns the robust covariance estimate `sig` of the multivariate data contained in `x`.

`[sig,mu] = robustcov(x)` also returns an estimate of the robust Minimum Covariance Determinant (MCD) mean, `mu`.

`[sig,mu,mah] = robustcov(x)` also returns the robust distances `mah`, computed as the Mahalanobis distances of the observations using the robust estimates of the mean and covariance.

`[sig,mu,mah,outliers] = robustcov(x)` also returns the indices of the observations retained as outliers in the sample data, `outliers`.

`[sig,mu,mah,outliers,s] = robustcov(x)` also returns a structure `s` that contains information about the estimate.

`[___] = robustcov(x,Name,Value)` returns any of the arguments shown in the previous syntaxes, using additional options specified by one or more `Name,Value` pair arguments. For example, you can specify which robust estimator to use or the start method to use for the attractors.

Examples

Detect Outliers Using Distance-Distance Plots

Use a Gaussian copula to generate random data points from a bivariate distribution.

```
rng default
rho = [1,0.05;0.05,1];
u = copularnd('Gaussian',rho,50);
```

Modify 5 randomly selected observations to be outliers.

```
noise = randperm(50,5);
u(noise,1) = u(noise,1)*5;
```

Calculate the robust covariance matrices using the three available methods: Fast-MCD, Orthogonalized Gnanadesikan-Kettenring (OGK), and Olive-Hawkins.

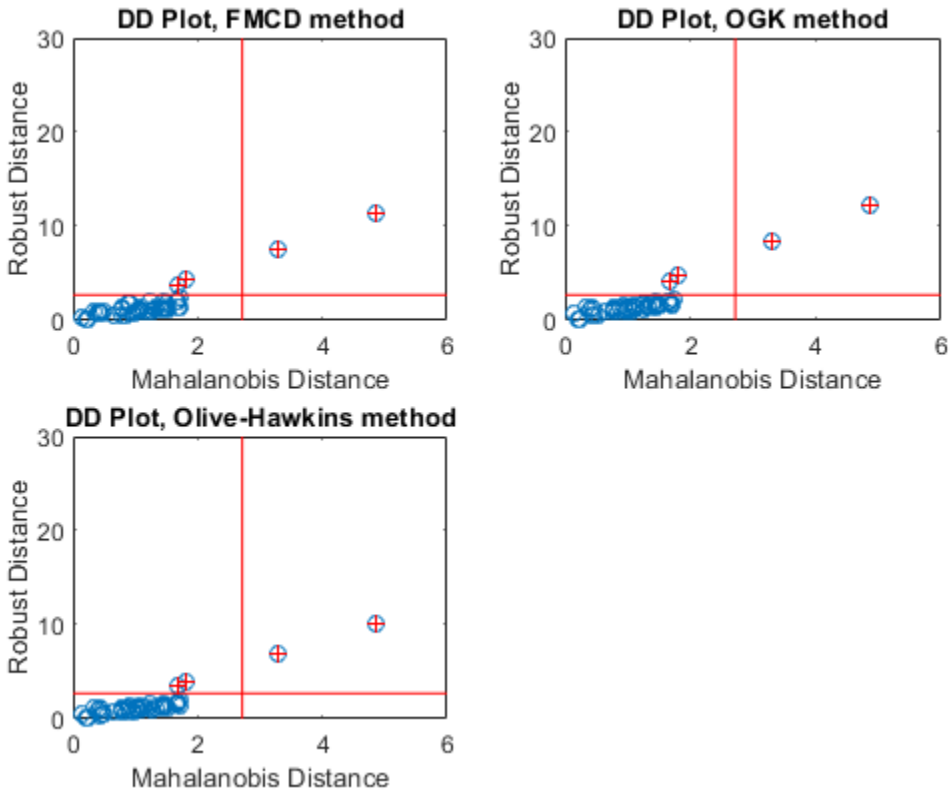
```
[Sfmcd, Mfmcd, dfmcd, Outfmcd] = robustcov(u);
[Sogk, Mogk, dogk, Outogk] = robustcov(u,'Method','ogk');
[Soh, Moh, doh, Outoh] = robustcov(u,'Method','olivehawkins');
```

Calculate the classical distance values for the sample data using the Mahalanobis measure.

```
d_classical = pdist2(u, mean(u),'mahal');
p = size(u,2);
chi2quantile = sqrt(chi2inv(0.975,p));
```

Create DD Plots for each robust covariance calculation method.

```
figure
subplot(2,2,1)
plot(d_classical, dfmcd, 'o')
line([chi2quantile, chi2quantile], [0, 30], 'color', 'r')
line([0, 6], [chi2quantile, chi2quantile], 'color', 'r')
hold on
plot(d_classical(Outfmcd), dfmcd(Outfmcd), 'r+')
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, FMCD method')
hold off

subplot(2,2,2)
plot(d_classical, dogk, 'o')
line([chi2quantile, chi2quantile], [0, 30], 'color', 'r')
line([0, 6], [chi2quantile, chi2quantile], 'color', 'r')
hold on
plot(d_classical(Outogk), dogk(Outogk), 'r+')
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, OGK method')
hold off

subplot(2,2,3)
plot(d_classical, doh, 'o')
line([chi2quantile, chi2quantile], [0, 30], 'color', 'r')
line([0, 6], [chi2quantile, chi2quantile], 'color', 'r')
hold on
plot(d_classical(Outoh), doh(Outoh), 'r+')
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, Olive-Hawkins method')
hold off
```



In a DD plot, the data points tend to cluster in a straight line that passes through the origin. Points that are far removed from this line are generally considered outliers. In each of the previous plots, the red '+' symbol indicates the data points that `robustcov` considers to be outliers.

Evaluate Data for Multivariate Normal Distribution

This example shows how to use `robustcov` to evaluate sample data for multivariate normal or other elliptically-contoured (EC) distributions.

Generate random sample data from a multivariate normal distribution. Calculate the Mahalanobis distances for the robust covariance estimates (using the Olive-Hawkins method) and the classical covariance estimates.

```
rng('default')
x1 = mvnrnd(zeros(1,3),eye(3),200);
[~, ~, d1] = robustcov(x1,'Method','olivehawkins');
d_classical1 = pdist2(x1,mean(x1),'mahalanobis');
```

Generate random sample data from an elliptically-contoured (EC) distribution. Calculate the Mahalanobis distances for the robust covariance estimates (using the Olive-Hawkins method) and the classical covariance estimates.

```
mu1 = [0 0 0];
sig1 = eye(3);
```

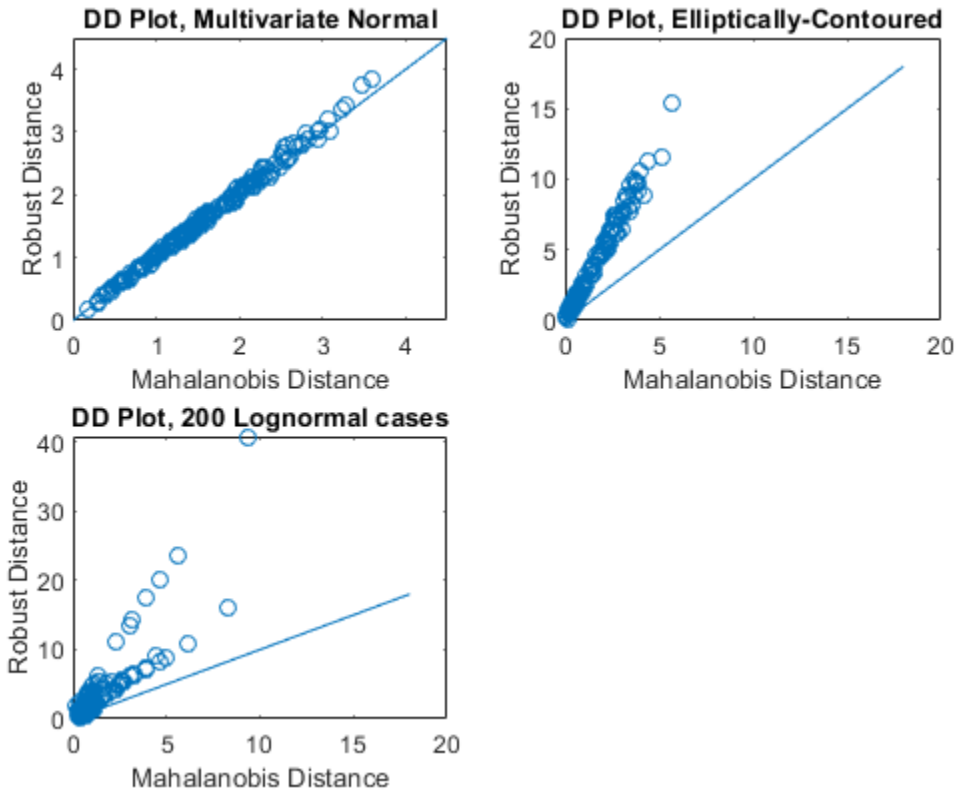
```
mu2 = [0 0 0];
sig2 = 25*eye(3);
x2 = [mvnrnd(mu1,sig1,120);mvnrnd(mu2,sig2,80)];
[~, ~, d2] = robustcov(x2, 'Method','olivehawkins');
d_classical2 = pdist2(x2, mean(x2), 'mahalanobis');
```

Generate random sample data from a multivariate lognormal distribution, which is neither multivariate normal or elliptically-contoured. Calculate the Mahalanobis distances for the robust covariance estimates (using the Olive-Hawkins method) and the classical covariance estimates.

```
x3 = exp(x1);
[~, ~, d3] = robustcov(x3, 'Method','olivehawkins');
d_classical3 = pdist2(x3, mean(x3), 'mahalanobis');
```

Create a D-D Plot for each of the three sets of sample data to compare.

```
figure
subplot(2,2,1)
plot(d_classical1,d1, 'o')
line([0 4.5], [0, 4.5])
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, Multivariate Normal')

subplot(2,2,2)
plot(d_classical2, d2, 'o')
line([0 18], [0, 18])
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, Elliptically-Contoured')

subplot(2,2,3)
plot(d_classical3, d3, 'o')
line([0 18], [0, 18])
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('DD Plot, 200 Lognormal cases')
```



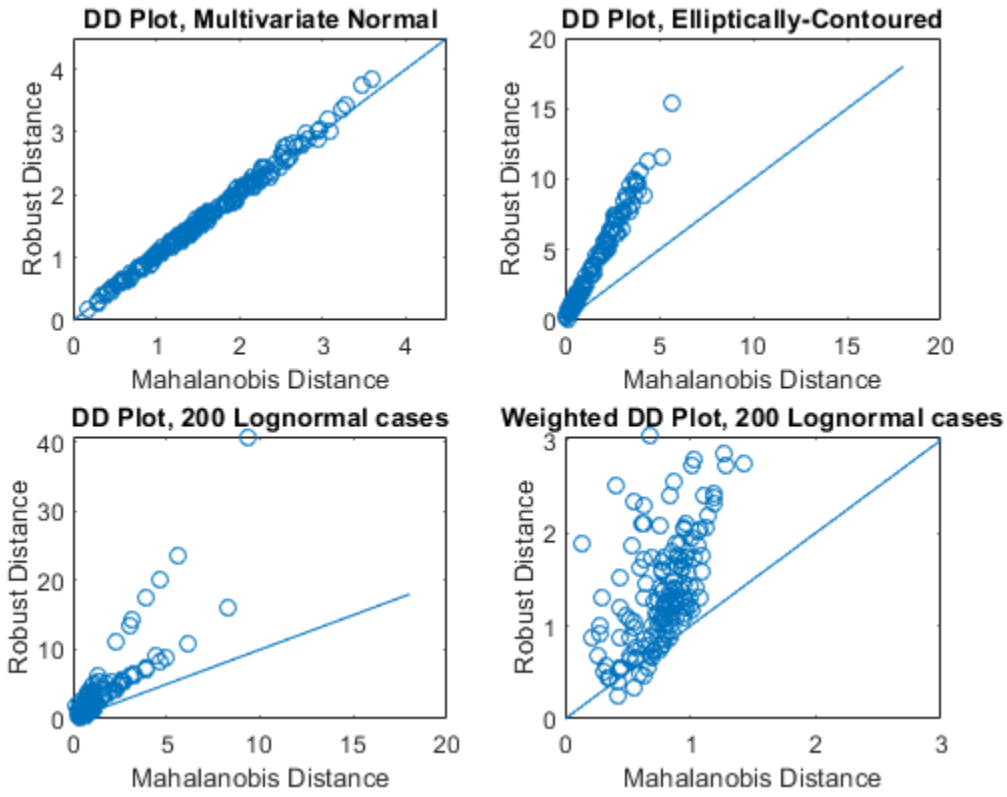
For data with a multivariate normal distribution (as shown in the upper left), the plotted points follow a straight, 45-degree line extending from the origin. For data with an elliptically-contoured distribution (as shown in the upper right), the plotted points follow a straight line, but are not at a 45-degree angle to the origin. For the lognormal distribution (as shown in the lower left), the plotted points do not follow a straight line.

It is difficult to identify any pattern in the lognormal distribution plot because most of the points are in the lower left of the plot. Use a weighted DD plot to magnify this corner and reveal features that are obscured when large robust distances exist.

```
d3_weighted = d3(d3 < sqrt(chi2inv(0.975,3)));
d_classical_weighted = d_classical3(d3 < sqrt(chi2inv(0.975,3)));
```

Add a fourth subplot to the figure to show the results of the weighting process on the lognormally distributed data.

```
subplot(2,2,4)
plot(d_classical_weighted, d3_weighted, 'o')
line([0 3], [0, 3])
xlabel('Mahalanobis Distance')
ylabel('Robust Distance')
title('Weighted DD Plot, 200 Lognormal cases')
```



The scale on this plot indicates that it represents a magnified view of the original DD plot for the lognormal data. This view more clearly shows the lack of pattern to the plot, which indicates that the data is neither multivariate normal nor elliptically contoured.

Compute Robust Covariance and Plot the Outliers

Use a Gaussian copula to generate random data points from a bivariate distribution.

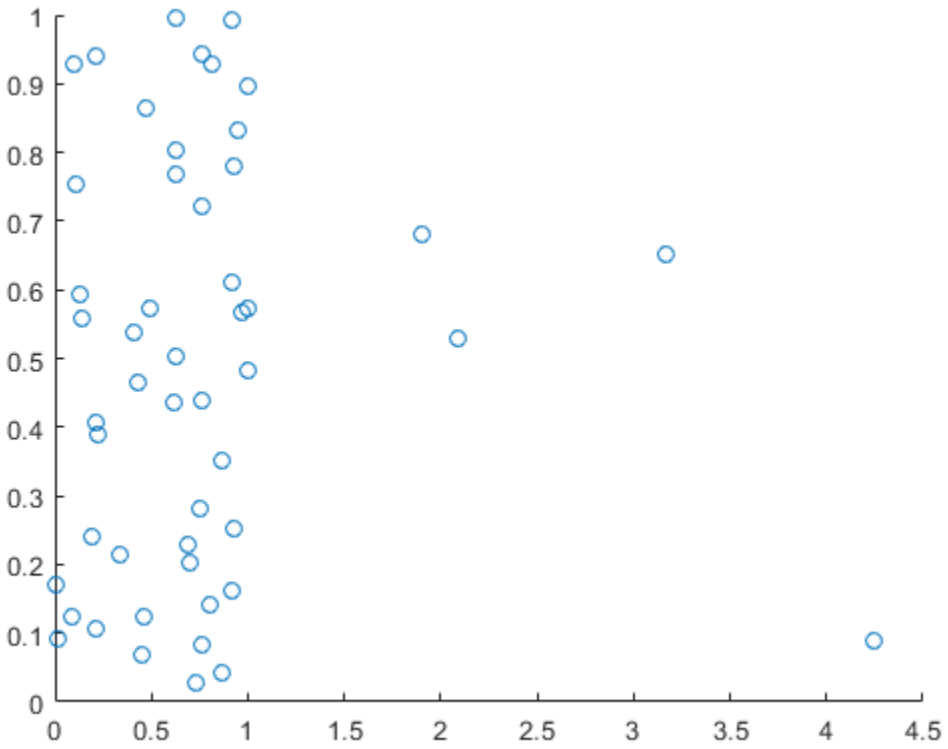
```
rng default
rho = [1,0.05;0.05,1];
u = copularnd('Gaussian',rho,50);
```

Modify 5 randomly selected observations to be outliers.

```
noise = randperm(50,5);
u(noise,1) = u(noise,1)*5;
```

Visualize the bivariate data using a scatter plot.

```
figure
scatter(u(:,1),u(:,2))
```



Most of the data points appear on the left side of the plot. However, some of the data points appear further to the right. These points are possible outliers that could affect the covariance matrix calculation.

Compare the classical and robust covariance matrices.

```
c = cov(u)
```

```
c = 2×2

    0.5523    0.0000
    0.0000    0.0913
```

```
rc = robustcov(u)
```

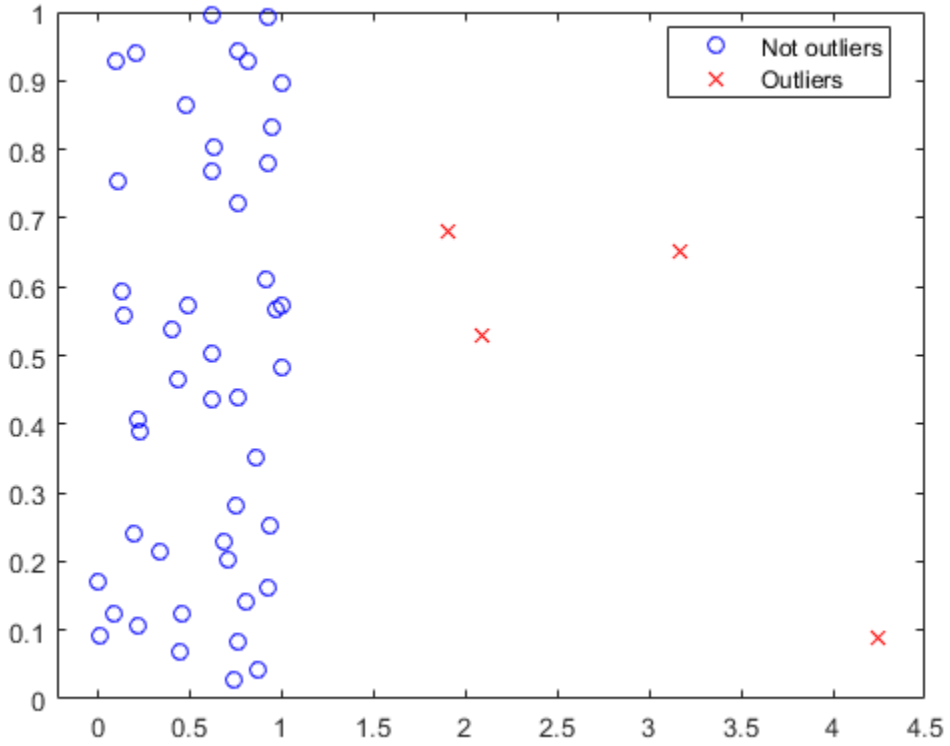
```
rc = 2×2

    0.1117    0.0364
    0.0364    0.1695
```

The classical and robust covariance matrices differ because the outliers present in the sample data influence the results.

Identify and plot the data points that `robustcov` considers outliers.

```
[sig,mu,mah,outliers] = robustcov(u);
figure
```

```
gscatter(u(:,1),u(:,2),outliers,'br','ox')
legend({'Not outliers','Outliers'})
```



`robustcov` identifies the data points on the right side of the plot as potential outliers, and treats them accordingly when calculating the robust covariance matrix.

Input Arguments

x — Sample data

matrix of numeric values

Sample data used to estimate the robust covariance matrix, specified as a matrix of numeric values. `x` is an *n*-by-*p* matrix where each row is an observation and each column is a variable.

`robustcov` removes any rows with missing predictor values when calculating the robust covariance matrix.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Method','ogk','NumOGKIterations',1` specifies the robust estimator as the Orthogonalized Gnanadesikan-Kettenring method and sets the number of orthogonalization iterations to 1.

For All Estimators

`Method` — Robust estimator

`'fmcd'` (default) | `'ogk'` | `'olivehawkins'`

Robust estimator, specified as one of the following.

Name	Value
`'fmcd'`	FAST-MCD (Minimum Covariance Determinant) on page 33-5727 method
`'ogk'`	Orthogonalized Gnanadesikan-Kettenring (OGK) on page 33-5728 estimate
`'olivehawkins'`	Concentration algorithm on page 33-5728 techniques, a family of fast, consistent and highly outlier-resistant methods

Example: `'Method','ogk'`

For the FMCD and OliveHawkins Methods Only

`OutlierFraction` — Outlier fraction

0.5 (default) | numeric value in the range [0,0.5]

Outlier fraction, specified as the comma-separated pair consisting of `'OutlierFraction'` and a numeric value in the range [0,0.5]. The value 1 – *OutlierFraction* specifies the fraction of observations over which to minimize the covariance determinant.

The algorithm chooses a subsample of size h = ceiling(n + p + 1) / 2), where n is the number of observations and p is the number of dimensions. `OutlierFraction` is the value for which the maximum possible breakdown is achieved, and controls the size of the subsets h over which the covariance determinant is minimized. The algorithm then chooses h to approximately equal (1 – *OutlierFraction*) × n observations per subset.

Example: `'OutlierFraction',0.25`

Data Types: `single` | `double`

`NumTrials` — Number of trials

positive integer value

Number of trials, specified as the comma-separated pair consisting of `'NumTrials'` and a positive integer value.

If `'Method'` is `'fmcd'`, then `NumTrials` is the number of random subsamples of size (p + 1) drawn from the sample data as starting points in the algorithm. p is the number of dimensions in the sample data. In this case, the default value for `NumTrials` is 500.

If `'Method'` is `'olivehawkins'`, then `NumTrials` is the number of trial fits, or attractors, to be used. In this case, the default value for `NumTrials` is 2. This option is only useful for non-deterministic starts.

Example: `'NumTrials',300`

Data Types: `single` | `double`

For the FMCD Method Only

`BiasCorrection` — Flag to apply small-sample correction factor

`1` (default) | `0`

Flag to apply small-sample correction factor, specified as the comma-separated pair consisting of `'BiasCorrection'` and either `1` or `0`. A `1` value indicates that `robustcov` corrects for bias in the covariance estimate for small samples. A `0` value indicates that `robustcov` does not apply this correction.

Example: `'BiasCorrection',0`

Data Types: `logical`

For the OGK Method Only

`NumOGKIterations` — Number of orthogonalization iterations

2 (default) | positive integer value

Number of orthogonalization iterations, specified as the comma-separated pair consisting of `'NumOGKIterations'` and a positive integer value. Generally, this value is set to 1 or 2, and further steps are unlikely to improve the estimation.

Example: `'NumIter',1`

Data Types: `single` | `double`

`UnivariateEstimator` — Function for computing univariate robust estimates

`'tauscale'` (default) | `'qn'`

Function for computing univariate robust estimates, specified as the comma-separated pair consisting of `'UnivariateEstimator'` and one of the following.

Name	Value
`'tauscale'`	Use the “tau-scale” estimate of Yohai and Zamar, which is a truncated standard deviation and a weighted mean.
`'qn'`	Use the Qn scale estimate of Croux and Rousseeuw.

Example: `'UnivariateEstimator','qn'`

For the OliveHawkins Method Only

`ReweightingMethod` — Method for reweighting

`'rfch'` (default) | `'rmvn'`

Method for reweighting in the efficiency step, specified as the comma-separated pair consisting of `'ReweightingMethod'` and one of the following.

Name	Value
'rfch'	Uses two reweighting steps. This is a standard method of reweighting to improve efficiency.
'rmvn'	Reweighted multivariate normal. Uses two reweighting steps that can be useful for estimating the true covariance matrix under a variety of outlier configurations when the clean data are multivariate normal.

Example: 'ReweightingMethod','rmvn'

NumConcentrationSteps — Number of concentration steps
10 (default) | positive integer value

Number of concentration steps, specified as the comma-separated pair consisting of 'NumConcentrationSteps' and a positive integer value.

Example: 'NumConcentrationSteps',8

Data Types: single | double

StartMethod — Start method for each attractor
'classical' (default) | 'medianball' | 'elemental' | function handle | cell array

Start method for each attractor, specified as the comma-separated pair consisting of 'Start' and one of the following.

Name	Value
'classical'	Use the classical estimator as the start. This is the DGK attractor which, used on its own, is known as the DGK estimator.
'medianball'	Use the Median Ball as the start. The Median Ball is (med(x),eye(p)). So 50% of cases furthest in Euclidean distance from the sample median are trimmed for computing the MB start. This is the MB attractor which, used on its own, is known as the MB estimator.
'elemental'	The attractor is generated by concentration where the start is a randomly selected elemental start: the classical estimator applied to a randomly selected “elemental set” of p + 1 cases. This “elemental” attractor is computationally efficient, but suffers from theoretical drawbacks, as it is inconsistent and zero breakdown.

By default, the attractor is chosen as follows: If one of the attractors is 'medianball', then any attractor whose location estimate has greater Euclidean distance from median(X) than half the data (in other words, is outside the median ball) is not used. Then the final attractor is chosen based on the MCD criterion.

You can also specify a function handle for a function that returns two output arguments used for computing the initial location and scatter estimates..

You can also specify a cell array containing any combination of the options given in the previous table and function handles. The number of attractors used is equal to the length of the cell array. This option allows more control over the algorithm and the ability to specify a custom number of attractors and starts.

Example: `'StartMethod','medianball'`

Output Arguments

`sig` — Robust covariance matrix estimates

numeric matrix

Robust covariance matrix estimates, returned as a *p*-by-*p* numeric matrix. *p* is the number of predictors contained in the sample data.

`mu` — Robust mean estimates

array of numeric values

Robust mean estimates, returned as a 1-by-*p* array of numeric values. *p* is the number of predictors contained in the sample data.

`mah` — Robust distances

array of numeric values

Robust distances, returned as a 1-by-*n* array of numeric values. `robustcov` removes any rows of `x` that contain missing data, so the number of rows of `mah` might be smaller than the number of rows in `x`.

`outliers` — Indices of outliers

array of logical values

Indices of observations retained as outliers in the sample data `x`, returned as a 1-by-*n* array of logical values. A `0` value indicates that the observation is not an outlier. A `1` value indicates that the observation is an outlier.

`robustcov` removes any rows of `x` that contain missing data, so the number of rows of `outliers` might be smaller than the number of rows in `x`.

`s` — Structure containing estimate information

structure

Structure containing estimate information, returned as a structure.

Algorithms

Minimum Covariance Determinant Estimate

Minimum covariance determinant (MCD) is the fastest estimator of multivariate location and scatter that is both consistent and robust. However, an exact evaluation of the MCD is impractical because it is computationally expensive to evaluate all possible subsets of the sample data. `robustcov` uses the FAST-MCD method to implement MCD [3]

The FAST-MCD method selects h observations out of n (where $n/2 < h \le n$) whose classical covariance matrix has the lowest possible determinant. The MCD mean is the mean of the h selected observations.

The MCD covariance is the covariance matrix of the h selected points, multiplied by a consistency factor to obtain consistency at the multivariate normal distribution, and by a correction factor to correct for bias at small sample sizes.

Orthogonalized Gnanadesikan-Kettenring Estimate

Orthogonalized Gnanadesikan-Kettenring (OGK) estimate is a positive definite estimate of the scatter starting from the Gnanadesikan and Kettering (GK) estimator, a pairwise robust scatter matrix that may be non-positive definite [1]. The estimate uses a form of principal components called an orthogonalization iteration on the pairwise scatter matrix, replacing its eigenvalues, which could be negative, with robust variances. This procedure can be iterated for improved results, and convergence is usually obtained after 2 or 3 iterations.

Olive Hawkins Estimate

The Olive-Hawkins estimate uses the “concentration algorithm” techniques proposed by Olive and Hawkins. This is a family of fast, consistent, and highly outlier-resistant methods. The estimate is a robust root n-consistent estimator of covariance for elliptically contoured distributions with fourth moments. This estimate is obtained by first generating trial estimates, or starts, and then using the concentration technique from each trial fit to obtain attractors.

Suppose (T_{0j},C_{0j}) is a start, then at the next iteration the classical mean and covariance estimators are computed from the approximately n / 2 cases (where n is the number of observations) with the smallest Mahalanobis distances based on the estimates from the previous iteration. This iteration can be continued for a fixed number of steps k, with the estimate at the last step, k, being the attractor. The final estimate is chosen based on a given criterion.

By default, two attractors are used. The first attractor is the Devlin-Gnanadesikan-Kettering (DGK) attractor, where the start used is the classical estimator. The second attractor is the Median Ball (MB) attractor, where the start used is `(median(x),eye(p))`, in other words the half set of data closest to `median(x)` in Euclidean distance. The MB attractor is used if the location estimator of the DGK attractor is outside of the median ball, and the attractor with the smallest determinant is used otherwise. The final mean estimate is the mean estimate of the chosen attractor, and the final covariance estimate is the covariance estimate of the chosen attractor, multiplied by a scaling factor to make the estimate consistent at the normal distribution.

References

[1] Maronna, R. and Zamar, R.H.. “Robust estimates of location and dispersion for high dimensional datasets.” *Technometrics*, Vol. 50, 2002.

[2] Pison, S. Van Aelst and G. Willems. “Small Sample Corrections for LTS and MCD.” *Metrika*, Vol. 55, 2002.

[3] Rousseeuw, P.J. and Van Driessen, K. “A fast algorithm for the minimum covariance determinant estimator.” *Technometrics*, Vol. 41, 1999.

[4] Olive, D.J. “A resistant estimator of multivariate location and dispersion.” *Computational Statistics and Data Analysis*, Vol. 46, pp. 99–102, 2004.

See Also

cov

Introduced in R2016a

robustdemo

Interactive robust regression

Syntax

```
robustdemo
robustdemo(x,y)
```

Description

`robustdemo` shows the difference between ordinary least squares and robust regression for data with a single predictor. With no input arguments, `robustdemo` displays a scatter plot of a sample of roughly linear data with one outlier. The bottom of the figure displays equations of lines fitted to the data using ordinary least squares and robust methods, together with estimates of the root mean squared errors.

Use the right mouse button to click on a point and view its least-squares leverage and robust weight.

Use the left mouse button to click-and-drag a point. The displays will update.

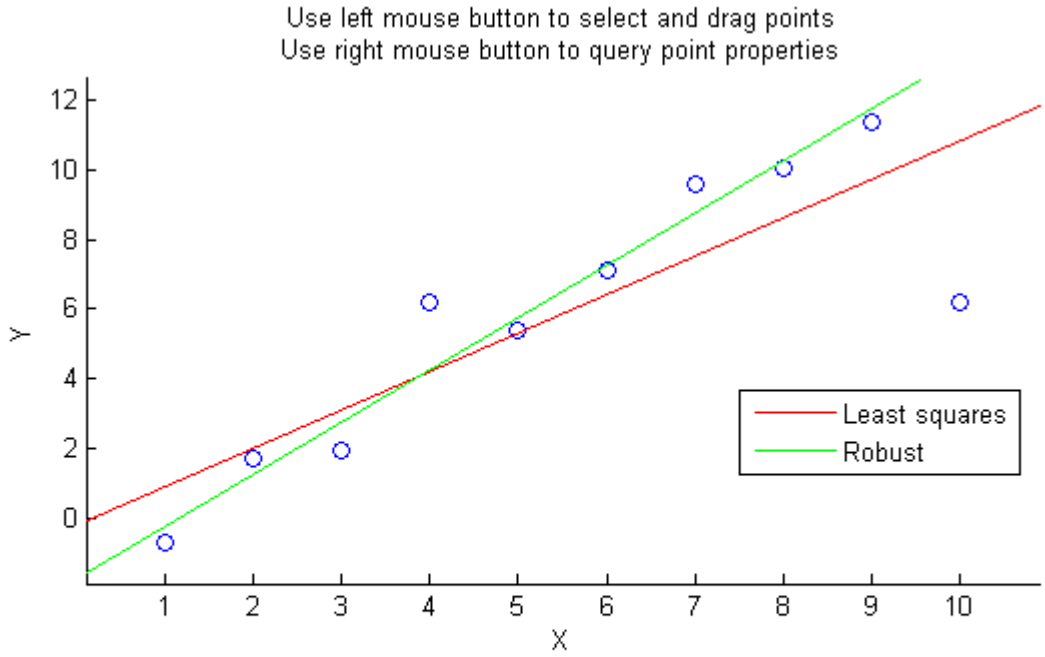
`robustdemo(x,y)` uses `x` and `y` data vectors you supply, in place of the sample data supplied with the function.

Examples

The following steps show you how to use `robustdemo`.

1 **Start the example.** To begin using `robustdemo` with the built-in data, simply type the function name:

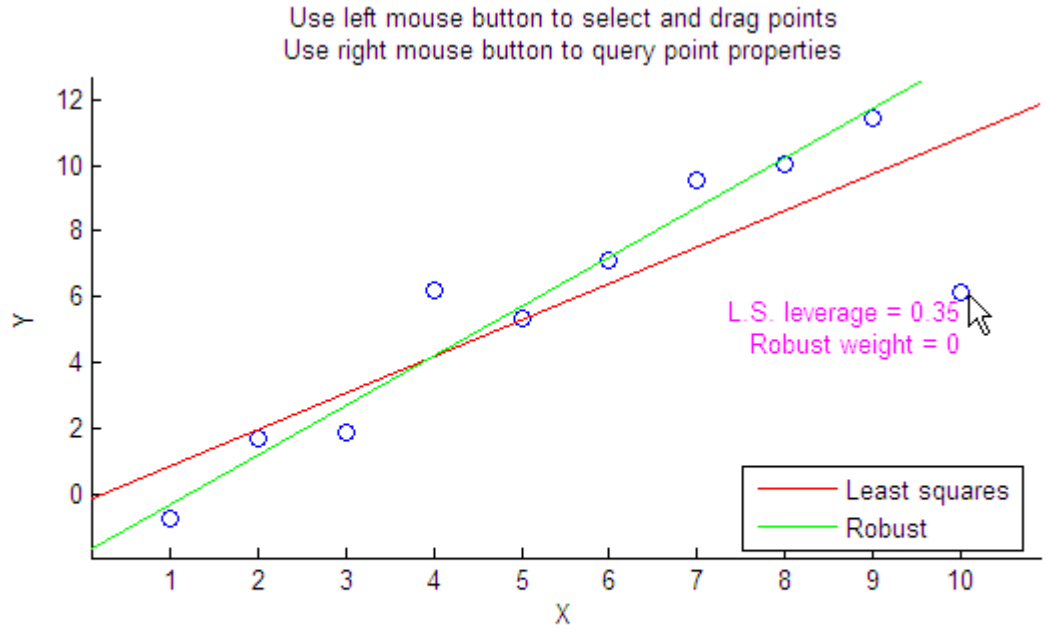
```
robustdemo
```



Least squares:	Y = -0.188327 + 1.10351*X	RMS error = 2.21375
Robust:	Y = -1.77278 + 1.50415*X	RMS error = 1.43663

The resulting figure shows a scatter plot with two fitted lines. The red line is the fit using ordinary least-squares regression. The green line is the fit using robust regression. At the bottom of the figure are the equations for the fitted lines, together with the estimated root mean squared errors for each fit.

2 **View leverages and robust weights.** Right-click on any data point to see its least-squares leverage and robust weight:

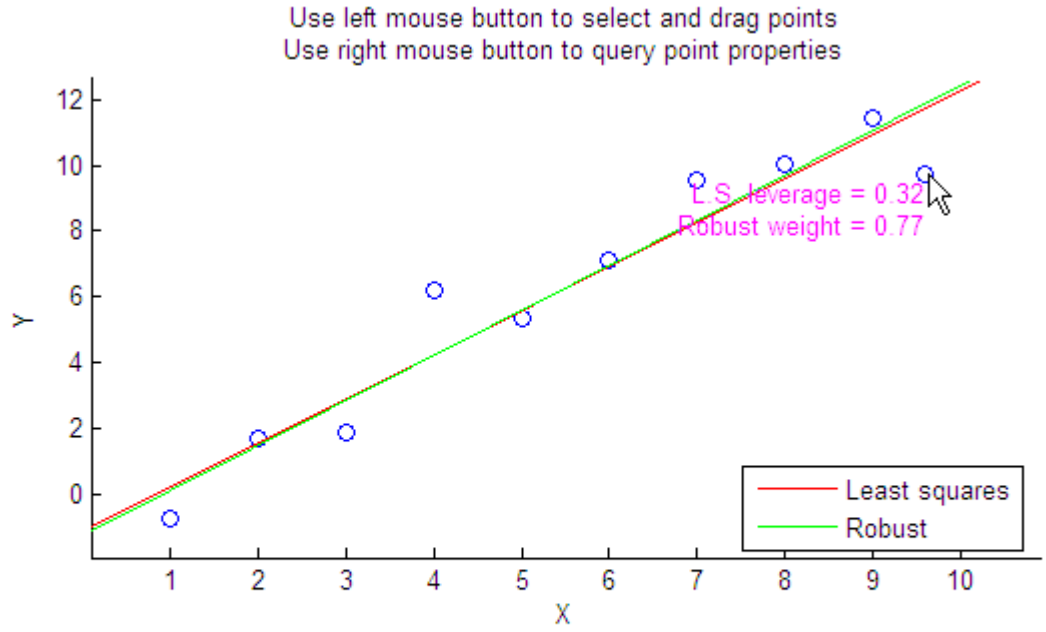


Least squares: Y = -0.188327 + 1.10351*X RMS error = 2.21375

Robust: Y = -1.77278 + 1.50415*X RMS error = 1.43663

In the built-in data, the right-most point has a relatively high leverage of 0.35. The point exerts a large influence on the least-squares fit, but its small robust weight shows that it is effectively excluded from the robust fit.

3 **See how changes in the data affect the fits.** With the left mouse button, click and hold on any data point and drag it to a new location. When you release the mouse button, the displays update:



Least squares:	Y = -1.0661 + 1.33785*X	RMS error = 1.21477
Robust:	Y = -1.18916 + 1.36459*X	RMS error = 1.27697

Bringing the right-most data point closer to the least-squares line makes the two fitted lines nearly identical. The adjusted right-most data point has significant weight in the robust fit.

See Also

`leverage` | `robustfit`

Introduced before R2006a

robustfit

Fit robust linear regression

Syntax

```
b = robustfit(X,y)
b = robustfit(X,y,wfun,tune,const)
[b,stats] = robustfit( ___ )
```

Description

`b = robustfit(X,y)` returns a vector `b` of coefficient estimates for a robust multiple linear regression of the responses in vector `y` on the predictors in matrix `X`.

`b = robustfit(X,y,wfun,tune,const)` specifies the fitting weight function options `wfun` and `tune`, and the indicator `const`, which determines if the model includes a constant term. You can pass in `[]` for `wfun`, `tune`, and `const` to use their default values.

`[b,stats] = robustfit( ___ )` also returns a structure `stats` containing estimated statistics, using any of the input argument combinations in previous syntaxes.

Examples

Estimate Robust Regression Coefficients

Estimate robust regression coefficients for a multiple linear model.

Load the `carsmall` data set. Specify car weight and horsepower as predictors and mileage per gallon as the response.

```
load carsmall
x1 = Weight;
x2 = Horsepower;
X = [x1 x2];
y = MPG;
```

Compute the robust regression coefficients.

```
b = robustfit(X,y)
```

```
b = 3×1

   47.1975
   -0.0068
   -0.0333
```

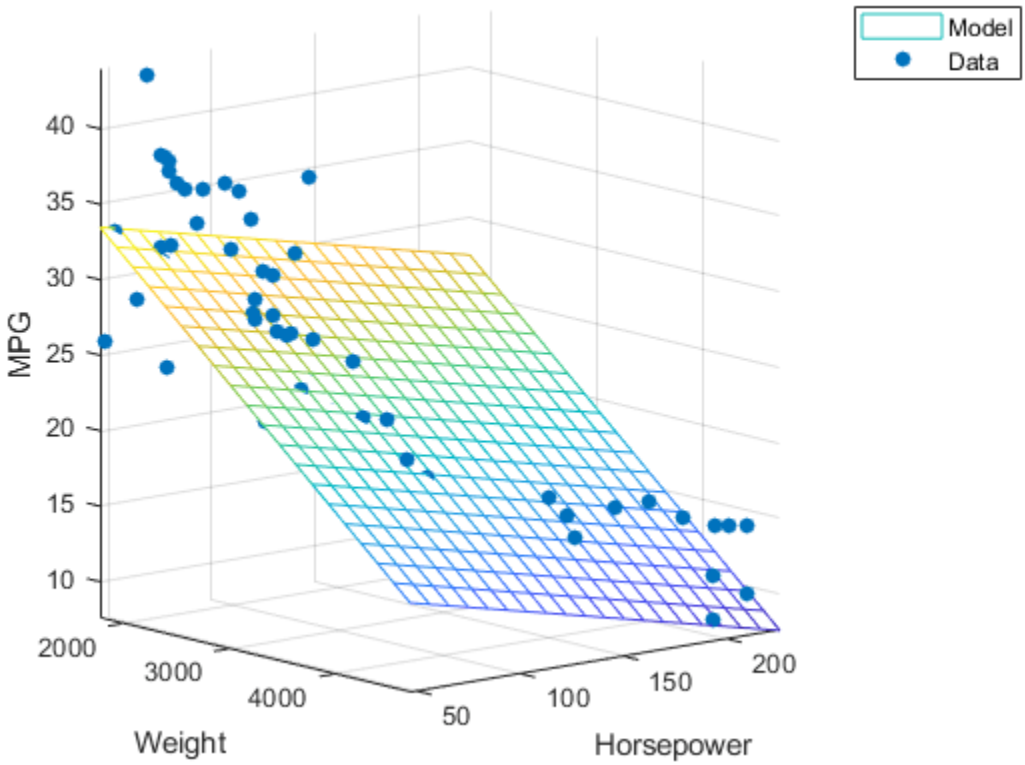
Plot the fitted model.

```
x1fit = linspace(min(x1),max(x1),20);
x2fit = linspace(min(x2),max(x2),20);
```

```
[X1FIT,X2FIT] = meshgrid(x1fit,x2fit);
YFIT = b(1) + b(2)*X1FIT + b(3)*X2FIT;
mesh(X1FIT,X2FIT,YFIT)
```

Plot the data.

```
hold on
scatter3(x1,x2,y,'filled')
hold off
xlabel('Weight')
ylabel('Horsepower')
zlabel('MPG')
legend('Model','Data')
view(50,10)
axis tight
```



Tune Robust Weight Function

Tune the weight function for robust regression by using different tuning constants.

Generate data with the trend $y = 10 - 2x$, and then change one value to simulate an outlier.

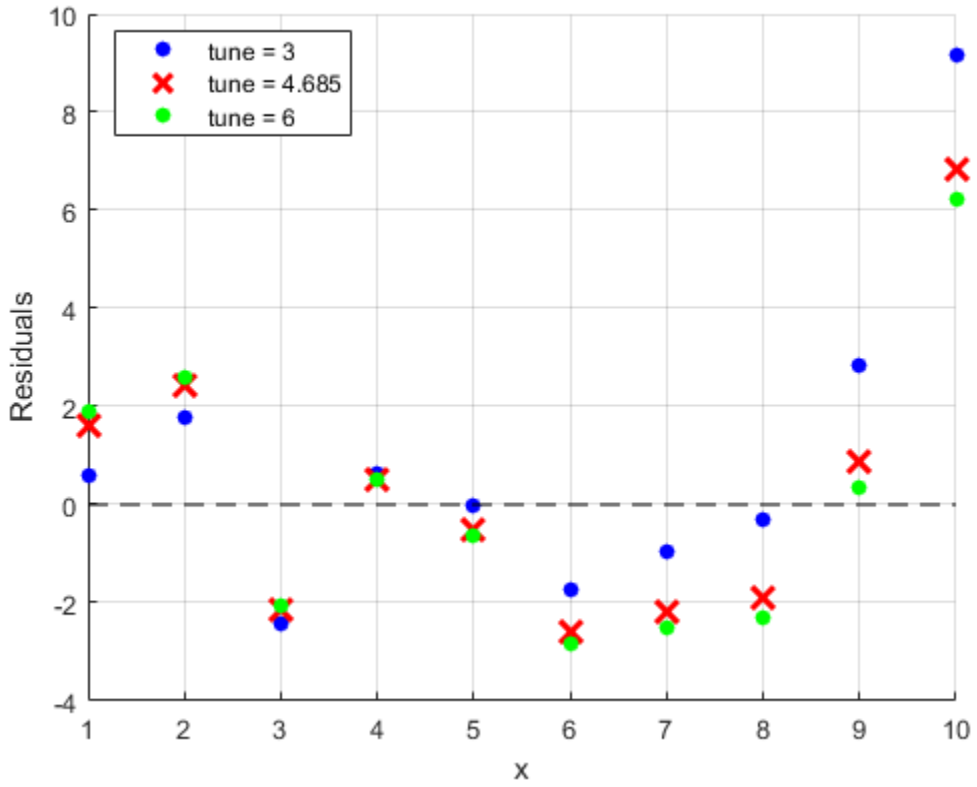
```
x = (1:10)';
rng ('default') % For reproducibility
y = 10 - 2*x + randn(10,1);
y(10) = 0;
```

Compute the robust regression residuals using the bisquare weight function for three different tuning constants. The default tuning constant is 4.685.

```
tune_const = [3 4.685 6];

for i = 1:length(tune_const)
    [~,stats] = robustfit(x,y,'bisquare',tune_const(i));
    resids(:,i) = stats.resid;
end
```

Create a plot of the residuals.

```
scatter(x,resids(:,1),'b','filled')
hold on
plot(resids(:,2),'rx','MarkerSize',10,'LineWidth',2)
scatter(x,resids(:,3),'g','filled')
plot([min(x) max(x)],[0 0],'--k')
hold off
grid on
xlabel('x')
ylabel('Residuals')
legend('tune = 3','tune = 4.685','tune = 6','Location','best')
```



Compute the root mean squared error (RMSE) of residuals for the three different tuning constants.

```
rmse = sqrt(mean(resids.^2))
```

```
rmse = 1×3
```

```
    3.2577    2.7576    2.7099
```

Because increasing the tuning constant decreases the downweight assigned to outliers, the RMSE decreases as the tuning constant increases.

Compare Robust and Least-Squares Regression

Generate data with the trend $y = 10 - 2x$, and then change one value to simulate an outlier.

```
x = (1:10)';
rng('default') % For reproducibility
y = 10 - 2*x + randn(10,1);
y(10) = 0;
```

Fit a straight line using ordinary least-squares regression. To compute coefficient estimates for a model with a constant term, include a column of ones in x.

```
bls = regress(y,[ones(10,1) x])
```

```
bls = 2×1

    7.8518
   -1.3644
```

Estimate a straight-line fit using robust regression. `robustfit` adds a constant term to the model by default.

```
[brob,stats] = robustfit(x,y);
brob
```

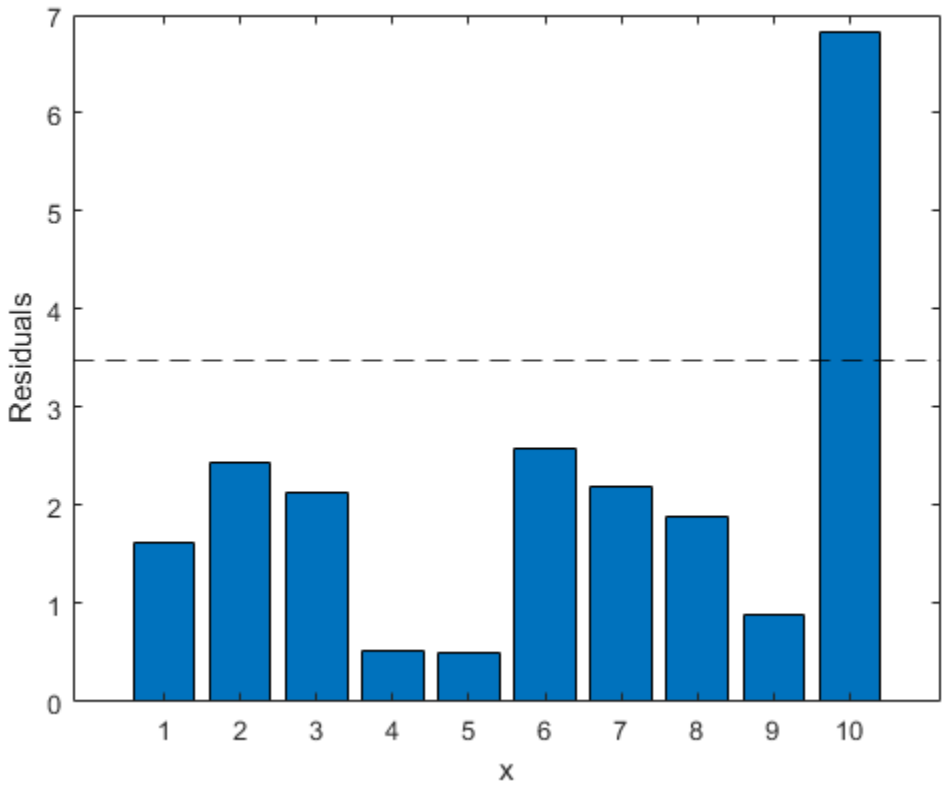
```
brob = 2×1

    8.4504
   -1.5278
```

Identify potential outliers by comparing the residuals to the median absolute deviation of the residuals.

```
outliers_ind = find(abs(stats.resid)>stats.mad_s);
```

Plot a bar graph of the residuals for robust regression.

```
bar(abs(stats.resid))
hold on
yline(stats.mad_s,'k--')
hold off
xlabel('x')
ylabel('Residuals')
```



Create a scatter plot of the data.

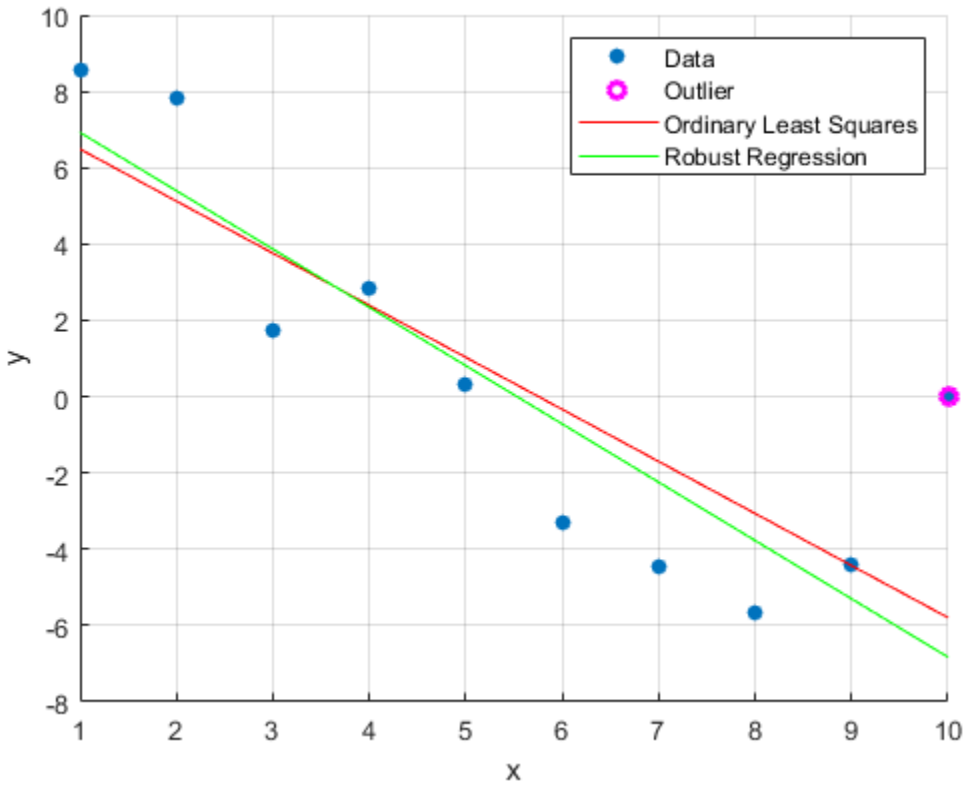
```
scatter(x,y,'filled')
```

Plot the outlier.

```
hold on
plot(x(outliers_ind),y(outliers_ind),'mo','LineWidth',2)
```

Plot the least-squares and robust fit.

```
plot(x,bls(1)+bls(2)*x,'r')
plot(x,brob(1)+brob(2)*x,'g')
hold off
xlabel('x')
ylabel('y')
legend('Data','Outlier','Ordinary Least Squares','Robust Regression')
grid on
```



The outlier influences the robust fit less than the least-squares fit.

Input Arguments

X — Predictor data

numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix. Rows of X correspond to observations, and columns correspond to predictor variables. X must have the same number of rows as y.

By default, `robustfit` adds a constant term to the model, unless you explicitly remove it by specifying `const` as `'off'`. So, do not include a column of 1s in X.

Data Types: `single` | `double`

y — Response data

numeric vector

Response data, specified as an *n*-by-1 numeric vector. Rows of y correspond to different observations. y must have the same number of rows as X.

Data Types: `single` | `double`

wfun — Robust fitting weight function

`'bisquare'` (default) | `'andrews'` | `'cauchy'` | `'fair'` | function handle | ...

Robust fitting weight function, specified as the name of a weight function described in the following table, or a function handle. `robustfit` uses the corresponding default tuning constant, unless otherwise specified by `tune`.

Weight Function	Description	Default Tuning Constant
`'andrews'`	`w = (abs(r)<pi) .* sin(r) ./ r`	1.339
`'bisquare'`	`w = (abs(r)<1) .* (1 - r.^2).^2` (also called biweight)	4.685
`'cauchy'`	`w = 1 ./ (1 + r.^2)`	2.385
`'fair'`	`w = 1 ./ (1 + abs(r))`	1.400
`'huber'`	`w = 1 ./ max(1, abs(r))`	1.345
`'logistic'`	`w = tanh(r) ./ r`	1.205
`'ols'`	Ordinary least squares (no weighting function)	None
`'talwar'`	`w = 1 * (abs(r)<1)`	2.795
`'welsch'`	`w = exp(-(r.^2))`	2.985
function handle	Custom weight function that accepts a vector `r` of scaled residuals, and returns a vector of weights the same size as `r`	1

The value *r* in the weight functions is

`r = resid/(tune*s*sqrt(1–h))`,

where

- `resid` is the vector of residuals from the previous iteration.
- `tune` is the tuning constant.
- `h` is the vector of leverage values from a least-squares fit.
- `s` is an estimate of the standard deviation of the error term given by `s = MAD/0.6745`.

`MAD` is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution. If `X` has *p* columns, the software excludes the smallest *p* absolute deviations when computing the median.

Data Types: `char` | `string` | `function handle`

`tune` — Tuning constant
positive scalar

Tuning constant, specified as a positive scalar. If you do not set `tune`, `robustfit` uses the corresponding default tuning constant for each weight function (see the table in `wfun`).

The default tuning constants of built-in weight functions give coefficient estimates that are approximately 95% as statistically efficient as the ordinary least-squares estimates, provided that the response has a normal distribution with no outliers. Decreasing the tuning constant increases the downweight assigned to large residuals; increasing the tuning constant decreases the downweight assigned to large residuals.

Data Types: `single` | `double`

`const` — Indicator for constant term

`'on'` (default) | `'off'`

Indicator for a constant term in the fit, specified as `'on'` or `'off'`. If `const` is `'on'`, then `robustfit` adds a first column of 1s to the predictor matrix `X`, and the output `b` becomes a (p + 1)-by-1 vector. If `const` is `'off'`, then `X` remains unchanged and `b` is a p-by-1 vector.

Data Types: `char` | `string`

Output Arguments

`b` — Coefficient estimates for robust multiple linear regression

numeric vector

Coefficient estimates for robust multiple linear regression, returned as a numeric vector. `b` is a p-by-1 vector, where p is the number of predictors in `X`.

By default, `robustfit` adds a constant term to the model, unless you explicitly remove it by specifying `const` as `'off'`.

`stats` — Model statistics

structure

Model statistics, returned as a structure. The following table describes the fields of the diagnostic statistics structure from the robust regression.

Field	Description
`ols_s`	Sigma estimate (root mean squared error) from ordinary least squares
`robust_s`	Robust estimate of sigma
`mad_s`	Estimate of sigma computed using the median absolute deviation of the residuals from their median; used for scaling residuals during iterative fitting
`s`	Final estimate of sigma, the largest between `robust_s` and a weighted average of `ols_s` and `robust_s`
`resid`	Residuals, observed minus fitted values (see “Raw Residuals” on page 11-80)
`rstud`	Studentized residuals, the residuals divided by an independent estimate of the residual standard deviation (see “Studentized Residuals” on page 11-80)
`se`	Standard error of the estimated coefficient value `b`
`covb`	Estimated covariance matrix for coefficient estimates
`coeffcorr`	Estimated correlation of coefficient estimates
`t`	t-statistic for each coefficient to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. Note that `t = b/se`.
`p`	p-values for the t-statistic of the hypothesis test that the corresponding coefficient is equal to zero or not
`w`	Vector of weights for a robust fit
`R`	R factor in the QR decomposition of `X`

Field	Description
dfe	Degrees of freedom for the error (residuals), equal to the number of observations minus the number of estimated coefficients
h	Vector of leverage values for a least-squares fit

More About

Leverage

Leverage is a measure of the effect of a particular observation on the regression predictions due to the position of that observation in the space of the inputs.

The leverage of observation i is the value of the ith diagonal term h_{ii} of the hat matrix H. The hat matrix H is defined in terms of the data matrix X:

$$H = X(X^TX)^{-1}X^T.$$

The hat matrix is also known as the *projection matrix* because it projects the vector of observations y onto the vector of predictions $\widehat{y}$, thus putting the "hat" on y.

Because the sum of the leverage values is p (the number of coefficients in the regression model), an observation i can be considered an outlier if its leverage substantially exceeds p/n, where n is the number of observations.

For more details, see “Hat Matrix and Leverage” on page 11-77.

Tips

- `robustfit` treats `NaN` values in `X` or `y` as missing values. `robustfit` omits observations with missing values from the robust fit.

Algorithms

- `robustfit` uses iteratively reweighted least squares to compute the coefficients `b`. The input `wfun` specifies the weights.
- `robustfit` estimates the variance-covariance matrix of the coefficient estimates `stats.covb` using the formula `inv(X'*X)*stats.s^2`. This estimate produces the standard error `stats.se` and correlation `stats.coeffcorr`.
- In a linear model, observed values of `y` and their residuals are random variables. Residuals have normal distributions with zero mean but with different variances at different values of the predictors. To put residuals on a comparable scale, `robustfit` “Studentizes” the residuals. That is, `robustfit` divides the residuals by an estimate of their standard deviation that is independent of their value. Studentized residuals have t-distributions with known degrees of freedom. `robustfit` returns the Studentized residuals in `stats.rstud`.

Alternative Functionality

`robustfit` is useful when you simply need the output arguments of the function or when you want to repeat fitting a model multiple times in a loop. If you need to investigate a robust fitted regression model further, create a linear regression model object `LinearModel` by using `fitlm`. Set the value for the name-value pair argument `'RobustOpts'` to `'on'`.

References

[1] DuMouchel, W. H., and F. L. O'Brien. “Integrating a Robust Option into a Multiple Regression Computing Environment.” *Computer Science and Statistics*: *Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[2] Holland, P. W., and R. E. Welsch. “Robust Regression Using Iteratively Reweighted Least-Squares.” *Communications in Statistics: Theory and Methods*, *A6*, 1977, pp. 813–827.

[3] Huber, P. J. *Robust Statistics*. Hoboken, NJ: John Wiley & Sons, Inc., 1981.

[4] Street, J. O., R. J. Carroll, and D. Ruppert. “A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares.” *The American Statistician*. Vol. 42, 1988, pp. 152–154.

See Also

`LinearModel` | `fitlm` | `regress` | `robustdemo`

Topics

“What Is a Linear Regression Model?” on page 11-6
“Reduce Outlier Effects Using Robust Regression” on page 11-104

Introduced before R2006a

rotatefactors

Rotate factor loadings

Syntax

```
B = rotatefactors(A)
B = rotatefactors(A,'Method','orthomax','Coeff',gamma)
B = rotatefactors(A,'Method','procrustes','Target',target)
B = rotatefactors(A,'Method','pattern','Target',target)
B = rotatefactors(A,'Method','promax')
[B,T] = rotatefactors(A,...)
```

Description

`B = rotatefactors(A)` rotates the *d*-by-*m* loadings matrix `A` to maximize the varimax criterion, and returns the result in `B`. Rows of `A` and `B` correspond to variables and columns correspond to factors, for example, the (*i*, *j*)th element of `A` is the coefficient for the *i* th variable on the *j* th factor. The matrix `A` usually contains principal component coefficients created with `pca` or `pcacov`, or factor loadings estimated with `factoran`.

`B = rotatefactors(A,'Method','orthomax','Coeff',gamma)` rotates `A` to maximize the orthomax criterion with the coefficient `gamma`, i.e., `B` is the orthogonal rotation of `A` that maximizes

```
sum(D*sum(B.^4,1) - GAMMA*sum(B.^2,1).^2)
```

The default value of 1 for `gamma` corresponds to varimax rotation. Other possibilities include `gamma` = 0, *m*/2, and *d*(*m* - 1)/(*d* + *m* - 2), corresponding to quartimax, equamax, and parsimax. You can also supply `'varimax'`, `'quartimax'`, `'equamax'`, or `'parsimax'` for the `'method'` parameter and omit the `'Coeff'` parameter.

If `'Method'` is `'orthomax'`, `'varimax'`, `'quartimax'`, `'equamax'`, or `'parsimax'`, then additional parameters are

- `'Normalize'` — Flag indicating whether the loadings matrix should be row-normalized for rotation. If `'on'` (the default), rows of `A` are normalized prior to rotation to have unit Euclidean norm, and unnormalized after rotation. If `'off'`, the raw loadings are rotated and returned.
- `'Reltol'` — Relative convergence tolerance in the iterative algorithm used to find `T`. The default is `sqrt(eps)`.
- `'Maxit'` — Iteration limit in the iterative algorithm used to find `T`. The default is `250`.

`B = rotatefactors(A,'Method','procrustes','Target',target)` performs an oblique procrustes rotation of `A` to the *d*-by-*m* target loadings matrix `target`.

`B = rotatefactors(A,'Method','pattern','Target',target)` performs an oblique rotation of the loadings matrix `A` to the *d*-by-*m* target pattern matrix `target`, and returns the result in `B`. `target` defines the "restricted" elements of `B`, i.e., elements of `B` corresponding to zero elements of `target` are constrained to have small magnitude, while elements of `B` corresponding to nonzero elements of `target` are allowed to take on any magnitude.

If `'Method'` is `'procrustes'` or `'pattern'`, an additional parameter is `'Type'`, the type of rotation. If `'Type'` is `'orthogonal'`, the rotation is orthogonal, and the factors remain

uncorrelated. If `'Type'` is `'oblique'` (the default), the rotation is oblique, and the rotated factors might be correlated.

When `'Method'` is `'pattern'`, there are restrictions on `target`. If `A` has *m* columns, then for orthogonal rotation, the *j*th column of `target` must contain at least *m* - *j* zeros. For oblique rotation, each column of `target` must contain at least *m* - 1 zeros.

`B = rotatefactors(A,'Method','promax')` rotates A to maximize the promax criterion, equivalent to an oblique Procrustes rotation with a target created by an orthomax rotation. Use the four orthomax parameters to control the orthomax rotation used internally by promax.

An additional parameter for 'promax' is `'Power'`, the exponent for creating promax target matrix. `'Power'` must be `1` or greater. The default is `4`.

`[B,T] = rotatefactors(A,...)` returns the rotation matrix `T` used to create `B`, that is, `B = A*T`. You can find the correlation matrix of the rotated factors by using `inv(T'*T)`. For orthogonal rotation, this is the identity matrix, while for oblique rotation, it has unit diagonal elements but nonzero off-diagonal elements.

Examples

```
rng('default') % for reproducibility
X = randn(100,10);

% Default (normalized varimax) rotation:
% first three principal components.
LPC = pca(X);
[L1,T] = rotatefactors(LPC(:,1:3));

% Equamax rotation:
% first three principal components.
[L2,T] = rotatefactors(LPC(:,1:3),...
                       'method','equamax');

% Promax rotation:
% first three factors.
LFA = factoran(X,3,'Rotate','none');
[L3,T] = rotatefactors(LFA(:,1:3),...
                       'method','promax',...
                       'power',2);

% Pattern rotation:
% first three factors.
Tgt = [1 1 1 1 1 0 1 0 1 1; ...
       0 0 0 1 1 1 0 0 0 0; ...
       1 0 0 1 0 1 1 1 1 0]';
[L4,T] = rotatefactors(LFA(:,1:3),...
                       'method','pattern',...
                       'target',Tgt);
inv(T'*T) % Correlation matrix of the rotated factors
ans =

    1.0000   -0.9593   -0.7098
   -0.9593    1.0000    0.5938
   -0.7098    0.5938    1.0000
```

References

[1] Harman, H. H. *Modern Factor Analysis*. 3rd ed. Chicago: University of Chicago Press, 1976.

[2] Lawley, D. N., and A. E. Maxwell. *Factor Analysis as a Statistical Method*. 2nd ed. New York: American Elsevier Publishing, 1971.

See Also

`biplot` | `factoran` | `pca` | `pcacov` | `procrustes`

Introduced before R2006a

rowexch

Row exchange

Syntax

```
dRE = rowexch(nfactors,nruns)
[dRE,X] = rowexch(nfactors,nruns)
[dRE,X] = rowexch(nfactors,nruns,model)
[dRE,X] = rowexch(...,param1,val1,param2,val2,...)
```

Description

`dRE = rowexch(nfactors,nruns)` uses a row-exchange algorithm to generate a *D*-optimal design `dRE` with `nruns` runs (the rows of `dRE`) for a linear additive model with `nfactors` factors (the columns of `dRE`). The model includes a constant term.

`[dRE,X] = rowexch(nfactors,nruns)` also returns the associated design matrix `X`, whose columns are the model terms evaluated at each treatment (row) of `dRE`.

`[dRE,X] = rowexch(nfactors,nruns,model)` uses the linear regression model specified in `model`. `model` is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

The order of the columns of `X` for a full quadratic model with *n* terms is:

1. The constant term
2. The linear terms in order 1, 2, ..., *n*
3. The interaction terms in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n*–1, *n*)
4. The squared terms in order 1, 2, ..., *n*

Other models use a subset of these terms, in the same order.

Alternatively, `model` can be a matrix specifying polynomial terms of arbitrary order. In this case, `model` should have one column for each factor and one row for each term in the model. The entries in any row of `model` are powers for the factors in the columns. For example, if a model has factors `X1`, `X2`, and `X3`, then a row `[0 1 2]` in `model` specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in `model` specifies a constant term, which can be omitted.

`[dRE,X] = rowexch(...,param1,val1,param2,val2,...)` specifies additional parameter/value pairs for the design. Valid parameters and their values are listed in the following table.

Parameter	Value
'bounds'	Lower and upper bounds for each factor, specified as a 2-by-nfactors matrix. Alternatively, this value can be a cell array containing nfactors elements, each element specifying the vector of allowable values for the corresponding factor.
'categorical'	Indices of categorical predictors.
'display'	Either 'on' or 'off' to control display of the iteration counter. The default is 'on'.
'excludefun'	Handle to a function that excludes undesirable runs. If the function is *f*, it must support the syntax $b = f(S)$, where S is a matrix of treatments with nfactors columns and b is a vector of Boolean values with the same number of rows as S. $b(i)$ is true if the ith row S should be excluded.
'init'	Initial design as an nruns-by-nfactors matrix. The default is a randomly selected set of points.
'levels'	Vector of number of levels for each factor.
'maxiter'	Maximum number of iterations. The default is 10.

options	A structure that specifies whether to run in parallel, and specifies the random stream or streams. Create the options structure with statset. This option requires Parallel Computing Toolbox. Option fields are:
	• UseParallel — Set to true to compute in parallel. Default is false.
	• UseSubstreams — Set to true to compute in parallel in a reproducible fashion. Default is false. To compute reproducibly, set Streams to a type allowing substreams: 'mlfg6331_64' or 'mrg32k3a'.
	• Streams — A RandStream object or cell array of such objects. If you do not specify Streams, rowexch uses the default stream or streams. If you choose to specify Streams, use a single object except in the case
	  • UseParallel is true
	  • UseSubstreams is false
	  In that case, use a cell array the same size as the Parallel pool.

'tries'	Number of times to try to generate a design from a new starting point. The algorithm uses random points for each try, except possibly the first. The default is 1.

Examples

Suppose you want a design to estimate the parameters in the following three-factor, seven-term interaction model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon$$

Use rowexch to generate a *D*-optimal design with seven runs:

```
nfactors = 3;
nruns = 7;
[dRE,X] = rowexch(nfactors,nruns,'interaction','tries',10)
dRE =
```

```
    -1    -1     1
     1    -1     1
     1    -1    -1
     1     1     1
    -1    -1    -1
    -1     1    -1
    -1     1     1
X =
     1    -1    -1     1     1    -1    -1
     1     1    -1     1    -1     1    -1
     1     1    -1    -1    -1    -1     1
     1     1     1     1     1     1     1
     1    -1    -1    -1     1     1     1
     1    -1     1    -1    -1     1    -1
     1    -1     1     1    -1    -1     1
```

Columns of the design matrix `X` are the model terms evaluated at each row of the design `dRE`. The terms appear in order from left to right: constant term, linear terms (1, 2, 3), interaction terms (12, 13, 23). Use `X` to fit the model, as described in “Linear Regression” on page 11-9, to response data measured at the design points in `dRE`.

Algorithms

Both `cordexch` and `rowexch` use iterative search algorithms. They operate by incrementally changing an initial design matrix X to increase $D = |X^TX|$ at each step. In both algorithms, there is randomness built into the selection of the initial design and into the choice of the incremental changes. As a result, both algorithms may return locally, but not globally, D-optimal designs. Run each algorithm multiple times and select the best result for your final design. Both functions have a `'tries'` parameter that automates this repetition and comparison.

At each step, the row-exchange algorithm exchanges an entire row of X with a row from a design matrix C evaluated at a candidate set of feasible treatments. The `rowexch` function automatically generates a C appropriate for a specified model, operating in two steps by calling the `candgen` and `candexch` functions in sequence. Provide your own C by calling `candexch` directly. In either case, if C is large, its static presence in memory can affect computation.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`candexch` | `candgen` | `cordexch`

Introduced before R2006a

rsmdemo

Interactive response surface demonstration

Syntax

```
rsmdemo
```

Description

`rsmdemo` opens a group of three graphical user interfaces for interactively investigating response surface methodology (RSM), nonlinear fitting, and the design of experiments.

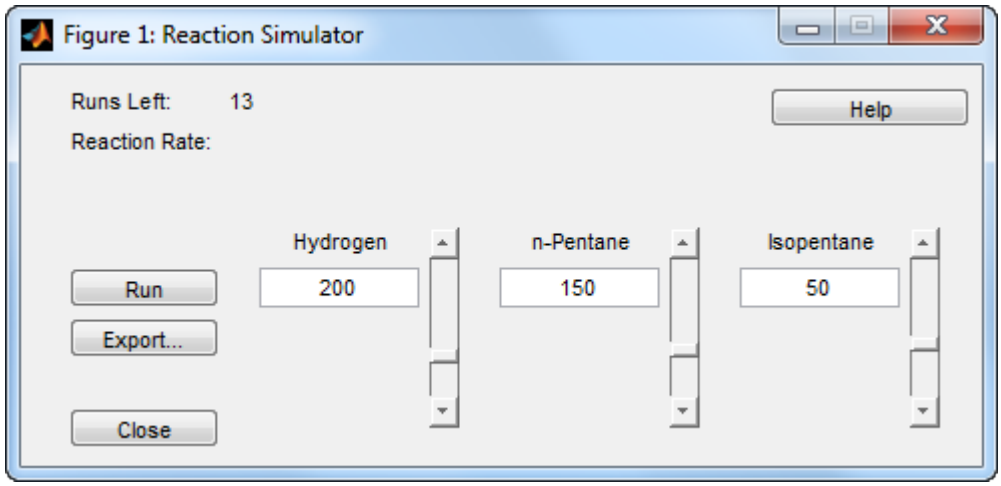
The interfaces allow you to collect and model data from a simulated chemical reaction. Experimental predictors are concentrations of three reactants (hydrogen, *n*-pentane, and isopentane) and the response is the reaction rate. The reaction rate is simulated by a Hougen-Watson model (Bates and Watts, [2] on page C-2, pp. 271–272):

$$rate = \frac{\beta_1 x_2 - x_3/\beta_5}{1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3}$$

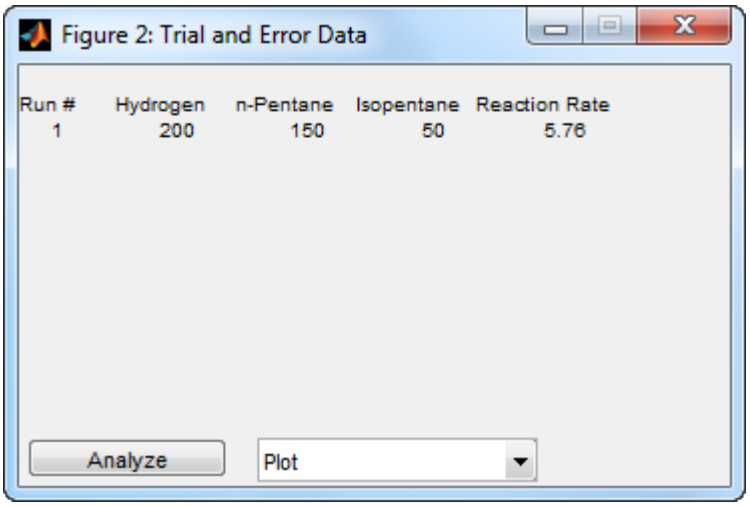
where *rate* is the reaction rate, x_1, x_2, and x_3 are the concentrations of hydrogen, *n*-pentane, and isopentane, respectively, and β_1, β_2, ... , β_5 are fixed parameters. Random errors are used to perturb the reaction rate for each combination of reactants.

Collect data using one of two methods:

1. Manually set reactant concentrations in the **Reaction Simulator** interface by editing the text boxes or by adjusting the associated sliders.

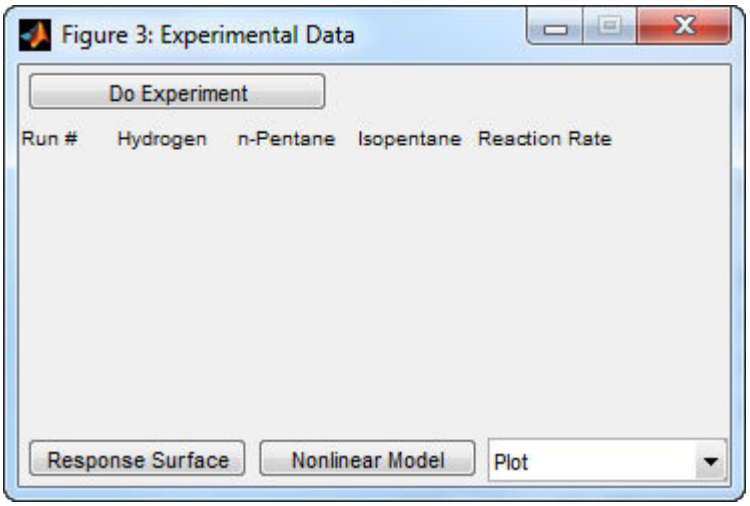


   When you click **Run**, the concentrations and simulated reaction rate are recorded on the **Trial and Error Data** interface.

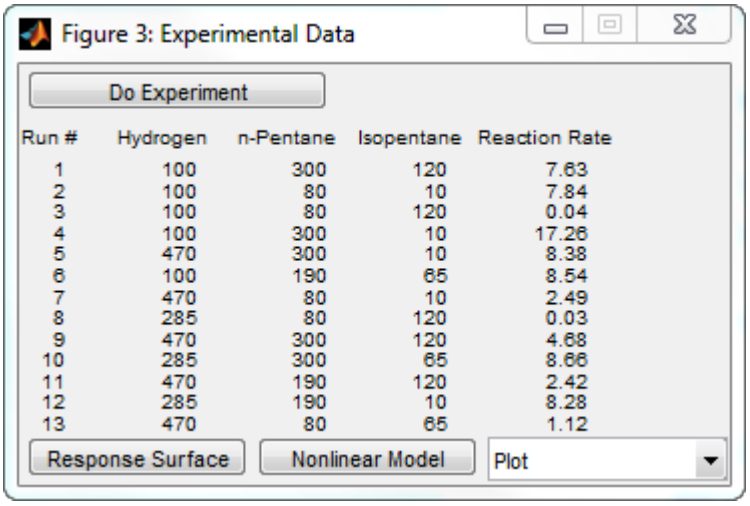


You are allowed up to 13 independent experimental runs for data collection.

2 Use a designed experiment to set reactant concentrations in the **Experimental Data** interface by clicking the **Do Experiment** button.



A 13-run *D*-optimal design for a full quadratic model is generated by the `cordexch` function, and the concentrations and simulated reaction rates are recorded on the same interface.



Once data is collected, scatter plots of reaction rates vs. individual predictors are generated by selecting one of the following from the **Plot** pop-up menu below the recorded data:

- **Hydrogen vs. Rate**
- **n-Pentane vs. Rate**
- **Isopentane vs. Rate**

Fit a response surface model to the data by clicking the **Analyze** button below the trial-and-error data or the **Response Surface** button below the experimental data. Both buttons load the data into the Response Surface Tool `rstool`. By default, trial-and-error data is fit with a linear additive model and experimental data is fit with a full quadratic model, but the models can be adjusted in the Response Surface Tool.

For experimental data, you have the additional option of fitting a Hougen-Watson model. Click the **Nonlinear Model** button to load the data and the model in `hougen` into the Nonlinear Fitting Tool `nlintool`.

See Also

`cordexch` | `hougen` | `nlintool` | `rstool`

Introduced before R2006a

rstool

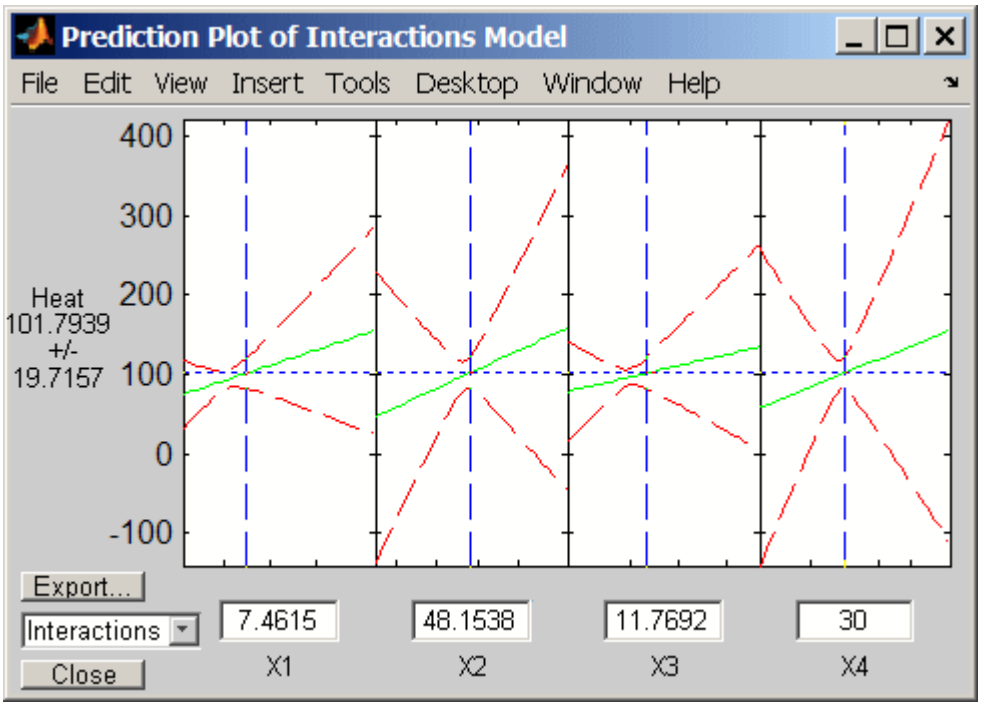
Interactive response surface modeling

Syntax

```
rstool
rstool(X,Y,model)
rstool(x,y,model,alpha)
rstool(x,y,model,alpha,xname,yname)
```

Description

`rstool` opens a graphical user interface for interactively investigating one-dimensional contours of multidimensional response surface models.



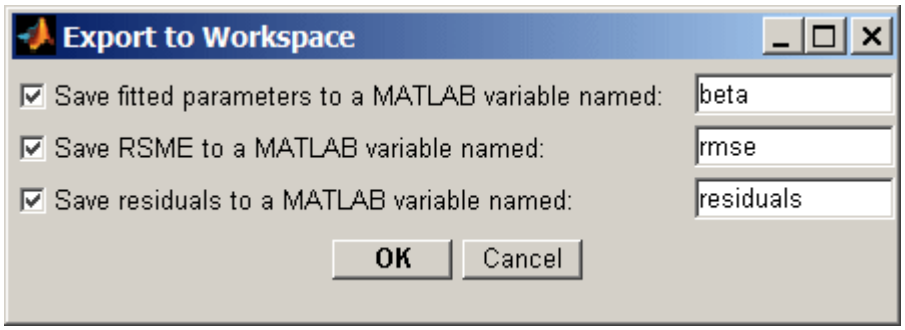
By default, the interface opens with the data from `hald.mat` and a fitted response surface with constant, linear, and interaction terms.

A sequence of plots is displayed, each showing a contour of the response surface against a single predictor, with all other predictors held fixed. `rstool` plots a 95% simultaneous confidence band for the fitted response surface as two red curves. Predictor values are displayed in the text boxes on the horizontal axis and are marked by vertical dashed blue lines in the plots. Predictor values are changed by editing the text boxes or by dragging the dashed blue lines. When you change the value of a predictor, all plots update to show the new point in predictor space.

The pop-up menu at the lower left of the interface allows you to choose among the following models:

- `Linear` — Constant and linear terms (the default)
- `Pure Quadratic` — Constant, linear, and squared terms
- `Interactions` — Constant, linear, and interaction terms
- `Full Quadratic` — Constant, linear, interaction, and squared terms

Click **Export** to open the following dialog box:



The dialog allows you to save information about the fit to MATLAB workspace variables with valid names.

`rstool(X,Y,model)` opens the interface with the predictor data in `X`, the response data in `Y`, and the fitted model `model`. Distinct predictor variables should appear in different columns of `X`. `Y` can be a vector, corresponding to a single response, or a matrix, with columns corresponding to multiple responses. `Y` must have as many elements (or rows, if it is a matrix) as `X` has rows.

The optional input `model` can be any one of the following:

- `'linear'` — Constant and linear terms (the default)
- `'purequadratic'` — Constant, linear, and squared terms
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms

To specify a polynomial model of arbitrary order, or a model without a constant term, use a matrix for `model` as described in `x2fx`.

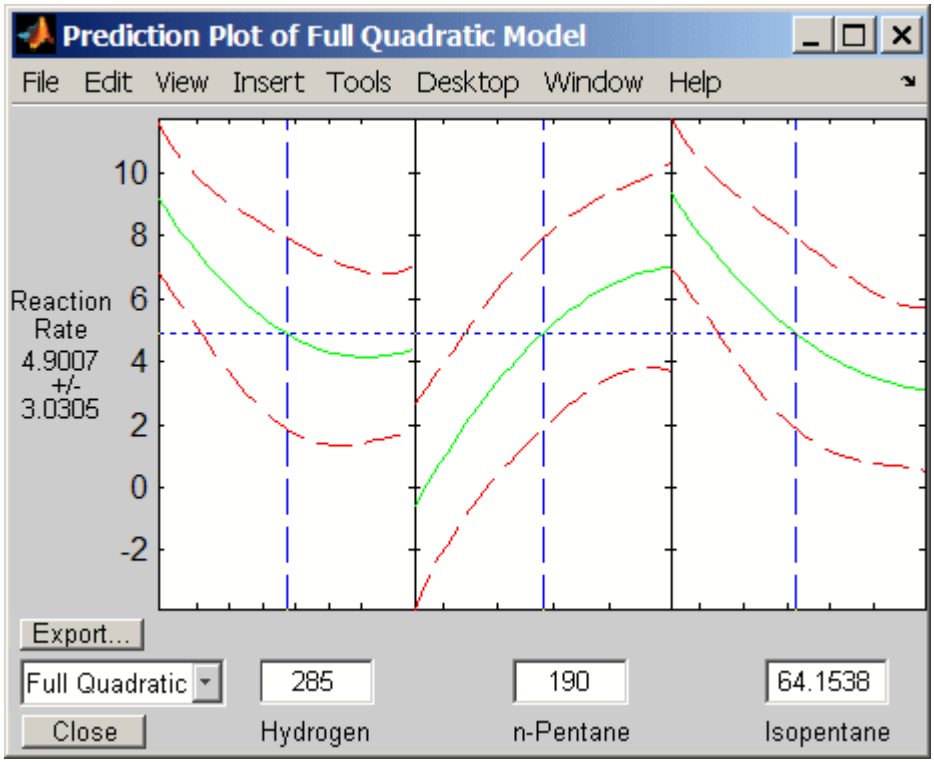
`rstool(x,y,model,alpha)` uses `100(1-alpha)`% global confidence intervals for new observations in the plots.

`rstool(x,y,model,alpha,xname,yname)` labels the axes using `xname` and `yname`. To label each subplot differently, `xname` and `yname` can be string arrays or cell arrays of character vectors.

Examples

The following uses `rstool` to visualize a quadratic response surface model of the 3-D chemical reaction data in `reaction.mat`:

```
load reaction
alpha = 0.01; % Significance level
rstool(reactants,rate,'quadratic',alpha,xn,yn)
```



The `rstool` interface is used by `rsmdemo` to visualize the results of simulated experiments with data like that in `reaction.mat`. As described in “Response Surface Designs” on page 28-8, `rsmdemo` uses a response surface model to generate simulated data at combinations of predictors specified by either the user or by a designed experiment.

See Also

`nlintool` | `rsmdemo` | `x2fx`

Introduced before R2006a

runstest

Run test for randomness

Syntax

```
h = runstest(x)
h = runstest(x,v)
h = runstest(x,'ud')
h = runstest( ___ ,Name,Value)
[h,p,stats] = runstest( ___ )
```

Description

`h = runstest(x)` returns a test decision for the null hypothesis that the values in the data vector `x` come in random order, against the alternative that they do not. The test is based on the number of runs of consecutive values above or below the mean of `x`. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`h = runstest(x,v)` returns a test decision based on the number of runs of consecutive values above or below the specified reference value `v`. Values exactly equal to `v` are discarded.

`h = runstest(x,'ud')` returns a test decision based on the number of runs up or down. Too few runs indicate a trend, while too many runs indicate an oscillation. Values exactly equal to the preceding value are discarded.

`h = runstest( ___ ,Name,Value)` returns a test decision using additional options specified by one or more name-value pair arguments. For example, you can change the significance level of the test, specify the algorithm used to calculate the *p*-value, or conduct a one-sided test.

`[h,p,stats] = runstest( ___ )` also returns the *p*-value of the test `p`, and a structure `stats` containing additional data about the test.

Examples

Test Data for Randomness Using Sample Median

Generate a vector of 40 random numbers from a standard normal distribution.

```
rng default;  % for reproducibility
x = randn(40,1);
```

Test whether the values in `x` appear in random order, using the sample median as the reference value.

```
[h,p] = runstest(x,median(x))
```

```
h = 0
```

```
p = 0.8762
```

The returned value of `h = 0` indicates that `runstest` does not reject the null hypothesis that the values in `x` are in random order at the default 5% significance level.

Input Arguments

`x` — Data vector

vector of scalar values

Data vector, specified as a vector of scalar values. `runstest` treats `NaN` values in `x` as missing values, and ignores them.

Data Types: `single` | `double`

`v` — Reference value

mean of `x` (default) | scalar value

Reference value, specified as a scalar value. If you specify a value for `v`, then `runstest` performs the hypothesis test based on the number of runs of consecutive values above or below `v`. `runstest` discards values exactly equal to `v`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'Method','Approximate','Tail','right'` specifies a right-tailed test with 1% significance level, which returns the approximate p-value.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Method` — Method used to compute *p*-value

`'exact'` | `'approximate'`

Method used to compute *p*-value, specified as the comma-separated pair consisting of `'Method'` and either `'exact'` to use an exact algorithm, or `'approximate'` to use a normal approximation. The default is `'exact'` for runs above/below, and for runs up/down when the length of `x` is less than or equal to 50. If `runstest` tests for runs up/down and the length of `x` is greater than 50, then the default is `'approximate'`, and the `'exact'` method is not available.

Example: `'Method','approximate'`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Two-tailed test (sequence is not random)
`'right'`	Right-tailed test (like values separate for runs above/below, direction alternates for runs up/down)
`'left'`	Left-tailed test (like values cluster for runs above/below, values trend for runs up/down)

Example: `'Tail','right'`

Output Arguments

`h` — Hypothesis test result

`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, then `runstest` rejects the null hypothesis at the `Alpha` significance level.
- If `h = 0`, then `runstest` fails to reject the null hypothesis at the `Alpha` significance level.

The result in `runstest` is based on the number of runs of consecutive values above or below the mean of `x`. Too few runs indicate a tendency for high and low values to cluster. Too many runs indicate a tendency for high and low values to alternate.

`runstest` uses a test statistic which is the difference between the number of runs and its mean, divided by its standard deviation. The test statistic is approximately normally distributed when the null hypothesis is true.

`p` — *p*-value

scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`p` is computed from either the test statistic or the exact distribution of the number of runs, depending on the value specified for the `'Method'` name-value pair argument.

`stats` — Test data

structure

Test data, returned as a structure with the following fields.

- `nruns` — The number of runs
- `n1` — The number of values above `v`
- `n0` — The number of values below `v`
- `z` — The test statistic

References

[1] Gibbons, Jean Dickinson, and Subhabrata Chakraborti. *Nonparametric Statistical Inference.* 5th ed. Boca Raton: CRC Press, 2011.

See Also

`signrank` | `signtest`

Introduced before R2006a

sampsizepwr

Sample size and power of test

Syntax

```
nout = sampsizepwr(testtype,p0,p1)
nout = sampsizepwr(testtype,p0,p1,pwr)

pwrout = sampsizepwr(testtype,p0,p1,[],n)

p1out = sampsizepwr(testtype,p0,[],pwr,n)

___ = sampsizepwr(testtype,p0,p1,pwr,n,Name,Value)
```

Description

`sampsizepwr` computes the sample size, power, or alternative parameter value for a hypothesis test, given the other two values. For example, you can compute the sample size required to obtain a particular power for a hypothesis test, given the parameter value of the alternative hypothesis.

`nout = sampsizepwr(testtype,p0,p1)` returns the sample size, `nout`, required for a two-sided test of the type specified by `testtype` to have a power (probability of rejecting the null hypothesis when the alternative hypothesis is true) of 0.90 when the significance level (probability of rejecting the null hypothesis when the null hypothesis is true) is 0.05. `p0` specifies parameter values under the null hypothesis. `p1` specifies the value, or an array of values, of the single parameter being tested under the alternative hypothesis.

`nout = sampsizepwr(testtype,p0,p1,pwr)` returns the sample size, `nout`, that corresponds to the specified power, `pwr`, and the parameter value under the alternative hypothesis, `p1`.

`pwrout = sampsizepwr(testtype,p0,p1,[],n)` returns the power achieved for a sample size of `n` when the true parameter value is `p1`.

`p1out = sampsizepwr(testtype,p0,[],pwr,n)` returns the parameter value detectable with the specified sample size, `n`, and the specified power, `pwr`.

`___ = sampsizepwr(testtype,p0,p1,pwr,n,Name,Value)` returns any of the previous arguments using one or more name-value pair arguments. For example, you can change the significance level of the test, or specify a right- or left-tailed test. The name-value pairs can appear in any order but must begin in the sixth argument position.

Examples

Compute Sample Size for Selected Power Value

A company runs a manufacturing process that fills empty bottles with 100 mL of liquid. To monitor quality, the company randomly selects several bottles and measures the volume of liquid inside.

Determine the sample size the company must use for a *t*-test to detect a difference between 100 mL and 102 mL with a power of 0.80. Assume that a standard deviation is 5 mL.

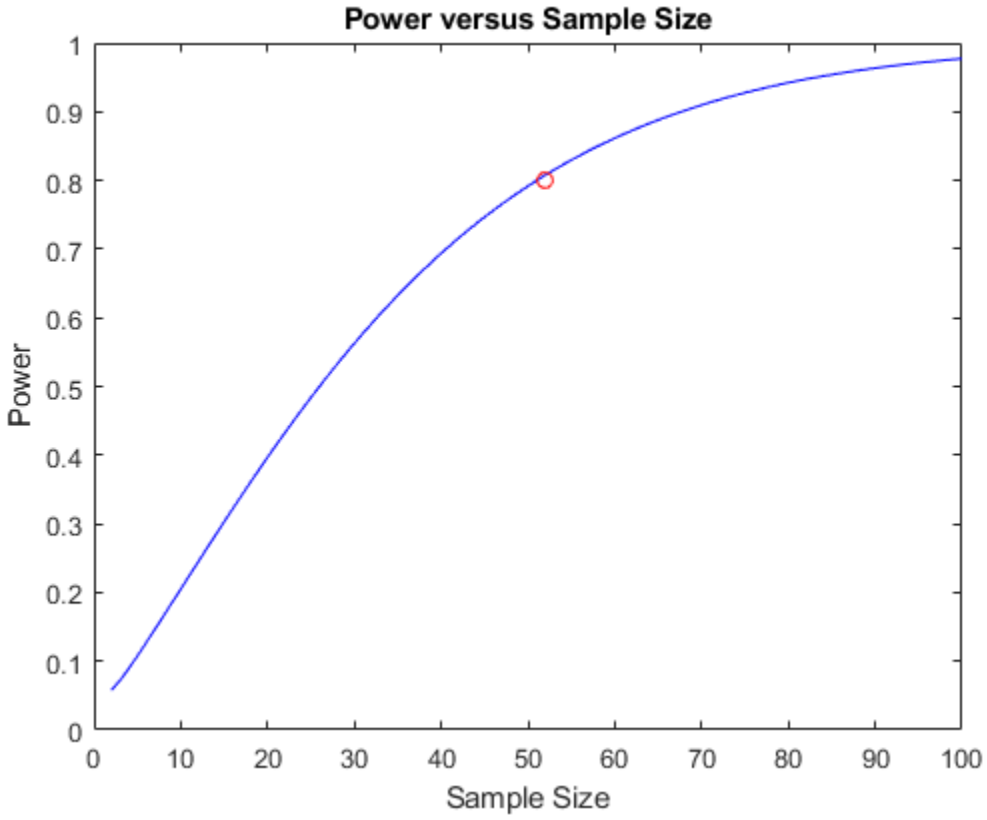
```
nout = sampsizepwr('t',[100 5],102,0.80)
```

```
nout = 52
```

The company must test 52 bottles to detect the difference between a mean volume of 100 mL and 102 mL with a power of 0.80.

Generate a power curve to visualize how the sample size affects the power of the test.

```
nn = 1:100;
pwrout = sampsizepwr('t',[100 5],102,[],nn);

figure;
plot(nn,pwrout,'b-',nout,0.8,'ro')
title('Power versus Sample Size')
xlabel('Sample Size')
ylabel('Power')
```



Compute Power and Sample Size for One-Sided Test

An employee wants to buy a house near her office. She decides to eliminate from consideration any house that has a mean morning commute time greater than 20 minutes. The null hypothesis for this

right-sided test is H0: $\mu = 20$, and the alternative hypothesis is HA: $\mu > 20$. The selected significance level is 0.05.

To determine the mean commute time, the employee takes a test drive from the house to her office during rush hour every morning for one week, so her total sample size is 5. She assumes that the standard deviation, σ, is equal to 5.

The employee decides that a true mean commute time of 25 minutes is too different from her targeted 20-minute limit, so she wants to detect a significant departure if the true mean is 25 minutes. Find the probability of incorrectly concluding that the mean commute time is no greater than 20 minutes.

Compute the power of the test, and then subtract the power from 1 to obtain β.

```
power = sampsizepwr('t',[20 5],25,[],5,'Tail','right');
beta = 1 - power
```

```
beta = 0.4203
```

The β value indicates a probability of 0.4203 that the employee concludes incorrectly that the morning commute is not greater than 20 minutes.

The employee decides that this risk is too high, and she wants no more than a 0.01 probability of reaching an incorrect conclusion. Calculate the number of test drives the employee must take to obtain a power of 0.99.

```
nout = sampsizepwr('t',[20 5],25,0.99,[],'Tail','right')
```

```
nout = 18
```

The results indicate that she must take 18 test drives from a candidate house to achieve this power level.

The employee decides that she only has time to take 10 test drives. She also accepts a 0.05 probability of making an incorrect conclusion. Calculate the smallest true parameter value that produces a detectable difference in mean commute time.

```
p1out = sampsizepwr('t',[20 5],[],0.95,10,'Tail','right')
```

```
p1out = 25.6532
```

Given the employee's target power level and sample size, her test detects a significant difference from a mean commute time of at least 25.6532 minutes.

Compute Sample Size for a Binomial Test

Compute the sample size, n, required to distinguish $p = 0.30$ from $p = 0.36$, using a binomial test with a power of 0.8.

```
napprox = sampsizepwr('p',0.30,0.36,0.8)
```

```
Warning: Values N>200 are approximate.  Plotting the power as a function
of N may reveal lower N values that have the required power.

napprox = 485
```

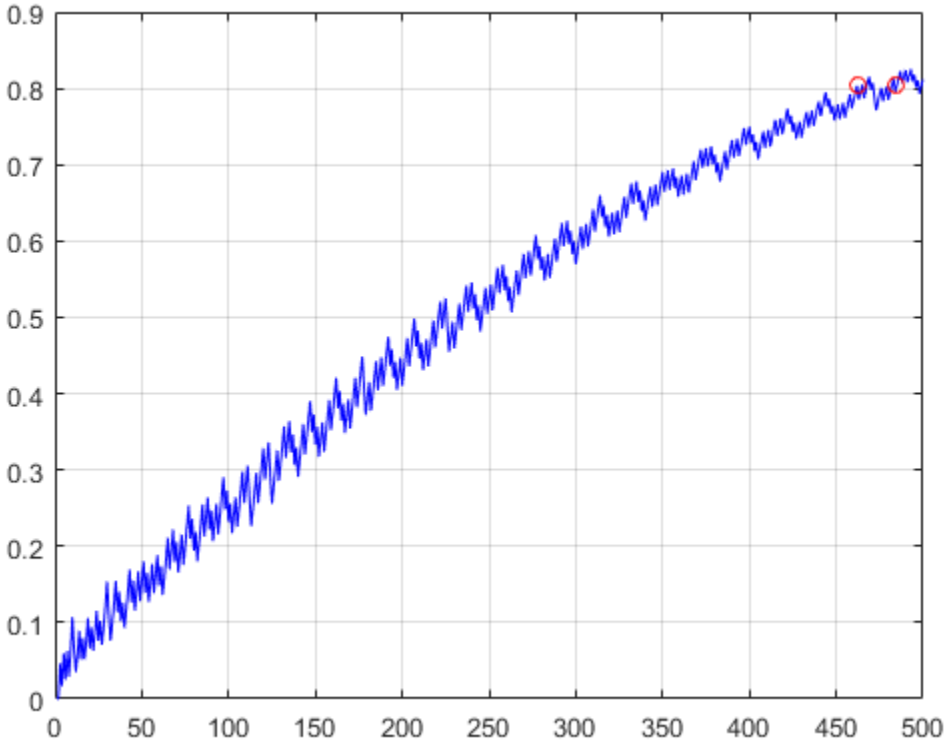
The result indicates that a power of 0.8 requires a sample size of 485. However, this result is approximate.

Make a plot to see if any smaller *n* values provide the required power of 0.8.

```
nn = 1:500;
pwrout = sampsizepwr('p',0.3,0.36,[],nn);
nexact = min(nn(pwrout>=0.8))
```

```
nexact = 462
```

```
figure
plot(nn,pwrout,'b-',[napprox nexact],pwrout([napprox nexact]),'ro')
grid on
```



The result indicates that a sample size of 462 also provides a power of 0.8 for this test.

Compute Power for a Two-Sample t-Test

A farmer wants to test the impact of two different types of fertilizer on the yield of his bean crops. He currently uses Fertilizer A, but believes that Fertilizer B might improve crop yield. Because Fertilizer B is more expensive than Fertilizer A, the farmer wants to limit the number of plans he treats with Fertilizer B in this experiment.

The farmer uses a 2:1 ratio of plants in each treatment group. He tests 10 plants with Fertilizer A, and 5 plants with Fertilizer B. The mean yield using Fertilizer A is 1.4 kg per plant, with a standard deviation of 0.2. The mean yield using Fertilizer B is 1.7 kg per plant. The significance level of the test is 0.05.

Compute the power of the test.

```
pwr = sampsizepwr('t2',[1.4 0.2],1.7,[],5,'Ratio',2)
```

```
pwr = 0.7165
```

The farmer wants to increase the power of the test to 0.90. Calculate how many plants he must treat with each type of fertilizer.

```
n = sampsizepwr('t2',[1.4 0.2],1.7,0.9,[])
```

```
n = 11
```

To increase the power of the test to 0.90, the farmer must test 11 plants with each type of fertilizer.

The farmer wants to reduce the number of plants he must treat with Fertilizer B, but keep the power of the test at 0.90 and maintain the initial 2:1 ratio of plants in each treatment group

Using a 2:1 ratio of plants in each treatment group, calculate how many plants the farmer must test to obtain a power of 0.90. Use the mean and standard deviation values obtained in the previous test.

```
[n1out,n2out] = sampsizepwr('t2',[1.4,0.2],1.7,0.9,[],'Ratio',2)
```

```
n1out = 8
```

```
n2out = 16
```

To obtain a power of 0.90, the farmer must treat 16 plants with Fertilizer A and 8 plants with Fertilizer B.

Input Arguments

`testtype` — Test type

`'z'` | `'t'` | `'t2'` | `'var'` | `'p'`

Test type, specified as one of the following.

- `'z'` — *z*-test for normally distributed data with known standard deviation.
- `'t'` — *t*-test for normally distributed data with unknown standard deviation.
- `'t2'` — Two-sample pooled *t*-test for normally distributed data with unknown standard deviation and equal variances.
- `'var'` — Chi-square test of variance for normally distributed data.
- `'p'` — Test of the *p* parameter (success probability) for a binomial distribution. The `'p'` test is a discrete test for which increasing the sample size does not always increase the power. For `n` values larger than 200, there may exist values smaller than the returned `n` value that also produce the specified power.

`p0` — Parameter value under null hypothesis

scalar value | two-element array of scalar values

Parameter value under the null hypothesis, specified as a scalar value or a two-element array of scalar values.

- If `testtype` is `'z'` or `'t'`, then `p0` is a two-element array `[mu0,sigma0]` of the mean and standard deviation, respectively, under the null hypothesis.
- If `testtype` is `'t2'`, then `p0` is a two-element array `[mu0,sigma0]` of the mean and standard deviation, respectively, of the first sample under the null and alternative hypotheses.
- If `testtype` is `'var'`, then `p0` is the variance under the null hypothesis.
- If `testtype` is `'p'`, then `p0` is the value of *p* under the null hypothesis.

Data Types: `single` | `double`

`p1` — Parameter value under alternative hypothesis
scalar value | array of scalar values | `[]`

Parameter value under the alternative hypothesis, specified as a scalar value or as an array of scalar values.

- If `testtype` is `'z'` or `'t'`, then `p1` is the value of the mean under the alternative hypothesis.
- If `testtype` is `'t2'`, then `p1` is the value of the mean of the second sample under the alternative hypothesis.
- If `testtype` is `'var'`, then `p1` is the variance under the alternative hypothesis.
- If `testtype` is `'p'`, then `p1` is the value of *p* under the alternative hypothesis.

If you specify `p1` as an array, then `sampsizepwr` returns an array for `nout` or `pwrout` that is the same length as `p1`.

To return the alternative parameter value, `p1out`, specify `p1` using empty brackets (`[]`), as shown in the syntax description on page 33-5761.

Data Types: `single` | `double`

`pwr` — Power of the test
0.90 (default) | scalar value in the range (0,1) | array of scalar values in the range (0,1) | `[]`

Power of the test, specified as a scalar value in the range (0,1) or as an array of scalar values in the range (0,1). The power of a test is the probability of rejecting the null hypothesis when the alternative hypothesis is true, given a particular significance level.

If you specify `pwr` as an array, then `sampsizepwr` returns an array for `nout` or `p1out` that is the same length as `pwr`.

To return a power value, `pwrout`, specify `pwr` using empty brackets (`[]`), as shown in the syntax description on page 33-5761.

Data Types: `single` | `double`

`n` — Sample size
positive integer value | array of positive integer values

Sample size, specified as a positive integer value or as an array of positive integer values.

If `testtype` is `'t2'`, then `sampsizepwr` assumes that the two sample sizes are equal. For unequal sample sizes, specify `n` as the smaller of the two sample sizes, and use the `'Ratio'` name-value pair

argument to indicate the sample size ratio. For example, if the smaller sample size is 5 and the larger sample size is 10, specify `n` as 5, and the `'Ratio'` name-value pair as 2.

If you specify `n` as an array, then `sampsizepwr` returns an array for `pwrout` or `p1out` that is the same length as `n`.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'Tail','right'` specifies a right-tailed test with a 0.01 significance level.

`Alpha` — Significance level
0.05 (default) | scalar value in the range (0,1)

Significance value of the test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Ratio` — Sample size ratio
1 (default) | scalar value greater than or equal to 1

Sample size ratio for a two-sample *t*-test, specified as the comma-separated pair consisting of `'Ratio'` and a scalar value greater than or equal to 1. The value of `Ratio` is equal to `n2/n1`, where `n2` is the larger sample size, and `n1` is the smaller sample size.

To return the power, `pwrout`, or alternative parameter value, `p1out`, specify the smaller of the two sample sizes for `n`, and use `'Ratio'` to indicate the sample size ratio.

Example: `'Ratio',2`

`Tail` — Test type
`'both'` (default) | `'right'` | `'left'`

Test type, specified as the comma-separated pair consisting of `'Tail'` and one of the following:

- `'both'` — Two-sided test for an alternative not equal to `p0`
- `'right'` — One-sided test for an alternative larger than `p0`
- `'left'` — One-sided test for an alternative smaller than `p0`

Example: `'Tail','right'`

Output Arguments

`nout` — Sample size
positive integer value | array of positive integer values

Sample size, returned as a positive integer value or as an array of positive integer values.

If `testtype` is `t2`, and you use the `'Ratio'` name-value pair argument to specify the ratio of the two unequal sample sizes, then `nout` returns the smaller of the two sample sizes.

Alternatively, to return both sample sizes, specify this argument as `[n1out,n2out]`. In this case, `sampsizepwr` returns the smaller sample size as `n1out`, and the larger sample size as `n2out`.

If you specify `pwr` or `p1` as an array, then `sampsizepwr` returns an array for `nout` that is the same length as `pwr` or `p1`.

`pwrout` — Power
scalar value in the range (0,1) | array of scalar values in the range (0,1)

Power achieved by the test, returned as a scalar value in the range (0,1) or as an array of scalar values in the range (0,1).

If you specify `n` or `p1` as an array, then `sampsizepwr` returns an array for `pwrout` that is the same length as `n` or `p1`.

`p1out` — Parameter value for the alternative hypothesis
scalar value | array of scalar values

Parameter value for the alternative hypothesis, returned as a scalar value or as an array of scalar values.

When computing `p1out` for the `'p'` test, if no alternative can be rejected for a given null hypothesis and significance level, the function displays a warning message and returns `NaN`.

See Also

`binocdf` | `ttest` | `ttest2` | `vartest` | `ztest`

Introduced in R2006b

saveCompactModel

(To be removed) Save model object in file for code generation

Note `saveCompactModel` will be removed in a future release. Use `saveLearnerForCoder` instead. To update your code, simply replace instances of `saveCompactModel` with `saveLearnerForCoder`.

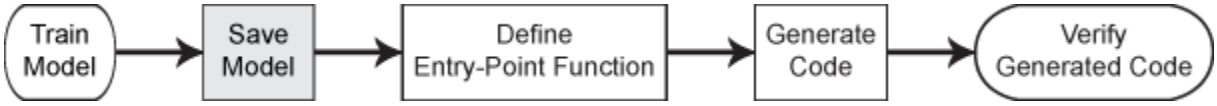
Syntax

```
saveCompactModel(Mdl,filename)
```

Description

To generate C/C++ code for the object functions (`predict`, `random`, `knnsearch`, or `rangesearch`) of machine learning models, use `saveCompactModel`, `loadCompactModel`, and `codegen`. After training a machine learning model, save the model by using `saveCompactModel`. Define an entry-point function that loads the model by using `loadCompactModel` and calls an object function. Then use `codegen` or the MATLAB Coder app to generate C/C++ code. Generating C/C++ code requires MATLAB Coder.

This flow chart shows the code generation workflow for the object functions of machine learning models. Use `saveCompactModel` for the highlighted step.



`saveCompactModel(Mdl,filename)` prepares a classification model, regression model, or nearest neighbor searcher (`Mdl`) for code generation and saves it in the MATLAB formatted binary file (MAT-file) named `filename`. You can pass `filename` to `loadCompactModel` to reconstruct the model object from the `filename` file.

Examples

Generate C/C++ Code for Prediction

After training a machine learning model, save the model by using `saveCompactModel`. Define an entry-point function that loads the model by using `loadCompactModel` and calls the `predict` function of the trained model. Then use `codegen` (MATLAB Coder) to generate C/C++ code.

This example briefly explains the code generation workflow for the prediction of machine learning models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details. To learn about the code generation for finding nearest neighbors using a nearest neighbor searcher model, see “Code Generation for Nearest Neighbor Searcher” on page 32-19.

Train Model

Load Fisher's iris data set. Remove all observed setosa irises data so that `X` and `Y` contain data for two classes only.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,:);
Y = species(inds);
```

Train a support vector machine (SVM) classification model using the processed data set.

```
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `SVMIris.mat` by using `saveCompactModel`.

```
saveCompactModel(Mdl,'SVMIris');
```

```
Warning: saveCompactModel will be removed in a future release. Use <a href="matlab:doc saveLearne
```

Define Entry-Point Function

Define an entry-point function named `classifyIrises` that does the following:

- Accept iris flower measurements with columns corresponding to `meas`, and return predicted labels.
- Load a trained SVM classification model.
- Predict labels using the loaded classification model for the iris flower measurements.

```
type classifyIrises.m % Display contents of classifyIrises.m file
```

```
function label = classifyIrises(X) %#codegen
%CLASSIFYIRISES Classify iris species using SVM Model
%   CLASSIFYIRISES classifies the iris flower measurements in X using the
%   compact SVM model in the file SVMIris.mat, and then returns class
%   labels in label.
CompactMdl = loadCompactModel('SVMIris');
label = predict(CompactMdl,X);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point

function at compile time. Pass `X` as the value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see "Specify Variable-Size Arguments for Code Generation" on page 32-45 and "Specify Properties of Entry-Point Function Inputs" (MATLAB Coder).

```
codegen classifyIrises -args {X}
```

```
Code generation successful.
```

`codegen` generates the MEX function `classifyIrises_mex` with a platform-dependent extension.

Verify Generated Code

Compare the labels classified using `predict`, `classifyIrises`, and `classifyIrises_mex`.

```
label1 = predict(Mdl,X);
label2 = classifyIrises(X);
```

```
Warning: loadCompactModel will be removed in a future release. Use <a href="matlab:doc loadLearne
```

```
label3 = classifyIrises_mex(X);
```

```
Warning: loadCompactModel will be removed in a future release. Use <a href="matlab:doc loadLearne
```

```
verify_label = isequal(label1,label2,label3)
```

```
verify_label = logical
   1
```

`isequal` returns logical 1 (true), which means all the inputs are equal. The labels classified all three ways are the same.

Input Arguments

`Mdl` — Machine learning model

full model object | compact model object

Machine learning model, specified as a full or compact model object, as given in the following tables of supported models.

- **Classification Model Object**

Model	Full/Compact Model Objects	Training Function
Binary decision tree for classification	`ClassificationTree`, `CompactClassificationTree`	`fitctree`
Discriminant analysis classification	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`	`fitcdiscr`
Ensemble classifier	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`	`fitcensemble`
k-nearest neighbor classification	`ClassificationKNN`	`fitcknn`
Linear model for binary classification of high-dimensional data	`ClassificationLinear`	`fitclinear`
Multiclass model for support vector machines (SVMs) or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`	`fitcecoc`
Naive Bayes classifier	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`	`fitcnb`
SVM for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	`fitcsvm`

- **Regression Model Object**

Model	Full/Compact Model Object	Training Function
Ensemble regression	`RegressionEnsemble`, `CompactRegressionEnsemble`, `RegressionBaggedEnsemble`	`fitrensemble`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`	`fitrgp`
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`	`fitglm`, `stepwiseglm`
Linear regression model	`LinearModel`, `CompactLinearModel`	`fitlm`, `stepwiselm`
Linear regression for high-dimensional data	`RegressionLinear`	`fitrlinear`
Regression tree	`RegressionTree`, `CompactRegressionTree`	`fitrtree`
SVM regression	`RegressionSVM`, `CompactRegressionSVM`	`fitrsvm`

- **Nearest Neighbor Searcher Object**

Model	Model Object	Training Function
Exhaustive nearest neighbor searcher	`ExhaustiveSearcher`	`ExhaustiveSearcher`, `createns`
Nearest neighbor searcher using *K*d-tree	`KDTreeSearcher`	`KDTreeSearcher`, `createns`

`filename` — File name
character vector | string scalar

File name, specified as a character vector or string scalar.

If the `filename` file exists, then `saveCompactModel` overwrites the file.

The extension of the `filename` file must be `.mat`. If `filename` has no extension, then `saveCompactModel` appends `.mat`.

If `filename` does not include a full path, then `saveCompactModel` saves the file to the current folder.

Example: `'SVMMdl'`

Data Types: `char` | `string`

Algorithms

`saveCompactModel` prepares a machine learning model (`Mdl`) for code generation. The function removes some properties that are not required for prediction.

- For a model that has a corresponding compact model, the `saveCompactModel` function applies the appropriate `compact` function to the model before saving it.
- For a model that does not have a corresponding compact model, such as `ClassificationKNN`, `ClassificationLinear`, `RegressionLinear`, `ExhaustiveSearcher`, and `KDTreeSearcher`, the `saveCompactModel` function removes properties such as hyperparameter optimization properties, training solver information, and others.

`loadCompactModel` loads the model saved by `saveCompactModel`.

Alternative Functionality

- Use a coder configurer created by `learnerCoderConfigurer` for the models listed in this table.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

  After training a machine learning model, create a coder configurer of the model. Use the object functions and properties of the configurer to configure code generation options and to generate code for the `predict` and `update` functions of the model. If you generate code using a coder configurer, you can update model parameters in the generated code without having to regenerate the code. For details, see “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80.

Compatibility Considerations

`saveCompactModel` will be removed
Warns starting in R2020b

`saveCompactModel` will be removed in a future release. Use `saveLearnerForCoder` instead.

`saveLearnerForCoder` and `loadLearnerForCoder` provide broader functionality, including fixed-point code generation for supported models.

This table shows how to update your code to use `saveLearnerForCoder`.

Not Recommended	Recommended
`saveCompactModel(Model,'MyModel');`	`saveLearnerForCoder(Model,'MyModel');`

See Also

`codegen` | `loadCompactModel` | `saveLearnerForCoder`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
“Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
“Code Generation for Nearest Neighbor Searcher” on page 32-19
“Specify Variable-Size Arguments for Code Generation” on page 32-45

Introduced in R2016b

saveLearnerForCoder

Save model object in file for code generation

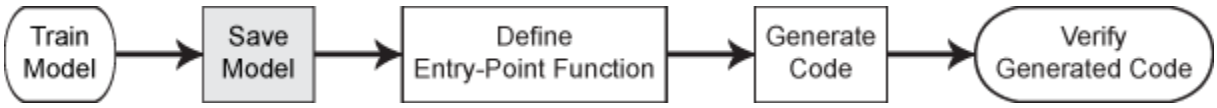
Syntax

```
saveLearnerForCoder(Mdl,filename)
```

Description

To generate C/C++ code for the object functions of machine learning models (including `predict`, `random`, `knnsearch`, `rangesearch`, and incremental learning functions), use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen`. After training a machine learning model, save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls an object function. Then use `codegen` or the MATLAB Coder app to generate C/C++ code. Generating C/C++ code requires MATLAB Coder.

This flow chart shows the code generation workflow for the object functions of machine learning models. Use `saveLearnerForCoder` for the highlighted step.



Fixed-point C/C++ code generation requires an additional step that defines the fixed-point data types of the variables required for prediction. Create a fixed-point data type structure by using the data type function generated by `generateLearnerDataTypeFcn`, and use the structure as an input argument of `loadLearnerForCoder` in an entry-point function. Generating fixed-point C/C++ code requires MATLAB Coder and Fixed-Point Designer.

This flow chart shows the fixed-point code generation workflow for the `predict` function of a machine learning model. Use `saveLearnerForCoder` for the highlighted step.

`saveLearnerForCoder(Mdl,filename)` prepares a classification model, regression model, or nearest neighbor searcher (`Mdl`) for code generation and saves it in the MATLAB formatted binary file (MAT-file) named `filename`. You can pass `filename` to `loadLearnerForCoder` to reconstruct the model object from the `filename` file.

Examples

Generate C/C++ Code for Prediction

After training a machine learning model, save the model by using `saveLearnerForCoder`. Define an entry-point function that loads the model by using `loadLearnerForCoder` and calls the `predict` function of the trained model. Then use `codegen` (MATLAB Coder) to generate C/C++ code.

This example briefly explains the code generation workflow for the prediction of machine learning models at the command line. For more details, see “Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9. You can also generate code using the MATLAB Coder app. See “Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22 for details. To learn about the code generation for finding nearest neighbors using a nearest neighbor searcher model, see “Code Generation for Nearest Neighbor Searcher” on page 32-19.

Train Model

Load Fisher's iris data set. Remove all observed setosa irises data so that `X` and `Y` contain data for two classes only.

```
load fisheriris
inds = ~strcmp(species,'setosa');
X = meas(inds,:);
Y = species(inds);
```

Train a support vector machine (SVM) classification model using the processed data set.

```
Mdl = fitcsvm(X,Y);
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `SVMIris.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'SVMIris');
```

Define Entry-Point Function

Define an entry-point function named `classifyIris` that does the following:

- Accept iris flower measurements with columns corresponding to `meas`, and return predicted labels.
- Load a trained SVM classification model.
- Predict labels using the loaded classification model for the iris flower measurements.

```
type classifyIris.m % Display contents of classifyIris.m file
```

```
function label = classifyIris(X) %#codegen
%CLASSIFYIRIS Classify iris species using SVM Model
%   CLASSIFYIRIS classifies the iris flower measurements in X using the SVM
%   model in the file SVMIris.mat, and then returns class labels in label.
Mdl = loadLearnerForCoder('SVMIris');
label = predict(Mdl,X);
end
```

Add the `%#codegen` compiler directive (or pragma) to the entry-point function after the function signature to indicate that you intend to generate code for the MATLAB algorithm. Adding this directive instructs the MATLAB Code Analyzer to help you diagnose and fix violations that would result in errors during code generation.

Note: If you click the button located in the upper-right section of this example and open this example in MATLAB®, then MATLAB® opens the example folder. This folder includes the entry-point function file.

Generate Code

Generate code for the entry-point function using `codegen` (MATLAB Coder). Because C and C++ are statically typed languages, you must determine the properties of all variables in the entry-point function at compile time. Pass `X` as the value of the `-args` option to specify that the generated code must accept an input that has the same data type and array size as the training data `X`. If the number of observations is unknown at compile time, you can also specify the input as variable-size by using `coder.typeof` (MATLAB Coder). For details, see “Specify Variable-Size Arguments for Code Generation” on page 32-45 and “Specify Properties of Entry-Point Function Inputs” (MATLAB Coder).

```
codegen classifyIris -args {X}
```

```
Code generation successful.
```

`codegen` generates the MEX function `classifyIris_mex` with a platform-dependent extension.

Verify Generated Code

Compare the labels classified using `predict`, `classifyIris`, and `classifyIris_mex`.

```
label1 = predict(Mdl,X);
label2 = classifyIris(X);
label3 = classifyIris_mex(X);
verify_label = isequal(label1,label2,label3)
```

```
verify_label = logical
   1
```

`isequal` returns logical 1 (true), which means all the inputs are equal. The labels classified all three ways are the same.

Generate Fixed-Point C/C++ Code for Prediction

After training a machine learning model, save the model using `saveLearnerForCoder`. For fixed-point code generation, specify the fixed-point data types of the variables required for prediction by using the data type function generated by `generateLearnerDataTypeFcn`. Then, define an entry-point function that loads the model by using both `loadLearnerForCoder` and the specified fixed-point data types, and calls the `predict` function of the model. Use `codegen` (MATLAB Coder) to generate fixed-point C/C++ code for the entry-point function, and then verify the generated code.

Before generating code using `codegen`, you can use `buildInstrumentedMex` (Fixed-Point Designer) and `showInstrumentationResults` (Fixed-Point Designer) to optimize the fixed-point data types to improve the performance of the fixed-point code. Record minimum and maximum values of named and internal variables for prediction by using `buildInstrumentedMex`. View the instrumentation results using `showInstrumentationResults`; then, based on the results, tune the fixed-point data type properties of the variables. For details regarding this optional step, see “Fixed-Point Code Generation for Prediction of SVM” on page 32-87.

Train Model

Load the `ionosphere` data set and train a binary SVM classification model.

```
load ionosphere
Mdl = fitcsvm(X,Y,'KernelFunction','gaussian');
```

`Mdl` is a `ClassificationSVM` model.

Save Model

Save the SVM classification model to the file `myMdl.mat` by using `saveLearnerForCoder`.

```
saveLearnerForCoder(Mdl,'myMdl');
```

Define Fixed-Point Data Types

Use `generateLearnerDataTypeFcn` to generate a function that defines the fixed-point data types of the variables required for prediction of the SVM model.

```
generateLearnerDataTypeFcn('myMdl',X)
```

`generateLearnerDataTypeFcn` generates the `myMdl_datatype` function.

Create a structure `T` that defines the fixed-point data types by using `myMdl_datatype`.

```
T = myMdl_datatype('Fixed')
```

```
T = struct with fields:
              XDataType: [0x0 embedded.fi]
          ScoreDataType: [0x0 embedded.fi]
    InnerProductDataType: [0x0 embedded.fi]
```

The structure `T` includes the fields for the named and internal variables required to run the `predict` function. Each field contains a fixed-point object, returned by `fi` (Fixed-Point Designer). The fixed-point object specifies fixed-point data type properties, such as word length and fraction length. For example, display the fixed-point data type properties of the predictor data.

```
T.XDataType
```

```
ans =

[]

          DataTypeMode: Fixed-point: binary point scaling
            Signedness: Signed
            WordLength: 16
        FractionLength: 14

        RoundingMethod: Floor
        OverflowAction: Wrap
           ProductMode: FullPrecision
  MaxProductWordLength: 128
               SumMode: FullPrecision
     MaxSumWordLength: 128
```

Define Entry-Point Function

Define an entry-point function named `myFixedPointPredict` that does the following:

- Accept the predictor data `X` and the fixed-point data type structure `T`.
- Load a fixed-point version of a trained SVM classification model by using both `loadLearnerForCoder` and the structure `T`.

- Predict labels and scores using the loaded model.

```
type myFixedPointPredict.m % Display contents of myFixedPointPredict.m file
```

```
function [label,score] = myFixedPointPredict(X,T) %#codegen
Mdl = loadLearnerForCoder('myMdl','DataType',T);
[label,score] = predict(Mdl,X);
end
```

Note: If you click the button located in the upper-right section of this example and open the example in MATLAB®, then MATLAB opens the example folder. This folder includes the entry-point function file.

Generate Code

The `XDataType` field of the structure `T` specifies the fixed-point data type of the predictor data. Convert `X` to the type specified in `T.XDataType` by using the `cast` (Fixed-Point Designer) function.

```
X_fx = cast(X,'like',T.XDataType);
```

Generate code for the entry-point function using `codegen`. Specify `X_fx` and constant folded `T` as input arguments of the entry-point function.

```
codegen myFixedPointPredict -args {X_fx,coder.Constant(T)}
```

```
Code generation successful.
```

`codegen` generates the MEX function `myFixedPointPredict_mex` with a platform-dependent extension.

Verify Generated Code

Pass predictor data to `predict` and `myFixedPointPredict_mex` to compare the outputs.

```
[labels,scores] = predict(Mdl,X);
[labels_fx,scores_fx] = myFixedPointPredict_mex(X_fx,T);
```

Compare the outputs from `predict` and `myFixedPointPredict_mex`.

```
verify_labels = isequal(labels,labels_fx)
```

```
verify_labels = logical
   1
```

`isequal` returns logical 1 (true), which means `labels` and `labels_fx` are equal. If the labels are not equal, you can compute the percentage of incorrectly classified labels as follows.

```
sum(strcmp(labels_fx,labels)==0)/numel(labels_fx)*100
```

```
ans = 0
```

Find the maximum of the relative differences between the score outputs.

```
relDiff_scores = max(abs((scores_fx.double(:,1)-scores(:,1))./scores(:,1)))
```

```
relDiff_scores = 0.0055
```

If you are not satisfied with the comparison results and want to improve the precision of the generated code, you can tune the fixed-point data types and regenerate the code. For details, see

"Tips" on page 33-2632 in `generateLearnerDataTypeFcn`, "Data Type Function" on page 33-2631, and "Fixed-Point Code Generation for Prediction of SVM" on page 32-87.

Input Arguments

`Mdl` — Machine learning model

full model object | compact model object

Machine learning model, specified as a full or compact model object, as given in the following tables of supported models. The tables also show whether each model supports fixed-point code generation.

- **Classification Model Object**

Model	Full/Compact Model Objects	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Binary decision tree for classification	`ClassificationTree`, `CompactClassificationTree`	Yes	Yes
Discriminant analysis classification	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`	No	Yes
Ensemble classifier	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`	Yes (Only for ensembles of decision trees)	Yes
Binary classification linear model for incremental learning	`incrementalClassificationLinear`	No	Yes
k-nearest neighbor classification	`ClassificationKNN`	No	Yes
Linear model for binary classification of high-dimensional data	`ClassificationLinear`	No	Yes
Multiclass model for support vector machines (SVMs) or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`	No	Yes
Naive Bayes classifier	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`	No	Yes
SVM for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	Yes	Yes

- **Regression Model Object**

Model	Full/Compact Model Object	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Ensemble regression	`RegressionEnsemble`, `CompactRegressionEnsemble`, `RegressionBaggedEnsemble`	Yes	Yes
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`	No	Yes (see "Tips" on page 33-3574)
Generalized linear model	`GeneralizedLinearModel`, `CompactGeneralizedLinearModel`	No	No
Linear regression model for incremental learning	`incrementalRegressionLinear`	No	Yes
Linear regression model	`LinearModel`, `CompactLinearModel`	No	Yes
Linear regression for high-dimensional data	`RegressionLinear`	No	No
Regression tree	`RegressionTree`, `CompactRegressionTree`	Yes	Yes
SVM regression	`RegressionSVM`, `CompactRegressionSVM`	Yes	Yes

- **Nearest Neighbor Searcher Object**

Model	Model Object	Fixed-Point Code Generation Support	Single-Precision Code Generation Support
Exhaustive nearest neighbor searcher	`ExhaustiveSearcher`	No	No
Nearest neighbor searcher using *K*d-tree	`KDTreeSearcher`	No	No

`filename` — File name
character vector | string scalar

File name, specified as a character vector or string scalar.

If the `filename` file exists, then `saveLearnerForCoder` overwrites the file.

The extension of the `filename` file must be `.mat`. If `filename` has no extension, then `saveLearnerForCoder` appends `.mat`.

If `filename` does not include a full path, then `saveLearnerForCoder` saves the file to the current folder.

Example: `'SVMMdl'`

Data Types: `char` | `string`

Algorithms

`saveLearnerForCoder` prepares a machine learning model (`Mdl`) for code generation. The function removes some unnecessary properties.

- For a model that has a corresponding compact model, the `saveLearnerForCoder` function applies the appropriate `compact` function to the model before saving it.
- For a model that does not have a corresponding compact model, such as `ClassificationKNN`, `ClassificationLinear`, `RegressionLinear`, `ExhaustiveSearcher`, and `KDTreeSearcher`, the `saveLearnerForCoder` function removes properties such as hyperparameter optimization properties, training solver information, and others.

`loadLearnerForCoder` loads the model saved by `saveLearnerForCoder`.

Alternative Functionality

- Use a coder configurer created by `learnerCoderConfigurer` for the models listed in this table.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

  After training a machine learning model, create a coder configurer of the model. Use the object functions and properties of the configurer to configure code generation options and to generate code for the `predict` and `update` functions of the model. If you generate code using a coder configurer, you can update model parameters in the generated code without having to regenerate the code. For details, see “Code Generation for Prediction and Update Using Coder Configurer” on page 32-80.

See Also

`codegen` | `generateLearnerDataTypeFcn` | `loadLearnerForCoder`

Topics

“Introduction to Code Generation” on page 32-2

“Code Generation for Prediction of Machine Learning Model at Command Line” on page 32-9
“Code Generation for Prediction of Machine Learning Model Using MATLAB Coder App” on page 32-22
“Code Generation for Nearest Neighbor Searcher” on page 32-19
“Fixed-Point Code Generation for Prediction of SVM” on page 32-87
“Specify Variable-Size Arguments for Code Generation” on page 32-45

Introduced in R2019b

scatterhist

Scatter plot with marginal histograms

Syntax

```
scatterhist(x,y)
scatterhist(x,y,Name,Value)

h = scatterhist( ___ )
```

Description

`scatterhist(x,y)` creates a 2-D scatter plot of the data in vectors `x` and `y`, and displays the marginal distributions of `x` and `y` as univariate histograms on the horizontal and vertical axes of the scatter plot, respectively.

`scatterhist(x,y,Name,Value)` creates the plot using additional options specified by one or more name-value pair arguments. For example, you can specify a grouping variable or change the display options.

`h = scatterhist( ___ )` returns a vector of three axis handles for the scatter plot, the histogram along the horizontal axis, and the histogram along the vertical axis, respectively, using any of the input arguments in the previous syntaxes.

Examples

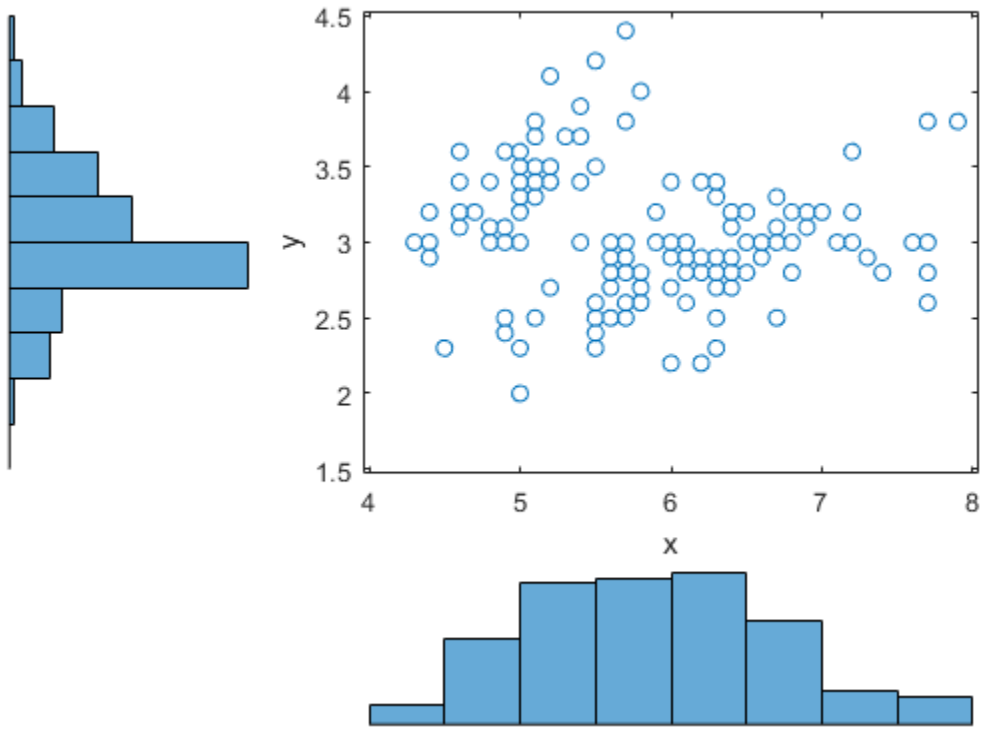
Create a `scatterhist` Plot

Load the sample data. Create data vector `x` from the first column of the data matrix, which contains sepal length measurements from iris flowers. Create data vector `y` from the second column of the data matrix, which contains sepal width measurements from the same flowers.

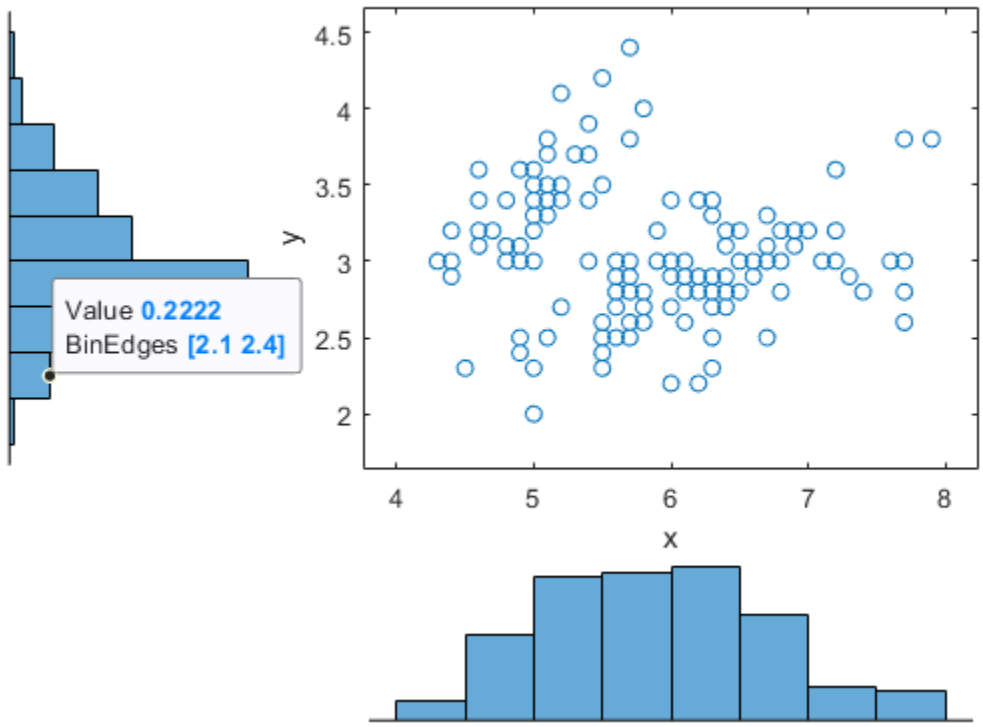
```
load fisheriris.mat;
x = meas(:,1);
y = meas(:,2);
```

Create a scatter plot and two marginal histograms to visualize the relationship between sepal length and sepal width.

```
scatterhist(x,y)
```



Display a data tip for a bin in a histogram. A data tip appears when you hover over a bin in a histogram.



The data tip displays the probability density function estimate of the selected bin and the lower and upper values for the bin edges.

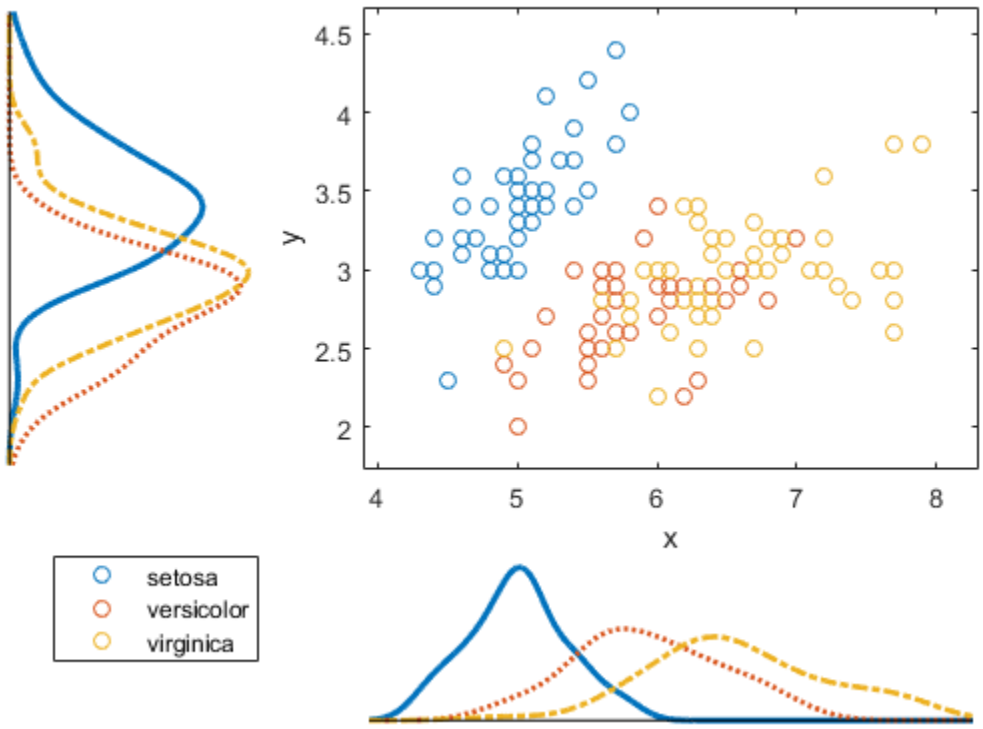
Plot Grouped Data

Load the sample data. Create data vector `x` from the first column of the data matrix, which contains sepal length measurements from three species of iris flowers. Create data vector `y` from the second column of the data matrix, which contains sepal width measurements from the same flowers.

```
load fisheriris.mat;
x = meas(:,1);
y = meas(:,2);
```

Create a scatter plot and six kernel density plots to visualize the relationship between sepal length and sepal width, grouped by species.

```
scatterhist(x,y,'Group',species,'Kernel','on')
```



The plot shows that the relationship between sepal length and width varies depending on the flower species.

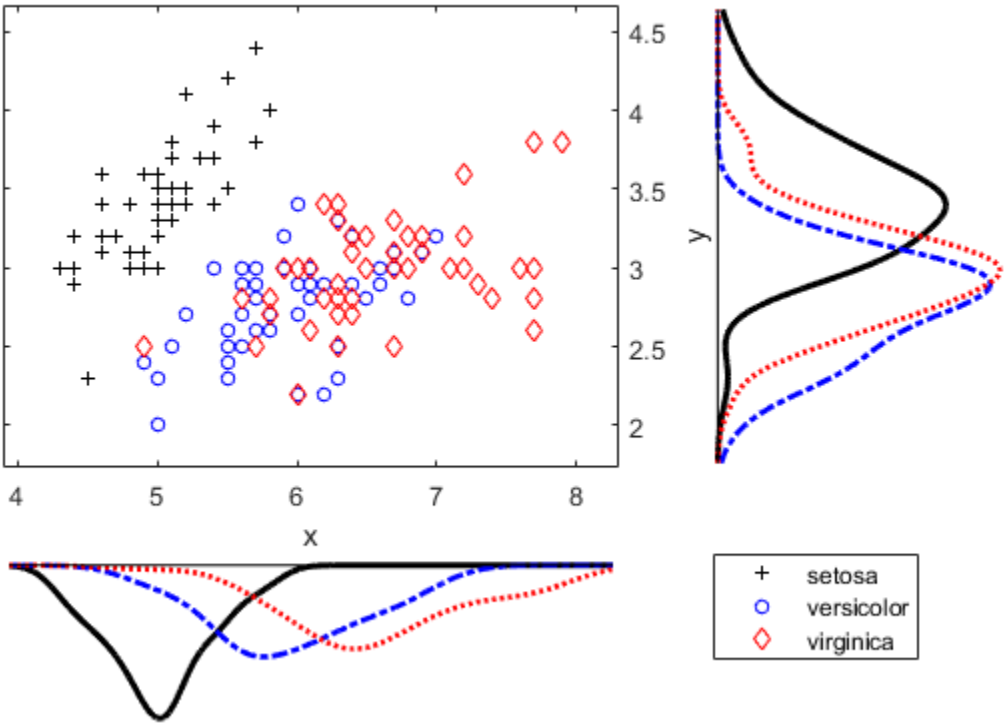
Customize the Plot Display

Load the sample data. Create data vector x from the first column of the data matrix, which contains sepal length measurements from three different species of iris flowers. Create data vector y from the second column of the data matrix, which contains sepal width measurements from the same flowers.

```
load fisheriris.mat;
x = meas(:,1);
y = meas(:,2);
```

Create a scatter plot and six kernel density plots to visualize the relationship between sepal length and sepal width as measured on three species of iris flowers, grouped by species. Customize the appearance of the plots.

```
scatterhist(x,y,'Group',species,'Kernel','on','Location','SouthEast',...
    'Direction','out','Color','kbr','LineStyle',{'-','-.',':'},...
    'LineWidth',[2,2,2],'Marker','+od','MarkerSize',[4,5,6]);
```



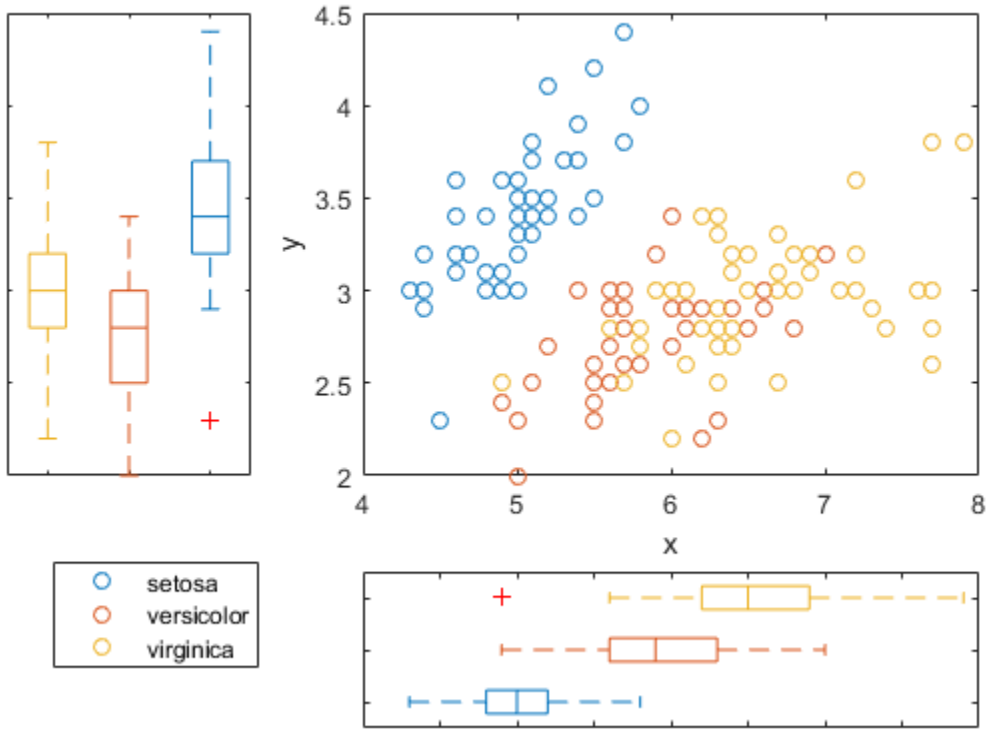
Customize Plots Using Axes Handles

Load the sample data. Create data vector x from the first column of the data matrix, which contains sepal length measurements from three species of iris flowers. Create data vector y from the second column of the data matrix, which contains sepal width measurements from the same flowers.

```
load fisheriris.mat;
x = meas(:,1);
y = meas(:,2);
```

Use axis handles to replace the marginal histograms with box plots.

```
h = scatterhist(x,y,'Group',species);
hold on;
clr = get(h(1),'colororder');
boxplot(h(2),x,species,'orientation','horizontal',...
     'label',{'','',''},'color',clr);
boxplot(h(3),y,species,'orientation','horizontal',...
     'label', {'','',''},'color',clr);
set(h(2:3),'XTickLabel','');
view(h(3),[270,90]);  % Rotate the Y plot
axis(h(1),'auto');  % Sync axes
hold off;
```



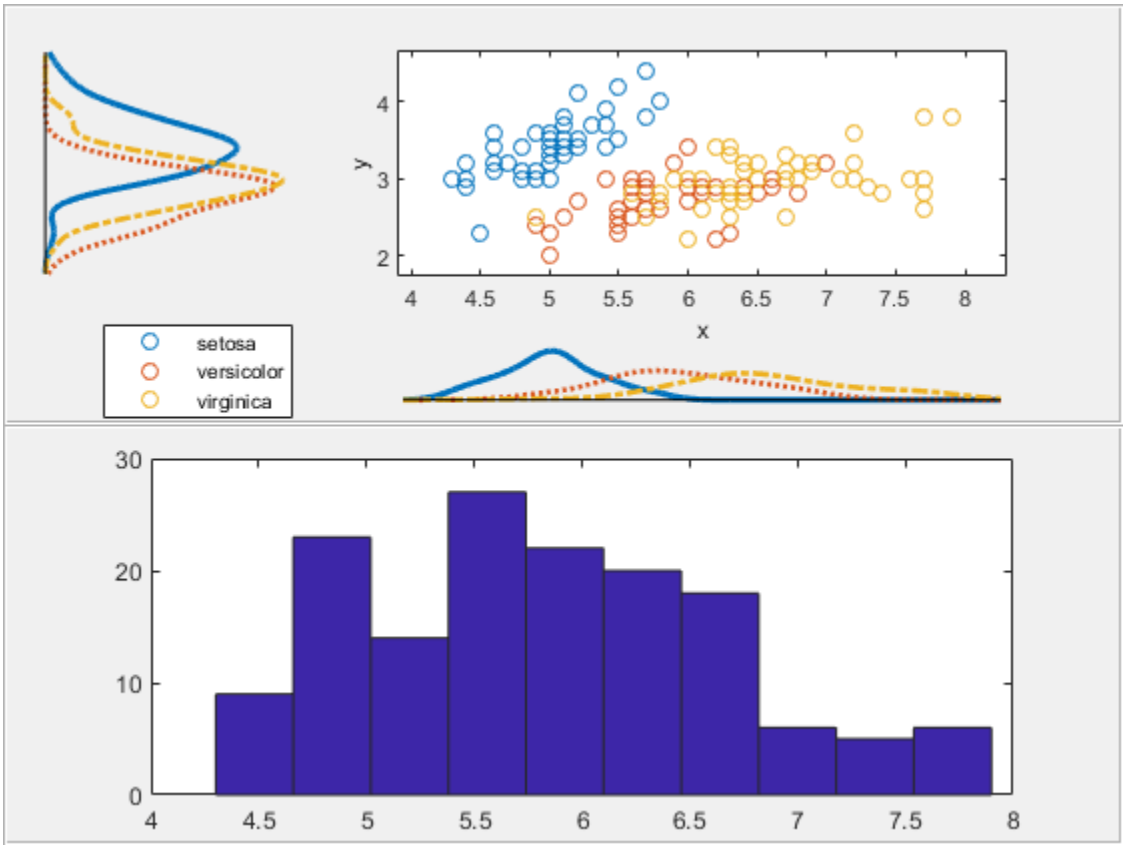
Create a `scatterhist` Plot in a Specified Parent Container

Load the sample data. Create data vector `x` from the first column of the data matrix, which contains sepal length measurements from iris flowers. Create data vector `y` from the second column of the data matrix, which contains sepal width measurements from the same flowers.

```
load fisheriris
x = meas(:,1);
y = meas(:,2);
```

Create a new figure and define two `uipanel` objects to divide the figure into two parts. In the upper half of the figure, plot the sample data using `scatterhist`. Include marginal kernel density plots grouped by species. In the lower half of the figure, plot a histogram of the sepal length measurements contained in `x`.

```
figure
hp1 = uipanel('position',[0 .5 1 .5]);
hp2 = uipanel('position',[0 0 1 .5]);
scatterhist(x,y,'Group',species,'Kernel','on','Parent',hp1);
axes('Parent',hp2);
hist(x);
```



Input Arguments

x — Sample data

vector

Sample data, specified as a vector. The data vectors `x` and `y` must be the same length.

If `x` or `y` contain `NaN` values, then `scatterhist`:

- Removes rows with `NaN` values in either `x` or `y` from both data vectors when generating the scatter plot
- Removes rows with `NaN` values only from the corresponding `x` or `y` data vector when generating the marginal histograms

Data Types: `single` | `double`

y — Sample data

vector

Sample data, specified as a vector. The data vectors `x` and `y` must be the same length.

If `x` or `y` contain `NaN` values, then `scatterhist`:

- Removes rows with `NaN` values in either `x` or `y` from both data vectors when generating the scatter plot

- Removes rows with `NaN` values only from the corresponding `x` or `y` data vector when generating the marginal histograms

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Location','SouthEast','Direction','out'` specifies a plot with histograms located below and to the right of the scatter plot, with the bars directed away from the scatter plot.

`NBins` — Number of bins for histograms

positive integer value | vector

Number of bins for histograms, specified as the comma-separated pair consisting of `'NBins'` and a positive integer value greater than or equal to 2, or vector of two such values. If the number of bins is specified as a positive integer value, that value is the number of bins for both the `x` and `y` histograms. If the number of bins is specified by a vector, the first value is the number of bins for the `x` data, and the second value is the number of bins for the `y` data. By default, the number of bins is computed based on the sample standard deviation using Scott's rule.

Example: `'NBins',[5,7]`

Data Types: `single` | `double`

`Location` — Location of marginal histograms

`'SouthWest'` (default) | `'SouthEast'` | `'NorthEast'` | `'NorthWest'`

Location of the marginal histograms in the figure, specified as the comma-separated pair consisting of `'Location'` and one of the following.

`'SouthWest'`	Plot the histograms below and to the left of the scatter plot.
`'SouthEast'`	Plot the histograms below and to the right of the scatter plot.
`'NorthEast'`	Plot the histograms above and to the right of the scatter plot.
`'NorthWest'`	Plot the histograms above and to the left of the scatter plot.

Example: `'Location','SouthEast'`

`Direction` — Direction of marginal histograms

`'in'` (default) | `'out'`

Direction of the marginal histograms, specified as the comma-separated pair consisting of `'Direction'` and one of the following.

`'in'`	Plot the histograms with the bars directed toward the scatter plot.
`'out'`	Plot the histograms with the bars directed away from the scatter plot.

Example: `'Direction','out'`

`Group` — Grouping variable
categorical array | logical or numeric vector | character array | string array | cell array of character vectors

Grouping variable, specified as the comma-separated pair consisting of `'Group'` and a categorical array, logical or numeric vector, character array, string array, or cell array of character vectors. Each unique value in a grouping variable defines a group.

For example, if `Gender` is a cell array of character vectors with values `'Male'` and `'Female'`, you can use `Gender` as a grouping variable to plot your data by gender.

The number of rows in the grouping variable must be equal to the length of `x`.

Example: `'Group',Gender`

Data Types: `categorical` | `single` | `double` | `logical` | `char` | `string` | `cell`

`PlotGroup` — Grouped plot indicator
`'on'` | `'off'`

Grouped plot indicator, specified as the comma-separated pair consisting of `'PlotGroup'` and one of the following.

`'on'`	Display grouped histograms or grouped kernel density plots. This is the default if a `Group` parameter is specified.
`'off'`	Display histograms or kernel density plots of the whole data set. This is the default if a `Group` parameter is not specified.

Example: `'PlotGroup','off'`

`Style` — Histogram display style
`'stairs'` | `'bar'`

Histogram display style, specified as the comma-separated pair consisting of `'PlotGroup'` and one of the following.

`'stairs'`	Display a stairstep plot that shows the outline of the histogram without filling the bars. This is the default if you specify a grouping variable that contains more than one group.
`'bar'`	Display a histogram bar plot. This is the default if you specify a grouping variable that contains only one group or if `PlotGroup` is specified as `'off'`.

Example: `'Style','bar'`

`Kernel` — Kernel density plot indicator
`'off'` (default) | `'on'` | `'overlay'`

Kernel density plot indicator, specified as the comma-separated pair consisting of `'Kernel'` and one of the following.

`'off'`	Display the marginal distributions as histograms.
`'on'`	Display the marginal distributions as kernel density plots.

`'overlay'`	Display the marginal distributions as kernel density plots overlaid onto histograms, similar to `histfit`.

Example: `'Kernel','overlay'`

`Bandwidth` — Bandwidth of kernel smoothing window
matrix

Bandwidth of kernel smoothing window, specified as the comma-separated pair consisting of `'Bandwidth'` and a matrix of size 2-by-*K*, where *K* is the number of unique groups. The first row of the matrix gives the bandwidth of each group in `x`, and the second row gives the bandwidth of each group in `y`. By default, `scatterhist` finds the optimal bandwidth for estimating normal densities. Specifying a different bandwidth value changes the smoothing characteristics of the resulting kernel density plot. The value specified is a scaling factor for the normal distribution used to generate the kernel density plot.

Example: `'Bandwidth',[.5,.2,.1;.15,.25,.35]`

Data Types: `single` | `double`

`Legend` — Legend visibility indicator
`'on'` | `'off'`

Legend visibility indicator, specified as the comma-separated pair consisting of `'Legend'` and one of the following.

`'on'`	Set legend visible. This is the default if a `Group` parameter is specified.
`'off'`	Set legend invisible. This is the default if a `Group` parameter is not specified.

Example: `'Legend','on'`

`Parent` — Parent container of the plot
`uipanel` container object | `figure` container object

Parent container for the plot, specified as a `uipanel` container object or `figure` container object. You can create panel container objects using `uipanel` or `figure`, respectively.

For example, if `h1` is a panel container object, specify the parent container of the plot as follows.

Example: `'Parent',h1`

`LineStyle` — Style of kernel density plot line
valid line style | string array or cell array of line styles

Style of kernel density plot line, specified as the comma-separated pair consisting of `'LineStyle'` and a valid line style or a string array or cell array of valid line styles. See `plot` for valid line styles. The default is a solid line. Use a string array or cell array to specify different line styles for each group. When the total number of groups exceeds the number of specified values, `scatterhist` cycles through the specified values.

Example: `'LineStyle',{'-',':','-.'}`

Data Types: `char` | `string` | `cell`

`LineWidth` — Width of kernel density plot line
`0.5` (default) | nonnegative scalar value | vector

Width of kernel density plot line, specified as the comma-separated pair consisting of `'LineWidth'` and a nonnegative scalar value or vector of nonnegative scalar values. The specified value is the size of the kernel density plot line measured in points. The default size is 0.5 points. Use a vector to specify different line widths for each group. When the total number of groups is greater than the number of specified values, `scatterhist` cycles through the specified values.

Example: `'LineWidth',[0.5,1,2]`

Data Types: `single` | `double`

`Color` — Marker color for each scatter plot group
character vector or string scalar of color names | matrix of RGB values

Marker color for each scatter plot group, specified as the comma-separated pair consisting of `'Color'` and a character vector or string scalar of color names, or a three-column matrix of RGB values in the range [0,1]. If you specify colors using a matrix, then each row of the matrix is an RGB triplet that represents a group. The three columns of the matrix represent the R value, G value, and B value, respectively. When the total number of groups exceeds the number of specified colors, `scatterhist` cycles through the specified colors.

This table lists the predefined colors and their equivalent RGB triplet values.

Option	Description	Equivalent RGB Triplet
`'red'` or `'r'`	Red	`[1 0 0]`
`'green'` or `'g'`	Green	`[0 1 0]`
`'blue'` or `'b'`	Blue	`[0 0 1]`
`'yellow'` or `'y'`	Yellow	`[1 1 0]`
`'magenta'` or `'m'`	Magenta	`[1 0 1]`
`'cyan'` or `'c'`	Cyan	`[0 1 1]`
`'white'` or `'w'`	White	`[1 1 1]`
`'black'` or `'k'`	Black	`[0 0 0]`

Example: `'Color','kcm'`

Example: `'Color',[.5,0,1;0,.5,.5]`

Data Types: `single` | `double` | `char` | `string`

`Marker` — Marker symbol for each scatterplot group
`'o'` (default) | character vector | string scalar

Marker symbol for each scatter plot group, specified as the comma-separated pair consisting of `'Marker'` and a character vector or string scalar of one or more valid marker symbols. See `plot` for valid symbols. The default is `'o'`, a circle. When the total number of groups exceeds the number of specified symbols, `scatterhist` cycles through the specified symbols.

Example: `'Marker','+do'`

Data Types: `char` | `string`

`MarkerSize` — Marker size for each scatter plot group
6 (default) | nonnegative scalar value | vector

Marker size for each scatter plot group, specified as the comma-separated pair consisting of `'MarkerSize'` and a nonnegative scalar value or a vector of nonnegative scalar values, measured in points. When the total number of groups exceeds the number of specified values, `scatterhist` cycles through the specified values.

Example: `'MarkerSize',10`

Data Types: `single` | `double`

Output Arguments

h — Axes handles

vector

Axes handles for the three plots, returned as a vector. The vector contains the handles for the scatter plot, the histogram along the horizontal axis, and the histogram along the vertical axis, respectively.

Alternative Functionality

Alternatively, you can create a `ScatterHistogramChart` object by using the `scatterhistogram` function.

- Explore the data interactively in the object by panning, zooming, and using data tips. Unlike the `scatterhist` function, `scatterhistogram` updates the marginal histograms based on the data within the current scatter plot limits.
- Control the appearance and behavior of the scatter histogram chart by changing the ScatterHistogramChart Properties.

See Also

`gscatter` | `histogram` | `scatterhistogram`

Topics

“Grouping Variables” on page 2-45

Introduced in R2007a

scramble

Scramble quasirandom point set

Syntax

```
ps = scramble(p,type)
ps = scramble(p,'clear')
ps = scramble(p)
```

Description

`ps = scramble(p,type)` returns a scrambled copy `ps` of the point set `p`, created using the scramble type specified by `type`. The point set `p` is either a `haltonset` or `sobolset` object, and each type of point set supports a different scramble type.

The scrambled point set `ps` is the same kind of object as `p`.

`ps = scramble(p,'clear')` removes the scramble setting from `p` and returns the result in `ps`.

`ps = scramble(p)` reapplies the existing scramble setting to `p`, which typically results in a different point set because of the randomness of the scrambling algorithms.

Examples

Create Halton Point Set

Generate a three-dimensional Halton point set, skip the first 1000 values, and then retain every 101st point.

```
p = haltonset(3,'Skip',1e3,'Leap',1e2)
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : none
```

Apply reverse-radix scrambling by using `scramble`.

```
p = scramble(p,'RR2')
```

```
p =
Halton point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
```

```
    ScrambleMethod : RR2
```

Generate the first four points by using `net`.

```
X0 = net(p,4)
```

```
X0 = 4×3

    0.0928    0.6950    0.0029
    0.6958    0.2958    0.8269
    0.3013    0.6497    0.4141
    0.9087    0.7883    0.2166
```

Generate every third point, up to the eleventh point, by using parenthesis indexing.

```
X = p(1:3:11,:)
```

```
X = 4×3

    0.0928    0.6950    0.0029
    0.9087    0.7883    0.2166
    0.3843    0.9840    0.9878
    0.6831    0.7357    0.7923
```

Scramble and Unscramble Sobol Point Set

Create and scramble a five-dimensional Sobol point set. Specify the `'MatousekAffineOwen'` scramble type.

```
p = sobolset(5);
ps = scramble(p,'MatousekAffineOwen');
```

Compare the first four points in the two point sets.

```
X = net(p,4)
```

```
X = 4×5

         0         0         0         0         0
    0.5000    0.5000    0.5000    0.5000    0.5000
    0.2500    0.7500    0.2500    0.7500    0.2500
    0.7500    0.2500    0.7500    0.2500    0.7500
```

```
X2 = net(ps,4)
```

```
X2 = 4×5

    0.6681    0.2784    0.2476    0.5688    0.0513
    0.4485    0.6735    0.5417    0.3285    0.9719
    0.9940    0.9606    0.3515    0.1586    0.4742
    0.1550    0.1202    0.9226    0.9262    0.5491
```

Remove the scramble setting from `ps` by using the `'clear'` option. The point set `clearps` matches the original point set `p`.

```
clearps = scramble(ps,'clear');
clearX = net(clearps,4)
```

```
clearX = 4×5

         0         0         0         0         0
    0.5000    0.5000    0.5000    0.5000    0.5000
    0.2500    0.7500    0.2500    0.7500    0.2500
    0.7500    0.2500    0.7500    0.2500    0.7500
```

Pass `ps` to the `scramble` function without additional input arguments. The software removes the scramble setting from `ps` and then reapplies it. Because of the randomness of the scrambling algorithm, the new scrambled point set `newps` differs from the original scrambled point set `ps`.

```
newps = scramble(ps);
newX = net(newps,4)
```

```
newX = 4×5

    0.6882    0.6261    0.9298    0.3314    0.4169
    0.2442    0.1978    0.4307    0.6286    0.8666
    0.7827    0.2868    0.5172    0.8430    0.1261
    0.2772    0.8576    0.0164    0.1404    0.5905
```

Input Arguments

p — Point set

`haltonset` object | `sobolset` object

Point set, specified as either a `haltonset` or `sobolset` object.

Example: `sobolset(5)`

type — Scramble type

`'RR2'` | `'MatousekAffineOwen'`

Scramble type, specified as `'RR2'` or `'MatousekAffineOwen'`. Different point sets support different scramble types, as indicated in this table.

Object	Scramble Type
`haltonset`	`'RR2'` — A permutation of the radical inverse coefficients derived by applying a reverse-radix operation to all of the possible coefficient values. The scramble is described in [1].
`sobolset`	`'MatousekAffineOwen'` — A random linear scramble combined with a random digital shift. The scramble is described in [2].

References

[1] Kocis, L., and W. J. Whiten. “Computational Investigations of Low-Discrepancy Sequences.” *ACM Transactions on Mathematical Software*. Vol. 23, No. 2, 1997, pp. 266–294.

[2] Matousek, J. “On the L2-Discrepancy for Anchored Boxes.” *Journal of Complexity*. Vol. 14, No. 4, 1998, pp. 527–556.

See Also

`haltonset` | `net` | `reduceDimensions` | `sobolset`

Introduced in R2008a

segment

Piecewise distribution segments containing input values

Syntax

```
s = segment(pd,x)
s = segment(pd,[],p)
```

Description

`s = segment(pd,x)` returns a vector `s` of positive integers indicating which segment in the piecewise distribution `pd` contains each quantile value in `x`.

The values 1, 2, and 3 in `s` indicate the lower tail, center, and upper tail segments in `pd`, respectively. If `pd` does not include a lower tail segment, then 1 and 2 indicate the center and upper tail segments, respectively.

`s = segment(pd,[],p)` returns a vector `s` of positive integers indicating which segment in the piecewise distribution `pd` contains each cumulative probability value in `p`.

Examples

Find Segment in `paretotails` Object

Generate a sample data set and create a `paretotails` object by fitting a piecewise distribution with Pareto tails to the generated data. Find the segment containing the specified quantile values by using the object function `segment`.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to x. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and generalized Pareto distributions (GPDs) for the lower and upper 10% of the data set.

```
pd = paretotails(x,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.33251    (0 < p < 0.1): lower tail, GPD(-0.0063504,0.567017)
   -1.33251 < x < 1.80149  (0.1 < p < 0.9): interpolated empirical cdf
        1.80149 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.24874,3.00974)
```

Find the segment containing the specified points by using the `segment` function.

```
xpts = -3:3;
s = segment(pd,xpts)
```

```
s = 1×7

     1     1     2     2     2     3     3
```

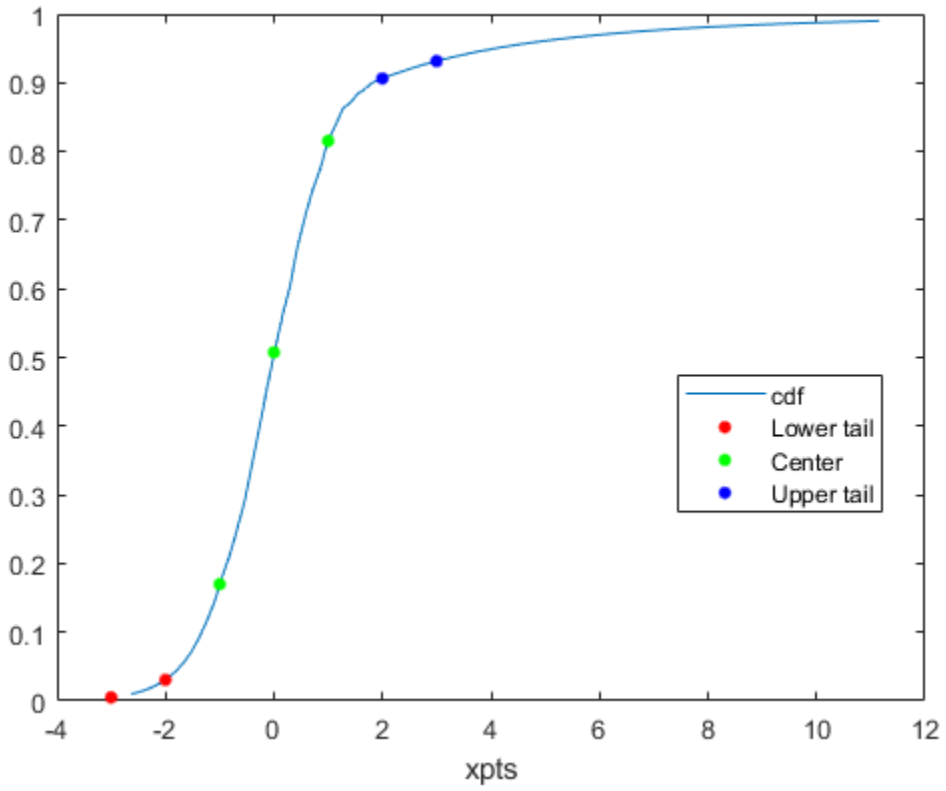
1, 2, and 3 indicate the lower tail, center, and upper tail segments in `pd`, respectively.

Draw the scatter plot of the points (`xpts`) grouped by their segments over the cumulative distribution function (cdf) plot. Plot the cdf of `pd`.

```
xgrid = linspace(icdf(pd,.01), icdf(pd,.99));
ygrid = cdf(pd,xgrid);
plot(xgrid,ygrid)
```

Superimpose the scatter plot of `xpts` by using `gscatter`.

```
hold on
gscatter(xpts,cdf(pd,xpts),s)
legend('cdf','Lower tail','Center','Upper tail')
hold off
```



Find Segment Containing Boundary Points

Generate a sample data set and create a `paretotails` object by fitting a piecewise distribution with Pareto tails to the generated data. Find the segment containing the boundary points by using the object function `segment`.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and generalized Pareto distributions (GPDs) for the lower and upper 10% of the data set.

```
pd = paretotails(x,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.33251    (0 < p < 0.1): lower tail, GPD(-0.0063504,0.567017)
   -1.33251 < x < 1.80149  (0.1 < p < 0.9): interpolated empirical cdf
        1.80149 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.24874,3.00974)
```

Return the boundary values between the piecewise segments by using the `boundary` function.

```
[p,q] = boundary(pd)
```

```
p = 2×1

    0.1000
    0.9000


q = 2×1

   -1.3325
    1.8015
```

The values in `p` are the cumulative probabilities at the boundaries, and the values in `q` are the corresponding quantiles.

Find the segment containing the boundary points by using the quantile values.

```
s1 = segment(pd,q)
```

```
s1 = 2×1

     2
     3
```

1, 2, and 3 indicate the lower tail, center, and upper tail segments in `pd`, respectively. The output `s1` implies that the first boundary between the lower tail segment and the center segment belongs to the

center segment, and the second boundary between the center segment and the upper tail segment belongs to the upper tail segment.

You can also use the cumulative probability values to find the corresponding segments.

```
s2 = segment(pd,[],[0;p;1])
```

```
s2 = 4×1

     1
     2
     3
     3
```

Input Arguments

pd — Piecewise distribution with Pareto tails

`paretotails` object

Piecewise distribution with Pareto tails, specified as a `paretotails` object.

x — Quantile

numeric vector

Quantile values, specified as a numeric vector.

Data Types: `single` | `double`

p — Cumulative probability

numeric vector of range `[0,1]` values

Cumulative probability values, specified as a numeric vector of range `[0,1]` values.

Data Types: `single` | `double`

See Also

`boundary` | `lowerparams` | `nsegments` | `paretotails` | `upperparams`

Topics

“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

selectModels

Class: `ClassificationLinear`

Choose subset of regularized, binary linear classification models

Syntax

```
SubMdl = selectModels(Mdl,idx)
```

Description

`SubMdl = selectModels(Mdl,idx)` returns a subset of trained, binary linear classification models from a set of binary linear classification models (`Mdl`) trained using various regularization strengths. The indices (`idx`) correspond to the regularization strengths in `Mdl.Lambda`, and specify which models to return.

Input Arguments

`Mdl` — Binary linear classification models trained using various regularization strengths

`ClassificationLinear` model object

Binary linear classification models trained using various regularization strengths, specified as a `ClassificationLinear` model object. You can create a `ClassificationLinear` model object using `fitclinear`.

Although `Mdl` is one model object, if `numel(Mdl.Lambda)` = $L \geq 2$, then you can think of `Mdl` as L trained models.

`idx` — Indices corresponding to regularization strengths

numeric vector of positive integers

Indices corresponding to regularization strengths, specified as a numeric vector of positive integers. Values of `idx` must be in the interval [1,L], where L = `numel(Mdl.Lambda)`.

Data Types: `double` | `single`

Output Arguments

`SubMdl` — Subset of binary linear classification models trained using various regularization strengths

`ClassificationLinear` model object

Subset of binary linear classification models trained using various regularization strengths, returned as a `ClassificationLinear` model object.

Examples

Find Good Lasso Penalty Using Classification Loss

To determine a good lasso-penalty strength for a linear classification model that uses a logistic regression learner, compare test-sample classification error rates.

Load the NLP data set. Preprocess the data as in “Specify Custom Classification Loss” on page 33-3644.

```
load nlpdata
Ystats = Y == 'stats';
X = X';

rng(10); % For reproducibility
Partition = cvpartition(Ystats,'Holdout',0.30);
testIdx = test(Partition);
XTest = X(:,testIdx);
YTest = Ystats(testIdx);
```

Create a set of 11 logarithmically-spaced regularization strengths from 10^{-6} through $10^{-0.5}$.

```
Lambda = logspace(-6,-0.5,11);
```

Train binary, linear classification models that use each of the regularization strengths. Optimize the objective function using SpaRSA. Lower the tolerance on the gradient of the objective function to `1e-8`.

```
CVMdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'CVPartition',Partition,'Learner','logistic','Solver','sparsa',...
    'Regularization','lasso','Lambda',Lambda,'GradientTolerance',1e-8)
```

```
CVMdl =
  ClassificationPartitionedLinear
    CrossValidatedModel: 'Linear'
           ResponseName: 'Y'
        NumObservations: 31572
                  KFold: 1
              Partition: [1x1 cvpartition]
             ClassNames: [0 1]
         ScoreTransform: 'none'


  Properties, Methods
```

Extract the trained linear classification model.

```
Mdl = CVMdl.Trained{1}
```

```
Mdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'logit'
              Beta: [34023x11 double]
              Bias: [1x11 double]
            Lambda: [1x11 double]
           Learner: 'logistic'
```

```
  Properties, Methods
```

`Mdl` is a `ClassificationLinear` model object. Because `Lambda` is a sequence of regularization strengths, you can think of `Mdl` as 11 models, one for each regularization strength in `Lambda`.

Estimate the test-sample classification error.

```
ce = loss(Mdl,X(:,testIdx),Ystats(testIdx),'ObservationsIn','columns');
```

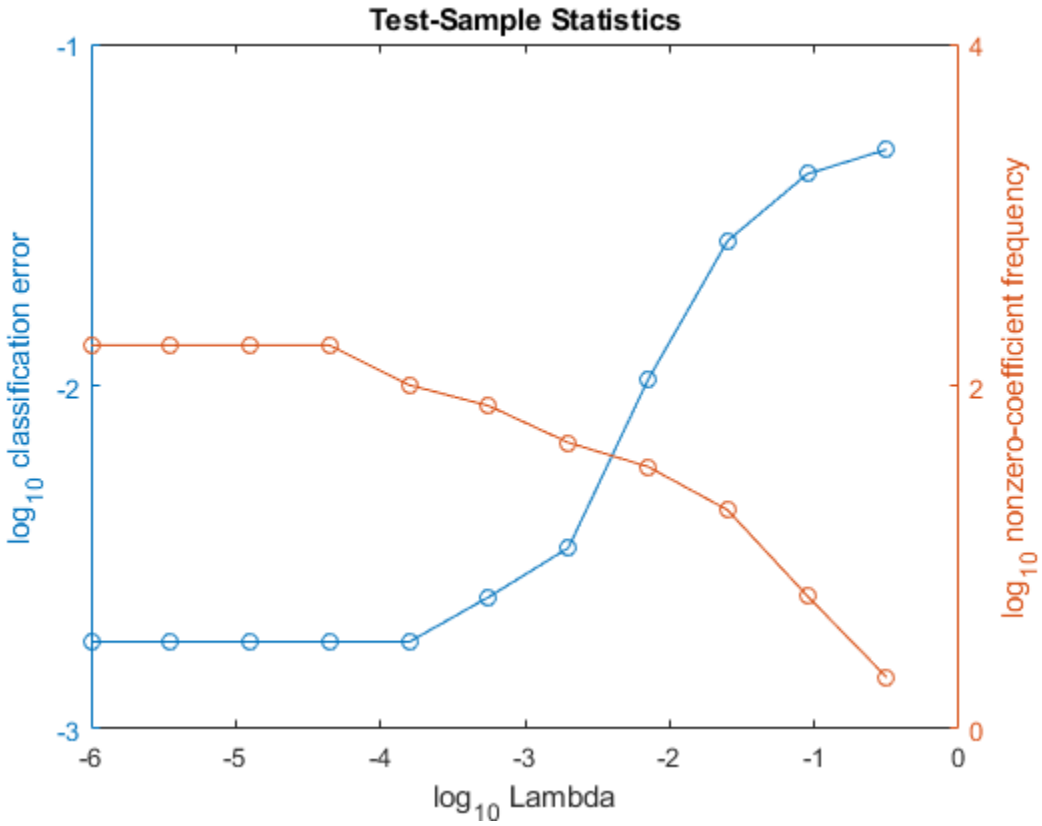
Because there are 11 regularization strengths, `ce` is a 1-by-11 vector of classification error rates.

Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a classifier. For each regularization strength, train a linear classification model using the entire data set and the same options as when you cross-validated the models. Determine the number of nonzero coefficients per model.

```
Mdl = fitclinear(X,Ystats,'ObservationsIn','columns',...
    'Learner','logistic','Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda,'GradientTolerance',1e-8);
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the test-sample error rates and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(ce),...
    log10(Lambda),log10(numNZCoeff + 1));
hL1.Marker = 'o';
hL2.Marker = 'o';
ylabel(h(1),'log_{10} classification error')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
title('Test-Sample Statistics')
hold off
```



Choose the index of the regularization strength that balances predictor variable sparsity and low classification error. In this case, a value between 10^{-4} to 10^{-1} should suffice.

```
idxFinal = 7;
```

Select the model from `Mdl` with the chosen regularization strength.

```
MdlFinal = selectModels(Mdl,idxFinal);
```

`MdlFinal` is a `ClassificationLinear` model containing one regularization strength. To estimate labels for new observations, pass `MdlFinal` and the new data to `predict`.

Tip

One way to build several predictive, binary linear classification models is:

1. Hold out a portion of the data for testing.
2. Train a binary, linear classification model using `fitclinear`. Specify a grid of regularization strengths using the `'Lambda'` name-value pair argument and supply the training data. `fitclinear` returns one `ClassificationLinear` model object, but it contains a model for each regularization strength.
3. To determine the quality of each regularized model, pass the returned model object and the held-out data to, for example, `loss`.

4. Identify the indices (`idx`) of a satisfactory subset of regularized models, and then pass the returned model and the indices to `selectModels`. `selectModels` returns one `ClassificationLinear` model object, but it contains `numel(idx)` regularized models.
5. To predict class labels for new data, pass the data and the subset of regularized models to `predict`.

See Also

`ClassificationLinear` | `fitclinear` | `loss` | `predict`

Introduced in R2016a

selectModels

Package: `classreg.learning.classif`

Choose subset of multiclass ECOC models composed of binary `ClassificationLinear` learners

Syntax

```
SubMdl = selectModels(Mdl,idx)
```

Description

`SubMdl = selectModels(Mdl,idx)` returns a subset of trained error-correcting output codes (ECOC) models composed of `ClassificationLinear` binary models from a set of multiclass ECOC models (`Mdl`) trained using various regularization strengths. The indices (`idx`) correspond to the regularization strengths in `Mdl.BinaryLearners{1}.Lambda` and specify which models to return.

`SubMdl` is returned as a `CompactClassificationECOC` model object.

Examples

Select Best Regularized Models

Choose a subset of trained ECOC models composed of linear binary learners with various regularization strengths.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels.

Create a set of 11 logarithmically spaced regularization strengths from 10^{-8} through 10^{-1}.

```
Lambda = logspace(-8,-1,11);
```

Create a linear classification model template that specifies optimizing the objective function using SpaRSA. Use lasso penalties with the strengths specified in `Lambda`.

```
t = templateLinear('Solver','sparsa','Regularization','lasso',...
    'Lambda',Lambda);
```

Hold out 30% of the data for testing. Identify the test-sample indices.

```
rng(1); % For reproducibility
cvp = cvpartition(Y,'Holdout',0.30);
idxTest = test(cvp);
```

Train an ECOC model composed of linear classification models. For quicker execution time, orient the predictor data so that individual observations correspond to columns.

```
X = X';
PMdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns','CVPartition',cvp);
Mdl = PMdl.Trained{1};
numel(Mdl.BinaryLearners{1}.Lambda)

ans = 11
```

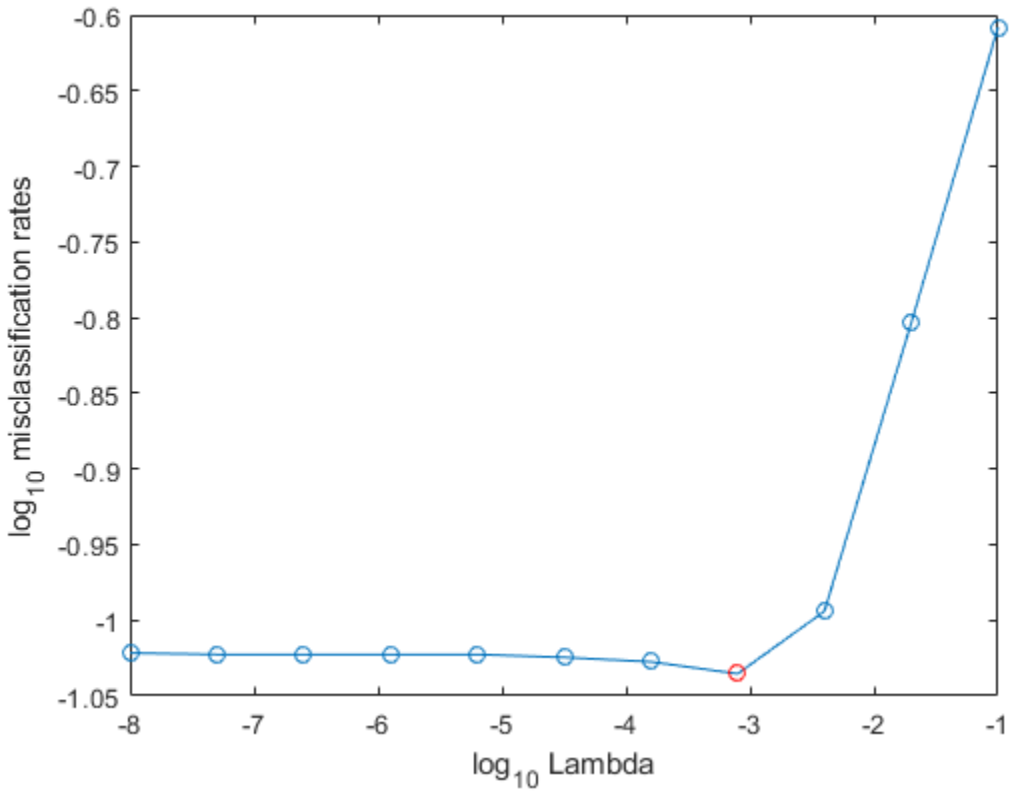
`Mdl` is a `CompactClassificationECOC` model object. Because `Lambda` is an 11-dimensional vector of regularization strengths, you can think of `Mdl` as eleven trained models, each corresponding to a regularization strength.

Estimate the test-sample misclassification rates for each regularized model.

```
ce = loss(Mdl,X(:,idxTest),Y(idxTest),'ObservationsIn','columns');
```

Plot the misclassification rates with respect to regularization strength on the log scale.

```
figure
plot(log10(Lambda),log10(ce),'-o')
ylabel('log_{10} misclassification rates')
xlabel('log_{10} Lambda')
[~,minCEIdx] = min(ce);
minLambda = Lambda(minCEIdx);
hold on
plot(log10(minLambda),log10(ce(minCEIdx)),'ro');
hold off
```



Several values of `Lambda` yield similarly small classification error values. Consider choosing greater values of `Lambda` (that still yield good classification rates) because they lead to predictor variable sparsity.

Select the four models with regularization strengths that occur around the point at which the classification error starts increasing.

```
idx = 7:10;
MdlFinal = selectModels(Mdl,idx)
```

```
MdlFinal =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [1x13 categorical]
    ScoreTransform: 'none'
    BinaryLearners: {78x1 cell}
      CodingMatrix: [13x78 double]


  Properties, Methods
```

```
LambdaFinal = MdlFinal.BinaryLearners{1}.Lambda
```

```
LambdaFinal = 1×4

    0.0002    0.0008    0.0040    0.0200
```

`MdlFinal` is a `CompactClassificationECOC` model object. You can think of it as four models trained using the four regularization strengths in `LambdaFinal`.

Input Arguments

`Mdl` — Multiclass ECOC model composed of binary linear classifiers

`CompactClassificationECOC` model object

Multiclass ECOC model composed of binary linear classifiers, trained using various regularization strengths, specified as a `CompactClassificationECOC` model object.

When creating `Mdl`, you must:

- Use `fitcecoc`.
- Specify `ClassificationLinear` binary learners (see `Learners`).
- Specify the same regularization strengths for each linear binary learner.

Although `Mdl` is one model object, if `numel(Mdl.BinaryLearners{1}.Lambda)` = $L \geq 2$, then you can think of `Mdl` as L trained models.

`idx` — Indices corresponding to regularization strengths

positive integer vector

Indices corresponding to regularization strengths, specified as a positive integer vector. Values of `idx` must be in the interval [1,L], where L = `numel(Mdl.BinaryLearners{1}.Lambda)`.

Data Types: `double` | `single`

Tips

- One way to build several predictive ECOC models composed of binary linear classification models is:

  1. Create a linear classification model template using `templateLinear` and specify a grid of regularization strengths using the `'Lambda'` name-value pair argument.
  2. Hold out a portion of the data for testing.
  3. Train an ECOC model using `fitcecoc`. Specify the template using the `'Learners'` name-value pair argument and supply the training data. `fitcecoc` returns one `CompactClassificationECOC` model object containing `ClassificationLinear` binary learners, but all binary learners contain a model for each regularization strength.
  4. To determine the quality of each regularized model, pass the returned model object and the held-out data to, for example, `loss`.
  5. Identify the indices (`idx`) of a satisfactory subset of regularized models, and then pass the returned model and the indices to `selectModels`. The function `selectModels` returns one `CompactClassificationECOC` model object, but it contains `numel(idx)` regularized models.
  6. To predict class labels for new data, pass the data and the subset of regularized models to `predict`.

See Also

`ClassificationLinear` | `CompactClassificationECOC` | `fitcecoc` | `loss` | `predict` | `templateLinear`

Introduced in R2016a

selectModels

Class: `RegressionLinear`

Select fitted regularized linear regression models

Syntax

```
SubMdl = selectModels(Mdl,idx)
```

Description

`SubMdl = selectModels(Mdl,idx)` returns a subset of trained linear regression models from a set of linear regression models (`Mdl`) trained using various regularization strengths. The indices `idx` correspond to the regularization strengths in `Mdl.Lambda`, and specify which models to return.

Input Arguments

`Mdl` — Linear regression models trained using various regularization strengths

`RegressionLinear` model object

Linear regression models trained using various regularization strengths, specified as a `RegressionLinear` model object. You can create a `RegressionLinear` model object using `fitrlinear`.

Although `Mdl` is one model object, if `numel(Mdl.Lambda)` = $L \geq 2$, then you can think of `Mdl` as L trained models.

`idx` — Indices corresponding to regularization strengths

numeric vector of positive integers

Indices corresponding to regularization strengths, specified as a numeric vector of positive integers. Values of `idx` must be in the interval [1,L], where L = `numel(Mdl.Lambda)`.

Data Types: `double` | `single`

Output Arguments

`SubMdl` — Subset of linear regression models trained using various regularization strengths

`RegressionLinear` model object

Subset of linear regression models trained using various regularization strengths, returned as a `RegressionLinear` model object.

Examples

Find Good Lasso Penalty Using Regression Loss

Simulate 10000 observations from this model

$$y = x_{100} + 2x_{200} + e\,.$$

- $X = \{x_1, \ldots, x_{1000}\}$ is a 10000-by-1000 sparse matrix with 10% nonzero standard normal elements.
- e is random normal error with mean 0 and standard deviation 0.3.

```
rng(1) % For reproducibility
n = 1e4;
d = 1e3;
nz = 0.1;
X = sprandn(n,d,nz);
Y = X(:,100) + 2*X(:,200) + 0.3*randn(n,1);
```

Create a set of 15 logarithmically-spaced regularization strengths from 10^{-4} through 10^{-1}.

```
Lambda = logspace(-4,-1,15);
```

Hold out 30% of the data for testing. Identify the test-sample indices.

```
cvp = cvpartition(numel(Y),'Holdout',0.30);
idxTest = test(cvp);
```

Train a linear regression model using lasso penalties with the strengths in `Lambda`. Specify the regularization strengths, optimizing the objective function using SpaRSA, and the data partition. To increase execution speed, transpose the predictor data and specify that the observations are in columns.

```
X = X';
CVMdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Solver','sparsa','Regularization','lasso','CVPartition',cvp);
Mdl1 = CVMdl.Trained{1};
numel(Mdl1.Lambda)
```

```
ans = 15
```

`Mdl1` is a `RegressionLinear` model. Because `Lambda` is a 15-dimensional vector of regularization strengths, you can think of `Mdl1` as 15 trained models, one for each regularization strength.

Estimate the test-sample mean squared error for each regularized model.

```
mse = loss(Mdl1,X(:,idxTest),Y(idxTest),'ObservationsIn','columns');
```

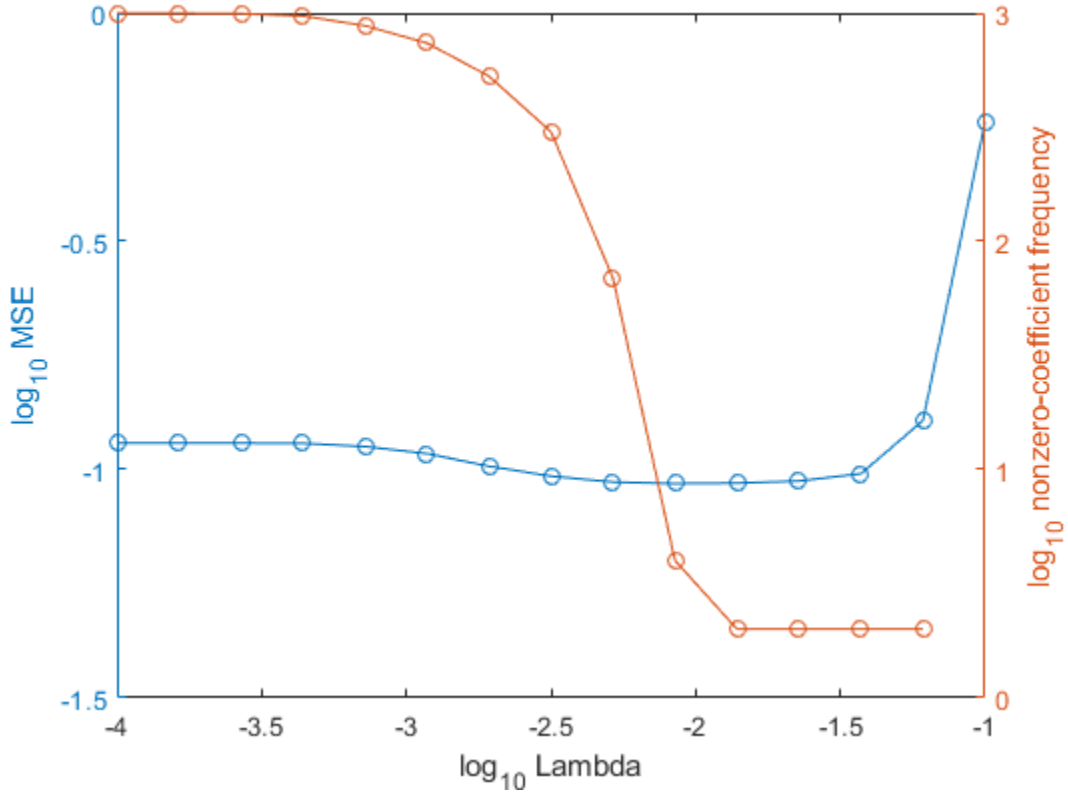
Higher values of `Lambda` lead to predictor variable sparsity, which is a good quality of a regression model. Retrain the model using the entire data set and all options used previously, except the data-partition specification. Determine the number of nonzero coefficients per model.

```
Mdl = fitrlinear(X,Y,'ObservationsIn','columns','Lambda',Lambda,...
    'Solver','sparsa','Regularization','lasso');
numNZCoeff = sum(Mdl.Beta~=0);
```

In the same figure, plot the MSE and frequency of nonzero coefficients for each regularization strength. Plot all variables on the log scale.

```
figure;
[h,hL1,hL2] = plotyy(log10(Lambda),log10(mse),...
    log10(Lambda),log10(numNZCoeff));
hL1.Marker = 'o';
```

```
hL2.Marker = 'o';
ylabel(h(1),'log_{10} MSE')
ylabel(h(2),'log_{10} nonzero-coefficient frequency')
xlabel('log_{10} Lambda')
hold off
```



Select the index or indices of `Lambda` that balance minimal classification error and predictor-variable sparsity (for example, `Lambda(11)`).

```
idx = 11;
MdlFinal = selectModels(Mdl,idx);
```

`MdlFinal` is a trained `RegressionLinear` model object that uses `Lambda(11)` as a regularization strength.

Tip

One way to build several predictive linear regression models is:

1. Hold out a portion of the data for testing.
2. Train a linear regression model using `fitrlinear`. Specify a grid of regularization strengths using the `'Lambda'` name-value pair argument and supply the training data. `fitrlinear` returns one `RegressionLinear` model object, but it contains a model for each regularization strength.

3. To determine the quality of each regularized model, pass the returned model object and the held-out data to, for example, `loss`.
4. Identify the indices (`idx`) of a satisfactory subset of regularized models, and then pass the returned model and the indices to `selectModels`. `selectModels` returns one `RegressionLinear` model object, but it contains `numel(idx)` regularized models.
5. To predict class labels for new data, pass the data and the subset of regularized models to `predict`.

See Also

`RegressionLinear` | `fitrlinear` | `loss` | `predict`

Introduced in R2016a

SemiSupervisedGraphModel

Semi-supervised graph-based model for classification

Description

You can use a semi-supervised graph-based method to label unlabeled data by using the `fitsemigraph` function. The resulting `SemiSupervisedGraphModel` object contains the fitted labels for the unlabeled observations (`FittedLabels`) and their scores (`LabelScores`). You can also use the `SemiSupervisedGraphModel` object as a classifier, trained on both the labeled and unlabeled data, to classify new data by using the `predict` function.

Creation

Create a `SemiSupervisedGraphModel` object by using `fitsemigraph`.

Properties

`FittedLabels` — Labels fitted to unlabeled data
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Labels fitted to the unlabeled data, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `FittedLabels` has the same data type as the class labels in the response variable in the call to `fitsemigraph`. (The software treats string arrays as cell arrays of character vectors.)

Each row of `FittedLabels` represents the fitted label of the corresponding row of `UnlabeledX` or `UnlabeledTbl`.

For more information on how `fitsemigraph` fits labels, see “Algorithms” on page 33-2062.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`LabelScores` — Scores for fitted labels
numeric matrix

This property is read-only.

Scores for the fitted labels, specified as a numeric matrix. `LabelScores` has size *u*-by-*K*, where *u* is the number of observations (or rows) in the unlabeled data and *K* is the number of classes in `ClassNames`.

`score(u,k)` is the likelihood that the observation `u` belongs to class `k`, where a higher score value indicates a higher likelihood.

For more information on how `fitsemigraph` computes label scores, see “Algorithms” on page 33-2062.

Data Types: `double`

`Method` — Labeling technique
`'labelpropagation'` | `'labelpropagationexact'` | `'labelspreading'` | `'labelspreadingexact'`

This property is read-only.

Labeling technique used to label the unlabeled data, specified as `'labelpropagation'`, `'labelpropagationexact'`, `'labelspreading'`, or `'labelspreadingexact'`.

Data Types: `char`

`CategoricalPredictors` — Categorical predictor indices
positive integer vector | `[]`

This property is read-only.

Categorical predictor indices, specified as a positive integer vector. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used to label the unlabeled data, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. The order of the elements of `ClassNames` determines the order of the classes.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`PredictorNames` — Predictor variable names
cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the predictor data.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

Object Functions

predict Label new data using semi-supervised graph-based classifier

Examples

Fit Labels to Unlabeled Data

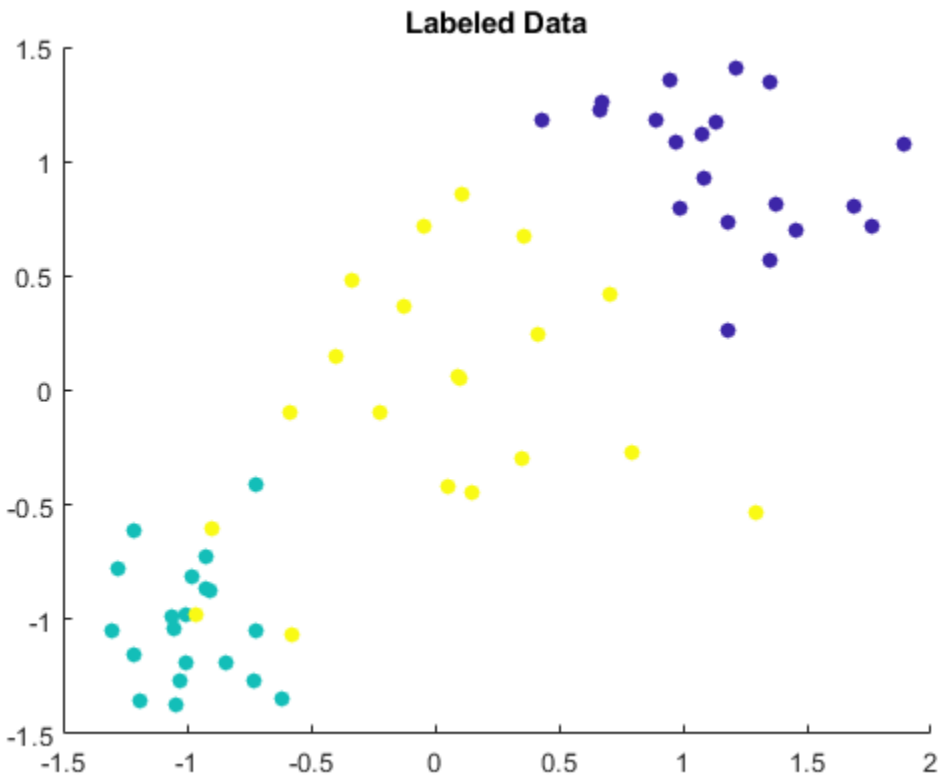
Fit labels to unlabeled data by using a semi-supervised graph-based method.

Randomly generate 60 observations of labeled data, with 20 observations in each of three classes.

```
rng('default') % For reproducibility

labeledX = [randn(20,2)*0.25 + ones(20,2);
            randn(20,2)*0.25 - ones(20,2);
            randn(20,2)*0.5];
Y = [ones(20,1); ones(20,1)*2; ones(20,1)*3];
```

Visualize the labeled data by using a scatter plot. Observations in the same class have the same color. Notice that the data is split into three clusters with very little overlap.

```
scatter(labeledX(:,1),labeledX(:,2),[],Y,'filled')
title('Labeled Data')
```



Randomly generate 300 additional observations of unlabeled data, with 100 observations per class. For the purposes of validation, keep track of the true labels for the unlabeled data.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
trueLabels = [ones(100,1); ones(100,1)*2; ones(100,1)*3];
```

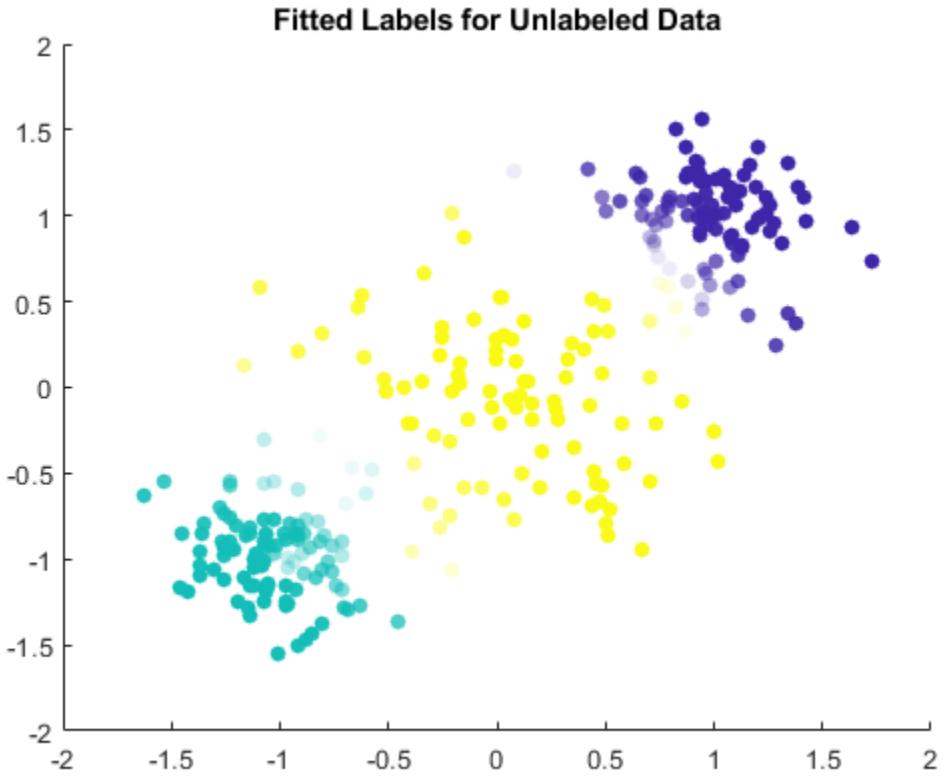
Fit labels to the unlabeled data by using a semi-supervised graph-based method. The function `fitsemigraph` returns a `SemiSupervisedGraphModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemigraph(labeledX,Y,unlabeledX)
```

```
Mdl =
  SemiSupervisedGraphModel with properties:

             FittedLabels: [300x1 double]
              LabelScores: [300x3 double]
                ClassNames: [1 2 3]
             ResponseName: 'Y'
    CategoricalPredictors: []
                   Method: 'labelpropagation'


  Properties, Methods
```

Visualize the fitted label results by using a scatter plot. Use the fitted labels to set the color of the observations, and use the maximum label scores to set the transparency of the observations. Observations with less transparency are labeled with greater confidence. Notice that observations that lie closer to the cluster boundaries are labeled with more uncertainty.

```
maxLabelScores = max(Mdl.LabelScores,[],2);
rescaledScores = rescale(maxLabelScores,0.05,0.95);
scatter(unlabeledX(:,1),unlabeledX(:,2),[],Mdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledScores);
title('Fitted Labels for Unlabeled Data')
```



Determine the accuracy of the labeling by using the true labels for the unlabeled data.

```
numWrongLabels = sum(trueLabels ~= Mdl.FittedLabels)
```

```
numWrongLabels = 10
```

Only 10 of the 300 observations in `unlabeledX` are mislabeled.

Classify New Data Using Model Trained on Labeled and Unlabeled Data

Use both labeled and unlabeled data to train a `SemiSupervisedGraphModel` object. Label new data using the trained model.

Randomly generate 15 observations of labeled data, with 5 observations in each of three classes.

```
rng('default') % For reproducibility
labeledX = [randn(5,2)*0.25 + ones(5,2);
            randn(5,2)*0.25 - ones(5,2);
            randn(5,2)*0.5];
Y = [ones(5,1); ones(5,1)*2; ones(5,1)*3];
```

Randomly generate 300 additional observations of unlabeled data, with 100 observations per class.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
```

Fit labels to the unlabeled data by using a semi-supervised graph-based method. Specify label spreading as the labeling algorithm, and use an automatically selected kernel scale factor. The function `fitsemigraph` returns a `SemiSupervisedGraphModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemigraph(labeledX,Y,unlabeledX,'Method','labelspreading', ...
    'KernelScale','auto')
```

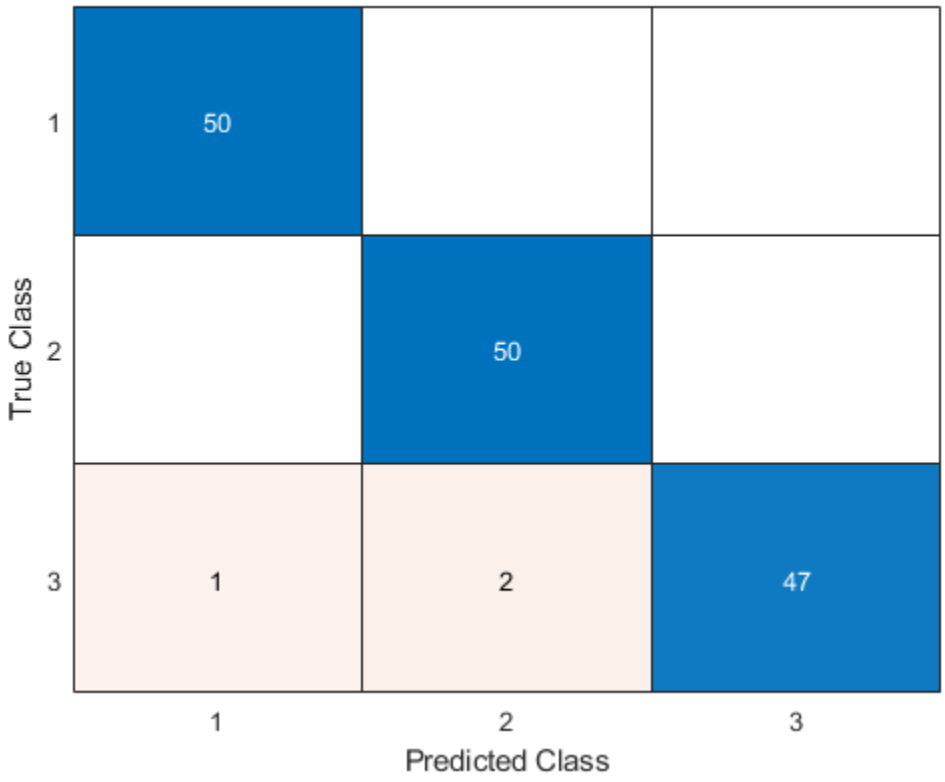
```
Mdl =
  SemiSupervisedGraphModel with properties:

             FittedLabels: [300x1 double]
              LabelScores: [300x3 double]
                ClassNames: [1 2 3]
             ResponseName: 'Y'
    CategoricalPredictors: []
                   Method: 'labelspreading'


  Properties, Methods
```

Randomly generate 150 observations of new data, with 50 observations per class. For the purposes of validation, keep track of the true labels for the new data.

```
newX = [randn(50,2)*0.25 + ones(50,2);
        randn(50,2)*0.25 - ones(50,2);
        randn(50,2)*0.5];
trueLabels = [ones(50,1); ones(50,1)*2; ones(50,1)*3];
```

Predict the labels for the new data by using the `predict` function of the `SemiSupervisedGraphModel` object. Compare the true labels to the predicted labels by using a confusion matrix.

```
predictedLabels = predict(Mdl,newX);
confusionchart(trueLabels,predictedLabels)
```



Only 3 of the 150 observations in `newX` are mislabeled.

See Also

`SemiSupervisedSelfTrainingModel` | `fitsemigraph` | `fitsemiself` | `predict`

Topics

“Label Data Using Semi-Supervised Learning Techniques” on page 18-250

Introduced in R2020b

SemiSupervisedSelfTrainingModel

Semi-supervised self-trained model for classification

Description

You can use a semi-supervised self-training method to label unlabeled data by using the `fitsemiself` function. The resulting `SemiSupervisedSelfTrainingModel` object contains the fitted labels for the unlabeled observations (`FittedLabels`) and their scores (`LabelScores`). You can also use the `SemiSupervisedSelfTrainingModel` object as a classifier, trained on both the labeled and unlabeled data, to classify new data by using the `predict` function.

Creation

Create a `SemiSupervisedSelfTrainingModel` object by using `fitsemiself`.

Properties

`FittedLabels` — Labels fitted to unlabeled data
categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Labels fitted to the unlabeled data, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `FittedLabels` has the same data type as the class labels in the response variable in the call to `fitsemiself`. (The software treats string arrays as cell arrays of character vectors.)

Each row of `FittedLabels` represents the fitted label of the corresponding observation of `UnlabeledX` or `UnlabeledTbl`.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`LabelScores` — Scores for fitted labels
numeric matrix

This property is read-only.

Scores for the fitted labels, specified as a numeric matrix. `LabelScores` has size *u*-by-*K*, where *u* is the number of observations in the unlabeled data and *K* is the number of classes in `ClassNames`.

`score(u,k)` is the likelihood that the observation `u` belongs to class `k`, where a higher score value indicates a higher likelihood. The range of score values depends on the underlying classifier `Learner`.

Data Types: `single` | `double`

`Learner` — Underlying classifier
classification model object

This property is read-only.

Underlying classifier, specified as a classification model object. `fitsemiself` uses this classifier in a loop to label and score the unlabeled data. You can use dot notation to display the parameter and hyperparameter values of the underlying classifier.

For example, if you specify `'Learner','svm'` in the call to `fitsemiself`, then you can enter `Mdl.Learner.KernelParameters` to display the kernel parameters of the final support vector machine (SVM) model trained on both the labeled and unlabeled data.

Note Because the `Mdl.Learner` model has some limitations (for example, lack of support for tabular data), avoid using it directly with its object functions, such as `loss` and `predict`. To predict on new data, use the `predict` object function of `SemiSupervisedSelfTrainingModel`.

`CategoricalPredictors` — Categorical predictor indices

positive integer vector | `[]`

This property is read-only.

Categorical predictor indices, specified as a positive integer vector. Assuming that the predictor data contains observations in rows, `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `double`

`ClassNames` — Unique class labels

categorical array | character array | logical vector | numeric vector | cell array of character vectors

This property is read-only.

Unique class labels used to label the unlabeled data, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. The order of the elements of `ClassNames` determines the order of the classes.

Data Types: `single` | `double` | `logical` | `char` | `cell` | `categorical`

`PredictorNames` — Predictor variable names

cell array of character vectors

This property is read-only.

Predictor variable names, specified as a cell array of character vectors. The order of the elements of `PredictorNames` corresponds to the order in which the predictor names appear in the predictor data.

Data Types: `cell`

`ResponseName` — Response variable name

character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

Object Functions

predict Label new data using semi-supervised self-trained classifier

Examples

Fit Labels to Unlabeled Data

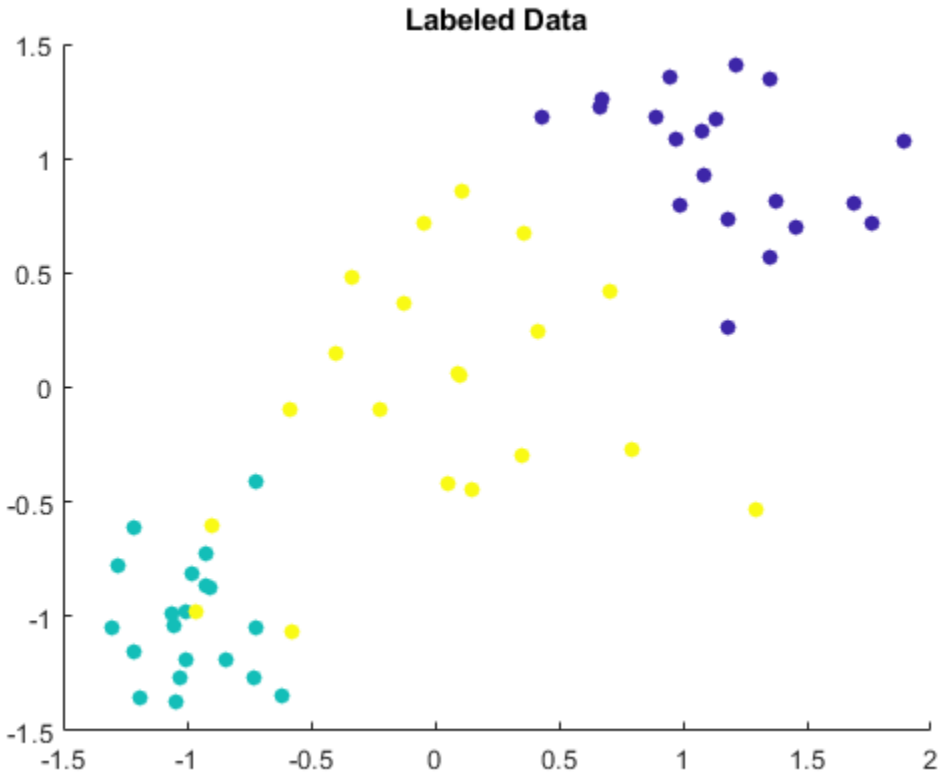
Fit labels to unlabeled data by using a semi-supervised self-training method.

Randomly generate 60 observations of labeled data, with 20 observations in each of three classes.

```
rng('default') % For reproducibility

labeledX = [randn(20,2)*0.25 + ones(20,2);
            randn(20,2)*0.25 - ones(20,2);
            randn(20,2)*0.5];
Y = [ones(20,1); ones(20,1)*2; ones(20,1)*3];
```

Visualize the labeled data by using a scatter plot. Observations in the same class have the same color. Notice that the data is split into three clusters with very little overlap.

```
scatter(labeledX(:,1),labeledX(:,2),[],Y,'filled')
title('Labeled Data')
```



Randomly generate 300 additional observations of unlabeled data, with 100 observations per class. For the purposes of validation, keep track of the true labels for the unlabeled data.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
trueLabels = [ones(100,1); ones(100,1)*2; ones(100,1)*3];
```

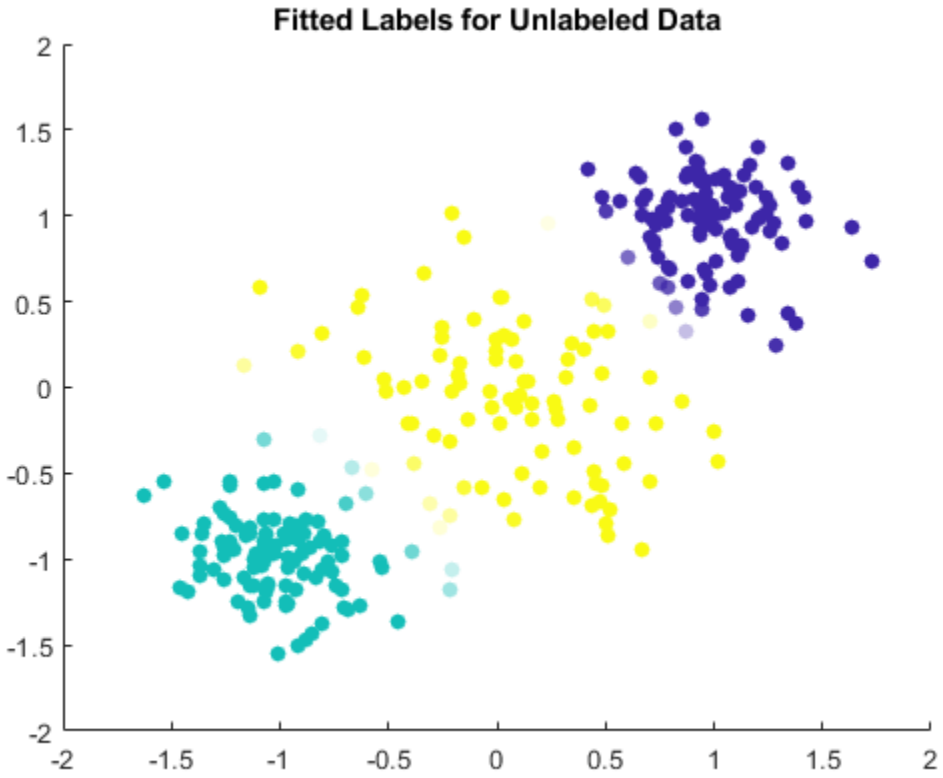
Fit labels to the unlabeled data by using a semi-supervised self-training method. The function `fitsemiself` returns a `SemiSupervisedSelfTrainingModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemiself(labeledX,Y,unlabeledX)
```

```
Mdl =
  SemiSupervisedSelfTrainingModel with properties:

             FittedLabels: [300x1 double]
              LabelScores: [300x3 double]
                ClassNames: [1 2 3]
             ResponseName: 'Y'
    CategoricalPredictors: []
                  Learner: [1x1 classreg.learning.classif.CompactClassificationECOC]


  Properties, Methods
```

Visualize the fitted label results by using a scatter plot. Use the fitted labels to set the color of the observations, and use the maximum label scores to set the transparency of the observations. Observations with less transparency are labeled with greater confidence. Notice that observations that lie closer to the cluster boundaries are labeled with more uncertainty.

```
maxLabelScores = max(Mdl.LabelScores,[],2);
rescaledScores = rescale(maxLabelScores,0.05,0.95);
scatter(unlabeledX(:,1),unlabeledX(:,2),[],Mdl.FittedLabels,'filled', ...
    'MarkerFaceAlpha','flat','AlphaData',rescaledScores);
title('Fitted Labels for Unlabeled Data')
```



Determine the accuracy of the labeling by using the true labels for the unlabeled data.

```
numWrongLabels = sum(trueLabels ~= Mdl.FittedLabels)
```

```
numWrongLabels = 8
```

Only 8 of the 300 observations in `unlabeledX` are mislabeled.

Classify New Data Using Model Trained on Labeled and Unlabeled Data

Use both labeled and unlabeled data to train a `SemiSupervisedSelfTrainingModel` object. Label new data using the trained model.

Randomly generate 15 observations of labeled data, with 5 observations in each of three classes.

```
rng('default') % For reproducibility
labeledX = [randn(5,2)*0.25 + ones(5,2);
            randn(5,2)*0.25 - ones(5,2);
            randn(5,2)*0.5];
Y = [ones(5,1); ones(5,1)*2; ones(5,1)*3];
```

Randomly generate 300 additional observations of unlabeled data, with 100 observations per class.

```
unlabeledX = [randn(100,2)*0.25 + ones(100,2);
              randn(100,2)*0.25 - ones(100,2);
              randn(100,2)*0.5];
```

Fit labels to the unlabeled data by using a semi-supervised self-training method. The function `fitsemiself` returns a `SemiSupervisedSelfTrainingModel` object whose `FittedLabels` property contains the fitted labels for the unlabeled data and whose `LabelScores` property contains the associated label scores.

```
Mdl = fitsemiself(labeledX,Y,unlabeledX)
```

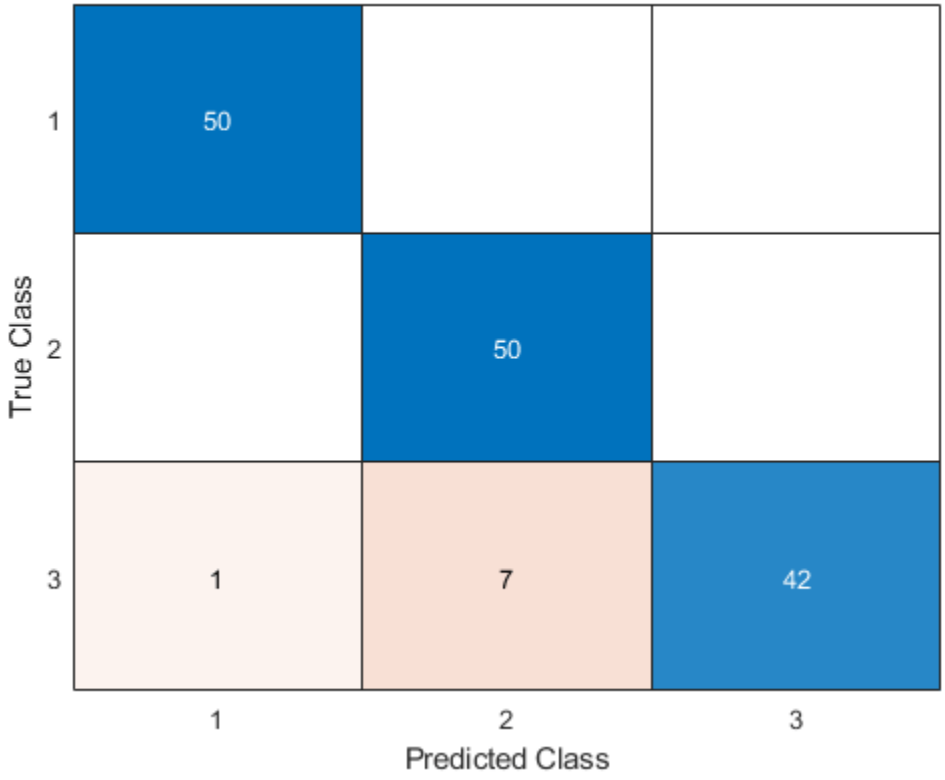
```
Mdl =
  SemiSupervisedSelfTrainingModel with properties:

             FittedLabels: [300x1 double]
              LabelScores: [300x3 double]
                ClassNames: [1 2 3]
             ResponseName: 'Y'
    CategoricalPredictors: []
                  Learner: [1x1 classreg.learning.classif.CompactClassificationECOC]


  Properties, Methods
```

Randomly generate 150 observations of new data, with 50 observations per class. For the purposes of validation, keep track of the true labels for the new data.

```
newX = [randn(50,2)*0.25 + ones(50,2);
        randn(50,2)*0.25 - ones(50,2);
        randn(50,2)*0.5];
trueLabels = [ones(50,1); ones(50,1)*2; ones(50,1)*3];
```

Predict the labels for the new data by using the `predict` function of the `SemiSupervisedSelfTrainingModel` object. Compare the true labels to the predicted labels by using a confusion matrix.

```
predictedLabels = predict(Mdl,newX);
confusionchart(trueLabels,predictedLabels)
```



Only 8 of the 150 observations in `newX` are mislabeled.

See Also

`SemiSupervisedGraphModel` | `fitsemigraph` | `fitsemiself` | `predict`

Topics

“Label Data Using Semi-Supervised Learning Techniques” on page 18-250

Introduced in R2020b

sequentialfs

Sequential feature selection using custom criterion

Syntax

```
inmodel = sequentialfs(fun,X,y)
inmodel = sequentialfs(fun,X,Y,Z,...)
[inmodel,history] = sequentialfs(fun,X,...)
[] = sequentialfs(...,param1,val1,param2,val2,...)
```

Description

`inmodel = sequentialfs(fun,X,y)` selects a subset of features from the data matrix `X` that best predict the data in `y` by sequentially selecting features until there is no improvement in prediction. Rows of `X` correspond to observations; columns correspond to variables or features. `y` is a column vector of response values or class labels for each observation in `X`. `X` and `y` must have the same number of rows. `fun` is a function handle to a function that defines the criterion used to select features and to determine when to stop. The output `inmodel` is a logical vector indicating which features are finally chosen.

Starting from an empty feature set, `sequentialfs` creates candidate feature subsets by sequentially adding each of the features not yet selected. For each candidate feature subset, `sequentialfs` performs 10-fold cross-validation by repeatedly calling `fun` with different training subsets of `X` and `y`, `XTRAIN` and `ytrain`, and test subsets of `X` and `y`, `XTEST` and `ytest`, as follows:

```
criterion = fun(XTRAIN,ytrain,XTEST,ytest)
```

`XTRAIN` and `ytrain` contain the same subset of rows of `X` and `Y`, while `XTEST` and `ytest` contain the complementary subset of rows. `XTRAIN` and `XTEST` contain the data taken from the columns of `X` that correspond to the current candidate feature set.

Each time it is called, `fun` must return a scalar value `criterion`. Typically, `fun` uses `XTRAIN` and `ytrain` to train or fit a model, then predicts values for `XTEST` using that model, and finally returns some measure of distance, or loss, of those predicted values from `ytest`. In the cross-validation calculation for a given candidate feature set, `sequentialfs` sums the values returned by `fun` and divides that sum by the total number of test observations. It then uses that mean value to evaluate each candidate feature subset.

Typical loss measures include sum of squared errors for regression models (`sequentialfs` computes the mean-squared error in this case), and the number of misclassified observations for classification models (`sequentialfs` computes the misclassification rate in this case).

Note `sequentialfs` divides the sum of the values returned by `fun` across all test sets by the total number of test observations. Accordingly, `fun` should not divide its output value by the number of test observations.

After computing the mean `criterion` values for each candidate feature subset, `sequentialfs` chooses the candidate feature subset that minimizes the mean criterion value. This process continues until adding more features does not decrease the criterion.

`inmodel = sequentialfs(fun,X,Y,Z,...)` allows any number of input variables `X`, `Y`, `Z`, ... `sequentialfs` chooses features (columns) only from `X`, but otherwise imposes no interpretation on `X`, `Y`, `Z`, ... All data inputs, whether column vectors or matrices, must have the same number of rows. `sequentialfs` calls `fun` with training and test subsets of `X`, `Y`, `Z`, ... as follows:

```
criterion = fun(XTRAIN,YTRAIN,ZTRAIN,...,
                XTEST,YTEST,ZTEST,...)
```

`sequentialfs` creates `XTRAIN`, `YTRAIN`, `ZTRAIN`, ... , `XTEST`, `YTEST`, `ZTEST`, ... by selecting subsets of the rows of `X`, `Y`, `Z`, ... `fun` must return a scalar value `criterion`, but may compute that value in any way. Elements of the logical vector `inmodel` correspond to columns of `X` and indicate which features are finally chosen.

`[inmodel,history] = sequentialfs(fun,X,...)` returns information on which feature is chosen at each step. `history` is a scalar structure with the following fields:

- `Crit` — A vector containing the criterion values computed at each step.
- `In` — A logical matrix in which row `i` indicates the features selected at step `i`.

`[] = sequentialfs(...,param1,val1,param2,val2,...)` specifies optional parameter name/value pairs from the following table.

Parameter	Value
`'cv'`	The validation method used to compute the criterion for each candidate feature subset. • When the value is a positive integer `k`, `sequentialfs` uses `k`-fold cross-validation without stratification. • When the value is an object of the `cvpartition` class, other forms of cross-validation can be specified. • When the value is `'resubstitution'`, the original data are passed to `fun` as both the training and test data to compute the criterion. • When the value is `'none'`, `sequentialfs` calls `fun` as `criterion = fun(X,Y,Z,...)`, without separating test and training sets. The default value is `10`, that is, 10-fold cross-validation without stratification. So-called wrapper methods use a function `fun` that implements a learning algorithm. These methods usually apply cross-validation to select features. So-called filter methods use a function `fun` that measures characteristics of the data (such as correlation) to select features.
`'mcreps'`	A positive integer indicating the number of Monte-Carlo repetitions for cross-validation. The default value is `1`. The value must be `1` if the value of `'cv'` is `'resubstitution'` or `'none'`.
`'direction'`	The direction of the sequential search. The default is `'forward'`. A value of `'backward'` specifies an initial candidate set including all features and an algorithm that removes features sequentially until the criterion increases.

Parameter	Value
'keepin'	A logical vector or a vector of column numbers specifying features that must be included. The default is empty.
'keepout'	A logical vector or a vector of column numbers specifying features that must be excluded. The default is empty.
'nfeatures'	The number of features at which sequentialfs should stop. inmodel includes exactly this many features. The default value is empty, indicating that sequentialfs should stop when a local minimum of the criterion is found. A nonempty value overrides values of 'MaxIter' and 'TolFun' in 'options'.
'nullmodel'	A logical value, indicating whether or not the null model (containing no features from X) should be included in feature selection and in the history output. The default is false.
'options'	Options structure for the iterative sequential search algorithm, as created by statset. sequentialfs uses the following statset parameters: • Display — Amount of information displayed by the algorithm. The default is 'off'. • MaxIter — Maximum number of iterations allowed. The default is Inf. • TolFun — Termination tolerance for the objective function value. The default is 1e-6 if 'direction' is 'forward'; 0 if 'direction' is 'backward'. • TolTypeFun — Use absolute or relative objective function tolerances. The default is 'rel'. • UseParallel — Set to true to compute in parallel. Default is false. • UseSubstreams — Set to true to compute in parallel in a reproducible fashion. Default is false. To compute reproducibly, set Streams to a type allowing substreams: 'mlfg6331_64' or 'mrg32k3a'. • Streams — A RandStream object or cell array consisting of one such object. If you do not specify Streams, sequentialfs uses the default stream. To compute in parallel, you need Parallel Computing Toolbox.

Examples

Perform sequential feature selection for classification of noisy features:

```
load fisheriris
rng('default') % For reproducibility
X = randn(150,10);
X(:,[1 3 5 7])= meas;
y = species;
```

```
c = cvpartition(y,'k',10);
opts = statset('Display','iter');
fun = @(XT,yT,Xt,yt)loss(fitcecoc(XT,yT),Xt,yt);

[fs,history] = sequentialfs(fun,X,y,'cv',c,'options',opts)

Start forward sequential feature selection:
Initial columns included:  none
Columns that can not be included:  none
Step 1, added column 5, criterion value 0.00266667
Step 2, added column 7, criterion value 0.00222222
Step 3, added column 1, criterion value 0.00177778
Step 4, added column 3, criterion value 0.000888889
Final columns included:  1 3 5 7

fs =

  1×10 logical array

   1   0   1   0   1   0   1   0   0   0


history =

  struct with fields:

      In: [4×10 logical]
    Crit: [0.0027 0.0022 0.0018 8.8889e-04]

history.In

ans =

  4×10 logical array

   0   0   0   0   1   0   0   0   0   0
   0   0   0   0   1   0   1   0   0   0
   1   0   0   0   1   0   1   0   0   0
   1   0   1   0   1   0   1   0   0   0
```

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`crossval` | `cvpartition` | `fscmrmr` | `fscnca` | `fsrnca` | `fsulaplacian` | `relieff` | `statset` | `stepwiselm`

Topics

“Introduction to Feature Selection” on page 15-49
“Select Subset of Features with Comparative Predictive Power” on page 15-61
“Selecting Features for Classifying High-dimensional Data” on page 15-171

Introduced in R2008a

set

Class: `dataset`

(Not Recommended) Set and display dataset array properties

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
set(A)
set(A,PropertyName)
A = set(A,PropertyName,PropertyValue,...)
B = set(A,PropertyName,value)
```

Description

`set(A)` displays all properties of the dataset array `A` and their possible values.

`set(A,PropertyName)` displays possible values for the property specified by *`PropertyName`*.

`A = set(A,PropertyName,PropertyValue,...)` sets property name/value pairs.

`B = set(A,PropertyName,value)` returns a dataset array `B` that is a copy of `A`, but with the property `'PropertyName'` set to the value `value`.

Note Using `set(A,'PropertyName',value)` without assigning to a variable does not modify A's properties. Use `A = set(A,'PropertyName',value)` to modify `A`.

Examples

Create a dataset array from Fisher's iris data and add a description:

```
load fisheriris
NumObs = size(meas,1);
NameObs = strcat({'Obs'},num2str((1:NumObs)','%-d'));
iris = dataset({nominal(species),'species'},...
               {meas,'SL','SW','PL','PW'},...
               'ObsNames',NameObs);
iris = set(iris,'Description','Fisher''s Iris Data');
get(iris)
   Description: 'Fisher's Iris Data'
   Units: {}
   DimNames: {'Observations' 'Variables'}
   UserData: []
   ObsNames: {150x1 cell}
   VarNames: {'species' 'SL' 'SW' 'PL' 'PW'}
```

See Also

get | summary

setDefaultYfit

Class: `CompactTreeBagger`

Set default value for `predict`

Syntax

```
B = setDefaultYfit(B,Yfit)
```

Description

`B = setDefaultYfit(B,Yfit)` sets the default prediction for ensemble `B` to `Yfit`. The default prediction must be specified as a character vector or string scalar for classification or as a numeric scalar for regression. This setting controls what predicted value `CompactTreeBagger` returns when no prediction is possible, for example when the `predict` method needs to predict for an observation which has only false values in the matrix supplied through `'UseInstanceForTree'` argument.

See Also

`predict`

setdiff

Class: `dataset`

(Not Recommended) Set difference for dataset array observations

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = setdiff(A,B)
C = setdiff(A,B,vars)
C = setxor(A,B,vars,setOrder)
[C,iA] = setxor( ___ )
```

Description

`C = setdiff(A,B)` for `dataset` arrays `A` and `B` returns the set of observations that are in `A` but not `B`, with repetitions removed. The observations in the dataset array `C` are sorted.

`C = setdiff(A,B,vars)` returns the set of observations that are in `A` but not `B`, considering only the variables specified in `vars`, with repetitions removed. The observations in the dataset array `C` are sorted by these variables. The values for variables not specified in `vars` for each observation in `C` are taken from the corresponding observation in `A`. If there are multiple observations in `A` that correspond to an observation in `C`, those values are taken from the first occurrence.

`C = setxor(A,B,vars,setOrder)` returns the observations in `C` in the order specified by `setOrder`.

`[C,iA] = setxor( ___ )` also returns the index vector `iA` such that `C = A(iA,:)`. If there are repeated observations in `A`, then `setxor` returns the index of the first occurrence. You can use any of the previous input arguments.

Input Arguments

`A,B`

Input dataset arrays.

`vars`

String array or cell array of character vectors containing variable names, or a vector of integers containing variable column numbers. `vars` indicates the variables that `setdiff` considers.

Specify `vars` as `[]` to use its default value of all variables.

`setOrder`

Flag indicating the sorting order for the observations in `C`. The possible values of `setOrder` are:

`'sorted'`	Observations in `C` are in sorted order (default).
`'stable'`	Observations in `C` are in the same order that they appear in `A`.

Output Arguments

`C`

Dataset array with the observations that are in `A` but not `B`, with repetitions removed. `C` is in sorted order (by default), or the order specified by `setOrder`.

`iA`

Index vector, indicating the observations from `A` that are in `C`. The vector `iA` contains the index to the first occurrence of any repeated observations in `A`.

Examples

Set Difference of Two Dataset Arrays

Create a scalar structure array, and then convert it into two dataset arrays.

```
S(1,1).Name = 'CLARK';
S(1,1).Gender = 'M';
S(1,1).SystolicBP = 124;
S(1,1).DiastolicBP = 93;

S(2,1).Name = 'BROWN';
S(2,1).Gender = 'F';
S(2,1).SystolicBP = 122;
S(2,1).DiastolicBP = 80;

S(3,1).Name = 'MARTIN';
S(3,1).Gender = 'M';
S(3,1).SystolicBP = 130;
S(3,1).DiastolicBP = 92;

A = struct2dataset(S(1:2));
B = struct2dataset(S(2:3));
```

The intersection of `A` and `B` is the second observation, with last name `BROWN`.

Return the set difference of `A` and `B`.

```
[C,iA] = setdiff(A,B)
```

```
C =
    Name             Gender       SystolicBP    DiastolicBP
    {'CLARK'}        {'M'}        124           93


iA = 1
```

The first observation in `A` is not present in `B`.

See Also

`dataset` | `intersect` | `ismember` | `setxor` | `sortrows` | `union` | `unique`

Topics

“Dataset Arrays” on page 2-112

setlabels

(Not Recommended) Assign labels to levels of nominal or ordinal arrays

Note The `nominal` and `ordinal` array data types are not recommended. To represent ordered and unordered discrete, nonnumeric data, use the "Categorical Arrays" data type instead.

Syntax

```
B = setlabels(A,labels)
B = setlabels(A,labels,levels)
```

Description

`B = setlabels(A,labels)` returns a `nominal` or `ordinal` array object of the same type as `A`, but with levels labeled in the order specified by `labels`.

`B = setlabels(A,labels,levels)` labels only the levels specified in `levels`.

Input Arguments

`A` — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, specified as a `nominal` or `ordinal` array object created with `nominal` or `ordinal`.

`labels` — Labels to assign

string array | cell array of character vectors | 2-D character matrix

Labels to assign to levels, specified as a string array, cell array of character vectors, or 2-D character matrix.

Data Types: `char` | `string` | `cell`

`levels` — Levels to assign labels

string array | cell array of character vectors | 2-D character matrix

Level to assign labels to, specified as a string array, cell array of character vectors, or 2-D character matrix.

Data Types: `char` | `string` | `cell`

Output Arguments

`B` — Nominal or ordinal array

`nominal` array | `ordinal` array

Nominal or ordinal array, returned as a `nominal` or `ordinal` array object.

See Also

`getlabels` | `nominal` | `ordinal`

Topics

“Change Category Labels” on page 2-7

Introduced in R2007a

setxor

Class: `dataset`

(Not Recommended) Set exclusive or for dataset array observations

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = setxor(A,B)
C = setxor(A,B,vars)
C = setxor(A,B,vars,setOrder)
[C,iA,iB] = setxor( ___ )
```

Description

`C = setxor(A,B)` for `dataset` arrays `A` and `B` returns the set of observations that are not in the intersection of the two arrays, with repetitions removed. The observations in the dataset array `C` are sorted.

`C = setxor(A,B,vars)` returns the set of observations that are not in the intersection of the two arrays, considering only the variables specified in `vars`, with repetitions removed. The observations in the dataset array `C` are sorted by these variables. The values for variables not specified in `vars` for each observation in `C` are taken from the corresponding observation in `A` or `B`. If there are multiple observations in `A` or `B` that correspond to an observation in `C`, those values are taken from the first occurrence.

`C = setxor(A,B,vars,setOrder)` returns the observations in `C` in the order specified by `setOrder`.

`[C,iA,iB] = setxor( ___ )` also returns index vectors `iA` and `iB` such that `C` is a sorted combination of the values `A(iA,:)` and `B(iB,:)`. If there are repeated observations in `A` or `B`, then `setxor` returns the index of the first occurrence. You can use any of the previous input arguments.

Input Arguments

`A,B`

Input dataset arrays.

`vars`

String array or cell array of character vectors containing variable names, or a vector of integers containing variable column numbers. `vars` indicates the variables in `A` and `B` that `setxor` considers.

Specify `vars` as `[]` to use its default value of all variables.

`setOrder`

Flag indicating the sorting order for the observations in `C`. The possible values of `setOrder` are:

`'sorted'`	Observations in `C` are in sorted order (default).
`'stable'`	Observations in `C` are in the same order that they appear in `A`, then `B`.

Output Arguments

`C`

Dataset array with the observations not in the intersection of `A` and `B`, with repetitions removed. `C` is in sorted order (by default), or the order specified by `setOrder`.

`iA`

Index vector, indicating the observations from `A` that are in `C`. The vector `iA` contains the index to the first occurrence of any repeated observations in `A`.

`iB`

Index vector, indicating the observations from `B` that are in `C`. The vector `iB` contains the index to the first occurrence of any repeated observations in `B`.

Examples

Symmetric Difference of Two Dataset Arrays

Create a scalar structure array, and then convert it into two dataset arrays.

```
S(1,1).Name = 'CLARK';
S(1,1).Gender = 'M';
S(1,1).SystolicBP = 124;
S(1,1).DiastolicBP = 93;

S(2,1).Name = 'BROWN';
S(2,1).Gender = 'F';
S(2,1).SystolicBP = 122;
S(2,1).DiastolicBP = 80;

S(3,1).Name = 'MARTIN';
S(3,1).Gender = 'M';
S(3,1).SystolicBP = 130;
S(3,1).DiastolicBP = 92;

A = struct2dataset(S(1:2));
B = struct2dataset(S(2:3));
```

The intersection of `A` and `B` is the second observation, with last name `BROWN`.

Return the symmetric difference of `A` and `B`.

```
[C,iA,iB] = setxor(A,B);
C
```

```
C =
    Name              Gender       SystolicBP    DiastolicBP
    {'CLARK' }        {'M'}        124           93
    {'MARTIN'}        {'M'}        130           92
```

```
[iA iB]
```

```
ans = 1×2

     1     2
```

The symmetric difference contains the first observation from `A`, and the second observation from `B`.

See Also

`dataset` | `intersect` | `ismember` | `setdiff` | `sortrows` | `union` | `unique`

Topics

“Dataset Arrays” on page 2-112

shapley

Shapley values

Description

The Shapley on page 33-5863 value of a feature for a query point explains the deviation of the prediction for the query point from the average prediction, due to the feature. For each query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average.

You can create a `shapley` object for a machine learning model with a specified query point (`queryPoint`). The software creates an object and computes the Shapley values of all features for the query point.

Use the Shapley values to explain the contribution of individual features to a prediction at the specified query point. Use the `plot` function to create a bar graph of the Shapley values. You can compute the Shapley values for another query point by using the `fit` function.

Creation

Syntax

```
explainer = shapley(blackbox)
explainer = shapley(blackbox,X)

explainer = shapley(___,'QueryPoint',queryPoint)

explainer = shapley(___,Name,Value)
```

Description

`explainer = shapley(blackbox)` creates a `shapley` object using a machine learning model object `blackbox` that contains predictor data. To compute Shapley values, use the `fit` function with `explainer`.

`explainer = shapley(blackbox,X)` creates a `shapley` object using the predictor data in `X`.

`explainer = shapley(___,'QueryPoint',queryPoint)` also computes the Shapley values for the query point `queryPoint` and stores the computed Shapley values in the `ShapleyValues` property of `explainer`. You can specify `queryPoint` in addition to any of the input argument combinations in the previous syntaxes.

`explainer = shapley(___,Name,Value)` specifies additional options using one or more name-value arguments. For example, specify `'UseParallel',true` to compute Shapley values in parallel.

Input Arguments

`blackbox` — Machine learning model to be interpreted
regression model object | classification model object | function handle

Machine learning model to be interpreted, specified as a full or compact regression or classification model object or a function handle.

- Full or compact model object — You can specify a full or compact regression or classification model object, which has a `predict` object function. The software uses the `predict` function to compute Shapley values.
  - If you specify a model object that does not contain predictor data (for example, a compact model), then you must provide the predictor data using `X`.
  - When you train a model, use a numeric matrix or table for the predictor data where rows correspond to individual observations.

  Regression Model Object

Supported Model	Full or Compact Regression Model Object
Ensemble of regression models	`RegressionEnsemble`, `RegressionBaggedEnsemble`, `CompactRegressionEnsemble`
Gaussian kernel regression model using random feature expansion	`RegressionKernel`
Gaussian process regression	`RegressionGP`, `CompactRegressionGP`
Generalized additive model	`RegressionGAM`, `CompactRegressionGAM`
Linear regression for high-dimensional data	`RegressionLinear`
Neural network regression model	`RegressionNeuralNetwork`, `CompactRegressionNeuralNetwork`
Regression tree	`RegressionTree`, `CompactRegressionTree`
Support vector machine regression	`RegressionSVM`, `CompactRegressionSVM`

Classification Model Object

Supported Model	Full or Compact Classification Model Object
Discriminant analysis classifier	`ClassificationDiscriminant`, `CompactClassificationDiscriminant`
Multiclass model for support vector machines or other classifiers	`ClassificationECOC`, `CompactClassificationECOC`
Ensemble of learners for classification	`ClassificationEnsemble`, `CompactClassificationEnsemble`, `ClassificationBaggedEnsemble`
Gaussian kernel classification model using random feature expansion	`ClassificationKernel`
Generalized additive model	`ClassificationGAM`, `CompactClassificationGAM`
k-nearest neighbor classifier	`ClassificationKNN`
Linear classification model	`ClassificationLinear`
Multiclass naive Bayes model	`ClassificationNaiveBayes`, `CompactClassificationNaiveBayes`
Neural network classifier	`ClassificationNeuralNetwork`, `CompactClassificationNeuralNetwork`
Support vector machine classifier for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`

- Function handle — You can specify a function handle that accepts predictor data and returns a column vector containing a prediction for each observation in the predictor data. The prediction is a predicted response for regression or a predicted score of a single class for classification. You must provide the predictor data using `X`.

`X` — Predictor data
numeric matrix | table

Predictor data, specified as a numeric matrix or table. Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables that makes up the columns of `X` must have the same order as the predictor variables that trained `blackbox`, stored in `blackbox.X`.
  - If you trained `blackbox` using a table, then `X` can be a numeric matrix if the table contains all numeric predictor variables.
- For a table:
  - If you trained `blackbox` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those in `Tbl`. However, the column order of `X` does not need to correspond to the column order of `Tbl`.

- If you trained `blackbox` using a numeric matrix, then the predictor names in `blackbox.PredictorNames` and the corresponding predictor variable names in `X` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `X` must be numeric vectors.
- `X` can contain additional variables (response variables, observation weights, and so on), but `shapley` ignores them.
- `shapley` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

If `blackbox` is a model object that does not contain predictor data or a function handle, you must provide `X`. If `blackbox` is a full machine learning model object and you specify this argument, then `shapley` does not use the predictor data in `blackbox`; it uses the specified predictor data only.

Data Types: `single` | `double`

`queryPoint` — Query point
row vector of numeric values | single-row table

Query point at which `shapley` explains a prediction, specified as a row vector of numeric values or a single-row table.

- For a row vector of numeric values:
  - The variables that makes up the columns of `queryPoint` must have the same order as `X` or the predictor variables that trained `blackbox`, stored in `blackbox.X`.
  - If you trained `blackbox` using a table, then `queryPoint` can be a numeric vector if the table contains all numeric variables.
- For a single-row table:
  - If you trained `blackbox` using a table (for example, `Tbl`), then all predictor variables in `queryPoint` must have the same variable names and data types as those in `Tbl`. However, the column order of `queryPoint` does not need to correspond to the column order of `Tbl`.
  - If you trained `blackbox` using a numeric matrix, then the predictor names in `blackbox.PredictorNames` and the corresponding predictor variable names in `queryPoint` must be the same. To specify predictor names during training, use the `'PredictorNames'` name-value argument. All predictor variables in `queryPoint` must be numeric vectors.
  - `queryPoint` can contain additional variables (response variables, observation weights, and so on), but `shapley` ignores them.
  - `shapley` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.

If `queryPoint` contains `NaN`s for continuous predictors and `'Method'` is `'conditional-kernel'`, then the Shapley values (`ShapleyValues`) in the returned object are `NaN`s. Otherwise, `shapley` handles `NaN`s in `queryPoint` in the same way as `blackbox` (the `predict` object function of `blackbox` or the function handle specified by `blackbox`).

Example: `blackbox.X(1,:)` specifies the query point as the first observation of the predictor data in the full machine learning model `blackbox`.

Data Types: `single` | `double` | `table`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `shapley(blackbox,'QueryPoint',q,'Method','conditional-kernel')` creates a `shapley` object and computes the Shapley values for the query point `q` using the extension to the kernelSHAP algorithm.

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `blackbox` uses a subset of input variables as predictors, then the software indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the variable names of the predictor data in the form of a table. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the variable names of the predictor data in the form of a table.
`'all'`	All predictors are categorical.

- If you specify `blackbox` as a function handle, then `shapley` identifies categorical predictors from the predictor data `X`. If the predictor data is in a table, `shapley` assumes that a variable is categorical if it is a logical vector, unordered categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix, `shapley` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.
- If you specify `blackbox` as a regression or classification model object, then `shapley` identifies categorical predictors by using the `CategoricalPredictors` property of the model object.

`shapley` supports an ordered categorical predictor when `blackbox` supports ordered categorical predictors and `'Method'` is `'interventional-kernel'`.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`MaxNumSubsets` — Maximum number of predictor subsets
`min(2^M,1024)` where `M` is the number of predictors (default) | positive integer

Maximum number of predictor subsets to use for Shapley value computation, specified as a positive integer.

For details on how `shapley` chooses the subsets to use, see “Complexity of Computing Shapley Values” on page 18-277.

Example: `'MaxNumSubsets',100`

Data Types: `single` | `double`

`Method` — Shapley value computation algorithm
`'interventional-kernel'` (default) | `'conditional-kernel'`

Shapley value computation algorithm, specified as `'interventional-kernel'` or `'conditional-kernel'`.

- `'interventional-kernel'` (default) — `shapley` uses the kernelSHAP algorithm [1] with an interventional value function.
- `'conditional-kernel'` — `shapley` uses the extension to the kernelSHAP algorithm [2] with a conditional value function.

For details about these algorithms, see “Shapley Value Computation Algorithms” on page 18-272.

Example: `'Method','conditional-kernel'`

Data Types: `char` | `string`

`UseParallel` — Flag to run in parallel
`false` (default) | `true`

Flag to run in parallel, specified as `true` or `false`. If you specify `'UseParallel',true`, the `shapley` function executes for-loop iterations in parallel by using `parfor`. This option requires Parallel Computing Toolbox.

Example: `'UseParallel',true`

Data Types: `logical`

Properties

`BlackboxModel` — Machine learning model to be interpreted
regression model object | classification model object | function handle

This property is read-only.

Machine learning model to be interpreted, specified as a regression or classification model object or a function handle.

The `blackbox` argument sets this property.

`BlackboxFitted` — Prediction for query point computed by machine learning model
scalar

This property is read-only.

Prediction for the query point computed by the machine learning model (`BlackboxModel`), specified as a scalar.

- If `BlackboxModel` is a model object, then `BlackboxFitted` is a predicted response for regression or a classified label for classification.
- If `BlackboxModel` is a function handle, then `BlackboxFitted` is a value returned by the function handle, either a predicted response for regression or a predicted score of a single class for classification.

Data Types: `single` | `double` | `categorical` | `logical` | `char` | `string` | `cell`

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers | `[]`

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

- If you specify `blackbox` using a function handle, then `shapley` identifies categorical predictors from the predictor data `X`. If you specify the `'CategoricalPredictors'` name-value argument, then the argument sets this property.
- If you specify `blackbox` as a regression or classification model object, then `shapley` determines this property by using the `CategoricalPredictors` property of the model object.

`shapley` supports an ordered categorical predictor when `blackbox` supports ordered categorical predictors and `'Method'` is `'interventional-kernel'`.

Data Types: `single` | `double`

`Method` — Shapley value computation algorithm
`'interventional-kernel'` | `'conditional-kernel'`

This property is read-only.

Shapley value computation algorithm, specified as `'interventional-kernel'` or `'conditional-kernel'`.

- `'interventional-kernel'` — `shapley` uses the kernelSHAP algorithm [1] with an interventional value function.
- `'conditional-kernel'` — `shapley` uses the extension to the kernelSHAP algorithm [2] with a conditional value function.

The `'Method'` argument of `shapley` or the `'Method'` argument of `fit` sets this property.

For details about these algorithms, see “Shapley Value Computation Algorithms” on page 18-272.

Data Types: `char` | `string`

`NumSubsets` — Number of predictor subsets
positive integer

This property is read-only.

Number of predictor subsets to use for Shapley value computation, specified as a positive integer.

The 'MaxNumSubsets' argument of shapley or the 'MaxNumSubsets' argument of fit sets this property.

For details on how shapley chooses the subsets to use, see “Complexity of Computing Shapley Values” on page 18-277.

Data Types: single | double

QueryPoint — Query point
row vector of numeric values | single-row table

This property is read-only.

Query point at which shapley explains a prediction using the Shapley values (ShapleyValues), specified as a row vector of numeric values or single-row table.

The queryPoint argument of shapley or the queryPoint argument of fit sets this property.

Data Types: single | double | table

ShapleyValues — Shapley values for query point
table

This property is read-only.

Shapley values for the query point (QueryPoint), specified as a table.

- For regression, the table has two columns. The first column contains the predictor variable names, and the second column contains the Shapley values of the predictors.
- For classification, the table has two or more columns, depending on the number of classes in BlackboxModel. The first column contains the predictor variable names, and the rest of the columns contain the Shapley values of the predictors for each class.

Data Types: table

X — Predictor data
numeric matrix | table

This property is read-only.

Predictor data, specified as a numeric matrix or table.

Each row of X corresponds to one observation, and each column corresponds to one variable.

- If you specify the X argument, then it sets this property.
- If you specify blackbox as a full machine learning model object and do not specify X, then this property value is the predictor data used to train blackbox.

If an observation contains NaNs for continuous predictors and Method is 'conditional-kernel', then shapley does not use the observation for the Shapley value computation. Otherwise, shapley handles NaNs in X in the same way as BlackboxModel (the predict object function of BlackboxModel or the function handle specified by BlackboxModel).

shapley stores all observations, including the rows with missing values, in this property.

Data Types: single | double | table

Object Functions

fit Compute Shapley values for query point
plot Plot Shapley values

Examples

Compute Shapley Values When Creating shapley Object

Train a classification model and create a `shapley` object. When you create a `shapley` object, specify a query point so that the software computes the Shapley values for the query point. Then create a bar graph of the Shapley values by using the object function `plot`.

Load the `CreditRating_Historical` data set. The data set contains customer IDs and their financial ratios, industry labels, and credit ratings.

```
tbl = readtable('CreditRating_Historical.dat');
```

Display the first three rows of the table.

```
head(tbl,3)
```

```
ans=3×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA     Industry    Rating
    _____    _____    _____    _______    ________    _____    ________    ______

    62394    0.013    0.104     0.036      0.447      0.142       3        {'BB'}
    48608    0.232    0.335     0.062      1.969      0.281       8        {'A' }
    42444    0.311    0.367     0.074      1.935      0.366       1        {'A' }
```

Train a blackbox model of credit ratings by using the `fitcecoc` function. Use the variables from the second through seventh columns in `tbl` as the predictor variables. A recommended practice is to specify the class names to set the order the classes.

```
blackbox = fitcecoc(tbl,'Rating', ...
    'PredictorNames',tbl.Properties.VariableNames(2:7), ...
    'CategoricalPredictors','Industry', ...
    'ClassNames',{'AAA' 'AA' 'A' 'BBB' 'BB' 'B' 'CCC'});
```

Create a `shapley` object that explains the prediction for the last observation. Specify a query point so that the software computes Shapley values and stores them in the `ShapleyValues` property.

```
queryPoint = tbl(end,:)
```

```
queryPoint=1×8 table
     ID      WC_TA    RE_TA    EBIT_TA    MVE_BVTD    S_TA    Industry    Rating
    _____    _____    _____    _______    ________    ____    ________    ______

    73104    0.239    0.463     0.065      2.924      0.34       2        {'AA'}
```

```
explainer = shapley(blackbox,'QueryPoint',queryPoint)
```

```
Warning: Computation can be slow because the predictor data has over 1000 observations. Use a sma
```

```
explainer =
  shapley with properties:
```

```
          BlackboxModel: [1x1 ClassificationECOC]
             QueryPoint: [1x8 table]
         BlackboxFitted: {'AA'}
          ShapleyValues: [6x8 table]
             NumSubsets: 64
                      X: [3932x6 table]
    CategoricalPredictors: 6
                 Method: 'interventional-kernel'
```

As the warning message indicates, the computation can be slow because the predictor data has over 1000 observations. For faster computation, use a smaller sample of the training set or specify `'UseParallel'` as `true`.

For a classification model, `shapley` computes Shapley values using the predicted class score for each class. Display the values in the `ShapleyValues` property.

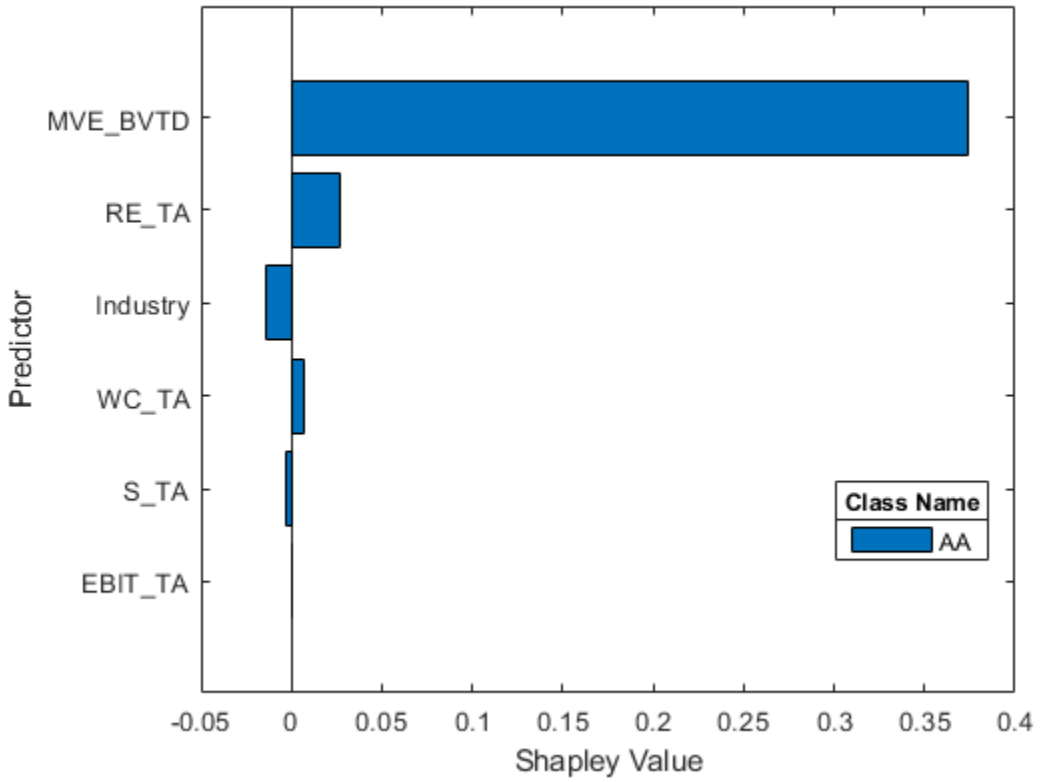
```
explainer.ShapleyValues
```

```
ans=6×8 table
    Predictor        AAA           AA            A            BBB            BB            B
    _________    __________    __________    __________    ___________    ___________    _____

    "WC_TA"         0.014715      0.006439      0.002669     0.00048882     -0.0079015     -0.0
    "RE_TA"         0.047918      0.026904      0.014759     -0.0031481       -0.02512     -0.0
    "EBIT_TA"      0.0003427    0.00015023    0.00011977     3.3904e-05    -0.00018925    -0.000
    "MVE_BVTD"       0.38334       0.37376       0.17563      -0.032136       -0.18729       -0
    "S_TA"        -0.0037303    -0.0026019    -8.9059e-05    -0.00081782    -5.4905e-05      0.0
    "Industry"     -0.028974     -0.013901      0.0010365       0.023298       0.026474       0.0
```

The `ShapleyValues` property contains the Shapley values of all features for each class.

Plot the Shapley values for the predicted class by using the `plot` function. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
f = figure;
plot(explainer)
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The horizontal bar graph shows the Shapley values for all variables, sorted by their absolute values. Each Shapley value explains the deviation of the score for the query point from the average score of the predicted class, due to the corresponding variable.

Create `shapley` Object and Compute Shapley Values Using `fit`

Train a regression model and create a `shapley` object. When you create a `shapley` object, if you do not specify a query point, then the software does not compute Shapley values. Use the object function `fit` to compute the Shapley values for the specified query point. Then create a bar graph of the Shapley values by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on, as well as the response variable `MPG`.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight,MPG);
```

Removing missing values in a training set can help reduce memory consumption and speed up training for the `fitrkernel` function. Remove missing values in `tbl`.

```
tbl = rmmissing(tbl);
```

Train a blackbox model of MPG by using the `fitrkernel` function

```
rng('default') % For reproducibility
mdl = fitrkernel(tbl,'MPG','CategoricalPredictors',[2 5]);
```

Create a `shapley` object. Specify the data set `tbl`, because `mdl` does not contain training data.

```
explainer = shapley(mdl,tbl)
```

```
explainer =
  shapley with properties:

            BlackboxModel: [1x1 RegressionKernel]
               QueryPoint: []
           BlackboxFitted: []
            ShapleyValues: []
               NumSubsets: 64
                        X: [392x7 table]
    CategoricalPredictors: [2 5]
                   Method: 'interventional-kernel'
```

`explainer` stores the training data `tbl` in the `X` property.

Compute the Shapley values of all predictor variables for the first observation in `tbl`.

```
queryPoint = tbl(1,:)
```

```
queryPoint=1×7 table
    Acceleration    Cylinders    Displacement    Horsepower    Model_Year    Weight    MPG
    ____________    _________    ____________    __________    __________    ______    ___

         12             8            307            130            70          3504     18
```

```
explainer = fit(explainer,queryPoint);
```

For a regression model, `shapley` computes Shapley values using the predicted response, and stores them in the `ShapleyValues` property. Display the values in the `ShapleyValues` property.

```
explainer.ShapleyValues
```

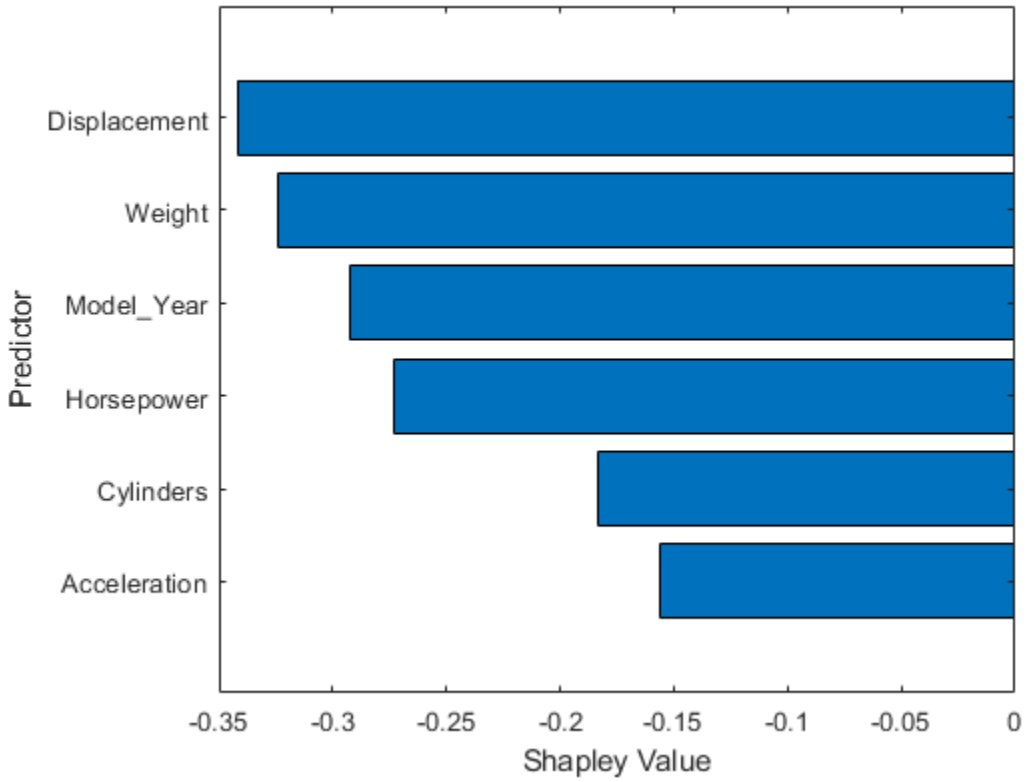
```
ans=6×2 table
      Predictor       ShapleyValue
    _______________    ____________

    "Acceleration"       -0.1561
    "Cylinders"         -0.18306
    "Displacement"      -0.34203
    "Horsepower"        -0.27291
    "Model_Year"         -0.2926
    "Weight"            -0.32402
```

Display the predicted response for the query point, and plot the Shapley values for the query point by using the `plot` function. To display an existing underscore in any predictor name, change the `TickLabelInterpreter` value of the axes to `'none'`.

```
explainer.BlackboxFitted
```

```
ans = 21.0495
```

```
f = figure;
plot(explainer)
f.CurrentAxes.TickLabelInterpreter = 'none';
```



The horizontal bar graph shows the Shapley values for all variables, sorted by their absolute values. Each Shapley value explains the deviation of the prediction for the query point from the average, due to the corresponding variable.

Specify Blackbox Model Using Function Handle

Train a regression model and create a `shapley` object using a function handle to the `predict` function of the model. Use the object function `fit` to compute the Shapley values for the specified query point. Then plot the Shapley values by using the object function `plot`.

Load the `carbig` data set, which contains measurements of cars made in the 1970s and early 1980s.

```
load carbig
```

Create a table containing the predictor variables `Acceleration`, `Cylinders`, and so on.

```
tbl = table(Acceleration,Cylinders,Displacement,Horsepower,Model_Year,Weight);
```

Train a blackbox model of `MPG` by using the `TreeBagger` function.

```
rng('default') % For reproducibility
Mdl = TreeBagger(100,tbl,MPG,'Method','regression','CategoricalPredictors',[2 5]);
```

`shapley` does not support a `TreeBagger` object directly, so you cannot specify the first input argument (blackbox model) of `shapley` as a `TreeBagger` object. Instead, you can use a function handle to the `predict` function. You can also specify options of the `predict` function using name-value arguments of the function.

Create the function handle to the `predict` function of the `TreeBagger` object `Mdl`. Specify the array of tree indices to use as `1:50`.

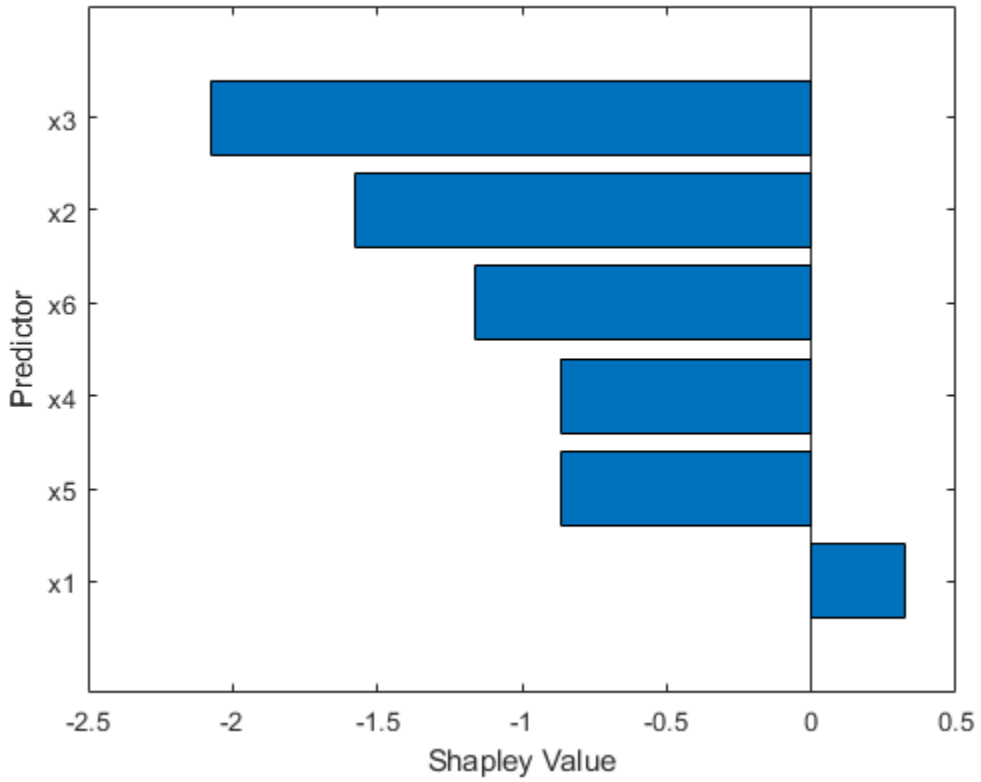
```
f = @(tbl) predict(Mdl,tbl,'Trees',1:50);
```

Create a `shapley` object using the function handle `f`. When you specify a blackbox model as a function handle, you must provide the predictor data. `tbl` includes categorical predictors (`Cylinder` and `Model_Year`) with the `double` data type. By default, `shapley` does not treat variables with the `double` data type as categorical predictors. Specify the second (`Cylinder`) and fifth (`Model_Year`) variables as categorical predictors.

```
explainer = shapley(f,tbl,'CategoricalPredictors',[2 5]);
explainer = fit(explainer,tbl(1,:));
```

Plot the Shapley values.

```
plot(explainer)
```



Display the predictor names in order of importance.

```
tbl.Properties.VariableNames([3 2 6 4 5 1])

ans = 1x6 cell
  Columns 1 through 4

    {'Displacement'}    {'Cylinders'}    {'Weight'}    {'Horsepower'}

  Columns 5 through 6

    {'Model_Year'}    {'Acceleration'}
```

More About

Shapley Values

In game theory, the Shapley value of a player is the average marginal contribution of the player in a cooperative game. In the context of machine learning prediction, the Shapley value of a feature for a query point explains the contribution of the feature to a prediction (response for regression or score of each class for classification) at the specified query point.

The Shapley value corresponds to the deviation of the prediction for the query point from the average prediction, due to the feature. For a query point, the sum of the Shapley values for all features corresponds to the total deviation of the prediction from the average.

For more details, see “Shapley Values for Machine Learning Model” on page 18-272.

References

[1] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[2] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, set the `'UseParallel'` name-value argument to `true` in the call to this function.

For more general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`lime` | `plotPartialDependence`

Topics

“Shapley Values for Machine Learning Model” on page 18-272
“Interpret Machine Learning Models” on page 18-256

Introduced in R2021a

shrink

Prune ensemble

Syntax

```
cmp = shrink(ens)
cmp = shrink(ens,Name,Value)
```

Description

`cmp = shrink(ens)` returns a compact shrunken version of `ens`, a regularized ensemble. `cmp` retains only learners with weights above a threshold.

`cmp = shrink(ens,Name,Value)` returns an ensemble with additional options specified by one or more `Name,Value` pair arguments. You can specify several name-value pair arguments in any order as `Name1,Value1,…,NameN,ValueN`.

Input Arguments

`ens`

A regression ensemble created with `fitrensemble`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`lambda`

Vector of nonnegative regularization parameter values for lasso. If `ens.Regularization` is nonempty (populate it with `regularize`), `shrink` regularizes `ens` using `lambda`. If `ens` contains a `Regularization` structure, you cannot pass `lambda`.

Default: `[]`

`threshold`

Lower cutoff on weights for weak learners, a numeric nonnegative scalar. `shrink` creates `cmp` from those learners with weights above `threshold`.

Default: `0`

`weightcolumn`

Column index of `ens.Regularization.TrainedWeights`, a positive integer. `shrink` creates `cmp` with learner weights from this column.

Default: `1`

Output Arguments

cmp

A regression ensemble of class `CompactRegressionEnsemble`. Use `cmp` for making predictions exactly as you use `ens`, with the `predict` method.

`shrink` orders the members of `cmp` from largest to smallest.

Examples

Shrink Bagged Regression Ensemble

Shrink a 300-member bagged regression ensemble, and view the number of members of the resulting ensemble.

Generate sample data.

```
X = rand(2000,20);
Y = repmat(-1,2000,1);
Y(sum(X(:,1:5),2)>2.5) = 1;
```

Shrink a 300-member bagged regression ensemble using `0.1` for the parameter `lambda`.

```
bag = fitrensemble(X,Y,'Method','Bag','NumLearningCycles',300);
cmp = shrink(bag,'lambda',0.1);
```

View the number of members of the resulting ensemble.

```
cmp.NumTrained
```

```
ans = 94
```

See Also

`cvshrink` | `predict` | `regularize`

Topics

“Ensemble Regularization” on page 18-70

signrank

Wilcoxon signed rank test

Syntax

```
p = signrank(x)
p = signrank(x,y)
p = signrank(x,y,Name,Value)
[p,h] = signrank( ___ )
[p,h,stats] = signrank( ___ )

[ ___ ] = signrank(x,m)
[ ___ ] = signrank(x,m,Name,Value)
```

Description

`p = signrank(x)` returns the *p*-value of a two-sided Wilcoxon signed rank test on page 33-5872.

`signrank` tests the null hypothesis that data in the vector `x` come from a distribution whose median is zero at the 5% significance level. The test assumes that the data in `x` come from a continuous distribution symmetric about its median.

`p = signrank(x,y)` returns the *p*-value of a paired, two-sided test for the null hypothesis that `x – y` comes from a distribution with zero median.

`p = signrank(x,y,Name,Value)` returns the *p*-value for the sign test with additional options specified by one or more `Name`,`Value` pair arguments.

`[p,h] = signrank( ___ )` also returns a logical value indicating the test decision. `h = 1` indicates a rejection of the null hypothesis, and `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level. You can use any of the input arguments in the previous syntaxes.

`[p,h,stats] = signrank( ___ )` also returns the structure `stats` with information about the test statistic.

`[ ___ ] = signrank(x,m)` returns any of the output arguments in the previous syntaxes for the null hypothesis that the data in `x` are observations from a distribution with median `m`.

`[ ___ ] = signrank(x,m,Name,Value)` returns any of the output arguments in the previous syntaxes for the signed rank test with additional options specified by one or more `Name`,`Value` pair arguments.

Examples

Test for Zero Median of a Single Population

Test the hypothesis of zero median.

Generate the sample data.

```
rng('default') % for reproducibility
x = randn(1,25) + 1.30;
```

Test the hypothesis that the data in x has zero median.

```
[p,h] = signrank(x)
```

```
p = 3.2229e-05
```

```
h = logical
   1
```

At the default 5% significance level, the value h = 1 indicates that the test rejects the null hypothesis of zero median.

Test the Median of Differences of Paired Samples

Test the hypothesis of zero median for the difference between paired samples.

Generate the sample data.

```
rng('default') % for reproducibility
x = lognrnd(2,.25,10,1);
y = x + trnd(2,10,1);
```

Test the hypothesis that x – y has zero median.

```
[p,h] = signrank(x,y)
```

```
p = 0.3223
```

```
h = logical
   0
```

The results indicate that the test fails to reject the null hypothesis of zero median in the difference at the default 5% significance level.

Signed Rank Test for Large Samples

Conduct a -sided test on a large sample using approximation.

Load the sample data.

```
load('gradespaired.mat');
```

Test the null hypothesis that the median of the grade differences of students before and after participating in a tutoring program is 0 against the alternate that it is less than 0.

```
[p,h,stats] = signrank(gradespaired(:,1),...
        gradespaired(:,2),'tail','left')
```

```
p = 0.0047
```

```
h = logical
   1


stats = struct with fields:
          zval: -2.5982
    signedrank: 2.0175e+03
```

Because the sample size is greater than 15, `signrank` uses an approximate method to calculate the *p*-value and also returns the value of the *z*-statistic. The value `h = 1` indicates that the test rejects the null hypothesis that there is no difference between the grade medians at the 5% significance level. There is enough statistical evidence to conclude that the median grade before the tutoring program is less than the median grade after the tutoring program.

Repeat the test using the exact method.

```
[p,h,stats] = signrank(gradespaired(:,1),gradespaired(:,2),...
        'tail','left','method','exact')
```

```
p = 0.0045
```

```
h = logical
   1


stats = struct with fields:
    signedrank: 2.0175e+03
```

The results obtained using the approximate method are consistent with the exact method.

Two-Sided Test for the Median of a Single Population

Load the sample data.

```
load mileage
```

The data contains the mileages per gallon for three different types of cars in columns 1 to 3.

Test the hypothesis that the median mileage for the type of cars in the second column differs from 33.

```
[p,h,stats] = signrank(mileage(:,2),33)
```

```
p = 0.0313
```

```
h = logical
   1


stats = struct with fields:
    signedrank: 21
```

At the 5% significance level, the results indicate that the median mileage for the second type of cars differs from 33. Note that `signrank` uses an exact method to calculate the *p*-value for small samples and does not return the *z*-statistic.

Right-Sided Test for the Median of a Single Population

Use the name-value pair arguments in `signrank`.

Load the sample data.

```
load mileage
```

The data contains the mileage per gallon for three different types of cars in columns 1 to 3.

Test the hypothesis that the median mileage for the type of cars in the second row are larger than 33.

```
[p,h,stats] = signrank(mileage(:,2),33,'tail','right')
```

```
p = 0.0156

h = logical
   1


stats = struct with fields:
    signedrank: 21
```

Repeat the same test at the 1% significance level using the approximate method.

```
[p,h,stats] = signrank(mileage(:,2),33,'tail','right',...
'alpha',0.01,'method','approximate')
```

```
p = 0.0180

h = logical
   0


stats = struct with fields:
          zval: 2.0966
    signedrank: 21
```

This result, `h = 0`, indicates that the null hypothesis cannot be rejected at the 1% significance level.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

y — Sample data

vector

Sample data, specified as a vector. `y` must be the same length as `x`.

Data Types: `single` | `double`

`m` — Hypothesized value of the median

scalar

Hypothesized value of the median, specified as a scalar.

Example: `signrank(x,10)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'alpha',0.01,'method','approximate','tail','right'` specifies a right-tailed signed rank test with 1% significance level, which returns the approximate p-value.

`alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level of the decision of a hypothesis test, specified as the comma-separated pair consisting of `'alpha'` and a scalar value in the range 0 to 1. Significance level of `h` is 100 * `alpha`%.

Example: `'alpha', 0.01`

Data Types: `double` | `single`

`method` — Computation method of `p`

`'exact'` | `'approximate'`

Computation method of `p`, specified as the comma-separated pair consisting of `'method'` and one of the following.

`'exact'`	Exact computation of the *p*-value, `p`. Default value for 15 or fewer observations in `x`, `x – m`, or `x – y` when `method` is unspecified.
`'approximate'`	Normal approximation while computing the *p*-value, `p`. Default value for more than 15 observations in `x`, `x – m`, or `x – y` when `'method'` is unspecified because the exact method can be slow on large samples.

Example: `'method','exact'`

`tail` — Type of test

`'both'` (default) | `'right'` | `'left'`

Type of test, specified as the comma-separated pair consisting of `'tail'` and one of the following:

`'both'`	Two-sided hypothesis test, which is the default test type. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median different than 0 or `m`. • For a two-sample test, the alternate hypothesis states that the data in `x – y` come from a distribution with median different than 0.

`'right'`	Right-tailed hypothesis test. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median greater than 0 or `m`. • For a two-sample test, the alternate hypothesis states the data in `x – y` come from a distribution with median greater than 0.
`'left'`	Left-tailed hypothesis test. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median less than 0 or `m`. • For a two-sample test, the alternate hypothesis states the data in `x – y` come from a distribution with median less than 0.

Example: `'tail','left'`

Output Arguments

p — *p*-value of the test

nonnegative scalar

p-value of the test, returned as a nonnegative scalar from 0 to 1. `p` is the probability of observing a test statistic as or more extreme than the observed value under the null hypothesis. `signrank` computes the two-sided *p*-value by doubling the most significant one-sided value.

h — Result of the hypothesis test

1 | 0

Result of the hypothesis test, returned as a logical value.

- If `h` = 1, this indicates the rejection of the null hypothesis at the 100 * `alpha`% significance level.
- If `h` = 0, this indicates a failure to reject the null hypothesis at the 100 * `alpha`% significance level.

stats — Test statistics

structure

Test statistics, returned as a structure. The test statistics stored in `stats` are:

- `signrank`: Value of the sign rank test statistic.
- `zval`: Value of the *z*- statistic on page 33-5873 (computed when `'method'` is `'approximate'`).

More About

Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is a nonparametric test for two populations when the observations are paired. In this case, the test statistic, W, is the sum of the ranks of positive differences between the observations in the two samples (that is, `x – y`). When you use the test for one sample, then W is the sum of the ranks of positive differences between the observations and the hypothesized median value M_0 (which is 0 when you use `signrank(x)` and `m` when you use `signrank(x,m)`).

z-Statistic

For large samples, or when `method` is `approximate`, the `signrank` function calculates the *p*-value using the *z*-statistic, given by

$$z = \frac{(W - n(n+1)/4)}{\sqrt{\frac{n(n+1)(2n+1) - tieadj}{24}}},$$

where *n* is the sample size of the difference `x − y` or `x – m`. For the two-sample case, `signrank` uses `[tie_rank,tieadj] = tiedrank(abs(diffxy),0,0,epsdiff)` to obtain the tie adjustment value `tieadj`.

Algorithms

`signrank` treats `NaN`s in `x` and `y` as missing values and ignores them.

For the two-sample case, `signrank` uses a tolerance based on the values `epsdiff = eps(x) + eps(y)`. `signrank` computes the absolute values of the differences (`abs(d(i))` where `d(i) = x(i) − y(i)`) and compares them to `epsdiff`. Values with an absolute value less than `epsdiff` (`abs(d(i)) < epsdiff(i)`) are treated as ties.

References

[1] Gibbons, J. D., and S. Chakraborti. *Nonparametric Statistical Inference*, 5th Ed., Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group, 2011.

[2] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

See Also

`ranksum` | `signtest` | `ttest` | `ztest`

Introduced before R2006a

signtest

Sign test

Syntax

```
p = signtest(x)
p = signtest(x,y)
p = signtest(x,y,Name,Value)
[p,h] = signtest( ___ )
[p,h,stats] = signtest( ___ )

[ ___ ] = signtest(x,m)
[ ___ ] = signtest(x,m,Name,Value)
```

Description

`p = signtest(x)` returns the *p*-value for a two-sided sign test on page 33-5880.

`signtest` tests the hypothesis that data in `x` has a continuous distribution with zero median against the alternative that the distribution does not have zero median at the 5% significance level.

`p = signtest(x,y)` returns the *p*-value of a two-sided sign test on page 33-5880. Here,`signtest` tests for the hypothesis that the data in `x – y` has a distribution with zero median against the alternative that the distribution does not have zero median. Note that a hypothesis of zero median for `x – y` is not equivalent to a hypothesis of equal median for `x` and `y`.

`p = signtest(x,y,Name,Value)` returns the *p*-value for the sign test with additional options specified by one or more `Name,Value` pair arguments.

`[p,h] = signtest( ___ )` also returns a logical value indicating the test decision. The value `h = 1` indicates a rejection of the null hypothesis, and `h = 0` indicates a failure to reject the null hypothesis at the 5% significance level. You can use any of the input arguments in the previous syntaxes.

`[p,h,stats] = signtest( ___ )` also returns the structure `stats` containing information about the test statistic.

`[ ___ ] = signtest(x,m)` returns any of the output arguments in the previous syntaxes for the test whether the data in `x` are observations from a distribution with median `m` against the alternative that the median is different from `m`.

`[ ___ ] = signtest(x,m,Name,Value)` returns any of the output arguments in the previous syntaxes for the sign test with additional options specified by one or more `Name,Value` pair arguments.

Examples

Test for Zero Median of a Single Population

Test the hypothesis of zero median.

Generate the sample data.

```
rng('default') % for reproducibility
x = randn(1,25);
```

The sampling distribution of `x` is symmetric with zero median.

Test the null hypothesis that `x` comes from a distribution with a median different from zero median.

```
[p,h,stats] = signtest(x,0)
```

```
p = 0.1078

h = logical
   0


stats = struct with fields:
    zval: NaN
    sign: 17
```

At the default 5% significance level, the result `h = 0` indicates that `signtest` fails to reject to the null hypothesis of zero median. `signtest` calculates the *p*-value using the exact method, hence it does not calculate `zval` and returns it as a `NaN`.

Test for Zero Median for the Difference of Paired Samples

Test the hypothesis of zero median for the difference between paired samples.

Generate the sample data.

```
rng('default') % for reproducibility
before = lognrnd(2,.25,10,1);
after = before + (lognrnd(0,.5,10,1) - 1);
```

The sampling distribution of the difference between `before` and `after` is symmetric with zero median.

Test the null hypothesis that the difference of `before` and `after` has zero median.

```
[p,h] = signtest(before,after)
```

```
p = 0.7539

h = logical
   0
```

At the default 5% significance level, the value `h = 0` indicates that `signtest` fails to reject to the null hypothesis of zero median in the difference.

Medians of Paired Samples

Test the hypothesis of zero median for the difference between two paired samples using the exact and approximate methods.

Generate the sample data.

```
rng('default') % for reproducibility
x = lognrnd(2,.25,15,1);
y = x + trnd(2,15,1);
display([x y])
```

```
    8.4521    7.8047
   11.6869   11.4094
    4.2009    5.1133
    9.1664   12.1655
    8.0020   10.0300
    5.3285    6.0153
    6.6300    5.1235
    8.0499    8.6737
   18.0763   19.2164
   14.7665   15.3380
    5.2726    8.4187
   15.7798   16.2093
    8.8583    8.5575
    7.2735    7.4783
    8.8347    7.8894
```

Test the hypothesis that x – y has zero median.

```
[p,h,stats] = signtest(x,y)
```

```
p = 0.3018

h = logical
   0


stats = struct with fields:
    zval: NaN
    sign: 5
```

At the default 5% significance level, the value h = 0 indicates that the test fails to reject the null hypothesis of zero median in the difference.

Repeat the test using the approximate method.

```
[p,h,stats] = signtest(x,y,'Method','approximate')
```

```
p = 0.3017

h = logical
   0


stats = struct with fields:
    zval: -1.0328
    sign: 5
```

The approximate *p*-value, which `signtest` obtains using the z-statistic, is really close to the exact *p*-value.

Test for Large Samples

Perform a left-sided sign test for large samples.

Load the sample data.

```
load gradespaired
```

Test the null hypothesis that the median of the grade differences before and after the tutoring program is 0 against the alternate that it is less than 0.

```
[p,h,stats] = signtest(gradespaired(:,1),gradespaired(:,2),'Tail','left')
```

```
p = 0.0013
```

```
h = logical
   1
```

```
stats = struct with fields:
    zval: -3.0110
    sign: 37
```

Because the sample size is large (greater than 100), `signtest` uses an approximate method to calculate the *p*-value and also returns the value of the *z*-statistic. The test rejects the null hypothesis that there is no difference between the grade medians at the 5% significance level.

Test for Median of a Single Population

Test the hypothesis that the population median is different from a specified value.

Load the sample data.

```
load lawdata
```

The data set has 15 observations for variables `gpa` and `lsat`.

Test the hypothesis that the median `lsat` score is higher than 570.

```
[p,h,stats] = signtest(lsat,570,'Tail','right')
```

```
p = 0.0176
```

```
h = logical
   1
```

```
stats = struct with fields:
    zval: NaN
```

```
    sign: 12
```

Both the *p*-value, 0.0176, and `h` = 1 indicate that at the 5% significance level the test concludes in favor of the alternate hypothesis.

Input Arguments

`x` — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

`y` — Sample data

vector

Sample data, specified as a vector. `y` must be the same length as `x`.

Data Types: `single` | `double`

`m` — Hypothesized value of the median

scalar

Hypothesized value of the median, specified as a scalar.

Example: `signtest(x,35)`

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alpha',0.01,'Method','approximate','Tail','right'` specifies a right-tailed sign test with 1% significance level, which returns the approximate p-value.

`Alpha` — Significance level

0.05 (default) | scalar value in the range 0 to 1

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range 0 to 1. The default value of `Alpha` is 0.05. Significance level of `h` is 100 * `Alpha`%.

Example: `'Alpha', 0.01`

Data Types: `double` | `single`

`Method` — *p*-value computation method

`'exact'` | `'approximate'`

p-value computation method, specified as the comma-separated pair consisting of `'Method'` and one of the following:

`'exact'`	Exact computation of the *p*-value, `p`.
`'approximate'`	Normal approximation for computing the *p*-value, `p`.

The default computation method is `'exact'`, if there are fewer than 100 observations and `'approximate'` if there are 100 observations or more.

Example: `'Method','exact'`

`Tail` — Type of test
`'both'` (default) | `'right'` | `'left'`

Type of test, specified as the comma-separated pair consisting of `'Tail'` and one of the following:

`'both'`	Two-sided hypothesis test, which is the default test type. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median different than zero (or `m`). • For a two-sample test, the alternate hypothesis states that the data in `x-y` come from a distribution with median different than zero.
`'right'`	Right-tailed hypothesis test. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median greater than zero (or `m`). • For a two-sample test, the alternate hypothesis states the data in `x-y` come from a distribution with median greater than zero.
`'left'`	Left-tailed hypothesis test. • For a one-sample test, the alternate hypothesis states that the data in `x` come from a continuous distribution with median less than zero (or `m`). • For a two-sample test, the alternative hypothesis states the data in `x-y` come from a distribution with median less than zero.

Example: `'Tail','left'`

Output Arguments

`p` — *p*-value of the test
nonnegative scalar

p-value of the test, returned as a nonnegative scalar from 0 to 1. `p` is the probability of observing a test statistic as or more extreme than the observed value under the null hypothesis. `signtest` computes the two-sided *p*-value by doubling the most significant one-sided value.

`h` — Result of the hypothesis test
1 | 0

Result of the hypothesis test, returned as a logical value.

- If `h` = 1, this indicates rejection of the null hypothesis at the 100 * `Alpha`% significance level.
- If `h` = 0, this indicates a failure to reject the null hypothesis at the 100 * `Alpha`% significance level.

`stats` — Test statistics
structure

Test statistics, returned as a structure. The test statistics stored in `stats` are:

- `sign`: Value of the sign test statistic.
- `zval`: Value of the z-statistic on page 33-5880 (computed only for large samples).

More About

Sign Test

The sign test is a nonparametric test for the median of a population or median of the difference of two populations.

For example, for tests on a single population median:

- If the test is two-sided, then the test statistic, S, is the minimum of the number of observations that are smaller or larger than the hypothesized median value, M_0.
- If the test is right-sided, then S is the number of observations that are larger than the hypothesized median value M_0.
- If the test is left-sided, then S is the number of observations that are smaller than the hypothesized median value M_0.

z-Statistic

For a large sample, `signtest` uses the z-statistic to approximate the p-value.

The `signtest` test statistic is the number of elements that are greater than `0` (for `signtest(x)` or `signtest(x-y)`), or `m` (for `signtest(x,m)`). Hence, the z-statistic of the sign test, with the continuity correction, is:

$$z = \frac{(S - E(S))}{\sqrt{V(S)}} = \frac{(S - (0.5)n - 0.5sign(npos - nneg))}{\sqrt{(0.5)(0.5)n}},$$

where *npos* and *nneg* are the number of positive and negative differences from the hypothesized median value, respectively.

Algorithms

For a one-sample test, `signtest` omits values in `x` that are zero or `NaN`.

For a two-sample test, `signtest` omits values in `x – y` that are zero or `NaN`.

References

[1] Gibbons, J. D., and S. Chakraborti. *Nonparametric Statistical Inference*, 5th Ed. Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group, 2011.

[2] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

See Also

`ranksum` | `signrank` | `ttest` | `ztest`

Introduced before R2006a

silhouette

Silhouette plot

Syntax

```
silhouette(X,clust)
silhouette(X,clust,Distance)
silhouette(X,clust,Distance,DistParameter)

s = silhouette( ___ )
[s,h] = silhouette( ___ )
```

Description

`silhouette(X,clust)` plots cluster silhouettes for the *n*-by-*p* input data matrix `X`, given the cluster assignment `clust` of each point (observation) in `X`.

`silhouette(X,clust,Distance)` plots the silhouettes using the inter-point distance metric specified in `Distance`.

`silhouette(X,clust,Distance,DistParameter)` accepts one or more additional distance metric parameter values when you specify `Distance` as a custom distance function handle `@`*`distfun`* that accepts the additional parameter values.

`s = silhouette( ___ )` returns the silhouette values in `s` for any of the input argument combinations in the previous syntaxes without plotting the cluster silhouettes.

`[s,h] = silhouette( ___ )` plots the silhouettes and returns the figure handle `h` in addition to the silhouette values in `s`.

Examples

Create Silhouette Plot

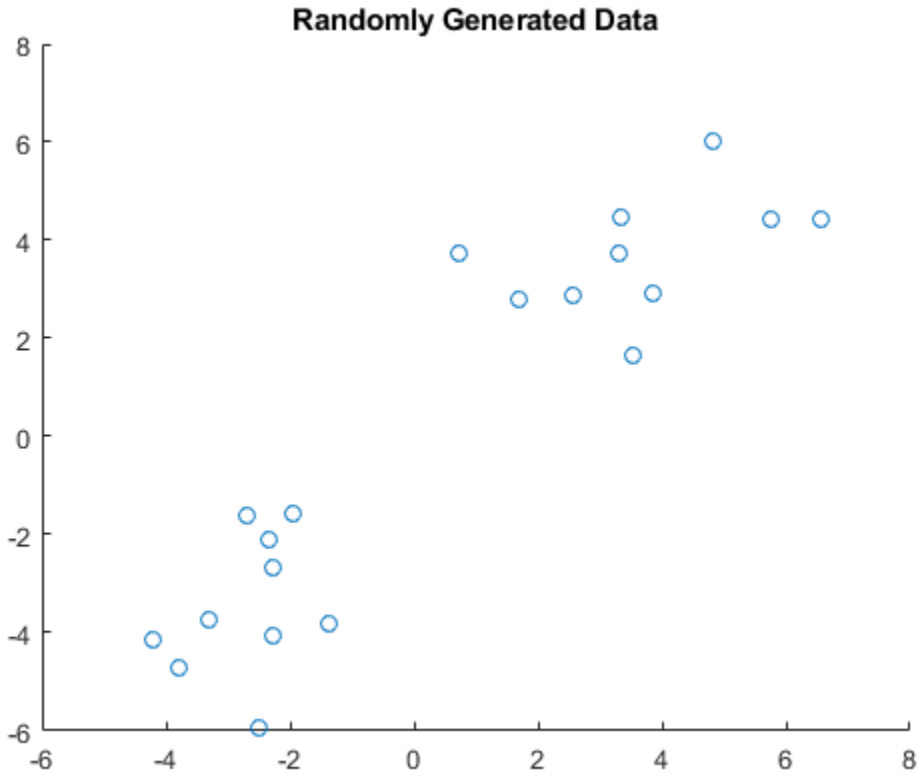
Create silhouette plots from clustered data using different distance metrics.

Generate random sample data.

```
rng('default')  % For reproducibility
X = [randn(10,2)+3;randn(10,2)-3];
```

Create a scatter plot of the data.

```
scatter(X(:,1),X(:,2));
title('Randomly Generated Data');
```



The scatter plot shows that the data appears to be split into two clusters of equal size.

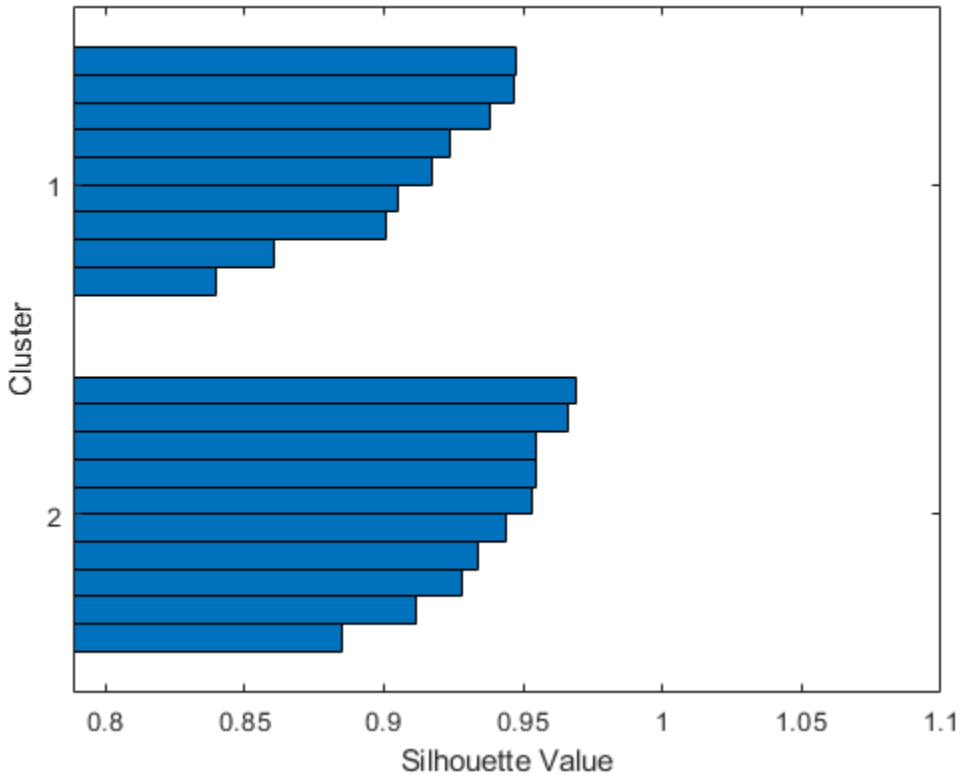
Partition the data into two clusters using `kmeans` with the default squared Euclidean distance metric.

```
clust = kmeans(X,2);
```

`clust` contains the cluster indices of the data.

Create a silhouette plot from the clustered data using the default squared Euclidean distance metric.

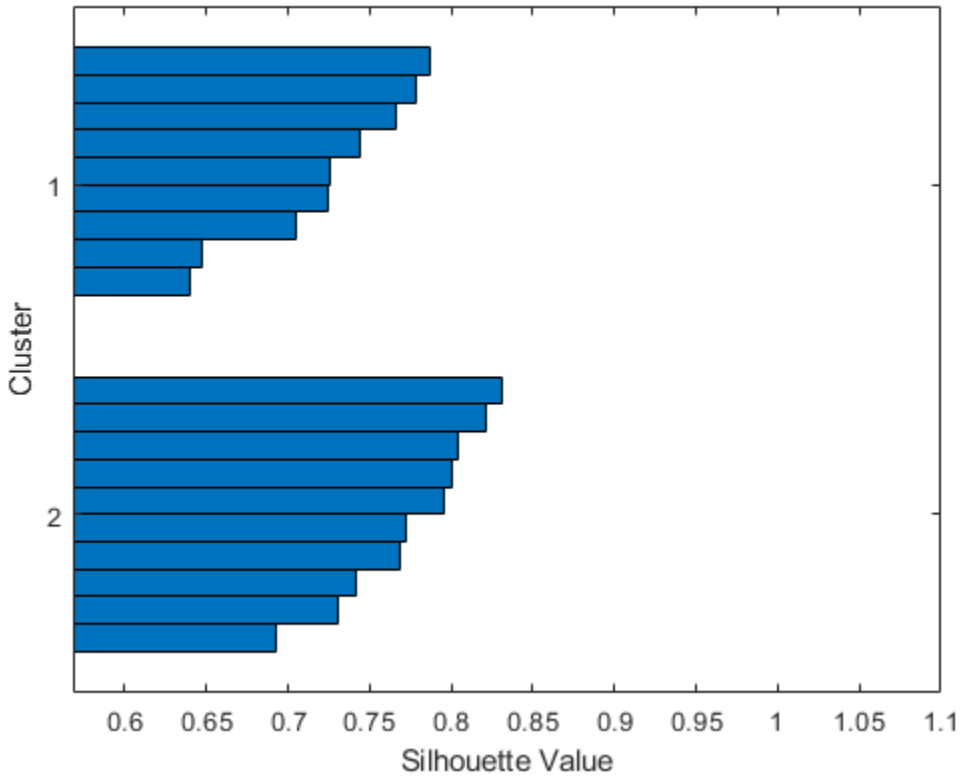
```
silhouette(X,clust)
```



The silhouette plot shows that the data is split into two clusters of equal size. All the points in the two clusters have large silhouette values (0.8 or greater), indicating that the clusters are well separated.

Create a silhouette plot from the clustered data using the Euclidean distance metric.

```
silhouette(X,clust,'Euclidean')
```



The silhouette plot shows that the data is split into two clusters of equal size. All the points in the two clusters have large silhouette values (0.6 or greater), indicating that the clusters are well separated.

Compute Silhouette Values

Compute the silhouette values from clustered data.

Generate random sample data.

```
rng('default')  % For reproducibility
X = [randn(10,2)+1;randn(10,2)-1];
```

Cluster the data in `X` based on the sum of absolute differences in distance by using `kmeans`.

```
clust = kmeans(X,2,'distance','cityblock');
```

`clust` contains the cluster indices of the data.

Compute the silhouette values from the clustered data. Specify the distance metric as `'cityblock'` to indicate that the `kmeans` clustering is based on the sum of absolute differences.

```
s = silhouette(X,clust,'cityblock')
```

```
s = 20×1
```

```
    0.0816
    0.5848
    0.1906
    0.2781
    0.3954
    0.4050
    0.0897
    0.5416
    0.6203
    0.6664
      ⋮
```

Find Silhouette Values Using Custom Distance Metric

Find silhouette values from clustered data using a custom chi-square distance metric. Verify that the chi-square distance metric is equivalent to the Euclidean distance metric, but with an optional scaling parameter.

Generate random sample data.

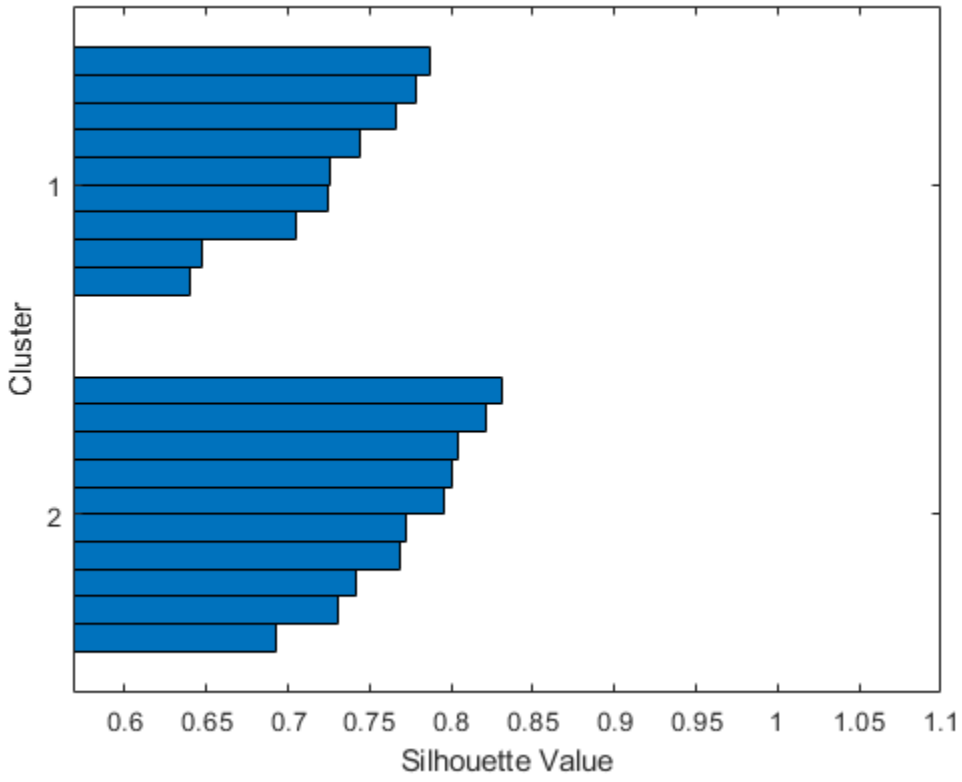
```
rng('default'); % For reproducibility
X = [randn(10,2)+3;randn(10,2)-3];
```

Cluster the data in `X` using `kmeans` with the default squared Euclidean distance metric.

```
clust = kmeans(X,2);
```

Find silhouette values and create a silhouette plot from the clustered data using the Euclidean distance metric.

```
[s,h] = silhouette(X,clust,'Euclidean')
```




```
s = 20×1

    0.6472
    0.7241
    0.5682
    0.7658
    0.7864
    0.6397
    0.7253
    0.7783
    0.7054
    0.7442
      ⋮
```

```
h =
  Figure (1) with properties:

      Number: 1
        Name: ''
       Color: [1 1 1]
    Position: [360 502 560 420]
       Units: 'pixels'

  Show all properties
```

The chi-square distance between J-dimensional points x and z is

$$\chi(x,z) = \sqrt{\sum_{j=1}^{J} w_j(x_j - z_j)^2},$$

where w_j is the weight associated with dimension j.

Set weights for each dimension and specify the chi-square distance function. The distance function must:

- Take as input arguments the n-by-p input data matrix `X`, one row of `X` (for example, `x`), and a scaling (or weight) parameter `w`.
- Calculate the distance from `x` to each row of `X`.
- Return a vector of length n. Each element of the vector is the distance between the observation corresponding to `x` and the observations corresponding to each row of `X`.

```
w = [0.4; 0.6]; % Set arbitrary weights for illustration
chiSqrDist = @(x,Z,w)sqrt((bsxfun(@minus,x,Z).^2)*w);
```

Find silhouette values from the clustered data using the custom distance metric `chiSqrDist`.

```
s1 = silhouette(X,clust,chiSqrDist,w)
```

```
s1 = 20×1

    0.6288
    0.7239
    0.6244
    0.7696
    0.7957
    0.6688
    0.7386
    0.7865
    0.7223
    0.7572
      ⋮
```

Set the weight for both dimensions to 1 to use `chiSqrDist` as the Euclidean distance metric. Find silhouette values and verify that they are the same as the values in `s`.

```
w2 = [1; 1];
s2 = silhouette(X,clust,chiSqrDist,w2);
AreValuesEqual = isequal(s2,s)
```

```
AreValuesEqual = logical
   1
```

The silhouette values are the same in `s` and `s2`.

Input Arguments

X — Input data

numeric matrix

Input data, specified as a numeric matrix of size *n*-by-*p*. Rows correspond to points, and columns correspond to coordinates.

Data Types: `single` | `double`

`clust` — Cluster assignment
categorical variable | numeric vector | character matrix | string array | cell array of character vectors

Cluster assignment, specified as a categorical variable, numeric vector, character matrix, string array, or cell array of character vectors containing a cluster name for each point in `X`.

`silhouette` treats `NaN`s and empty values in `clust` as missing values and ignores the corresponding rows of `X`.

Data Types: `single` | `double` | `char` | `string` | `cell` | `categorical`

`Distance` — Distance metric
`'sqEuclidean'` (default) | `'Euclidean'` | `'cityblock'` | function handle | vector of pairwise distances | ...

Distance metric, specified as a character vector, string scalar, or function handle, as described in this table.

Metric	Description
`'Euclidean'`	Euclidean distance
`'sqEuclidean'`	Squared Euclidean distance (default)
`'cityblock'`	Sum of absolute differences
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)
`'Hamming'`	Percentage of coordinates that differ
`'Jaccard'`	Percentage of nonzero coordinates that differ
Vector	A numeric row vector of pairwise distances, in the form created by the `pdist` function. `X` is not used in this case, and can safely be set to `[]`.

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D = distfun(X0,X,DistParameter)`
	`% calculation of distance`
	`...`
	where
	• `X0` is a `1`-by-*p* vector containing a single point (observation) of the input data matrix `X`.
	• `X` is an *n*-by-*p* matrix of points.
	• `DistParameter` represents one or more additional parameter values specific to `@`*`distfun`*.
	• `D` is an *n*-by-`1` vector of distances, and `D(k)` is the distance between observations `X0` and `X(k,:)`.

For more information, see “Distance Metrics” on page 18-12.

Example: `'cosine'`

Data Types: `char` | `string` | `function_handle` | `single` | `double`

`DistParameter` — Distance metric parameter value
positive scalar | numeric vector | numeric matrix

Distance metric parameter value, specified as a positive scalar, numeric vector, or numeric matrix. This argument is valid only when you specify a custom distance function handle `@distfun` that accepts one or more parameter values in addition to the input parameters `X0` and `X`.

Example: `silhouette(X,clust,distfun,p1,p2)` where `p1` and `p2` are additional distance metric parameter values for `@distfun`

Data Types: `single` | `double`

Output Arguments

`s` — Silhouette values
n-by-1 vector of values ranging from –1 to 1

Silhouette values, returned as an *n*-by-1 vector of values ranging from –1 to 1. A silhouette value measures how similar a point is to points in its own cluster, when compared to points in other clusters. Values range from –1 to 1. A high silhouette value indicates that a point is well matched to its own cluster, and poorly matched to other clusters.

Data Types: `single` | `double`

`h` — Figure handle
scalar

Figure handle, returned as a scalar. You can use the figure handle to query and modify figure properties. For more information, see Figure.

More About

Silhouette Value

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters.

The silhouette value `Si` for the `i`th point is defined as

```
Si = (bi-ai)/ max(ai,bi)
```

where `ai` is the average distance from the `i`th point to the other points in the same cluster as `i`, and `bi` is the minimum average distance from the `i`th point to points in a different cluster, minimized over clusters.

The silhouette value ranges from –1 to 1. A high silhouette value indicates that `i` is well matched to its own cluster, and poorly matched to other clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution might have too many or too few clusters. You can use silhouette values as a clustering evaluation criterion with any distance metric.

References

[1] Kaufman L., and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

See Also

`dendrogram` | `evalclusters` | `kmeans` | `linkage` | `pdist`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

SilhouetteEvaluation

Package: clustering.evaluation
Superclasses: ClusterCriterion

Silhouette criterion clustering evaluation object

Description

SilhouetteEvaluation is an object consisting of sample data, clustering data, and silhouette criterion values used to evaluate the optimal number of data clusters. Create a silhouette criterion clustering evaluation object using evalclusters.

Construction

eva = evalclusters(x,clust,'Silhouette') creates a silhouette criterion clustering evaluation object.

eva = evalclusters(x,clust,'Silhouette',Name,Value) creates a silhouette criterion clustering evaluation object using additional options specified by one or more name-value pair arguments.

Input Arguments

x — Input data

matrix

Input data, specified as an *N*-by-*P* matrix. *N* is the number of observations, and *P* is the number of variables.

Data Types: single | double

clust — Clustering algorithm

'kmeans' | 'linkage' | 'gmdistribution' | matrix of clustering solutions | function handle

Clustering algorithm, specified as one of the following.

'kmeans'	Cluster the data in x using the kmeans clustering algorithm, with 'EmptyAction' set to 'singleton' and 'Replicates' set to 5.
'linkage'	Cluster the data in x using the clusterdata agglomerative clustering algorithm, with 'Linkage' set to 'ward'.
'gmdistribution'	Cluster the data in x using the gmdistribution Gaussian mixture distribution algorithm, with 'SharedCov' set to true and 'Replicates' set to 5.

If criterion is 'CalinskiHarabasz', 'DaviesBouldin', or 'silhouette', you can specify a clustering algorithm using a function handle. The function must be of the form C = clustfun(DATA,K), where DATA is the data to be clustered, and K is the number of clusters. The output of clustfun must be one of the following:

- A vector of integers representing the cluster index for each observation in `DATA`. There must be `K` unique values in this vector.
- A numeric *n*-by-*K* matrix of score for *n* observations and *K* classes. In this case, the cluster index for each observation is determined by taking the largest score value in each row.

If `criterion` is `'CalinskiHarabasz'`, `'DaviesBouldin'`, or `'silhouette'`, you can also specify `clust` as a *n*-by-*K* matrix containing the proposed clustering solutions. *n* is the number of observations in the sample data, and *K* is the number of proposed clustering solutions. Column *j* contains the cluster indices for each of the *N* points in the *j*th clustering solution.

Data Types: `single` | `double` | `char` | `string` | `function_handle`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'KList',[1:5],'Distance','cityblock'` specifies to test 1, 2, 3, 4, and 5 clusters using the city block distance metric.

`ClusterPriors` — Prior probabilities for each cluster

`'empirical'` (default) | `'equal'`

Prior probabilities for each cluster, specified as the comma-separated pair consisting of `'ClusterPriors'` and one of the following.

`'empirical'`	Compute the overall silhouette value for the clustering solution by averaging the silhouette values for all points. Each cluster contributes to the overall silhouette value proportionally to its size.
`'equal'`	Compute the overall silhouette value for the clustering solution by averaging the silhouette values for all points within each cluster, and then averaging those values across all clusters. Each cluster contributes equally to the overall silhouette value, regardless of its size.

Example: `'ClusterPriors','empirical'`

`Distance` — Distance metric

`'sqEuclidean'` (default) | `'Euclidean'` | `'cityblock'` | vector | function | ...

Distance metric used for computing the criterion values, specified as the comma-separated pair consisting of `'Distance'` and one of the following.

`'sqEuclidean'`	Squared Euclidean distance
`'Euclidean'`	Euclidean distance. This option is not valid for the `kmeans` clustering algorithm.
`'cityblock'`	Sum of absolute differences
`'cosine'`	One minus the cosine of the included angle between points (treated as vectors)
`'correlation'`	One minus the sample correlation between points (treated as sequences of values)

`'Hamming'`	Percentage of coordinates that differ. This option is only valid for the `Silhouette` criterion.
`'Jaccard'`	Percentage of nonzero coordinates that differ. This option is only valid for the `Silhouette` criterion.

For detailed information about each distance metric, see `pdist`.

You can also specify a function for the distance metric using a function handle. The distance function must be of the form `d2 = distfun(XI,XJ)`, where `XI` is a 1-by-*n* vector corresponding to a single row of the input matrix `X`, and `XJ` is an m_2-by-*n* matrix corresponding to multiple rows of `X`. `distfun` must return an m_2-by-1 vector of distances `d2`, whose *k*th element is the distance between `XI` and `XJ(k,:)`.

`Distance` only accepts a function handle if the clustering algorithm `clust` accepts a function handle as the distance metric. For example, the `kmeans` clustering algorithm does not accept a function handle as the distance metric. Therefore, if you use the `kmeans` algorithm and then specify a function handle for `Distance`, the software errors.

- If `criterion` is `'silhouette'`, you can also specify `Distance` as the output vector created by the function `pdist`.
- When `clust` is `'kmeans'` or `'gmdistribution'`, `evalclusters` uses the distance metric specified for `Distance` to cluster the data.
- If `clust` is `'linkage'`, and `Distance` is either `'sqEuclidean'` or `'Euclidean'`, then the clustering algorithm uses the Euclidean distance and Ward linkage.
- If `clust` is `'linkage'` and `Distance` is any other metric, then the clustering algorithm uses the specified distance metric and average linkage.
- In all other cases, the distance metric specified for `Distance` must match the distance metric used in the clustering algorithm to obtain meaningful results.

Example: `'Distance','Euclidean'`

Data Types: `single` | `double` | `char` | `string` | `function_handle`

`KList` — List of number of clusters to evaluate
vector

List of number of clusters to evaluate, specified as the comma-separated pair consisting of `'KList'` and a vector of positive integer values. You must specify `KList` when `clust` is a clustering algorithm name or a function handle. When `criterion` is `'gap'`, `clust` must be a character vector, a string scalar, or a function handle, and you must specify `KList`.

Example: `'KList',[1:6]`

Data Types: `single` | `double`

Properties

`ClusteringFunction`

Clustering algorithm used to cluster the input data, stored as a valid clustering algorithm name or function handle. If the clustering solutions are provided in the input, `ClusteringFunction` is empty.

`ClusterPriors`

Prior probabilities for each cluster, stored as valid prior probability name.

`ClusterSilhouettes`

Silhouette values corresponding to each proposed number of clusters in `InspectedK`, stored as a cell array of vectors.

`CriterionName`

Name of the criterion used for clustering evaluation, stored as a valid criterion name.

`CriterionValues`

Criterion values corresponding to each proposed number of clusters in `InspectedK`, stored as a vector of numerical values.

`Distance`

Distance metric used for clustering data, stored as a valid distance metric name.

`InspectedK`

List of the number of proposed clusters for which to compute criterion values, stored as a vector of positive integer values.

`Missing`

Logical flag for excluded data, stored as a column vector of logical values. If `Missing` equals `true`, then the corresponding value in the data matrix `x` is not used in the clustering solution.

`NumObservations`

Number of observations in the data matrix `X`, minus the number of missing (`NaN`) values in `X`, stored as a positive integer value.

`OptimalK`

Optimal number of clusters, stored as a positive integer value.

`OptimalY`

Optimal clustering solution corresponding to `OptimalK`, stored as a column vector of positive integer values. If the clustering solutions are provided in the input, `OptimalY` is empty.

`X`

Data used for clustering, stored as a matrix of numerical values.

Methods

Inherited Methods

addK	Evaluate additional numbers of clusters
plot	Plot clustering evaluation object criterion values
compact	Compact clustering evaluation object

Examples

Evaluate the Clustering Solution Using Silhouette Criterion

Evaluate the optimal number of clusters using the silhouette clustering evaluation criterion.

Generate sample data containing random numbers from three multivariate distributions with different parameter values.

```
rng('default');  % For reproducibility
mu1 = [2 2];
sigma1 = [0.9 -0.0255; -0.0255 0.9];

mu2 = [5 5];
sigma2 = [0.5 0 ; 0 0.3];

mu3 = [-2, -2];
sigma3 = [1 0 ; 0 0.9];

N = 200;

X = [mvnrnd(mu1,sigma1,N);...
     mvnrnd(mu2,sigma2,N);...
     mvnrnd(mu3,sigma3,N)];
```

Evaluate the optimal number of clusters using the silhouette criterion. Cluster the data using `kmeans`.

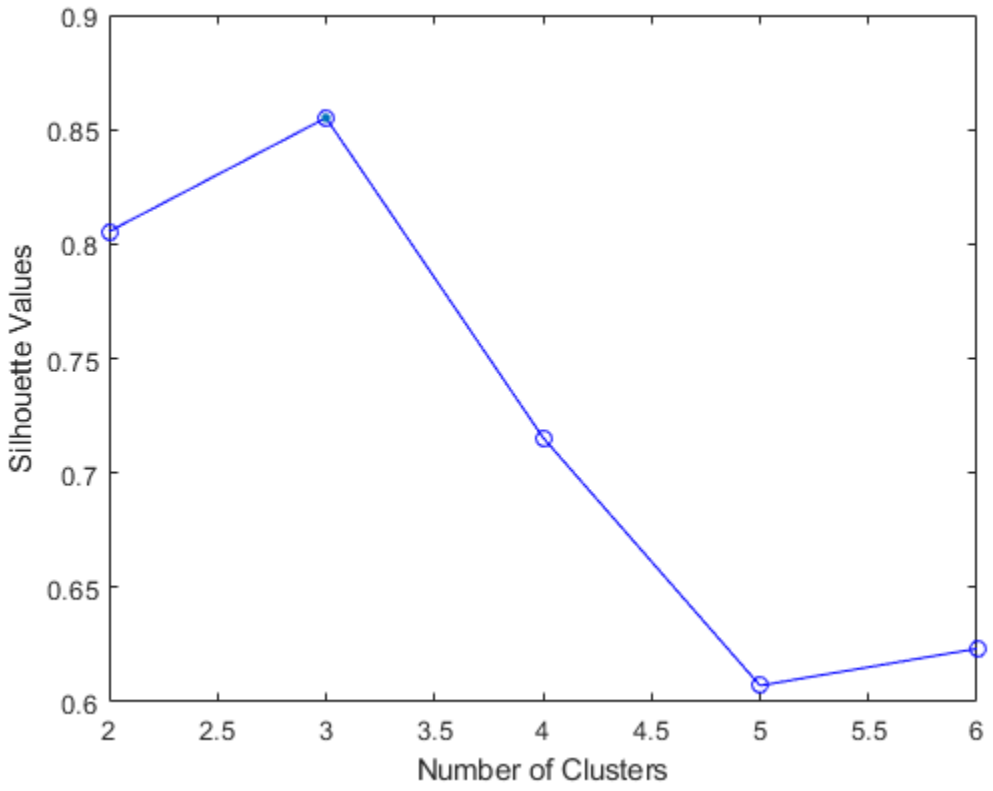
```
E = evalclusters(X,'kmeans','silhouette','klist',[1:6])
```

```
E =
  SilhouetteEvaluation with properties:

    NumObservations: 600
        InspectedK: [1 2 3 4 5 6]
    CriterionValues: [NaN 0.8055 0.8551 0.7155 0.6071 0.6232]
           OptimalK: 3
```

The `OptimalK` value indicates that, based on the silhouette criterion, the optimal number of clusters is three.

Plot the silhouette criterion values for each number of clusters tested.

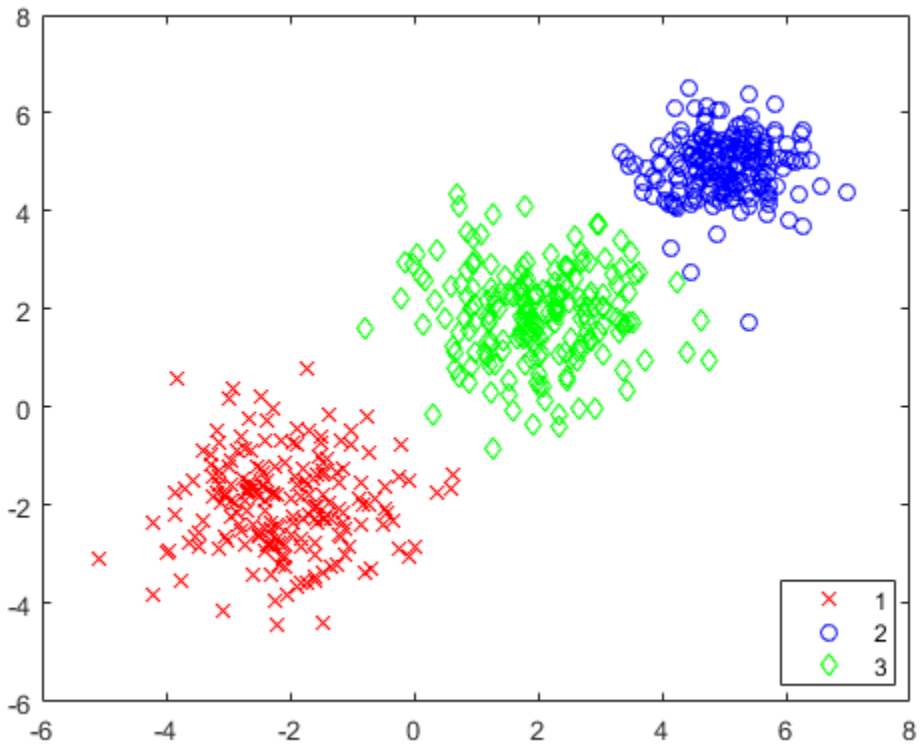
```
figure;
plot(E)
```



The plot shows that the highest silhouette value occurs at three clusters, suggesting that the optimal number of clusters is three.

Create a grouped scatter plot to visually examine the suggested clusters.

```
figure;
gscatter(X(:,1),X(:,2),E.OptimalY,'rbg','xod')
```



The plot shows three distinct clusters within the data: Cluster 1 is in the lower-left corner, cluster 2 is in the upper-right corner, and cluster 3 is near the center of the plot.

More About

Silhouette Value

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters.

The silhouette value `Si` for the `i`th point is defined as

```
Si = (bi-ai)/ max(ai,bi)
```

where `ai` is the average distance from the `i`th point to the other points in the same cluster as `i`, and `bi` is the minimum average distance from the `i`th point to points in a different cluster, minimized over clusters.

The silhouette value ranges from `−1` to `1`. A high silhouette value indicates that `i` is well matched to its own cluster, and poorly matched to other clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution might have too many or too few clusters. You can use silhouette values as a clustering evaluation criterion with any distance metric.

References

[1] Kaufman L. and P. J. Rouseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

[2] Rouseeuw, P. J. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.” *Journal of Computational and Applied Mathematics*. Vol. 20, No. 1, 1987, pp. 53–65.

See Also

`CalinskiHarabaszEvaluation` | `DaviesBouldinEvaluation` | `GapEvaluation` | `evalclusters` | `silhouette`

Topics
Class Attributes
Property Attributes

single

Class: `dataset`

(Not Recommended) Convert dataset variables to single array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = single(A)
B = single(A,vars)
```

Description

`B = single(A)` returns the contents of the dataset `A`, converted to one single array. The classes of the variables in the dataset must support the conversion.

`B = single(A,vars)` returns the contents of the dataset variables specified by `vars`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

See Also

`dataset` | `double` | `replacedata`

size

Class: `dataset`

(Not Recommended) Size of dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
D = SIZE(A)
[NOBS,NVARS] = SIZE(A)
[M1,M2,M3,...,MN] = SIZE(A)
M = size(A,dim)
```

Description

`D = SIZE(A)` returns the two-element row vector `D = [NOBS,NVARS]` containing the number of observations and number of variables in the dataset `A`. A dataset array always has two dimensions.

`[NOBS,NVARS] = SIZE(A)` returns the numbers of observations and variables in the dataset `A` as separate output variables.

`[M1,M2,M3,...,MN] = SIZE(A)`, for `N > 2`, returns `M1 = NOBS`, `M2 = NVARS`, and `M3,..,MN = 1`.

`M = size(A,dim)` returns the length of the dimension specified by the scalar `dim`:

- `M = size(A,1)` returns `NOBS`
- `M = size(A,2)` returns `NVARS`
- `M = size(A,k)` returns `1` for `k > 2`

See Also

`length` | `ndims` | `numel`

slicesample

Slice sampler

Syntax

```
rnd = slicesample(initial,nsamples,'pdf',pdf)
rnd = slicesample(initial,nsamples,'logpdf',logpdf)
[rnd,neval] = slicesample(initial,...)
[rnd,neval] = slicesample(initial,...,Name,Value)
```

Description

`rnd = slicesample(initial,nsamples,'pdf',pdf)` generates `nsamples` random samples using the slice sampling method (see “Algorithms” on page 33-5905). `pdf` gives the target probability density function (pdf). `initial` is a row vector or scalar containing the initial value of the random sample sequences.

`rnd = slicesample(initial,nsamples,'logpdf',logpdf)` generates samples using the logarithm of the pdf.

`[rnd,neval] = slicesample(initial,...)` returns the average number of function evaluations that occurred in the slice sampling.

`[rnd,neval] = slicesample(initial,...,Name,Value)` generates random samples with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`initial`

Initial point, a scalar or row vector. Set `initial` so `pdf(initial)` is a strictly positive scalar. `length(initial)` is the number of dimensions of each sample.

`nsamples`

Positive integer, the number of samples that `slicesample` generates.

`pdf`

Handle to a function that generates the probability density function, specified with `@`. `pdf` can be unnormalized, meaning it need not integrate to `1`.

`logpdf`

Handle to a function that generates the logarithm of the probability density function, specified with `@`. `logpdf` can be the logarithm of an unnormalized pdf.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`burnin`

Nonnegative integer, the number of samples to generate and discard before generating the samples to return. The slice sampling algorithm is a Markov chain whose stationary distribution is proportional to that of the `pdf` argument. Set `burnin` to a high enough value that you believe the Markov chain approximately reaches stationarity after `burnin` samples.

Default: `0`

`thin`

Positive integer, where `slicesample` discards every `thin - 1` samples and returns the next. The slice sampling algorithm is a Markov chain, so the samples are serially correlated. To reduce the serial correlation, choose a larger value of `thin`.

Default: `1`

`width`

Width of the interval around the current sample, a scalar or vector of positive values. `slicesample` begins with this interval and searches for an appropriate region containing the points of `pdf` that evaluate to a large enough value.

- If `width` is a scalar and the samples have multiple dimensions, `slicesample` uses `width` for each dimension.
- If `width` is a vector, it should have the same length as `initial`.

Default: `10`

Output Arguments

`rnd`

`nsamples`-by-`length(initial)` matrix, where each row is one sample.

`neval`

Scalar, the mean number of function evaluations per sample. `neval` includes the `burnin` and `thin` evaluations, not just the evaluations of samples returned in `rnd`. Therefore the total number of function evaluations is

`neval*(nsamples*thin + burnin)`.

Examples

Generate Random Samples From a Multimodal Density

This example shows how to generate random samples from a multimodal density using `slicesample`.

Define a function proportional to a multimodal density.

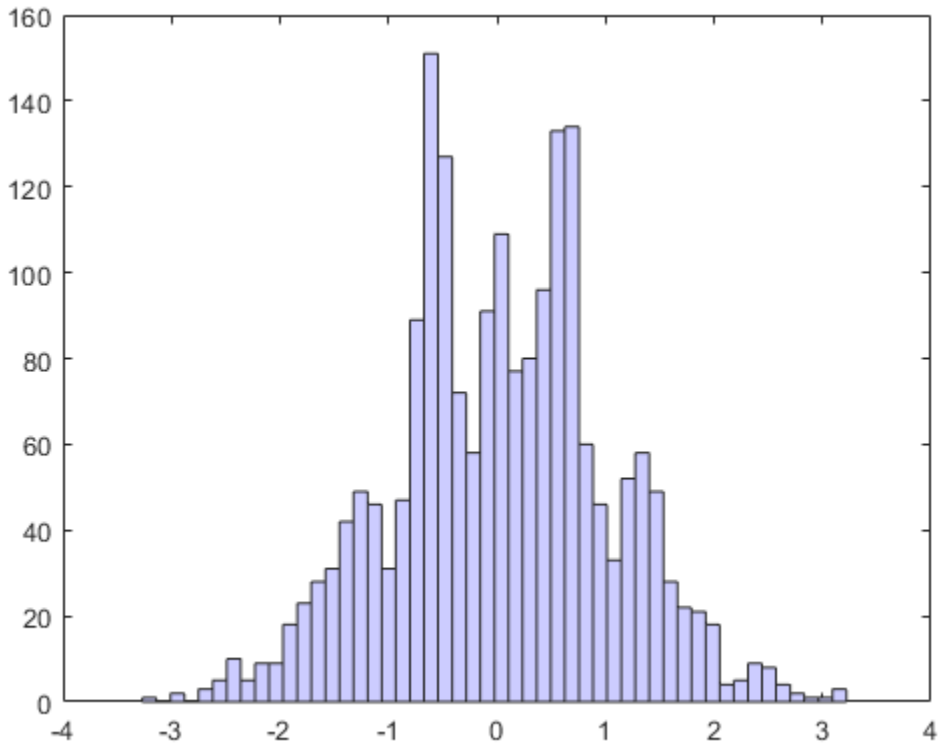
```
rng default  % For reproducibility
f = @(x) exp(-x.^2/2).*(1 + (sin(3*x)).^2).*...
    (1 + (cos(5*x).^2));
area = integral(f,-5,5);
```

Generate 2000 samples from the density, using a burn-in period of 1000, and keeping one in five samples.

```
N = 2000;
x = slicesample(1,N,'pdf',f,'thin',5,'burnin',1000);
```

Plot a histogram of the sample.

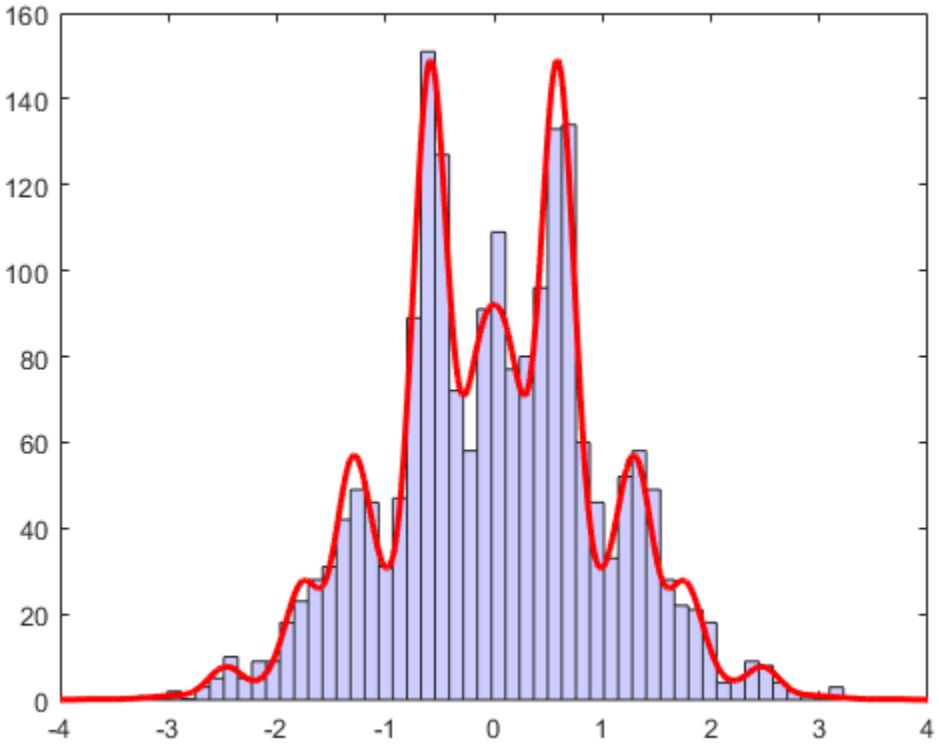
```
[binheight,bincenter] = hist(x,50);
h = bar(bincenter,binheight,'hist');
h.FaceColor = [.8 .8 1];
```



Scale the density to have the same area as the histogram, and superimpose it on the histogram.

```
hold on
h = gca;
```

```
xd = h.XLim;
xgrid = linspace(xd(1),xd(2),1000);
binwidth = (bincenter(2)-bincenter(1));
y = (N*binwidth/area) * f(xgrid);
plot(xgrid,y,'r','LineWidth',2)
hold off
```



The samples seem to fit the theoretical distribution well, so the `burnin` value seems adequate.

Tips

- There are no definitive suggestions for choosing appropriate values for `burnin`, `thin`, or `width`. Choose starting values of `burnin` and `thin`, and increase them, if necessary, to give the requisite independence and marginal distributions. See Neal [1] for details of the effect of adjusting `width`.

Algorithms

At each point in the sequence of random samples, `slicesample` selects the next point by “slicing” the density to form a neighborhood around the previous point where the density is above some value. Consequently, the sample points are not independent. Nearby points in the sequence tend to be closer together than they would be from a sample of independent values. For many purposes, the entire set of points can be used as a sample from the target distribution. However, when this type of serial correlation is a problem, the `burnin` and `thin` parameters can help reduce that correlation.

`slicesample` uses the slice sampling algorithm of Neal [1]. For numerical stability, it converts a `pdf` function into a `logpdf` function. The algorithm to resize the support region for each level, called "stepping-out" and "stepping-in," was suggested by Neal.

References

[1] Neal, Radford M. "Slice Sampling." *Ann. Stat.* Vol. 31, No. 3, pp. 705–767, 2003. Available at Project Euclid.

See Also

`mhsample` | `rand` | `randsample`

Topics

"Representing Sampling Distributions Using Markov Chain Samplers" on page 7-9

Introduced in R2006a

skewness

Skewness

Syntax

```
y = skewness(X)
y = skewness(X,flag)
y = skewness(X,flag,'all')
y = skewness(X,flag,dim)
y = skewness(X,flag,vecdim)
```

Description

`y = skewness(X)` returns the sample skewness of `X`.

- If `X` is a vector, then `skewness(X)` returns a scalar value that is the skewness of the elements in `X`.
- If `X` is a matrix, then `skewness(X)` returns a row vector containing the sample skewness of each column in `X`.
- If `X` is a multidimensional array, then `skewness(X)` operates along the first nonsingleton dimension of `X`.

`y = skewness(X,flag)` specifies whether to correct for bias (`flag = 0`) or not (`flag = 1`, the default). When `X` represents a sample from a population, the skewness of `X` is biased, meaning it tends to differ from the population skewness by a systematic amount based on the sample size. You can set `flag` to `0` to correct for this systematic bias.

`y = skewness(X,flag,'all')` returns the skewness of all elements of `X`.

`y = skewness(X,flag,dim)` returns the skewness along the operating dimension `dim` of `X`.

`y = skewness(X,flag,vecdim)` returns the skewness over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `skewness(X,1,[1 2])` returns a 1-by-1-by-4 array. Each element of the output array is the biased skewness of the elements on the corresponding page of `X`.

Examples

Find Skewness of Matrix

Set the random seed for reproducibility of the results.

```
rng('default')
```

Generate a matrix with 5 rows and 4 columns.

```
X = randn(5,4)
```

```
X = 5×4
```

```
    0.5377   -1.3077   -1.3499   -0.2050
    1.8339   -0.4336    3.0349   -0.1241
   -2.2588    0.3426    0.7254    1.4897
    0.8622    3.5784   -0.0631    1.4090
    0.3188    2.7694    0.7147    1.4172
```

Find the sample skewness of `X`.

```
y = skewness(X)
```

```
y = 1×4

   -0.9362    0.2333    0.4363   -0.4075
```

`y` is a row vector containing the sample skewness of each column in `X`.

Correct for Bias in Sample Skewness

For an input vector, correct for bias in the calculation of skewness by specifying the `flag` input argument.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Generate a vector of length 10.

```
x = randn(10,1)
```

```
x = 10×1

    0.5377
    1.8339
   -2.2588
    0.8622
    0.3188
   -1.3077
   -0.4336
    0.3426
    3.5784
    2.7694
```

Find the biased skewness of `x`. By default, `skewness` sets the value of `flag` to `1` for computing the biased skewness.

```
y1 = skewness(x) % flag is 1 by default
```

```
y1 = 0.1061
```

Find the bias-corrected skewness of `x` by setting the value of `flag` to `0`.

```
y2 = skewness(x,0)
```

```
y2 = 0.1258
```

Find Skewness Along Given Dimension

Find the skewness along different dimensions for a multidimensional array.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4,3,2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
   -0.2050    1.4172    1.6302
```

Find the skewness of `X` along the default dimension.

```
Y1 = skewness(X)
```

```
Y1 =
Y1(:,:,1) =

   -0.8084   -0.5578   -1.0772


Y1(:,:,2) =

   -0.0403   -1.1472   -0.6632
```

By default, skewness operates along the first dimension of `X` whose size does not equal 1. In this case, this dimension is the first dimension of `X`. Therefore, `Y1` is a 1-by-3-by-2 array.

Find the biased skewness of `X` along the second dimension.

```
Y2 = skewness(X,1,2)
```

```
Y2 =
Y2(:,:,1) =

    0.6956
   -0.5575
```

```
    0.0049
    0.6033


Y2(:,:,2) =

   -0.6969
    0.1828
    0.7071
   -0.6714
```

`Y2` is a 4-by-1-by-2 array.

Find the biased skewness of `X` along the third dimension.

```
Y3 = skewness(X,1,3)
```

```
Y3 = 4×3
10^-15 ×

         0    0.1597    0.5062
    0.1952         0         0
         0   -0.2130         0
    0.3654         0    0.4807
```

`Y3` is a 4-by-3 matrix.

Find Skewness Along Vector of Dimensions

Find the skewness over multiple dimensions by using the `'all'` and `vecdim` input arguments.

Set the random seed for reproducibility of the results.

```
rng('default')
```

Create a 4-by-3-by-2 array of random numbers.

```
X = randn([4 3 2])
```

```
X =
X(:,:,1) =

    0.5377    0.3188    3.5784
    1.8339   -1.3077    2.7694
   -2.2588   -0.4336   -1.3499
    0.8622    0.3426    3.0349


X(:,:,2) =

    0.7254   -0.1241    0.6715
   -0.0631    1.4897   -1.2075
    0.7147    1.4090    0.7172
```

```
   -0.2050    1.4172    1.6302
```

Find the biased skewness of X.

```
yall = skewness(X,1,'all')
```

```
yall = 0.0916
```

`yall` is the biased skewness of the entire input data set X.

Find the biased skewness of each page of X by specifying the first and second dimensions.

```
ypage = skewness(X,1,[1 2])
```

```
ypage =
ypage(:,:,1) =

    0.1070


ypage(:,:,2) =

   -0.6263
```

For example, `ypage(1,1,2)` is the biased skewness of the elements in `X(:,:,2)`.

Find the biased skewness of the elements in each `X(:,i,:)` slice by specifying the first and third dimensions.

```
ycol = skewness(X,1,[1 3])
```

```
ycol = 1×3

   -1.0755   -0.3108   -0.2209
```

For example, `ycol(3)` is the biased skewness of the elements in `X(:,3,:)`.

Input Arguments

X — Input data

vector | matrix | multidimensional array

Input data that represents a sample from a population, specified as a vector, matrix, or multidimensional array.

- If X is a vector, then `skewness(X)` returns a scalar value that is the skewness of the elements in X.
- If X is a matrix, then `skewness(X)` returns a row vector containing the sample skewness of each column in X.
- If X is a multidimensional array, then `skewness(X)` operates along the first nonsingleton dimension of X.

To specify the operating dimension when X is a matrix or an array, use the `dim` input argument.

`skewness` treats `NaN` values in `X` as missing values and removes them.

Data Types: `single` | `double`

`flag` — Indicator for bias
1 (default) | 0

Indicator for the bias, specified as `0` or `1`.

- If `flag` is `1` (default), then the skewness of `X` is biased, meaning it tends to differ from the population skewness by a systematic amount based on the sample size.
- If `flag` is `0`, then `skewness` corrects for the systematic bias.

Data Types: `single` | `double` | `logical`

`dim` — Dimension
positive integer

Dimension along which to operate, specified as a positive integer. If you do not specify a value for `dim`, then the default is the first dimension of `X` whose size does not equal 1.

Consider the skewness of a matrix `X`:

- If `dim` is equal to 1, then `skewness` returns a row vector that contains the sample skewness of each column in `X`.
- If `dim` is equal to 2, then `skewness` returns a column vector that contains the sample skewness of each row in `X`.

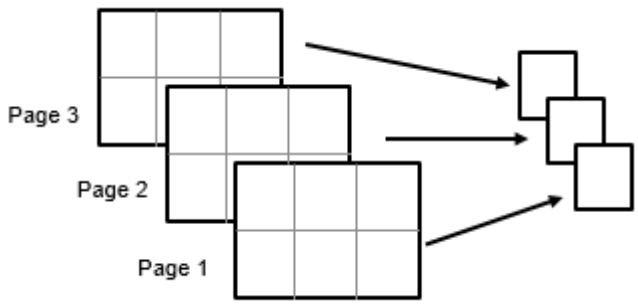
If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `skewness` returns an array of `NaN`s the same size as `X`.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `y` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `y`.

For example, if `X` is a 2-by-3-by-3 array, then `skewness(X,1,[1 2])` returns a 1-by-1-by-3 array. Each element of the output array is the biased skewness of the elements on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

y — Skewness

scalar | vector | matrix | multidimensional array

Skewness, returned as a scalar, vector, matrix, or multidimensional array.

Algorithms

Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data spreads out more to the left of the mean than to the right. If skewness is positive, the data spreads out more to the right. The skewness of the normal distribution on page B-119 (or any perfectly symmetric distribution) is zero.

The skewness of a distribution is defined as

$$s = \frac{E(x-\mu)^3}{\sigma^3},$$

where μ is the mean of x, σ is the standard deviation of x, and $E(t)$ represents the expected value of the quantity t. The `skewness` function computes a sample version of this population value.

When you set `flag` to `1`, the skewness is biased, and the following equation applies:

$$s_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^3}.$$

When you set `flag` to `0`, `skewness` corrects for the systematic bias, and the following equation applies:

$$s_0 = \frac{\sqrt{n(n-1)}}{n-2}s_1.$$

This bias-corrected equation requires that `X` contain at least three elements.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`kurtosis` | `mean` | `moment` | `std` | `var`

Topics

"Normal Distribution" on page B-119

Introduced before R2006a

sobolset

Sobol quasirandom point set

Description

`sobolset` is a quasirandom point set object that produces points from the Sobol sequence. The Sobol sequence is a base-2 digital sequence that fills space in a highly uniform manner.

Creation

Syntax

```
p = sobolset(d)
p = sobolset(d,Name,Value)
```

Description

`p = sobolset(d)` constructs a `d`-dimensional point set `p`, which is a `sobolset` object with default property settings. The input argument `d` corresponds to the `Dimensions` property of `p`.

`p = sobolset(d,Name,Value)` sets properties on page 33-5915 of `p` using one or more name-value pair arguments. Enclose each property name in quotes. For example, `sobolset(5,'Leap',2)` creates a five-dimensional point set from the first point, fourth point, seventh point, tenth point, and so on.

The returned object `p` encapsulates properties of a Sobol quasirandom sequence. The point set is finite, with a length determined by the `Skip` and `Leap` properties and by limits on the size of the point set indices (maximum value of 2^{53}). Values of the point set are generated whenever you access `p` using `net` or parenthesis indexing. Values are not stored within `p`.

Properties

`Dimensions` — Number of dimensions

positive integer scalar in interval [1,1111]

This property is read-only.

Number of dimensions of the points in the point set, specified as a positive integer scalar in the interval [1,1111]. For example, each point in the point set `p` with `p.Dimensions = 5` has five values.

Use the `d` input argument to specify the number of dimensions when you create a point set using the `sobolset` function.

Use the `reduceDimensions` object function to reduce the number of dimensions after you create a point set.

`Leap` — Interval between points

`0` (default) | positive integer scalar

Interval between points in the sequence, specified as a positive integer scalar. In other words, the `Leap` property of a point set specifies the number of points in the sequence to leap over and omit for every point taken. The default `Leap` value is 0, which corresponds to taking every point from the sequence.

Leaping is a technique used to improve the quality of a point set. However, you must choose the `Leap` values with care. Many `Leap` values create sequences that fail to touch on large sub-hyper-rectangles of the unit hypercube and, therefore, fail to be a uniform quasirandom point set. For more information, see [4].

Example: `p = sobolset(__,'Leap',50);`

Example: `p.Leap = 100;`

`PointOrder` — Point generation method
`'standard'` (default) | `'graycode'`

Point generation method, specified as `'standard'` or `'graycode'`. The `PointOrder` property specifies the order in which the Sobol sequence points are produced. When `PointOrder` is set to `'standard'`, the points produced match the original Sobol sequence implementation. When `PointOrder` is set to `'graycode'`, the sequence is generated by an implementation that uses the Gray code of the index instead of the index itself.

You can use the `'graycode'` option for faster sequence generation, but the software then changes the order of the generated points. For more information on the Gray code implementation, see [1].

Example: `p = sobolset(__,'PointOrder','graycode');`

Example: `p.PointOrder = 'standard';`

`ScrambleMethod` — Settings that control scrambling
`0x0` structure (default) | structure with `Type` and `Options` fields

Settings that control the scrambling of the sequence, specified as a structure with these fields:

- `Type` — A character vector containing the name of the scramble
- `Options` — A cell array of parameter values for the scramble

Use the `scramble` object function to set scrambles. For a list of valid scramble types, see the `type` input argument of `scramble`. An error occurs if you set an invalid scramble type for a given point set.

The `ScrambleMethod` property also accepts an empty matrix as a value. The software then clears all scrambling and sets the property to contain a `0x0` structure.

`Skip` — Number of initial points in sequence to omit
`0` (default) | positive integer scalar

Number of initial points in the sequence to omit from the point set, specified as a positive integer scalar.

Initial points of a sequence sometimes exhibit undesirable properties. For example, the first point is often `(0,0,0,...)`, which can cause the sequence to be unbalanced because the counterpart of the point, `(1,1,1,...)`, never appears. Also, initial points often exhibit correlations among different dimensions, and these correlations disappear later in the sequence.

Example: `p = sobolset(__,'Skip',2e3);`

Example: `p.Skip = 1e3;`

Type — Sequence type
`'Sobol'` (default)

This property is read-only.

Sequence type on which the quasirandom point set `p` is based, specified as `'Sobol'`.

Object Functions

net	Generate quasirandom point set
reduceDimensions	Reduce dimensions of Sobol point set
scramble	Scramble quasirandom point set

You can also use the following MATLAB functions with a `sobolset` object. The software treats the point set object like a matrix of multidimensional points.

length	Length of largest array dimension
size	Array size

Examples

Create Sobol Point Set

Generate a three-dimensional Sobol point set, skip the first 1000 values, and then retain every 101st point.

```
p = sobolset(3,'Skip',1e3,'Leap',1e2)
```

```
p =
Sobol point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : none
        PointOrder : standard
```

Apply a random linear scramble combined with a random digital shift by using `scramble`.

```
p = scramble(p,'MatousekAffineOwen')
```

```
p =
Sobol point set in 3 dimensions (89180190640991 points)

Properties:
              Skip : 1000
              Leap : 100
    ScrambleMethod : MatousekAffineOwen
        PointOrder : standard
```

Generate the first four points by using `net`.

```
X0 = net(p,4)
```

```
X0 = 4×3

    0.7601    0.5919    0.9529
    0.1795    0.0856    0.0491
    0.5488    0.0785    0.8483
    0.3882    0.8771    0.8755
```

Generate every third point, up to the eleventh point, by using parenthesis indexing.

```
X = p(1:3:11,:)
```

```
X = 4×3

    0.7601    0.5919    0.9529
    0.3882    0.8771    0.8755
    0.6905    0.4951    0.8464
    0.1955    0.5679    0.3192
```

Tips

- The `Skip` and `Leap` properties are useful for parallel applications. For example, if you have a Parallel Computing Toolbox license, you can partition a sequence of points across *N* different workers by using the function `labindex`. On each *n*th worker, set the `Skip` property of the point set to *n* – 1 and the `Leap` property to *N* – 1. The following code shows how to partition a sequence across three workers.

  ```
  Nworkers = 3;
  p = sobolset(10,'Leap',Nworkers-1);
  spmd(Nworkers)
      p.Skip = labindex - 1;

      % Compute something using points 1,4,7...
      % or points 2,5,8... or points 3,6,9...
  end
  ```

Algorithms

Sobol Sequence Generation

Consider a default `sobolset` object `p` that contains `d`-dimensional points. Each `p(i,:)` is a point in a Sobol sequence. The *j*th coordinate of the *i*th point, `p(i,j)`, is equal to

$$\begin{cases} 0, & i = 1 \\ \gamma_i(1)v_j(1) \oplus \gamma_i(2)v_j(2) \oplus \ldots, & i > 1. \end{cases}$$

- The $\gamma_i(n)$ values are 0s or 1s such that

  $$i - 1 = \sum_{n=1} \gamma_i(n) 2^{n-1}.$$

  In other words, the $\gamma_i(n)$ values are the binary digits of the integer i – 1.

- The $v_j(n)$ values are called direction numbers. They are uniquely defined for each coordinate j. For more details on these values, see “Direction Numbers Generation” on page 33-5919.
- The $\oplus$ operator is the bitwise exclusive-or operator. For two numbers expressed in binary, the $\oplus$ operator compares the digits in each position. For a given digit position, the $\oplus$ operator returns a 1 if the digits in that position differ and returns a 0 if the digits in that position are the same.
  - For example, $19 \oplus 24 = (10011)_2 \oplus (11000)_2 = (01011)_2 = 11$.
  - Similarly, $\frac{1}{2} \oplus \frac{3}{4} = (0.1)_2 \oplus (0.11)_2 = (0.01)_2 = \frac{1}{4}$.

For more information, see [3].

Direction Numbers Generation

The set of direction numbers $v_j(n)$ depends on the coordinate j. Define the direction numbers in terms of $m_j(n)$ values:

$$v_j(n) := \frac{m_j(n)}{2^n}.$$

For each j, you can generate the direction numbers by selecting the following:

- A primitive polynomial in $\mathbb{Z}_2$ of some degree s_j

  $$x^{s_j} + a_j(1)x^{s_j-1} + a_j(2)x^{s_j-2} + \ldots + a_j(s_j-1)x + 1.$$

  Each coefficient in the polynomial is either 0 or 1.
- s_j initial direction numbers. For each initial direction number, the corresponding $m_j(n)$ value must be either 1 or an odd number less than 2^n.

The remaining direction numbers are determined by the following recurrence relation, which uses the coefficients of the primitive polynomial, the previous direction numbers, and the $\oplus$ bitwise exclusive-or operator.

$$m_j(n) := 2a_j(1)m_j(n-1) \oplus 2^2 a_j(2)m_j(n-2) \oplus \ldots \oplus 2^{s_j-1}a_j(s_j-1)m_j(n-s_j+1) \oplus 2^{s_j}m_j(n-s_j) \oplus m_j(n-s_j).$$

`sobolset` uses the same primitive polynomials and initial direction numbers described in [3]. These parameters are provided for the first 1111 dimensions.

References

[1] Bratley, P., and B. L. Fox. “Algorithm 659 Implementing Sobol's Quasirandom Sequence Generator.” *ACM Transactions on Mathematical Software*. Vol. 14, No. 1, 1988, pp. 88–100.

[2] Hong, H. S., and F. J. Hickernell. “Algorithm 823: Implementing Scrambled Digital Sequences.” *ACM Transactions on Mathematical Software*. Vol. 29, No. 2, 2003, pp. 95–109.

[3] Joe, S., and F. Y. Kuo. “Remark on Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator.” *ACM Transactions on Mathematical Software*. Vol. 29, No. 1, 2003, pp. 49–57.

[4] Kocis, L., and W. J. Whiten. “Computational Investigations of Low-Discrepancy Sequences.” *ACM Transactions on Mathematical Software*. Vol. 23, No. 2, 1997, pp. 266–294.

[5] Matousek, J. “On the L2-Discrepancy for Anchored Boxes.” *Journal of Complexity*. Vol. 14, No. 4, 1998, pp. 527–556.

See Also

`haltonset` | `net` | `reduceDimensions` | `scramble`

Topics

“Generating Quasi-Random Numbers” on page 7-12

Introduced in R2008a

sortClasses

Package: `mlearnlib.graphics.chart`

Sort classes of confusion matrix chart

Syntax

```
sortClasses(cm,order)
```

Description

`sortClasses(cm,order)` sorts the classes of the confusion matrix chart `cm` in the order specified by `order`. You can sort the classes in their natural order, by the values along the diagonal of the confusion matrix, or in fixed order that you specify.

Examples

Sort Classes by Precision or Recall

Create a confusion matrix chart and sort the classes of the chart according to the class-wise true positive rate (recall) or the class-wise positive predictive value (precision).

Load and inspect the `arrhythmia` data set.

```
load arrhythmia
isLabels = unique(Y);
nLabels = numel(isLabels)
```

```
nLabels = 13
```

```
tabulate(categorical(Y))
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

The data contains 16 distinct labels that describe various degrees of arrhythmia, but the response (Y) includes only 13 distinct labels.

Train a classification tree and predict the resubstitution response of the tree.

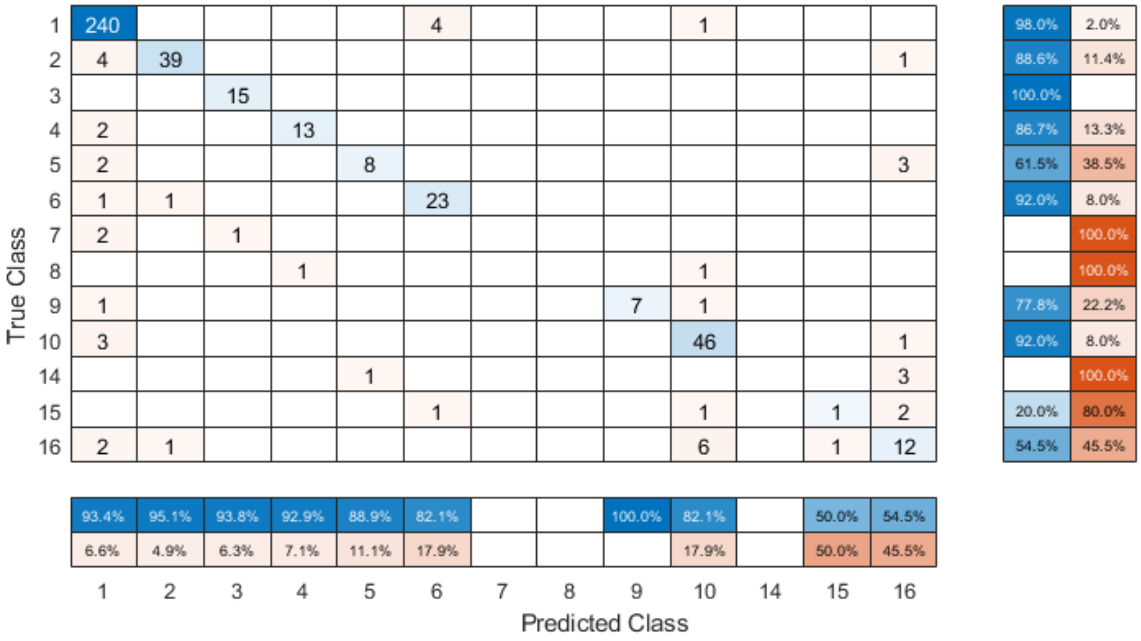
```
Mdl = fitctree(X,Y);
predictedY = resubPredict(Mdl);
```

Create a confusion matrix chart from the true labels `Y` and the predicted labels `predictedY`. Specify `'RowSummary'` as `'row-normalized'` to display the true positive rates and false positive rates in the row summary. Also, specify `'ColumnSummary'` as `'column-normalized'` to display the positive predictive values and false discovery rates in the column summary.

```
fig = figure;
cm = confusionchart(Y,predictedY,'RowSummary','row-normalized','ColumnSummary','column-normalized
```

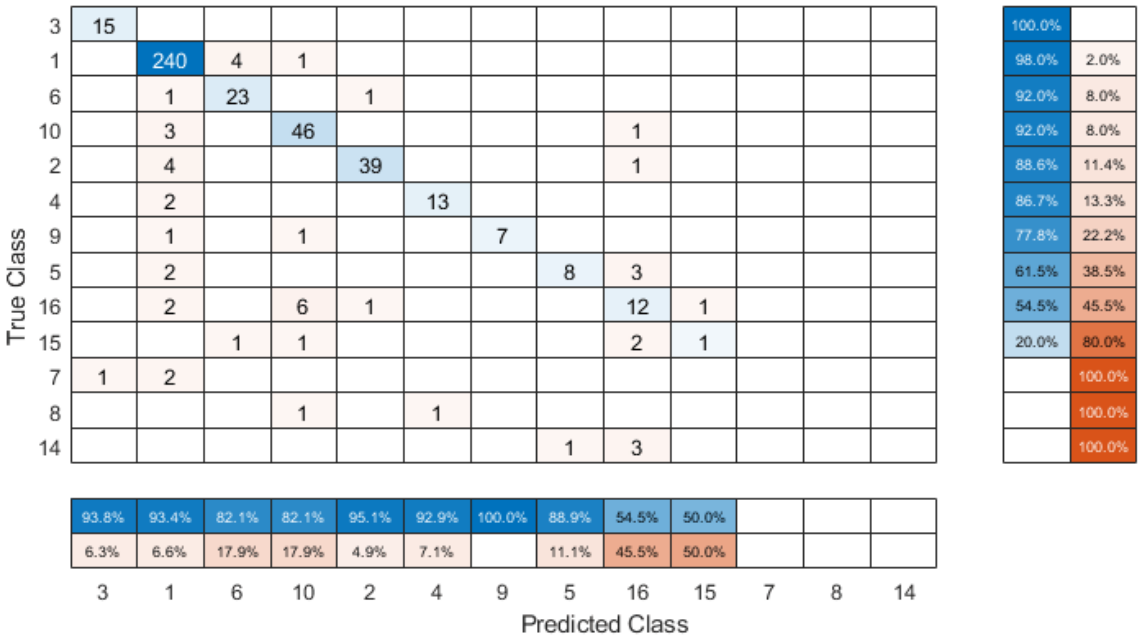
Resize the container of the confusion chart so percentages appear in the row summary.

```
fig_Position = fig.Position;
fig_Position(3) = fig_Position(3)*1.5;
fig.Position = fig_Position;
```



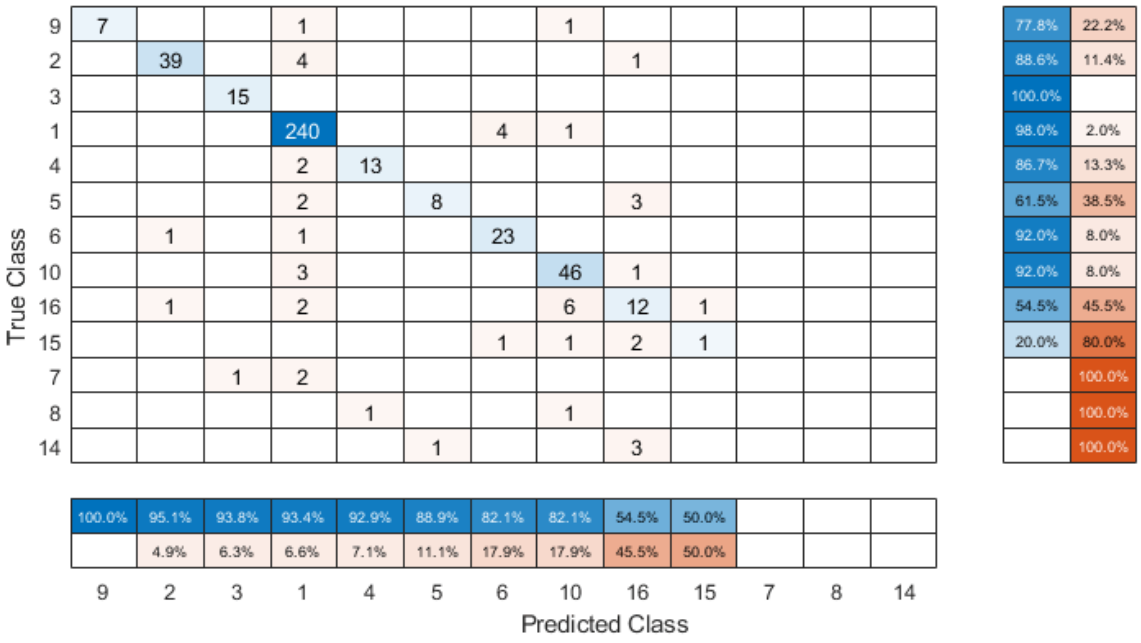
To sort the confusion matrix according to the true positive rate, normalize the cell values across each row by setting the `Normalization` property to `'row-normalized'` and then use `sortClasses`. After sorting, reset the `Normalization` property back to `'absolute'` to display the total number of observations in each cell.

```
cm.Normalization = 'row-normalized';
sortClasses(cm,'descending-diagonal')
cm.Normalization = 'absolute';
```



To sort the confusion matrix according to the positive predictive value, normalize the cell values across each column by setting the `Normalization` property to `'column-normalized'` and then use `sortClasses`. After sorting, reset the `Normalization` property back to `'absolute'` to display the total number of observations in each cell.

```
cm.Normalization = 'column-normalized';
sortClasses(cm,'descending-diagonal')
cm.Normalization = 'absolute';
```



Sort Classes to Cluster Similar Classes

Create a confusion matrix chart by using the `confusionchart` function, and sort the classes to cluster similar classes by using the `'cluster'` option of the `sortClasses` function. This example also shows how to cluster by using the `pdist`, `linkage`, and `optimalleaforder` functions.

Generate a sample data set that contains eight distinct classes.

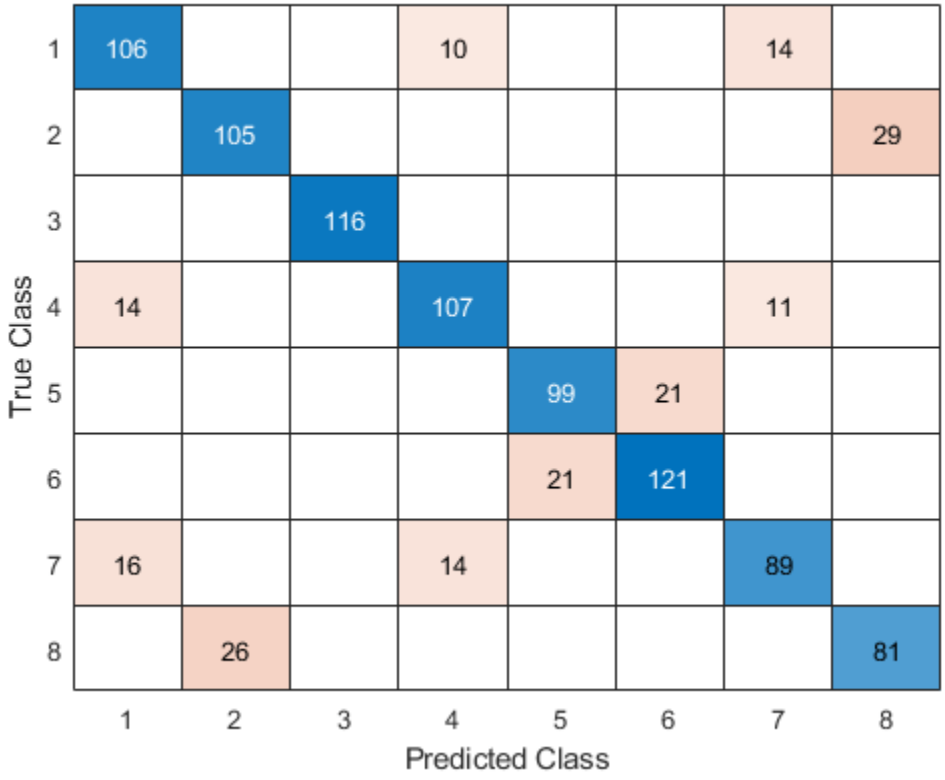
```
rng('default') % For reproducibility
trueLabels = randi(8,1000,1);
predictedLabels = trueLabels;
```

Insert confusion among classes {1,4,7}, {2,8}, and {5,6} for the first 200 samples.

```
rename = [4 8 3 7 6 5 1 2];
predictedLabels(1:100) = rename(predictedLabels(1:100));
rename = [7 8 3 1 6 5 4 2];
predictedLabels(101:200) = rename(predictedLabels(101:200));
```

Create a confusion matrix chart from the true labels `trueLabels` and the predicted labels `predictedLabels`.

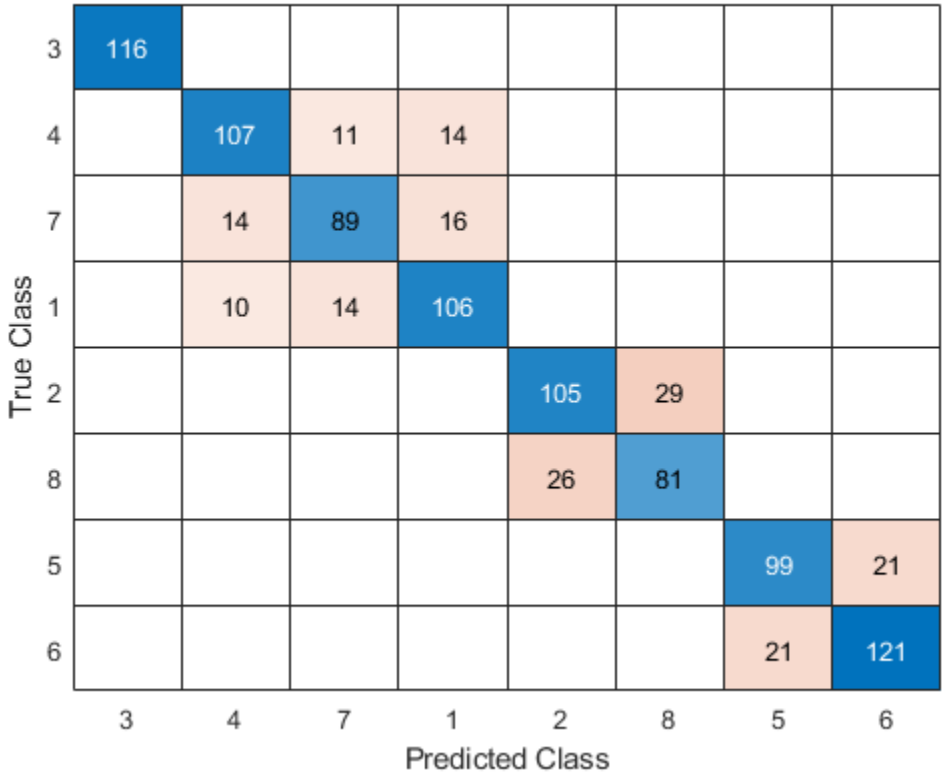
```
figure
cm1 = confusionchart(trueLabels,predictedLabels);
```



Cluster Using `'cluster'`

Sort the classes to cluster similar classes by using the `'cluster'` option.

```
sortClasses(cm1,'cluster')
```



Cluster Using `pdist`, `linkage`, and `optimalleaforder`

Instead of using the `'cluster'` option, you can use the `pdist`, `linkage`, and `optimalleaforder` functions to cluster confusion matrix values. You can customize clustering by using the options of these functions. For details, see the corresponding function reference pages.

Suppose you have a confusion matrix and class labels.

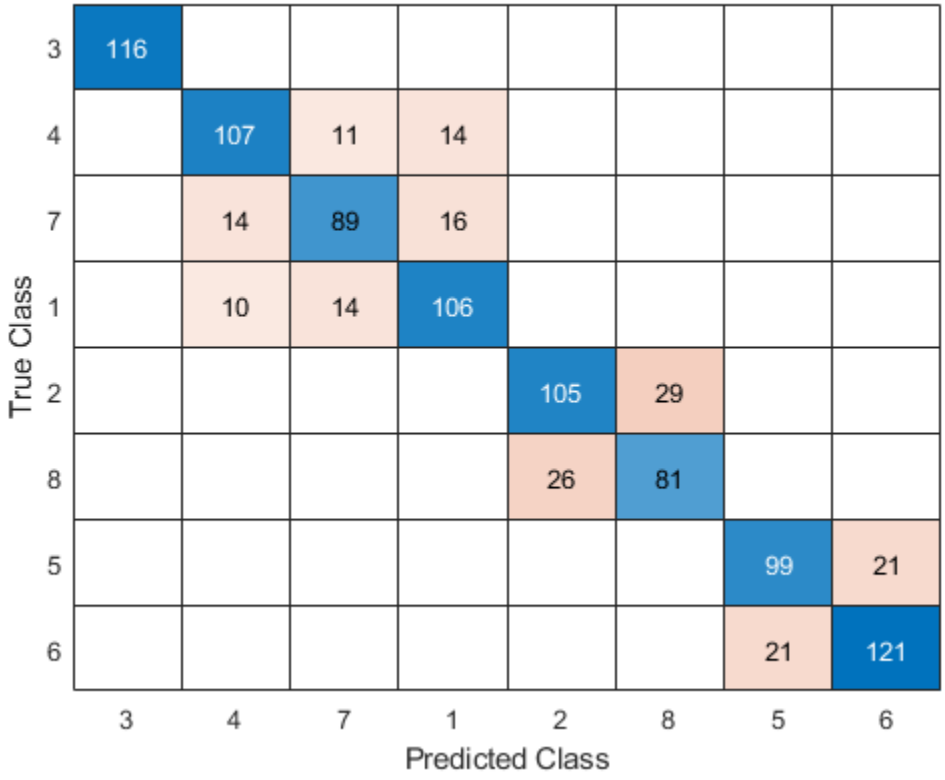
```
m = confusionmat(trueLabels,predictedLabels);
labels = [1 2 3 4 5 6 7 8];
```

Compute the clustered matrix and find the corresponding class labels by using `pdist`, `linkage`, and `optimalleaforder`. The `pdist` function computes the Euclidean distance `D` between pairs of the confusion matrix values. The `optimalleaforder` function returns an optimal leaf ordering for the hierarchical binary cluster tree `linkage(D)` using the distance `D`.

```
D = pdist(m);
idx = optimalleaforder(linkage(D),D);
clusteredM = m(idx,idx);
clusteredLabels = labels(idx);
```

Create a confusion matrix chart using the clustered matrix and the corresponding class labels. Then, sort the classes using the class labels.

```
cm2 = confusionchart(clusteredM,clusteredLabels);
sortClasses(cm2,clusteredLabels)
```



The sorted confusion matrix chart `cm2`, which you created by using `pdist`, `linkage`, and `optimalleaforder`, is identical to the sorted confusion matrix chart `cm1`, which you created by using the `'cluster'` option.

Sort Classes in Fixed Order

Create a confusion matrix chart and sort the classes of the chart in a fixed order.

Load Fisher's iris data set.

```
load fisheriris
X = meas([51:150,1:50],:);
Y = species([51:150,1:50],:);
```

`X` is a numeric matrix that contains four petal measurements for 150 irises. `Y` is a cell array of character vectors that contains the corresponding iris species.

Train a *k*-nearest neighbor (KNN) classifier, where the number of nearest neighbors in the predictors (*k*) is 5. A good practice is to standardize numeric predictor data.

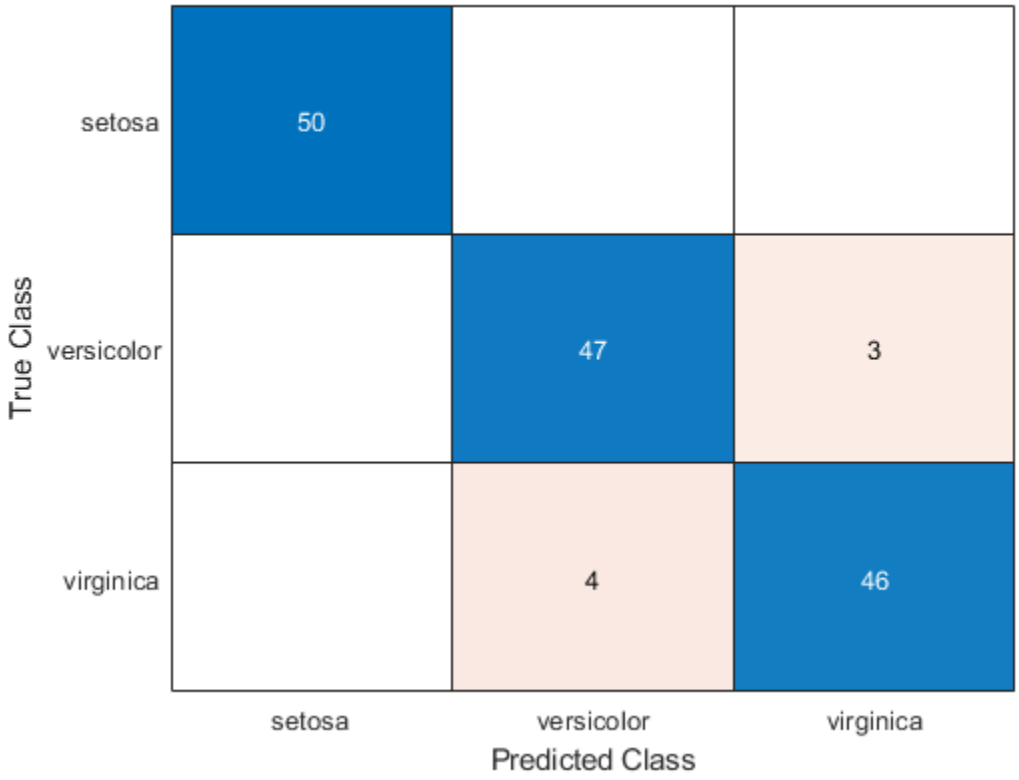
```
Mdl = fitcknn(X,Y,'NumNeighbors',5,'Standardize',1);
```

Predict the labels of the training data.

```
predictedY = resubPredict(Mdl);
```

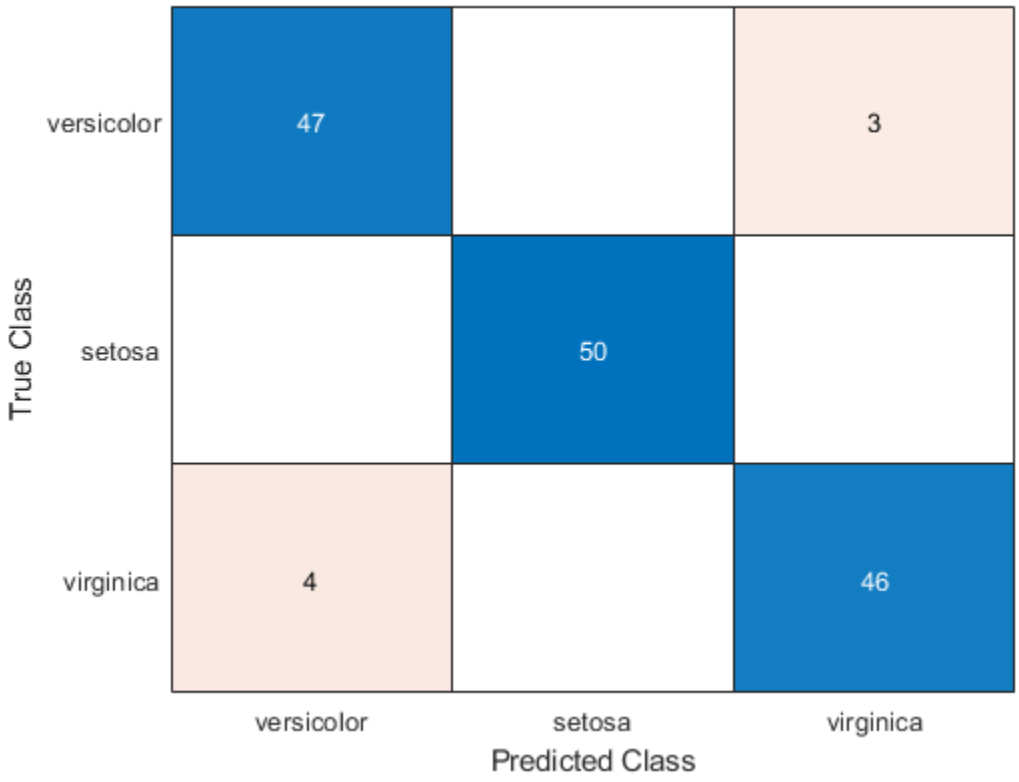
Create a confusion matrix chart from the true labels `Y` and the predicted labels `predictedY`.

```
cm = confusionchart(Y,predictedY);
```



By default, `confusionchart` sorts the classes into their natural order as defined by `sort`. In this example, the class labels are character vectors, so `confusionchart` sorts the classes alphabetically. Reorder the classes of the confusion matrix chart in a fixed order.

```
sortClasses(cm,["versicolor","setosa","virginica"])
```



Input Arguments

`cm` — Confusion matrix chart

`ConfusionMatrixChart` object

Confusion matrix chart, specified as a `ConfusionMatrixChart` object. To create a confusion matrix chart, use `confusionchart`,

`order` — Order in which to sort classes

`'auto'` | `'ascending-diagonal'` | `'descending-diagonal'` | `'cluster'` | array

Order in which to sort the classes of the confusion matrix chart, specified as one of these values:

- `'auto'` — Sorts the classes into their natural order as defined by the `sort` function. For example, if the class labels of the confusion matrix chart are a string vector, then sort alphabetically. If the class labels are an ordinal categorical vector, then use the order of the class labels.
- `'ascending-diagonal'` — Sort the classes so that the values along the diagonal of the confusion matrix increase from top left to bottom right.
- `'descending-diagonal'` — Sort the classes so that the values along the diagonal of the confusion matrix decrease from top left to bottom right.
- `'cluster'` — Sort the classes to cluster similar classes. You can customize clustering by using the `pdist`, `linkage`, and `optimalleaforder` functions. For details, see “Sort Classes to Cluster Similar Classes” on page 33-5924.

- Array — Sort the classes in a unique order specified by a categorical vector, numeric vector, string vector, character array, cell array of character vectors, or logical vector. The array must be a permutation of the `ClassLabels` property of the confusion matrix chart.

Example: `sortClasses(cm,'ascending-diagonal')`

Example: `sortClasses(cm,["owl","cat","toad"])`

See Also

Functions
`categorical` | `confusionchart` | `linkage` | `optimalleaforder` | `pdist`

Properties
ConfusionMatrixChart Properties

Introduced in R2018b

sortrows

Class: `dataset`

(Not Recommended) Sort rows of dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = sortrows(A)
B = sortrows(A,vars)
B = sortrows(A,'obsnames')
B = sortrows(A,vars,mode)
[B,idx] = sortrows(A)
```

Description

`B = sortrows(A)` returns a copy of the dataset array `A`, with the observations sorted in ascending order by all of the variables in `A`. The observations in `B` are sorted first by the first variable, next by the second variable, and so on. Each variable in `A` must be a valid input to `sort`, or, if a variable has multiple columns, to the MATLAB `sortrows` function or to its own`sortrows` method.

`B = sortrows(A,vars)` sorts the observations in `A` by the variables specified by `vars`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector.

`B = sortrows(A,'obsnames')` sorts the observations in `A` by observation name.

`B = sortrows(A,vars,mode)` sorts in the direction specified by `mode`. When `mode` is `'ascend'` (the default) or `'descend'`, `sortrows` sorts `A` by the variables specified by `vars` in ascending or descending order, respectively. `mode` can also be a string array or cell array containing `'ascend'` or `'descend'`, to specify a different sorting direction for each variable in `vars`. Specify `[]` for `vars` to sort using all variables.

`[B,idx] = sortrows(A)` also returns an index vector `idx` such that `B = A(idx,:)`.

Examples

Sort the data in `hospital.mat` by age and then by last name:

```
load hospital
hospital(1:5,1:3)
ans =
            LastName       Sex       Age
 YPL-320    'SMITH'        Male      38
 GLI-532    'JOHNSON'      Male      43
 PNI-258    'WILLIAMS'     Female    38
 MIJ-579    'JONES'        Female    40
```

```
 XLK-030    'BROWN'        Female    49

hospital = sortrows(hospital,{'Age','LastName'});
hospital(1:5,1:3)
ans =
            LastName       Sex       Age
 REV-997    'ALEXANDER'    Male      25
 FZR-250    'HALL'         Male      25
 LIM-480    'HILL'         Female    25
 XUE-826    'JACKSON'      Male      25
 SCQ-914    'JAMES'        Male      25
```

Sort the data in `hospital` by gender in ascending order, and age in descending order.

```
hospital = sortrows(hospital,{'Sex','Age'},{'ascend','descend'});
hospital(1:5,1:3)
ans =

               LastName        Sex       Age
    XLK-030    'BROWN'         Female    49
    GGU-691    'HUGHES'        Female    49
    KKL-155    'ADAMS'         Female    48
    HQO-561    'BRYANT'        Female    48
    BKD-785    'CLARK'         Female    48

hospital(end-4:end,1:3)
ans =

               LastName           Sex     Age
    VNL-702    'MOORE'            Male    28
    REV-997    'ALEXANDER'        Male    25
    FZR-250    'HALL'             Male    25
    XUE-826    'JACKSON'          Male    25
    SCQ-914    'JAMES'            Male    25
```

See Also

`dataset` | `unique`

Topics

“Sort Observations in Dataset Arrays” on page 2-82
“Dataset Arrays” on page 2-112

sparsefilt

Feature extraction by using sparse filtering

Syntax

```
Mdl = sparsefilt(X,q)
Mdl = sparsefilt(X,q,Name,Value)
```

Description

`Mdl = sparsefilt(X,q)` returns a sparse filtering model object that contains the results from applying sparse filtering to the table or matrix of predictor data `X` containing p variables. `q` is the number of features to extract from `X`, therefore `sparsefilt` learns a p-by-`q` matrix of transformation weights. For undercomplete or overcomplete feature representations, `q` can be less than or greater than the number of predictor variables, respectively.

- To access the learned transformation weights, use `Mdl.TransformWeights`.
- To transform `X` to the new set of features by using the learned transformation, pass `Mdl` and `X` to `transform`.

`Mdl = sparsefilt(X,q,Name,Value)` uses additional options specified by one or more `Name,Value` pair arguments. For example, you can standardize the predictor data or apply L^2 regularization.

Examples

Create Sparse Filter

Create a `SparseFiltering` object by using the `sparsefilt` function.

Load the `SampleImagePatches` image patches.

```
data = load('SampleImagePatches');
size(data.X)
```

```
ans = 1×2

        5000         363
```

There are 5,000 image patches, each containing 363 features.

Extract 100 features from the data.

```
rng default % For reproducibility
Q = 100;
obj = sparsefilt(data.X,Q,'IterationLimit',100)
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

```
obj =
  SparseFiltering
           ModelParameters: [1x1 struct]
             NumPredictors: 363
        NumLearnedFeatures: 100
                        Mu: []
                     Sigma: []
                   FitInfo: [1x1 struct]
          TransformWeights: [363x100 double]
    InitialTransformWeights: []


  Properties, Methods
```

`sparsefilt` issues a warning because it stopped due to reaching the iteration limit, instead of reaching a step-size limit or a gradient-size limit. You can still use the learned features in the returned object by calling the `transform` function.

Restart `sparsefilt`

Continue optimizing a sparse filter.

Load the `SampleImagePatches` image patches.

```
data = load('SampleImagePatches');
size(data.X)
```

```
ans = 1×2

        5000         363
```

There are 5,000 image patches, each containing 363 features.

Extract 100 features from the data and use an iteration limit of 20.

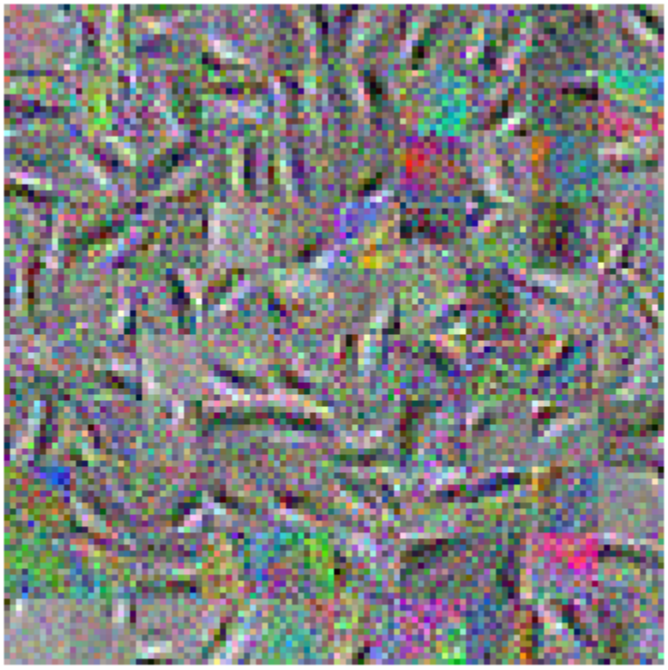
```
rng default % For reproducibility
q = 100;
Mdl = sparsefilt(data.X,q,'IterationLimit',20);
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

View the resulting transformation matrix as image patches.

```
wts = Mdl.TransformWeights;
W = reshape(wts,[11,11,3,q]);
[dx,dy,~,~] = size(W);
for f = 1:q
    Wvec = W(:,:,:,f);
    Wvec = Wvec(:);
    Wvec =(Wvec - min(Wvec))/(max(Wvec) - min(Wvec));
    W(:,:,:,f) = reshape(Wvec,dx,dy,3);
end
m   = ceil(sqrt(q));
n   = m;
```

```
img = zeros(m*dx,n*dy,3);
f   = 1;
for i = 1:m
    for j = 1:n
        if (f <= q)
            img((i-1)*dx+1:i*dx,(j-1)*dy+1:j*dy,:) = W(:,:,:,f);
            f = f+1;
        end
    end
end
imshow(img,'InitialMagnification',300);
```



The image patches appear noisy. To clean up the noise, try more iterations. Restart the optimization from where it stopped for another 40 iterations.

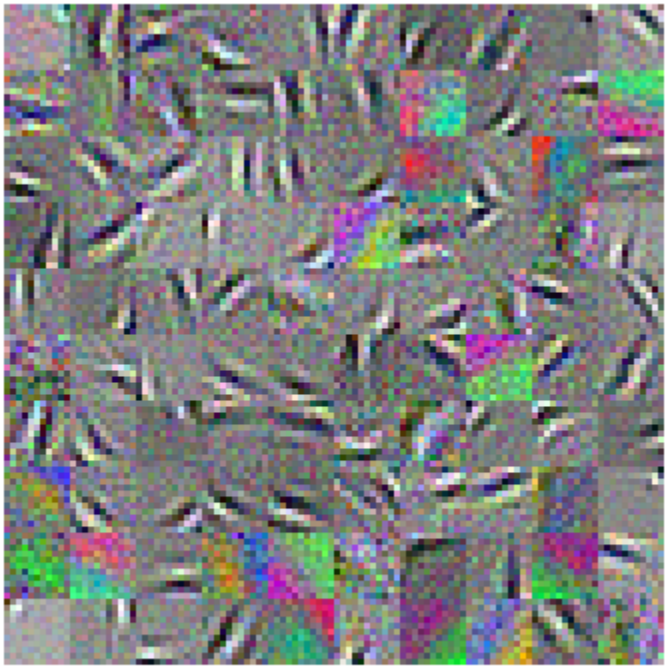
```
Mdl = sparsefilt(data.X,q,'IterationLimit',40,'InitialTransformWeights',wts);
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

View the updated transformation matrix as image patches.

```
wts = Mdl.TransformWeights;
W = reshape(wts,[11,11,3,q]);
[dx,dy,~,~] = size(W);
for f = 1:q
    Wvec = W(:,:,:,f);
    Wvec = Wvec(:);
    Wvec =(Wvec - min(Wvec))/(max(Wvec) - min(Wvec));
```

```
    W(:,:,:,f) = reshape(Wvec,dx,dy,3);
end
m   = ceil(sqrt(q));
n   = m;
img = zeros(m*dx,n*dy,3);
f   = 1;
for i = 1:m
    for j = 1:n
        if (f <= q)
            img((i-1)*dx+1:i*dx,(j-1)*dy+1:j*dy,:) = W(:,:,:,f);
            f = f+1;
        end
    end
end
imshow(img,'InitialMagnification',300);
```



These images are less noisy.

Input Arguments

X — Predictor data

numeric matrix | table

Predictor data, specified as an *n*-by-*p* numeric matrix or table. Rows correspond to individual observations and columns correspond to individual predictor variables. If X is a table, then all of its variables must be numeric vectors.

Data Types: `single` | `double` | `table`

q — Number of features to extract

positive integer

Number of features to extract from the predictor data, specified as a positive integer.

`sparsefilt` stores a *p*-by-`q` transform weight matrix in `Mdl.TransformWeights`. Therefore, setting very large values for `q` can result in greater memory consumption and increased computation time.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Standardize',true,'Lambda',1` standardizes the predictor data and applies a penalty of 1 to the transform weight matrix.

IterationLimit — Maximum number of iterations

`1000` (default) | positive integer

Maximum number of iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

Example: `'IterationLimit',1e6`

Data Types: `single` | `double`

VerbosityLevel — Verbosity level

`0` (default) | nonnegative integer

Verbosity level for monitoring algorithm convergence, specified as the comma-separated pair consisting of `'VerbosityLevel'` and a value in this table.

Value	Description
`0`	`sparsefilt` does not display convergence information at the command line.
Positive integer	`sparsefilt` displays convergence information at the command line.

Convergence Information

Heading	Meaning
FUN VALUE	Objective function value.
NORM GRAD	Norm of the gradient of the objective function.
NORM STEP	Norm of the iterative step, meaning the distance between the previous point and the current point.
CURV	OK means the weak Wolfe condition is satisfied. This condition is a combination of sufficient decrease of the objective function and a curvature condition.
GAMMA	Inner product of the step times the gradient difference, divided by the inner product of the gradient difference with itself. The gradient difference is the gradient at the current point minus the gradient at the previous point. Gives diagnostic information on the objective function curvature.
ALPHA	Step direction multiplier, which differs from 1 when the algorithm performed a line search.
ACCEPT	YES means the algorithm found an acceptable step to take.

Example: `'VerbosityLevel',1`

Data Types: `single` | `double`

`Lambda` — L^2 regularization coefficient value
`0` (default) | positive numeric scalar

L^2 regularization coefficient value for the transform weight matrix, specified as the comma-separated pair consisting of `'Lambda'` and a positive numeric scalar. If you specify `0`, the default, then there is no regularization term in the objective function.

Example: `'Lambda',0.1`

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictor data
`false` (default) | `true`

Flag to standardize the predictor data, specified as the comma-separated pair consisting of `'Standardize'` and `true` (`1`) or `false` (`0`).

If `Standardize` is `true`, then:

- `sparsefilt` centers and scales each column of the predictor data (`X`) by the column mean and standard deviation, respectively.
- `sparsefilt` extracts new features by using the standardized predictor matrix, and stores the predictor variable means and standard deviations in properties `Mu` and `Sigma` of `Mdl`.

Example: `'Standardize',true`

Data Types: `logical`

`InitialTransformWeights` — Transformation weights that initialize optimization
`randn(p,q)` (default) | numeric matrix

Transformation weights that initialize optimization, specified as the comma-separated pair consisting of `'InitialTransformWeights'` and a *p*-by-`q` numeric matrix. *p* must be the number of columns or variables in `X` and `q` is the value of `q`.

Tip You can continue optimizing a previously returned transform weight matrix by passing it as an initial value in another call to `sparsefilt`. The output model object `Mdl` stores a learned transform weight matrix in the `TransformWeights` property.

Example: `'InitialTransformWeights',Mdl.TransformWeights`

Data Types: `single` | `double`

`GradientTolerance` — Relative convergence tolerance on gradient norm
`1e-6` (default) | positive numeric scalar

Relative convergence tolerance on gradient norm, specified as the comma-separated pair consisting of `'GradientTolerance'` and a positive numeric scalar. This gradient is the gradient of the objective function.

Example: `'GradientTolerance',1e-4`

Data Types: `single` | `double`

`StepTolerance` — Absolute convergence tolerance on step size
`1e-6` (default) | positive numeric scalar

Absolute convergence tolerance on the step size, specified as the comma-separated pair consisting of `'StepTolerance'` and a positive numeric scalar.

Example: `'StepTolerance',1e-4`

Data Types: `single` | `double`

Output Arguments

`Mdl` — Learned sparse filtering model
`SparseFiltering` model object

Learned sparse filtering model, returned as a `SparseFiltering` model object.

To access properties of `Mdl`, use dot notation. For example:

- To access the learned transform weights, use `Mdl.TransformWeights`.
- To access the fitting information structure, use `Mdl.FitInfo`.

To find sparse filtering coefficients for new data, use the `transform` function.

Algorithms

The `sparsefilt` function creates a nonlinear transformation of input features to output features. The transformation is based on optimizing an objective function that encourages the representation of each example by as few output features as possible while at the same time keeping the output features equally active across examples.

For details, see “Sparse Filtering Algorithm” on page 15-130.

See Also

`SparseFiltering` | `rica` | `transform`

Topics

“Feature Extraction Workflow” on page 15-135
“Feature Extraction” on page 15-130

Introduced in R2017a

SparseFiltering

Feature extraction by sparse filtering

Description

`SparseFiltering` uses sparse filtering to learn a transformation that maps input predictors to new predictors.

Creation

Create a `SparseFiltering` object using the `sparsefilt` function.

Properties

`FitInfo` — Fitting history
structure

This property is read-only.

Fitting history, returned as a structure with two fields:

- `Iteration` — Iteration numbers from 0 through the final iteration.
- `Objective` — Objective function value at each corresponding iteration. Iteration 0 corresponds to the initial values, before any fitting.

Data Types: `struct`

`InitialTransformWeights` — Initial feature transformation weights
`p`-by-`q` matrix

This property is read-only.

Initial feature transformation weights, returned as a `p`-by-`q` matrix, where `p` is the number of predictors passed in `X` and `q` is the number of features that you want. These weights are the initial weights passed to the creation function. The data type is single when the training data `X` is single.

Data Types: `single` | `double`

`ModelParameters` — Parameters used for training model
structure

This property is read-only.

Parameters used for training the model, returned as a structure. The structure contains a subset of the fields that corresponds to the `sparsefilt` name-value pairs that were in effect during model creation:

- `IterationLimit`

- `VerbosityLevel`
- `Lambda`
- `Standardize`
- `GradientTolerance`
- `StepTolerance`

For details, see the `sparsefilt` name-value pairs in the documentation.

Data Types: `struct`

`Mu` — Predictor means when standardizing
`p`-by-`1` vector

This property is read-only.

Predictor means when standardizing, returned as a `p`-by-`1` vector. This property is nonempty when the `Standardize` name-value pair is `true` at model creation. The value is the vector of predictor means in the training data. The data type is single when the training data `X` is single.

Data Types: `single` | `double`

`NumLearnedFeatures` — Number of output features
positive integer

This property is read-only.

Number of output features, returned as a positive integer. This value is the `q` argument passed to the creation function, which is the requested number of features to learn.

Data Types: `double`

`NumPredictors` — Number of input predictors
positive integer

This property is read-only.

Number of input predictors, returned as a positive integer. This value is the number of predictors passed in `X` to the creation function.

Data Types: `double`

`Sigma` — Predictor standard deviations when standardizing
`p`-by-`1` vector

This property is read-only.

Predictor standard deviations when standardizing, returned as a `p`-by-`1` vector. This property is nonempty when the `Standardize` name-value pair is `true` at model creation. The value is the vector of predictor standard deviations in the training data. The data type is single when the training data `X` is single.

Data Types: `single` | `double`

`TransformWeights` — Feature transformation weights
`p`-by-`q` matrix

This property is read-only.

Feature transformation weights, returned as a p-by-q matrix, where p is the number of predictors passed in X and q is the number of features that you want. The data type is single when the training data X is single.

Data Types: `single` | `double`

Object Functions

transform Transform predictors into extracted features

Examples

Create Sparse Filter

Create a `SparseFiltering` object by using the `sparsefilt` function.

Load the `SampleImagePatches` image patches.

```
data = load('SampleImagePatches');
size(data.X)
```

```
ans = 1×2

        5000         363
```

There are 5,000 image patches, each containing 363 features.

Extract 100 features from the data.

```
rng default % For reproducibility
Q = 100;
obj = sparsefilt(data.X,Q,'IterationLimit',100)
```

```
Warning: Solver LBFGS was not able to converge to a solution.

obj =
  SparseFiltering
           ModelParameters: [1x1 struct]
             NumPredictors: 363
         NumLearnedFeatures: 100
                         Mu: []
                      Sigma: []
                    FitInfo: [1x1 struct]
          TransformWeights: [363x100 double]
    InitialTransformWeights: []


  Properties, Methods
```

`sparsefilt` issues a warning because it stopped due to reaching the iteration limit, instead of reaching a step-size limit or a gradient-size limit. You can still use the learned features in the returned object by calling the `transform` function.

See Also

`ReconstructionICA` | `rica` | `sparsefilt` | `transform`

Topics

“Feature Extraction Workflow” on page 15-135
“Feature Extraction” on page 15-130

Introduced in R2017a

spectralcluster

Spectral clustering

Syntax

```
idx = spectralcluster(X,k)

idx = spectralcluster(S,k,'Distance','precomputed')

idx = spectralcluster( ___ ,Name,Value)
[idx,V] = spectralcluster( ___ )
[idx,V,D] = spectralcluster( ___ )
```

Description

`idx = spectralcluster(X,k)` partitions observations in the *n*-by-*p* data matrix `X` into `k` clusters using the spectral clustering algorithm (see Algorithms on page 33-5959). `spectralcluster` returns an *n*-by-1 vector `idx` containing cluster indices of each observation.

`idx = spectralcluster(S,k,'Distance','precomputed')` returns a vector of cluster indices for `S`, the similarity matrix on page 33-5958 (or adjacency matrix) of a similarity graph on page 33-5958. `S` can be the output of `adjacency`.

To use a similarity matrix as the first input, you must specify `'Distance','precomputed'`.

`idx = spectralcluster( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments in addition to the input arguments in previous syntaxes. For example, you can specify `'SimilarityGraph','epsilon'` to construct a similarity graph using the radius search method.

`[idx,V] = spectralcluster( ___ )` also returns the eigenvectors `V` corresponding to the `k` smallest eigenvalues of the Laplacian matrix on page 33-5959.

`[idx,V,D] = spectralcluster( ___ )` also returns a vector `D` containing the `k` smallest eigenvalues of the Laplacian matrix.

Examples

Perform Spectral Clustering on Input Data

Cluster a 2-D circular data set using spectral clustering with the default Euclidean distance metric.

Generate synthetic data that contains two noisy circles.

```
rng('default') % For reproducibility

% Parameters for data generation
N = 300;  % Size of each cluster
r1 = 2;   % Radius of first circle
r2 = 4;   % Radius of second circle
```

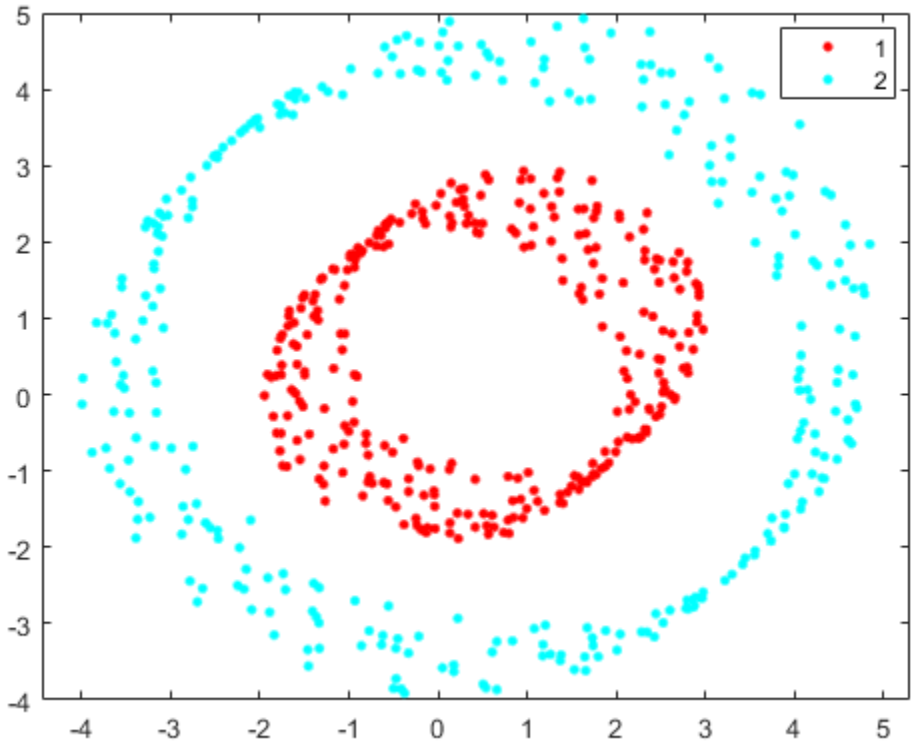
```
theta = linspace(0,2*pi,N)';

X1 = r1*[cos(theta),sin(theta)]+ rand(N,1);
X2 = r2*[cos(theta),sin(theta)]+ rand(N,1);
X = [X1;X2]; % Noisy 2-D circular data set
```

Find two clusters in the data by using spectral clustering.

```
idx = spectralcluster(X,2);
```

Visualize the result of clustering.

```
gscatter(X(:,1),X(:,2),idx);
```



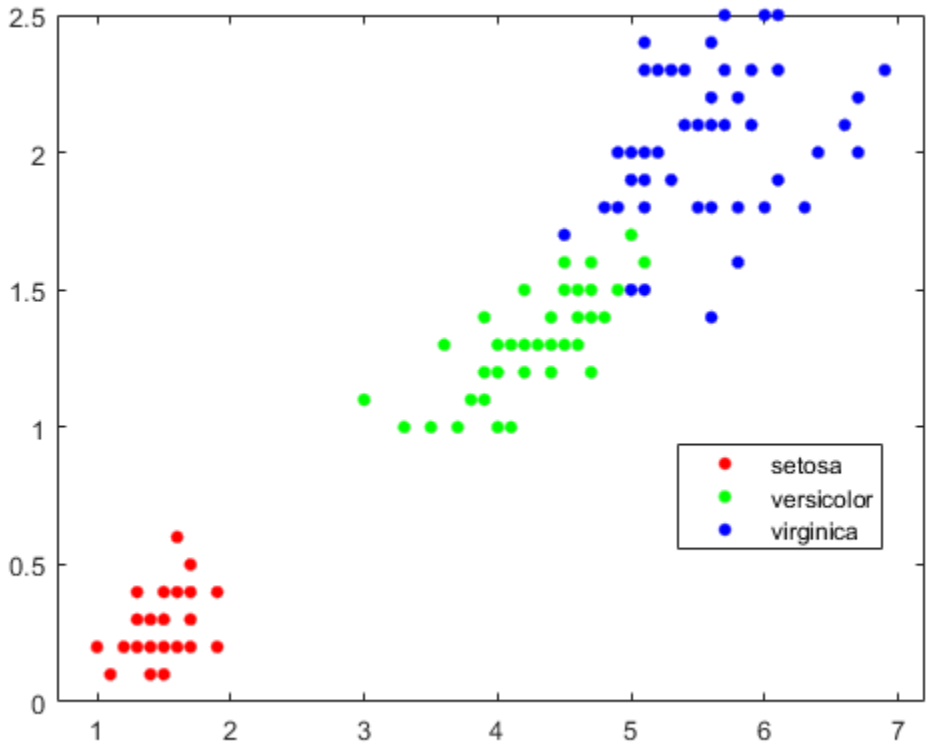
The `spectralcluster` function correctly identifies the two clusters in the data set.

Perform Spectral Clustering on Similarity Matrix

Compute a similarity matrix from Fisher's iris data set and perform spectral clustering on the similarity matrix.

Load Fisher's iris data set. Use the petal lengths and widths as features to consider for clustering.

```
load fisheriris
X = meas(:,3:4);
gscatter(X(:,1),X(:,2),species);
```



Find the distance between each pair of observations in X by using the `pdist` and `squareform` functions with the default Euclidean distance metric.

```
dist_temp = pdist(X);
dist = squareform(dist_temp);
```

Construct the similarity matrix and confirm that it is symmetric.

```
S = exp(-dist.^2);
issymmetric(S)
```

```
ans = logical
   1
```

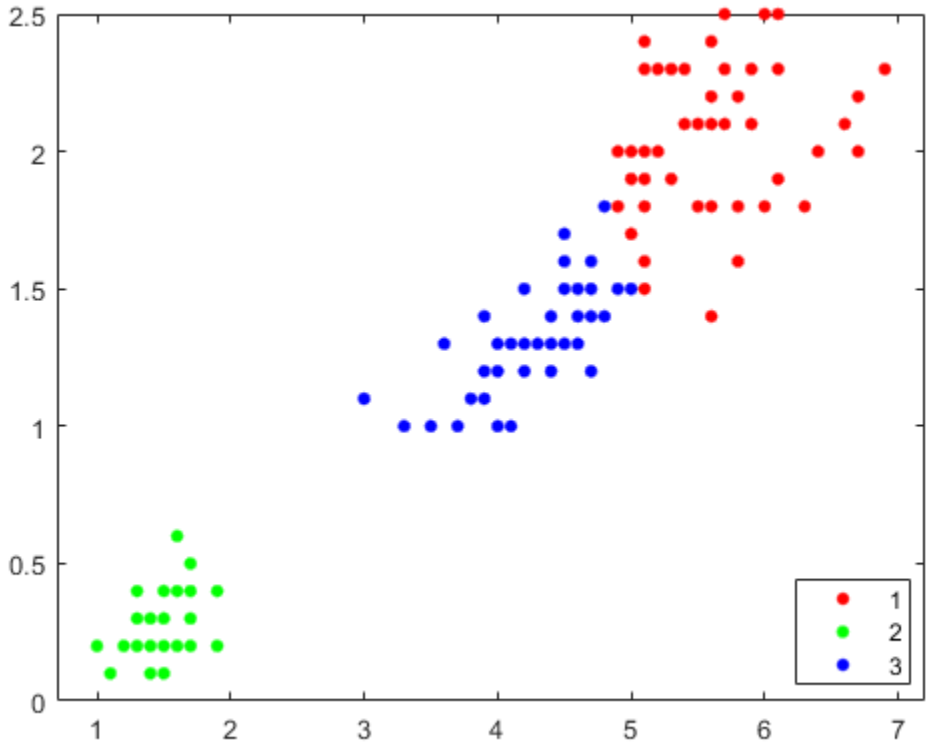
Perform spectral clustering. Specify `'Distance','precomputed'` to perform clustering using the similarity matrix. Specify `k=3` clusters, and set the `'LaplacianNormalization'` name-value pair argument to use the normalized symmetric Laplacian matrix.

```
k = 3; % Number of clusters
rng('default') % For reproducibility
idx = spectralcluster(S,k,'Distance','precomputed','LaplacianNormalization','symmetric');
```

`idx` contains the cluster indices for each observation in X.

Visualize the result of clustering.

```
gscatter(X(:,1),X(:,2),idx);
```



Tabulate the clustering results.

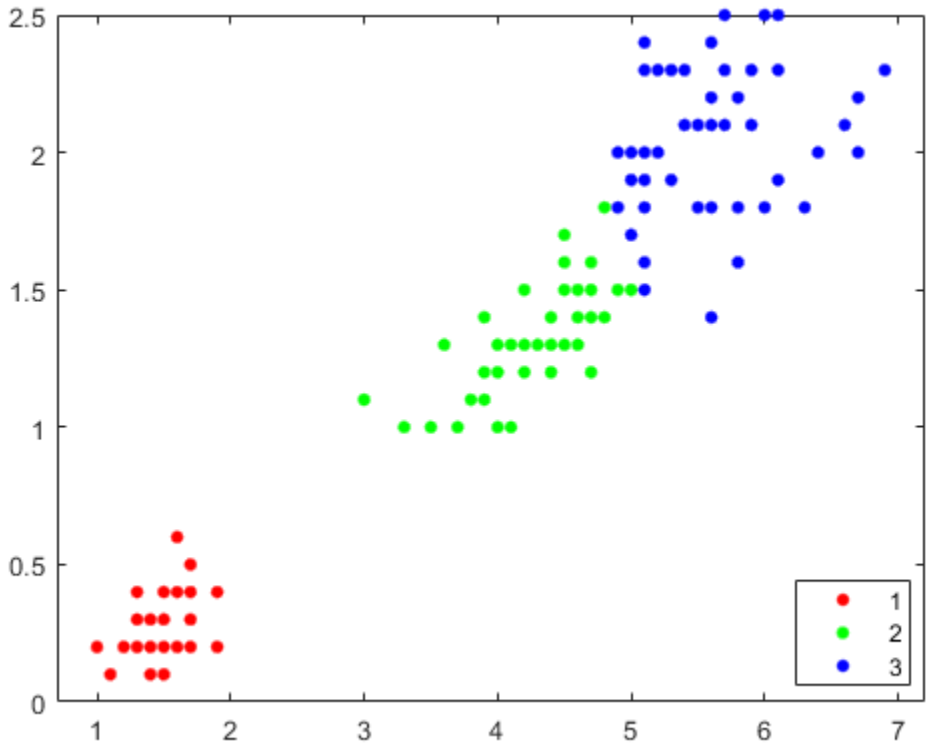
```
tabulate(idx)
```

```
  Value    Count   Percent
      1       48     32.00%
      2       50     33.33%
      3       52     34.67%
```

The `Percent` column shows the percentage of data points assigned to the three clusters.

Repeat spectral clustering using the data as input to `spectralcluster`. Specify `'NumNeighbors'` as `size(X,1)`, which corresponds to creating the similarity matrix `S` by connecting each point to all the remaining points.

```
idx2 = spectralcluster(X,k,'NumNeighbors',size(X,1),'LaplacianNormalization','symmetric');
gscatter(X(:,1),X(:,2),idx2);
```




```
tabulate(idx2)
```

```
  Value    Count   Percent
      1       50     33.33%
      2       52     34.67%
      3       48     32.00%
```

The clustering results for both approaches are the same. The order of cluster assignments is different, even though the data points are clustered in the same way.

Cluster Using Radius Search for Similarity Graph

Find clusters in a data set, based on a specified search radius for creating a similarity graph.

Create data with 3 clusters, each containing 500 points.

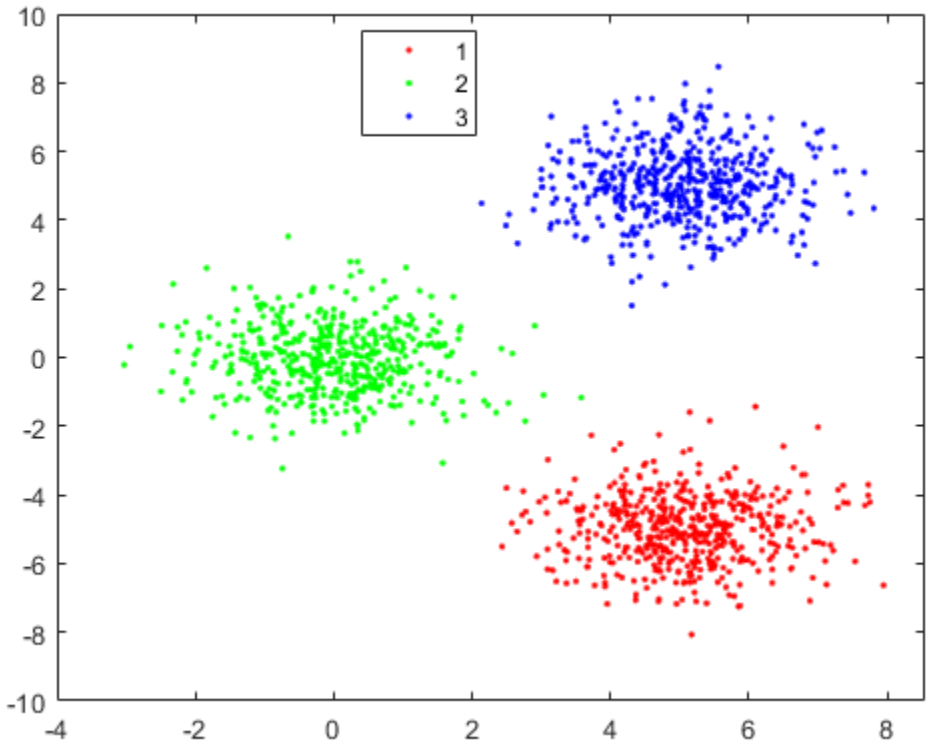
```
rng('default') % For reproducibility
N = 500;
X = [mvnrnd([0 0],eye(2),N); ...
    mvnrnd(5*[1 -1],eye(2),N); ...
    mvnrnd(5*[1 1],eye(2),N)];
```

Specify a search radius of 2 for creating a similarity graph, and find 3 clusters in the data.

```
idx = spectralcluster(X,3,'SimilarityGraph','epsilon','Radius',2);
```

Visualize the result of clustering.

```
gscatter(X(:,1),X(:,2),idx);
```



Find Eigenvalues and Eigenvectors of Laplacian Matrix

Find the eigenvalues and eigenvectors of the Laplacian matrix and use the values to confirm clustering results.

Randomly generate sample data with three well-separated clusters, each containing 100 points.

```
rng('default'); % For reproducibility
n = 100;
X = [randn(n,2)*0.5+3;
     randn(n,2)*0.5
     randn(n,2)*0.5-3];
```

Estimate the number of clusters in the data by using the eigenvalues of the Laplacian matrix. Compute the five smallest eigenvalues (in magnitude) of the Laplacian matrix.

```
[~,~,D_temp] = spectralcluster(X,5)
```

```
D_temp = 5×1

   -0.0000
```

```
   -0.0000
   -0.0000
    0.0277
    0.0296
```

Only the first three eigenvalues are approximately zero. The number of zero eigenvalues is a good indicator of the number of connected components in a similarity graph and, therefore, is a good estimate of the number of clusters in your data. So, `k=3` is a good estimate of the number of clusters in `X`.

Find `k=3` clusters and return the three smallest eigenvalues and corresponding eigenvectors of the Laplacian matrix.

```
[idx,V,D] = spectralcluster(X,3)
```

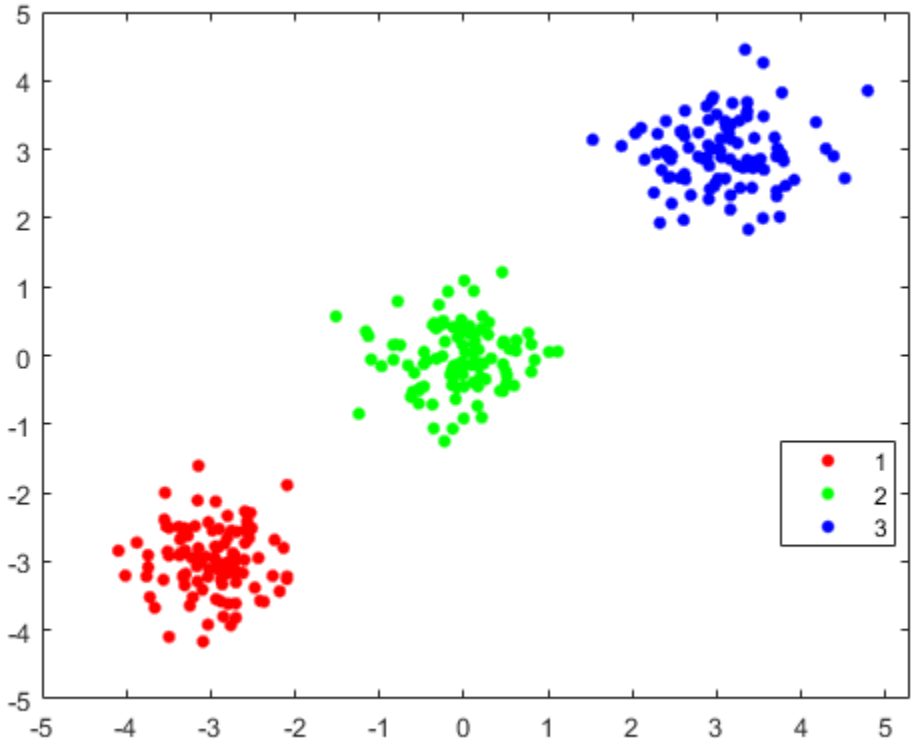
```
idx = 300×1

     3
     3
     3
     3
     3
     3
     3
     3
     3
     3
       ⋮


V = 300×3

   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
   -0.0000   -0.0000   -0.1000
      ⋮


D = 3×1
10^-16 ×

   -0.3308
   -0.3747
   -0.4167
```

Elements of `D` correspond to the three smallest eigenvalues of the Laplacian matrix. The columns of `V` contain the eigenvectors corresponding to the eigenvalues in `D`. For well-separated clusters, the eigenvectors are indicator vectors. The eigenvectors have values of zero (or close to zero) for points

that do not belong to a particular cluster, and nonzero values for points that belong to a particular cluster.

Visualize the result of clustering.

```
gscatter(X(:,1),X(:,2),idx);
```



Input Arguments

X — Input data

numeric matrix

Input data, specified as an *n*-by-*p* numeric matrix. The rows of `X` correspond to observations (or points), and the columns correspond to variables.

The software treats `NaN`s in `X` as missing data and ignores any row of `X` containing at least one `NaN`. The `spectralcluster` function returns `NaN` values for the corresponding row in the output arguments `idx` and `V`.

Data Types: `single` | `double`

S — Similarity matrix

symmetric matrix

Similarity matrix, specified as an *n*-by-*n* symmetric matrix, where *n* is the number of observations. A similarity matrix (or adjacency matrix) represents the input data by modeling local neighborhood

relationships among the data points. The values in a similarity matrix represent the edges (or connections) between nodes (data points) that are connected in a similarity graph on page 33-5958. For more information, see “Similarity Matrix” on page 33-5958.

`S` must not contain any `NaN` values.

To use a similarity matrix as the first input of `spectralcluster`, you must specify `'Distance','precomputed'`.

Data Types: `single` | `double`

k — Number of clusters

positive integer

Number of clusters in the data, specified as a positive integer.

For details about how to estimate the number of clusters, see “Tips” on page 33-5959.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `spectralcluster(X,3,'SimilarityGraph','epsilon','Radius',5)` specifies 3 clusters and uses the radius search method with a search radius of 5 to construct a similarity graph.

`Distance` — Distance metric

character vector | string scalar | function handle

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a character vector, string scalar, or function handle, as described in this table.

Value	Description
`'precomputed'`	Precomputed distance. You must specify this option if the first input to `spectralcluster` is a similarity matrix `S`.
`'euclidean'`	Euclidean distance (default)
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between observations is scaled by dividing by the corresponding element of the standard deviation computed from `X`. Use the `Scale` name-value pair argument to specify a different scaling factor.
`'mahalanobis'`	Mahalanobis distance using the sample covariance of `X`, `C = cov(X,'omitrows')`. Use the `Cov` name-value pair argument to specify a different covariance matrix.
`'cityblock'`	City block distance
`'minkowski'`	Minkowski distance. The default exponent is 2. Use the `P` name-value pair argument to specify a different exponent, where `P` is a positive scalar value.
`'chebychev'`	Chebychev distance (maximum coordinate difference)

Value	Description
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors)
`'correlation'`	One minus the sample correlation between observations (treated as sequences of values)
`'hamming'`	Hamming distance, which is the percentage of coordinates that differ
`'jaccard'`	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values)

`@`*`distfun`*	Custom distance function handle. A distance function has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of distance`
	`...`
	where
	• `ZI` is a `1`-by-`n` vector containing a single observation.
	• `ZJ` is an `m2`-by-`n` matrix containing multiple observations. `distfun` must accept a matrix `ZJ` with an arbitrary number of observations.
	• `D2` is an `m2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.
	If your data is not sparse, you can generally compute distance more quickly by using a built-in distance instead of a function handle.

For more information, see “Distance Metrics” on page 18-12.

When you use the `'seuclidean'`, `'minkowski'`, or `'mahalanobis'` distance metric, you can specify the additional name-value pair argument `'Scale'`, `'P'`, or `'Cov'`, respectively, to control the distance metric.

Example: `spectralcluster(X,5,'Distance','minkowski','P',3)` specifies 5 clusters and uses of the Minkowski distance metric with an exponent of 3 to perform the clustering algorithm.

P — Exponent for Minkowski distance metric
2 (default) | positive scalar

Exponent for the Minkowski distance metric, specified as the comma-separated pair consisting of `'P'` and a positive scalar.

This argument is valid only if `'Distance'` is `'minkowski'`.

Example: `'P',3`

Data Types: `single` | `double`

Cov — Covariance matrix for Mahalanobis distance metric
`cov(X,'omitrows')` (default) | positive definite matrix

Covariance matrix for the Mahalanobis distance metric, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix.

This argument is valid only if `'Distance'` is `'mahalanobis'`.

Example: `'Cov',eye(4)`

Data Types: `single` | `double`

`Scale` — Scaling factors for standardized Euclidean distance metric
`std(X,'omitnan')` (default) | numeric vector of nonnegative values

Scaling factors for the standardized Euclidean distance metric, specified as the comma-separated pair consisting of `'Scale'` and a numeric vector of nonnegative values.

`Scale` has length *p* (the number of columns in `X`), because each dimension (column) of `X` has a corresponding value in `Scale`. For each dimension of `X`, `spectralcluster` uses the corresponding value in `Scale` to standardize the difference between observations.

This argument is valid only if `'Distance'` is `'seuclidean'`.

Data Types: `single` | `double`

`SimilarityGraph` — Type of similarity graph
`'knn'` (default) | `'epsilon'`

Type of similarity graph to construct from the input data `X`, specified as the comma-separated pair consisting of `'SimilarityGraph'` and one of these values.

Value	Description	Graph-Specific Name-Value Pair Arguments
`'knn'`	(Default) Construct the graph using nearest neighbors.	`'NumNeighbors'` — Number of nearest neighbors used to construct the similarity graph `'KNNGraphType'` — Type of nearest neighbor graph
`'epsilon'`	Construct the graph using a radius search. You must specify a value for `Radius` if you use this option.	`'Radius'` — Search radius for the nearest neighbors used to construct the similarity graph

For more information, see “Similarity Graph” on page 33-5958.

This argument is valid only if `'Distance'` is not `'precomputed'`.

Example: `'SimilarityGraph','epsilon'`

`NumNeighbors` — Number of nearest neighbors
`log(size(X,1))` (default) | positive integer

Number of nearest neighbors used to construct the similarity graph, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer.

This argument is valid only if `'SimilarityGraph'` is `'knn'`. For more information, see “Similarity Graph” on page 33-5958.

Example: `'NumNeighbors',10`

Data Types: `single` | `double`

KNNGraphType — Type of nearest neighbor graph
`'complete'` (default) | `'mutual'`

Type of nearest neighbor graph, specified as the comma-separated pair consisting of `'KNNGraphType'` and one of these values.

Value	Description
`'complete'`	(Default) Connects two points *i* and *j*, when either *i* is a nearest neighbor of *j* or *j* is a nearest neighbor of *i*. This option leads to a denser representation of the similarity matrix.
`'mutual'`	Connects two points *i* and *j*, when *i* is a nearest neighbor of *j* and *j* is a nearest neighbor of *i*. This option leads to a sparser representation of the similarity matrix.

This argument is valid only if `'SimilarityGraph'` is `'knn'`.

Example: `'KNNGraphType','mutual'`

Radius — Search radius
nonnegative scalar

Search radius for the nearest neighbors used to construct the similarity graph, specified as the comma-separated pair consisting of `'Radius'` and a nonnegative scalar.

You must specify this argument if `'SimilarityGraph'` is `'epsilon'`. For more information, see “Similarity Graph” on page 33-5958.

Example: `'Radius',5`

Data Types: `single` | `double`

KernelScale — Scale factor
1 (default) | `'auto'` | positive scalar

Scale factor for the kernel, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software uses the scale factor to transform distances to similarity measures. For more information, see “Similarity Graph” on page 33-5958.

- The `'auto'` option is supported only for the `'euclidean'` and `'seuclidean'` distance metrics.
- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. To reproduce results, set a random number seed using `rng` before calling `spectralcluster`.

This argument is valid only if `'Distance'` is not `'precomputed'`.

Example: `'KernelScale','auto'`

Data Types: `double` | `single` | `char` | `string`

LaplacianNormalization — Method to normalize Laplacian matrix
`'randomwalk'` (default) | `'symmetric'` | `'none'`

Method to normalize the Laplacian matrix L, specified as the comma-separated pair consisting of `'LaplacianNormalization'` and one of these values.

Value	Description
`'none'`	Use Laplacian matrix L without normalization.
`'randomwalk'`	(Default) Use the normalized random-walk Laplacian matrix L_{rw} (Shi-Malik [2]). $L_{rw} = D_g^{-1}L$. The matrix D_g is the degree matrix on page 33-5959.
`'symmetric'`	Use the normalized symmetric Laplacian matrix L_s (Ng-Jordan-Weiss [3]). $L_s = D_g^{-1/2}LD_g^{-1/2}$. The matrix D_g is the degree matrix.

For more information, see “Laplacian Matrix” on page 33-5959.

Example: `'LaplacianNormalization','randomwalk'`

`ClusterMethod` — Clustering method
`'kmeans'` (default) | `'kmedoids'`

Clustering method to cluster the eigenvectors of the Laplacian matrix, specified as the comma-separated pair consisting of `'ClusterMethod'` and either `'kmeans'` or `'kmedoids'`.

- `'kmeans'` — Perform k-means clustering by using the `kmeans` function.
- `'kmedoids'` — Perform k-medoids clustering by using the `kmedoids` function.

`kmeans` and `kmedoids` involve randomness in their algorithms. Therefore, to reproduce the results of `spectralcluster`, you must set the seed of the random number generator by using `rng`.

Example: `'ClusterMethod','kmedoids'`

Output Arguments

`idx` — Cluster indices
numeric column vector

Cluster indices, returned as a numeric column vector. `idx` has n rows, and each row of `idx` indicates the cluster assignment of the corresponding row (or observation) in `X`.

Data Types: `double`

`V` — Eigenvectors
numeric matrix

Eigenvectors, returned as an n-by-`k` numeric matrix. The columns of `V` are the eigenvectors corresponding to the `k` smallest eigenvalues of the Laplacian matrix on page 33-5959. These eigenvectors are a low-dimensional representation of the input data `X` in a new space where clusters are more widely separated.

For well-separated clusters, the eigenvectors are indicator vectors. That is, the eigenvectors have values of zero (or close to zero) for points that do not belong to a given cluster, and nonzero values for points that belong to a particular cluster.

Data Types: `single` | `double`

`D` — Eigenvalues
numeric vector

Eigenvalues, returned as a `k`-by-1 numeric vector that contains the `k` smallest eigenvalues of the Laplacian matrix. The number of zero eigenvalues in `D` is an indicator of the number of connected components in the similarity graph and, therefore, is a good estimate of the number of clusters in your data.

Data Types: `single` | `double`

More About

Similarity Graph

A similarity graph models the local neighborhood relationships between data points in `X` as an undirected graph. The nodes in the graph represent data points, and the edges, which are directionless, represent the connections between the data points.

If the pairwise distance $Dist_{i,j}$ between any two nodes i and j is positive (or larger than a certain threshold), then the similarity graph connects the two nodes using an edge [1]. The edge between the two nodes is weighted by the pairwise similarity $S_{i,j}$, where $S_{i,\,j} = \exp\left(-\left(\frac{Dist_{i,\,j}}{\sigma}\right)^2\right)$, for a specified kernel scale σ value.

`spectralcluster` supports these two methods of constructing a similarity graph:

- *Nearest neighbor* method (if `'SimilarityGraph'` is `'knn'`(default)): `spectralcluster` connects points in `X` that are nearest neighbors. You can use the `'NumNeighbors'` and `'KNNGraphType'` name-value pair arguments to specify the options for constructing the nearest neighbor graph.
  - Use `'NumNeighbors'` to specify the number of nearest neighbors.
  - Use `'KNNGraphType'` to specify whether to make a `'complete'` or `'mutual'` connection of points.
- *Radius search* method (if `'SimilarityGraph'` is `'epsilon'`): `spectralcluster` connects points whose pairwise distances are smaller than a search radius. You must specify the search radius for nearest neighbors used to construct the similarity graph by using the `'Radius'` name-value pair argument.

Similarity Matrix

A similarity matrix is a matrix representation of a similarity graph on page 33-5958. The n-by-n matrix $S = (S_{i,\,j})_{i,\,j\,=\,1,\,\ldots,\,n}$ contains pairwise similarity values between connected nodes in the similarity graph. The similarity matrix of a graph is also called an adjacency matrix.

The similarity matrix is symmetric because the edges of the similarity graph are directionless. A value of $S_{i,\,j} = 0$ means that nodes i and j of the similarity graph are not connected.

Degree Matrix

A degree matrix D_g is an n-by-n diagonal matrix obtained by summing the rows of the similarity matrix on page 33-5958 S. That is, the ith diagonal element of D_g is $D_g(i,i) = \sum_{j=1}^{n} S_{i,j}$.

Laplacian Matrix

A Laplacian matrix is one way of representing a similarity graph on page 33-5958. The `spectralcluster` function supports the unnormalized Laplacian matrix, the normalized Laplacian matrix using the Shi-Malik method [2], and the normalized Laplacian matrix using the Ng-Jordan-Weiss method [3].

- The *unnormalized Laplacian matrix L* is the difference between the degree matrix on page 33-5959 and the similarity matrix on page 33-5958.

  $$L = D_g - S.$$

- The *normalized random-walk Laplacian matrix (Shi-Malik)* is defined as:

  $$L_{rw} = D_g^{-1}L.$$

  To derive L_{rw}, solve the generalized eigenvalue problem $Lv = \lambda D_g v$, where v is a column vector of length n, and λ is a scalar. The values of λ that satisfy the equation are the generalized eigenvalues of the matrix $L_{rw} = D_g^{-1}L$.

  You can use the MATLAB function `eigs` to solve the generalized eigenvalue problem.

- The *normalized symmetric Laplacian matrix (Ng-Jordan-Weiss)* is defined as:

  $$L_s = D_g^{-1/2}LD_g^{-1/2}.$$

Use the `'LaplacianNormalization'` name-value pair argument to specify the method to normalize the Laplacian matrix.

Tips

- Consider using spectral clustering when the clusters in your data do not naturally correspond to convex regions.
- From the spectral clustering algorithm, you can estimate the number of clusters `k` as:
  - The number of eigenvalues of the Laplacian matrix that are equal to `0`.
  - The number of connected components in your similarity graph representation. Use `graph` to create a similarity graph from a similarity matrix, and use `conncomp` to find the number of connected components in the graph.

  For an example, see “Estimate Number of Clusters and Perform Spectral Clustering” on page 16-27.

Algorithms

Spectral clustering is a graph-based algorithm for clustering data points (or observations in `X`). The algorithm involves constructing a graph, finding its Laplacian matrix on page 33-5959, and using this

matrix to find k eigenvectors to split the graph k ways. By default, the algorithm for `spectralcluster` computes the normalized random-walk Laplacian matrix using the method described by Shi-Malik [2]. `spectralcluster` also supports the unnormalized Laplacian matrix and the normalized symmetric Laplacian matrix which uses the Ng-Jordan-Weiss method [3]. `spectralcluster` implements clustering as follows:

1. For each data point in `X`, define a local neighborhood using either the radius search method or nearest neighbor method, as specified by the `'SimilarityGraph'` name-value pair argument (see "Similarity Graph" on page 33-5958). Then, find the pairwise distances $Dist_{i,j}$ for all points i and j in the neighborhood.
2. Convert the distances to similarity measures using the kernel transformation $S_{i,j} = \exp\left(-\left(\frac{Dist_{i,j}}{\sigma}\right)^2\right)$. The matrix S is the similarity matrix on page 33-5958, and σ is the scale factor for the kernel, as specified using the `'KernelScale'` name-value pair argument.
3. Calculate the unnormalized Laplacian matrix on page 33-5959 L , the normalized random-walk Laplacian matrix L_{rw}, or the normalized symmetric Laplacian matrix L_s, depending on the value of the `'LaplacianNormalization'` name-value pair argument.
4. Create a matrix $V \in \mathbb{R}^{n \times k}$ containing columns $v_1, \ldots, v_k$, where the columns are the k eigenvectors that correspond to the k smallest eigenvalues of the Laplacian matrix. If using L_s, normalize each row of V to have unit length.
5. Treating each row of V as a point, cluster the n points using k-means clustering (default) or k-medoids clustering, as specified by the `'ClusterMethod'` name-value pair argument.
6. Assign the original points in `X` to the same clusters as their corresponding rows in V.

References

[1] Von Luxburg, U. "A Tutorial on Spectral Clustering." *Statistics and Computing Journal*. Vol.17, Number 4, 2007, pp. 395–416.

[2] Shi, J., and J. Malik. "Normalized cuts and image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 22, 2000, pp. 888–905.

[3] Ng, A.Y., M. Jordan, and Y. Weiss. "On spectral clustering: Analysis and an algorithm." In *Proceedings of the Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 849–856.

See Also

`adjacency` | `eigs` | `kmeans` | `kmedoids` | `pdist` | `squareform`

Topics

"Partition Data Using Spectral Clustering" on page 16-26

Introduced in R2019b

squareform

Format distance matrix

Syntax

```
ZOut = squareform(yIn)
yOut = squareform(ZIn)

ZOut = squareform(yIn,'tomatrix')
yOut = squareform(ZIn,'tovector')
```

Description

`ZOut = squareform(yIn)` converts `yIn`, a pairwise distance vector of length $m(m–1)/2$ for m observations, into `ZOut`, an m-by-m symmetric matrix with zeros along the diagonal.

The pairwise distances in `yIn` are arranged in the order (2,1), (3,1), ..., (m,1), (3,2), ..., (m,2), ..., (m,m–1). The pairwise distance between the `i`th and `j`th observations is in `ZOut(i,j)` and `yIn((i–1)*(m–i/2)+j–i)` for $i \le j$.

`yOut = squareform(ZIn)` converts `ZIn`, a square, symmetric matrix with zeros along the diagonal, into `yOut`, a vector containing the `ZIn` elements below the diagonal.

`ZOut = squareform(yIn,'tomatrix')` forces `squareform` to treat `yIn` as a vector and converts `yIn` into a matrix.

`yOut = squareform(ZIn,'tovector')` forces `squareform` to treat `ZIn` as a matrix and converts `ZIn` into a vector. If `ZIn` is a scalar (1-by-1), then `ZIn` must be zero.

The previous two syntaxes are useful when the input argument is a scalar. If you do not specify either `'tomatrix'` or `'tovector'`, then the default is `'tomatrix'`.

Examples

Compute Euclidean Distance and Convert Distance Vector to Matrix

Compute the Euclidean distance between pairs of observations, and convert the distance vector to a matrix using `squareform`.

Create a matrix with three observations and two variables.

```
rng('default') % For reproducibility
X = rand(3,2);
```

Compute the Euclidean distance.

```
D = pdist(X)
```

```
D = 1×3
```

```
    0.2954    1.0670    0.9448
```

The pairwise distances are arranged in the order (2,1), (3,1), (3,2). You can easily locate the distance between observations `i` and `j` by using `squareform`.

```
Z = squareform(D)
```

```
Z = 3×3

         0    0.2954    1.0670
    0.2954         0    0.9448
    1.0670    0.9448         0
```

`squareform` returns a symmetric matrix where `Z(i,j)` corresponds to the pairwise distance between observations `i` and `j`. For example, you can find the distance between observations 2 and 3.

```
Z(2,3)
```

```
ans = 0.9448
```

Pass `Z` to the `squareform` function to reproduce the output of the `pdist` function.

```
y = squareform(Z)
```

```
y = 1×3

    0.2954    1.0670    0.9448
```

The outputs `y` from `squareform` and `D` from `pdist` are the same.

Input Arguments

`yIn` — Input distance vector

numeric vector | logical vector

Input distance vector, specified as a numeric or logical vector of length $m(m–1)/2$, where m is the number of observations.

The pairwise distances in `yIn` are arranged in the order (2,1), (3,1), ..., (m,1), (3,2), ..., (m,2), ..., (m,m–1), i.e., the lower-left triangle of the m-by-m distance matrix in column order. The pairwise distance between observations i and j is in `yIn((i–1)*(m–i/2)+j–i)` for $i \le j$.

You can create `yIn` by using the `pdist` function. m is the number of observations in the input data of `pdist`.

Data Types: `single` | `double` | `logical`

`ZIn` — Input distance matrix

numeric matrix | logical matrix

Input distance matrix, specified as a numeric or logical matrix. `ZIn` is an m-by-m symmetric matrix with zeros along the diagonal, where m is the number of observations. `ZIn(i,j)` denotes the distance between the `i`th and `j`th observations.

Data Types: `single` | `double` | `logical`

Output Arguments

`yOut` — Distance vector

numeric vector | logical vector

Distance vector, returned as a numeric or logical vector of length $m(m–1)/2$, where m is the number of observations.

The pairwise distances in `yOut` are arranged in the order (2,1), (3,1), ..., (m,1), (3,2), ..., (m,2), ..., (m,m–1), i.e., the lower-left triangle of the m-by-m distance matrix in column order. The pairwise distance between observations i and j is in `yOut((i–1)*(m–i/2)+j–i)` for $i \leq j$.

`yOut` has the same format as the output from the `pdist` function.

`ZOut` — Distance matrix

numeric matrix | logical matrix

Distance matrix, returned as a numeric or logical matrix. `ZOut` is an m-by-m symmetric matrix with zeros along the diagonal, where m is the number of observations. `ZOut(i,j)` denotes the distance between the `i`th and `j`th observations.

Tips

- You can use `squareform` to format a vector or matrix that is similar to a distance vector or matrix, such as the correlation coefficient matrix (`corrcoef`).

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations: The conversion direction `'tomatrix'` or `'tovector'` must be a compile-time constant. For example, to specify the conversion direction as `'tovector'`, include `coder.Constant('tovector')` in the `-args` value of `codegen`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

`pdist`

Introduced before R2006a

stack

Class: `dataset`

(Not Recommended) Stack dataset array from multiple variables into single variable

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = stack(A,datavars)
[B,iA] = stack(A,datavars)
B = stack(A,datavars,Parameter,value)
```

Description

`B = stack(A,datavars)` stacks multiple variables in dataset array `A` into a single variable in `B`. In general, `B` contains fewer variables but more observations than `A`.

`datavars` specifies a group of `m` data variables in `A`. `stack` creates a single data variable in `B` by interleaving their values, and if `A` has `n` observations, then `B` has `m`-by-`n` observations. In other words, `stack` takes the `m` data values from each observation in `A` and stacks them up to create `m` observations in `B`. `datavars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. `stack` also creates a grouping variable in `B` to indicate which of the `m` data variables in `A` each observation in `B` corresponds to.

`stack` assigns values for the "per-variable properties (e.g., `Units` and `VarDescription`) for the new data variable in `B` from the corresponding property values for the first variable listed in `datavars`.

`stack` copies the remaining variables from `A` to `B` without stacking, by replicating each of their values `m` times. These variables are typically grouping variables. Because their values are constant across each group of `m` observations in `B`, they identify which observation in `A` an observation in `B` came from.

`[B,iA] = stack(A,datavars)` returns an index vector `iA` indicating the correspondence between observations in `B` and those in `A`. `stack` creates `B(j,:)` using `A(iA(j),datavarss)`.

For more information on grouping variables, see “Grouping Variables” on page 2-45.

Input Arguments

`B = stack(A,datavars,Parameter,value)` uses the following parameter name/value pairs to control how `stack` converts variables in `A` to variables in `B`:

`'ConstVars'`	Variables in `A` to copy to `B` without stacking. `ConstVars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. The default is all variables in `A` not specified in `datavars`.
`'NewDataVarName'`	A name for the data variable to be created in `B`. The default is a concatenation of the names of the `m` variables that are stacked up.
`'IndVarName'`	A name for the grouping variable to create in `B` to indicate the source of each value in the new data variable. The default is based on the `'NewDataVarName'` parameter.

You can also specify multiple groups of data variables in `A`, each of which becomes a variable in `B`. All groups must contain the same number of variables. Use a string array or cell array of character vectors to contain multiple parameter values for `datavars` or to contain multiple values for `'NewDataVarName'`.

Examples

Combine several variables for estimated influenza rates into a single variable. Then unstack the estimated influenza rates by date.

```
load flu

% FLU has a 'Date' variable, and 10 variables for estimated influenza rates
% (in 9 different regions, estimated from Google searches, plus a
% nationwide estimate from the CDC). Combine those 10 variables into an
% array that has a single data variable, 'FluRate', and an indicator
% variable, 'Region', that says which region each estimate is from.
[flu2,iflu] = stack(flu, 2:11, 'NewDataVarName','FluRate', ...
    'IndVarName','Region')

% The second observation in FLU is for 10/16/2005.  Find the observations
% in FLU2 that correspond to that date.
flu(2,:)
flu2(iflu==2,:)

% Use the 'Date' variable from that array to split 'FluRate' into 52
% separate variables, each containing the estimated influenza rates for
% each unique date.  The new array has one observation for each region.  In
% effect, this is the original array FLU "on its side".
dateNames = cellstr(datestr(flu.Date,'mmm_DD_YYYY'));
[flu3,iflu2] = unstack(flu2, 'FluRate', 'Date', ...
    'NewDataVarNames',dateNames)

% Since observations in FLU3 represent regions, IFLU2 indicates the first
% occurrence in FLU2 of each region.
flu2(iflu2,:)
```

See Also

`join` | `unstack`

Topics
“Grouping Variables” on page 2-45

State property

Class: `qrandstream`

Current state of the stream

Description

The `State` property of a quasi-random stream contains the index into the associated point set of the next point to draw in the stream. Getting and resetting the `State` property allows you to return a stream to a previous state. The initial value of `State` is `1`.

Examples

```
Q = qrandstream('sobol', 5);
s = Q.State;
u1 = qrand(Q, 10)
Q.State = s;
u2 = qrand(Q, 10) % contains exactly the same values as u1
```

See Also

`qrand`

statget

Access values in statistics options structure

Syntax

```
val = statget(options,param)
val = statget(options,param,default)
```

Description

`val = statget(options,param)` returns the value of the parameter specified by `param` in the statistics options structure `options`. The input `param` is a character vector or a string scalar of the parameter name. If the parameter is undefined in `options`, `statget` returns `[]`. You need to type only enough leading characters to define the parameter name uniquely. `statget` ignores case for parameter names. For available `options`, see Inputs.

`val = statget(options,param,default)` returns `default` if the specified parameter is undefined in the optimization options structure `options`.

Input Arguments

`DerivStep`

Relative difference used in finite difference derivative calculations. A positive scalar, or a vector of positive scalars the same size as the vector of parameters estimated by the Statistics and Machine Learning Toolbox function using the options structure.

`Display`

Amount of information displayed by the algorithm.

- `'off'` — Displays no information.
- `'final'` — Displays the final output.
- `'iter'` — Displays iterative output to the command window for some functions; otherwise displays the final output.

`FunValCheck`

Check for invalid values, such as `NaN` or `Inf`, from the objective function.

- `'off'`
- `'on'`

`GradObj`

Flags whether the objective function returns a gradient vector as a second output.

- `'off'`
- `'on'`

`Jacobian`

Flags whether the objective function returns a Jacobian as a second output.

- `'off'`
- `'on'`

`MaxFunEvals`

Maximum number of objective function evaluations allowed. Positive integer.

`MaxIter`

Maximum number of iterations allowed. Positive integer.

`OutputFcn`

The solver calls all output functions after each iteration.

- Function handle specified using @
- a cell array with function handles
- an empty array (default)

`Robust`

Invoke robust fitting option.

- `'off'`
- `'on'`

`RobustWgtFun`

A weight function for robust fitting. Valid only when `Robust` is `'on'`. Can also be a function handle that accepts a normalized residual as input and returns the robust weights as output.

- `'bisquare'`
- `'andrews'`
- `'cauchy'`
- `'fair'`
- `'huber'`
- `'logistic'`
- `'talwar'`
- `'welsch'`

`Streams`

A single instance of the `RandStream` class, or a cell array of `RandStream` instances. The Streams option is accepted by some functions to govern what stream(s) to use in generating random numbers within the function. If `'UseSubstreams'` is `true`, the Streams value must be a scalar, or must be empty. If `'UseParallel'` is `true` and `'UseSubstreams'` is `false`, then the Streams argument must either be empty, or its length must match the number of processors used in the computation: equal to the *parpool* size if a *parpool* is open, a scalar otherwise.

TolBnd

Parameter bound tolerance. Positive scalar.

TolFun

Termination tolerance for the objective function value. Positive scalar.

TolTypeFun

Use `TolFun` for absolute or relative objective function tolerances.

- `'abs'`
- `'rel'`

TolTypeX

Use `TolX` for absolute or relative parameter tolerances.

- `'abs'`
- `'rel'`

TolX

Termination tolerance for the parameters. Positive scalar.

Tune

The tuning constant used in robust fitting to normalize the residuals before applying the weight function. The default value depends upon the weight function. This parameter is necessary if you specify the weight function as a function handle. Positive scalar.

UseParallel

Flag indicating whether eligible functions should use capabilities of the Parallel Computing Toolbox (PCT), if the capabilities are available. That is, if the PCT is installed, and a PCT `parpool` is in effect. Valid values are `false` (the default), for serial computation, and `true`, for parallel computation.

UseSubstreams

Flag indicating whether the random number generator in eligible functions should use `Substream` property of the `RandStream` class. `false` (default) or `true`. When `true`, high level iterations within the function will set the `Substream` property to the value of the iteration. This behavior helps to generate reproducible random number streams in parallel and/or serial mode computation.

WgtFun

A weight function for robust fitting. Valid only when `Robust` is `'on'`. Can also be a function handle that accepts a normalized residual as input and returns the robust weights as output.

- `'bisquare'`
- `'andrews'`
- `'cauchy'`
- `'fair'`

- `'huber'`
- `'logistic'`
- `'talwar'`
- `'welsch'`

Examples

This statement returns the value of the `Display` statistics options parameter from the structure called `my_options`.

```
val = statget(my_options,'Display')
```

Return the value of the `Display` statistics options parameter from the structure called `my_options` (as in the previous example). If the `Display` parameter is undefined, `statget` returns the value `'final'`.

```
optnew = statget(my_options,'Display','final');
```

See Also

`statset`

Introduced before R2006a

statset

Create statistics options structure

Syntax

```
statset
statset(statfun)
options = statset(...)
options = statset(fieldname1,val1,fieldname2,val2,...)
options = statset(oldopts,fieldname1,val1,fieldname2,val2,...)
options = statset(oldopts,newopts)
```

Description

`statset` with no input arguments and no output arguments displays all fields of a statistics options structure and their possible values.

`statset(statfun)` displays fields and default values used by the Statistics and Machine Learning Toolbox function `statfun`. Specify `statfun` using a character vector, a string scalar, or a function handle.

`options = statset(...)` creates a statistics options structure `options`. With no input arguments, all fields of the options structure are an empty array (`[]`). With a specified `statfun`, function-specific fields are default values and the remaining fields are `[]`. Function-specific fields set to `[]` indicate that the function is to use its default value for that parameter. For available `options`, see Inputs.

`options = statset(fieldname1,val1,fieldname2,val2,...)` creates an options structure in which the named fields have the specified values. Any unspecified values are `[]`. Use character vectors or string scalars for field names. For named values, you must input the complete character vector or string scalar for the value. If you provide an invalid character vector or string scalar for a value, `statset` uses the default.

`options = statset(oldopts,fieldname1,val1,fieldname2,val2,...)` creates a copy of `oldopts` with the named parameters changed to the specified values.

`options = statset(oldopts,newopts)` combines an existing options structure, `oldopts`, with a new options structure, `newopts`. Any parameters in `newopts` with nonempty values overwrite corresponding parameters in `oldopts`.

Input Arguments

`DerivStep`

Relative difference used in finite difference derivative calculations. A positive scalar, or a vector of positive scalars the same size as the vector of parameters estimated by the Statistics and Machine Learning Toolbox function using the options structure.

`Display`

Amount of information displayed by the algorithm.

- `'off'` — Displays no information.
- `'final'` — Displays the final output.
- `'iter'` — Displays iterative output to the command window for some functions; otherwise displays the final output.

`FunValCheck`

Check for invalid values, such as `NaN` or `Inf`, from the objective function.

- `'off'`
- `'on'`

`GradObj`

Flags whether the objective function returns a gradient vector as a second output.

- `'off'`
- `'on'`

`Jacobian`

Flags whether the objective function returns a Jacobian as a second output.

- `'off'`
- `'on'`

`MaxFunEvals`

Maximum number of objective function evaluations allowed. Positive integer.

`MaxIter`

Maximum number of iterations allowed. Positive integer.

`OutputFcn`

The solver calls all output functions after each iteration.

- Function handle specified using @
- a cell array with function handles
- an empty array (default)

`Robust`

(Not recommended) Invoke robust fitting option.

- `'off'`
- `'on'`

`Robust` is not recommended. Use `RobustWgtFun` for robust fitting.

RobustWgtFun

Weight function for robust fitting. Can also be a function handle that accepts a normalized residual as input and returns the robust weights as output. If you use a function handle, give a `Tune` constant. See “Robust Options” on page 33-5976.

Streams

A single instance of the `RandStream` class, or a cell array of `RandStream` instances. The Streams option is accepted by some functions to govern what stream(s) to use in generating random numbers within the function. If `'UseSubstreams'` is `true`, the Streams value must be a scalar, or must be empty. If `'UseParallel'` is `true` and `'UseSubstreams'` is `false`, then the Streams argument must either be empty, or its length must match the number of processors used in the computation: equal to the *parpool* size if a *parpool* is open, a scalar otherwise.

TolBnd

Parameter bound tolerance. Positive scalar.

TolFun

Termination tolerance for the objective function value. Positive scalar.

TolTypeFun

Use `TolFun` for absolute or relative objective function tolerances.

- `'abs'`
- `'rel'`

TolTypeX

Use `TolX` for absolute or relative parameter tolerances.

- `'abs'`
- `'rel'`

TolX

Termination tolerance for the parameters. Positive scalar.

Tune

Tuning constant used in robust fitting to normalize the residuals before applying the weight function. The default value depends upon the weight function. This parameter is necessary if you specify the weight function as a function handle. Positive scalar. See “Robust Options” on page 33-5976.

UseParallel

Flag indicating whether eligible functions should use capabilities of the Parallel Computing Toolbox (PCT), if the capabilities are available. That is, if the PCT is installed, and a PCT `parpool` is in effect. Valid values are `false` (the default), for serial computation, and `true`, for parallel computation.

UseSubstreams

Flag indicating whether the random number generator in eligible functions should use `Substream` property of the `RandStream` class. `false` (default) or `true`. When `true`, high level iterations within

the function will set the `Substream` property to the value of the iteration. This behavior helps to generate reproducible random number streams in parallel and/or serial mode computation.

`WgtFun`

(Not recommended) Weight function for robust fitting. Valid only when `Robust` is `'on'`. Can also be a function handle that accepts a normalized residual as input and returns the robust weights as output. See “Robust Options” on page 33-5976.

`WgtFun` is not recommended. Use `RobustWgtFun` instead.

Examples

Suppose you want to change the default parameter values for the function `evfit`, which fits an extreme value distribution to data. The defaults parameter values are:

```
statset('evfit')
ans =
          Display: 'off'
      MaxFunEvals: []
          MaxIter: []
           TolBnd: []
           TolFun: []
       TolTypeFun: []
             TolX: 1.0000e-06
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: []
        OutputFcn: []
```

The only parameters that `evfit` uses are `Display` and `TolX`. To create an options structure with the value of `TolX` set to `1e-8`, enter:

```
options = statset('TolX',1e-8)
% Pass options to evfit:
mu = 1;
sigma = 1;
data = evrnd(mu,sigma,1,100);

paramhat = evfit(data,[],[],[],options)
```

More About

Robust Options

Weight Function	Equation	Default Tuning Constant
`'andrews'`	`w = (abs(r)<pi) .* sin(r) ./ r`	1.339
`'bisquare'` (default)	`w = (abs(r)<1) .* (1 - r.^2).^2`	4.685
`'cauchy'`	`w = 1 ./ (1 + r.^2)`	2.385
`'fair'`	`w = 1 ./ (1 + abs(r))`	1.400
`'huber'`	`w = 1 ./ max(1, abs(r))`	1.345
`'logistic'`	`w = tanh(r) ./ r`	1.205
`'talwar'`	`w = 1 * (abs(r)<1)`	2.795
`'welsch'`	`w = exp(-(r.^2))`	2.985
`[]`	No robust fitting	—

See Also

`statget`

Introduced before R2006a

std

Package: prob

Standard deviation of probability distribution

Syntax

```
s = std(pd)
```

Description

`s = std(pd)` returns the standard deviation `s` of the probability distribution `pd`.

Examples

Standard Deviation of a Fitted Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Fit a normal distribution object to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the standard deviation of the fitted distribution.

```
s = std(pd)
```

```
s = 8.7202
```

For a normal distribution, the standard deviation is equal to the parameter `sigma`.

Standard Deviation of a Skewed Distribution

Create a Weibull probability distribution object

```
pd = makedist('Weibull','a',5,'b',2)
```

```
pd =
  WeibullDistribution
```

```
  Weibull distribution
    A = 5
    B = 2
```

Compute the standard deviation of the distribution.

```
s = std(pd)
```

```
s = 2.3163
```

Standard Deviation of a Triangular Distribution

Create a triangular distribution object.

```
pd = makedist('Triangular','a',-3,'b',1,'c',3)
```

```
pd =
  TriangularDistribution

A = -3, B = 1, C = 3
```

Compute the standard deviation of the distribution.

```
s = std(pd)
```

```
s = 1.2472
```

Standard Deviation of a Kernel Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades;
x = grades(:,1);
```

Create a probability distribution object by fitting a kernel distribution to the data.

```
pd = fitdist(x,'Kernel')
```

```
pd =
  KernelDistribution

    Kernel = normal
    Bandwidth = 3.61677
    Support = unbounded
```

Compute the standard deviation of the fitted distribution.

```
s = std(pd)
```

```
s = 9.4069
```

Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

s — Standard deviation

nonnegative scalar value

Standard deviation of the probability distribution, returned as a nonnegative scalar value.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `fitdist` | `makedist` | `mean` | `var`

Topics

“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

step

Improve generalized linear regression model by adding or removing terms

Syntax

```
NewMdl = step(mdl)
NewMdl = step(mdl,Name,Value)
```

Description

`NewMdl = step(mdl)` returns a generalized linear regression model based on `mdl` using stepwise regression to add or remove one predictor.

`NewMdl = step(mdl,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the criterion to use to add or remove terms and the maximum number of steps to take.

Examples

Modify Generalized Linear Regression Model Using `step`

Fit a Poisson regression model using random data and a single predictor, and then use `step` to improve the model by adding or removing predictor terms.

Generate sample data that has 20 predictor variables. Use three of the predictors to generate the Poisson response variable.

```
rng('default') % For reproducibility
X = randn(100,20);
mu = exp(X(:,[5 10 15])*[.4;.2;.3] + 1);
y = poissrnd(mu);
```

Create a generalized linear regression model of Poisson data using `X(:,2)` as the only predictor.

```
mdl = fitglm(X,y,'y ~ x2','Distribution','poisson')
```

```
mdl =
Generalized linear regression model:
    log(y) ~ 1 + x2
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)      1.1386    0.056722     20.073    1.2817e-89
    x2             0.010768    0.056564    0.19037       0.84902


100 observations, 98 error degrees of freedom
```

```
Dispersion: 1
Chi^2-statistic vs. constant model: 0.0362, p-value = 0.849
```

Improve `mdl` by using `step`. Specify `'NSteps'` as 5 to allow at most 5 steps of stepwise regression.

```
mdl1 = step(mdl,'NSteps',5)
```

```
1. Adding x5, Deviance = 134.4375, Chi2Stat = 52.21338, PValue = 4.978574e-13
2. Adding x15, Deviance = 106.1925, Chi2Stat = 28.24496, PValue = 1.068927e-07
3. Adding x10, Deviance = 94.708, Chi2Stat = 11.4845, PValue = 0.000701792
4. Removing x2, Deviance = 95.021, Chi2Stat = 0.31263, PValue = 0.57607

mdl1 =
Generalized linear regression model:
    log(y) ~ 1 + x5 + x10 + x15
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.0115     0.064275    15.737    8.4217e-56
    x5             0.39508     0.066665    5.9263    3.0977e-09
    x10            0.18863      0.05534    3.4085     0.0006532
    x15            0.29295     0.053269    5.4995    3.8089e-08


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 91.7, p-value = 9.61e-20
```

`step` adds the three predictor variables used to generate the response variable to the model and removes `X(:,2)` from the model.

Input Arguments

`mdl` — Generalized linear regression model

`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Criterion','aic','Upper','quadratic','Verbose',2` instructs `step` to use the Akaike information criterion, include (at most) the quadratic terms in the model, and display the evaluation process and the decision taken at each step.

`Criterion` — Criterion to add or remove terms

`'Deviance'` (default) | `'sse'` | `'aic'` | `'bic'` | `'rsquared'` | `'adjrsquared'`

Criterion to add or remove terms, specified as the comma-separated pair consisting of `'Criterion'` and one of these values:

- `'Deviance'` — *p*-value for an *F*-test or chi-squared test of the change in the deviance that results from adding or removing the term. The *F*-test tests a single model, and the chi-squared test compares two different models.
- `'sse'` — *p*-value for an *F*-test of the change in the sum of squared error that results from adding or removing the term.
- `'aic'` — Change in the value of the Akaike information criterion (AIC).
- `'bic'` — Change in the value of the Bayesian information criterion (BIC).
- `'rsquared'` — Increase in the value of R^2.
- `'adjrsquared'` — Increase in the value of adjusted R^2.

Example: `'Criterion','bic'`

`Lower` — Model specification describing terms that cannot be removed
`'constant'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'Y ~ terms'`

Model specification describing terms that cannot be removed from the model, specified as the comma-separated pair consisting of `'Lower'` and one of these values:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and x_1*x_2^2 terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-5986, specifying terms in the model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.

- A character vector or string scalar “Formula” on page 33-5986 in the form
  `'Y ~ terms'`,
  where the `terms` are in “Wilkinson Notation” on page 33-5986. The variable names in the formula must be valid MATLAB identifiers.

Example: `'Lower','linear'`

Data Types: `char` | `string` | `single` | `double`

`NSteps` — Maximum number of steps to take
1 (default) | positive integer

Maximum number of steps to take, specified as the comma-separated pair consisting of `'NSteps'` and a positive integer.

Example: `'NSteps',5`

Data Types: `single` | `double`

`PEnter` — Threshold for criterion to add term
scalar value

Threshold for the criterion to add a term, specified as the comma-separated pair consisting of `'PEnter'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'Deviance'`	0.05	If the *p*-value of the *F*-statistic or chi-squared statistic is less than `PEnter` (*p*-value to enter), add the term to the model.
`'SSE'`	0.05	If the *p*-value of the *F*-statistic is less than `PEnter`, add the term to the model.
`'AIC'`	0	If the change in the AIC of the model is less than `PEnter`, add the term to the model.
`'BIC'`	0	If the change in the BIC of the model is less than `PEnter`, add the term to the model.
`'Rsquared'`	0.1	If the increase in the R-squared value of the model is greater than `PEnter`, add the term to the model.
`'AdjRsquared'`	0	If the increase in the adjusted R-squared value of the model is greater than `PEnter`, add the term to the model.

For more information, see the `Criterion` name-value pair argument.

Example: `'PEnter',0.075`

`PRemove` — Threshold for criterion to remove term
scalar value

Threshold for the criterion to remove a term, specified as the comma-separated pair consisting of `'PRemove'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'Deviance'`	0.10	If the *p*-value of the *F*-statistic or chi-squared statistic is greater than `PRemove` (*p*-value to remove), remove the term from the model.
`'SSE'`	0.10	If the *p*-value of the *F*-statistic is greater than `PRemove`, remove the term from the model.
`'AIC'`	0.01	If the change in the AIC of the model is greater than `PRemove`, remove the term from the model.
`'BIC'`	0.01	If the change in the BIC of the model is greater than `PRemove`, remove the term from the model.
`'Rsquared'`	0.05	If the increase in the R-squared value of the model is less than `PRemove`, remove the term from the model.
`'AdjRsquared'`	-0.05	If the increase in the adjusted R-squared value of the model is less than `PRemove`, remove the term from the model.

At each step, the `step` function also checks whether a term is redundant (linearly dependent) with other terms in the current model. When a term is linearly dependent on other terms in the current model, the `step` function removes the redundant term, regardless of the criterion value.

For more information, see the `Criterion` name-value pair argument.

Example: `'PRemove',0.05`

`Upper` — Model specification describing largest set of terms in fit

`'interactions'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'Y ~ terms'`

Model specification describing the largest set of terms in the fit, specified as the comma-separated pair consisting of `'Upper'` and one of these values:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.

Value	Model Type
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and $x_1*x_2^2$ terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-5986, specifying terms in the model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar “Formula” on page 33-5986 in the form
  `'Y ~ terms'`,
  where the `terms` are in “Wilkinson Notation” on page 33-5986. The variable names in the formula must be valid MATLAB identifiers.

Example: `'Upper','quadratic'`

Data Types: `char` | `string` | `single` | `double`

`Verbose` — Control for display of information
`1` (default) | `0` | `2`

Control for the display of information, specified as the comma-separated pair consisting of `'Verbose'` and one of these values:

- `0` — Suppress all display.
- `1` — Display the action taken at each step.
- `2` — Display the evaluation process and the action taken at each step.

Example: `'Verbose',2`

Output Arguments

`NewMdl` — Generalized linear regression model
`GeneralizedLinearModel` object

Generalized linear regression model, returned as a `GeneralizedLinearModel` object.

To overwrite the input argument `mdl`, assign the new model to `mdl`.

```
mdl = step(mdl);
```

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'`*`y ~ terms`*`'`.

- *`y`* is the response name.
- *`terms`* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.
- `'y ~ x1 + x2 + x3 − 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include `−1` in the formula.

A formula includes a constant term unless you explicitly remove the term using `−1`.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.

- – means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
–x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 – x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- Stepwise regression is a systematic method for adding and removing terms from a linear or generalized linear model based on their statistical significance in explaining the response variable. The method begins with an initial model, specified using `modelspec`, and then compares the explanatory power of incrementally larger and smaller models.

  The `step` function uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add to the model or remove from the model based on the value of the `'Criterion'` name-value pair argument.

  The default value of `'Criterion'` for a linear regression model is `'sse'`. In this case, `stepwiselm` and `step` of `LinearModel` use the *p*-value of an *F*-statistic to test models with and without a potential term at each step. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the function adds the term to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the function removes the term from the model.

  Stepwise regression takes these steps when `'Criterion'` is `'sse'`:

  1. Fit the initial model.
  2. Examine a set of available terms not in the model. If any of the terms have *p*-values less than an entrance tolerance (that is, if it is unlikely a term would have a zero coefficient if added to the model), add the term with the smallest *p*-value and repeat this step; otherwise, go to step 3.

3 If any of the available terms in the model have *p*-values greater than an exit tolerance (that is, the hypothesis of a zero coefficient cannot be rejected), remove the term with the largest *p*-value and return to step 2; otherwise, end the process.

At any stage, the function will not add a higher-order term if the model does not also include all lower-order terms that are subsets of the higher-order term. For example, the function will not try to add the term `X1:X2^2` unless both `X1` and `X2^2` are already in the model. Similarly, the function will not remove lower-order terms that are subsets of higher-order terms that remain in the model. For example, the function will not try to remove `X1` or `X2^2` if `X1:X2^2` remains in the model.

The default value of `'Criterion'` for a generalized linear model is `'Deviance'`. `stepwiseglm` and `step` of `GeneralizedLinearModel` follow a similar procedure for adding or removing terms.

You can specify other criteria by using the `'Criterion'` name-value pair argument. For example, you can specify the change in the value of the Akaike information criterion, Bayesian information criterion, R-squared, or adjusted R-squared as the criterion to add or remove terms.

Depending on the terms included in the initial model, and the order in which the function adds and removes terms, the function might build different models from the same set of potential terms. The function terminates when no single step improves the model. However, a different initial model or a different sequence of steps does not guarantee a better fit. In this sense, stepwise models are locally optimal, but might not be globally optimal.

- `step` treats a categorical predictor as follows:
  - A model with a categorical predictor that has *L* levels (categories) includes *L* – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `step` treats the group of *L* – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with *L* levels consist of the element-wise product of the *L* – 1 indicator variables with the continuous predictor.
  - Interaction terms between two categorical predictors with *L* and *M* levels consist of the (*L* – 1)*(*M* – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
  - You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

  Therefore, if `step` adds or removes a categorical predictor, the function actually adds or removes the group of indicator variables in one step. Similarly, if `step` adds or removes an interaction term with a categorical predictor, the function actually adds or removes the group of interaction terms including the categorical predictor.

- `step` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `tbl`, `X`, and `Y` to be missing values. `step` does not use observations with missing values in the fit. The `ObservationInfo` property of a fitted model indicates whether or not `step` uses each observation in the fit.

Alternative Functionality

- Use `stepwiseglm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `addTerms` or `removeTerms` to add or remove specific terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`GeneralizedLinearModel` | `addTerms` | `removeTerms` | `stepwiseglm`

Topics

“Plots to Understand Predictor Effects and How to Modify a Model” on page 12-21
“Generalized Linear Models” on page 12-9

Introduced in R2012a

step

Improve linear regression model by adding or removing terms

Syntax

```
NewMdl = step(mdl)
NewMdl = step(mdl,Name,Value)
```

Description

`NewMdl = step(mdl)` returns a linear regression model based on `mdl` using stepwise regression to add or remove one predictor.

`NewMdl = step(mdl,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the criterion to use to add or remove terms and the maximum number of steps to take.

Examples

Modify Linear Regression Model Using `step`

Fit a linear regression model and use `step` to improve the model by adding or removing terms. This example also describes how the `step` function treats a categorical predictor.

Load the `carsmall` data set, and create a table using the `Weight`, `Model_Year`, and `MPG` variables.

```
load carsmall
tbl1 = table(MPG,Weight);
tbl1.Year = categorical(Model_Year);
```

Create a linear regression model of `MPG` as a function of `Weight`.

```
mdl1 = fitlm(tbl1,'MPG ~ Weight')
```

```
mdl1 =
Linear regression model:
    MPG ~ 1 + Weight

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   __________    _________    _______    __________

    (Intercept)        49.238       1.6411     30.002    2.7015e-49
    Weight         -0.0086119    0.0005348    -16.103    1.6434e-28


Number of observations: 94, Error degrees of freedom: 92
Root Mean Squared Error: 4.13
R-squared: 0.738,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 259, p-value = 1.64e-28
```

Adjust the model to include up to `'quadratic'` terms by using `step`. Specify `'NSteps'` as 5 to allow at most 5 steps of stepwise regression. Specify `'Verbose'` as 2 to display the evaluation process and the decision taken at each step.

```
NewMdl1 = step(mdl1,'Upper','quadratic','NSteps',5,'Verbose',2)
```

```
   pValue for adding Year is 8.2284e-15
   pValue for adding Weight^2 is 0.15454
1. Adding Year, FStat = 47.5136, pValue = 8.22836e-15
   pValue for adding Weight:Year is 0.0071637
   pValue for adding Weight^2 is 0.0022303
2. Adding Weight^2, FStat = 9.9164, pValue = 0.0022303
   pValue for adding Weight:Year is 0.19519
   pValue for removing Year is 2.9042e-16

NewMdl1 =
Linear regression model:
    MPG ~ 1 + Weight + Year + Weight^2

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                   __________    __________    _______    __________

    (Intercept)        54.206        4.7117     11.505    2.6648e-19
    Weight          -0.016404     0.0031249    -5.2493    1.0283e-06
    Year_76            2.0887       0.71491     2.9215     0.0044137
    Year_82            8.1864       0.81531     10.041    2.6364e-16
    Weight^2       1.5573e-06    4.9454e-07      3.149     0.0022303


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.78
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 172, p-value = 5.52e-41
```

`step` creates two indicator variables, `Year_76` and `Year_82`, because `Year` includes three distinct values. `step` does not consider the square terms of indicator variables because the square of an indicator variable is itself.

Because `'Verbose'` is 2, `step` displays the evaluation process:

- `step` computes the *p*-values for adding `Year` or `Weight^2`. The *p*-value for `Year` is less than both the *p*-value for `Weight^2` and the default threshold value of 0.05; therefore, `step` adds `Year` to the model.
- `step` computes the *p*-values for adding `Weight:Year` or `Weight^2`. Because the *p*-value for `Weight^2` is less than the *p*-value for `Weight:Year`, the `step` function adds `Weight^2` to the model.
- After adding the quadratic term, `step` computes the *p*-value for adding `Weight:Year` again, but the *p*-value is greater than the threshold value. Therefore, `step` does not add the term to the model. `step` does not examine adding `Weight^3` because of the upper bound specified by the `'Upper'` name-value pair argument.
- `step` looks for terms to remove. `step` already examined `Weight^2`, so it computes only the *p*-value for removing `Year`. Because the *p*-value is less than the default threshold value of 0.10, `step` does not remove the term.
- Although the maximum allowed number of steps is 5, `step` terminates the process after two steps because the model does not improve by adding or removing a term.

`step` treats the two indicator variables as one predictor variable and adds `Year` in one step. To treat the two indicator variables as two distinct predictor variables, use `dummyvar` to create separate categorical variables.

```
temp_Year = dummyvar(tbl1.Year);
Year_76 = temp_Year(:,2);
Year_82 = temp_Year(:,3);
```

Create a table containing `MPG`, `Weight`, `Year_76`, and `Year_82`.

```
tbl2 = table(MPG,Weight,Year_76,Year_82);
```

Create a linear regression model of `MPG` as a function of `Weight`, and use `step` to improve the model.

```
mdl2 = fitlm(tbl2,'MPG ~ Weight');
NewMdl2 = step(mdl2,'Upper','quadratic','NSteps',5)
```

```
1. Adding Year_82, FStat = 83.1956, pValue = 1.76163e-14
2. Adding Weight:Year_82, FStat = 8.0641, pValue = 0.0055818
3. Adding Year_76, FStat = 8.1284, pValue = 0.0054157

NewMdl2 =
Linear regression model:
    MPG ~ 1 + Year_76 + Weight*Year_82

Estimated Coefficients:
                       Estimate         SE         tStat       pValue
                      __________    __________    _______    __________

    (Intercept)           38.844        1.5294     25.397     1.503e-42
    Weight             -0.006272    0.00042673    -14.698    1.5622e-25
    Year_76               2.0395       0.71537      2.851     0.0054157
    Year_82               19.607        3.8731     5.0623    2.2163e-06
    Weight:Year_82    -0.0046268     0.0014979    -3.0888     0.0026806


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.79
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 171, p-value = 6.54e-41
```

The model `NewMdl2` includes the interaction term `Weight:Year_82` instead of `Weight^2`, the term included in `NewMdl1`.

Input Arguments

`mdl` — Linear regression model

`LinearModel` object

Linear regression model, specified as a `LinearModel` object created using `fitlm` or `stepwiselm`.

You can use `step` only if you create `mdl` by using `fitlm` with the `'RobustOpts'` name-value pair argument set to the default `'off'`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Criterion','aic','Upper','quadratic','Verbose',2` instructs `step` to use the Akaike information criterion, include (at most) the quadratic terms in the model, and display the evaluation process and the decision taken at each step.

`Criterion` — Criterion to add or remove terms

`'sse'` (default) | `'aic'` | `'bic'` | `'rsquared'` | `'adjrsquared'`

Criterion to add or remove terms, specified as the comma-separated pair consisting of `'Criterion'` and one of these values:

- `'sse'` — *p*-value for an *F*-test of the change in the sum of squared error that results from adding or removing the term
- `'aic'` — Change in the value of Akaike information criterion (AIC)
- `'bic'` — Change in the value of Bayesian information criterion (BIC)
- `'rsquared'` — Increase in the value of R^2
- `'adjrsquared'` — Increase in the value of adjusted R^2

Example: `'Criterion','bic'`

`Lower` — Model specification describing terms that cannot be removed

`'constant'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'Y ~ terms'`

Model specification describing terms that cannot be removed from the model, specified as the comma-separated pair consisting of `'Lower'` and one of these values:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.

Value	Model Type
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and $x_1*x_2^2$ terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a "Terms Matrix" on page 33-5997, specifying terms in the model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar "Formula" on page 33-5997 in the form
  `'Y ~ terms'`,
  where the `terms` are in "Wilkinson Notation" on page 33-5998. The variable names in the formula must be valid MATLAB identifiers.

Example: `'Lower','linear'`

Data Types: `single` | `double` | `char` | `string`

`NSteps` — Maximum number of steps to take
1 (default) | positive integer

Maximum number of steps to take, specified as the comma-separated pair consisting of `'NSteps'` and a positive integer.

Example: `'NSteps',5`

Data Types: `single` | `double`

`PEnter` — Threshold for criterion to add term
scalar value

Threshold for the criterion to add a term, specified as the comma-separated pair consisting of `'PEnter'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'SSE'`	0.05	If the *p*-value of the *F*-statistic is less than `PEnter` (*p*-value to enter), add the term to the model.
`'AIC'`	0	If the change in the AIC of the model is less than `PEnter`, add the term to the model.
`'BIC'`	0	If the change in the BIC of the model is less than `PEnter`, add the term to the model.

Criterion	Default Value	Decision
'Rsquared'	0.1	If the increase in the R-squared value of the model is greater than PEnter, add the term to the model.
'AdjRsquared'	0	If the increase in the adjusted R-squared value of the model is greater than PEnter, add the term to the model.

For more information, see the Criterion name-value pair argument.

Example: 'PEnter',0.075

PRemove — Threshold for criterion to remove term
scalar value

Threshold for the criterion to remove a term, specified as the comma-separated pair consisting of 'PRemove' and a scalar value, as described in this table.

Criterion	Default Value	Decision
'SSE'	0.10	If the *p*-value of the *F*-statistic is greater than PRemove (*p*-value to remove), remove the term from the model.
'AIC'	0.01	If the change in the AIC of the model is greater than PRemove, remove the term from the model.
'BIC'	0.01	If the change in the BIC of the model is greater than PRemove, remove the term from the model.
'Rsquared'	0.05	If the increase in the R-squared value of the model is less than PRemove, remove the term from the model.
'AdjRsquared'	-0.05	If the increase in the adjusted R-squared value of the model is less than PRemove, remove the term from the model.

At each step, the step function also checks whether a term is redundant (linearly dependent) with other terms in the current model. When any term is linearly dependent with other terms in the current model, the step function removes the redundant term, regardless of the criterion value.

For more information, see the Criterion name-value pair argument.

Example: 'PRemove',0.05

Upper — Model specification describing largest set of terms in fit
`'interactions'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'Y ~ terms'`

Model specification describing the largest set of terms in the fit, specified as the comma-separated pair consisting of `'Upper'` and one of these values:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and $x_1*x_2^2$ terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-5997, specifying terms in the model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar “Formula” on page 33-5997 in the form
  `'Y ~ terms'`,
  where the `terms` are in “Wilkinson Notation” on page 33-5998. The variable names in the formula must be valid MATLAB identifiers.

Example: `'Upper','quadratic'`

Data Types: `single` | `double` | `char` | `string`

Verbose — Control for display of information
`1` (default) | `0` | `2`

Control for the display of information, specified as the comma-separated pair consisting of `'Verbose'` and one of these values:

- `0` — Suppress all display.

- `1` — Display the action taken at each step.
- `2` — Display the evaluation process and the action taken at each step.

Example: `'Verbose',2`

Output Arguments

`NewMdl` — Linear regression model

`LinearModel` object

Linear regression model, returned as a `LinearModel` object

To overwrite the input argument `mdl`, assign the new model to `mdl`.

```
mdl = step(mdl);
```

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'`*`y ~ terms`*`'`.

- *`y`* is the response name.
- *`terms`* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.

- `'y ~ x1 + x2 + x3 − 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include −1 in the formula.

A formula includes a constant term unless you explicitly remove the term using −1.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.
- − means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
−x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 − x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Algorithms

- Stepwise regression is a systematic method for adding and removing terms from a linear or generalized linear model based on their statistical significance in explaining the response variable. The method begins with an initial model, specified using `modelspec`, and then compares the explanatory power of incrementally larger and smaller models.

  The `step` function uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add to the model or remove from the model based on the value of the `'Criterion'` name-value pair argument.

  The default value of `'Criterion'` for a linear regression model is `'sse'`. In this case, `stepwiselm` and `step` of `LinearModel` use the *p*-value of an *F*-statistic to test models with and

without a potential term at each step. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the function adds the term to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the function removes the term from the model.

Stepwise regression takes these steps when `'Criterion'` is `'sse'`:

1. Fit the initial model.
2. Examine a set of available terms not in the model. If any of the terms have *p*-values less than an entrance tolerance (that is, if it is unlikely a term would have a zero coefficient if added to the model), add the term with the smallest *p*-value and repeat this step; otherwise, go to step 3.
3. If any of the available terms in the model have *p*-values greater than an exit tolerance (that is, the hypothesis of a zero coefficient cannot be rejected), remove the term with the largest *p*-value and return to step 2; otherwise, end the process.

At any stage, the function will not add a higher-order term if the model does not also include all lower-order terms that are subsets of the higher-order term. For example, the function will not try to add the term `X1:X2^2` unless both `X1` and `X2^2` are already in the model. Similarly, the function will not remove lower-order terms that are subsets of higher-order terms that remain in the model. For example, the function will not try to remove `X1` or `X2^2` if `X1:X2^2` remains in the model.

The default value of `'Criterion'` for a generalized linear model is `'Deviance'`. `stepwiseglm` and `step` of `GeneralizedLinearModel` follow a similar procedure for adding or removing terms.

You can specify other criteria by using the `'Criterion'` name-value pair argument. For example, you can specify the change in the value of the Akaike information criterion, Bayesian information criterion, R-squared, or adjusted R-squared as the criterion to add or remove terms.

Depending on the terms included in the initial model, and the order in which the function adds and removes terms, the function might build different models from the same set of potential terms. The function terminates when no single step improves the model. However, a different initial model or a different sequence of steps does not guarantee a better fit. In this sense, stepwise models are locally optimal, but might not be globally optimal.

- `step` treats a categorical predictor as follows:
  - A model with a categorical predictor that has *L* levels (categories) includes *L* – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `step` treats the group of *L* – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.

- Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
- Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
- You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Therefore, if `step` adds or removes a categorical predictor, the function actually adds or removes the group of indicator variables in one step. Similarly, if `step` adds or removes an interaction term with a categorical predictor, the function actually adds or removes the group of interaction terms including the categorical predictor.

Alternative Functionality

- Use `stepwiselm` to specify terms in a starting model and continue improving the model until no single step of adding or removing a term is beneficial.
- Use `addTerms` or `removeTerms` to add or remove specific terms.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- This function supports model objects fitted with GPU array input arguments.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`LinearModel` | `addTerms` | `removeTerms` | `stepwiselm`

Topics

“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression” on page 11-9
“Stepwise Regression” on page 11-99

Introduced in R2012a

stepwise

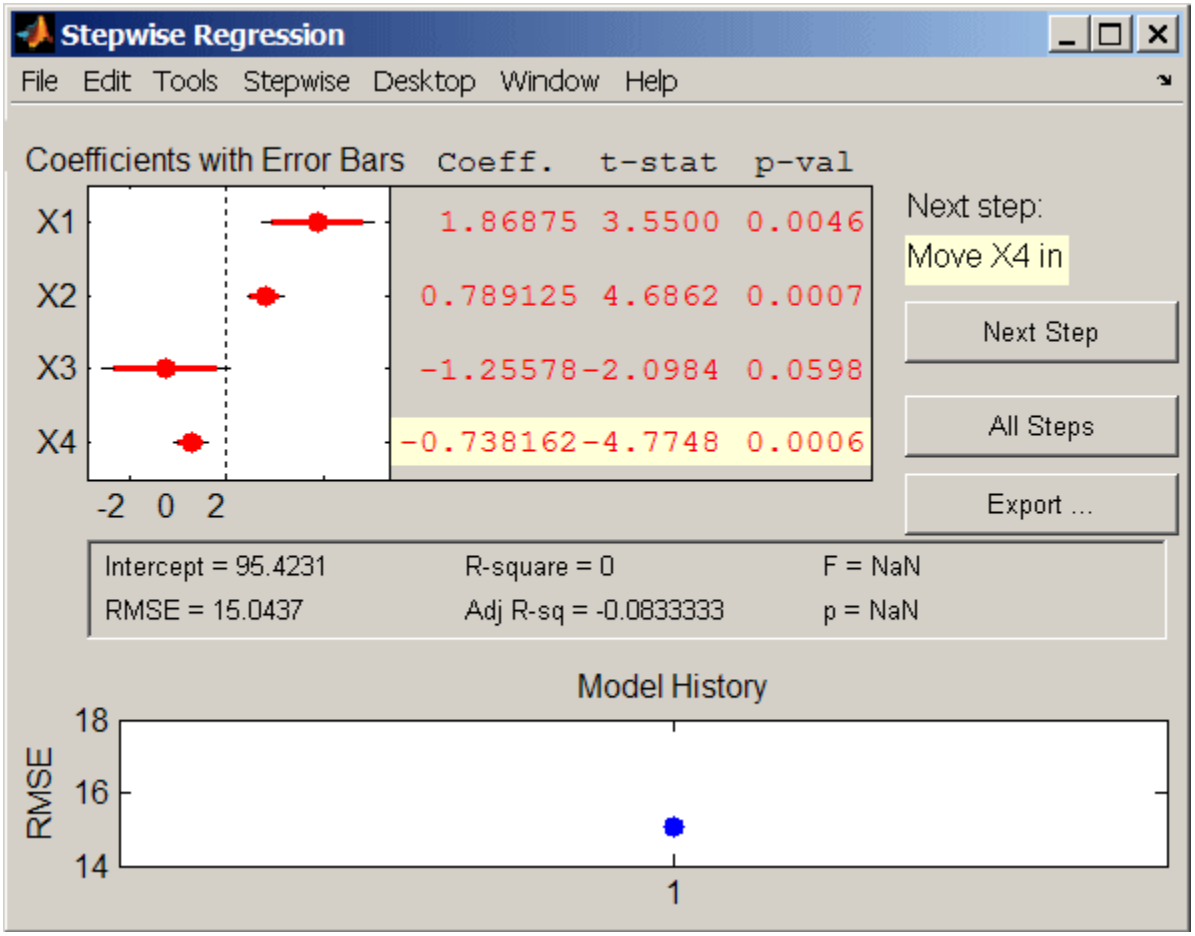
Interactive stepwise regression

Syntax

```
stepwise
stepwise(X,y)
stepwise(X,y,inmodel,penter,premove)
```

Description

`stepwise` uses the sample data in `hald.mat` to display a graphical user interface for performing stepwise regression of the response values in `heat` on the predictive terms in `ingredients`.



The upper left of the interface displays estimates of the coefficients for all potential terms, with horizontal bars indicating 90% (colored) and 95% (grey) confidence intervals. The red color indicates that, initially, the terms are not in the model. Values displayed in the table are those that would result if the terms were added to the model.

The middle portion of the interface displays summary statistics for the entire model. These statistics are updated with each step.

The lower portion of the interface, **Model History**, displays the RMSE for the model. The plot tracks the RMSE from step to step, so you can compare the optimality of different models. Hover over the blue dots in the history to see which terms were in the model at a particular step. Click on a blue dot in the history to open a copy of the interface initialized with the terms in the model at that step.

Initial models, as well as entrance/exit tolerances for the *p*-values of *F*-statistics, are specified using additional input arguments to `stepwise`. Defaults are an initial model with no terms, an entrance tolerance of 0.05, and an exit tolerance of 0.10.

To center and scale the input data (compute *z*-scores) to improve conditioning of the underlying least-squares problem, select `Scale Inputs` from the **Stepwise** menu.

You proceed through a stepwise regression in one of two ways:

1. Click **Next Step** to select the recommended next step. The recommended next step either adds the most significant term or removes the least significant term. When the regression reaches a local minimum of RMSE, the recommended next step is "Move no terms." You can perform all of the recommended steps at once by clicking **All Steps**.
2. Click a line in the plot or in the table to toggle the state of the corresponding term. Clicking a red line, corresponding to a term not currently in the model, adds the term to the model and changes the line to blue. Clicking a blue line, corresponding to a term currently in the model, removes the term from the model and changes the line to red.

To call `addedvarplot` and produce an added variable plot from the `stepwise` interface, select **Added Variable Plot** from the **Stepwise** menu. A list of terms is displayed. Select the term you want to add, and then click **OK**.

Click **Export** to display a dialog box that allows you to select information from the interface to save to the MATLAB workspace. Check the information you want to export and, optionally, change the names of the workspace variables to be created. Click **OK** to export the information.

`stepwise(X,y)` displays the interface using the *p* predictive terms in the *n*-by-*p* matrix `X` and the response values in the *n*-by-1 vector `y`. Distinct predictive terms should appear in different columns of `X`.

Note `stepwise` automatically includes a constant term in all models. Do not enter a column of 1s directly into `X`.

`stepwise` treats `NaN` values in either `X` or `y` as missing values, and ignores them.

`stepwise(X,y,inmodel,penter,premove)` additionally specifies the initial model (`inmodel`) and the entrance (`penter`) and exit (`premove`) tolerances for the *p*-values of *F*-statistics. `inmodel` is either a logical vector with length equal to the number of columns of `X`, or a vector of indices, with values ranging from 1 to the number of columns in `X`. The value of `penter` must be less than or equal to the value of `premove`.

Algorithms

Stepwise regression is a systematic method for adding and removing terms from a multilinear model based on their statistical significance in a regression. The method begins with an initial model and

then compares the explanatory power of incrementally larger and smaller models. At each step, the *p* value of an *F*-statistic is computed to test models with and without a potential term. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model. The method proceeds as follows:

1. Fit the initial model.
2. If any terms not in the model have *p*-values less than an entrance tolerance (that is, if it is unlikely that they would have zero coefficient if added to the model), add the one with the smallest *p* value and repeat this step; otherwise, go to step 3.
3. If any terms in the model have *p*-values greater than an exit tolerance (that is, if it is unlikely that the hypothesis of a zero coefficient can be rejected), remove the one with the largest *p* value and go to step 2; otherwise, end.

Depending on the terms included in the initial model and the order in which terms are moved in and out, the method may build different models from the same set of potential terms. The method terminates when no single step improves the model. There is no guarantee, however, that a different initial model or a different sequence of steps will not lead to a better fit. In this sense, stepwise models are locally optimal, but may not be globally optimal.

See Also

`addedvarplot` | `regress` | `stepwisefit`

Introduced before R2006a

stepwiseglm

Create generalized linear regression model by stepwise regression

Syntax

```
mdl = stepwiseglm(tbl)
mdl = stepwiseglm(X,y)
mdl = stepwiseglm(___,modelspec)
mdl = stepwiseglm(___,modelspec,Name,Value)
```

Description

`mdl = stepwiseglm(tbl)` creates a generalized linear model of a table or dataset array `tbl` using stepwise regression to add or remove predictors, starting from a constant model. `stepwiseglm` uses the last variable of `tbl` as the response variable. `stepwiseglm` uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add the model to or remove from the model, based on the value of the `'Criterion'` argument.

`mdl = stepwiseglm(X,y)` creates a generalized linear model of the responses `y` to a data matrix `X`.

`mdl = stepwiseglm(___,modelspec)` specifies the starting model `modelspec` using any of the input argument combinations in previous syntaxes.

`mdl = stepwiseglm(___,modelspec,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the categorical variables, the smallest or largest set of terms to use in the model, the maximum number of steps to take, or the criterion that `stepwiseglm` uses to add or remove terms.

Examples

Generalized Linear Model Using Stepwise Algorithm

Create response data using just three of 20 predictors, and create a generalized linear model using stepwise algorithm to see if it uses just the correct predictors.

Create data with 20 predictors, and Poisson response using just three of the predictors, plus a constant.

```
rng('default') % for reproducibility
X = randn(100,20);
mu = exp(X(:,[5 10 15])*[.4;.2;.3] + 1);
y = poissrnd(mu);
```

Fit a generalized linear model using the Poisson distribution.

```
mdl =  stepwiseglm(X,y,...
    'constant','upper','linear','Distribution','poisson')
```

```
1. Adding x5, Deviance = 134.439, Chi2Stat = 52.24814, PValue = 4.891229e-13
2. Adding x15, Deviance = 106.285, Chi2Stat = 28.15393, PValue = 1.1204e-07
3. Adding x10, Deviance = 95.0207, Chi2Stat = 11.2644, PValue = 0.000790094

mdl =
Generalized linear regression model:
    log(y) ~ 1 + x5 + x10 + x15
    Distribution = Poisson

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     1.0115     0.064275    15.737    8.4217e-56
    x5             0.39508     0.066665    5.9263    3.0977e-09
    x10            0.18863      0.05534    3.4085     0.0006532
    x15            0.29295     0.053269    5.4995    3.8089e-08


100 observations, 96 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 91.7, p-value = 9.61e-20
```

The starting model is the constant model. `stepwiseglm` by default uses deviance of the model as the criterion. It first adds `x5` into the model, as the *p*-value for the test statistic, deviance (the differences in the deviances of the two models), is less than the default threshold value 0.05. Then, it adds `x15` because given `x5` is in the model, when `x15` is added, the *p*-value for chi-squared test is smaller than 0.05. It then adds `x10` because given `x5` and `x15` are in the model, when `x10` is added, the *p*-value for the chi-square test statistic is again less than 0.05.

Input Arguments

`tbl` — Input data

table | dataset array

Input data including predictor and response variables, specified as a table or dataset array. The predictor variables and response variable can be numeric, logical, categorical, character, or string. The response variable can have a data type other than numeric only if `'Distribution'` is `'binomial'`.

- By default, `stepwiseglm` takes the last variable as the response variable and the others as the predictor variables.
- To set a different column as the response variable, use the `ResponseVar` name-value pair argument.
- To use a subset of the columns as predictors, use the `PredictorVars` name-value pair argument.
- To define a model specification, set the `modelspec` argument using a formula or terms matrix. The formula or terms matrix specifies which columns to use as the predictor or response variables.

The variable names in a table do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot specify `modelspec` using a formula.

- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiseglm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

You can verify the variable names in `tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

X — Predictor variables
matrix

Predictor variables, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

By default, there is a constant term in the model, unless you explicitly remove it, so do not include a column of 1s in `X`.

Data Types: `single` | `double`

y — Response variable
vector | matrix

Response variable, specified as a vector or matrix.

- If `'Distribution'` is not `'binomial'`, then `y` must be an *n*-by-1 vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`. The data type must be single or double.
- If `'Distribution'` is `'binomial'`, then `y` can be an *n*-by-1 vector or *n*-by-2 matrix with counts in column 1 and `BinomialSize` in column 2.

Data Types: `single` | `double` | `logical` | `categorical`

`modelspec` — Starting model
`'constant'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'y ~ terms'`

Starting model for `stepwiseglm`, specified as one of the following:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.

Value	Model Type
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and $x_1*x_2^2$ terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a "Terms Matrix" on page 33-6014, specifying terms in the model, where *t* is the number of terms and *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar "Formula" on page 33-6015 in the form
  `'y ~ terms'`,
  where the `terms` are in "Wilkinson Notation" on page 33-6015. The variable names in the formula must be variable names in `tbl` or variable names specified by `Varnames`. Also, the variable names must be valid MATLAB identifiers.

  The software determines the order of terms in a fitted model by using the order of terms in `tbl` or `X`. Therefore, the order of terms in the model can be different from the order of terms in the specified formula.

If you want to specify the smallest or largest set of terms in the model that `stepwiselm` fits, use the `Lower` and `Upper` name-value pair arguments.

Data Types: `char` | `string` | `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Criterion','aic','Distribution','poisson','Upper','interactions'` specifies Akaike Information Criterion as the criterion to add or remove variables to the model, Poisson distribution as the distribution of the response variable, and a model with all possible interactions as the largest model to consider as the fit.

`BinomialSize` — Number of trials for binomial distribution

1 (default) | numeric scalar | numeric vector | character vector | string scalar

Number of trials for binomial distribution, that is the sample size, specified as the comma-separated pair consisting of `'BinomialSize'` and the variable name in `tbl`, a numeric scalar, or a numeric vector of the same length as the response. This is the parameter `n` for the fitted binomial distribution. `BinomialSize` applies only when the `Distribution` parameter is `'binomial'`.

If `BinomialSize` is a scalar value, that means all observations have the same number of trials.

As an alternative to `BinomialSize`, you can specify the response as a two-column matrix with counts in column 1 and `BinomialSize` in column 2.

Data Types: `single` | `double` | `char` | `string`

`CategoricalVars` — Categorical variable list
string array | cell array of character vectors | logical or numeric index vector

Categorical variable list, specified as the comma-separated pair consisting of `'CategoricalVars'` and either a string array or cell array of character vectors containing categorical variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are categorical.

- If data is in a table or dataset array `tbl`, then, by default, `stepwiseglm` treats all categorical values, logical values, character arrays, string arrays, and cell arrays of character vectors as categorical variables.
- If data is in matrix `X`, then the default value of `'CategoricalVars'` is an empty matrix `[]`. That is, no variable is categorical unless you specify it as categorical.

For example, you can specify the second and third variables out of six as categorical using either of the following:

Example: `'CategoricalVars',[2,3]`

Example: `'CategoricalVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`Criterion` — Criterion to add or remove terms
`'Deviance'` (default) | `'sse'` | `'aic'` | `'bic'` | `'rsquared'` | `'adjrsquared'`

Criterion to add or remove terms, specified as the comma-separated pair consisting of `'Criterion'` and one of these values:

- `'Deviance'` — *p*-value for an *F*-test or chi-squared test of the change in the deviance that results from adding or removing the term. The *F*-test tests a single model, and the chi-squared test compares two different models.
- `'sse'` — *p*-value for an *F*-test of the change in the sum of squared error that results from adding or removing the term.
- `'aic'` — Change in the value of the Akaike information criterion (AIC).
- `'bic'` — Change in the value of the Bayesian information criterion (BIC).
- `'rsquared'` — Increase in the value of R^2.
- `'adjrsquared'` — Increase in the value of adjusted R^2.

Example: `'Criterion','bic'`

`DispersionFlag` — Indicator to compute dispersion parameter
`false` for `'binomial'` and `'poisson'` distributions (default) | `true`

Indicator to compute dispersion parameter for `'binomial'` and `'poisson'` distributions, specified as the comma-separated pair consisting of `'DispersionFlag'` and one of the following.

`true`	Estimate a dispersion parameter when computing standard errors. The estimated dispersion parameter value is the sum of squared Pearson residuals divided by the degrees of freedom for error (DFE).
`false`	Default. Use the theoretical value of 1 when computing standard errors.

The fitting function always estimates the dispersion for other distributions.

Example: `'DispersionFlag',true`

`Distribution` — Distribution of the response variable
`'normal'` (default) | `'binomial'` | `'poisson'` | `'gamma'` | `'inverse gaussian'`

Distribution of the response variable, specified as the comma-separated pair consisting of `'Distribution'` and one of the following.

`'normal'`	Normal distribution
`'binomial'`	Binomial distribution
`'poisson'`	Poisson distribution
`'gamma'`	Gamma distribution
`'inverse gaussian'`	Inverse Gaussian distribution

Example: `'Distribution','gamma'`

`Exclude` — Observations to exclude
logical or numeric index vector

Observations to exclude from the fit, specified as the comma-separated pair consisting of `'Exclude'` and a logical or numeric index vector indicating which observations to exclude from the fit.

For example, you can exclude observations 2 and 3 out of 6 using either of the following examples.

Example: `'Exclude',[2,3]`

Example: `'Exclude',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical`

`Intercept` — Indicator for constant term
`true` (default) | `false`

Indicator for the constant term (intercept) in the fit, specified as the comma-separated pair consisting of `'Intercept'` and either `true` to include or `false` to remove the constant term from the model.

Use `'Intercept'` only when specifying the model using a character vector or string scalar, not a formula or matrix.

Example: `'Intercept',false`

`Link` — Link function
canonical link function (default) | scalar value | structure

Link function to use in place of the canonical link function, specified as the comma-separated pair consisting of `'Link'` and one of the following.

Link Function Name	Link Function	Mean (Inverse) Function
`'identity'`	$f(\mu) = \mu$	$\mu = Xb$
`'log'`	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
`'logit'`	$f(\mu) = \log(\mu/(1-\mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'probit'`	$f(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function of the standard normal distribution.	$\mu = \Phi(Xb)$
`'comploglog'`	$f(\mu) = \log(-\log(1 - \mu))$	$\mu = 1 - \exp(-\exp(Xb))$
`'reciprocal'`	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
`p` (a number)	$f(\mu) = \mu^p$	$\mu = Xb^{1/p}$
`S` (a structure) with three fields. Each field holds a function handle that accepts a vector of inputs and returns a vector of the same size: • `S.Link` — The link function • `S.Inverse` — The inverse link function • `S.Derivative` — The derivative of the link function	$f(\mu) =$ `S.Link`(μ)	$\mu =$ `S.Inverse`(Xb)

The link function defines the relationship $f(\mu) = X*b$ between the mean response μ and the linear combination of predictors $X*b$.

For more information on the canonical link functions, see “Canonical Function” on page 33-6016.

Example: `'Link','probit'`

Data Types: `char` | `string` | `single` | `double` | `struct`

`Lower` — Model specification describing terms that cannot be removed from model
`'constant'` (default) | character vector | string scalar | terms matrix

Model specification describing terms that cannot be removed from the model, specified as the comma-separated pair consisting of `'Lower'` and one of the options for `modelspec` naming the model.

Example: `'Lower','linear'`

`NSteps` — Maximum number of steps to take
no limit (default) | positive integer

Maximum number of steps to take, specified as the comma-separated pair consisting of `'NSteps'` and a positive integer.

Example: `'NSteps',5`

Data Types: `single` | `double`

`Offset` — Offset variable

[ ] (default) | numeric vector | character vector | string scalar

Offset variable in the fit, specified as the comma-separated pair consisting of `'Offset'` and the variable name in `tbl` or a numeric vector with the same length as the response.

`stepwiseglm` uses `Offset` as an additional predictor with a coefficient value fixed at 1. In other words, the formula for fitting is

$f(\mu) = \texttt{Offset} + X*b$,

where f is the link function, μ is the mean response, and $X*b$ is the linear combination of predictors X. The `Offset` predictor has coefficient `1`.

For example, consider a Poisson regression model. Suppose the number of counts is known for theoretical reasons to be proportional to a predictor `A`. By using the log link function and by specifying `log(A)` as an offset, you can force the model to satisfy this theoretical constraint.

Data Types: `single` | `double` | `char` | `string`

`PEnter` — Threshold for criterion to add term

scalar value

Threshold for the criterion to add a term, specified as the comma-separated pair consisting of `'PEnter'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'Deviance'`	0.05	If the *p*-value of the *F*-statistic or chi-squared statistic is less than `PEnter` (*p*-value to enter), add the term to the model.
`'SSE'`	0.05	If the *p*-value of the *F*-statistic is less than `PEnter`, add the term to the model.
`'AIC'`	0	If the change in the AIC of the model is less than `PEnter`, add the term to the model.
`'BIC'`	0	If the change in the BIC of the model is less than `PEnter`, add the term to the model.
`'Rsquared'`	0.1	If the increase in the R-squared value of the model is greater than `PEnter`, add the term to the model.
`'AdjRsquared'`	0	If the increase in the adjusted R-squared value of the model is greater than `PEnter`, add the term to the model.

For more information, see the `Criterion` name-value pair argument.

Example: `'PEnter',0.075`

`PredictorVars` — Predictor variables
string array | cell array of character vectors | logical or numeric index vector

Predictor variables to use in the fit, specified as the comma-separated pair consisting of `'PredictorVars'` and either a string array or cell array of character vectors of the variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are predictor variables.

The string values or character vectors should be among the names in `tbl`, or the names you specify using the `'VarNames'` name-value pair argument.

The default is all variables in `X`, or all variables in `tbl` except for `ResponseVar`.

For example, you can specify the second and third variables as the predictor variables using either of the following examples.

Example: `'PredictorVars',[2,3]`

Example: `'PredictorVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`PRemove` — Threshold for criterion to remove term
scalar value

Threshold for the criterion to remove a term, specified as the comma-separated pair consisting of `'PRemove'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'Deviance'`	0.10	If the *p*-value of the *F*-statistic or chi-squared statistic is greater than `PRemove` (*p*-value to remove), remove the term from the model.
`'SSE'`	0.10	If the *p*-value of the *F*-statistic is greater than `PRemove`, remove the term from the model.
`'AIC'`	0.01	If the change in the AIC of the model is greater than `PRemove`, remove the term from the model.
`'BIC'`	0.01	If the change in the BIC of the model is greater than `PRemove`, remove the term from the model.
`'Rsquared'`	0.05	If the increase in the R-squared value of the model is less than `PRemove`, remove the term from the model.

Criterion	Default Value	Decision
'AdjRsquared'	-0.05	If the increase in the adjusted R-squared value of the model is less than PRemove, remove the term from the model.

At each step, the `stepwiseglm` function also checks whether a term is redundant (linearly dependent) with other terms in the current model. When a term is linearly dependent on other terms in the current model, the `stepwiseglm` function removes the redundant term, regardless of the criterion value.

For more information, see the `Criterion` name-value pair argument.

Example: `'PRemove',0.05`

`ResponseVar` — Response variable
last column in `tbl` (default) | character vector or string scalar containing variable name | logical or numeric index vector

Response variable to use in the fit, specified as the comma-separated pair consisting of `'ResponseVar'` and either a character vector or string scalar containing the variable name in the table or dataset array `tbl`, or a logical or numeric index vector indicating which column is the response variable. You typically need to use `'ResponseVar'` when fitting a table or dataset array `tbl`.

For example, you can specify the fourth variable, say `yield`, as the response out of six variables, in one of the following ways.

Example: `'ResponseVar','yield'`

Example: `'ResponseVar',[4]`

Example: `'ResponseVar',logical([0 0 0 1 0 0])`

Data Types: `single` | `double` | `logical` | `char` | `string`

`Upper` — Model specification describing largest set of terms in fit
`'interactions'` (default) | character vector | string scalar | terms matrix

Model specification describing the largest set of terms in the fit, specified as the comma-separated pair consisting of `'Upper'` and one of the options for `modelspec` naming the model.

Example: `'Upper','quadratic'`

`VarNames` — Names of variables
`{'x1','x2',...,'xn','y'}` (default) | string array | cell array of character vectors

Names of variables, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors including the names for the columns of `X` first, and the name for the response variable `y` last.

`'VarNames'` is not applicable to variables in a table or dataset array, because those variables already have names.

The variable names do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiseglm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

Before specifying `'VarNames',varNames`, you can verify the variable names in `varNames` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Example: `'VarNames',{'Horsepower','Acceleration','Model_Year','MPG'}`

Data Types: `string` | `cell`

`Verbose` — Control for display of information
`1` (default) | `0` | `2`

Control for the display of information, specified as the comma-separated pair consisting of `'Verbose'` and one of these values:

- `0` — Suppress all display.
- `1` — Display the action taken at each step.
- `2` — Display the evaluation process and the action taken at each step.

Example: `'Verbose',2`

`Weights` — Observation weights
`ones(n,1)` (default) | *n*-by-1 vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations.

Data Types: `single` | `double`

Output Arguments

`mdl` — Generalized linear regression model
`GeneralizedLinearModel` object

Generalized linear regression model, specified as a `GeneralizedLinearModel` object created using `fitglm` or `stepwiseglm`.

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'y ~ terms'`.

- *y* is the response name.
- *terms* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.
- `'y ~ x1 + x2 + x3 − 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include `−1` in the formula.

A formula includes a constant term unless you explicitly remove the term using `−1`.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- `+` means include the next variable.
- `−` means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- `*` defines an interaction and all lower-order terms.
- `^` raises the predictor to a power, exactly as in `*` repeated, so `^` includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
`1`	Constant (intercept) term
`x1^k`, where `k` is a positive integer	$x1$, $x1^2$, ..., $x1^k$

Wilkinson Notation	Terms in Standard Notation
`x1 + x2`	x1, x2
`x1*x2`	x1, x2, x1*x2
`x1:x2`	x1*x2 only
`-x2`	Do not include x2
`x1*x2 + x3`	x1, x2, x3, x1*x2
`x1 + x2 + x3 + x1:x2`	x1, x2, x3, x1*x2
`x1*x2*x3 - x1:x2:x3`	x1, x2, x3, x1*x2, x1*x3, x2*x3
`x1*(x2 + x3)`	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Canonical Function

The default link function for a generalized linear model is the canonical link function.

Distribution	Canonical Link Function Name	Link Function	Mean (Inverse) Function
`'normal'`	`'identity'`	$f(\mu) = \mu$	$\mu = Xb$
`'binomial'`	`'logit'`	$f(\mu) = \log(\mu/(1 - \mu))$	$\mu = \exp(Xb) / (1 + \exp(Xb))$
`'poisson'`	`'log'`	$f(\mu) = \log(\mu)$	$\mu = \exp(Xb)$
`'gamma'`	`-1`	$f(\mu) = 1/\mu$	$\mu = 1/(Xb)$
`'inverse gaussian'`	`-2`	$f(\mu) = 1/\mu^2$	$\mu = (Xb)^{-1/2}$

Tips

- The generalized linear model `mdl` is a standard linear model unless you specify otherwise with the `Distribution` name-value pair.
- For other methods such as `devianceTest`, or properties of the `GeneralizedLinearModel` object, see `GeneralizedLinearModel`.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- Stepwise regression is a systematic method for adding and removing terms from a linear or generalized linear model based on their statistical significance in explaining the response variable. The method begins with an initial model, specified using `modelspec`, and then compares the explanatory power of incrementally larger and smaller models.

  The `stepwiseglm` function uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add to the model or remove from the model based on the value of the `'Criterion'` name-value pair argument.

The default value of `'Criterion'` for a linear regression model is `'sse'`. In this case, `stepwiselm` and `step` of `LinearModel` use the *p*-value of an *F*-statistic to test models with and without a potential term at each step. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the function adds the term to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the function removes the term from the model.

Stepwise regression takes these steps when `'Criterion'` is `'sse'`:

1. Fit the initial model.
2. Examine a set of available terms not in the model. If any of the terms have *p*-values less than an entrance tolerance (that is, if it is unlikely a term would have a zero coefficient if added to the model), add the term with the smallest *p*-value and repeat this step; otherwise, go to step 3.
3. If any of the available terms in the model have *p*-values greater than an exit tolerance (that is, the hypothesis of a zero coefficient cannot be rejected), remove the term with the largest *p*-value and return to step 2; otherwise, end the process.

At any stage, the function will not add a higher-order term if the model does not also include all lower-order terms that are subsets of the higher-order term. For example, the function will not try to add the term `X1:X2^2` unless both `X1` and `X2^2` are already in the model. Similarly, the function will not remove lower-order terms that are subsets of higher-order terms that remain in the model. For example, the function will not try to remove `X1` or `X2^2` if `X1:X2^2` remains in the model.

The default value of `'Criterion'` for a generalized linear model is `'Deviance'`. `stepwiseglm` and `step` of `GeneralizedLinearModel` follow a similar procedure for adding or removing terms.

You can specify other criteria by using the `'Criterion'` name-value pair argument. For example, you can specify the change in the value of the Akaike information criterion, Bayesian information criterion, R-squared, or adjusted R-squared as the criterion to add or remove terms.

Depending on the terms included in the initial model, and the order in which the function adds and removes terms, the function might build different models from the same set of potential terms. The function terminates when no single step improves the model. However, a different initial model or a different sequence of steps does not guarantee a better fit. In this sense, stepwise models are locally optimal, but might not be globally optimal.

- `stepwiseglm` treats a categorical predictor as follows:
  - A model with a categorical predictor that has *L* levels (categories) includes *L* – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `stepwiseglm` treats the group of *L* – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.

- Interaction terms between a continuous predictor and a categorical predictor with L levels consist of the element-wise product of the L – 1 indicator variables with the continuous predictor.
- Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
- You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Therefore, if `stepwiseglm` adds or removes a categorical predictor, the function actually adds or removes the group of indicator variables in one step. Similarly, if `stepwiseglm` adds or removes an interaction term with a categorical predictor, the function actually adds or removes the group of interaction terms including the categorical predictor.

- `stepwiseglm` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `tbl`, `X`, and `Y` to be missing values. `stepwiseglm` does not use observations with missing values in the fit. The `ObservationInfo` property of a fitted model indicates whether or not `stepwiseglm` uses each observation in the fit.

Alternatives

- Use `fitglm` to create a model with a fixed specification. Use `step`, `addTerms`, or `removeTerms` to adjust a fitted model.

References

[1] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

[2] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[3] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

See Also

`GeneralizedLinearModel` | `fitglm` | `predict`

Topics

“Compare large and small stepwise models” on page 11-99
“Generalized Linear Models” on page 12-9
“Sequential Feature Selection” on page 15-61

Introduced in R2013b

stepwiselm

Perform stepwise regression

Syntax

```
mdl = stepwiselm(tbl)
mdl = stepwiselm(X,y)
mdl = stepwiselm(___,modelspec)
mdl = stepwiselm(___,Name,Value)
```

Description

`mdl = stepwiselm(tbl)` creates a linear model for the variables in the table or dataset array `tbl` using stepwise regression to add or remove predictors, starting from a constant model. `stepwiselm` uses the last variable of `tbl` as the response variable. `stepwiselm` uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add the model to or remove from the model, based on the value of the `'Criterion'` argument.

`mdl = stepwiselm(X,y)` creates a linear model of the responses `y` to the predictor variables in the data matrix `X`.

`mdl = stepwiselm(___,modelspec)` specifies the starting model `modelspec` using any of the input argument combinations in previous syntaxes.

`mdl = stepwiselm(___,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify the categorical variables, the smallest or largest set of terms to use in the model, the maximum number of steps to take, or the criterion that `stepwiselm` uses to add or remove terms.

Examples

Fit Linear Model Using Stepwise Regression

Load the `hald` data set, which measures the effect of cement composition on its hardening heat.

```
load hald
```

This data set includes the variables `ingredients` and `heat`. The matrix `ingredients` contains the percent composition of four chemicals present in the cement. The vector `heat` contains the values for the heat hardening after 180 days for each cement sample.

Fit a stepwise linear regression model to the data. Specify 0.06 as the threshold for the criterion to add a term to the model.

```
 mdl = stepwiselm(ingredients,heat,'PEnter',0.06)
```

```
1. Adding x4, FStat = 22.7985, pValue = 0.000576232
2. Adding x1, FStat = 108.2239, pValue = 1.105281e-06
3. Adding x2, FStat = 5.0259, pValue = 0.051687
4. Removing x4, FStat = 1.8633, pValue = 0.2054
```

```
mdl =
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     52.577       2.2862    22.998    5.4566e-10
    x1              1.4683       0.1213    12.105    2.6922e-07
    x2             0.66225     0.045855    14.442     5.029e-08


Number of observations: 13, Error degrees of freedom: 10
Root Mean Squared Error: 2.41
R-squared: 0.979,  Adjusted R-Squared: 0.974
F-statistic vs. constant model: 230, p-value = 4.41e-09
```

By default, the starting model is a constant model. `stepwiselm` performs forward selection and adds the `x4`, `x1`, and `x2` terms (in that order), because the corresponding *p*-values are less than the `PEnter` value of 0.06. `stepwiselm` then uses backward elimination and removes `x4` from the model because, once `x2` is in the model, the *p*-value of `x4` is greater than the default value of `PRemove`, 0.1.

Stepwise Regression Using Specified Model Formula and Variables

Perform stepwise regression using variables stored in a dataset array. Specify the starting model using Wilkinson notation, and identify the response and predictor variables using optional arguments.

Load the sample data.

```
load hospital
```

The hospital dataset array includes the gender, age, weight, and smoking status of patients.

Fit a linear model with a starting model of a constant term and `Smoker` as the predictor variable. Specify the response variable, `Weight`, and categorical predictor variables, `Sex`, `Age`, and `Smoker`.

```
mdl = stepwiselm(hospital,'Weight~1+Smoker',...
'ResponseVar','Weight','PredictorVars',{'Sex','Age','Smoker'},...
'CategoricalVar',{'Sex','Smoker'})
```

```
1. Adding Sex, FStat = 770.0158, pValue = 6.262758e-48
2. Removing Smoker, FStat = 0.21224, pValue = 0.64605

mdl =
Linear regression model:
    Weight ~ 1 + Sex

Estimated Coefficients:
                   Estimate      SE      tStat       pValue
                   ________    ______    ______    ___________

    (Intercept)     130.47     1.1995    108.77    5.2762e-104
    Sex_Male         50.06     1.7496    28.612     2.2464e-49
```

```
Number of observations: 100, Error degrees of freedom: 98
Root Mean Squared Error: 8.73
R-squared: 0.893,  Adjusted R-Squared: 0.892
F-statistic vs. constant model: 819, p-value = 2.25e-49
```

At each step, `stepwiselm` searches for terms to add and remove. At first step, stepwise algorithm adds `Sex` to the model with a *p*-value of 6.26e-48. Then, removes Smoker from the model, since given `Sex` in the model, the variable `Smoker` becomes redundant. `stepwiselm` only includes `Sex` in the final linear model. The weight of the patients do not seem to differ significantly according to age or the status of smoking.

Stepwise Regression Using Terms Matrix

Load a sample data set and define the matrix of predictors.

```
load carsmall
X = [Acceleration,Weight];
```

Define the starting model and the upper model using terms matrices.

```
T_starting = [0 0 0] % a constant model
```

```
T_starting = 1×3

     0     0     0
```

```
T_upper = [0 0 0;1 0 0;0 1 0;1 1 0] % a linear model with interactions
```

```
T_upper = 4×3

     0     0     0
     1     0     0
     0     1     0
     1     1     0
```

Create a linear regression model using stepwise regression. Specify the starting model and the upper bound of the model using the terms matrices, and specify `'Verbose'` as 2 to display the evaluation process and the decision taken at each step.

```
mdl = stepwiselm(X,MPG,T_starting,'upper',T_upper,'Verbose',2)
```

```
   pValue for adding x1 is 4.0973e-06
   pValue for adding x2 is 1.6434e-28
1. Adding x2, FStat = 259.3087, pValue = 1.643351e-28
   pValue for adding x1 is 0.18493
   No candidate terms to remove

mdl =
Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
```

```
                    ___________    _________    _______    __________

    (Intercept)         49.238        1.6411     30.002    2.7015e-49
    x2              -0.0086119     0.0005348    -16.103    1.6434e-28


Number of observations: 94, Error degrees of freedom: 92
Root Mean Squared Error: 4.13
R-squared: 0.738,  Adjusted R-Squared: 0.735
F-statistic vs. constant model: 259, p-value = 1.64e-28
```

Stepwise Regression with Categorical Predictor

Fit a linear regression model with a categorical predictor using stepwise regression. `stepwiselm` adds or removes a group of indicator variables in one step to add or removes a categorical predictor. This example also shows how to create indicator variables manually and pass them to `stepwiselm` so that `stepwiselm` treats each indicator variable as a separate predictor.

Load the `carsmall` data set, and create a table using the `Weight`, `Model_Year`, and `MPG` variables.

```
load carsmall
Year = categorical(Model_Year);
tbl1 = table(MPG,Weight,Year);
```

Fit a linear regression model of `MPG` using stepwise regression. Specify the starting model as a function of `Weight`. Set the upper bound of the model to `'poly21'`, meaning the model can include (at most) a constant and the terms `Weight`, `Weight^2`, `Year`, and `Weight*Year`. Specify `'Verbose'` as 2 to display the evaluation process and the decision taken at each step.

```
mdl1 = stepwiselm(tbl1,'MPG ~ Weight','Upper','poly21','Verbose',2)
```

```
   pValue for adding Year is 8.2284e-15
   pValue for adding Weight^2 is 0.15454
1. Adding Year, FStat = 47.5136, pValue = 8.22836e-15
   pValue for adding Weight^2 is 0.0022303
   pValue for adding Weight:Year is 0.0071637
2. Adding Weight^2, FStat = 9.9164, pValue = 0.0022303
   pValue for adding Weight:Year is 0.19519
   pValue for removing Year is 2.9042e-16

mdl1 =
Linear regression model:
    MPG ~ 1 + Weight + Year + Weight^2

Estimated Coefficients:
                    Estimate         SE         tStat       pValue
                    ___________    __________    _______    __________

    (Intercept)         54.206        4.7117     11.505    2.6648e-19
    Weight           -0.016404     0.0031249    -5.2493    1.0283e-06
    Year_76             2.0887       0.71491     2.9215     0.0044137
    Year_82             8.1864       0.81531     10.041    2.6364e-16
    Weight^2        1.5573e-06    4.9454e-07      3.149     0.0022303
```

```
Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.78
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 172, p-value = 5.52e-41
```

`stepwiselm` creates two indicator variables, `Year_76` and `Year_82`, because `Year` includes three distinct values.

Because `'Verbose'` is 2, `stepwiselm` displays the evaluation process:

- `stepwiselm` creates a model as a function of `Weight`.
- `stepwiselm` computes the *p*-values for adding `Year` or `Weight^2`. The *p*-value for `Year` is less than both the *p*-value for `Weight^2` and the default threshold value of 0.05; therefore, `stepwiselm` adds `Year` to the model.
- `stepwiselm` computes the *p*-values for adding `Weight:Year` or `Weight^2`. Because the *p*-value for `Weight^2` is less than the *p*-value for `Weight:Year`, the `stepwiselm` function adds `Weight^2` to the model.
- After adding the quadratic term, `stepwiselm` computes the *p*-value for adding `Weight:Year` again, but the *p*-value is greater than the threshold value. Therefore, `stepwiselm` does not add the term to the model. `stepwiselm` does not examine adding `Weight^3` because of the upper bound specified by the `'Upper'` name-value pair argument.
- `stepwiselm` looks for terms to remove. `stepwiselm` already examined `Weight^2`, so it computes only the *p*-value for removing `Year`. Because the *p*-value is less than the default threshold value of 0.10, `stepwiselm` does not remove the term.
- Although the maximum allowed number of steps is 5, `stepwiselm` terminates the process after two steps because the model does not improve by adding or removing a term.

`stepwiselm` treats the two indicator variables as one predictor variable and adds `Year` in one step. To treat the two indicator variables as two distinct predictor variables, use `dummyvar` to create separate categorical variables.

```
temp_Year = dummyvar(Year);
Year_76 = logical(temp_Year(:,2));
Year_82 = logical(temp_Year(:,3));
```

Create a table containing `MPG`, `Weight`, `Year_76`, and `Year_82`.

```
tbl2 = table(MPG,Weight,Year_76,Year_82);
```

Create a stepwise linear regression model from the same starting model used for `mdl1`.

```
mdl2 = stepwiselm(tbl2,'MPG ~ Weight','Upper','poly211')
```

```
1. Adding Year_82, FStat = 83.1956, pValue = 1.76163e-14
2. Adding Weight:Year_82, FStat = 8.0641, pValue = 0.0055818
3. Adding Year_76, FStat = 8.1284, pValue = 0.0054157

mdl2 =
Linear regression model:
    MPG ~ 1 + Year_76 + Weight*Year_82

Estimated Coefficients:
                        Estimate         SE         tStat       pValue
                       __________    __________    _______    __________
```

```
    (Intercept)             38.844        1.5294     25.397     1.503e-42
    Weight                -0.006272    0.00042673    -14.698    1.5622e-25
    Year_76_1                2.0395       0.71537      2.851     0.0054157
    Year_82_1                19.607        3.8731     5.0623    2.2163e-06
    Weight:Year_82_1    -0.0046268     0.0014979    -3.0888     0.0026806


Number of observations: 94, Error degrees of freedom: 89
Root Mean Squared Error: 2.79
R-squared: 0.885,  Adjusted R-Squared: 0.88
F-statistic vs. constant model: 171, p-value = 6.54e-41
```

The model `mdl2` includes the interaction term `Weight:Year_82_1` instead of `Weight^2`, the term included in `mdl1`.

Input Arguments

`tbl` — Input data

table | dataset array

Input data including predictor and response variables, specified as a table or dataset array. The predictor variables can be numeric, logical, categorical, character, or string. The response variable must be numeric or logical.

- By default, `stepwiselm` takes the last variable as the response variable and the others as the predictor variables.
- To set a different column as the response variable, use the `ResponseVar` name-value pair argument.
- To use a subset of the columns as predictors, use the `PredictorVars` name-value pair argument.
- To define a model specification, set the `modelspec` argument using a formula or terms matrix. The formula or terms matrix specifies which columns to use as the predictor or response variables.

The variable names in a table do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot specify `modelspec` using a formula.
- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiselm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

You can verify the variable names in `tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

`X` — Predictor variables

matrix

Predictor variables, specified as an *n*-by-*p* matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

By default, there is a constant term in the model, unless you explicitly remove it, so do not include a column of 1s in X.

Data Types: `single` | `double`

y — Response variable

vector

Response variable, specified as an *n*-by-1 vector, where *n* is the number of observations. Each entry in y is the response for the corresponding row of X.

Data Types: `single` | `double` | `logical`

modelspec — Starting model

`'constant'` (default) | character vector or string scalar naming the model | *t*-by-(*p* + 1) terms matrix | character vector or string scalar formula in the form `'y ~ terms'`

Starting model for the stepwise regression, specified as one of the following:

- A character vector or string scalar naming the model.

Value	Model Type
`'constant'`	Model contains only a constant (intercept) term.
`'linear'`	Model contains an intercept and linear term for each predictor.
`'interactions'`	Model contains an intercept, linear term for each predictor, and all products of pairs of distinct predictors (no squared terms).
`'purequadratic'`	Model contains an intercept term and linear and squared terms for each predictor.
`'quadratic'`	Model contains an intercept term, linear and squared terms for each predictor, and all products of pairs of distinct predictors.
`'poly`*`ijk`*`'`	Model is a polynomial with all terms up to degree *`i`* in the first predictor, degree *`j`* in the second predictor, and so on. Specify the maximum degree for each predictor by using numerals 0 though 9. The model contains interaction terms, but the degree of each interaction term does not exceed the maximum value of the specified degrees. For example, `'poly13'` has an intercept and x_1, x_2, x_2^2, x_2^3, x_1*x_2, and x_1*x_2^2 terms, where x_1 and x_2 are the first and second predictors, respectively.

- A *t*-by-(*p* + 1) matrix, or a “Terms Matrix” on page 33-6031, specifying terms in the model, where *t* is the number of terms and *p* is the number of predictor variables, and +1 accounts for the response variable. A terms matrix is convenient when the number of predictors is large and you want to generate the terms programmatically.
- A character vector or string scalar “Formula” on page 33-6031 in the form
  `'y ~ terms'`,
  where the `terms` are in “Wilkinson Notation” on page 33-6031. The variable names in the formula must be variable names in `tbl` or variable names specified by `Varnames`. Also, the variable names must be valid MATLAB identifiers.

The software determines the order of terms in a fitted model by using the order of terms in `tbl` or `X`. Therefore, the order of terms in the model can be different from the order of terms in the specified formula.

If you want to specify the smallest or largest set of terms in the model that `stepwiselm` fits, use the `Lower` and `Upper` name-value pair arguments.

Data Types: `char` | `string` | `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Criterion','aic','Upper','interactions','Verbose',1` instructs `stepwiselm` to use the Akaike information criterion, display the action it takes at each step, and include at most the interaction terms in the model.

`CategoricalVars` — Categorical variable list
string array | cell array of character vectors | logical or numeric index vector

Categorical variable list, specified as the comma-separated pair consisting of `'CategoricalVars'` and either a string array or cell array of character vectors containing categorical variable names in the table or dataset array `tbl`, or a logical or numeric index vector indicating which columns are categorical.

- If data is in a table or dataset array `tbl`, then, by default, `stepwiselm` treats all categorical values, logical values, character arrays, string arrays, and cell arrays of character vectors as categorical variables.
- If data is in matrix `X`, then the default value of `'CategoricalVars'` is an empty matrix `[]`. That is, no variable is categorical unless you specify it as categorical.

For example, you can specify the second and third variables out of six as categorical using either of the following:

Example: `'CategoricalVars',[2,3]`

Example: `'CategoricalVars',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical` | `string` | `cell`

`Criterion` — Criterion to add or remove terms
`'sse'` (default) | `'aic'` | `'bic'` | `'rsquared'` | `'adjrsquared'`

Criterion to add or remove terms, specified as the comma-separated pair consisting of `'Criterion'` and one of these values:

- `'sse'` — *p*-value for an *F*-test of the change in the sum of squared error that results from adding or removing the term
- `'aic'` — Change in the value of Akaike information criterion (AIC)
- `'bic'` — Change in the value of Bayesian information criterion (BIC)
- `'rsquared'` — Increase in the value of R^2
- `'adjrsquared'` — Increase in the value of adjusted R^2

Example: `'Criterion','bic'`

`Exclude` — Observations to exclude
logical or numeric index vector

Observations to exclude from the fit, specified as the comma-separated pair consisting of `'Exclude'` and a logical or numeric index vector indicating which observations to exclude from the fit.

For example, you can exclude observations 2 and 3 out of 6 using either of the following examples.

Example: `'Exclude',[2,3]`

Example: `'Exclude',logical([0 1 1 0 0 0])`

Data Types: `single` | `double` | `logical`

`Intercept` — Indicator for constant term
`true` (default) | `false`

Indicator for the constant term (intercept) in the fit, specified as the comma-separated pair consisting of `'Intercept'` and either `true` to include or `false` to remove the constant term from the model.

Use `'Intercept'` only when specifying the model using a character vector or string scalar, not a formula or matrix.

Example: `'Intercept',false`

`Lower` — Model specification describing terms that cannot be removed from model
`'constant'` (default) | character vector | string scalar | terms matrix

Model specification describing terms that cannot be removed from the model, specified as the comma-separated pair consisting of `'Lower'` and one of the options for `modelspec` naming the model.

Example: `'Lower','linear'`

`NSteps` — Maximum number of steps to take
no limit (default) | positive integer

Maximum number of steps to take, specified as the comma-separated pair consisting of `'NSteps'` and a positive integer.

Example: `'NSteps',5`

Data Types: `single` | `double`

`PEnter` — Threshold for criterion to add term
scalar value

Threshold for the criterion to add a term, specified as the comma-separated pair consisting of `'PEnter'` and a scalar value, as described in this table.

Criterion	Default Value	Decision
`'SSE'`	0.05	If the *p*-value of the *F*-statistic is less than `PEnter` (*p*-value to enter), add the term to the model.

Criterion	Default Value	Decision
'AIC'	0	If the change in the AIC of the model is less than PEnter, add the term to the model.
'BIC'	0	If the change in the BIC of the model is less than PEnter, add the term to the model.
'Rsquared'	0.1	If the increase in the R-squared value of the model is greater than PEnter, add the term to the model.
'AdjRsquared'	0	If the increase in the adjusted R-squared value of the model is greater than PEnter, add the term to the model.

For more information, see the Criterion name-value pair argument.

Example: 'PEnter',0.075

PredictorVars — Predictor variables
string array | cell array of character vectors | logical or numeric index vector

Predictor variables to use in the fit, specified as the comma-separated pair consisting of 'PredictorVars' and either a string array or cell array of character vectors of the variable names in the table or dataset array tbl, or a logical or numeric index vector indicating which columns are predictor variables.

The string values or character vectors should be among the names in tbl, or the names you specify using the 'VarNames' name-value pair argument.

The default is all variables in X, or all variables in tbl except for ResponseVar.

For example, you can specify the second and third variables as the predictor variables using either of the following examples.

Example: 'PredictorVars',[2,3]

Example: 'PredictorVars',logical([0 1 1 0 0 0])

Data Types: single | double | logical | string | cell

PRemove — Threshold for criterion to remove term
scalar value

Threshold for the criterion to remove a term, specified as the comma-separated pair consisting of 'PRemove' and a scalar value, as described in this table.

Criterion	Default Value	Decision
'SSE'	0.10	If the *p*-value of the *F*-statistic is greater than PRemove (*p*-value to remove), remove the term from the model.

Criterion	Default Value	Decision
'AIC'	0.01	If the change in the AIC of the model is greater than PRemove, remove the term from the model.
'BIC'	0.01	If the change in the BIC of the model is greater than PRemove, remove the term from the model.
'Rsquared'	0.05	If the increase in the R-squared value of the model is less than PRemove, remove the term from the model.
'AdjRsquared'	-0.05	If the increase in the adjusted R-squared value of the model is less than PRemove, remove the term from the model.

At each step, the `stepwiselm` function also checks whether a term is redundant (linearly dependent) with other terms in the current model. When any term is linearly dependent with other terms in the current model, the `stepwiselm` function removes the redundant term, regardless of the criterion value.

For more information, see the `Criterion` name-value pair argument.

Example: `'PRemove',0.05`

`ResponseVar` — Response variable

last column in `tbl` (default) | character vector or string scalar containing variable name | logical or numeric index vector

Response variable to use in the fit, specified as the comma-separated pair consisting of `'ResponseVar'` and either a character vector or string scalar containing the variable name in the table or dataset array `tbl`, or a logical or numeric index vector indicating which column is the response variable. You typically need to use `'ResponseVar'` when fitting a table or dataset array `tbl`.

For example, you can specify the fourth variable, say `yield`, as the response out of six variables, in one of the following ways.

Example: `'ResponseVar','yield'`

Example: `'ResponseVar',[4]`

Example: `'ResponseVar',logical([0 0 0 1 0 0])`

Data Types: `single` | `double` | `logical` | `char` | `string`

`Upper` — Model specification describing largest set of terms in fit

`'interactions'` (default) | character vector | string scalar | terms matrix

Model specification describing the largest set of terms in the fit, specified as the comma-separated pair consisting of `'Upper'` and one of the options for `modelspec` naming the model.

Example: `'Upper','quadratic'`

`VarNames` — Names of variables
`{'x1','x2',...,'xn','y'}` (default) | string array | cell array of character vectors

Names of variables, specified as the comma-separated pair consisting of `'VarNames'` and a string array or cell array of character vectors including the names for the columns of `X` first, and the name for the response variable `y` last.

`'VarNames'` is not applicable to variables in a table or dataset array, because those variables already have names.

The variable names do not have to be valid MATLAB identifiers. However, if the names are not valid, you cannot use a formula when you fit or adjust a model; for example:

- You cannot use a formula to specify the terms to add or remove when you use the `addTerms` function or the `removeTerms` function, respectively.
- You cannot use a formula to specify the lower and upper bounds of the model when you use the `step` or `stepwiselm` function with the name-value pair arguments `'Lower'` and `'Upper'`, respectively.

Before specifying `'VarNames',varNames`, you can verify the variable names in `varNames` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Example: `'VarNames',{'Horsepower','Acceleration','Model_Year','MPG'}`

Data Types: `string` | `cell`

`Verbose` — Control for display of information
`1` (default) | `0` | `2`

Control for the display of information, specified as the comma-separated pair consisting of `'Verbose'` and one of these values:

- `0` — Suppress all display.
- `1` — Display the action taken at each step.
- `2` — Display the evaluation process and the action taken at each step.

Example: `'Verbose',2`

`Weights` — Observation weights
`ones(n,1)` (default) | *n*-by-1 vector of nonnegative scalar values

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and an *n*-by-1 vector of nonnegative scalar values, where *n* is the number of observations.

Data Types: `single` | `double`

Output Arguments

`mdl` — Linear model
`LinearModel` object

Linear model representing a least-squares fit of the response to the data, returned as a `LinearModel` object.

For the properties and methods of the linear model object, `mdl`, see the `LinearModel` class page.

More About

Terms Matrix

A terms matrix `T` is a *t*-by-(*p* + 1) matrix specifying terms in a model, where *t* is the number of terms, *p* is the number of predictor variables, and +1 accounts for the response variable. The value of `T(i,j)` is the exponent of variable `j` in term `i`.

For example, suppose that an input includes three predictor variables `x1`, `x2`, and `x3` and the response variable `y` in the order `x1`, `x2`, `x3`, and `y`. Each row of `T` represents one term:

- `[0 0 0 0]` — Constant term or intercept
- `[0 1 0 0]` — `x2`; equivalently, `x1^0 * x2^1 * x3^0`
- `[1 0 1 0]` — `x1*x3`
- `[2 0 0 0]` — `x1^2`
- `[0 1 2 0]` — `x2*(x3^2)`

The `0` at the end of each term represents the response variable. In general, a column vector of zeros in a terms matrix represents the position of the response variable. If you have the predictor and response variables in a matrix and column vector, then you must include `0` for the response variable in the last column of each row.

Formula

A formula for model specification is a character vector or string scalar of the form `'y ~ terms'`.

- *y* is the response name.
- *terms* represents the predictor terms in a model using Wilkinson notation.

To represent predictor and response variables, use the variable names of the table input `tbl` or the variable names specified by using `VarNames`. The default value of `VarNames` is `{'x1','x2',...,'xn','y'}`.

For example:

- `'y ~ x1 + x2 + x3'` specifies a three-variable linear model with intercept.
- `'y ~ x1 + x2 + x3 – 1'` specifies a three-variable linear model without intercept. Note that formulas include a constant (intercept) term by default. To exclude a constant term from the model, you must include `–1` in the formula.

A formula includes a constant term unless you explicitly remove the term using `–1`.

Wilkinson Notation

Wilkinson notation describes the terms present in a model. The notation relates to the terms present in a model, not to the multipliers (coefficients) of those terms.

Wilkinson notation uses these symbols:

- + means include the next variable.

- – means do not include the next variable.
- `:` defines an interaction, which is a product of terms.
- * defines an interaction and all lower-order terms.
- ^ raises the predictor to a power, exactly as in * repeated, so ^ includes lower-order terms as well.
- `()` groups terms.

This table shows typical examples of Wilkinson notation.

Wilkinson Notation	Terms in Standard Notation
1	Constant (intercept) term
x1^k, where k is a positive integer	x1, $x1^2$, ..., $x1^k$
x1 + x2	x1, x2
x1*x2	x1, x2, x1*x2
x1:x2	x1*x2 only
–x2	Do not include x2
x1*x2 + x3	x1, x2, x3, x1*x2
x1 + x2 + x3 + x1:x2	x1, x2, x3, x1*x2
x1*x2*x3 – x1:x2:x3	x1, x2, x3, x1*x2, x1*x3, x2*x3
x1*(x2 + x3)	x1, x2, x3, x1*x2, x1*x3

For more details, see “Wilkinson Notation” on page 11-91.

Tips

- You cannot use robust regression with stepwise regression. Check your data for outliers before using `stepwiselm`.
- For other methods such as `anova`, or properties of the `LinearModel` object, see `LinearModel`.
- After training a model, you can generate C/C++ code that predicts responses for new data. Generating C/C++ code requires MATLAB Coder. For details, see “Introduction to Code Generation” on page 32-2.

Algorithms

- Stepwise regression is a systematic method for adding and removing terms from a linear or generalized linear model based on their statistical significance in explaining the response variable. The method begins with an initial model, specified using `modelspec`, and then compares the explanatory power of incrementally larger and smaller models.

  The `stepwiselm` function uses forward and backward stepwise regression to determine a final model. At each step, the function searches for terms to add to the model or remove from the model based on the value of the `'Criterion'` name-value pair argument.

  The default value of `'Criterion'` for a linear regression model is `'sse'`. In this case, `stepwiselm` and `step` of `LinearModel` use the *p*-value of an *F*-statistic to test models with and without a potential term at each step. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to

reject the null hypothesis, the function adds the term to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the function removes the term from the model.

Stepwise regression takes these steps when `'Criterion'` is `'sse'`:

1. Fit the initial model.
2. Examine a set of available terms not in the model. If any of the terms have *p*-values less than an entrance tolerance (that is, if it is unlikely a term would have a zero coefficient if added to the model), add the term with the smallest *p*-value and repeat this step; otherwise, go to step 3.
3. If any of the available terms in the model have *p*-values greater than an exit tolerance (that is, the hypothesis of a zero coefficient cannot be rejected), remove the term with the largest *p*-value and return to step 2; otherwise, end the process.

At any stage, the function will not add a higher-order term if the model does not also include all lower-order terms that are subsets of the higher-order term. For example, the function will not try to add the term `X1:X2^2` unless both `X1` and `X2^2` are already in the model. Similarly, the function will not remove lower-order terms that are subsets of higher-order terms that remain in the model. For example, the function will not try to remove `X1` or `X2^2` if `X1:X2^2` remains in the model.

The default value of `'Criterion'` for a generalized linear model is `'Deviance'`. `stepwiseglm` and `step` of `GeneralizedLinearModel` follow a similar procedure for adding or removing terms.

You can specify other criteria by using the `'Criterion'` name-value pair argument. For example, you can specify the change in the value of the Akaike information criterion, Bayesian information criterion, R-squared, or adjusted R-squared as the criterion to add or remove terms.

Depending on the terms included in the initial model, and the order in which the function adds and removes terms, the function might build different models from the same set of potential terms. The function terminates when no single step improves the model. However, a different initial model or a different sequence of steps does not guarantee a better fit. In this sense, stepwise models are locally optimal, but might not be globally optimal.

- `stepwiselm` treats a categorical predictor as follows:
  - A model with a categorical predictor that has *L* levels (categories) includes *L* – 1 indicator variables. The model uses the first category as a reference level, so it does not include the indicator variable for the reference level. If the data type of the categorical predictor is `categorical`, then you can check the order of categories by using `categories` and reorder the categories by using `reordercats` to customize the reference level. For more details about creating indicator variables, see “Automatic Creation of Dummy Variables” on page 2-49.
  - `stepwiselm` treats the group of *L* – 1 indicator variables as a single variable. If you want to treat the indicator variables as distinct predictor variables, create indicator variables manually by using `dummyvar`. Then use the indicator variables, except the one corresponding to the reference level of the categorical variable, when you fit a model. For the categorical predictor `X`, if you specify all columns of `dummyvar(X)` and an intercept term as predictors, then the design matrix becomes rank deficient.
  - Interaction terms between a continuous predictor and a categorical predictor with *L* levels consist of the element-wise product of the *L* – 1 indicator variables with the continuous predictor.

- Interaction terms between two categorical predictors with L and M levels consist of the (L – 1)*(M – 1) indicator variables to include all possible combinations of the two categorical predictor levels.
- You cannot specify higher-order terms for a categorical predictor because the square of an indicator is equal to itself.

Therefore, if `stepwiselm` adds or removes a categorical predictor, the function actually adds or removes the group of indicator variables in one step. Similarly, if `stepwiselm` adds or removes an interaction term with a categorical predictor, the function actually adds or removes the group of interaction terms including the categorical predictor.

- `stepwiselm` considers `NaN`, `''` (empty character vector), `""` (empty string), `<missing>`, and `<undefined>` values in `tbl`, `X`, and `Y` to be missing values. `stepwiselm` does not use observations with missing values in the fit. The `ObservationInfo` property of a fitted model indicates whether or not `stepwiselm` uses each observation in the fit.

Alternative Functionality

- You can construct a model using `fitlm`, and then manually adjust the model using `step`, `addTerms`, or `removeTerms`.

See Also

`LinearModel` | `fitlm` | `step`

Topics

“Stepwise Regression” on page 11-99
“Linear Regression” on page 11-9
“Linear Regression Workflow” on page 11-35
“Interpret Linear Regression Results” on page 11-50
“Linear Regression with Categorical Covariates” on page 2-52
“Sequential Feature Selection” on page 15-61

Introduced in R2013b

stepwisefit

Fit linear regression model using stepwise regression

Syntax

```
b = stepwisefit(X,y)
b = stepwisefit(X,y,Name,Value)
[b,se,pval] = stepwisefit(___)
[b,se,pval,finalmodel,stats] = stepwisefit(___)
[b,se,pval,finalmodel,stats,nextstep,history] = stepwisefit(___)
```

Description

`b = stepwisefit(X,y)` returns a vector `b` of coefficient estimates from stepwise regression of the response vector `y` on the predictor variables in matrix `X`. `stepwisefit` begins with an initial constant model and takes forward or backward steps to add or remove variables, until a stopping criterion is satisfied.

`b = stepwisefit(X,y,Name,Value)` specifies additional options using one or more name-value pair arguments. For example, you can specify a nonconstant initial model, or a maximum number of steps that `stepwisefit` can take.

`[b,se,pval] = stepwisefit(___)` returns the coefficient estimates `b`, standard errors `se`, and *p*-values `pval` using any of the input argument combinations in previous syntaxes.

`[b,se,pval,finalmodel,stats] = stepwisefit(___)` also returns a specification of the variables in the final regression model `finalmodel`, and statistics `stats` about the final model.

`[b,se,pval,finalmodel,stats,nextstep,history] = stepwisefit(___)` also returns the recommended next step `nextstep` and information `history` about all the steps taken.

Examples

Stepwise Regression with Default Arguments

Perform a basic stepwise regression and obtain the coefficient estimates.

Load the `hald` data set.

```
load hald
whos % Check variables loaded in workspace
```

```
  Name              Size            Bytes  Class     Attributes

  Description      22x58             2552  char
  hald             13x5               520  double
  heat             13x1               104  double
  ingredients      13x4               416  double
```

This data set contains observations of the heat evolved during cement hardening for various mixtures of four cement ingredients. The response variable is `heat`. The matrix `ingredients` contains four columns of predictors.

Run `stepwisefit` beginning with only a constant term in the model and using the default entry and exit tolerances of 0.05 and 0.10, respectively.

```
b = stepwisefit(ingredients,heat)
```

```
Initial columns included: none
Step 1, added column 4, p=0.000576232
Step 2, added column 1, p=1.10528e-06
Final columns included:  1 4
    {'Coeff'  }    {'Std.Err.'}    {'Status'}    {'P'         }
    {[ 1.4400]}    {[  0.1384]}    {'In'    }    {[1.1053e-06]}
    {[ 0.4161]}    {[  0.1856]}    {'Out'   }    {[    0.0517]}
    {[-0.4100]}    {[  0.1992]}    {'Out'   }    {[    0.0697]}
    {[-0.6140]}    {[  0.0486]}    {'In'    }    {[1.8149e-07]}

b = 4×1

    1.4400
    0.4161
   -0.4100
   -0.6140
```

The `stepwisefit` display shows that columns 1 and 4 are included in the final model. The output `b` includes estimates for all columns, even those that do not appear in the final model. `stepwisefit` computes the estimate for column 2 (or 3) by fitting a model consisting of the final model plus column 2 (or 3).

Tune the Stepwise Procedure

Load the `carsmall` data set, which contains various car measurements.

```
load carsmall
whos
```

```
  Name               Size            Bytes  Class     Attributes

  Acceleration      100x1               800  double
  Cylinders         100x1               800  double
  Displacement      100x1               800  double
  Horsepower        100x1               800  double
  MPG               100x1               800  double
  Mfg               100x13             2600  char
  Model             100x33             6600  char
  Model_Year        100x1               800  double
  Origin            100x7              1400  char
  Weight            100x1               800  double
```

Perform stepwise regression with four continuous variables and the response variable `MPG`.

```
X = [Acceleration Cylinders Displacement Horsepower];
y = MPG;
b4_default = stepwisefit(X,y) % Stepwise regression with default arguments
```

```
Initial columns included: none
Step 1, added column 2, p=1.59001e-25
Step 2, added column 4, p=0.00364266
Step 3, added column 1, p=0.0161414
Final columns included:  1 2 4
    {'Coeff'  }    {'Std.Err.'}    {'Status'}    {'P'         }
    {[-0.4517]}    {[  0.1842]}    {'In'    }    {[    0.0161]}
    {[-2.6407]}    {[  0.4823]}    {'In'    }    {[4.0003e-07]}
    {[ 0.0148]}    {[  0.0157]}    {'Out'   }    {[    0.3472]}
    {[-0.0772]}    {[  0.0204]}    {'In'    }    {[2.6922e-04]}

b4_default = 4×1

   -0.4517
   -2.6407
    0.0148
   -0.0772
```

The term `Displacement` never enters the model. Determine if it is highly correlated with the other three terms by computing the term correlation matrix.

```
corrcoef(X,'Rows','complete') % To exclude rows with missing values from calculation
```

```
ans = 4×4

    1.0000   -0.6438   -0.6968   -0.6968
   -0.6438    1.0000    0.9517    0.8622
   -0.6968    0.9517    1.0000    0.9134
   -0.6968    0.8622    0.9134    1.0000
```

The third row of the correlation matrix corresponds to `Displacement`. This term is highly correlated with the other three terms, especially `Cylinders` (`0.95`) and `Horsepower` (`0.91`).

Redefine the input matrix `X` to include `Weight`. Specify an initial model containing the terms `Displacement` and `Horsepower` by using the `'InModel'` name-value pair argument.

```
X = [Acceleration Cylinders Displacement Horsepower Weight];
inmodel = [false false true true false];
b5_inmodel = stepwisefit(X,y,'InModel',inmodel)
```

```
Initial columns included: 3 4
Step 1, added column 5, p=1.06457e-06
Step 2, added column 2, p=0.00410234
Final columns included:  2 3 4 5
    {'Coeff'  }    {'Std.Err.'}    {'Status'}    {'P'         }
    {[-0.0912]}    {[  0.2032]}    {'Out'   }    {[    0.6548]}
    {[-2.3223]}    {[  0.7879]}    {'In'    }    {[    0.0041]}
    {[ 0.0252]}    {[  0.0145]}    {'In'    }    {[    0.0862]}
    {[-0.0449]}    {[  0.0231]}    {'In'    }    {[    0.0555]}
    {[-0.0050]}    {[  0.0012]}    {'In'    }    {[1.0851e-04]}

b5_inmodel = 5×1
```

```
   -0.0912
   -2.3223
    0.0252
   -0.0449
   -0.0050
```

The final model consists of terms 2–5. However, `Displacement` and `Horsepower` estimates have *p*-values greater than `0.05` in the final model. You can tune the stepwise algorithm to behave more conservatively by using the `'PRemove'` name-value pair argument. For example, setting `'PRemove'` to `0.05` (instead of the default `0.1`) results in a smaller final model with only two terms, each with a *p*-value less than `0.05`.

```
b5_inmodel_premove = stepwisefit(X,y,'InModel',inmodel,'PRemove',0.05)
```

```
Initial columns included: 3 4
Step 1, added column 5, p=1.06457e-06
Step 2, added column 2, p=0.00410234
Step 3, removed column 3, p=0.0862131
Step 4, removed column 4, p=0.239239
Final columns included:  2 5
    {'Coeff'  }    {'Std.Err.'}    {'Status'}    {'P'         }
    {[-0.0115]}    {[  0.1656]}    {'Out'   }    {[    0.9449]}
    {[-1.6037]}    {[  0.5146]}    {'In'    }    {[    0.0025]}
    {[ 0.0101]}    {[  0.0124]}    {'Out'   }    {[    0.4186]}
    {[-0.0234]}    {[  0.0198]}    {'Out'   }    {[    0.2392]}
    {[-0.0055]}    {[  0.0011]}    {'In'    }    {[3.9038e-06]}

b5_inmodel_premove = 5×1

   -0.0115
   -1.6037
    0.0101
   -0.0234
   -0.0055
```

Center and scale each column (compute the *z*-scores) before fitting by using the `'Scale'` name-value pair argument. The scaling does not change the model selected, the signs of coefficient estimates, or their *p*-values. However, the scaling does scale the coefficient estimates.

```
b5_inmodel_premove_scale = stepwisefit(X,y,'InModel',inmodel,'PRemove',0.05,'Scale','on')
```

```
Initial columns included: 3 4
Step 1, added column 5, p=1.06457e-06
Step 2, added column 2, p=0.00410234
Step 3, removed column 3, p=0.0862131
Step 4, removed column 4, p=0.239239
Final columns included:  2 5
    {'Coeff'  }    {'Std.Err.'}    {'Status'}    {'P'         }
    {[-0.0370]}    {[  0.5339]}    {'Out'   }    {[    0.9449]}
    {[-2.8136]}    {[  0.9028]}    {'In'    }    {[    0.0025]}
    {[ 1.1155]}    {[  1.3726]}    {'Out'   }    {[    0.4186]}
    {[-1.0617]}    {[  0.8961]}    {'Out'   }    {[    0.2392]}
    {[-4.4406]}    {[  0.9028]}    {'In'    }    {[3.9038e-06]}

b5_inmodel_premove_scale = 5×1
```

```
   -0.0370
   -2.8136
    1.1155
   -1.0617
   -4.4406
```

Usually, you scale to compare estimates of terms that are measured in different scales, such as `Horsepower` and `Weight`. In this case, increasing `Horsepower` by one standard deviation leads to an expected drop of `1` in `MPG`, whereas increasing `Weight` by one standard deviation leads to an expected drop of `4.4` in `MPG`.

Retrieve Detailed Output from Stepwise Regression

Load the `imports-85` data set. This data set contains characteristics of cars imported in 1985. For a list of all column names, see the variable `Description` in the workspace or type `Description` at the command line.

```
load imports-85
whos
```

```
  Name               Size            Bytes  Class     Attributes

  Description        9x79             1422  char
  X                205x26            42640  double
```

Choose a subset of continuous variables to use in stepwise regression, consisting of the predictor variables `engine-size`, `bore`, `stroke`, `compression-ratio`, `horsepower`, `peak-rpm`, `city-mpg`, and `highway-mpg`, and the response variable `price`.

```
varnames = ["engine-size","bore","stroke","compression-ratio","horsepower","peak-rpm","city-mpg"
dataTbl = array2table(X(:,8:16),'VariableNames',varnames); % Create data table with variable name
Xstepw = dataTbl{:,{'engine-size','bore','stroke','compression-ratio','horsepower','peak-rpm','ci
ystepw = dataTbl{:,{'price'}}; % Response vector
```

Run `stepwisefit` of the variable `price` on the other eight variables, first with the default constant initial model and then with an initial model including `highway-mpg`. Omit the display of step information.

```
[betahat_def,se_def,pval_def,finalmodel_def,stats_def] = stepwisefit(Xstepw,ystepw,'Display',"of
inmodel = [false false false false false false false true];
[betahat_in,se_in,pval_in,finalmodel_in,stats_in] = stepwisefit(Xstepw,ystepw,'InModel',inmodel,
```

Inspect the final models returned by `stepwisefit`.

```
finalmodel_def
```

```
finalmodel_def = 1x8 logical array

   1   0   1   1   0   1   1   0
```

```
finalmodel_in
```

```
finalmodel_in = 1x8 logical array
```

```
   1   0   1   1   0   1   0   1
```

The default model drops `highway-mpg` (term 8) from the model and includes `city-mpg` (term 7) instead. Compare the root mean squared errors (RMSEs) of these two final models.

```
stats_def.rmse
```

```
ans = 3.3033e+03
```

```
stats_in.rmse
```

```
ans = 3.3324e+03
```

The model resulting from the default arguments has a slightly lower RMSE. Note that a full specification of the final model consists of the term estimates plus the intercept estimate.

```
betahat_def % Term estimates
```

```
betahat_def = 8×1
10^3 ×

    0.1559
   -0.2242
   -2.8578
    0.3904
    0.0222
    0.0024
   -0.2414
    0.0793
```

```
stats_def.intercept % Intercept estimate
```

```
ans = -7.3506e+03
```

Retrieve the history of the default run of `stepwisefit` and the recommended next step. Omit the display of step information.

```
[~,~,~,~,~,nextstep_def,history_def]=stepwisefit(Xstepw,ystepw,'Display',"off");
nextstep_def
```

```
nextstep_def = 0
```

No further steps are recommended (`nextstep_def` is `0`).

```
history_def.('in')
```

```
ans = 7x8 logical array

   1   0   0   0   0   0   0   0
   1   0   0   0   1   0   0   0
   1   0   0   1   1   0   0   0
   1   0   1   1   1   0   0   0
   1   0   1   1   1   1   0   0
   1   0   1   1   1   1   1   0
   1   0   1   1   0   1   1   0
```

The algorithm performs a total of seven steps. The output shows that `engine-size` (term `1`) is added in step `1`, `horsepower` (term `5`) is added in step `2`, and so on.

Input Arguments

`X` — Predictor variables

numeric matrix

Predictor variables, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictor variables. Each column of `X` represents one variable, and each row represents one observation.

`stepwisefit` always includes a constant term in the model. Therefore, do not include a column of 1s in `X`.

Data Types: `single` | `double`

`y` — Response variable

numeric or logical vector

Response variable, specified as an *n*-by-1 numeric or logical vector, where *n* is the number of observations. Each entry in `y` is the response for the corresponding row of `X`.

Data Types: `single` | `double` | `logical`

Note `stepwisefit` treats `NaN` values in either `X` or `y` as missing and ignores all rows containing these values.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'PEnter',0.10,'PRemove',0.15,'MaxIter',8` instructs `stepwisefit` to use entry and exit tolerances of `0.10` and `0.15`, respectively, and to take a maximum of 8 steps.

`InModel` — Terms for initial model

logical vector

Terms for the initial model, specified as the comma-separated pair consisting of `'InModel'` and a logical vector specifying terms to include in the initial model. The default is to include no terms.

Example: `'InModel',[true false false true]`

Data Types: `logical`

`PEnter` — Tolerance for adding terms to model

`0.05` (default) | positive scalar

Tolerance for adding terms to the model, specified as the comma-separated pair consisting of `'PEnter'` and a positive scalar specifying the maximum *p*-value for a term to be added. The default is 0.05.

Example: 'PEnter',0.10

Data Types: single | double

PRemove — Tolerance for removing terms from model
maximum of PEnter and 0.10 (default) | positive scalar

Tolerance for removing terms from the model, specified as the comma-separated pair consisting of 'PRemove' and a positive scalar specifying the minimum *p*-value for a term to be removed. The default is the maximum of PEnter and 0.10.

Note PRemove is not allowed to be smaller than PEnter because that would cause stepwisefit to enter an infinite loop, where a variable is repeatedly added to the model and removed from the model.

Example: 'PRemove',0.15

Data Types: single | double

Display — Indicator for displaying step information
'on' (default) | 'off'

Indicator for displaying step information, specified as the comma-separated pair consisting of 'Display' and 'on' or 'off'.

- 'on' displays information about each step in the Command Window (default).
- 'off' omits the display.

Example: 'Display','off'

MaxIter — Maximum number of steps
Inf (default) | positive integer

Maximum number of steps, specified as the comma-separated pair consisting of 'MaxIter' and a positive integer or Inf (default). Inf allows the algorithm to run until no single step improves the model.

Example: 'MaxIter',12

Data Types: double

Keep — Terms to keep in their initial state
logical vector

Terms to keep in their initial state, specified as the comma-separated pair consisting of 'Keep' and a logical vector. The value true for a term specified to be in (or out of) the initial model forces that term to remain in (or out of) the final model. The value false for a term does not force that term to remain in (or out of) the final model. The default is to specify no terms to keep in their initial state.

Example: 'Keep',[true true false false]

Data Types: logical

Scale — Indicator for centering and scaling terms
'off' (default) | 'on'

Indicator for centering and scaling terms, specified as the comma-separated pair consisting of `'Scale'` and `'off'` or `'on'`.

- `'off'` does not center and scale the terms (default).
- `'on'` centers and scales each column of `X` (computes the *z*-scores) before fitting.

Example: `'Scale','on'`

Output Arguments

b — Estimated coefficients

numeric vector

Estimated coefficients, returned as a numeric vector corresponding to the terms in `X`. The `stepwisefit` function calculates the values in `b` as follows:

- If a term is included in the final model, then its corresponding value in `b` is the estimate resulting from fitting the final model.
- If a term is excluded from the final model, then its corresponding value in `b` is the estimate resulting from fitting the final model plus that term.

Note To obtain a full specification of the fitted model, you also need the estimated intercept in addition to `b`. The estimated intercept is provided as a field in the output argument `stats`. For more details see “stepwisefit Fitted Model” on page 33-6045.

se — Standard errors

numeric vector

Standard errors, returned as a numeric vector corresponding to the estimates in `b`.

pval — *p*-values

numeric vector

p-values, returned as a numeric vector that results from testing whether elements of `b` are `0`.

finalmodel — Final model

logical vector

Final model, returned as a logical vector with length equal to the number of columns in `X`, indicating which terms are in the final model.

stats — Additional statistics

structure

Additional statistics, returned as a structure with the following fields. All statistics pertain to the final model except where noted.

Field	Description
`source`	Character vector `'stepwisefit'`
`dfe`	Degrees of freedom for error

Field	Description
df0	Degrees of freedom for the regression
SStotal	Total sum of squares of the response
SSresid	Sum of squares of the residuals
fstat	*F*-statistic for testing the final model vs. no model (mean only)
pval	*p*-value of the *F*-statistic
rmse	Root mean squared error
xr	Residuals for terms not in the final model, computed by subtracting from each term the predicted response of the final model
yr	Residuals for the response using predictors in the final model
B	Coefficients for terms in the final model, with the value for each term not in the model set to the value that would be obtained by adding that term to the model
SE	Standard errors for coefficient estimates
TSTAT	*t* statistics for coefficient estimates
PVAL	*p*-values for coefficient estimates
intercept	Estimated intercept
wasnan	Rows in the data that contain NaN values

nextstep — Recommended next step
nonnegative integer

Recommended next step, returned as a nonnegative integer equal to the index of the next term to add to or remove from the model, or 0 if no further steps are recommended.

history — Information on steps taken
structure

Information on steps taken, returned as a structure with the following fields.

Field	Description
B	Matrix of regression coefficients, where each column is one step and each row is one coefficient vector
rmse	Root mean squared errors for the model at each step
df0	Degrees of freedom for the regression at each step
in	Logical array indicating which predictors are in the model at each step, where each row is one step and each column is one predictor

More About

`stepwisefit` Fitted Model

The final `stepwisefit` fitted model is

$$\hat{y} = \text{stats.intercept} + X(:,\text{finalmodel}) * b(\text{finalmodel}).$$

Here,

- $\hat{y}$ is the predicted mean response.
- `stats.intercept` is the estimated intercept.
- X`(:,finalmodel)` is the input matrix for the terms in the final model.
- `b(finalmodel)` is the vector of coefficient estimates for the terms in the final model.

Algorithms

Stepwise regression is a method for adding terms to and removing terms from a multilinear model based on their statistical significance. This method begins with an initial model and then takes successive steps to modify the model by adding or removing terms. At each step, the *p*-value of an *F*-statistic is computed to test models with and without a potential term. If a term is not currently in the model, the null hypothesis is that the term would have a zero coefficient if added to the model. If there is sufficient evidence to reject the null hypothesis, the term is added to the model. Conversely, if a term is currently in the model, the null hypothesis is that the term has a zero coefficient. If there is insufficient evidence to reject the null hypothesis, the term is removed from the model. The method proceeds as follows:

1. Fit the initial model.
2. If any terms not in the model have *p*-values less than an entry tolerance, add the one with the smallest *p*-value and repeat this step. For example, assume the initial model is the default constant model and the entry tolerance is the default `0.05`. The algorithm first fits all models consisting of the constant plus another term and identifies the term that has the smallest *p*-value, for example term `4`. If the term `4` *p*-value is less than `0.05`, then term `4` is added to the model. Next, the algorithm performs a search among all models consisting of the constant, term `4`, and another term. If a term not in the model has a *p*-value less than `0.05`, the term with the smallest *p*-value is added to the model and the process is repeated. When no further terms exist that can be added to the model, the algorithm proceeds to step 3.
3. If any terms in the model have *p*-values greater than an exit tolerance, remove the one with the largest *p*-value and go to step 2; otherwise, end.

In each step of the algorithm, `stepwisefit` uses the method of least squares to estimate the model coefficients. After adding a term to the model at an earlier stage, the algorithm might subsequently drop that term if it is no longer helpful in combination with other terms added later. The method terminates when no single step improves the model. However, the final model is not guaranteed to be optimal, which means having the best fit to the data. A different initial model or a different sequence of steps might lead to a better fit. In this sense, stepwise models are locally optimal, but are not necessarily globally optimal.

Alternative Functionality

- You can create a model using `fitlm`, and then manually adjust the model using `step`, `addTerms`, and `removeTerms`.
- Use `stepwiselm` if you have data in a table, you have a mix of continuous and categorical predictors, or you want to specify model formulas that can potentially include higher-order and interaction terms.
- Use `stepwiseglm` to create stepwise generalized linear models (for example, if you have a binary response variable and want to fit a classification model).

References

[1] Draper, Norman R., and Harry Smith. *Applied Regression Analysis*. Hoboken, NJ: Wiley-Interscience, 1998. pp. 307–312.

See Also

`addedvarplot` | `regress` | `stepwise` | `stepwiseglm` | `stepwiselm`

Introduced before R2006a

subsasgn

Class: `dataset`

(Not Recommended) Subscripted assignment to dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Description

`A = subsasgn(A,S,B)` is called for the syntax `A(i,j)=B`, `A{i,j}=B`, or `A.var=B` when `A` is a dataset array. `S` is a structure array with the fields:

`type`	`'()'`, `'{}'`, or `'.'` specifying the subscript type.
`subs`	Cell array or character vector containing the actual subscripts.

`A(i,j) = B` assigns the contents of the dataset array `B` to a subset of the observations and variables in the dataset array `A`. `i` and `j` are one of the following types:

- positive integers
- vectors of positive integers
- observation/variable names
- cell arrays containing one or more observation/variable names
- logical vectors

The assignment does not use observation names, variable names, or any other properties of `B` to modify properties of `A`; however properties of `A` are extended with default values if the assignment expands the number of observations or variables in `A`. Elements of `B` are assigned into `A` by position, not by matching names.

`A{i,j} = B` assigns the value `B` into an element of the dataset array `A`. `i` and `J` are positive integers, or logical vectors. Cell indexing cannot assign into multiple dataset elements, that is, the subscripts `i` and `j` must each refer to only a single observation or variable. `B` is cast to the type of the target variable if necessary. If the dataset element already exists, `A{i,j}` may also be followed by further subscripting as supported by the variable.

For dataset variables that are cell arrays, assignments such as `A{1,'CellVar'} = B` assign into the contents of the target dataset element in the same way that `{}`-indexing of an ordinary cell array does.

For dataset variables that are `n`-D arrays, i.e., each observation is a matrix or array, an assignment such as `A{1,'ArrayVar'} = B` assigns into the second and following dimensions of the target dataset element, i.e., the assignment adds a leading singleton dimension to `B` to account for the observation dimension of the dataset variable.

`A.var = B or A.(varname) = B` assigns `B` to a dataset variable. `var` is a variable name literal, or `varname` is a character variable containing a variable name. If the dataset variable already exists, the

assignment completely replaces that variable. To assign into an element of the variable, `A.var` or `A.(varname)` may be followed by further subscripting as supported by the variable. In particular, `A.var(obsnames,...) = B` and `A.var{obsnames,...} = B` (when supported by `var`) provide assignment into a dataset variable using observation names.

`A.properties.propertyname = P` assigns to a dataset property. `propertyname` is one of the following:

- `'ObsNames'`
- `'VarNames'`
- `'Description'`
- `'Units'`
- `'DimNames'`
- `'UserData'`
- `'VarDescription'`

To assign into an element of the property, `A.properties.propertyname` may also be followed by further subscripting as supported by the property.

You cannot assign multiple values into dataset variables or properties using assignments such as `[A.CellVar{1:2}] = B`, `[A.StructVar(1:2).field] = B`, or `[A.Properties.ObsNames{1:2}] = B`. Use multiple assignments of the form `A.CellVar{1} = B` instead.

Similarly, if a dataset variable is a cell array with multiple columns or is an n-D cell array, then the contents of that variable for a single observation consists of multiple cells, and you cannot assign to all of them using the syntax `A{1,'CellVar'} = B`. Use multiple assignments of the form `[A.CellVar{1,1}] = B` instead.

See Also

`dataset` | `set` | `subsref`

subsref

Class: `dataset`

(Not Recommended) Subscripted reference for dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
B = subsref(A,S)
```

Description

`B = subsref(A,S)` is called for the syntax `A(i,j)`, `A{i,j}`, or `A.var` when `A` is a dataset array. `S` is a structure array with the fields:

`type`	Character vector containing `'()'`, `'{}'`, or `'.'` specifying the subscript type.
`subs`	Cell array or character vector containing the actual subscripts.

`B = A(i,j)` returns a dataset array that contains a subset of the observations and variables in the dataset array `A`. `i` and `j` are one of the following types:

- positive integers
- vectors of positive integers
- observation/variable names
- cell arrays containing one or more observation/variable names
- logical vectors

`B` contains the same property values as `A`, subsetted for observations or variables where appropriate.

`B = A{i,j}` returns an element of a dataset variable. `i` and `j` are positive integers, or logical vectors. Cell indexing cannot return multiple dataset elements, that is, the subscripts `i` and `j` must each refer to only a single observation or variable. `A{i,j}` may also be followed by further subscripting as supported by the variable.

For dataset variables that are cell arrays, expressions such as `A{1,'CellVar'}` return the contents of the referenced dataset element in the same way that `{}`-indexing on an ordinary cell array does. If the dataset variable is a single column of cells, the contents of a single cell is returned. If the dataset variable has multiple columns or is `n`-D, multiple outputs containing the contents of multiple cells are returned.

For dataset variables that are `n`-D arrays, i.e., each observation is a matrix or an array, expressions such as `A{1,'ArrayVar'}` return `A.ArrayVar(1,:,...)` with the leading singleton dimension squeezed out.

`B = A.var` or `A.(varname)` returns a dataset variable. `var` is a variable name literal, or `varname` is a character variable containing a variable name. `A.var` or `A.(varname)` may also be followed by further subscripting as supported by the variable. In particular, `A.var(obsnames,...)` and `A.var{obsnames,...}` (when supported by `var`) provide subscripting into a dataset variable using observation names.

`P = A.Properties.propertyname` returns a dataset property. `propertyname` is one of the following:

- `'ObsNames'`
- `'VarNames'`
- `'Description'`
- `'Units'`
- `'DimNames'`
- `'UserData'`
- `'VarDescription'`

`A.properties.propertyname` may also be followed by further subscripting as supported by the property.

Limitations

Subscripting expressions such as `A.CellVar{1:2}`, `A.StructVar(1:2).field`, or `A.Properties.ObsNames{1:2}` are valid, but result in `subsref` returning multiple outputs in the form of a comma-separated list. If you explicitly assign to output arguments on the left-hand side of an assignment, for example, `[cellval1,cellval2] = A.CellVar{1:2}`, those variables will receive the corresponding values. However, if there are no output arguments, only the first output in the comma-separated list is returned.

Similarly, if a dataset variable is a cell array with multiple columns or is an n-D cell array, then subscripting expressions such as `A{1,'CellVar'}` result in `subsref` returning the contents of multiple cells. You should explicitly assign to output arguments on the left-hand side of an assignment, for example, `[cellval1,cellval2] = A{1,'CellVar'}`.

See Also

`dataset` | `set` | `subsasgn`

summary

Class: `dataset`

(Not Recommended) Print summary of dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
summary(A)
s = summary(A)
```

Description

`summary(A)` prints a summary of a dataset array and the variables that it contains.

`s = summary(A)` returns a scalar structure `s` that contains a summary of the dataset `A` and the variables that `A` contains. For more information on the fields in `s`, see Outputs.

Summary information depends on the type of the variables in the data set:

- For numerical variables, `summary` computes a five-number summary of the data, giving the minimum, the first quartile, the median, the third quartile, and the maximum.
- For logical variables, `summary` counts the number of `true`s and `false`s in the data.
- For categorical variables, `summary` counts the number of data at each level.

Output Arguments

The following list describes the fields in the structure `s`:

- `Description` — A character array containing the dataset description.
- `Variables` — A structure array with one element for each dataset variable in A. Each element has the following fields:
  - `Name` — A character vector containing the name of the variable.
  - `Description` — A character vector containing the variable's description.
  - `Units` — A character vector containing the variable's units.
  - `Size` — A numeric vector containing the size of the variable.
  - `Class` — A character vector containing the class of the variable.
  - `Data` — A scalar structure containing the following fields.

    For numeric variables:

    - `Probabilities` — A numeric vector containing the probabilities [0.0 .25 .50 .75 1.0] and NaN (if any are present in the corresponding dataset variable).

- `Quantiles` — A numeric vector containing the values that correspond to 'Probabilities' for the corresponding dataset variable, and a count of NaNs (if any are present).

For logical variables:

- `Values` — The logical vector [true false].
- `Counts` — A numeric vector of counts for each logical value.

For categorical variables:

- `Levels` — A cell array containing the labels for each level of the corresponding dataset variable.
- `Counts` — A numeric vector of counts for each level.

`'Data'` is empty if variable is not numeric, categorical, or logical. If a dataset variable has more than one column, then the corresponding `'Quantiles'` or `'Counts'` field is a matrix or an array.

Examples

Summarize Fisher's iris data:

```
load fisheriris
species = nominal(species);
data = dataset(species,meas);
summary(data)
species: [150x1 nominal]
  setosa   versicolor   virginica
      50           50          50
meas: [150x4 double]
  min       4.3000         2         1    0.1000
  1st Q     5.1000    2.8000    1.6000    0.3000
  median    5.8000         3    4.3500    1.3000
  3rd Q     6.4000    3.3000    5.1000    1.8000
  max       7.9000    4.4000    6.9000    2.5000
```

Summarize the data in `hospital.mat`:

```
load hospital
summary(hospital)

Dataset array created from the data file hospital.dat.

The first column of the file ("id") is used for observation
names.  Other columns ("sex" and "smoke") have been
converted from their original coded values into categorical
and logical variables.  Two sets of columns ("sys" and
"dia", "trial1" through "trial4") have been combined into
single variables with multivariate observations.  Column
headers have been replaced with more descriptive variable
names.  Units have been added where appropriate.


LastName: [100x1 cell array of character vectors]
Sex: [100x1 nominal]
     Female      Male
```

```
         53        47

Age: [100x1 double, Units = Yrs]
     min      1st Q      median      3rd Q      max
      25         32          39         44       50

Weight: [100x1 double, Units = Lbs]
     min      1st Q         median        3rd Q        max
     111      130.5000      142.5000      180.5000     202

Smoker: [100x1 logical]
     true      false
       34         66

BloodPressure: [100x2 double, Units = mm Hg]
Systolic/Diastolic
     min              109           68
     1st Q       117.5000      77.5000
     median           122      81.5000
     3rd Q       127.5000           89
     max              138           99

Trials: [100x1 cell, Units = Counts]
From zero to four measurement trials performed
```

See Also

get | grpstats | set

struct2dataset

(Not Recommended) Convert structure array to dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = struct2dataset(S)
ds = struct2dataset(S,Name,Value)
```

Description

`ds = struct2dataset(S)` converts a structure array to a dataset array.

`ds = struct2dataset(S,Name,Value)` performs the conversion using additional options specified by one or more `Name,Value` pair arguments.

Examples

Convert Scalar Structure Array to Dataset Array

Convert a scalar structure array to a dataset array using the default options.

Create a structure array to convert.

```
S.Name = {'CLARK';'BROWN';'MARTIN'};
S.Gender = {'M';'F';'M'};
S.SystolicBP = [124;122;130];
S.DiastolicBP = [93;80;92];
S
```

```
S = struct with fields:
           Name: {3x1 cell}
         Gender: {3x1 cell}
     SystolicBP: [3x1 double]
    DiastolicBP: [3x1 double]
```

The scalar structure array has four fields, each with three rows.

Convert the structure array to a dataset array.

```
ds = struct2dataset(S)
```

```
ds =
    Name              Gender       SystolicBP    DiastolicBP
    {'CLARK' }        {'M'}        124           93
    {'BROWN' }        {'F'}        122           80
```

```
    {'MARTIN'}        {'M'}        130           92
```

The structure field names in `S` become the variable names in the output dataset array. The size of `ds` is 3-by-4.

Convert Nonscalar Structure Array to Dataset Array

Convert a nonscalar structure array to a dataset array, using one of the structure fields for observation names.

Create a nonscalar structure array to convert.

```
S(1,1).Name = 'CLARK';
S(1,1).Gender = 'M';
S(1,1).SystolicBP = 124;
S(1,1).DiastolicBP = 93;

S(2,1).Name = 'BROWN';
S(2,1).Gender = 'F';
S(2,1).SystolicBP = 122;
S(2,1).DiastolicBP = 80;

S(3,1).Name = 'MARTIN';
S(3,1).Gender = 'M';
S(3,1).SystolicBP = 130;
S(3,1).DiastolicBP = 92;

S
```

```
S=3×1 struct array with fields:
    Name
    Gender
    SystolicBP
    DiastolicBP
```

This is a 3-by-1 structure array with 4 fields.

Convert the structure array to a dataset array, using the `Name` field for observation names.

```
ds = struct2dataset(S,'ReadObsNames','Name')
```

```
ds =
              Gender       SystolicBP    DiastolicBP
    CLARK     {'M'}        124           93
    BROWN     {'F'}        122           80
    MARTIN    {'M'}        130           92
```

The size of `ds` is 3-by-3 because the structure field `Name` is used for observation names, and not as a dataset array variable.

```
ds.Properties.DimNames
```

```
ans = 1x2 cell
    {'Name'}    {'Variables'}
```

```
ds.Properties.ObsNames
```

```
ans = 3x1 cell
    {'CLARK' }
    {'BROWN' }
    {'MARTIN'}
```

Input Arguments

S — Input structure array

structure array

Input structure array to convert to a dataset array, specified as a scalar structure array with *N* fields, each with *M* rows, or a nonscalar *M*-by-1 structure array with *N* fields.

Data Types: `struct`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ReadObsNames','myField'` specifies that the structure field, `myField`, contains observation names.

ReadObsNames — Name of structure field containing observation names for dataset array

`false` (default) | character vector | string scalar

Name of structure field containing observation names for the output dataset array, specified as the comma-separated pair consisting of `'ReadObsNames'` and a character vector or string scalar containing a field name from the input structure array, `S`. When you specify a field name, `struct2dataset` uses that field to create observation names, and sets `ds.Properties.DimNames` equal to `{ReadObsNames,'Variables'}`.

For example, to specify that observation names are in the structure field, `Names`, use

Example: `'ReadObsNames','Names'`

By default, or if `ReadObsNames` is equal to `false`, `struct2dataset` does not create observation names unless you specify names using the name-value pair argument `ObsNames`.

ObsNames — Observation names for dataset array

string array | cell array of character vectors

Observation names for the output dataset array, specified as the comma-separated pair consisting of `'ObsNames'` and a string array or cell array of character vectors containing observation names. The names do not need to be valid MATLAB identifiers, but they must be unique.

AsScalar — Indicator for how to treat scalar structure

`false` | `true`

Indicator for how to treat a scalar input structure array, specified as the comma-separated pair consisting of `'AsScalar'` and either `true` or `false`. The default value is `true` if `S` is a scalar structure array, and `false` otherwise.

By default, `struct2dataset` converts a scalar structure array with *N* fields, each with *M* rows, into an *M*-by-*N* dataset array.

If instead you set `AsScalar` equal to `false` for a scalar input structure array, then `struct2dataset` converts `S` to a dataset array with *N* observations.

Output Arguments

`ds` — Output dataset array

dataset array

Output dataset array, returned by default with *M* observations and *N* variables.

- If `S` is a scalar structure array with *N* fields, each with *M* rows, then `ds` is an *M*-by-*N* dataset array.
- If `S` is a nonscalar *M*-by-1 structure array with *N* fields, then `ds` is an *M*-by-*N* dataset array.
- If `S` is a scalar structure array with *N* fields, each with *M* rows, and `AsScalar` is set equal to `false`, then `ds` is a dataset array with *N* observations.

See Also

`cell2dataset` | `dataset` | `dataset2struct`

Topics

“Create a Dataset Array from Workspace Variables” on page 2-57
“Create a Dataset Array from a File” on page 2-62
“Dataset Arrays” on page 2-112

Introduced in R2012b

surfht

Interactive contour plot

Syntax

```
surfht(z)
surfht(x,y,z)
```

Description

`surfht(z)` creates an interactive contour plot of the data in matrix `z`. `surfht` treats the values in `z` as the height above the plane.

`surfht(x,y,z)` creates an interactive contour plot of the data in matrix `z`, using the *x*-axis values contained in `x` and the *y*-axis values contained in `y`.

Examples

Create an Interactive Contour Plot

This example shows how to use `surfht` to create an interactive contour plot.

Create a grid of the (*x*,*y*) domain from (-2,-2) to (2,2) using `meshgrid`.

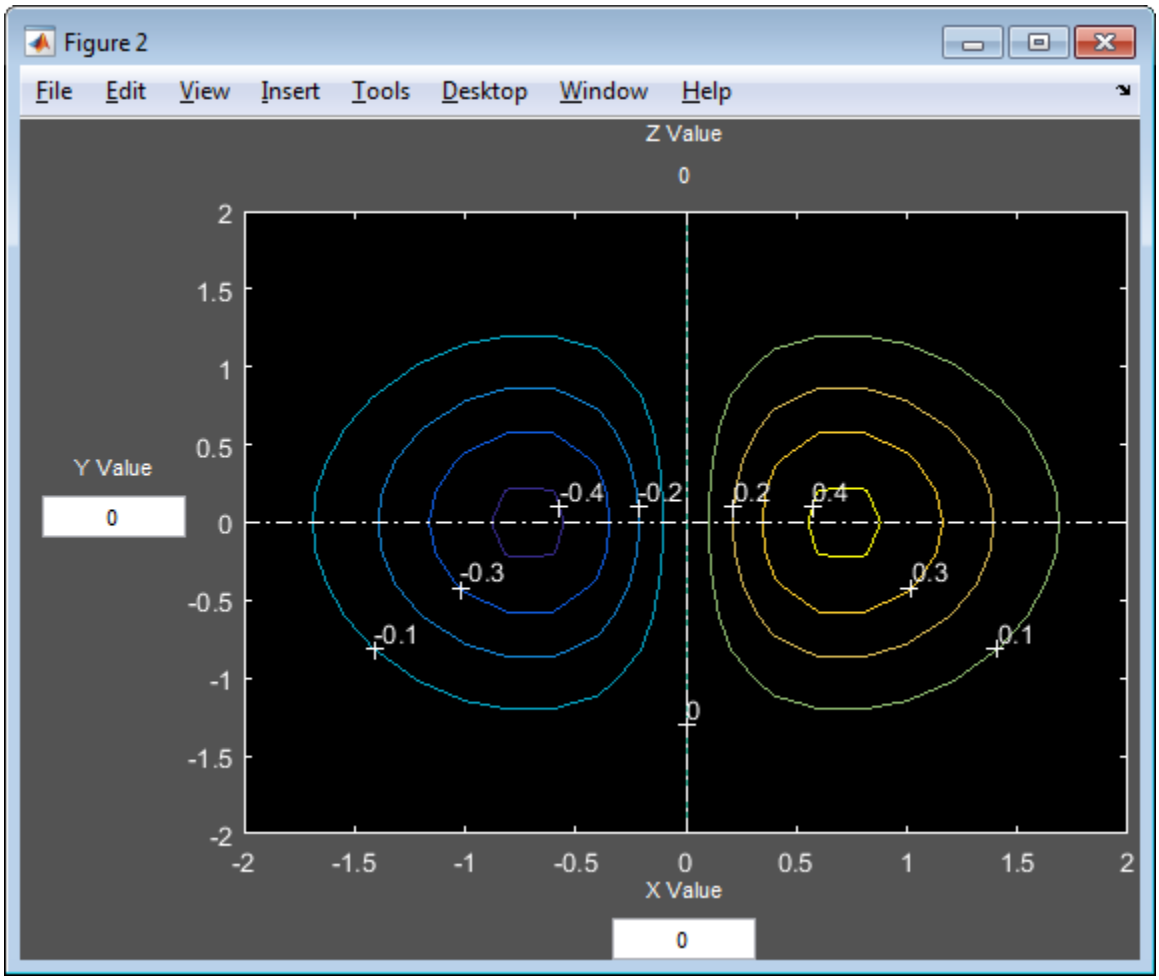
```
[x,y] = meshgrid(-2:0.2:2,-2:0.2:2);
```

Evaluate the function $z(x,y) = x \times \exp(-x^2 - y^2)$ over this domain.

```
z = x.*exp(-x.^2 - y.^2);
```

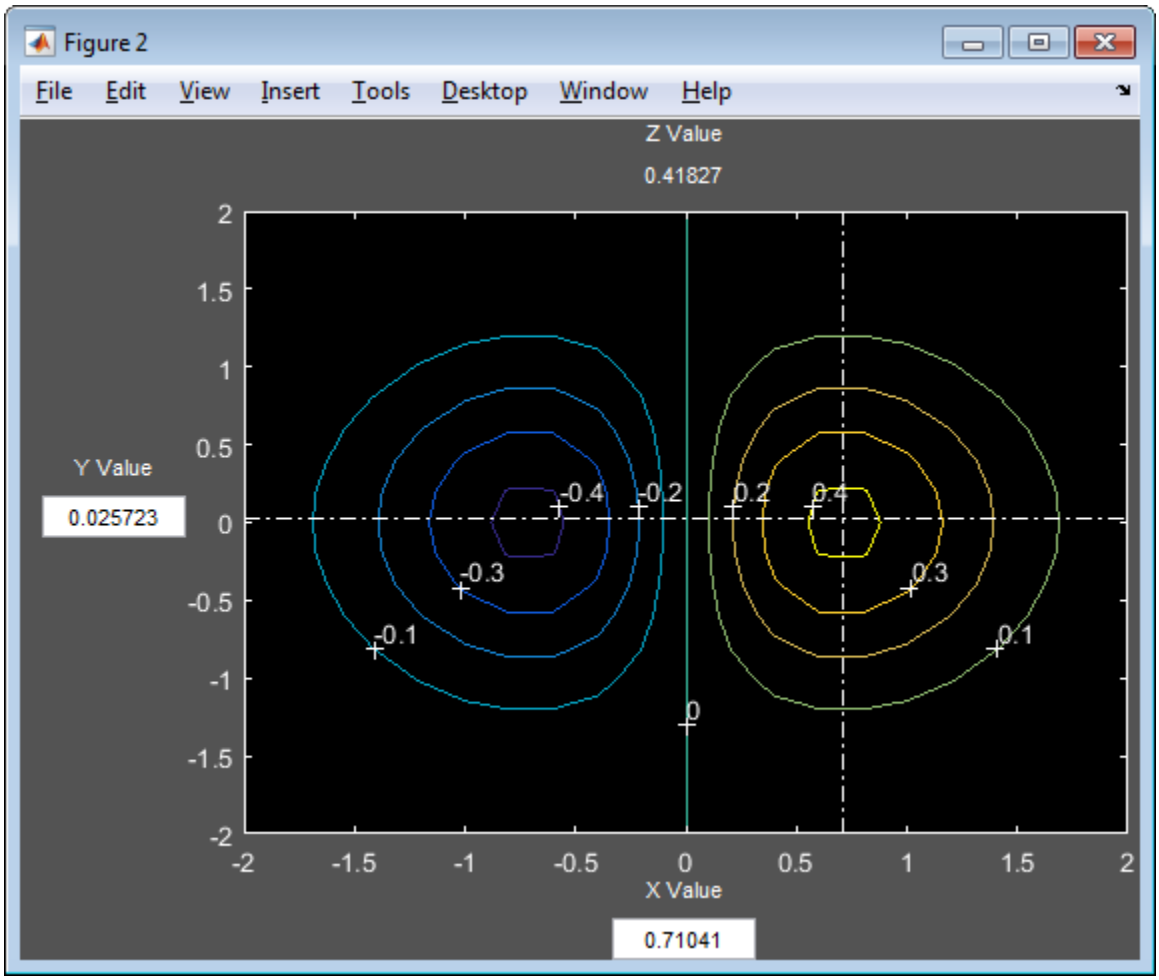
Open an interactive contour plot. Since `meshgrid` creates a grid of the *x* and *y* values, open the plot using the first row of *x* and the first column of *y*.

```
surfht(x(1,:),y(:,1),z)
```



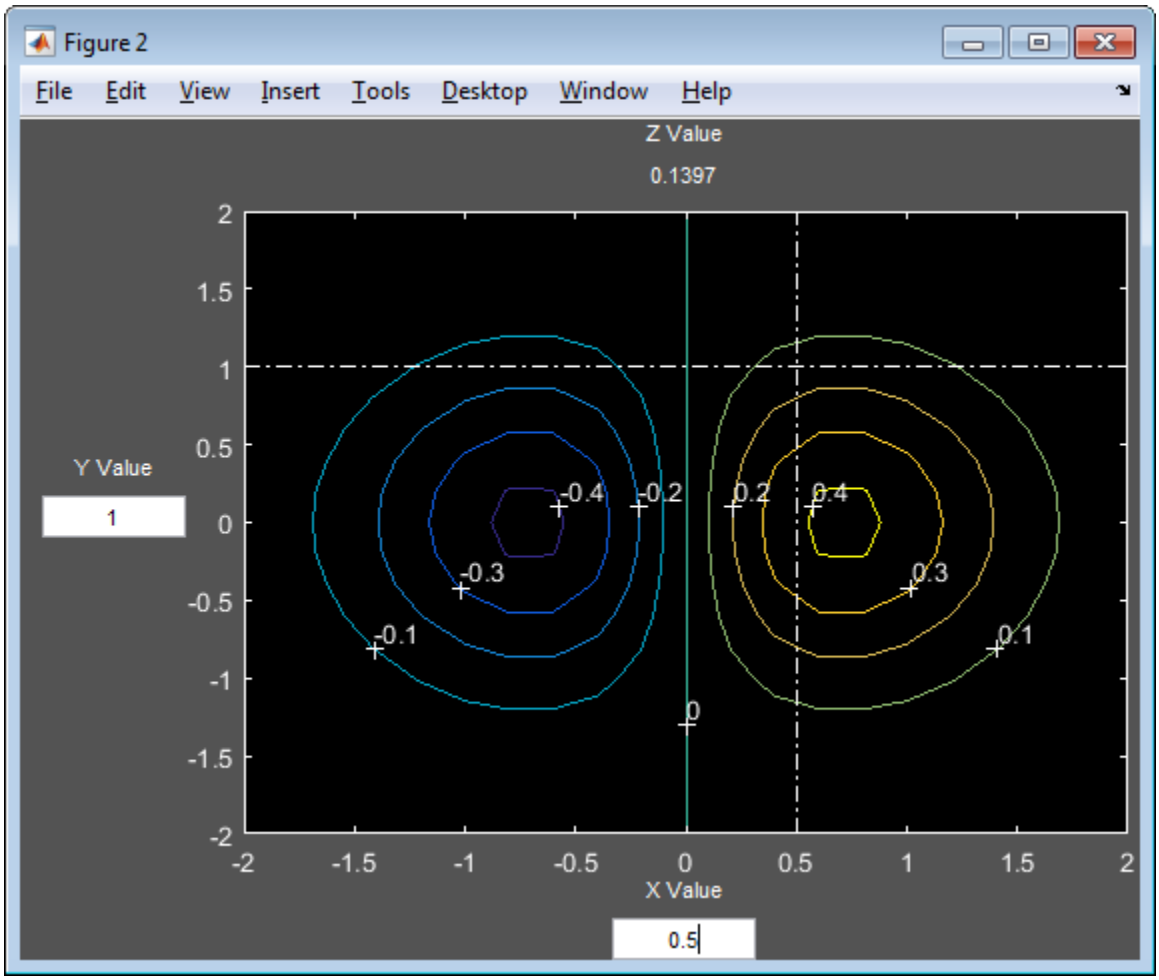
The figure shows a contour plot of the z values along the specified x- and y-axes.

Click the plot to evaluate z at the (x,y) coordinates indicated by the intersecting white lines.



For example, at $x = 0.71041$ and $y = 0.025723$, the value of z is 0.41827.

Alternatively, enter values in the fields labeled `X Value` and `Y Value` to evaluate z at the specified coordinates. For example, evaluate z at $x = 0.5$ and $y = 1$.



The value of *z* is 0.1397.

Input Arguments

z — *z*-axis values for contour plot

numeric matrix

z-axis values for contour plot, specified as a numeric matrix.

`surfht` treats the values in `z` as the height above the plane. By default, the *x*-axis values of the plot are the column indices of `z`, and the *y*-axis values of the plot are the row indices of `z`. To change the *x*- and *y*-axis values, specify `x` and `y`, respectively.

Data Types: `single` | `double`

x — *x*-axis values for contour plot

column indices of `z` (default) | numeric vector

x-axis values for contour plot, specified as a numeric vector. The length of `x` must match the number of columns in `z`.

Data Types: `single` | `double`

y — *y*-axis values for contour plot
row indices of `z` (default) | numeric vector

y-axis values for contour plot, specified as a numeric vector. The length of `y` must match the number of rows in `z`.

Data Types: `single` | `double`

Tips

- The intersection of the vertical and horizontal reference lines on the interactive plot defines the current *x* value and *y* value.
- Drag the dotted white reference lines to watch the interpolated *z* value (at the top of the plot) update simultaneously.
- Alternatively, obtain a specific interpolated *z* value by typing the *x* value and *y* value into editable text fields on the *x*-axis and *y*-axis, respectively.

See Also

`contour` | `meshgrid` | `surf`

Introduced before R2006a

survival

Calculate survival of Cox proportional hazards model

Syntax

```
s = survival(coxMdl)
s = survival(coxMdl,X)
s = survival(coxMdl,X,Stratification)
s = survival( ___ ,'Time',T)
[s,Tout] = survival( ___ )
```

Description

`s = survival(coxMdl)` estimates the baseline survival function of a Cox proportional hazards model `coxMdl`. The survival function at time `t` is the estimated probability of survival until time `t`. The term baseline refers to the survival function at the determined baseline of the predictors. This value is stored in `coxMdl.Baseline`, and the default value is the mean of the data set used for training.

`s = survival(coxMdl,X)` estimates the survival function when the predictors have the values in `X`. In this case, `s` is a column for each row of `X`.

`s = survival(coxMdl,X,Stratification)` estimates the survival function for the given value of the stratification variable `Stratification`. You must have one row in `Stratification` for each row in `X`.

Note When you train `coxMdl` using stratification variables and pass predictor variables `X`, `survival` also requires you to pass stratification variables.

`s = survival( ___ ,'Time',T)` computes the survival at times `T` using any of the input argument combinations in the previous syntaxes.

`[s,Tout] = survival( ___ )` also returns the times `Tout` at which each survival estimate is calculated.

Examples

Calculate Survival

Perform a Cox proportional hazards regression on the `lightbulb` data set, which contains simulated lifetimes of light bulbs. The first column of the light bulb data contains the lifetime (in hours) of two different types of bulbs. The second column contains a binary variable indicating whether the bulb is fluorescent or incandescent; 0 indicates the bulb is fluorescent, and 1 indicates it is incandescent. The third column contains the censoring information, where 0 indicates the bulb was observed until failure, and 1 indicates the observation was censored.

Fit a Cox proportional hazards model for the lifetime of the light bulbs, accounting for censoring. The predictor variable is the type of bulb.

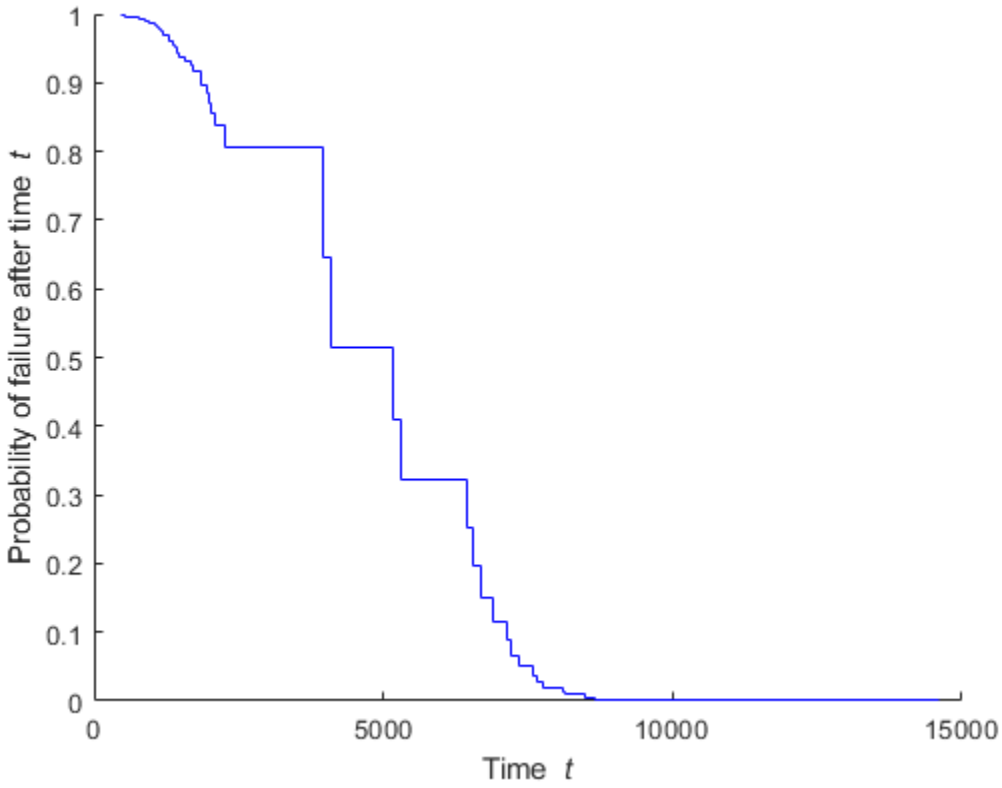
```
load lightbulb
coxMdl = fitcox(lightbulb(:,2),lightbulb(:,1), ...
    'Censoring',lightbulb(:,3));
```

Calculate the baseline survival function as a function of time `t`, meaning the probability that a light bulb fails after time `t`. By default, the baseline is calculated for the mean of the predictor, which in this case is `mean(lightbulb(:,2)) = 0.5`. Return the times `Tout` at which the survival function is calculated.

```
[s,Tout] = survival(coxMdl);
```

Plot the survival as a stairstep graph of time. (The times `Tout` are also in `coxMdl.Hazard(:,1)`.)

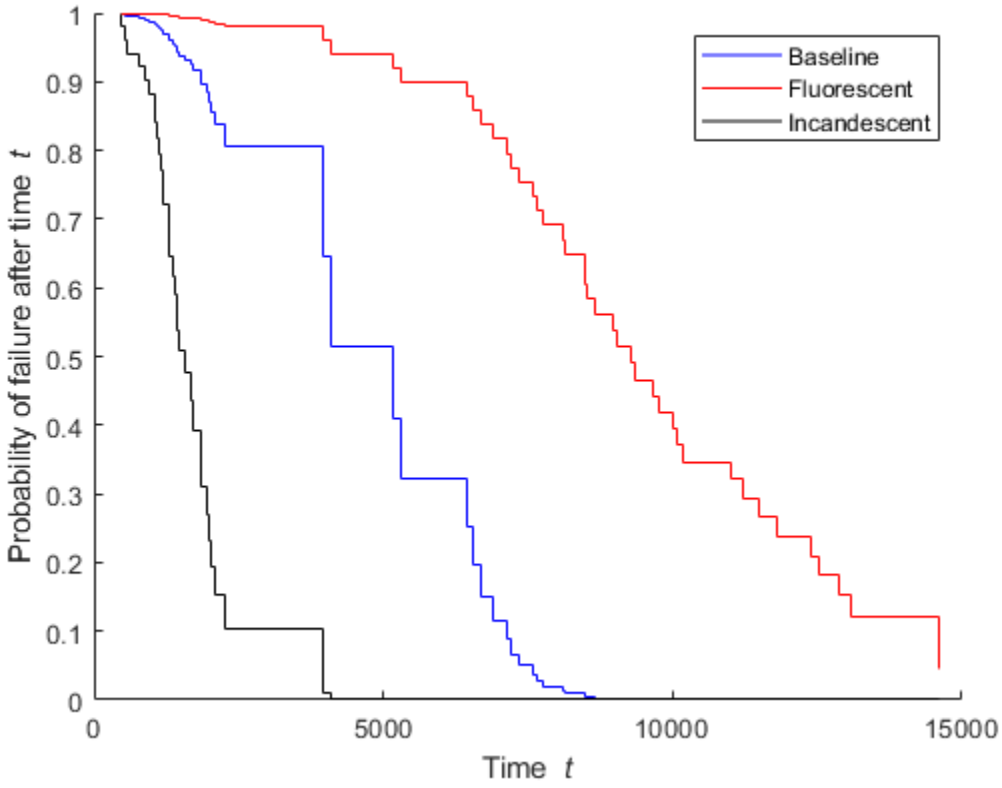
```
hold on;
stairs(Tout,s,'b-')
xlabel 'Time \it t'
ylabel 'Probability of failure after time \it t'
```



Overlay the plot with the survival functions for fluorescent and incandescent bulbs.

```
s_fluorescent = survival(coxMdl,0);
s_incandescent = survival(coxMdl,1);
stairs(Tout,s_fluorescent,'r-')
stairs(Tout,s_incandescent,'k-')
```

```
legend('Baseline','Fluorescent','Incandescent')
hold off
```



To create plots without first creating the survival data, use `plotSurvival`.

Survival for Stratified Model

Load the `coxModel` data. (This simulated data is generated in the example “Cox Proportional Hazards Model Object” on page 14-39.) The model named `coxMdl` has three stratification levels (1, 2, and 3) and a predictor `X` with three categorical values (1, 1/20, and 1/100).

```
load coxModel
```

Calculate the survival function for `X = 1` at the three stratification levels.

```
c1 = categorical(1);
X = [c1;c1;c1];
stratification = [1;2;3];
s = survival(coxMdl,X,stratification);
```

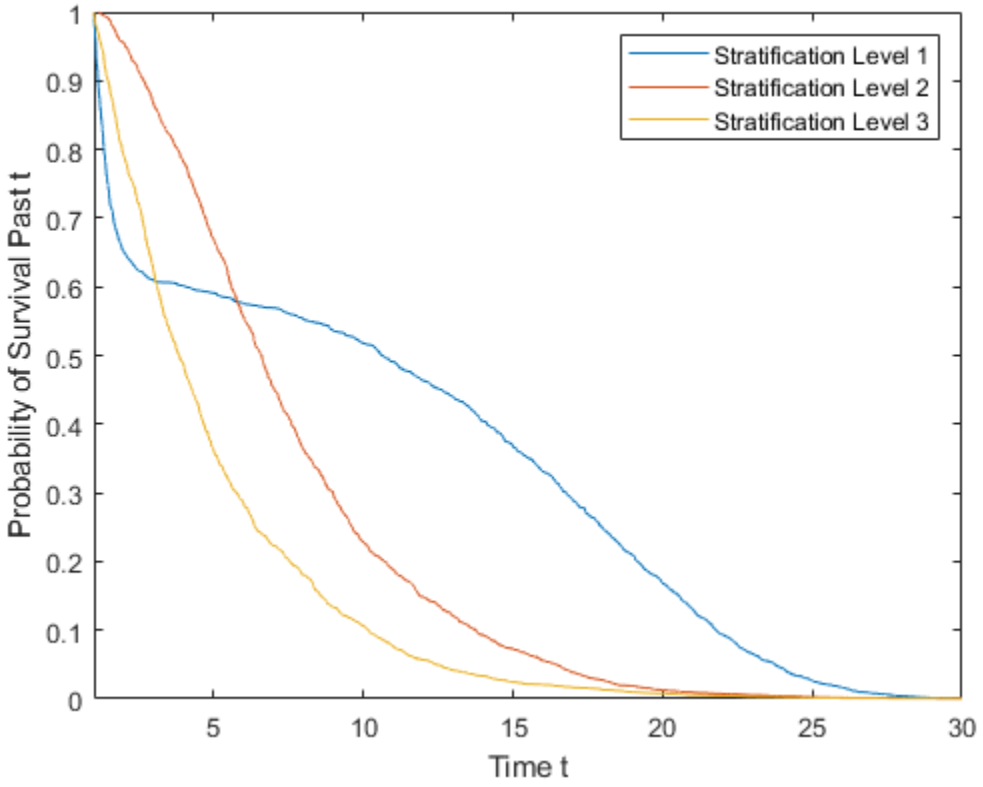
Plot the three survival functions. First, find the times for the three stratification levels.

```
t1 = find(coxMdl.Hazard(:,3) == 1);
t1 = coxMdl.Hazard(t1,1);
t2 = find(coxMdl.Hazard(:,3) == 2);
```

```
t2 = coxMdl.Hazard(t2,1);
t3 = find(coxMdl.Hazard(:,3) == 3);
t3 = coxMdl.Hazard(t3,1);
```

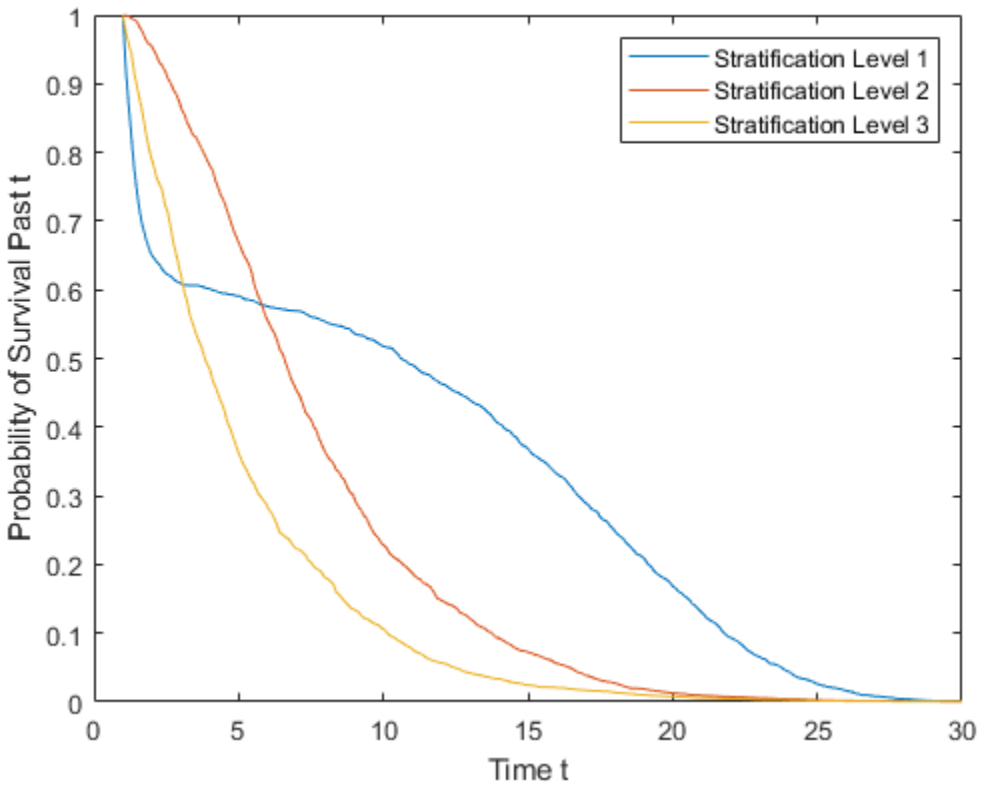
Plot the survival for the three levels. View the plot for times 1 through 30.

```
plot(t1,s{1},t2,s{2},t3,s{3})
xlim([1,30])
legend('Stratification Level 1','Stratification Level 2','Stratification Level 3','Location','no
xlabel('Time t')
ylabel('Probability of Survival Past t')
```



Alternatively, evaluate the survival for times 1 through 30 by specifying the `Time` argument.

```
t = linspace(1,30,300);
st = survival(coxMdl,X,stratification,'Time',t);
figure
plot(t,st{1},t,st{2},t,st{3})
legend('Stratification Level 1','Stratification Level 2','Stratification Level 3','Location','no
xlabel('Time t')
ylabel('Probability of Survival Past t')
```



Input Arguments

`coxMdl` — Fitted Cox proportional hazards model

`CoxModel` object

Fitted Cox proportional hazards model, specified as a `CoxModel` object. Create `coxMdl` using `fitcox`.

`X` — Predictors for model

mean of predictors used for training, but `0` for all categorical predictors (default) | array of predictors of type used for training

Predictors for the model, specified as an array of predictors of the same type used for training `coxMdl`. Each row of `X` represents one set of predictors.

Data Types: `double` | `table` | `categorical`

`Stratification` — Stratification level

variable or variables of type used for training

Stratification level, specified as a variable or variables of the same type used for training `coxMdl`. Specify the same number of rows in `Stratification` as in `X`.

Data Types: `single` | `double` | `logical` | `char` | `string` | `table` | `cell` | `categorical`

T — Times for survival estimates
`coxMdl.Hazard(:,1)` (default) | real vector

Times for survival estimates, specified as a real vector. `survival` sorts the specified times and converts them to a column vector, if necessary. The resulting values are linearly interpolated from times in the training data.

Example: `0:40`

Data Types: `double`

Output Arguments

s — Survival estimates
numeric column vector | cell array of numeric column vectors

Survival estimates, returned as a numeric column vector or a cell array of numeric column vectors.

- For a nonstratified model, `s` is a sorted numeric column vector of estimated probabilities.
- For a stratified model, `s` is a cell array of sorted numeric column vectors of the estimated probabilities for each stratification level.

`survival` returns a column of survival estimates for each row of `X`.

Tout — Times for survival estimates
`mdl.Hazard(:,1)` (default) | numeric column vector | cell array of numeric column vectors

Times for survival estimates, returned as one of the following.

- For a nonstratified model, `Tout` is a sorted numeric column vector of times in the training set.
- For a stratified model, `Tout` is a cell array of sorted numeric column vectors of the training times in the training set for each stratification level.

The `coxMdl.Hazard(:,1)` vector contains the times for both stratified and nonstratified models. For stratified models, the times for different stratification levels are separated by a 0 entry.

Data Types: `double` | `cell`

See Also

`CoxModel` | `fitcox` | `hazardratio` | `plotSurvival`

Topics
“Cox Proportional Hazards Model Object” on page 14-39

Introduced in R2021a

table2dataset

(Not Recommended) Convert table to dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = table2dataset(t)
```

Description

`ds = table2dataset(t)` converts a table to a dataset array.

Examples

Convert a Table to a Dataset Array

Load the sample data, which contains nutritional information for 77 cereals.

```
load cereal;
```

Create a table containing the calorie, protein, fat, and name data for the first five cereals. Label the variables.

```
Calories = Calories(1:5);
Protein = Protein(1:5);
Fat = Fat(1:5);
Name = Name(1:5);

cereal = table(Calories,Protein,Fat,'RowNames',Name)
```

```
cereal=5×3 table
                              Calories    Protein    Fat
                              ________    _______    ___

    100% Bran                    70          4        1
    100% Natural Bran           120          3        5
    All-Bran                     70          4        1
    All-Bran with Extra Fiber    50          4        0
    Almond Delight              110          2        2
```

Convert the table to a dataset array.

```
ds = table2dataset(cereal)
```

```
ds =
                              Calories    Protein    Fat
    100% Bran                  70          4         1
```

```
100% Natural Bran            120         3          5
All-Bran                      70         4          1
All-Bran with Extra Fiber     50         4          0
Almond Delight               110         2          2
```

Input Arguments

`t` — Input table

table

Input table to convert to a dataset array, specified as a table. Each variable in `t` becomes a variable in the output dataset array `ds`.

Data Types: `table`

Output Arguments

`ds` — Output dataset array

dataset array

Output dataset array, returned as a dataset array containing the variables from the input table `t`.

See Also

`dataset` | `table`

Topics

“Dataset Arrays” on page 2-112
“Tables”

Introduced in R2013b

tabulate

Frequency table

Syntax

```
tabulate(x)
tbl = tabulate(x)
```

Description

`tabulate(x)` displays a frequency table of the data in the vector `x`. For each unique value in `x`, the `tabulate` function shows the number of instances and percentage of that value in `x`. See `tbl`.

`tbl = tabulate(x)` returns the frequency table `tbl` as a numeric matrix when `x` is numeric and as a cell array otherwise.

Examples

Tabulate Data Vector

Create a frequency table for a vector of data.

Load the `patients` data set. Display the first five entries of the `Gender` variable. Each value indicates the gender of a patient.

```
load patients
Gender(1:5)
```

```
ans = 5x1 cell
    {'Male'  }
    {'Male'  }
    {'Female'}
    {'Female'}
    {'Female'}
```

Generate a frequency table that shows the number and percentage of `Male` and `Female` patients in the data set.

```
tabulate(Gender)
```

```
   Value    Count   Percent
    Male       47     47.00%
  Female       53     53.00%
```

Tabulate Positive Integer Vector

Create a frequency table for a vector of positive integers. By default, if a vector `x` contains only positive integers, then `tabulate` returns 0 counts for the integers between `1` and `max(x)` that do not

appear in `x`. To avoid this behavior, convert the vector `x` to a `categorical` vector before calling `tabulate`.

Load the `patients` data set. Display the first five entries of the `Height` variable. Each value indicates the height, in inches, of a patient.

```
load patients
Height(1:5)
```

```
ans = 5×1

    71
    69
    64
    67
    64
```

Create a frequency table that shows, in its second and third columns, the number and percentage of patients in the data set that have a particular height. Display the first five entries and the last five entries of the matrix that `tabulate` returns. `tbl` contains one row for each height between `1` and `72` inches, where `72` is the maximum height value in `Height`.

```
tbl = tabulate(Height);
first = tbl(1:5,:)
```

```
first = 5×3

     1     0     0
     2     0     0
     3     0     0
     4     0     0
     5     0     0
```

```
last = tbl(end-4:end,:)
```

```
last = 5×3

    68    15    15
    69     8     8
    70    11    11
    71    10    10
    72     4     4
```

Generate a frequency table that shows `Count` and `Percent` values only for heights that appear in the `Height` variable. Convert `Height` to a `categorical` variable, and then call the `tabulate` function.

```
newHeight = categorical(Height);
tabulate(newHeight)
```

```
  Value    Count   Percent
     60        1      1.00%
     62        3      3.00%
     63        7      7.00%
     64       12     12.00%
     65        8      8.00%
```

```
    66       15     15.00%
    67        6      6.00%
    68       15     15.00%
    69        8      8.00%
    70       11     11.00%
    71       10     10.00%
    72        4      4.00%
```

Create Table Array from Tabulated Data

Create a frequency table from a character array by using `tabulate`. Convert the resulting cell array to a `table` array, and visualize the results.

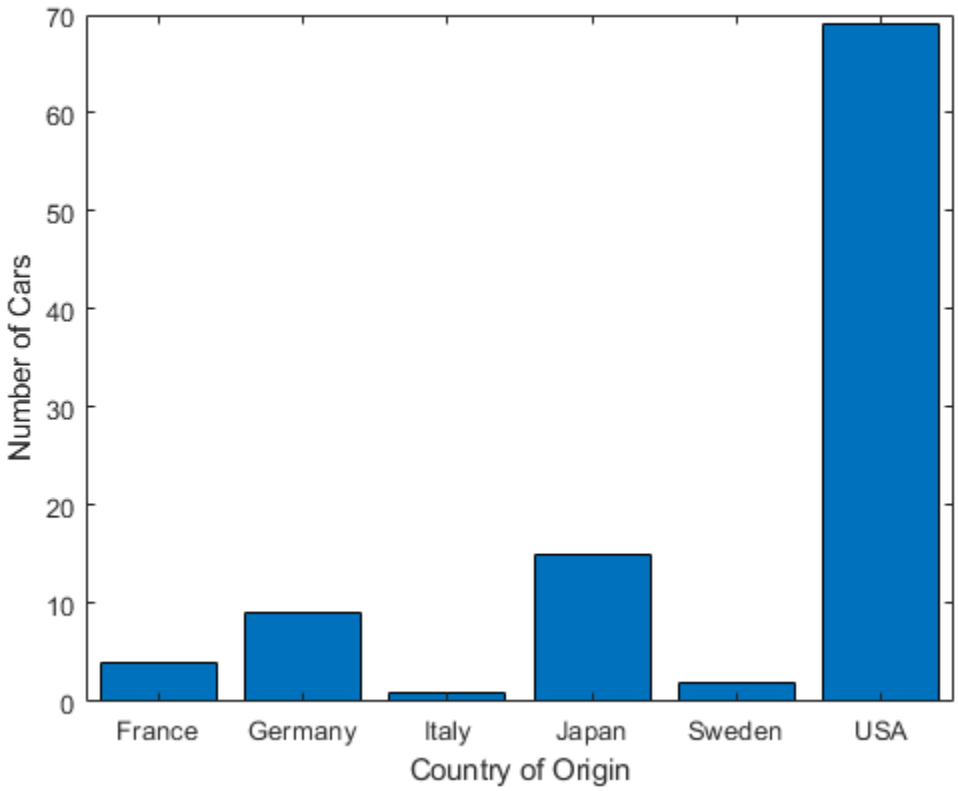
Load the `carsmall` data set. Tabulate the data in the `Origin` variable, which shows the country of origin of each car in the data set. Convert the resulting cell array `tbl` to a `table` array `t`. Change the `Value` column to a `categorical` vector.

```
load carsmall
tbl = tabulate(Origin);
t = cell2table(tbl,'VariableNames', ...
    {'Value','Count','Percent'});
t.Value = categorical(t.Value)
```

```
t=6×3 table
     Value     Count    Percent
    _______    _____    _______

    USA         69        69
    France       4         4
    Japan       15        15
    Germany      9         9
    Sweden       2         2
    Italy        1         1
```

Create a bar graph from the frequency table.

```
bar(t.Value,t.Count)
xlabel('Country of Origin')
ylabel('Number of Cars')
```



Tabulate Data with Missing Values

Create a frequency table from a numeric vector with `NaN` values.

Load the `carsmall` data set. The `MPG` variable contains the miles per gallon measurement of 100 cars. For six of the cars, the `MPG` value is missing (`NaN`).

```
load carsmall
numcars = length(MPG)
```

```
numcars = 100
```

```
nanindex = isnan(MPG);
numMissingMPG = length(MPG(nanindex))
```

```
numMissingMPG = 6
```

Create a frequency table using `MPG`. Convert the matrix output from `tabulate` to a table, and label the table columns.

```
tbl = tabulate(MPG);
t = array2table(tbl,'VariableNames', ...
    {'Value','Count','Percent'})
```

```
t=37×3 table
    Value    Count    Percent
```

```
_____    _____    _______

   9        1      1.0638
  10        2      2.1277
  11        1      1.0638
  13        4      4.2553
  14        5      5.3191
14.5        1      1.0638
  15        5      5.3191
15.5        1      1.0638
  16        2      2.1277
16.5        2      2.1277
  17        1      1.0638
17.5        2      2.1277
  18        4      4.2553
18.5        1      1.0638
  19        2      2.1277
  20        2      2.1277
   ⋮
```

The frequency table displays data only for the 94 cars with numeric `MPG` values. `tabulate` calculates the percentage of each `MPG` value in this subset of cars, not the entire set of 100 cars.

```
tnumcars = sum(t.Count)
```

```
tnumcars = 94
```

Input Arguments

x — Input data

numeric vector | logical vector | categorical vector | character array | string array | cell array of character vectors

Input data, specified as a numeric vector, logical vector, categorical vector, character array, string array, or cell array of character vectors.

- If `x` is a numeric vector, then `tbl` is a numeric matrix.
- If `x` is a logical vector, categorical vector, character array, string array, or cell array of character vectors, then `tbl` is a cell array.

Note If the elements of `x` are positive integers, then the frequency table includes `0` counts for the integers between `1` and `max(x)` that do not appear in `x`. For an example, see “Tabulate Positive Integer Vector” on page 33-6071.

Data Types: `single` | `double` | `logical` | `categorical` | `char` | `string` | `cell`

Output Arguments

tbl — Frequency table

numeric matrix | cell array

Frequency table, returned as a numeric matrix or cell array. `tbl` includes the following information.

Column	Description
1st column (Value)	Unique values of `x`
2nd column (Count)	Number of instances of each value
3rd column (Percent)	Percentage of each value

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

See Also

`bar` | `histogram` | `pareto`

Topics

“Grouping Variables” on page 2-45

Introduced before R2006a

tblread

Read tabular data from file

Syntax

```
[data,varnames,casenames] = tblread
[data,varnames,casenames] = tblread(filename)
[data,varnames,casenames] = tblread(filename,delimiter)
```

Description

`[data,varnames,casenames] = tblread` displays the File Open dialog box for interactive selection of a tabular data file. The file format has variable names in the first row, case names in the first column and data starting in the (2, 2) position. Outputs are:

- `data` — Numeric matrix with a value for each variable-case pair
- `varnames` — Character matrix containing the variable names in the first row of the file
- `casenames` — Character matrix containing the names of each case in the first column of the file

`[data,varnames,casenames] = tblread(filename)` allows command line specification of the name of a file in the current folder, or the complete path name of any file, using the character vector or string scalar *`filename`*.

`[data,varnames,casenames] = tblread(filename,delimiter)` reads from the file using *`delimiter`* as the delimiting character. Accepted values for *`delimiter`* are:

- `' '` or `'space'`
- `'\t'` or `'tab'`
- `','` or `'comma'`
- `';'` or `'semi'`
- `'|'` or `'bar'`

The default value of *`delimiter`* is `'space'`.

Examples

```
[data,varnames,casenames] = tblread('sat.dat')
data =
  470  530
  520  480

varnames =
Male
Female

casenames =
Verbal
Quantitative
```

See Also

`caseread` | `readtable` | `tblwrite` | `tdfread`

Introduced before R2006a

tblwrite

Write tabular data to file

Syntax

```
tblwrite(data,varnames,casenames)
tblwrite(data,varnames,casenames,filename)
tblwrite(data,varnames,casenames,filename,delimiter)
```

Description

`tblwrite(data,varnames,casenames)` displays the **File Open** dialog box for interactive specification of the tabular data output file. The file format has variable names in the first row, case names in the first column and `data` starting in the (2,2) position.

`varnames` is a character matrix or string array containing the variable names. `casenames` is a character matrix or string array containing the names of each case in the first column. `data` is a character matrix or string array with a value for each variable-case pair.

`tblwrite(data,varnames,casenames,filename)` specifies a file in the current folder, or the complete path name of any file in the character vector or string scalar *`filename`*.

`tblwrite(data,varnames,casenames,filename,delimiter)` writes to the file using *`delimiter`* as the delimiting character. The following table lists the accepted values for *`delimiter`* and their equivalent names.

Value	Name
`' '`	`'space'`
`'\t'`	`'tab'`
`','`	`'comma'`
`';'`	`'semi'`
`'|'`	`'bar'`

The default value of *`delimiter`* is `'space'`.

Examples

Continuing the example from `tblread`:

```
tblwrite(data,varnames,casenames,'sattest.dat')
type sattest.dat
                     Male  Female
Verbal               470     530
Quantitative        520     480
```

See Also

`casewrite` | `tblread` | `writetable`

Introduced before R2006a

tcdf

Student's *t* cumulative distribution function

Syntax

```
p = tcdf(x,nu)

p = tcdf(x,nu,'upper')
```

Description

`p = tcdf(x,nu)` returns the cumulative distribution function (cdf) of the Student's *t* distribution with `nu` degrees of freedom, evaluated at the values in `x`.

`p = tcdf(x,nu,'upper')` returns the complement of the cdf, evaluated at the values in `x` with `nu` degrees of freedom, using an algorithm that more accurately computes the extreme upper-tail probabilities than subtracting the lower tail value from 1.

Examples

Compute Student's *t* Distribution cdf

Generate a random sample of size `100` from a normally distributed population with mean `1` and standard deviation `2`.

```
rng default   % For reproducibility
mu = 1;
n = 100;
sigma = 2;
x = normrnd(mu,sigma,n,1);
```

Compute the sample mean, sample standard deviation, and *t*-score of the sample.

```
xbar = mean(x);
s = std(x);
t = (xbar-mu)/(s/sqrt(n))
```

```
t = 1.0589
```

Use `tcdf` to compute the probability of a sample of size `100` having a larger *t*-score than the *t*-score of the sample.

```
p = 1-tcdf(t,n-1)
```

```
p = 0.1461
```

This probability is the same as the *p* value returned by a *t* test with null hypothesis that the sample comes from a normal population with mean `1` and alternative hypothesis that the mean is greater than `1`.

```
[h,ptest] = ttest(x,mu,0.05,'right');
ptest
```

```
ptest = 0.1461
```

Compute Complementary cdf (Tail Distribution)

Determine the probability that an observation from the Student's *t* distribution with degrees of freedom `99` falls on the interval `[10 Inf]`.

```
p1 = 1 - tcdf(10,99)
```

```
p1 = 0
```

`tcdf(10,99)` is nearly `1`, so `p1` becomes `0`. Specify `'upper'` so that `tcdf` computes the extreme upper-tail probabilities more accurately.

```
p2 = tcdf(10,99,'upper')
```

```
p2 = 5.4699e-17
```

You can also use `'upper'` to compute a right-tailed *p*-value.

Input Arguments

x — Values at which to evaluate cdf

scalar value | array of scalar values

Values at which to evaluate the cdf, specified as a scalar value or an array of scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `tcdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `p` is the cdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[-1,0,3,4]`

Data Types: `single` | `double`

nu — Degrees of freedom

positive scalar value | array of positive scalar values

Degrees of freedom for the Student's *t* distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the cdf at multiple values, specify `x` using an array.
- To evaluate the cdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `tcdf` expands each scalar input into a constant array of the same size as the array

inputs. Each element in p is the cdf value of the distribution specified by the corresponding element in nu, evaluated at the corresponding element in x.

Example: `[9,19,49,99]`

Data Types: `single` | `double`

Output Arguments

p — cdf values

scalar value | array of scalar values

cdf values evaluated at the values in x, returned as a scalar value or an array of scalar values. p is the same size as x and nu after any necessary scalar expansion. Each element in p is the cdf value of the distribution specified by the corresponding element in nu, evaluated at the corresponding element in x.

More About

Student's *t* cdf

The Student's *t* distribution is a one-parameter family of curves. The parameter ν is the degrees of freedom. The Student's *t* distribution has zero mean.

The cdf of the Student's *t* distribution is

$$p = F(x|\nu) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} dt,$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result *p* is the probability that a single observation from the *t* distribution with ν degrees of freedom falls in the interval $[-\infty, x]$.

For more information, see "Student's t Distribution" on page B-149.

Alternative Functionality

- `tcdf` is a function specific to the Student's *t* distribution. Statistics and Machine Learning Toolbox also offers the generic function `cdf`, which supports various probability distributions. To use `cdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `tcdf` is faster than the generic function `cdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `tinv` | `tpdf` | `trnd` | `tstat` | `ttest` | `ttest2`

Topics

“Student's t Distribution” on page B-149

Introduced before R2006a

tdfread

Read tab-delimited file

Syntax

```
tdfread
tdfread(filename)
tdfread(filename,delimiter)

s = tdfread( ___ )
```

Description

`tdfread` opens the Select File to Open dialog box for interactive selection of a data file, and reads the data from the file you select. `tdfread` can read data from tab-delimited text files with `.txt`, `.dat`, or `.csv` file extensions.

Select a file that has variable names in the first row and values separated by tabs in the remaining rows. `tdfread` creates a variable in the workspace for each column of the file, and names each variable according to its first row value.

- If a column contains only numeric data in all rows except the first, then `tdfread` creates a `double` variable.
- Otherwise, `tdfread` creates a `char` variable.

After importing all values, `tdfread` displays information about the imported variables, such as their size, bytes, and class.

`tdfread(filename)` creates variables from the data in `filename`, which is either the name of a file in the current folder or the complete path name of a file.

`tdfread(filename,delimiter)` indicates that the character specified by `delimiter` separates values in the file.

`s = tdfread( ___ )` returns a structure `s` in which each field contains a variable. Specify any of the input argument combinations in the previous syntaxes.

Examples

Create Workspace Variables from Text File

Display the contents of the `sat2.dat` file. Note that the first row of the file contains the variable names.

```
type sat2.dat
```

```
Test,Gender,Score
Verbal,Male,470
Verbal,Female,530
```

```
Quantitative,Male,520
Quantitative,Female,480
```

In the workspace, create the variables `Gender`, `Score`, and `Test` from the columns of the file. Because commas separate the values in the file, specify `','` as the delimiter.

```
tdfread('sat2.dat',',')
```

```
  Name        Size            Bytes  Class     Attributes

  Gender      4x6                48  char
  Score       4x1                32  double
  Test        4x12               96  char
```

Input Arguments

`filename` — Name of file to read

character vector | string scalar

Name of the file to read, specified as a character vector or string scalar.

Depending on the location of the file, `filename` has one of these forms.

Location of File	Form
Current folder or folder on the MATLAB path	Specify the name of the file in `filename`. **Example:** `'myTextFile.txt'`
Folder that is not the current folder or a folder on the MATLAB path	Specify the full or relative path name in `filename`. **Example:** `'C:\myFolder\myTextFile.txt'`

Example: `'sat2.dat'`

Data Types: `char` | `string`

`delimiter` — Delimiter character

`'\t'` or `'tab'` (default) | `'bar'` | `'comma'` | `'semi'` | `'space'` | ...

Delimiter character, specified as one of the values in this table.

Value	Description

`'|'`
 	
`'bar'`
	Vertical bar

`','`
 	
`'comma'`
	Comma

`';'`
 	
`'semi'`
	Semicolon

Value	Description

`' '`
 	
`'space'`
	Space

`'\t'`
 	
`'tab'`
	Tab

Example: `','`

Data Types: `char` | `string`

Alternative Functionality

Consider using the `readtable`, `readmatrix`, or `readcell` MATLAB functions to import data. These functions provide more flexible data importing options than `tdfread`.

See Also

Import Tool | `readcell` | `readmatrix` | `readtable` | `textscan`

Introduced before R2006a

templateDiscriminant

Discriminant analysis classifier template

Syntax

```
t = templateDiscriminant()
t = templateDiscriminant(Name,Value)
```

Description

`t = templateDiscriminant()` returns a discriminant analysis learner template suitable for training ensembles or error-correcting output code (ECOC) multiclass models.

If you specify a default template, then the software uses default values for all input arguments during training.

Specify `t` as a learner in `fitcensemble` or `fitcecoc`.

`t = templateDiscriminant(Name,Value)` creates a template with additional options specified by one or more name-value pair arguments.

For example, you can specify the discriminant type or the regularization parameter.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a Discriminant Analysis Template for Ensemble Learning

Create a nondefault discriminant analysis template for use in `fitcensemble`.

Load Fisher's iris data set.

```
load fisheriris
```

Create a template for pseudolinear discriminant analysis.

```
t = templateDiscriminant('DiscrimType','pseudoLinear')
```

```
t =
Fit template for classification Discriminant.

    DiscrimType: 'pseudoLinear'
          Gamma: []
          Delta: []
     FillCoeffs: []
     SaveMemory: []
        Version: 1
         Method: 'Discriminant'
```

```
            Type: 'classification'
```

All properties of the template object are empty except for `DiscrimType`, `Method`, and `Type`. When trained on, the software fills in the empty properties with their respective default values.

Specify `t` as a weak learner for a classification ensemble.

```
Mdl = fitcensemble(meas,species,'Method','Subspace','Learners',t);
```

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl)
```

```
L = 0.0400
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'DiscrimType','pseudoLinear','SaveMemory','on'` specifies a template for pseudolinear discriminant analysis that does not store the full covariance matrix.

`Delta` — Linear coefficient threshold
`0` (default) | nonnegative scalar value

Linear coefficient threshold, specified as the comma-separated pair consisting of `'Delta'` and a nonnegative scalar value. If a coefficient of `Mdl` has magnitude smaller than `Delta`, `Mdl` sets this coefficient to `0`, and you can eliminate the corresponding predictor from the model. Set `Delta` to a higher value to eliminate more predictors.

`Delta` must be `0` for quadratic discriminant models.

Data Types: `single` | `double`

`DiscrimType` — Discriminant type
`'linear'` (default) | `'quadratic'` | `'diaglinear'` | `'diagquadratic'` | `'pseudolinear'` | `'pseudoquadratic'`

Discriminant type, specified as the comma-separated pair consisting of `'DiscrimType'` and a character vector or string scalar in this table.

Value	Description	Predictor Covariance Treatment
`'linear'`	Regularized linear discriminant analysis (LDA)	• All classes have the same covariance matrix. • $\widehat{\Sigma}_\gamma = (1-\gamma)\widehat{\Sigma} + \gamma \mathrm{diag}\left(\widehat{\Sigma}\right)$. $\widehat{\Sigma}$ is the empirical, pooled covariance matrix and γ is the amount of regularization.
`'diaglinear'`	LDA	All classes have the same, diagonal covariance matrix.
`'pseudolinear'`	LDA	All classes have the same covariance matrix. The software inverts the covariance matrix using the pseudo inverse.
`'quadratic'`	Quadratic discriminant analysis (QDA)	The covariance matrices can vary among classes.
`'diagquadratic'`	QDA	The covariance matrices are diagonal and can vary among classes.
`'pseudoquadratic'`	QDA	The covariance matrices can vary among classes. The software inverts the covariance matrix using the pseudo inverse.

Note To use regularization, you must specify `'linear'`. To specify the amount of regularization, use the `Gamma` name-value pair argument.

Example: `'DiscrimType','quadratic'`

`FillCoeffs` — `Coeffs` property flag

`'on'` | `'off'`

`Coeffs` property flag, specified as the comma-separated pair consisting of `'FillCoeffs'` and `'on'` or `'off'`. Setting the flag to `'on'` populates the `Coeffs` property in the classifier object. This can be computationally intensive, especially when cross-validating. The default is `'on'`, unless you specify a cross-validation name-value pair, in which case the flag is set to `'off'` by default.

Example: `'FillCoeffs','off'`

`Gamma` — Amount of regularization

scalar value in the interval [0,1]

Amount of regularization to apply when estimating the covariance matrix of the predictors, specified as the comma-separated pair consisting of `'Gamma'` and a scalar value in the interval [0,1]. `Gamma` provides finer control over the covariance matrix structure than `DiscrimType`.

- If you specify `0`, then the software does not use regularization to adjust the covariance matrix. That is, the software estimates and uses the unrestricted, empirical covariance matrix.

  - For linear discriminant analysis, if the empirical covariance matrix is singular, then the software automatically applies the minimal regularization required to invert the covariance matrix. You can display the chosen regularization amount by entering `Mdl.Gamma` at the command line.
  - For quadratic discriminant analysis, if at least one class has an empirical covariance matrix that is singular, then the software throws an error.
- If you specify a value in the interval (0,1), then you must implement linear discriminant analysis, otherwise the software throws an error. Consequently, the software sets `DiscrimType` to `'linear'`.
- If you specify `1`, then the software uses maximum regularization for covariance matrix estimation. That is, the software restricts the covariance matrix to be diagonal. Alternatively, you can set `DiscrimType` to `'diagLinear'` or `'diagQuadratic'` for diagonal covariance matrices.

Example: `'Gamma',1`

Data Types: `single` | `double`

`SaveMemory` — Flag to save covariance matrix
`'off'` (default) | `'on'`

Flag to save covariance matrix, specified as the comma-separated pair consisting of `'SaveMemory'` and either `'on'` or `'off'`. If you specify `'on'`, then `fitcdiscr` does not store the full covariance matrix, but instead stores enough information to compute the matrix. The `predict` method computes the full covariance matrix for prediction, and does not store the matrix. If you specify `'off'`, then `fitcdiscr` computes and stores the full covariance matrix in `Mdl`.

Specify `SaveMemory` as `'on'` when the input matrix contains thousands of predictors.

Example: `'SaveMemory','on'`

Output Arguments

`t` — Discriminant analysis classification template
template object

Discriminant analysis classification template suitable for training ensembles or error-correcting output code (ECOC) multiclass models, returned as a template object. Pass `t` to `fitcensemble` or `fitcecoc` to specify how to create the discriminant analysis classifier for the ensemble or ECOC model, respectively.

If you display `t` to the Command Window, then all unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

See Also

`ClassificationDiscriminant` | `fitcecoc` | `fitcensemble` | `predict`

Introduced in R2014a

templateECOC

Error-correcting output codes learner template

Syntax

```
t = templateECOC()
t = templateECOC(Name,Value)
```

Description

`t = templateECOC()` returns an error-correcting output codes (ECOC) classification learner template.

If you specify a default template, then the software uses default values for all input arguments during training.

`t = templateECOC(Name,Value)` returns a template with additional options specified by one or more name-value pair arguments.

For example, you can specify a coding design, whether to fit posterior probabilities, or the types of binary learners.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a Default ECOC Classification Learner Template

Use `templateECOC` to create a default ECOC template.

```
t = templateECOC()
```

```
t =
Fit template for classification ECOC.

    BinaryLearners: ''
           Coding: ''
      FitPosterior: []
           Options: []
    VerbosityLevel: []
     NumConcurrent: []
           Version: 1
            Method: 'ECOC'
              Type: 'classification'
```

All properties of the template object are empty except for `Method` and `Type`. When you pass `t` to `testckfold`, the software fills in the empty properties with their respective default values. For

example, the software fills the `BinaryLearners` property with `'SVM'`. For details on other default values, see `fitcecoc`.

`t` is a plan for an ECOC learner. When you create it, no computation occurs. You can pass `t` to `testckfold` to specify a plan for an ECOC classification model to statistically compare with another model.

Statistically Compare Performance of Two ECOC Classification Models

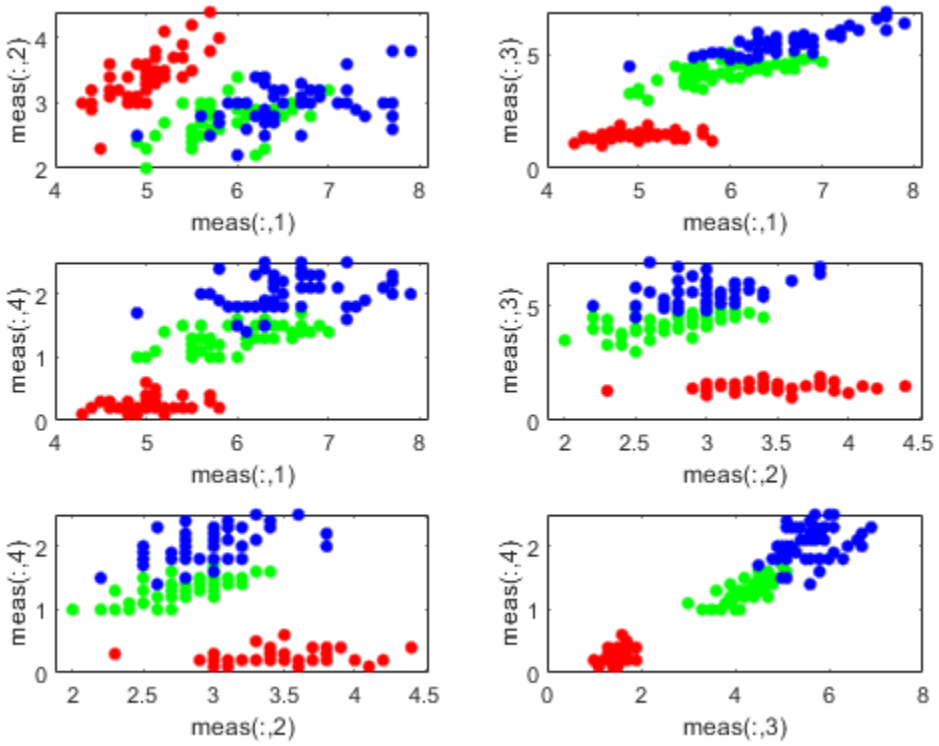
One way to select predictors or features is to train two models where one that uses a subset of the predictors that trained the other. Statistically compare the predictive performances of the models. If there is sufficient evidence that model trained on fewer predictors performs better than the model trained using more of the predictors, then you can proceed with a more efficient model.

Load Fisher's iris data set. Plot all 2-dimensional combinations of predictors.

```
load fisheriris
d = size(meas,2); % Number of predictors
pairs = nchoosek(1:d,2)
```

```
pairs = 6×2

     1     2
     1     3
     1     4
     2     3
     2     4
     3     4
```

```
for j = 1:size(pairs,1)
    subplot(3,2,j)
    gscatter(meas(:,pairs(j,1)),meas(:,pairs(j,2)),species)
    xlabel(sprintf('meas(:,%d)',pairs(j,1)))
    ylabel(sprintf('meas(:,%d)',pairs(j,2)))
    legend off
end
```



Based on the scatterplot, `meas(:,3)` and `meas(:,4)` seem like they separate the groups well.

Create an ECOC template. Specify to use a one-versus-all coding design.

```
t = templateECOC('Coding','onevsall');
```

By default, the ECOC model uses linear SVM binary learners. You can choose other, supported algorithms by specifying them using the `'Learners'` name-value pair argument.

Test whether an ECOC model that is just trained using predictors 3 and 4 performs at most as well as an ECOC model that is trained using all predictors. Rejecting this null hypothesis means that the ECOC model trained using predictors 3 and 4 performs better than the ECOC model trained using all predictors. Suppose C_1 represents the classification error of the ECOC model trained using predictors 3 and 4 and C_2 represents the classification error of the ECOC model trained using all predictors, then the test is:

$$H_0: C_1 \geq C_2$$
$$H_1: C_1 < C_2$$

By default, `testckfold` conducts a 5-by-2 k-fold F test, which is not appropriate as a one-tailed test. Specify to conduct a 5-by-2 k-fold t test.

```
rng(1); % For reproducibility
[h,pValue] = testckfold(t,t,meas(:,pairs(6,:)),meas,species,...
    'Alternative','greater','Test','5x2t')
```

```
h = logical
   0


pValue = 0.8940
```

The `h = 0` indicates that there is not enough evidence to suggest that the model trained using predictors 3 and 4 is more accurate than the model trained using all predictors.

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Coding','ternarycomplete','FitPosterior',true,'Learners','tree'` specifies a ternary complete coding design, to transform scores to posterior probabilities, and to grow classification trees for all binary learners.

`Coding` — Coding design
`'onevsone'` (default) | `'allpairs'` | `'binarycomplete'` | `'denserandom'` | `'onevsall'` | `'ordinal'` | `'sparserandom'` | `'ternarycomplete'` | numeric matrix

Coding design name, specified as the comma-separated pair consisting of `'Coding'` and a numeric matrix or a value in this table.

Value	Number of Binary Learners	Description
`'allpairs'` and `'onevsone'`	$K(K – 1)/2$	For each binary learner, one class is positive, another is negative, and the software ignores the rest. This design exhausts all combinations of class pair assignments.
`'binarycomplete'`	$2^{(K – 1)} – 1$	This design partitions the classes into all binary combinations, and does not ignore any classes. For each binary learner, all class assignments are `-1` and `1` with at least one positive and negative class in the assignment.
`'denserandom'`	Random, but approximately $10 \log_2 K$	For each binary learner, the software randomly assigns classes into positive or negative classes, with at least one of each type. For more details, see “Random Coding Design Matrices” on page 33-1646.

Value	Number of Binary Learners	Description
`'onevsall'`	K	For each binary learner, one class is positive and the rest are negative. This design exhausts all combinations of positive class assignments.
`'ordinal'`	K – 1	For the first binary learner, the first class is negative, and the rest positive. For the second binary learner, the first two classes are negative, the rest positive, and so on.
`'sparserandom'`	Random, but approximately 15 $\log_2 K$	For each binary learner, the software randomly assigns classes as positive or negative with probability 0.25 for each, and ignores classes with probability 0.5. For more details, see “Random Coding Design Matrices” on page 33-1646.
`'ternarycomplete'`	$\left(3^K - 2^{(K+1)} + 1\right)/2$	This design partitions the classes into all ternary combinations. All class assignments are `0`, `-1`, and `1` with at least one positive and one negative class in the assignment.

You can also specify a coding design using a custom coding matrix. The custom coding matrix is a K-by-L matrix. Each row corresponds to a class and each column corresponds to a binary learner. The class order (rows) corresponds to the order in `ClassNames`. Compose the matrix by following these guidelines:

- Every element of the custom coding matrix must be `-1`, `0`, or `1`, and the value must correspond to a dichotomous class assignment. This table describes the meaning of `Coding(i,j)`, that is, the class that learner `j` assigns to observations in class `i`.

Value	Dichotomous Class Assignment
–1	Learner `j` assigns observations in class `i` to a negative class.
`0`	Before training, learner `j` removes observations in class `i` from the data set.
`1`	Learner `j` assigns observations in class `i` to a positive class.

- Every column must contain at least one `-1` or `1`.
- For all column indices `i`,`j` such that `i` ≠ `j`, `Coding(:,i)` cannot equal `Coding(:,j)` and `Coding(:,i)` cannot equal `-Coding(:,j)`.

- All rows of the custom coding matrix must be different.

For more details on the form of custom coding design matrices, see “Custom Coding Design Matrices” on page 33-1644.

Example: `'Coding','ternarycomplete'`

Data Types: `char` | `string` | `double` | `single` | `int16` | `int32` | `int64` | `int8`

`FitPosterior` — Flag indicating whether to transform scores to posterior probabilities
`false` or `0` (default) | `true` or `1`

Flag indicating whether to transform scores to posterior probabilities, specified as the comma-separated pair consisting of `'FitPosterior'` and a `true` (`1`) or `false` (`0`).

If `FitPosterior` is `true`, then the software transforms binary-learner classification scores to posterior probabilities. You can obtain posterior probabilities by using `kfoldPredict`, `predict`, or `resubPredict`.

`fitcecoc` does not support fitting posterior probabilities if:

- The ensemble method is `AdaBoostM2`, `LPBoost`, `RUSBoost`, `RobustBoost`, or `TotalBoost`.
- The binary learners (`Learners`) are linear or kernel classification models that implement SVM. To obtain posterior probabilities for linear or kernel classification models, implement logistic regression instead.

Example: `'FitPosterior',true`

Data Types: `logical`

`Learners` — Binary learner templates
`'svm'` (default) | `'discriminant'` | `'kernel'` | `'knn'` | `'linear'` | `'naivebayes'` | `'tree'` | template object | cell vector of template objects

Binary learner templates, specified as the comma-separated pair consisting of `'Learners'` and a character vector, string scalar, template object, or cell vector of template objects. Specifically, you can specify binary classifiers such as SVM, and the ensembles that use `GentleBoost`, `LogitBoost`, and `RobustBoost`, to solve multiclass problems. However, `fitcecoc` also supports multiclass models as binary classifiers.

- If `Learners` is a character vector or string scalar, then the software trains each binary learner using the default values of the specified algorithm. This table summarizes the available algorithms.

Value	Description
`'discriminant'`	Discriminant analysis. For default options, see `templateDiscriminant`.
`'kernel'`	Kernel classification model. For default options, see `templateKernel`.
`'knn'`	*k*-nearest neighbors. For default options, see `templateKNN`.
`'linear'`	Linear classification model. For default options, see `templateLinear`.

Value	Description
'naivebayes'	Naive Bayes. For default options, see templateNaiveBayes.
'svm'	SVM. For default options, see templateSVM.
'tree'	Classification trees. For default options, see templateTree.

- If Learners is a template object, then each binary learner trains according to the stored options. You can create a template object using:
  - templateDiscriminant, for discriminant analysis.
  - templateEnsemble, for ensemble learning. You must at least specify the learning method (Method), the number of learners (NLearn), and the type of learner (Learners). You cannot use the AdaBoostM2 ensemble method for binary learning.
  - templateKernel, for kernel classification.
  - templateKNN, for *k*-nearest neighbors.
  - templateLinear, for linear classification.
  - templateNaiveBayes, for naive Bayes.
  - templateSVM, for SVM.
  - templateTree, for classification trees.
- If Learners is a cell vector of template objects, then:
  - Cell *j* corresponds to binary learner *j* (in other words, column *j* of the coding design matrix), and the cell vector must have length *L*. *L* is the number of columns in the coding design matrix. For details, see Coding.
  - To use one of the built-in loss functions for prediction, then all binary learners must return a score in the same range. For example, you cannot include default SVM binary learners with default naive Bayes binary learners. The former returns a score in the range (-∞,∞), and the latter returns a posterior probability as a score. Otherwise, you must provide a custom loss as a function handle to functions such as predict and loss.
  - You cannot specify linear classification model learner templates with any other template.
  - Similarly, you cannot specify kernel classification model learner templates with any other template.

By default, the software trains learners using default SVM templates.

Example: 'Learners','tree'

Output Arguments

t — ECOC classification template

template object

ECOC classification template, returned as a template object. Pass t to testckfold to specify how to create an ECOC classifier whose predictive performance you want to compare with another classifier.

If you display t to the Command Window, then all, unspecified options appear empty ([]). However, the software replaces empty options with their corresponding default values during training.

See Also

`ClassificationECOC` | `designecoc` | `fitcecoc` | `predict` | `templateDiscriminant` | `templateEnsemble` | `templateKNN` | `templateSVM` | `templateTree` | `testckfold`

Introduced in R2015a

templateEnsemble

Ensemble learning template

Syntax

```
t = templateEnsemble(Method,NLearn,Learners)
t = templateEnsemble(Method,NLearn,Learners,Name,Value)
```

Description

`t = templateEnsemble(Method,NLearn,Learners)` returns an ensemble learning template that specifies to use the ensemble aggregation method `Method`, `NLearn` learning cycles, and weak learners `Learners`.

All other options of the template (`t`) specific to ensemble learning appear empty, but the software uses their corresponding default values during training.

`t = templateEnsemble(Method,NLearn,Learners,Name,Value)` returns an ensemble template with additional options specified by one or more name-value pair arguments.

For example, you can specify the number of predictors in each random subspace learner, learning rate for shrinkage, or the target classification error for `RobustBoost`.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those options that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create an Ensemble Learning Template

Use `templateEnsemble` to specify an ensemble learning template. You must specify the ensemble method, the number of learning cycles, and the type of weak learners. For this example, specify the AdaBoostM1 method, 100 learners, and classification tree weak learners.

```
t = templateEnsemble('AdaBoostM1',100,'tree')
```

```
t =
Fit template for classification AdaBoostM1.

                Type: 'classification'
              Method: 'AdaBoostM1'
    LearnerTemplates: 'Tree'
              NLearn: 100
           LearnRate: []
```

All properties of the template object are empty except for `Method`, `Type`, `LearnerTemplates`, and `NLearn`. When trained on, the software fills in the empty properties with their respective default values. For example, the software fills the `LearnRate` property with `1`.

`t` is a plan for an ensemble learner, and no computation takes place when you specify it. You can pass `t` to `fitcecoc` to specify ensemble binary learners for ECOC multiclass learning.

Create an Ensemble Template for ECOC Multiclass Learning

Create an ensemble template for use in `fitcecoc`.

Load the arrhythmia data set.

```
load arrhythmia
tabulate(categorical(Y));
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

```
rng(1); % For reproducibility
```

Some classes have small relative frequencies in the data.

Create a template for a AdaBoostM1 ensemble of classification trees, and specify to use 100 learners and a shrinkage of 0.1. By default, boosting grows stumps (i.e., one node having a set of leaves). Since there are classes with small frequencies, the trees must be leafy enough to be sensitive to the minority classes. Specify the minimum number of leaf node observations to 3.

```
tTree = templateTree('MinLeafSize',20);
t = templateEnsemble('AdaBoostM1',100,tTree,'LearnRate',0.1);
```

All properties of the template objects are empty except for `Method` and `Type`, and the corresponding properties of the name-value pair argument values in the function calls. When you pass `t` to the training function, the software fills in the empty properties with their respective default values.

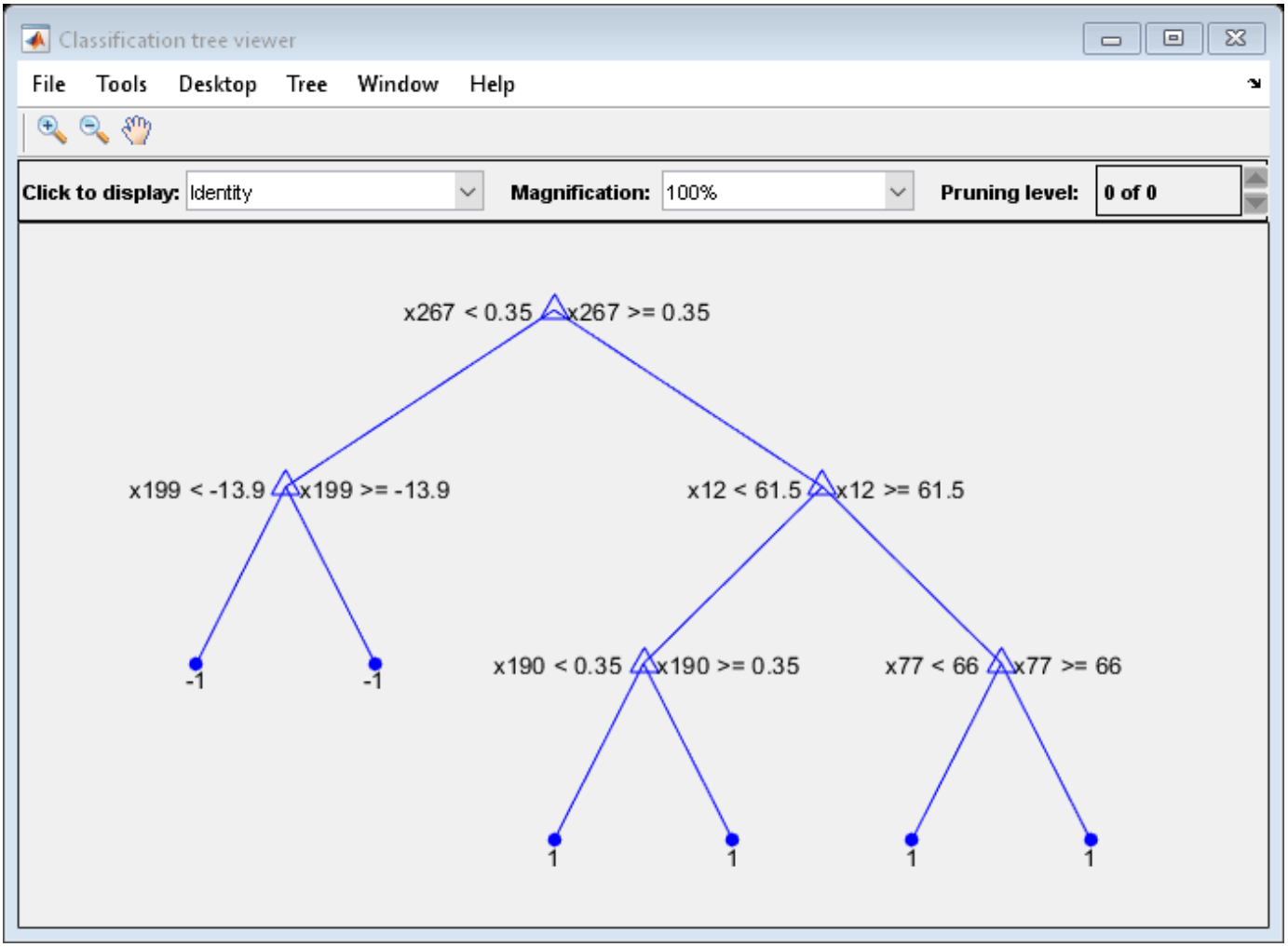
Specify `t` as a binary learner for an ECOC multiclass model. Train using the default one-versus-one coding design.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

- `Mdl` is a `ClassificationECOC` multiclass model.
- `Mdl.BinaryLearners` is a 78-by-1 cell array of `CompactClassificationEnsemble` models.
- `Mdl.BinaryLearners{j}.Trained` is a 100-by-1 cell array of `CompactClassificationTree` models, for `j` = 1,...,78.

You can verify that one of the binary learners contains a weak learner that isn't a stump by using `view`.

```
view(Mdl.BinaryLearners{1}.Trained{1},'Mode','graph')
```



Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl,'LossFun','classiferror')
```

```
L = 0.0819
```

Speed Up Training ECOC Classifiers Using Binning and Parallel Computing

Train a one-versus-all ECOC classifier using a `GentleBoost` ensemble of decision trees with surrogate splits. To speed up training, bin numeric predictors and use parallel computing. Binning is valid only when `fitcecoc` uses a tree learner. After training, estimate the classification error using 10-fold cross-validation. Note that parallel computing requires Parallel Computing Toolbox™.

Load Sample Data

Load and inspect the `arrhythmia` data set.

```
load arrhythmia
[n,p] = size(X)
```

```
n = 452
```

```
p = 279
```

```
isLabels = unique(Y);
nLabels = numel(isLabels)
```

```
nLabels = 13
```

```
tabulate(categorical(Y))
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

The data set contains 279 predictors, and the sample size of 452 is relatively small. Of the 16 distinct labels, only 13 are represented in the response (`Y`). Each label describes various degrees of arrhythmia, and 54.20% of the observations are in class 1.

Train One-Versus-All ECOC Classifier

Create an ensemble template. You must specify at least three arguments: a method, a number of learners, and the type of learner. For this example, specify `'GentleBoost'` for the method, `100` for the number of learners, and a decision tree template that uses surrogate splits because there are missing observations.

```
tTree = templateTree('surrogate','on');
tEnsemble = templateEnsemble('GentleBoost',100,tTree);
```

`tEnsemble` is a template object. Most of its properties are empty, but the software fills them with their default values during training.

Train a one-versus-all ECOC classifier using the ensembles of decision trees as binary learners. To speed up training, use binning and parallel computing.

- Binning (`'NumBins',50`) — When you have a large training data set, you can speed up training (a potential decrease in accuracy) by using the `'NumBins'` name-value pair argument. This argument is valid only when `fitcecoc` uses a tree learner. If you specify the `'NumBins'` value, then the software bins every numeric predictor into a specified number of equiprobable bins, and

then grows trees on the bin indices instead of the original data. You can try `'NumBins',50` first, and then change the `'NumBins'` value depending on the accuracy and training speed.

- Parallel computing (`'Options',statset('UseParallel',true)`) — With a Parallel Computing Toolbox license, you can speed up the computation by using parallel computing, which sends each binary learner to a worker in the pool. The number of workers depends on your system configuration. When you use decision trees for binary learners, `fitcecoc` parallelizes training using Intel® Threading Building Blocks (TBB) for dual-core systems and above. Therefore, specifying the `'UseParallel'` option is not helpful on a single computer. Use this option on a cluster.

Additionally, specify that the prior probabilities are $1/K$, where $K = 13$ is the number of distinct classes.

```
options = statset('UseParallel',true);
Mdl = fitcecoc(X,Y,'Coding','onevsall','Learners',tEnsemble,...
                'Prior','uniform','NumBins',50,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

`Mdl` is a `ClassificationECOC` model.

Cross-Validation

Cross-validate the ECOC classifier using 10-fold cross-validation.

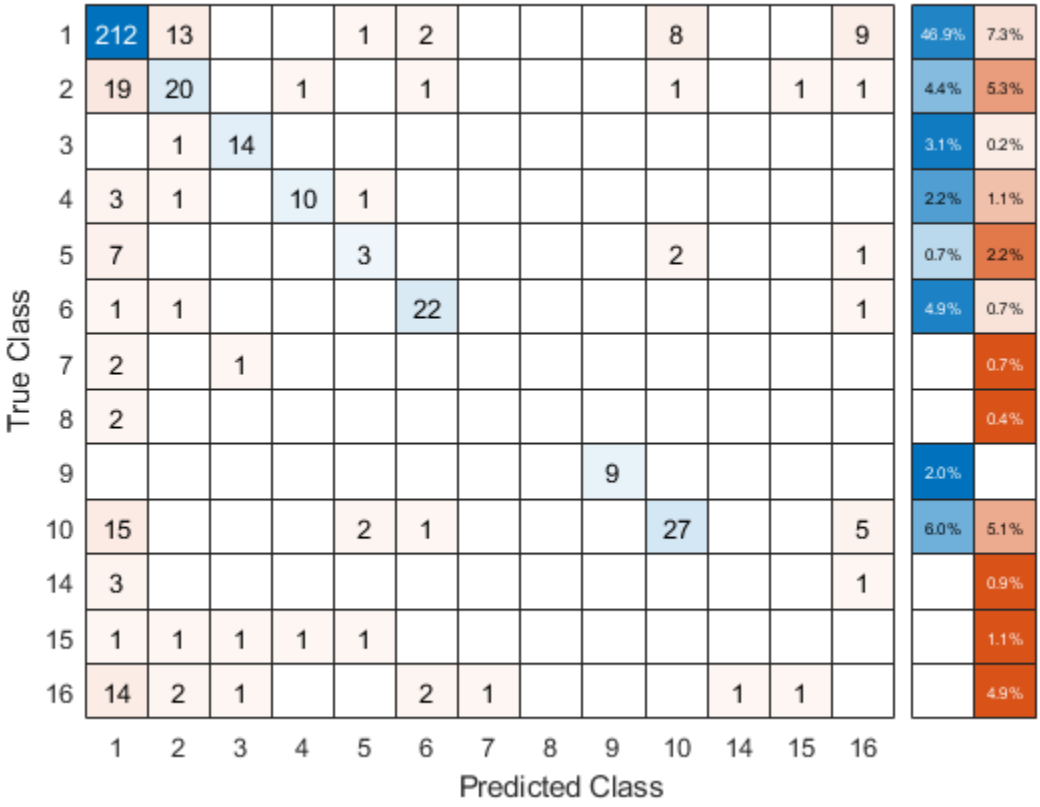
```
CVMdl = crossval(Mdl,'Options',options);
```

```
Warning: One or more folds do not contain points from all the groups.
```

`CVMdl` is a `ClassificationPartitionedECOC` model. The warning indicates that some classes are not represented while the software trains at least one fold. Therefore, those folds cannot predict labels for the missing classes. You can inspect the results of a fold using cell indexing and dot notation. For example, access the results of the first fold by entering `CVMdl.Trained{1}`.

Use the cross-validated ECOC classifier to predict validation-fold labels. You can compute the confusion matrix by using `confusionchart`. Move and resize the chart by changing the inner position property to ensure that the percentages appear in the row summary.

```
oofLabel = kfoldPredict(CVMdl,'Options',options);
ConfMat = confusionchart(Y,oofLabel,'RowSummary','total-normalized');
ConfMat.InnerPosition = [0.10 0.12 0.85 0.85];
```



Reproduce Binned Data

Reproduce binned predictor data by using the `BinEdges` property of the trained model and the `discretize` function.

```
X = Mdl.X; % Predictor data
Xbinned = zeros(size(X));
edges = Mdl.BinEdges;
% Find indices of binned predictors.
idxNumeric = find(~cellfun(@isempty,edges));
if iscolumn(idxNumeric)
    idxNumeric = idxNumeric';
end
for j = idxNumeric
    x = X(:,j);
    % Convert x to array if x is a table.
    if istable(x)
        x = table2array(x);
    end
    % Group x into bins by using the discretize function.
    xbinned = discretize(x,[-inf; edges{j}; inf]);
    Xbinned(:,j) = xbinned;
end
```

`Xbinned` contains the bin indices, ranging from 1 to the number of bins, for numeric predictors. `Xbinned` values are `0` for categorical predictors. If `X` contains `NaN`s, then the corresponding `Xbinned` values are `NaN`s.

Input Arguments

`Method` — Ensemble aggregation method

`'AdaBoostM1'` | `'LogitBoost'` | `'GentleBoost'` | `'RUSBoost'` | `'Subspace'` | `'Bag'` | `'AdaBoostM2'` | `'LSBoost'` | ...

Ensemble aggregation method, specified as one of the method names in this list.

- For classification with two classes:
  - `'AdaBoostM1'`
  - `'LogitBoost'`
  - `'GentleBoost'`
  - `'RobustBoost'` (requires Optimization Toolbox)
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For classification with three or more classes:
  - `'AdaBoostM2'`
  - `'LPBoost'` (requires Optimization Toolbox)
  - `'TotalBoost'` (requires Optimization Toolbox)
  - `'RUSBoost'`
  - `'Subspace'`
  - `'Bag'`
- For regression:
  - `'LSBoost'`
  - `'Bag'`

If you specify `'Method','Bag'`, then specify the problem type using the `Type` name-value pair argument, because you can specify `'Bag'` for classification and regression problems.

For details about ensemble aggregation algorithms and examples, see “Ensemble Algorithms” on page 18-39 and “Choose an Applicable Ensemble Aggregation Method” on page 18-32.

`NLearn` — Number of ensemble learning cycles

positive integer | `'AllPredictorCombinations'`

Number of ensemble learning cycles, specified as a positive integer or `'AllPredictorCombinations'`.

- If you specify a positive integer, then, at every learning cycle, the software trains one weak learner for every template object in `Learners`. Consequently, the software trains `NLearn*numel(Learners)` learners.
- If you specify `'AllPredictorCombinations'`, then set `Method` to `'Subspace'` and specify one learner only in `Learners`. With these settings, the software trains learners for all possible combinations of predictors taken `NPredToSample` at a time. Consequently, the software trains `nchoosek(size(X,2),NPredToSample)` learners.

For more details, see “Tips” on page 33-6111.

Data Types: `single` | `double` | `char` | `string`

`Learners` — Weak learners to use in ensemble

weak-learner name | weak-learner template object | cell vector of weak-learner template objects

Weak learners to use in the ensemble, specified as a weak-learner name, weak-learner template object, or cell array of weak-learner template objects.

Weak Learner	Weak-Learner Name	Template Object Creation Function	Method Settings
Discriminant analysis	`'Discriminant'`	`templateDiscriminant`	Recommended for `'Subspace'`
k nearest neighbors	`'KNN'`	`templateKNN`	For `'Subspace'` only
Decision tree	`'Tree'`	`templateTree`	All methods except `'Subspace'`

For more details, see `NLearn` and “Tips” on page 33-6111.

Example: For an ensemble composed of two types of classification trees, supply `{t1 t2}`, where `t1` and `t2` are classification tree templates.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'LearningRate',0.05,'NPrint',5` specifies to use `0.05` as the learning rate and to display a message to the command line every time it trains `5` learners.

General Ensemble Options

`NPrint` — Printout frequency

`'off'` (default) | positive integer

Printout frequency, specified as the comma-separated pair consisting of `'NPrint'` and a positive integer or `'off'`.

To track the number of *weak learners* or *folds* that the software trained so far, specify a positive integer. That is, if you specify the positive integer *m* and:

- Do not specify any cross-validation option of the fitting function (for example, `CrossVal`), then the software displays a message to the command line every time it completes training *m* weak learners.

- A cross-validation option, then the software displays a message to the command line every time it finishes training *m* folds.

If you specify `'off'`, then the software does not display a message when it completes training weak learners.

Tip When training an ensemble of many weak learners on a large data set, specify a positive integer for `NPrint`.

Example: `'NPrint',5`

Data Types: `single` | `double` | `char` | `string`

Type — Supervised learning type

`'classification'` | `'regression'`

Supervised learning type, specified as the comma-separated pair consisting of `'Type'` and `'classification'` or `'regression'`.

- If `Method` is `'bag'`, then the supervised learning type is ambiguous. Therefore, specify `Type` when bagging.
- Otherwise, the value of `Method` determines the supervised learning type.

Example: `'Type','classification'`

Sampling Options for Boosting Methods and Bagging

FResample — Fraction of training set to resample

1 (default) | positive scalar in (0,1]

Fraction of the training set to resample for every weak learner, specified as the comma-separated pair consisting of `'FResample'` and a positive scalar in (0,1].

To use `'FResample'`, specify `'bag'` for `Method` or set `Resample` to `'on'`.

Example: `'FResample',0.75`

Data Types: `single` | `double`

Replace — Flag indicating to sample with replacement

`'on'` (default) | `'off'`

Flag indicating sampling with replacement, specified as the comma-separated pair consisting of `'Replace'` and `'off'` or `'on'`.

- For `'on'`, the software samples the training observations with replacement.
- For `'off'`, the software samples the training observations without replacement. If you set `Resample` to `'on'`, then the software samples training observations assuming uniform weights. If you also specify a boosting method, then the software boosts by reweighting observations.

Unless you set `Method` to `'bag'` or set `Resample` to `'on'`, `Replace` has no effect.

Example: `'Replace','off'`

`Resample` — Flag indicating to resample

`'off'` | `'on'`

Flag indicating to resample, specified as the comma-separated pair consisting of `'Resample'` and `'off'` or `'on'`.

- If `Method` is a boosting method, then:
  - `'Resample','on'` specifies to sample training observations using updated weights as the multinomial sampling probabilities.
  - `'Resample','off'`(default) specifies to reweight observations at every learning iteration.
- If `Method` is `'bag'`, then `'Resample'` must be `'on'`. The software resamples a fraction of the training observations (see `FResample`) with or without replacement (see `Replace`).

If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of 1 for `FResample`.

AdaBoostM1, AdaBoostM2, LogitBoost, GentleBoost, and LSBoost Method Options

`LearnRate` — Learning rate for shrinkage

1 (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

RUSBoost Method Options

`LearnRate` — Learning rate for shrinkage

1 (default) | numeric scalar in (0,1]

Learning rate for shrinkage, specified as the comma-separated pair consisting of `'LearnRate'` and a numeric scalar in the interval (0,1].

To train an ensemble using shrinkage, set `LearnRate` to a value less than `1`, for example, `0.1` is a popular choice. Training an ensemble using shrinkage requires more learning iterations, but often achieves better accuracy.

Example: `'LearnRate',0.1`

Data Types: `single` | `double`

`RatioToSmallest` — Sampling proportion with respect to lowest-represented class

positive numeric scalar | numeric vector of positive values

Sampling proportion with respect to the lowest-represented class, specified as the comma-separated pair consisting of `'RatioToSmallest'` and a numeric scalar or numeric vector of positive values with length equal to the number of distinct classes in the training data.

Suppose that there are *K* classes in the training data and the lowest-represented class has *m* observations in the training data.

- If you specify the positive numeric scalar *s*, then the software samples *s***m* observations from each class, that is, it uses the same sampling proportion for each class. For more details, see "Algorithms" on page 33-6112.
- If you specify the numeric vector `[s1,s2,...,sK]`, then the software samples *si***m* observations from class *i*, *i* = 1,...,*K*. The elements of `RatioToSmallest` correspond to the order of the class names specified using the `ClassNames` name-value pair argument of the fitting function (see "Tips" on page 33-6111).

The default value is `ones(K,1)`, which specifies to sample *m* observations from each class.

Example: `'RatioToSmallest',[2,1]`

Data Types: `single` | `double`

LPBoost and TotalBoost Method Options

`MarginPrecision` — Margin precision to control convergence speed

`0.1` (default) | numeric scalar in [0,1]

Margin precision to control convergence speed, specified as the comma-separated pair consisting of `'MarginPrecision'` and a numeric scalar in the interval [0,1]. `MarginPrecision` affects the number of boosting iterations required for convergence.

Tip To train an ensemble using many learners, specify a small value for `MarginPrecision`. For training using a few learners, specify a large value.

Example: `'MarginPrecision',0.5`

Data Types: `single` | `double`

RobustBoost Method Options

`RobustErrorGoal` — Target classification error

`0.1` (default) | nonnegative numeric scalar

Target classification error, specified as the comma-separated pair consisting of `'RobustErrorGoal'` and a nonnegative numeric scalar. The upper bound on possible values depends on the values of `RobustMarginSigma` and `RobustMaxMargin`. However, the upper bound cannot exceed `1`.

Tip For a particular training set, usually there is an optimal range for `RobustErrorGoal`. If you set it too low or too high, then the software can produce a model with poor classification accuracy. Try cross-validating to search for the appropriate value.

Example: `'RobustErrorGoal',0.05`

Data Types: `single` | `double`

`RobustMarginSigma` — Classification margin distribution spread

`0.1` (default) | positive numeric scalar

Classification margin distribution spread over the training data, specified as the comma-separated pair consisting of `'RobustMarginSigma'` and a positive numeric scalar. Before specifying `RobustMarginSigma`, consult the literature on `RobustBoost`, for example, [19].

Example: `'RobustMarginSigma',0.5`

Data Types: `single` | `double`

`RobustMaxMargin` — Maximal classification margin
`0` (default) | nonnegative numeric scalar

Maximal classification margin in the training data, specified as the comma-separated pair consisting of `'RobustMaxMargin'` and a nonnegative numeric scalar. The software minimizes the number of observations in the training data having classification margins below `RobustMaxMargin`.

Example: `'RobustMaxMargin',1`

Data Types: `single` | `double`

Random Subspace Method Options

`NPredToSample` — Number of predictors to sample
`1` (default) | positive integer

Number of predictors to sample for each random subspace learner, specified as the comma-separated pair consisting of `'NPredToSample'` and a positive integer in the interval 1,...,*p*, where *p* is the number of predictor variables (`size(X,2)` or `size(Tbl,2)`).

Data Types: `single` | `double`

Output Arguments

`t` — Classification template for ensemble learning
classification template object

Classification template for ensemble learning, returned as a template object. You can pass `t` to, for example, `fitcecoc` to specify how to create the ensemble learning classifier for the ECOC model.

If you display `t` in the Command Window, then all, unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

Tips

- `NLearn` can vary from a few dozen to a few thousand. Usually, an ensemble with good predictive power requires from a few hundred to a few thousand weak learners. However, you do not have to train an ensemble for that many cycles at once. You can start by growing a few dozen learners, inspect the ensemble performance and then, if necessary, train more weak learners using `resume` for classification problems, or `resume` for regression problems.
- Ensemble performance depends on the ensemble setting and the setting of the weak learners. That is, if you specify weak learners with default parameters, then the ensemble can perform. Therefore, like ensemble settings, it is good practice to adjust the parameters of the weak learners using templates, and to choose values that minimize generalization error.
- If you specify to resample using `Resample`, then it is good practice to resample to entire data set. That is, use the default setting of `1` for `FResample`.

- In classification problems (that is, `Type` is `'classification'`):
  - If the ensemble aggregation method (`Method`) is `'bag'` and:
    - The misclassification cost is highly imbalanced, then, for in-bag samples, the software oversamples unique observations from the class that has a large penalty.
    - The class prior probabilities are highly skewed, the software oversamples unique observations from the class that has a large prior probability.

    For smaller sample sizes, these combinations can result in a very low relative frequency of out-of-bag observations from the class that has a large penalty or prior probability. Consequently, the estimated out-of-bag error is highly variable and it might be difficult to interpret. To avoid large estimated out-of-bag error variances, particularly for small sample sizes, set a more balanced misclassification cost matrix using the `Cost` name-value pair argument of the fitting function, or a less skewed prior probability vector using `Prior` name-value pair argument of the fitting function.
  - Because the order of some input and output arguments correspond to the distinct classes in the training data, it is good practice to specify the class order using the `ClassNames` name-value pair argument of the fitting function.
    - To quickly determine the class order, remove all observations from the training data that are unclassified (that is, have a missing label), obtain and display an array of all the distinct classes, and then specify the array for `ClassNames`. For example, suppose the response variable (`Y`) is a cell array of labels. This code specifies the class order in the variable `classNames`.

      ```
      Ycat = categorical(Y);
      classNames = categories(Ycat)
      ```

      `categorical` assigns `<undefined>` to unclassified observations and `categories` excludes `<undefined>` from its output. Therefore, if you use this code for cell arrays of labels or similar code for categorical arrays, then you do not have to remove observations with missing labels to obtain a list of the distinct classes.
    - To specify that order should be from lowest-represented label to most-represented, then quickly determine the class order (as in the previous bullet), but arrange the classes in the list by frequency before passing the list to `ClassNames`. Following from the previous example, this code specifies the class order from lowest- to most-represented in `classNamesLH`.

      ```
      Ycat = categorical(Y);
      classNames = categories(Ycat);
      freq = countcats(Ycat);
      [~,idx] = sort(freq);
      classNamesLH = classNames(idx);
      ```

Algorithms

- For details of ensemble aggregation algorithms, see “Ensemble Algorithms” on page 18-39.
- If you specify `Method` to be a boosting algorithm and `Learners` to be decision trees, then the software grows stumps by default. A decision stump is one root node connected to two terminal, leaf nodes. You can adjust tree depth by specifying the `MaxNumSplits`, `MinLeafSize`, and `MinParentSize` name-value pair arguments using `templateTree`.

- The software generates in-bag samples by oversampling classes with large misclassification costs and undersampling classes with small misclassification costs. Consequently, out-of-bag samples have fewer observations from classes with large misclassification costs and more observations from classes with small misclassification costs. If you train a classification ensemble using a small data set and a highly skewed cost matrix, then the number of out-of-bag observations per class might be very low. Therefore, the estimated out-of-bag error might have a large variance and might be difficult to interpret. The same phenomenon can occur for classes with large prior probabilities.
- For the RUSBoost ensemble aggregation method (`Method`), the name-value pair argument `RatioToSmallest` specifies the sampling proportion for each class with respect to the lowest-represented class. For example, suppose that there are 2 classes in the training data, *A* and *B*. *A* have 100 observations and *B* have 10 observations. Also, suppose that the lowest-represented class has *m* observations in the training data.
  - If you set `'RatioToSmallest',2`, then *s***m* = `2*10 = 20`. Consequently, the software trains every learner using 20 observations from class *A* and 20 observations from class *B*. If you set `'RatioToSmallest',[2 2]`, then you will obtain the same result.
  - If you set `'RatioToSmallest',[2,1]`, then *s1***m* = `2*10 = 20` and *s2***m* = `1*10 = 10`. Consequently, the software trains every learner using 20 observations from class *A* and 10 observations from class *B*.
- For ensembles of decision trees, and for dual-core systems and above, `fitcensemble` and `fitrensemble` parallelize training using Intel Threading Building Blocks (TBB). For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.

See Also

`ClassificationECOC` | `ClassificationEnsemble` | `fitcecoc` | `fitcensemble` | `fitrensemble` | `templateDiscriminant` | `templateKNN` | `templateTree`

Topics

“Supervised Learning Workflow and Algorithms” on page 18-3
“Framework for Ensemble Learning” on page 18-31

Introduced in R2014b

templateKernel

Kernel model template

Syntax

```
t = templateKernel()
t = templateKernel(Name,Value)
```

Description

`templateKernel` creates a template suitable for fitting a Gaussian kernel classification model for nonlinear classification.

The template specifies the binary learner model, number of dimensions of expanded space, kernel scale, box constraint, and regularization strength, among other parameters. After creating the template, train the model by passing the template and data to `fitcecoc`.

`t = templateKernel()` returns a kernel model template.

If you create a default template, then the software uses default values for all input arguments during training.

`t = templateKernel(Name,Value)` returns a template with additional options specified by one or more name-value pair arguments. For example, you can implement logistic regression or specify the number of dimensions of the expanded space.

If you display `t` in the Command Window, then some properties of `t` appear empty (`[]`). During training, the software uses default values for the empty properties.

Examples

Create Default Kernel Model Template

Create a default kernel model template and use it to train an error-correcting output codes (ECOC) multiclass model.

Load Fisher's iris data set.

```
load fisheriris
```

Create a default kernel model template.

```
t = templateKernel()
```

```
t =
Fit template for classification Kernel.

             BetaTolerance: []
                 BlockSize: []
             BoxConstraint: []
```

```
                Epsilon: []
    NumExpansionDimensions: []
         GradientTolerance: []
        HessianHistorySize: []
            IterationLimit: []
               KernelScale: []
                    Lambda: []
                   Learner: 'svm'
              LossFunction: []
                    Stream: []
             VerbosityLevel: []
                   Version: 1
                    Method: 'Kernel'
                      Type: 'classification'
```

During training, the software fills in the empty properties with their respective default values.

Specify `t` as a binary learner for an ECOC multiclass model.

```
Mdl = fitcecoc(meas,species,'Learners',t)
```

```
Mdl =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: {'setosa'  'versicolor'  'virginica'}
    ScoreTransform: 'none'
    BinaryLearners: {3x1 cell}
      CodingMatrix: [3x3 double]


  Properties, Methods
```

`Mdl` is a `CompactClassificationECOC` multiclass classifier.

Specify Kernel Model Template Options

Create a kernel model template with additional options to implement logistic regression with a kernel scale parameter selected by a heuristic procedure.

```
t = templateKernel('Learner','logistic','KernelScale','auto')
```

```
t =
Fit template for classification Kernel.

             BetaTolerance: []
                 BlockSize: []
             BoxConstraint: []
                   Epsilon: []
    NumExpansionDimensions: []
         GradientTolerance: []
        HessianHistorySize: []
            IterationLimit: []
               KernelScale: 'auto'
                    Lambda: []
```

```
          Learner: 'logistic'
     LossFunction: []
           Stream: []
    VerbosityLevel: []
          Version: 1
           Method: 'Kernel'
             Type: 'classification'
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example:
`'Learner','logistic','NumExpansionDimensions',2^15,'KernelScale','auto'`
specifies to implement logistic regression after mapping the predictor data to the 2^15 dimensional space using feature expansion with a kernel scale parameter selected by a heuristic procedure.

Kernel Classification Options

`Learner` — Linear classification model type

`'svm'` (default) | `'logistic'`

Linear classification model type, specified as the comma-separated pair consisting of `'Learner'` and `'svm'` or `'logistic'`.

In the following table, $f(x) = T(x)\beta + b$.

- x is an observation (row vector) from p predictor variables.
- $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
- β is a vector of m coefficients.
- b is the scalar bias.

Value	Algorithm	Response Range	Loss Function
`'svm'`	Support vector machine	$y \in \{-1,1\}$; 1 for the positive class and –1 otherwise	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$
`'logistic'`	Logistic regression	Same as `'svm'`	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$

Example: `'Learner','logistic'`

`NumExpansionDimensions` — Number of dimensions of expanded space

`'auto'` (default) | positive integer

Number of dimensions of the expanded space, specified as the comma-separated pair consisting of `'NumExpansionDimensions'` and `'auto'` or a positive integer. For `'auto'`, the `templateKernel`

function selects the number of dimensions using `2.^ceil(min(log2(p)+5,15))`, where `p` is the number of predictors.

For details, see “Random Feature Expansion” on page 33-6120.

Example: `'NumExpansionDimensions',2^15`

Data Types: `char` | `string` | `single` | `double`

`KernelScale` — Kernel scale parameter
1 (default) | `'auto'` | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software obtains a random basis for random feature expansion by using the kernel scale parameter. For details, see “Random Feature Expansion” on page 33-6120.

If you specify `'auto'`, then the software selects an appropriate kernel scale parameter using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed by using `rng` before training.

Example: `'KernelScale','auto'`

Data Types: `char` | `string` | `single` | `double`

`BoxConstraint` — Box constraint
1 (default) | positive scalar

Box constraint, specified as the comma-separated pair consisting of `'BoxConstraint'` and a positive scalar.

This argument is valid only when `'Learner'` is `'svm'`(default) and you do not specify a value for the regularization term strength `'Lambda'`. You can specify either `'BoxConstraint'` or `'Lambda'` because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations.

Example: `'BoxConstraint',100`

Data Types: `single` | `double`

`Lambda` — Regularization term strength
`'auto'` (default) | nonnegative scalar

Regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and `'auto'` or a nonnegative scalar.

For `'auto'`, the value of `'Lambda'` is $1/n$, where n is the number of observations.

You can specify either `'BoxConstraint'` or `'Lambda'` because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$.

Example: `'Lambda',0.01`

Data Types: `char` | `string` | `single` | `double`

Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e–5` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of optimization iterations
positive integer

Maximum number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The default value is 1000 if the transformed data fits in memory, as specified by the `BlockSize` name-value pair argument. Otherwise, the default value is 100.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Other Kernel Classification Options

`BlockSize` — Maximum amount of allocated memory
`4e^3` (4GB) (default) | positive scalar

Maximum amount of allocated memory (in megabytes), specified as the comma-separated pair consisting of `'BlockSize'` and a positive scalar.

If `templateKernel` requires more memory than the value of `'BlockSize'` to hold the transformed predictor data, then the software uses a block-wise strategy. For details about the block-wise strategy, see “Algorithms” on page 33-6723.

Example: `'BlockSize',1e4`

Data Types: `single` | `double`

`RandomStream` — Random number stream
global stream (default) | random stream object

Random number stream for reproducibility of data transformation, specified as the comma-separated pair consisting of `'RandomStream'` and a random stream object. For details, see “Random Feature Expansion” on page 33-6120.

Use `'RandomStream'` to reproduce the random basis functions that `templateKernel` uses to transform the predictor data to a high-dimensional space. For details, see “Managing the Global Stream Using RandStream” and “Creating and Controlling a Random Number Stream”.

Example: `'RandomStream',RandStream('mlfg6331_64')`

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of the history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. At each iteration, `templateKernel` composes the Hessian approximation by using statistics from the latest `HessianHistorySize` iterations.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `0` or `1`. `Verbose` controls the display of diagnostic information at the command line.

Value	Description
`0`	`templateKernel` does not display diagnostic information.
`1`	`templateKernel` displays the value of the objective function, gradient magnitude, and other diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

Output Arguments

`t` — Kernel model template
template object

Kernel model template, returned as a template object. To train a kernel classification model for multiclass problems, pass `t` to `fitcecoc`.

If you display `t` in the Command Window, then some properties appear empty (`[]`). The software replaces the empty properties with their corresponding default values during training.

More About

Random Feature Expansion

Random feature expansion, such as Random Kitchen Sinks[1] and Fastfood[2], is a scheme to approximate Gaussian kernels of the kernel classification algorithm to use for big data in a computationally efficient way. Random feature expansion is more practical for big data applications that have large training sets, but can also be applied to smaller data sets that fit in memory.

The kernel classification algorithm searches for an optimal hyperplane that separates the data into two classes after mapping features into a high-dimensional space. Nonlinear features that are not linearly separable in a low-dimensional space can be separable in the expanded high-dimensional space. All the calculations for hyperplane classification use only dot products. You can obtain a nonlinear classification model by replacing the dot product x_1x_2' with the nonlinear kernel function $G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle$, where x_i is the ith observation (row vector) and $\varphi(x_i)$ is a transformation that maps x_i to a high-dimensional space (called the "kernel trick"). However, evaluating $G(x_1,x_2)$ (Gram matrix) for each pair of observations is computationally expensive for a large data set (large n).

The random feature expansion scheme finds a random transformation so that its dot product approximates the Gaussian kernel. That is,

$$G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle \approx T(x_1)T(x_2)',$$

where $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$). The Random Kitchen Sink scheme uses the random transformation

$$T(x) = m^{-1/2}\exp(iZx')',$$

where $Z \in \mathbb{R}^{m \times p}$ is a sample drawn from $N\left(0, \sigma^{-2}\right)$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage. The Fastfood scheme introduces another random basis V instead of Z using Hadamard matrices combined with Gaussian scaling matrices. This random basis reduces the computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

The `templateKernel` function uses the Fastfood scheme for random feature expansion and uses linear classification to train a Gaussian kernel classification model. Unlike solvers in the `templateSVM` function, which require computation of the n-by-n Gram matrix, the solver in `templateKernel` only needs to form a matrix of size n-by-m, with m typically much less than n for big data.

Box Constraint

A box constraint is a parameter that controls the maximum penalty imposed on margin-violating observations, and aids in preventing overfitting (regularization). Increasing the box constraint can lead to longer training times.

The box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations.

Algorithms

`templateKernel` minimizes the regularized objective function using a Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver with ridge (L_2) regularization. To find the type of LBFGS solver used for training, type `FitInfo.Solver` in the Command Window.

- `'LBFGS-fast'` — LBFGS solver.
- `'LBFGS-blockwise'` — LBFGS solver with a block-wise strategy. If `templateKernel` requires more memory than the value of `BlockSize` to hold the transformed predictor data, then it uses a block-wise strategy.
- `'LBFGS-tall'` — LBFGS solver with a block-wise strategy for tall arrays.

When `templateKernel` uses a block-wise strategy, `templateKernel` implements LBFGS by distributing the calculation of the loss and gradient among different parts of the data at each iteration. Also, `templateKernel` refines the initial estimates of the linear coefficients and the bias term by fitting the model locally to parts of the data and combining the coefficients by averaging. If you specify `'Verbose',1`, then `templateKernel` displays diagnostic information for each data pass and stores the information in the `History` field of `FitInfo`.

When `templateKernel` does not use a block-wise strategy, the initial estimates are zeros. If you specify `'Verbose',1`, then `templateKernel` displays diagnostic information for each iteration and stores the information in the `History` field of `FitInfo`.

References

[1] Rahimi, A., and B. Recht. “Random Features for Large-Scale Kernel Machines.” *Advances in Neural Information Processing Systems*. Vol. 20, 2008, pp. 1177–1184.

[2] Le, Q., T. Sarlós, and A. Smola. “Fastfood — Approximating Kernel Expansions in Loglinear Time.” *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, No. 3, 2013, pp. 244–252.

[3] Huang, P. S., H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. “Kernel methods match Deep Neural Networks on TIMIT.” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 205–209.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations when you train a model by passing a kernel model template and tall arrays to `fitcecoc`:

- The default values for these name-value pair arguments are different when you work with tall arrays.
  - `'Verbose'` — Default value is `1`.
  - `'BetaTolerance'` — Default value is relaxed to `1e–3`.
  - `'GradientTolerance'` — Default value is relaxed to `1e–5`.
  - `'IterationLimit'` — Default value is relaxed to `20`.

- If `'KernelScale'` is `'auto'`, then `templateKernel` uses the random stream controlled by `tallrng` for subsampling. For reproducibility, you must set a random number seed for both the global stream and the random stream controlled by `tallrng`.
- If `'Lambda'` is `'auto'`, then `templateKernel` might take an extra pass through the data to calculate the number of observations.
- `templateKernel` uses a block-wise strategy. For details, see “Algorithms” on page 33-6121.

For more information, see “Tall Arrays”.

See Also

`fitcecoc` | `fitckernel`

Introduced in R2018b

templateKNN

k-nearest neighbor classifier template

Syntax

```
t = templateKNN()
t = templateKNN(Name,Value)
```

Description

`t = templateKNN()` returns a *k*-nearest neighbor (KNN) learner template suitable for training ensembles or error-correcting output code (ECOC) multiclass models.

If you specify a default template, then the software uses default values for all input arguments during training.

Specify `t` as a learner in `fitcensemble` or `fitcecoc`.

`t = templateKNN(Name,Value)` creates a template with additional options specified by one or more name-value pair arguments.

For example, you can specify the nearest neighbor search method, the number of nearest neighbors to find, or the distance metric.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a *k*-Nearest Neighbor Template for Ensemble

Create a nondefault *k*-nearest neighbor template for use in `fitcensemble`.

Load Fisher's iris data set.

```
load fisheriris
```

Create a template for a 5-nearest neighbor search, and specify to standardize the predictors.

```
t = templateKNN('NumNeighbors',5,'Standardize',1)
```

```
t =
Fit template for classification KNN.

       NumNeighbors: 5
           NSMethod: ''
           Distance: ''
         BucketSize: ''
        IncludeTies: []
```

```
     DistanceWeight: []
         BreakTies: []
          Exponent: []
               Cov: []
             Scale: []
    StandardizeData: 1
            Version: 1
             Method: 'KNN'
               Type: 'classification'
```

All properties of the template object are empty except for `NumNeighbors`, `Method`, `StandardizeData`, and `Type`. When you specify `t` as a learner, the software fills in the empty properties with their respective default values.

Specify `t` as a weak learner for a classification ensemble.

```
Mdl = fitcensemble(meas,species,'Method','Subspace','Learners',t);
```

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl)
```

```
L = 0.0600
```

Create a *k*-Nearest Neighbor Template for ECOC Multiclass Learning

Create a nondefault *k*-nearest neighbor template for use in `fitcecoc`.

Load Fisher's iris data set.

```
load fisheriris
```

Create a template for a 5-nearest neighbor search, and specify to standardize the predictors.

```
t = templateKNN('NumNeighbors',5,'Standardize',1)
```

```
t =
Fit template for classification KNN.

       NumNeighbors: 5
           NSMethod: ''
           Distance: ''
         BucketSize: ''
        IncludeTies: []
     DistanceWeight: []
         BreakTies: []
          Exponent: []
               Cov: []
             Scale: []
    StandardizeData: 1
            Version: 1
             Method: 'KNN'
               Type: 'classification'
```

All properties of the template object are empty except for `NumNeighbors`, `Method`, `StandardizeData`, and `Type`. When you specify `t` as a learner, the software fills in the empty properties with their respective default values.

Specify `t` as a binary learner for an ECOC multiclass model.

```
Mdl = fitcecoc(meas,species,'Learners',t);
```

By default, the software trains `Mdl` using the one-versus-one coding design.

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl,'LossFun','classiferror')
```

```
L = 0.0467
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'NumNeighbors',4,'Distance','minkowski'` specifies a 4-nearest neighbor classifier template using the Minkowski distance measure.

`BreakTies` — Tie-breaking algorithm

`'smallest'` (default) | `'nearest'` | `'random'`

Tie-breaking algorithm used by the `predict` method if multiple classes have the same smallest cost, specified as the comma-separated pair consisting of `'BreakTies'` and one of the following:

- `'smallest'` — Use the smallest index among tied groups.
- `'nearest'` — Use the class with the nearest neighbor among tied groups.
- `'random'` — Use a random tiebreaker among tied groups.

By default, ties occur when multiple classes have the same number of nearest points among the `K` nearest neighbors.

Example: `'BreakTies','nearest'`

`BucketSize` — Maximum data points in node

`50` (default) | positive integer value

Maximum number of data points in the leaf node of the *k*d-tree, specified as the comma-separated pair consisting of `'BucketSize'` and a positive integer value. This argument is meaningful only when `NSMethod` is `'kdtree'`.

Example: `'BucketSize',40`

Data Types: `single` | `double`

`Cov` — Covariance matrix

`cov(X,'omitrows')` (default) | positive definite matrix of scalar values

Covariance matrix, specified as the comma-separated pair consisting of `'Cov'` and a positive definite matrix of scalar values representing the covariance matrix when computing the Mahalanobis distance. This argument is only valid when `'Distance'` is `'mahalanobis'`.

You cannot simultaneously specify `'Standardize'` and either of `'Scale'` or `'Cov'`.

Data Types: `single` | `double`

`Distance` — Distance metric
`'cityblock'` | `'chebychev'` | `'correlation'` | `'cosine'` | `'euclidean'` | `'hamming'` | function handle | ...

Distance metric, specified as the comma-separated pair consisting of `'Distance'` and a valid distance metric name or function handle. The allowable distance metric names depend on your choice of a neighbor-searcher method (see `NSMethod`).

NSMethod	Distance Metric Names
`exhaustive`	Any distance metric of `ExhaustiveSearcher`
`kdtree`	`'cityblock'`, `'chebychev'`, `'euclidean'`, or `'minkowski'`

This table includes valid distance metrics of `ExhaustiveSearcher`.

Distance Metric Names	Description
`'cityblock'`	City block distance.
`'chebychev'`	Chebychev distance (maximum coordinate difference).
`'correlation'`	One minus the sample linear correlation between observations (treated as sequences of values).
`'cosine'`	One minus the cosine of the included angle between observations (treated as vectors).
`'euclidean'`	Euclidean distance.
`'hamming'`	Hamming distance, percentage of coordinates that differ.
`'jaccard'`	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ.
`'mahalanobis'`	Mahalanobis distance, computed using a positive definite covariance matrix `C`. The default value of `C` is the sample covariance matrix of `X`, as computed by `cov(X,'omitrows')`. To specify a different value for `C`, use the `'Cov'` name-value pair argument.
`'minkowski'`	Minkowski distance. The default exponent is `2`. To specify a different exponent, use the `'Exponent'` name-value pair argument.
`'seuclidean'`	Standardized Euclidean distance. Each coordinate difference between `X` and a query point is scaled, meaning divided by a scale value `S`. The default value of `S` is the standard deviation computed from `X`, `S = std(X,'omitnan')`. To specify another value for `S`, use the `Scale` name-value pair argument.
`'spearman'`	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).

Distance Metric Names	Description

`@distfun`	Distance function handle. *`distfun`* has the form
	`function D2 = distfun(ZI,ZJ)`
	`% calculation of  distance`
	`...`
	where
	• `ZI` is a `1`-by-`N` vector containing one row of `X` or `Y`.
	• `ZJ` is an `M2`-by-`N` matrix containing multiple rows of `X` or `Y`.
	• `D2` is an `M2`-by-`1` vector of distances, and `D2(k)` is the distance between observations `ZI` and `ZJ(k,:)`.

If you specify `CategoricalPredictors` as `'all'`, then the default distance metric is `'hamming'`. Otherwise, the default distance metric is `'euclidean'`.

For definitions, see “Distance Metrics” on page 18-12.

Example: `'Distance','minkowski'`

Data Types: `char` | `string` | `function_handle`

`DistanceWeight` — Distance weighting function
`'equal'` (default) | `'inverse'` | `'squaredinverse'` | function handle

Distance weighting function, specified as the comma-separated pair consisting of `'DistanceWeight'` and either a function handle or one of the values in this table.

Value	Description
`'equal'`	No weighting
`'inverse'`	Weight is 1/distance
`'squaredinverse'`	Weight is $1/\text{distance}^2$
`@fcn`	*`fcn`* is a function that accepts a matrix of nonnegative distances, and returns a matrix the same size containing nonnegative distance weights. For example, `'squaredinverse'` is equivalent to `@(d)d.^(-2)`.

Example: `'DistanceWeight','inverse'`

Data Types: `char` | `string` | `function_handle`

`Exponent` — Minkowski distance exponent
2 (default) | positive scalar value

Minkowski distance exponent, specified as the comma-separated pair consisting of `'Exponent'` and a positive scalar value. This argument is only valid when `'Distance'` is `'minkowski'`.

Example: `'Exponent',3`

Data Types: `single` | `double`

`IncludeTies` — Tie inclusion flag
`false` (default) | `true`

Tie inclusion flag, specified as the comma-separated pair consisting of `'IncludeTies'` and a logical value indicating whether `predict` includes all the neighbors whose distance values are equal to the `K`th smallest distance. If `IncludeTies` is `true`, `predict` includes all these neighbors. Otherwise, `predict` uses exactly `K` neighbors.

Example: `'IncludeTies',true`

Data Types: `logical`

`NSMethod` — Nearest neighbor search method
`'kdtree'` | `'exhaustive'`

Nearest neighbor search method, specified as the comma-separated pair consisting of `'NSMethod'` and `'kdtree'` or `'exhaustive'`.

- `'kdtree'` — Creates and uses a *k*d-tree to find nearest neighbors. `'kdtree'` is valid when the distance metric is one of the following:
  - `'euclidean'`
  - `'cityblock'`
  - `'minkowski'`
  - `'chebychev'`
- `'exhaustive'` — Uses the exhaustive search algorithm. When predicting the class of a new point `xnew`, the software computes the distance values from all points in `X` to `xnew` to find nearest neighbors.

The default is `'kdtree'` when `X` has `10` or fewer columns, `X` is not sparse or a `gpuArray`, and the distance metric is a `'kdtree'` type; otherwise, `'exhaustive'`.

Example: `'NSMethod','exhaustive'`

`NumNeighbors` — Number of nearest neighbors to find
`1` (default) | positive integer value

Number of nearest neighbors in `X` to find for classifying each point when predicting, specified as the comma-separated pair consisting of `'NumNeighbors'` and a positive integer value.

Example: `'NumNeighbors',3`

Data Types: `single` | `double`

`Scale` — Distance scale
`std(X,'omitnan')` (default) | vector of nonnegative scalar values

Distance scale, specified as the comma-separated pair consisting of `'Scale'` and a vector containing nonnegative scalar values with length equal to the number of columns in `X`. Each coordinate difference between `X` and a query point is scaled by the corresponding element of `Scale`. This argument is only valid when `'Distance'` is `'seuclidean'`.

You cannot simultaneously specify `'Standardize'` and either of `'Scale'` or `'Cov'`.

Data Types: `single` | `double`

`Standardize` — Flag to standardize predictors
`false` (default) | `true`

Flag to standardize the predictors, specified as the comma-separated pair consisting of `'Standardize'` and `true` (`1`) or `false` (`0`).

If you set `'Standardize',true`, then the software centers and scales each column of the predictor data (`X`) by the column mean and standard deviation, respectively.

The software does not standardize categorical predictors, and throws an error if all predictors are categorical.

You cannot simultaneously specify `'Standardize',1` and either of `'Scale'` or `'Cov'`.

It is good practice to standardize the predictor data.

Example: `'Standardize',true`

Data Types: `logical`

Output Arguments

`t` — *k*NN classification template

template object

*k*NN classification template suitable for training ensembles or error-correcting output code (ECOC) multiclass models, returned as a template object. Pass `t` to `fitcensemble` or `fitcecoc` to specify how to create the KNN classifier for the ensemble or ECOC model, respectively.

If you display `t` to the Command Window, then all, unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

See Also

`ClassificationKNN` | `ExhaustiveSearcher` | `fitcecoc` | `fitcensemble`

Topics

“Random Subspace Classification” on page 18-104

Introduced in R2014a

templateLinear

Linear classification learner template

Syntax

```
t = templateLinear()
t = templateLinear(Name,Value)
```

Description

`templateLinear` creates a template suitable for fitting a linear classification model to high-dimensional data for multiclass problems.

The template specifies the binary learner model, regularization type and strength, and solver, among other things. After creating the template, train the model by passing the template and data to `fitcecoc`.

`t = templateLinear()` returns a linear classification learner template.

If you specify a default template, then the software uses default values for all input arguments during training.

`t = templateLinear(Name,Value)` returns a template with additional options specified by one or more name-value pair arguments. For example, you can specify to implement logistic regression, specify the regularization type or strength, or specify the solver to use for objective-function minimization.

If you display `t` in the Command Window, then all options appear empty (`[]`) except options that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Train Multiclass Linear Classification Model

Train an ECOC model composed of multiple binary, linear classification models.

Load the NLP data set.

```
load nlpdata
```

`X` is a sparse matrix of predictor data, and `Y` is a categorical vector of class labels. There are more than two classes in the data.

Create a default linear-classification-model template.

```
t = templateLinear();
```

To adjust the default values, see the “Name-Value Pair Arguments” on page 33-6131 on `templateLinear` page.

Train an ECOC model composed of multiple binary, linear classification models that can identify the product given the frequency distribution of words on a documentation web page. For faster training time, transpose the predictor data, and specify that observations correspond to columns.

```
X = X';
rng(1); % For reproducibility
Mdl = fitcecoc(X,Y,'Learners',t,'ObservationsIn','columns')
```

```
Mdl =
  CompactClassificationECOC
      ResponseName: 'Y'
        ClassNames: [1x13 categorical]
    ScoreTransform: 'none'
    BinaryLearners: {78x1 cell}
      CodingMatrix: [13x78 double]


  Properties, Methods
```

Alternatively, you can train an ECOC model composed of default linear classification models using `'Learners','Linear'`.

To conserve memory, `fitcecoc` returns trained ECOC models composed of linear classification learners in `CompactClassificationECOC` model objects.

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Learner','logistic','Regularization','lasso','CrossVal','on'` specifies to implement logistic regression with a lasso penalty, and to implement 10-fold cross-validation.

Linear Classification Options

Lambda — Regularization term strength

`'auto'` (default) | nonnegative scalar | vector of nonnegative values

Regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and `'auto'`, a nonnegative scalar, or a vector of nonnegative values.

- For `'auto'`, `Lambda` = 1/*n*.
  - If you specify a cross-validation, name-value pair argument (e.g., `CrossVal`), then *n* is the number of in-fold observations.
  - Otherwise, *n* is the training sample size.
- For a vector of nonnegative values, `templateLinear` sequentially optimizes the objective function for each distinct value in `Lambda` in ascending order.

- If `Solver` is `'sgd'` or `'asgd'` and `Regularization` is `'lasso'`, `templateLinear` does not use the previous coefficient estimates as a warm start on page 33-1767 for the next optimization iteration. Otherwise, `templateLinear` uses warm starts.
- If `Regularization` is `'lasso'`, then any coefficient estimate of 0 retains its value when `templateLinear` optimizes using subsequent values in `Lambda`.
- `templateLinear` returns coefficient estimates for each specified regularization strength.

Example: `'Lambda',10.^(-(10:-2:2))`

Data Types: `char` | `string` | `double` | `single`

`Learner` — Linear classification model type
`'svm'` (default) | `'logistic'`

Linear classification model type, specified as the comma-separated pair consisting of `'Learner'` and `'svm'` or `'logistic'`.

In this table, $f(x) = x\beta + b$.

- β is a vector of p coefficients.
- x is an observation from p predictor variables.
- b is the scalar bias.

Value	Algorithm	Response Range	Loss Function
`'svm'`	Support vector machine	$y \in \{-1,1\}$; 1 for the positive class and –1 otherwise	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$
`'logistic'`	Logistic regression	Same as `'svm'`	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$

Example: `'Learner','logistic'`

`Regularization` — Complexity penalty type
`'lasso'` | `'ridge'`

Complexity penalty type, specified as the comma-separated pair consisting of `'Regularization'` and `'lasso'` or `'ridge'`.

The software composes the objective function for minimization from the sum of the average loss function (see `Learner`) and the regularization term in this table.

Value	Description
`'lasso'`	Lasso (L1) penalty: $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$
`'ridge'`	Ridge (L2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

To specify the regularization term strength, which is λ in the expressions, use `Lambda`.

The software excludes the bias term (β_0) from the regularization penalty.

If `Solver` is `'sparsa'`, then the default value of `Regularization` is `'lasso'`. Otherwise, the default is `'ridge'`.

Tip

- For predictor variable selection, specify `'lasso'`. For more on variable selection, see “Introduction to Feature Selection” on page 15-49.
- For optimization accuracy, specify `'ridge'`.

Example: `'Regularization','lasso'`

`Solver` — Objective function minimization technique

`'sgd'` | `'asgd'` | `'dual'` | `'bfgs'` | `'lbfgs'` | `'sparsa'` | string array | cell array of character vectors

Objective function minimization technique, specified as the comma-separated pair consisting of `'Solver'` and a character vector or string scalar, a string array, or a cell array of character vectors with values from this table.

Value	Description	Restrictions
`'sgd'`	Stochastic gradient descent (SGD) [4][2]	
`'asgd'`	Average stochastic gradient descent (ASGD) [7]	
`'dual'`	Dual SGD for SVM [1][6]	`Regularization` must be `'ridge'` and `Learner` must be `'svm'`.
`'bfgs'`	Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (BFGS) [3]	Inefficient if `X` is very high-dimensional.
`'lbfgs'`	Limited-memory BFGS (LBFGS) [3]	`Regularization` must be `'ridge'`.
`'sparsa'`	Sparse Reconstruction by Separable Approximation (SpaRSA) [5]	`Regularization` must be `'lasso'`.

If you specify:

- A ridge penalty (see `Regularization`) and the predictor data set contains 100 or fewer predictor variables, then the default solver is `'bfgs'`.
- An SVM model (see `Learner`), a ridge penalty, and the predictor data set contains more than 100 predictor variables, then the default solver is `'dual'`.
- A lasso penalty and the predictor data set contains 100 or fewer predictor variables, then the default solver is `'sparsa'`.

Otherwise, the default solver is `'sgd'`.

For more details on which solver to choose, see “Tips” on page 33-6141.

Example: `'Solver',{'sgd','lbfgs'}`

`Beta` — Initial linear coefficient estimates

`zeros(p,1)` (default) | numeric vector | numeric matrix

Initial linear coefficient estimates (β), specified as the comma-separated pair consisting of `'Beta'` and a p-dimensional numeric vector or a p-by-L numeric matrix. p is the number of predictor variables in `X` and L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a p-dimensional vector, then the software optimizes the objective function L times using this process.

  1. The software optimizes using `Beta` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The software optimizes again using the resulting estimate from the previous optimization as a warm start on page 33-1767, and the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.

- If you specify a p-by-L matrix, then the software optimizes the objective function L times. At iteration j, the software uses `Beta(:,j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.

If you set `'Solver','dual'`, then the software ignores `Beta`.

Data Types: `single` | `double`

`Bias` — Initial intercept estimate

numeric scalar | numeric vector

Initial intercept estimate (b), specified as the comma-separated pair consisting of `'Bias'` and a numeric scalar or an L-dimensional numeric vector. L is the number of regularization-strength values (for more details, see `Lambda`).

- If you specify a scalar, then the software optimizes the objective function L times using this process.

  1. The software optimizes using `Bias` as the initial value and the minimum value of `Lambda` as the regularization strength.
  2. The uses the resulting estimate as a warm start to the next optimization iteration, and uses the next smallest value in `Lambda` as the regularization strength.
  3. The software implements step 2 until it exhausts all values in `Lambda`.

- If you specify an L-dimensional vector, then the software optimizes the objective function L times. At iteration j, the software uses `Bias(j)` as the initial value and, after it sorts `Lambda` in ascending order, uses `Lambda(j)` as the regularization strength.
- By default:

  - If `Learner` is `'logistic'`, then let g_j be 1 if `Y(j)` is the positive class, and -1 otherwise. `Bias` is the weighted average of the g for training or, for cross-validation, in-fold observations.
  - If `Learner` is `'svm'`, then `Bias` is 0.

Data Types: `single` | `double`

`FitBias` — Linear model intercept inclusion flag

`true` (default) | `false`

Linear model intercept inclusion flag, specified as the comma-separated pair consisting of `'FitBias'` and `true` or `false`.

Value	Description
`true`	The software includes the bias term b in the linear model, and then estimates it.
`false`	The software sets $b = 0$ during estimation.

Example: `'FitBias',false`

Data Types: `logical`

`PostFitBias` — Flag to fit linear model intercept after optimization

`false` (default) | `true`

Flag to fit the linear model intercept after optimization, specified as the comma-separated pair consisting of `'PostFitBias'` and `true` or `false`.

Value	Description
`false`	The software estimates the bias term b and the coefficients β during optimization.
`true`	To estimate b, the software: **1** Estimates β and b using the model **2** Estimates classification scores **3** Refits b by placing the threshold on the classification scores that attains maximum accuracy

If you specify `true`, then `FitBias` must be true.

Example: `'PostFitBias',true`

Data Types: `logical`

`Verbose` — Verbosity level

`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `0` or `1`. `Verbose` controls the display of diagnostic information at the command line.

Value	Description
`0`	`templateLinear` does not display diagnostic information.
`1`	`templateLinear` periodically displays the value of the objective function, gradient magnitude, and other diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

SGD and ASGD Solver Options

`BatchSize` — Mini-batch size
positive integer

Mini-batch size, specified as the comma-separated pair consisting of `'BatchSize'` and a positive integer. At each iteration, the software estimates the gradient using `BatchSize` observations from the training data.

- If the predictor data is a numeric matrix, then the default value is `10`.
- If the predictor data is a sparse matrix, then the default value is `max([10,ceil(sqrt(ff))])`, where `ff = numel(X)/nnz(X)`, that is, the fullness factor of `X`.

Example: `'BatchSize',100`

Data Types: `single` | `double`

`LearnRate` — Learning rate
positive scalar

Learning rate, specified as the comma-separated pair consisting of `'LearnRate'` and a positive scalar. `LearnRate` controls the optimization step size by scaling the subgradient.

- If `Regularization` is `'ridge'`, then `LearnRate` specifies the initial learning rate γ_0. `templateLinear` determines the learning rate for iteration t, γ_t, using

  $$\gamma_t = \frac{\gamma_0}{(1 + \lambda \gamma_0 t)^c}.$$

  - λ is the value of `Lambda`.
  - If `Solver` is `'sgd'`, then $c = 1$.
  - If `Solver` is `'asgd'`, then c is 0.75 [7].
- If `Regularization` is `'lasso'`, then, for all iterations, `LearnRate` is constant.

By default, `LearnRate` is `1/sqrt(1+max((sum(X.^2,obsDim))))`, where `obsDim` is `1` if the observations compose the columns of the predictor data `X`, and `2` otherwise.

Example: `'LearnRate',0.01`

Data Types: `single` | `double`

`OptimizeLearnRate` — Flag to decrease learning rate
`true` (default) | `false`

Flag to decrease the learning rate when the software detects divergence (that is, over-stepping the minimum), specified as the comma-separated pair consisting of `'OptimizeLearnRate'` and `true` or `false`.

If `OptimizeLearnRate` is `'true'`, then:

1. For the few optimization iterations, the software starts optimization using `LearnRate` as the learning rate.
2. If the value of the objective function increases, then the software restarts and uses half of the current value of the learning rate.

3 The software iterates step 2 until the objective function decreases.

Example: `'OptimizeLearnRate',true`

Data Types: `logical`

`TruncationPeriod` — Number of mini-batches between lasso truncation runs

`10` (default) | positive integer

Number of mini-batches between lasso truncation runs, specified as the comma-separated pair consisting of `'TruncationPeriod'` and a positive integer.

After a truncation run, the software applies a soft threshold to the linear coefficients. That is, after processing k = `TruncationPeriod` mini-batches, the software truncates the estimated coefficient j using

$$\widehat{\beta}_j^* = \begin{cases} \widehat{\beta}_j - u_t & \text{if } \widehat{\beta}_j > u_t, \\ 0 & \text{if } \left|\widehat{\beta}_j\right| \le u_t, \\ \widehat{\beta}_j + u_t & \text{if } \widehat{\beta}_j < -u_t. \end{cases}$$

- For SGD, $\widehat{\beta}_j$ is the estimate of coefficient j after processing k mini-batches. $u_t = k\gamma_t\lambda$. γ_t is the learning rate at iteration t. λ is the value of `Lambda`.
- For ASGD, $\widehat{\beta}_j$ is the averaged estimate coefficient j after processing k mini-batches, $u_t = k\lambda$.

If `Regularization` is `'ridge'`, then the software ignores `TruncationPeriod`.

Example: `'TruncationPeriod',100`

Data Types: `single` | `double`

SGD and ASGD Convergence Controls

`BatchLimit` — Maximal number of batches

positive integer

Maximal number of batches to process, specified as the comma-separated pair consisting of `'BatchLimit'` and a positive integer. When the software processes `BatchLimit` batches, it terminates optimization.

- By default:
  - The software passes through the data `PassLimit` times.
  - If you specify multiple solvers, and use (A)SGD to get an initial approximation for the next solver, then the default value is `ceil(1e6/BatchSize)`. `BatchSize` is the value of the `'BatchSize'` name-value pair argument.
- If you specify `'BatchLimit'` and `'PassLimit'`, then the software chooses the argument that results in processing the fewest observations.
- If you specify `'BatchLimit'` but not `'PassLimit'`, then the software processes enough batches to complete up to one entire pass through the data.

Example: `'BatchLimit',100`

Data Types: `single` | `double`

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`NumCheckConvergence` — Number of batches to process before next convergence check
positive integer

Number of batches to process before next convergence check, specified as the comma-separated pair consisting of `'NumCheckConvergence'` and a positive integer.

To specify the batch size, see `BatchSize`.

The software checks for convergence about 10 times per pass through the entire data set by default.

Example: `'NumCheckConvergence',100`

Data Types: `single` | `double`

`PassLimit` — Maximal number of passes
`1` (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer.

The software processes all observations when it completes one pass through the data.

When the software passes through the data `PassLimit` times, it terminates optimization.

If you specify `'BatchLimit'` and `PassLimit`, then the software chooses the argument that results in processing the fewest observations.

Example: `'PassLimit',5`

Data Types: `single` | `double`

Dual SGD Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `DeltaGradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`DeltaGradientTolerance` — Gradient-difference tolerance
`1` (default) | nonnegative scalar

Gradient-difference tolerance between upper and lower pool Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862 violators, specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar.

- If the magnitude of the KKT violators is less than `DeltaGradientTolerance`, then the software terminates optimization.
- If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'DeltaGapTolerance',1e-2`

Data Types: `double` | `single`

`NumCheckConvergence` — Number of passes through entire data set to process before next convergence check
`5` (default) | positive integer

Number of passes through entire data set to process before next convergence check, specified as the comma-separated pair consisting of `'NumCheckConvergence'` and a positive integer.

Example: `'NumCheckConvergence',100`

Data Types: `single` | `double`

`PassLimit` — Maximal number of passes
`10` (default) | positive integer

Maximal number of passes through the data, specified as the comma-separated pair consisting of `'PassLimit'` and a positive integer.

When the software completes one pass through the data, it has processed all observations.

When the software passes through the data `PassLimit` times, it terminates optimization.

Example: `'PassLimit',5`

Data Types: `single` | `double`

BFGS, LBFGS, and SpaRSA Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e-4` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in `Solver`, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

If the software converges for the last solver specified in the software, then optimization terminates. Otherwise, the software uses the next solver specified in `Solver`.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. That is, at each iteration, the software composes the Hessian using statistics from the latest `HessianHistorySize` iterations.

The software does not support `'HessianHistorySize'` for SpaRSA.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`IterationLimit` — Maximal number of optimization iterations
`1000` (default) | positive integer

Maximal number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer. `IterationLimit` applies to these values of `Solver`: `'bfgs'`, `'lbfgs'`, and `'sparsa'`.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Output Arguments

`t` — Linear classification model learner template

template object

Linear classification model learner template, returned as a template object. To train a linear classification model using high-dimensional data for multiclass problems, pass `t` to `fitcecoc`.

If you display `t` to the Command Window, then all, unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

More About

Warm Start

A warm start is initial estimates of the beta coefficients and bias term supplied to an optimization routine for quicker convergence.

Tips

- It is a best practice to orient your predictor matrix so that observations correspond to columns and to specify `'ObservationsIn','columns'`. As a result, you can experience a significant reduction in optimization-execution time.
- For better optimization accuracy if the predictor data is high-dimensional and `Regularization` is `'ridge'`, set any of these combinations for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'dual'` if `Learner` is `'svm'`
  - `{'sgd','lbfgs'}`
  - `{'asgd','lbfgs'}`
  - `{'dual','lbfgs'}` if `Learner` is `'svm'`

  Other combinations can result in poor optimization accuracy.
- For better optimization accuracy if the predictor data is moderate- through low-dimensional and `Regularization` is `'ridge'`, set `Solver` to `'bfgs'`.
- If `Regularization` is `'lasso'`, set any of these combinations for `Solver`:
  - `'sgd'`
  - `'asgd'`
  - `'sparsa'`
  - `{'sgd','sparsa'}`
  - `{'asgd','sparsa'}`
- When choosing between SGD and ASGD, consider that:
  - SGD takes less time per iteration, but requires more iterations to converge.
  - ASGD requires fewer iterations to converge, but takes more time per iteration.

- If the predictor data has few observations, but many predictor variables, then:
  - Specify `'PostFitBias',true`.
  - For SGD or ASGD solvers, set `PassLimit` to a positive integer that is greater than 1, for example, 5 or 10. This setting often results in better accuracy.
- For SGD and ASGD solvers, `BatchSize` affects the rate of convergence.
  - If `BatchSize` is too small, then the software achieves the minimum in many iterations, but computes the gradient per iteration quickly.
  - If `BatchSize` is too large, then the software achieves the minimum in fewer iterations, but computes the gradient per iteration slowly.
- Large learning rate (see `LearnRate`) speed-up convergence to the minimum, but can lead to divergence (that is, over-stepping the minimum). Small learning rates ensure convergence to the minimum, but can lead to slow termination.
- If `Regularization` is `'lasso'`, then experiment with various values of `TruncationPeriod`. For example, set `TruncationPeriod` to `1`, `10`, and then `100`.
- For efficiency, the software does not standardize predictor data. To standardize the predictor data (`X`), enter

  ```
  X = bsxfun(@rdivide,bsxfun(@minus,X,mean(X,2)),std(X,0,2));
  ```

  The code requires that you orient the predictors and observations as the rows and columns of `X`, respectively. Also, for memory-usage economy, the code replaces the original predictor data the standardized data.

References

[1] Hsieh, C. J., K. W. Chang, C. J. Lin, S. S. Keerthi, and S. Sundararajan. “A Dual Coordinate Descent Method for Large-Scale Linear SVM.” *Proceedings of the 25th International Conference on Machine Learning, ICML ’08*, 2001, pp. 408–415.

[2] Langford, J., L. Li, and T. Zhang. “Sparse Online Learning Via Truncated Gradient.” *J. Mach. Learn. Res.*, Vol. 10, 2009, pp. 777–801.

[3] Nocedal, J. and S. J. Wright. *Numerical Optimization*, 2nd ed., New York: Springer, 2006.

[4] Shalev-Shwartz, S., Y. Singer, and N. Srebro. “Pegasos: Primal Estimated Sub-Gradient Solver for SVM.” *Proceedings of the 24th International Conference on Machine Learning, ICML ’07*, 2007, pp. 807–814.

[5] Wright, S. J., R. D. Nowak, and M. A. T. Figueiredo. “Sparse Reconstruction by Separable Approximation.” *Trans. Sig. Proc.*, Vol. 57, No 7, 2009, pp. 2479–2493.

[6] Xiao, Lin. “Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization.” *J. Mach. Learn. Res.*, Vol. 11, 2010, pp. 2543–2596.

[7] Xu, Wei. “Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent.” *CoRR*, abs/1107.2490, 2011.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations when you train a model by passing a linear model template and tall arrays to `fitcecoc`:

- The default values for these name-value pair arguments are different when you work with tall arrays.
  - `'Lambda'` — Can be `'auto'` (default) or a scalar
  - `'Regularization'` — Supports only `'ridge'`
  - `'Solver'` — Supports only `'lbfgs'`
  - `'FitBias'` — Supports only `true`
  - `'Verbose'` — Default value is `1`
  - `'BetaTolerance'` — Default value is relaxed to `1e–3`
  - `'GradientTolerance'` — Default value is relaxed to `1e–3`
  - `'IterationLimit'` — Default value is relaxed to `20`
- When `fitcecoc` uses a `templateLinear` object with tall arrays, the only available solver is LBFGS. The software implements LBFGS by distributing the calculation of the loss and gradient among different parts of the tall array at each iteration. If you do not specify initial values for `Beta` and `Bias`, the software refines the initial estimates of the parameters by fitting the model locally to parts of the data and combining the coefficients by averaging.

For more information, see “Tall Arrays”.

See Also

`fitcecoc` | `fitclinear` | `fitrlinear`

Introduced in R2016a

templateNaiveBayes

Naive Bayes classifier template

Syntax

```
t = templateNaiveBayes()
t = templateNaiveBayes(Name,Value)
```

Description

`t = templateNaiveBayes()` returns a naive Bayes on page 33-6148 template suitable for training error-correcting output code (ECOC) multiclass models.

If you specify a default template, then the software uses default values for all input arguments during training.

Specify `t` as a learner in `fitcecoc`.

`t = templateNaiveBayes(Name,Value)` returns a template with additional options specified by one or more name-value pair arguments. All properties of `t` are empty, except those you specify using `Name,Value` pair arguments.

For example, you can specify distributions for the predictors.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a Default Naive Bayes Template

Use `templateNaiveBayes` to specify a default naive Bayes template.

```
t = templateNaiveBayes()
```

```
t =
Fit template for classification NaiveBayes.

    DistributionNames: [1x0 double]
               Kernel: []
              Support: []
                Width: []
              Version: 1
               Method: 'NaiveBayes'
                 Type: 'classification'
```

All properties of the template object are empty except for `Method` and `Type`. When you pass `t` to the training function, the software fills in the empty properties with their respective default values. For

example, the software fills the `DistributionNames` property with a 1-by- `D` cell array of character vectors with `'normal'` in each cell, where `D` is the number of predictors. For details on other default values, see `fitcnb`.

`t` is a plan for a naive Bayes learner, and no computation occurs when you specify it. You can pass `t` to `fitcecoc` to specify naive Bayes binary learners for ECOC multiclass learning.

Create a Naive Bayes Template for ECOC Multiclass Learning

Create a nondefault naive Bayes template for use in `fitcecoc`.

Load Fisher's iris data set.

```
load fisheriris
```

Create a template for naive Bayes binary classifiers, and specify kernel distributions for all predictors.

```
t = templateNaiveBayes('DistributionNames','kernel')
```

```
t =
Fit template for classification NaiveBayes.

    DistributionNames: 'kernel'
               Kernel: []
              Support: []
                Width: []
              Version: 1
               Method: 'NaiveBayes'
                 Type: 'classification'
```

All properties of the template object are empty except for `DistributionNames`, `Method`, and `Type`. When you pass `t` to the training function, the software fills in the empty properties with their respective default values.

Specify `t` as a binary learner for an ECOC multiclass model.

```
Mdl = fitcecoc(meas,species,'Learners',t);
```

By default, the software trains `Mdl` using the one-versus-one coding design.

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl,'LossFun','classiferror')
```

```
L = 0.0333
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'DistributionNames','mn'` specifies to treat all predictors as token counts for a multinomial model.

`DistributionNames` — Data distributions
`'kernel'` | `'mn'` | `'mvmn'` | `'normal'` | string array | cell array of character vectors

Data distributions `fitcnb` uses to model the data, specified as the comma-separated pair consisting of `'DistributionNames'` and a character vector or string scalar, a string array, or a cell array of character vectors with values from this table.

Value	Description
`'kernel'`	Kernel smoothing density estimate.
`'mn'`	Multinomial distribution. If you specify `mn`, then all features are components of a multinomial distribution. Therefore, you cannot include `'mn'` as an element of a string array or a cell array of character vectors. For details, see “Algorithms” on page 33-1799.
`'mvmn'`	Multivariate multinomial distribution. For details, see “Algorithms” on page 33-1799.
`'normal'`	Normal (Gaussian) distribution.

If you specify a character vector or string scalar, then the software models all the features using that distribution. If you specify a 1-by-*P* string array or cell array of character vectors, then the software models feature *j* using the distribution in element *j* of the array.

By default, the software sets all predictors specified as categorical predictors (using the `CategoricalPredictors` name-value pair argument) to `'mvmn'`. Otherwise, the default distribution is `'normal'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'DistributionNames','mn'`

Example: `'DistributionNames',{'kernel','normal','kernel'}`

`Kernel` — Kernel smoother type
`'normal'` (default) | `'box'` | `'epanechnikov'` | `'triangle'` | string array | cell array of character vectors

Kernel smoother type, specified as the comma-separated pair consisting of `'Kernel'` and a character vector or string scalar, a string array, or a cell array of character vectors.

This table summarizes the available options for setting the kernel smoothing density region. Let $I\{u\}$ denote the indicator function.

Value	Kernel	Formula
`'box'`	Box (uniform)	$f(x) = 0.5I\{\|x\| \le 1\}$
`'epanechnikov'`	Epanechnikov	$f(x) = 0.75\left(1 - x^2\right)I\{\|x\| \le 1\}$
`'normal'`	Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-0.5x^2\right)$
`'triangle'`	Triangular	$f(x) = (1 - \|x\|)I\{\|x\| \le 1\}$

If you specify a 1-by-*P* string array or cell array, with each element of the array containing any value in the table, then the software trains the classifier using the kernel smoother type in element *j* for feature *j* in `X`. The software ignores elements of `Kernel` not corresponding to a predictor whose distribution is `'kernel'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'Kernel',{'epanechnikov','normal'}`

`Support` — Kernel smoothing density support
`'unbounded'` (default) | `'positive'` | string array | cell array | numeric row vector

Kernel smoothing density support, specified as the comma-separated pair consisting of `'Support'` and `'positive'`, `'unbounded'`, a string array, a cell array, or a numeric row vector. The software applies the kernel smoothing density to the specified region.

This table summarizes the available options for setting the kernel smoothing density region.

Value	Description
1-by-2 numeric row vector	For example, `[L,U]`, where `L` and `U` are the finite lower and upper bounds, respectively, for the density support.
`'positive'`	The density support is all positive real values.
`'unbounded'`	The density support is all real values.

If you specify a 1-by-*P* string array or cell array, with each element in the string array containing any text value in the table and each element in the cell array containing any value in the table, then the software trains the classifier using the kernel support in element *j* for feature *j* in `X`. The software ignores elements of `Kernel` not corresponding to a predictor whose distribution is `'kernel'`.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'KSSupport',{[-10,20],'unbounded'}`

Data Types: `char` | `string` | `cell` | `double`

`Width` — Kernel smoothing window width
matrix of numeric values | numeric column vector | numeric row vector | scalar

Kernel smoothing window width, specified as the comma-separated pair consisting of `'Width'` and a matrix of numeric values, numeric column vector, numeric row vector, or scalar.

Suppose there are *K* class levels and *P* predictors. This table summarizes the available options for setting the kernel smoothing window width.

Value	Description
K-by-*P* matrix of numeric values	Element (*k*,*j*) specifies the width for predictor *j* in class *k*.
K-by-1 numeric column vector	Element *k* specifies the width for all predictors in class *k*.
1-by-*P* numeric row vector	Element *j* specifies the width in all class levels for predictor *j*.
scalar	Specifies the bandwidth for all features in all classes.

By default, the software selects a default width automatically for each combination of predictor and class by using a value that is optimal for a Gaussian distribution. If you specify `Width` and it contains `NaN`s, then the software selects widths for the elements containing `NaN`s.

You must specify that at least one predictor has distribution `'kernel'` to additionally specify `Kernel`, `Support`, or `Width`.

Example: `'Width',[NaN NaN]`

Data Types: `double` | `struct`

Output Arguments

`t` — Naive Bayes classification template

template object

Naive Bayes classification template suitable for training error-correcting output code (ECOC) multiclass models, returned as a template object. Pass `t` to `fitcecoc` to specify how to create the naive Bayes classifier for the ECOC model.

If you display `t` to the Command Window, then all, unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

More About

Naive Bayes

Naive Bayes is a classification algorithm that applies density estimation to the data.

The algorithm leverages Bayes theorem, and (naively) assumes that the predictors are conditionally independent, given the class. Although the assumption is usually violated in practice, naive Bayes classifiers tend to yield posterior distributions that are robust to biased class density estimates, particularly where the posterior is 0.5 (the decision boundary) [1].

Naive Bayes classifiers assign observations to the most probable class (in other words, the maximum a posteriori decision rule). Explicitly, the algorithm takes these steps:

1. Estimate the densities of the predictors within each class.
2. Model posterior probabilities according to Bayes rule. That is, for all $k = 1,...,K$,

$$\widehat{P}(Y = k \mid X_1, .., X_P) = \frac{\pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)}{\sum_{k=1}^{K} \pi(Y = k) \prod_{j=1}^{P} P\left(X_j \middle| Y = k\right)},$$

where:

- Y is the random variable corresponding to the class index of an observation.
- $X_1,...,X_P$ are the random predictors of an observation.
- $\pi(Y = k)$ is the prior probability that a class index is k.

3 Classify an observation by estimating the posterior probability for each class, and then assign the observation to the class yielding the maximum posterior probability.

If the predictors compose a multinomial distribution, then the posterior probability $\widehat{P}(Y = k|X_1, .., X_P) \propto \pi(Y = k)P_{mn}(X_1, ..., X_P|Y = k)$, where $P_{mn}(X_1, ..., X_P|Y = k)$ is the probability mass function of a multinomial distribution.

Algorithms

- If you specify `'DistributionNames','mn'` when training `Mdl` using `fitcnb`, then the software fits a multinomial distribution using the bag-of-tokens model on page 33-1798. The software stores the probability that token j appears in class k in the property `DistributionParameters{k,j}`. Using additive smoothing [2], the estimated probability is

  $$P(\text{token } j \,|\, \text{class } k) = \frac{1 + c_{j|k}}{P + c_k},$$

  where:

  - $c_{j|k} = n_k \frac{\sum_{i:\, y_i \in \text{class } k} x_{ij} w_i}{\sum_{i:\, y_i \in \text{class } k} w_i}$, which is the weighted number of occurrences of token j in class k.
  - n_k is the number of observations in class k.
  - w_i is the weight for observation i. The software normalizes weights within a class such that they sum to the prior probability for that class.
  - $c_k = \sum_{j=1}^{P} c_{j|k}$, which is the total weighted number of occurrences of all tokens in class k.
- If you specify `'DistributionNames','mvmn'` when training `Mdl` using `fitcnb`, then:

  1 For each predictor, the software collects a list of the unique levels, stores the sorted list in `CategoricalLevels`, and considers each level a bin. Each predictor/class combination is a separate, independent multinomial random variable.

  2 For predictor j in class k, the software counts instances of each categorical level using the list stored in `CategoricalLevels{j}`.

  3 The software stores the probability that predictor j, in class k, has level L in the property `DistributionParameters{k,j}`, for all levels in `CategoricalLevels{j}`. Using additive smoothing [2], the estimated probability is

  $$P(\text{predictor } j = L \,|\, \text{class } k) = \frac{1 + m_{j|k}(L)}{m_j + m_k},$$

  where:

- $m_{j|k}(L) = n_k \frac{\sum_{i:y_i \in \text{class } k} I\{x_{ij} = L\} w_i}{\sum_{i:y_i \in \text{class } k} w_i}$, which is the weighted number of observations for which predictor j equals L in class k.
- n_k is the number of observations in class k.
- $I\{x_{ij} = L\} = 1$ if $x_{ij} = L$, 0 otherwise.
- w_i is the weight for observation i. The software normalizes weights within a class such that they sum to the prior probability for that class.
- m_j is the number of distinct levels in predictor j.
- m_k is the weighted number of observations in class k.

References

[1] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[2] Manning, C. D., P. Raghavan, and M. Schütze. *Introduction to Information Retrieval*, NY: Cambridge University Press, 2008.

See Also

`ClassificationECOC` | `ClassificationNaiveBayes` | `fitcecoc` | `fitcnb`

Introduced in R2014b

templateSVM

Support vector machine template

Syntax

```
t = templateSVM()
t = templateSVM(Name,Value)
```

Description

`t = templateSVM()` returns a support vector machine (SVM) learner template suitable for training error-correcting output code (ECOC) multiclass models.

If you specify a default template, then the software uses default values for all input arguments during training.

Specify `t` as a binary learner, or one in a set of binary learners, in `fitcecoc` to train an ECOC multiclass classifier.

`t = templateSVM(Name,Value)` returns a template with additional options specified by one or more name-value pair arguments.

For example, you can specify the box constraint, the kernel function, or whether to standardize the predictors.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a Default Support Vector Machine Template

Use `templateSVM` to specify a default SVM template.

```
t = templateSVM()
```

```
t =
Fit template for classification SVM.

                    Alpha: [0x1 double]
             BoxConstraint: []
                 CacheSize: []
             CachingMethod: ''
                ClipAlphas: []
    DeltaGradientTolerance: []
                   Epsilon: []
              GapTolerance: []
              KKTTolerance: []
            IterationLimit: []
```

```
           KernelFunction: ''
              KernelScale: []
             KernelOffset: []
     KernelPolynomialOrder: []
                 NumPrint: []
                       Nu: []
          OutlierFraction: []
        RemoveDuplicates: []
         ShrinkagePeriod: []
                  Solver: ''
         StandardizeData: []
      SaveSupportVectors: []
          VerbosityLevel: []
                 Version: 2
                  Method: 'SVM'
                    Type: 'classification'
```

All properties of the template object are empty except for `Method` and `Type`. When you pass `t` to the training function, the software fills in the empty properties with their respective default values. For example, the software fills the `KernelFunction` property with `'linear'`. For details on other default values, see `fitcsvm`.

`t` is a plan for an SVM learner, and no computation occurs when you specify it. You can pass `t` to `fitcecoc` to specify SVM binary learners for ECOC multiclass learning. However, by default, `fitcecoc` uses default SVM binary learners.

Create an SVM Template for ECOC Multiclass Learning

Create a nondefault SVM template for use in `fitcecoc`.

Load Fisher's iris data set.

```
load fisheriris
```

Create a template for SVM binary classifiers, and specify to use a Gaussian kernel function.

```
t = templateSVM('KernelFunction','gaussian')
```

```
t =
Fit template for classification SVM.

                    Alpha: [0x1 double]
             BoxConstraint: []
                 CacheSize: []
             CachingMethod: ''
                ClipAlphas: []
    DeltaGradientTolerance: []
                   Epsilon: []
             GapTolerance: []
             KKTTolerance: []
            IterationLimit: []
            KernelFunction: 'gaussian'
               KernelScale: []
              KernelOffset: []
```

```
    KernelPolynomialOrder: []
                  NumPrint: []
                       Nu: []
            OutlierFraction: []
          RemoveDuplicates: []
           ShrinkagePeriod: []
                    Solver: ''
           StandardizeData: []
       SaveSupportVectors: []
           VerbosityLevel: []
                  Version: 2
                   Method: 'SVM'
                     Type: 'classification'
```

All properties of the template object are empty except for `DistributionNames`, `Method`, and `Type`. When trained on, the software fills in the empty properties with their respective default values.

Specify `t` as a binary learner for an ECOC multiclass model.

```
Mdl = fitcecoc(meas,species,'Learners',t);
```

`Mdl` is a `ClassificationECOC` multiclass classifier. By default, the software trains `Mdl` using the one-versus-one coding design.

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl,'LossFun','classiferror')
```

```
L = 0.0200
```

Retain and Discard Support Vectors of SVM Binary Learners

When you train an ECOC model with linear SVM binary learners, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties of the binary learners by default. You can choose instead to retain the support vectors and related values, and then discard them from the model later.

Load Fisher's iris data set.

```
load fisheriris
rng(1); % For reproducibility
```

Train an ECOC model using the entire data set. Specify retaining the support vectors by passing in the appropriate SVM template.

```
t = templateSVM('SaveSupportVectors',true);
MdlSV = fitcecoc(meas,species,'Learners',t);
```

`MdlSV` is a trained `ClassificationECOC` model with linear SVM binary learners. By default, `fitcecoc` implements a one-versus-one coding design, which requires three binary learners for three-class learning.

Access the estimated α (alpha) values using dot notation.

```
alpha = cell(3,1);
alpha{1} = MdlSV.BinaryLearners{1}.Alpha;
alpha{2} = MdlSV.BinaryLearners{2}.Alpha;
alpha{3} = MdlSV.BinaryLearners{3}.Alpha;
alpha
```

```
alpha=3×1 cell array
    { 3x1 double}
    { 3x1 double}
    {23x1 double}
```

`alpha` is a 3-by-1 cell array that stores the estimated values of α.

Discard the support vectors and related values from the ECOC model.

```
Mdl = discardSupportVectors(MdlSV);
```

`Mdl` is similar to `MdlSV`, except that the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties of all the linear SVM binary learners are empty (`[]`).

```
areAllEmpty = @(x)isempty([x.Alpha x.SupportVectors x.SupportVectorLabels]);
cellfun(areAllEmpty,Mdl.BinaryLearners)
```

```
ans = 3x1 logical array

   1
   1
   1
```

Compare the sizes of the two ECOC models.

```
vars = whos('Mdl','MdlSV');
100*(1 - vars(1).bytes/vars(2).bytes)
```

```
ans = 4.7075
```

`Mdl` is about 5% smaller than `MdlSV`.

Reduce your memory usage by compacting `Mdl` and then clearing `Mdl` and `MdlSV` from the workspace.

```
CompactMdl = compact(Mdl);
clear Mdl MdlSV;
```

Predict the label for a random row of the training data using the more efficient SVM model.

```
idx = randsample(size(meas,1),1)
```

```
idx = 63
```

```
predictedLabel = predict(CompactMdl,meas(idx,:))
```

```
predictedLabel = 1x1 cell array
    {'versicolor'}
```

```
trueLabel = species(idx)
```

```
trueLabel = 1x1 cell array
    {'versicolor'}
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'BoxConstraint',0.1,'KernelFunction','gaussian','Standardize',1` specifies a box constraint of `0.1`, to use the Gaussian (RBF) kernel, and to standardize the predictors.

`BoxConstraint` — Box constraint
1 (default) | positive scalar

Box constraint on page 33-1862, specified as the comma-separated pair consisting of `'BoxConstraint'` and a positive scalar.

For one-class learning, the software always sets the box constraint to `1`.

For more details on the relationships and algorithmic behavior of `BoxConstraint`, `Cost`, `Prior`, `Standardize`, and `Weights`, see "Algorithms" on page 33-1865.

Example: `'BoxConstraint',100`

Data Types: `double` | `single`

`CacheSize` — Cache size
`1000` (default) | `'maximal'` | positive scalar

Cache size, specified as the comma-separated pair consisting of `'CacheSize'` and `'maximal'` or a positive scalar.

If `CacheSize` is `'maximal'`, then the software reserves enough memory to hold the entire *n*-by-*n* Gram matrix on page 33-1862.

If `CacheSize` is a positive scalar, then the software reserves `CacheSize` megabytes of memory for training the model.

Example: `'CacheSize','maximal'`

Data Types: `double` | `single` | `char` | `string`

`ClipAlphas` — Flag to clip alpha coefficients
`true` (default) | `false`

Flag to clip alpha coefficients, specified as the comma-separated pair consisting of `'ClipAlphas'` and either `true` or `false`.

Suppose that the alpha coefficient for observation j is α_j and the box constraint of observation j is C_j, $j = 1,...,n$, where n is the training sample size.

Value	Description
`true`	At each iteration, if α_j is near 0 or near C_j, then MATLAB sets α_j to 0 or to C_j, respectively.
`false`	MATLAB does not change the alpha coefficients during optimization.

MATLAB stores the final values of α in the `Alpha` property of the trained SVM model object.

`ClipAlphas` can affect SMO and ISDA convergence.

Example: `'ClipAlphas',false`

Data Types: `logical`

`DeltaGradientTolerance` — Tolerance for gradient difference
nonnegative scalar

Tolerance for the gradient difference between upper and lower violators obtained by Sequential Minimal Optimization (SMO) or Iterative Single Data Algorithm (ISDA), specified as the comma-separated pair consisting of `'DeltaGradientTolerance'` and a nonnegative scalar.

If `DeltaGradientTolerance` is `0`, then the software does not use the tolerance for the gradient difference to check for optimization convergence.

The default values are:

- `1e-3` if the solver is SMO (for example, you set `'Solver','SMO'`)
- `0` if the solver is ISDA (for example, you set `'Solver','ISDA'`)

Example: `'DeltaGradientTolerance',1e-2`

Data Types: `double` | `single`

`GapTolerance` — Feasibility gap tolerance
`0` (default) | nonnegative scalar

Feasibility gap tolerance obtained by SMO or ISDA, specified as the comma-separated pair consisting of `'GapTolerance'` and a nonnegative scalar.

If `GapTolerance` is `0`, then the software does not use the feasibility gap tolerance to check for optimization convergence.

Example: `'GapTolerance',1e-2`

Data Types: `double` | `single`

`IterationLimit` — Maximal number of numerical optimization iterations
`1e6` (default) | positive integer

Maximal number of numerical optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The software returns a trained model regardless of whether the optimization routine successfully converges. `Mdl.ConvergenceInfo` contains convergence information.

Example: `'IterationLimit',1e8`

Data Types: `double` | `single`

`KernelFunction` — Kernel function

`'linear'` | `'gaussian'` | `'rbf'` | `'polynomial'` | function name

Kernel function used to compute the elements of the Gram matrix on page 33-1862, specified as the comma-separated pair consisting of `'KernelFunction'` and a kernel function name. Suppose $G(x_j,x_k)$ is element (j,k) of the Gram matrix, where x_j and x_k are p-dimensional vectors representing observations j and k in `X`. This table describes supported kernel function names and their functional forms.

Kernel Function Name	Description	Formula
`'gaussian'` or `'rbf'`	Gaussian or Radial Basis Function (RBF) kernel, default for one-class learning	$G(x_j, x_k) = \exp\left(-\|x_j - x_k\|^2\right)$
`'linear'`	Linear kernel, default for two-class learning	$G(x_j, x_k) = x_j' x_k$
`'polynomial'`	Polynomial kernel. Use `'PolynomialOrder',q` to specify a polynomial kernel of order q.	$G(x_j, x_k) = (1 + x_j' x_k)^q$

You can set your own kernel function, for example, `kernel`, by setting `'KernelFunction','kernel'`. The value `kernel` must have this form.

```
function G = kernel(U,V)
```

where:

- `U` is an m-by-p matrix. Columns correspond to predictor variables, and rows correspond to observations.
- `V` is an n-by-p matrix. Columns correspond to predictor variables, and rows correspond to observations.
- `G` is an m-by-n Gram matrix on page 33-1862 of the rows of `U` and `V`.

`kernel.m` must be on the MATLAB path.

It is a good practice to avoid using generic names for kernel functions. For example, call a sigmoid kernel function `'mysigmoid'` rather than `'sigmoid'`.

Example: `'KernelFunction','gaussian'`

Data Types: `char` | `string`

`KernelOffset` — Kernel offset parameter

nonnegative scalar

Kernel offset parameter, specified as the comma-separated pair consisting of `'KernelOffset'` and a nonnegative scalar.

The software adds `KernelOffset` to each element of the Gram matrix.

The defaults are:

- `0` if the solver is SMO (that is, you set `'Solver','SMO'`)

- `0.1` if the solver is ISDA (that is, you set `'Solver','ISDA'`)

Example: `'KernelOffset',0`

Data Types: `double` | `single`

`KernelScale` — Kernel scale parameter
`1` (default) | `'auto'` | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. The software divides all elements of the predictor matrix `X` by the value of `KernelScale`. Then, the software applies the appropriate kernel norm to compute the Gram matrix.

- If you specify `'auto'`, then the software selects an appropriate scale factor using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed using `rng` before training.
- If you specify `KernelScale` and your own kernel function, for example, `'KernelFunction','kernel'`, then the software throws an error. You must apply scaling within `kernel`.

Example: `'KernelScale','auto'`

Data Types: `double` | `single` | `char` | `string`

`KKTTolerance` — Karush-Kuhn-Tucker complementarity conditions violation tolerance
nonnegative scalar

Karush-Kuhn-Tucker (KKT) complementarity conditions on page 33-1862 violation tolerance, specified as the comma-separated pair consisting of `'KKTTolerance'` and a nonnegative scalar.

If `KKTTolerance` is `0`, then the software does not use the KKT complementarity conditions violation tolerance to check for optimization convergence.

The default values are:

- `0` if the solver is SMO (for example, you set `'Solver','SMO'`)
- `1e-3` if the solver is ISDA (for example, you set `'Solver','ISDA'`)

Example: `'KKTTolerance',1e-2`

Data Types: `double` | `single`

`NumPrint` — Number of iterations between optimization diagnostic message output
`1000` (default) | nonnegative integer

Number of iterations between optimization diagnostic message output, specified as the comma-separated pair consisting of `'NumPrint'` and a nonnegative integer.

If you specify `'Verbose',1` and `'NumPrint',numprint`, then the software displays all optimization diagnostic messages from SMO and ISDA every `numprint` iterations in the Command Window.

Example: `'NumPrint',500`

Data Types: `double` | `single`

`OutlierFraction` — Expected proportion of outliers in training data
`0` (default) | numeric scalar in the interval [0,1)

Expected proportion of outliers in the training data, specified as the comma-separated pair consisting of `'OutlierFraction'` and a numeric scalar in the interval [0,1).

Suppose that you set `'OutlierFraction',outlierfraction`, where `outlierfraction` is a value greater than 0.

- For two-class learning, the software implements robust learning. In other words, the software attempts to remove 100*`outlierfraction`% of the observations when the optimization algorithm converges. The removed observations correspond to gradients that are large in magnitude.
- For one-class learning, the software finds an appropriate bias term such that `outlierfraction` of the observations in the training set have negative scores.

Example: `'OutlierFraction',0.01`

Data Types: `double` | `single`

`PolynomialOrder` — Polynomial kernel function order
3 (default) | positive integer

Polynomial kernel function order, specified as the comma-separated pair consisting of `'PolynomialOrder'` and a positive integer.

If you set `'PolynomialOrder'` and `KernelFunction` is not `'polynomial'`, then the software throws an error.

Example: `'PolynomialOrder',2`

Data Types: `double` | `single`

`SaveSupportVectors` — Store support vectors, their labels, and the estimated *α* coefficients
`true` | `false`

Store support vectors, their labels, and the estimated *α* coefficients as properties of the resulting model, specified as the comma-separated pair consisting of `'SaveSupportVectors'` and `true` or `false`.

If `SaveSupportVectors` is `true`, the resulting model stores the support vectors in the `SupportVectors` property, their labels in the `SupportVectorLabels` property, and the estimated *α* coefficients in the `Alpha` property of the compact, SVM learners.

If `SaveSupportVectors` is `false` and `KernelFunction` is `'linear'`, the resulting model does not store the support vectors and the related estimates.

To reduce memory consumption by compact SVM models, specify `SaveSupportVectors`.

For linear, SVM binary learners in an ECOC model, the default value is `false`. Otherwise, the default value is `true`.

Example: `'SaveSupportVectors',true`

Data Types: `logical`

ShrinkagePeriod — Number of iterations between reductions of active set
0 (default) | nonnegative integer

Number of iterations between reductions of the active set, specified as the comma-separated pair consisting of 'ShrinkagePeriod' and a nonnegative integer.

If you set 'ShrinkagePeriod',0, then the software does not shrink the active set.

Example: 'ShrinkagePeriod',1000

Data Types: double | single

Solver — Optimization routine
'ISDA' | 'L1QP' | 'SMO'

Optimization routine, specified as the comma-separated pair consisting of 'Solver' and a value in this table.

Value	Description
'ISDA'	Iterative Single Data Algorithm (see [30])
'L1QP'	Uses quadprog to implement $L1$ soft-margin minimization by quadratic programming. This option requires an Optimization Toolbox license. For more details, see "Quadratic Programming Definition" (Optimization Toolbox).
'SMO'	Sequential Minimal Optimization (see [17])

The default value is 'ISDA' if you set 'OutlierFraction' to a positive value for two-class learning, and 'SMO' otherwise.

Example: 'Solver','ISDA'

Standardize — Flag to standardize predictor data
false (default) | true

Flag to standardize the predictor data, specified as the comma-separated pair consisting of 'Standardize' and true (1) or false (0).

If you set 'Standardize',true:

- The software centers and scales each column of the predictor data (X) by the weighted column mean and standard deviation, respectively (for details on weighted standardizing, see "Algorithms" on page 33-1865). MATLAB does not standardize the data contained in the dummy variable columns generated for categorical predictors.
- The software trains the classifier using the standardized predictor matrix, but stores the unstandardized data in the classifier property X.

Example: 'Standardize',true

Data Types: logical

Verbose — Verbosity level
0 (default) | 1 | 2

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. The value of `Verbose` controls the amount of optimization information that the software displays in the Command Window and saves the information as a structure to `Mdl.ConvergenceInfo.History`.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software does not display or save convergence information.
`1`	The software displays diagnostic messages and saves convergence criteria every `numprint` iterations, where `numprint` is the value of the name-value pair argument `'NumPrint'`.
`2`	The software displays diagnostic messages and saves convergence criteria at every iteration.

Example: `'Verbose',1`

Data Types: `double` | `single`

Output Arguments

`t` — SVM classification template

template object

SVM classification template suitable for training error-correcting output code (ECOC) multiclass models, returned as a template object. Pass `t` to `fitcecoc` to specify how to create the SVM classifier for the ECOC model.

If you display `t` to the Command Window, then all, unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

Tip

By default and for efficiency, `fitcecoc` empties the `Alpha`, `SupportVectorLabels`, and `SupportVectors` properties for all linear SVM binary learners. `fitcecoc` lists `Beta`, rather than `Alpha`, in the model display.

To store `Alpha`, `SupportVectorLabels`, and `SupportVectors`, pass a linear SVM template that specifies storing support vectors to `fitcecoc`. For example, enter:

```
t = templateSVM('SaveSupportVectors',true)
Mdl = fitcecoc(X,Y,'Learners',t);
```

You can remove the support vectors and related values by passing the resulting `ClassificationECOC` model to `discardSupportVectors`.

References

[1] Christianini, N., and J. C. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[2] Fan, R.-E., P.-H. Chen, and C.-J. Lin. “Working set selection using second order information for training support vector machines.” *Journal of Machine Learning Research*, Vol 6, 2005, pp. 1889–1918.

[3] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, Second Edition. NY: Springer, 2008.

[4] Kecman V., T. -M. Huang, and M. Vogt. “Iterative Single Data Algorithm for Training Kernel Machines from Huge Data Sets: Theory and Performance.” In *Support Vector Machines: Theory and Applications*. Edited by Lipo Wang, 255–274. Berlin: Springer-Verlag, 2005.

[5] Scholkopf, B., J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. “Estimating the Support of a High-Dimensional Distribution.” *Neural Comput.*, Vol. 13, Number 7, 2001, pp. 1443–1471.

[6] Scholkopf, B., and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, Adaptive Computation and Machine Learning*. Cambridge, MA: The MIT Press, 2002.

See Also

`ClassificationECOC` | `ClassificationSVM` | `fitcecoc` | `fitcsvm`

Introduced in R2014b

templateTree

Create decision tree template

Syntax

```
t = templateTree
t = templateTree(Name,Value)
```

Description

`t = templateTree` returns a default decision tree learner template suitable for training an ensemble (boosted and bagged decision trees) or error-correcting output code (ECOC) multiclass model. Specify `t` as a learner using:

- `fitcensemble` for classification ensembles
- `fitrensemble` for regression ensembles
- `fitcecoc` for ECOC model classification

If you specify a default decision tree template, then the software uses default values for all input arguments during training. It is good practice to specify the type of decision tree, e.g., for a classification tree template, specify `'Type','classification'`. If you specify the type of decision tree and display `t` in the Command Window, then all options except `Type` appear empty (`[]`).

`t = templateTree(Name,Value)` creates a template with additional options specified by one or more name-value pair arguments.

For example, you can specify the algorithm used to find the best split on a categorical predictor, the split criterion, or the number of predictors selected for each split.

If you display `t` in the Command Window, then all options appear empty (`[]`), except those that you specify using name-value pair arguments. During training, the software uses default values for empty options.

Examples

Create a Classification Template with Surrogate Splits

Create a decision tree template with surrogate splits, and use the template to train an ensemble using sample data.

Load Fisher's iris data set.

```
load fisheriris
```

Create a decision tree template of tree stumps with surrogate splits.

```
t = templateTree('Surrogate','on','MaxNumSplits',1)
```

```
t =
Fit template for Tree.
```

```
       Surrogate: 'on'
    MaxNumSplits: 1
```

Options for the template object are empty except for `Surrogate` and `MaxNumSplits`. When you pass `t` to the training function, the software fills in the empty options with their respective default values.

Specify `t` as a weak learner for a classification ensemble.

```
Mdl = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t)
```

```
Mdl =
  ClassificationEnsemble
              ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
          NumObservations: 150
               NumTrained: 100
                   Method: 'AdaBoostM2'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [100x1 double]
       FitInfoDescription: {2x1 cell}


  Properties, Methods
```

Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl)
```

```
L = 0.0333
```

Optimize Regression Ensemble Using Cross-Validation

One way to create an ensemble of boosted regression trees that has satisfactory predictive performance is to tune the decision tree complexity level using cross-validation. While searching for an optimal complexity level, tune the learning rate to minimize the number of learning cycles as well.

This example manually finds optimal parameters by using the cross-validation option (the `'KFold'` name-value pair argument) and the `kfoldLoss` function. Alternatively, you can use the `'OptimizeHyperparameters'` name-value pair argument to optimize hyperparameters automatically. See “Optimize Regression Ensemble” on page 33-2314.

Load the `carsmall` data set. Choose the number of cylinders, volume displaced by the cylinders, horsepower, and weight as predictors of fuel economy.

```
load carsmall
Tbl = table(Cylinders,Displacement,Horsepower,Weight,MPG);
```

The default values of the tree depth controllers for boosting regression trees are:

- `10` for `MaxNumSplits`.

- `5` for `MinLeafSize`
- `10` for `MinParentSize`

To search for the optimal tree-complexity level:

1. Cross-validate a set of ensembles. Exponentially increase the tree-complexity level for subsequent ensembles from decision stump (one split) to at most n - 1 splits. n is the sample size. Also, vary the learning rate for each ensemble between 0.1 to 1.
2. Estimate the cross-validated mean-squared error (MSE) for each ensemble.
3. For tree-complexity level j, $j = 1 \ldots J$, compare the cumulative, cross-validated MSE of the ensembles by plotting them against number of learning cycles. Plot separate curves for each learning rate on the same figure.
4. Choose the curve that achieves the minimal MSE, and note the corresponding learning cycle and learning rate.

Cross-validate a deep regression tree and a stump. Because the data contain missing values, use surrogate splits. These regression trees serve as benchmarks.

```
rng(1) % For reproducibility
MdlDeep = fitrtree(Tbl,'MPG','CrossVal','on','MergeLeaves','off', ...
    'MinParentSize',1,'Surrogate','on');
MdlStump = fitrtree(Tbl,'MPG','MaxNumSplits',1,'CrossVal','on', ...
    'Surrogate','on');
```

Cross-validate an ensemble of 150 boosted regression trees using 5-fold cross-validation. Using a tree template:

- Vary the maximum number of splits using the values in the sequence $\{2^0, 2^1, \ldots, 2^m\}$. m is such that 2^m is no greater than n - 1.
- Turn on surrogate splits.

For each variant, adjust the learning rate using each value in the set {0.1, 0.25, 0.5, 1}.

```
n = size(Tbl,1);
m = floor(log2(n - 1));
learnRate = [0.1 0.25 0.5 1];
numLR = numel(learnRate);
maxNumSplits = 2.^(0:m);
numMNS = numel(maxNumSplits);
numTrees = 150;
Mdl = cell(numMNS,numLR);

for k = 1:numLR
    for j = 1:numMNS
        t = templateTree('MaxNumSplits',maxNumSplits(j),'Surrogate','on');
        Mdl{j,k} = fitrensemble(Tbl,'MPG','NumLearningCycles',numTrees, ...
            'Learners',t,'KFold',5,'LearnRate',learnRate(k));
    end
end
```

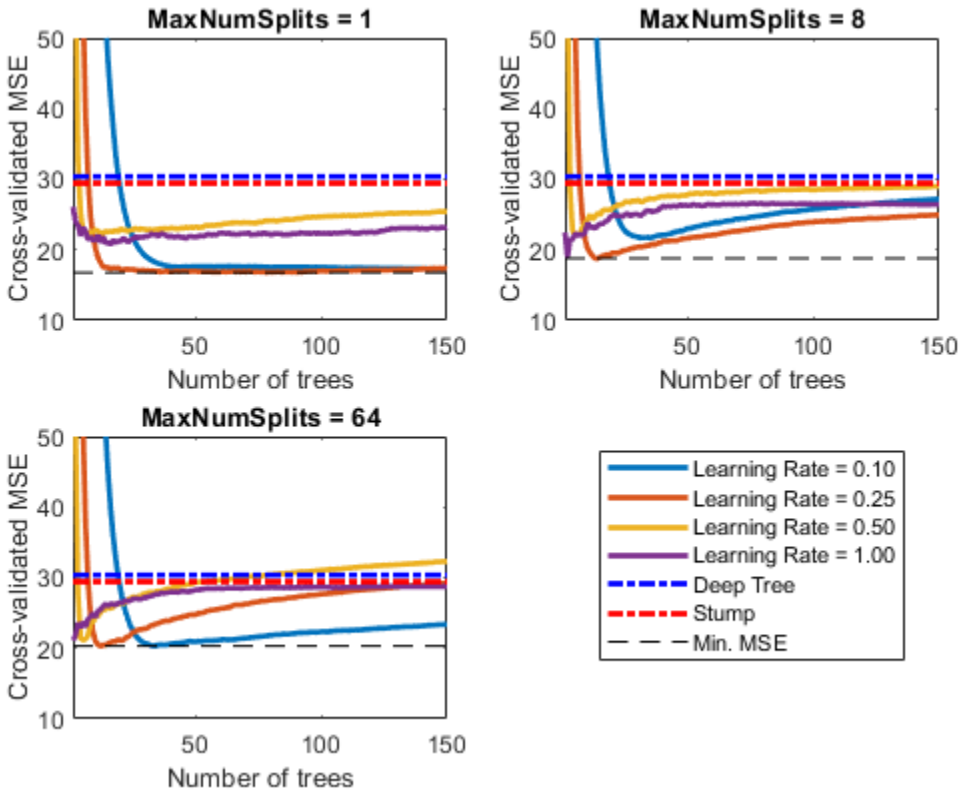
Estimate the cumulative, cross-validated MSE of each ensemble.

```
kflAll = @(x)kfoldLoss(x,'Mode','cumulative');
errorCell = cellfun(kflAll,Mdl,'Uniform',false);
```

```
error = reshape(cell2mat(errorCell),[numTrees numel(maxNumSplits) numel(learnRate)]);
errorDeep = kfoldLoss(MdlDeep);
errorStump = kfoldLoss(MdlStump);
```

Plot how the cross-validated MSE behaves as the number of trees in the ensemble increases. Plot the curves with respect to learning rate on the same plot, and plot separate plots for varying tree-complexity levels. Choose a subset of tree complexity levels to plot.

```
mnsPlot = [1 round(numel(maxNumSplits)/2) numel(maxNumSplits)];
figure;
for k = 1:3
    subplot(2,2,k)
    plot(squeeze(error(:,mnsPlot(k),:)),'LineWidth',2)
    axis tight
    hold on
    h = gca;
    plot(h.XLim,[errorDeep errorDeep],'-.b','LineWidth',2)
    plot(h.XLim,[errorStump errorStump],'-.r','LineWidth',2)
    plot(h.XLim,min(min(error(:,mnsPlot(k),:))).*[1 1],'--k')
    h.YLim = [10 50];
    xlabel('Number of trees')
    ylabel('Cross-validated MSE')
    title(sprintf('MaxNumSplits = %0.3g', maxNumSplits(mnsPlot(k))))
    hold off
end
hL = legend([cellstr(num2str(learnRate','Learning Rate = %0.2f')); ...
        'Deep Tree';'Stump';'Min. MSE']);
hL.Position(1) = 0.6;
```



Each curve contains a minimum cross-validated MSE occurring at the optimal number of trees in the ensemble.

Identify the maximum number of splits, number of trees, and learning rate that yields the lowest MSE overall.

```
[minErr,minErrIdxLin] = min(error(:));
[idxNumTrees,idxMNS,idxLR] = ind2sub(size(error),minErrIdxLin);
fprintf('\nMin. MSE = %0.5f',minErr)
```

```
Min. MSE = 16.77593
```

```
fprintf('\nOptimal Parameter Values:\nNum. Trees = %d',idxNumTrees);
```

```
Optimal Parameter Values:
Num. Trees = 78
```

```
fprintf('\nMaxNumSplits = %d\nLearning Rate = %0.2f\n',...
    maxNumSplits(idxMNS),learnRate(idxLR))
```

```
MaxNumSplits = 1
Learning Rate = 0.25
```

Create a predictive ensemble based on the optimal hyperparameters and the entire training set.

```
tFinal = templateTree('MaxNumSplits',maxNumSplits(idxMNS),'Surrogate','on');
MdlFinal = fitrensemble(Tbl,'MPG','NumLearningCycles',idxNumTrees, ...
    'Learners',tFinal,'LearnRate',learnRate(idxLR))
```

```
MdlFinal =
  RegressionEnsemble
           PredictorNames: {1x4 cell}
             ResponseName: 'MPG'
    CategoricalPredictors: []
        ResponseTransform: 'none'
          NumObservations: 94
               NumTrained: 78
                   Method: 'LSBoost'
             LearnerNames: {'Tree'}
     ReasonForTermination: 'Terminated normally after completing the requested number of training
                  FitInfo: [78x1 double]
       FitInfoDescription: {2x1 cell}
           Regularization: []


  Properties, Methods
```

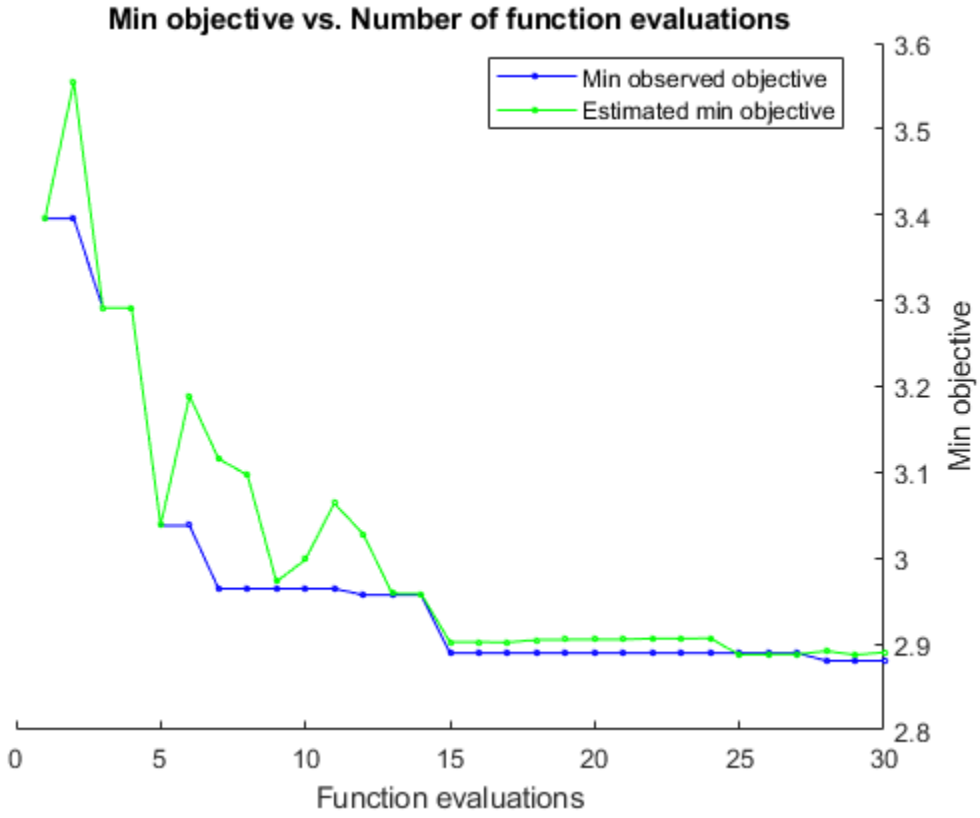
`MdlFinal` is a `RegressionEnsemble`. To predict the fuel economy of a car given its number of cylinders, volume displaced by the cylinders, horsepower, and weight, you can pass the predictor data and `MdlFinal` to `predict`.

Instead of searching optimal values manually by using the cross-validation option (`'KFold'`) and the `kfoldLoss` function, you can use the `'OptimizeHyperparameters'` name-value pair argument. When you specify `'OptimizeHyperparameters'`, the software finds optimal parameters automatically using Bayesian optimization. The optimal values obtained by using `'OptimizeHyperparameters'` can be different from those obtained using manual search.

```
t = templateTree('Surrogate','on');
mdl = fitrensemble(Tbl,'MPG','Learners',t, ...
    'OptimizeHyperparameters',{'NumLearningCycles','LearnRate','MaxNumSplits'})
```

```
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   | NumLearningC-|    Learn
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    | ycles        |
|=====================================================================================================
|    1 | Best   |      3.3955 |       1.274 |      3.3955 |      3.3955 |           26 |     0.01
|    2 | Accept |      6.0976 |      6.5794 |      3.3955 |      3.5549 |          170 |    0.001
|    3 | Best   |      3.2914 |      10.972 |      3.2914 |      3.2917 |          273 |      0.6
|    4 | Accept |      6.1839 |      3.1871 |      3.2914 |      3.2915 |           80 |    0.001
|    5 | Best   |      3.0379 |     0.89567 |      3.0379 |      3.0384 |           18 |      0.2
|    6 | Accept |      3.3628 |     0.49197 |      3.0379 |      3.1888 |           10 |      0.1
|    7 | Best   |      2.9646 |     0.52503 |      2.9646 |      3.1158 |           10 |      0.2
|    8 | Accept |      3.0528 |     0.50685 |      2.9646 |      3.0968 |           10 |      0.2
|    9 | Accept |      2.9789 |     0.47815 |      2.9646 |      2.9727 |           10 |      0.2
|   10 | Accept |      3.0605 |     0.46351 |      2.9646 |      2.9984 |           10 |      0.8
|   11 | Accept |       3.161 |     0.88622 |      2.9646 |      3.0639 |           21 |      0.4
|   12 | Best   |      2.9571 |     0.46911 |      2.9571 |      3.0282 |           10 |      0.4
|   13 | Accept |      6.1344 |     0.47308 |      2.9571 |      2.9588 |           10 |     0.01
|   14 | Accept |      2.9729 |       0.466 |      2.9571 |      2.9586 |           10 |      0.5
|   15 | Best   |      2.8895 |     0.46415 |      2.8895 |      2.9022 |           10 |      0.3
|   16 | Accept |      2.9254 |      2.5157 |      2.8895 |      2.9023 |           69 |      0.1
|   17 | Accept |      2.9254 |      1.3488 |      2.8895 |       2.902 |           35 |        0
|   18 | Accept |      2.9271 |     0.62763 |      2.8895 |      2.9048 |           14 |      0.3
|   19 | Accept |      2.9254 |      4.0751 |      2.8895 |      2.9051 |          116 |     0.06
|   20 | Accept |      2.8966 |       7.672 |      2.8895 |      2.9053 |          223 |      0.1
|=====================================================================================================
```

Iter	Eval result	Objective: log(1+loss)	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	NumLearningC-ycles	Learn
21	Accept	2.9346	1.9984	2.8895	2.9054	49	0.2
22	Accept	3.0867	6.7252	2.8895	2.9063	123	0.0
23	Accept	2.8982	18.831	2.8895	2.9064	460	0.05
24	Accept	2.915	19.45	2.8895	2.9066	498	0
25	Best	2.8894	23.436	2.8894	2.8875	475	0.08
26	Accept	2.9043	15.603	2.8894	2.8879	267	0.00
27	Accept	2.9468	1.1468	2.8894	2.8879	22	0.2
28	Best	2.8805	7.1523	2.8805	2.8918	146	0.1
29	Accept	3.338	0.76699	2.8805	2.8874	11	0.1
30	Accept	2.9067	4.6689	2.8805	2.8903	97	0.1




```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 192.1184 seconds
Total objective function evaluation time: 144.1501

Best observed feasible point:
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           146             0.19004          1
```

```
Observed objective function value = 2.8805
Estimated objective function value = 2.8903
Function evaluation time = 7.1523

Best estimated feasible point (according to models):
    NumLearningCycles    LearnRate    MaxNumSplits
    _________________    _________    ____________

           146            0.19004          1

Estimated objective function value = 2.8903
Estimated function evaluation time = 6.1119

mdl =
  RegressionEnsemble
                        PredictorNames: {1x4 cell}
                          ResponseName: 'MPG'
                 CategoricalPredictors: []
                     ResponseTransform: 'none'
                       NumObservations: 94
    HyperparameterOptimizationResults: [1x1 BayesianOptimization]
                             NumTrained: 146
                                 Method: 'LSBoost'
                           LearnerNames: {'Tree'}
                  ReasonForTermination: 'Terminated normally after completing the requested numbe
                                FitInfo: [146x1 double]
                      FitInfoDescription: {2x1 cell}
                          Regularization: []


  Properties, Methods
```

Unbiased Estimates of Predictor Importance Using Parallel Computing

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a random forest of 500 regression trees using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits. To reproduce random predictor selections, set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

```
rng('default'); % For reproducibility
t = templateTree('PredictorSelection','curvature','Surrogate','on', ...
    'Reproducible',true); % For reproducibility of random predictor selections
Mdl = fitrensemble(X,'MPG','Method','bag','NumLearningCycles',500, ...
    'Learners',t);
```

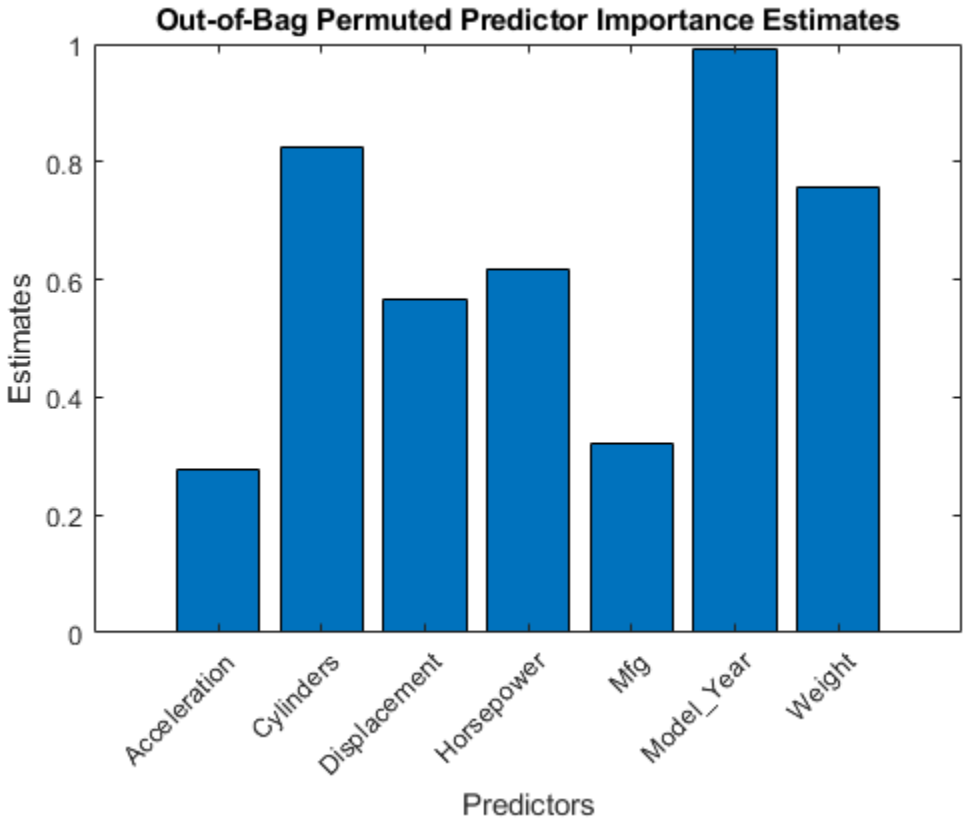
Estimate predictor importance measures by permuting out-of-bag observations. Perform calculations in parallel.

```
options = statset('UseParallel',true);
imp = oobPermutedPredictorImportance(Mdl,'Options',options);
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).
```

Compare the estimates using a bar graph.

```
figure;
bar(imp);
title('Out-of-Bag Permuted Predictor Importance Estimates');
ylabel('Estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Model_Year` is the most important predictor, followed by `Cylinders`. Compare these results to the results in “Estimate Importance of Predictors” on page 33-4355.

Create an Ensemble Template for ECOC Multiclass Learning

Create an ensemble template for use in `fitcecoc`.

Load the arrhythmia data set.

```
load arrhythmia
tabulate(categorical(Y));
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
```

```
     15        5      1.11%
     16       22      4.87%
```

```
rng(1); % For reproducibility
```

Some classes have small relative frequencies in the data.

Create a template for a AdaBoostM1 ensemble of classification trees, and specify to use 100 learners and a shrinkage of 0.1. By default, boosting grows stumps (i.e., one node having a set of leaves). Since there are classes with small frequencies, the trees must be leafy enough to be sensitive to the minority classes. Specify the minimum number of leaf node observations to 3.

```
tTree = templateTree('MinLeafSize',20);
t = templateEnsemble('AdaBoostM1',100,tTree,'LearnRate',0.1);
```

All properties of the template objects are empty except for `Method` and `Type`, and the corresponding properties of the name-value pair argument values in the function calls. When you pass `t` to the training function, the software fills in the empty properties with their respective default values.

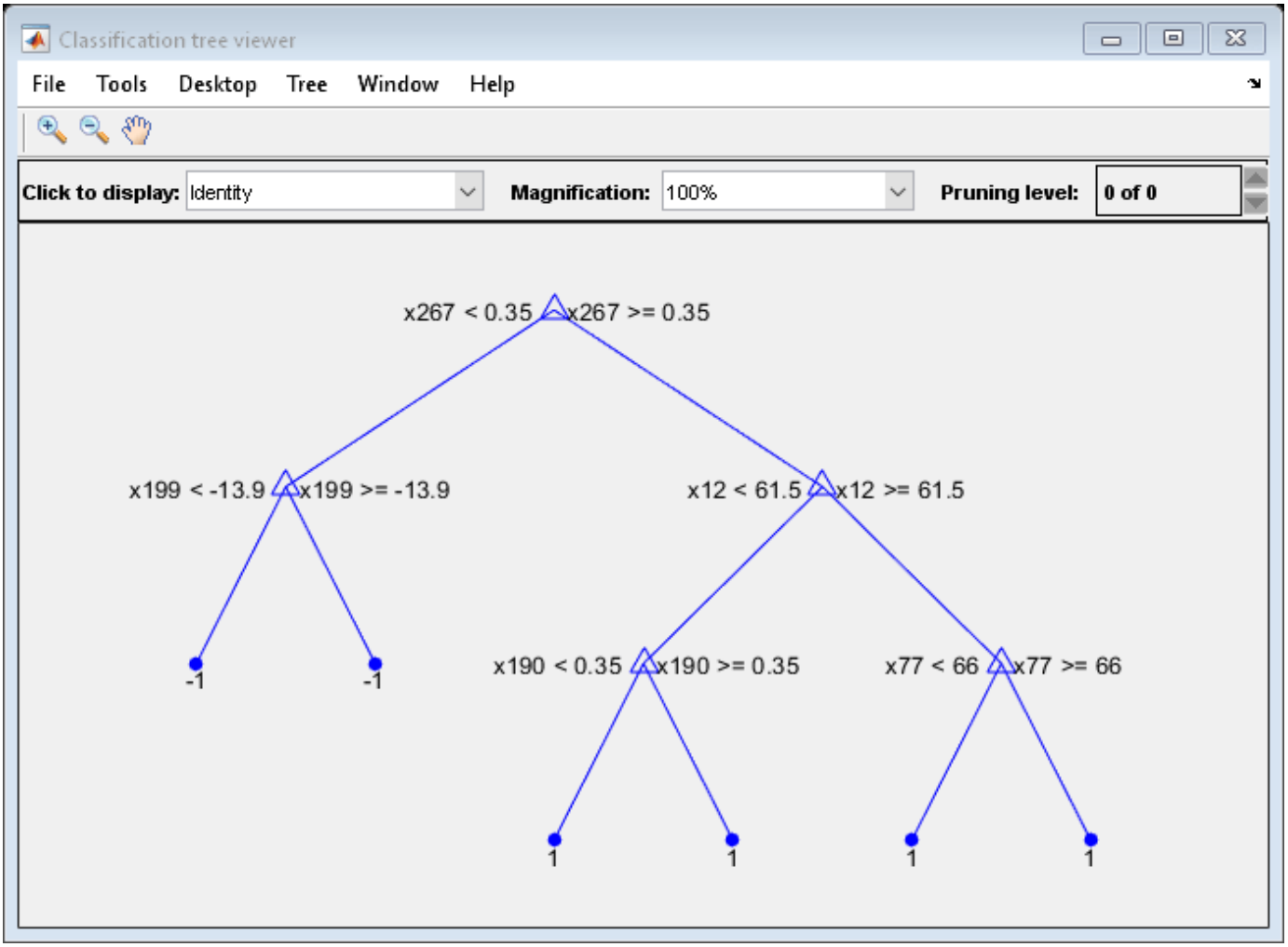
Specify `t` as a binary learner for an ECOC multiclass model. Train using the default one-versus-one coding design.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

- `Mdl` is a `ClassificationECOC` multiclass model.
- `Mdl.BinaryLearners` is a 78-by-1 cell array of `CompactClassificationEnsemble` models.
- `Mdl.BinaryLearners{j}.Trained` is a 100-by-1 cell array of `CompactClassificationTree` models, for `j` = 1,...,78.

You can verify that one of the binary learners contains a weak learner that isn't a stump by using `view`.

```
view(Mdl.BinaryLearners{1}.Trained{1},'Mode','graph')
```



Display the in-sample (resubstitution) misclassification error.

```
L = resubLoss(Mdl,'LossFun','classiferror')
```

```
L = 0.0819
```

Input Arguments

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Surrogate','on','NumVariablesToSample','all'` specifies a template with surrogate splits, and uses all available predictors at each split.

For Classification Trees and Regression Trees

`MaxNumSplits` — Maximal number of decision splits
positive integer

Maximal number of decision splits (or branch nodes) per tree, specified as the comma-separated pair consisting of `'MaxNumSplits'` and a positive integer. `templateTree` splits `MaxNumSplits` or fewer branch nodes. For more details on splitting behavior, see “Algorithms” on page 33-6181.

For bagged decision trees and decision tree binary learners in ECOC models, the default is *n* – *1*, where *n* is the number of observations in the training sample. For boosted decision trees, the default is `10`.

Example: `'MaxNumSplits',5`

Data Types: `single` | `double`

`MergeLeaves` — Leaf merge flag
`'off'` | `'on'`

Leaf merge flag, specified as the comma-separated pair consisting of `'MergeLeaves'` and either `'on'` or `'off'`.

When `'on'`, the decision tree merges leaves that originate from the same parent node, and that provide a sum of risk values greater or equal to the risk associated with the parent node. When `'off'`, the decision tree does not merge leaves.

For boosted and bagged decision trees, the defaults are `'off'`. For decision tree binary learners in ECOC models, the default is `'on'`.

Example: `'MergeLeaves','on'`

`MinLeafSize` — Minimum observations per leaf
positive integer value

Minimum observations per leaf, specified as the comma-separated pair consisting of `'MinLeafSize'` and a positive integer value. Each leaf has at least `MinLeafSize` observations per tree leaf. If you supply both `MinParentSize` and `MinLeafSize`, the decision tree uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

For boosted and bagged decision trees, the defaults are `1` for classification and `5` for regression. For decision tree binary learners in ECOC models, the default is `1`.

Example: `'MinLeafSize',2`

`MinParentSize` — Minimum observations per branch node
positive integer value

Minimum observations per branch node, specified as the comma-separated pair consisting of `'MinParentSize'` and a positive integer value. Each branch node in the tree has at least `MinParentSize` observations. If you supply both `MinParentSize` and `MinLeafSize`, the decision tree uses the setting that gives larger leaves: `MinParentSize = max(MinParentSize,2*MinLeafSize)`.

- If you specify `MinLeafSize`, then the default value for `'MinParentSize'` is `10`.
- If you do not specify `MinLeafSize`, then the default value changes depending on the training model. For boosted and bagged decision trees, the default value is `2` for classification and `10` for regression. For decision tree binary learners in ECOC models, the default value is `10`.

Example: `'MinParentSize',4`

`NumVariablesToSample` — Number of predictors to select at random for each split
positive integer value | `'all'`

Number of predictors to select at random for each split, specified as the comma-separated pair consisting of `'NumVariablesToSample'` and a positive integer value. Alternatively, you can specify `'all'` to use all available predictors.

If the training data includes many predictors and you want to analyze predictor importance, then specify `'NumVariablesToSample'` as `'all'`. Otherwise, the software might not select some predictors, underestimating their importance.

To reproduce the random selections, you must set the seed of the random number generator by using `rng` and specify `'Reproducible',true`.

For boosted decision trees and decision tree binary learners in ECOC models, the default is `'all'`. The default for bagged decision trees is the square root of the number of predictors for classification, or one third of the number of predictors for regression.

Example: `'NumVariablesToSample',3`

Data Types: `single` | `double` | `char` | `string`

`PredictorSelection` — Algorithm used to select the best split predictor
`'allsplits'` (default) | `'curvature'` | `'interaction-curvature'`

Algorithm used to select the best split predictor at each node, specified as the comma-separated pair consisting of `'PredictorSelection'` and a value in this table.

Value	Description
`'allsplits'`	Standard CART — Selects the split predictor that maximizes the split-criterion gain over all possible splits of all predictors [1].
`'curvature'`	Curvature test — Selects the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response [3][4]. Training speed is similar to standard CART.
`'interaction-curvature'`	Interaction test — Chooses the split predictor that minimizes the *p*-value of chi-square tests of independence between each predictor and the response, and that minimizes the *p*-value of a chi-square test of independence between each pair of predictors and response [3]. Training speed can be slower than standard CART.

For `'curvature'` and `'interaction-curvature'`, if all tests yield *p*-values greater than 0.05, then MATLAB stops splitting nodes.

Tip

- The curvature and interaction tests are not recommended for boosting decision trees. To train an ensemble of boosted trees that has greater accuracy, use standard CART instead.
- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [4]. If the predictor data set is heterogeneous, or if there are predictors that have relatively fewer distinct values than other variables, then consider specifying the curvature or interaction test.

- If there are predictors that have relatively fewer distinct values than other predictors, for example, if the predictor data set is heterogeneous.
- If an analysis of predictor importance is your goal. For more on predictor importance estimation, see `oobPermutedPredictorImportance` for classification problems, `oobPermutedPredictorImportance` for regression problems, and "Introduction to Feature Selection" on page 15-49.

- Trees grown using standard CART are not sensitive to predictor variable interactions. Also, such trees are less likely to identify important variables in the presence of many irrelevant predictors than the application of the interaction test. Therefore, to account for predictor interactions and identify importance variables in the presence of many irrelevant variables, specify the interaction test [3].
- Prediction speed is unaffected by the value of `'PredictorSelection'`.

For details on how `templateTree` selects split predictors, see "Node Splitting Rules" on page 33-1928 (classification), "Node Splitting Rules" on page 33-2413 (regression), and "Choose Split Predictor Selection Technique" on page 19-14.

Example: `'PredictorSelection','curvature'`

`Prune` — Flag to estimate optimal sequence of pruned subtrees
`'off'` | `'on'`

Flag to estimate the optimal sequence of pruned subtrees, specified as the comma-separated pair consisting of `'Prune'` and `'on'` or `'off'`.

If `Prune` is `'on'`, then the software trains the classification tree learners without pruning them, but estimates the optimal sequence of pruned subtrees for each learner in the ensemble or decision tree binary learner in ECOC models. Otherwise, the software trains the classification tree learners without estimating the optimal sequence of pruned subtrees.

For boosted and bagged decision trees, the default is `'off'`.

For decision tree binary learners in ECOC models, the default is `'on'`.

Example: `'Prune','on'`

`PruneCriterion` — Pruning criterion
`'error'` | `'impurity'` | `'mse'`

Pruning criterion, specified as the comma-separated pair consisting of `'PruneCriterion'` and a pruning criterion valid for the tree type.

- For classification trees, you can specify `'error'` (default) or `'impurity'`. If you specify `'impurity'`, then `templateTree` uses the impurity measure specified by the `'SplitCriterion'` name-value pair argument.
- For regression trees, you can specify only `'mse'`(default).

Example: `'PruneCriterion','impurity'`

`Reproducible` — Flag to enforce reproducibility
`false` (logical `0`) (default) | `true` (logical `1`)

Flag to enforce reproducibility over repeated runs of training a model, specified as the comma-separated pair consisting of `'Reproducible'` and either `false` or `true`.

If `'NumVariablesToSample'` is not `'all'`, then the software selects predictors at random for each split. To reproduce the random selections, you must specify `'Reproducible',true` and set the seed of the random number generator by using `rng`. Note that setting `'Reproducible'` to `true` can slow down training.

Example: `'Reproducible',true`

Data Types: `logical`

`SplitCriterion` — Split criterion
`'gdi'` | `'twoing'` | `'deviance'` | `'mse'`

Split criterion, specified as the comma-separated pair consisting of `'SplitCriterion'` and a split criterion valid for the tree type.

- For classification trees:
  - `'gdi'` for Gini's diversity index (default)
  - `'twoing'` for the twoing rule
  - `'deviance'` for maximum deviance reduction (also known as cross entropy)
- For regression trees:
  - `'mse'` for mean squared error (default)

Example: `'SplitCriterion','deviance'`

`Surrogate` — Surrogate decision splits
`'off'` (default) | `'on'` | `'all'` | positive integer value

Surrogate decision splits flag, specified as the comma-separated pair consisting of `'Surrogate'` and one of `'off'`, `'on'`, `'all'`, or a positive integer value.

- When `'off'`, the decision tree does not find surrogate splits at the branch nodes.
- When `'on'`, the decision tree finds at most 10 surrogate splits at each branch node.
- When set to `'all'`, the decision tree finds all surrogate splits at each branch node. The `'all'` setting can consume considerable time and memory.
- When set to a positive integer value, the decision tree finds at most the specified number of surrogate splits at each branch node.

Use surrogate splits to improve the accuracy of predictions for data with missing values. This setting also lets you compute measures of predictive association between predictors.

Example: `'Surrogate','on'`

Data Types: `single` | `double` | `char` | `string`

`Type` — Decision tree type
`'classification'` | `'regression'`

Decision tree type, specified as a value in the table

Value	Description
`'classification'`	Grow classification tree learners. The fitting functions `fitcensemble` and `fitcecoc` set this value when you pass `t` to them.
`'regression'`	Grow regression tree learners. The fitting function `fitrensemble` sets this value when you pass `t` to it.

Tip Although `t` infers `Type` from the fitting function to which it is supplied, the following occur when you set `Type`:

- The display of `t` shows all options. Each unspecified option is an empty array `[]`.
- `templateTree` checks specifications for errors.

Example: `'Type','classification'`

Data Types: `char` | `string`

For Classification Trees Only

`AlgorithmForCategorical` — Algorithm for best categorical predictor split
`'Exact'` | `'PullLeft'` | `'PCA'` | `'OVAbyClass'`

Algorithm to find the best split on a categorical predictor for data with C categories for data and $K \geq 3$ classes, specified as the comma-separated pair consisting of `'AlgorithmForCategorical'` and one of the following.

Value	Description
`'Exact'`	Consider all $2^{C-1} - 1$ combinations.
`'PullLeft'`	Start with all C categories on the right branch. Consider moving each category to the left branch as it achieves the minimum impurity for the K classes among the remaining categories. From this sequence, choose the split that has the lowest impurity.
`'PCA'`	Compute a score for each category using the inner product between the first principal component of a weighted covariance matrix (of the centered class probability matrix) and the vector of class probabilities for that category. Sort the scores in ascending order, and consider all $C - 1$ splits.

Value	Description
`'OVAbyClass'`	Start with all C categories on the right branch. For each class, order the categories based on their probability for that class. For the first class, consider moving each category to the left branch in order, recording the impurity criterion at each move. Repeat for the remaining classes. From this sequence, choose the split that has the minimum impurity.

The software selects the optimal subset of algorithms for each split using the known number of classes and levels of a categorical predictor. For two classes, it always performs the exact search. Use the `'AlgorithmForCategorical'` name-value pair argument to specify a particular algorithm.

For more details, see “Splitting Categorical Predictors in Classification Trees” on page 19-25.

Example: `'AlgorithmForCategorical','PCA'`

`MaxNumCategories` — Maximum category levels in split node
`10` (default) | nonnegative scalar value

Maximum category levels in the split node, specified as the comma-separated pair consisting of `'MaxNumCategories'` and a nonnegative scalar value. A classification tree splits a categorical predictor using the exact search algorithm if the predictor has at most `MaxNumCategories` levels in the split node. Otherwise, it finds the best categorical split using one of the inexact algorithms. Note that passing a small value can increase computation time and memory overload.

Example: `'MaxNumCategories',8`

For Regression Trees Only

`QuadraticErrorTolerance` — Quadratic error tolerance
`1e-6` (default) | positive scalar value

Quadratic error tolerance per node, specified as the comma-separated pair consisting of `'QuadraticErrorTolerance'` and a positive scalar value. A regression tree stops splitting nodes when the weighted mean squared error per node drops below `QuadraticErrorTolerance`$*\varepsilon$, where ε is the weighted mean squared error of all n responses computed before growing the decision tree.

$$\varepsilon = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2 .$$

w_i is the weight of observation i, given that the weights of all the observations sum to one ($\sum_{i=1}^{n} w_i = 1$), and

$$\bar{y} = \sum_{i=1}^{n} w_i y_i$$

is the weighted average of all the responses.

Example: `'QuadraticErrorTolerance',1e-4`

Output Arguments

`t` — Decision tree template for classification or regression

template object

Decision tree template for classification or regression suitable for training an ensemble (boosted and bagged decision trees) or error-correcting output code (ECOC) multiclass model, returned as a template object. Pass `t` to `fitcensemble`, or `fitrensemble`, or `fitcecoc` to specify how to create the decision tree for the classification ensemble, regression ensemble, or ECOC model, respectively.

If you display `t` in the Command Window, then all unspecified options appear empty (`[]`). However, the software replaces empty options with their corresponding default values during training.

Algorithms

- To accommodate `MaxNumSplits`, the software splits all nodes in the current layer, and then counts the number of branch nodes. A layer is the set of nodes that are equidistant from the root node. If the number of branch nodes exceeds `MaxNumSplits`, then the software follows this procedure.

  1. Determine how many branch nodes in the current layer need to be unsplit so that there would be at most `MaxNumSplits` branch nodes.
  2. Sort the branch nodes by their impurity gains.
  3. Unsplit the desired number of least successful branches.
  4. Return the decision tree grown so far.

  This procedure aims at producing maximally balanced trees.
- The software splits branch nodes layer by layer until at least one of these events occurs.

  - There are `MaxNumSplits` + 1 branch nodes.
  - A proposed split causes the number of observations in at least one branch node to be fewer than `MinParentSize`.
  - A proposed split causes the number of observations in at least one leaf node to be fewer than `MinLeafSize`.
  - The algorithm cannot find a good split within a layer (i.e., the pruning criterion (see `PruneCriterion`), does not improve for all proposed splits in a layer). A special case of this event is when all nodes are pure (i.e., all observations in the node have the same class).
  - For values `'curvature'` or `'interaction-curvature'` of `PredictorSelection`, all tests yield *p*-values greater than 0.05.

  `MaxNumSplits` and `MinLeafSize` do not affect splitting at their default values. Therefore, if you set `'MaxNumSplits'`, then splitting might stop due to the value of `MinParentSize` before `MaxNumSplits` splits occur.
- For details on selecting split predictors and node-splitting algorithms when growing decision trees, see “Algorithms” on page 33-1928 for classification trees and “Algorithms” on page 33-2413 for regression trees.

References

[1] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[2] Coppersmith, D., S. J. Hong, and J. R. M. Hosking. “Partitioning Nominal Attributes in Decision Trees.” *Data Mining and Knowledge Discovery*, Vol. 3, 1999, pp. 197–217.

[3] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[4] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

See Also

`ClassificationTree` | `RegressionTree` | `fitcecoc` | `fitcensemble` | `fitctree` | `fitrensemble` | `templateEnsemble`

Introduced in R2014a

test

Test indices for cross-validation

Syntax

```
idx = test(c)
idx = test(c,i)
```

Description

`idx = test(c)` returns the test indices `idx` for a `cvpartition` object `c` of type `'holdout'` or `'resubstitution'`.

- If `c.Type` is `'holdout'`, then `idx` specifies the observations in the test set.
- If `c.Type` is `'resubstitution'`, then `idx` specifies all observations.

`idx = test(c,i)` returns the test indices for repetition `i` of a `cvpartition` object `c` of type `'kfold'` or `'leaveout'`.

- If `c.Type` is `'kfold'`, then `idx` specifies the observations in the `i`th test set or fold.
- If `c.Type` is `'leaveout'`, then `idx` specifies the observation reserved for testing at repetition `i`.

Examples

Identify Test Indices in Holdout Partition

Identify the observations that are in the test (holdout) set of a `cvpartition` object.

Partition 10 observations for holdout validation. Select approximately 30% of the observations to be in the test set.

```
rng('default') % For reproducibility
c = cvpartition(10,'Holdout',0.30)
```

```
c =
Hold-out cross validation partition
   NumObservations: 10
       NumTestSets: 1
         TrainSize: 7
          TestSize: 3
```

Identify the test set observations. Observations that correspond to 1s are in the test set.

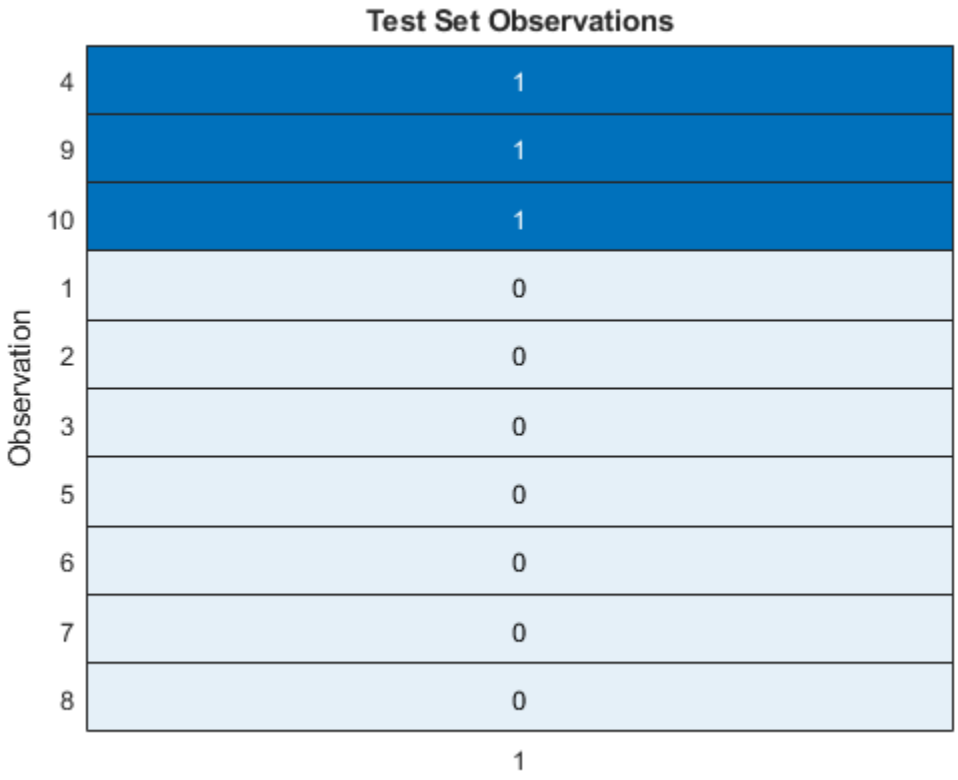
```
holdout = test(c)
```

```
holdout = 10x1 logical array

   0
   0
   0
```

```
     1
     0
     0
     0
     0
     1
     1
```

Visualize the results. The fourth, ninth, and tenth observations are in the test set.

```
h = heatmap(double(holdout),'ColorbarVisible','off');
sorty(h,'1','descend')
ylabel('Observation')
title('Test Set Observations')
```



Identify Test Indices in k-Fold Partition

Identify the observations that are in the test sets, or folds, of a `cvpartition` object for 3-fold cross-validation.

Partition 10 observations for 3-fold cross-validation. Notice that `c` contains three repetitions of training and test data.

```
rng('default') % For reproducibility
c = cvpartition(10,'KFold',3)

c =
K-fold cross validation partition
   NumObservations: 10
       NumTestSets: 3
         TrainSize: 7  6  7
          TestSize: 3  4  3
```

Identify the test set observations for each repetition of training and test data. Observations that correspond to 1s are in the corresponding test set (fold).

```
fold1 = test(c,1)

fold1 = 10x1 logical array

   1
   1
   0
   0
   0
   0
   0
   0
   1
   0


fold2 = test(c,2);
fold3 = test(c,3);
```

Visualize the results. The first, second, and ninth observations are in the first test set. The third, sixth, eighth, and tenth observations are in the second test set. The fourth, fifth, and seventh observations are in the third test set.

```
data = [fold1,fold2,fold3];
h = heatmap(double(data),'ColorbarVisible','off');
sorty(h,{'1','2','3'},'descend')
xlabel('Repetition')
ylabel('Observation')
title('Test Set Observations')
```

Input Arguments

`c` — Validation partition

`cvpartition` object

Validation partition, specified as a `cvpartition` object. The validation partition type of `c`, `c.Type`, is `'kfold'`, `'holdout'`, `'leaveout'`, or `'resubstitution'`.

`i` — Repetition index

positive integer scalar

Repetition index, specified as a positive integer scalar. Specifying `i` indicates to find the observations in the `i`th test set (fold).

Data Types: `single` | `double`

Output Arguments

`idx` — Indices for test set observations

logical vector

Indices for test set observations, returned as a logical vector. A value of 1 indicates that the corresponding observation is in the test set. A value of 0 indicates that the corresponding observation is in the training set.

See Also

`cvpartition` | `training`

Introduced in R2008a

testcholdout

Compare predictive accuracies of two classification models

Syntax

```
h = testcholdout(YHat1,YHat2,Y)
h = testcholdout(YHat1,YHat2,Y,Name,Value)
[h,p,e1,e2] = testcholdout( ___ )
```

Description

`testcholdout` statistically assesses the accuracies of two classification models. The function first compares their predicted labels against the true labels, and then it detects whether the difference between the misclassification rates is statistically significant.

You can assess whether the accuracies of the classification models are different, or whether one classification model performs better than another. `testcholdout` can conduct several McNemar test on page 33-6198 variations, including the asymptotic test, the exact-conditional test, and the mid-*p*-value test. For cost-sensitive assessment on page 33-6196, available tests include a chi-square test (requires an Optimization Toolbox license) and a likelihood ratio test.

`h = testcholdout(YHat1,YHat2,Y)` returns the test decision, by conducting the mid-*p*-value McNemar test on page 33-6198, from testing the null hypothesis that the predicted class labels `YHat1` and `YHat2` have equal accuracy for predicting the true class labels `Y`. The alternative hypothesis is that the labels have unequal accuracy.

`h = 1` indicates to reject the null hypothesis at the 5% significance level. `h = 0` indicates to not reject the null hypothesis at 5% level.

`h = testcholdout(YHat1,YHat2,Y,Name,Value)` returns the result of the hypothesis test with additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the type of alternative hypothesis, specify the type of test, or supply a cost matrix.

`[h,p,e1,e2] = testcholdout( ___ )` returns the *p*-value for the hypothesis test (`p`) and the respective classification loss on page 33-6200 of each set of predicted class labels (`e1` and `e2`) using any of the input arguments in the previous syntaxes.

Examples

Compare Accuracies of Two Different Classification Models

Train two classification models using different algorithms. Conduct a statistical test comparing the misclassification rates of the two models on a held-out set.

Load the `ionosphere` data set.

```
load ionosphere
```

Create a partition that evenly splits the data into training and testing sets.

```
rng(1);                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);            % Training-set indices
idxTest = test(CVP);                 % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Train an SVM model and an ensemble of 100 bagged classification trees. For the SVM model, specify to use the radial basis function kernel and a heuristic procedure to determine the kernel scale.

```
MdlSVM = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true,...
    'KernelFunction','RBF','KernelScale','auto');
t = templateTree('Reproducible',true);  % For reproducibility of random predictor selections
MdlBag = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','Bag','Learners',t);
```

`MdlSVM` is a trained `ClassificationSVM` model. `MdlBag` is a trained `ClassificationBaggedEnsemble` model.

Label the test-set observations using the trained models.

```
YhatSVM = predict(MdlSVM,X(idxTest,:));
YhatBag = predict(MdlBag,X(idxTest,:));
```

`YhatSVM` and `YhatBag` are vectors continuing the predicted class labels of the respective models.

Test whether the two models have equal predictive accuracies.

```
h = testcholdout(YhatSVM,YhatBag,Y(idxTest))
```

```
h = logical
   0
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Assess Whether One Classification Model Classifies Better Than Another

Train two classification models using the same algorithm, but adjust a hyperparameter to make the algorithm more complex. Conduct a statistical test to assess whether the simpler model has better accuracy in held-out data than the more complex model.

Load the `ionosphere` data set.

```
load ionosphere;
```

Create a partition that evenly splits the data into training and testing sets.

```
rng(1);                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);            % Training-set indices
idxTest = test(CVP);                 % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Train two SVM models: one that uses a linear kernel (the default for binary classification) and one that uses the radial basis function kernel. Use the default kernel scale of 1.

```
MdlLinear = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true);
MdlRBF = fitcsvm(X(idxTrain,:),Y(idxTrain),'Standardize',true,...
    'KernelFunction','RBF');
```

`MdlLinear` and `MdlRBF` are trained `ClassificationSVM` models.

Label the test-set observations using the trained models.

```
YhatLinear = predict(MdlLinear,X(idxTest,:));
YhatRBF = predict(MdlRBF,X(idxTest,:));
```

`YhatLinear` and `YhatRBF` are vectors continuing the predicted class labels of the respective models.

Test the null hypothesis that the simpler model (`MdlLinear`) is at most as accurate as the more complex model (`MdlRBF`). Because the test-set size is large, conduct the asymptotic McNemar test, and compare the results with the mid- *p*-value test (the cost-insensitive testing default). Request to return *p*-values and misclassification rates.

```
Asymp = zeros(4,1); % Preallocation
MidP = zeros(4,1);

[Asymp(1),Asymp(2),Asymp(3),Asymp(4)] = testcholdout(YhatLinear,YhatRBF,Y(idxTest),...
    'Alternative','greater','Test','asymptotic');
[MidP(1),MidP(2),MidP(3),MidP(4)] = testcholdout(YhatLinear,YhatRBF,Y(idxTest),...
    'Alternative','greater');
table(Asymp,MidP,'RowNames',{'h' 'p' 'e1' 'e2'})
```

```
ans=4×2 table
            Asymp          MidP
          __________    __________

    h              1             1
    p     7.2801e-09    2.7649e-10
    e1       0.13714       0.13714
    e2       0.33143       0.33143
```

The *p*-value is close to zero for both tests, which indicates strong evidence to reject the null hypothesis that the simpler model is less accurate than the more complex model. No matter what test you specify, `testcholdout` returns the same type of misclassification measure for both models.

Conduct Cost-Sensitive Comparison of Two Classification Models

For data sets with imbalanced class representations, or if the false-positive and false-negative costs are imbalanced, you can statistically compare the predictive performance of two classification models by including a cost matrix in the analysis.

Load the `arrhythmia` data set. Determine the class representations in the data.

```
load arrhythmia;
Y = categorical(Y);
tabulate(Y);
```

```
  Value    Count   Percent
      1      245     54.20%
      2       44      9.73%
      3       15      3.32%
      4       15      3.32%
      5       13      2.88%
      6       25      5.53%
      7        3      0.66%
      8        2      0.44%
      9        9      1.99%
     10       50     11.06%
     14        4      0.88%
     15        5      1.11%
     16       22      4.87%
```

There are 16 classes, however some are not represented in the data set (for example, class 13). Most observations are classified as not having arrhythmia (class 1). The data set is highly discrete with imbalanced classes.

Combine all observations with arrhythmia (classes 2 through 15) into one class. Remove those observations with unknown arrhythmia status (class 16) from the data set.

```
idx = (Y ~= '16');
Y = Y(idx);
X = X(idx,:);
Y(Y ~= '1') = 'WithArrhythmia';
Y(Y == '1') = 'NoArrhythmia';
Y = removecats(Y);
```

Create a partition that evenly splits the data into training and test sets.

```
rng(1);                              % For reproducibility
CVP = cvpartition(Y,'holdout',0.5);
idxTrain = training(CVP);            % Training-set indices
idxTest = test(CVP);                 % Test-set indices
```

`CVP` is a cross-validation partition object that specifies the training and test sets.

Create a cost matrix such that misclassifying a patient with arrhythmia into the "no arrhythmia" class is five times worse than misclassifying a patient without arrhythmia into the arrhythmia class. Classifying correctly incurs no cost. The rows indicate the true class and the columns indicate predicted class. When you conduct a cost-sensitive analysis, a good practice is to specify the order of the classes.

```
Cost = [0 1;5 0];
ClassNames = {'NoArrhythmia','WithArrhythmia'};
```

Train two boosting ensembles of 50 classification trees, one that uses AdaBoostM1 and another that uses LogitBoost. Because there are missing values in the data set, specify to use surrogate splits. Train the models using the cost matrix.

```
t = templateTree('Surrogate','on');
numTrees = 50;
MdlAda = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','AdaBoostM1',...
    'NumLearningCycles',numTrees,'Learners',t,...
    'Cost',Cost,'ClassNames',ClassNames);
MdlLogit = fitcensemble(X(idxTrain,:),Y(idxTrain),'Method','LogitBoost',...
```

```
    'NumLearningCycles',numTrees,'Learners',t,...
    'Cost',Cost,'ClassNames',ClassNames);
```

`MdlAda` and `MdlLogit` are trained `ClassificationEnsemble` models.

Label the test-set observations using the trained models.

```
YhatAda = predict(MdlAda,X(idxTest,:));
YhatLogit = predict(MdlLogit,X(idxTest,:));
```

`YhatLinear` and `YhatRBF` are vectors containing the predicted class labels of the respective models.

Test whether the AdaBoostM1 ensemble (`MdlAda`) and the LogitBoost ensemble (`MdlLogit`) have equal predictive accuracy. Supply the cost matrix. Conduct the asymptotic, likelihood ratio, cost-sensitive test (the default when you pass in a cost matrix). Request to return *p*-values and misclassification costs.

```
[h,p,e1,e2] = testcholdout(YhatAda,YhatLogit,Y(idxTest),'Cost',Cost)
```

```
h = logical
   0


p = 0.3334

e1 = 0.5581

e2 = 0.4698
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Input Arguments

YHat1 — Predicted class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Predicted class labels of the first classification model, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

If `YHat1` is a character array, then each element must correspond to one row of the array.

`YHat1`, `YHat2`, and `Y` must have equal lengths.

It is a best practice for `YHat1`, `YHat2`, and `Y` to share the same data type.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

YHat2 — Predicted class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Predicted class labels of the second classification model, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

If `YHat2` is a character array, then each element must correspond to one row of the array.

`YHat1`, `YHat2`, and `Y` must have equal lengths.

It is a best practice for `YHat1`, `YHat2`, and `Y` to share the same data type.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Y — True class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

True class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

If `Y` is a character array, then each element must correspond to one row of the array.

`YHat1`, `YHat2`, and `Y` must have equal lengths.

It is a best practice for `YHat1`, `YHat2`, and `Y` to share the same data type.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alternative','greater','Test','asymptotic','Cost',[0 2;1 0]` specifies to test whether the first set of first predicted class labels is more accurate than the second set, to conduct the asymptotic McNemar test, and to penalize misclassifying observations with the true label `ClassNames{1}` twice as much as for misclassifying observations with the true label `ClassNames{2}`.

`Alpha` — Hypothesis test significance level
`0.05` (default) | scalar value in the interval (0,1)

Hypothesis test significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the interval (0,1).

Example: `'Alpha',0.1`

Data Types: `single` | `double`

`Alternative` — Alternative hypothesis to assess
`'unequal'` (default) | `'greater'` | `'less'`

Alternative hypothesis to assess, specified as the comma-separated pair consisting of `'Alternative'` and one of the values listed in the table.

Value	Alternative hypothesis
`'unequal'` (default)	For predicting `Y`, `YHat1` and `YHat2` have unequal accuracies.
`'greater'`	For predicting `Y`, `YHat1` is more accurate than `YHat2`.
`'less'`	For predicting `Y`, `YHat1` is less accurate than `YHat2`.

Example: `'Alternative','greater'`

`ClassNames` — Class names
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class names, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. You must set `ClassNames` using the data type of `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of any input argument dimension that corresponds to class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost`.
- Select a subset of classes for testing. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train and test models using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default is the set of all distinct class names in `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure array.

- If you specify the square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- If you specify the structure `S`, then `S` must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`. You can use this field to specify the order of the classes.
  - `S.ClassificationCosts`, which contains the cost matrix, with rows and columns ordered as in `S.ClassNames`.

If you specify `Cost`, then `testcholdout` cannot conduct one-sided, exact, or mid-*p* tests. You must also specify `'Alternative','unequal','Test','asymptotic'`. For cost-sensitive testing options, see the `CostTest` name-value pair argument.

A best practice is to supply the same cost matrix used to train the classification models.

The default is `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `single` | `double` | `struct`

`CostTest` — Cost-sensitive test type
`'likelihood'` (default) | `'chisquare'`

Cost-sensitive test type, specified as the comma-separated pair consisting of `'CostTest'` and `'chisquare'` or `'likelihood'`. Unless you specify a cost matrix using the `Cost` name-value pair argument, `testcholdout` ignores `CostTest`.

This table summarizes the available options for cost-sensitive testing.

Value	Asymptotic test type	Requirements
`'chisquare'`	Chi-square test	Optimization Toolbox license to implement `quadprog`
`'likelihood'`	Likelihood ratio test	None

For more details, see “Cost-Sensitive Testing” on page 33-6196.

Example: `'CostTest','chisquare'`

`Test` — Test to conduct
`'asymptotic'` | `'exact'` | `'midp'`

Test to conduct, specified as the comma-separated pair consisting of `'Test'` and `'asymptotic'`, `'exact'`, and `'midp'`. This table summarizes the available options for cost-insensitive testing.

Value	Description
`'asymptotic'`	Asymptotic McNemar test
`'exact'`	Exact-conditional McNemar test
`'midp'` (default)	Mid-*p*-value McNemar test

For more details, see “McNemar Tests” on page 33-6198.

For cost-sensitive testing, `Test` must be `'asymptotic'`. When you specify the `Cost` name-value pair argument, and choose a cost-sensitive test using the `CostTest` name-value pair argument, `'asymptotic'` is the default.

Example: `'Test','asymptotic'`

Note `NaN`s, `<undefined>` values, empty character vectors (`''`), empty strings (`""`), and `<missing>` values indicate missing data values. `testcholdout`:

- Treats missing values in `YHat1` and `YHat2` as misclassified observations.
- Removes missing values in `Y` and the corresponding values of `YHat1` and `YHat2`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as a logical value.

`h = 1` indicates the rejection of the null hypothesis at the `Alpha` significance level.

`h = 0` indicates failure to reject the null hypothesis at the `Alpha` significance level.

Data Types: `logical`

p — *p*-value
scalar in the interval [0,1]

p-value of the test, returned as a scalar in the interval [0,1]. `p` is the probability that a random test statistic is at least as extreme as the observed test statistic, given that the null hypothesis is true.

`testcholdout` estimates `p` using the distribution of the test statistic, which varies with the type of test. For details on test statistics derived from the available variants of the McNemar test, see "McNemar Tests" on page 33-6198. For details on test statistics derived from cost-sensitive tests, see "Cost-Sensitive Testing" on page 33-6196.

e1 — Classification loss
scalar

Classification loss on page 33-6200 that summarizes the accuracy of the first set of class labels (`YHat1`) predicting the true class labels (`Y`), returned as a scalar.

For cost-insensitive testing, `e1` is the misclassification rate. That is, `e1` is the proportion of misclassified observations, which is a scalar in the interval [0,1].

For cost-sensitive testing, `e1` is the misclassification cost. That is, `e1` is the weighted average of the misclassification costs, in which the weights are the respective estimated proportions of misclassified observations.

e2 — Classification loss
scalar

Classification loss on page 33-6200 that summarizes the accuracy of the second set of class labels (`YHat2`) predicting the true class labels (`Y`), returned as a scalar.

For cost-insensitive testing, `e2` is the misclassification rate. That is, `e2` is the proportion of misclassified observations, which is a scalar in the interval [0,1].

For cost-sensitive testing, `e2` is the misclassification cost. That is, `e2` is the weighted average of the costs of misclassification, in which the weights are the respective estimated proportions of misclassified observations.

More About

Cost-Sensitive Testing

Conduct cost-sensitive testing when the cost of misclassification is imbalanced. By conducting a cost-sensitive analysis, you can account for the cost imbalance when you train the classification models and when you statistically compare them.

If the cost of misclassification is imbalanced, then the misclassification rate tends to be a poorly performing classification loss. Use misclassification cost instead to compare classification models.

Misclassification costs are often imbalanced in applications. For example, consider classifying subjects based on a set of predictors into two categories: healthy and sick. Misclassifying a sick subject as healthy poses a danger to the subject's life. However, misclassifying a healthy subject as sick typically causes some inconvenience, but does not pose significant danger. In this situation, you

assign misclassification costs such that misclassifying a sick subject as healthy is more costly than misclassifying a healthy subject as sick.

The definitions that follow summarize the cost-sensitive tests. In the definitions:

- n_{ijk} and $\widehat{\pi}_{ijk}$ are the number and estimated proportion of test-sample observations with the following characteristics. k is the true class, i is the label assigned by the first classification model, and j is the label assigned by the second classification model. The unknown true value of $\widehat{\pi}_{ijk}$ is π_{ijk}. The test-set sample size is $\sum_{i,j,k} n_{ijk} = n_{test}$. Additionally, $\sum_{i,j,k} \pi_{ijk} = \sum_{i,j,k} \widehat{\pi}_{ijk} = 1$.
- c_{ij} is the relative cost of assigning label j to an observation with true class i. $c_{ii} = 0$, $c_{ij} \geq 0$, and, for at least one (i,j) pair, $c_{ij} > 0$.
- All subscripts take on integer values from 1 through K, which is the number of classes.
- The expected difference in the misclassification costs of the two classification models is

  $$\delta = \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} \left(c_{ki} - c_{kj}\right)\pi_{ijk}.$$

- The hypothesis test is

  $$H_0: \delta = 0$$
  $$H_1: \delta \neq 0.$$

  The available cost-sensitive tests are appropriate for two-tailed testing.

Available asymptotic tests that address imbalanced costs are a chi-square test and a likelihood ratio test.

- Chi-square test — The chi-square test statistic is based on the Pearson and Neyman chi-square test statistics, but with a Laplace correction factor to account for any $n_{ijk} = 0$. The test statistic is

  $$t^*_{\chi^2} = \sum_{i \neq j} \sum_{k} \frac{\left(n_{ijk} + 1 - \left(n_{test} + K^3\right)\widehat{\pi}^{(1)}_{ijk}\right)^2}{n_{ijk} + 1}.$$

  If $1 - F_{\chi^2}\left(t^*_{\chi^2}; 1\right) < \alpha$, then reject H_0.

  - $\widehat{\pi}^{(1)}_{ijk}$ are estimated by minimizing $t^*_{\chi^2}$ under the constraint that $\delta = 0$.
  - $F_{\chi^2}(x; 1)$ is the χ^2 cdf with one degree of freedom evaluated at x.
- Likelihood ratio test — The likelihood ratio test is based on N_{ijk}, which are binomial random variables with sample size n_{test} and success probability π_{ijk}. The random variables represent the random number of observations with: true class k, label i assigned by the first classification model, and label j assigned by the second classification model. Jointly, the distribution of the random variables is multinomial.

  The test statistic is

  $$t^*_{LRT} = 2\log\left[\frac{P\left(\bigcap_{i,j,k} N_{ijk} = n_{ijk}; n_{test}, \widehat{\pi}_{ijk} = \widehat{\pi}^{(2)}_{ijk}\right)}{P\left(\bigcap_{i,j,k} N_{ijk} = n_{ijk}; n_{test}, \widehat{\pi}_{ijk} = \widehat{\pi}^{(3)}_{ijk}\right)}\right].$$

If $1 - F_{\chi^2}\left(t^*_{LRT}; 1\right) < \alpha$, then reject H_0.

- $\hat{\pi}^{(2)}_{ijk} = \frac{n_{ijk}}{n_{test}}$ is the unrestricted MLE of π_{ijk}.
- $\hat{\pi}^{(3)}_{ijk} = \frac{n_{ijk}}{n_{test} + \lambda(c_{ki} - c_{kj})}$ is the MLE under the null hypothesis that $\delta = 0$. λ is the solution to $\sum_{i,j,k} \frac{n_{ijk}(c_{ki} - c_{kj})}{n_{test} + \lambda(c_{ki} - c_{kj})} = 0$.
- $F_{\chi^2}(x; 1)$ is the χ^2 cdf with one degree of freedom evaluated at x.

McNemar Tests

McNemar Tests are hypothesis tests that compare two population proportions while addressing the issues resulting from two dependent, matched-pair samples.

One way to compare the predictive accuracies of two classification models is:

1. Partition the data into training and test sets.
2. Train both classification models using the training set.
3. Predict class labels using the test set.
4. Summarize the results in a two-by-two table similar to this figure.

		Model 2		
		Correct	Incorrect	
Model 1	Correct	n_{11}	n_{12}	$n_{1\bullet}$
	Incorrect	n_{21}	n_{22}	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	n_{test}

n_{ii} are the number of concordant pairs, that is, the number of observations that both models classify the same way (correctly or incorrectly). n_{ij}, $i \neq j$, are the number of discordant pairs, that is, the number of observations that models classify differently (correctly or incorrectly).

The misclassification rates for Models 1 and 2 are $\hat{\pi}_{2\bullet} = n_{2\bullet}/n$ and $\hat{\pi}_{\bullet 2} = n_{\bullet 2}/n$, respectively. A two-sided test for comparing the accuracy of the two models is

$H_0: \pi_{\bullet 2} = \pi_{2\bullet}$

$H_1: \pi_{\bullet 2} \neq \pi_{2\bullet}$.

The null hypothesis suggests that the population exhibits marginal homogeneity, which reduces the null hypothesis to $H_0: \pi_{12} = \pi_{21}$. Also, under the null hypothesis, $N_{12} \sim \text{Binomial}(n_{12} + n_{21}, 0.5)$ [1].

These facts are the basis for the available McNemar test variants: the asymptotic, exact-conditional, and mid-p-value McNemar tests. The definitions that follow summarize the available variants.

- Asymptotic — The asymptotic McNemar test statistics and rejection regions (for significance level α) are:
  - For one-sided tests, the test statistic is

    $$t_{a1}^* = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

    If $1 - \Phi\left(\left|t_1^*\right|\right) < \alpha$, where Φ is the standard Gaussian cdf, then reject H_0.
  - For two-sided tests, the test statistic is

    $$t_{a2}^* = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

    If $1 - F_{\chi^2}\left(t_2^*; m\right) < \alpha$, where $F_{\chi^2}(x; m)$ is the χ_m^2 cdf evaluated at x, then reject H_0.

  The asymptotic test requires large-sample theory, specifically, the Gaussian approximation to the binomial distribution.
  - The total number of discordant pairs, $n_d = n_{12} + n_{21}$, must be greater than 10 ([1], Ch. 10.1.4).
  - In general, asymptotic tests do not guarantee nominal coverage. The observed probability of falsely rejecting the null hypothesis can exceed α, as suggested in simulation studies in [18]. However, the asymptotic McNemar test performs well in terms of statistical power.
- Exact-Conditional — The exact-conditional McNemar test statistics and rejection regions (for significance level α) are ([36], [38]):
  - For one-sided tests, the test statistic is

    $$t_1^* = n_{12}.$$

    If $F_{\text{Bin}}\left(t_1^*; n_d, 0.5\right) < \alpha$, where $F_{\text{Bin}}(x; n, p)$ is the binomial cdf with sample size n and success probability p evaluated at x, then reject H_0.
  - For two-sided tests, the test statistic is

    $$t_2^* = \min(n_{12}, n_{21}).$$

    If $F_{\text{Bin}}\left(t_2^*; n_d, 0.5\right) < \alpha/2$, then reject H_0.

  The exact-conditional test always attains nominal coverage. Simulation studies in [18] suggest that the test is conservative, and then show that the test lacks statistical power compared to other variants. For small or highly discrete test samples, consider using the mid-p-value test ([1], Ch. 3.6.3).
- Mid-p-value test — The mid-p-value McNemar test statistics and rejection regions (for significance level α) are ([32]):

- For one-sided tests, the test statistic is

  $t_1^* = n_{12}$.

  If $F_{\text{Bin}}\left(t_1^* - 1; n_{12} + n_{21}, 0.5\right) + 0.5 f_{\text{Bin}}\left(t_1^*; n_{12} + n_{21}, 0.5\right) < \alpha$, where $F_{\text{Bin}}(x; n, p)$ and $f_{\text{Bin}}(x; n, p)$ are the binomial cdf and pdf, respectively, with sample size n and success probability p evaluated at x, then reject H_0.
- For two-sided tests, the test statistic is

  $t_2^* = \min(n_{12}, n_{21})$.

  If $F_{\text{Bin}}\left(t_2^* - 1; n_{12} + n_{21} - 1, 0.5\right) + 0.5 f_{\text{Bin}}\left(t_2^*; n_{12} + n_{21}, 0.5\right) < \alpha/2$, then reject H_0.

The mid-*p*-value test addresses the over-conservative behavior of the exact-conditional test. The simulation studies in [18] demonstrate that this test attains nominal coverage, and has good statistical power.

Classification Loss

Classification losses indicate the accuracy of a classification model or set of predicted labels. Two classification losses are the misclassification rate and cost.

`testcholdout` returns the classification losses (see `e1` and `e2`) under the alternative hypothesis (that is, the unrestricted classification losses). n_{ijk} is the number of test-sample observations with: true class k, label i assigned by the first classification model, and label j assigned by the second classification model. The corresponding estimated proportion is $\widehat{\pi}_{ijk} = \frac{n_{ijk}}{n_{test}}$. The test-set sample size is $\sum_{i,j,k} n_{ijk} = n_{test}$. The indices are taken from 1 through K, the number of classes.

- The misclassification rate, or classification error, is a scalar in the interval [0,1] representing the proportion of misclassified observations. That is, the misclassification rate for the first classification model is

  $$e_1 = \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{i \neq k} \widehat{\pi}_{ijk}.$$

  For the misclassification rate of the second classification model (e_2), switch the indices i and j in the formula.

  Classification accuracy decreases as the misclassification rate increases to 1.
- The misclassification cost is a nonnegative scalar that is a measure of classification quality relative to the values of the specified cost matrix. Its interpretation depends on the specified costs of misclassification. The misclassification cost is the weighted average of the costs of misclassification (specified in a cost matrix, C) in which the weights are the respective estimated proportions of misclassified observations. The misclassification cost for the first classification model is

  $$e_1 = \sum_{j=1}^{K} \sum_{k=1}^{K} \sum_{i \neq k} \widehat{\pi}_{ijk} c_{ki},$$

where c_{kj} is the cost of classifying an observation into class j if its true class is k. For the misclassification cost of the second classification model (e_2), switch the indices i and j in the formula.

In general, for a fixed cost matrix, classification accuracy decreases as the misclassification cost increases.

Tips

- It is a good practice to obtain predicted class labels by passing any trained classification model and new predictor data to the `predict` method. For example, for predicted labels from an SVM model, see `predict`.
- Cost-sensitive tests perform numerical optimization, which requires additional computational resources. The likelihood ratio test conducts numerical optimization indirectly by finding the root of a Lagrange multiplier in an interval. For some data sets, if the root lies close to the boundaries of the interval, then the method can fail. Therefore, if you have an Optimization Toolbox license, consider conducting the cost-sensitive chi-square test instead. For more details, see `CostTest` and “Cost-Sensitive Testing” on page 33-6196.

References

[1] Agresti, A. *Categorical Data Analysis*, 2nd Ed. John Wiley & Sons, Inc.: Hoboken, NJ, 2002.

[2] Fagerlan, M.W., S. Lydersen, and P. Laake. “The McNemar Test for Binary Matched-Pairs Data: Mid-p and Asymptotic Are Better Than Exact Conditional.” *BMC Medical Research Methodology*. Vol. 13, 2013, pp. 1–8.

[3] Lancaster, H.O. “Significance Tests in Discrete Distributions.” *JASA*, Vol. 56, Number 294, 1961, pp. 223–234.

[4] McNemar, Q. “Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages.” *Psychometrika*, Vol. 12, Number 2, 1947, pp. 153–157.

[5] Mosteller, F. “Some Statistical Problems in Measuring the Subjective Response to Drugs.” *Biometrics*, Vol. 8, Number 3, 1952, pp. 220–226.

See Also

`testckfold`

Topics

“Hypothesis Tests”

Introduced in R2015a

testckfold

Compare accuracies of two classification models by repeated cross-validation

Syntax

```
h = testckfold(C1,C2,X1,X2)
h = testckfold(C1,C2,X1,X2,Y)
h = testckfold(___,Name,Value)
[h,p,e1,e2] = testckfold(___)
```

Description

`testckfold` statistically assesses the accuracies of two classification models by repeatedly cross-validating the two models, determining the differences in the classification loss, and then formulating the test statistic by combining the classification loss differences. This type of test is particularly appropriate when sample size is limited.

You can assess whether the accuracies of the classification models are different, or whether one classification model performs better than another. Available tests include a 5-by-2 paired *t* test, a 5-by-2 paired *F* test, and a 10-by-10 repeated cross-validation *t* test. For more details, see "Repeated Cross-Validation Tests" on page 33-6218. To speed up computations, `testckfold` supports parallel computing (requires a Parallel Computing Toolbox license).

`h = testckfold(C1,C2,X1,X2)` returns the test decision that results from conducting a 5-by-2 paired *F* cross-validation test. The null hypothesis is the classification models `C1` and `C2` have equal accuracy in predicting the true class labels using the predictor and response data in the tables `X1` and `X2`. `h = 1` indicates to reject the null hypothesis at the 5% significance level.

`testckfold` conducts the cross-validation test by applying `C1` and `C2` to all predictor variables in `X1` and `X2`, respectively. The true class labels in `X1` and `X2` must be the same. The response variable names in `X1`, `X2`, `C1.ResponseName`, and `C2.ResponseName` must be the same.

For examples of ways to compare models, see "Tips" on page 33-6221.

`h = testckfold(C1,C2,X1,X2,Y)` applies the full classification model or classification templates `C1` and `C2` to all predictor variables in the tables or matrices of data `X1` and `X2`, respectively. `Y` is the table variable name corresponding to the true class labels, or an array of true class labels.

`h = testckfold(___,Name,Value)` uses any of the input arguments in the previous syntaxes and additional options specified by one or more `Name,Value` pair arguments. For example, you can specify the type of alternative hypothesis, the type of test, or the use of parallel computing.

`[h,p,e1,e2] = testckfold(___)` also returns the *p*-value for the hypothesis test (`p`) and the respective classification losses on page 33-6220 for each cross-validation run and fold (`e1` and `e2`).

Examples

Compare Classification Tree Predictor-Selection Algorithms

At each node, `fitctree` chooses the best predictor to split using an exhaustive search by default. Alternatively, you can choose to split the predictor that shows the most evidence of dependence with the response by conducting curvature tests. This example statistically compares classification trees grown via exhaustive search for the best splits and grown by conducting curvature tests with interaction.

Load the `census1994` data set.

```
load census1994.mat
rng(1) % For reproducibility
```

Grow a default classification tree using the training set, `adultdata`, which is a table. The response-variable name is `'salary'`.

```
C1 = fitctree(adultdata,'salary')
```

```
C1 =
  ClassificationTree
             PredictorNames: {1x14 cell}
               ResponseName: 'salary'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
             ScoreTransform: 'none'
            NumObservations: 32561


  Properties, Methods
```

`C1` is a full `ClassificationTree` model. Its `ResponseName` property is `'salary'`. `C1` uses an exhaustive search to find the best predictor to split on based on maximal splitting gain.

Grow another classification tree using the same data set, but specify to find the best predictor to split using the curvature test with interaction.

```
C2 = fitctree(adultdata,'salary','PredictorSelection','interaction-curvature')
```

```
C2 =
  ClassificationTree
             PredictorNames: {1x14 cell}
               ResponseName: 'salary'
    CategoricalPredictors: [2 4 6 7 8 9 10 14]
                ClassNames: [<=50K    >50K]
             ScoreTransform: 'none'
            NumObservations: 32561


  Properties, Methods
```

`C2` also is a full `ClassificationTree` model with `ResponseName` equal to `'salary'`.

Conduct a 5-by-2 paired *F* test to compare the accuracies of the two models using the training set. Because the response-variable names in the data sets and the `ResponseName` properties are all equal, and the response data in both sets are equal, you can omit supplying the response data.

```
h = testckfold(C1,C2,adultdata,adultdata)
```

```
h = logical
   0
```

`h = 0` indicates to not reject the null hypothesis that `C1` and `C2` have the same accuracies at 5% level.

Compare Accuracies of Two Different Classification Models

Conduct a statistical test comparing the misclassification rates of the two models using a 5-by-2 paired *F* test.

Load Fisher's iris data set.

```
load fisheriris;
```

Create a naive Bayes template and a classification tree template using default options.

```
C1 = templateNaiveBayes;
C2 = templateTree;
```

`C1` and `C2` are template objects corresponding to the naive Bayes and classification tree algorithms, respectively.

Test whether the two models have equal predictive accuracies. Use the same predictor data for each model. `testckfold` conducts a 5-by-2, two-sided, paired *F* test by default.

```
rng(1); % For reproducibility
h = testckfold(C1,C2,meas,meas,species)
```

```
h = logical
   0
```

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies.

Compare Classification Accuracies of Simple and Complex Models

Conduct a statistical test to assess whether a simpler model has better accuracy than a more complex model using a 10-by-10 repeated cross-validation *t* test.

Load Fisher's iris data set. Create a cost matrix that penalizes misclassifying a setosa iris twice as much as misclassifying a virginica iris as a versicolor.

```
load fisheriris;
tabulate(species)
```

```
       Value    Count   Percent
      setosa       50     33.33%
```

```
  versicolor       50     33.33%
   virginica       50     33.33%
```

```
Cost = [0 2 2;2 0 1;2 1 0];
ClassNames  = {'setosa' 'versicolor' 'virginica'};...
    % Specifies the order of the rows and columns in Cost
```

The empirical distribution of the classes is uniform, and the classification cost is slightly imbalanced.

Create two ECOC templates: one that uses linear SVM binary learners and one that uses SVM binary learners equipped with the RBF kernel.

```
tSVMLinear = templateSVM('Standardize',true); % Linear SVM by default
tSVMRBF = templateSVM('KernelFunction','RBF','Standardize',true);
C1 = templateECOC('Learners',tSVMLinear);
C2 = templateECOC('Learners',tSVMRBF);
```

`C1` and `C2` are ECOC template objects. `C1` is prepared for linear SVM. `C2` is prepared for SVM with an RBF kernel training.

Test the null hypothesis that the simpler model (`C1`) is at most as accurate as the more complex model (`C2`) in terms of classification costs. Conduct the 10-by-10 repeated cross-validation test. Request to return *p*-values and misclassification costs.

```
rng(1); % For reproducibility
[h,p,e1,e2] = testckfold(C1,C2,meas,meas,species,...
    'Alternative','greater','Test','10x10t','Cost',Cost,...
    'ClassNames',ClassNames)
```

```
h = logical
   0


p = 0.1077

e1 = 10×10

         0         0         0    0.0667         0    0.0667    0.1333         0    0.1333
    0.0667    0.0667         0         0         0         0    0.0667         0    0.0667    0.0
         0         0         0         0         0    0.0667    0.0667    0.0667    0.0667    0.0
    0.0667    0.0667         0    0.0667         0    0.0667         0         0    0.0667
    0.0667    0.0667    0.0667         0    0.0667    0.0667         0         0         0
         0         0    0.1333         0         0    0.0667         0         0    0.0667    0.0
    0.0667    0.0667         0         0    0.0667         0         0    0.0667         0    0.0
    0.0667         0    0.0667    0.0667         0    0.1333         0    0.0667         0
         0    0.0667    0.1333    0.0667    0.0667         0         0         0         0
         0    0.0667    0.0667    0.0667    0.0667         0         0    0.0667         0


e2 = 10×10

         0         0         0    0.1333         0    0.0667    0.1333         0    0.2667
    0.0667    0.0667         0    0.1333         0         0         0    0.1333    0.1333    0.0
    0.1333    0.1333         0         0         0    0.0667         0    0.0667    0.0667    0.0
         0    0.1333         0    0.0667    0.1333    0.1333         0         0    0.0667
    0.0667    0.0667    0.0667         0    0.0667    0.1333    0.1333         0         0    0.0
    0.0667         0    0.0667    0.0667         0    0.0667    0.1333         0    0.0667    0.0
    0.2000    0.0667         0         0    0.0667         0         0    0.1333         0    0.0
    0.2000         0         0    0.1333         0    0.1333         0    0.0667         0
```

```
         0    0.0667    0.0667    0.0667    0.1333         0    0.2000         0         0
    0.0667    0.0667         0    0.0667    0.1333         0         0    0.0667    0.1333    0.0
```

The *p*-value is slightly greater than 0.10, which indicates to retain the null hypothesis that the simpler model is at most as accurate as the more complex model. This result is consistent for any significance level (`Alpha`) that is at most 0.10.

`e1` and `e2` are 10-by-10 matrices containing misclassification costs. Row *r* corresponds to run *r* of the repeated cross validation. Column *k* corresponds to test-set fold *k* within a particular cross-validation run. For example, element (2,4) of `e2` is 0.1333. This value means that in cross-validation run 2, when the test set is fold 4, the estimated test-set misclassification cost is 0.1333.

Select Features Using Statistical Accuracy Comparison

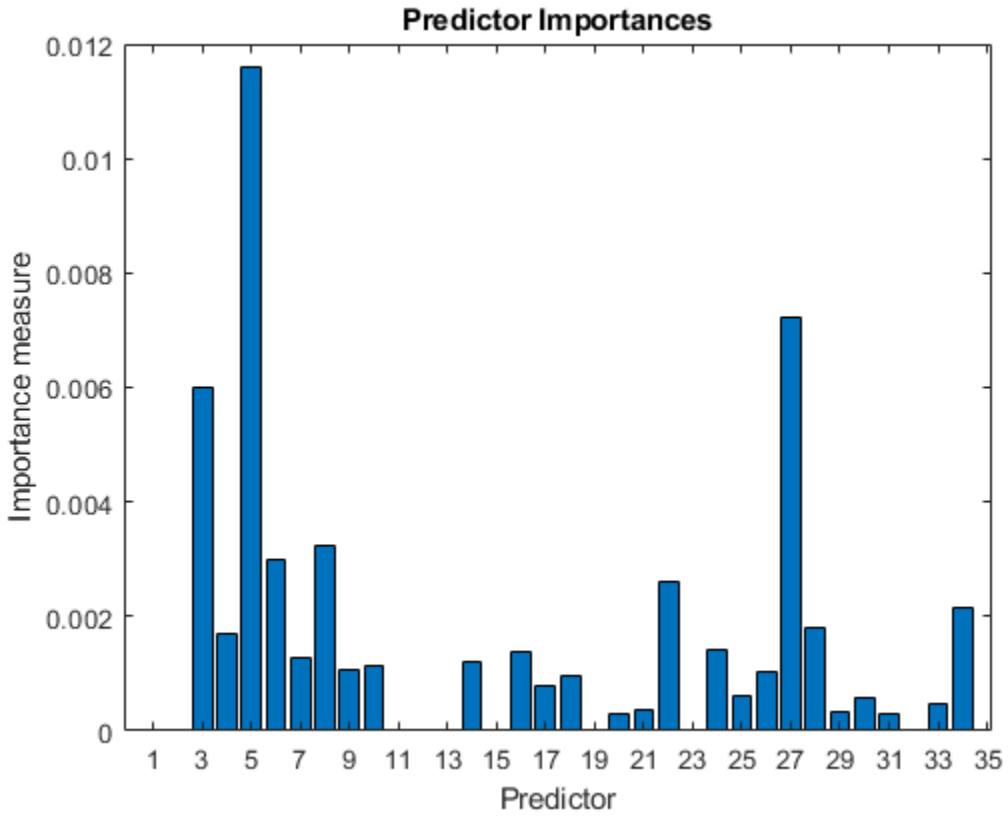
Reduce classification model complexity by selecting a subset of predictor variables (features) from a larger set. Then, statistically compare the accuracy between the two models.

Load the `ionosphere` data set.

```
load ionosphere
```

Train an ensemble of 100 boosted classification trees using AdaBoostM1 and the entire set of predictors. Inspect the importance measure for each predictor.

```
t = templateTree('MaxNumSplits',1); % Weak-learner template tree object
C = fitcensemble(X,Y,'Method','AdaBoostM1','Learners',t);
predImp = predictorImportance(C);

bar(predImp)
h = gca;
h.XTick = 1:2:h.XLim(2);
title('Predictor Importances')
xlabel('Predictor')
ylabel('Importance measure')
```



Identify the top five predictors in terms of their importance.

```
[~,idxSort] = sort(predImp,'descend');
idx5 = idxSort(1:5);
```

Test whether the two models have equal predictive accuracies. Specify the reduced data set and then the full predictor data. Use parallel computing to speed up computations.

```
s = RandStream('mlfg6331_64');
Options = statset('UseParallel',true,'Streams',s,'UseSubstreams',true);

[h,p,e1,e2] = testckfold(C,C,X(:,idx5),X,Y,'Options',Options)
```

```
Starting parallel pool (parpool) using the 'local' profile ...
Connected to the parallel pool (number of workers: 6).

h = logical
   0


p = 0.4161

e1 = 5×2

    0.0686    0.0795
    0.0800    0.0625
    0.0914    0.0568
    0.0400    0.0739
```

```
    0.0914    0.0966


e2 = 5×2

    0.0914    0.0625
    0.1257    0.0682
    0.0971    0.0625
    0.0800    0.0909
    0.0914    0.1193
```

`testckfold` treats trained classification models as templates, and so it ignores all fitted parameters in `C`. That is, `testckfold` cross validates `C` using only the specified options and the predictor data to estimate the out-of-fold classification losses.

`h = 0` indicates to not reject the null hypothesis that the two models have equal predictive accuracies. This result favors the simpler ensemble.

Input Arguments

`C1` — Classification model template or trained classification model

classification model template object | trained classification model object

Classification model template or trained classification model, specified as any classification model template object or trained classification model object described in these tables.

Template Type	Returned By
Classification tree	`templateTree`
Discriminant analysis	`templateDiscriminant`
Ensemble (boosting, bagging, and random subspace)	`templateEnsemble`
Error-correcting output codes (ECOC), multiclass classification model	`templateECOC`
*k*NN	`templateKNN`
Naive Bayes	`templateNaiveBayes`
Support Vector Machine (SVM)	`templateSVM`

Trained Model Type	Model Object	Returned By
Classification tree	`ClassificationTree`	`fitctree`
Discriminant analysis	`ClassificationDiscriminant`	`fitcdiscr`
Ensemble of bagged classification models	`ClassificationBaggedEnsemble`	`fitcensemble`
Ensemble of classification models	`ClassificationEnsemble`	`fitcensemble`
ECOC model	`ClassificationECOC`	`fitcecoc`

Trained Model Type	Model Object	Returned By
Generalized additive model (GAM)	`ClassificationGAM`	`fitcgam`
*k*NN	`ClassificationKNN`	`fitcknn`
Naive Bayes	`ClassificationNaiveBayes`	`fitcnb`
Neural network	`ClassificationNeuralNetwork` (with observations in rows)	`fitcnet`
SVM	`ClassificationSVM`	`fitcsvm`

For efficiency, supply a classification model template object instead of a trained classification model object.

C2 — Classification model template or trained model

classification model template object | trained classification model object

Classification model template or trained classification model, specified as any classification model template object or trained classification model object described in these tables.

Template Type	Returned By
Classification tree	`templateTree`
Discriminant analysis	`templateDiscriminant`
Ensemble (boosting, bagging, and random subspace)	`templateEnsemble`
Error-correcting output codes (ECOC), multiclass classification model	`templateECOC`
*k*NN	`templateKNN`
Naive Bayes	`templateNaiveBayes`
Support Vector Machine (SVM)	`templateSVM`

Trained Model Type	Model Object	Returned By
Classification tree	`ClassificationTree`	`fitctree`
Discriminant analysis	`ClassificationDiscriminant`	`fitcdiscr`
Ensemble of bagged classification models	`ClassificationBaggedEnsemble`	`fitcensemble`
Ensemble of classification models	`ClassificationEnsemble`	`fitcensemble`
ECOC model	`ClassificationECOC`	`fitcecoc`
Generalized additive model (GAM)	`ClassificationGAM`	`fitcgam`
*k*NN	`ClassificationKNN`	`fitcknn`
Naive Bayes	`ClassificationNaiveBayes`	`fitcnb`
Neural network	`ClassificationNeuralNetwork` (with observations in rows)	`fitcnet`

Trained Model Type	Model Object	Returned By
SVM	`ClassificationSVM`	`fitcsvm`

For efficiency, supply a classification model template object instead of a trained classification model object.

X1 — Data used to apply to first full classification model or template
numeric matrix | table

Data used to apply to the first full classification model or template, `C1`, specified as a numeric matrix or table.

Each row of `X1` corresponds to one observation, and each column corresponds to one variable. `testckfold` does not support multicolumn variables and cell arrays other than cell arrays of character vectors.

`X1` and `X2` must be of the same data type, and `X1`, `X2`, `Y` must have the same number of observations.

If you specify `Y` as an array, then `testckfold` treats all columns of `X1` as separate predictor variables.

Data Types: `double` | `single` | `table`

X2 — Data used to apply to second full classification model or template
numeric matrix | table

Data used to apply to the second full classification model or template, `C2`, specified as a numeric matrix or table.

Each row of `X2` corresponds to one observation, and each column corresponds to one variable. `testckfold` does not support multicolumn variables and cell arrays other than cell arrays of character vectors.

`X1` and `X2` must be of the same data type, and `X1`, `X2`, `Y` must have the same number of observations.

If you specify `Y` as an array, then `testckfold` treats all columns of `X2` as separate predictor variables.

Data Types: `double` | `single` | `table`

Y — True class labels
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors | character vector | string scalar

True class labels, specified as a categorical, character, or string array, a logical or numeric vector, a cell array of character vectors, or a character vector or string scalar.

- For a character vector or string scalar, `X1` and `X2` must be tables, their response variables must have the same name and values, and `Y` must be the common variable name. For example, if `X1.Labels` and `X2.Labels` are the response variables, then `Y` is `'Labels'` and `X1.Labels` and `X2.Labels` must be equivalent.
- For all other supported data types, `Y` is an array of true class labels.
  - If `Y` is a character array, then each element must correspond to one row of the array.

- `X1`, `X2`, `Y` must have the same number of observations (rows).

- If both of these statements are true, then you can omit supplying `Y`.

  - `X1` and `X2` are tables containing the same response variable (values and name).
  - `C1` and `C2` are full classification models containing `ResponseName` properties specifying the response variable names in `X1` and `X2`.

  Consequently, `testckfold` uses the common response variable in the tables. For example, if the response variables in the tables are `X1.Labels` and `X2.Labels`, and the values of `C1.ResponseName` and `C2.ResponseName` are `'Labels'`, then you do not have to supply `Y`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Alternative','greater','Test','10x10t','Options',statsset('UseParallel',true)` specifies to test whether the first set of first predicted class labels is more accurate than the second set, to conduct the 10-by-10 t test, and to use parallel computing for cross-validation.

`Alpha` — Hypothesis test significance level

`0.05` (default) | scalar value in the interval (0,1)

Hypothesis test significance level, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the interval (0,1).

Example: `'Alpha',0.1`

Data Types: `single` | `double`

`Alternative` — Alternative hypothesis to assess

`'unequal'` (default) | `'greater'` | `'less'`

Alternative hypothesis to assess, specified as the comma-separated pair consisting of `'Alternative'` and one of the values listed in the table.

Value	Alternative Hypothesis Description	Supported Tests
`'unequal'` (default)	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` and `C2` applied to `X2` have unequal accuracies.	`'5x2F'`, `'5x2t'`, and `'10x10t'`
`'greater'`	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` is more accurate than `C2` applied to `X2`.	`'5x2t'` and `'10x10t'`
`'less'`	For predicting `Y`, the set of predictions resulting from `C1` applied to `X1` is less accurate than `C2` applied to `X2`.	`'5x2t'` and `'10x10t'`

For details on supported tests, see `Test`.

Example: `'Alternative','greater'`

`X1CategoricalPredictors` — Flag identifying categorical predictors
`[]` (default) | logical vector | numeric vector | `'all'`

Flag identifying categorical predictors in the first test-set predictor data (`X1`), specified as the comma-separated pair consisting of `'X1CategoricalPredictors'` and one of the following:

- A numeric vector with indices from `1` through `p`, where `p` is the number of columns of `X1`.
- A logical vector of length `p`, where a `true` entry means that the corresponding column of `X1` is a categorical variable.
- `'all'`, meaning all predictors are categorical.

Specification of `X1CategoricalPredictors` is appropriate if:

- At least one predictor is categorical and `C1` is a classification tree, an ensemble of classification trees, an ECOC model, or a naive Bayes classification model.
- All predictors are categorical and `C1` is a *k*NN classification model.

If you specify `X1CategoricalPredictors` for any other case, then `testckfold` throws an error. For example, the function cannot train SVM learners using categorical predictors.

The default is `[]`, which indicates that there are no categorical predictors.

Example: `'X1CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string`

`X2CategoricalPredictors` — Flag identifying categorical predictors
`[]` (default) | logical vector | numeric vector | `'all'`

Flag identifying categorical predictors in the second test-set predictor data (`X2`), specified as the comma-separated pair consisting of `'X2CategoricalPredictors'` and one of the following:

- A numeric vector with indices from `1` through `p`, where `p` is the number of columns of `X2`.
- A logical vector of length `p`, where a `true` entry means that the corresponding column of `X2` is a categorical variable.
- `'all'`, meaning all predictors are categorical.

Specification of `X2CategoricalPredictors` is appropriate if:

- At least one predictor is categorical and `C2` is a classification tree, an ensemble of classification trees, an ECOC model, or a naive Bayes classification model.
- All predictors are categorical and `C2` is a *k*NN classification model.

If you specify `X2CategoricalPredictors` for any other case, then `testckfold` throws an error. For example, the function cannot train SVM learners using categorical predictors.

The default is `[]`, which indicates that there are no categorical predictors.

Example: `'X2CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string`

`ClassNames` — Class names
categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class names, specified as the comma-separated pair consisting of `'ClassNames'` and a categorical, character, or string array, logical or numeric vector, or cell array of character vectors. You must set `ClassNames` using the data type of `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of any input argument dimension that corresponds to class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost`.
- Select a subset of classes for testing. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train and test models using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default is the set of all distinct class names in `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell` | `categorical`

`Cost` — Classification cost
square matrix | structure array

Classification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure array.

- If you specify the square matrix `Cost`, then `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `Cost`, additionally specify the `ClassNames` name-value pair argument.
- If you specify the structure `S`, then `S` must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`. You can use this field to specify the order of the classes.
  - `S.ClassificationCosts`, which contains the cost matrix, with rows and columns ordered as in `S.ClassNames`

For cost-sensitive testing use, `testcholdout`.

It is a best practice to supply the same cost matrix used to train the classification models.

The default is `Cost(i,j) = 1` if `i ~= j`, and `Cost(i,j) = 0` if `i = j`.

Example: `'Cost',[0 1 2 ; 1 0 2; 2 2 0]`

Data Types: `double` | `single` | `struct`

`LossFun` — Loss function
`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and `'classiferror'`, `'binodeviance'`, `'exponential'`, `'hinge'`, or a function handle.

- The following table lists the available loss functions.

Value	Loss Function
'binodeviance'	Binomial deviance
'classiferror'	Classification error
'exponential'	Exponential loss
'hinge'	Hinge loss

- Specify your own function using function handle notation.

  Suppose that `n = size(X,1)` is the sample size and there are `K` unique classes. Your function must have the signature `lossvalue = lossfun(C,S,W,Cost)`, where:

  - The output argument `lossvalue` is a scalar.
  - *`lossfun`* is the name of your function.
  - `C` is an `n`-by-`K` logical matrix with rows indicating which class the corresponding observation belongs to. The column order corresponds to the class order in the `ClassNames` name-value pair argument.

    Construct `C` by setting `C(p,q) = 1` if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.
  - `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in the `ClassNames` name-value pair argument. `S` is a matrix of classification scores.
  - `W` is an `n`-by-1 numeric vector of observation weights. If you pass `W`, the software normalizes the weights to sum to `1`.
  - `Cost` is a `K`-by-`K` numeric matrix of classification costs. For example, `Cost = ones(K) - eye(K)` specifies a cost of `0` for correct classification and a cost of `1` for misclassification.

  Specify your function using `'LossFun',@lossfun`.

`Options` — Parallel computing options
`[]` (default) | structure array returned by `statset`

Parallel computing options, specified as the comma-separated pair consisting of `'Options'` and a structure array returned by `statset`. These options require Parallel Computing Toolbox. `testckfold` uses `'Streams'`, `'UseParallel'`, and `'UseSubtreams'` fields.

This table summarizes the available options.

Option	Description
`'Streams'`	A `RandStream` object or cell array of such objects. If you do not specify `Streams`, the software uses the default stream or streams. If you specify `Streams`, use a single object except when the following are true: • You have an open parallel pool. • `UseParallel` is `true`. • `UseSubstreams` is `false`. In that case, use a cell array of the same size as the parallel pool. If a parallel pool is not open, then the software tries to open one (depending on your preferences), and `Streams` must supply a single random number stream.
`'UseParallel'`	If you have Parallel Computing Toolbox, then you can invoke a pool of workers by setting `'UseParallel',true`.
`'UseSubstreams'`	Set to `true` to compute in parallel using the stream specified by `'Streams'`. Default is `false`. For example, set `Streams` to a type allowing substreams, such as `'mlfg6331_64'` or `'mrg32k3a'`.

Example: `'Options',statset('UseParallel',true)`

Data Types: `struct`

`Prior` — Prior probabilities

`'empirical'` (default) | `'uniform'` | numeric vector | structure

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and `'empirical'`, `'uniform'`, a numeric vector, or a structure.

This table summarizes the available options for setting prior probabilities.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Specify the order using the `ClassNames` name-value pair argument. The software normalizes the elements such that they sum to `1`.

Value	Description
structure	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities. The software normalizes the elements such that they sum to `1`.

Example: `'Prior',struct('ClassNames',{{'setosa','versicolor'}},'ClassProbs',[1,2])`

Data Types: `char` | `string` | `single` | `double` | `struct`

`Test` — Test to conduct
`'5x2F'` (default) | `'5x2t'` | `'10x10t'`

Test to conduct, specified as the comma-separated pair consisting of `'Test'` and one of he following: `'5x2F'`, `'5x2t'`, `'10x10t'`.

Value	Description	Supported Alternative Hypothesis
`'5x2F'` (default)	5-by-2 paired *F* test. Appropriate for two-sided testing only.	`'unequal'`
`'5x2t'`	5-by-2 paired *t* test	`'unequal'`, `'less'`, `'greater'`
`'10x10t'`	10-by-10 repeated cross-validation *t* test	`'unequal'`, `'less'`, `'greater'`

For details on the available tests, see “Repeated Cross-Validation Tests” on page 33-6218. For details on supported alternative hypotheses, see `Alternative`.

Example: `'Test','10x10t'`

`Verbose` — Verbosity level
`0` (default) | `1` | `2`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and `0`, `1`, or `2`. `Verbose` controls the amount of diagnostic information that the software displays in the Command Window during training of each cross-validation fold.

This table summarizes the available verbosity level options.

Value	Description
`0`	The software does not display diagnostic information.
`1`	The software displays diagnostic messages every time it implements a new cross-validation run.

Value	Description
2	The software displays diagnostic messages every time it implements a new cross-validation run, and every time it trains on a particular fold.

Example: `'Verbose',1`

Data Types: `double` | `single`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector.

The size of `Weights` must equal the number of rows of `X1`. The software weighs the observations in each row of `X` with the corresponding weight in `Weights`.

The software normalizes `Weights` to sum up to the value of the prior probability in the respective class.

Data Types: `double` | `single`

Notes:

- `testckfold` treats trained classification models as templates. Therefore, it ignores all fitted parameters in the model. That is, `testckfold` cross-validates using only the options specified in the model and the predictor data.
- The repeated cross-validation tests depend on the assumption that the test statistics are asymptotically normal under the null hypothesis. Highly imbalanced cost matrices (for example, `Cost = [0 100;1 0]`) and highly discrete response distributions (that is, most of the observations are in a small number of classes) might violate the asymptotic normality assumption. For cost-sensitive testing, use `testcholdout`.
- `NaN`s, `<undefined>` values, empty character vectors (`''`), empty strings (`""`), and `<missing>` values indicate missing data values.
  - For the treatment of missing values in `X1` and `X2`, see the appropriate classification model training function reference page: `fitctree`, `fitcdiscr`, `fitcensemble`, `fitcecoc`, `fitcgam`, `fitcknn`, `fitcnb`, or `fitcsvm`.
  - `testckfold` removes missing values in `Y` and the corresponding rows of `X1` and `X2`.

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as a logical value.

`h = 1` indicates the rejection of the null hypothesis at the `Alpha` significance level.

`h = 0` indicates failure to reject the null hypothesis at the `Alpha` significance level.

Data Types: `logical`

p — *p*-value
scalar in the interval [0,1]

p-value of the test, returned as a scalar in the interval [0,1]. `p` is the probability that a random test statistic is at least as extreme as the observed test statistic, given that the null hypothesis is true.

`testckfold` estimates `p` using the distribution of the test statistic, which varies with the type of test. For details on test statistics, see “Repeated Cross-Validation Tests” on page 33-6218.

e1 — Classification losses
numeric matrix

Classification losses on page 33-6220, returned as a numeric matrix. The rows of `e1` correspond to the cross-validation run and the columns correspond to the test fold.

`testckfold` applies the first test-set predictor data (`X1`) to the first classification model (`C1`) to estimate the first set of class labels.

`e1` summarizes the accuracy of the first set of class labels predicting the true class labels (`Y`) for each cross-validation run and fold. The meaning of the elements of `e1` depends on the type of classification loss.

e2 — Classification losses
numeric matrix

Classification losses on page 33-6220, returned as a numeric matrix. The rows of `e2` correspond to the cross-validation run and the columns correspond to the test fold.

`testckfold` applies the second test-set predictor data (`X2`) to the second classification model (`C2`) to estimate the second set of class labels.

`e2` summarizes the accuracy of the second set of class labels predicting the true class labels (`Y`) for each cross-validation run and fold. The meaning of the elements of `e2` depends on the type of classification loss.

More About

Repeated Cross-Validation Tests

Repeated cross-validation tests form the test statistic for comparing the accuracies of two classification models by combining the classification loss differences resulting from repeatedly cross-validating the data. Repeated cross-validation tests are useful when sample size is limited.

To conduct an *R*-by-*K* test:

1. Randomly divide (stratified by class) the predictor data sets and true class labels into *K* sets, *R* times. Each division is called a run and each set within a run is called a fold. Each run contains the complete, but divided, data sets.
2. For runs $r = 1$ through *R*, repeat these steps for $k = 1$ through *K*:
    a. Reserve fold *k* as a test set, and train the two classification models using their respective predictor data sets on the remaining $K - 1$ folds.

b Predict class labels using the trained models and their respective fold k predictor data sets.

c Estimate the classification loss by comparing the two sets of estimated labels to the true labels. Denote e_{crk} as the classification loss when the test set is fold k in run r of classification model c.

d Compute the difference between the classification losses of the two models:

$$\widehat{\delta}_{rk} = e_{1rk} - e_{2rk}.$$

At the end of a run, there are K classification losses per classification model.

3 Combine the results of step 2. For each $r = 1$ through R:

- Estimate the within-fold averages of the differences and their average: $\bar{\delta}_r = \frac{1}{K}\sum_{k=1}^{K}\widehat{\delta}_{kr}$.
- Estimate the overall average of the differences: $\bar{\delta} = \frac{1}{KR}\sum_{r=1}^{R}\sum_{k=1}^{K}\widehat{\delta}_{rk}$.
- Estimate the within-fold variances of the differences: $s_r^2 = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\delta}_{rk} - \bar{\delta}_r\right)^2$.
- Estimate the average of the within-fold differences: $\bar{s}^2 = \frac{1}{R}\sum_{r=1}^{R}s_r^2$.
- Estimate the overall sample variance of the differences: $S^2 = \frac{1}{KR-1}\sum_{r=1}^{R}\sum_{k=1}^{K}\left(\widehat{\delta}_{rk} - \bar{\delta}\right)^2$.

Compute the test statistic. All supported tests described here assume that, under H_0, the estimated differences are independent and approximately normally distributed, with mean 0 and a finite, common standard deviation. However, these tests violate the independence assumption, and so the test-statistic distributions are approximate.

- For $R = 2$, the test is a paired test. The two supported tests are a paired t and F test.
  - The test statistic for the paired t test is

$$t^*_{paired} = \frac{\widehat{\delta}_{11}}{\sqrt{\bar{s}^2}}.$$

    t^*_{paired} has a t-distribution with R degrees of freedom under the null hypothesis.

    To reduce the effects of correlation between the estimated differences, the quantity $\widehat{\delta}_{11}$ occupies the numerator rather than $\bar{\delta}$.

    5-by-2 paired t tests can be slightly conservative [4].
  - The test statistic for the paired F test is

$$F^*_{paired} = \frac{\frac{1}{RK}\sum_{r=1}^{R}\sum_{k=1}^{K}\left(\widehat{\delta}_{rk}\right)^2}{\bar{s}^2}.$$

F^*_{paired} has an F distribution with RK and R degrees of freedom.

A 5-by-2 paired F test has comparable power to the 5-by-2 t test, but is more conservative [1].

- For $R > 2$, the test is a repeated cross-validation test. The test statistic is

$$t^*_{CV} = \frac{\bar{\delta}}{S/\sqrt{\nu + 1}}.$$

t^*_{CV} has a t distribution with ν degrees of freedom. If the differences were truly independent, then $\nu = RK - 1$. In this case, the degrees of freedom parameter must be optimized.

For a 10-by-10 repeated cross-validation t test, the optimal degrees of freedom between 8 and 11 ([2] and [3]). `testckfold` uses $\nu = 10$.

The advantage of repeated cross-validation tests over paired tests is that the results are more repeatable [3]. The disadvantage is that they require high computational resources.

Classification Loss

Classification losses indicate the accuracy of a classification model or set of predicted labels. In general, for a fixed cost matrix, classification accuracy decreases as classification loss increases.

`testckfold` returns the classification losses (see `e1` and `e2`) under the alternative hypothesis (that is, the unrestricted classification losses). In the definitions that follow:

- The classification losses focus on the first classification model. The classification losses for the second model are similar.
- n_{test} is the test-set sample size.
- $I(x)$ is the indicator function. If x is a true statement, then $I(x) = 1$. Otherwise, $I(x) = 0$.
- $\widehat{p}_{1j}$ is the predicted class assignment of classification model 1 for observation j.
- y_j is the true class label of observation j.
- Binomial deviance has the form

$$e_1 = \frac{\sum_{j=1}^{n_{test}} w_j \log\left(1 + \exp\left(-2y'_j f(X_j)\right)\right)}{\sum_{j=1}^{n_{test}} w_j}$$

where:

- $y_j = 1$ for the positive class and -1 for the negative class.
- $f(X_j)$ is the classification score.

The binomial deviance has connections to the maximization of the binomial likelihood function. For details on binomial deviance, see [5].

- Exponential loss is similar to binomial deviance and has the form

$$e_1 = \frac{\sum_{j=1}^{n_{test}} w_j \exp\left(-y_j f(X_j)\right)}{\sum_{j=1}^{n_{test}} w_j}.$$

y_j and $f(X_j)$ take the same forms here as in the binomial deviance formula.

- Hinge loss has the form

$$e_1 = \frac{\sum_{j=1}^{n} w_j \max\left\{0, 1 - y_j' f\left(X_j\right)\right\}}{\sum_{j=1}^{n} w_j},$$

y_j and $f(X_j)$ take the same forms here as in the binomial deviance formula.

Hinge loss linearly penalizes for misclassified observations and is related to the SVM objective function used for optimization. For more details on hinge loss, see [5].

- Misclassification rate, or classification error, is a scalar in the interval [0,1] representing the proportion of misclassified observations. That is, the misclassification rate for the first classification model is

$$e_1 = \frac{\sum_{j=1}^{n_{test}} w_j I(\widehat{p}_{1j} \neq y_j)}{\sum_{j=1}^{n_{test}} w_j}.$$

Tips

- Examples of ways to compare models include:
  - Compare the accuracies of a simple classification model and a more complex model by passing the same set of predictor data.
  - Compare the accuracies of two different models using two different sets of predictors.
  - Perform various types of Feature Selection on page 15-49. For example, you can compare the accuracy of a model trained using a set of predictors to the accuracy of one trained on a subset or different set of predictors. You can arbitrarily choose the set of predictors, or use a feature selection technique like PCA or sequential feature selection (see `pca` and `sequentialfs`).
- If both of these statements are true, then you can omit supplying `Y`.
  - `X1` and `X2` are tables containing the response variable and use the same response variable name.
  - `C1` and `C2` are full classification models containing equal `ResponseName` properties (e.g. `strcmp(C1.ResponseName,C2.ResponseName) = 1`).

  Consequently, `testckfold` uses the common response variable in the tables.
- One way to perform cost-insensitive feature selection is:

1. Create a classification model template that characterizes the first classification model (`C1`).
2. Create a classification model template that characterizes the second classification model (`C2`).
3. Specify two predictor data sets. For example, specify `X1` as the full predictor set and `X2` as a reduced set.
4. Enter `testckfold(C1,C2,X1,X2,Y,'Alternative','less')`. If `testckfold` returns `1`, then there is enough evidence to suggest that the classification model that uses fewer predictors performs better than the model that uses the full predictor set.

Alternatively, you can assess whether there is a significant difference between the accuracies of the two models. To perform this assessment, remove the `'Alternative','less'` specification in step 4.`testckfold` conducts a two-sided test, and `h = 0` indicates that there is not enough evidence to suggest a difference in the accuracy of the two models.

- The tests are appropriate for the misclassification rate classification loss on page 33-6220, but you can specify other loss functions (see `LossFun`). The key assumptions are that the estimated classification losses are independent and normally distributed with mean 0 and finite common variance under the two-sided null hypothesis. Classification losses other than the misclassification rate can violate this assumption.
- Highly discrete data, imbalanced classes, and highly imbalanced cost matrices can violate the normality assumption of classification loss differences.

Algorithms

If you specify to conduct the 10-by-10 repeated cross-validation *t* test using `'Test','10x10t'`, then `testckfold` uses 10 degrees of freedom for the *t* distribution to find the critical region and estimate the *p*-value. For more details, see [2] and [3].

Alternatives

Use `testcholdout`:

- For test sets with larger sample sizes
- To implement variants of the McNemar test to compare two classification model accuracies
- For cost-sensitive testing using a chi-square or likelihood ratio test. The chi-square test uses `quadprog`, which requires an Optimization Toolbox license.

References

[1] Alpaydin, E. “Combined 5 x 2 CV F Test for Comparing Supervised Classification Learning Algorithms.” *Neural Computation*, Vol. 11, No. 8, 1999, pp. 1885–1992.

[2] Bouckaert. R. “Choosing Between Two Learning Algorithms Based on Calibrated Tests.” *International Conference on Machine Learning*, 2003, pp. 51–58.

[3] Bouckaert, R., and E. Frank. “Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms.” *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference*, 2004, pp. 3–12.

[4] Dietterich, T. “Approximate statistical tests for comparing supervised classification learning algorithms.” *Neural Computation*, Vol. 10, No. 7, 1998, pp. 1895–1923.

[5] Hastie, T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, 2nd Ed. New York: Springer, 2008.

Extended Capabilities

Automatic Parallel Support

Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`templateDiscriminant` | `templateECOC` | `templateEnsemble` | `templateKNN` | `templateNaiveBayes` | `templateSVM` | `templateTree` | `testcholdout`

Topics

“Hypothesis Tests”

Introduced in R2015a

tiedrank

Rank adjusted for ties

Syntax

```
[R,TIEADJ] = tiedrank(X)
[R,TIEADJ] = tiedrank(X,1)
[R,TIEADJ] = tiedrank(X,0,1)
```

Description

`[R,TIEADJ] = tiedrank(X)` computes the ranks of the values in the vector `X`. If any `X` values are tied, `tiedrank` computes their average rank. The return value `TIEADJ` is an adjustment for ties required by the nonparametric tests `signrank` and `ranksum`, and for the computation of Spearman's rank correlation.

`[R,TIEADJ] = tiedrank(X,1)` computes the ranks of the values in the vector `X`. `TIEADJ` is a vector of three adjustments for ties required in the computation of Kendall's tau. `tiedrank(X,0)` is the same as `tiedrank(X)`.

`[R,TIEADJ] = tiedrank(X,0,1)` computes the ranks from each end, so that the smallest and largest values get rank 1, the next smallest and largest get rank 2, etc. These ranks are used in the Ansari-Bradley test.

Examples

Counting from smallest to largest, the two 20 values are 2nd and 3rd, so they both get rank 2.5 (average of 2 and 3):

```
tiedrank([10 20 30 40 20])
ans =
    1.0000    2.5000    4.0000    5.0000    2.5000
```

Algorithms

`tiedrank` treats `NaN`s in `X` as missing values and ignores them. The rank of `NaN`s in the output argument `R` is `NaN`.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`ansaribradley` | `corr` | `partialcorr` | `ranksum` | `signrank`

Introduced before R2006a

tinv

Student's *t* inverse cumulative distribution function

Syntax

```
x = tinv(p,nu)
```

Description

`x = tinv(p,nu)` returns the inverse cumulative distribution function (icdf) of the Student's *t* distribution evaluated at the probability values in `p` using the corresponding degrees of freedom in `nu`.

Examples

Compute Student's *t* icdf

Find the 95th percentile of the Student's *t* distribution with `50` degrees of freedom.

```
p = .95;
nu = 50;
x = tinv(p,nu)
```

```
x = 1.6759
```

Compute Student's *t* icdf for Multiple Distributions

Compute the 99th percentile of the Student's *t* distribution for `1` to `6` degrees of freedom.

```
percentile = tinv(0.99,1:6)
```

```
percentile = 1×6

   31.8205    6.9646    4.5407    3.7469    3.3649    3.1427
```

Compute Confidence Interval Using Student's *t* icdf

Find a 95% confidence interval estimating the mean of a population by using `tinv`.

Generate a random sample of size `100` drawn from a normal population with mean `10` and standard deviation `2`.

```
mu = 10;
sigma = 2;
n = 100;
```

```
rng default   % For reproducibility
x = normrnd(mu,sigma,n,1);
```

Compute the sample mean, standard error, and degrees of freedom.

```
xbar = mean(x);
se = std(x)/sqrt(n);
nu = n - 1;
```

Find the upper and lower confidence bounds for the 95% confidence interval.

```
conf = 0.95;
alpha = 1 - conf;
pLo = alpha/2;
pUp = 1 - alpha/2;
```

Compute the critical values for the confidence bounds.

```
crit = tinv([pLo pUp], nu);
```

Determine the confidence interval for the population mean.

```
ci = xbar + crit*se
```

```
ci = 1×2

    9.7849   10.7075
```

This confidence interval is the same as the `ci` value returned by a `t` test of a null hypothesis that the sample comes from a normal population with mean `mu`.

```
[h,p,ci2] = ttest(x,mu,'Alpha',alpha);
ci2
```

```
ci2 = 2×1

    9.7849
   10.7075
```

Input Arguments

p — Probability values at which to evaluate icdf

scalar value in `[0,1]` | array of scalar values

Probability values at which to evaluate the icdf, specified as a scalar value or an array of scalar values, where each element is in the range `[0,1]`.

- To evaluate the icdf at multiple values, specify `p` using an array.
- To evaluate the icdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `p` and `nu` are arrays, then the array sizes must be the same. In this case, `tinv` expands each scalar input into a constant array of the same size as the array inputs. Each element in `x` is the icdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding probability in `p`.

Example: `[0.1 0.5 0.9]`

Data Types: `single` | `double`

nu — Degrees of freedom
positive scalar value | array of positive scalar values

Degrees of freedom for the Student's *t* distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the icdf at multiple values, specify p using an array.
- To evaluate the icdfs of multiple distributions, specify nu using an array.

If either or both of the input arguments p and nu are arrays, then the array sizes must be the same. In this case, `tinv` expands each scalar input into a constant array of the same size as the array inputs. Each element in x is the icdf value of the distribution specified by the corresponding element in nu, evaluated at the corresponding probability in p.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

Output Arguments

x — icdf values
scalar value | array of scalar values

icdf values evaluated at the probabilities in p, returned as a scalar value or an array of scalar values. x is the same size as p and nu after any necessary scalar expansion. Each element in x is the icdf value of the distribution specified by the corresponding element in nu, evaluated at the corresponding probability in p.

More About

Student's *t* icdf

The Student's *t* distribution is a one-parameter family of curves. The parameter ν is the degrees of freedom. The Student's *t* distribution has zero mean.

The *t* inverse function is defined in terms of the Student's *t* cdf as

$$x = F^{-1}(p|\nu) = \left\{x : F(x|\nu) = p\right\},$$

where

$$p = F(x|\nu) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} dt,$$

ν is the degrees of freedom, and $\Gamma(\cdot)$ is the Gamma function. The result *x* is the solution of the integral equation where you supply the probability *p*.

For more information, see “Student's t Distribution” on page B-149.

Alternative Functionality

- `tinv` is a function specific to the Student's *t* distribution. Statistics and Machine Learning Toolbox also offers the generic function `icdf`, which supports various probability distributions. To use `icdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `tinv` is faster than the generic function `icdf`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`icdf` | `tcdf` | `tpdf` | `trnd` | `tstat` | `ttest` | `ttest2`

Topics

“Student's t Distribution” on page B-149

Introduced before R2006a

tpdf

Student's *t* probability density function

Syntax

```
y = tpdf(x,nu)
```

Description

`y = tpdf(x,nu)` returns the probability density function (pdf) of the Student's *t* distribution with `nu` degrees of freedom, evaluated at the values in `x`.

Examples

Compute Student's *t* Distribution pdf

The value of the pdf at the mode is an increasing function of the degrees of freedom.

The mode of the Student's *t* distribution is at $x = 0$. Compute the pdf at the mode for degrees of freedom `1` to `6`.

```
tpdf(0,1:6)
```

```
ans = 1×6

    0.3183    0.3536    0.3676    0.3750    0.3796    0.3827
```

The *t* distribution converges to the standard normal distribution as the degrees of freedom approach infinity.

Compute the difference between the pdfs of the standard normal distribution and the Student's *t* distribution pdf with `30` degrees of freedom.

```
difference = tpdf(-2.5:2.5,30)-normpdf(-2.5:2.5)
```

```
difference = 1×6

    0.0035   -0.0006   -0.0042   -0.0042   -0.0006    0.0035
```

Input Arguments

x — Values at which to evaluate pdf

scalar value | array of scalar values

Values at which to evaluate the pdf, specified as a scalar value or an array of scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.

- To evaluate the pdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `tpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[-1 0 3 4]`

Data Types: `single` | `double`

`nu` — degrees of freedom
positive scalar value | array of positive scalar values

Degrees of freedom for the Student's *t* distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `nu` using an array.

If either or both of the input arguments `x` and `nu` are arrays, then the array sizes must be the same. In this case, `tpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

Output Arguments

`y` — pdf values
scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `p` is the same size as `x` and `nu` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding element in `nu`, evaluated at the corresponding element in `x`.

More About

Student's *t* pdf

The Student's *t* distribution is a one-parameter family of curves. The parameterν is the degrees of freedom. The Student's *t* distribution has zero mean.

The pdf of the Student's *t* distribution is

$$y = f(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result y is the probability of observing a particular value of x from the Student's t distribution with ν degrees of freedom.

For more information, see "Student's t Distribution" on page B-149.

Alternative Functionality

- `tpdf` is a function specific to the Student's t distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, specify the probability distribution name and its parameters. Note that the distribution-specific function `tpdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`pdf` | `tcdf` | `tinv` | `trnd` | `tstat` | `ttest` | `ttest2`

Topics

"Student's t Distribution" on page B-149

Introduced before R2006a

training

Training indices for cross-validation

Syntax

```
idx = training(c)
idx = training(c,i)
```

Description

`idx = training(c)` returns the training indices `idx` for a `cvpartition` object `c` of type `'holdout'` or `'resubstitution'`.

- If `c.Type` is `'holdout'`, then `idx` specifies the observations in the training set.
- If `c.Type` is `'resubstitution'`, then `idx` specifies all observations.

`idx = training(c,i)` returns the training indices for repetition `i` of a `cvpartition` object `c` of type `'kfold'` or `'leaveout'`.

- If `c.Type` is `'kfold'`, then `idx` specifies the observations in the `i`th training set.
- If `c.Type` is `'leaveout'`, then `idx` specifies the observations reserved for training at repetition `i`.

Examples

Identify Training Indices in Holdout Partition

Identify the observations that are in the training set of a `cvpartition` object for holdout validation.

Partition 10 observations for holdout validation. Select approximately 30% of the observations to be in the test (holdout) set.

```
rng('default') % For reproducibility
c = cvpartition(10,'Holdout',0.30)
```

```
c =
Hold-out cross validation partition
   NumObservations: 10
       NumTestSets: 1
         TrainSize: 7
          TestSize: 3
```

Identify the training set observations. Observations that correspond to 1s are in the training set.

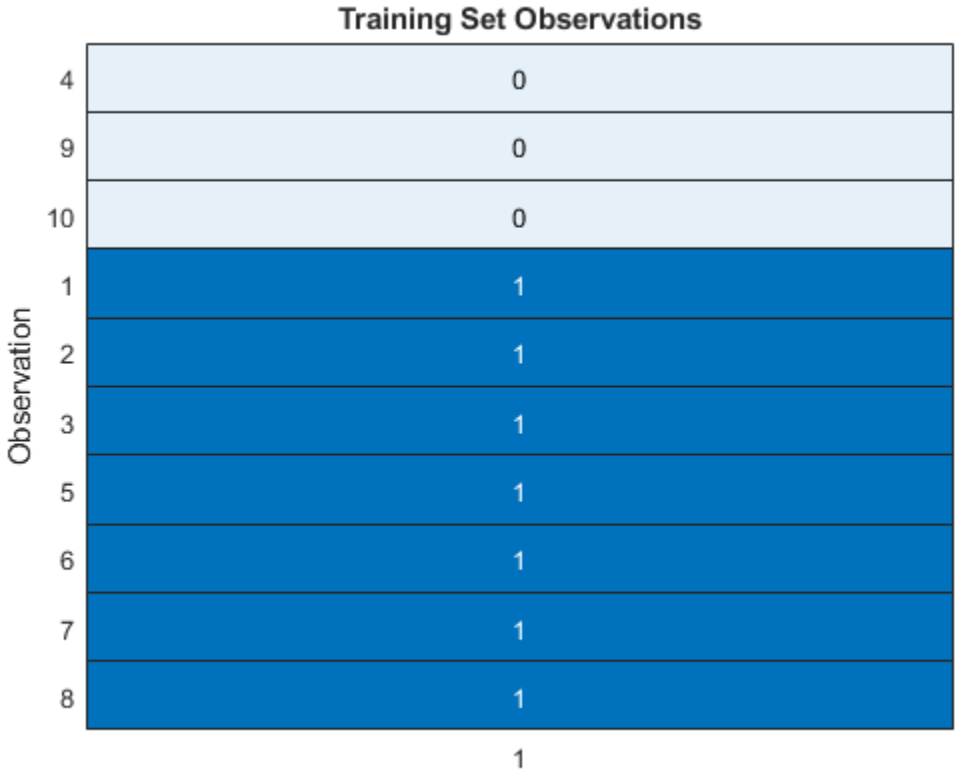
```
set = training(c)
```

```
set = 10x1 logical array

   1
   1
```

```
1
0
1
1
1
1
0
0
```

Visualize the results. All observations except the fourth, ninth, and tenth are in the training set.

```
h = heatmap(double(set),'ColorbarVisible','off');
sorty(h,'1','ascend')
ylabel('Observation')
title('Training Set Observations')
```



Identify Training Indices in k-Fold Partition

Identify the observations that are in the training sets of a `cvpartition` object for 3-fold cross-validation.

Partition 10 observations for 3-fold cross-validation. Notice that `c` contains three repetitions of training and test data.

```
rng('default') % For reproducibility
c = cvpartition(10,'KFold',3)

c =
K-fold cross validation partition
   NumObservations: 10
       NumTestSets: 3
         TrainSize: 7  6  7
          TestSize: 3  4  3
```

Identify the training set observations for each repetition of training and test data. Observations that correspond to 1s are in the corresponding training set.

```
set1 = training(c,1)

set1 = 10x1 logical array

   0
   0
   1
   1
   1
   1
   1
   1
   0
   1


set2 = training(c,2);
set3 = training(c,3);
```

Visualize the results. All observations except the first, second, and ninth are in the first training set. All observations except the third, sixth, eighth, and tenth are in the second training set. All observations except the fourth, fifth, and seventh are in the third training set.

```
data = [set1,set2,set3];
h = heatmap(double(data),'ColorbarVisible','off');
sorty(h,{'1','2','3'},'ascend')
xlabel('Repetition')
ylabel('Observation')
title('Training Set Observations')
```

Input Arguments

`c` — Validation partition

`cvpartition` object

Validation partition, specified as a `cvpartition` object. The validation partition type of `c`, `c.Type`, is `'kfold'`, `'holdout'`, `'leaveout'`, or `'resubstitution'`.

`i` — Repetition index

positive integer scalar

Repetition index, specified as a positive integer scalar. Specifying `i` indicates to find the observations in the `i`th training set.

Data Types: `single` | `double`

Output Arguments

`idx` — Indices for training set observations

logical vector

Indices for training set observations, returned as a logical vector. A value of 1 indicates that the corresponding observation is in the training set. A value of 0 indicates that the corresponding observation is in the test set.

See Also

`cvpartition` | `test`

Introduced in R2008a

transform

Transform predictors into extracted features

Syntax

```
z = transform(Mdl,x)
```

Description

`z = transform(Mdl,x)` transforms the data `x` into the features `z` via the model `Mdl`.

Examples

Transform Data to Learned Features

Create a feature transformation model with 100 features from the `SampleImagePatches` data.

```
rng('default') % For reproducibility
data = load('SampleImagePatches');
q = 100;
X = data.X;
Mdl = sparsefilt(X,q)
```

```
Warning: Solver LBFGS was not able to converge to a solution.
```

```
Mdl =
  SparseFiltering
           ModelParameters: [1x1 struct]
             NumPredictors: 363
        NumLearnedFeatures: 100
                        Mu: []
                     Sigma: []
                   FitInfo: [1x1 struct]
          TransformWeights: [363x100 double]
    InitialTransformWeights: []


  Properties, Methods
```

`sparsefilt` issues a warning because it stopped due to reaching the iteration limit, instead of reaching a step-size limit or a gradient-size limit. You can still use the learned features in the returned object by calling the `transform` function.

Transform the first five rows of the input data `X` to the new feature space.

```
y = transform(Mdl,X(1:5,:));
size(y)
```

```
ans = 1×2
```

```
     5   100
```

Input Arguments

`Mdl` — Feature extraction model

`SparseFiltering` object | `ReconstructionICA` object

Feature extraction model, specified as a `SparseFiltering` object or as a `ReconstructionICA` object. Create `Mdl` by using the `sparsefilt` function or the `rica` function.

`x` — Predictor data

matrix with `p` columns | table of numeric values with `p` columns

Predictor data, specified as a matrix with `p` columns or as a table of numeric values with `p` columns. Here, `p` is the number of predictors in the model, which is `Mdl.NumPredictors`. Each row of the input matrix or table represents one data point to transform.

Data Types: `single` | `double` | `table`

Output Arguments

`z` — Transformed data

`n`-by-`q` matrix

Transformed data, returned as an `n`-by-`q` matrix. Here, `n` is the number of rows in the input data `x`, and `q` is the number of features, which is `Mdl.NumLearnedFeatures`.

Algorithms

`transform` converts data to predicted features by using the learned weight matrix `W` to map input predictors to output features.

- For `rica`, input data `X` maps linearly to output features `XW`. See “Reconstruction ICA Algorithm” on page 15-132.
- For `sparsefilt`, input data maps nonlinearly to output features $\widehat{F}$(`X`,`W`). See “Sparse Filtering Algorithm” on page 15-130.

Caution The result of `transform` for sparse filtering depends on the number of data points. In particular, the result of applying `transform` to each row of a matrix separately differs from the result of applying `transform` to the entire matrix at once.

See Also

`ReconstructionICA` | `SparseFiltering` | `rica` | `sparsefilt`

Topics

“Feature Extraction Workflow” on page 15-135
“Feature Extraction” on page 15-130

Introduced in R2017a

transform

Transform new data using generated features

Syntax

```
NewTbl = transform(Transformer,Tbl)
NewTbl = transform(Transformer,Tbl,Index)
```

Description

`NewTbl = transform(Transformer,Tbl)` returns a table with transformed features generated by the `FeatureTransformer` object `Transformer`. The input `Tbl` must contain the required variables, whose data types must match those of the variables originally passed to `gencfeatures` when `Transformer` was created.

`NewTbl = transform(Transformer,Tbl,Index)` returns a subset of the transformed features, where `Index` indicates the features to return.

Examples

Compute Cross-Validation Loss Using Generated Features

Generate features to train a linear classifier. Compute the cross-validation classification error of the model by using the `crossval` function.

Load the `ionosphere` data set, and create a table containing the predictor data.

```
load ionosphere
Tbl = array2table(X);
```

Create a random partition for stratified 5-fold cross-validation.

```
rng("default") % For reproducibility of the partition
cvp = cvpartition(Y,"KFold",5);
```

Compute the cross-validation classification loss for a linear model trained on the original features in `Tbl`.

```
CVMdl = fitclinear(Tbl,Y,"CVPartition",cvp);
cvloss = kfoldLoss(CVMdl)
```

```
cvloss = 0.1339
```

Create the custom function `myloss` (shown at the end of this example). This function generates 20 features from the training data, and then applies the same training set transformations to the test data. The function then fits a linear classifier to the training data and computes the test set loss.

Note: If you use the live script file for this example, the `myloss` function is already included at the end of the file. Otherwise, you need to create this function at the end of your .m file or add it as a file on the MATLAB® path.

Compute the cross-validation classification loss for a linear model trained on features generated from the predictors in `Tbl`.

```
newcvloss = mean(crossval(@myloss,Tbl,Y,"Partition",cvp))
```

```
newcvloss = 0.0770
```

```
function testloss = myloss(TrainTbl,trainY,TestTbl,testY)
[Transformer,NewTrainTbl] = gencfeatures(TrainTbl,trainY,20);
NewTestTbl = transform(Transformer,TestTbl);
Mdl = fitclinear(NewTrainTbl,trainY);
testloss = loss(Mdl,NewTestTbl,testY, ...
    "LossFun","classiferror");
end
```

Train Model Using Subset of Generated Features

Train a linear classifier using only the numeric generated features returned by `gencfeatures`.

Load the `patients` data set. Create a table from a subset of the variables.

```
load patients
Tbl = table(Age,Diastolic,Height,SelfAssessedHealthStatus, ...
    Smoker,Systolic,Weight,Gender);
```

Partition the data into training and test sets. Use approximately 70% of the observations as training data, and 30% of the observations as test data. Partition the data using `cvpartition`.

```
rng("default")
c = cvpartition(Tbl.Gender,"Holdout",0.30);
TrainTbl = Tbl(training(c),:);
TestTbl = Tbl(test(c),:);
```

Use the training data to generate 25 new features. Specify the minimum redundancy maximum relevance (MRMR) feature selection method for selecting new features.

```
Transformer = gencfeatures(TrainTbl,"Gender",25, ...
    "FeatureSelectionMethod","mrmr")
```

```
Transformer =
  FeatureTransformer with properties:

                    Type: 'classification'
           TargetLearner: 'linear'
    NumEngineeredFeatures: 24
      NumOriginalFeatures: 1
         TotalNumFeatures: 25
```

Inspect the generated features.

```
Info = describe(Transformer)
```

```
Info=25×4 table
                                        Type        IsOriginal         InputVariables
```

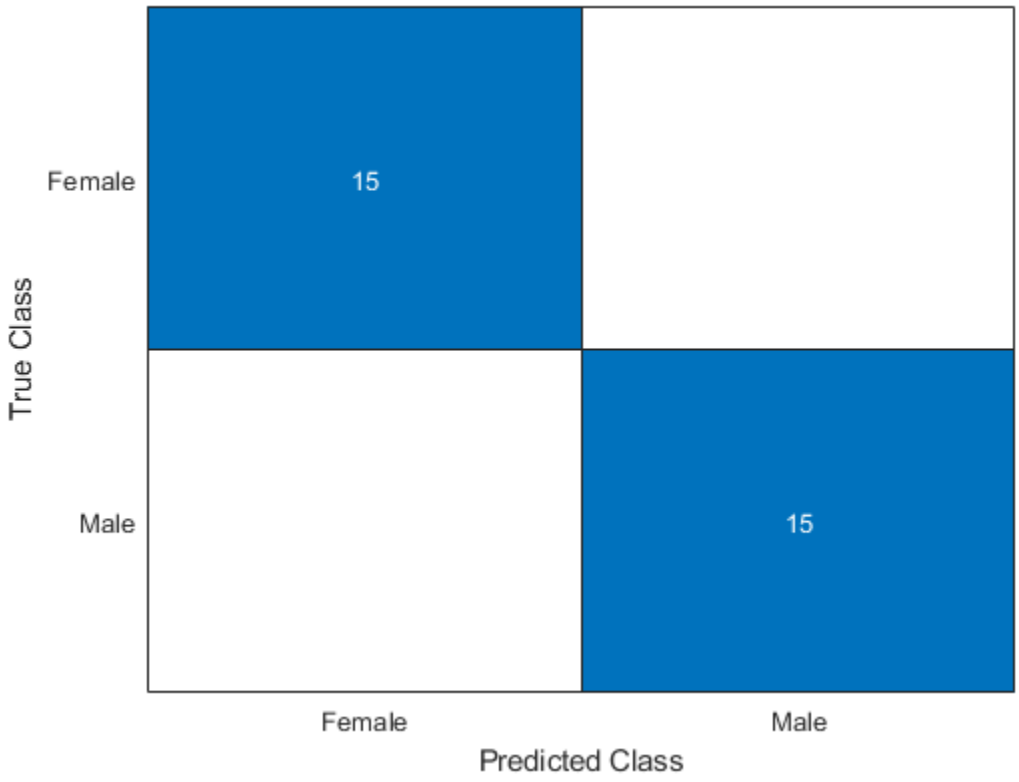
```
c(SelfAssessedHealthStatus)    Categorical      true        SelfAssessedHealthStatus    "Vari
eb5(Weight)                    Categorical      false       Weight                      "Equa
zsc(sqrt(Systolic))            Numeric          false       Systolic                    "sqrt
zsc(sin(Systolic))             Numeric          false       Systolic                    "sin(
zsc(Systolic./Weight)          Numeric          false       Systolic, Weight            "Syst
zsc(Age+Weight)                Numeric          false       Age, Weight                 "Age
zsc(Age./Weight)               Numeric          false       Age, Weight                 "Age
zsc(Diastolic.*Weight)         Numeric          false       Diastolic, Weight           "Dias
q6(Height)                     Categorical      false       Height                      "Equi
zsc(Systolic+Weight)           Numeric          false       Systolic, Weight            "Syst
zsc(Diastolic-Weight)          Numeric          false       Diastolic, Weight           "Dias
zsc(Age-Weight)                Numeric          false       Age, Weight                 "Age
zsc(Height./Weight)            Numeric          false       Height, Weight              "Heig
zsc(Height.*Weight)            Numeric          false       Height, Weight              "Heig
zsc(Diastolic+Weight)          Numeric          false       Diastolic, Weight           "Dias
zsc(Age.*Weight)               Numeric          false       Age, Weight                 "Age
  ⋮
```

Transform the training and test sets, but retain only the numeric predictors.

```
numericIdx = (Info.Type == "Numeric");
NewTrainTbl = transform(Transformer,TrainTbl,numericIdx);
NewTestTbl = transform(Transformer,TestTbl,numericIdx);
```

Train a linear model using the transformed training data. Visualize the accuracy of the model's test set predictions by using a confusion matrix.

```
Mdl = fitclinear(NewTrainTbl,TrainTbl.Gender);
testLabels = predict(Mdl,NewTestTbl);
confusionchart(TestTbl.Gender,testLabels)
```



Input Arguments

`Transformer` — Feature transformer
`FeatureTransformer` object

Feature transformer, specified as a `FeatureTransformer` object.

`Tbl` — Features to transform
table

Features to transform, specified as a table. The rows must correspond to observations, and the columns must correspond to the predictors used to generate the transformed features stored in `Transformer`. You can enter `describe(Transformer).InputVariables` to see the list of features that `Tbl` must contain.

Data Types: `table`

`Index` — Features to return
numeric vector | logical vector | string array | cell array of character vectors

Features to return, specified as a numeric or logical vector indicating the position of the features, or a string array or cell array of character vectors indicating the names of the features.

Example: `1:12`

Data Types: `single` | `double` | `logical` | `string` | `cell`

Output Arguments

`NewTbl` — Transformed features

table

Transformed features, returned as a table. Each row corresponds to an observation, and each column corresponds to a generated feature.

See Also

`FeatureTransformer` | `describe` | `gencfeatures`

Topics

“Automated Feature Engineering for Classification” on page 18-190

Introduced in R2021a

TreeBagger class

Bag of decision trees

Description

`TreeBagger` bags an ensemble of decision trees for either classification or regression. Bagging stands for bootstrap aggregation. Every tree in the ensemble is grown on an independently drawn bootstrap replica of input data. Observations not included in this replica are "out of bag" for this tree.

`TreeBagger` relies on the `ClassificationTree` and `RegressionTree` functionality for growing individual trees. In particular, `ClassificationTree` and `RegressionTree` accepts the number of features selected at random for each decision split as an optional input argument. That is, `TreeBagger` implements the random forest algorithm [1].

For regression problems, `TreeBagger` supports mean and quantile regression (that is, quantile regression forest [2]).

- To predict mean responses or estimate the mean-squared error given data, pass a `TreeBagger` model and the data to `predict` or `error`, respectively. To perform similar operations for out-of-bag observations, use `oobPredict` or `oobError`.
- To estimate quantiles of the response distribution or the quantile error given data, pass a `TreeBagger` model and the data to `quantilePredict` or `quantileError`, respectively. To perform similar operations for out-of-bag observations, use `oobQuantilePredict` or `oobQuantileError`.

Construction

TreeBagger	Create bag of decision trees

Object Functions

append	Append new trees to ensemble
compact	Compact ensemble of decision trees
error	Error (misclassification probability or MSE)
fillprox	Proximity matrix for training data
growTrees	Train additional trees and add to ensemble
margin	Classification margin
mdsprox	Multidimensional scaling of proximity matrix
meanMargin	Mean classification margin
oobError	Out-of-bag error
oobMargin	Out-of-bag margins
oobMeanMargin	Out-of-bag mean margins
oobPredict	Ensemble predictions for out-of-bag observations
oobQuantileError	Out-of-bag quantile loss of bag of regression trees
oobQuantilePredict	Quantile predictions for out-of-bag observations from bag of regression trees
partialDependence	Compute partial dependence

plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses using ensemble of bagged decision trees
quantileError	Quantile loss using bag of regression trees
quantilePredict	Predict response quantile using bag of regression trees

Properties

`ClassNames`

A cell array containing the class names for the response variable `Y`. This property is empty for regression trees.

`ComputeOOBPrediction`

A logical flag specifying whether out-of-bag predictions for training observations should be computed. The default is `false`.

If this flag is `true`, the following properties are available:

- `OOBIndices`
- `OOBInstanceWeight`

If this flag is `true`, the following methods can be called:

- `oobError`
- `oobMargin`
- `oobMeanMargin`

`ComputeOOBPredictorImportance`

A logical flag specifying whether out-of-bag estimates of variable importance should be computed. The default is `false`. If this flag is `true`, then `ComputeOOBPrediction` is true as well.

If this flag is `true`, the following properties are available:

- `OOBPermutedPredictorDeltaError`
- `OOBPermutedPredictorDeltaMeanMargin`
- `OOBPermutedPredictorCountRaiseMargin`

`Cost`

Square matrix, where `Cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`. The number of rows and columns in `Cost` is the number of unique classes in the response.

This property is:

- read-only
- empty (`[]`) for ensembles of regression trees

DefaultYfit

Default value returned by `predict` and `oobPredict`. The `DefaultYfit` property controls what predicted value is returned when no prediction is possible. For example, when `oobPredict` needs to predict for an observation that is in-bag for all trees in the ensemble.

- For classification, you can set this property to either `''` or `'MostPopular'`. If you choose `'MostPopular'` (the default), the property value becomes the name of the most probably class in the training data. If you choose `''`, the in-bag observations are excluded from computation of the out-of-bag error and margin.
- For regression, you can set this property to any numeric scalar. The default value is the mean of the response for the training data. If you set this property to `NaN`, the in-bag observations are excluded from computation of the out-of-bag error and margin.

DeltaCriterionDecisionSplit

A numeric array of size 1-by-*Nvars* of changes in the split criterion summed over splits on each variable, averaged across the entire ensemble of grown trees.

InBagFraction

Fraction of observations that are randomly selected with replacement for each bootstrap replica. The size of each replica is *Nobs*×`InBagFraction`, where *Nobs* is the number of observations in the training set. The default value is 1.

MergeLeaves

A logical flag specifying whether decision tree leaves with the same parent are merged for splits that do not decrease the total risk. The default value is `false`.

Method

Method used by trees. The possible values are `'classification'` for classification ensembles, and `'regression'` for regression ensembles.

MinLeafSize

Minimum number of observations per tree leaf. By default, `MinLeafSize` is 1 for classification and 5 for regression. For decision tree training, the `MinParent` value is set equal to `2*MinLeafSize`.

NumTrees

Scalar value equal to the number of decision trees in the ensemble.

NumPredictorSplit

A numeric array of size 1-by-*Nvars*, where every element gives a number of splits on this predictor summed over all trees.

NumPredictorsToSample

Number of predictor or feature variables to select at random for each decision split. By default, `NumPredictorsToSample` is equal to the square root of the total number of variables for classification, and one third of the total number of variables for regression.

OOBIndices

Logical array of size *Nobs*-by-*NumTrees*, where *Nobs* is the number of observations in the training data and *NumTrees* is the number of trees in the ensemble. A `true` value for the (*i*,*j*) element indicates that observation *i* is out-of-bag for tree *j*. In other words, observation *i* was not selected for the training data used to grow tree *j*.

OOBInstanceWeight

Numeric array of size *Nobs*-by-1 containing the number of trees used for computing the out-of-bag response for each observation. *Nobs* is the number of observations in the training data used to create the ensemble.

OOBPermutedPredictorCountRaiseMargin

A numeric array of size 1-by-*Nvars* containing a measure of variable importance for each predictor variable (feature). For any variable, the measure is the difference between the number of raised margins and the number of lowered margins if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble. This property is empty for regression trees.

OOBPermutedPredictorDeltaError

A numeric array of size 1-by-*Nvars* containing a measure of importance for each predictor variable (feature). For any variable, the measure is the increase in prediction error if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble.

OOBPermutedPredictorDeltaMeanMargin

A numeric array of size 1-by-*Nvars* containing a measure of importance for each predictor variable (feature). For any variable, the measure is the decrease in the classification margin if the values of that variable are permuted across the out-of-bag observations. This measure is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble. This property is empty for regression trees.

OutlierMeasure

A numeric array of size *Nobs*-by-1, where *Nobs* is the number of observations in the training data, containing outlier measures for each observation.

Prior

Numeric vector of prior probabilities for each class. The order of the elements of `Prior` corresponds to the order of the classes in `ClassNames`.

This property is:

- read-only
- empty (`[]`) for ensembles of regression trees

Proximity

A numeric matrix of size *Nobs*-by-*Nobs*, where *Nobs* is the number of observations in the training data, containing measures of the proximity between observations. For any two observations, their proximity is defined as the fraction of trees for which these observations land on the same leaf. This is a symmetric matrix with 1s on the diagonal and off-diagonal elements ranging from 0 to 1.

Prune

The `Prune` property is true if decision trees are pruned and false if they are not. Pruning decision trees is not recommended for ensembles. The default value is false.

SampleWithReplacement

A logical flag specifying if data are sampled for each decision tree with replacement. This property is `true` if `TreeBagger` samples data with replacement and `false` otherwise. Default value is `true`.

TreeArguments

Cell array of arguments for `fitctree` or `fitrtree`. These arguments are used by `TreeBagger` when growing new trees for the ensemble.

Trees

A cell array of size *NumTrees*-by-1 containing the trees in the ensemble.

SurrogateAssociation

A matrix of size *Nvars*-by-*Nvars* with predictive measures of variable association, averaged across the entire ensemble of grown trees. If you grew the ensemble setting `'surrogate'` to `'on'`, this matrix for each tree is filled with predictive measures of association averaged over the surrogate splits. If you grew the ensemble setting `'surrogate'` to `'off'` (default), `SurrogateAssociation` is diagonal.

PredictorNames

A cell array containing the names of the predictor variables (features). `TreeBagger` takes these names from the optional `'names'` parameter. The default names are `'x1'`, `'x2'`, etc.

W

Numeric vector of weights of length *Nobs*, where *Nobs* is the number of observations (rows) in the training data. `TreeBagger` uses these weights for growing every decision tree in the ensemble. The default `W` is `ones(Nobs,1)`.

X

A table or numeric matrix of size *Nobs*-by-*Nvars*, where *Nobs* is the number of observations (rows) and *Nvars* is the number of variables (columns) in the training data. If you train the ensemble using a table of predictor values, then `X` is a table. If you train the ensemble using a matrix of predictor values, then `X` is a matrix. This property contains the predictor (or feature) values.

Y

A size *Nobs* array of response data. Elements of `Y` correspond to the rows of `X`. For classification, `Y` is the set of true class labels. Labels can be any grouping variable on page 2-45, that is, a numeric or

logical vector, character matrix, string array, cell array of character vectors, or categorical vector. `TreeBagger` converts labels to a cell array of character vectors for classification. For regression, `Y` is a numeric vector.

Examples

Train Ensemble of Bagged Classification Trees

Load Fisher's iris data set.

```
load fisheriris
```

Train an ensemble of bagged classification trees using the entire data set. Specify `50` weak learners. Store which observations are out of bag for each tree.

```
rng(1); % For reproducibility
Mdl = TreeBagger(50,meas,species,'OOBPrediction','On',...
    'Method','classification')
```

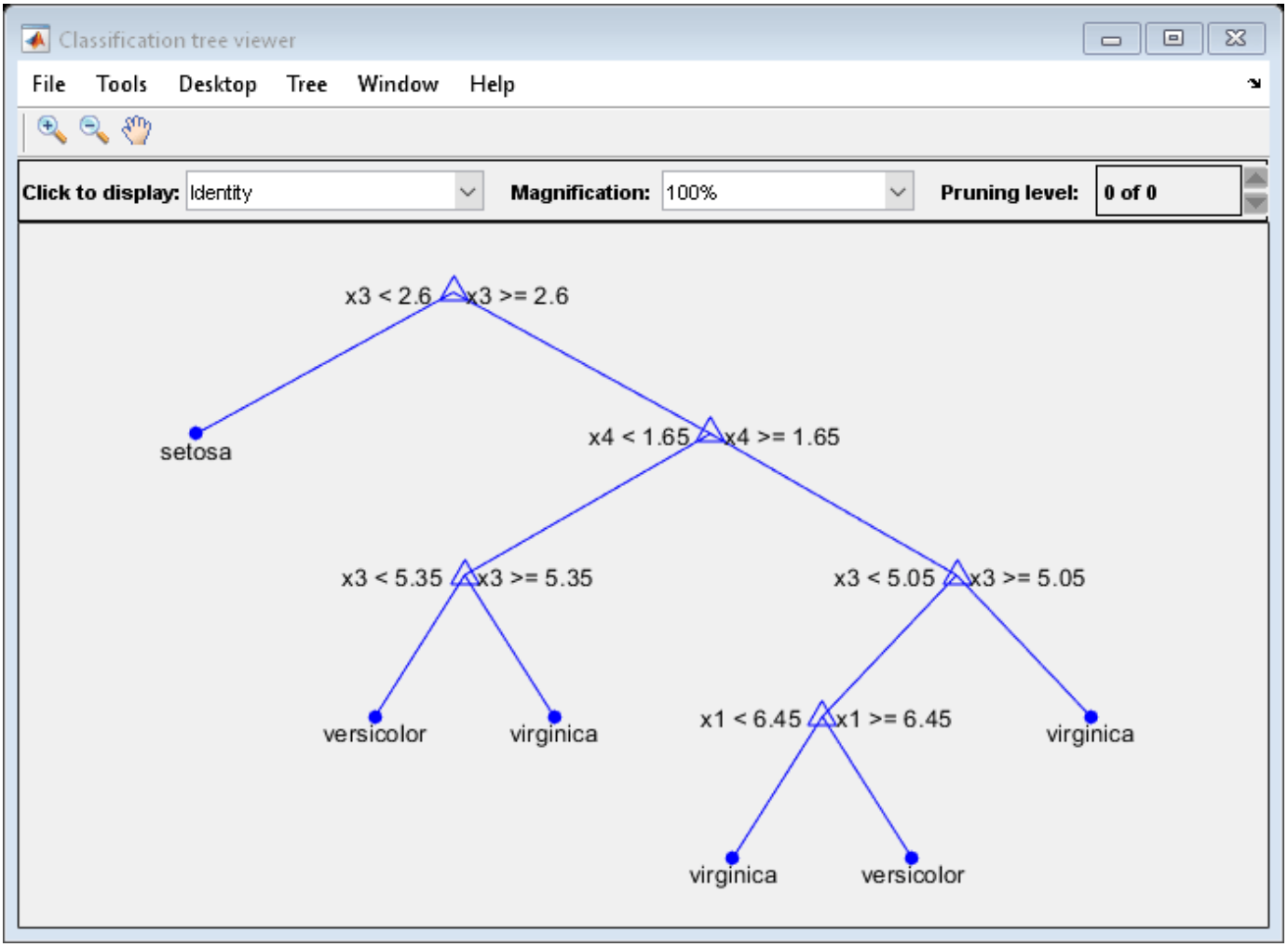
```
Mdl =
  TreeBagger
Ensemble with 50 bagged decision trees:
                    Training X:              [150x4]
                    Training Y:              [150x1]
                        Method:       classification
                NumPredictors:                    4
       NumPredictorsToSample:                    2
                   MinLeafSize:                   1
                InBagFraction:                    1
       SampleWithReplacement:                     1
        ComputeOOBPrediction:                     1
 ComputeOOBPredictorImportance:                   0
                     Proximity:                  []
                    ClassNames:        'setosa'    'versicolor'     'virginica'

  Properties, Methods
```

`Mdl` is a `TreeBagger` ensemble.

`Mdl.Trees` stores a 50-by-1 cell vector of the trained classification trees (`CompactClassificationTree` model objects) that compose the ensemble.

Plot a graph of the first trained classification tree.

```
view(Mdl.Trees{1},'Mode','graph')
```

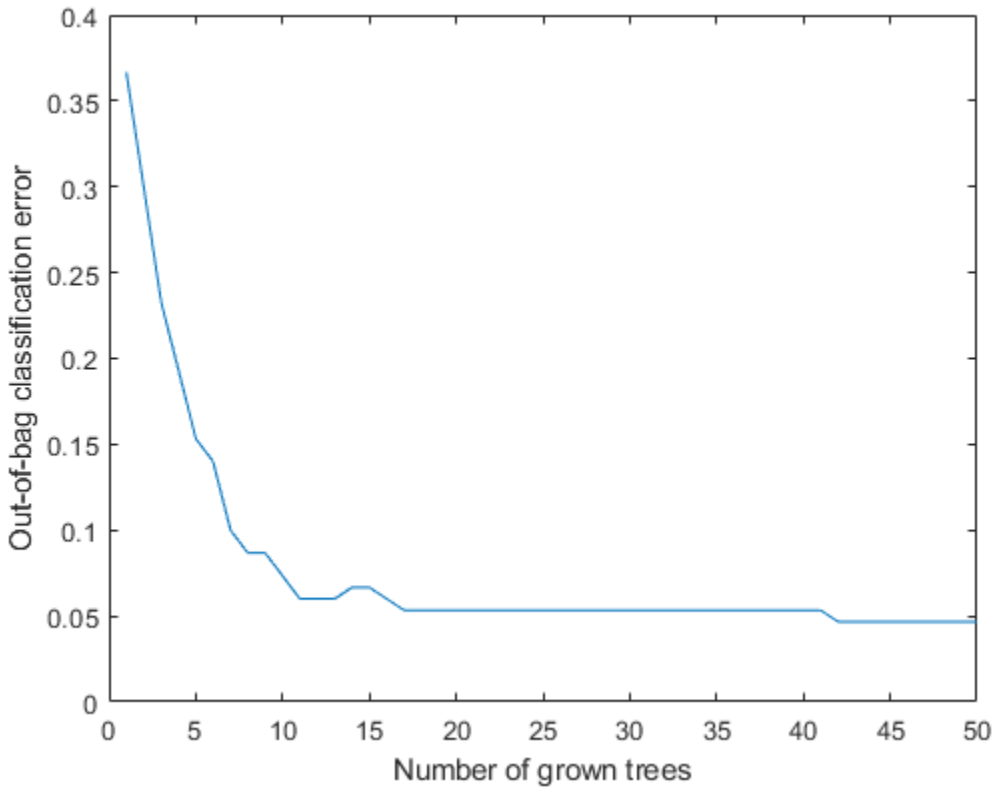


By default, `TreeBagger` grows deep trees.

`Mdl.OOBIndices` stores the out-of-bag indices as a matrix of logical values.

Plot the out-of-bag error over the number of grown classification trees.

```
figure;
oobErrorBaggedEnsemble = oobError(Mdl);
plot(oobErrorBaggedEnsemble)
xlabel 'Number of grown trees';
ylabel 'Out-of-bag classification error';
```



The out-of-bag error decreases with the number of grown trees.

To label out-of-bag observations, pass `Mdl` to `oobPredict`.

Train Ensemble of Bagged Regression Trees

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression');
```

`Mdl` is a `TreeBagger` ensemble.

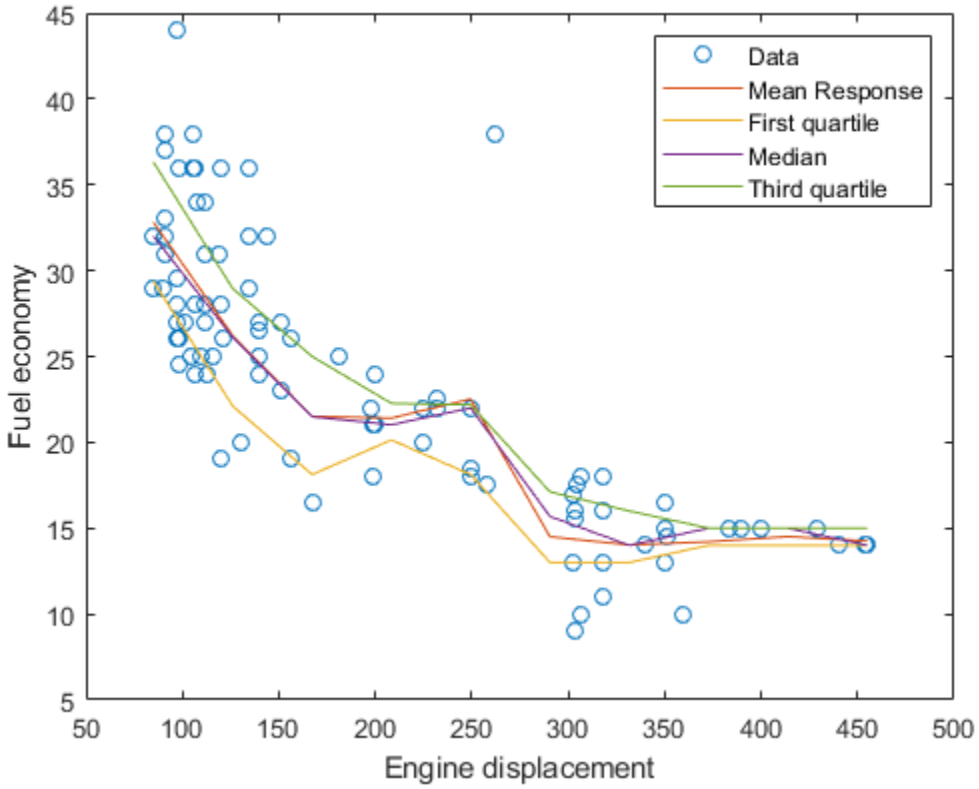
Using a trained bag of regression trees, you can estimate conditional mean responses or perform quantile regression to predict conditional quantiles.

For ten equally-spaced engine displacements between the minimum and maximum in-sample displacement, predict conditional mean responses and conditional quartiles.

```
predX = linspace(min(Displacement),max(Displacement),10)';
mpgMean = predict(Mdl,predX);
mpgQuartiles = quantilePredict(Mdl,predX,'Quantile',[0.25,0.5,0.75]);
```

Plot the observations, and estimated mean responses and quartiles in the same figure.

```
figure;
plot(Displacement,MPG,'o');
hold on
plot(predX,mpgMean);
plot(predX,mpgQuartiles);
ylabel('Fuel economy');
xlabel('Engine displacement');
legend('Data','Mean Response','First quartile','Median','Third quartile');
```



Unbiased Predictor Importance Estimates

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
```

```
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
rng('default'); % For reproducibility
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

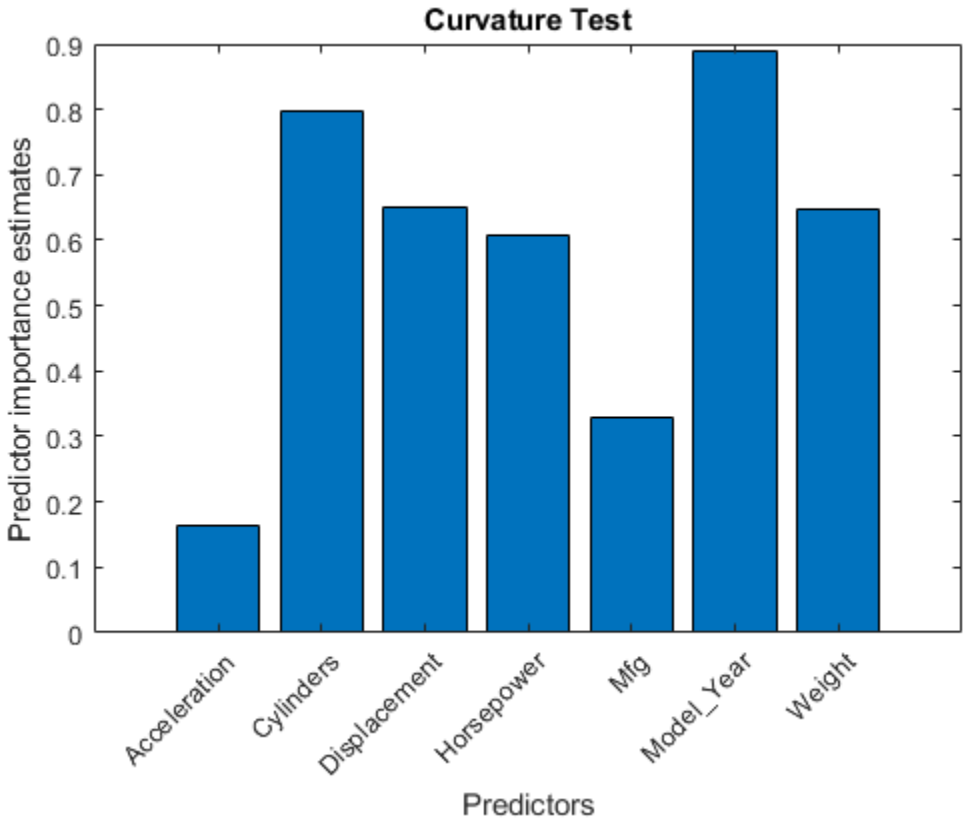
Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a random forest of 200 regression trees using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits. Store the out-of-bag information for predictor importance estimation.

```
Mdl = TreeBagger(200,X,'MPG','Method','regression','Surrogate','on',...
    'PredictorSelection','curvature','OOBPredictorImportance','on');
```

`TreeBagger` stores predictor importance estimates in the property `OOBPermutedPredictorDeltaError`. Compare the estimates using a bar graph.

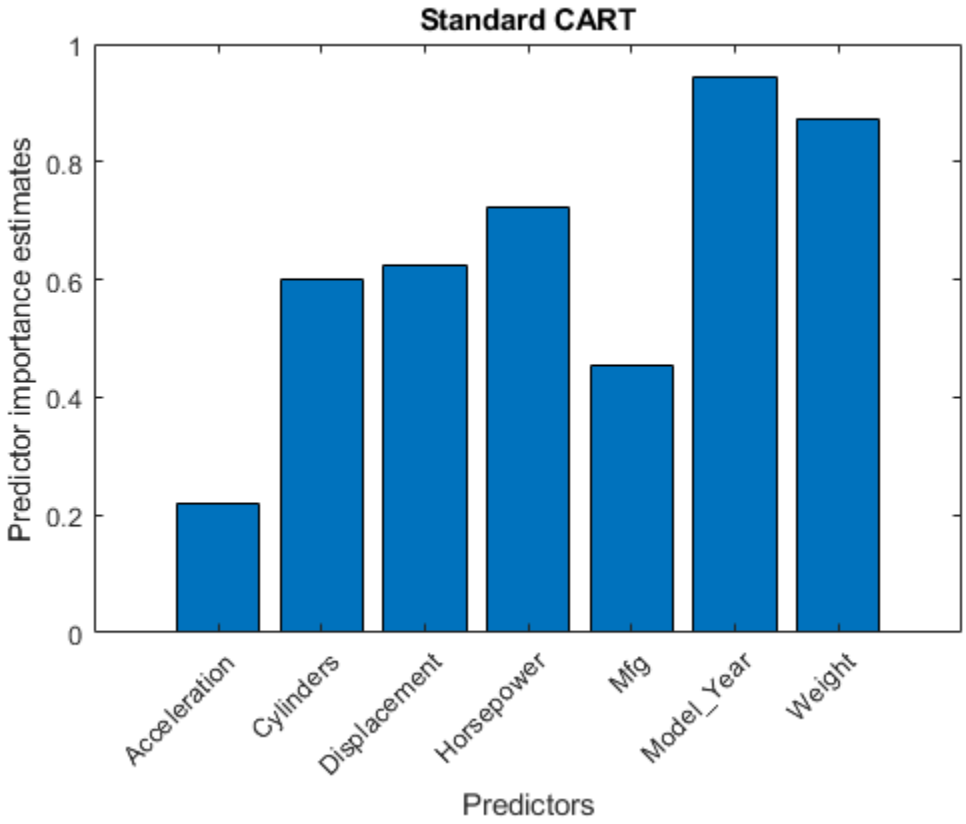
```
imp = Mdl.OOBPermutedPredictorDeltaError;

figure;
bar(imp);
title('Curvature Test');
ylabel('Predictor importance estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Model_Year` is the most important predictor, followed by `Weight`.

Compare the `imp` to predictor importance estimates computed from a random forest that grows trees using standard CART.

```
MdlCART = TreeBagger(200,X,'MPG','Method','regression','Surrogate','on',...
    'OOBPredictorImportance','on');

impCART = MdlCART.OOBPermutedPredictorDeltaError;

figure;
bar(impCART);
title('Standard CART');
ylabel('Predictor importance estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Weight`, a continuous predictor, is the most important. The next two most importance predictor are `Model_Year` followed closely by `Horsepower`, which is a continuous predictor.

Copy Semantics

Value. To learn how this affects your use of the class, see Comparing Handle and Value Classes in the MATLAB Object-Oriented Programming documentation.

Tip

For a `TreeBagger` model object `B`, the `Trees` property stores a cell vector of `B.NumTrees` `CompactClassificationTree` or `CompactRegressionTree` model objects. For a textual or graphical display of tree `t` in the cell vector, enter

```
view(B.Trees{t})
```

Alternative Functionality

Statistics and Machine Learning Toolbox offers three objects for bagging and random forest:

- `ClassificationBaggedEnsemble` created by `fitcensemble` for classification
- `RegressionBaggedEnsemble` created by `fitrensemble` for regression

- `TreeBagger` created by `TreeBagger` for classification and regression

For details about the differences between `TreeBagger` and bagged ensembles (`ClassificationBaggedEnsemble` and `RegressionBaggedEnsemble`), see “Comparison of TreeBagger and Bagged Ensembles” on page 18-44.

References

[1] Breiman, L. "Random Forests." *Machine Learning* 45, pp. 5–32, 2001.

[2] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

See Also

`CompactTreeBagger` | `TreeBagger` | `compact` | `error` | `oobError` | `oobPredict` | `predict` | `view` | `view`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
“Framework for Ensemble Learning” on page 18-31
“Decision Trees” on page 19-2
“Grouping Variables” on page 2-45

TreeBagger

Class: `TreeBagger`

Create bag of decision trees

Individual decision trees tend to overfit. Bootstrap-aggregated (bagged) decision trees combine the results of many decision trees, which reduces the effects of overfitting and improves generalization. `TreeBagger` grows the decision trees in the ensemble using bootstrap samples of the data. Also, `TreeBagger` selects a random subset of predictors to use at each decision split as in the random forest algorithm [1].

By default, `TreeBagger` bags classification trees. To bag regression trees instead, specify `'Method','regression'`.

For regression problems, `TreeBagger` supports mean and quantile regression (that is, quantile regression forest [5]).

Syntax

```
Mdl = TreeBagger(NumTrees,Tbl,ResponseVarName)
Mdl = TreeBagger(NumTrees,Tbl,formula)
Mdl = TreeBagger(NumTrees,Tbl,Y)
B = TreeBagger(NumTrees,X,Y)
B = TreeBagger(NumTrees,X,Y,Name,Value)
```

Description

`Mdl = TreeBagger(NumTrees,Tbl,ResponseVarName)` returns an ensemble of `NumTrees` bagged classification trees trained using the sample data in the table `Tbl`. `ResponseVarName` is the name of the response variable in `Tbl`.

`Mdl = TreeBagger(NumTrees,Tbl,formula)` returns an ensemble of bagged classification trees trained using the sample data in the table `Tbl`. `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`. Specify `Formula` using Wilkinson notation. For more information, see “Wilkinson Notation” on page 11-91.

`Mdl = TreeBagger(NumTrees,Tbl,Y)` returns an ensemble of classification trees using the predictor variables in table `Tbl` and class labels in vector `Y`.

`Y` is an array of response data. Elements of `Y` correspond to the rows of `Tbl`. For classification, `Y` is the set of true class labels. Labels can be any grouping variable on page 2-45, that is, a numeric or logical vector, character matrix, string array, cell array of character vectors, or categorical vector. `TreeBagger` converts labels to a cell array of character vectors. For regression, `Y` is a numeric vector. To grow regression trees, you must specify the name-value pair `'Method','regression'`.

`B = TreeBagger(NumTrees,X,Y)` creates an ensemble `B` of `NumTrees` decision trees for predicting response `Y` as a function of predictors in the numeric matrix of training data, `X`. Each row in `X` represents an observation and each column represents a predictor or feature.

`B = TreeBagger(NumTrees,X,Y,Name,Value)` specifies optional parameter name-value pairs:

`'InBagFraction'`	Fraction of input data to sample with replacement from the input data for growing each new tree. Default value is 1.
`'Cost'`	Square matrix `C`, where `C(i,j)` is the cost of classifying a point into class `j` if its true class is `i` (i.e., the rows correspond to the true class and the columns correspond to the predicted class). The order of the rows and columns of `Cost` corresponds to the order of the classes in the `ClassNames` property of the trained `TreeBagger` model `B`. Alternatively, `cost` can be a structure `S` having two fields: • `S.ClassNames` containing the group names as a categorical variable, character array, string array, or cell array of character vectors • `S.ClassificationCosts` containing the cost matrix `C` The default value is `C(i,j) = 1` if `i ~= j`, and `C(i,j) = 0` if `i = j`. If `Cost` is highly skewed, then, for in-bag samples, the software oversamples unique observations from the class that has a large penalty. For smaller sample sizes, this might cause a very low relative frequency of out-of-bag observations from the class that has a large penalty. Therefore, the estimated out-of-bag error is highly variable, and might be difficult to interpret.
`'SampleWithReplacement'`	`'on'` to sample with replacement or `'off'` to sample without replacement. If you sample without replacement, you need to set `'InBagFraction'` to a value less than one. Default is `'on'`.
`'OOBPrediction'`	`'on'` to store info on what observations are out of bag for each tree. This info can be used by `oobPrediction` to compute the predicted class probabilities for each tree in the ensemble. Default is `'off'`.
`'OOBPredictorImportance'`	`'on'` to store out-of-bag estimates of feature importance in the ensemble. Default is `'off'`. Specifying `'on'` also sets the `'OOBPrediction'` value to `'on'`. If an analysis of predictor importance is your goal, then also specify `'PredictorSelection','curvature'` or `'PredictorSelection','interaction-curvature'`. For more details, see `fitctree` or `fitrtree`.
`'Method'`	Either `'classification'` or `'regression'`. Regression requires a numeric `Y`.
`'NumPredictorsToSample'`	Number of variables to select at random for each decision split. Default is the square root of the number of variables for classification and one third of the number of variables for regression. Valid values are `'all'` or a positive integer. Setting this argument to any valid value but `'all'` invokes Breiman's random forest algorithm [1].
`'NumPrint'`	Number of training cycles (grown trees) after which `TreeBagger` displays a diagnostic message showing training progress. Default is no diagnostic messages.
`'MinLeafSize'`	Minimum number of observations per tree leaf. Default is 1 for classification and 5 for regression.

`'Options'`	A structure that specifies options that govern the computation when growing the ensemble of decision trees. One option requests that the computation of decision trees on multiple bootstrap replicates uses multiple processors, if the Parallel Computing Toolbox is available. Two options specify the random number streams to use in selecting bootstrap replicates. You can create this argument with a call to `statset`. You can retrieve values of the individual fields with a call to `statget`. Applicable `statset` parameters are:
	• `'UseParallel'` — If `true` and Parallel Computing Toolbox is installed, then the software uses an existing parallel pool for parallel trees, or, depending on parallel preferences, the software opens and uses a new pool if none is currently open. Otherwise, the software computes in serial. Default is `false`, meaning serial computation.
	For dual-core systems and above, `TreeBagger` parallelizes training using Intel Threading Building Blocks (TBB). Therefore, using the `'UseParallel'` option on a single computer may not speed up computation much and may consume more memory than in serial. For details on Intel TBB, see https://software.intel.com/en-us/intel-tbb.
	• `'UseSubstreams'` — If `true` select each bootstrap replicate using a separate Substream of the random number generator (aka Stream). This option is available only with `RandStream` types that support Substreams: `'mlfg6331_64'` or `'mrg32k3a'`. Default is `false`, do not use a different Substream to compute each bootstrap replicate.
	• `Streams` — A `RandStream` object or cell array of such objects. If you do not specify `Streams`, `TreeBagger` uses the default stream or streams. If you choose to specify `Streams`, use a single object except in the case
	  • `UseParallel` is `true`
	  • `UseSubstreams` is `false`
	In that case, use a cell array the same size as the Parallel pool.

`'Prior'`	Prior probabilities for each class. Specify as one of:
	• A character vector or string scalar:
	  • `'Empirical'` determines class probabilities from class frequencies in `Y`. If you pass observation weights, they are used to compute the class probabilities. This is the default.
	  • `'Uniform'` sets all class probabilities equal.
	• A vector (one scalar value for each class). The order of the elements `Prior` corresponds to the order of the classes in the `ClassNames` property of the trained `TreeBagger` model `B`.
	• A structure `S` with two fields:
	  • `S.ClassNames` containing the class names as a categorical variable, character array, string array, or cell array of character vectors
	  • `S.ClassProbs` containing a vector of corresponding probabilities
	If you set values for both `Weights` and `Prior`, the weights are renormalized to add up to the value of the prior probability in the respective class.
	If `Prior` is highly skewed, then, for in-bag samples, the software oversamples unique observations from the class that has a large prior probability. For smaller sample sizes, this might cause a very low relative frequency of out-of-bag observations from the class that has a large prior probability. Therefore, the estimated out-of-bag error is highly variable, and might be difficult to interpret.

`'PredictorNames'`	Predictor variable names, specified as the comma-separated pair consisting of `'PredictorNames'` and a string array or cell array of unique character vectors. The functionality of `'PredictorNames'` depends on the way you supply the training data.
	• If you supply `X` and `Y`, then you can use `'PredictorNames'` to give the predictor variables in `X` names.
	  • The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
	  • By default, `PredictorNames` is `{'x1','x2',...}`.
	• If you supply `Tbl`, then you can use `'PredictorNames'` to choose which predictor variables to use in training. That is, `TreeBagger` uses the predictor variables in `PredictorNames` and the response only in training.
	  • `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
	  • By default, `PredictorNames` contains the names of all predictor variables.
	  • It good practice to specify the predictors for training using one of `'PredictorNames'` or `formula` only.

`'CategoricalPredictors'`	Categorical predictors list, specified as the comma-separated pair consisting of `'CategoricalPredictors'` and one of the following. • A numeric vector with indices from `1` to `p`, where `p` is the number of columns of `X`. • A logical vector of length `p`, where a `true` entry means that the corresponding column of `X` is a categorical variable. • A string array or cell array of character vectors, where each element in the array is the name of a predictor variable. The names must match entries in `PredictorNames` values. • A character matrix, where each row of the matrix is a name of a predictor variable. The names must match entries in `PredictorNames` values. Pad the names with extra blanks so each row of the character matrix has the same length. • `'all'`, meaning all predictors are categorical.
`'ChunkSize'`	Chunk size, specified as the comma-separated pair consisting of `'ChunkSize'` and a positive integer. The chunk size specifies the number of observations in each chunk of data. The default value is `50000`. **Note** This option only applies when using `TreeBagger` on tall arrays. See “Extended Capabilities” on page 33-0 for more information.

In addition to the optional arguments above, `TreeBagger` accepts these optional `fitctree` and `fitrtree` arguments.

Supported `fitctree` arguments	Supported `fitrtree` arguments
`AlgorithmForCategorical`	`MaxNumSplits`
`MaxNumCategories`	`MergeLeaves`
`MaxNumSplits`	`PredictorSelection`
`MergeLeaves`	`Prune`
`PredictorSelection`	`PruneCriterion`
`Prune`	`QuadraticErrorTolerance`
`PruneCriterion`	`SplitCriterion`
`SplitCriterion`	`Surrogate`
`Surrogate`	`Weights`
`'Weights'`	

Examples

Train Ensemble of Bagged Classification Trees

Load Fisher's iris data set.

```
load fisheriris
```

Train an ensemble of bagged classification trees using the entire data set. Specify `50` weak learners. Store which observations are out of bag for each tree.

```
rng(1); % For reproducibility
Mdl = TreeBagger(50,meas,species,'OOBPrediction','On',...
    'Method','classification')
```

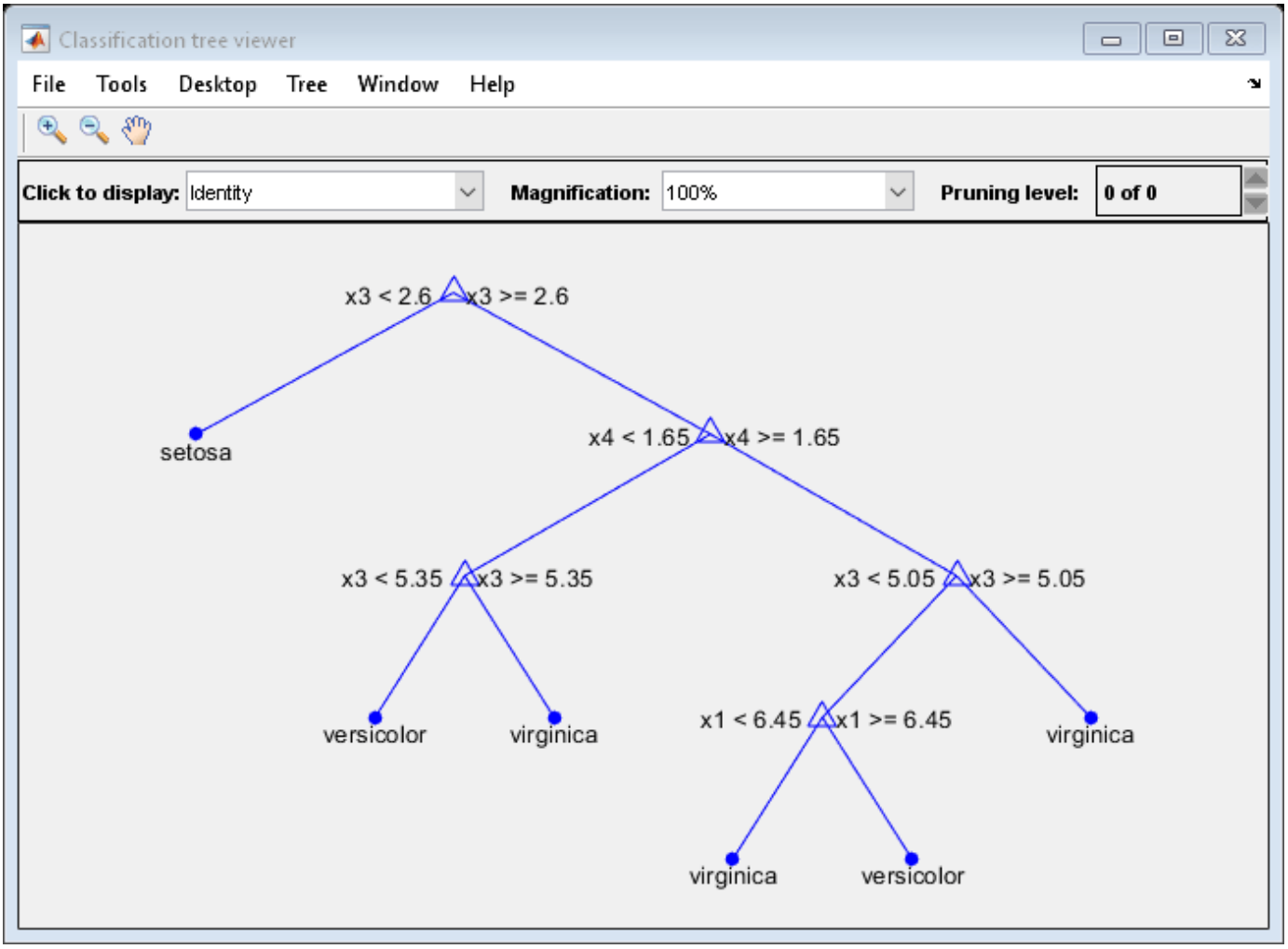
```
Mdl =
  TreeBagger
Ensemble with 50 bagged decision trees:
                    Training X:              [150x4]
                    Training Y:              [150x1]
                        Method:       classification
                NumPredictors:                    4
         NumPredictorsToSample:                    2
                     MinLeafSize:                 1
                InBagFraction:                    1
         SampleWithReplacement:                   1
          ComputeOOBPrediction:                   1
 ComputeOOBPredictorImportance:                   0
                      Proximity:                  []
                    ClassNames:        'setosa'    'versicolor'     'virginica'

  Properties, Methods
```

`Mdl` is a `TreeBagger` ensemble.

`Mdl.Trees` stores a 50-by-1 cell vector of the trained classification trees (`CompactClassificationTree` model objects) that compose the ensemble.

Plot a graph of the first trained classification tree.

```
view(Mdl.Trees{1},'Mode','graph')
```

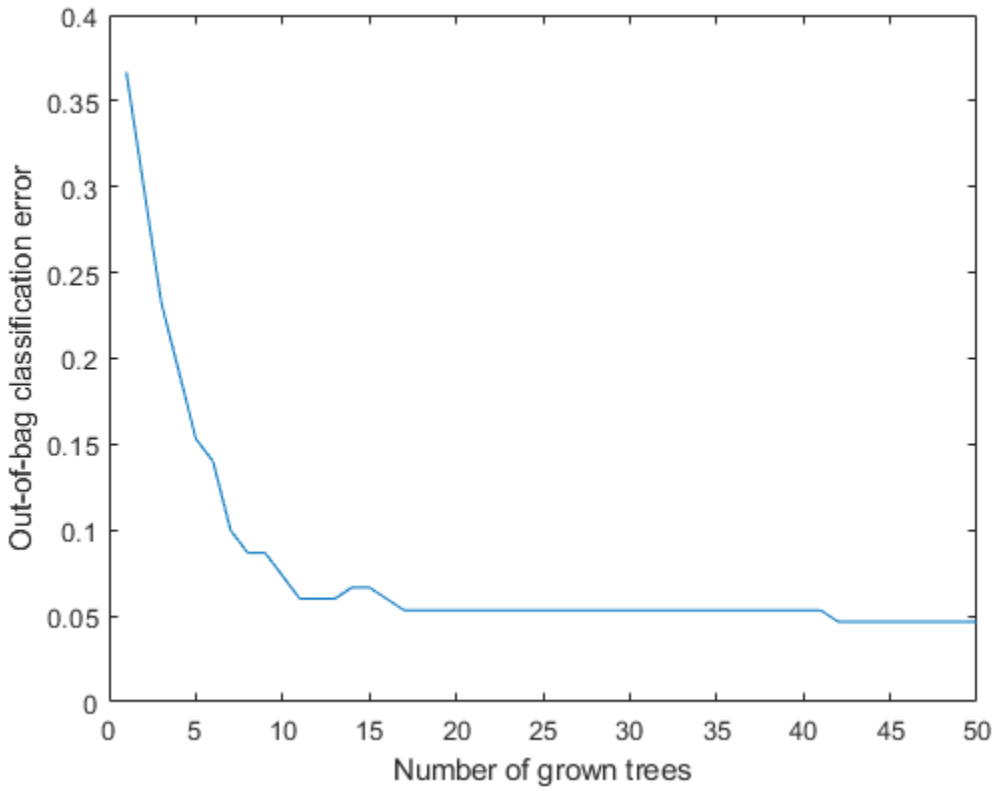


By default, `TreeBagger` grows deep trees.

`Mdl.OOBIndices` stores the out-of-bag indices as a matrix of logical values.

Plot the out-of-bag error over the number of grown classification trees.

```
figure;
oobErrorBaggedEnsemble = oobError(Mdl);
plot(oobErrorBaggedEnsemble)
xlabel 'Number of grown trees';
ylabel 'Out-of-bag classification error';
```



The out-of-bag error decreases with the number of grown trees.

To label out-of-bag observations, pass `Mdl` to `oobPredict`.

Train Ensemble of Bagged Regression Trees

Load the `carsmall` data set. Consider a model that predicts the fuel economy of a car given its engine displacement.

```
load carsmall
```

Train an ensemble of bagged regression trees using the entire data set. Specify 100 weak learners.

```
rng(1); % For reproducibility
Mdl = TreeBagger(100,Displacement,MPG,'Method','regression');
```

`Mdl` is a `TreeBagger` ensemble.

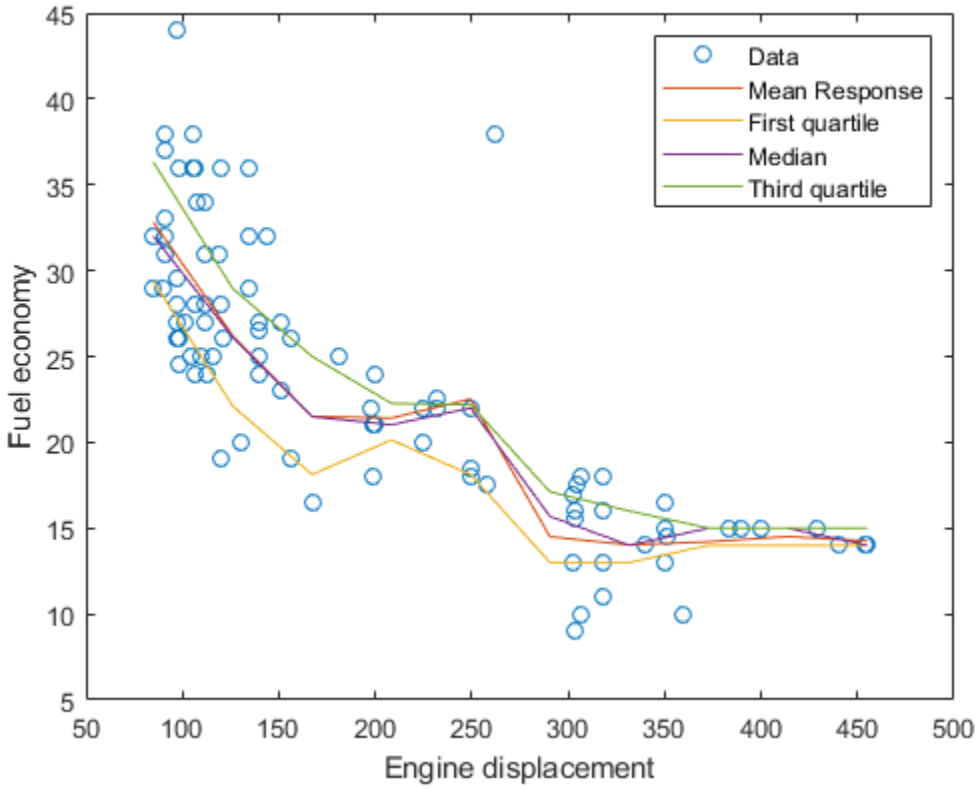
Using a trained bag of regression trees, you can estimate conditional mean responses or perform quantile regression to predict conditional quantiles.

For ten equally-spaced engine displacements between the minimum and maximum in-sample displacement, predict conditional mean responses and conditional quartiles.

```
predX = linspace(min(Displacement),max(Displacement),10)';
mpgMean = predict(Mdl,predX);
mpgQuartiles = quantilePredict(Mdl,predX,'Quantile',[0.25,0.5,0.75]);
```

Plot the observations, and estimated mean responses and quartiles in the same figure.

```
figure;
plot(Displacement,MPG,'o');
hold on
plot(predX,mpgMean);
plot(predX,mpgQuartiles);
ylabel('Fuel economy');
xlabel('Engine displacement');
legend('Data','Mean Response','First quartile','Median','Third quartile');
```



Unbiased Predictor Importance Estimates

Load the `carsmall` data set. Consider a model that predicts the mean fuel economy of a car given its acceleration, number of cylinders, engine displacement, horsepower, manufacturer, model year, and weight. Consider `Cylinders`, `Mfg`, and `Model_Year` as categorical variables.

```
load carsmall
Cylinders = categorical(Cylinders);
Mfg = categorical(cellstr(Mfg));
Model_Year = categorical(Model_Year);
```

```
X = table(Acceleration,Cylinders,Displacement,Horsepower,Mfg,...
    Model_Year,Weight,MPG);
rng('default'); % For reproducibility
```

Display the number of categories represented in the categorical variables.

```
numCylinders = numel(categories(Cylinders))
```

```
numCylinders = 3
```

```
numMfg = numel(categories(Mfg))
```

```
numMfg = 28
```

```
numModelYear = numel(categories(Model_Year))
```

```
numModelYear = 3
```

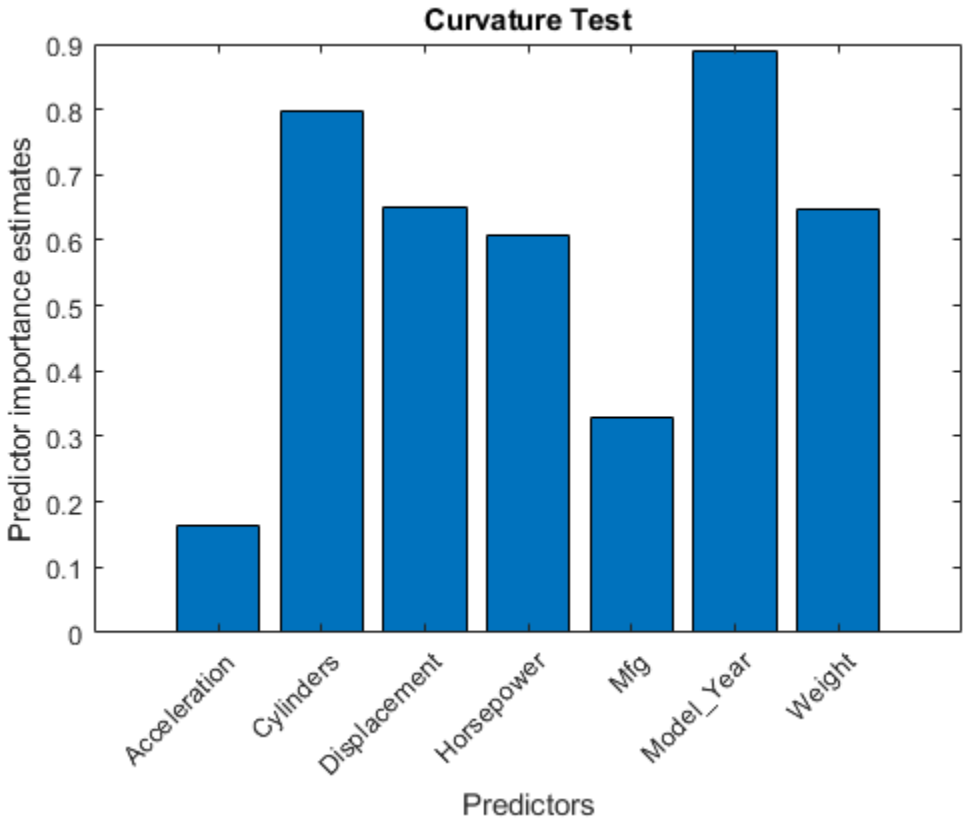
Because there are 3 categories only in `Cylinders` and `Model_Year`, the standard CART, predictor-splitting algorithm prefers splitting a continuous predictor over these two variables.

Train a random forest of 200 regression trees using the entire data set. To grow unbiased trees, specify usage of the curvature test for splitting predictors. Because there are missing values in the data, specify usage of surrogate splits. Store the out-of-bag information for predictor importance estimation.

```
Mdl = TreeBagger(200,X,'MPG','Method','regression','Surrogate','on',...
    'PredictorSelection','curvature','OOBPredictorImportance','on');
```

`TreeBagger` stores predictor importance estimates in the property `OOBPermutedPredictorDeltaError`. Compare the estimates using a bar graph.

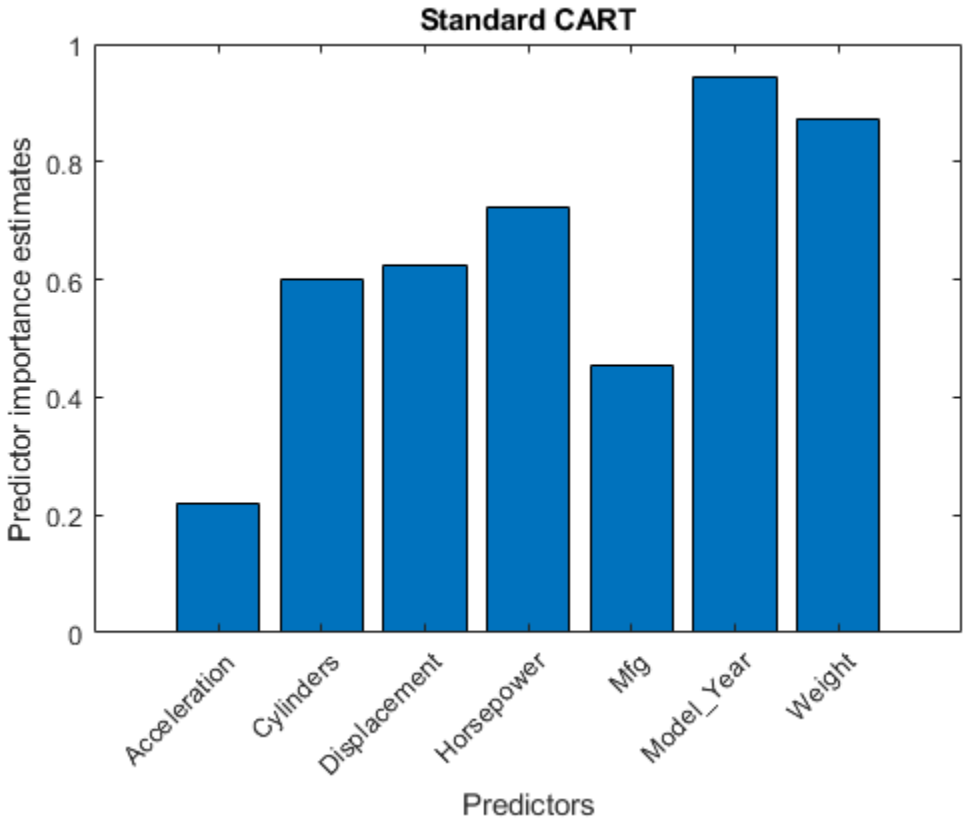
```
imp = Mdl.OOBPermutedPredictorDeltaError;

figure;
bar(imp);
title('Curvature Test');
ylabel('Predictor importance estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Model_Year` is the most important predictor, followed by `Weight`.

Compare the `imp` to predictor importance estimates computed from a random forest that grows trees using standard CART.

```
MdlCART = TreeBagger(200,X,'MPG','Method','regression','Surrogate','on',...
    'OOBPredictorImportance','on');

impCART = MdlCART.OOBPermutedPredictorDeltaError;

figure;
bar(impCART);
title('Standard CART');
ylabel('Predictor importance estimates');
xlabel('Predictors');
h = gca;
h.XTickLabel = Mdl.PredictorNames;
h.XTickLabelRotation = 45;
h.TickLabelInterpreter = 'none';
```



In this case, `Weight`, a continuous predictor, is the most important. The next two most importance predictor are `Model_Year` followed closely by `Horsepower`, which is a continuous predictor.

Train Ensemble of Bagged Classification Trees on Tall Array

Train an ensemble of bagged classification trees for observations in a tall array, and find the misclassification probability of each tree in the model for weighted observations. The sample data set `airlinesmall.csv` is a large data set that contains a tabular file of airline flight data.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore that references the location of the folder containing the data set. Select a subset of the variables to work with, and treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Create a tall table that contains the data in the datastore.

```
ds = datastore('airlinesmall.csv');
ds.SelectedVariableNames = {'Month','DayofMonth','DayOfWeek',...
                            'DepTime','ArrDelay','Distance','DepDelay'};
```

```
ds.TreatAsMissing = 'NA';
tt  = tall(ds) % Tall table
```

```
tt =

  Mx7 tall table

    Month    DayofMonth    DayOfWeek    DepTime    ArrDelay    Distance    DepDelay
    _____    __________    _________    _______    ________    ________    ________

     10          21            3          642          8          308          12
     10          26            1         1021          8          296           1
     10          23            5         2055         21          480          20
     10          23            5         1332         13          296          12
     10          22            4          629          4          373          -1
     10          28            3         1446         59          308          63
     10           8            4          928          3          447          -2
     10          10            6          859         11          954          -1
      :           :            :            :          :            :           :
      :           :            :            :          :            :           :
```

Determine the flights that are late by 10 minutes or more by defining a logical variable that is true for a late flight. This variable contains the class labels. A preview of this variable includes the first few rows.

```
Y = tt.DepDelay > 10 % Class labels
```

```
Y =

  Mx1 tall logical array

   1
   0
   1
   1
   0
   1
   0
   0
   :
   :
```

Create a tall array for the predictor data.

```
X = tt{:,1:end-1} % Predictor data
```

```
X =

  Mx6 tall double matrix

         10          21           3         642           8         308
         10          26           1        1021           8         296
         10          23           5        2055          21         480
         10          23           5        1332          13         296
         10          22           4         629           4         373
         10          28           3        1446          59         308
         10           8           4         928           3         447
         10          10           6         859          11         954
```

```
          :           :           :           :           :           :
          :           :           :           :           :           :
```

Create a tall array for the observation weights by arbitrarily assigning double weights to the observations in class 1.

```
W = Y+1; % Weights
```

Remove rows in `X`, `Y`, and `W` that contain missing data.

```
R = rmmissing([X Y W]); % Data with missing entries removed
X = R(:,1:end-2);
Y = R(:,end-1);
W = R(:,end);
```

Train an ensemble of 20 bagged decision trees using the entire data set. Specify a weight vector and uniform prior probabilities. For reproducibility, set the seeds of the random number generators using `rng` and `tallrng`. The results can vary depending on the number of workers and the execution environment for the tall arrays. For details, see “Control Where Your Code Runs”.

```
rng('default')
tallrng('default')
tMdl = TreeBagger(20,X,Y,'Weights',W,'Prior','Uniform')
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.9 sec
Evaluation completed in 2.3 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 5.4 sec
Evaluation completed in 5.8 sec
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 12 sec
Evaluation completed in 12 sec

tMdl =
  CompactTreeBagger
Ensemble with 20 bagged decision trees:
              Method:       classification
       NumPredictors:                    6
          ClassNames: '0' '1'

  Properties, Methods
```

`tMdl` is a `CompactTreeBagger` ensemble with 20 bagged decision trees.

Calculate the misclassification probability of each tree in the model. Attribute a weight contained in the vector `W` to each observation by using the `'Weights'` name-value pair argument.

```
terr = error(tMdl,X,Y,'Weights',W)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 9.5 sec
Evaluation completed in 9.6 sec

terr = 20×1

    0.1420
    0.1214
```

```
    0.1115
    0.1078
    0.1037
    0.1027
    0.1005
    0.0997
    0.0981
    0.0983
      ⋮
```

Find the average misclassification probability for the ensemble of decision trees.

```
avg_terr = mean(terr)
```

```
avg_terr = 0.1022
```

Tips

- Avoid large estimated out-of-bag error variances by setting a more balanced misclassification cost matrix or a less skewed prior probability vector.
- The `Trees` property of `B` stores a cell array of `B.NumTrees` `CompactClassificationTree` or `CompactRegressionTree` model objects. For a textual or graphical display of tree `t` in the cell array, enter

  ```
  view(B.Trees{t})
  ```
- Standard CART tends to select split predictors containing many distinct values, e.g., continuous variables, over those containing few distinct values, e.g., categorical variables [4]. Consider specifying the curvature or interaction test if any of the following are true:

  - If there are predictors that have relatively fewer distinct values than other predictors, for example, if the predictor data set is heterogeneous.
  - If an analysis of predictor importance is your goal. `TreeBagger` stores predictor importance estimates in the `OOBPermutedPredictorDeltaError` property of `Mdl`.

  For more information on predictor selection, see `PredictorSelection` for classification trees or `PredictorSelection` for regression trees.

Algorithms

- `TreeBagger` generates in-bag samples by oversampling classes with large misclassification costs and undersampling classes with small misclassification costs. Consequently, out-of-bag samples have fewer observations from classes with large misclassification costs and more observations from classes with small misclassification costs. If you train a classification ensemble using a small data set and a highly skewed cost matrix, then the number of out-of-bag observations per class might be very low. Therefore, the estimated out-of-bag error might have a large variance and might be difficult to interpret. The same phenomenon can occur for classes with large prior probabilities.
- For details on selecting split predictors and node-splitting algorithms when growing decision trees, see “Algorithms” on page 33-1928 for classification trees and “Algorithms” on page 33-2413 for regression trees.

Alternative Functionality

Statistics and Machine Learning Toolbox offers three objects for bagging and random forest:

- `ClassificationBaggedEnsemble` created by `fitcensemble` for classification
- `RegressionBaggedEnsemble` created by `fitrensemble` for regression
- `TreeBagger` created by `TreeBagger` for classification and regression

For details about the differences between `TreeBagger` and bagged ensembles (`ClassificationBaggedEnsemble` and `RegressionBaggedEnsemble`), see “Comparison of TreeBagger and Bagged Ensembles” on page 18-44.

References

[1] Breiman, L. "Random Forests." *Machine Learning* 45, pp. 5–32, 2001.

[2] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[3] Loh, W.Y. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” *Statistica Sinica*, Vol. 12, 2002, pp. 361–386.

[4] Loh, W.Y. and Y.S. Shih. “Split Selection Methods for Classification Trees.” *Statistica Sinica*, Vol. 7, 1997, pp. 815–840.

[5] Meinshausen, N. “Quantile Regression Forests.” *Journal of Machine Learning Research*, Vol. 7, 2006, pp. 983–999.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function supports tall arrays for out-of-memory data with the limitations:

- Supported syntaxes for tall `X`, `Y`, `Tbl` are:
  - `B = TreeBagger(NumTrees,Tbl,Y)`
  - `B = TreeBagger(NumTrees,X,Y)`
  - `B = TreeBagger(___,Name,Value)`
- For tall arrays, `TreeBagger` supports classification but does not support regression.
- Supported name-value pairs are:
  - `'NumPredictorsToSample'` — Default value is the square root of the number of variables for classification.
  - `'MinLeafSize'` — Default value is `1` if the number of observations is less than 50,000. If the number of observations is 50,000 or greater, then the default value is `max(1,min(5,floor(0.01*NobsChunk)))`, where `NobsChunk` is the number of observations in a chunk.
  - `'ChunkSize'` (only for tall arrays) — Default value is `50000`.

In addition, `TreeBagger` supports these optional arguments of `fitctree`:

- `'AlgorithmForCategorical'`
- `'CategoricalPredictors'`
- `'Cost'` — The columns of the cost matrix `C` cannot contain `Inf` or `NaN` values.
- `'MaxNumCategories'`
- `'MaxNumSplits'`
- `'MergeLeaves'`
- `'PredictorNames'`
- `'PredictorSelection'`
- `'Prior'`
- `'Prune'`
- `'PruneCriterion'`
- `'SplitCriterion'`
- `'Surrogate'`
- `'Weights'`

- For tall data, `TreeBagger` returns a `CompactTreeBagger` object that contains most of the same properties as a full `TreeBagger` object. The main difference is that the compact object is more memory efficient. The compact object does not include properties that include the data, or that include an array of the same size as the data.
- The number of trees contained in the returned `CompactTreeBagger` object can differ from the number of trees specified as input to the `TreeBagger` function. `TreeBagger` determines the number of trees to return based on factors that include the size of the input data set and the number of data chunks available to grow trees.
- Supported `CompactTreeBagger` methods are:

  - `combine`
  - `error`
  - `margin`
  - `meanMargin`
  - `predict`
  - `setDefaultYfit`

  The `error`, `margin`, `meanMargin`, and `predict` methods do not support the name-value pair arguments `'Trees'`, `'TreeWeights'`, or `'UseInstanceForTree'`. The `meanMargin` method additionally does not support `'Weights'`.
- `TreeBagger` creates a random forest by generating trees on disjoint chunks of the data. When more data is available than is required to create the random forest, the data is subsampled. For a similar example, see Random Forests for Big Data (Genuer, Poggi, Tuleau-Malot, Villa-Vialaneix 2015).

  Depending on how the data is stored, it is possible that some chunks of data contain observations from only a few classes out of all the classes. In this case, `TreeBagger` might produce inferior results compared to the case where each chunk of data contains observations from most of the classes.

- During training of the `TreeBagger` algorithm, the speed, accuracy, and memory usage depend on a number of factors. These factors include values for `NumTrees`, `'ChunkSize'`, `'MinLeafSize'`, and `'MaxNumSplits'`.

  For an *n*-by-*p* tall array `X`, `TreeBagger` implements sampling during training. This sampling depends on these variables:

  - Number of trees `NumTrees`
  - Chunk size `'ChunkSize'`
  - Number of observations *n*
  - Number of chunks *r* (approximately equal to *n*/`'ChunkSize'`)

  Because the value of *n* is fixed for a given `X`, your settings for `NumTrees` and `'ChunkSize'` determine how `TreeBagger` samples `X`.

  1. If *r* > `NumTrees`, then `TreeBagger` samples `'ChunkSize' * NumTrees` observations from `X`, and trains one tree per chunk (with each chunk containing `'ChunkSize'` number of observations). This scenario is the most common when you work with tall arrays.
  2. If *r* ≤ `NumTrees`, then `TreeBagger` trains approximately `NumTrees`/*r* trees in each chunk, using bootstrapping within the chunk.
  3. If *n* ≤ `'ChunkSize'`, then `TreeBagger` uses bootstrapping to generate samples (each of size *n*) on which to train individual trees.

- When specifying a value for `NumTrees`, consider the following:

  - If you run your code on Apache Spark, and your data set is distributed with Hadoop® Distributed File System (HDFS™), start by specifying a value for `NumTrees` that is at least twice the number of partitions in HDFS for your data set. This setting prevents excessive data communication among Apache Spark executors and can improve performance of the `TreeBagger` algorithm.
  - `TreeBagger` copies fitted trees into the client memory in the resulting `CompactTreeBagger` model. Therefore, the amount of memory available to the client creates an upper bound on the value you can set for `NumTrees`. You can tune the values of `'MinLeafSize'` and `'MaxNumSplits'` for more efficient speed and memory usage at the expense of some predictive accuracy. After tuning, if the value of `NumTrees` is less than twice the number of partitions in HDFS for your data set, then consider repartitioning your data in HDFS to have larger partitions.

  After specifying a value for `NumTrees`, set `'ChunkSize'` to ensure that `TreeBagger` uses most of the data to grow trees. Ideally, `'ChunkSize' * NumTrees` should approximate *n*, the number of rows in your data. Note that the memory available in the workers for training individual trees can also determine an upper bound for `'ChunkSize'`.

  You can adjust the Apache Spark memory properties to avoid out-of-memory errors and support your workflow. See `parallel.cluster.Hadoop` (Parallel Computing Toolbox) for more information.

For more information, see “Tall Arrays for Out-of-Memory Data”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To run in parallel, specify the `'Options'` name-value argument in the call to this function and set the `'UseParallel'` field of the options structure to `true` using `statset`.

For example: `'Options',statset('UseParallel',true)`

For more information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`TreeBagger` | `compact` | `error` | `fitctree` | `fitrtree` | `oobError` | `oobPredict` | `predict` | `statset` | `view` | `view`

Topics

“Bootstrap Aggregation (Bagging) of Regression Trees Using TreeBagger” on page 18-113
“Bootstrap Aggregation (Bagging) of Classification Trees Using TreeBagger” on page 18-124
“Framework for Ensemble Learning” on page 18-31
“Decision Trees” on page 19-2
“Grouping Variables” on page 2-45

Introduced in R2009a

trimmean

Mean, excluding outliers

Syntax

```
m = trimmean(X,percent)
m = trimmean(X,percent,flag)
m = trimmean( ___ ,'all')
m = trimmean( ___ ,dim)
m = trimmean( ___ ,vecdim)
```

Description

`m = trimmean(X,percent)` returns the mean of values of `X`, computed after removing the outliers of `X`. For example, if `X` is a vector that has `n` values, `m` is the mean of `X` excluding the highest and lowest `k` data values, where `k = n*(percent/100)/2`.

- If `X` is a vector, then `trimmean(X,percent)` is the mean of all the values of `X`, computed after removing the outliers.
- If `X` is a matrix, then `trimmean(X,percent)` is a row vector of column means, computed after removing the outliers.
- If `X` is a multidimensional array, then `trimmean` operates along the first nonsingleton dimension of `X`.

`m = trimmean(X,percent,flag)` specifies how to trim when `k` (half the number of outliers) is not an integer.

`m = trimmean( ___ ,'all')` returns the trimmed mean of all the values in `X` using any of the input argument combinations in the previous syntaxes.

`m = trimmean( ___ ,dim)` returns the trimmed mean along the operating dimension `dim` of `X`.

`m = trimmean( ___ ,vecdim)` returns the trimmed mean over the dimensions specified in the vector `vecdim`. For example, if `X` is a 2-by-3-by-4 array, then `trimmean(X,10,[1 2])` returns a 1-by-1-by-4 array. Each value of the output array is the mean of the middle 90% of the values on the corresponding page of `X`.

Examples

Efficiency of Trimmed Mean

Find the relative efficiency of the 10% trimmed mean to the sample mean for a given data set.

Generate a 100-by-100 matrix of random numbers from the standard normal distribution. This matrix represents 100 samples, each containing 100 data points.

```
rng default;  % For reproducibility
X = normrnd(0,1,100,100);
```

Compute the sample mean and the 10% trimmed mean for each column of the data matrix.

```
m = mean(X); % Sample mean
trim = trimmean(X,10); % Trimmed mean
```

Compute the relative efficiency of the trimmed mean to the sample mean. The relative efficiency is the variance of the sample mean divided by the variance of the trimmed mean.

```
vm = var(m) % Variance of the sample mean
```

```
vm = 0.0094
```

```
vtrim = var(trim) % Variance of the trimmed mean
```

```
vtrim = 0.0097
```

```
efficiency = vm/vtrim % Relative efficiency of the trimmed mean to the sample mean
```

```
efficiency = 0.9663
```

The sample mean has a smaller variance than the trimmed mean (`efficiency < 1`). Therefore, the trimmed mean is less efficient than the sample mean.

Control Trimming for Distribution with Outliers

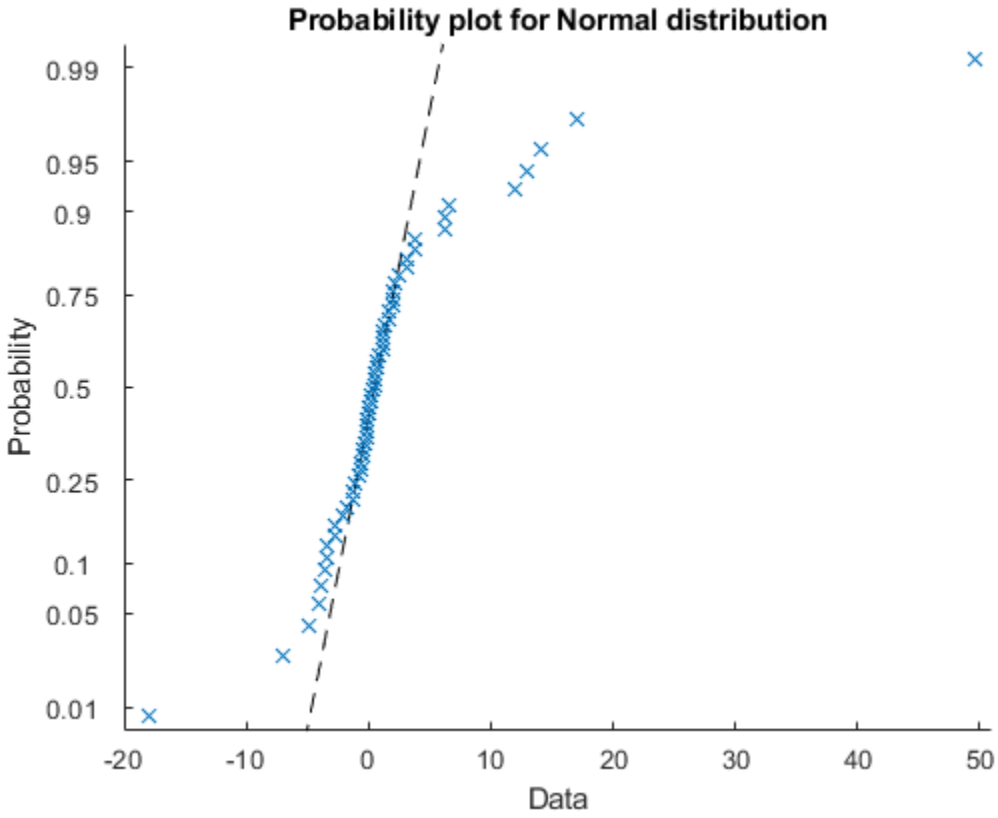
Control the trimming for a distribution with outliers when `k` (half the number of outliers to be trimmed) is not an integer.

Generate a vector of random numbers from the Student's *t* distribution with degrees of freedom equal to 1. The Student's *t* distribution tends to have outliers.

```
rng default;  % For reproducibility
nu = 1; % Degrees of freedom
n = 60; % Number of rows
m = 1;  % Number of columns
x = trnd(nu,n,m); % Vector
```

Visualize the distribution using a normal probability plot.

```
probplot(x)
```



Although the distribution is symmetric around zero, several outliers affect the mean.

Find the mean of the data.

```
mn = mean(x)
```

```
mn = 1.6452
```

Find the 33% trimmed mean of the data.

```
trim = trimmean(x,33)
```

```
trim = 0.4940
```

The 33% trimmed mean is closer to zero, which is more representative of the data. For the 33% trimmed mean, `k` is not an integer (`k = 60*(33/100)/2` gives a value of `9.9`). Therefore, `trimmean` rounds `k` to the nearest integer (`10`) by default.

Control trimming by rounding `k` down to the next smaller integer (`9`). Specify the control for trimming to `'floor'`.

```
trim = trimmean(x,33,'floor')
```

```
trim = 0.4933
```

Find Trimmed Mean Along Given Dimension

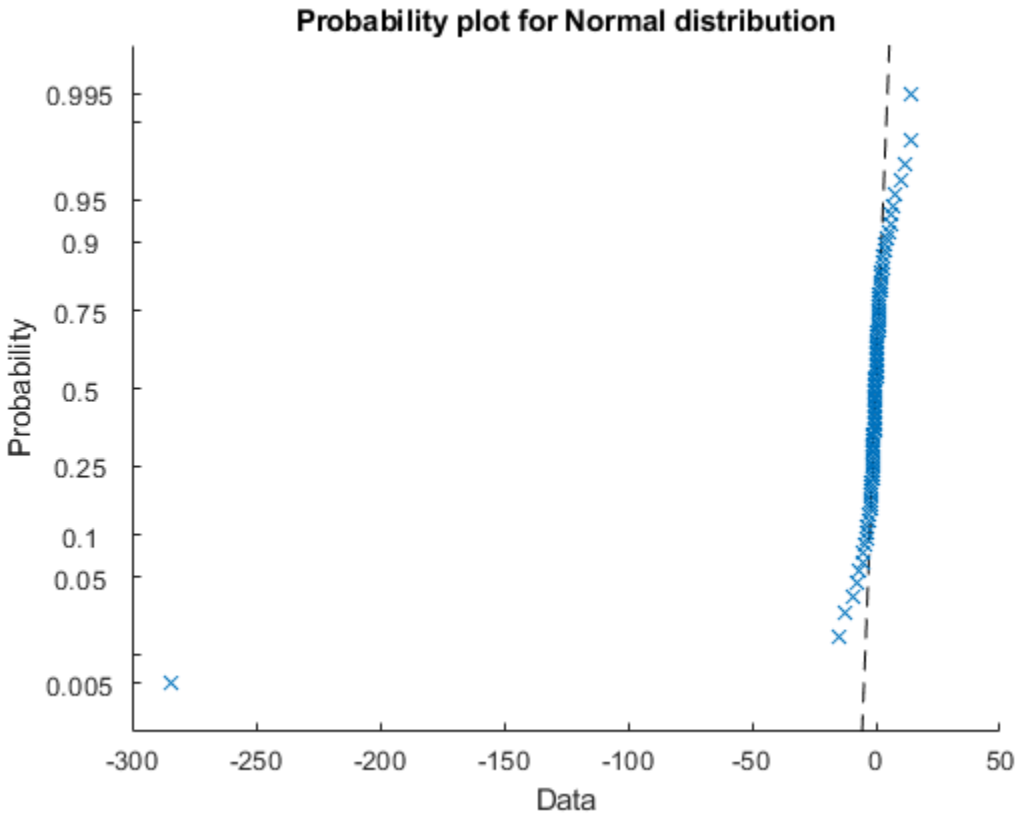
Find the trimmed mean along different dimensions for a matrix.

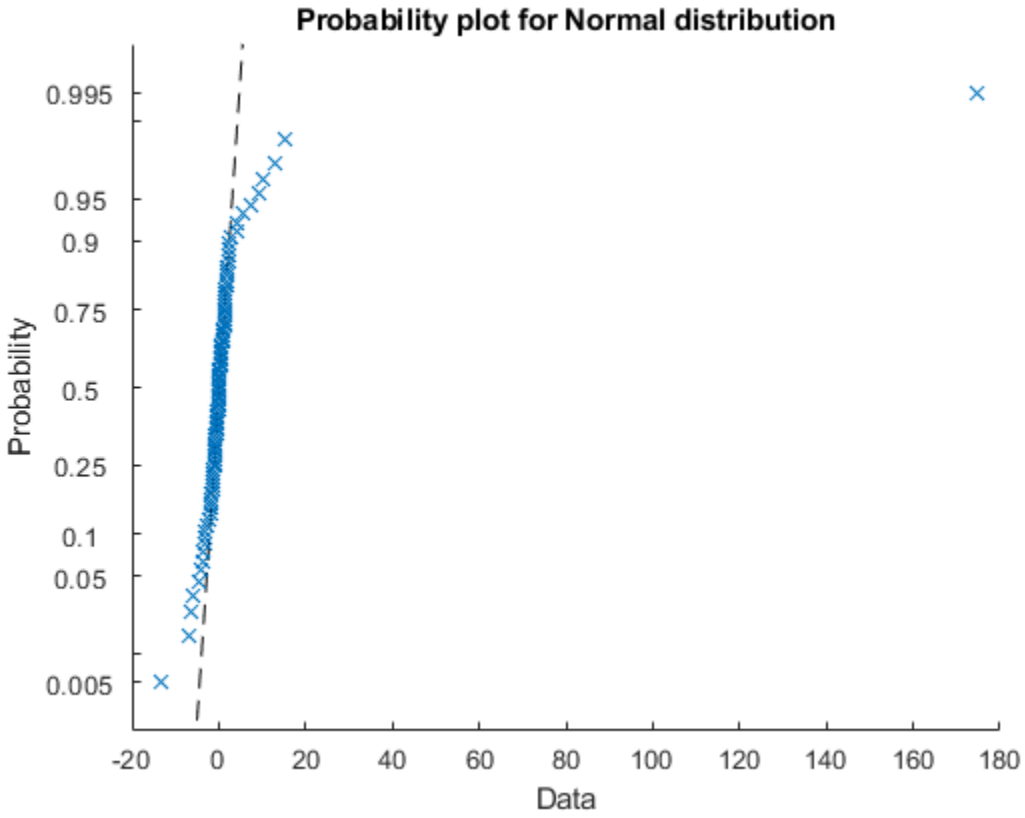
Generate a matrix of random numbers from the Student's *t* distribution. The Student's *t* distribution tends to have outliers.

```
rng('default')
nu = 1; % Degrees of freedom
n = 2; % Number of rows
m = 100;  % Number of columns
X = trnd(nu,n,m);
```

Visualize the distribution for each row of X using a normal probability plot.

```
for i = 1:n
    figure()
    probplot(X(i,:))
end
```





Find the mean for each row of X.

```
mn = mean(X,2)
```

```
mn = 2×1

   -2.7379
    2.0087
```

Find the 30% trimmed mean for each row of X. Specify `dim = 2` as the operating dimension.

```
trim = trimmean(X,30,2)
```

```
trim = 2×1

   -0.0868
    0.1115
```

The 30% trimmed mean of each row is closer to zero, which is more representative of the data.

Trimmed Mean Along Vector of Dimensions

Calculate the trimmed mean over multiple dimensions by using the `'all'` and `vecdim` input arguments.

Create a 5-by-4-by-2 array with some outlier values.

```
X = reshape(1:40,[5 4 2]);
X([3 37]) = -100
```

```
X =
X(:,:,1) =

     1     6    11    16
     2     7    12    17
  -100     8    13    18
     4     9    14    19
     5    10    15    20


X(:,:,2) =

    21    26    31    36
    22    27    32  -100
    23    28    33    38
    24    29    34    39
    25    30    35    40
```

Find the 10% trimmed mean of `X`.

```
mall = trimmean(X,10,'all')
```

```
mall = 19.4722
```

`mall` is the mean of the middle 90% of the values in `X`.

Find the 10% trimmed mean for each page of `X`.

```
mpage = trimmean(X,10,[1 2])
```

```
mpage =
mpage(:,:,1) =

   10.3889


mpage(:,:,2) =

   29.6111
```

For example, `mpage(1,1,2)` is the mean of the middle 90% of the values in `X(:,:,2)`.

Input Arguments

X — Input data

vector | matrix | multidimensional array

Input data that represents a sample from a population, specified as a vector, matrix, or multidimensional array.

- If X is a vector, then `trimmean(X,percent)` is the mean of all the values of X, computed after removing the outliers.
- If X is a matrix, then `trimmean(X,percent)` is a row vector of column means, computed after removing the outliers.
- If X is a multidimensional array, then `trimmean` operates along the first nonsingleton dimension of X.

To specify the operating dimension when X is a matrix or an array, use the `dim` input argument.

`trimmean` treats `NaN` values in X as missing values and removes them.

Data Types: `single` | `double`

`percent` — Percentage
scalar

Percentage of input data to be trimmed, specified as a scalar between `0` and `100`.

`trimmean` uses the value of `percent` to determine the number of outliers (highest and lowest `k` values in X) to remove from X before computing the mean. For X with `n` values, `k = n*(percent/100)/2`.

Data Types: `single` | `double`

`flag` — Control for trimming
`'round'` (default) | `'floor'` | `'weighted'`

Control for trimming when `k` (half the number of outliers) is not an integer, specified as one of the values in this table.

Value	Description
`'round'`	Round `k` to the nearest integer (round to a smaller integer if `k` is a half integer). This value is the default.
`'floor'`	Round `k` down to the next smaller integer.
`'weighted'`	If `k = i + f`, where `i` is an integer and `f` is a fraction, compute a weighted mean with weight `(1 − f)` for the `(i + 1)th` and `(n − i)th` values, and full weight for the values between them.

Data Types: `char` | `string`

`dim` — Dimension
positive integer scalar

Dimension along which to operate, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension of X whose size does not equal 1.

Consider a two-dimensional array X:

- If `dim` is equal to 1, then `trimmean(X,percent,1)` returns a row vector containing the trimmed mean for each column in X.
- If `dim` is equal to 2, then `trimmean(X,percent,2)` returns a column vector containing the trimmed mean for each row in X.

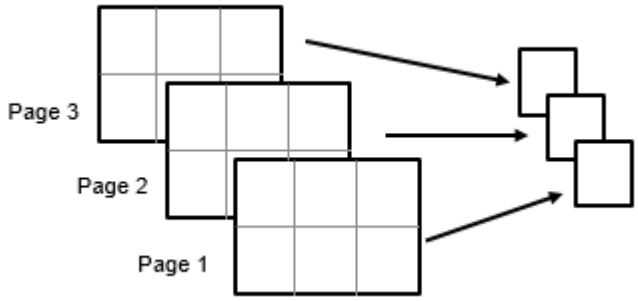
If `dim` is greater than `ndims(X)` or if `size(X,dim)` is 1, then `trimmean` returns X.

Data Types: `single` | `double`

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `m` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `m`.

For example, if `X` is a 2-by-3-by-3 array, then `trimmean(X,10,[1 2])` returns a 1-by-1-by-3 array. Each element of the output is the mean of the middle 90% of the values on the corresponding page of `X`.



Data Types: `single` | `double`

Output Arguments

`m` — Trimmed mean
scalar | vector | matrix | multidimensional array

Trimmed mean values, returned as a scalar, vector, matrix, or multidimensional array.

Tips

- The trimmed mean is a robust estimate of the location of a data sample. If the data contains outliers, then the trimmed mean represents the center of the data better than the sample mean. However, if all the data is from the same probability distribution, then the trimmed mean is less efficient than the sample mean as an estimator of the data location.

Extended Capabilities

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

geomean | harmmean | mean | median

Introduced before R2006a

trnd

Student's *t* random numbers

Syntax

```
r = trnd(nu)
r = trnd(nu,sz1,...,szN)
r = trnd(nu,sz)
```

Description

`r = trnd(nu)` generates a random number from the Student's *t* distribution with `nu` degrees of freedom.

`r = trnd(nu,sz1,...,szN)` generates an array of random numbers from the Student's *t* distribution, where `sz1,...,szN` indicates the size of each dimension.

`r = trnd(nu,sz)` generates an array of random numbers from the Student's *t* distribution, where size vector `sz` specifies `size(r)`.

Examples

Generate Student's *t* Distribution Random Number

Generate a single random number from the Student's *t* distribution with `10` degrees of freedom.

```
nu = 10;
r = trnd(nu)
```

```
r = 1.0585
```

Generate Student's *t* Distribution Random Numbers

Generate a 1-by-6 array of Student's *t* random numbers with `1` degree of freedom.

```
nu1 = ones(1,6); % 1-by-6 array of ones
r1 = trnd(nu1)
```

```
r1 = 1×6

    0.2108    7.8450  -11.0511    0.4134    4.3293   -0.8323
```

If you specify `nu` as a scalar, it expands into a constant array with dimensions specified by `sz1,...,szn`.

Generate a 2-by-6 array of Student's *t* random numbers with `3` degrees of freedom

```
nu2 = 3;
sz1 = 2;
sz2 = 6;
r2 = trnd(nu2,sz1,sz2)
```

```
r2 = 2×6

    0.9257    0.3379    0.6477   -2.2792   -2.8371    0.3632
   -0.2996   -0.6845   -1.2554   -0.5134    1.0458   -0.5521
```

If you specify both `nu` and `sz` as arrays, then the dimensions specified by `sz` must match the dimension of `nu`.

Generate a 1-by-6 array of Student's *t* random numbers with `1` to `6` degrees of freedom.

```
nu3 = 1:6;
sz = [1 6];
r3 = trnd(nu3,sz)
```

```
r3 = 1×6

    1.3609    0.1845   -4.0246   -0.8724   -0.7507    2.3493
```

Input Arguments

`nu` — Degrees of freedom

scalar value | array of scalar values

Degrees of freedom for the Student's *t* distribution, specified as a scalar value or an array of scalar values.

To generate random numbers from multiple distributions, specify `nu` using an array. Each element in `r` is the random number generated from the distribution specified by the corresponding degrees of freedom in `nu`.

Example: `[9 19 49 99]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)

integers

Size of each dimension, specified as separate arguments of integers.

If `nu` is an array, then the specified dimensions `sz1,...,szN` must match the dimensions of `nu`. The default values of `sz1,...,szN` are the dimensions of `nu`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `trnd` ignores trailing dimensions with a size of `1`. For example, `trnd(5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the distribution with `5` degrees of freedom.

Example: `3,5`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If `nu` is an array, then the specified dimensions `sz` must match the dimensions of `nu`. The default values of `sz` are the dimensions of `nu`.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `trnd` ignores trailing dimensions with a size of `1`. For example, `trnd(5,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the distribution with `5` degrees of freedom.

Example: `[3 5]`

Data Types: `single` | `double`

Output Arguments

`r` — Student's *t* random numbers
scalar value | array of scalar values

Student's *t* random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding degrees of freedom in `nu`.

Alternative Functionality

- `trnd` is a function specific to the Student's *t* distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Note that the distribution-specific function `trnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence of numbers returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`random` | `tcdf` | `tinv` | `tpdf` | `tstat`

Topics

“Student's t Distribution” on page B-149

Introduced before R2006a

truncate

Package: `prob`

Truncate probability distribution object

Syntax

```
t = truncate(pd,lower,upper)
```

Description

`t = truncate(pd,lower,upper)` returns a probability distribution `t`, which is the probability distribution `pd` truncated to the specified interval with lower limit, `lower`, and upper limit, `upper`.

Examples

Truncate a Probability Distribution

Create a standard normal probability distribution object.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

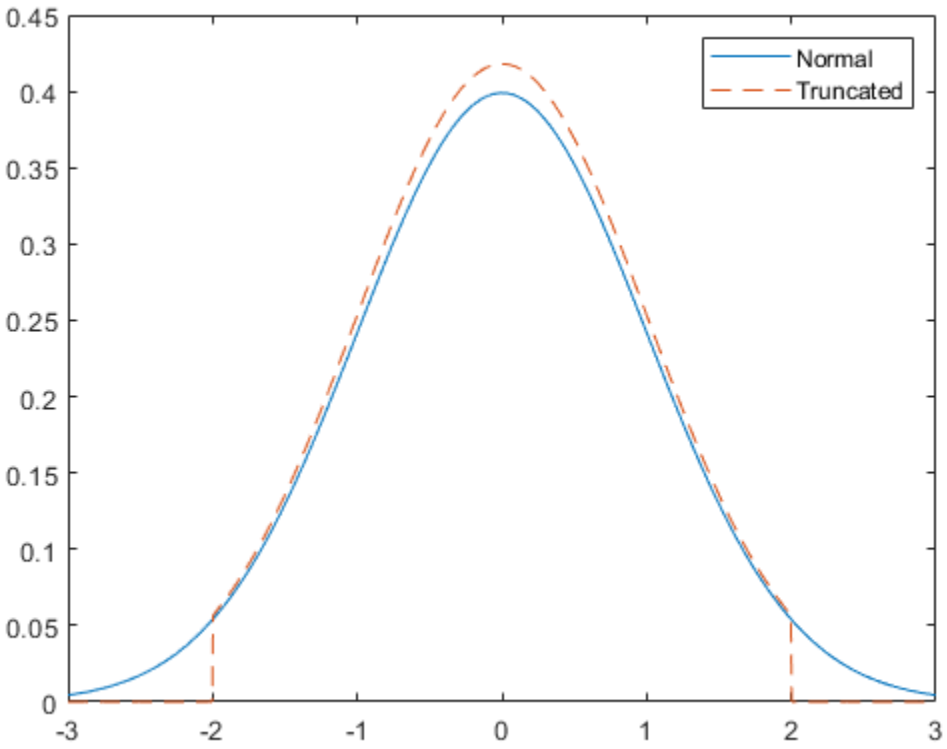
Truncate the distribution to have a lower limit of -2 and an upper limit of 2.

```
t = truncate(pd,-2,2)
```

```
t =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
  Truncated to the interval [-2, 2]
```

Plot the pdf of the original and truncated distributions for a visual comparison.

```
x = linspace(-3,3,1000);
figure
plot(x,pdf(pd,x))
hold on
plot(x,pdf(t,x),'LineStyle','--')
legend('Normal','Truncated')
hold off
```



Generate Random Numbers from a Truncated Distribution

Create a standard normal probability distribution object.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Truncate the distribution by restricting it to positive values. Set the lower limit to 0 and the upper limit to infinity.

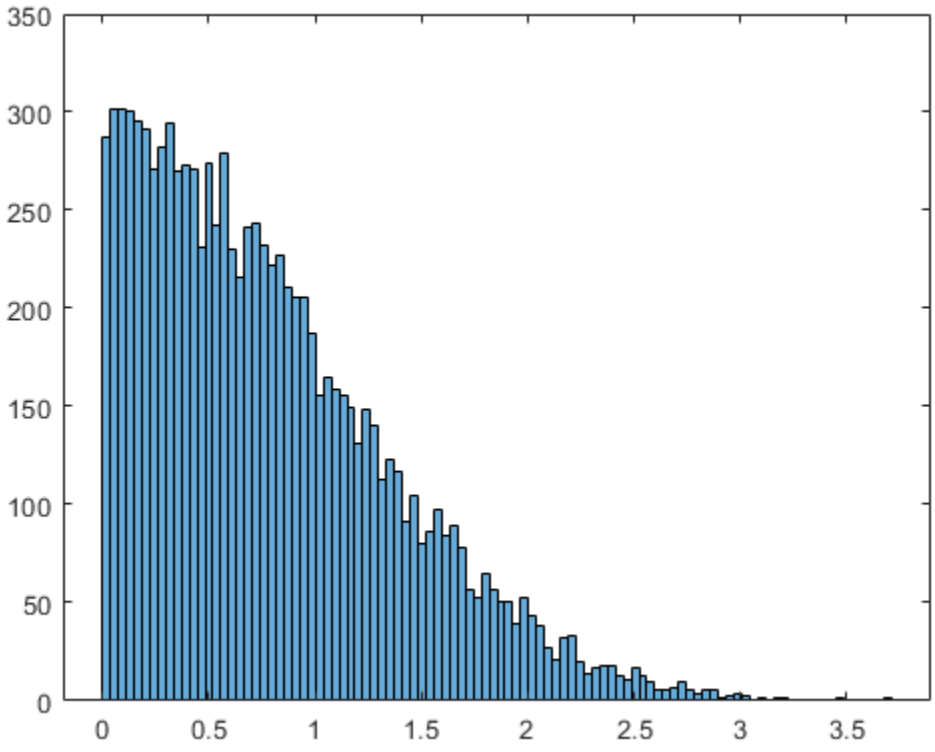
```
t = truncate(pd,0,inf)
```

```
t =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

```
  Truncated to the interval [0, Inf]
```

Generate random numbers from the truncated distribution and visualize with a histogram.

```
r = random(t,10000,1);
histogram(r,100)
```



Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

lower — Lower truncation limit

scalar value

Lower truncation limit, specified as a scalar value.

Data Types: `single` | `double`

`upper` — Upper truncation limit
scalar value

Upper truncation limit, specified as a scalar value.

Data Types: `single` | `double`

Output Arguments

`t` — Truncated distribution
probability distribution object

Truncated distribution, returned as a probability distribution object. The probability distribution function (pdf) of `t` is 0 outside the truncation interval. Inside the truncation interval, the pdf of `t` is equal to the pdf of `pd`, but divided by the probability assigned to that interval by `pd`.

The object properties of `t` are the same as those of `pd` with these exceptions:

- The `Truncation` property of `t` stores the truncation interval.
- The `IsTruncated` property of `t` is 1.
- The `InputData` property of `t` is empty. For a fitted distribution object, the `InputData` property stores the data used for distribution fitting. The truncated distribution object does not store the input data.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see “Code Generation for Probability Distribution Objects” on page 32-82.
- A truncated probability distribution object cannot be an input argument of an entry-point function. To evaluate a truncated distribution using object functions such as `cdf`, `pdf`, `mean`, and so on, call `truncate` and one or more of these object functions within a single entry-point function.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics
“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced in R2013a

tsne

t-Distributed Stochastic Neighbor Embedding

Syntax

```
Y = tsne(X)
Y = tsne(X,Name,Value)
[Y,loss] = tsne( ___ )
```

Description

`Y = tsne(X)` returns a matrix of two-dimensional embeddings of the high-dimensional rows of `X`.

`Y = tsne(X,Name,Value)` modifies the embeddings using options specified by one or more name-value pair arguments.

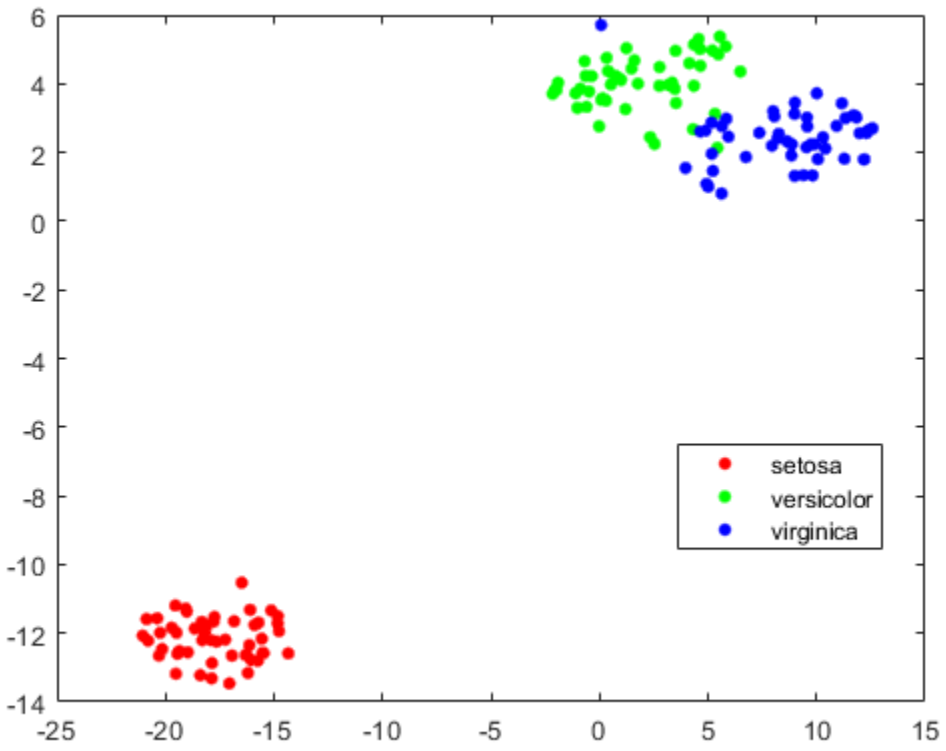
`[Y,loss] = tsne( ___ )`, for any input arguments, also returns the Kullback-Leibler divergence between the joint distributions that model the data `X` and the embedding `Y`.

Examples

Visualize Fisher Iris Data

The Fisher iris data set has four-dimensional measurements of irises, and corresponding classification into species. Visualize this data by reducing the dimension using `tsne`.

```
load fisheriris
rng default % for reproducibility
Y = tsne(meas);
gscatter(Y(:,1),Y(:,2),species)
```

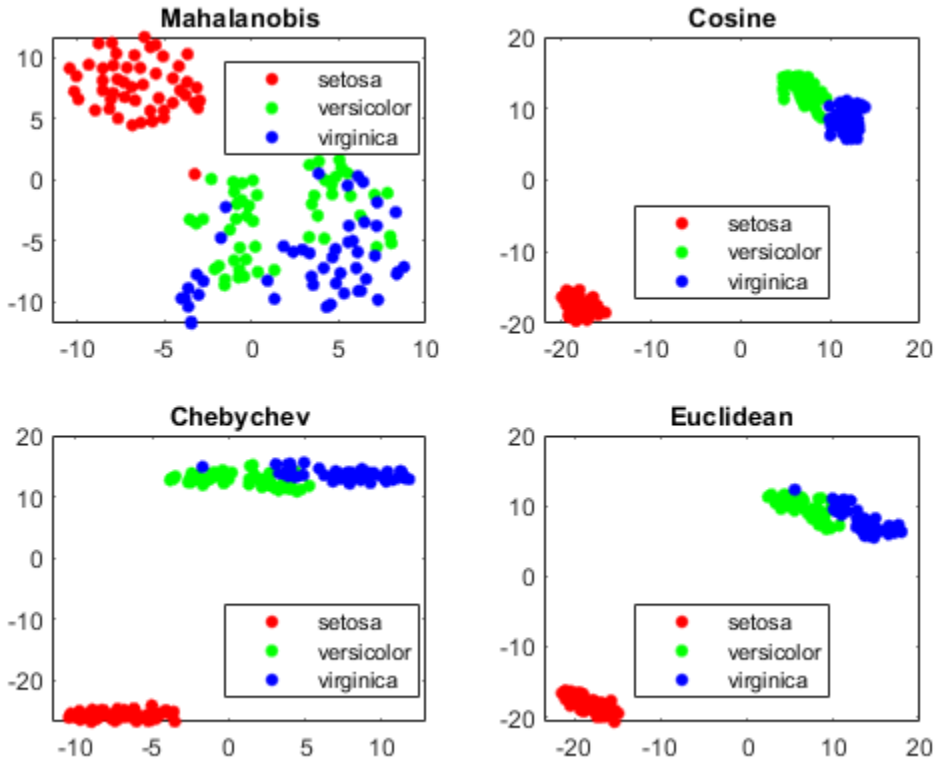


Compare Distance Metrics

Use various distance metrics to try to obtain a better separation between species in the Fisher iris data.

```
load fisheriris

rng('default') % for reproducibility
Y = tsne(meas,'Algorithm','exact','Distance','mahalanobis');
subplot(2,2,1)
gscatter(Y(:,1),Y(:,2),species)
title('Mahalanobis')

rng('default') % for fair comparison
Y = tsne(meas,'Algorithm','exact','Distance','cosine');
subplot(2,2,2)
gscatter(Y(:,1),Y(:,2),species)
title('Cosine')

rng('default') % for fair comparison
Y = tsne(meas,'Algorithm','exact','Distance','chebychev');
subplot(2,2,3)
gscatter(Y(:,1),Y(:,2),species)
title('Chebychev')
```

```
rng('default') % for fair comparison
Y = tsne(meas,'Algorithm','exact','Distance','euclidean');
subplot(2,2,4)
gscatter(Y(:,1),Y(:,2),species)
title('Euclidean')
```



In this case, the cosine, Chebychev, and Euclidean distance metrics give reasonably good separation of clusters. But the Mahalanobis distance metric does not give a good separation.

Plot Results with NaN Input Data

`tsne` removes input data rows that contain any `NaN` entries. Therefore, you must remove any such rows from your classification data before plotting.

For example, change a few random entries in the Fisher iris data to `NaN`.

```
load fisheriris
rng default % for reproducibility
meas(rand(size(meas)) < 0.05) = NaN;
```

Embed the four-dimensional data into two dimensions using `tsne`.

```
Y = tsne(meas,'Algorithm','exact');
```

```
Warning: Rows with NaN missing values in X or 'InitialY' values are removed.
```

Determine how many rows were eliminated from the embedding.

```
length(species)-length(Y)
```

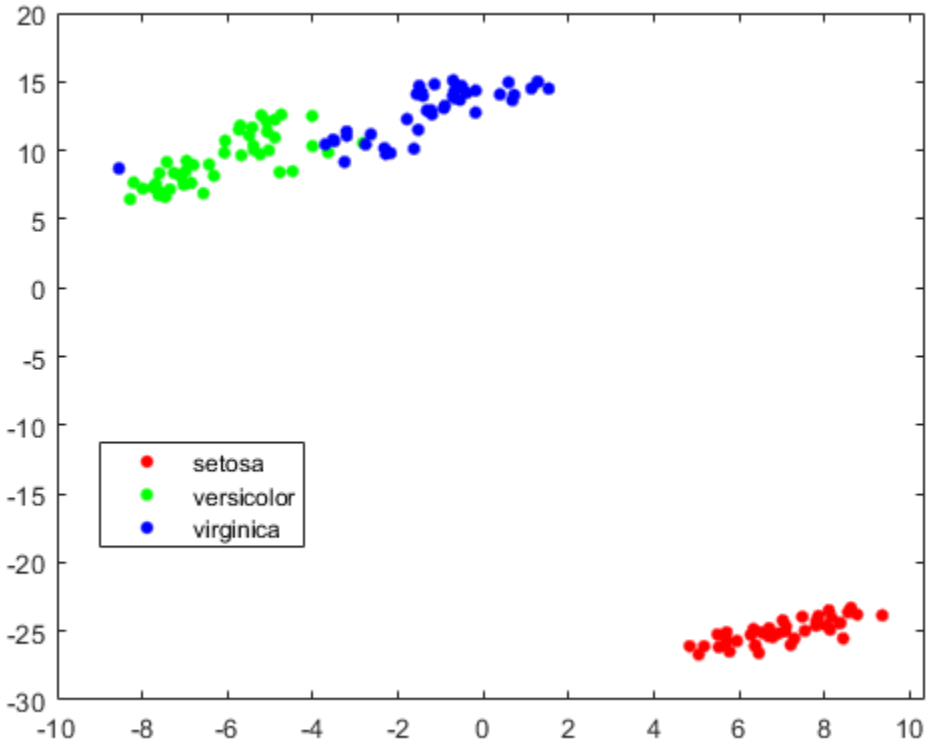
```
ans = 22
```

Prepare to plot the result by locating the rows of `meas` that have no `NaN` values.

```
goodrows = not(any(isnan(meas),2));
```

Plot the results using only the rows of `species` that correspond to rows of `meas` with no `NaN` values.

```
gscatter(Y(:,1),Y(:,2),species(goodrows))
```



Compare t-SNE Loss

Find both 2-D and 3-D embeddings of the Fisher iris data, and compare the loss for each embedding. It is likely that the loss is lower for a 3-D embedding, because this embedding has more freedom to match the original data.

```
load fisheriris
rng default % for reproducibility
[Y,loss] = tsne(meas,'Algorithm','exact');
rng default % for fair comparison
[Y2,loss2] = tsne(meas,'Algorithm','exact','NumDimensions',3);
fprintf('2-D embedding has loss %g, and 3-D embedding has loss %g.\n',loss,loss2)
```

```
2-D embedding has loss 0.1255, and 3-D embedding has loss 0.0956204.
```

As expected, the 3-D embedding has lower loss.

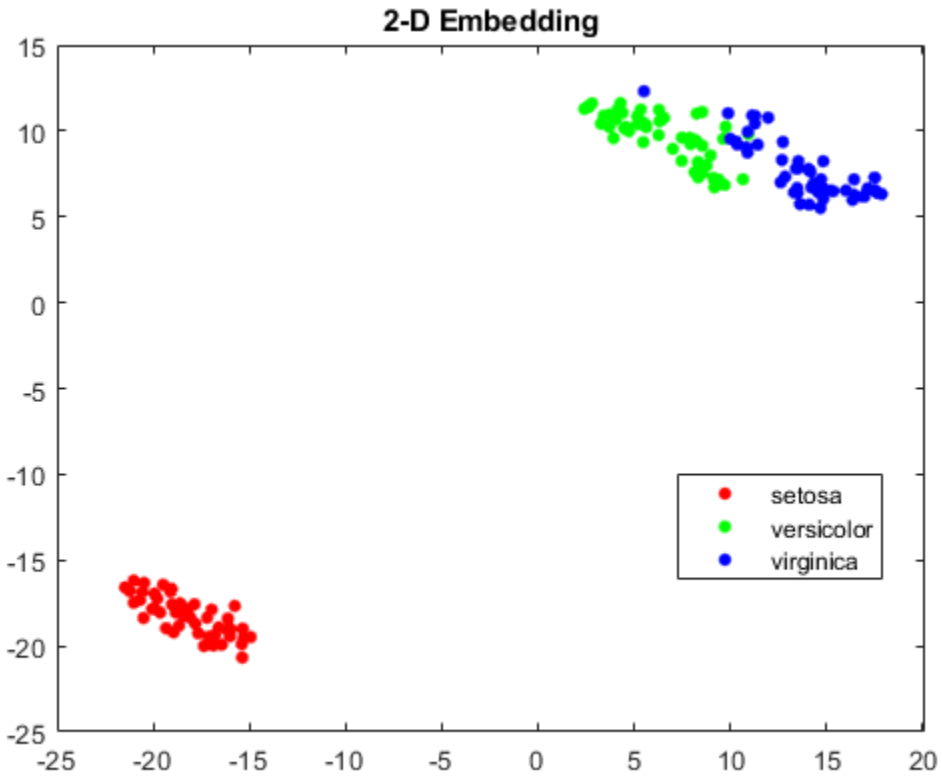
View the embeddings. Use RGB colors `[1 0 0]`, `[0 1 0]`, and `[0 0 1]`.

For the 3-D plot, convert the species to numeric values using the `categorical` command, then convert the numeric values to RGB colors using the `sparse` function as follows. If `v` is a vector of positive integers 1, 2, or 3, corresponding to the species data, then the command

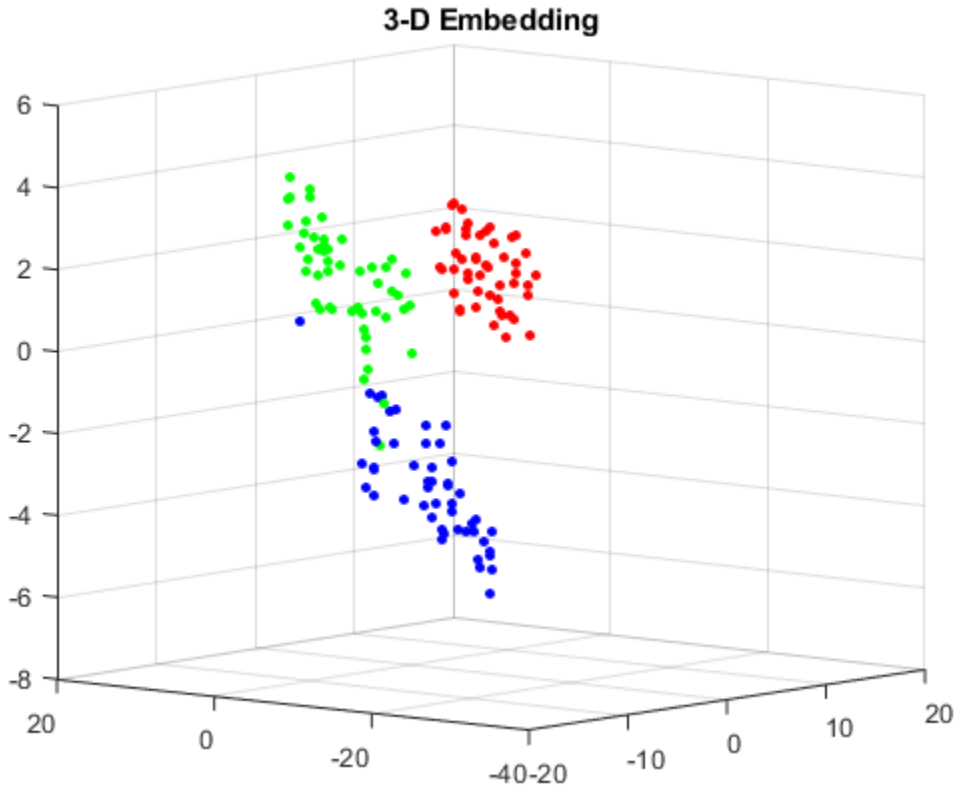
```
sparse(1:numel(v),v,ones(size(v)))
```

is a sparse matrix whose rows are the RGB colors of the species.

```
gscatter(Y(:,1),Y(:,2),species,eye(3))
title('2-D Embedding')
```




```
figure
v = double(categorical(species));
c = full(sparse(1:numel(v),v,ones(size(v)),numel(v),3));
scatter3(Y2(:,1),Y2(:,2),Y2(:,3),15,c,'filled')
title('3-D Embedding')
view(-50,8)
```



Input Arguments

X — Data points

n-by-m matrix

Data points, specified as an n-by-m matrix, where each row is one m-dimensional point.

`tsne` removes rows of X that contain any `NaN` values before creating an embedding. See “Plot Results with NaN Input Data” on page 33-6298.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `Y = tsne(X,'Algorithm','Exact','NumPCAComponents',50)`

Algorithm Control

`Algorithm` — `tsne` algorithm

`'barneshut'` (default) | `'exact'`

`tsne` algorithm, specified as `'barneshut'` or `'exact'`. The `'exact'` algorithm optimizes the Kullback-Leibler divergence of distributions between the original space and the embedded space. The

'barneshut' algorithm performs an approximate optimization that is faster and uses less memory when the number of data rows is large.

Note For the 'barneshut' algorithm, tsne uses knnsearch to find the nearest neighbors.

Example: 'exact'

Distance — Distance metric
'euclidean' (default) | 'seuclidean' | 'cityblock' | 'chebychev' | 'minkowski' | 'mahalanobis' | 'cosine' | 'correlation' | 'spearman' | 'hamming' | 'jaccard' | function handle

Distance metric, specified by one of the following. For definitions of the distance metrics, see pdist.

- 'euclidean' — Euclidean distance.
- 'seuclidean' — Standardized Euclidean distance. Each coordinate difference between rows in X and the query matrix is scaled by dividing by the corresponding element of the standard deviation computed from S = std(X,'omitnan').
- 'cityblock' — City block distance.
- 'chebychev' — Chebychev distance, which is the maximum coordinate difference.
- 'minkowski' — Minkowski distance with exponent 2. This is the same as Euclidean distance.
- 'mahalanobis' — Mahalanobis distance, computed using the positive definite covariance matrix cov(X,'omitrows').
- 'cosine' — 1 minus the cosine of the included angle between observations (treated as vectors).
- 'correlation' — One minus the sample linear correlation between observations (treated as sequences of values).
- 'spearman' — One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
- 'hamming' — Hamming distance, which is the percentage of coordinates that differ.
- 'jaccard' — One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
- custom distance function — A distance function specified using @ (for example, @distfun). For details, see “More About” on page 33-6305.

In all cases, tsne uses squared pairwise distances to calculate the Gaussian kernel in the joint distribution of X.

Example: 'mahalanobis'

Exaggeration — Size of natural clusters in data
4 (default) | scalar value 1 or greater

Size of natural clusters in data, specified as a scalar value 1 or greater.

A large exaggeration makes tsne learn larger joint probabilities of Y and creates relatively more space between clusters in Y. tsne uses exaggeration in the first 99 optimization iterations.

If the value of Kullback-Leibler divergence increases in the early stage of the optimization, try reducing the exaggeration. See “tsne Settings” on page 15-117.

Example: `10`

Data Types: `single` | `double`

`NumDimensions` — Dimension of the output `Y`
`2` (default) | positive integer

Dimension of the output `Y`, specified as a positive integer. Generally, set `NumDimensions` to `2` or `3`.

Example: `3`

Data Types: `single` | `double`

`NumPCAComponents` — PCA dimension reduction
`0` (default) | nonnegative integer

PCA dimension reduction, specified as a nonnegative integer. Before `tsne` embeds the high-dimensional data, it first reduces the dimensionality of the data to `NumPCAComponents` using the `pca` function. When `NumPCAComponents` is `0`, `tsne` does not use PCA.

Example: `50`

Data Types: `single` | `double`

`Perplexity` — Effective number of local neighbors of each point
`30` (default) | positive scalar

Effective number of local neighbors of each point, specified as a positive scalar. See “t-SNE Algorithm” on page 15-104.

Larger perplexity causes `tsne` to use more points as nearest neighbors. Use a larger value of `Perplexity` for a large dataset. Typical `Perplexity` values are from `5` to `50`. In the Barnes-Hut algorithm, `tsne` uses `min(3*Perplexity,N-1)` as the number of nearest neighbors. See “tsne Settings” on page 15-117.

Example: `10`

Data Types: `single` | `double`

`Standardize` — Normalize input data
`false` (default) | `true`

Normalize input data, specified as `false` or `true`. When `true`, `tsne` centers and scales `X` by dividing the columns by their standard deviations.

When features in `X` are on different scales, set `'Standardize'` to `true`. Do this because the learning process is based on nearest neighbors, so features with large scales can override the contribution of features with small scales.

Example: `true`

Data Types: `logical`

Optimization Control

`InitialY` — Initial embedded points
`1e-4*randn(N,NumDimensions)` (default) | n-by-`NumDimensions` real matrix

Initial embedded points, specified as an `n`-by-`NumDimensions` real matrix, where `n` is the number of rows of `X`. The `tsne` optimization algorithm uses these points as initial values.

Data Types: `single` | `double`

`LearnRate` — Learning rate for optimization process
`500` (default) | positive scalar

Learning rate for optimization process, specified as a positive scalar. Typically, set values from `100` through `1000`.

When `LearnRate` is too small, `tsne` can converge to a poor local minimum. When `LearnRate` is too large, the optimization can initially have the Kullback-Leibler divergence increase rather than decrease. See “tsne Settings” on page 15-117.

Example: `1000`

Data Types: `single` | `double`

`NumPrint` — Iterative display frequency
`20` (default) | positive integer

Iterative display frequency, specified as a positive integer. When the `Verbose` name-value pair is not `0`, `tsne` returns iterative display after every `NumPrint` iterations. If the `Options` name-value pair contains a nonempty `'OutputFcn'` entry, then output functions run after every `NumPrint` iterations.

Example: `20`

Data Types: `single` | `double`

`Options` — Optimization options
structure containing the fields `'MaxIter'`, `'OutputFcn'`, and `'TolFun'`

Optimization options, specified as a structure containing the fields `'MaxIter'`, `'OutputFcn'`, and `'TolFun'`. Create `'Options'` using `statset` or `struct`.

- `'MaxIter'` — Positive integer specifying the maximum number of optimization iterations. Default: `1000`.
- `'OutputFcn'` — Function handle or cell array of function handles specifying one or more functions to call after every `NumPrint` optimization iterations. For syntax details, see “t-SNE Output Function” on page 15-110. Default: `[]`.
- `'TolFun'` — Stopping criterion for the optimization. The optimization exits when the norm of the gradient of the Kullback-Leibler divergence is less than `'TolFun'`. Default: `1e-10`.

Example: `options = statset('MaxIter',500)`

Data Types: `struct`

`Theta` — Barnes-Hut tradeoff parameter
`0.5` (default) | scalar from 0 through 1

Barnes-Hut tradeoff parameter, specified as a scalar from 0 through 1. Higher values give a faster but less accurate optimization. Applies only when `Algorithm` is `'barneshut'`.

Example: `0.1`

Data Types: `single` | `double`

`Verbose` — Iterative display
`0` (default) | `1` | `2`

Iterative display, specified as `0`, `1`, or `2`. When `Verbose` is not `0`, `tsne` prints a summary table of the Kullback-Leibler divergence and the norm of its gradient every `NumPrint` iterations.

When `Verbose` is `2`, `tsne` also prints the variances of Gaussian kernels. `tsne` uses these kernels in its computation of the joint probability of `X`. If you see a large difference in the scales of the minimum and maximum variances, you can sometimes get more suitable results by rescaling `X`.

Example: 2

Data Types: `single` | `double`

Output Arguments

`Y` — Embedded points

`n`-by-`NumDimensions` matrix

Embedded points, returned as an `n`-by-`NumDimensions` matrix. Each row represents one embedded point. `n` is the number of rows of data `X` that do not contain any `NaN` entries. See “Plot Results with NaN Input Data” on page 33-6298.

`loss` — Kullback-Leibler divergence

nonnegative scalar

Kullback-Leibler divergence between modeled input and output distributions, returned as a nonnegative scalar. For details, see “t-SNE Algorithm” on page 15-104.

More About

Custom Distance Function

The syntax of a custom distance function is as follows.

```
function D2 = distfun(ZI,ZJ)
```

`tsne` passes `ZI` and `ZJ` to your function, and your function computes the distance.

- `ZI` is a 1-by-*n* vector containing a single row from `X` or `Y`.
- `ZJ` is an *m*-by-*n* matrix containing multiple rows of `X` or `Y`.

Your function returns `D2`, which is an *m*-by-1 vector of distances. The *j*th element of `D2` is the distance between the observations `ZI` and `ZJ(j,:)`.

Tip If your data are not sparse, then usually the built-in distance functions are faster than a function handle.

Algorithms

`tsne` constructs a set of embedded points in a low-dimensional space whose relative similarities mimic those of the original high-dimensional points. The embedded points show the clustering in the original data.

Roughly, the algorithm models the original points as coming from a Gaussian distribution, and the embedded points as coming from a Student's *t* distribution. The algorithm tries to minimize the Kullback-Leibler divergence between these two distributions by moving the embedded points.

For details, see "t-SNE" on page 15-104.

See Also

`gscatter` | `knnsearch` | `pca` | `pdist` | `statset`

Topics

"Visualize High-Dimensional Data Using t-SNE" on page 15-113
"t-SNE Custom Output Function" on page 15-111
"tsne Settings" on page 15-117
"t-SNE" on page 15-104
"Dimensionality Reduction and Feature Extraction"

Introduced in R2017a

tstat

Student's *t* mean and variance

Syntax

```
[m,v] = tstat(nu)
```

Description

`[m,v] = tstat(nu)` returns the mean and variance of the Student's *t* distribution with `nu` degrees of freedom.

Examples

Compute Mean and Variance of Student's *t* Distribution

Compute the mean and variance for Student's *t* distribution with degrees of freedom *nu* equal to `1` to `30`.

```
nu = reshape(1:30,6,5);
[m,v] = tstat(nu)
```

```
m = 6×5

   NaN     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0
     0     0     0     0     0


v = 6×5

       NaN    1.4000    1.1818    1.1176    1.0870
       NaN    1.3333    1.1667    1.1111    1.0833
    3.0000    1.2857    1.1538    1.1053    1.0800
    2.0000    1.2500    1.1429    1.1000    1.0769
    1.6667    1.2222    1.1333    1.0952    1.0741
    1.5000    1.2000    1.1250    1.0909    1.0714
```

Note that the mean is undefined for `1` degree of freedom, and variance is undefined for `1` and `2` degrees of freedom.

Input Arguments

nu — Degrees of freedom

positive scalar value | array of positive scalar values

Degrees of freedom for the Student's *t* distribution, specified as a positive scalar value or an array of positive scalar values.

Example: `[9,19,49,99]`

Data Types: `single` | `double`

Output Arguments

`m` — Mean

scalar value | array of scalar values

Mean of the Student's *t* distribution with the degrees of freedom specified in `nu`, returned as a scalar value or an array of scalar values. `m` is the same size as `nu`.

`v` — Variance

scalar value | array of scalar values

Variance of the Student's *t* distribution with the degrees of freedom specified in `nu`, returned as a scalar value or an array of scalar values. `v` is the same size as `nu`.

More About

Mean and Variance of Student's *t* Distribution

The parameters of the Student's *t* distribution depend on the degrees of freedom.

The mean of the Student's *t* distribution is $\mu = 0$ for degrees of freedom ν greater than 1. If ν equals 1, then the mean is undefined.

The variance of the Student's *t* distribution is $\frac{\nu}{\nu - 2}$ for degrees of freedom ν greater than 2. If ν is less than or equal to 2, then the variance is undefined.

For more information, see "Student's t Distribution" on page B-149.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`tcdf` | `tinv` | `tpdf` | `trnd`

Topics

"Student's t Distribution" on page B-149

Introduced before R2006a

ttest

One-sample and paired-sample *t*-test

Syntax

```
h = ttest(x)

h = ttest(x,y)
h = ttest(x,y,Name,Value)

h = ttest(x,m)
h = ttest(x,m,Name,Value)

[h,p] = ttest(___)
[h,p,ci,stats] = ttest(___)
```

Description

`h = ttest(x)` returns a test decision for the null hypothesis that the data in `x` comes from a normal distribution with mean equal to zero and unknown variance, using the one-sample *t*-test on page 33-6315. The alternative hypothesis is that the population distribution does not have a mean equal to zero. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = ttest(x,y)` returns a test decision for the null hypothesis that the data in `x − y` comes from a normal distribution with mean equal to zero and unknown variance, using the paired-sample *t*-test.

`h = ttest(x,y,Name,Value)` returns a test decision for the paired-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`h = ttest(x,m)` returns a test decision for the null hypothesis that the data in `x` comes from a normal distribution with mean `m` and unknown variance. The alternative hypothesis is that the mean is not `m`.

`h = ttest(x,m,Name,Value)` returns a test decision for the one-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = ttest(___)` also returns the *p*-value, `p`, of the test, using any of the input arguments from the previous syntax groups.

`[h,p,ci,stats] = ttest(___)` also returns the confidence interval `ci` for the mean of `x`, or of `x − y` for the paired *t*-test, and the structure `stats` containing information about the test statistic.

Examples

t-Test for Mean Equal to Zero

Load the sample data. Create a vector containing the third column of the stock returns data.

```
load stockreturns
x = stocks(:,3);
```

Test the null hypothesis that the sample data comes from a population with mean equal to zero.

```
[h,p,ci,stats] = ttest(x)
```

```
h = 1

p = 0.0106

ci = 2×1

   -0.7357
   -0.0997


stats = struct with fields:
    tstat: -2.6065
       df: 99
       sd: 1.6027
```

The returned value `h = 1` indicates that `ttest` rejects the null hypothesis at the 5% significance level.

t-Test at Different Significance Level

Load the sample data. Create a vector containing the third column of the stock returns data.

```
load stockreturns
x = stocks(:,3);
```

Test the null hypothesis that the sample data are from a population with mean equal to zero at the 1% significance level.

```
h = ttest(x,0,'Alpha',0.01)
```

```
h = 0
```

The returned value `h = 0` indicates that `ttest` does not reject the null hypothesis at the 1% significance level.

Paired-Sample t-Test

Load the sample data. Create vectors containing the first and second columns of the data matrix to represent students' grades on two exams.

```
load examgrades
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the pairwise difference between data vectors `x` and `y` has a mean equal to zero.

```
[h,p] = ttest(x,y)
```

```
h = 0
```

```
p = 0.9805
```

The returned value of `h = 0` indicates that `ttest` does not reject the null hypothesis at the default 5% significance level.

Paired-Sample t-Test at Different Significance Level

Load the sample data. Create vectors containing the first and second columns of the data matrix to represent students' grades on two exams.

```
load examgrades
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the pairwise difference between data vectors `x` and `y` has a mean equal to zero at the 1% significance level.

```
[h,p] = ttest(x,y,'Alpha',0.01)
```

```
h = 0
```

```
p = 0.9805
```

The returned value of `h = 0` indicates that `ttest` does not reject the null hypothesis at the 1% significance level.

t-Test for a Hypothesized Mean

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that sample data comes from a distribution with mean `m = 75`.

```
h = ttest(x,75)
```

```
h = 0
```

The returned value of `h = 0` indicates that `ttest` does not reject the null hypothesis at the 5% significance level.

One-Sided t-Test

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the data comes from a population with mean equal to 65, against the alternative that the mean is greater than 65.

```
h = ttest(x,65,'Tail','right')
```

```
h = 1
```

The returned value of `h = 1` indicates that `ttest` rejects the null hypothesis at the 5% significance level, in favor of the alternate hypothesis that the data comes from a population with a mean greater than 65.

Input Arguments

x — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array on page 33-6315. `ttest` performs a separate *t*-test along each column and returns a vector of results. If `y` sample data is specified, `x` and `y` must be the same size.

Data Types: `single` | `double`

y — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array on page 33-6315. If `y` sample data is specified, `x` and `y` must be the same size.

Data Types: `single` | `double`

m — Hypothesized population mean

0 (default) | scalar value

Hypothesized population mean, specified as a scalar value.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01` conducts a right-tailed hypothesis test at the 1% significance level.

Alpha — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension
first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix along which to test the means, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the column means, while `'Dim',2` tests the row means.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis
`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of:

- `'both'` — Test against the alternative hypothesis that the population mean is not `m`.
- `'right'` — Test against the alternative hypothesis that the population mean is greater than `m`.
- `'left'` — Test against the alternative hypothesis that the population mean is less than `m`.

`ttest` tests the null hypothesis that the population mean is `m` against the specified alternative hypothesis.

Example: `'Tail','right'`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`ci` — Confidence interval
vector

Confidence interval for the true population mean, returned as a two-element vector containing the lower and upper boundaries of the 100 × (1 – `Alpha`)% confidence interval.

`stats` — Test statistics
structure

Test statistics, returned as a structure containing the following:

- `tstat` — Value of the test statistic.
- `df` — Degrees of freedom of the test.
- `sd` — Estimated population standard deviation. For a paired *t*-test, `sd` is the standard deviation of `x − y`.

More About

One-Sample t-Test

The one-sample *t*-test is a parametric test of the location parameter when the population standard deviation is unknown.

The test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ is the hypothesized population mean, s is the sample standard deviation, and n is the sample size. Under the null hypothesis, the test statistic has Student's t distribution with n – 1 degrees of freedom.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

Tips

- Use `sampsizepwr` to calculate:
  - The sample size that corresponds to specified power and parameter values;
  - The power achieved for a particular sample size, given the true parameter value;
  - The parameter value detectable with the specified sample size and power.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also

`sampsizepwr` | `ttest2` | `ztest`

Introduced before R2006a

ttest2

Two-sample *t*-test

Syntax

```
h = ttest2(x,y)
h = ttest2(x,y,Name,Value)
[h,p] = ttest2(___)
[h,p,ci,stats] = ttest2(___)
```

Description

`h = ttest2(x,y)` returns a test decision for the null hypothesis that the data in vectors `x` and `y` comes from independent random samples from normal distributions with equal means and equal but unknown variances, using the two-sample *t*-test on page 33-6321. The alternative hypothesis is that the data in `x` and `y` comes from populations with unequal means. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = ttest2(x,y,Name,Value)` returns a test decision for the two-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct the test without assuming equal variances.

`[h,p] = ttest2(___)` also returns the *p*-value, `p`, of the test, using any of the input arguments in the previous syntaxes.

`[h,p,ci,stats] = ttest2(___)` also returns the confidence interval on the difference of the population means, `ci`, and the structure `stats` containing information about the test statistic.

Examples

Two-Sample t-Test for Equal Means

Load the data set. Create vectors containing the first and second columns of the data matrix to represent students' grades on two exams.

```
load examgrades
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the two data samples are from populations with equal means.

```
[h,p,ci,stats] = ttest2(x,y)
```

```
h = 0
```

```
p = 0.9867
```

```
ci = 2×1
```

```
   -1.9438
```

```
    1.9771


stats = struct with fields:
    tstat: 0.0167
       df: 238
       sd: 7.7084
```

The returned value of `h = 0` indicates that `ttest2` does not reject the null hypothesis at the default 5% significance level.

t-Test for Equal Means Without Assuming Equal Variances

Load the data set. Create vectors containing the first and second columns of the data matrix to represent students' grades on two exams.

```
load examgrades
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the two data vectors are from populations with equal means, without assuming that the populations also have equal variances.

```
[h,p] = ttest2(x,y,'Vartype','unequal')
```

```
h = 0

p = 0.9867
```

The returned value of `h = 0` indicates that `ttest2` does not reject the null hypothesis at the default 5% significance level even if equal variances are not assumed.

Input Arguments

x — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array. `ttest2` treats `NaN` values as missing data and ignores them.

- If `x` and `y` are specified as vectors, they do not need to be the same length.
- If `x` and `y` are specified as matrices, they must have the same number of columns. `ttest2` performs a separate *t*-test along each column and returns a vector of results.
- If `x` and `y` are specified as multidimensional arrays on page 33-6321, they must have the same size along all but the first nonsingleton dimension on page 33-6321.

Data Types: `single` | `double`

y — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array. `ttest2` treats `NaN` values as missing data and ignores them.

- If `x` and `y` are specified as vectors, they do not need to be the same length.
- If `x` and `y` are specified as matrices, they must have the same number of columns. `ttest2` performs a separate *t*-test along each column and returns a vector of results.
- If `x` and `y` are specified as multidimensional arrays on page 33-6321, they must have the same size along all but the first nonsingleton dimension on page 33-6321. `ttest2` works along the first nonsingleton dimension.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01,'Vartype','unequal'` specifies a right-tailed test at the 1% significance level, and does not assume that `x` and `y` have equal population variances.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension

first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix along which to test the means, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the column means, while `'Dim',2` tests the row means.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of:

- `'both'` — Test against the alternative hypothesis that the population means are not equal.
- `'right'` — Test against the alternative hypothesis that the population mean of `x` is greater than the population mean of `y`.
- `'left'` — Test against the alternative hypothesis that the population mean of `x` is less than the population mean of `y`.

`ttest2` tests the null hypothesis that the population means are equal against the specified alternative hypothesis.

Example: `'Tail','right'`

`Vartype` — Variance type
`'equal'` (default) | `'unequal'`

Variance type, specified as the comma-separated pair consisting of `'Vartype'` and one of the following.

`'equal'`	Conduct test using the assumption that `x` and `y` are from normal distributions with unknown but equal variances.
`'unequal'`	Conduct test using the assumption that `x` and `y` are from normal distributions with unknown and unequal variances. This is called the Behrens-Fisher problem. `ttest2` uses Satterthwaite's approximation for the effective degrees of freedom.

`Vartype` must be a single variance type, even when `x` is a matrix or a multidimensional array.

Example: `'Vartype','unequal'`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`ci` — Confidence interval
vector

Confidence interval for the difference in population means of `x` and `y`, returned as a two-element vector containing the lower and upper boundaries of the 100 × (1 – `Alpha`)% confidence interval.

`stats` — Test statistics
structure

Test statistics for the two-sample *t*-test, returned as a structure containing the following:

- `tstat` — Value of the test statistic.
- `df` — Degrees of freedom of the test.
- `sd` — Pooled estimate of the population standard deviation (for the equal variance case) or a vector containing the unpooled estimates of the population standard deviations (for the unequal variance case).

More About

Two-Sample *t*-test

The two-sample *t*-test is a parametric test that compares the location parameter of two independent data samples.

The test statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}},$$

where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y are the sample standard deviations, and n and m are the sample sizes.

In the case where it is assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has Student's t distribution with $n + m - 2$ degrees of freedom, and the sample standard deviations are replaced by the pooled standard deviation

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}.$$

In the case where it is not assumed that the two data samples are from populations with equal variances, the test statistic under the null hypothesis has an approximate Student's t distribution with a number of degrees of freedom given by Satterthwaite's approximation. This test is sometimes called Welch's t-test.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

Tips

- Use `sampsizepwr` to calculate:
  - The sample size that corresponds to specified power and parameter values;
  - The power achieved for a particular sample size, given the true parameter value;
  - The parameter value detectable with the specified sample size and power.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`sampsizepwr` | `ttest` | `ztest`

Introduced before R2006a

BetaDistribution

Beta probability distribution object

Description

A `BetaDistribution` object consist of parameters, a model description, and sample data for a beta probability distribution.

The beta distribution describes a family of curves that are unique in that they are nonzero only on the interval (0,1). A more general version of the distribution assigns parameters to the endpoints of the interval.

The beta distribution uses the following parameters.

Parameter	Description	Support
a	First shape parameter	$a > 0$
b	Second shape parameter	$b > 0$

Creation

There are several ways to create a `BetaDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

a — First shape parameter
positive scalar value

First shape parameter of the beta distribution, specified as a positive scalar value.

Data Types: `single` | `double`

b — Second shape parameter
positive scalar value

Second shape parameter of the beta distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Beta Distribution Object Using Default Parameters

Create a beta distribution object using the default parameter values.

```
pd = makedist('Beta')
```

```
pd =
  BetaDistribution

  Beta distribution
    a = 1
    b = 1
```

Create a Beta Distribution Object Using Specified Parameters

Create a beta distribution object by specifying the parameter values.

```
pd = makedist('Beta','a',2,'b',4)
```

```
pd =
  BetaDistribution

  Beta distribution
    a = 2
    b = 4
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 0.3333
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see “Code Generation” on page 33-2252 of `fitdist`.
- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Probability Distribution Objects” on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Beta Distribution” on page B-6

Introduced in R2013a

BinomialDistribution

Binomial probability distribution object

Description

A `BinomialDistribution` object consists of parameters, a model description, and sample data for a binomial probability distribution

The binomial distribution models the total number of successes in repeated trials from an infinite population under the following conditions:

- Only two outcomes are possible for each of n trials.
- The probability of success for each trial is constant.
- All trials are independent of each other.

The binomial distribution uses the following parameters.

Parameter	Description	Support
`N`	Number of trials	positive integer
`p`	Probability of success	$0 \le p \le 1$

Creation

There are several ways to create a `BinomialDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`N` — Number of trials
positive integer value

Number of trials for the binomial distribution, specified as a positive integer value.

Data Types: `single` | `double`

`p` — Probability of success
positive scalar value in the range `[0,1]`

Probability of success of any individual trial for the binomial distribution, specified as a positive scalar value in the range `[0,1]`.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Negative Binomial Distribution Object Using Default Parameters

Create a negative binomial distribution object using the default parameter values.

```
pd = makedist('NegativeBinomial')
```

```
pd =
  NegativeBinomialDistribution

  Negative Binomial distribution
    R =   1
    P = 0.5
```

Create a Binomial Distribution Object Using Specified Parameters

Create a binomial distribution object by specifying the parameter values.

```
pd = makedist('Binomial','N',30,'p',0.25)
```

```
pd =
  BinomialDistribution

  Binomial distribution
    N =   30
    p = 0.25
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 7.5000
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Binomial Distribution” on page B-10
“Bernoulli Distribution” on page B-2

Introduced in R2013a

BirnbaumSaundersDistribution

Birnbaum-Saunders probability distribution object

Description

A `BirnbaumSaundersDistribution` object consists of parameters, a model description, and sample data for a Birnbaum-Saunders probability distribution.

The Birnbaum-Saunders distribution was originally proposed as a lifetime model for materials subject to cyclic patterns of stress and strain, where the ultimate failure of the material comes from the growth of a prominent flaw. In materials science, Miner's Rule suggests that the damage occurring after n cycles, at a stress level with an expected lifetime of N cycles, is proportional to n / N. Whenever Miner's Rule applies, the Birnbaum-Saunders model is a reasonable choice for a lifetime distribution model.

The Birnbaum-Saunders distribution uses the following parameters.

Parameter	Description	Support
`beta`	scale parameter	$\beta > 0$
`gamma`	shape parameter	$\gamma > 0$

Creation

There are several ways to create a `BirnbaumSaundersDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`beta` — Scale parameter
positive scalar value

Scale parameter of the Birnbaum-Saunders distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`gamma` — Shape parameter
positive scalar value

Shape parameter of the Birnbaum-Saunders distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Birnbaum-Saunders Distribution Object Using Default Parameters

Create a Birnbaum-Saunders distribution object using the default parameter values.

```
pd = makedist('BirnbaumSaunders')
```

```
pd =
  BirnbaumSaundersDistribution

  Birnbaum-Saunders distribution
     beta = 1
    gamma = 1
```

Create a Birnbaum-Saunders Distribution Object Using Specified Parameter Values

Create a Birnbaum-Saunders distribution object by specifying the parameter values.

```
pd = makedist('BirnbaumSaunders','beta',2,'gamma',5)
```

```
pd =
  BirnbaumSaundersDistribution

  Birnbaum-Saunders distribution
     beta = 2
    gamma = 5
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 27
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Birnbaum-Saunders Distribution” on page B-18

Introduced in R2013a

BurrDistribution

Burr probability distribution object

Description

A `BurrDistribution` object consists of parameters, a model description, and sample data for a Burr probability distribution.

The Burr distribution is a three-parameter family of distributions on the positive real line. It can fit a wide range of empirical data, and is used in various fields such as finance, hydrology, and reliability to model a variety of data types.

The Burr distribution uses the following parameters.

Parameter	Description	Support
`alpha`	Scale parameter	$\alpha > 0$
`c`	First shape parameter	$c > 0$
`k`	Second shape parameter	$k > 0$

Creation

There are several ways to create a `BurrDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`alpha` — Scale parameter
positive scalar value

Scale parameter of the Burr distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`c` — First shape parameter
positive scalar value

First shape parameter of the Burr distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`k` — Second shape parameter
positive scalar value

Second shape parameter of the Burr distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution

`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters

positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates

matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters

array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values

vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function

icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Burr Distribution Object Using Default Parameters

Create a Burr distribution object using the default parameter values.

```
pd = makedist('Burr')
```

```
pd =
  BurrDistribution

  Burr distribution
    alpha = 1
        c = 1
        k = 1
```

Create a Burr Distribution Object Using Specified Parameters

Create a Burr distribution object by specifying parameter values.

```
pd = makedist('Burr','alpha',1,'c',2,'k',5)
```

```
pd =
  BurrDistribution

  Burr distribution
    alpha = 1
        c = 2
        k = 5
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 0.4295
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics
“Burr Type XII Distribution” on page B-19

Introduced in R2013a

ExponentialDistribution

Exponential probability distribution object

Description

An `ExponentialDistribution` object consists of parameters, a model description, and sample data for an exponential probability distribution.

The exponential distribution is used to model events that occur randomly over time, and its main application area is studies of lifetimes. It is a special case of the gamma distribution with the shape parameter $a = 1$.

The exponential distribution uses the following parameters.

Parameter	Description	Support
`mu`	Mean	$\mu > 0$

Creation

There are several ways to create a `ExponentialDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameter

`mu` — Mean
positive scalar value

Mean of the exponential distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create an Exponential Distribution Object Using Default Parameters

Create an exponential distribution object using the default parameter values.

```
pd = makedist('Exponential')
```

```
pd =
  ExponentialDistribution

  Exponential distribution
    mu = 1
```

Create an Exponential Distribution Object Using Specified Parameters

Create an exponential distribution object by specifying the parameter values.

```
pd = makedist('Exponential','mu',2)
```

```
pd =
  ExponentialDistribution

  Exponential distribution
    mu = 2
```

Compute the variance of the distribution.

```
v = var(pd)
```

```
v = 4
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see “Code Generation” on page 33-2252 of `fitdist`.
- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Probability Distribution Objects” on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics
“Exponential Distribution” on page B-33

Introduced in R2013a

ExtremeValueDistribution

Extreme value probability distribution object

Description

An `ExtremeValueDistribution` object consists of parameters, a model description, and sample data for an extreme value probability distribution.

The extreme value distribution is appropriate for modeling the smallest value from a distribution whose tails decay exponentially fast, for example, the normal distribution. It can also model the largest value from a distribution, such as the normal or exponential distributions, by using the negative of the original values.

The extreme value distribution uses the following parameters.

Parameter	Description	Support
`mu`	Location parameter	$-\infty < \mu < \infty$
`sigma`	Scale parameter	$\sigma \geq 0$

Creation

There are several ways to create a `ExtremeValueDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Location parameter
scalar value

Location parameter of the extreme value distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter
nonnegative scalar value

Scale parameter of the extreme value distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create an Extreme Value Distribution Object Using Default Parameters

Create an extreme value distribution object using the default parameter values.

```
pd = makedist('ExtremeValue')
```

```
pd =
  ExtremeValueDistribution

  Extreme Value distribution
       mu = 0
    sigma = 1
```

Create an Extreme Value Distribution Object Using Specified Parameters

Create an extreme value distribution object by specifying the parameter values.

```
pd = makedist('ExtremeValue','mu',-1,'sigma',2)
```

```
pd =
  ExtremeValueDistribution

  Extreme Value distribution
       mu = -1
    sigma =  2
```

Compute the standard deviation for the distribution.

```
s = std(pd)
```

```
s = 2.5651
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see “Code Generation” on page 33-2252 of `fitdist`.

- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Probability Distribution Objects” on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Extreme Value Distribution” on page B-40

Introduced in R2013a

GammaDistribution

Gamma probability distribution object

Description

A `GammaDistribution` object consists of parameters, a model description, and sample data for a gamma probability distribution.

The gamma distribution is a two-parameter family of distributions used to model sums of exponentially distributed random variables. The chi-square and the exponential distributions, which are special cases of the gamma distribution, are one-parameter distributions that fix one of the two gamma parameters.

The gamma distribution uses the following parameters.

Parameter	Description	Support
a	Shape parameter	$a > 0$
b	Scale parameter	$b > 0$

Creation

There are several ways to create a `GammaDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

a — Shape parameter
positive scalar value

Shape parameter for the gamma distribution, specified as a positive scalar value.

Data Types: `single` | `double`

b — Scale parameter
nonnegative scalar value

Scale parameter for the gamma distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name

character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting

structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions

cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names

cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Gamma Distribution Object Using Default Parameters

Create a gamma distribution object using the default parameter values.

```
pd = makedist('Gamma')
```

```
pd =
  GammaDistribution

  Gamma distribution
    a = 1
    b = 1
```

Create a Gamma Distribution Object Using Specified Parameters

Create a gamma distribution object by specifying the parameter values.

```
pd = makedist('Gamma','a',2,'b',4)
```

```
pd =
  GammaDistribution

  Gamma distribution
    a = 2
    b = 4
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 8
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Gamma Distribution” on page B-47

Introduced in R2013a

GeneralizedExtremeValueDistribution

Generalized extreme value probability distribution object

Description

A `GeneralizedExtremeValueDistribution` object consists of parameters, a model description, and sample data for a generalized extreme value probability distribution.

The generalized extreme value distribution is often used to model the smallest or largest value among a large set of independent, identically distributed random values representing measurements or observations. It combines three simpler distributions into a single form, allowing a continuous range of possible shapes that include all three of the simpler distributions.

The three distribution types correspond to the limiting distribution of block maxima from different classes of underlying distributions:

- Type 1 — Distributions whose tails decrease exponentially, such as the normal distribution
- Type 2 — Distributions whose tails decrease as a polynomial, such as Student's t distribution
- Type 3 — Distributions whose tails are finite, such as the beta distribution

The generalized extreme value distribution uses the following parameters.

Parameter	Description	Support
`k`	Shape parameter	$-\infty \le k \le \infty$
`sigma`	Scale parameter	$\sigma \ge 0$
`mu`	Location parameter	$-\infty \le \mu \le \infty$

Creation

There are several ways to create a `GeneralizedExtremeValueDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

k — Shape parameter
scalar value

Shape parameter of the generalized extreme value distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter
nonnegative scalar value

Scale parameter of the generalized extreme value distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

`mu` — Location parameter
scalar value

Location parameter of the generalized extreme value distribution, specified as a scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Generalized Extreme Value Distribution Object Using Default Parameters

Create a generalized extreme value distribution object using the default parameter values.

```
pd = makedist('GeneralizedExtremeValue')
```

```
pd =
  GeneralizedExtremeValueDistribution

  Generalized Extreme Value distribution
        k = 0
    sigma = 1
       mu = 0
```

Create a Generalized Extreme Value Distribution Object Using Specified Parameters

Create a generalized extreme value distribution object by specifying values for the parameters.

```
pd = makedist('GeneralizedExtremeValue','k',0,'sigma',2,'mu',1)
```

```
pd =
  GeneralizedExtremeValueDistribution

  Generalized Extreme Value distribution
        k = 0
    sigma = 2
       mu = 1
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 2.1544
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Generalized Extreme Value Distribution” on page B-55

Introduced in R2013a

GeneralizedParetoDistribution

Generalized Pareto probability distribution object

Description

A `GeneralizedParetoDistribution` object consists of parameters, a model description, and sample data for a generalized Pareto probability distribution.

The generalized Pareto distribution is used to model the tails of another distribution. It allows a continuous range of possible shapes that include both the exponential and Pareto distributions as special cases. It has three basic forms, each corresponding to a limiting distribution of exceedance data from a different class of underlying distributions.

- Distributions whose tails decrease exponentially, such as the normal, lead to a generalized Pareto shape parameter of zero.
- Distributions whose tails decrease polynomially, such as the Student's *t*, lead to a positive shape parameter.
- Distributions whose tails are finite, such as the beta, lead to a negative shape parameter.

The generalized Pareto distribution uses the following parameters.

Parameter	Description	Support
`k`	Shape parameter	$-\infty < k < \infty$
`sigma`	Scale parameter	$\sigma \geq 0$
`theta`	Location parameter	$-\infty < \theta < \infty$

Creation

There are several ways to create a `GeneralizedParetoDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

k — Shape parameter
scalar value

Shape parameter for the generalized Pareto distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter
nonnegative scalar value

Scale parameter for the generalized Pareto distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

`theta` — Location parameter
scalar value

Location parameter for the generalized Pareto distribution, specified as a scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Generalized Pareto Distribution Object Using Default Parameters

Create a generalized Pareto distribution object using the default parameter values.

```
pd = makedist('GeneralizedPareto')
```

```
pd =
  GeneralizedParetoDistribution

  Generalized Pareto distribution
        k = 1
    sigma = 1
    theta = 1
```

Create a Generalized Pareto Distribution Object Using Specified Parameters

Create a generalized Pareto distribution object by specifying parameter values.

```
pd = makedist('GeneralizedPareto','k',0,'sigma',2,'theta',1)
```

```
pd =
  GeneralizedParetoDistribution

  Generalized Pareto distribution
        k = 0
    sigma = 2
    theta = 1
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 3
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Generalized Pareto Distribution” on page B-59
“Nonparametric and Empirical Probability Distributions” on page 5-30

Introduced in R2013a

HalfNormalDistribution

Half-normal probability distribution object

Description

A `HalfNormalDistribution` object consists of parameters, a model description, and sample data for a half-normal probability distribution.

The half-normal distribution is a special case of the folded normal and truncated normal distribution. Applications of the half-normal distribution include modeling measurement data and lifetime data.

The half-normal distribution uses the following parameters:

Parameter	Description	Support
`mu`	Location	$-\infty < \mu < \infty$
`sigma`	Scale	$\sigma \geq 0$

For more information about the half-normal distribution, see “Half-Normal Distribution” on page B-68.

Creation

There are several ways to create a `HalfNormalDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Location parameter
scalar value

Location parameter of the half-normal distribution, specified as a scalar value. The `mu` parameter is also the lower limit of the half-normal distribution.

The Statistics and Machine Learning Toolbox implementation of the half-normal distribution assumes a fixed value for the location parameter μ. You can specify a value for the μ parameter when creating a `HalfNormalDistribution` object.

Data Types: `single` | `double`

`sigma` — Scale parameter
nonnegative scalar value

Scale parameter of the half-normal distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution

`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters

positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates

matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters

array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values

vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval

vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution

median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Half-Normal Distribution Object Using Default Parameters

```
pd = makedist('HalfNormal')
```

```
pd =
  HalfNormalDistribution

  Half Normal distribution
       mu = 0
    sigma = 1
```

Create a Half-Normal Distribution Object Using Specified Parameters

Create a half-normal distribution object. Specify `mu` equal to 0 and `sigma` equal to 1.5.

```
pd = makedist('HalfNormal','mu',0,'sigma',1.5)
```

```
pd =
  HalfNormalDistribution

  Half Normal distribution
       mu =   0
    sigma = 1.5
```

Compute the mean and standard deviation of the distribution.

```
m = mean(pd)
```

```
m = 1.1968
```

```
s = std(pd)
```

```
s = 0.9042
```

Fit a Half-Normal Distribution Object

Generate 100 random numbers from a standard normal distribution and compute their absolute value.

```
rng default  % For reproducibility
x = abs(random(makedist('Normal'),100,1));
```

Fit a half-normal distribution object to the sample data.

```
pd = fitdist(x,'HalfNormal')
```

```
pd =
  HalfNormalDistribution

  Half Normal distribution
       mu =      0
    sigma = 1.1631   [1.02184, 1.35006]
```

Calculate the mean of the fitted half-normal distribution using the probability distribution object.

```
m = mean(pd)
```

```
m = 0.9280
```

Calculate the mean of the half-normal distribution by substituting the fitted `mu` and `sigma` parameter values into the formula

$$mean = \mu + \sigma\sqrt{\frac{2}{\pi}}.$$

```
mcalc = pd.mu + pd.sigma*(sqrt(2/pi))
```

```
mcalc = 0.9280
```

References

[1] Cooray, K. and M.M.A. Ananda. “A Generalization of the Half-Normal Distribution with Applications to Lifetime Data.” *Communications in Statistics – Theory and Methods*. Vol. 37, Number 9, 2008, pp. 1323–1337.

[2] Pewsey, A. "Large-Sample Inference for the General Half-Normal Distribution." *Communications in Statistics – Theory and Methods*. Vol. 31, Number 7, 2002, pp. 1045–1054.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Half-Normal Distribution” on page B-68

Introduced in R2016a

InverseGaussianDistribution

Inverse Gaussian probability distribution object

Description

An `InverseGaussianDistribution` object consists of parameters, a model description, and sample data for an inverse Gaussian probability distribution.

Also known as the Wald distribution, the inverse Gaussian is used to model nonnegative positively skewed data. Inverse Gaussian distributions have many similarities to standard Gaussian (normal) distributions, which lead to applications in inferential statistics.

The inverse Gaussian distribution uses the following parameters.

Parameter	Description	Support
`mu`	Scale parameter	$\mu > 0$
`lambda`	Shape parameter	$\lambda > 0$

Creation

There are several ways to create a `InverseGaussianDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Scale parameter
positive scalar value

Scale parameter for the inverse Gaussian distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`lambda` — Shape parameter
positive scalar value

Shape parameter for the inverse Gaussian distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create an Inverse Gaussian Distribution Object Using Default Parameters

Create an inverse Gaussian distribution object using the default parameter values.

```
pd = makedist('InverseGaussian')
```

```
pd =
  InverseGaussianDistribution

  Inverse Gaussian distribution
        mu = 1
    lambda = 1
```

Create an Inverse Gaussian Distribution Object Using Specified Parameters

Create an inverse Gaussian distribution object by specifying parameter values.

```
pd = makedist('InverseGaussian','mu',2,'lambda',4)
```

```
pd =
  InverseGaussianDistribution

  Inverse Gaussian distribution
        mu = 2
    lambda = 4
```

Compute the standard deviation of the distribution.

```
s = std(pd)
```

```
s = 1.4142
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Inverse Gaussian Distribution” on page B-75

Introduced in R2013a

KernelDistribution

Kernel probability distribution object

Description

A `KernelDistribution` object consists of parameters, a model description, and sample data for a nonparametric kernel-smoothing distribution.

The kernel distribution is a nonparametric estimation of the probability density function (pdf) of a random variable.

The kernel distribution uses the following options.

Option	Description	Possible Values
`Kernel`	Kernel function type	`normal`, `box`, `triangle`, `epanechnikov`
`BandWidth`	Kernel smoothing parameter	`BandWidth > 0`

Creation

There are several ways to create a `KernelDistribution` probability distribution object.

- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`Kernel` — Kernel smoother type
`'normal'` | `'box'` | `'triangle'` | `'epanechnikov'`

Kernel function type, specified as a valid kernel function type name.

`BandWidth` — Bandwidth of kernel smoothing window
positive scalar value

Bandwidth of the kernel smoothing window, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

Object Functions

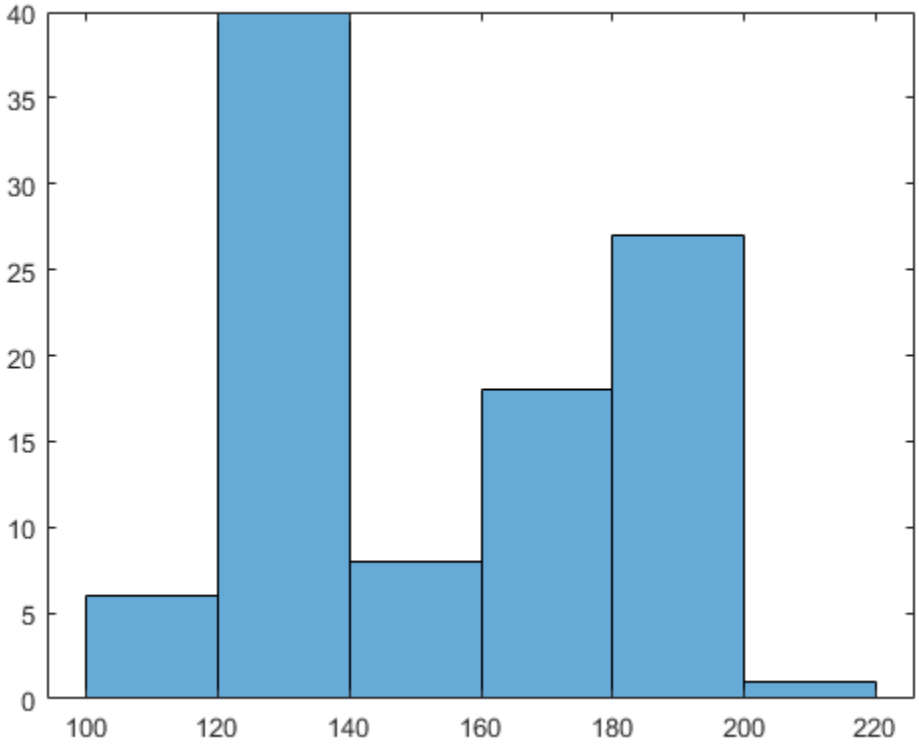
cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
pdf	Probability density function
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Fit a Kernel Distribution Object to Data

Load the sample data. Visualize the patient weight data using a histogram.

```
load hospital
histogram(hospital.Weight)
```



The histogram shows that the data has two modes, one for female patients and one for male patients.

Create a probability distribution object by fitting a kernel distribution to the patient weight data.

```
pd_kernel = fitdist(hospital.Weight,'Kernel')
```

```
pd_kernel =
  KernelDistribution

    Kernel = normal
    Bandwidth = 14.3792
    Support = unbounded
```

For comparison, create another probability distribution object by fitting a normal distribution to the patient weight data.

```
pd_normal = fitdist(hospital.Weight,'Normal')
```

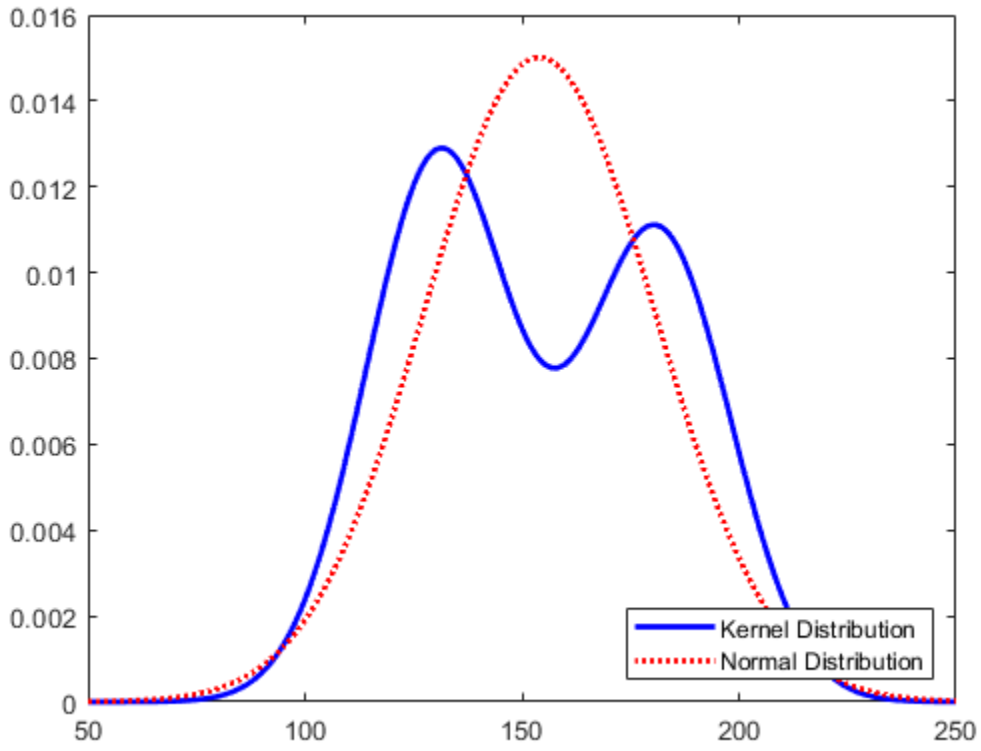
```
pd_normal =
  NormalDistribution

  Normal distribution
       mu =     154   [148.728, 159.272]
    sigma = 26.5714   [23.3299, 30.8674]
```

Define the x values and compute the pdf of each distribution.

```
x = 50:1:250;
pdf_kernel = pdf(pd_kernel,x);
pdf_normal = pdf(pd_normal,x);
```

Plot the pdf of each distribution.

```
plot(x,pdf_kernel,'Color','b','LineWidth',2);
hold on;
plot(x,pdf_normal,'Color','r','LineStyle',':','LineWidth',2);
legend('Kernel Distribution','Normal Distribution','Location','SouthEast');
hold off;
```



Fitting a kernel distribution instead of a unimodal distribution such as the normal reveals the separate modes for the female and male patients.

See Also

Distribution Fitter | `fitdist`

Topics

“Fit Kernel Distribution Object to Data” on page 5-36
“Fit Probability Distribution Objects to Grouped Data” on page 5-92
“Compare Multiple Distribution Fits” on page 5-87
“Kernel Distribution” on page B-78
“Nonparametric and Empirical Probability Distributions” on page 5-30

Introduced in R2013a

LogisticDistribution

Logistic probability distribution object

Description

A `LogisticDistribution` object consists of parameters, a model description, and sample data for a logistic probability distribution.

The logistic distribution is used for growth models and in logistic regression. It has longer tails and a higher kurtosis than the normal distribution.

The logistic distribution uses the following parameters.

Parameter	Description	Support
`mu`	Mean	$-\infty < \mu < \infty$
`sigma`	Scale parameter	$\sigma \geq 0$

Creation

There are several ways to create a `LogisticDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Mean
scalar value

Mean of the logistic distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter
nonnegative scalar value

Scale parameter of the logistic distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Logistic Distribution Object Using Default Parameters

Create a logistic distribution object using the default parameter values.

```
pd = makedist('Logistic')
```

```
pd =
  LogisticDistribution

  Logistic distribution
       mu = 0
    sigma = 1
```

Create a Logistic Distribution Object Using Specified Parameters

Create a logistic distribution object by specifying parameter values.

```
pd = makedist('Logistic','mu',2,'sigma',4)
```

```
pd =
  LogisticDistribution

  Logistic distribution
       mu = 2
    sigma = 4
```

Compute the standard deviation of the distribution.

```
s = std(pd)
```

```
s = 7.2552
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Compare Multiple Distribution Fits” on page 5-87
“Logistic Distribution” on page B-85

Introduced in R2013a

LoglogisticDistribution

Loglogistic probability distribution object

Description

A `LoglogisticDistribution` object consists of parameters, a model description, and sample data for a loglogistic probability distribution.

The loglogistic distribution is closely related to the logistic distribution. If x is distributed loglogistically with parameters μ and σ, then $\log(x)$ is distributed logistically with mean and standard deviation. This distribution is often used in survival analysis to model events that experience an initial rate increase, followed by a rate decrease.

The loglogistic distribution uses the following parameters.

Parameter	Description	Support
`mu`	Mean of logarithmic values	$\mu > 0$
`sigma`	Scale parameter of logarithmic values	$\sigma > 0$

Creation

There are several ways to create a `LoglogisticDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Mean of logarithmic values
positive scalar value

Mean of logarithmic values for the loglogistic distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter of logarithmic values
positive scalar value

Scale parameter of logarithmic values for the loglogistic distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Loglogistic Distribution Object Using Default Parameters

Create a loglogistic distribution object using the default parameter values.

```
pd = makedist('Loglogistic')
```

```
pd =
  LoglogisticDistribution

  Log-Logistic distribution
       mu = 0
    sigma = 1
```

Create a Loglogistic Distribution Object Using Specified Parameters

Create a loglogistic distribution object by specifying the parameter values.

```
pd = makedist('Loglogistic','mu',5,'sigma',2)
```

```
pd =
  LoglogisticDistribution

  Log-Logistic distribution
       mu = 5
    sigma = 2
```

Generate random numbers from the loglogistic distribution and compute their log values.

```
rng(19) % for reproducibility
x = random(pd,10000,1);
logx = log(x);
```

Compute the mean of the log values.

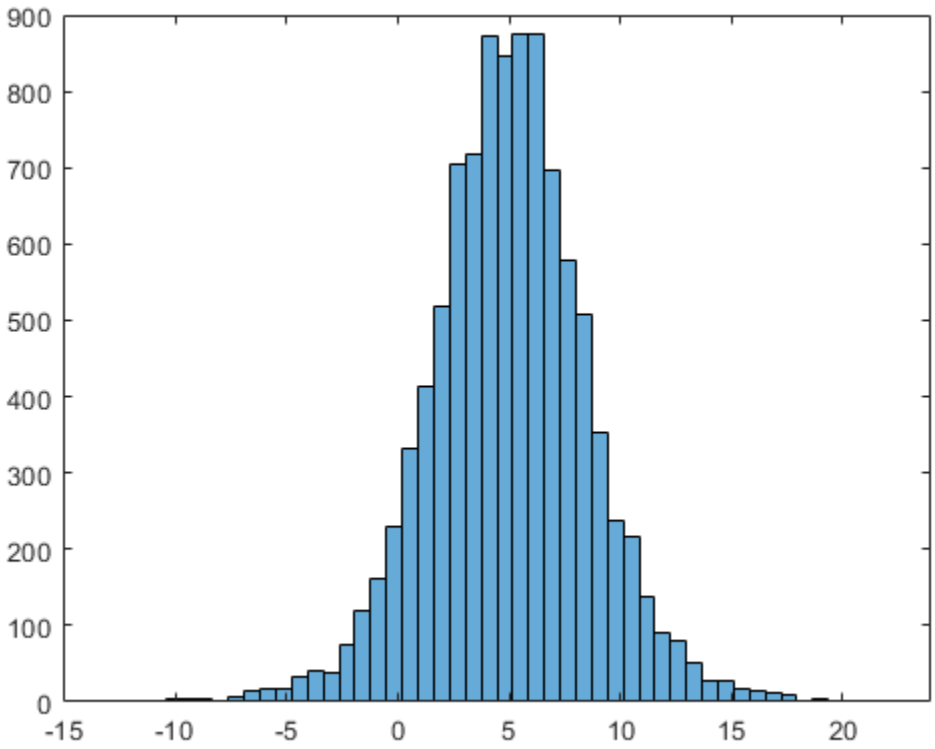
```
m = mean(logx)
```

```
m = 4.9828
```

The mean of the log of `x` is equal to the `mu` parameter of `x`, since `x` has a loglogistic distribution.

Plot `logx`.

```
histogram(logx,50)
```



The plot shows that the log values of `x` have a logistic distribution.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Loglogistic Distribution” on page B-86

Introduced in R2013a

LognormalDistribution

Lognormal probability distribution object

Description

A `LognormalDistribution` object consists of parameters, a model description, and sample data for a lognormal probability distribution.

The lognormal distribution, sometimes called the Galton distribution, is a probability distribution whose logarithm has a normal distribution. The lognormal distribution is applicable when the quantity of interest must be positive, because log(x) exists only when x is positive.

The lognormal distribution uses the following parameters.

Parameter	Description	Support
`mu` (μ)	Mean of logarithmic values	$-\infty < \mu < \infty$
`sigma` (σ)	Standard deviation of logarithmic values	$\sigma \geq 0$

Creation

There are several ways to create a `LognormalDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Mean of logarithmic values
scalar value

Mean of logarithmic values for the lognormal distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Standard deviation of logarithmic values
nonnegative scalar value

Standard deviation of logarithmic values for the lognormal distribution, specified as a nonnegative scalar value.

You can specify `sigma` to be zero when you create an object by using `makedist`. Some object functions support an object `pd` with zero standard deviation. For example, `random(pd)` always returns `exp(mu)`.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution

median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Lognormal Distribution Object Using Default Parameters

Create a lognormal distribution object using the default parameter values.

```
pd = makedist('Lognormal')
```

```
pd =
  LognormalDistribution

  Lognormal distribution
       mu = 0
    sigma = 1
```

Create Lognormal Distribution Object Using Specified Parameters

Create a lognormal distribution object by specifying the parameter values.

```
pd = makedist('Lognormal','mu',5,'sigma',2)
```

```
pd =
  LognormalDistribution

  Lognormal distribution
       mu = 5
    sigma = 2
```

Compute the mean of the lognormal distribution.

```
mean(pd)
```

```
ans = 1.0966e+03
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see "Code Generation" on page 33-2252 of `fitdist`.
- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "Code Generation for Probability Distribution Objects" on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

"Lognormal Distribution" on page B-88

Introduced in R2013a

MultinomialDistribution

Multinomial probability distribution object

Description

A `MultinomialDistribution` object consists of parameters and a model description for a multinomial probability distribution.

The multinomial distribution is a generalization of the binomial distribution. While the binomial distribution gives the probability of the number of "successes" in n independent trials of a two-outcome process, the multinomial distribution gives the probability of each combination of outcomes in n independent trials of a k-outcome process. The probability of each outcome in any one trial is given by the fixed probabilities $p_1, \ldots, p_k$.

The multinomial distribution uses the following parameters.

Parameter	Description	Support
`probabilities`	Outcome probabilities	$0 \le \text{probabilities}(i) \le 1$; $\sum_{\text{all}(i)} \text{probabilities}(i) = 1$

Creation

Create a `MultinomialDistribution` probability distribution with specified parameter values object using `makedist`.

Properties

Distribution Parameter

`Probabilities` — outcome probabilities
vector of scalar values in the range `[0,1]`

Outcome probabilities for the multinomial distribution, stored as a vector of scalar values in the range `[0,1]`. The values in `probabilities` must sum to 1.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
pdf	Probability density function
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Multinomial Distribution Object Using Default Parameters

Create a multinomial distribution object using the default parameter values.

```
pd = makedist('Multinomial')
```

```
pd =
  MultinomialDistribution

  Probabilities:
    0.5000    0.5000
```

Create a Multinomial Distribution Object Using Specified Parameters

Create a multinomial distribution object for a distribution with three possible outcomes. Outcome 1 has a probability of 1/2, outcome 2 has a probability of 1/3, and outcome 3 has a probability of 1/6.

```
pd = makedist('Multinomial','probabilities',[1/2 1/3 1/6])
```

```
pd =
  MultinomialDistribution

  Probabilities:
    0.5000    0.3333    0.1667
```

Generate a random outcome from the distribution.

```
rng('default');  % for reproducibility
r = random(pd)
```

```
r = 2
```

The result of this trial is outcome 2. By default, the number of trials in each experiment, *n*, equals 1.

Generate random outcomes from the distribution when the number of trials in each experiment, *n*, equals 1, and the experiment is repeated ten times.

```
rng('default');  % for reproducibility
r = random(pd,10,1)

r = 10×1

     2
     3
     1
     3
     2
     1
     1
     2
     3
     3
```

Each element in the array is the outcome of an individual experiment that contains one trial.

Generate random outcomes from the distribution when the number of trials in each experiment, *n*, equals 5, and the experiment is repeated ten times.

```
rng('default');  % for reproducibility
r = random(pd,10,5)

r = 10×5

     2     1     2     2     1
     3     3     1     1     1
     1     3     3     1     2
     3     1     3     1     2
     2     2     2     1     1
     1     1     2     2     1
     1     1     2     2     1
     2     3     1     1     2
     3     2     2     3     2
     3     3     1     1     2
```

Each element in the resulting matrix is the outcome of one trial. The columns correspond to the five trials in each experiment, and the rows correspond to the ten experiments. For example, in the first experiment (corresponding to the first row), 2 of the 5 trials resulted in outcome 1, and 3 of the 5 trials resulted in outcome 2.

See Also

`makedist`

Topics

“Multinomial Probability Distribution Objects” on page 5-95
“Multinomial Probability Distribution Functions” on page 5-98
“Multinomial Distribution” on page B-96

Introduced in R2013a

NakagamiDistribution

Nakagami probability distribution object

Description

A `NakagamiDistribution` object consists of parameters, a model description, and sample data for a Nakagami probability distribution.

The Nakagami distribution is commonly used in communication theory to model scattered signals that reach a receiver using multiple paths.

The Nakagami distribution uses the following parameters.

Parameter	Description	Support
`mu`	Shape parameter	$\mu > 0$
`omega`	Scale parameter	$\omega > 0$

Creation

There are several ways to create a `NakagamiDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Shape parameter
positive scalar value

Shape parameter for the Nakagami distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`omega` — Scale parameter
positive scalar value

Scale parameter for the Nakagami distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Nakagami Distribution Object Using Default Parameters

Create a Nakagami distribution object using the default parameter values.

```
pd = makedist('Nakagami')
```

```
pd =
  NakagamiDistribution

  Nakagami distribution
       mu = 1
    omega = 1
```

Create a Nakagami Distribution Object Using Specified Parameters

Create a Nakagami distribution object by specifying parameter values.

```
pd = makedist('Nakagami','mu',5,'omega',2)
```

```
pd =
  NakagamiDistribution

  Nakagami distribution
       mu = 5
    omega = 2
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 1.3794
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Nakagami Distribution” on page B-108

Introduced in R2013a

NegativeBinomialDistribution

Negative binomial distribution object

Description

A `NegativeBinomialDistribution` object consists of parameters, a model description, and sample data for a negative binomial probability distribution.

The negative binomial distribution models the number of failures x before a specified number of successes, R, is reached in a series of independent, identical trials. This distribution can also model count data, in which case R does not need to be an integer value.

The negative binomial distribution uses the following parameters.

Parameter	Description	Support
R	Number of successes	$r > 0$
p	Probability of success	$0 < p \le 1$

Creation

There are several ways to create a `NegativeBinomialDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

R — Number of successes
positive scalar value

Number of successes for the negative binomial distribution, specified as a positive scalar value.

Data Types: `single` | `double`

p — Probability of success
positive scalar value in the range (0,1]

Probability of success of any individual trial for the negative binomial distribution, specified as a positive scalar value in the range (0,1].

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution

`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters

positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates

matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters

array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values

vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval

vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution

paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Negative Binomial Distribution Object Using Default Parameters

Create a negative binomial distribution object using the default parameter values.

```
pd = makedist('NegativeBinomial')
```

```
pd =
  NegativeBinomialDistribution

  Negative Binomial distribution
    R =   1
    P = 0.5
```

Create a Negative Binomial Distribution Object Using Specified Parameters

Create a negative binomial distribution object by specifying the parameter values.

```
pd = makedist('NegativeBinomial','R',5,'p',.1)
```

```
pd =
  NegativeBinomialDistribution

  Negative Binomial distribution
    R =   5
    P = 0.1
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 45
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Negative Binomial Distribution” on page B-109

Introduced in R2013a

NormalDistribution

Normal probability distribution object

Description

A `NormalDistribution` object consists of parameters, a model description, and sample data for a normal probability distribution.

The normal distribution, sometimes called the Gaussian distribution, is a two-parameter family of curves. The usual justification for using the normal distribution for modeling is the Central Limit theorem, which states (roughly) that the sum of independent samples from any distribution with finite mean and variance converges to the normal distribution as the sample size goes to infinity.

The normal distribution uses the following parameters.

Parameter	Description	Support
`mu` (μ)	Mean	$-\infty < \mu < \infty$
`sigma` (σ)	Standard deviation	$\sigma \geq 0$

Creation

There are several ways to create a `NormalDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Mean
scalar value

Mean of the normal distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Standard deviation
nonnegative scalar value

Standard deviation of the normal distribution, specified as a nonnegative scalar value.

You can specify `sigma` to be zero when you create an object by using `makedist`. Some object functions support an object `pd` with zero standard deviation. For example, `random(pd)` always returns `mu`, and `cdf(pd,x)` returns either 0 or 1. The output is 0 if `x` is smaller than `mu`, and 1 otherwise. `mean`, `std`, and `var` return the mean, standard deviation, and variance of `pd`, respectively.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution

`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters

positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates

matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters

array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values

vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval

vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution

median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Normal Distribution Object Using Default Parameters

Create a normal distribution object using the default parameter values.

```
pd = makedist('Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Create a Normal Distribution Object Using Specified Parameters

Create a normal distribution object by specifying the parameter values.

```
pd = makedist('Normal','mu',75,'sigma',10)
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75
    sigma = 10
```

Compute the interquartile range of the distribution.

```
r = iqr(pd)
```

```
r = 13.4898
```

Fit Normal Distribution Object

Load the sample data and create a vector containing the first column of student exam grade data.

```
load examgrades
x = grades(:,1);
```

Create a normal distribution object by fitting it to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see "Code Generation" on page 33-2252 of `fitdist`.
- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "Code Generation for Probability Distribution Objects" on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

"Compare Multiple Distribution Fits" on page 5-87
"Normal Distribution" on page B-119

Introduced in R2013a

PiecewiseLinearDistribution

Piecewise linear probability distribution object

Description

A `PiecewiseLinearDistribution` object consists of a model description for a piecewise linear probability distribution.

The piecewise linear distribution is a nonparametric probability distribution created using a piecewise linear representation of the cumulative distribution function (cdf). The options specified for the piecewise linear distribution specify the form of the cdf. The probability density function (pdf) is a step function.

The piecewise linear distribution uses the following parameters.

Parameter	Description
`x`	Vector of *x* values at which the cdf changes slope
`Fx`	Vector of cdf values that correspond to each value in `x`

Creation

Create a `PiecewiseLinearDistribution` probability distribution with specified parameter values object using `makedist`.

Properties

Distribution Parameters

x — Data values
vector of scalar values

Data values at which the cumulative distribution function (cdf) changes slope, specified as a vector of scalar values.

Data Types: `single` | `double`

Fx — cdf value
vector of scalar values

cdf value at each value in `x`, specified as a vector of scalar values.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
pdf	Probability density function
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Piecewise Linear Distribution Object Using Default Parameters

Create a piecewise linear distribution object using the default parameter values.

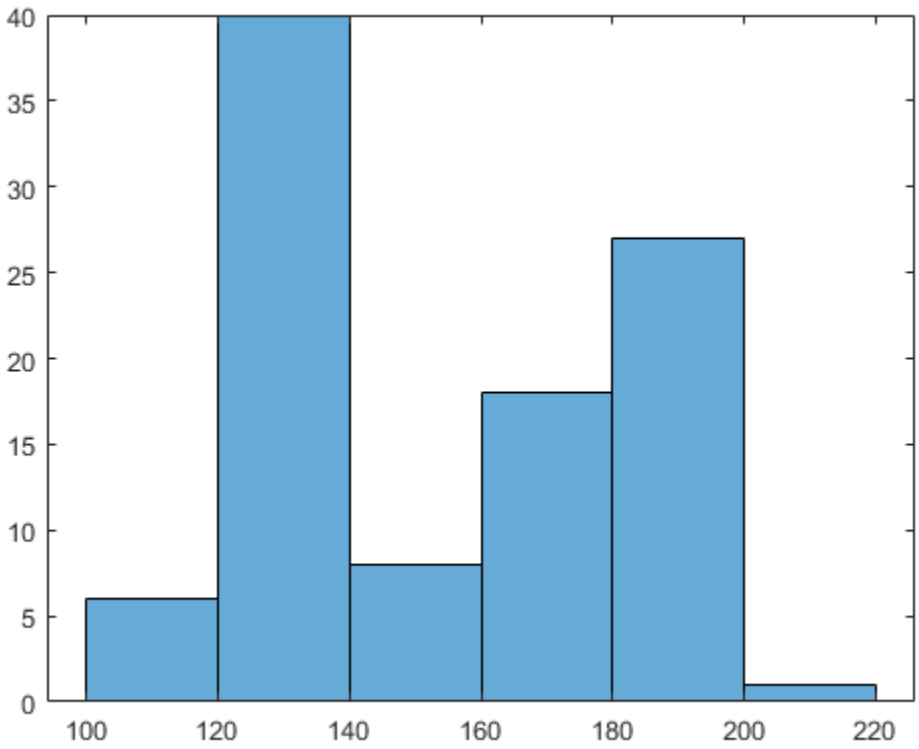
```
pd = makedist('PiecewiseLinear')
```

```
pd =
  PiecewiseLinearDistribution

F(0) = 0
F(1) = 1
```

Create a Piecewise Linear Distribution Object Using Specified Parameters

Load the sample data. Visualize the patient weight data using a histogram.

```
load hospital
histogram(hospital.Weight)
```



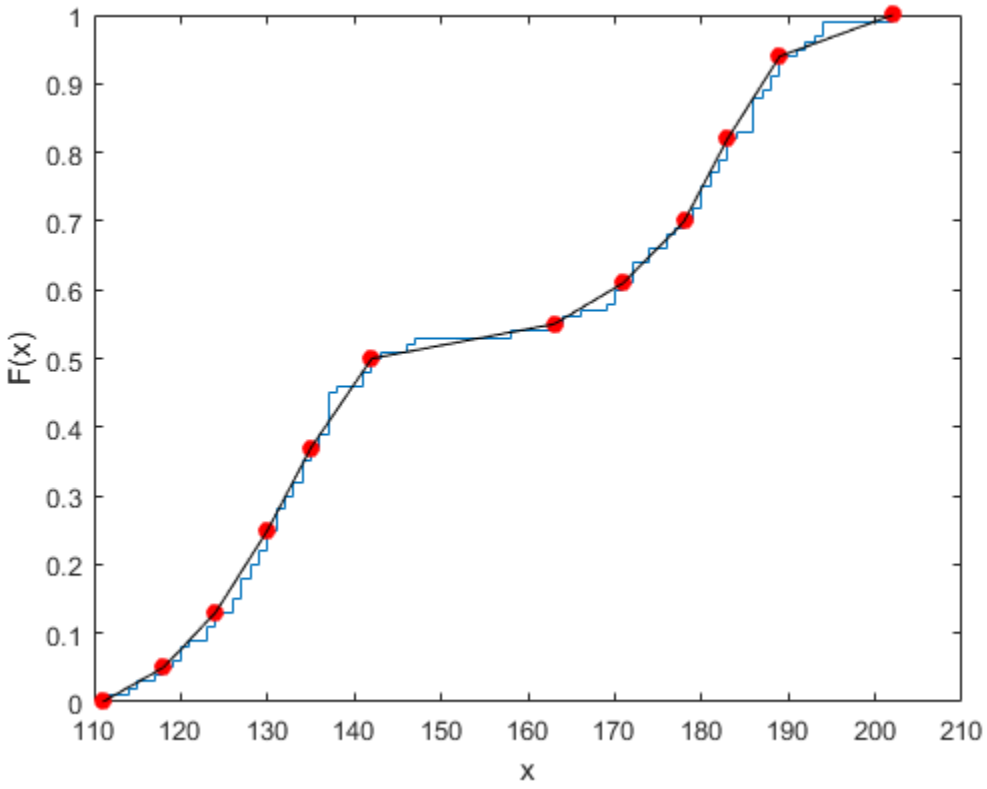
The histogram shows that the data has two modes, one for female patients and one for male patients.

Compute the empirical cumulative distribution function (ecdf) for the data.

```
[f,x] = ecdf(hospital.Weight);
```

Construct a piecewise linear approximation to the ecdf and plot both functions.

```
f = f(1:5:end); % keep a less dense grid of points
x = x(1:5:end);

figure;
ecdf(hospital.Weight)
hold on
plot(x,f,'ro','MarkerFace','r') % overlay grid
plot(x,f,'k') % show interpolation
```



Create a piecewise linear probability distribution object using the piecewise approximation of the ecdf.

```
pd = makedist('PiecewiseLinear','x',x,'Fx',f)
```

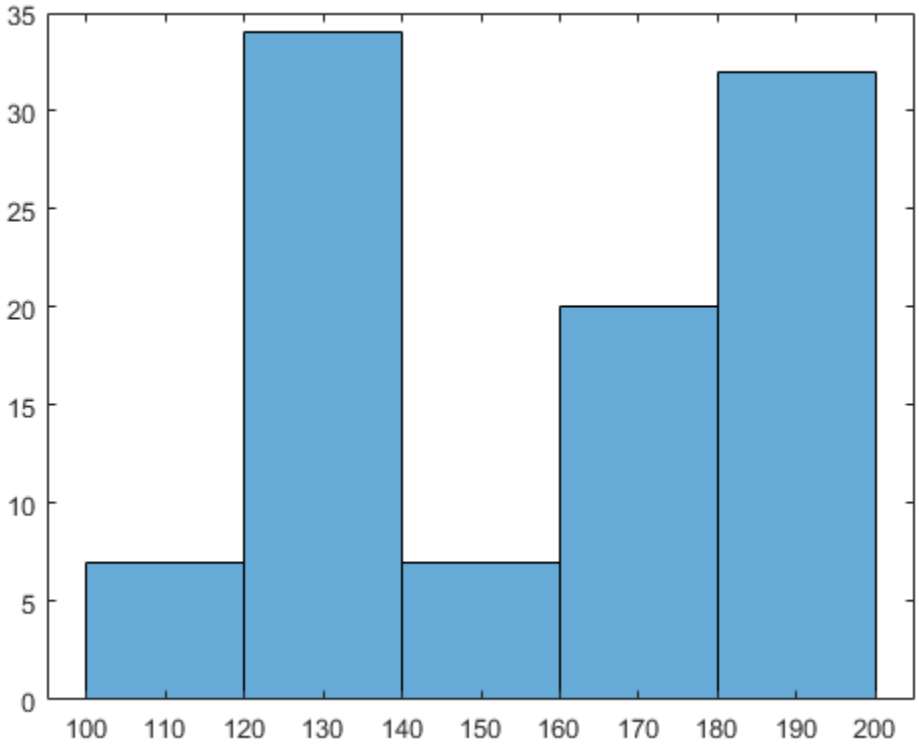
```
pd =
  PiecewiseLinearDistribution

F(111) = 0
F(118) = 0.05
F(124) = 0.13
F(130) = 0.25
F(135) = 0.37
F(142) = 0.5
F(163) = 0.55
F(171) = 0.61
F(178) = 0.7
F(183) = 0.82
F(189) = 0.94
F(202) = 1
```

Generate 100 random numbers from the distribution.

```
rw = random(pd,100,1);
```

Plot the random numbers to visually compare their distribution to the original data.

```
figure;
histogram(rw)
```



The random numbers generated from the piecewise linear distribution have the same bimodal distribution as the original data.

See Also

`makedist`

Topics

“Nonparametric and Empirical Probability Distributions” on page 5-30
“Piecewise Linear Distribution” on page B-130

Introduced in R2013a

PoissonDistribution

Poisson probability distribution object

Description

A `PoissonDistribution` object consists of parameters, a model description, and sample data for a Poisson probability distribution.

The Poisson distribution is appropriate for applications that involve counting the number of times a random event occurs in a given amount of time, distance, area, etc. If the number of counts follows the Poisson distribution, then the interval between individual counts follows the exponential distribution.

The Poisson distribution uses the following parameters.

Parameter	Description	Support
`lambda`	Mean	$\lambda \geq 0$

Creation

There are several ways to create a `PoissonDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameter

`lambda` — Mean
nonnegative scalar value

Mean of the Poisson distribution, stored as a nonnegative scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Poisson Distribution Object Using Default Parameters

Create a Poisson distribution object using the default parameter values.

```
pd = makedist('Poisson')
```

```
pd =
  PoissonDistribution

  Poisson distribution
    lambda = 1
```

Create a Poisson Distribution Object Using Specified Parameters

Create a Poisson distribution object by specifying the parameter values.

```
pd = makedist('Poisson','lambda',5)
```

```
pd =
  PoissonDistribution

  Poisson distribution
    lambda = 5
```

Compute the variance of the distribution.

```
v = var(pd)
```

```
v = 5
```

For the Poisson distribution, both the mean and variance are equal to the parameter lambda.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Poisson Distribution” on page B-131

Introduced in R2013a

RayleighDistribution

Rayleigh probability distribution object

Description

A `RayleighDistribution` object consists of parameters, a model description, and sample data for a normal probability distribution.

The Rayleigh distribution is a special case of the Weibull distribution. It is often used in communication theory to model scattered signals that reach a receiver by multiple paths.

The Rayleigh distribution uses the following parameter.

Parameter	Description	Support
`b`	Defining parameter	$b > 0$

Creation

There are several ways to create a `RayleighDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameter

`b` — Defining parameter
positive scalar value

Defining parameter for the Rayleigh distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Rayleigh Distribution Object Using Default Parameters

Create a Rayleigh distribution object using the default parameter values.

```
pd = makedist('Rayleigh')

pd =
  RayleighDistribution

  Rayleigh distribution
    B = 1
```

Create a Rayleigh Distribution Object Using Specified Parameters

Create a Rayleigh distribution object by specifying the parameter values.

```
pd = makedist('Rayleigh','b',3)

pd =
  RayleighDistribution

  Rayleigh distribution
    B = 3
```

Compute the mean of the distribution.

```
m = mean(pd)

m = 3.7599
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Rayleigh Distribution” on page B-137

Introduced in R2013a

RicianDistribution

Rian probability distribution object

Description

A `RicianDistribution` object consists of parameters, a model description, and sample data for a Rician probability distribution.

The Rician distribution is used in communications theory to model scattered signals that reach a receiver using multiple paths.

The Rician distribution uses the following parameters.

Name	Description	Support
`s`	Noncentrality parameter	$s \geq 0$
`sigma`	Scale parameter	$\sigma > 0$

Creation

There are several ways to create a `RicianDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`s` — Noncentrality parameter
nonnegative scalar value

Noncentrality parameter of the Rician distribution, specified as a nonnegative scalar value.

Data Types: `single` | `double`

`sigma` — scale parameter
positive scalar value

Scale parameter for the Rician distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Rician Distribution Object Using Default Parameters

Create a Rician distribution object using the default parameter values.

```
pd = makedist('Rician')
```

```
pd =
  RicianDistribution

  Rician distribution
        s = 1
    sigma = 1
```

Create a Rician Distribution Object Using Specified Parameters

Create a Rician distribution object by specifying the parameter values.

```
pd = makedist('Rician','s',0,'sigma',2)
```

```
pd =
  RicianDistribution

  Rician distribution
        s = 0
    sigma = 2
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 2.5066
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Rician Distribution” on page B-139

Introduced in R2013a

StableDistribution

Stable probability distribution object

Description

`StableDistribution` is an object consisting of parameters, a model description, and sample data for a stable probability distribution.

The stable distribution uses the following parameters.

Parameter	Description	Support
`alpha`	First shape parameter	$0 < \alpha \leq 2$
`beta`	Second shape parameter	$-1 \leq \beta \leq 1$
`gam`	Scale parameter	$0 < \gamma < \infty$
`delta`	Location parameter	$-\infty < \delta < \infty$

Creation

There are several ways to create a `StableDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`alpha` — First shape parameter
scalar value in the range (0,2]

First shape parameter of the stable distribution, specified as a scalar value in the range (0,2].

Data Types: `single` | `double`

`beta` — Second shape parameter
scalar value in the range [-1,1]

Second shape parameter of the stable distribution, specified as a scalar value in the range [-1,1].

Data Types: `single` | `double`

`gam` — Scale parameter
scalar value in the range (0,∞)

Scale parameter of the stable distribution, specified as a scalar value in the range (0,∞).

Data Types: `single` | `double`

`delta` — Location parameter
scalar value

Location parameter of the stable distribution, specified as a scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Stable Distribution Object Using Default Parameters

Create a stable distribution object using the default parameter values.

```
pd = makedist('Stable')
```

```
pd =
  StableDistribution

  Stable distribution
    alpha = 2
     beta = 0
      gam = 1
    delta = 0
```

Create a Stable Distribution Object Using Specified Parameters

Create a stable distribution object by specifying the parameter values `alpha = 0.5`, `beta = 0`, `gam = 1`, and `delta = 0`.

```
pd = makedist('Stable','alpha',0.5,'beta',0,'gam',1,'delta',0);
```

Calculate the mean of the distribution.

```
m = mean(pd)
```

```
m = NaN
```

The mean of the stable distribution is undefined for values of `alpha` less than or equal to 1.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics
“Stable Distribution” on page B-140

Introduced in R2016a

tLocationScaleDistribution

t Location-Scale probability distribution object

Description

A `tLocationScaleDistribution` object consists of parameters, a model description, and sample data for a *t* location-scale probability distribution.

The *t* location-scale distribution is useful for modeling data distributions with heavier tails (more prone to outliers) than the normal distribution. It approaches the normal distribution as ν approaches infinity, and smaller values of ν yield heavier tails.

The *t* location-scale distribution uses the following parameters.

Parameter	Description	Support
`mu`	Location parameter	$-\infty < \mu < \infty$
`sigma`	Scale parameter	$\sigma > 0$
`nu`	Shape parameter	$\nu > 0$

Creation

There are several ways to create a `tLocationScaleDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

`mu` — Location parameter
scalar value

Location parameter of the *t* location-scale distribution, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Scale parameter
positive scalar value

Scale parameter of the *t* location-scale distribution, specified as a positive scalar value.

Data Types: `single` | `double`

`nu` — Degrees of freedom
positive scalar value

Degrees of freedom of the *t* location-scale distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution

`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters

positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates

matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters

array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values

vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function

icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a t Location-Scale Distribution Object Using Default Parameters

Create a *t* location scale distribution object using the default parameter values.

```
pd = makedist('tLocationScale')
```

```
pd =
  tLocationScaleDistribution

  t Location-Scale distribution
       mu = 0
    sigma = 1
       nu = 5
```

Create a t Location-Scale Distribution Object Using Specified Parameters

Create a *t* location-scale distribution object by specifying the parameter values.

```
pd = makedist('tLocationScale','mu',-2,'sigma',1,'nu',20)
```

```
pd =
  tLocationScaleDistribution

  t Location-Scale distribution
       mu = -2
    sigma =  1
       nu = 20
```

Compute the interquartile range of the distribution.

```
r = iqr(pd)
```

```
r = 1.3739
```

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics
“Represent Cauchy Distribution Using t Location-Scale” on page 5-104
“t Location-Scale Distribution” on page B-156

Introduced in R2013a

TriangularDistribution

Triangular probability distribution object

Description

A `TriangularDistribution` object consists of parameters and a model description for a triangular probability distribution.

The triangular distribution is frequently used in simulations when limited sample data is available. The lower and upper limits represent the smallest and largest values, and the location of the peak represents an estimate of the mode.

The triangular distribution uses the following parameters.

Parameter	Description	Support
a	Lower limit	$a \le b$
b	Peak location	$a \le b \le c$
c	Upper limit	$c \ge b$

Creation

Create a `TriangularDistribution` probability distribution with specified parameter values object using `makedist`.

Properties

Distribution Parameters

a — Lower limit
scalar value

Lower limit for the triangular distribution, specified as a scalar value.

Data Types: `single` | `double`

b — Peak location
scalar value

Location of the peak for the triangular distribution, specified as a scalar value greater than or equal to `a`.

Data Types: `single` | `double`

c — Upper limit
scalar value

Upper limit for the triangular distribution, specified as a scalar value greater than or equal to `b`.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
pdf	Probability density function
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Triangular Distribution Object Using Default Parameters

Create a triangular distribution object using the default parameter values.

```
pd = makedist('Triangular')
```

```
pd =
  TriangularDistribution

A = 0, B = 0.5, C = 1
```

Create a Triangular Distribution Object Using Specified Parameters

Create a triangular distribution object by specifying parameter values.

```
pd = makedist('Triangular','a',-2,'b',1,'c',5)
```

```
pd =
  TriangularDistribution

A = -2, B = 1, C = 5
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 1.3333
```

See Also

makedist

Topics

“Generate Random Numbers Using the Triangular Distribution” on page 5-47
“Triangular Distribution” on page B-158
“Nonparametric and Empirical Probability Distributions” on page 5-30

Introduced in R2013a

UniformDistribution

Uniform probability distribution object

Description

A `UniformDistribution` object consists of parameters and a model description for a uniform probability distribution.

The uniform distribution has a constant probability density function between its two parameters, lower (the minimum) and upper (the maximum). This distribution is appropriate for representing round-off errors in values tabulated to a particular number of decimal places.

The uniform distribution uses the following parameters.

Parameter	Description	Support
`lower`	Lower parameter	$-\infty < lower < upper$
`upper`	Upper parameter	$lower < upper < \infty$

Creation

Create a `UniformDistribution` probability distribution with specified parameter values object using `makedist`.

Properties

Distribution Parameters

`lower` — Lower parameter
scalar value

Lower parameter for the uniform distribution, specified as a scalar value.

Data Types: `single` | `double`

`upper` — Upper parameter
scalar value

Upper parameter for the uniform distribution, specified as a scalar value greater than `lower`.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
pdf	Probability density function
random	Random numbers
std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Uniform Distribution Object Using Default Parameters

Create a uniform distribution object using the default parameter values.

```
pd = makedist('Uniform')
```

```
pd =
  UniformDistribution

  Uniform distribution
    Lower = 0
    Upper = 1
```

Create a Uniform Distribution Object Using Specified Parameters

Create a uniform distribution object by specifying parameter values.

```
pd = makedist('Uniform','Lower',-4,'Upper',2)
```

```
pd =
  UniformDistribution

  Uniform distribution
    Lower = -4
    Upper =  2
```

Compute the interquartile range of the distribution

```
r = iqr(pd)
```

```
r = 3
```

See Also

`makedist`

Topics
“Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101
“Uniform Distribution (Continuous)” on page B-163

Introduced in R2013a

WeibullDistribution

Weibull probability distribution object

Description

A `WeibullDistribution` object consists of parameters, a model description, and sample data for a Weibull probability distribution.

The Weibull distribution is used in reliability and lifetime modeling, and to model the breaking strength of materials.

The Weibull distribution uses the following parameters.

Parameter	Description	Support
`a`	Scale parameter	$a > 0$
`b`	Shape parameter	$b > 0$

Creation

There are several ways to create a `WeibullDistribution` probability distribution object.

- Create a distribution with specified parameter values using `makedist`.
- Fit a distribution to data using `fitdist`.
- Interactively fit a distribution to data using the **Distribution Fitter** app.

Properties

Distribution Parameters

a — Scale parameter
positive scalar value

Scale parameter of the Weibull distribution, specified as a positive scalar value.

Data Types: `single` | `double`

b — Shape parameter
positive scalar value

Shape parameter of the Weibull distribution, specified as a positive scalar value.

Data Types: `single` | `double`

Distribution Characteristics

`IsTruncated` — Logical flag for truncated distribution
`0` | `1`

This property is read-only.

Logical flag for truncated distribution, specified as a logical value. If `IsTruncated` equals `0`, the distribution is not truncated. If `IsTruncated` equals `1`, the distribution is truncated.

Data Types: `logical`

`NumParameters` — Number of parameters
positive integer value

This property is read-only.

Number of parameters for the probability distribution, specified as a positive integer value.

Data Types: `double`

`ParameterCovariance` — Covariance matrix of the parameter estimates
matrix of scalar values

This property is read-only.

Covariance matrix of the parameter estimates, specified as a *p*-by-*p* matrix, where *p* is the number of parameters in the distribution. The (`i`,`j`) element is the covariance between the estimates of the `i`th parameter and the `j`th parameter. The (`i`,`i`) element is the estimated variance of the `i`th parameter. If parameter `i` is fixed rather than estimated by fitting the distribution to data, then the (`i`,`i`) elements of the covariance matrix are 0.

Data Types: `double`

`ParameterIsFixed` — Logical flag for fixed parameters
array of logical values

This property is read-only.

Logical flag for fixed parameters, specified as an array of logical values. If `0`, the corresponding parameter in the `ParameterNames` array is not fixed. If `1`, the corresponding parameter in the `ParameterNames` array is fixed.

Data Types: `logical`

`ParameterValues` — Distribution parameter values
vector of scalar values

This property is read-only.

Distribution parameter values, specified as a vector.

Data Types: `single` | `double`

`Truncation` — Truncation interval
vector of scalar values

This property is read-only.

Truncation interval for the probability distribution, specified as a vector containing the lower and upper truncation boundaries.

Data Types: `single` | `double`

Other Object Properties

`DistributionName` — Probability distribution name
character vector

This property is read-only.

Probability distribution name, specified as a character vector.

Data Types: `char`

`InputData` — Data used for distribution fitting
structure

This property is read-only.

Data used for distribution fitting, specified as a structure containing the following:

- `data`: Data vector used for distribution fitting.
- `cens`: Censoring vector, or empty if none.
- `freq`: Frequency vector, or empty if none.

Data Types: `struct`

`ParameterDescription` — Distribution parameter descriptions
cell array of character vectors

This property is read-only.

Distribution parameter descriptions, specified as a cell array of character vectors. Each cell contains a short description of one distribution parameter.

Data Types: `char`

`ParameterNames` — Distribution parameter names
cell array of character vectors

This property is read-only.

Distribution parameter names, specified as a cell array of character vectors.

Data Types: `char`

Object Functions

cdf	Cumulative distribution function
icdf	Inverse cumulative distribution function
iqr	Interquartile range
mean	Mean of probability distribution
median	Median of probability distribution
negloglik	Negative loglikelihood of probability distribution
paramci	Confidence intervals for probability distribution parameters
pdf	Probability density function
proflik	Profile likelihood function for probability distribution
random	Random numbers

std	Standard deviation of probability distribution
truncate	Truncate probability distribution object
var	Variance of probability distribution

Examples

Create a Weibull Distribution Object Using Default Parameters

Create a Weibull distribution object using the default parameter values.

```
pd = makedist('Weibull')
```

```
pd =
  WeibullDistribution

  Weibull distribution
    A = 1
    B = 1
```

Create a Weibull Distribution Object Using Specified Parameter Values

Create a Weibull distribution object by specifying the parameter values.

```
pd = makedist('Weibull','a',2,'b',5)
```

```
pd =
  WeibullDistribution

  Weibull distribution
    A = 2
    B = 5
```

Compute the mean of the distribution.

```
m = mean(pd)
```

```
m = 1.8363
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- You must create a probability distribution object by fitting a probability distribution to sample data from the `fitdist` function. For the usage notes and limitations of `fitdist`, see “Code Generation” on page 33-2252 of `fitdist`.
- These object functions support code generation: `cdf`, `icdf`, `iqr`, `mean`, `median`, `pdf`, `std`, `truncate`, and `var`.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “Code Generation for Probability Distribution Objects” on page 32-82.

See Also

Distribution Fitter | `fitdist` | `makedist`

Topics

“Fit Probability Distribution Objects to Grouped Data” on page 5-92
“Compare Multiple Distribution Fits” on page 5-87
“Estimate Parameters of Three-Parameter Weibull Distribution” on page B-175
“Weibull Distribution” on page B-170

Introduced in R2013a

union

Class: `dataset`

(Not Recommended) Set union for dataset array observations

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = union(A,B)
C = union(A,B,vars)
C = union(A,B,vars,setOrder)
[C,iA,iB] = union( ___ )
```

Description

`C = union(A,B)` for `dataset` arrays `A` and `B` returns the combined set of observations from the two arrays, with repetitions removed. The observations in the dataset array `C` are sorted.

`C = union(A,B,vars)` returns the combined set of observations from the two arrays, with repetitions of unique combinations of the variables specified in `vars` removed. The observations in the dataset array `C` are sorted by those variables.

The values for variables not specified in `vars` for each observation in `C` are taken from the corresponding observation in `A` or `B`, or from `A` if there are common observations in both `A` and `B`. If there are multiple observations in `A` or `B` that correspond to an observation in `C`, those values are taken from the first occurrence.

`C = union(A,B,vars,setOrder)` returns the observations in `C` in the order specified by `setOrder`.

`[C,iA,iB] = union( ___ )` also returns index vectors `iA` and `iB` such that `C` is a sorted combination of the values `A(iA,:)` and `B(iB,:)`. If there are common observations in `A` and `B`, then `union` returns only the index from `A`, in `iA`. If there are repeated observations in `A` or `B`, then the index of the first occurrence is returned. You can use any of the previous input arguments.

Input Arguments

`A,B`

Input dataset arrays.

`vars`

String array or cell array of character vectors containing variable names, or a vector of integers containing variable column numbers. `vars` indicates the variables for which `union` removes repetitions of unique combinations of the variables.

Specify `vars` as `[]` to use its default value of all variables.

`setOrder`

Flag indicating the sorting order for the observations in `C`. The possible values of `setOrder` are:

`'sorted'`	Observations in `C` are in sorted order (default).
`'stable'`	Observations in `C` are in the same order that they appear in `A`, then `B`.

Output Arguments

`C`

Dataset array with the combined observations of `A` and `B`, with repetitions removed. `C` is in sorted order (by default), or the order specified by `setOrder`.

`iA`

Index vector, indicating the observations in `A` that contribute to the union. `iA` contains the index to the first occurrence of any repeated observations in `A`.

`iB`

Index vector, indicating the observations in `B` that contribute to the union. If there are common observations in `A` and `B`, then `union` returns only the index from `A`, in `iA`. `iB` contains the index to the first occurrence of any repeated observations in `B`.

Examples

Union of Two Dataset Arrays

Load sample data.

```
A = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'));
B = dataset('XLSFile',fullfile(matlabroot,'help/toolbox/stats/examples','hospitalSmall.xlsx'),'Sh
[length(A) length(B)]
```

```
ans =

    14     8
```

The first dataset array, `A`, has 14 observations. The second dataset array, `B`, has 8 observations.

Return the union.

```
C = union(A,B);
length(C)
```

```
ans =

    21
```

The union of the two dataset arrays has 21 observations, indicating that there was one observation replicated in A and B.

See Also

`dataset` | `intersect` | `ismember` | `setdiff` | `setxor` | `sortrows` | `unique`

Topics

“Create a Dataset Array from a File” on page 2-62
“Merge Dataset Arrays” on page 2-85
“Dataset Arrays” on page 2-112

unique

Class: `dataset`

(Not Recommended) Unique observations in dataset array

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
C = unique(A)
[C,ia,ic] = unique(A)
C = unique(A,vars)
[C,ia,ic] = unique(A,vars)
[...] = unique(A,vars,occurrence)
[...] = unique(...,'R2012a')
[...] = unique(...,'legacy')
[...] = unique(A,vars,setOrder)
```

Description

Note The behavior of `dataset.unique` is consistent with the MATLAB function `unique`. For a demonstration of using the `'legacy'` flag to preserve the behavior from R2012b and prior in your existing code, see the documentation for `unique`.

`C = unique(A)` returns a copy of the dataset A that contains only the sorted unique observations. `A` must contain only variables whose class has a `unique` method, including:

- numeric
- character
- logical
- categorical
- string
- cell arrays of character vectors

For a variable with multiple columns, its class's `unique` method must support the `'rows'` flag.

`[C,ia,ic] = unique(A)` also returns index vectors `ia` and `ic` such that `C = A(ia,:)` and `A = C(ic,:)`.

`C = unique(A,vars)` returns a dataset that contains only one observation for each unique combination of values for the variables in `A` specified in `vars`. `vars` is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. `C` includes all variables from `A`. The values in `C` for the variables not specified in `vars` are taken from the last occurrence among observations in `A` with each unique combination of values for the variables specified in `vars`.

`[C,ia,ic] = unique(A,vars)` also returns index vectors `ia` and `ic` such that `C = A(ia,:)` and `A(:,vars) = C(ic,vars)`.

`[...] = unique(A,vars,occurrence)` specifies which index is returned in `ia` in the case of repeated observations in `A`. The default value is `occurrence='first'`, which returns the index of the first occurrence of each repeated observation in `A`. `occurrence='last'` returns the index of the last occurrence of each repeated observation in `A`. The values in `C` for variables not specified in `vars` are taken from the observations `A(ia,:)`. Specify `vars` as `[]` to use the default value of all variables.

`[...] = unique(...,'R2012a')` adopts the future behavior of `unique`. You can specify the flag as the final argument with any previous syntax that accepts `A`, `vars`, or `occurrence`.

`[...] = unique(...,'legacy')` preserves the current behavior of `unique`. You can specify the flag as the final argument with any previous syntax that accepts `A`, `vars`, or `occurrence`.

`[...] = unique(A,vars,setOrder)` returns the observations of `C` in a specific order. `setOrder='sorted'` returns the values of `C` in sorted order. `setOrder='stable'` returns the values of `C` in the same order as `A`. If there are repeated observations in `A`, then `ia` returns the index of the first occurrence of each repeated observation. Specify `vars` as `[]` to use the default value of all variables.

See Also

`dataset` | `set` | `subsasgn`

unidcdf

Discrete uniform cumulative distribution function

Syntax

```
p = unidcdf(x,N)
p = unidcdf(x,N,'upper')
```

Description

`p = unidcdf(x,N)` returns the discrete uniform cdf at each value in `x` using the corresponding maximum observable value in `N`. `x` and `N` can be vectors, matrices, or multidimensional arrays that have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The maximum observable values in `N` must be positive integers.

`p = unidcdf(x,N,'upper')` returns the complement of the discrete uniform cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The discrete uniform cdf is

$$p = F(x \mid N) = \frac{floor(x)}{N} I_{(1,\ldots,N)}(x)$$

The result, p, is the probability that a single observation from the discrete uniform distribution with maximum N will be a positive integer less than or equal to x. The values x do not need to be integers.

Examples

Compute Discrete Uniform Distribution cdf

What is the probability of drawing a number 20 or less from a hat with the numbers from 1 to 50 inside?

```
probability = unidcdf(20,50)
```

```
probability = 0.4000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`cdf` | `mle` | `unidinv` | `unidpdf` | `unidrnd` | `unidstat`

Topics

“Uniform Distribution (Discrete)” on page B-168

Introduced before R2006a

unidinv

Discrete uniform inverse cumulative distribution function

Syntax

```
X = unidinv(P,N)
```

Description

`X = unidinv(P,N)` returns the smallest positive integer `X` such that the discrete uniform cdf evaluated at `X` is equal to or exceeds `P`. You can think of `P` as the probability of drawing a number as large as `X` out of a hat with the numbers `1` through `N` inside.

`P` and `N` can be vectors, matrices, or multidimensional arrays that have the same size, which is also the size of `X`. A scalar input for `N` or `P` is expanded to a constant array with the same dimensions as the other input. The values in `P` must lie on the interval `[0 1]` and the values in `N` must be positive integers.

Examples

```
x = unidinv(0.7,20)
x =
    14

y = unidinv(0.7 + eps,20)
y =
    15
```

A small change in the first parameter produces a large jump in output. The cdf and its inverse are both step functions. The example shows what happens at a step.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`icdf` | `unidcdf` | `unidpdf` | `unidrnd` | `unidstat`

Topics

“Uniform Distribution (Discrete)” on page B-168

Introduced before R2006a

unidpdf

Discrete uniform probability density function

Syntax

```
Y = unidpdf(X,N)
```

Description

`Y = unidpdf(X,N)` computes the discrete uniform pdf at each of the values in `X` using the corresponding maximum observable value in `N`. `X` and `N` can be vectors, matrices, or multidimensional arrays that have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in `N` must be positive integers.

The discrete uniform pdf is

$$y = f(x|N) = \frac{1}{N} I_{(1,\ldots,N)}(x)$$

You can think of y as the probability of observing any one number between 1 and n.

Examples

For fixed `n`, the uniform discrete pdf is a constant.

```
y = unidpdf(1:6,10)
y =
  0.1000  0.1000  0.1000  0.1000  0.1000  0.1000
```

Now fix `x`, and vary `n`.

```
likelihood = unidpdf(5,4:9)
likelihood =
  0  0.2000  0.1667  0.1429  0.1250  0.1111
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`pdf` | `unidcdf` | `unidinv` | `unidrnd` | `unidstat`

Topics
“Uniform Distribution (Discrete)” on page B-168

Introduced before R2006a

unidrnd

Random numbers from discrete uniform distribution

Syntax

```
r = unidrnd(n)
r = unidrnd(n,sz1,...,szN)
r = unidrnd(n,sz)
```

Description

`r = unidrnd(n)` generates random numbers from the discrete uniform distribution specified by its maximum value `n`.

`n` can be a scalar, vector, matrix, or multidimensional array.

`r = unidrnd(n,sz1,...,szN)` generates an array of random numbers from the discrete uniform distribution with the scalar maximum value `n`, where `sz1,...,szN` indicates the size of each dimension.

`r = unidrnd(n,sz)` generates an array of random numbers from the discrete uniform distribution with the scalar maximum value `n`, where vector `sz` specifies `size(r)`.

Examples

Array of Random Numbers from Several Discrete Uniform Distributions

Generate an array of random numbers from the discrete uniform distributions. For each distribution, specify its maximum value.

Specify the maximum values of the distributions.

```
n = 1:10:100;
```

Generate random numbers from the discrete uniform distributions.

```
r = unidrnd(n)
```

```
r = 1×10

     1    10     3    29    26     5    17    39    78    88
```

Array of Random Numbers from One Discrete Uniform Distribution

Generate an array of random numbers from one discrete uniform distribution. Here, the maximum value `n` is a scalar.

Use the `unidrnd` function to generate random numbers from the discrete uniform distribution with the maximum value 100. The function returns one number.

```
R_scalar = unidrnd(100)
```

```
R_scalar = 82
```

Generate a 2-by-3 array of random numbers from the same distribution by specifying the required array dimensions.

```
R_array = unidrnd(100,2,3)
```

```
R_array = 2×3

    91    92    10
    13    64    28
```

Alternatively, specify the required array dimensions as a vector.

```
R_array = unidrnd(100,[2,3])
```

```
R_array = 2×3

    55    97    98
    96    16    96
```

Input Arguments

`n` — Maximum value

positive integer | array of positive integers

Maximum value, specified as a positive integer or array of positive integers.

Example: `unidrnd(10)`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)

integers

Size of each dimension, specified as separate arguments of integers. For example, specifying `5,3,2` generates a 5-by-3-by-2 array of random numbers from the discrete uniform distribution.

If `n` is an array, then the specified dimensions `sz1,...,szN` must match the dimensions of `n`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `unidrnd` ignores trailing dimensions with a size of 1. For example, `unidrnd(n,3,1,1,1)` produces a 3-by-1 vector of random numbers.

Example: `5,3,2`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers. For example, specifying `[5 3 2]` generates a 5-by-3-by-2 array of random numbers from the discrete uniform distribution.

If `n` is an array, then the specified dimensions `sz` must match the dimensions of `n`.

- If you specify a single value `[sz1]`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `unidrnd` ignores trailing dimensions with a size of 1. For example, `unidrnd(n,[3 1 1 1])` produces a 3-by-1 vector of random numbers.

Example: `[5 3 2]`

Data Types: `single` | `double`

Output Arguments

`r` — Random numbers from discrete uniform distribution
scalar value | array of scalar values

Random numbers from the discrete uniform distribution, returned as a scalar value or an array of scalar values.

Data Types: `single` | `double`

Alternative Functionality

- `unidrnd` is a function specific to discrete uniform distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Note that the distribution-specific function `unidrnd` is faster than the generic function `random`.
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers than MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`random` | `unidcdf` | `unidinv` | `unidpdf` | `unidstat`

Topics

“Uniform Distribution (Discrete)” on page B-168

Introduced before R2006a

unidstat

Discrete uniform mean and variance

Syntax

```
[M,V] = unidstat(N)
```

Description

`[M,V] = unidstat(N)` returns the mean and variance of the discrete uniform distribution with minimum value `1` and maximum value `N`.

The mean of the discrete uniform distribution with parameter N is $(N + 1)/2$. The variance is $(N^2 - 1)/12$.

Examples

```
[m,v] = unidstat(1:6)
m =
  1.0000  1.5000  2.0000  2.5000  3.0000  3.5000
v =
  0  0.2500  0.6667  1.2500  2.0000  2.9167
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`unidcdf` | `unidinv` | `unidpdf` | `unidrnd`

Topics

“Uniform Distribution (Discrete)” on page B-168

Introduced before R2006a

unifcdf

Continuous uniform cumulative distribution function

Syntax

```
p = unifcdf(x,a,b)
p = unifcdf(x,a,b,'upper')
```

Description

`p = unifcdf(x,a,b)` returns the uniform cdf at each value in `x` using the corresponding lower endpoint (minimum), `a` and upper endpoint (maximum), `b`. `x`, `a`, and `b` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant matrix with the same dimensions as the other inputs.

`p = unifcdf(x,a,b,'upper')` returns the complement of the uniform cdf at each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities.

The uniform cdf is

$$p = F(x \mid a, b) = \frac{x - a}{b - a} I_{[a, b]}(x)$$

The standard uniform distribution has `a = 0` and `b = 1`.

Examples

Compute Uniform Distribution cdf

What is the probability that an observation from a standard uniform distribution will be less than 0.75?

```
probability = unifcdf(0.75)
```

```
probability = 0.7500
```

What is the probability that an observation from a uniform distribution with `a = -1` and `b = 1` will be less than 0.75?

```
probability = unifcdf(0.75,-1,1)
```

```
probability = 0.8750
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

See Also
`cdf` | `unifinv` | `unifit` | `unifpdf` | `unifrnd` | `unifstat`

Topics
“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unifinv

Continuous uniform inverse cumulative distribution function

Syntax

```
X = unifinv(P,A,B)
```

Description

`X = unifinv(P,A,B)` computes the inverse of the uniform cdf with parameters `A` and `B` (the minimum and maximum values, respectively) at the corresponding probabilities in `P`. `P`, `A`, and `B` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs.

The inverse of the uniform cdf is

$$x = F^{-1}(p|a,b) = a + p(a-b)I_{[0,1]}(p)$$

The standard uniform distribution has A = `0` and B = `1`.

Examples

What is the median of the standard uniform distribution?

```
median_value = unifinv(0.5)
median_value =
  0.5000
```

What is the 99th percentile of the uniform distribution between -1 and 1?

```
percentile = unifinv(0.99,-1,1)
percentile =
  0.9800
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`icdf` | `unifcdf` | `unifit` | `unifpdf` | `unifrnd` | `unifstat`

Topics

“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unifit

Continuous uniform parameter estimates

Syntax

```
[ahat,bhat] = unifit(data)
[ahat,bhat,ACI,BCI] = unifit(data)
[ahat,bhat,ACI,BCI] = unifit(data,alpha)
```

Description

`[ahat,bhat] = unifit(data)` returns the maximum likelihood estimates (MLEs) of the parameters of the uniform distribution given the data in `data`.

`[ahat,bhat,ACI,BCI] = unifit(data)` also returns 95% confidence intervals, `ACI` and `BCI`, which are matrices with two rows. The first row contains the lower bound of the interval for each column of the matrix `data`. The second row contains the upper bound of the interval.

`[ahat,bhat,ACI,BCI] = unifit(data,alpha)` controls the confidence level by using `alpha`. For example, if `alpha = 0.01`, then `ACI` and `BCI` are 99% confidence intervals.

Examples

```
r = unifrnd(10,12,100,2);
[ahat,bhat,aci,bci] = unifit(r)
ahat =
  10.0154  10.0060
bhat =
  11.9989  11.9743
aci =
  9.9551  9.9461
  10.0154  10.0060
bci =
  11.9989  11.9743
  12.0592  12.0341
```

See Also

`mle` | `unifcdf` | `unifinv` | `unifpdf` | `unifrnd` | `unifstat`

Topics

“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unifpdf

Continuous uniform probability density function

Syntax

```
y = unifpdf(x)
y = unifpdf(x,a,b)
```

Description

`y = unifpdf(x)` returns the probability density function (pdf) of the standard uniform distribution, evaluated at the values in `x`.

`y = unifpdf(x,a,b)` returns the pdf of the continuous uniform distribution on the interval [`a`, `b`], evaluated at the values in `x`.

Examples

Compute Standard and Continuous Uniform pdf

The pdf of the standard uniform distribution is constant on the interval `[0,1]`.

Compute the pdf of 0.2, 0.4,...,1 in the standard uniform distribution.

```
x = 0.2:0.2:1;
y = unifpdf(x)
```

```
y = 1×5

     1     1     1     1     1
```

If `x` is not between `a` and `b`, `unifpdf` returns `0`.

Compute the pdf of `1` through `5` in the continuous uniform distribution on the interval `[2,4]`.

```
x2 = 1:5;
unifpdf(x2,2,4)
```

```
ans = 1×5

         0    0.5000    0.5000    0.5000         0
```

If the parameter `a` is larger than `b`, `unifpdf` returns `NaN` regardless of the `x` input.

```
unifpdf(5,10,1)
```

```
ans = NaN
```

Input Arguments

x — Values at which to evaluate pdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the pdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `unifpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

a — Lower endpoint

scalar value | array of scalar values

Lower endpoint of the uniform distribution, specified as a scalar value or an array of scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `unifpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[0 -1 7 9]`

Data Types: `single` | `double`

b — Upper endpoint

scalar value | array of scalar values

Upper endpoint of the uniform distribution, specified as a scalar value or an array of scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `unifpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 1 10 10]`

Data Types: `single` | `double`

Output Arguments

y — pdf values

scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x`, `a`, and `b` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

More About

Continuous Uniform pdf

The continuous uniform distribution is a two-parameter family of curves with a constant pdf on its interval of support, [`a`, `b`]. The parameters a and b are the endpoints of the interval.

The continuous uniform pdf is

$$y = f(x|a,b) = \frac{1}{b-a} I_{[a,b]}(x).$$

The standard uniform distribution occurs when $a = 0$ and $b = 1$.

For more information, see “Uniform Distribution (Continuous)” on page B-163.

Alternative Functionality

- `unifpdf` is a function specific to the continuous uniform distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create a `UniformDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `unifpdf` is faster than the generic function `pdf`.
- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`UniformDistribution` | `pdf` | `unifcdf` | `unifinv` | `unifit` | `unifrnd` | `unifstat`

Topics
“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unifrnd

Continuous uniform random numbers

Syntax

```
r = unifrnd(a,b)
r = unifrnd(a,b,sz1,...,szN)
r = unifrnd(a,b,sz)
```

Description

`r = unifrnd(a,b)` generates a random number from the continuous uniform distribution with the lower endpoints `a` and upper endpoint `b`.

`r = unifrnd(a,b,sz1,...,szN)` generates an array of uniform random numbers, where `sz1,...,szN` indicates the size of each dimension.

`r = unifrnd(a,b,sz)` generates an array of uniform random numbers, where the size vector `sz` specifies `size(r)`.

Examples

Generate Uniform Random Number

Generate a random number from the continuous uniform distribution with the lower parameter `0` and upper parameter `1`.

```
r = unifrnd(0,1)
```

```
r = 0.8147
```

Generate Uniform Random Numbers

Generate `5` random numbers from the continuous uniform distributions on the intervals (0,1), (0,2),..., (0,5).

```
a1 = 0;
b1 = 1:5;
r1 = unifrnd(a1,b1)
```

```
r1 = 1×5

    0.8147    1.8116    0.3810    3.6535    3.1618
```

By default, `unifrnd` generates an array that is the same size as `a` and `b` after any necessary scalar expansion so that all scalars are expanded to match the dimensions of the other inputs.

If you specify array dimensions `sz1,...,szN`, they must match the dimensions of `a` and `b` after any necessary scalar expansion.

Generate a 2-by-3 array of random numbers from the continuous uniform distribution with the lower parameter `0` and upper parameter `1`.

```
sz = [2 3];
r2 = unifrnd(0,1,sz)
```

```
r2 = 2×3

    0.0975    0.5469    0.9649
    0.2785    0.9575    0.1576
```

Generate 6 random numbers on the intervals (0,1), (1,2),..., (5,6).

```
a3 = 0:5;
b3 = 1:6;
r3 = unifrnd(a3,b3,1,6)
```

```
r3 = 1×6

    0.9706    1.9572    2.4854    3.8003    4.1419    5.4218
```

Input Arguments

a — Lower endpoint

scalar value | array of scalar values

Lower endpoint of the uniform distribution, specified as a scalar value or an array of scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If both `a` and `b` are arrays, then the array sizes must be the same. If either `a` or `b` is a scalar, then `unifrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[0 -1 7 9]`

Data Types: `single` | `double`

b — Upper endpoint

scalar value | array of scalar values

Upper endpoint of the uniform distribution, specified as a scalar value or an array of scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If both `a` and `b` are arrays, then the array sizes must be the same. If either `a` or `b` is a scalar, then `unifrnd` expands the scalar argument into a constant array of the same size as the other argument. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[1 1 10 10]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as separate arguments of integers.

If either `a` or `b` is an array, then the specified dimensions `sz1,...,szN` must match the common dimensions of `a` and `b` after any necessary scalar expansion. The default values of `sz1,...,szN` are the common dimensions.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `unifrnd` ignores trailing dimensions with a size of 1. For example, `unifrnd(–3,5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the uniform distribution with lower endpoint –3 and upper endpoint 5.

Example: `2,3`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If either `a` or `b` is an array, then the specified dimensions `sz` must match the common dimensions of `a` and `b` after any necessary scalar expansion. The default values of `sz` are the common dimensions.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `unifrnd` ignores trailing dimensions with a size of 1. For example, `unifrnd(–3,5,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the uniform distribution with lower endpoint –3 and upper endpoint 5.

Example: `[2 3]`

Data Types: `single` | `double`

Output Arguments

`r` — Uniform random numbers
scalar value | array of scalar values

Uniform random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Alternative Functionality

- `unifrnd` is a function specific to the continuous uniform distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, create a `UniformDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its

parameters. Note that the distribution-specific function `unifrnd` is faster than the generic function `random`.

- Use `rand` to generate numbers from the uniform distribution on the interval (0,1).
- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`UniformDistribution` | `rand` | `random` | `unifcdf` | `unifinv` | `unifit` | `unifpdf` | `unifstat`

Topics

“Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101
“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unifstat

Continuous uniform mean and variance

Syntax

```
[M,V] = unifstat(A,B)
```

Description

`[M,V] = unifstat(A,B)` returns the mean of and variance for the continuous uniform distribution using the corresponding lower endpoint (minimum), `A` and upper endpoint (maximum), `B`. Vector or matrix inputs for `A` and `B` must have the same size, which is also the size of `M` and `V`. A scalar input for `A` or `B` is expanded to a constant matrix with the same dimensions as the other input.

The mean of the continuous uniform distribution with parameters a and b is $(a + b)/2$, and the variance is $(a - b)^2/12$.

Examples

```
a = 1:6;
b = 2.*a;
[m,v] = unifstat(a,b)
m =
  1.5000  3.0000  4.5000  6.0000  7.5000  9.0000
v =
  0.0833  0.3333  0.7500  1.3333  2.0833  3.0000
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`unifcdf` | `unifinv` | `unifit` | `unifpdf` | `unifrnd`

Topics

“Uniform Distribution (Continuous)” on page B-163

Introduced before R2006a

unstack

Class: `dataset`

(Not Recommended) Unstack dataset array from single variable into multiple variables

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
A = unstack(B,datavar,indvar)
[A,iB] = unstack(B,datavar,indvar)
A = unstack(B,datavar,indvar,'Parameter',value)
```

Description

`A = unstack(B,datavar,indvar)` unstacks a single variable in dataset array `B` into multiple variables in `A`. In general `A` contains more variables, but fewer observations, than `B`.

`datavar` specifies the data variable in `B` to unstack. `indvar` specifies an indicator variable in `B` that determines which variable in `A` each value in `datavar` is unstacked into. `unstack` treats the remaining variables in `B` as grouping variables. Each unique combination of their values defines a group of observations in `B` that will be unstacked into a single observation in `A`.

`unstack` creates `m` data variables in `A`, where `m` is the number of group levels in `indvar`. The values in `indvar` indicate which of those `m` variables receive which values from `datavar`. The `j`-th data variable in `A` contains the values from `datavar` that correspond to observations whose `indvar` value was the `j`-th of the `m` possible levels. Elements of those `m` variables for which no corresponding data value in `B` exists contain a default value.

`datavar` is a positive integer, a character vector, a string scalar, or a logical vector containing a single true value. `indvar` is a positive integer, a variable name, or a logical vector containing a single true value.

`[A,iB] = unstack(B,datavar,indvar)` returns an index vector `iB` indicating the correspondence between observations in `A` and those in `B`. For each observation in `A`, `iB` contains the index of the first in the corresponding group of observations in `B`.

For more information on grouping variables, see “Grouping Variables” on page 2-45.

Input Arguments

`A = unstack(B,datavar,indvar,'Parameter',value)` uses the following parameter name/value pairs to control how `unstack` converts variables in `B` to variables in `A`:

'GroupVars'	Grouping variables in B that define groups of observations. groupvars is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. The default is all variables in B not listed in datavar or indvar.
'NewDataVarNames'	A string array or cell array of character vectors containing names for the data variables unstack should create in A. Default is the group names of the grouping variable specified in indvar.
'AggregationFun'	A function handle that accepts a subset of values from datavar and returns a single value. stack applies this function to observations from the same group that have the same value of indvar. The function must aggregate the data values into a single value, and in such cases it is not possible to recover B from A using stack. The default is @sum for numeric data variables. For non-numeric variables, there is no default, and you must specify 'AggregationFun' if multiple observations in the same group have the same values of indvar.
'ConstVars'	Variables in B to copy to A without unstacking. The values for these variables in A are taken from the first observation in each group in B, so these variables should typically be constant within each group. ConstVars is a positive integer, a vector of positive integers, a character vector, a string array, a cell array of character vectors, or a logical vector. The default is no variables.

You can also specify more than one data variable in B, each of which becomes a set of m variables in A. In this case, specify datavar as a vector of positive integers, a string array or cell array containing variable names, or a logical vector. You may specify only one variable with indvar. The names of each set of data variables in A are the name of the corresponding data variable in B concatenated with the names specified in 'NewDataVarNames'. The function specified in 'AggregationFun' must return a value with a single row.

Examples

Combine several variables for estimated influenza rates into a single variable. Then unstack the estimated influenza rates by date.

```
load flu

% FLU has a 'Date' variable, and 10 variables for estimated influenza rates
% (in 9 different regions, estimated from Google searches, plus a
% nationwide estimate from the CDC). Combine those 10 variables into an
% array that has a single data variable, 'FluRate', and an indicator
% variable, 'Region', that says which region each estimate is from.
[flu2,iflu] = stack(flu, 2:11, 'NewDataVarName','FluRate', ...
    'IndVarName','Region')

% The second observation in FLU is for 10/16/2005.  Find the observations
% in FLU2 that correspond to that date.
flu(2,:)
flu2(iflu==2,:)

% Use the 'Date' variable from that array to split 'FluRate' into 52
```

```
% separate variables, each containing the estimated influenza rates for
% each unique date.  The new array has one observation for each region.  In
% effect, this is the original array FLU "on its side".
dateNames = cellstr(datestr(flu.Date,'mmm_DD_YYYY'));
[flu3,iflu2] = unstack(flu2, 'FluRate', 'Date', ...
    'NewDataVarNames',dateNames)

% Since observations in FLU3 represent regions, IFLU2 indicates the first
% occurrence in FLU2 of each region.
flu2(iflu2,:)
```

See Also

`join` | `stack`

Topics

“Grouping Variables” on page 2-45

update

Package: `classreg.learning.classif`

Update model parameters for code generation

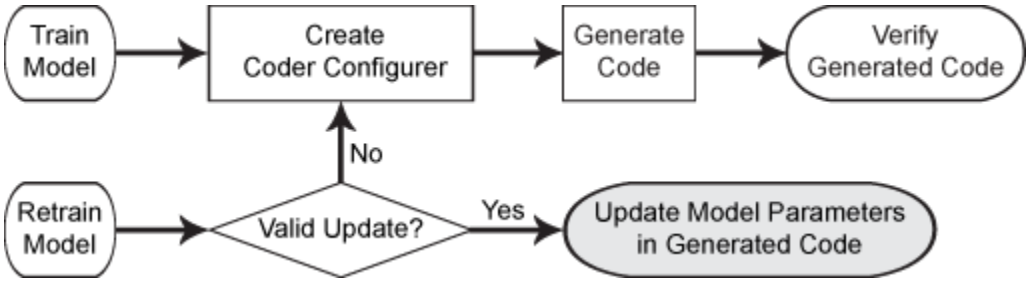
Syntax

```
updatedMdl = update(Mdl,params)
```

Description

Generate C/C++ code for the `predict` and `update` functions of a machine learning model by using a coder configurer object. Create this object by using `learnerCoderConfigurer` and its object function `generateCode`. Then you can use the `update` function to update model parameters in the generated code without having to regenerate the code. This feature reduces the effort required to regenerate, redeploy, and reverify C/C++ code when you retrain a model with new data or settings.

This flow chart shows the code generation workflow using a coder configurer. Use `update` for the highlighted step.



If you do not generate code, then you do not need to use the `update` function. When you retrain a model in MATLAB, the returned model already includes modified parameters.

`updatedMdl = update(Mdl,params)` returns an updated version of `Mdl` that contains new parameters in `params`.

After retraining a model, use the `validatedUpdateInputs` function to detect modified parameters in the retrained model and validate whether the modified parameter values satisfy the coder attributes of the parameters. Use the output of `validatedUpdateInputs`, the validated parameters, as the input `params` to update model parameters.

Examples

Update Parameters of SVM Classification Model in Generated Code

Train a SVM model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data.

Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`). Train a binary SVM classification model using the first 50 observations.

```
load ionosphere
Mdl = fitcsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `ClassificationSVM` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted labels and scores.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:),'NumOutputs',2);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM classification model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 34];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 34 predictors, so the value of the `SizeVector` attribute must be 34 and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 34];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrain
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
VariableDimensions attribute for SupportVectorLabels has been modified to satisfy configuration c
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM classification model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `ClassificationSVMModel` for the two entry-point functions in the `codegen\mex\ClassificationSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score] = predict(Mdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
```

Compare `label` and `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`score_mex` might include round-off differences compared with `score`. In this case, compare `score_mex` and `score`, allowing a small tolerance.

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `score` and `score_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitcsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score] = predict(retrainedMdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `labels` and `labels_mex` are equal, and the score values are equal within the tolerance.

Update Parameters of ECOC Classification Model in Generated Code

Train an error-correcting output codes (ECOC) model using SVM binary learners and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the ECOC model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using different settings, and update parameters in the generated code without regenerating the code.

Train Model

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create an SVM binary learner template to use a Gaussian kernel function and to standardize predictor data.

```
t = templateSVM('KernelFunction','gaussian','Standardize',true);
```

Train a multiclass ECOC model using the template `t`.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

`Mdl` is a `ClassificationECOC` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationECOC` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns the first two outputs of the `predict` function, which are the predicted labels and negated average binary losses.

```
configurer = learnerCoderConfigurer(Mdl,X,'NumOutputs',2)
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

`configurer` is a `ClassificationECOCCoderConfigurer` object, which is a coder configurer of a `ClassificationECOC` object. The display shows the tunable input arguments of `predict` and `update`: `X`, `BinaryLearners`, `Prior`, and `Cost`.

Specify Coder Attributes of Parameters

Specify the coder attributes of `predict` arguments (predictor data and the name-value pair arguments `'Decoding'` and `'BinaryLoss'`) and `update` arguments (support vectors of the SVM learners) so that you can use these arguments as the input arguments of `predict` and `update` in the generated code.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector`

attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 4];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 4 predictors, so the second value of the `SizeVector` attribute must be 4 and the second value of the `VariableDimensions` attribute must be `false`.

Next, modify the coder attributes of `BinaryLoss` and `Decoding` to use the `'BinaryLoss'` and `'Decoding'` name-value pair arguments in the generated code. Display the coder attributes of `BinaryLoss`.

```
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

             Value: 'hinge'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 0
```

To use a nondefault value in the generated code, you must specify the value before generating the code. Specify the `Value` attribute of `BinaryLoss` as `'exponential'`.

```
configurer.BinaryLoss.Value = 'exponential';
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

             Value: 'exponential'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 1
```

If you modify attribute values when `Tunability` is `false` (logical 0), the software sets the `Tunability` to `true` (logical 1).

Display the coder attributes of `Decoding`.

```
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: 'lossweighted'
```

```
    SelectedOption: 'Built-in'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 1
        Tunability: 0
```

Specify the `IsConstant` attribute of `Decoding` as `false` so that you can use all available values in `BuiltInOptions` in the generated code.

```
configurer.Decoding.IsConstant = false;
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: [1x1 LearnerCoderInput]
    SelectedOption: 'NonConstant'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 0
        Tunability: 1
```

The software changes the `Value` attribute of `Decoding` to a `LearnerCoderInput` object so that you can use both `'lossweighted'` and `'lossbased'` as the value of `'Decoding'`. Also, the software sets the `SelectedOption` to `'NonConstant'` and the `Tunability` to `true`.

Finally, modify the coder attributes of `SupportVectors` in `BinaryLearners`. Display the coder attributes of `SupportVectors`.

```
configurer.BinaryLearners.SupportVectors
```

```
ans =
  LearnerCoderInput with properties:

            SizeVector: [54 4]
    VariableDimensions: [1 0]
              DataType: 'double'
            Tunability: 1
```

The default value of `VariableDimensions` is `[true false]` because each learner has a different number of support vectors. If you retrain the ECOC model using new data or different settings, the number of support vectors in the SVM learners can vary. Therefore, increase the upper bound of the number of support vectors.

```
configurer.BinaryLearners.SupportVectors.SizeVector = [150 4];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Display the coder configurer.

```
configurer
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
        BinaryLoss: [1x1 EnumeratedInput]
          Decoding: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

The display now includes `BinaryLoss` and `Decoding` as well.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the ECOC classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationECOCModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationECOCModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationECOCModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument. Because you specified `'Decoding'` as a tunable input argument by changing the `IsConstant` attribute before generating the code, you also need to specify it in the call to the MEX function, even though `'lossweighted'` is the default value of `'Decoding'`.

```
[label,NegLoss] = predict(Mdl,X,'BinaryLoss','exponential');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
```

Compare `label` to `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`NegLoss_mex` might include round-off differences compared to `NegLoss`. In this case, compare `NegLoss_mex` to `NegLoss`, allowing a small tolerance.

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `NegLoss` and `NegLoss_mex` are equal within the tolerance `1e−8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using a different setting. Specify `'KernelScale'` as `'auto'` so that the software selects an appropriate scale factor using a heuristic procedure.

```
t_new = templateSVM('KernelFunction','gaussian','Standardize',true,'KernelScale','auto');
retrainedMdl = fitcecoc(X,Y,'Learners',t_new);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationECOCModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` to the outputs from the `predict` function in the updated MEX function.

```
[label,NegLoss] = predict(retrainedMdl,X,'BinaryLoss','exponential','Decoding','lossbased');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `label` and `label_mex` are equal, and `NegLoss` and `NegLoss_mex` are equal within the tolerance.

Update Parameters of SVM Regression Model in Generated Code

Train a support vector machine (SVM) model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `carsmall` data set and train an SVM regression model using the first 50 observations.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `RegressionSVM` object.

Create Coder Configurer

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:));
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM regression model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM regression model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 2];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of

observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains two predictors, so the value of the `SizeVector` attribute must be two and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 2];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `RegressionSVMModel` for the two entry-point functions in the `codegen\mex\RegressionSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
yfit = predict(Mdl,X);
yfit_mex = RegressionSVMModel('predict',X);
```

`yfit_mex` might include round-off differences compared with `yfit`. In this case, compare `yfit` and `yfit_mex`, allowing a small tolerance.

```
find(abs(yfit-yfit_mex) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e–6`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
yfit = predict(retrainedMdl,X);
yfit_mex = RegressionSVMModel('predict',X);
find(abs(yfit-yfit_mex) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e-6`.

Update Parameters of Regression Tree Model in Generated Code

Train a regression tree using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Load the `carbig` data set, and train a regression tree model using half of the observations.

```
load carbig
X = [Displacement Horsepower Weight];
Y = MPG;

rng('default') % For reproducibility
```

```
n = length(Y);
idxTrain = randsample(n,n/2);
XTrain = X(idxTrain,:);
YTrain = Y(idxTrain);

Mdl = fitrtree(XTrain,YTrain);
```

`Mdl` is a `RegressionTree` object.

Create Coder Configurer

Create a coder configurer for the `RegressionTree` model by using `learnerCoderConfigurer`. Specify the predictor data `XTrain`. The `learnerCoderConfigurer` function uses the input `XTrain` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted responses and node numbers for the predictions.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'NumOutputs',2);
```

`configurer` is a `RegressionTreeCoderConfigurer` object, which is a coder configurer of a `RegressionTree` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the regression tree model parameters so that you can update the parameters in the generated code after retraining the model.

Specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 3];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

The size of the first dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 3 predictors, the value of the `SizeVector` attribute must be `3` and the value of the `VariableDimensions` attribute must be `0`.

If you retrain the tree model using new data or different settings, the number of nodes in the tree can vary. Therefore, specify the first dimension of the `SizeVector` attribute of one of these properties so that you can update the number of nodes in the generated code: `Children`, `CutPoint`, `CutPredictorIndex`, or `NodeMean`. The software then modifies the other properties automatically.

For example, set the first value of the `SizeVector` attribute of the `NodeMean` property to `Inf`. The software modifies the `SizeVector` and `VariableDimensions` attributes of `Children`, `CutPoint`,

and `CutPredictorIndex` to match the new upper bound on the number of nodes in the tree. Additionally, the first value of the `VariableDimensions` attribute of `NodeMean` changes to `1`.

```
configurer.NodeMean.SizeVector = [Inf 1];
```

```
SizeVector attribute for Children has been modified to satisfy configuration constraints.
SizeVector attribute for CutPoint has been modified to satisfy configuration constraints.
SizeVector attribute for CutPredictorIndex has been modified to satisfy configuration constraints
VariableDimensions attribute for Children has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPoint has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPredictorIndex has been modified to satisfy configuration co
```

```
configurer.NodeMean.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the regression tree model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionTreeModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[Yfit,node] = predict(Mdl,XTrain);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',XTrain);
```

Compare `Yfit` to `Yfit_mex` and `node` to `node_mex`.

```
max(abs(Yfit-Yfit_mex),[],'all')
```

```
ans = 0
```

```
isequal(node,node_mex)
```

```
ans = logical
   1
```

In general, `Yfit_mex` might include round-off differences compared to `Yfit`. In this case, the comparison confirms that `Yfit` and `Yfit_mex` are equal.

`isequal` returns logical 1 (`true`) if all the input arguments are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same node numbers.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrtree(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionTreeModel('update',params)
```

Verify Generated Code

Compare the output arguments from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[Yfit,node] = predict(retrainedMdl,X);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',X);

max(abs(Yfit-Yfit_mex),[],'all')

ans = 0

isequal(node,node_mex)

ans = logical
   1
```

The comparison confirms that the predicted responses and node numbers are equal.

Input Arguments

`Mdl` — Machine learning model

model object

Machine learning model, specified as a model object, as given in this table of supported models.

Model	Model Object
Binary decision tree for multiclass classification	`CompactClassificationTree`

Model	Model Object
SVM for one-class and binary classification	`CompactClassificationSVM`
Linear model for binary classification	`ClassificationLinear`
Multiclass model for SVMs and linear models	`CompactClassificationECOC`
Binary decision tree for regression	`CompactRegressionTree`
Support vector machine (SVM) regression	`CompactRegressionSVM`
Linear regression	`RegressionLinear`

For the code generation usage notes and limitations of a machine learning model, see the Code Generation section of the model object page.

`params` — Parameters to update
structure

Parameters to update in the machine learning model, specified as a structure with a field for each parameter to update.

Create `params` by using the `validatedUpdateInputs` function. This function detects modified parameters in the retrained model, validates whether the modified parameter values satisfy the coder attributes of the parameters, and returns the parameters to update as a structure.

The set of parameters that you can update varies depending on the machine learning model, as described in this table.

Model	Parameters to Update
Binary decision tree for multiclass classification	`Children`, `ClassProbability`, `Cost`, `CutPoint`, `CutPredictorIndex`, `Prior`
SVM for one-class and binary classification	• If you train an SVM classification model (`Mdl`) with a linear kernel function and discard support vectors by using `discardSupportVectors`, then `params` can include `Beta`, `Bias`, `Cost`, `Mu`, `Prior`, `Scale` (kernel scale parameter in `KernelParameters`), and `Sigma`. • Otherwise, `params` can include `Alpha`, `Beta`, `Cost`, `Mu`, `Prior`, `Scale` (kernel scale parameter in `KernelParameters`), `Sigma`, `SupportVectorLabels`, and `SupportVectors`.
	• If `Mdl` is a one-class SVM classification model, then `params` cannot include `Cost` or `Prior`.
Linear model for binary classification	`Beta`, `Bias`, `Cost`, `Prior`
Multiclass model for SVMs and linear models	`BinaryLearners`, `Cost`, `Prior`
Binary decision tree for regression	`Children`, `CutPoint`, `CutPredictorIndex`, `NodeMean`

Model	Parameters to Update
SVM regression	• If you train an SVM regression model (`Mdl`) with a linear kernel function and discard support vectors by using `discardSupportVectors`, then `params` can include `Beta`, `Bias`, `Mu`, `Scale` (kernel scale parameter in `KernelParameters`), and `Sigma`. • Otherwise, `params` can include `Alpha`, `Bias`, `Mu`, `Scale` (kernel scale parameter in `KernelParameters`), `Sigma`, and `SupportVectors`.
Linear regression	`Beta`, `Bias`

Output Arguments

`updatedMdl` — Updated machine learning model

model object

Updated machine learning model, returned as a model object that is the same type of object as `Mdl`. The output `updatedMdl` is an updated version of the input `Mdl` that contains new parameters in `params`.

Tips

- If you modify any of the name-value pair arguments listed in this table when you retrain a model, you cannot use `update` to update the parameters. You must generate C/C++ code again.

Model	Arguments Not Supported for Update
Binary decision tree for multiclass classification	Arguments of `fitctree` — `'ClassNames'`, `'ScoreTransform'`
SVM for one-class and binary classification	Arguments of `fitcsvm` — `'ClassNames'`, `'KernelFunction'`, `'PolynomialOrder'`, `'ScoreTransform'`, `'Standardize'`
Linear model for binary classification	Arguments of `fitclinear` — `'ClassNames'`, `'ScoreTransform'`
Multiclass model for SVMs and linear models	Arguments of `fitcecoc` — `'ClassNames'`, `'Coding'`, `'ScoreTransform'` If you specify the binary learners in `fitcecoc` as template objects (see `'Learners'`), then for each binary learner, you cannot modify the following: • Arguments of `templateSVM` — `'KernelFunction'`, `'PolynomialOrder'`, `'Standardize'` • Arguments of `templateLinear` — `'Learner'` (because modifying the model type changes the score transform of the binary learner)
Binary decision tree for regression	Arguments of `fitrtree` — `'ResponseTransform'`

Model	Arguments Not Supported for Update
SVM regression	Arguments of `fitrsvm` — `'KernelFunction'`, `'PolynomialOrder'`, `'ResponseTransform'`, `'Standardize'`
Linear regression	Arguments of `fitrlinear` — `'ResponseTransform'`

- In the coder configurer workflow, you use `generateCode` to create both the `update.m` entry-point function and the MEX function for the entry-point function. Assuming the name of the MEX function is `myModel`, you call `update` using this syntax.

  ```
  myModel('update',params)
  ```

  To see how the syntax described on this page is used in the entry-point function, display the contents of the `update.m` and `initialize.m` files by using the `type` function.

  ```
  type update.m
  type initialize.m
  ```

  For an example that shows the contents of the `update.m` and `initialize.m` files, see “Generate Code Using Coder Configurer” on page 33-2610.

Algorithms

In the coder configurer workflow, the `Mdl` input argument of `update` is a model returned by `loadLearnerForCoder`. This model and the `updatedMdl` object are reduced classification or regression models that primarily contain properties required for prediction.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Create a coder configurer by using `learnerCoderConfigurer` and then generate code for `predict` and `update` by using the object function `generateCode`.
- For the code generation usage notes and limitations of the machine learning model `Mdl`, see the Code Generation section of the model object page.

Model	Model Object
Binary decision tree for multiclass classification	`CompactClassificationTree`
SVM for one-class and binary classification	`CompactClassificationSVM`
Linear model for binary classification	`ClassificationLinear`
Multiclass model for SVMs and linear models	`CompactClassificationECOC`
Binary decision tree for regression	`CompactRegressionTree`
Support vector machine (SVM) regression	`CompactRegressionSVM`
Linear regression	`RegressionLinear`

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

`generateCode` | `learnerCoderConfigurer` | `validatedUpdateInputs`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

updateMetrics

Update performance metrics in linear model for incremental learning given new data

Syntax

```
Mdl = updateMetrics(Mdl,X,Y)
Mdl = updateMetrics(Mdl,X,Y,Name,Value)
```

Description

Given streaming data, `updateMetrics` measures the performance of a configured incremental learning model for linear regression (`incrementalRegressionLinear` object) or linear binary classification (`incrementalClassificationLinear` object). `updateMetrics` stores the performance metrics in the output model.

`updateMetrics` allows for flexible incremental learning. After you call the function to update model performance metrics on an incoming chunk of data, you can perform other actions before you train the model to the data. For example, you can decide whether you need to train the model based on its performance on a chunk of data. Alternatively, you can both update model performance metrics and train the model on the data as it arrives, in one call, by using the `updateMetricsAndFit` function.

To measure the model performance on a specified batch of data, call `loss` instead.

`Mdl = updateMetrics(Mdl,X,Y)` returns an incremental learning model `Mdl`, which is the input incremental learning model `Mdl` modified to contain the model performance metrics on the incoming predictor and response data, `X` and `Y` respectively.

When the input model is warm (`Mdl.IsWarm` is `true`), `updateMetrics` overwrites previously computed metrics, stored in the `Metrics` property, with the new values. Otherwise, `updateMetrics` stores `NaN` values in `Metrics` instead.

The input and output models have the same data type.

`Mdl = updateMetrics(Mdl,X,Y,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify that the columns of the predictor data matrix correspond to observations, and set observation weights.

Examples

Track Performance of Incremental Model

Train a linear model for binary classification by using `fitclinear`, convert it to an incremental learner, and then track its performance to streaming data.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
```

```
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Train Linear Model for Binary Classification

Fit a linear model for binary classification to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitclinear(X(idxtt,:),Y(idxtt))
```

```
TTMdl =
  ClassificationLinear
      ResponseName: 'Y'
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2999
            Lambda: 8.2967e-05
           Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `ClassificationLinear` model object representing a traditionally trained linear model for binary classification.

Convert Trained Model

Convert the traditionally trained classification model to a binary classification linear model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationLinear

            IsWarm: 1
           Metrics: [1x2 table]
        ClassNames: [0 1]
    ScoreTransform: 'none'
              Beta: [60x1 double]
              Bias: -0.2999
           Learner: 'svm'


  Properties, Methods
```

```
IncrementalMdl.IsWarm
```

```
ans = logical
   1
```

The incremental model is warm. Therefore, `updateMetrics` can track model performance metrics given data.

Track Performance Metrics

Track the model performance on the rest of the data by using the `updateMetrics` function. Simulate a data stream by processing 50 observations at a time. At each iteration:

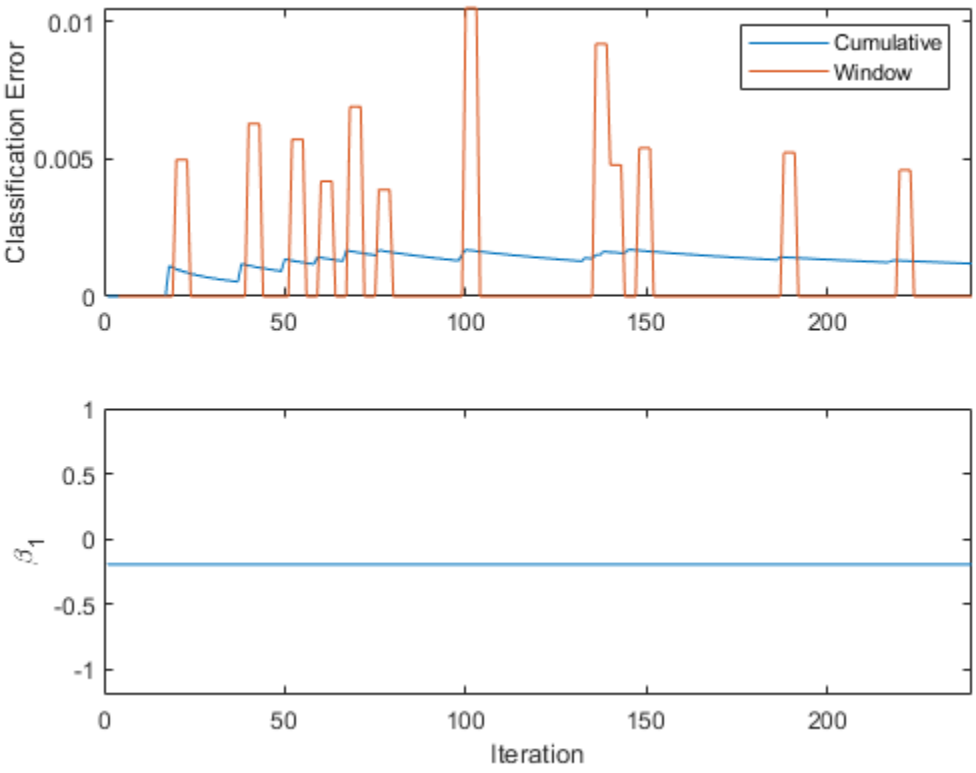
1. Call `updateMetrics` to update the cumulative and window classification error of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model.
2. Store the classification error and first coefficient β_1.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = [IncrementalMdl.Beta(1); zeros(nchunk,1)];
Xil = X(idxil,:);
Yil = Y(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(idx,:),Yil(idx));
    ce{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
    beta1(j + 1) = IncrementalMdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object that has tracked the model performance to observations in the data stream.

Plot a trace plot of the performance metrics and estimated coefficient β_1.

```
figure;
subplot(2,1,1)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
legend(h,ce.Properties.VariableNames)
subplot(2,1,2)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable, whereas the window loss jumps.

β_1 does not change because `updateMetrics` does not fit the model to the data.

Configure Incremental Model to Track Performance Metrics

Create a incremental linear SVM model for binary classification. Specify an estimation period of 5,000 observations and the SGD solver.

```
Mdl = incrementalClassificationLinear('EstimationPeriod',5000,'Solver','sgd')
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
              Beta: [0x1 double]
              Bias: 0
           Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model. All its properties are read-only.

Determine whether the model is warm and the size of the metrics warm-up period by querying model properties.

```
isWarm = Mdl.IsWarm
```

```
isWarm = logical
   0
```

```
mwp = Mdl.MetricsWarmupPeriod
```

```
mwp = 1000
```

`Mdl.IsWarm` is `0`; therefore, `Mdl` is not warm.

Determine the number of observations incremental fitting functions, such as `fit`, must process before measuring the performance of the model.

```
numObsBeforeMetrics = Mdl.MetricsWarmupPeriod + Mdl.EstimationPeriod
```

```
numObsBeforeMetrics = 6000
```

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Perform incremental learning. At each iteration, perform the following actions:

- Simulate a data stream by processing a chunk of 50 observations.
- Measure model performance metrics on the incoming chunk using `updateMetrics`. Overwrite the input model.
- Fit the model to the incoming chunk. Overwrite the input model.
- Store β_1 and the misclassification error rate to see how they evolve during incremental learning.

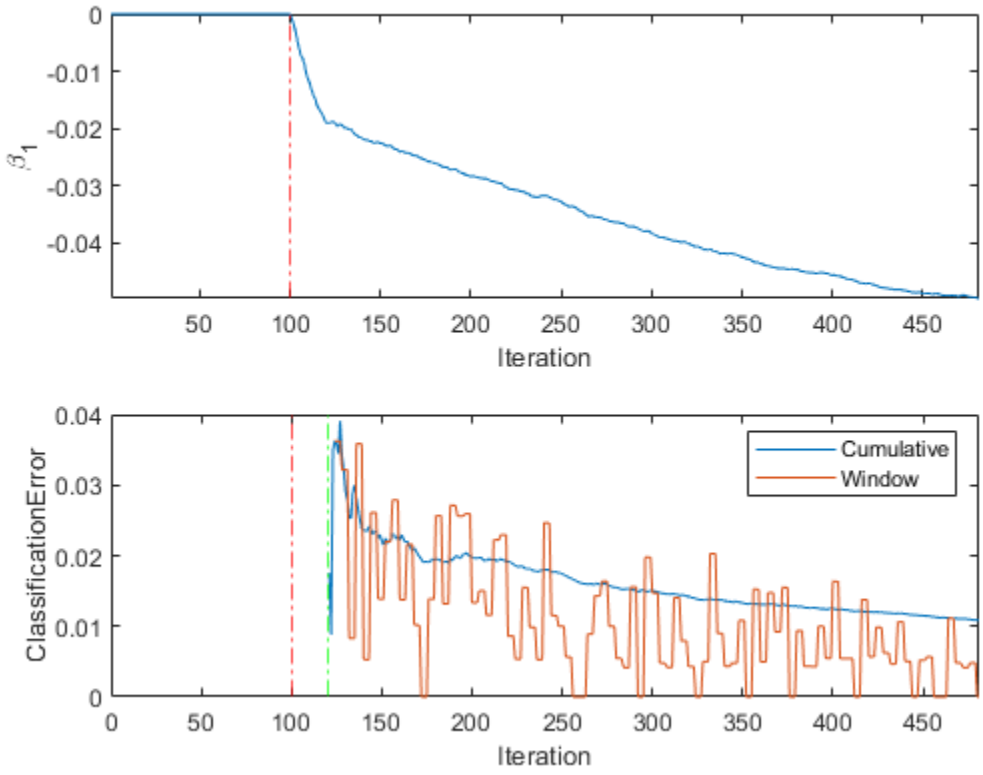
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
```

```
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    Mdl = fit(Mdl,X(idx,:),Y(idx));
    beta1(j) = Mdl.Beta(1);
end
```

`IncrementalMdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream.

To see how the parameters evolved during incremental learning, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
axis tight
subplot(2,1,2)
plot(ce.Variables);
ylabel('ClassificationError')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline(numObsBeforeMetrics/numObsPerChunk,'g-.');
xlabel('Iteration')
xlim([0 nchunk]);
legend(ce.Properties.VariableNames)
```



```
mdlIsWarm = numObsBeforeMetrics/numObsPerChunk
```

```
mdlIsWarm = 120
```

The plot suggests that `fit` does not fit the model to the data or update the parameters until after the estimation period. Also, `updateMetrics` does not track the classification error until after the estimation and metrics warm-up periods (120 chunks).

Perform Conditional Training

Incrementally train a linear regression model only when its performance degrades.

Load and shuffle the 2015 NYC housing data set. For more details on the data, see NYC Open Data.

```
load NYCHousing2015

rng(1) % For reproducibility
n = size(NYCHousing2015,1);
shuffidx = randsample(n,n);
NYCHousing2015 = NYCHousing2015(shuffidx,:);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data, and transpose the data to speed up computations.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}]';
```

Configure a linear regression model for incremental learning so that it does not have an estimation or metrics warm-up period. Specify a metrics window size of 1000 observations. Fit the configured model to the first 100 observations, and specify that the observations are oriented along the columns of the data.

```
Mdl = incrementalRegressionLinear('EstimationPeriod',0,'MetricsWarmupPeriod',0,...
    'MetricsWindowSize',1000);
numObsPerChunk = 100;
Mdl = fit(Mdl,X(:,1:numObsPerChunk),Y(1:numObsPerChunk),'ObservationsIn','columns');
```

`Mdl` is an `incrementalRegressionLinear` model object.

Perform incremental learning, with conditional fitting, by following this procedure for each iteration:

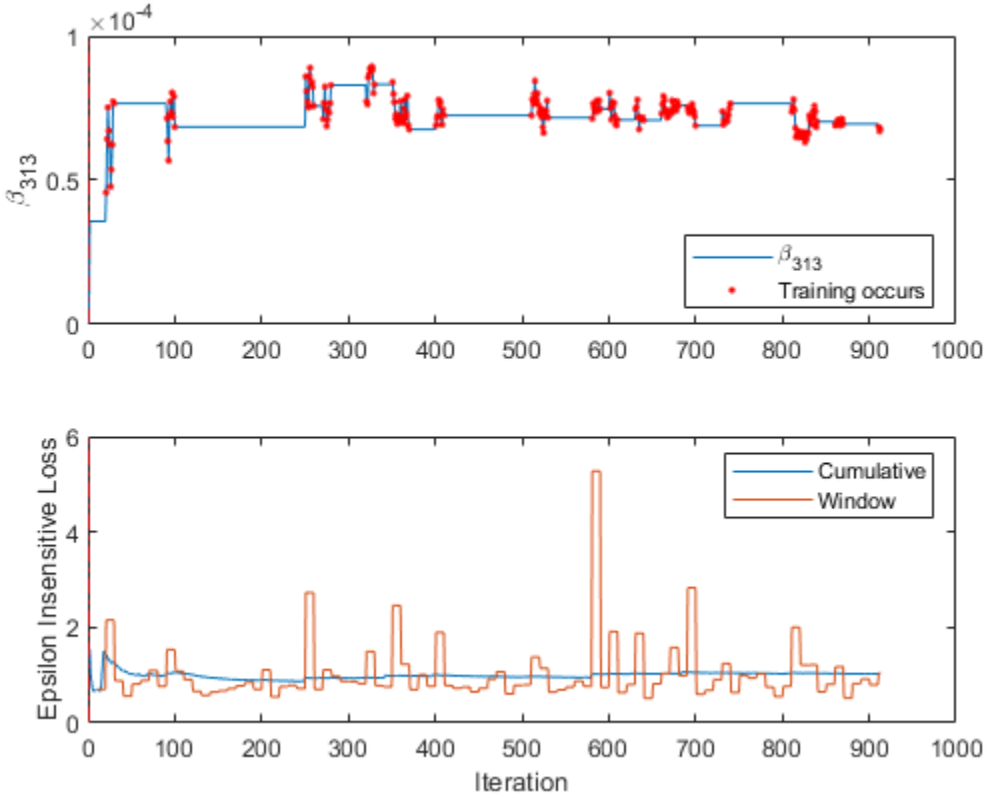
- Simulate a data stream by processing a chunk of 100 observations.
- Update the model performance by computing the epsilon insensitive loss, within a 200 observation window. Specify that the observations are oriented along the columns of the data.
- Fit the model to the chunk of data only when the loss more than doubles from the minimum loss experienced. Specify that the observations are oriented along the columns of the data.
- When tracking performance and fitting, overwrite the previous incremental model.
- Store the epsilon insensitive loss and β_{313} to see the how the loss and coefficient evolve during training.
- Track when `fit` trains the model.

```
% Preallocation
n = numel(Y) - numObsPerChunk;
nchunk = floor(n/numObsPerChunk);
beta313 = zeros(nchunk,1);
ei = array2table(nan(nchunk,2),'VariableNames',["Cumulative" "Window"]);
trained = false(nchunk,1);

% Incremental fitting
for j = 2:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(:,idx),Y(idx),'ObservationsIn','columns');
    ei{j,:} = Mdl.Metrics{"EpsilonInsensitiveLoss",:};
    minei = min(ei{:,2});
    pdiffloss = (ei{j,2} - minei)/minei*100;
    if pdiffloss > 100
        Mdl = fit(Mdl,X(:,idx),Y(idx),'ObservationsIn','columns');
        trained(j) = true;
    end
    beta313(j) = Mdl.Beta(end);
end
```

`Mdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

To see how the model performance and β_{313} evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(beta313)
hold on
plot(find(trained),beta313(trained),'r.')
ylabel('\beta_{313}')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
legend('\beta_{313}','Training occurs','Location','southeast')
hold off
subplot(2,1,2)
plot(ei.Variables)
ylabel('Epsilon Insensitive Loss')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xlabel('Iteration')
legend(ei.Properties.VariableNames)
```



The trace plot of β_{313} shows periods of constant values, during which the loss did not double from the minimum experienced.

Input Arguments

`Mdl` — Incremental learning model whose performance is measured

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Incremental learning model whose performance is measured, specified as an `incrementalClassificationLinear` or `incrementalRegressionLinear` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

If `Mdl.IsWarm` is `false`, `updateMetrics` does not track the performance of the model. You must fit `Mdl` to `Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod` observations by passing `Mdl` and the data to `fit` before `updateMetrics` can track performance metrics. For more details, see “Algorithms” on page 33-6515.

`X` — Chunk of predictor data

floating-point matrix

Chunk of predictor data with which to measure the model performance, specified as a floating-point matrix of *n* observations and `Mdl.NumPredictors` predictor variables. The value of the

`'ObservationsIn'` name-value pair argument determines the orientation of the variables and observations.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `updateMetrics` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `updateMetrics` issues an error.
- `updateMetrics` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

Y — Chunk of labels

categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels with which to measure the model performance, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors for classification problems; or a floating-point vector for regression problems.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

For classification problems:

- `updateMetrics` supports binary classification only.
- When the `ClassNames` property of the input model `Mdl` is nonempty, the following conditions apply:
  - If `Y` contains a label that is not a member of `Mdl.ClassNames`, `updateMetrics` issues an error.
  - The data type of `Y` and `Mdl.ClassNames` must be the same.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Note

If an observation (predictor or label) or weight `Weight` contains at least one missing (`NaN`) value, `updateMetrics` ignores the observation. Consequently, `updateMetrics` uses fewer than *n* observations to compute the model performance.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ObservationsIn','columns','Weights',W` specifies that the columns of the predictor matrix correspond to observations, and the vector `W` contains observation weights to apply during incremental learning.

`ObservationsIn` — Predictor data observation dimension
`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`.

Data Types: `char` | `string`

`Weights` — Chunk of observation weights
floating-point vector of positive values

Chunk of observation weights, specified as the comma-separated pair consisting of `'Weights'` and a floating-point vector of positive values. `updateMetrics` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(`*n*`,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-6515.

Data Types: `double` | `single`

Output Arguments

`Mdl` — Updated incremental learning model
`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Updated incremental learning model, returned as an incremental learning model object of the same data type as the input model `Mdl`, either `incrementalClassificationLinear` or `incrementalRegressionLinear`.

If the model is not warm, `updateMetrics` does not compute performance metrics. As a result, the `Metrics` property of `Mdl` remains completely composed of `NaN` values. If the model is warm, `updateMetrics` computes the cumulative and window performance metrics on the new data `X` and `Y`, and overwrites the corresponding elements of `Mdl.Metrics`. All other properties of the input model `Mdl` carry over to the output model `Mdl`. For more details, see “Algorithms” on page 33-6515.

Tips

- Unlike traditional training, incremental learning might not have a separate test (holdout) set. Therefore, to treat each incoming chunk of data as a test set, pass the incremental model and each incoming chunk to `updateMetrics` before training the model on the same data using `fit`.

Algorithms

Performance Metrics

- `updateMetrics` tracks only model performance metrics, specified by the row labels of the table in `Mdl.Metrics`, from new data when the incremental model is warm (`IsWarm` property is `true`). An incremental model is warm after the `fit` function fits the incremental model to `Mdl.MetricsWarmupPeriod` observations, which is the metrics warm-up period.

  If `Mdl.EstimationPeriod` > 0, the functions estimate hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` updates the metrics at the following frequencies:

  - `Cumulative` — The function computes cumulative metrics since the start of model performance tracking. The function updates metrics every time you call it and bases the calculation on the entire supplied data set.
  - `Window` — The function computes metrics based on all observations within a window determined by the `Mdl.MetricsWindowSize` property. `Mdl.MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `Mdl.MetricsWindowSize` is 20, the function computes metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `Mdl.MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `Mdl.MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `Mdl.MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the function uses the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

Observation Weights

For classification problems, if the prior class probability distribution is known (in other words, the prior distribution is not empirical), `updateMetrics` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that observation weights are the respective prior class probabilities by default.

For regression problems or if the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `updateMetrics`.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `updateMetrics` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `updateMetrics` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for updateMetrics, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `updateMetrics`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For usage notes and limitations of the model object, see `incrementalClassificationLinear` or `incrementalRegressionLinear`.
`X`	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `Y`. • The number of predictor variables must equal to `Mdl.NumPredictors`. • `X` must be `single` or `double`.
`Y`	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `X`. • For classification problems, all labels in `Y` must be represented in `Mdl.ClassNames`. • `Y` and `Mdl.ClassNames` must have the same data type.

- The following restrictions apply:
  - If you configure `Mdl` to shuffle data (`Mdl.Shuffle` is `true`, or `Mdl.Solver` is `'sgd'` or `'asgd'`), the `updateMetrics` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB. Therefore, if you fit `Mdl` before updating the performance metrics, the metrics computed in MATLAB and those computed by the generated code might not be equal.
  - Use a homogeneous data type for all floating-point input arguments and object properties, specifically, either `single` or `double`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Objects
`incrementalClassificationLinear` | `incrementalRegressionLinear`

Functions
`fit` | `loss` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2020b

updateMetrics

Update performance metrics in naive Bayes classification model for incremental learning given new data

Syntax

```
Mdl = updateMetrics(Mdl,X,Y)
Mdl = updateMetrics(Mdl,X,Y,'Weights',Weights)
```

Description

Given streaming data, `updateMetrics` measures the performance of a configured naive Bayes classification model for incremental learning (`incrementalClassificationNaiveBayes` object). `updateMetrics` stores the performance metrics in the output model.

`updateMetrics` allows for flexible incremental learning. After you call the function to update model performance metrics on an incoming chunk of data, you can perform other actions before you train the model to the data. For example, you can decide whether you need to train the model based on its performance on a chunk of data. Alternatively, you can both update model performance metrics and train the model on the data as it arrives, in one call, by using the `updateMetricsAndFit` function.

To measure the model performance on a specified batch of data, call `loss` instead.

`Mdl = updateMetrics(Mdl,X,Y)` returns a naive Bayes classification model for incremental learning `Mdl`, which is the input naive Bayes classification model for incremental learning `Mdl` modified to contain the model performance metrics on the incoming predictor and response data, `X` and `Y` respectively.

When the input model is warm (`Mdl.IsWarm` is `true`), `updateMetrics` overwrites previously computed metrics, stored in the `Metrics` property, with the new values. Otherwise, `updateMetrics` stores `NaN` values in `Metrics` instead.

The input and output models have the same data type.

`Mdl = updateMetrics(Mdl,X,Y,'Weights',Weights)` specifies observation weights `Weights`.

Examples

Track Performance of Incremental Model

Train a naive Bayes classification model by using `fitcnb`, convert it to an incremental learner, and then track its performance to streaming data.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
```

```
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Train Naive Bayes Classification Model

Fit a naive Bayes classification model to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitcnb(X(idxtt,:),Y(idxtt))
```

```
TTMdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 12053
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained model.

Convert Trained Model

Convert the traditionally trained classification model to a naive Bayes classification model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes

                    IsWarm: 1
                   Metrics: [1×2 table]
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

The incremental model is warm. Therefore, `updateMetrics` can track model performance metrics given data.

Track Performance Metrics

Track the model performance on the rest of the data by using the `updateMetrics` function. Simulate a data stream by processing 50 observations at a time. At each iteration:

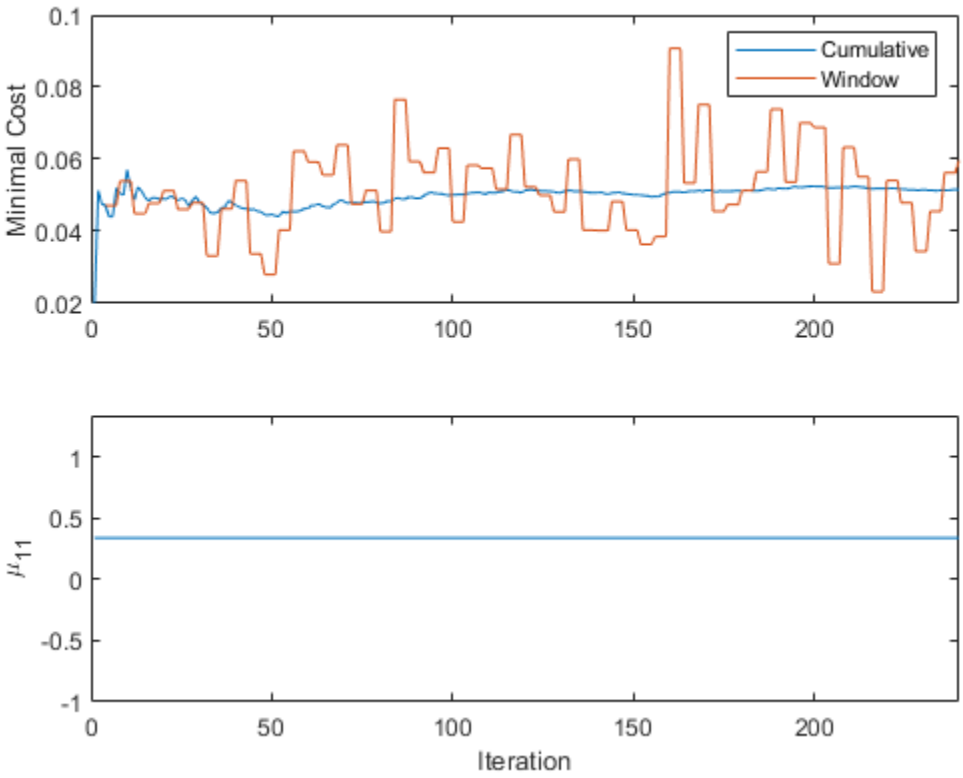
1. Call `updateMetrics` to update the cumulative and window minimal cost of the model given the incoming chunk of observations. Overwrite the previous incremental model to update the losses in the `Metrics` property. Note that the function does not fit the model to the chunk of data—the chunk is "new" data for the model.
2. Store the minimal cost and mean of the first predictor in the first class μ_{11}.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu11 = [IncrementalMdl.DistributionParameters{1,1}(1); zeros(nchunk,1)];
Xil = X(idxil,:);
Yil = Y(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetrics(IncrementalMdl,Xil(idx,:),Yil(idx));
    mc{j,:} = IncrementalMdl.Metrics{"MinimalCost",:};
    mu11(j + 1) = IncrementalMdl.DistributionParameters{1,1}(1);
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object that has tracked the model performance to observations in the data stream.

Plot a trace plot of the performance metrics and estimated coefficient μ_{11}.

```
figure;
subplot(2,1,1)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
legend(h,mc.Properties.VariableNames)
subplot(2,1,2)
plot(mu11)
ylabel('\mu_{11}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss is stable, whereas the window loss jumps.

μ_{11} does not change because `updateMetrics` does not fit the model to the data.

Configure Incremental Model to Track Performance Metrics and Specify Weights

Create a naive Bayes classification model for incremental learning by calling `incrementalClassificationNaiveBayes` and specifying a maximum of 5 expected classes in the data. Specify tracking misclassification error rate in addition to minimal cost.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5,'Metrics',"classiferror");
```

`Mdl` is an `incrementalClassificationNaiveBayes` model. All its properties are read-only.

Determine whether the model is warm and the size of the metrics warm-up period by querying model properties.

```
isWarm = Mdl.IsWarm
```

```
isWarm = logical
   0
```

```
mwp = Mdl.MetricsWarmupPeriod
```

```
mwp = 1000
```

`Mdl.IsWarm` is `0`; therefore, `Mdl` is not warm.

Determine the number of observations incremental fitting functions, such as `fit`, must process before measuring the performance of the model.

```
numObsBeforeMetrics = Mdl.MetricsWarmupPeriod
```

```
numObsBeforeMetrics = 1000
```

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Suppose that the data collected when the subject was not moving (`Y <= 2`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from a still subject, and 1 to a moving subject.

```
W = ones(n,1) + ~Y;
```

Implement incremental learning by performing the following actions at each iteration:

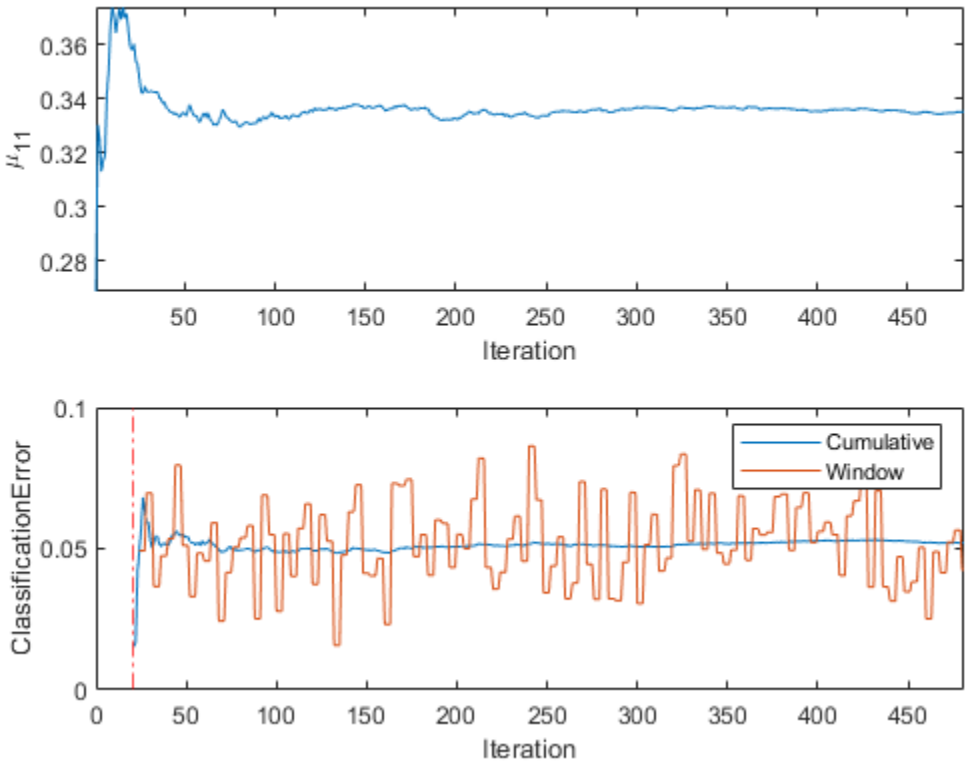
- Simulate a data stream by processing a chunk of 50 observations.
- Measure model performance metrics on the incoming chunk using `updateMetrics`. Specify the corresponding observation weights and overwrite the input model.
- Fit the model to the incoming chunk. Specify the corresponding observation weights and overwrite the input model.
- Store μ_{11} and the misclassification error rate to see how they evolve during incremental learning.

```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu11 = zeros(nchunk,1);

% Incremental learning
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx),'Weights',W(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    Mdl = fit(Mdl,X(idx,:),Y(idx),'Weights',W(idx));
    mu11(j) = Mdl.DistributionParameters{1,1}(1);
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the parameters evolved during incremental learning, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(mu11)
ylabel('\mu_{11}')
xlabel('Iteration')
axis tight
subplot(2,1,2)
plot(ce.Variables);
ylabel('ClassificationError')
xline(numObsBeforeMetrics/numObsPerChunk,'r-.');
xlabel('Iteration')
xlim([0 nchunk]);
legend(ce.Properties.VariableNames)
```




```
mdlIsWarm = numObsBeforeMetrics/numObsPerChunk
```

```
mdlIsWarm = 20
```

The plot suggests that `fit` always fits the model to the data, and `updateMetrics` does not track the classification error until after the metrics warm-up period (20 chunks).

Perform Conditional Training

Incrementally train a naive Bayes classification model only when its performance degrades.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1) % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Configure a naive Bayes classification model for incremental learning so that the maximum number of expected classes is 5, the tracked performance metric includes the misclassification error rate, and the metrics window size of 1000. Fit the configured model to the first 1000 observations.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5,'MetricsWindowSize',1000,...
    'Metrics','classiferror');
initobs = 1000;
Mdl = fit(Mdl,X(1:initobs,:),Y(1:initobs));
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object.

Perform incremental learning, with conditional fitting, by following this procedure for each iteration:

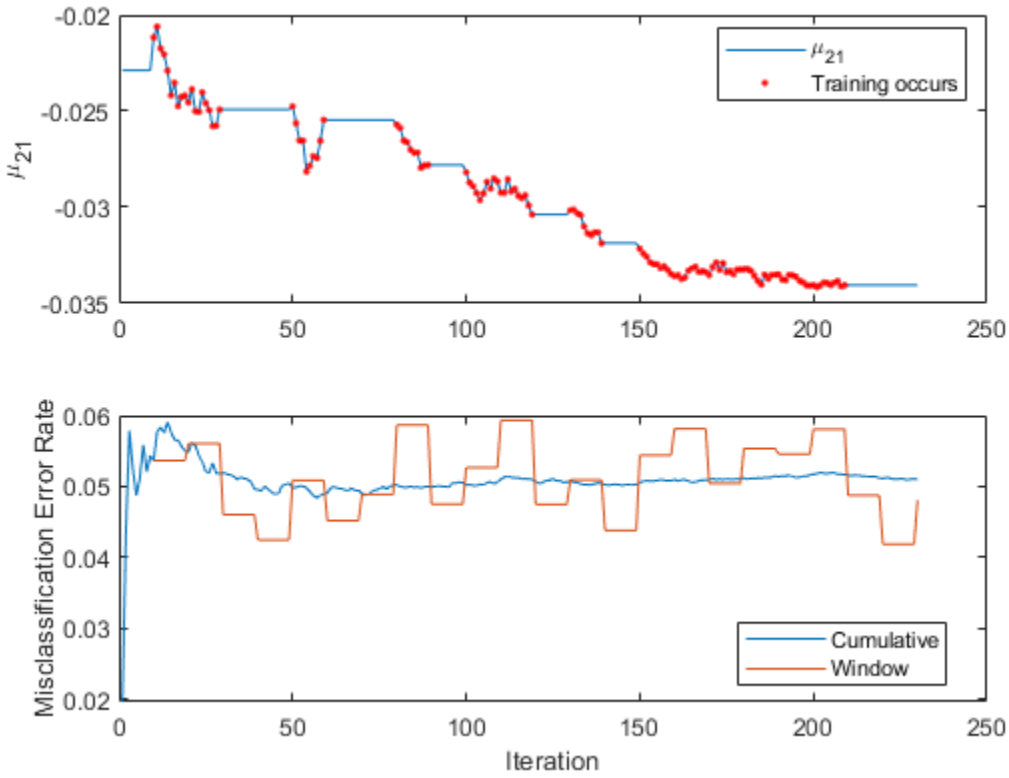
- Simulate a data stream by processing a chunk of 100 observations at a time.
- Update the model performance on the incoming chunk of data.
- Fit the model to the chunk of data only when the misclassification error rate is greater than 0.05.
- When tracking performance and fitting, overwrite the previous incremental model.
- Store the misclassification error rate and the mean of the first predictor in the second class μ_{21} to see how they evolve during training.
- Track when `fit` trains the model.

```
% Preallocation
numObsPerChunk = 100;
nchunk = floor((n - initobs)/numObsPerChunk);
mu21 = zeros(nchunk,1);
ce = array2table(nan(nchunk,2),'VariableNames',["Cumulative" "Window"]);
trained = false(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1 + initobs);
    iend   = min(n,numObsPerChunk*j + initobs);
    idx = ibegin:iend;
    Mdl = updateMetrics(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    if ce{j,2} > 0.05
        Mdl = fit(Mdl,X(idx,:),Y(idx));
        trained(j) = true;
    end
    mu21(j) = Mdl.DistributionParameters{2,1}(1);
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

To see how the model performance and μ_{21} evolved during training, plot them on separate subplots.

```
subplot(2,1,1)
plot(mu21)
hold on
plot(find(trained),mu21(trained),'r.')
ylabel('\mu_{21}')
legend('\mu_{21}','Training occurs','Location','best')
hold off
subplot(2,1,2)
plot(ce.Variables)
ylabel('Misclassification Error Rate')
xlabel('Iteration')
legend(ce.Properties.VariableNames,'Location','best')
```



The trace plot of μ_{21} shows periods of constant values, during which the loss within the previous 1000 observation window is at most 0.05.

Input Arguments

`Mdl` — Incremental learning model whose performance is measured

`incrementalClassificationNaiveBayes` model object

Incremental learning model whose performance is measured, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by

converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

If `Mdl.IsWarm` is `false`, `updateMetrics` does not track the performance of the model. Before `updateMetrics` can track performance metrics, you must perform all the following actions:

- Fit the input model `Mdl` to all expected classes (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`)
- Fit the input model `Mdl` to `Mdl.MetricsWarmupPeriod` observations by passing `Mdl` and the data to `fit` . For more details, see “Performance Metrics” on page 33-6527.

`X` — Chunk of predictor data
floating-point matrix

Chunk of predictor data with which to measure the model performance

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*`j`*`)` is the label of observation *j* (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `updateMetrics` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `updateMetrics` issues an error.
- `updateMetrics` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`Y` — Chunk of labels
categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels with which to measure the model performance, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*`j`*`)` is the label of observation *j* (row or column) in `X`. `updateMetrics` issues an error when at least one of the conditions is met:

- `Y` contains a newly encountered label and the maximum number of classes has been reached previously (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`).
- The data types of `Y` and `Mdl.ClassNames` are different.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

`Weights` — Chunk of observation weights
floating-point vector of positive values

Chunk of observation weights, specified as a floating-point vector of positive values. `updateMetrics` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(n,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-6528.

Data Types: `double` | `single`

Note

If an observation (predictor or label) or weight `Weight` contains at least one missing (`NaN`) value, `updateMetrics` ignores the observation. Consequently, `updateMetrics` uses fewer than *n* observations to compute the model performance.

Output Arguments

`Mdl` — Updated naive Bayes classification model for incremental learning

`incrementalClassificationNaiveBayes` model object

Updated naive Bayes classification model for incremental learning, returned as an incremental learning model object of the same data type as the input model `Mdl`, an `incrementalClassificationNaiveBayes` object.

If the model is not warm, `updateMetrics` does not compute performance metrics. As a result, the `Metrics` property of `Mdl` remains completely composed of `NaN` values. If the model is warm, `updateMetrics` computes the cumulative and window performance metrics on the new data `X` and `Y`, and overwrites the corresponding elements of `Mdl.Metrics`. All other properties of the input model `Mdl` carry over to the output model `Mdl`. For more details, see “Performance Metrics” on page 33-6527.

Tips

- Unlike traditional training, incremental learning might not have a separate test (holdout) set. Therefore, to treat each incoming chunk of data as a test set, pass the incremental model and each incoming chunk to `updateMetrics` before training the model on the same data using `fit`.

Algorithms

Performance Metrics

- `updateMetrics` tracks only model performance metrics, specified by the row labels of the table in `Mdl.Metrics`, from new data when the incremental model is warm (`IsWarm` property is `true`). An incremental model is warm when the `fit` function performs both of the following actions:
  - Fit the incremental model to `Mdl.MetricsWarmupPeriod` observations, which is the metrics warm-up period.
  - Fit the incremental model to all expected classes (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`)

- `Mdl.Metrics` stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetrics` updates the metrics at the following frequencies:

  - `Cumulative` — The function computes cumulative metrics since the start of model performance tracking. The function updates metrics every time you call it and bases the calculation on the entire supplied data set.
  - `Window` — The function computes metrics based on all observations within a window determined by the `Mdl.MetricsWindowSize` property. `Mdl.MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `Mdl.MetricsWindowSize` is 20, the function computes metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `Mdl.MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `Mdl.MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `Mdl.MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the function uses the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

Observation Weights

For each conditional predictor distribution, `updateMetrics` computes the weighted average and standard deviation.

If the prior class probability distribution is known (in other words, the prior distribution is not empirical), `updateMetrics` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that the default observation weights are the respective prior class probabilities.

If the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `updateMetrics`.

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `loss` | `updateMetricsAndFit`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Flexible Workflow” on page 26-26

Introduced in R2021a

updateMetricsAndFit

Update performance metrics in linear model for incremental learning given new data and train model

Syntax

```
Mdl = updateMetricsAndFit(Mdl,X,Y)
Mdl = updateMetricsAndFit(Mdl,X,Y,Name,Value)
```

Description

Given streaming data, `updateMetricsAndFit` first evaluates the performance of a configured incremental learning model for linear regression (`incrementalRegressionLinear` object) or linear binary classification (`incrementalClassificationLinear` object) by calling `updateMetrics` on incoming data. Then `updateMetricsAndFit` fits the model to that data by calling `fit`. In other words, `updateMetricsAndFit` performs prequential evaluation because it treats each incoming chunk of data as a test set, and tracks performance metrics measured cumulatively and over a specified window [1].

`updateMetricsAndFit` provides a simple way to update model performance metrics and train the model on each chunk of data. Alternatively, you can perform the operations separately by calling `updateMetrics` and then `fit`, which allows for more flexibility (for example, you can decide whether you need to train the model based on its performance on a chunk of data).

`Mdl = updateMetricsAndFit(Mdl,X,Y)` returns an incremental learning model `Mdl`, which is the input incremental learning model `Mdl` with the following modifications:

1. `updateMetricsAndFit` measures the model performance on the incoming predictor and response data, `X` and `Y` respectively. When the input model is warm (`Mdl.IsWarm` is `true`), `updateMetricsAndFit` overwrites previously computed metrics, stored in the `Metrics` property, with the new values. Otherwise, `updateMetricsAndFit` stores `NaN` values in `Metrics` instead.
2. `updateMetricsAndFit` fits the modified model to the incoming data by following this procedure:
   a. Initialize the solver with the configurations and linear model coefficient and bias estimates of the input model `Mdl`.
   b. Fit the model to the data, and store the updated coefficient estimates and configurations in the output model `Mdl`.

The input and output models have the same data type.

`Mdl = updateMetricsAndFit(Mdl,X,Y,Name,Value)` uses additional options specified by one or more name-value pair arguments. For example, you can specify that the columns of the predictor data matrix correspond to observations, and set observation weights.

Examples

Update Performance Metrics and Train Model on Data Stream

Create a default incremental linear SVM model for binary classification.

```
Mdl = incrementalClassificationLinear()
```

```
Mdl =
  incrementalClassificationLinear

            IsWarm: 0
           Metrics: [1x2 table]
        ClassNames: [1x0 double]
    ScoreTransform: 'none'
              Beta: [0x1 double]
              Bias: 0
           Learner: 'svm'


  Properties, Methods
```

`Mdl` is an `incrementalClassificationLinear` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Responses can be one of five classes: Sitting, Standing, Walking, Running, or Dancing. Dichotomize the response by identifying whether the subject is moving (`actid` > 2).

```
Y = Y > 2;
```

Fit the incremental model to the training data by using the `updateMetricsAndfit` function. At each iteration:

- Simulate a data stream by processing a chunk of 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store β_1, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

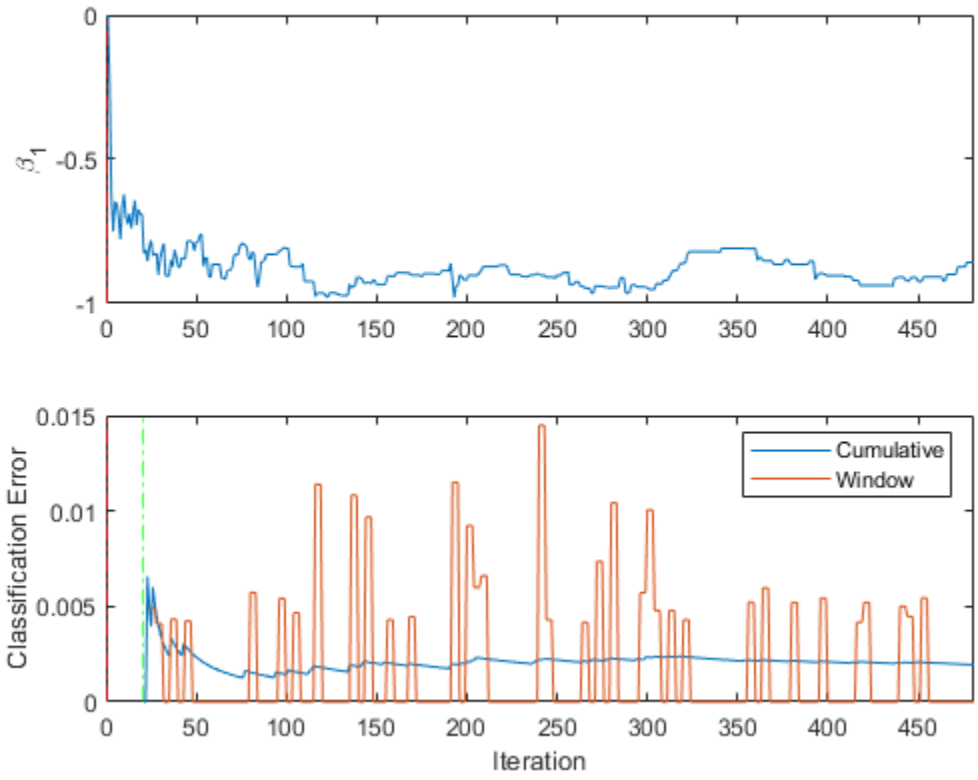
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
ce = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta1 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
```

```
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    ce{j,:} = Mdl.Metrics{"ClassificationError",:};
    beta1(j + 1) = Mdl.Beta(1);
end
```

`Mdl` is an `incrementalClassificationLinear` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit` checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and β_1 evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(beta1)
ylabel('\beta_1')
xlim([0 nchunk]);
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
subplot(2,1,2)
h = plot(ce.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
xline(Mdl.EstimationPeriod/numObsPerChunk,'r-.');
xline((Mdl.EstimationPeriod + Mdl.MetricsWarmupPeriod)/numObsPerChunk,'g-.');
legend(h,ce.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fits β_1 during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Specify Orientation of Observations and Observation Weights

Train a linear regression model by using `fitrlinear`, convert it to an incremental learner, track its performance, and fit it to streaming data. Carry over training options from traditional to incremental learning.

Load and Preprocess Data

Load the 2015 NYC housing data set, and shuffle the data. For more details on the data, see NYC Open Data.

```
load NYCHousing2015
rng(1); % For reproducibility
n = size(NYCHousing2015,1);
idxshuff = randsample(n,n);
NYCHousing2015 = NYCHousing2015(idxshuff,:);
```

Suppose that the data collected from Manhattan (`BOROUGH = 1`) was collected using a new method that doubles its quality. Create a weight variable that attributes 2 to observations collected from Manhattan, and 1 to all other observations.

```
n = size(NYCHousing2015,1);
NYCHousing2015.W = ones(n,1) + (NYCHousing2015.BOROUGH == 1);
```

Extract the response variable `SALEPRICE` from the table. For numerical stability, scale `SALEPRICE` by `1e6`.

```
Y = NYCHousing2015.SALEPRICE/1e6;
NYCHousing2015.SALEPRICE = [];
```

Create dummy variable matrices from the categorical predictors.

```
catvars = ["BOROUGH" "BUILDINGCLASSCATEGORY" "NEIGHBORHOOD"];
dumvarstbl = varfun(@(x)dummyvar(categorical(x)),NYCHousing2015,...
    'InputVariables',catvars);
dumvarmat = table2array(dumvarstbl);
NYCHousing2015(:,catvars) = [];
```

Treat all other numeric variables in the table as linear predictors of sales price. Concatenate the matrix of dummy variables to the rest of the predictor data. Transpose the resulting predictor matrix.

```
idxnum = varfun(@isnumeric,NYCHousing2015,'OutputFormat','uniform');
X = [dumvarmat NYCHousing2015{:,idxnum}]';
```

Train Linear Regression Model

Fit a linear regression model to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitrlinear(X(:,idxtt),Y(idxtt),'ObservationsIn','columns',...
    'Weights',NYCHousing2015.W(idxtt))
```

```
TTMdl =
  RegressionLinear
         ResponseName: 'Y'
    ResponseTransform: 'none'
                 Beta: [313x1 double]
                 Bias: 0.1116
               Lambda: 2.1977e-05
              Learner: 'svm'


  Properties, Methods
```

`TTMdl` is a `RegressionLinear` model object representing a traditionally trained linear regression model.

Convert Trained Model

Convert the traditionally trained linear regression model to a linear regression model for incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl)
```

```
IncrementalMdl =
  incrementalRegressionLinear

              IsWarm: 1
             Metrics: [1x2 table]
    ResponseTransform: 'none'
                Beta: [313x1 double]
                Bias: 0.1116
             Learner: 'svm'


  Properties, Methods
```

Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetrics` and `fit` functions. At each iteration:

1. Simulate a data stream by processing a chunk of 500 observations.
2. Call `updateMetricsAndFit` to update the cumulative and window epsilon insensitive loss of the model given the incoming chunk of observations, and then fit the model to the data. Overwrite the previous incremental model to update the losses in the `Metrics` property. Specify that the observations are oriented in columns, and specify the observation weights.
3. Store the losses and last estimated coefficient β_{313}.

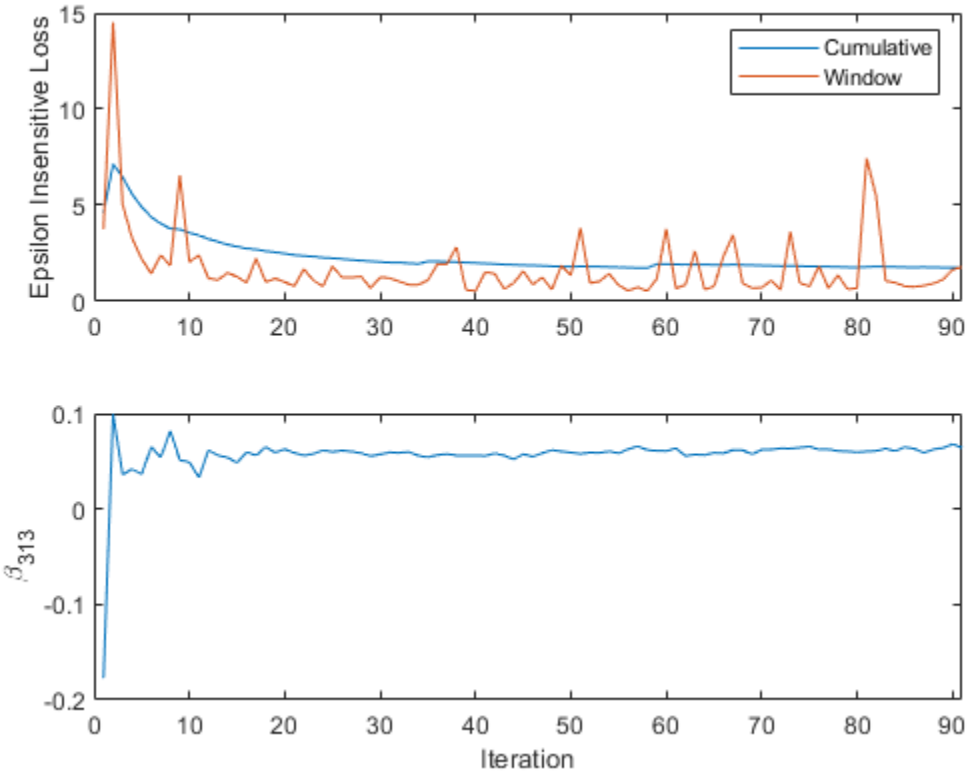
```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 500;
nchunk = floor(nil/numObsPerChunk);
ei = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
beta313 = [IncrementalMdl.Beta(end); zeros(nchunk,1)];
Xil = X(:,idxil);
Yil = Y(idxil);
Wil = NYCHousing2015.W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(:,idx),Yil(idx),...
        'ObservationsIn','columns','Weights',Wil(idx));
    ei{j,:} = IncrementalMdl.Metrics{"EpsilonInsensitiveLoss",:};
    beta313(j + 1) = IncrementalMdl.Beta(end);
end
```

`IncrementalMdl` is an `incrementalRegressionLinear` model object trained on all the data in the stream.

Plot a trace plot of the performance metrics and estimated coefficient β_{313}.

```
figure;
subplot(2,1,1)
h = plot(ei.Variables);
xlim([0 nchunk]);
```

```
ylabel('Epsilon Insensitive Loss')
legend(h,ei.Properties.VariableNames)
subplot(2,1,2)
plot(beta313)
ylabel('\beta_{313}')
xlim([0 nchunk]);
xlabel('Iteration')
```



The cumulative loss gradually changes with each iteration (chunk of 500 observations), whereas the window loss jumps. Because the metrics window is 200 by default, `updateMetricsAndFit` measures the performance based on the latest 200 observations in each 500 observation chunk.

β_{313} changes, but levels off quickly, as `fit` processes chunks of observations.

Input Arguments

`Mdl` — Incremental learning model whose performance is measured and is fit to data

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Incremental learning model whose performance is measured and then the model is fit to data, specified as an `incrementalClassificationLinear` or `incrementalRegressionLinear` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

If `Mdl.IsWarm` is `false`, `updateMetricsAndFit` does not track the performance of the model. For more details, see “Algorithms” on page 33-6539.

`X` — Chunk of predictor data

floating-point matrix

Chunk of predictor data with which to measure the model performance and then to fit the model to, specified as a floating-point matrix of *n* observations and `Mdl.NumPredictors` predictor variables. The value of the `'ObservationsIn'` name-value pair argument determines the orientation of the variables and observations.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `updateMetricsAndFit` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `updateMetricsAndFit` issues an error.
- `updateMetricsAndFit` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see “Dummy Variables” on page 2-48.

Data Types: `single` | `double`

`Y` — Chunk of labels

categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels with which to measure the model performance and then fit the model to, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors for classification problems; or a floating-point vector for regression problems.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

For classification problems:

- `updateMetricsAndFit` supports binary classification only.
- When the `ClassNames` property of the input model `Mdl` is nonempty, the following conditions apply:
  - If `Y` contains a label that is not a member of `Mdl.ClassNames`, `updateMetricsAndFit` issues an error.
  - The data type of `Y` and `Mdl.ClassNames` must be the same.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Note

- If an observation (predictor or label) or weight `Weight` contains at least one missing (`NaN`) value, `updateMetricsAndFit` ignores the observation. Consequently, `updateMetricsAndFit` uses fewer than *n* observations to compute the model performance.
- The chunk size *n* and the stochastic gradient descent (SGD) hyperparameter batch size (`Mdl.BatchSize`) can be different values. If *n* < `Mdl.BatchSize`, `updateMetricsAndFit` uses the *n* available observations when it applies SGD.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'ObservationsIn','columns','Weights',W` specifies that the columns of the predictor matrix correspond to observations, and the vector `W` contains observation weights to apply during incremental learning.

`ObservationsIn` — Predictor data observation dimension

`'rows'` (default) | `'columns'`

Predictor data observation dimension, specified as the comma-separated pair consisting of `'ObservationsIn'` and `'columns'` or `'rows'`.

Data Types: `char` | `string`

`Weights` — Chunk of observation weights

floating-point vector of positive values

Chunk of observation weights, specified as the comma-separated pair consisting of `'Weights'` and a floating-point vector of positive values. `updateMetricsAndFit` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(`*n*`,1)`.

For more details, including normalization schemes, see “Observation Weights” on page 33-6540.

Data Types: `double` | `single`

Output Arguments

`Mdl` — Updated incremental learning model

`incrementalClassificationLinear` model object | `incrementalRegressionLinear` model object

Updated incremental learning model, returned as an incremental learning model object of the same data type as the input model `Mdl`, either `incrementalClassificationLinear` or `incrementalRegressionLinear`.

When you call `updateMetricsAndFit`, the following conditions apply:

- If the model is not warm, `updateMetricsAndFit` does not compute performance metrics. As a result, the `Metrics` property of `Mdl` remains completely composed of `NaN` values. For more details, see “Algorithms” on page 33-6539.
- If `Mdl.EstimationPeriod` > 0, `updateMetricsAndFit` estimates hyperparameters using the first `Mdl.EstimationPeriod` observations passed to it; the function does not train the input model using that data. However, if an incoming chunk of *n* observations is greater than or equal to the number of observations remaining in the estimation period *m*, `updateMetricsAndFit` estimates hyperparameters using the first *n* – *m* observations, and fits the input model to the remaining *m* observations. Consequently, the software updates the `Beta` and `Bias` properties, hyperparameter properties, and recordkeeping properties such as `NumTrainingObservations`.

For classification problems, if the `ClassNames` property of the input model `Mdl` is an empty array, `updateMetricsAndFit` sets the `ClassNames` property of the output model `Mdl` to `unique(Y)`.

Algorithms

Performance Metrics

- `updateMetricsAndFit` tracks model performance metrics, specified by the row labels of the table in `Mdl.Metrics`, from new data when the incremental model is warm (`IsWarm` property is `true`). An incremental model is warm after an incremental fitting, like `updateMetricsAndFit`, fits the incremental model to `Mdl.MetricsWarmupPeriod` observations, which is the metrics warm-up period.

  If `Mdl.EstimationPeriod` > 0, `updateMetricsAndFit` estimates hyperparameters before fitting the model to data. Therefore, the functions must process an additional `EstimationPeriod` observations before the model starts the metrics warm-up period.
- The `Metrics` property of the incremental model stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetricsAndFit` updates the metrics at the following frequencies:

  - `Cumulative` — The function computes cumulative metrics since the start of model performance tracking. The function updates metrics every time you call the function and bases the calculation on the entire supplied data set.
  - `Window` — The function computes metrics based on all observations within a window determined by the `Mdl.MetricsWindowSize` property. `Mdl.MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `Mdl.MetricsWindowSize` is 20, the function computes metrics based on the last 20 observations in the supplied data (`X((end – 20 + 1):end,:)` and `Y((end – 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `Mdl.MetricsWindowSize` and a buffer of observation weights.
    2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
    3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes

a batch of observations, the latest incoming `Mdl.MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `Mdl.MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the function uses the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

Observation Weights

For classification problems, if the prior class probability distribution is known (in other words, the prior distribution is not empirical), `updateMetricsAndFit` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that observation weights are the respective prior class probabilities by default.

For regression problems or if the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `updateMetricsAndFit`.

References

[1] Bifet, Albert, Ricard Gavaldá, Geoffrey Holmes, and Bernhard Pfahringer. Machine Learning for Data Streams with Practical Example in MOA. Cambridge, MA: The MIT Press, 2007.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- Use `saveLearnerForCoder`, `loadLearnerForCoder`, and `codegen` to generate code for the `updateMetricsAndFit` function. Save a trained model by using `saveLearnerForCoder`. Define an entry-point function that loads the saved model by using `loadLearnerForCoder` and calls the `updateMetricsAndFit` function. Then use `codegen` to generate code for the entry-point function.
- To generate single-precision C/C++ code for updateMetricsAndFit, specify the name-value argument `'DataType','single'` when you call the `loadLearnerForCoder` function.
- This table contains notes about the arguments of `updateMetricsAndFit`. Arguments not included in this table are fully supported.

Argument	Notes and Limitations
`Mdl`	For usage notes and limitations of the model object, see `incrementalClassificationLinear` or `incrementalRegressionLinear`.

Argument	Notes and Limitations
X	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `Y`. • The number of predictor variables must equal to `Mdl.NumPredictors`. • `X` must be `single` or `double`.
Y	• Batch-to-batch, the number of observations can be a variable size, but must equal the number of observations in `X`. • For classification problems, all labels in `Y` must be represented in `Mdl.ClassNames`. • `Y` and `Mdl.ClassNames` must have the same data type.

- The following restrictions apply:
  - If you configure `Mdl` to shuffle data (`Mdl.Shuffle` is `true`, or `Mdl.Solver` is `'sgd'` or `'asgd'`), the `updateMetricsAndFit` function randomly shuffles each incoming batch of observations before it fits the model to the batch. The order of the shuffled observations might not match the order generated by MATLAB. Therefore, the fitted coefficients computed in MATLAB and by the generated code might not be equal.
  - Use a homogeneous data type for all floating-point input arguments and object properties, specifically, either `single` or `double`.

For more information, see “Introduction to Code Generation” on page 32-2.

See Also

Objects
`incrementalClassificationLinear` | `incrementalRegressionLinear`

Functions
`fit` | `updateMetrics`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19
“Initialize Incremental Learning Model from Logistic Regression Model Trained in Classification Learner” on page 26-36
“Initialize Incremental Learning Model from SVM Regression Model Trained in Regression Learner” on page 26-30

Introduced in R2020b

updateMetricsAndFit

Update performance metrics in naive Bayes classification model for incremental learning given new data and train model

Syntax

```
Mdl = updateMetricsAndFit(Mdl,X,Y)
Mdl = updateMetricsAndFit(Mdl,X,Y,'Weights',Weights)
```

Description

Given streaming data, `updateMetricsAndFit` first evaluates the performance of a configured naive Bayes classification model for incremental learning (`incrementalClassificationNaiveBayes` object) by calling `updateMetrics` on incoming data. Then `updateMetricsAndFit` fits the model to that data by calling `fit`. In other words, `updateMetricsAndFit` performs prequential evaluation because it treats each incoming chunk of data as a test set, and tracks performance metrics measured cumulatively and over a specified window [1].

`updateMetricsAndFit` provides a simple way to update model performance metrics and train the model on each chunk of data. Alternatively, you can perform the operations separately by calling `updateMetrics` and then `fit`, which allows for more flexibility (for example, you can decide whether you need to train the model based on its performance on a chunk of data).

`Mdl = updateMetricsAndFit(Mdl,X,Y)` returns a naive Bayes classification model for incremental learning `Mdl`, which is the input naive Bayes classification model for incremental learning `Mdl` with the following modifications:

1. `updateMetricsAndFit` measures the model performance on the incoming predictor and response data, `X` and `Y` respectively. When the input model is warm (`Mdl.IsWarm` is `true`), `updateMetricsAndFit` overwrites previously computed metrics, stored in the `Metrics` property, with the new values. Otherwise, `updateMetricsAndFit` stores `NaN` values in `Metrics` instead.
2. `updateMetricsAndFit` fits the modified model to the incoming data by updating the conditional posterior mean and standard deviation of each predictor variable, given the class, and stores the new estimates, among other configurations, in the output model `Mdl`.

The input and output models have the same data type.

`Mdl = updateMetricsAndFit(Mdl,X,Y,'Weights',Weights)` specifies observation weights `Weights`.

Examples

Update Performance Metrics and Train Model on Data Stream

Create a naive Bayes classification model for incremental learning by calling `incrementalClassificationNaiveBayes` and specifying a maximum of 5 expected classes in the data.

```
Mdl = incrementalClassificationNaiveBayes('MaxNumClasses',5)

Mdl =
  incrementalClassificationNaiveBayes

                   IsWarm: 0
                  Metrics: [1×2 table]
               ClassNames: [1×0 double]
           ScoreTransform: 'none'
        DistributionNames: 'normal'
    DistributionParameters: {}


  Properties, Methods
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object. All its properties are read-only.

`Mdl` must be fit to data before you can use it to perform any other operations.

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
n = numel(actid);
rng(1); % For reproducibility
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Implement incremental learning by performing the following actions at each iteration:

- Simulate a data stream by processing a chunk of 50 observations.
- Overwrite the previous incremental model with a new one fitted to the incoming observation.
- Store the conditional mean of the first predictor in the first class μ_{11}, the cumulative metrics, and the window metrics to see how they evolve during incremental learning.

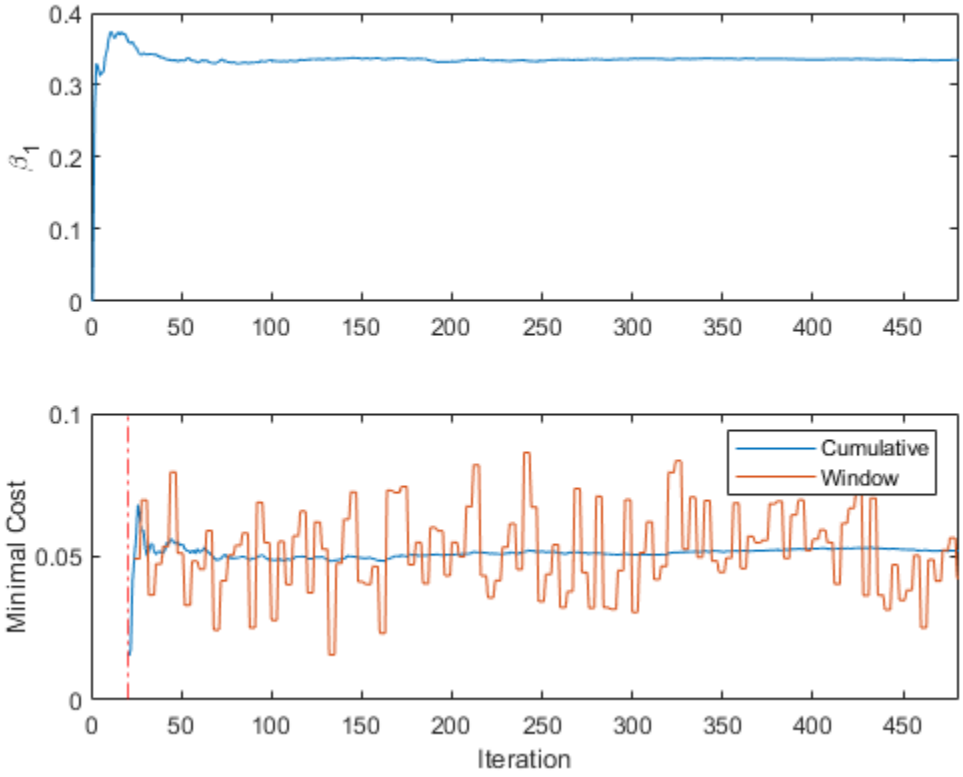
```
% Preallocation
numObsPerChunk = 50;
nchunk = floor(n/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
mu11 = zeros(nchunk,1);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(n,numObsPerChunk*(j-1) + 1);
    iend   = min(n,numObsPerChunk*j);
    idx = ibegin:iend;
    Mdl = updateMetricsAndFit(Mdl,X(idx,:),Y(idx));
    mc{j,:} = Mdl.Metrics{"MinimalCost",:};
    mu11(j + 1) = Mdl.DistributionParameters{1,1}(1);
end
```

`Mdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream. During incremental learning and after the model is warmed up, `updateMetricsAndFit`

checks the performance of the model on the incoming observation, and then fits the model to that observation.

To see how the performance metrics and μ_{11} evolved during training, plot them on separate subplots.

```
figure;
subplot(2,1,1)
plot(mu11)
ylabel('\beta_1')
xlim([0 nchunk]);
subplot(2,1,2)
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Minimal Cost')
xline(Mdl.MetricsWarmupPeriod/numObsPerChunk,'r-.');
legend(h,mc.Properties.VariableNames)
xlabel('Iteration')
```



The plot suggests that `updateMetricsAndFit` does the following:

- Fits μ_{11} during all incremental learning iterations.
- Compute performance metrics after the metrics warm-up period only.
- Compute the cumulative metrics during each iteration.
- Compute the window metrics after processing 500 observations.

Specify Observation Weights

Train a naive Bayes classification model by using `fitcnb`, convert it to an incremental learner, track its performance on streaming data and fit it to the data in one call. Specify observation weights.

Load and Preprocess Data

Load the human activity data set. Randomly shuffle the data.

```
load humanactivity
rng(1); % For reproducibility
n = numel(actid);
idx = randsample(n,n);
X = feat(idx,:);
Y = actid(idx);
```

For details on the data set, enter `Description` at the command line.

Suppose that the data collected when the subject was not moving (`Y <= 2`) has double the quality than when the subject was moving. Create a weight variable that attributes 2 to observations collected from a still subject, and 1 to a moving subject.

```
W = ones(n,1) + ~Y;
```

Train Naive Bayes Classification Model

Fit a naive Bayes classification model to a random sample of half the data.

```
idxtt = randsample([true false],n,true);
TTMdl = fitcnb(X(idxtt,:),Y(idxtt),'Weights',W(idxtt))
```

```
TTMdl =
  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: [1 2 3 4 5]
           ScoreTransform: 'none'
          NumObservations: 12053
         DistributionNames: {1×60 cell}
    DistributionParameters: {5×60 cell}


  Properties, Methods
```

`TTMdl` is a `ClassificationNaiveBayes` model object representing a traditionally trained naive Bayes classification model.

Convert Trained Model

Convert the traditionally trained model to a naive Bayes classification for incremental learning. Specify tracking the misclassification error rate during incremental learning.

```
IncrementalMdl = incrementalLearner(TTMdl,'Metrics',"classiferror")
```

```
IncrementalMdl =
  incrementalClassificationNaiveBayes
```

```
                  IsWarm: 1
                 Metrics: [2×2 table]
              ClassNames: [1 2 3 4 5]
          ScoreTransform: 'none'
       DistributionNames: {1×60 cell}
  DistributionParameters: {5×60 cell}


  Properties, Methods
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model. Because class names are specified in `Mdl.ClassNames`, labels encountered during incremental learning must be in `Mdl.ClassNames`.

Separately Track Performance Metrics and Fit Model

Perform incremental learning on the rest of the data by using the `updateMetricsAndfit` function. At each iteration:

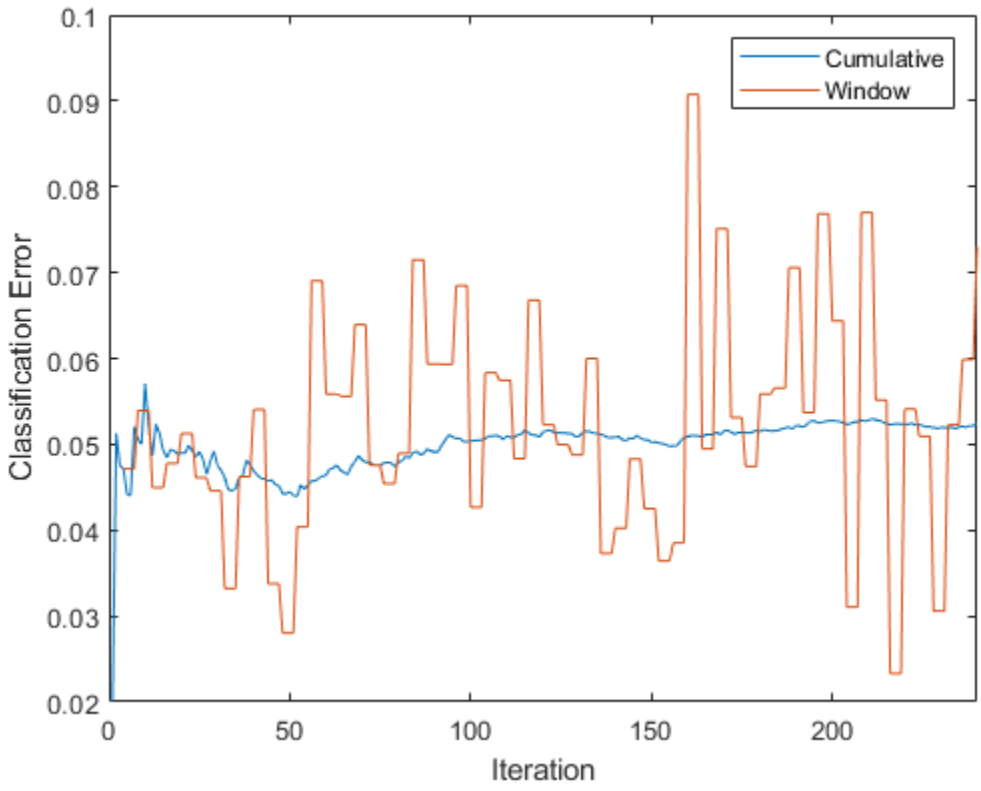
1. Simulate a data stream by processing 50 observations at a time.
2. Call `updateMetricsAndFit` to update the cumulative and window performance metrics of the model given the incoming chunk of observations, and then fit the model to the data. Overwrite the previous incremental model to update the losses in the `Metrics` property. Specify the observation weights.
3. Store the misclassification error rate.

```
% Preallocation
idxil = ~idxtt;
nil = sum(idxil);
numObsPerChunk = 50;
nchunk = floor(nil/numObsPerChunk);
mc = array2table(zeros(nchunk,2),'VariableNames',["Cumulative" "Window"]);
Xil = X(idxil,:);
Yil = Y(idxil);
Wil = W(idxil);

% Incremental fitting
for j = 1:nchunk
    ibegin = min(nil,numObsPerChunk*(j-1) + 1);
    iend   = min(nil,numObsPerChunk*j);
    idx = ibegin:iend;
    IncrementalMdl = updateMetricsAndFit(IncrementalMdl,Xil(idx,:),Yil(idx),...
        'Weights',Wil(idx));
    mc{j,:} = IncrementalMdl.Metrics{"ClassificationError",:};
end
```

`IncrementalMdl` is an `incrementalClassificationNaiveBayes` model object trained on all the data in the stream.

Plot a trace plot of the misclassification error rate.

```
h = plot(mc.Variables);
xlim([0 nchunk]);
ylabel('Classification Error')
legend(h,mc.Properties.VariableNames)
xlabel('Iteration')
```



The cumulative loss initially jump, but stabilizes around 0.05, whereas the window loss jumps.

Input Arguments

`Mdl` — Naive Bayes classification model for incremental learning whose performance is measured and is fit to data

`incrementalClassificationNaiveBayes` model object

Naive Bayes classification model for incremental learning whose performance is measured and then the model is fit to data, specified as an `incrementalClassificationNaiveBayes` model object. You can create `Mdl` directly or by converting a supported, traditionally trained machine learning model using the `incrementalLearner` function. For more details, see the corresponding reference page.

If `Mdl.IsWarm` is `false`, `updateMetricsAndFit` does not track the performance of the model. For more details, see “Performance Metrics” on page 33-6549.

`X` — Chunk of predictor data

floating-point matrix

Chunk of predictor data with which to measure the model performance and then to fit the model to, specified as an *n*-by-`Mdl.NumPredictors` floating-point matrix.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`.

Note

- If `Mdl.NumPredictors` = 0, `updateMetricsAndFit` infers the number of predictors from `X`, and sets the congruent property of the output model. Otherwise, if the number of predictor variables in the streaming data changes from `Mdl.NumPredictors`, `updateMetricsAndFit` issues an error.
- `updateMetricsAndFit` supports only floating-point input predictor data. If the input model `Mdl` represents a converted, traditionally trained model fit to categorical data, use `dummyvar` to convert each categorical variable to a numeric matrix of dummy variables, and concatenate all dummy variable matrices and any other numeric predictors. For more details, see "Dummy Variables" on page 2-48.

Data Types: `single` | `double`

Y — Chunk of labels

categorical array | character array | string array | logical vector | floating-point vector | cell array of character vectors

Chunk of labels with which to measure the model performance and then fit the model to, specified as a categorical, character, or string array, logical or floating-point vector, or cell array of character vectors.

The length of the observation labels `Y` and the number of observations in `X` must be equal; `Y(`*j*`)` is the label of observation *j* (row or column) in `X`. `updateMetricsAndFit` issues an error when at least one of the conditions is met:

- `Y` contains a newly encountered label and the maximum number of classes has been reached previously (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`).
- The data types of `Y` and `Mdl.ClassNames` are different.

Data Types: `char` | `string` | `cell` | `categorical` | `logical` | `single` | `double`

Weights — Chunk of observation weights

floating-point vector of positive values

Chunk of observation weights, specified as a floating-point vector of positive values. `updateMetricsAndFit` weighs the observations in `X` with the corresponding values in `Weights`. The size of `Weights` must equal *n*, which is the number of observations in `X`.

By default, `Weights` is `ones(`*n*`,1)`.

For more details, including normalization schemes, see "Observation Weights" on page 33-6550.

Data Types: `double` | `single`

Note

If an observation (predictor or label) or weight `Weight` contains at least one missing (`NaN`) value, `updateMetricsAndFit` ignores the observation. Consequently, `updateMetricsAndFit` uses fewer than *n* observations to compute the model performance.

Output Arguments

`Mdl` — Updated naive Bayes classification model for incremental learning

`incrementalClassificationNaiveBayes` model object

Updated naive Bayes classification model for incremental learning, returned as an incremental learning model object of the same data type as the input model `Mdl`, `incrementalClassificationNaiveBayes`.

If the model is not warm, `updateMetricsAndFit` does not compute performance metrics. As a result, the `Metrics` property of `Mdl` remains completely composed of `NaN` values. If the model is warm, `updateMetricsAndFit` computes the cumulative and window performance metrics on the new data `X` and `Y`, and overwrites the corresponding elements of `Mdl.Metrics`. All other properties of the input model `Mdl` carry over to the output model `Mdl`. For more details, see “Performance Metrics” on page 33-6549.

If the `ClassNames` property of the input model `Mdl` is an empty array, `updateMetricsAndFit` sets the `ClassNames` property of the output model `Mdl` to `unique(Y)`. If the maximum number of classes is not reached, `updateMetricsAndFit` appends to `Mdl.ClassNames` any newly encountered labels in `Y`.

Algorithms

Performance Metrics

- `updateMetricsAndFit` tracks model performance metrics, specified by the row labels of the table in `Mdl.Metrics`, from new data when the incremental model is warm (`IsWarm` property is `true`). An incremental model is warm when an incremental fitting, like `updateMetricsAndFit` performs both of the following actions:
  - Fit the incremental model to `Mdl.MetricsWarmupPeriod` observations, which is the metrics warm-up period.
  - Fit the incremental model to all expected classes (see `MaxNumClasses` and `ClassNames` arguments of `incrementalClassificationNaiveBayes`)
- `Mdl.Metrics` stores two forms of each performance metric as variables (columns) of a table, `Cumulative` and `Window`, with individual metrics in rows. When the incremental model is warm, `updateMetricsAndFit` updates the metrics at the following frequencies:
  - `Cumulative` — The function computes cumulative metrics since the start of model performance tracking. The function updates metrics every time you call the function and bases the calculation on the entire supplied data set.
  - `Window` — The function computes metrics based on all observations within a window determined by the `Mdl.MetricsWindowSize` property. `Mdl.MetricsWindowSize` also determines the frequency at which the software updates `Window` metrics. For example, if `Mdl.MetricsWindowSize` is 20, the function computes metrics based on the last 20 observations in the supplied data (`X((end − 20 + 1):end,:)` and `Y((end − 20 + 1):end)`).

    Incremental functions that track performance metrics within a window use the following process:

    1. For each specified metric, store a buffer of length `Mdl.MetricsWindowSize` and a buffer of observation weights.

2. Populate elements of the metrics buffer with the model performance based on batches of incoming observations, and store corresponding observations weights in the weights buffer.
3. When the buffer is filled, overwrite `Mdl.Metrics.Window` with the weighted average performance in the metrics window. If the buffer is overfilled when the function processes a batch of observations, the latest incoming `Mdl.MetricsWindowSize` observations enter the buffer, and the earliest observations are removed from the buffer. For example, suppose `Mdl.MetricsWindowSize` is 20, the metrics buffer has 10 values from a previously processed batch, and 15 values are incoming. To compose the length 20 window, the function uses the measurements from the 15 incoming observations and the latest 5 measurements from the previous batch.

Observation Weights

For each conditional predictor distribution, `updateMetricsAndFit` computes the weighted average and standard deviation.

If the prior class probability distribution is known (in other words, the prior distribution is not empirical), `updateMetricsAndFit` normalizes observation weights to sum to the prior class probabilities in the respective classes. This action implies that the default observation weights are the respective prior class probabilities.

If the prior class probability distribution is empirical, the software normalizes the specified observation weights to sum to 1 each time you call `updateMetricsAndFit`.

References

[1] Bifet, Albert, Ricard Gavaldá, Geoffrey Holmes, and Bernhard Pfahringer. Machine Learning for Data Streams with Practical Example in MOA. Cambridge, MA: The MIT Press, 2007.

See Also

Objects
`incrementalClassificationNaiveBayes`

Functions
`fit` | `updateMetrics`

Topics
“Incremental Learning Overview” on page 26-2
“Configure Incremental Learning Model” on page 26-8
“Implement Incremental Learning for Classification Using Succinct Workflow” on page 26-19

Introduced in R2021a

upperparams

Upper Pareto tail parameters

Syntax

```
params = upperparams(pd)
```

Description

`params = upperparams(pd)` returns the two-element vector `params`, which includes the shape and scale parameters of the generalized Pareto distribution (GPD) in the upper tail of `pd`.

`upperparams` does not return the location parameter of the GPD. The location parameter is the quantile value corresponding to the upper tail cumulative probability. Use the `boundary` function to return the location parameter.

Examples

Parameters of Upper Pareto Tail

Generate a sample data set and fit a piecewise distribution with Pareto tails to the data by using `paretotails`. Find the distribution parameters of the upper Pareto tail by using the object function `upperparams`.

Generate a sample data set containing 20% outliers.

```
rng('default');  % For reproducibility
left_tail = -exprnd(1,100,1);
right_tail = exprnd(5,100,1);
center = randn(800,1);
x = [left_tail;center;right_tail];
```

Create a `paretotails` object by fitting a piecewise distribution to `x`. Specify the boundaries of the tails using the lower and upper tail cumulative probabilities so that a fitted object consists of the empirical distribution for the middle 80% of the data set and GPDs for the lower and upper 10% of the data set.

```
pd = paretotails(x,0.1,0.9)
```

```
pd =
Piecewise distribution with 3 segments
      -Inf < x < -1.33251    (0 < p < 0.1): lower tail, GPD(-0.0063504,0.567017)
   -1.33251 < x < 1.80149  (0.1 < p < 0.9): interpolated empirical cdf
        1.80149 < x < Inf    (0.9 < p < 1): upper tail, GPD(0.24874,3.00974)
```

Return the shape and scale parameters of the fitted GPD of the upper tail by using the `upperparams` function.

```
params = upperparams(pd)
```

```
params = 1×2

    0.2487    3.0097
```

You can also get the upper Pareto tail parameters by using the `UpperParameters` property. Access the `UpperParameters` property by using dot notation.

```
pd.UpperParameters

ans = 1×2

    0.2487    3.0097
```

The location parameter of the GPD is equal to the quantile value of the upper tail cumulative probability. Return the location parameter by using the `boundary` function.

```
[p,q] = boundary(pd)

p = 2×1

    0.1000
    0.9000


q = 2×1

   -1.3325
    1.8015
```

The values in `p` are the cumulative probabilities at the boundaries, and the values in `q` are the corresponding quantiles. `q(1)` is the location parameter of the GPD of the upper tail.

Use the `lowerparams` function or the `LowerParameters` property to get the lower Pareto tail parameters.

Input Arguments

pd — Piecewise distribution with Pareto tails

`paretotails` object

Piecewise distribution with Pareto tails, specified as a `paretotails` object.

See Also

`boundary` | `gpfit` | `lowerparams` | `nsegments` | `paretotails` | `segment`

Topics

“Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
“Nonparametric and Empirical Probability Distributions” on page 5-30
“Nonparametric Estimates of Cumulative Distribution Functions and Their Inverses” on page 5-181
“Generalized Pareto Distribution” on page B-59

Introduced in R2007a

validatedUpdateInputs

Package: `classreg.learning.coder.config.svm`

Validate and extract machine learning model parameters to update

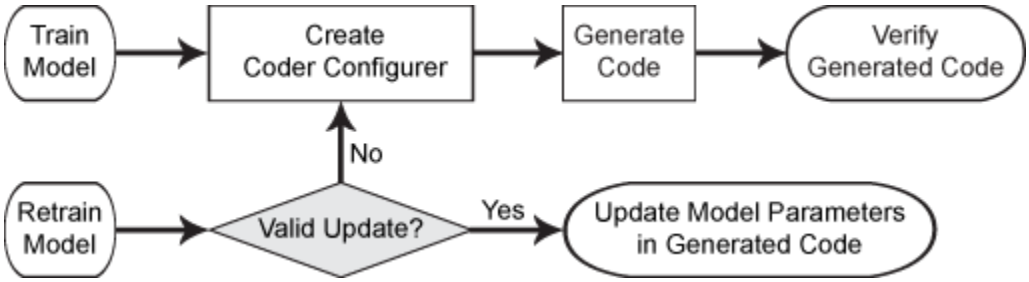
Syntax

```
params = validatedUpdateInputs(configurer,retrainedMdl)
```

Description

Generate C/C++ code for the `predict` and `update` functions of a machine learning model by using a coder configurer object. Create this object by using `learnerCoderConfigurer` and its object function `generateCode`. After retraining the model with new data or settings, you can update model parameters in the generated code without having to regenerate the code. Use `validatedUpdateInputs` to validate and extract the model parameters to update. This function helps you identify potential problems before you update the model parameters in the generated code. You can use the output of `validatedUpdateInputs`, the validated parameters, as an input argument of the `update` function to update model parameters.

This flow chart shows the code generation workflow using a coder configurer. Use `validatedUpdateInputs` for the highlighted step.



`params = validatedUpdateInputs(configurer,retrainedMdl)` returns the validated machine learning model parameters to update. `validatedUpdateInputs` detects the modified model parameters in `retrainedMdl` and validates whether they satisfy the coder attributes stored in `configurer`.

Examples

Update Parameters of SVM Classification Model in Generated Code

Train a SVM model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`). Train a binary SVM classification model using the first 50 observations.

```
load ionosphere
Mdl = fitcsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `ClassificationSVM` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted labels and scores.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:),'NumOutputs',2);
```

`configurer` is a `ClassificationSVMCoderConfigurer` object, which is a coder configurer of a `ClassificationSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM classification model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 34];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 34 predictors, so the value of the `SizeVector` attribute must be 34 and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 34];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

```
configurer.SupportVectors.VariableDimensions = [true false];
```

```
VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
VariableDimensions attribute for SupportVectorLabels has been modified to satisfy configuration c
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM classification model (`Mdl`) with default settings.

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `ClassificationSVMModel` for the two entry-point functions in the `codegen\mex\ClassificationSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[label,score] = predict(Mdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
```

Compare `label` and `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`score_mex` might include round-off differences compared with `score`. In this case, compare `score_mex` and `score`, allowing a small tolerance.

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `score` and `score_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitcsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[label,score] = predict(retrainedMdl,X);
[label_mex,score_mex] = ClassificationSVMModel('predict',X);
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(score-score_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `labels` and `labels_mex` are equal, and the score values are equal within the tolerance.

Update Parameters of ECOC Classification Model in Generated Code

Train an error-correcting output codes (ECOC) model using SVM binary learners and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the ECOC model parameters. Use the object function of the coder configurer to generate C code that predicts labels for new predictor data. Then retrain the model using different settings, and update parameters in the generated code without regenerating the code.

Train Model

Load Fisher's iris data set.

```
load fisheriris
X = meas;
Y = species;
```

Create an SVM binary learner template to use a Gaussian kernel function and to standardize predictor data.

```
t = templateSVM('KernelFunction','gaussian','Standardize',true);
```

Train a multiclass ECOC model using the template `t`.

```
Mdl = fitcecoc(X,Y,'Learners',t);
```

`Mdl` is a `ClassificationECOC` object.

Create Coder Configurer

Create a coder configurer for the `ClassificationECOC` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns the first two outputs of the `predict` function, which are the predicted labels and negated average binary losses.

```
configurer = learnerCoderConfigurer(Mdl,X,'NumOutputs',2)
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]

   Code Generation Parameters:
        NumOutputs: 2
    OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

`configurer` is a `ClassificationECOCCoderConfigurer` object, which is a coder configurer of a `ClassificationECOC` object. The display shows the tunable input arguments of `predict` and `update`: `X`, `BinaryLearners`, `Prior`, and `Cost`.

Specify Coder Attributes of Parameters

Specify the coder attributes of `predict` arguments (predictor data and the name-value pair arguments `'Decoding'` and `'BinaryLoss'`) and `update` arguments (support vectors of the SVM learners) so that you can use these arguments as the input arguments of `predict` and `update` in the generated code.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 4];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains 4 predictors, so the second value of the `SizeVector` attribute must be 4 and the second value of the `VariableDimensions` attribute must be `false`.

Next, modify the coder attributes of `BinaryLoss` and `Decoding` to use the `'BinaryLoss'` and `'Decoding'` name-value pair arguments in the generated code. Display the coder attributes of `BinaryLoss`.

```
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

             Value: 'hinge'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 0
```

To use a nondefault value in the generated code, you must specify the value before generating the code. Specify the `Value` attribute of `BinaryLoss` as `'exponential'`.

```
configurer.BinaryLoss.Value = 'exponential';
configurer.BinaryLoss
```

```
ans =
  EnumeratedInput with properties:

             Value: 'exponential'
    SelectedOption: 'Built-in'
    BuiltInOptions: {1x7 cell}
        IsConstant: 1
        Tunability: 1
```

If you modify attribute values when `Tunability` is `false` (logical 0), the software sets the `Tunability` to `true` (logical 1).

Display the coder attributes of `Decoding`.

```
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: 'lossweighted'
    SelectedOption: 'Built-in'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 1
        Tunability: 0
```

Specify the `IsConstant` attribute of `Decoding` as `false` so that you can use all available values in `BuiltInOptions` in the generated code.

```
configurer.Decoding.IsConstant = false;
configurer.Decoding
```

```
ans =
  EnumeratedInput with properties:

             Value: [1x1 LearnerCoderInput]
    SelectedOption: 'NonConstant'
    BuiltInOptions: {'lossweighted'  'lossbased'}
        IsConstant: 0
        Tunability: 1
```

The software changes the `Value` attribute of `Decoding` to a `LearnerCoderInput` object so that you can use both `'lossweighted'` and `'lossbased'` as the value of `'Decoding'`. Also, the software sets the `SelectedOption` to `'NonConstant'` and the `Tunability` to `true`.

Finally, modify the coder attributes of `SupportVectors` in `BinaryLearners`. Display the coder attributes of `SupportVectors`.

```
configurer.BinaryLearners.SupportVectors
```

```
ans =
  LearnerCoderInput with properties:

            SizeVector: [54 4]
    VariableDimensions: [1 0]
              DataType: 'double'
            Tunability: 1
```

The default value of `VariableDimensions` is `[true false]` because each learner has a different number of support vectors. If you retrain the ECOC model using new data or different settings, the number of support vectors in the SVM learners can vary. Therefore, increase the upper bound of the number of support vectors.

```
configurer.BinaryLearners.SupportVectors.SizeVector = [150 4];
```

```
SizeVector attribute for Alpha has been modified to satisfy configuration constraints.
SizeVector attribute for SupportVectorLabels has been modified to satisfy configuration constrai
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` and `SupportVectorLabels` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Display the coder configurer.

```
configurer
```

```
configurer =
  ClassificationECOCCoderConfigurer with properties:

   Update Inputs:
```

```
    BinaryLearners: [1x1 ClassificationSVMCoderConfigurer]
             Prior: [1x1 LearnerCoderInput]
              Cost: [1x1 LearnerCoderInput]

   Predict Inputs:
                 X: [1x1 LearnerCoderInput]
        BinaryLoss: [1x1 EnumeratedInput]
          Decoding: [1x1 EnumeratedInput]

   Code Generation Parameters:
        NumOutputs: 2
     OutputFileName: 'ClassificationECOCModel'


  Properties, Methods
```

The display now includes `BinaryLoss` and `Decoding` as well.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the ECOC classification model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'ClassificationECOCModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `ClassificationECOCModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\ClassificationECOCModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument. Because you specified `'Decoding'` as a tunable input argument by changing the `IsConstant` attribute before generating the code, you also need to specify it in the call to the MEX function, even though `'lossweighted'` is the default value of `'Decoding'`.

```
[label,NegLoss] = predict(Mdl,X,'BinaryLoss','exponential');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
```

Compare `label` to `label_mex` by using `isequal`.

```
isequal(label,label_mex)
```

```
ans = logical
   1
```

`isequal` returns logical 1 (`true`) if all the inputs are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same labels.

`NegLoss_mex` might include round-off differences compared to `NegLoss`. In this case, compare `NegLoss_mex` to `NegLoss`, allowing a small tolerance.

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `NegLoss` and `NegLoss_mex` are equal within the tolerance `1e–8`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using a different setting. Specify `'KernelScale'` as `'auto'` so that the software selects an appropriate scale factor using a heuristic procedure.

```
t_new = templateSVM('KernelFunction','gaussian','Standardize',true,'KernelScale','auto');
retrainedMdl = fitcecoc(X,Y,'Learners',t_new);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
ClassificationECOCModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` to the outputs from the `predict` function in the updated MEX function.

```
[label,NegLoss] = predict(retrainedMdl,X,'BinaryLoss','exponential','Decoding','lossbased');
[label_mex,NegLoss_mex] = ClassificationECOCModel('predict',X,'BinaryLoss','exponential','Decodi
isequal(label,label_mex)
```

```
ans = logical
   1
```

```
find(abs(NegLoss-NegLoss_mex) > 1e-8)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `label` and `label_mex` are equal, and `NegLoss` and `NegLoss_mex` are equal within the tolerance.

Update Parameters of SVM Regression Model in Generated Code

Train a support vector machine (SVM) model using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the SVM model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the whole data set and update parameters in the generated code without regenerating the code.

Train Model

Load the `carsmall` data set and train an SVM regression model using the first 50 observations.

```
load carsmall
X = [Horsepower,Weight];
Y = MPG;
Mdl = fitrsvm(X(1:50,:),Y(1:50));
```

`Mdl` is a `RegressionSVM` object.

Create Coder Configurer

Create a coder configurer for the `RegressionSVM` model by using `learnerCoderConfigurer`. Specify the predictor data `X`. The `learnerCoderConfigurer` function uses the input `X` to configure the coder attributes of the `predict` function input.

```
configurer = learnerCoderConfigurer(Mdl,X(1:50,:));
```

`configurer` is a `RegressionSVMCoderConfigurer` object, which is a coder configurer of a `RegressionSVM` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the SVM regression model parameters so that you can update the parameters in the generated code after retraining the model. This example specifies the coder attributes of predictor data that you want to pass to the generated code and the coder attributes of the support vectors of the SVM regression model.

First, specify the coder attributes of `X` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 2];
configurer.X.VariableDimensions = [true false];
```

The size of the first dimension is the number of observations. In this case, the code specifies that the upper bound of the size is `Inf` and the size is variable, meaning that `X` can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. `X` contains two predictors, so the value of the `SizeVector` attribute must be two and the value of the `VariableDimensions` attribute must be `false`.

If you retrain the SVM model using new data or different settings, the number of support vectors can vary. Therefore, specify the coder attributes of `SupportVectors` so that you can update the support vectors in the generated code.

```
configurer.SupportVectors.SizeVector = [250 2];

SizeVector attribute for Alpha has been modified to satisfy configuration constraints.

configurer.SupportVectors.VariableDimensions = [true false];

VariableDimensions attribute for Alpha has been modified to satisfy configuration constraints.
```

If you modify the coder attributes of `SupportVectors`, then the software modifies the coder attributes of `Alpha` to satisfy configuration constraints. If the modification of the coder attributes of one parameter requires subsequent changes to other dependent parameters to satisfy configuration constraints, then the software changes the coder attributes of the dependent parameters.

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Use `generateCode` to generate code for the `predict` and `update` functions of the SVM regression model (`Mdl`) with default settings.

```
generateCode(configurer)

generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionSVMModel.mat'
Code generation successful.
```

`generateCode` generates the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively. Then `generateCode` creates a MEX function named `RegressionSVMModel` for the two entry-point functions in the `codegen\mex\RegressionSVMModel` folder and copies the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
yfit = predict(Mdl,X);
yfit_mex = RegressionSVMModel('predict',X);
```

`yfit_mex` might include round-off differences compared with `yfit`. In this case, compare `yfit` and `yfit_mex`, allowing a small tolerance.

```
find(abs(yfit-yfit_mex) > 1e-6)

ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e–6`.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrsvm(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionSVMModel('update',params)
```

Verify Generated Code

Compare the outputs from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
yfit = predict(retrainedMdl,X);
yfit_mex = RegressionSVMModel('predict',X);
find(abs(yfit-yfit_mex) > 1e-6)
```

```
ans =

  0x1 empty double column vector
```

The comparison confirms that `yfit` and `yfit_mex` are equal within the tolerance `1e-6`.

Update Parameters of Regression Tree Model in Generated Code

Train a regression tree using a partial data set and create a coder configurer for the model. Use the properties of the coder configurer to specify coder attributes of the model parameters. Use the object function of the coder configurer to generate C code that predicts responses for new predictor data. Then retrain the model using the entire data set, and update parameters in the generated code without regenerating the code.

Train Model

Load the `carbig` data set, and train a regression tree model using half of the observations.

```
load carbig
X = [Displacement Horsepower Weight];
Y = MPG;

rng('default') % For reproducibility
n = length(Y);
idxTrain = randsample(n,n/2);
XTrain = X(idxTrain,:);
YTrain = Y(idxTrain);

Mdl = fitrtree(XTrain,YTrain);
```

`Mdl` is a `RegressionTree` object.

Create Coder Configurer

Create a coder configurer for the `RegressionTree` model by using `learnerCoderConfigurer`. Specify the predictor data `XTrain`. The `learnerCoderConfigurer` function uses the input `XTrain`

to configure the coder attributes of the `predict` function input. Also, set the number of outputs to 2 so that the generated code returns predicted responses and node numbers for the predictions.

```
configurer = learnerCoderConfigurer(Mdl,XTrain,'NumOutputs',2);
```

`configurer` is a `RegressionTreeCoderConfigurer` object, which is a coder configurer of a `RegressionTree` object.

Specify Coder Attributes of Parameters

Specify the coder attributes of the regression tree model parameters so that you can update the parameters in the generated code after retraining the model.

Specify the coder attributes of the `X` property of `configurer` so that the generated code accepts any number of observations. Modify the `SizeVector` and `VariableDimensions` attributes. The `SizeVector` attribute specifies the upper bound of the predictor data size, and the `VariableDimensions` attribute specifies whether each dimension of the predictor data has a variable size or fixed size.

```
configurer.X.SizeVector = [Inf 3];
configurer.X.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

The size of the first dimension is the number of observations. Setting the value of the `SizeVector` attribute to `Inf` causes the software to change the value of the `VariableDimensions` attribute to `1`. In other words, the upper bound of the size is `Inf` and the size is variable, meaning that the predictor data can have any number of observations. This specification is convenient if you do not know the number of observations when generating code.

The size of the second dimension is the number of predictor variables. This value must be fixed for a machine learning model. Because the predictor data contains 3 predictors, the value of the `SizeVector` attribute must be `3` and the value of the `VariableDimensions` attribute must be `0`.

If you retrain the tree model using new data or different settings, the number of nodes in the tree can vary. Therefore, specify the first dimension of the `SizeVector` attribute of one of these properties so that you can update the number of nodes in the generated code: `Children`, `CutPoint`, `CutPredictorIndex`, or `NodeMean`. The software then modifies the other properties automatically.

For example, set the first value of the `SizeVector` attribute of the `NodeMean` property to `Inf`. The software modifies the `SizeVector` and `VariableDimensions` attributes of `Children`, `CutPoint`, and `CutPredictorIndex` to match the new upper bound on the number of nodes in the tree. Additionally, the first value of the `VariableDimensions` attribute of `NodeMean` changes to `1`.

```
configurer.NodeMean.SizeVector = [Inf 1];
```

```
SizeVector attribute for Children has been modified to satisfy configuration constraints.
SizeVector attribute for CutPoint has been modified to satisfy configuration constraints.
SizeVector attribute for CutPredictorIndex has been modified to satisfy configuration constraints
VariableDimensions attribute for Children has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPoint has been modified to satisfy configuration constraints
VariableDimensions attribute for CutPredictorIndex has been modified to satisfy configuration co
```

```
configurer.NodeMean.VariableDimensions
```

```
ans = 1x2 logical array

   1   0
```

Generate Code

To generate C/C++ code, you must have access to a C/C++ compiler that is configured properly. MATLAB Coder locates and uses a supported, installed compiler. You can use `mex -setup` to view and change the default compiler. For more details, see “Change Default Compiler”.

Generate code for the `predict` and `update` functions of the regression tree model (`Mdl`).

```
generateCode(configurer)
```

```
generateCode creates these files in output folder:
'initialize.m', 'predict.m', 'update.m', 'RegressionTreeModel.mat'
Code generation successful.
```

The `generateCode` function completes these actions:

- Generate the MATLAB files required to generate code, including the two entry-point functions `predict.m` and `update.m` for the `predict` and `update` functions of `Mdl`, respectively.
- Create a MEX function named `RegressionTreeModel` for the two entry-point functions.
- Create the code for the MEX function in the `codegen\mex\RegressionTreeModel` folder.
- Copy the MEX function to the current folder.

Verify Generated Code

Pass some predictor data to verify whether the `predict` function of `Mdl` and the `predict` function in the MEX function return the same predicted responses. To call an entry-point function in a MEX function that has more than one entry point, specify the function name as the first input argument.

```
[Yfit,node] = predict(Mdl,XTrain);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',XTrain);
```

Compare `Yfit` to `Yfit_mex` and `node` to `node_mex`.

```
max(abs(Yfit-Yfit_mex),[],'all')
```

```
ans = 0
```

```
isequal(node,node_mex)
```

```
ans = logical
   1
```

In general, `Yfit_mex` might include round-off differences compared to `Yfit`. In this case, the comparison confirms that `Yfit` and `Yfit_mex` are equal.

`isequal` returns logical 1 (`true`) if all the input arguments are equal. The comparison confirms that the `predict` function of `Mdl` and the `predict` function in the MEX function return the same node numbers.

Retrain Model and Update Parameters in Generated Code

Retrain the model using the entire data set.

```
retrainedMdl = fitrtree(X,Y);
```

Extract parameters to update by using `validatedUpdateInputs`. This function detects the modified model parameters in `retrainedMdl` and validates whether the modified parameter values satisfy the coder attributes of the parameters.

```
params = validatedUpdateInputs(configurer,retrainedMdl);
```

Update parameters in the generated code.

```
RegressionTreeModel('update',params)
```

Verify Generated Code

Compare the output arguments from the `predict` function of `retrainedMdl` and the `predict` function in the updated MEX function.

```
[Yfit,node] = predict(retrainedMdl,X);
[Yfit_mex,node_mex] = RegressionTreeModel('predict',X);

max(abs(Yfit-Yfit_mex),[],'all')

ans = 0

isequal(node,node_mex)

ans = logical
   1
```

The comparison confirms that the predicted responses and node numbers are equal.

Input Arguments

`configurer` — Coder configurer

coder configurer object

Coder configurer of a machine learning model, specified as a coder configurer object created by using `learnerCoderConfigurer`.

Model	Coder Configurer Object
Binary decision tree for multiclass classification	`ClassificationTreeCoderConfigurer`
SVM for one-class and binary classification	`ClassificationSVMCoderConfigurer`
Linear model for binary classification	`ClassificationLinearCoderConfigurer`
Multiclass model for SVMs and linear models	`ClassificationECOCCoderConfigurer`
Binary decision tree for regression	`RegressionTreeCoderConfigurer`
Support vector machine (SVM) regression	`RegressionSVMCoderConfigurer`
Linear regression	`RegressionLinearCoderConfigurer`

`retrainedMdl` — Retrained machine learning model

full model object | compact model object

Retrained machine learning model, specified as a full or compact model object, as given in this table of supported models.

Model	Full/Compact Model Object	Training Function
Binary decision tree for multiclass classification	`ClassificationTree`, `CompactClassificationTree`	`fitctree`
SVM for one-class and binary classification	`ClassificationSVM`, `CompactClassificationSVM`	`fitcsvm`
Linear model for binary classification	`ClassificationLinear`	`fitclinear`
Multiclass model for SVMs and linear models	`ClassificationECOC`, `CompactClassificationECOC`	`fitcecoc`
Binary decision tree for regression	`RegressionTree`, `CompactRegressionTree`	`fitrtree`
Support vector machine (SVM) regression	`RegressionSVM`, `CompactRegressionSVM`	`fitrsvm`
Linear regression	`RegressionLinear`	`fitrlinear`

Output Arguments

`params` — Validated parameters to update

structure

Validated parameters to update in the machine learning model, specified as a structure with a field for each parameter extracted from `retrainedMdl`.

The model parameters in `params` include all parameters listed in the `UpdateInputs` property of `configurer`, which is the list of tunable model parameters.

You can use `params` as an input argument of `update` to update model parameters.

Tips

- `validatedUpdateInputs` returns an error message if you modify any of the name-value pair arguments listed in this table when you retrain the model `retrainedMdl`. In this case, you cannot use `update` to update the parameters. You must generate C/C++ code again.

Model	Arguments Not Supported for Update
Binary decision tree for multiclass classification	Arguments of `fitctree` — `'ClassNames'`, `'ScoreTransform'`
SVM for one-class and binary classification	Arguments of `fitcsvm` — `'ClassNames'`, `'KernelFunction'`, `'PolynomialOrder'`, `'ScoreTransform'`, `'Standardize'`

Model	Arguments Not Supported for Update
Linear model for binary classification	Arguments of `fitclinear` — `'ClassNames'`, `'ScoreTransform'`
Multiclass model for SVMs and linear models	Arguments of `fitcecoc` — `'ClassNames'`, `'Coding'`, `'ScoreTransform'` If you specify the binary learners in `fitcecoc` as template objects (see `'Learners'`), then for each binary learner, you cannot modify the following: • Arguments of `templateSVM` — `'KernelFunction'`, `'PolynomialOrder'`, `'Standardize'` • Arguments of `templateLinear` — `'Learner'` (because modifying the model type changes the score transform of the binary learner)
Binary decision tree for regression	Arguments of `fitrtree` — `'ResponseTransform'`
SVM regression	Arguments of `fitrsvm` — `'KernelFunction'`, `'PolynomialOrder'`, `'ResponseTransform'`, `'Standardize'`
Linear regression	Arguments of `fitrlinear` — `'ResponseTransform'`

- `validatedUpdateInputs` displays a warning message if the machine learning model parameters in `configurer` and `retrainedMdl` are identical.

See Also

`generateCode` | `learnerCoderConfigurer` | `update`

Topics

“Introduction to Code Generation” on page 32-2
“Code Generation for Prediction and Update Using Coder Configurer” on page 32-80

Introduced in R2018b

var

Package: prob

Variance of probability distribution

Syntax

```
v = var(pd)
```

Description

`v = var(pd)` returns the variance `v` of the probability distribution `pd`.

Examples

Variance of a Fitted Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades
x = grades(:,1);
```

Fit a normal distribution object to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

Compute the variance of the fitted distribution.

```
v = var(pd)
```

```
v = 76.0419
```

For a normal distribution, the variance is equal to the square of the parameter `sigma`.

Variance of a Skewed Distribution

Create a Weibull probability distribution object.

```
pd = makedist('Weibull','a',5,'b',2)
```

```
pd =
  WeibullDistribution
```

```
  Weibull distribution
    A = 5
    B = 2
```

Compute the variance of the distribution.

```
v = var(pd)
```

```
v = 5.3650
```

Variance of a Triangular Distribution

Create a triangular distribution object.

```
pd = makedist('Triangular','a',-3,'b',1,'c',3)
```

```
pd =
  TriangularDistribution

A = -3, B = 1, C = 3
```

Compute the variance of the distribution.

```
v = var(pd)
```

```
v = 1.5556
```

Variance of a Kernel Distribution

Load the sample data. Create a vector containing the first column of students' exam grade data.

```
load examgrades;
x = grades(:,1);
```

Fit a kernel distribution object to the data.

```
pd = fitdist(x,'Kernel')
```

```
pd =
  KernelDistribution

    Kernel = normal
    Bandwidth = 3.61677
    Support = unbounded
```

Compute the variance of the fitted distribution.

```
v = var(pd)
```

```
v = 88.4893
```

Input Arguments

pd — Probability distribution

probability distribution object

Probability distribution, specified as a probability distribution object created using one of the following.

Function or App	Description
`makedist`	Create a probability distribution object using specified parameter values.
`fitdist`	Fit a probability distribution object to sample data.
Distribution Fitter	Fit a probability distribution to sample data using the interactive Distribution Fitter app and export the fitted object to the workspace.

Output Arguments

v — Variance

nonnegative scalar value

Variance of the probability distribution, returned as a nonnegative scalar value.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The input argument `pd` can be a fitted probability distribution object for beta, exponential, extreme value, lognormal, normal, and Weibull distributions. Create `pd` by fitting a probability distribution to sample data from the `fitdist` function. For an example, see "Code Generation for Probability Distribution Objects" on page 32-82.

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

See Also

Distribution Fitter | `fitdist` | `makedist` | `mean` | `std`

Topics

"Working with Probability Distributions" on page 5-3
"Supported Distributions" on page 5-14

Introduced in R2013a

vartest

Chi-square variance test

Syntax

```
h = vartest(x,v)
h = vartest(x,v,Name,Value)
[h,p] = vartest( ___ )
[h,p,ci,stats] = vartest( ___ )
```

Description

`h = vartest(x,v)` returns a test decision for the null hypothesis that the data in vector `x` comes from a normal distribution with variance `v`, using the chi-square variance test on page 33-6577. The alternative hypothesis is that `x` comes from a normal distribution with a different variance. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = vartest(x,v,Name,Value)` performs the chi-square variance test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = vartest( ___ )` also returns the *p*-value of the test, `p`, using any of the input arguments in the previous syntaxes.

`[h,p,ci,stats] = vartest( ___ )` also returns the confidence interval for the true variance, `ci`, and the structure `stats` containing information about the test statistic.

Examples

Chi-Squared Test for Specified Variance

Load the sample data. Create a vector containing the first column of the students' exam grades matrix.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the data comes from a distribution with a variance of 25.

```
[h,p,ci,stats] = vartest(x,25)
```

```
h = 1
```

```
p = 0
```

```
ci = 2×1
```

```
   59.8936
   99.7688
```

```
stats = struct with fields:
    chisqstat: 361.9597
           df: 119
```

The returned value `h = 1` indicates that `vartest` rejects the null hypothesis at the default 5% significance level. `ci` shows the lower and upper boundaries of the 95% confidence interval for the true variance, and suggests that the true variance is greater than 25.

Chi-Squared Test Using One-Sided Hypothesis

Load the sample data. Create a vector containing the first column of the students' exam grades matrix.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the data comes from a distribution with a variance of 25, against the alternative hypothesis that the variance is greater than 25.

```
[h,p] = vartest(x,25,'Tail','right')
```

```
h = 1
```

```
p = 2.4269e-26
```

The returned value of `h = 1` indicates that `vartest` rejects the null hypothesis at the default 5% significance level, in favor of the alternative hypothesis that the variance is greater than 25.

Input Arguments

x — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array. For matrices, `vartest` performs separate tests along each column of `x`, and returns a row vector of results. For multidimensional arrays on page 33-6577, `vartest` works along the first nonsingleton dimension on page 33-6577 of `x`.

Data Types: `single` | `double`

v — Hypothesized variance

nonnegative scalar value

Hypothesized variance, specified as a nonnegative scalar value.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01` specifies a right-tailed hypothesis test at the 1% significance level.

`Alpha` — Significance level
`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension
first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix to test along, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the data in each column for equality to the hypothesized variance, while `'Dim',2` tests the data in each row.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis
`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternative hypothesis that the population variance is not `v`.
`'right'`	Test the alternative hypothesis that the population variance is greater than `v`.
`'left'`	Test the alternative hypothesis that the population variance is less than `v`.

Example: `'Tail','right'`

Output Arguments

`h` — Hypothesis test result
`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`ci` — Confidence interval
vector

Confidence interval for the true variance, returned as a two-element vector containing the lower and upper boundaries of the 100 × (1 – `Alpha`)% confidence interval.

`stats` — Test statistics
structure

Test statistics for the chi-square variance test, returned as a structure containing:

- `chisqstat` — Value of the test statistic.
- `df` — Degrees of freedom of the test.

More About

Chi-Square Variance Test

The chi-square variance test is used to test whether the variance of a population is equal to a hypothesized value.

The test statistic is

$$T = (n-1)\left(\frac{s}{\sigma_0}\right)^2,$$

where n is the sample size, s is the sample standard deviation, and σ_0 is the hypothesized standard deviation. The denominator is the ratio of the sample standard deviation to the hypothesized standard deviation. The further this ratio deviates from 1, the more likely you are to reject the null hypothesis. The test statistic T has a chi-square distribution with n – 1 degrees of freedom under the null hypothesis.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

Tips

- Use `sampsizepwr` to calculate:
  - The sample size that corresponds to specified power and parameter values;
  - The power achieved for a particular sample size, given the true parameter value;
  - The parameter value detectable with the specified sample size and power.

Extended Capabilities

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`sampsizepwr` | `vartest2` | `vartestn`

Introduced before R2006a

vartest2

Two-sample *F*-test for equal variances

Syntax

```
h = vartest2(x,y)
h = vartest2(x,y,Name,Value)
[h,p] = vartest2(___)
[h,p,ci,stats] = vartest2(___)
```

Description

`h = vartest2(x,y)` returns a test decision for the null hypothesis that the data in vectors `x` and `y` comes from normal distributions with the same variance, using the two-sample *F*-test on page 33-6582. The alternative hypothesis is that they come from normal distributions with different variances. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = vartest2(x,y,Name,Value)` returns a test decision for the two-sample *F*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = vartest2(___)` also returns the *p*-value of the test, `p`, using any of the input arguments in the previous syntaxes.

`[h,p,ci,stats] = vartest2(___)` also returns the confidence interval for the true variance ratio, `ci`, and the structure `stats` containing information about the test statistic.

Examples

Test for Equal Variances

Load the sample data. Create vectors containing the first and second columns of the data matrix to represent students&' grades on two exams.

```
load examgrades;
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the data in `x` and `y` comes from distributions with the same variance.

```
[h,p,ci,stats] = vartest2(x,y)
```

```
h = 1
```

```
p = 0.0019
```

```
ci = 2×1
```

```
    1.2383
    2.5494
```

```
stats = struct with fields:
    fstat: 1.7768
      df1: 119
      df2: 119
```

The returned result `h = 1` indicates that `vartest2` rejects the null hypothesis at the default 5% significance level. `ci` contains the lower and upper boundaries of the 95% confidence interval for the true variance ratio. `stats` contains the value of the test statistic for the *F*-test and the numerator and denominator degrees of freedom.

One-Sided Hypothesis Test

Load the sample data. Create vectors containing the first and second columns of the data matrix to represent students' grades on two exams.

```
load examgrades;
x = grades(:,1);
y = grades(:,2);
```

Test the null hypothesis that the data in `x` and `y` comes from distributions with the same variance, against the alternative that the population variance of `x` is greater than that of `y`.

```
vartest2(x,y,'Tail','right')
```

```
ans = 1
```

The returned result `h = 1` indicates that `vartest2` rejects the null hypothesis at the default 5% significance level, in favor of the alternative hypothesis that the population variance of `x` is greater than that of `y`.

Input Arguments

x — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array on page 33-6582.

- If `x` and `y` are vectors, they do not need to be the same length.
- If `x` and `y` are matrices, they must have the same number of columns, but do not need to have the same number of rows. `vartest2` performs separate tests along each column and returns a vector of the results.
- If `x` and `y` are multidimensional arrays, they must have the same number of dimensions, and the same size along all but the first nonsingleton dimension on page 33-6583.

Data Types: `single` | `double`

y — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array on page 33-6582.

- If `x` and `y` are vectors, they do not need to be the same length.
- If `x` and `y` are matrices, they must have the same number of columns, but do not need to have the same number of rows. `vartest2` performs separate tests along each column and returns a vector of the results.
- If `x` and `y` are multidimensional arrays, they must have the same number of dimensions, and the same size along all but the first nonsingleton dimension on page 33-6583.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01` specifies a right-tailed hypothesis test at the 1% significance level.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension

first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix to test along, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the data in each column for variance equality, while `'Dim',2` tests the data in each row.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate using the *F*-test, specified as the comma-separated pair consisting of `'Tail'` and one of the following.

`'both'`	Test the alternative hypothesis that the population variances are not equal.
`'right'`	Test the alternative hypothesis that the population variance of `x` is greater than that of `y`.
`'left'`	Test the alternative hypothesis that the population variance of `x` is less than that of `y`.

Example: `'Tail','right'`

Output Arguments

h — Hypothesis test result

1 | 0

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.
- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

p — *p*-value

scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

ci — Confidence interval

vector

Confidence interval for the true ratio of the population variances, returned as a two-element vector containing the lower and upper boundaries of the 100 × (1 – `Alpha`)% confidence interval.

stats — Test statistics

structure

Test statistics for the hypothesis test, returned as a structure containing:

- `fstat` — Value of the test statistic.
- `df1` — Numerator degrees of freedom of the test.
- `df2` — Denominator degrees of freedom of the test.

More About

Two-Sample *F*-Test

The two-sample *F*-test is used to test if the variances of two populations are equal.

The test statistic is

$$F = \frac{{s_1}^2}{{s_2}^2},$$

where s_1 and s_2 are the sample standard deviations. The test statistic is a ratio of the two sample variances. The further this ratio deviates from 1, the more likely you are to reject the null hypothesis. Under the null hypothesis, the test statistic F has a F-distribution with numerator degrees of freedom equal to N_1 – 1 and denominator degrees of freedom equal to N_2 – 1, where N_1 and N_2 are the sample sizes of the two data sets.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`vartest` | `vartestn`

Introduced before R2006a

vartestn

Multiple-sample tests for equal variances

Syntax

```
vartestn(x)
vartestn(x,Name,Value)

vartestn(x,group)
vartestn(x,group,Name,Value)

p = vartestn( ___ )
[p,stats] = vartestn( ___ )
```

Description

`vartestn(x)` returns a summary table of statistics and a box plot for a Bartlett test of the null hypothesis that the columns of data vector `x` come from normal distributions with the same variance. The alternative hypothesis is that not all columns of data have the same variance.

`vartestn(x,Name,Value)` returns a summary table of statistics and a box plot for a test of unequal variances with additional options specified by one or more name-value pair arguments. For example, you can specify a different type of hypothesis test or change the display settings for the test results.

`vartestn(x,group)` returns a summary table of statistics and a box plot for a Bartlett test of the null hypothesis that the data in each categorical group comes from normal distributions with the same variance. The alternative hypothesis is that not all groups have the same variance.

`vartestn(x,group,Name,Value)` returns a summary table of statistics and a box plot for a test of unequal variances with additional options specified by one or more name-value pair arguments. For example, you can specify a different type of hypothesis test or change the display settings for the test results.

`p = vartestn( ___ )` also returns the *p*-value of the test, `p`, using any of the input arguments in the previous syntaxes.

`[p,stats] = vartestn( ___ )` also returns the structure `stats` containing information about the test statistic.

Examples

Test Data for Equal Variances

Load the sample data.

```
load examgrades
```

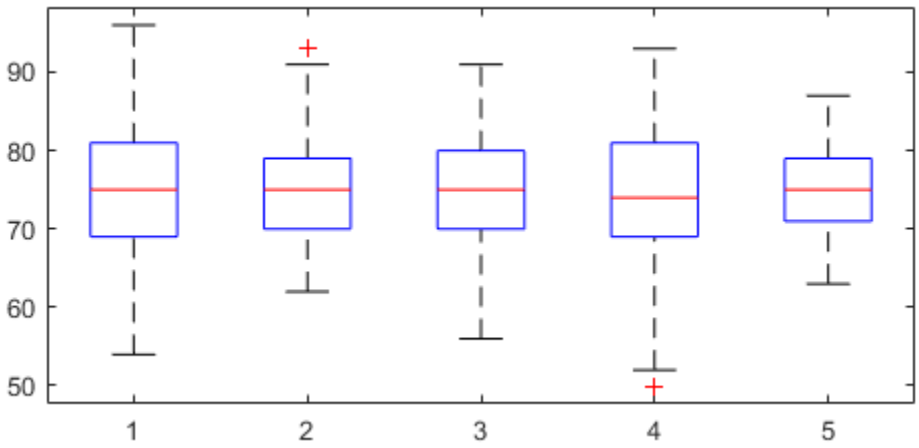
Test the null hypothesis that the variances are equal across the five columns of data in the students' exam grades matrix, `grades`.

```
vartestn(grades)
```

Group Summary Table

Group	Count	Mean	Std Dev
1	120	75.0083	8.7202
2	120	74.9917	6.54204
3	120	74.9917	7.43091
4	120	75.0333	8.60128
5	120	74.9917	5.25884
Pooled	600	75.0033	7.42558

Bartlett's statistic	38.7332
Degrees of freedom	4
p-value	0




```
ans = 7.9086e-08
```

The low *p*-value, `p = 0`, indicates that `vartestn` rejects the null hypothesis that the variances are equal across all five columns, in favor of the alternative hypothesis that at least one column has a different variance.

Test Grouped Data for Equal Variances

Load the sample data.

```
load carsmall
```

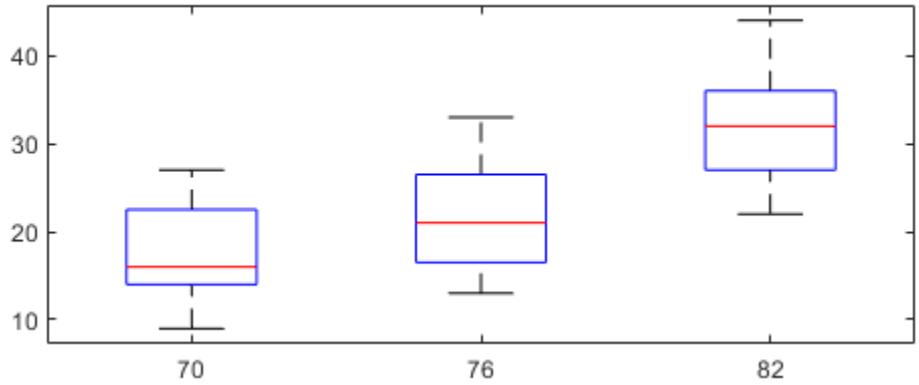
Test the null hypothesis that the variances in miles per gallon (`MPG`) are equal across different model years.

```
vartestn(MPG,Model_Year)
```

Group Summary Table

Group	Count	Mean	Std Dev
70	29	17.6897	5.33923
76	34	21.5735	5.8893
82	31	31.7097	5.39255
Pooled	94	23.7181	5.562

Bartlett's statistic	0.36619
Degrees of freedom	2
p-value	0.83269




```
ans = 0.8327
```

The high *p*-value, `p = 0.83269`, indicates that `vartestn` does not reject the null hypothesis that the variances in miles per gallon (`MPG`) are equal across different model years.

Test for Equal Variances Using Levene's Test

Load the sample data.

```
load carsmall
```

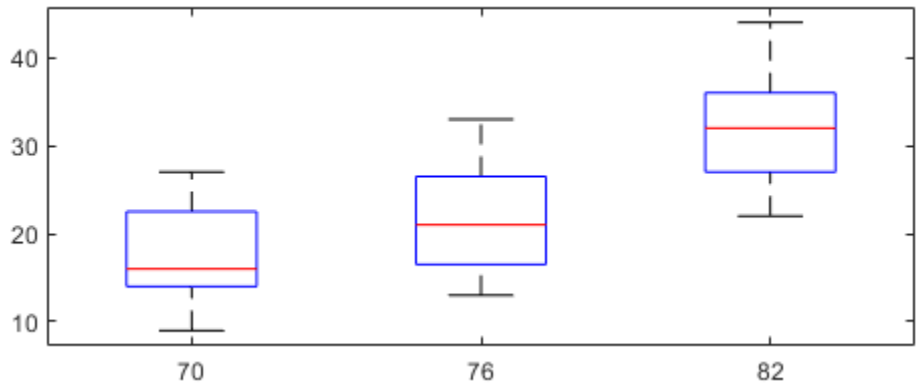
Use Levene's test to test the null hypothesis that the variances in miles per gallon (`MPG`) are equal across different model years.

```
p = vartestn(MPG,Model_Year,'TestType','LeveneAbsolute')
```

Group Summary Table

Group	Count	Mean	Std Dev
70	29	17.6897	5.33923
76	34	21.5735	5.8893
82	31	31.7097	5.39255
Pooled	94	23.7181	5.562

Levene's statistic (absolute)	0.46126
Degrees of freedom	2, 91
p-value	0.63195




```
p = 0.6320
```

The high *p*-value, `p = 0.63195`, indicates that `vartestn` does not reject the null hypothesis that the variances in miles per gallon (`MPG`) are equal across different model years.

Test for Equal Variances Using the Brown-Forsythe Test

Load the sample data.

```
load examgrades
```

Test the null hypothesis that the variances are equal across the five columns of data in the students' exam grades matrix, `grades`, using the Brown-Forsythe test. Suppress the display of the summary table of statistics and the box plot.

```
[p,stats] = vartestn(grades,'TestType','BrownForsythe','Display','off')
```

```
p = 1.3121e-06
```

```
stats = struct with fields:
    fstat: 8.4160
       df: [4 595]
```

The small *p*-value, `p = 1.3121e-06`, indicates that `vartestn` rejects the null hypothesis that the variances are equal across all five columns, in favor of the alternative hypothesis that at least one column has a different variance.

Input Arguments

x — Sample data

matrix | column vector

Sample data, specified as a matrix or column vector. If a grouping variable `group` is specified, then `x` must be a column vector. If a grouping variable is not specified, `x` must be a matrix. In either case, `vartestn` treats `NaN` values as missing values and ignores them.

Data Types: `single` | `double`

group — Grouping variable

categorical array | logical or numeric vector | character array | string array | cell array of character vectors

Grouping variable, specified as a categorical array, logical or numeric vector, character array, string array, or cell array of character vectors with one row for each element of `x`. Each unique value in a grouping variable defines a group. `vartestn` treats `NaN` values as missing values and ignores them.

For example, if `Gender` is a cell array of character vectors with values `'Male'` and `'Female'`, you can use `Gender` as a grouping variable to test your data by gender.

Example: `Gender`

Data Types: `categorical` | `single` | `double` | `logical` | `string` | `cell` | `char`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'TestType','BrownForsythe','Display','off'` specifies a Brown-Forsythe test and omits the plot of the results.

Display — Display settings for test results

`'on'` (default) | `'off'`

Display settings for test results, specified as the comma-separated pair consisting of `'Display'` and one of the following.

`'on'`	Display a box plot and table of summary statistics.
`'off'`	Do not display a box plot and table of summary statistics.

Example: `'display','off'`

TestType — Type of hypothesis test

`'Bartlett'` (default) | `'LeveneQuadratic'` | `'LeveneAbsolute'` | `'BrownForsythe'` | `'OBrien'`

Type of hypothesis test to perform, specified as the comma-separated pair consisting of `'TestType'` and one of the following.

`'Bartlett'`	Bartlett's test.
`'LeveneQuadratic'`	Levene's test computed by performing ANOVA on the squared deviations of the data values from their group means.
`'LeveneAbsolute'`	Levene's test computed by performing ANOVA on the absolute deviations of the data values from their group means.
`'BrownForsythe'`	Brown-Forsythe test computed by performing ANOVA on the absolute deviations of the data values from the group medians.
`'OBrien'`	O'Brien's modification of Levene's test with `W` = 0.5.

Example: `'TestType','OBrien'`

Output Arguments

`p` — *p*-value

scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`stats` — Test statistics

structure

Test statistics for the hypothesis test, returned as a structure containing:

- `chistat`: Value of the test statistic.
- `df`: Degrees of freedom of the test.

More About

Bartlett's Test

Bartlett's test is used to test whether multiple data samples have equal variances, against the alternative that at least two of the data samples do not have equal variances.

The test statistic is

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i-1)\ln s_i^2}{1+(1/(3(k-1)))\left(\left(\sum_{i=1}^{k}1/(N_i-1)\right)-1/(N-k)\right)},$$

where s_i^2 is the variance of the ith group, N is the total sample size, N_i is the sample size of the ith group, k is the number of groups, and s_p^2 is the pooled variance. The pooled variance is defined as

$$s_p^2 = \sum_{i=1}^{k}(N_i-1)s_i^2/(N-k)\,.$$

The test statistic has a chi-square distribution with k – 1 degrees of freedom under the null hypothesis.

Bartlett's test is sensitive to departures from normality. If your data comes from a nonnormal distribution, Levene's test could provide a more accurate result.

Levene, Brown-Forsythe, and O'Brien Tests

The Levene, Brown-Forsythe, and O'Brien tests are used to test if multiple data samples have equal variances, against the alternative that at least two of the data samples do not have equal variances.

The test statistic is

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i\left(\bar{Z}_{i\cdot} - \bar{Z}_{\cdot\cdot}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(Z_{ij} - \bar{Z}_{i\cdot}\right)^2},$$

where N_i is the sample size of the ith group, and k is the number of groups. Depending on the type of test specified with the `TestType` name-value pair arguments, Z_{ij} can have one of four definitions:

- If you specify `LeveneAbsolute`, `vartestn` uses $Z_{ij} = \left|Y_{ij} - \bar{Y}_{i\cdot}\right|$, where $\bar{Y}_{i\cdot}$ is the mean of the ith subgroup.
- If you specify `LeveneQuadratic`, `vartestn` uses $Z_{ij}2 = \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$, where $\bar{Y}_{i\cdot}$ is the mean of the ith subgroup.
- If you specify `BrownForsythe`, `vartestn` uses $Z_{ij} = \left|Y_{ij} - \tilde{Y}_{i\cdot}\right|$, where $\tilde{Y}_{i\cdot}$ is the median of the ith subgroup.
- If you specify `OBrien`, `vartestn` uses

  $$Z_{ij} = \frac{(0.5 + n_i - 2)n_i\left(y_{ij} - \bar{y}_i\right)^2 - \ 0.5(n_i - 1)\sigma_i 2}{(n_i - 1)(n_i - 2)},$$

  where n_i is the size of the ith group, σ_i^2 is its sample variance.

In all cases, the test statistic has an F-distribution with k – 1 numerator degrees of freedom, and N – k denominator degrees of freedom.

The Levene, Brown-Forsythe, and O'Brien tests are less sensitive to departures from normality than Bartlett's test, so they are useful alternatives if you suspect the samples come from nonnormal distributions.

See Also

`anova1` | `vartest` | `vartest2`

Introduced before R2006a

vertcat

Class: `dataset`

(Not Recommended) Vertical concatenation for dataset arrays

Note The `dataset` data type is not recommended. To work with heterogeneous data, use the MATLAB® `table` data type instead. See MATLAB `table` documentation for more information.

Syntax

```
ds = vertcat(ds1, ds2, ...)
```

Description

`ds = vertcat(ds1, ds2, ...)` vertically concatenates the dataset arrays `ds1, ds2, ...`. Observation names, when present, must be unique across datasets. `vertcat` fills in default observation names for the output when some of the inputs have names and some do not.

Variable names for all dataset arrays must be identical except for order. `vertcat` concatenates by matching variable names. `vertcat` assigns values for the "per-variable" properties (e.g., `Units` and `VarDescription`) in `ds` from the corresponding property values in `ds1`.

See Also

`cat` | `horzcat`

compact

Class: `clustering.evaluation.ClusterCriterion`
Package: `clustering.evaluation`

Compact clustering evaluation object

Syntax

```
c = compact(eva)
```

Description

`c = compact(eva)` returns a compact clustering evaluation object, which contains a subset of information about the clustering solution in `eva`. Compacting a clustering evaluation object reduces the memory requirements of the object, which is useful when clustering a large data set.

Input Arguments

eva — Clustering evaluation data

clustering evaluation object

Clustering evaluation data, specified as a clustering evaluation object. Create a clustering evaluation object using `evalclusters`.

Output Arguments

c — Compact clustering evaluation object

clustering evaluation object

Compact clustering evaluation object, returned as a clustering evaluation object. The compact object includes the clustering evaluation results, In the compact object, the properties for the input data `X`, optimal clustering solution `OptimalY`, and the list of excluded data `Missing` are empty.

Examples

Create a Compact Clustering Evaluation Object

Create a compact clustering evaluation object from a full clustering evaluation object.

Load the sample data.

```
load fisheriris;
```

The data contains length and width measurements from the sepals and petals of three species of iris flowers.

Create a clustering evaluation object. Cluster the data using `kmeans`, and evaluate the optimal number of clusters using the gap criterion.

```
rng('default');  % For reproducibility
eva = evalclusters(meas,'kmeans','Gap','KList',[1:6])

eva =
  GapEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [0.0720 0.5928 0.8762 1.0114 1.0534 1.0720]
           OptimalK: 5
```

Create a compact clustering evaluation object from `eva`.

```
c = compact(eva)

c =
  GapEvaluation with properties:

    NumObservations: 150
         InspectedK: [1 2 3 4 5 6]
    CriterionValues: [0.0720 0.5928 0.8762 1.0114 1.0534 1.0720]
           OptimalK: 5
```

The displayed output of the compact object `c` is the same as the original object `eva`, but some properties not shown in the display are different. For example, in the compact object, the properties `x`, `OptimalY`, and `Missing` are empty.

Display the optimal clustering solution `OptimalY` for `c`.

```
c.OptimalY

ans =

     []
```

See Also

`evalclusters`

view

View classification tree

Syntax

```
view(tree)
view(tree,Name,Value)
```

Description

`view(tree)` returns a text description of `tree`, a decision tree.

`view(tree,Name,Value)` describes `tree` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`tree`

A classification tree or compact classification tree created by `fitctree` or `compact`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Mode`

Value describing the display of `tree`, either `'graph'` or `'text'`. `'graph'` opens a user interface displaying `tree`, and containing controls for querying the tree. `'text'` sends output to the Command Window describing `tree`.

Default: `'text'`

Examples

View Trained Classification Tree

View textual and graphical displays of a trained classification tree.

Load Fisher's iris data set.

```
load fisheriris
```

Train a classification tree using all measurements.

```
Mdl = fitctree(meas,species);
```

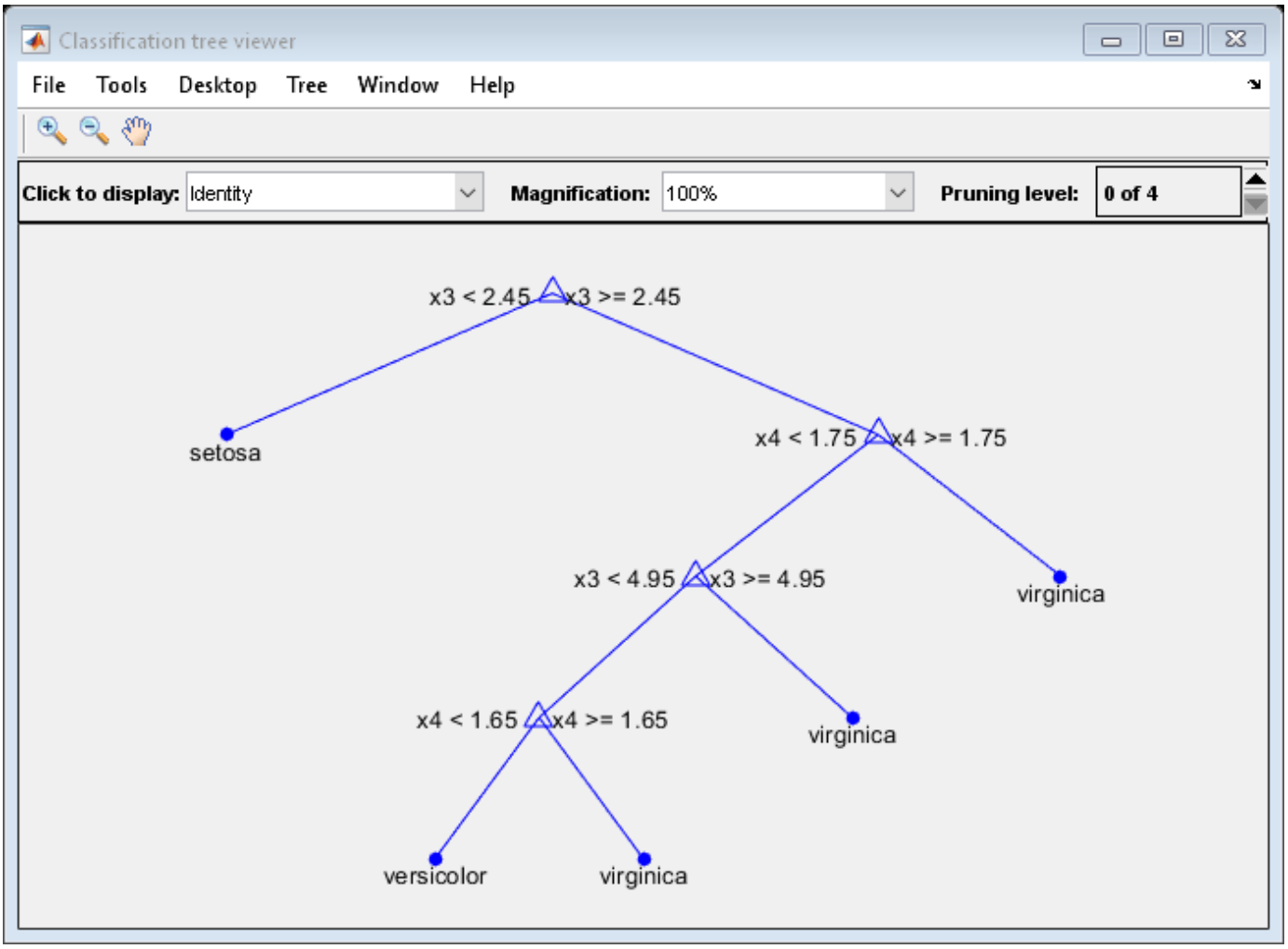
View textual display of the trained classification tree.

```
view(Mdl)
```

```
Decision tree for classification
1  if x3<2.45 then node 2 elseif x3>=2.45 then node 3 else setosa
2  class = setosa
3  if x4<1.75 then node 4 elseif x4>=1.75 then node 5 else versicolor
4  if x3<4.95 then node 6 elseif x3>=4.95 then node 7 else versicolor
5  class = virginica
6  if x4<1.65 then node 8 elseif x4>=1.65 then node 9 else versicolor
7  class = virginica
8  class = versicolor
9  class = virginica
```

View graphical display of the trained classification tree.

```
view(Mdl,'Mode','graph');
```



View Tree from Bag of Trees

Load Fisher's iris data set.

```
load fisheriris
```

Grow a bag of 100 classification trees using all measurements.

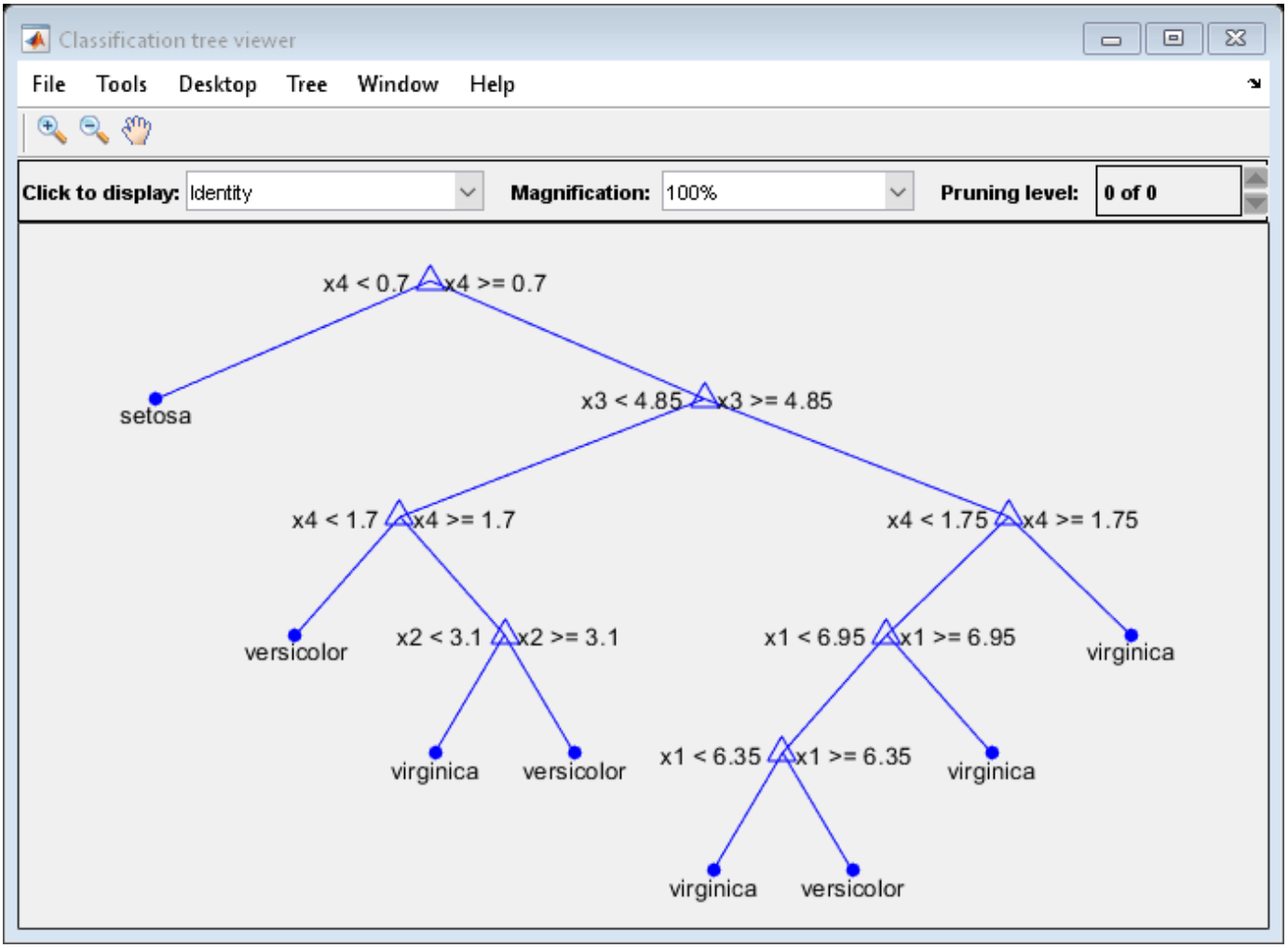
```
rng(1) % For reproducibility
Mdl = TreeBagger(100,meas,species);
```

Alternatively, you can use `fitcensemble` to grow a bag of classification trees.

`Mdl` is a `TreeBagger` model object. `Mdl.Trees` stores the bag of 100 trained classification trees in a 100-by-1 cell array. That is, each cell in `Mdl.Trees` contains a `CompactClassificationTree` model object.

View a graph of the 10th classification tree in the bag.

```
Tree10 = Mdl.Trees{10};
view(Tree10,'Mode','graph');
```



By default, the software grows deep trees for bags of trees.

View Tree from Boosted Ensemble

Load Fisher's iris data set.

```
load fisheriris
```

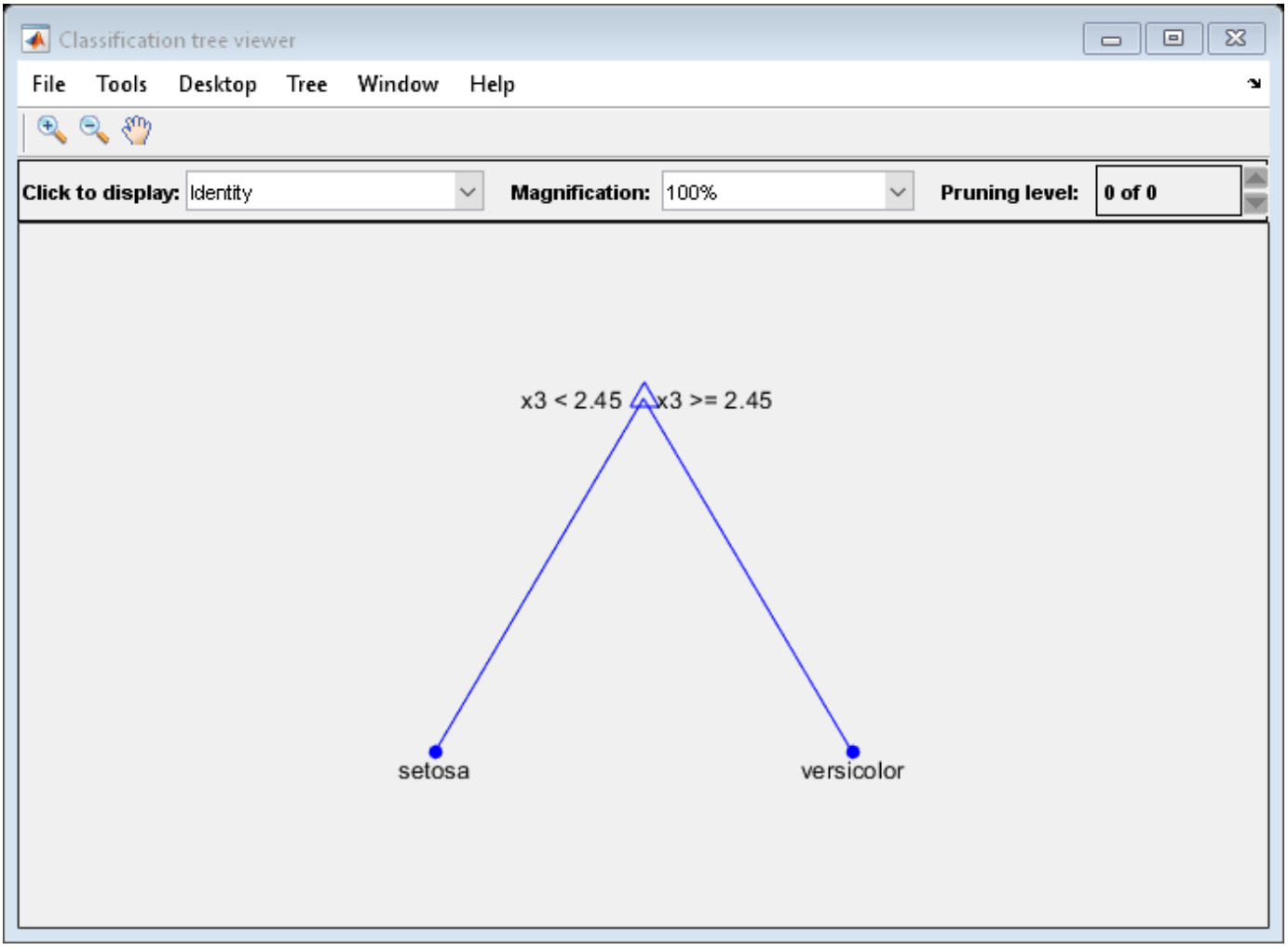
Boost an ensemble of 100 classification trees using all measurements. Specify tree stumps as the weak learners.

```
t = templateTree('MaxNumSplits',1);
Mdl = fitcensemble(meas,species,'Method','AdaBoostM2','Learners',t);
```

`Mdl` is a `ClassificationEnsemble` model object. `Mdl.Trained` stores the ensemble of 100 trained classification trees in a 100-by-1 cell array. That is, each cell in `Mdl.Trained` contains a `CompactClassificationTree` model object.

View a graph of the 10th classification tree in the ensemble.

```
Tree10 = Mdl.Trained{10};
view(Tree10,'Mode','graph');
```



The graph shows a tree stump because you specified stumps as the weak learners for the ensemble. However, this behavior is not the default for `fitcensemble`. By default, `fitcensemble` grows

shallow trees for boosted ensembles of trees. That is, `'Learners'` is `templateTree('MaxNumSplits',10)`.

Tip

To view tree `t` from an ensemble of trees, enter one of these lines of code

```
view(Ens.Trained{t})
view(Bag.Trees{t})
```

- `Ens` is a full ensemble returned by `fitcensemble` or a compact ensemble returned by `compact`.
- `Bag` is a full bag of trees returned by `TreeBagger` or a compact bag of trees returned by `compact`.

To save `tree` in the Command Window, get a figure handle by using the `findall` and `setdiff` functions, and then save `tree` using the function `saveas`.

```
before = findall(groot,'Type','figure'); % Find all figures
view(Mdl,'Mode','graph')
after = findall(groot,'Type','figure');
h = setdiff(after,before); % Get the figure handle of the tree viewer
saveas(h,'a.png')
```

See Also

`ClassificationTree` | `fitctree`

view

View regression tree

Syntax

```
view(tree)
view(tree,Name,Value)
```

Description

`view(tree)` returns a text description of `tree`, a decision tree.

`view(tree,Name,Value)` describes `tree` with additional options specified by one or more `Name,Value` pair arguments.

Input Arguments

`tree`

A regression tree or compact regression tree created by `fitrtree` or `compact`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

`Mode`

Display of `tree`, either `'graph'` or `'text'`. `'graph'` opens a GUI displaying `tree`, and containing controls for querying the tree. `'text'` sends output to the Command Window describing `tree`.

Default: `'text'`

Examples

View Trained Regression Tree

View textual and graphical displays of a trained regression tree.

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

Train a regression tree using all measurements.

```
Mdl = fitrtree(X,Y);
```

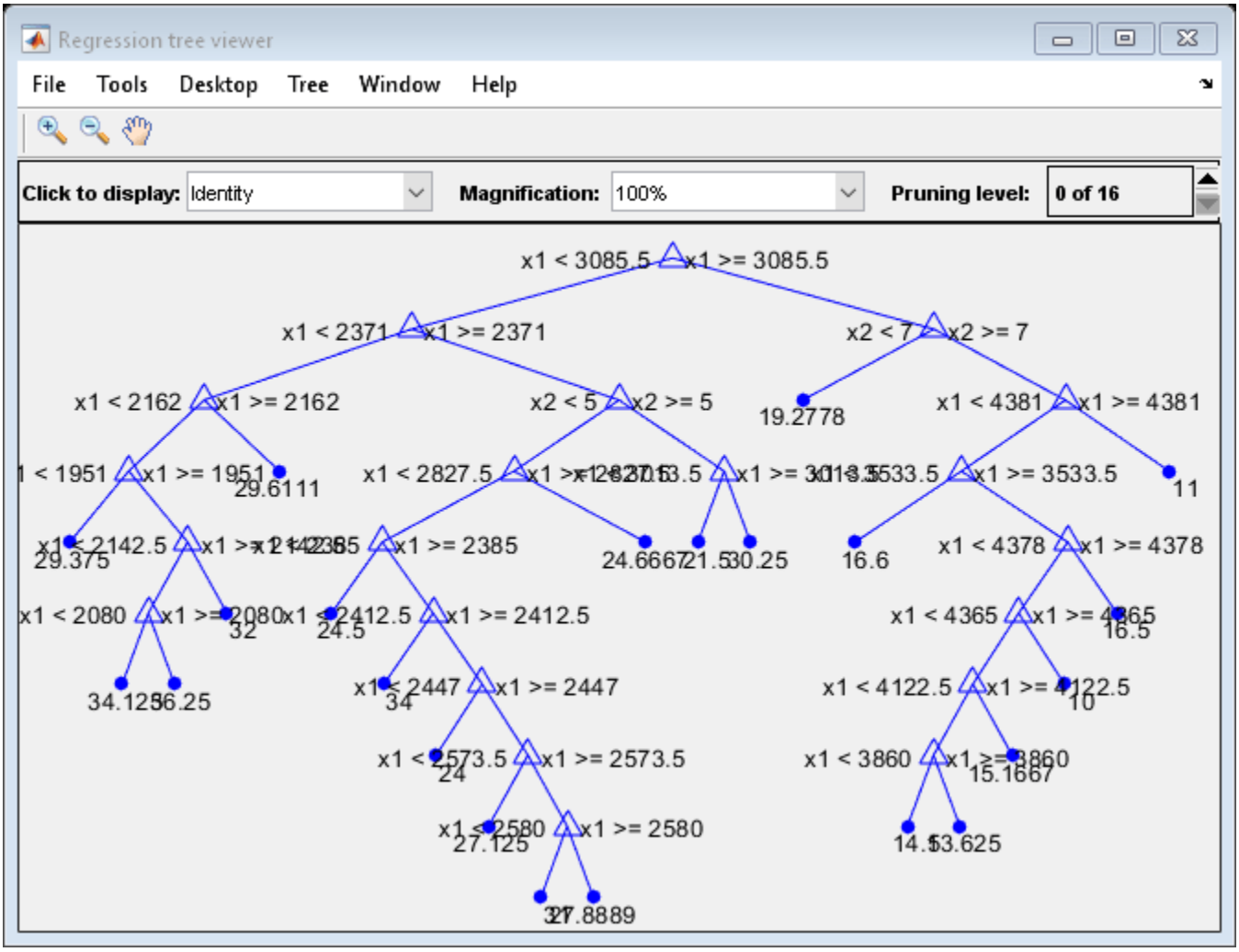
View textual display of the trained regression tree.

```
view(Mdl)
```

```
Decision tree for regression
 1  if x1<3085.5 then node 2 elseif x1>=3085.5 then node 3 else 23.7181
 2  if x1<2371 then node 4 elseif x1>=2371 then node 5 else 28.7931
 3  if x2<7 then node 6 elseif x2>=7 then node 7 else 15.5417
 4  if x1<2162 then node 8 elseif x1>=2162 then node 9 else 32.0741
 5  if x2<5 then node 10 elseif x2>=5 then node 11 else 25.9355
 6  fit = 19.2778
 7  if x1<4381 then node 12 elseif x1>=4381 then node 13 else 14.2963
 8  if x1<1951 then node 14 elseif x1>=1951 then node 15 else 33.3056
 9  fit = 29.6111
10  if x1<2827.5 then node 16 elseif x1>=2827.5 then node 17 else 27.2143
11  if x1<3013.5 then node 18 elseif x1>=3013.5 then node 19 else 23.25
12  if x1<3533.5 then node 20 elseif x1>=3533.5 then node 21 else 14.8696
13  fit = 11
14  fit = 29.375
15  if x1<2142.5 then node 22 elseif x1>=2142.5 then node 23 else 34.4286
16  if x1<2385 then node 24 elseif x1>=2385 then node 25 else 27.6389
17  fit = 24.6667
18  fit = 21.5
19  fit = 30.25
20  fit = 16.6
21  if x1<4378 then node 26 elseif x1>=4378 then node 27 else 14.3889
22  if x1<2080 then node 28 elseif x1>=2080 then node 29 else 34.8333
23  fit = 32
24  fit = 24.5
25  if x1<2412.5 then node 30 elseif x1>=2412.5 then node 31 else 28.0313
26  if x1<4365 then node 32 elseif x1>=4365 then node 33 else 14.2647
27  fit = 16.5
28  fit = 34.125
29  fit = 36.25
30  fit = 34
31  if x1<2447 then node 34 elseif x1>=2447 then node 35 else 27.6333
32  if x1<4122.5 then node 36 elseif x1>=4122.5 then node 37 else 14.5313
33  fit = 10
34  fit = 24
35  if x1<2573.5 then node 38 elseif x1>=2573.5 then node 39 else 27.8929
36  if x1<3860 then node 40 elseif x1>=3860 then node 41 else 14.15
37  fit = 15.1667
38  fit = 27.125
39  if x1<2580 then node 42 elseif x1>=2580 then node 43 else 28.2
40  fit = 14.5
41  fit = 13.625
42  fit = 31
43  fit = 27.8889
```

View graphical display of the trained regression tree.

```
view(Mdl,'Mode','graph');
```



View Tree from Bag of Trees

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

Grow a bag of 100 regression trees using all measurements.

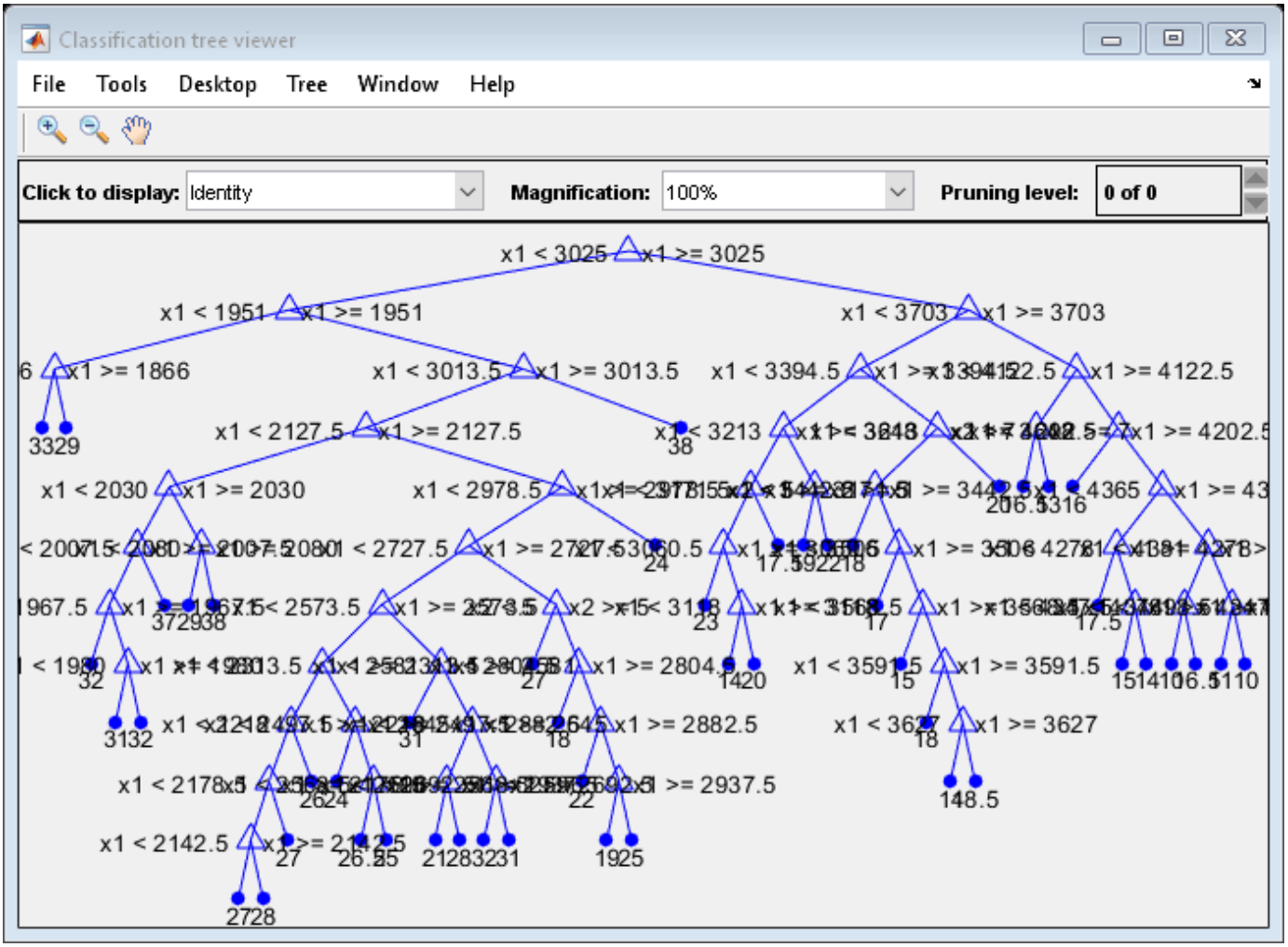
```
rng(1) % For reproducibility
Mdl = TreeBagger(100,X,Y);
```

Alternatively, you can use `fitrensemble` to grow a bag of regression trees.

`Mdl` is a `TreeBagger` model object. `Mdl.Trees` stores the bag of 100 trained regression trees in a 100-by-1 cell array. That is, each cell in `Mdl.Trees` contains a `CompactRegressionTree` model object.

View a graph of the 10th regression tree in the bag.

```
Tree10 = Mdl.Trees{10};
view(Tree10,'Mode','graph');
```



By default, the software grows deep trees for bags of trees.

View Tree from Boosted Ensemble

Load the `carsmall` data set. Consider a model that explains a car's fuel economy (`MPG`) using its weight (`Weight`) and number of cylinders (`Cylinders`).

```
load carsmall
X = [Weight Cylinders];
Y = MPG;
```

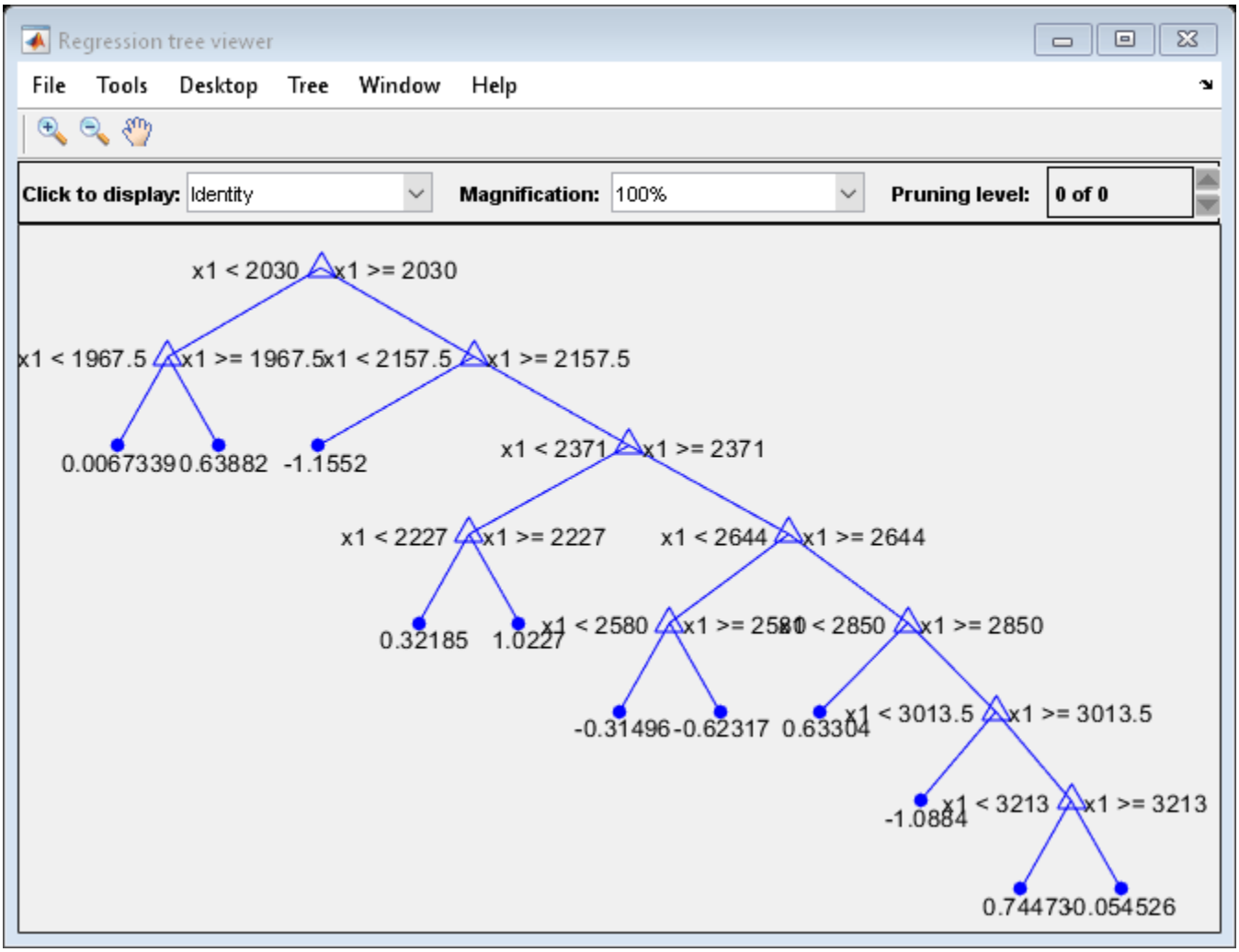
Boost an ensemble of 100 regression trees using all measurements.

```
Mdl = fitrensemble(X,Y,'Method','LSBoost');
```

`Mdl` is a `RegressionEnsemble` model object. `Mdl.Trained` stores the ensemble of 100 trained regression trees in a 100-by-1 cell array. That is, each cell in `Mdl.Trained` contains a `CompactRegressionTree` model object.

View a graph of the 10th regression tree in the ensemble.

```
Tree10 = Mdl.Trained{10};
view(Tree10,'Mode','graph');
```



By default, `fitrensemble` grows shallow trees for boosted ensembles of trees. That is, `'Learners'` is `templateTree('MaxNumSplits',10)`.

Tip

To view tree *`t`* from an ensemble of trees, enter one of these lines of code

```
view(Ens.Trained{t})
view(Bag.Trees{t})
```

- `Ens` is a full ensemble returned by `fitrensemble` or a compact ensemble returned by `compact`.
- `Bag` is a full bag of trees returned by `TreeBagger` or a compact bag of trees returned by `compact`.

To save `tree` in the Command Window, get a figure handle by using the `findall` and `setdiff` functions, and then save `tree` using the function `saveas`.

```
before = findall(groot,'Type','figure'); % Find all figures
view(Mdl,'Mode','graph')
after = findall(groot,'Type','figure');
h = setdiff(after,before); % Get the figure handle of the tree viewer
saveas(h,'a.png')
```

See Also

`RegressionTree` | `fitrtree`

wblcdf

Weibull cumulative distribution function

Syntax

```
p = wblcdf(x,a,b)
[p,plo,pup] = wblcdf(x,a,b,pcov,alpha)
[p,plo,pup] = wblcdf( ___ ,'upper')
```

Description

`p = wblcdf(x,a,b)` returns the cdf of the Weibull distribution with scale parameter `a` and shape parameter `b`, at each value in `x`. `x`, `a`, and `b` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other inputs. The default values for `a` and `b` are both `1`. The parameters `a` and `b` must be positive.

`[p,plo,pup] = wblcdf(x,a,b,pcov,alpha)` returns confidence bounds for `p` when the input parameters `a` and `b` are estimates. `pcov` is the 2-by-2 covariance matrix of the estimated parameters. `alpha` has a default value of 0.05, and specifies 100(1 - `alpha`)% confidence bounds. `plo` and `pup` are arrays of the same size as `p` containing the lower and upper confidence bounds.

`[p,plo,pup] = wblcdf( ___ ,'upper')` returns the complement of the Weibull cdf for each value in `x`, using an algorithm that more accurately computes the extreme upper tail probabilities. You can use `'upper'` with any of the previous syntaxes.

The function `wblcdf` computes confidence bounds for `p` using a normal approximation to the distribution of the estimate

$$\widehat{b}\left(\log x - \log\widehat{a}\right)$$

and then transforms those bounds to the scale of the output `p`. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `pcov` from large samples, but in smaller samples other methods of computing the confidence bounds might be more accurate.

The Weibull cdf is

$$p = F(x \mid a, b) = \int_0^x b a^{-b} t^{b-1} e^{-\left(\frac{t}{a}\right)^b} dt = 1 - e^{-\left(\frac{x}{a}\right)^b}.$$

Examples

Weibull Distribution cdf

What is the probability that a value from a Weibull distribution with parameters `a = 0.15` and `b = 0.8` is less than 0.5?

```
probability = wblcdf(0.5, 0.15, 0.8)
```

```
probability = 0.9272
```

How sensitive is this result to small changes in the parameters?

```
[A, B] = meshgrid(0.1:0.05:0.2,0.2:0.05:0.3);
probability = wblcdf(0.5, A, B)
```

```
probability = 3×3

    0.7484    0.7198    0.6991
    0.7758    0.7411    0.7156
    0.8022    0.7619    0.7319
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`cdf` | `wblfit` | `wblinv` | `wbllike` | `wblpdf` | `wblplot` | `wblrnd` | `wblstat`

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wblfit

Weibull parameter estimates

Syntax

```
parmHat = wblfit(x)

[parmHat,parmCI] = wblfit(x)
[parmHat,parmCI] = wblfit(x,alpha)

[ ___ ] = wblfit(x,alpha,censoring)
[ ___ ] = wblfit(x,alpha,censoring,freq)
[ ___ ] = wblfit(x,alpha,censoring,freq,options)
```

Description

`parmHat = wblfit(x)` returns the estimates of Weibull distribution parameters (shape and scale), given the sample data in `x`.

`[parmHat,parmCI] = wblfit(x)` also returns the 95% confidence intervals for the parameter estimates.

`[parmHat,parmCI] = wblfit(x,alpha)` specifies the confidence level for the confidence intervals to be `100(1–alpha)`%.

`[ ___ ] = wblfit(x,alpha,censoring)` specifies whether each value in `x` is right-censored or not. Use the logical vector `censoring` in which 1 indicates observations that are right-censored and 0 indicates observations that are fully observed.

`[ ___ ] = wblfit(x,alpha,censoring,freq)` specifies the frequency or weights of observations.

`[ ___ ] = wblfit(x,alpha,censoring,freq,options)` specifies optimization options for the iterative algorithm `wblfit` to use to compute maximum likelihood estimates (MLEs) with censoring. Create `options` by using the function `statset`.

You can pass in `[]` for `alpha`, `censoring`, and `freq` to use their default values.

Examples

Estimate Parameters of Weibull Distribution

Generate 100 random numbers from the Weibull distribution with scale `0.8` and shape `3`.

```
x = wblrnd(0.8,3,100,1);
```

Estimate the parameters of the Weibull distribution from the data.

```
parmHat = wblfit(x)
```

```
parmHat = 1×2
```

```
    0.7751    2.9433
```

Estimate Parameters of Weibull Distribution with Confidence Intervals

Generate 100 random numbers from the Weibull distribution with scale 1 and shape 2.

```
x = wblrnd(1,2,100,1);
```

Find the 95% confidence intervals estimating the parameters of the Weibull distribution from the data.

```
[parmHat,parmCI] = wblfit(x)
```

```
parmHat = 1×2

    0.9536    1.9622


parmCI = 2×2

    0.8583    1.6821
    1.0596    2.2890
```

The top row of `parmCI` contains the lower bounds of the confidence intervals and the bottom row contains the upper bounds of the confidence intervals.

Estimate Weibull Parameters with Algorithm Options

Generate 100 Weibull random variables from the distribution with scale 2 and shape 5.

```
x = wblrnd(2,5,100,1);
```

Display the algorithm parameters for `wblfit`.

```
statset('wblfit')
```

```
ans = struct with fields:
          Display: 'off'
      MaxFunEvals: []
          MaxIter: []
           TolBnd: []
           TolFun: []
       TolTypeFun: []
             TolX: 1.0000e-06
         TolTypeX: []
          GradObj: []
         Jacobian: []
        DerivStep: []
      FunValCheck: []
           Robust: []
     RobustWgtFun: []
```

```
           WgtFun: []
             Tune: []
      UseParallel: []
    UseSubstreams: []
          Streams: {}
        OutputFcn: []
```

Specify algorithm parameters using name-value pair arguments of the function `statset`. Change how results are displayed (`Display`), and set the termination tolerance for parameters (`TolX`).

```
options = statset('Display','iter','TolX',1e-4); % Optimization options
```

Find the MLEs using the new algorithm parameters.

```
parmhat = wblfit(x,[],[],[],options)
```

```
 Func-count    x          f(x)             Procedure
    2        0.193283    -0.0172927        initial
    3        0.205467    0.00262429        interpolation
    4        0.203862   2.99018e-05        interpolation
    5        0.203862   2.99018e-05        interpolation

Zero found in the interval [0.193283, 0.386565]

parmhat = 1×2

    1.9624    4.9050
```

`wblfit` displays information about the iterations.

Input Arguments

x — Sample data

vector

Sample data, specified as a vector.

Data Types: `single` | `double`

alpha — Significance level

0.05 (default) | scalar in the range (0,1)

Significance level for the confidence intervals, specified as a scalar in the range (0,1). The confidence level is `100(1–alpha)`%, where `alpha` is the probability that the confidence intervals do not contain the true value.

Example: `0.01`

Data Types: `single` | `double`

censoring — Indicator for censoring

array of 0s (default) | logical vector

Indicator for the censoring of each value in `x`, specified as a logical vector of the same size as `x`. Use 1 for observations that are right-censored and 0 for observations that are fully observed.

The default is an array of 0s, meaning that all observations are fully observed.

Data Types: `logical`

`freq` — Frequency or weights of observations
array of 1s (default) | nonnegative vector

Frequency or weights of observations, specified as a nonnegative vector that is the same size as `x`. The `freq` input argument typically contains nonnegative integer counts for the corresponding elements in `x`, but can contain any nonnegative values.

To obtain the weighted MLEs for a data set with censoring, specify weights of observations, normalized to the number of observations in `x`.

The default is an array of 1s, meaning one observation per element of `x`.

Data Types: `single` | `double`

`options` — Optimization options
`statset('wblfit')` (default) | structure

Optimization options, specified as a structure. `options` determines the control parameters for the iterative algorithm that `wblfit` uses to compute MLEs for censored data.

Create `options` by using the function `statset` or by creating a structure array containing the fields and values described in this table.

Field Name	Value	Default Value
`Display`	Amount of information displayed by the algorithm. • `'off'` — Displays no information • `'final'` — Displays the final output • `'iter'` — Displays iterative output	`'off'`
`TolX`	Termination tolerance for the parameters, specified as a positive scalar	`1e-8`

You can also enter `statset('wblfit')` in the Command Window to see the names and default values of the fields that `wblfit` includes in the `options` structure.

Example: `statset('Display','iter')` specifies to display the information from each step of the iterative algorithm.

Data Types: `struct`

Output Arguments

`parmHat` — Estimate of parameters
1-by-2 row vector

Estimate of the parameters *a* (scale) and *b* (shape) of the Weibull distribution, returned as a row vector.

`parmCI` — Confidence intervals for parameters
2-by-2 matrix

Confidence intervals for the mean parameters of the Weibull distribution, returned as a 2-by-2 matrix vector containing the lower and upper bounds of the `100(1–alpha)`% confidence interval.

The first and second rows correspond to the lower and upper bounds of the confidence intervals, respectively.

Alternative Functionality

`wblfit` is a function specific to Weibull distribution. Statistics and Machine Learning Toolbox also offers the generic functions `mle`, `fitdist`, and `paramci` and the **Distribution Fitter** app, which support various probability distributions.

- `mle` returns MLEs and the confidence intervals of MLEs for the parameters of various probability distributions. You can specify the probability distribution name or a custom probability density function.
- Create a `WeibullDistribution` probability distribution object by fitting the distribution to data using the `fitdist` function or the **Distribution Fitter** app. The object properties `a` and `b` store the parameter estimates. To obtain the confidence intervals for the parameter estimates, pass the object to `paramci`.

Extended Capabilities

C/C++ Code Generation
Generate C and C++ code using MATLAB® Coder™.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`mle` | `wblcdf` | `wblinv` | `wbllike` | `wblpdf` | `wblplot` | `wblrnd` | `wblstat`

Topics
“Weibull Distribution” on page B-170

Introduced before R2006a

wblinv

Weibull inverse cumulative distribution function

Syntax

```
X = wblinv(P,A,B)
[X,XLO,XUP] = wblinv(P,A,B,PCOV,alpha)
```

Description

`X = wblinv(P,A,B)` returns the inverse cumulative distribution function (cdf) for a Weibull distribution with scale parameter `A` and shape parameter `B`, evaluated at the values in `P`. `P`, `A`, and `B` can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array of the same size as the other inputs. The default values for `A` and `B` are both `1`.

`[X,XLO,XUP] = wblinv(P,A,B,PCOV,alpha)` returns confidence bounds for `X` when the input parameters `A` and `B` are estimates. `PCOV` is a 2-by-2 matrix containing the covariance matrix of the estimated parameters. `alpha` has a default value of 0.05, and specifies 100(1 - `alpha`)% confidence bounds. `XLO` and `XUP` are arrays of the same size as `X` containing the lower and upper confidence bounds.

The function `wblinv` computes confidence bounds for `X` using a normal approximation to the distribution of the estimate

$$\log\widehat{a} + \frac{\log q}{\widehat{b}}$$

where q is the `P`th quantile from a Weibull distribution with scale and shape parameters both equal to 1. The computed bounds give approximately the desired confidence level when you estimate `mu`, `sigma`, and `PCOV` from large samples, but in smaller samples other methods of computing the confidence bounds might be more accurate.

The inverse of the Weibull cdf is

$$x = F^{-1}(p \mid a, b) = -a[\ln(1-p)]^{1/b}.$$

Examples

The lifetimes (in hours) of a batch of light bulbs has a Weibull distribution with parameters `a = 200` and `b = 6`.

Find the median lifetime of the bulbs:

```
life = wblinv(0.5, 200, 6)
life =
 188.1486
```

Generate 100 random values from this distribution, and estimate the 90th percentile (with confidence bounds) from the random sample

```
x = wblrnd(200,6,100,1);
p = wblfit(x)
[nlogl,pcov] = wbllike(p,x)
[q90,q90lo,q90up] = wblinv(0.9,p(1),p(2),pcov)
p =

  204.8918    6.3920


nlogl =

  496.8915


pcov =

   11.3392    0.5233
    0.5233    0.2573


q90 =

  233.4489


q90lo =

  226.0092


q90up =

  241.1335
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

icdf | wblcdf | wblfit | wbllike | wblpdf | wblplot | wblrnd | wblstat

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wbllike

Weibull negative log-likelihood

Syntax

```
nlogL = wbllike(params,data)
[logL,AVAR] = wbllike(params,data)
[...] = wbllike(params,data,censoring)
[...] = wbllike(params,data,censoring,freq)
```

Description

`nlogL = wbllike(params,data)` returns the Weibull log-likelihood. `params(1)` is the scale parameter, `A`, and `params(2)` is the shape parameter, `B`.

`[logL,AVAR] = wbllike(params,data)` also returns `AVAR`, which is the asymptotic variance-covariance matrix of the parameter estimates if the values in `params` are the maximum likelihood estimates. `AVAR` is the inverse of Fisher's information matrix. The diagonal elements of `AVAR` are the asymptotic variances of their respective parameters.

`[...] = wbllike(params,data,censoring)` accepts a Boolean vector, `censoring`, of the same size as `data`, which is 1 for observations that are right-censored and 0 for observations that are observed exactly.

`[...] = wbllike(params,data,censoring,freq)` accepts a frequency vector, `freq`, of the same size as `data`. `freq` typically contains integer frequencies for the corresponding elements in `data`, but can contain any nonnegative values. Pass in `[]` for `censoring` to use its default value.

The Weibull negative log-likelihood for uncensored data is

$$(-\log L) = -\log \prod_{i=1} f(a,b\,|\,x_i) = -\sum_{i=1}^{n} \log f(a,b\,|\,x_i)$$

where f is the Weibull pdf.

`wbllike` is a utility function for maximum likelihood estimation.

Examples

This example continues the example from `wblfit`.

```
r = wblrnd(0.5,0.8,100,1);
[logL, AVAR] = wbllike(wblfit(r),r)
logL =
  47.3349
AVAR =
  0.0048  0.0014
  0.0014  0.0040
```

References

[1] Patel, J. K., C. H. Kapadia, and D. B. Owen. *Handbook of Statistical Distributions*. New York: Marcel Dekker, 1976.

See Also

`wblcdf` | `wblfit` | `wblinv` | `wblpdf` | `wblplot` | `wblrnd` | `wblstat`

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wblpdf

Weibull probability density function

Syntax

```
y = wblpdf(x)
y = wblpdf(x,a)
y = wblpdf(x,a,b)
```

Description

`y = wblpdf(x)` returns the probability density function (pdf) of the Weibull distribution with unit parameters, evaluated at the values in `x`.

`y = wblpdf(x,a)` returns the pdf of the Weibull distribution with scale parameter `a` and unit shape, evaluated at the values in `x`. This is equivalent to the pdf of the exponential distribution.

`y = wblpdf(x,a,b)` returns the pdf of the Weibull distribution with scale parameter `a` and shape parameter `b`, evaluated at the values in `x`.

Examples

Compute Weibull pdf

Compute the density of the observed value `3` in the Weibull distribution unit scale and shape.

```
y1 = wblpdf(3)
```

```
y1 = 0.0498
```

Compute the density of the observed value `3` in the Weibull distributions with scale parameter `2` and shape parameters `1` through `5`.

```
y2 = wblpdf(3,2,1:5)
```

```
y2 = 1×5

    0.1116    0.1581    0.1155    0.0427    0.0064
```

Compare Weibull and Exponential pdfs

The exponential distribution with parameter `mu` is a special case of the Weibull distribution, where `a = mu` and `b = 1`.

Compute the density of sample observations in the exponential distributions with means `1` through `5` using `expcdf`.

```
x = 0.2:0.2:1;
mu = 1:5;
y1 = exppdf(x,mu)
```

```
y1 = 1×5

    0.8187    0.4094    0.2729    0.2047    0.1637
```

Compute the density of the same sample observations using `wblpdf` where the scale parameter is equal to `mu` and the shape parameter is `1`.

```
y2 = wblpdf(x,mu)
```

```
y2 = 1×5

    0.8187    0.4094    0.2729    0.2047    0.1637
```

The two functions return the same values.

Input Arguments

x — Values at which to evaluate pdf

nonnegative scalar value | array of nonnegative scalar values

Values at which to evaluate the pdf, specified as a nonnegative scalar value or an array of nonnegative scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `wblpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[3 4 7 9]`

Data Types: `single` | `double`

a — Scale parameter

1 (default) | positive scalar value | array of positive scalar values

Scale parameter of the Weibull distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `wblpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

b — Shape parameter
1 (default) | positive scalar value | array of positive scalar values

Shape parameter of the Weibull distribution, specified as a positive scalar value or an array of positive scalar values.

- To evaluate the pdf at multiple values, specify `x` using an array.
- To evaluate the pdfs of multiple distributions, specify `a` and `b` using arrays.

If one or more of the input arguments `x`, `a`, and `b` are arrays, then the array sizes must be the same. In this case, `wblpdf` expands each scalar input into a constant array of the same size as the array inputs. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

Example: `[1 1 2 2]`

Data Types: `single` | `double`

Output Arguments

y — pdf values
scalar value | array of scalar values

pdf values evaluated at the values in `x`, returned as a scalar value or an array of scalar values. `y` is the same size as `x`, `a`, and `b` after any necessary scalar expansion. Each element in `y` is the pdf value of the distribution specified by the corresponding elements in `a` and `b`, evaluated at the corresponding element in `x`.

More About

Weibull pdf

The Weibull distribution is a two-parameter family of curves. The parameters a and b are scale and shape, respectively.

The Weibull pdf is

$$f(x|a,b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-(x/a)^b}.$$

Some instances refer to the Weibull distribution with a single parameter, which corresponds to `wblpdf` with $a = 1$.

For more information, see “Weibull Distribution” on page B-170.

Alternative Functionality

- `wblpdf` is a function specific to the Weibull distribution. Statistics and Machine Learning Toolbox also offers the generic function `pdf`, which supports various probability distributions. To use `pdf`, create a `WeibullDistribution` probability distribution object and pass the object as an input argument or specify the probability distribution name and its parameters. Note that the distribution-specific function `wblpdf` is faster than the generic function `pdf`.

- Use the **Probability Distribution Function** app to create an interactive plot of the cumulative distribution function (cdf) or probability density function (pdf) for a probability distribution.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`WeibullDistribution` | `pdf` | `wblcdf` | `wblfit` | `wblinv` | `wbllike` | `wblplot` | `wblrnd` | `wblstat`

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wblplot

Weibull probability plot

Syntax

```
wblplot(x)
wblplot(ax,x)
h = wblplot( ___ )
```

Description

`wblplot(x)` creates a Weibull probability plot comparing the distribution of the data in `x` to the Weibull distribution.

`wblplot` plots each data point in `x` using plus sign (`'+'`) markers and draws two reference lines that represent the theoretical distribution. A solid reference line connects the first and third quartiles of the data, and a dashed reference line extends the solid line to the ends of the data. If the sample data has a Weibull distribution, then the data points appear along the reference line. A distribution other than Weibull introduces curvature in the data plot.

`wblplot(ax,x)` adds a Weibull probability plot into the axes specified by `ax`.

`h = wblplot( ___ )` returns graphics handles corresponding to the plotted lines, using any of the input arguments in the previous syntaxes.

Examples

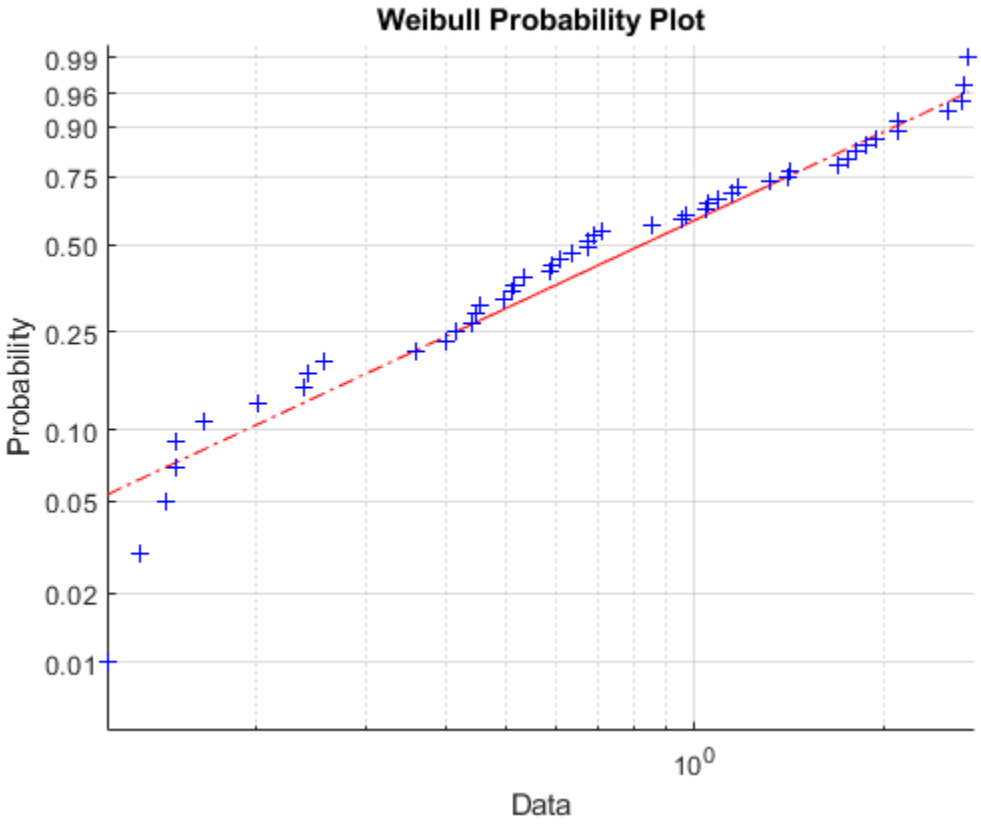
Create Weibull Probability Plot

Generate a vector `r` containing 50 random numbers from the Weibull distribution with the scale parameter 1.2 and the shape parameter 1.5.

```
rng('default')  % For reproducibility
r = wblrnd(1.2,1.5,50,1);
```

Create a Weibull probability plot to visually determine if the data comes from a Weibull distribution.

```
wblplot(r)
```



The plot indicates that the data likely comes from a Weibull distribution.

Test for Weibull Distribution

Generate two sample data sets, one from a Weibull distribution and another from a lognormal distribution. Perform the Lilliefors test to assess whether each data set is from a Weibull distribution. Confirm the test decision by performing a visual comparison using a Weibull probability plot (`wblplot`).

Generate samples from a Weibull distribution.

```
rng('default')
data1 = wblrnd(0.5,2,[500,1]);
```

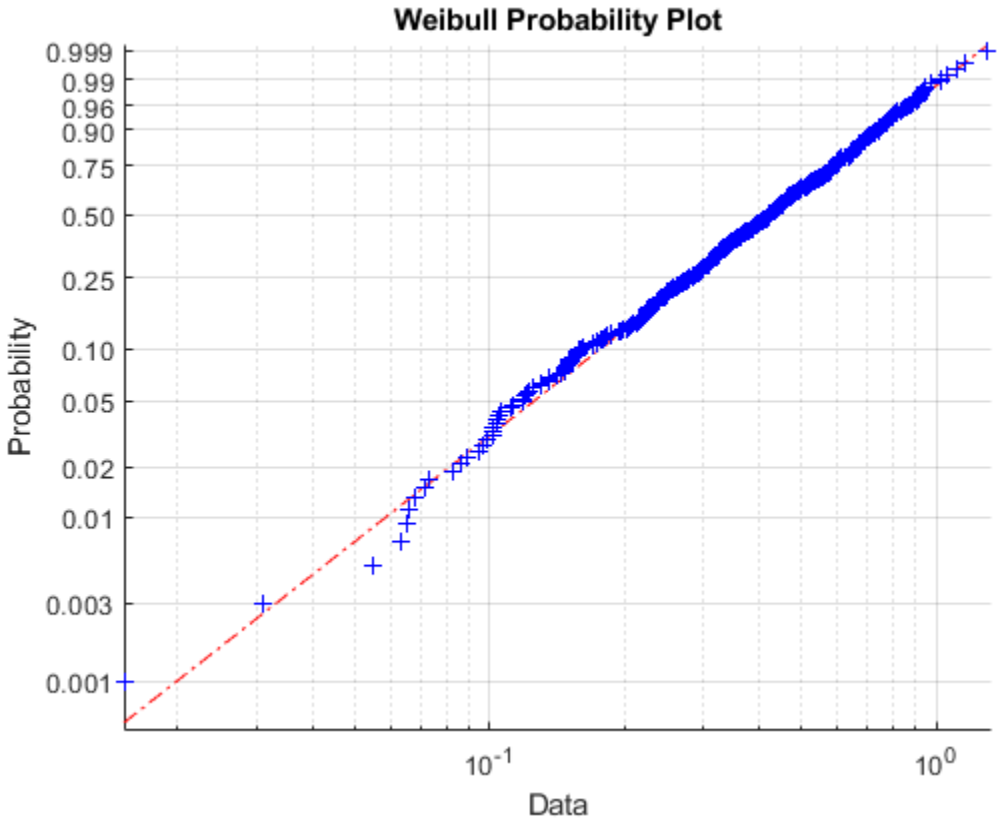
Perform the Lilliefors test by using the `lillietest`. To test data for a Weibull distribution, test if the logarithm of the data has an extreme value distribution.

```
h1 = lillietest(log(data1),'Distribution','extreme value')
```

```
h1 = 0
```

The returned value of `h1 = 0` indicates that `lillietest` fails to reject the null hypothesis at the default 5% significance level. Confirm the test decision using a Weibull probability plot.

```
wblplot(data1)
```



The plot indicates that the data follows a Weibull distribution.

Generate samples from a lognormal distribution.

```
data2 =lognrnd(5,2,[500,1]);
```

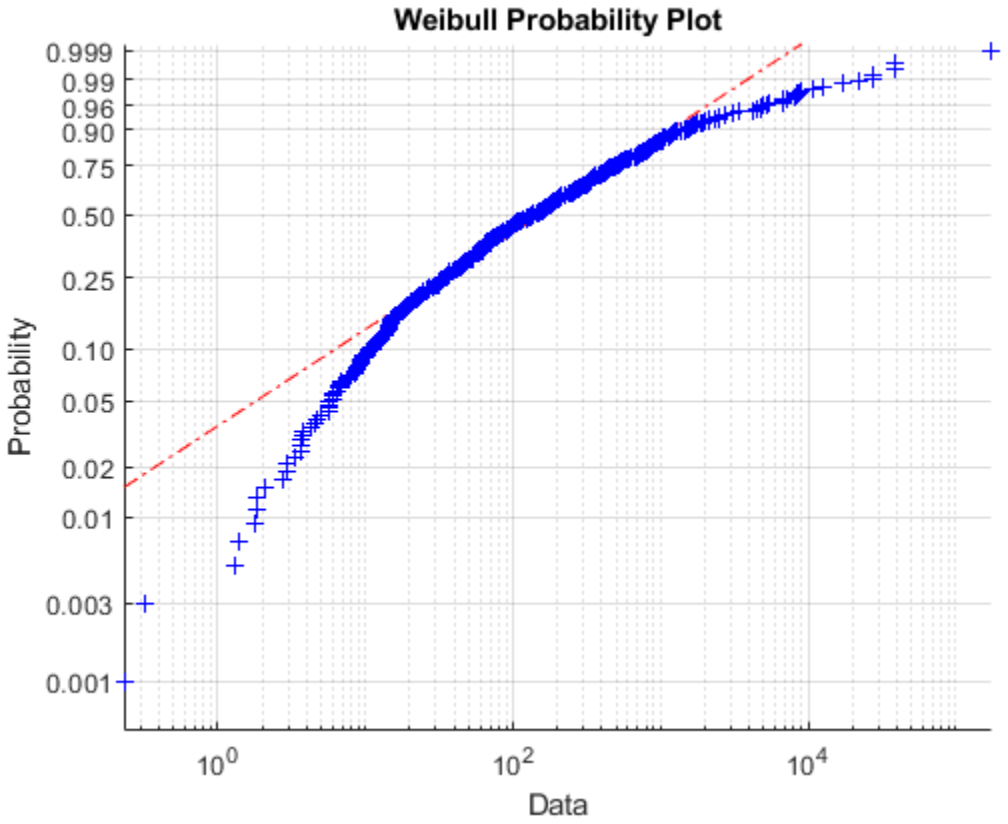
Perform the Lilliefors test.

```
h2 = lillietest(log(data2),'Distribution','extreme value')
```

```
h2 = 1
```

The returned value of `h2 = 1` indicates that `lillietest` rejects the null hypothesis at the default 5% significance level. Confirm the test decision using a Weibull probability plot.

```
wblplot(data2)
```



The plot indicates that the data does not follow a Weibull distribution.

Input Arguments

x — Sample data

numeric vector | numeric matrix

Sample data, specified as a numeric vector or numeric matrix. `wblplot` displays each value in `x` using the symbol `'+'`. If `x` is a matrix, then `wblplot` displays a separate line for each column of `x`.

Data Types: `single` | `double`

ax — Target axes

`Axes` object | `UIAxes` object

Target axes, specified as an `Axes` object or a `UIAxes` object. `wblplot` adds an additional plot into the axes specified by `ax`. For details, see Axes Properties and UIAxes Properties.

Use `gca` to return the current axes for the current figure.

Output Arguments

h — Graphics handles for line objects

vector of `Line` graphics handles

Graphics handles for line objects, returned as a vector of `Line` graphics handles. Graphics handles are unique identifiers that you can use to query and modify the properties of a specific line on the plot. For each column of `x`, `wblplot` returns three handles:

- The line representing the data points. `wblplot` represents each data point in `x` using plus sign (`'+'`) markers.
- The line joining the first and third quartiles of each column of `x`, represented as a solid line.
- The extrapolation of the quartile line, extended to the minimum and maximum values of `x`, represented as a dashed line.

To view and set properties of line objects, use dot notation. For information on using dot notation, see "Access Property Values". For information on the `Line` properties that you can set, see Primitive Line.

Algorithms

`wblplot` matches the quantiles of sample data to the quantiles of a Weibull distribution. The sample data is sorted, scaled logarithmically, and plotted on the x-axis. The y-axis represents the quantiles of the Weibull distribution, converted into probability values. Therefore, the y-axis scaling is not linear.

Where the x-axis value is the ith sorted value from a sample of size N, the y-axis value is the midpoint between evaluation points of the empirical cumulative distribution function of the data. The midpoint is equal to $\frac{(i - 0.5)}{N}$.

`wblplot` superimposes a reference line to assess the linearity of the plot. The line goes through the first and third quartiles of the data.

Alternative Functionality

You can use the `probplot` function to create a probability plot. The `probplot` function enables you to indicate censored data and specify the distribution for a probability plot.

See Also

`ecdf` | `normplot` | `probplot` | `wblcdf` | `wblfit` | `wblrnd`

Topics

"Distribution Plots" on page 4-7
"Hypothesis Testing" on page 8-5
"Weibull Distribution" on page B-170

Introduced before R2006a

wblrnd

Weibull random numbers

Syntax

```
r = wblrnd(a,b)
r = wblrnd(a,b,sz1,...,szN)
r = wblrnd(a,b,sz)
```

Description

`r = wblrnd(a,b)` generates a random number from the Weibull distribution with scale `a` and shape `b`.

`r = wblrnd(a,b,sz1,...,szN)` generates an array of random numbers from the Weibull distribution, where `sz1,...,szN` indicates the size of each dimension.

`r = wblrnd(a,b,sz)` generates an array of random numbers from the Weibull distribution where size vector `sz` specifies `size(r)`.

Examples

Generate Weibull Random Number

Generate a single random number from the Weibull distribution with scale `4` and shape `3`.

```
r = wblrnd(4,3)
```

```
r = 2.3582
```

Generate Weibull Random Numbers

Generate a 1-by-5 array of random numbers drawn from the Weibull distributions with scale `3` and shape values `1` through `5`.

```
a1 = 3;
b1 = 1:5;
r1 = wblrnd(a1,b1)
```

```
r1 = 1×5

    0.6147    0.9437    3.8195    1.6459    2.5666
```

If you specify array dimensions, they must match the dimensions of `a` and `b` after any scalar expansion.

Generate a 1-by-6 array of random numbers drawn from the Weibull distributions with scale values 1 through 6 and shape values 5 through 10, respectively.

```
a2 = 1:6;
b2 = 5:10;
sz1 = 1;
sz2 = 6;
r2 = wblrnd(a2,b2,sz1,sz2)
```

```
r2 = 1×6

    1.1841    2.0836    2.7912    2.7026    3.4531    6.3799
```

Generate a 2-by-3 array of random numbers from the Weibull distribution with scale 4 and shape 5.

```
sz = [2 3];
r3 = wblrnd(4,5,sz)
```

```
r3 = 2×3

    1.9817    3.7486    4.5729
    2.1395    2.9624    3.8841
```

Input Arguments

a — Scale parameter

positive scalar value | array of positive scalar values

Scale parameter of the Weibull distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If either or both of the input arguments `a` and `b` are arrays, then the array sizes must be the same. In this case, `wblrnd` expands each scalar input into a constant array of the same size as the array inputs. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[1 2 3 5]`

Data Types: `single` | `double`

b — Shape parameter

positive scalar value | array of positive scalar values

Shape parameter of the Weibull distribution, specified as a positive scalar value or an array of positive scalar values.

To generate random numbers from multiple distributions, specify `a` and `b` using arrays. If either or both of the input arguments `a` and `b` are arrays, then the array sizes must be the same. In this case, `wblrnd` expands each scalar input into a constant array of the same size as the array inputs. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Example: `[1 1 2 2]`

Data Types: `single` | `double`

`sz1,...,szN` — Size of each dimension (as separate arguments)
integers

Size of each dimension, specified as separate arguments of integers.

If `a` and `b` are arrays, then the specified dimensions `sz1,...,szN` must match the dimensions of `a` and `b`. The default values of `sz1,...,szN` are the dimensions of `a` and `b`.

- If you specify a single value `sz1`, then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `wblrnd` ignores trailing dimensions with a size of `1`. For example, `wblrnd(2,5,3,1,1,1)` produces a 3-by-1 vector of random numbers from the distribution with scale `2` and shape `5`.

Example: `3,5`

Data Types: `single` | `double`

`sz` — Size of each dimension (as a row vector)
row vector of integers

Size of each dimension, specified as a row vector of integers.

If `a` and `b` are arrays, then the specified dimensions `sz` must match the dimensions of `a` and `b`. The default values of `sz` are the dimensions of `a` and `b`.

- If you specify a single value [`sz1`], then `r` is a square matrix of size `sz1`-by-`sz1`.
- If the size of any dimension is `0` or negative, then `r` is an empty array.
- Beyond the second dimension, `wblrnd` ignores trailing dimensions with a size of `1`. For example, `wblrnd(2,5,[3 1 1 1])` produces a 3-by-1 vector of random numbers from the distribution with scale `2` and shape `5`.

Example: `[3 5]`

Data Types: `single` | `double`

Output Arguments

`r` — Weibull random numbers
scalar value | array of scalar values

Weibull random numbers, returned as a scalar value or an array of scalar values with the dimensions specified by `sz1,...,szN` or `sz`. Each element in `r` is the random number generated from the distribution specified by the corresponding elements in `a` and `b`.

Alternative Functionality

- `wblrnd` is a function specific to the Weibull distribution. Statistics and Machine Learning Toolbox also offers the generic function `random`, which supports various probability distributions. To use `random`, specify the probability distribution name and its parameters. Note that the distribution-specific function `wblrnd` is faster than the generic function `random`.

- To generate random numbers interactively, use `randtool`, a user interface for random number generation.

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

The generated code can return a different sequence of numbers from the sequence returned by MATLAB if either of the following is true:

- The output is nonscalar.
- An input parameter is invalid for the distribution.

For more information on code generation, see “Introduction to Code Generation” on page 32-2 and “General Code Generation Workflow” on page 32-5.

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`WeibullDistribution` | `random` | `wblcdf` | `wblfit` | `wblinv` | `wbllike` | `wblpdf` | `wblplot` | `wblstat`

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wblstat

Weibull mean and variance

Syntax

```
[M,V] = wblstat(A,B)
```

Description

`[M,V] = wblstat(A,B)` returns the mean of and variance for the Weibull distribution with scale parameter, `A` and shape parameter, `B`. Vector or matrix inputs for `A` and `B` must have the same size, which is also the size of `M` and `V`. A scalar input for `A` or `B` is expanded to a constant matrix with the same dimensions as the other input.

The mean of the Weibull distribution with parameters a and b is

$$a\left[\Gamma\left(1+b^{-1}\right)\right]$$

and the variance is

$$a^2\left[\Gamma\left(1+2b^{-1}\right)-\Gamma\left(1+b^{-1}\right)^2\right]$$

Examples

```
[m,v] = wblstat(1:4,1:4)
m =
  1.0000  1.7725  2.6789  3.6256
v =
  1.0000  0.8584  0.9480  1.0346

wblstat(0.5,0.7)
ans =
  0.6329
```

Extended Capabilities

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

See Also

`wblcdf` | `wblfit` | `wblinv` | `wbllike` | `wblpdf` | `wblplot` | `wblrnd`

Topics

“Weibull Distribution” on page B-170

Introduced before R2006a

wishrnd

Wishart random numbers

Syntax

```
W = wishrnd(Sigma,df)
W = wishrnd(Sigma,df,D)
[W,D] = wishrnd(Sigma,df)
```

Description

`W = wishrnd(Sigma,df)` generates a random matrix `W` having the Wishart distribution with covariance matrix `Sigma` and with `df` degrees of freedom. The inverse of `W` has the Inverse Wishart distribution with parameters `Tau = inv(Sigma)` and `df` degrees of freedom.

`W = wishrnd(Sigma,df,D)` expects `D` to be the Cholesky factor of `Sigma`. If you call `wishrnd` multiple times using the same value of `Sigma`, it's more efficient to supply `D` instead of computing it each time.

`[W,D] = wishrnd(Sigma,df)` returns `D` so you can provide it as input in future calls to `wishrnd`.

This function defines the parameter `Sigma` so that the mean of the output matrix is `Sigma*df`

See Also

`iwishrnd`

Topics

“Wishart Distribution” on page B-178

Introduced before R2006a

xptread

Create table from data stored in SAS XPORT format file

Syntax

```
data = xptread
data = xptread(filename)
[data,missing] = xptread(filename)
xptread(...,'ReadObsNames',true)
```

Description

`data = xptread` displays a dialog box for selecting a file, then reads data from the file into a table. The file must be in the `SAS XPORT` format.

`data = xptread(filename)` retrieves data from a `SAS XPORT` format file `filename`. For example, to open a `SAS XPORT` file named `sample.xpt`, type `data = xptread('sample.xpt')` at the command prompt.

`[data,missing] = xptread(filename)` returns a nominal array, `missing`, that contains the missing data type information from the `XPORT` format file.

The `XPORT` format allows for 28 missing data types, represented in the file by an upper case letter or the characters `'.'` or `'_'`. If the `XPORT` file contains missing values, `xptread` converts them to `NaN` values in the output table, `data`. However, if you need the specific missing types, you can recover this information by specifying a second output, `missing`. The entries in `missing` are one of the 28 missing data type values from the `XPORT` format file (`'.'`, `'_'`, `'A'`,...,`'Z'`), or are undefined for values that are not present at all in the `XPORT` format file. The outputs `missing` and `data` are the same size.

`xptread(...,'ReadObsNames',true)` treats the first variable in the file as observation names. The default value is false.

`xptread` only supports single data sets per file. `xptread` does not support compressed files.

See Also

`table`

Introduced in R2009b

x2fx

Convert predictor matrix to design matrix

Syntax

```
D = x2fx(X,model)
D = x2fx(X,model,categ)
D = x2fx(X,model,categ,catlevels)
```

Description

`D = x2fx(X,model)` converts a matrix of predictors `X` to a design matrix `D` for regression analysis. Distinct predictor variables should appear in different columns of `X`.

The optional input `model` controls the regression model. By default, `x2fx` returns the design matrix for a linear additive model with a constant term. `model` is one of the following:

- `'linear'` — Constant and linear terms. This is the default.
- `'interaction'` — Constant, linear, and interaction terms
- `'quadratic'` — Constant, linear, interaction, and squared terms
- `'purequadratic'` — Constant, linear, and squared terms

If `X` has *n* columns, the order of the columns of `D` for a full quadratic model is:

1. The constant term
2. The linear terms (the columns of `X`, in order 1, 2, ..., *n*)
3. The interaction terms (pairwise products of the columns of `X`, in order (1, 2), (1, 3), ..., (1, *n*), (2, 3), ..., (*n*–1, *n*))
4. The squared terms (in order 1, 2, ..., *n*)

Other models use a subset of these terms, in the same order.

Alternatively, `model` can be a matrix specifying polynomial terms of arbitrary order. In this case, `model` should have one column for each column in `X` and one row for each term in the model. The entries in any row of `model` are powers for the corresponding columns of `X`. For example, if `X` has columns `X1`, `X2`, and `X3`, then a row `[0 1 2]` in `model` specifies the term `(X1.^0).*(X2.^1).*(X3.^2)`. A row of all zeros in `model` specifies a constant term, which can be omitted.

`D = x2fx(X,model,categ)` treats columns with numbers listed in the vector `categ` as categorical variables. Terms involving categorical variables produce dummy variable columns in `D`. Dummy variables are computed under the assumption that possible categorical levels are completely enumerated by the unique values that appear in the corresponding column of `X`.

`D = x2fx(X,model,categ,catlevels)` accepts a vector `catlevels` the same length as `categ`, specifying the number of levels in each categorical variable. In this case, values in the corresponding column of `X` must be integers in the range from 1 to the specified number of levels. Not all of the levels need to appear in `X`.

Examples

Convert Predictor Matrix to Design Matrix

Convert two predictors X1 and X2 (the columns of X) into a design matrix for a full quadratic model with terms `constant`, X1, X2, `X1.*X2`, `X1.^2`, and `X2.^2`.

```
X = [1 10
     2 20
     3 10
     4 20
     5 15
     6 15];
D = x2fx(X,'quadratic')
```

```
D = 6×6

     1     1    10    10     1   100
     1     2    20    40     4   400
     1     3    10    30     9   100
     1     4    20    80    16   400
     1     5    15    75    25   225
     1     6    15    90    36   225
```

Convert two predictors X1 and X2 (the columns of X) into a design matrix for a quadratic model with terms `constant`, X1, X2, `X1.*X2`, and `X1.^2`.

```
X = [1 10
     2 20
     3 10
     4 20
     5 15
     6 15];
model = [0 0
         1 0
         0 1
         1 1
         2 0];
D = x2fx(X,model)
```

```
D = 6×5

     1     1    10    10     1
     1     2    20    40     4
     1     3    10    30     9
     1     4    20    80    16
     1     5    15    75    25
     1     6    15    90    36
```

See Also

`candexch` | `candgen` | `cordexch` | `regstats` | `rowexch` | `rstool`

Introduced before R2006a

zscore

Standardized *z*-scores

Syntax

```
Z = zscore(X)
Z = zscore(X,flag)
Z = zscore(X,flag,'all')
Z = zscore(X,flag,dim)
Z = zscore(X,flag,vecdim)
[Z,mu,sigma] = zscore( ___ )
```

Description

`Z = zscore(X)` returns the *z*-score on page 33-6643 for each element of `X` such that columns of `X` are centered to have mean 0 and scaled to have standard deviation 1. `Z` is the same size as `X`.

- If `X` is a vector, then `Z` is a vector of *z*-scores.
- If `X` is a matrix, then `Z` is a matrix of the same size as `X`, and each column of `Z` has mean 0 and standard deviation 1.
- For multidimensional arrays on page 33-6644, *z*-scores in `Z` are computed along the first nonsingleton dimension on page 33-6644 of `X`.

`Z = zscore(X,flag)` scales `X` using the standard deviation indicated by `flag`.

- If `flag` is 0 (default), then `zscore` scales `X` using the sample standard deviation on page 33-6644, with *n* - 1 in the denominator of the standard deviation formula. `zscore(X,0)` is the same as `zscore(X)`.
- If `flag` is 1, then `zscore` scales `X` using the population standard deviation on page 33-6644, with *n* in the denominator of standard deviation formula.

`Z = zscore(X,flag,'all')` standardizes `X` by using the mean and standard deviation of all the values in `X`.

`Z = zscore(X,flag,dim)` standardizes `X` along the operating dimension `dim`. For example, for a matrix `X`, if `dim` = 1, then `zscore` uses the means and standard deviations along the columns of `X`, if `dim` = 2, then `zscore` uses the means and standard deviations along the rows of `X`.

`Z = zscore(X,flag,vecdim)` standardizes `X` over the dimensions specified by the vector `vecdim`. For example, if `X` is a matrix, then `zscore(X,0,[1 2])` is equivalent to `zscore(X,0,'all')` because every element of a matrix is contained in the array slice defined by dimensions 1 and 2.

`[Z,mu,sigma] = zscore( ___ )` also returns the means and standard deviations used for centering and scaling, `mu` and `sigma`, respectively. You can use any of the input arguments in the previous syntaxes.

Examples

Z-Scores of Two Data Vectors

Compute and plot the *z*-scores of two data vectors, and then compare the results.

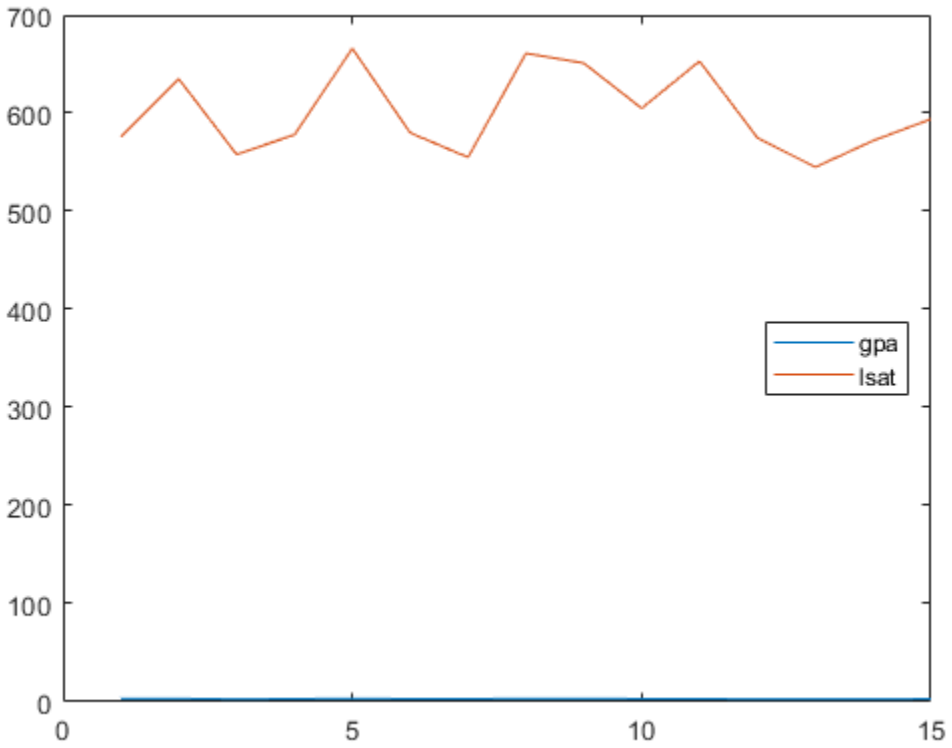
Load the sample data.

```
load lawdata
```

Two variables load into the workspace: `gpa` and `lsat`.

Plot both variables on the same axes.

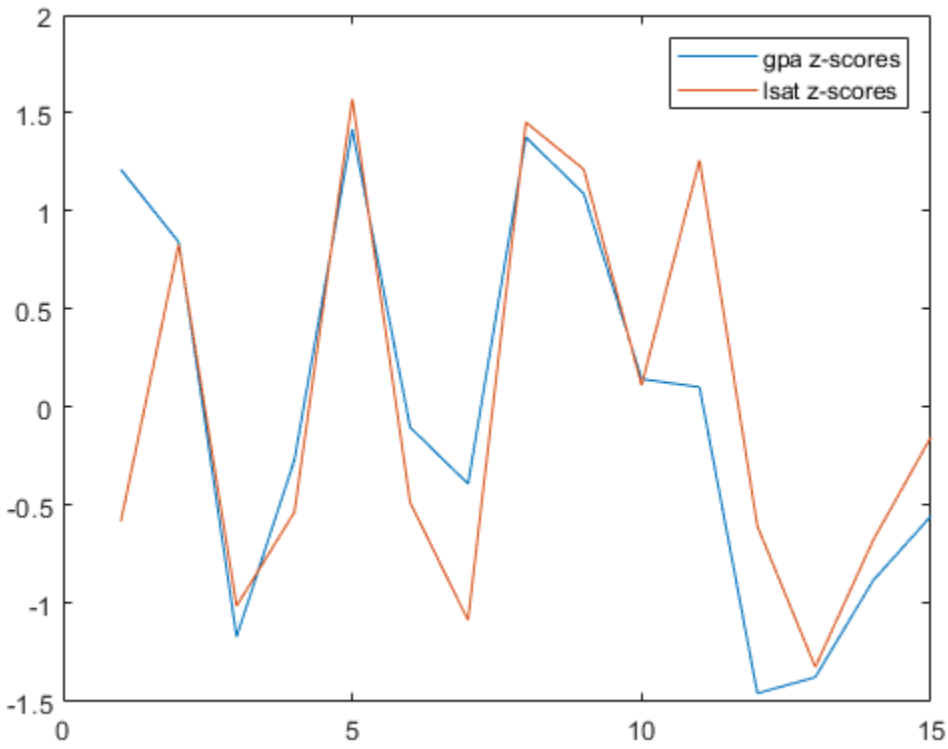
```
plot([gpa,lsat])
legend('gpa','lsat','Location','East')
```



It is difficult to compare these two measures because they are on a very different scale.

Plot the *z*-scores of `gpa` and `lsat` on the same axes.

```
Zgpa = zscore(gpa);
Zlsat = zscore(lsat);
plot([Zgpa, Zlsat])
legend('gpa z-scores','lsat z-scores','Location','Northeast')
```



Now, you can see the relative performance of individuals with respect to both their `gpa` and `lsat` results. For example, the third individual's `gpa` and `lsat` results are both one standard deviation below the sample mean. The eleventh individual's `gpa` is around the sample mean but has an `lsat` score almost 1.25 standard deviations above the sample average.

Check the mean and standard deviation of the *z*-scores you created.

```
mean([Zgpa,Zlsat])
```

```
ans = 1×2
10^-14 ×

   -0.1088    0.0357
```

```
std([Zgpa,Zlsat])
```

```
ans = 1×2

     1     1
```

By definition, *z*-scores of `gpa` and `lsat` have mean 0 and standard deviation 1.

Z-Scores for a Population vs. Sample

Load the sample data.

```
load lawdata
```

Two variables load into the workspace: `gpa` and `lsat`.

Compute the *z*-scores of `gpa` using the population formula for standard deviation.

```
Z1 = zscore(gpa,1); % population formula
Z0 = zscore(gpa,0); % sample formula
disp([Z1 Z0])
```

```
    1.2554    1.2128
    0.8728    0.8432
   -1.2100   -1.1690
   -0.2749   -0.2656
    1.4679    1.4181
   -0.1049   -0.1013
   -0.4024   -0.3888
    1.4254    1.3771
    1.1279    1.0896
    0.1502    0.1451
    0.1077    0.1040
   -1.5076   -1.4565
   -1.4226   -1.3743
   -0.9125   -0.8815
   -0.5724   -0.5530
```

For a sample from a population, the population standard deviation formula with *n* in the denominator corresponds to the maximum likelihood estimate of the population standard deviation, and might be biased. The sample standard deviation formula, on the other hand, is the unbiased estimator of the population standard deviation for a sample.

Z-Scores of a Data Matrix

Compute *z*-scores using the mean and standard deviation computed along the columns or rows of a data matrix.

Load the sample data.

```
load flu
```

The dataset array `flu` is loaded in the workplace. `flu` has 52 observations on 11 variables. The first variable contains dates (in weeks). The other variables contain the flu estimates for different regions in the U.S.

Convert the dataset array to a data matrix.

```
flu2 = double(flu(:,2:end));
```

The new data matrix, `flu2`, is a 52-by-10 double data matrix. The rows correspond to the weeks and the columns correspond to the U.S. regions in the data set array `flu`.

Standardize the flu estimate for each region (the *columns* of `flu2`).

```
Z1 = zscore(flu2,[ ],1);
```

You can see the *z*-scores in the variable editor by double-clicking on the matrix `Z1` created in the workspace.

Standardize the flu estimate for each week (the *rows* of `flu2`).

```
Z2 = zscore(flu2,[ ],2);
```

Z-Scores of Multidimensional Array

Find the z-scores of a multidimensional array by specifying to standardize the data along different dimensions. Compare the results when using the `'all'`, `dim`, and `vecdim` input arguments.

Create a 3-by-4-by-2 array.

```
X = reshape(1:24,[3 4 2])
```

```
X =
X(:,:,1) =

     1     4     7    10
     2     5     8    11
     3     6     9    12


X(:,:,2) =

    13    16    19    22
    14    17    20    23
    15    18    21    24
```

Standardize `X` by using the mean and standard deviation of all the values in `X`.

```
Zall = zscore(X,0,'all')
```

```
Zall =
Zall(:,:,1) =

   -1.6263   -1.2021   -0.7778   -0.3536
   -1.4849   -1.0607   -0.6364   -0.2121
   -1.3435   -0.9192   -0.4950   -0.0707


Zall(:,:,2) =

    0.0707    0.4950    0.9192    1.3435
    0.2121    0.6364    1.0607    1.4849
    0.3536    0.7778    1.2021    1.6263
```

The resulting multidimensional array of z-scores has mean 0 and standard deviation 1. For example, compute the mean and standard deviation of `Zall`.

```
mZall = mean(Zall(:,:,:),'all')

mZall = -9.2519e-18

sZall = std(Zall(:,:,:),0,'all')

sZall = 1.0000
```

Now standardize X along the second dimension.

```
Zdim = zscore(X,0,2)

Zdim =
Zdim(:,:,1) =

   -1.1619   -0.3873    0.3873    1.1619
   -1.1619   -0.3873    0.3873    1.1619
   -1.1619   -0.3873    0.3873    1.1619


Zdim(:,:,2) =

   -1.1619   -0.3873    0.3873    1.1619
   -1.1619   -0.3873    0.3873    1.1619
   -1.1619   -0.3873    0.3873    1.1619
```

The elements in each row of each page of `Zdim` have mean 0 and standard deviation 1. For example, compute the mean and standard deviation of the first row of the second page of `Zdim`.

```
mZdim = mean(Zdim(1,:,2),'all')

mZdim = 0

sZdim = std(Zdim(1,:,2),0,'all')

sZdim = 1
```

Finally, standardize X based on the second and third dimensions.

```
Zvecdim = zscore(X,0,[2 3])

Zvecdim =
Zvecdim(:,:,1) =

   -1.4289   -1.0206   -0.6124   -0.2041
   -1.4289   -1.0206   -0.6124   -0.2041
   -1.4289   -1.0206   -0.6124   -0.2041


Zvecdim(:,:,2) =

    0.2041    0.6124    1.0206    1.4289
    0.2041    0.6124    1.0206    1.4289
    0.2041    0.6124    1.0206    1.4289
```

The elements in each `Zvecdim(i,:,:)` slice have mean 0 and standard deviation 1. For example, compute the mean and standard deviation of the elements in `Zvecdim(1,:,:)`.

```
mZvecdim = mean(Zvecdim(1,:,:),'all')
```

```
mZvecdim = 2.7756e-17
```

```
sZvecdim = std(Zvecdim(1,:,:),0,'all')
```

```
sZvecdim = 1
```

Z-Scores, Mean, and Standard Deviation

Return the mean and standard deviation used to compute the *z*-scores.

Load the sample data.

```
load lawdata
```

Two variables load into the workspace: `gpa` and `lsat`.

Return the *z*-scores, mean, and standard deviation of `gpa`.

```
[Z,gpamean,gpastdev] = zscore(gpa)
```

```
Z = 15×1

    1.2128
    0.8432
   -1.1690
   -0.2656
    1.4181
   -0.1013
   -0.3888
    1.3771
    1.0896
    0.1451
      ⋮
```

```
gpamean = 3.0947
```

```
gpastdev = 0.2435
```

Input Arguments

X — Input data

vector | matrix | multidimensional array

Input data, specified as a vector, matrix, or multidimensional array.

Data Types: `double` | `single`

`flag` — Indicator for the standard deviation

0 (default) | 1

Indicator for the standard deviation used to compute the *z*-scores, specified as 0 or 1.

- If `flag` is 0 (default), then `zscore` scales `X` using the sample standard deviation on page 33-6644. `zscore(X,0)` is the same as `zscore(X)`.
- If `flag` is 1, then `zscore` scales `X` using the population standard deviation on page 33-6644.

`dim` — Dimension
positive integer scalar

Dimension along which to calculate the *z*-scores of `X`, specified as a positive integer scalar. If you do not specify a value, then the default value is the first array dimension whose size does not equal 1.

For example, for a matrix `X`, if `dim` = 1, then `zscore` uses the means and standard deviations along the columns of `X`, and if `dim` = 2, then `zscore` uses the means and standard deviations along the rows of `X`.

`vecdim` — Vector of dimensions
positive integer vector

Vector of dimensions along which to calculate the *z*-scores of `X`, specified as a positive integer vector. Each element of `vecdim` represents a dimension of the input array `X`. The output `Z` has the same dimensions as `X`, but the mean `mu` and standard deviation `sigma` each have length 1 in the operating dimensions. The other dimension lengths are the same for `X`, `mu`, and `sigma`.

For example, if `X` is a 2-by-3-by-3 array, then `zscore(X,0,[1 2])` uses the means and standard deviations along the pages of `X` to standardize the values of `X`.

Data Types: `single` | `double`

Output Arguments

`Z` — *z*-scores
vector | matrix | multidimensional array

z-scores, returned as a vector, matrix, or multidimensional array. `Z` has the same dimensions as `X`.

The values of `Z` depend on whether you specify `'all'`, `dim`, or `vecdim`. If you do not specify any of these input arguments, then the following conditions apply:

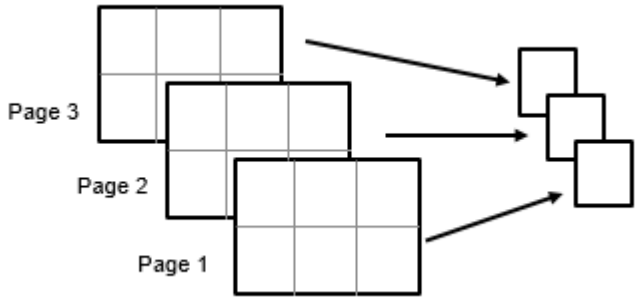
- If `X` is a vector, then `Z` is a vector of *z*-scores with mean 0 and variance 1.
- If `X` is an array, then `zscore` standardizes along the first nonsingleton dimension of `X`.

For an example that demonstrates the differences in `Z` when you use `'all'`, `dim`, and `vecdim`, see “Z-Scores of Multidimensional Array” on page 33-6639.

`mu` — Mean
scalar | vector | matrix | multidimensional array

Mean of `X` used to compute the *z*-scores, returned as a scalar, vector, matrix, or multidimensional array. `mu` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `mu`.

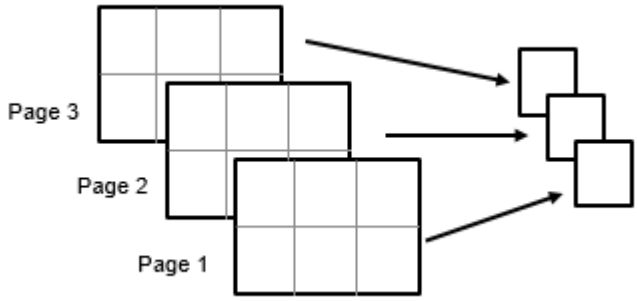
For example, if `X` is a 2-by-3-by-3 array and `vecdim` is `[1 2]`, then `mu` is a 1-by-1-by-3 array of means. Each value in `mu` corresponds to the mean of a page in `X`.



`sigma` — Standard deviation
scalar | vector | matrix | multidimensional array

Standard deviation of `X` used to compute the *z*-scores, returned as a scalar, vector, matrix, or multidimensional array. `sigma` has length 1 in the specified operating dimensions. The other dimension lengths are the same for `X` and `sigma`.

For example, if `X` is a 2-by-3-by-3 array and `vecdim` is `[1 2]`, then `sigma` is a 1-by-1-by-3 array of standard deviations. Each value in `sigma` corresponds to the standard deviation of a page in `X`.



More About

Z-Score

For a random variable X with mean μ and standard deviation σ, the z-score of a value x is

$$z = \frac{(x - \mu)}{\sigma}.$$

For sample data with mean $\bar{X}$ and standard deviation S, the z-score of a data point x is

$$z = \frac{\left(x - \bar{X}\right)}{S}.$$

z-scores measure the distance of a data point from the mean in terms of the standard deviation. This is also called *standardization* of data. The standardized data set has mean 0 and standard deviation 1, and retains the shape properties of the original data set (same skewness and kurtosis).

You can use z-scores to put data on the same scale before further analysis. This lets you to compare two or more data sets with different units.

Multidimensional Array

A *multidimensional array* is an array with more than two dimensions. For example, if X is a 1-by-3-by-4 array, then X is a three-dimensional array.

First Nonsingleton Dimension

A *first nonsingleton dimension* is the first dimension of an array whose size is not equal to 1. For example, if X is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of X.

Sample Standard Deviation

The *sample standard deviation S* is given by

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}{n-1}}.$$

S is the square root of an unbiased estimator of the variance of the population from which X is drawn, as long as X consists of independent, identically distributed samples. $\bar{X}$ is the sample mean.

Notice that the denominator in this variance formula is n – 1.

Population Standard Deviation

If the data is the entire population of values, then you can use the *population standard deviation*,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \mu\right)^2}{n}}.$$

If X is a random sample from a population, then the mean μ is estimated by the sample mean, and σ is the biased maximum likelihood estimator of the population standard deviation.

Notice that the denominator in this variance formula is n.

Algorithms

`zscore` returns `NaN`s for any sample containing `NaN`s.

`zscore` returns `0`s for any sample that is constant (all values are the same). For example, if X is a vector of the same numeric value, then Z is a vector of `0`s.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

This function fully supports tall arrays. For more information, see “Tall Arrays”.

C/C++ Code Generation

Generate C and C++ code using MATLAB® Coder™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.
- The `dim` input argument must be a compile-time constant.
- If you do not specify the `dim` input argument, the working (or operating) dimension can be different in the generated code. As a result, run-time errors can occur. For more details, see "Automatic dimension restriction" (MATLAB Coder).

For more information on code generation, see "Introduction to Code Generation" on page 32-2 and "General Code Generation Workflow" on page 32-5.

GPU Arrays
Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

Usage notes and limitations:

- The `'all'` and `vecdim` input arguments are not supported.

For more information, see "Run MATLAB Functions on a GPU" (Parallel Computing Toolbox).

See Also
`mean` | `normalize` | `rescale` | `std`

Introduced before R2006a

ztest

z-test

Syntax

```
h = ztest(x,m,sigma)
h= ztest(x,m,sigma,Name,Value)
[h,p] = ztest(___)
[h,p,ci,zval] = ztest(___)
```

Description

`h = ztest(x,m,sigma)` returns a test decision for the null hypothesis that the data in the vector `x` comes from a normal distribution with mean `m` and a standard deviation `sigma`, using the *z*-test on page 33-6649. The alternative hypothesis is that the mean is not `m`. The result `h` is `1` if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h= ztest(x,m,sigma,Name,Value)` returns a test decision for the *z*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = ztest(___)` also returns the *p*-value of the test, using any of the input arguments from previous syntaxes.

`[h,p,ci,zval] = ztest(___)` also returns the confidence interval of the population mean, `ci`, and the value of the test statistic, `zval`.

Examples

z-Test for a Hypothesized Mean

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the data comes from a normal distribution with mean `m = 75` and standard deviation `sigma = 10`.

```
[h,p,ci,zval] = ztest(x,75,10)
```

```
h = 0
```

```
p = 0.9927
```

```
ci = 2×1
```

```
   73.2191
   76.7975
```

```
zval = 0.0091
```

The returned value of `h = 0` indicates that `ztest` does not reject the null hypothesis at the default 5% significance level.

One-Sided z-Test

Load the sample data. Create a vector containing the first column of the students' exam grades data.

```
load examgrades
x = grades(:,1);
```

Test the null hypothesis that the data comes from a normal distribution with mean `m = 65` and standard deviation `sigma = 10`, against the alternative that the mean is greater than 65.

```
[h,p] = ztest(x,65,10,'Tail','right')
```

```
h = 1
```

```
p = 2.8596e-28
```

The returned value of `h = 1` indicates that `ztest` rejects the null hypothesis at the default 5% significance level, in favor of the alternative hypothesis that the population mean is greater than 65.

Input Arguments

`x` — Sample data

vector | matrix | multidimensional array

Sample data, specified as a vector, matrix, or multidimensional array.

- If `x` is specified as a vector, `ztest` returns a single value for each output argument.
- If `x` is specified as a matrix, `ztest` performs a separate *z*-test along each column of `x` and returns a vector of results.
- If `x` is specified as a multidimensional array on page 33-6649, `ztest` works along the first nonsingleton dimension on page 33-6649 of `x`.

In all cases, `ztest` treats `NaN` values as missing data and ignores them.

Data Types: `single` | `double`

`m` — Hypothesized mean

scalar value

Hypothesized mean, specified as a scalar value.

Data Types: `single` | `double`

`sigma` — Population standard deviation

scalar value

Population standard deviation, specified as a scalar value.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Tail','right','Alpha',0.01` specifies a right-tailed hypothesis test at the 1% significance level.

`Alpha` — Significance level

`0.05` (default) | scalar value in the range (0,1)

Significance level of the hypothesis test, specified as the comma-separated pair consisting of `'Alpha'` and a scalar value in the range (0,1).

Example: `'Alpha',0.01`

Data Types: `single` | `double`

`Dim` — Dimension

first nonsingleton dimension (default) | positive integer value

Dimension of the input matrix along which to test the means, specified as the comma-separated pair consisting of `'Dim'` and a positive integer value. For example, specifying `'Dim',1` tests the column means, while `'Dim',2` tests the row means.

Example: `'Dim',2`

Data Types: `single` | `double`

`Tail` — Type of alternative hypothesis

`'both'` (default) | `'right'` | `'left'`

Type of alternative hypothesis to evaluate, specified as the comma-separated pair consisting of `'Tail'` and one of:

- `'both'` — Test against the alternative hypothesis that the population mean is not `m`.
- `'right'` — Test against the alternative hypothesis that the population mean is greater than `m`.
- `'left'` — Test against the alternative hypothesis that the population mean is less than `m`.

`ztest` tests the null hypothesis that the population mean is `m` against the specified alternative hypothesis.

Example: `'Tail','right'`

Output Arguments

`h` — Hypothesis test result

`1` | `0`

Hypothesis test result, returned as `1` or `0`.

- If `h = 1`, this indicates the rejection of the null hypothesis at the `Alpha` significance level.

- If `h = 0`, this indicates a failure to reject the null hypothesis at the `Alpha` significance level.

`p` — *p*-value
scalar value in the range [0,1]

p-value of the test, returned as a scalar value in the range [0,1]. `p` is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of `p` cast doubt on the validity of the null hypothesis.

`ci` — Confidence interval
vector

Confidence interval for the true population mean, returned as a two-element vector containing the lower and upper boundaries of the 100 × (1 – `Alpha`)% confidence interval.

`zval` — Test statistic
nonnegative scalar value

Test statistic, returned as a nonnegative scalar value.

More About

z-Test

The *z*-test is a parametric hypothesis test used to determine whether a sample data set comes from a population with a particular mean. The test assumes that the sample data comes from a population with a normal distribution and a known standard deviation.

The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}},$$

where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, and n is the sample size. Under the null hypothesis, the test statistic has a standard normal distribution.

Multidimensional Array

A multidimensional array has more than two dimensions. For example, if `x` is a 1-by-3-by-4 array, then `x` is a three-dimensional array.

First Nonsingleton Dimension

The first nonsingleton dimension is the first dimension of an array whose size is not equal to 1. For example, if `x` is a 1-by-2-by-3-by-4 array, then the second dimension is the first nonsingleton dimension of `x`.

Tips

- Use `sampsizepwr` to calculate:
  - The sample size that corresponds to specified power and parameter values;
  - The power achieved for a particular sample size, given the true parameter value;

- The parameter value detectable with the specified sample size and power.

Extended Capabilities

GPU Arrays

Accelerate code by running on a graphics processing unit (GPU) using Parallel Computing Toolbox™.

This function fully supports GPU arrays. For more information, see “Run MATLAB Functions on a GPU” (Parallel Computing Toolbox).

See Also

`sampsizepwr` | `ttest` | `ttest2`

Introduced before R2006a

hmcSampler

Hamiltonian Monte Carlo (HMC) sampler

Syntax

```
hmc = hmcSampler(logpdf,startpoint)
hmc = hmcSampler( ___ ,Name,Value)
```

Description

`hmc = hmcSampler(logpdf,startpoint)` creates a Hamiltonian Monte Carlo (HMC) sampler, returned as a `HamiltonianSampler` object. `logpdf` is a function handle that evaluates the logarithm of the probability density of the equilibrium distribution and its gradient. The column vector `startpoint` is the initial point from which to start HMC sampling.

After you create the sampler, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics using the methods of the `HamiltonianSampler` class. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

`hmc = hmcSampler( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments. Specify name-value pair arguments after all other input arguments.

Examples

Create Hamiltonian Monte Carlo Sampler

Create a Hamiltonian Monte Carlo (HMC) sampler to sample from a normal distribution.

First, save a function `normalDistGrad` on the MATLAB® path that returns the multivariate normal log probability density and its gradient (`normalDistGrad` is defined at the end of this example). Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
means = [1;-3];
standevs = [1;2];
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point for the HMC sampler.

```
startpoint = randn(2,1);
```

Create the HMC sampler and display its properties.

```
smp = hmcSampler(logpdf,startpoint);
```

```
smp
```

```
smp =
  HamiltonianSampler with properties:
```

```
                StepSize: 0.1000
                NumSteps: 50
              MassVector: [2x1 double]
            JitterMethod: 'jitter-both'
  StepSizeTuningMethod: 'dual-averaging'
MassVectorTuningMethod: 'iterative-sampling'
                 LogPDF: @(theta)normalDistGrad(theta,means,standevs)
          VariableNames: {2x1 cell}
             StartPoint: [2x1 double]
```

The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

Input Arguments

`logpdf` — Logarithm of target density and its gradient

function handle

Logarithm of target density and its gradient, specified as a function handle.

`logpdf` must return two output arguments: `[lpdf,glpdf] = logpdf(X)`. Here, `lpdf` is the base-e log probability density (up to an additive constant), `glpdf` is the gradient of the log density, and the point `X` is a column vector with the same number of elements as `startpoint`.

The input argument `X` to `logpdf` must be unconstrained, meaning that every element of `X` can be any real number. Transform any constrained sampling parameters into unconstrained variables before using the HMC sampler.

If the `'UseNumericalGradient'` value is set to `true`, then `logpdf` does not need to return the gradient as the second output. Using a numerical gradient can be easier since `logpdf` does not need to compute the gradient, but it can make sampling slower.

Data Types: `function_handle`

`startpoint` — Initial point to start sampling from

numeric column vector

Initial point to start sampling from, specified as a numeric column vector.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'VariableNames',{'Intercept','Beta'},'MassVectorTuningMethod','hessian'` specifies sampling variable names and the mass vector tuning method to be `'hessian'`.

`StepSize` — Step size of Hamiltonian dynamics
`0.1` (default) | positive scalar

Step size of Hamiltonian dynamics, specified as the comma-separated pair consisting of `'StepSize'` and a positive scalar.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. This argument controls the step size of that leapfrog integration.

You can automatically tune the step size using `tuneSampler`.

Example: `'StepSize',0.2`

Data Types: `single` | `double`

`NumSteps` — Number of steps of Hamiltonian dynamics
`50` (default) | positive integer

Number of steps of Hamiltonian dynamics, specified as the comma-separated pair consisting of `'NumSteps'` and a positive integer.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. This argument controls the number of steps of that leapfrog integration.

You can automatically tune the number of steps using `tuneSampler`.

Example: `'NumSteps',20`

Data Types: `single` | `double`

`MassVector` — Mass vector of momentum variables
`ones(size(startpoint,1),1)` (default) | numeric column vector

Mass vector of momentum variables, specified as the comma-separated pair consisting of `'MassVector'` and a numeric column vector with positive values and the same length as `startpoint`.

The “masses” of the momentum variables associated with the variables of interest control the Hamiltonian dynamics in each Markov chain proposal.

You can automatically tune the mass vector using `tuneSampler`.

Example: `'MassVector',rand(3,1)`

Data Types: `single` | `double`

`JitterMethod` — Method for jittering step size and number of steps
`'jitter-both'` (default) | `'jitter-numsteps'` | `'none'`

Method for jittering the step size and the number of steps, specified as the comma-separated pair consisting of `'JitterMethod'` and one of the following values.

Value	Description
'jitter-both'	Randomly jitter the step size and number of steps for each leapfrog trajectory.
'jitter-numsteps'	Jitter only the number of steps of each leapfrog trajectory.
'none'	Perform no jittering.

With jittering, the sampler randomly selects the step size or the number of steps of each leapfrog trajectory as values smaller than the 'StepSize' and 'NumSteps' values. Use jittering to improve the stability of the leapfrog integration of the Hamiltonian dynamics.

Example: 'JitterMethod','jitter-both'

StepSizeTuningMethod — Method for tuning sampler step size
'dual-averaging' (default) | 'none'

Method for tuning the sampler step size, specified as the comma-separated pair consisting of 'StepSizeTuningMethod' and 'dual-averaging' or 'none'.

If the 'StepSizeTuningMethod' value is set to 'dual-averaging', then tuneSampler tunes the leapfrog step size of the HMC sampler to achieve a certain acceptance ratio for a fixed value of the simulation length. The simulation length equals the step size multiplied by the number of steps. To set the target acceptance ratio, use the 'TargetAcceptanceRatio' name-value pair argument of the tuneSampler method.

Example: 'StepSizeTuningMethod','none'

MassVectorTuningMethod — Method for tuning sampler mass vector
'iterative-sampling' (default) | 'hessian' | 'none'

Method for tuning the sampler mass vector, specified as the comma-separated pair consisting of 'MassVectorTuningMethod' and one of the following values.

Value	Description
'iterative-sampling'	Tune the MassVector via successive approximations by drawing samples using a sequence of mass vector estimates.
'hessian'	Set the MassVector equal to the negative diagonal Hessian of the logpdf at the startpoint.
'none'	Perform no tuning of the MassVector.

To perform the tuning, use the tuneSampler method.

Example: 'MassVectorTuningMethod','hessian'

CheckGradient — Flag for checking analytical gradient
true (or 1) (default) | false (or 0)

Flag for checking the analytical gradient, specified as the comma-separated pair consisting of 'CheckGradient' and either true (or 1) or false (or 0).

If `'CheckGradient'` is `true`, then the sampler calculates the numerical gradient at the `startpoint` and compares it to the analytical gradient returned by `logpdf`.

Example: `'CheckGradient',true`

`VariableNames` — Sampling variable names
`{'x1','x2',...}` (default) | string array | cell array of character vectors

Sampling variable names, specified as the comma-separated pair consisting of `'VariableNames'` and a string array or a cell array of character vectors. Elements of the array must be unique. The length of the array must be the same as the length of `startpoint`.

Supply a `'VariableNames'` value to label the components of the vector you want to sample using the HMC sampler.

Example: `'VariableNames',{'Intercept','Beta'}`

Data Types: `string` | `cell`

`UseNumericalGradient` — Flag for using numerical gradient
`false` (or `0`) (default) | `true` (or `1`)

Flag for using numerical gradient, specified as the comma-separated pair consisting of `'UseNumericalGradient'` and either `true` (or `1`) or `false` (or `0`).

If you set the `'UseNumericalGradient'` value to `true`, then the HMC sampler numerically estimates the gradient from the log density returned by `logpdf`. In this case, the `logpdf` function does not need to return the gradient of the log density as the second output. Using a numerical gradient makes HMC sampling slower.

Example: `'UseNumericalGradient',true`

Output Arguments

`hmc` — Hamiltonian Monte Carlo sampler
`HamiltonianSampler` object

Hamiltonian Monte Carlo sampler, returned as a `HamiltonianSampler` object.

See Also

Functions
`mhsample` | `slicesample`

Classes
`HamiltonianSampler`

Topics
“Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26
“Representing Sampling Distributions Using Markov Chain Samplers” on page 7-9

Introduced in R2017a

HamiltonianSampler class

Hamiltonian Monte Carlo (HMC) sampler

Description

A Hamiltonian Monte Carlo (HMC) sampler is a gradient-based Markov Chain Monte Carlo sampler that you can use to generate samples from a probability density *P(x)*. HMC sampling requires specification of *log P(x)* and its gradient.

The parameter vector *x* must be unconstrained, meaning that every element of *x* can be any real number. To sample constrained parameters, transform these parameters into unconstrained variables before using the HMC sampler.

After creating a sampler, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics using the methods of this class. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

Construction

`hmc = hmcSampler(logpdf,startpoint)` creates a Hamiltonian Monte Carlo (HMC) sampler, returned as a `HamiltonianSampler` object. `logpdf` is a function handle that evaluates the logarithm of the probability density of the equilibrium distribution and its gradient. The column vector `startpoint` is the initial point from which to start HMC sampling.

`hmc = hmcSampler( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments. Specify name-value pair arguments after all other input arguments.

Input Arguments

`logpdf` — Logarithm of target density and its gradient

function handle

Logarithm of target density and its gradient, specified as a function handle.

`logpdf` must return two output arguments: `[lpdf,glpdf] = logpdf(X)`. Here, `lpdf` is the base-e log probability density (up to an additive constant), `glpdf` is the gradient of the log density, and the point `X` is a column vector with the same number of elements as `startpoint`.

The input argument `X` to `logpdf` must be unconstrained, meaning that every element of `X` can be any real number. Transform any constrained sampling parameters into unconstrained variables before using the HMC sampler.

If the `'UseNumericalGradient'` value is set to `true`, then `logpdf` does not need to return the gradient as the second output. Using a numerical gradient can be easier since `logpdf` does not need to compute the gradient, but it can make sampling slower.

Data Types: `function_handle`

`startpoint` — Initial point to start sampling from

numeric column vector

Initial point to start sampling from, specified as a numeric column vector.

Data Types: `single` | `double`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'VariableNames', {'Intercept','Beta'},'MassVectorTuningMethod','hessian'` specifies sampling variable names and the mass vector tuning method to be `'hessian'`.

`StepSize` — Step size of Hamiltonian dynamics
`0.1` (default) | positive scalar

Step size of Hamiltonian dynamics, specified as the comma-separated pair consisting of `'StepSize'` and a positive scalar.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. This argument controls the step size of that leapfrog integration.

You can automatically tune the step size using `tuneSampler`.

Example: `'StepSize',0.2`

`NumSteps` — Number of steps of Hamiltonian dynamics
`50` (default) | positive integer

Number of steps of Hamiltonian dynamics, specified as the comma-separated pair consisting of `'NumSteps'` and a positive integer.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. This argument controls the number of steps of that leapfrog integration.

You can automatically tune the number of steps using `tuneSampler`.

Example: `'NumSteps',20`

`MassVector` — Mass vector of momentum variables
`ones(size(startpoint,1),1)` (default) | numeric column vector

Mass vector of momentum variables, specified as the comma-separated pair consisting of `'MassVector'` and a numeric column vector with positive values and the same length as `startpoint`.

The “masses” of the momentum variables associated with the variables of interest control the Hamiltonian dynamics in each Markov chain proposal.

You can automatically tune the mass vector using `tuneSampler`.

Example: `'MassVector',rand(3,1)`

`JitterMethod` — Method for jittering step size and number of steps
`'jitter-both'` (default) | `'jitter-numsteps'` | `'none'`

Method for jittering the step size and number of steps, specified as the comma-separated pair consisting of `'JitterMethod'` and one of the following:

Value	Description
`'jitter-both'`	Randomly jitter the step size and number of steps for each leapfrog trajectory.
`'jitter-numsteps'`	Jitter only the number of steps of each leapfrog trajectory.
`'none'`	Perform no jittering.

With jittering, the sampler randomly selects the step size or the number of steps of each leapfrog trajectory as values smaller than the `'StepSize'` and `'NumSteps'` values. Use jittering to improve the stability of the leapfrog integration of the Hamiltonian dynamics.

Example: `'JitterMethod','jitter-both'`

`StepSizeTuningMethod` — Method for tuning sampler step size
`'dual-averaging'` (default) | `'none'`

Method for tuning the sampler step size, specified as the comma-separated pair consisting of `'StepSizeTuningMethod'` and `'dual-averaging'` or `'none'`.

If the `'StepSizeTuningMethod'` value is set to `'dual-averaging'`, then `tuneSampler` tunes the leapfrog step size of the HMC sampler to achieve a certain acceptance ratio for a fixed value of the simulation length. The simulation length equals the step size multiplied by the number of steps. To set the target acceptance ratio, use the `'TargetAcceptanceRatio'` name-value pair argument of the `tuneSampler` method.

Example: `'StepSizeTuningMethod','none'`

`MassVectorTuningMethod` — Method for tuning sampler mass vector
`'iterative-sampling'` (default) | `'hessian'` | `'none'`

Method for tuning the sampler mass vector, specified as the comma-separated pair consisting of `'MassVectorTuningMethod'` and one of the following values

Value	Description
`'iterative-sampling'`	Tune the `MassVector` via successive approximations by drawing samples using a sequence of mass vector estimates.
`'hessian'`	Set the `MassVector` equal to the negative diagonal Hessian of the `logpdf` at the `startpoint`.
`'none'`	Perform no tuning of the `MassVector`.

To perform the tuning, use the `tuneSampler` method.

Example: `'MassVectorTuningMethod','hessian'`

`CheckGradient` — Flag for checking analytical gradient
`true` (or `1`) (default) | `false` (or `0`)

Flag for checking the analytical gradient, specified as the comma-separated pair consisting of `'CheckGradient'` and either `true` (or `1`) or `false` (or `0`).

If `'CheckGradient'` is `true`, then the sampler calculates the numerical gradient at the `startpoint` and compares it to the analytical gradient returned by `logpdf`.

Example: `'CheckGradient',true`

`VariableNames` — Sampling variable names
`{'x1','x2',...}` (default) | string array | cell array of character vectors

Sampling variable names, specified as the comma-separated pair consisting of `'VariableNames'` and a string array or cell array of character vectors. Elements of the array must be unique. The length of the array must be the same as the length of `startpoint`.

Supply a `'VariableNames'` value to label the components of the vector you want to sample using the HMC sampler.

Example: `'VariableNames',{'Intercept','Beta'}`

`UseNumericalGradient` — Flag for using numerical gradient
`false` (or `0`) (default) | `true` (or `1`)

Flag for using numerical gradient, specified as the comma-separated pair consisting of `'UseNumericalGradient'` and either `true` (or `1`) or `false` (or `0`).

If you set the `'UseNumericalGradient'` value to `true`, then the HMC sampler numerically estimates the gradient from the log density returned by `logpdf`. In this case, the `logpdf` function does not need to return the gradient of the log density as the second output. Using a numerical gradient makes HMC sampling slower.

Example: `'UseNumericalGradient',true`

Properties

`StepSize` — Step size of Hamiltonian dynamics
`0.1` (default) | positive scalar

Step size of Hamiltonian dynamics, specified as a positive scalar.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. The value of this property controls the step size of that leapfrog integration.

`NumSteps` — Number of steps of Hamiltonian dynamics
`50` (default) | positive integer

Number of steps of Hamiltonian dynamics, specified as a positive integer.

To propose a new state for the Markov chain, the HMC sampler integrates the Hamiltonian dynamics using leapfrog integration. The value of this property controls the number of steps of that leapfrog integration.

`MassVector` — Mass vector of momentum variables
`ones(size(startpoint,1),1)` (default) | numeric column vector

Mass vector of momentum variables, specified as a numeric column vector with positive values and the same length as `startpoint`.

The “masses” of the momentum variables associated with the variables of interest control the Hamiltonian dynamics in each Markov chain proposal.

`JitterMethod` — Method for jittering step size and number of steps
`'jitter-both'` (default) | `'jitter-numsteps'` | `'none'`

Method for jittering the step size and the number of steps, specified as one of the following values.

Value	Description
`'jitter-both'`	Randomly jitter the step size and number of steps of each leapfrog trajectory.
`'jitter-numsteps'`	Jitter only the number of steps of each leapfrog trajectory.
`'none'`	Perform no jittering.

With jittering, the sampler randomly selects the step size or the number of steps of each leapfrog trajectory as values smaller than the `'StepSize'` and `'NumSteps'` values. Use jittering to improve the stability of the leapfrog integration of the Hamiltonian dynamics.

`StepSizeTuningMethod` — Method for tuning sampler step size
`'dual-averaging'` (default) | `'none'`

Method for tuning the sampler step size, specified as `'dual-averaging'` or `'none'`.

If `StepSizeTuningMethod` equals `'dual-averaging'`, then `tuneSampler` tunes the leapfrog step size of the HMC sampler to achieve a certain acceptance ratio for a fixed value of the simulation length. The simulation length equals the step size multiplied by the number of steps. To set the target acceptance ratio, use the `'TargetAcceptanceRatio'` name-value pair argument of the `tuneSampler` method.

`MassVectorTuningMethod` — Method for tuning sampler mass vector
`'iterative-sampling'` (default) | `'hessian'` | `'none'`

Method for tuning the sampler mass vector, specified as one of the following values.

Value	Description
`'iterative-sampling'`	Tune the `MassVector` via successive approximations by drawing samples using a sequence of mass vector estimates.
`'hessian'`	Set the `MassVector` equal to the negative diagonal Hessian of the `logpdf` at the `startpoint`.
`'none'`	Perform no tuning of the `MassVector`.

To perform the tuning, use the `tuneSampler` method.

`LogPDF` — Logarithm of target density and its gradient
function handle

Logarithm of target density and its gradient, specified as a function handle.

`LogPDF` returns two output arguments: `[lpdf,glpdf] = LogPDF(X)`. Here, `lpdf` is the base-e log probability density (up to an additive constant) and `glpdf` is the gradient of the log density at the point `X`. The input argument `X` must be a column vector with the same number of elements as the `StartPoint` property.

If you set the `'UseNumericalGradient'` value to `true` when creating the sampler, then `LogPDF` returns the numerical gradient in `glpdf`.

`StartPoint` — Initial point to start sampling from
numeric column vector

Initial point to start sampling from, specified as a numeric column vector.

`VariableNames` — Sampling variable names
`{'x1','x2',...}` (default) | cell array of unique character vectors

Sampling variable names, specified as a cell array of unique character vectors.

Methods

estimateMAP	Estimate maximum of log probability density
tuneSampler	Tune Hamiltonian Monte Carlo (HMC) sampler
drawSamples	Generate Markov chain using Hamiltonian Monte Carlo (HMC)
diagnostics	Markov Chain Monte Carlo diagnostics

Examples

Create Hamiltonian Monte Carlo Sampler

Create a Hamiltonian Monte Carlo (HMC) sampler to sample from a normal distribution.

First, save a function `normalDistGrad` on the MATLAB® path that returns the multivariate normal log probability density and its gradient (`normalDistGrad` is defined at the end of this example). Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
means = [1;-3];
standevs = [1;2];
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point for the HMC sampler.

```
startpoint = randn(2,1);
```

Create the HMC sampler and display its properties.

```
smp = hmcSampler(logpdf,startpoint);

smp
```

```
smp =
  HamiltonianSampler with properties:

                 StepSize: 0.1000
                 NumSteps: 50
               MassVector: [2x1 double]
             JitterMethod: 'jitter-both'
      StepSizeTuningMethod: 'dual-averaging'
    MassVectorTuningMethod: 'iterative-sampling'
                   LogPDF: @(theta)normalDistGrad(theta,means,standevs)
            VariableNames: {2x1 cell}
               StartPoint: [2x1 double]
```

The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

See Also

Functions
`hmcSampler` | `mhsample` | `slicesample`

Topics
“Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26

Introduced in R2017a

estimateMAP

Class: `HamiltonianSampler`

Estimate maximum of log probability density

Syntax

```
xhat = estimateMAP(smp)
[xhat,fitinfo] = estimateMAP(smp)
[xhat,fitinfo] = estimateMAP(___,Name,Value)
```

Description

`xhat = estimateMAP(smp)` returns the maximum-a-posteriori (MAP) estimate of the log probability density of the Monte Carlo sampler `smp`.

`[xhat,fitinfo] = estimateMAP(smp)` returns additional fitting information in `fitinfo`.

`[xhat,fitinfo] = estimateMAP(___,Name,Value)` specifies additional options using one or more name-value pair arguments. Specify name-value pair arguments after all other input arguments.

Input Arguments

`smp` — Hamiltonian Monte Carlo sampler

`HamiltonianSampler` object

Hamiltonian Monte Carlo sampler, specified as a `HamiltonianSampler` object.

`estimateMAP` estimates the maximum of the log probability density specified in `smp.LogPDF`.

Use the `hmcSampler` function to create a sampler.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'IterationLimit',100,'StepTolerance',1e-5` estimates the MAP point using an iteration limit of 100 and a step size convergence tolerance of `1e-5`.

`StartPoint` — Initial point to start optimization from

`smp.StartPoint` (default) | numeric column vector

Initial point to start optimization from, specified as a numeric column vector with the same number of elements as the `StartPoint` property of the sampler `smp`.

Example: `'StartPoint',randn(size(smp.StartPoint))`

`IterationLimit` — Maximum number of optimization iterations

`1000` (default) | positive integer

Maximum number of optimization iterations, specified as a positive integer.

Example: `'IterationLimit',100`

`VerbosityLevel` — Verbosity level of Command Window output

`0` (default) | positive integer

Verbosity level of Command Window output during function maximization, specified as `0` or a positive integer.

- With the value set to `0`, `estimateMAP` displays no details on the optimization.
- With the value set to a positive integer, `estimateMAP` displays convergence information at each iteration.

Convergence Information

Heading	Meaning
`FUN VALUE`	Objective function value.
`NORM GRAD`	Norm of the gradient of the objective function.
`NORM STEP`	Norm of the iterative step, meaning the distance between the previous point and the current point.
`CURV`	`OK` means the weak Wolfe condition is satisfied. This condition is a combination of sufficient decrease of the objective function and a curvature condition.
`GAMMA`	Inner product of the step times the gradient difference, divided by the inner product of the gradient difference with itself. The gradient difference is the gradient at the current point minus the gradient at the previous point. Gives diagnostic information on the objective function curvature.
`ALPHA`	Step direction multiplier, which differs from `1` when the algorithm performed a line search.
`ACCEPT`	`YES` means the algorithm found an acceptable step to take.

Example: `'VerbosityLevel',1`

`GradientTolerance` — Relative gradient convergence tolerance

`1e-6` (default) | positive scalar

Relative gradient convergence tolerance, specified as a positive scalar.

Let `tau = max(1,min(abs(f),infnormg0))`, where `f` is the current objective function value and `infnormg0` is the initial gradient norm. If the norm of the objective function gradient is smaller than `tau` times the `'GradientTolerance'` value, then the maximization is considered to have converged to a local optimum.

Example: `'GradientTolerance',1e-4`

`StepTolerance` — Step size convergence tolerance

`1e-6` (default) | positive scalar

Step size convergence tolerance, specified as a positive scalar.

If the proposed step size is smaller than the `'StepTolerance'` value, then the maximization is considered to have converged to a local optimum.

Example: `'StepTolerance',1e-5`

Output Arguments

`xhat` — MAP point estimate

numeric vector

MAP point estimate, returned as a numeric vector of the same size as `smp.StartPoint`.

`fitinfo` — Fitting information

structure

Fitting information for the MAP computation, returned as a structure with these fields:

Field	Description
`Iteration`	Iteration indices from 0 through the final iteration.
`Objective`	Negative log probability density at each iteration. The MAP point is computed by minimizing the negative log probability density. You can check that the final values are all similar, indicating that the function optimization has converged to a local optimum.
`Gradient`	Gradient of the negative log probability density at the final iteration.

Data Types: `struct`

Examples

Estimate MAP Point of HMC Sampler

Create a Hamiltonian Monte Carlo sampler for a normal distribution and estimate the maximum-a-posteriori (MAP) point of the log probability density.

First, save a function `normalDistGrad` on the MATLAB® path that returns the multivariate normal log probability density and its gradient (`normalDistGrad` is defined at the end of this example). Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
means = [1;-1];
standevs = [1;0.3];
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point and create the HMC sampler.

```
startpoint = zeros(2,1);
smp = hmcSampler(logpdf,startpoint);
```

Estimate the MAP point (the point where the probability density has its maximum). Show more information during optimization by setting the `'VerbosityLevel'` value to 1.

```
[xhat,fitinfo] = estimateMAP(smp,'VerbosityLevel',1);
```

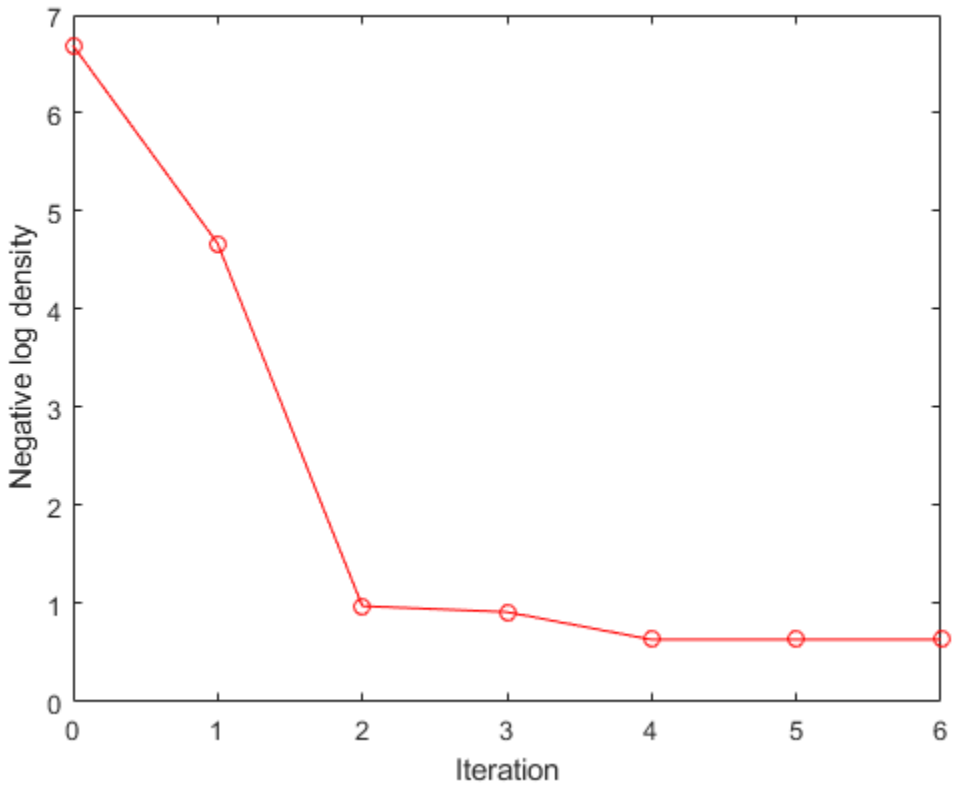
```
 o Solver = LBFGS, HessianHistorySize = 15, LineSearchMethod = weakwolfe

|=====================================================================================================
|   ITER   |   FUN VALUE   |  NORM GRAD  |  NORM STEP  |  CURV  |    GAMMA    |    ALPHA    | AC
|=====================================================================================================
|        0 |  6.689460e+00 |   1.111e+01 |   0.000e+00 |        |   9.000e-03 |   0.000e+00 |
|        1 |  4.671622e+00 |   8.889e+00 |   2.008e-01 |    OK  |   9.006e-02 |   2.000e+00 |
|        2 |  9.759850e-01 |   8.268e-01 |   8.215e-01 |    OK  |   9.027e-02 |   1.000e+00 |
|        3 |  9.158025e-01 |   7.496e-01 |   7.748e-02 |    OK  |   5.910e-01 |   1.000e+00 |
|        4 |  6.339508e-01 |   3.104e-02 |   7.472e-01 |    OK  |   9.796e-01 |   1.000e+00 |
|        5 |  6.339043e-01 |   3.668e-05 |   3.762e-03 |    OK  |   9.599e-02 |   1.000e+00 |
|        6 |  6.339043e-01 |   2.488e-08 |   3.333e-06 |    OK  |   9.015e-02 |   1.000e+00 |

         Infinity norm of the final gradient = 2.488e-08
              Two norm of the final step     = 3.333e-06, TolX   = 1.000e-06
Relative infinity norm of the final gradient = 2.488e-08, TolFun = 1.000e-06
EXIT: Local minimum found.
```

To further check that the optimization has converged to a local minimum, plot the `fitinfo.Objective` field. This field contains the values of the negative log density at each iteration of the function optimization. The final values are all very similar, so the optimization has converged.

```
fitinfo
```

```
fitinfo = struct with fields:
    Iteration: [7x1 double]
    Objective: [7x1 double]
     Gradient: [2x1 double]
```

```
plot(fitinfo.Iteration,fitinfo.Objective,'ro-');
xlabel('Iteration');
ylabel('Negative log density');
```



Display the MAP estimate. It is indeed equal to the `means` variable, which is the exact maximum.

```
xhat
```

```
xhat = 2×1

    1.0000
   -1.0000
```

```
means
```

```
means = 2×1

     1
    -1
```

The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

Tips

- First create a Hamiltonian Monte Carlo sampler using the `hmcSampler` function, and then use `estimateMAP` to estimate the MAP point.
- After creating an HMC sampler, you can tune the sampler, draw samples, and check convergence diagnostics using the other methods of the `HamiltonianSampler` class. Using the MAP estimate as a starting point in the `tuneSampler` and `drawSamles` methods can lead to more efficient tuning and sampling. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

Algorithms

- `estimateMAP` uses a limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) quasi-Newton optimizer to search for the maximum of the log probability density. See Nocedal and Wright [1].

References

[1] Nocedal, J. and S. J. Wright. *Numerical Optimization*, Second Edition. Springer Series in Operations Research, Springer Verlag, 2006.

See Also

Functions
`hmcSampler`

Classes
`HamiltonianSampler`

Topics
“Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26

Introduced in R2017a

tuneSampler

Class: `HamiltonianSampler`

Tune Hamiltonian Monte Carlo (HMC) sampler

Syntax

```
tunedSmp = tuneSampler(smp)
[tunedSmp,tuningInfo] = tuneSampler(smp)
[tunedSmp,tuningInfo] = tuneSampler( ___ ,Name,Value)
```

Description

`tunedSmp = tuneSampler(smp)` returns a tuned Hamiltonian Monte Carlo (HMC) sampler.

First, `tuneSampler` tunes the mass vector of the HMC sampler `smp`. Then, it tunes the step size and number of steps of the leapfrog integrations to achieve a certain target acceptance ratio.

You can use the tuned sampler to create Markov chains using the `drawSamples` method.

`[tunedSmp,tuningInfo] = tuneSampler(smp)` returns additional tuning information in `tuningInfo`.

`[tunedSmp,tuningInfo] = tuneSampler( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments. Specify name-value pair arguments after all other input arguments.

Input Arguments

`smp` — Hamiltonian Monte Carlo sampler

`HamiltonianSampler` object

Hamiltonian Monte Carlo sampler to tune, specified as a `HamiltonianSampler` object.

Use the `hmcSampler` function to create a sampler.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'StepSizeTuningMethod','dual-averaging','MassVectorTuningMethod','hessian'` tunes an HMC sampler using the specified methods for tuning the step size and mass vector of the sampler.

`StartPoint` — Initial point to start tuning from

`smp.StartPoint` (default) | numeric column vector

Initial point to start tuning from, specified as a numeric column vector with the same number of elements as the `StartPoint` property of the sampler `smp`.

Example: `'StartPoint',randn(size(smp.StartPoint))`

`StepSizeTuningMethod` — Method for tuning sampler step size
`smp.StepSizeTuningMethod` (default) | `'dual-averaging'` | `'none'`

Method for tuning sampler step size, specified as the comma-separated pair consisting of `'StepSizeTuningMethod'` and `'dual-averaging'` or `'none'`.

If `'StepSizeTuningMethod'` is set to `'dual-averaging'`, then `tuneSampler` tunes the leapfrog step size of the HMC sampler to achieve a target acceptance ratio for a fixed value of the simulation length. The simulation length equals the step size multiplied by the number of steps. To specify the target acceptance ratio, set the `'TargetAcceptanceRatio'` value.

To change the simulation length, set `smp.StepSize = a` and `smp.NumSteps = b`, for some values of `a` and `b`. This gives a simulation length of `a*b`.

Example: `'StepSizeTuningMethod','none'`

`MassVectorTuningMethod` — Method for tuning sampler mass vector
`smp.MassVectorTuningMethod` (default) | `'iterative-sampling'` | `'hessian'` | `'none'`

Method for tuning the sampler mass vector, specified as the comma-separated pair consisting of `'MassVectorTuningMethod'` and one of the following values.

Value	Description
`'iterative-sampling'`	Tune the `MassVector` via successive approximations by drawing samples using a sequence of mass vector estimates.
`'hessian'`	Set the `MassVector` equal to the negative diagonal Hessian of the `logpdf` at the `startpoint`.
`'none'`	Perform no tuning of the `MassVector`.

Example: `'MassVectorTuningMethod','hessian'`

`NumStepSizeTuningIterations` — Number of step size tuning iterations
`100` (default) | positive integer

Number of step size tuning iterations, specified as a positive integer.

If the `'StepSizeTuningMethod'` value is `'none'`, then `tuneSampler` does not tune the step size.

Example: `'NumStepSizeTuningIterations',50`

`TargetAcceptanceRatio` — Target acceptance ratio
`0.65` (default) | scalar from 0 through 1

Target acceptance ratio of the Markov chain, specified as a scalar from 0 through 1.

`tuneSampler` tunes the step size and number of steps of the leapfrog integration to achieve the specified target acceptance ratio for a fixed value of the simulation length. The simulation length is the leapfrog integration step size multiplied by the number of integration steps.

If the `'StepSizeTuningMethod'` value is `'none'`, then `tuneSampler` does not tune the step size.

To change the simulation length, set `smp.StepSize = a` and `smp.NumSteps = b`, for some values of `a` and `b`. This gives a simulation length of `a*b`.

Example: `'TargetAcceptanceRatio',0.55`

`NumStepsLimit` — Maximum number of leapfrog steps
`2000` (default) | positive integer

Maximum number of leapfrog steps allowed during step size tuning, specified as a positive integer.

If the `'StepSizeTuningMethod'` value is `'none'`, then `tuneSampler` does not tune the step size.

Example: `'NumStepsLimit',1000`

`VerbosityLevel` — Verbosity level of Command Window output
`0` (default) | nonnegative integer

Verbosity level of Command Window output during sampler tuning, specified as a nonnegative integer.

- With the value set to `0`, `tuneSampler` displays no details of the tuning.
- With the value set to `1`, `tuneSampler` displays details of the step size tuning.
- With the value set to `2` or larger, `tuneSampler` displays details of the step size and mass vector tuning.

Heading	Description
`ITER`	Iteration number.
`LOG PDF`	Log probability density at the current iteration.
`STEP SIZE`	Leapfrog integration step size at the current iteration.
`NUM STEPS`	Number of leapfrog integration steps at the current iteration.
`ACC RATIO`	Acceptance ratio, that is, the fraction of proposals which are accepted.
`DIVERGENT`	Number of times the sampler failed to generate a valid proposal due to the leapfrog iterations generating `NaN`s or `Infs`. `tuneSampler` searches for a good value of the integration step size. While doing so, the algorithm can encounter regions of instability and report nonzero values in the `DIVERGENT` column. This behavior is normal and is not a problem in itself.

Example: `'VerbosityLevel',1`

`NumPrint` — Verbose output frequency
`100` (default) | positive integer

Verbose output frequency, specified as a positive integer.

If the `'VerbosityLevel'` value is a positive integer, `tuneSampler` outputs tuning details every `'NumPrint'` iterations.

Example: `'NumPrint',50`

Output Arguments

`tunedSmp` — Tuned Hamiltonian Monte Carlo sampler

`HamiltonianSampler` object

Tuned Hamiltonian Monte Carlo sampler, returned as a `HamiltonianSampler` object.

`tuningInfo` — Tuning information

structure

Tuning information, returned as a structure with these fields.

Field	Description
`MassVector`	Tuned mass vector
`StepSize`	Tuned leapfrog step size
`NumSteps`	Tuned value of the number of leapfrog integration steps
`MassVectorTuningInfo`	Structure with additional information on the mass vector tuning
`StepSizeTuningInfo`	Structure with additional information on the step size tuning

If you tune the mass vector using the `'iterative-sampling'` method, then `MassVectorTuningInfo` has the following fields.

Field	Description
`MassVector`	Tuned mass vector
`IterativeSamplingMassVectorProfile`	*P*-by-*K* matrix of mass vectors used during the *K* iterations, where *P* is the number of sampling variables
`IterativeSamplingNumSamples`	*K*-by-1 vector of the number of samples drawn for each of the *K* iterations

If you tune the mass vector using the `'hessian'` method, then `MassVectorTuningInfo` has the following fields.

Field	Description
`MassVector`	Tuned mass vector
`NegativeDiagonalHessian`	Negative diagonal Hessian of `logpdf` at the tuning start point. If some elements are negative, this field can be different from the `MassVector` field.
`HessianPoint`	Point at which the Hessian is evaluated

If the `MassVectorTuningMethod` value is `'none'`, then `MassVectorTuningInfo` is empty.

If you tune the step size using the `'dual-averaging'` method, then `StepSizeTuningInfo` has the following fields.

Field	Description
`StepSize`	Tuned step size
`NumSteps`	Tuned value of the number of steps
`StepSizeProfile`	Column vector containing the step sizes at each tuning iteration
`AcceptanceRatio`	Final acceptance ratio achieved during tuning

If the step size is not tuned, then`StepSizeTuningInfo` is empty.

Data Types: `struct`

Examples

Tune Hamiltonian Monte Carlo Sampler

Tune the parameters of a Hamiltonian Monte Carlo (HMC) sampler.

Define the number of parameters to sample and their means.

```
NumParams = 9;
means = [1:NumParams]';
standevs = 1;
```

First, save a function `normalDistGrad` on the MATLAB® path that returns the multivariate normal log probability density and its gradient (`normalDistGrad` is defined at the end of this example). Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point and create the HMC sampler.

```
startpoint = randn(NumParams,1);
smp = hmcSampler(logpdf,startpoint);
```

It is important to select good values for the sampler parameters to get efficient sampling. The best way to find good values is to automatically tune the `MassVector`, `StepSize`, and `NumSteps` parameters using `tuneSampler`. The method:

1. Tunes the `MassVector` of the sampler.

2. Tunes `StepSize` and `NumSteps` for a fixed simulation length to achieve a certain acceptance ratio. The default target acceptance ratio of 0.65 is good in most cases.

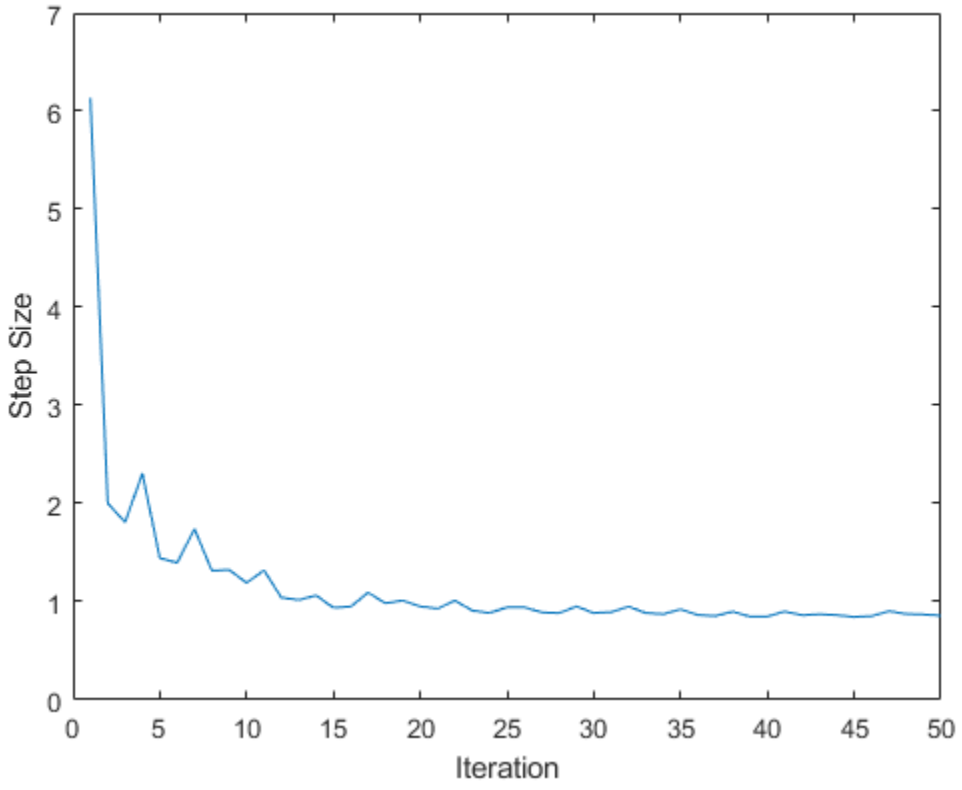
```
[smp,info] = tuneSampler(smp,'NumStepSizeTuningIterations',50,'VerbosityLevel',1,'NumPrint',10);
```

```
o Tuning mass vector using method: iterative-sampling
Finished mass vector tuning iteration 1 of 5.
Finished mass vector tuning iteration 2 of 5.
Finished mass vector tuning iteration 3 of 5.
```

```
Finished mass vector tuning iteration 4 of 5.
Finished mass vector tuning iteration 5 of 5.

o Tuning step size using method: dual-averaging. Target acceptance ratio = 0.65

o Initial step size for dual-averaging = 2

|=========================================================================================|
|   ITER   |    LOG PDF    |  STEP SIZE  |  NUM STEPS  |  ACC RATIO  |  DIVERGENT   |
|=========================================================================================|
|       10 | -1.710457e+01 |   1.193e+00 |           4 |   5.000e-01 |            0 |
|       20 | -9.152514e+00 |   9.527e-01 |           5 |   5.500e-01 |            0 |
|       30 | -1.068923e+01 |   8.856e-01 |           6 |   5.333e-01 |            0 |
|       40 | -1.290816e+01 |   8.506e-01 |           6 |   5.750e-01 |            0 |
|       50 | -1.770386e+01 |   8.581e-01 |           6 |   6.000e-01 |            0 |
```

Plot the evolution of the step size during tuning to ensure that the step size tuning has converged. Display the achieved acceptance ratio.

```
figure;
plot(info.StepSizeTuningInfo.StepSizeProfile);
xlabel('Iteration');
ylabel('Step Size');
```




```
accratio = info.StepSizeTuningInfo.AcceptanceRatio

accratio = 0.6000
```

The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

Tips

- After creating an HMC sampler using the `hmcSampler` function, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics using the methods of the `HamiltonianSampler` class. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

See Also

Functions
`hmcSampler`

Classes
`HamiltonianSampler`

Topics
“Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26

Introduced in R2017a

drawSamples

Class: `HamiltonianSampler`

Generate Markov chain using Hamiltonian Monte Carlo (HMC)

Syntax

```
chain = drawSamples(smp)
[chain,endpoint,accratio] = drawSamples(smp)
[chain,endpoint,accratio] = drawSamples( ___ ,Name,Value)
```

Description

`chain = drawSamples(smp)` generates a Markov chain by drawing samples using the Hamiltonian Monte Carlo sampler `smp`.

`[chain,endpoint,accratio] = drawSamples(smp)` also returns the final state of the Markov chain in `endpoint` and the fraction of accepted proposals in `accratio`.

`[chain,endpoint,accratio] = drawSamples( ___ ,Name,Value)` specifies additional options using one or more name-value pair arguments. Specify name-value pair arguments after all other input arguments.

Input Arguments

`smp` — Hamiltonian Monte Carlo sampler

`HamiltonianSampler` object

Hamiltonian Monte Carlo sampler, specified as a `HamiltonianSampler` object.

`drawSamples` draws samples from the target log probability density in `smp.LogPDF`. Use the `hmcSampler` function to create a sampler.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'Burnin',500,'NumSamples',2000` generates a Markov chain by discarding 500 burn-in samples and then drawing 2000 samples.

`Burnin` — Number of burn-in samples to discard

`1000` (default) | positive integer

Number of burn-in samples to discard from the beginning of the Markov chain, specified as a positive integer.

Example: `'Burnin',500`

`NumSamples` — Number of samples to draw
`1000` (default) | positive integer

Number of samples to draw from the Markov chain using the HMC sampler, specified as a positive integer.

The `drawSamples` method generates this number of samples after the burn-in period.

Example: `'NumSamples',2000`

`ThinSize` — Markov chain thinning size
`1` (default) | positive integer

Markov chain thinning size, specified as a positive integer.

Only one out of the `'ThinSize'` number of samples are kept. The rest of the samples are discarded.

Example: `'ThinSize',5`

`StartPoint` — Initial point to start sampling from
`smp.StartPoint` (default) | numeric column vector

Initial point to start sampling from, specified as a numeric column vector with the same number of elements as the `StartPoint` property of the sampler `smp`.

Example: `'StartPoint',randn(5,1)`

`VerbosityLevel` — Verbosity level of Command Window output
`0` (default) | positive integer

Verbosity level of Command Window output during sampling, specified as `0` or a positive integer.

With the value set to `0`, `drawSamples` displays no details during sampling.

With the value set to a positive integer, `drawSamples` displays details of the sampling. To set the output frequency, use the `'NumPrint'` name-value pair argument.

`drawSamples` displays the output as a table with these columns.

Heading	Description
`ITER`	Iteration number
`LOG PDF`	Log probability density at the current iteration
`STEP SIZE`	Leapfrog integration step size at the current iteration. If the step size is jittered, it can vary between iterations.
`NUM STEPS`	Number of leapfrog integration steps at the current iteration. If the number of steps is jittered, it can vary between iterations
`ACC RATIO`	Acceptance ratio, that is, the fraction of proposals that are accepted. The acceptance ratio is calculated from the beginning of sampling, including the burn-in period.

Heading	Description
DIVERGENT	Number of times the sampler failed to generate a valid proposal due to the leapfrog iterations generating NaNs or Infs. When drawing samples, a nonzero value in the DIVERGENT column indicates that the chosen step size is above the stability threshold for some region of state space. To fix this issue, try to set the StepSize to a smaller value, draw new samples, and check that all values in the DIVERGENT column equal 0.

Example: 'VerbosityLevel',1

NumPrint — Verbose output frequency
100 (default) | positive integer

Verbose output frequency, specified as a positive integer.

If the 'VerbosityLevel' value is a positive integer, then drawSamples outputs sampling details every 'NumPrint' iterations.

Example: 'NumPrint',200

Output Arguments

chain — Markov chain generated using Hamiltonian Monte Carlo
numeric matrix

Markov chain generated using Hamiltonian Monte Carlo, returned as a numeric matrix.

Each row of chain is a sample, and each column represents one sampling variable.

endpoint — Final state of Markov chain
numeric column vector

Final state of the Markov chain, returned as a numeric column vector of the same length as smp.StartPoint.

accratio — Acceptance ratio
numeric scalar

Acceptance ratio of the Markov chain proposals, returned as a numeric scalar. The acceptance ratio is calculated from the beginning of sampling, including the burn-in period.

Examples

Draw Samples Using HMC Sampler

Create MCMC chains for a multivariate normal distribution using a Hamiltonian Monte Carlo (HMC) sampler.

Define the number of parameters to sample and their means.

```
NumParams = 100;
means = randn(NumParams,1);
standevs = 0.1;
```

First, save a function `normalDistGrad` on the MATLAB® path that returns the multivariate normal log probability density and its gradient (`normalDistGrad` is defined at the end of this example). Then, call the function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point of the sampler. Create the HMC sampler and tune its parameters.

```
startpoint = randn(NumParams,1);
smp = hmcSampler(logpdf,startpoint);
smp = tuneSampler(smp);
```

Draw samples from the posterior density, using a few independent chains. Choose different, randomly distributed starting points for each chain. Specify the number of burn-in samples to discard from the beginning of the Markov chain and the number of samples to generate after the burn-in. Set the `'VerbosityLevel'` to print details during sampling for the first chain.

```
NumChains  = 4;
chains     = cell(NumChains,1);
Burnin     = 500;
NumSamples = 2000;
for c = 1:NumChains
    if c == 1
        showOutput = 1;
    else
        showOutput = 0;
    end
    chains{c} = drawSamples(smp,'Burnin',Burnin,'NumSamples',NumSamples,...
        'Start',randn(size(startpoint)),'VerbosityLevel',showOutput,'NumPrint',500);
end
```

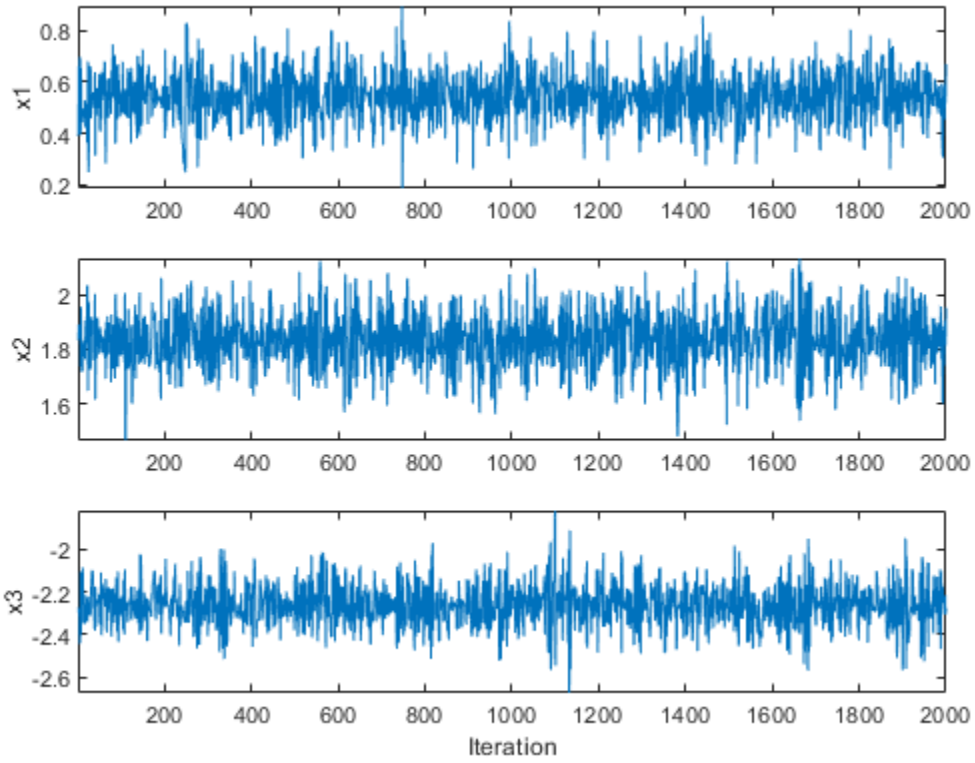
```
|=============================================================================|
|   ITER   |    LOG PDF    |  STEP SIZE  |  NUM STEPS  |  ACC RATIO  |  DIVERGENT  |
|=============================================================================|
|      500 |  8.450463e+01 |   4.776e-01 |           5 |   9.060e-01 |           0 |
|     1000 |  8.034444e+01 |   4.776e-01 |           9 |   8.810e-01 |           0 |
|     1500 |  9.156276e+01 |   4.776e-01 |           2 |   8.867e-01 |           0 |
|     2000 |  8.027782e+01 |   2.817e-02 |           6 |   8.890e-01 |           0 |
|     2500 |  9.892440e+01 |   4.648e-01 |           2 |   8.904e-01 |           0 |
```

After obtaining a random sample, investigate issues such as convergence and mixing to determine whether the samples represent a reasonable set of random realizations from the target distribution. To examine the output, plot the trace plots of the samples for the first few variables using the first chain.

A number of burn-in samples have been removed to reduce the effect of the sampling starting point. Furthermore, the trace plots look like high-frequency noise, without any visible long-range correlation between the samples. This indicates that the chain is well mixed.

```
for p = 1:3
    subplot(3,1,p);
    plot(chains{1}(:,p));
```

```
    ylabel(smp.VariableNames(p))
    axis tight
end
xlabel('Iteration')
```



The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as the `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

Tips

- After creating an HMC sampler using the `hmcSampler` function, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics using the methods of the `HamiltonianSampler` class. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

See Also

Functions
`hmcSampler`

Classes
`HamiltonianSampler`

Topics
Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26

Introduced in R2017a

diagnostics

Class: `HamiltonianSampler`

Markov Chain Monte Carlo diagnostics

Syntax

```
tbl = diagnostics(smp,chains)
tbl = diagnostics(smp,chains,'MaxLag',maxlag)
```

Description

`tbl = diagnostics(smp,chains)` returns Markov Chain Monte Carlo diagnostics for the chains in `chains`.

`tbl = diagnostics(smp,chains,'MaxLag',maxlag)` specifies the maximum number of autocorrelation lags to use for computing effective sample sizes.

Input Arguments

`smp` — Hamiltonian Monte Carlo sampler

`HamiltonianSampler` object

Hamiltonian Monte Carlo sampler, specified as a `HamiltonianSampler` object.

Use the `hmcSampler` function to create a sampler.

`chains` — MCMC chains

matrix | cell array

MCMC chains, specified as one of the following:

- A matrix, where each row is a sample and each column a parameter.
- A cell array of matrices, where the chain `chains{i}` is a matrix where each row is a sample and each column a parameter.

The number of parameters (that is, matrix columns) must equal the number of elements of the `StartPoint` property of the `smp` sampler.

`maxlag` — Maximum number of autocorrelation lags

`100` (default) | positive integer

Maximum number of autocorrelation lags for computing effective sample sizes, specified as a positive integer.

The effective sample size calculation uses lags of `1,2,...,maxlag` for each chain in `chains` that has more than `maxlag` samples.

For chains with `maxlag` or fewer samples, the calculation uses N_i - 1 lags, where N_i is the number of samples of chain i.

Example: `'MaxLag',50`

Output Arguments

`tbl` — MCMC diagnostics

table

MCMC diagnostics, computed using all the chains in `chains` and returned as a table with these columns.

Column	Description
`Name`	Variable name
`Mean`	Posterior mean estimate
`MCSE`	Estimate of the Monte Carlo standard error (the standard deviation of the posterior mean estimate)
`SD`	Estimate of the posterior standard deviation
`Q5`	Estimate of the 5th quantile of the marginal posterior distribution
`Q95`	Estimate of the 95th quantile of the marginal posterior distribution
`ESS`	Effective sample size for the posterior mean estimate. Larger effective sample sizes lead to more accurate results. If the samples are independent, then the effective sample size is equal to the number of samples.
`RHat`	Gelman-Rubin convergence statistic. As a rule of thumb, values of `RHat` less than 1.1 are interpreted as a sign that the chains have converged to the target distribution. If `RHat` for any variable is larger than 1.1, try drawing more Monte Carlo samples.

Examples

Compute Markov Chain Monte Carlo Diagnostics

Create MCMC chains using a Hamiltonian Monte Carlo (HMC) sampler and compute MCMC diagnostics.

First, save a function on the MATLAB® path that returns the multivariate normal log probability density and its gradient. In this example, that function is called `normalDistGrad` and is defined at the end of the example. Then, call this function with arguments to define the `logpdf` input argument to the `hmcSampler` function.

```
means = [1;-2;2];
standevs = [1;2;0.5];
logpdf = @(theta)normalDistGrad(theta,means,standevs);
```

Choose a starting point. Create the HMC sampler and tune its parameters.

```
startpoint = randn(3,1);
smp = hmcSampler(logpdf,startpoint);
smp = tuneSampler(smp);
```

Draw samples from the posterior density, using a few independent chains. Choose different, randomly distributed starting points for each chain. Specify the number of burn-in samples to discard from the beginning of the Markov chain and the number of samples to generate after the burn-in.

```
NumChains  = 4;
chains     = cell(NumChains,1);
Burnin     = 500;
NumSamples = 1000;
for c = 1:NumChains
    chains{c} = drawSamples(smp,'Burnin',Burnin,'NumSamples',NumSamples,...
        'Start',randn(size(startpoint)));
end
```

Compute MCMC diagnostics and display the results. Compare the true means in `means` with the column titled `Mean` in the `MCMCdiagnostics` table. The true posterior means are within a few Monte Carlo standard errors (MCSEs) of the estimated posterior means. The HMC sampler has accurately recovered the true means. Similarly, the estimated standard deviations in the column `SD` are very near the true standard deviations in `standev`.

```
MCMCdiagnostics = diagnostics(smp,chains)
```

```
MCMCdiagnostics=3×8 table
     Name      Mean        MCSE         SD          Q5         Q95        ESS      RHat
    ______    _______    ________    _______    ________    ______    ______    ____

    {'x1'}     1.0038    0.016474    0.96164    -0.58601     2.563    3407.4     1
    {'x2'}    -2.0435    0.034933      1.999     -5.3476    1.1851    3274.5     1
    {'x3'}     1.9957    0.008209    0.49693      1.2036    2.8249    3664.5     1
```

```
means
```

```
means = 3×1

     1
    -2
     2
```

```
standevs
```

```
standevs = 3×1

    1.0000
    2.0000
    0.5000
```

The `normalDistGrad` function returns the logarithm of the multivariate normal probability density with means in `Mu` and standard deviations in `Sigma`, specified as scalars or columns vectors the same length as the `startpoint`. The second output argument is the corresponding gradient.

```
function [lpdf,glpdf] = normalDistGrad(X,Mu,Sigma)
Z = (X - Mu)./Sigma;
lpdf = sum(-log(Sigma) - .5*log(2*pi) - .5*(Z.^2));
glpdf = -Z./Sigma;
end
```

Tips

- After creating an HMC sampler using the `hmcSampler` function, you can compute MAP (maximum-a-posteriori) point estimates, tune the sampler, draw samples, and check convergence diagnostics using the methods of the `HamiltonianSampler` class. For an example of this workflow, see Bayesian Linear Regression Using Hamiltonian Monte Carlo on page 7-26.

See Also

Functions
`hmcSampler`

Classes
`HamiltonianSampler`

Topics
“Bayesian Linear Regression Using Hamiltonian Monte Carlo” on page 7-26

Introduced in R2017a

Classification Learner

Train models to classify data using supervised machine learning

Description

The **Classification Learner** app trains models to classify data. Using this app, you can explore supervised machine learning using various classifiers. You can explore your data, select features, specify validation schemes, train models, and assess results. You can perform automated training to search for the best classification model type, including decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, ensemble, and neural network classification.

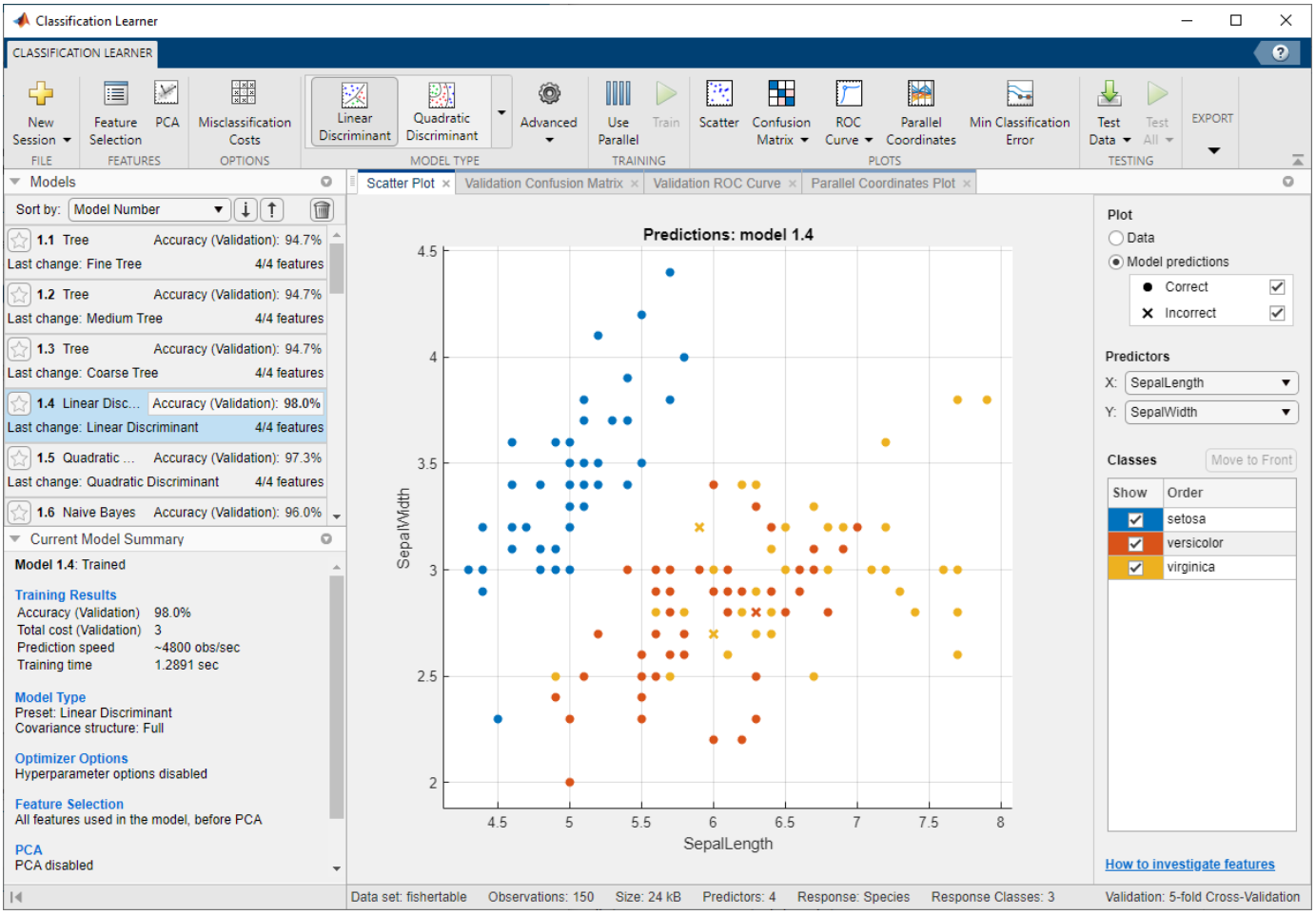
You can perform supervised machine learning by supplying a known set of input data (observations or examples) and known responses to the data (e.g., labels or classes). You use the data to train a model that generates predictions for the response to new data. To use the model with new data, or to learn about programmatic classification, you can export the model to the workspace or generate MATLAB code to recreate the trained model.

Tip To get started, in the Classifier list, try **All Quick-To-Train** to train a selection of models. See “Automated Classifier Training” on page 23-10.

Required Products

- MATLAB
- Statistics and Machine Learning Toolbox

Note: When you use Classification Learner in MATLAB Online, you can train models in parallel using a Cloud Center cluster (requires Parallel Computing Toolbox). For more information, see “Use Parallel Computing Toolbox with Cloud Center Cluster in MATLAB Online” (Parallel Computing Toolbox).



Open the Classification Learner App

- MATLAB Toolstrip: On the **Apps** tab, under **Machine Learning**, click the app icon.
- MATLAB command prompt: Enter `classificationLearner`.

Examples

- “Train Classification Models in Classification Learner App” on page 23-10
- “Select Data and Validation for Classification Problem” on page 23-18
- “Automated Classifier Training” on page 23-10
- “Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
- “Choose Classifier Options” on page 23-22
- “Assess Classifier Performance in Classification Learner” on page 23-65
- “Export Classification Model to Predict New Data” on page 23-77

Programmatic Use

`classificationLearner` opens the Classification Learner app or brings focus to the app if it is already open.

`classificationLearner(Tbl,ResponseVarName)` opens the Classification Learner app and populates the New Session from Arguments dialog box with the data contained in the table `Tbl`. The `ResponseVarName` argument, specified as a character vector or string scalar, is the name of the response variable in `Tbl` that contains the class labels. The remaining variables in `Tbl` are the predictor variables.

`classificationLearner(Tbl,Y)` opens the Classification Learner app and populates the New Session from Arguments dialog box with the predictor variables in the table `Tbl` and the class labels in the vector `Y`. You can specify the response `Y` as a categorical array, character array, string array, logical vector, numeric vector, or cell array of character vectors.

`classificationLearner(X,Y)` opens the Classification Learner app and populates the New Session from Arguments dialog box with the *n*-by-*p* predictor matrix `X` and the *n* class labels in the vector `Y`. Each row of `X` corresponds to one observation, and each column corresponds to one variable. The length of `Y` and the number of rows of `X` must be equal.

`classificationLearner(___,Name,Value)` specifies cross-validation options using one or more of the following name-value arguments in addition to any of the input argument combinations in the previous syntaxes. For example, you can specify `'KFold',10` to use a 10-fold cross-validation scheme.

- `'CrossVal'`, specified as `'on'` (default) or `'off'`, is the cross-validation flag. If you specify `'on'`, then the app uses 5-fold cross-validation. If you specify `'off'`, then the app uses resubstitution validation.

  You can override the `'CrossVal'` cross-validation setting by using the `'Holdout'` or `'KFold'` name-value argument. You can specify only one of these arguments at a time.
- `'Holdout'`, specified as a numeric scalar in the range [0.05,0.5], is the fraction of the data used for holdout validation. The app uses the remaining data for training.
- `'KFold'`, specified as a positive integer in the range [2,50], is the number of folds to use for cross-validation.

See Also

Apps
Regression Learner

Functions
`fitcdiscr` | `fitcecoc` | `fitcensemble` | `fitcknn` | `fitcnet` | `fitcsvm` | `fitctree` | `fitglm`

Topics
“Train Classification Models in Classification Learner App” on page 23-10
“Select Data and Validation for Classification Problem” on page 23-18
“Automated Classifier Training” on page 23-10
“Feature Selection and Feature Transformation Using Classification Learner App” on page 23-42
“Choose Classifier Options” on page 23-22
“Assess Classifier Performance in Classification Learner” on page 23-65

“Export Classification Model to Predict New Data” on page 23-77

Introduced in R2015a

Regression Learner

Train regression models to predict data using supervised machine learning

Description

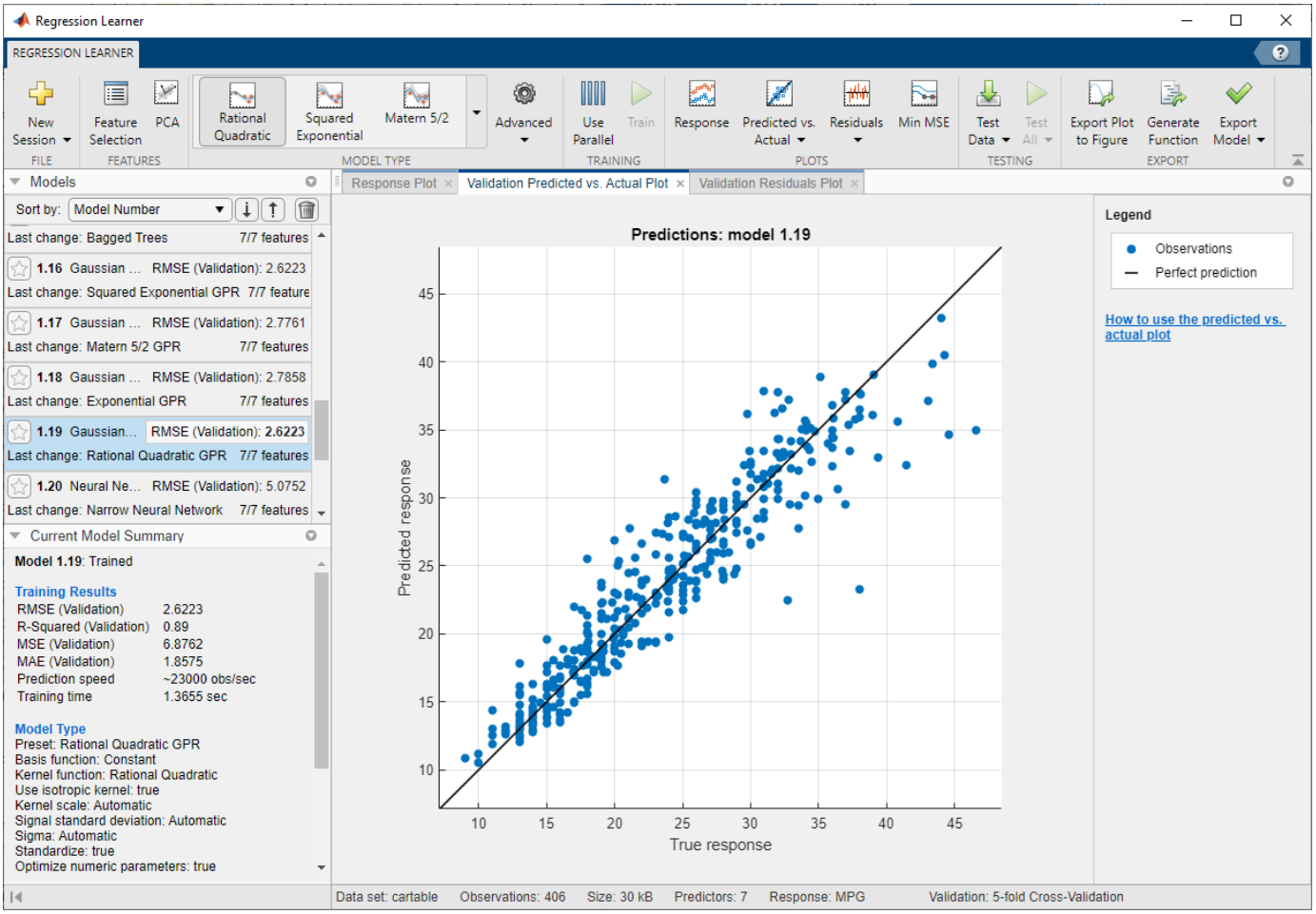
The **Regression Learner** app trains regression models to predict data. Using this app, you can explore your data, select features, specify validation schemes, train models, and assess results. You can perform automated training to search for the best regression model type, including linear regression models, regression trees, Gaussian process regression models, support vector machines, ensembles of regression trees, and neural network regression models.

Perform supervised machine learning by supplying a known set of observations of input data (predictors) and known responses. Use the observations to train a model that generates predicted responses for new input data. To use the model with new data, or to learn about programmatic regression, you can export the model to the workspace or generate MATLAB code to recreate the trained model.

Required Products

- MATLAB
- Statistics and Machine Learning Toolbox

Note: When you use Regression Learner in MATLAB Online, you can train models in parallel using a Cloud Center cluster (requires Parallel Computing Toolbox). For more information, see “Use Parallel Computing Toolbox with Cloud Center Cluster in MATLAB Online” (Parallel Computing Toolbox).



Open the Regression Learner App

- MATLAB Toolstrip: On the **Apps** tab, under **Machine Learning**, click the app icon.
- MATLAB command prompt: Enter `regressionLearner`.

Examples

- Train Regression Models in Regression Learner App on page 24-2
- Select Data and Validation for Regression Problem on page 24-8
- Automated Regression Model Training on page 24-2
- Choose Regression Model Options on page 24-12
- Feature Selection and Feature Transformation on page 24-26
- Assess Model Performance in Regression Learner on page 24-42
- Export Regression Model to Predict New Data on page 24-54
- Train Regression Trees Using Regression Learner App on page 24-60

Programmatic Use

`regressionLearner` opens the Regression Learner app or brings focus to the app if it is already open.

`regressionLearner(Tbl,ResponseVarName)` opens the Regression Learner app and populates the New Session from Arguments dialog box with the data contained in the table `Tbl`. The `ResponseVarName` argument, specified as a character vector or string scalar, is the name of the variable in `Tbl` that contains the response values. The remaining variables in `Tbl` are the predictor variables.

`regressionLearner(Tbl,Y)` opens the Regression Learner app and populates the New Session from Arguments dialog box with the predictor variables in the table `Tbl` and the response values in the numeric vector `Y`.

`regressionLearner(X,Y)` opens the Regression Learner app and populates the New Session from Arguments dialog box with the *n*-by-*p* predictor matrix `X` and the *n* response values in the vector `Y`. Each row of `X` corresponds to one observation, and each column corresponds to one variable. The length of `Y` and the number of rows of `X` must be equal.

`regressionLearner( ___ ,Name,Value)` specifies cross-validation options using one or more of the following name-value arguments in addition to any of the input argument combinations in the previous syntaxes. For example, you can specify `'KFold',10` to use a 10-fold cross-validation scheme.

- `'CrossVal'`, specified as `'on'` (default) or `'off'`, is the cross-validation flag. If you specify `'on'`, then the app uses 5-fold cross-validation. If you specify `'off'`, then the app uses resubstitution validation.

  You can override the `'CrossVal'` cross-validation setting by using the `'Holdout'` or `'KFold'` name-value argument. You can specify only one of these arguments at a time.
- `'Holdout'`, specified as a numeric scalar in the range [0.05,0.5], is the fraction of the data used for holdout validation. The app uses the remaining data for training.
- `'KFold'`, specified as a positive integer in the range [2,50], is the number of folds to use for cross-validation.

See Also

Apps
Classification Learner

Functions
`fitlm` | `fitrensemble` | `fitrgp` | `fitrnet` | `fitrsvm` | `fitrtree` | `stepwiselm`

Topics
Train Regression Models in Regression Learner App on page 24-2
Select Data and Validation for Regression Problem on page 24-8
Automated Regression Model Training on page 24-2
Choose Regression Model Options on page 24-12
Feature Selection and Feature Transformation on page 24-26
Assess Model Performance in Regression Learner on page 24-42
Export Regression Model to Predict New Data on page 24-54

Train Regression Trees Using Regression Learner App on page 24-60

Introduced in R2017a

Distribution Fitter

Fit probability distributions to data

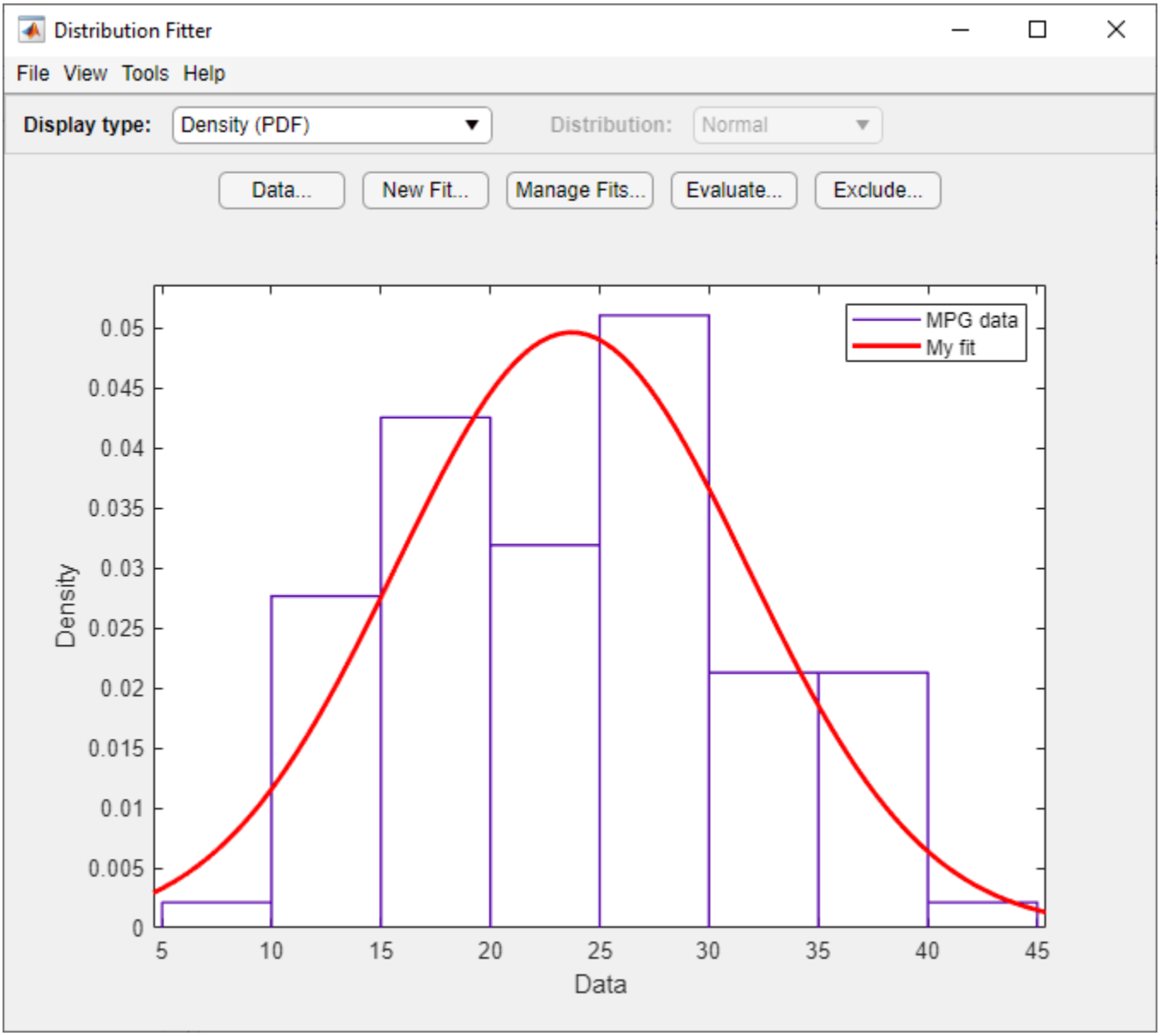
Description

The **Distribution Fitter** app interactively fits probability distributions to data imported from the MATLAB workspace. You can choose from 22 built-in probability distributions or create your own custom distribution. The app displays plots of the fitted distribution superimposed on a histogram of the data. Available plots include probability density function (pdf), cumulative distribution function (cdf), probability plots, and survivor functions. You can export the fitted parameter values to the workspace as a probability distribution object, and use object functions to perform further analyses. For more information on working with these objects, see “Working with Probability Distributions” on page 5-3. For the programmatic work flow of the Distribution Fitter app, see `distributionFitter`.

Required Products

- MATLAB
- Statistics and Machine Learning Toolbox

Note: Distribution Fitter does not provide printing, custom distribution defining, or code generating functionality in MATLAB Online.



Open the Distribution Fitter App

- MATLAB Toolstrip: On the **Apps** tab, under **Math, Statistics and Optimization**, click the app icon.
- MATLAB command prompt: Enter `distributionFitter`.

Examples

- “Fit a Distribution Using the Distribution Fitter App” on page 5-71

Parameters

Data

`Data` — Data to import from the workspace
list of variables | valid MATLAB expression

Specify the data to import by selecting a variable from the drop-down list. If the variable is a matrix, the app imports the first column of the matrix by default. To select a different column or row of the matrix, click **Select Column or Row**. Alternatively, you can enter any valid MATLAB expression in the field.

`Censoring` — Variable containing censoring data
list of variables

Specify the censoring data by selecting a variable from the drop-down list. If the variable is a matrix, the app imports the first column of the matrix by default. To select a different column or row of the matrix, click **Select Column or Row**. This parameter is optional.

`Frequency` — Variable containing frequency data
list of variables

Specify the frequency data by selecting a variable from the drop-down list. If the variable is a matrix, the app imports the first column of the matrix by default. To select a different column or row of the matrix, click **Select Column or Row**. This parameter is optional.

`Data set name` — Data set name
character vector

Specify a name for the data set as a character vector, or accept the default name.

`Manage data sets` — Manage previously imported data sets
list of data sets

Manage previously imported data sets. Click the data set of interest, then click the buttons below this pane to view the data (**View**), set the bin rules (**Set Bin Rules**), rename the data set (**Rename**), or delete the data set (**Delete**).

`Data preview` — Preview plot of data
histogram plot

Display a preview plot of the variable selected from the **Data** drop-down menu.

New fit

`Fit name` — Name of fit
character vector

Specify a name for the fit or accept the default name.

`Data` — Data set to fit
list of data sets

Specify the data to fit by selecting a data set from the drop-down list.

`Distribution` — Distribution to fit
`Normal` (default) | `Exponential` | `Weibull` | `Non-parametric` | `...`

Specify the distribution to fit by selecting a distribution name from the drop-down list.

`Exclusion rule` — Data exclusion rule
list of exclusion rules

Specify a rule to exclude some data values by selecting an exclusion rule from the drop-down list. To populate this drop-down list, you must first define exclusion rules by clicking **Exclude** in the main window of the app. This parameter is optional

Manage fits

`Plot` — Flag to plot fitted distribution
checked (default) | unchecked

Specify which fit or fits to plot in the main window by selecting the **Plot** check box next to each fit. Clear the **Plot** check box to remove a fit from the plot.

`Conf counds` — Flag to plot confidence bounds
unchecked (default) | checked

If you select **Plot** for a particular fit, you can select **Conf bounds** to display the confidence bounds for that fit on the plot in the main window. Clearing the **Conf bounds** check box removes the confidence intervals from the plot. The Distribution Fitter app displays confidence bounds only if the Display Type in the main window is set to **Cumulative probability (CDF)**, **Quantile (inverse CDF)**, **Survivor function**, or **Cumulative hazard**.

Evaluate

`Fit` — Fit to evaluate
list of fits

Select one or more fits from the list to evaluate.

`Function` — Available functions to fit
`Density (PDF)` (default) | `Cumulative probability (CDF)` | `Quantile (inverse CDF)` | `Survivor function` | `Cumulative hazard` | `Hazard rate`

Specify the type of probability function to evaluate from the drop-down list. Available probability functions include the probability density function (pdf), cumulative distribution function (cdf), quantile (inverse cdf), survival function, cumulative hazard, and hazard rate.

`At x =` — Values at which to evaluate function
numeric vector

Specify a numeric vector of values at which to evaluate the function. If you specify **Function** as **Quantile (inverse CDF)**, this field name changes to **At p =** and you enter a vector of probability values.

`Compute confidence bounds` — Flag to compute confidence bounds
unchecked (default) | checked

Select **Compute confidence bounds** to compute the confidence bounds for the selected fit. This check box is enabled only if you specify **Function** as **Cumulative probability (CDF)**, **Quantile (inverse CDF)**, **Survivor function**, or **Cumulative hazard**. This parameter is optional.

`Level` — Level for confidence bounds
95% (default) | numeric value

Specify the level at which to compute the confidence bounds. This check box is enabled only if you specify **Function** as **Cumulative probability (CDF)**, **Quantile (inverse CDF)**, **Survivor function**, or **Cumulative hazard**.

`Plot function` — Flag to plot function
unchecked (default) | checked

Select **Plot function** to display a plot of the distribution function, evaluated at the points that you enter in the **At x =** field, in a new window. This parameter is optional.

Exclude

`Exclusion rule name` — Name of exclusion rule
character vector

Specify a name for the exclusion rule as a character vector.

`Exclude sections` — Define data exclusion rules numerically
numeric value

Specify lower and upper limits for the data numerically.

`Exclude graphically` — Define data exclusion rules graphically
list of variables

Specify lower and upper limits for the data by selecting a variable from the **Select data** drop-down list and clicking **Exclude graphically**. An interactive plot opens in a new window, where you can add lower or upper limits by clicking and dragging a boundary on the plot.

`Existing exclusion rules` — List of existing exclusion rules
list of exclusion rules

Select an existing exclusion rule from the list. You can copy, view, rename, or delete exclusion rules by clicking the appropriate button.

Programmatic Use

`distributionFitter` opens the Distribution Fitter app, or brings focus to the app if it is already open.

`distributionFitter(y)` opens the Distribution Fitter app populated with the data specified by the vector `y`.

`distributionFitter(y,cens)` uses the vector `cens` to specify whether the observation `y(j)` is censored, `(cens(j)==1)`, or observed exactly, `(cens(j)==0)`. If `cens` is omitted or empty, then no `y` values are censored.

`distributionFitter(y,cens,freq)` uses the vector `freq` to specify the frequency of each element contained in `y`. If `freq` is omitted or empty, then all values in `y` have a frequency of 1.

`distributionFitter(y,cens,freq,dsname)` creates a data set with the name `dsname` using the data vector, `y`, censoring indicator, `cens`, and frequency vector, `freq`. Specify `dsname` as a character vector or string scalar, for example, `'mydata'`.

See Also

Functions

`distributionFitter` | `fitdist` | `makedist`

Topics

“Fit a Distribution Using the Distribution Fitter App” on page 5-71
“Model Data Using the Distribution Fitter App” on page 5-51
“Working with Probability Distributions” on page 5-3
“Supported Distributions” on page 5-14

Introduced before R2006a

fitckernel

Fit Gaussian kernel classification model using random feature expansion

Syntax

```
Mdl = fitckernel(X,Y)

Mdl = fitckernel(Tbl,ResponseVarName)
Mdl = fitckernel(Tbl,formula)
Mdl = fitckernel(Tbl,Y)

Mdl = fitckernel( ___ ,Name,Value)
[Mdl,FitInfo] = fitckernel( ___ )
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitckernel( ___ )
```

Description

`fitckernel` trains or cross-validates a binary Gaussian kernel classification model for nonlinear classification. `fitckernel` is more practical for big data applications that have large training sets but can also be applied to smaller data sets that fit in memory.

`fitckernel` maps data in a low-dimensional space into a high-dimensional space, then fits a linear model in the high-dimensional space by minimizing the regularized objective function. Obtaining the linear model in the high-dimensional space is equivalent to applying the Gaussian kernel to the model in the low-dimensional space. Available linear classification models include regularized support vector machine (SVM) and logistic regression models.

To train a nonlinear SVM model for binary classification of in-memory data, see `fitcsvm`.

`Mdl = fitckernel(X,Y)` returns a binary Gaussian kernel classification model trained using the predictor data in `X` and the corresponding class labels in `Y`. The `fitckernel` function maps the predictors in a low-dimensional space into a high-dimensional space, then fits a binary SVM model to the transformed predictors and class labels. This linear model is equivalent to the Gaussian kernel classification model in the low-dimensional space.

`Mdl = fitckernel(Tbl,ResponseVarName)` returns a kernel classification model `Mdl` trained using the predictor variables contained in the table `Tbl` and the class labels in `Tbl.ResponseVarName`.

`Mdl = fitckernel(Tbl,formula)` returns a kernel classification model trained using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitckernel(Tbl,Y)` returns a kernel classification model using the predictor variables in the table `Tbl` and the class labels in vector `Y`.

`Mdl = fitckernel( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can implement logistic regression, specify the number of dimensions of the expanded space, or specify to cross-validate.

`[Mdl,FitInfo] = fitckernel( ___ )` also returns the fit information in the structure array `FitInfo` using any of the input arguments in the previous syntaxes. You cannot request `FitInfo` for cross-validated models.

`[Mdl,FitInfo,HyperparameterOptimizationResults] = fitckernel( ___ )` also returns the hyperparameter optimization results `HyperparameterOptimizationResults` when you optimize hyperparameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Examples

Train Kernel Classification Model

Train a binary kernel classification model using SVM.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
[n,p] = size(X)
```

```
n = 351
```

```
p = 34
```

```
resp = unique(Y)
```

```
resp = 2x1 cell
    {'b'}
    {'g'}
```

Train a binary kernel classification model that identifies whether the radar return is bad (`'b'`) or good (`'g'`). Extract a fit summary to determine how well the optimization algorithm fits the model to the data.

```
rng('default') % For reproducibility
[Mdl,FitInfo] = fitckernel(X,Y)
```

```
Mdl =
  ClassificationKernel
              ResponseName: 'Y'
                ClassNames: {'b'  'g'}
                   Learner: 'svm'
    NumExpansionDimensions: 2048
               KernelScale: 1
                    Lambda: 0.0028
            BoxConstraint: 1


  Properties, Methods


FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'hinge'
                  Lambda: 0.0028
```

```
           BetaTolerance: 1.0000e-04
       GradientTolerance: 1.0000e-06
          ObjectiveValue: 0.2604
       GradientMagnitude: 0.0028
    RelativeChangeInBeta: 8.2512e-05
                 FitTime: 0.4139
                 History: []
```

`Mdl` is a `ClassificationKernel` model. To inspect the in-sample classification error, you can pass `Mdl` and the training data or new data to the `loss` function. Or, you can pass `Mdl` and new predictor data to the `predict` function to predict class labels for new observations. You can also pass `Mdl` and the training data to the `resume` function to continue training.

`FitInfo` is a structure array containing optimization information. Use `FitInfo` to determine whether optimization termination measurements are satisfactory.

For better accuracy, you can increase the maximum number of optimization iterations (`'IterationLimit'`) and decrease the tolerance values (`'BetaTolerance'` and `'GradientTolerance'`) by using the name-value pair arguments. Doing so can improve measures like `ObjectiveValue` and `RelativeChangeInBeta` in `FitInfo`. You can also optimize model parameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Cross-Validate Kernel Classification Model

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
rng('default') % For reproducibility
```

Cross-validate a binary kernel classification model. By default, the software uses 10-fold cross-validation.

```
CVMdl = fitckernel(X,Y,'CrossVal','on')
```

```
CVMdl =
  ClassificationPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 351
                  KFold: 10
              Partition: [1x1 cvpartition]
             ClassNames: {'b'  'g'}
         ScoreTransform: 'none'


  Properties, Methods
```

```
numel(CVMdl.Trained)
```

```
ans = 10
```

`CVMdl` is a `ClassificationPartitionedKernel` model. Because `fitckernel` implements 10-fold cross-validation, `CVMdl` contains 10 `ClassificationKernel` models that the software trains on training-fold (in-fold) observations.

Estimate the cross-validated classification error.

```
kfoldLoss(CVMdl)
```

```
ans = 0.0940
```

The classification error rate is approximately 9%.

Optimize Kernel Classifier

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value pair argument.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

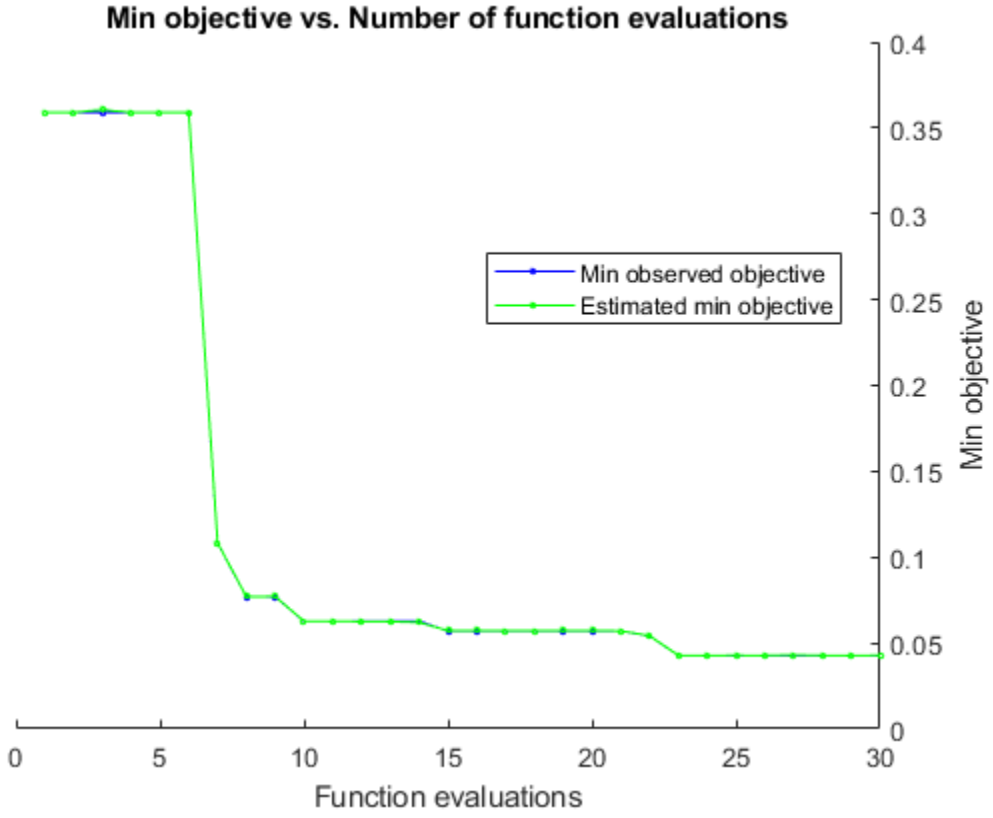
```
load ionosphere
```

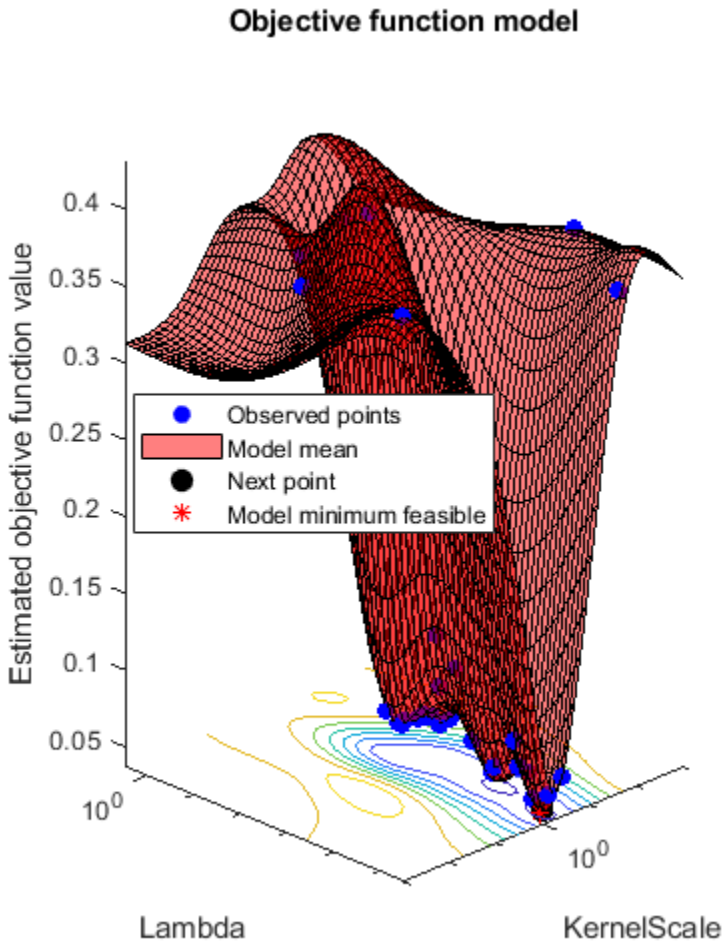
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization. Specify `'OptimizeHyperparameters'` as `'auto'` so that `fitckernel` finds optimal values of the `'KernelScale'` and `'Lambda'` name-value pair arguments. For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng('default')
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitckernel(X,Y,'OptimizeHyperparameters','auto
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName','expected-improvement-p
```

```
|=====================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  KernelScale |       La
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |     0.35897 |      1.2518 |     0.35897 |     0.35897 |       64.836 |   4.481
|    2 | Accept |     0.35897 |      2.1157 |     0.35897 |     0.35897 |     0.036335 |     0.0
|    3 | Accept |     0.39601 |      1.4585 |     0.35897 |     0.36053 |    0.0022147 |   6.8254
|    4 | Accept |     0.35897 |     0.75618 |     0.35897 |     0.35898 |       5.1259 |      0.2
|    5 | Accept |     0.35897 |      1.1115 |     0.35897 |     0.35897 |      0.24853 |      0.1
|    6 | Accept |     0.35897 |     0.44584 |     0.35897 |     0.35897 |       885.09 |   0.0005
|    7 | Best   |     0.10826 |      1.1051 |     0.10826 |     0.10833 |       8.0346 |    0.004
|    8 | Best   |    0.076923 |      1.3917 |    0.076923 |    0.076999 |       7.0902 |    0.003
|    9 | Accept |    0.091168 |      1.3441 |    0.076923 |    0.077059 |       9.1504 |    0.002
|   10 | Best   |    0.062678 |      1.4936 |    0.062678 |    0.062723 |       3.5487 |    0.002
|   11 | Accept |    0.062678 |       1.432 |    0.062678 |    0.062741 |       2.3869 |     0.00
|   12 | Accept |     0.41026 |      1.8381 |    0.062678 |    0.062536 |      0.14075 |    0.002
|   13 | Accept |    0.062678 |      1.5015 |    0.062678 |    0.062532 |       3.4215 |    0.003
|   14 | Accept |    0.062678 |      1.5722 |    0.062678 |    0.061956 |       3.2928 |    0.003
|   15 | Best   |     0.05698 |      1.3448 |     0.05698 |    0.057204 |       5.0598 |    0.002
|   16 | Accept |    0.062678 |      1.9683 |     0.05698 |    0.057186 |       5.3401 |    0.001
|   17 | Accept |     0.05698 |      1.3237 |     0.05698 |    0.057118 |        1.813 |    0.000
|   18 | Accept |    0.059829 |      1.5347 |     0.05698 |    0.057092 |       1.5122 |    0.004
|   19 | Accept |    0.059829 |      1.4056 |     0.05698 |     0.05718 |       1.9277 |    0.005
```

```
|   20 | Accept |    0.065527 |      1.2771 |     0.05698 |    0.057189 |       1.4064 |    0.009
|======================================================================================================
| Iter | Eval   | Objective   | Objective   | BestSoFar   | BestSoFar   |  KernelScale |        La
|      | result |             | runtime     | (observed)  | (estim.)    |              |
|======================================================================================================
|   21 | Accept |     0.05698 |      1.6767 |     0.05698 |    0.057033 |       5.1719 |    0.002
|   22 | Best   |    0.054131 |      3.6305 |    0.054131 |    0.054176 |       1.9618 |    6.5704
|   23 | Best   |    0.042735 |      1.6584 |    0.042735 |    0.042763 |       1.9463 |    1.0169
|   24 | Accept |    0.082621 |      1.5806 |    0.042735 |    0.042775 |       1.0661 |    1.3245
|   25 | Accept |    0.054131 |      3.4914 |    0.042735 |    0.042789 |        3.288 |    2.0035
|   26 | Accept |    0.062678 |      2.0516 |    0.042735 |    0.042769 |        2.657 |    3.0334
|   27 | Accept |    0.059829 |      2.2748 |    0.042735 |    0.043054 |       2.0381 |    1.9791
|   28 | Accept |    0.042735 |      2.9023 |    0.042735 |    0.042764 |       3.5043 |     0.000
|   29 | Accept |    0.054131 |      1.2828 |    0.042735 |    0.042764 |       1.3897 |    3.2288
|   30 | Accept |    0.062678 |      3.0746 |    0.042735 |    0.042792 |       2.2414 |     0.000
```






```
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 91.3356 seconds
Total objective function evaluation time: 51.2956

Best observed feasible point:
    KernelScale      Lambda
    ___________    __________

      1.9463       1.0169e-05

Observed objective function value = 0.042735
Estimated objective function value = 0.043106
Function evaluation time = 1.6584

Best estimated feasible point (according to models):
    KernelScale     Lambda
    ___________    _________
```

```
      3.5043       0.0001237

Estimated objective function value = 0.042792
Estimated function evaluation time = 3.2219

Mdl =
  ClassificationKernel
              ResponseName: 'Y'
                ClassNames: {'b'  'g'}
                   Learner: 'svm'
    NumExpansionDimensions: 2048
               KernelScale: 3.5043
                    Lambda: 1.2370e-04
            BoxConstraint: 23.0320


  Properties, Methods


FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'hinge'
                  Lambda: 1.2370e-04
           BetaTolerance: 1.0000e-04
       GradientTolerance: 1.0000e-06
          ObjectiveValue: 0.0426
       GradientMagnitude: 0.0028
    RelativeChangeInBeta: 8.9154e-05
                 FitTime: 0.7436
                 History: []


HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                     ObjectiveFcn: @createObjFcn/inMemoryObjFcn
             VariableDescriptions: [4x1 optimizableVariable]
                          Options: [1x1 struct]
                     MinObjective: 0.0427
                  XAtMinObjective: [1x2 table]
            MinEstimatedObjective: 0.0428
         XAtMinEstimatedObjective: [1x2 table]
          NumObjectiveEvaluations: 30
                 TotalElapsedTime: 91.3356
                        NextPoint: [1x2 table]
                           XTrace: [30x2 table]
                   ObjectiveTrace: [30x1 double]
                 ConstraintsTrace: []
                    UserDataTrace: {30x1 cell}
     ObjectiveEvaluationTimeTrace: [30x1 double]
               IterationTimeTrace: [30x1 double]
                       ErrorTrace: [30x1 double]
                 FeasibilityTrace: [30x1 logical]
      FeasibilityProbabilityTrace: [30x1 double]
              IndexOfMinimumTrace: [30x1 double]
           ObjectiveMinimumTrace: [30x1 double]
    EstimatedObjectiveMinimumTrace: [30x1 double]
```

For big data, the optimization procedure can take a long time. If the data set is too large to run the optimization procedure, you can try to optimize the parameters using only partial data. Use the `datasample` function and specify `'Replace','false'` to sample data without replacement.

Input Arguments

`X` — Predictor data

numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

`Y` — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

- `fitckernel` supports only binary classification. Either `Y` must contain exactly two distinct classes, or you must specify two classes for training by using the `ClassNames` name-value pair argument.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.
- A good practice is to specify the class order by using the `ClassNames` name-value pair argument.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if the response variable `Y` is stored as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If `Y` is a character array, then each element of the response variable must correspond to one row of the array.

A good practice is to specify the order of the classes by using the `ClassNames` name-value argument.

Data Types: `char` | `string`

`formula` — Explanatory model of response variable and subset of predictor variables
character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response variable
- At least one missing value in a predictor observation (row in `X` or `Tbl`)
- `NaN` value or `0` weight (`'Weights'`)

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `Mdl = fitckernel(X,Y,'Learner','logistic','NumExpansionDimensions',2^15,'KernelScal`

e','auto') implements logistic regression after mapping the predictor data to the 2^15 dimensional space using feature expansion with a kernel scale parameter selected by a heuristic procedure.

Kernel Classification Options

Learner — Linear classification model type

'svm' (default) | 'logistic'

Linear classification model type, specified as the comma-separated pair consisting of 'Learner' and 'svm' or 'logistic'.

In the following table, $f(x) = T(x)\beta + b$.

- x is an observation (row vector) from p predictor variables.
- $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
- β is a vector of m coefficients.
- b is the scalar bias.

Value	Algorithm	Response Range	Loss Function
'svm'	Support vector machine	$y \in \{-1,1\}$; 1 for the positive class and –1 otherwise	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$
'logistic'	Logistic regression	Same as 'svm'	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$

Example: 'Learner','logistic'

NumExpansionDimensions — Number of dimensions of expanded space

'auto' (default) | positive integer

Number of dimensions of the expanded space, specified as the comma-separated pair consisting of 'NumExpansionDimensions' and 'auto' or a positive integer. For 'auto', the fitckernel function selects the number of dimensions using 2.^ceil(min(log2(p)+5,15)), where p is the number of predictors.

For details, see “Random Feature Expansion” on page 33-6722.

Example: 'NumExpansionDimensions',2^15

Data Types: char | string | single | double

KernelScale — Kernel scale parameter

1 (default) | 'auto' | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of 'KernelScale' and 'auto' or a positive scalar. The software obtains a random basis for random feature expansion by using the kernel scale parameter. For details, see “Random Feature Expansion” on page 33-6722.

If you specify 'auto', then the software selects an appropriate kernel scale parameter using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed by using rng before training.

Example: 'KernelScale','auto'

Data Types: char | string | single | double

BoxConstraint — Box constraint
1 (default) | positive scalar

Box constraint, specified as the comma-separated pair consisting of 'BoxConstraint' and a positive scalar.

This argument is valid only when 'Learner' is 'svm'(default) and you do not specify a value for the regularization term strength 'Lambda'. You can specify either 'BoxConstraint' or 'Lambda' because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations.

Example: 'BoxConstraint',100

Data Types: single | double

Lambda — Regularization term strength
'auto' (default) | nonnegative scalar

Regularization term strength, specified as the comma-separated pair consisting of 'Lambda' and 'auto' or a nonnegative scalar.

For 'auto', the value of 'Lambda' is $1/n$, where n is the number of observations.

You can specify either 'BoxConstraint' or 'Lambda' because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$.

Example: 'Lambda',0.01

Data Types: char | string | single | double

Cross-Validation Options

CrossVal — Flag to train cross-validated classifier
'off' (default) | 'on'

Flag to train a cross-validated classifier, specified as the comma-separated pair consisting of 'Crossval' and 'on' or 'off'.

If you specify 'on', then the software trains a cross-validated classifier with 10 folds.

You can override this cross-validation setting using the CVPartition, Holdout, KFold, or Leaveout name-value pair argument. You can use only one cross-validation name-value pair argument at a time to create a cross-validated model.

Example: 'Crossval','on'

CVPartition — Cross-validation partition
[] (default) | cvpartition partition object

Cross-validation partition, specified as a cvpartition partition object created by cvpartition. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. If you specify `'Leaveout','on'`, then, for each of the *n* observations (where *n* is the number of observations excluding missing observations), the software completes these steps:

1. Reserve the observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in the cells of an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four name-value pair arguments only: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term

`1e–5` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'BetaTolerance',1e–6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance

`1e–6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'GradientTolerance',1e–5`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of optimization iterations

positive integer

Maximum number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The default value is 1000 if the transformed data fits in memory, as specified by the `BlockSize` name-value pair argument. Otherwise, the default value is 100.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Other Kernel Classification Options

`BlockSize` — Maximum amount of allocated memory

`4e^3` (4GB) (default) | positive scalar

Maximum amount of allocated memory (in megabytes), specified as the comma-separated pair consisting of `'BlockSize'` and a positive scalar.

If `fitckernel` requires more memory than the value of `'BlockSize'` to hold the transformed predictor data, then the software uses a block-wise strategy. For details about the block-wise strategy, see “Algorithms” on page 33-6723.

Example: `'BlockSize',1e4`

Data Types: `single` | `double`

`RandomStream` — Random number stream
global stream (default) | random stream object

Random number stream for reproducibility of data transformation, specified as the comma-separated pair consisting of `'RandomStream'` and a random stream object. For details, see “Random Feature Expansion” on page 33-6722.

Use `'RandomStream'` to reproduce the random basis functions that `fitckernel` uses to transform the predictor data to a high-dimensional space. For details, see “Managing the Global Stream Using RandStream” and “Creating and Controlling a Random Number Stream”.

Example: `'RandomStream',RandStream('mlfg6331_64')`

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of the history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. At each iteration, `fitckernel` composes the Hessian approximation by using statistics from the latest `HessianHistorySize` iterations.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `0` or `1`. `Verbose` controls the display of diagnostic information at the command line.

Value	Description
`0`	`fitckernel` does not display diagnostic information.
`1`	`fitckernel` displays and stores the value of the objective function, gradient magnitude, and other diagnostic information. `FitInfo.History` contains the diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

Other Classification Options

`CategoricalPredictors` — Categorical predictors list
vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitckernel` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitckernel` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitckernel` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitckernel` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered categorical variable, `fitckernel` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitckernel` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`ClassNames` — Names of classes to use for training

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Names of classes to use for training, specified as a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. `ClassNames` must have the same data type as the response variable in `Tbl` or `Y`.

If `ClassNames` is a character array, then each element must correspond to one row of the array.

Use `ClassNames` to:

- Specify the order of the classes during training.
- Specify the order of any input or output argument dimension that corresponds to the class order. For example, use `ClassNames` to specify the order of the dimensions of `Cost` or the column order of classification scores returned by `predict`.

- Select a subset of classes for training. For example, suppose that the set of all distinct class names in `Y` is `{'a','b','c'}`. To train the model using observations from classes `'a'` and `'c'` only, specify `'ClassNames',{'a','c'}`.

The default value for `ClassNames` is the set of all distinct class names in the response variable in `Tbl` or `Y`.

Example: `'ClassNames',{'b','g'}`

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification cost
square matrix | structure array

Misclassification cost, specified as the comma-separated pair consisting of `'Cost'` and a square matrix or structure.

- If you specify the square matrix `cost` (`'Cost',cost`), then `cost(i,j)` is the cost of classifying a point into class `j` if its true class is `i`. That is, the rows correspond to the true class, and the columns correspond to the predicted class. To specify the class order for the corresponding rows and columns of `cost`, use the `ClassNames` name-value pair argument.
- If you specify the structure `S` (`'Cost',S`), then it must have two fields:
  - `S.ClassNames`, which contains the class names as a variable of the same data type as `Y`
  - `S.ClassificationCosts`, which contains the cost matrix with rows and columns ordered as in `S.ClassNames`

The default value for `Cost` is `ones(`K`) - eye(`K`)`, where K is the number of distinct classes.

`fitckernel` uses `Cost` to adjust the prior class probabilities specified in `Prior`. Then, `fitckernel` uses the adjusted prior probabilities for training and resets the cost matrix to its default.

Example: `'Cost',[0 2; 1 0]`

Data Types: `single` | `double` | `struct`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitckernel` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.

- By default, `PredictorNames` contains the names of all predictor variables.
- A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`Prior` — Prior probabilities

`'empirical'` (default) | `'uniform'` | numeric vector | structure array

Prior probabilities for each class, specified as the comma-separated pair consisting of `'Prior'` and `'empirical'`, `'uniform'`, a numeric vector, or a structure array.

This table summarizes the available options for setting prior probabilities.

Value	Description
`'empirical'`	The class prior probabilities are the class relative frequencies in `Y`.
`'uniform'`	All class prior probabilities are equal to 1/*K*, where *K* is the number of classes.
numeric vector	Each element is a class prior probability. Order the elements according to their order in `Y`. If you specify the order using the `'ClassNames'` name-value pair argument, then order the elements accordingly.
structure array	A structure `S` with two fields: • `S.ClassNames` contains the class names as a variable of the same type as `Y`. • `S.ClassProbs` contains a vector of corresponding prior probabilities.

`fitckernel` normalizes the prior probabilities in `Prior` to sum to 1.

Example: `'Prior',struct('ClassNames',{{'setosa','versicolor'}},'ClassProbs',1:2)`

Data Types: `char` | `string` | `double` | `single` | `struct`

`ResponseName` — Response variable name

`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ScoreTransform` — Score transformation

`'none'` (default) | `'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | function handle | ...

Score transformation, specified as a character vector, string scalar, or function handle.

This table summarizes the available character vectors and string scalars.

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

For a MATLAB function or a function you define, use its function handle for the score transform. The function handle must accept a matrix (the original scores) and return a matrix of the same size (the transformed scores).

Example: `'ScoreTransform','logit'`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

nonnegative numeric vector | name of variable in `Tbl`

Observation weights, specified as a nonnegative numeric vector or the name of a variable in `Tbl`. The software weights each observation in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of observations in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if the weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors or the response variable when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

The software normalizes `Weights` to sum to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of these values:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'KernelScale','Lambda'}`.
- `'all'` — Optimize all eligible parameters.
- Cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitckernel` by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair argument.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitckernel` are:

- `KernelScale` — `fitckernel` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.
- `Lambda` — `fitckernel` searches among positive values, by default log-scaled in the range `[1e-3,1e3]/n`, where `n` is the number of observations.
- `Learner` — `fitckernel` searches among `'svm'` and `'logistic'`.
- `NumExpansionDimensions` — `fitckernel` searches among positive integers, by default log-scaled in the range `[100,10000]`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example:

```
load fisheriris
params = hyperparameters('fitckernel',meas,species);
params(2).Range = [1e-4,1e6];
```

Pass `params` as the value of `'OptimizeHyperparameters'`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Kernel Classifier” on page 33-6703.

Example: `'OptimizeHyperparameters','auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`

Field Name	Values	Default
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	

Field Name	Values	Default
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained kernel classification model

`ClassificationKernel` model object | `ClassificationPartitionedKernel` cross-validated model object

Trained kernel classification model, returned as a `ClassificationKernel` model object or `ClassificationPartitionedKernel` cross-validated model object.

If you set any of the name-value pair arguments `CrossVal`, `CVPartition`, `Holdout`, `KFold`, or `Leaveout`, then `Mdl` is a `ClassificationPartitionedKernel` cross-validated classifier. Otherwise, `Mdl` is a `ClassificationKernel` classifier.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.NumExpansionDimensions` in the Command Window to display the number of dimensions of the expanded space.

Note Unlike other classification models, and for economical memory usage, a `ClassificationKernel` model object does not store the training data or training process details (for example, convergence history).

`FitInfo` — Optimization details

structure array

Optimization details, returned as a structure array including fields described in this table. The fields contain final values or name-value pair argument specifications.

Field	Description
`Solver`	Objective function minimization technique: `'LBFGS-fast'`, `'LBFGS-blockwise'`, or `'LBFGS-tall'`. For details, see “Algorithms” on page 33-6723.
`LossFunction`	Loss function. Either `'hinge'` or `'logit'` depending on the type of linear classification model. See `Learner`.
`Lambda`	Regularization term strength. See `Lambda`.
`BetaTolerance`	Relative tolerance on the linear coefficients and the bias term. See `BetaTolerance`.
`GradientTolerance`	Absolute gradient tolerance. See `GradientTolerance`.
`ObjectiveValue`	Value of the objective function when optimization terminates. The classification loss plus the regularization term compose the objective function.

Field	Description
`GradientMagnitude`	Infinite norm of the gradient vector of the objective function when optimization terminates. See `GradientTolerance`.
`RelativeChangeInBeta`	Relative changes in the linear coefficients and the bias term when optimization terminates. See `BetaTolerance`.
`FitTime`	Elapsed, wall-clock time (in seconds) required to fit the model to the data.
`History`	History of optimization information. This field is empty (`[]`) if you specify `'Verbose',0`. For details, see `Verbose` and "Algorithms" on page 33-6723.

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.ObjectiveValue` in the Command Window.

A good practice is to examine `FitInfo` to assess whether convergence is satisfactory.

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table of hyperparameters and associated values

Cross-validation optimization of hyperparameters, returned as a `BayesianOptimization` object or a table of hyperparameters and associated values. The output is nonempty when the value of `'OptimizeHyperparameters'` is not `'none'`. The output value depends on the `Optimizer` field value of the `'HyperparameterOptimizationOptions'` name-value pair argument:

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Limitations

- `fitckernel` does not accept initial conditions for the vector of coefficients beta (β) and bias term (b) used to determine the decision function, $f(x) = T(x)\beta + b$.
- `fitckernel` does not support standardization.

More About

Random Feature Expansion

Random feature expansion, such as Random Kitchen Sinks[1] and Fastfood[2], is a scheme to approximate Gaussian kernels of the kernel classification algorithm to use for big data in a computationally efficient way. Random feature expansion is more practical for big data applications that have large training sets, but can also be applied to smaller data sets that fit in memory.

The kernel classification algorithm searches for an optimal hyperplane that separates the data into two classes after mapping features into a high-dimensional space. Nonlinear features that are not linearly separable in a low-dimensional space can be separable in the expanded high-dimensional

space. All the calculations for hyperplane classification use only dot products. You can obtain a nonlinear classification model by replacing the dot product x_1x_2' with the nonlinear kernel function $G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle$, where x_i is the ith observation (row vector) and $\varphi(x_i)$ is a transformation that maps x_i to a high-dimensional space (called the "kernel trick"). However, evaluating $G(x_1,x_2)$ (Gram matrix) for each pair of observations is computationally expensive for a large data set (large n).

The random feature expansion scheme finds a random transformation so that its dot product approximates the Gaussian kernel. That is,

$$G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle \approx T(x_1)T(x_2)',$$

where $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$). The Random Kitchen Sink scheme uses the random transformation

$$T(x) = m^{-1/2}\exp(iZx')',$$

where $Z \in \mathbb{R}^{m \times p}$ is a sample drawn from $N\left(0, \sigma^{-2}\right)$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage. The Fastfood scheme introduces another random basis V instead of Z using Hadamard matrices combined with Gaussian scaling matrices. This random basis reduces the computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

The `fitckernel` function uses the Fastfood scheme for random feature expansion and uses linear classification to train a Gaussian kernel classification model. Unlike solvers in the `fitcsvm` function, which require computation of the n-by-n Gram matrix, the solver in `fitckernel` only needs to form a matrix of size n-by-m, with m typically much less than n for big data.

Box Constraint

A box constraint is a parameter that controls the maximum penalty imposed on margin-violating observations, and aids in preventing overfitting (regularization). Increasing the box constraint can lead to longer training times.

The box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations.

Algorithms

`fitckernel` minimizes the regularized objective function using a Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver with ridge (L_2) regularization. To find the type of LBFGS solver used for training, type `FitInfo.Solver` in the Command Window.

- `'LBFGS-fast'` — LBFGS solver.
- `'LBFGS-blockwise'` — LBFGS solver with a block-wise strategy. If `fitckernel` requires more memory than the value of `BlockSize` to hold the transformed predictor data, then it uses a block-wise strategy.
- `'LBFGS-tall'` — LBFGS solver with a block-wise strategy for tall arrays.

When `fitckernel` uses a block-wise strategy, `fitckernel` implements LBFGS by distributing the calculation of the loss and gradient among different parts of the data at each iteration. Also, `fitckernel` refines the initial estimates of the linear coefficients and the bias term by fitting the model locally to parts of the data and combining the coefficients by averaging. If you specify

'Verbose',1, then fitckernel displays diagnostic information for each data pass and stores the information in the History field of FitInfo.

When fitckernel does not use a block-wise strategy, the initial estimates are zeros. If you specify 'Verbose',1, then fitckernel displays diagnostic information for each iteration and stores the information in the History field of FitInfo.

References

[1] Rahimi, A., and B. Recht. "Random Features for Large-Scale Kernel Machines." *Advances in Neural Information Processing Systems*. Vol. 20, 2008, pp. 1177–1184.

[2] Le, Q., T. Sarlós, and A. Smola. "Fastfood — Approximating Kernel Expansions in Loglinear Time." *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, No. 3, 2013, pp. 244–252.

[3] Huang, P. S., H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. "Kernel methods match Deep Neural Networks on TIMIT." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 205–209.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- fitckernel does not support tall table data.
- Some name-value pair arguments have different defaults compared to the default values for the in-memory fitckernel function. Supported name-value pair arguments, and any differences, are:
  - 'Learner'
  - 'NumExpansionDimensions'
  - 'KernelScale'
  - 'BoxConstraint'
  - 'Lambda'
  - 'BetaTolerance' — Default value is relaxed to 1e–3.
  - 'GradientTolerance' — Default value is relaxed to 1e–5.
  - 'IterationLimit' — Default value is relaxed to 20.
  - 'BlockSize'
  - 'RandomStream'
  - 'HessianHistorySize'
  - 'Verbose' — Default value is 1.
  - 'ClassNames'
  - 'Cost'
  - 'Prior'
  - 'ScoreTransform'

- `'Weights'` — Value must be a tall array.
- `'OptimizeHyperparameters'`
- `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.

- If `'KernelScale'` is `'auto'`, then `fitckernel` uses the random stream controlled by `tallrng` for subsampling. For reproducibility, you must set a random number seed for both the global stream and the random stream controlled by `tallrng`.
- If `'Lambda'` is `'auto'`, then `fitckernel` might take an extra pass through the data to calculate the number of observations in `X`.
- `fitckernel` uses a block-wise strategy. For details, see “Algorithms” on page 33-6723.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`ClassificationKernel` | `ClassificationPartitionedKernel` | `bayesopt` | `bestPoint` | `fitclinear` | `fitcsvm` | `predict` | `resume` | `templateKernel`

Topics
“Train SVM Classifiers Using a Gaussian Kernel” on page 18-156
“Bayesian Optimization with Tall Arrays” on page 30-9

Introduced in R2017b

ClassificationKernel

Gaussian kernel classification model using random feature expansion

Description

`ClassificationKernel` is a trained model object for a binary Gaussian kernel classification model using random feature expansion. `ClassificationKernel` is more practical for big data applications that have large training sets but can also be applied to smaller data sets that fit in memory.

Unlike other classification models, and for economical memory usage, `ClassificationKernel` model objects do not store the training data. However, they do store information such as the number of dimensions of the expanded space, the kernel scale parameter, prior-class probabilities, and the regularization strength.

You can use trained `ClassificationKernel` models to continue training using the training data and to predict labels or classification scores for new data. For details, see `resume` and `predict`.

Creation

Create a `ClassificationKernel` object using the `fitckernel` function. This function maps data in a low-dimensional space into a high-dimensional space, then fits a linear model in the high-dimensional space by minimizing the regularized objective function. The linear model in the high-dimensional space is equivalent to the model with a Gaussian kernel in the low-dimensional space. Available linear classification models include regularized support vector machine (SVM) and logistic regression models.

Properties

Kernel Classification Properties

`Learner` — Linear classification model type
`'logistic'` | `'svm'`

Linear classification model type, specified as `'logistic'` or `'svm'`.

In the following table, $f(x) = T(x)\beta + b$.

- x is an observation (row vector) from p predictor variables.
- $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
- β is a vector of m coefficients.
- b is the scalar bias.

Value	Algorithm	Loss Function	FittedLoss Value
'logistic'	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$	'logit'
'svm'	Support vector machine	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$	'hinge'

NumExpansionDimensions — Number of dimensions of expanded space
positive integer

Number of dimensions of the expanded space, specified as a positive integer.

Data Types: single | double

KernelScale — Kernel scale parameter
positive scalar

Kernel scale parameter, specified as a positive scalar.

Data Types: char | single | double

BoxConstraint — Box constraint
positive scalar

Box constraint, specified as a positive scalar.

Data Types: double | single

Lambda — Regularization term strength
nonnegative scalar

Regularization term strength, specified as a nonnegative scalar.

Data Types: single | double

FittedLoss — Loss function used to fit linear model
'hinge' | 'logit'

This property is read-only.

Loss function used to fit the linear model, specified as 'hinge' or 'logit'.

Value	Algorithm	Loss Function	Learner Value
'hinge'	Support vector machine	Hinge: $\ell[y, f(x)] = \max[0, 1 - yf(x)]$	'svm'
'logit'	Logistic regression	Deviance (logistic): $\ell[y, f(x)] = \log\{1 + \exp[-yf(x)]\}$	'logistic'

Regularization — Complexity penalty type
'ridge (L2)'

Complexity penalty type, which is always 'ridge (L2)'.

The software composes the objective function for minimization from the sum of the average loss function (see FittedLoss) and the regularization term, ridge (L_2) penalty.

The ridge (L_2) penalty is

$$\frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

where λ specifies the regularization term strength (see `Lambda`). The software excludes the bias term (β_0) from the regularization penalty.

Other Classification Properties

`CategoricalPredictors` — Indices of categorical predictors
vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ClassNames` — Unique class labels
categorical array | character array | logical vector | numeric vector | cell array of character vectors

Unique class labels used in training, specified as a categorical or character array, logical or numeric vector, or cell array of character vectors. `ClassNames` has the same data type as the class labels `Y`. (The software treats string arrays as cell arrays of character vectors.) `ClassNames` also determines the class order.

Data Types: `categorical` | `char` | `logical` | `single` | `double` | `cell`

`Cost` — Misclassification costs
square numeric matrix

This property is read-only.

Misclassification costs, specified as a square numeric matrix. `Cost` has K rows and columns, where K is the number of classes.

`Cost(`*`i`*`,`*`j`*`)` is the cost of classifying a point into class *`j`* if its true class is *`i`*. The order of the rows and columns of `Cost` corresponds to the order of the classes in `ClassNames`.

Data Types: `double`

`ModelParameters` — Parameters used for training model
structure

Parameters used for training the `ClassificationKernel` model, specified as a structure.

Access fields of `ModelParameters` using dot notation. For example, access the relative tolerance on the linear coefficients and the bias term by using `Mdl.ModelParameters.BetaTolerance`.

Data Types: `struct`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns used as predictor variables in the training data `X` or `Tbl`.

Data Types: `cell`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`Prior` — Prior class probabilities
numeric vector

This property is read-only.

Prior class probabilities, specified as a numeric vector. `Prior` has as many elements as classes in `ClassNames`, and the order of the elements corresponds to the elements of `ClassNames`.

Data Types: `double`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ScoreTransform` — Score transformation function to apply to predicted scores
`'doublelogit'` | `'invlogit'` | `'ismax'` | `'logit'` | `'none'` | function handle | ...

Score transformation function to apply to predicted scores, specified as a function name or function handle.

For kernel classification models and before the score transformation, the predicted classification score for the observation x (row vector) is $f(x) = T(x)\beta + b$.

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

To change the score transformation function to *`function`*, for example, use dot notation.

- For a built-in function, enter this code and replace *`function`* with a value from the table.

  ```
  Mdl.ScoreTransform = 'function';
  ```

Value	Description
`'doublelogit'`	$1/(1 + e^{-2x})$

Value	Description
`'invlogit'`	$\log(x / (1 - x))$
`'ismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to 0
`'logit'`	$1/(1 + e^{-x})$
`'none'` or `'identity'`	x (no transformation)
`'sign'`	–1 for $x < 0$ 0 for $x = 0$ 1 for $x > 0$
`'symmetric'`	$2x - 1$
`'symmetricismax'`	Sets the score for the class with the largest score to 1, and sets the scores for all other classes to –1
`'symmetriclogit'`	$2/(1 + e^{-x}) - 1$

- For a MATLAB function, or a function that you define, enter its function handle.

  ```
  Mdl.ScoreTransform = @function;
  ```

  `function` must accept a matrix of the original scores for each class, and then return a matrix of the same size representing the transformed scores for each class.

Data Types: `char` | `function_handle`

Object Functions

edge	Classification edge for Gaussian kernel classification model
lime	Local interpretable model-agnostic explanations (LIME)
loss	Classification loss for Gaussian kernel classification model
margin	Classification margins for Gaussian kernel classification model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict labels for Gaussian kernel classification model
resume	Resume training of Gaussian kernel classification model
shapley	Shapley values

Examples

Train Kernel Classification Model

Train a binary kernel classification model using SVM.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
[n,p] = size(X)
```

```
n = 351
```

```
p = 34
```

```
resp = unique(Y)

resp = 2x1 cell
    {'b'}
    {'g'}
```

Train a binary kernel classification model that identifies whether the radar return is bad (`'b'`) or good (`'g'`). Extract a fit summary to determine how well the optimization algorithm fits the model to the data.

```
rng('default') % For reproducibility
[Mdl,FitInfo] = fitckernel(X,Y)

Mdl =
  ClassificationKernel
             ResponseName: 'Y'
               ClassNames: {'b'  'g'}
                  Learner: 'svm'
    NumExpansionDimensions: 2048
               KernelScale: 1
                    Lambda: 0.0028
            BoxConstraint: 1


  Properties, Methods


FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'hinge'
                  Lambda: 0.0028
           BetaTolerance: 1.0000e-04
       GradientTolerance: 1.0000e-06
          ObjectiveValue: 0.2604
       GradientMagnitude: 0.0028
    RelativeChangeInBeta: 8.2512e-05
                 FitTime: 0.4139
                 History: []
```

`Mdl` is a `ClassificationKernel` model. To inspect the in-sample classification error, you can pass `Mdl` and the training data or new data to the `loss` function. Or, you can pass `Mdl` and new predictor data to the `predict` function to predict class labels for new observations. You can also pass `Mdl` and the training data to the `resume` function to continue training.

`FitInfo` is a structure array containing optimization information. Use `FitInfo` to determine whether optimization termination measurements are satisfactory.

For better accuracy, you can increase the maximum number of optimization iterations (`'IterationLimit'`) and decrease the tolerance values (`'BetaTolerance'` and `'GradientTolerance'`) by using the name-value pair arguments. Doing so can improve measures like `ObjectiveValue` and `RelativeChangeInBeta` in `FitInfo`. You can also optimize model parameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Predict Class Labels and Resume Training

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 20% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.20);
trainingInds = training(Partition); % Indices for the training set
XTrain = X(trainingInds,:);
YTrain = Y(trainingInds);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a binary kernel classification model that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

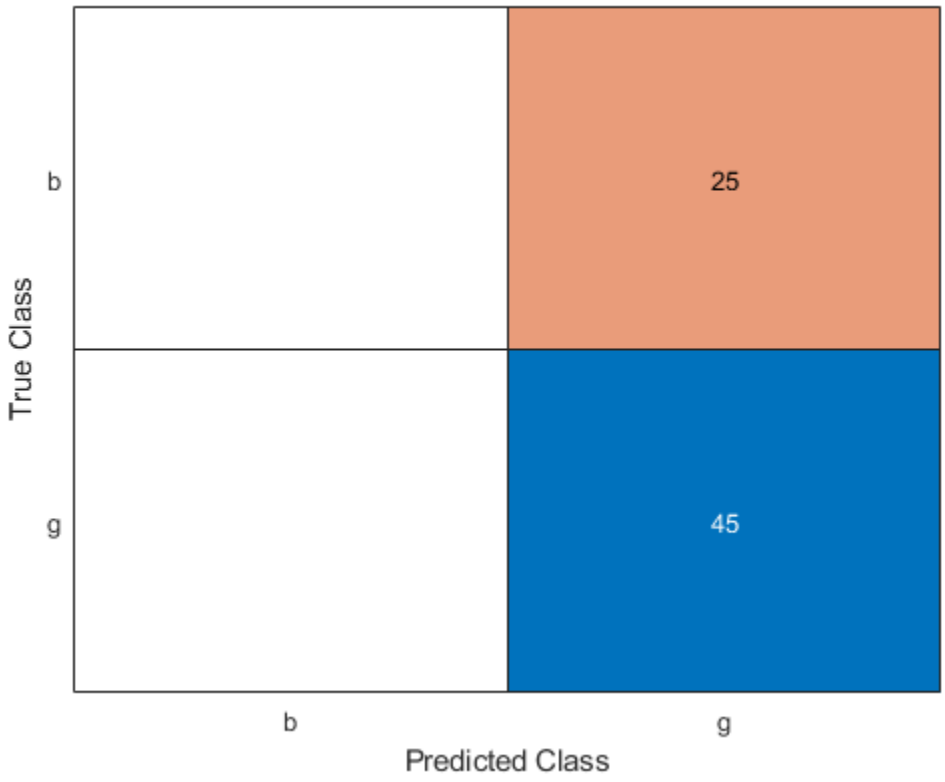
```
Mdl = fitckernel(XTrain,YTrain,'IterationLimit',5,'Verbose',1);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            0 |  1.000000e+00 |  0.000000e+00 |  2.811388e-01 |
|  LBFGS |      1 |            1 |  7.585395e-01 |  4.000000e+00 |  3.594306e-01 |   1.000000e+00
|  LBFGS |      1 |            2 |  7.160994e-01 |  1.000000e+00 |  2.028470e-01 |   6.923988e-01
|  LBFGS |      1 |            3 |  6.825272e-01 |  1.000000e+00 |  2.846975e-02 |   2.388909e-01
|  LBFGS |      1 |            4 |  6.699435e-01 |  1.000000e+00 |  1.779359e-02 |   1.325304e-01
|  LBFGS |      1 |            5 |  6.535619e-01 |  1.000000e+00 |  2.669039e-01 |   4.112952e-01
|=============================================================================================================
```

`Mdl` is a `ClassificationKernel` model.

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set.

```
label = predict(Mdl,XTest);
ConfusionTest = confusionchart(YTest,label);
```




```
L = loss(Mdl,XTest,YTest)
```

```
L = 0.3594
```

`Mdl` misclassifies all bad radar returns as good returns.

Continue training by using `resume`. This function continues training with the same options used for training `Mdl`.

```
UpdatedMdl = resume(Mdl,XTrain,YTrain);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient   |    Relative
|        |        |               |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |             0 |  6.535619e-01 |  0.000000e+00 |  2.669039e-01 |
|  LBFGS |      1 |             1 |  6.132547e-01 |  1.000000e+00 |  6.355537e-03 |   1.522092e-01
|  LBFGS |      1 |             2 |  5.938316e-01 |  4.000000e+00 |  3.202847e-02 |   1.498036e-01
|  LBFGS |      1 |             3 |  4.169274e-01 |  1.000000e+00 |  1.530249e-01 |   7.234253e-01
|  LBFGS |      1 |             4 |  3.679212e-01 |  5.000000e-01 |  2.740214e-01 |   2.495886e-01
|  LBFGS |      1 |             5 |  3.332261e-01 |  1.000000e+00 |  1.423488e-02 |   9.558680e-02
|  LBFGS |      1 |             6 |  3.235335e-01 |  1.000000e+00 |  7.117438e-03 |   7.137260e-02
|  LBFGS |      1 |             7 |  3.112331e-01 |  1.000000e+00 |  6.049822e-02 |   1.252157e-01
|  LBFGS |      1 |             8 |  2.972144e-01 |  1.000000e+00 |  7.117438e-03 |   5.796240e-02
|  LBFGS |      1 |             9 |  2.837450e-01 |  1.000000e+00 |  8.185053e-02 |   1.484733e-01
|  LBFGS |      1 |            10 |  2.797642e-01 |  1.000000e+00 |  3.558719e-02 |   5.856842e-02
|  LBFGS |      1 |            11 |  2.771280e-01 |  1.000000e+00 |  2.846975e-02 |   2.349433e-02
|  LBFGS |      1 |            12 |  2.741570e-01 |  1.000000e+00 |  3.914591e-02 |   3.113194e-02
```

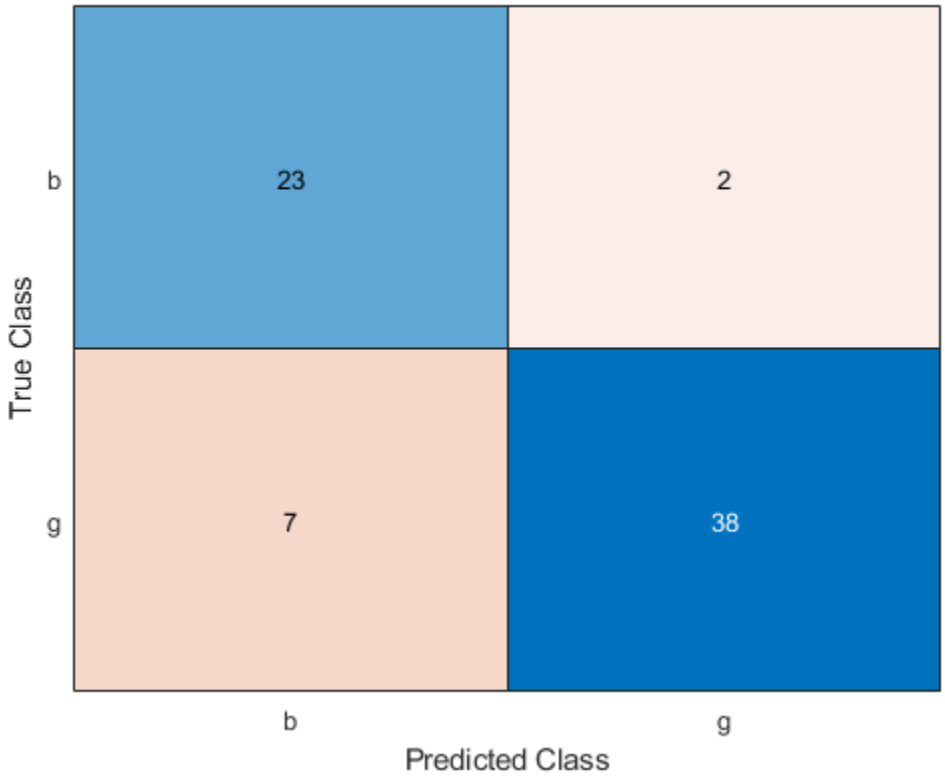
```
|  LBFGS |      1 |           13 |  2.725701e-01 |  5.000000e-01 |  1.067616e-01 |   8.729821e-02
|  LBFGS |      1 |           14 |  2.667147e-01 |  1.000000e+00 |  3.914591e-02 |   3.491723e-02
|  LBFGS |      1 |           15 |  2.621152e-01 |  1.000000e+00 |  7.117438e-03 |   5.104726e-02
|  LBFGS |      1 |           16 |  2.601652e-01 |  1.000000e+00 |  3.558719e-02 |   3.764904e-02
|  LBFGS |      1 |           17 |  2.589052e-01 |  1.000000e+00 |  3.202847e-02 |   3.655744e-02
|  LBFGS |      1 |           18 |  2.583185e-01 |  1.000000e+00 |  7.117438e-03 |   6.490571e-02
|  LBFGS |      1 |           19 |  2.556482e-01 |  1.000000e+00 |  9.252669e-02 |   4.601390e-02
|  LBFGS |      1 |           20 |  2.542643e-01 |  1.000000e+00 |  7.117438e-02 |   4.141838e-02
|=========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=========================================================================================================
|  LBFGS |      1 |           21 |  2.532117e-01 |  1.000000e+00 |  1.067616e-02 |   1.661720e-02
|  LBFGS |      1 |           22 |  2.529890e-01 |  1.000000e+00 |  2.135231e-02 |   1.231678e-02
|  LBFGS |      1 |           23 |  2.523232e-01 |  1.000000e+00 |  3.202847e-02 |   1.958586e-02
|  LBFGS |      1 |           24 |  2.506736e-01 |  1.000000e+00 |  1.779359e-02 |   2.474613e-02
|  LBFGS |      1 |           25 |  2.501995e-01 |  1.000000e+00 |  1.779359e-02 |   2.514352e-02
|  LBFGS |      1 |           26 |  2.488242e-01 |  1.000000e+00 |  3.558719e-03 |   1.531810e-02
|  LBFGS |      1 |           27 |  2.485295e-01 |  5.000000e-01 |  3.202847e-02 |   1.229760e-02
|  LBFGS |      1 |           28 |  2.482244e-01 |  1.000000e+00 |  4.270463e-02 |   8.970983e-03
|  LBFGS |      1 |           29 |  2.479714e-01 |  1.000000e+00 |  3.558719e-03 |   7.393900e-03
|  LBFGS |      1 |           30 |  2.477316e-01 |  1.000000e+00 |  3.202847e-02 |   3.268087e-03
|  LBFGS |      1 |           31 |  2.476178e-01 |  2.500000e-01 |  3.202847e-02 |   5.445890e-03
|  LBFGS |      1 |           32 |  2.474874e-01 |  1.000000e+00 |  1.779359e-02 |   3.535903e-03
|  LBFGS |      1 |           33 |  2.473980e-01 |  1.000000e+00 |  7.117438e-03 |   2.821725e-03
|  LBFGS |      1 |           34 |  2.472935e-01 |  1.000000e+00 |  3.558719e-03 |   2.699880e-03
|  LBFGS |      1 |           35 |  2.471418e-01 |  1.000000e+00 |  3.558719e-03 |   1.242523e-02
|  LBFGS |      1 |           36 |  2.469862e-01 |  1.000000e+00 |  2.846975e-02 |   7.895605e-03
|  LBFGS |      1 |           37 |  2.469598e-01 |  1.000000e+00 |  2.135231e-02 |   6.657676e-03
|  LBFGS |      1 |           38 |  2.466941e-01 |  1.000000e+00 |  3.558719e-02 |   4.654690e-03
|  LBFGS |      1 |           39 |  2.466660e-01 |  5.000000e-01 |  1.423488e-02 |   2.885769e-03
|  LBFGS |      1 |           40 |  2.465605e-01 |  1.000000e+00 |  3.558719e-03 |   4.562565e-03
|=========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=========================================================================================================
|  LBFGS |      1 |           41 |  2.465362e-01 |  1.000000e+00 |  1.423488e-02 |   5.652180e-03
|  LBFGS |      1 |           42 |  2.463528e-01 |  1.000000e+00 |  3.558719e-03 |   2.389759e-03
|  LBFGS |      1 |           43 |  2.463207e-01 |  1.000000e+00 |  1.511170e-03 |   3.738286e-03
|  LBFGS |      1 |           44 |  2.462585e-01 |  5.000000e-01 |  7.117438e-02 |   2.321693e-03
|  LBFGS |      1 |           45 |  2.461742e-01 |  1.000000e+00 |  7.117438e-03 |   2.599725e-03
|  LBFGS |      1 |           46 |  2.461434e-01 |  1.000000e+00 |  3.202847e-02 |   3.186923e-03
|  LBFGS |      1 |           47 |  2.461115e-01 |  1.000000e+00 |  7.117438e-03 |   1.530711e-03
|  LBFGS |      1 |           48 |  2.460814e-01 |  1.000000e+00 |  1.067616e-02 |   1.811714e-03
|  LBFGS |      1 |           49 |  2.460533e-01 |  5.000000e-01 |  1.423488e-02 |   1.012252e-03
|  LBFGS |      1 |           50 |  2.460111e-01 |  1.000000e+00 |  1.423488e-02 |   4.166762e-03
|  LBFGS |      1 |           51 |  2.459414e-01 |  1.000000e+00 |  1.067616e-02 |   3.271946e-03
|  LBFGS |      1 |           52 |  2.458809e-01 |  1.000000e+00 |  1.423488e-02 |   1.846440e-03
|  LBFGS |      1 |           53 |  2.458479e-01 |  1.000000e+00 |  1.067616e-02 |   1.180871e-03
|  LBFGS |      1 |           54 |  2.458146e-01 |  1.000000e+00 |  1.455008e-03 |   1.422954e-03
|  LBFGS |      1 |           55 |  2.457878e-01 |  1.000000e+00 |  7.117438e-03 |   1.880892e-03
|  LBFGS |      1 |           56 |  2.457519e-01 |  1.000000e+00 |  2.491103e-02 |   1.074764e-03
|  LBFGS |      1 |           57 |  2.457420e-01 |  1.000000e+00 |  7.473310e-02 |   9.511878e-04
|  LBFGS |      1 |           58 |  2.457212e-01 |  1.000000e+00 |  3.558719e-03 |   3.718564e-04
|  LBFGS |      1 |           59 |  2.457089e-01 |  1.000000e+00 |  4.270463e-02 |   6.237270e-04
|  LBFGS |      1 |           60 |  2.457047e-01 |  5.000000e-01 |  1.423488e-02 |   3.647573e-04
|=========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
```

```
|        |        |              |              |              |   magnitude   | change in Beta
|=============================================================================================
|  LBFGS |      1 |           61 |  2.456991e-01 |  1.000000e+00 |  1.423488e-02 |   5.666884e-04
|  LBFGS |      1 |           62 |  2.456898e-01 |  1.000000e+00 |  1.779359e-02 |   4.697056e-04
|  LBFGS |      1 |           63 |  2.456792e-01 |  1.000000e+00 |  1.779359e-02 |   5.984927e-04
|  LBFGS |      1 |           64 |  2.456603e-01 |  1.000000e+00 |  1.403782e-03 |   5.414985e-04
|  LBFGS |      1 |           65 |  2.456482e-01 |  1.000000e+00 |  3.558719e-03 |   6.506293e-04
|  LBFGS |      1 |           66 |  2.456358e-01 |  1.000000e+00 |  1.476262e-03 |   1.284139e-03
|  LBFGS |      1 |           67 |  2.456124e-01 |  1.000000e+00 |  3.558719e-03 |   8.636596e-04
|  LBFGS |      1 |           68 |  2.455980e-01 |  1.000000e+00 |  1.067616e-02 |   9.861527e-04
|  LBFGS |      1 |           69 |  2.455780e-01 |  1.000000e+00 |  1.067616e-02 |   5.102487e-04
|  LBFGS |      1 |           70 |  2.455633e-01 |  1.000000e+00 |  3.558719e-03 |   1.228077e-03
|  LBFGS |      1 |           71 |  2.455449e-01 |  1.000000e+00 |  1.423488e-02 |   7.864590e-04
|  LBFGS |      1 |           72 |  2.455261e-01 |  1.000000e+00 |  3.558719e-02 |   1.090815e-03
|  LBFGS |      1 |           73 |  2.455142e-01 |  1.000000e+00 |  1.067616e-02 |   1.701506e-03
|  LBFGS |      1 |           74 |  2.455075e-01 |  1.000000e+00 |  1.779359e-02 |   1.504577e-03
|  LBFGS |      1 |           75 |  2.455008e-01 |  1.000000e+00 |  3.914591e-02 |   1.144021e-03
|  LBFGS |      1 |           76 |  2.454943e-01 |  1.000000e+00 |  2.491103e-02 |   3.015254e-04
|  LBFGS |      1 |           77 |  2.454918e-01 |  5.000000e-01 |  3.202847e-02 |   9.837523e-04
|  LBFGS |      1 |           78 |  2.454870e-01 |  1.000000e+00 |  1.779359e-02 |   4.328953e-04
|  LBFGS |      1 |           79 |  2.454865e-01 |  5.000000e-01 |  3.558719e-03 |   7.126815e-04
|  LBFGS |      1 |           80 |  2.454775e-01 |  1.000000e+00 |  5.693950e-02 |   8.992562e-04
|=============================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================
|  LBFGS |      1 |           81 |  2.454686e-01 |  1.000000e+00 |  1.183730e-03 |   1.590246e-04
|  LBFGS |      1 |           82 |  2.454612e-01 |  1.000000e+00 |  2.135231e-02 |   1.389570e-04
|  LBFGS |      1 |           83 |  2.454506e-01 |  1.000000e+00 |  3.558719e-03 |   6.162089e-04
|  LBFGS |      1 |           84 |  2.454436e-01 |  1.000000e+00 |  1.423488e-02 |   1.877414e-03
|  LBFGS |      1 |           85 |  2.454378e-01 |  1.000000e+00 |  1.423488e-02 |   3.370852e-04
|  LBFGS |      1 |           86 |  2.454249e-01 |  1.000000e+00 |  1.423488e-02 |   8.133615e-04
|  LBFGS |      1 |           87 |  2.454101e-01 |  1.000000e+00 |  1.067616e-02 |   3.872088e-04
|  LBFGS |      1 |           88 |  2.453963e-01 |  1.000000e+00 |  1.779359e-02 |   5.670260e-04
|  LBFGS |      1 |           89 |  2.453866e-01 |  1.000000e+00 |  1.067616e-02 |   1.444984e-03
|  LBFGS |      1 |           90 |  2.453821e-01 |  1.000000e+00 |  7.117438e-03 |   2.457270e-03
|  LBFGS |      1 |           91 |  2.453790e-01 |  5.000000e-01 |  6.761566e-02 |   8.228766e-04
|  LBFGS |      1 |           92 |  2.453603e-01 |  1.000000e+00 |  2.135231e-02 |   1.084233e-03
|  LBFGS |      1 |           93 |  2.453540e-01 |  1.000000e+00 |  2.135231e-02 |   2.060005e-04
|  LBFGS |      1 |           94 |  2.453482e-01 |  1.000000e+00 |  1.779359e-02 |   1.560883e-04
|  LBFGS |      1 |           95 |  2.453461e-01 |  1.000000e+00 |  1.779359e-02 |   1.614693e-03
|  LBFGS |      1 |           96 |  2.453371e-01 |  1.000000e+00 |  3.558719e-02 |   2.145835e-04
|  LBFGS |      1 |           97 |  2.453305e-01 |  1.000000e+00 |  4.270463e-02 |   7.602088e-04
|  LBFGS |      1 |           98 |  2.453283e-01 |  2.500000e-01 |  2.135231e-02 |   3.422253e-04
|  LBFGS |      1 |           99 |  2.453246e-01 |  1.000000e+00 |  3.558719e-03 |   3.872561e-04
|  LBFGS |      1 |          100 |  2.453214e-01 |  1.000000e+00 |  3.202847e-02 |   1.732237e-04
|=============================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================
|  LBFGS |      1 |          101 |  2.453168e-01 |  1.000000e+00 |  1.067616e-02 |   3.065286e-04
|  LBFGS |      1 |          102 |  2.453155e-01 |  5.000000e-01 |  4.626335e-02 |   3.402368e-04
|  LBFGS |      1 |          103 |  2.453136e-01 |  1.000000e+00 |  1.779359e-02 |   2.215029e-04
|  LBFGS |      1 |          104 |  2.453119e-01 |  1.000000e+00 |  3.202847e-02 |   4.142355e-04
|  LBFGS |      1 |          105 |  2.453093e-01 |  1.000000e+00 |  1.423488e-02 |   2.186007e-04
|  LBFGS |      1 |          106 |  2.453090e-01 |  1.000000e+00 |  2.846975e-02 |   1.338602e-03
|  LBFGS |      1 |          107 |  2.453048e-01 |  1.000000e+00 |  1.423488e-02 |   3.208296e-04
|  LBFGS |      1 |          108 |  2.453040e-01 |  1.000000e+00 |  3.558719e-02 |   1.294488e-03
```

```
|  LBFGS |      1 |          109 |  2.452977e-01 |  1.000000e+00 |  1.423488e-02 |   8.328380e-04
|  LBFGS |      1 |          110 |  2.452934e-01 |  1.000000e+00 |  2.135231e-02 |   5.149259e-04
|  LBFGS |      1 |          111 |  2.452886e-01 |  1.000000e+00 |  1.779359e-02 |   3.650664e-04
|  LBFGS |      1 |          112 |  2.452854e-01 |  1.000000e+00 |  1.067616e-02 |   2.633981e-04
|  LBFGS |      1 |          113 |  2.452836e-01 |  1.000000e+00 |  1.067616e-02 |   1.804300e-04
|  LBFGS |      1 |          114 |  2.452817e-01 |  1.000000e+00 |  7.117438e-03 |   4.251642e-04
|  LBFGS |      1 |          115 |  2.452741e-01 |  1.000000e+00 |  1.779359e-02 |   9.018440e-04
|  LBFGS |      1 |          116 |  2.452691e-01 |  1.000000e+00 |  2.135231e-02 |   9.941716e-05
|========================================================================================================
```

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set.

```
UpdatedLabel = predict(UpdatedMdl,XTest);
UpdatedConfusionTest = confusionchart(YTest,UpdatedLabel);
```




```
UpdatedL = loss(UpdatedMdl,XTest,YTest)
```

```
UpdatedL = 0.1284
```

The classification error decreases after `resume` updates the classification model with more iterations.

See Also

`ClassificationLinear` | `fitckernel` | `fitclinear`

Introduced in R2017b

edge

Classification edge for Gaussian kernel classification model

Syntax

```
e = edge(Mdl,X,Y)

e = edge(Mdl,Tbl,ResponseVarName)
e = edge(Mdl,Tbl,Y)

e = edge( ___ ,'Weights',weights)
```

Description

`e = edge(Mdl,X,Y)` returns the classification edge on page 33-6741 for the binary Gaussian kernel classification model `Mdl` using the predictor data in `X` and the corresponding class labels in `Y`.

`e = edge(Mdl,Tbl,ResponseVarName)` returns the classification edge for the trained kernel classifier `Mdl` using the predictor data in table `Tbl` and the class labels in `Tbl.ResponseVarName`.

`e = edge(Mdl,Tbl,Y)` returns the classification edge for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

`e = edge( ___ ,'Weights',weights)` returns the weighted classification edge using the observation weights supplied in `weights`. Specify the weights after any of the input argument combinations in previous syntaxes.

Examples

Estimate Test-Set Edge

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model using the training set.

```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds));
```

Estimate the training-set edge and the test-set edge.

```
eTrain = edge(Mdl,X(trainingInds,:),Y(trainingInds))
```

```
eTrain = 2.1703

eTest = edge(Mdl,X(testInds,:),Y(testInds))

eTest = 1.5643
```

Feature Selection Using Test-Set Edges

Perform feature selection by comparing test-set edges from multiple models. Based solely on this criterion, the classifier with the highest edge is the best classifier.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
XTrain = X(trainingInds,:);
YTrain = Y(trainingInds);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds);
```

Randomly choose half of the predictor variables.

```
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.5*p));
```

Train two binary kernel classification models: one that uses all of the predictors, and one that uses half of the predictors.

```
Mdl = fitckernel(XTrain,YTrain);
PMdl = fitckernel(XTrain(:,idxPart),YTrain);
```

`Mdl` and `PMdl` are `ClassificationKernel` models.

Estimate the test-set edge for each classifier.

```
fullEdge = edge(Mdl,XTest,YTest)

fullEdge = 1.6335

partEdge = edge(PMdl,XTest(:,idxPart),YTest)

partEdge = 2.0205
```

Based on the test-set edges, the classifier that uses half of the predictors is the better model.

Input Arguments

`Mdl` — Binary kernel classification model

`ClassificationKernel` model object

Binary kernel classification model, specified as a `ClassificationKernel` model object. You can create a `ClassificationKernel` model object using `fitckernel`.

`X` — Predictor data

n-by-*p* numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors used to train `Mdl`.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

`Y` — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `edge` must also be in a table.

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

`weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as a numeric vector or the name of a variable in `Tbl`.

- If `weights` is a numeric vector, then the size of `weights` must be equal to the number of rows in `X` or `Tbl`.
- If `weights` is the name of a variable in `Tbl`, you must specify `weights` as a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, `edge` computes the weighted classification edge on page 33-6741. The software weights the observations in each row of `X` or `Tbl` with the corresponding weights in `weights`.

`edge` normalizes `weights` to sum up to the value of the prior probability in the respective class.

Data Types: `single` | `double` | `char` | `string`

Output Arguments

e — Classification edge
numeric scalar

Classification edge on page 33-6741, returned as a numeric scalar.

More About

Classification Edge

The classification edge is the weighted mean of the classification margins.

One way to choose among multiple classifiers, for example to perform feature selection, is to choose the classifier that yields the greatest edge.

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b\,.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `edge` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationKernel` | `fitckernel` | `margin` | `predict`

Introduced in R2017b

loss

Classification loss for Gaussian kernel classification model

Syntax

```
L = loss(Mdl,X,Y)

L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss( ___ ,Name,Value)
```

Description

`L = loss(Mdl,X,Y)` returns the classification loss on page 33-6747 for the binary Gaussian kernel classification model `Mdl` using the predictor data in `X` and the corresponding class labels in `Y`.

`L = loss(Mdl,Tbl,ResponseVarName)` returns the classification loss for the model `Mdl` using the predictor data in `Tbl` and the true class labels in `Tbl.ResponseVarName`.

`L = loss(Mdl,Tbl,Y)` returns the classification loss for the model `Mdl` using the predictor data in table `Tbl` and the true class labels in `Y`.

`L = loss( ___ ,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify a classification loss function and observation weights. Then, `loss` returns the weighted classification loss using the specified loss function.

Examples

Estimate Test-Set Classification Loss

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model using the training set.

```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds));
```

Estimate the training-set classification error and the test-set classification error.

```
ceTrain = loss(Mdl,X(trainingInds,:),Y(trainingInds))
```

```
ceTrain = 0.0067

ceTest = loss(Mdl,X(testInds,:),Y(testInds))

ceTest = 0.1140
```

Specify Custom Classification Loss

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model using the training set.

```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds));
```

Create an anonymous function that measures linear loss, that is,

$$L = \frac{\sum_j -w_j y_j f_j}{\sum_j w_j}.$$

w_j is the weight for observation j, y_j is response j (-1 for the negative class, and 1 otherwise), and f_j is the raw classification score of observation j.

```
linearloss = @(C,S,W,Cost)sum(-W.*sum(S.*C,2))/sum(W);
```

Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `'LossFun'` name-value pair argument.

Estimate the training-set classification loss and the test-set classification loss using the linear loss function.

```
ceTrain = loss(Mdl,X(trainingInds,:),Y(trainingInds),'LossFun',linearloss)

ceTrain = -1.0851

ceTest = loss(Mdl,X(testInds,:),Y(testInds),'LossFun',linearloss)

ceTest = -0.7821
```

Input Arguments

`Mdl` — Binary kernel classification model

`ClassificationKernel` model object

Binary kernel classification model, specified as a `ClassificationKernel` model object. You can create a `ClassificationKernel` model object using `fitckernel`.

`X` — Predictor data

n-by-*p* numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors used to train `Mdl`.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

`Y` — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `L = loss(Mdl,X,Y,'LossFun','quadratic','Weights',weights)` returns the weighted classification loss using the quadratic loss function.

LossFun — Loss function

`'classiferror'` (default) | `'binodeviance'` | `'exponential'` | `'hinge'` | `'logit'` | `'mincost'` | `'quadratic'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or a function handle.

- This table lists the available loss functions. Specify one using its corresponding value.

Value	Description
`'binodeviance'`	Binomial deviance
`'classiferror'`	Misclassified rate in decimal
`'exponential'`	Exponential loss
`'hinge'`	Hinge loss
`'logit'`	Logistic loss
`'mincost'`	Minimal expected misclassification cost (for classification scores that are posterior probabilities)
`'quadratic'`	Quadratic loss

  `'mincost'` is appropriate for classification scores that are posterior probabilities. For kernel classification models, logistic regression learners return posterior probabilities as classification scores by default, but SVM learners do not (see `predict`).

- To specify a custom loss function, use function handle notation. The function must have this form:

  `lossvalue = lossfun(C,S,W,Cost)`

  - The output argument `lossvalue` is a scalar.
  - You specify the function name (*`lossfun`*).
  - `C` is an `n`-by-`K` logical matrix with rows indicating the class to which the corresponding observation belongs. `n` is the number of observations in `Tbl` or `X`, and `K` is the number of distinct classes (`numel(Mdl.ClassNames)`. The column order corresponds to the class order in `Mdl.ClassNames`. Create `C` by setting `C(p,q) = 1`, if observation `p` is in class `q`, for each row. Set all other elements of row `p` to `0`.

- `S` is an `n`-by-`K` numeric matrix of classification scores. The column order corresponds to the class order in `Mdl.ClassNames`. `S` is a matrix of classification scores, similar to the output of `predict`.
- `W` is an `n`-by-1 numeric vector of observation weights.
- `Cost` is a `K`-by-`K` numeric matrix of misclassification costs. For example, `Cost = ones(K) – eye(K)` specifies a cost of `0` for correct classification and `1` for misclassification.

Example: `'LossFun',@lossfun`

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If `Weights` is a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `Tbl`.
- If `Weights` is the name of a variable in `Tbl`, you must specify `Weights` as a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, `loss` computes the weighted classification loss on page 33-6747 and normalizes the weights to sum up to the value of the prior probability in the respective class.

Data Types: `double` | `single` | `char` | `string`

Output Arguments

`L` — Classification loss
numeric scalar

Classification loss on page 33-6747, returned as a numeric scalar. The interpretation of `L` depends on `Weights` and `LossFun`.

More About

Classification Loss

Classification loss functions measure the predictive inaccuracy of classification models. When you compare the same type of loss among many models, a lower loss indicates a better predictive model.

Suppose the following:

- L is the weighted average classification loss.
- n is the sample size.
- y_j is the observed class label. The software codes it as –1 or 1, indicating the negative or positive class (or the first or second class in the `ClassNames` property), respectively.
- $f(X_j)$ is the positive-class classification score for observation (row) j of the predictor data X.
- $m_j = y_j f(X_j)$ is the classification score for classifying observation j into the class corresponding to y_j. Positive values of m_j indicate correct classification and do not contribute much to the average

loss. Negative values of m_j indicate incorrect classification and contribute significantly to the average loss.

- The weight for observation j is w_j. The software normalizes the observation weights so that they sum to the corresponding prior class probability. The software also normalizes the prior probabilities so that they sum to 1. Therefore,

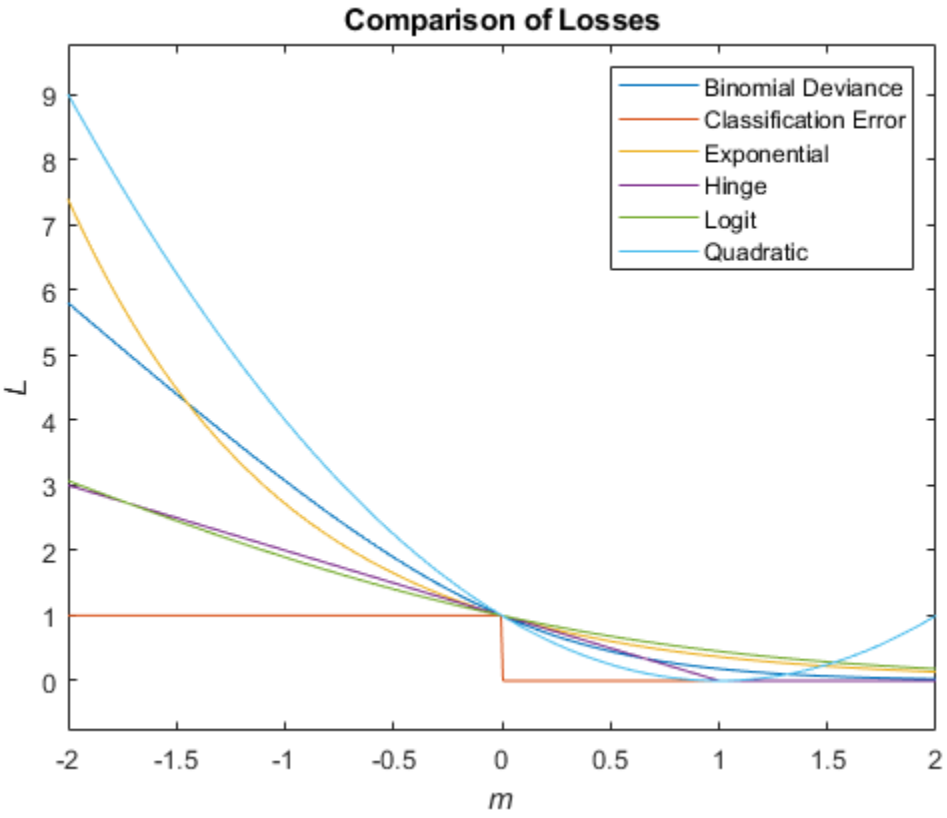
$$\sum_{j=1}^{n} w_j = 1.$$

This table describes the supported loss functions that you can specify by using the `'LossFun'` name-value argument.

Loss Function	Value of LossFun	Equation
Binomial deviance	`'binodeviance'`	$L = \sum_{j=1}^{n} w_j \log\{1 + \exp[-2m_j]\}.$
Exponential loss	`'exponential'`	$L = \sum_{j=1}^{n} w_j \exp(-m_j).$
Misclassified rate in decimal	`'classiferror'`	$L = \sum_{j=1}^{n} w_j I\{\hat{y}_j \neq y_j\}.$ $\hat{y}_j$ is the class label corresponding to the class with the maximal score. $I\{\cdot\}$ is the indicator function.
Hinge loss	`'hinge'`	$L = \sum_{j=1}^{n} w_j \max\{0, 1 - m_j\}.$
Logit loss	`'logit'`	$L = \sum_{j=1}^{n} w_j \log(1 + \exp(-m_j)).$

Loss Function	Value of LossFun	Equation
Minimal expected misclassification cost	`'mincost'`	`'mincost'` is appropriate only if classification scores are posterior probabilities. The software computes the weighted minimal expected classification cost using this procedure for observations $j = 1,...,n$. **1** Estimate the expected misclassification cost of classifying the observation X_j into the class k: $\gamma_{jk} = \left(f(X_j)'C\right)_k$. $f(X_j)$ is the column vector of class posterior probabilities for binary and multiclass classification for the observation X_j. C is the cost matrix stored in the `Cost` property of the model. **2** For observation j, predict the class label corresponding to the minimal expected misclassification cost: $\widehat{y}_j = \underset{k = 1,...,K}{\operatorname{argmin}} \gamma_{jk}$. **3** Using C, identify the cost incurred (c_j) for making the prediction. The weighted average of the minimal expected misclassification cost loss is $L = \sum_{j=1}^{n} w_j c_j$. If you use the default cost matrix (whose element value is 0 for correct classification and 1 for incorrect classification), then the `'mincost'` loss is equivalent to the `'classiferror'` loss.
Quadratic loss	`'quadratic'`	$L = \sum_{j=1}^{n} w_j \left(1 - m_j\right)^2$.

This figure compares the loss functions (except `'mincost'`) over the score m for one observation. Some functions are normalized to pass through the point (0,1).



Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `loss` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationKernel` | `fitckernel` | `predict`

Introduced in R2017b

margin

Classification margins for Gaussian kernel classification model

Syntax

```
m = margin(Mdl,X,Y)

m = margin(Mdl,Tbl,ResponseVarName)
m = margin(Mdl,Tbl,Y)
```

Description

`m = margin(Mdl,X,Y)` returns the classification margins on page 33-6755 for the binary Gaussian kernel classification model `Mdl` using the predictor data in `X` and the corresponding class labels in `Y`.

`m = margin(Mdl,Tbl,ResponseVarName)` returns the classification margins for the trained kernel classifier `Mdl` using the predictor data in table `Tbl` and the class labels in `Tbl.ResponseVarName`.

`m = margin(Mdl,Tbl,Y)` returns the classification margins for the classifier `Mdl` using the predictor data in table `Tbl` and the class labels in vector `Y`.

Examples

Estimate Test-Set Margins

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 30% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.30);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model using the training set.

```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds));
```

Estimate the training-set margins and test-set margins.

```
mTrain = margin(Mdl,X(trainingInds,:),Y(trainingInds));
mTest = margin(Mdl,X(testInds,:),Y(testInds));
```

Plot both sets of margins using box plots.

```
boxplot([mTrain; mTest],[zeros(size(mTrain,1),1); ones(size(mTest,1),1)], ...
    'Labels',{'Training set','Test set'});
title('Training-Set and Test-Set Margins')
```

The margin distribution of the training set is situated higher than the margin distribution of the test set.

Feature Selection Using Test-Set Margins

Perform feature selection by comparing test-set margins from multiple models. Based solely on this criterion, the classifier with the larger margins is the better classifier.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
XTrain = X(trainingInds,:);
YTrain = Y(trainingInds);
testInds = test(Partition); % Indices for the test set
```

```
XTest = X(testInds,:);
YTest = Y(testInds);
```

Randomly choose 10% of the predictor variables.

```
p = size(X,2); % Number of predictors
idxPart = randsample(p,ceil(0.1*p));
```

Train two binary kernel classification models: one that uses all of the predictors, and one that uses the random 10%.

```
Mdl = fitckernel(XTrain,YTrain);
PMdl = fitckernel(XTrain(:,idxPart),YTrain);
```

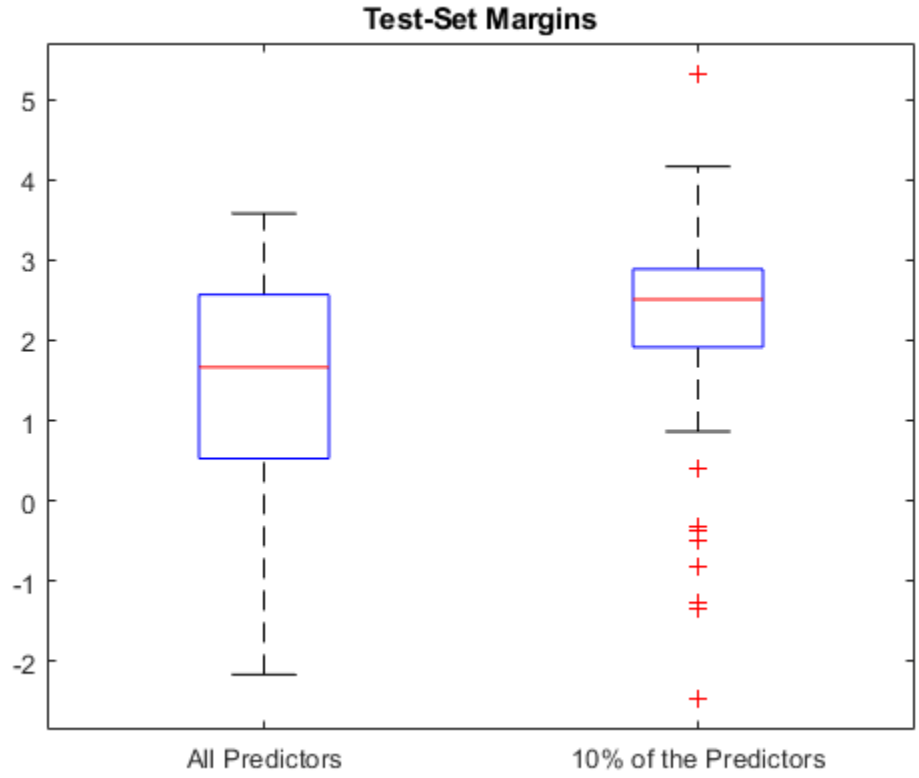
`Mdl` and `PMdl` are `ClassificationKernel` models.

Estimate the test-set margins for each classifier.

```
fullMargins = margin(Mdl,XTest,YTest);
partMargins = margin(PMdl,XTest(:,idxPart),YTest);
```

Plot the distribution of the margin sets using box plots.

```
boxplot([fullMargins partMargins], ...
    'Labels',{'All Predictors','10% of the Predictors'});
title('Test-Set Margins')
```



The margin distribution of `PMdl` is situated higher than the margin distribution of `Mdl`. Therefore, the `PMdl` model is the better classifier.

Input Arguments

`Mdl` — Binary kernel classification model

`ClassificationKernel` model object

Binary kernel classification model, specified as a `ClassificationKernel` model object. You can create a `ClassificationKernel` model object using `fitckernel`.

`X` — Predictor data

n-by-*p* numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors used to train `Mdl`.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

`Y` — Class labels

categorical array | character array | string array | logical vector | numeric vector | cell array of character vectors

Class labels, specified as a categorical, character, or string array; logical or numeric vector; or cell array of character vectors.

- The data type of `Y` must be the same as the data type of `Mdl.ClassNames`. (The software treats string arrays as cell arrays of character vectors.)
- The distinct classes in `Y` must be a subset of `Mdl.ClassNames`.
- If `Y` is a character array, then each element must correspond to one row of the array.
- The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `margin` must also be in a table.

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

The response variable must be a categorical, character, or string array; a logical or numeric vector; or a cell array of character vectors. If the response variable is a character array, then each element must correspond to one row of the array.

Data Types: `char` | `string`

Output Arguments

`m` — Classification margins

numeric column vector

Classification margins on page 33-6755, returned as an *n*-by-1 numeric column vector, where *n* is the number of observations in `X`.

More About

Classification Margin

The classification margin for binary classification is, for each observation, the difference between the classification score for the true class and the classification score for the false class.

The software defines the classification margin for binary classification as

$$m = 2yf(x).$$

x is an observation. If the true label of x is the positive class, then y is 1, and –1 otherwise. $f(x)$ is the positive-class classification score for the observation x. The classification margin is commonly defined as $m = yf(x)$.

If the margins are on the same scale, then they serve as a classification confidence measure. Among multiple classifiers, those that yield greater margins are better.

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `margin` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`ClassificationKernel` | `edge` | `fitckernel` | `predict`

Introduced in R2017b

predict

Predict labels for Gaussian kernel classification model

Syntax

```
Label = predict(Mdl,X)
[Label,Score] = predict(Mdl,X)
```

Description

`Label = predict(Mdl,X)` returns a vector of predicted class labels for the predictor data in the matrix or table `X`, based on the binary Gaussian kernel classification model `Mdl`.

`[Label,Score] = predict(Mdl,X)` also returns classification scores on page 33-6763 for both classes.

Examples

Predict Training Set Labels

Predict the training set labels using a binary kernel classification model, and display the confusion matrix for the resulting classification.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Train a binary kernel classification model that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

```
rng('default') % For reproducibility
Mdl = fitckernel(X,Y);
```

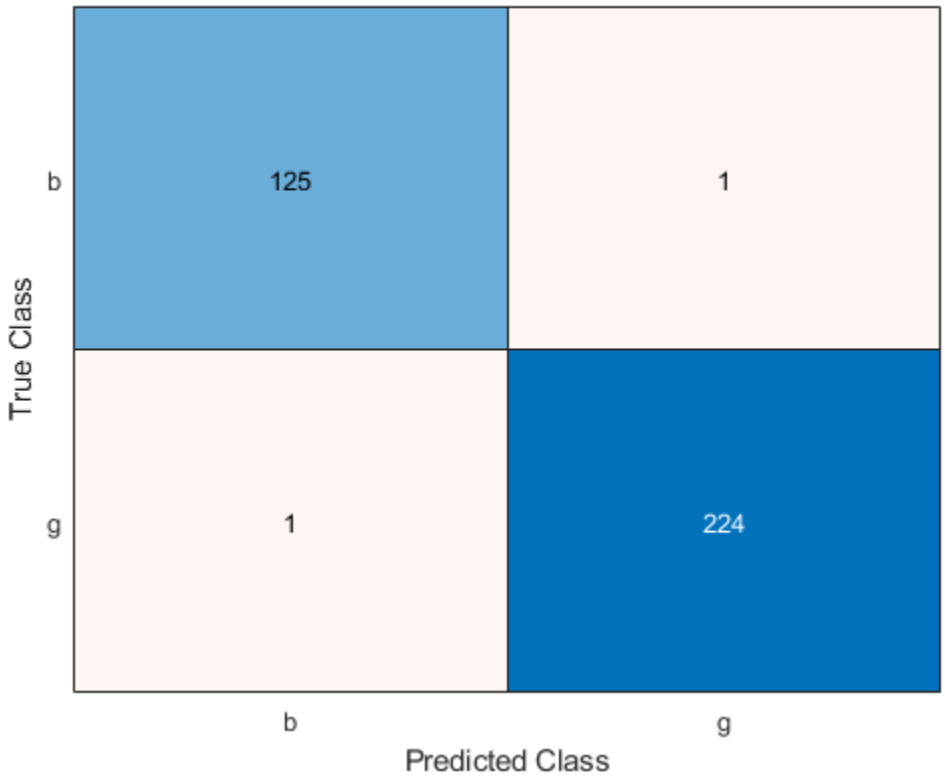
`Mdl` is a `ClassificationKernel` model.

Predict the training set, or resubstitution, labels.

```
label = predict(Mdl,X);
```

Construct a confusion matrix.

```
ConfusionTrain = confusionchart(Y,label);
```



The model misclassifies one radar return for each class.

Predict Test Set Labels

Predict the test set labels using a binary kernel classification model, and display the confusion matrix for the resulting classification.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 15% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.15);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model using the training set. A good practice is to define the class order.

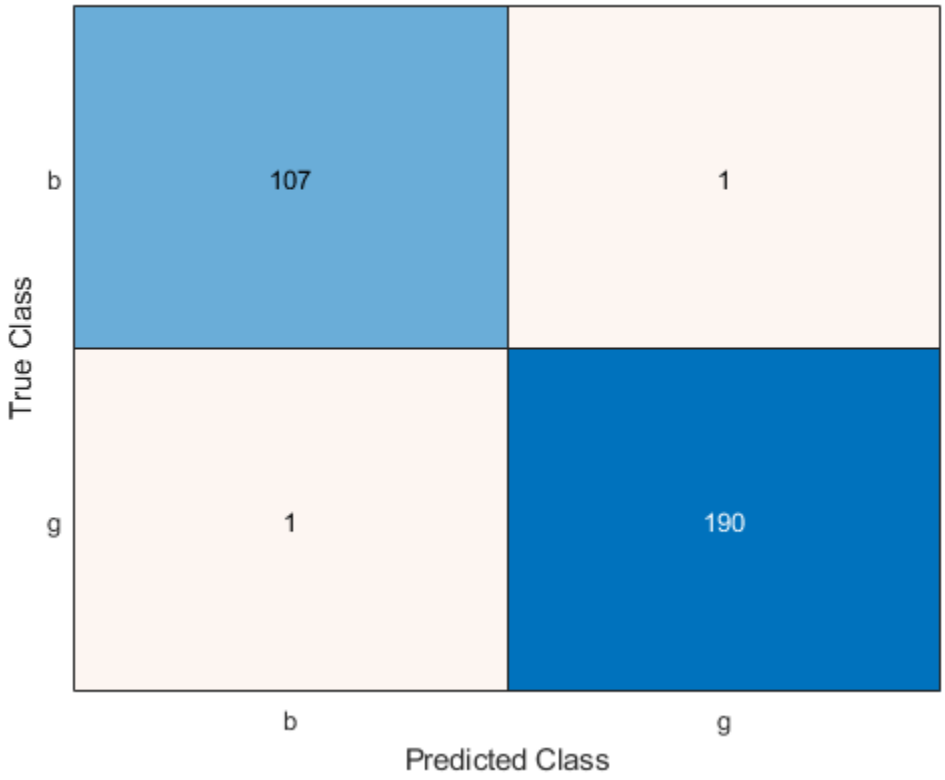
```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds),'ClassNames',{'b','g'});
```

Predict the training-set labels and the test set labels.

```
labelTrain = predict(Mdl,X(trainingInds,:));
labelTest = predict(Mdl,X(testInds,:));
```

Construct a confusion matrix for the training set.

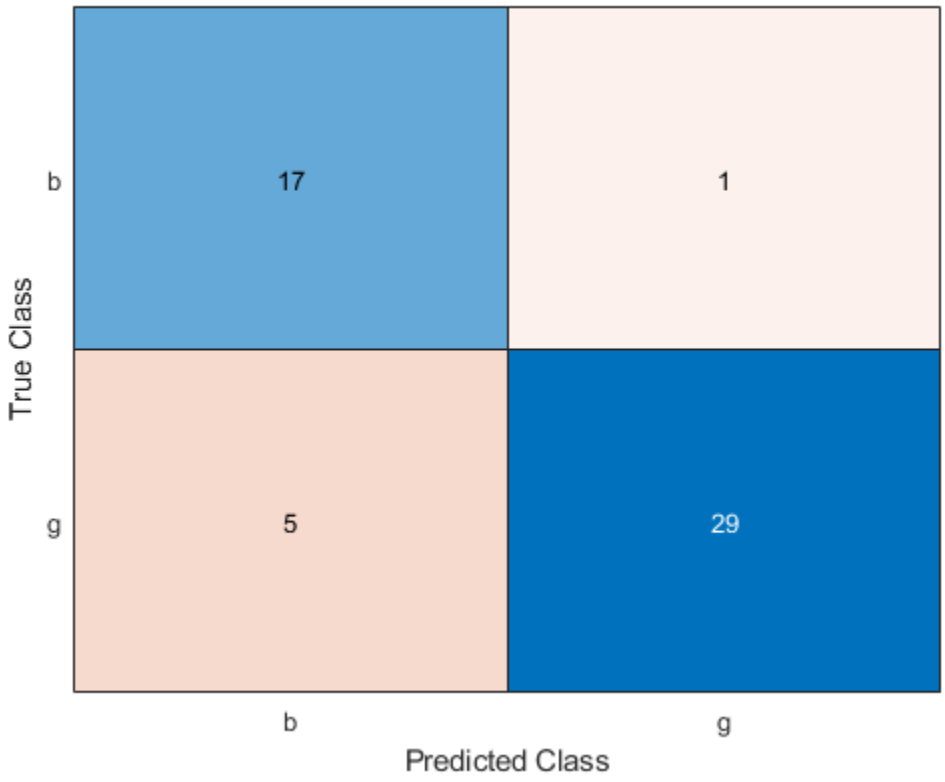
```
ConfusionTrain = confusionchart(Y(trainingInds),labelTrain);
```



The model misclassifies only one radar return for each class.

Construct a confusion matrix for the test set.

```
ConfusionTest = confusionchart(Y(testInds),labelTest);
```



The model misclassifies one bad radar return as being a good return, and five good radar returns as being bad returns.

Estimate Posterior Class Probabilities

Estimate posterior class probabilities for a test set, and determine the quality of the model by plotting a receiver operating characteristic (ROC) curve. Kernel classification models return posterior probabilities for logistic regression learners only.

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 30% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.30);
trainingInds = training(Partition); % Indices for the training set
testInds = test(Partition); % Indices for the test set
```

Train a binary kernel classification model. Fit logistic regression learners.

```
Mdl = fitckernel(X(trainingInds,:),Y(trainingInds), ...
    'ClassNames',{'b','g'},'Learner','logistic');
```

Predict the posterior class probabilities for the test set.

```
[~,posterior] = predict(Mdl,X(testInds,:));
```

Because `Mdl` has one regularization strength, the output `posterior` is a matrix with two columns and rows equal to the number of test-set observations. Column `i` contains posterior probabilities of `Mdl.ClassNames(i)` given a particular observation.

Obtain false and true positive rates, and estimate the area under the curve (AUC). Specify that the second class is the positive class.

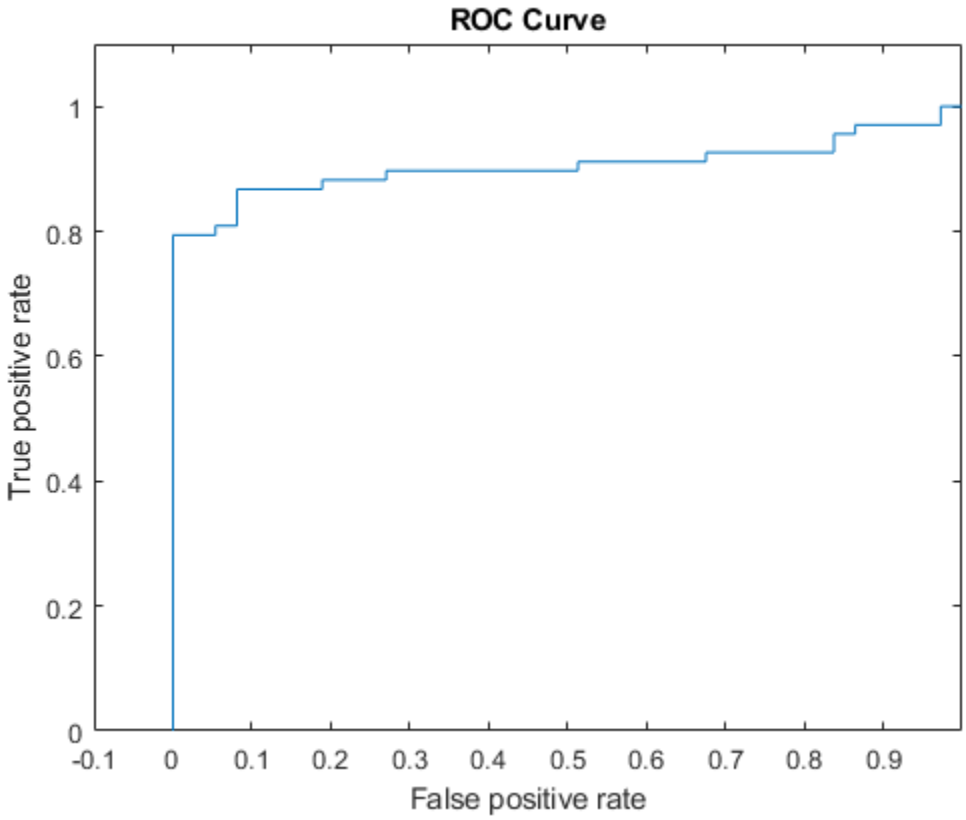
```
[fpr,tpr,~,auc] = perfcurve(Y(testInds),posterior(:,2),Mdl.ClassNames(2));
auc
```

```
auc = 0.9042
```

The AUC is close to `1`, which indicates that the model predicts labels well.

Plot an ROC curve.

```
figure;
plot(fpr,tpr)
h = gca;
h.XLim(1) = -0.1;
h.YLim(2) = 1.1;
xlabel('False positive rate')
ylabel('True positive rate')
title('ROC Curve')
```



Input Arguments

`Mdl` — Binary kernel classification model

`ClassificationKernel` model object

Binary kernel classification model, specified as a `ClassificationKernel` model object. You can create a `ClassificationKernel` model object using `fitckernel`.

`X` — Predictor data to be classified

numeric matrix | table

Predictor data to be classified, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`) and `Tbl` contains all numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors by using the `CategoricalPredictors` name-value pair argument of `fitckernel`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitckernel`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Data Types: `table` | `double` | `single`

Output Arguments

`Label` — Predicted class labels

categorical array | character array | logical matrix | numeric matrix | cell array of character vectors

Predicted class labels, returned as a categorical or character array, logical or numeric matrix, or cell array of character vectors.

`Label` has *n* rows, where *n* is the number of observations in `X`, and has the same data type as the observed class labels (`Y`) used to train `Mdl`. (The software treats string arrays as cell arrays of character vectors.)

`predict` classifies observations into the class yielding the highest score.

`Score` — Classification scores
numeric array

Classification scores on page 33-6763, returned as an *n*-by-2 numeric array, where *n* is the number of observations in `X`. `Score(i,j)` is the score for classifying observation `i` into class `j`. `Mdl.ClassNames` stores the order of the classes.

If `Mdl.Learner` is `'logistic'`, then classification scores are posterior probabilities.

More About

Classification Score

For kernel classification models, the raw classification score for classifying the observation x, a row vector, into the positive class is defined by

$$f(x) = T(x)\beta + b\,.$$

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β is the estimated column vector of coefficients.
- b is the estimated scalar bias.

The raw classification score for classifying x into the negative class is $-f(x)$. The software classifies observations into the class that yields a positive score.

If the kernel classification model consists of logistic regression learners, then the software applies the `'logit'` score transformation to the raw classification scores (see `ScoreTransform`).

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `predict` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also
`ClassificationKernel` | `confusionchart` | `fitckernel` | `perfcurve` | `resume`

Introduced in R2017b

resume

Resume training of Gaussian kernel classification model

Syntax

```
UpdatedMdl = resume(Mdl,X,Y)

UpdatedMdl = resume(Mdl,Tbl,ResponseVarName)
UpdatedMdl = resume(Mdl,Tbl,Y)

UpdatedMdl = resume(___,Name,Value)
[UpdatedMdl,FitInfo] = resume(___)
```

Description

`UpdatedMdl = resume(Mdl,X,Y)` continues training with the same options used to train `Mdl`, including the training data (predictor data in `X` and class labels in `Y`) and the feature expansion. The training starts at the current estimated parameters in `Mdl`. The function returns a new binary Gaussian kernel classification model `UpdatedMdl`.

`UpdatedMdl = resume(Mdl,Tbl,ResponseVarName)` continues training with the predictor data in `Tbl` and the true class labels in `Tbl.ResponseVarName`.

`UpdatedMdl = resume(Mdl,Tbl,Y)` continues training with the predictor data in table `Tbl` and the true class labels in `Y`.

`UpdatedMdl = resume(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can modify convergence control options, such as convergence tolerances and the maximum number of additional optimization iterations.

`[UpdatedMdl,FitInfo] = resume(___)` also returns the fit information in the structure array `FitInfo`.

Examples

Predict Class Labels and Resume Training

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 20% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.20);
trainingInds = training(Partition); % Indices for the training set
XTrain = X(trainingInds,:);
YTrain = Y(trainingInds);
```

```
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a binary kernel classification model that identifies whether the radar return is bad (`'b'`) or good (`'g'`).

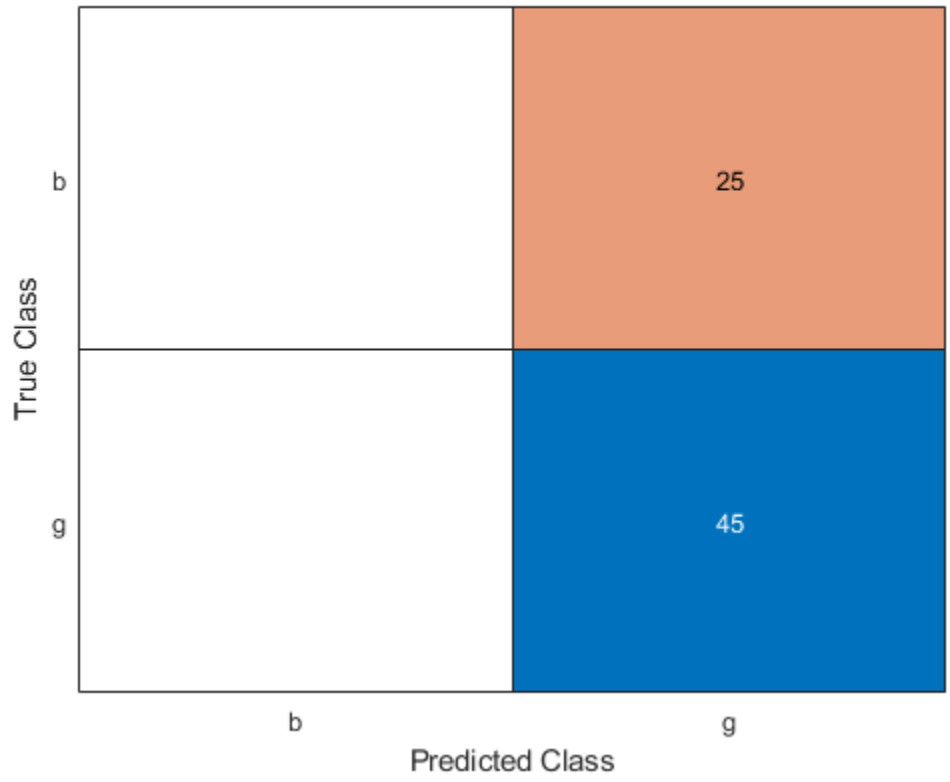
```
Mdl = fitckernel(XTrain,YTrain,'IterationLimit',5,'Verbose',1);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective    |     Step      |    Gradient    |    Relative
|        |        |               |                |               |   magnitude    | change in Beta
|=============================================================================================================
|  LBFGS |      1 |             0 |  1.000000e+00  |  0.000000e+00 |  2.811388e-01  |
|  LBFGS |      1 |             1 |  7.585395e-01  |  4.000000e+00 |  3.594306e-01  |   1.000000e+00
|  LBFGS |      1 |             2 |  7.160994e-01  |  1.000000e+00 |  2.028470e-01  |   6.923988e-01
|  LBFGS |      1 |             3 |  6.825272e-01  |  1.000000e+00 |  2.846975e-02  |   2.388909e-01
|  LBFGS |      1 |             4 |  6.699435e-01  |  1.000000e+00 |  1.779359e-02  |   1.325304e-01
|  LBFGS |      1 |             5 |  6.535619e-01  |  1.000000e+00 |  2.669039e-01  |   4.112952e-01
|=============================================================================================================
```

`Mdl` is a `ClassificationKernel` model.

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set.

```
label = predict(Mdl,XTest);
ConfusionTest = confusionchart(YTest,label);
```



```
L = loss(Mdl,XTest,YTest)
```

```
L = 0.3594
```

`Mdl` misclassifies all bad radar returns as good returns.

Continue training by using `resume`. This function continues training with the same options used for training `Mdl`.

```
UpdatedMdl = resume(Mdl,XTrain,YTrain);
```

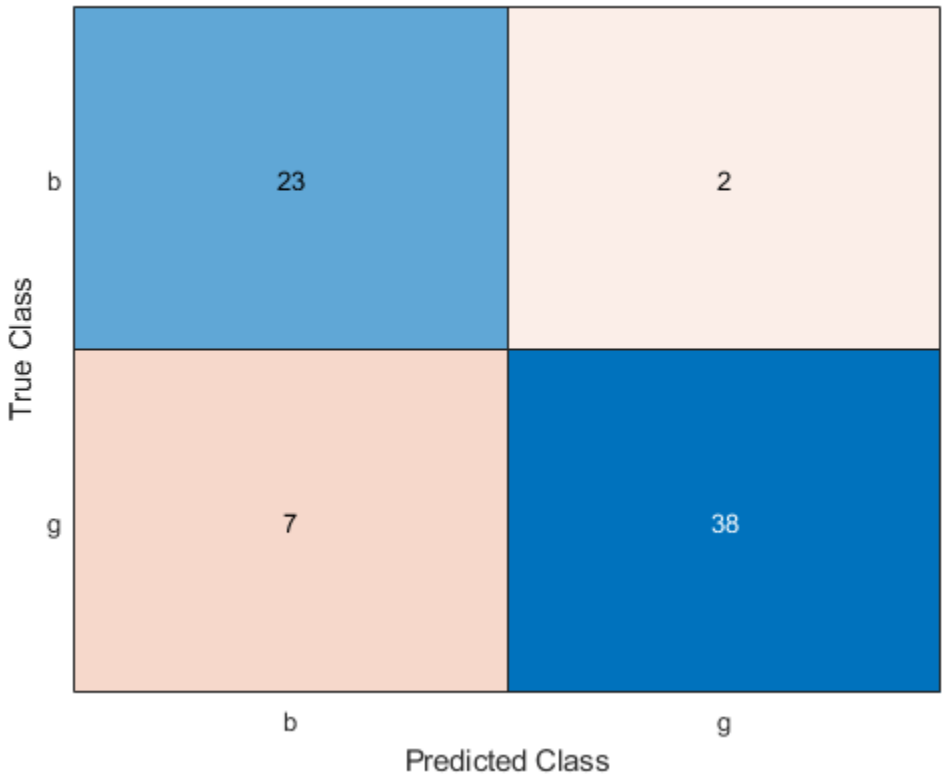
```
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient   |    Relative
|        |        |               |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |             0 |  6.535619e-01 |  0.000000e+00 |  2.669039e-01 |
|  LBFGS |      1 |             1 |  6.132547e-01 |  1.000000e+00 |  6.355537e-03 |   1.522092e-01
|  LBFGS |      1 |             2 |  5.938316e-01 |  4.000000e+00 |  3.202847e-02 |   1.498036e-01
|  LBFGS |      1 |             3 |  4.169274e-01 |  1.000000e+00 |  1.530249e-01 |   7.234253e-01
|  LBFGS |      1 |             4 |  3.679212e-01 |  5.000000e-01 |  2.740214e-01 |   2.495886e-01
|  LBFGS |      1 |             5 |  3.332261e-01 |  1.000000e+00 |  1.423488e-02 |   9.558680e-02
|  LBFGS |      1 |             6 |  3.235335e-01 |  1.000000e+00 |  7.117438e-03 |   7.137260e-02
|  LBFGS |      1 |             7 |  3.112331e-01 |  1.000000e+00 |  6.049822e-02 |   1.252157e-01
|  LBFGS |      1 |             8 |  2.972144e-01 |  1.000000e+00 |  7.117438e-03 |   5.796240e-02
|  LBFGS |      1 |             9 |  2.837450e-01 |  1.000000e+00 |  8.185053e-02 |   1.484733e-01
|  LBFGS |      1 |            10 |  2.797642e-01 |  1.000000e+00 |  3.558719e-02 |   5.856842e-02
|  LBFGS |      1 |            11 |  2.771280e-01 |  1.000000e+00 |  2.846975e-02 |   2.349433e-02
|  LBFGS |      1 |            12 |  2.741570e-01 |  1.000000e+00 |  3.914591e-02 |   3.113194e-02
|  LBFGS |      1 |            13 |  2.725701e-01 |  5.000000e-01 |  1.067616e-01 |   8.729821e-02
|  LBFGS |      1 |            14 |  2.667147e-01 |  1.000000e+00 |  3.914591e-02 |   3.491723e-02
|  LBFGS |      1 |            15 |  2.621152e-01 |  1.000000e+00 |  7.117438e-03 |   5.104726e-02
|  LBFGS |      1 |            16 |  2.601652e-01 |  1.000000e+00 |  3.558719e-02 |   3.764904e-02
|  LBFGS |      1 |            17 |  2.589052e-01 |  1.000000e+00 |  3.202847e-02 |   3.655744e-02
|  LBFGS |      1 |            18 |  2.583185e-01 |  1.000000e+00 |  7.117438e-03 |   6.490571e-02
|  LBFGS |      1 |            19 |  2.556482e-01 |  1.000000e+00 |  9.252669e-02 |   4.601390e-02
|  LBFGS |      1 |            20 |  2.542643e-01 |  1.000000e+00 |  7.117438e-02 |   4.141838e-02
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient   |    Relative
|        |        |               |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            21 |  2.532117e-01 |  1.000000e+00 |  1.067616e-02 |   1.661720e-02
|  LBFGS |      1 |            22 |  2.529890e-01 |  1.000000e+00 |  2.135231e-02 |   1.231678e-02
|  LBFGS |      1 |            23 |  2.523232e-01 |  1.000000e+00 |  3.202847e-02 |   1.958586e-02
|  LBFGS |      1 |            24 |  2.506736e-01 |  1.000000e+00 |  1.779359e-02 |   2.474613e-02
|  LBFGS |      1 |            25 |  2.501995e-01 |  1.000000e+00 |  1.779359e-02 |   2.514352e-02
|  LBFGS |      1 |            26 |  2.488242e-01 |  1.000000e+00 |  3.558719e-03 |   1.531810e-02
|  LBFGS |      1 |            27 |  2.485295e-01 |  5.000000e-01 |  3.202847e-02 |   1.229760e-02
|  LBFGS |      1 |            28 |  2.482244e-01 |  1.000000e+00 |  4.270463e-02 |   8.970983e-03
|  LBFGS |      1 |            29 |  2.479714e-01 |  1.000000e+00 |  3.558719e-03 |   7.393900e-03
|  LBFGS |      1 |            30 |  2.477316e-01 |  1.000000e+00 |  3.202847e-02 |   3.268087e-03
|  LBFGS |      1 |            31 |  2.476178e-01 |  2.500000e-01 |  3.202847e-02 |   5.445890e-03
|  LBFGS |      1 |            32 |  2.474874e-01 |  1.000000e+00 |  1.779359e-02 |   3.535903e-03
|  LBFGS |      1 |            33 |  2.473980e-01 |  1.000000e+00 |  7.117438e-03 |   2.821725e-03
|  LBFGS |      1 |            34 |  2.472935e-01 |  1.000000e+00 |  3.558719e-03 |   2.699880e-03
|  LBFGS |      1 |            35 |  2.471418e-01 |  1.000000e+00 |  3.558719e-03 |   1.242523e-02
|  LBFGS |      1 |            36 |  2.469862e-01 |  1.000000e+00 |  2.846975e-02 |   7.895605e-03
|  LBFGS |      1 |            37 |  2.469598e-01 |  1.000000e+00 |  2.135231e-02 |   6.657676e-03
|  LBFGS |      1 |            38 |  2.466941e-01 |  1.000000e+00 |  3.558719e-02 |   4.654690e-03
```

```
|  LBFGS |      1 |           39 |  2.466660e-01 |  5.000000e-01 |  1.423488e-02 |   2.885769e-03
|  LBFGS |      1 |           40 |  2.465605e-01 |  1.000000e+00 |  3.558719e-03 |   4.562565e-03
|========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |      Step     |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|========================================================================================================
|  LBFGS |      1 |           41 |  2.465362e-01 |  1.000000e+00 |  1.423488e-02 |   5.652180e-03
|  LBFGS |      1 |           42 |  2.463528e-01 |  1.000000e+00 |  3.558719e-03 |   2.389759e-03
|  LBFGS |      1 |           43 |  2.463207e-01 |  1.000000e+00 |  1.511170e-03 |   3.738286e-03
|  LBFGS |      1 |           44 |  2.462585e-01 |  5.000000e-01 |  7.117438e-02 |   2.321693e-03
|  LBFGS |      1 |           45 |  2.461742e-01 |  1.000000e+00 |  7.117438e-03 |   2.599725e-03
|  LBFGS |      1 |           46 |  2.461434e-01 |  1.000000e+00 |  3.202847e-02 |   3.186923e-03
|  LBFGS |      1 |           47 |  2.461115e-01 |  1.000000e+00 |  7.117438e-03 |   1.530711e-03
|  LBFGS |      1 |           48 |  2.460814e-01 |  1.000000e+00 |  1.067616e-02 |   1.811714e-03
|  LBFGS |      1 |           49 |  2.460533e-01 |  5.000000e-01 |  1.423488e-02 |   1.012252e-03
|  LBFGS |      1 |           50 |  2.460111e-01 |  1.000000e+00 |  1.423488e-02 |   4.166762e-03
|  LBFGS |      1 |           51 |  2.459414e-01 |  1.000000e+00 |  1.067616e-02 |   3.271946e-03
|  LBFGS |      1 |           52 |  2.458809e-01 |  1.000000e+00 |  1.423488e-02 |   1.846440e-03
|  LBFGS |      1 |           53 |  2.458479e-01 |  1.000000e+00 |  1.067616e-02 |   1.180871e-03
|  LBFGS |      1 |           54 |  2.458146e-01 |  1.000000e+00 |  1.455008e-03 |   1.422954e-03
|  LBFGS |      1 |           55 |  2.457878e-01 |  1.000000e+00 |  7.117438e-03 |   1.880892e-03
|  LBFGS |      1 |           56 |  2.457519e-01 |  1.000000e+00 |  2.491103e-02 |   1.074764e-03
|  LBFGS |      1 |           57 |  2.457420e-01 |  1.000000e+00 |  7.473310e-02 |   9.511878e-04
|  LBFGS |      1 |           58 |  2.457212e-01 |  1.000000e+00 |  3.558719e-03 |   3.718564e-04
|  LBFGS |      1 |           59 |  2.457089e-01 |  1.000000e+00 |  4.270463e-02 |   6.237270e-04
|  LBFGS |      1 |           60 |  2.457047e-01 |  5.000000e-01 |  1.423488e-02 |   3.647573e-04
|========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |      Step     |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|========================================================================================================
|  LBFGS |      1 |           61 |  2.456991e-01 |  1.000000e+00 |  1.423488e-02 |   5.666884e-04
|  LBFGS |      1 |           62 |  2.456898e-01 |  1.000000e+00 |  1.779359e-02 |   4.697056e-04
|  LBFGS |      1 |           63 |  2.456792e-01 |  1.000000e+00 |  1.779359e-02 |   5.984927e-04
|  LBFGS |      1 |           64 |  2.456603e-01 |  1.000000e+00 |  1.403782e-03 |   5.414985e-04
|  LBFGS |      1 |           65 |  2.456482e-01 |  1.000000e+00 |  3.558719e-03 |   6.506293e-04
|  LBFGS |      1 |           66 |  2.456358e-01 |  1.000000e+00 |  1.476262e-03 |   1.284139e-03
|  LBFGS |      1 |           67 |  2.456124e-01 |  1.000000e+00 |  3.558719e-03 |   8.636596e-04
|  LBFGS |      1 |           68 |  2.455980e-01 |  1.000000e+00 |  1.067616e-02 |   9.861527e-04
|  LBFGS |      1 |           69 |  2.455780e-01 |  1.000000e+00 |  1.067616e-02 |   5.102487e-04
|  LBFGS |      1 |           70 |  2.455633e-01 |  1.000000e+00 |  3.558719e-03 |   1.228077e-03
|  LBFGS |      1 |           71 |  2.455449e-01 |  1.000000e+00 |  1.423488e-02 |   7.864590e-04
|  LBFGS |      1 |           72 |  2.455261e-01 |  1.000000e+00 |  3.558719e-02 |   1.090815e-03
|  LBFGS |      1 |           73 |  2.455142e-01 |  1.000000e+00 |  1.067616e-02 |   1.701506e-03
|  LBFGS |      1 |           74 |  2.455075e-01 |  1.000000e+00 |  1.779359e-02 |   1.504577e-03
|  LBFGS |      1 |           75 |  2.455008e-01 |  1.000000e+00 |  3.914591e-02 |   1.144021e-03
|  LBFGS |      1 |           76 |  2.454943e-01 |  1.000000e+00 |  2.491103e-02 |   3.015254e-04
|  LBFGS |      1 |           77 |  2.454918e-01 |  5.000000e-01 |  3.202847e-02 |   9.837523e-04
|  LBFGS |      1 |           78 |  2.454870e-01 |  1.000000e+00 |  1.779359e-02 |   4.328953e-04
|  LBFGS |      1 |           79 |  2.454865e-01 |  5.000000e-01 |  3.558719e-03 |   7.126815e-04
|  LBFGS |      1 |           80 |  2.454775e-01 |  1.000000e+00 |  5.693950e-02 |   8.992562e-04
|========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |      Step     |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|========================================================================================================
|  LBFGS |      1 |           81 |  2.454686e-01 |  1.000000e+00 |  1.183730e-03 |   1.590246e-04
|  LBFGS |      1 |           82 |  2.454612e-01 |  1.000000e+00 |  2.135231e-02 |   1.389570e-04
|  LBFGS |      1 |           83 |  2.454506e-01 |  1.000000e+00 |  3.558719e-03 |   6.162089e-04
|  LBFGS |      1 |           84 |  2.454436e-01 |  1.000000e+00 |  1.423488e-02 |   1.877414e-03
```

```
|  LBFGS |      1 |           85 |  2.454378e-01 |  1.000000e+00 |  1.423488e-02 |   3.370852e-04
|  LBFGS |      1 |           86 |  2.454249e-01 |  1.000000e+00 |  1.423488e-02 |   8.133615e-04
|  LBFGS |      1 |           87 |  2.454101e-01 |  1.000000e+00 |  1.067616e-02 |   3.872088e-04
|  LBFGS |      1 |           88 |  2.453963e-01 |  1.000000e+00 |  1.779359e-02 |   5.670260e-04
|  LBFGS |      1 |           89 |  2.453866e-01 |  1.000000e+00 |  1.067616e-02 |   1.444984e-03
|  LBFGS |      1 |           90 |  2.453821e-01 |  1.000000e+00 |  7.117438e-03 |   2.457270e-03
|  LBFGS |      1 |           91 |  2.453790e-01 |  5.000000e-01 |  6.761566e-02 |   8.228766e-04
|  LBFGS |      1 |           92 |  2.453603e-01 |  1.000000e+00 |  2.135231e-02 |   1.084233e-03
|  LBFGS |      1 |           93 |  2.453540e-01 |  1.000000e+00 |  2.135231e-02 |   2.060005e-04
|  LBFGS |      1 |           94 |  2.453482e-01 |  1.000000e+00 |  1.779359e-02 |   1.560883e-04
|  LBFGS |      1 |           95 |  2.453461e-01 |  1.000000e+00 |  1.779359e-02 |   1.614693e-03
|  LBFGS |      1 |           96 |  2.453371e-01 |  1.000000e+00 |  3.558719e-02 |   2.145835e-04
|  LBFGS |      1 |           97 |  2.453305e-01 |  1.000000e+00 |  4.270463e-02 |   7.602088e-04
|  LBFGS |      1 |           98 |  2.453283e-01 |  2.500000e-01 |  2.135231e-02 |   3.422253e-04
|  LBFGS |      1 |           99 |  2.453246e-01 |  1.000000e+00 |  3.558719e-03 |   3.872561e-04
|  LBFGS |      1 |          100 |  2.453214e-01 |  1.000000e+00 |  3.202847e-02 |   1.732237e-04
|========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|========================================================================================================
|  LBFGS |      1 |          101 |  2.453168e-01 |  1.000000e+00 |  1.067616e-02 |   3.065286e-04
|  LBFGS |      1 |          102 |  2.453155e-01 |  5.000000e-01 |  4.626335e-02 |   3.402368e-04
|  LBFGS |      1 |          103 |  2.453136e-01 |  1.000000e+00 |  1.779359e-02 |   2.215029e-04
|  LBFGS |      1 |          104 |  2.453119e-01 |  1.000000e+00 |  3.202847e-02 |   4.142355e-04
|  LBFGS |      1 |          105 |  2.453093e-01 |  1.000000e+00 |  1.423488e-02 |   2.186007e-04
|  LBFGS |      1 |          106 |  2.453090e-01 |  1.000000e+00 |  2.846975e-02 |   1.338602e-03
|  LBFGS |      1 |          107 |  2.453048e-01 |  1.000000e+00 |  1.423488e-02 |   3.208296e-04
|  LBFGS |      1 |          108 |  2.453040e-01 |  1.000000e+00 |  3.558719e-02 |   1.294488e-03
|  LBFGS |      1 |          109 |  2.452977e-01 |  1.000000e+00 |  1.423488e-02 |   8.328380e-04
|  LBFGS |      1 |          110 |  2.452934e-01 |  1.000000e+00 |  2.135231e-02 |   5.149259e-04
|  LBFGS |      1 |          111 |  2.452886e-01 |  1.000000e+00 |  1.779359e-02 |   3.650664e-04
|  LBFGS |      1 |          112 |  2.452854e-01 |  1.000000e+00 |  1.067616e-02 |   2.633981e-04
|  LBFGS |      1 |          113 |  2.452836e-01 |  1.000000e+00 |  1.067616e-02 |   1.804300e-04
|  LBFGS |      1 |          114 |  2.452817e-01 |  1.000000e+00 |  7.117438e-03 |   4.251642e-04
|  LBFGS |      1 |          115 |  2.452741e-01 |  1.000000e+00 |  1.779359e-02 |   9.018440e-04
|  LBFGS |      1 |          116 |  2.452691e-01 |  1.000000e+00 |  2.135231e-02 |   9.941716e-05
|========================================================================================================
```

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set.

```
UpdatedLabel = predict(UpdatedMdl,XTest);
UpdatedConfusionTest = confusionchart(YTest,UpdatedLabel);
```




```
UpdatedL = loss(UpdatedMdl,XTest,YTest)
```

```
UpdatedL = 0.1284
```

The classification error decreases after `resume` updates the classification model with more iterations.

Resume Training with Modified Convergence Control Training Options

Load the `ionosphere` data set. This data set has 34 predictors and 351 binary responses for radar returns, either bad (`'b'`) or good (`'g'`).

```
load ionosphere
```

Partition the data set into training and test sets. Specify a 20% holdout sample for the test set.

```
rng('default') % For reproducibility
Partition = cvpartition(Y,'Holdout',0.20);
trainingInds = training(Partition); % Indices for the training set
XTrain = X(trainingInds,:);
YTrain = Y(trainingInds);
testInds = test(Partition); % Indices for the test set
XTest = X(testInds,:);
YTest = Y(testInds);
```

Train a binary kernel classification model with relaxed convergence control training options by using the name-value pair arguments `'BetaTolerance'` and `'GradientTolerance'`.

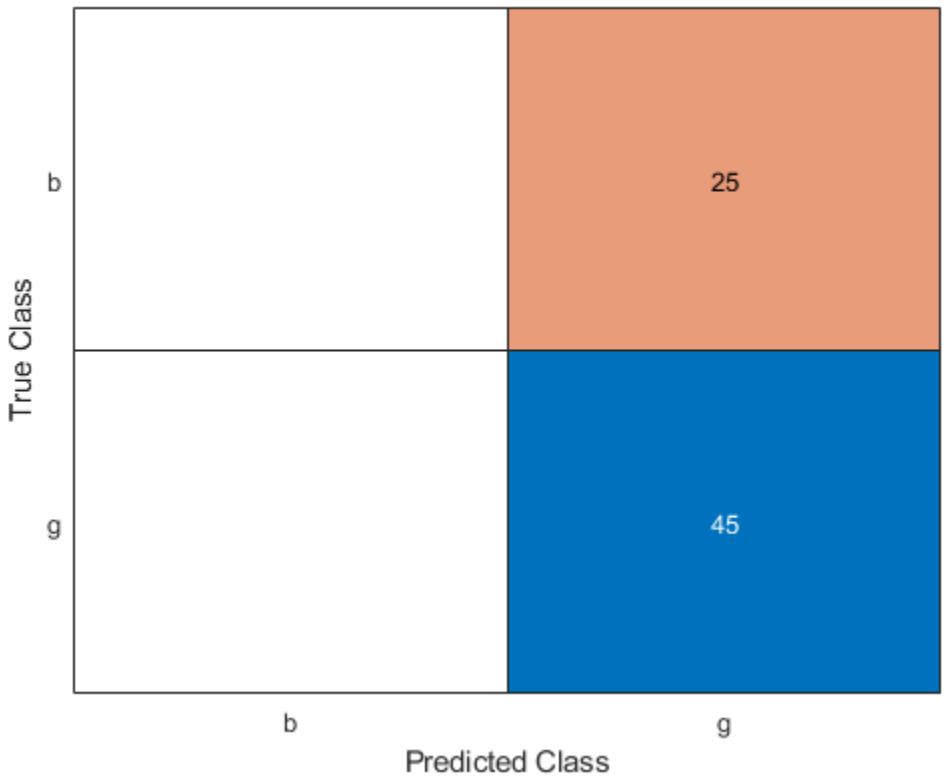
```
[Mdl,FitInfo] = fitckernel(XTrain,YTrain,'Verbose',1, ...
    'BetaTolerance',1e-1,'GradientTolerance',1e-1);
```

```
|=========================================================================================================
| Solver |  Pass  |   Iteration   |   Objective    |     Step       |    Gradient    |    Relative
|        |        |               |                |                |   magnitude    | change in Beta
|=========================================================================================================
|  LBFGS |      1 |             0 |  1.000000e+00  |  0.000000e+00  |  2.811388e-01  |
|  LBFGS |      1 |             1 |  7.585395e-01  |  4.000000e+00  |  3.594306e-01  |   1.000000e+00
|  LBFGS |      1 |             2 |  7.160994e-01  |  1.000000e+00  |  2.028470e-01  |   6.923988e-01
|  LBFGS |      1 |             3 |  6.825272e-01  |  1.000000e+00  |  2.846975e-02  |   2.388909e-01
|=========================================================================================================
```

`Mdl` is a `ClassificationKernel` model.

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set

```
label = predict(Mdl,XTest);
ConfusionTest = confusionchart(YTest,label);
```




```
L = loss(Mdl,XTest,YTest)
```

```
L = 0.3594
```

`Mdl` misclassifies all bad radar returns as good returns.

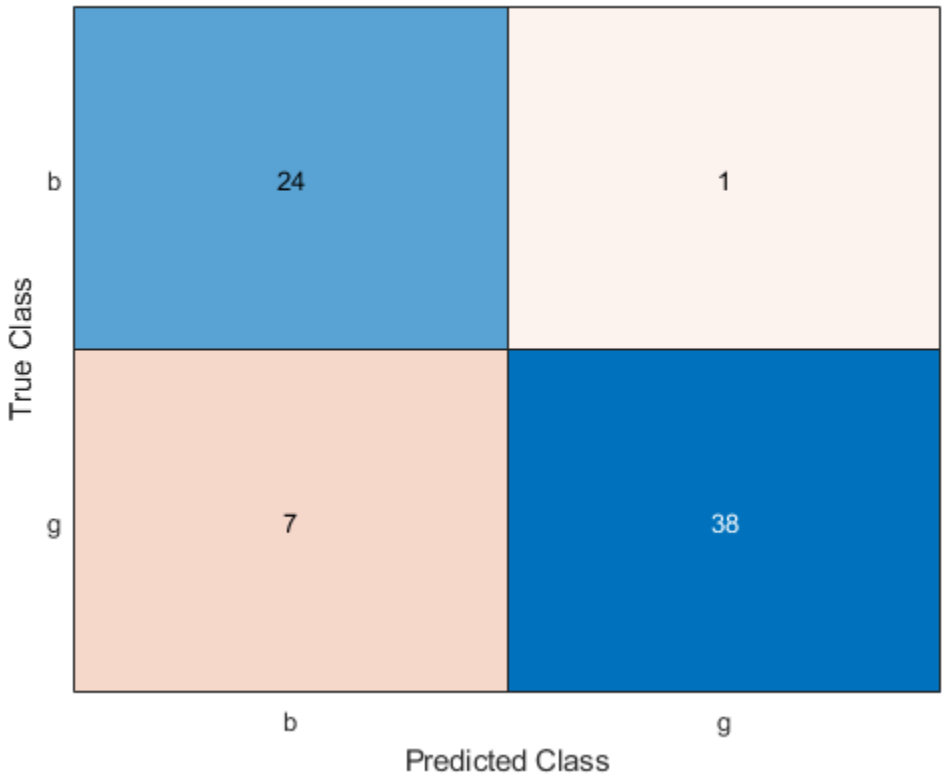
Continue training by using `resume` with modified convergence control training options.

```
[UpdatedMdl,UpdatedFitInfo] = resume(Mdl,XTrain,YTrain, ...
    'BetaTolerance',1e-2,'GradientTolerance',1e-2);
```

```
|=========================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient   |    Relative
|        |        |               |               |               |   magnitude   | change in Beta
|=========================================================================================================
|  LBFGS |      1 |             0 |  6.825272e-01 |  0.000000e+00 |  2.846975e-02 |
|  LBFGS |      1 |             1 |  6.692805e-01 |  2.000000e+00 |  2.846975e-02 |   1.389258e-01
|  LBFGS |      1 |             2 |  6.466824e-01 |  1.000000e+00 |  2.348754e-01 |   4.149425e-01
|  LBFGS |      1 |             3 |  5.441382e-01 |  2.000000e+00 |  1.743772e-01 |   5.344538e-01
|  LBFGS |      1 |             4 |  5.222333e-01 |  1.000000e+00 |  3.309609e-01 |   7.530878e-01
|  LBFGS |      1 |             5 |  3.776579e-01 |  1.000000e+00 |  1.103203e-01 |   6.532621e-01
|  LBFGS |      1 |             6 |  3.523520e-01 |  1.000000e+00 |  5.338078e-02 |   1.384232e-01
|  LBFGS |      1 |             7 |  3.422319e-01 |  5.000000e-01 |  3.202847e-02 |   9.703897e-02
|  LBFGS |      1 |             8 |  3.341895e-01 |  1.000000e+00 |  3.202847e-02 |   5.009485e-02
|  LBFGS |      1 |             9 |  3.199302e-01 |  1.000000e+00 |  4.982206e-02 |   8.038014e-02
|  LBFGS |      1 |            10 |  3.017904e-01 |  1.000000e+00 |  1.423488e-02 |   2.845012e-01
|  LBFGS |      1 |            11 |  2.853480e-01 |  1.000000e+00 |  3.558719e-02 |   9.799137e-02
|  LBFGS |      1 |            12 |  2.753979e-01 |  1.000000e+00 |  3.914591e-02 |   9.975305e-02
|  LBFGS |      1 |            13 |  2.647492e-01 |  1.000000e+00 |  3.914591e-02 |   9.713710e-02
|  LBFGS |      1 |            14 |  2.639242e-01 |  1.000000e+00 |  1.423488e-02 |   6.721803e-02
|  LBFGS |      1 |            15 |  2.617385e-01 |  1.000000e+00 |  1.779359e-02 |   2.625089e-02
|  LBFGS |      1 |            16 |  2.598600e-01 |  1.000000e+00 |  7.117438e-02 |   3.338724e-02
|  LBFGS |      1 |            17 |  2.594176e-01 |  1.000000e+00 |  1.067616e-02 |   2.441171e-02
|  LBFGS |      1 |            18 |  2.579350e-01 |  1.000000e+00 |  3.202847e-02 |   2.979246e-02
|  LBFGS |      1 |            19 |  2.570669e-01 |  1.000000e+00 |  1.779359e-02 |   4.432998e-02
|  LBFGS |      1 |            20 |  2.552954e-01 |  1.000000e+00 |  1.769940e-03 |   1.899895e-02
|=========================================================================================================
```

Predict the test-set labels, construct a confusion matrix for the test set, and estimate the classification error for the test set.

```
UpdatedLabel = predict(UpdatedMdl,XTest);
UpdatedConfusionTest = confusionchart(YTest,UpdatedLabel);
```




```
UpdatedL = loss(UpdatedMdl,XTest,YTest)
```

```
UpdatedL = 0.1140
```

The classification error decreases after `resume` updates the classification model with smaller convergence tolerances.

Display the outputs `FitInfo` and `UpdatedFitInfo`.

```
FitInfo
```

```
FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'hinge'
                  Lambda: 0.0036
           BetaTolerance: 0.1000
       GradientTolerance: 0.1000
          ObjectiveValue: 0.6825
       GradientMagnitude: 0.0285
    RelativeChangeInBeta: 0.2389
                 FitTime: 0.0545
                 History: [1x1 struct]
```

```
UpdatedFitInfo
```

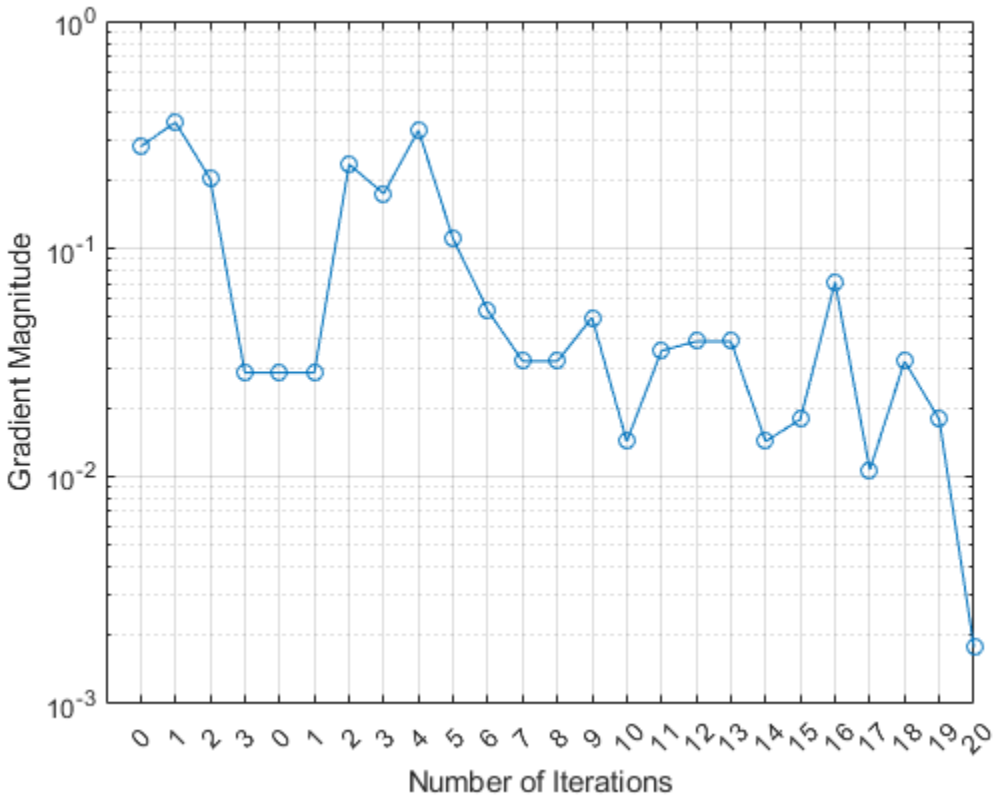
```
UpdatedFitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
```

```
          LossFunction: 'hinge'
                Lambda: 0.0036
         BetaTolerance: 0.0100
     GradientTolerance: 0.0100
        ObjectiveValue: 0.2553
     GradientMagnitude: 0.0018
 RelativeChangeInBeta: 0.0190
               FitTime: 0.1603
               History: [1x1 struct]
```

Both trainings terminate because the software satisfies the absolute gradient tolerance.

Plot the gradient magnitude versus the number of iterations by using `UpdatedFitInfo.History.GradientMagnitude`. Note that the `History` field of `UpdatedFitInfo` includes the information in the `History` field of `FitInfo`.

```
semilogy(UpdatedFitInfo.History.GradientMagnitude,'o-')
ax = gca;
ax.XTick = 1:25;
ax.XTickLabel = UpdatedFitInfo.History.IterationNumber;
grid on
xlabel('Number of Iterations')
ylabel('Gradient Magnitude')
```



The first training terminates after three iterations because the gradient magnitude becomes less than `1e-1`. The second training terminates after 20 iterations because the gradient magnitude becomes less than `1e-2`.

Input Arguments

`Mdl` — Binary kernel classification model

`ClassificationKernel` model object

Binary kernel classification model, specified as a `ClassificationKernel` model object. You can create a `ClassificationKernel` model object using `fitckernel`.

`X` — Predictor data used to train `Mdl`

n-by-*p* numeric matrix

Predictor data used to train `Mdl`, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors.

Data Types: `single` | `double`

`Y` — Class labels used to train `Mdl`

categorical array | character array | string array | logical vector | vector of numeric values | cell array of character vectors

Class labels used to train `Mdl`, specified as a categorical, character, or string array, logical or numeric vector, or cell array of character vectors.

Data Types: `categorical` | `char` | `string` | `logical` | `single` | `double` | `cell`

`Tbl` — Sample data used to train `Mdl`

table

Sample data used to train `Mdl`, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `Mdl` using sample data contained in a table, then the input data for `resume` must also be in a table.

`ResponseVarName` — Name of response variable used to train `Mdl`

name of variable in `Tbl`

Name of the response variable used to train `Mdl`, specified as the name of a variable in `Tbl`. The `ResponseVarName` value must match the name `Mdl.ResponseName`.

Data Types: `char` | `string`

Note `resume` should run only on the same training data and observation weights used to train `Mdl`. The `resume` function uses the same training options used to train `Mdl`, including feature expansion.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `UpdatedMdl = resume(Mdl,X,Y,'GradientTolerance',1e-5)` resumes training with the same options used to train `Mdl`, except the absolute gradient tolerance.

`Weights` — Observation weights used to train `Mdl`
numeric vector | name of variable in `Tbl`

Observation weights used to train `Mdl`, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If `Weights` is a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `Tbl`.
- If `Weights` is the name of a variable in `Tbl`, you must specify `Weights` as a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply weights, `resume` normalizes the weights to sum up to the value of the prior probability in the respective class.

Data Types: `double` | `single` | `char` | `string`

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`BetaTolerance` value used to train `Mdl` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

By default, the value is the same `BetaTolerance` value used to train `Mdl`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`GradientTolerance` value used to train `Mdl` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

By default, the value is the same `GradientTolerance` value used to train `Mdl`.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of additional optimization iterations
positive integer

Maximum number of additional optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The default value is 1000 if the transformed data fits in memory (`Mdl.ModelParameters.BlockSize`), which you specify by using the name-value pair argument when training `Mdl`. Otherwise, the default value is 100.

Note that the default value is not the value used to train `Mdl`.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Output Arguments

`UpdatedMdl` — Updated kernel classification model
`ClassificationKernel` model object

Updated kernel classification model, returned as a `ClassificationKernel` model object.

`FitInfo` — Optimization details
structure array

Optimization details, returned as a structure array including fields described in this table. The fields contain final values or name-value pair argument specifications.

Field	Description
`Solver`	Objective function minimization technique: `'LBFGS-fast'`, `'LBFGS-blockwise'`, or `'LBFGS-tall'`. For details, see “Algorithms” on page 33-6723 of `fitckernel`.
`LossFunction`	Loss function. Either `'hinge'` or `'logit'` depending on the type of linear classification model. See `Learner` of `fitckernel`.
`Lambda`	Regularization term strength. See `Lambda` of `fitckernel`.
`BetaTolerance`	Relative tolerance on the linear coefficients and the bias term. See `BetaTolerance`.
`GradientTolerance`	Absolute gradient tolerance. See `GradientTolerance`.
`ObjectiveValue`	Value of the objective function when optimization terminates. The classification loss plus the regularization term compose the objective function.
`GradientMagnitude`	Infinite norm of the gradient vector of the objective function when optimization terminates. See `GradientTolerance`.
`RelativeChangeInBeta`	Relative changes in the linear coefficients and the bias term when optimization terminates. See `BetaTolerance`.
`FitTime`	Elapsed, wall-clock time (in seconds) required to fit the model to the data.

Field	Description
`History`	History of optimization information. This field also includes the optimization information from training `Mdl`. This field is empty (`[]`) if you specify `'Verbose',0` when training `Mdl`. For details, see `Verbose` and "Algorithms" on page 33-6723 of `fitckernel`.

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.ObjectiveValue` in the Command Window.

A good practice is to examine `FitInfo` to assess whether convergence is satisfactory.

More About

Random Feature Expansion

Random feature expansion, such as Random Kitchen Sinks[1] and Fastfood[2], is a scheme to approximate Gaussian kernels of the kernel classification algorithm to use for big data in a computationally efficient way. Random feature expansion is more practical for big data applications that have large training sets, but can also be applied to smaller data sets that fit in memory.

The kernel classification algorithm searches for an optimal hyperplane that separates the data into two classes after mapping features into a high-dimensional space. Nonlinear features that are not linearly separable in a low-dimensional space can be separable in the expanded high-dimensional space. All the calculations for hyperplane classification use only dot products. You can obtain a nonlinear classification model by replacing the dot product x_1x_2' with the nonlinear kernel function $G(x_1,x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle$, where x_i is the ith observation (row vector) and $\varphi(x_i)$ is a transformation that maps x_i to a high-dimensional space (called the "kernel trick"). However, evaluating $G(x_1,x_2)$ (Gram matrix) for each pair of observations is computationally expensive for a large data set (large n).

The random feature expansion scheme finds a random transformation so that its dot product approximates the Gaussian kernel. That is,

$$G(x_1,x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle \approx T(x_1)T(x_2)',$$

where $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$). The Random Kitchen Sink scheme uses the random transformation

$$T(x) = m^{-1/2}\exp(iZx')',$$

where $Z \in \mathbb{R}^{m \times p}$ is a sample drawn from $N\left(0, \sigma^{-2}\right)$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage. The Fastfood scheme introduces another random basis V instead of Z using Hadamard matrices combined with Gaussian scaling matrices. This random basis reduces the computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

The `fitckernel` function uses the Fastfood scheme for random feature expansion and uses linear classification to train a Gaussian kernel classification model. Unlike solvers in the `fitcsvm` function, which require computation of the n-by-n Gram matrix, the solver in `fitckernel` only needs to form a matrix of size n-by-m, with m typically much less than n for big data.

References

[1] Rahimi, A., and B. Recht. “Random Features for Large-Scale Kernel Machines.” *Advances in Neural Information Processing Systems*. Vol. 20, 2008, pp. 1177–1184.

[2] Le, Q., T. Sarlós, and A. Smola. “Fastfood — Approximating Kernel Expansions in Loglinear Time.” *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, No. 3, 2013, pp. 244–252.

[3] Huang, P. S., H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. “Kernel methods match Deep Neural Networks on TIMIT.” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 205–209.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `resume` does not support tall `table` data.
- The default value for the `'IterationLimit'` name-value pair argument is relaxed to 20 when working with tall arrays.
- `resume` uses a block-wise strategy. For details, see “Algorithms” on page 33-6723 of `fitckernel`.

For more information, see “Tall Arrays”.

See Also

`ClassificationKernel` | `fitckernel` | `predict`

Introduced in R2017b

fitrkernel

Fit Gaussian kernel regression model using random feature expansion

Syntax

```
Mdl = fitrkernel(X,Y)

Mdl = fitrkernel(Tbl,ResponseVarName)
Mdl = fitrkernel(Tbl,formula)
Mdl = fitrkernel(Tbl,Y)

Mdl = fitrkernel(___,Name,Value)
[Mdl,FitInfo] = fitrkernel(___)
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrkernel(___)
```

Description

`fitrkernel` trains or cross-validates a Gaussian kernel regression model for nonlinear regression. `fitrkernel` is more practical to use for big data applications that have large training sets, but can also be applied to smaller data sets that fit in memory.

`fitrkernel` maps data in a low-dimensional space into a high-dimensional space, then fits a linear model in the high-dimensional space by minimizing the regularized objective function. Obtaining the linear model in the high-dimensional space is equivalent to applying the Gaussian kernel to the model in the low-dimensional space. Available linear regression models include regularized support vector machine (SVM) and least-squares regression models.

To train a nonlinear SVM regression model on in-memory data, see `fitrsvm`.

`Mdl = fitrkernel(X,Y)` returns a compact Gaussian kernel regression model trained using the predictor data in `X` and the corresponding responses in `Y`.

`Mdl = fitrkernel(Tbl,ResponseVarName)` returns a kernel regression model `Mdl` trained using the predictor variables contained in the table `Tbl` and the response values in `Tbl.ResponseVarName`.

`Mdl = fitrkernel(Tbl,formula)` returns a kernel regression model trained using the sample data in the table `Tbl`. The input argument `formula` is an explanatory model of the response and a subset of predictor variables in `Tbl` used to fit `Mdl`.

`Mdl = fitrkernel(Tbl,Y)` returns a kernel regression model using the predictor variables in the table `Tbl` and the response values in vector `Y`.

`Mdl = fitrkernel(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can implement least-squares regression, specify the number of dimension of the expanded space, or specify cross-validation options.

`[Mdl,FitInfo] = fitrkernel(___)` also returns the fit information in the structure array `FitInfo` using any of the input arguments in the previous syntaxes. You cannot request `FitInfo` for cross-validated models.

`[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrkernel( ___ )` also returns the hyperparameter optimization results when you optimize hyperparameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Examples

Train Gaussian Kernel Regression Model

Train a kernel regression model for a tall array by using SVM.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore that references the folder location with the data. The data can be contained in a single file, a collection of files, or an entire folder. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Select a subset of the variables to use. Create a tall table on top of the datastore.

```
varnames = {'ArrTime','DepTime','ActualElapsedTime'};
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames',varnames);
t = tall(ds);
```

Specify `DepTime` and `ArrTime` as the predictor variables (`X`) and `ActualElapsedTime` as the response variable (`Y`). Select the observations for which `ArrTime` is later than `DepTime`.

```
daytime = t.ArrTime>t.DepTime;
Y = t.ActualElapsedTime(daytime);     % Response data
X = t{daytime,{'DepTime' 'ArrTime'}}; % Predictor data
```

Standardize the predictor variables.

```
Z = zscore(X); % Standardize the data
```

Train a default Gaussian kernel regression model with the standardized predictors. Extract a fit summary to determine how well the optimization algorithm fits the model to the data.

```
[Mdl,FitInfo] = fitrkernel(Z,Y)
```

```
Found 6 chunks.
|=====================================================================================|
| Solver | Iteration  /  |   Objective   |   Gradient    | Beta relative |
|        | Data Pass     |               |   magnitude   |    change     |
|=====================================================================================|
|   INIT |     0 /     1 |  4.313465e+01 |  6.296907e-02 |           NaN |
|  LBFGS |     0 /     2 |  3.704335e+01 |  1.789316e-02 |  9.985854e-01 |
|  LBFGS |     1 /     3 |  3.703211e+01 |  2.880402e-02 |  1.044172e-03 |
|  LBFGS |     2 /     4 |  3.701616e+01 |  2.297788e-02 |  5.115891e-04 |
|  LBFGS |     2 /     5 |  3.700183e+01 |  1.750937e-02 |  1.023672e-03 |
|  LBFGS |     3 /     6 |  3.679055e+01 |  4.815047e-02 |  1.113182e-02 |
|  LBFGS |     4 /     7 |  3.637852e+01 |  1.058657e-01 |  2.994089e-02 |
|  LBFGS |     5 /     8 |  3.565372e+01 |  1.406536e-01 |  7.033477e-02 |
```

```
|  LBFGS |     6 /     9 |  3.478061e+01 |  1.479288e-01 |  1.185262e-01 |
|  LBFGS |     7 /    10 |  3.616955e+01 |  1.544917e-01 |  2.790848e-01 |
|  LBFGS |     7 /    11 |  3.459534e+01 |  1.212256e-01 |  1.229242e-01 |
|  LBFGS |     8 /    12 |  3.379859e+01 |  8.791025e-02 |  5.417481e-02 |
|  LBFGS |     9 /    13 |  3.339981e+01 |  3.077806e-02 |  4.638645e-02 |
|  LBFGS |    10 /    14 |  3.325224e+01 |  3.082755e-02 |  2.867793e-02 |
|  LBFGS |    11 /    15 |  3.320036e+01 |  4.168377e-02 |  9.376887e-03 |
|  LBFGS |    12 /    16 |  3.309321e+01 |  5.018195e-02 |  1.831484e-02 |
|  LBFGS |    13 /    17 |  3.288069e+01 |  4.506485e-02 |  3.732443e-02 |
|  LBFGS |    14 /    18 |  3.245691e+01 |  3.787163e-02 |  1.036929e-01 |
|  LBFGS |    15 /    19 |  3.210116e+01 |  2.418833e-02 |  1.190984e-01 |
|  LBFGS |    16 /    20 |  3.190585e+01 |  2.666398e-02 |  3.921991e-02 |
|=========================================================================|
| Solver | Iteration  /  |   Objective   |   Gradient    | Beta relative |
|        | Data Pass     |               |   magnitude   |    change     |
|=========================================================================|
|  LBFGS |    17 /    21 |  3.172622e+01 |  2.548259e-02 |  3.805655e-02 |
|  LBFGS |    18 /    22 |  3.154538e+01 |  1.280266e-02 |  4.363429e-02 |
|  LBFGS |    19 /    23 |  3.138533e+01 |  1.446779e-02 |  8.822868e-02 |
|  LBFGS |    20 /    24 |  3.283513e+01 |  2.218528e-01 |  1.318597e-01 |
|  LBFGS |    20 /    25 |  3.158782e+01 |  1.019184e-01 |  6.992082e-02 |
|  LBFGS |    20 /    26 |  3.136869e+01 |  4.678412e-02 |  3.603399e-02 |
|=========================================================================|

Mdl =
  RegressionKernel
            PredictorNames: {'x1'  'x2'}
              ResponseName: 'Y'
                   Learner: 'svm'
    NumExpansionDimensions: 64
               KernelScale: 1
                    Lambda: 8.5385e-06
             BoxConstraint: 1
                   Epsilon: 5.9303


  Properties, Methods


FitInfo = struct with fields:
                    Solver: 'LBFGS-tall'
              LossFunction: 'epsiloninsensitive'
                    Lambda: 8.5385e-06
           BetaTolerance: 1.0000e-03
       GradientTolerance: 1.0000e-05
          ObjectiveValue: 31.3687
       GradientMagnitude: 0.0468
    RelativeChangeInBeta: 0.0360
                 FitTime: 50.9380
                 History: [1x1 struct]
```

`Mdl` is a `RegressionKernel` model. To inspect the regression error, you can pass `Mdl` and the training data or new data to the `loss` function. Or, you can pass `Mdl` and new predictor data to the `predict` function to predict responses for new observations. You can also pass `Mdl` and the training data to the `resume` function to continue training.

`FitInfo` is a structure array containing optimization information. Use `FitInfo` to determine whether optimization termination measurements are satisfactory.

For improved accuracy, you can increase the maximum number of optimization iterations (`'IterationLimit'`) and decrease the tolerance values (`'BetaTolerance'` and `'GradientTolerance'`) by using the name-value pair arguments of `fitrkernel`. Doing so can improve measures like `ObjectiveValue` and `RelativeChangeInBeta` in `FitInfo`. You can also optimize model parameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Cross-Validate Kernel Regression Model

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:5);
Y = R(:,end);
```

Standardize the predictor variables.

```
Z = zscore(X);
```

Cross-validate a kernel regression model using 5-fold cross-validation.

```
Mdl = fitrkernel(Z,Y,'Kfold',5)
```

```
Mdl =
  RegressionPartitionedKernel
    CrossValidatedModel: 'Kernel'
           ResponseName: 'Y'
        NumObservations: 392
                  KFold: 5
              Partition: [1x1 cvpartition]
      ResponseTransform: 'none'


  Properties, Methods
```

```
numel(Mdl.Trained)
```

```
ans = 5
```

`Mdl` is a `RegressionPartitionedKernel` model. Because `fitrkernel` implements five-fold cross-validation, `Mdl` contains five `RegressionKernel` models that the software trains on training-fold (in-fold) observations.

Examine the cross-validation loss (mean squared error) for each fold.

```
kfoldLoss(Mdl,'mode','individual')
```

```
ans = 5×1

   13.0610
   14.0975
   24.0104
   21.1223
   24.3979
```

Optimize Kernel Regression

Optimize hyperparameters automatically using the `'OptimizeHyperparameters'` name-value pair argument.

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:5);
Y = R(:,end);
```

Standardize the predictor variables.

```
Z = zscore(X);
```

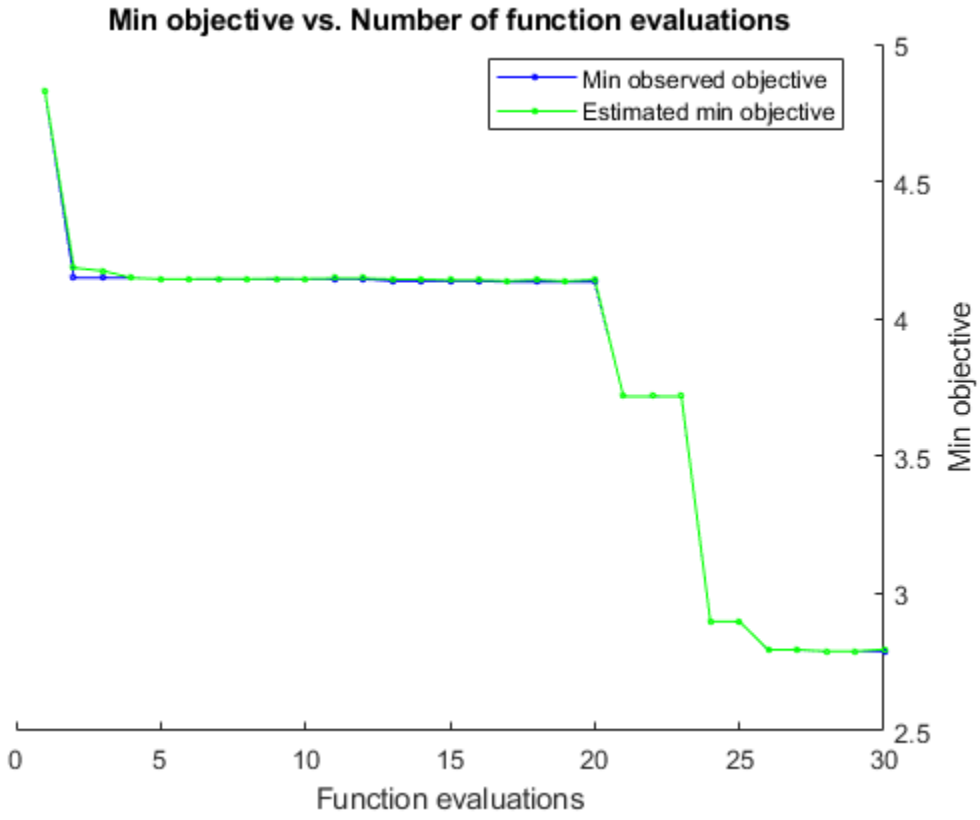
Find hyperparameters that minimize five-fold cross-validation loss by using automatic hyperparameter optimization. Specify `'OptimizeHyperparameters'` as `'auto'` so that `fitrkernel` finds the optimal values of the `'KernelScale'`, `'Lambda'`, and `'Epsilon'` name-value pair arguments. For reproducibility, set the random seed and use the `'expected-improvement-plus'` acquisition function.

```
rng('default')
[Mdl,FitInfo,HyperparameterOptimizationResults] = fitrkernel(Z,Y,'OptimizeHyperparameters','auto
    'HyperparameterOptimizationOptions',struct('AcquisitionFunctionName','expected-improvement-p
```

```
|=====================================================================================================
| Iter | Eval   | Objective:  | Objective   | BestSoFar   | BestSoFar   |  KernelScale |       La
|      | result | log(1+loss) | runtime     | (observed)  | (estim.)    |              |
|=====================================================================================================
|    1 | Best   |      4.8295 |       2.028 |      4.8295 |      4.8295 |     0.011518 |   6.8068
|    2 | Best   |      4.1488 |     0.33789 |      4.1488 |      4.1855 |       477.57 |     0.00
|    3 | Accept |      4.1521 |      0.2055 |      4.1488 |      4.1747 |    0.0080478 |    0.005
|    4 | Accept |      4.1506 |     0.31201 |      4.1488 |      4.1488 |      0.10935 |      0.1
```

5	Best	4.1446	0.33156	4.1446	4.1446	326.29	2
6	Accept	4.1521	0.29472	4.1446	4.1447	719.11	0.1
7	Accept	4.1501	0.41261	4.1446	4.1461	0.052426	2
8	Accept	4.1521	0.23016	4.1446	4.1447	990.71	0.01
9	Accept	4.1521	0.2329	4.1446	4.1465	415.85	0.05
10	Accept	4.1454	0.30527	4.1446	4.1455	972.49	1
11	Accept	4.1495	0.29501	4.1446	4.1473	121.79	1
12	Accept	4.1521	0.24626	4.1446	4.1474	985.81	0.8
13	Best	4.1374	0.26584	4.1374	4.1441	167.34	2
14	Accept	4.1434	0.25606	4.1374	4.1437	74.527	
15	Accept	4.1402	0.24839	4.1374	4.1407	877.17	2
16	Accept	4.1436	0.32898	4.1374	4.1412	0.0010354	0.01
17	Best	4.1346	0.39461	4.1346	4.1375	0.0010362	0.01
18	Accept	4.1521	0.30181	4.1346	4.1422	0.0010467	0.009
19	Accept	4.1508	0.38787	4.1346	4.1367	760.12	0.001
20	Accept	4.1435	0.62452	4.1346	4.143	0.020647	0.008

Iter	Eval result	Objective: log(1+loss)	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	KernelScale	La
21	Best	3.7172	0.45412	3.7172	3.7174	818.08	2.5529
22	Accept	4.1521	0.32339	3.7172	3.7177	0.006272	2.5598
23	Accept	4.0567	0.35643	3.7172	3.7176	940.43	2.6941
24	Best	2.8979	1.3783	2.8979	2.8979	37.141	2.5677
25	Accept	4.1521	0.27897	2.8979	2.898	13.817	2.5755
26	Best	2.795	1.1906	2.795	2.7953	20.022	2.6098
27	Accept	2.8284	1.1201	2.795	2.7956	17.252	2.7719
28	Best	2.7896	1.1718	2.7896	2.7898	11.432	7.621
29	Accept	2.822	2.8227	2.7896	2.7899	8.5133	2.5872
30	Accept	2.8061	0.9128	2.7896	2.7965	15.823	6.1956




```
_____________________________________________________________
Optimization completed.
MaxObjectiveEvaluations of 30 reached.
Total function evaluations: 30
Total elapsed time: 55.1082 seconds
Total objective function evaluation time: 18.0491

Best observed feasible point:
    KernelScale     Lambda      Epsilon
    ___________    _________    _______

      11.432       7.621e-06     2.094

Observed objective function value = 2.7896
Estimated objective function value = 2.7987
Function evaluation time = 1.1718

Best estimated feasible point (according to models):
    KernelScale      Lambda      Epsilon
    ___________    __________    _______

      15.823       6.1956e-06    2.0085

Estimated objective function value = 2.7965
Estimated function evaluation time = 1.2521

Mdl =
  RegressionKernel
```

```
              ResponseName: 'Y'
                   Learner: 'svm'
    NumExpansionDimensions: 256
               KernelScale: 15.8229
                    Lambda: 6.1956e-06
            BoxConstraint: 411.7488
                   Epsilon: 2.0085


  Properties, Methods


FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'epsiloninsensitive'
                  Lambda: 6.1956e-06
           BetaTolerance: 1.0000e-04
       GradientTolerance: 1.0000e-06
          ObjectiveValue: 1.3582
       GradientMagnitude: 0.0051
    RelativeChangeInBeta: 5.3944e-05
                 FitTime: 0.1737
                 History: []


HyperparameterOptimizationResults =
  BayesianOptimization with properties:

                     ObjectiveFcn: @createObjFcn/inMemoryObjFcn
              VariableDescriptions: [5x1 optimizableVariable]
                           Options: [1x1 struct]
                      MinObjective: 2.7896
                   XAtMinObjective: [1x3 table]
             MinEstimatedObjective: 2.7965
          XAtMinEstimatedObjective: [1x3 table]
           NumObjectiveEvaluations: 30
                  TotalElapsedTime: 55.1082
                         NextPoint: [1x3 table]
                            XTrace: [30x3 table]
                    ObjectiveTrace: [30x1 double]
                  ConstraintsTrace: []
                     UserDataTrace: {30x1 cell}
      ObjectiveEvaluationTimeTrace: [30x1 double]
                IterationTimeTrace: [30x1 double]
                        ErrorTrace: [30x1 double]
                  FeasibilityTrace: [30x1 logical]
       FeasibilityProbabilityTrace: [30x1 double]
               IndexOfMinimumTrace: [30x1 double]
             ObjectiveMinimumTrace: [30x1 double]
    EstimatedObjectiveMinimumTrace: [30x1 double]
```

For big data, the optimization procedure can take a long time. If the data set is too large to run the optimization procedure, you can try to optimize the parameters using only partial data. Use the `datasample` function and specify `'Replace','false'` to sample data without replacement.

Input Arguments

X — Predictor data

numeric matrix

Predictor data to which the regression model is fit, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictor variables.

The length of `Y` and the number of observations in `X` must be equal.

Data Types: `single` | `double`

Y — Response data

numeric vector

Response data, specified as an *n*-dimensional numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

Tbl — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain one additional column for the response variable. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

- If `Tbl` contains the response variable, and you want to use all remaining variables in `Tbl` as predictors, then specify the response variable by using `ResponseVarName`.
- If `Tbl` contains the response variable, and you want to use only a subset of the remaining variables in `Tbl` as predictors, then specify a formula by using `formula`.
- If `Tbl` does not contain the response variable, then specify a response variable by using `Y`. The length of the response variable and the number of rows in `Tbl` must be equal.

Data Types: `table`

ResponseVarName — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector.

You must specify `ResponseVarName` as a character vector or string scalar. For example, if `Tbl` stores the response variable `Y` as `Tbl.Y`, then specify it as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Y`, as predictors when training the model.

Data Types: `char` | `string`

formula — Explanatory model of response variable and subset of predictor variables

character vector | string scalar

Explanatory model of the response variable and a subset of the predictor variables, specified as a character vector or string scalar in the form `'Y~x1+x2+x3'`. In this form, `Y` represents the response variable, and `x1`, `x2`, and `x3` represent the predictor variables.

To specify a subset of variables in `Tbl` as predictors for training the model, use a formula. If you specify a formula, then the software does not use any variables in `Tbl` that do not appear in `formula`.

The variable names in the formula must be both variable names in `Tbl` (`Tbl.Properties.VariableNames`) and valid MATLAB identifiers. You can verify the variable names in `Tbl` by using the `isvarname` function. If the variable names are not valid, then you can convert them by using the `matlab.lang.makeValidName` function.

Data Types: `char` | `string`

Note The software treats `NaN`, empty character vector (`''`), empty string (`""`), `<missing>`, and `<undefined>` elements as missing values, and removes observations with any of these characteristics:

- Missing value in the response variable
- At least one missing value in a predictor observation (row in `X` or `Tbl`)
- `NaN` value or `0` weight (`'Weights'`)

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Note You cannot use any cross-validation name-value pair argument along with the `'OptimizeHyperparameters'` name-value pair argument. You can modify the cross-validation for `'OptimizeHyperparameters'` only by using the `'HyperparameterOptimizationOptions'` name-value pair argument.

Example: `Mdl = fitrkernel(X,Y,'Learner','leastsquares','NumExpansionDimensions',2^15,'KernelScale','auto')` implements least-squares regression after mapping the predictor data to the `2^15` dimensional space using feature expansion with a kernel scale parameter selected by a heuristic procedure.

Kernel Regression Options

`BoxConstraint` — Box constraint

1 (default) | positive scalar

Box constraint on page 33-6801, specified as the comma-separated pair consisting of `'BoxConstraint'` and a positive scalar.

This argument is valid only when `'Learner'` is `'svm'`(default) and you do not specify a value for the regularization term strength `'Lambda'`. You can specify either `'BoxConstraint'` or `'Lambda'` because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations (rows in `X`).

Example: `'BoxConstraint',100`

Data Types: `single` | `double`

`Epsilon` — Half width of epsilon-insensitive band
`'auto'` (default) | nonnegative scalar value

Half the width of the epsilon-insensitive band, specified as the comma-separated pair consisting of `'Epsilon'` and `'auto'` or a nonnegative scalar value.

For `'auto'`, the `fitrkernel` function determines the value of `Epsilon` as `iqr(Y)/13.49`, which is an estimate of a tenth of the standard deviation using the interquartile range of the response variable `Y`. If `iqr(Y)` is equal to zero, then `fitrkernel` sets the value of `Epsilon` to 0.1.

`'Epsilon'` is valid only when `Learner` is `svm`.

Example: `'Epsilon',0.3`

Data Types: `single` | `double`

`NumExpansionDimensions` — Number of dimensions of expanded space
`'auto'` (default) | positive integer

Number of dimensions of the expanded space, specified as the comma-separated pair consisting of `'NumExpansionDimensions'` and `'auto'` or a positive integer. For `'auto'`, the `fitrkernel` function selects the number of dimensions using `2.^ceil(min(log2(p)+5,15))`, where `p` is the number of predictors.

Example: `'NumExpansionDimensions',2^15`

Data Types: `char` | `string` | `single` | `double`

`KernelScale` — Kernel scale parameter
`1` (default) | `'auto'` | positive scalar

Kernel scale parameter, specified as the comma-separated pair consisting of `'KernelScale'` and `'auto'` or a positive scalar. MATLAB obtains the random basis for random feature expansion by using the kernel scale parameter. For details, see “Random Feature Expansion” on page 33-6800.

If you specify `'auto'`, then MATLAB selects an appropriate kernel scale parameter using a heuristic procedure. This heuristic procedure uses subsampling, so estimates can vary from one call to another. Therefore, to reproduce results, set a random number seed by using `rng` before training.

Example: `'KernelScale','auto'`

Data Types: `char` | `string` | `single` | `double`

`Lambda` — Regularization term strength
`'auto'` (default) | nonnegative scalar

Regularization term strength, specified as the comma-separated pair consisting of `'Lambda'` and `'auto'` or a nonnegative scalar.

For `'auto'`, the value of `'Lambda'` is $1/n$, where n is the number of observations (rows in `X`).

You can specify either `'BoxConstraint'` or `'Lambda'` because the box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$.

Example: `'Lambda',0.01`

Data Types: `char` | `string` | `single` | `double`

`Learner` — Linear regression model type
`'svm'` (default) | `'leastsquares'`

Linear regression model type, specified as the comma-separated pair consisting of `'Learner'` and `'svm'` or `'leastsquares'`.

In the following table, $f(x) = T(x)\beta + b$.

- x is an observation (row vector) from p predictor variables.
- $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
- β is a vector of m coefficients.
- b is the scalar bias.

Value	Algorithm	Response range	Loss function
`'leastsquares'`	Linear regression via ordinary least squares	$y \in (-\infty,\infty)$	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$
`'svm'`	Support vector machine regression	Same as `'leastsquares'`	Epsilon-insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$

Example: `'Learner','leastsquares'`

`Verbose` — Verbosity level
`0` (default) | `1`

Verbosity level, specified as the comma-separated pair consisting of `'Verbose'` and either `0` or `1`. `Verbose` controls the amount of diagnostic information `fitrkernel` displays at the command line.

Value	Description
`0`	`fitrkernel` does not display diagnostic information.
`1`	`fitrkernel` displays and stores the value of the objective function, gradient magnitude, and other diagnostic information. `FitInfo.History` contains the diagnostic information.

Example: `'Verbose',1`

Data Types: `single` | `double`

`BlockSize` — Maximum amount of allocated memory
`4e^3` (4GB) (default) | positive scalar

Maximum amount of allocated memory (in megabytes), specified as the comma-separated pair consisting of `'BlockSize'` and a positive scalar.

If `fitrkernel` requires more memory than the value of `BlockSize` to hold the transformed predictor data, then MATLAB uses a block-wise strategy. For details about the block-wise strategy, see “Algorithms” on page 33-6801.

Example: `'BlockSize',1e4`

Data Types: `single` | `double`

`RandomStream` — Random number stream

global stream (default) | random stream object

Random number stream for reproducibility of data transformation, specified as the comma-separated pair consisting of `'RandomStream'` and a random stream object. For details, see “Random Feature Expansion” on page 33-6800.

Use `'RandomStream'` to reproduce the random basis functions that `fitrkernel` uses to transform the data in `X` to a high-dimensional space. For details, see “Managing the Global Stream Using RandStream” and “Creating and Controlling a Random Number Stream”.

Example: `'RandomStream',RandStream('mlfg6331_64')`

Other Regression Options

`CategoricalPredictors` — Categorical predictors list

vector of positive integers | logical vector | character matrix | string array | cell array of character vectors | `'all'`

Categorical predictors list, specified as one of the values in this table.

Value	Description
Vector of positive integers	Each entry in the vector is an index value corresponding to the column of the predictor data that contains a categorical variable. The index values are between 1 and `p`, where `p` is the number of predictors used to train the model. If `fitrkernel` uses a subset of input variables as predictors, then the function indexes the predictors using only the subset. The `'CategoricalPredictors'` values do not count the response variable, the observation weight variable, and any other variables that the function does not use.
Logical vector	A `true` entry means that the corresponding column of predictor data is a categorical variable. The length of the vector is `p`.
Character matrix	Each row of the matrix is the name of a predictor variable. The names must match the entries in `PredictorNames`. Pad the names with extra blanks so each row of the character matrix has the same length.
String array or cell array of character vectors	Each element in the array is the name of a predictor variable. The names must match the entries in `PredictorNames`.
`'all'`	All predictors are categorical.

By default, if the predictor data is in a table (`Tbl`), `fitrkernel` assumes that a variable is categorical if it is a logical vector, categorical vector, character array, string array, or cell array of character vectors. If the predictor data is a matrix (`X`), `fitrkernel` assumes that all predictors are continuous. To identify any other predictors as categorical predictors, specify them by using the `'CategoricalPredictors'` name-value argument.

For the identified categorical predictors, `fitrkernel` creates dummy variables using two different schemes, depending on whether a categorical variable is unordered or ordered. For an unordered

categorical variable, `fitrkernel` creates one dummy variable for each level of the categorical variable. For an ordered categorical variable, `fitrkernel` creates one less dummy variable than the number of categories. For details, see “Automatic Creation of Dummy Variables” on page 2-49.

Example: `'CategoricalPredictors','all'`

Data Types: `single` | `double` | `logical` | `char` | `string` | `cell`

`PredictorNames` — Predictor variable names
string array of unique names | cell array of unique character vectors

Predictor variable names, specified as a string array of unique names or cell array of unique character vectors. The functionality of `PredictorNames` depends on the way you supply the training data.

- If you supply `X` and `Y`, then you can use `PredictorNames` to assign names to the predictor variables in `X`.
  - The order of the names in `PredictorNames` must correspond to the column order of `X`. That is, `PredictorNames{1}` is the name of `X(:,1)`, `PredictorNames{2}` is the name of `X(:,2)`, and so on. Also, `size(X,2)` and `numel(PredictorNames)` must be equal.
  - By default, `PredictorNames` is `{'x1','x2',...}`.
- If you supply `Tbl`, then you can use `PredictorNames` to choose which predictor variables to use in training. That is, `fitrkernel` uses only the predictor variables in `PredictorNames` and the response variable during training.
  - `PredictorNames` must be a subset of `Tbl.Properties.VariableNames` and cannot include the name of the response variable.
  - By default, `PredictorNames` contains the names of all predictor variables.
  - A good practice is to specify the predictors for training using either `'PredictorNames'` or `formula`, but not both.

Example: `'PredictorNames',`
`{'SepalLength','SepalWidth','PetalLength','PetalWidth'}`

Data Types: `string` | `cell`

`ResponseName` — Response variable name
`'Y'` (default) | character vector | string scalar

Response variable name, specified as a character vector or string scalar.

- If you supply `Y`, then you can use `'ResponseName'` to specify a name for the response variable.
- If you supply `ResponseVarName` or `formula`, then you cannot use `'ResponseName'`.

Example: `'ResponseName','response'`

Data Types: `char` | `string`

`ResponseTransform` — Response transformation
`'none'` (default) | function handle

Response transformation, specified as either `'none'` or a function handle. The default is `'none'`, which means `@(y)y`, or no transformation. For a MATLAB function or a function you define, use its function handle for the response transformation. The function handle must accept a vector (the original response values) and return a vector of the same size (the transformed response values).

Example: Suppose you create a function handle that applies an exponential transformation to an input vector by using `myfunction = @(y)exp(y)`. Then, you can specify the response transformation as `'ResponseTransform',myfunction`.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights
vector of scalar values | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a vector of scalar values or the name of a variable in `Tbl`. The software weights each observation (or row) in `X` or `Tbl` with the corresponding value in `Weights`. The length of `Weights` must equal the number of rows in `X` or `Tbl`.

If you specify the input data as a table `Tbl`, then `Weights` can be the name of a variable in `Tbl` that contains a numeric vector. In this case, you must specify `Weights` as a character vector or string scalar. For example, if weights vector `W` is stored as `Tbl.W`, then specify it as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `W`, as predictors when training the model.

By default, `Weights` is `ones(n,1)`, where `n` is the number of observations in `X` or `Tbl`.

`fitrkernel` normalizes the weights to sum to 1.

Data Types: `single` | `double` | `char` | `string`

Cross-Validation Options

`CrossVal` — Cross-validation flag
`'off'` (default) | `'on'`

Cross-validation flag, specified as the comma-separated pair consisting of `'Crossval'` and `'on'` or `'off'`.

If you specify `'on'`, then the software implements 10-fold cross-validation.

You can override this cross-validation setting using the `CVPartition`, `Holdout`, `KFold`, or `Leaveout` name-value pair argument. You can use only one cross-validation name-value pair argument at a time to create a cross-validated model.

Example: `'Crossval','on'`

`CVPartition` — Cross-validation partition
`[]` (default) | `cvpartition` partition object

Cross-validation partition, specified as a `cvpartition` partition object created by `cvpartition`. The partition object specifies the type of cross-validation and the indexing for the training and validation sets.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: Suppose you create a random partition for 5-fold cross-validation on 500 observations by using `cvp = cvpartition(500,'KFold',5)`. Then, you can specify the cross-validated model by using `'CVPartition',cvp`.

`Holdout` — Fraction of data for holdout validation
scalar value in the range (0,1)

Fraction of the data used for holdout validation, specified as a scalar value in the range (0,1). If you specify `'Holdout',p`, then the software completes these steps:

1. Randomly select and reserve `p*100`% of the data as validation data, and train the model using the rest of the data.
2. Store the compact, trained model in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Holdout',0.1`

Data Types: `double` | `single`

`KFold` — Number of folds
`10` (default) | positive integer value greater than 1

Number of folds to use in a cross-validated model, specified as a positive integer value greater than 1. If you specify `'KFold',k`, then the software completes these steps:

1. Randomly partition the data into `k` sets.
2. For each set, reserve the set as validation data, and train the model using the other `k` – 1 sets.
3. Store the `k` compact, trained models in a `k`-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can specify only one of these four name-value arguments: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'KFold',5`

Data Types: `single` | `double`

`Leaveout` — Leave-one-out cross-validation flag
`'off'` (default) | `'on'`

Leave-one-out cross-validation flag, specified as the comma-separated pair consisting of `'Leaveout'` and `'on'` or `'off'`. If you specify `'Leaveout','on'`, then, for each of the *n* observations (where *n* is the number of observations excluding missing observations), the software completes these steps:

1. Reserve the observation as validation data, and train the model using the other *n* – 1 observations.
2. Store the *n* compact, trained models in the cells of an *n*-by-1 cell vector in the `Trained` property of the cross-validated model.

To create a cross-validated model, you can use one of these four name-value pair arguments only: `CVPartition`, `Holdout`, `KFold`, or `Leaveout`.

Example: `'Leaveout','on'`

Convergence Controls

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`1e-5` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`1e-6` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\|\nabla \mathscr{L}_t\|_\infty = \max|\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`HessianHistorySize` — Size of history buffer for Hessian approximation
`15` (default) | positive integer

Size of the history buffer for Hessian approximation, specified as the comma-separated pair consisting of `'HessianHistorySize'` and a positive integer. At each iteration, `fitrkernel` composes the Hessian by using statistics from the latest `HessianHistorySize` iterations.

Example: `'HessianHistorySize',10`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of optimization iterations
positive integer

Maximum number of optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The default value is 1000 if the transformed data fits in memory, as specified by `BlockSize`. Otherwise, the default value is 100.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Hyperparameter Optimization Options

`OptimizeHyperparameters` — Parameters to optimize
`'none'` (default) | `'auto'` | `'all'` | string array or cell array of eligible parameter names | vector of `optimizableVariable` objects

Parameters to optimize, specified as the comma-separated pair consisting of `'OptimizeHyperparameters'` and one of these values:

- `'none'` — Do not optimize.
- `'auto'` — Use `{'KernelScale','Lambda','Epsilon'}`.
- `'all'` — Optimize all eligible parameters.
- Cell array of eligible parameter names.
- Vector of `optimizableVariable` objects, typically the output of `hyperparameters`.

The optimization attempts to minimize the cross-validation loss (error) for `fitrkernel` by varying the parameters. To control the cross-validation type and other aspects of the optimization, use the `HyperparameterOptimizationOptions` name-value pair argument.

Note `'OptimizeHyperparameters'` values override any values you set using other name-value pair arguments. For example, setting `'OptimizeHyperparameters'` to `'auto'` causes the `'auto'` values to apply.

The eligible parameters for `fitrkernel` are:

- `Epsilon` — `fitrkernel` searches among positive values, by default log-scaled in the range `[1e-3,1e2]*iqr(Y)/1.349`.
- `KernelScale` — `fitrkernel` searches among positive values, by default log-scaled in the range `[1e-3,1e3]`.
- `Lambda` — `fitrkernel` searches among positive values, by default log-scaled in the range `[1e-3,1e3]/n`, where `n` is the number of observations.
- `Learner` — `fitrkernel` searches among `'svm'` and `'leastsquares'`.
- `NumExpansionDimensions` — `fitrkernel` searches among positive integers, by default log-scaled in the range `[100,10000]`.

Set nondefault parameters by passing a vector of `optimizableVariable` objects that have nondefault values. For example:

```
load carsmall
params = hyperparameters('fitrkernel',[Horsepower,Weight],MPG);
params(2).Range = [1e-4,1e6];
```

Pass `params` as the value of `'OptimizeHyperparameters'`.

By default, iterative display appears at the command line, and plots appear according to the number of hyperparameters in the optimization. For the optimization and plots, the objective function is log(1 + cross-validation loss) for regression and the misclassification rate for classification. To control the iterative display, set the `Verbose` field of the `'HyperparameterOptimizationOptions'` name-value pair argument. To control the plots, set the `ShowPlots` field of the `'HyperparameterOptimizationOptions'` name-value pair argument.

For an example, see “Optimize Kernel Regression” on page 33-6783.

Example: `'OptimizeHyperparameters','auto'`

`HyperparameterOptimizationOptions` — Options for optimization
structure

Options for optimization, specified as the comma-separated pair consisting of `'HyperparameterOptimizationOptions'` and a structure. This argument modifies the effect of the `OptimizeHyperparameters` name-value pair argument. All fields in the structure are optional.

Field Name	Values	Default
`Optimizer`	• `'bayesopt'` — Use Bayesian optimization. Internally, this setting calls `bayesopt`. • `'gridsearch'` — Use grid search with `NumGridDivisions` values per dimension. • `'randomsearch'` — Search at random among `MaxObjectiveEvaluations` points. `'gridsearch'` searches in a random order, using uniform sampling without replacement from the grid. After optimization, you can get a table in grid order by using the command `sortrows(Mdl.HyperparameterOptimizationResults)`.	`'bayesopt'`
`AcquisitionFunctionName`	• `'expected-improvement-per-second-plus'` • `'expected-improvement'` • `'expected-improvement-plus'` • `'expected-improvement-per-second'` • `'lower-confidence-bound'` • `'probability-of-improvement'` Acquisition functions whose names include `per-second` do not yield reproducible results because the optimization depends on the runtime of the objective function. Acquisition functions whose names include `plus` modify their behavior when they are overexploiting an area. For more details, see “Acquisition Function Types” on page 10-3.	`'expected-improvement-per-second-plus'`
`MaxObjectiveEvaluations`	Maximum number of objective function evaluations.	`30` for `'bayesopt'` or `'randomsearch'`, and the entire grid for `'gridsearch'`
`MaxTime`	Time limit, specified as a positive real. The time limit is in seconds, as measured by `tic` and `toc`. Run time can exceed `MaxTime` because `MaxTime` does not interrupt function evaluations.	`Inf`
`NumGridDivisions`	For `'gridsearch'`, the number of values in each dimension. The value can be a vector of positive integers giving the number of values for each dimension, or a scalar that applies to all dimensions. This field is ignored for categorical variables.	`10`

Field Name	Values	Default
`ShowPlots`	Logical value indicating whether to show plots. If `true`, this field plots the best objective function value against the iteration number. If there are one or two optimization parameters, and if `Optimizer` is `'bayesopt'`, then `ShowPlots` also plots a model of the objective function against the parameters.	`true`
`SaveIntermediateResults`	Logical value indicating whether to save results when `Optimizer` is `'bayesopt'`. If `true`, this field overwrites a workspace variable named `'BayesoptResults'` at each iteration. The variable is a `BayesianOptimization` object.	`false`
`Verbose`	Display to the command line. • `0` — No iterative display • `1` — Iterative display • `2` — Iterative display with extra information For details, see the `bayesopt` `Verbose` name-value pair argument.	`1`
`UseParallel`	Logical value indicating whether to run Bayesian optimization in parallel, which requires Parallel Computing Toolbox. Due to the nonreproducibility of parallel timing, parallel Bayesian optimization does not necessarily yield reproducible results. For details, see “Parallel Bayesian Optimization” on page 10-7.	`false`
`Repartition`	Logical value indicating whether to repartition the cross-validation at every iteration. If `false`, the optimizer uses a single partition for the optimization. `true` usually gives the most robust results because this setting takes partitioning noise into account. However, for good results, `true` requires at least twice as many function evaluations.	`false`
Use no more than one of the following three field names.		
`CVPartition`	A `cvpartition` object, as created by `cvpartition`.	`'Kfold',5` if you do not specify any cross-validation field
`Holdout`	A scalar in the range `(0,1)` representing the holdout fraction.	
`Kfold`	An integer greater than 1.	

Example: `'HyperparameterOptimizationOptions',struct('MaxObjectiveEvaluations',60)`

Data Types: `struct`

Output Arguments

`Mdl` — Trained kernel regression model

`RegressionKernel` model object | `RegressionPartitionedKernel` cross-validated model object

Trained kernel regression model, returned as a `RegressionKernel` model object or `RegressionPartitionedKernel` cross-validated model object.

If you set any of the name-value pair arguments `CrossVal`, `CVPartition`, `Holdout`, `KFold`, or `Leaveout`, then `Mdl` is a `RegressionPartitionedKernel` cross-validated model. Otherwise, `Mdl` is a `RegressionKernel` model.

To reference properties of `Mdl`, use dot notation. For example, enter `Mdl.NumExpansionDimensions` in the Command Window to display the number of dimensions of the expanded space.

Note Unlike other regression models, and for economical memory usage, a `RegressionKernel` model object does not store the training data or training process details (for example, convergence history).

`FitInfo` — Optimization details

structure array

Optimization details, returned as a structure array including fields described in this table. The fields contain final values or name-value pair argument specifications.

Field	Description
`Solver`	Objective function minimization technique: `'LBFGS-fast'`, `'LBFGS-blockwise'`, or `'LBFGS-tall'`. For details, see “Algorithms” on page 33-6801.
`LossFunction`	Loss function. Either mean squared error (MSE) or epsilon-insensitive, depending on the type of linear regression model. See `Learner`.
`Lambda`	Regularization term strength. See `Lambda`.
`BetaTolerance`	Relative tolerance on the linear coefficients and the bias term. See `BetaTolerance`.
`GradientTolerance`	Absolute gradient tolerance. See `GradientTolerance`.
`ObjectiveValue`	Value of the objective function when optimization terminates. The regression loss plus the regularization term compose the objective function.
`GradientMagnitude`	Infinite norm of the gradient vector of the objective function when optimization terminates. See `GradientTolerance`.
`RelativeChangeInBeta`	Relative changes in the linear coefficients and the bias term when optimization terminates. See `BetaTolerance`.
`FitTime`	Elapsed, wall-clock time (in seconds) required to fit the model to the data.

Field	Description
`History`	History of optimization information. This field also includes the optimization information from training `Mdl`. This field is empty (`[]`) if you specify `'Verbose',0`. For details, see `Verbose` and "Algorithms" on page 33-6801.

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.ObjectiveValue` in the Command Window.

Examine the information provided by `FitInfo` to assess whether convergence is satisfactory.

`HyperparameterOptimizationResults` — Cross-validation optimization of hyperparameters
`BayesianOptimization` object | table of hyperparameters and associated values

Cross-validation optimization of hyperparameters, returned as a `BayesianOptimization` object or a table of hyperparameters and associated values. The output is nonempty when the value of `'OptimizeHyperparameters'` is not `'none'`. The output value depends on the `Optimizer` field value of the `'HyperparameterOptimizationOptions'` name-value pair argument:

Value of `Optimizer` Field	Value of `HyperparameterOptimizationResults`
`'bayesopt'` (default)	Object of class `BayesianOptimization`
`'gridsearch'` or `'randomsearch'`	Table of hyperparameters used, observed objective function values (cross-validation loss), and rank of observations from lowest (best) to highest (worst)

Limitations

- `fitrkernel` does not accept initial conditions for the linear coefficients beta (β) and bias term (b) used to determine the decision function, $f(x) = T(x)\beta + b$.
- `fitrkernel` does not support standardization.

More About

Random Feature Expansion

Random feature expansion, such as Random Kitchen Sinks[1] and Fastfood[2], is a scheme to approximate Gaussian kernels of the kernel regression algorithm for big data in a computationally efficient way. Random feature expansion is more practical for big data applications that have large training sets but can also be applied to smaller data sets that fit in memory.

The kernel regression algorithm searches for an optimal function that deviates from each response data point (y_i) by values no greater than the epsilon margin (ε) after mapping the predictor data into a high-dimensional space.

Some regression problems cannot be described adequately using a linear model. In such cases, obtain a nonlinear regression model by replacing the dot product x_1x_2' with a nonlinear kernel function $G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle$, where x_i is the ith observation (row vector) and $\varphi(x_i)$ is a transformation that maps x_i to a high-dimensional space (called the "kernel trick"). However,

evaluating $G(x_1,x_2)$, the Gram matrix, for each pair of observations is computationally expensive for a large data set (large n).

The random feature expansion scheme finds a random transformation so that its dot product approximates the Gaussian kernel. That is,

$$G(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle \approx T(x_1)T(x_2)',$$

where $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$). The Random Kitchen Sink[1] scheme uses the random transformation

$$T(x) = m^{-1/2}\exp(iZx')',$$

where $Z \in \mathbb{R}^{m \times p}$ is a sample drawn from $N\left(0, \sigma^{-2}\right)$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage. The Fastfood[2] scheme introduces another random basis V instead of Z using Hadamard matrices combined with Gaussian scaling matrices. This random basis reduces computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

You can specify values for m and σ^2, using the `NumExpansionDimensions` and `KernelScale` name-value pair arguments of `fitrkernel`, respectively.

The `fitrkernel` function uses the Fastfood scheme for random feature expansion and uses linear regression to train a Gaussian kernel regression model. Unlike solvers in the `fitrsvm` function, which require computation of the n-by-n Gram matrix, the solver in `fitrkernel` only needs to form a matrix of size n-by-m, with m typically much less than n for big data.

Box Constraint

A box constraint is a parameter that controls the maximum penalty imposed on observations that lie outside the epsilon margin (ε), and helps to prevent overfitting (regularization). Increasing the box constraint can lead to longer training times.

The box constraint (C) and the regularization term strength (λ) are related by $C = 1/(\lambda n)$, where n is the number of observations.

Algorithms

`fitrkernel` minimizes the regularized objective function using a Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) solver with ridge (L_2) regularization. To find the type of LBFGS solver used for training, type `FitInfo.Solver` in the Command Window.

- `'LBFGS-fast'` — LBFGS solver.
- `'LBFGS-blockwise'` — LBFGS solver with a block-wise strategy. If `fitrkernel` requires more memory than the value of `BlockSize` to hold the transformed predictor data, then it uses a block-wise strategy.
- `'LBFGS-tall'` — LBFGS solver with a block-wise strategy for tall arrays.

When `fitrkernel` uses a block-wise strategy, `fitrkernel` implements LBFGS by distributing the calculation of the loss and gradient among different parts of the data at each iteration. Also, `fitrkernel` refines the initial estimates of the linear coefficients and the bias term by fitting the model locally to parts of the data and combining the coefficients by averaging. If you specify `'Verbose',1`, then `fitrkernel` displays diagnostic information for each data pass and stores the information in the `History` field of `FitInfo`.

When `fitrkernel` does not use a block-wise strategy, the initial estimates are zeros. If you specify `'Verbose',1`, then `fitrkernel` displays diagnostic information for each iteration and stores the information in the `History` field of `FitInfo`.

References

[1] Rahimi, A., and B. Recht. “Random Features for Large-Scale Kernel Machines.” *Advances in Neural Information Processing Systems*. Vol. 20, 2008, pp. 1177–1184.

[2] Le, Q., T. Sarlós, and A. Smola. “Fastfood — Approximating Kernel Expansions in Loglinear Time.” *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, No. 3, 2013, pp. 244–252.

[3] Huang, P. S., H. Avron, T. N. Sainath, V. Sindhwani, and B. Ramabhadran. “Kernel methods match Deep Neural Networks on TIMIT.” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 205–209.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `fitrkernel` does not support tall `table` data.
- Some name-value pair arguments have different defaults compared to the default values for the in-memory `fitrkernel` function. Supported name-value pair arguments, and any differences, are:
  - `'BoxConstraint'`
  - `'Epsilon'`
  - `'NumExpansionDimensions'`
  - `'KernelScale'`
  - `'Lambda'`
  - `'Learner'`
  - `'Verbose'` — Default value is `1`.
  - `'BlockSize'`
  - `'RandomStream'`
  - `'ResponseTransform'`
  - `'Weights'` — Value must be a tall array.
  - `'BetaTolerance'` — Default value is relaxed to `1e–3`.
  - `'GradientTolerance'` — Default value is relaxed to `1e–5`.
  - `'HessianHistorySize'`
  - `'IterationLimit'` — Default value is relaxed to `20`.
  - `'OptimizeHyperparameters'`
  - `'HyperparameterOptimizationOptions'` — For cross-validation, tall optimization supports only `'Holdout'` validation. By default, the software selects and reserves 20% of the data as holdout validation data, and trains the model using the rest of the data. You can specify

a different value for the holdout fraction by using this argument. For example, specify `'HyperparameterOptimizationOptions',struct('Holdout',0.3)` to reserve 30% of the data as validation data.

- If `'KernelScale'` is `'auto'`, then `fitrkernel` uses the random stream controlled by `tallrng` for subsampling. For reproducibility, you must set a random number seed for both the global stream and the random stream controlled by `tallrng`.
- If `'Lambda'` is `'auto'`, then `fitrkernel` might take an extra pass through the data to calculate the number of observations in `X`.
- `fitrkernel` uses a block-wise strategy. For details, see “Algorithms” on page 33-6801.

For more information, see “Tall Arrays”.

Automatic Parallel Support
Accelerate code by automatically running computation in parallel using Parallel Computing Toolbox™.

To perform parallel hyperparameter optimization, use the `'HyperparameterOptimizationOptions', struct('UseParallel',true)` name-value argument in the call to this function.

For more information on parallel hyperparameter optimization, see “Parallel Bayesian Optimization” on page 10-7.

For general information about parallel computing, see “Run MATLAB Functions with Automatic Parallel Support” (Parallel Computing Toolbox).

See Also

`RegressionKernel` | `RegressionPartitionedKernel` | `bayesopt` | `bestPoint` | `fitrlinear` | `fitrsvm` | `loss` | `predict` | `resume`

Topics
“Understanding Support Vector Machine Regression” on page 25-2

Introduced in R2018a

RegressionKernel

Gaussian kernel regression model using random feature expansion

Description

`RegressionKernel` is a trained model object for Gaussian kernel regression using random feature expansion. `RegressionKernel` is more practical for big data applications that have large training sets but can also be applied to smaller data sets that fit in memory.

Unlike other regression models, and for economical memory usage, `RegressionKernel` model objects do not store the training data. However, they do store information such as the dimension of the expanded space, the kernel scale parameter, and the regularization strength.

You can use trained `RegressionKernel` models to continue training using the training data, predict responses for new data, and compute the mean squared error or epsilon-insensitive loss. For details, see `resume`, `predict`, and `loss`.

Creation

Create a `RegressionKernel` object using the `fitrkernel` function. This function maps data in a low-dimensional space into a high-dimensional space, then fits a linear model in the high-dimensional space by minimizing the regularized objective function. Obtaining the linear model in the high-dimensional space is equivalent to applying the Gaussian kernel to the model in the low-dimensional space. Available linear regression models include regularized support vector machines (SVM) and least-squares regression models.

Properties

Kernel Regression Properties

`Epsilon` — Half width of epsilon-insensitive band
nonnegative scalar

Half the width of the epsilon-insensitive band, specified as a nonnegative scalar.

If `Learner` is not `'svm'`, then `Epsilon` is an empty array (`[]`).

Data Types: `single` | `double`

`Learner` — Linear regression model type
`'svm'` (default) | `'leastsquares'`

Linear regression model type, specified as `'leastsquares'` or `'svm'`.

In the following table, $f(x) = T(x)\beta + b$.

- x is an observation (row vector) from p predictor variables.

- $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
- β is a vector of m coefficients.
- b is the scalar bias.

Value	Algorithm	Loss function	FittedLoss Value
`'leastsquares'`	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	`'mse'`
`'svm'`	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	`'epsiloninsensitive'`

`NumExpansionDimensions` — Number of dimensions of expanded space
positive integer

Number of dimensions of the expanded space, specified as a positive integer.

Data Types: `single` | `double`

`KernelScale` — Kernel scale parameter
positive scalar

Kernel scale parameter, specified as a positive scalar.

Data Types: `single` | `double`

`BoxConstraint` — Box constraint
positive scalar

Box constraint, specified as a positive scalar.

Data Types: `double` | `single`

`Lambda` — Regularization term strength
nonnegative scalar

Regularization term strength, specified as a nonnegative scalar.

Data Types: `single` | `double`

`FittedLoss` — Loss function used to fit linear model
`'epsiloninsensitive'` | `'mse'`

Loss function used to fit the linear model, specified as `'epsiloninsensitive'` or `'mse'`.

Value	Algorithm	Loss function	Learner Value
`'epsiloninsensitive'`	Support vector machine regression	Epsilon insensitive: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$	`'svm'`

Value	Algorithm	Loss function	Learner Value
`'mse'`	Linear regression through ordinary least squares	Mean squared error (MSE): $\ell[y, f(x)] = \frac{1}{2}[y - f(x)]^2$	`'leastsquares'`

`Regularization` — Complexity penalty type
`'lasso (L1)'` | `'ridge (L2)'`

Complexity penalty type, specified as `'lasso (L1)'` or `'ridge (L2)'`.

The software composes the objective function for minimization from the sum of the average loss function (see `FittedLoss`) and a regularization value from this table.

Value	Description
`'lasso (L1)'`	Lasso (L_1) penalty: $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$
`'ridge (L2)'`	Ridge (L_2) penalty: $\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2$

λ specifies the regularization term strength (see `Lambda`).

The software excludes the bias term (β_0) from the regularization penalty.

Other Regression Properties

`CategoricalPredictors` — Indices of categorical predictors
vector of positive integers

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`ModelParameters` — Parameters used for training model
structure

Parameters used for training the `RegressionKernel` model, specified as a structure.

Access fields of `ModelParameters` using dot notation. For example, access the relative tolerance on the linear coefficients and the bias term by using `Mdl.ModelParameters.BetaTolerance`.

Data Types: `struct`

`PredictorNames` — Predictor names
cell array of character vectors

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns used as predictor variables in the training data `X` or `Tbl`.

Data Types: `cell`

`ExpandedPredictorNames` — Expanded predictor names
cell array of character vectors

Expanded predictor names, specified as a cell array of character vectors.

If the model uses encoding for categorical variables, then `ExpandedPredictorNames` includes the names that describe the expanded variables. Otherwise, `ExpandedPredictorNames` is the same as `PredictorNames`.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function to apply to predicted responses
`'none'` | function handle

Response transformation function to apply to predicted responses, specified as `'none'` or a function handle.

For kernel regression models and before the response transformation, the predicted response for the observation x (row vector) is $f(x) = T(x)\beta + b$.

- $T(\cdot)$ is a transformation of an observation for feature expansion.
- β corresponds to `Mdl.Beta`.
- b corresponds to `Mdl.Bias`.

For a MATLAB function or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @function`, where `function` accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `function_handle`

Object Functions

lime	Local interpretable model-agnostic explanations (LIME)
loss	Regression loss for Gaussian kernel regression model
partialDependence	Compute partial dependence
plotPartialDependence	Create partial dependence plot (PDP) and individual conditional expectation (ICE) plots
predict	Predict responses for Gaussian kernel regression model
resume	Resume training of Gaussian kernel regression model
shapley	Shapley values

Examples

Train Gaussian Kernel Regression Model

Train a kernel regression model for a tall array by using SVM.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore that references the folder location with the data. The data can be contained in a single file, a collection of files, or an entire folder. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Select a subset of the variables to use. Create a tall table on top of the datastore.

```
varnames = {'ArrTime','DepTime','ActualElapsedTime'};
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
    'SelectedVariableNames',varnames);
t = tall(ds);
```

Specify `DepTime` and `ArrTime` as the predictor variables (`X`) and `ActualElapsedTime` as the response variable (`Y`). Select the observations for which `ArrTime` is later than `DepTime`.

```
daytime = t.ArrTime>t.DepTime;
Y = t.ActualElapsedTime(daytime);     % Response data
X = t{daytime,{'DepTime' 'ArrTime'}}; % Predictor data
```

Standardize the predictor variables.

```
Z = zscore(X); % Standardize the data
```

Train a default Gaussian kernel regression model with the standardized predictors. Extract a fit summary to determine how well the optimization algorithm fits the model to the data.

```
[Mdl,FitInfo] = fitrkernel(Z,Y)
```

```
Found 6 chunks.
|=====================================================================================|
| Solver | Iteration  /  |   Objective   |   Gradient    | Beta relative |
|        | Data Pass     |               |   magnitude   |    change     |
|=====================================================================================|
|   INIT |     0 /     1 |  4.313465e+01 |  6.296907e-02 |           NaN |
|  LBFGS |     0 /     2 |  3.704335e+01 |  1.789316e-02 |  9.985854e-01 |
|  LBFGS |     1 /     3 |  3.703211e+01 |  2.880402e-02 |  1.044172e-03 |
|  LBFGS |     2 /     4 |  3.701616e+01 |  2.297788e-02 |  5.115891e-04 |
|  LBFGS |     2 /     5 |  3.700183e+01 |  1.750937e-02 |  1.023672e-03 |
|  LBFGS |     3 /     6 |  3.679055e+01 |  4.815047e-02 |  1.113182e-02 |
|  LBFGS |     4 /     7 |  3.637852e+01 |  1.058657e-01 |  2.994089e-02 |
|  LBFGS |     5 /     8 |  3.565372e+01 |  1.406536e-01 |  7.033477e-02 |
|  LBFGS |     6 /     9 |  3.478061e+01 |  1.479288e-01 |  1.185262e-01 |
|  LBFGS |     7 /    10 |  3.616955e+01 |  1.544917e-01 |  2.790848e-01 |
|  LBFGS |     7 /    11 |  3.459534e+01 |  1.212256e-01 |  1.229242e-01 |
|  LBFGS |     8 /    12 |  3.379859e+01 |  8.791025e-02 |  5.417481e-02 |
|  LBFGS |     9 /    13 |  3.339981e+01 |  3.077806e-02 |  4.638645e-02 |
|  LBFGS |    10 /    14 |  3.325224e+01 |  3.082755e-02 |  2.867793e-02 |
|  LBFGS |    11 /    15 |  3.320036e+01 |  4.168377e-02 |  9.376887e-03 |
|  LBFGS |    12 /    16 |  3.309321e+01 |  5.018195e-02 |  1.831484e-02 |
|  LBFGS |    13 /    17 |  3.288069e+01 |  4.506485e-02 |  3.732443e-02 |
|  LBFGS |    14 /    18 |  3.245691e+01 |  3.787163e-02 |  1.036929e-01 |
|  LBFGS |    15 /    19 |  3.210116e+01 |  2.418833e-02 |  1.190984e-01 |
|  LBFGS |    16 /    20 |  3.190585e+01 |  2.666398e-02 |  3.921991e-02 |
```

```
|=====================================================================|
| Solver | Iteration  /  |   Objective   |   Gradient    | Beta relative |
|        | Data Pass     |               |   magnitude   |    change     |
|=====================================================================|
|  LBFGS |    17 /    21 |  3.172622e+01 |  2.548259e-02 |  3.805655e-02 |
|  LBFGS |    18 /    22 |  3.154538e+01 |  1.280266e-02 |  4.363429e-02 |
|  LBFGS |    19 /    23 |  3.138533e+01 |  1.446779e-02 |  8.822868e-02 |
|  LBFGS |    20 /    24 |  3.283513e+01 |  2.218528e-01 |  1.318597e-01 |
|  LBFGS |    20 /    25 |  3.158782e+01 |  1.019184e-01 |  6.992082e-02 |
|  LBFGS |    20 /    26 |  3.136869e+01 |  4.678412e-02 |  3.603399e-02 |
|=====================================================================|

Mdl =
  RegressionKernel
            PredictorNames: {'x1'  'x2'}
              ResponseName: 'Y'
                   Learner: 'svm'
    NumExpansionDimensions: 64
               KernelScale: 1
                    Lambda: 8.5385e-06
            BoxConstraint: 1
                  Epsilon: 5.9303


  Properties, Methods


FitInfo = struct with fields:
                  Solver: 'LBFGS-tall'
            LossFunction: 'epsiloninsensitive'
                  Lambda: 8.5385e-06
           BetaTolerance: 1.0000e-03
       GradientTolerance: 1.0000e-05
          ObjectiveValue: 31.3687
       GradientMagnitude: 0.0468
    RelativeChangeInBeta: 0.0360
                 FitTime: 50.9380
                 History: [1x1 struct]
```

`Mdl` is a `RegressionKernel` model. To inspect the regression error, you can pass `Mdl` and the training data or new data to the `loss` function. Or, you can pass `Mdl` and new predictor data to the `predict` function to predict responses for new observations. You can also pass `Mdl` and the training data to the `resume` function to continue training.

`FitInfo` is a structure array containing optimization information. Use `FitInfo` to determine whether optimization termination measurements are satisfactory.

For improved accuracy, you can increase the maximum number of optimization iterations (`'IterationLimit'`) and decrease the tolerance values (`'BetaTolerance'` and `'GradientTolerance'`) by using the name-value pair arguments of `fitrkernel`. Doing so can improve measures like `ObjectiveValue` and `RelativeChangeInBeta` in `FitInfo`. You can also optimize model parameters by using the `'OptimizeHyperparameters'` name-value pair argument.

Estimate Sample Loss and Resume Training

Resume training a Gaussian kernel regression model for more iterations to improve the regression loss.

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:5);
Y = R(:,end);
```

Reserve 10% of the observations as a holdout sample. Extract the training and test indices from the partition definition.

```
rng(10)  % For reproducibility
N = length(Y);
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Standardize the training data and train a kernel regression model. Set the iteration limit to 5 and specify `'Verbose',1` to display diagnostic information.

```
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training data
tr_sigma(tr_sigma==0) = 1;
Mdl = fitrkernel(Ztrain,Ytrain,'IterationLimit',5,'Verbose',1)
```

```
|=============================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================
|  LBFGS |      1 |            0 |  5.691016e+00 |  0.000000e+00 |  5.852758e-02 |
|  LBFGS |      1 |            1 |  5.086537e+00 |  8.000000e+00 |  5.220869e-02 |   9.846711e-02
|  LBFGS |      1 |            2 |  3.862301e+00 |  5.000000e-01 |  3.796034e-01 |   5.998808e-01
|  LBFGS |      1 |            3 |  3.460613e+00 |  1.000000e+00 |  3.257790e-01 |   1.615091e-01
|  LBFGS |      1 |            4 |  3.136228e+00 |  1.000000e+00 |  2.832861e-02 |   8.006254e-02
|  LBFGS |      1 |            5 |  3.063978e+00 |  1.000000e+00 |  1.475038e-02 |   3.314455e-02
|=============================================================================================
```

```
Mdl =
  RegressionKernel
             ResponseName: 'Y'
                  Learner: 'svm'
    NumExpansionDimensions: 256
               KernelScale: 1
                    Lambda: 0.0028
```

```
              BoxConstraint: 1
                    Epsilon: 0.8617


  Properties, Methods
```

`Mdl` is a `RegressionKernel` model.

Standardize the test data using the same mean and standard deviation of the training data columns. Estimate the epsilon-insensitive error for the test set.

```
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Standardize the test data
Ytest = Y(idxTest);

L = loss(Mdl,Ztest,Ytest,'LossFun','epsiloninsensitive')
```

```
L = 2.0674
```

Continue training the model by using `resume`. This function continues training with the same options used for training `Mdl`.

```
UpdatedMdl = resume(Mdl,Ztrain,Ytrain);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient    |    Relative
|        |        |               |               |               |   magnitude    | change in Beta
|=============================================================================================================
|  LBFGS |      1 |             0 |  3.063978e+00 |  0.000000e+00 |  1.475038e-02 |
|  LBFGS |      1 |             1 |  3.007822e+00 |  8.000000e+00 |  1.391637e-02 |   2.603966e-02
|  LBFGS |      1 |             2 |  2.817171e+00 |  5.000000e-01 |  5.949008e-02 |   1.918084e-01
|  LBFGS |      1 |             3 |  2.807294e+00 |  2.500000e-01 |  6.798867e-02 |   2.973097e-02
|  LBFGS |      1 |             4 |  2.791060e+00 |  1.000000e+00 |  2.549575e-02 |   1.639328e-02
|  LBFGS |      1 |             5 |  2.767821e+00 |  1.000000e+00 |  6.154419e-03 |   2.468903e-02
|  LBFGS |      1 |             6 |  2.738163e+00 |  1.000000e+00 |  5.949008e-02 |   9.476263e-02
|  LBFGS |      1 |             7 |  2.719146e+00 |  1.000000e+00 |  1.699717e-02 |   1.849972e-02
|  LBFGS |      1 |             8 |  2.705941e+00 |  1.000000e+00 |  3.116147e-02 |   4.152590e-02
|  LBFGS |      1 |             9 |  2.701162e+00 |  1.000000e+00 |  5.665722e-03 |   9.401466e-03
|  LBFGS |      1 |            10 |  2.695341e+00 |  5.000000e-01 |  3.116147e-02 |   4.968046e-02
|  LBFGS |      1 |            11 |  2.691277e+00 |  1.000000e+00 |  8.498584e-03 |   1.017446e-02
|  LBFGS |      1 |            12 |  2.689972e+00 |  1.000000e+00 |  1.983003e-02 |   9.938921e-03
|  LBFGS |      1 |            13 |  2.688979e+00 |  1.000000e+00 |  1.416431e-02 |   6.606316e-03
|  LBFGS |      1 |            14 |  2.687787e+00 |  1.000000e+00 |  1.621956e-03 |   7.089542e-03
|  LBFGS |      1 |            15 |  2.686539e+00 |  1.000000e+00 |  1.699717e-02 |   1.169701e-02
|  LBFGS |      1 |            16 |  2.685356e+00 |  1.000000e+00 |  1.133144e-02 |   1.069310e-02
|  LBFGS |      1 |            17 |  2.685021e+00 |  5.000000e-01 |  1.133144e-02 |   2.104248e-02
|  LBFGS |      1 |            18 |  2.684002e+00 |  1.000000e+00 |  2.832861e-03 |   6.175231e-03
|  LBFGS |      1 |            19 |  2.683507e+00 |  1.000000e+00 |  5.665722e-03 |   3.724026e-03
|  LBFGS |      1 |            20 |  2.683343e+00 |  5.000000e-01 |  5.665722e-03 |   9.549119e-03
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient    |    Relative
|        |        |               |               |               |   magnitude    | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            21 |  2.682897e+00 |  1.000000e+00 |  5.665722e-03 |   7.172867e-03
|  LBFGS |      1 |            22 |  2.682682e+00 |  1.000000e+00 |  2.832861e-03 |   2.587726e-03
|  LBFGS |      1 |            23 |  2.682485e+00 |  1.000000e+00 |  2.832861e-03 |   2.953648e-03
|  LBFGS |      1 |            24 |  2.682326e+00 |  1.000000e+00 |  2.832861e-03 |   7.777294e-03
```

```
|  LBFGS |      1 |           25 |  2.681914e+00 |  1.000000e+00 |  2.832861e-03 |   2.778555e-03
|  LBFGS |      1 |           26 |  2.681867e+00 |  5.000000e-01 |  1.031085e-03 |   3.638352e-03
|  LBFGS |      1 |           27 |  2.681725e+00 |  1.000000e+00 |  5.665722e-03 |   1.515199e-03
|  LBFGS |      1 |           28 |  2.681692e+00 |  5.000000e-01 |  1.314940e-03 |   1.850055e-03
|  LBFGS |      1 |           29 |  2.681625e+00 |  1.000000e+00 |  2.832861e-03 |   1.456903e-03
|  LBFGS |      1 |           30 |  2.681594e+00 |  5.000000e-01 |  2.832861e-03 |   8.704875e-04
|  LBFGS |      1 |           31 |  2.681581e+00 |  5.000000e-01 |  8.498584e-03 |   3.934768e-04
|  LBFGS |      1 |           32 |  2.681579e+00 |  1.000000e+00 |  8.498584e-03 |   1.847866e-03
|  LBFGS |      1 |           33 |  2.681553e+00 |  1.000000e+00 |  9.857038e-04 |   6.509825e-04
|  LBFGS |      1 |           34 |  2.681541e+00 |  5.000000e-01 |  8.498584e-03 |   6.635528e-04
|  LBFGS |      1 |           35 |  2.681499e+00 |  1.000000e+00 |  5.665722e-03 |   6.194735e-04
|  LBFGS |      1 |           36 |  2.681493e+00 |  5.000000e-01 |  1.133144e-02 |   1.617763e-03
|  LBFGS |      1 |           37 |  2.681473e+00 |  1.000000e+00 |  9.869233e-04 |   8.418484e-04
|  LBFGS |      1 |           38 |  2.681469e+00 |  1.000000e+00 |  5.665722e-03 |   1.069722e-03
|  LBFGS |      1 |           39 |  2.681432e+00 |  1.000000e+00 |  2.832861e-03 |   8.501930e-04
|  LBFGS |      1 |           40 |  2.681423e+00 |  2.500000e-01 |  1.133144e-02 |   9.543716e-04
|=========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=========================================================================================================
|  LBFGS |      1 |           41 |  2.681416e+00 |  1.000000e+00 |  2.832861e-03 |   8.763251e-04
|  LBFGS |      1 |           42 |  2.681413e+00 |  5.000000e-01 |  2.832861e-03 |   4.101888e-04
|  LBFGS |      1 |           43 |  2.681403e+00 |  1.000000e+00 |  5.665722e-03 |   2.713209e-04
|  LBFGS |      1 |           44 |  2.681392e+00 |  1.000000e+00 |  2.832861e-03 |   2.115241e-04
|  LBFGS |      1 |           45 |  2.681383e+00 |  1.000000e+00 |  2.832861e-03 |   2.872858e-04
|  LBFGS |      1 |           46 |  2.681374e+00 |  1.000000e+00 |  8.498584e-03 |   5.771001e-04
|  LBFGS |      1 |           47 |  2.681353e+00 |  1.000000e+00 |  2.832861e-03 |   3.160871e-04
|  LBFGS |      1 |           48 |  2.681334e+00 |  5.000000e-01 |  8.498584e-03 |   1.045502e-03
|  LBFGS |      1 |           49 |  2.681314e+00 |  1.000000e+00 |  7.878714e-04 |   1.505118e-03
|  LBFGS |      1 |           50 |  2.681306e+00 |  1.000000e+00 |  2.832861e-03 |   4.756894e-04
|  LBFGS |      1 |           51 |  2.681301e+00 |  1.000000e+00 |  1.133144e-02 |   3.664873e-04
|  LBFGS |      1 |           52 |  2.681288e+00 |  1.000000e+00 |  2.832861e-03 |   1.449821e-04
|  LBFGS |      1 |           53 |  2.681287e+00 |  2.500000e-01 |  1.699717e-02 |   2.357176e-04
|  LBFGS |      1 |           54 |  2.681282e+00 |  1.000000e+00 |  5.665722e-03 |   2.046663e-04
|  LBFGS |      1 |           55 |  2.681278e+00 |  1.000000e+00 |  2.832861e-03 |   2.546349e-04
|  LBFGS |      1 |           56 |  2.681276e+00 |  2.500000e-01 |  1.307940e-03 |   1.966786e-04
|  LBFGS |      1 |           57 |  2.681274e+00 |  5.000000e-01 |  1.416431e-02 |   1.005310e-04
|  LBFGS |      1 |           58 |  2.681271e+00 |  5.000000e-01 |  1.118892e-03 |   1.147324e-04
|  LBFGS |      1 |           59 |  2.681269e+00 |  1.000000e+00 |  2.832861e-03 |   1.332914e-04
|  LBFGS |      1 |           60 |  2.681268e+00 |  2.500000e-01 |  1.132045e-03 |   5.441369e-05
|=========================================================================================================
```

Estimate the epsilon-insensitive error for the test set using the updated model.

```
UpdatedL = loss(UpdatedMdl,Ztest,Ytest,'LossFun','epsiloninsensitive')
```

```
UpdatedL = 1.8933
```

The regression error decreases by a factor of about `0.08` after `resume` updates the regression model with more iterations.

See Also

`RegressionLinear` | `fitrkernel` | `fitrlinear`

Introduced in R2018a

loss

Regression loss for Gaussian kernel regression model

Syntax

```
L = loss(Mdl,X,Y)

L = loss(Mdl,Tbl,ResponseVarName)
L = loss(Mdl,Tbl,Y)

L = loss(___,Name,Value)
```

Description

`L = loss(Mdl,X,Y)` returns the mean squared error (MSE) for the Gaussian kernel regression model `Mdl` using the predictor data in `X` and the corresponding responses in `Y`.

`L = loss(Mdl,Tbl,ResponseVarName)` returns the MSE for the model `Mdl` using the predictor data in `Tbl` and the true responses in `Tbl.ResponseVarName`.

`L = loss(Mdl,Tbl,Y)` returns the MSE for the model `Mdl` using the predictor data in table `Tbl` and the true responses in `Y`.

`L = loss(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can specify a regression loss function and observation weights. Then, `loss` returns the weighted regression loss using the specified loss function.

Examples

Calculate Sample Loss for Gaussian Kernel Regression Model

Train a Gaussian kernel regression model for a tall array, then calculate the resubstitution mean squared error and epsilon-insensitive error.

When you perform calculations on tall arrays, MATLAB® uses either a parallel pool (default if you have Parallel Computing Toolbox™) or the local MATLAB session. To run the example using the local MATLAB session when you have Parallel Computing Toolbox, change the global execution environment by using the `mapreducer` function.

```
mapreducer(0)
```

Create a datastore that references the folder location with the data. The data can be contained in a single file, a collection of files, or an entire folder. Treat `'NA'` values as missing data so that `datastore` replaces them with `NaN` values. Select a subset of the variables to use. Create a tall table on top of the datastore.

```
varnames = {'ArrTime','DepTime','ActualElapsedTime'};
ds = datastore('airlinesmall.csv','TreatAsMissing','NA',...
```

```
    'SelectedVariableNames',varnames);
t = tall(ds);
```

Specify `DepTime` and `ArrTime` as the predictor variables (`X`) and `ActualElapsedTime` as the response variable (`Y`). Select the observations for which `ArrTime` is later than `DepTime`.

```
daytime = t.ArrTime>t.DepTime;
Y = t.ActualElapsedTime(daytime);     % Response data
X = t{daytime,{'DepTime' 'ArrTime'}}; % Predictor data
```

Standardize the predictor variables.

```
Z = zscore(X); % Standardize the data
```

Train a default Gaussian kernel regression model with the standardized predictors. Set `'Verbose',0` to suppress diagnostic messages.

```
[Mdl,FitInfo] = fitrkernel(Z,Y,'Verbose',0)
```

```
Mdl =
  RegressionKernel
            PredictorNames: {'x1'  'x2'}
              ResponseName: 'Y'
                   Learner: 'svm'
    NumExpansionDimensions: 64
               KernelScale: 1
                    Lambda: 8.5385e-06
            BoxConstraint: 1
                   Epsilon: 5.9303


  Properties, Methods


FitInfo = struct with fields:
                    Solver: 'LBFGS-tall'
              LossFunction: 'epsiloninsensitive'
                    Lambda: 8.5385e-06
             BetaTolerance: 1.0000e-03
       GradientTolerance: 1.0000e-05
          ObjectiveValue: 30.7814
       GradientMagnitude: 0.0191
    RelativeChangeInBeta: 0.0228
                 FitTime: 56.1280
                 History: []
```

`Mdl` is a trained `RegressionKernel` model, and the structure array `FitInfo` contains optimization details.

Determine how well the trained model generalizes to new predictor values by estimating the resubstitution mean squared error and epsilon-insensitive error.

```
lossMSE = loss(Mdl,Z,Y) % Resubstitution mean squared error
```

```
lossMSE =

  MxNx... tall array
```

```
    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

```
lossEI = loss(Mdl,Z,Y,'LossFun','epsiloninsensitive') % Resubstitution epsilon-insensitive error
```

```
lossEI =

  MxNx... tall array

    ?    ?    ?    ...
    ?    ?    ?    ...
    ?    ?    ?    ...
    :    :    :
    :    :    :
```

Evaluate the tall arrays and bring the results into memory by using `gather`.

```
[lossMSE,lossEI] = gather(lossMSE,lossEI)
```

```
Evaluating tall expression using the Local MATLAB Session:
- Pass 1 of 1: Completed in 1.6 sec
Evaluation completed in 1.9 sec

lossMSE = 2.8851e+03

lossEI = 28.0050
```

Specify Custom Regression Loss

Specify a custom regression loss (Huber loss) for a Gaussian kernel regression model.

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Weight,Cylinders,Horsepower,Model_Year];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]);
X = R(:,1:4);
Y = R(:,end);
```

Reserve 10% of the observations as a holdout sample. Extract the training and test indices from the partition definition.

```
rng(10)  % For reproducibility
N = length(Y);
cvp = cvpartition(N,'Holdout',0.1);
```

```
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Standardize the training data and train the regression kernel model.

```
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training data
tr_sigma(tr_sigma==0) = 1;
Mdl = fitrkernel(Ztrain,Ytrain)
```

```
Mdl =
  RegressionKernel
             ResponseName: 'Y'
                  Learner: 'svm'
    NumExpansionDimensions: 128
               KernelScale: 1
                    Lambda: 0.0028
            BoxConstraint: 1
                  Epsilon: 0.8617


  Properties, Methods
```

`Mdl` is a `RegressionKernel` model.

Create an anonymous function that measures Huber loss ($\delta = 1$), that is,

$$L = \frac{1}{\sum w_j} \sum_{j=1}^{n} w_j \ell_j,$$

where

$$\ell_j = \begin{cases} 0.5\hat{e}_j^2; & |\hat{e}_j| \le 1 \\ |\hat{e}_j| - 0.5; & |\hat{e}_j| > 1 \end{cases}.$$

$\hat{e}_j$ is the residual for observation j. Custom loss functions must be written in a particular form. For rules on writing a custom loss function, see the `'LossFun'` name-value pair argument.

```
huberloss = @(Y,Yhat,W)sum(W.*((0.5*(abs(Y-Yhat)<=1).*(Y-Yhat).^2) + ...
    ((abs(Y-Yhat)>1).*abs(Y-Yhat)-0.5)))/sum(W);
```

Estimate the training set regression loss using the Huber loss function.

```
eTrain = loss(Mdl,Ztrain,Ytrain,'LossFun',huberloss)
```

```
eTrain = 1.7210
```

Standardize the test data using the same mean and standard deviation of the training data columns. Estimate the test set regression loss using the Huber loss function.

```
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Standardize the test data
Ytest = Y(idxTest);
```

```
eTest = loss(Mdl,Ztest,Ytest,'LossFun',huberloss)
```

```
eTest = 1.3062
```

Input Arguments

`Mdl` — Kernel regression model

`RegressionKernel` model object

Kernel regression model, specified as a `RegressionKernel` model object. You can create a `RegressionKernel` model object using `fitrkernel`.

`X` — Predictor data

n-by-*p* numeric matrix

Predictor data, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors. *p* must be equal to the number of predictors used to train `Mdl`.

Data Types: `single` | `double`

`Y` — Response data

numeric vector

Response data, specified as an *n*-dimensional numeric vector. The length of `Y` must be equal to the number of observations in `X` or `Tbl`.

Data Types: `single` | `double`

`Tbl` — Sample data

table

Sample data used to train the model, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName` or `Y`.

If you train `Mdl` using sample data contained in a table, then the input data for `loss` must also be in a table.

`ResponseVarName` — Response variable name

name of variable in `Tbl`

Response variable name, specified as the name of a variable in `Tbl`. The response variable must be a numeric vector. If `Tbl` contains the response variable used to train `Mdl`, then you do not need to specify `ResponseVarName`.

If you specify `ResponseVarName`, then you must specify it as a character vector or string scalar. For example, if the response variable is stored as `Tbl.Y`, then specify `ResponseVarName` as `'Y'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.Y`, as predictors.

Data Types: `char` | `string`

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `L = loss(Mdl,X,Y,'LossFun','epsiloninsensitive','Weights',weights)` returns the weighted regression loss using the epsilon-insensitive loss function.

`LossFun` — Loss function

`'mse'` (default) | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in loss function name or a function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar. Also, in the table, $f(x) = T(x)\beta + b$.
  - x is an observation (row vector) from p predictor variables.
  - $T(\cdot)$ is a transformation of an observation (row vector) for feature expansion. $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$).
  - β is a vector of m coefficients.
  - b is the scalar bias.

Value	Description
`'epsiloninsensitive'`	Epsilon-insensitive loss: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$
`'mse'`	MSE: $\ell[y, f(x)] = [y - f(x)]^2$

  `'epsiloninsensitive'` is appropriate for SVM learners only.
- Specify your own function by using function handle notation.

  Let `n` be the number of observations in `X`. Your function must have this signature:

  `lossvalue = lossfun(Y,Yhat,W)`

  - The output argument `lossvalue` is a scalar.
  - You choose the function name (*`lossfun`*).
  - `Y` is an n-dimensional vector of observed responses. `loss` passes the input argument `Y` in for `Y`.
  - `Yhat` is an n-dimensional vector of predicted responses, which is similar to the output of `predict`.
  - `W` is an `n`-by-1 numeric vector of observation weights.

  Specify your function using `'LossFun',@lossfun`.

Data Types: `char` | `string` | `function_handle`

`Weights` — Observation weights

`ones(size(X,1),1)` (default) | numeric vector | name of variable in `Tbl`

Observation weights, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If `Weights` is a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `Tbl`.
- If `Weights` is the name of a variable in `Tbl`, you must specify `Weights` as a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply the observation weights, `loss` computes the weighted regression loss, that is, the "Weighted Mean Squared Error" on page 33-6819 or "Epsilon-Insensitive Loss Function" on page 33-6820.

`loss` normalizes `Weights` to sum to 1.

Data Types: `double` | `single` | `char` | `string`

Output Arguments

L — Regression loss

numeric scalar

Regression loss, returned as a numeric scalar. The interpretation of `L` depends on `Weights` and `LossFun`. For example, if you use the default observation weights and specify `'epsiloninsensitive'` as the loss function, then `L` is the epsilon-insensitive loss.

More About

Weighted Mean Squared Error

The weighted mean squared error is calculated as follows:

$$\text{mse} = \frac{\sum_{j=1}^{n} w_j \left(f\left(x_j\right) - y_j\right)^2}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of observations.
- x_j is the jth observation (row of predictor data).
- y_j is the observed response to x_j.
- $f(x_j)$ is the response prediction of the Gaussian kernel regression model `Mdl` to x_j.
- w is the vector of observation weights.

Each observation weight in w is equal to `ones(`n`,1)/`n by default. You can specify different values for the observation weights by using the `'Weights'` name-value pair argument. `loss` normalizes `Weights` to sum to 1.

Epsilon-Insensitive Loss Function

The epsilon-insensitive loss function ignores errors that are within the distance epsilon (ε) of the function value. The function is formally described as:

$$Loss_{\varepsilon} = \begin{cases} 0\,,\ if\ |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon\,,\ otherwise\,. \end{cases}$$

The mean epsilon-insensitive loss is calculated as follows:

$$Loss = \frac{\sum_{j=1}^{n} w_j \max\left(0, \left|y_j - f\left(x_j\right)\right| - \varepsilon\right)}{\sum_{j=1}^{n} w_j},$$

where:

- n is the number of observations.
- x_j is the jth observation (row of predictor data).
- y_j is the observed response to x_j.
- $f(x_j)$ is the response prediction of the Gaussian kernel regression model `Mdl` to x_j.
- w is the vector of observation weights.

Each observation weight in w is equal to `ones(`n`,1)/`n by default. You can specify different values for the observation weights by using the `'Weights'` name-value pair argument. `loss` normalizes `Weights` to sum to 1.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `loss` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`RegressionKernel` | `fitrkernel` | `predict` | `resume`

Introduced in R2018a

predict

Predict responses for Gaussian kernel regression model

Syntax

```
YFit = predict(Mdl,X)
```

Description

`YFit = predict(Mdl,X)` returns a vector of predicted responses for the predictor data in the matrix or table `X`, based on the binary Gaussian kernel regression model `Mdl`.

Examples

Predict Test Set Responses

Predict the test set responses using a Gaussian kernel regression model for the `carbig` data set.

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Weight,Cylinders,Horsepower,Model_Year];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]);
X = R(:,1:4);
Y = R(:,end);
```

Reserve 10% of the observations as a holdout sample. Extract the training and test indices from the partition definition.

```
rng(10)  % For reproducibility
N = length(Y);
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Standardize the training data and train the regression kernel model.

```
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training data
tr_sigma(tr_sigma==0) = 1;
Mdl = fitrkernel(Ztrain,Ytrain)
```

```
Mdl =
  RegressionKernel
             ResponseName: 'Y'
                  Learner: 'svm'
    NumExpansionDimensions: 128
               KernelScale: 1
                    Lambda: 0.0028
            BoxConstraint: 1
                  Epsilon: 0.8617


  Properties, Methods
```

`Mdl` is a `RegressionKernel` model.

Standardize the test data using the same mean and standard deviation of the training data columns. Predict responses for the test set.

```
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Standardize the test data
Ytest = Y(idxTest);

YFit = predict(Mdl,Ztest);
```

Create a table containing the first 10 observed response values and predicted response values.

```
table(Ytest(1:10),YFit(1:10),'VariableNames', ...
    {'ObservedValue','PredictedValue'})
```

```
ans=10×2 table
    ObservedValue    PredictedValue
    _____________    ______________

          18              17.616
          14              25.799
          24              24.141
          25              25.018
          14              13.637
          14              14.557
          18              18.584
          27              26.096
          21              25.031
          13              13.324
```

Estimate the test set regression loss using the mean squared error loss function.

```
L = loss(Mdl,Ztest,Ytest)
```

```
L = 9.2664
```

Input Arguments

`Mdl` — Kernel regression model

`RegressionKernel` model object

Kernel regression model, specified as a `RegressionKernel` model object. You can create a `RegressionKernel` model object using `fitrkernel`.

`X` — Predictor data used to generate responses
numeric matrix | table

Predictor data used to generate responses, specified as a numeric matrix or table.

Each row of `X` corresponds to one observation, and each column corresponds to one variable.

- For a numeric matrix:
  - The variables in the columns of `X` must have the same order as the predictor variables that trained `Mdl`.
  - If you trained `Mdl` using a table (for example, `Tbl`) and `Tbl` contains all numeric predictor variables, then `X` can be a numeric matrix. To treat numeric predictors in `Tbl` as categorical during training, identify categorical predictors using the `CategoricalPredictors` name-value pair argument of `fitrkernel`. If `Tbl` contains heterogeneous predictor variables (for example, numeric and categorical data types) and `X` is a numeric matrix, then `predict` throws an error.
- For a table:
  - `predict` does not support multicolumn variables or cell arrays other than cell arrays of character vectors.
  - If you trained `Mdl` using a table (for example, `Tbl`), then all predictor variables in `X` must have the same variable names and data types as those that trained `Mdl` (stored in `Mdl.PredictorNames`). However, the column order of `X` does not need to correspond to the column order of `Tbl`. Also, `Tbl` and `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.
  - If you trained `Mdl` using a numeric matrix, then the predictor names in `Mdl.PredictorNames` and corresponding predictor variable names in `X` must be the same. To specify predictor names during training, see the `PredictorNames` name-value pair argument of `fitrkernel`. All predictor variables in `X` must be numeric vectors. `X` can contain additional variables (response variables, observation weights, and so on), but `predict` ignores them.

Data Types: `double` | `single` | `table`

Output Arguments

`YFit` — Predicted responses
numeric vector

Predicted responses, returned as a numeric vector.

`YFit` is an *n*-by-*1* vector of the same data type as the response data (`Y`) used to train `Mdl`, where *n* is the number of observations in `X`.

Extended Capabilities

Tall Arrays
Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `predict` does not support tall `table` data.

For more information, see “Tall Arrays”.

See Also

`RegressionKernel` | `fitrkernel` | `loss` | `resume`

Introduced in R2018a

resume

Resume training of Gaussian kernel regression model

Syntax

```
UpdatedMdl = resume(Mdl,X,Y)

UpdatedMdl = resume(Mdl,Tbl,ResponseVarName)
UpdatedMdl = resume(Mdl,Tbl,Y)

UpdatedMdl = resume(___,Name,Value)
[UpdatedMdl,FitInfo] = resume(___)
```

Description

`UpdatedMdl = resume(Mdl,X,Y)` continues training with the same options used to train `Mdl`, including the training data (predictor data in `X` and response data in `Y`) and the feature expansion. The training starts at the current estimated parameters in `Mdl`. The function returns a new Gaussian kernel regression model `UpdatedMdl`.

`UpdatedMdl = resume(Mdl,Tbl,ResponseVarName)` continues training with the predictor data in `Tbl` and the true responses in `Tbl.ResponseVarName`.

`UpdatedMdl = resume(Mdl,Tbl,Y)` continues training with the predictor data in table `Tbl` and the true responses in `Y`.

`UpdatedMdl = resume(___,Name,Value)` specifies options using one or more name-value pair arguments in addition to any of the input argument combinations in previous syntaxes. For example, you can modify convergence control options, such as convergence tolerances and the maximum number of additional optimization iterations.

`[UpdatedMdl,FitInfo] = resume(___)` also returns the fit information in the structure array `FitInfo`.

Examples

Estimate Sample Loss and Resume Training

Resume training a Gaussian kernel regression model for more iterations to improve the regression loss.

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:5);
Y = R(:,end);
```

Reserve 10% of the observations as a holdout sample. Extract the training and test indices from the partition definition.

```
rng(10)  % For reproducibility
N = length(Y);
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Standardize the training data and train a kernel regression model. Set the iteration limit to 5 and specify `'Verbose',1` to display diagnostic information.

```
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training data
tr_sigma(tr_sigma==0) = 1;
Mdl = fitrkernel(Ztrain,Ytrain,'IterationLimit',5,'Verbose',1)
```

```
|=============================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================
|  LBFGS |      1 |            0 |  5.691016e+00 |  0.000000e+00 |  5.852758e-02 |
|  LBFGS |      1 |            1 |  5.086537e+00 |  8.000000e+00 |  5.220869e-02 |   9.846711e-02
|  LBFGS |      1 |            2 |  3.862301e+00 |  5.000000e-01 |  3.796034e-01 |   5.998808e-01
|  LBFGS |      1 |            3 |  3.460613e+00 |  1.000000e+00 |  3.257790e-01 |   1.615091e-01
|  LBFGS |      1 |            4 |  3.136228e+00 |  1.000000e+00 |  2.832861e-02 |   8.006254e-02
|  LBFGS |      1 |            5 |  3.063978e+00 |  1.000000e+00 |  1.475038e-02 |   3.314455e-02
|=============================================================================================

Mdl =
  RegressionKernel
             ResponseName: 'Y'
                  Learner: 'svm'
    NumExpansionDimensions: 256
               KernelScale: 1
                    Lambda: 0.0028
            BoxConstraint: 1
                  Epsilon: 0.8617


  Properties, Methods
```

`Mdl` is a `RegressionKernel` model.

Standardize the test data using the same mean and standard deviation of the training data columns. Estimate the epsilon-insensitive error for the test set.

```
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Standardize the test data
```

```
Ytest = Y(idxTest);

L = loss(Mdl,Ztest,Ytest,'LossFun','epsiloninsensitive')

L = 2.0674
```

Continue training the model by using `resume`. This function continues training with the same options used for training `Mdl`.

```
UpdatedMdl = resume(Mdl,Ztrain,Ytrain);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient    |    Relative
|        |        |               |               |               |   magnitude    | change in Beta
|=============================================================================================================
|  LBFGS |      1 |             0 |  3.063978e+00 |  0.000000e+00 |  1.475038e-02 |
|  LBFGS |      1 |             1 |  3.007822e+00 |  8.000000e+00 |  1.391637e-02 |   2.603966e-02
|  LBFGS |      1 |             2 |  2.817171e+00 |  5.000000e-01 |  5.949008e-02 |   1.918084e-01
|  LBFGS |      1 |             3 |  2.807294e+00 |  2.500000e-01 |  6.798867e-02 |   2.973097e-02
|  LBFGS |      1 |             4 |  2.791060e+00 |  1.000000e+00 |  2.549575e-02 |   1.639328e-02
|  LBFGS |      1 |             5 |  2.767821e+00 |  1.000000e+00 |  6.154419e-03 |   2.468903e-02
|  LBFGS |      1 |             6 |  2.738163e+00 |  1.000000e+00 |  5.949008e-02 |   9.476263e-02
|  LBFGS |      1 |             7 |  2.719146e+00 |  1.000000e+00 |  1.699717e-02 |   1.849972e-02
|  LBFGS |      1 |             8 |  2.705941e+00 |  1.000000e+00 |  3.116147e-02 |   4.152590e-02
|  LBFGS |      1 |             9 |  2.701162e+00 |  1.000000e+00 |  5.665722e-03 |   9.401466e-03
|  LBFGS |      1 |            10 |  2.695341e+00 |  5.000000e-01 |  3.116147e-02 |   4.968046e-02
|  LBFGS |      1 |            11 |  2.691277e+00 |  1.000000e+00 |  8.498584e-03 |   1.017446e-02
|  LBFGS |      1 |            12 |  2.689972e+00 |  1.000000e+00 |  1.983003e-02 |   9.938921e-03
|  LBFGS |      1 |            13 |  2.688979e+00 |  1.000000e+00 |  1.416431e-02 |   6.606316e-03
|  LBFGS |      1 |            14 |  2.687787e+00 |  1.000000e+00 |  1.621956e-03 |   7.089542e-03
|  LBFGS |      1 |            15 |  2.686539e+00 |  1.000000e+00 |  1.699717e-02 |   1.169701e-02
|  LBFGS |      1 |            16 |  2.685356e+00 |  1.000000e+00 |  1.133144e-02 |   1.069310e-02
|  LBFGS |      1 |            17 |  2.685021e+00 |  5.000000e-01 |  1.133144e-02 |   2.104248e-02
|  LBFGS |      1 |            18 |  2.684002e+00 |  1.000000e+00 |  2.832861e-03 |   6.175231e-03
|  LBFGS |      1 |            19 |  2.683507e+00 |  1.000000e+00 |  5.665722e-03 |   3.724026e-03
|  LBFGS |      1 |            20 |  2.683343e+00 |  5.000000e-01 |  5.665722e-03 |   9.549119e-03
|=============================================================================================================
| Solver |  Pass  |   Iteration   |   Objective   |     Step      |    Gradient    |    Relative
|        |        |               |               |               |   magnitude    | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            21 |  2.682897e+00 |  1.000000e+00 |  5.665722e-03 |   7.172867e-03
|  LBFGS |      1 |            22 |  2.682682e+00 |  1.000000e+00 |  2.832861e-03 |   2.587726e-03
|  LBFGS |      1 |            23 |  2.682485e+00 |  1.000000e+00 |  2.832861e-03 |   2.953648e-03
|  LBFGS |      1 |            24 |  2.682326e+00 |  1.000000e+00 |  2.832861e-03 |   7.777294e-03
|  LBFGS |      1 |            25 |  2.681914e+00 |  1.000000e+00 |  2.832861e-03 |   2.778555e-03
|  LBFGS |      1 |            26 |  2.681867e+00 |  5.000000e-01 |  1.031085e-03 |   3.638352e-03
|  LBFGS |      1 |            27 |  2.681725e+00 |  1.000000e+00 |  5.665722e-03 |   1.515199e-03
|  LBFGS |      1 |            28 |  2.681692e+00 |  5.000000e-01 |  1.314940e-03 |   1.850055e-03
|  LBFGS |      1 |            29 |  2.681625e+00 |  1.000000e+00 |  2.832861e-03 |   1.456903e-03
|  LBFGS |      1 |            30 |  2.681594e+00 |  5.000000e-01 |  2.832861e-03 |   8.704875e-04
|  LBFGS |      1 |            31 |  2.681581e+00 |  5.000000e-01 |  8.498584e-03 |   3.934768e-04
|  LBFGS |      1 |            32 |  2.681579e+00 |  1.000000e+00 |  8.498584e-03 |   1.847866e-03
|  LBFGS |      1 |            33 |  2.681553e+00 |  1.000000e+00 |  9.857038e-04 |   6.509825e-04
|  LBFGS |      1 |            34 |  2.681541e+00 |  5.000000e-01 |  8.498584e-03 |   6.635528e-04
|  LBFGS |      1 |            35 |  2.681499e+00 |  1.000000e+00 |  5.665722e-03 |   6.194735e-04
|  LBFGS |      1 |            36 |  2.681493e+00 |  5.000000e-01 |  1.133144e-02 |   1.617763e-03
|  LBFGS |      1 |            37 |  2.681473e+00 |  1.000000e+00 |  9.869233e-04 |   8.418484e-04
|  LBFGS |      1 |            38 |  2.681469e+00 |  1.000000e+00 |  5.665722e-03 |   1.069722e-03
```

```
|  LBFGS |      1 |           39 |  2.681432e+00 |  1.000000e+00 |  2.832861e-03 |   8.501930e-04
|  LBFGS |      1 |           40 |  2.681423e+00 |  2.500000e-01 |  1.133144e-02 |   9.543716e-04
|========================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |      Step     |    Gradient   |     Relative
|        |        |              |               |               |   magnitude   | change in Beta
|========================================================================================================
|  LBFGS |      1 |           41 |  2.681416e+00 |  1.000000e+00 |  2.832861e-03 |   8.763251e-04
|  LBFGS |      1 |           42 |  2.681413e+00 |  5.000000e-01 |  2.832861e-03 |   4.101888e-04
|  LBFGS |      1 |           43 |  2.681403e+00 |  1.000000e+00 |  5.665722e-03 |   2.713209e-04
|  LBFGS |      1 |           44 |  2.681392e+00 |  1.000000e+00 |  2.832861e-03 |   2.115241e-04
|  LBFGS |      1 |           45 |  2.681383e+00 |  1.000000e+00 |  2.832861e-03 |   2.872858e-04
|  LBFGS |      1 |           46 |  2.681374e+00 |  1.000000e+00 |  8.498584e-03 |   5.771001e-04
|  LBFGS |      1 |           47 |  2.681353e+00 |  1.000000e+00 |  2.832861e-03 |   3.160871e-04
|  LBFGS |      1 |           48 |  2.681334e+00 |  5.000000e-01 |  8.498584e-03 |   1.045502e-03
|  LBFGS |      1 |           49 |  2.681314e+00 |  1.000000e+00 |  7.878714e-04 |   1.505118e-03
|  LBFGS |      1 |           50 |  2.681306e+00 |  1.000000e+00 |  2.832861e-03 |   4.756894e-04
|  LBFGS |      1 |           51 |  2.681301e+00 |  1.000000e+00 |  1.133144e-02 |   3.664873e-04
|  LBFGS |      1 |           52 |  2.681288e+00 |  1.000000e+00 |  2.832861e-03 |   1.449821e-04
|  LBFGS |      1 |           53 |  2.681287e+00 |  2.500000e-01 |  1.699717e-02 |   2.357176e-04
|  LBFGS |      1 |           54 |  2.681282e+00 |  1.000000e+00 |  5.665722e-03 |   2.046663e-04
|  LBFGS |      1 |           55 |  2.681278e+00 |  1.000000e+00 |  2.832861e-03 |   2.546349e-04
|  LBFGS |      1 |           56 |  2.681276e+00 |  2.500000e-01 |  1.307940e-03 |   1.966786e-04
|  LBFGS |      1 |           57 |  2.681274e+00 |  5.000000e-01 |  1.416431e-02 |   1.005310e-04
|  LBFGS |      1 |           58 |  2.681271e+00 |  5.000000e-01 |  1.118892e-03 |   1.147324e-04
|  LBFGS |      1 |           59 |  2.681269e+00 |  1.000000e+00 |  2.832861e-03 |   1.332914e-04
|  LBFGS |      1 |           60 |  2.681268e+00 |  2.500000e-01 |  1.132045e-03 |   5.441369e-05
|========================================================================================================
```

Estimate the epsilon-insensitive error for the test set using the updated model.

```
UpdatedL = loss(UpdatedMdl,Ztest,Ytest,'LossFun','epsiloninsensitive')
```

```
UpdatedL = 1.8933
```

The regression error decreases by a factor of about `0.08` after `resume` updates the regression model with more iterations.

Resume Training with Modified Convergence Control Training Options

Load the `carbig` data set.

```
load carbig
```

Specify the predictor variables (`X`) and the response variable (`Y`).

```
X = [Acceleration,Cylinders,Displacement,Horsepower,Weight];
Y = MPG;
```

Delete rows of `X` and `Y` where either array has `NaN` values. Removing rows with `NaN` values before passing data to `fitrkernel` can speed up training and reduce memory usage.

```
R = rmmissing([X Y]); % Data with missing entries removed
X = R(:,1:5);
Y = R(:,end);
```

Reserve 10% of the observations as a holdout sample. Extract the training and test indices from the partition definition.

```
rng(10)  % For reproducibility
N = length(Y);
cvp = cvpartition(N,'Holdout',0.1);
idxTrn = training(cvp); % Training set indices
idxTest = test(cvp);    % Test set indices
```

Standardize the training data and train a kernel regression model with relaxed convergence control training options by using the name-value pair arguments `'BetaTolerance'` and `'GradientTolerance'`. Specify `'Verbose',1` to display diagnostic information.

```
Xtrain = X(idxTrn,:);
Ytrain = Y(idxTrn);
[Ztrain,tr_mu,tr_sigma] = zscore(Xtrain); % Standardize the training data
tr_sigma(tr_sigma==0) = 1;
[Mdl,FitInfo] = fitrkernel(Ztrain,Ytrain,'Verbose',1, ...
    'BetaTolerance',2e-2,'GradientTolerance',2e-2);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            0 |  5.691016e+00 |  0.000000e+00 |  5.852758e-02 |
|  LBFGS |      1 |            1 |  5.086537e+00 |  8.000000e+00 |  5.220869e-02 |   9.846711e-02
|  LBFGS |      1 |            2 |  3.862301e+00 |  5.000000e-01 |  3.796034e-01 |   5.998808e-01
|  LBFGS |      1 |            3 |  3.460613e+00 |  1.000000e+00 |  3.257790e-01 |   1.615091e-01
|  LBFGS |      1 |            4 |  3.136228e+00 |  1.000000e+00 |  2.832861e-02 |   8.006254e-02
|  LBFGS |      1 |            5 |  3.063978e+00 |  1.000000e+00 |  1.475038e-02 |   3.314455e-02
|=============================================================================================================
```

`Mdl` is a `RegressionKernel` model.

Standardize the test data using the same mean and standard deviation of the training data columns. Estimate the epsilon-insensitive error for the test set.

```
Xtest = X(idxTest,:);
Ztest = (Xtest-tr_mu)./tr_sigma; % Standardize the test data
Ytest = Y(idxTest);

L = loss(Mdl,Ztest,Ytest,'LossFun','epsiloninsensitive')
```

```
L = 2.0674
```

Continue training the model by using `resume` with modified convergence control options.

```
[UpdatedMdl,UpdatedFitInfo] = resume(Mdl,Ztrain,Ytrain, ...
    'BetaTolerance',2e-3,'GradientTolerance',2e-3);
```

```
|=============================================================================================================
| Solver |  Pass  |   Iteration  |   Objective   |     Step      |    Gradient   |    Relative
|        |        |              |               |               |   magnitude   | change in Beta
|=============================================================================================================
|  LBFGS |      1 |            0 |  3.063978e+00 |  0.000000e+00 |  1.475038e-02 |
|  LBFGS |      1 |            1 |  3.007822e+00 |  8.000000e+00 |  1.391637e-02 |   2.603966e-02
|  LBFGS |      1 |            2 |  2.817171e+00 |  5.000000e-01 |  5.949008e-02 |   1.918084e-01
|  LBFGS |      1 |            3 |  2.807294e+00 |  2.500000e-01 |  6.798867e-02 |   2.973097e-02
```

```
|  LBFGS |      1 |            4 |  2.791060e+00 |  1.000000e+00 |  2.549575e-02 |   1.639328e-02
|  LBFGS |      1 |            5 |  2.767821e+00 |  1.000000e+00 |  6.154419e-03 |   2.468903e-02
|  LBFGS |      1 |            6 |  2.738163e+00 |  1.000000e+00 |  5.949008e-02 |   9.476263e-02
|  LBFGS |      1 |            7 |  2.719146e+00 |  1.000000e+00 |  1.699717e-02 |   1.849972e-02
|  LBFGS |      1 |            8 |  2.705941e+00 |  1.000000e+00 |  3.116147e-02 |   4.152590e-02
|  LBFGS |      1 |            9 |  2.701162e+00 |  1.000000e+00 |  5.665722e-03 |   9.401466e-03
|  LBFGS |      1 |           10 |  2.695341e+00 |  5.000000e-01 |  3.116147e-02 |   4.968046e-02
|  LBFGS |      1 |           11 |  2.691277e+00 |  1.000000e+00 |  8.498584e-03 |   1.017446e-02
|  LBFGS |      1 |           12 |  2.689972e+00 |  1.000000e+00 |  1.983003e-02 |   9.938921e-03
|  LBFGS |      1 |           13 |  2.688979e+00 |  1.000000e+00 |  1.416431e-02 |   6.606316e-03
|  LBFGS |      1 |           14 |  2.687787e+00 |  1.000000e+00 |  1.621956e-03 |   7.089542e-03
|=========================================================================================================
```

Estimate the epsilon-insensitive error for the test set using the updated model.

```
UpdatedL = loss(UpdatedMdl,Ztest,Ytest,'LossFun','epsiloninsensitive')
```

```
UpdatedL = 1.8891
```

The regression error decreases after `resume` updates the regression model with smaller convergence tolerances.

Display the outputs `FitInfo` and `UpdatedFitInfo`.

```
FitInfo
```

```
FitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'epsiloninsensitive'
                  Lambda: 0.0028
           BetaTolerance: 0.0200
       GradientTolerance: 0.0200
          ObjectiveValue: 3.0640
       GradientMagnitude: 0.0148
    RelativeChangeInBeta: 0.0331
                 FitTime: 0.0431
                 History: [1x1 struct]
```

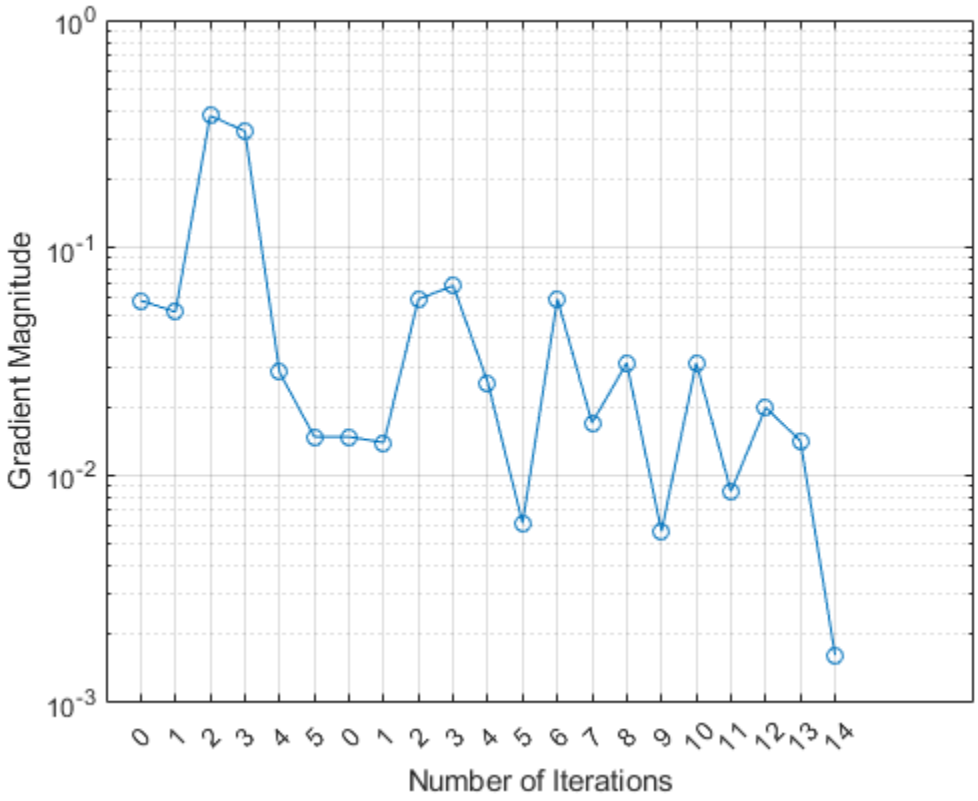
```
UpdatedFitInfo
```

```
UpdatedFitInfo = struct with fields:
                  Solver: 'LBFGS-fast'
            LossFunction: 'epsiloninsensitive'
                  Lambda: 0.0028
           BetaTolerance: 0.0020
       GradientTolerance: 0.0020
          ObjectiveValue: 2.6878
       GradientMagnitude: 0.0016
    RelativeChangeInBeta: 0.0071
                 FitTime: 0.0631
                 History: [1x1 struct]
```

Both trainings terminate because the software satisfies the absolute gradient tolerance.

Plot the gradient magnitude versus the number of iterations by using `UpdatedFitInfo.History.GradientMagnitude`. Note that the `History` field of `UpdatedFitInfo` includes the information in the `History` field of `FitInfo`.

```
semilogy(UpdatedFitInfo.History.GradientMagnitude,'o-')
ax = gca;
ax.XTick = 1:21;
ax.XTickLabel = UpdatedFitInfo.History.IterationNumber;
grid on
xlabel('Number of Iterations')
ylabel('Gradient Magnitude')
```



The first training terminates after five iterations because the gradient magnitude becomes less than `2e-2`. The second training terminates after 14 iterations because the gradient magnitude becomes less than `2e-3`.

Input Arguments

`Mdl` — Kernel regression model

`RegressionKernel` model object

Kernel regression model, specified as a `RegressionKernel` model object. You can create a `RegressionKernel` model object using `fitrkernel`.

`X` — Predictor data used to train `Mdl`

n-by-*p* numeric matrix

Predictor data used to train `Mdl`, specified as an *n*-by-*p* numeric matrix, where *n* is the number of observations and *p* is the number of predictors.

Data Types: `single` | `double`

Y — Response data used to train `Mdl`
numeric vector

Response data used to train `Mdl`, specified as a numeric vector.

Data Types: `double` | `single`

`Tbl` — Sample data used to train `Mdl`
table

Sample data used to train `Mdl`, specified as a table. Each row of `Tbl` corresponds to one observation, and each column corresponds to one predictor variable. Optionally, `Tbl` can contain additional columns for the response variable and observation weights. `Tbl` must contain all of the predictors used to train `Mdl`. Multicolumn variables and cell arrays other than cell arrays of character vectors are not allowed.

If you trained `Mdl` using sample data contained in a table, then the input data for `resume` must also be in a table.

`ResponseVarName` — Name of response variable used to train `Mdl`
name of variable in `Tbl`

Name of the response variable used to train `Mdl`, specified as the name of a variable in `Tbl`. The `ResponseVarName` value must match the name `Mdl.ResponseName`.

Data Types: `char` | `string`

Note `resume` should run only on the same training data and observation weights (`Weights`) used to train `Mdl`. The `resume` function uses the same training options, such as feature expansion, used to train `Mdl`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `UpdatedMdl = resume(Mdl,X,Y,'BetaTolerance',1e-3)` resumes training with the same options used to train `Mdl`, except the relative tolerance on the linear coefficients and the bias term.

`Weights` — Observation weights used to train `Mdl`
numeric vector | name of variable in `Tbl`

Observation weights used to train `Mdl`, specified as the comma-separated pair consisting of `'Weights'` and a numeric vector or the name of a variable in `Tbl`.

- If `Weights` is a numeric vector, then the size of `Weights` must be equal to the number of rows in `X` or `Tbl`.
- If `Weights` is the name of a variable in `Tbl`, you must specify `Weights` as a character vector or string scalar. For example, if the weights are stored as `Tbl.W`, then specify `Weights` as `'W'`. Otherwise, the software treats all columns of `Tbl`, including `Tbl.W`, as predictors.

If you supply the observation weights, `resume` normalizes `Weights` to sum to 1.

Data Types: `double` | `single` | `char` | `string`

`BetaTolerance` — Relative tolerance on linear coefficients and bias term
`BetaTolerance` value used to train `Mdl` (default) | nonnegative scalar

Relative tolerance on the linear coefficients and the bias term (intercept), specified as the comma-separated pair consisting of `'BetaTolerance'` and a nonnegative scalar.

Let $B_t = [\beta_t' \ b_t]$, that is, the vector of the coefficients and the bias term at optimization iteration t. If $\left\| \frac{B_t - B_{t-1}}{B_t} \right\|_2 <$ BetaTolerance, then optimization terminates.

If you also specify `GradientTolerance`, then optimization terminates when the software satisfies either stopping criterion.

By default, the value is the same `BetaTolerance` value used to train `Mdl`.

Example: `'BetaTolerance',1e-6`

Data Types: `single` | `double`

`GradientTolerance` — Absolute gradient tolerance
`GradientTolerance` value used to train `Mdl` (default) | nonnegative scalar

Absolute gradient tolerance, specified as the comma-separated pair consisting of `'GradientTolerance'` and a nonnegative scalar.

Let $\nabla \mathscr{L}_t$ be the gradient vector of the objective function with respect to the coefficients and bias term at optimization iteration t. If $\| \nabla \mathscr{L}_t \|_\infty = \max |\nabla \mathscr{L}_t| <$ GradientTolerance, then optimization terminates.

If you also specify `BetaTolerance`, then optimization terminates when the software satisfies either stopping criterion.

By default, the value is the same `GradientTolerance` value used to train `Mdl`.

Example: `'GradientTolerance',1e-5`

Data Types: `single` | `double`

`IterationLimit` — Maximum number of additional optimization iterations
positive integer

Maximum number of additional optimization iterations, specified as the comma-separated pair consisting of `'IterationLimit'` and a positive integer.

The default value is 1000 if the transformed data fits in memory (`Mdl.ModelParameters.BlockSize`), which you specify by using the `'BlockSize'` name-value pair argument when training `Mdl` with `fitrkernel`. Otherwise, the default value is 100.

Note that the default value is not the value used to train `Mdl`.

Example: `'IterationLimit',500`

Data Types: `single` | `double`

Output Arguments

`UpdatedMdl` — Updated kernel regression model

`RegressionKernel` model object

Updated kernel regression model, returned as a `RegressionKernel` model object.

`FitInfo` — Optimization details

structure array

Optimization details, returned as a structure array including fields described in this table. The fields contain final values or name-value pair argument specifications.

Field	Description
`Solver`	Objective function minimization technique: `'LBFGS-fast'`, `'LBFGS-blockwise'`, or `'LBFGS-tall'`. For details, see the "Algorithms" on page 33-6801 section of `fitrkernel`.
`LossFunction`	Loss function. Either mean squared error (MSE) or epsilon-insensitive, depending on the type of linear regression model. See `Learner` of `fitrkernel`.
`Lambda`	Regularization term strength. See `Lambda` of `fitrkernel`.
`BetaTolerance`	Relative tolerance on the linear coefficients and the bias term. See `BetaTolerance`.
`GradientTolerance`	Absolute gradient tolerance. See `GradientTolerance`.
`ObjectiveValue`	Value of the objective function when optimization terminates. The regression loss plus the regularization term compose the objective function.
`GradientMagnitude`	Infinite norm of the gradient vector of the objective function when optimization terminates. See `GradientTolerance`.
`RelativeChangeInBeta`	Relative changes in the linear coefficients and the bias term when optimization terminates. See `BetaTolerance`.
`FitTime`	Elapsed, wall-clock time (in seconds) required to fit the model to the data.
`History`	History of optimization information. This field also includes the optimization information from training `Mdl`. This field is empty (`[]`) if you specify `'Verbose',0` when training `Mdl`. For details, see `Verbose` and the "Algorithms" on page 33-6801 section of `fitrkernel`.

To access fields, use dot notation. For example, to access the vector of objective function values for each iteration, enter `FitInfo.ObjectiveValue` in the Command Window.

Examine the information provided by `FitInfo` to assess whether convergence is satisfactory.

More About

Random Feature Expansion

Random feature expansion, such as Random Kitchen Sinks[1] and Fastfood[2], is a scheme to approximate Gaussian kernels of the kernel regression algorithm for big data in a computationally

efficient way. Random feature expansion is more practical for big data applications that have large training sets but can also be applied to smaller data sets that fit in memory.

The kernel regression algorithm searches for an optimal function that deviates from each response data point (y_i) by values no greater than the epsilon margin (ε) after mapping the predictor data into a high-dimensional space.

Some regression problems cannot be described adequately using a linear model. In such cases, obtain a nonlinear regression model by replacing the dot product x_1x_2' with a nonlinear kernel function $G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle$, where x_i is the ith observation (row vector) and $\varphi(x_i)$ is a transformation that maps x_i to a high-dimensional space (called the "kernel trick"). However, evaluating $G(x_1,x_2)$, the Gram matrix, for each pair of observations is computationally expensive for a large data set (large n).

The random feature expansion scheme finds a random transformation so that its dot product approximates the Gaussian kernel. That is,

$$G(x_1, x_2) = \langle\varphi(x_1), \varphi(x_2)\rangle \approx T(x_1)T(x_2)',$$

where $T(x)$ maps x in $\mathbb{R}^p$ to a high-dimensional space ($\mathbb{R}^m$). The Random Kitchen Sink[1] scheme uses the random transformation

$$T(x) = m^{-1/2}\exp(iZx')',$$

where $Z \in \mathbb{R}^{m \times p}$ is a sample drawn from $N\left(0, \sigma^{-2}\right)$ and σ^2 is a kernel scale. This scheme requires $O(mp)$ computation and storage. The Fastfood[2] scheme introduces another random basis V instead of Z using Hadamard matrices combined with Gaussian scaling matrices. This random basis reduces computation cost to $O(m\log p)$ and reduces storage to $O(m)$.

You can specify values for m and σ^2, using the `NumExpansionDimensions` and `KernelScale` name-value pair arguments of `fitrkernel`, respectively.

The `fitrkernel` function uses the Fastfood scheme for random feature expansion and uses linear regression to train a Gaussian kernel regression model. Unlike solvers in the `fitrsvm` function, which require computation of the n-by-n Gram matrix, the solver in `fitrkernel` only needs to form a matrix of size n-by-m, with m typically much less than n for big data.

Extended Capabilities

Tall Arrays

Calculate with arrays that have more rows than fit in memory.

Usage notes and limitations:

- `resume` does not support tall `table` data.
- The default value for the `'IterationLimit'` name-value pair argument is relaxed to 20 when you work with tall arrays.
- `resume` uses a block-wise strategy. For details, see the "Algorithms" on page 33-6801 section of `fitrkernel`.

For more information, see "Tall Arrays".

See Also

`RegressionKernel` | `fitrkernel` | `loss` | `predict`

Introduced in R2018a

RegressionPartitionedKernel

Cross-validated kernel model for regression

Description

`RegressionPartitionedKernel` is a set of kernel regression models trained on cross-validated folds. To obtain a cross-validated, kernel regression model, use `fitrkernel` and specify one of the cross-validation options. You can estimate the predictive quality of the model, or how well the linear regression model generalizes, using one or more of these "kfold" methods: `kfoldPredict` and `kfoldLoss`.

Every "kfold" method uses models trained on *training-fold* observations to predict the response for *validation-fold* observations. For example, suppose that you cross-validate using five folds. In this case, the software randomly assigns each observation into five groups of equal size (roughly). The training fold contains four of the groups (that is, roughly 4/5 of the data) and the validation fold contains the other group (that is, roughly 1/5 of the data). In this case, cross-validation proceeds as follows:

1. The software trains the first model (stored in `CVMdl.Trained{1}`) using the observations in the last four groups and reserves the observations in the first group for validation.
2. The software trains the second model (stored in `CVMdl.Trained{2}`) using the observations in the first group and the last three groups. The software reserves the observations in the second group for validation.
3. The software proceeds in a similar fashion for the third through the fifth models.

If you validate by calling `kfoldPredict`, it computes predictions for the observations in group 1 using the first model, group 2 for the second model, and so on. In short, the software estimates a response for every observation using the model trained without that observation.

Note `RegressionPartitionedKernel` model objects do not store the predictor data set.

Creation

Create a `RegressionPartitionedKernel` object using the `fitrkernel` function. Use one of the `'CrossVal'`, `'CVPartition'`, `'Holdout'`, `'KFold'`, or `'Leaveout'` name-value pair arguments in the call to `fitrkernel`. For details, see the `fitrkernel` function reference page.

Properties

Cross-Validation Properties

`CrossValidatedModel` — Cross-validated model name
character vector

This property is read-only.

Cross-validated model name, specified as a character vector.

For example, `'Kernel'` specifies a cross-validated kernel model.

Data Types: `char`

`KFold` — Number of cross-validated folds
positive integer scalar

This property is read-only.

Number of cross-validated folds, specified as a positive integer scalar.

Data Types: `double`

`ModelParameters` — Cross-validation parameter values
object

This property is read-only.

Cross-validation parameter values, specified as an object. The parameter values correspond to the name-value pair argument values used to cross-validate the kernel regression model. `ModelParameters` does not contain estimated parameters.

`NumObservations` — Number of observations
positive numeric scalar

This property is read-only.

Number of observations in the training data, specified as a positive numeric scalar.

Data Types: `double`

`Partition` — Data partition
`cvpartition` model

This property is read-only.

Data partition indicating how the software splits the data into cross-validation folds, specified as a `cvpartition` model.

`Trained` — Kernel regression models trained on cross-validation folds
cell array of `RegressionKernel` models

This property is read-only.

Kernel regression models trained on cross-validation folds, specified as a cell array of `RegressionKernel` models. `Trained` has k cells, where k is the number of folds.

Data Types: `cell`

`W` — Observation weights
numeric vector

This property is read-only.

Observation weights used to cross-validate the model, specified as a numeric vector. `W` has `NumObservations` elements.

The software normalizes the weights used for training so that `sum(W,'omitnan')` is `1`.

Data Types: `single` | `double`

`Y` — Observed response values
numeric vector

This property is read-only.

Observed response values used to cross-validate the model, specified as a numeric vector. `Y` has `NumObservations` elements.

Each row of `Y` represents the observed response of the corresponding observation in the predictor data.

Data Types: `single` | `double`

Other Regression Properties

`CategoricalPredictors` — Categorical predictor indices
vector of positive integers

This property is read-only.

Categorical predictor indices, specified as a vector of positive integers. `CategoricalPredictors` contains index values corresponding to the columns of the predictor data that contain categorical predictors. If none of the predictors are categorical, then this property is empty (`[]`).

Data Types: `single` | `double`

`PredictorNames` — Predictor names
cell array of character vectors

This property is read-only.

Predictor names in order of their appearance in the predictor data, specified as a cell array of character vectors. The length of `PredictorNames` is equal to the number of columns used as predictor variables in the training data `X` or `Tbl`.

Data Types: `cell`

`ResponseName` — Response variable name
character vector

This property is read-only.

Response variable name, specified as a character vector.

Data Types: `char`

`ResponseTransform` — Response transformation function
`'none'` | function handle

Response transformation function, specified as `'none'` or a function handle. `ResponseTransform` describes how the software transforms raw response values predicted by the model.

For a MATLAB function, or a function that you define, enter its function handle. For example, you can enter `Mdl.ResponseTransform = @`*`function`*, where *`function`* accepts a numeric vector of the original responses and returns a numeric vector of the same size containing the transformed responses.

Data Types: `char` | `string` | `function_handle`

Object Functions

kfoldLoss	Regression loss for cross-validated kernel regression model
kfoldPredict	Predict responses for observations in cross-validated kernel regression model

Examples

Create Cross-Validated Kernel Regression Model

Simulate sample data.

```
rng(0,'twister'); % For reproducibility
n = 1000;
x = linspace(-10,10,n)';
y = 1 + x*2e-2 + sin(x)./x + 0.2*randn(n,1);
```

Cross-validate a kernel regression model.

```
CVMdl = fitrkernel(x,y,'Kfold',5);
```

`CVMdl` is a `RegressionPartitionedKernel` 5-fold cross-validated model. `CVMdl.Trained` contains a cell vector of five `RegressionKernel` models. Display the trained property.

```
CVMdl.Trained
```

```
ans=5×1 cell array
    {1x1 RegressionKernel}
    {1x1 RegressionKernel}
    {1x1 RegressionKernel}
    {1x1 RegressionKernel}
    {1x1 RegressionKernel}
```

Each cell contains a kernel regression model trained on four folds, then tested on the remaining fold.

Predict responses for observations in the validation folds and estimate the generalization error by passing `CVMdl` to `kfoldPredict` and `kfoldLoss`, respectively.

```
yHat = kfoldPredict(CVMdl);
L = kfoldLoss(CVMdl)
```

```
L = 0.1887
```

`kfoldLoss` computes the average mean squared error for all the folds by default. The estimated mean squared error is 0.1887.

See Also

`RegressionKernel` | `fitrkernel`

Introduced in R2018b

kfoldLoss

Package: `classreg.learning.partition`

Regression loss for cross-validated kernel regression model

Syntax

```
L = kfoldLoss(CVMdl)
L = kfoldLoss(CVMdl,Name,Value)
```

Description

`L = kfoldLoss(CVMdl)` returns the regression loss obtained by the cross-validated kernel regression model `CVMdl`. For every fold, `kfoldLoss` computes the regression loss for observations in the validation fold, using a model trained on observations in the training fold.

`L = kfoldLoss(CVMdl,Name,Value)` returns the mean squared error (MSE) with additional options specified by one or more name-value pair arguments. For example, you can specify the regression-loss function or which folds to use for loss calculation.

Examples

Compute Loss for Cross-Validated Kernel Regression Models

Simulate sample data:

```
rng(0,'twister'); % For reproducibility
n = 1000;
x = linspace(-10,10,n)';
y = 1 + x*2e-2 + sin(x)./x + 0.2*randn(n,1);
```

Cross-validate a kernel regression model.

```
CVMdl = fitrkernel(x,y,'Kfold',5);
```

`fitrkernel` implements 5-fold cross-validation. `CVMdl` is a `RegressionPartitionedKernel` model. It contains the property `Trained`, which is a 5-by-1 cell array holding 5 `RegressionKernel` models that the software trained using the training set.

Compute the epsilon-insensitive loss for each fold for observations that `fitrkernel` did not use in training the folds.

```
L = kfoldLoss(CVMdl,'LossFun','epsiloninsensitive','Mode','individual')
```

```
L = 5×1

    0.2812
    0.3223
    0.3073
    0.3117
```

```
    0.2576
```

Input Arguments

`CVMdl` — Cross-validated kernel regression model

`RegressionPartitionedKernel` model object

Cross-validated kernel regression model, specified as a `RegressionPartitionedKernel` model object. You can create a `RegressionPartitionedKernel` model using `fitrkernel` and specifying any of the cross-validation name-value pair arguments, for example, `CrossVal`.

Name-Value Pair Arguments

Specify optional comma-separated pairs of `Name,Value` arguments. `Name` is the argument name and `Value` is the corresponding value. `Name` must appear inside quotes. You can specify several name and value pair arguments in any order as `Name1,Value1,...,NameN,ValueN`.

Example: `'LossFun','epsiloninsensitive','Mode','individual'` specifies `kfoldLoss` to return the epsilon-insensitive loss for each fold.

`Folds` — Fold indices to use for response prediction

`1:CVMdl.KFold` (default) | numeric vector of positive integers

Fold indices to use for response prediction, specified as the comma-separated pair consisting of `'Folds'` and a numeric vector of positive integers. The elements of `Folds` must range from `1` through `CVMdl.KFold`.

Example: `'Folds',[1 4 10]`

Data Types: `single` | `double`

`LossFun` — Loss function

`'mse'` (default) | `'epsiloninsensitive'` | function handle

Loss function, specified as the comma-separated pair consisting of `'LossFun'` and a built-in, loss-function name or function handle.

- The following table lists the available loss functions. Specify one using its corresponding character vector or string scalar. Also, in the table, $f(x) = x\beta + b$.
  - β is a vector of p coefficients.
  - x is an observation from p predictor variables.
  - b is the scalar bias.

Value	Description
`'epsiloninsensitive'`	Epsilon-insensitive loss: $\ell[y, f(x)] = \max[0, \lvert y - f(x)\rvert - \varepsilon]$
`'mse'`	MSE: $\ell[y, f(x)] = [y - f(x)]^2$

  `'epsiloninsensitive'` is appropriate for SVM learners only.
- Specify your own function using function handle notation.

Assume that `n` is the number of observations in `X`. Your function must have this signature

```
lossvalue = lossfun(Y,Yhat,W)
```

where:

- The output argument `lossvalue` is a scalar.
- You specify the function name (*`lossfun`*).
- `Y` is an `n`-dimensional vector of observed responses. `kfoldLoss` passes the input argument `Y` in for `Y`.
- `Yhat` is an `n`-dimensional vector of predicted responses, which is similar to the output of `predict`.
- `W` is an `n`-by-1 numeric vector of observation weights.

Data Types: `char` | `string` | `function_handle`

`Mode` — Loss aggregation level
`'average'` (default) | `'individual'`

Loss aggregation level, specified as the comma-separated pair consisting of `'Mode'` and `'average'` or `'individual'`.

Value	Description
`'average'`	Returns losses averaged over all folds
`'individual'`	Returns losses for each fold

Example: `'Mode','individual'`

Output Arguments

`L` — Cross-validated regression losses
numeric scalar | numeric vector

Cross-validated regression losses, returned as a numeric scalar or vector. The interpretation of `L` depends on `LossFun`.

- If `Mode` is `'average'`, then `L` is a scalar.
- Otherwise, `L` is a *k*-by-1 vector, where *k* is the number of folds. `L(`*`j`*`)` is the average regression loss over fold *`j`*.

To estimate `L`, `kfoldLoss` uses the data that created `CVMdl`.

See Also
`RegressionKernel` | `RegressionPartitionedKernel` | `fitrkernel` | `kfoldPredict`

Introduced in R2018b

kfoldPredict

Package: `classreg.learning.partition`

Predict responses for observations in cross-validated kernel regression model

Syntax

```
YHat = kfoldPredict(CVMdl)
```

Description

`YHat = kfoldPredict(CVMdl)` returns cross-validated predicted responses by the cross-validated kernel regression model `CVMdl`. That is, for every fold, `kfoldPredict` predicts the responses for observations that it holds out in validation fold while it trains using all other observations in the training fold.

Examples

Predict Responses from Cross-Validated Kernel Regression Models

Simulate sample data:

```
rng(0,'twister'); % For reproducibility
n = 1000;
x = linspace(-10,10,n)';
y = 1 + x*2e-2 + sin(x)./x + 0.2*randn(n,1);
```

Cross-validate a kernel regression model.

```
CVMdl = fitrkernel(x,y,'CrossVal','on');
```

By default, `fitrkernel` implements 10-fold cross-validation. `CVMdl` is a `RegressionPartitionedKernel` model. It contains the property `Trained`, which is a 10-by-1 cell array holding 10 `RegressionKernel` models that the software trained using the training set.

Predict responses for observations that `fitrkernel` did not use in training the folds.

```
yHat = kfoldPredict(CVMdl);
```

`yHat` is a numeric vector. Display the first five predicted responses.

```
yHat(1:5)
```

```
ans = 5×1

    1.0769
    1.0744
    1.0758
    1.0781
    1.0795
```

Input Arguments

`CVMdl` — Cross-validated kernel regression model

`RegressionPartitionedKernel` model object

Cross-validated kernel regression model, specified as a `RegressionPartitionedKernel` model object. You can create a `RegressionPartitionedKernel` model using `fitrkernel` and specifying any of the one of the cross-validation name-value pair arguments, for example, `CrossVal`.

To obtain estimates, `kfoldPredict` applies the same data used to cross-validate the kernel regression model (see `X` input argument on `fitrkernel` page).

Output Arguments

`YHat` — Cross-validated predicted responses

numeric vector

Cross-validated predicted responses, returned as an *n*-by-1 numeric array, where *n* is the number of observations in the predictor data used to create `CVMdl` (see `X` input argument on `fitrkernel` page).

See Also

`RegressionKernel` | `RegressionPartitionedKernel` | `fitrkernel` | `kfoldLoss`

Introduced in R2018b

A

Sample Data Sets

Sample Data Sets

Statistics and Machine Learning Toolbox software includes the sample data sets in the following table.

To load a data set into the MATLAB workspace, type:

```
load filename
```

where *filename* is one of the files listed in the table.

Data sets contain individual data variables, description variables with references, and dataset arrays encapsulating the data set and its description, as appropriate.

File	Description of Data Set
acetylene.mat	Chemical reaction data with correlated predictors
arrhythmia.mat	Cardiac arrhythmia data from the UCI machine learning repository
carbig.mat	Measurements of cars, 1970–1982
carsmall.mat	Subset of carbig.mat. Measurements of cars, 1970, 1976, 1982
census1994.mat	Adult data from the UCI machine learning repository
cereal.mat	Breakfast cereal ingredients
cities.mat	Quality of life ratings for U.S. metropolitan areas
discrim.mat	A version of cities.mat used for discriminant analysis
examgrades.mat	Exam grades on a scale of 0–100
fisheriris.mat	Fisher's 1936 iris data
flu.mat	Google Flu Trends estimated ILI (influenza-like illness) percentage for various regions of the US, and CDC weighted ILI percentage based on sentinel provider reports
gas.mat	Gasoline prices around the state of Massachusetts in 1993
hald.mat	Heat of cement vs. mix of ingredients
hogg.mat	Bacteria counts in different shipments of milk
hospital.mat	Simulated hospital data
humanactivity.mat	Human activity recognition data of five activities: sitting, standing, walking, running, and dancing
imports-85.mat	1985 Auto Imports Database from the UCI repository
ionosphere.mat	Ionosphere dataset from the UCI machine learning repository
kmeansdata.mat	Four-dimensional clustered data
lawdata.mat	Grade point average and LSAT scores from 15 law schools
mileage.mat	Mileage data for three car models from two factories
moore.mat	Biochemical oxygen demand on five predictors
morse.mat	Recognition of Morse code distinctions by non-coders
nlpdata.mat	Natural language processing data extracted from the MathWorks® documentation

File	Description of Data Set
`ovariancancer.mat`	Grouped observations on 4000 predictors [1][2]
`parts.mat`	Dimensional run-out on 36 circular parts
`polydata.mat`	Sample data for polynomial fitting
`popcorn.mat`	Popcorn yield by popper type and brand
`reaction.mat`	Reaction kinetics for Hougen-Watson model
`spectra.mat`	NIR spectra and octane numbers of 60 gasoline samples
`stockreturns.mat`	Simulated stock returns

References

[1] Conrads, Thomas P., Vincent A. Fusaro, Sally Ross, Don Johann, Vinodh Rajapakse, Ben A. Hitt, Seth M. Steinberg, et al. "High-Resolution Serum Proteomic Features for Ovarian Cancer Detection." *Endocrine-Related Cancer* 11 (2004): 163–78.

[2] Petricoin, Emanuel F., Ali M. Ardekani, Ben A. Hitt, Peter J. Levine, Vincent A. Fusaro, Seth M. Steinberg, Gordon B. Mills, et al. “Use of Proteomic Patterns in Serum to Identify Ovarian Cancer.” *The Lancet* 359, no. 9306 (February 2002): 572–77.

B

Probability Distributions

Bernoulli Distribution

In this section...

Overview

The Bernoulli distribution is a discrete probability distribution with only two possible values for the random variable. Each instance of an event with a Bernoulli distribution is called a Bernoulli trial.

Parameters

The Bernoulli distribution uses the following parameter.

Parameter	Description	Support
p	Probability of success	$0 \le p \le 1$

Probability Density Function

The probability density function (pdf) of the Bernoulli distribution is

$$f(x\left|p\right.) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \end{cases}.$$

For discrete distributions, the pdf is also known as the probability mass function (pmf).

For an example, see “Compute Bernoulli Distribution pdf” on page B-3.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the Bernoulli distribution is

$$F(x\left|p\right.) = \begin{cases} 1-p, & x = 0 \\ 1, & x = 1 \end{cases}.$$

For an example, see “Compute Bernoulli Distribution cdf” on page B-3.

Descriptive Statistics

The mean of the Bernoulli distribution is p.

The variance of the Bernoulli distribution is $p(1 - p)$.

Examples

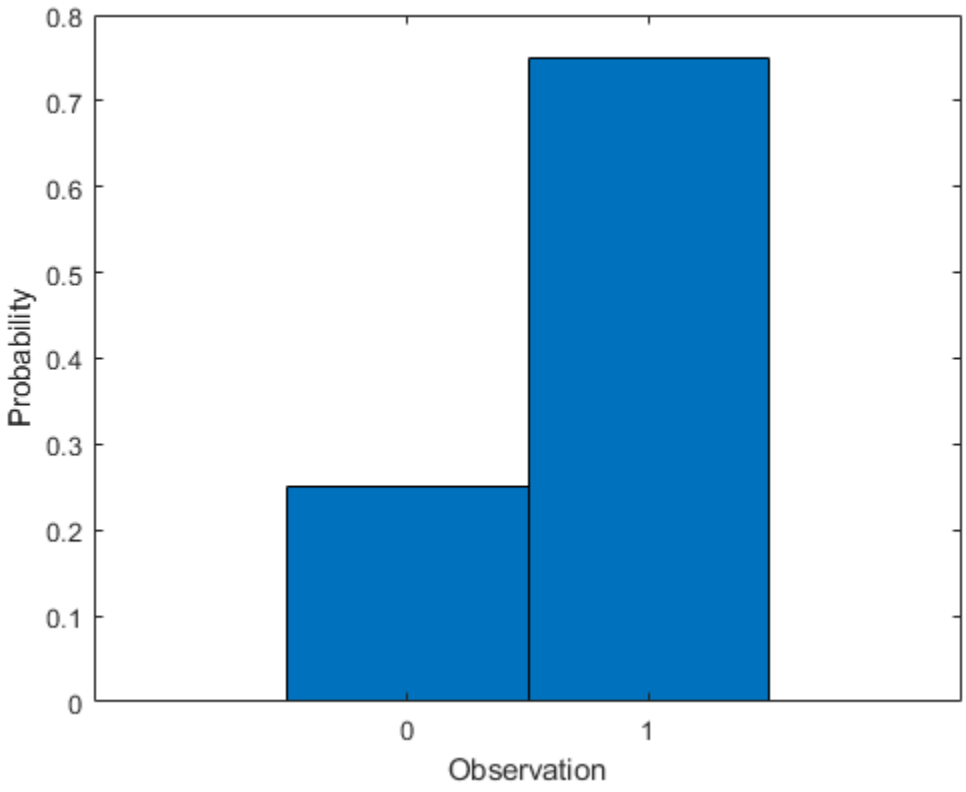
Compute Bernoulli Distribution pdf

The Bernoulli distribution is a special case of the binomial distribution, where `N = 1`. Use `binopdf` to compute the pdf of the Bernoulli distribution with the probability of success `0.75`.

```
p = 0.75;
x = 0:1;
y = binopdf(0:1,1,p);
```

Plot the pdf with bars of width `1`.

```
figure
bar(x,y,1)
xlabel('Observation')
ylabel('Probability')
```



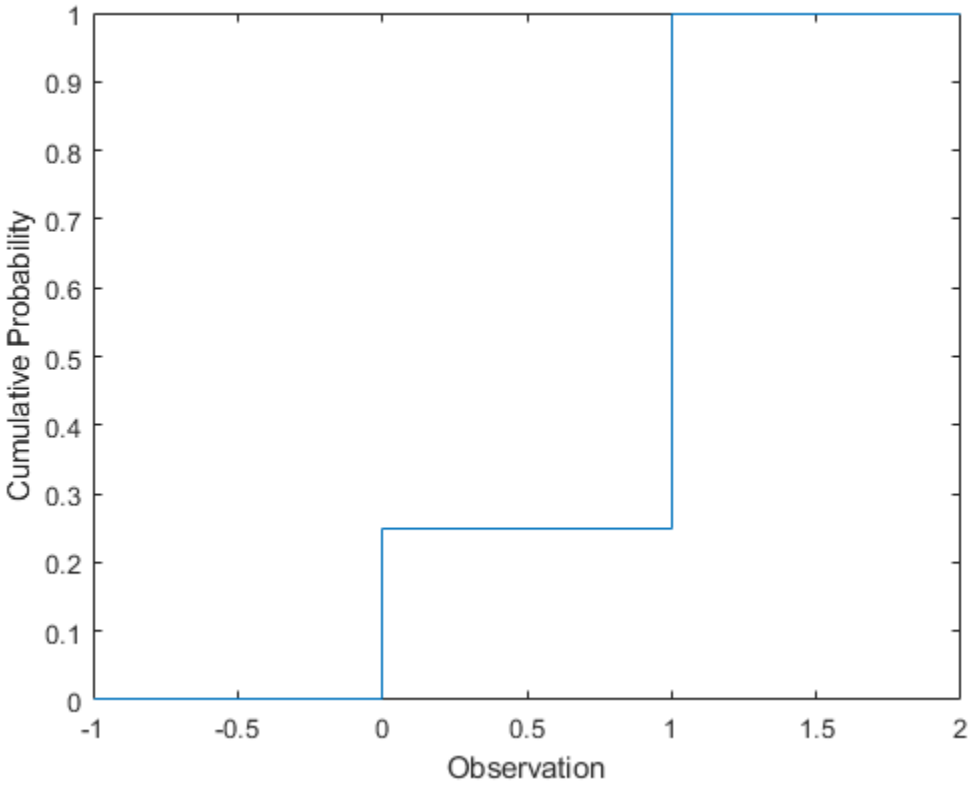
Compute Bernoulli Distribution cdf

The Bernoulli distribution is a special case of the binomial distribution, where `N = 1`. Use `binocdf` to compute the cdf of the Bernoulli distribution with the probability of success `0.75`.

```
p = 0.75;
y = binocdf(-1:2,1,p);
```

Plot the cdf.

```
figure
stairs(-1:2,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Related Distributions

- Binomial Distribution on page B-10 — The binomial distribution is a two-parameter discrete distribution that models the total number of successes in repeated Bernoulli trials. The Bernoulli distribution occurs as a binomial distribution with $N = 1$.
- Geometric Distribution on page B-63 — The geometric distribution is a one-parameter discrete distribution that models the total number of failures before the first success in repeated Bernoulli trials.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

See Also

`BinomialDistribution`

More About

- “Binomial Distribution” on page B-10
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Beta Distribution

In this section...

Overview

The beta distribution describes a family of curves that are unique in that they are nonzero only on the interval (0 1). A more general version of the function assigns parameters to the endpoints of the interval.

Statistics and Machine Learning Toolbox provides several ways to work with the beta distribution. You can use the following approaches to estimate parameters from sample data, compute the pdf, cdf, and icdf, generate random numbers, and more.

- Fit a probability distribution object to sample data, or create a probability distribution object with specified parameter values. See `Using BetaDistribution Objects` for more information.
- Work with data input from matrices, tables, and dataset arrays using probability distribution functions. See “Supported Distributions” on page 5-14 for a list of beta distribution functions.
- Interactively fit, explore, and generate random numbers from the distribution using an app or user interface.

For more information on each of these options, see “Working with Probability Distributions” on page 5-3.

Parameters

The beta distribution uses the following parameters.

Parameter	Description	Support
a	First shape parameter	$a > 0$
b	Second shape parameter	$b > 0$

Probability Density Function

Definition

The probability density function (pdf) of the beta distribution is

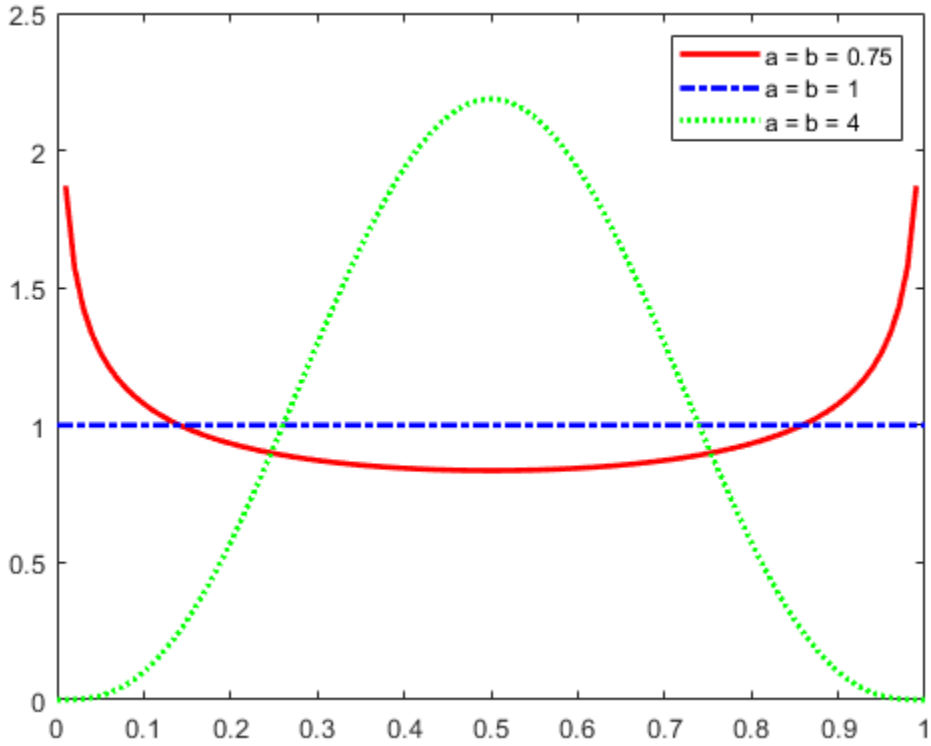
$$y = f(x \mid a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{[0,1]}(x)$$

where $B(\cdot)$ is the Beta function. The indicator function $I_{(0,1)}(x)$ ensures that only values of x in the range (0,1) have nonzero probability.

Plot

This plot shows how changing the value of the parameters alters the shape of the pdf. The constant pdf (the flat line) shows that the standard uniform distribution is a special case of the beta distribution, which occurs when `a = b = 1`.

```
X = 0:.01:1;
y1 = betapdf(X,0.75,0.75);
y2 = betapdf(X,1,1);
y3 = betapdf(X,4,4);

figure
plot(X,y1,'Color','r','LineWidth',2)
hold on
plot(X,y2,'LineStyle','-.','Color','b','LineWidth',2)
plot(X,y3,'LineStyle',':','Color','g','LineWidth',2)
legend({'a = b = 0.75','a = b = 1','a = b = 4'},'Location','NorthEast');
hold off
```



Relationship to Other Distributions

The beta distribution has a functional relationship with the t distribution. If Y is an observation from Student's t distribution with ν degrees of freedom, then the following transformation generates X, which is beta distributed.

$$X = \frac{1}{2} + \frac{1}{2}\frac{Y}{\sqrt{\nu + Y^2}}$$

If $Y \sim t(\nu)$, then $X \sim \beta\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$

This relationship is used to compute values of the t cdf and inverse function as well as generating t distributed random numbers.

Cumulative Distribution Function

The beta cdf is the same as the incomplete beta function.

Example

Suppose you are collecting data that has hard lower and upper bounds of zero and one respectively. Parameter estimation is the process of determining the parameters of the beta distribution that fit this data best in some sense.

One popular criterion of goodness is to maximize the likelihood function. The likelihood has the same form as the beta pdf. But for the pdf, the parameters are known constants and the variable is x. The likelihood function reverses the roles of the variables. Here, the sample values (the x's) are already observed. So they are the fixed constants. The variables are the unknown parameters. Maximum likelihood estimation (MLE) involves calculating the values of the parameters that give the highest likelihood given the particular set of data.

The function `betafit` returns the MLEs and confidence intervals for the parameters of the beta distribution. Here is an example using random numbers from the beta distribution with `a = 5` and `b = 0.2`.

```
rng default  % For reproducibility
r = betarnd(5,0.2,100,1);
[phat, pci] = betafit(r)
```

```
phat = 1×2

    7.4911    0.2135


pci = 2×2

    5.0861    0.1744
   11.0334    0.2614
```

The MLE for parameter `a` is 7.4911, compared to the true value of 5. The 95% confidence interval for `a` goes from 5.0861 to 11.0334, which does not include the true value. While this is an unlikely result, it does sometimes happen when estimating distribution parameters.

Similarly the MLE for parameter `b` is 0.2135, compared to the true value of 0.2. The 95% confidence interval for `b` goes from 0.1744 to 0.2614, which does include the true value. In this made-up example you know the “true value.” In experimentation you do not.

See Also

`BetaDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Binomial Distribution

In this section...

Overview

The binomial distribution is a two-parameter family of curves. The binomial distribution is used to model the total number of successes in a fixed number of independent trials that have the same probability of success, such as modeling the probability of a given number of heads in ten flips of a fair coin.

Statistics and Machine Learning Toolbox offers several ways to work with the binomial distribution.

- Create a probability distribution object `BinomialDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the binomial distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`binocdf`, `binopdf`, `binoinv`, `binostat`, `binofit`, `binornd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple binomial distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Binomial'`) and parameters.

Parameters

The binomial distribution uses the following parameters.

Parameter	Description	Support
N	Number of trials	Positive integer
p	Probability of success in a single trial	$0 \le p \le 1$

The sum of two binomial random variables that both have the same parameter p is also a binomial random variable with N equal to the sum of the number of trials.

Probability Density Function

The probability density function (pdf) of the binomial distribution is

$$f(x|N,p) = \binom{N}{x} p^x (1-p)^{N-x} \quad ; \quad x = 0, 1, 2, \ldots, N,$$

where x is the number of successes in N trials of a Bernoulli process with the probability of success p. The result is the probability of exactly x successes in N trials. For discrete distributions, the pdf is also known as the probability mass function (pmf).

For an example, see “Compute Binomial Distribution pdf” on page B-12.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the binomial distribution is

$$F(x|N,p) = \sum_{i=0}^{x} \binom{N}{i} p^i (1-p)^{N-i} \quad ; \quad x = 0, 1, 2, \ldots, N,$$

where x is the number of successes in N trials of a Bernoulli process with the probability of success p. The result is the probability of at most x successes in N trials.

For an example, see “Compute Binomial Distribution cdf” on page B-13.

Descriptive Statistics

The mean of the binomial distribution is Np.

The variance of the binomial distribution is $Np(1 – p)$.

Example

Fit Binomial Distribution to Data

Generate a binomial random number that counts the number of successes in `100` trials with the probability of success `0.9` in each trial.

```
x = binornd(100,0.9)
```

```
x = 85
```

Fit a binomial distribution to data using `fitdist`.

```
pd = fitdist(x,'Binomial','NTrials',100)
```

```
pd =
  BinomialDistribution

  Binomial distribution
    N =  100
    p = 0.85   [0.764692, 0.913546]
```

`fitdist` returns a `BinomialDistribution` object. The interval next to `p` is the 95% confidence interval estimating `p`.

Estimate the parameter `p` using the distribution functions.

```
[phat,pci] = binofit(x,100) % Distribution-specific function

phat = 0.8500

pci = 1×2

    0.7647    0.9135

[phat2,pci2] = mle(x,'distribution','Binomial',"NTrials",100) % Generic distribution function

phat2 = 0.8500

pci2 = 2×1

    0.7647
    0.9135
```

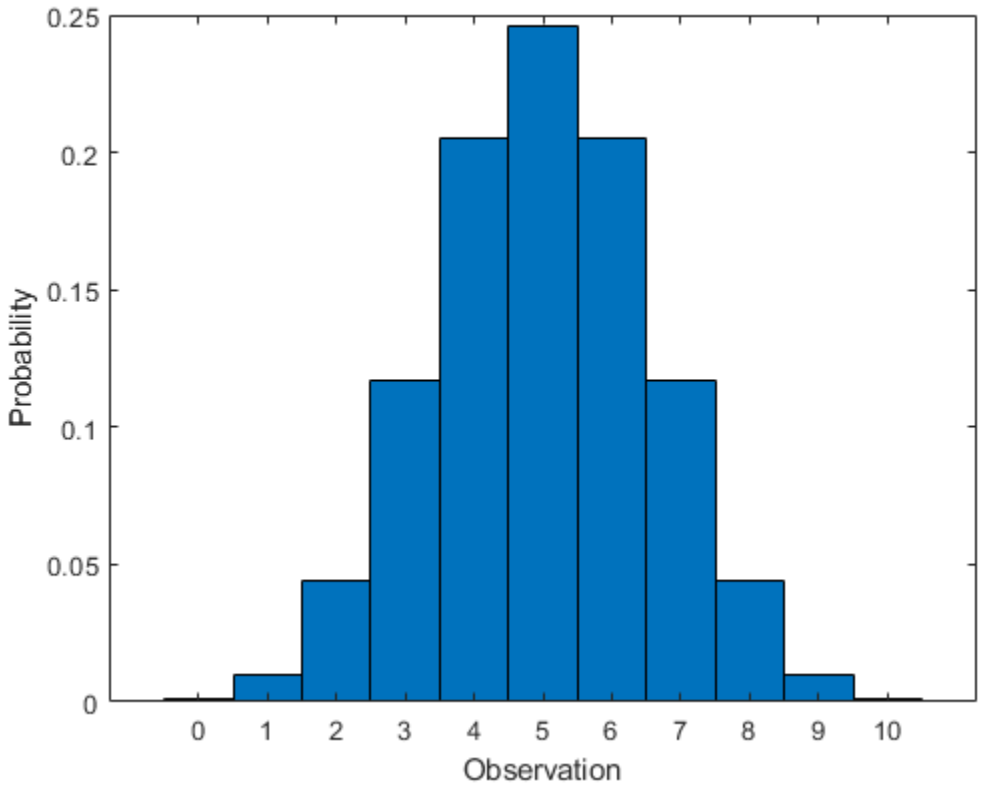
Compute Binomial Distribution pdf

Compute the pdf of the binomial distribution with `10` trials and the probability of success `0.5`.

```
x = 0:10;
y = binopdf(x,10,0.5);
```

Plot the pdf with bars of width `1`.

```
figure
bar(x,y,1)
xlabel('Observation')
ylabel('Probability')
```



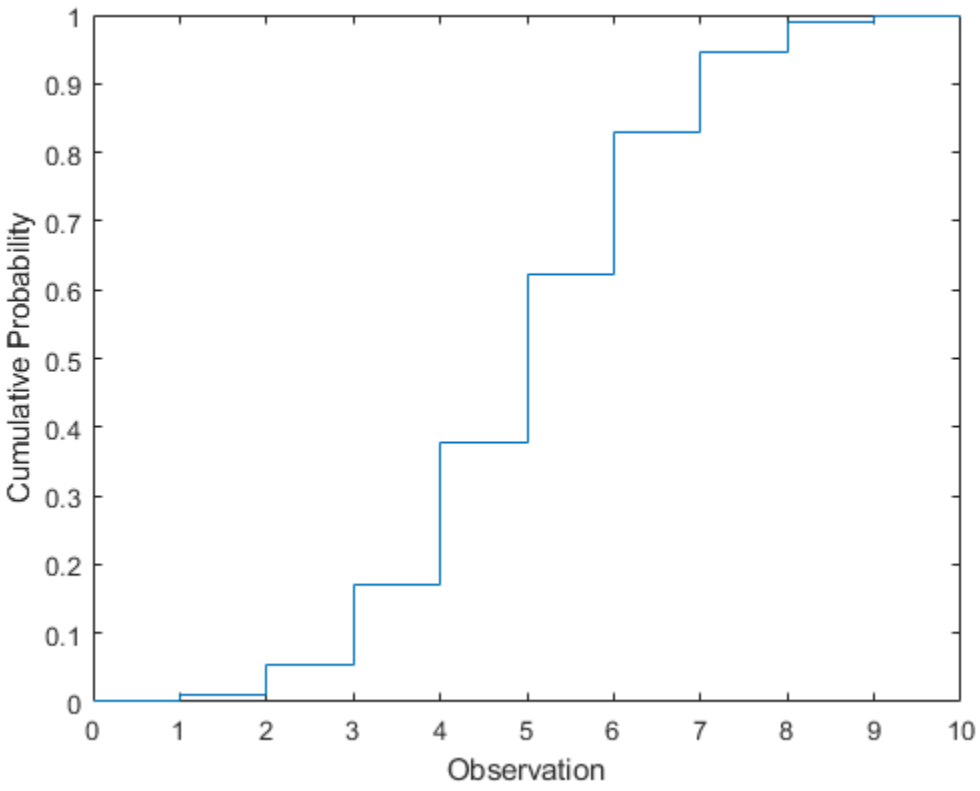
Compute Binomial Distribution cdf

Compute the cdf of the binomial distribution with `10` trials and the probability of success `0.5`.

```
x = 0:10;
y = binocdf(x,10,0.5);
```

Plot the cdf.

```
figure
stairs(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Compare Binomial and Normal Distribution pdfs

When `N` is large, the binomial distribution with parameters `N` and `p` can be approximated by the normal distribution with mean `N*p` and variance `N*p*(1–p)` provided that `p` is not too large or too small.

Compute the pdf of the binomial distribution counting the number of successes in `50` trials with the probability `0.6` in a single trial .

```
N = 50;
p = 0.6;
x1 = 0:N;
y1 = binopdf(x1,N,p);
```

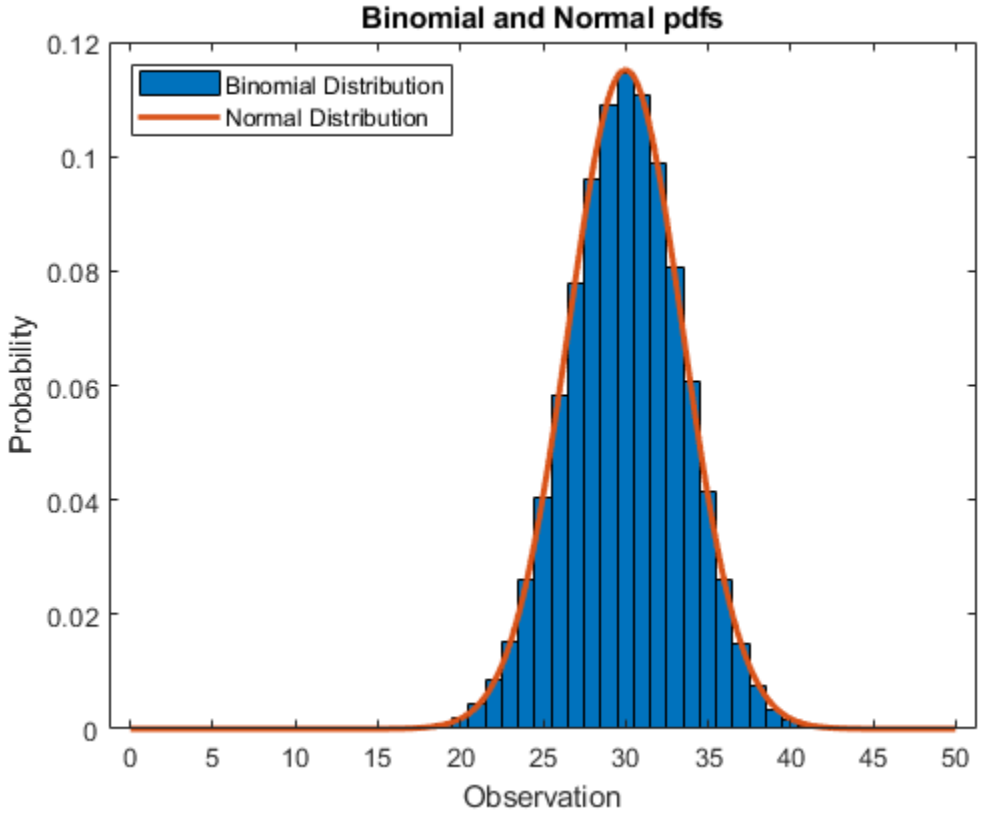
Compute the pdf of the corresponding normal distribution.

```
mu = N*p;
sigma = sqrt(N*p*(1-p));
x2 = 0:0.1:N;
y2 = normpdf(x2,mu,sigma);
```

Plot the pdfs on the same axis.

```
figure
bar(x1,y1,1)
hold on
plot(x2,y2,'LineWidth',2)
```

```
xlabel('Observation')
ylabel('Probability')
title('Binomial and Normal pdfs')
legend('Binomial Distribution','Normal Distribution','location','northwest')
hold off
```



The pdf of the normal distribution closely approximates the pdf of the binomial distribution.

Compare Binomial and Poisson Distribution pdfs

When `p` is small, the binomial distribution with parameters `N` and `p` can be approximated by the Poisson distribution with mean `N*p`, provided that `N*p` is also small.

Compute the pdf of the binomial distribution counting the number of successes in `20` trials with the probability of success `0.05` in a single trial.

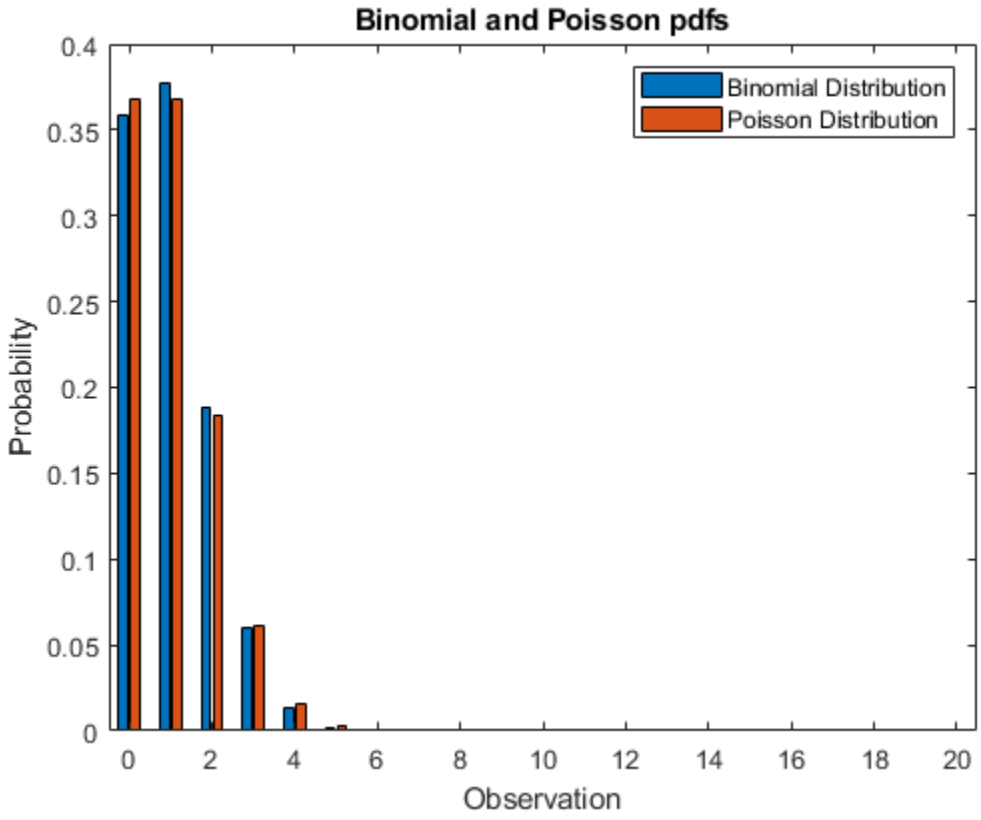
```
N = 20;
p = 0.05;
x = 0:N;
y1 = binopdf(x,N,p);
```

Compute the pdf of the corresponding Poisson distribution.

```
mu = N*p;
y2 = poisspdf(x,mu);
```

Plot the pdfs on the same axis.

```
figure
bar(x,[y1; y2])
xlabel('Observation')
ylabel('Probability')
title('Binomial and Poisson pdfs')
legend('Binomial Distribution','Poisson Distribution','location','northeast')
```



The pdf of the Poisson distribution closely approximates the pdf of the binomial distribution.

Related Distributions

- Bernoulli Distribution on page B-2 — The Bernoulli distribution is a one-parameter discrete distribution that models the success of a single trial, and occurs as a binomial distribution with $N = 1$.
- Multinomial Distribution on page B-96 — The multinomial distribution is a discrete distribution that generalizes the binomial distribution when each trial has more than two possible outcomes.
- “Normal Distribution” on page B-119 — The normal distribution is a two-parameter continuous distribution that has parameters μ (mean) and σ (standard deviation). As N increases, the binomial distribution can be approximated by a normal distribution with $\mu = Np$ and $\sigma^2 = Np(1 - p)$. See “Compare Binomial and Normal Distribution pdfs” on page B-14.
- “Poisson Distribution” on page B-131 — The Poisson distribution is a one-parameter discrete distribution that takes nonnegative integer values. The parameter λ is both the mean and the variance of the distribution. The Poisson distribution is the limiting case of a binomial distribution

where N approaches infinity and p goes to zero while $Np = \lambda$. See “Compare Binomial and Poisson Distribution pdfs” on page B-15.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

[3] Loader, Catherine. *Fast and Accurate Computation of Binomial Probabilities*. July 9, 2000.

See Also

`BinomialDistribution` | `binocdf` | `binofit` | `binoinv` | `binopdf` | `binornd` | `binostat` | `fitdist` | `makedist`

More About

- “Bernoulli Distribution” on page B-2
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Birnbaum-Saunders Distribution

In this section...

Definition

The Birnbaum-Saunders distribution has the density function

$$\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\left(\sqrt{x/\beta}-\sqrt{\beta/x}\right)^2}{2\gamma^2}\right\}\left(\frac{\left(\sqrt{x/\beta}+\sqrt{\beta/x}\right)}{2\gamma x}\right)$$

with scale parameter $\beta > 0$ and shape parameter $\gamma > 0$, for $x > 0$.

If x has a Birnbaum-Saunders distribution with parameters β and γ, then

$$\frac{\left(\sqrt{x/\beta}-\sqrt{\beta/x}\right)}{\gamma}$$

has a standard normal distribution.

Background

The Birnbaum-Saunders distribution was originally proposed as a lifetime model for materials subject to cyclic patterns of stress and strain, where the ultimate failure of the material comes from the growth of a prominent flaw. In materials science, Miner's Rule suggests that the damage occurring after n cycles, at a stress level with an expected lifetime of N cycles, is proportional to n / N. Whenever Miner's Rule applies, the Birnbaum-Saunders model is a reasonable choice for a lifetime distribution model.

Parameters

To estimate distribution parameters, us `mle` or the Distribution Fitter app.

See Also

`BirnbaumSaundersDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Burr Type XII Distribution

In this section...

Definition

The Burr type XII distribution is a three-parameter family of distributions on the positive real line. The cumulative distribution function (cdf) of the Burr distribution is

$$F(x\left|\alpha,c,k\right.)=1-\frac{1}{\left(1+\left(\frac{x}{\alpha}\right)^{c}\right)^{k}},\quad x>0,\ \alpha>0,\ c>0,\ k>0,$$

where c and k are the shape parameters and α is the scale parameter. The probability density function (pdf) is

$$f(x\left|\alpha,c,k\right.)=\frac{\frac{kc}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}}{\left(1+\left(\frac{x}{\alpha}\right)^{c}\right)^{k+1}},\quad x>0,\ \alpha>0,\ c>0,\ k>0\ .$$

The density of the Burr type XII distribution is L-shaped if $c \le 1$ and unimodal, otherwise.

Background

Burr distribution was first discussed by Burr (1942) as a two-parameter family. An additional scale parameter was introduced by Tadikamalla (1980). It is a flexible distribution family that can express a wide range of distribution shapes. The Burr distribution includes, overlaps, or has as a limiting case, many commonly used distributions such as gamma, lognormal, loglogistic, bell-shaped, and J-shaped beta distributions (but not U-shaped). Some compound distributions also correspond to the Burr distribution. For example, compounding a Weibull distribution with a gamma distribution for its scale parameter results in a Burr distribution. Similarly, compounding an exponential distribution with a gamma distribution for its rate parameter, 1/μ, also yields a Burr distribution. The Burr distribution also has two asymptotic limiting cases: Weibull and Pareto Type I.

The Burr distribution can fit a wide range of empirical data. Different values of its parameters cover a broad set of skewness and kurtosis. Hence, it is used in various fields such as finance, hydrology, and reliability to model a variety of data types. Examples of data modeled by the Burr distribution are household income, crop prices, insurance risk, travel time, flood levels, and failure data.

The survival and hazard functions of Burr type XII distribution are, respectively,

$$S(x|\alpha,c,k) = \frac{1}{\left[1+\left(\frac{x}{\alpha}\right)^c\right]^k}$$

and

$$h(x|\alpha,c,k) = \frac{\frac{kc}{\alpha}\left(\frac{x}{\alpha}\right)^{c-1}}{1+\left(\frac{x}{\alpha}\right)^c}.$$

If $c > 1$, the hazard function h(x) is non-monotonic with a mode at $x = \alpha(c - 1)^{1/c}$.

Parameters

The three-parameter Burr distribution is defined by its scale parameter α and shape parameters c and k. You can estimate the parameters using `mle` or `fitdist`. Both functions support censored data for Burr distribution.

Generate sample data from a Burr distribution with scale parameter 0.5 and shape parameters 2 and 5.

```
rng('default')
R = random('burr',0.5,2,5,1000,1);
```

Estimate the parameters and the confidence intervals.

```
[phat,pci] = mle(R,'distribution','burr')
```

```
phat =

    0.4154    2.1217    4.0550


pci =

    0.2985    1.9560    2.4079
    0.5782    2.3014    6.8288
```

The default 95% confidence intervals for the parameters include the true parameter values.

The three-parameter Burr distribution converges asymptotically to one of the two limiting forms as its parameters diverge:

- If $k \to 0$, $c \to \infty$, $ck = \lambda$, then the Burr distribution reduces to a two-parameter Pareto distribution with the cdf

  $$F_P = 1 - \left(\frac{x}{\alpha}\right)^{-\lambda},\ x \geq \alpha.$$

- If $k \to \infty$, $\alpha \to \infty$, $\alpha/k^{1/c} = \theta$, then the Burr distribution reduces to a two-parameter Weibull distribution with the cdf

  $$F_W(x|c,\theta) = 1 - \exp\left[-\left(\frac{x}{\theta}\right)^c\right].$$

If `mle` or `fitdist` detects such divergence, it returns an error message, but informs you of the limiting distribution and corresponding parameter estimates for that distribution.

Fit a Burr Distribution and Draw the cdf

This example shows how to fit a Burr distribution to data, draw the cdf, and construct a histogram with a Burr distribution fit.

1. Load the sample data.

```
load arrhythmia
```

The fifth column in X contains a measurement obtained from electrocardiograms, called QRS duration.

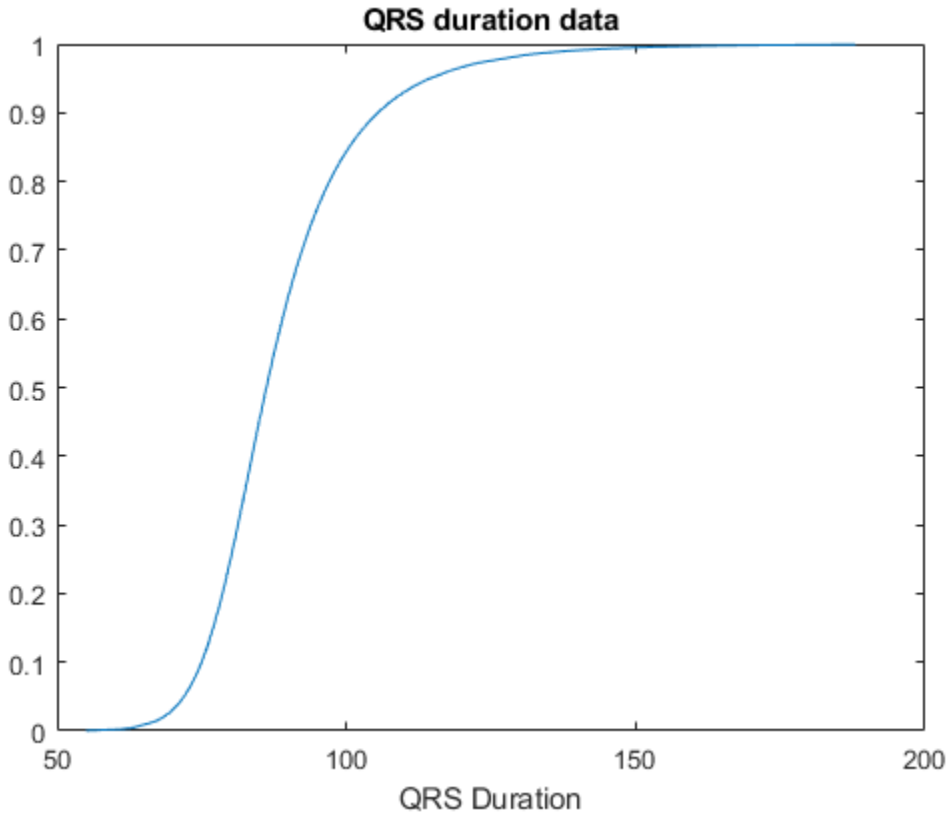
2. Fit a Burr distribution to the QRS duration data, and get the parameter estimates.

```
PD = fitdist(X(:,5),'burr');
```

PD has the maximum likelihood estimates of the Burr distribution parameters in the property `Param`. The estimates are $\alpha = 80.4515$, $c = 18.9251$, $k = 0.4492$.

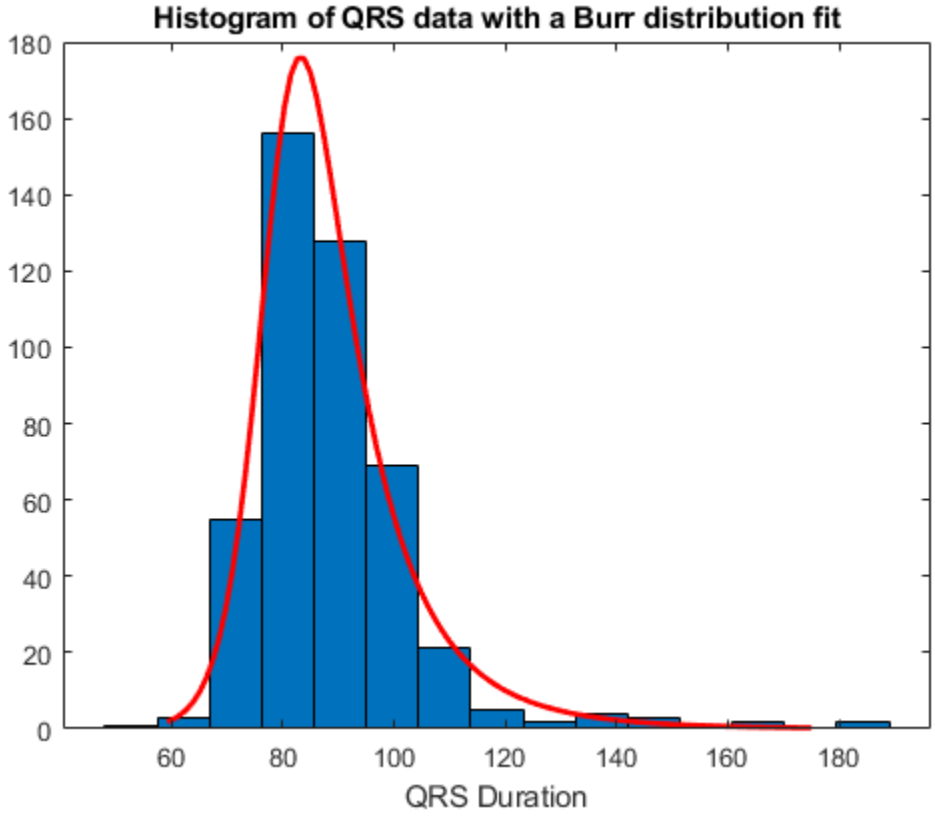
3. Plot the cdf of the QRS duration data.

```
QRScdf=cdf('burr',sortrows(X(:,5)),80.4515,18.9251,0.4492);
plot(sortrows(X(:,5)),QRScdf)
title('QRS duration data')
xlabel('QRS Duration')
```



4. Draw the histogram of QRS duration data with 15 bins and the pdf of the Burr distribution fit.

```
histfit(X(:,5),15,'burr')
title('Histogram of QRS data with a Burr distribution fit')
xlabel('QRS Duration')
```



Compare Lognormal and Burr Distribution pdfs

Compare the lognormal pdf to the Burr pdf using income data generated from a lognormal distribution.

Generate the income data.

```
rng('default') % For reproducibility
y = random('Lognormal',log(25000),0.65,[500,1]);
```

Fit a Burr distribution.

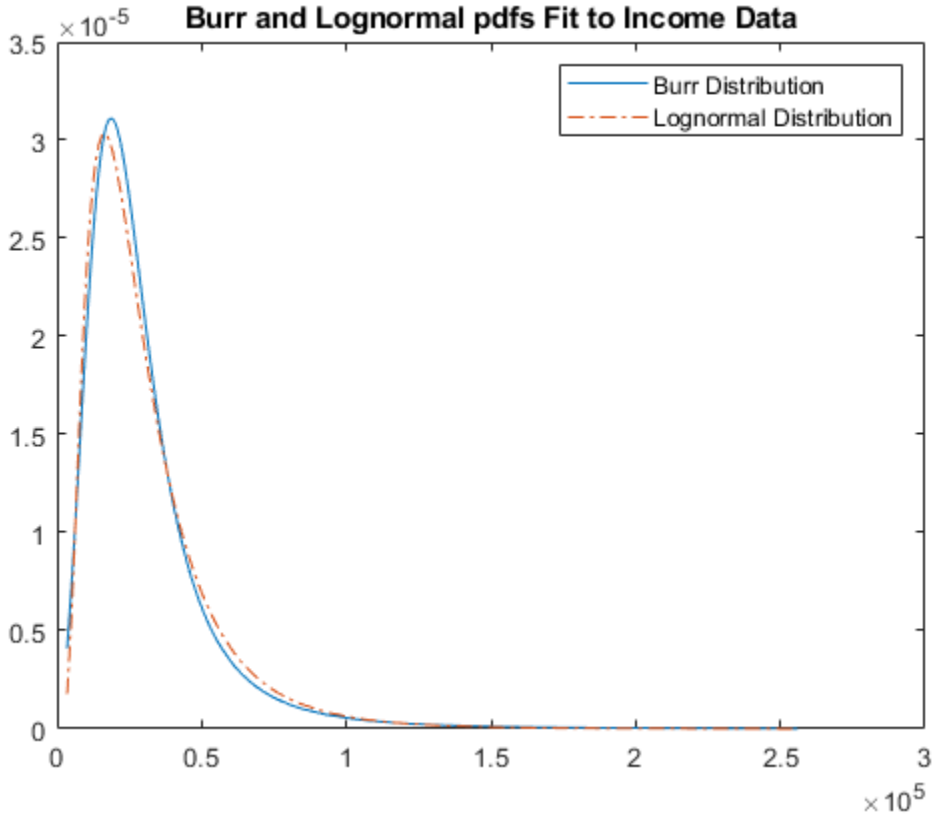
```
pd = fitdist(y,'burr')
```

```
pd =
  BurrDistribution

  Burr distribution
    alpha = 26007.2   [21165.5, 31956.4]
        c = 2.63743   [2.3053, 3.0174]
```

```
k = 1.09658   [0.775479, 1.55064]
```

Plot both the Burr and lognormal pdfs of income data on the same figure.

```
p_burr = pdf(pd,sortrows(y));
p_lognormal = pdf('Lognormal',sortrows(y),log(25000),0.65);
plot(sortrows(y),p_burr,'-',sortrows(y),p_lognormal,'-.')
title('Burr and Lognormal pdfs Fit to Income Data')
legend('Burr Distribution','Lognormal Distribution')
```

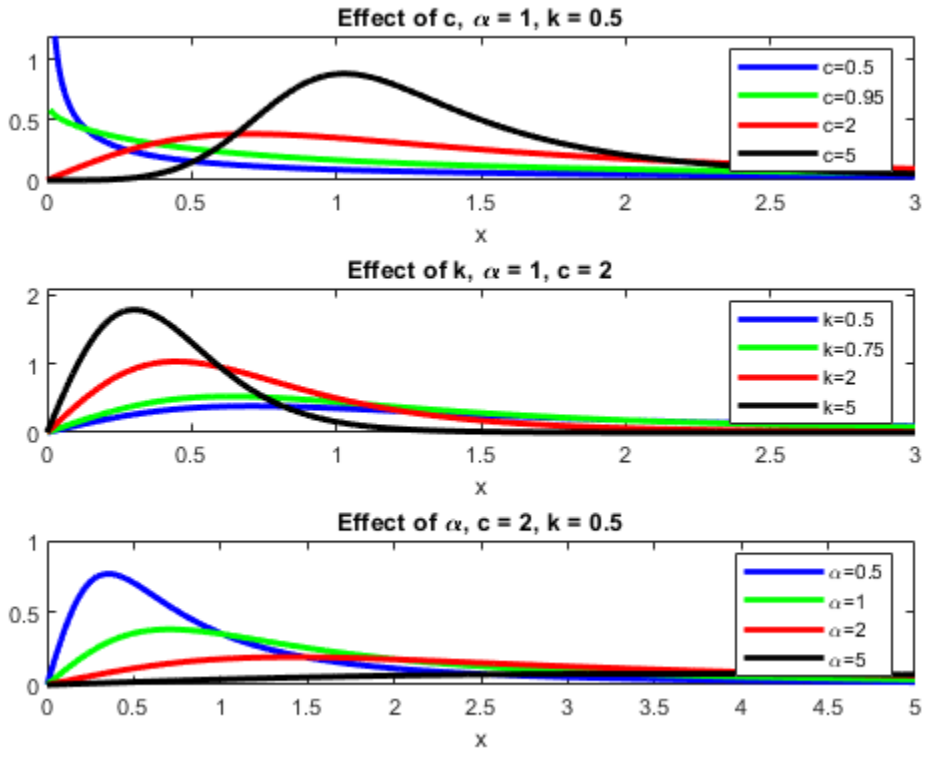


Burr pdf for Various Parameters

This example shows how to create a variety of shapes for probability density functions of the Burr distribution.

```
X = 0:0.01:5;
c = [0.5 0.95 2 5];
k = [0.5 0.75 2 5];
alpha = [0.5 1 2 5];
colors = ['b';'g';'r';'k'];

figure
for i = 1:1:4
pdf1(i,:) = pdf('burr',X,1,c(i),0.5);
pdf2(i,:) = pdf('burr',X,1,2,k(i));
```

```
pdf3(i,:) = pdf('burr',X,alpha(i),2,0.5);

axC = subplot(3,1,1);
pC(i) = plot(X,pdf1(i,:),colors(i),'LineWidth',2);
title('Effect of c, \alpha = 1, k = 0.5'),xlabel('x')
hold on

axK = subplot(3,1,2);
pK(i) = plot(X,pdf2(i,:),colors(i),'LineWidth',2);
title('Effect of k, \alpha = 1, c = 2'),xlabel('x')
hold on

axAlpha = subplot(3,1,3);
pAlpha(i) = plot(X,pdf3(i,:),colors(i),'LineWidth',2);
title('Effect of \alpha, c = 2, k = 0.5'),xlabel('x')
hold on
end

set(axC,'XLim',[0 3],'YLim',[0 1.2]);
set(axK,'XLim',[0 3],'YLim',[0 2.1]);
set(axAlpha,'XLim',[0 5],'YLim',[0 1]);

legend(axC,'c=0.5','c=0.95','c=2','c=5');
legend(axK,'k=0.5','k=0.75','k=2','k=5');
legend(axAlpha,'\alpha=0.5','\alpha=1','\alpha=2','\alpha=5');
```



This figure illustrates how the shape and scale of the Burr distribution changes for different values of its parameters.

Survival and Hazard Functions of Burr Distribution

This example shows how to find and plot the survival and hazard functions for a sample coming from a Burr distribution.

Generate the data.

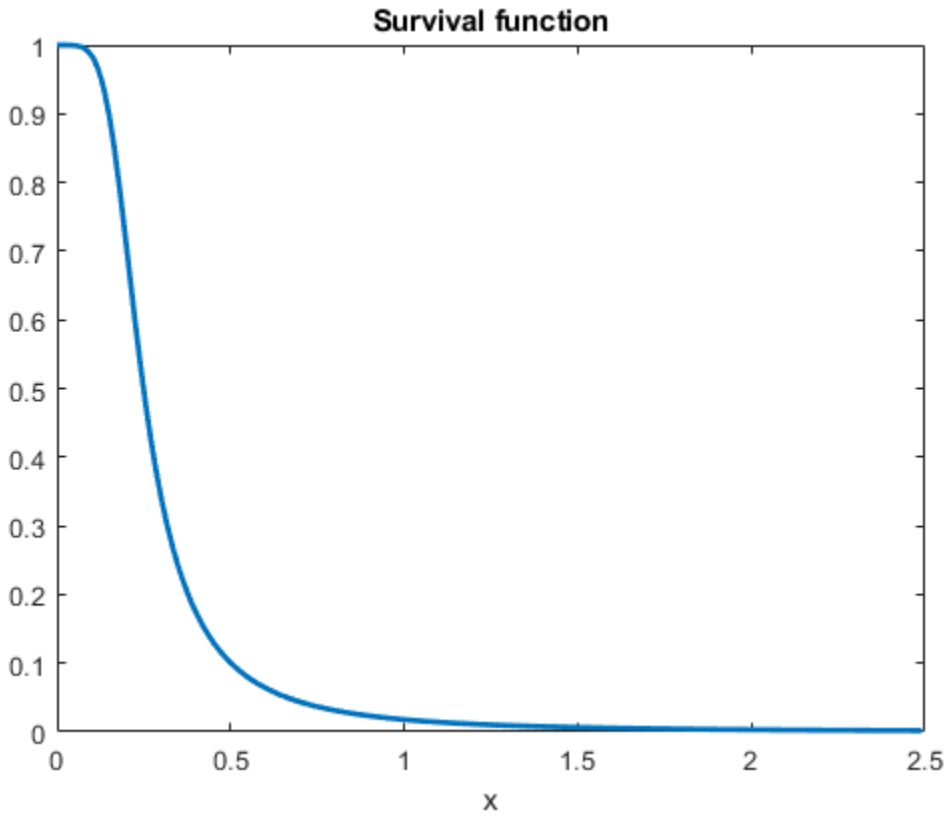
```
X = 0:0.015:2.5;
```

Evaluate the pdf and cdf of data in X.

```
Xpdf = pdf('burr',X,0.2,5,0.5);
Xcdf = cdf('burr',X,0.2,5,0.5);
```

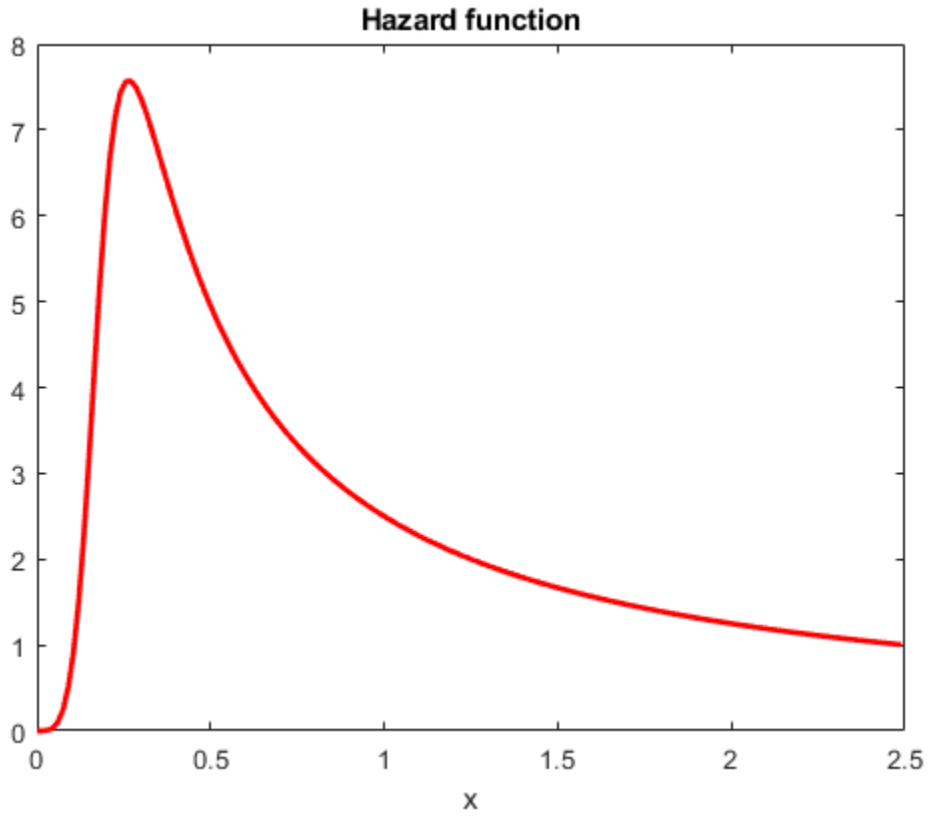
Evaluate and plot the survival function of data in X.

```
S = 1.-Xcdf; % survival function
plot(X,S,'LineWidth',2)
title('Survival function')
xlabel('x')
```



Evaluate and plot the hazard function of data in X.

```
H = Xpdf./S; % hazard function
plot(X,H,'r','LineWidth',2)
title('Hazard function')
xlabel('x')
```



Divergence of Parameter Estimates

This example shows how to interpret the display when the parameter estimates diverge when fitting a Burr distribution to input data.

1. Generate sample data from the Weibull distribution with parameters 0.5 and 2.

```
rng('default') % for reproducibility
X = wblrnd(0.5,2,100,1);
```

2. Fit a Burr distribution.

```
PD = fitdist(X,'burr');
```

```
Error using addburr>burrfit (line 566)
The data are not fit by a Burr distribution with finite parameters.
The maximum likelihood fit is provided by the k->Inf, alpha->Inf
limiting form of the Burr distribution: a Weibull distribution
with the parameters below.
    a (scale): 0.476817
    b (shape): 1.96219
```

```
Error in prob.BurrDistribution.fit (line 246)
            p = burrfit(x,0.05,cens,freq,opt);

Error in fitdist>localfit (line 238)
pd = feval(fitter,x,'cens',c,'freq',f,varargin{:});

Error in fitdist (line 185)
    pd = localfit(dist,fitter,x,cens,freq,args{:});
```

The error message tells you that the Weibull family seems to fit the data better and gives you the parameter estimates from a Weibull fit. You can use those estimates directly. If you need covariance estimates for the parameters or other information about the fit, you can refit a Weibull distribution to the data.

3. Fit a Weibull distribution to the data and find the confidence intervals for the parameter estimates.

```
PD = fitdist(X,'weibull');
paramci(PD)

ans =

    0.4291    1.6821
    0.5298    2.2890
```

These are the 95% confidence intervals of the parameter estimates for the Weibull distribution fit.

References

[1] Burr, Irving W. “Cumulative frequency functions.” *The Annals of Mathematical Statistics*, Vol. 13, Number 2, 1942, pp. 215–232.

[2] Tadikamalla, Pandu R. “A look at the Burr and related distributions.” *International Statistical Review*, Vol. 48, Number 3, 1980, pp. 337–344.

[3] Rodriguez, Robert N. “A guide to the Burr type XII distributions.” *Biometrika*, Vol. 64, Number 1, 1977, pp. 129–134.

[4] AL-Hussaini, Essam K. “A characterization of the Burr type XII distribution”. *Appl. Math. Lett.* Vol. 4, Number 1, 1991, pp. 59–61.

[5] Grammig, Joachim and Kai-Oliver Maurer. “Non-monotonic hazard functions and the autoregressive conditional duration model.” *Econometrics Journal*, Vol. 3, 2000, pp. 16–38.

See Also

`BurrDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Chi-Square Distribution

In this section...

Overview

The chi-square (χ^2) distribution is a one-parameter family of curves. The chi-square distribution is commonly used in hypothesis testing, particularly the chi-square test for goodness of fit.

Statistics and Machine Learning Toolbox offers multiple ways to work with the chi-square distribution.

- Use distribution-specific functions (`chi2cdf`, `chi2inv`, `chi2pdf`, `chi2rnd`, `chi2stat`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple chi-square distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Chisquare'`) and parameters.

Parameters

The chi-square distribution uses the following parameter.

Parameter	Description	Support
nu (ν)	Degrees of freedom	ν = 1, 2, 3,...

The degrees of freedom parameter is typically an integer, but chi-square functions accept any positive value.

The sum of two chi-square random variables with degrees of freedom ν_1 and ν_2 is a chi-square random variable with degrees of freedom $\nu = \nu_1 + \nu_2$.

Probability Density Function

The probability density function (pdf) of the chi-square distribution is

$$y = f(x|\nu) = \frac{x^{(\nu-2)/2}e^{-x/2}}{2^{\frac{\nu}{2}}\Gamma(\nu/2)},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function.

For an example, see “Compute Chi-Square Distribution pdf” on page B-29.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the chi-square distribution is

$$p = F(x|\nu) = \int_0^x \frac{t^{(\nu-2)/2}e^{-t/2}}{2^{\nu/2}\Gamma(\nu/2)}dt,$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result p is the probability that a single observation from the chi-square distribution with ν degrees of freedom falls in the interval $[0, x]$.

For an example, see “Compute Chi-Square Distribution cdf” on page B-30.

Inverse Cumulative Distribution Function

The inverse cumulative distribution function (icdf) of the chi-square distribution is

$$x = F^{-1}(p|\nu) = \{x : F(x|\nu) = p\},$$

where

$$p = F(x|\nu) = \int_0^x \frac{t^{(\nu-2)/2}e^{-t/2}}{2^{\nu/2}\Gamma(\nu/2)}dt,$$

ν is the degrees of freedom, and $\Gamma(\cdot)$ is the Gamma function. The result p is the probability that a single observation from the chi-square distribution with ν degrees of freedom falls in the interval $[0, x]$.

Descriptive Statistics

The mean of the chi-square distribution is ν.

The variance of the chi-square distribution is 2ν.

Examples

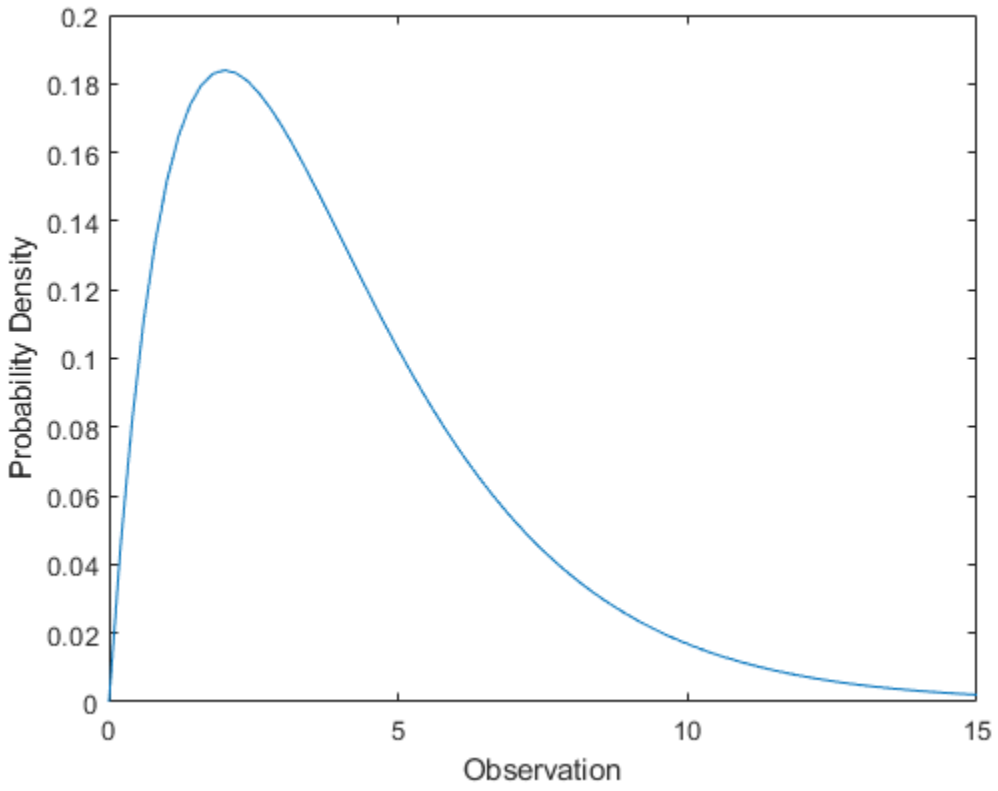
Compute Chi-Square Distribution pdf

Compute the pdf of a chi-square distribution with 4 degrees of freedom.

```
x = 0:0.2:15;
y = chi2pdf(x,4);
```

Plot the pdf.

```
figure;
plot(x,y)
xlabel('Observation')
ylabel('Probability Density')
```



The chi-square distribution is skewed to the right, especially for few degrees of freedom.

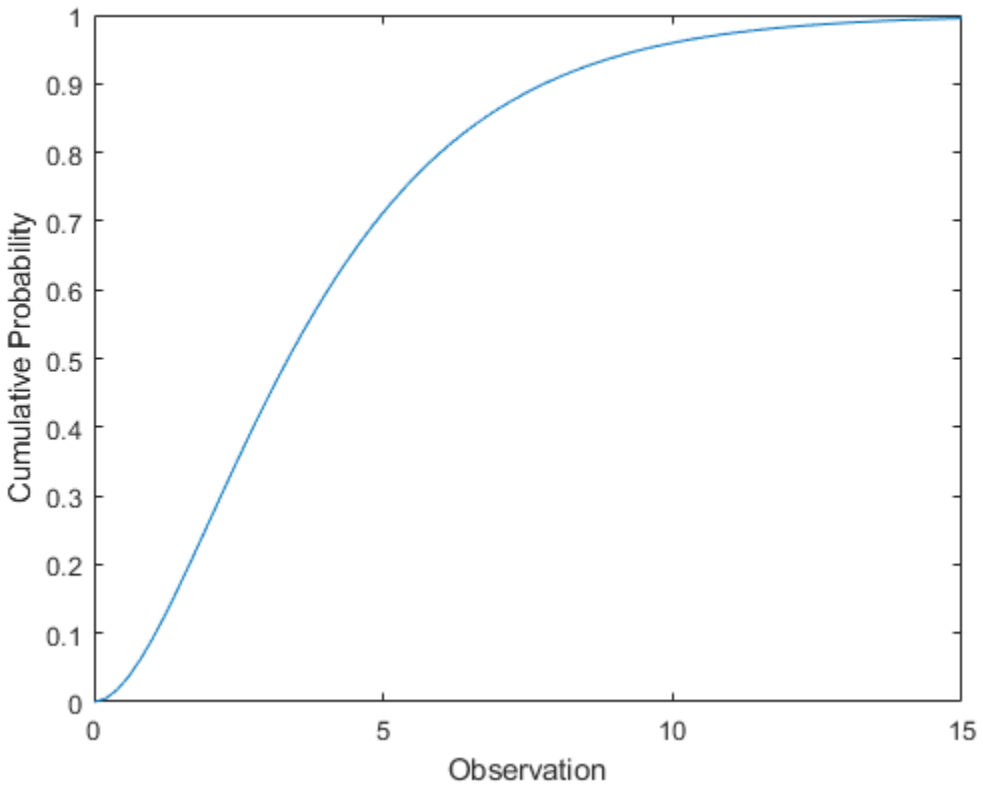
Compute Chi-Square Distribution cdf

Compute the cdf of a chi-square distribution with 4 degrees of freedom.

```
x = 0:0.2:15;
y = chi2cdf(x,4);
```

Plot the cdf.

```
figure;
plot(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Related Distributions

- "F Distribution" on page B-45 — The F distribution is a two-parameter distribution that has parameters ν_1 (numerator degrees of freedom) and ν_2 (denominator degrees of freedom). The F distribution can be defined as the ratio $F = \frac{\chi^2_1 / \nu_1}{\chi^2_2 / \nu_2}$, where ${\chi^2}_1$ and ${\chi^2}_2$ are both chi-square distributed with ν_1 and ν_2 degrees of freedom, respectively.
- "Gamma Distribution" on page B-47 — The gamma distribution is a two-parameter continuous distribution that has parameters a (shape) and b (scale). The chi-square distribution is equal to the gamma distribution with $2a = \nu$ and $b = 2$.
- "Noncentral Chi-Square Distribution" on page B-113 — The noncentral chi-square distribution is a two-parameter continuous distribution that has parameters ν (degrees of freedom) and δ (noncentrality). The noncentral chi-square distribution is equal to the chi-square distribution when $\delta = 0$.
- "Normal Distribution" on page B-119 — The normal distribution is a two-parameter continuous distribution that has parameters μ (mean) and σ (standard deviation). The standard normal distribution occurs when $\mu = 0$ and $\sigma = 1$.

If Z_1, Z_2, ..., Z_n are standard normal random variables, then $\sum_{i=1}^{n} Z_i^2$ has a chi-square distribution with degrees of freedom $\nu = n - 1$.

If a set of n observations is normally distributed with variance σ^2 and sample variance s^2, then $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution with degrees of freedom $\nu = n - 1$. This relationship is used to calculate confidence intervals for the estimate of the normal parameter σ^2 in the function `normfit`.

- "Student's t Distribution" on page B-149 — The Student's t distribution is a one-parameter continuous distribution that has parameter ν (degrees of freedom). If Z has a standard normal distribution and χ^2 has a chi-square distribution with degrees of freedom ν, then $t = \frac{Z}{\sqrt{\chi^2/\nu}}$ has a Student's t distribution with degrees of freedom ν.
- "Wishart Distribution" on page B-178 — The Wishart distribution is a higher dimensional analog of the chi-square distribution.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[4] Kreyszig, Erwin. *Introductory Mathematical Statistics: Principles and Methods*. New York: Wiley, 1970.

See Also

`chi2cdf` | `chi2gof` | `chi2inv` | `chi2pdf` | `chi2rnd` | `chi2stat`

More About

- "Working with Probability Distributions" on page 5-3
- "Supported Distributions" on page 5-14

Exponential Distribution

In this section...

Overview

The exponential distribution is a one-parameter family of curves. The exponential distribution models wait times when the probability of waiting an additional period of time is independent of how long you have already waited. For example, the probability that a light bulb will burn out in its next minute of use is relatively independent of how many minutes it has already burned.

Statistics and Machine Learning Toolbox offers several ways to work with the exponential distribution.

- Create a probability distribution object `ExponentialDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the exponential distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`expcdf`, `exppdf`, `expinv`, `explike`, `expstat`, `expfit`, `exprnd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple exponential distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Exponential'`) and parameters.

Parameters

The exponential distribution uses the following parameter.

Parameter	Description	Support
`mu` (μ)	Mean	μ `> 0`

The parameter μ is also equal to the standard deviation of the exponential distribution.

The standard exponential distribution has μ=1.

A common alternative parameterization of the exponential distribution is to use λ defined as the mean number of events in an interval as opposed to μ, which is the mean wait time for an event to occur. λ and μ are reciprocals.

Parameter Estimation

The *likelihood function* is the probability density function (pdf) viewed as a function of the parameters. The *maximum likelihood estimates* (MLEs) are the parameter estimates that maximize the likelihood function for fixed values of `x`.

The maximum likelihood estimator of μ for the exponential distribution is $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$, where $\bar{x}$ is the sample mean for samples x_1, x_2, ..., x_n. The sample mean is an unbiased estimator of the parameter μ.

To fit the exponential distribution to data and find a parameter estimate, use `expfit`, `fitdist`, or `mle`. Unlike `expfit` and `mle`, which return parameter estimates, `fitdist` returns the fitted probability distribution object `ExponentialDistribution`. The object property `mu` stores the parameter estimate.

For an example, see “Fit Exponential Distribution to Data” on page B-35.

Probability Density Function

The pdf of the exponential distribution is

$$y = f(x \mid \mu) = \frac{1}{\mu} e^{\frac{-x}{\mu}}.$$

For an example, see “Compute Exponential Distribution pdf” on page B-35.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the exponential distribution is

$$p = F(x \mid u) = \int_0^x \frac{1}{\mu} e^{\frac{-t}{\mu}} dt = 1 - e^{\frac{-x}{\mu}}.$$

The result p is the probability that a single observation from the exponential distribution with mean μ falls in the interval $[0, x]$.

For an example, see “Compute Exponential Distribution cdf” on page B-36.

Inverse Cumulative Distribution Function

The inverse cumulative distribution function (icdf) of the exponential distribution is

$$x = F^{-1}(p \mid \mu) = -\mu \ln(1 - p).$$

The result x is the value such that an observation from an exponential distribution with parameter μ falls in the range $[0\ x]$ with probability p.

Hazard Function

The hazard function (instantaneous failure rate) is the ratio of the pdf and the complement of the cdf. If $f(t)$ and $F(t)$ are the pdf and cdf of a distribution (respectively), then the hazard rate is

$h(t) = \frac{f(t)}{1 - F(t)}$. Substituting the pdf and cdf of the exponential distribution for $f(t)$ and $F(t)$ yields a constant λ. The exponential distribution is the only continuous distribution with a constant hazard function. λ is the reciprocal of μ and can be interpreted as the rate at which events occur in any given interval. Consequently, when you model survival times, the probability that an item will survive an extra unit of time is independent of the current age of the item.

For an example, see “Exponentially Distributed Lifetimes” on page B-37.

Examples

Fit Exponential Distribution to Data

Generate a sample of `100` of exponentially distributed random numbers with mean `700`.

```
x = exprnd(700,100,1); % Generate sample
```

Fit an exponential distribution to data using `fitdist`.

```
pd = fitdist(x,'exponential')
```

```
pd =
  ExponentialDistribution

  Exponential distribution
    mu = 641.934   [532.598, 788.966]
```

`fitdist` returns an `ExponentialDistribution` object. The interval next to the parameter estimate is the 95% confidence interval for the distribution parameter.

Estimate the parameter using the distribution functions.

```
[muhat,muci] = expfit(x) % Distribution specific function
```

```
muhat = 641.9342

muci = 2×1

  532.5976
  788.9660
```

```
[muhat2,muci2] = mle(x,'distribution','exponential') % Generic distribution function
```

```
muhat2 = 641.9342

muci2 = 2×1

  532.5976
  788.9660
```

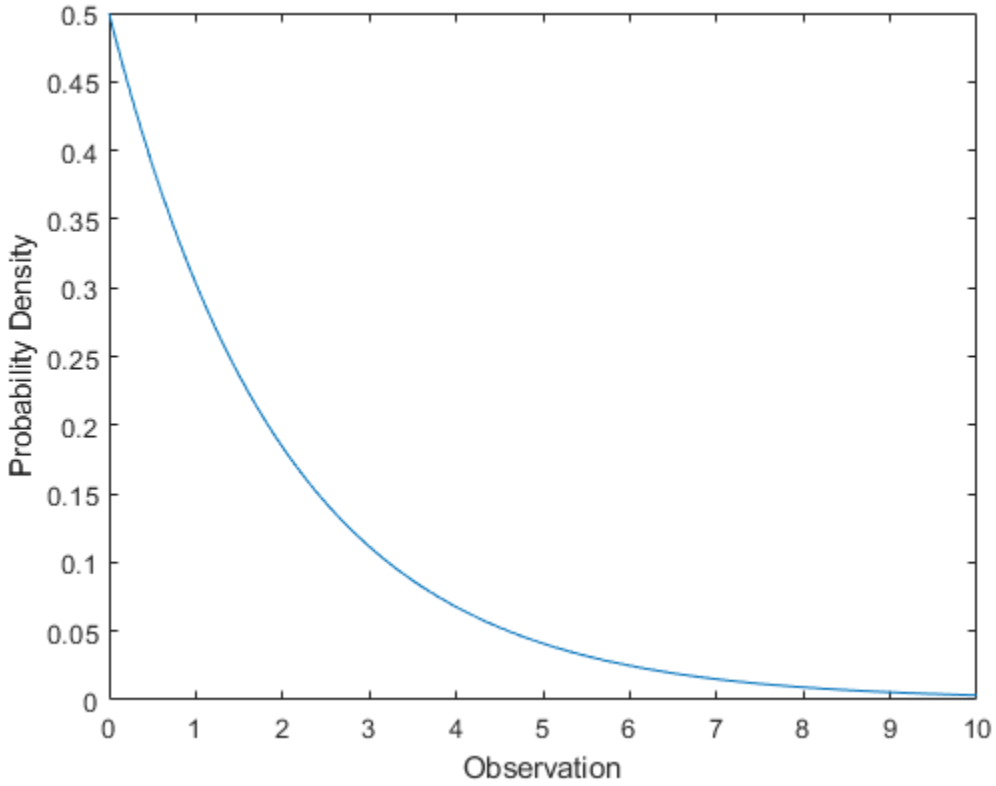
Compute Exponential Distribution pdf

Compute the pdf of an exponential distribution with parameter `mu = 2`.

```
x = 0:0.1:10;
y = exppdf(x,2);
```

Plot the pdf.

```
figure;
plot(x,y)
xlabel('Observation')
ylabel('Probability Density')
```



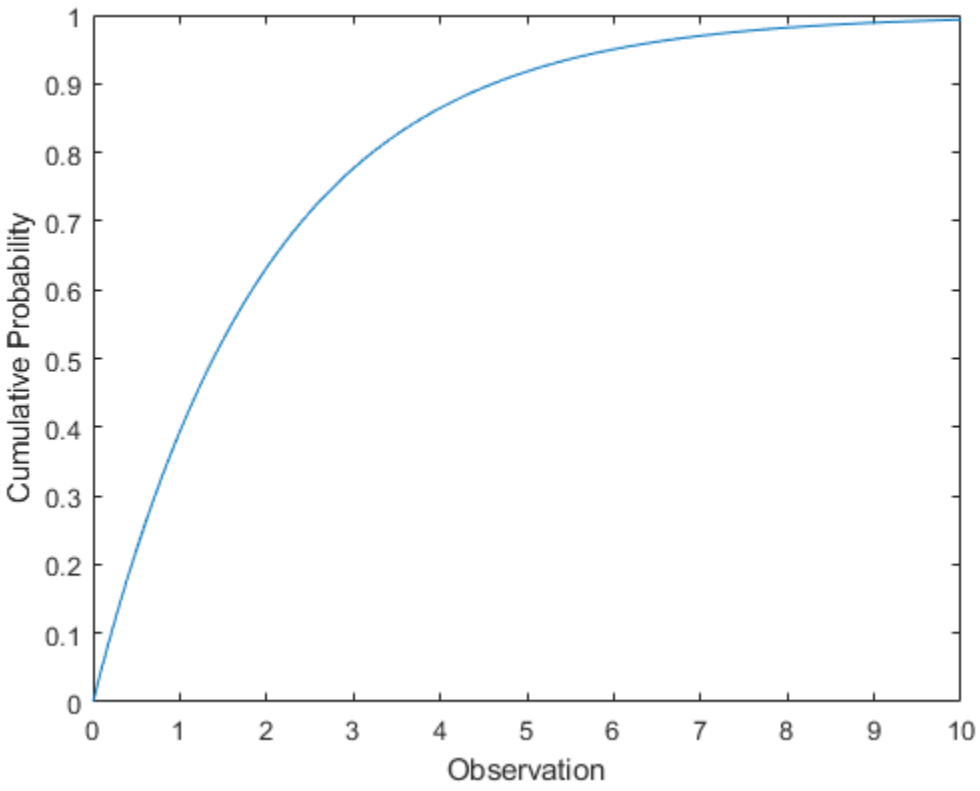
Compute Exponential Distribution cdf

Compute the cdf of an exponential distribution with parameter `mu = 2`.

```
x = 0:0.1:10;
y = expcdf(x,2);
```

Plot the cdf.

```
figure;
plot(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Exponentially Distributed Lifetimes

Compute the hazard function of the exponential distribution with mean `mu = 2` at the values one through five.

```
x = 1:5;
lambda1 = exppdf(x,2)./(1-expcdf(x,2))
```

```
lambda1 = 1×5

    0.5000    0.5000    0.5000    0.5000    0.5000
```

The hazard function (instantaneous rate of failure to survival) of the exponential distribution is constant and always equals `1/mu`. This constant is often denoted by λ.

Evaluate the hazard functions of the exponential distributions with means one through five at `x = 3`.

```
mu = 1:5;
lambda2 = exppdf(3,mu)./(1-expcdf(3,mu))
```

```
lambda2 = 1×5

    1.0000    0.5000    0.3333    0.2500    0.2000
```

The probability that an item with an exponentially distributed lifetime survive one more unit of time is independent of how long it has survived.

Compute the probability of an item surviving one more year at various ages when the mean survival time is `10` years.

```
x2 = 5:5:25;
x3 = x2 + 1;
deltap = (expcdf(x3,10)-expcdf(x2,10))./(1-expcdf(x2,10))
```

```
deltap = 1×5

    0.0952    0.0952    0.0952    0.0952    0.0952
```

The probability of surviving one more year is the same regardless of how long an item has already survived.

Related Distributions

- "Burr Type XII Distribution" on page B-19 — The Burr distribution is a three-parameter continuous distribution. An exponential distribution compounded with a gamma distribution on the mean yields a Burr distribution.
- "Gamma Distribution" on page B-47 — The gamma distribution is a two-parameter continuous distribution that has parameters a (shape) and b (scale). When $a = 1$, the gamma distribution is equal to the exponential distribution with mean $\mu = b$. The sum of k exponentially distributed random variables with mean μ has a gamma distribution with parameters $a = k$ and $\mu = b$.
- "Geometric Distribution" on page B-63 — The geometric distribution is a one-parameter discrete distribution that models the total number of failures before the first success in repeated Bernoulli trials. The geometric distribution is a discrete analog of the exponential distribution and is the only discrete distribution with a constant hazard function.
- "Generalized Pareto Distribution" on page B-59 — The generalized Pareto distribution is a three-parameter continuous distribution that has parameters k (shape), σ (scale), and θ (threshold). When both $k = 0$ and $\theta = 0$, the generalized Pareto distribution is equal to the exponential distribution with mean $\mu = \sigma$.
- "Poisson Distribution" on page B-131 — The Poisson distribution is a one-parameter discrete distribution that takes nonnegative integer values. The parameter λ is both the mean and the variance of the distribution. The Poisson distribution models counts of the number of times a random event occurs in a given amount of time. In such a model, the amount of time between occurrences is modeled by the exponential distribution with mean $\frac{1}{\lambda}$.
- "Weibull Distribution" on page B-170 — The Weibull distribution is a two-parameter continuous distribution that has parameters a (scale) and b (shape). The Weibull distribution is also used to model lifetimes, but it does not have a constant hazard rate. When $b = 1$, the Weibull distribution is equal to the exponential distribution with mean $\mu = a$.

  For an example, see "Compare Exponential and Weibull Distribution Hazard Functions" on page B-174.

References

[1] Crowder, Martin J., ed. *Statistical Analysis of Reliability Data*. Reprinted. London: Chapman & Hall, 1995.

[2] Kotz, Samuel, and Saralees Nadarajah. *Extreme Value Distributions: Theory and Applications*. London : River Edge, NJ: Imperial College Press; Distributed by World Scientific, 2000.

[3] Meeker, William Q., and Luis A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Applied Probability and Statistics Section. New York: Wiley, 1998.

[4] Lawless, Jerald F. *Statistical Models and Methods for Lifetime Data*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley-Interscience, 2003.

See Also

`ExponentialDistribution` | `expcdf` | `expfit` | `expinv` | `explike` | `exppdf` | `exprnd` | `expstat` | `fitdist` | `makedist`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Extreme Value Distribution

In this section...

Definition

The probability density function for the extreme value distribution with location parameter μ and scale parameter σ is

$$y = f(x|\mu,\sigma) = \sigma^{-1}\exp\left(\frac{x-\mu}{\sigma}\right)\exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)$$

This form of the probability density function is suitable for modeling the minimum value. To model the maximum value, use the negative of the original values.

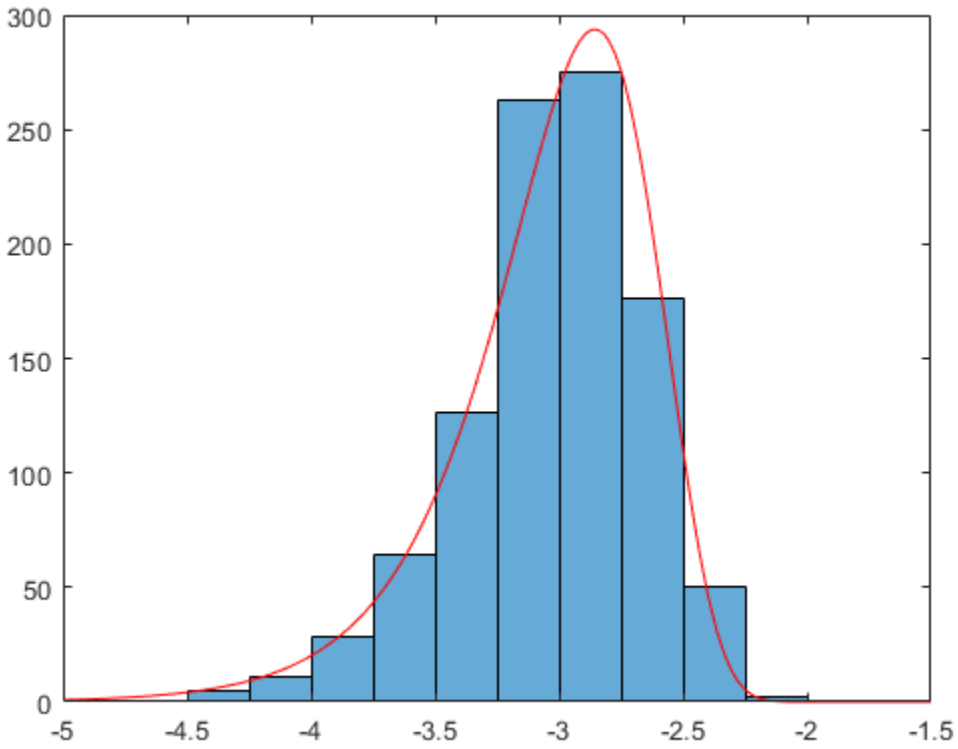
If T has a Weibull distribution on page B-170 with parameters a and b, then $\log T$ has an extreme value distribution with parameters $\mu = \log a$ and $\sigma = 1/b$.

Background

Extreme value distributions are often used to model the smallest or largest value among a large set of independent, identically distributed random values representing measurements or observations. The extreme value distribution is appropriate for modeling the smallest value from a distribution whose tails decay exponentially fast, such as, the normal distribution. It can also model the largest value from a distribution, such as the normal or exponential distributions, by using the negative of the original values.

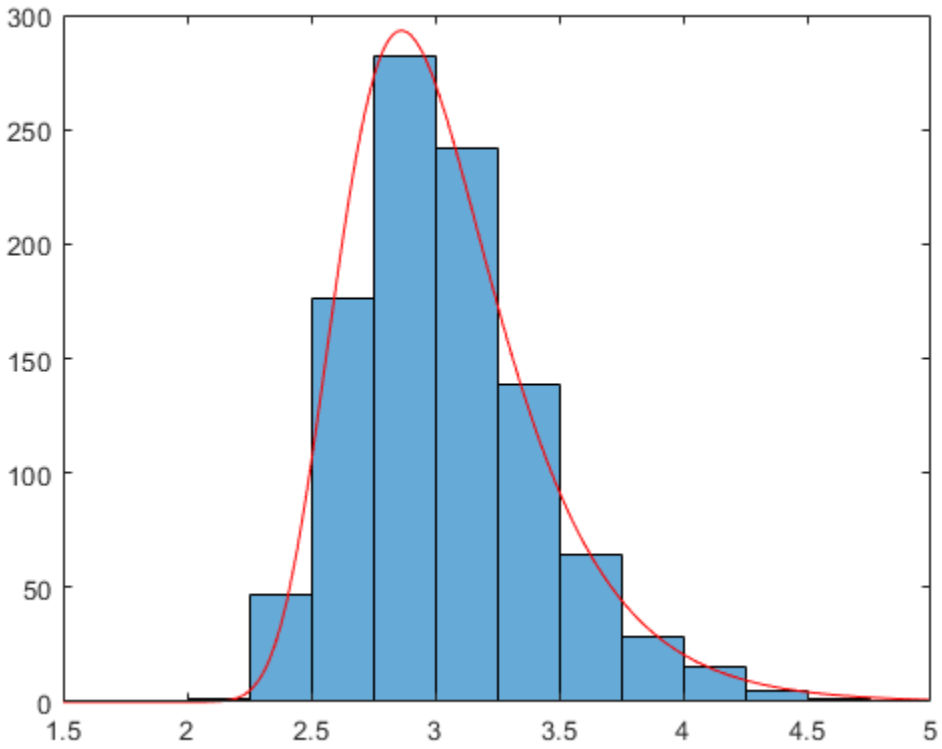
For example, the following fits an extreme value distribution to minimum values taken over 1000 sets of 500 observations from a normal distribution.

```
rng default;  % For reproducibility
xMinima = min(randn(1000,500), [], 2);
paramEstsMinima = evfit(xMinima);
y = linspace(-5,-1.5,1001);
histogram(xMinima,-4.75:.25:-1.75);
p = evpdf(y,paramEstsMinima(1),paramEstsMinima(2));
line(y,.25*length(xMinima)*p,'color','r')
```



The following fits an extreme value distribution to the maximum values in each set of observations.

```
rng default;  % For reproducibility
xMaxima = max(randn(1000,500), [], 2);
paramEstsMaxima = evfit(-xMaxima);
y = linspace(1.5,5,1001);
histogram(xMaxima,1.75:.25:4.75);
p = evpdf(-y,paramEstsMaxima(1),paramEstsMaxima(2));
line(y,.25*length(xMaxima)*p,'color','r')
```



Although the extreme value distribution is most often used as a model for extreme values, you can also use it as a model for other types of continuous data. For example, extreme value distributions are closely related to the Weibull distribution. If `T` has a Weibull distribution, then `log(T)` has a type 1 extreme value distribution.

Parameters

The function `evfit` returns the maximum likelihood estimates (MLEs) and confidence intervals for the parameters of the extreme value distribution. The following example shows how to fit some sample data using `evfit`, including estimates of the mean and variance from the fitted distribution.

Suppose you want to model the size of the smallest washer in each batch of 1000 from a manufacturing process. If you believe that the sizes are independent within and between each batch, you can fit an extreme value distribution to measurements of the minimum diameter from a series of eight experimental batches. The following code returns the MLEs of the distribution parameters as `parmhat` and the confidence intervals as the columns of `parmci`.

```
x = [19.774 20.141 19.44 20.511 21.377 19.003 19.66 18.83];
[parmhat, parmci] = evfit(x)
```

```
parmhat =
   20.2506    0.8223

parmci =
   19.644 0.49861
   20.857 1.3562
```

You can find mean and variance of the extreme value distribution with these parameters using the function `evstat`.

```
[meanfit, varfit] = evstat(parmhat(1),parmhat(2))

meanfit =
   19.776

varfit =
   1.1123
```

Examples

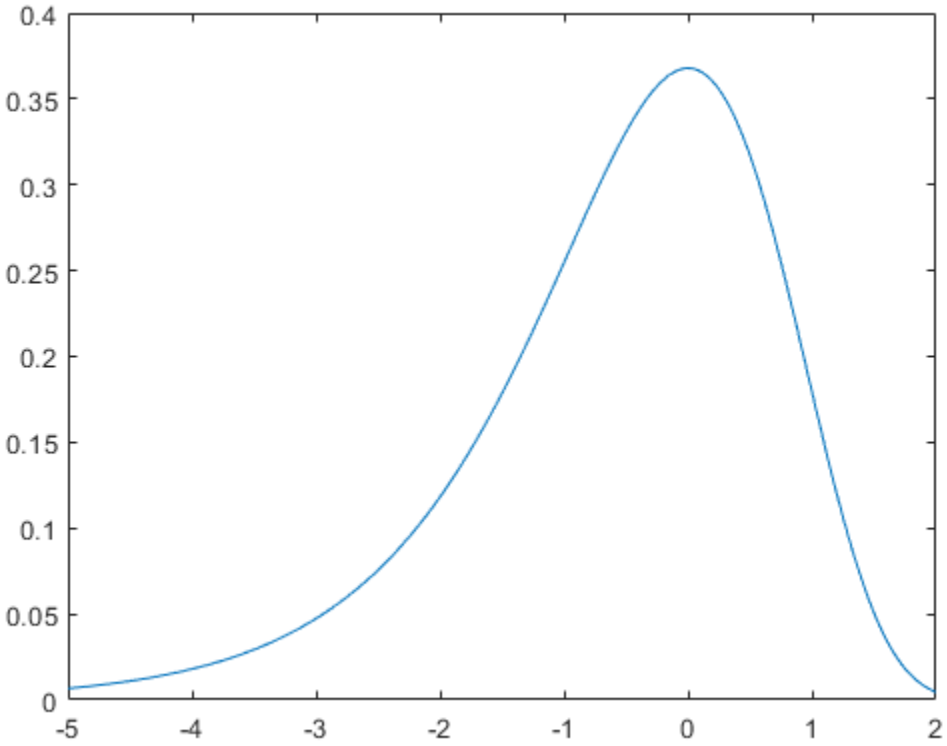
Compute the Extreme Value Distribution pdf

Compute the pdf of an extreme value distribution.

```
t = [-5:.01:2];
y = evpdf(t);
```

Plot the pdf.

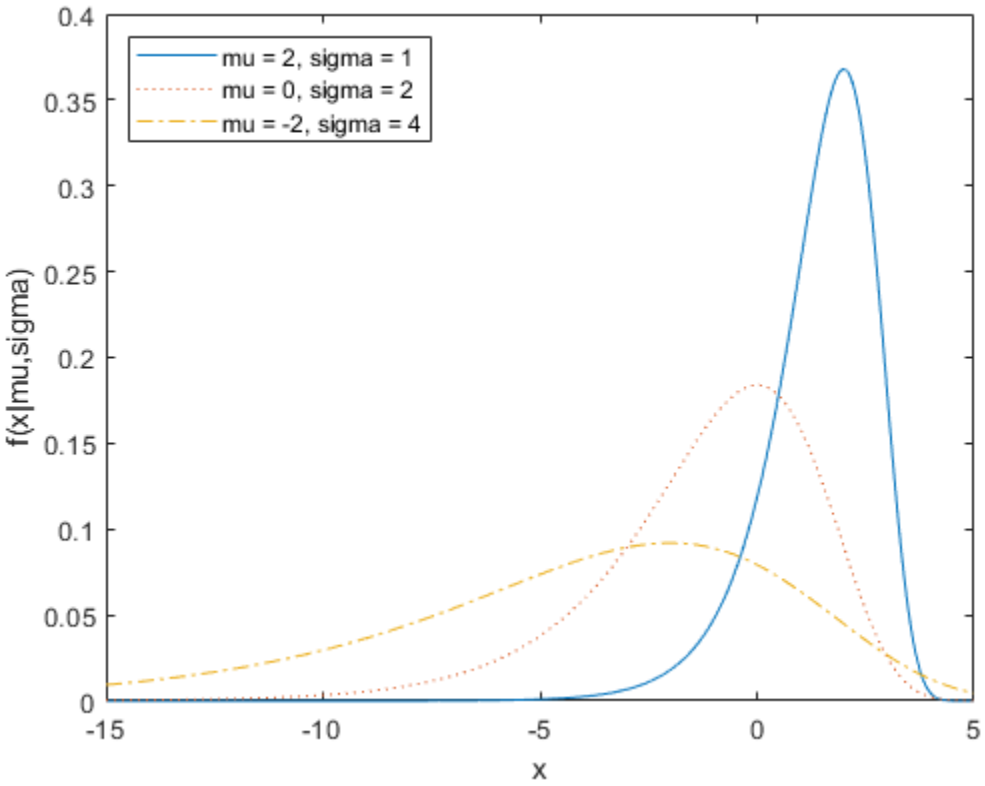
```
figure;
plot(t,y)
```



The extreme value distribution is skewed to the left, and its general shape remains the same for all parameter values. The location parameter, `mu`, shifts the distribution along the real line, and the scale parameter, `sigma`, expands or contracts the distribution.

The following plots the probability function for different combinations of `mu` and `sigma`.

```
x = -15:.01:5;
plot(x,evpdf(x,2,1),'-', ...
     x,evpdf(x,0,2),':', ...
     x,evpdf(x,-2,4),'-.');
legend({'mu = 2, sigma = 1', ...
        'mu = 0, sigma = 2', ...
        'mu = -2, sigma = 4'}, ...
       'Location','NW')
xlabel('x')
ylabel('f(x|mu,sigma)')
```



See Also

`ExtremeValueDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

F Distribution

In this section...

Definition

The pdf for the F distribution is

$$y = f(x|\nu_1,\nu_2) = \frac{\Gamma\left[\frac{(\nu_1+\nu_2)}{2}\right]}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\frac{x^{\frac{\nu_1-2}{2}}}{\left[1+\left(\frac{\nu_1}{\nu_2}\right)x\right]^{\frac{\nu_1+\nu_2}{2}}}$$

where $\Gamma(\cdot)$ is the Gamma function.

Background

The F distribution has a natural relationship with the chi-square distribution. If χ_1 and χ_2 are both chi-square with ν_1 and ν_2 degrees of freedom respectively, then the statistic F below is F-distributed.

$$F(\nu_1,\nu_2) = \frac{\chi_1/\nu_1}{\chi_2/\nu_2}$$

The two parameters, ν_1 and ν_2, are the numerator and denominator degrees of freedom. That is, ν_1 and ν_2 are the number of independent pieces of information used to calculate χ_1 and χ_2, respectively.

Examples

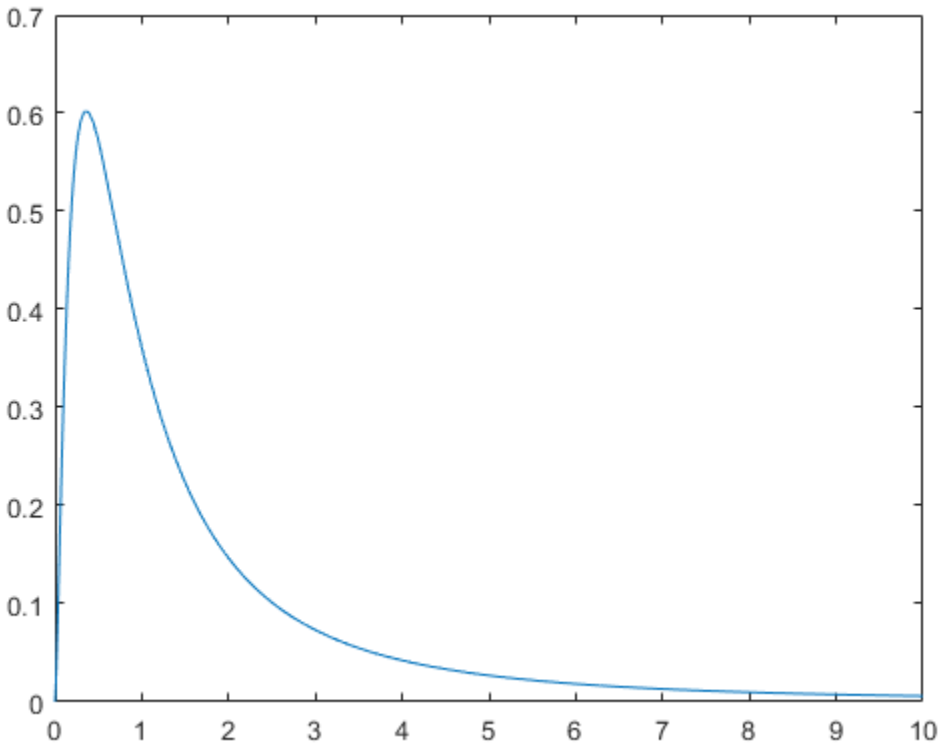
Compute the F Distribution pdf

Compute the pdf of an F distribution with 5 numerator degrees of freedom and 3 denominator degrees of freedom.

```
x = 0:0.01:10;
y = fpdf(x,5,3);
```

Plot the pdf.

```
figure;
plot(x,y)
```



The plot shows that the *F* distribution exists on positive real numbers and is skewed to the right.

See Also

`fcdf` | `finv` | `fpdf` | `frnd` | `fstat` | `random`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Gamma Distribution

In this section...

Overview

The gamma distribution is a two-parameter family of curves. The gamma distribution models sums of exponentially distributed random variables and generalizes both the chi-square and exponential distributions.

Statistics and Machine Learning Toolbox offers several ways to work with the gamma distribution.

- Create a probability distribution object `GammaDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the gamma distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`gamcdf`, `gampdf`, `gaminv`, `gamlike`, `gamstat`, `gamfit`, `gamrnd`, `randg`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple gamma distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Gamma'`) and parameters.

Parameters

The gamma distribution uses the following parameters.

Parameter	Description	Support
`a`	Shape	a `> 0`
`b`	Scale	b `> 0`

The standard gamma distribution has unit scale.

The sum of two gamma random variables with shape parameters a_1 and a_2 both with scale parameter b is a gamma random variable with shape parameter $a = a_1 + a_2$ and scale parameter b.

Parameter Estimation

The *likelihood function* is the probability density function (pdf) viewed as a function of the parameters. The *maximum likelihood estimates* (MLEs) are the parameter estimates that maximize the likelihood function for fixed values of `x`.

The maximum likelihood estimators of a and b for the gamma distribution are the solutions to the simultaneous equations

$$\log\widehat{a} - \psi(\widehat{a}) = \log\left(\bar{x}/\left(\prod_{i=1}^{n} x_i\right)^{1/n}\right)$$

$$\widehat{b} = \frac{\bar{x}}{\widehat{a}}$$

where $\bar{x}$ is the sample mean for the sample x_1, x_2, ..., x_n, and Ψ is the digamma function `psi`.

To fit the gamma distribution to data and find parameter estimates, use `gamfit`, `fitdist`, or `mle`. Unlike `gamfit` and `mle`, which return parameter estimates, `fitdist` returns the fitted probability distribution object `GammaDistribution`. The object properties `a` and `b` store the parameter estimates.

For an example, see “Fit Gamma Distribution to Data” on page B-49.

Probability Density Function

The pdf of the gamma distribution is

$$y = f(x\,|\,a,b) = \frac{1}{b^a\Gamma(a)}x^{a-1}e^{\frac{-x}{b}},$$

where $\Gamma(\cdot)$ is the Gamma function.

For an example, see “Compute Gamma Distribution pdf” on page B-50.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the gamma distribution is

$$p = F(x\,|\,a,b) = \frac{1}{b^a\Gamma(a)}\int_0^x t^{a-1}e^{\frac{-t}{b}}dt\,.$$

The result p is the probability that a single observation from the gamma distribution with parameters a and b falls in the interval $[0\ x]$.

For an example, see “Compute Gamma Distribution cdf” on page B-51.

The gamma cdf is related to the incomplete gamma function `gammainc` by

$$f(x\,|\,a,b) = \text{gammainc}\left(\frac{x}{b},a\right).$$

Inverse Cumulative Distribution Function

The inverse cumulative distribution function (icdf) of the gamma distribution in terms of the gamma cdf is

$$x = F^{-1}(p|a,b) = \left\{x : F(x|a,b) = p\right\},$$

where

$$p = F(x|a,b) = \frac{1}{b^a \Gamma(a)} \int_0^x t^{a-1} e^{\frac{-t}{b}} dt\,.$$

The result x is the value such that an observation from the gamma distribution with parameters a and b falls in the range $[0\ x]$ with probability p.

The preceding integral equation has no known analytical solution. `gaminv` uses an iterative approach (Newton's method) to converge on the solution.

Descriptive Statistics

The mean of the gamma distribution is ab.

The variance of the gamma distribution is ab^2.

Examples

Fit Gamma Distribution to Data

Generate a sample of `100` gamma random numbers with shape `3` and scale `5`.

```
x = gamrnd(3,5,100,1);
```

Fit a gamma distribution to data using `fitdist`.

```
pd = fitdist(x,'gamma')
```

```
pd =
  GammaDistribution

  Gamma distribution
    a =  2.7783   [2.1374, 3.61137]
    b = 5.73438   [4.30198, 7.64372]
```

`fitdist` returns a `GammaDistribution` object. The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

Estimate the parameters `a` and `b` using the distribution functions.

```
[muhat,muci] = gamfit(x) % Distribution specific function
```

```
muhat = 1×2
```

```
    2.7783    5.7344


muci = 2×2

    2.1374    4.3020
    3.6114    7.6437


[muhat2,muci2] = mle(x,'distribution','gamma') % Generic function

muhat2 = 1×2

    2.7783    5.7344


muci2 = 2×2

    2.1374    4.3020
    3.6114    7.6437
```

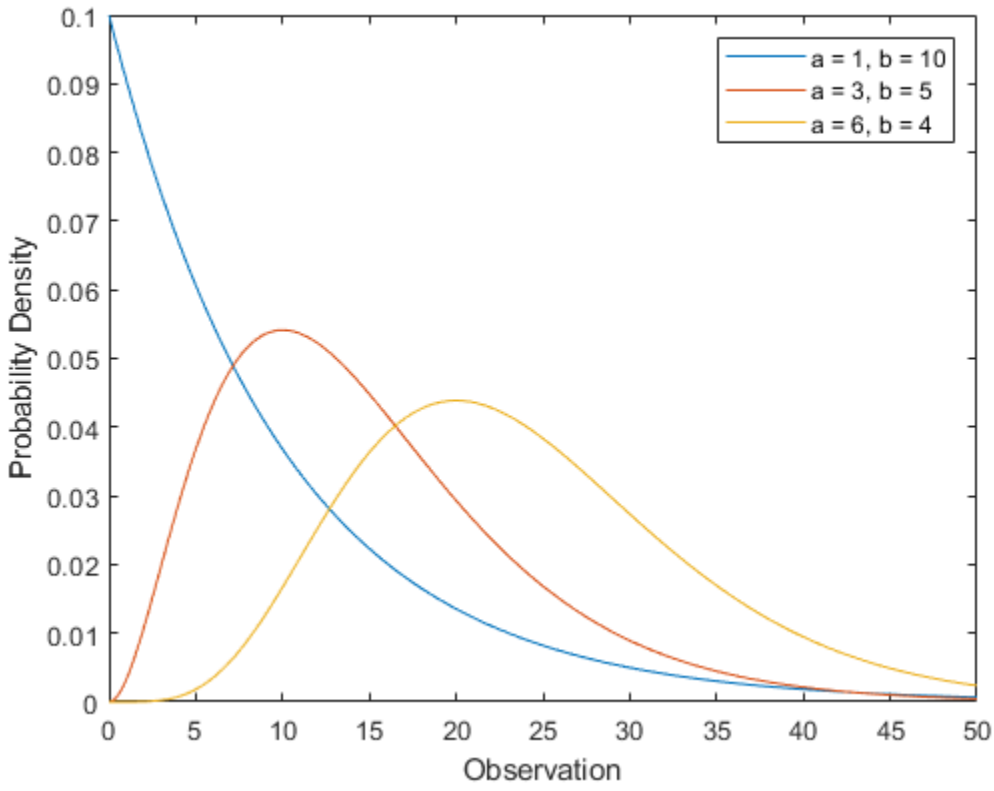
Compute Gamma Distribution pdf

Compute the pdfs of the gamma distribution with several shape and scale parameters.

```
x = 0:0.1:50;
y1 = gampdf(x,1,10);
y2 = gampdf(x,3,5);
y3 = gampdf(x,6,4);
```

Plot the pdfs.

```
figure;
plot(x,y1)
hold on
plot(x,y2)
plot(x,y3)
hold off
xlabel('Observation')
ylabel('Probability Density')
legend('a = 1, b = 10','a = 3, b = 5','a = 6, b = 4')
```



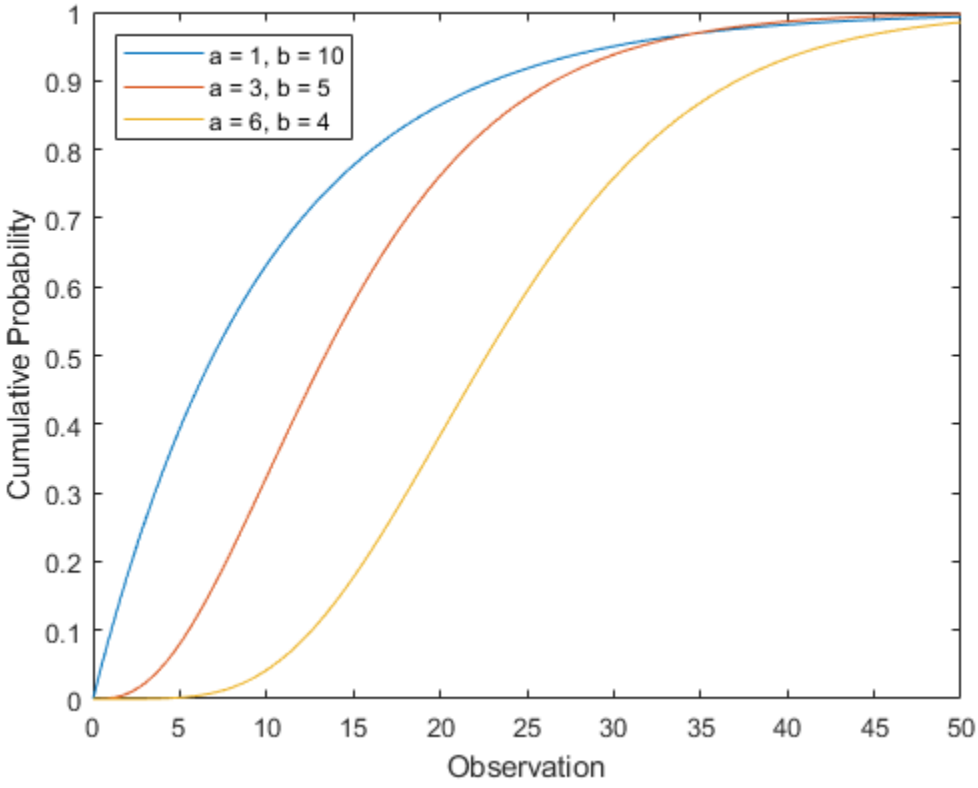
Compute Gamma Distribution cdf

Compute the cdfs of the gamma distribution with several shape and scale parameters.

```
x = 0:0.1:50;
y1 = gamcdf(x,1,10);
y2 = gamcdf(x,3,5);
y3 = gamcdf(x,6,4);
```

Plot the cdfs.

```
figure;
plot(x,y1)
hold on
plot(x,y2)
plot(x,y3)
hold off
xlabel('Observation')
ylabel('Cumulative Probability')
legend('a = 1, b = 10','a = 3, b = 5','a = 6, b = 4',"Location","northwest")
```



Compare Gamma and Normal Distribution pdfs

The gamma distribution has the shape parameter a and the scale parameter b. For a large a, the gamma distribution closely approximates the normal distribution with mean $\mu = ab$ and variance $\sigma^2 = ab^2$.

Compute the pdf of a gamma distribution with parameters `a = 100` and `b = 5`.

```
a = 100;
b = 5;
x = 250:750;
y_gam = gampdf(x,a,b);
```

For comparison, compute the mean, standard deviation, and pdf of the normal distribution that gamma approximates.

```
mu = a*b
```

```
mu = 500
```

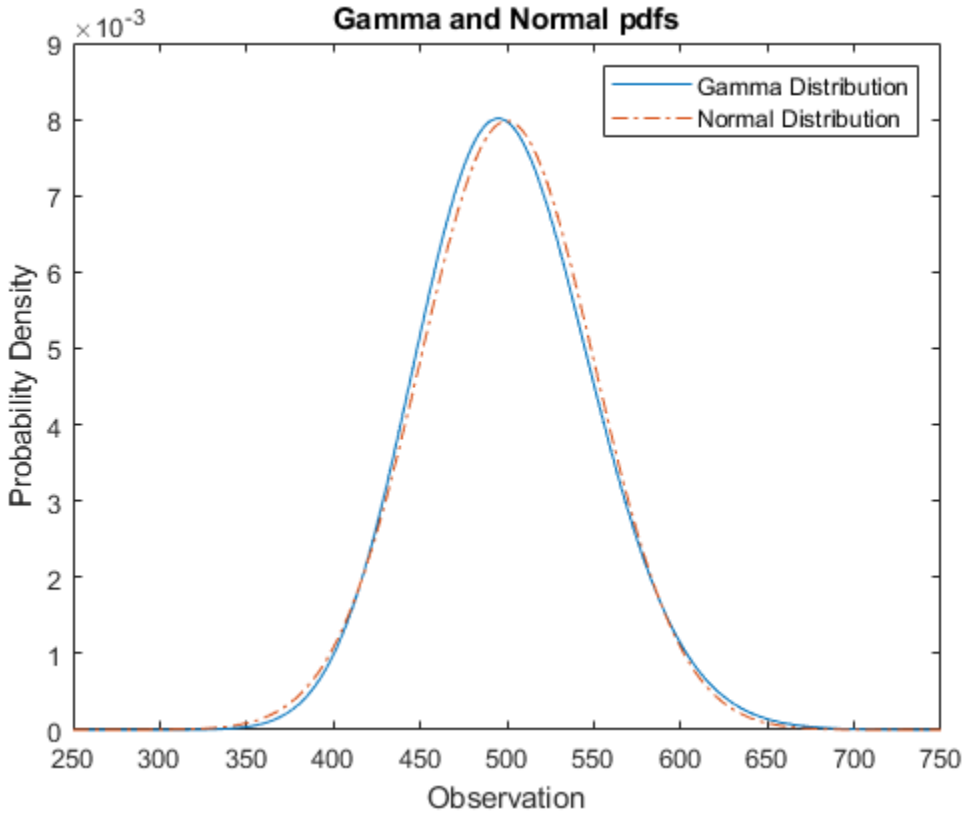
```
sigma = sqrt(a*b^2)
```

```
sigma = 50
```

```
y_norm = normpdf(x,mu,sigma);
```

Plot the pdfs of the gamma distribution and the normal distribution on the same figure.

```
plot(x,y_gam,'-',x,y_norm,'-.')
title('Gamma and Normal pdfs')
xlabel('Observation')
ylabel('Probability Density')
legend('Gamma Distribution','Normal Distribution')
```



The pdf of the normal distribution approximates the pdf of the gamma distribution.

Related Distributions

- "Beta Distribution" on page B-6 — The beta distribution is a two-parameter continuous distribution that has parameters a (first shape parameter) and b (second shape parameter). If X_1 and X_2 have standard gamma distributions with shape parameters a_1 and a_2 respectively, then $Y = \frac{X_1}{X_1 + X_2}$ has a beta distribution with shape parameters a_1 and a_2.
- "Chi-Square Distribution" on page B-28 — The chi-square distribution is a one-parameter continuous distribution that has parameter ν (degrees of freedom). The chi-square distribution is equal to the gamma distribution with $2a = \nu$ and $b = 2$.
- "Exponential Distribution" on page B-33 — The exponential distribution is a one-parameter continuous distribution that has parameter μ (mean). The exponential distribution is equal to the gamma distribution with $a = 1$ and $b = \mu$. The sum of k exponentially distributed random variables with mean μ is the gamma distribution with parameters $a = k$ and $\mu = b$.

- “Nakagami Distribution” on page B-108 — The Nakagami distribution is a two-parameter continuous distribution with shape parameter μ and scale parameter ω. If x has a Nakagami distribution, then x^2 has a gamma distribution with $a = \mu$ and $ab = \omega$.
- “Normal Distribution” on page B-119 — The normal distribution is a two-parameter continuous distribution that has parameters μ (mean) and σ (standard deviation). When a is large, the gamma distribution closely approximates a normal distribution with $\mu = ab$ and $\sigma^2 = ab^2$. For an example, see “Compare Gamma and Normal Distribution pdfs” on page B-52.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

[3] Hahn, Gerald J., and Samuel S. Shapiro. *Statistical Models in Engineering*. Wiley Classics Library. New York: Wiley, 1994.

[4] Lawless, Jerald F. *Statistical Models and Methods for Lifetime Data*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley-Interscience, 2003.

[5] Meeker, William Q., and Luis A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Applied Probability and Statistics Section. New York: Wiley, 1998.

[6] Marsaglia, George, and Wai Wan Tsang. “A Simple Method for Generating Gamma Variables.” *ACM Transactions on Mathematical Software* 26, no. 3 (September 1, 2000): 363–72. https://doi.org/10.1007/978-1-4613-8643-8.

See Also

`GammaDistribution` | `fitdist` | `gamcdf` | `gamfit` | `gaminv` | `gamlike` | `gampdf` | `gamrnd` | `gamstat` | `makedist` | `randg`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Generalized Extreme Value Distribution

In this section...

Definition

The probability density function for the generalized extreme value distribution with location parameter μ, scale parameter σ, and shape parameter $k \neq 0$ is

$$y = f(x|k,\mu,\sigma) = \left(\frac{1}{\sigma}\right)\exp\left(-\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right)\left(1+k\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{k}}$$

for

$$1 + k\frac{(x-\mu)}{\sigma} > 0$$

$k > 0$ corresponds to the Type II case, while $k < 0$ corresponds to the Type III case. For $k = 0$, corresponding to the Type I case, the density is

$$y = f(x|0,\mu,\sigma) = \left(\frac{1}{\sigma}\right)\exp\left(-\exp\left(-\frac{(x-\mu)}{\sigma}\right)-\frac{(x-\mu)}{\sigma}\right)$$

Background

Like the extreme value distribution, the generalized extreme value distribution is often used to model the smallest or largest value among a large set of independent, identically distributed random values representing measurements or observations. For example, you might have batches of 1000 washers from a manufacturing process. If you record the size of the largest washer in each batch, the data are known as block maxima (or minima if you record the smallest). You can use the generalized extreme value distribution as a model for those block maxima.

The generalized extreme value combines three simpler distributions into a single form, allowing a continuous range of possible shapes that includes all three of the simpler distributions. You can use any one of those distributions to model a particular dataset of block maxima. The generalized extreme value distribution allows you to “let the data decide” which distribution is appropriate.

The three cases covered by the generalized extreme value distribution are often referred to as the Types I, II, and III. Each type corresponds to the limiting distribution of block maxima from a different class of underlying distributions. Distributions whose tails decrease exponentially, such as the normal, lead to the Type I. Distributions whose tails decrease as a polynomial, such as Student's *t*, lead to the Type II. Distributions whose tails are finite, such as the beta, lead to the Type III.

Types I, II, and III are sometimes also referred to as the Gumbel, Frechet, and Weibull types, though this terminology can be slightly confusing. The Type I (Gumbel) and Type III (Weibull) cases actually correspond to the mirror images of the usual Gumbel and Weibull distributions, for example, as

computed by the functions `evcdf` and `evfit` , or `wblcdf` and `wblfit`, respectively. Finally, the Type II (Frechet) case is equivalent to taking the reciprocal of values from a standard Weibull distribution.

Parameters

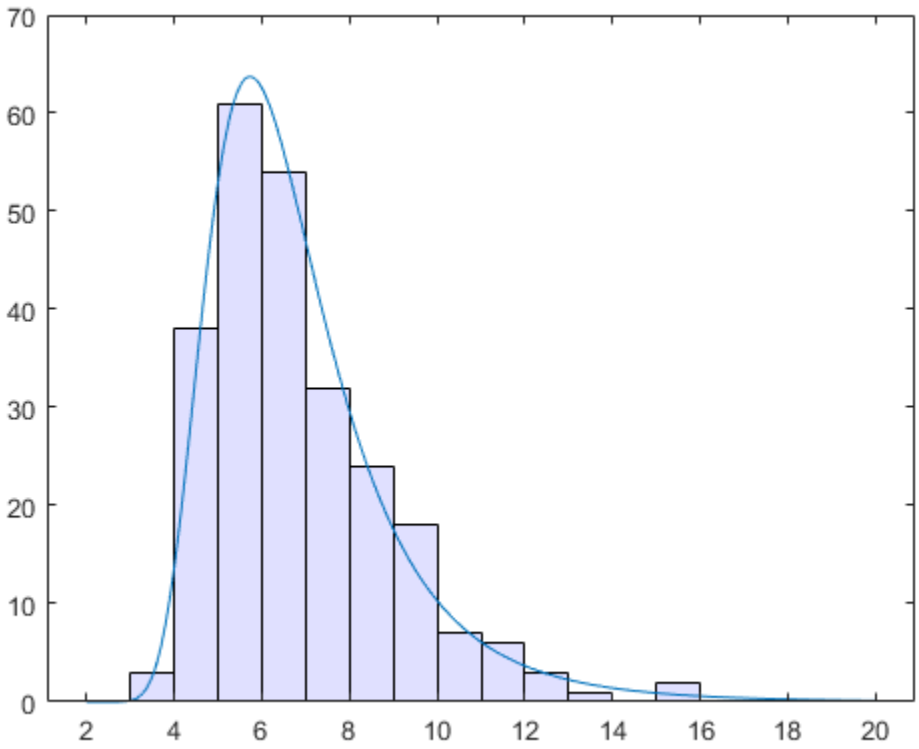
If you generate 250 blocks of 1000 random values drawn from Student's *t* distribution with 5 degrees of freedom, and take their maxima, you can fit a generalized extreme value distribution to those maxima.

```
blocksize = 1000;
nblocks = 250;
rng default  % For reproducibility
t = trnd(5,blocksize,nblocks);
x = max(t); % 250 column maxima
paramEsts = gevfit(x)
```

```
paramEsts = 1×3

    0.1185    1.4530    5.8929
```

Notice that the shape parameter estimate (the first element) is positive, which is what you would expect based on block maxima from a Student's *t* distribution.

```
histogram(x,2:20,'FaceColor',[.8 .8 1]);
xgrid = linspace(2,20,1000);
line(xgrid,nblocks*...
     gevpdf(xgrid,paramEsts(1),paramEsts(2),paramEsts(3)));
```

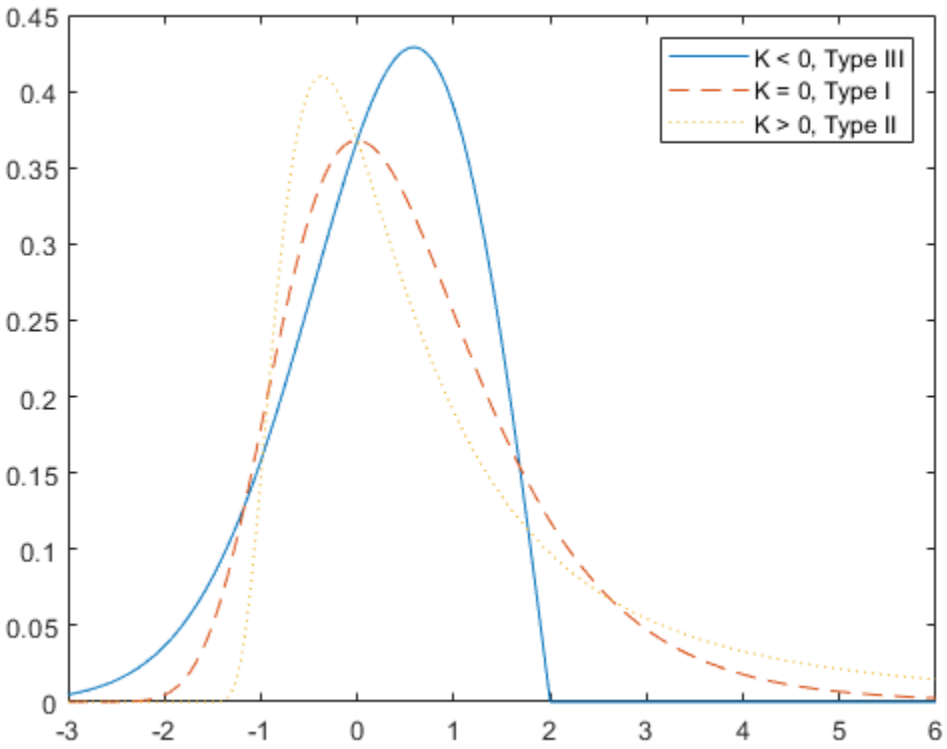


Examples

Compute the Generalized Extreme Value Distribution pdf

Generate examples of probability density functions for the three basic forms of the generalized extreme value distribution.

```
x = linspace(-3,6,1000);
y1 = gevpdf(x,-.5,1,0);
y2 = gevpdf(x,0,1,0);
y3 = gevpdf(x,.5,1,0);
plot(x,y1,'-', x,y2,'--', x,y3,':')
legend({'K < 0, Type III' 'K = 0, Type I' 'K > 0, Type II'})
```



Notice that for `k > 0`, the distribution has zero probability density for `x` such that $x < -\sigma/k + \mu$.

For `k < 0`, the distribution has zero probability density for $x > -\sigma/k + \mu$.

For `k = 0`, there is no upper or lower bound.

See Also

`GeneralizedExtremeValueDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Generalized Pareto Distribution

In this section...

Definition

The probability density function for the generalized Pareto distribution with shape parameter $k \neq 0$, scale parameter σ, and threshold parameter θ, is

$$y = f(x \mid k, \sigma, \theta) = \left(\frac{1}{\sigma}\right)\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}}$$

for $\theta < x$, when $k > 0$, or for $\theta < x < \theta - \sigma/k$ when $k < 0$.

For $k = 0$, the density is

$$y = f(x \mid 0, \sigma, \theta) = \left(\frac{1}{\sigma}\right)e^{-\frac{(x-\theta)}{\sigma}}$$

for $\theta < x$.

If $k = 0$ and $\theta = 0$, the generalized Pareto distribution is equivalent to the exponential distribution. If $k > 0$ and $\theta = \sigma/k$, the generalized Pareto distribution is equivalent to the Pareto distribution with a scale parameter equal to σ/k and a shape parameter equal to $1/k$.

Background

Like the exponential distribution, the generalized Pareto distribution is often used to model the tails of another distribution. For example, you might have washers from a manufacturing process. If random influences in the process lead to differences in the sizes of the washers, a standard probability distribution, such as the normal, could be used to model those sizes. However, while the normal distribution might be a good model near its mode, it might not be a good fit to real data in the tails and a more complex model might be needed to describe the full range of the data. On the other hand, only recording the sizes of washers larger (or smaller) than a certain threshold means you can fit a separate model to those tail data, which are known as *exceedances*. You can use the generalized Pareto distribution in this way, to provide a good fit to extremes of complicated data.

The generalized Pareto distribution allows a continuous range of possible shapes that includes both the exponential and Pareto distributions as special cases. You can use either of those distributions to model a particular dataset of exceedances. The generalized Pareto distribution allows you to “let the data decide” which distribution is appropriate.

The generalized Pareto distribution has three basic forms, each corresponding to a limiting distribution of exceedance data from a different class of underlying distributions.

- Distributions whose tails decrease exponentially, such as the normal, lead to a generalized Pareto shape parameter of zero.
- Distributions whose tails decrease as a polynomial, such as Student's *t*, lead to a positive shape parameter.
- Distributions whose tails are finite, such as the beta, lead to a negative shape parameter.

The generalized Pareto distribution is used in the tails of distribution fit objects of the `paretotails` object.

Parameters

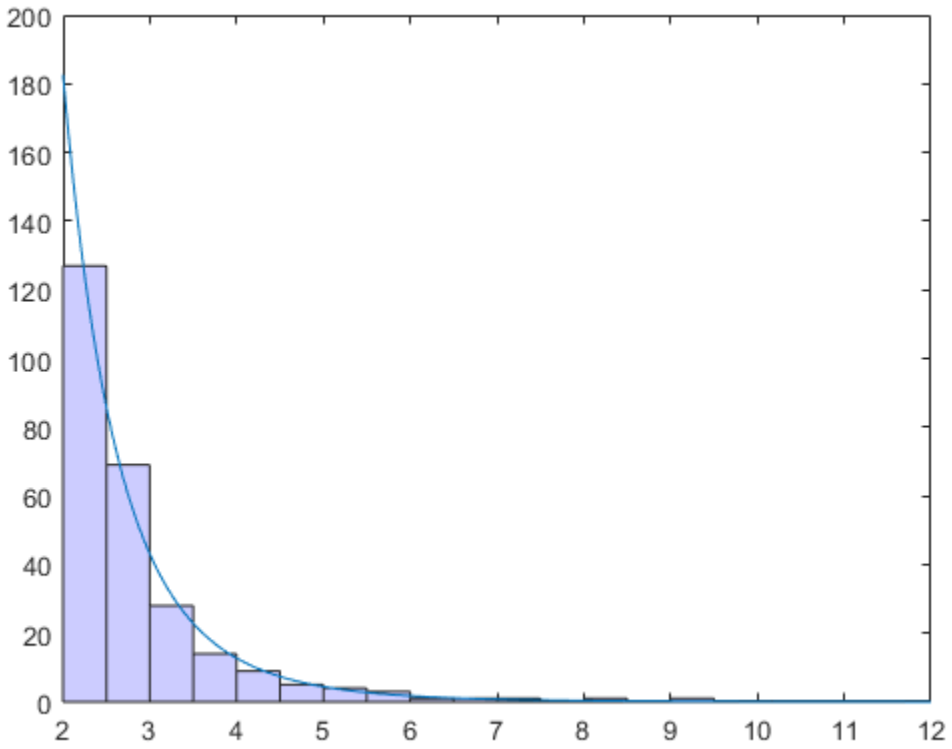
If you generate a large number of random values from a Student's *t* distribution with 5 degrees of freedom, and then discard everything less than 2, you can fit a generalized Pareto distribution to those exceedances.

```
rng default  % For reproducibility
t = trnd(5,5000,1);
y = t(t > 2) - 2;
paramEsts = gpfit(y)
```

```
paramEsts = 1×2

    0.1445    0.7225
```

Notice that the shape parameter estimate (the first element) is positive, which is what you would expect based on exceedances from a Student's *t* distribution.

```
hist(y+2,2.25:.5:11.75);
h = findobj(gca,'Type','patch');
h.FaceColor = [.8 .8 1];
xgrid = linspace(2,12,1000);
line(xgrid,.5*length(y)*...
     gppdf(xgrid,paramEsts(1),paramEsts(2),2));
```



Examples

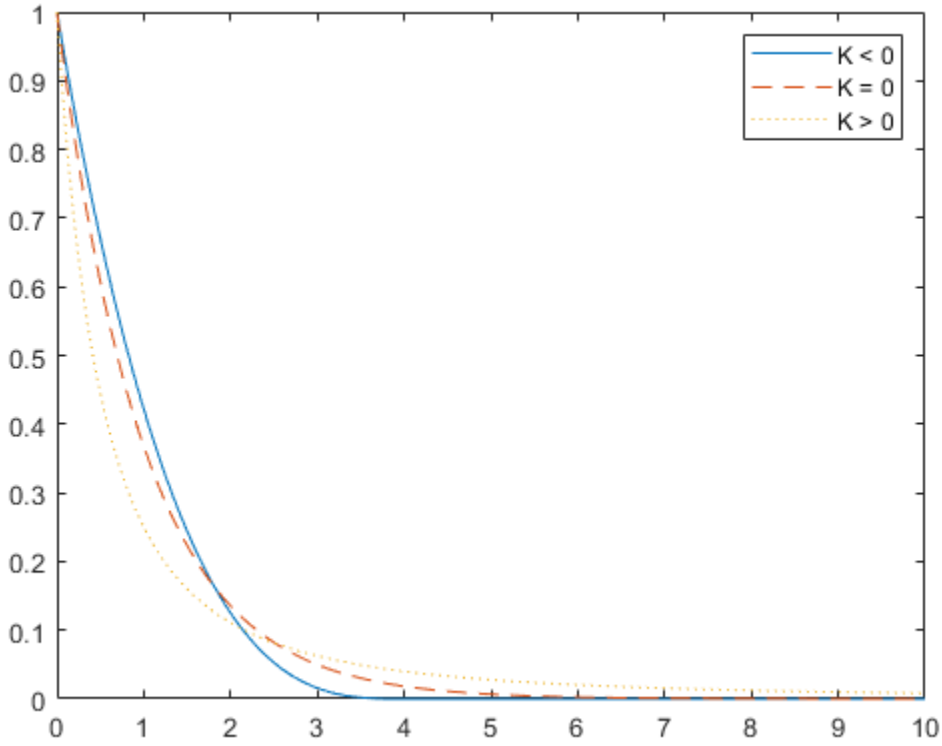
Compute Generalized Pareto Distribution pdf

Compute the pdf of three generalized Pareto distributions. The first has shape parameter `k = -0.25`, the second has `k = 0`, and the third has `k = 1`.

```
x = linspace(0,10,1000);
y1 = gppdf(x,-.25,1,0);
y2 = gppdf(x,0,1,0);
y3 = gppdf(x,1,1,0);
```

Plot the three pdfs on the same figure.

```
figure;
plot(x,y1,'-', x,y2,'--', x,y3,':')
legend({'K < 0' 'K = 0' 'K > 0'});
```



References

[1] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[2] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

See Also

`GeneralizedParetoDistribution`

More About

- “Fit a Nonparametric Distribution with Pareto Tails” on page 5-43
- “Nonparametric and Empirical Probability Distributions” on page 5-30
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Geometric Distribution

In this section...

Overview

The geometric distribution is a one-parameter family of curves that models the number of failures before one success in a series of independent trials, where each trial results in either success or failure, and the probability of success in any individual trial is constant. For example, if you toss a coin, the geometric distribution models the number of tails observed before the result is heads. The geometric distribution is discrete, existing only on the nonnegative integers.

Statistics and Machine Learning Toolbox offers multiple ways to work with the geometric distribution.

- Use distribution-specific functions (`geocdf`, `geopdf`, `geoinv`, `geostat`, `geornd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple geometric distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `mle`, `random`) with a specified distribution name (`'Geometric'`) and parameters.

Parameters

The geometric distribution uses the following parameter.

Parameter	Description	Support
p	Probability of success	$0 \le p \le 1$

Probability Density Function

The probability density function (pdf) of the geometric distribution is

$$y = f(x \mid p) = p(1-p)^x \quad ; \quad x = 0, 1, 2, \ldots ,$$

where p is the probability of success, and x is the number of failures before the first success. The result y is the probability of observing exactly x trials before a success, when the probability of success in any given trial is p. For discrete distributions, the pdf is also known as the probability mass function (pmf).

For an example, see “Compute Geometric Distribution pdf” on page B-64.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the geometric distribution is

$$y = F(x|p) = 1 - (1-p)^{x+1};\ x = 0, 1, 2, \ldots,$$

where p is the probability of success, and x is the number of failures before the first success. The result y is the probability of observing up to x trials before a success, when the probability of success in any given trial is p.

For an example, see “Compute Geometric Distribution cdf” on page B-65.

Descriptive Statistics

The mean of the geometric distribution is mean $= \frac{1-p}{p}$, and the variance of the geometric distribution is var $= \frac{1-p}{p^2}$, where p is the probability of success.

Hazard Function

The hazard function (instantaneous failure rate) is the ratio of the pdf and the complement of the cdf. If $f(t)$ and $F(t)$ are the pdf and cdf of a distribution (respectively), then the hazard rate is $h(t) = \frac{f(t)}{1-F(t)}$. Substituting the pdf and cdf of the geometric distribution for $f(t)$ and $F(t)$ above yields a constant equal to the reciprocal of the mean. The geometric distribution is the only discrete distribution with constant hazard function. Consequently, the probability of observing a success is independent of the number of failures already observed.

Examples

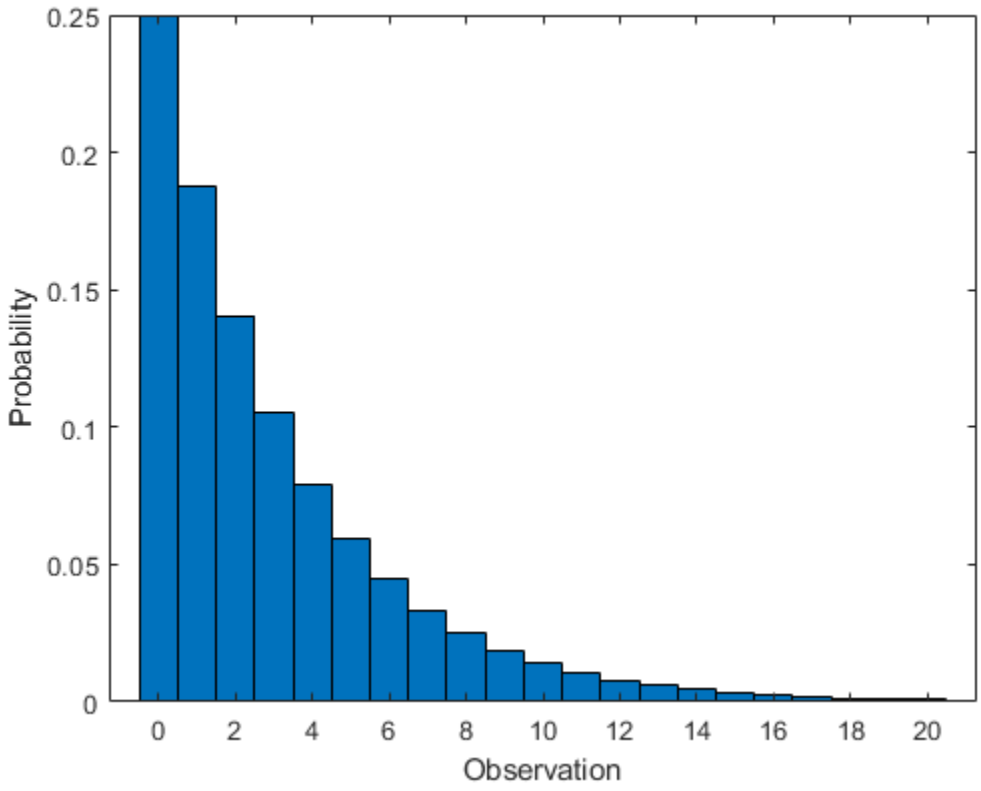
Compute Geometric Distribution pdf

Compute the pdf of the geometric distribution with the probability of success `0.25`.

```
x = 0:20;
y = geopdf(x,0.25);
```

Plot the pdf with bars of width `1`.

```
figure
bar(x,y,1)
xlabel('Observation')
ylabel('Probability')
```



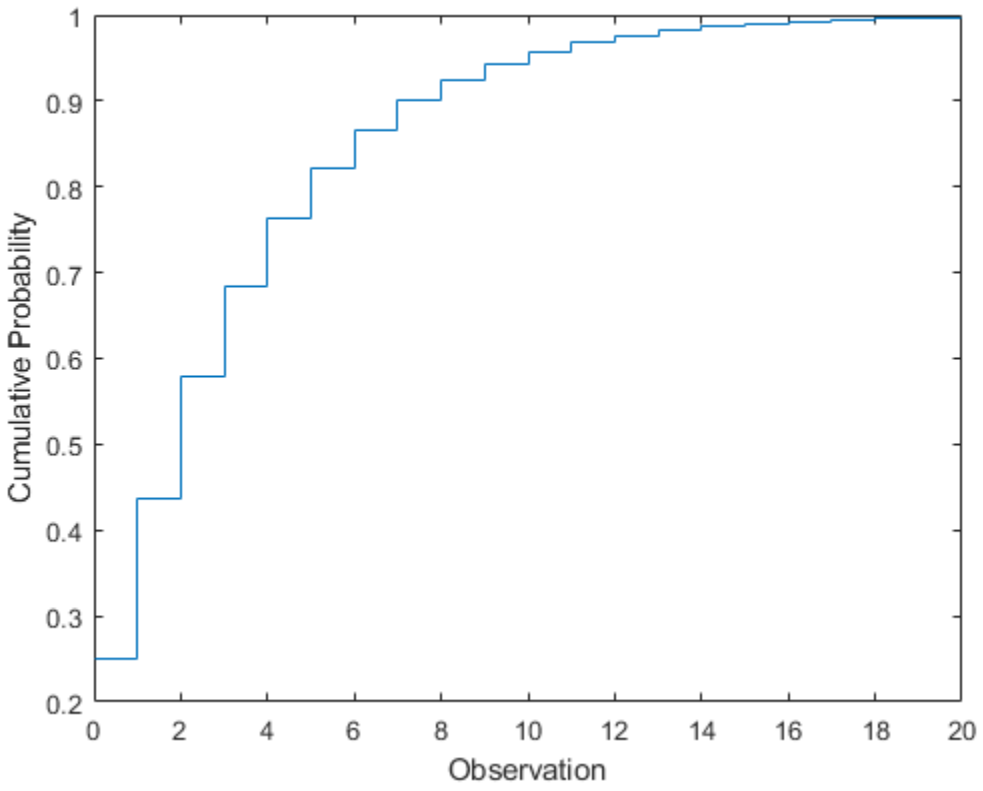
Compute Geometric Distribution cdf

Compute the cdf of the geometric distribution with the probability of success `0.25`.

```
x = 0:20;
y = geocdf(x,0.25);
```

Plot the cdf.

```
figure
stairs(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Compute Geometric Distribution Probabilities

Assume that the probability of a five-year-old car battery not starting in cold weather is 0.03. The driver attempts to start the car every morning during a span of cold weather lasting 25 days. Model this scenario with a geometric distribution, where the event to observe is the car not starting.

Compute the cdf of 25 to find the probability of the car not starting during one of the 25 days.

```
x = 25;
p = 0.03;
notstart = geocdf(x,p)
```

```
notstart = 0.5470
```

Compute the complement to find the probability of the car starting every day for all 25 days.

```
start = 1 - notstart
```

```
start = 0.4530
```

Related Distributions

- “Exponential Distribution” on page B-33 — The exponential distribution is a one-parameter continuous distribution that has parameter μ (mean). The exponential distribution is a continuous analog of the geometric and is the only distribution other than geometric with a constant hazard function.

- Negative Binomial Distribution on page B-109 — The negative binomial distribution is a two-parameter discrete distribution that has parameters r and p, and models the number of failures observed before r successes with probability p of success in a single trial. The geometric distribution occurs as the negative binomial distribution with $r = 1$.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

See Also

`NegativeBinomialDistribution` | `geocdf` | `geoinv` | `geopdf` | `geornd` | `geostat`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Half-Normal Distribution

In this section...

Overview

The half-normal distribution is a special case of the folded normal and truncated normal distributions. Some applications of the half-normal distribution include modeling measurement data and lifetime data.

Parameters

The half-normal distribution uses the following parameters:

Parameter	Description
$-\infty < \mu < \infty$	Location parameter
$\sigma \geq 0$	Scale parameter

The support for the half-normal distribution is $x \geq \mu$.

Use `makedist` with specified parameter values to create a half-normal probability distribution object `HalfNormalDistribution`. Use `fitdist` to fit a half-normal probability distribution object to sample data. Use `mle` to estimate the half-normal distribution parameter values from sample data without creating a probability distribution object. For more information about working with probability distributions, see “Working with Probability Distributions” on page 5-3.

The Statistics and Machine Learning Toolbox implementation of the half-normal distribution assumes a fixed value for the location parameter μ. Therefore, neither `fitdist` nor `mle` estimates the value of the parameter μ when fitting a half-normal distribution to sample data. You can specify a value for the μ parameter by using the name-value pair argument `'mu'`. The default value for the `'mu'` argument is 0 in both `fitdist` and `mle`.

Probability Density Function

The probability density function (pdf) of the half-normal distribution is

$$y = f(x \mid \mu, \sigma) = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};\ x \geq \mu\,,$$

where μ is the location parameter and σ is the scale parameter. If $x \leq \mu$, then the pdf is undefined.

To compute the pdf of the half-normal distribution, create a `HalfNormalDistribution` probability distribution object using `fitdist` or `makedist`, then use the `pdf` method to work with the object.

PDF of Half-Normal Probability Distribution

This example shows how changing the values of the `mu` and `sigma` parameters alters the shape of the pdf.

Create four probability distribution objects with different parameters.

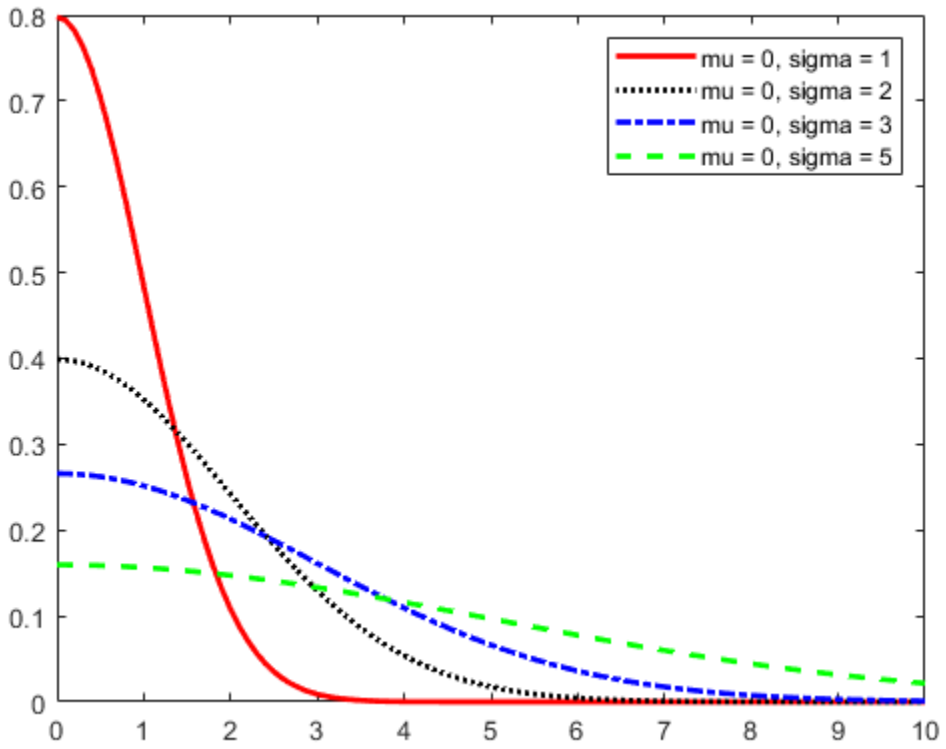
```
pd1 = makedist('HalfNormal');
pd2 = makedist('HalfNormal','mu',0,'sigma',2);
pd3 = makedist('HalfNormal','mu',0,'sigma',3);
pd4 = makedist('HalfNormal','mu',0,'sigma',5);
```

Compute the probability density functions (pdfs) of each distribution.

```
x = 0:0.1:10;
pdf1 = pdf(pd1,x);
pdf2 = pdf(pd2,x);
pdf3 = pdf(pd3,x);
pdf4 = pdf(pd4,x);
```

Plot the pdfs on the same figure.

```
figure;
plot(x,pdf1,'r','LineWidth',2)
hold on;
plot(x,pdf2,'k:','LineWidth',2);
plot(x,pdf3,'b-.','LineWidth',2);
plot(x,pdf4,'g--','LineWidth',2);
legend({'mu = 0, sigma = 1','mu = 0, sigma = 2',...
    'mu = 0, sigma = 3','mu = 0, sigma = 5'},'Location','NE');
hold off;
```



As `sigma` increases, the curve flattens and the peak value becomes smaller.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the half-normal distribution is

$$y = F(x) = erf\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) = 2\Phi\left(\frac{x-\mu}{\sigma}\right) - 1 \; ; \; x \geq \mu \, ,$$

where μ is the location parameter, σ is the scale parameter, $erf(\bullet)$ is the error function, and $\Phi(\bullet)$ is the cdf of the standard normal distribution. If $x \leq \mu$, then the cdf is undefined.

To compute the cdf of the half-normal distribution, create a `HalfNormalDistribution` probability distribution object using `fitdist` or `makedist`, then use the `cdf` method to work with the object.

CDF of Half-Normal Probability Distribution

This example shows how changing the values of the `mu` and `sigma` parameters alters the shape of the cdf.

Create four probability distribution objects with different parameters.

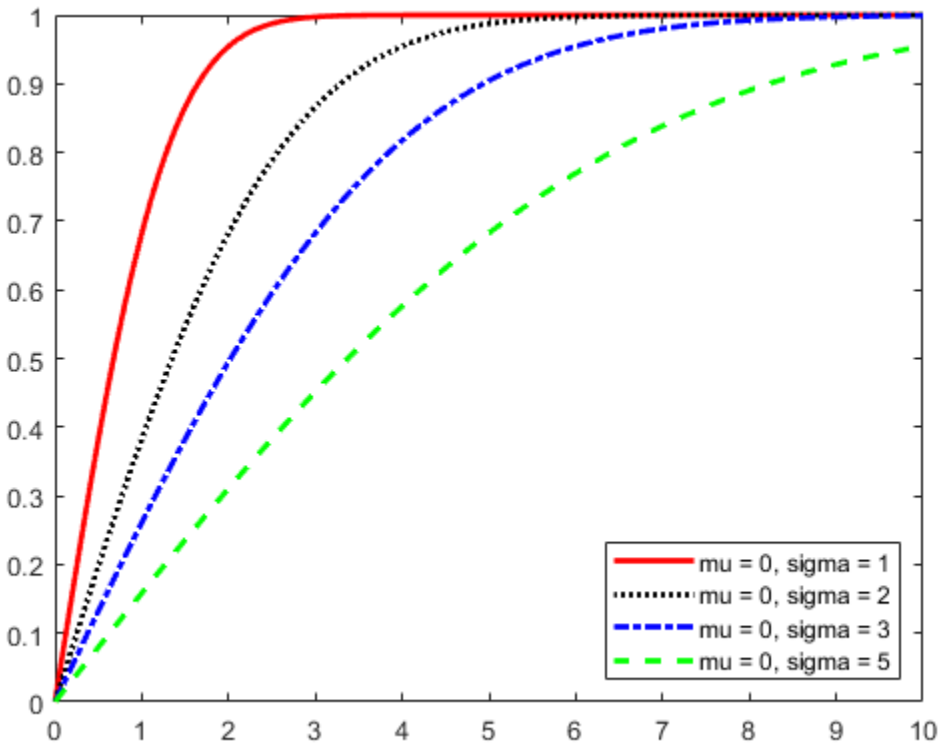
```
pd1 = makedist('HalfNormal');
pd2 = makedist('HalfNormal','mu',0,'sigma',2);
pd3 = makedist('HalfNormal','mu',0,'sigma',3);
pd4 = makedist('HalfNormal','mu',0,'sigma',5);
```

Compute the cumulative distribution functions (cdfs) for each probability distribution.

```
x = 0:0.1:10;
cdf1 = cdf(pd1,x);
cdf2 = cdf(pd2,x);
cdf3 = cdf(pd3,x);
cdf4 = cdf(pd4,x);
```

Plot all four cdfs on the same figure.

```
figure;
plot(x,cdf1,'r','LineWidth',2)
hold on;
plot(x,cdf2,'k:','LineWidth',2);
plot(x,cdf3,'b-.','LineWidth',2);
plot(x,cdf4,'g--','LineWidth',2);
legend({'mu = 0, sigma = 1','mu = 0, sigma = 2',...
    'mu = 0, sigma = 3','mu = 0, sigma = 5'},'Location','SE');
hold off;
```



As `sigma` increases, the curve of the cdf flattens.

Descriptive Statistics

The mean of the half-normal distribution is

$$\text{mean} = \mu + \sigma\sqrt{\frac{2}{\pi}},$$

where μ is the location parameter and σ is the scale parameter.

The variance of the half-normal distribution is

$$\text{var} = \sigma^2\left(1 - \frac{2}{\pi}\right),$$

where σ is the scale parameter.

Relationship to Other Distributions

If a random variable Z has a standard normal distribution with a mean μ equal to zero and standard deviation σ equal to one, then $X = \mu + \sigma|Z|$ has a half-normal distribution with parameters μ and σ.

References

[1] Cooray, K. and M.M.A. Ananda. “A Generalization of the Half-Normal Distribution with Applications to Lifetime Data.” *Communications in Statistics – Theory and Methods*. Vol. 37, Number 9, 2008, pp. 1323–1337.

[2] Pewsey, A. “Large-Sample Inference for the General Half-Normal Distribution.” *Communications in Statistics – Theory and Methods*. Vol. 31, Number 7, 2002, pp. 1045–1054.

See Also

`HalfNormalDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Hypergeometric Distribution

In this section...

Definition

The hypergeometric pdf is

$$y = f(x \mid M, K, n) = \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}$$

Background

The hypergeometric distribution models the total number of successes in a fixed-size sample drawn without replacement from a finite population.

The distribution is discrete, existing only for nonnegative integers less than the number of samples or the number of possible successes, whichever is greater. The hypergeometric distribution differs from the binomial only in that the population is finite and the sampling from the population is without replacement.

The hypergeometric distribution has three parameters that have direct physical interpretations.

- M is the size of the population.
- K is the number of items with the desired characteristic in the population.
- n is the number of samples drawn.

Sampling “without replacement” means that once a particular sample is chosen, it is removed from the relevant population for all subsequent selections.

Examples

Compute and Plot Hypergeometric Distribution CDF

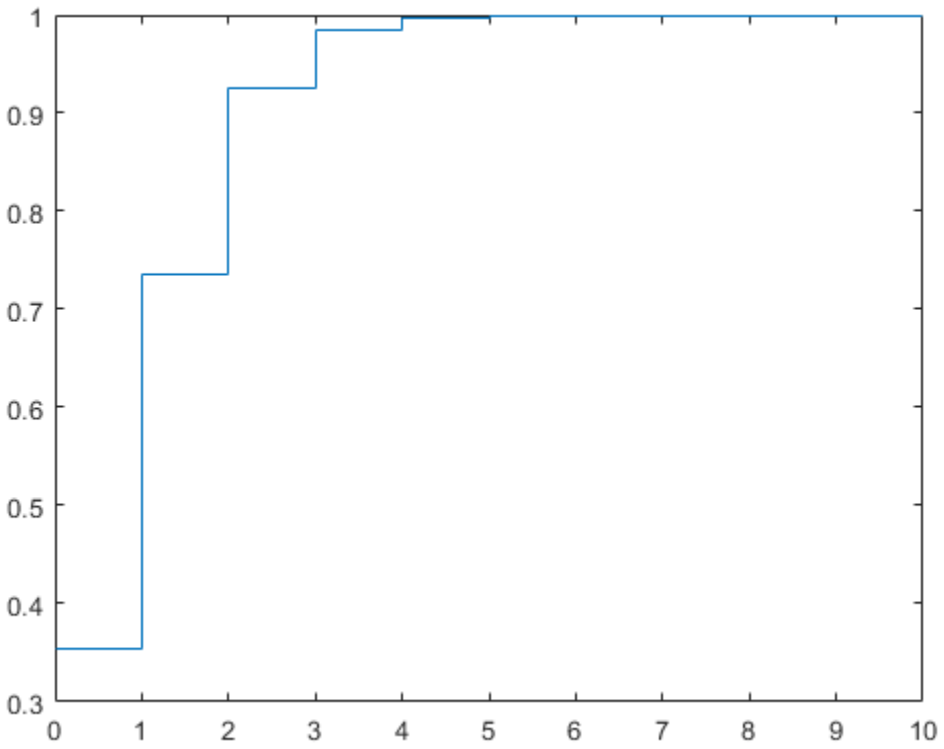
This example shows how to compute and plot the cdf of a hypergeometric distribution.

Compute the cdf of a hypergeometric distribution that draws 20 samples from a group of 1000 items, when the group contains 50 items of the desired type.

```
x = 0:10;
y = hygecdf(x,1000,50,20);
```

Plot the cdf.

```
stairs(x,y)
```



The x-axis of the plot shows the number of items drawn that are of the desired type. The y-axis shows the corresponding cdf values.

See Also

`hygecdf` | `hygeinv` | `hygepdf` | `hygernd` | `hygestat` | `random`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Inverse Gaussian Distribution

In this section...

Definition

The inverse Gaussian distribution has the density function

$$\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\}$$

Background

Also known as the Wald distribution, the inverse Gaussian is used to model nonnegative positively skewed data. The distribution originated in the theory of Brownian motion, but has been used to model diverse phenomena. Inverse Gaussian distributions have many similarities to standard Gaussian (normal) distributions, which lead to applications in inferential statistics.

Parameters

To estimate distribution parameters, use `mle` or the Distribution Fitter app.

See Also

`InverseGaussianDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Inverse Wishart Distribution

Definition

The probability density function of the d-dimensional Inverse Wishart distribution is given by

$$y = f(X, \Sigma, \nu) = \frac{|T|^{(\nu/2)} e^{\left(-\frac{1}{2}\text{trace}\left(TX^{-1}\right)\right)}}{2^{(\nu d)/2} \Pi^{(d(d-1))/4} |X|^{(\nu+d+1)/2} \Gamma(\nu/2) \ldots \Gamma(\nu-(d-1))/2},$$

where X and T are d-by-d symmetric positive definite matrices, and ν is a scalar greater than or equal to d. While it is possible to define the Inverse Wishart for singular T, the density cannot be written as above.

If a random matrix has a Wishart distribution with parameters T^{-1} and ν, then the inverse of that random matrix has an inverse Wishart distribution with parameters T and ν. The mean of the distribution is given by

$$\frac{1}{\nu - d - 1} T$$

where d is the number of rows and columns in T.

Only random matrix generation is supported for the inverse Wishart, including both singular and nonsingular T.

Background

The inverse Wishart distribution is based on the Wishart distribution on page B-178. In Bayesian statistics it is used as the conjugate prior for the covariance matrix of a multivariate normal distribution.

Example

Notice that the sampling variability is quite large when the degrees of freedom is small.

```
Tau = [1 .5; .5 2];
df = 10; S1 = iwishrnd(Tau,df)*(df-2-1)

S1 =
       1.7959      0.64107
      0.64107       1.5496

df = 1000; S2 = iwishrnd(Tau,df)*(df-2-1)

S2 =
       0.9842      0.50158
      0.50158       2.1682
```

See Also

`iwishrnd`

More About

- “Wishart Distribution” on page B-178
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Kernel Distribution

In this section...

Overview

A kernel distribution is a nonparametric representation of the probability density function (pdf) of a random variable. You can use a kernel distribution when a parametric distribution cannot properly describe the data, or when you want to avoid making assumptions about the distribution of the data. A kernel distribution is defined by a smoothing function and a bandwidth value, which control the smoothness of the resulting density curve.

Kernel Density Estimator

The kernel density estimator is the estimated pdf of a random variable. For any real values of x, the kernel density estimator's formula is given by

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

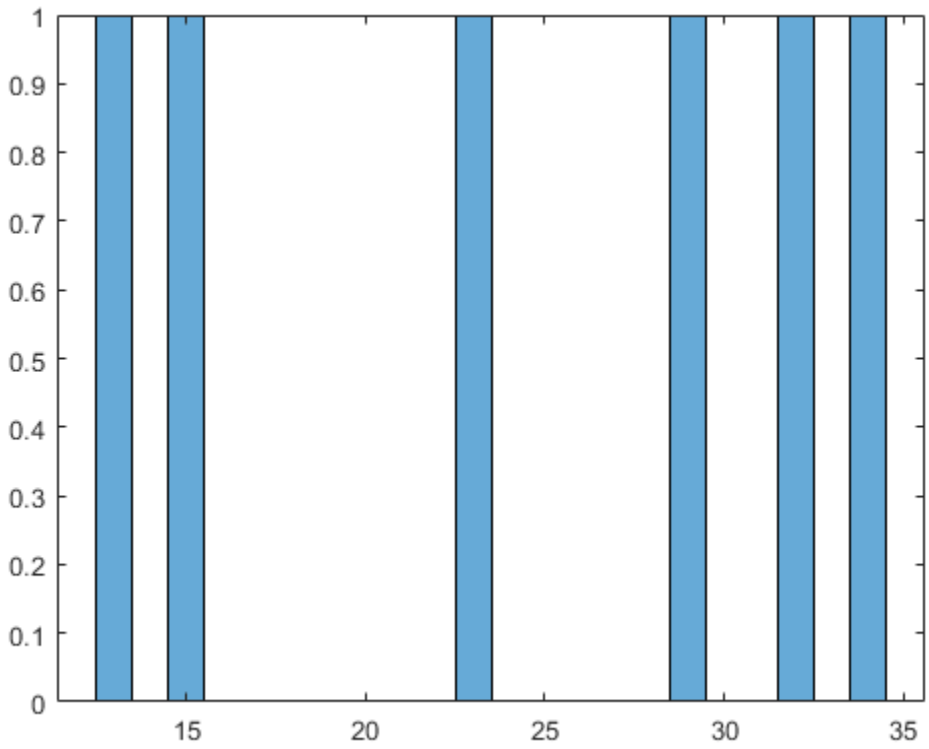
where x_1, x_2, ..., x_n are random samples from an unknown distribution, n is the sample size, $K(\cdot)$ is the kernel smoothing function, and h is the bandwidth.

Kernel Smoothing Function

The kernel smoothing function defines the shape of the curve used to generate the pdf. Similar to a histogram, the kernel distribution builds a function to represent the probability distribution using the sample data. But unlike a histogram, which places the values into discrete bins, a kernel distribution sums the component smoothing functions for each data value to produce a smooth, continuous probability curve. The following plots show a visual comparison of a histogram and a kernel distribution generated from the same sample data.

A histogram represents the probability distribution by establishing bins and placing each data value in the appropriate bin.

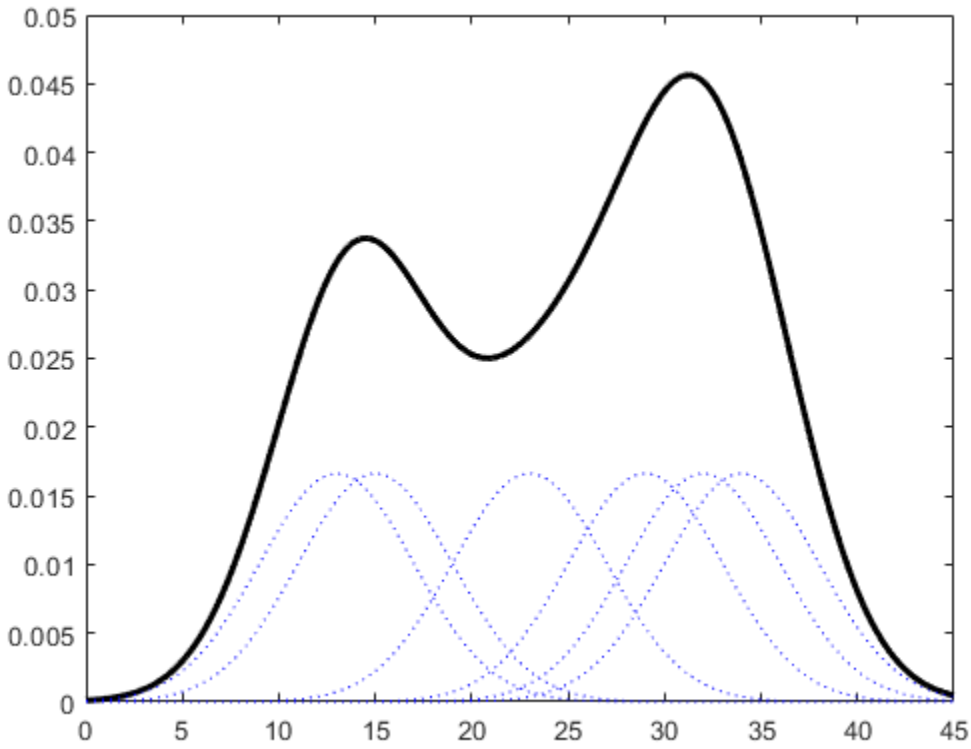
```
SixMPG = [13;15;23;29;32;34];
figure
histogram(SixMPG)
```



Because of this bin count approach, the histogram produces a discrete probability density function. This might be unsuitable for certain applications, such as generating random numbers from a fitted distribution.

Alternatively, the kernel distribution builds the pdf by creating an individual probability density curve for each data value, then summing the smooth curves. This approach creates one smooth, continuous probability density function for the data set.

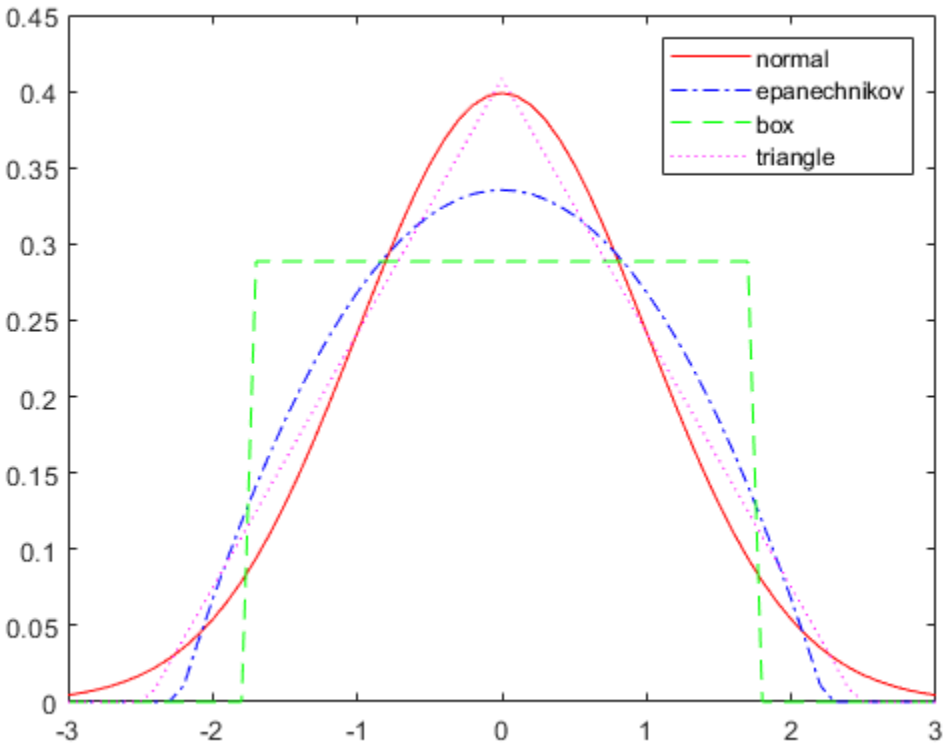
```
figure
pdSix = fitdist(SixMPG,'Kernel','BandWidth',4);
x = 0:.1:45;
ySix = pdf(pdSix,x);
plot(x,ySix,'k-','LineWidth',2)

% Plot each individual pdf and scale its appearance on the plot
hold on
for i=1:6
    pd = makedist('Normal','mu',SixMPG(i),'sigma',4);
    y = pdf(pd,x);
    y = y/6;
    plot(x,y,'b:')
end
hold off
```



The smaller dashed curves are the probability distributions for each value in the sample data, scaled to fit the plot. The larger solid curve is the overall kernel distribution of the `SixMPG` data. The kernel smoothing function refers to the shape of those smaller component curves, which have a normal distribution in this example.

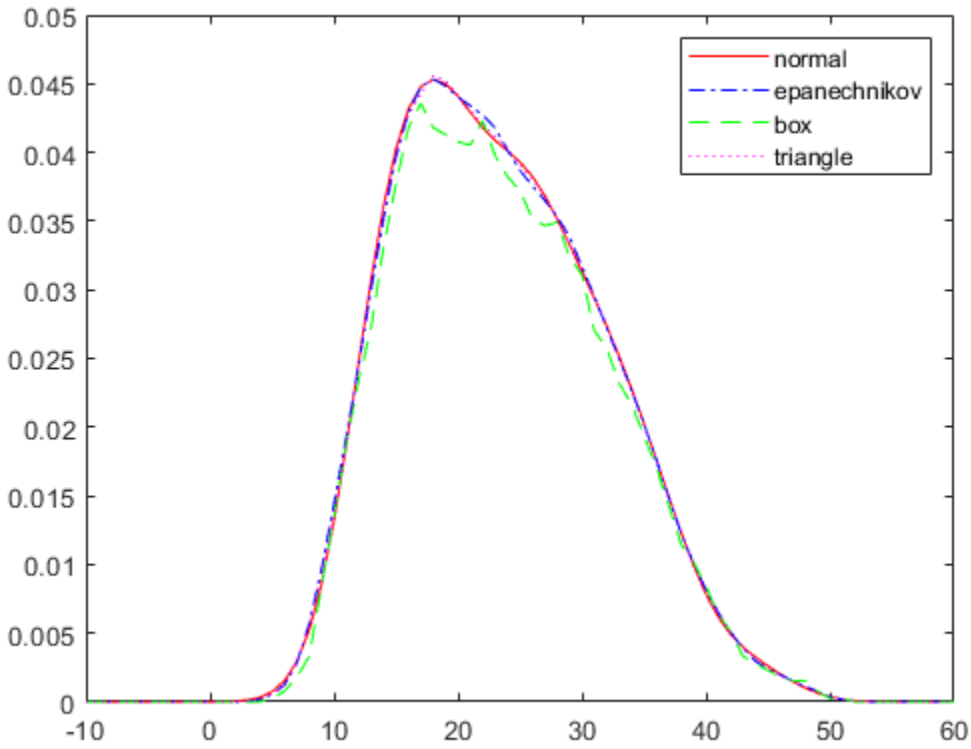
You can choose one of several options for the kernel smoothing function. This plot shows the shapes of the available smoothing functions.

```
% Set plot specifications
hname = {'normal' 'epanechnikov' 'box' 'triangle'};
colors = {'r' 'b' 'g' 'm'};
lines = {'-','-.','--',':'};

% Generate a sample of each kernel smoothing function and plot
data = [0];
figure
for j=1:4
    pd = fitdist(data,'kernel','Kernel',hname{j});
    x = -3:.1:3;
    y = pdf(pd,x);
    plot(x,y,'Color',colors{j},'LineStyle',lines{j})
    hold on
end
legend(hname)
hold off
```



To understand the effect of different kernel smoothing functions on the shape of the resulting pdf estimate, compare plots of the mileage data (`MPG`) from `carbig.mat` using each available kernel function.

```
load carbig
% Set plot specifications
hname = {'normal' 'epanechnikov' 'box' 'triangle'};
colors = {'r' 'b' 'g' 'm'};
lines = {'-','-.','--',':'};

% Generate kernel distribution objects and plot
figure
for j=1:4
    pd = fitdist(MPG,'kernel','Kernel',hname{j});
    x = -10:1:60;
    y = pdf(pd,x);
    plot(x,y,'Color',colors{j},'LineStyle',lines{j})
    hold on
end
legend(hname)
hold off
```



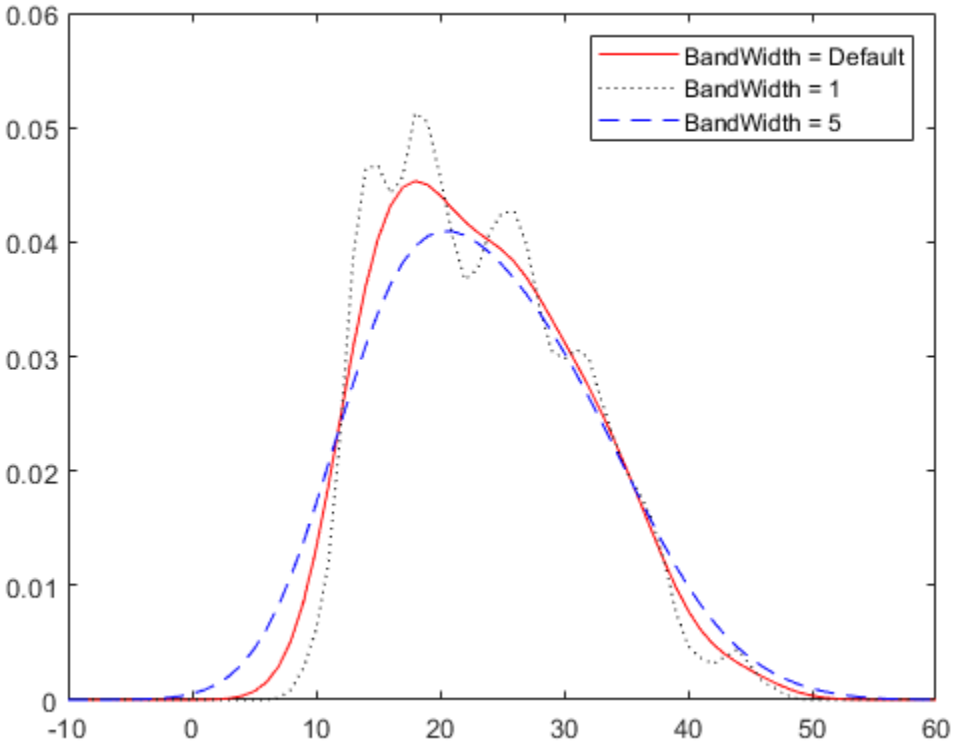
Each density curve uses the same input data, but applies a different kernel smoothing function to generate the pdf. The density estimates are roughly comparable, but the shape of each curve varies slightly. For example, the box kernel produces a density curve that is less smooth than the others.

Bandwidth

The choice of bandwidth value controls the smoothness of the resulting probability density curve. This plot shows the density estimate for the `MPG` data, using a normal kernel smoothing function with three different bandwidths.

```
% Create kernel distribution objects
load carbig
pd1 = fitdist(MPG,'kernel');
pd2 = fitdist(MPG,'kernel','BandWidth',1);
pd3 = fitdist(MPG,'kernel','BandWidth',5);

% Compute each pdf
x = -10:1:60;
y1 = pdf(pd1,x);
y2 = pdf(pd2,x);
y3 = pdf(pd3,x);

% Plot each pdf
plot(x,y1,'Color','r','LineStyle','-')
hold on
```

```
plot(x,y2,'Color','k','LineStyle',':')
plot(x,y3,'Color','b','LineStyle','--')
legend({'BandWidth = Default','BandWidth = 1','BandWidth = 5'})
hold off
```



The default bandwidth, which is theoretically optimal for estimating densities for the normal distribution [1], produces a reasonably smooth curve. Specifying a smaller bandwidth produces a very rough curve, but reveals that there might be two major peaks in the data. Specifying a larger bandwidth produces a curve nearly identical to the kernel function, and is so smooth that it obscures potentially important features of the data.

References

[1] Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press Inc., 1997.

See Also

`KernelDistribution` | `ksdensity`

More About

- “Fit Kernel Distribution Object to Data” on page 5-36
- “Fit Kernel Distribution Using ksdensity” on page 5-39

- “Fit Distributions to Grouped Data Using ksdensity” on page 5-41
- “Nonparametric and Empirical Probability Distributions” on page 5-30
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Logistic Distribution

In this section...

Overview

The logistic distribution is used for growth models and in logistic regression. It has longer tails and a higher kurtosis than the normal distribution.

Parameters

The logistic distribution uses the following parameters.

Parameter	Description	Support
`mu`	Mean	$-\infty < \mu < \infty$
`sigma`	Scale parameter	$\sigma \geq 0$

Probability Density Function

The probability density function (pdf) is

$$f(x\left|\mu,\sigma\right.) = \frac{\exp\left\{\frac{x-\mu}{\sigma}\right\}}{\sigma\left(1+\exp\left\{\frac{x-\mu}{\sigma}\right\}\right)^2} \quad ; \quad -\infty < x < \infty \ .$$

Relationship to Other Distributions

The loglogistic distribution is closely related to the logistic distribution. If x is distributed loglogistically with parameters μ and σ, then $\log(x)$ is distributed logistically with parameters μ and σ.

See Also

`LogisticDistribution`

More About

- “Compare Multiple Distribution Fits” on page 5-87
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Loglogistic Distribution

In this section...

Overview

The loglogistic distribution is a probability distribution whose logarithm has a logistic distribution. This distribution is often used in survival analysis to model events that experience an initial rate increase, followed by a rate decrease. It is also known as the Fisk distribution in economics applications.

Parameters

The loglogistic distribution uses the following parameters.

Parameter	Description	Support
`mu`	Mean of logarithmic values	$\mu > 0$
`sigma`	Scale parameter of logarithmic values	$\sigma > 0$

Probability Density Function

The probability density function (pdf) is

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma}\frac{1}{x}\frac{e^{z}}{\left(1+e^{z}\right)^{2}} \quad ; \quad x \geq 0\,,$$

where $z = \dfrac{\log(x) - \mu}{\sigma}$.

Relationship to Other Distributions

The loglogistic distribution is closely related to the logistic distribution. If x is distributed loglogistically with parameters μ and σ, then $\log(x)$ is distributed logistically with parameters μ and σ. The relationship is similar to that between the lognormal on page B-88 and normal on page B-119 distribution.

See Also

`LoglogisticDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Lognormal Distribution

In this section...

Overview

The lognormal distribution, sometimes called the Galton distribution, is a probability distribution whose logarithm has a normal distribution. The lognormal distribution is applicable when the quantity of interest must be positive, because log(x) exists only when x is positive.

Statistics and Machine Learning Toolbox offers several ways to work with the lognormal distribution.

- Create a probability distribution object `LognormalDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the lognormal distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`logncdf`, `lognpdf`, `logninv`, `lognlike`, `lognstat`, `lognfit`, `lognrnd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple lognormal distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Lognormal'`) and parameters.

Parameters

The lognormal distribution uses these parameters.

Parameter	Description	Support
`mu` (μ)	Mean of logarithmic values	$-\infty < \mu < \infty$
`sigma` (σ)	Standard deviation of logarithmic values	$\sigma \geq 0$

If X follows the lognormal distribution with parameters μ and σ, then log(X) follows the normal distribution with mean μ and standard deviation σ.

Parameter Estimation

To fit the lognormal distribution to data and find the parameter estimates, use `lognfit`, `fitdist`, or `mle`.

- For uncensored data, `lognfit` and `fitdist` find the unbiased estimates of the distribution parameters, and `mle` finds the maximum likelihood estimates.

- For censored data, `lognfit`, `fitdist`, and `mle` find the maximum likelihood estimates.

Unlike `lognfit` and `mle`, which return parameter estimates, `fitdist` returns the fitted probability distribution object `LognormalDistribution`. The object properties `mu` and `sigma` store the parameter estimates.

Descriptive Statistics

The mean m and variance v of a lognormal random variable are functions of the lognormal distribution parameters μ and σ:

$$m = \exp\left(\mu + \sigma^2/2\right)$$
$$v = \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right)$$

Also, you can compute the lognormal distribution parameters μ and σ from the mean m and variance v:

$$\mu = \log\left(m^2/\sqrt{v + m^2}\right)$$
$$\sigma = \sqrt{\log\left(v/m^2 + 1\right)}$$

Probability Density Function

The probability density function (pdf) of the lognormal distribution is

$$y = f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{\frac{-(\log x - \mu)^2}{2\sigma^2}\right\}, \quad \text{for } x > 0.$$

For an example, see “Compute Lognormal Distribution pdf” on page B-89.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the lognormal distribution is

$$p = F(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^x \frac{1}{t}\exp\left\{\frac{-(\log t - \mu)^2}{2\sigma^2}\right\}dt, \quad \text{for } x > 0.$$

For an example, see “Compute Lognormal Distribution cdf” on page B-90.

Examples

Compute Lognormal Distribution pdf

Suppose the income of a family of four in the United States follows a lognormal distribution with `mu = log(20,000)` and `sigma = 1`. Compute and plot the income density.

Create a lognormal distribution object by specifying the parameter values.

```
pd = makedist('Lognormal','mu',log(20000),'sigma',1)
```

```
pd =
  LognormalDistribution
```

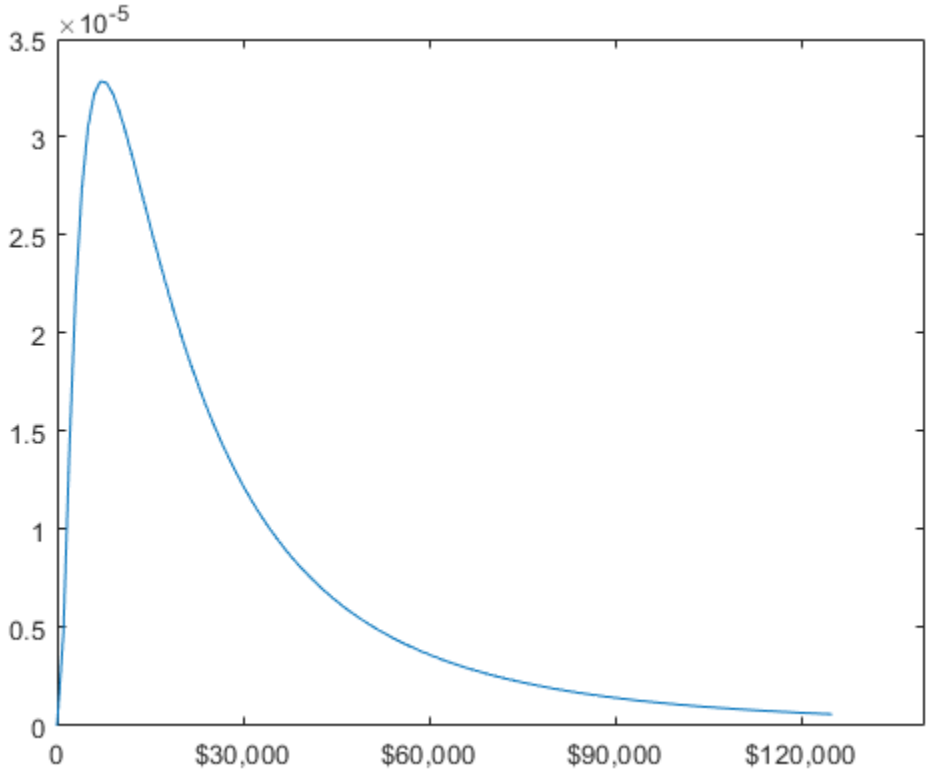
```
  Lognormal distribution
       mu = 9.90349
    sigma =       1
```

Compute the pdf values.

```
x = (10:1000:125010)';
y = pdf(pd,x);
```

Plot the pdf.

```
plot(x,y)
h = gca;
h.XTick = [0 30000 60000 90000 120000];
h.XTickLabel = {'0','$30,000','$60,000',...
                    '$90,000','$120,000'};
```



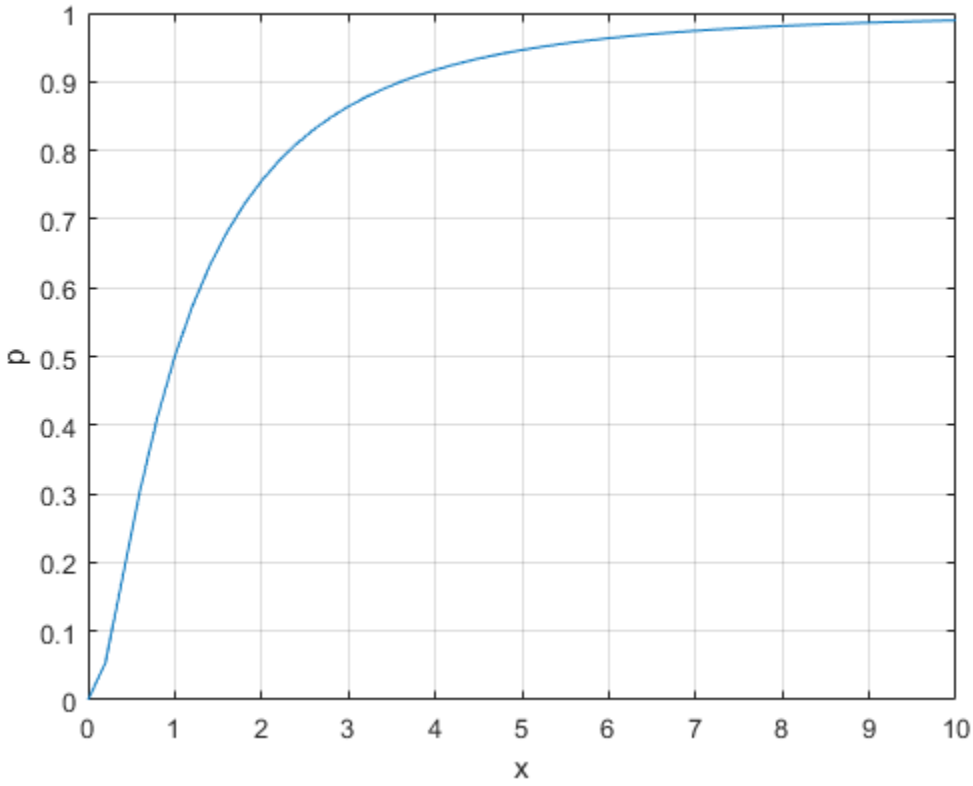
Compute Lognormal Distribution cdf

Compute the cdf values evaluated at the values in `x` for the lognormal distribution with mean `mu` and standard deviation `sigma`.

```
x = 0:0.2:10;
mu = 0;
sigma = 1;
p = logncdf(x,mu,sigma);
```

Plot the cdf.

```
plot(x,p)
grid on
xlabel('x')
ylabel('p')
```



Relationship Between Normal and Lognormal Distributions

If X follows the lognormal distribution with parameters μ and σ, then $\log(X)$ follows the normal distribution with mean μ and standard deviation σ. Use distribution objects to inspect the relationship between normal and lognormal distributions.

Create a lognormal distribution object by specifying the parameter values.

```
pd = makedist('Lognormal','mu',5,'sigma',2)
```

```
pd =
  LognormalDistribution

  Lognormal distribution
       mu = 5
    sigma = 2
```

Compute the mean of the lognormal distribution.

```
mean(pd)
```

```
ans = 1.0966e+03
```

The mean of the lognormal distribution is not equal to the `mu` parameter. The mean of the logarithmic values is equal to `mu`. Confirm this relationship by generating random numbers.

Generate random numbers from the lognormal distribution and compute their log values.

```
rng('default');  % For reproducibility
x = random(pd,10000,1);
logx = log(x);
```

Compute the mean of the logarithmic values.

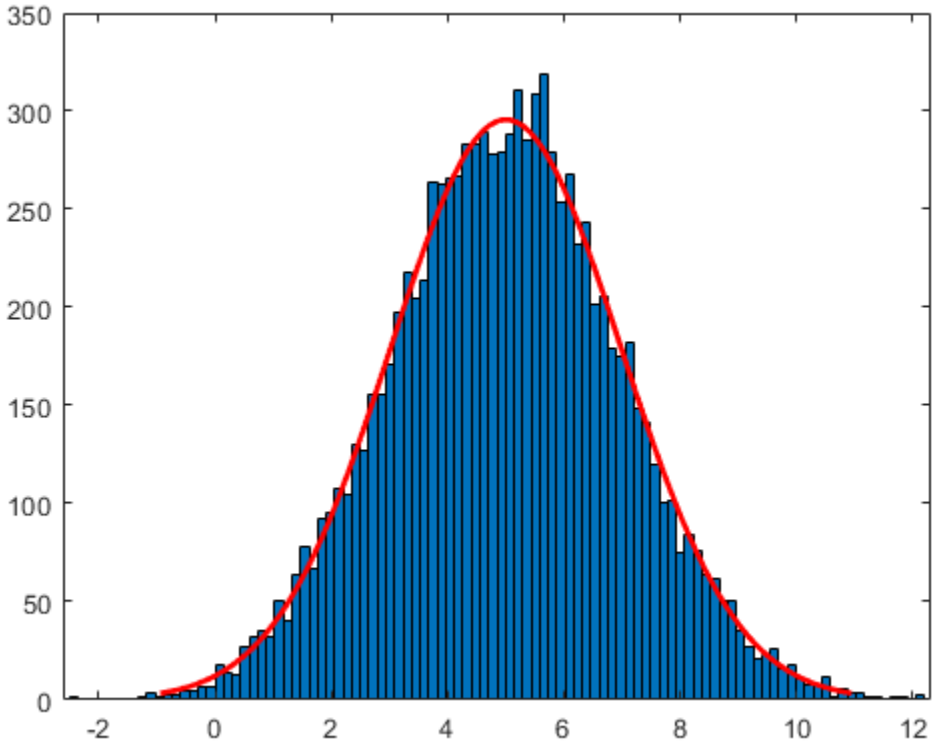
```
m = mean(logx)
```

```
m = 5.0033
```

The mean of the log of `x` is close to the `mu` parameter of `x`, because `x` has a lognormal distribution.

Construct a histogram of `logx` with a normal distribution fit.

```
histfit(logx)
```



The plot shows that the log values of `x` are normally distributed.

`histfit` uses `fitdist` to fit a distribution to data. Use `fitdist` to obtain parameters used in fitting.

```
pd_normal = fitdist(logx,'Normal')

pd_normal =
  NormalDistribution

  Normal distribution
       mu = 5.00332   [4.96445, 5.04219]
    sigma = 1.98296   [1.95585, 2.01083]
```

The estimated normal distribution parameters are close to the lognormal distribution parameters 5 and 2.

Compare Lognormal and Burr Distribution pdfs

Compare the lognormal pdf to the Burr pdf using income data generated from a lognormal distribution.

Generate the income data.

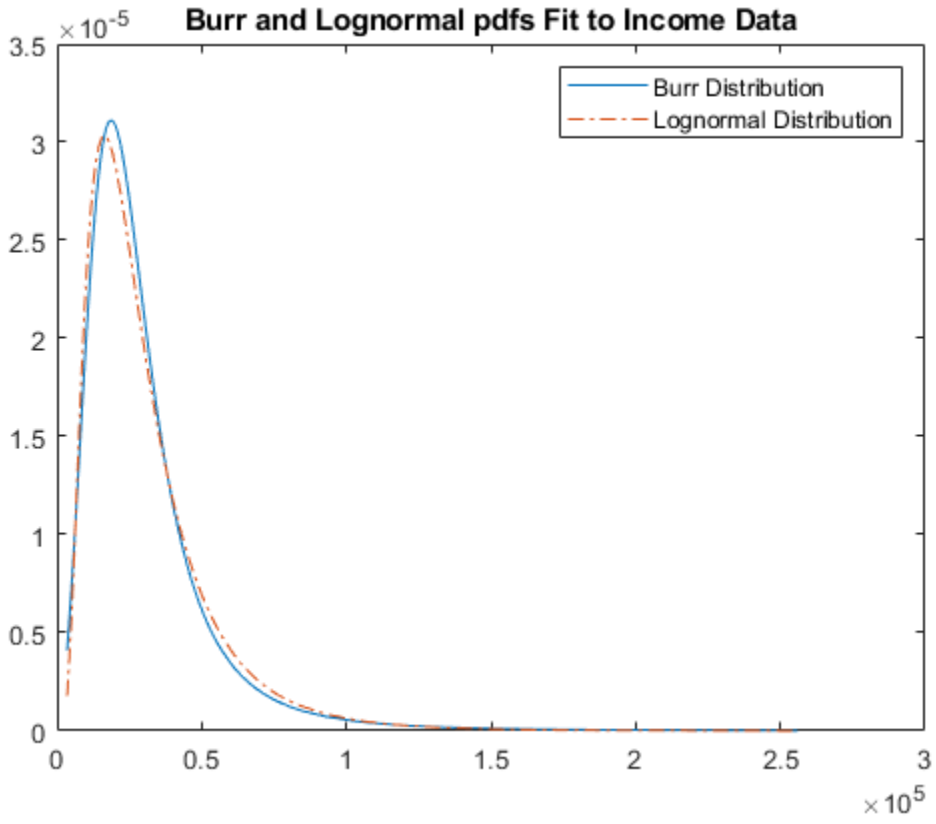
```
rng('default') % For reproducibility
y = random('Lognormal',log(25000),0.65,[500,1]);
```

Fit a Burr distribution.

```
pd = fitdist(y,'burr')

pd =
  BurrDistribution

  Burr distribution
    alpha = 26007.2   [21165.5, 31956.4]
        c = 2.63743   [2.3053, 3.0174]
        k = 1.09658   [0.775479, 1.55064]
```

Plot both the Burr and lognormal pdfs of income data on the same figure.

```
p_burr = pdf(pd,sortrows(y));
p_lognormal = pdf('Lognormal',sortrows(y),log(25000),0.65);
plot(sortrows(y),p_burr,'-',sortrows(y),p_lognormal,'-.')
title('Burr and Lognormal pdfs Fit to Income Data')
legend('Burr Distribution','Lognormal Distribution')
```



Related Distributions

- "Normal Distribution" on page B-119 — The lognormal distribution is closely related to the normal distribution. If X is distributed lognormally with parameters μ and σ, then $\log(x)$ is distributed normally with mean μ and standard deviation σ. See "Relationship Between Normal and Lognormal Distributions" on page B-91.
- "Burr Type XII Distribution" on page B-19 — The Burr distribution is a flexible distribution family that can express a wide range of distribution shapes. It has as a limiting case many commonly used distributions such as gamma, lognormal, loglogistic, bell-shaped, and J-shaped beta distributions (but not U-shaped). See "Compare Lognormal and Burr Distribution pdfs" on page B-93.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[3] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[4] Marsaglia, G., and W. W. Tsang. “A Fast, Easily Implemented Method for Sampling from Decreasing or Symmetric Unimodal Density Functions.” *SIAM Journal on Scientific and Statistical Computing*. Vol. 5, Number 2, 1984, pp. 349–359.

[5] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

[6] Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* 3rd ed., New York: McGraw-Hill, 1974. pp. 540–541.

See Also

`LognormalDistribution` | `logncdf` | `lognfit` | `logninv` | `lognlike` | `lognpdf` | `lognrnd` | `lognstat`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Multinomial Distribution

In this section...

Overview

Multinomial distribution models the probability of each combination of successes in a series of independent trials. Use this distribution when there are more than two possible mutually exclusive outcomes for each trial, and each outcome has a fixed probability of success.

Parameter

Multinomial distribution uses the following parameter.

Parameter	Description	Constraints
`probabilities`	Outcome probabilities	$0 \le \text{probabilities}(i) \le 1$; $\sum_{\text{all}(i)} \text{probabilities}(i) = 1$

Probability Density Function

The multinomial pdf is

$$f(x|n,p) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k},$$

where k is the number of possible mutually exclusive outcomes for each trial, and n is the total number of trials. The vector $x = (x_1 \ldots x_k)$ is the number of observations of each k outcome, and contains nonnegative integer components that sum to n. The vector $p = (p_1 \ldots p_k)$ is the fixed probability of each k outcome, and contains nonnegative scalar components that sum to 1.

Descriptive Statistics

The expected number of observations of outcome i in n trials is

$$\mathrm{E}\{x_i\} = np_i\,,$$

where p_i is the fixed probability of outcome i.

The variance is of outcome i is

$$\mathrm{var}(x_i) = np_i(1 - p_i)\ .$$

The covariance of outcomes i and j is

$$\mathrm{cov}(x_i, x_j) = -np_ip_j,\ i \neq j.$$

Relationship to Other Distributions

The multinomial distribution is a generalization of the binomial distribution on page B-10. While the binomial distribution gives the probability of the number of "successes" in n independent trials of a two-outcome process, the multinomial distribution gives the probability of each combination of outcomes in n independent trials of a k-outcome process. The probability of each outcome in any one trial is given by the fixed probabilities $p_1,..., p_k$.

See Also

`MultinomialDistribution`

More About

- "Multinomial Probability Distribution Objects" on page 5-95
- "Multinomial Probability Distribution Functions" on page 5-98
- "Working with Probability Distributions" on page 5-3
- "Supported Distributions" on page 5-14

Multivariate Normal Distribution

In this section...

Overview

The multivariate normal distribution is a generalization of the univariate normal distribution on page B-119 to two or more variables. It is a distribution for random vectors of correlated variables, where each vector element has a univariate normal distribution. In the simplest case, no correlation exists among variables, and elements of the vectors are independent univariate normal random variables.

Because it is easy to work with, the multivariate normal distribution is often used as a model for multivariate data.

Statistics and Machine Learning Toolbox provides several functionalities related to the multivariate normal distribution.

- Generate random numbers from the distribution using `mvnrnd`.
- Evaluate the probability density function (pdf) at specific values using `mvnpdf`.
- Evaluate the cumulative distribution function (cdf) at specific values using `mvncdf`.

Parameters

The multivariate normal distribution uses the parameters in this table.

Parameter	Description	Univariate Normal Analogue
μ	Mean vector	Mean μ (scalar)
Σ	Covariance matrix — Diagonal elements contain the variances for each variable, and off-diagonal elements contain the covariances between variables	Variance σ^2 (scalar)

Note that in the one-dimensional case, Σ is the variance, not the standard deviation. For more information on the parameters of the univariate normal distribution, see “Parameters” on page B-119.

Probability Density Function

The probability density function (pdf) of the d-dimensional multivariate normal distribution is

$$y = f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^d}} \exp\left(-\frac{1}{2}(x\text{-}\mu)\ \Sigma^{-1}(x\text{-}\mu)'\right)$$

where x and μ are 1-by-d vectors and Σ is a d-by-d symmetric, positive definite matrix.

Note that Statistics and Machine Learning Toolbox:

- Supports singular Σ for random vector generation only. The pdf cannot be written in the same form when Σ is singular.
- Uses x and μ oriented as row vectors rather than column vectors.

For an example, see “Bivariate Normal Distribution pdf” on page B-99.

Cumulative Distribution Function

The multivariate normal cumulative distribution function (cdf) evaluated at x is defined as the probability that a random vector v, distributed as multivariate normal, lies within the semi-infinite rectangle with upper limits defined by x,

$$\Pr\{v(1) \le x(1), v(2) \le x(2), ..., v(d) \le x(d)\} .$$

Although the multivariate normal cdf has no closed form, `mvncdf` can compute cdf values numerically.

For an example, see “Bivariate Normal Distribution cdf” on page B-100.

Examples

Bivariate Normal Distribution pdf

Compute and plot the pdf of a bivariate normal distribution with parameters `mu = [0 0]` and `Sigma = [0.25 0.3; 0.3 1]`.

Define the parameters `mu` and `Sigma`.

```
mu = [0 0];
Sigma = [0.25 0.3; 0.3 1];
```

Create a grid of evenly spaced points in two-dimensional space.

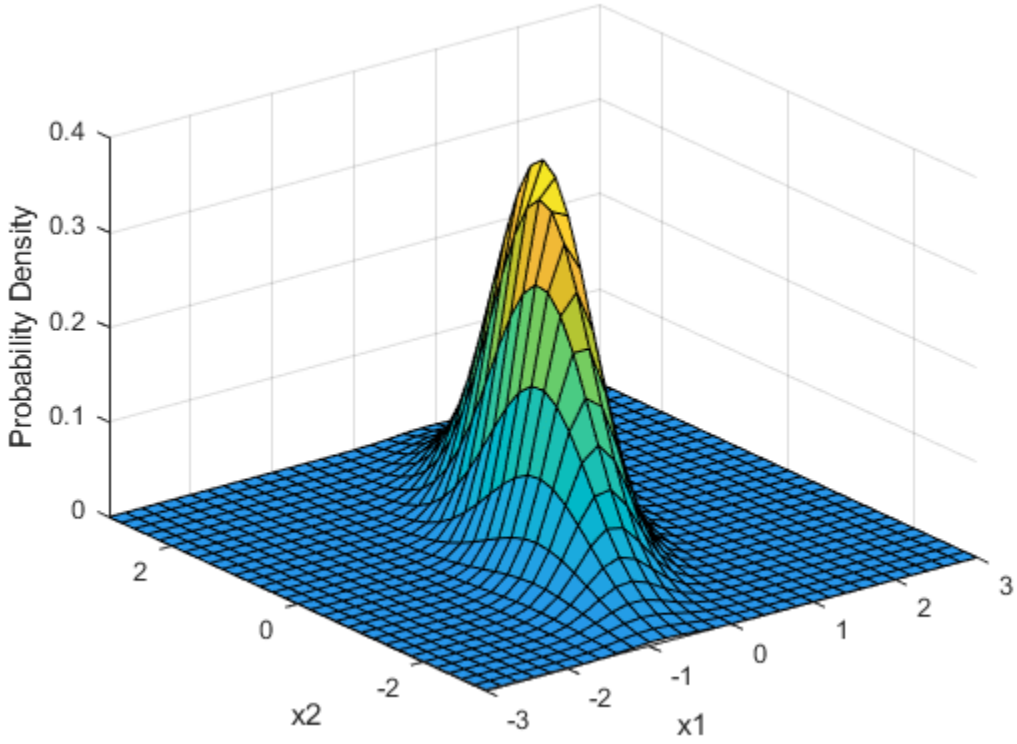
```
x1 = -3:0.2:3;
x2 = -3:0.2:3;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
```

Evaluate the pdf of the normal distribution at the grid points.

```
y = mvnpdf(X,mu,Sigma);
y = reshape(y,length(x2),length(x1));
```

Plot the pdf values.

```
surf(x1,x2,y)
caxis([min(y(:))-0.5*range(y(:)),max(y(:))])
axis([-3 3 -3 3 0 0.4])
xlabel('x1')
ylabel('x2')
zlabel('Probability Density')
```



Bivariate Normal Distribution cdf

Compute and plot the cdf of a bivariate normal distribution.

Define the mean vector `mu` and the covariance matrix `Sigma`.

```
mu = [1 -1];
Sigma = [.9 .4; .4 .3];
```

Create a grid of 625 evenly spaced points in two-dimensional space.

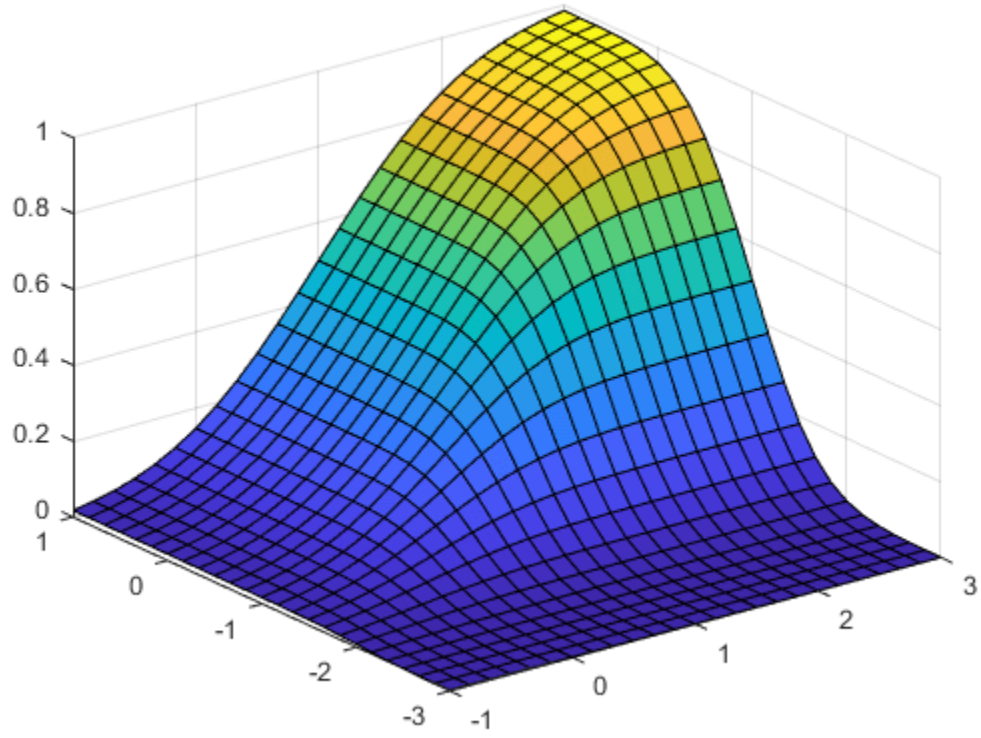
```
[X1,X2] = meshgrid(linspace(-1,3,25)',linspace(-3,1,25)');
X = [X1(:) X2(:)];
```

Evaluate the cdf of the normal distribution at the grid points.

```
p = mvncdf(X,mu,Sigma);
```

Plot the cdf values.

```
Z = reshape(p,25,25);
surf(X1,X2,Z)
```



Probability over Rectangular Region

Compute the probability over the unit square of a bivariate normal distribution, and create a contour plot of the results.

Define the bivariate normal distribution parameters `mu` and `Sigma`.

```
mu = [0 0];
Sigma = [0.25 0.3; 0.3 1];
```

Compute the probability over the unit square.

```
p = mvncdf([0 0],[1 1],mu,Sigma)
```

```
p = 0.2097
```

To visualize the result, first create a grid of evenly spaced points in two-dimensional space.

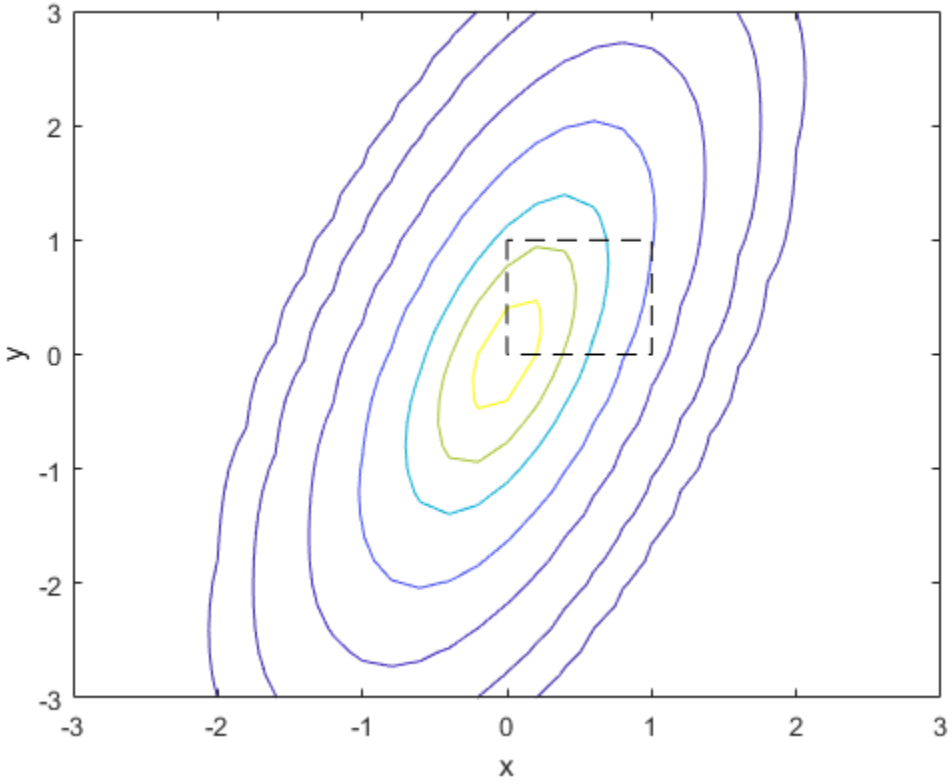
```
x1 = -3:.2:3;
x2 = -3:.2:3;
[X1,X2] = meshgrid(x1,x2);
X = [X1(:) X2(:)];
```

Then, evaluate the pdf of the normal distribution at the grid points.

```
y = mvnpdf(X,mu,Sigma);
y = reshape(y,length(x2),length(x1));
```

Finally, create a contour plot of the multivariate normal distribution that includes the unit square.

```
contour(x1,x2,y,[0.0001 0.001 0.01 0.05 0.15 0.25 0.35])
xlabel('x')
ylabel('y')
line([0 0 1 1 0],[1 0 0 1 1],'Linestyle','--','Color','k')
```



Computing a multivariate cumulative probability requires significantly more work than computing a univariate probability. By default, the `mvncdf` function computes values to less than full machine precision, and returns an estimate of the error as an optional second output. View the error estimate in this case.

```
[p,err] = mvncdf([0 0],[1 1],mu,Sigma)
```

```
p = 0.2097
```

```
err = 1.0000e-08
```

References

[1] Kotz, S., N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions: Volume 1: Models and Applications.* 2nd ed. New York: John Wiley & Sons, Inc., 2000.

See Also

`NormalDistribution` | `mvncdf` | `mvnpdf` | `mvnrnd`

More About

- “Normal Distribution” on page B-119
- “Supported Distributions” on page 5-14

Multivariate t Distribution

In this section...

Definition

The probability density function of the d-dimensional multivariate Student's t distribution is given by

$$f(x,\Sigma,\nu) = \frac{1}{|\Sigma|^{1/2}}\frac{1}{\sqrt{(\nu\pi)^d}}\frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)}\left(1+\frac{x'\,\Sigma^{-1}x}{\nu}\right)^{-(\nu+d)/2}.$$

where x is a 1-by-d vector, Σ is a d-by-d symmetric, positive definite matrix, and ν is a positive scalar. While it is possible to define the multivariate Student's t for singular Σ, the density cannot be written as above. For the singular case, only random number generation is supported. Note that while most textbooks define the multivariate Student's t with x oriented as a column vector, for the purposes of data analysis software, it is more convenient to orient x as a row vector, and Statistics and Machine Learning Toolbox software uses that orientation.

Background

The multivariate Student's t distribution is a generalization of the univariate Student's t to two or more variables. It is a distribution for random vectors of correlated variables, each element of which has a univariate Student's t distribution. In the same way as the univariate Student's t distribution can be constructed by dividing a standard univariate normal random variable by the square root of a univariate chi-square random variable, the multivariate Student's t distribution can be constructed by dividing a multivariate normal random vector having zero mean and unit variances by a univariate chi-square random variable.

The multivariate Student's t distribution is parameterized with a correlation matrix, Σ, and a positive scalar degrees of freedom parameter, ν. ν is analogous to the degrees of freedom parameter of a univariate Student's t distribution. The off-diagonal elements of Σ contain the correlations between variables. Note that when Σ is the identity matrix, variables are uncorrelated; however, they are not independent.

The multivariate Student's t distribution is often used as a substitute for the multivariate normal distribution in situations where it is known that the marginal distributions of the individual variables have fatter tails than the normal.

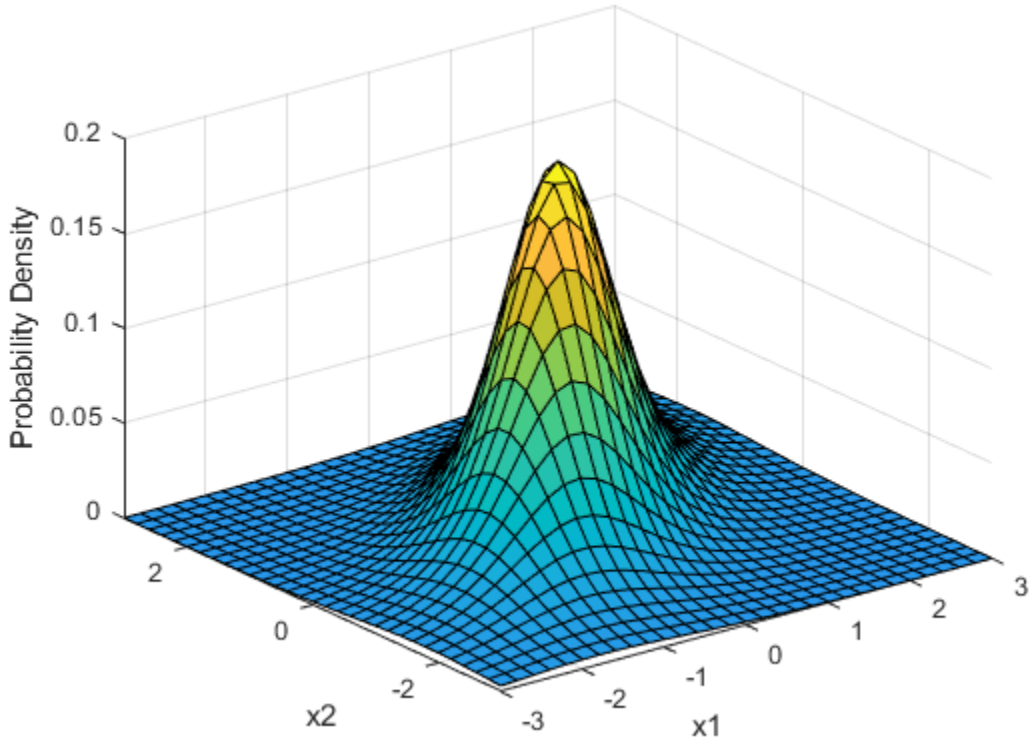
Example

Plot PDF and CDF of Multivariate t-Distribution

Plot the pdf of a bivariate Student's t distribution. You can use this distribution for a higher number of dimensions as well, although visualization is not easy.

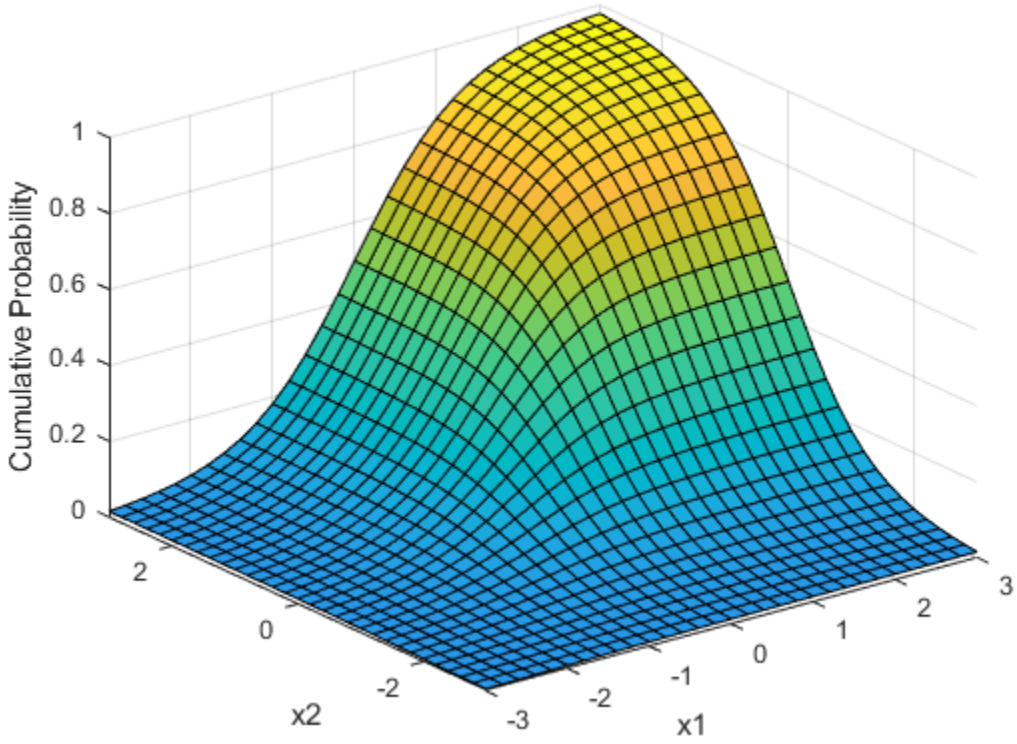
```
Rho = [1 .6; .6 1];
nu = 5;
```

```
x1 = -3:.2:3; x2 = -3:.2:3;
[X1,X2] = meshgrid(x1,x2);
F = mvtpdf([X1(:) X2(:)],Rho,nu);
F = reshape(F,length(x2),length(x1));
surf(x1,x2,F);
caxis([min(F(:))-.5*range(F(:)),max(F(:))]);
axis([-3 3 -3 3 0 .2])
xlabel('x1'); ylabel('x2'); zlabel('Probability Density');
```



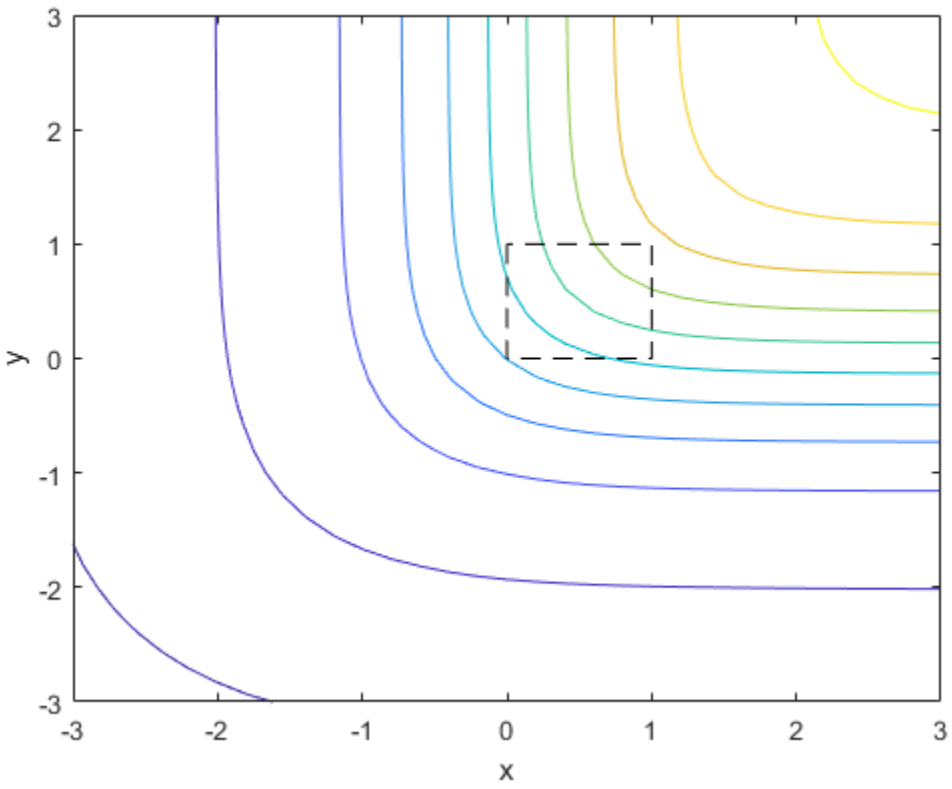
Plot the cdf of a bivariate Student's *t* distribution.

```
F = mvtcdf([X1(:) X2(:)],Rho,nu);
F = reshape(F,length(x2),length(x1));
surf(x1,x2,F);
caxis([min(F(:))-.5*range(F(:)),max(F(:))]);
axis([-3 3 -3 3 0 1])
xlabel('x1'); ylabel('x2'); zlabel('Cumulative Probability');
```



Since the bivariate Student's *t* distribution is defined on the plane, you can also compute cumulative probabilities over rectangular regions. For example, this contour plot illustrates the computation that follows, of the probability contained within the unit square shown in the figure.

```
contour(x1,x2,F,[.0001 .001 .01 .05:.1:.95 .99 .999 .9999]);
xlabel('x'); ylabel('y');
line([0 0 1 1 0],[1 0 0 1 1],'linestyle','--','color','k');
```



Compute the value of the probability contained within the unit square.

```
F = mvtcdf([0 0],[1 1],Rho,nu)
```

```
F = 0.1401
```

Computing a multivariate cumulative probability requires significantly more work than computing a univariate probability. By default, the `mvtcdf` function computes values to less than full machine precision and returns an estimate of the error, as an optional second output.

```
[F,err] = mvtcdf([0 0],[1 1],Rho,nu)
```

```
F = 0.1401
```

```
err = 1.0000e-08
```

See Also

`mvtcdf` | `mvtpdf` | `mvtrnd`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Nakagami Distribution

In this section...

Definition

The Nakagami distribution has the density function

$$2\left(\frac{\mu}{\omega}\right)^{\mu}\frac{1}{\Gamma(\mu)}x^{(2\mu-1)}e^{\frac{-\mu}{\omega}x^2}$$

with shape parameter μ and scale parameter $\omega > 0$, for $x > 0$. If x has a Nakagami distribution with parameters μ and ω, then x^2 has a gamma distribution with shape parameter μ and scale parameter ω/μ.

Background

In communications theory, Nakagami distributions, Rician distributions on page B-139, and Rayleigh distributions on page B-137 are used to model scattered signals that reach a receiver by multiple paths. Depending on the density of the scatter, the signal will display different fading characteristics. Rayleigh and Nakagami distributions are used to model dense scatters, while Rician distributions model fading with a stronger line-of-sight. Nakagami distributions can be reduced to Rayleigh distributions, but give more control over the extent of the fading.

Parameters

To estimate distribution parameters, use `mle` or the Distribution Fitter app.

See Also

`NakagamiDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Negative Binomial Distribution

In this section...

Definition

When the r parameter is an integer, the negative binomial pdf is

$$y = f(x \mid r, p) = \binom{r + x - 1}{x} p^r q^x I_{(0, 1, \ldots)}(x)$$

where $q = 1 - p$. When r is not an integer, the binomial coefficient in the definition of the pdf is replaced by the equivalent expression

$$\frac{\Gamma(r + x)}{\Gamma(r)\Gamma(x + 1)}$$

Background

In its simplest form (when r is an integer), the negative binomial distribution models the number of failures x before a specified number of successes is reached in a series of independent, identical trials. Its parameters are the probability of success in a single trial, p, and the number of successes, r. A special case of the negative binomial distribution, when $r = 1$, is the geometric distribution on page B-63, which models the number of failures before the first success.

More generally, r can take on non-integer values. This form of the negative binomial distribution has no interpretation in terms of repeated trials, but, like the Poisson distribution on page B-131, it is useful in modeling count data. The negative binomial distribution is more general than the Poisson distribution because it has a variance that is greater than its mean, making it suitable for count data that do not meet the assumptions of the Poisson distribution. In the limit, as r increases to infinity, the negative binomial distribution approaches the Poisson distribution.

Parameters

Suppose you are collecting data on the number of auto accidents on a busy highway, and would like to be able to model the number of accidents per day. Because these are count data, and because there are a very large number of cars and a small probability of an accident for any specific car, you might think to use the Poisson distribution. However, the probability of having an accident is likely to vary from day to day as the weather and amount of traffic change, and so the assumptions needed for the Poisson distribution are not met. In particular, the variance of this type of count data sometimes exceeds the mean by a large amount. The data below exhibit this effect: most days have few or no accidents, and a few days have a large number.

```
accident = [2  3  4  2  3  1  12  8  14  31  23  1  10  7  0];
m = mean(accident)
```

```
m = 8.0667
```

```
v = var(accident)
```

```
v = 79.3524
```

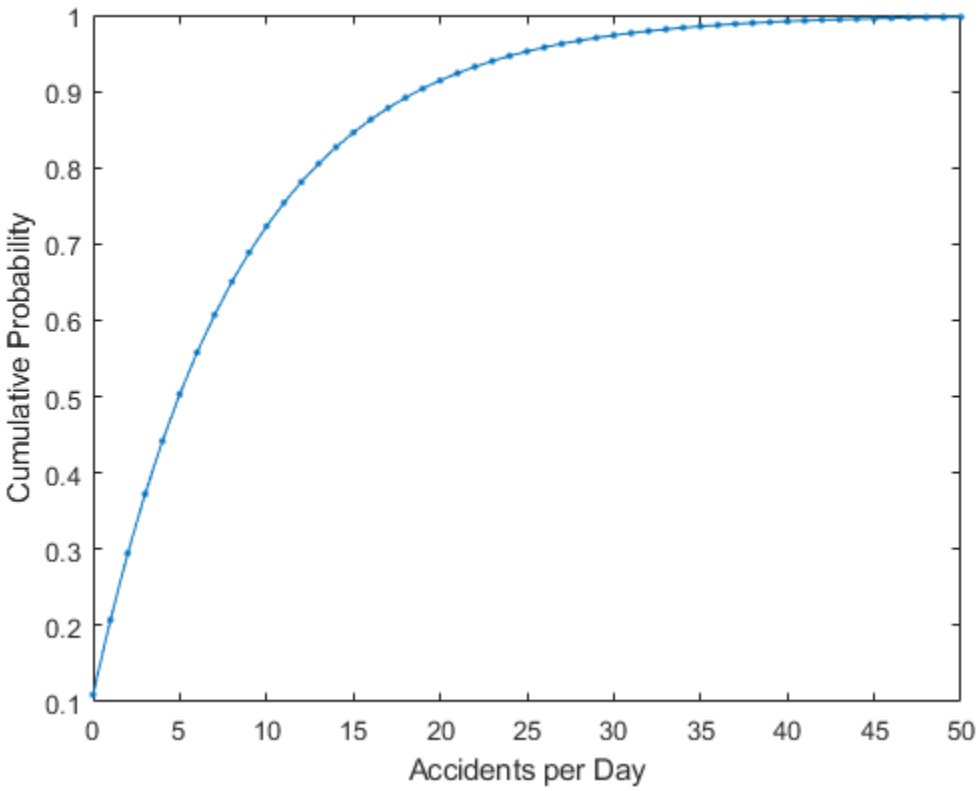
The negative binomial distribution is more general than the Poisson, and is often suitable for count data when the Poisson is not. The function `nbinfit` returns the maximum likelihood estimates (MLEs) and confidence intervals for the parameters of the negative binomial distribution. Here are the results from fitting the `accident` data:

```
[phat,pci] = nbinfit(accident)
```

```
phat = 1×2

    1.0060    0.1109


pci = 2×2

    0.2152    0.0171
    1.7968    0.2046
```

It is difficult to give a physical interpretation in this case to the individual parameters. However, the estimated parameters can be used in a model for the number of daily accidents. For example, a plot of the estimated cumulative probability function shows that while there is an estimated 10% chance of no accidents on a given day, there is also about a 10% chance that there will be 20 or more accidents.

```
plot(0:50,nbincdf(0:50,phat(1),phat(2)),'.-');
xlabel('Accidents per Day')
ylabel('Cumulative Probability')
```

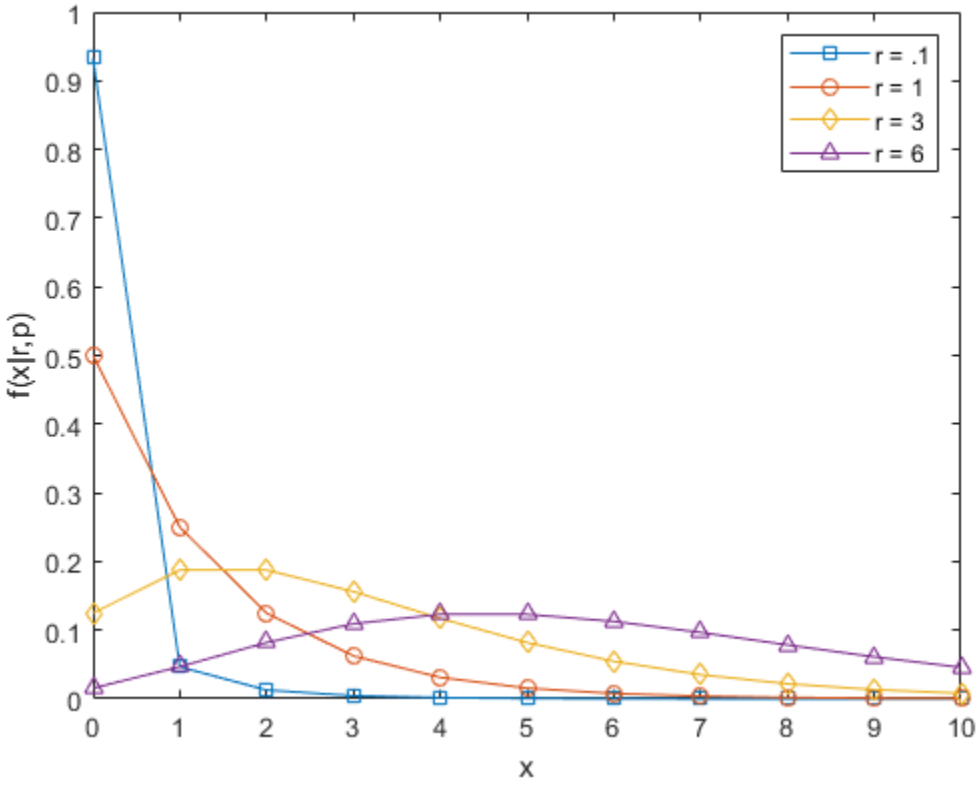


Example

Compute and Plot Negative Binomial Distribution PDF

Compute and plot the pdf using four different values for the parameter `r`, the desired number of successes: `.1`, `1`, `3`, and `6`. In each case, the probability of success `p` is `.5`.

```
x = 0:10;
plot(x,nbinpdf(x,.1,.5),'s-', ...
     x,nbinpdf(x,1,.5),'o-', ...
     x,nbinpdf(x,3,.5),'d-', ...
     x,nbinpdf(x,6,.5),'^-');
legend({'r = .1' 'r = 1' 'r = 3' 'r = 6'})
xlabel('x')
ylabel('f(x|r,p)')
```



The plot shows that the negative binomial distribution can take on a variety of shapes, ranging from very skewed to nearly symmetric, depending on the value of `r`.

See Also

`NegativeBinomialDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Noncentral Chi-Square Distribution

In this section...

Definition

There are many equivalent formulas for the noncentral chi-square distribution function. One formulation uses a modified Bessel function of the first kind. Another uses the generalized Laguerre polynomials. The cumulative distribution function is computed using a weighted sum of χ^2 probabilities with the weights equal to the probabilities of a Poisson distribution. The Poisson parameter is one-half of the noncentrality parameter of the noncentral chi-square

$$F(x\left|\nu,\delta\right.) = \sum_{j=0}^{\infty}\left(\frac{\left(\frac{1}{2}\delta\right)^j}{j!}e^{\frac{-\delta}{2}}\right)\Pr\left[\chi^2_{\nu+2j} \le x\right]$$

where δ is the noncentrality parameter.

Background

The χ^2 distribution is actually a simple special case of the noncentral chi-square distribution. One way to generate random numbers with a χ^2 distribution (with ν degrees of freedom) is to sum the squares of ν standard normal random numbers (mean equal to zero.)

What if the normally distributed quantities have a mean other than zero? The sum of squares of these numbers yields the noncentral chi-square distribution. The noncentral chi-square distribution requires two parameters: the degrees of freedom and the noncentrality parameter. The noncentrality parameter is the sum of the squared means of the normally distributed quantities.

The noncentral chi-square has scientific application in thermodynamics and signal processing. The literature in these areas may refer to it as the “Rician Distribution” on page B-139 or generalized “Rayleigh Distribution” on page B-137.

Examples

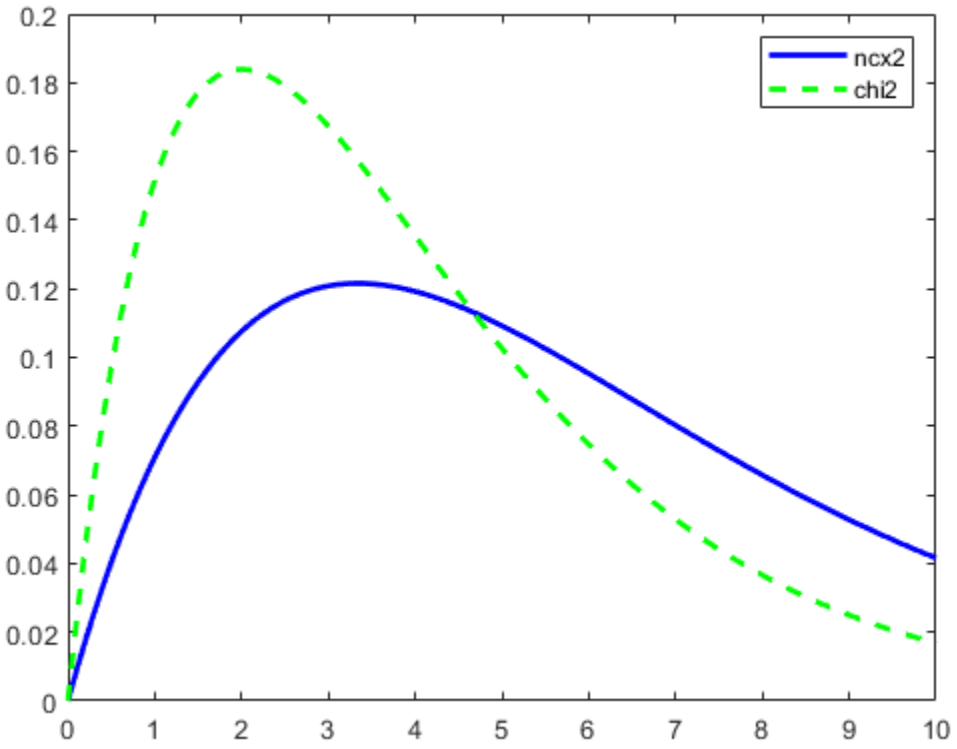
Compute Noncentral Chi-Square Distribution pdf

Compute the pdf of a noncentral chi-square distribution with degrees of freedom `V = 4` and noncentrality parameter `DELTA = 2`. For comparison, also compute the pdf of a chi-square distribution with the same degrees of freedom.

```
x = (0:0.1:10)';
ncx2 = ncx2pdf(x,4,2);
chi2 = chi2pdf(x,4);
```

Plot the pdf of the noncentral chi-square distribution on the same figure as the pdf of the chi-square distribution.

```
figure;
plot(x,ncx2,'b-','LineWidth',2)
hold on
plot(x,chi2,'g--','LineWidth',2)
legend('ncx2','chi2')
```



See Also

`ncx2cdf` | `ncx2inv` | `ncx2pdf` | `ncx2rnd` | `ncx2stat` | `random`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Noncentral F Distribution

In this section...

Definition

Similar to the noncentral χ^2 distribution, the toolbox calculates noncentral F distribution probabilities as a weighted sum of incomplete beta functions using Poisson probabilities as the weights.

$$F(x\left|\nu_1,\nu_2,\delta\right.)=\sum_{j=0}^{\infty}\left(\frac{\left(\frac{1}{2}\delta\right)^j}{j!}e^{\frac{-\delta}{2}}\right)I\left(\frac{\nu_1\cdot x}{\nu_2+\nu_1\cdot x}\left|\frac{\nu_1}{2}+j,\frac{\nu_2}{2}\right.\right)$$

$I(x|a,b)$ is the incomplete beta function with parameters a and b, and δ is the noncentrality parameter.

Background

As with the χ^2 distribution, the F distribution is a special case of the noncentral F distribution. The F distribution is the result of taking the ratio of χ^2 random variables each divided by its degrees of freedom.

If the numerator of the ratio is a noncentral chi-square random variable divided by its degrees of freedom, the resulting distribution is the noncentral F distribution.

The main application of the noncentral F distribution is to calculate the power of a hypothesis test relative to a particular alternative.

Examples

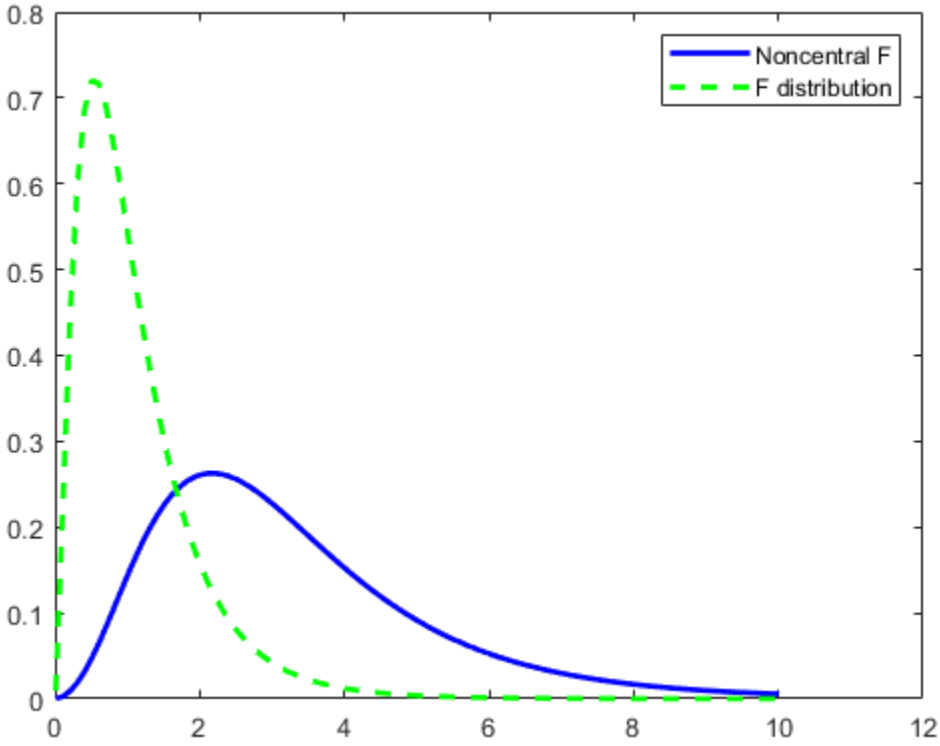
Compute Noncentral F Distribution pdf

Compute the pdf of a noncentral F distribution with degrees of freedom `NU1 = 5` and `NU2 = 20`, and noncentrality parameter `DELTA = 10`. For comparison, also compute the pdf of an F distribution with the same degrees of freedom.

```
x = (0.01:0.1:10.01)';
p1 = ncfpdf(x,5,20,10);
p = fpdf(x,5,20);
```

Plot the pdf of the noncentral F distribution and the pdf of the F distribution on the same figure.

```
figure;
plot(x,p1,'b-','LineWidth',2)
hold on
plot(x,p,'g--','LineWidth',2)
legend('Noncentral F','F distribution')
```



See Also

`ncfcdf` | `ncfinv` | `ncfpdf` | `ncfrnd` | `ncfstat` | `random`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Noncentral t Distribution

In this section...

Definition

The most general representation of the noncentral t distribution is quite complicated. Johnson and Kotz [67] give a formula for the probability that a noncentral t variate falls in the range [–u, u].

$$P(-u < x < u \mid \nu, \delta) = \sum_{j=0}^{\infty} \left(\frac{\left(\frac{1}{2}\delta^2\right)^j}{j!} e^{\frac{-\delta^2}{2}} \right) I\left(\frac{u^2}{\nu + u^2} \middle| \frac{1}{2} + j, \frac{\nu}{2} \right)$$

$I(x|\nu,\delta)$ is the incomplete beta function with parameters ν and δ. δ is the noncentrality parameter, and ν is the number of degrees of freedom.

Background

The noncentral t distribution is a generalization of Student's t distribution.

Student's t distribution with n – 1 degrees of freedom models the t-statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation of a random sample of size n from a normal population with mean μ. If the population mean is actually μ_0, then the t-statistic has a noncentral t distribution with noncentrality parameter

$$\delta = \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}$$

The noncentrality parameter is the normalized difference between μ_0 and μ.

The noncentral t distribution gives the probability that a t test will correctly reject a false null hypothesis of mean μ when the population mean is actually μ_0; that is, it gives the power of the t test. The power increases as the difference μ_0 – μ increases, and also as the sample size n increases.

Examples

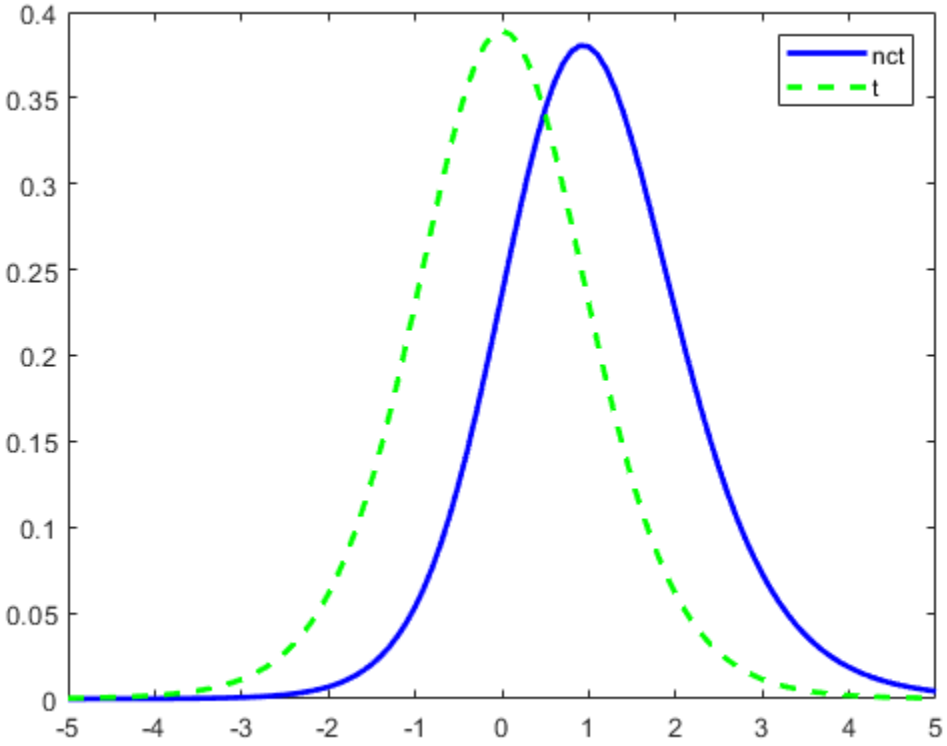
Compute Noncentral t Distribution pdf

Compute the pdf of a noncentral t distribution with degrees of freedom `V = 10` and noncentrality parameter `DELTA = 1`. For comparison, also compute the pdf of a t distribution with the same degrees of freedom.

```
x = (-5:0.1:5)';
nct = nctpdf(x,10,1);
t = tpdf(x,10);
```

Plot the pdf of the noncentral *t* distribution and the pdf of the *t* distribution on the same figure.

```
plot(x,nct,'b-','LineWidth',2)
hold on
plot(x,t,'g--','LineWidth',2)
legend('nct','t')
```



See Also

`nctcdf` | `nctinv` | `nctpdf` | `nctrnd` | `nctstat` | `random`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Normal Distribution

In this section...

Overview

The normal distribution, sometimes called the Gaussian distribution, is a two-parameter family of curves. The usual justification for using the normal distribution for modeling is the Central Limit theorem, which states (roughly) that the sum of independent samples from any distribution with finite mean and variance converges to the normal distribution as the sample size goes to infinity.

Statistics and Machine Learning Toolbox offers several ways to work with the normal distribution.

- Create a probability distribution object `NormalDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the normal distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`normcdf`, `normpdf`, `norminv`, `normlike`, `normstat`, `normfit`, `normrnd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple normal distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Normal'`) and parameters.

Parameters

The normal distribution uses these parameters.

Parameter	Description	Support
`mu` (μ)	Mean	$-\infty < \mu < \infty$
`sigma` (σ)	Standard deviation	$\sigma \geq 0$

The standard normal distribution has zero mean and unit standard deviation. If z is standard normal, then $\sigma z + \mu$ is also normal with mean μ and standard deviation σ. Conversely, if x is normal with mean μ and standard deviation σ, then $z = (x - \mu) / \sigma$ is standard normal.

Parameter Estimation

The *maximum likelihood estimates* (MLEs) are the parameter estimates that maximize the likelihood function. The maximum likelihood estimators of μ and σ^2 for the normal distribution, respectively, are

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

and

$$s^2_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

$\bar{x}$ is the sample mean for samples x_1, x_2, ..., x_n. The sample mean is an unbiased estimator of the parameter μ. However, s^2_{MLE} is a biased estimator of the parameter σ^2, meaning that its expected value does not equal the parameter.

The *minimum variance unbiased estimator* (MVUE) is commonly used to estimate the parameters of the normal distribution. The MVUE is the estimator that has the minimum variance of all unbiased estimators of a parameter. The MVUEs of the parameters μ and σ^2 for the normal distribution are the sample mean $\bar{x}$ and sample variance s^2, respectively.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

To fit the normal distribution to data and find the parameter estimates, use `normfit`, `fitdist`, or `mle`.

- For uncensored data, `normfit` and `fitdist` find the unbiased estimates, and `mle` finds the maximum likelihood estimates.
- For censored data, `normfit`, `fitdist`, and `mle` find the maximum likelihood estimates.

Unlike `normfit` and `mle`, which return parameter estimates, `fitdist` returns the fitted probability distribution object `NormalDistribution`. The object properties `mu` and `sigma` store the parameter estimates.

For an example, see “Fit Normal Distribution Object” on page B-121.

Probability Density Function

The normal probability density function (pdf) is

$$y = f(x \,|\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad \text{for } x \in \mathbb{R}.$$

The *likelihood function* is the pdf viewed as a function of the parameters. The maximum likelihood estimates (MLEs) are the parameter estimates that maximize the likelihood function for fixed values of `x`.

For an example, see “Compute and Plot the Normal Distribution pdf” on page B-121.

Cumulative Distribution Function

The normal cumulative distribution function (cdf) is

$$p = F(x \,|\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt, \quad \text{for } x \in \mathbb{R}.$$

p is the probability that a single observation from a normal distribution with parameters μ and σ falls in the interval $(-\infty,x]$.

The standard normal cumulative distribution function $\Phi(x)$ is functionally related to the error function `erf`.

$$\Phi(x) = \frac{1}{2}\left(1 - \mathrm{erf}\left(-\frac{x}{\sqrt{2}}\right)\right)$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt = 2\Phi(\sqrt{2}x) - 1.$$

For an example, see “Plot Standard Normal Distribution cdf” on page B-122

Examples

Fit Normal Distribution Object

Load the sample data and create a vector containing the first column of student exam grade data.

```
load examgrades
x = grades(:,1);
```

Create a normal distribution object by fitting it to the data.

```
pd = fitdist(x,'Normal')
```

```
pd =
  NormalDistribution

  Normal distribution
       mu = 75.0083   [73.4321, 76.5846]
    sigma =  8.7202   [7.7391, 9.98843]
```

The intervals next to the parameter estimates are the 95% confidence intervals for the distribution parameters.

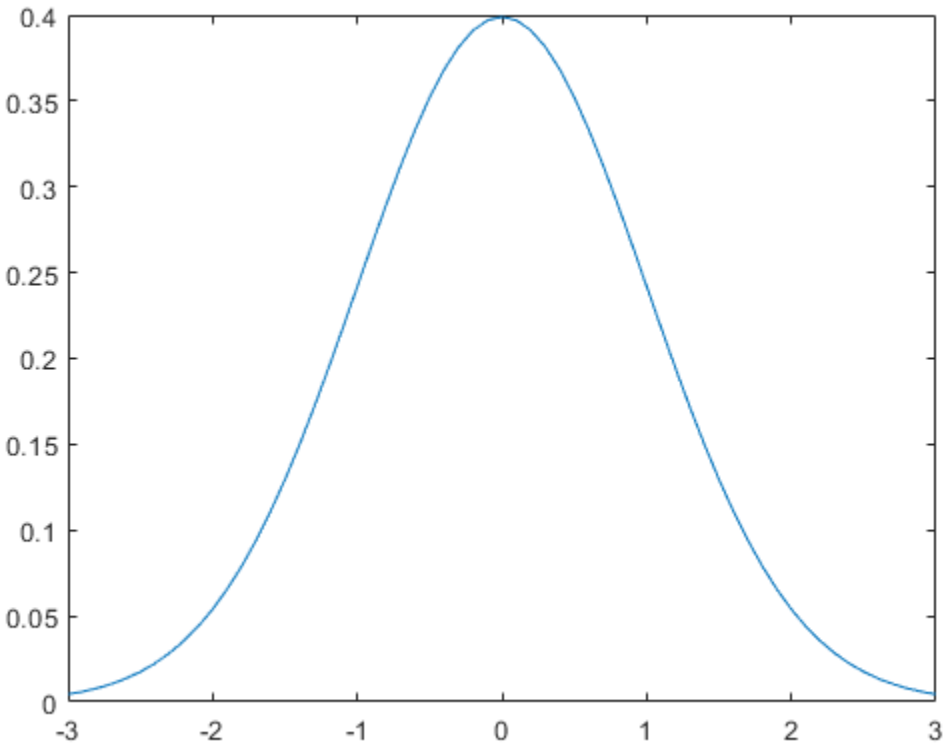
Compute and Plot the Normal Distribution pdf

Compute the pdf of a standard normal distribution, with parameters μ equal to 0 and σ equal to 1.

```
x = [-3:.1:3];
y = normpdf(x,0,1);
```

Plot the pdf.

```
plot(x,y)
```



Plot Standard Normal Distribution cdf

Create a standard normal distribution object.

```
pd = makedist('Normal')
```

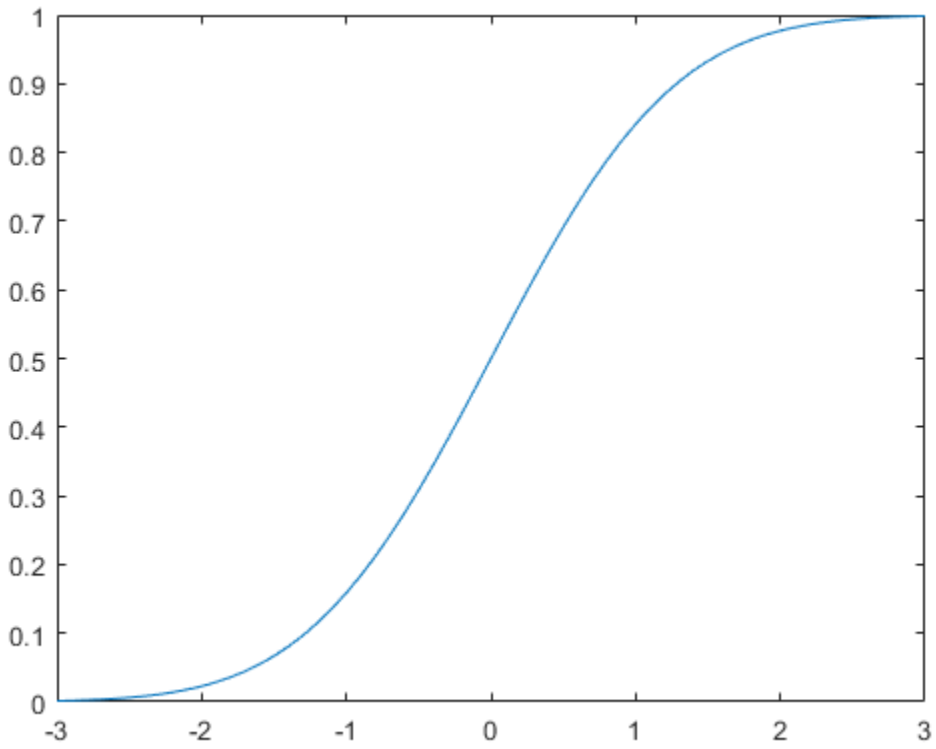
```
pd =
  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1
```

Specify the x values and compute the cdf.

```
x = -3:.1:3;
p = cdf(pd,x);
```

Plot the cdf of the standard normal distribution.

```
plot(x,p)
```



Compare Gamma and Normal Distribution pdfs

The gamma distribution has the shape parameter a and the scale parameter b. For a large a, the gamma distribution closely approximates the normal distribution with mean $\mu = ab$ and variance $\sigma^2 = ab^2$.

Compute the pdf of a gamma distribution with parameters `a = 100` and `b = 5`.

```
a = 100;
b = 5;
x = 250:750;
y_gam = gampdf(x,a,b);
```

For comparison, compute the mean, standard deviation, and pdf of the normal distribution that gamma approximates.

```
mu = a*b
```

```
mu = 500
```

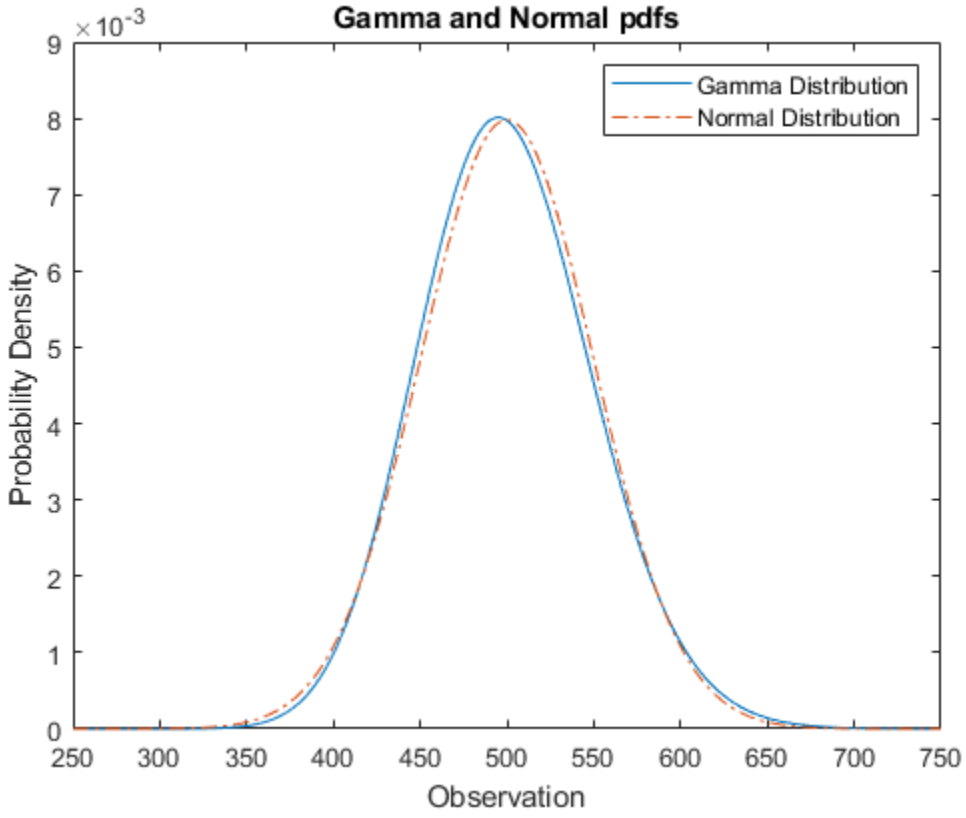
```
sigma = sqrt(a*b^2)
```

```
sigma = 50
```

```
y_norm = normpdf(x,mu,sigma);
```

Plot the pdfs of the gamma distribution and the normal distribution on the same figure.

```
plot(x,y_gam,'-',x,y_norm,'-.')
title('Gamma and Normal pdfs')
xlabel('Observation')
ylabel('Probability Density')
legend('Gamma Distribution','Normal Distribution')
```



The pdf of the normal distribution approximates the pdf of the gamma distribution.

Relationship Between Normal and Lognormal Distributions

If X follows the lognormal distribution with parameters μ and σ, then $\log(X)$ follows the normal distribution with mean μ and standard deviation σ. Use distribution objects to inspect the relationship between normal and lognormal distributions.

Create a lognormal distribution object by specifying the parameter values.

```
pd = makedist('Lognormal','mu',5,'sigma',2)
```

```
pd =
  LognormalDistribution

  Lognormal distribution
       mu = 5
    sigma = 2
```

Compute the mean of the lognormal distribution.

```
mean(pd)
```

```
ans = 1.0966e+03
```

The mean of the lognormal distribution is not equal to the `mu` parameter. The mean of the logarithmic values is equal to `mu`. Confirm this relationship by generating random numbers.

Generate random numbers from the lognormal distribution and compute their log values.

```
rng('default');  % For reproducibility
x = random(pd,10000,1);
logx = log(x);
```

Compute the mean of the logarithmic values.

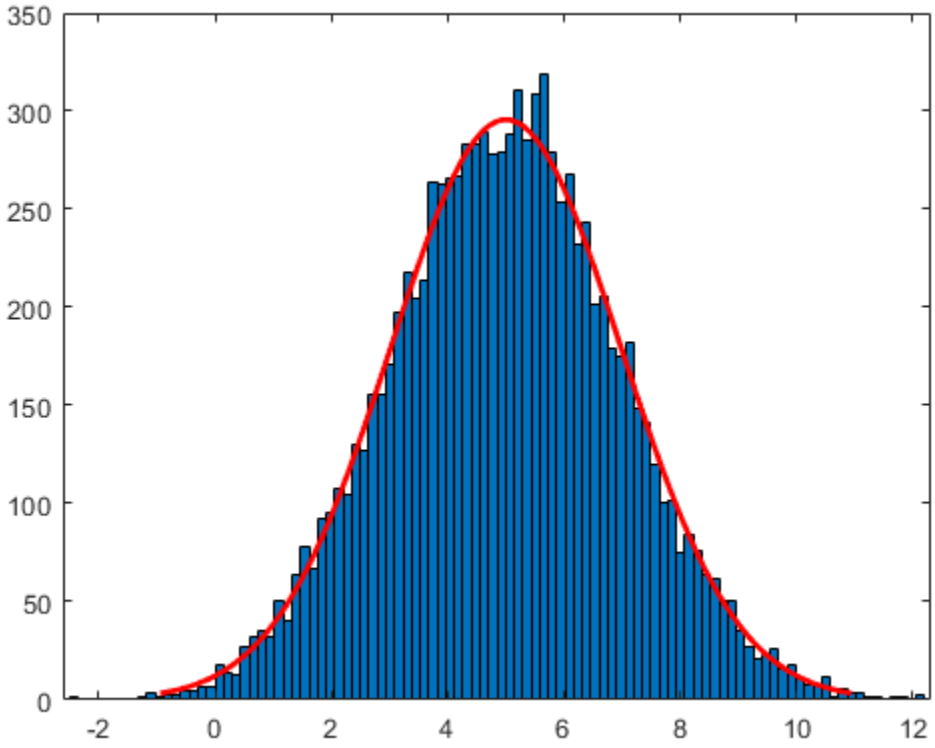
```
m = mean(logx)
```

```
m = 5.0033
```

The mean of the log of `x` is close to the `mu` parameter of `x`, because `x` has a lognormal distribution.

Construct a histogram of `logx` with a normal distribution fit.

```
histfit(logx)
```



The plot shows that the log values of `x` are normally distributed.

`histfit` uses `fitdist` to fit a distribution to data. Use `fitdist` to obtain parameters used in fitting.

```
pd_normal = fitdist(logx,'Normal')

pd_normal =
  NormalDistribution

  Normal distribution
       mu = 5.00332   [4.96445, 5.04219]
    sigma = 1.98296   [1.95585, 2.01083]
```

The estimated normal distribution parameters are close to the lognormal distribution parameters 5 and 2.

Compare Student's t and Normal Distribution pdfs

The Student's t distribution is a family of curves depending on a single parameter ν (the degrees of freedom). As the degrees of freedom ν approach infinity, the t distribution approaches the standard normal distribution.

Compute the pdfs for the Student's t distribution with the parameter `nu = 5` and the Student's t distribution with the parameter `nu = 15`.

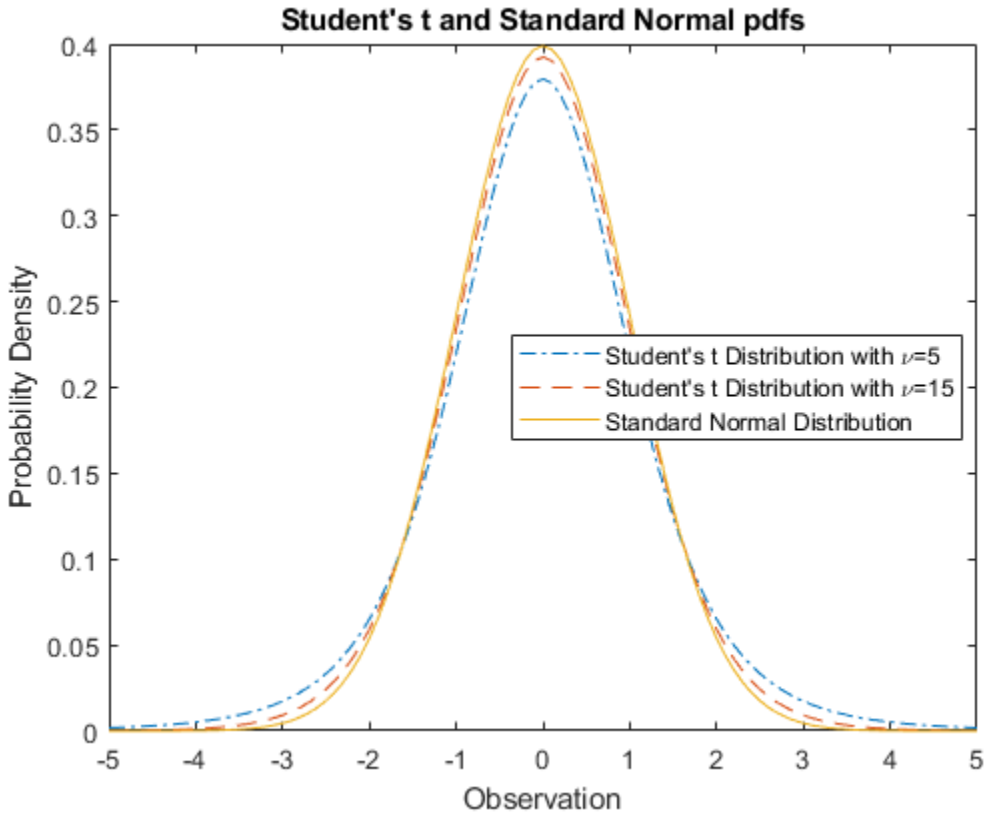
```
x = [-5:0.1:5];
y1 = tpdf(x,5);
y2 = tpdf(x,15);
```

Compute the pdf for a standard normal distribution.

```
z = normpdf(x,0,1);
```

Plot the Student's t pdfs and the standard normal pdf on the same figure.

```
plot(x,y1,'-.',x,y2,'--',x,z,'-')
legend('Student''s t Distribution with \nu=5', ...
    'Student''s t Distribution with \nu=15', ...
    'Standard Normal Distribution','Location','best')
xlabel('Observation')
ylabel('Probability Density')
title('Student''s t and Standard Normal pdfs')
```



The standard normal pdf has shorter tails than the Student's *t* pdfs.

Related Distributions

- "Binomial Distribution" on page B-10 — The binomial distribution models the total number of successes in n repeated trials with the probability of success p. As n increases, the binomial distribution can be approximated by a normal distribution with $\mu = np$ and $\sigma^2 = np(1–p)$. See "Compare Binomial and Normal Distribution pdfs" on page B-14.
- "Birnbaum-Saunders Distribution" on page B-18 — If x has a Birnbaum-Saunders distribution with parameters β and γ, then

  $$\frac{\left(\sqrt{x/\beta} - \sqrt{\beta/x}\right)}{\gamma}$$

  has a standard normal distribution.
- "Chi-Square Distribution" on page B-28 — The chi-square distribution is the distribution of the sum of squared, independent, standard normal random variables. If a set of n observations is normally distributed with variance σ^2, and s^2 is the sample variance, then $(n–1)s^2/\sigma^2$ has a chi-square distribution with $n–1$ degrees of freedom. The `normfit` function uses this relationship to calculate confidence intervals for the estimate of the normal parameter σ^2 .
- "Extreme Value Distribution" on page B-40 — The extreme value distribution is appropriate for modeling the smallest or largest value from a distribution whose tails decay exponentially fast, such as, the normal distribution.

- “Gamma Distribution” on page B-47 — The gamma distribution has the shape parameter a and the scale parameter b. For a large a, the gamma distribution closely approximates the normal distribution with mean $\mu = ab$ and variance $\sigma^2 = ab^2$. The gamma distribution has density only for positive real numbers. See “Compare Gamma and Normal Distribution pdfs” on page B-123.
- “Half-Normal Distribution” on page B-68 — The half-normal distribution is a special case of the folded normal and truncated normal distributions. If a random variable Z has a standard normal distribution, then $X = \mu + \sigma|Z|$ has a half-normal distribution with parameters μ and σ.
- “Logistic Distribution” on page B-85 — The logistic distribution is used for growth models and in logistic regression. It has longer tails and a higher kurtosis than the normal distribution.
- “Lognormal Distribution” on page B-88 — If X follows the lognormal distribution with parameters μ and σ, then $\log(X)$ follows the normal distribution with mean μ and standard deviation σ. See “Relationship Between Normal and Lognormal Distributions” on page B-124.
- “Multivariate Normal Distribution” on page B-98 — The multivariate normal distribution is a generalization of the univariate normal to two or more variables. It is a distribution for random vectors of correlated variables, in which each element has a univariate normal distribution. In the simplest case, there is no correlation among variables, and elements of the vectors are independent, univariate normal random variables.
- “Poisson Distribution” on page B-131 — The Poisson distribution is a one-parameter discrete distribution that takes nonnegative integer values. The parameter, λ, is both the mean and the variance of the distribution. As λ increase, the Poisson distribution can be approximated by a normal distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$.
- “Rayleigh Distribution” on page B-137 — The Rayleigh distribution is a special case of the Weibull distribution with applications in communications theory. If the component velocities of a particle in the x and y directions are two independent normal random variables with zero means and equal variances, then the distance the particle travels per unit time follows the Rayleigh distribution.
- “Stable Distribution” on page B-140 — The normal distribution is a special case of the stable distribution. The stable distribution with the first shape parameter $\alpha = 2$ corresponds to the normal distribution.

  $$N\left(\mu,\sigma^2\right) = S\left(2,0,\frac{\sigma}{\sqrt{2}},\mu\right).$$

- “Student's t Distribution” on page B-149 — The Student’s t distribution is a family of curves depending on a single parameter ν (the degrees of freedom). As the degrees of freedom ν goes to infinity, the t distribution approaches the standard normal distribution. See “Compare Student's t and Normal Distribution pdfs” on page B-126.

  If x is a random sample of size n from a normal distribution with mean μ, then the statistic

  $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

  where $\bar{x}$ is the sample mean and s is the sample standard deviation, has the Student's t distribution with n–1 degrees of freedom.
- “t Location-Scale Distribution” on page B-156 — The t location-scale distribution is useful for modeling data distributions with heavier tails (more prone to outliers) than the normal distribution. It approaches the normal distribution as the shape parameter ν approaches infinity.

References

[1] Abramowitz, M., and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1964.

[2] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 1993.

[3] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 1982.

[4] Marsaglia, G., and W. W. Tsang. "A Fast, Easily Implemented Method for Sampling from Decreasing or Symmetric Unimodal Density Functions." *SIAM Journal on Scientific and Statistical Computing*. Vol. 5, Number 2, 1984, pp. 349–359.

[5] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

See Also

`NormalDistribution` | `erf` | `normcdf` | `normfit` | `norminv` | `normlike` | `normpdf` | `normrnd` | `normstat`

More About

- "Working with Probability Distributions" on page 5-3
- "Supported Distributions" on page 5-14
- "Compare Multiple Distribution Fits" on page 5-87
- "Multivariate Distributions" on page 5-20

Piecewise Linear Distribution

Overview

The piecewise linear distribution creates a nonparametric representation of the cumulative distribution function (cdf) by linearly connecting the known cdf values from the sample data.

Parameters

The piecewise linear distribution uses the following parameters.

Parameter	Description
x	Vector of x values at which the cdf changes slope
Fx	Vector of cdf values that correspond to each value in x

Cumulative Distribution Function

The piecewise linear distribution constructs a continuous cumulative distribution function (cdf) by connecting the empirical cdf values such that the cdf increases linearly between x(j) and x(j + 1).

Relationship to Other Distributions

The piecewise linear distribution is a continuous version of the discrete empirical cumulative distribution function (ecdf).

See Also

`PiecewiseLinearDistribution`

More About

- “Nonparametric and Empirical Probability Distributions” on page 5-30
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Poisson Distribution

In this section...

Overview

The Poisson distribution is a one-parameter family of curves that models the number of times a random event occurs. This distribution is appropriate for applications that involve counting the number of times a random event occurs in a given amount of time, distance, area, and so on. Sample applications that involve Poisson distributions include the number of Geiger counter clicks per second, the number of people walking into a store in an hour, and the number of packets lost over a network per minute.

Statistics and Machine Learning Toolbox offers several ways to work with the Poisson distribution.

- Create a probability distribution object `PoissonDistribution` by fitting a probability distribution to sample data or by specifying parameter values. Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the Poisson distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`poisscdf`, `poisspdf`, `poissinv`, `poisstat`, `poissfit`, `poissrnd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple Poisson distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Poisson'`) and parameters.

Parameters

The Poisson distribution uses the following parameter.

Parameter	Description	Support
`lambda` (λ)	Mean	$\lambda \geq 0$

The parameter λ is also equal to the variance of the Poisson distribution.

The sum of two Poisson random variables with parameters λ_1 and λ_2 is a Poisson random variable with parameter $\lambda = \lambda_1 + \lambda_2$.

Probability Density Function

The probability density function (pdf) of the Poisson distribution is

$$f(x\left|\lambda\right.) = \frac{\lambda^x}{x!}e^{-\lambda};\ x = 0, 1, 2, ..., \infty\ .$$

The result is the probability of exactly x occurrences of the random event. For discrete distributions, the pdf is also known as the probability mass function (pmf).

For an example, see “Compute Poisson Distribution pdf” on page B-132.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the Poisson distribution is

$$p = F(x\left|\lambda\right.) = e^{-\lambda}\sum_{i=0}^{floor(x)}\frac{\lambda^i}{i!}.$$

The result is the probability of at most x occurrences of the random event.

For an example, see “Compute Poisson Distribution cdf” on page B-133.

Examples

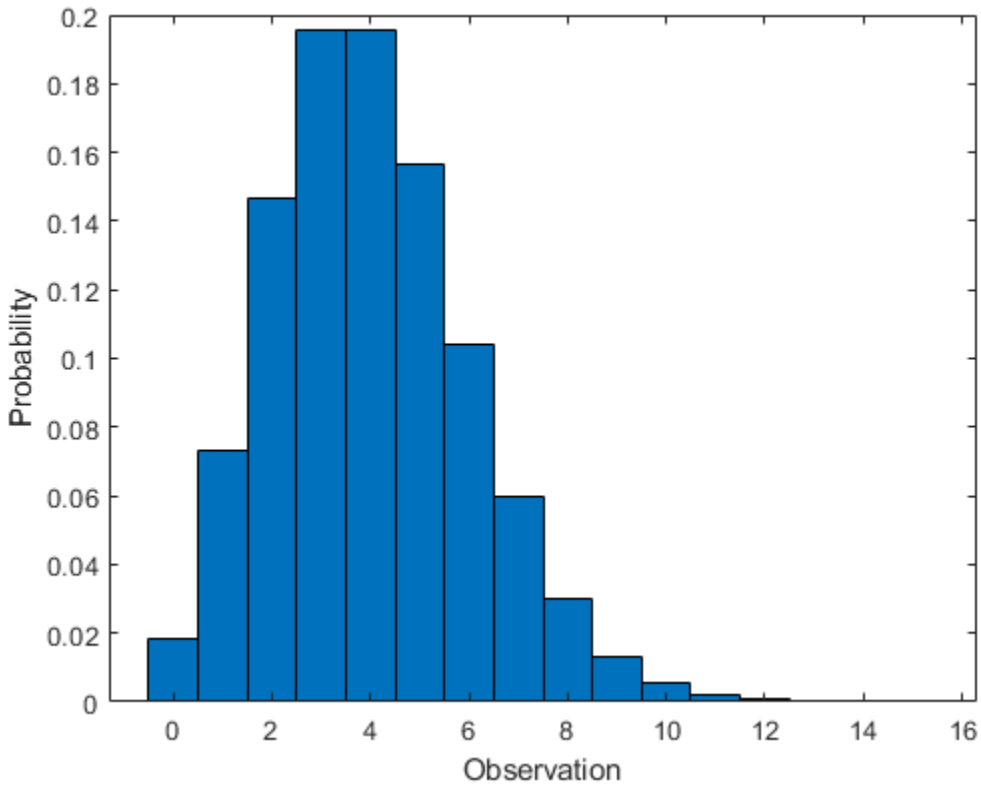
Compute Poisson Distribution pdf

Compute the pdf of the Poisson distribution with parameter `lambda = 4`.

```
x = 0:15;
y = poisspdf(x,4);
```

Plot the pdf with bars of width `1`.

```
figure
bar(x,y,1)
xlabel('Observation')
ylabel('Probability')
```



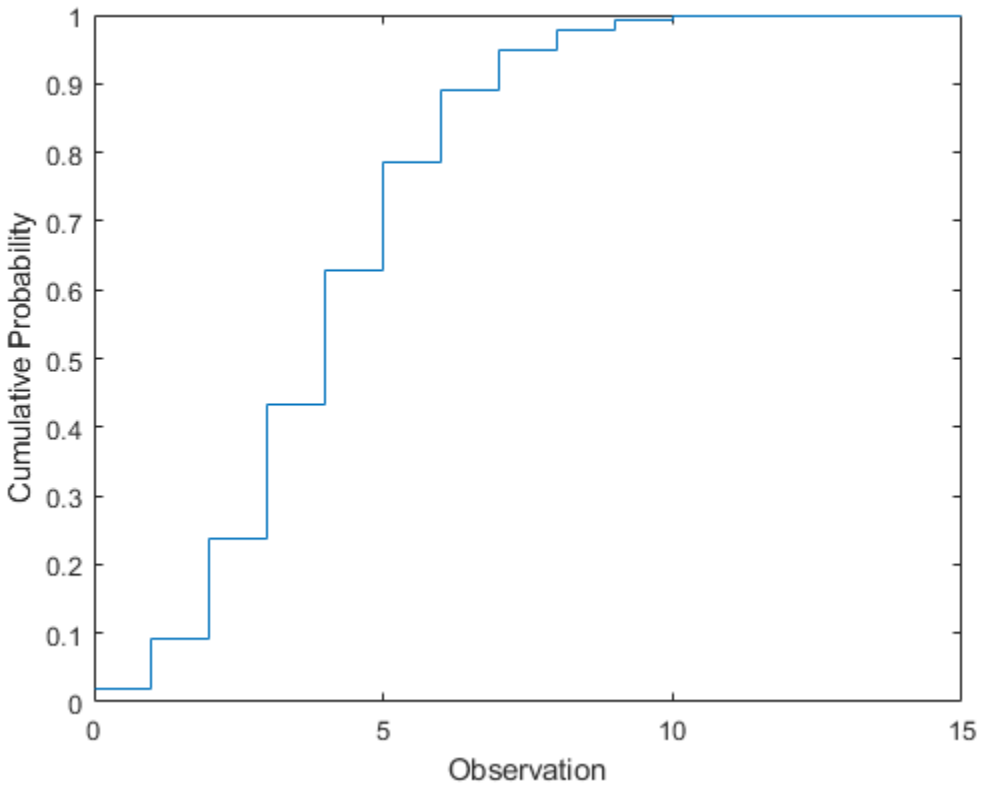
Compute Poisson Distribution cdf

Compute the cdf of the Poisson distribution with parameter `lambda = 4`.

```
x = 0:15;
y = poisscdf(x,4);
```

Plot the cdf.

```
figure
stairs(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Compare Poisson and Normal Distribution pdfs

When `lambda` is large, the Poisson distribution can be approximated by the normal distribution with mean `lambda` and variance `lambda`.

Compute the pdf of the Poisson distribution with parameter `lambda = 50`.

```
lambda = 50;
x1 = 0:100;
y1 = poisspdf(x1,lambda);
```

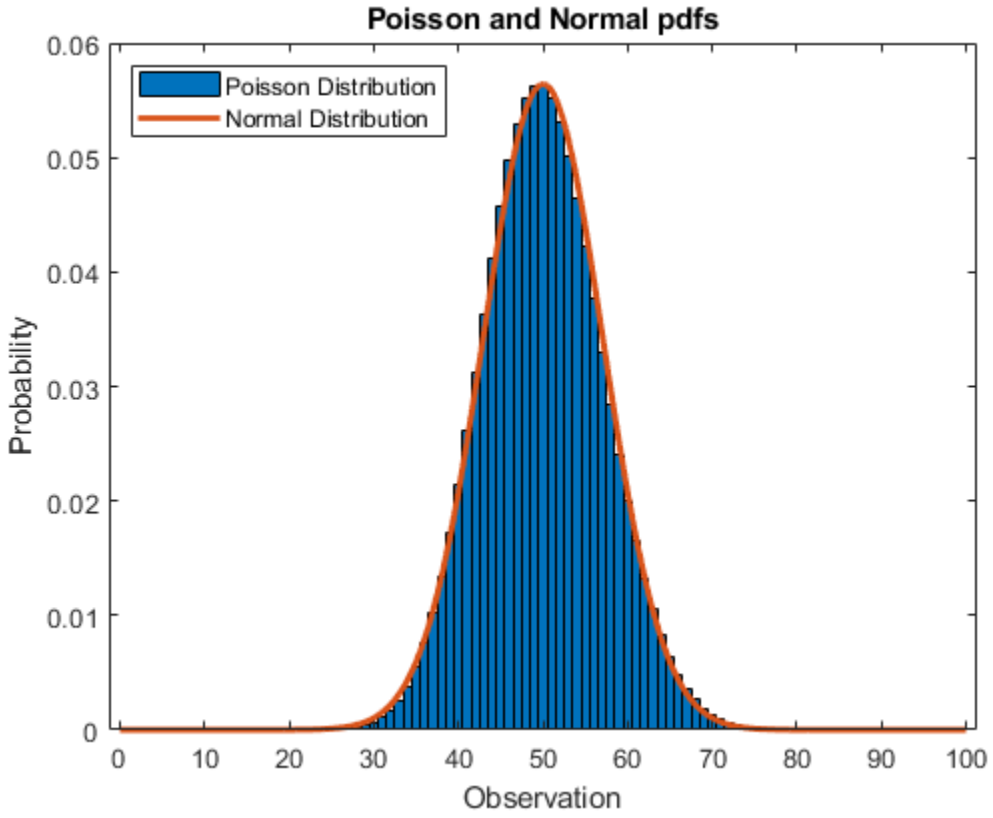
Compute the pdf of the corresponding normal distribution.

```
mu = lambda;
sigma = sqrt(lambda);
x2 = 0:0.1:100;
y2 = normpdf(x2,mu,sigma);
```

Plot the pdfs on the same axis.

```
figure
bar(x1,y1,1)
hold on
plot(x2,y2,'LineWidth',2)
xlabel('Observation')
ylabel('Probability')
title('Poisson and Normal pdfs')
```

```
legend('Poisson Distribution','Normal Distribution','location','northwest')
hold off
```



The pdf of the normal distribution closely approximates the pdf of the Poisson distribution.

Related Distributions

- "Binomial Distribution" on page B-10 — The binomial distribution is a two-parameter discrete distribution that counts the number of successes in N independent trials with the probability of success p. The Poisson distribution is the limiting case of a binomial distribution where N approaches infinity and p goes to zero while $Np = \lambda$. See "Compare Binomial and Poisson Distribution pdfs" on page B-15.
- "Exponential Distribution" on page B-33 — The exponential distribution is a one-parameter continuous distribution that has parameter μ (mean). The Poisson distribution models counts of the number of times a random event occurs in a given amount of time. In such a model, the amount of time between occurrences is modeled by the exponential distribution with mean $\frac{1}{\lambda}$.
- "Normal Distribution" on page B-119 — The normal distribution is a two-parameter continuous distribution that has parameters μ (mean) and σ (standard deviation). When λ is large, the Poisson distribution can be approximated by the normal distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$. See "Compare Poisson and Normal Distribution pdfs" on page B-134.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

[4] Loader, Catherine. *Fast and Accurate Computation of Binomial Probabilities*. July 9, 2000.

See Also

`PoissonDistribution` | `poisscdf` | `poissfit` | `poissinv` | `poisspdf` | `poissrnd` | `poisstat`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Rayleigh Distribution

In this section...

Definition

The Rayleigh pdf is

$$y = f(x \mid b) = \frac{x}{b^2} e\left(\frac{-x^2}{2b^2}\right)$$

Background

The Rayleigh distribution is a special case of the Weibull distribution on page B-170. If A and B are the parameters of the Weibull distribution, then the Rayleigh distribution with parameter b is equivalent to the Weibull distribution with parameters$A = \sqrt{2}b$ and $B = 2$.

If the component velocities of a particle in the x and y directions are two independent normal random variables with zero means and equal variances, then the distance the particle travels per unit time is distributed Rayleigh.

In communications theory, Nakagami distributions on page B-108, Rician distributions on page B-139, and Rayleigh distributions are used to model scattered signals that reach a receiver by multiple paths. Depending on the density of the scatter, the signal will display different fading characteristics. Rayleigh and Nakagami distributions are used to model dense scatters, while Rician distributions model fading with a stronger line-of-sight. Nakagami distributions can be reduced to Rayleigh distributions, but give more control over the extent of the fading.

Parameters

The `raylfit` function returns the MLE of the Rayleigh parameter. This estimate is

$$b = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2}$$

Examples

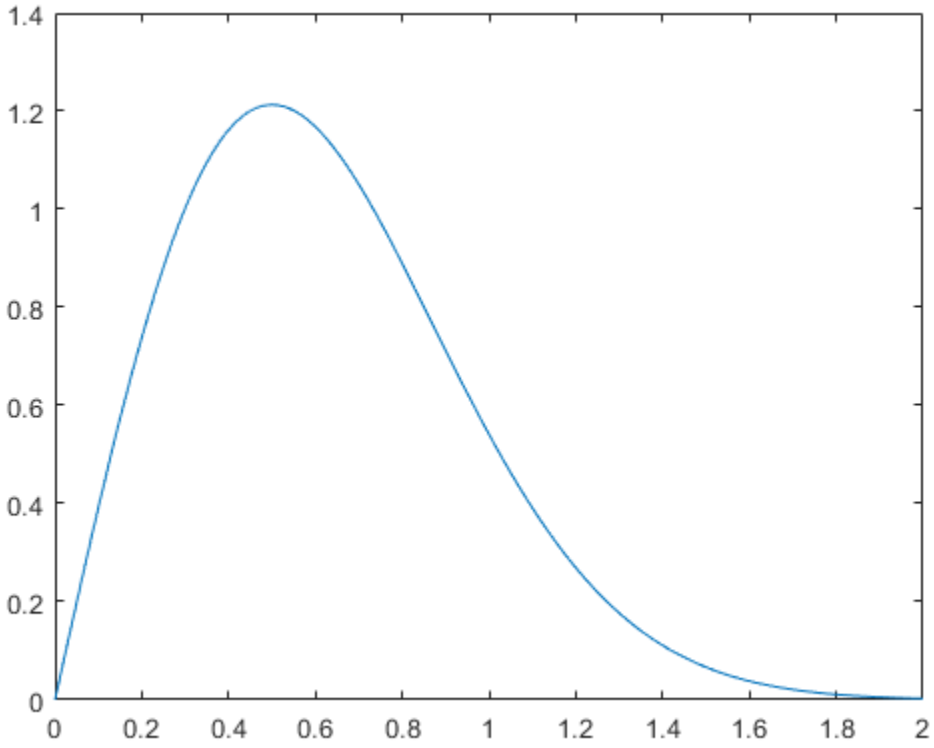
Compute and Plot Rayleigh Distribution pdf

Compute the pdf of a Rayleigh distribution with parameter `B = 0.5`.

```
x = [0:0.01:2];
p = raylpdf(x,0.5);
```

Plot the pdf.

```
figure;
plot(x,p)
```



See Also

`RayleighDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Rician Distribution

In this section...

Definition

The Rician distribution has the density function

$$I_0\left(\frac{xs}{\sigma^2}\right)\frac{x}{\sigma^2}e^{-\left(\frac{x^2+s^2}{2\sigma^2}\right)}$$

with noncentrality parameter $s \geq 0$ and scale parameter $\sigma > 0$, for $x > 0$. I_0 is the zero-order modified Bessel function of the first kind. If x has a Rician distribution with parameters s and σ, then $(x/\sigma)^2$ has a noncentral chi-square distribution with two degrees of freedom and noncentrality parameter $(s/\sigma)^2$.

Background

In communications theory, Nakagami distributions on page B-108, Rician distributions, and Rayleigh distributions on page B-137 are used to model scattered signals that reach a receiver by multiple paths. Depending on the density of the scatter, the signal will display different fading characteristics. Rayleigh and Nakagami distributions are used to model dense scatters, while Rician distributions model fading with a stronger line-of-sight. Nakagami distributions can be reduced to Rayleigh distributions, but give more control over the extent of the fading.

Parameters

To estimate distribution parameters, use `mle` or the Distribution Fitter app.

See Also

`RicianDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Stable Distribution

In this section...

Overview

Stable distributions are a class of probability distributions suitable for modeling heavy tails and skewness. A linear combination of two independent, identically-distributed stable-distributed random variables has the same distribution as the individual variables. In other words, if X_1, X_2, ..., X_n are independent and identically distributed stable random variables, then for every n

$$X_1 + X_2 + ... + X_n \overset{d}{=} c_n X + d_n$$

where the constant $c_n > 0$ and $d_n \in \mathbb{R}$.

The stable distribution is an application of the Generalized Central Limit Theorem, which states that the limit of normalized sums of independent identically distributed variables is stable.

Several different parameterizations exist for the stable distribution. The implementation in Statistics and Machine Learning Toolbox uses the parameterization described in [2]. In this case, a random variable X has the stable distribution $S(\alpha, \beta, \gamma, \delta; 0)$ if its characteristic function is given by:

$$E\left(e^{itX}\right) = \begin{cases} \exp\left(-\gamma^{\alpha}|t|^{\alpha}\left[1 + i\beta\,\mathrm{sign}(t)\tan\frac{\pi\alpha}{2}\left((\gamma|t|)^{1-\alpha} - 1\right)\right] + i\delta t\right) & for\ \alpha \neq 1, \\ \exp\left(-\gamma|t|\left[1 + i\beta\,\mathrm{sign}(t)\frac{2}{\pi}\ln(\gamma|t|)\right] + i\delta t\right) & for\ \alpha = 1 \end{cases}$$

Parameters

The stable distribution uses the following parameters.

Parameter	Description	Support
`alpha`	First shape parameter	$0 < \alpha \leq 2$
`beta`	Second shape parameter	$-1 \leq \beta \leq 1$
`gam`	Scale parameter	$0 < \gamma < \infty$
`delta`	Location parameter	$-\infty < \delta < \infty$

The first shape parameter, α, describes the tails of the distribution. The software computes the densities of the stable distribution using the direct integration method. As explained in [1], numerical difficulties exist with accurately computing the pdf and cdf when the α parameter is close to 1 or 0. If α is close to 1 (specifically, $0 < |\alpha - 1| < 0.02$), then the software rounds α to 1. If α is close to 0, then the densities may not be accurate.

The second shape parameter, β, describes the skewness of the distribution. If $\beta = 0$, then the distribution is symmetric. If $\beta > 0$, then the distribution is right-skewed. If $\beta < 0$, then the distribution is left-skewed. When α is small, the skewness of β is significant. As α increases, the effect of β decreases.

Probability Density Function

Definition

Most members of the stable distribution family do not have an explicit probability density function (pdf). Instead, the pdf is described in terms of the characteristic function [2].

Some special cases of the stable distribution, such as the normal, Cauchy, and Lévy distributions, have closed-form density functions. See Relationship to Other Distributions on page B-146 for more information.

Use `pdf` to calculate the probability density function for the stable distribution. The software computes the pdf using the direct integration method. As explained in [1], numerical difficulties exist with accurately computing the pdf when the α parameter is close to 1 or 0. If α is close to 1 (specifically, $0 < |\alpha - 1| < 0.02$), then the software rounds α to 1. If α is close to 0, then the densities may not be accurate.

Compare PDFs of Stable Distributions

The following plot compares the probability density functions for stable distributions with different `alpha` values. In each case, `beta = 0`, `gam = 1`, and `delta = 0`.

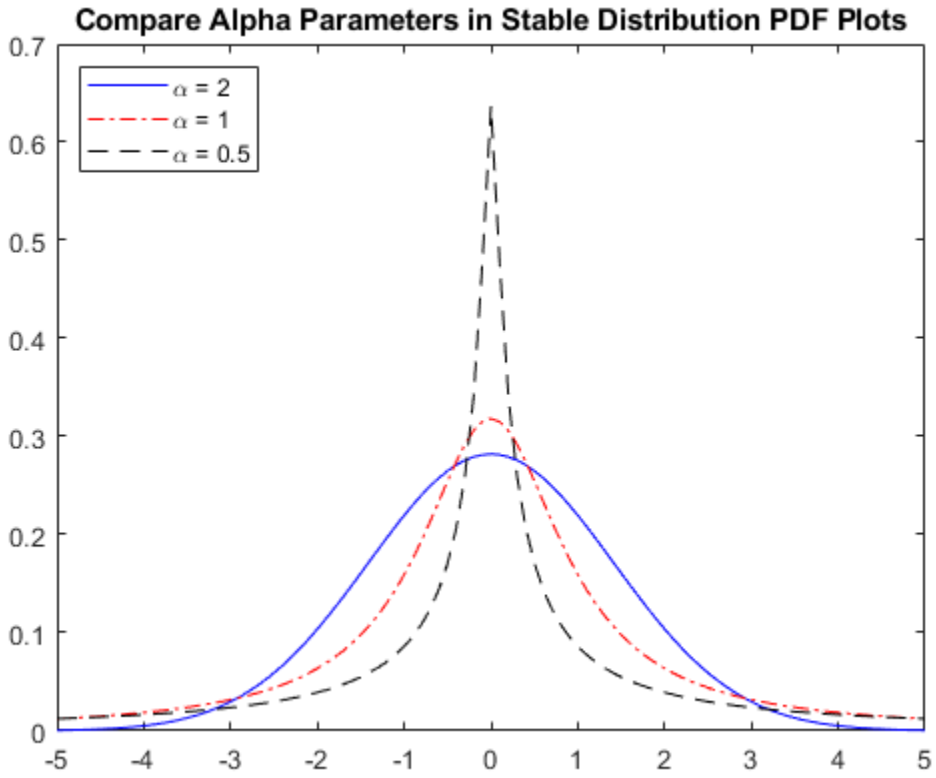
```
pd1 = makedist('Stable','alpha',2,'beta',0,'gam',1,'delta',0);
pd2 = makedist('Stable','alpha',1,'beta',0,'gam',1,'delta',0);
pd3 = makedist('Stable','alpha',0.5,'beta',0,'gam',1,'delta',0);
```

Calculate the pdf for each distribution.

```
x = -5:.1:5;
pdf1 = pdf(pd1,x);
pdf2 = pdf(pd2,x);
pdf3 = pdf(pd3,x);
```

Plot all three pdf functions on the same figure for visual comparison.

```
figure
plot(x,pdf1,'b-');
hold on
plot(x,pdf2,'r-.');
plot(x,pdf3,'k--');
title('Compare Alpha Parameters in Stable Distribution PDF Plots')
legend('\alpha = 2','\alpha = 1','\alpha = 0.5','Location','northwest')
hold off
```



The plot illustrates the effect of the `alpha` parameter on the tails of the distribution.

The next plot compares the probability density functions for stable distributions with different `beta` values. In each case, `alpha = 0.5`, `gam = 1`, and `delta = 0`.

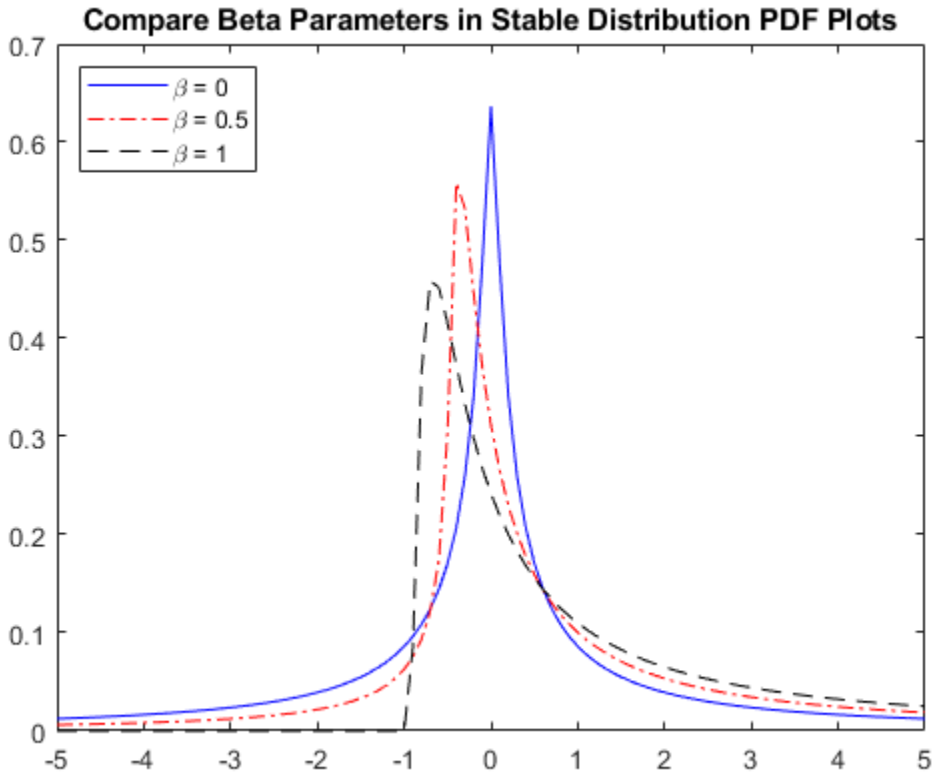
```
pd1 = makedist('Stable','alpha',0.5,'beta',0,'gam',1,'delta',0);
pd2 = makedist('Stable','alpha',0.5,'beta',0.5,'gam',1,'delta',0);
pd3 = makedist('Stable','alpha',0.5,'beta',1,'gam',1,'delta',0);
```

Calculate the pdf for each distribution.

```
x = -5:.1:5;
pdf1 = pdf(pd1,x);
pdf2 = pdf(pd2,x);
pdf3 = pdf(pd3,x);
```

Plot all three pdf functions on the same figure for visual comparison.

```
figure
plot(x,pdf1,'b-');
hold on
plot(x,pdf2,'r-.');
plot(x,pdf3,'k--');
title('Compare Beta Parameters in Stable Distribution PDF Plots')
legend('\beta = 0','\beta = 0.5','\beta = 1','Location','northwest')
hold off
```



Random Number Generation

Use `random` to generate random numbers from the stable distribution. The software generates random numbers for the stable distribution using the method proposed in [3]

Cumulative Distribution Function

Definition

Most members of the stable distribution family do not have an explicit cumulative distribution function (cdf). Instead, the cdf is described in terms of the characteristic function [2].

Use `cdf` to calculate the cumulative distribution function for the stable distribution. The software computes the cdf using the direct integration method. As explained in [1], numerical difficulties exist with accurately computing the cdf when the α parameter is close to 1 or 0. If α is close to 1 (specifically, $0 < |\alpha - 1| < 0.02$), then the software rounds α to 1. If α is close to 0, then the densities may not be accurate.

Compare CDFs of Stable Distributions

The following plot compares the cumulative distribution functions for stable distributions with different `alpha` values. In each case, `beta = 0`, `gam = 1`, and `delta = 0`.

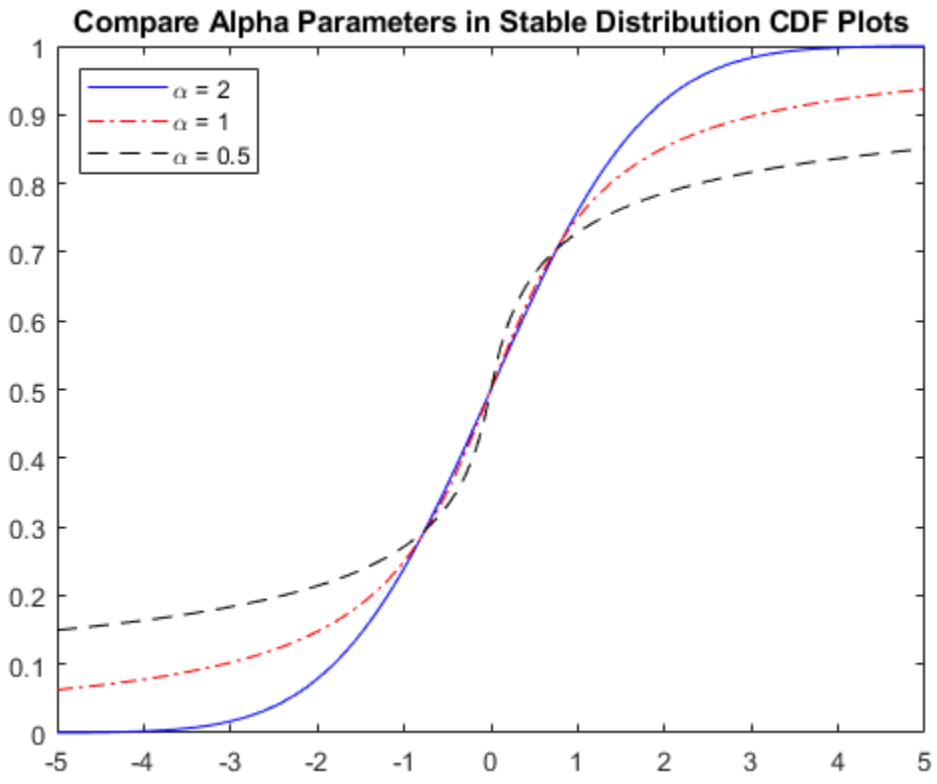
```
pd1 = makedist('Stable','alpha',2,'beta',0,'gam',1,'delta',0);
pd2 = makedist('Stable','alpha',1,'beta',0,'gam',1,'delta',0);
pd3 = makedist('Stable','alpha',0.5,'beta',0,'gam',1,'delta',0);
```

Calculate the cdf for each distribution.

```
x = -5:.1:5;
cdf1 = cdf(pd1,x);
cdf2 = cdf(pd2,x);
cdf3 = cdf(pd3,x);
```

Plot all three cdf functions on the same figure for visual comparison.

```
figure
plot(x,cdf1,'b-');
hold on
plot(x,cdf2,'r-.');
plot(x,cdf3,'k--');
title('Compare Alpha Parameters in Stable Distribution CDF Plots')
legend('\alpha = 2','\alpha = 1','\alpha = 0.5','Location','northwest')
hold off
```



The plot illustrates the effect of the `alpha` parameter on the shape of the cdf.

The next plot compares the cumulative distribution functions for stable distributions with different `beta` values. In all cases, `alpha = 0.5`, `gam = 1`, and `delta = 0`.

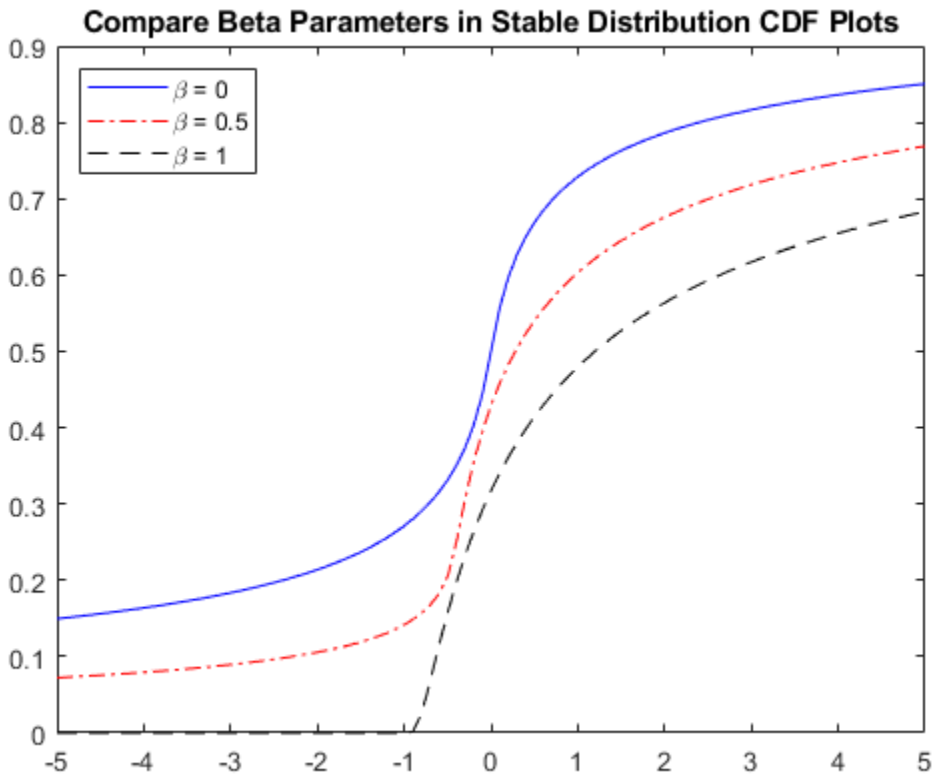
```
pd1 = makedist('Stable','alpha',0.5,'beta',0,'gam',1,'delta',0);
pd2 = makedist('Stable','alpha',0.5,'beta',0.5,'gam',1,'delta',0);
pd3 = makedist('Stable','alpha',0.5,'beta',1,'gam',1,'delta',0);
```

Calculate the cdf for each distribution.

```
x = -5:.1:5;
cdf1 = cdf(pd1,x);
cdf2 = cdf(pd2,x);
cdf3 = cdf(pd3,x);
```

Plot all three pdf functions on the same figure for visual comparison.

```
figure
plot(x,cdf1,'b-');
hold on
plot(x,cdf2,'r-.');
plot(x,cdf3,'k--');
title('Compare Beta Parameters in Stable Distribution CDF Plots')
legend('\beta = 0','\beta = 0.5','\beta = 1','Location','northwest')
hold off
```



Descriptive Statistics

The mean of the stable distribution is undefined for values of $\alpha \leq 1$. For $\alpha > 1$, the mean of the stable distribution is

$$\text{mean} = \delta - \beta\gamma\tan\left(\frac{\pi\alpha}{2}\right).$$

Use `mean` to calculate the mean of the stable distribution.

The variance of the stable distribution is undefined for values of α < 2. For α = 2, the variance of the stable distribution is

$$\text{var} = 2\gamma^2 .$$

Use `var` to calculate the variance of the stable distribution.

Relationship to Other Distributions

The stable distribution has three special cases: The normal distribution, the Cauchy distribution, and the Lévy distribution. These distributions are notable because they have closed-form probability density functions.

Normal Distribution

The normal, or Gaussian, distribution is a special case of the stable distribution. The stable distribution with $\alpha = 2$ corresponds to the normal distribution. In other words,

$$N\left(\mu, \sigma^2\right) = S\left(2, 0, \frac{\sigma}{\sqrt{2}}, \mu\right) .$$

μ is the mean and σ is the standard deviation of the normal distribution.

Although the value of β has no effect when $\alpha = 2$, the normal distribution is usually associated with $\beta = 0$.

The probability density function for the normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty .$$

A plot of the density for a normal distribution is symmetric and has a bell-shaped curve.

Cauchy Distribution

The Cauchy distribution is a special case of the stable distribution with α = 1 and β = 0. In other words,

$$Cauchy(\delta, \gamma) = S(1, 0, \gamma, \delta) ,$$

where γ is the scale parameter and δ is location parameter of the Cauchy distribution.

The probability density function for the Cauchy distribution is

$$f(x) = \frac{1}{\pi}\frac{\gamma}{\gamma^2 + (x-\delta)^2}, \quad -\infty < x < \infty .$$

A plot of the density for a Cauchy distribution is symmetric and has a bell-shaped curve, but has heavier tails than the density of a normal distribution.

Lévy Distribution

The Lévy distribution is a special case of the stable distribution where α = 0.5 and β = 1. In other words,

$$Lévy(\delta,\gamma) = S(0.5,1,\gamma,\gamma+\delta)\ .$$

where γ is the scale parameter and δ is location parameter of the Lévy distribution.

The probability density function for the Lévy distribution is

$$f(x) = \sqrt{\frac{\gamma}{2\pi}}\frac{1}{(x-\delta)^{3/2}}\exp\left(-\frac{\gamma}{2(x-\delta)}\right),\ \delta < x < \infty\ .$$

A plot of the density for a Lévy distribution is highly skewed and has heavy tails.

Comparison Plot for Stable Distributions

The following plot compares the probability density functions for the standard normal, Cauchy, and Lévy distributions.

Create a probability distribution object for the standard normal, Cauchy, and Lévy distributions.

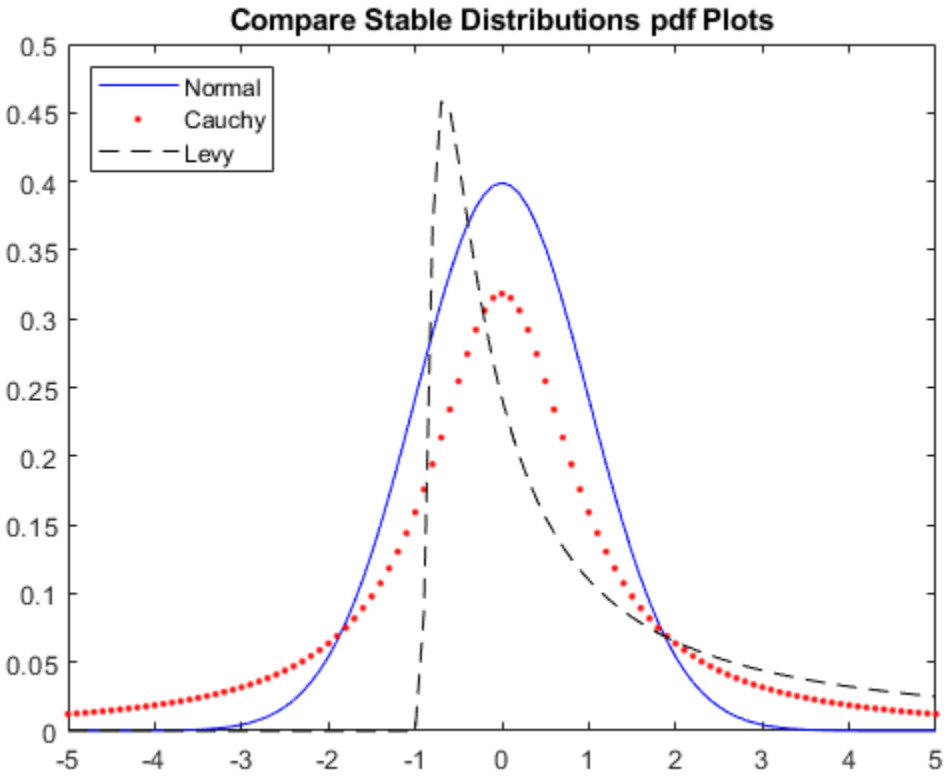
```
pd_norm = makedist('Stable','alpha',2,'beta',0,'gam',1/sqrt(2),'delta',0);
pd_cauchy = makedist('Stable','alpha',1,'beta',0,'gam',1,'delta',0);
pd_levy = makedist('Stable','alpha',0.5,'beta',1,'gam',1,'delta',0);
```

Calculate the pdf for each distribution.

```
x = -5:.1:5;
pdf_norm = pdf(pd_norm,x);
pdf_cauchy = pdf(pd_cauchy,x);
pdf_levy = pdf(pd_levy,x);
```

Plot all three pdf functions on the same figure for visual comparison.

```
figure
plot(x,pdf_norm,'b-');
hold on
plot(x,pdf_cauchy,'r.');
plot(x,pdf_levy,'k--');
title('Compare Stable Distributions pdf Plots')
legend('Normal','Cauchy','Levy','Location','northwest')
hold off
```



References

[1] Nolan, John P. “Numerical calculation of stable densities and distribution functions.” *Communications in Statistics: Stochastic Models*. Vol. 13, No. 4, 1997, pp. 759–774.

[2] Nolan, John P. *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-52915-4.

[3] Weron, A. and R. Weron. “Computer simulation of Lévy α-stable variables and processes.” *Lecture Notes in Physics*. Vol. 457, 1995, pp. 379–392.

See Also

`StableDistribution`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Student's t Distribution

In this section...

Overview

The Student’s *t* distribution is a one-parameter family of curves. This distribution is typically used to test a hypothesis regarding the population mean when the population standard deviation is unknown.

Statistics and Machine Learning Toolbox offers multiple ways to work with the Student’s *t* distribution.

- Use distribution-specific functions (`tcdf`, `tinv`, `tpdf`, `trnd`, `tstat`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple Student’s *t* distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'T'`) and parameters.

Parameters

The Student’s *t* distribution uses the following parameter.

Parameter	Description	Support
nu (ν)	Degrees of freedom	ν = 1, 2, 3,...

Probability Density Function

The pdf of the Student's *t* distribution is

$$y = f(x \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}},$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result y is the probability of observing a particular value of x from the Student’s *t* distribution with ν degrees of freedom.

For an example, see “Compute and Plot Student's t Distribution pdf” on page B-150.

Cumulative Distribution Function

The cdf of the Student's t distribution is

$$p = F(x|\nu) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} dt,$$

where ν is the degrees of freedom and $\Gamma(\cdot)$ is the Gamma function. The result p is the probability that a single observation from the t distribution with ν degrees of freedom falls in the interval $[-\infty, x]$.

For an example, see "Compute and Plot Student's t Distribution cdf" on page B-151.

Inverse Cumulative Distribution Function

The t inverse function is defined in terms of the Student's t cdf as

$$x = F^{-1}(p|\nu) = \left\{x : F(x|\nu) = p\right\},$$

where

$$p = F(x|\nu) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}} dt,$$

ν is the degrees of freedom, and $\Gamma(\cdot)$ is the Gamma function. The result x is the solution of the integral equation where you supply the probability p.

For an example, see "Compute Student's t icdf" on page B-152.

Descriptive Statistics

The mean of the Student's t distribution is $\mu = 0$ for degrees of freedom ν greater than 1. If ν equals 1, then the mean is undefined.

The variance of the Student's t distribution is $\frac{\nu}{\nu-2}$ for degrees of freedom ν greater than 2. If ν is less than or equal to 2, then the variance is undefined.

Examples

Compute and Plot Student's t Distribution pdf

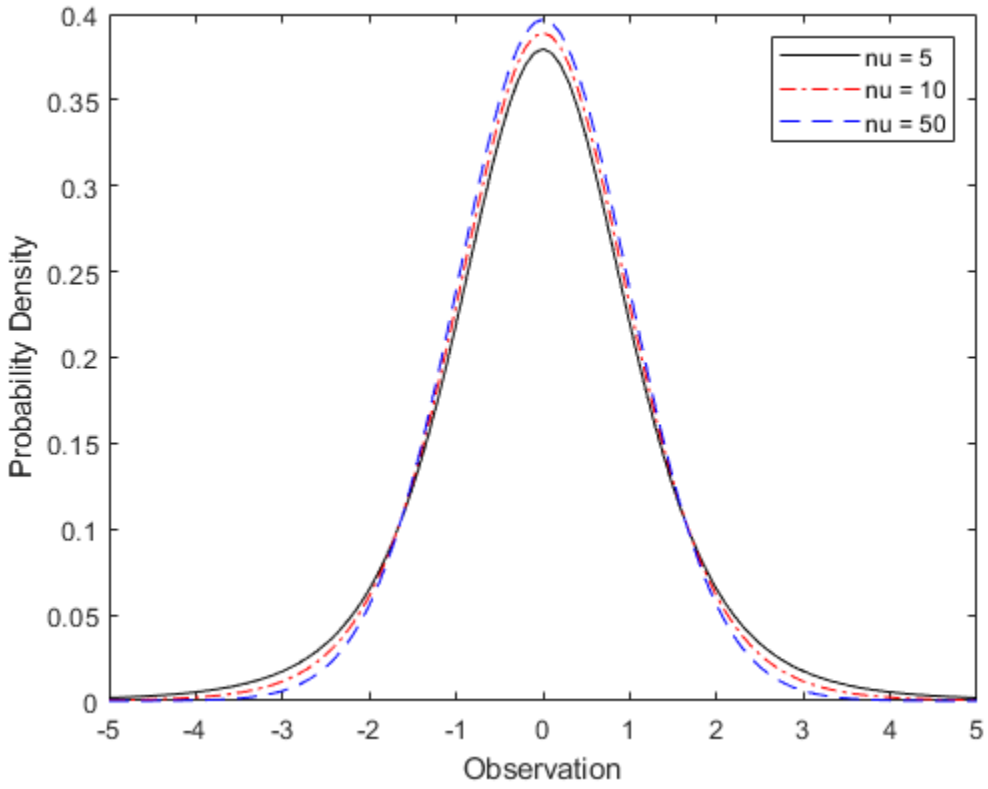
Compute the pdf of a Student's t distribution with degrees of freedom equal to `5`, `10`, and `50`.

```
x = [-5:.1:5];
y1 = tpdf(x,5);
```

```
y2 = tpdf(x,10);
y3 = tpdf(x,50);
```

Plot the pdf for all three choices nu on the same axis.

```
figure;
plot(x,y1,'Color','black','LineStyle','-')
hold on
plot(x,y2,'Color','red','LineStyle','-.')
plot(x,y3,'Color','blue','LineStyle','--')
xlabel('Observation')
ylabel('Probability Density')
legend({'nu = 5','nu = 10','nu = 50'})
hold off
```



Compute and Plot Student's t Distribution cdf

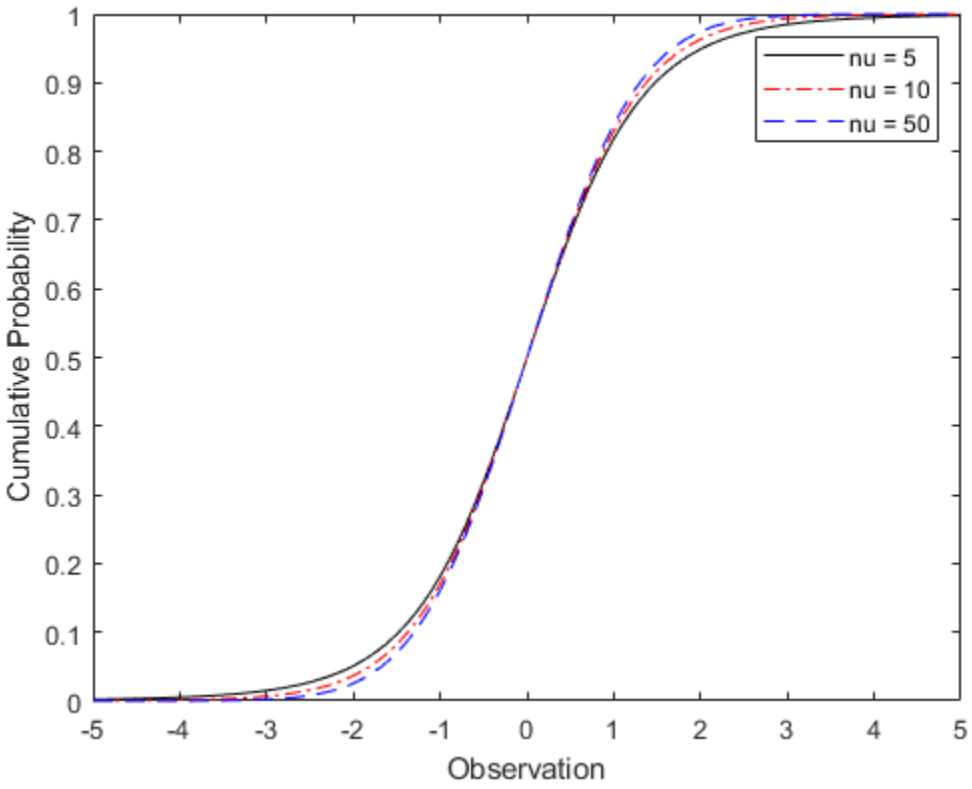
Compute the cdf of a Student's *t* distribution with degrees of freedom equal to 5, 10, and 50.

```
x = [-5:.1:5];
y1 = tcdf(x,5);
y2 = tcdf(x,10);
y3 = tcdf(x,50);
```

Plot the cdf for all three choices of nu on the same axis.

```
figure;
plot(x,y1,'Color','black','LineStyle','-')
```

```
hold on
plot(x,y2,'Color','red','LineStyle','-.')
plot(x,y3,'Color','blue','LineStyle','--')
xlabel('Observation')
ylabel('Cumulative Probability')
legend({'nu = 5','nu = 10','nu = 50'})
hold off
```



Compute Student's t icdf

Find the 95th percentile of the Student's *t* distribution with `50` degrees of freedom.

```
p = .95;
nu = 50;
x = tinv(p,nu)
```

```
x = 1.6759
```

Compare Student's t and Normal Distribution pdfs

The Student's *t* distribution is a family of curves depending on a single parameter ν (the degrees of freedom). As the degrees of freedom ν approach infinity, the *t* distribution approaches the standard normal distribution.

Compute the pdfs for the Student's *t* distribution with the parameter `nu = 5` and the Student's *t* distribution with the parameter `nu = 15`.

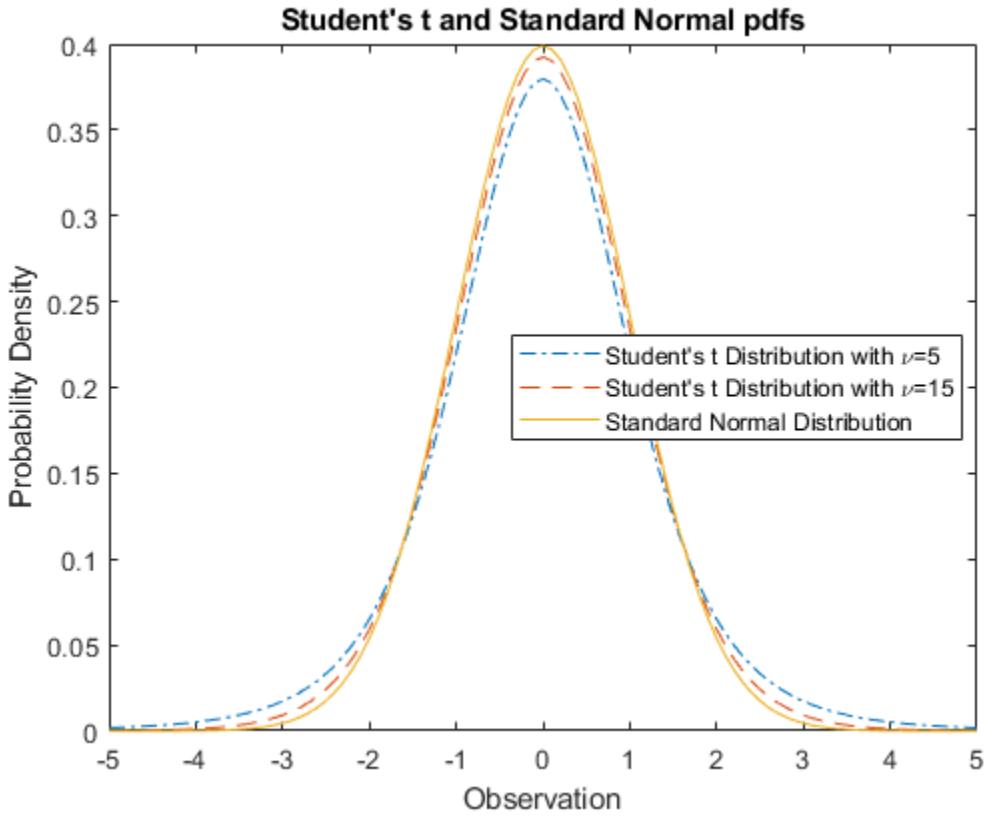
```
x = [-5:0.1:5];
y1 = tpdf(x,5);
y2 = tpdf(x,15);
```

Compute the pdf for a standard normal distribution.

```
z = normpdf(x,0,1);
```

Plot the Student's *t* pdfs and the standard normal pdf on the same figure.

```
plot(x,y1,'-.',x,y2,'--',x,z,'-')
legend('Student''s t Distribution with \nu=5', ...
    'Student''s t Distribution with \nu=15', ...
    'Standard Normal Distribution','Location','best')
xlabel('Observation')
ylabel('Probability Density')
title('Student''s t and Standard Normal pdfs')
```



The standard normal pdf has shorter tails than the Student's *t* pdfs.

Related Distributions

- "Beta Distribution" on page B-6 — The beta distribution is a two-parameter continuous distribution that has parameters a (first shape parameter) and b (second shape parameter). If Y has a Student's t distribution with ν degrees of freedom, then $X = \frac{1}{2} + \frac{1}{2}\frac{Y}{\sqrt{\nu + Y^2}}$ has beta

distribution with the shape parameters $a = \nu/2$ and $b = \nu/2$. This relationship is used to compute values of the t cdf and inverse functions, and to generate t distributed random numbers.

- Cauchy Distribution — The Cauchy distribution is a two-parameter continuous distribution with the parameters γ (scale) and δ (location). It is a special case of the "Stable Distribution" on page B-140 with the shape parameters $\alpha = 1$ and $\beta = 0$. The standard Cauchy distribution (unit scale and location zero) is the Student's t distribution with degrees of freedom ν equal to 1. The standard Cauchy distribution has an undefined mean and variance.

  For an example, see "Generate Cauchy Random Numbers Using Student's t" on page 5-107.

- "Chi-Square Distribution" on page B-28 — The chi-square distribution is a one-parameter continuous distribution that has the parameter ν (degrees of freedom). If Z has a standard normal distribution and χ^2 has a chi-square distribution with degrees of freedom ν, then $t = \frac{Z}{\sqrt{\chi^2/\nu}}$ has a Student's t distribution with degrees of freedom ν.

- "Noncentral t Distribution" on page B-117 — The noncentral t distribution is a two-parameter continuous distribution that generalizes the Student's t distribution and has the parameters ν (degrees of freedom) and δ (noncentrality). Setting $\delta = 0$ yields the Student's t distribution.

- "Normal Distribution" on page B-119 — The normal distribution is a two-parameter continuous distribution with the parameters μ (mean) and σ (standard deviation).

  As the degrees of freedom ν approach infinity, the Student's t distribution approaches the standard normal distribution (zero mean and unit standard deviation).

  For an example, see "Compare Student's t and Normal Distribution pdfs" on page B-152

  If x is a random sample of size n from a normal distribution with mean μ, then the statistic $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, where $\bar{x}$ is the sample mean and s is the sample standard deviation, has a Student's t distribution with n —1 degrees of freedom.

  For an example, see "Compute Student's t Distribution cdf" on page 33-6081.

- "t Location-Scale Distribution" on page B-156 — The t location-scale distribution is a three-parameter continuous distribution with the parameters μ (mean), σ (scale), and ν (shape). If x has a t location-scale distribution with the parameters μ, σ, and ν, then $\frac{x - \mu}{\sigma}$ has a Student's t distribution with ν degrees of freedom.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

[4] Kreyszig, Erwin. *Introductory Mathematical Statistics: Principles and Methods*. New York: Wiley, 1970.

See Also

`tcdf` | `tinv` | `tpdf` | `trnd` | `tstat` | `ttest` | `ttest2`

More About

- “Generate Cauchy Random Numbers Using Student's t” on page 5-107
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

t Location-Scale Distribution

In this section...

Overview

The *t* location-scale distribution is useful for modeling data distributions with heavier tails (more prone to outliers) than the normal distribution. It approaches the normal distribution as ν approaches infinity, and smaller values of ν yield heavier tails.

Parameters

The *t* location-scale distribution uses the following parameters.

Parameter	Description	Support
μ	Location parameter	$-\infty < \mu < \infty$
σ	Scale parameter	$\sigma > 0$
ν	Shape parameter	$\nu > 0$

To estimate distribution parameters, use `mle`. Alternatively, fit a `tLocationScaleDistribution` object to data using `fitdist` or the **Distribution Fitter** app.

Probability Density Function

The probability density function (pdf) of the *t* location-scale distribution is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left[\frac{\nu+\left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)}$$

where $\Gamma(\bullet)$ is the gamma function, μ is the location parameter, σ is the scale parameter, and ν is the shape parameter .

To compute the probability density function, use `pdf` and specify `'tLocationScale'`. Alternatively, you can create a `tLocationScaleDistribution` object using `fitdist` or `makedist`, then use the `pdf` to work with the object.

Cumulative Distribution Function

To compute the probability density function, use `cdf` and specify `'tLocationScale'`. Alternatively, you can create a `tLocationScaleDistribution` object using `fitdist` or `makedist`, then use the `cdf` to work with the object.

Descriptive Statistics

The mean of the *t* location-scale distribution is

$$\text{mean} = \mu,$$

where μ is the location parameter. The mean is only defined for shape parameter values $\nu > 1$. For other values of ν, the mean is undefined.

The variance of the *t* location-scale distribution is

$$\text{var} = \sigma^2 \frac{\nu}{\nu - 2},$$

where μ is the location parameter and ν is the shape parameter. The variance is only defined for values of $\nu > 2$. For other values of ν, the variance is undefined.

To compute the mean and variance, create a `tLocationScaleDistribution` object using `fitdist` or `makedist`. You can also use the **Distribution Fitter** app.

Relationship to Other Distributions

If x has a *t* location-scale distribution, with parameters μ, σ, and ν, then

$$\frac{x - \mu}{\sigma}$$

has a Student's *t* distribution with ν degrees of freedom.

See Also

`tLocationScaleDistribution`

More About

- “Represent Cauchy Distribution Using t Location-Scale” on page 5-104
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Triangular Distribution

In this section...

Overview

The triangular distribution provides a simplistic representation of the probability distribution when limited sample data is available. Its parameters are the minimum, maximum, and peak of the data. Common applications include business and economic simulations, project management planning, natural phenomena modeling, and audio dithering.

Parameters

The triangular distribution uses the following parameters.

Parameter	Description	Constraints
a	Lower limit	$a \le b$
b	Peak location	$a \le b \le c$
c	Upper limit	$c \ge b$

Parameter Estimation

Typically, you estimate triangular distribution parameters using subjectively reasonable values based on the sample data. You can estimate the lower and upper limit parameters a and c using the minimum and maximum values of the sample data, respectively. You can estimate the peak location parameter b using the sample mean, median, mode, or any other subjectively reasonable estimate of the population mode.

Probability Density Function

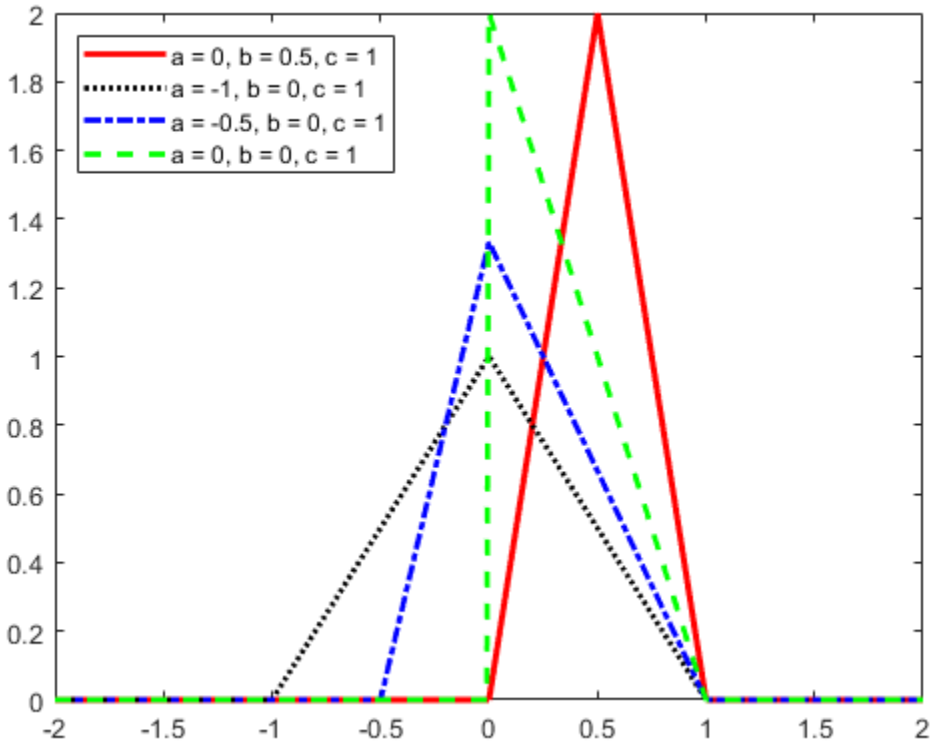
The probability density function (pdf) of the triangular distribution is

$$f(x \mid a,b,c) = \begin{cases} \dfrac{2(x-a)}{(c-a)(b-a)} & ; \quad a \le x \le b \\ \dfrac{2(c-x)}{(c-a)(c-b)} & ; \quad b < x \le c \\ 0 & ; \quad x < a \,,\, x > c \end{cases} .$$

This plot shows how changing the value of the parameters a, b, and c alters the shape of the pdf.

```
% Create four distribution objects with different parameters
pd1 = makedist('Triangular');
pd2 = makedist('Triangular','a',-1,'b',0,'c',1);
```

```
pd3 = makedist('Triangular','a',-.5,'b',0,'c',1);
pd4 = makedist('Triangular','a',0,'b',0,'c',1);

% Compute the pdfs
x = -2:.01:2;
pdf1 = pdf(pd1,x);
pdf2 = pdf(pd2,x);
pdf3 = pdf(pd3,x);
pdf4 = pdf(pd4,x);

% Plot the pdfs
figure;
plot(x,pdf1,'r','LineWidth',2)
hold on;
plot(x,pdf2,'k:','LineWidth',2);
plot(x,pdf3,'b-.','LineWidth',2);
plot(x,pdf4,'g--','LineWidth',2);
legend({'a = 0, b = 0.5, c = 1','a = -1, b = 0, c = 1',...
    'a = -0.5, b = 0, c = 1','a = 0, b = 0, c = 1'},'Location','NW');
hold off;
```



As the distance between *a* and *c* increases, the density at any particular value within the distribution boundaries decreases. Because the density function integrates to 1, the height of the pdf plot decreases as its width increases. The location of the peak parameter *b* determines whether the pdf skews right or left, or if it is symmetrical.

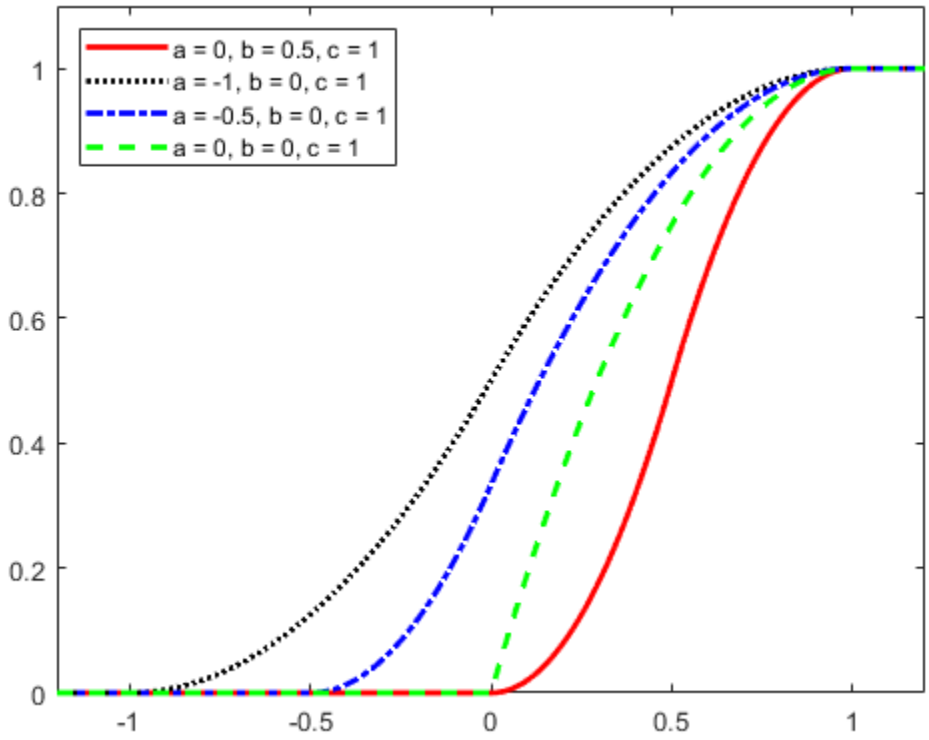
Cumulative Distribution Function

The cumulative distribution function (cdf) of the triangular distribution is

$$F(x \mid a,b,c) = \begin{cases} 0 & , \quad x < a \\ \dfrac{(x-a)^2}{(c-a)(b-a)} & , \quad a \le x \le b \\ 1 - \dfrac{(c-x)^2}{(c-a)(c-b)} & , \quad b < x \le c \\ 1 & , \quad x > c \end{cases}.$$

This plot shows how changing the value of the parameters a, b, and c alters the shape of the cdf.

```
% Create four distribution objects with different parameters
pd1 = makedist('Triangular');
pd2 = makedist('Triangular','a',-1,'b',0,'c',1);
pd3 = makedist('Triangular','a',-.5,'b',0,'c',1);
pd4 = makedist('Triangular','a',0,'b',0,'c',1);

% Compute the cdfs
x = -1.2:.01:1.2;
cdf1 = cdf(pd1,x);
cdf2 = cdf(pd2,x);
cdf3 = cdf(pd3,x);
cdf4 = cdf(pd4,x);

% Plot the cdfs
figure;
plot(x,cdf1,'r','LineWidth',2)
xlim([-1.2 1.2]);
ylim([0 1.1]);hold on;
plot(x,cdf2,'k:','LineWidth',2);
plot(x,cdf3,'b-.','LineWidth',2);
plot(x,cdf4,'g--','LineWidth',2);
legend({'a = 0, b = 0.5, c = 1','a = -1, b = 0, c = 1',...
    'a = -0.5, b = 0, c = 1','a = 0, b = 0, c = 1'},'Location','NW');
hold off;
```



Descriptive Statistics

The mean and variance of the triangular distribution are related to the parameters a, b, and c.

The mean is

$$\text{mean} = \left(\frac{a+b+c}{3}\right).$$

The variance is

$$\text{var} = \left(\frac{a^2+b^2+c^2-ab-ac-bc}{18}\right).$$

See Also

`TriangularDistribution`

More About

- “Generate Random Numbers Using the Triangular Distribution” on page 5-47
- “Nonparametric and Empirical Probability Distributions” on page 5-30
- “Working with Probability Distributions” on page 5-3

- “Supported Distributions” on page 5-14

Uniform Distribution (Continuous)

In this section...

Overview

The uniform distribution (also called the rectangular distribution) is a two-parameter family of curves that is notable because it has a constant probability distribution function (pdf) between its two bounding parameters. This distribution is appropriate for representing the distribution of round-off errors in values tabulated to a particular number of decimal places. The uniform distribution is used in random number generating techniques such as the inversion method.

Statistics and Machine Learning Toolbox offers several ways to work with the uniform distribution.

- Create a probability distribution object `UniformDistribution` by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Use distribution-specific functions (`unifcdf`, `unifpdf`, `unifinv`, `unifit`, `unifstat`, `unifrnd`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple uniform distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Uniform'`) and parameters.

Parameters

The uniform distribution uses the following parameters.

Parameter	Description	Support
a	Lower endpoint	$-\infty < a < b$
b	Upper endpoint	$a < b < \infty$

The standard uniform distribution has $a = 0$ and $b = 1$.

Parameter Estimation

The *maximum likelihood estimates* (MLEs) are the parameter estimates that maximize the likelihood function. The maximum likelihood estimators of a and b for the uniform distribution are the sample minimum and maximum, respectively.

To fit the uniform distribution to data and find parameter estimates, use `unifit` or `mle`.

Probability Density Function

The pdf of the uniform distribution is

$$f(x\,|\,a,b) = \begin{cases} \left(\frac{1}{b-a}\right) & ; \quad a \le x \le b \\ 0 & ; \quad otherwise \end{cases} .$$

The pdf is constant between a and b.

For an example, see “Compute Continuous Uniform Distribution pdf” on page B-164.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the uniform distribution is

$$F(x\,|\,a,b) = \begin{cases} 0 & ; \quad x < a \\ \frac{x-a}{b-a} & ; \quad a \le x < b \\ 1 & ; \quad x \ge b \end{cases} .$$

The result p is the probability that a single observation from a uniform distribution with parameters a and b falls in the interval $[a\ x]$.

For an example, see “Compute Continuous Uniform Distribution cdf” on page B-165.

Descriptive Statistics

The mean of the uniform distribution is $\mu = \frac{1}{2}(a + b)$.

The variance of the uniform distribution is $\sigma^2 = \frac{1}{12}(b - a)^2$.

Random Number Generation

You can use the standard uniform distribution to generate random numbers for any other continuous distribution by the inversion method. The inversion method relies on the principle that continuous cumulative distribution functions (cdfs) range uniformly over the open interval (0, 1). If u is a uniform random number on (0, 1), then $x = \mathsf{F}^{-1}(u)$ generates a random number x from the continuous distribution with the specified cdf F.

For an example, see “Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101.

Examples

Compute Continuous Uniform Distribution pdf

Create three uniform distribution objects with different parameters.

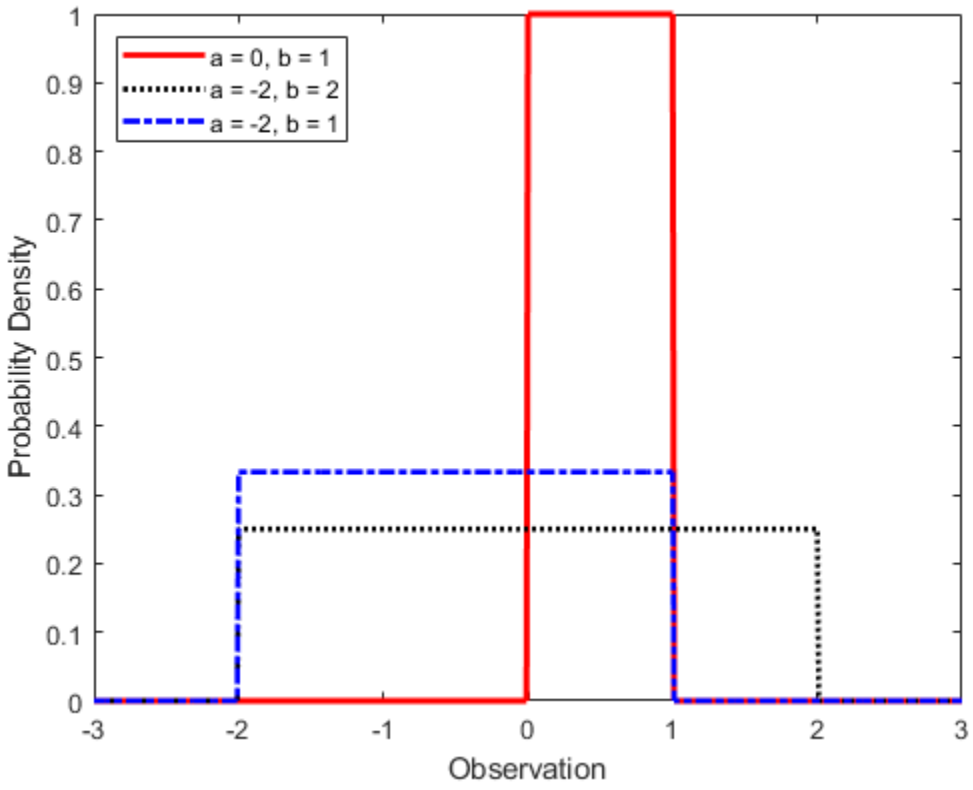
```
pd1 = makedist('Uniform');                       % Standard uniform distribution
pd2 = makedist('Uniform','lower',-2,'upper',2); % Uniform distribution with a = -2 and b = 2
pd3 = makedist('Uniform','lower',-2,'upper',1); % Uniform distribution with a = -2 and b = 1
```

Compute the pdfs for the three uniform distributions.

```
x = -3:.01:3;
pdf1 = pdf(pd1,x);
pdf2 = pdf(pd2,x);
pdf3 = pdf(pd3,x);
```

Plot the pdfs on the same axis.

```
figure;
plot(x,pdf1,'r','LineWidth',2);
hold on;
plot(x,pdf2,'k:','LineWidth',2);
plot(x,pdf3,'b-.','LineWidth',2);
legend({'a = 0, b = 1','a = -2, b = 2','a = -2, b = 1'},'Location','northwest');
xlabel('Observation')
ylabel('Probability Density')
hold off;
```



As the width of the interval `(a,b)` increases, the height of each pdf decreases.

Compute Continuous Uniform Distribution cdf

Create three uniform distribution objects with different parameters.

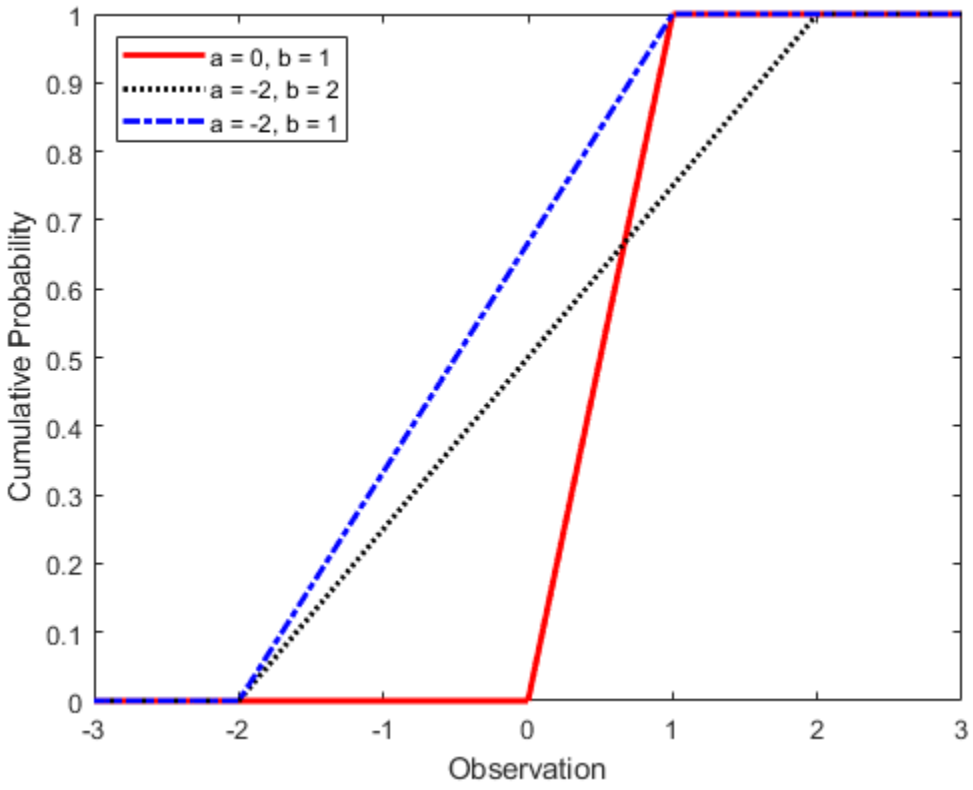
```
pd1 = makedist('Uniform');                        % Standard uniform distribution
pd2 = makedist('Uniform','lower',-2,'upper',2); % Uniform distribution with a = -2 and b = 2
pd3 = makedist('Uniform','lower',-2,'upper',1); % Uniform distribution with a = -2 and b = 1
```

Compute the cdfs for the three uniform distributions.

```
x = -3:.01:3;
cdf1 = cdf(pd1,x);
cdf2 = cdf(pd2,x);
cdf3 = cdf(pd3,x);
```

Plot the cdfs on the same axis.

```
figure;
plot(x,cdf1,'r','LineWidth',2);
hold on;
plot(x,cdf2,'k:','LineWidth',2);
plot(x,cdf3,'b-.','LineWidth',2);
legend({'a = 0, b = 1','a = -2, b = 2','a = -2, b = 1'},'Location','NW');
xlabel('Observation')
ylabel('Cumulative Probability')
hold off;
```



As the width of the interval `(a,b)` increases, the slope of each cdf decreases.

Related Distributions

- “Beta Distribution” on page B-6 — The beta distribution is a two-parameter continuous distribution that has parameters *a* (first shape parameter) and *b* (second shape parameter). The standard uniform distribution is equal to the beta distribution with unit parameters.
- “Triangular Distribution” on page B-158 — The triangular distribution is a three-parameter continuous distribution that has parameters *a* (lower limit), *b* (peak), and *c* (upper limit). The sum of two random variables with a standard uniform distribution has a triangular distribution with $a = 0$, $b = 1$, and $c = 0$.

References

[1] Abramowitz, Milton, and Irene A. Stegun, eds. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. 9. Dover print.; [Nachdr. der Ausg. von 1972]. Dover Books on Mathematics. New York, NY: Dover Publ, 2013.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

See Also

`UniformDistribution` | `fitdist` | `makedist` | `unifcdf` | `unifinv` | `unifit` | `unifpdf` | `unifrnd` | `unifstat`

More About

- “Generate Random Numbers Using Uniform Distribution Inversion” on page 5-101
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Uniform Distribution (Discrete)

In this section...

Definition

The discrete uniform pdf is

$$y = f(x|N) = \frac{1}{N} I_{(1, \ldots, N)}(x)$$

Background

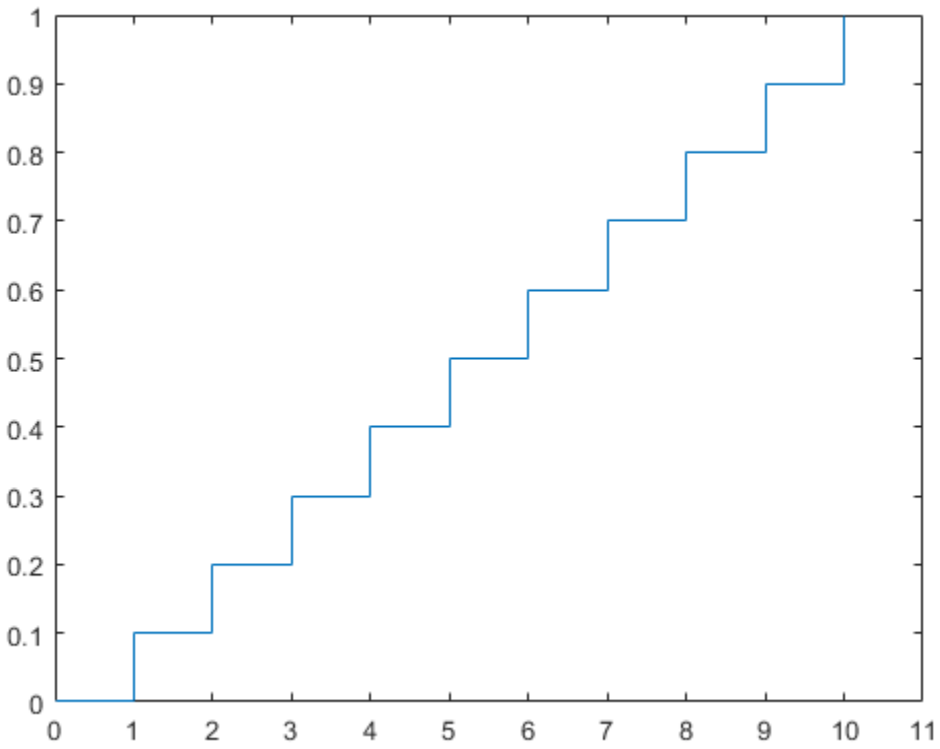
The discrete uniform distribution is a simple distribution that puts equal weight on the integers from one to N.

Examples

Plot a Discrete Uniform Distribution cdf

As for all discrete distributions, the cdf is a step function. The plot shows the discrete uniform cdf for $N = 10$.

```
x = 0:10;
y = unidcdf(x,10);
figure;
stairs(x,y)
h = gca;
h.XLim = [0 11];
```



Generate Discrete Uniform Random Numbers

Pick a random sample of 10 from a list of 553 items.

```
rng default;  % for reproducibility
numbers = unidrnd(553,1,10)
```

```
numbers = 1×10

   451   501    71   506   350    54   155   303   530   534
```

See Also

`random` | `unidcdf` | `unidinv` | `unidpdf` | `unidrnd` | `unidstat`

More About

- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Weibull Distribution

In this section...

Overview

The Weibull distribution is a two-parameter family of curves. This distribution is named for Waloddi Weibull, who offered it as an appropriate analytical tool for modeling the breaking strength of materials. Current usage also includes reliability and lifetime modeling. The Weibull distribution is more flexible than the exponential distribution for these purposes, because the exponential distribution has a constant hazard function.

Statistics and Machine Learning Toolbox offers several ways to work with the Weibull distribution.

- Create a probability distribution object `WeibullDistribution` by fitting a probability distribution to sample data (`fitdist`) or by specifying parameter values (`makedist`). Then, use object functions to evaluate the distribution, generate random numbers, and so on.
- Work with the Weibull distribution interactively by using the **Distribution Fitter** app. You can export an object from the app and use the object functions.
- Use distribution-specific functions (`wblcdf`, `wblpdf`, `wblinv`, `wbllike`, `wblstat`, `wblfit`, `wblrnd`, `wblplot`) with specified distribution parameters. The distribution-specific functions can accept parameters of multiple Weibull distributions.
- Use generic distribution functions (`cdf`, `icdf`, `pdf`, `random`) with a specified distribution name (`'Weibull'`) and parameters.

Parameters

The Weibull distribution uses these parameters.

Parameter	Description	Support
`a`	Scale	$a > 0$
`b`	Shape	$b > 0$

The standard Weibull distribution has unit scale.

Parameter Estimation

The *likelihood function* is the probability density function (pdf) viewed as a function of the parameters. The *maximum likelihood estimates* (MLEs) are the parameter estimates that maximize

the likelihood function for fixed values of `x`. The maximum likelihood estimators of a and b for the Weibull distribution are the solution of the simultaneous equations

$$\widehat{a} = \left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i^{\widehat{b}}\right]^{\frac{1}{\widehat{b}}}$$

$$\widehat{b} = \frac{n}{\left(\frac{1}{\widehat{a}}\right)\sum_{i=1}^{n} x_i^{\widehat{b}}\log x_i - \sum_{i=1}^{n}\log x_i}.$$

$\hat{a}$ and $\widehat{b}$ are unbiased estimators of the parameters a and b.

To fit the Weibull distribution to data and find parameter estimates, use `wblfit`, `fitdist`, or `mle`. Unlike `wblfit` and `mle`, which return parameter estimates, `fitdist` returns the fitted probability distribution object `WeibullDistribution`. The object properties `a` and `b` store the parameter estimates.

For an example, see “Fit Weibull Distribution to Data and Estimate Parameters” on page B-172.

Probability Density Function

The pdf of the Weibull distribution is

$$f(x\,|\,a,b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-(x/a)^b}.$$

For an example, see “Compute Weibull Distribution pdf” on page B-172.

Cumulative Distribution Function

The cumulative distribution function (cdf) of the Weibull distribution is

$$p = F(x\,|\,a,b) = \int_0^x ba^{-b}t^{b-1}e^{-\left(\frac{t}{a}\right)^b}dt = 1 - e^{-\left(\frac{x}{a}\right)^b}.$$

The result p is the probability that a single observation from a Weibull distribution with parameters a and b falls in the interval $[0\ x]$.

For an example, see “Compute Weibull Distribution cdf” on page B-173.

Inverse Cumulative Distribution Function

The inverse cdf of the Weibull distribution is

$$x = F^{-1}(p\,|\,a,b) = -a[\ln(1-p)]^{1/b}.$$

The result x is the value where an observation from a Weibull distribution with parameters a and b falls in the range $[0\ x]$ with probability p.

Hazard Function

The hazard function (instantaneous failure rate) is the ratio of the pdf and the complement of the cdf. If $f(t)$ and $F(t)$ are the pdf and cdf of a distribution, then the hazard rate is $h(t) = \frac{f(t)}{1 - F(t)}$. Substituting the pdf and cdf of the exponential distribution for $f(t)$ and $F(t)$ above yields the function $\frac{b}{a^b} x^{b-1}$.

For an example, see “Compare Exponential and Weibull Distribution Hazard Functions” on page B-174.

Examples

Fit Weibull Distribution to Data and Estimate Parameters

Simulate the tensile strength data of a thin filament using the Weibull distribution with the scale parameter value `0.5` and the shape parameter value `2`.

```
rng('default');                  % For reproducibility
strength = wblrnd(0.5,2,100,1);  % Simulated strengths
```

Compute the MLEs and confidence intervals for the Weibull distribution parameters.

```
[param,ci] = wblfit(strength)
```

```
param = 1×2

    0.4768    1.9622


ci = 2×2

    0.4291    1.6821
    0.5298    2.2890
```

The estimated scale parameter is `0.4768`, with the 95% confidence interval `(0.4291,0.5298)`.

The estimated shape parameter is `1.9622`, with the 95% confidence interval `(1.6821,2.2890)`.

The default confidence interval for each parameter contains the true value.

Compute Weibull Distribution pdf

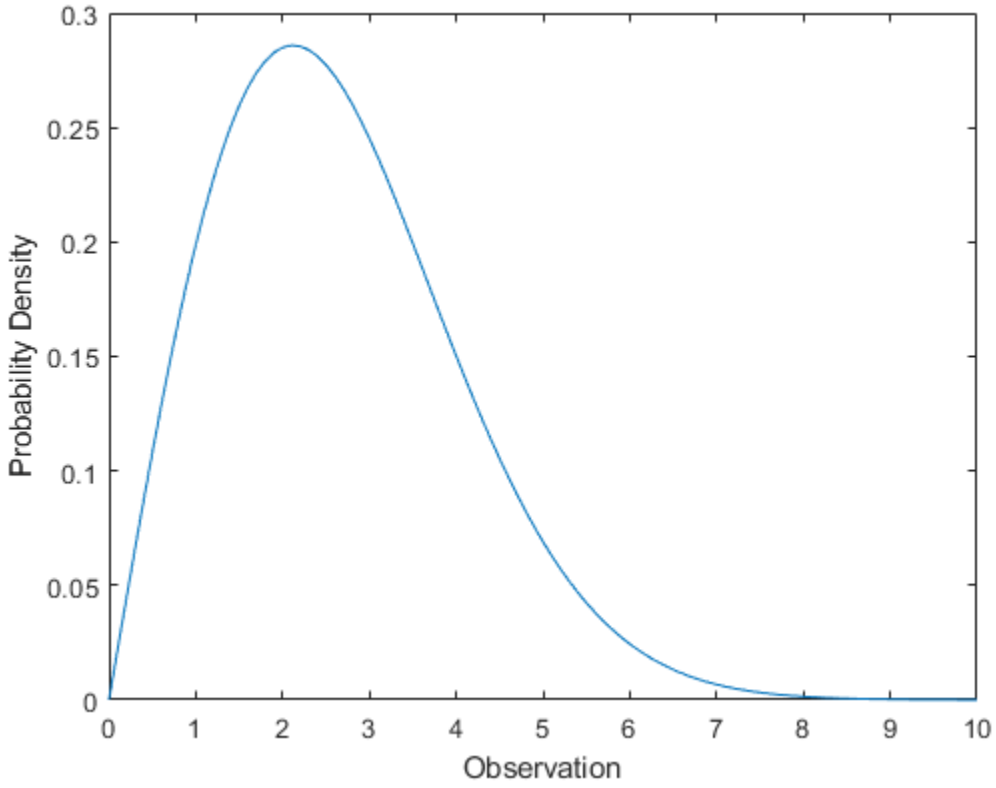
Compute the pdf of the Weibull distribution with the scale parameter value `3` and the shape parameter value `2`.

```
x = 0:0.1:10;
y = wblpdf(x,3,2);
```

Plot the pdf.

```
figure;
plot(x,y)
```

```
xlabel('Observation')
ylabel('Probability Density')
```



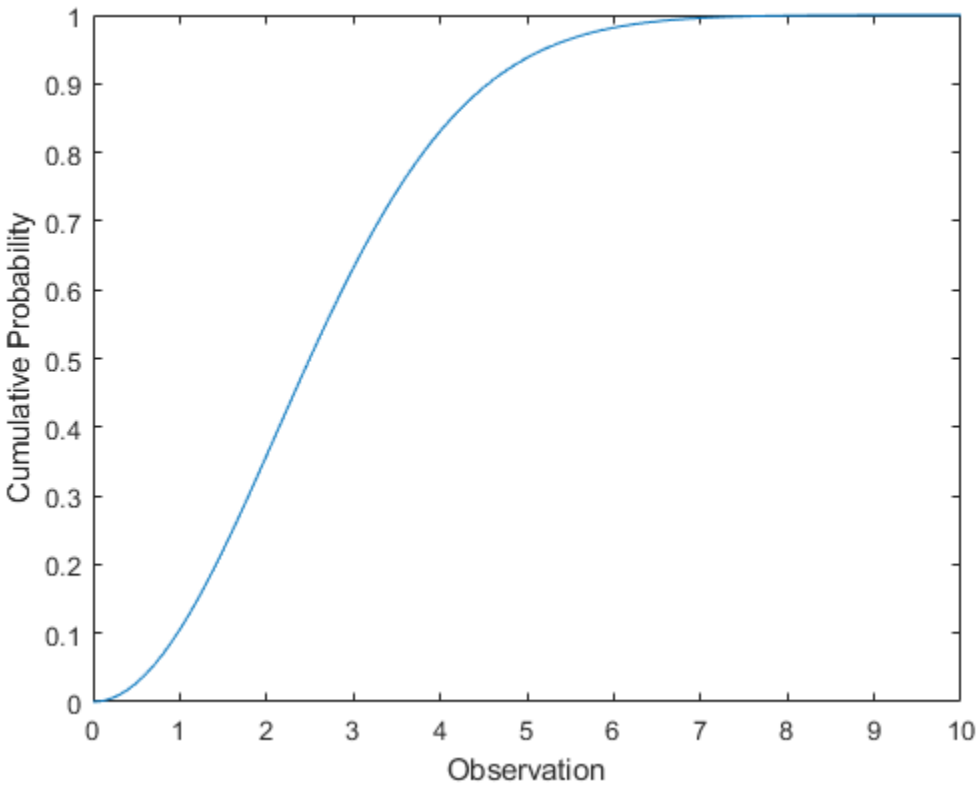
Compute Weibull Distribution cdf

Compute the cdf of the Weibull distribution with the scale parameter value 3 and the shape parameter value 2.

```
x = 0:0.1:10;
y = wblcdf(x,3,2);
```

Plot the cdf.

```
figure;
plot(x,y)
xlabel('Observation')
ylabel('Cumulative Probability')
```



Compare Exponential and Weibull Distribution Hazard Functions

The exponential distribution has a constant hazard function, which is not generally the case for the Weibull distribution. In this example, the Weibull hazard rate increases with age (a reasonable assumption).

Compute the hazard function for the Weibull distribution with the scale parameter value 1 and the shape parameter value 2.

```
t = 0:0.1:4.5;
h1 = wblpdf(t,1,2)./(1-wblcdf(t,1,2));
```

Compute the mean of the Weibull distribution with scale parameter value 1 and shape parameter value 2.

```
mu = wblstat(1,2)
```

```
mu = 0.8862
```

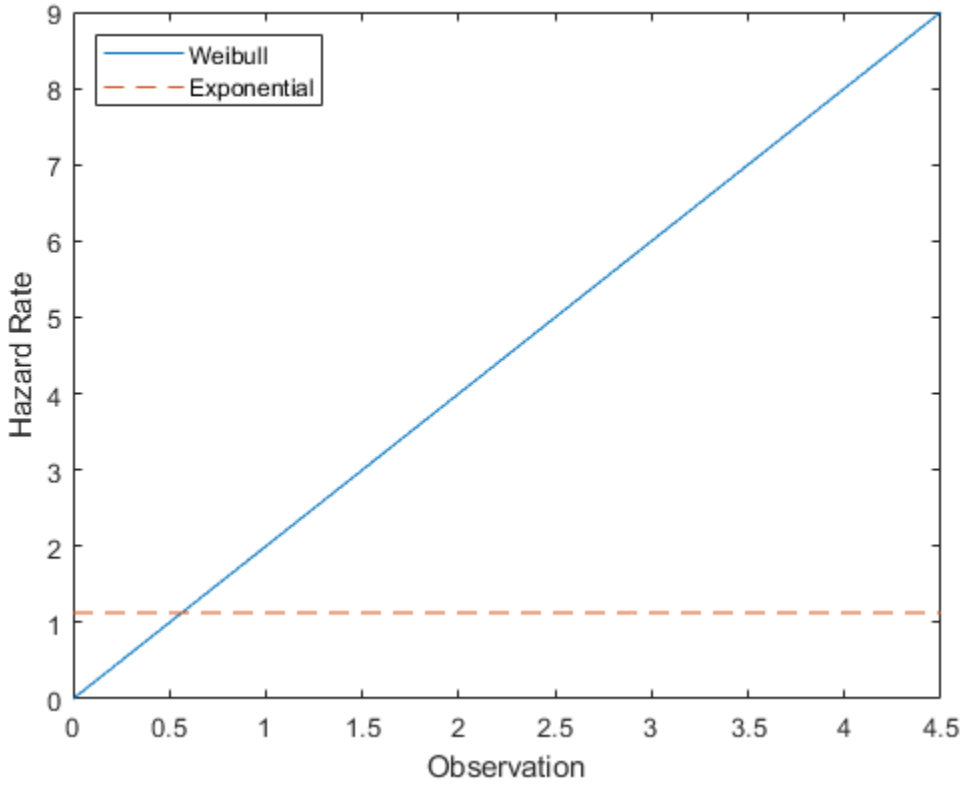
Compute the hazard function for the exponential distribution with mean `mu`.

```
h2 = exppdf(t,mu)./(1-expcdf(t,mu));
```

Plot both hazard functions on the same axis.

```
plot(t,h1,'-',t,h2,'--')
xlabel('Observation')
```

```
ylabel('Hazard Rate')
legend('Weibull','Exponential','location','northwest')
```



Estimate Parameters of Three-Parameter Weibull Distribution

Statistics and Machine Learning Toolbox™ uses a two-parameter Weibull distribution with a scale parameter a and a shape parameter b. The Weibull distribution can take one more parameter, a location parameter c. The pdf becomes

$$f(x|a,b,c) = \begin{cases} \frac{b}{a}\left(\frac{x-c}{a}\right)^{b-1}\exp\left(-\left(\frac{x-c}{a}\right)^{b}\right) & \text{if } x > c, \\ 0 & \text{if } x \le c, \end{cases}$$

where a and b are positive values, and c is a real value.

Generate sample data of size 1000 from a three-parameter Weibull distribution with the scale parameter 1, shape parameter 1, and location parameter 10.

```
rng('default') % For reproducibility
data = wblrnd(1,1,[1000,1]) + 10;
```

Define a probability density function for a three-parameter Weibull distribution.

```
custompdf = @(x,a,b,c) (x>c).*(b/a).*(((x-c)/a).^(b-1)).*exp(-((x-c)/a).^b);
```

`mle` estimates the parameters from data. If `mle` does not converge with default statistics options, modify them by using the name-value pair argument `'Options'`.

Create a statistics options structure `opt` by using the function `statset`.

```
opt = statset('MaxIter',1e5,'MaxFunEvals',1e5,'FunValCheck','off');
```

The option `opt` includes the following options:

- `'MaxIter',1e5` — Increase the maximum number of iterations to `1e5`.
- `'MaxFunEvals',1e5` — Increase the maximum number of object function evaluations to `1e5`.
- `'FunValCheck','off'` — Turn off checking for invalid object function values.

For a distribution with a region that has zero probability density, `mle` might try some parameters that have zero density, and it will fail to estimate parameters. To avoid this problem, you can turn off the option that checks for invalid function values by using `'FunValCheck','off'`.

Use `mle` to estimate the parameters. Note that the Weibull probability density function is positive only for $x > c$. This constraint also implies that a location parameter c is smaller than the minimum of the sample data. Include the lower and upper bounds of parameters by using the name-value pair arguments `'LowerBound'` and `'UpperBound'`, respectively.

```
params = mle(data,'pdf',custompdf,'start',[5 5 5],...
     'Options',opt,'LowerBound',[0 0 -Inf],'UpperBound',[Inf Inf min(data)])
```

```
params = 1×3

    1.0258    1.0618   10.0004
```

If the scale parameter b is smaller than 1, the probability density of the Weibull distribution approaches infinity as x goes to c, where c is the location parameter. The maximum of the likelihood function is infinite. `mle` may find satisfactory estimates in some cases, but the global maximum is degenerate when $b < 1$.

Related Distributions

- "Exponential Distribution" on page B-33 — The exponential distribution is a one-parameter continuous distribution that has parameter μ (mean). This distribution is also used for lifetime modeling. When $b = 1$, the Weibull distribution is equal to the exponential distribution with mean $\mu = a$.
- "Extreme Value Distribution" on page B-40 — The extreme value distribution is a two-parameter continuous distribution with parameters μ (location) and σ (scale). If X has a Weibull distribution with parameters a and b, then $\log X$ has an extreme value distribution with parameters $\mu = \log a$ and $\sigma = 1/b$. This relationship is used to fit data to a Weibull distribution.
- "Rayleigh Distribution" on page B-137 — The Rayleigh distribution is a one-parameter continuous distribution that has parameter b (scale). If A and B are the parameters of the Weibull distribution, then the Rayleigh distribution with parameter b is equivalent to the Weibull distribution with parameters$A = \sqrt{2}b$ and $B = 2$.
- Three-Parameter Weibull Distribution — The three-parameter Weibull distribution adds a location parameter that is zero in the two-parameter case. If X has a two-parameter Weibull distribution, then $Y = X + c$ has a three-parameter Weibull distribution with the added location parameter c.

For an example, see “Estimate Parameters of Three-Parameter Weibull Distribution” on page B-175.

References

[1] Crowder, Martin J., ed. *Statistical Analysis of Reliability Data*. Reprinted. London: Chapman & Hall, 1995.

[2] Devroye, Luc. *Non-Uniform Random Variate Generation*. New York, NY: Springer New York, 1986. https://doi.org/10.1007/978-1-4613-8643-8

[3] Evans, Merran, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. 2nd ed. New York: J. Wiley, 1993.

[4] Lawless, Jerald F. *Statistical Models and Methods for Lifetime Data*. 2nd ed. Wiley Series in Probability and Statistics. Hoboken, N.J: Wiley-Interscience, 2003.

[5] Meeker, William Q., and Luis A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Applied Probability and Statistics Section. New York: Wiley, 1998.

See Also

`WeibullDistribution` | `wblcdf` | `wblfit` | `wblinv` | `wbllike` | `wblpdf` | `wblplot` | `wblrnd` | `wblstat`

More About

- “Fit Probability Distribution Objects to Grouped Data” on page 5-92
- “Compare Multiple Distribution Fits” on page 5-87
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Wishart Distribution

In this section...

Overview

The Wishart distribution is a generalization of the univariate chi-square distribution to two or more variables. It is a distribution for symmetric positive semidefinite matrices, typically covariance matrices, the diagonal elements of which are each chi-square random variables. In the same way as the chi-square distribution can be constructed by summing the squares of independent, identically distributed, zero-mean univariate normal random variables, the Wishart distribution can be constructed by summing the inner products of independent, identically distributed, zero-mean multivariate normal random vectors. The Wishart distribution is often used as a model for the distribution of the sample covariance matrix for multivariate normal random data, after scaling by the sample size.

Only random matrix generation is supported for the Wishart distribution, including both singular and nonsingular Σ.

Parameters

The Wishart distribution is parameterized with a symmetric, positive semidefinite matrix, Σ, and a positive scalar degrees of freedom parameter, ν. ν is analogous to the degrees of freedom parameter of a univariate chi-square distribution, and Σν is the mean of the distribution.

Probability Density Function

The probability density function of the *d*-dimensional Wishart distribution is given by

$$y = f(X, \Sigma, \nu) = \frac{|X|^{((\nu\text{-}d\text{-}1)/2)} e^{\left(-\frac{1}{2}\text{trace}\left(\Sigma^{-1}X\right)\right)}}{2^{(\nu d)/2}\pi^{(d(d-1))/4}|\sum|^{\nu/2}\Gamma(\nu/2)\ldots\Gamma(\nu\text{-}(d\text{-}1))/2}$$

where X and Σ are d-by-d symmetric positive definite matrices, and ν is a scalar greater than d – 1. While it is possible to define the Wishart for singular Σ, the density cannot be written as above.

Example

If x is a bivariate normal random vector with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & .5 \\ .5 & 2 \end{pmatrix}$$

then you can use the Wishart distribution to generate a sample covariance matrix without explicitly generating x itself. Notice how the sampling variability is quite large when the degrees of freedom is small.

```
Sigma = [1 .5; .5 2];
df = 10; S1 = wishrnd(Sigma,df)/df

S1 =
       1.7959      0.64107
      0.64107       1.5496

df = 1000; S2 = wishrnd(Sigma,df)/df

S2 =
       0.9842      0.50158
      0.50158       2.1682
```

See Also

`wishrnd`

More About

- “Inverse Wishart Distribution” on page B-76
- “Working with Probability Distributions” on page 5-3
- “Supported Distributions” on page 5-14

Bibliography

Bibliography

[1] Aas, Kjersti, Martin. Jullum, and Anders Løland. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *arXiv:1903.10464* (2019).

[2] Atkinson, A. C., and A. N. Donev. *Optimum Experimental Designs*. New York: Oxford University Press, 1992.

[3] Bates, D. M., and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. Hoboken, NJ: John Wiley & Sons, Inc., 1988.

[4] Belsley, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics*. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

[5] Berry, M. W., et al. "Algorithms and Applications for Approximate Nonnegative Matrix Factorization." *Computational Statistics and Data Analysis*. Vol. 52, No. 1, 2007, pp. 155–173.

[6] Bookstein, Fred L. *Morphometric Tools for Landmark Data*. Cambridge, UK: Cambridge University Press, 1991.

[7] Bouye, E., V. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli. "Copulas for Finance: A Reading Guide and Some Applications." Working Paper. Groupe de Recherche Operationnelle, Credit Lyonnais, 2000.

[8] Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis*. New York: Oxford University Press, 1997.

[9] Box, G. E. P., and N. R. Draper. *Empirical Model-Building and Response Surfaces*. Hoboken, NJ: John Wiley & Sons, Inc., 1987.

[10] Box, G. E. P., W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. Hoboken, NJ: Wiley-Interscience, 1978.

[11] Bratley, P., and B. L. Fox. "ALGORITHM 659 Implementing Sobol's Quasirandom Sequence Generator." *ACM Transactions on Mathematical Software*. Vol. 14, No. 1, 1988, pp. 88–100.

[12] Breiman, L. "Random Forests." *Machine Learning*. Vol. 4, 2001, pp. 5–32.

[13] Breiman, L., J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.

[14] Bulmer, M. G. *Principles of Statistics*. Mineola, NY: Dover Publications, Inc., 1979.

[15] Bury, K.. *Statistical Distributions in Engineering*. Cambridge, UK: Cambridge University Press, 1999.

[16] Chatterjee, S., and A. S. Hadi. "Influential Observations, High Leverage Points, and Outliers in Linear Regression." *Statistical Science*. Vol. 1, 1986, pp. 379–416.

[17] Collett, D. *Modeling Binary Data*. New York: Chapman & Hall, 2002.

[18] Conover, W. J. *Practical Nonparametric Statistics*. Hoboken, NJ: John Wiley & Sons, Inc., 1980.

[19] Cook, R. D., and S. Weisberg. *Residuals and Influence in Regression*. New York: Chapman & Hall/CRC Press, 1983.

[20] Cox, D. R., and D. Oakes. *Analysis of Survival Data*. London: Chapman & Hall, 1984.

[21] Davidian, M., and D. M. Giltinan. *Nonlinear Models for Repeated Measurements Data*. New York: Chapman & Hall, 1995.

[22] Deb, P., and M. Sefton. "The Distribution of a Lagrange Multiplier Test of Normality." *Economics Letters*. Vol. 51, 1996, pp. 123–130.

[23] de Jong, S. "SIMPLS: An Alternative Approach to Partial Least Squares Regression." *Chemometrics and Intelligent Laboratory Systems*. Vol. 18, 1993, pp. 251–263.

[24] Demidenko, E. *Mixed Models: Theory and Applications*. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

[25] Delyon, B., M. Lavielle, and E. Moulines, *Convergence of a stochastic approximation version of the EM algorithm*, Annals of Statistics, 27, 94-128, 1999.

[26] Dempster, A. P., N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society*. Series B, Vol. 39, No. 1, 1977, pp. 1–37.

[27] Devroye, L. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag, 1986.

[28] Dobson, A. J. *An Introduction to Generalized Linear Models*. New York: Chapman & Hall, 1990.

[29] Dunn, O.J., and V.A. Clark. *Applied Statistics: Analysis of Variance and Regression*. New York: Wiley, 1974.

[30] Draper, N. R., and H. Smith. *Applied Regression Analysis*. Hoboken, NJ: Wiley-Interscience, 1998.

[31] Drezner, Z. "Computation of the Trivariate Normal Integral." *Mathematics of Computation*. Vol. 63, 1994, pp. 289–294.

[32] Drezner, Z., and G. O. Wesolowsky. "On the Computation of the Bivariate Normal Integral." *Journal of Statistical Computation and Simulation*. Vol. 35, 1989, pp. 101–107.

[33] DuMouchel, W. H., and F. L. O'Brien. "Integrating a Robust Option into a Multiple Regression Computing Environment." *Computer Science and Statistics*: *Proceedings of the 21st Symposium on the Interface*. Alexandria, VA: American Statistical Association, 1989.

[34] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press, 1998.

[35] Efron, B., and R. J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[36] Embrechts, P., C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. New York: Springer, 1997.

[37] Evans, M., N. Hastings, and B. Peacock. *Statistical Distributions*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 1993, pp. 50–52, 73–74, 102–105, 147, 148.

[38] Friedman, J. H. "Greedy function approximation: a gradient boosting machine." *The Annals of Statistics*. Vol. 29, No. 5, 2001, pp. 1189-1232.

[39] Genz, A. “Numerical Computation of Rectangular Bivariate and Trivariate Normal and t Probabilities.” *Statistics and Computing*. Vol. 14, No. 3, 2004, pp. 251–260.

[40] Genz, A., and F. Bretz. “Comparison of Methods for the Computation of Multivariate t Probabilities.” *Journal of Computational and Graphical Statistics*. Vol. 11, No. 4, 2002, pp. 950–971.

[41] Genz, A., and F. Bretz. “Numerical Computation of Multivariate t Probabilities with Application to Power Calculation of Multiple Contrasts.” *Journal of Statistical Computation and Simulation*. Vol. 63, 1999, pp. 361–378.

[42] Gibbons, J. D. *Nonparametric Statistical Inference*. New York: Marcel Dekker, 1985.

[43] Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.” *Journal of Computational and Graphical Statistics*. Vol. 24, No. 1, 2015, pp. 44-65.

[44] Goodall, C. R. “Computation Using the QR Decomposition.” *Handbook in Statistics.* Vol. 9, Amsterdam: Elsevier/North-Holland, 1993.

[45] Goodnight, J.H., and F.M. Speed. *Computing Expected Mean Squares*. Cary, NC: SAS Institute, 1978.

[46] Hahn, Gerald J., and S. S. Shapiro. *Statistical Models in Engineering*. Hoboken, NJ: John Wiley & Sons, Inc., 1994, p. 95.

[47] Hald, A. *Statistical Theory with Engineering Applications*. Hoboken, NJ: John Wiley & Sons, Inc., 1960.

[48] Harman, H. H. *Modern Factor Analysis*. 3rd Ed. Chicago: University of Chicago Press, 1976.

[49] Hastie, T., R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. New York: Springer, 2001.

[50] Hill, P. D. “Kernel estimation of a distribution function.” *Communications in Statistics – Theory and Methods*. Vol. 14, Issue 3, 1985, pp. 605-620.

[51] Hochberg, Y., and A. C. Tamhane. *Multiple Comparison Procedures*. Hoboken, NJ: John Wiley & Sons, 1987.

[52] Hoerl, A. E., and R. W. Kennard. “Ridge Regression: Applications to Nonorthogonal Problems.” *Technometrics*. Vol. 12, No. 1, 1970, pp. 69–82.

[53] Hoerl, A. E., and R. W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*. Vol. 12, No. 1, 1970, pp. 55–67.

[54] Hogg, R. V., and J. Ledolter. *Engineering Statistics*. New York: MacMillan, 1987.

[55] Holland, P. W., and R. E. Welsch. “Robust Regression Using Iteratively Reweighted Least-Squares.” *Communications in Statistics: Theory and Methods*, *A6*, 1977, pp. 813–827.

[56] Hollander, M., and D. A. Wolfe. *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc., 1999.

[57] Hong, H. S., and F. J. Hickernell. “ALGORITHM 823: Implementing Scrambled Digital Sequences.” *ACM Transactions on Mathematical Software*. Vol. 29, No. 2, 2003, pp. 95–109.

[58] Huang, P. S., H. Avron, and T. N. Sainath, V. Sindhwani, and B. Ramabhadran. “Kernel methods match Deep Neural Networks on TIMIT.” *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 205–209.

[59] Huber, P. J. *Robust Statistics*. Hoboken, NJ: John Wiley & Sons, Inc., 1981.

[60] Jackson, J. E. *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley and Sons, 1991.

[61] Jain, A., and R. Dubes. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

[62] Jarque, C. M., and A. K. Bera. “A test for normality of observations and regression residuals.” *International Statistical Review*. Vol. 55, No. 2, 1987, pp. 163–172.

[63] Joe, S., and F. Y. Kuo. “Remark on Algorithm 659: Implementing Sobol's Quasirandom Sequence Generator.” *ACM Transactions on Mathematical Software*. Vol. 29, No. 1, 2003, pp. 49–57.

[64] Johnson, N., and S. Kotz. *Distributions in Statistics: Continuous Univariate Distributions-2*. Hoboken, NJ: John Wiley & Sons, Inc., 1970, pp. 130–148, 189–200, 201–219.

[65] Johnson, N. L., N. Balakrishnan, and S. Kotz. *Continuous Multivariate Distributions*. Vol. 1. Hoboken, NJ: Wiley-Interscience, 2000.

[66] Johnson, N. L., S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Vol. 1, Hoboken, NJ: Wiley-Interscience, 1993.

[67] Johnson, N. L., S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Vol. 2, Hoboken, NJ: Wiley-Interscience, 1994.

[68] Johnson, N. L., S. Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. Hoboken, NJ: Wiley-Interscience, 1997.

[69] Johnson, N. L., S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. Hoboken, NJ: Wiley-Interscience, 1993.

[70] Jolliffe, I. T. *Principal Component Analysis*. 2nd ed., New York: Springer-Verlag, 2002.

[71] Jones, M.C. "Simple boundary correction for kernel density estimation." *Statistics and Computing*. Vol. 3, Issue 3, 1993, pp. 135-146.

[72] Jöreskog, K. G. "Some Contributions to Maximum Likelihood Factor Analysis." *Psychometrika*. Vol. 32, 1967, pp. 443–482.

[73] Kaufman L., and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, Inc., 1990.

[74] Kempka, Michał, Wojciech Kotłowski, and Manfred K. Warmuth. "Adaptive Scale-Invariant Online Algorithms for Learning Linear Models." *CoRR* (February 2019). https://arxiv.org/abs/1902.07528.

[75] Kendall, David G. "A Survey of the Statistical Theory of Shape." *Statistical Science*. Vol. 4, No. 2, 1989, pp. 87–99.

[76] Klein, J. P., and M. L. Moeschberger. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2003.

[77] Kleinbaum, D. G., and M. Klein. *Survival Analysis*. Statistics for Biology and Health. 2nd edition. Springer, 2005.

[78] Kocis, L., and W. J. Whiten. "Computational Investigations of Low-Discrepancy Sequences." *ACM Transactions on Mathematical Software*. Vol. 23, No. 2, 1997, pp. 266–294.

[79] Kotz, S., and S. Nadarajah. *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press, 2000.

[80] Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*. New York: Oxford University Press, 1988.

[81] Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Hoboken, NJ: Wiley-Interscience, 2002.

[82] Lawley, D. N., and A. E. Maxwell. *Factor Analysis as a Statistical Method*. 2nd ed. New York: American Elsevier Publishing, 1971.

[83] Le, Q., T. Sarlós, and A. Smola. "Fastfood — Approximating Kernel Expansions in Loglinear Time." *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28, No. 3, 2013, pp. 244–252.

[84] Lilliefors, H. W. "On the Kolmogorov-Smirnov test for normality with mean and variance unknown." *Journal of the American Statistical Association*. Vol. 62, 1967, pp. 399–402.

[85] Lilliefors, H. W. "On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown." *Journal of the American Statistical Association*. Vol. 64, 1969, pp. 387–389.

[86] Lindstrom, M. J., and D. M. Bates. "Nonlinear mixed-effects models for repeated measures data." *Biometrics*. Vol. 46, 1990, pp. 673–687.

[87] Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed., Hoboken, NJ: John Wiley & Sons, Inc., 2002.

[88] Lundberg, Scott M., and S. Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30 (2017): 4765–774.

[89] Mardia, K. V., J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Burlington, MA: Academic Press, 1980.

[90] Marquardt, D.W. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation." *Technometrics*. Vol. 12, No. 3, 1970, pp. 591–612.

[91] Marquardt, D. W., and R.D. Snee. "Ridge Regression in Practice." *The American Statistician*. Vol. 29, No. 1, 1975, pp. 3–20.

[92] Marsaglia, G., and W. W. Tsang. "A Simple Method for Generating Gamma Variables." *ACM Transactions on Mathematical Software*. Vol. 26, 2000, pp. 363–372.

[93] Marsaglia, G., W. Tsang, and J. Wang. "Evaluating Kolmogorov's Distribution." *Journal of Statistical Software*. Vol. 8, Issue 18, 2003.

[94] Martinez, W. L., and A. R. Martinez. *Computational Statistics with MATLAB*. New York: Chapman & Hall/CRC Press, 2002.

[95] Massey, F. J. “The Kolmogorov-Smirnov Test for Goodness of Fit.” *Journal of the American Statistical Association*. Vol. 46, No. 253, 1951, pp. 68–78.

[96] Matousek, J. “On the L2-Discrepancy for Anchored Boxes.” *Journal of Complexity*. Vol. 14, No. 4, 1998, pp. 527–556.

[97] McLachlan, G., and D. Peel. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons, Inc., 2000.

[98] McCullagh, P., and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[99] McGill, R., J. W. Tukey, and W. A. Larsen. “Variations of Boxplots.” *The American Statistician*. Vol. 32, No. 1, 1978, pp. 12–16.

[100] Meeker, W. Q., and L. A. Escobar. *Statistical Methods for Reliability Data*. Hoboken, NJ: John Wiley & Sons, Inc., 1998.

[101] Meng, Xiao-Li, and Donald B. Rubin. “Maximum Likelihood Estimation via the ECM Algorithm.” *Biometrika*. Vol. 80, No. 2, 1993, pp. 267–278.

[102] Meyers, R. H., and D.C. Montgomery. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Hoboken, NJ: John Wiley & Sons, Inc., 1995.

[103] Miller, L. H. “Table of Percentage Points of Kolmogorov Statistics.” *Journal of the American Statistical Association*. Vol. 51, No. 273, 1956, pp. 111–121.

[104] Milliken, G. A., and D. E. Johnson. *Analysis of Messy Data, Volume 1: Designed Experiments*. Boca Raton, FL: Chapman & Hall/CRC Press, 1992.

[105] Montgomery, D. *Introduction to Statistical Quality Control*. Hoboken, NJ: John Wiley & Sons, 1991, pp. 369–374.

[106] Montgomery, D. C. *Design and Analysis of Experiments*. Hoboken, NJ: John Wiley & Sons, Inc., 2001.

[107] Mood, A. M., F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* 3rd ed., New York: McGraw-Hill, 1974. pp. 540–541.

[108] Moore, J. *Total Biochemical Oxygen Demand of Dairy Manures*. Ph.D. thesis. University of Minnesota, Department of Agricultural Engineering, 1975.

[109] Mosteller, F., and J. Tukey. *Data Analysis and Regression*. Upper Saddle River, NJ: Addison-Wesley, 1977.

[110] Nelson, L. S. “Evaluating Overlapping Confidence Intervals.” *Journal of Quality Technology*. Vol. 21, 1989, pp. 140–141.

[111] Patel, J. K., C. H. Kapadia, and D. B. Owen. *Handbook of Statistical Distributions*. New York: Marcel Dekker, 1976.

[112] Pinheiro, J. C., and D. M. Bates. “Approximations to the log-likelihood function in the nonlinear mixed-effects model.” *Journal of Computational and Graphical Statistics*. Vol. 4, 1995, pp. 12–35.

[113] Rahimi, A., and B. Recht. “Random Features for Large-Scale Kernel Machines.” *Advances in Neural Information Processing Systems*. Vol 20, 2008, pp. 1177–1184.

[114] Rice, J. A. *Mathematical Statistics and Data Analysis*. Pacific Grove, CA: Duxbury Press, 1994.

[115] Rosipal, R., and N. Kramer. "Overview and Recent Advances in Partial Least Squares." *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop (SLSFS 2005), Revised Selected Papers (Lecture Notes in Computer Science 3940)*. Berlin, Germany: Springer-Verlag, 2006, pp. 34–51.

[116] Sachs, L. *Applied Statistics: A Handbook of Techniques*. New York: Springer-Verlag, 1984, p. 253.

[117] Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2015.

[118] Searle, S. R., F. M. Speed, and G. A. Milliken. "Population marginal means in the linear model: an alternative to least-squares means." *American Statistician*. 1980, pp. 216–221.

[119] Seber, G. A. F. and A. J. Lee. *Linear Regression Analysis*. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2003.

[120] Seber, G. A. F. *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons, Inc., 1984.

[121] Seber, G. A. F., and C. J. Wild. *Nonlinear Regression*. Hoboken, NJ: Wiley-Interscience, 2003.

[122] Sexton, Joe, and A. R. Swensen. "ECM Algorithms that Converge at the Rate of EM." *Biometrika*. Vol. 87, No. 3, 2000, pp. 651–662.

[123] Silverman, B.W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[124] Snedecor, G. W., and W. G. Cochran. *Statistical Methods*. Ames, IA: Iowa State Press, 1989.

[125] Spath, H. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*. Translated by J. Goldschmidt. New York: Halsted Press, 1985.

[126] Stein, M. "Large sample properties of simulations using latin hypercube sampling." *Technometrics*. Vol. 29, No. 2, 1987, pp. 143–151. Correction, Vol. 32, p. 367.

[127] Stephens, M. A. "Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables." *Journal of the Royal Statistical Society*. Series B, Vol. 32, No. 1, 1970, pp. 115–122.

[128] Street, J. O., R. J. Carroll, and D. Ruppert. "A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares." *The American Statistician*. Vol. 42, 1988, pp. 152–154.

[129] Student. "On the Probable Error of the Mean." *Biometrika*. Vol. 6, No. 1, 1908, pp. 1–25.

[130] Vellemen, P. F., and D. C. Hoaglin. *Application, Basics, and Computing of Exploratory Data Analysis*. Pacific Grove, CA: Duxbury Press, 1981.

[131] Weibull, W. "A Statistical Theory of the Strength of Materials." *Ingeniors Vetenskaps Akademiens Handlingar*. Stockholm: Royal Swedish Institute for Engineering Research, No. 151, 1939.

[132] Zahn, C. T. “Graph-theoretical methods for detecting and describing Gestalt clusters.” *IEEE Transactions on Computers*. Vol. C-20, Issue 1, 1971, pp. 68–86.